

CS163: Deep Learning for Computer Vision

Lecture 1: Course Introduction

Bolei Zhou

Deep Learning for Computer Vision

Deep Learning for Computer Vision

Building artificial systems
that process, perceive, and
reason about visual data

Deep Learning for Computer Vision

Building artificial systems
that process, perceive, and
reason about **visual data**

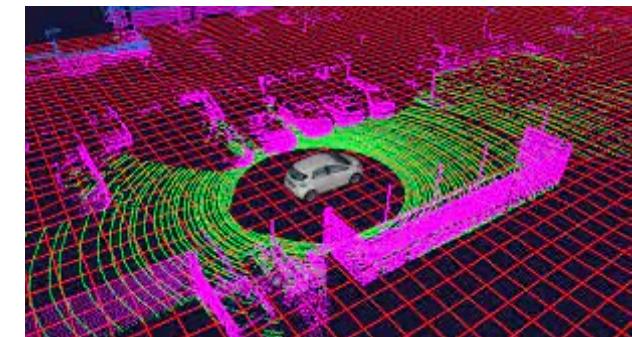
image



video



3D point clouds



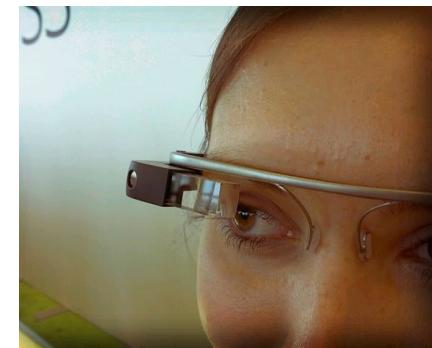
Computer Vision is everywhere!



Left to right:
[Image by Roger H Goun](#) is licensed under CC BY 2.0
[Image](#) is CCO 1.0 public domain
[Image](#) is CCO 1.0 public domain
[Image](#) is CCO 1.0 public domain

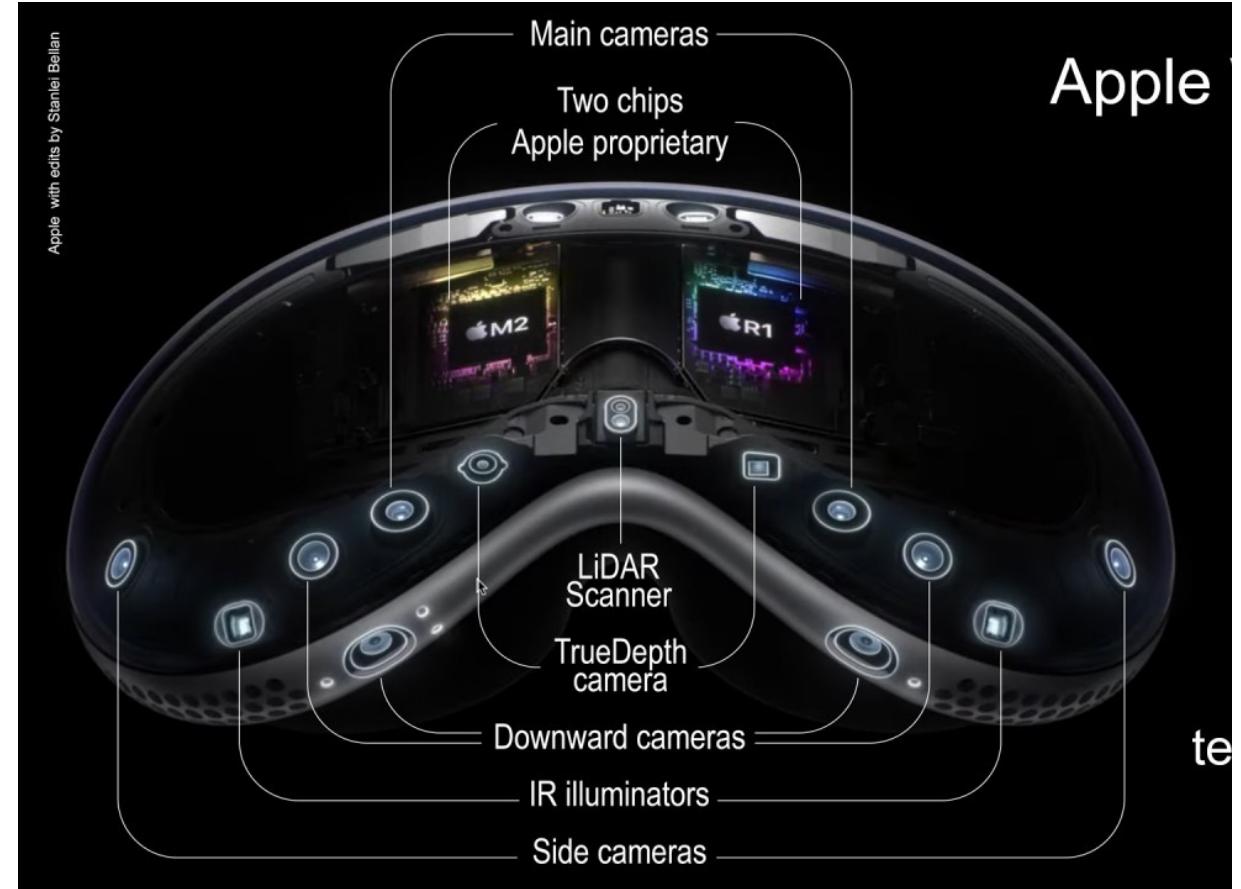


Left to right:
[Image is free to use](#)
[Image](#) is CCO 1.0 public domain
[Image by NASA](#) is licensed under CC BY 2.0
[Image](#) is CCO 1.0 public domain



Bottom row, left to right:
[Image](#) is CCO 1.0 public domain
[Image by Derek Keats](#) is licensed under CC BY 2.0; changes made
[Image](#) is public domain
[Image](#) is licensed under CC-BY 2.0; changes made

Apple's Vision Pro will be on market on Feb.2



Deep Learning for Computer Vision

Building artificial systems that
learn from data and experience

Deep Learning for Computer Vision

Hierarchical learning algorithms
with many “layers”, (very) loosely
inspired by the brain

Artificial Intelligence

Artificial Intelligence

Machine Learning

Computer
Vision

Artificial Intelligence

Computer
Vision

Machine Learning

Deep
Learning

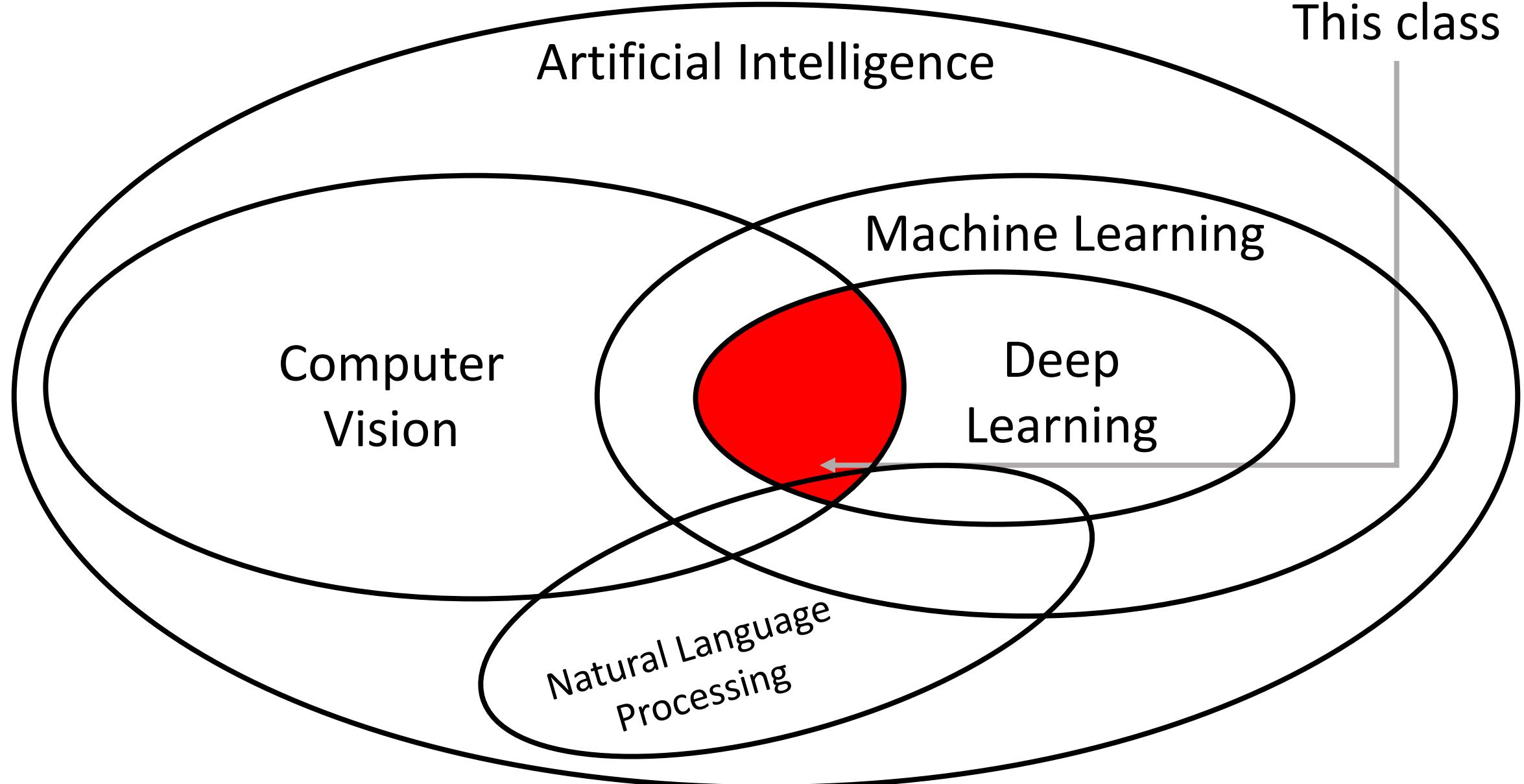
Artificial Intelligence

Computer
Vision

Machine Learning

Deep
Learning

This class



Artificial Intelligence

This class

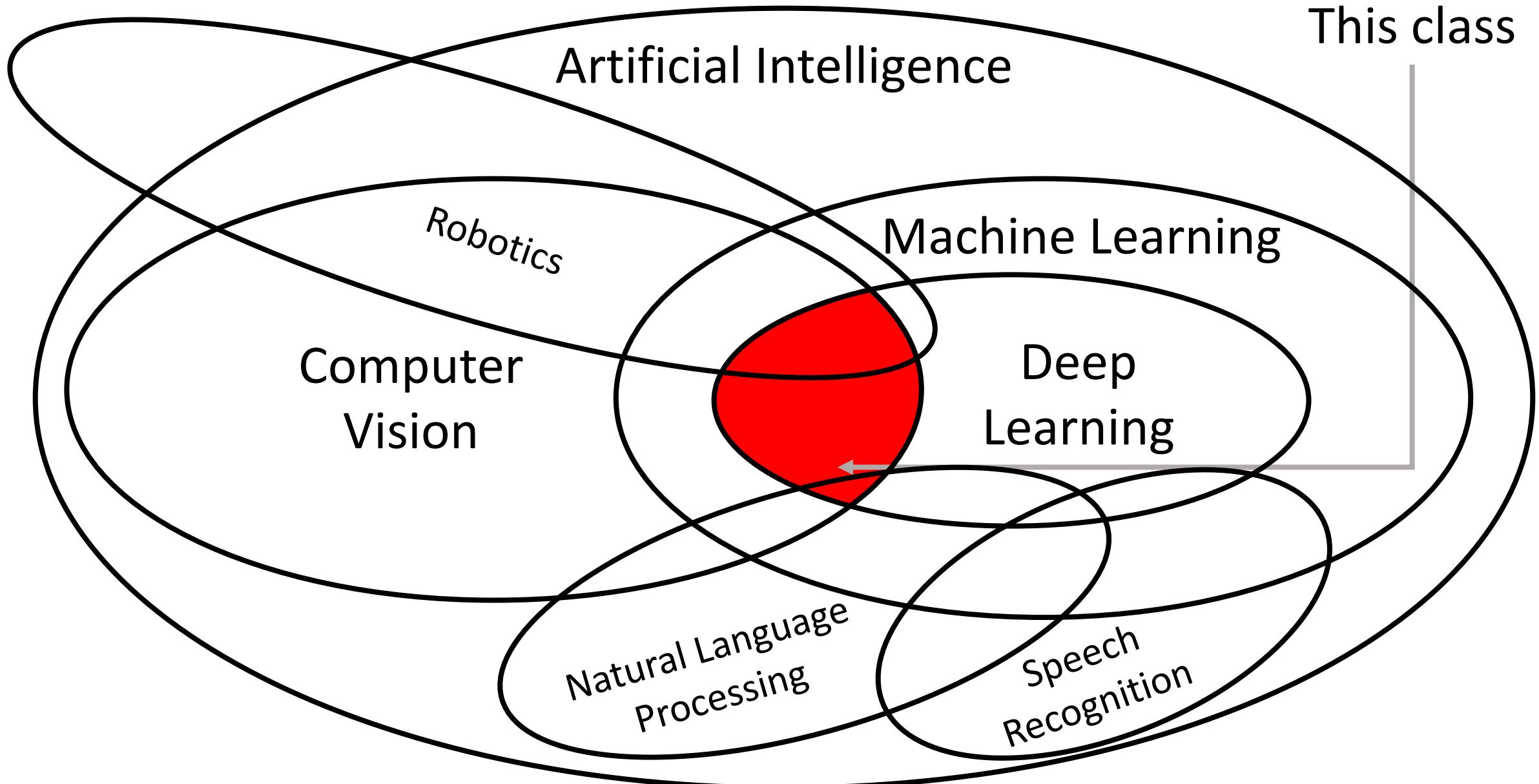
Computer
Vision

Machine Learning

Deep
Learning

Natural Language
Processing

Speech
Recognition



Today's Agenda

- A brief history of computer vision and deep learning
- My personal journey in computer vision
- Course overview and logistics

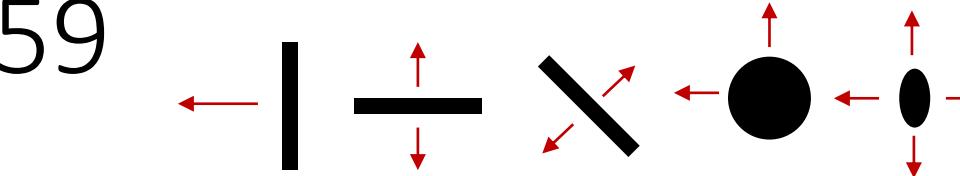
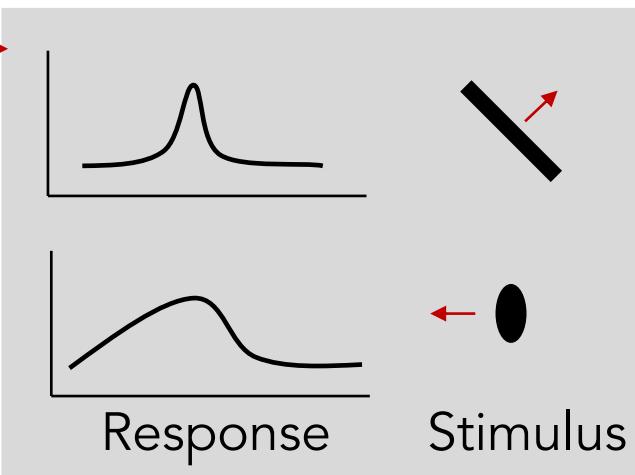
Hubel and Wiesel, 1959

Measure
brain activity

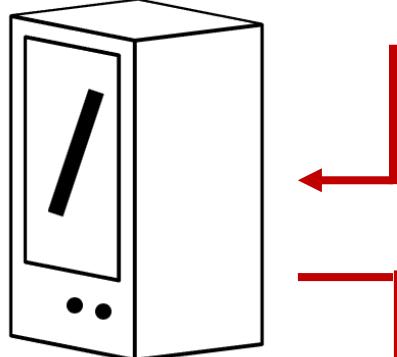


Cat image by CNX OpenStax is licensed under CC BY 4.0; changes made

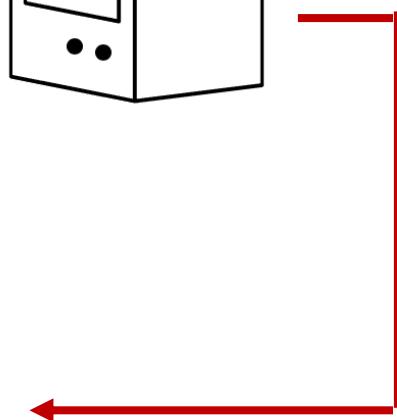
1959
Hubel & Wiesel



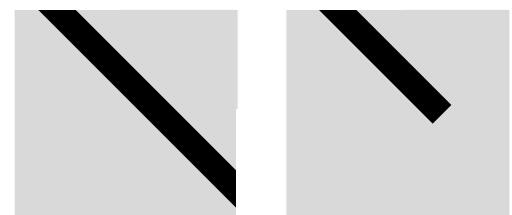
Simple cells:
Response to light orientation



Complex cells:
Response to light orientation and movement



Hypercomplex cells:
response to movement with an end point



No response

<https://www.youtube.com/watch?v=lOHayh06LJ4>



Won Nobel Price

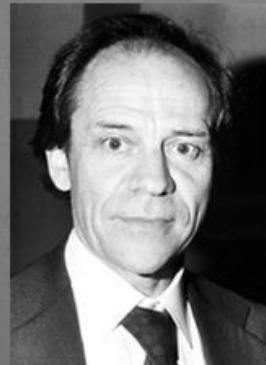
1981 (Shared)

- David H. Hubel (US), Torsten N. Wiesel (Sweden)

**“for their discoveries concerning information processing in
the visual system”**

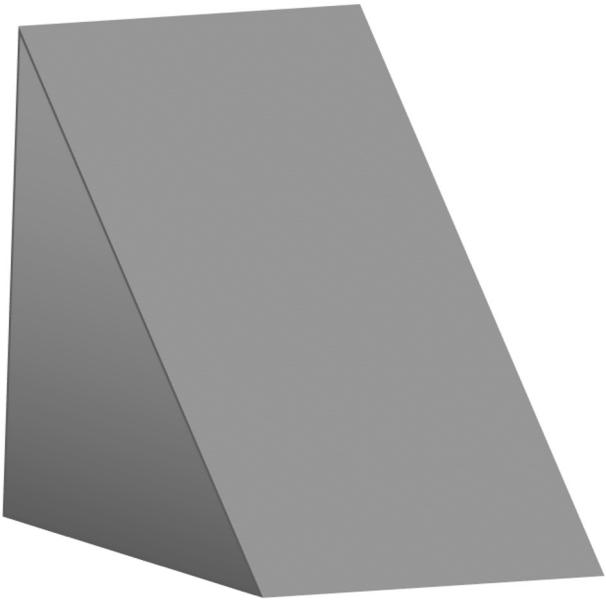


D. H. Hubel

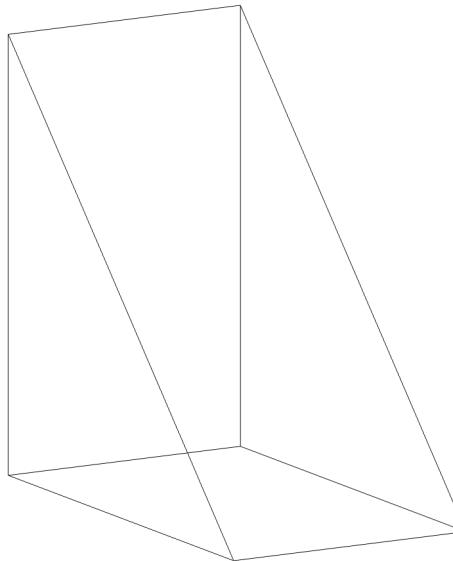


T. N. Wiesel

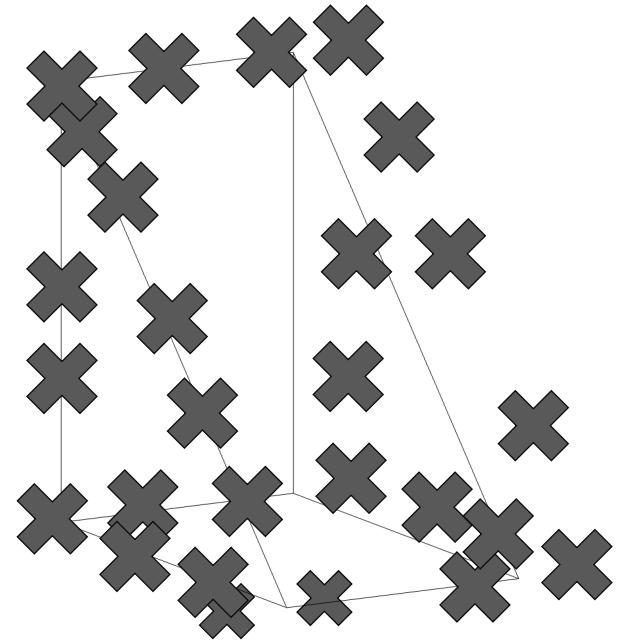
Larry Roberts, 1963



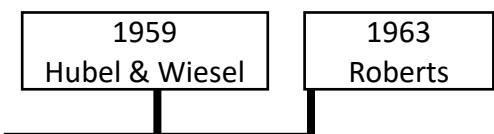
(a) Original picture



(b) Differentiated picture

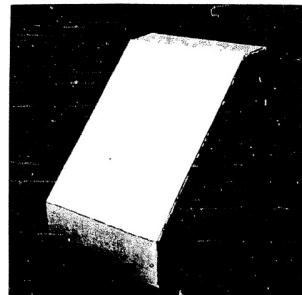
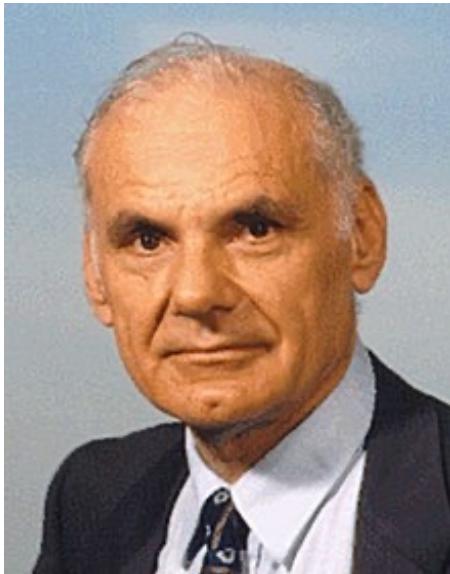


(c) Feature points selected

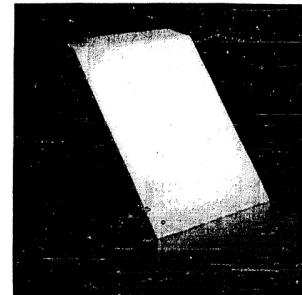


Larry Roberts, 1963

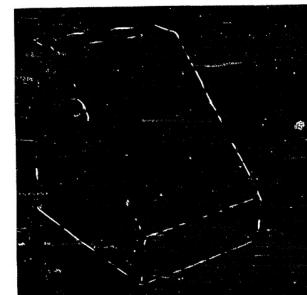
- First PhD thesis on computer vision at MIT: [Machine Perception of Three-Dimensional Solids](#)
- Later went on to be “Founding Father of the Internet” (per Wikipedia)



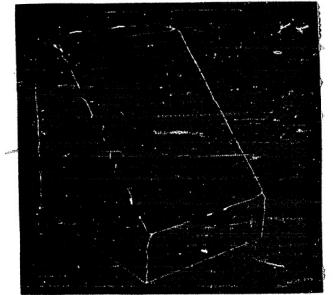
A. Original Picture



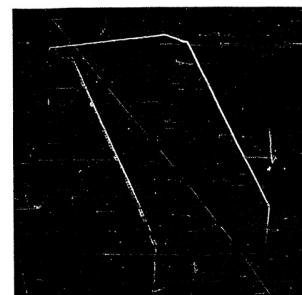
B. Computer Display of Picture
(Reflected by mistake)



