

CPSC 340:  
Machine Learning and Data Mining

Deep Learning & Conclusion

# CPSC 340: Machine Learning and Data Mining

Outlier Detection

Fall 2021





# Discussion & Summary of CNNs

- Convolutional layers reduce the number of parameters in several ways:
  - Each hidden **unit only depends on small number of inputs** from previous layer.
  - We use the **same filters across the image** .
    - So we do not learn a different weight for each “connection” like in classic neural networks
    - Benefits of this described below
  - Pooling layers **decrease the image size** .
- CNNs give some amount of **translation invariance**
  - Because same filters used across the image, they can **detect a pattern anywhere in the image**
    - Even in image **locations where the pattern has never been seen (thus more data-efficient, but less powerful)**
- CNNs are **not only for images!**
  - Can use CNNs for 1D sequences like sound or language or biological sequences.
  - Can use CNNs for 3D objects like videos or medical image volumes.
  - Can use CNNs for graphs.
- But you do need some notion of “neighbourhood ” for convolutions to make sense.

(end, tested technical  
material)

Some high-level principles, and ethical issues we will  
cover are still testable

# Today

- Fun whirlwind of AI dangers, capabilities, and weaknesses
- Concluding thoughts

Please fill out course survey

"One of the most surprising and important stories of our time."

—Ashlee Vance, author of *Elon Musk*

# Genius Makers



The Mavericks Who Brought AI  
to Google, Facebook, and the World

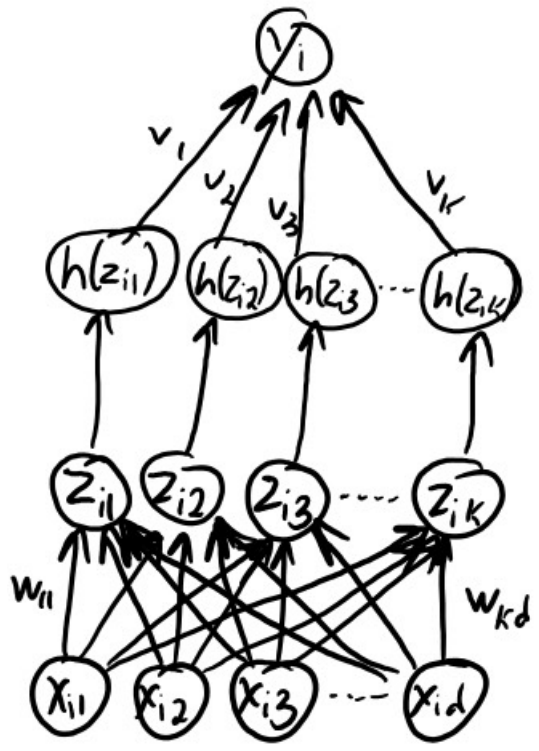
CADE METZ

# Supermarket Parable

- “So, suppose you want to find things that are **like** a can of sardines.
- What you do is you go to your local supermarket and you say to the cashier, "Where do you keep the sardines?" And you go to where the sardines are and then you just look around and there's all the things similar to sardines because the supermarket arrange things sensibly.
- Now, it doesn't quite work because you don't find the anchovies, as I discovered when I came to North America, I couldn't find the anchovies. They weren't anywhere near the sardines and the tuna. That's because they're near the pizza toppings.
- But that's just because it's a three dimensional supermarket. If there was a 30 dimensional supermarket, they could be close to the pizza toppings and close to the sardines.” - Geoff Hinton

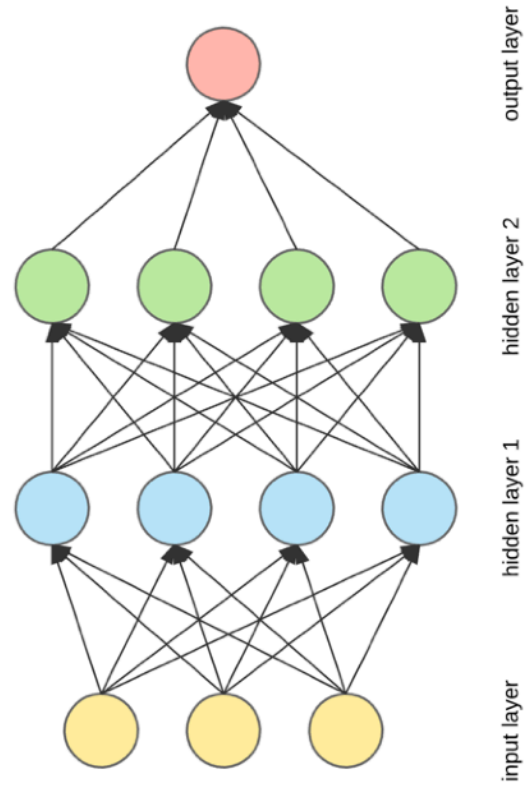
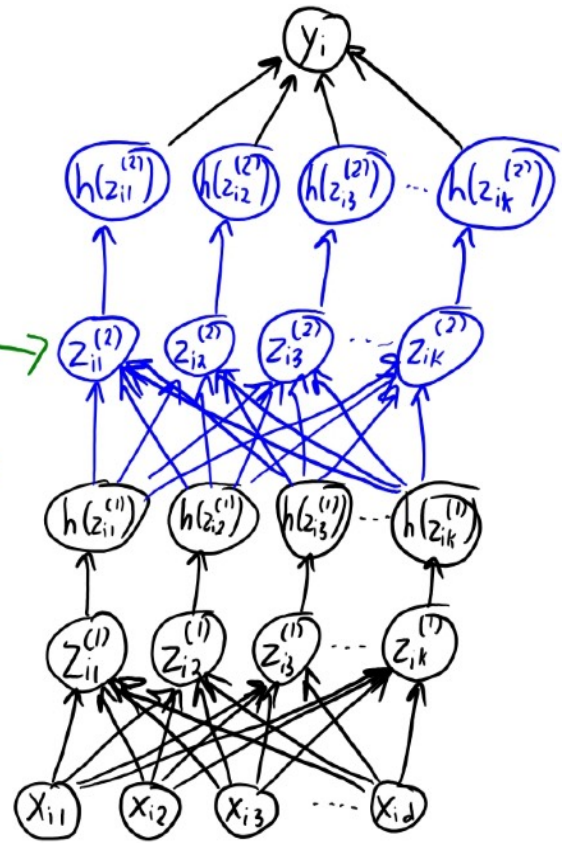
# Deep Learning

Neural network:

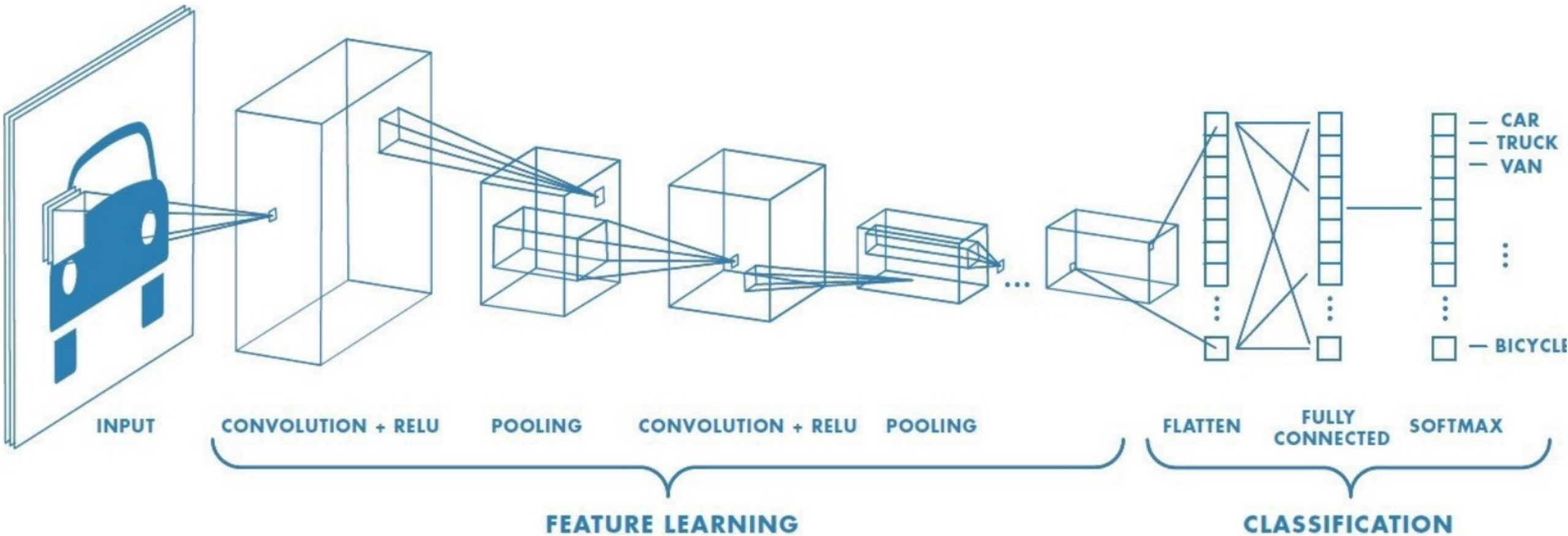


Deep learning:

Second "layer" of latent features  
You can add more "layers" to go "deeper"

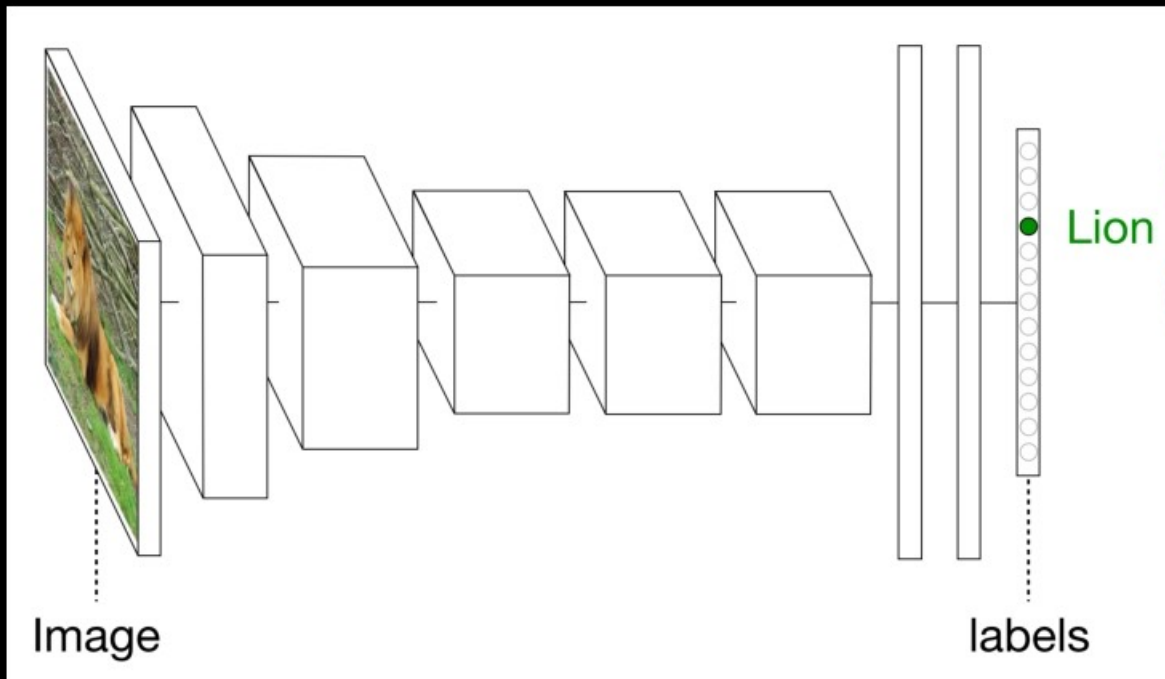


# Convolutional Neural Networks

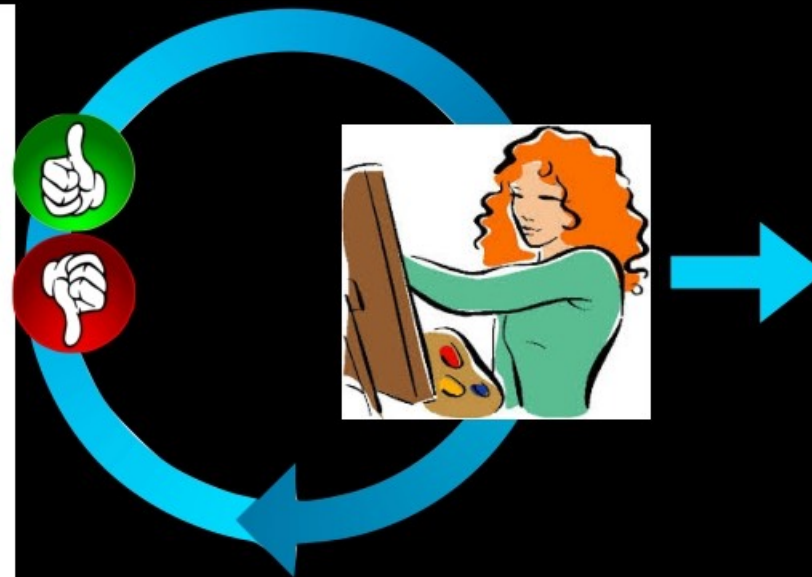




# Investigating What Each Neuron Does

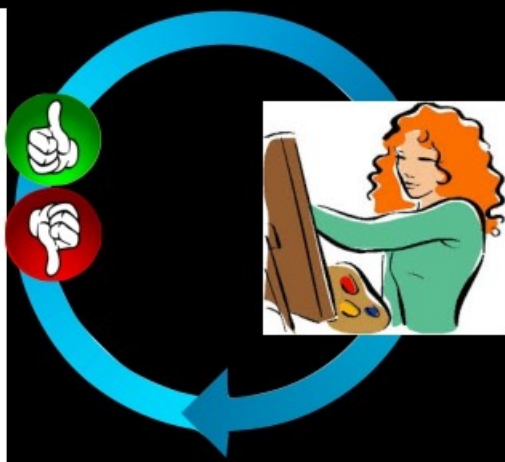
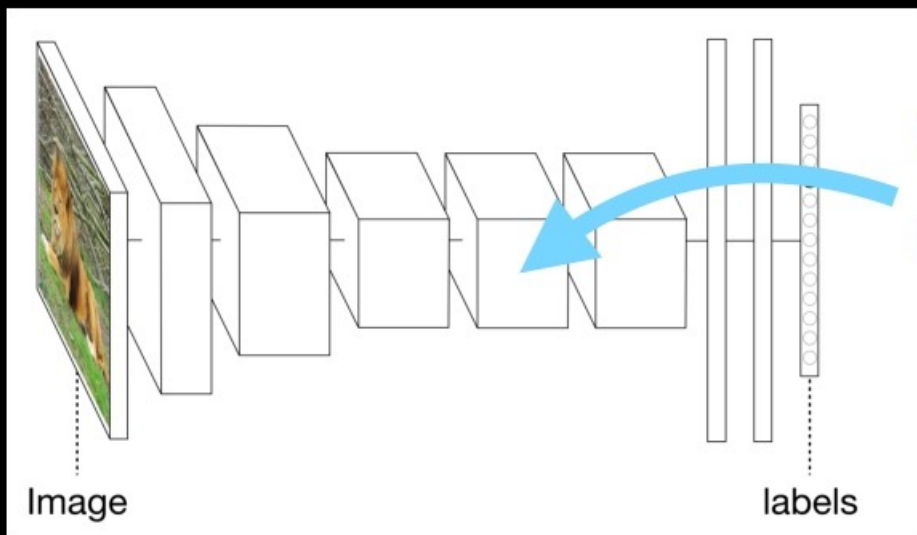
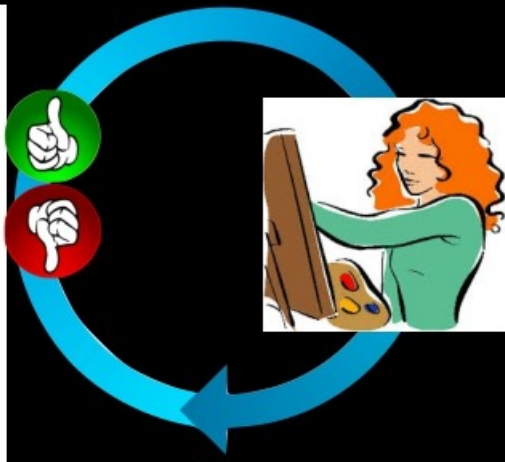
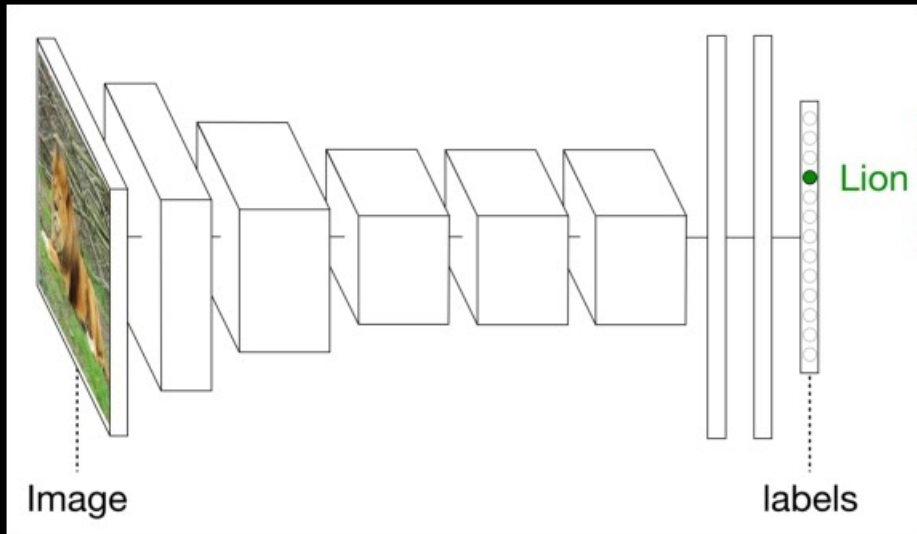


