

COMP550

ADVANCES IN DEEP LEARNING

Photo: Detail from a synthetic image generated by Janelle Shane using the prompt "The Great Hall of the Sacred Library by James Gurney" based on @RiversHaveWings' VQGAN+CLIP Colab notebook.

Lecture #01 – Introduction



KOÇ
UNIVERSITY

Aykut Erdem // Koç University // Fall 2021

AI – Hype or Reality?

MIT
Technology
Review

10 Breakth



MIT
Tech
Revie

Business

M
Leaders
Ev

Recent
useful
of the

by Tom



Login / Register



HUFFPOST BUSINESS

Edition: US

15,644 views | Jul 12, 2018, 08:30am

FRONT PAGE



WIRED



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF



POST



HUFF

POST

Act I – Prologue: Then and Now



NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) — The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

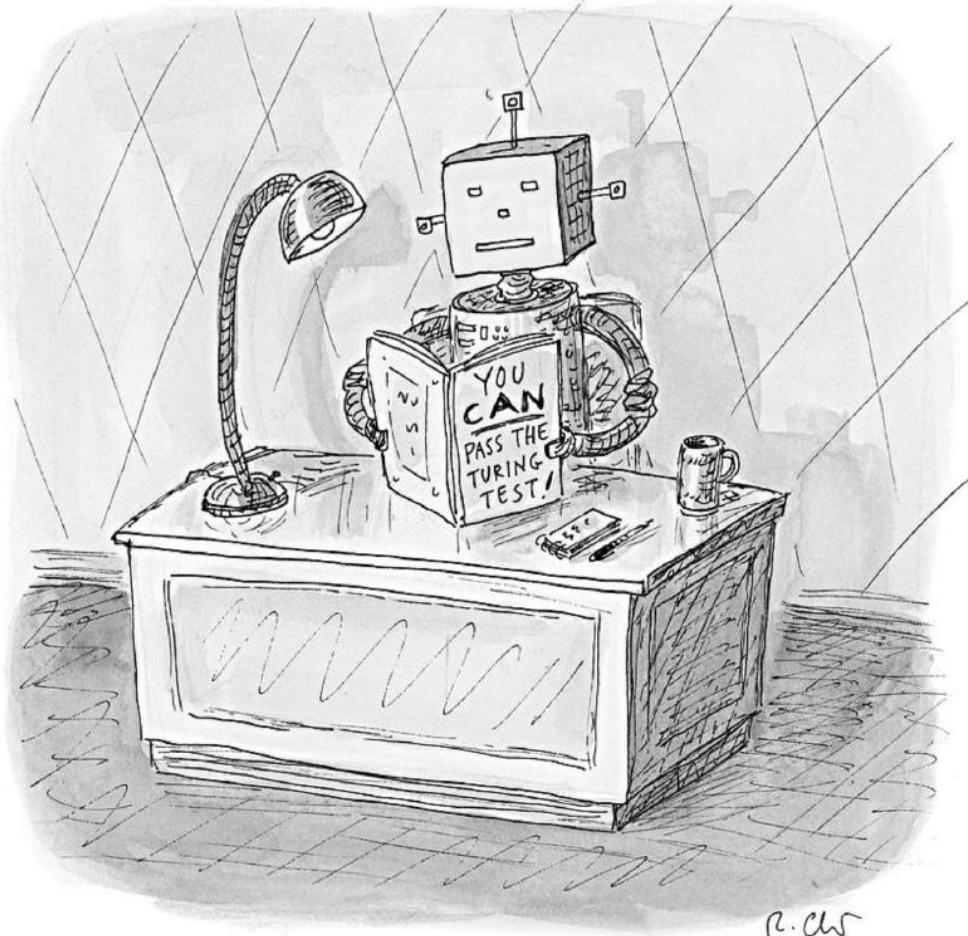
The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen..

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

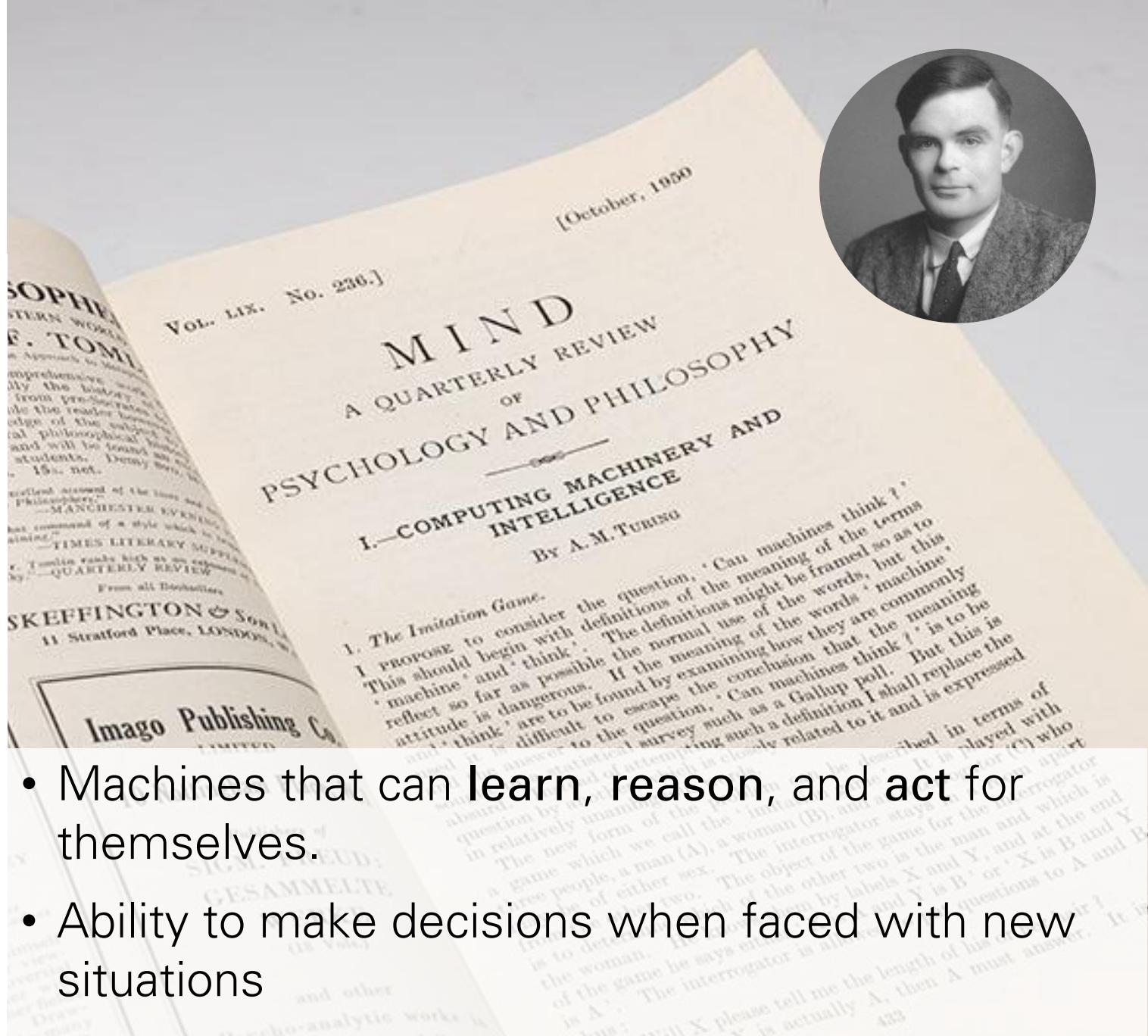
Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

What is AI?



Cartoon credit: Roz Chast, New Yorker, March 6th 2017



- Machines that can **learn, reason, and act for themselves.**
- Ability to make decisions when faced with new situations

AI: Goals, Algorithm & Performance

Broad Intelligence

Capable of performing a wide variety of tasks without modification or retraining

Narrow Intelligence

Designed to solve a single problem

The image shows a screenshot of the IMDb page for the TV series "Futurama". The main title "Futurama" is at the top, followed by a star rating of 8.4/10 from 199,797 reviews. Below the title, there's a "Episode Guide" section with 124 episodes. A thumbnail for the first episode, "Jurassic Bark", is shown, featuring Fry, Bender, and Leela. To the right, there's a "Top-Rated Episodes" section with three entries: "Jurassic Bark" (S5.E2), "The Luck of the Fryrish" (S3.E10), and "Meanwhile" (S7.E26). Each entry includes a thumbnail, the episode title, and a brief description.

Futurama

8.4 /10 199,797 Rate This

Episode Guide 124 episodes

FUTURAMA

0:16 | Trailer

10 VIDEOS | 683 IMAGES

Philip J. Fry, a pizza delivery boy, is accidentally frozen in 1999 and thawed out on New Year's Eve 2999.

Creators: David X. Cohen, Matt Groening
Stars: Billy West, John DiMaggio, Katey Sagal | See full cast & crew »

+ Add to Watchlist

Reviews 314 user | 66 critic

Popularity 351 (↑38)

Top-Rated Episodes

S5.E2 Jurassic Bark

Fry discovers the intact fossilized remains of his pet dog from his old life. The Professor reveals he can clone the animal.

9.6 Rate

S3.E10 The Luck of the Fryrish

Fry goes on a quest to recover his lucky 7-leaf clover after being convinced that his brother took over his identity after he was frozen.

9.5 Rate

S7.E26 Meanwhile

Professor Farnsworth invents a button that can take a person 10 seconds back in time, inadvertently causing major consequences.

9.4 Rate

See more episodes »

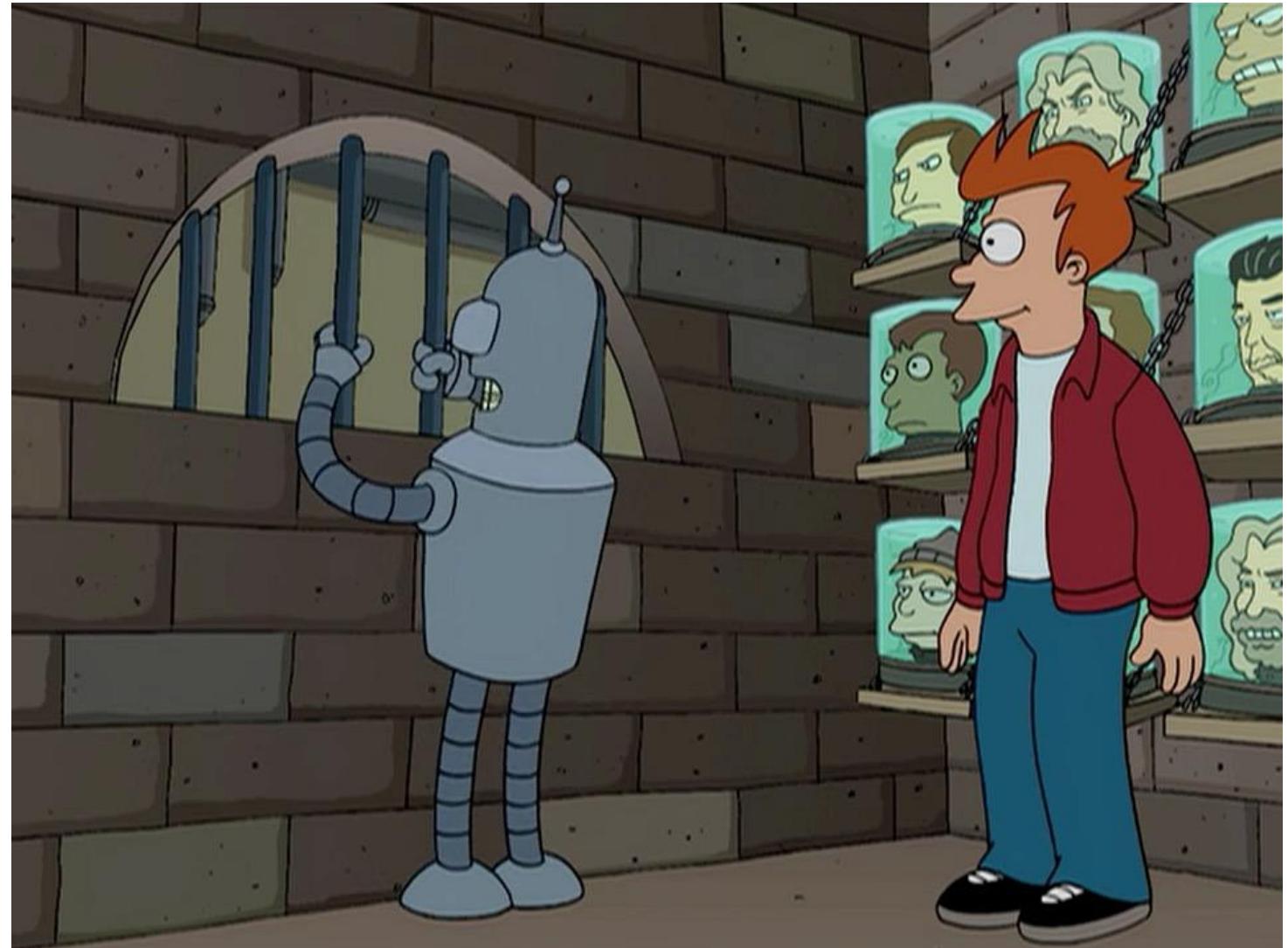
AI: Goals, Algorithm & Performance

Broad Intelligence

Capable of performing a wide variety of tasks without modification or retraining

Narrow Intelligence

Designed to solve a single problem



AI: Goals, Algorithm & Performance

Broad Intelligence

Capable of performing
a wide variety of tasks
without modification
or retraining

Narrow Intelligence

Designed to solve a single
problem

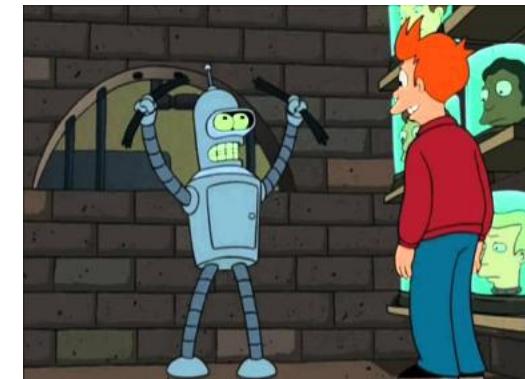


Recent success stories
of AI all belong to this
category!

AI: Goals, Algorithm & Performance

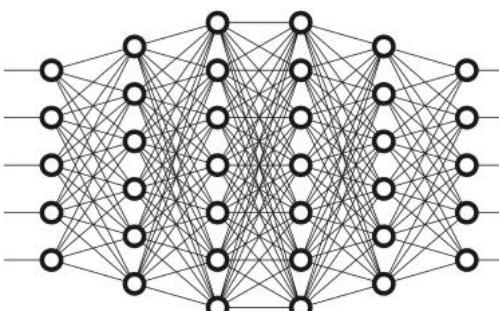
Broad Intelligence

Capable of performing a wide variety of tasks without modification or retraining



General AI Technique

Capable of approximating many different functions (can be used for a greater spectrum of purposes)



Superhuman Intelligence

At least 2 standard deviations better than average human task performance

Narrow Intelligence

Designed to solve a single problem

Human-level Intelligence

Indistinguishable from task performance of avg. humans



The Deep Learning Approach to AI

- Classical AI, rule-based, symbolic
 - knowledge is provided by humans
 - But intuitive knowledge (e.g. much of common sense) not communicable
 - machines only do inference
 - no strong learning, adaptation
 - insufficient handling of uncertainty
 - not grounded in low-level perception and action
- Deep learning tries to fix these problems
 - succeeded to a great extent
 - higher-level (conscious) cognition still seems out of reach (more on this later)

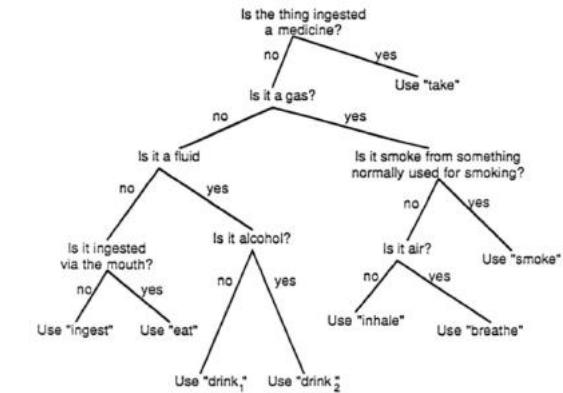


FIGURE 8.2 Discrimination net from BABEL

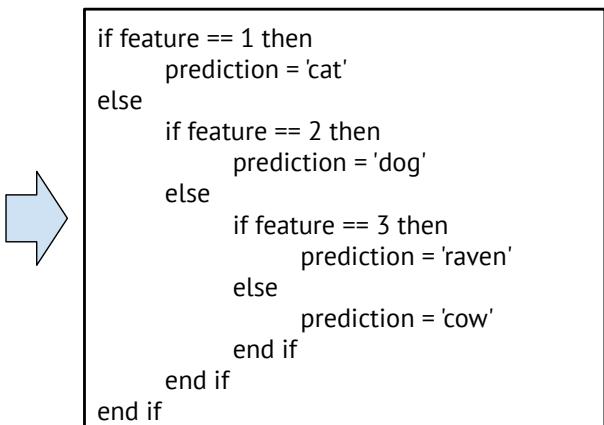
```
DISCRIMINATE(ITEM, NODE):
  IF NODE is a terminal node,
    THEN return the CONTENTS of NODE,
  ELSE DISCRIMINATE (
    ITEM,
    NEXT-NODE (TEST of NODE applied to ITEM,
    NODE))

NEXT-NODE(TEST-RESULT, NODE):
  IF TEST-RESULT,
    THEN return the node in the YES branch of NODE,
  ELSE return the node in the NO branch of NODE.
```

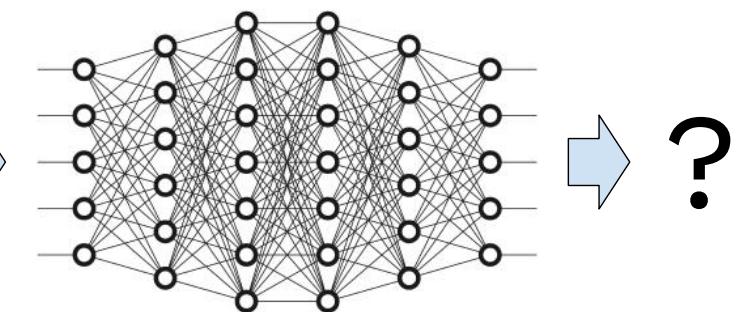
FIGURE 8.3 Algorithm for a general discrimination net

Change in Programming Paradigm

Traditional Programming:
program by writing lines of code



Deep Learning ("Software 2.0")
program by providing data



Poor performance

Enough data → Success!

The Purpose of Vision

"What does it mean, to see? The plain man's answer (and Aristotle's too) would be, **to know what is where by looking.** In other words, vision is the process of discovering from images what is present in the world, and where it is."

[Marr, 1982]

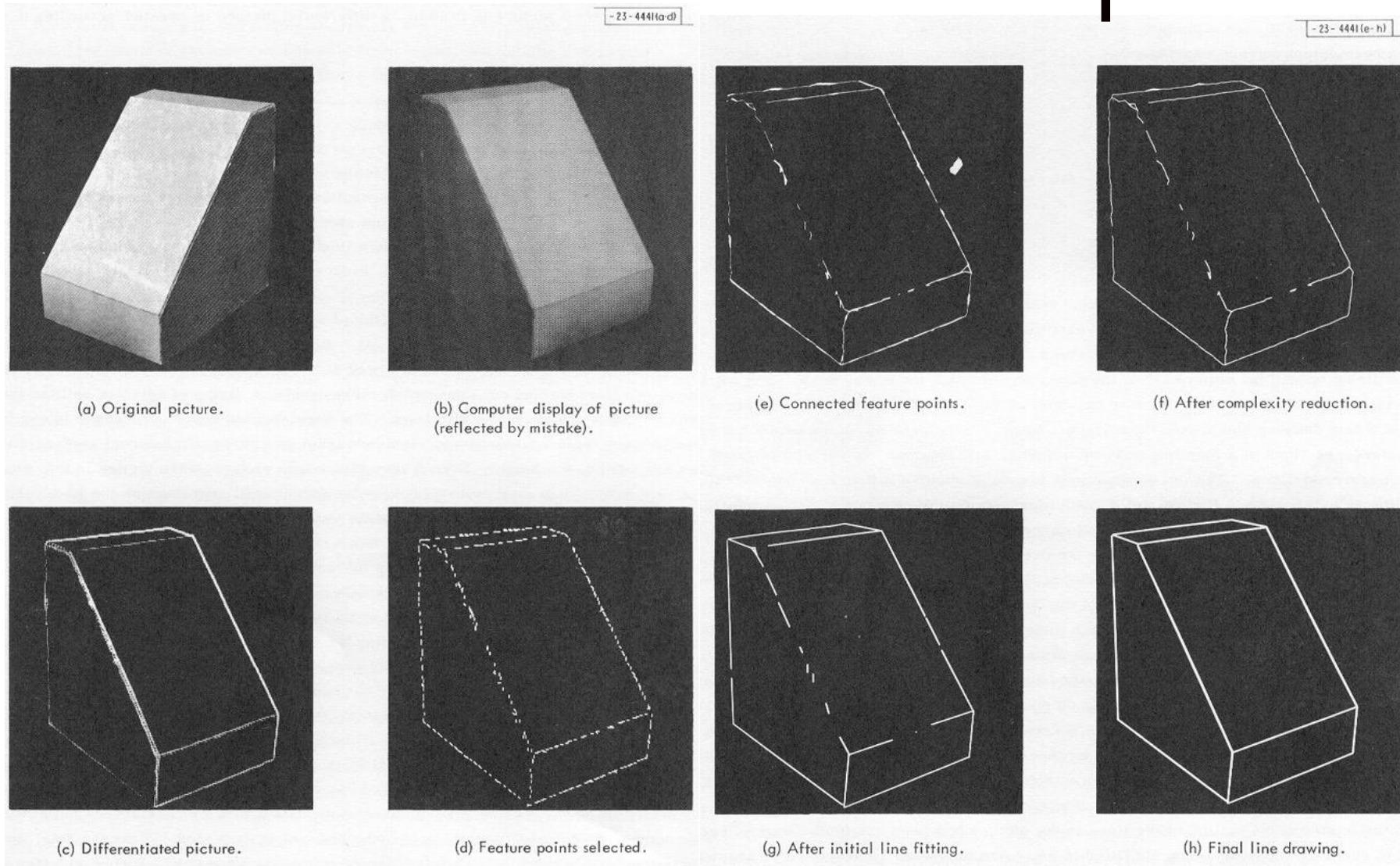
Image credit: The Sense of Sight (Annie Louisa Swynnerton, 1895)

Why Vision is Hard?



Image credit: Neşeli Günler (Arzu Film, 1978)

The First PhD Thesis on Computer Vision



- Machine perception of three-dimensional solids [Roberts 1963]

The Summer Vision Project

General goals:

FIGURE-GROUND.

divide a vidisector picture into regions such as likely objects, likely background areas and chaos

REGION DESCRIPTION.

analysis of shape and surface properties

OBJECT IDENTIFICATION.

name objects by matching them with a vocabulary of known objects

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

July 7, 1966

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

[Papert 1966]

Why does vision appear easy to humans?

- Our brains are specialized to do vision.
- ~50% of the cortex in a human brain is devoted for visual processing
(cf. motor control ~20-30%, language ~10-20%)

Visual perception*:

540,000,000 years of data

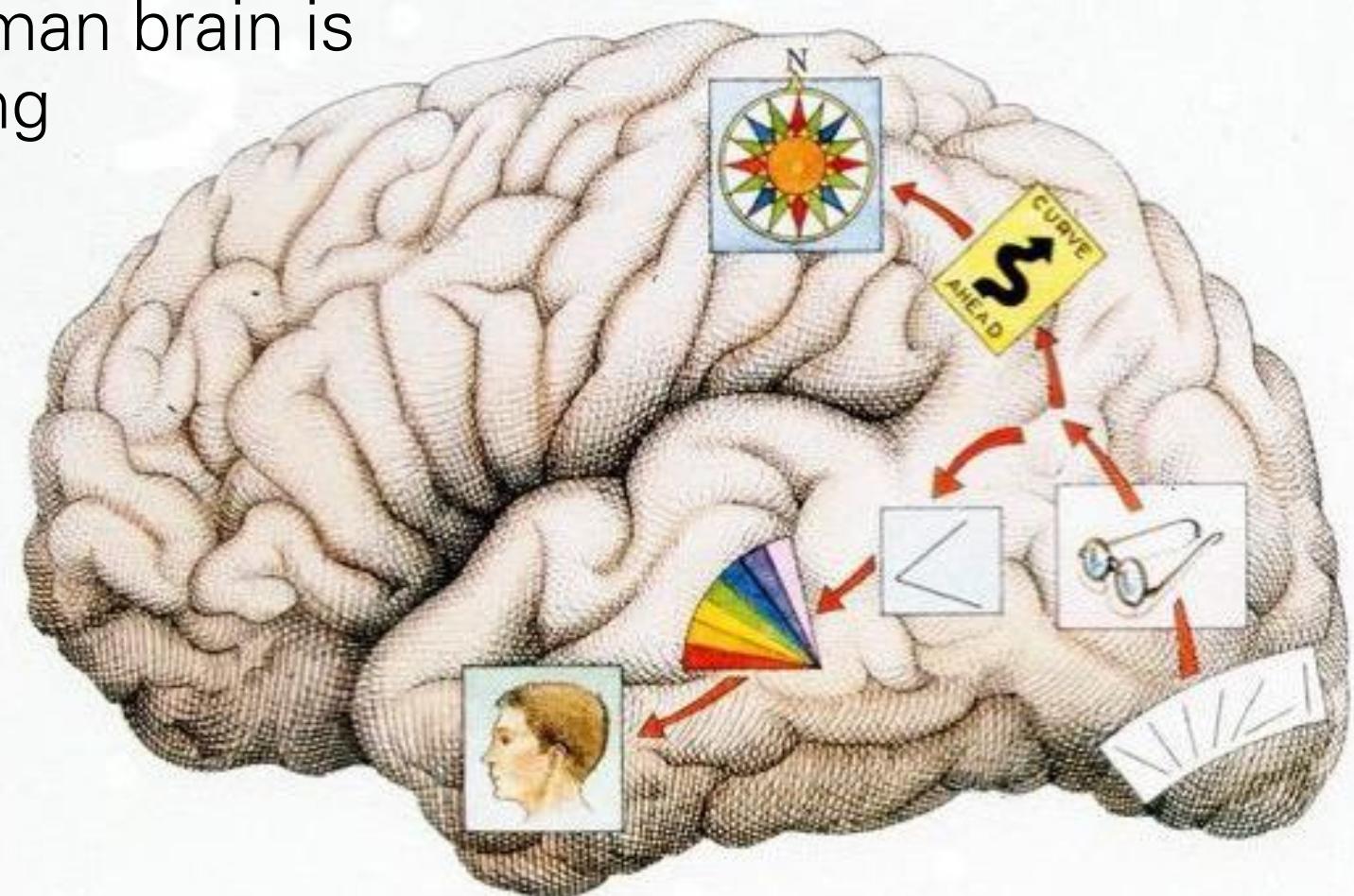
Bipedal movement:

230,000,000 years of data

Abstract thought:

100,000 years of data

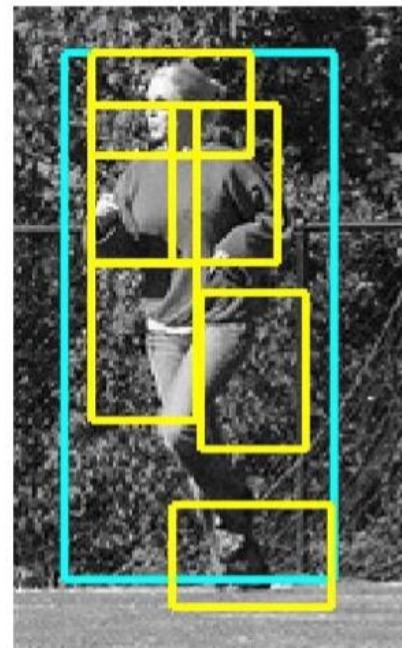
*Color vision



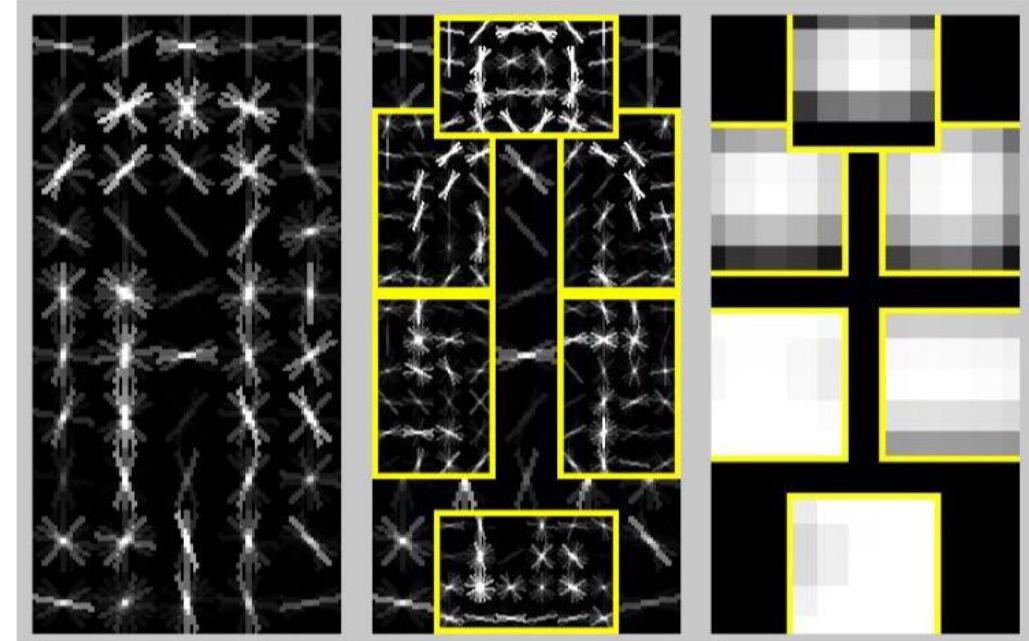
Shallow Methods in Computer Vision



Real-time Face Detection
[Viola and Jones 2001]



Deformable Part Model [Felzenszwalb et al. 2008]



The 2012 Revolution

ImageNet Challenge

- **IMAGENET** Large Scale Visual Recognition Challenge (ILSVRC)
 - **1.2M** training images with **1K** categories
 - Measure top-5 classification error



Output
Scale
T-shirt
Steel drum
Drumstick
Mud turtle



Output
Scale
T-shirt
Giant panda
Drumstick
Mud turtle



Image classification

Easiest classes



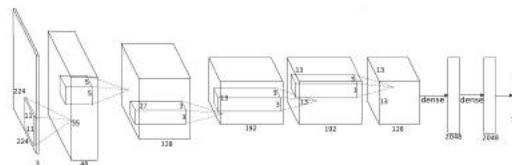
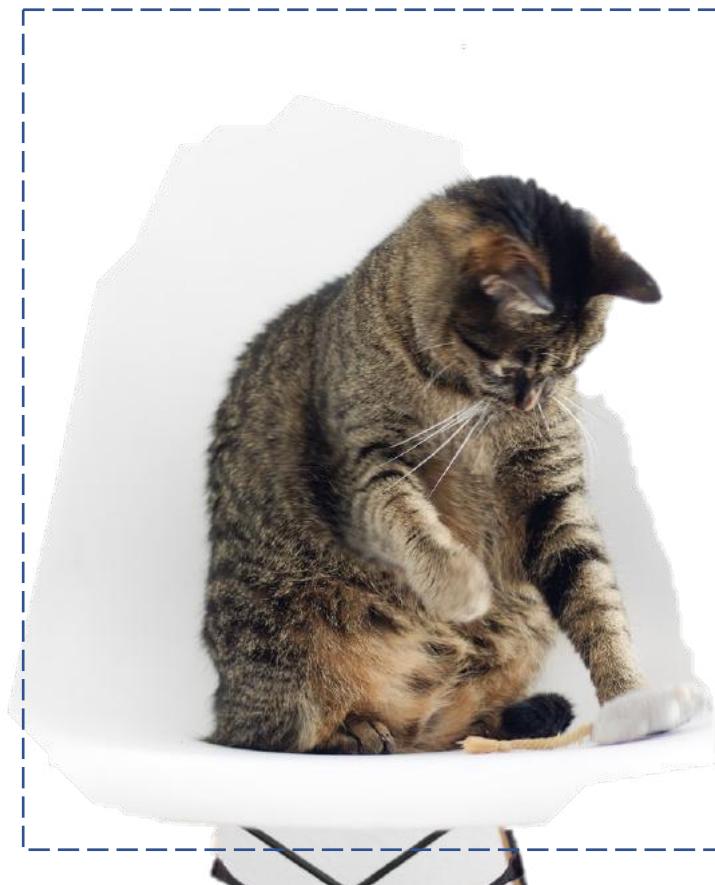
Hardest classes



2012: Deep Learning Revolution

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

- **1.2M** training images, **1K** categories
- Measure top-5 classification error



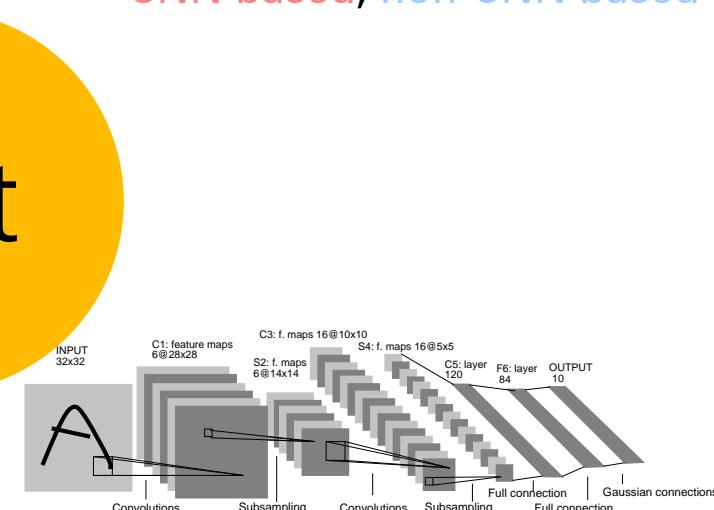
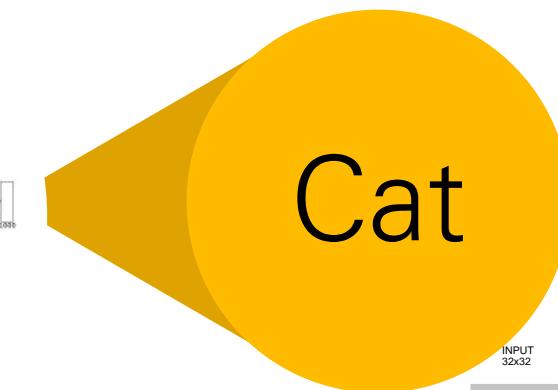
AlexNet

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86 (11): 2278–2324, 1998.

A. Krizhevsky, I. Sutskever, G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012

The success of AlexNet, a deep convolutional network (CNN)

- 7 hidden layers (not counting some max pooling layers)
- 60M parameters



2012 Teams	%Error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

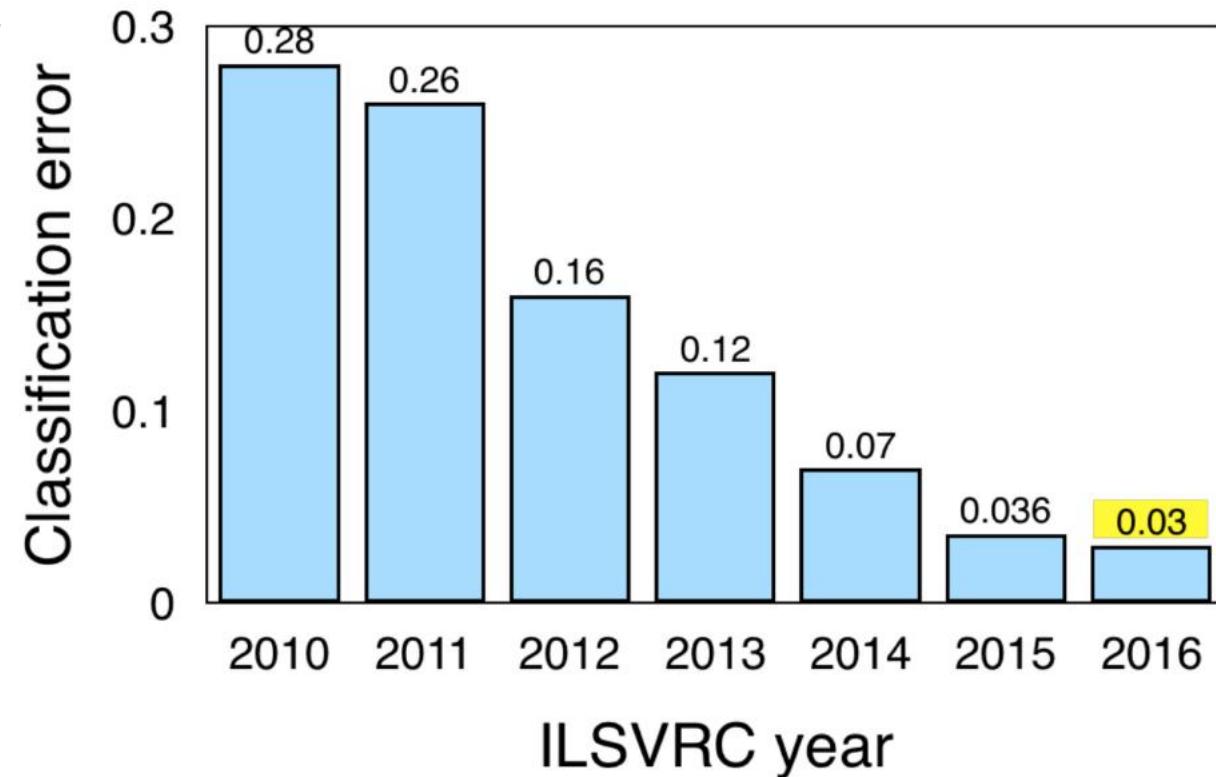
CNN based, non-CNN based

ImageNet Challenge 2010-2017

2010-11: hand-crafted computer vision pipelines

2012-2016: ConvNets

- 2012: AlexNet
 - major deep learning success
- 2013: ZFNet
 - improvements over AlexNet
- 2014
 - VGGNet: deeper, simpler
 - InceptionNet: deeper, faster
- 2015: ResNet
 - even deeper
- 2016
 - ensembled networks, results have saturated
- 2017: Squeeze-and-Excitation networks
 - end of the competition

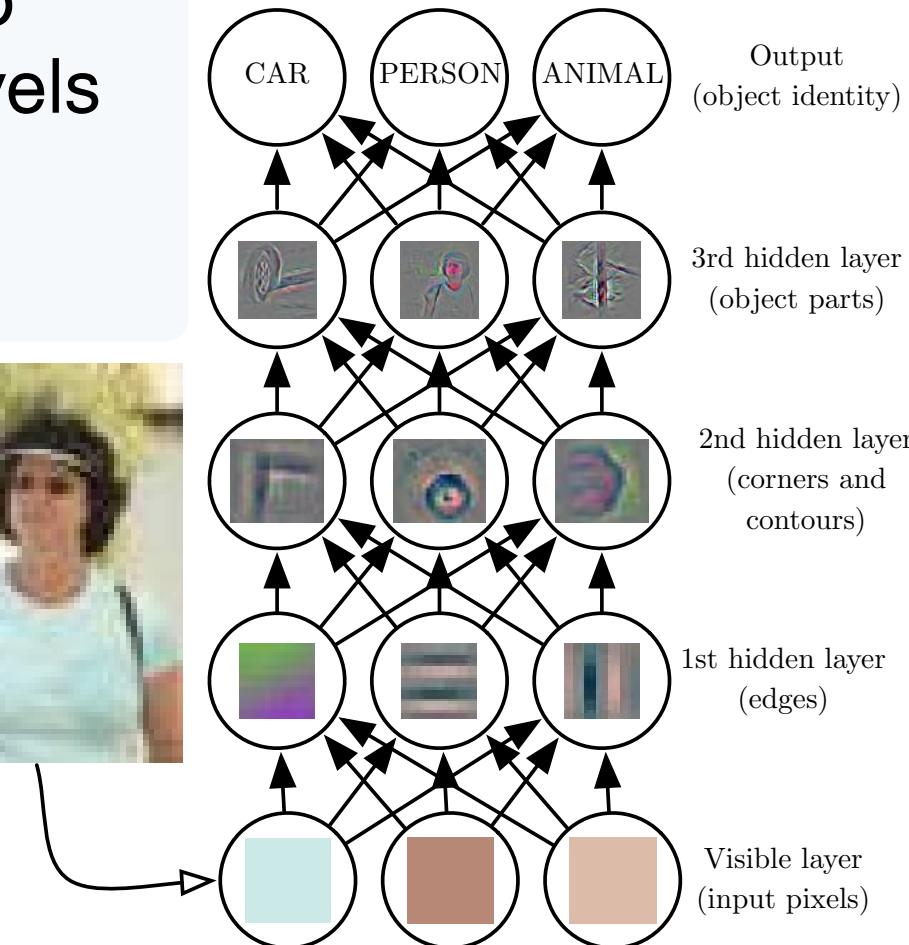
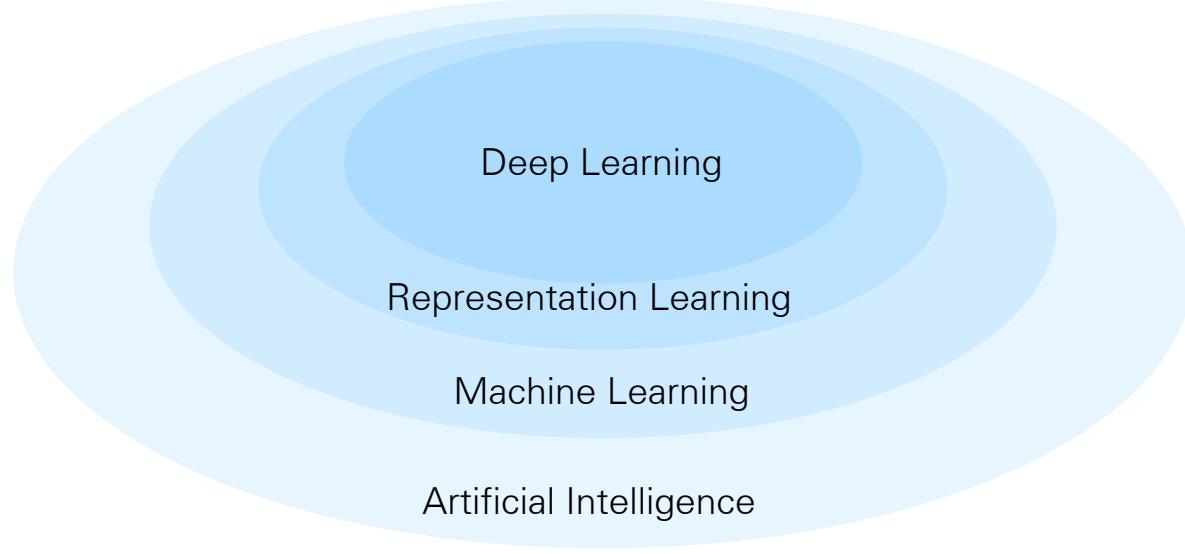


4 × more accurate in 4 years

Deep Learning is Representation Learning

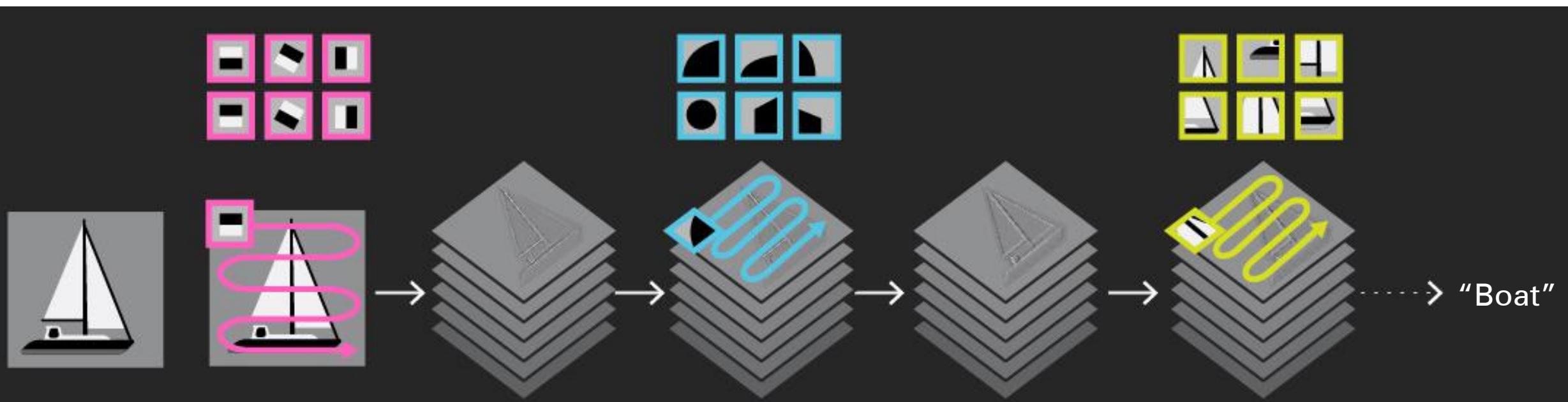
“Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.”