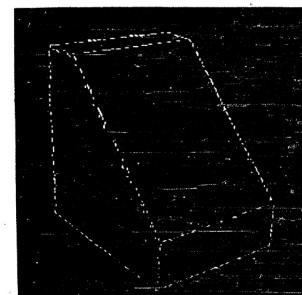
E. Connected Feature Points



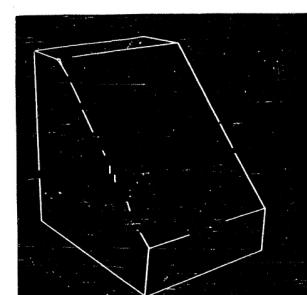
F. After Complexity Reduction



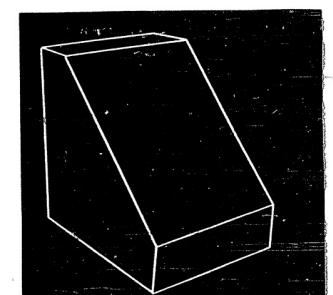
C. Differentiated Picture



D. Feature Points Selected

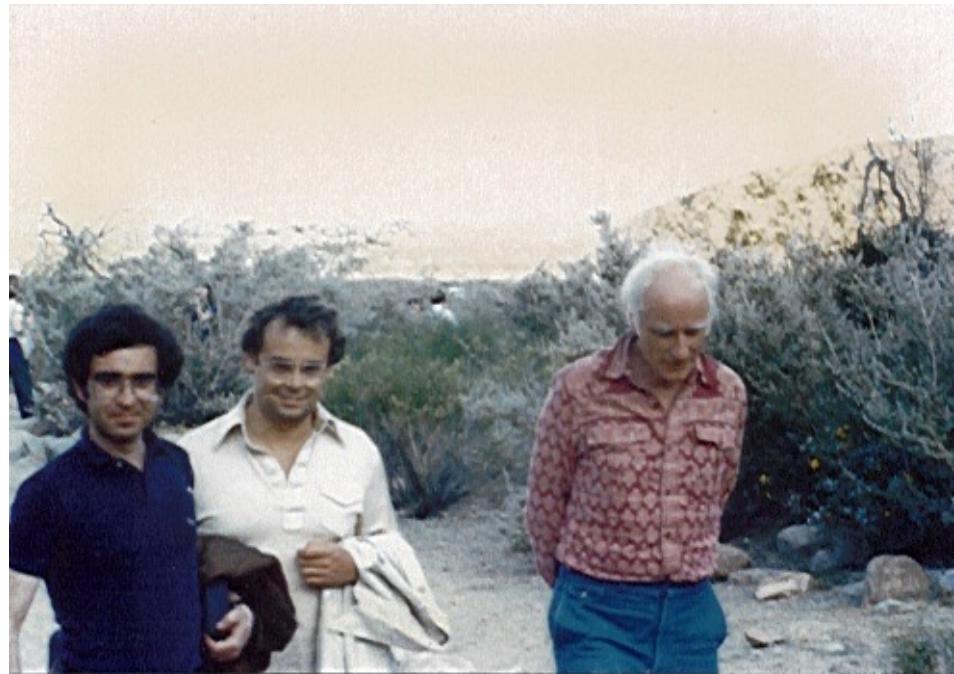
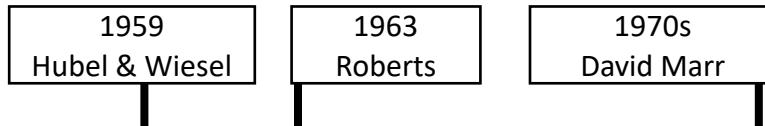


G. After Initial Line Fitting H. Final Line Drawing

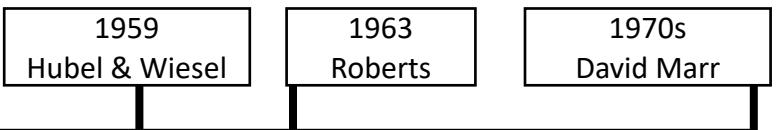
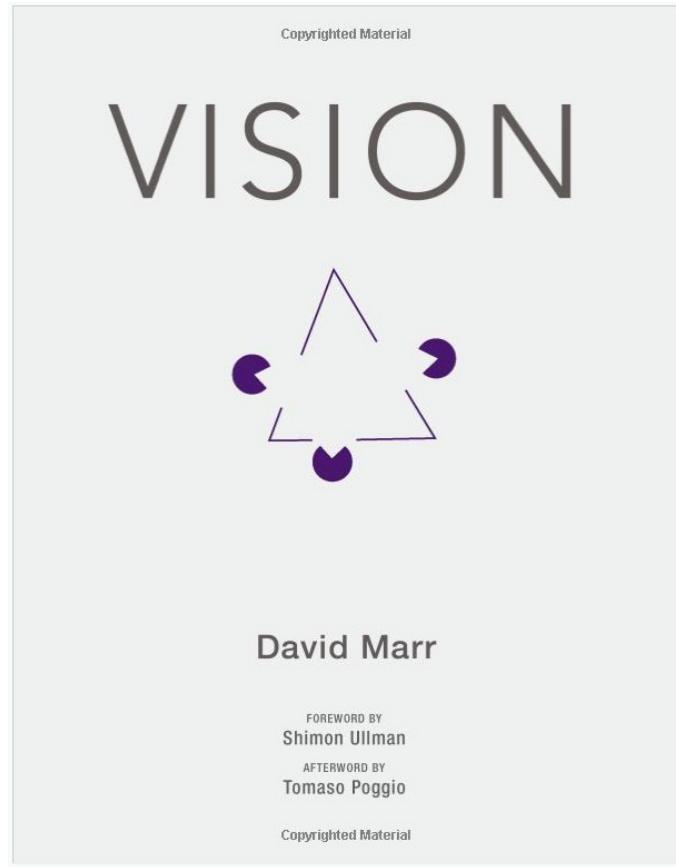


David Marr, 1970s

- British neuroscientist, physiologist
- Tragically died young in 1980 at the age of 35
- Best paper award at ICCV is named after him, Marr Prize

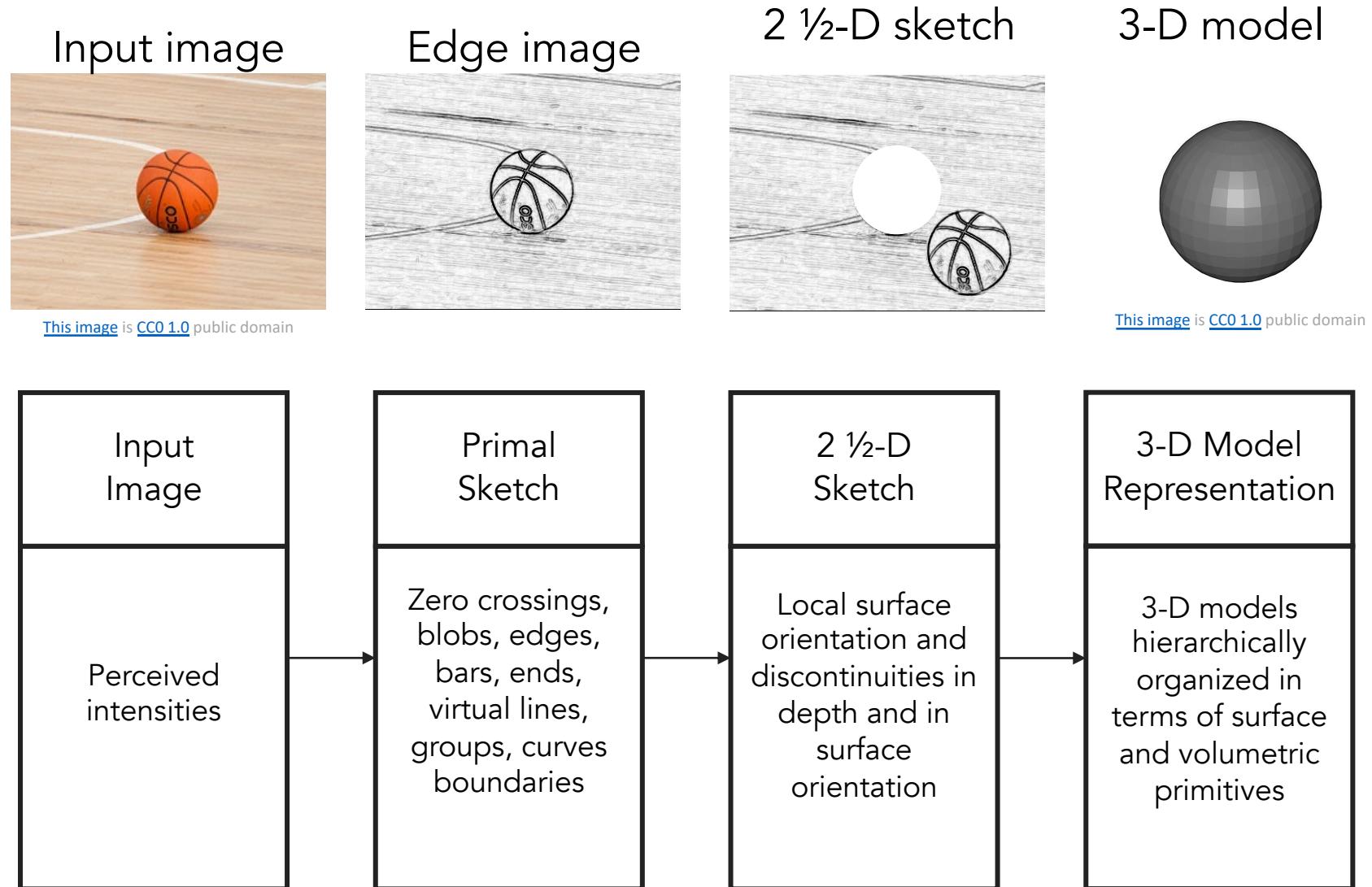


Tomaso Poggio, David Marr and Francis Crick, Borrego Desert, circa 1979



A Computational Investigation into the Human Representation and Processing of Visual Information

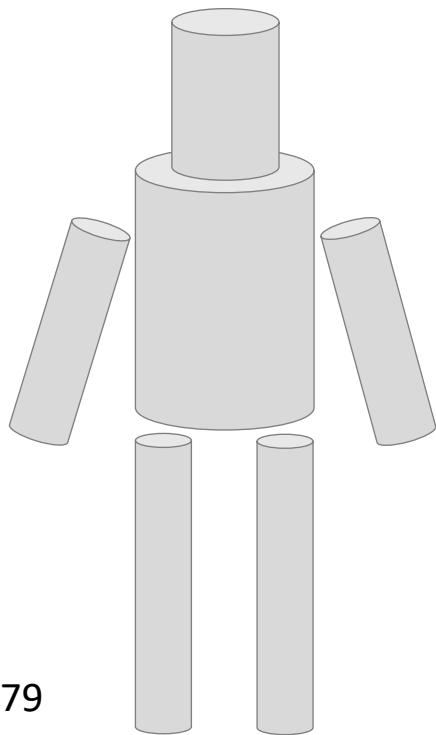
<https://mitpress.mit.edu/books/vision>



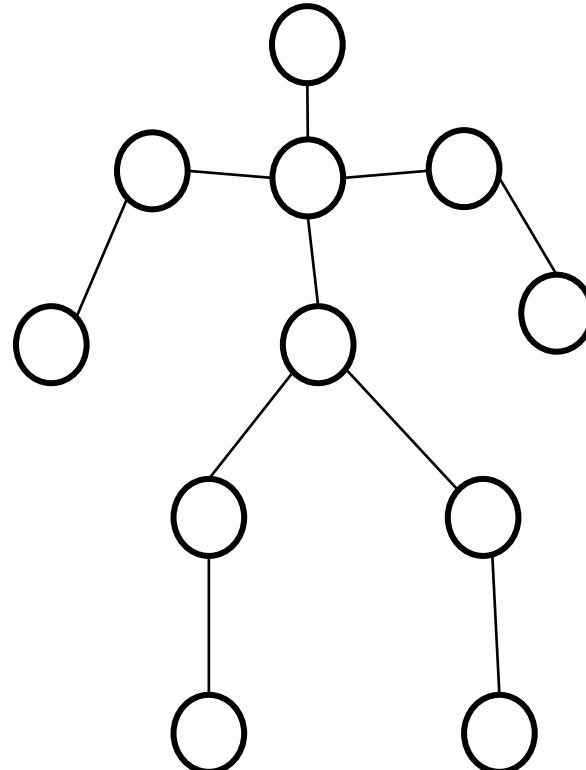
Stages of Visual Representation, David Marr, 1970s

Lecture 1 - 23

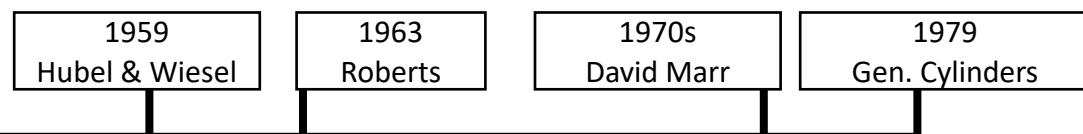
Recognition via Parts (1970s)



Generalized Cylinders,
Brooks and Binford, 1979



Pictorial Structures,
Fischler and Elshlager, 1973



Recognition via Edge Detection (1980s)



1959
Hubel & Wiesel

1963
Roberts

1970s
David Marr

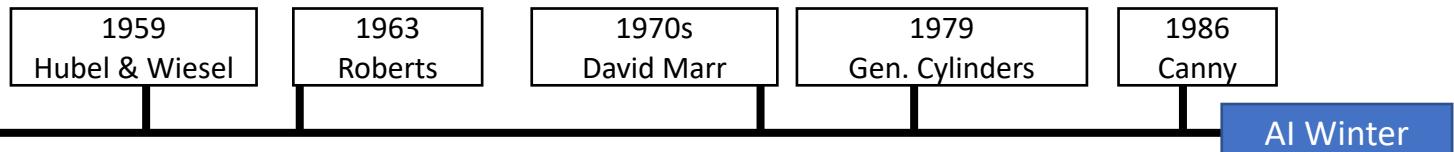
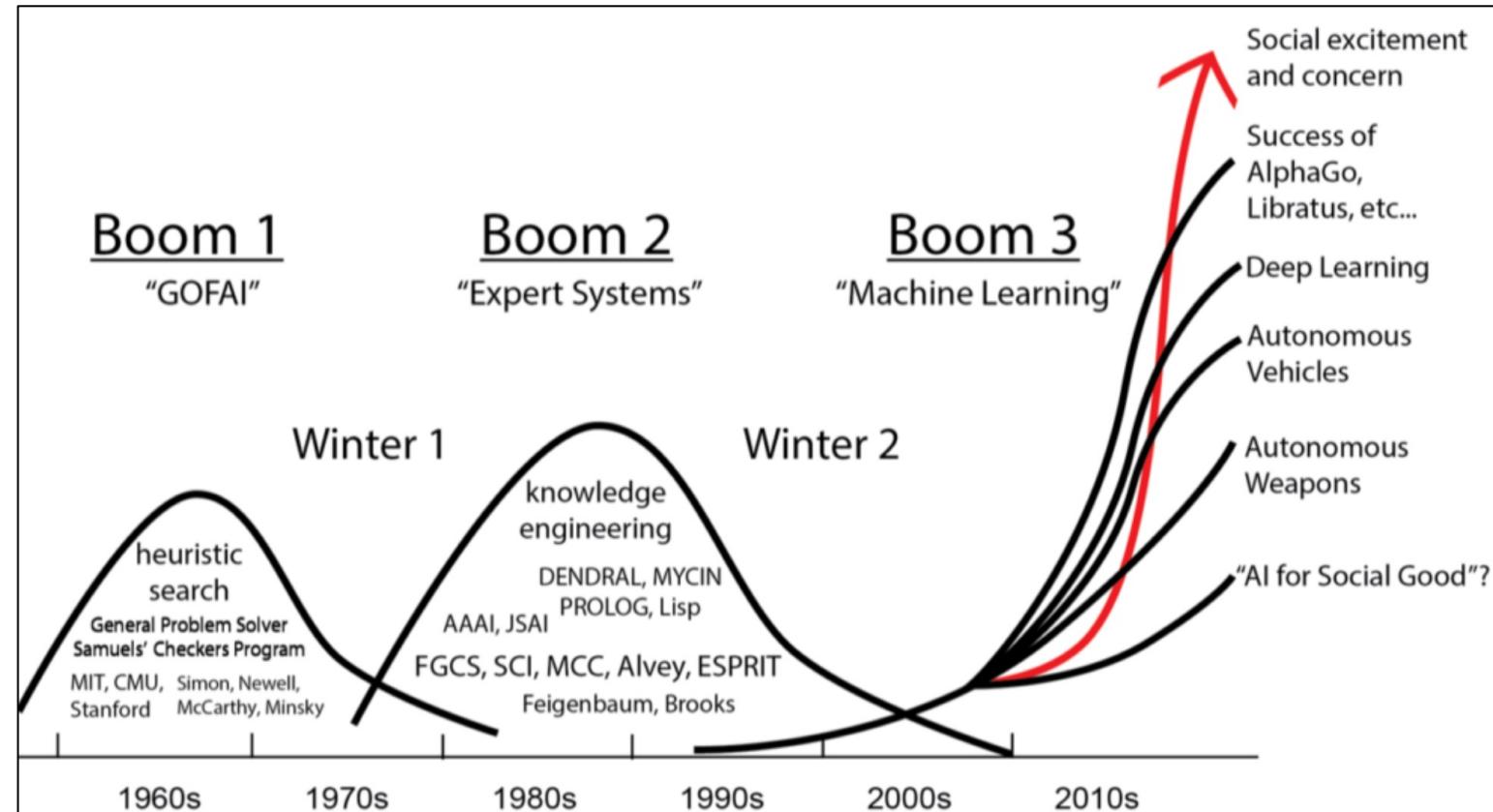
1979
Gen. Cylinders

1986
Canny

John Canny, 1986
David Lowe, 1987

AI Winter in late 1970s and early 1990s

As a result of hype due to over-inflated promises by developers, unnaturally high expectations from end-users, and extensive promotion in the media.



Recognition via Grouping (1990s)



1959
Hubel & Wiesel

1963
Roberts

1970s
David Marr

1979
Gen. Cylinders

1986
Canny

1997
Norm. Cuts

AI Winter

Normalized Cuts, Shi and Malik, 1997

Recognition via Matching (2000s)



[Image](#) is public domain



[Image](#) is public domain

1959
Hubel & Wiesel

1963
Roberts

1970s
David Marr

1979
Gen. Cylinders

1986
Canny

1997
Norm. Cuts

1999
SIFT

AI Winter

SIFT, David
Lowe, 1999

Face Detection

Viola and Jones, 2001

One of the first successful applications of machine learning to vision



1959
Hubel & Wiesel

1963
Roberts

1970s
David Marr

1979
Gen. Cylinders

1986
Canny

1997
Norm. Cuts

1999
SIFT

2001
V&J

AI Winter

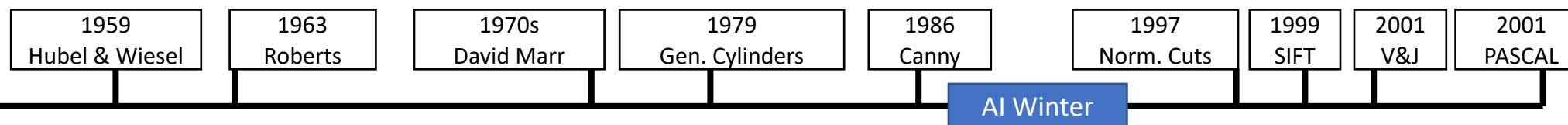
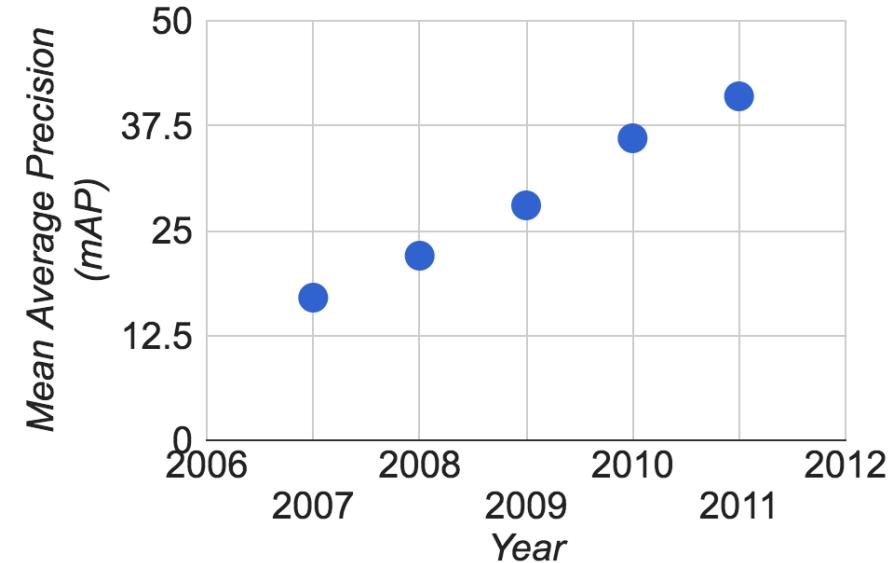
PASCAL Visual Object Challenge

[Image](#) is CC0 1.0 public domain



[Image](#) is CC0 1.0 public domain

Pascal VOC 2007



IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:
1,000 object classes
1,431,167 images



Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

Deng et al, 2009
Russakovsky et al. IJCV 2015

1959
Hubel & Wiesel

1963
Roberts

1970s
David Marr

1979
Gen. Cylinders

1986
Canny

1997
Norm. Cuts

1999
SIFT

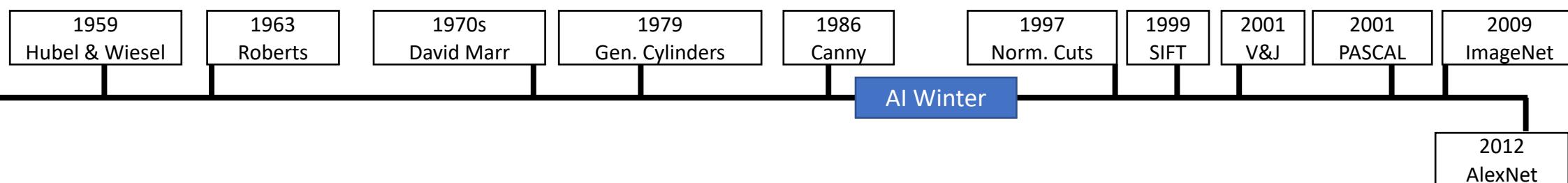
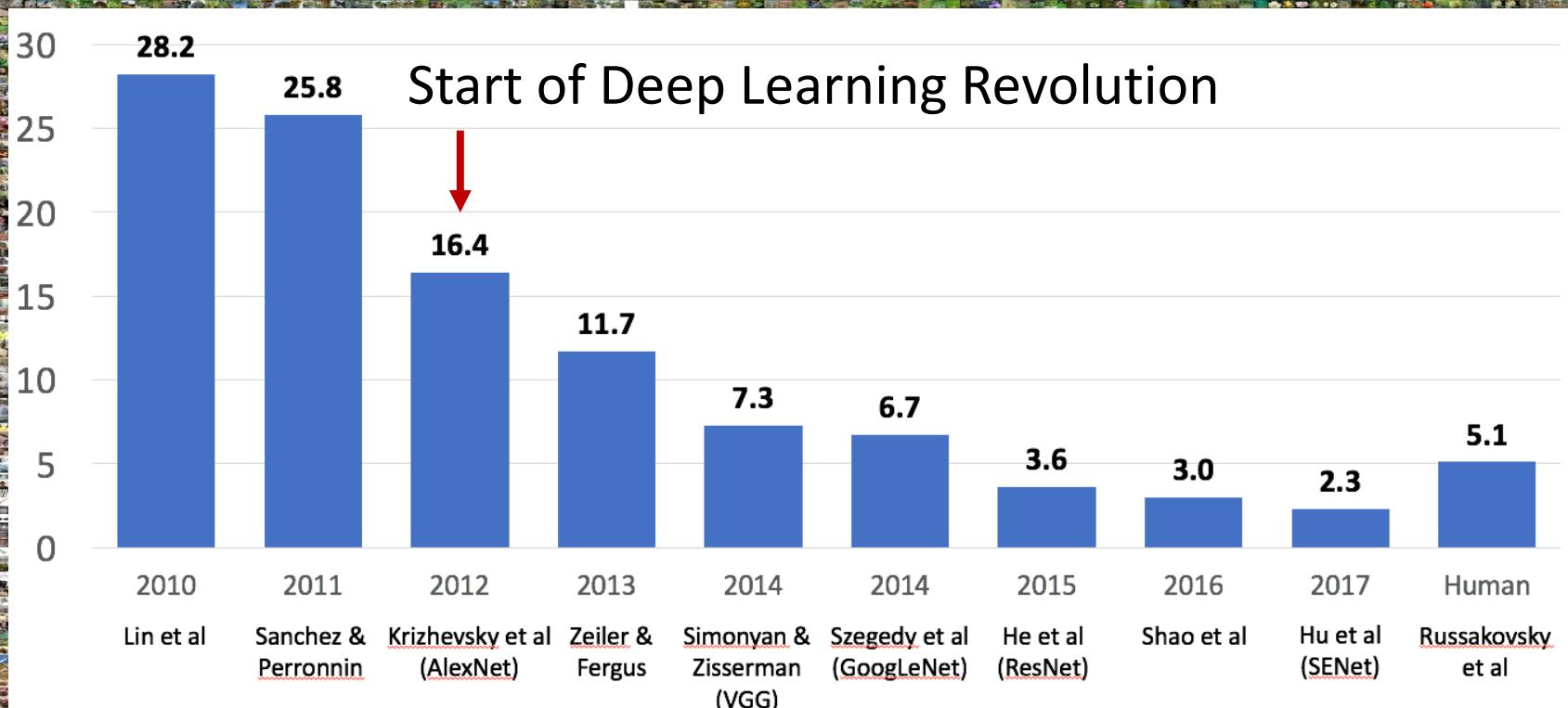
2001
V&J

