

CSE 599 Empirical Foundations of Machine Learning

University of Washington, Autumn 2021

Normally class starts at 10 am, today 10:05 so people can find the room.

Welcome!

Introduction



Instructor: Ludwig Schmidt

MIT (PhD) → Berkeley (postdoc) → UW (faculty) - started this fall.

Research interests: exactly this class!



TA: Mitchell Wortsman

UW (3rd year PhD student advised by Ali Farhadi)

Research interests: still narrowing it down ...

1. Logistics

2. Background & motivation

3. Course outline

1. Logistics

2. Background & motivation

3. Course outline

Basics

Room: CSE2 G04 (Gates building)

Time: Tuesday / Thursday 10 - 11:20 am

Website: <https://mlfoundations.github.io/au21/> (announcements, material, etc.)

Registration: Now available! (see link on website)

Please provide feedback if you see things we can improve or suggestions for topics

Ask questions any time!

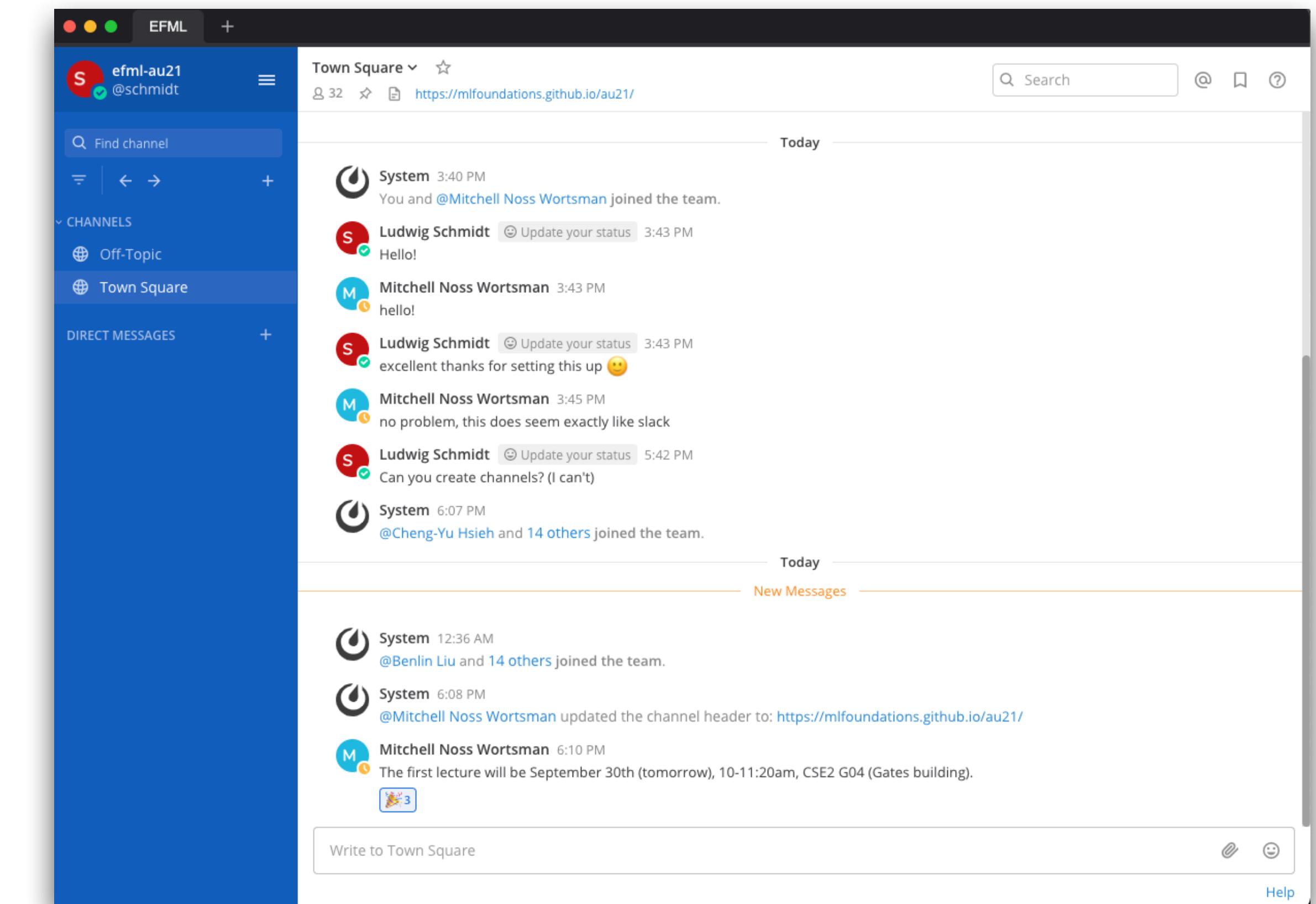
Communication: Mattermost

Similar to Slack but FERPA-compliant (Family Educational Rights and Privacy Act)

Should be accessible by anyone at UW
(may require a request if not CSE)

**Please log in if you have not already
done so! (It's easy)**

Feel free to ask any questions related to
the course, post papers, etc.



(Remote) Attendance

In-person attendance is **strongly** encouraged.

→ Experience will be better, especially for discussions.

We have a Zoom link for a few people who cannot
join in person, but the link is secret :-)



zoom

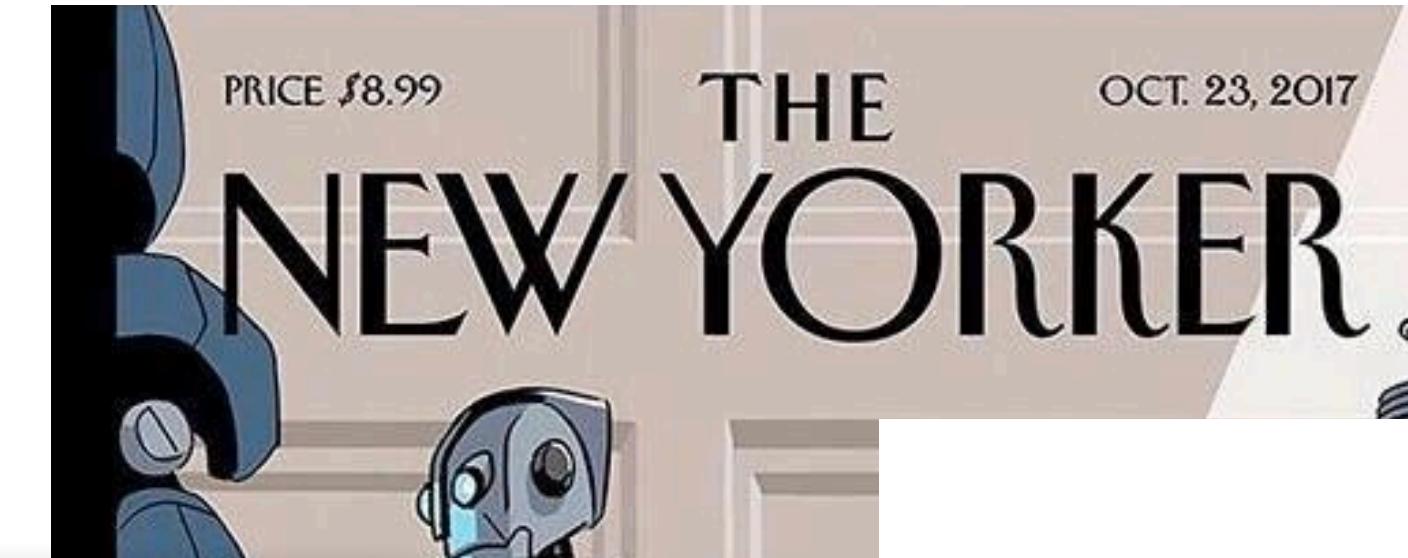
If you cannot join for a specific session, message
Mitchell and me the day before and we will send you the link.

1. Logistics

2. Background & motivation

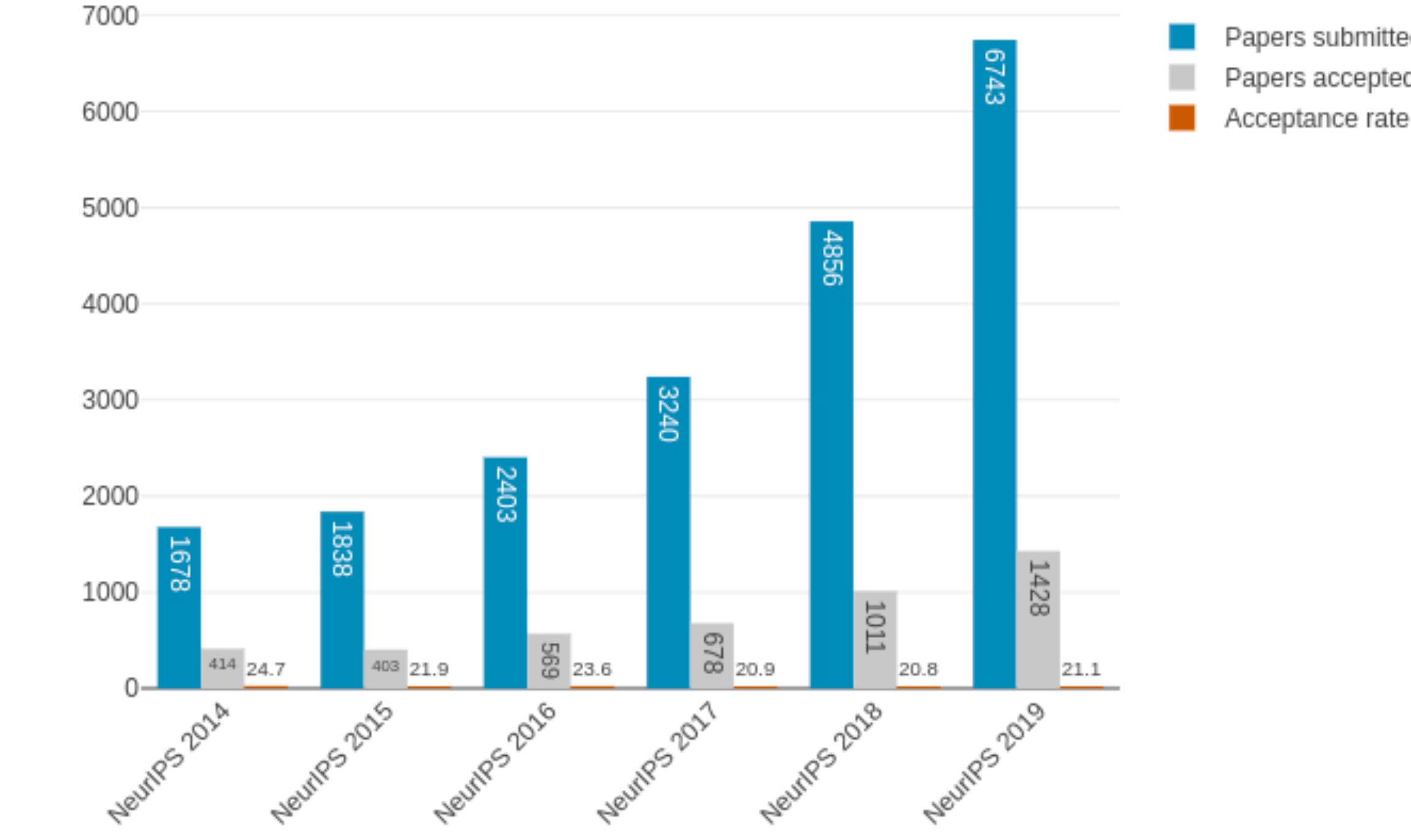
3. Course outline

Explosive Growth in ML



The image is a screenshot of The New York Times Magazine website. On the left, there's a dark sidebar with the word 'FEATURE' and the title 'The Great A.I. Awakening'. Below the title is a short description: 'How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.' To the right of the sidebar is a large graphic: a white circle containing a smaller circle divided into four quadrants of different colors (red, green, yellow, blue). To the right of the graphic is a cartoon illustration of a blue robot standing over a small brown dog. The top navigation bar includes 'The New York Times Magazine' and 'Account ▾'.

Statistics of acceptance rate NeurIPS



Berkeley News

CAMPUS & COMMUNITY, CAMPUS NEWS

Berkeley inaugurates Division of Data Science and Information, connecting teams across research from all corners of campus

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

HOW GOOGLE IS REMAKING ITSELF AS A “MACHINE LEARNING FIRST” COMPANY

STEVEN LEVY BACKCHANNEL 06.22.2016 12:00 AM

How Google is Remaking Itself as a “Machine Learning First” Company

If you want to build artificial intelligence into every product, you better retrain your army of coders. Check.

AAAS

Become a Member

Science

Contents ▾

News ▾

Careers ▾

Journals ▾

SHARE



205

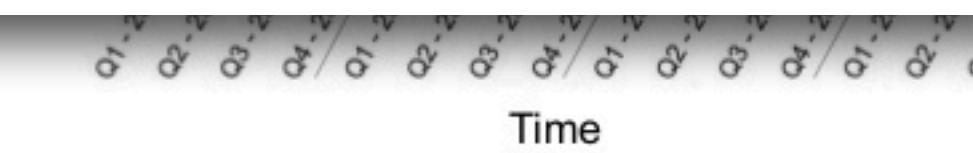


Senator Charles Schumer (D-NY) unveiled his artificial intelligence plan last week at a meeting of the National Security Commission on Artificial Intelligence. ALEX WONG/GETTY IMAGES

United States should make a massive investment in AI, top Senate Democrat says

By Jeffrey Mervis | Nov. 11, 2019, 11:45 AM

The top Democrat in the U.S. Senate wants the government to create a new agency that would invest an additional \$100 billion over 5 years on basic research in artificial intelligence (AI). Senator Charles Schumer (D-NY) says the initiative would enable the United States to keep pace with China and Russia in a critical research arena and plug gaps in what U.S. companies are unwilling to finance.





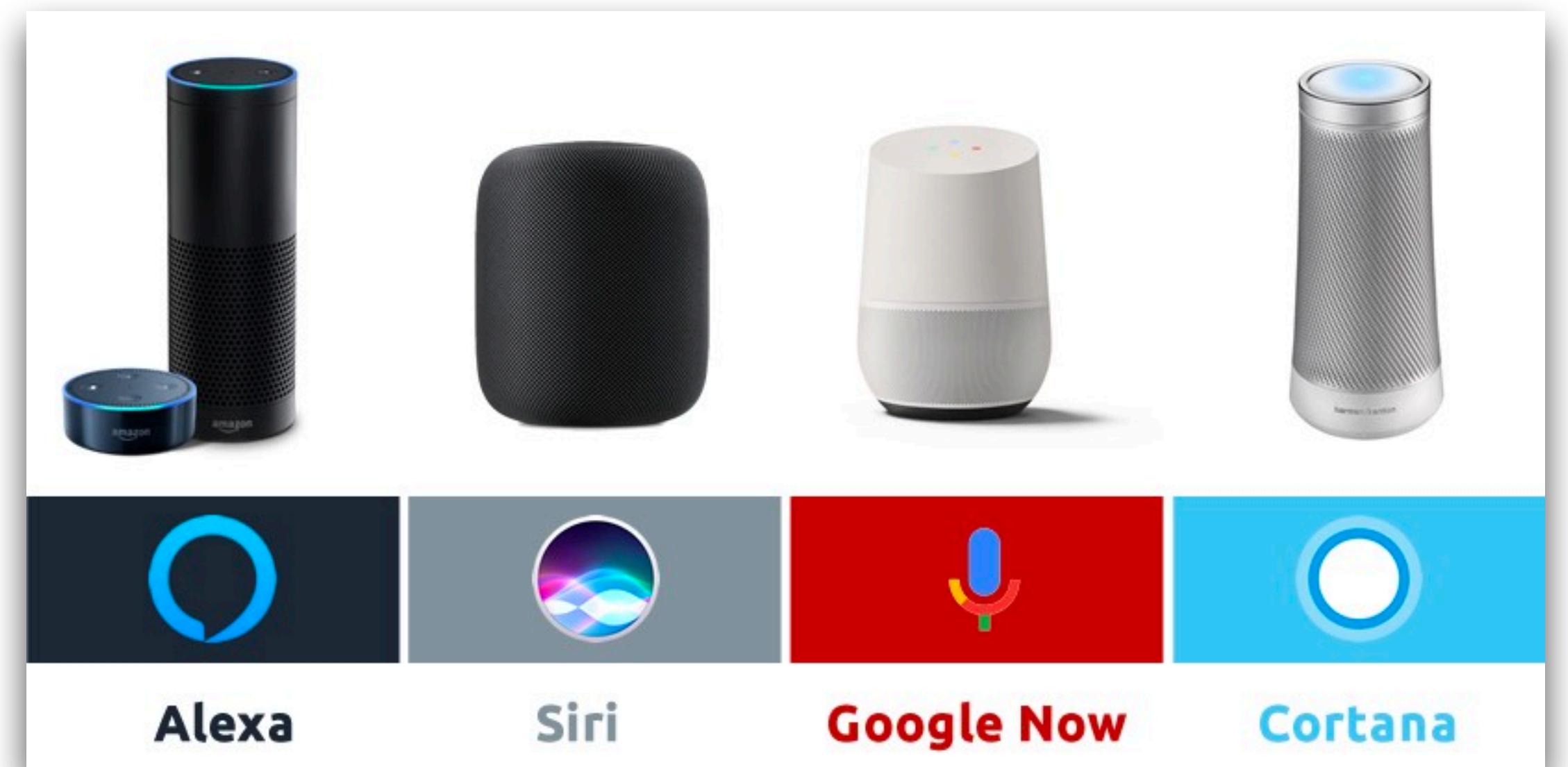
Self-driving cars



Robotics



Medical imaging



Voice assistants

What are the key advancements?