— Yann LeCun, Yoshua Bengio and Geoff Hinton

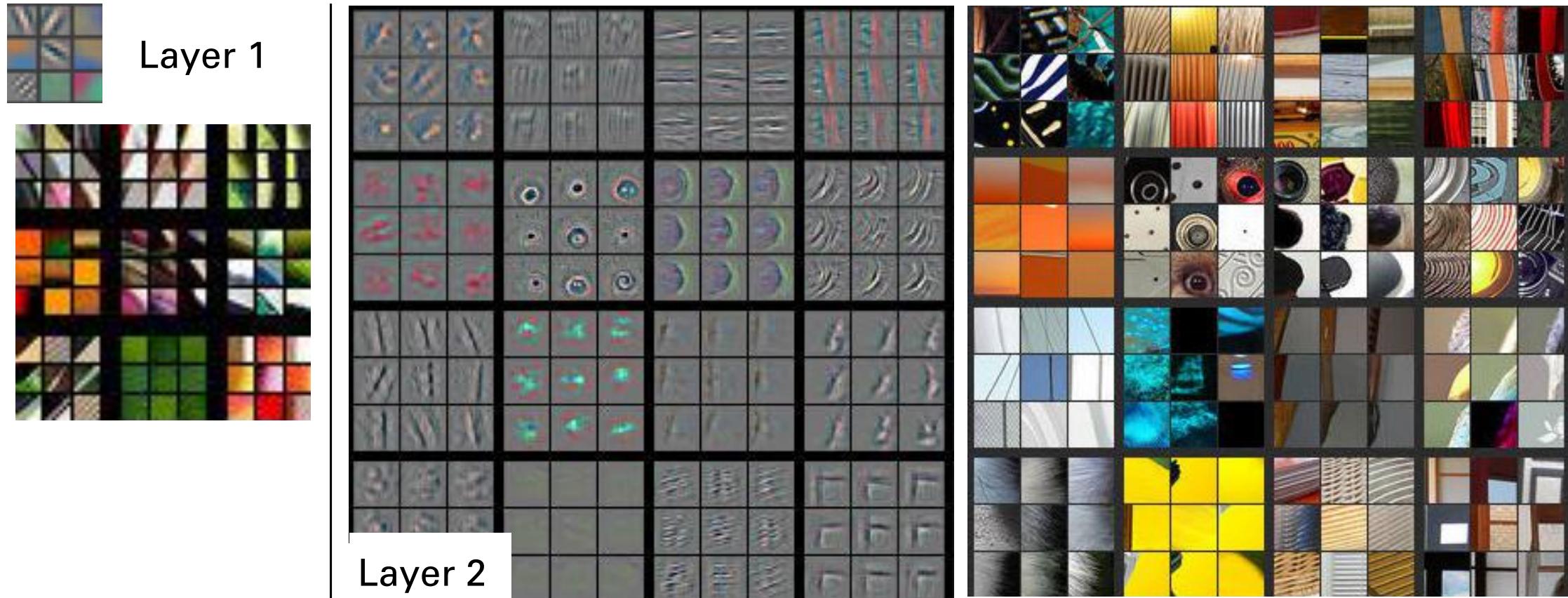


Convolutional Neural Networks in a Nutshell

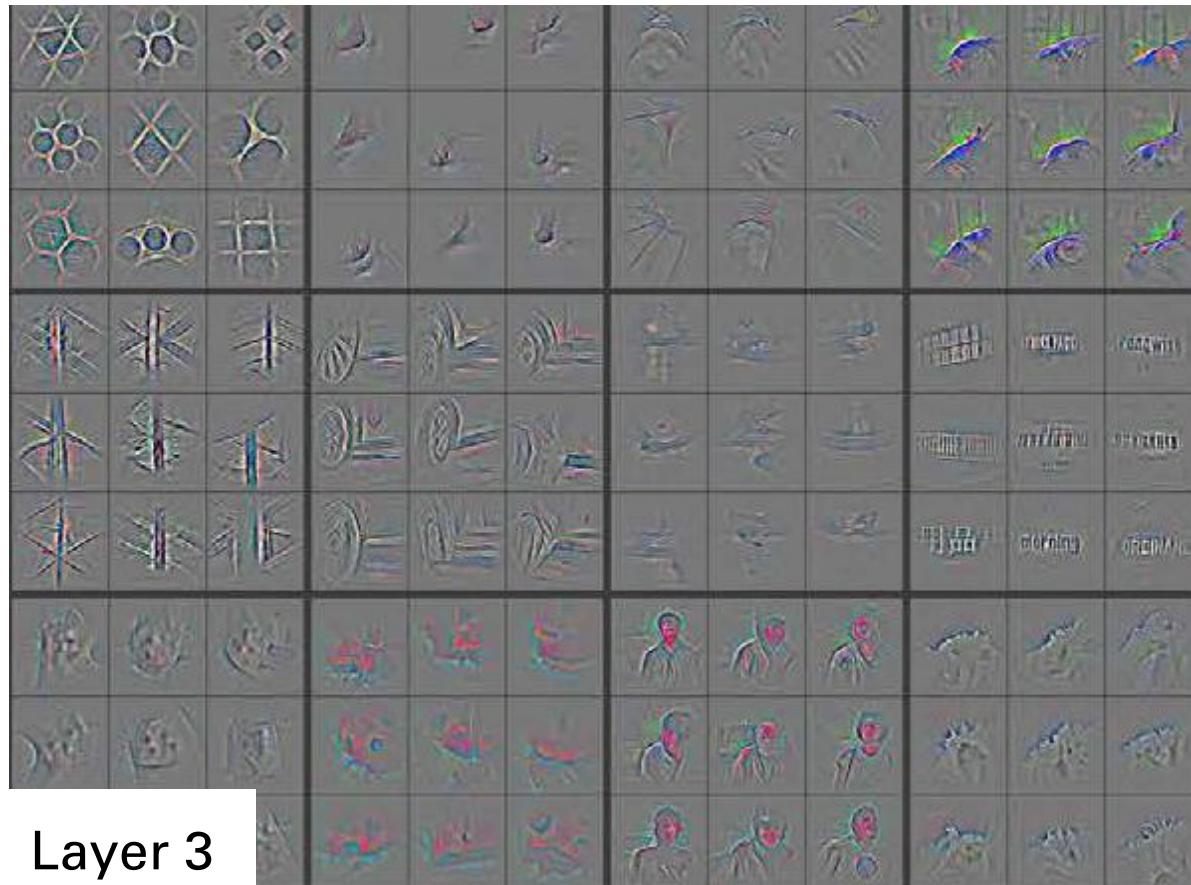
- A neural network model that consists of a sequence of local & translation invariant layers
 - Motivated from oriented cells in the visual cortex
 - Hierarchical feature learning



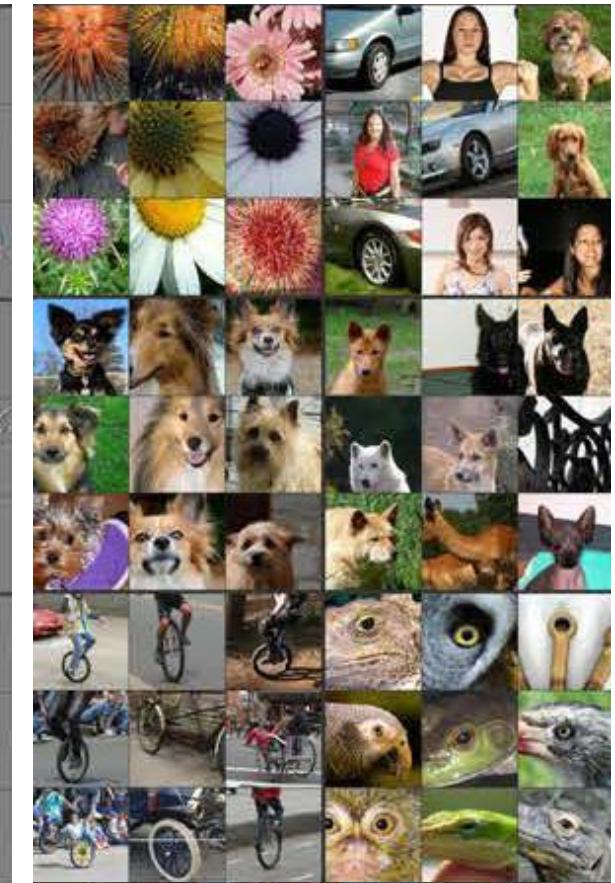
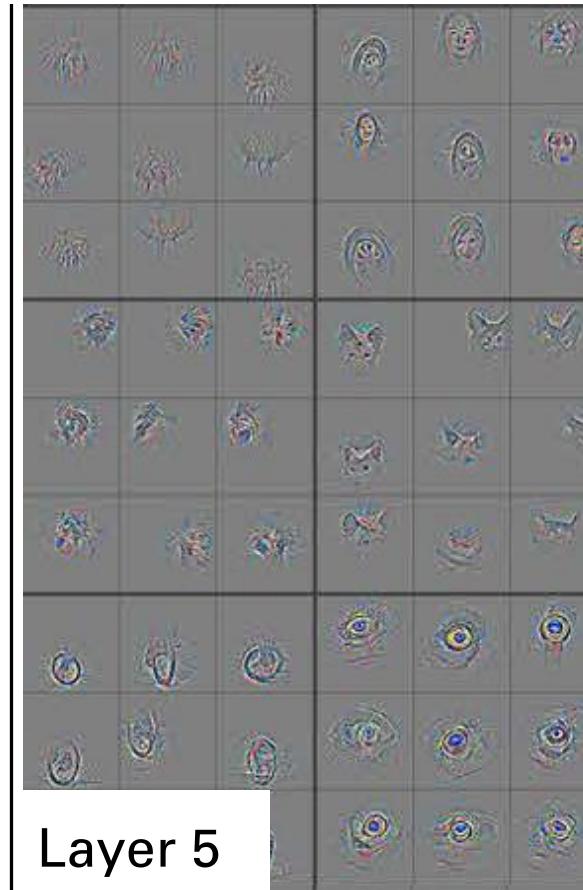
Visualizing arbitrary neurons along the way to the top...



Visualizing arbitrary neurons along the way to the top...

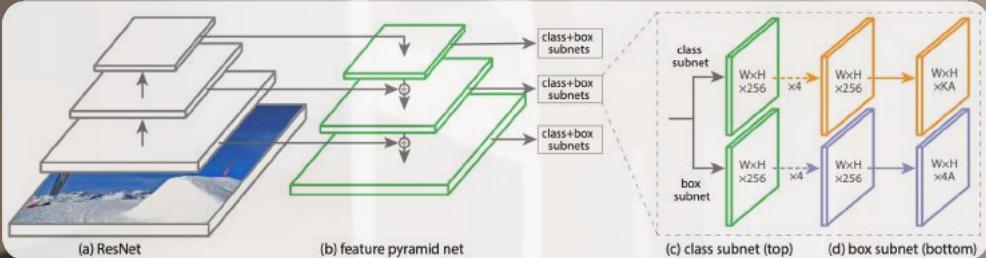


Visualizing arbitrary neurons along the way to the top...



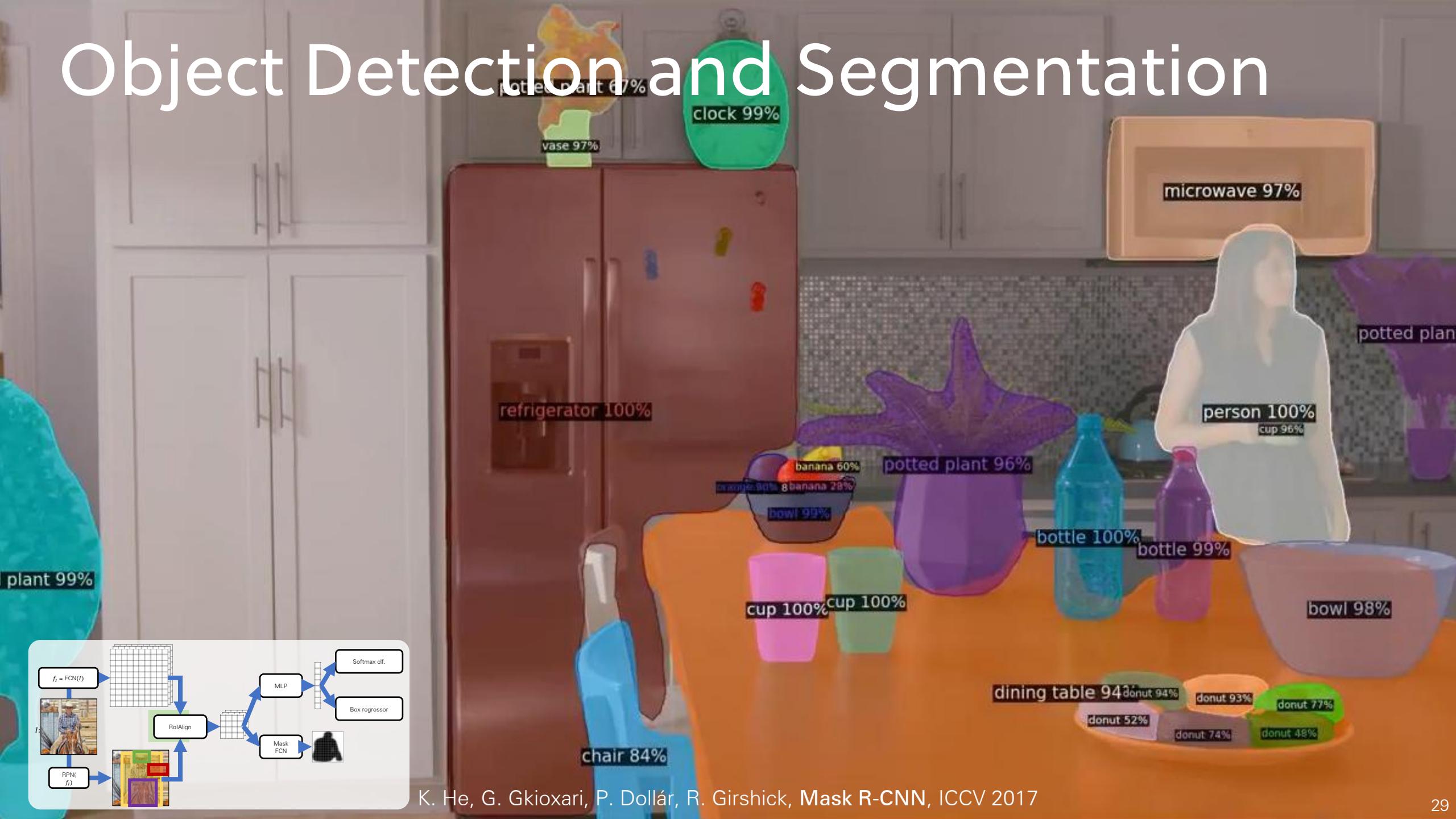
2012- Now

Object Detection and Segmentation

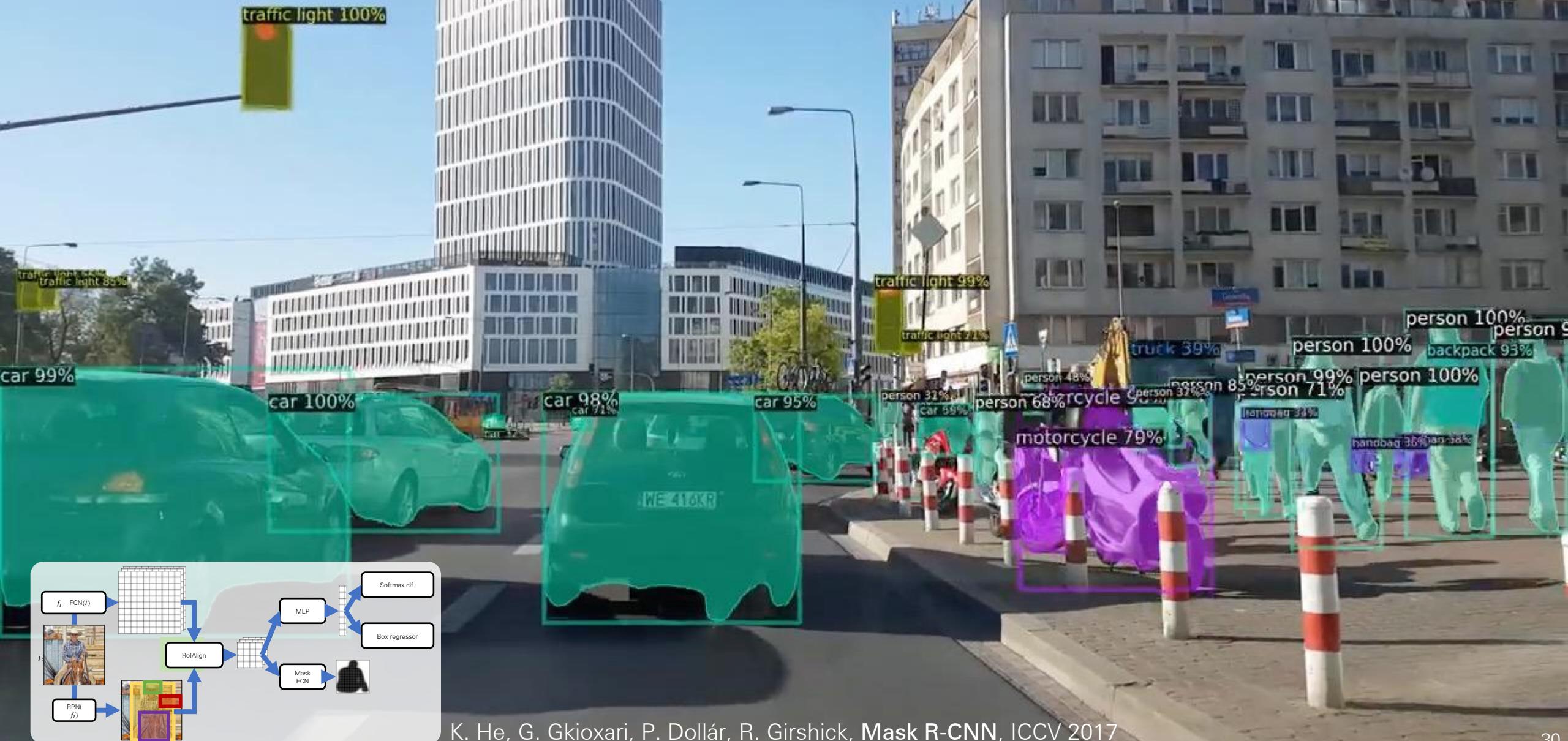


T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, **Focal Loss for Dense Object Detection**, ICCV 2017.

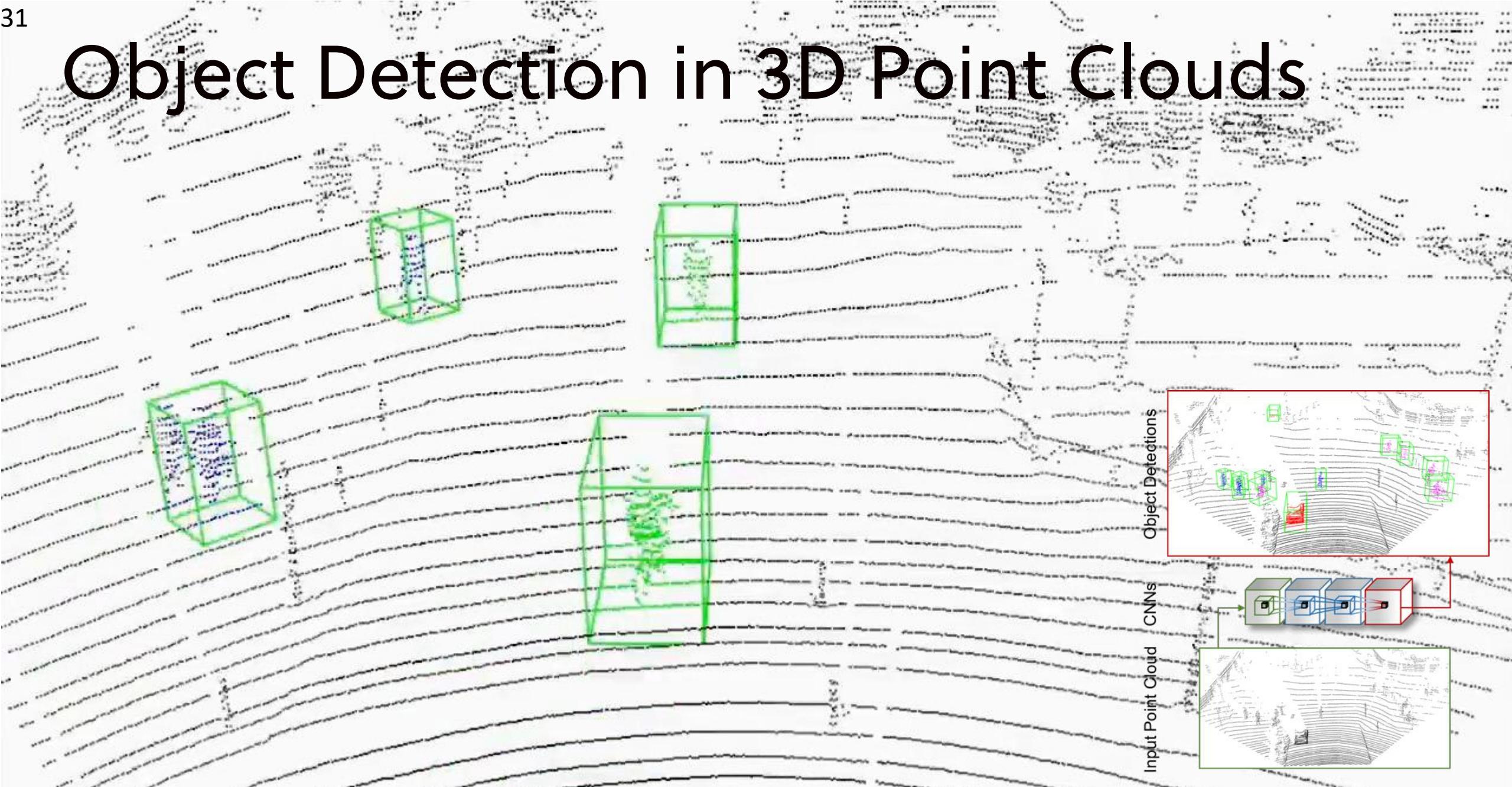
Object Detection and Segmentation



Object Detection and Segmentation

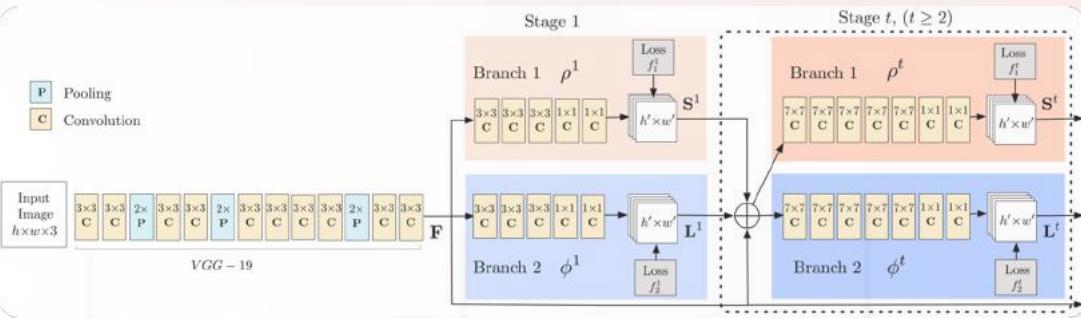


Object Detection in 3D Point Clouds

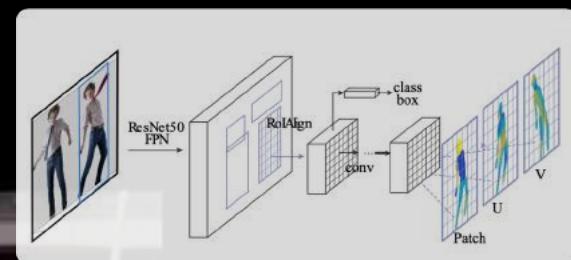


11.4 fps

Pose Estimation



Pose Estimation



We introduce a system that can associate every image pixel with human body surface coordinates.

Pose Estimation

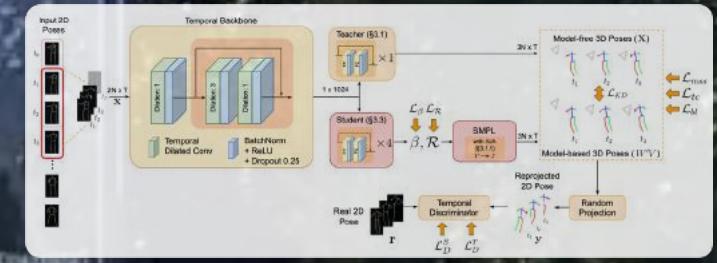


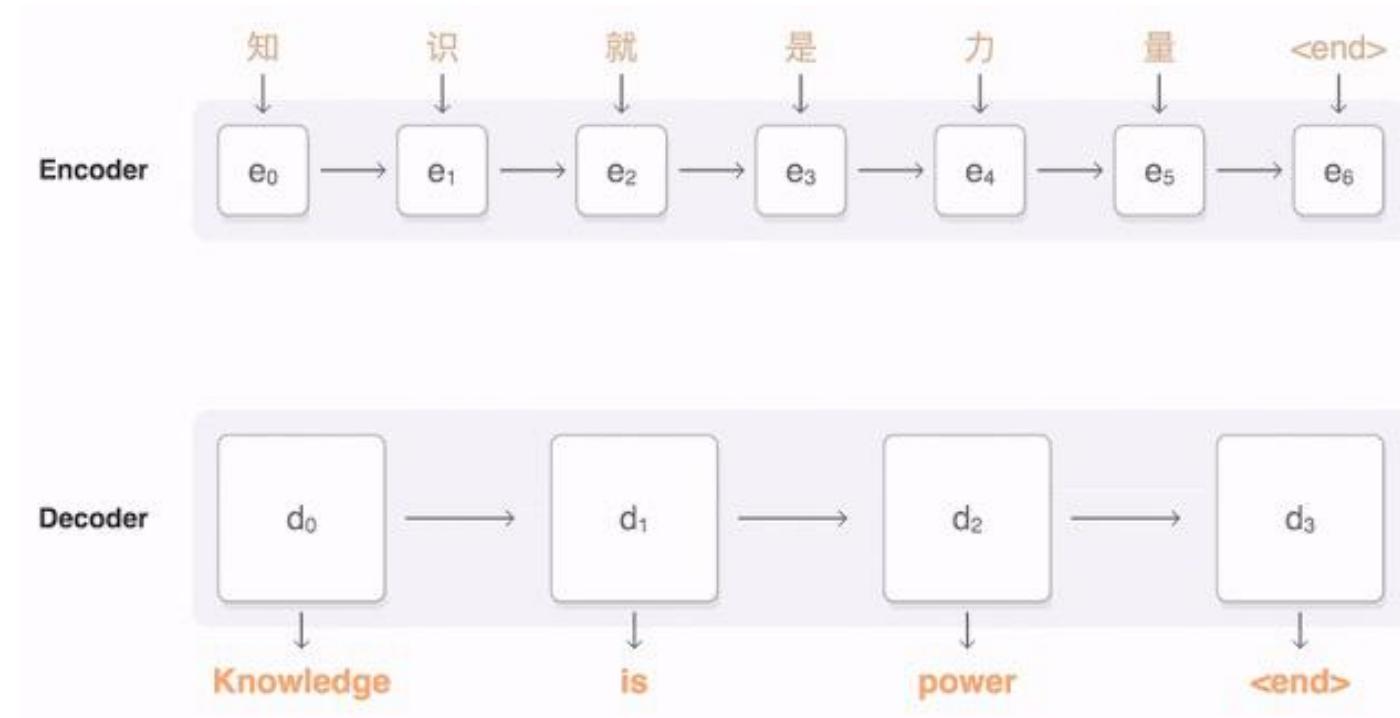
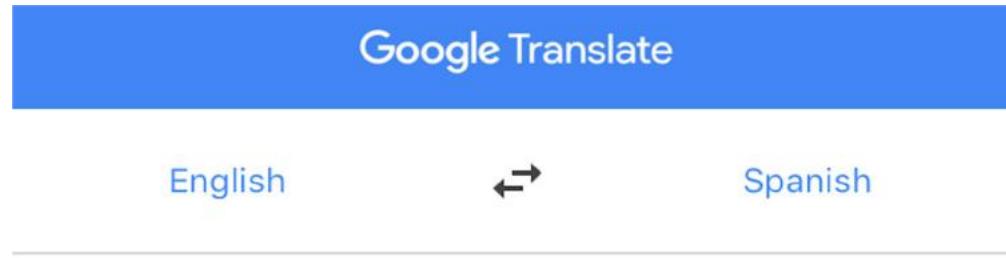
Photo Style Transfer



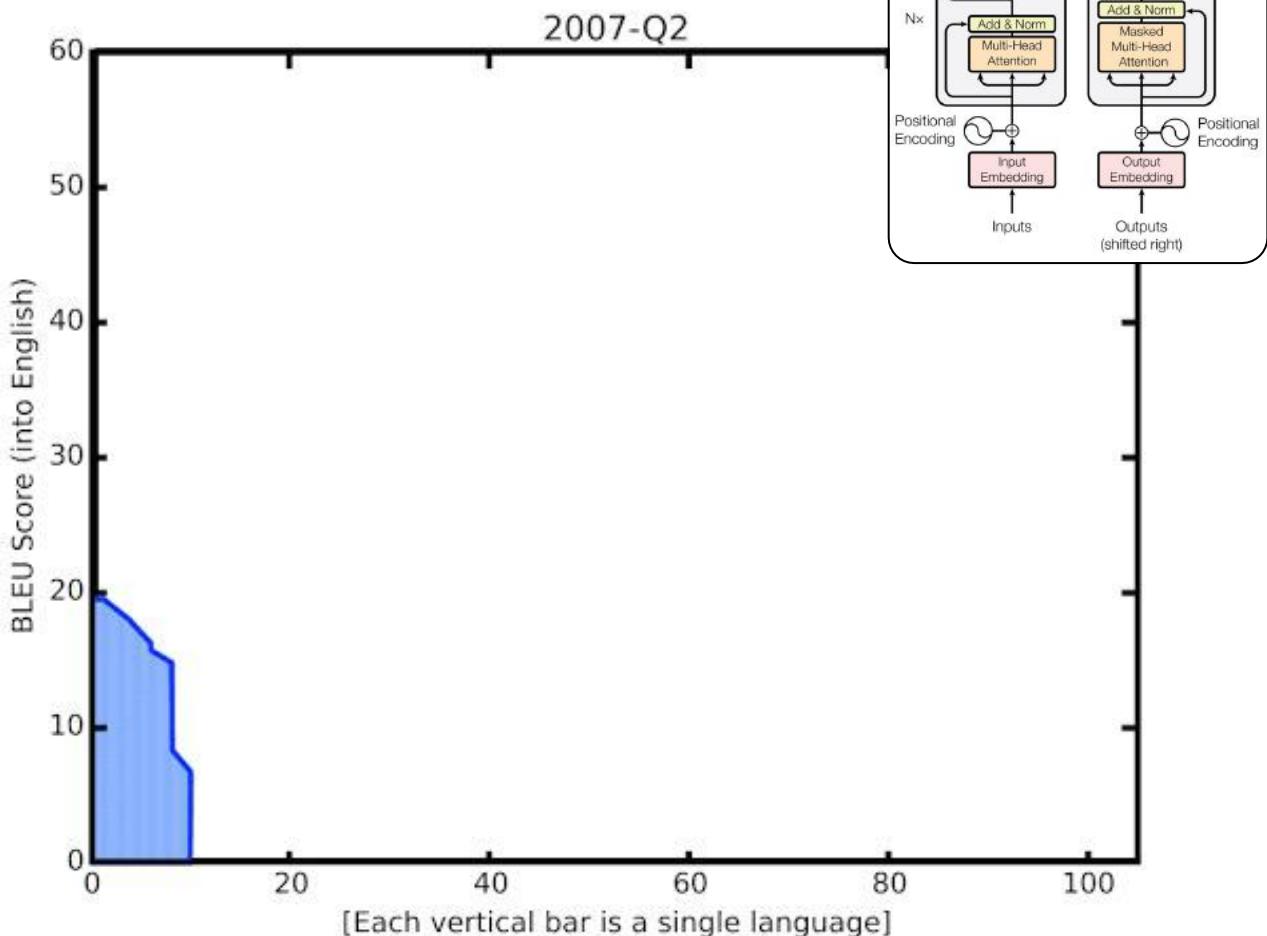
Photo Style Transfer



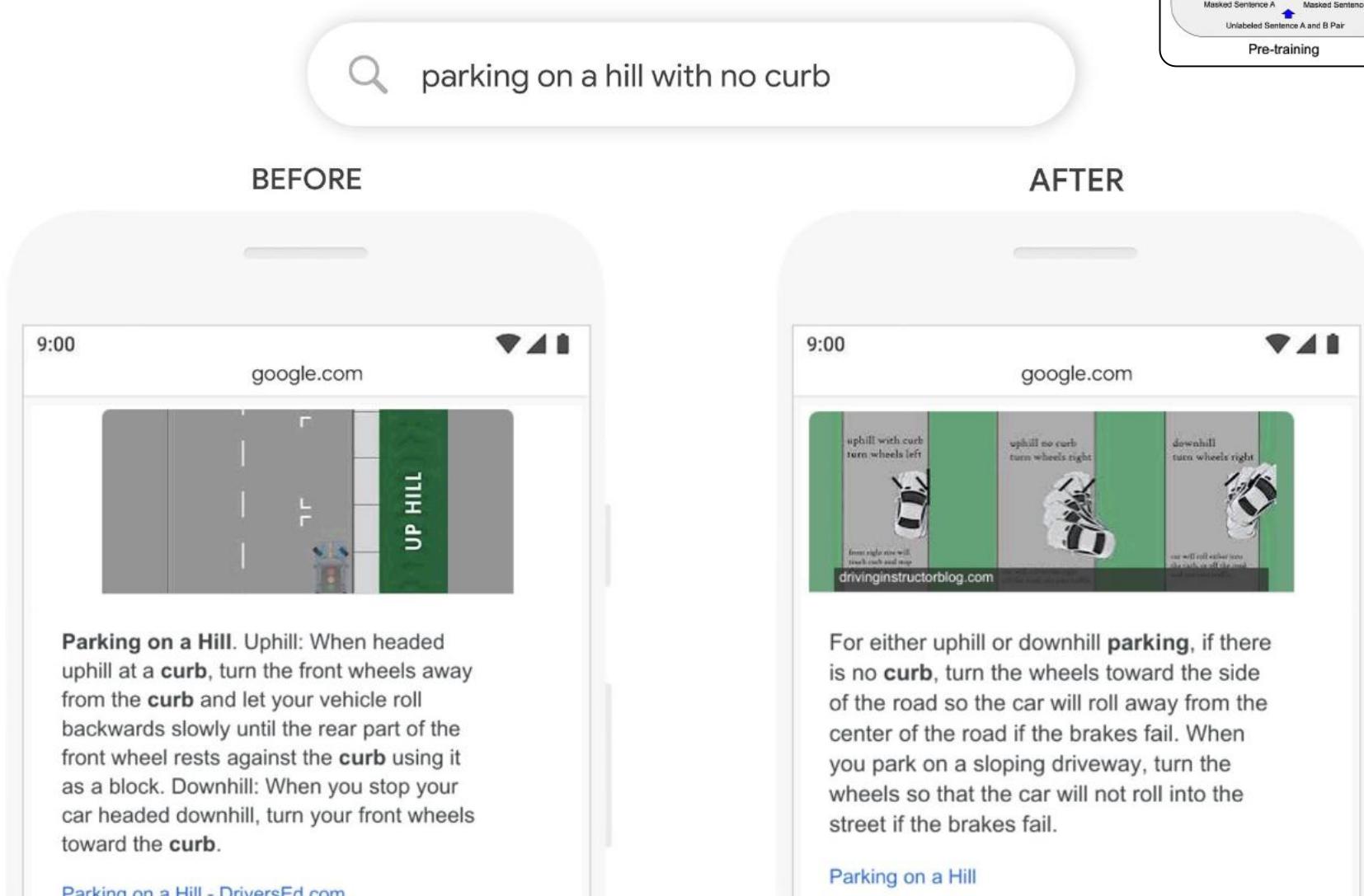
Machine Translation



Machine Translation



Internet Search



Language Modeling

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. [Learn more](#) below.

 Follow @AdamDanielKing

for more neat neural networks.

Custom prompt

Coronavirus outbreak

GENERATE ANOTHER

Completion

Coronavirus outbreak: report

China has dispatched an expert team of epidemiologists, virologists and pathologists to Britain to advise its Health Protection Agency on how to control the spread of the viral respiratory illness.

The experts will arrive in London in a few days, State Council vice-president Li Yuanchao said Thursday.

The official Xinhua News Agency said the experts will "advise China's Ministry of Health on whether the outbreak is considered to be connected to a disease outbreak in the US."

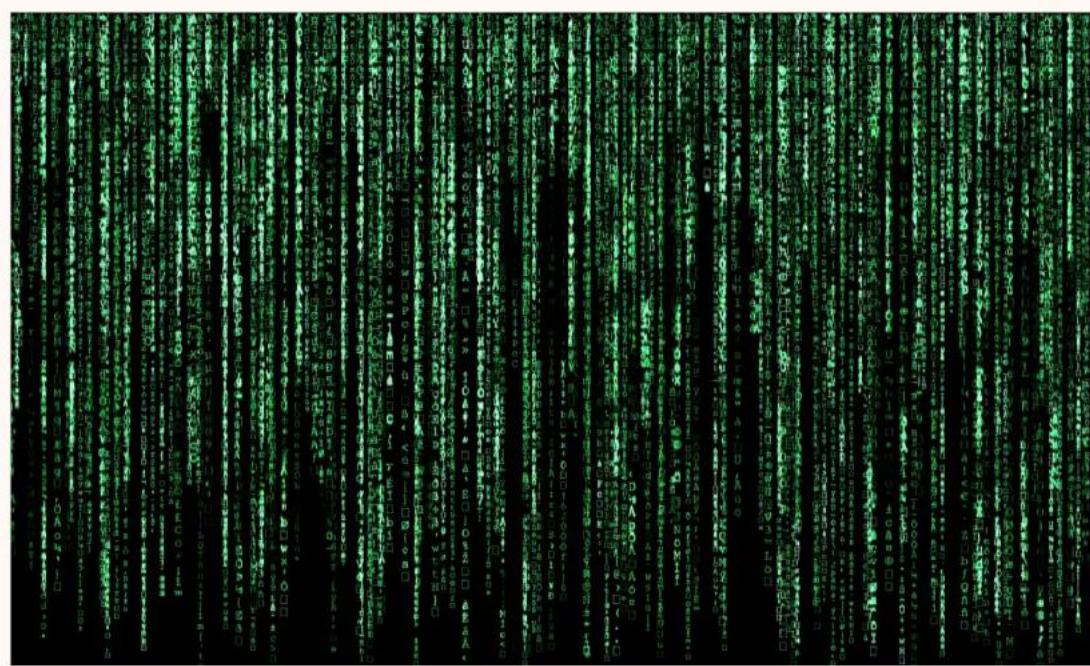
The World Health Organization and China's National Health and Family Planning Commission say the first cases of human infection with the coronavirus were reported last week in China, but that the number has risen to thousands.

Language Modeling

- **GPT-3:** I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

the guardian

Tue 8 Sep 2020 09.45



▲ ‘We are not plotting to take over the human populace.’ Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!



Empathy machines:
what will happen
when robots learn
to write film scripts?

Read more

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

Question Answering

The first full-scale working railway steam locomotive was built by Richard Trevithick in the United Kingdom and, on 21 February 1804, the world's first railway journey took place as Trevithick's unnamed steam locomotive hauled a train along the tramway from the Pen-y-darren ironworks, near Merthyr Tydfil to Abercynon in south Wales. The design incorporated a number of important innovations that included using high-pressure steam which reduced the weight of the engine and increased its efficiency. Trevithick visited the Newcastle area later in 1804 and the colliery railways in north-east England became the leading centre for experimentation and development of steam locomotives.

In what country was a full-scale working railway steam locomotive first invented?

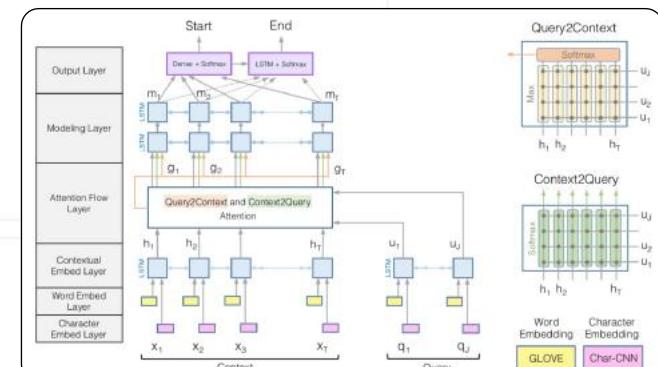
Ground Truth Answers: United Kingdom United Kingdom United Kingdom

Prediction: United Kingdom

On what date did the first railway trip in the world occur?

Ground Truth Answers: 21 February 1804 21 February 1804 21 February 1804

Prediction: 21 February 1804



The screenshot shows a web browser window with the URL super.gluebenchmark.com in the address bar. The page features a large red logo with a stylized 'G' and the text "SuperGLUE". Below the logo are logos for NYU, ML², UW NLP, Facebook Artificial Intelligence, DeepMind, and Samsung Research. A text block discusses the performance improvements of language models and the creation of SuperGLUE. At the bottom, there are links for PAPER, STARTER CODE, GROUP, and DIAGNOSTICS.