2001
PASCAL

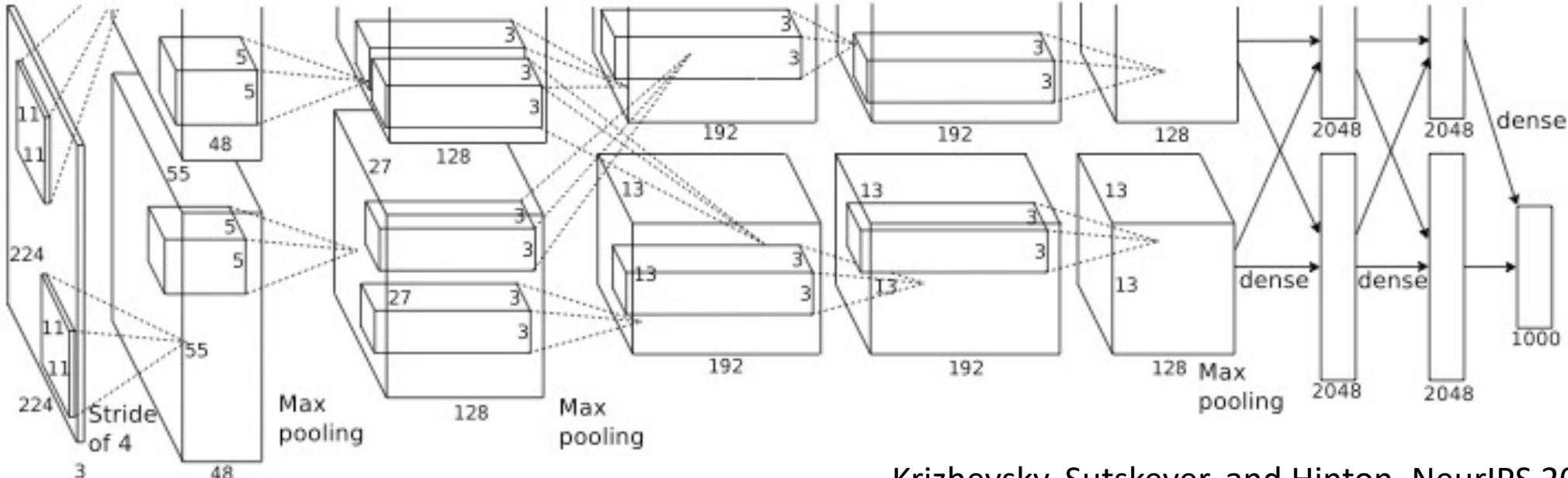
2009
ImageNet

AI Winter

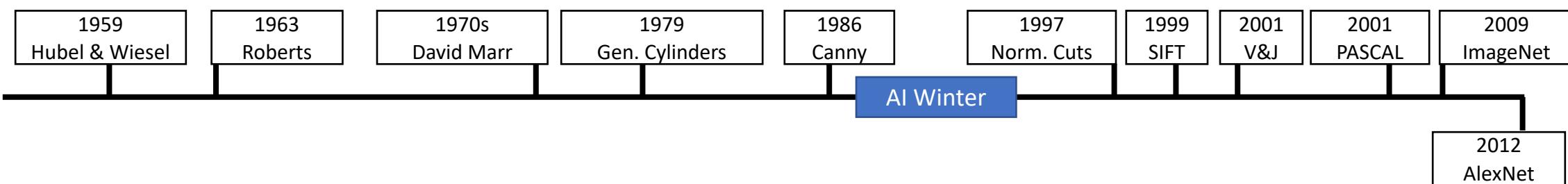
IMAGENET Large Scale Visual Recognition Challenge



AlexNet: Deep Learning Approach for CV



Krizhevsky, Sutskever, and Hinton, NeurIPS 2012



Perceptron

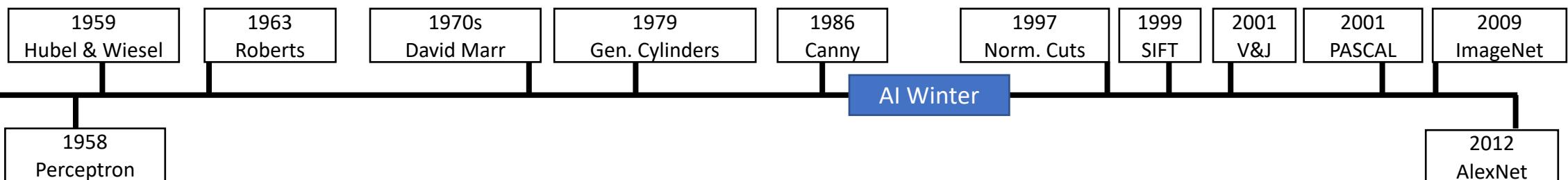
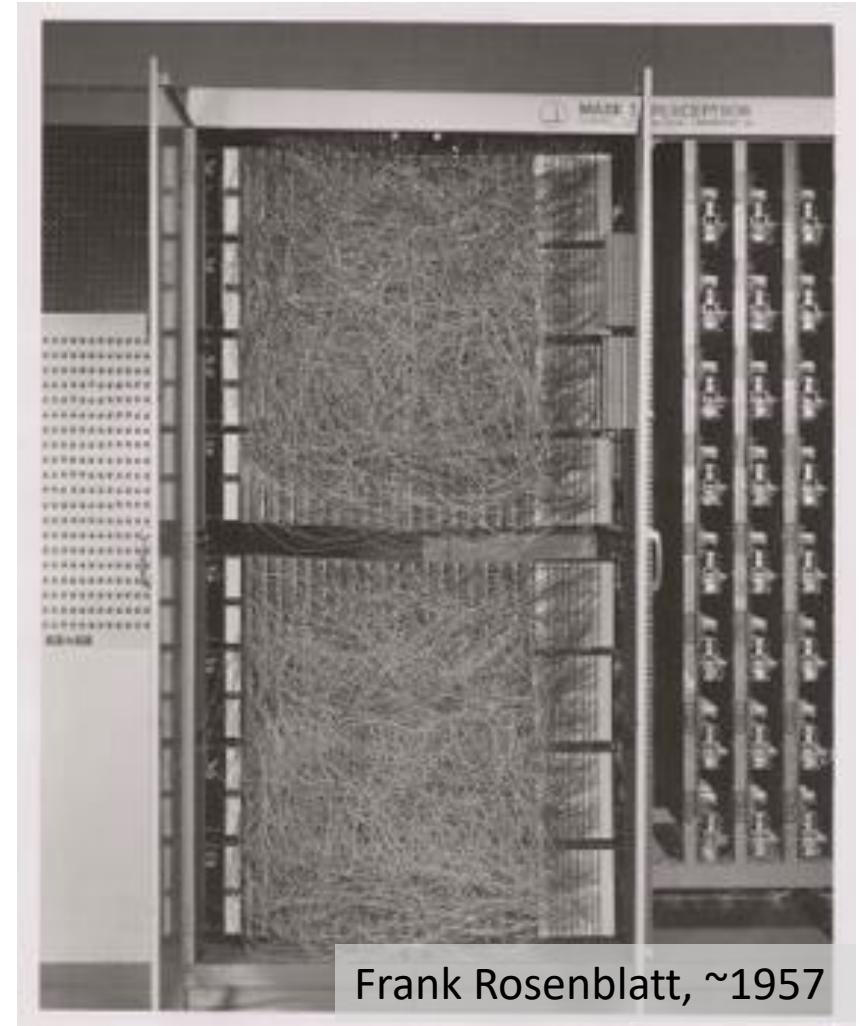
One of the earliest algorithms that could learn from data

Implemented in hardware! Weights stored in potentiometers,
updated with electric motors during learning

Connected to a camera that used 20x20 cadmium sulfide
photocells to make a 400-pixel image

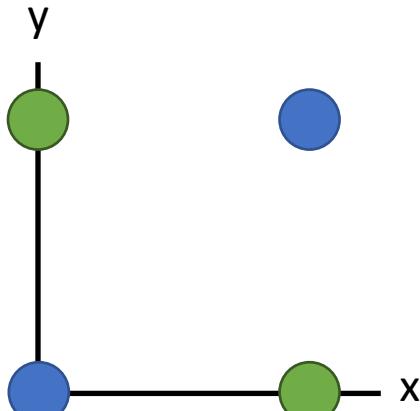
Could learn to recognize letters of the alphabet

Today we would recognize it as a **linear classifier**

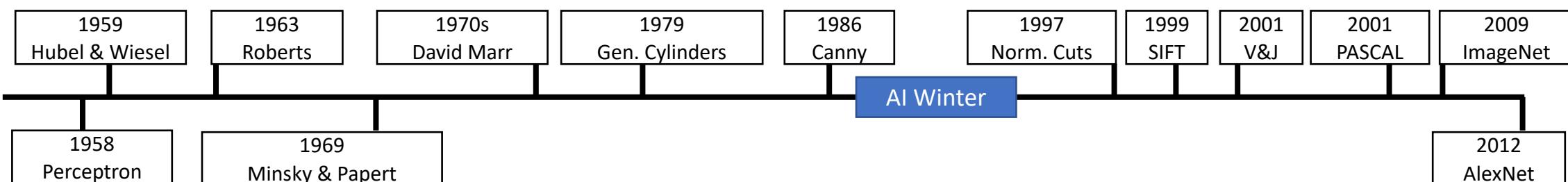
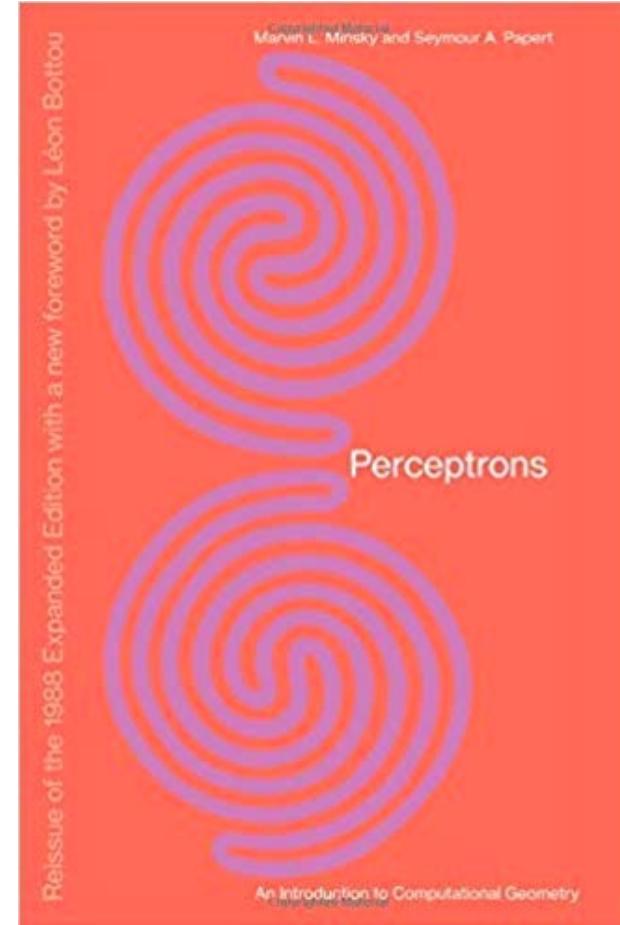


Minsky and Papert, 1969

X	Y	F(x,y)
0	0	0
0	1	1
1	0	1
1	1	0



Showed that Perceptrons could not learn the XOR function
Caused a lot of disillusionment in the field

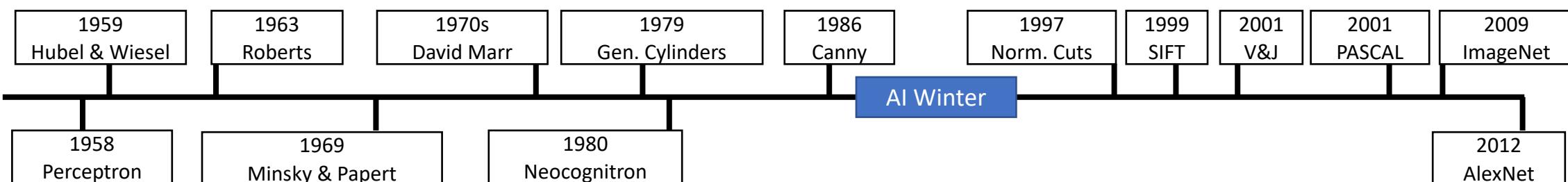
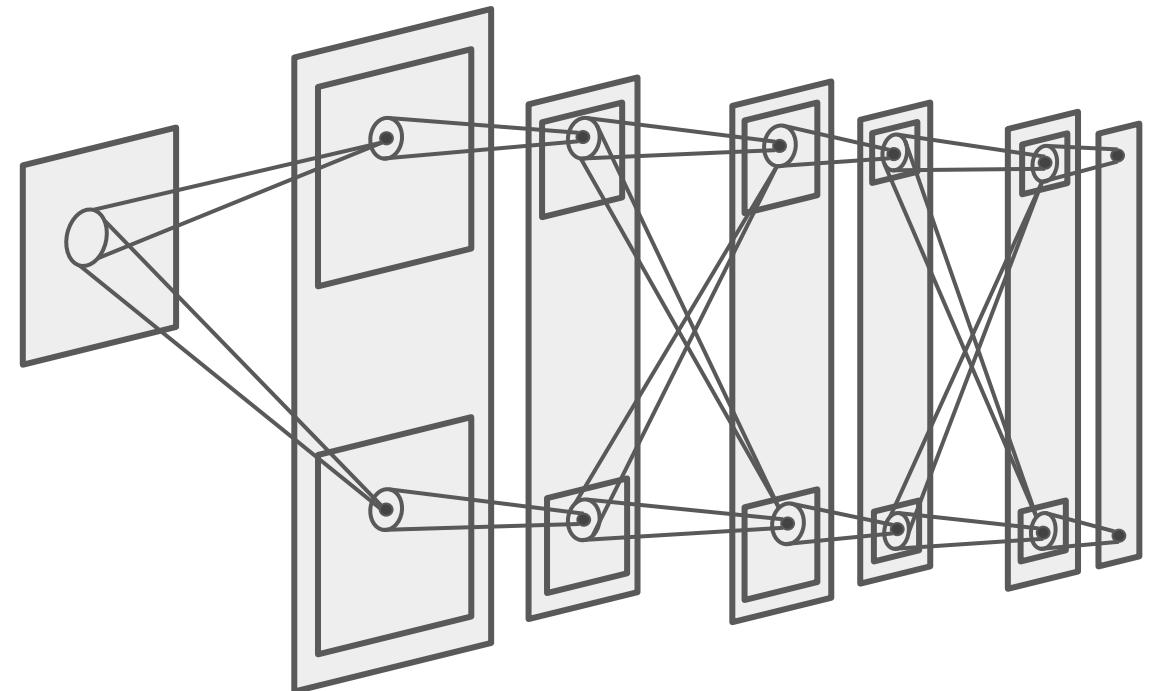


Neocognitron: Fukushima, 1980

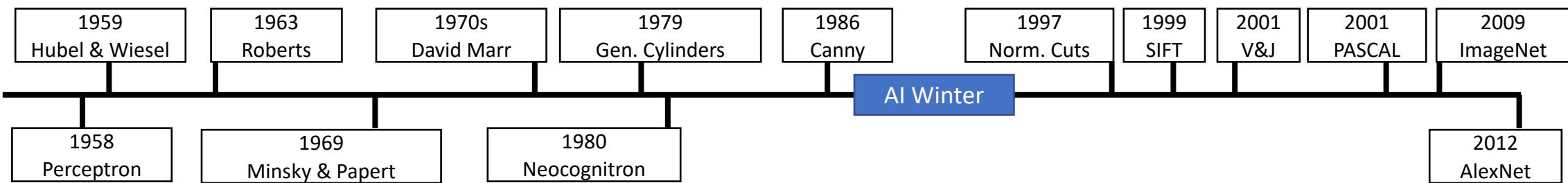
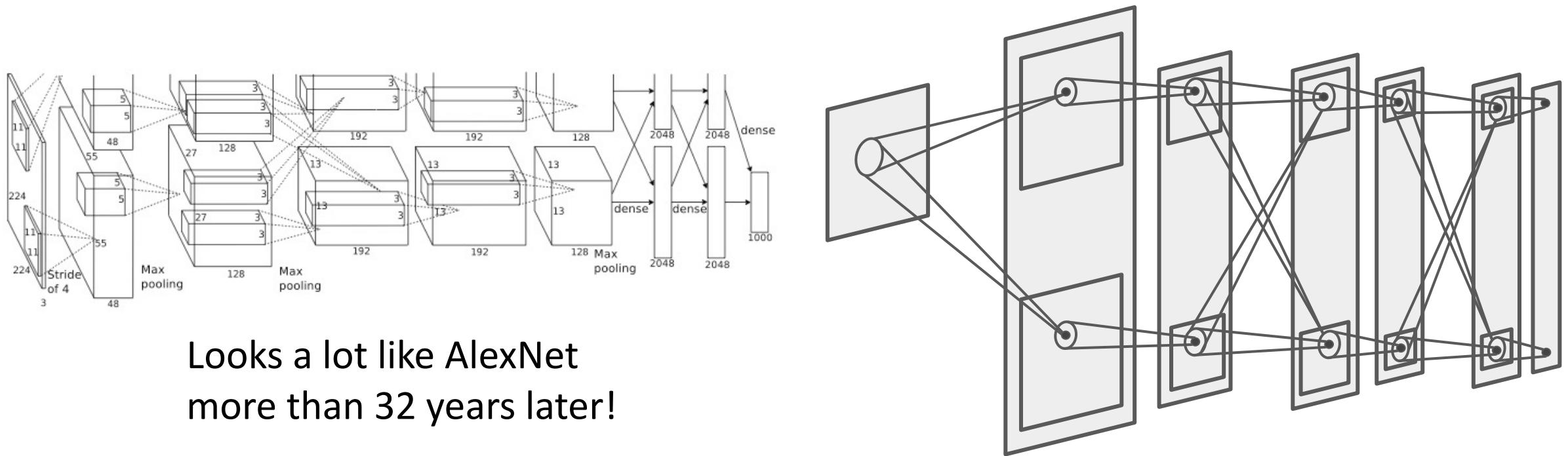
Computational model the visual system,
directly inspired by Hubel and Wiesel's
hierarchy of complex and simple cells

Interleaved simple cells (convolution)
and complex cells (pooling)

No practical training algorithm



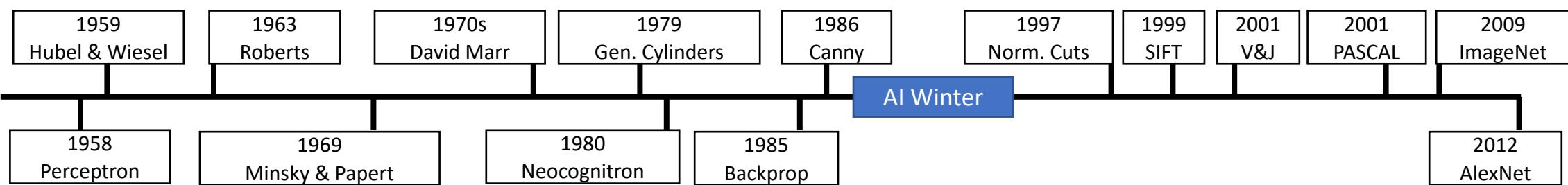
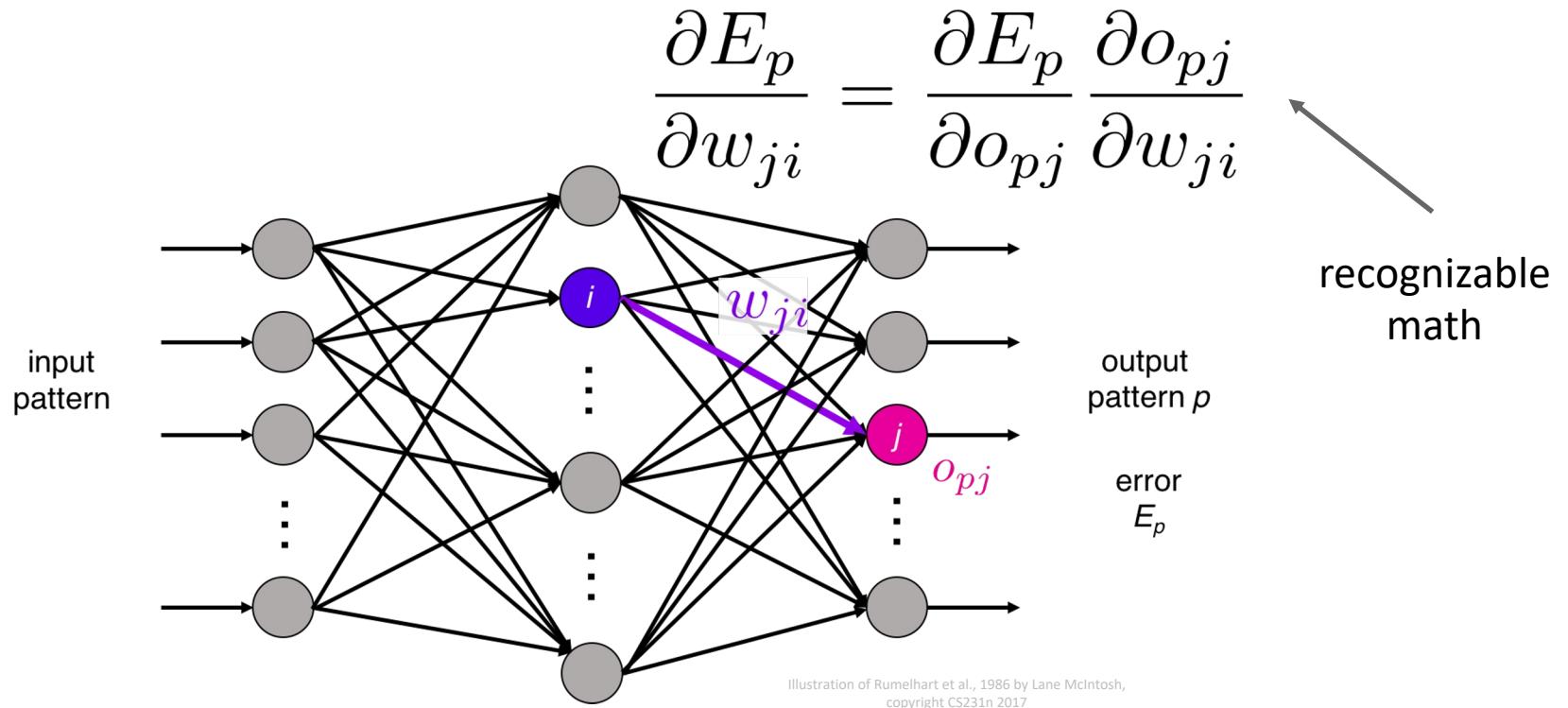
Neocognitron: Fukushima, 1980



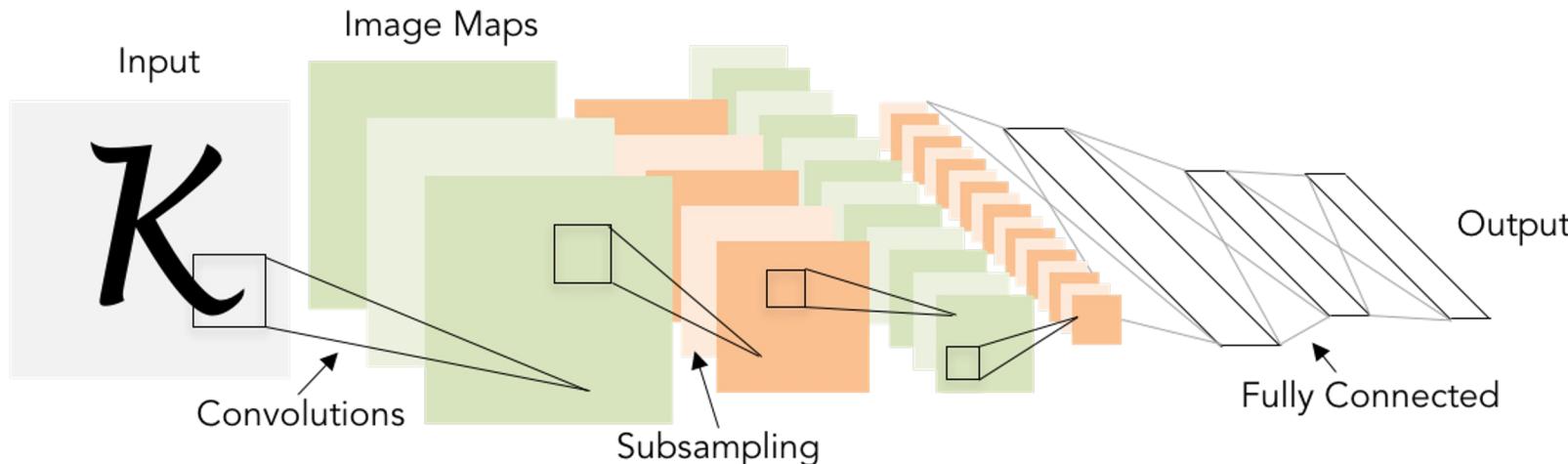
Backprop: Rumelhart, Hinton, and Williams, 1986