Progress in multiple areas of machine learning with similar approach: **deep learning**

- Computer vision
- Automatic speech recognition
- Natural language processing
- Game playing (Go, Atari, Starcraft, DotA)

Focus today: **computer vision**

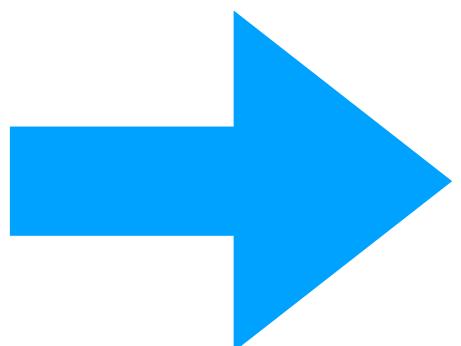


[Deng, Dong, Socher, Li, Li, Fei-Fei'09]

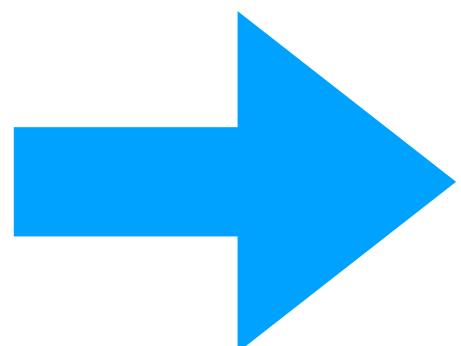
[Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, Berg Fei-Fei'15] 14

ImageNet

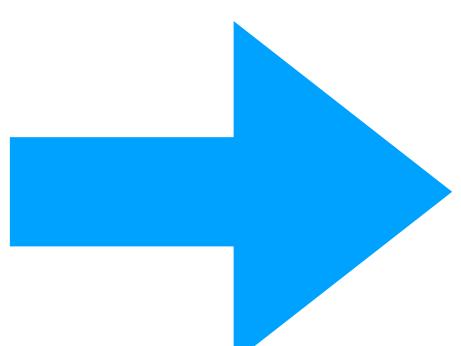
Large **image classification** dataset: 1.2 mio training images, 1,000 image classes.



Golden retriever

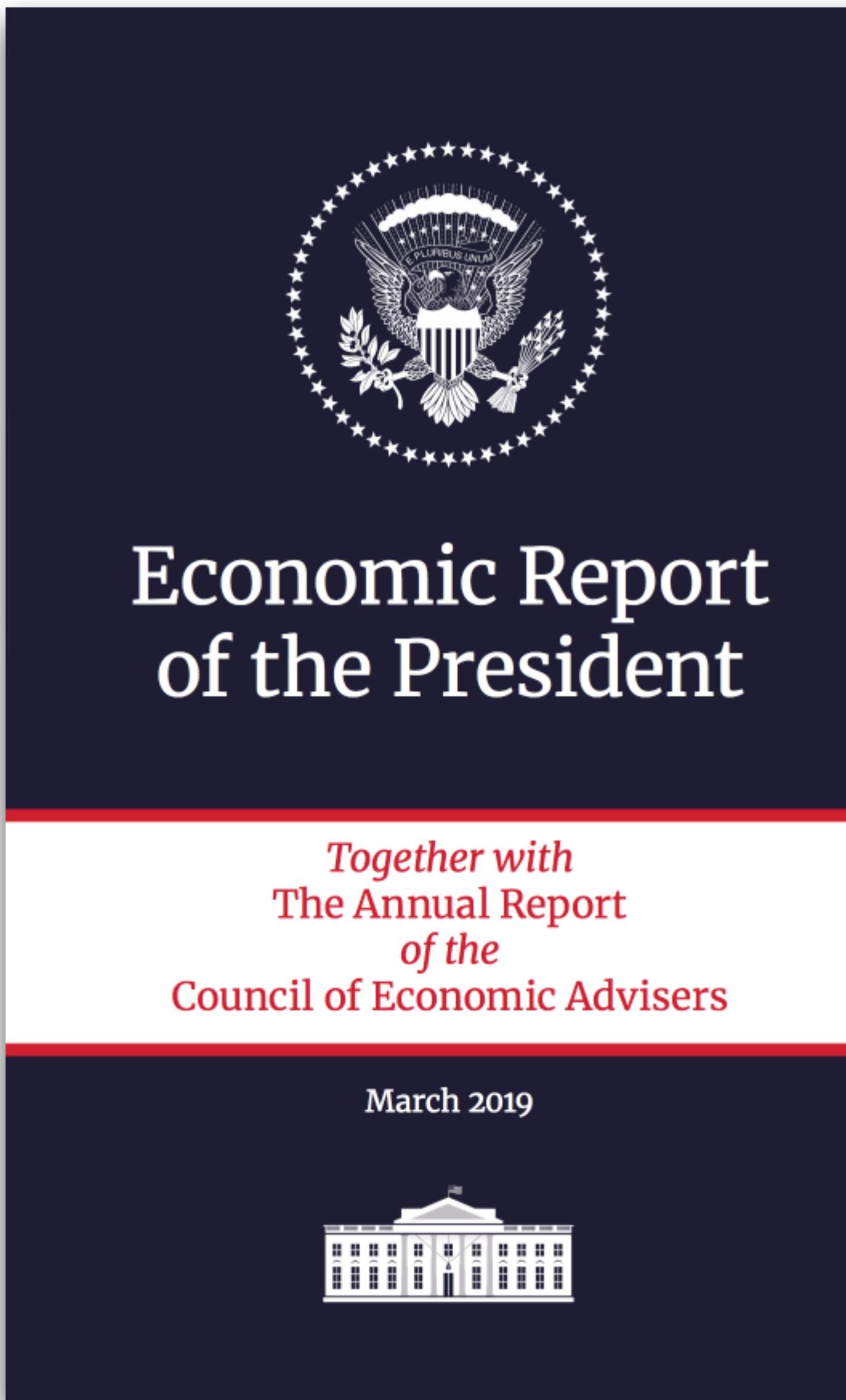


Great white shark

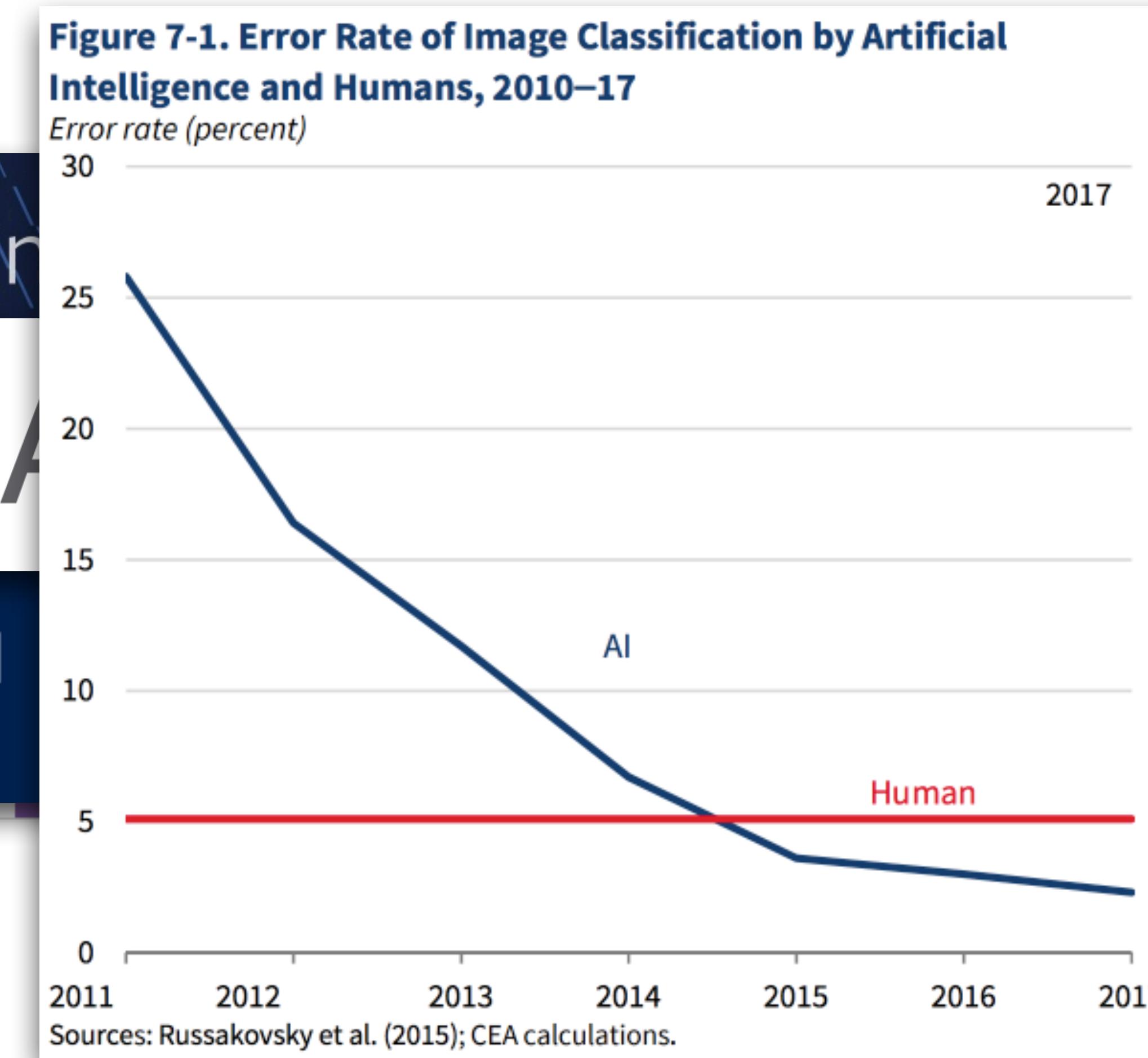


Minibus

ImageNet

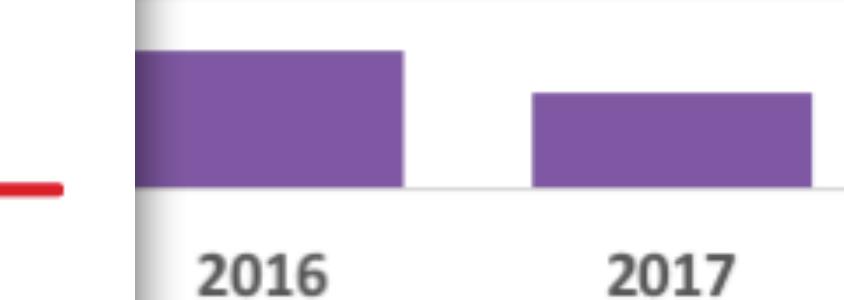


last decade:



*that the following
most impactful paper in
the field of deep learning and computer vision
was published in the last two years.”*

CACM June 2017



ImageNet History

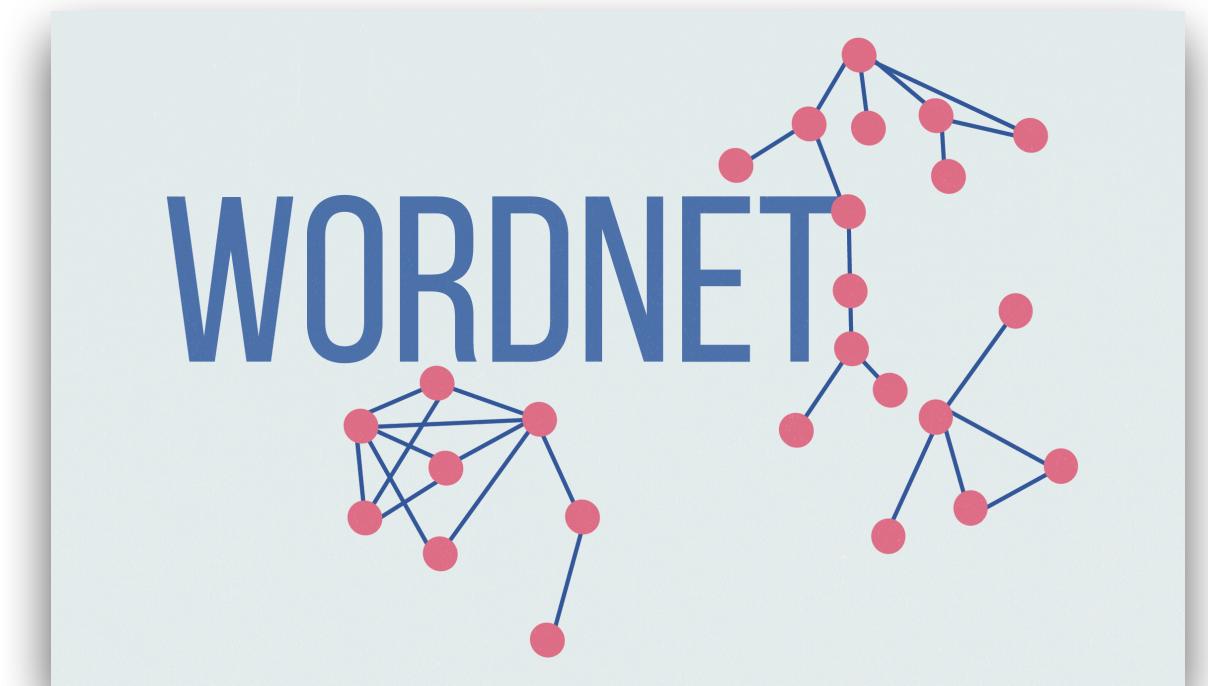
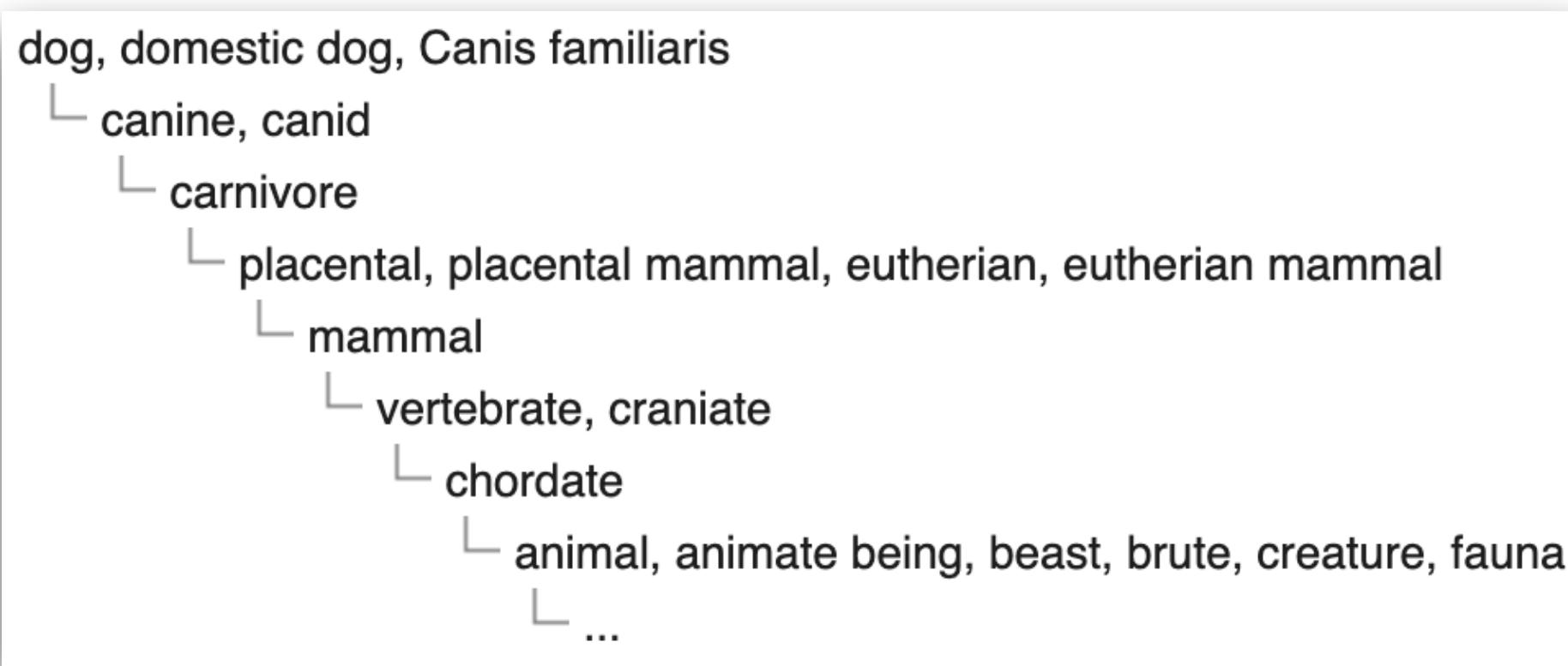
Key person: **Fei-Fei Li**

Assistant prof at Princeton starting 2007

Princeton is also home to the **WordNet** project



Hierarchical database of words in English and other languages



ImageNet History

Fei-Fei's vision (2006 – 2007):

- Humans know thousands of visual categories (neuroscience).
- If we want human-like computer vision, we need correspondingly large datasets.

→ Let's populate all of WordNet with around 1,000 images per node!