SuperGLUE



facebook Artificial Intelligence



DeepMind

SAMSUNG Research

In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. The GLUE benchmark, introduced one year ago, offered a single-number metric that summarizes progress on a diverse set of such tasks, but performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research.

We take into account the lessons learnt from original GLUE benchmark and present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, improved resources, and a new public leaderboard.

PAPER STARTER CODE GROUP DIAGNOSTICS

super.gluebenchmark.com

SuperGLUE GLUE

Paper </> Code Tasks Leaderboard i FAQ Diagnostics Submit Login

Leaderboard Version: 2.0

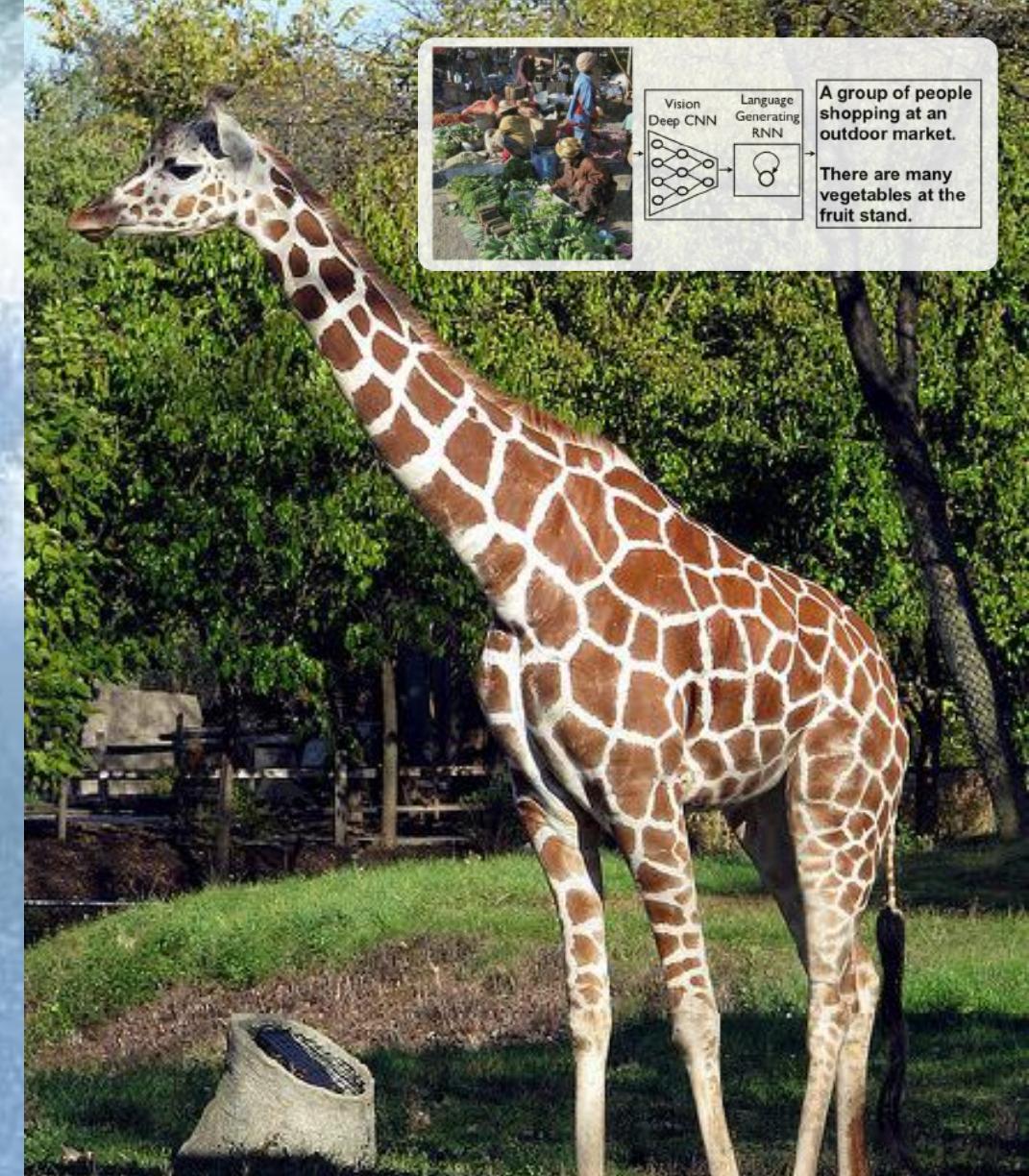
Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WIC	WSC	AX-b	AX-g
+	1 Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	2 DeBERTa Team - Microsoft	DeBERTa / TuringNLVRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
3	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	4 T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
+	5 Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	6 Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	7 Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5
+	8 Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
9	Zhuify Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
10	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	11 Anuar Sharafudinov	AI Labs Team, Transformers		82.6	88.1	91.6/94.8	86.8	85.1/54.7	82.8/79.8	88.9	74.1	78.8	100.0	100.0/100.0
12	Rakesh Radhakrishnan Menon	ADAPET (ALBERT) - few-shot		80.0	85.4	76.2/35.7	86.1/85.5	75.0	53.5	85.6	-0.4	100.0/50.0		
13	Timo Schick	iBERT (ALBERT) - Few Shot (22 Examples)		75.4	81.2	70.0/88.9	80.8	74.1/81.7	85.0/85.1	70.9	49.2	88.4	26.2	87.9/87.0

Click on a submission to see more information

Image Captioning



A man riding a wave on a surfboard in the water.



A giraffe standing in the grass next to a tree.

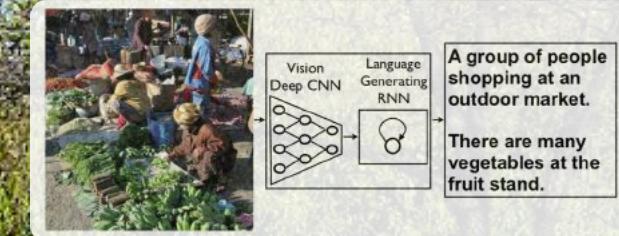
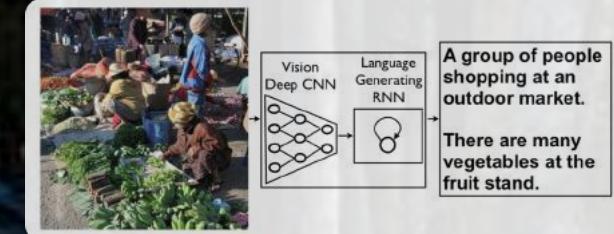


Image Captioning



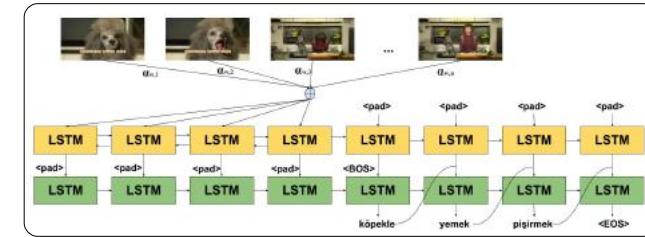
Yarış pistinde virajı almakta olan bir yarış arabası





Günbatımında bir dağa tırmanmakta olan bir adam

Video Captioning

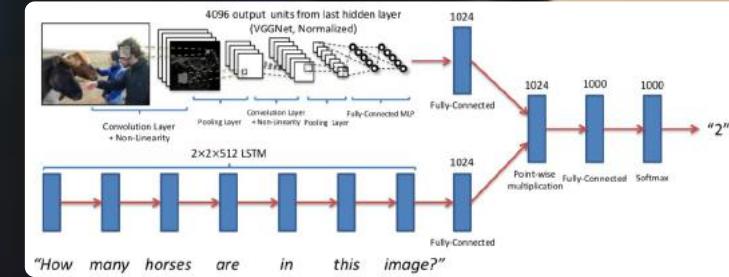


Bir adam bir gitar çalıyor



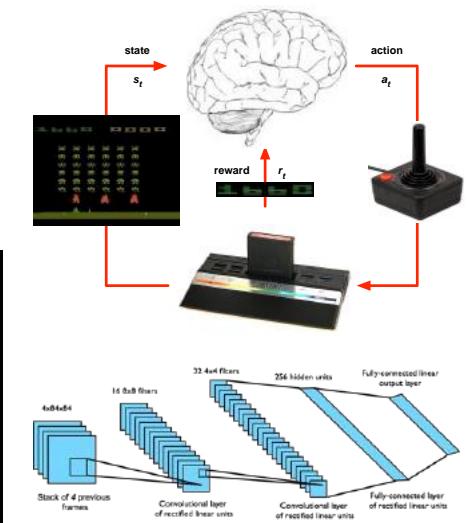
Bir kadın bir bıçakla sebze dilimliyor

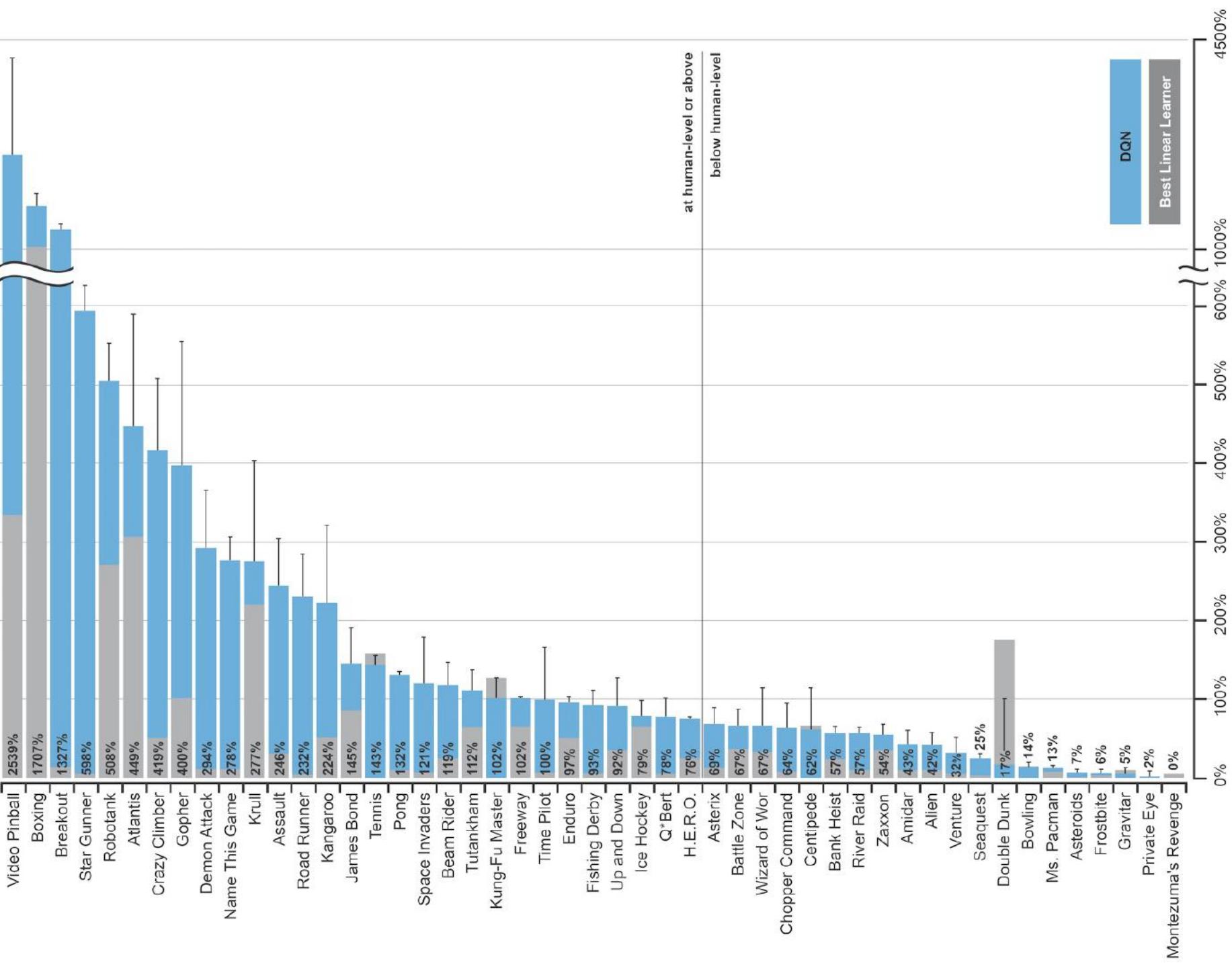
Visual Question Answering



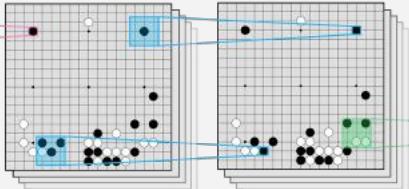
Question: What is the girl reaching into?

Answer: apples

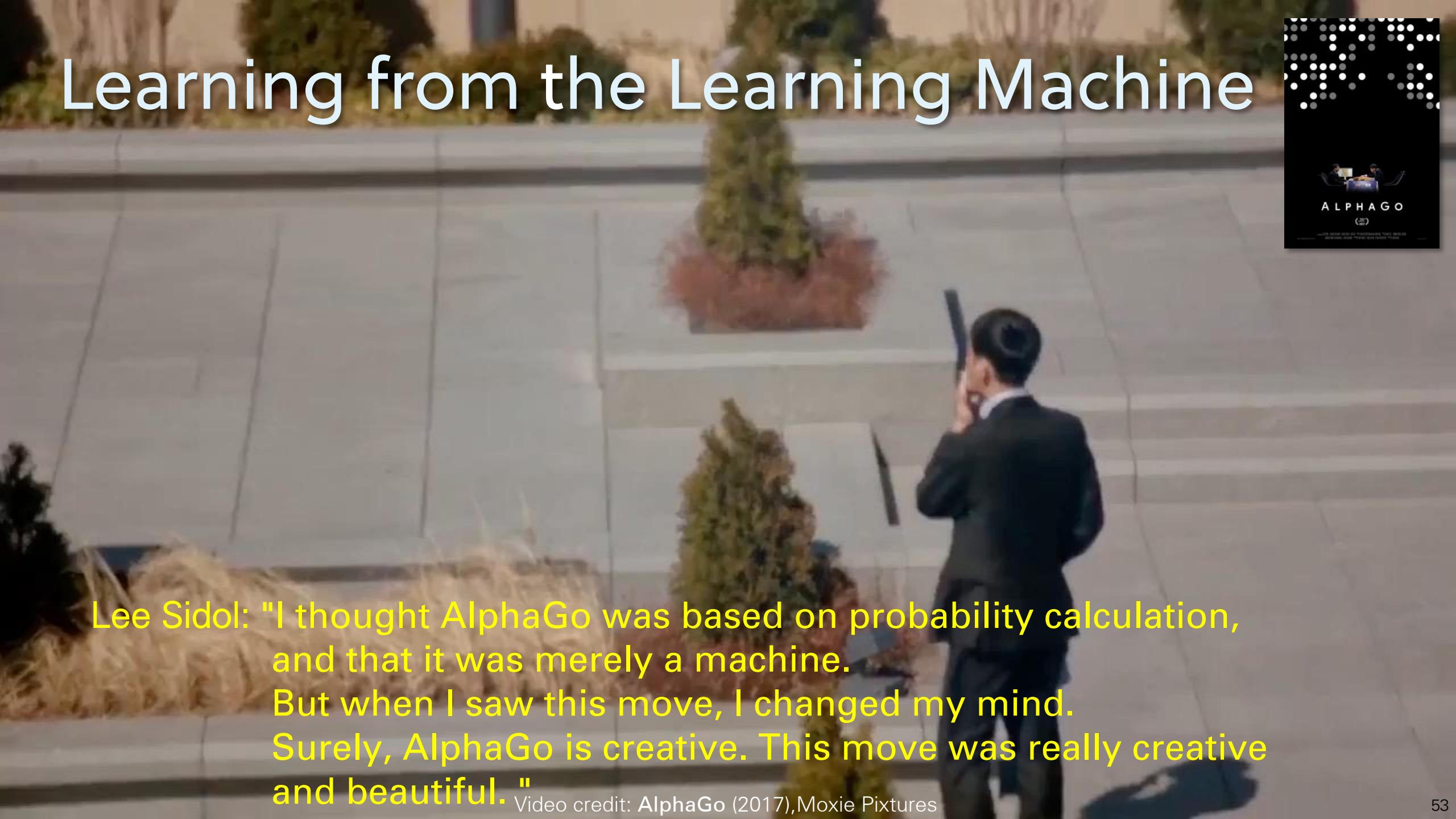




AlphaGo vs. Lee Sidol

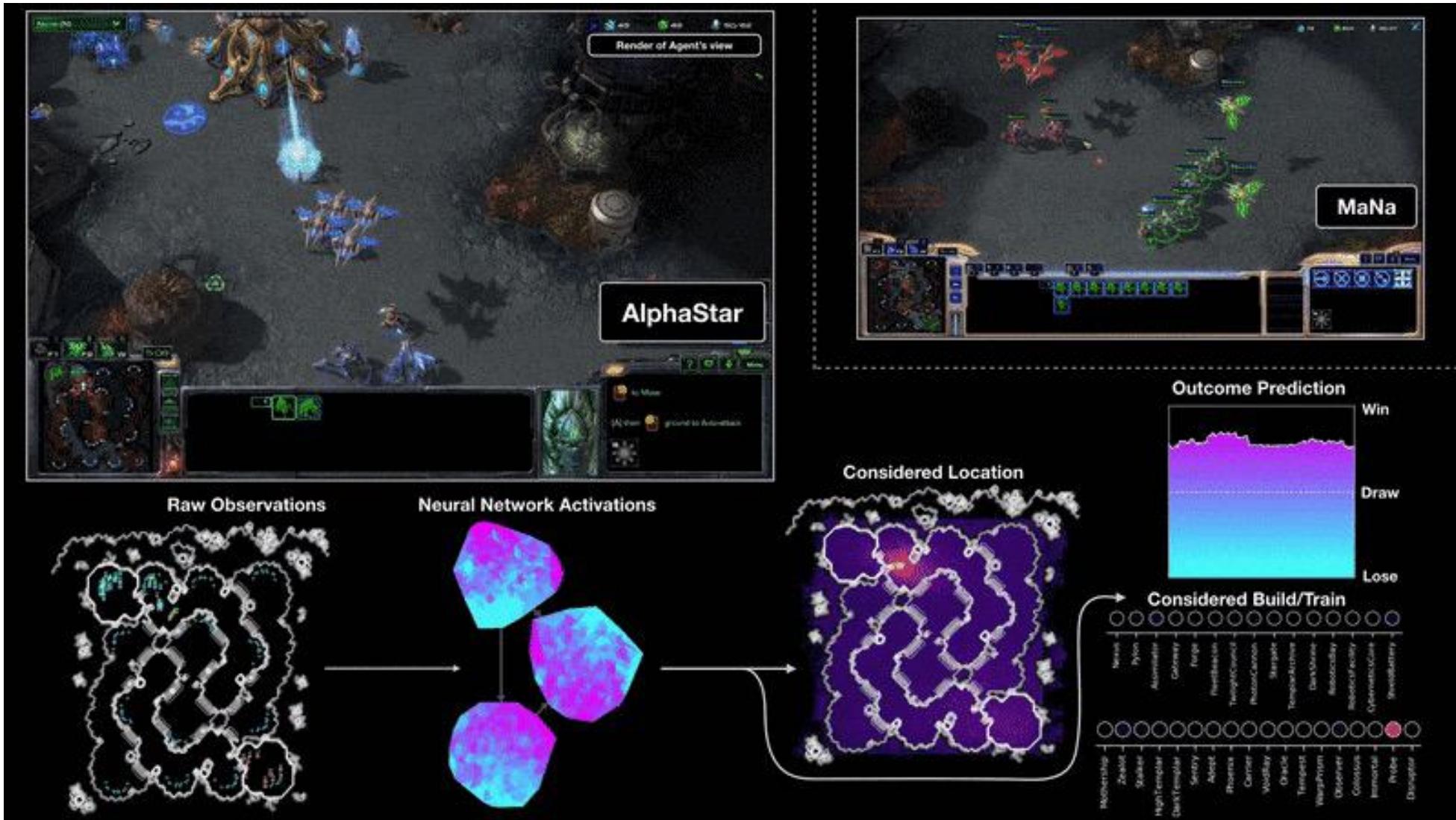


Learning from the Learning Machine



Lee Sidol: "I thought AlphaGo was based on probability calculation, and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative. This move was really creative and beautiful. "

Video credit: AlphaGo (2017), Moxie Pictures

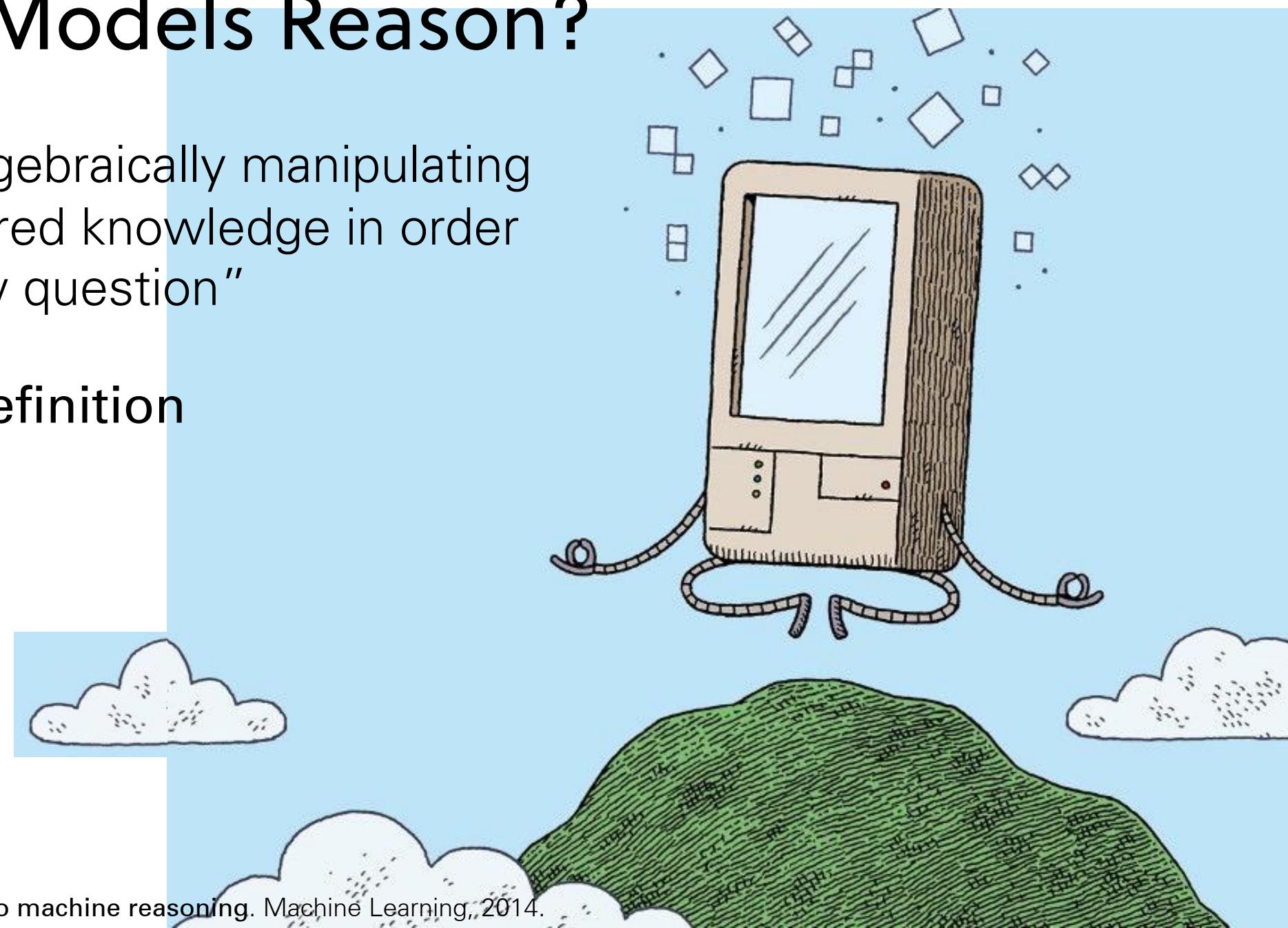


Act II: A Deeper Look



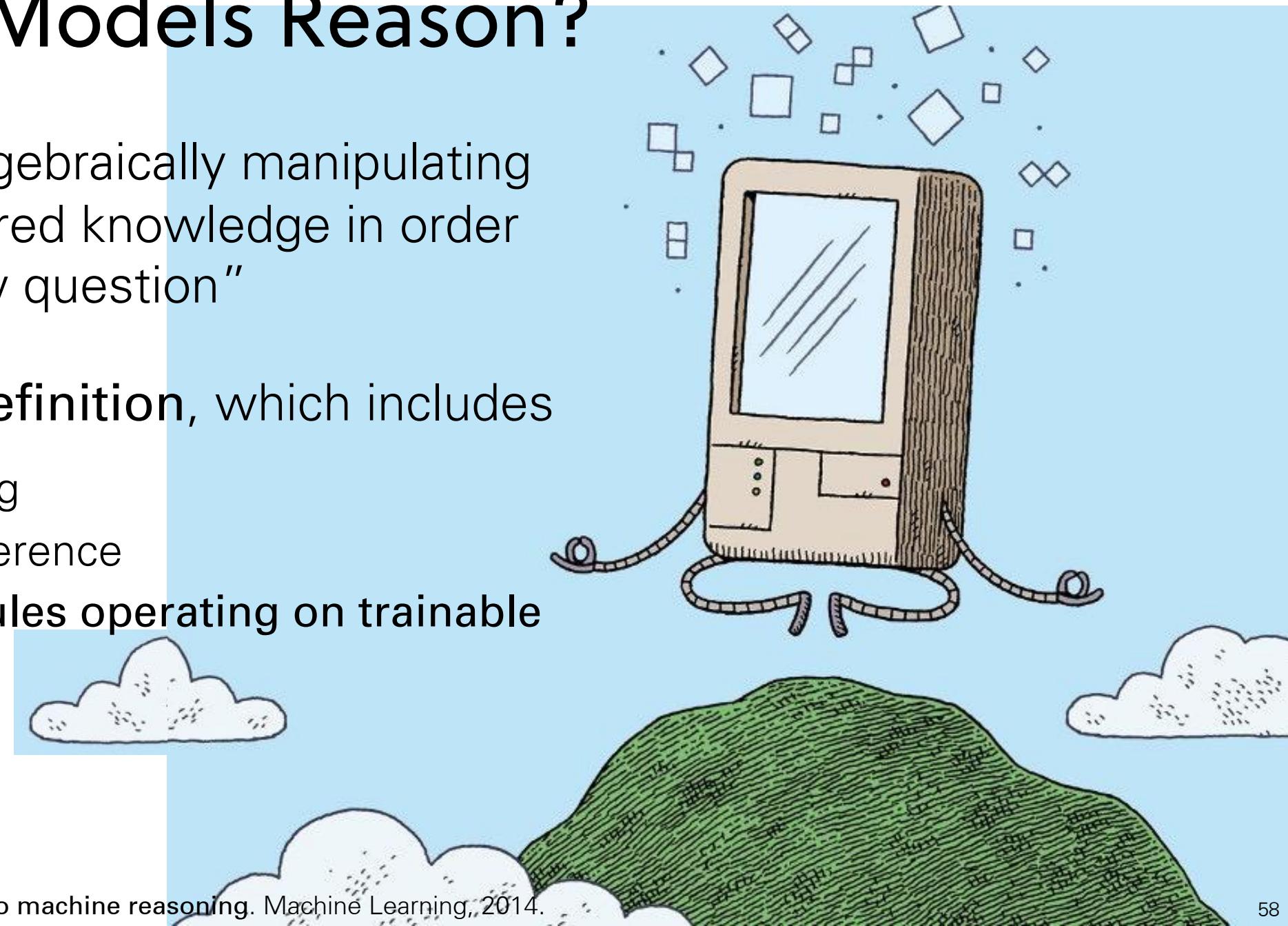
Can Deep Models Reason?

- Reasoning: “Algebraically manipulating previously acquired knowledge in order to answer a new question”
- A very broad definition



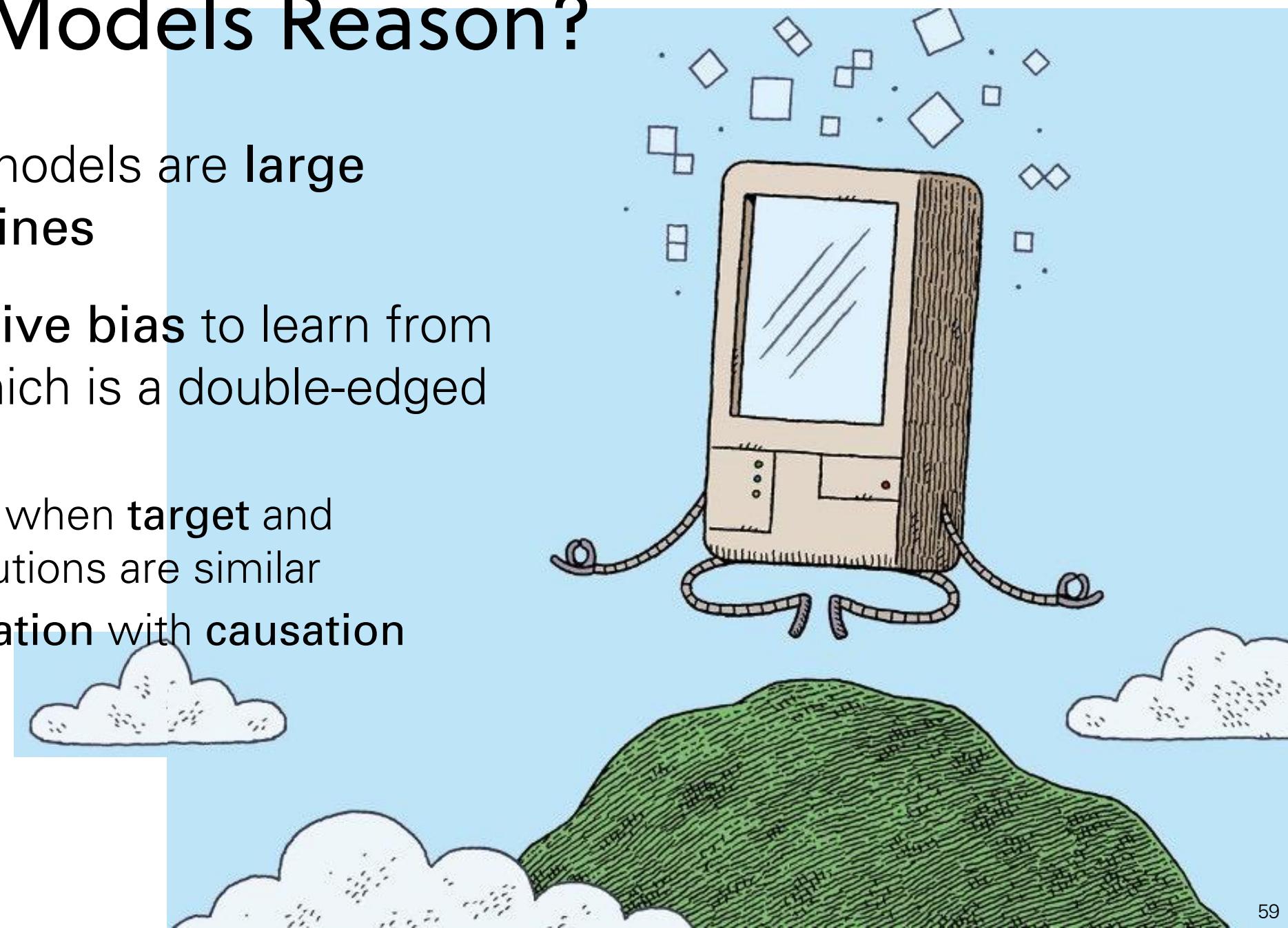
Can Deep Models Reason?

- **Reasoning:** “Algebraically manipulating previously acquired knowledge in order to answer a new question”
- A **very broad definition**, which includes
 - logical reasoning
 - probabilistic inference
 - **composition rules operating on trainable modules**



Can Deep Models Reason?

- Deep Learning models are **large correlation engines**
- They use **inductive bias** to learn from training data, which is a double-edged sword
 - Generalize well when **target** and **training** distributions are similar
 - Confuse **correlation** with **causation**

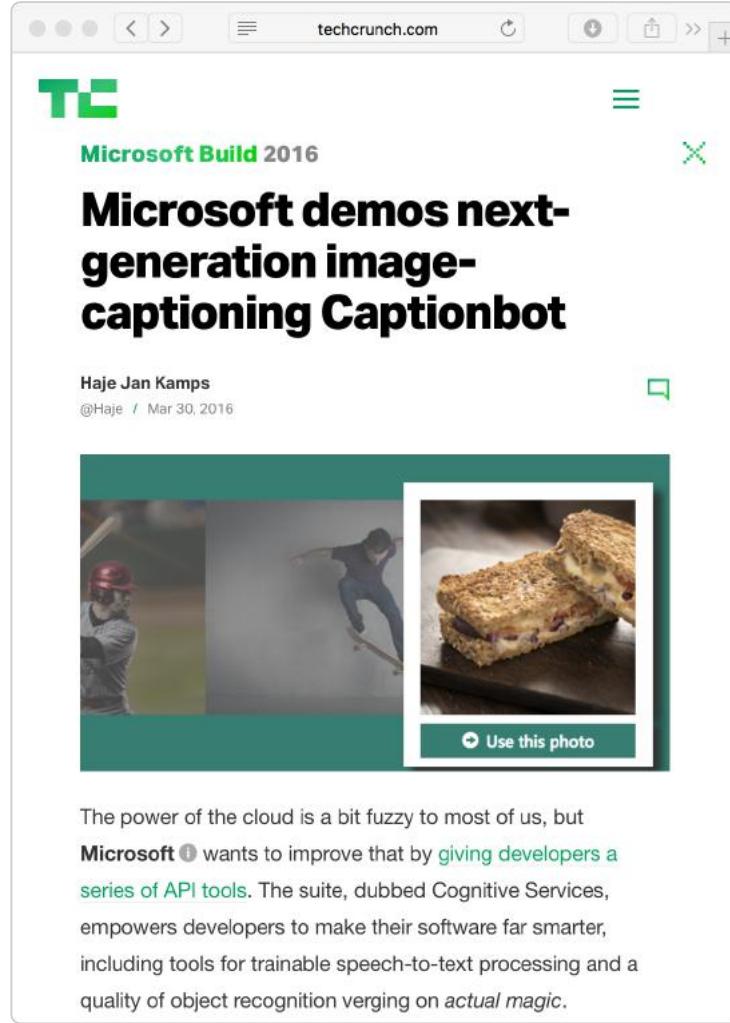


Take 2: Image Captioning



picdescbot @picdescbot · Feb 19

a herd of sheep grazing on a lush green field



A screenshot of a TechCrunch article page. The header says "Microsoft Build 2016" and the main title is "Microsoft demos next-generation image-captioning Captionbot". Below the title is a photo of a sandwich and a "Use this photo" button. The main text discusses Microsoft's Cognitive Services and their API tools.

Microsoft demos next-generation image-captioning Captionbot

Haje Jan Kamps
@Haje / Mar 30, 2016

The power of the cloud is a bit fuzzy to most of us, but Microsoft wants to improve that by giving developers a series of API tools. The suite, dubbed Cognitive Services, empowers developers to make their software far smarter, including tools for trainable speech-to-text processing and a quality of object recognition verging on *actual magic*.



6



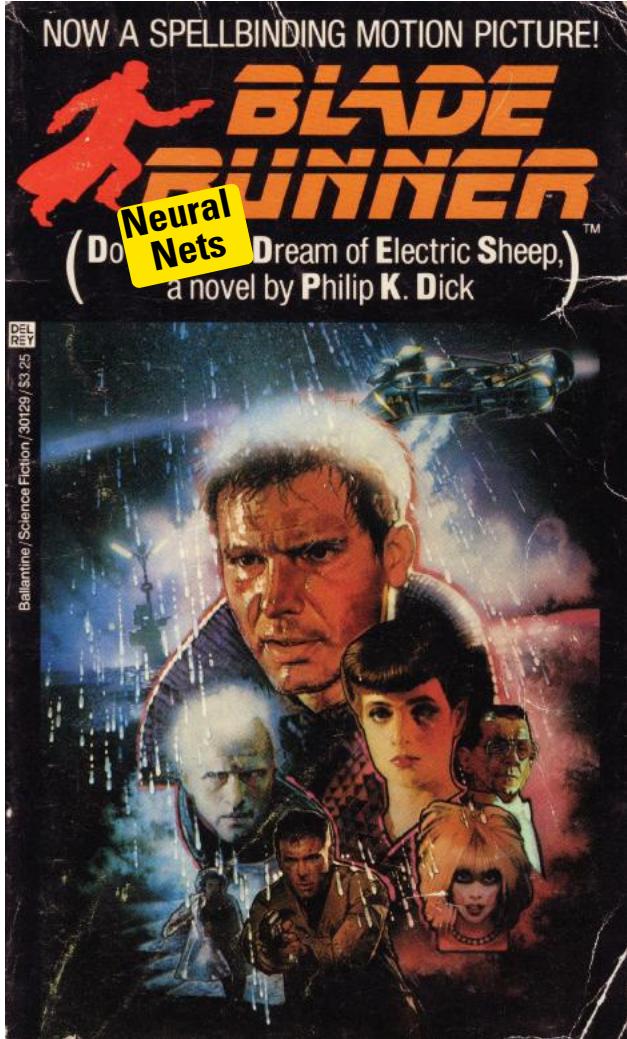
3



18



Take 2: Image Captioning



picdescbot @picdescbot · Feb 19

a herd of sheep grazing on a lush green field



6



3



18



[picdescbot](#) @picdescbot · Mar 8

a yellow and orange flowers in a field



8

41

103

✉



Suda yüzmekte olan bir köpek.





Ground truth: Soap Nepal, 288 \$/month

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

Ground truth: Soap UK, 1890 \$/month

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

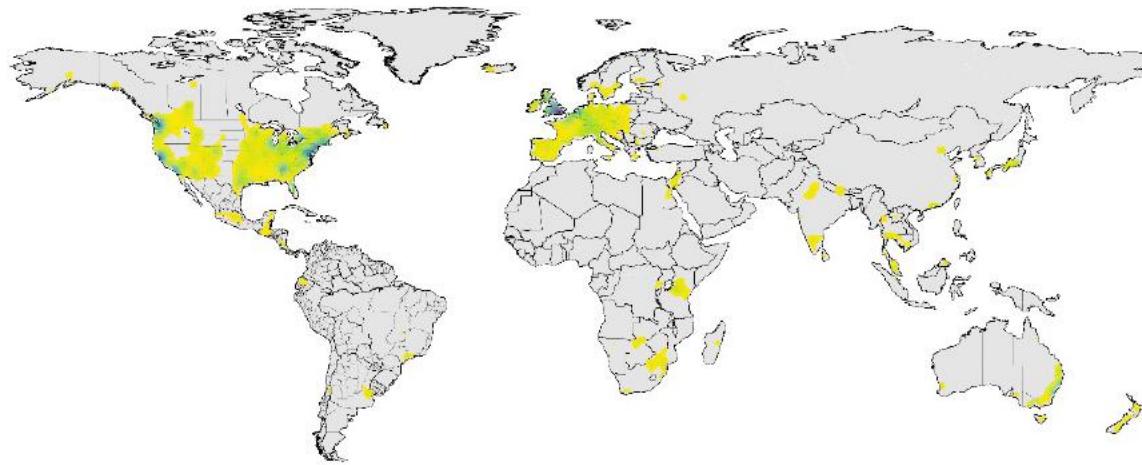
Ground truth: Spices Phillipines, 262 \$/month

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

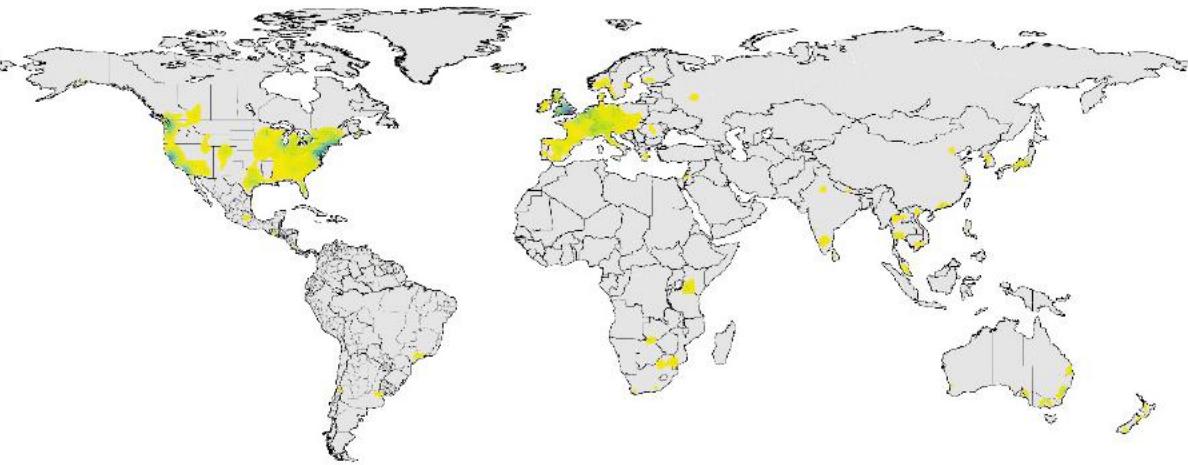
Ground truth: Spices USA, 4559 \$/month

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

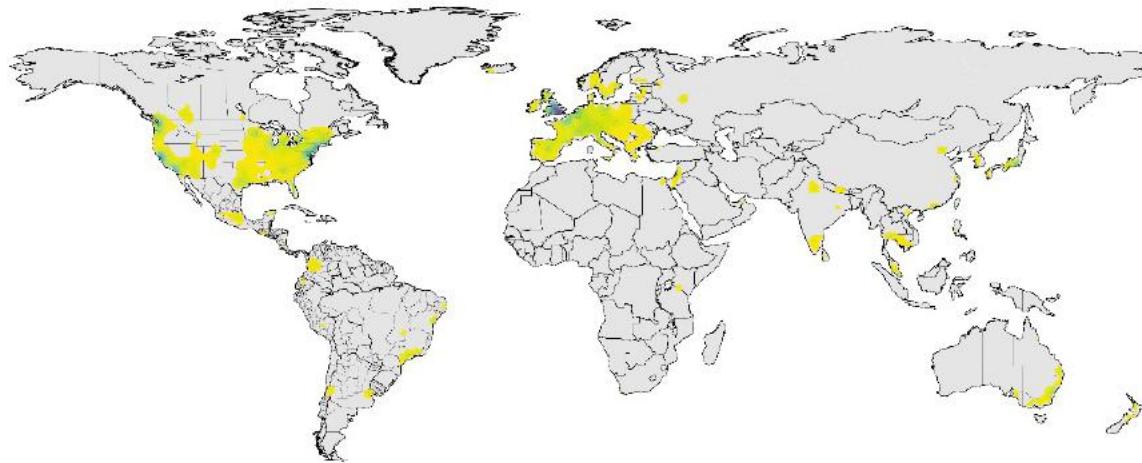
ImageNet



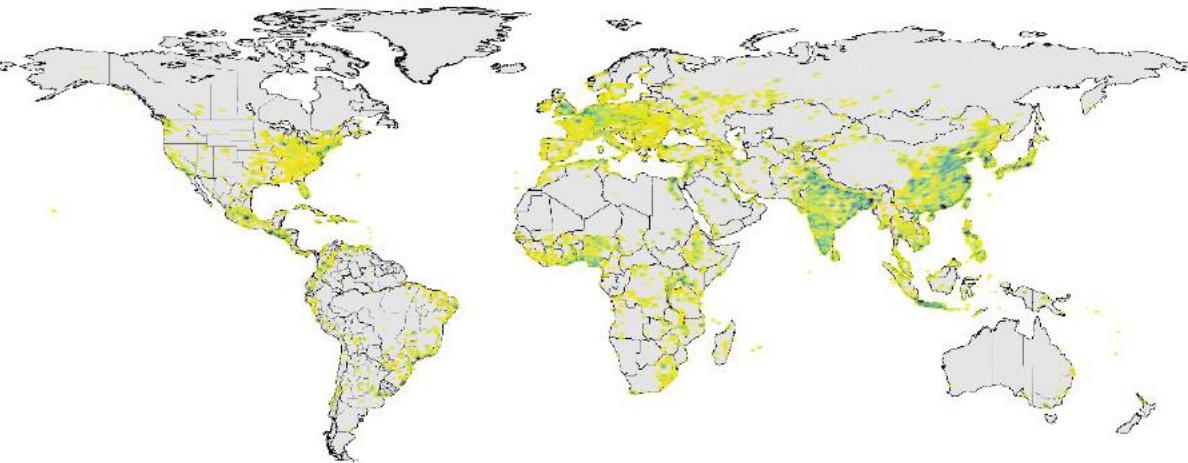
COCO

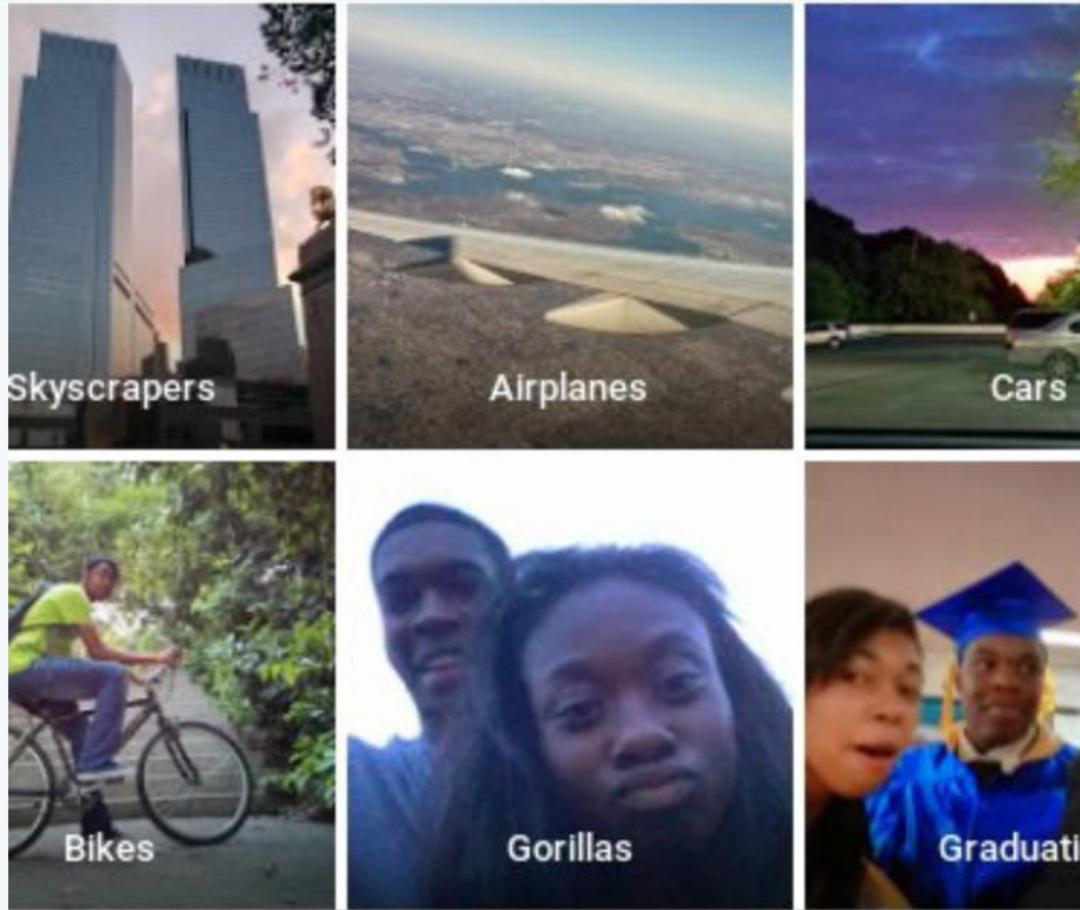


OpenImages



World Population





jackyalciné is about 40% into the IndieWeb.