Introduced backpropagation
for computing gradients in
neural networks

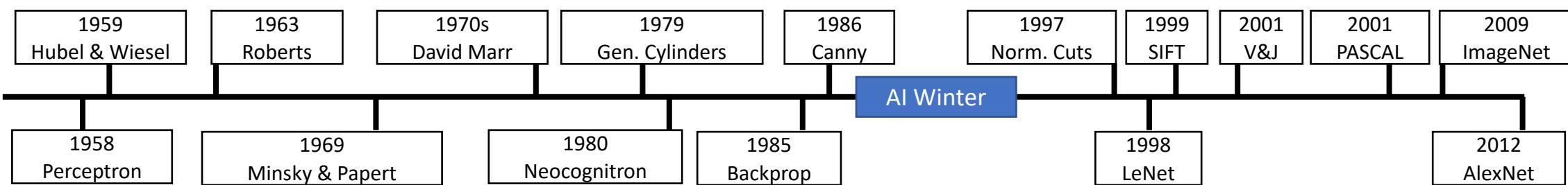
Successfully trained
perceptrons with multiple
layers

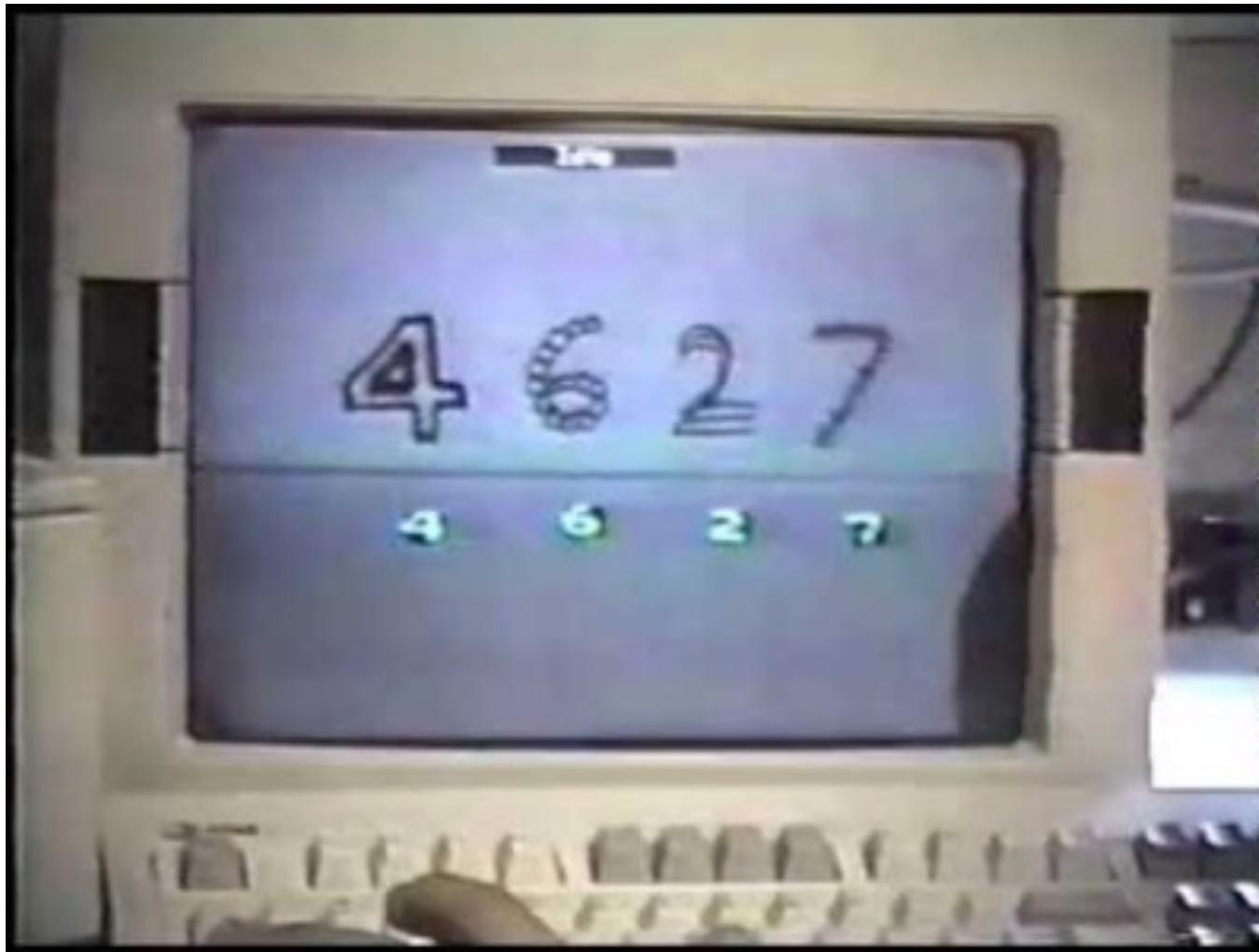


Convolutional Networks: LeCun et al, 1998



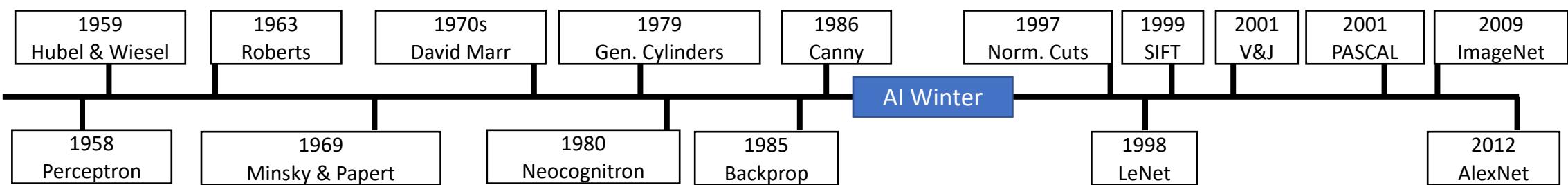
Applied backprop algorithm to a Neocognitron-like architecture
Learned to recognize handwritten digits
Was deployed in a commercial system by NEC, processed handwritten checks
Very similar to our modern convolutional networks!





ConvNet Demo from 1993

https://www.youtube.com/watch?v=FwFduRA_L6Q

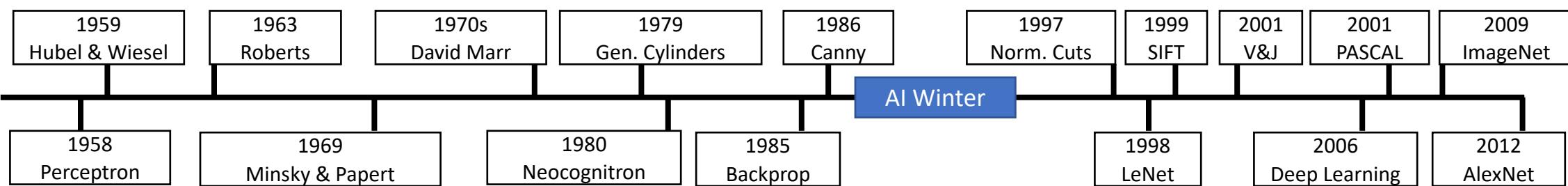
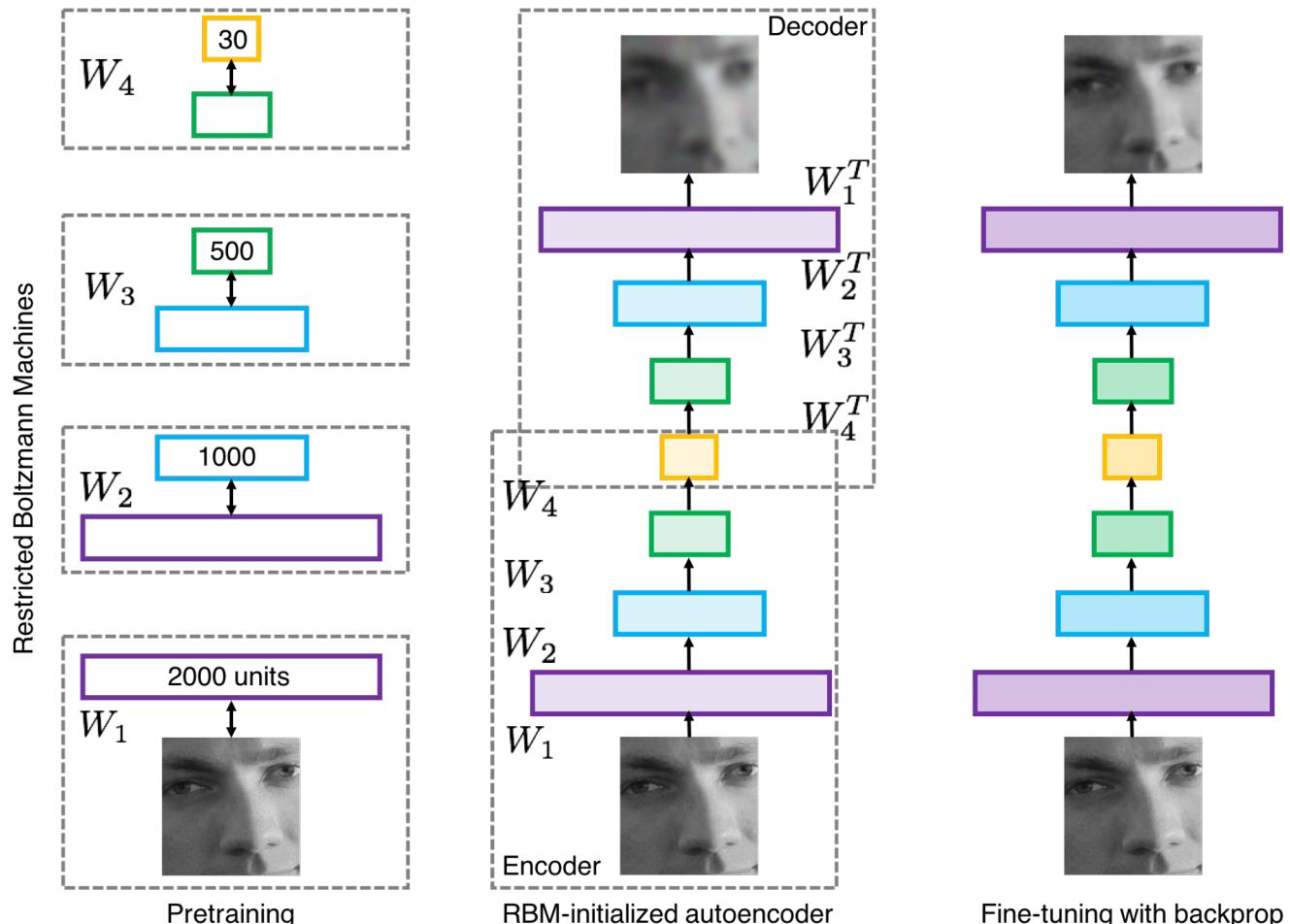


2000s: “Deep Learning”

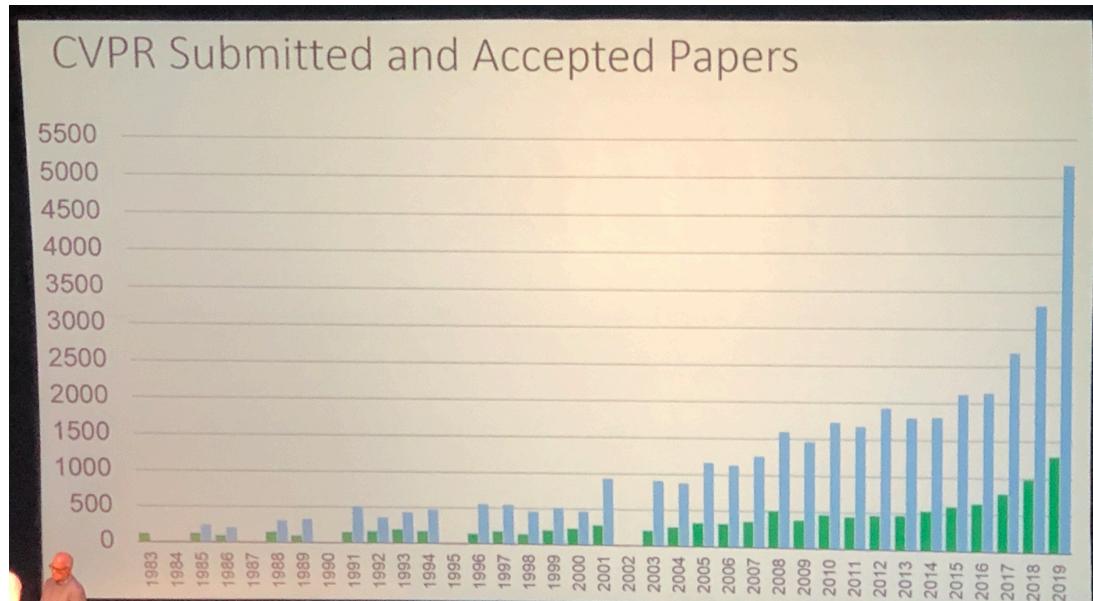
People tried to train neural networks that were deeper and deeper

Not a mainstream research topic at this time

Hinton and Salakhutdinov, 2006
Bengio et al, 2007
Lee et al, 2009
Glorot and Bengio, 2010



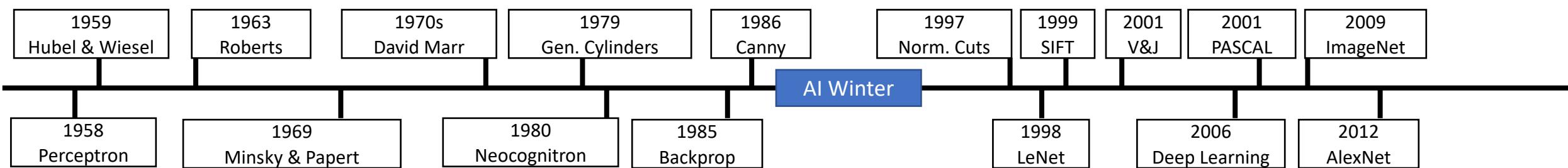
2012 to Present: Deep Learning Explosion



No. CVPR'20 Submissions: ~7,500 (+50% increase)
No. CVPR'21 Submissions: 8,161 (9% increase)
No. CVPR'24 submissions: ~11,000

I submitted my first paper to CVPR'11
~1600 submissions

Publications at top Computer Vision conference CVPR



2012 to Present: Neural Nets are everywhere

Image Classification

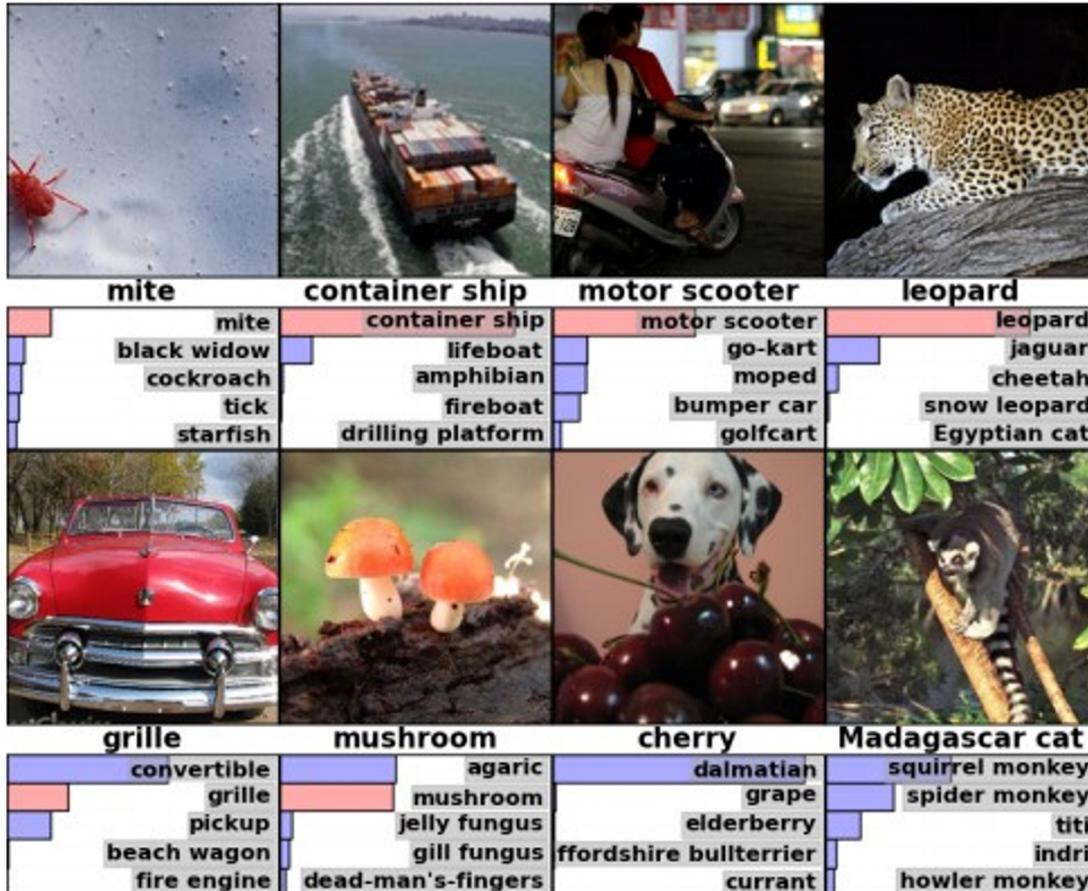
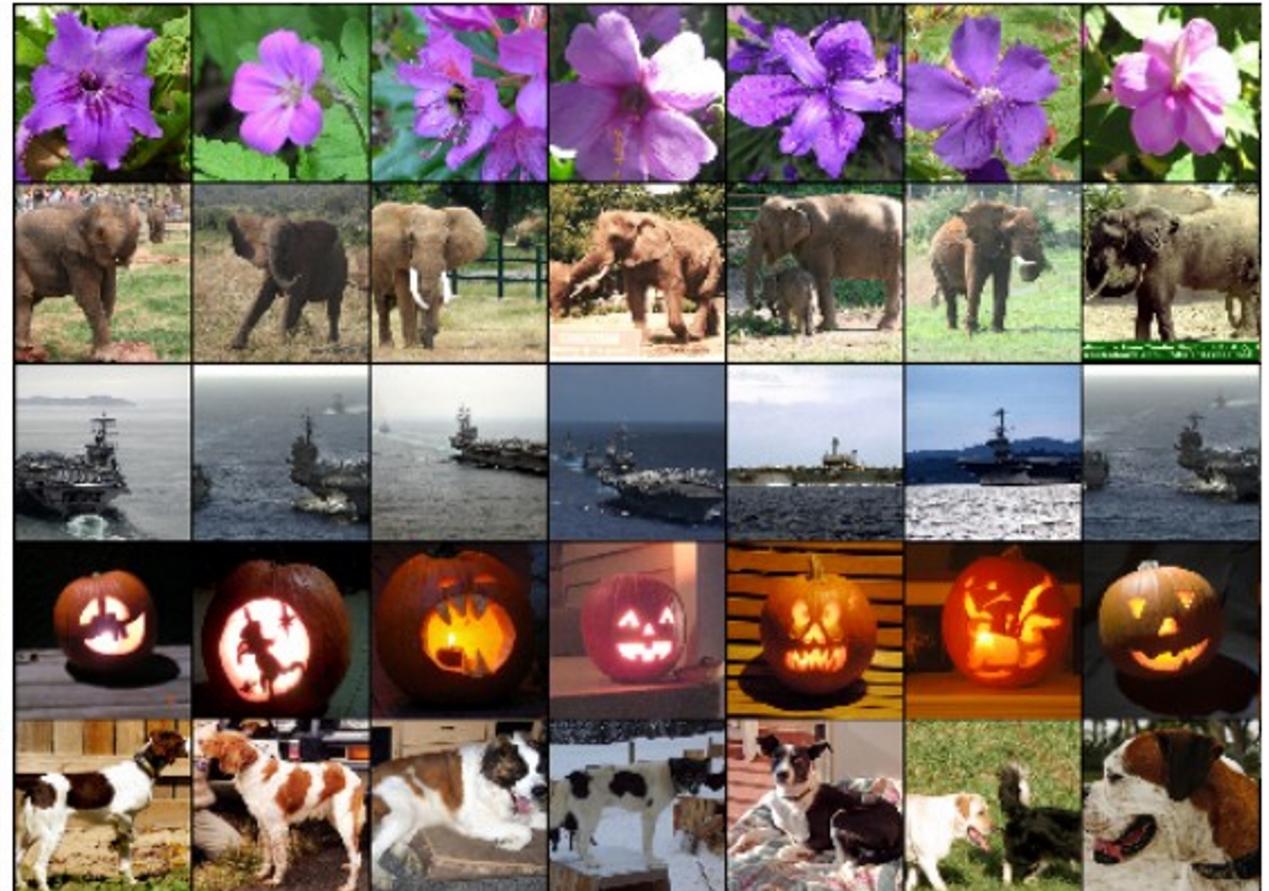


