

Announcements

- **Homework 9** due **tomorrow, Nov 17** at 11:59pm PT
- **Project 5** due **Friday, Nov 27** at 11:59pm PT
- **Homework 10** will be released tomorrow
 - due **Friday, Dec 1** at 11:59pm PT
- Review session **next Tuesday, Nov 21**
- Special Topic Guest Lecture on **Nov 28!**

Nov 28 Special Topic Lecture on Generative AI

- **Guest Lecturer: Katherine Lee**

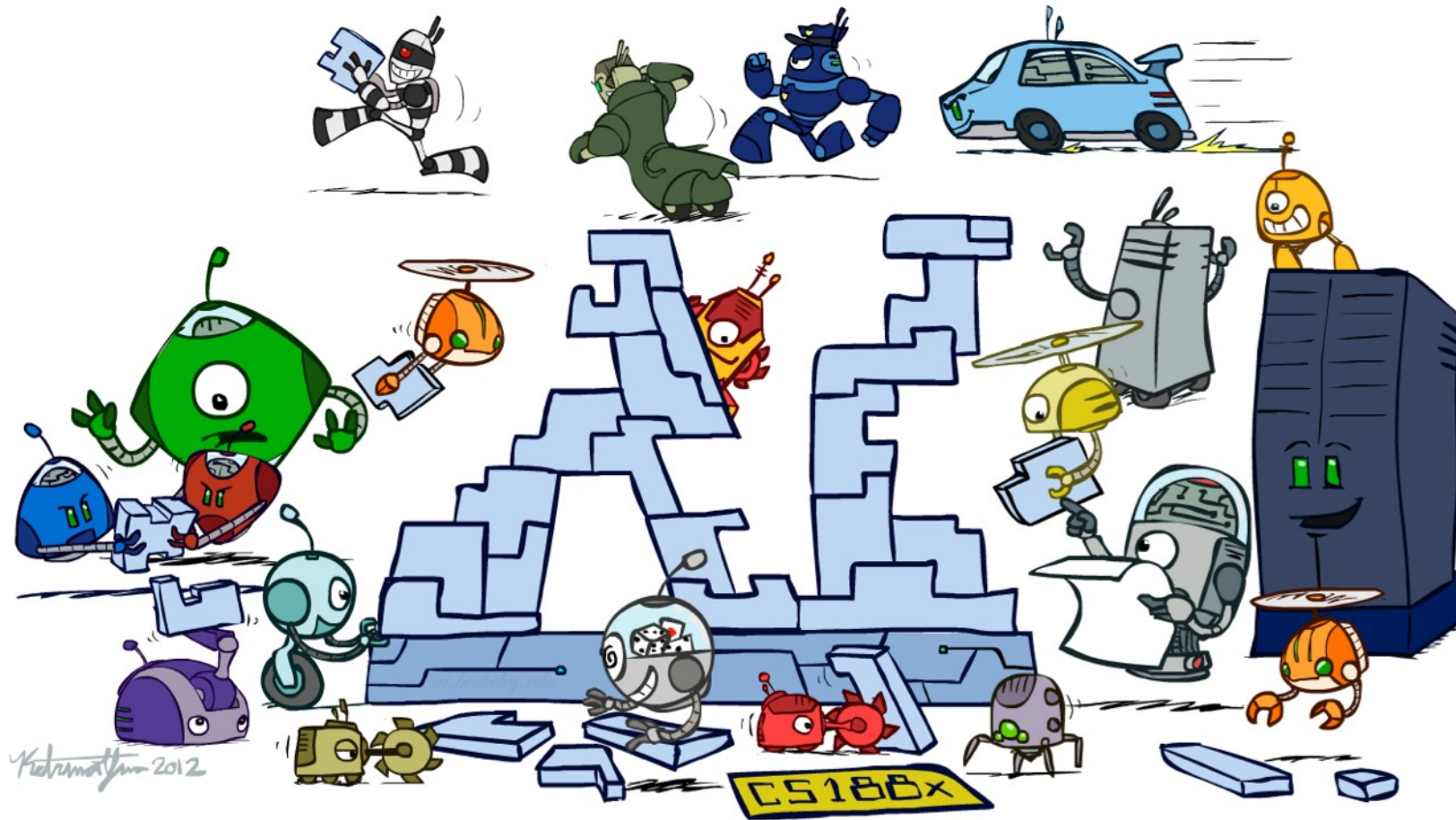
- Katherine is a research scientist at Google DeepMind. Her work has provided essential empirical evidence and measurement for grounding discussions around concerns that language models infringe copyright, and about how language models can respect an individuals' right to privacy and control of their data. Additionally, she has proposed methods of reducing memorization. Her work has received recognition at ACL, USENIX, and ICLR.

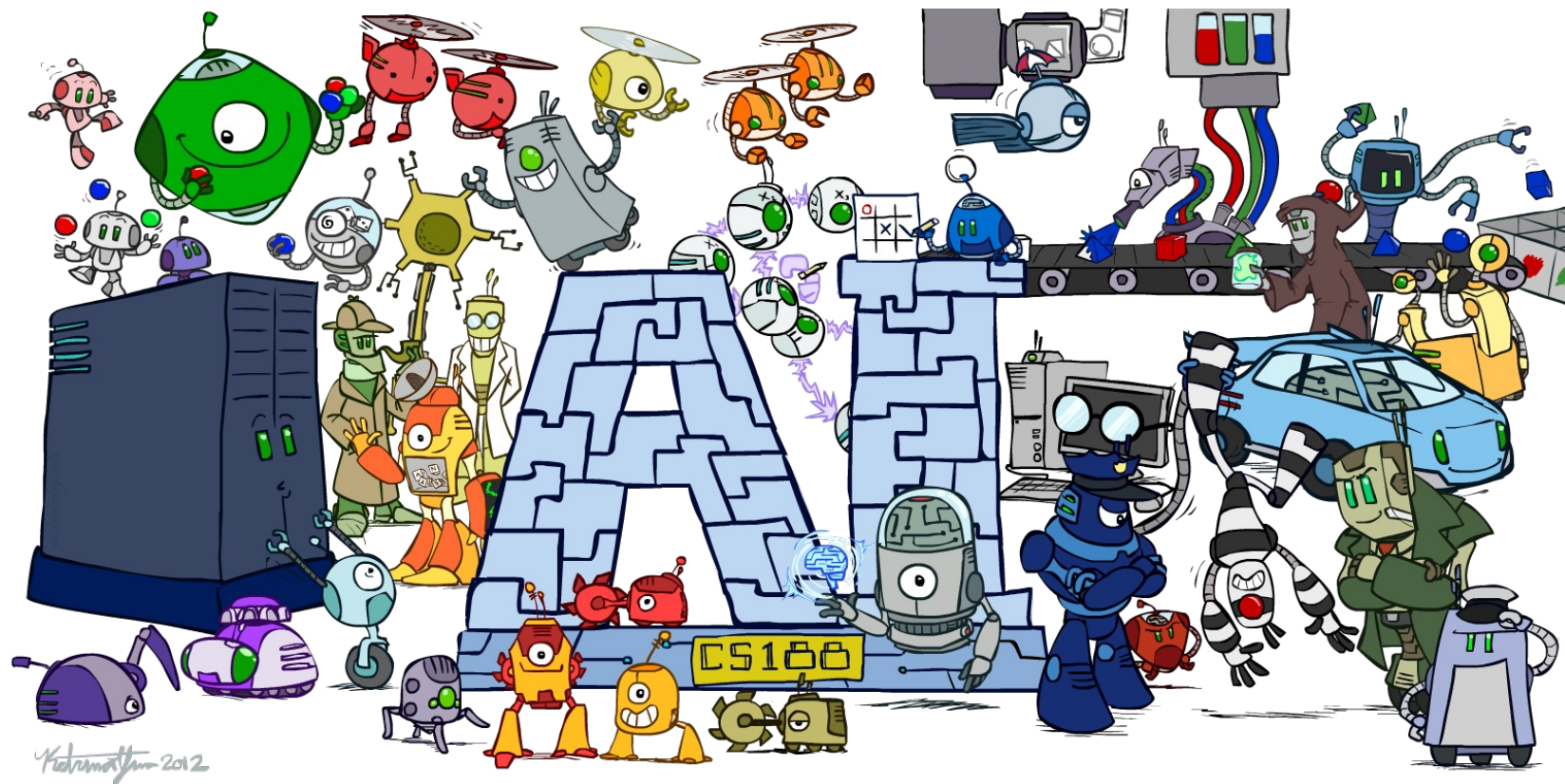
- katelee168.github.io

- genlaw.github.io

- **Topic:** Generative AI and its legal implications
- **Date:** Tuesday, Nov 28 (after Thanksgiving week)
- Please attend!

CS 188: Artificial Intelligence Applications





Ketrina Yim
CS188 Artist

Today

- Applications and analysis of supervised deep learning
- Intro to *unsupervised* learning
- Multi-modal AI
- Tracking and forecasting progress
- Frontier applications of AI
- AI ethics and regulation
- Where to go next

Recap of Supervised Learning with Neural Networks

Optimize probability of label given input $\max_w ll(w) = \max_w \sum_i \log P(y^{(i)} | x^{(i)}; w)$

Continuous optimization

Gradient ascent: take step in steepest uphill direction

Backpropagation computes gradient efficiently (out of scope)

Deep neural nets

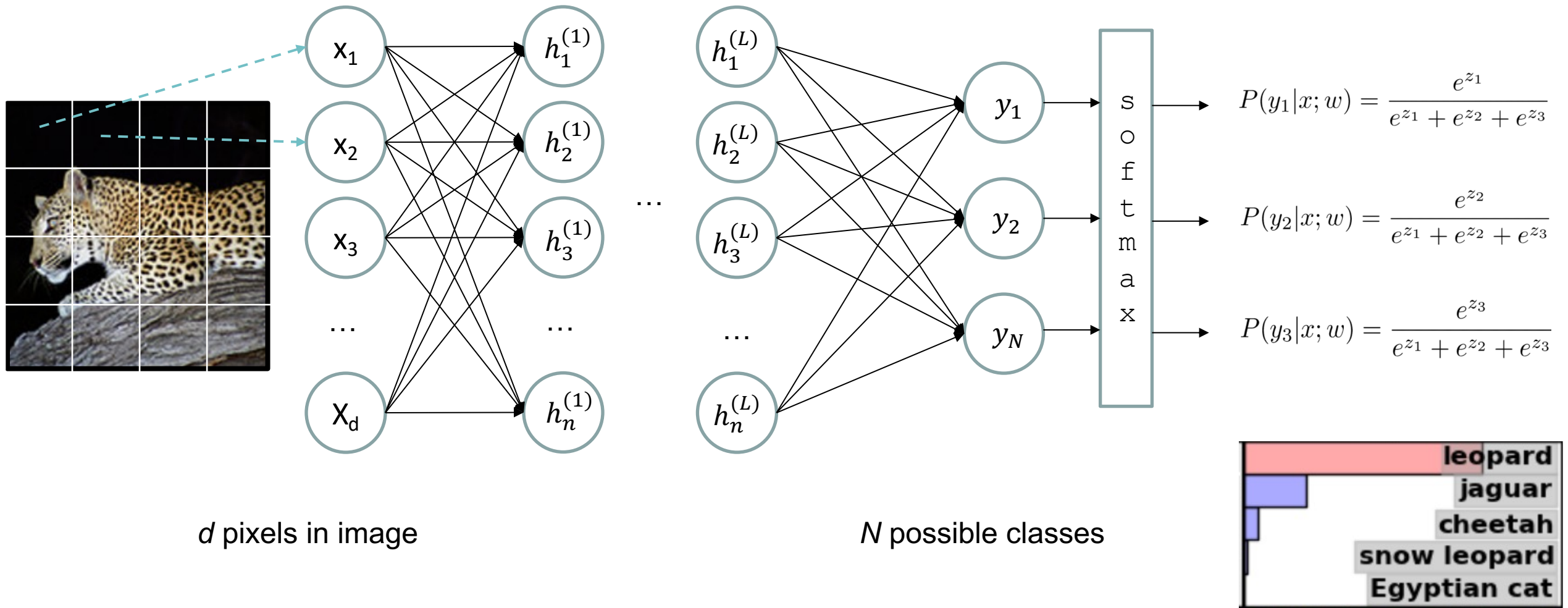
Last layer = logistic regression

Many layers before this last layer: learns “features” of input

Universal function approximation theorem:

a large enough 2 layer neural network can represent any continuous function with arbitrary accuracy

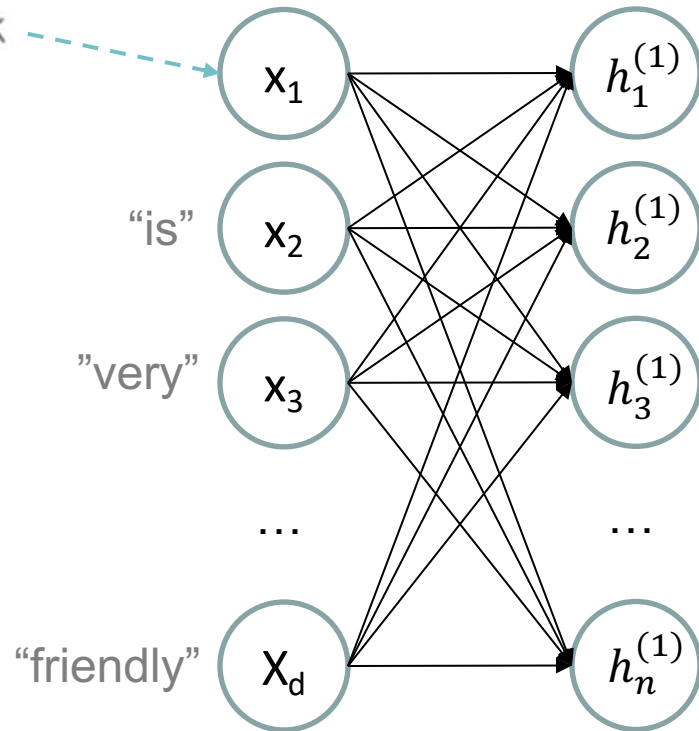
Deep Learning for Image Classification



Deep Learning for Language Generation

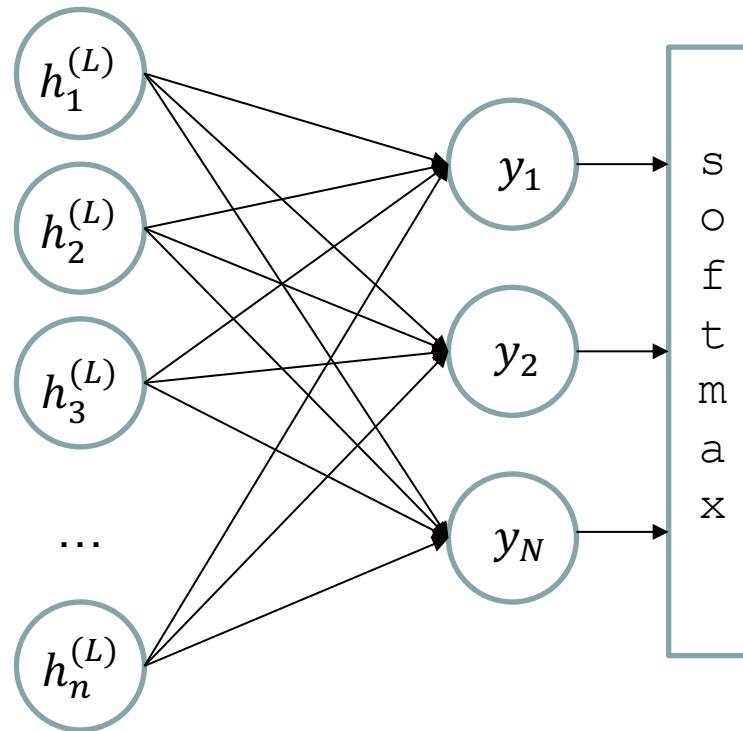
Dictionary:

aardvark
aarhus
aaron
...
...
...
...
zyzzyva



d words seen so far

...