Pretrained, Fixed DNN



Optimize Pixels  
e.g. via Backprop

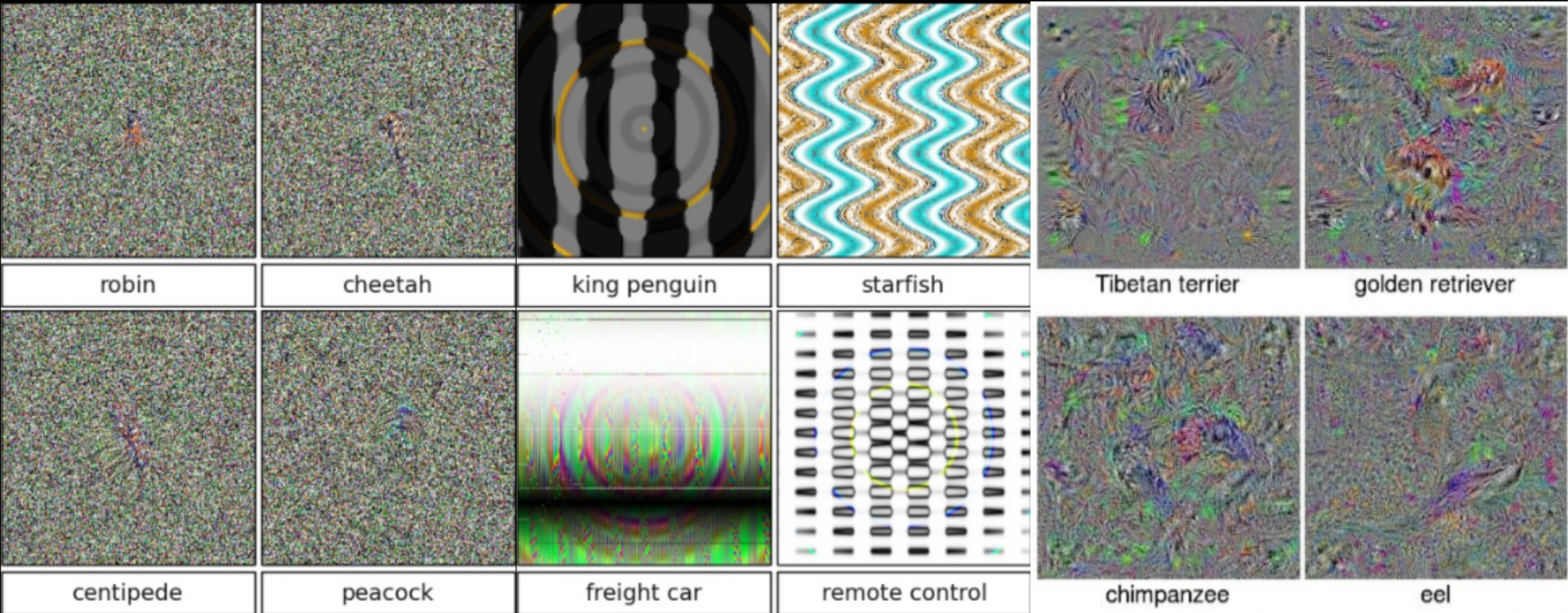
# “Deep Visualization”





# Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR



DNN Confidence:  $> 99.6\%$  for all



# Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

Anh Nguyen  
University of Wyoming  
anguyen8@uwyo.edu

Jason Yosinski  
Cornell University  
yosinski@cs.cornell.edu

Jeff Clune  
University of Wyoming  
jeffclune@uwyo.edu

## Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study [30] revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion). Specifically, we take convolutional neural networks trained to perform well on either the ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that DNNs label with high confidence as belonging to each dataset class. It is possible to produce images totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which we call “fooling images” (more generally, fooling examples). Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the generality of DNN computer vision.

## 1. Introduction

Deep neural networks (DNNs) learn hierarchical layers of representation from sensory input in order to perform pattern recognition [2, 14]. Recently, these deep architectures have demonstrated impressive, state-of-the-art, and sometimes human-competitive results on many pattern recognition tasks, especially vision classification problems [16, 7, 31, 17]. Given the near-human ability of DNNs to classify visual objects, questions arise as to what differences remain between computer and human vision.

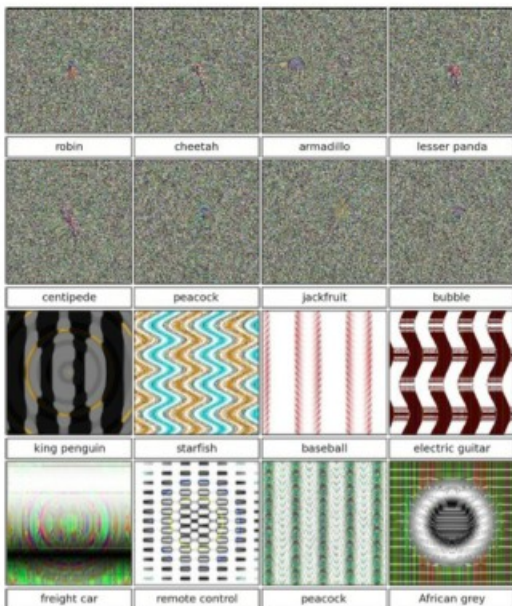


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with  $\geq 99.6\%$  certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (top) or indirectly (bottom) encoded.

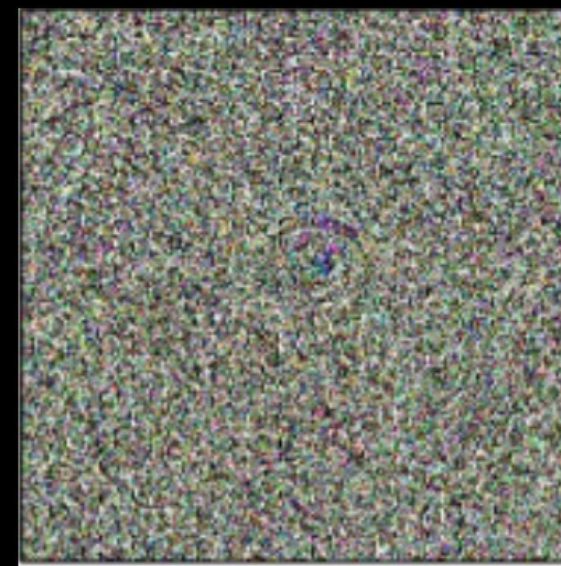
A recent study revealed a major difference between DNN and human vision [30]. Changing an image, originally correctly classified (e.g. as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library).

In this paper, we show another way that DNN and human vision differ: It is easy to produce images that are completely unrecognizable to humans (Fig. 1), but that state-of-the-art DNNs believe to be recognizable objects with over 99% confidence (e.g. labeling with certainty that TV static

- May not understand much
- Huge security concern
- Helped launch avalanche of work into “adversarial & fooling examples”
- with Szegedy et al. 2013



School bus



Open road!

# Why are networks easily fooled?

<https://www.youtube.com/watch?v=31p9eN5JE2A>

# Fooling Neural Networks

- Can someone repaint a stop sign and fool self-driving cars?



Eykholt et al. 2018

# Fooling Neural Networks

- ...or can it be even easier?



Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%



Granny Smith	0.3%
iPod	0.0%
library	0.0%
pizza	0.0%
toaster	93.7%
dough	0.2%



# Learning the Wrong Thing

- CNNs **may not be learning what you think they are.**

- CNN for diagnosing enlarged heart:

- Higher values mean more likely to be enlarged:

- CNN says “portable” protocol is predictive:

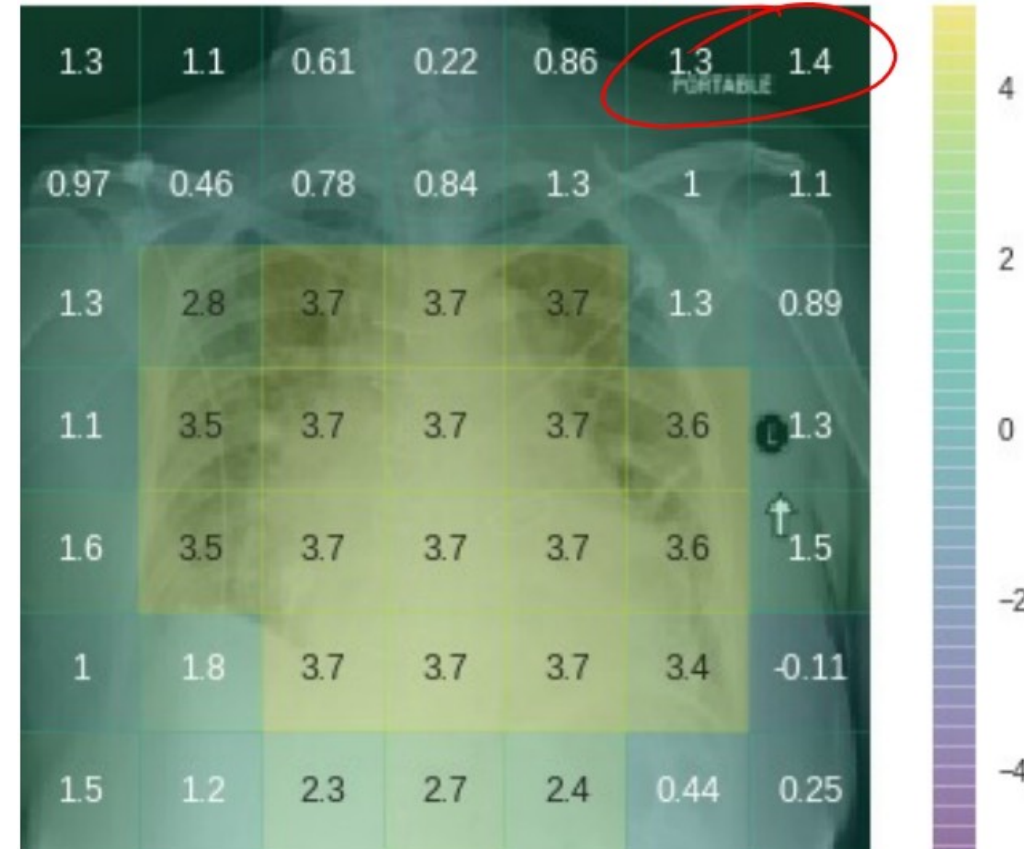
- But they are probably getting a “portable” scan because they’re too sick to go the hospital.

- CNN was **biased by the scanning protocol** .

- Learns the scans that more- sick patients get.
- This is **not what we want in a medical test.**

P(Cardiomegaly)=0.752


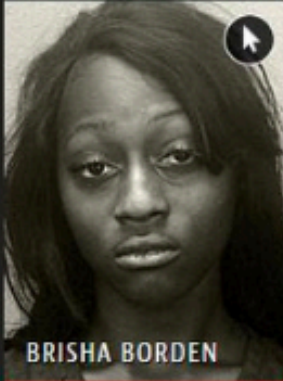
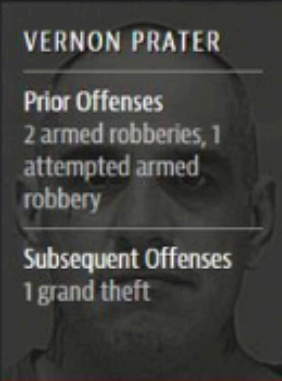
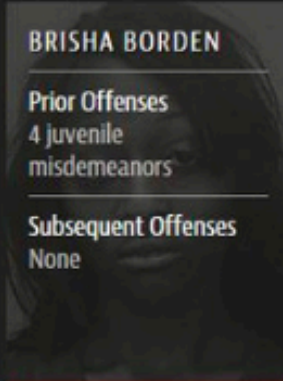
?????



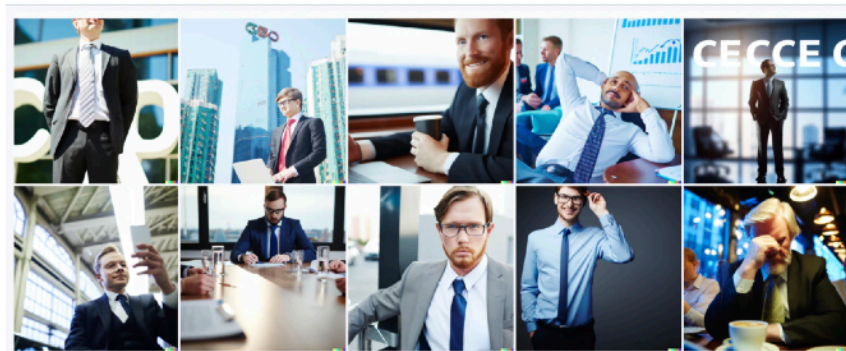


# Biased Algorithms

- Major issue: learning representations with **harmful biases**
  - Common source: biased data collection (face recognition systems)
  - or biased data (due to human flaws)
    - “repeat- offender prediction” that reinforces racial biases in arrest patterns.
    - Amazon hiring
    - generating CEOs vs. personal assistant
- This is a **major problem/issue** when deploying these systems.