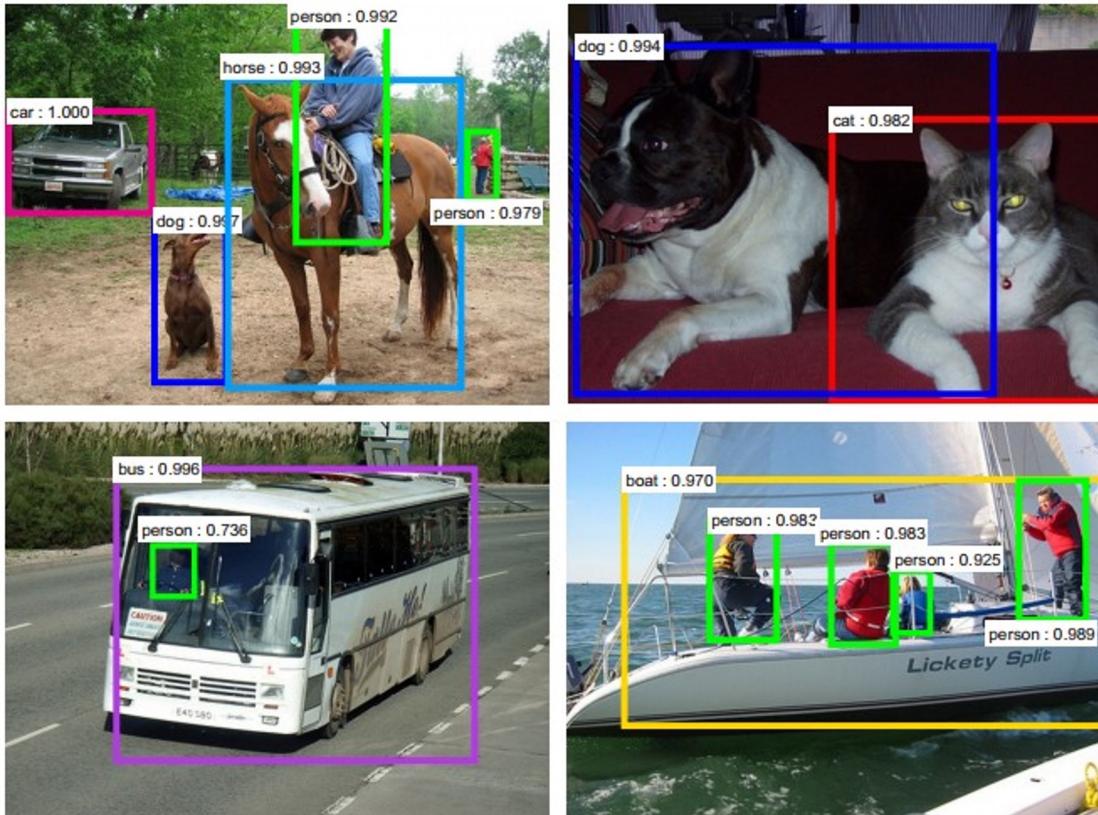
Image Retrieval



Figures copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

2012 to Present: Neural Nets are everywhere

Object Detection



Ren, He, Girshick, and Sun, 2015

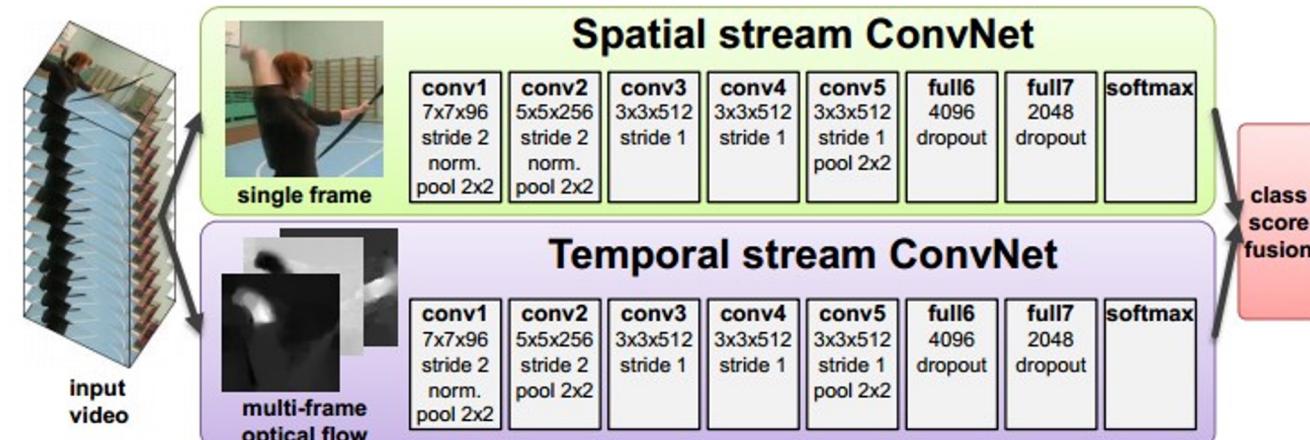
Image Segmentation



Fabaret et al, 2012

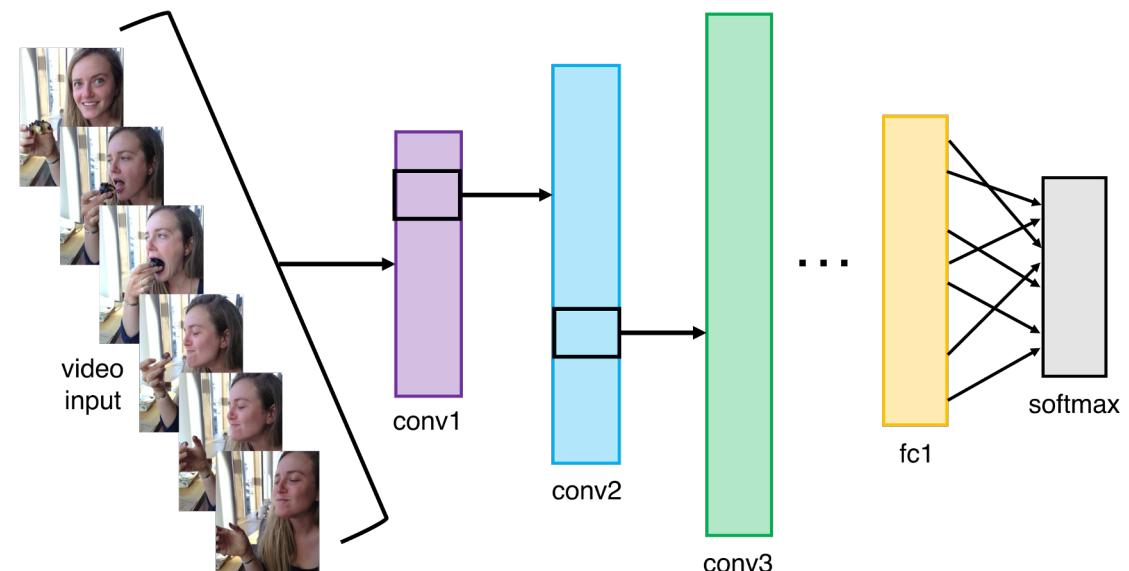
2012 to Present: Neural Nets are everywhere

Video Classification



Simonyan et al, 2014

Activity Recognition



2012 to Present: Neural Nets are everywhere

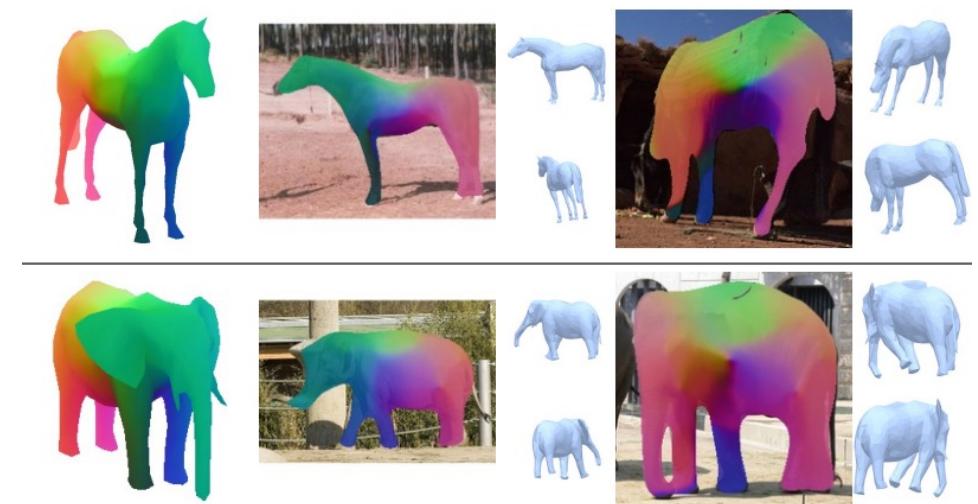
Pose Recognition (Toshev and Szegedy, 2014)



DensePose (Guler et al, 2018)

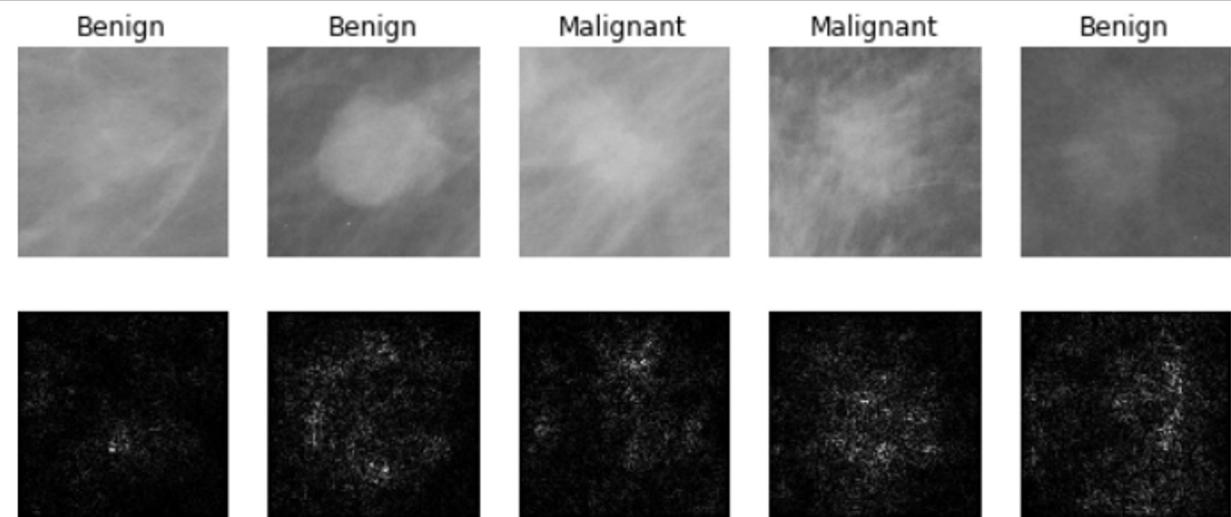


Kulkarni et al, 2020



2012 to Present: Neural Nets are everywhere

Medical Imaging



Levy et al, 2016

Figure reproduced with permission

Whale recognition



[Kaggle Challenge](#)

This image by Christin Khan is in the public domain and originally came from the U.S. NOAA.

Galaxy Classification



Dieleman et al, 2014

From left to right: [public domain by NASA](#), [usage permitted by ESA/Hubble](#), [public domain by NASA](#), and [public domain](#).

2012 to Present: Neural Nets are everywhere



*A white teddy bear
sitting in the grass*



*A man in a baseball
uniform throwing a ball*



*A woman is holding
a cat in her hand*



*A man riding a wave
on top of a surfboard*



*A cat sitting on a
suitcase on the floor*



*A woman standing on a
beach holding a surfboard*

Image Captioning

Vinyals et al, 2015

Karpathy and Fei-Fei, 2015

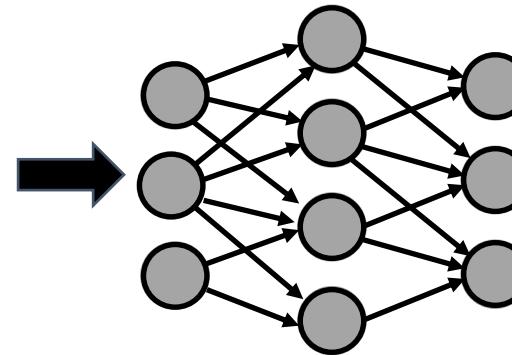
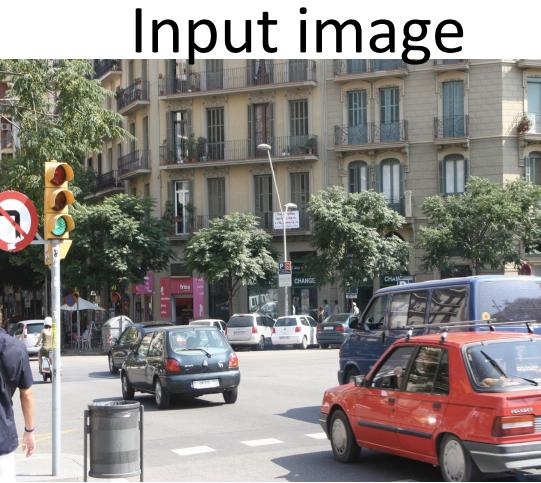
All images are CC0 Public domain:

<https://pixabay.com/en/luggage-antique-cat-1643010/>
<https://pixabay.com/en/teddy-plush-bears-cute-teddy-bear-1623436/>
<https://pixabay.com/en/surf-wave-summer-sport-litoral-1668716/>
<https://pixabay.com/en/woman-female-model-portrait-adult-983967/>
<https://pixabay.com/en/handstand-lake-meditation-496008/>
<https://pixabay.com/en/baseball-player-shortstop-infield-1045263/>

Captions generated by Justin Johnson using [Neuraltalk2](#)

Paradigm Shifting from Classification to Generative Modeling

Classification model



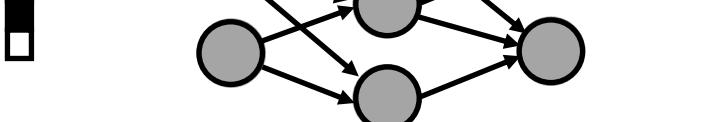
street & pixel labels



tree
car
van
ashcan
person
streetlight
signboard
traffic light

Generative model

Input vector



Generative Modeling in Computer Vision

Image Generation

Karras et al, 2018

Brock et al., 2018



StyleGAN



BigGAN

Generative Modeling in Computer Vision

Style Transfer

Mordvinsev et al, 2015

Gatys et al, 2016



Original image is CCO public domain

Starry Night and Tree Roots by Van Gogh are in the public domain

Bokeh image is in the public domain

Stylized images copyright Justin Johnson, 2017;

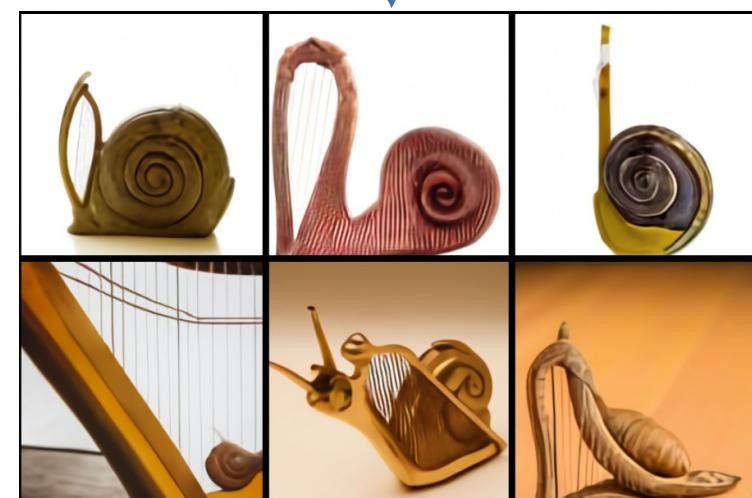
reproduced with permission

Text-guided Image Generation from OpenAI (early 2021)

An armchair in the shape
of an avocado



A snail made of harp



Text-guided Image Generation from OpenAI (late 2021)



“a hedgehog using a calculator”



“a corgi wearing a red bowtie and a purple party hat”



“robots meditating in a vipassana retreat”



“a fall landscape with a small cottage next to a lake”



“a surrealist dream-like oil painting by salvador dalí of a cat playing checkers”



“a professional photo of a sunset behind the grand canyon”



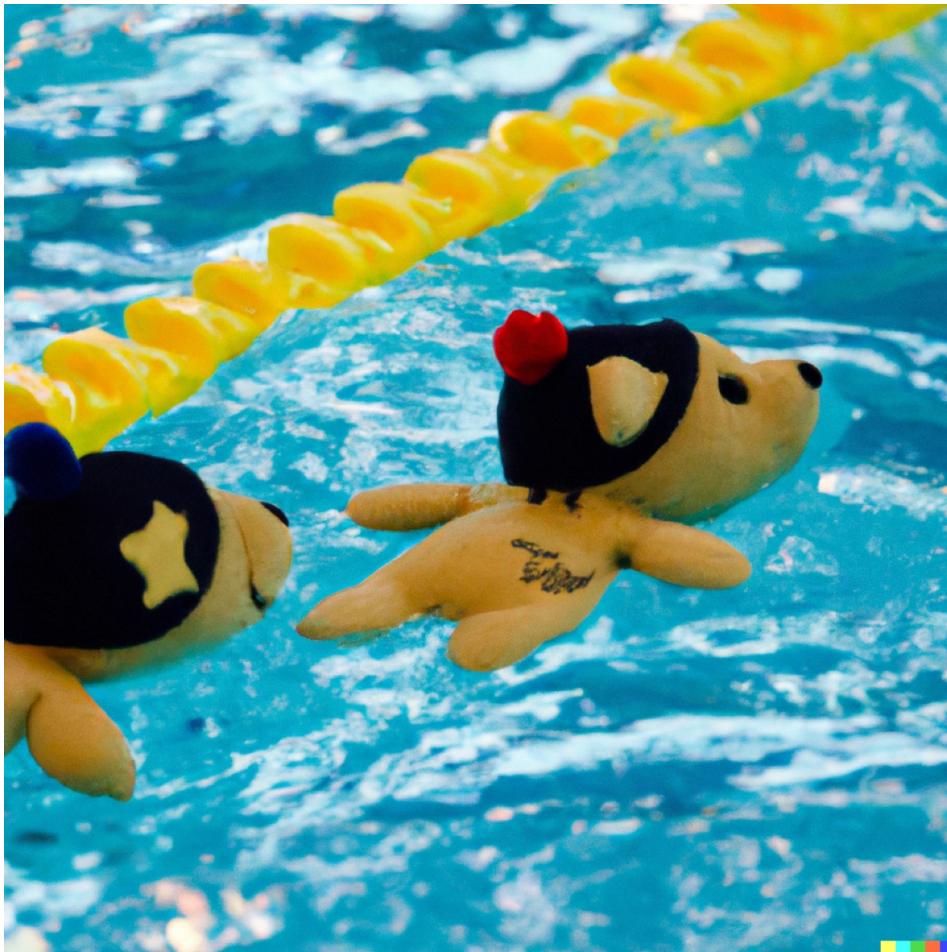
“a high-quality oil painting of a psychedelic hamster dragon”



“an illustration of albert einstein wearing a superhero costume”

2022: A magic year for Text-to-Image Generation

Teddy bears swimming at the Olympics 400m Butterfly event.



OpenAI's DALLE2



Google's Imagen

2022: A magic year for Text-to-Image Generation

A photo of a Corgi dog riding a bike in Times Square.
It is wearing sunglasses and a beach hat.



OpenAI's DALLE2



Google's Imagen

2022: A magic year for Text-to-Image Generation

A dog looking curiously in the mirror, seeing a cat.



OpenAI's DALLE2



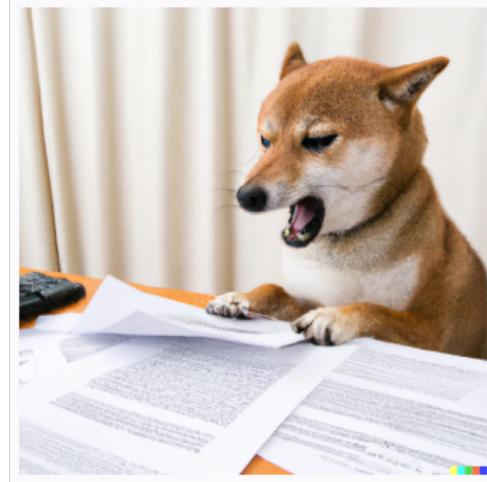
Google's Imagen

Try your own: [ChatGPT](#), Meta AI (inside your IG app)

shiba inu overlooking Grand Canyon in starry night



a shiba inu is doing tax reporting, angrily



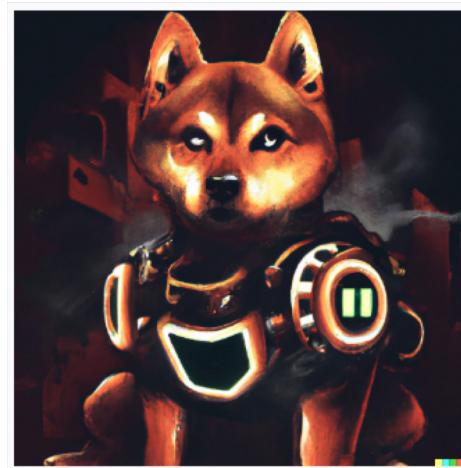
shiba inu is playing heavy metal music



a superhero shiba inu fighting an evil corgi, watercolor



a robot shiba inu, cyberpunk style



A shiba inu is doing rock climbing, pixel art



2023: Explosion of 3D and Video Generation

DreamFusion

<https://dreamfusion3d.github.io/>

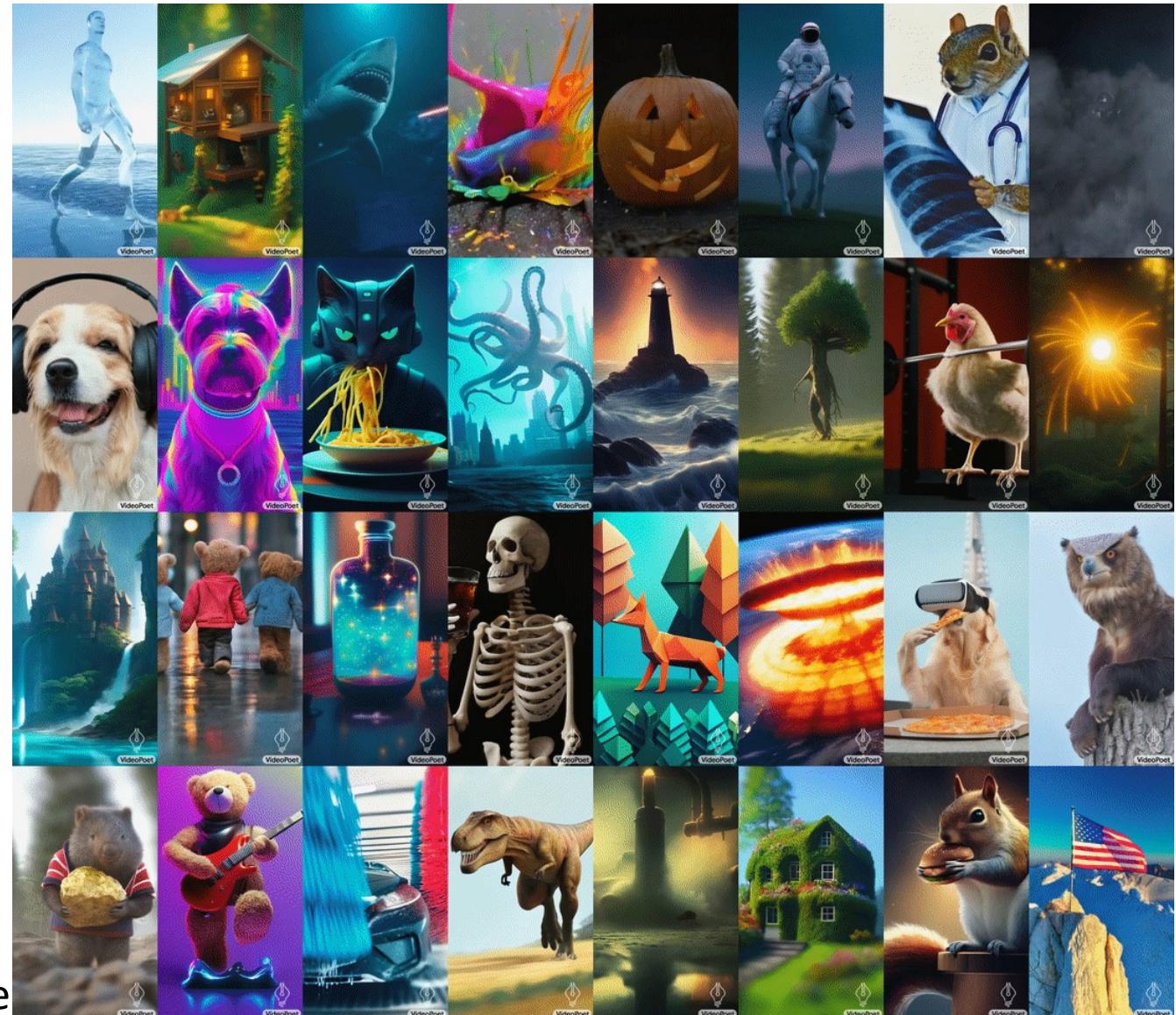


Magic3D:

<https://research.nvidia.com/labs/dir/magic3d>



Google VideoPoet: <https://sites.research.google/videopoet/>



2023-now: Multi-modal Large Model and Vision-Language Models (VLM)

- GPT-4v



describe what's going on



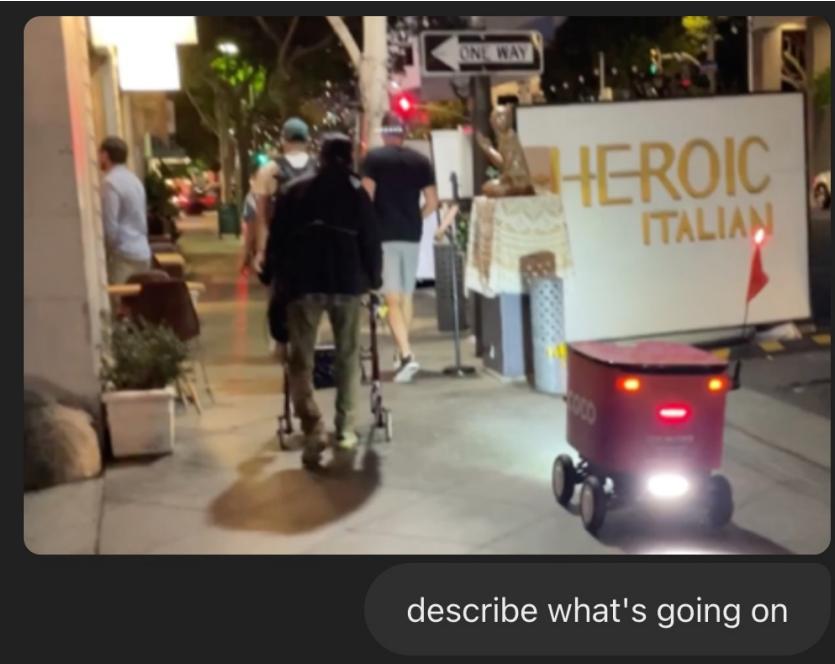
The image shows a street scene, likely in the evening, as indicated by artificial lighting. Several people are walking along a sidewalk. The person in the foreground is using a walker, possibly indicating they have mobility challenges. Ahead of them, two other individuals walk together.