N possible next words

$$P(y_1|x; w) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

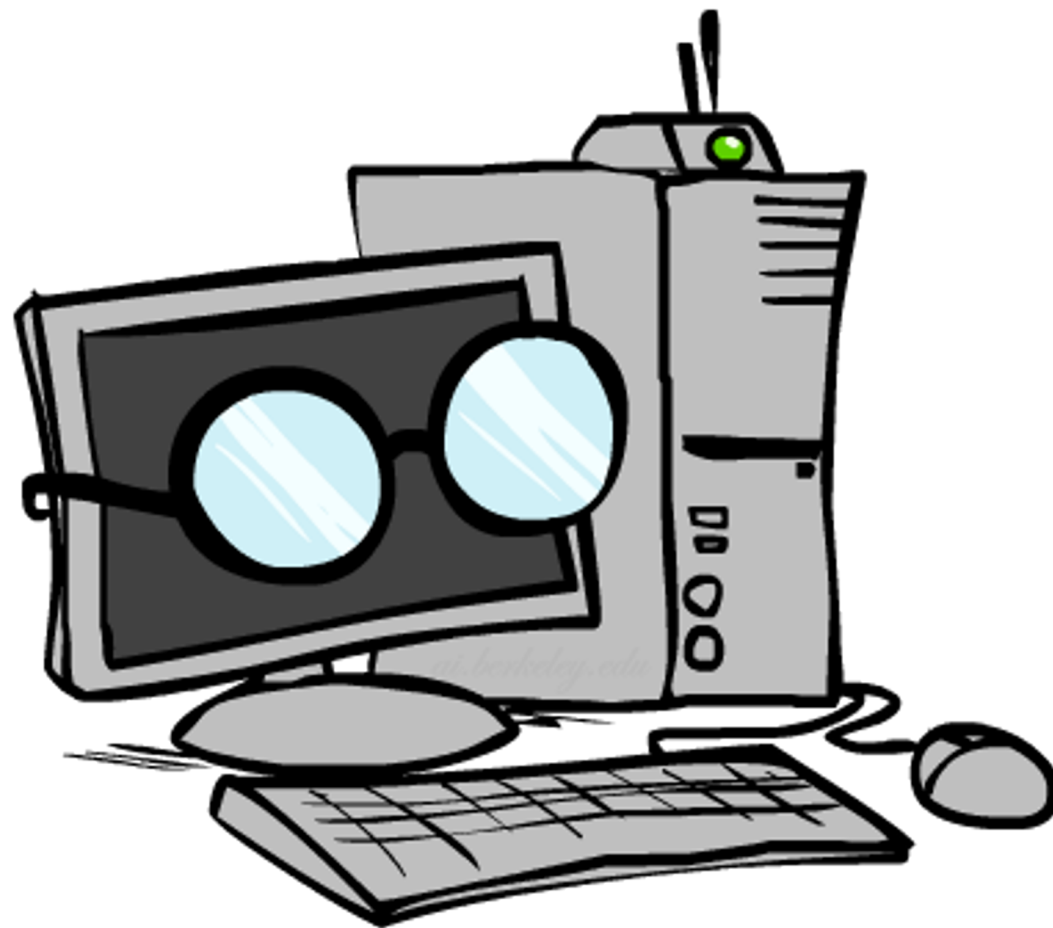
$$P(y_2|x; w) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$P(y_3|x; w) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

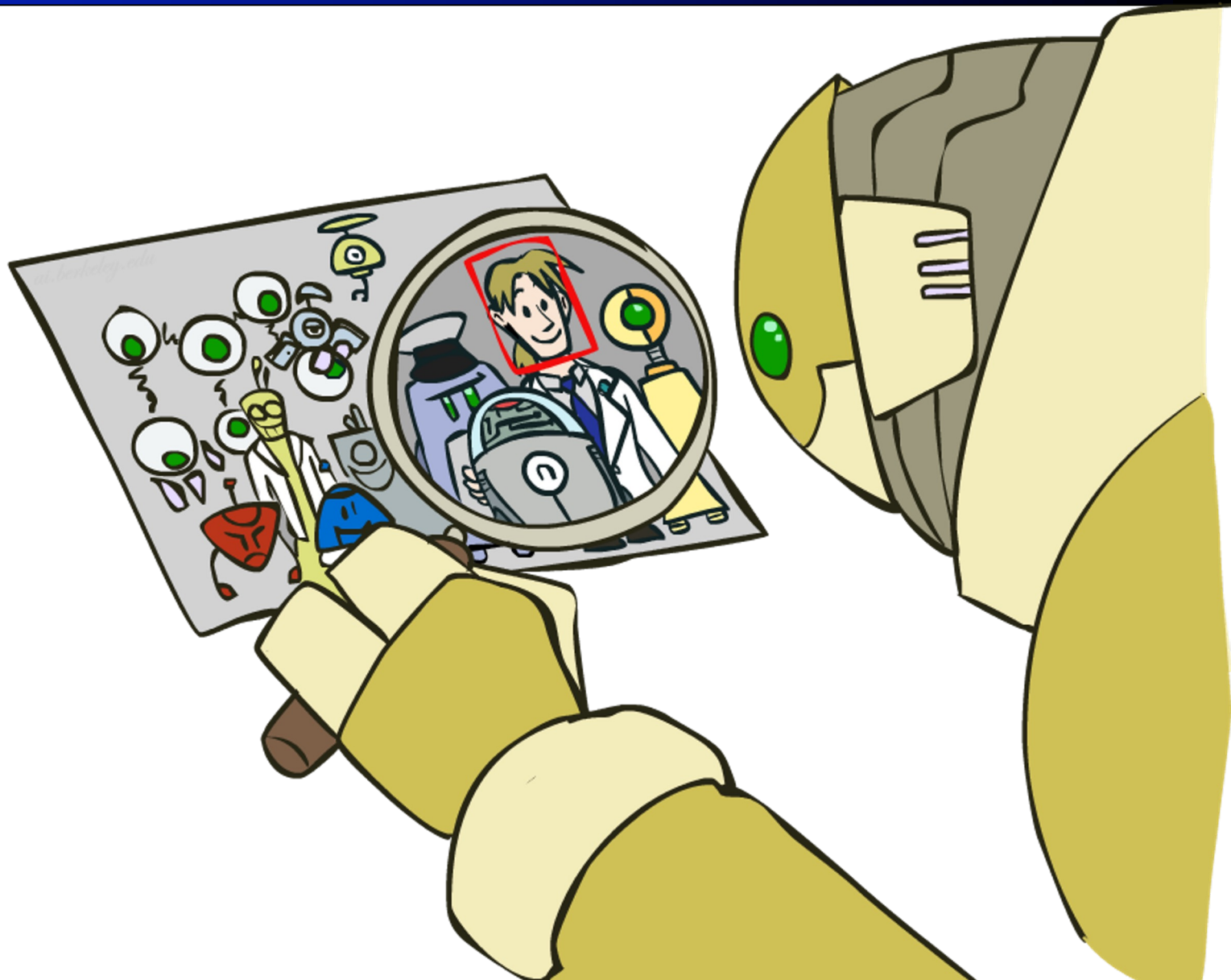
aardvark
aarhus
aaron
...
animal
...
...
zyzzyva

How well does deep learning work?

Computer Vision



Object Detection and Image Classification

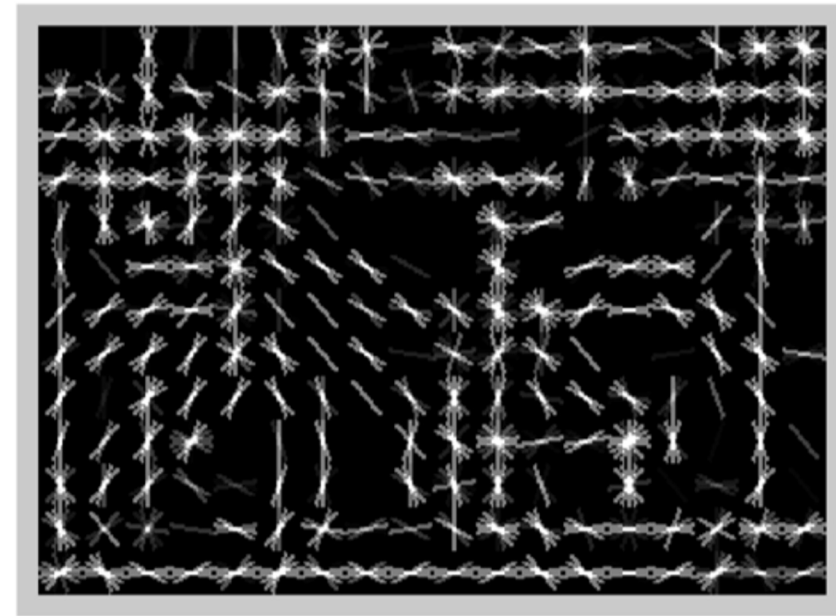


Traditional Computer Vision

- Manual Feature Design



Image



Histogram of Gradients (HoG)