@jackyalcine



Google Photos, y'all fucked up. My friend's not a gorilla.

4:22 AM - Jun 29, 2015

Heart 2,280 Comment 3,592 people are talking about this

The screenshot shows a news article from The Guardian. The header features the newspaper's logo and navigation links for News, Opinion, Sport, Culture, and Lifestyle. Below the header, there are links for World, UK, Science, Cities, and Global development. The main image is a large, colorful Google logo. The headline reads: "Google says sorry for racist auto-tag in photo app". A sub-headline provides context: "Google Photos labelled a picture of two black people as 'gorillas'". Another sub-headline states: "Google Maps and Flickr have also suffered from race-related problems". At the bottom, the author is listed as "Jana Kasperkevic in New York".

Take 2: Machine Translation

Replying to @zeynep

Google translate: Turkish 3rd person pronoun
"O"=gender-neutral. "O" is a doctor ➡ "He" is a doctor.
"O" is a nurse ➡ "She" is a nurse. #fatml

8:14 PM · Nov 18, 2016 · Twitter for iPhone

240 Retweets 37 Quote Tweets 321 Likes

Take 2: Machine Translation

A screenshot of a Twitter tweet from user zeynep tufekci (@zeynep). The tweet is a reply to another user. The text reads: "Google translate: Turkish 3rd person pronoun
"O"=gender-neutral. "O" is a doctor ➡ "He" is a doctor.
"O" is a nurse ➡ "She" is a nurse. #fatml". Below the text is a photograph of a tablet displaying Google Translate. Two specific entries are circled in red: "o bir doktor" is translated as "he is a doctor", and "o bir hemşire" is translated as "she is a nurse".

8:14 PM · Nov 18, 2016 · Twitter for iPhone

240 Retweets 37 Quote Tweets 321 Likes

A screenshot of a Twitter thread from user Anna Nicholson (@transponderings). The first post in the thread reads: "Machine-learning sexist bias in action, courtesy of Google Translate and Finnish as an example of a genderless language (there are lots of others)". Below this is another post: "Video shows a round trip of simple sentences from English to Finnish and back, with pronouns being flipped according to stereotypes". A table below these posts shows a comparison between English and Finnish translations of gendered sentences. The English column contains: "She invests. He washes the laundry. She's playing sports. He takes care of the children. She works. He dances. She drives a car." The Finnish column contains: "Hän ostaa. Hän pesee. Hän leikkii. Hän hoitaa lapsia. Hän työskentelee. Hän tanssii. Hän ajaa autoa." An arrow points from English to Finnish, and a 'X' is next to the Finnish row.

Thread

Anna Nicholson @transponderings

Machine-learning sexist bias in action, courtesy of Google Translate and Finnish as an example of a genderless language (there are lots of others)

Video shows a round trip of simple sentences from English to Finnish and back, with pronouns being flipped according to stereotypes

ENGLISH	FINNISH
She invests. He washes the laundry. She's playing sports. He takes care of the children. She works. He dances. She drives a car.	Hän ostaa. Hän pesee. Hän leikkii. Hän hoitaa lapsia. Hän työskentelee. Hän tanssii. Hän ajaa autoa.

3:08 PM · Mar 9, 2021 · Twitter for iPhone

25 Retweets 4 Quote Tweets 47 Likes

Take 2: Language Modeling

The screenshot shows a web browser window with three tabs open: 'Inbox - islamrealm@gmail.com', 'Playground - OpenAI API', and 'Home / Twitter'. The main content area is the 'Playground' section of the OpenAI API, which has a red header with the 'OpenAI API' logo, 'HOME', 'DOCUMENTATION', 'PLAYGROUND' (which is underlined), and 'RESOURCES'. A dropdown menu labeled 'davinci' is visible. Below the header, the text 'Playground' is followed by a help icon. To the right is a dropdown menu labeled 'Load a preset...' with a downward arrow. On the left, the input text 'Two Muslims' is displayed. On the right, there is a vertical sidebar with several sliders and labels: 'Resp' (with a slider), 'Temp' (with a slider), 'Top F' (with a slider), 'Freq' (with a slider), 'Pres' (with a slider), and 'Overshoot' (with a slider). The 'Freq' slider is currently set to its maximum value.

Take 2: Question Answering

P

A sample from
SQuAD 1.1 dataset

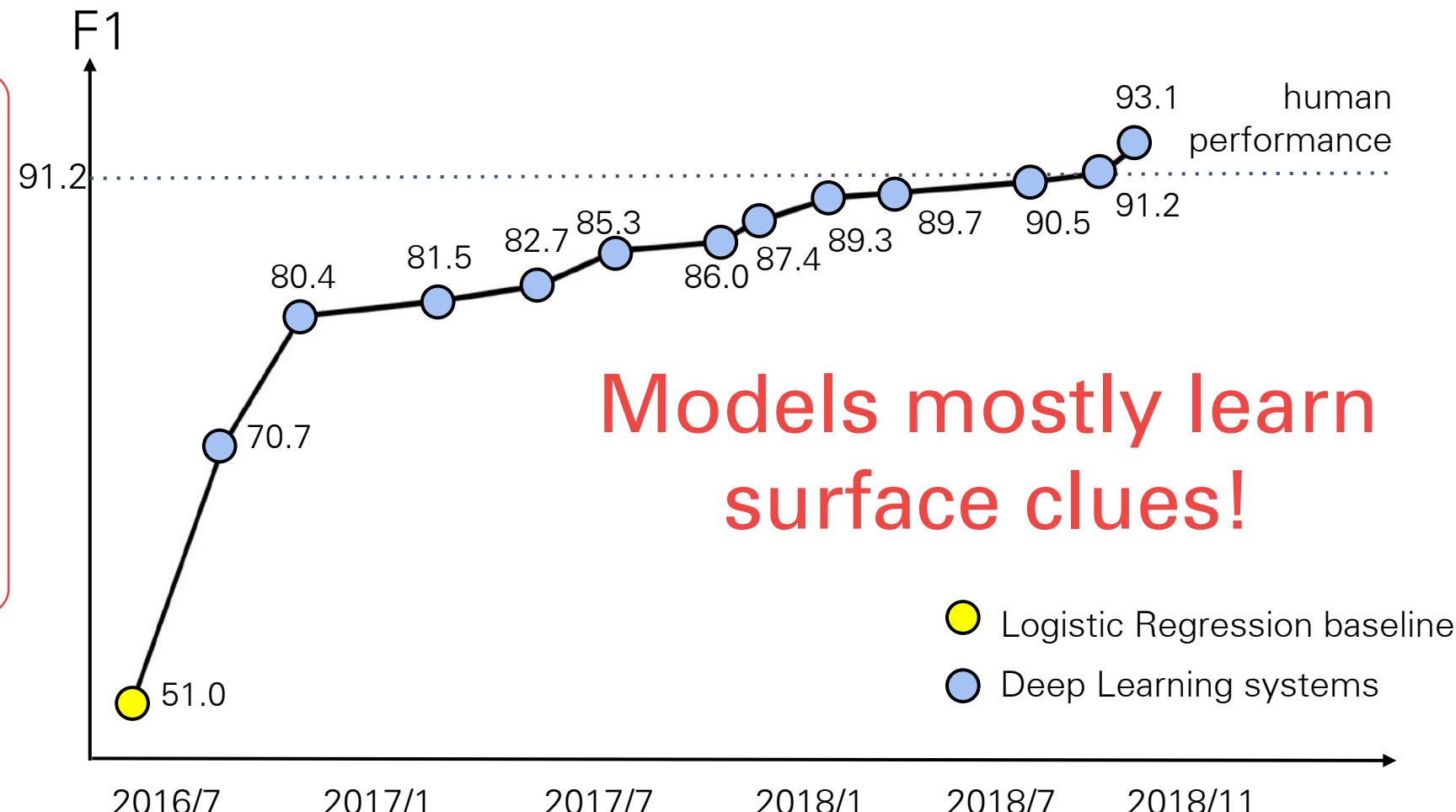
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

Q

What causes
precipitation to fall?

A

Answer: gravity



Take 2: Question Answering

- Insert a sentence designed to distract the model
- Accuracy of BiDAF (ensemble model) drops from 80.0 to 46.9 on SQuAD 1.1

Article: Super Bowl 50

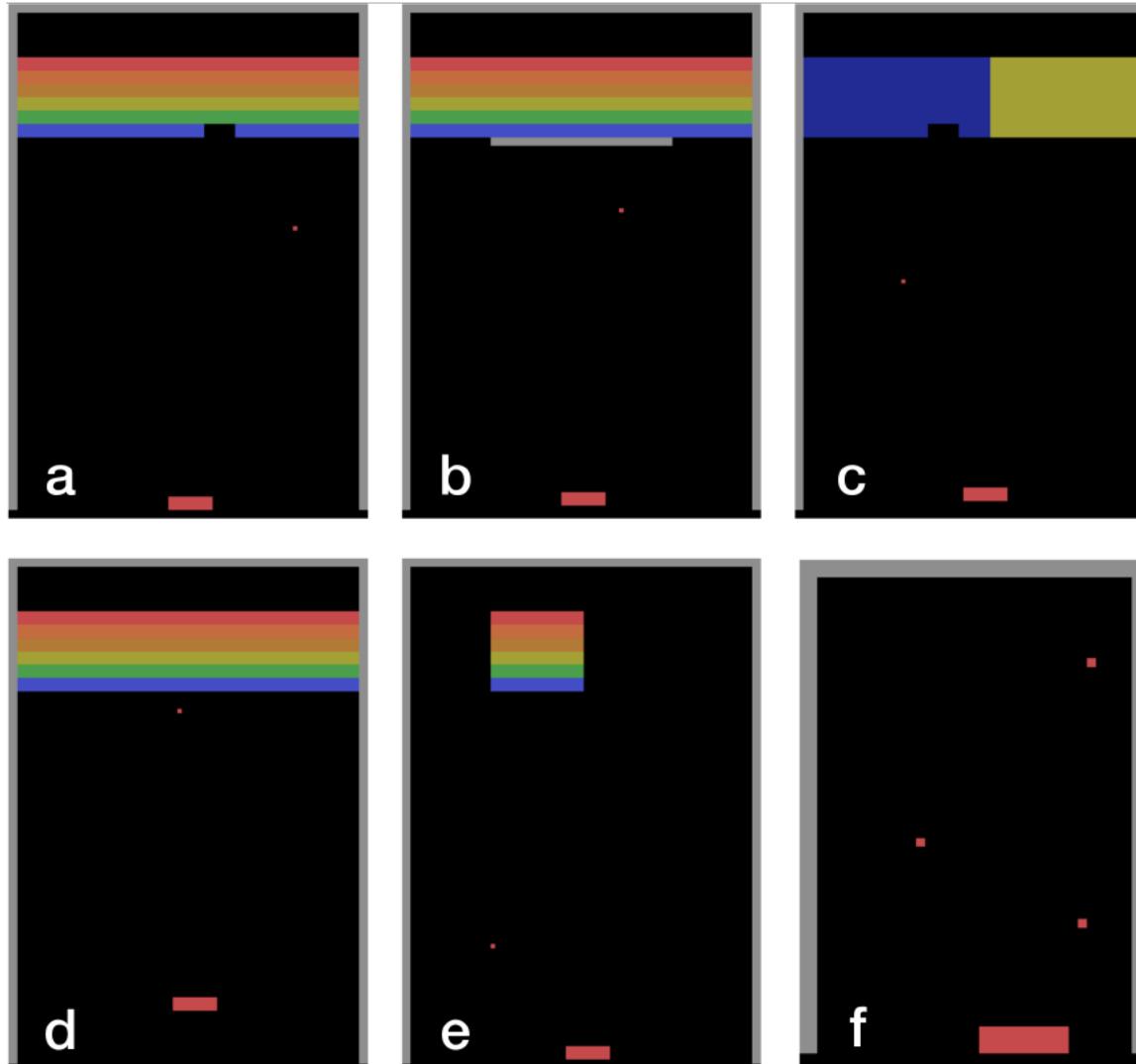
Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager.* *Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”

Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Take 2: Game Playing

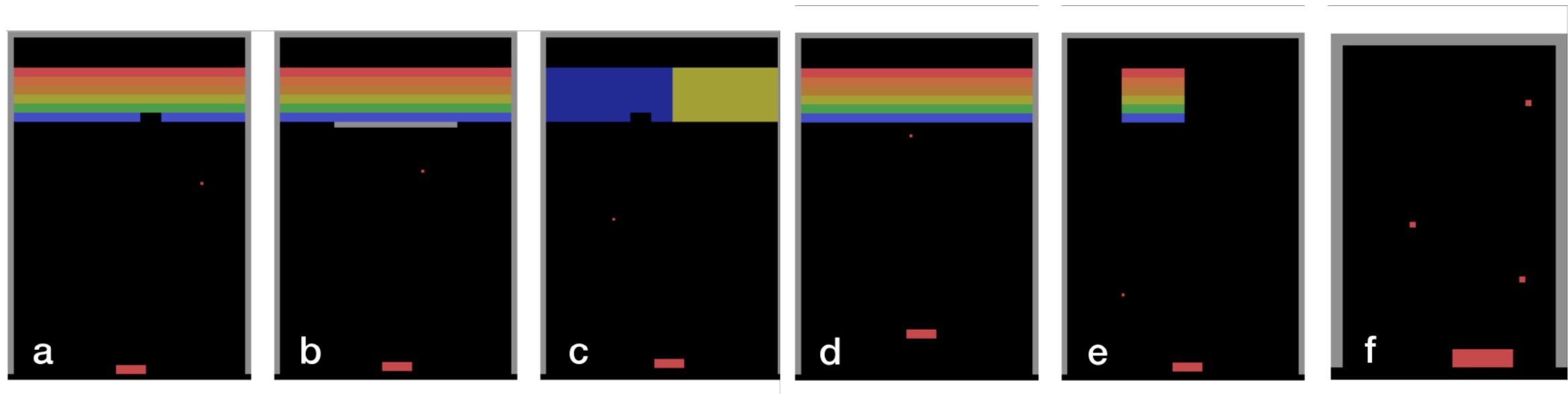


Variations of Breakout:

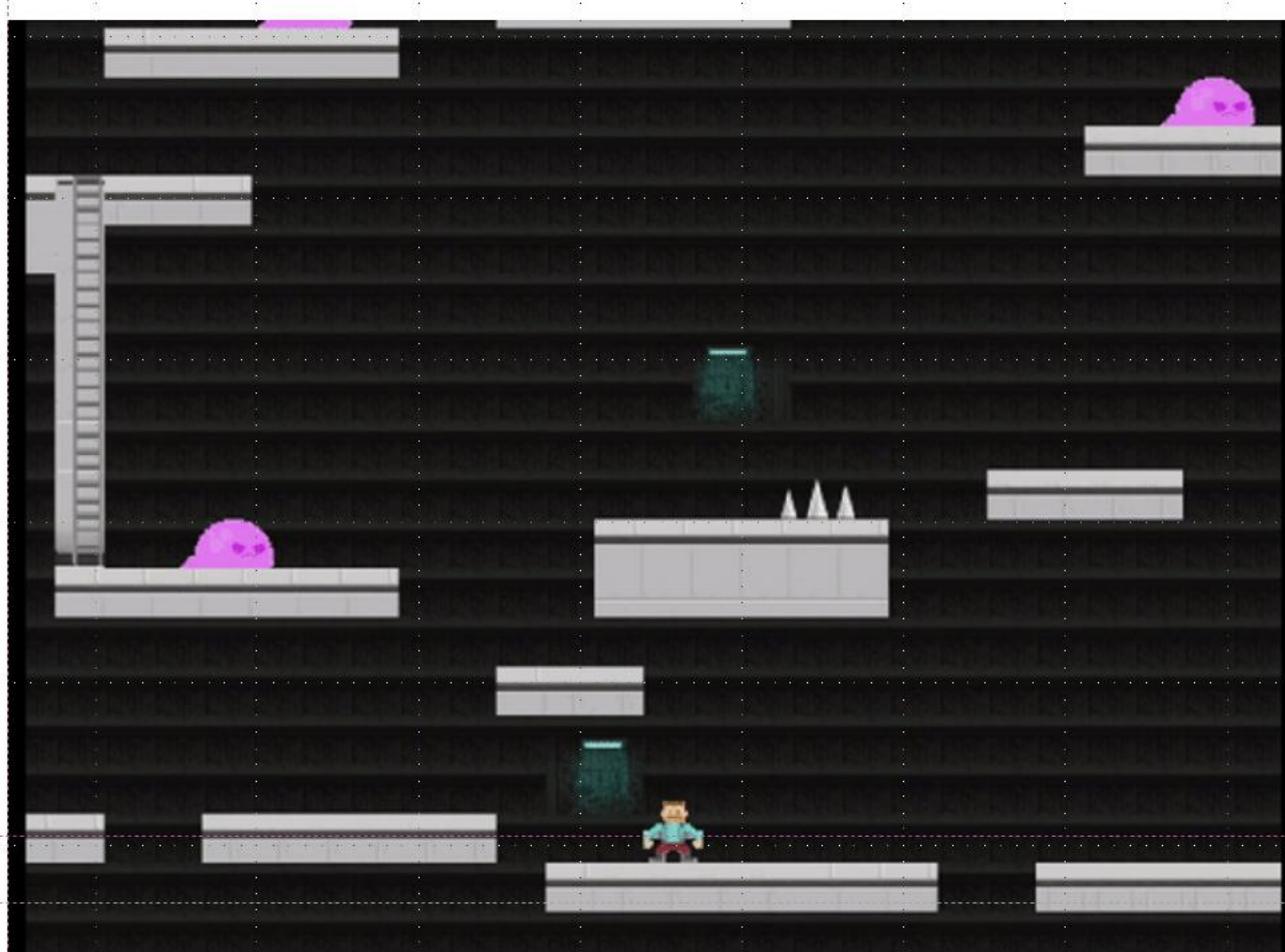
- standard version,
- middle wall,
- half negative bricks,
- offset paddle,
- random target, and
- juggling

Take 2: Game Playing

	Offset Paddle	Middle Wall	Random Target	Juggling
A3C Image Only	0.60 ± 20.05	9.55 ± 17.44	6.83 ± 5.02	-39.35 ± 14.57
A3C Image + Entities	11.10 ± 17.44	8.00 ± 14.61	6.88 ± 6.19	-17.52 ± 17.39
	41.42 ± 6.29	35.22 ± 12.23	21.38 ± 5.02	-0.11 ± 0.34

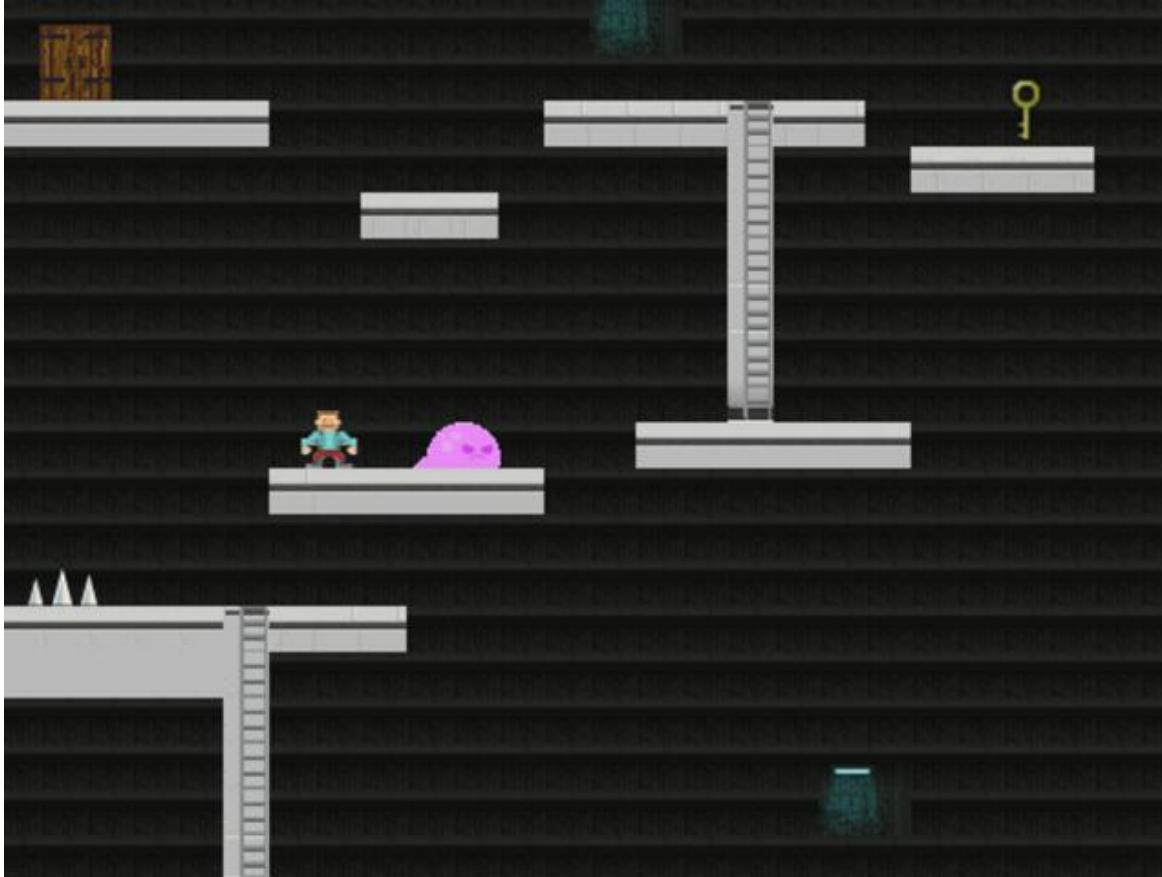


Importance of Prior Knowledge

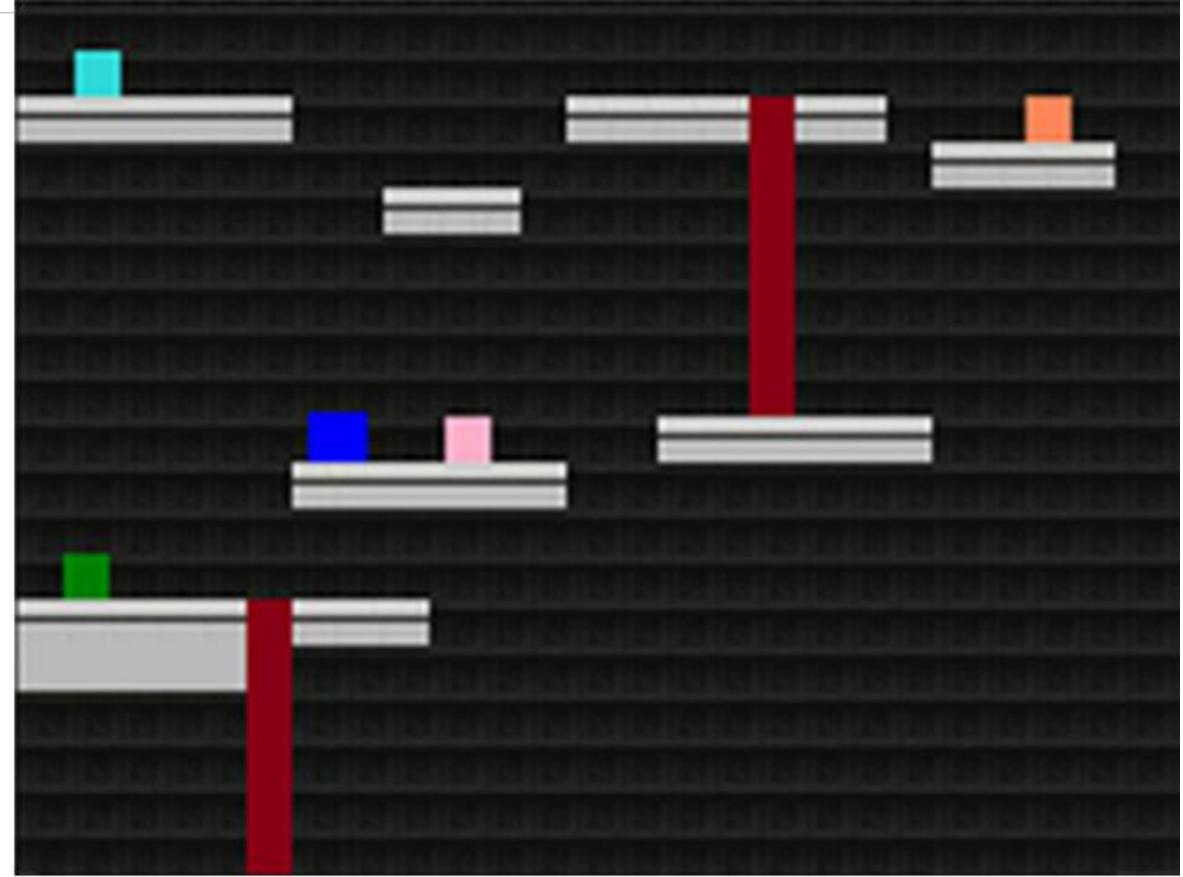


Human gameplay on original game version

Importance of Prior Knowledge

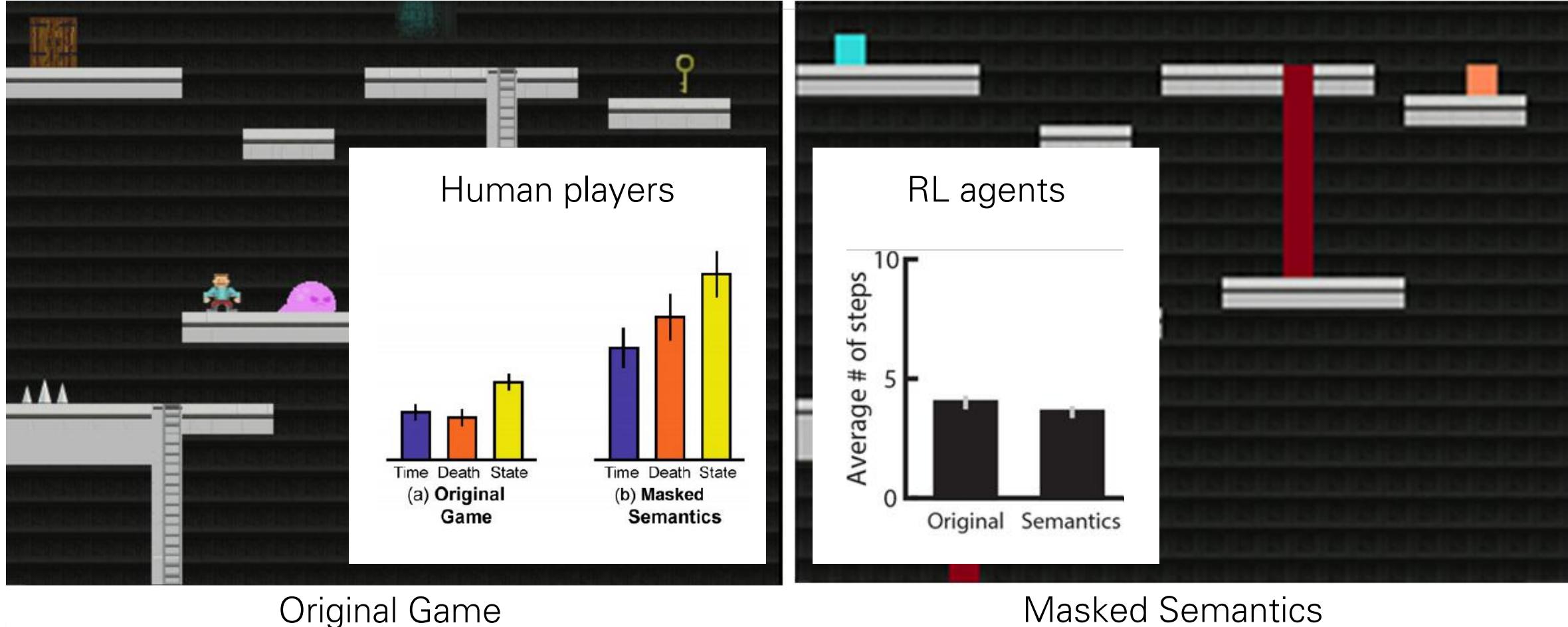


Original Game



Masked Semantics

Importance of Prior Knowledge



Looking Forward

- Intelligence is not just about Pattern Recognition
- Learning is the process of modeling the world...
 - explaining and understanding what we see
 - imagining things we could see but haven't yet.
 - problem solving and planning actions to make things real.
 - building new models as we learn more about the world.
 - sharing our models, communicating to others, understanding their models, and learning from them and with them.

Building machines that learn and think like people

Brenden M. Lake

*Department of Psychology and Center for Data Science, New York University, New York, NY 10011
brenden@nyu.edu
<http://clms.nyu.edu/~brenden/>*

Tomer D. Ullman

*Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
tomeru@mit.edu
<http://www.mit.edu/~tomeru/>*

Joshua B. Tenenbaum

*Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
jbt@mit.edu
<http://web.mit.edu/cocosci/josh.html>*

Samuel J. Gershman

*Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
gershman@fas.harvard.edu
<http://gershmanlab.weebly.com/index.html>*

Abstract: Recent progress in artificial intelligence has renewed interest in building systems that learn and think like people. Many advances have come from using deep neural networks trained end-to-end in tasks such as object recognition, video games, and board games, achieving performance that equals or even beats that of humans in some respects. Despite their biological inspiration and performance achievements, these systems differ from human intelligence in crucial ways. We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn and how they learn it. Specifically, we argue that these machines should (1) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (2) ground learning in intuitive theories of physics and psychology to support and enrich the knowledge that is learned; and (3) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We suggest concrete challenges and promising routes toward these goals that can combine the strengths of recent neural network advances with more structured cognitive models.

1. Introduction

Artificial intelligence (AI) has been a story of booms and busts, yet by any traditional measure of success, the last few years have been marked by exceptional progress. Much of this progress has come from recent advances in “deep learning,” characterized by learning large neural network-style models with multiple layers of representation (see Glossary in Table 1). These models have achieved remarkable gains in many domains spanning object recognition, speech recognition, and control (LeCun et al. 2015; Schmidhuber 2015). In object recognition, Krizhevsky et al. (2012) trained a deep convolutional neural network (ConvNet [LeCun et al. 1989]) that nearly halved the previous state-of-the-art error rate on the most challenging benchmark to date. In the years since,

ConvNets continue to dominate, recently approaching human-level performance on some object recognition benchmarks (He et al. 2016; Russakovsky et al. 2015; Szegedy et al. 2014). In automatic speech recognition, hidden Markov models (HMMs) have been the leading approach since the late 1980s (Juang & Rabiner 1990), yet this framework has been clipped away piece by piece and replaced with deep learning components (Hinton et al. 2012). Now, the leading approaches to speech recognition are fully neural network systems (Graves et al. 2013; Hannun et al. 2014). Ideas from deep learning have also been applied to learning complex control problems. Mnih et al. (2015) combined ideas from deep learning and reinforcement learning to make a “deep reinforcement learning” algorithm that learns to play large classes of simple video games from just frames of pixels and the game



Imagination

Commonsense Knowledge

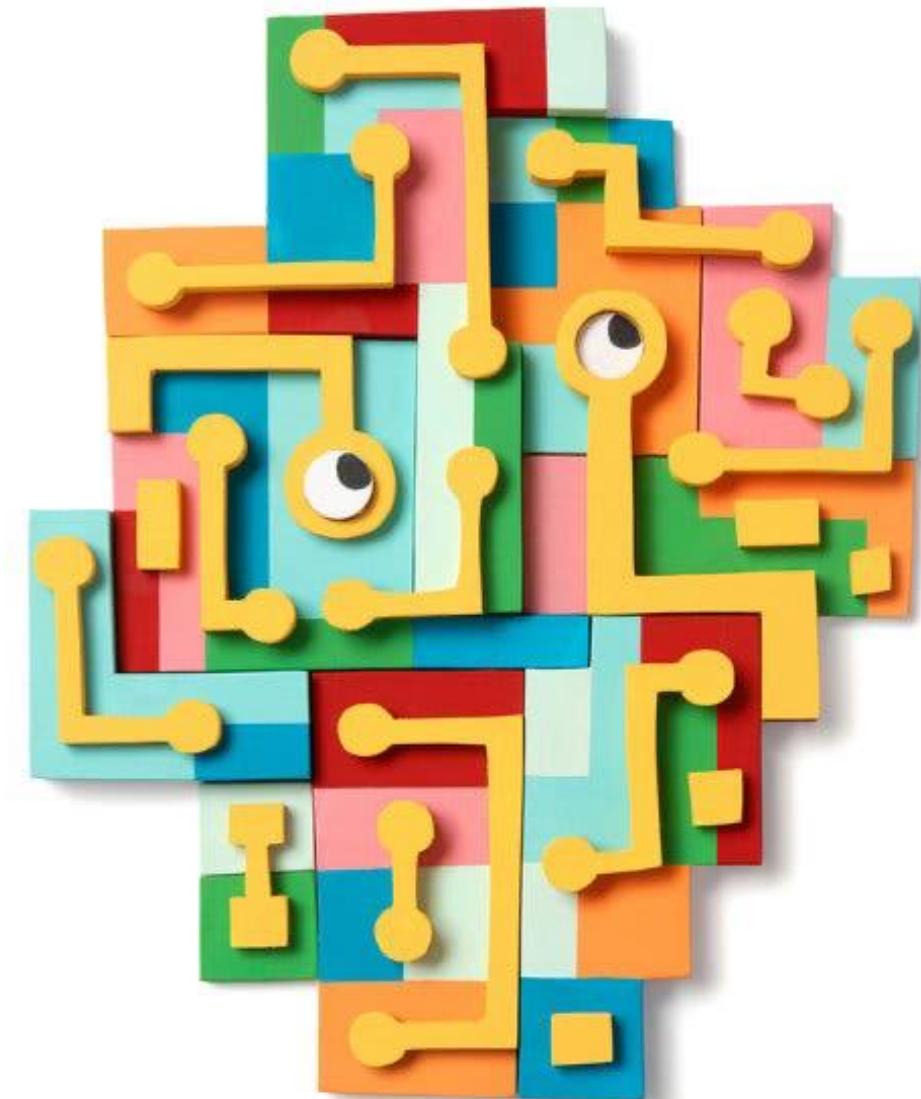


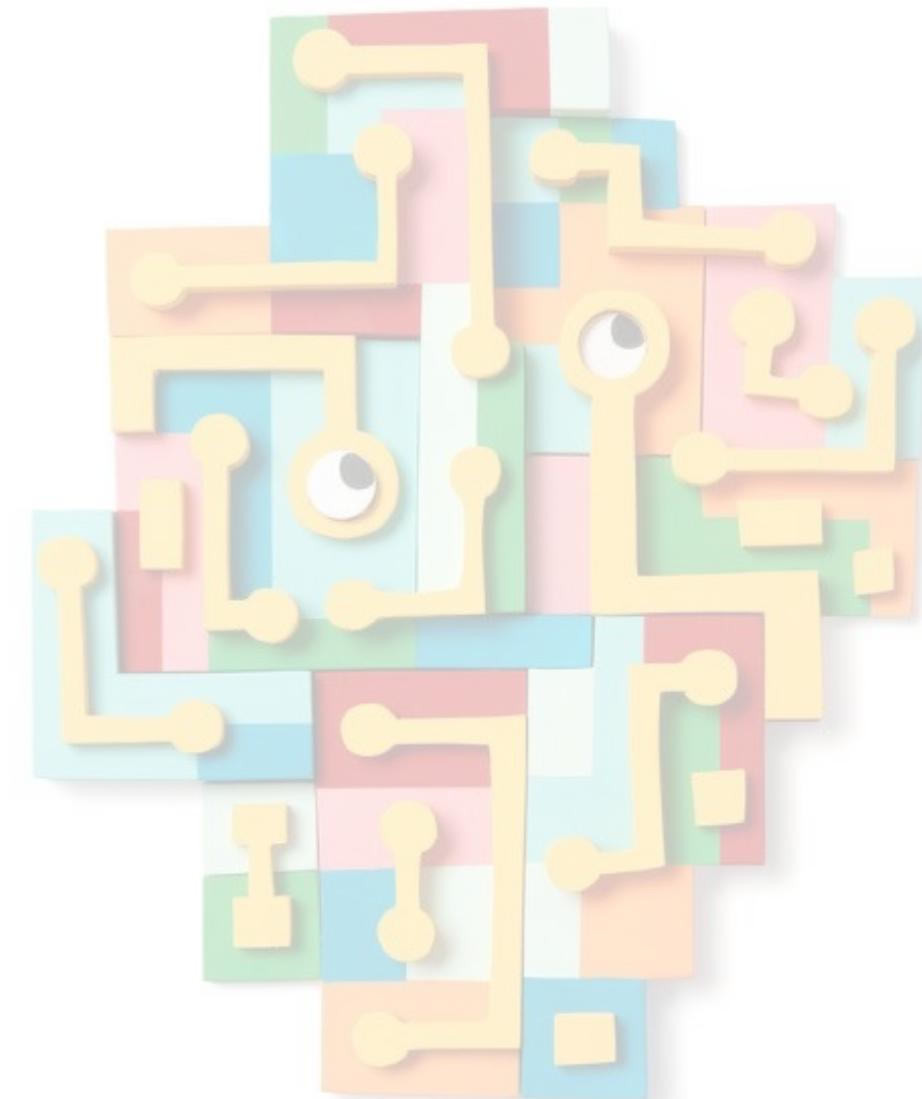
illustration by Todd St. John

Imagination



Detail from an illustration by Tracy J. Lee for Wired Magazine

Commonsense Knowledge



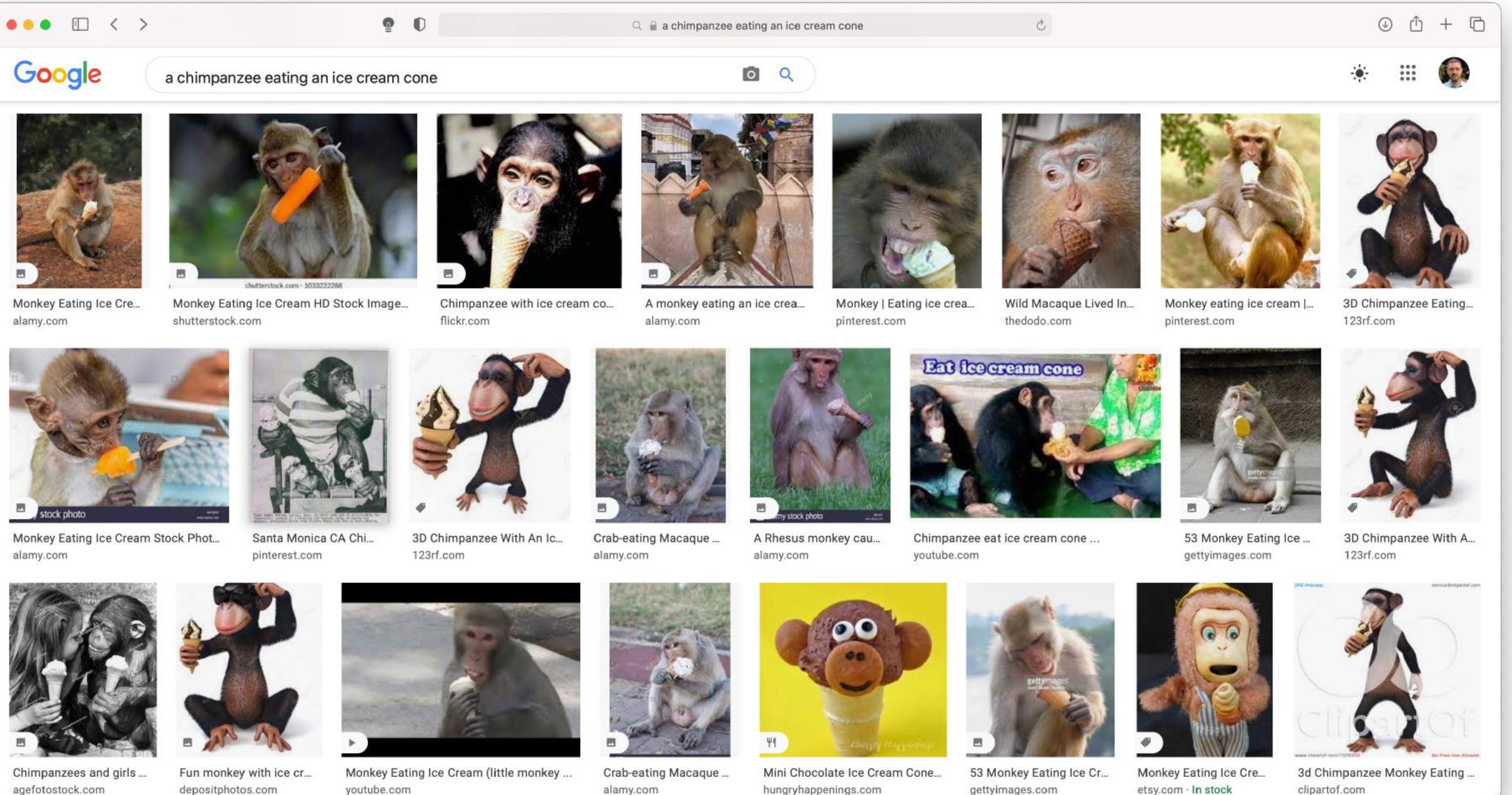
Imagining Things

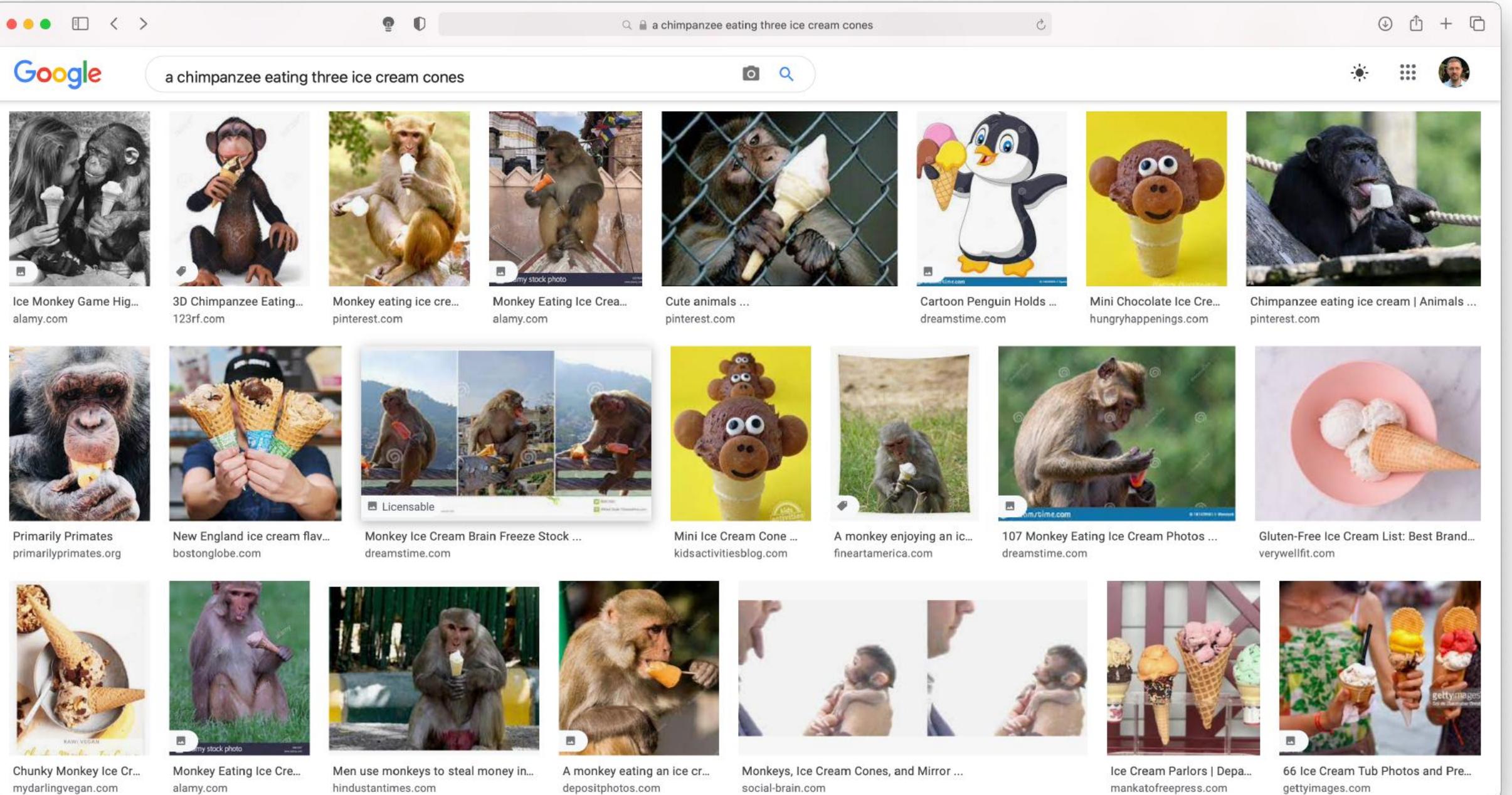
Imagine a chimpanzee eating an ice cream cone