In the lower right corner, there is a small autonomous delivery robot with headlights and a red flag, typically used for transporting food or small packages in urban environments. It has branding on it, likely from a delivery service, and is navigating the sidewalk along with the pedestrians. There is a sign on the right that reads "HEROIC ITALIAN," which could be a nearby restaurant.

The image captures the coexistence of people and autonomous delivery robots in public urban spaces, illustrating the growing role of such technologies in everyday settings.

2023-now: Multi-modal Large Model and Vision-Language Models (VLM)

- GPT-4v

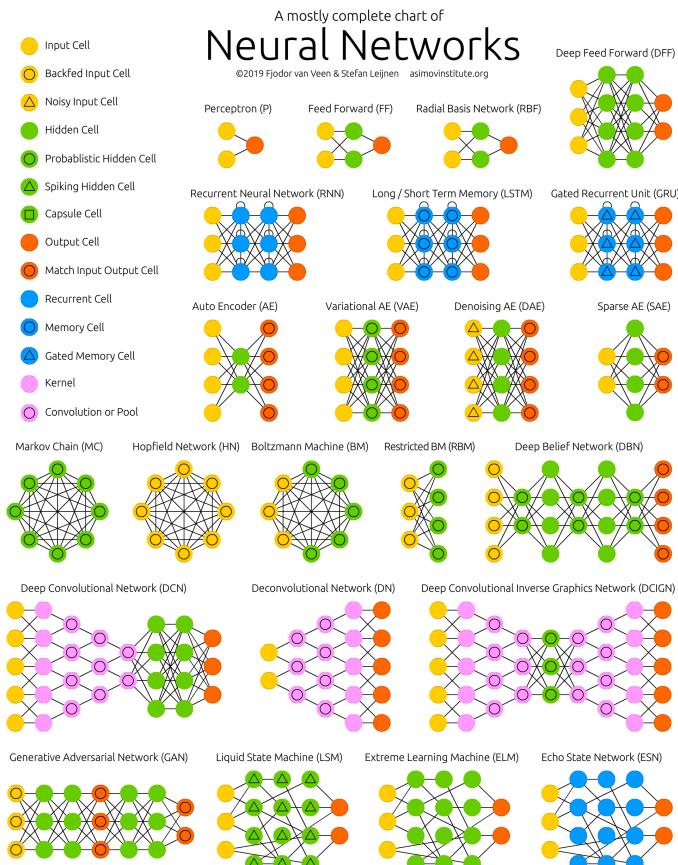


A generated image of a nighttime street scene. In the foreground, a small autonomous delivery robot with a heart-shaped logo on its front is navigating the sidewalk. An elderly man with a cane is walking away from the camera. In the background, there are cars on the road, streetlights, and storefronts, including a "RESTAURANT". A white text bubble at the top right says "can you generate a similar image". Below the image, a caption reads: "Here is the generated image based on the scene you described. It features a nighttime street with pedestrians and a small autonomous delivery robot navigating the sidewalk, along with streetlights and storefronts in the background. Let me know if you need any modifications!"

Why deep learning now?

Algorithms

New architectures



Data

Large-scale datasets (both labeled and unlabeled)



2009: 1.2 million images

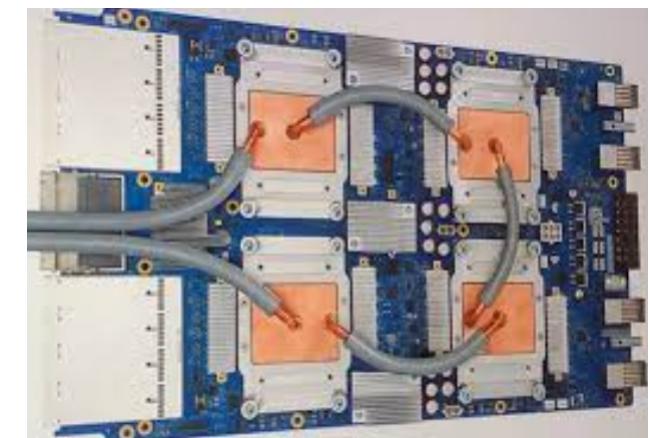


2014: 120k annotated images



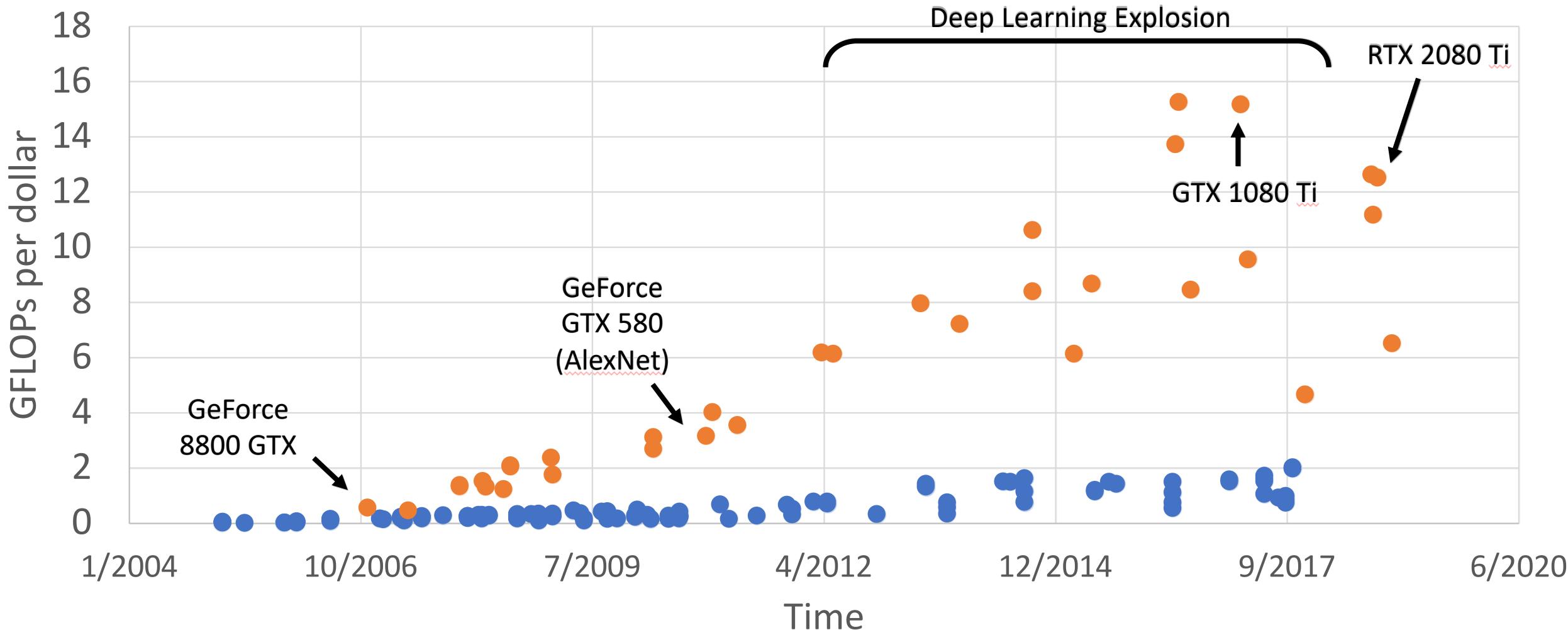
Computation

GPU, TPU, and others



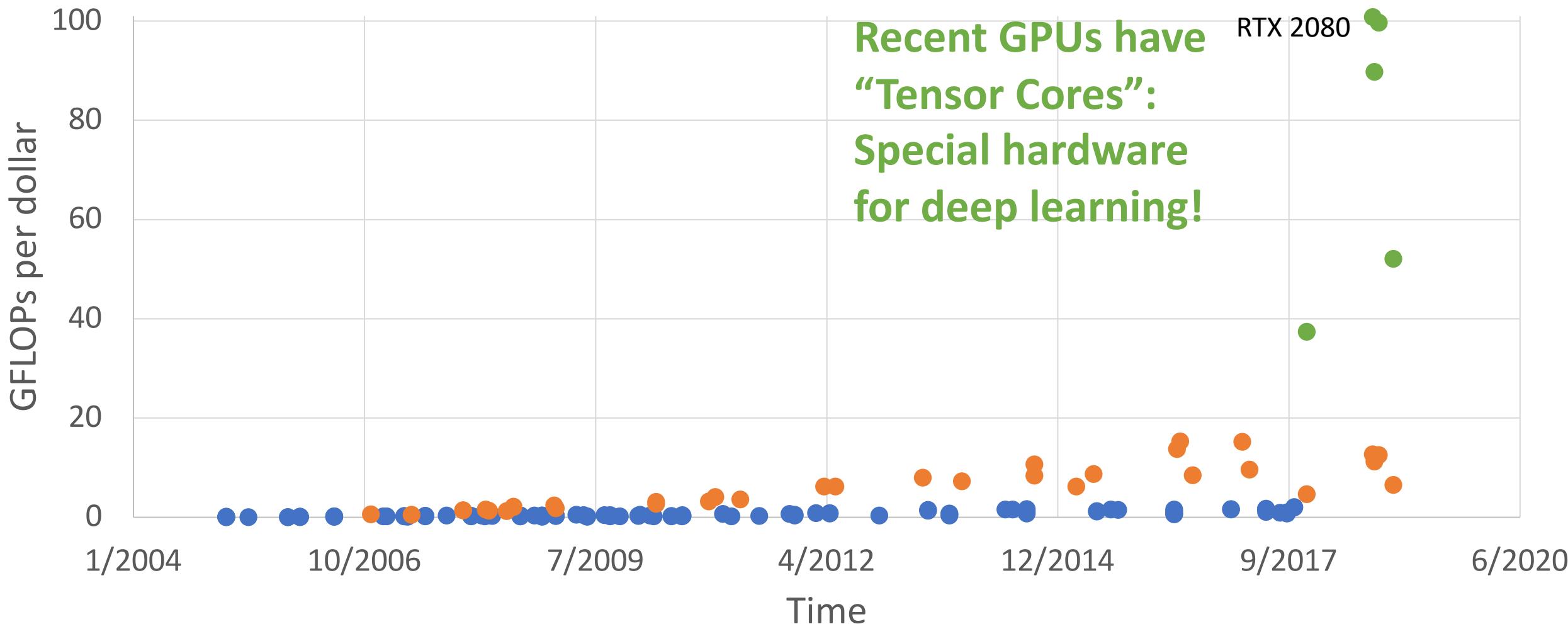
GFLOPs per Dollar

● CPU ● GPU FP32



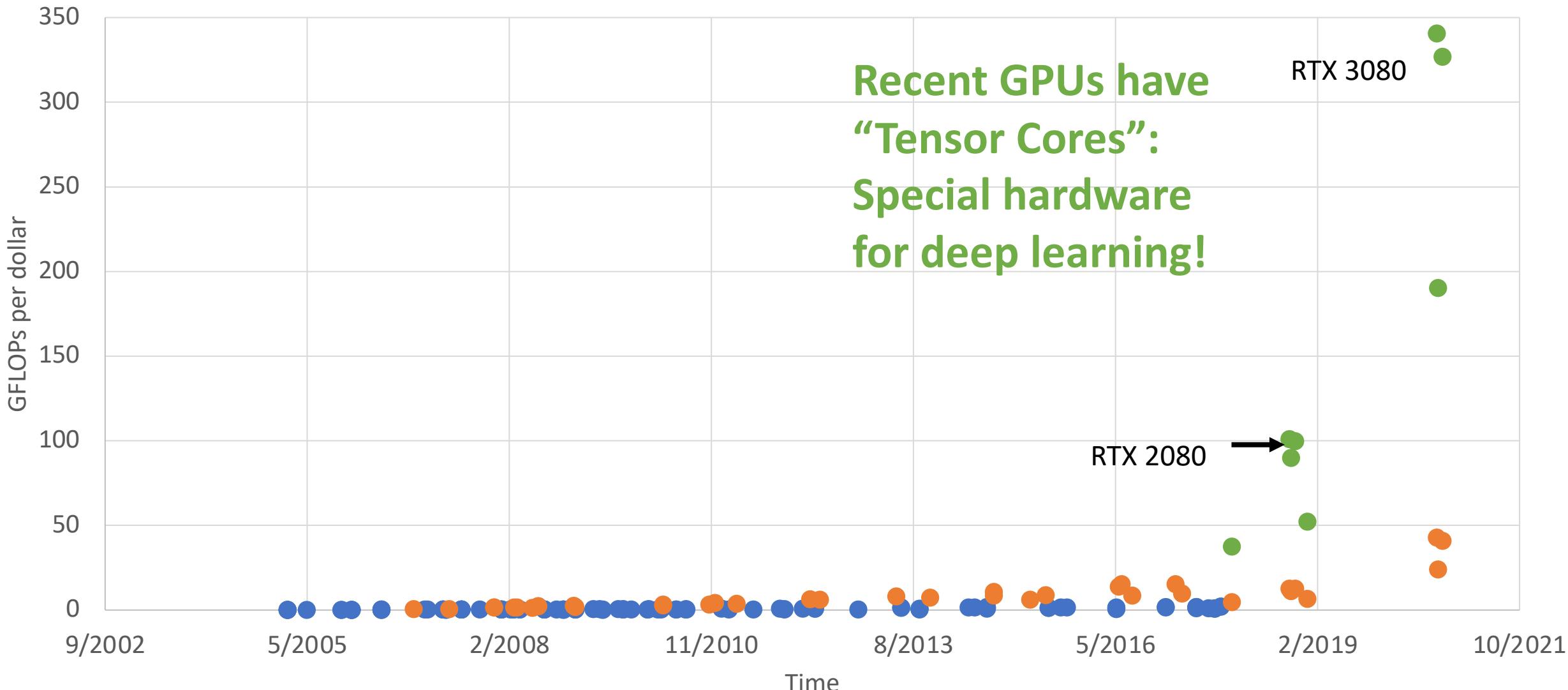
GFLOPs per Dollar

● CPU ● GPU FP32 ● GPU Tensor Core



GFLOPs per Dollar

● CPU ● GPU FP32 ● GPU Tensor Core



Not easy to buy the latest GPU

Home > News > Components > Graphics Cards

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.



By [Michael Kan](#) January 18, 2024

f X ↗ ...



; > Computers & Accessories > Computer Components > Internal Components > Graphics Cards



Up H100 Graphics Card,
80GB HBM2e
Memory, Deep
Learning, Data Center,
Compute GPU

Brand: Generic

3.2 ★★★★☆ 9 ratings

| [Search this page](#)

\$30,009⁹⁹

30-day refund/replacement

Graphics NVIDIA
Processor

Lecture 1 - 65



Elon Musk @elonmusk · Jun 2
Replying to @AravSrinivas
Given the pace of technology improvement, it's not worth sinking 1GW of power into H100s.

The @xAI 100k **H100** liquid-cooled training cluster will be online in a few months.

Next big step would probably be ~300k B200s with CX8 networking next summer.

275 555 3.6K 1.4M

FORTUNE

TECH · NVIDIA

Nvidia CEO Jensen Huang says AI chip shortage is making his customers tense and emotional

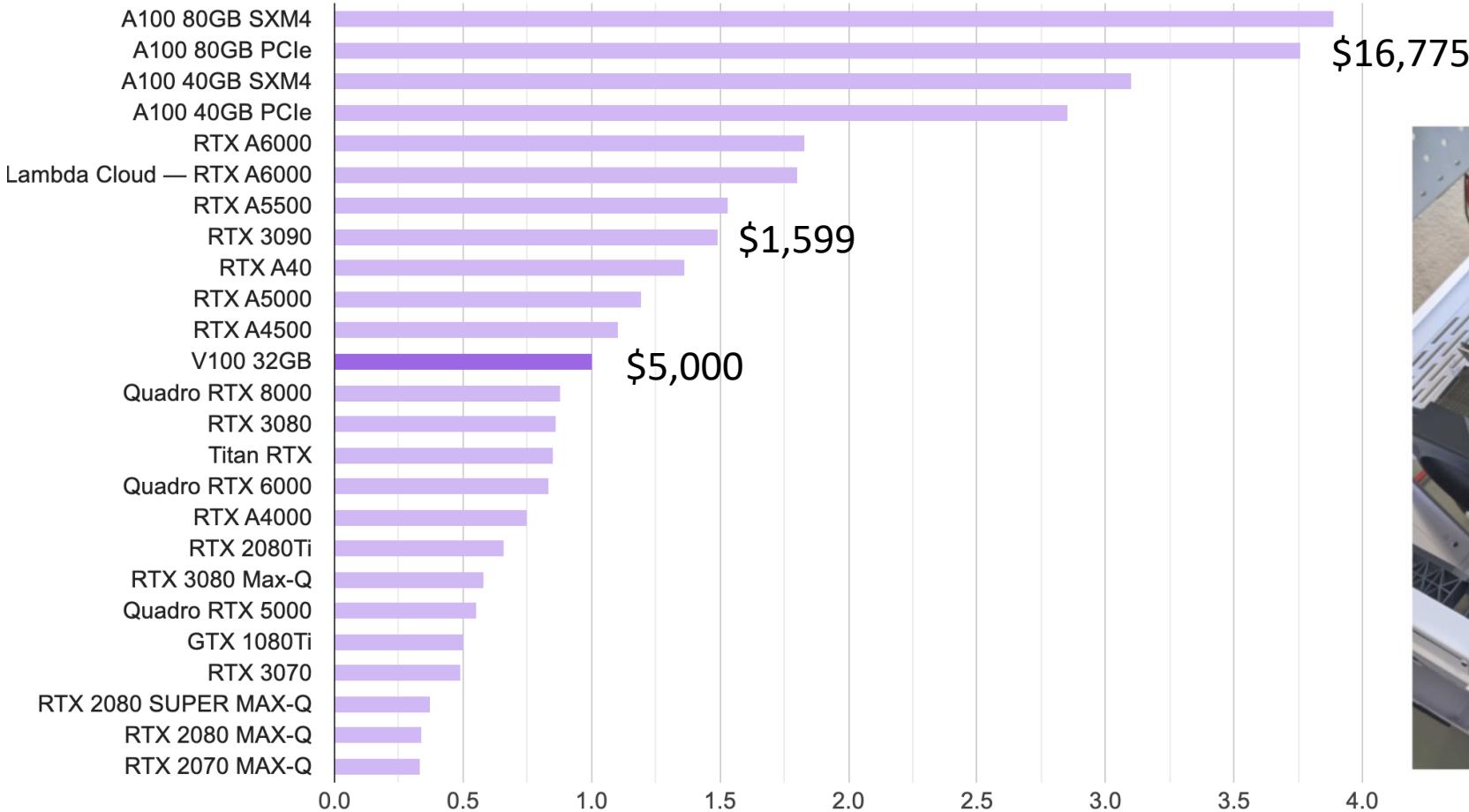
BY [CHRISTIAAN HETZNER](#)

September 12, 2024 at 5:19 AM PDT



<https://lambdalabs.com/gpu-benchmarks>

Lastest H100: \$30,000 (4x speedup over A100 on GenAI)



Revolution of Large Model (LM) = \$\$\$ Business

- Large language models (LLMs) like OpenAI's GPT-3, Google's PaLM, and NVIDIA's Megatron require huge amount of computation
- It would take 355 years to train GPT-3 on a single NVIDIA Tesla V100 GPU which is already \$5000 per unit
- Using 1,024x A100 GPUs, researchers calculated that OpenAI could have trained GPT-3 in as little as 34 days.
- Estimated that it cost around \$5M in compute time to train GPT-3.
- ChatGPT is based on GPT-4 and GPT-3.5
- “Sam Altman stated that the cost of training GPT-4 was more than \$100 million”

https://www.reddit.com/r/GPT3/comments/p1xf10/how_many_days_did_it_take_to_train_gpt3_is/

2018 Turing Award



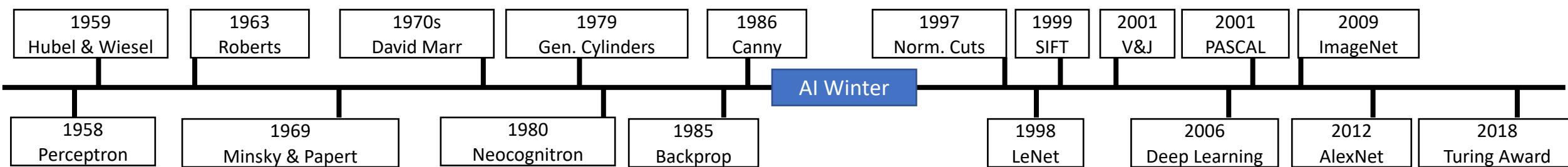
Yann LeCun



Geoffrey Hinton



Yoshua Bengio



Must Read: A Nature Review paper by the Big 3

- https://creativecoding.soe.ucsc.edu/courses/cs523/slides/week3/DeepLearning_Lecun.pdf

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun^{1,2}, Yoshua Bengio³ & Geoffrey Hinton^{4,5}

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Machine-learning technology powers many aspects of modern society: from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine-learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search. Increasingly, these applications make use of a class of techniques called deep learning.

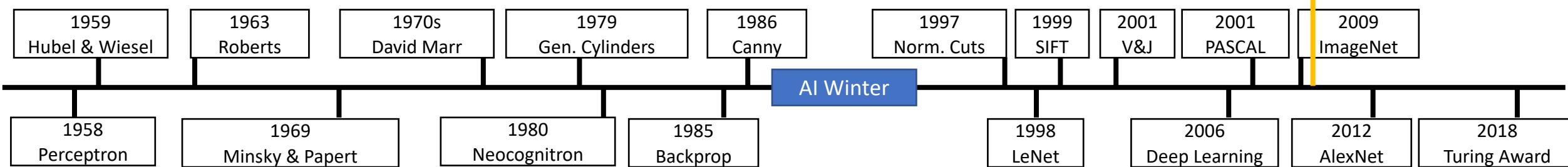
Conventional machine-learning techniques were limited in their

intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government. In addition to beating records in image recognition^{1–4} and speech recognition^{5–7}, it has beaten other machine-learning techniques at predicting the activity of potential drug molecules⁸, analysing particle accelerator data^{9,10}, reconstructing brain circuits¹¹, and predicting the effects of mutations in non-coding DNA on gene expression and disease^{12,13}. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding¹⁴, particularly topic classification, sentiment analysis, question answering¹⁵ and lan-

Personal Journey in Computer Vision



Me started doing CV
research in 2010



Started as a Pre-deep learning CV researcher

- My first CVPR paper in 2011
- Research topic (2011-2013): Collective motion detection in videos



Zhou et al. Coherent Filtering. ECCV'12

2013- 2015: (Deep) Visual Scene Understanding

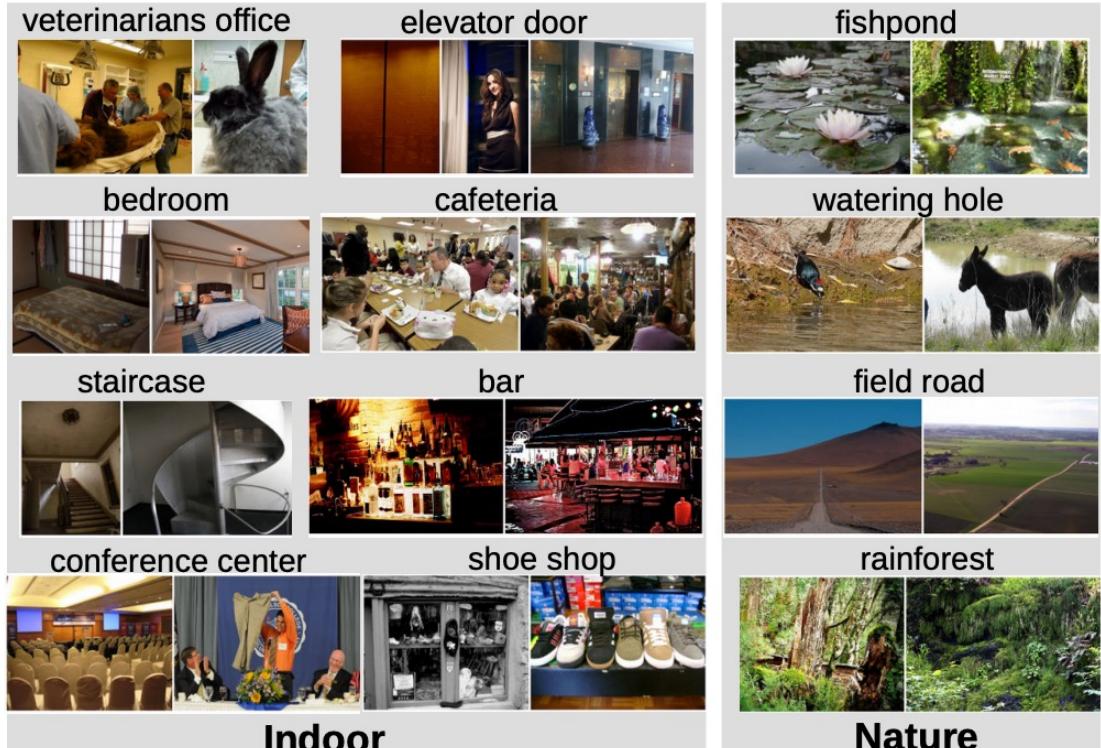
- From object understanding to scene understanding (PhD Thesis work)