Image Classification: Performance

ImageNet Error Rate 2010-2014

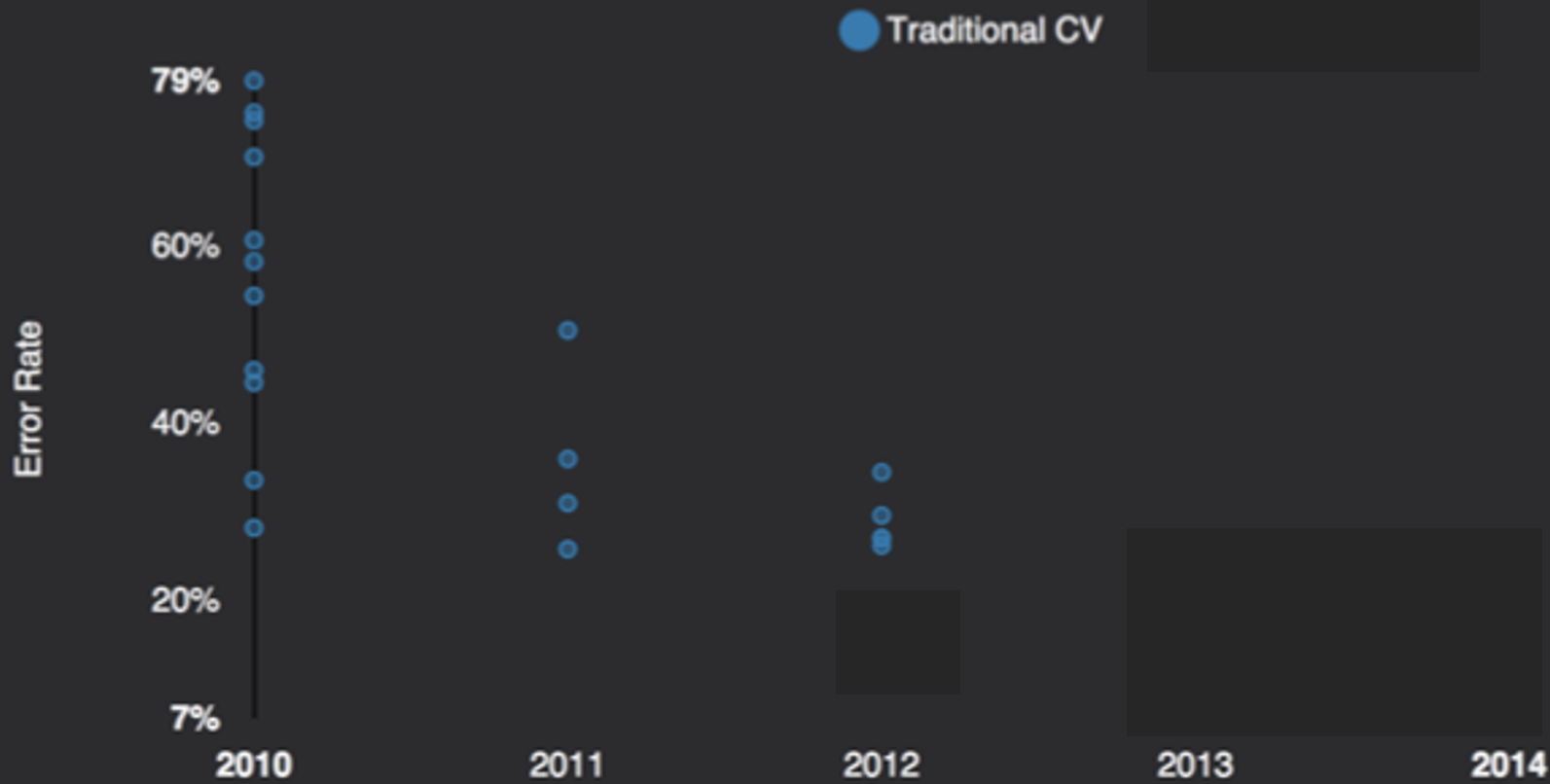


Image Classification: Performance

ImageNet Error Rate 2010-2014

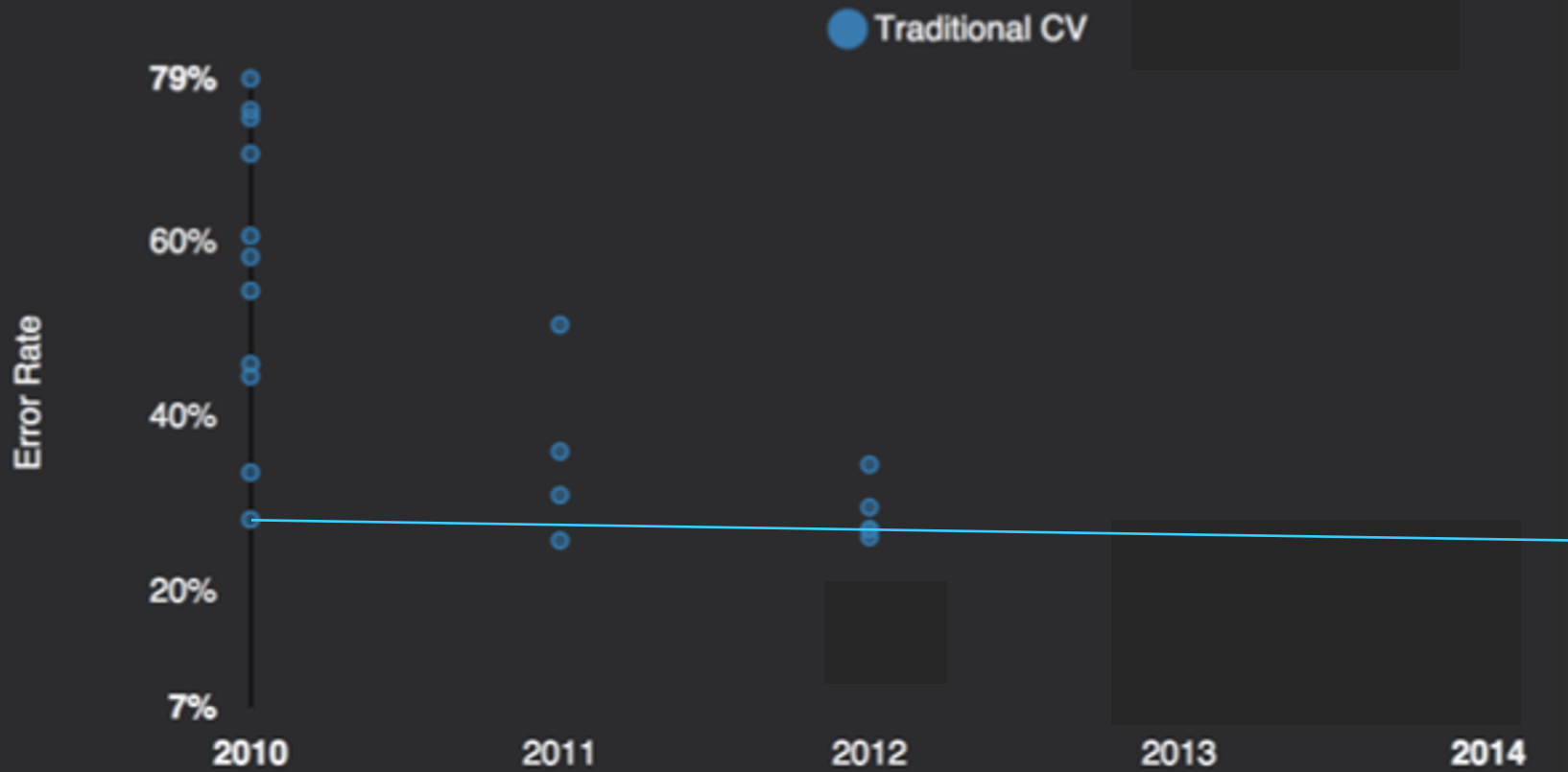


Image Classification: Performance

ImageNet Error Rate 2010-2014

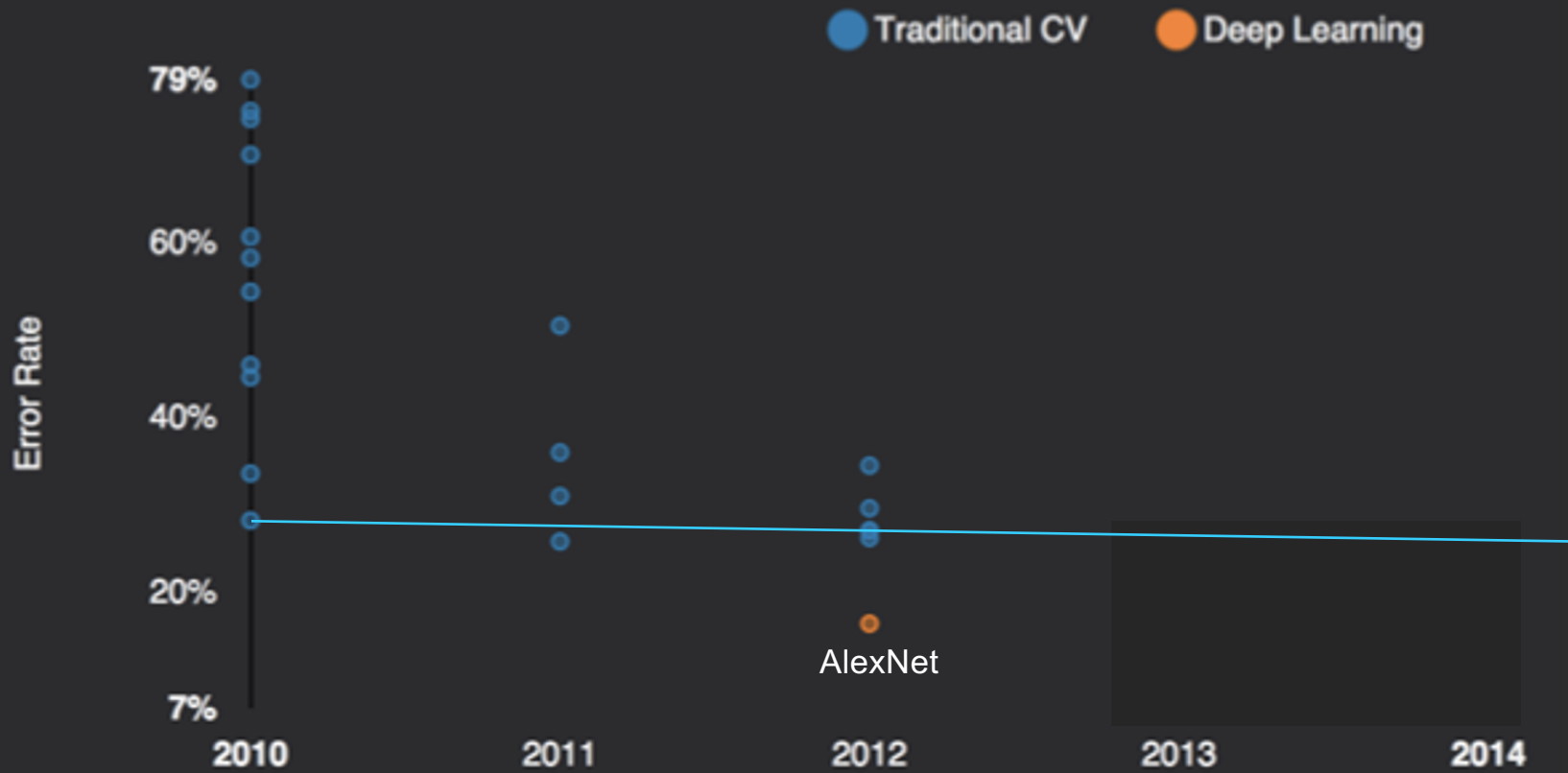
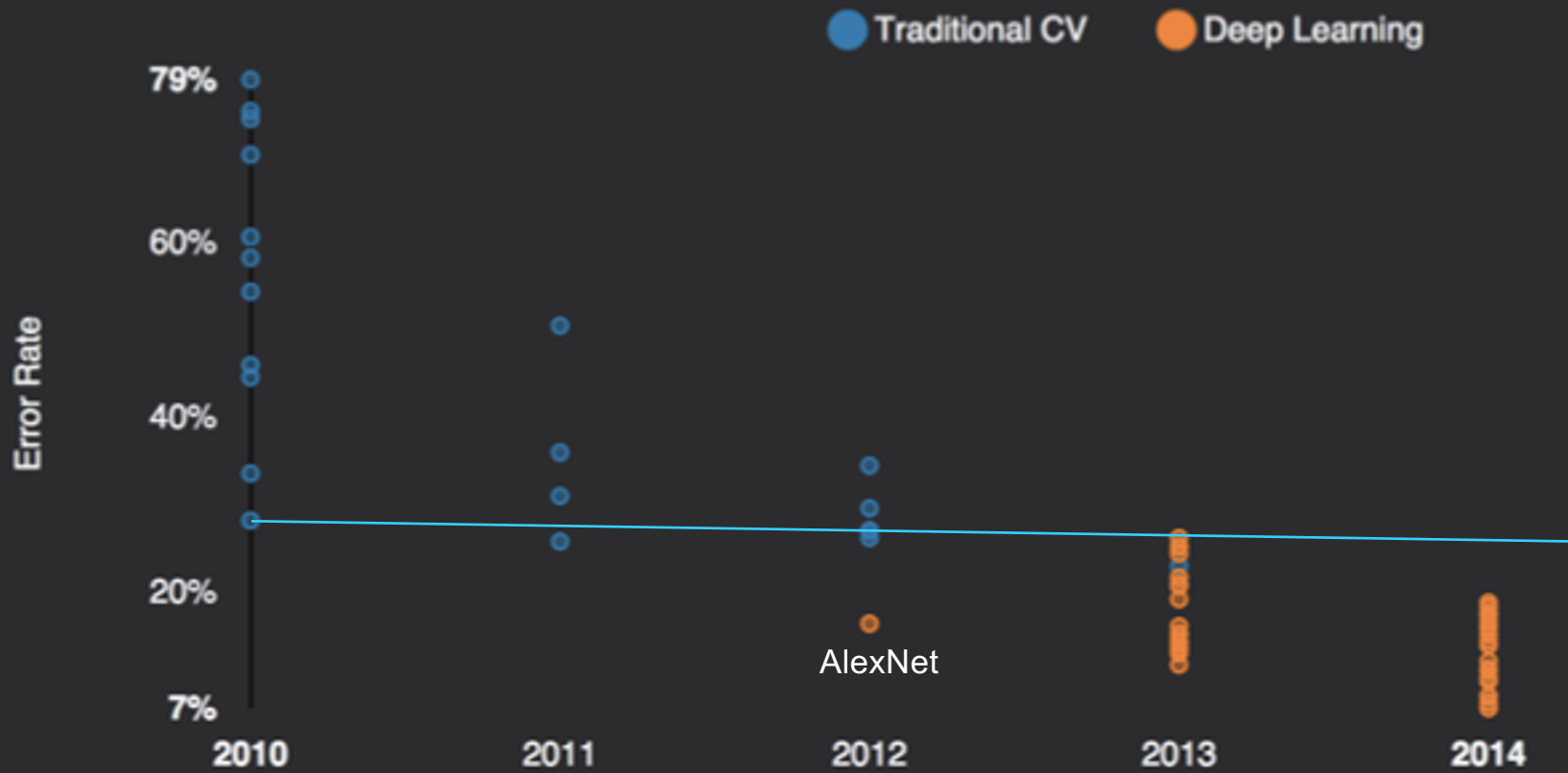


Image Classification: Performance

ImageNet Error Rate 2010-2014



graph credit Matt Zeiler, Clarifai

Image Classification: Performance

ImageNet Error Rate 2010-2014

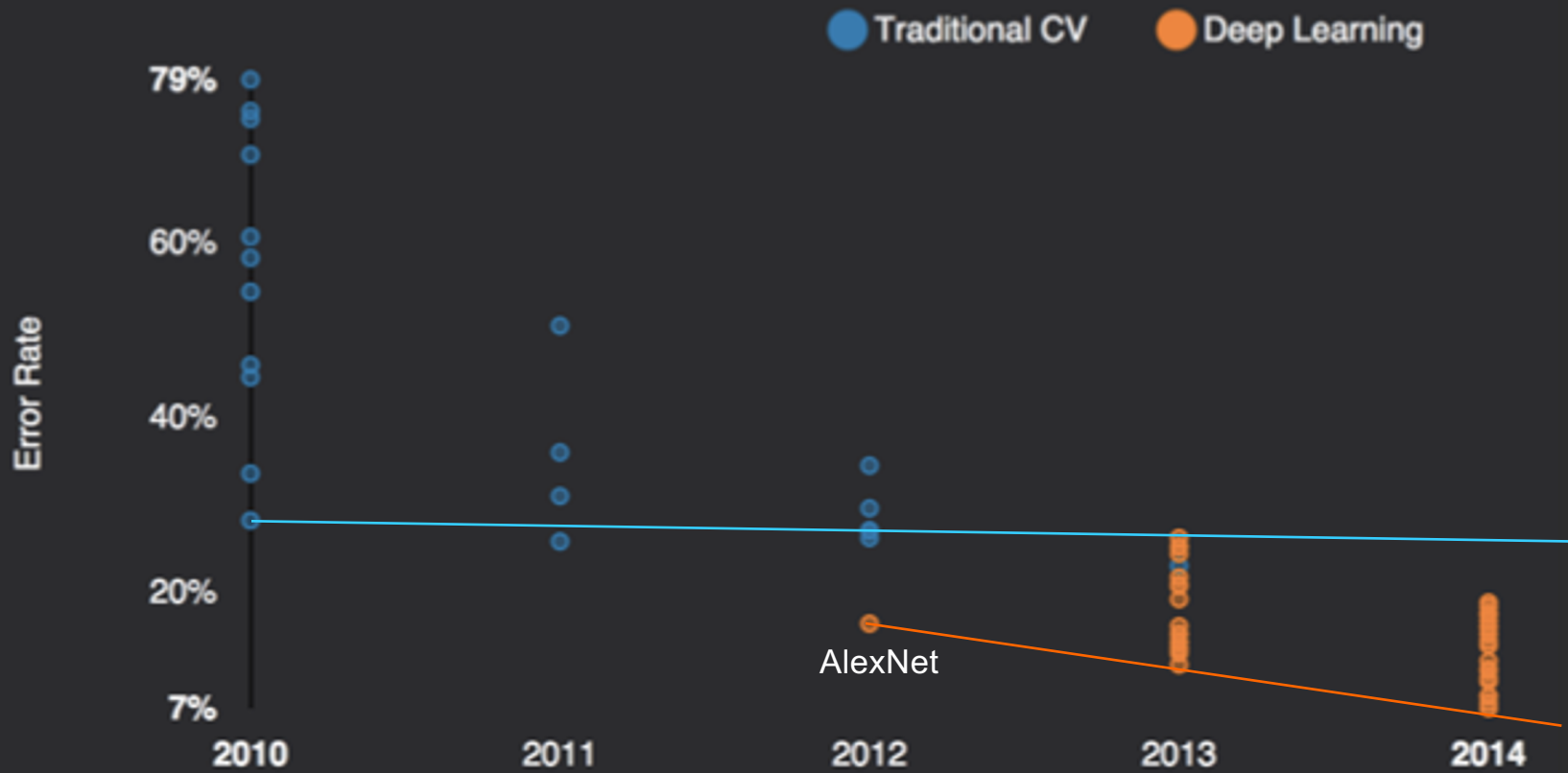


Image Classification: Performance

Leaderboard

Dataset

View

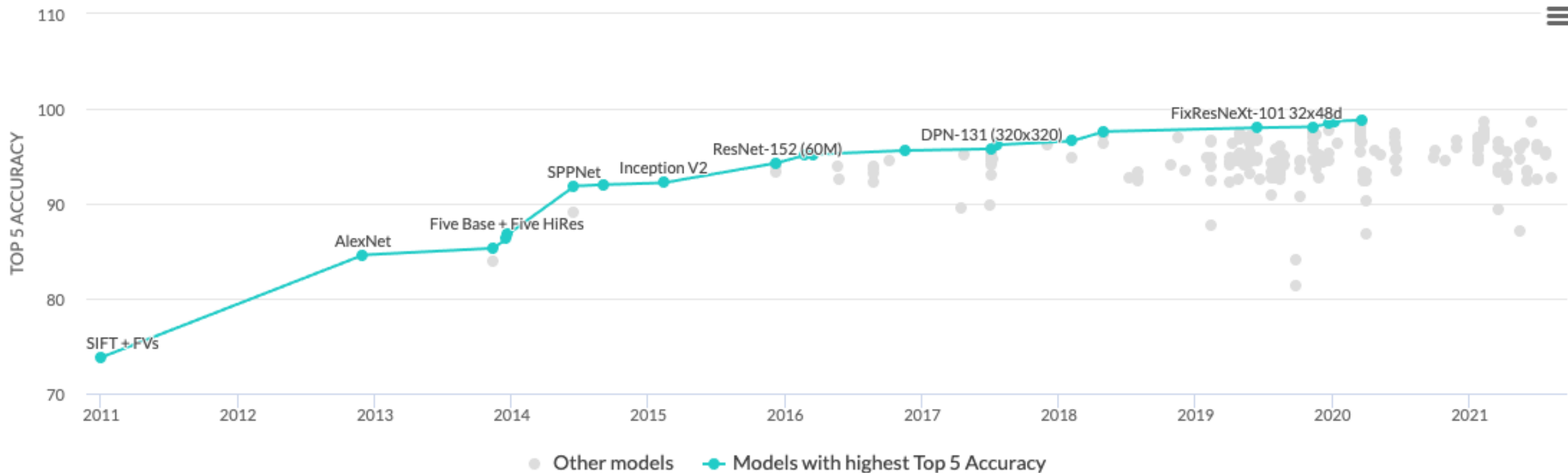
Top 5 Accuracy

by

Date

for

All models



Language Understanding: Performance

DATASET	METRIC	OUR RESULT	PREVIOUS RECORD	HUMAN
Winograd Schema Challenge	accuracy (+)	70.70%	63.7%	92%+
LAMBADA	accuracy (+)	63.24%	59.23%	95%+
LAMBADA	perplexity (-)	8.6	99	~1-2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	93.30%	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	89.05%	82.3%	92%
Penn Tree Bank	perplexity (-)	35.76	46.54	unknown
WikiText-2	perplexity (-)	18.34	39.14	unknown

Protein Prediction: AlphaFold

nature

View all Nature Research

Explore content ▾ Journal information ▾ Publish with us ▾ Subscribe

nature > news > article

NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

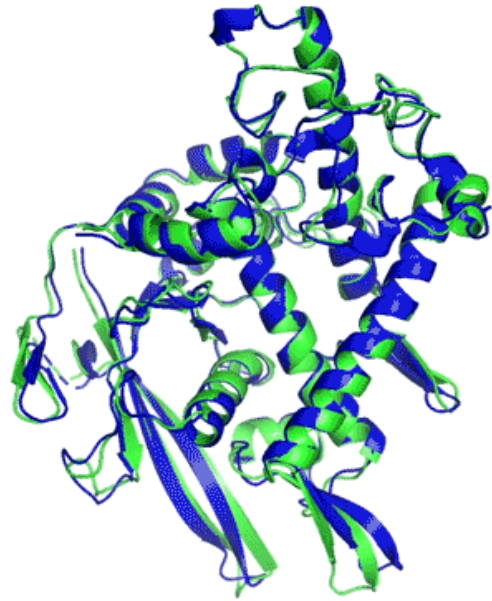
[Ewen Callaway](#)



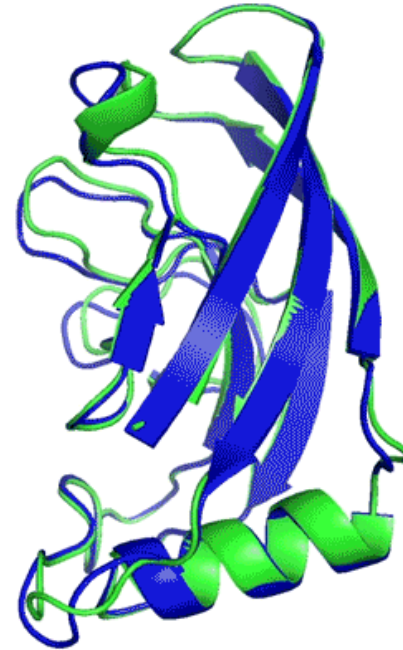
A protein's function is determined by its 3D shape. Credit: DeepMind