→ About 50 million images for about 50,000 classes (nouns in WordNet)

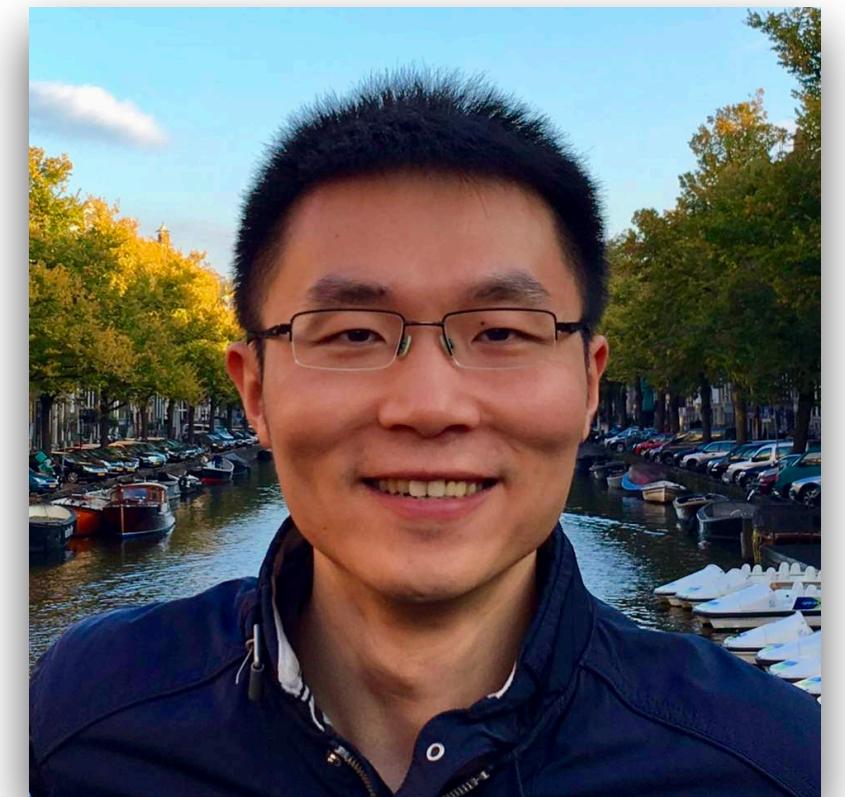
(Planned) ImageNet is 1000x larger!

Context: **PASCAL VOC**

- Most active object detection / classification dataset from 2005 - 2012
- Largest version (2012): 12,000 images total for 20 classes

Building ImageNet

Main student: Jia Deng (now back at Princeton as faculty)



Where do you get 50 million images?

→ Internet! (increasing amount of consumer photos)



How do you label them?

→ Internet! (Crowdsourcing platforms)

- + lots of **clever** task design
- + lots of **hard** work

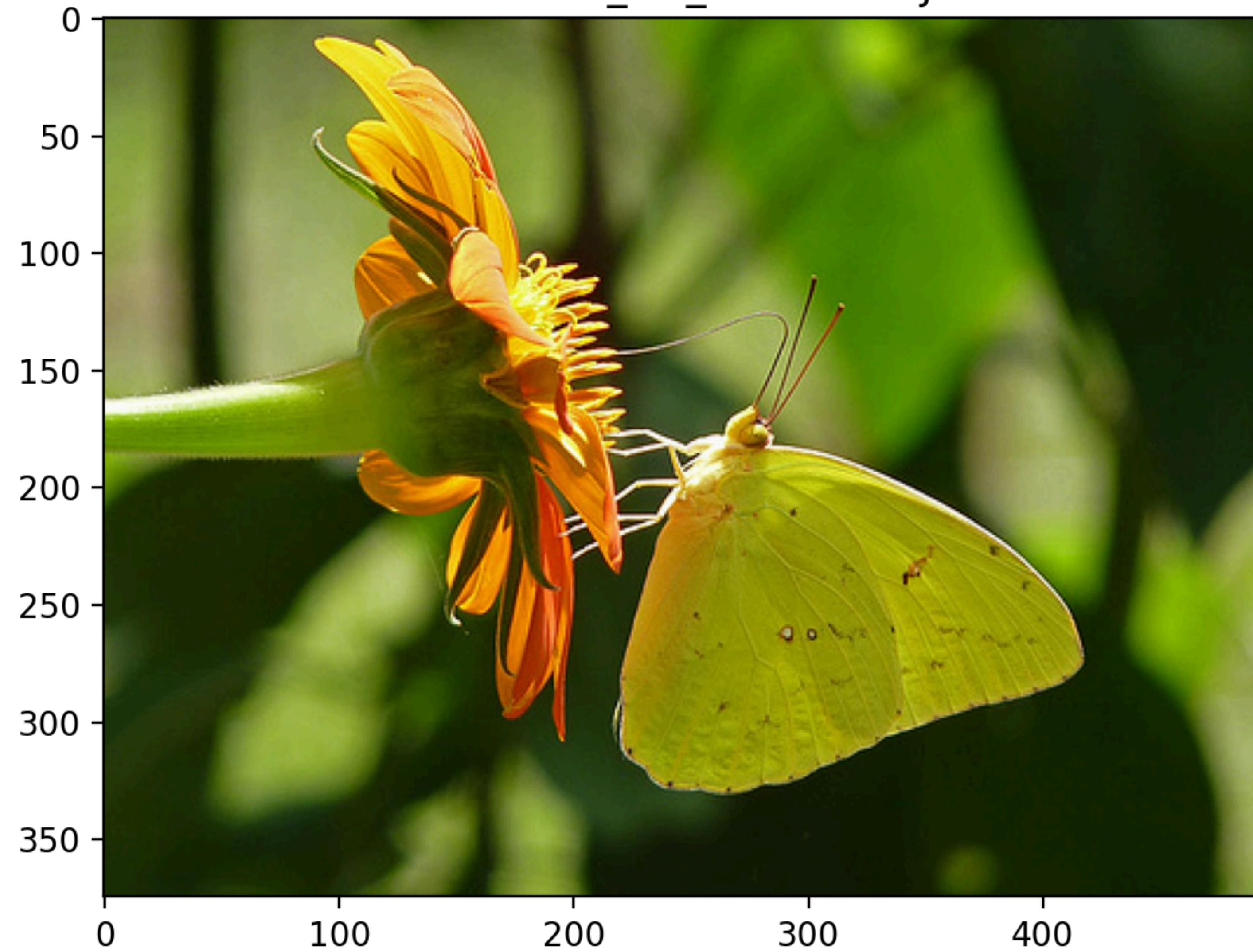


[Deng, Dong, Socher, Li, Li, Fei-Fei'09]