2013- 2015: (Deep) Visual Scene Understanding

places ● ● ●

10 million images from >400 scene categories



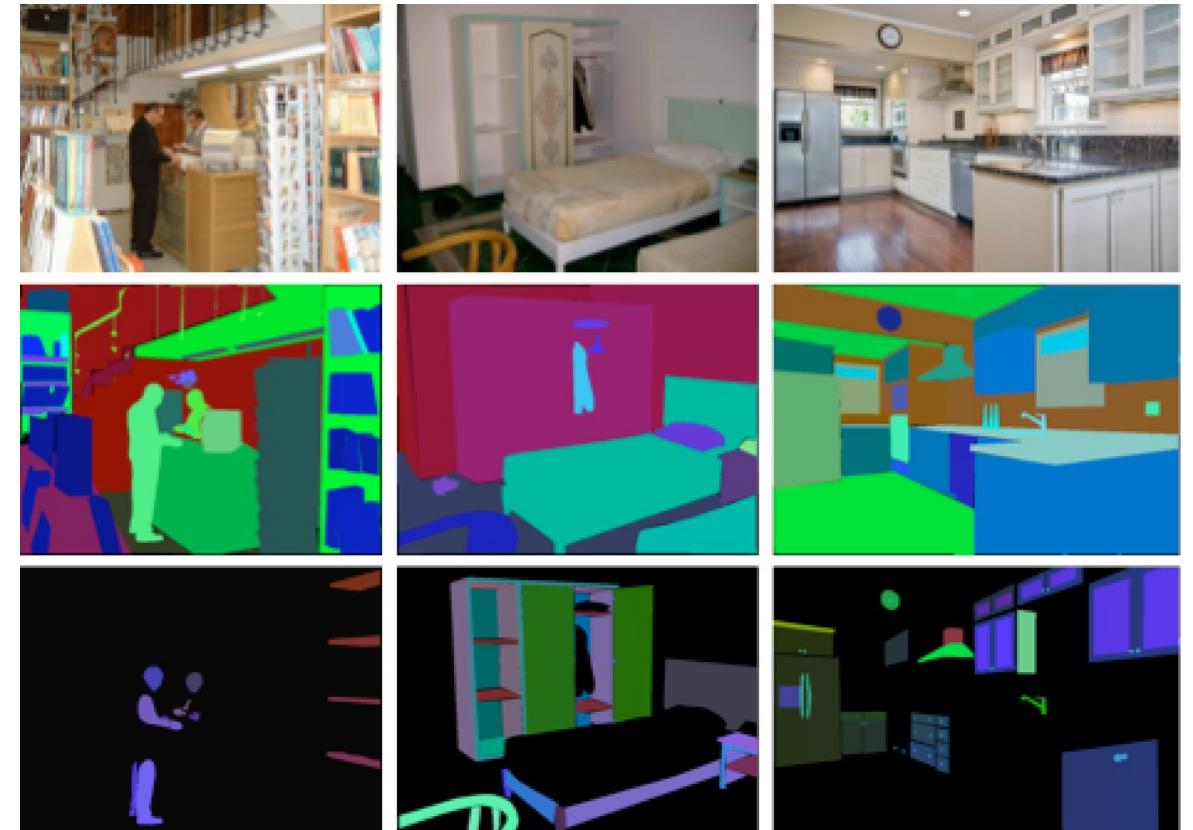
<http://places2.csail.mit.edu>

Scene recognition demo

A screenshot of a mobile phone screen showing a scene recognition demo. The top status bar shows signal strength, AT&T connectivity, 11:05 PM, and the URL places2.csail.mit.edu. The main interface asks "What scene is your image from? Upload to find out!" with a "Choose File" button and a thumbnail showing "1 photo". Below this is a photograph of a conference room with people seated around a long table. At the bottom, it displays "Predictions from PlacesCNN:" followed by "Type: indoor" and "Scene categories: conference_room (0.456)".

2013- 2015: (Deep) Visual Scene Understanding

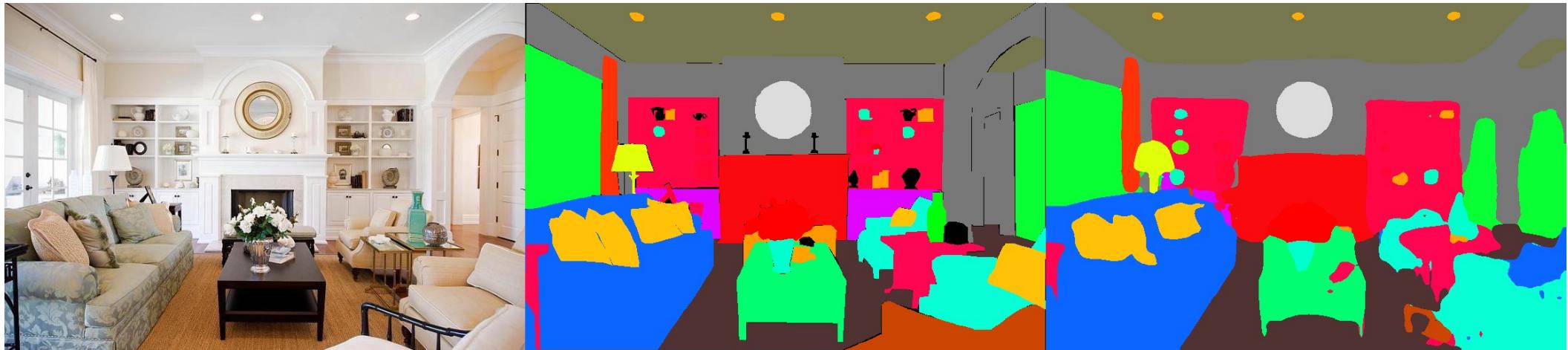
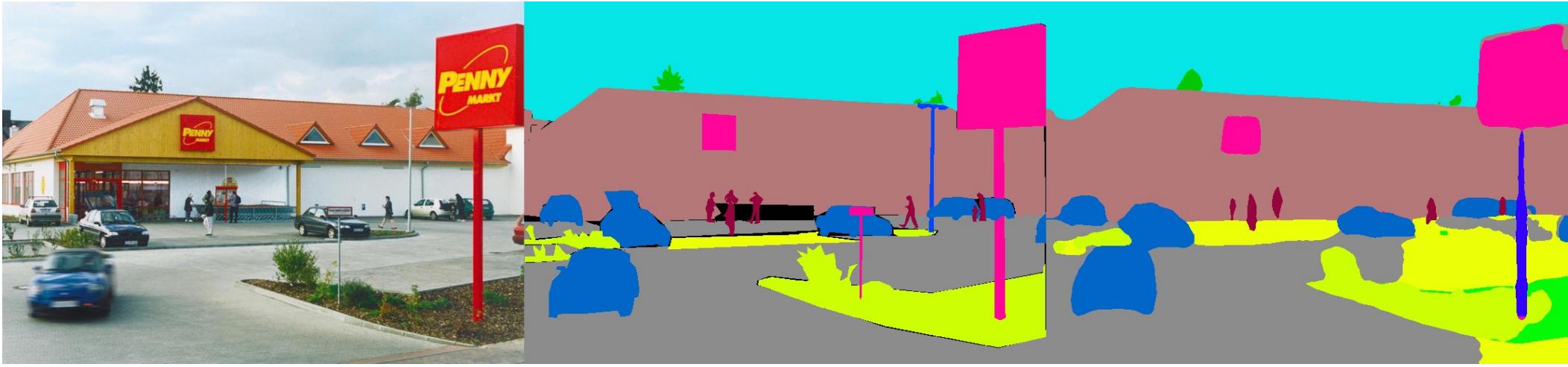
ADE20K: pixel-wise scene parsing benchmark & networks



20,000 fully annotated images by one annotator's
>5 years effort

<https://groups.csail.mit.edu/vision/datasets/ADE20K/>

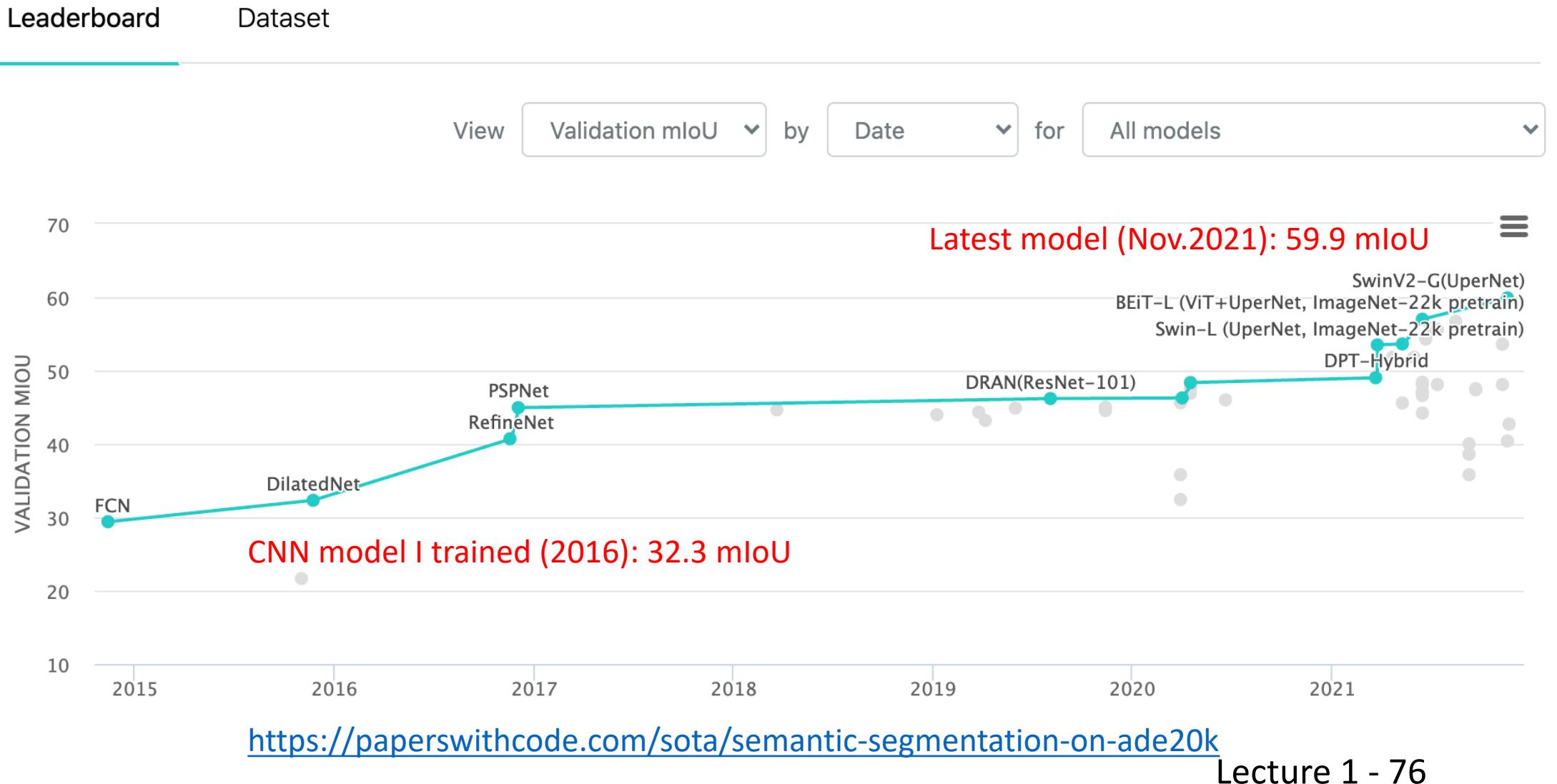
2013- 2015: (Deep) Visual Scene Understanding



<https://colab.research.google.com/github/CSAILVision/semantic-segmentation-pytorch/blob/master/notebooks/DemoSegmenter.ipynb>

<https://github.com/CSAILVision/semantic-segmentation-pytorch>

Progress for semantic segmentation on ADE20K

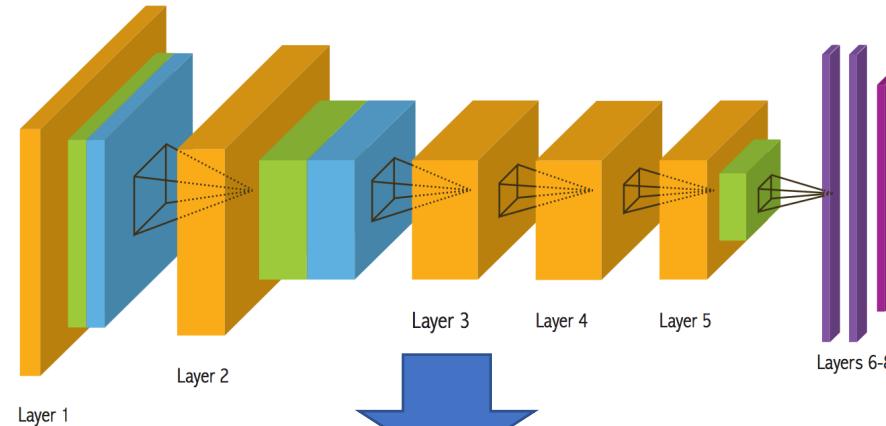


2015-2018: Understanding Deep Neural Networks

[Zhou et al., CVPR'16]



Convolutional Neural Network (CNN)



Explaining Prediction

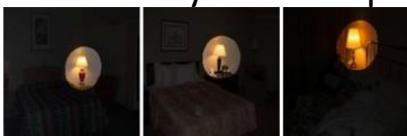
Cafeteria (0.9)

Interpreting Internal Representations

Unit5 at Layer3 : Trademark



Unit2 at Layer4: Lamp



[Zhou et al., NIPS'14]

[Zhou et al., ICLR'15]

[Bau*, Zhou* et al., CVPR'17]

2015-2018: Understanding Deep Neural Networks

Class Activation Mapping (Zhou et al, CVPR'16)

Prediction: Conference Center



Prediction: Indoor Booth



2015-2018: Understanding Deep Neural Networks

Prediction: Martial Art gym



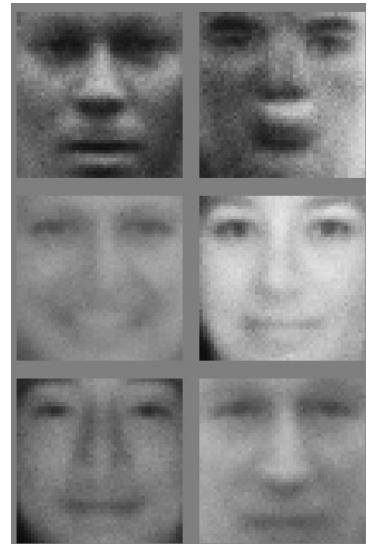
2015-2018: Understanding Deep Neural Networks

Class Activation Mapping (Zhou et al, CVPR'16)



2018-2021: Controlling Generative Models (GANs)

2014



2015



2017



2018



2018



2019



GAN

(Goodfellow et al)

DCGAN

(Radford et al)

PG-GAN

(Karras et al)

StyleGAN

(Karras et al)

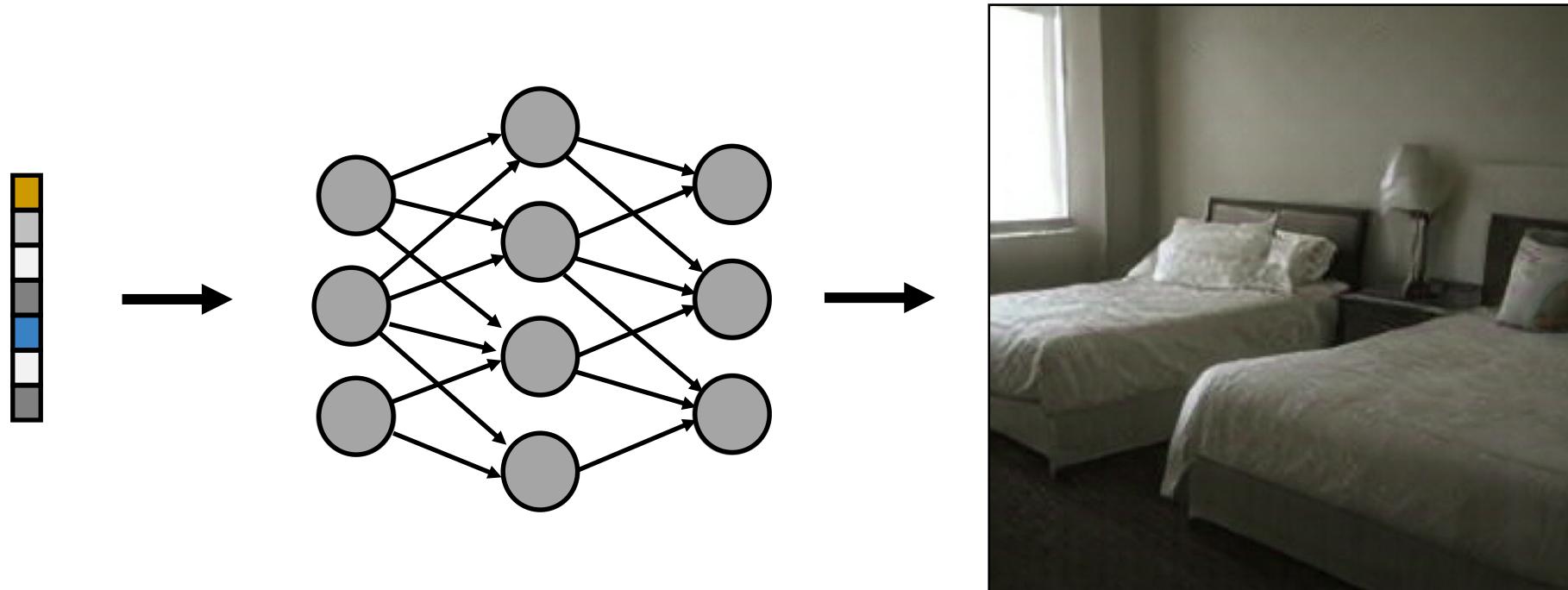
BigGAN

(Brock et al)

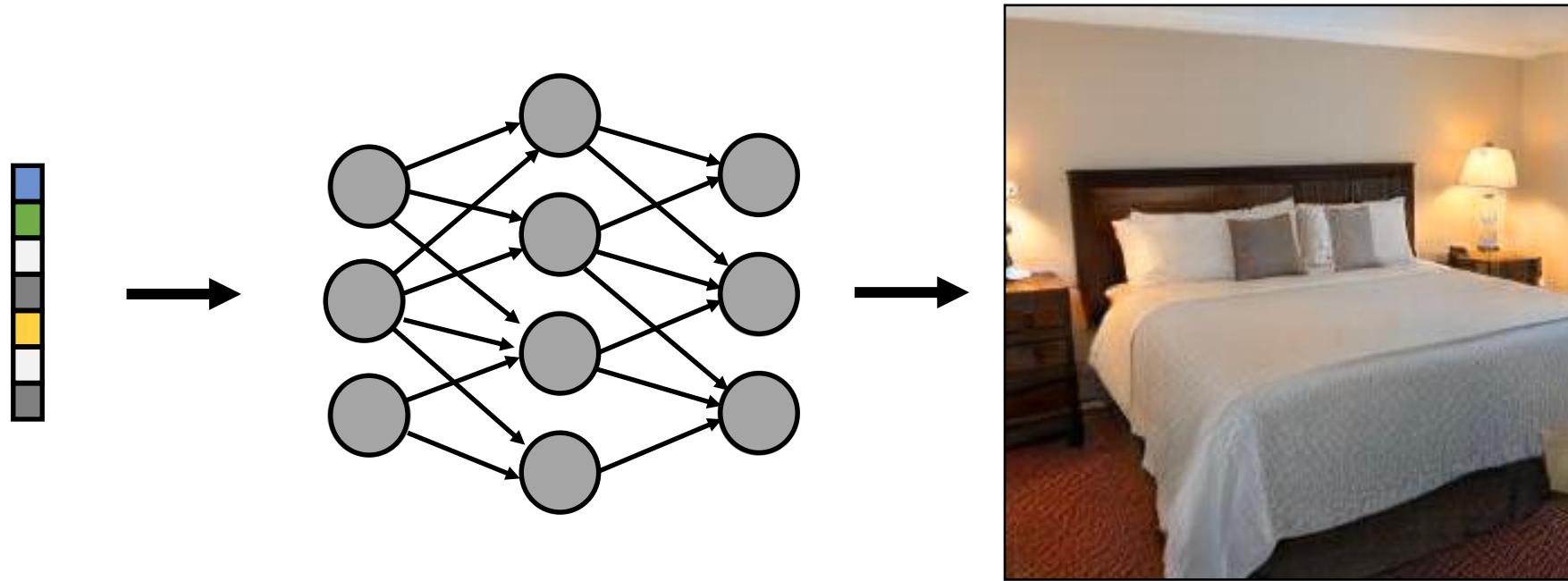
StyleGAN v2

(Karras et al)

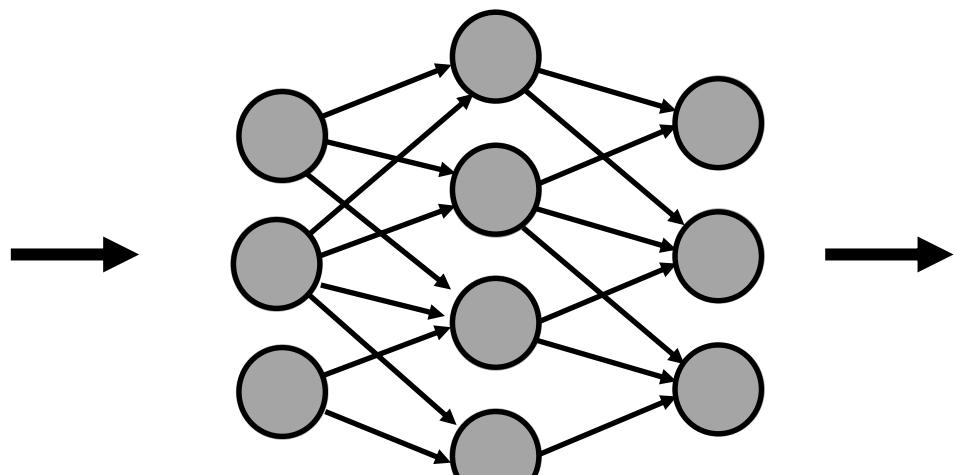
2018-2021: Controlling Generative Models (GANs)



2018-2021: Controlling Generative Models (GANs)



2018-2021: Controlling with Generative Models (GANs)



Steerability of generative models:
How to steer the model to customize output?