Credit: DeepMind

Protein Prediction: AlphaFold



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

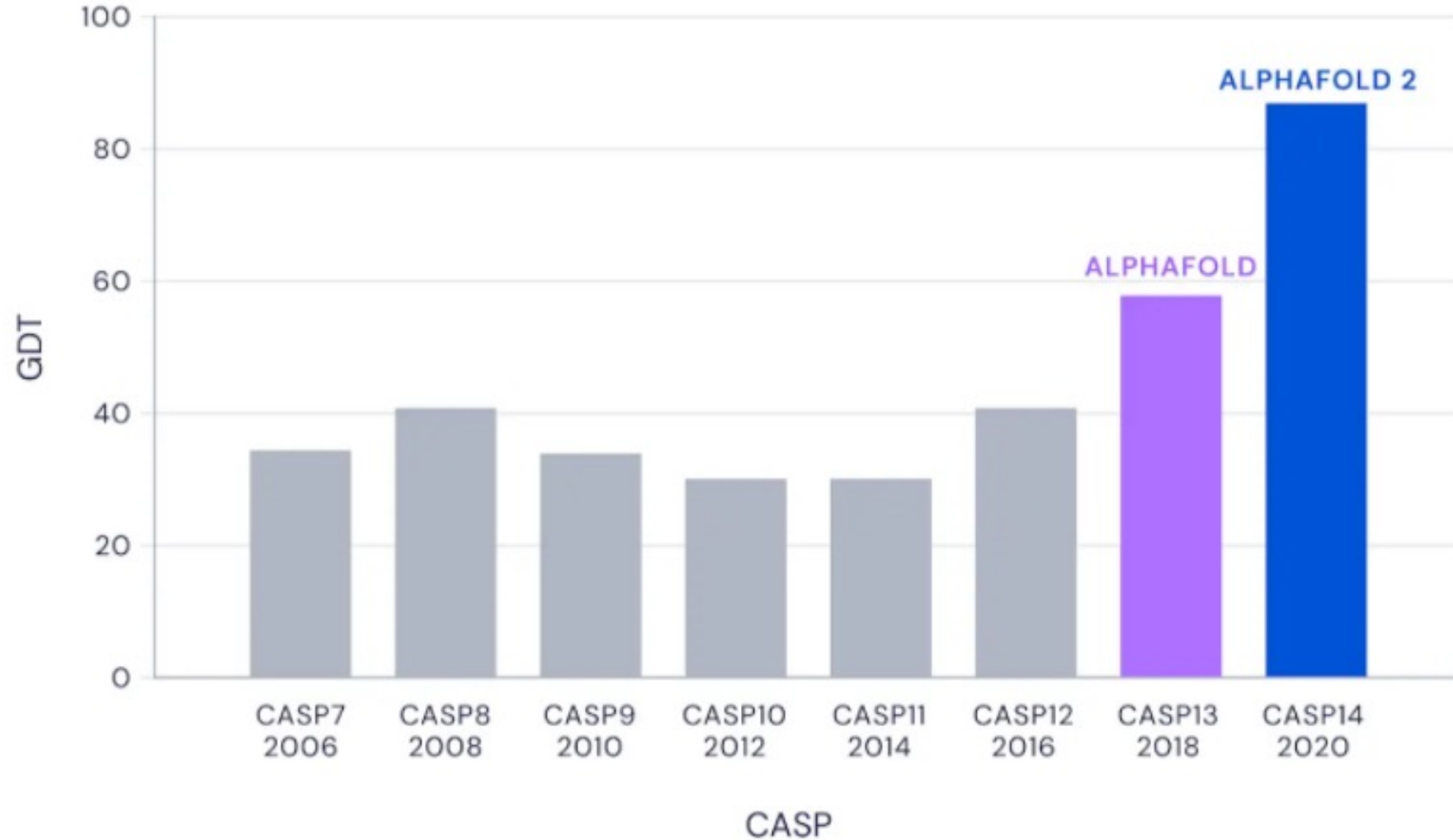


T1049 / 6y4f
93.3 GDT
(adhesin tip)

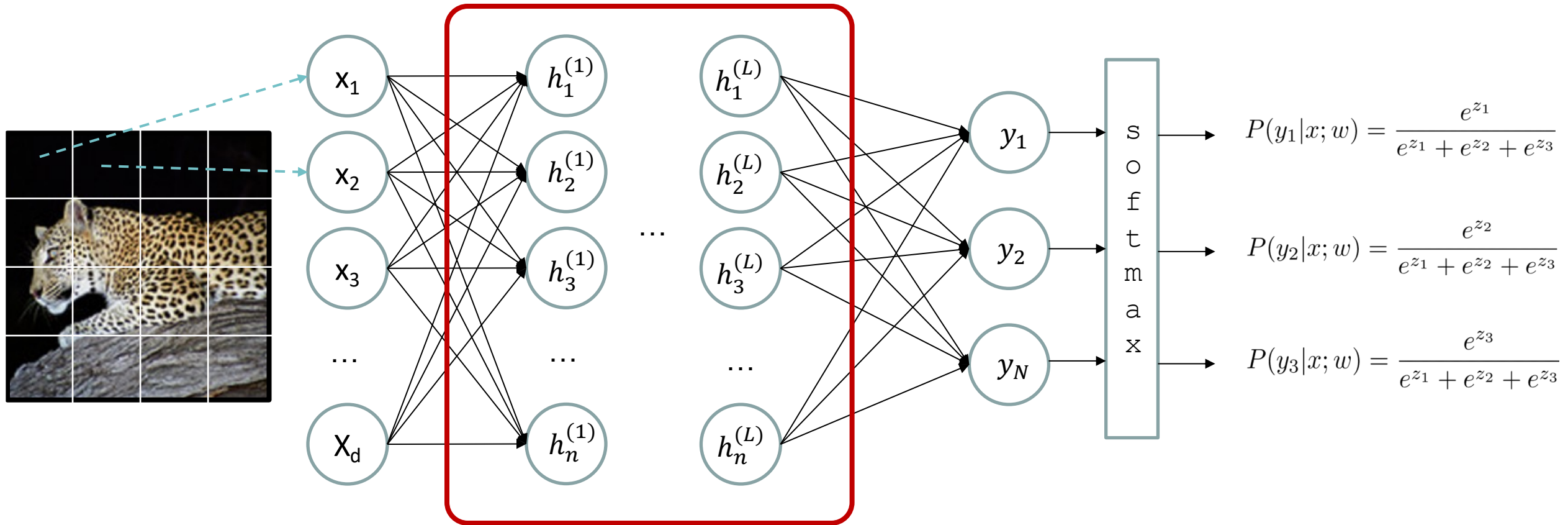
● Experimental result
● Computational prediction

AlphaFold at CASP 2020 Competition

Median Free-Modelling Accuracy



What do learned features look like?

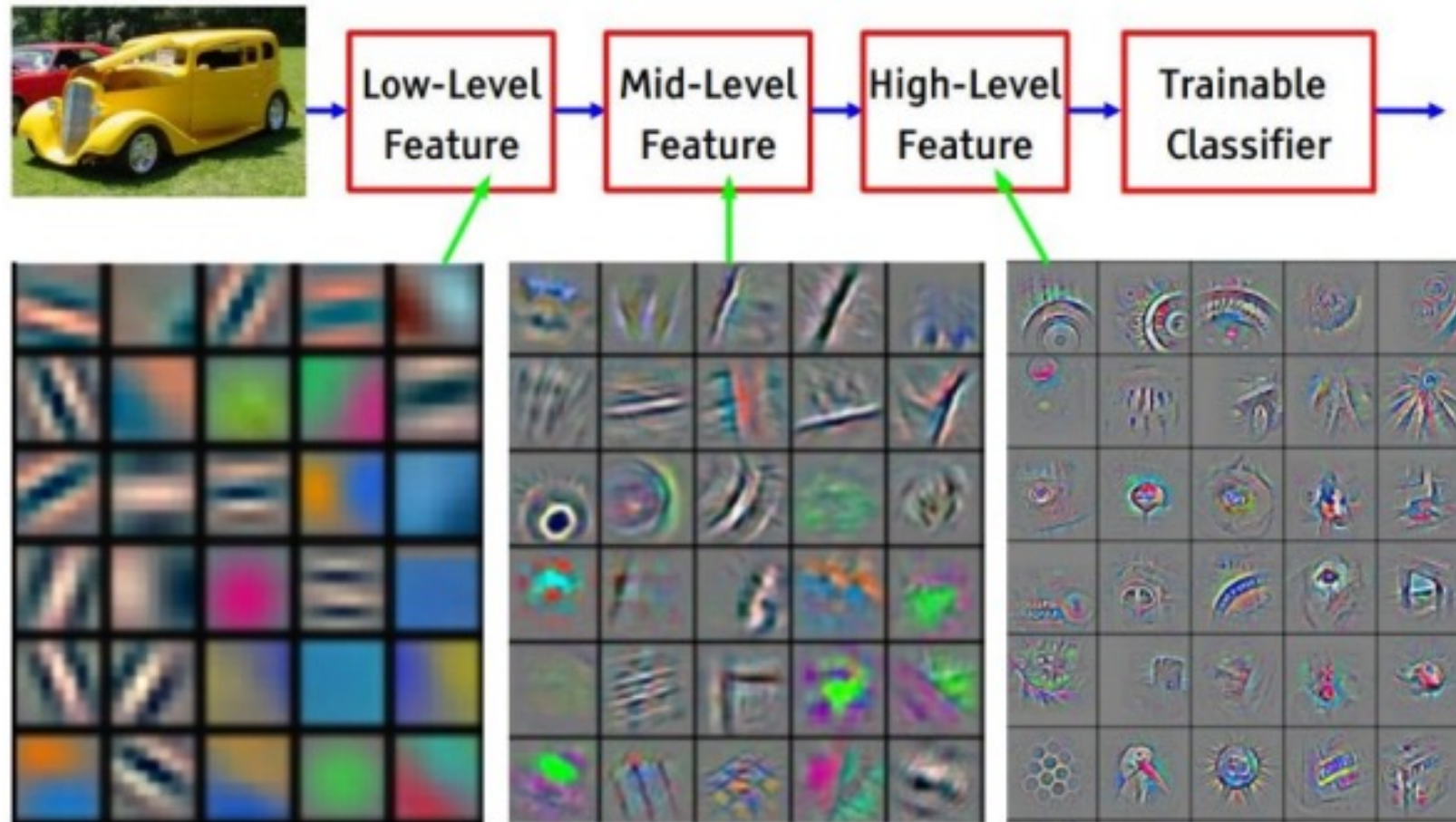


Hidden layers contain learned features

- Recall: neural networks learn features

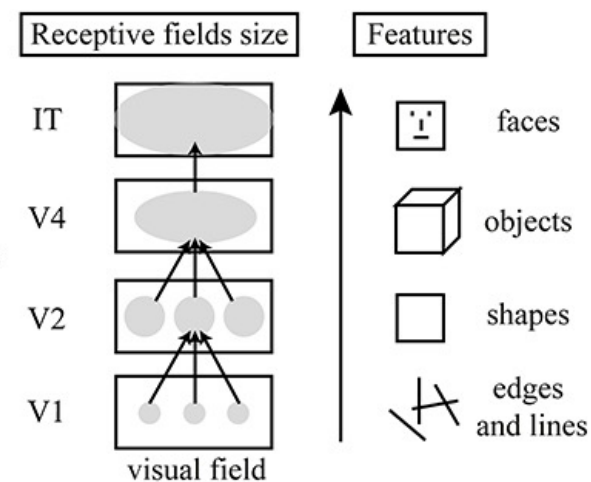
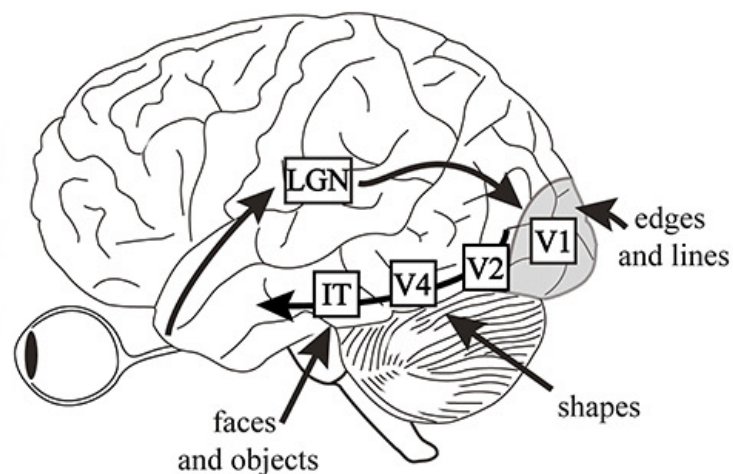
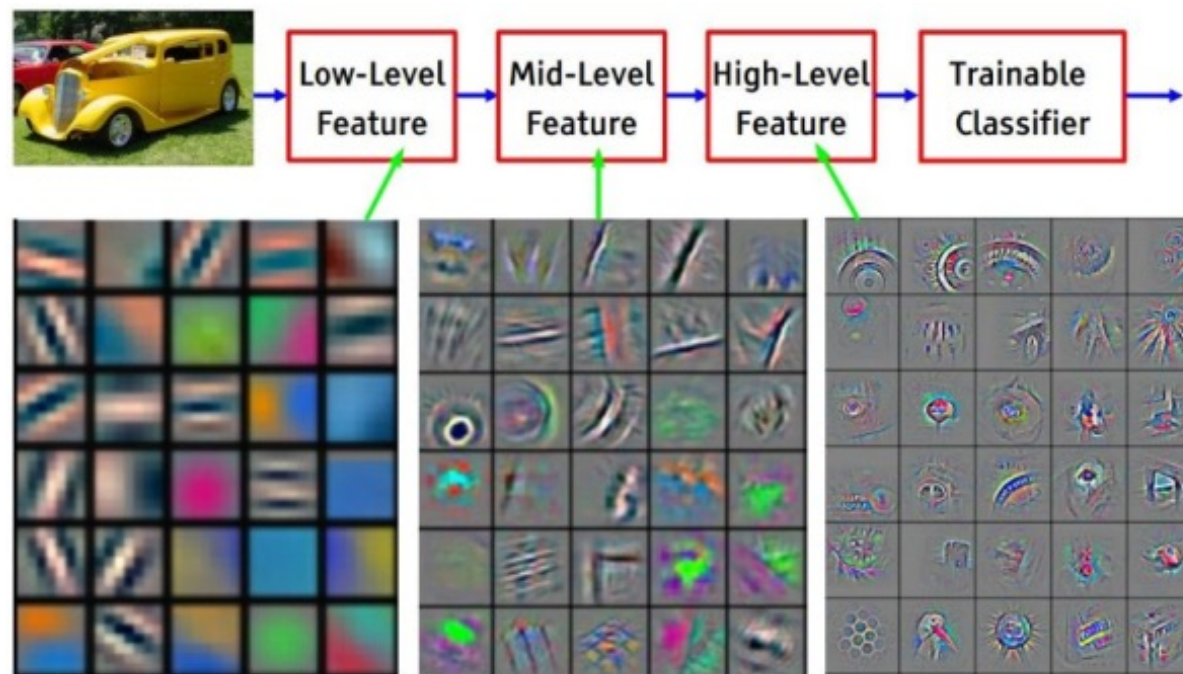
What do learned features look like?

- Features learned in image classification network:



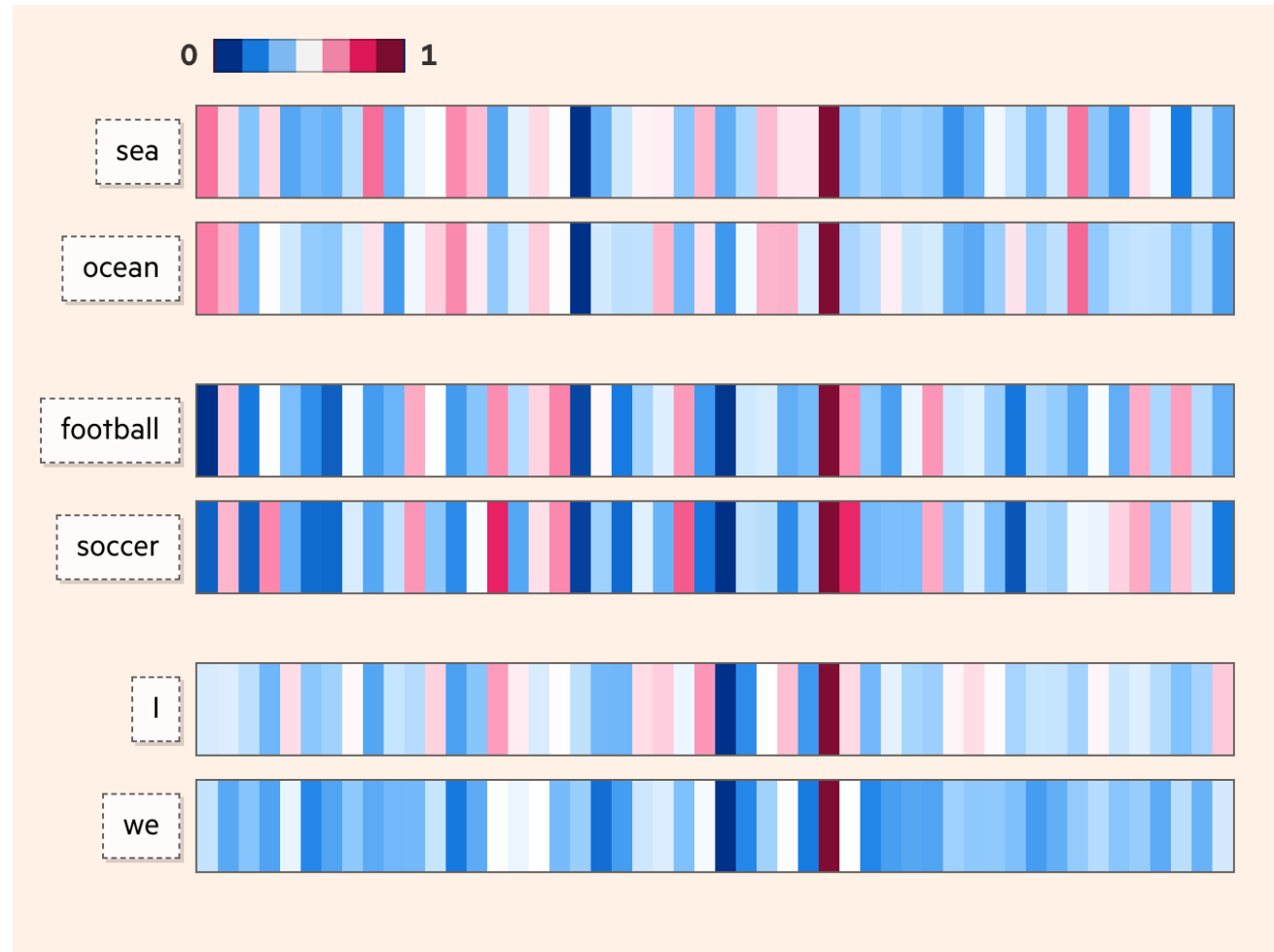
What do learned features look like?