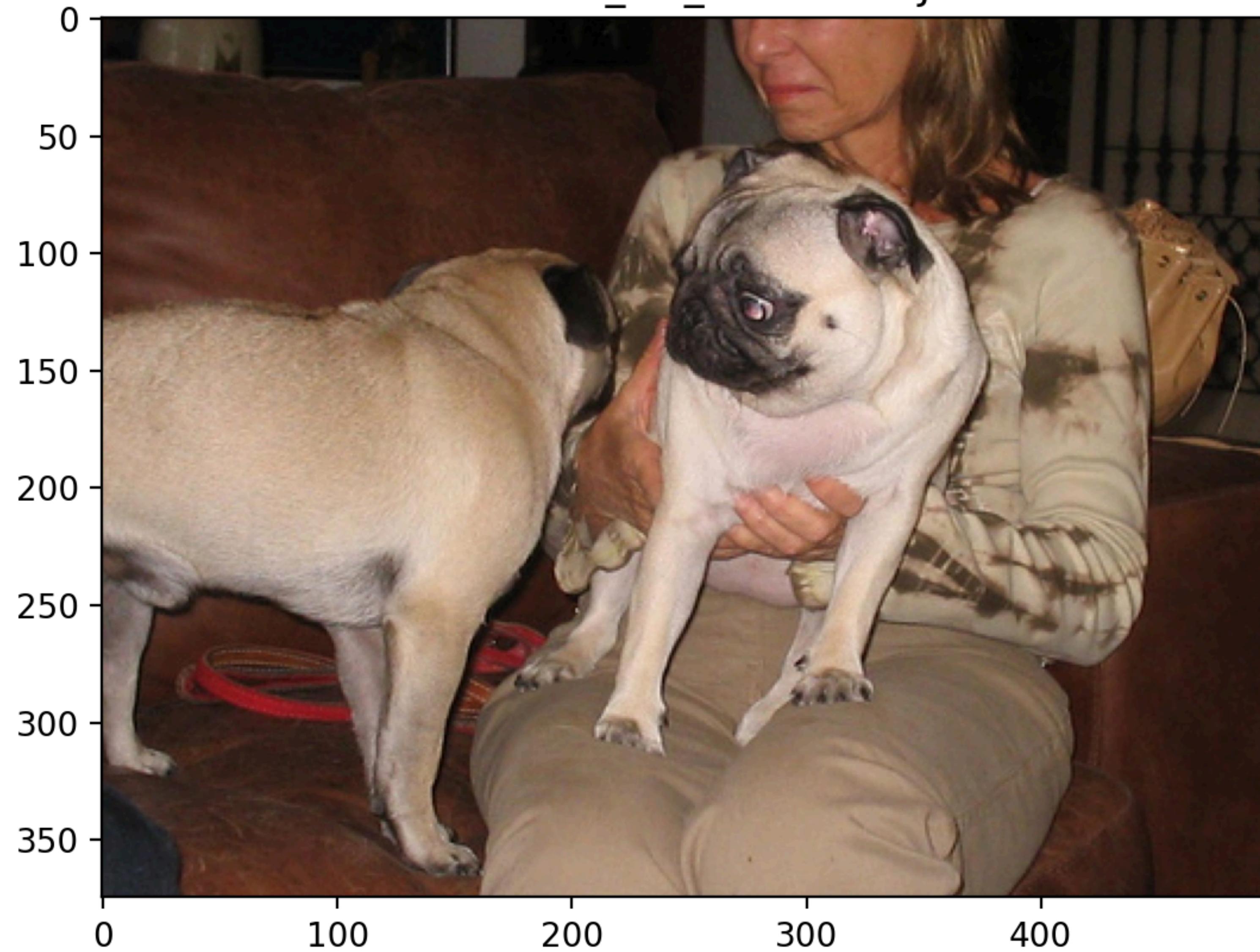
ILSVRC2012_val_00000293.jpeg



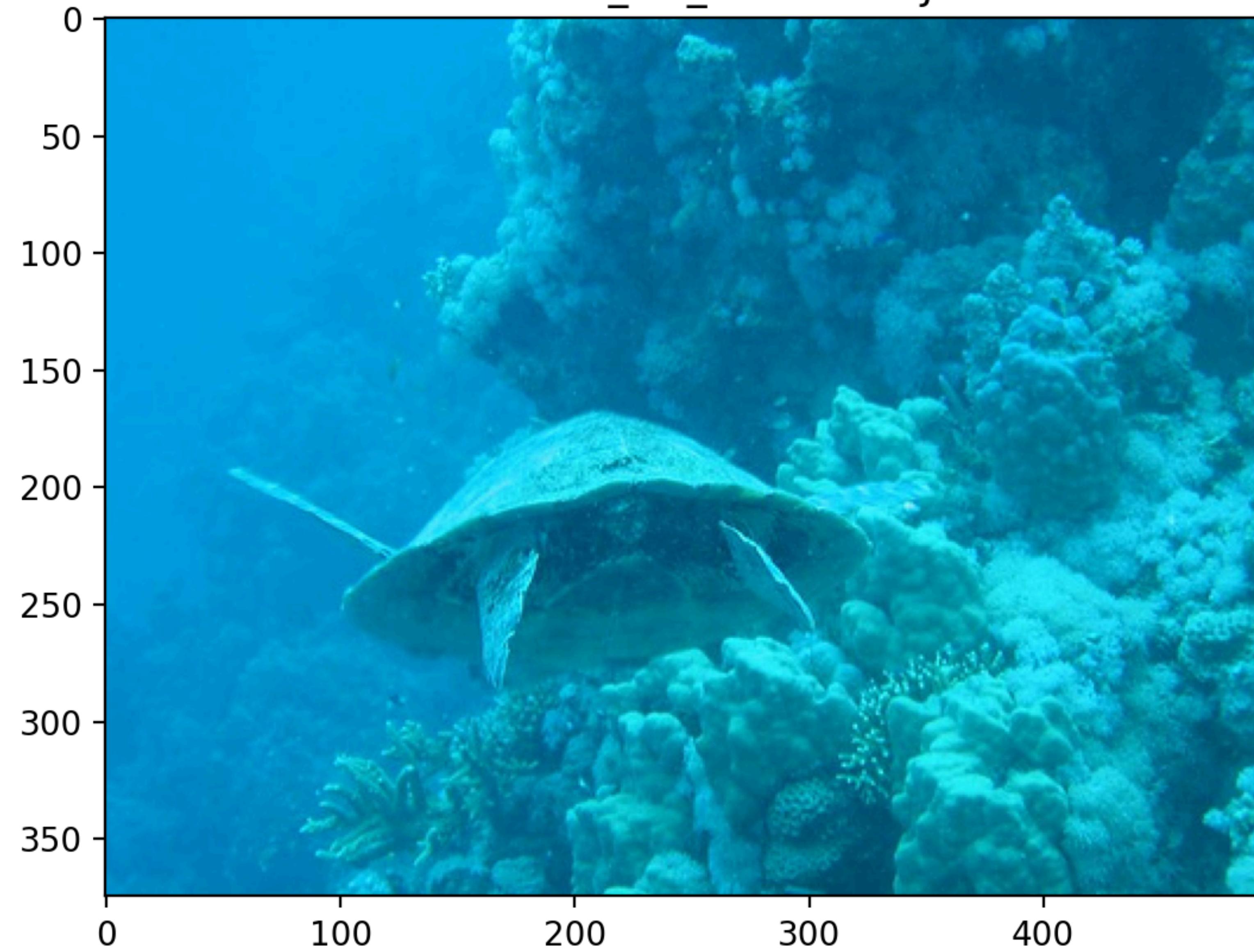
ILSVRC2012_val_00025559.JPEG



ILSVRC2012_val_00047583.jpeg



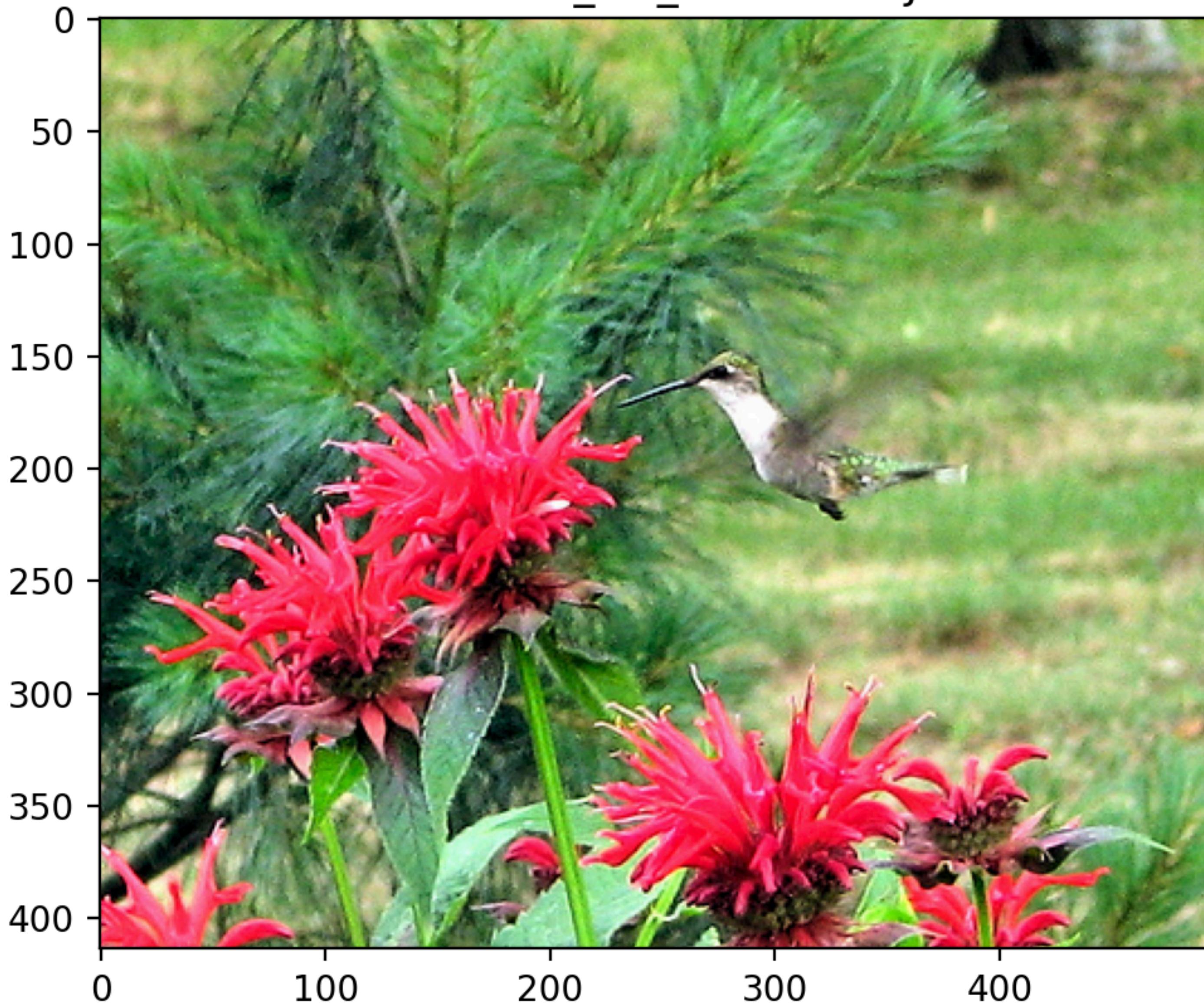
ILSVRC2012_val_00033445.jpeg



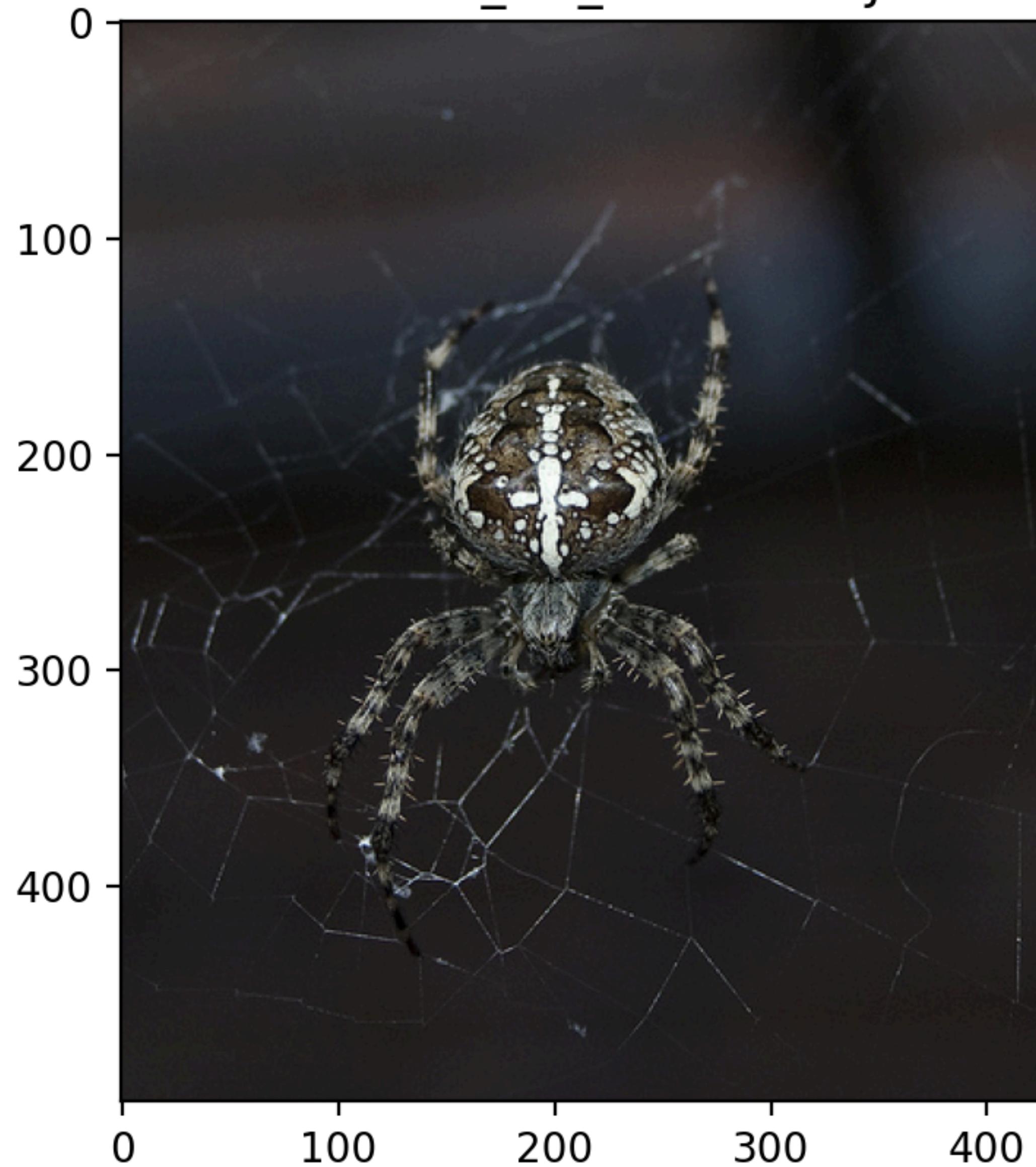
ILSVRC2012_val_00027126.JPEG



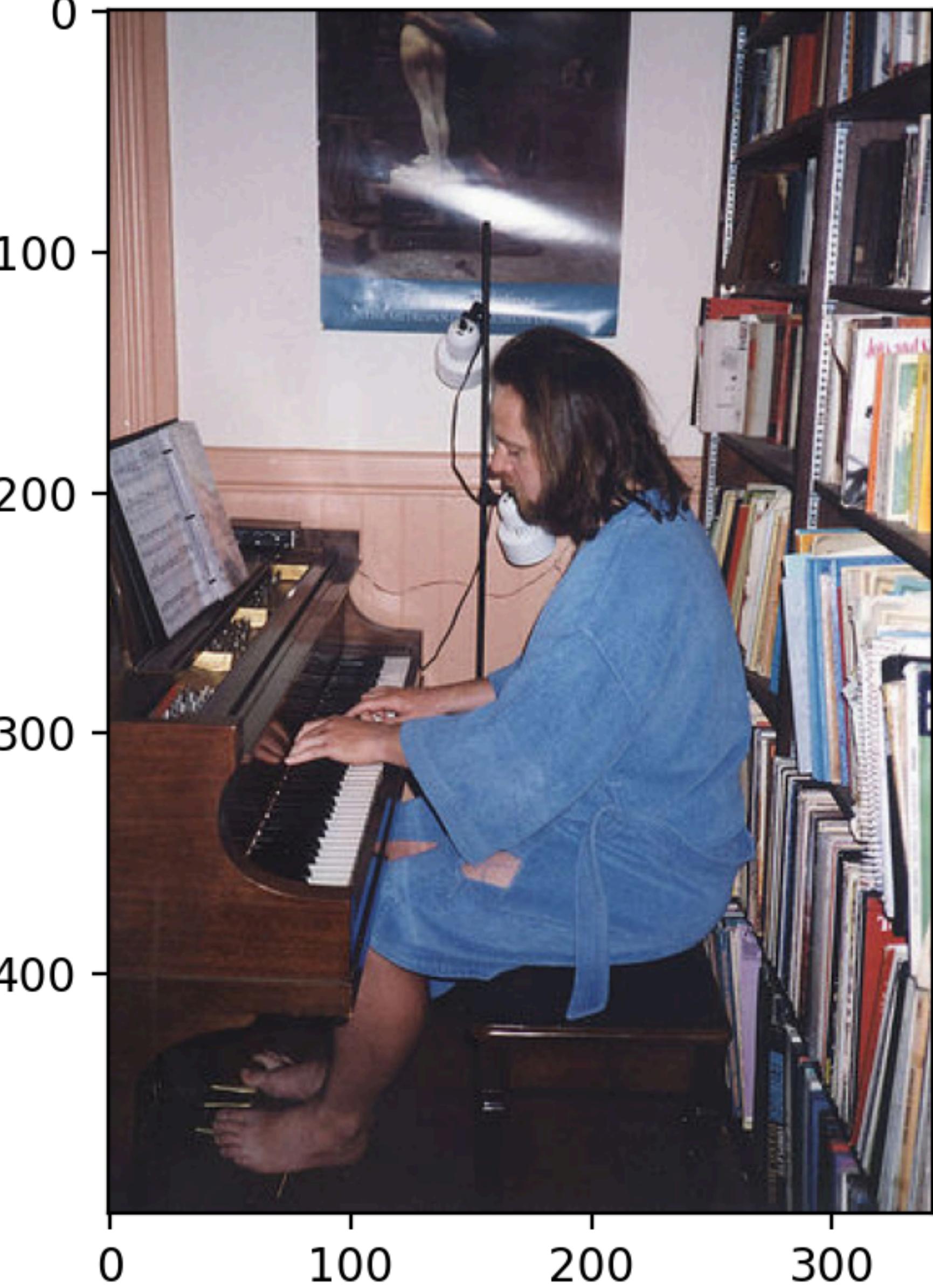
ILSVRC2012_val_00013085.jpeg



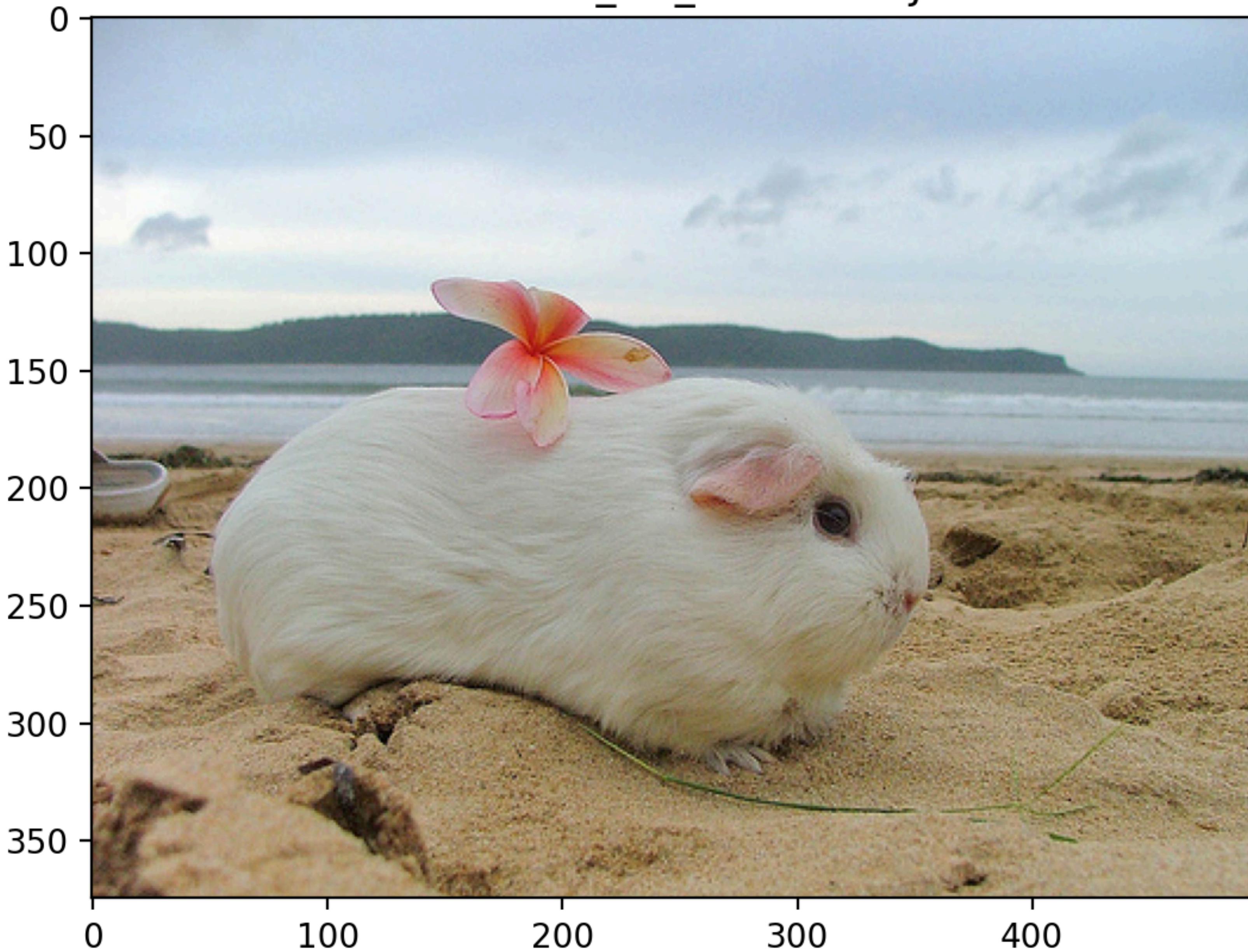
ILSVRC2012_val_00035593.jpeg



ILSVRC2012_val_00012694.JPEG



ILSVRC2012_val_00009233.jpeg



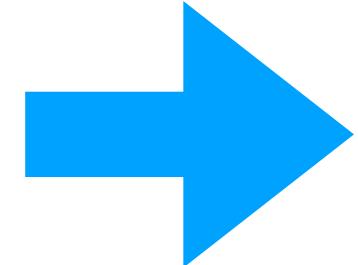
ILSVRC2012_val_00016541.jpeg



ImageNet Competition

ImageNet was about 10% done (already 5 million images!)

Alex Berg (prof at UNC and research scientist at FAIR)



Let's make it a competition!

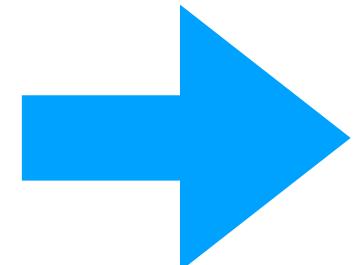


ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Olga Russakovsky (student then postdoc at Stanford)



“Small” version of ImageNet: 1,000 classes, 1.2 million images



“ImageNet” has become equivalent to ILSVRC 2012

IMAGENET Large Scale Visual Recognition Challenge 2010 (ILSVRC2010)

Held as a "taster competition" in conjunction with PASCAL Visual Object Classes Challenge 2010 (VOC2010).

[Registration](#) [Download](#) [Introduction](#) [Data](#) [Task](#) [Development kit](#) [Timetable](#) [Features](#) [Submission](#) [Citation^{new}](#) [Organizers](#)
[Contact](#)

News

- September 2, 2014: [A new paper](#) which describes the collection of the ImageNet Large Scale Visual Recognition Challenge dataset, analyzes the results of the past five years of the challenge, and even compares current computer accuracy with human accuracy is now available. *Please cite it when reporting ILSVRC2010 results or using the dataset.*
- For latest challenge, please visit [here](#).
- September 16, 2010: Slides for [overview of results](#) are available, along with slides from the two winning teams:

Winner: NEC-UIUC

Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu (NEC). LiangLiang Cao, Zhen Li, Min-Hsuan Tsai, Xi Zhou, Thomas Huang (UIUC). Tong Zhang (Rutgers).

[PDF] **NB: This is unpublished work. Please contact the authors if you plan to make use of any of the ideas presented.**

Honorable mention: XRCE

Jorge Sanchez, Florent Perronnin, Thomas Mensink (XRCE)

[PDF] **NB: This is unpublished work. Please contact the authors if you plan to make use of any of the ideas presented.**

- September 3, 2010: [Full results](#) are available. Please join us at the [VOC workshop](#) at ECCV 2010 on 9/11/2010 at Crete, Greece. At the workshop we will provide an overview of the results and invite winning teams to present their methods. We look forward to seeing you there.
- August 9, 2010: Submission deadline is extended to [4:59pm PDT, August 30, 2010](#). There will be no further extensions.
- August 8, 2010: [Submission site](#) is up.
- June 16, 2010: Test data is available for [download!](#).
- May 3, 2010: Training data, validation data and development kit are available for [download!](#).
- May 3, 2010: [Registration](#) is up!. Please register to stay updated.
- Mar 18, 2010: We are preparing to run the ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC2010)

ImageNet Classification Task

Training data: 1.2 million images for 1,000 classes (roughly class-balanced)

Validation set: 50,000 images for 1,000 classes (exactly class-balanced)

Test set: 150,000 images for 1,000 classes (exactly class-balanced, hidden labels)

Evaluation metric: **Top-5 accuracy**

- Five predictions per image
- Prediction counts as correct if the image label is among the five predictions

Why? Sometimes multiple labels per image, sometimes unclear class boundaries.
+ task is already hard enough

ILSVRC2012_val_00016541.jpeg



n03950228

WordNet ID (wnid)