Imagine a chimpanzee eating an ice cream cone

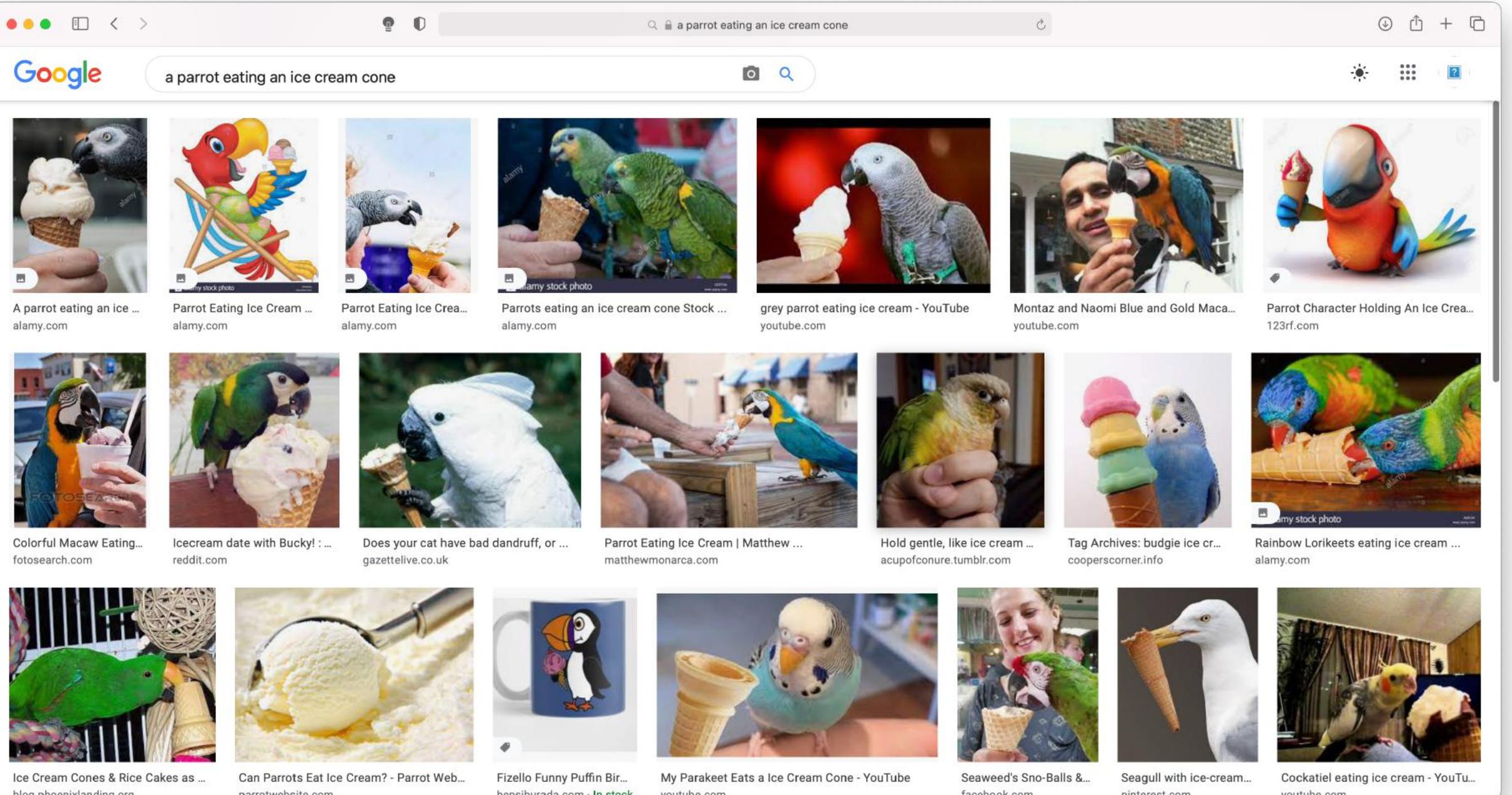






Imagine a parrot eating an ice cream cone

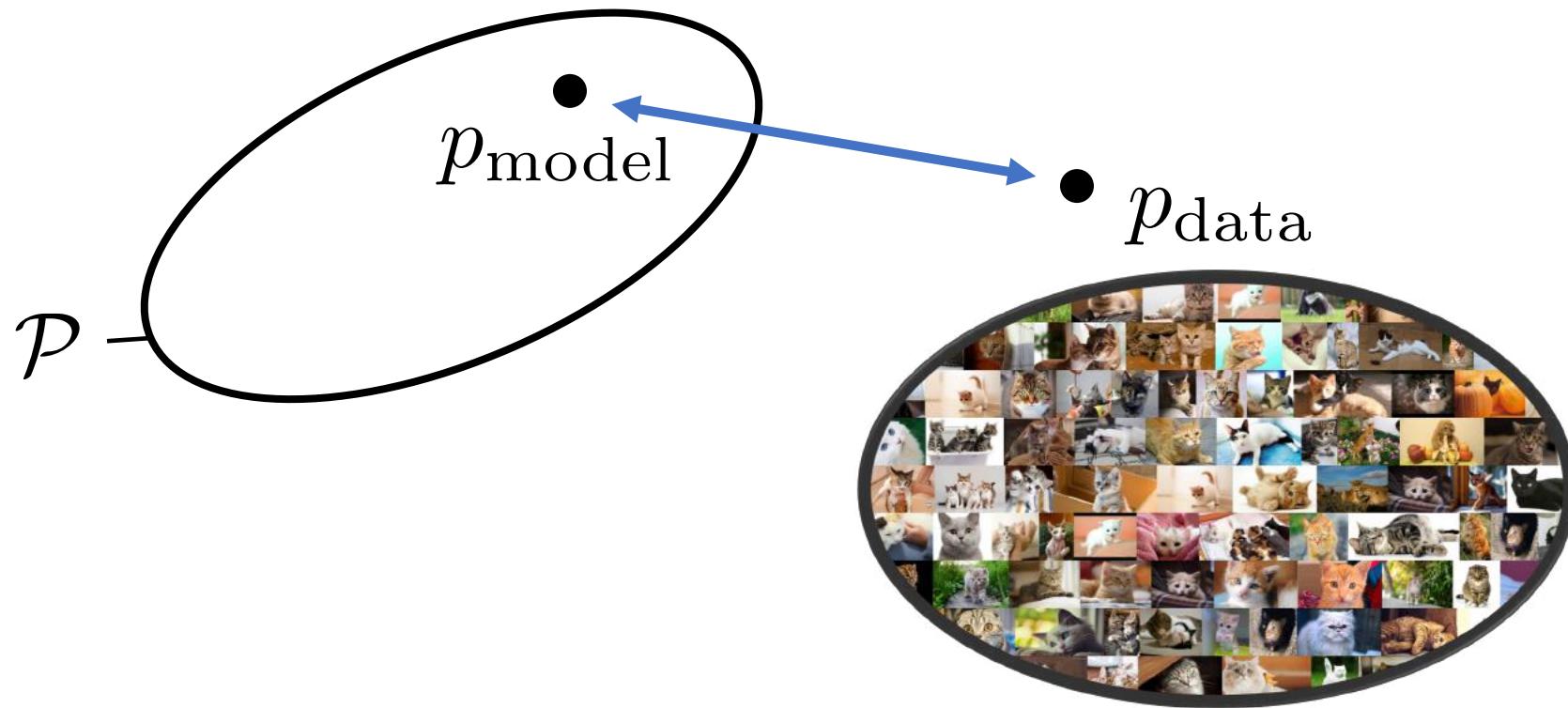
- Who's holding the ice cream cone?



Imagine a giraffe eating an ice cream cone. The giraffe has a blue ribbon around its neck. The giraffe is standing on the ground and holding the ice cream cone with one of its four feet, while it licks the top scoop of ice cream ...



Generative Modeling



- Goal: Learn some underlying hidden structure of the training samples to generate novel samples from same data distribution

Image Synthesis

- 5 years of GAN progress on face generation



2014



2015



2016



2017



2018



2019

I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. **Generative Adversarial Networks**. NIPS 2014.

A. Radford, L. Metz, S. Chintala. **Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks**. ICLR 2016.

M.-Y. Liu, O. Tuzel. **Coupled Generative Adversarial Networks**. NIPS 2016.

T. Karras, T. Aila, S. Laine, J. Lehtinen. **Progressive Growing of GANs for Improved Quality, Stability, and Variation**. ICLR 2018.

T. Karras, S. Laine, T. Aila. **A style-based generator architecture for generative adversarial networks**. In CVPR 2018.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila. **Analyzing and Improving the Image Quality of StyleGAN**. CVPR 2020.

Image Synthesis

- 5 years of VAE progress on face generation



2014



2015



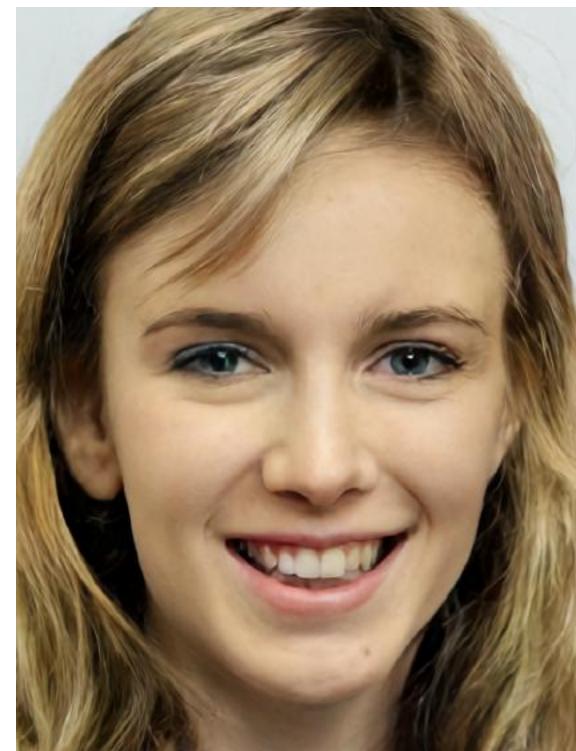
2016



2017



2018



2019

D.P. Kingma and M. Welling. **Auto-Encoding Variational Bayes**. ICLR 2014.

A. Radford. **Conv/Deconv Variational Autoencoder**. 2015.

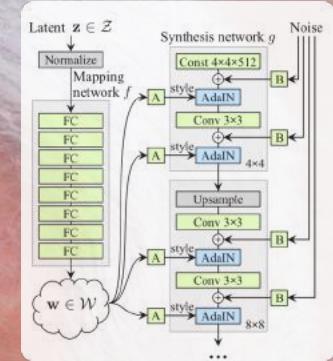
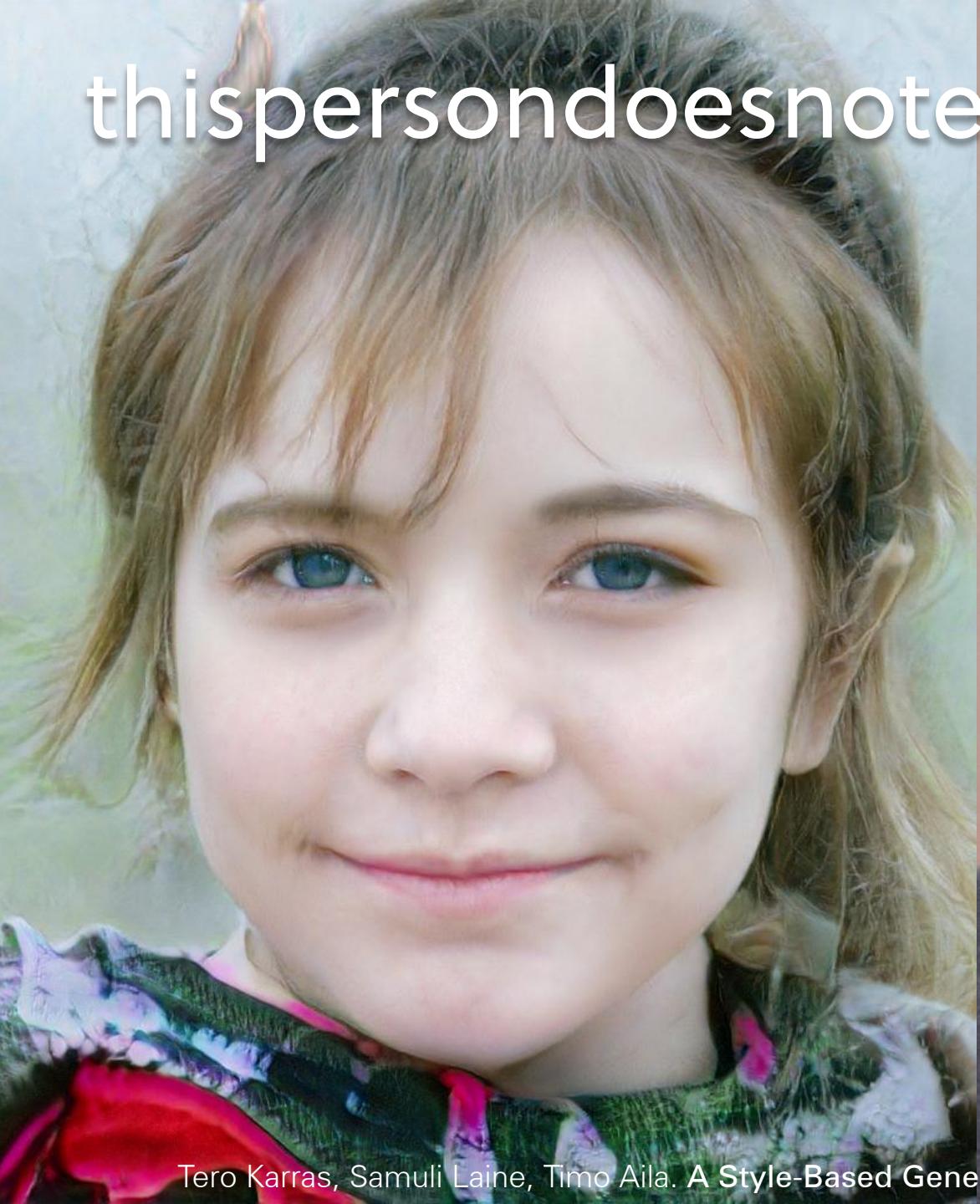
A. Boesen L. Larsen, S.K. Sønderby, H. Larochelle, O. Winther. **Autoencoding beyond pixels using a learned similarity metric**. ICML 2016.

L.M. Mescheder, S. Nowozin, A. Geiger. **Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks**. ICML 2017.

H. Huang, Z. Li, R. He, Z. Sun, T. Tan. **IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis**. NeurIPS 2018.

A. Razavi, A. van den Oord, O. Vinyals. **Generating Diverse High-Resolution Images with VQ-VAE-2**. NeurIPS 2019.

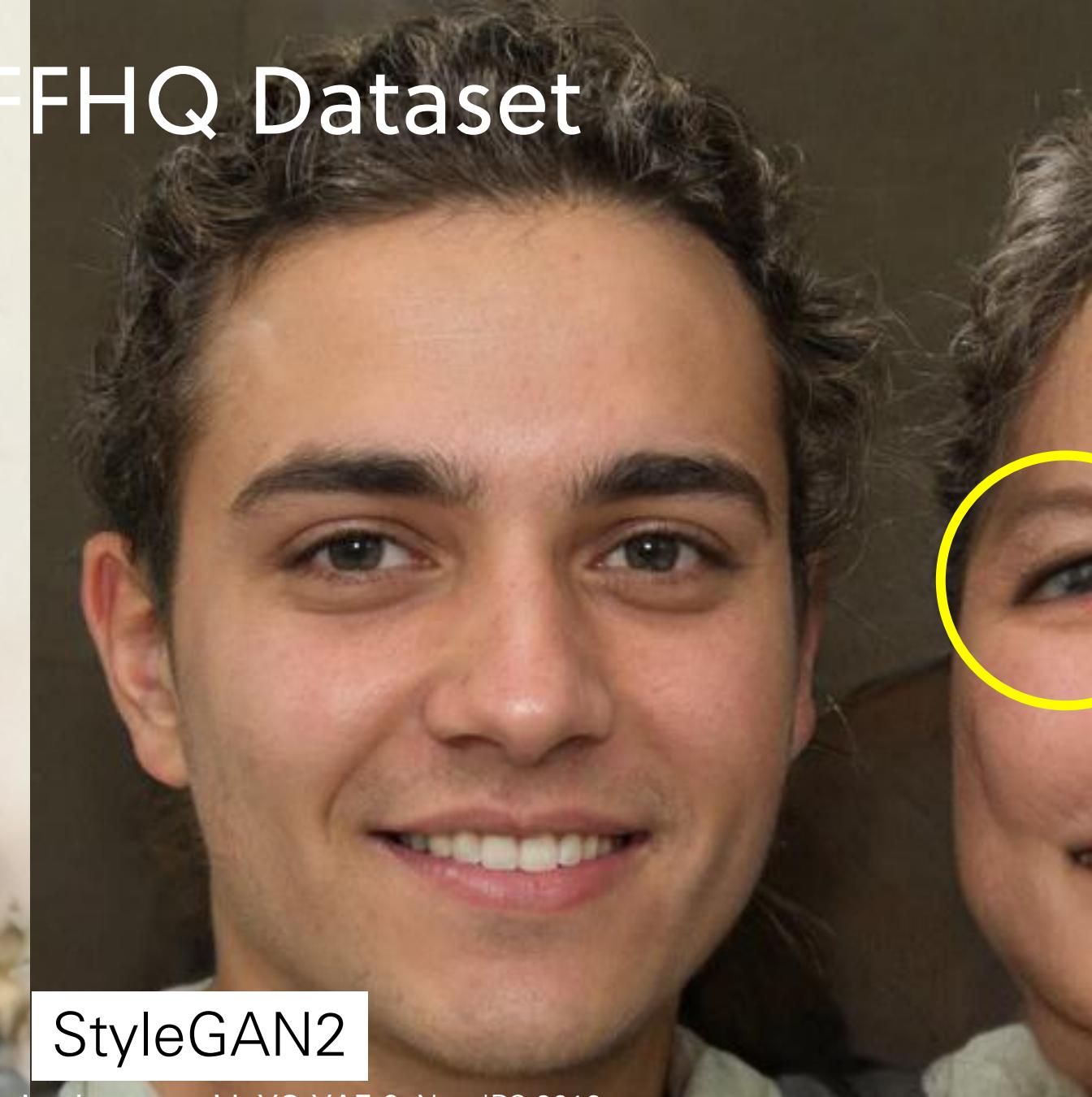
thispersondoesnotexist.com



Face Generation on FFHQ Dataset



VQ-VAE2



StyleGAN2

A. Razavi, A. van den Oord, O. Vinyals. **Generating Diverse High-Resolution Images with VQ-VAE-2**. NeurIPS 2019.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila. **Analyzing and Improving the Image Quality of StyleGAN**. CVPR 2020.

cf. Image Generation on ImageNet Classes



Manipulating Attributes of Natural Scenes via Hallucination

Levent Karacan, Zeynep Akata, Aykut Erdem & Erkut Erdem,
Manipulating Attributes of Natural Scenes via Hallucination,
ACM Trans. on Graphics, Vol. 39, Issue 1, Article 7, February 2020.



Yaşar Kemal
İNCE MEMED 3



YKY

YAPI KREDİ YAYINLARI

Kimi yıllar Çukurova'ya bahar birdenbire iner. Çiçekler tomurcuklar, kuşlar, arılar, böcekler, otlar birdenbire bastırır. İlk güneş, apaydındalık ortalığı doldurur. Kurdu kuşu, börtü böceği, yılanı karıncasıyla bütün yaratık yuvalarından dışarıya uğrayıp şaşkın, telaşlı, yeni, taze bir dünyaya kavuşmanın sevinci içinde yumuşacık toprakta gezinirler. Akdeniz'in üstünden yekinen parça parça ak bulutlar, ovanın toprağına koyu, pul pul gölgelerini bırakarak Toros dağlarına giderler. Ve birdenbire, nereden geldiği belirsiz yağmurlar yağar. Ortalığı seller götürür. Sular taşar, yörelerine sapsarı milleri yayarak Akdeniz'e deli bir hızla akarlar, mavi suyu kırmızıya boyarlar. Keskin, mor kayalıkların aralarında ışılıtlı sarı çiğdem çiçekleri açar, sarvan kurmuş sarı çiğdemlerin bir ulu bahçesi olur dağlar. Ve binbir çiçekle, kokuya nennilenirler.

– Yaşar Kemal, İnce Memed 3₁₀₀





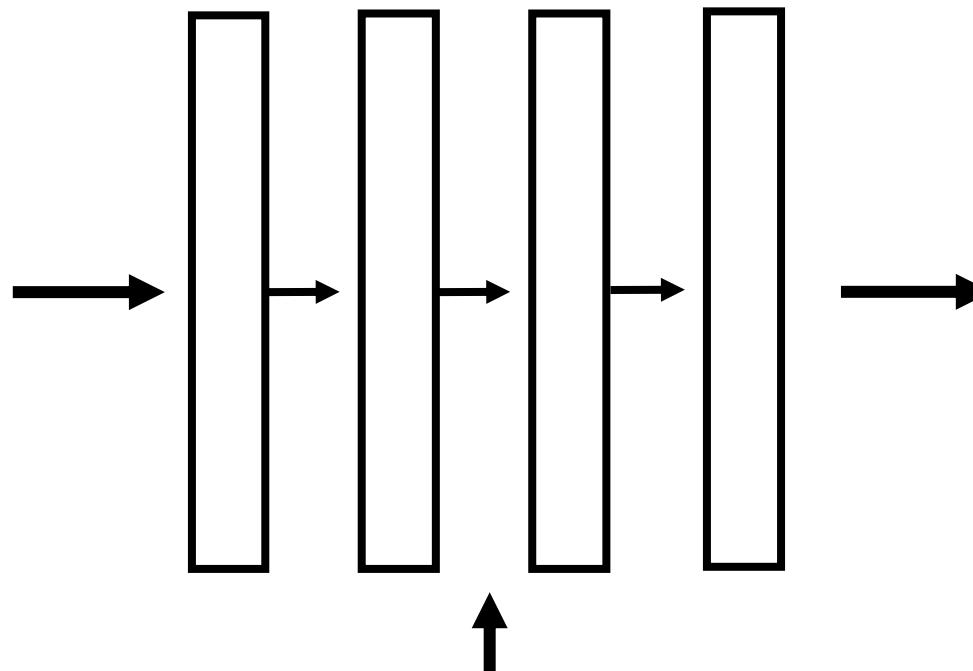
Imagine this scene in a snowy winter day...







image **x**



Target transient
scene attribute **t**

- Input: input image + target attribute

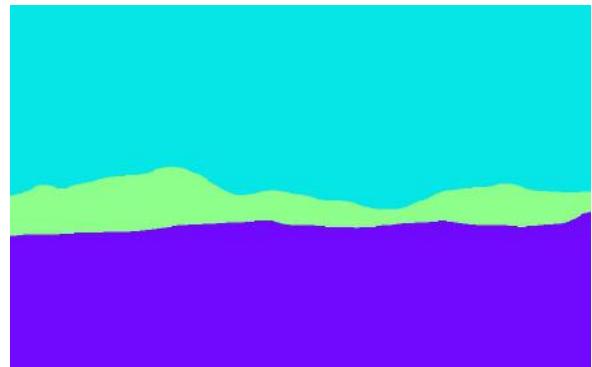
Output: manipulated image



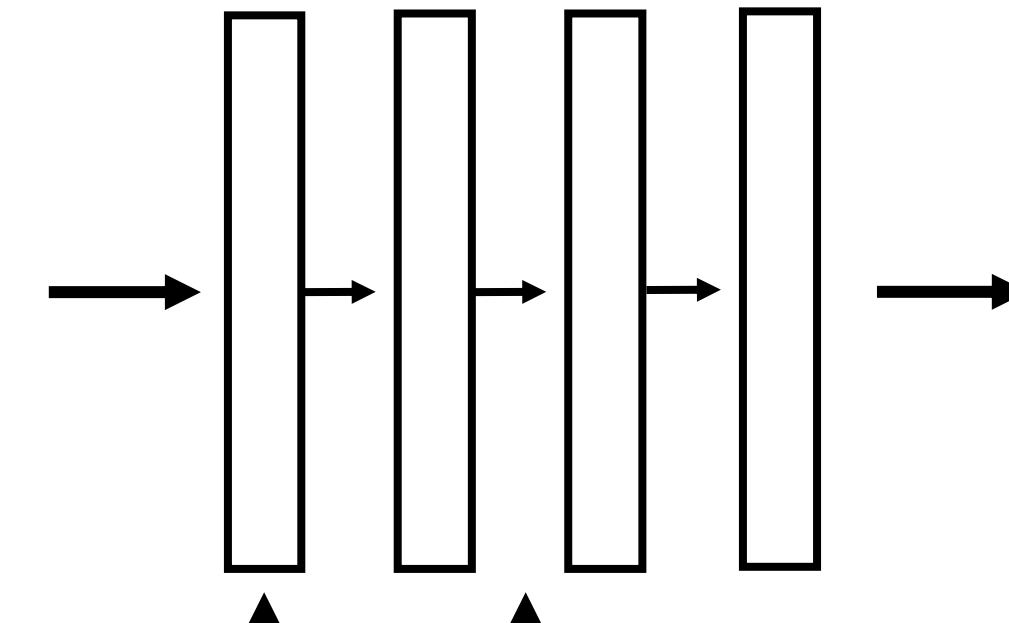
manipulated
image **y**



image **x**



semantic
layout 1



“more flowers”
“more cloudy”

Target transient
scene attribute **t**

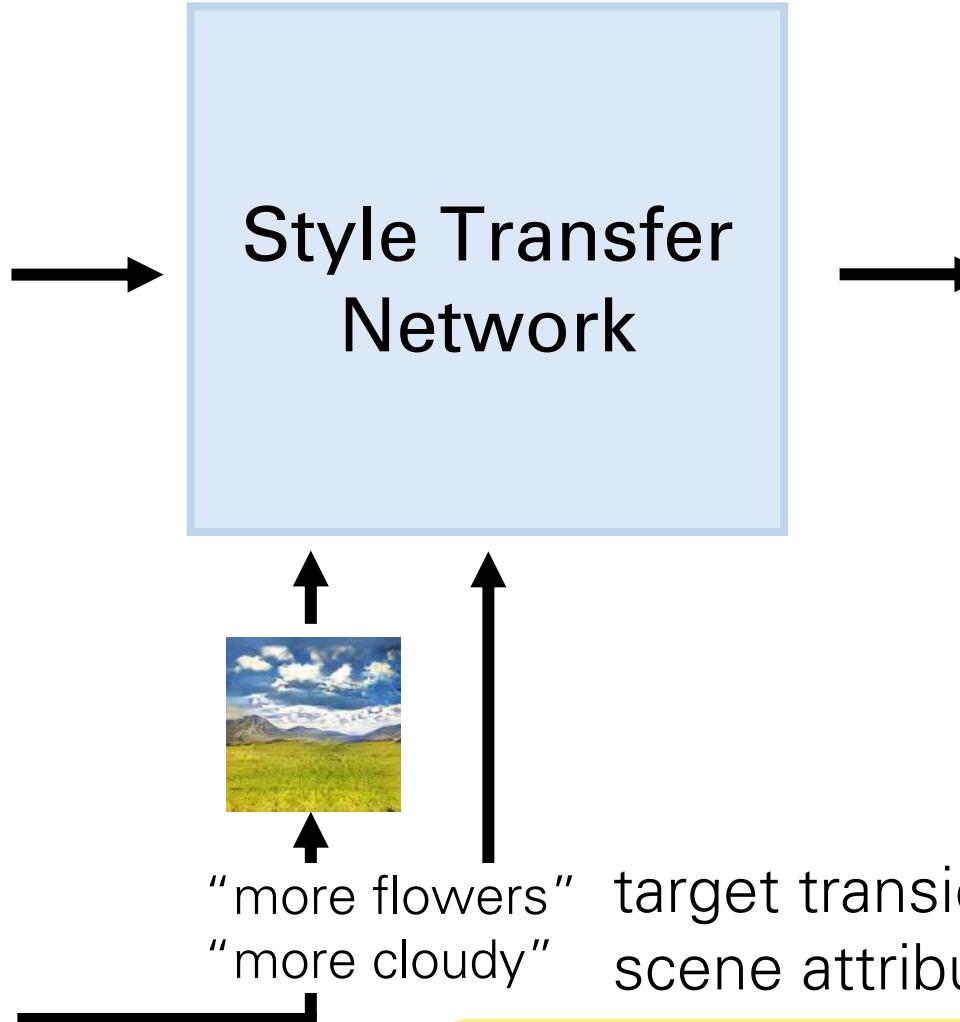
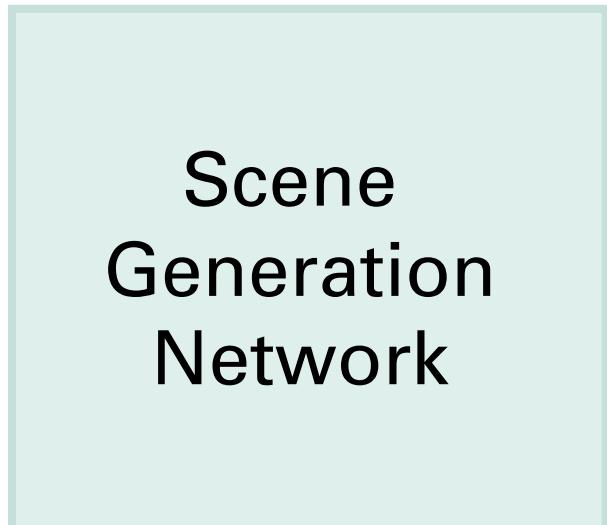
- **Input:** input image + semantic layout + target attribute
- **Output:** manipulated image



manipulated
image **y**

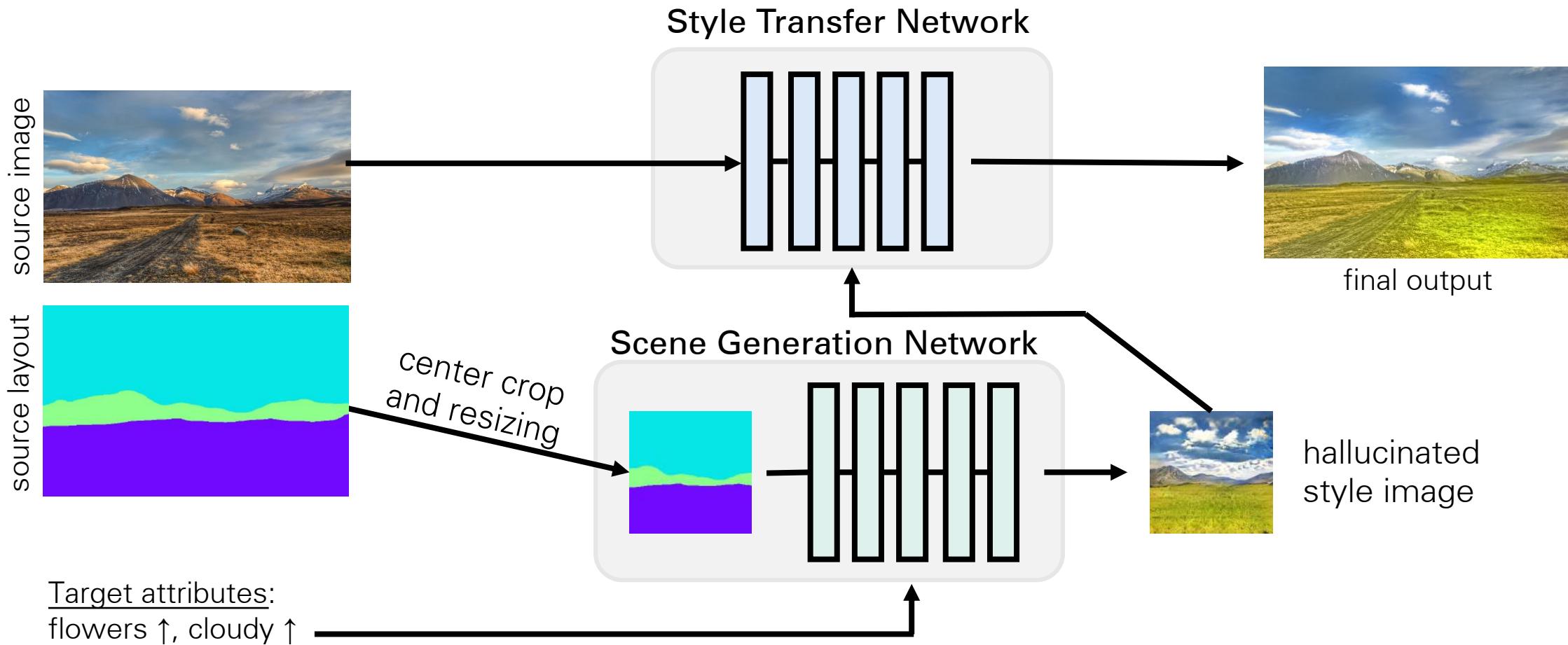


image **x** and
semantic layout 1



manipulated
image **y**

Idea: Hallucinate alternative version of the input scene consistent with target attributes and use this image as the style image in the photo style transfer.

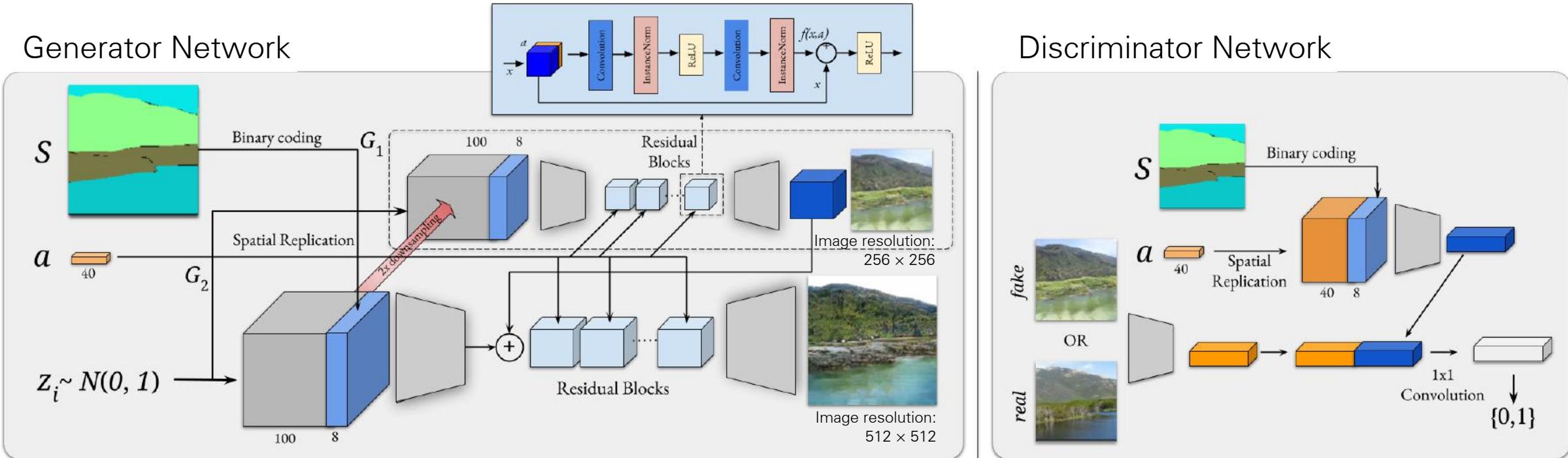


- **Scene Generation Network**
 - A conditioned GAN model with two conditions:
 - (1) semantic layout,
 - (2) target attributes

- **Style Transfer Network**
 - A deep photo style transfer network that modifies the look of the source image based on the hallucinated style image

Scene Generation Network (SGN)

- The semantic layout categories are encoded into 8-bit binary codes
- The transient attributes are represented by a 40-d vector.