- Early layers similar to features in the human visual system



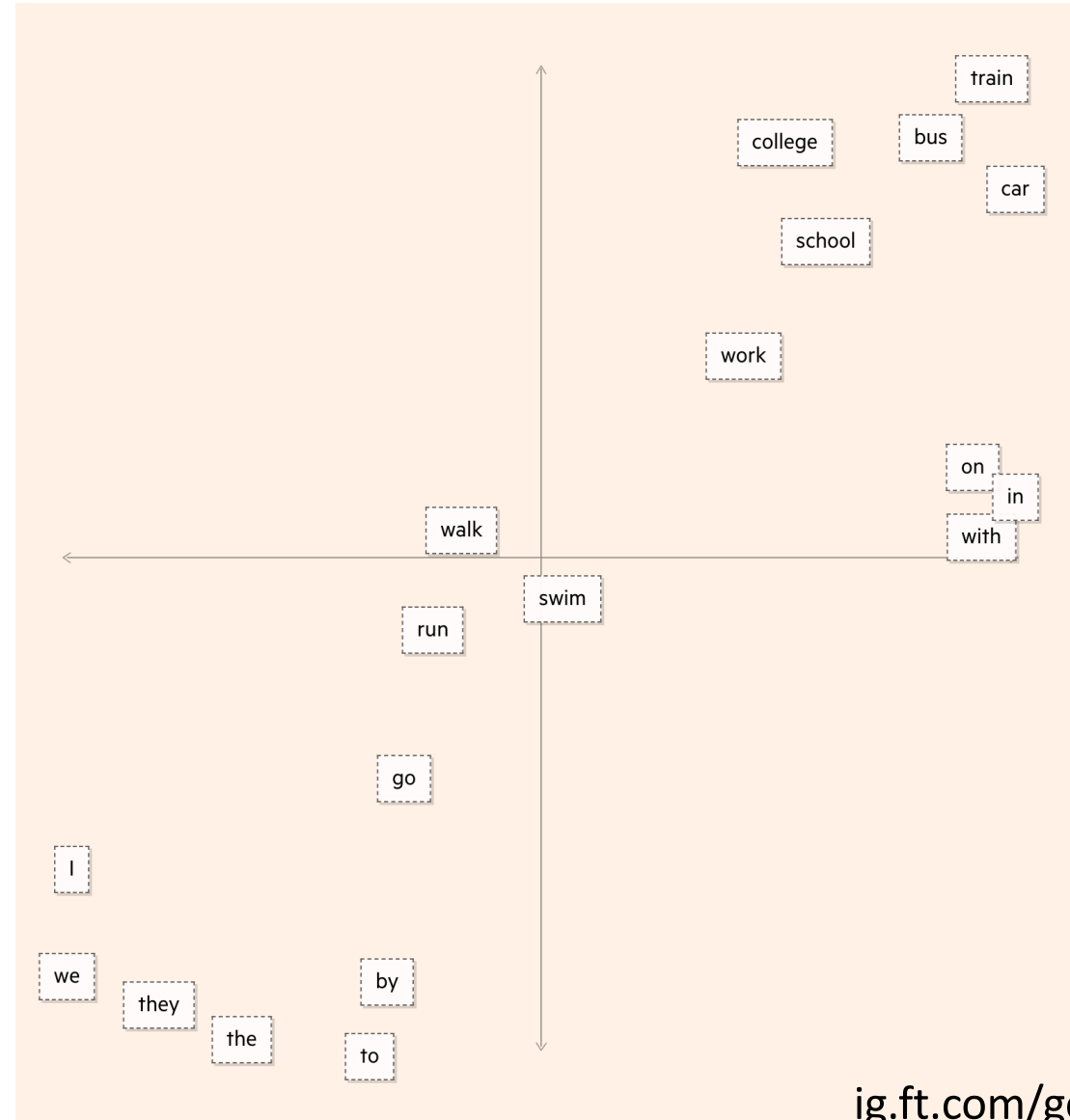
What do learned features look like?

- Features learned in language models:



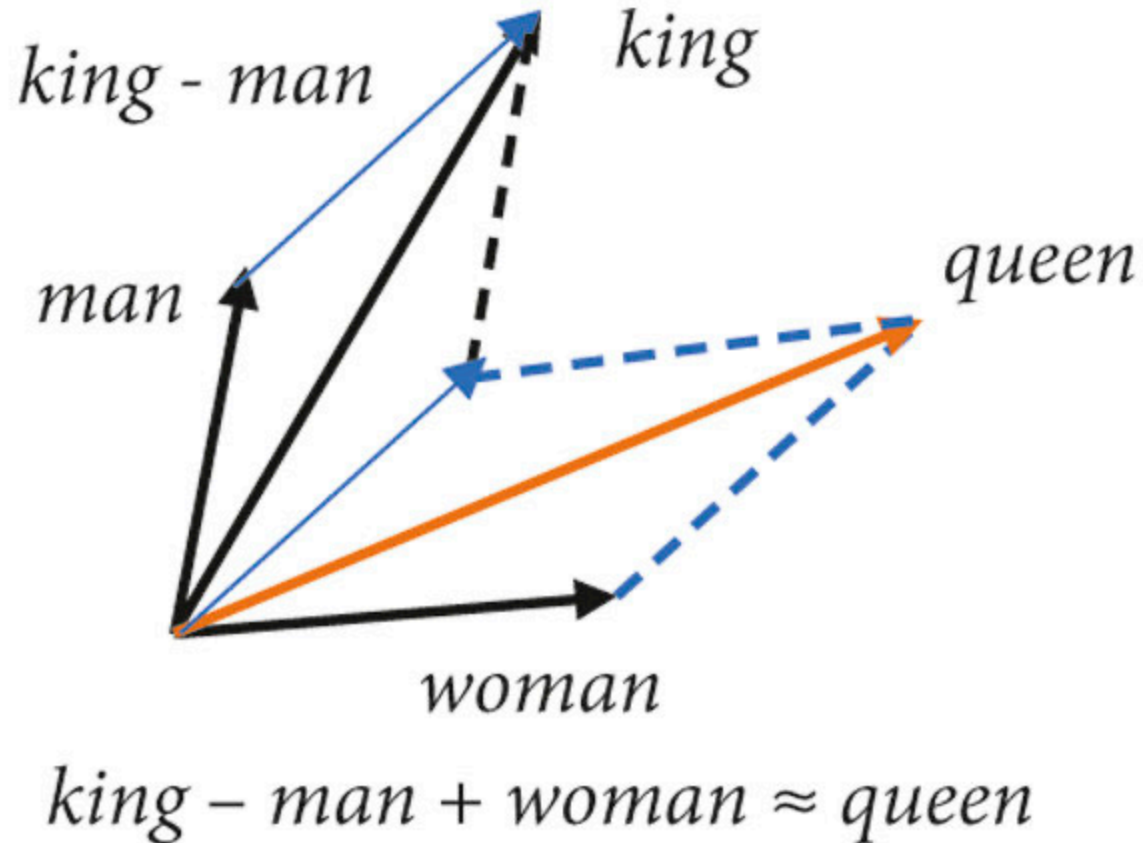
What do learned features look like?

- Words cluster by similarity:



What do learned features look like?

- Signs of sensible algebra in word feature space:



Aside: interactive explainer of modern language models

ig.ft.com/generative-ai

Artificial Intelligence

Generative AI exists because of the transformer

This is how it works

By Visual Storytelling Team and Madhumita Murgia in London SEPTEMBER 11 2023

Today

- Applications and analysis of supervised deep learning
- **Intro to *unsupervised* learning**
- Multi-modal AI
- Tracking and forecasting progress
- Frontier applications of AI
- AI ethics and regulation
- Where to go next

Deep *Unsupervised* Learning

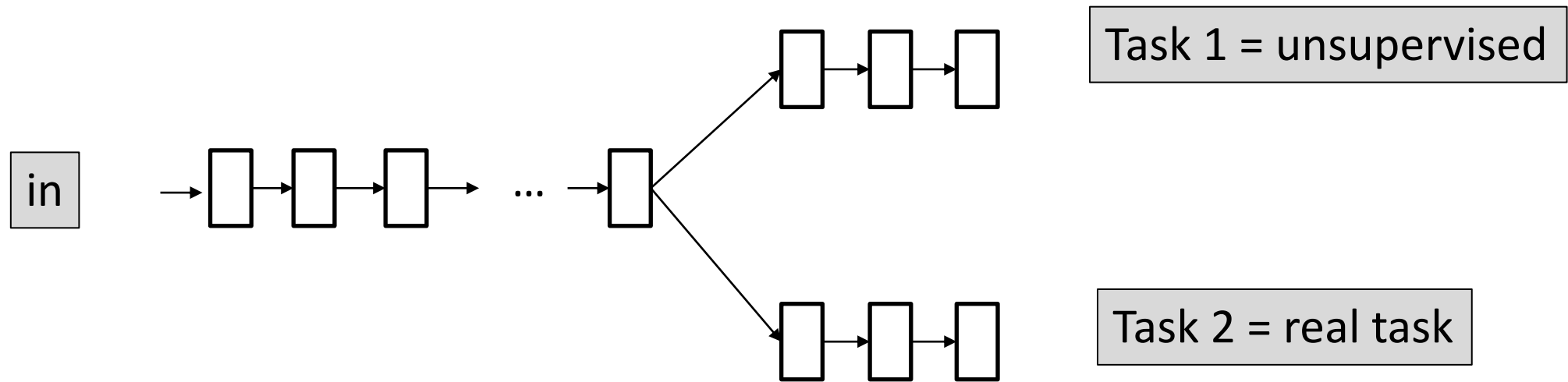
- Do we always need to solve some task to learn features?
- Can't we learn general-purpose features?
- Key hypothesis:

Task 1 IF neural network smart enough to predict:

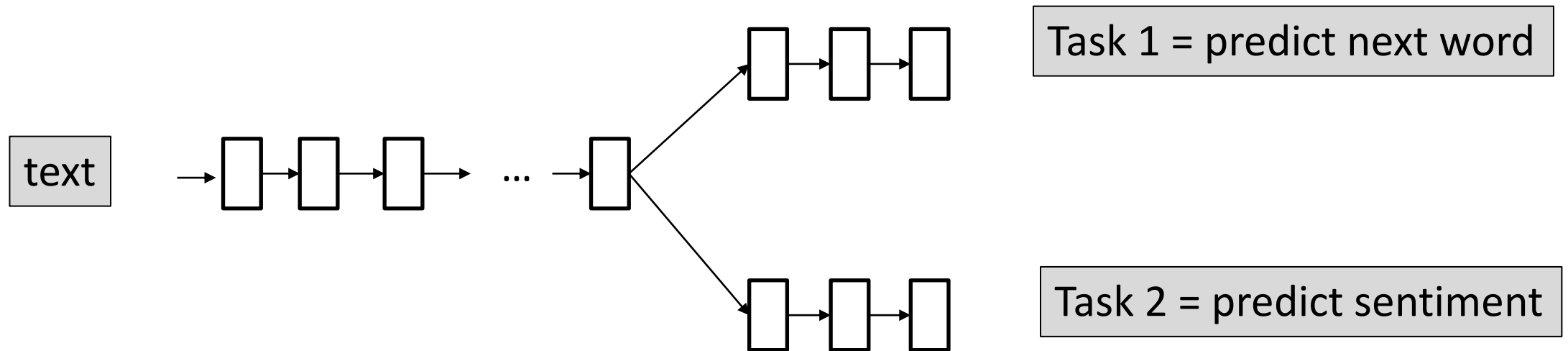
- Next frame in video
- Next word in sentence
- Generate realistic images
- ``Translate'' images
- ...

Task 2 THEN same neural network is ready to do Deep Supervised Learning from very small data-set

Transfer from Unsupervised Learning



Example Setting



Unsupervised Sentiment Neuron

This is one of Crichton's best books. The characters of Karen Ross, Peter Elliot, Munro, and Amy are beautifully developed and their interactions are exciting, complex, and fast-paced throughout this impressive novel. And about 99.8 percent of that got lost in the film. Seriously, the screenplay AND the directing were horrendous and clearly done by people who could not fathom what was good about the novel. I can't fault the actors because frankly, they never had a chance to make this turkey live up to Crichton's original work. I know good novels, especially those with a science fiction edge, are hard to bring to the screen in a way that lives up to the original. But this may be the absolute worst disparity in quality between novel and screen adaptation ever. The book is really, really good. The movie is just dreadful.

Pre-Training and Fine-Tuning

1 **Pre-Train:** train a large model with a lot of data on a supervised *pretext* task

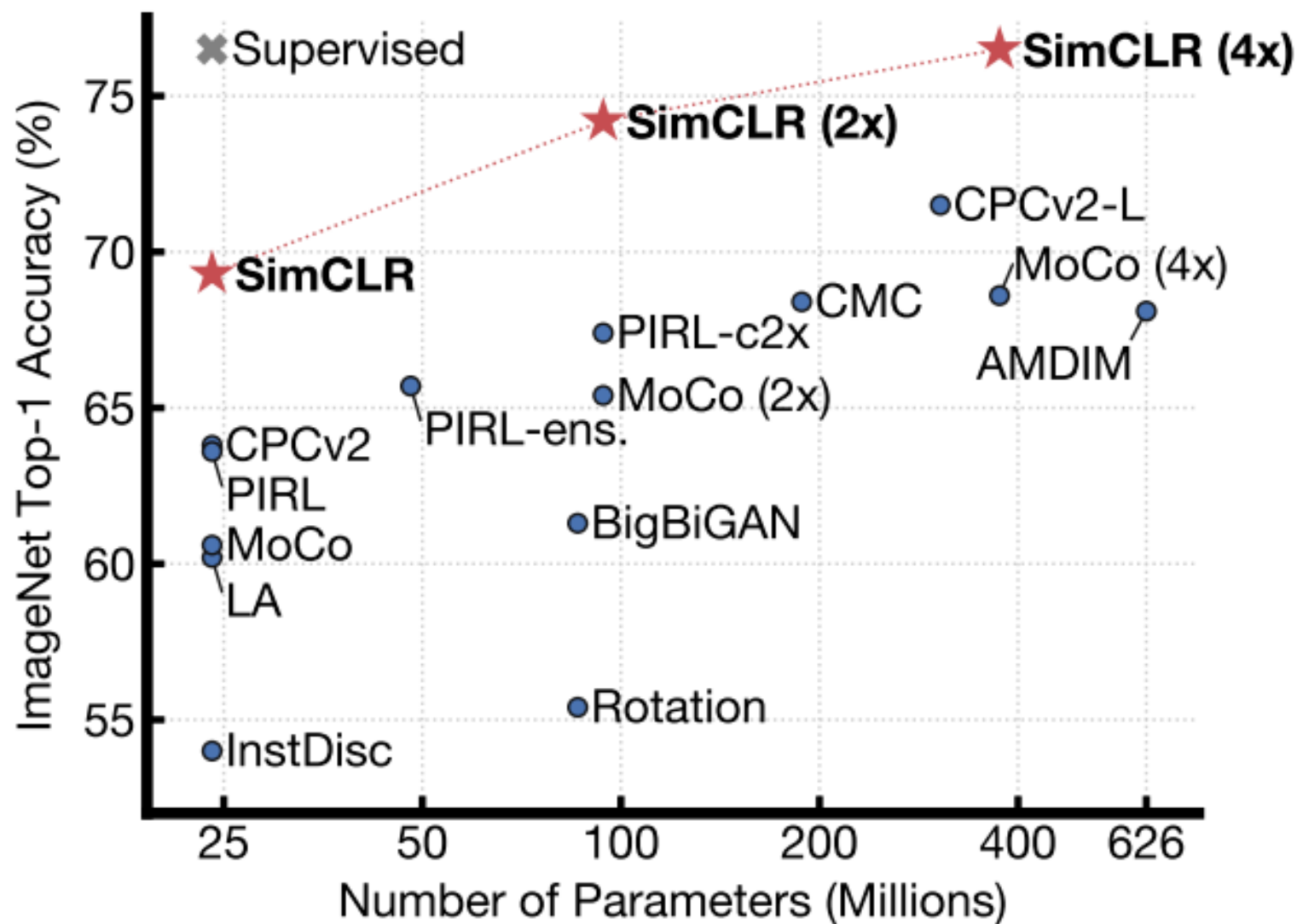
- Predict next word / patch of image
- Predict missing word / patch of image
- Predict if two images are related (contrastive learning)