pitcher, ewer

Synonym set

ILSVRC2012_val_00007151.jpeg



© 1999 Ledbetter

n02488702

colobus, colobus monkey

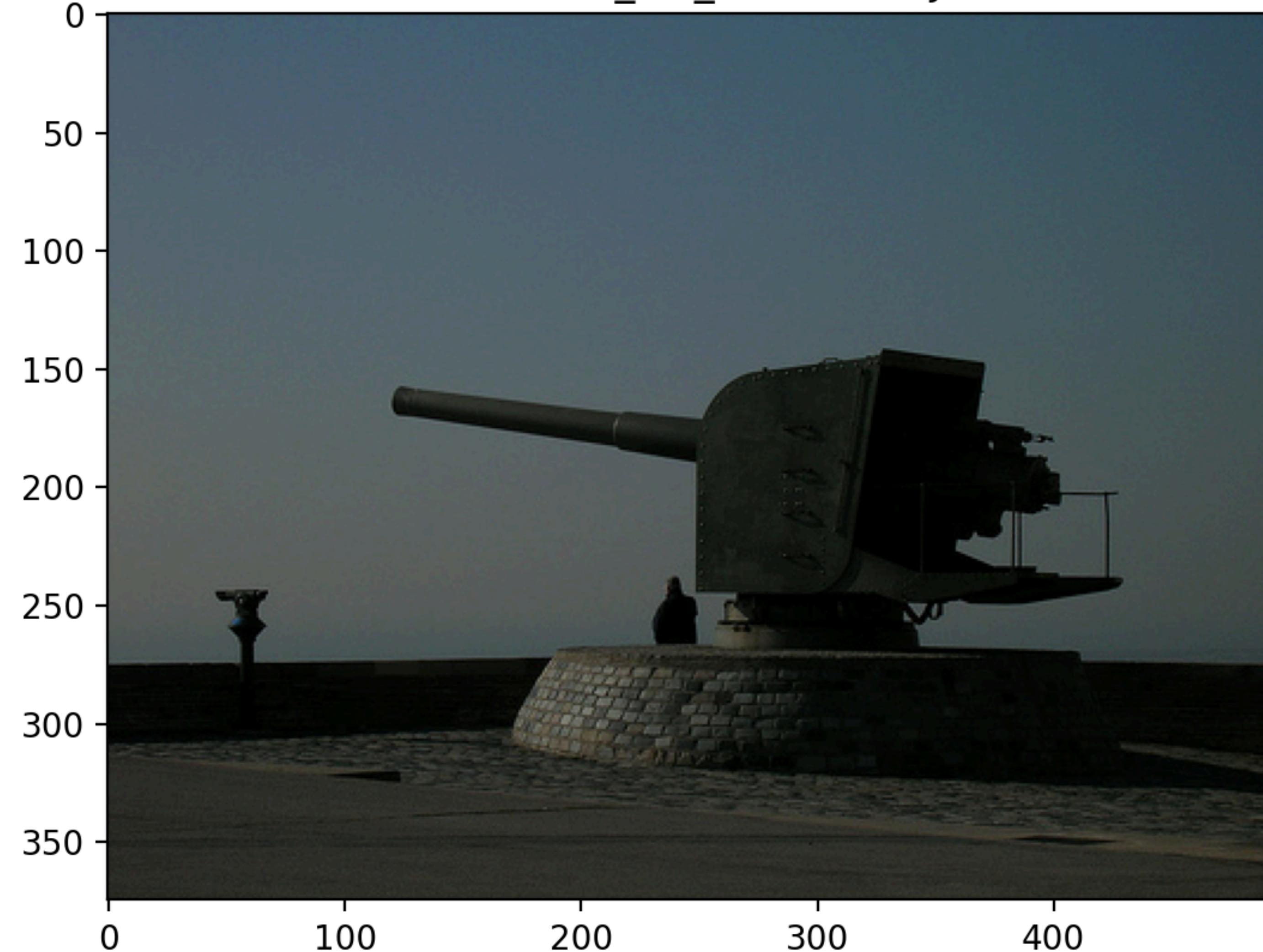
ILSVRC2012_val_00042060.jpeg



n03026506

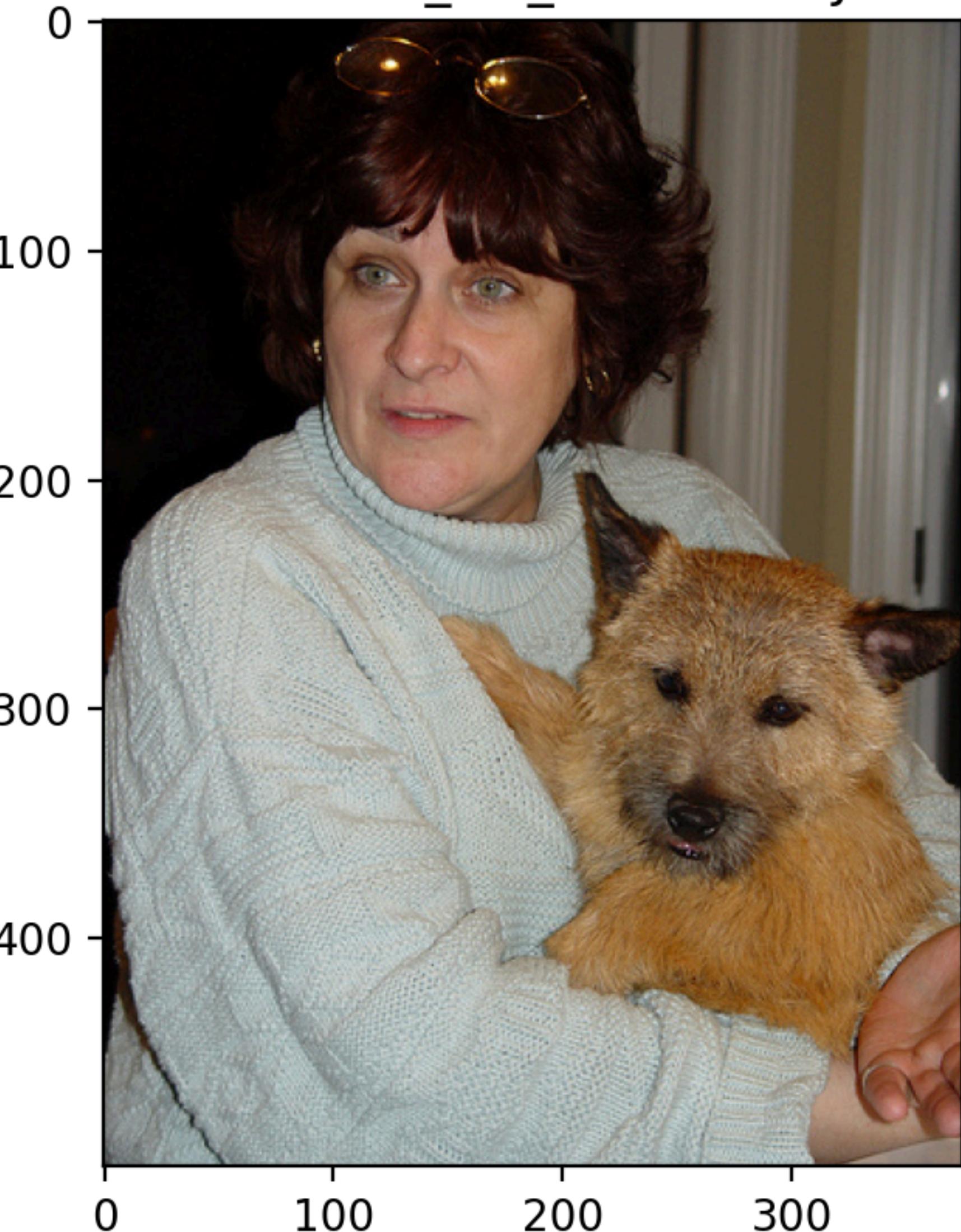
Christmas stocking

ILSVRC2012_val_00001902.jpeg



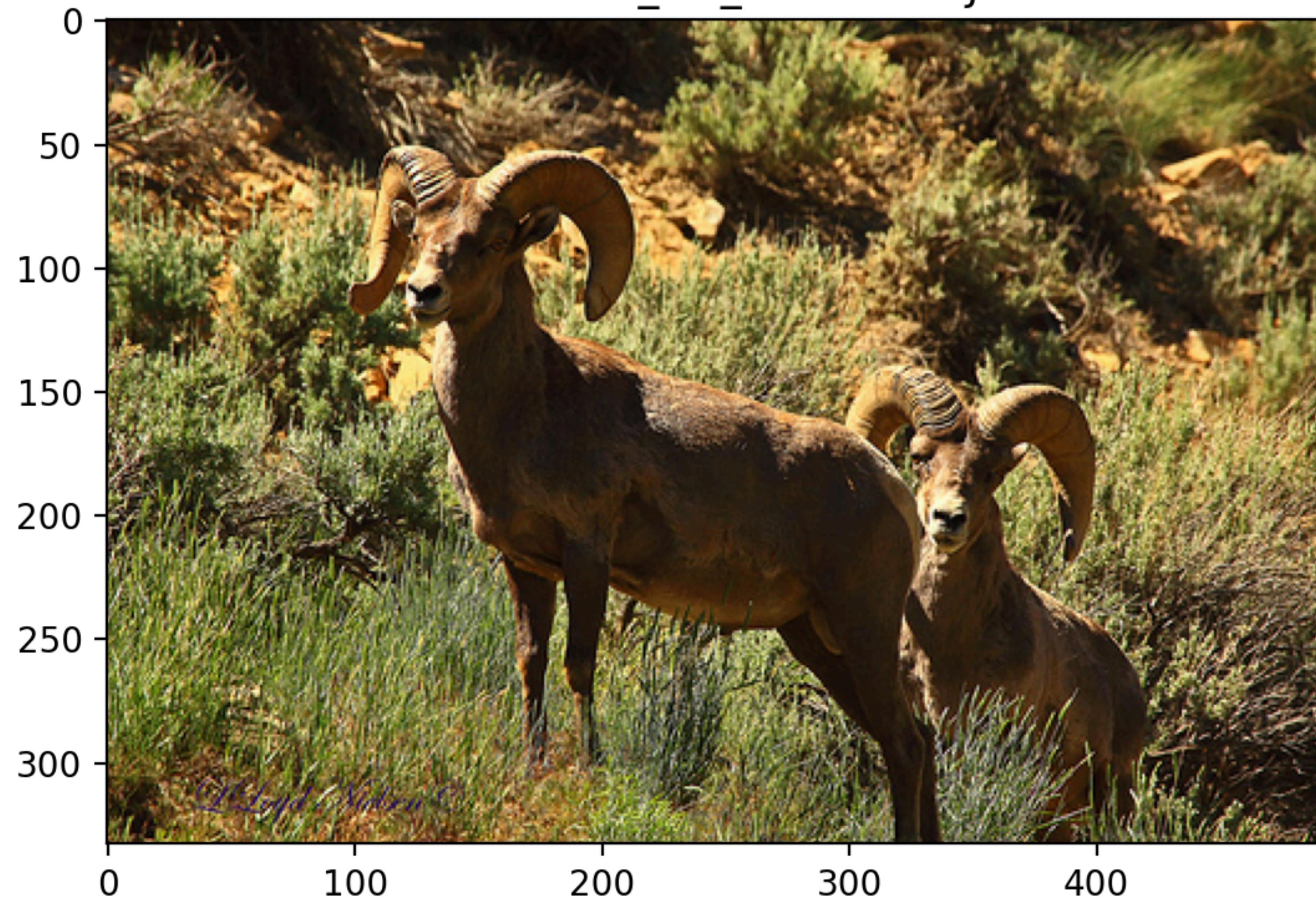
n02950826 cannon

ILSVRC2012_val_00007880.JPEG



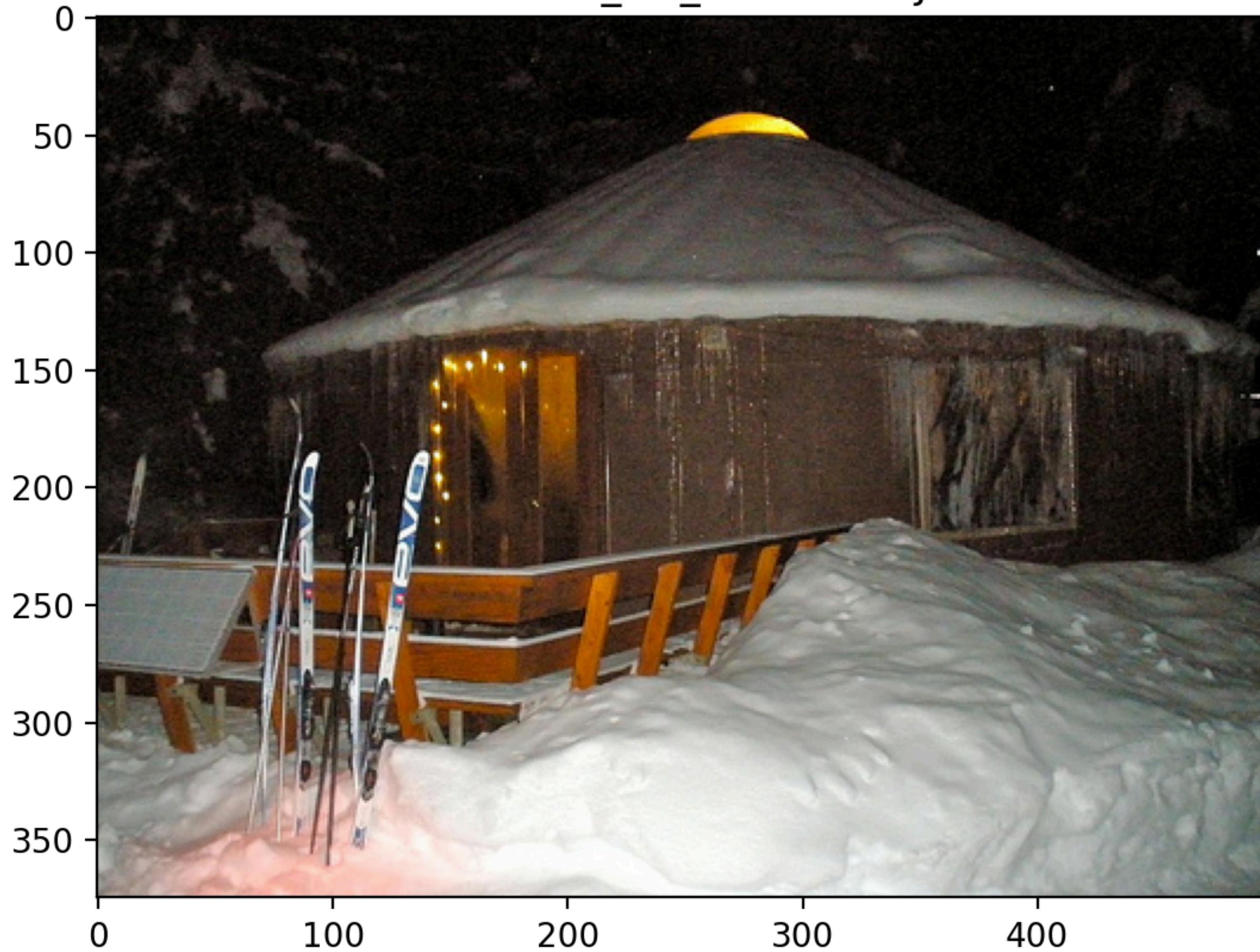
n02094258 Norwich terrier

ILSVRC2012_val_00016391.jpeg



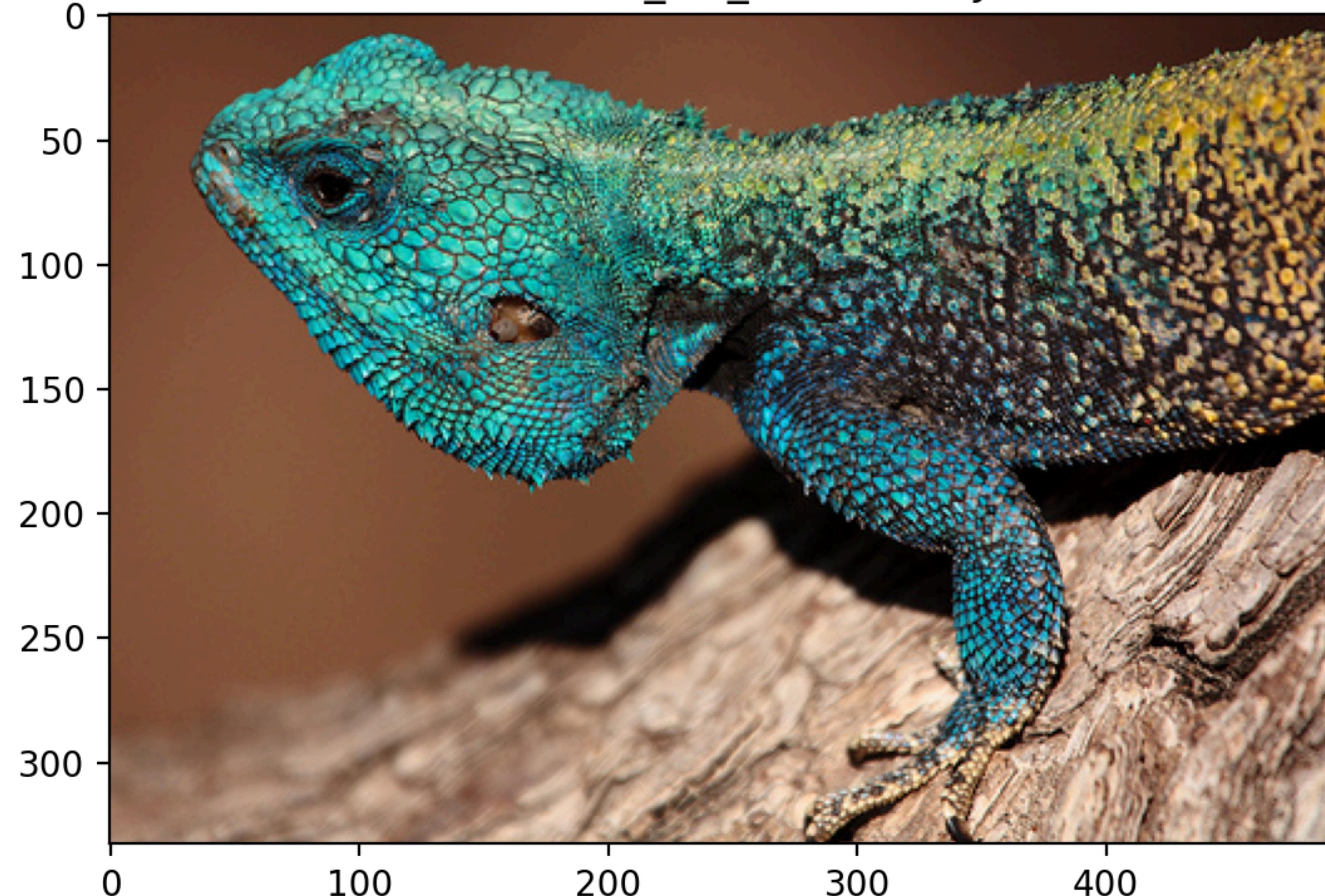
n02412080 ram, tup

ILSVRC2012_val_00020151.jpeg



n04613696 yurt

ILSVRC2012_val_00041169.jpeg



n01687978 agama

ILSVRC2012_val_00037836.jpeg



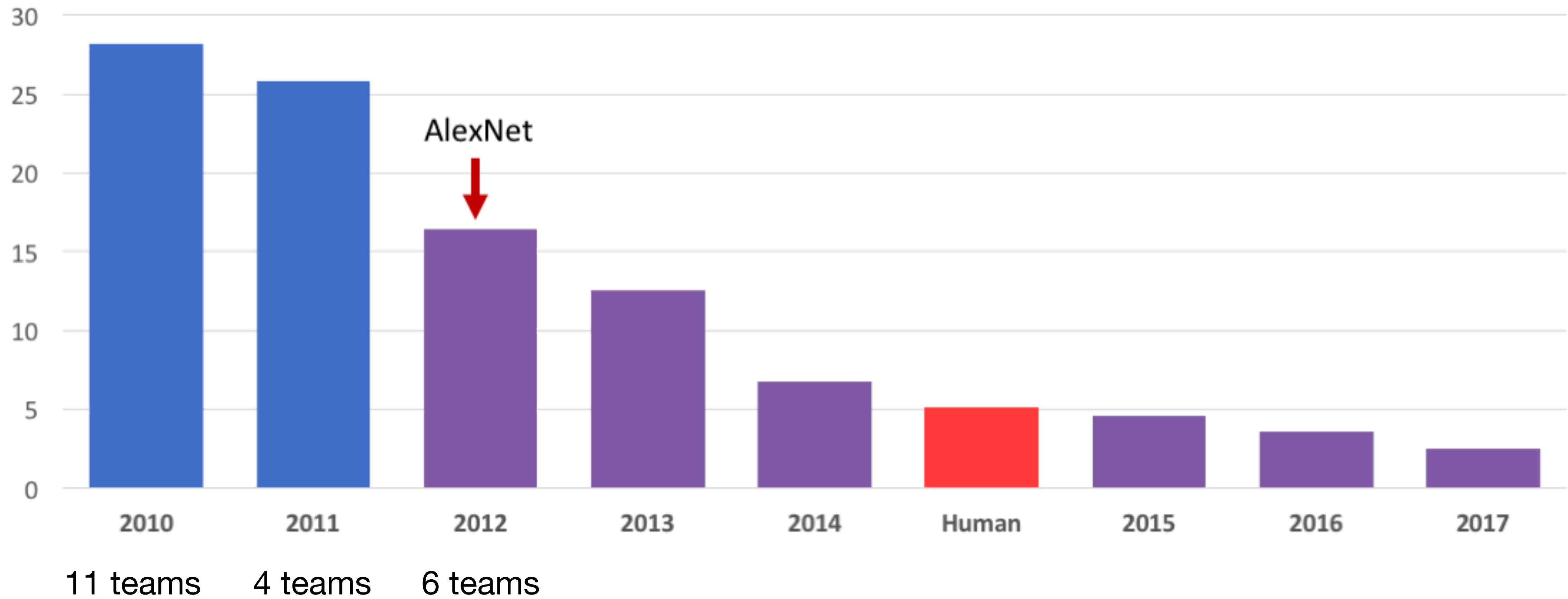
n02134418 sloth bear, *Melursus ursinus*, *Ursus ursinus*