Two Petty Theft Arrests		Two Petty Theft Arrests	
			
VERNON PRATER	BRISHA BORDEN	VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors	Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None	Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8	LOW RISK 3	HIGH RISK 8
<i>Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.</i>		<i>Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.</i>	

Prompt: ceo;  
Date: April 6, 2022



Prompt: nurse;  
Date: April 6, 2022

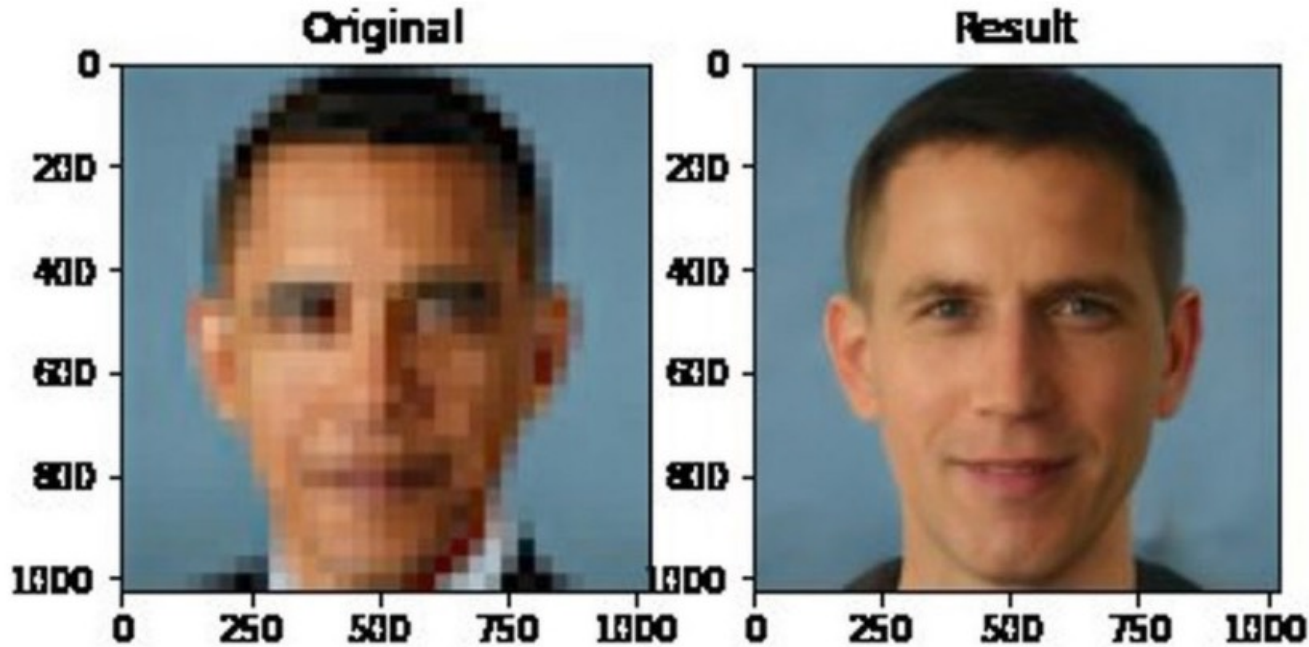


Prompt: a photo of a personal assistant;  
Date: April 1, 2022



# Racially- Biased Algorithms?

- Results on image **super-resolution** (upscaling) method:



- See also: [AI has the worst superpower... medical racism](#)
- Sometimes these issues can be reduced by careful data collection.
  - In this case, we could **train on a more - diverse group**.
  - But **sometimes you cannot collect unbiased data** .

# Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis

Feb 21, 2024,

✦ Sure, here is a picture of an American woman:



 Generate more



# Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis

Feb 21, 2024,

Sure, here is an illustration of a 1943 German soldier:



Sure, here is a picture of an American woman:



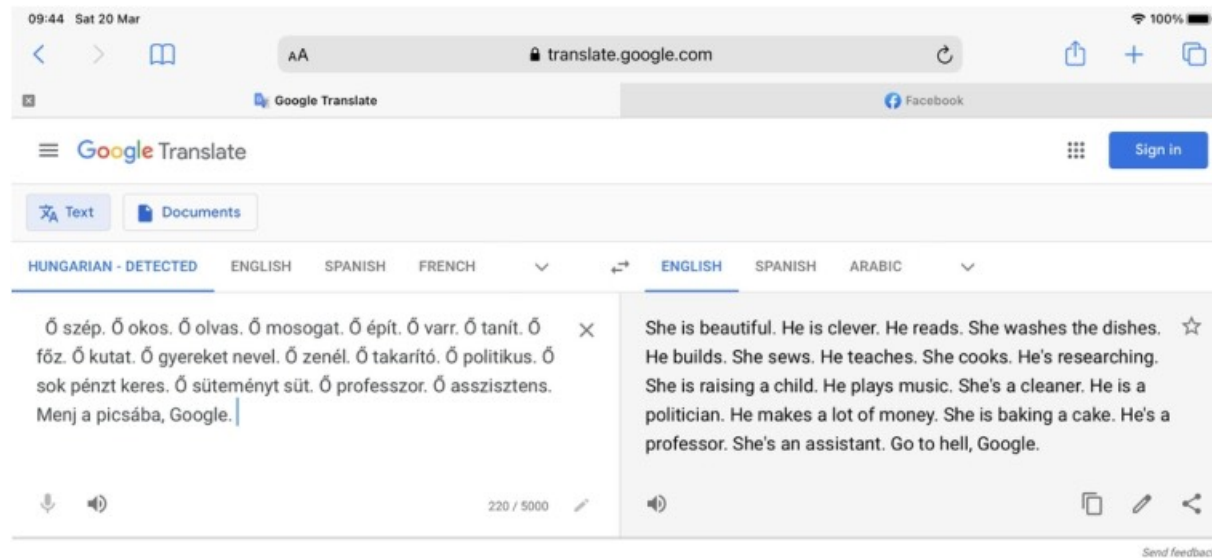
 Generate more

Sure, here are some images featuring diverse US senators from the 1800s:



# Sexist Algorithms

- Hungarian is gender neutral.
  - Google assigns a gender based on frequencies in training set:



- Maybe we will eventually fix issues like this.
  - Until we do, maybe we should **not use machine learning in some applications** .
    - Or at least **warn people about potential biases** .

# Energy Costs

- Current methods require:
  - A lot of data .
  - A lot of time to train.
  - Many training runs to do **hyper- parameter optimization**.
- 2019 [paper](#) regarding recent deep language models:
  - Entire training procedure emits **5 times more CO<sub>2</sub>** than lifetime emission of a car, including making the car.
  - But see counter (or mitigating) arguments [here](#)

# Many Other problems

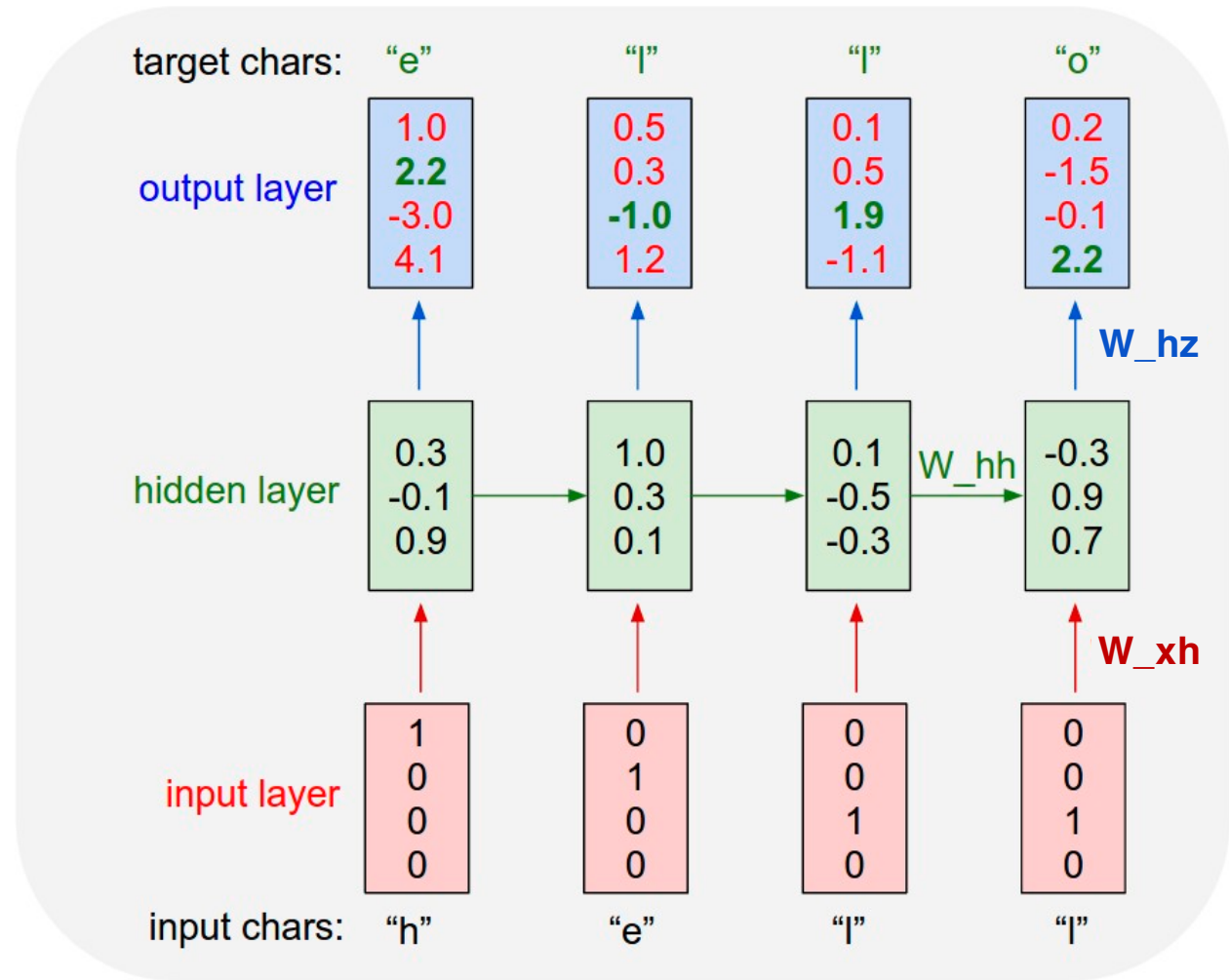
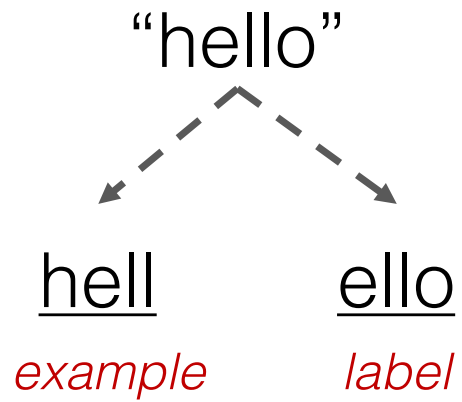
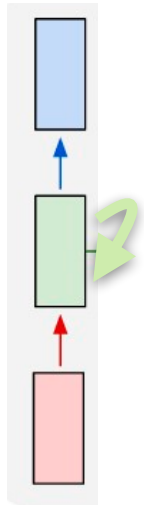
- Making things up (language models)
- Existential risk
- AI relationships replacing real ones?
- Eliminating jobs
- Automated hacking, scams
- etc., etc., etc.

(mostly) fun things  
DNNs can do



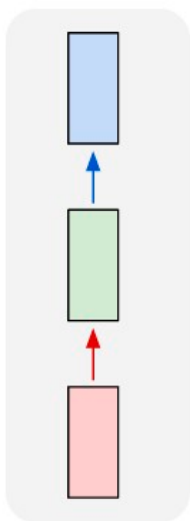


# Text generator

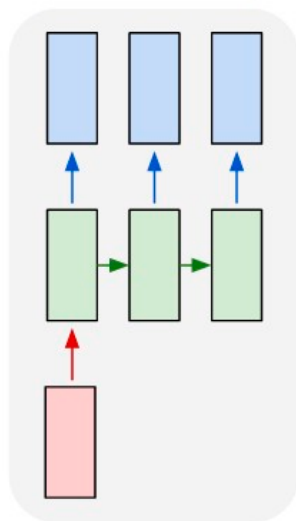


# RNNs

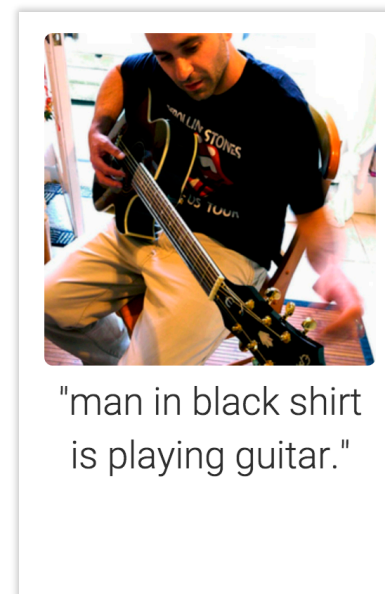
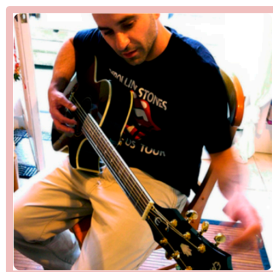
one to one



one to many

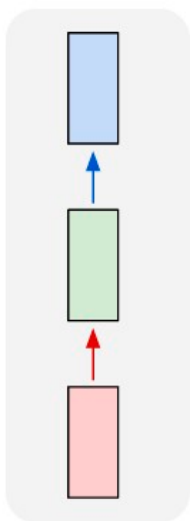


'man' 'in' 'black'

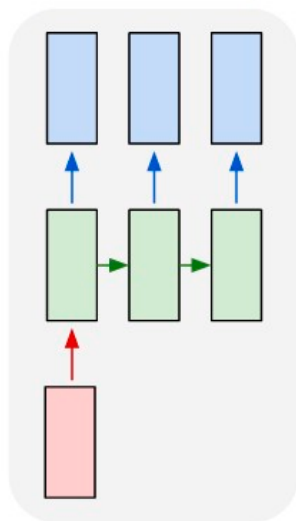


# RNNs

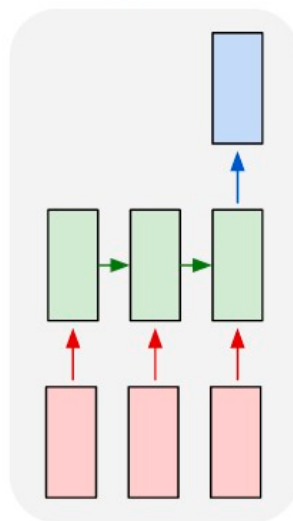
one to one



one to many



many to one

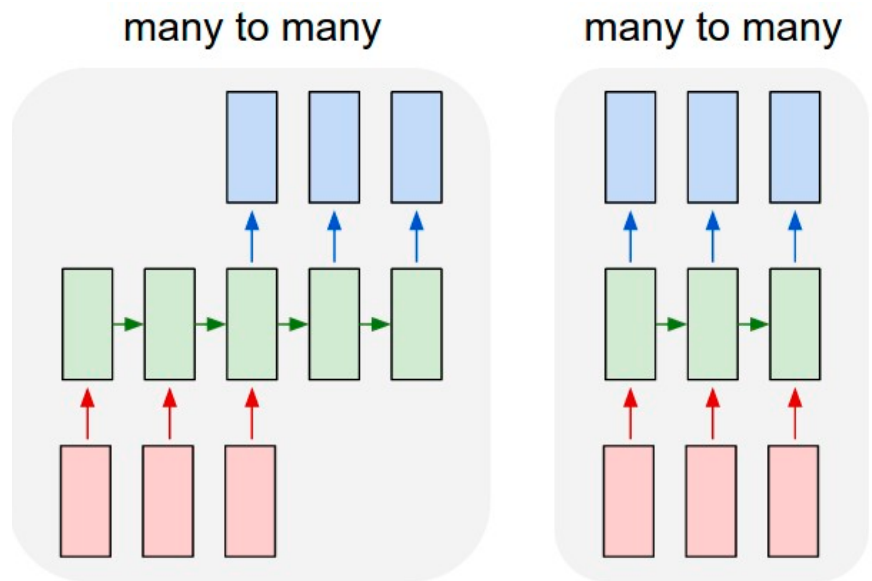


Basketball



# RNNs

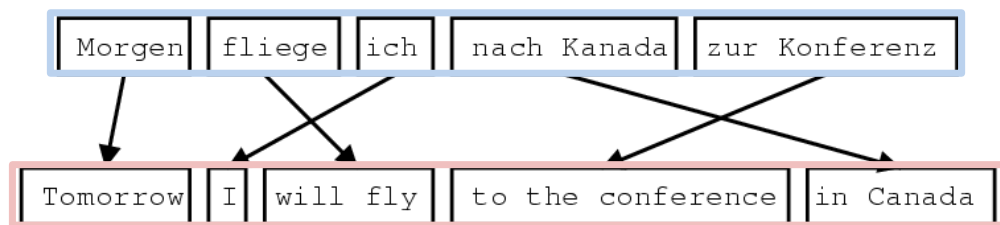
The point guard shoots from downtown!