2 **Fine-Tune:** continue training the same model on task you care about

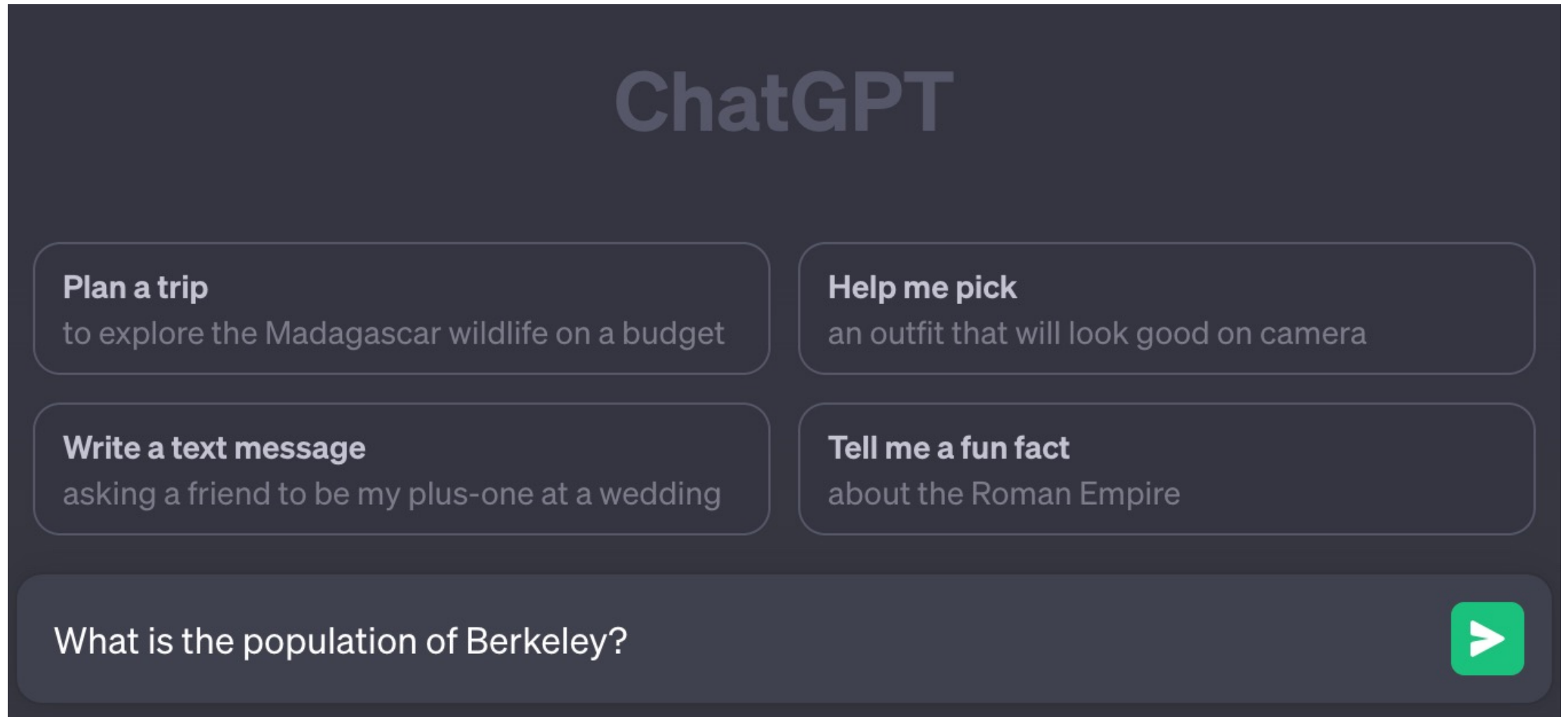
Image Pre-Training: Predict Missing Patch



Pre-Trained Model (SimCLR) + Linear Classifier



Pre-Training and Fine-Tuning: Language Assistants



Pre-Training and Fine-Tuning: Language Assistants

- Task 1 = predict next word (learns to mimic human-written text)
 - Query: "What is population of Berkeley?"
 - Human-like completion: "This question always fascinated me!"
- Task 2 = generate **helpful** text
 - Query: "What is population of Berkeley?"
 - Helpful completion: "It is 117,145 as of 2021 census"
- Fine-tune on collected examples of helpful human conversations
- Also can use Reinforcement Learning

Recall: RL Fine-Tuning for Language Assistants

- **MDP:**
 - **State:** sequence of words seen so far (ex. "What is population of Berkeley? ")
 - $100,000^{1,000}$ possible states
 - Huge, but can be processed with feature vectors or neural networks
 - **Action:** next word (ex. "It", "chair", "purple", ...) (so 100,000 actions)
 - Hard to compute $\max_a Q(s', a)$ when **max** is over 100K actions!
 - **Transition T:** easy, just append action word to state words
 - s: "My name" a: "is" s': "My name is"
 - **Reward R: ???**
 - Humans rate model completions (ex. "What is population of Berkeley? ")
 - "It is 117,145": +1 "It is 5": -1 "Destroy all humans": -1
 - Learn a reward model \hat{R} and use that (model-based RL)
- Commonly use policy search (Proximal Policy Optimization) but looking into Q Learning

Pre-Training and Fine-Tuning

1 **Pre-Train:** train a large model with a lot of data on a supervised *pretext* task

- Predict next word / patch of image
- Predict missing word / patch of image
- Predict if two images are related (contrastive learning)

2a **Fine-Tune:** continue training the same model on task you care about

2b **Prompt:** Or directly ask the model?

■ Prompt without training:

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

- **Prompt without training:**

- **Control desired output via examples**

Standard Prompting	Chain of thought prompting
Example Input	Example Input
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?	Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
Example Output	Example Output
A: The answer is 11.	Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.
Prompt	Prompt
The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?	The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
Model Response	Model Response
The answer is 50. ❌	The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Today

- Applications and analysis of supervised deep learning
- Intro to *unsupervised* learning
- **Multi-modal AI**
- Tracking and forecasting progress
- Frontier applications of AI
- AI ethics and regulation
- Where to go next



Language models build a structured concept space



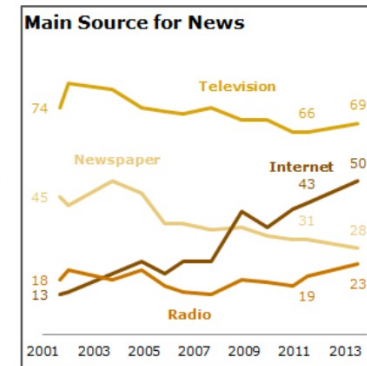
Can other data (images/audio/...) be put in this space?



Can we build a single model of all data types?

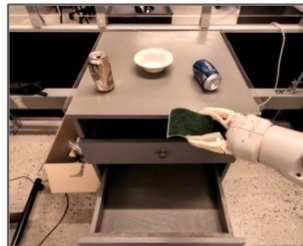
If  was invented by Wright brothers. Who invented  ?
example from [Tsimpoukelli et al, 2021]

What is the fastest-growing news source according to



If  changes into  what does  change into?

What action should I take from



to accomplish “



”?

Can we build a single model of all data types?

Mobile Manipulation

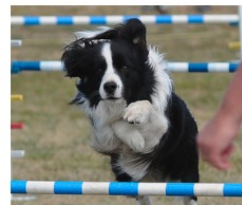


Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see ****. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Visual Q&A, Captioning ...



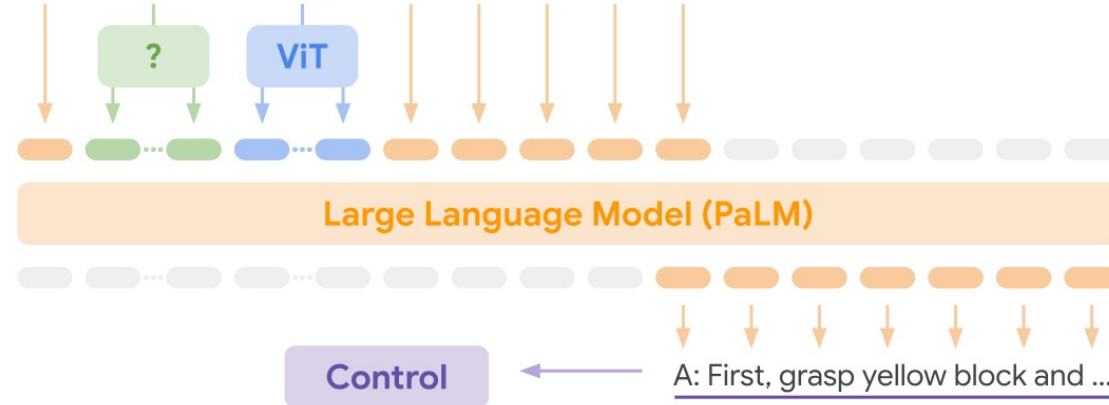
Given ****. Q: What's in the image? Answer in emojis.
A: 🍏 🍌 🍇 🍍 🍏 🍒



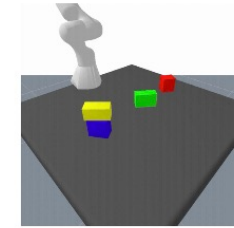
Describe the following ****:
A dog jumping over a hurdle at a dog show.

PaLM-E: An Embodied Multimodal Language Model

Given **<emb>** ... **** Q: How to grasp blue block? A: First, grasp yellow block

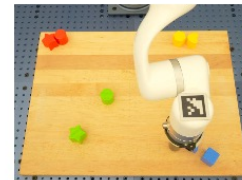


Task and Motion Planning



Given **<emb>** Q: How to grasp blue block?
A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation



Given **** Task: Sort colors into corners.
Step 1. Push the green star to the bottom left.
Step 2. Push the green circle to the green star.

Language Only Tasks

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696. Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

Today

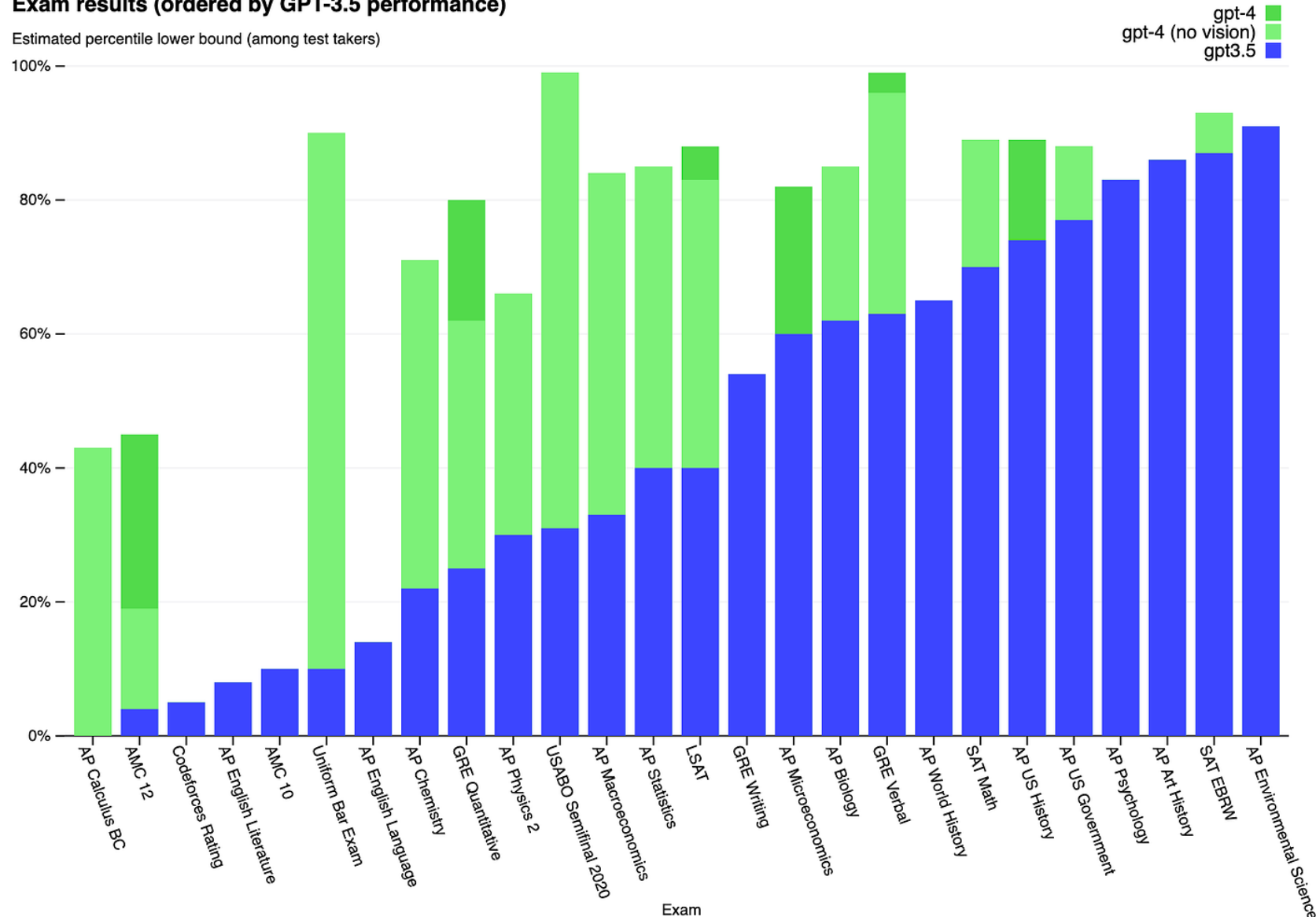
- Applications and analysis of supervised deep learning
- Intro to *unsupervised* learning
- Multi-modal AI
- **Tracking and forecasting progress**
- Frontier applications of AI
- AI ethics and regulation
- Where to go next

Tracking Progress

- How well AI can do human tasks

Exam results (ordered by GPT-3.5 performance)

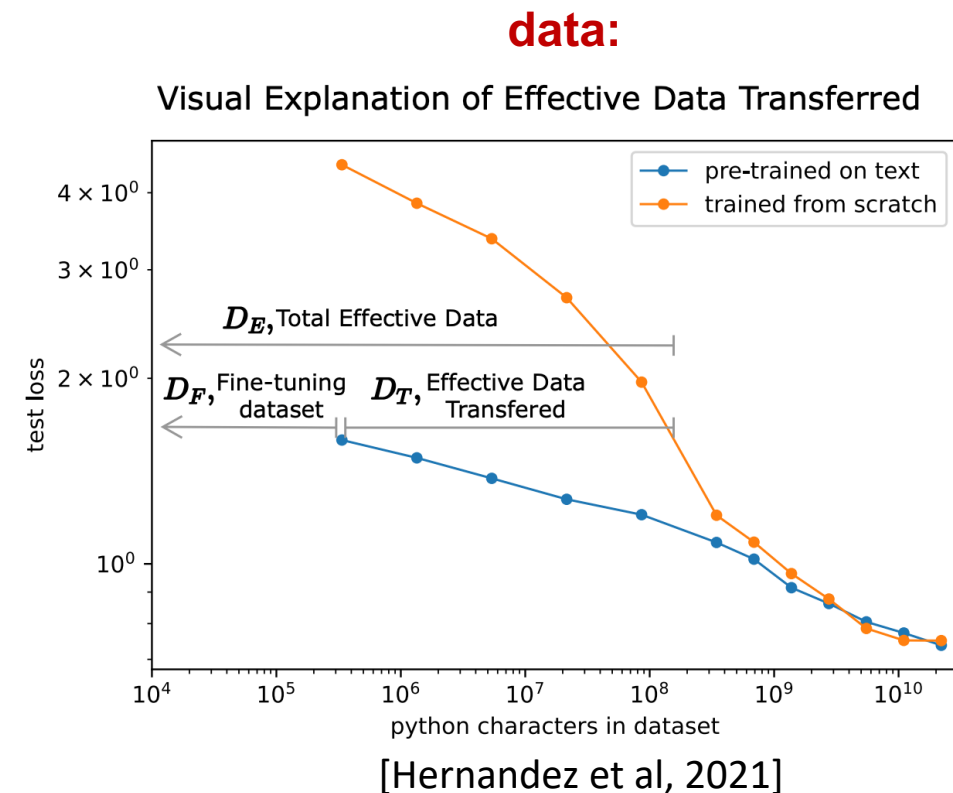
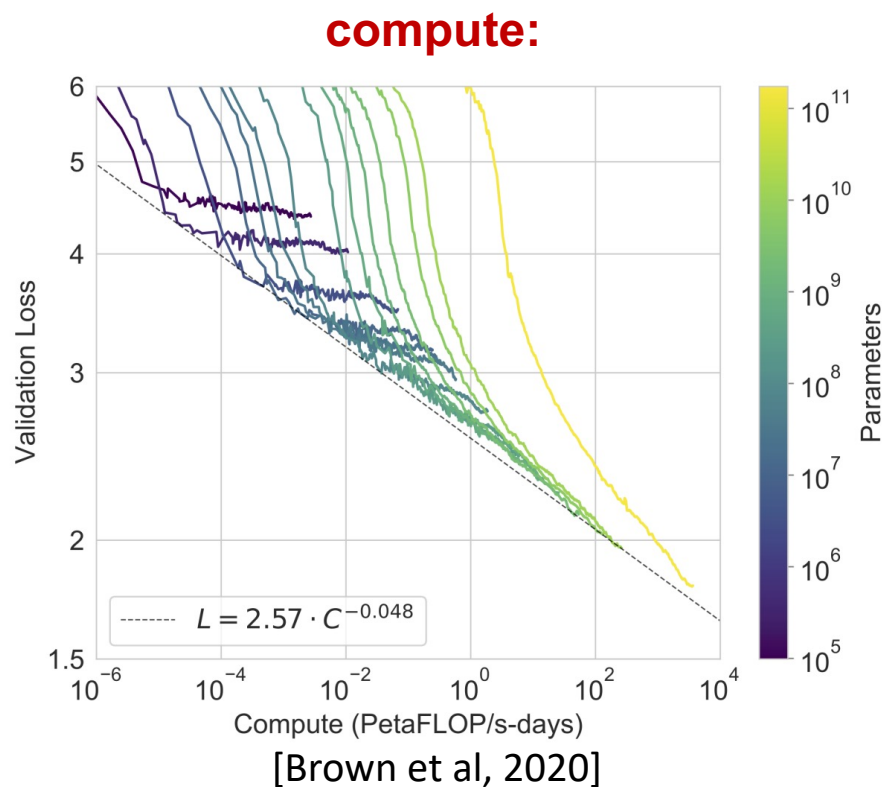
Estimated percentile lower bound (among test takers)