ILSVRC2012_val_00013247.JPG



n04591713 wine bottle

ILSVRC top-5 Error on ImageNet



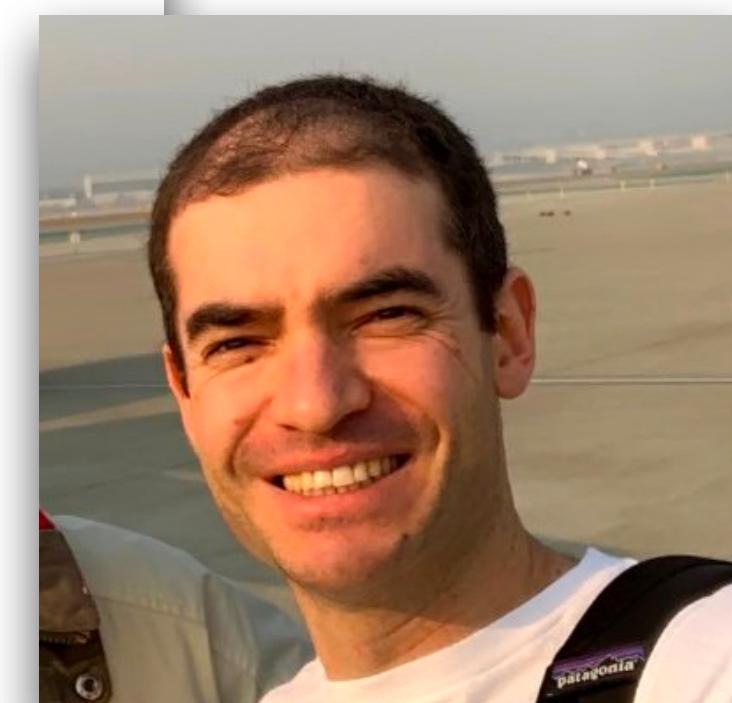
AlexNet

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca



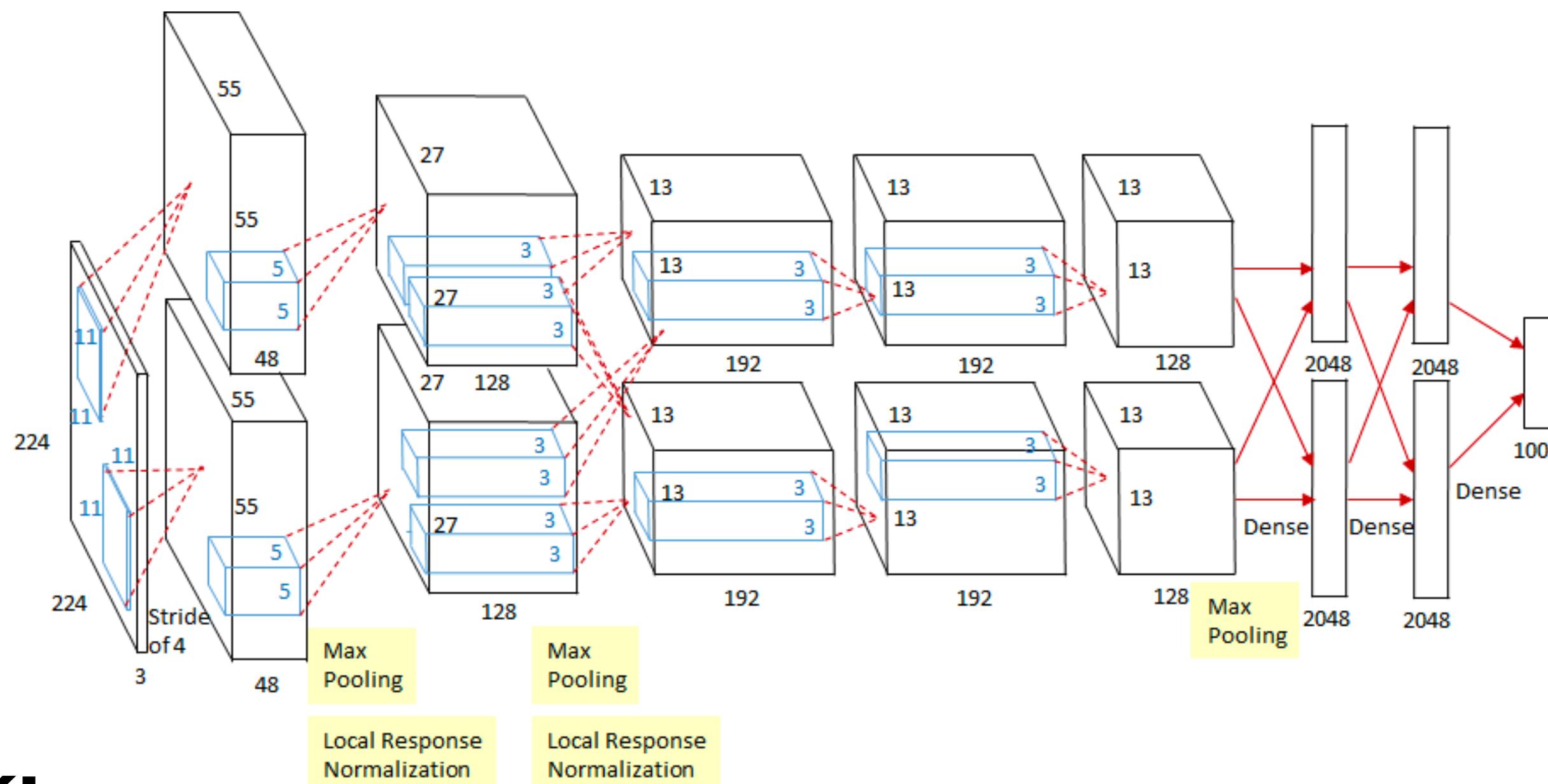
Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

AlexNet

Large **convolutional neural network** (CNN)

Basic idea like in the late 80s, many “tricks” to get it to work on ImageNet



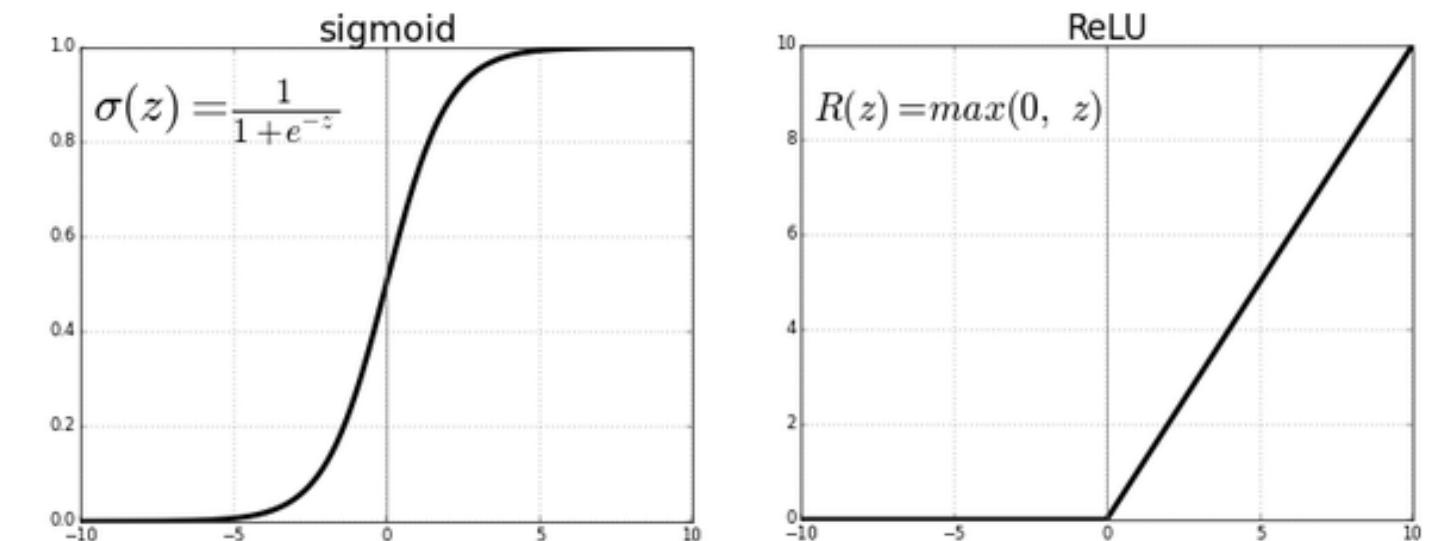
Basic building block:

Structured, learnable linear layer followed by a simple element-wise non-linearity

Repeat the building block several times, add a classification loss at the end.

AlexNet Ingredients

ReLU (rectified linear unit) non-linearity

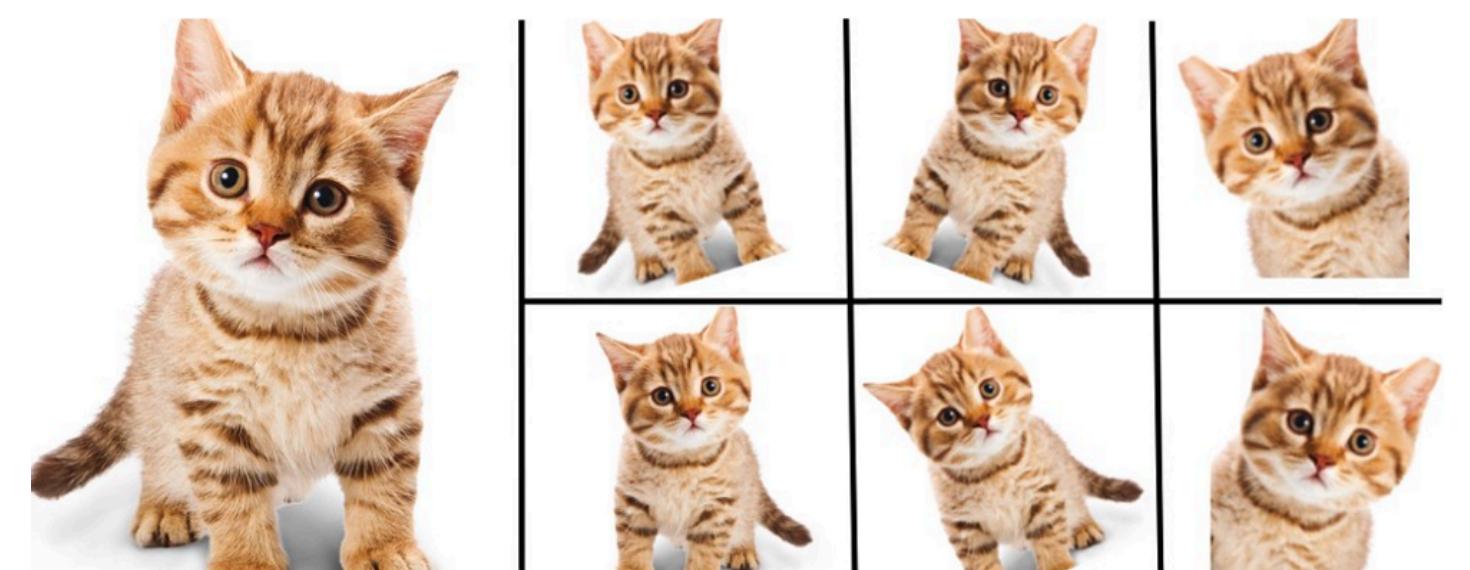


Local response normalization

Training on GPUs



Overlapping pooling



Dropout

Data augmentation

Why these? Each change lead to 0 - 2 percentage points of accuracy improvement.

AlexNet Background

Alex' Masters thesis: “Learning Multiple Layers of Features from Tiny Images”

Built a smaller image classification dataset **CIFAR-10**

- 50,000 images
- 10 classes
- 32x32 pixels
- Subset of a large dataset TinyImages (80 million images)



Alex worked on fast neural network implementations for CIFAR-10.

- Good results, so they decided to scale up the approach
- Alex tuned the model for **one year** on ImageNet

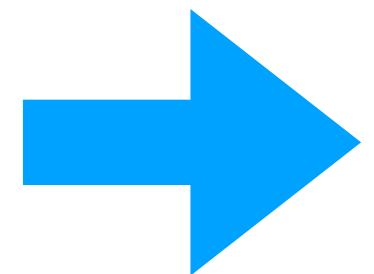
AlexNet Results

Model	Top-1	Top-5
<i>Sparse coding [2]</i>	47.1%	28.2%
<i>SIFT + FVs [24]</i>	45.7%	25.7%
CNN	37.5%	17.0%

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

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

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were “pre-trained” to classify the entire ImageNet 2011 Fall release. See Section 6 for details.



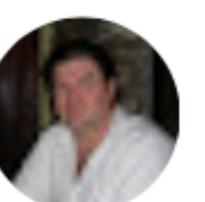
About 9 percentage points improvement over previous state-of-the art

Immediate Controversy in 2012

 **Yann LeCun** ▶ Public Oct 13, 2012

+[Alex Krizhevsky](#)'s talk at the ImageNet ECCV workshop yesterday made a bit of a splash. The room was overflowing with people standing and sitting on the floor. There was a lively series of comments afterwards, with +[Alyosha Efros](#), Jitendra Malik, and I doing much of the talking.

 **Svetlana Lazebnik** +1 Oct 13, 2012
Too bad I couldn't be there! Any take-away points for those of us who couldn't attend? +[Alyosha Efros](#), I'd love to get your take as well!

 **Yann LeCun** Oct 13, 2012
+[Svetlana Lazebnik](#): Our friend +[Alyosha Efros](#) said that ImageNet is the wrong task, wrong dataset, wrong everything. You know him ;-)
Still, he likes the idea of feature learning.



Alyosha Efros +11

Something like that... :) I do like feature learning, the less supervised -- the better. So, I am excited that people are working in this direction, but I am not ready to declare success until they can show improvement on PASCAL detection. Basically, I think ImageNet is just too easy (+[Yann LeCun](#) did confirm that it's easier than PASCAL in terms of objects being more centered and little scale variation). In my view, the important thing to look at is chance performance. Chance on PASCAL detection is something like 1 in a million. Chance on Imagenet classification is 1 in 200 (easier than Caltech-256!!!). Chance on ImageNet detection is lower but still maybe around 1 in a thousand or so.

When chance is so high, the temptation for a classifier to overfit to the bias in the data is too great. The fact that "t-short" category turned out to be one of the easiest ones for all the classifiers in the competition should give us pause as to whether

th



Yann LeCun +16

This is not a religious war between deep learning and computer vision. Everyone wins when someone improves a result on some benchmark. No one should feel "defeated", and no one should give up unless they no longer believe in what they are doing. Progress is always exciting, particularly when it comes from a brand new way of doing things, rather than from a carefully tweaked combination of existing methods.

Oct
14,
2012



Geoffrey Hinton +31

I predicted that some vision people would say that the task was too easy if a neural net was successful. Luckily I know Jitendra so I asked him in advance whether this task would really count as doing proper object recognition and he said it would, though he also said it would be good to do localization too. To his credit, Andrew Zisserman says our result is impressive.