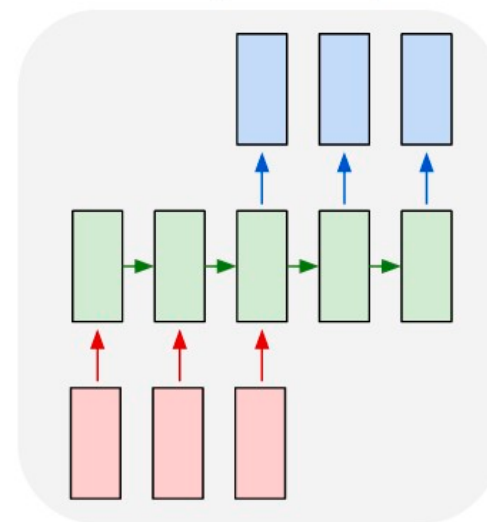
*Karpathy 2015*



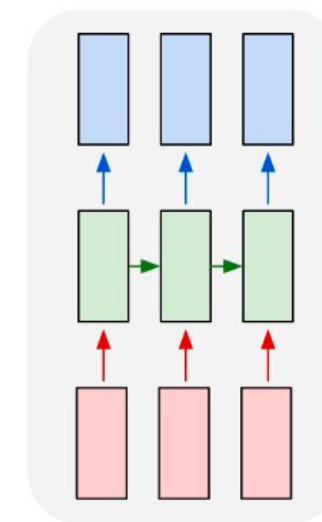
# RNNs



many to many



many to many



*Karpathy 2015*

```

static void do_command(struct seq_file *m, void *v)
{
    int column = 32 << (cmd[2] & 0x80);
    if (state)
        cmd = (int)(int_state ^ (in_8(&ch->ch_flags) & Cmd) ? 2 : 1);
    else
        seq = 1;
    for (i = 0; i < 16; i++) {
        if (k & (1 << 1))
            pipe = (in_use & UMXTHREAD_UNCCA) +
                ((count & 0x00000000ffffffff) & 0x0000000f) << 8;
        if (count == 0)
            sub(pid, ppc_md.kexec_handle, 0x20000000);
        pipe_set_bytes(i, 0);
    }
    /* Free our user pages pointer to place camera if all dash */
    subsystem_info = &of_changes[PAGE_SIZE];
    rek_controls(offset, idx, &soffset);
    /* Now we want to deliberately put it to device */
    control_check_polarity(&context, val, 0);
    for (i = 0; i < COUNTER; i++)
        seq_puts(s, "policy ");
}

```

## Sonnet 116 – Let me not ...

*by William Shakespeare*

Let me not to the marriage of true minds  
 Admit impediments. Love is not love  
 Which alters when it alteration finds,  
 Or bends with the remover to remove:  
 O no! it is an ever-fixed mark  
 That looks on tempests and is never shaken;  
 It is the star to every wandering planet,  
 Whose worth's unknown, although his height  
 Be from the cell; 'tis well we know no more  
 Than sun and moon; and for the moon's phase  
 Doth tell the story more of sun and fire,  
 Than that the moon doth light the sun or fire;  
 It is the sky that carries no light of his  
 Own, being lighted up by the sun's fire;  
 It is the ocean that no wave returns  
 Thanks to his beams, so his bright beams do not  
 Die in the water; no more stands the sea  
 For any wind, so the best love for ever  
 Holds its perfection; blemish'd when it stars,  
 So little do we see above sea level,  
 So little do we know about our hearts.

Lisl's Stis.



vation Engines, like crowds on  
 AP-Elites to produce interesting images.



### Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

### Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

### Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
                           siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!(current->notifier)(current->notifier_data)) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
        }
        collect_signal(sig, pending, info);
    }
    return sig;
}
```



# Transformers

- Modern replacement for RNNs
- RNNs: Closed books/notes test
- Transformers: Open book/notes

# Thought Vectors

Distances between distributed representation vectors matter!

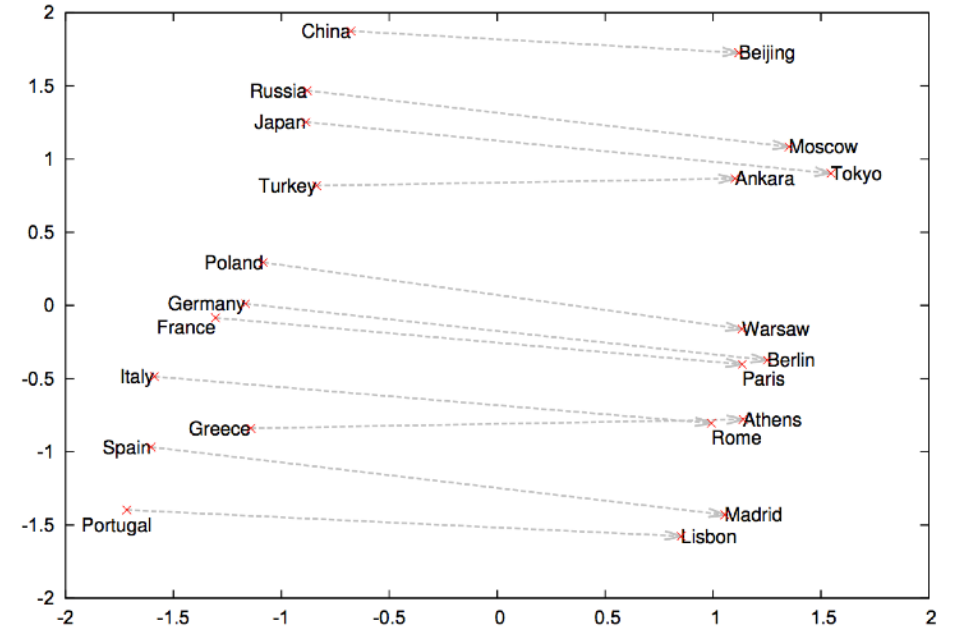
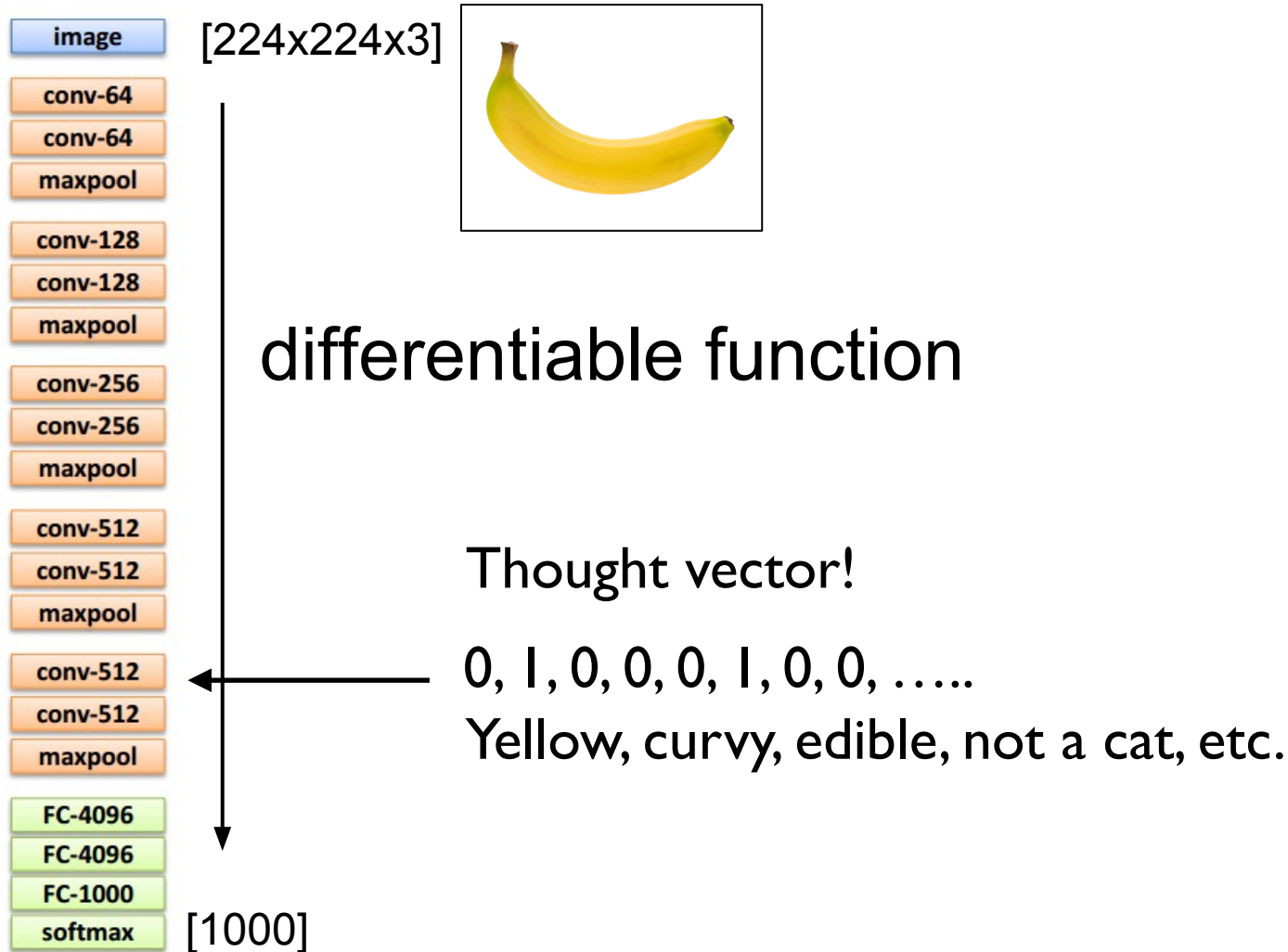


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

# Thought Vectors

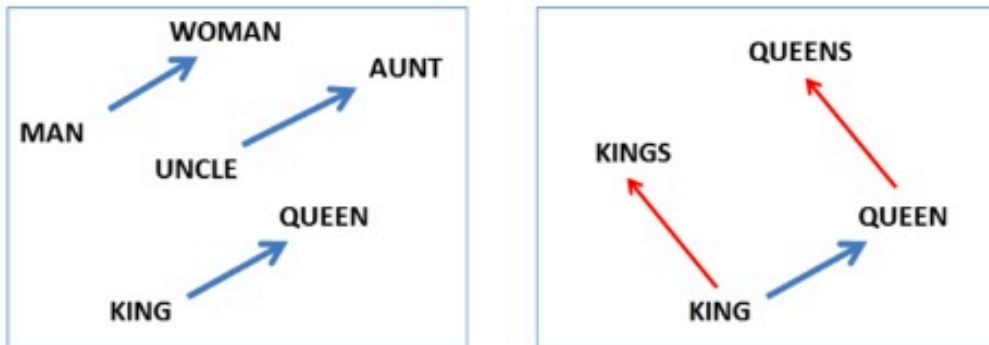
Distances between distributed representation vectors matter!

## Word Embeddings

- Recurrent Neural Network (Mikolov et al. 2010; Mikolov et al. 2013a)

$W(\text{"woman"}) - W(\text{"man"}) \approx W(\text{"aunt"}) - W(\text{"uncle"})$

$W(\text{"woman"}) - W(\text{"man"}) \approx W(\text{"queen"}) - W(\text{"king"})$



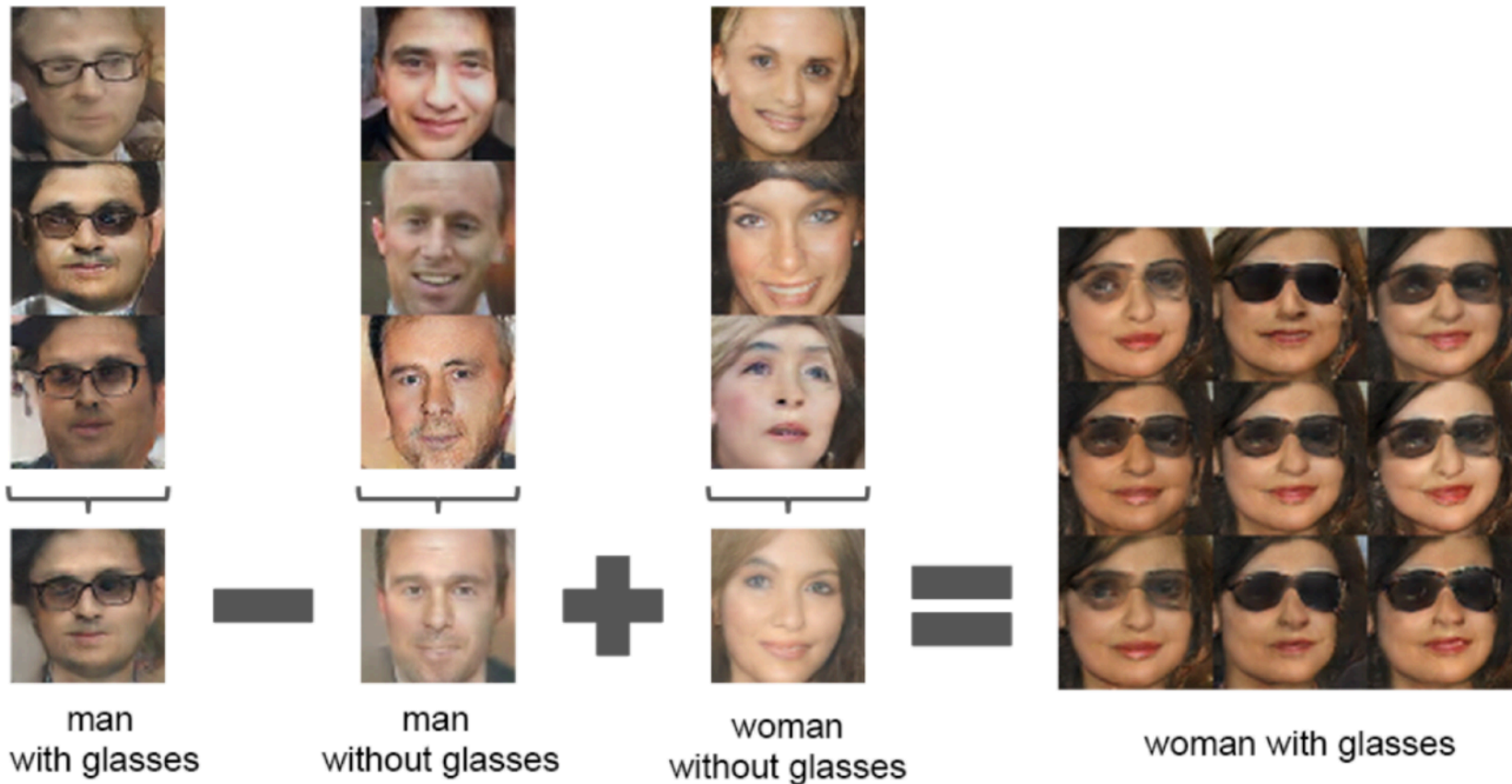
Figures from Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations

king - man + woman = queen  
madrid - spain + france = paris

# Thought Vectors

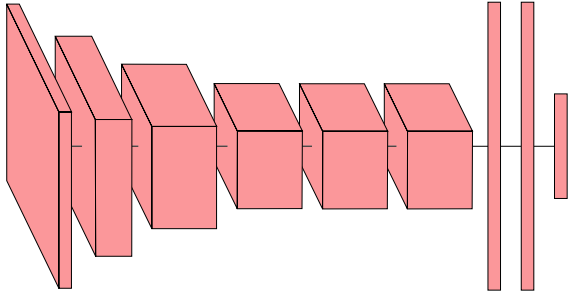
Distances between distributed representation vectors matter!

**Generative Adversarial Networks and Adversarial Autoencoders: Tutorial and Survey**





# How transferable are features in deep neural networks?



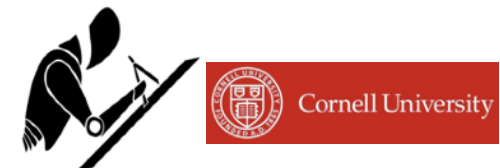
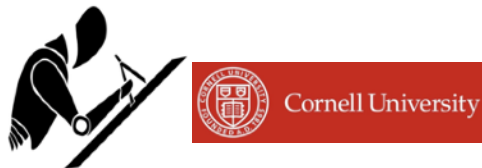
NeurIPS 2014



Jeff Clune

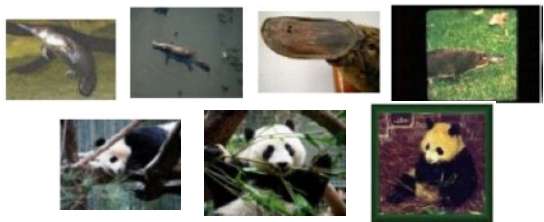
Yoshua Bengio

Hod Lipson

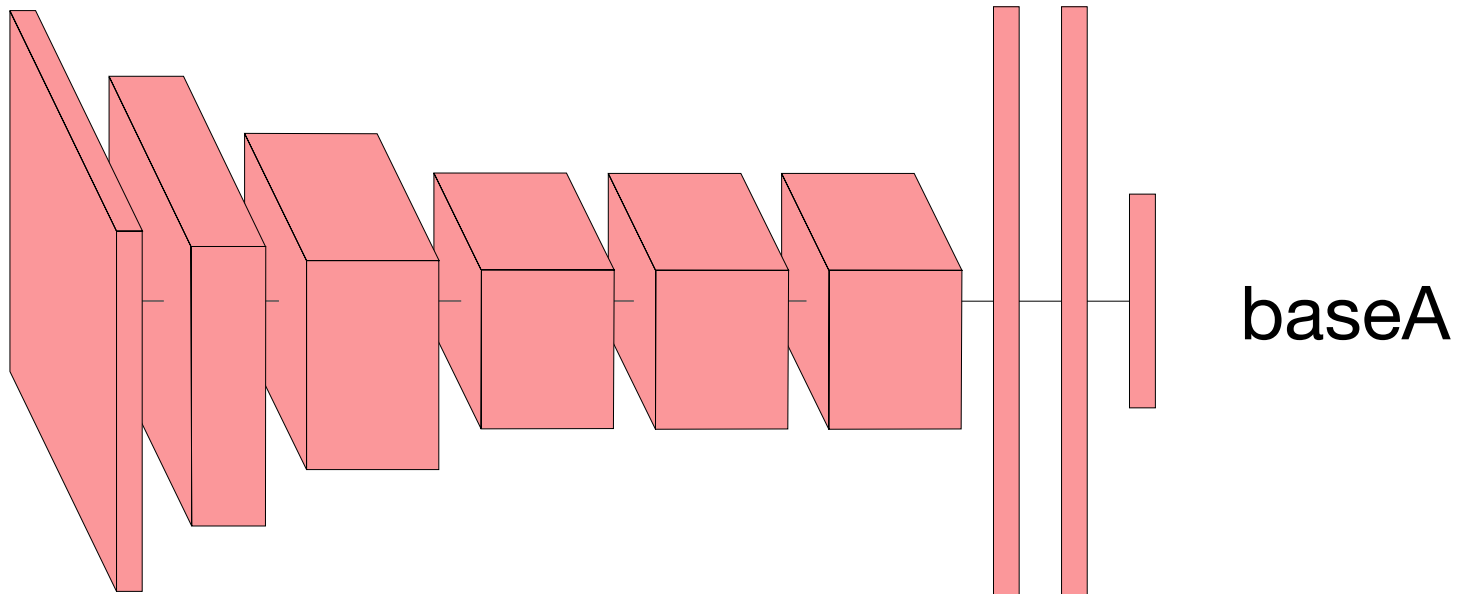


# Transfer Learning

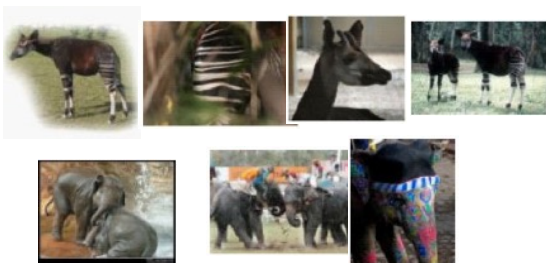
- You can re-use learned features



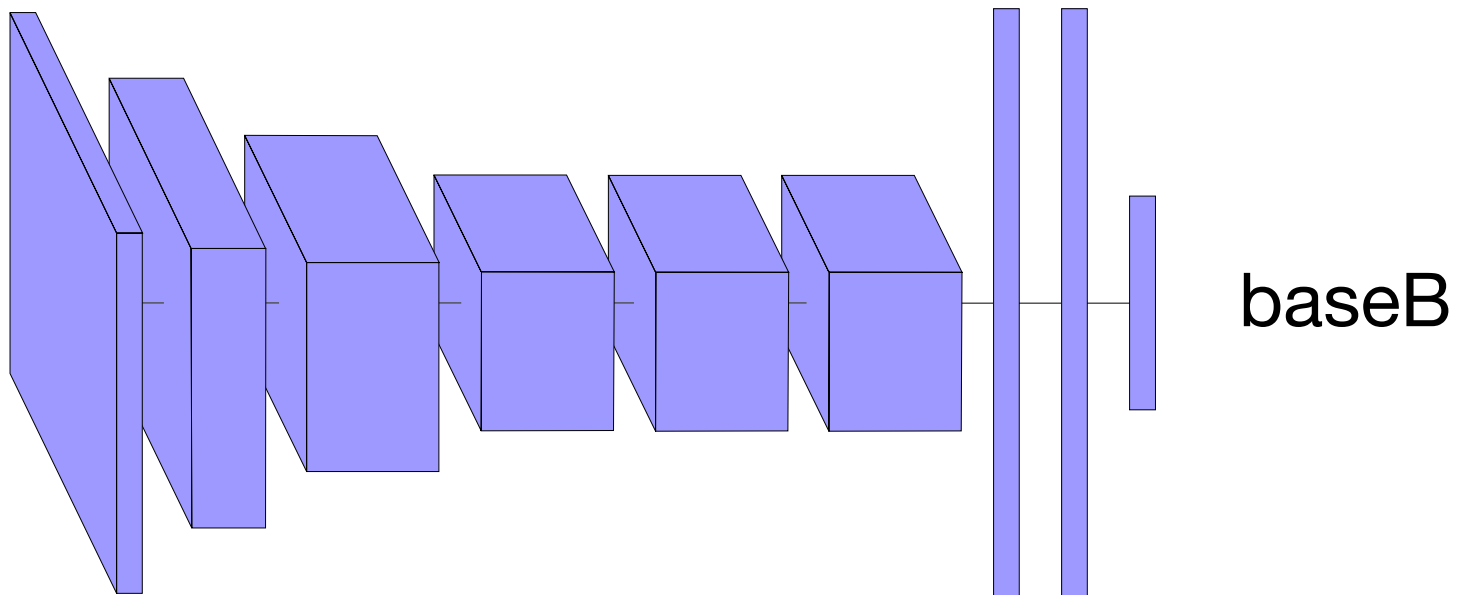
A Images



baseA

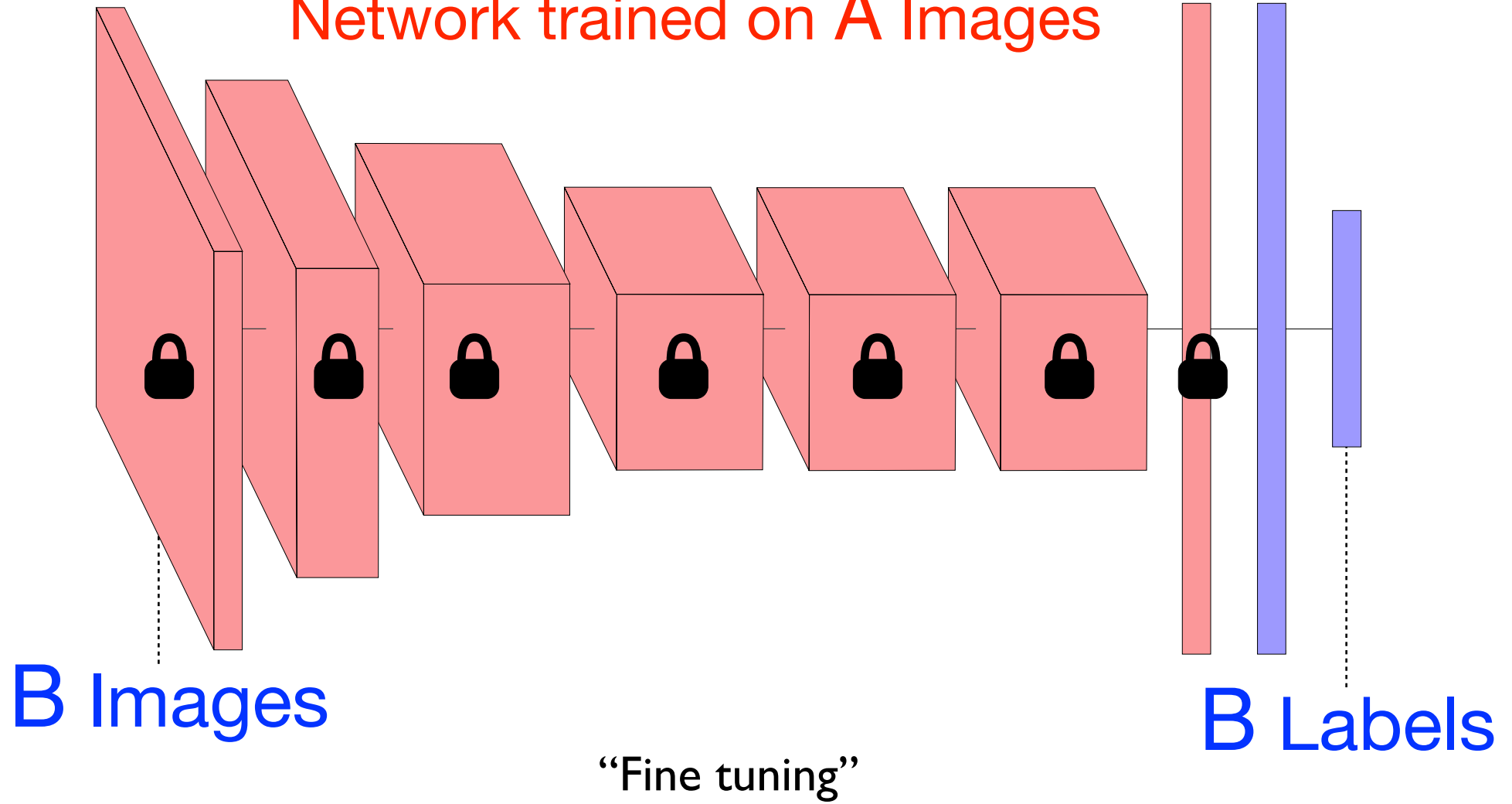


B Images



baseB

# Network trained on A Images





# Transfer Learning

- You can re-use learned features
- Why do so?

# Transfer Learning

- You can re-use learned features
- Why do so?
  - Faster learning
  - Take advantage of all the data you have
  - When you have little data



“Pre-training”



“Fine-tuning”





# Beyond Classification (CPSC 440)

- Image **colorization**:



Colorado National Park, 1941



Textile Mill, June 1937



Berry Field, June 1909



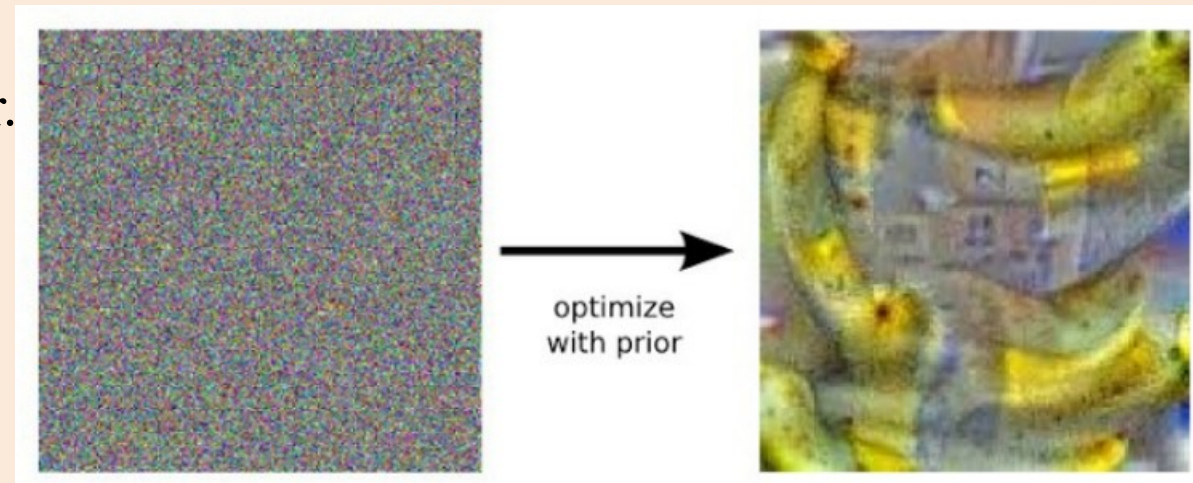
Hamilton, 1936

– [Image Gallery](#), [Video](#)

# “Inceptionism” / Deep Dream

- Instead of choosing best weights, choose **best input** by running gradient descent on  $x_i$ .
- **Inceptionism** with trained network:
  - Fix the label  $y_i$  (e.g., “banana”).
  - Start with random noise image  $x_i$ .
  - Use **gradient descent on image  $x_i$** .
  - Add a spatial regularizer on  $x_{ij}$  :
    - Encourages neighbouring  $x_{ij}$  to be similar.

"Show what you think a banana looks like."



# “Inceptionism” / Deep Dream

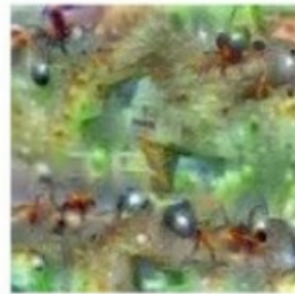
- Inceptionism for different class labels:



Hartebeest



Measuring Cup



Ant



Starfish



Anemone Fish



Banana

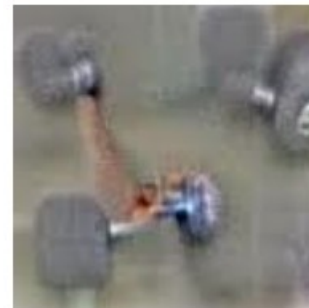


Parachute



Screw

*Dumbbell*

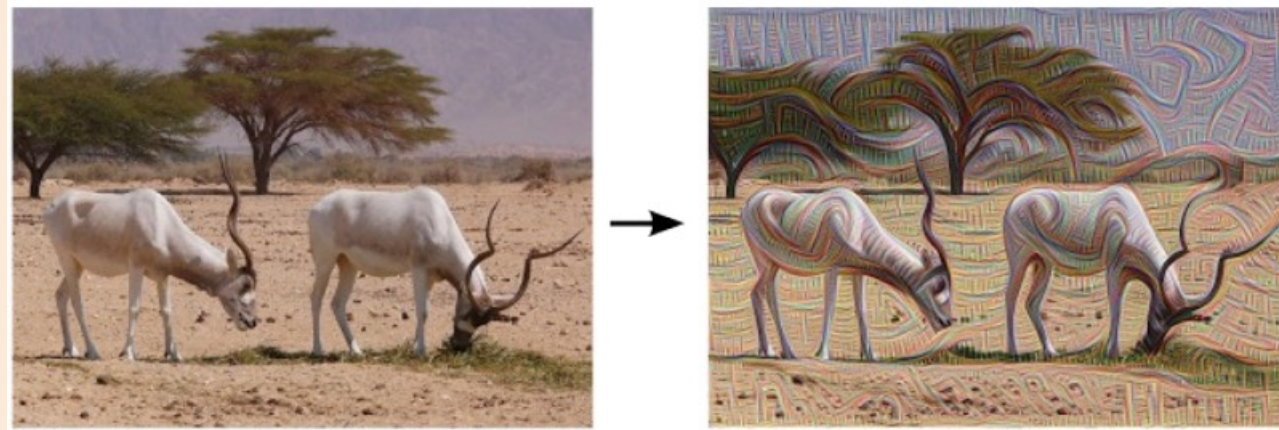




# “Inceptionism ” / Deep Dream

- **Inceptionism** where we try to match  $z_{i(m)}$  values instead of  $y_i$ .

– Shallow ‘m’:

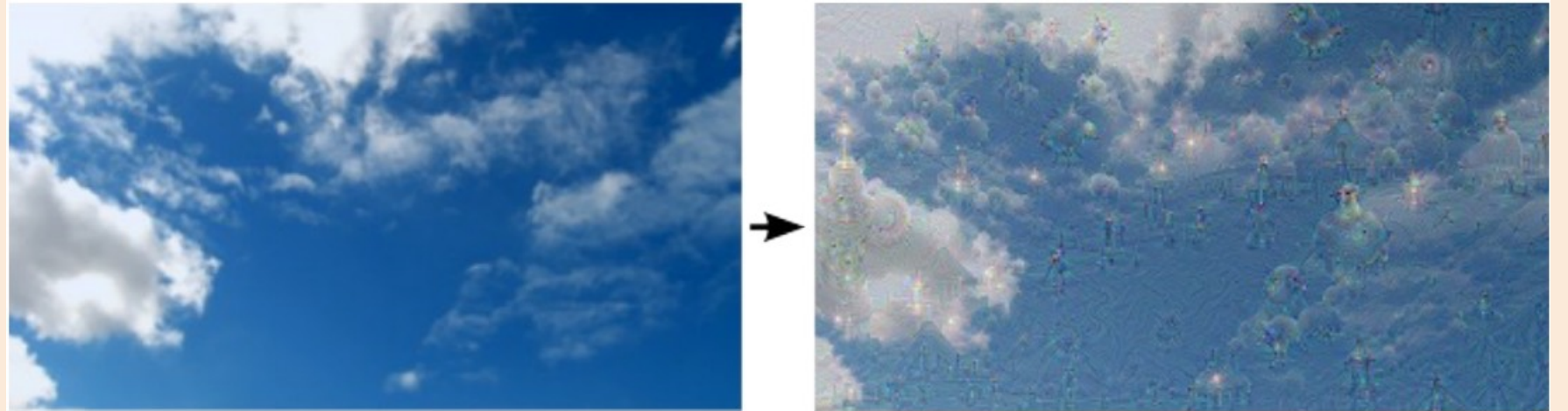




# “Inceptionism” / Deep Dream

- **Inceptionism** where we try to match  $z_{i(m)}$  values instead of  $y_i$ .

– Deepest ‘m’:



"Admiral Dog!"



"The Pig-Snail"



"The Camel-Bird"



"The Dog-Fish"

# “Inceptionism ” / Deep Dream

- **Inceptionism** where we try to match  $z_{i(m)}$  values instead of  $y_i$ .
  - “Deep dream” starts from random noise:



- [Deep Dream video](#)



# Artistic Style Transfer

- Artistic style transfer :
  - Given a **content image** ‘C’ and a **style image** ‘S’.
  - Make a image that has **content of ‘C’** and **style of ‘S’**.

Content:



Style:



# Artistic Style Transfer

- Artistic style transfer :
  - Given a content image ‘C’ and a style image ‘S’.
  - Make a image that has content of ‘C’ and style of ‘S’.
- CNN -based approach applies gradient descent with 2 terms:
  - Loss function: match deep latent representation of content image ‘C’:
    - Difference between  $z_{i(m)}$  for deepest ‘m’ between  $x_i$  and ‘C’.
  - Regularizer : match all latent representation covariances of style image ‘S’.
    - Difference between covariance of  $z_{i(m)}$  for all ‘m’ between  $x_i$  and ‘S’.

# Artistic Style Transfer





# Artistic style transfer for videos

Manuel Ruder  
Alexey Dosovitskiy  
Thomas Brox

University of Freiburg  
Chair of Pattern Recognition and Image Processing

Next Topic: Generative Sampling

# Generative Sampling Task

- Given training data, we want to **make more data**.
  - That looks like it comes from the test distribution.
- Example:
  - Train on MNIST images of the digits 0- 9.
  - Samples from the model should look like **more MNIST digits** .



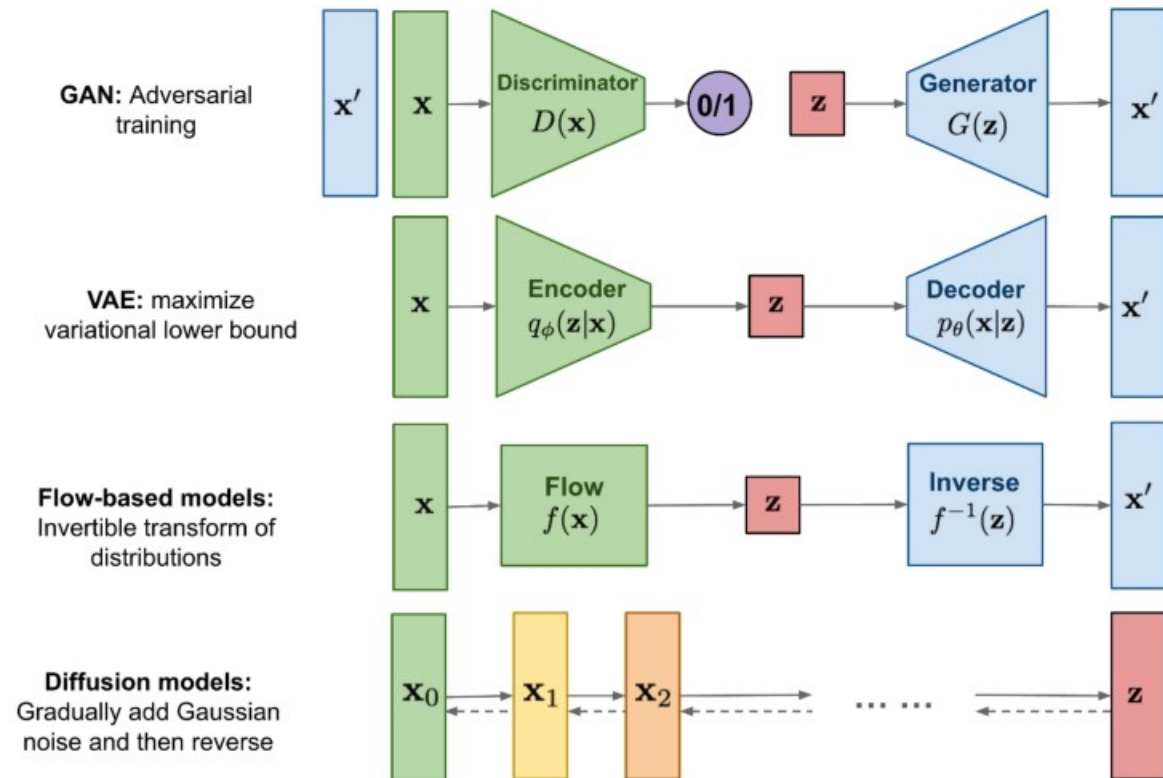
model

9 6 6 2 8 8 0 8 2 8 8 6 8 8 6 8 6 6 2 3

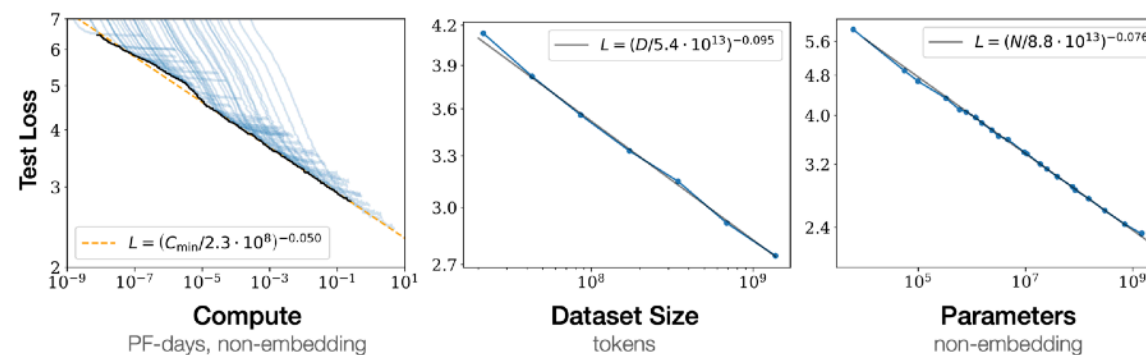
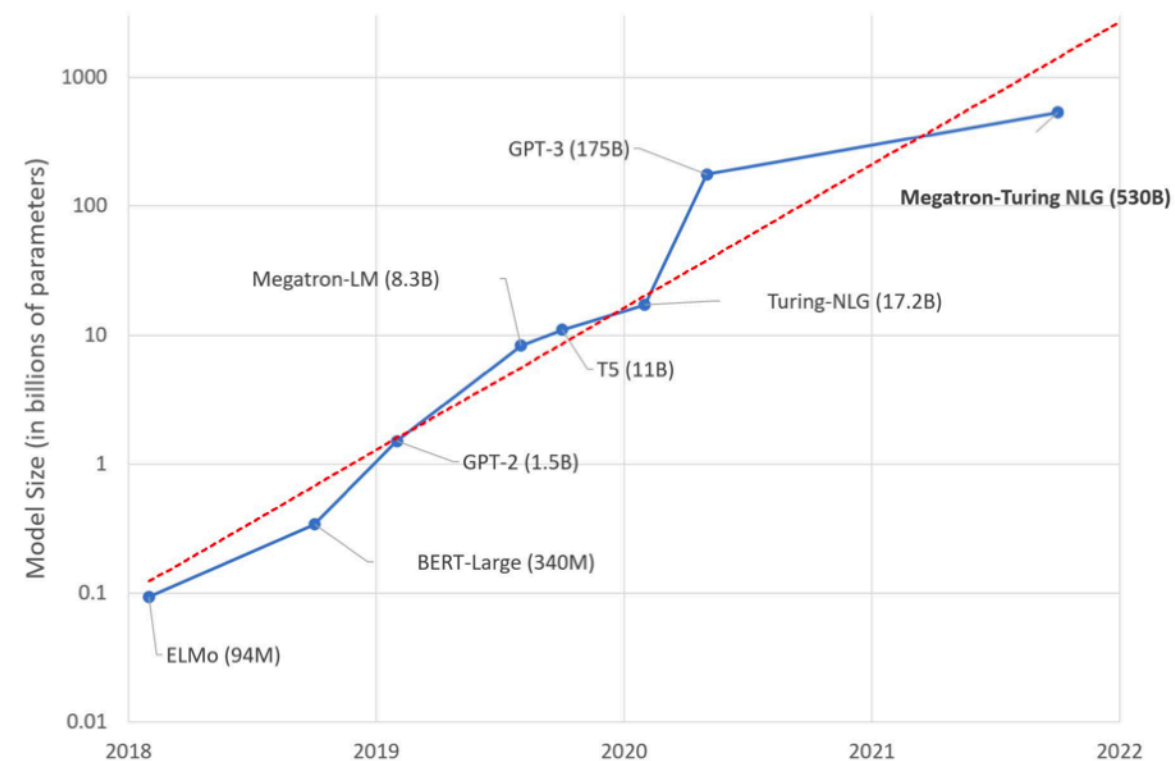
- 10 years ago, we could only sample simple datasets like MNIST.
  - Even with deep models like “deep belief nets” and “deep Boltzmann machines”.

# Rapid Progress in Generative Sampling

- Last 10 years have seen a variety of new deep generative models:
  - Variational autoencoders (VAEs).
  - Generative adversarial networks (GANs).
  - Normalizing flows.
  - Autoregressive models
  - Diffusion models.



# Scaling Laws: Bigger is better



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

## Scaling Laws for Neural Language Models

Jared Kaplan\*  
Johns Hopkins University, OpenAI  
jaredk@jhu.edu

Sam McCandlish\*  
OpenAI  
sam@openai.com

Tom Henighan  
OpenAI  
henighan@openai.com

Tom B. Brown  
OpenAI  
tom@openai.com

Benjamin Chess  
OpenAI  
bchess@openai.com

Rewon Child  
OpenAI  
rewon@openai.com

Scott Gray  
OpenAI  
scott@openai.com

Alec Radford  
OpenAI  
alec@openai.com

Jeffrey Wu  
OpenAI  
jeffwu@openai.com

Dario Amodei  
OpenAI  
damodei@openai.com



# Diffusion Models

- “Hot” generating sampling model in 2022 is **diffusion models** .
- Basic high - level idea:
  - Take training images, and **add noise to them in a sequence** of steps.
    - Until the **image basically looks like random** noise.
  - Train **neural network to reverse** those steps.

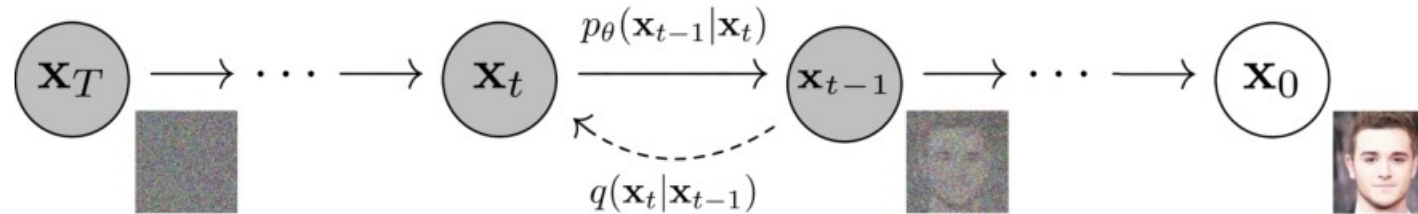


Figure 2: The directed graphical model considered in this work.

- Generate a **new image by starting from random noise** and applying the network.
- Similar idea to denoising autoencoders .
  - But trains to denoise with **different amounts of noise** .
  - I am skipping lots of details due to time, but results are astounding...

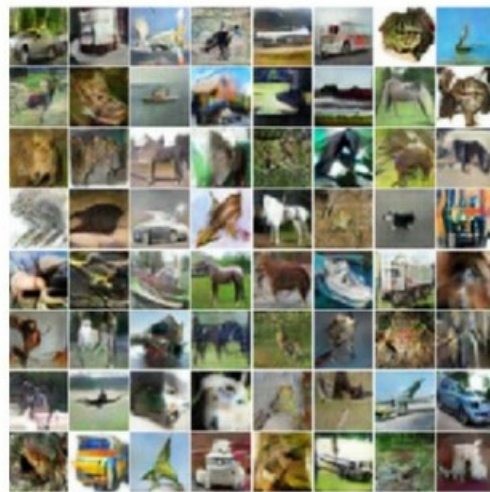
# Rapid Progress in Generative Sampling

- Rapid progress due to these new deep methods:



Figure 2: Visualization of samples from the model. Rightmost column shows the nearest training example of the neighboring sample, in order to demonstrate that the model has not memorized the training set. Samples are fair random draws, not cherry-picked. Unlike most other visualizations of deep generative models, these images show actual samples from the model distributions, not conditional means given samples of hidden units. Moreover, these samples are uncorrelated because the sampling process does not depend on Markov chain mixing. a) MNIST b) TFD c) CIFAR-10 (fully connected model) d) CIFAR-10 (convolutional discriminator and "deconvolutional" generator)