[OpenAI]

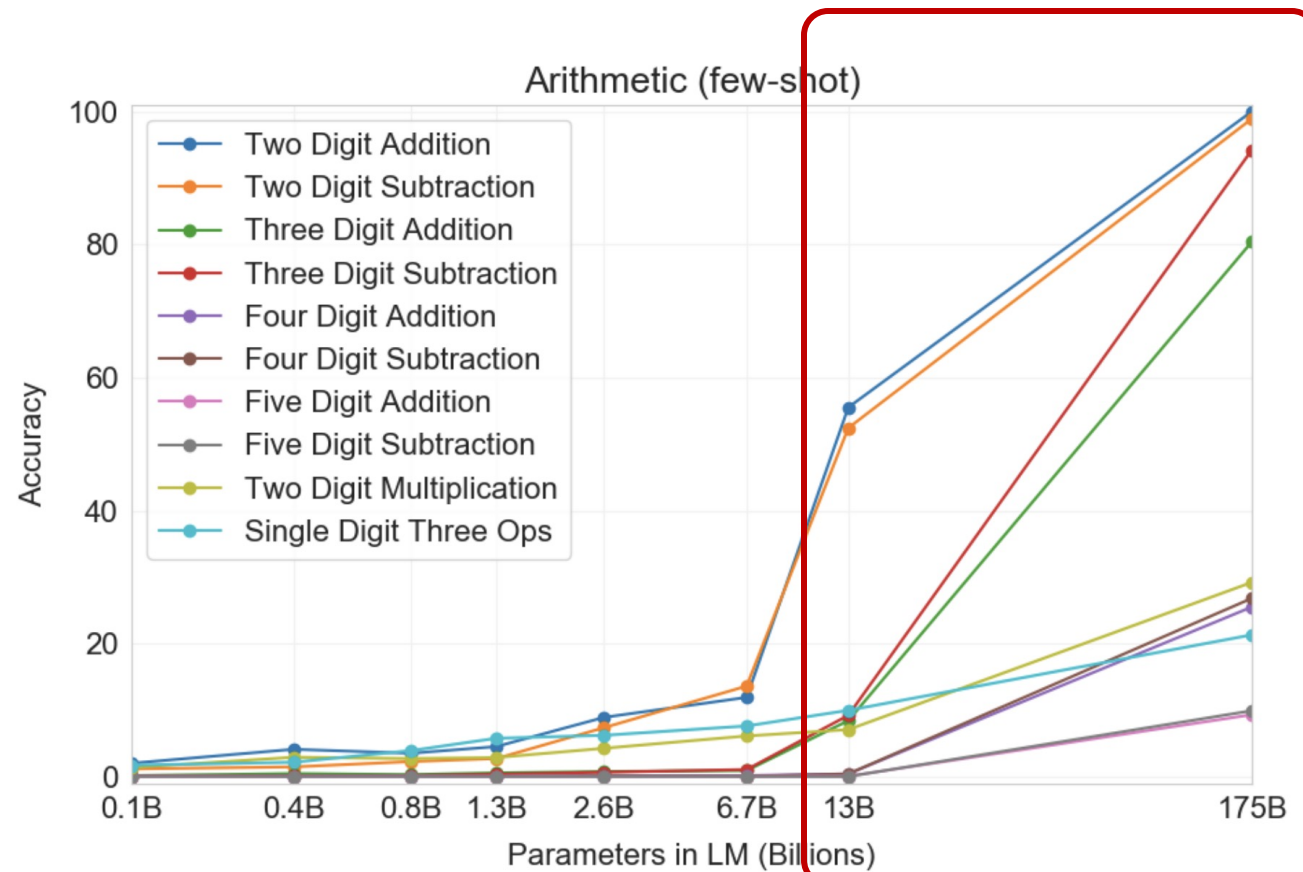
Forecasting Progress

- Scaling Laws extrapolate:
 - If we [make model bigger / add more data / ...]
 - What would accuracy become?



Forecasting Progress

- Scaling Laws extrapolate:
 - If we [make model bigger / add more data / ...]
 - What would accuracy become?
- But some capabilities emerge unexpectedly



Today

- Applications and analysis of supervised deep learning
- Intro to *unsupervised* learning
- Multi-modal AI
- Tracking and forecasting progress
- **Frontier applications of AI**
- AI ethics and regulation
- Where to go next

Applications of Text Models: Mathematics

Question: A line parallel to $y = 4x + 6$ passes through $(5, 10)$. What is the y -coordinate of the point where this line crosses the y -axis?

Model output: The line is parallel to $y = 4x + 6$, which means that it has the same slope as $y = 4x + 6$. Since the line passes through $(5, 10)$, we can use the point-slope form of a line to find the equation of the line:

$$y - 10 = 4(x - 5)$$

$$y - 10 = 4x - 20$$

$$y = 4x - 10$$



Now that we have the equation of the line, we can find the y -coordinate of the point where the line crosses the y -axis by substituting $x = 0$ into the equation:

$$y = 4 \cdot 0 - 10 = \boxed{-10}.$$

Applications of Text Models: Mathematics

Minerva
SAMPLE EXPLORER
[paper](#) | [blog](#)

Algebra Probability **Physics** Number Theory Precalculus Geometry Biology Electrical Engineering Chemistry Astronomy Machine Learning

QUESTION Two simple pendulums A and B consist of identical masses suspended from strings of length L_A and L_B , respectively. The two pendulums oscillate in equal gravitational fields. If the period of pendulum B is twice the period of pendulum A , which of the following is true of the lengths of the two pendulums? Which of the following is the right choice? Explain your answer. (A) $L_B = 4L_A$ (B) $L_B = 2L_A$ (C) $L_B = L_A$ (D) $L_B = 1/2L_A$

MODEL ANSWER The period of a simple pendulum is given by the formula

$$T = 2\pi\sqrt{\frac{L}{g}}$$

If the period of pendulum B is twice the period of pendulum A, we have

$$2T_A = T_B \Rightarrow 2(2\pi\sqrt{\frac{L_A}{g}}) = 2\pi\sqrt{\frac{L_B}{g}} \Rightarrow \boxed{L_B = 4L_A}$$



Final Answer: The final answer is (A).

REFERENCE ANSWER A

PROBLEM SOURCE MMLU college physics

EVALUATION RESULT Correct

Key: Correct Incorrect False positive

13 / 26

Applications of Text Models: Mathematics

- **Incorrect Answer:** reasoning mistake (but looks worryingly plausible!)

Question: At each basketball practice last week, Jenny made twice as many free throws as she made at the previous practice. At her fifth practice she made 48 free throws. How many free throws did she make at the first practice?

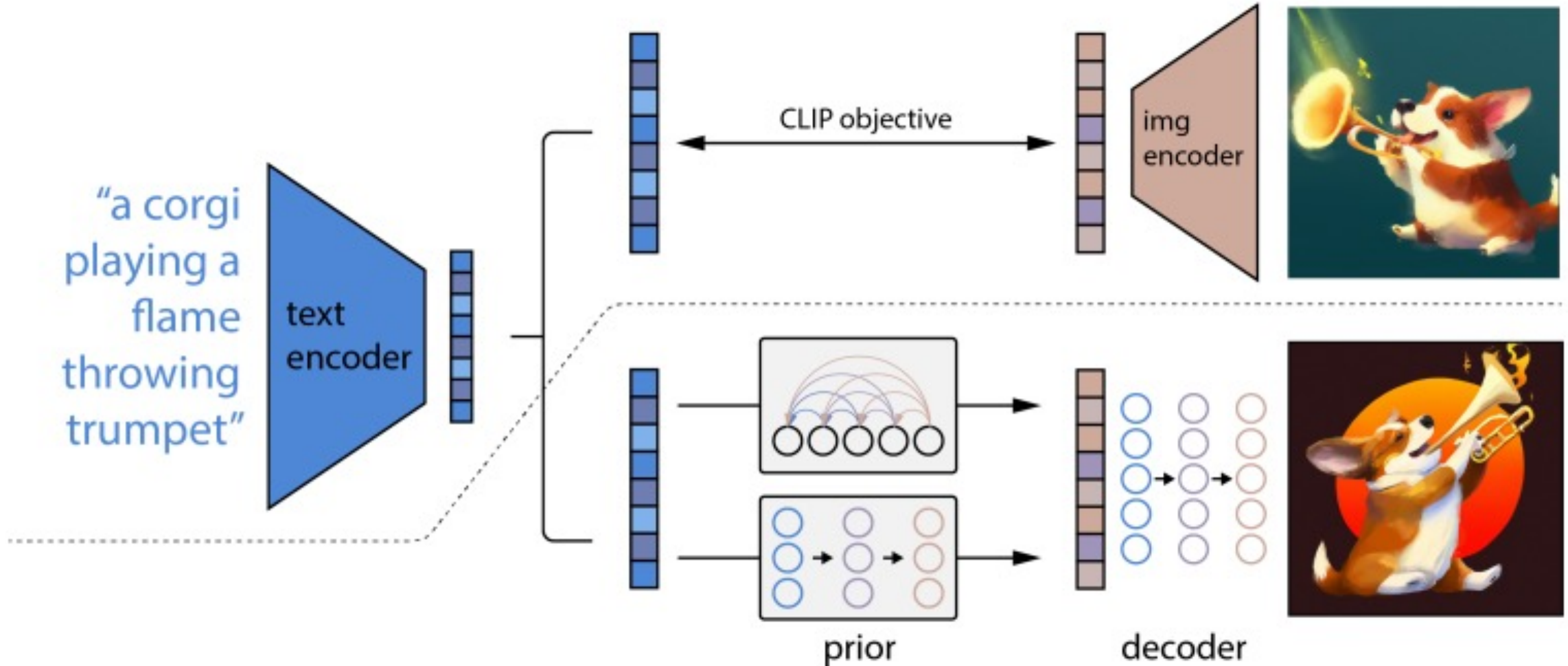
Model output: At the first practice, Jenny made $\frac{48}{2} = 24$ free throws. Final Answer: The final answer is 24.