Oct
15,
2012

I think its pretty amazing to claim that a vision task is "just too easy" when we succeed even though some really good vision

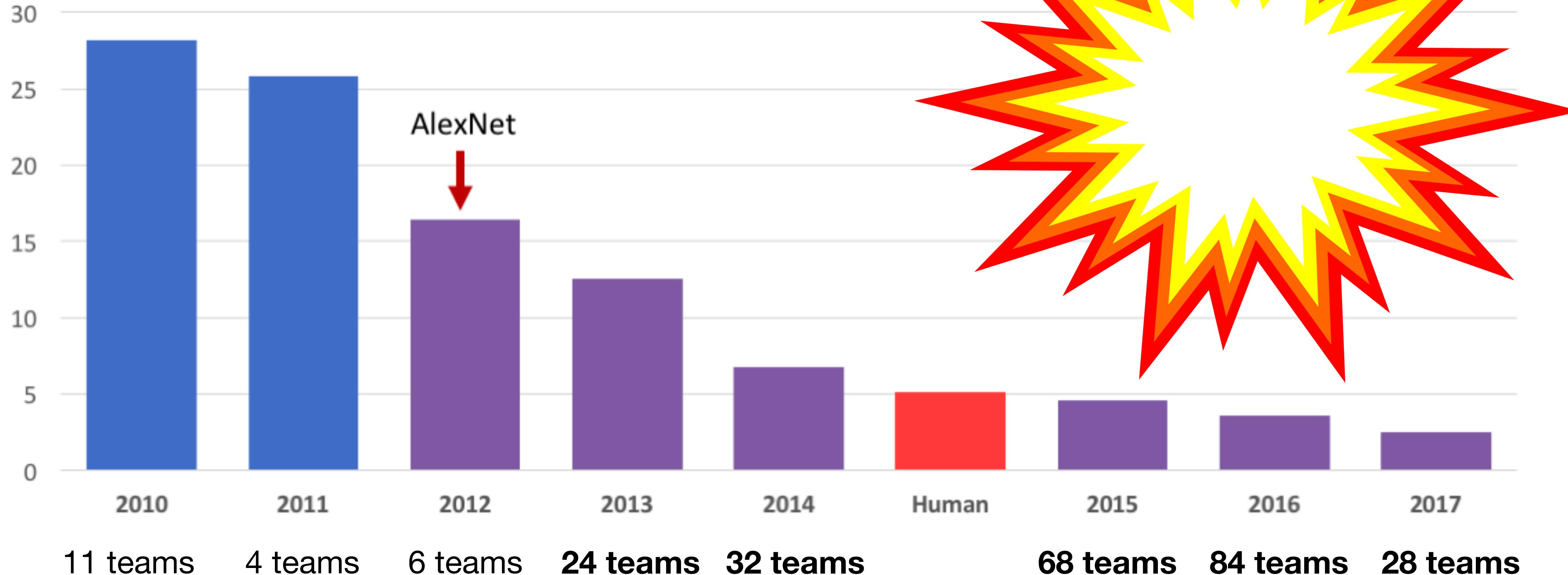
Oct
16,
2012

and at it and failed to do nearly as well. I also think discredit a system that gets about 84% correct by get 0.5% correct by chance is a bit desperate.

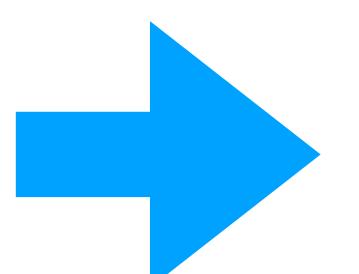
NOTE: Alyosha is a great scientist.

When he's wrong, he's happy to admit it and he is wrong in interesting ways.

ILSVRC top-5 Error on ImageNet



Large improvement, new method



Tremendous interest from the community

Impact on ImageNet

Effectively every team switches to convolutional neural networks.

Subsequent networks

- VGG (2014): up to 19 layers (AlexNet: 8 layers), more parameters
- ResNet (2015): 150 layers, more parameters
- Wide ResNets, ResNeXT, SE-ResNet, EfficientNet, AmoebaNet, MobileNet, Inception, NASNet, DenseNet, SqueezeNet, etc.

Training times **increase** to weeks on dozens of GPUs (\$30k) ...

... and decrease by orders of magnitude (\$100 for a ResNet)

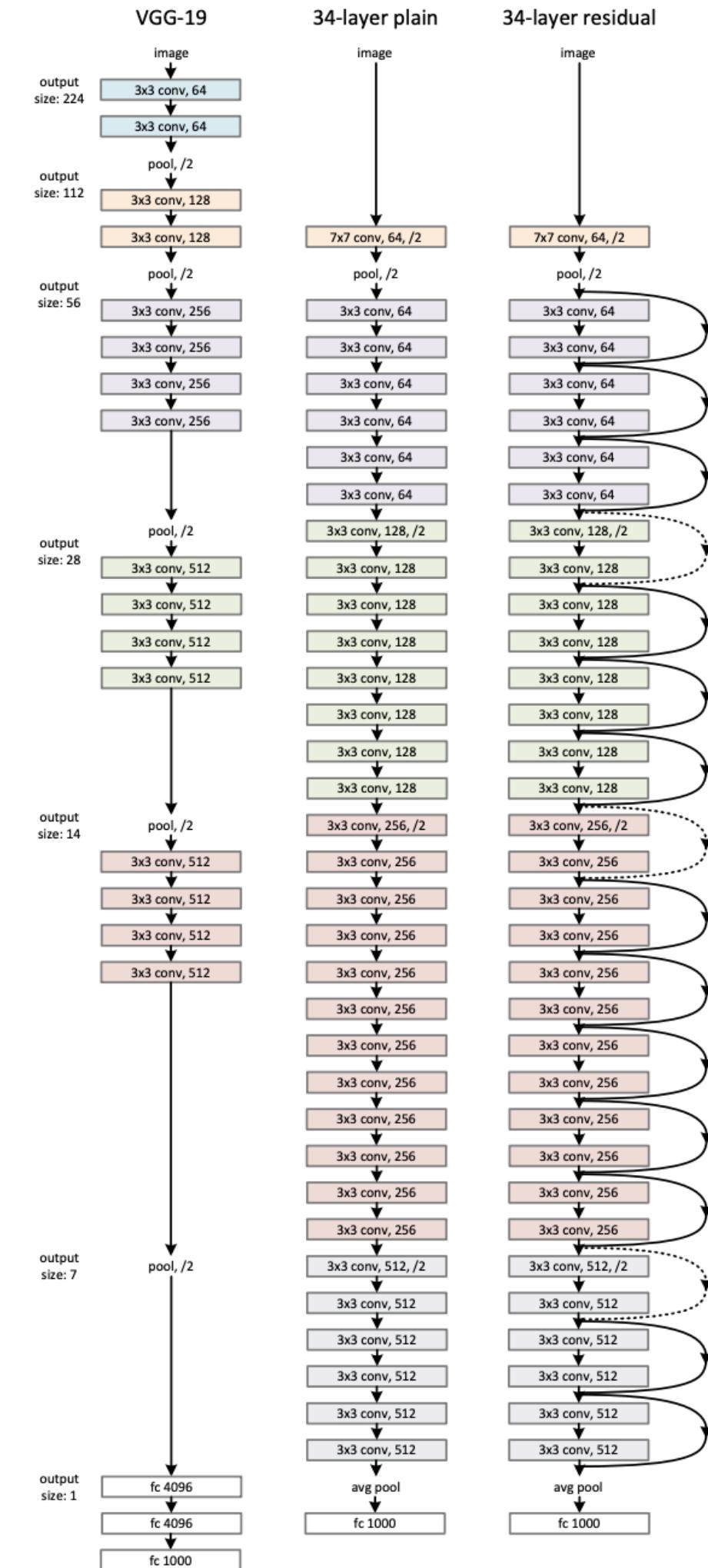
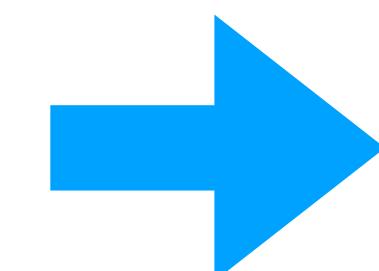
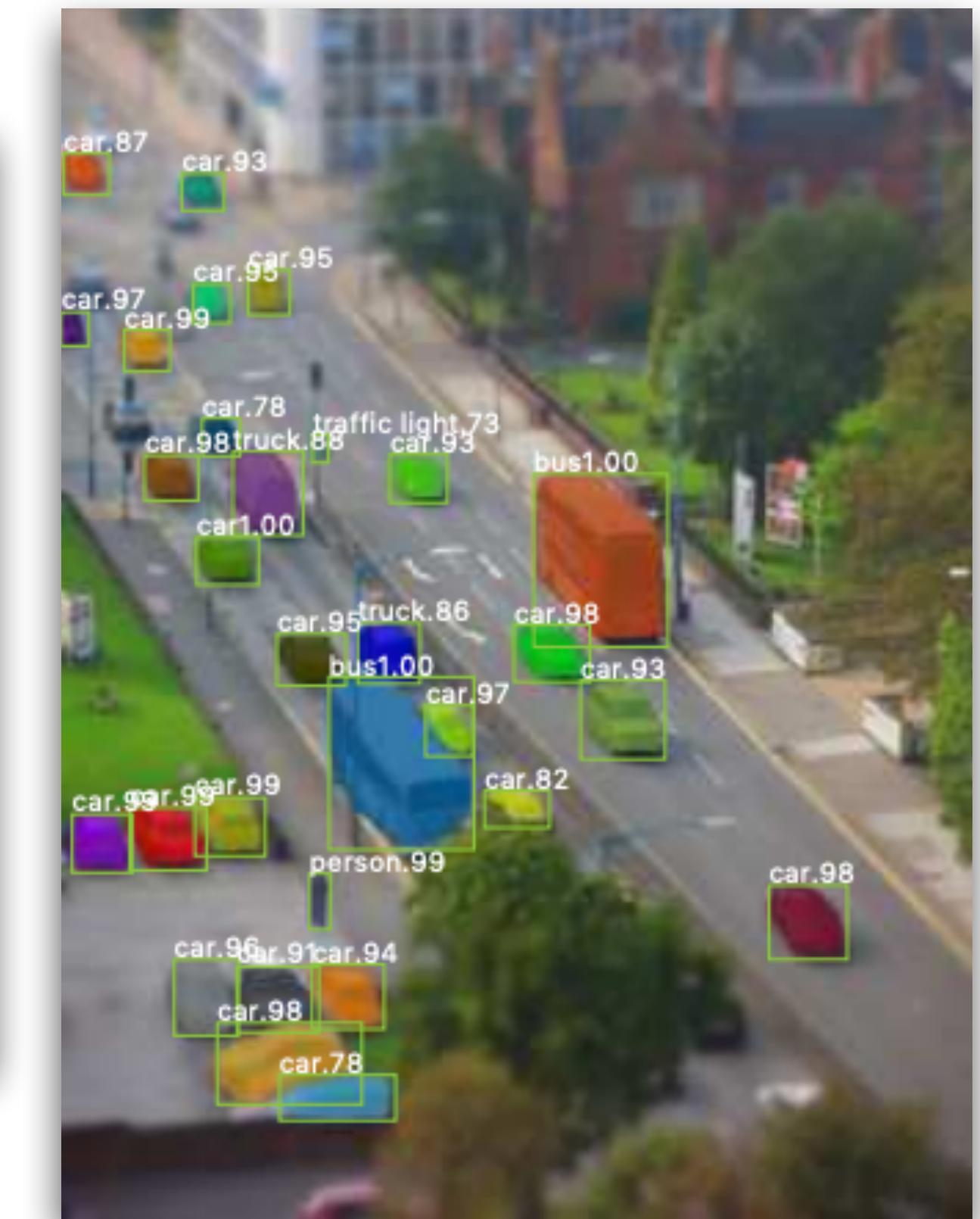


Figure 3. Example network architectures for ImageNet. **Left:** the VGG-19 model [41] (19.6 billion FLOPs) as a reference. **Middle:** a plain network with 34 parameter layers (3.6 billion FLOPs). **Right:** a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase dimensions. **Table 1** shows more details and other variants.

Impact on Computer Vision

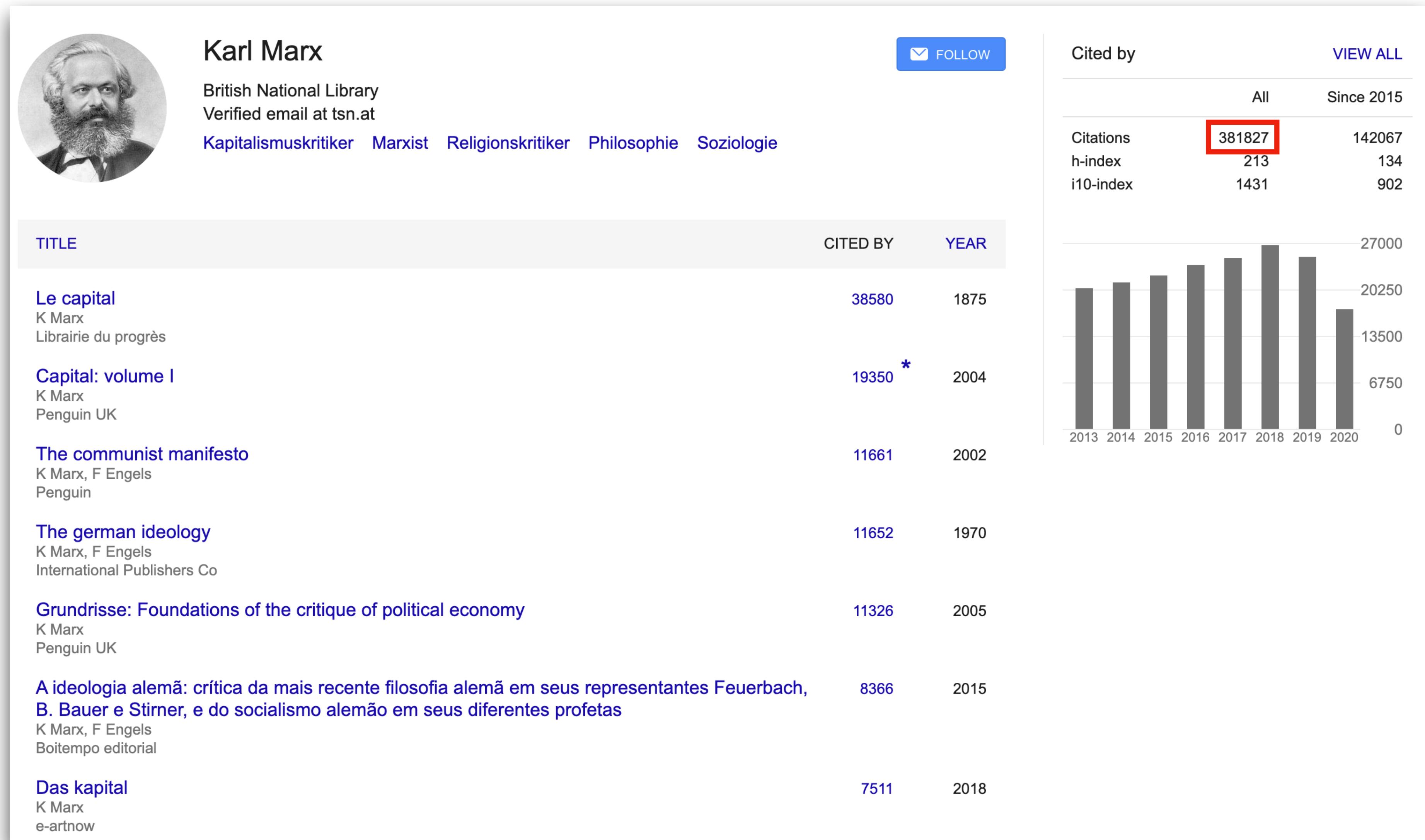
Effectively the entire field switches to convolutional neural networks.

- Object detection
 - Image segmentation
 - Pose estimation
 - 3D reconstruction
 - Image inpainting
 - Generative models
 - etc.