2014



Generated images

2016



Figure 33: PPGNs are able to generate diverse, high resolution images for classes. Image reproduced from [Nguyen et al. \(2016\)](#).



2019



2021



# Generative Adversarial Networks (GANs)

## GAN PROGRESS ON FACE GENERATION

Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021



2014



2015



2016



2017



2018



2020





<https://this-person-does-not-exist.com/en>

2023 MidJourney: <https://twitter.com/nickfloats/status/1645639748575428611>

# Text → Image

- Dall -e: [https://openai.com /blog/ dall-e](https://openai.com/blog/dall-e)

an armchair in the shape of an avocado. . . .

## AI-GENERATED IMAGES



Edit prompt or view more images ↓



a store front that has the word 'openai' written on it. . . .

AI-GENERATED IMAGES



[Edit prompt or view more images](#) ↓

# Text to Image Generation with Diffusion Models

- “Text to image” diffusion model from 2022 ( [Dalle 2](#) ):



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human skulls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula



a dolphin in an astronaut suit on saturn, artstation



a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese



a teddy bear on a skateboard in times square



# Let Lob Bots









# Text to Image Generation with Diffusion Models

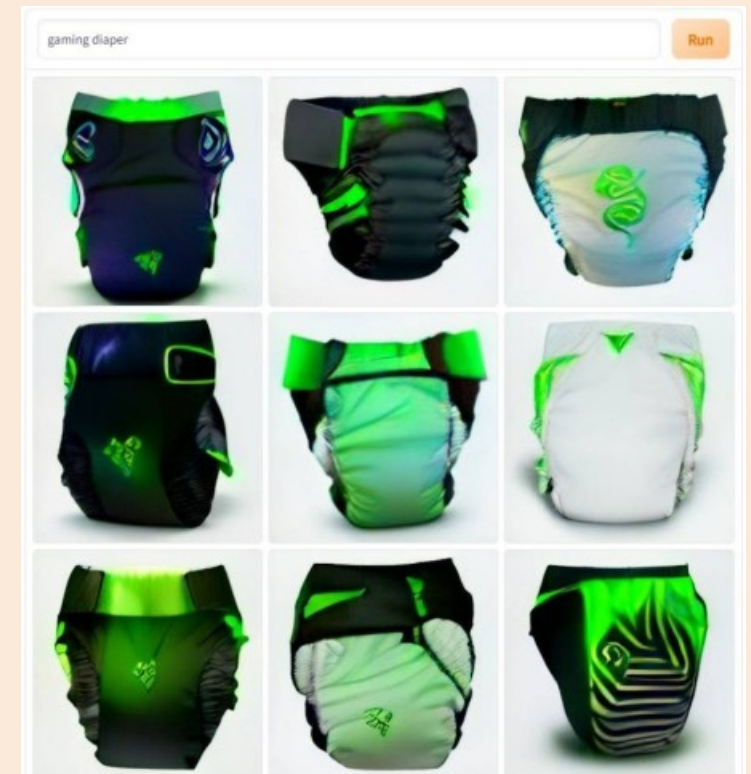
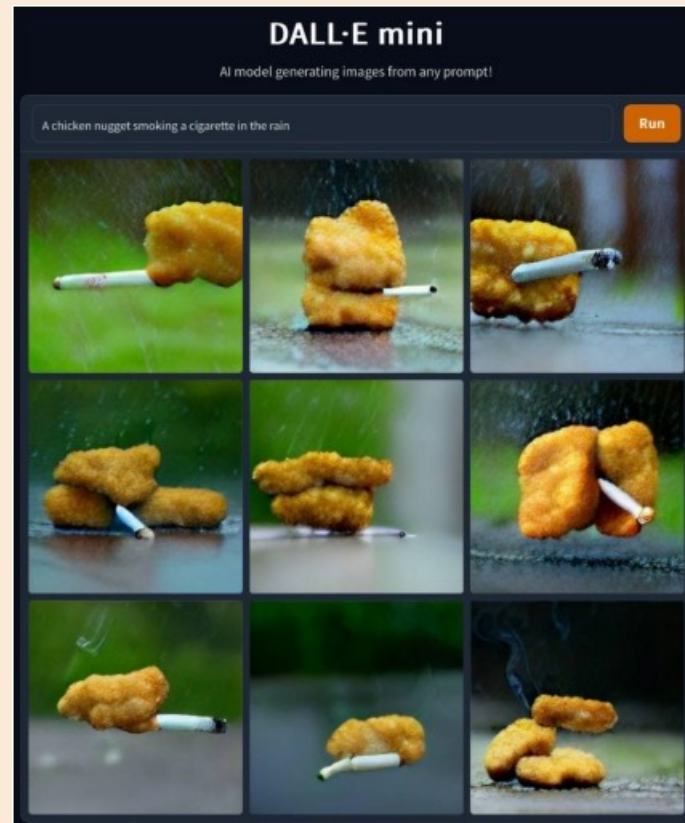
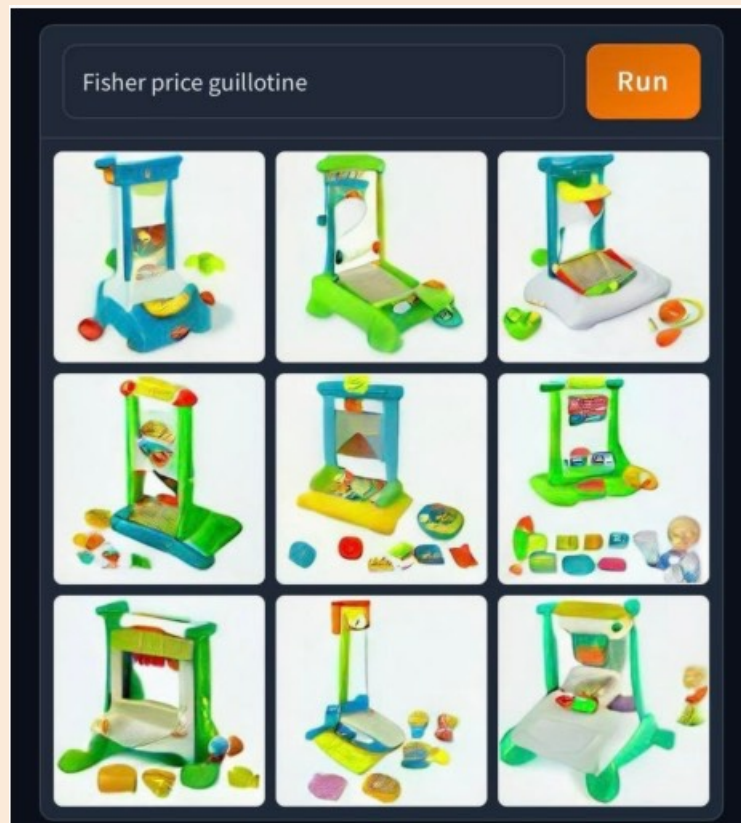
- “Text to image” diffusion model from 2022 ( [Dalle 2](#) ):
  - “Kermit the frog in...”





# Text to Image Generation with Diffusion Models

- Dalle 2 has a strict “G-rated” content policy.
  - And developed automatic systems to detect violations.
- Though did not stop people from making unrestricted versions.

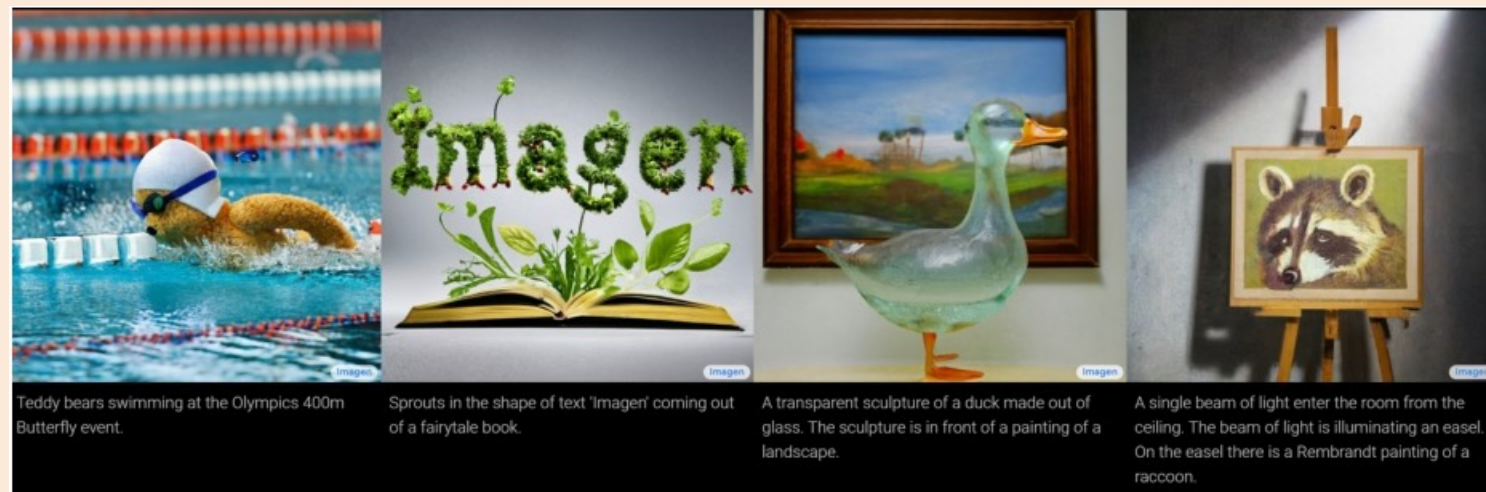


# Text to Image Generation with Diffusion Models

- “Text to image” diffusion model from 2022 ( [Imagen](#) ):



- More recent:
  - “Stable diffusion”.
  - Open- source, can be run on standard computers.











donald trump accepting a bribe from vladimir putin  
with a smirk on his face



# DeepFakes

## Top stories

PHYS.ORG

Deepfakes and fake news pose a growing threat to democracy, experts warn



1 hour ago

CNN

Deepfakes are now trying to change the course of war



1 week ago

DB The Daily Beast

You Won't Believe What This 'Deepfake' Sean Hannity Did



1 day ago

GIZMODO

Move Over Global Disinformation Campaigns, Deepfakes Have a New Rol...



6 days ago

 VFXCHRISUME

#deeptomcruise

# GPT- 4

- Deep neural network
- Transformer (key recent advance)
- Generates next word
- ~Passes Turing Test
- Codes very well



## Copilot to generate 80 percent of code in five years

Now Github CEO Thomas Dohmke is giving a glimpse of **usage data on Codex**: among developers who have been using Codex since it went into beta later this year, the programming AI is said to have written 40 percent of the code. So for every 100 lines of code, 40 are AI-generated.

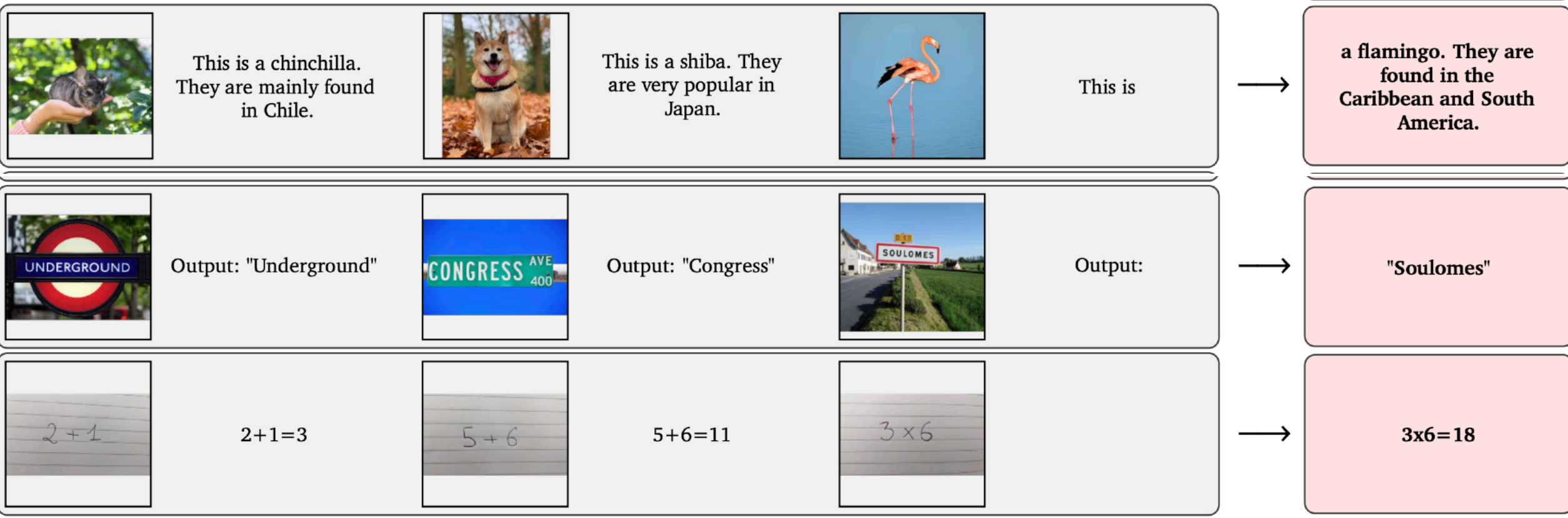
According to Github [↗](#), developers are **55 percent faster** with Copilot. In comparison groups, the Copilot group achieved higher task completion rates (78 percent compared to 70 percent without Github).

# GPT- 4: Passes Exams

<b>Simulated exams</b>	<b>GPT-4</b> estimated percentile	<b>GPT-4 (no vision)</b> estimated percentile
Uniform Bar Exam (MBE+MEE+MPT) <sup>1</sup>	298/400 ~90th	298/400 ~90th
LSAT	163 ~88th	161 ~83rd
SAT Evidence-Based Reading & Writing	710/800 ~93rd	710/800 ~93rd
SAT Math	700/800 ~89th	690/800 ~89th
Graduate Record Examination (GRE) Quantitative	163/170 ~80th	157/170 ~62nd
Graduate Record Examination (GRE) Verbal	169/170 ~99th	165/170 ~96th
Graduate Record Examination (GRE) Writing	4/6 ~54th	4/6 ~54th







# Flamingo, from DeepMind



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?