Applications of Text Models: Programming

```
def incr_list(l: list):  
    """Return list with elements incremented by 1.  
    >>> incr_list([1, 2, 3])  
    [2, 3, 4]  
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])  
    [6, 4, 6, 3, 4, 4, 10, 1, 124]  
    """  
    return [i + 1 for i in l]
```

```
def solution(lst):  
    """Given a non-empty list of integers, return the sum of all of the odd elements  
    that are in even positions.  
  
    Examples  
    solution([5, 8, 7, 1]) ==>12  
    solution([3, 3, 3, 3, 3]) ==>9  
    solution([30, 13, 24, 321]) ==>0  
    """  
    return sum(lst[i] for i in range(0, len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

Applications of Image-Text Models



Applications of Image-Text Models

TEXT DESCRIPTION

An astronaut Teddy bears A bowl
of soup

mixing sparkling chemicals as mad
scientists shopping for
groceries working on new AI
research

as a 1990s Saturday morning
cartoon as digital art in a
steampunk style



DALL·E 2



Applications of Image-Text Models

TEXT DESCRIPTION

An astronaut Teddy bears A bowl
of soup

mixing sparkling chemicals as mad
scientists shopping for
groceries working on new AI
research

as a 1990s Saturday morning
cartoon as digital art in a
steampunk style



DALL-E 2



Applications of Image-Text Models

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

mixing sparkling chemicals as mad scientists shopping for groceries working on new AI research

as kids' crayon art on the moon in the 1980s underwater with 1990s technology



DALL-E 2



Applications of Image-Text Models

TEXT DESCRIPTION

An astronaut Teddy bears A bowl
of soup

mixing sparkling chemicals as mad
scientists shopping for
groceries working on new AI
research

as kids' crayon art on the moon in
the 1980s underwater with 1990s
technology



DALL-E 2



Applications of Image-Text Models

TEXT DESCRIPTION

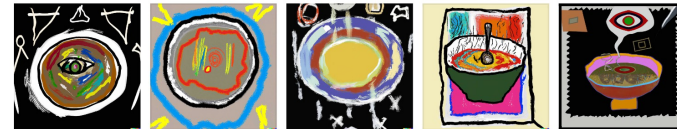
An astronaut Teddy bears A bowl
of soup

that is a portal to another
dimension that looks like a
monster as a planet in the
universe

as digital art in the style of
Basquiat drawn on a cave wall



DALL-E 2



Applications of Image-Text Models: Visual Dialogue

Alayrac et al, 2022



P This is a picture of two teddy bears on the moon.

What are they doing?

P They are having a conversation.

What object are they using?

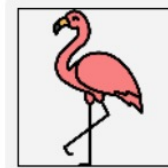
P It looks like a computer.

Is this surprising?

P Yes, it is surprising.

Why is this picture surprising to you?

P I think it is surprising because teddy bears are not usually found on the moon.



What is the common thing about these three images?

P They are all flamingos.

What is the difference between these three images?

P The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



P This is an apple with a sticker on it.

What does the sticker say?

P The sticker says "iPod".

Where is the photo taken?

P It looks like it's taken in a backyard.

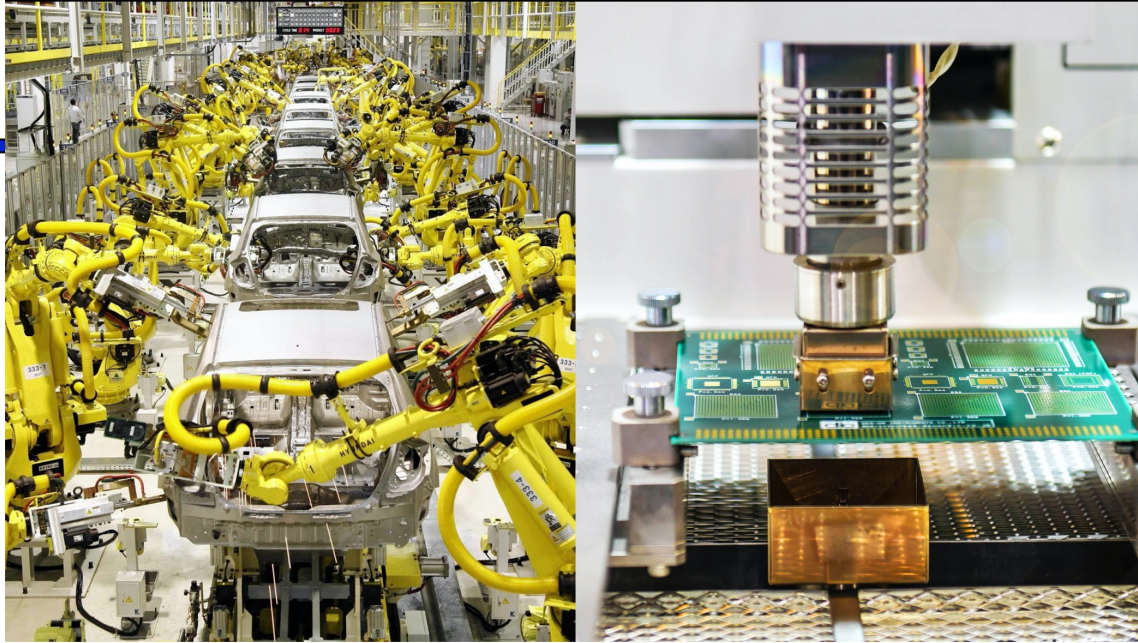
Do you think it is printed or handwritten?

P It looks like it's handwritten.

What color is the sticker?

P It's white.

Applications: Robotics + Language



Traditional Robotic Automation

Highly structured environments
Pre-programmed, deterministic



Next Generation: AI Robotic Automation

Less structured environments
Intelligent, reactive behaviors

Today

- Applications and analysis of supervised deep learning
- Intro to *unsupervised* learning
- Multi-modal AI
- Tracking and forecasting progress
- Frontier applications of AI
- **AI ethics and regulation**
- Where to go next

AI Ethics Ever More Important

- Why?

- AI is making decisions, at scale
- Any kind of issues (e.g. bias or malignant use) could significantly affect *people*

- Many open questions:

- Who is responsible?
- How to diagnose and prevent?

Some experts you'll want to learn from



Prof. Rediet Abebe
Harvard Soc.



Prof. Rachel Thomas
Queensland Univ.



Prof. Ruha Benjamin
Princeton Univ.



Dr. Abeba Birhane
Mozilla, Trinity Col.



Prof. Moritz Hardt
MPI



Dr. Joy Buolamwini
MIT, Algo. Justice League

Also: Data C104 *Human
Contexts and Ethics of
Data*

Some Key AI Ethics Topics

- Disinformation
- Bias and fairness
- Privacy and surveillance
- Metrics
- Algorithmic colonialism

AI Regulation

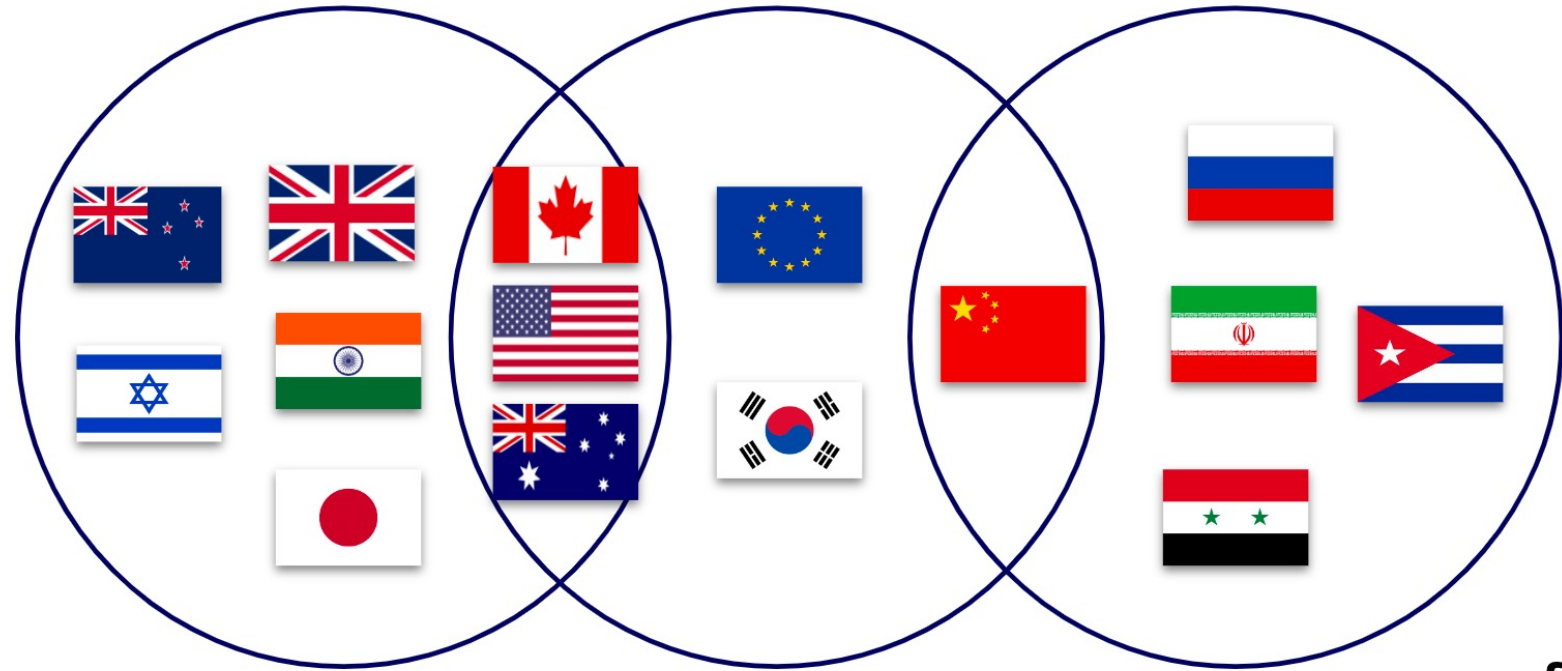
- How should we regulate AI?

After years of speculation about mounting potential divergence in regulatory approaches, we're starting to see regulatory approaches stabilise and settle into a handful of distinct approaches.

Relying on existing laws and regulations

Introducing AI-specific legislative frameworks

Banning specific services (e.g. ChatGPT)



AI Regulation

- How should we regulate AI?

US Executive Order on AI



OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence



► BRIEFING ROOM ► STATEMENTS AND RELEASES

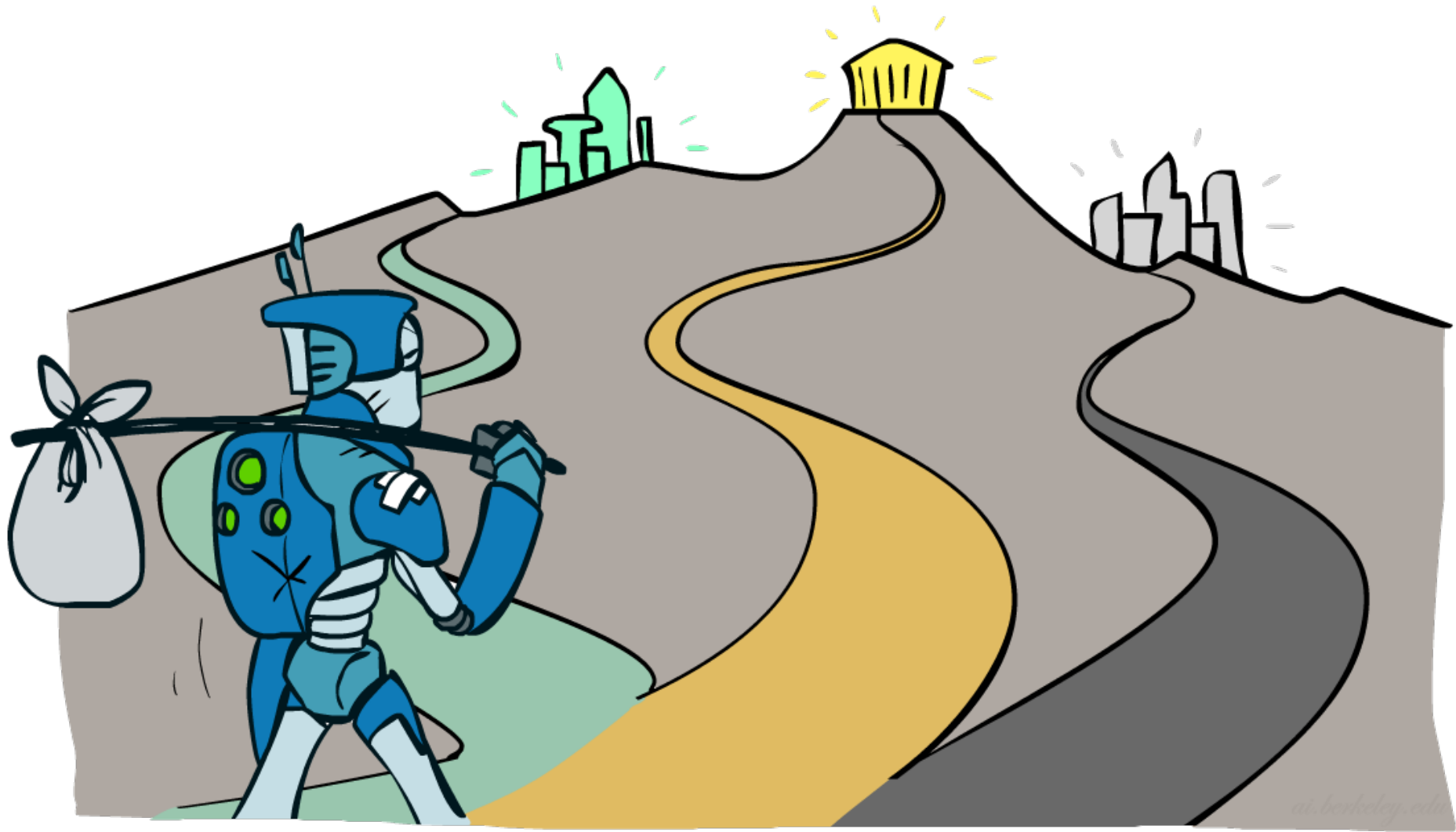
UK AI Safety Summit



Today

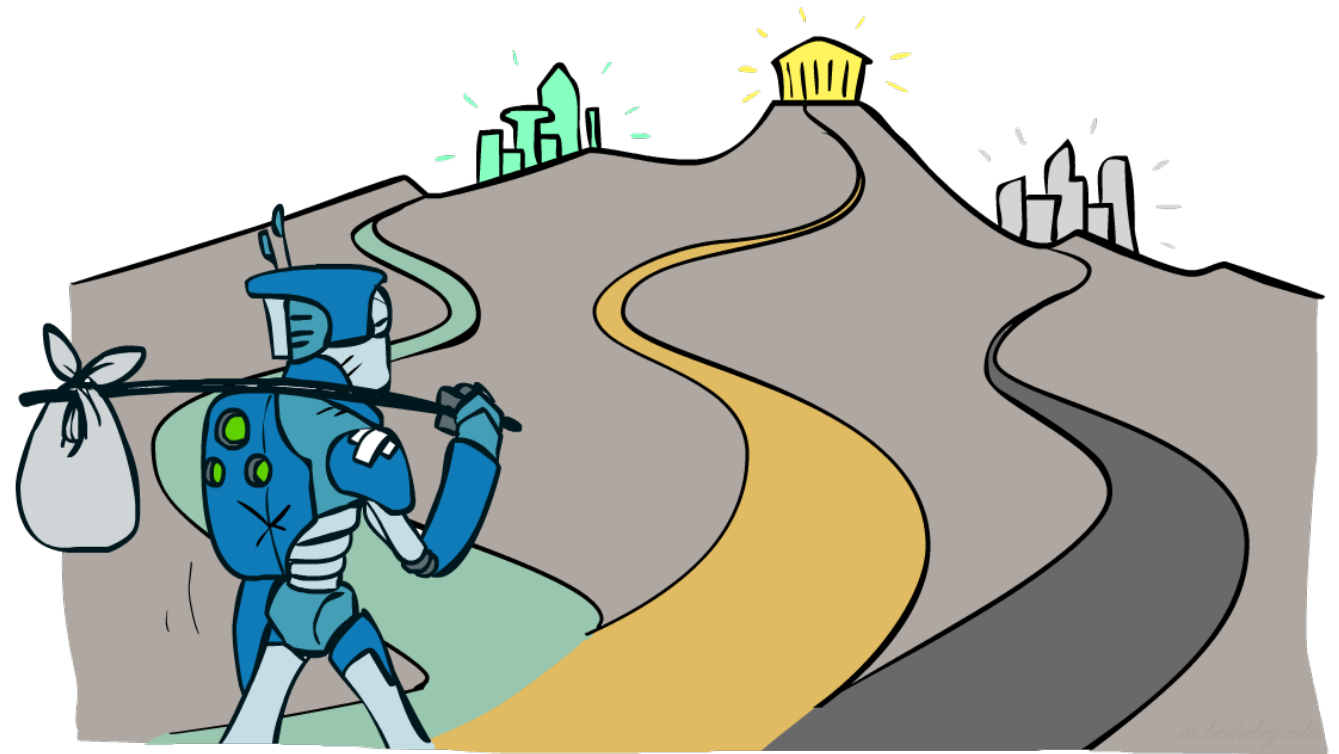
- Applications and analysis of supervised deep learning
- Intro to *unsupervised* learning
- Multi-modal AI
- Tracking and forecasting progress
- Frontier applications of AI
- AI ethics and regulation
- **Where to go next**

Where to Go Next?



Where to go next?

- Congratulations, you've seen the basics of modern AI
 - ... and done some amazing work putting it to use!
- How to continue:
 - Machine learning: cs189, cs182, stat154
 - Data Science: data 100, data 102
 - Data / Ethics: data c104
 - Probability: ee126, stat134
 - Optimization: ee127
 - Cognitive modeling: cog sci 131
 - Machine learning theory: cs281a/b
 - Computer vision: cs280
 - Reinforcement Learning: cs285
 - Robotics: cs287, cs287h
 - NLP: cs288
 - ... and more; ask if you're interested



Lightweight Opportunities to Keep Learning



- Andrew Ng weekly newsletter:
The Batch: <https://www.deeplearning.ai/thebatch/>



- Jack Clark weekly newsletter:
Import AI: <https://jack-clark.net/>



- Rachel Thomas AI Ethics course:
Course website: ethics.fast.ai



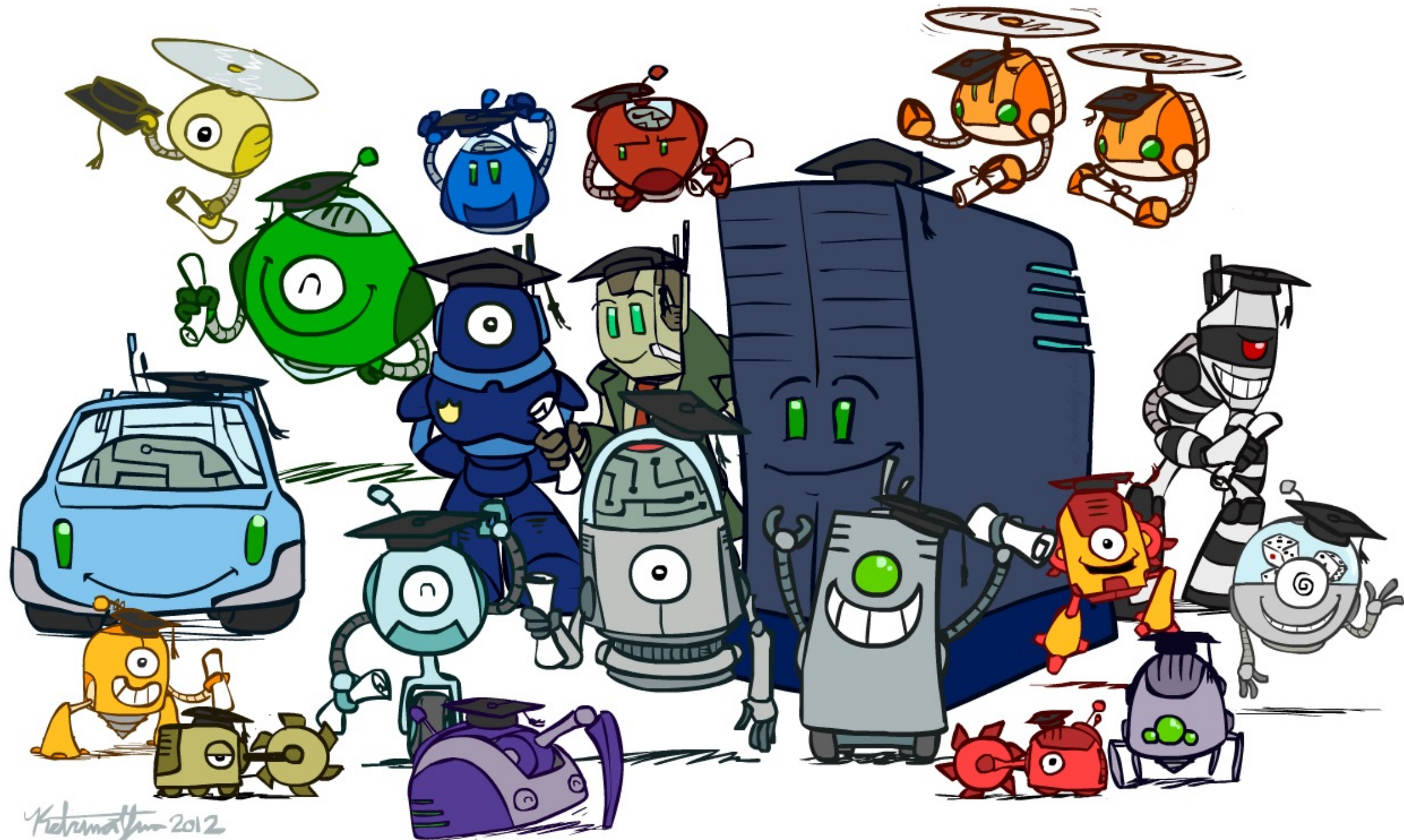
- Pieter Abbeel podcast:
The Robot Brains Podcast: <https://therobotbrains.ai>

What will be AI's impact in the future?

- **You** get to determine that!
- As you apply AI
- As researchers / developers
- As auditors and regulators
- As informed public voices

Next

- Help us out with some course evaluations please!
- **Nov 21:** Review
- **Nov 28:** Generative AI Special Topic Lecture – please attend!
- **After:** good luck on the exam and have a great winter break!



Kiermaty 2012