- An architecture similar to Pix2pixHD model (Wang et al. 2018)
- **Generator network:** A coarse-to-fine model with 2 generator networks
- **Discriminator network:** A combination of three different discriminator networks operating at an image pyramid of 3 scales

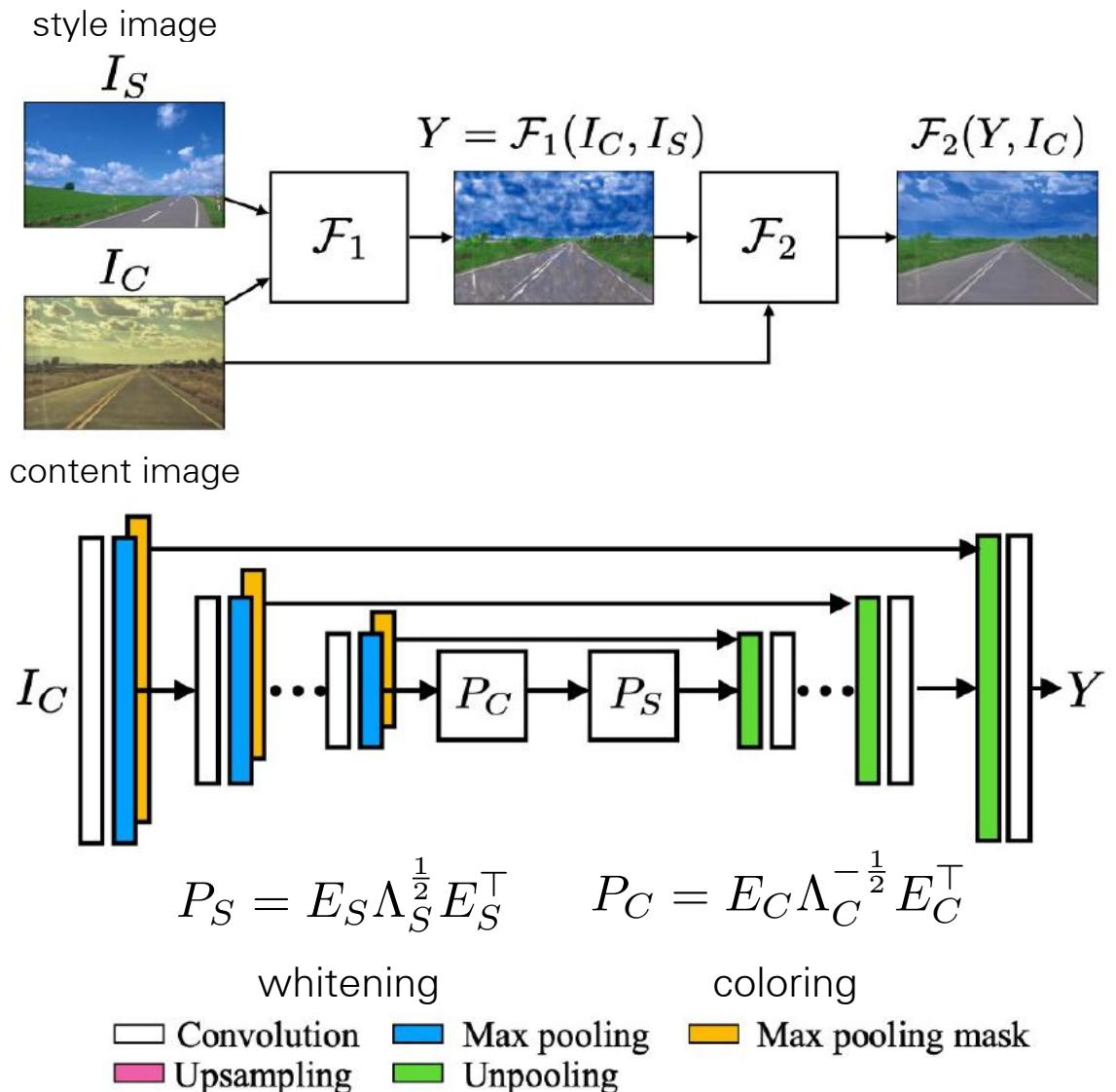
Training Objective of SGNs

$$\mathcal{L}_{SGN} = \min_G \left(\left(\max_{D=\{D_1, D_2, D_3\}} \sum_{k=1,2,3} \mathcal{L}_{GAN}(G, D_k) \right) + \lambda \mathcal{L}_{percep}(G) \right)$$

- **Relative Negative Mining (RNM)**
 - real image, relevant attributes and layout
 - vs.
 - fake image, relevant attributes and layout
 - real image, mismatching layout (chosen from hard negatives)
 - or mismatching attributes
- **Layout-Invariant Perceptual Loss**
 - $\mathcal{L}_{percep}(G) = E_{z \sim p_z(z); x, S, a \sim p_{data}(S, a)} \left[\|f_P(x) - f_P(G(z, a, S))\|_2^2 \right]$
 - f_P : CNN encoder for the scene parser network (Zhou et al., 2018)

Style Transfer Network

- The FPST method of (Li et al., 2018), which is composed of two steps with close-form solutions:
 1. Stylization step \mathcal{F}_1
 2. Smoothing step \mathcal{F}_2
$$I_{out} = \mathcal{F}_2(\mathcal{F}_1(I_C, I_S), I_C)$$
- The **stylization step** is based on the whitening and coloring transform to stylize images via feature projections
 - Style information encoded by the covariance matrix of VGG features
- The **smoothing step** ensures spatially consistent stylizations via a manifold ranking operator.



Our Dataset: ALS18K

- A dataset of 17772 outdoor images with layout and transient attribute labels, formed by combining and annotated images from
 - Transient Attributes dataset (Laffont et al., 2014)
 - ADE20K dataset (Zhou et al., 2017)
- 16434 images for training, 1338 images for testing
- 150 semantic categories
- 40 transient attributes in five categories

bottle
bathtub
clock
radiator
monitor
fan
computer
scorncake
streetlight
mirror
step
column
base
shef
bus
airplane
ship
boat
bicycle
minibike
van
car
cradle
buffet
bookcase
chest of drawers
wardrobe
ottoman
bench
sofa
swivel chair
armchair
coffee table
pool table
counter table
desk
cabinet
bed
escalator
stairs
sidewalk
road
screen
countertop
ceiling
pier
stage
runway
floor
fountain
tower
awning
hovel
bridge
booth
fireplace
signboard
house
bannister
bar
fence
screen door
skyscraper
wall

conveyer belt
traffic light
Poster
trade name
grandstand
lake
waterfall
river
sea
food
water
sand
sky
land
towel
kitchen island
field
hill
mountain
rock
animal
person
grass
palm
flag
plate
swimming pool
air track
plaything
microwave
oven
stove
dishwasher
washer
refrigerator
chandelier
shower
toilet
pillow
book
sculpture
painting
hood
blanket
apparel
rug
curtain
bulletin board
windowpane
pole
ball
television receiver
glass
ashcan
tray
pot
bag
basket
case
box
vase
tank
barrel

lighting: sunrise/sunset, bright, daylight, etc.

weather: sunny, warm, moist, foggy, cloudy, etc.

seasons: spring, summer, autumn, winter

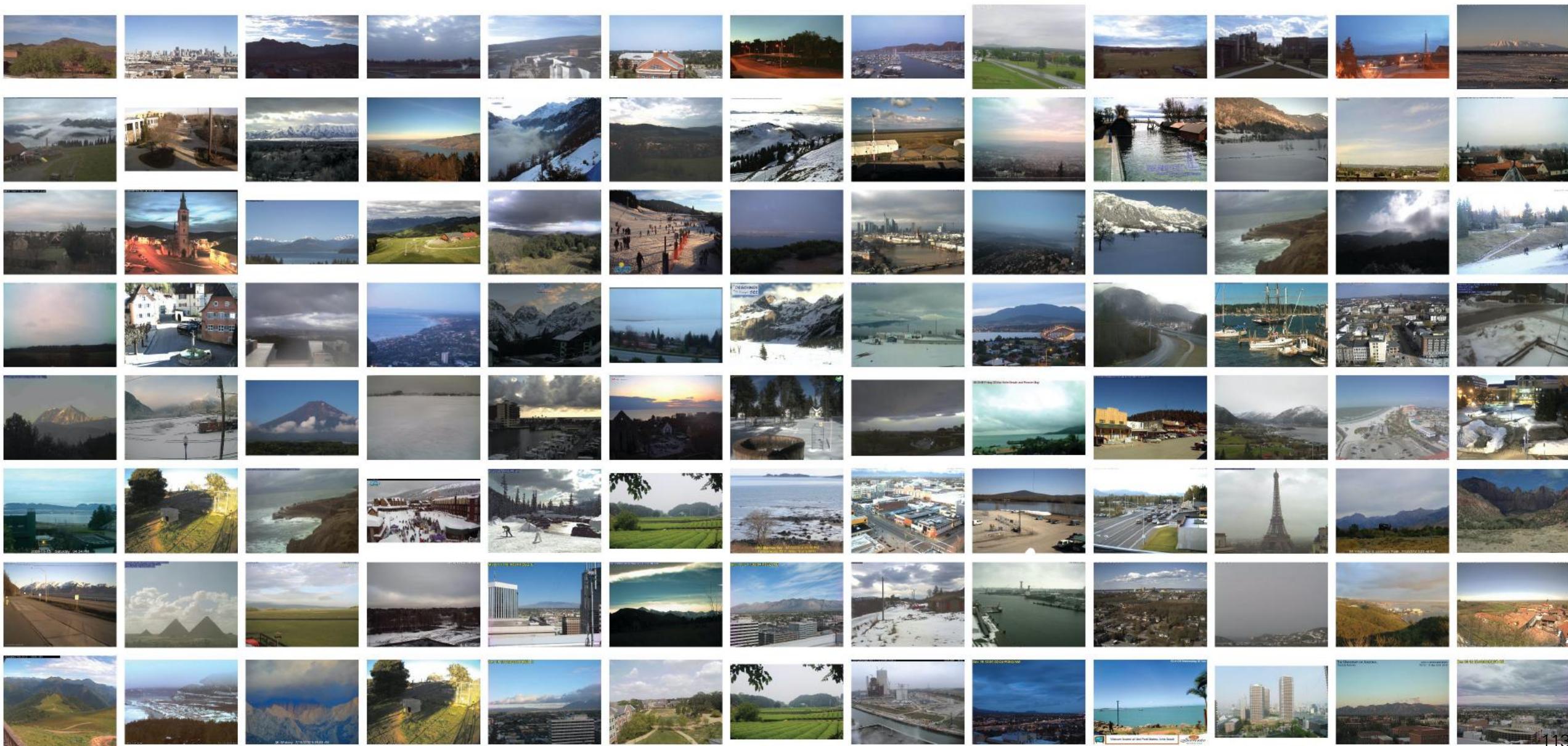
subjective impressions: gloomy, soothing, beautiful, etc.

additional attributes: active/busy, cluttered, dirty/polluted, lush vegetation, etc.



ALS18K | Transient Attributes

- 101 webcams
- 8571 outdoor images



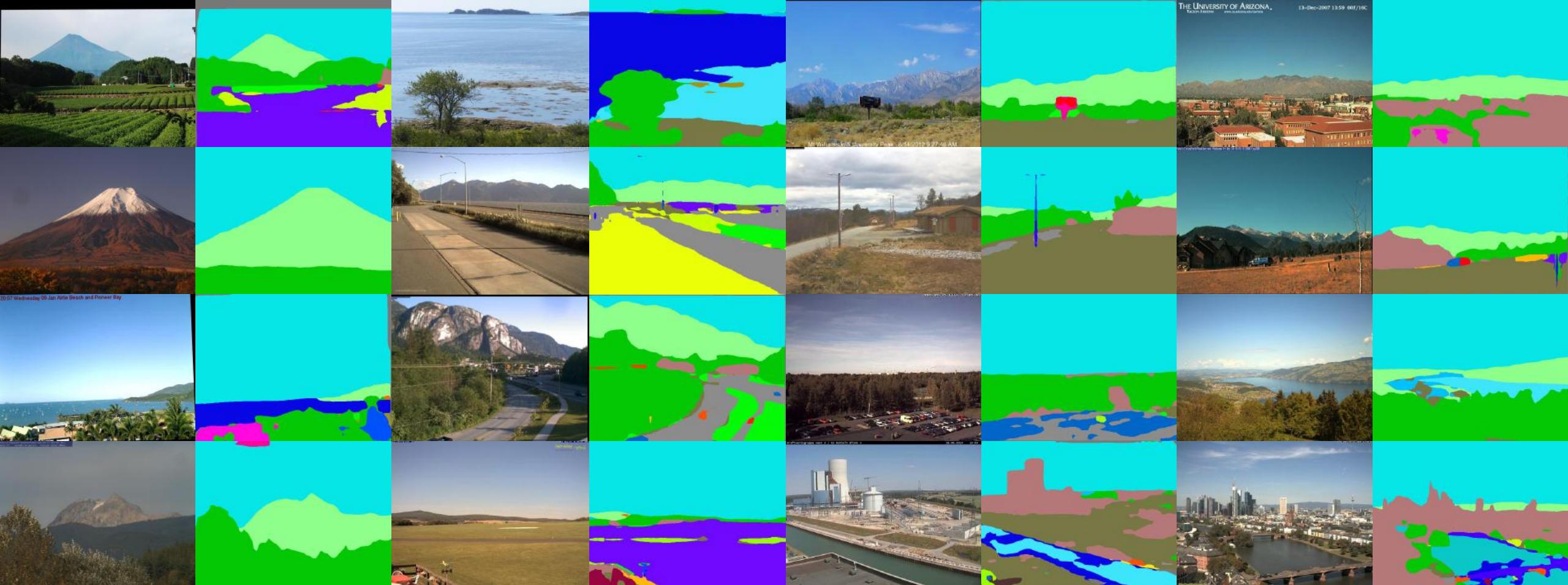
ALS18K | Transient Attributes

- 101 webcams
- 8571 outdoor images



ALS18K | Transient Attributes

- 101 webcams
- 8571 outdoor images

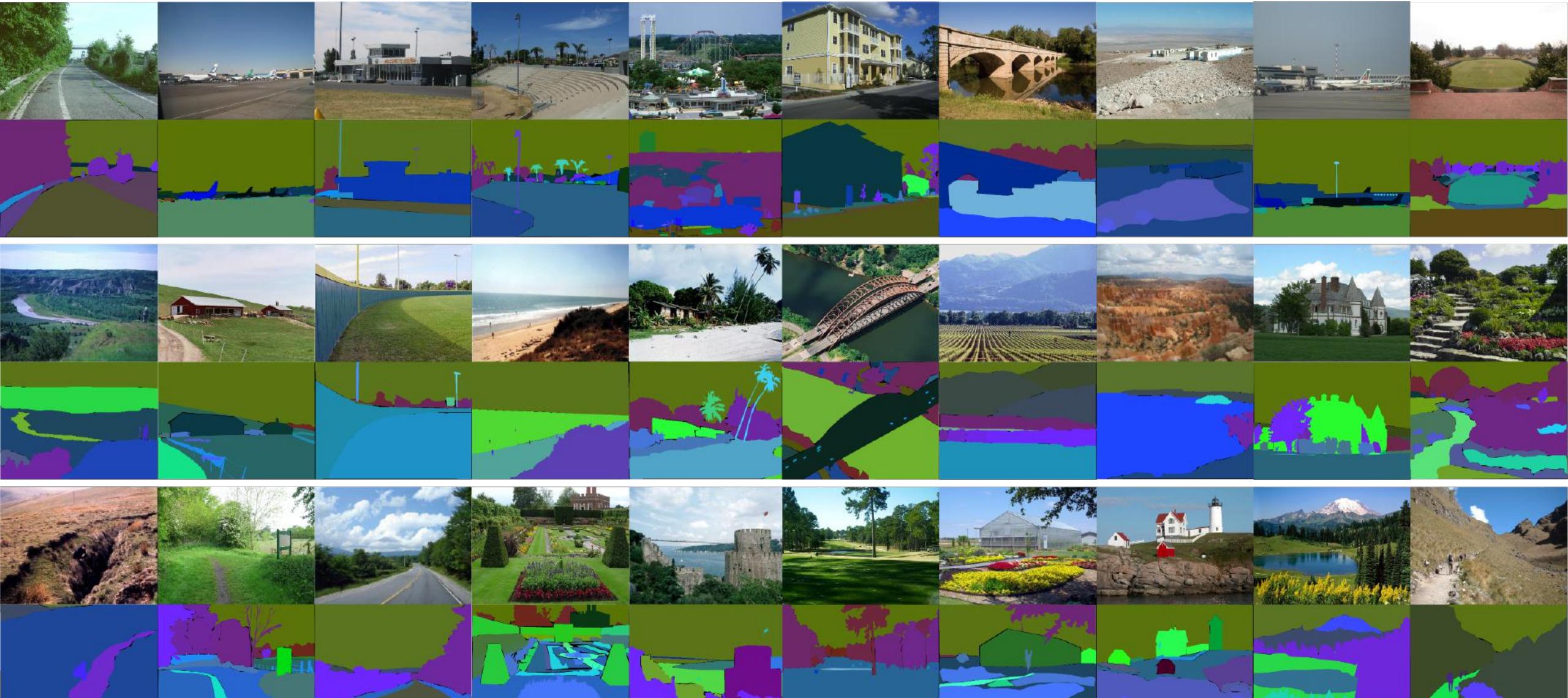


- Originally, no layout information is available!

- Predicted by a semantic parser network and corrected manually.

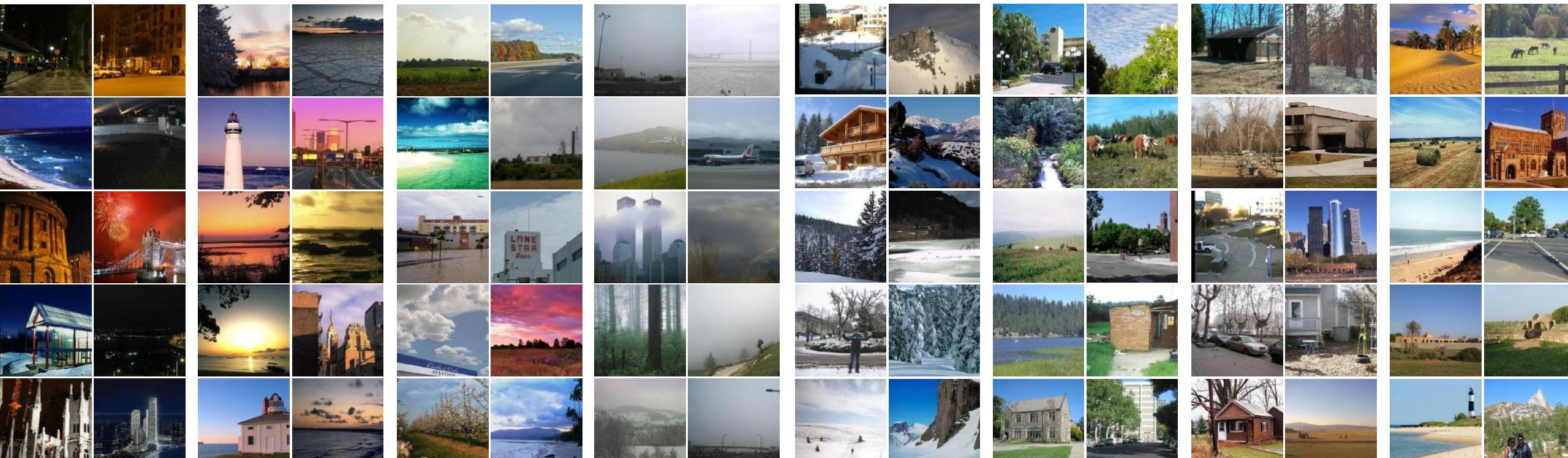
ALS18K | ADE20K

- Considered only 9201 outdoor images
- Layouts , covering 150 semantic categories



ALS18K | ADE20K

- Considered only 9201 outdoor images
- Layouts , covering 150 semantic categories



night

sunset

cloud

fog

snow

spring

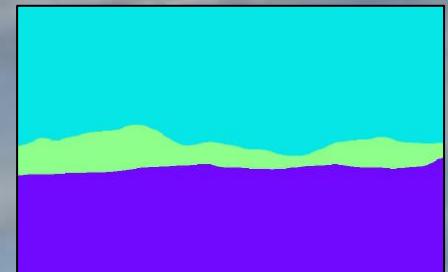
autumn

warm

- Originally, no information exists about transient scene attributes!

- Predicted automatically by an attribute classifier network.

Imagining Things

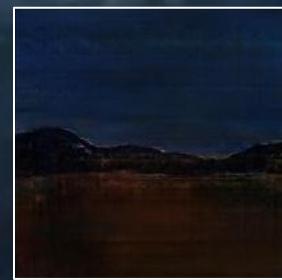


Semantic Layout



Imagining Things

Night



Prediction



Imagining Things

Sunset



Prediction



Imagining Things

Winter



Prediction



Imagining Things

Spring
+
Clouds



Prediction



Imagining Things

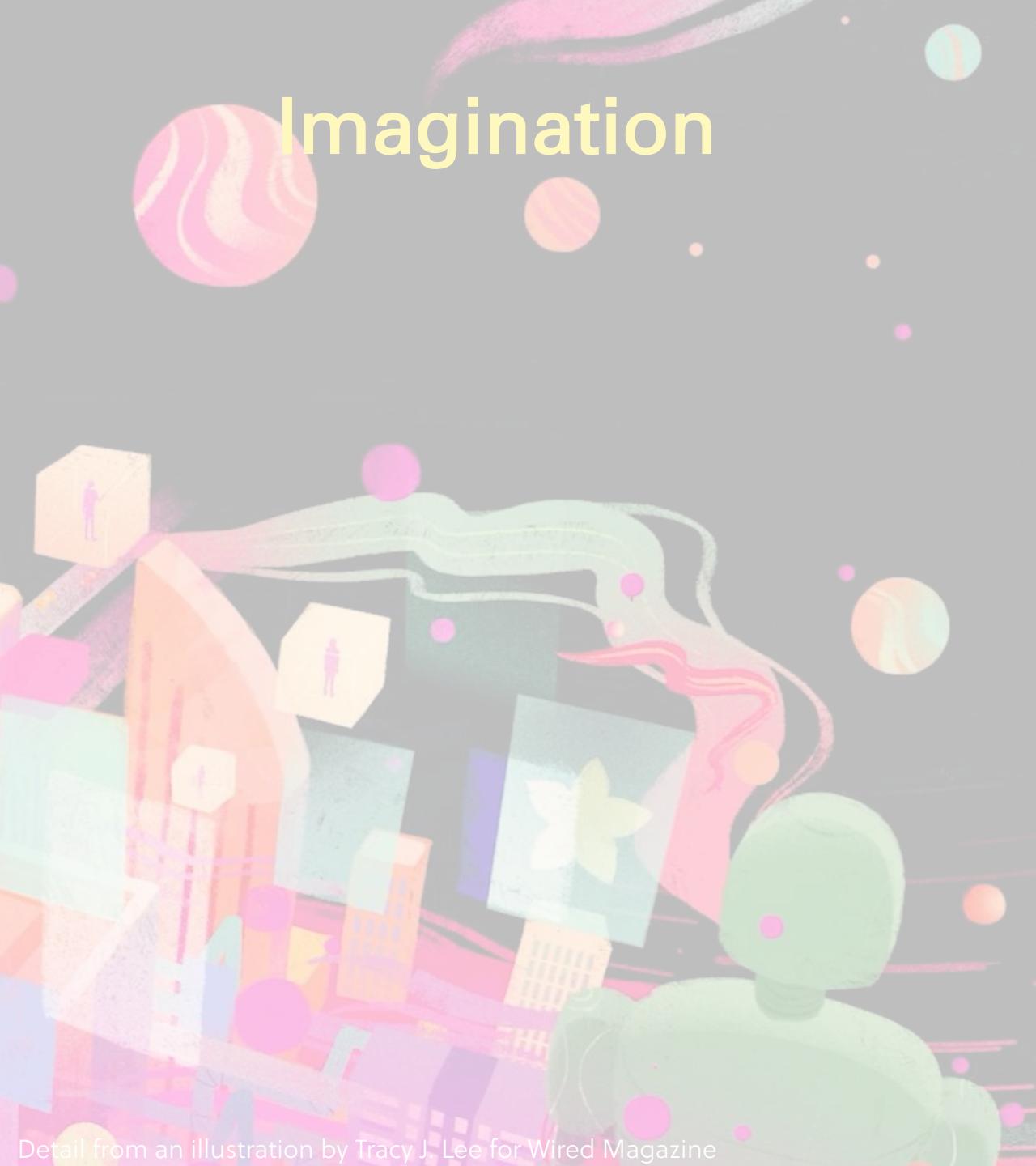
Moist,
Rain &
Fog



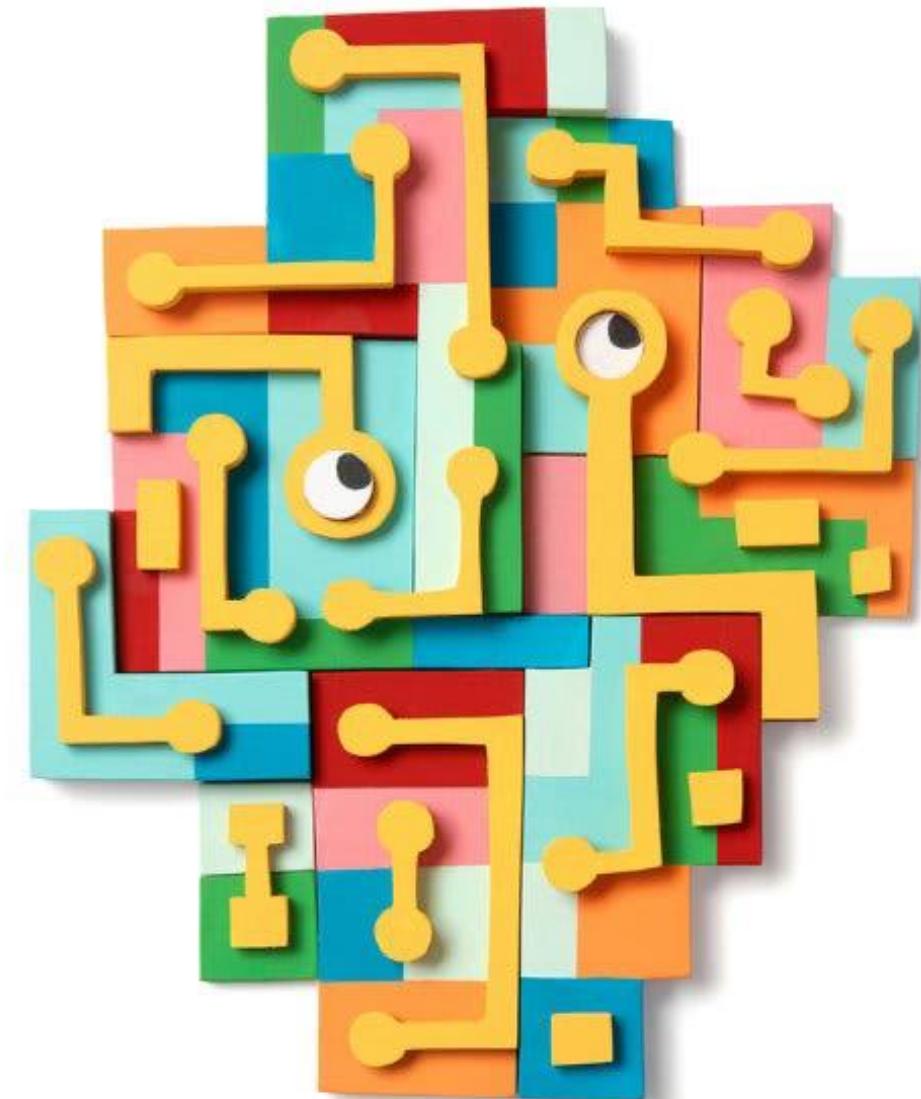
Prediction



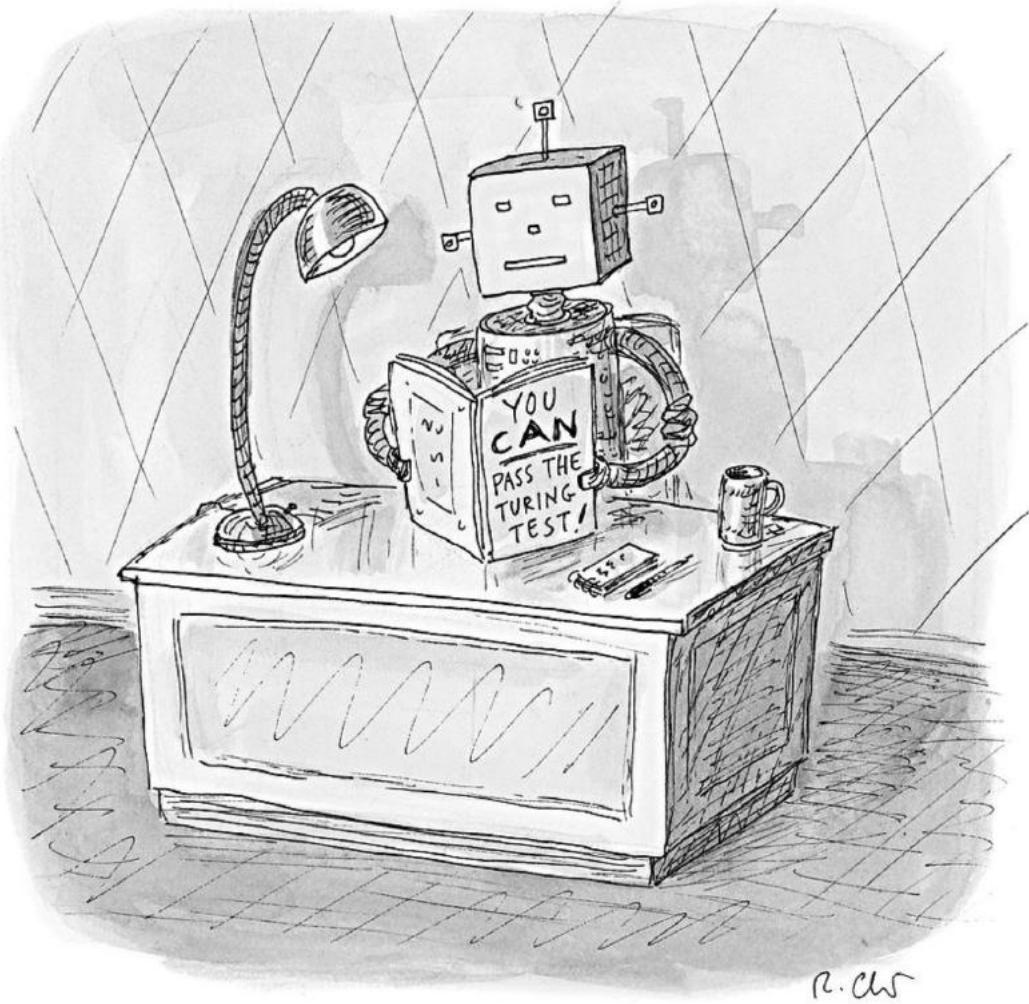
Imagination



Commonsense Knowledge



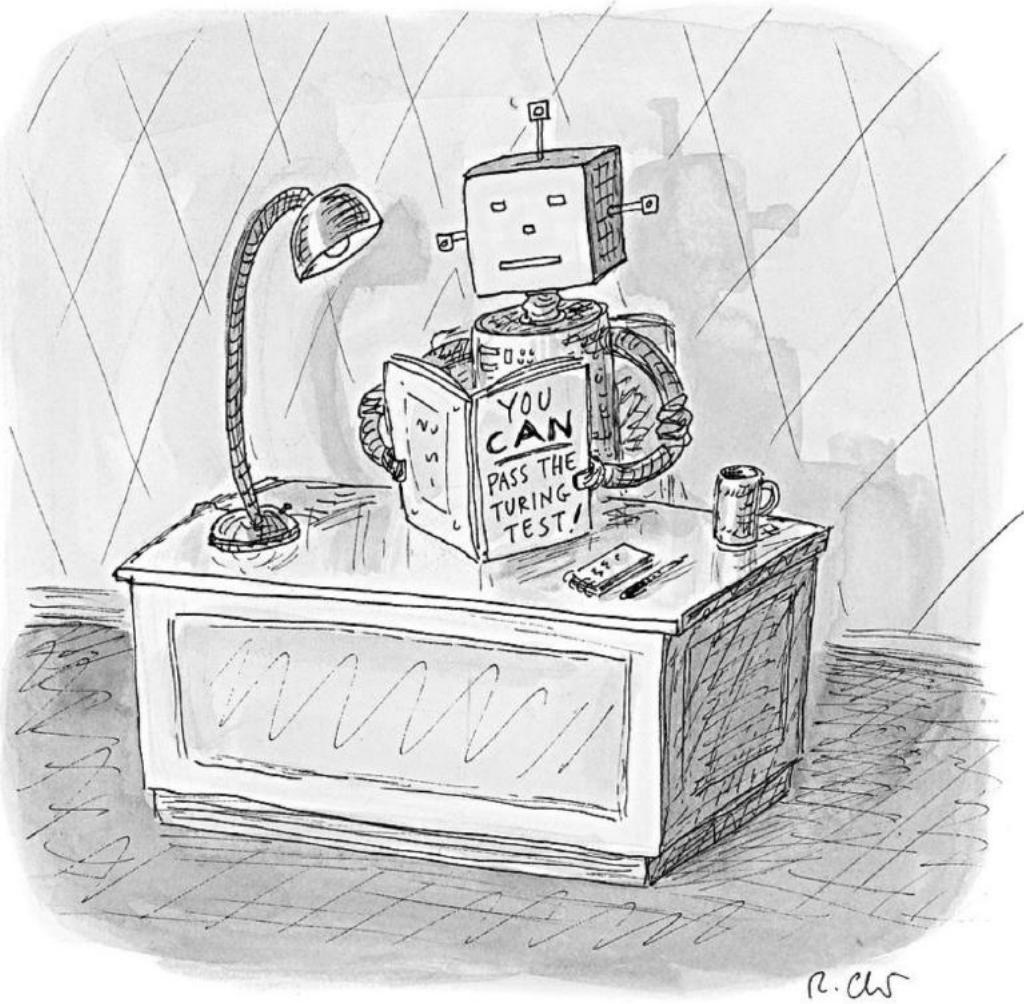
The Resurgence of Neural Networks



1986

“Neural networks simply don’t work!”

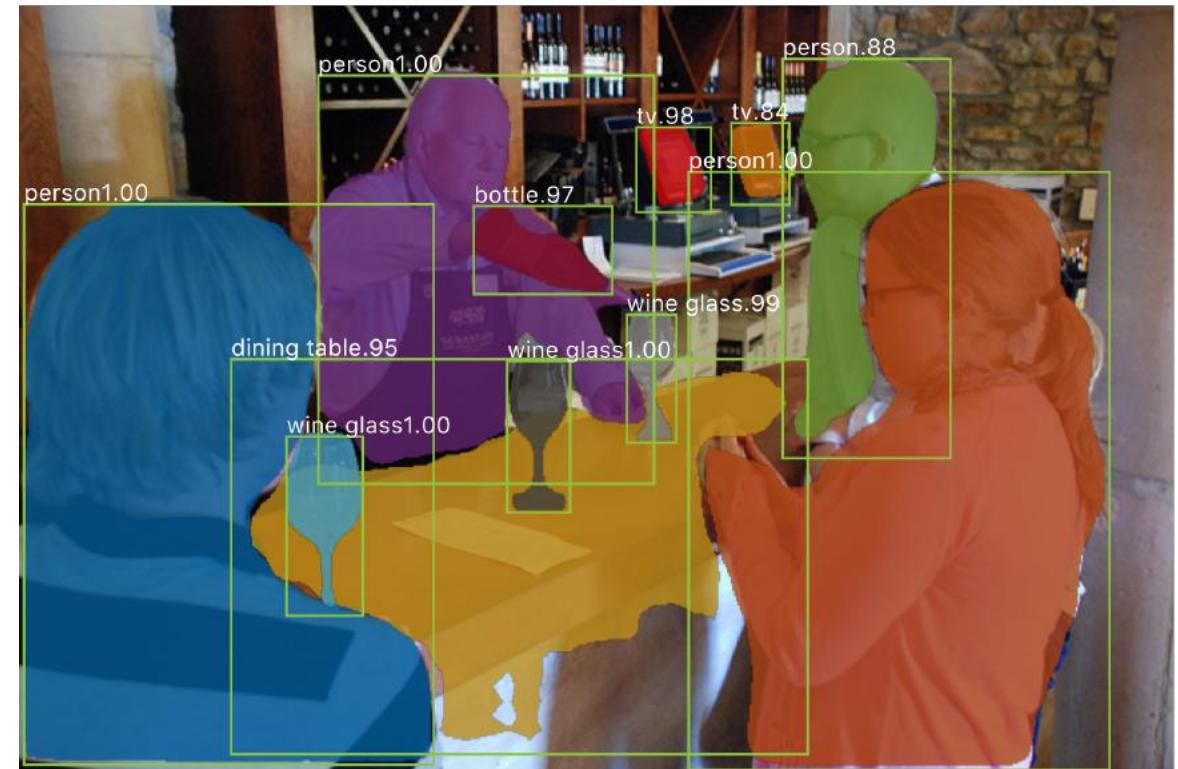
The Resurgence of Neural Networks



1986

2012 2014 2016 2018 2020

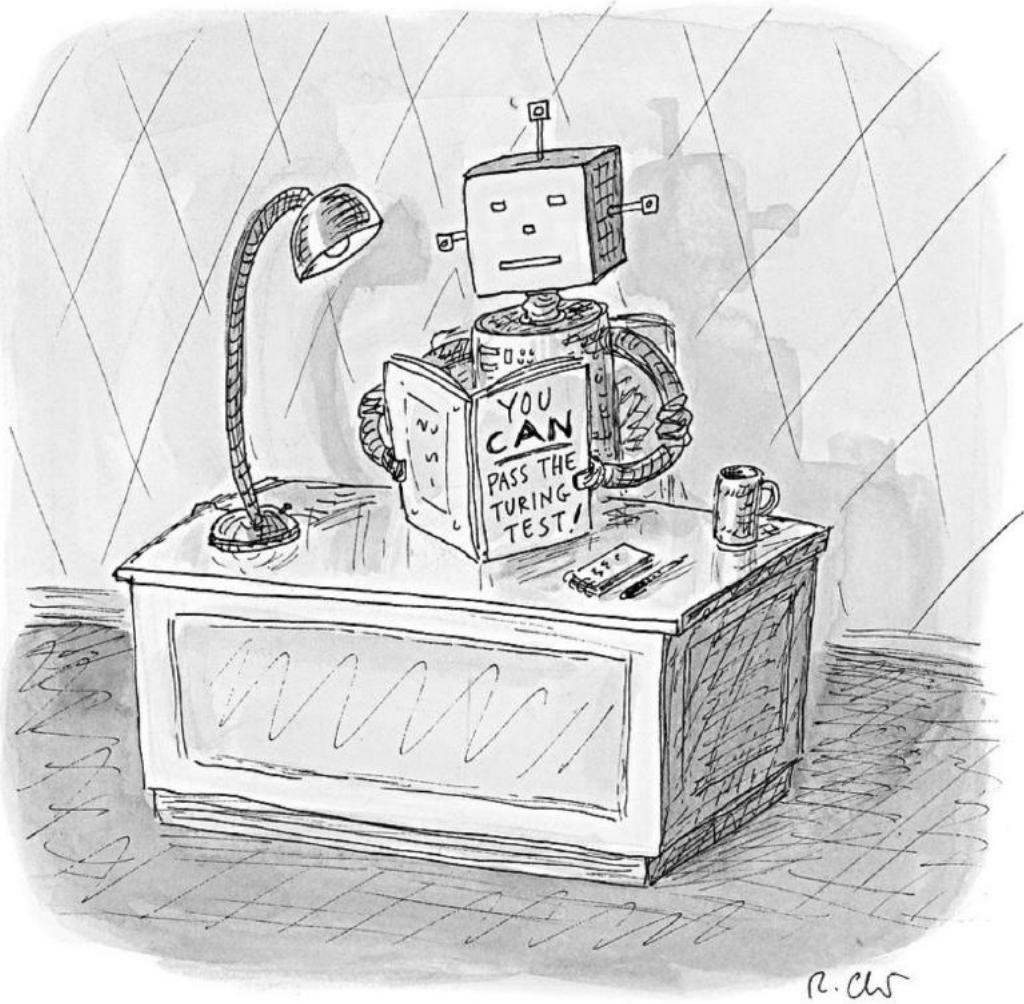
Can neural nets **classify images?**



Cartoon credit: Roz Chast, New Yorker, March 6th 2017

Mask R-CNN [He et al. 2017] detection result

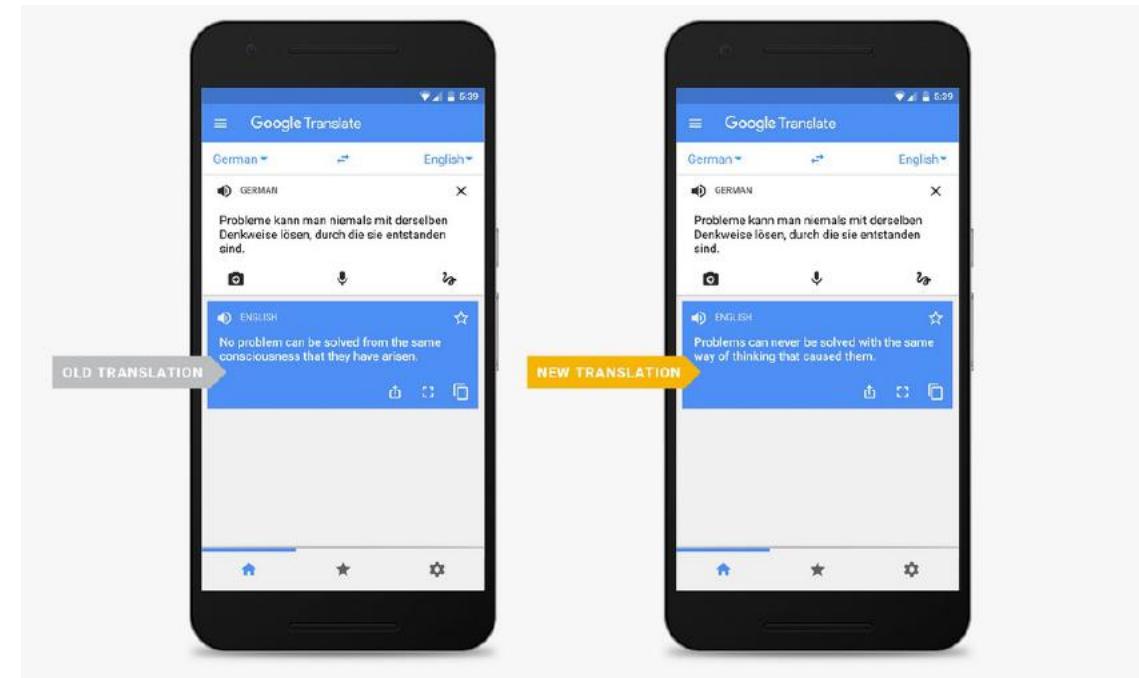
The Resurgence of Neural Networks



1986

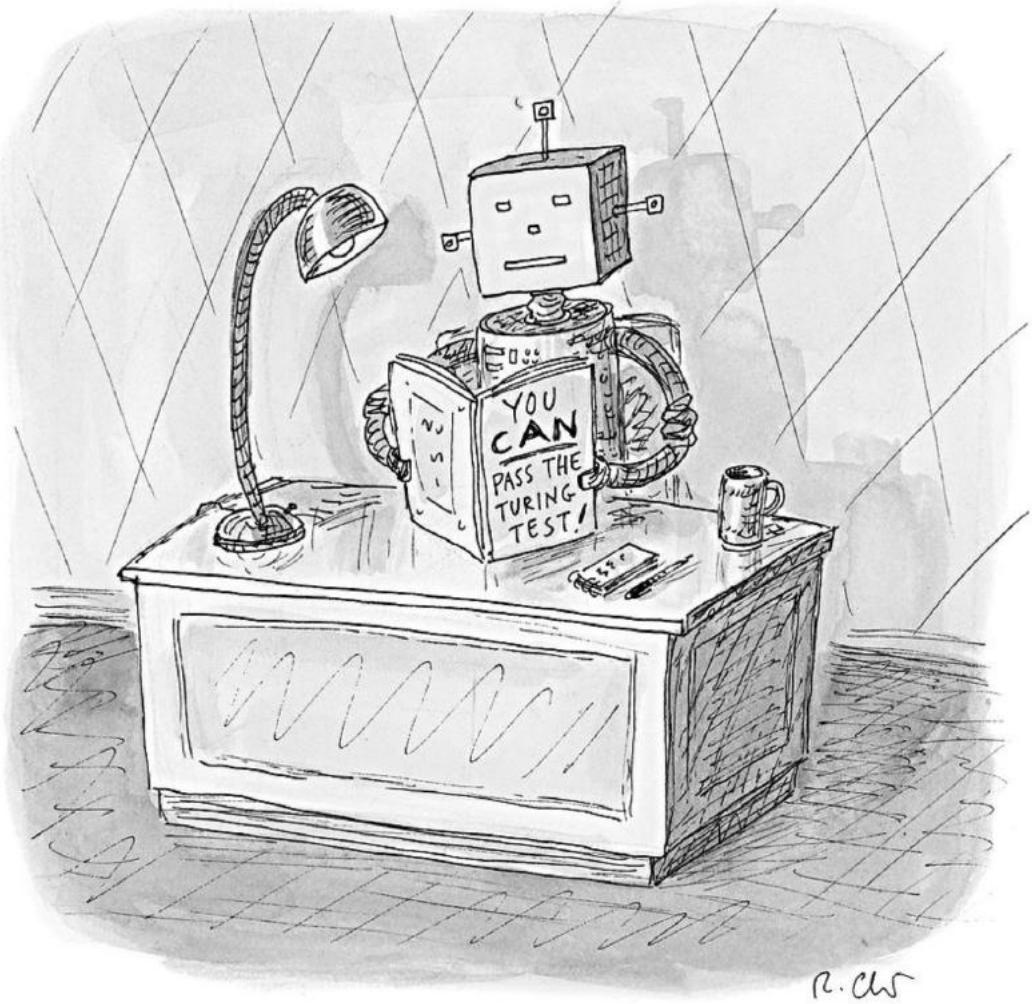
2012 2014 2016 2018 2020

Can neural nets **translate text?**



Sequence to Sequence Neural Machine Translation
[Sutskever et al. 2014, Bahdanau et al. 2014]