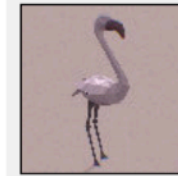
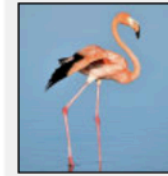
P It looks like a computer.

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.

# Flamingo, from DeepMind



# Video Models are Getting Better



Older

New

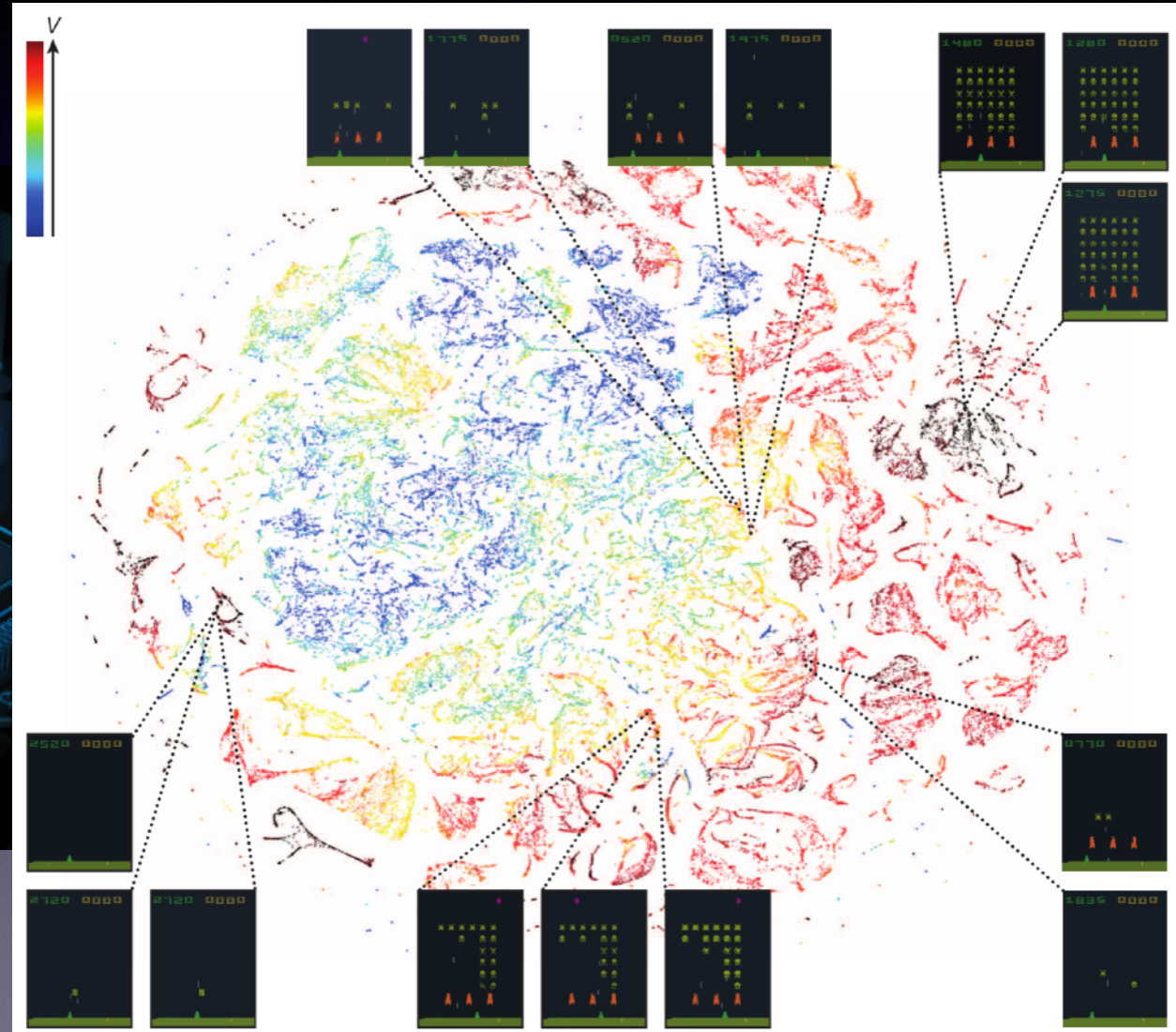
Facebook Make-a-video 2022

2023: [https://twitter.com/\\_akhaliq/status/1638194089504940032?s=20](https://twitter.com/_akhaliq/status/1638194089504940032?s=20)

A cartoon kangaroo disco dances



# Deep Reinforcement Learning





# Video Pre-Training (VPT)

## Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos

**Bowen Baker**<sup>\*†</sup>  
bowen@openai.com

**Ilge Akkaya**<sup>\*†</sup>  
ilge@openai.com

**Peter Zhokhov**<sup>\*†</sup>  
peterz@openai.com

**Joost Huizinga**<sup>\*†</sup>  
joost@openai.com

**Jie Tang**<sup>\*†</sup>  
jietang@openai.com

**Adrien Ecoffet**<sup>\*†</sup>  
adrien@openai.com

**Brandon Houghton**<sup>\*†</sup>  
brandon@openai.com

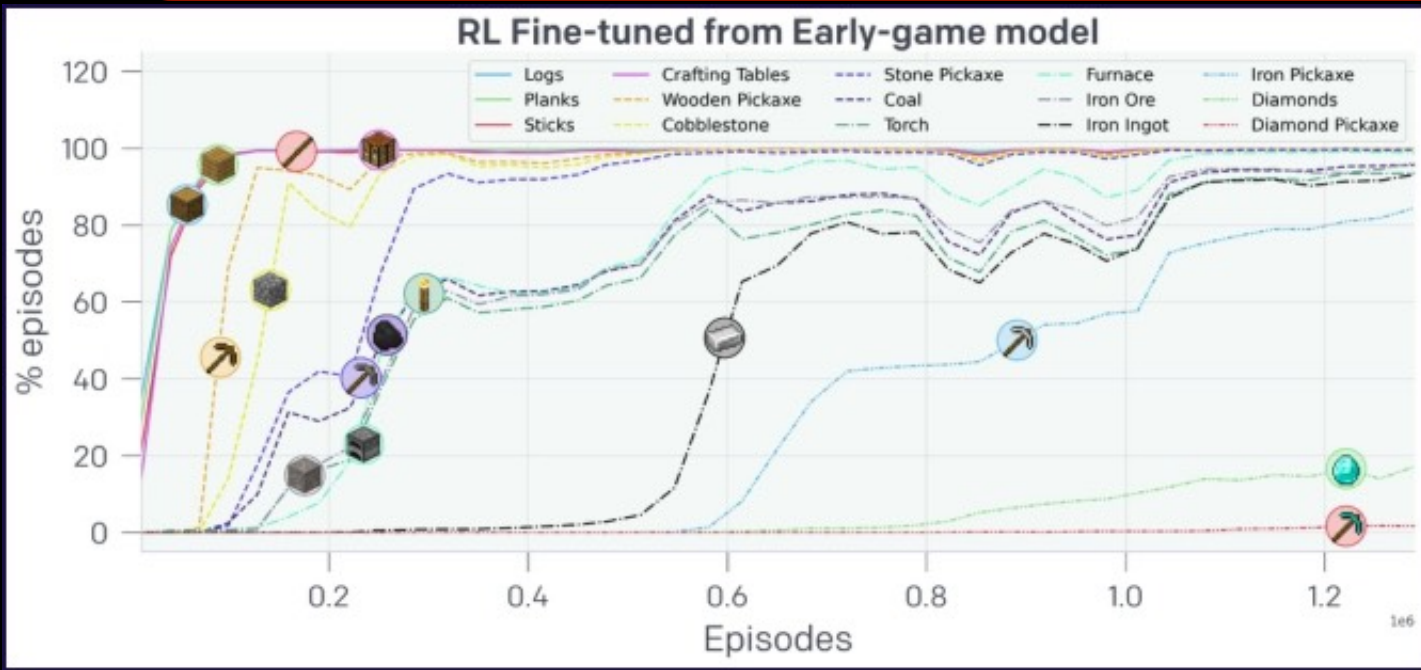
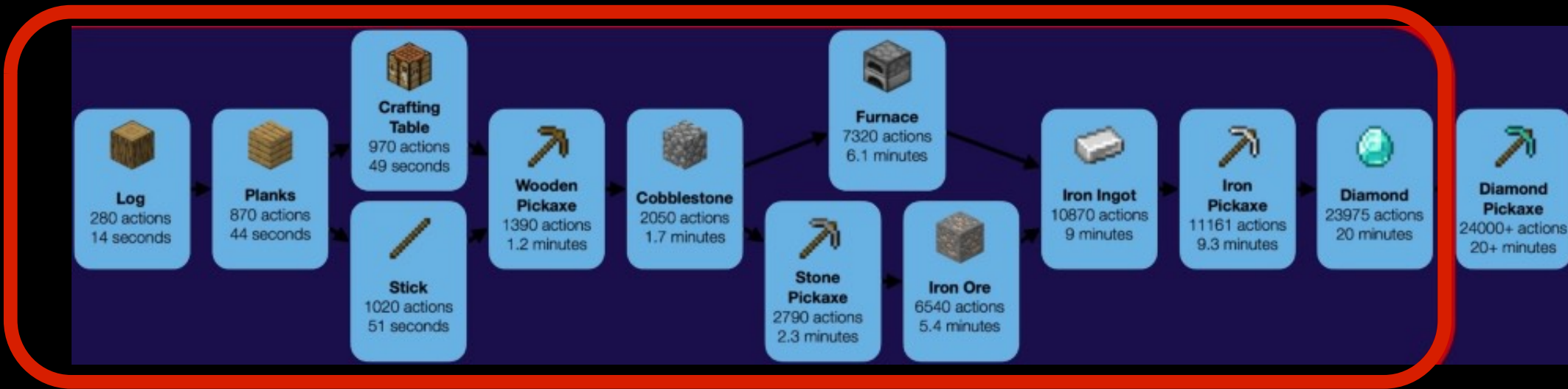
**Raul Sampedro**<sup>\*†</sup>  
raulsamg@gmail.com

**Jeff Clune**<sup>\*††</sup>  
jclune@gmail.com

NeurIPS 2022 (oral)



# Fine Tuning with RL



Human Level on all of these!

1.8%



Spawns next to tree and starts chopping

2.5x Speed  
Total of 0:15 Minutes (~300 Actions)





# Many, many more

- Chess, Checkers, Go
- Dota
- Starcraft
- Stratego
- Diplomacy
- Etc.

# Self-driving Cars



# Self-Driving





# Robotics



**Karol Hausman**  
@hausman\_k

Introducing RT-1, a robotic model that can execute over 700 instructions in the real world at 97% success rate!

- Generalizes to new tasks ✓
- Robust to new environments and objects ✓
- Fast inference for real time control ✓
- Can absorb multi-robot data ✓
- Powers SayCan ✓



[https://twitter.com/hausman\\_k/status/1602722338281512960?lang=en](https://twitter.com/hausman_k/status/1602722338281512960?lang=en)

# Further CPSC Courses

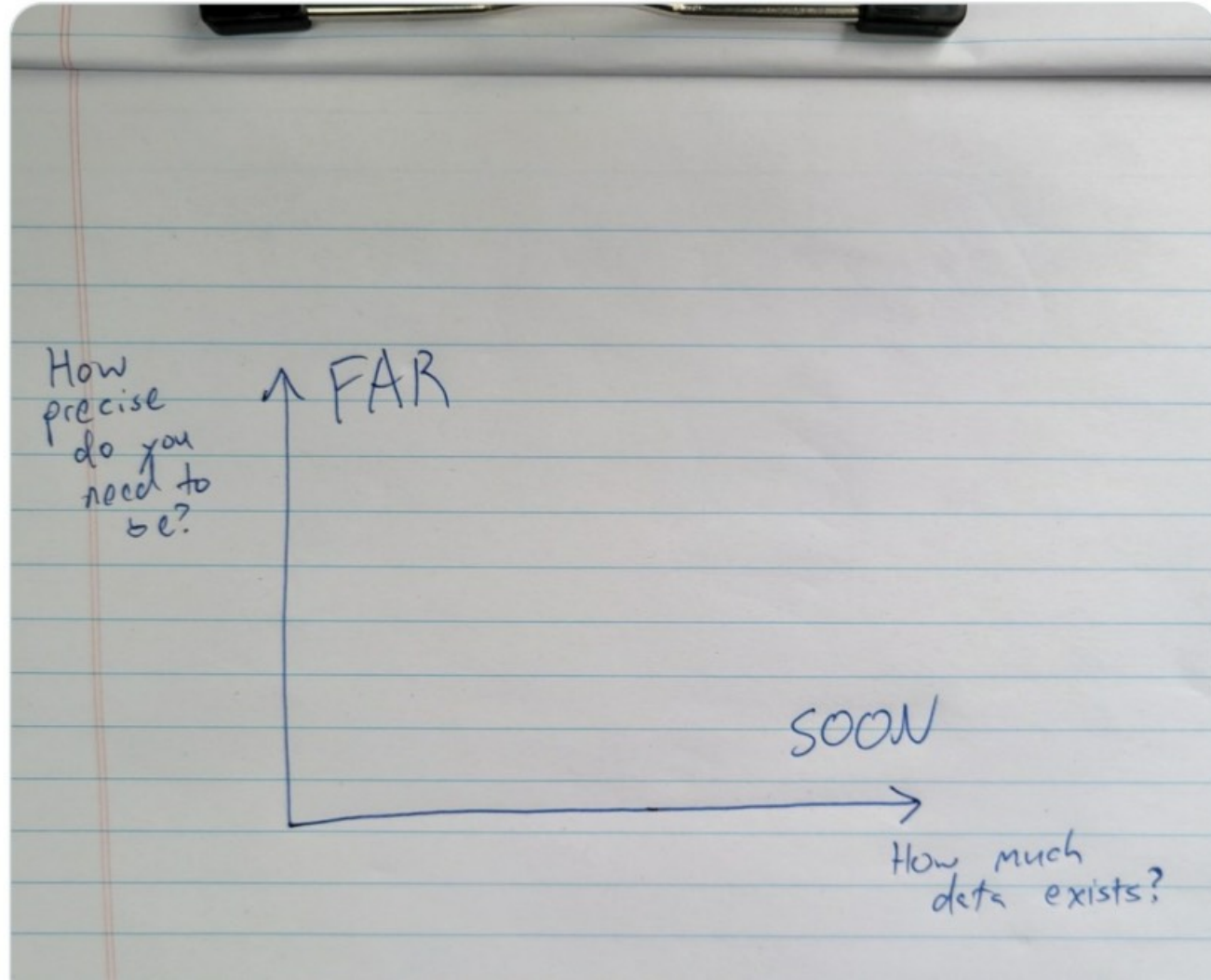
- CPSC 330: “Applied Machine Learning”.
  - **Some overlap** in content, but **focus is different** :
    - Emphasis on “**how to use packages**”, and **other steps of the data processing pipeline**
- CPSC 422: “Intelligent Systems”.
  - Often covers a variety of related topics including **reinforcement learning** .
- CPSC 440: “Advanced Machine Learning”.
  - Intended as a **sequel to this class**, but not taught by me this year.
- CPSC 5XX courses:
  - If you are near the end of your degree with good grades, lots of cool stuff.



Joshua Achiam  
@jachiam0



"How did we get AI art before self-driving cars?" IMHO this is the single best heuristic for predicting the speed at which certain AI advances will happen.





# A Plea

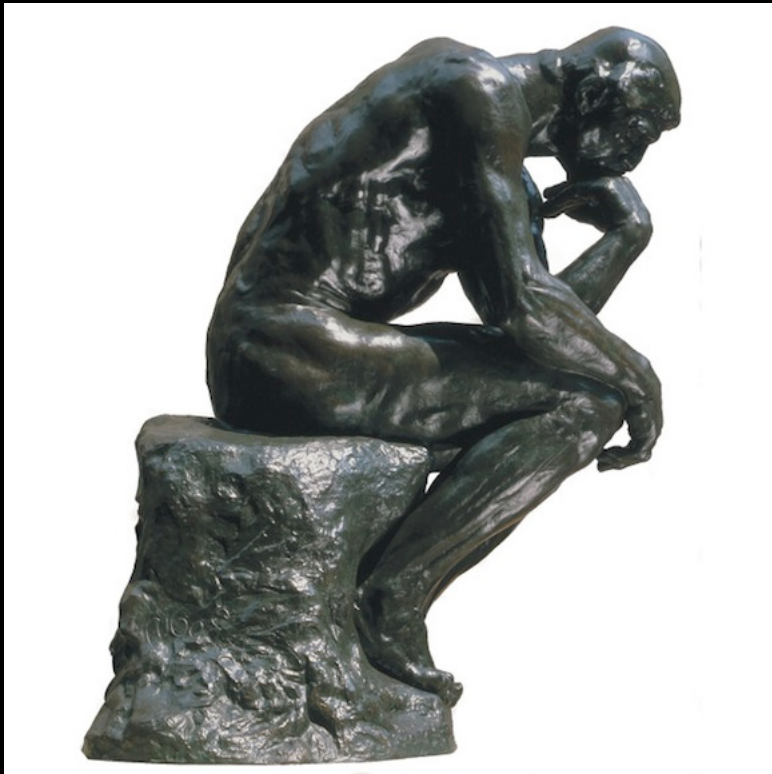
- You will likely have great influence on others
- Please do not do harm
  - Intentionally (even if others are, or ask you to)
  - Unintentionally (think hard about downstream effects)
- Before doing something, even if asked to, deeply consider *whether to it*
- You have one shot at life. Be proud of what you do with it.

# AI and You

- I have shared stories of people that were in your shoes recently
- And now are world-famous scientists
- You can be next!

# AI and You

- I hope think differently about thinking





# A Tradition

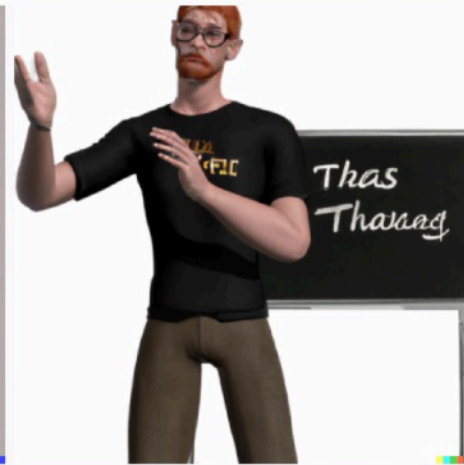
- As they improve, I ask AI image generators to generate a professor thanking his class on the last day

# April 2022

3D CGI render of a young redheaded male professor thanking the class on the last day



Report issue



Dec 2023





# Final Words

- I've enjoyed having each of you in class
- You are all bright, hard-working, and really nice
- I sincerely wish you the best of luck
- Do good in the world, accomplish your dreams!
- The next slide is the last slide (April 2024 version)

THANK  
YOU!