2018-2021: Controlling Generative Models (GANs)

Changing the view of the scenes



Improving the green space



Yang, Shen, Zhou. IJCV'20

2018-2021: Controlling Generative Models (GANs)

Human-in-the-loop AI
content creation



SeFa: Closed-Form Factorization of Latent Semantics in GANs

Random Sample



Recent model: controlling diffusion model

Another line of research

2019-now: Embodied AI and Machine Autonomy



<https://decisionforce.github.io/metadrive/>

Lecture 1 - 88

Very recent work: [MetaUrban](#)



2020-now: Human-in-the-loop Learning

Human-in-the-loop Reinforcement Learning

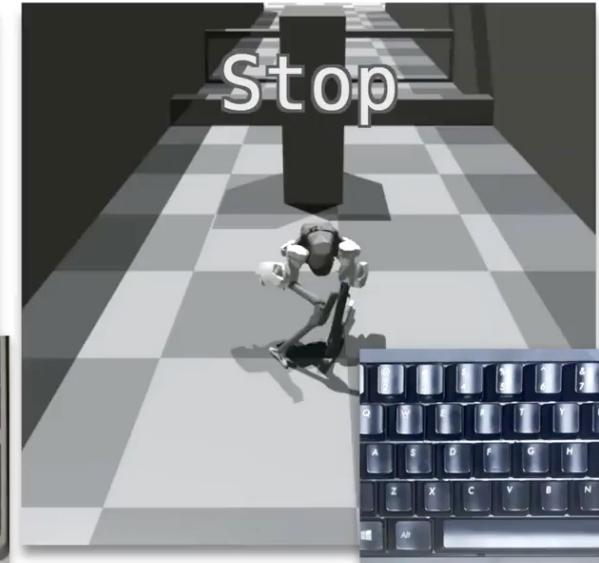


Human-in-the-loop inference

A1 Quadruped Robot



Cassie Bipedal Robot

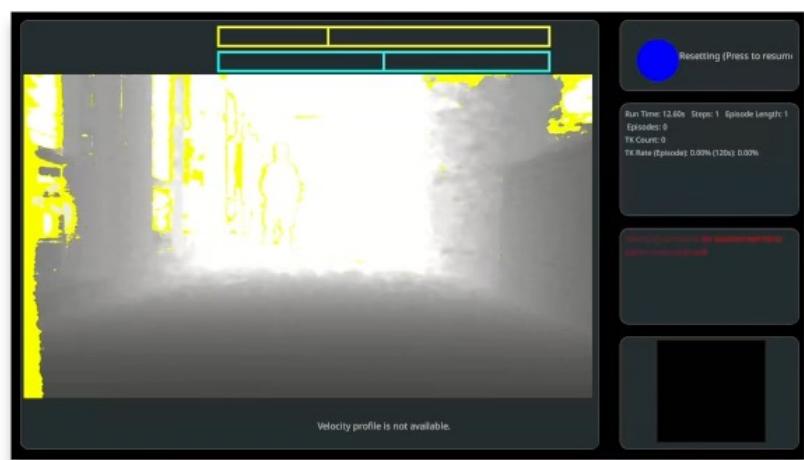


NeurIPS'22. <https://metadriverse.github.io/policydissect/>

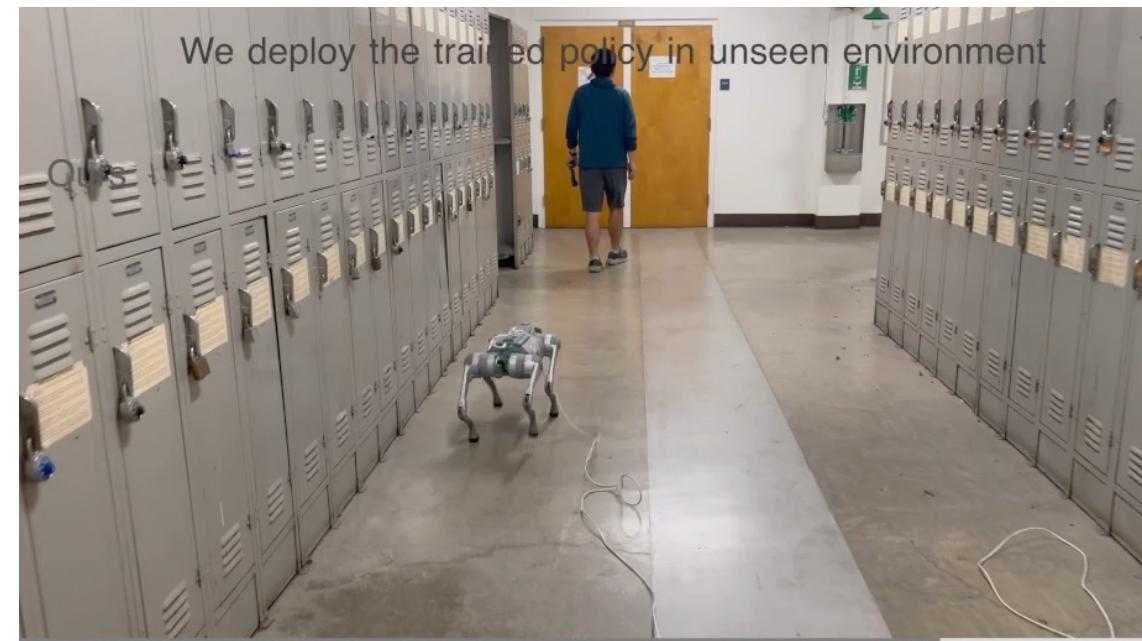
NeurIPS'23. <https://metadriverse.github.io/pvp/>

Lecture 1 - 90

2024-now: Real-world Robotic Mobility



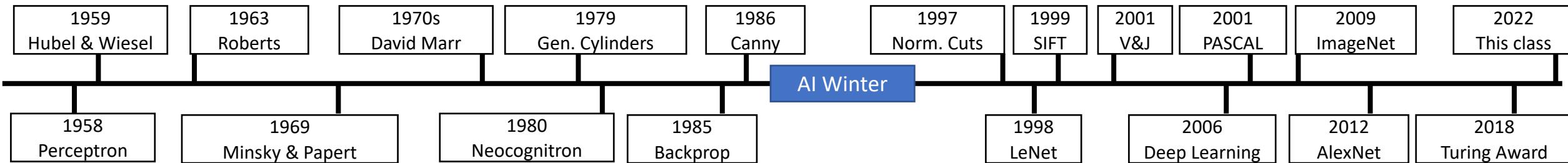
Train - Following ★ Eval - Following ★ Train - Nav. ★ Eval - Nav. ★ Compare BC ★ Zero-shot Deploy



Train - Following ★ Eval - Following ★ Train - Nav. ★ Eval - Nav. ★ Compare BC ★ Zero-shot Deploy

Today's Agenda

- A brief history of computer vision and deep learning
- Personal journey in computer vision research
- Course overview and logistics



Teaching Assistants, Learning Assistant, Readers

Teaching Assistant (TA)



Zhizheng Liu
zhizheng@cs.ucla.edu)
PhD student at CS



Sicheng Mo
smo3@cs.ucla.edu)
PhD student at CS

Learning Assistant (LA)



Joe Lin
CS Undergrad



Krystof Latka
CS Undergrad

Course Information

- Course Master Page:
<https://docs.google.com/document/d/1oeQWkJmCuHfJKAY4QAKgU6o3u33CJcgLxXZInDap7ug/edit?usp=sharing>
 - Syllabus, schedule, assignment link, etc
- Bruinlearn course site: <https://bruinlearn.ucla.edu/courses/191395>
 - Modules: where we share the slides and you submit assignment
- Piazza: <https://piazza.com/ucla/fall2024/cs163>
 - (Almost) all questions about the course should go here
 - Use private questions if you want to post code
 - Learning Assistants (LAs) will do question-answering, TAs will assist and oversee.

Week 1: Overview

Lecture 1: Course introduction

Lecture 2: Image classification and machine learning basics

Week 2: Image classification and linear neural networks

Lecture 3: Linear classifier for image classification

Lecture 4: Regularization + Optimization

Week 3: Foundation on neural networks

Assignment 2 out

Lecture 5: Neural Networks

Lecture 6: Backpropagation

Week 4: Convolutional neural networks

Lecture 7: Convolutional Neural Networks

Lecture 8: CNN architectures

Week 5: Training and understanding neural networks

Lecture 9: Training NNs

Lecture 10: Visualizing and understanding neural networks

Week 6: RNN, Attention, Transformer, ViT

Assignment 3 out

Lecture 11: RNN and Attention

Lecture 12: Transformer and ViT

Week 7: Object detection and dense prediction (Monday is Veterans Day)

Lecture 13: Object detection

Week 8: Dense prediction, Generative models: intro

Assignment 4 out

Lecture 14: Dense prediction (semantic segmentation, pose estimation)

Lecture 16: Generative models: intro

Week 9: Generative models: Advances

(Thanksgiving week)

Lecture 17: Generative models: advances

Lecture 18: A taste of CV research by various student researchers (will be pre-recorded as is too close to Thanksgiving break...)

Week 10: CV research and course summary

Lecture 19: Recent progress and how to do CV research

Lecture 20: Course Recap

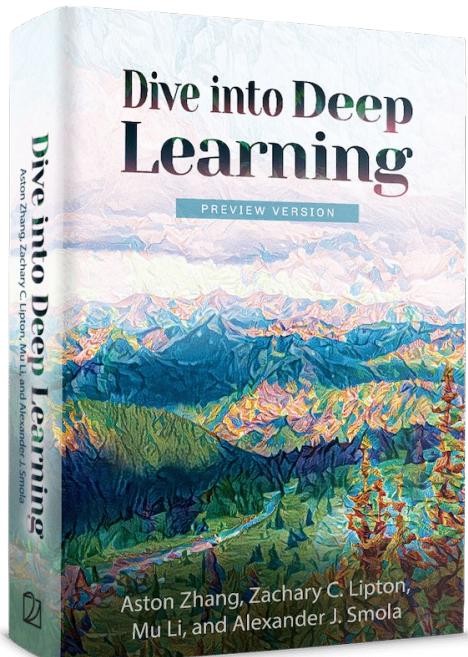
Piazza Etiquette

- Post only short snippets of code (< 20-30 lines)
- Ask a specific and concrete question
- Explain what you have tried so far, and what happened
- See StackOverflow guide on asking good questions:
<https://stackoverflow.com/help/how-to-ask>
- The reply turnaround time from TAs is about 36 hours (unfortunately we have only 2 LAs and 2 TAs for 160 students)

Optional Textbooks (all three freely available online)

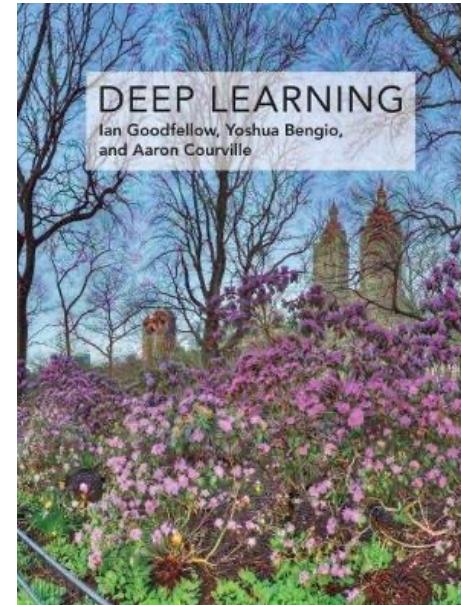
Dive into Deep Learning

- <https://d2l.ai/>
- Detailed code and practice



Deep Learning

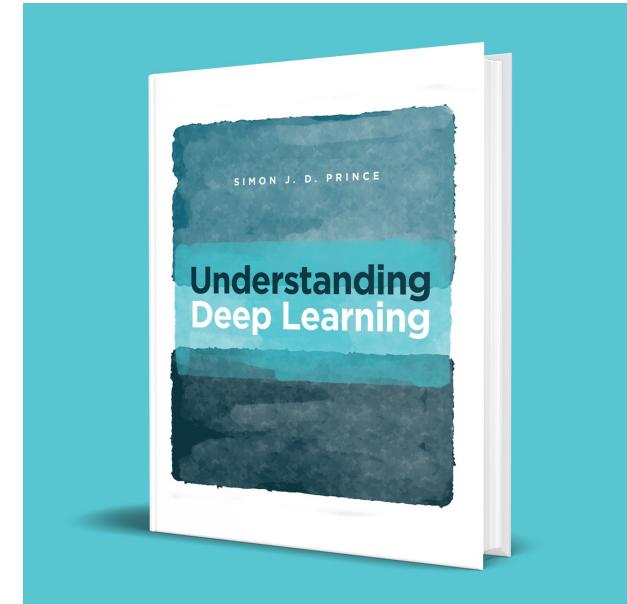
- <https://www.deeplearningbook.org/>



(latest)

Understanding Deep Learning

<https://udlbook.github.io/udlbook/>



Grading

- 4 programming assignments: 40%
 - Homework assignments will be based on Python, PyTorch, and Google Colab
- Final: 50%
- Course Project: 10%
- Late policy
 - 3 free late days to use on assignments
 - Once free late days are exhausted, 25% penalty per day

Collaboration Policy

- **Rule 1:** Don't look at solutions or code that are not your own; everything you submit should be your own work
- **Rule 2:** Don't share your solution code with others; however discussing ideas or general strategies is fine and encouraged
- **Rule 3:** Indicate in your submissions anyone you worked with
- Turning in something late / incomplete is better than violating the honor code

Course Structure

- First half: Fundamentals
 - Details of how to implement and train different types of networks
 - Image classification as the core CV application
 - How to train and debug the networks
- Second half: Various CV applications and trendy topics
 - Less detailed: provide overview and references, but skip some details
 - Object detection, image segmentation, videos
 - Attention, Transformers
 - Generative models: GANs, VAEs, NeRF, Diffusion, etc

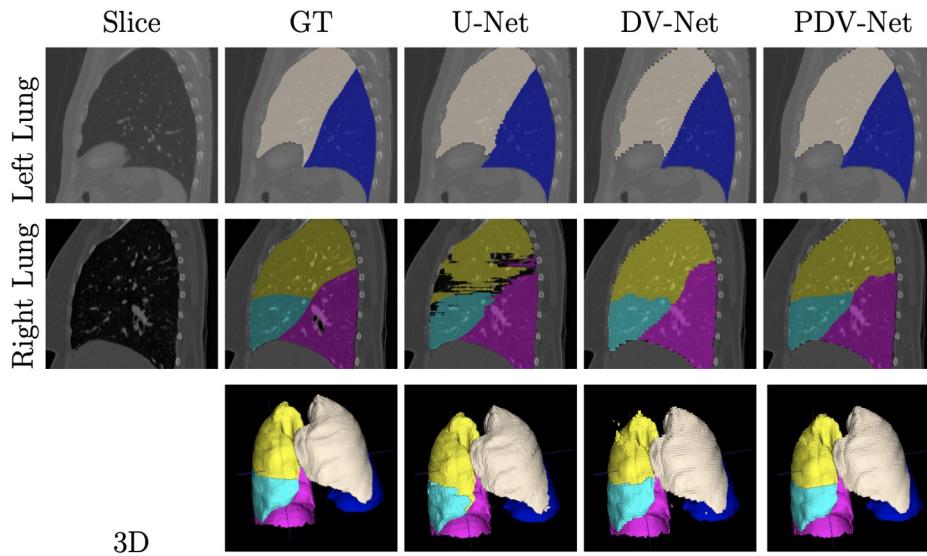
Homework assignment

- <https://github.com/UCLAdeepvision/CS163-Assessments-2024Fall>
- Assignment 1 is released today
- Due: Sunday, Oct 20
- Start early!

Assignment 1	OUT: Monday, Sept 30	DUE: Sunday, Oct 20
Assignment 2	OUT: Monday, Oct 14	DUE: Sunday, Nov 3
Assignment 3	OUT: Monday, Nov 4	DUE: Sunday, Nov 23
Assignment 4	OUT: Monday, Nov 18	DUE: Sunday, Dec 8

Course Project

- Computer vision is a very broad field, and cannot be fully covered in an introductory course
- This project gives you the opportunity to learn about a topic/problem in CV that **you** are passionate about



<https://ucladeepvision.github.io/CS188-Projects-2022Winter/2022/01/27/team07-medical-image-segmentation.html>

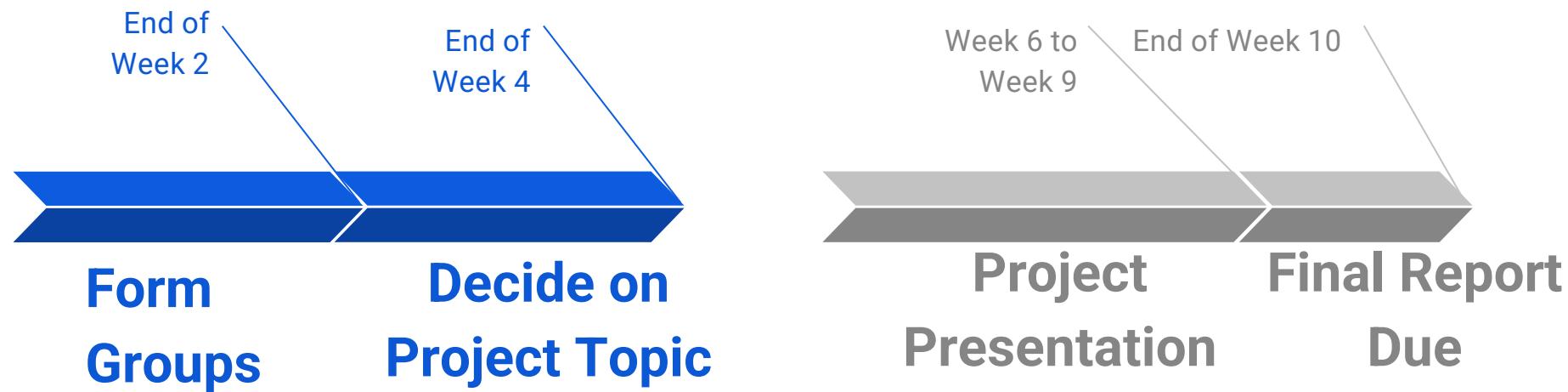


<https://ucladeepvision.github.io/CS188-Projects-2023Winter/2023/03/28/team39-AstroCV.html>

Course Project

- Up to 4 students per group
- 10% of final grade (5% presentation + 5% final report)
- Presentation (~15min) held in-person during discussion sections
 - Focus is on introducing the topic to those unfamiliar with the field
- Final report due at the end of the quarter
 - Formatted as a technical blog article – will be made available to general public
- **Extra Credit Opportunity!**
 - Implementing/running existing approaches and including them in the presentation/report
 - Google Cloud credit will be provided in case students need to run compute-heavy code

Course Project Timeline (Tentative, see [Google Doc](#) for latest)



During discussion section
Sign-up sheet will be made available to select presentation slot

Final report will be in the webpage format

See previous year's course project reports:

<https://ucladeepvision.github.io/CS188-Projects-2024Winter/>

<https://ucladeepvision.github.io/CS188-Projects-2023Winter/>

<https://ucladeepvision.github.io/CS188-Projects-2022Winter/>

Pratosh Menon, Bulent Yesilyurt, Hayden D'Souza, Riley Bruins on Mar 22, 2024

Deep Learning for Prostate Segmentation

We aim to analyze how we can use deep learning technique for prostate image segmentation. Prostate cancer is the second most common form of cancer for men worldwide and the fifth leading cause of death for men globally [3]. However, this is a statistic that can be considerably changed with early stage detection. In fact, the cancer is completely curable within 5 years if we catch it early. To this end, we explore how we can use existing deep learning architectures to help with prostate image segmentation to catch early prostate cancer in patients.

William Zhou, Leon Lenk, Maxine Wu, Artin Kim on Mar 22, 2024

Multimodal Vision-Language Models: Applications to

LaTeX OCR

LaTeX is widely utilized in scientific and mathematical fields. Our objective was to develop a pipeline capable of transforming hand-written equations into LaTeX code. To achieve this, we devised a two-step model. Initially, we employed an R-CNN model to delineate bounding boxes around equations on a standard ruled piece of paper, utilizing a custom dataset we generated. Subsequently, we passed these selected regions into a TrOCR model pre-trained on Im2LaTeX-100k, a dataset comprising rendered LaTeX images. We further fine-tuned the model on a handwritten mathematical expressions dataset on Kaggle, which is a collection of the CROHME handwritten digit competition datasets over three years [6] [7] [8]. Our model successfully generated the ground LaTeX accurately for 4 out of 8 hand-drawn examples we produced. For the remaining 4 examples, it produced LaTeX similar to the ground truth, albeit with minor errors.

William Zhao, Suhas Nagar, Lucas Jeong on Mar 22, 2024

Navigating the Future: A Comparative Analysis of Trajectory Prediction Models

Trajectory prediction is a challenging task due to the multimodal nature of human behavior and the complexity in multi-agent systems. In this technical report we explore three machine learning approaches that aim to tackle these challenges: Social GAN, Social-STGCNN, and EvolveGraph. Social GAN uses variety loss to generate diverse trajectories and a pooling module to model subtle social cues. Social-STGCNN models social interactions explicitly through a graphical structure. EvolveGraph establishes a framework for forecasting the evolution of the interaction graph. We compare the advantages and disadvantages of these approaches at the end of the report.

Botao Xia on Mar 22, 2024

Super-resolution via diffusion method

Super resolution enhances image resolution from low to high, with modern techniques like convolutional neural networks and diffusion models like SR3 significantly improving image detail and quality. This Post explore a simplified implementation of SR3. View code [\[Here\]](#)

Colin Melendez on Mar 22, 2024

NeRFs for Synthesizing Novel Views of 3D Scenes

In the domain of generative images, NeRFs are a powerful tool for generating novel views of 3d scenes with an extremely high degree of detail. here, we will review the basics of Neural Radiance Field (NeRF) models, look at how they can be used to generate novel views, and investigate an optimization to the original design with the KiloNeRF model to see how we can improve on some of its shortcomings.

Fuyi Yang, John Tran on Mar 22, 2024

Deep Learning Based Image-to-Image Translation Techniques

The goal of this project is to explore and understand the problem of image to image translation. Two approaches addressing this topic will be analyzed: CycleGAN and FreeControl. An implementation of CycleGAN is also discussed.

Liyu Zerihun on Mar 22, 2024

Exploring CLIP for Zero-Shot Classification

In this blog article, we'll delve into CLIP(Contrastive Language-Image Pre-training), focusing primarily on its application in zero-shot classification. Unlike examining a pre-trained version, we will embark on training CLIP ourselves, crafting the core components of the model and employing a distinct, smaller dataset for training purposes. We'll also introduce custom loss functions and dissect specific elements of CLIP, such as its contrastive loss mechanism. This article aims to dissect the architecture and its implications, making minor adjustments to better grasp what drives its effectiveness.

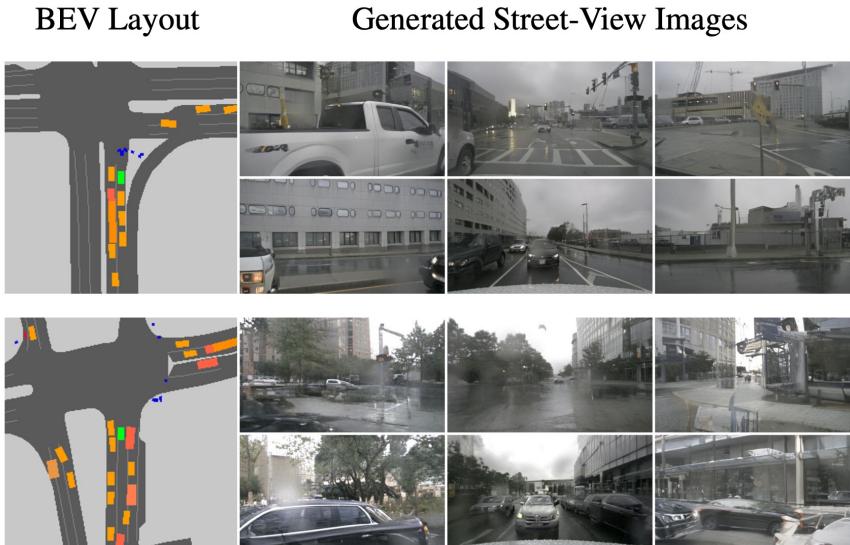
Cloud computing for course project/assignment

- Computational resource is valuable, no free food and you pay what you get
- Each student will get \$50 Google educational cloud credit
 - Instruction will be sent in week 2
- Consider to buy the Google Colab Pro version or on-demand
 - \$9.99 per month (see <https://colab.research.google.com/signup>), or pay as you go
 - You can cancel it after you finish the course
 - Well, computing resource is like electricity, you have to pay :/

Student achievement after the course

- 1st time CS188: ~70 students to start with, and 37 students finished
- 2nd time CS188: ~110 students to start with, and 80 students finished
- 3rd time CS188: ~160 students signed up, and 140 students finished
- Some students became student researchers and continue their research career in computer vision and deep learning

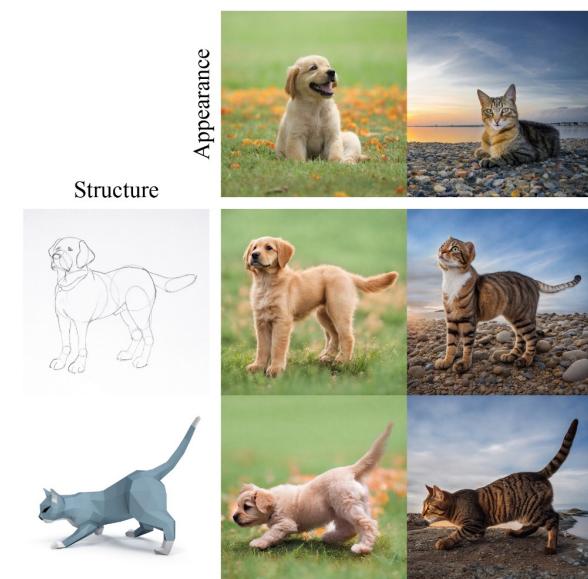
Alex Swerdlow's work on street-view image generation BEVGen (published at RAL)



Wenjie Mo's work on human-in-the-loop driving (published at NeurIPS'23)



Jordan Lin's first-authored paper [Ctrl-X to NeurIPS'24!](#)



Attendance

- We don't record attendance
- Feel free to skip the first few lectures if you have already learned how to train neural networks
- ZOOM link will be provided, as well as recording (but it happened before that I forgot to click the recording button), but the quality is bad.
- If you have any class overlap and you will miss more than 1/3 in-person lectures, you are highly suggested to drop the course

Next time: Image Classification