The Resurgence of Neural Networks



1986

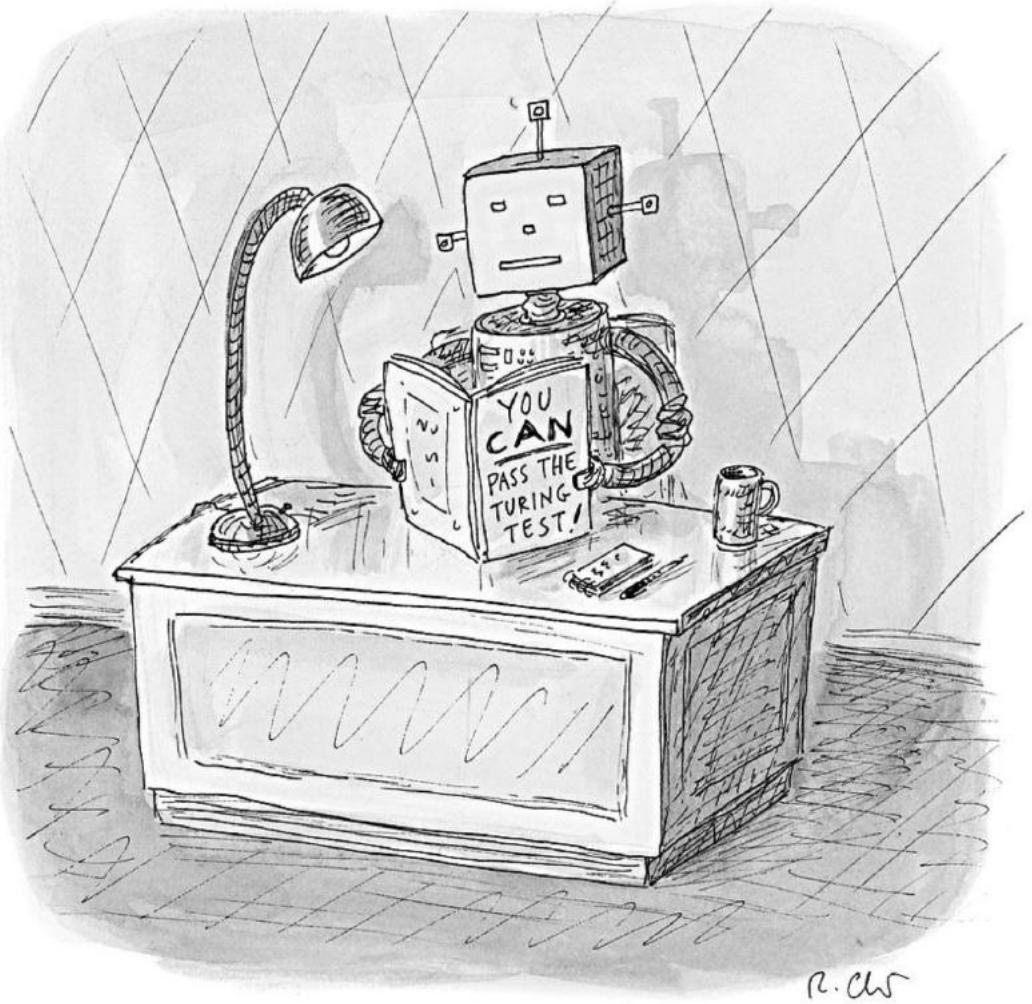
2012 2014 2016 2018 2020

Can neural nets **play Go?**



AlphaGo beats Lee Sedol (March 9-15, 2016)

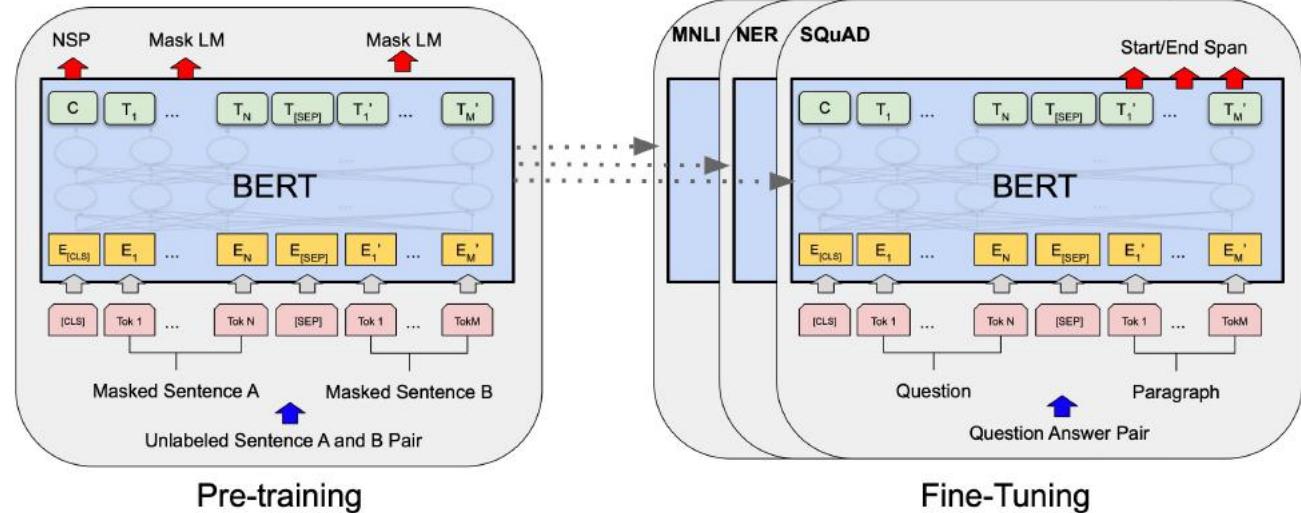
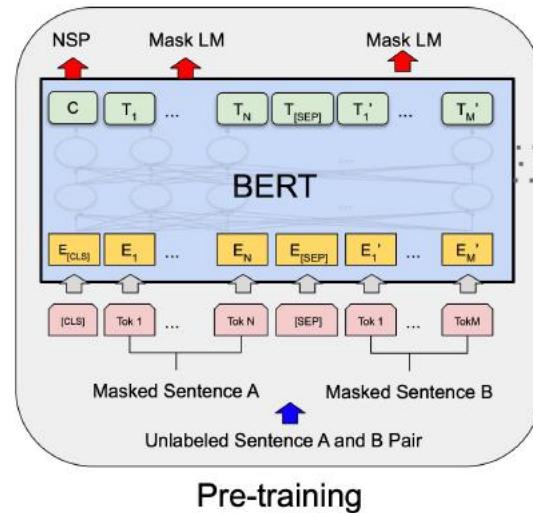
The Resurgence of Neural Networks



1986

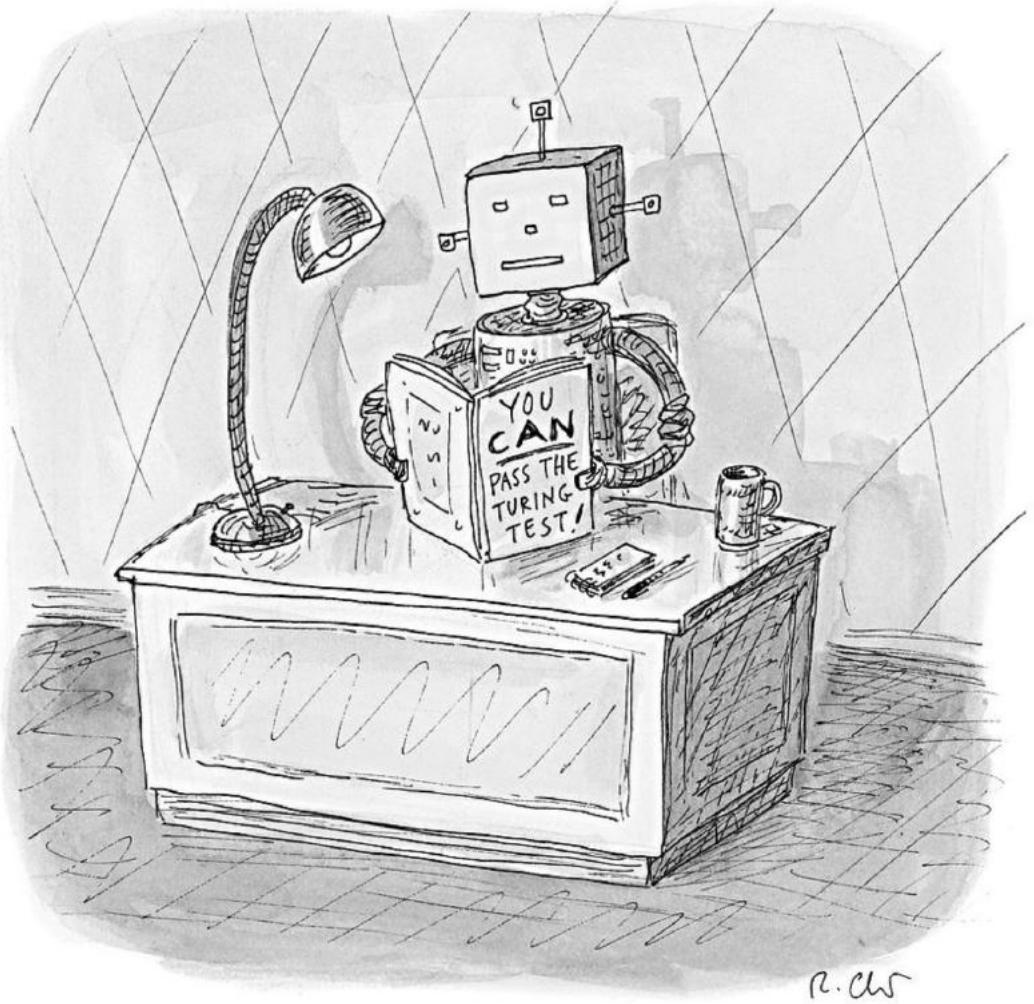
2012 2014 2016 2018 2020

Can neural nets **understand**
encyclopedic knowledge?



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Devlin et al. 2018]

Looking Forward



1986

2012 2014 2016 2018 2020

Can neural nets learn commonsense knowledge?



Types of Knowledge

- Propositional knowledge
 - Who is Bram Stoker?

Information Extraction



IBM Watson vs. Jeopardy

Types of Knowledge

- **Propositional knowledge**
 - Who is Bram Stoker?
- **Commonsense knowledge**
 - the basic level of **practical knowledge** and **reasoning**
 - concerning **everyday situations** and **events**
 - that are **commonly** shared among **most** people.

Information Extraction

Types of Knowledge

- **Propositional knowledge**

- Who is Bram Stoker?

Information Extraction

- **Commonsense knowledge**

- How to do things?

(i) How-to Knowledge



Types of Knowledge

- **Propositional knowledge**
 - Who is Bram Stoker?
- **Commonsense knowledge**
 - How to do things?
 - What other people think or want?

Information Extraction

(ii) Social Norms



© Warneken & Tomasello

Types of Knowledge

- **Propositional knowledge**
 - Who is Bram Stoker?
- **Commonsense knowledge**
 - How to do things?
 - What other people think or want?
 - How the world operates at a physical level?

Information Extraction

(iii) Intuitive Physics



- The trophy doesn't fit in the brown suitcase because it's too small.
What is too small?

Answer 0: the trophy

Answer 1: the suitcase

- Joan made sure to thank Susan for all the help she had received. Who had received the help?

Answer 0: Joan

Answer 1: Susan

The Winograd Schema Challenge

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

Leora Morgenstern
S.A.I.C
Arlington, VA 22203
leora.morgenstern@saic.com

Abstract

In this paper, we present an alternative to the Turing Test that has some conceptual and practical advantages. A Winograd schema is a pair of sentences that differ only in one or two words and that contain a referential ambiguity that is resolved in opposite directions in the two sentences. We have compiled a collection of Winograd schemas, designed so that the correct answer is obvious to the human reader, but cannot easily be found using selectional restrictions or statistical techniques over text corpora. A contestant in the Winograd Schema Challenge is presented with a collection of one sentence from each pair, and required to achieve human-level accuracy in choosing the correct disambiguation.

1 Introduction

The well-known Turing Test was first proposed by Alan Turing (1950) as a practical way to defuse what seemed to him to be a pointless argument about whether or not machines could think. In nutshell, he proposes that instead of asking such a vague question and then getting caught up in a debate about what it means to really be thinking, we should focus on *observable behaviour* and ask whether a machine would be capable of producing behaviour that we would say required thought in people. The sort of behaviour he had in mind was participating in a natural conversation in English over a teletype in what he calls the Imitation Game. The idea, roughly, is that if an interrogator were unable to tell after a long, free-flowing and unrestricted conversation with a machine whether she was dealing with a person or a machine, then we should be prepared to say that the machine was thinking. Requiring more of the machine, such that as that it looks a certain way, or biologically, or have a certain causal history, is just arbitrary chauvinism.

It is not our intent to defend Turing's argument here (but see the Discussion section below). For our purposes, we simply accept the argument and the emphasis Turing places on intelligent behaviour, counter to critics such as Searle (2008). We also accept that typed English text is a sufficient medium for displaying intelligent behaviour, counter to critics such as Harnad (1989). That is, assuming that *any* sort of behaviour is going to be judged sufficient for show-

ing the presence of thinking (or understanding, or intelligence, or whatever appropriate mental attribute), we assume that typed English text, despite its limitations, will be a rich enough medium.

2 The trouble with Turing

The Turing Test does have some troubling aspects, however. First, note the central role of *deception*. Consider the case of a future intelligent machine trying to pass the test. It must converse with an interrogator and not just show its stuff, but fool her into thinking she is dealing with a *person*. This is just a game, of course, so it's not really lying. But to imitate a person well without being evasive, the machine will need to assume a false identity (to answer "How tall are you?" or "Tell me about your parents."). All other things being equal, we should much prefer a test that did not depend on chicanery of this sort. Or to put it differently, a machine should be able to show us that it is thinking without having to pretend to be somebody or to have some property (like being tall) that it does not have.

We might also question whether a *conversation* in English is the right sort of test. Free-form conversations are no doubt the best way to get to know someone, to find out what they think about something, and therefore *that* they are thinking about something. But conversations are so adaptable and can be so wide-ranging that they facilitate deception and trickery.

Consider, for example, ELIZA (Weizenbaum 1966), where a program (usually included as part of the normal Emacs distribution), using very simple means, was able to fool some people into believing they were conversing with a psychiatrist. The deception works at least in part because we are extremely forgiving in terms of what we will accept as legitimate conversation. A Rogerian psychiatrist may say very little except to encourage a patient to keep on talking, but it may be enough, at least for a while.

Consider also the Loebner competition (Shieber 1994), a restricted version of the Turing Test that has attracted considerable publicity. In this case, we have a more balanced conversation taking place than with ELIZA. What is striking about transcripts of these conversations is the fluidity of the responses from the subjects: elaborate wordplay, puns, jokes, quotations, clever asides, emotional outbursts, points of order. Everything, it would seem, except clear and direct

Types of Knowledge

- **Propositional knowledge**

- Who is Bram Stoker?

- **Commonsense knowledge**



RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes

Semih Yagcioglu, Aykut Erdem, Erkut Erdem and Nazli Ikizler-Cinbis

Hacettepe University Computer Vision Lab

Dept. of Computer Engineering, Hacettepe University, Ankara, TURKEY

semih.yagcioglu@hacettepe.edu.tr, {aykut,erkut,nazli}@cs.hacettepe.edu.tr

[Yagcioglu et al., EMNLP 2018]

Information Extraction

(i) How-to Knowledge

(ii) Social Norms

(iii) Intuitive Physics

Types of Knowledge

- Propositional knowledge

- Who is Bram Stoker?

- Commonsense knowledge

 **CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions**

Tayfun Ates^{1,†}, Muhammed Samil Atesoglu¹, Cagatay Yigit¹, Ilker Keser², Mert Kobas³, Erkut Erdem¹, Aykut Erdem², Tilbe Goksun³, Deniz Yuret²

¹ Department of Computer Engineering, Hacettepe University, Ankara, Turkey

² Department of Computer Engineering, Koç University, Istanbul, Turkey

³ Department of Psychology, Koç University, Istanbul, Turkey

[†]Correspondence: tates@hacettepe.edu.tr

018]

Information Extraction

(i) How-to Knowledge

(ii) Social Norms

(iii) Intuitive Physics



RecipeQA

RecipeQA: A Challenge Dataset for Multimodal Comprehension
of Cooking Recipes

Semih Yagcioglu, Aykut Erdem, Erkut Erdem & Nazli Ikizler-Cinbis
EMNLP 2018



Why Use Cooking Recipes?

Roasted Butternut Squash

Ingredients

- a butternut squash
- olive oil
- salt & pepper
- (optional but tasty) fresh sage

Instructions

1. Peel your squash and remove the seeds
2. Cut into small pieces and pop in the oven
3. Roast at 400 F° for 25-30 minutes, or until soft.



procedural
knowledge

Why Use Cooking Recipes?

Roasted Butternut Squash

Ingredients

- a butternut squash
- olive oil
- salt & pepper
- (optional but tasty) fr

Cut
what?

Instructions

1. Peel your squash and remove the seeds
2. Cut into small pieces and pop in the oven
3. Roast at 400 F° for 25-30 minutes, or until soft.

procedural
knowledge

Why Use Cooking Recipes?

Roast

Ingred

- a b

- olive oil

- salt & pepper

- (optional but tasty) fr

Instructions

1. Peel your squash and remove the seeds
2. **Cut into small pieces** and pop in the oven
3. Roast at 400 F° for 25-30 minutes, or until soft.



visual grounding

Cut
what?

procedural
knowledge

Why Use Cooking Recipes?

Roast



Ingredients



• a butternut



• olive oil



• salt & pepper



• (optional but tasty) fr

Instructions

1. Peel your squash and remove seeds.
2. Cut into small pieces and place in a roasting pan.
3. Roast at 400 F° for 25-30 minutes, or until soft.

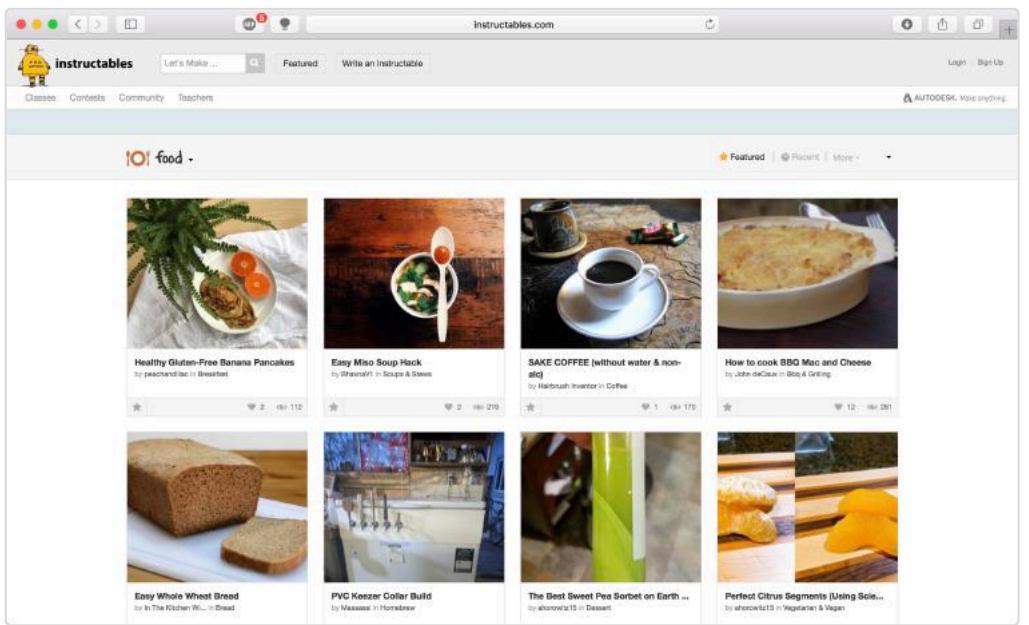
Cut
what

(Commonsense)
knowledge to fill in
the gap..

knowledge

RecipeQA

- We crawled the food category of instructables.com
- Recipes with **more than 2 steps**
 - Each step has a title, a passage and include zero or more images
- Each recipe is **copyright free!**



~20k recipes

Four Types of Multiple Choice Questions

- Visual Cloze
- Text Cloze
- Coherence
- Ordering
- 36K multiple choice questions

		train	val	test
	# of recipes	15847	1963	1969
	... avg. # of steps	5.99	6.01	6.00
	... avg. # of tokens (titles)	17.79	17.40	17.67
	... avg. # of tokens (descr.)	443.01	440.51	435.33
	... avg. # of images	12.67	12.74	12.65
	# of question-answers	29657	3562	3567
	... textual cloze	7837	961	963
	... visual cloze	7144	842	848
	... visual coherence	7118	830	851
	... visual ordering	7558	929	905

An Example Ordering Question

P

Step 1: Gather Supplies

To make Turkish style coffee, you will need the following: A Turkish Ibrik/Cezve You can find them all over the web. The ibrik is the main tool you will need to brew Turkish coffee. It is also called a cezve. A hot stove Ground Turkish coffee. I used normal coffee beans, but what makes it Turkish is the way it is ground. Go to the supermarket, and the machine should have a setting where you can grind it "Turkish." It was the finest setting at my supermarket. Turkish ground coffee should be very fine. Of course, real Turkish coffee would be super, but we are in a recession.

Step 2: Prepping

If you want your coffee to be sweeter, then first add a little bit of sugar to ibrik. Next, take your ibrik and fill it to the start of the neck with water. Sometimes, there is something indicating where the the water line should be, but on mine I just had to guess. By not filling the ibrik up to the top, you give room for the foam to rise, which I will describe later. Once you have your water in the ibrik, add the coffee grounds to the ibrik. Now, for the amount of grounds you add. I added two heaping teaspoons. There is really no definite amount, because some people like it strong, others not so strong. I like mine strong, so I used two heaping teaspoons. When you add the grounds to the ibrik, make sure you just drop them right on the top. Do not stir, just keep the grinds laying on the top. The grinds form a seal between the air and the water, so it takes longer to boil, and results in the grinds foaming at the top.

Step 3: First Boil

Notice how I said first boil. There are going to be multiple boils. Place it on a low heat. The slower, the better. After a couple minutes, pay very close attention to the ibrik. The grounds will start to boil and foam up. Once they begin to foam and fill up the neck, remove the ibrik from the heat source. Now, let the foam settle.

Step 4: Second, Third, Fourth Boil

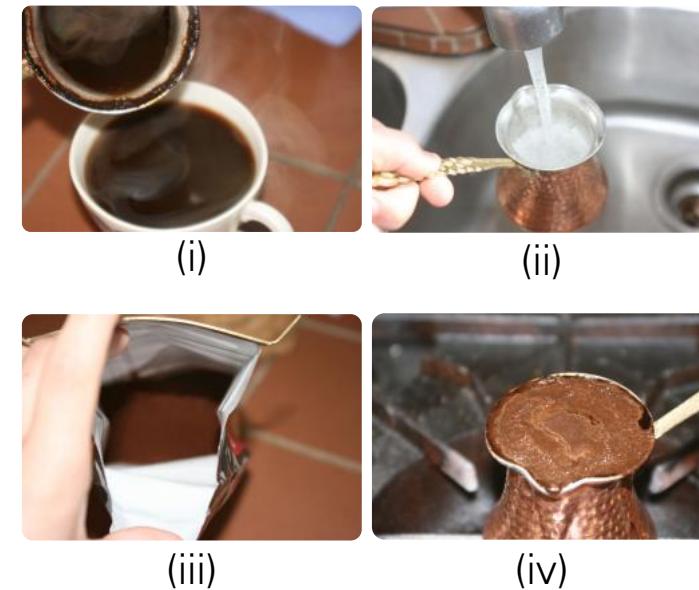
Once the froth has settled from the first boil, place the ibrik back onto the heat source. Let the grounds foam up again, and remove it from the heat source just like you did with the first boil. The third and fourth boils are completely optional. They may add a little more flavor to the coffee, and a little more heat. Once you have completed all of your boils, there will be a little foam on the top of your ibrik, but most of the grounds have floated to the bottom where they will usually stay.

Step 5: Serve!

Time to serve! As you can see on your ibrik, the bottom is curved out. This keeps the grounds from coming out when you pour your coffee. I add cream to my coffee, maybe a little sugar if you didn't before you started. Take the foam, and place it on top of your drink. It is very tasty as well! Enjoy!



How to brew Turkish Coffee



Q

Question: What is the correct order of the images?

A

Answer: (iii) - (ii) - (iv) - (i)

Baseline Results

	Visual Cloze	Textual Cloze	Visual Coherence	Visual Ordering
Hasty Student	27.35	26.89	65.80	40.88
Impatient Reader (Text only)	—	28.03	—	—
Impatient Reader (Multimodal)	27.36	29.07	28.08	26.74
Human*	77.60	73.60	81.60	64.00

* Based on 100 questions sampled randomly from the validation set.

Baseline Results

	Visual Classification	Text Classification	Visual Reasoning	Text Reasoning	Visual Classification
Hasty Student	27.36	29.07	28.08	26.74	27.36
Impatient Reader (Text only)					
Impatient Reader (Multimodal)	27.36	29.07	28.08	26.74	27.36
Human*	77.60	73.60	81.60	64.00	77.60

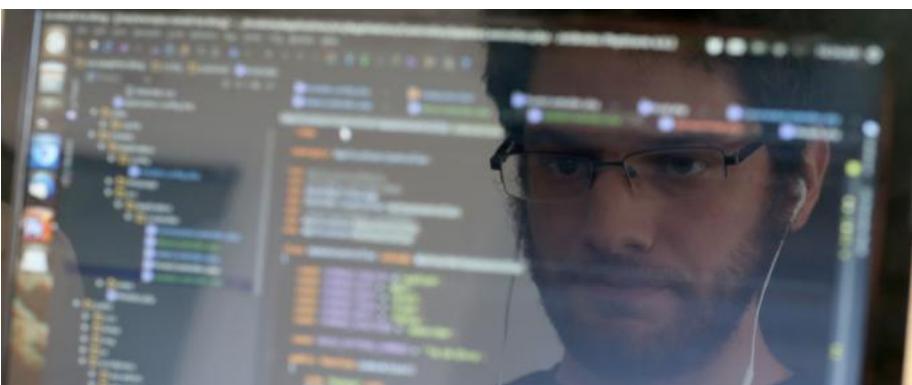
There is large gap (>45)
between humans and
our neural baseline!

* Based on 100 questions sampled randomly from the validation set.

Cookbooks, Wikipedia, and auto-generated Spanglish: The quirky ways AI researchers gather data

Here are four of the most creative data collection methods used by experts at the leading annual conference on natural-language processing.

by Karen Hao November 5, 2018



MIT Technology Review

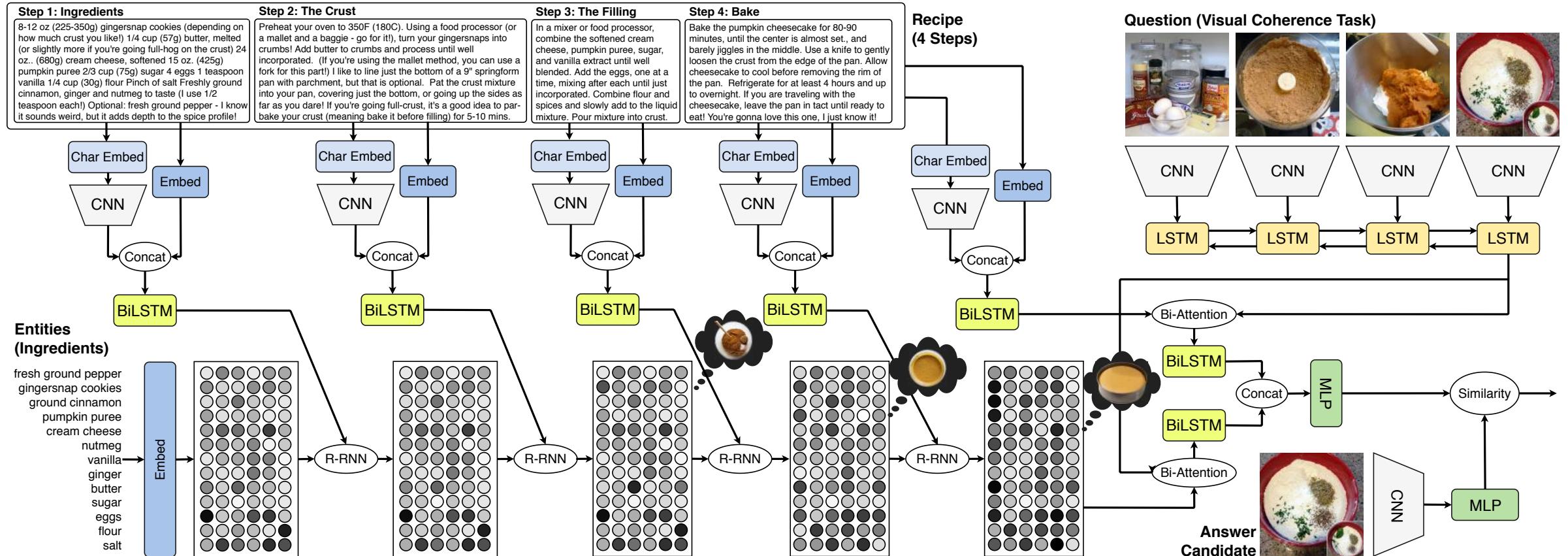
COOKBOOKS

Recipes are great for making food, but they can also provide nourishment for machines. They all follow a similar step-by-step pattern, and they often include pictures that correspond with the text—an excellent source of structured data for teaching machines to comprehend text and images at the same time. That's why researchers at Hacettepe University in Turkey **compiled a giant data set** of around 20,000 illustrated cooking recipes. They hope it will be a new resource for benchmarking the performance of joint image-text comprehension.

What they call “RecipeQA” will build on previous research that has focused on machine reading comprehension and visual comprehension separately. In the former, the machine must understand a question and a related passage to find the answer; in the latter, it searches for the answer in a related photo instead. Having text and photos side by side increases the complexity of the task because the photos and text may share complementary or redundant information.

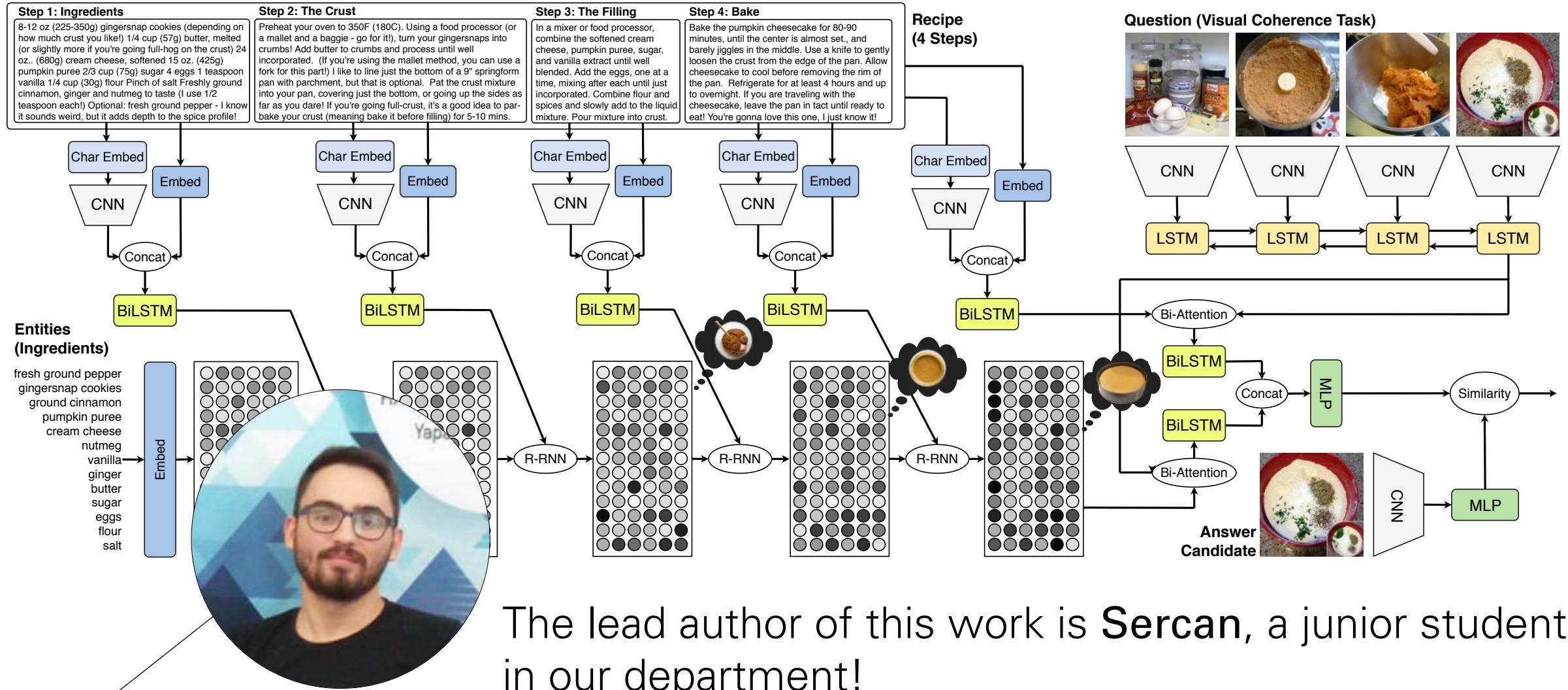
Featured on MIT Technology Review's webpage as one of four of the most creative datasets presented at EMNLP 2018!

Procedural Reasoning Networks



- The model uses an **explicit memory unit** to keep track of the state changes of the ingredients

Procedural Reasoning Networks



The lead author of this work is **Sercan**, a junior student in our department!

Procedural Reasoning Networks

Model	Single-task Training				Multi-task Training				All
	Cloze	Coherence	Ordering	Average	Cloze	Coherence	Ordering		
Hasty Student	27.35	65.80	40.88	44.68	—	—	—	—	
Impatient Reader	27.36	28.08	26.74	27.39	—	—	—	—	
BiDAF	53.95	48.82	62.42	55.06	44.62	36.00	63.93	48.67	
BiDAF w/ static memory	51.82	45.88	60.90	52.87	47.81	40.23	62.94	50.59	
PRN	56.31	53.64	62.77	57.57	46.45	40.58	62.67	50.17	
Human*	77.60	81.60	64.00	74.40	—				

* Based on 100 questions sampled randomly from the validation set.

```

graph TD
    Input[Image of a bowl of soup] --> CNN[CNN]
    CNN --> LSTM[LSTM]
    LSTM --> MLP[MLP]
    MLP --> Bilinear[bilinenarity]
    Bilinear --> Scores[Multi-task Training Scores]
  
```



CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions

Tayfun Ates, Muhammed Samil Atesoglu, Cagatay Yigit, Ilker Keser, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, Deniz Yuret, 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2020

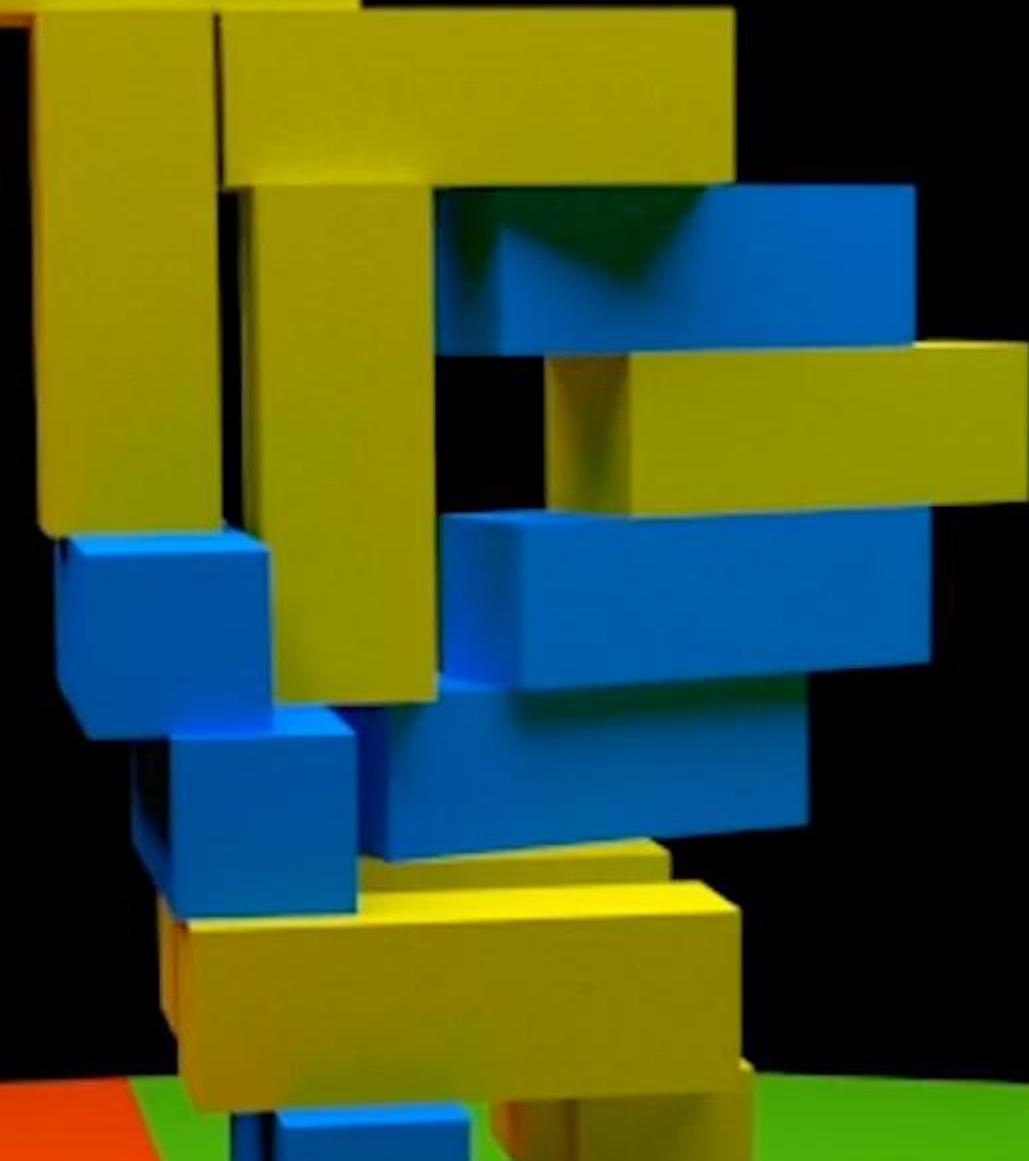


Intuitive Physics

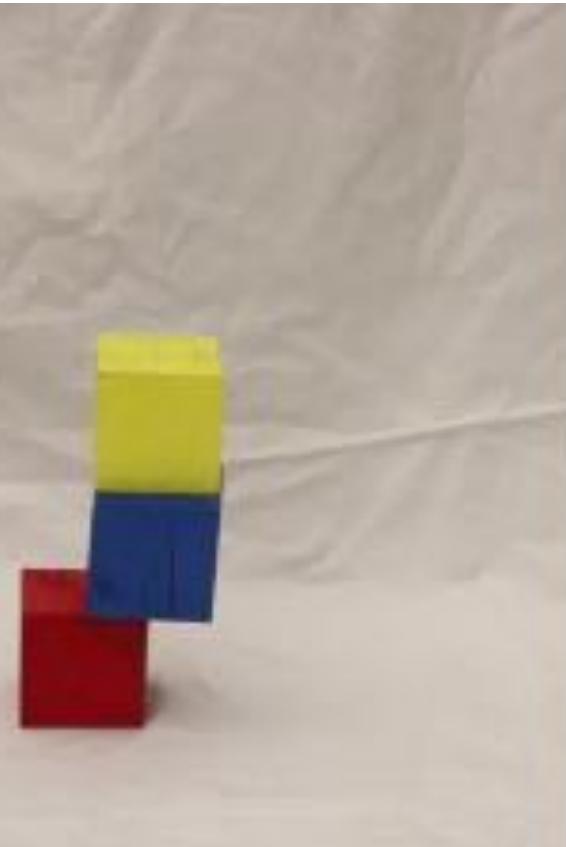
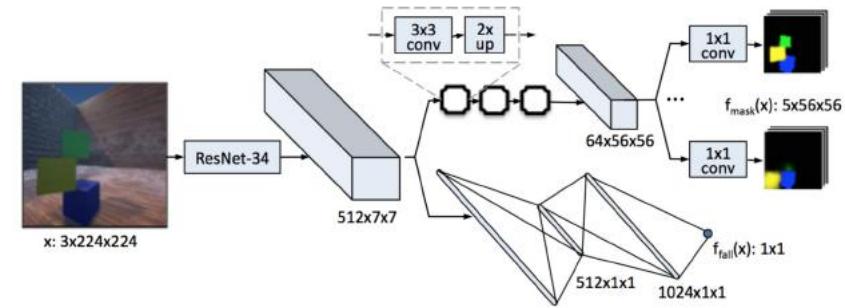
- Commonsense understanding of how the world operates at a physical level
- Helps us to perceive, understand and act with our environment



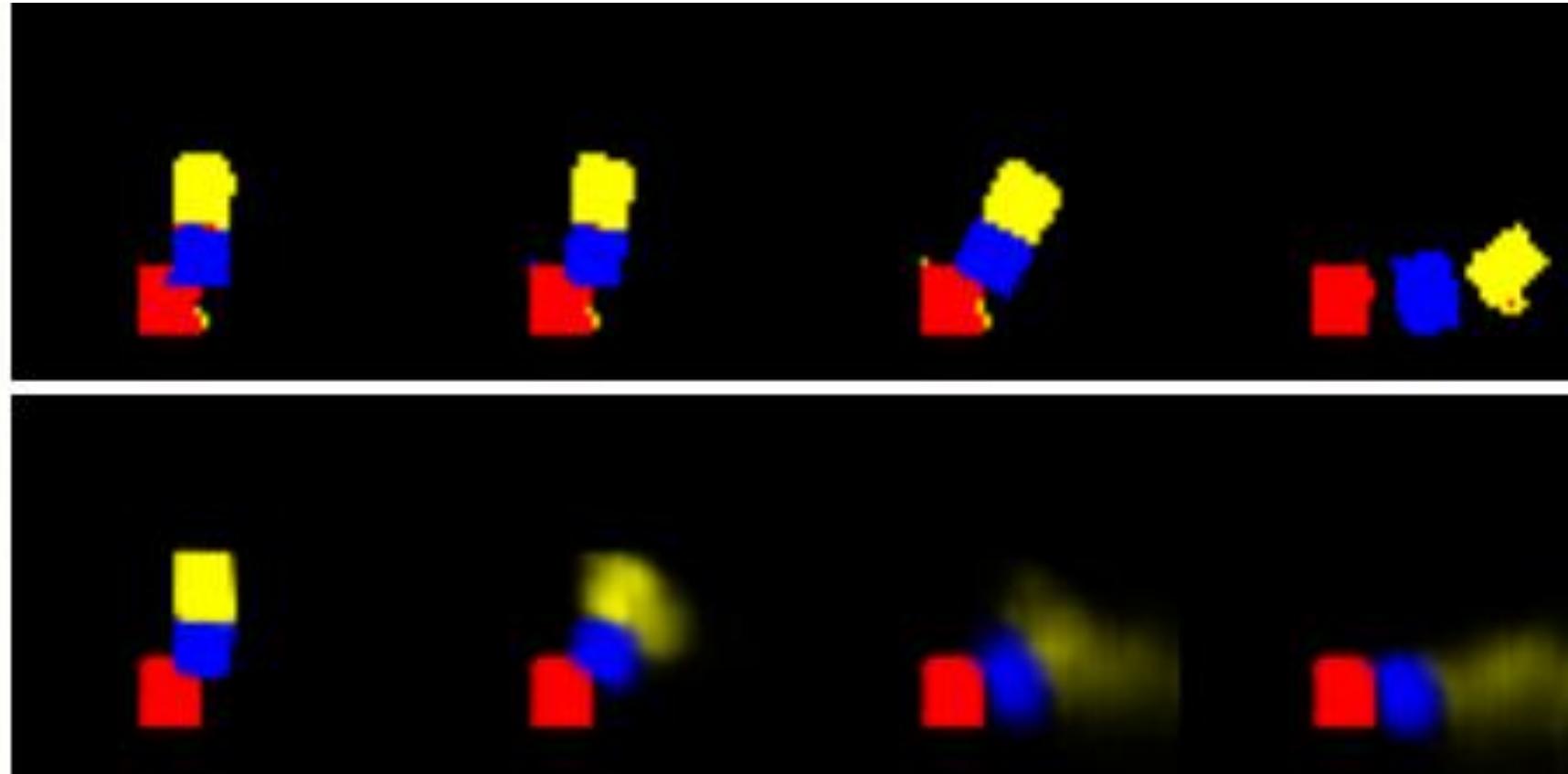
Intuitive Physics



Intuitive Physics



Initial frame



PhysNet predictions of the future

The CRAFT Dataset

- A new VQA dataset involving causal dynamic scenes.
- Motivated by theories of force dynamics in cognitive linguistics.
 - Leonard Talmy. Force Dynamics in language and cognition. *Cognitive Science*, 12, 1, 49–100, 1988.
 - Phillip Wolff and Aron K Barbey. Causal reasoning with forces. *Front. Hum. Neurosci.*, 9:1, 2015.

CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions

Tayfun Ates^{1,†}, Muhammed Samil Atesoglu¹, Cagatay Yigit¹, Ilker Kesenci², Mert Kobas³, Ercut Erdem¹, Aykut Erdem², Tilde Goksun³, Deniz Yuret²

¹ Department of Computer Engineering, Hacettepe University, Ankara, Turkey

² Department of Computer Engineering, Koç University, Istanbul, Turkey

³ Department of Psychology, Koç University, Istanbul, Turkey

[†]Correspondence: tates@hacettepe.edu.tr

Abstract

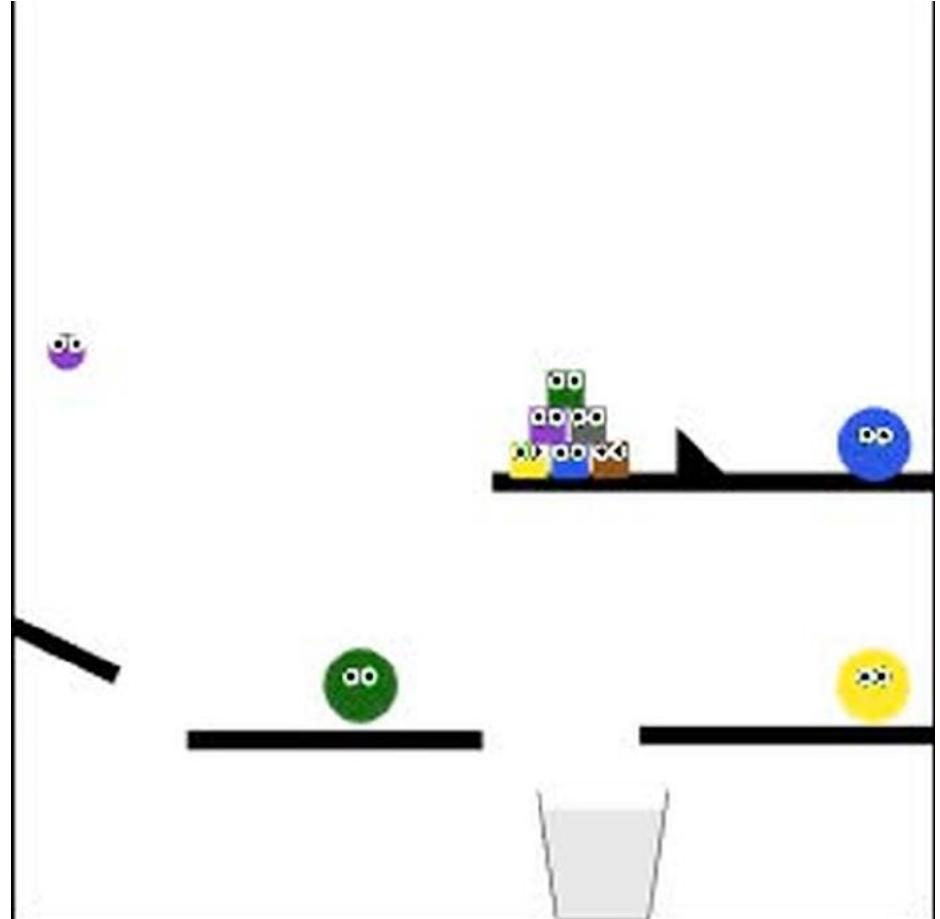
Recent advances in Artificial Intelligence and deep learning have revived the interest in studying the gap between the reasoning capabilities of humans and machines. In this ongoing work, we introduce CRAFT, a new visual question answering dataset that requires causal reasoning about physical forces and object interactions. It contains 38K video and question pairs that are generated from 3K videos from 10 different virtual environments, containing different number of objects in motion that interact with each other. Two question categories from CRAFT include previously studied *descriptive* and *counterfactual* questions. Besides, inspired by the theory of force dynamics from the field of human cognitive psychology, we introduce new question categories that involve understanding the intentions of objects through the notions of *cause*, *enable*, and *prevent*. Our preliminary results demonstrate that even though these tasks are very intuitive for humans, the implemented baselines could not cope with the underlying challenges.

1 Introduction

The collection of abilities of humans to understand and make approximate predictions about physical environments consisting of various objects that are in steady state or in motion is known as *intuitive physics* [Kubricht et al., 2017]. Cognitive scientists have extensively studied the factors that affect infants' or adults' ability of physical reasoning [Baillargeon, 1995, 2008; Téglás et al., 2011; Battaglia et al., 2013]. Some of these abilities have also been studied for other animals such as chicks (*Gallus gallus*) [Chiandetti and Vallortigara, 2011]. Recent advances in machine learning have enabled computers to understand what type of object is present in a specified image (*classification*), which bounding box best wraps that object (*detection*), what its exact boundaries are (*segmentation*). Although these artificial vision systems have shown astounding progress in the past decade, there are some areas in which these systems are still significantly below human performance. One such area includes the capability of humans to reason about physical actions of objects by observing their environment. This is a recent research direction for which cognitive and machine learning scientists are working together to bring similar capabilities to artificially intelligent robots so that they acquire similar intuitions and better understand their surroundings.

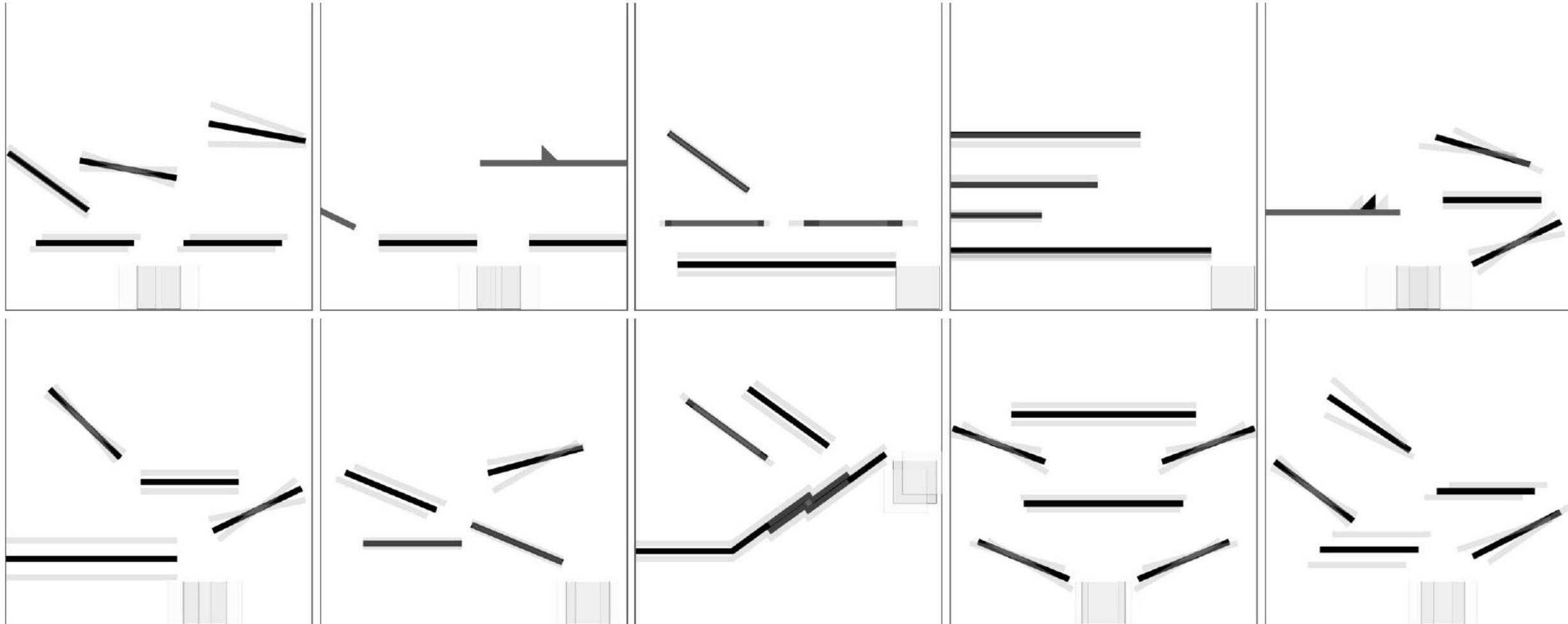
A crucial point that is worth mentioning here is that improving physical reasoning capabilities can also make agents better anticipate the results of their actions in their physical environments. They can gain the ability to consider counterfactual actions without actually performing them. They can estimate what will happen if they perform a specific action. One of the recent examples in this

The CRAFT Dataset



- A new VQA dataset involving causal dynamic scenes.
- Nearly 38K question and video pairs from 3K videos
- 10 sec long video sequences with 256×256 pixels frames
- Box2D is used for simulation physics
- 10 unique scene layouts

Scene Layouts



- 10 unique scene layouts – randomized by performing random perturbations on the scene elements

Scene Elements

Dynamic Objects

- square □, triangle △, circle ○
- gray ■, red ■, blue ■, green ■, brown ■, purple ■, cyan ■, yellow ■
- 2 sizes (small ■, large ■)

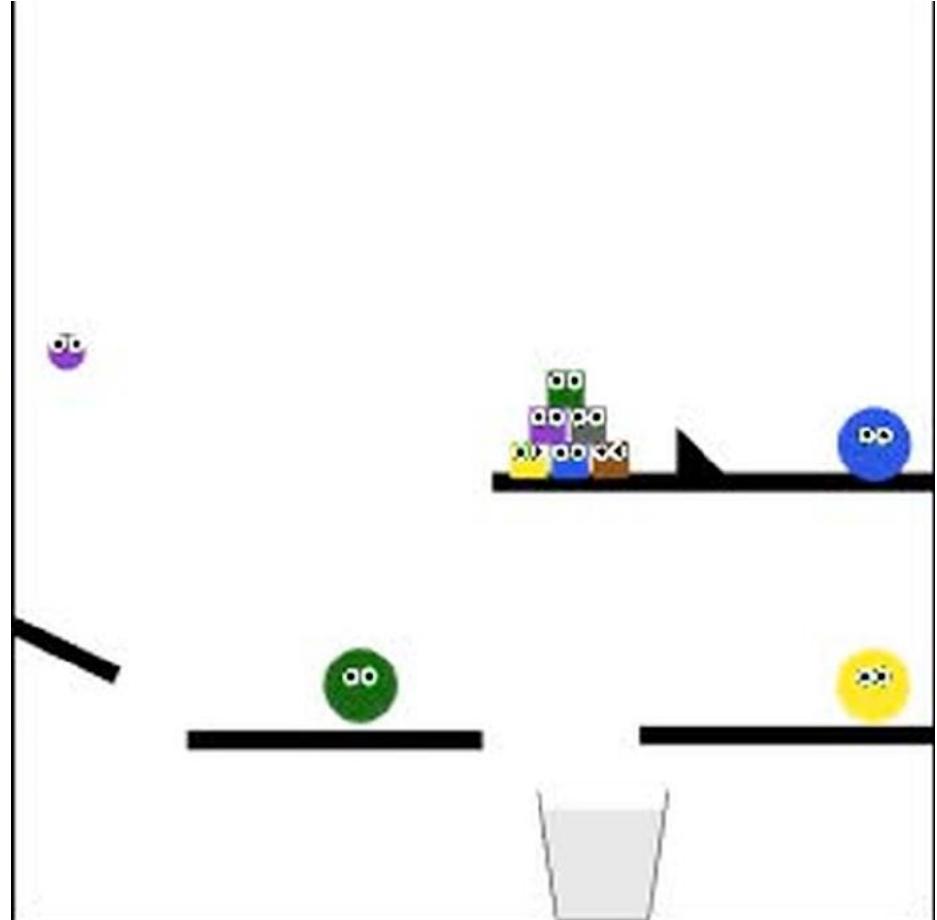
Attributes are in continuous change throughout the sequence due to gravity or other interactions that they are subject to, until they rest.

Static Scene Elements

- Ramp ▲, platform ■, basket ■, left wall ↗, right wall ↙, ground ↑
- All black in color
- Single size

Attributes such as position or orientation are decided at the beginning of the simulation and they are fixed throughout the video sequence

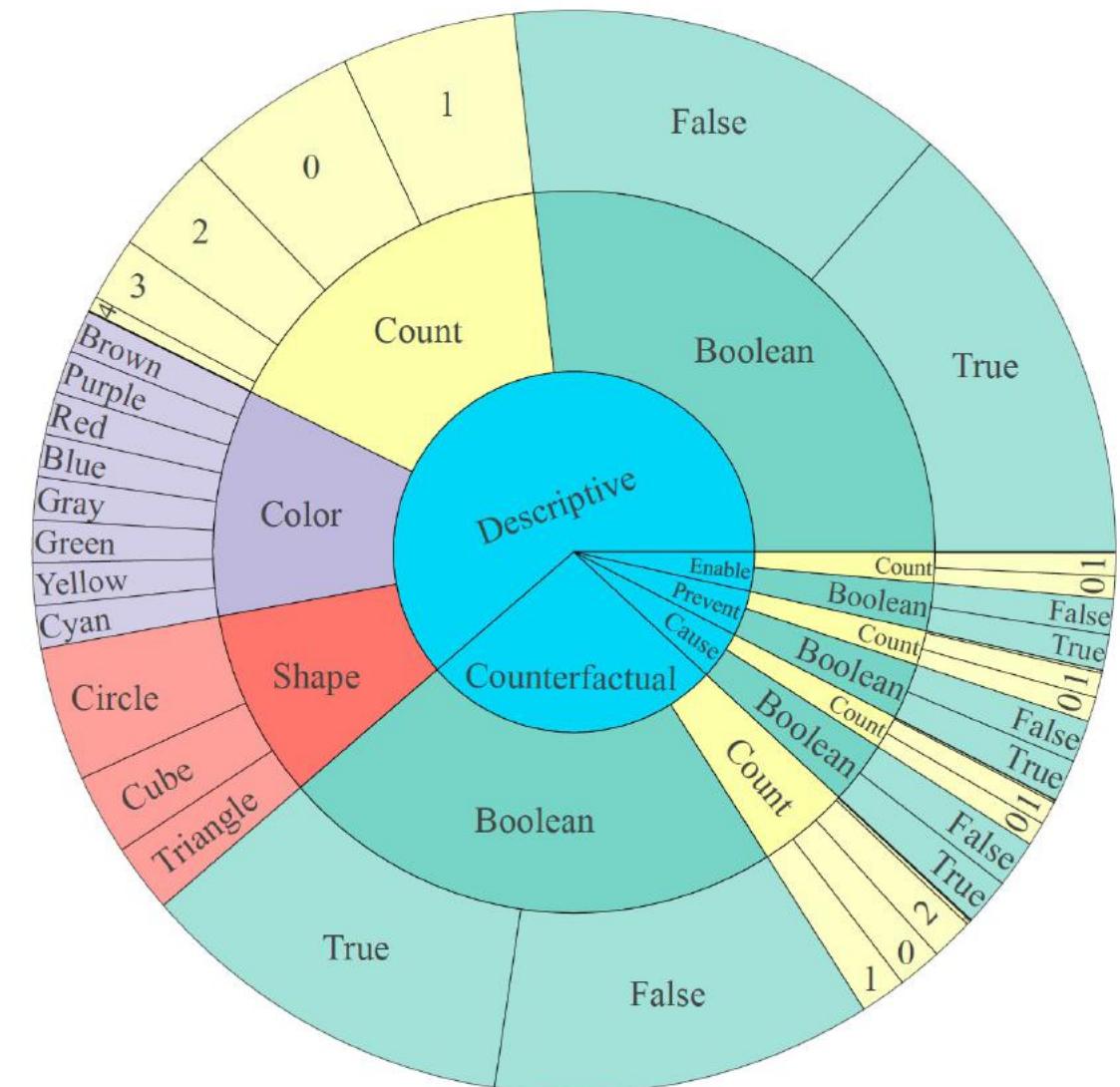
Event Types



- Video starts
- Blue circle hits the gray square
- Blue circle starts touching the brown square
 -
 -
- Yellow circle hits the green circle
- Purple square hits the yellow circle
 -
 -
- Gray square enters the basket
- Video end

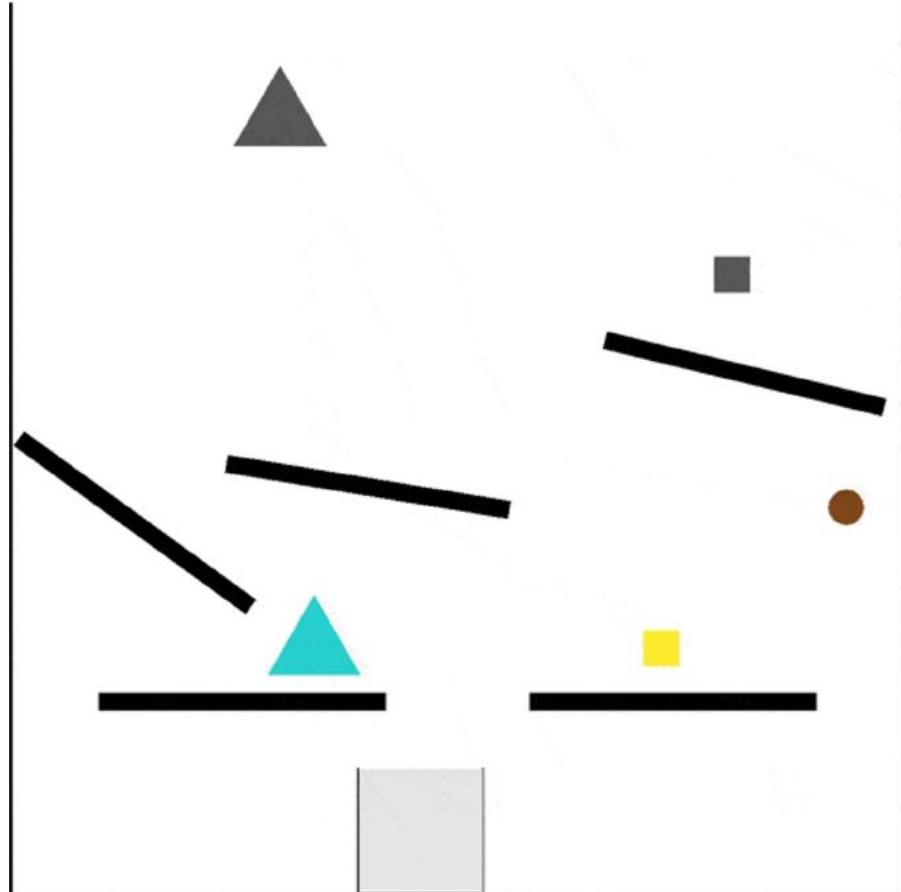
Dataset Statistics

- 650 scene layout and question type pairs
- We try keeping overall answer distribution as uniform as possible
- Four question types:
 - Descriptive
 - Counterfactual
 - Cause
 - Enable
 - Prevent



Descriptive Questions

- Each question requires extracting attributes of objects.
- Some of them (e.g. counting) need temporal reasoning too.



Question: How many object fall to the ground?

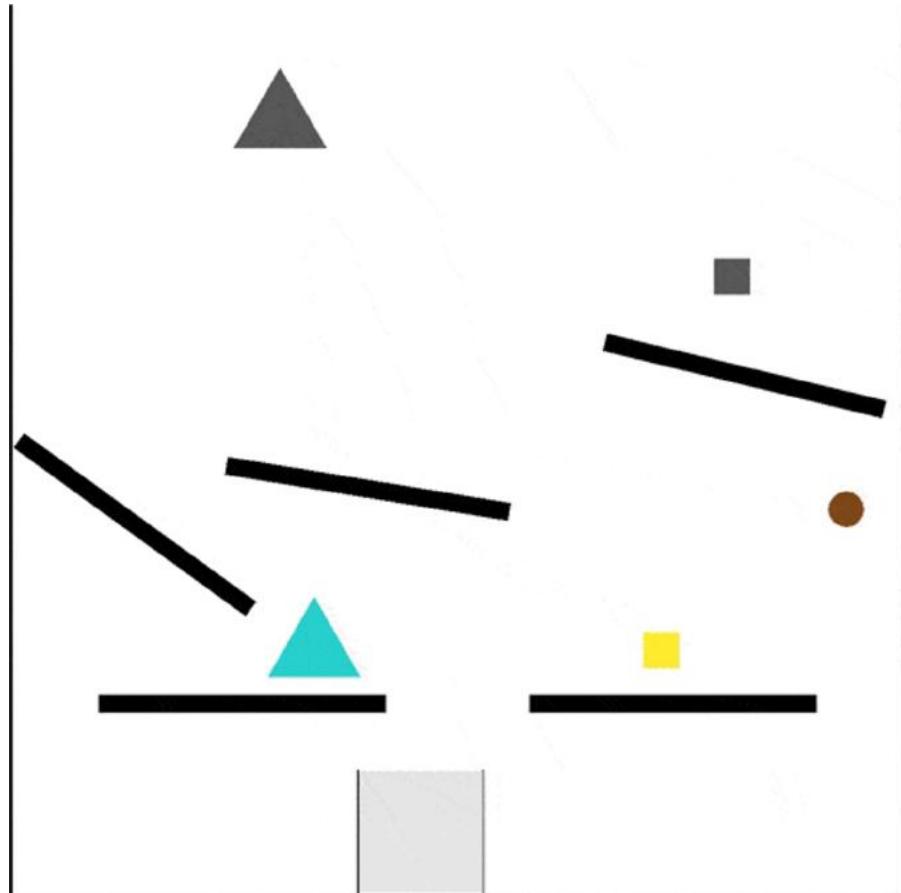
Answer: 2

Question: "After entering the basket, does the small yellow square collide with other objects?"

Answer: True

Descriptive Questions

- Each question requires extracting attributes of objects.
- Some of them (e.g. counting) need temporal reasoning too.



Question: How many object fall to the ground?

Program:

```
Count (
    FilterDynamic (
        FilterObjectsFromEvents (
            FilterCollideGround (
                Events ()
            )
        )
    )
)
```

Answer: 2

Question: "After entering the basket, does the small yellow square collide with other objects?

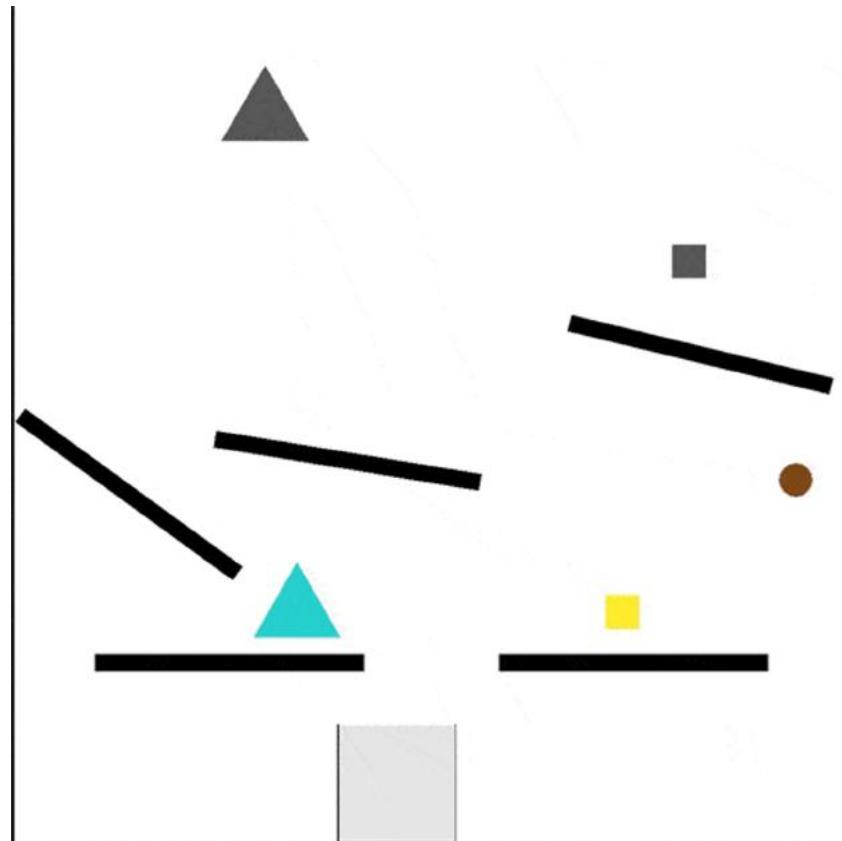
Program:

```
Var QueryObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ), "Yellow" ), "Cube" )
Var SmallYellowCubeEvents = FilterEvents ( Events(), QueryObject )
Exist (
    FilterAfter (
        FilterCollisionWithDynamics ( SmallYellowCubeEvents ),
        FilterFirst (
            FilterEnterBasket ( SmallYellowCubeEvents )
        )
    )
)
```

Answer: True

Counterfactual Questions

- Each question requires understanding what would happen if one of the objects was removed from the scene.



Question: How many object fall to the ground, if the small yellow box is removed?

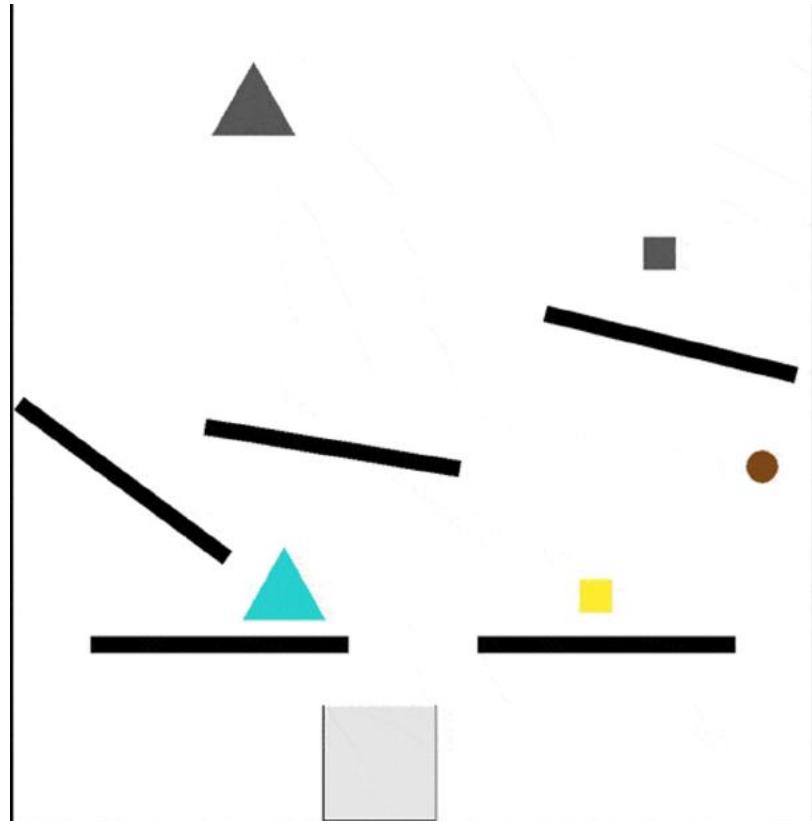
Answer: 1

Question: Does the small gray box enter the basket, if any other single one of the objects is removed?

Answer: True

Cause Questions

- Each question requires reasoning about whether an object is causing an event to occur or not (mentally comparing the original input sequence with the counterfactual version).

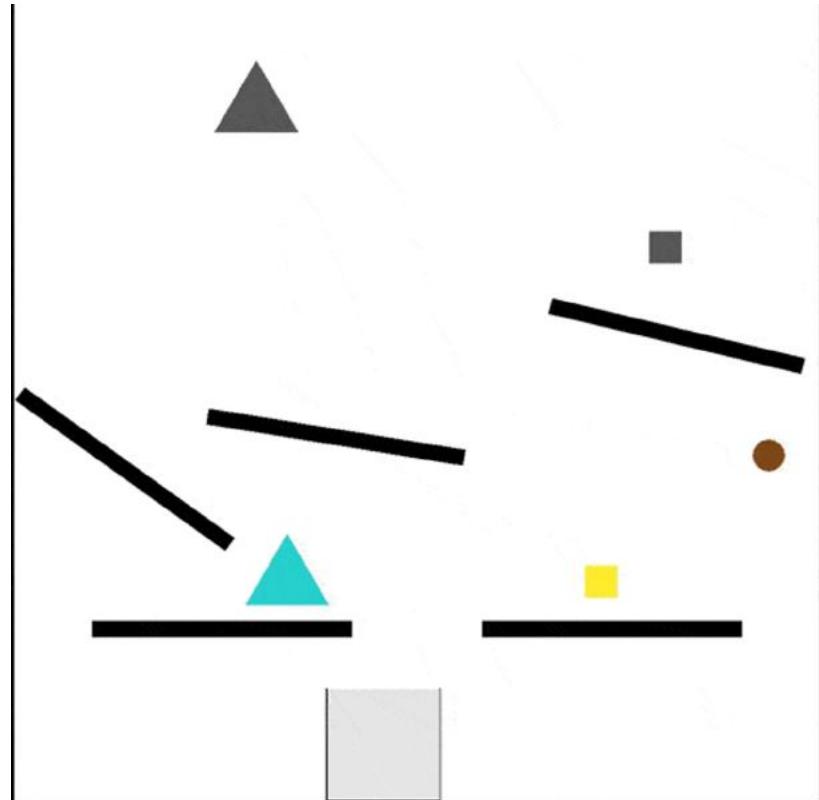


Question: Does the **small brown circle** cause the **tiny yellow square** to **enter** the basket?

Answer: True

Enable Questions

- Each question requires counterfactual reasoning (mentally simulation to see whether an event will still happen when an object is removed from the scene)

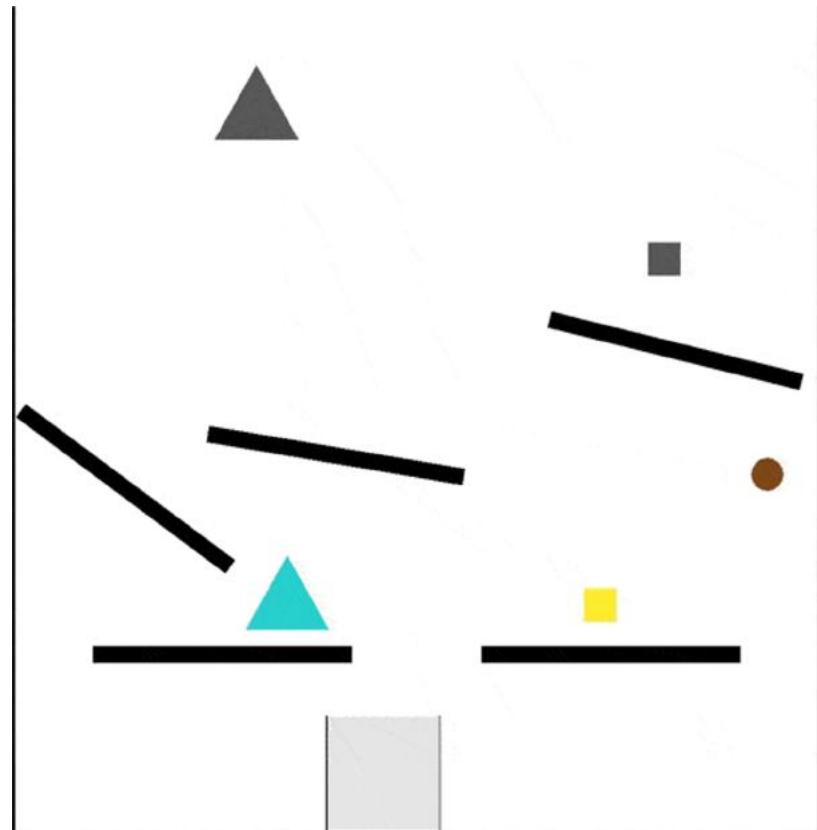


Question: How many objects does the small gray block enable to enter the basket?

Answer: 0

Prevent Questions

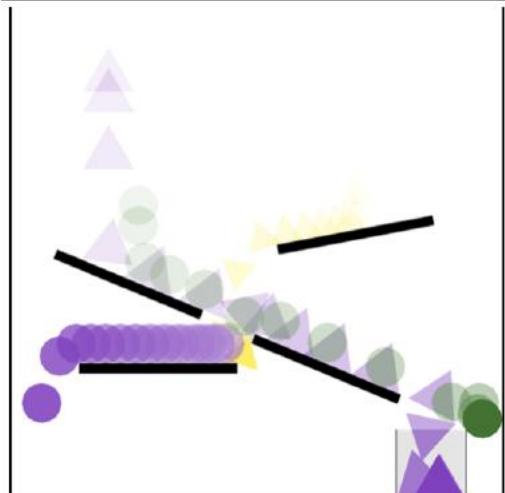
- Each question requires counterfactual reasoning (mentally simulating what happens next when an object is removed from the scene and comparing with original outcome)



Question: Does the **small yellow square** prevent the **tiny brown circle** from **entering** the basket?

Answer: **True**

Qualitative Results

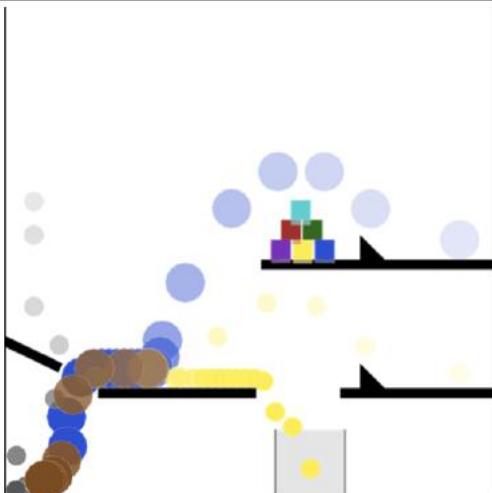


What is the shape of first object to collide with the tiny yellow triangle?

Correct Answer: "Circle"

LSTM Prediction: "Triangle"

LSTM+CNN (Last Frame) Prediction: "Triangle"

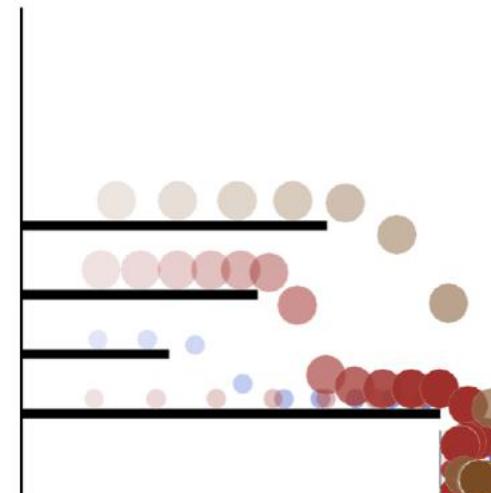


What color is the object that the small gray circle first collides with?

Correct Answer: "Brown"

LSTM Prediction: "Green"

LSTM+CNN (Last Frame) Prediction: "Cyan"

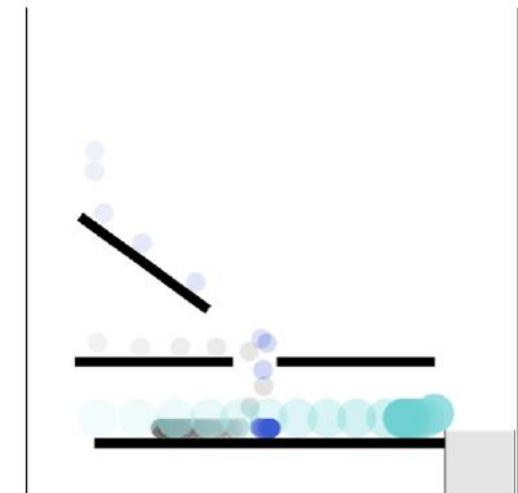


Is the big red circle is kept by the large brown circle from colliding with the bucket?

Correct Answer: "True"

LSTM Prediction: "False"

LSTM+CNN (First Frame) Prediction: "False"



If the big cyan circle is removed, how many objects enter the basket?

Correct Answer: "0"

LSTM Answer: "1"

LSTM+CNN (First Frame) Prediction: "1"

Quantitative Results

Model	Input	Train	Val	Cause	Counterfactual	Test			
						Descriptive	Enable	Prevent	All
MFA	Question	28.61	27.86	30.50	42.36	21.54	27.31	27.54	28.00
AT-MFA	Question	42.83	41.29	45.60	47.11	38.14	47.60	44.59	41.48
LSTM	Question	58.73	44.76	53.77	52.99	39.16	52.77	55.08	44.65
LSTM-CNN	Question + First Frame	93.14	47.68	44.34	50.34	46.32	52.03	53.77	47.83
LSTM-CNN	Question + Last Frame	90.94	53.34	50.63	56.71	53.99	52.40	51.48	54.42
Human*	Question + Video	—	—	66.67	74.07	93.01	59.09	90.90	85.89

*Based on responses from 12 adults on 489 questions sampled from the test split

Quantitative Results

Model	Input	Train	Val	Test					
				Cause	Counterfactual	Descriptive	Enable	Prevent	All
MFA	Question	88.81	87.88	88.50	10.00	81.51	27.31	27.54	28.00
AT-MFA	Question						7.60	44.59	41.48
LSTM	Question						7.77	55.08	44.65
LSTM-CNN	Question + First Frame						10.03	53.77	47.83
LSTM-CNN	Question + Last Frame						11.40	51.48	54.42
Human*	Question + Video	–	–	66.67	74.07	93.01	59.09	90.90	85.89

There is large gap (>30)
between humans and
our neural baseline!

*Based on responses from 12 adults on 489 questions sampled from the test split

End Credits

“It’s clear that to make and assess progress in AI more effectively, we will need to develop a better vocabulary for talking about what machines can do. And more generally, we will need a better scientific understanding of intelligence as it manifests in different systems in nature. This will require AI researchers to engage more deeply with other scientific disciplines that study intelligence.

...

In 1892, the psychologist William James said of psychology at the time, “This is no science; it is only the hope of a science” [84]. This is a perfect characterization of today’s AI.”



Why AI is Harder Than We Think

Melanie Mitchell
Santa Fe Institute
Santa Fe, NM, USA
mm@santafe.edu

Abstract

Since its beginning in the 1950s, the field of artificial intelligence has cycled several times between periods of optimistic predictions and massive investment (“AI spring”) and periods of disappointment, loss of confidence, and reduced funding (“AI winter”). Even with today’s seemingly fast pace of AI breakthroughs, the development of long-promised technologies such as self-driving cars, housekeeping robots, and conversational companions has turned out to be much harder than many people expected. One reason for these repeating cycles is our limited understanding of the nature and complexity of intelligence itself. In this paper I describe four fallacies in common assumptions made by AI researchers, which can lead to overconfident predictions about the field. I conclude by discussing the open questions spurred by these fallacies, including the age-old challenge of imbuing machines with humanlike common sense.

Introduction

The year 2020 was supposed to herald the arrival of self-driving cars. Five years earlier, a headline in *The Guardian* predicted that “From 2020 you will become a permanent backseat driver” [1]. In 2016 *Business Insider* assured us that “10 million self-driving cars will be on the road by 2020” [2]. Tesla Motors CEO Elon Musk promised in 2019 that “A year from now, we’ll have over a million cars with full self-driving, software...everything” [3]. And 2020 was the target announced by several automobile companies to bring self-driving cars to market [4, 5, 6].

Despite attempts to redefine “full self-driving” into existence [7], none of these predictions has come true. It’s worth quoting AI expert Drew McDermott on what can happen when over-optimism about AI systems—in particular, self-driving cars—turns out to be wrong:

Perhaps expectations are too high, and... this will eventually result in disaster. [S]uppose that five years from now [funding] collapses miserably as autonomous vehicles fail to roll. Every startup company fails. And there’s a big backlash so that you can’t get money for anything connected with AI. Everybody hurriedly changes the names of their research projects to something else. This condition [is] called the “AI Winter” [8].

What’s most notable is that McDermott’s warning is from 1984, when, like today, the field of AI was awash with confident optimism about the near future of machine intelligence. McDermott was writing about a cyclical pattern in the field. New, apparent breakthroughs would lead AI practitioners to predict rapid progress, successful commercialization, and the near-term prospects of “true AI”. Governments and companies would get caught up in the enthusiasm, and would shower the field with research and development funding. AI Spring would be in bloom. When progress stalled, the enthusiasm, funding, and jobs would dry up. AI Winter would arrive. Indeed, about five years after McDermott’s warning, a new AI winter set in.

In this chapter I explore the reasons for the repeating cycle of overconfidence followed by disappointment in expectations about AI. I argue that over-optimism among the public, the media, and even experts can