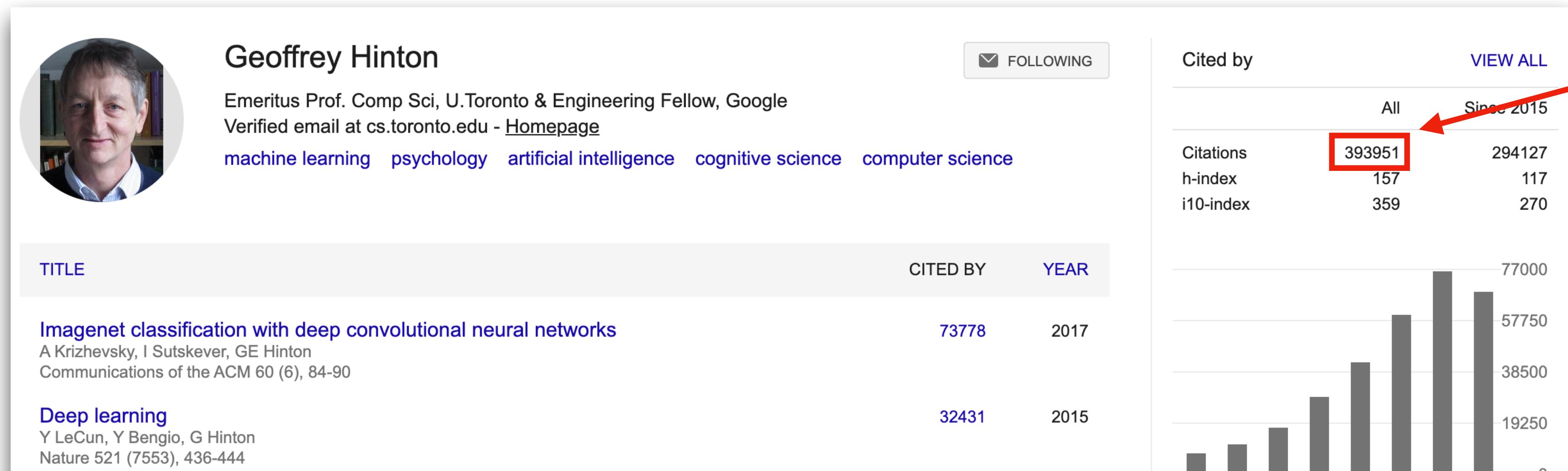


Deep learning revolution in computer vision

Historical Comparison - Revolutions



Historical Comparison - Revolutions



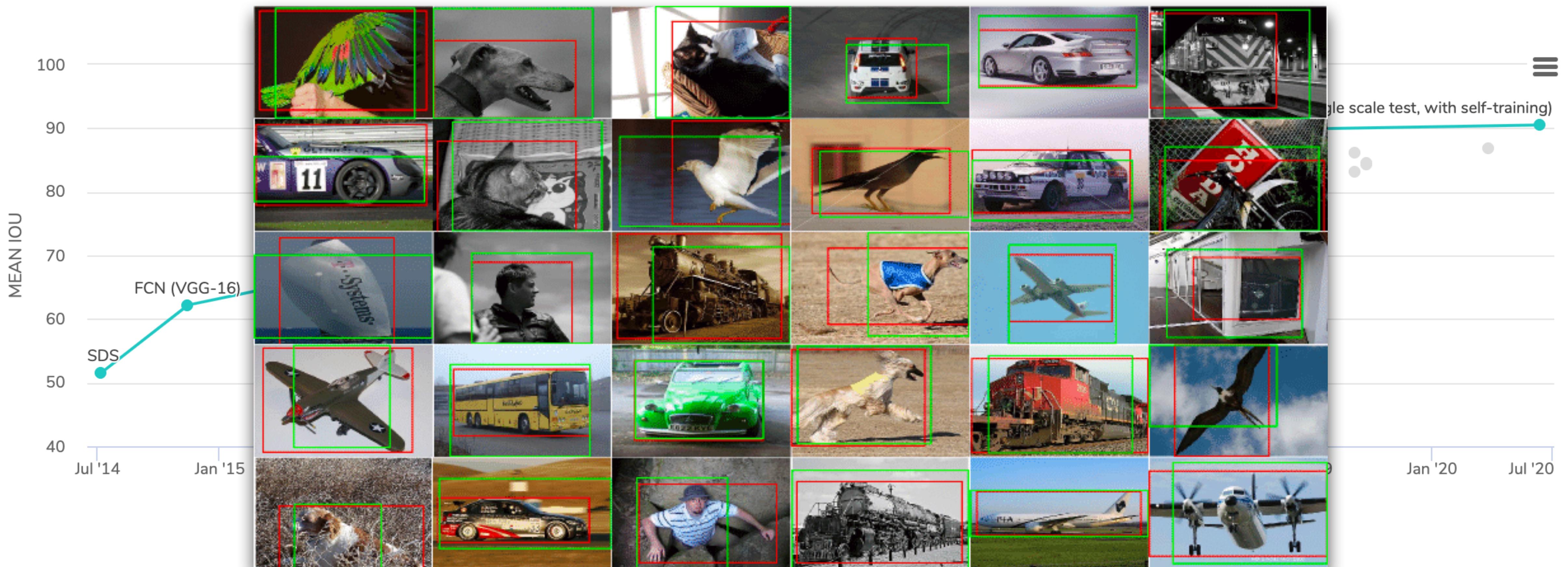
CAVEAT: DO NOT MEASURE SCIENCE BY CITATION COUNT



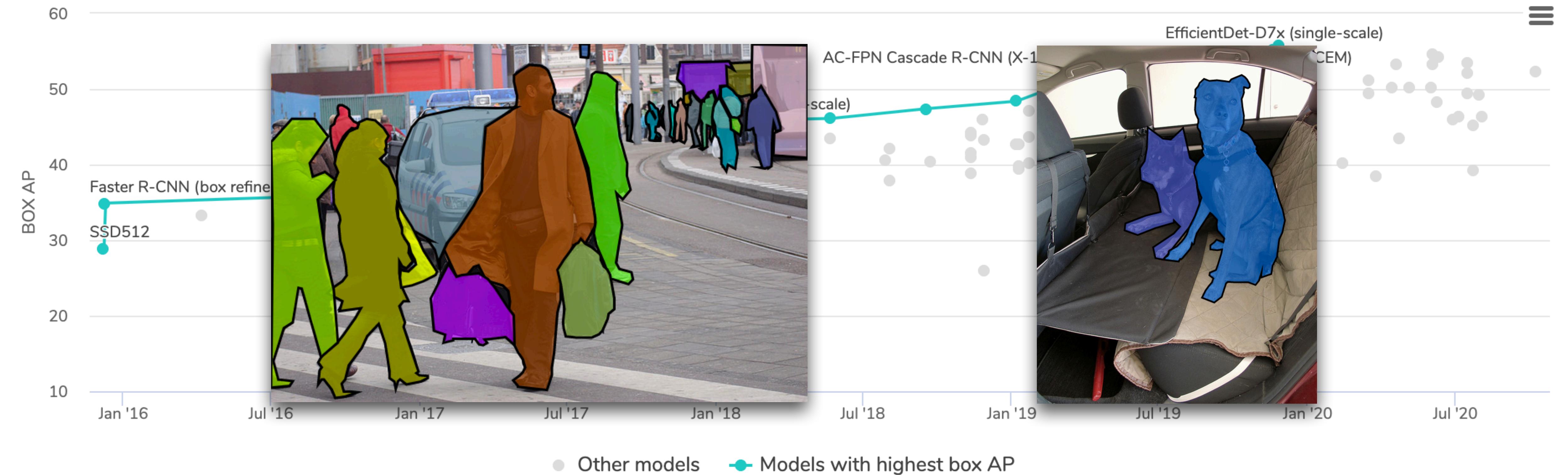
Learning internal representations by error propagation DE Rumelhart, GE Hinton, RJ Williams MIT Press, Cambridge, MA 1 (318)	26912	1986	George E. Dahl Google Brain
Dropout: a simple way to prevent neural networks from overfitting N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov The journal of machine learning research 15 (1), 1929-1958	23994	2014	Abdelrahman Mohamed Research scientist, Facebook AI ...
Learning representations by back-propagating errors DE Rumelhart, GE Hinton, RJ Williams Nature 323 (6088), 533-536	23115	1986	Vinod Nair Research Scientist, DeepMind

Similar Performance Trends for Many Other Datasets

Object detection (PASCAL VOC)

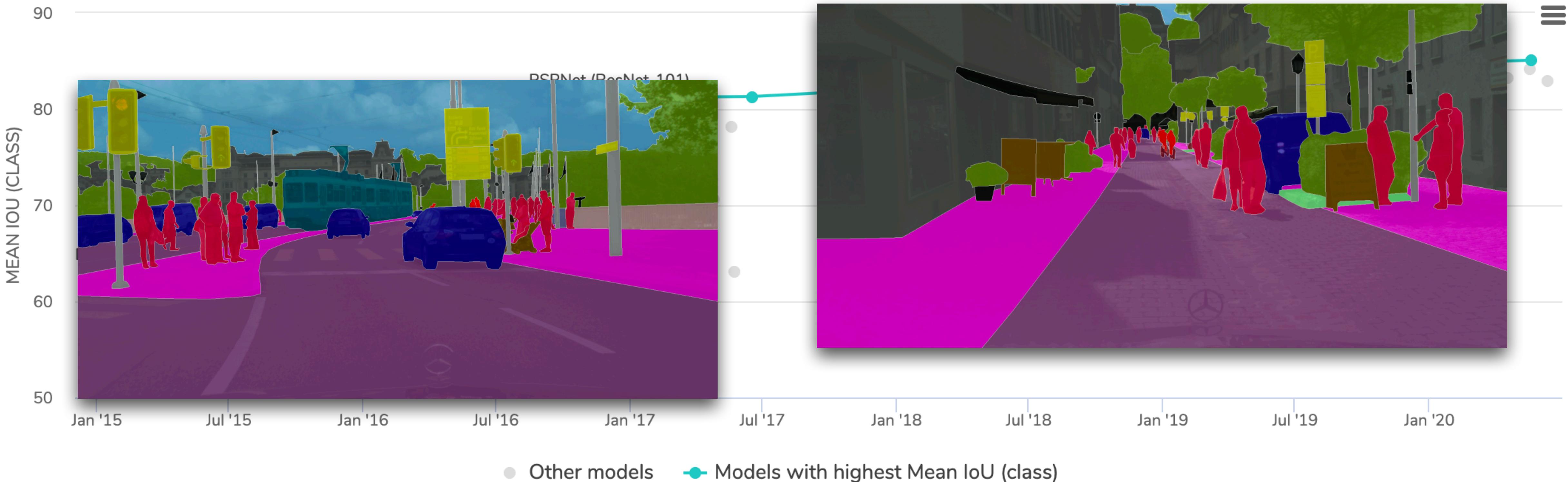


Object Detection (MS COCO)

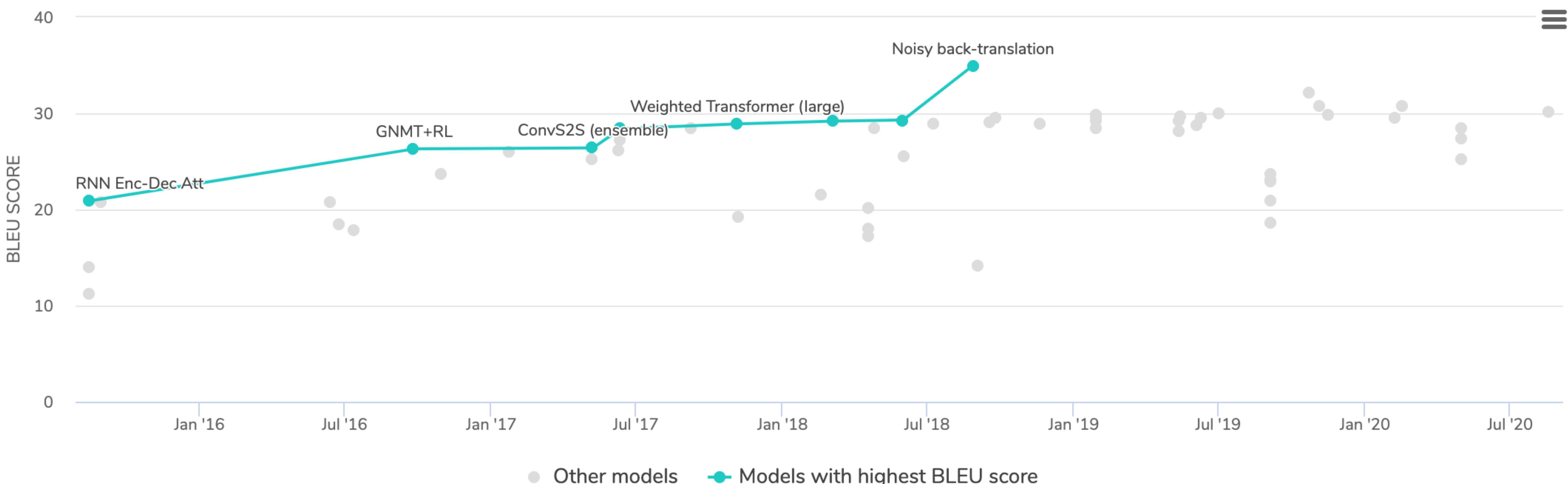


<https://paperswithcode.com/sota>

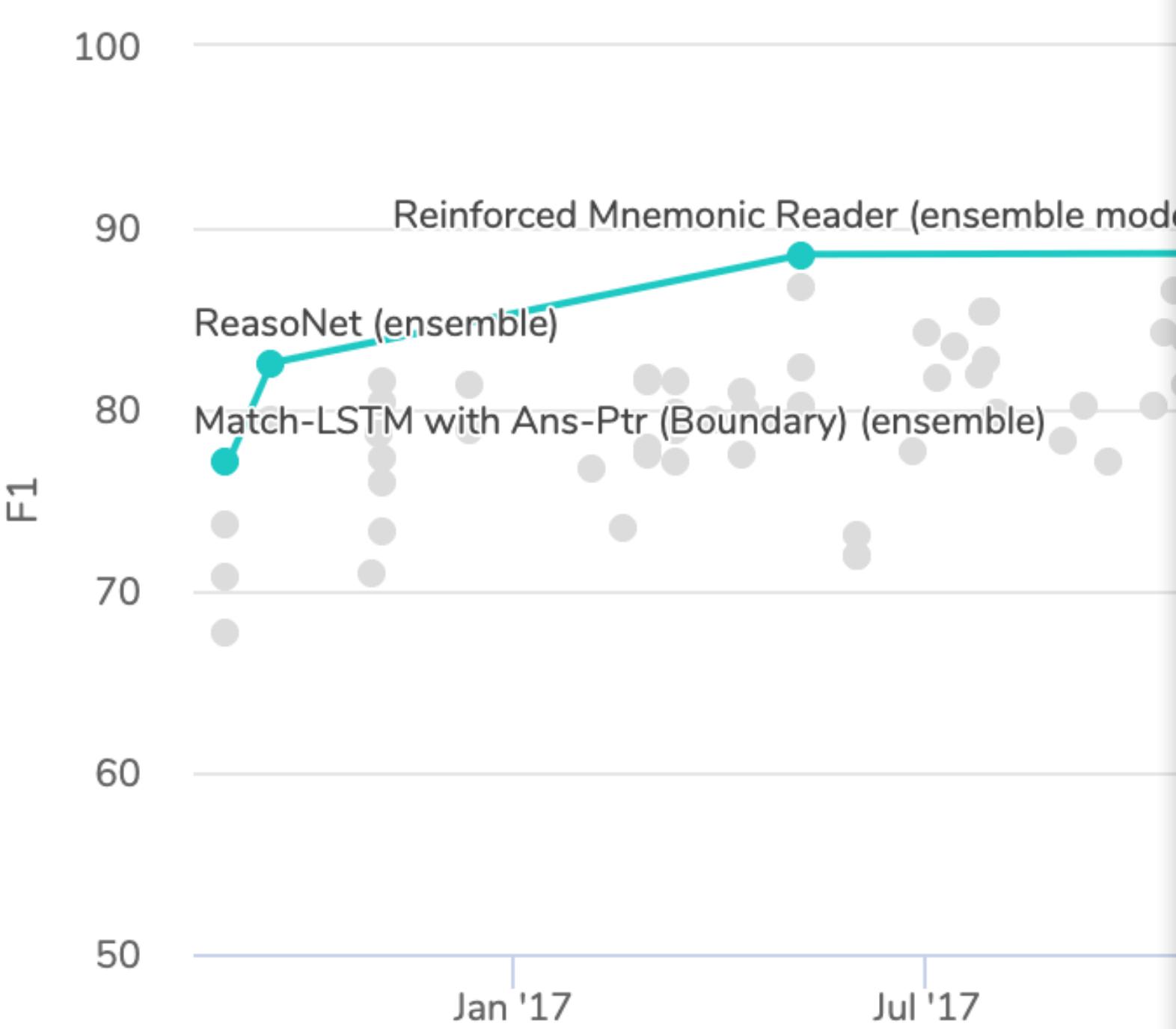
Semantic Segmentation (Cityscapes)



Machine Translation (WMT EN-DE)



Question Answering (SQuAD 1.1)

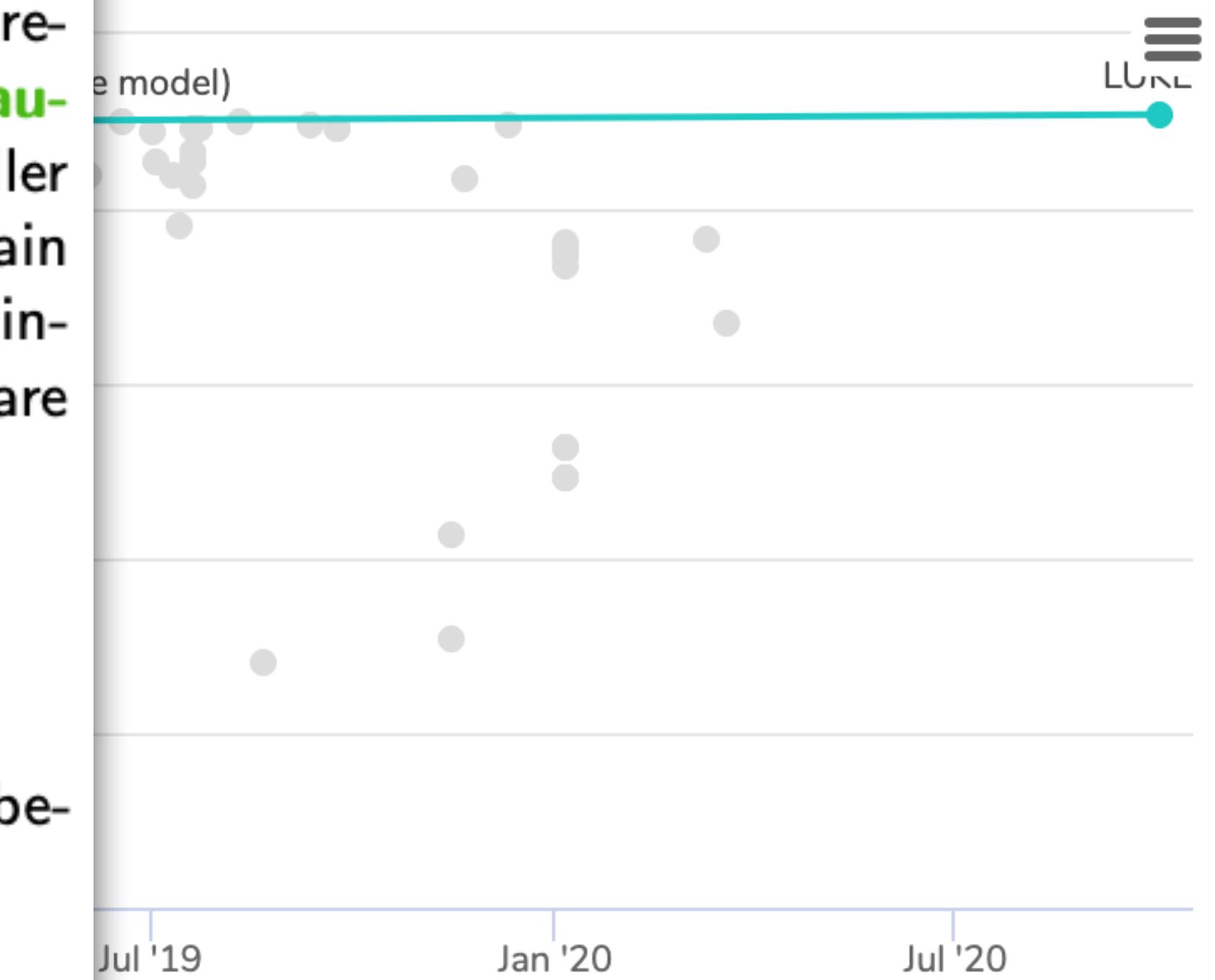


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or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

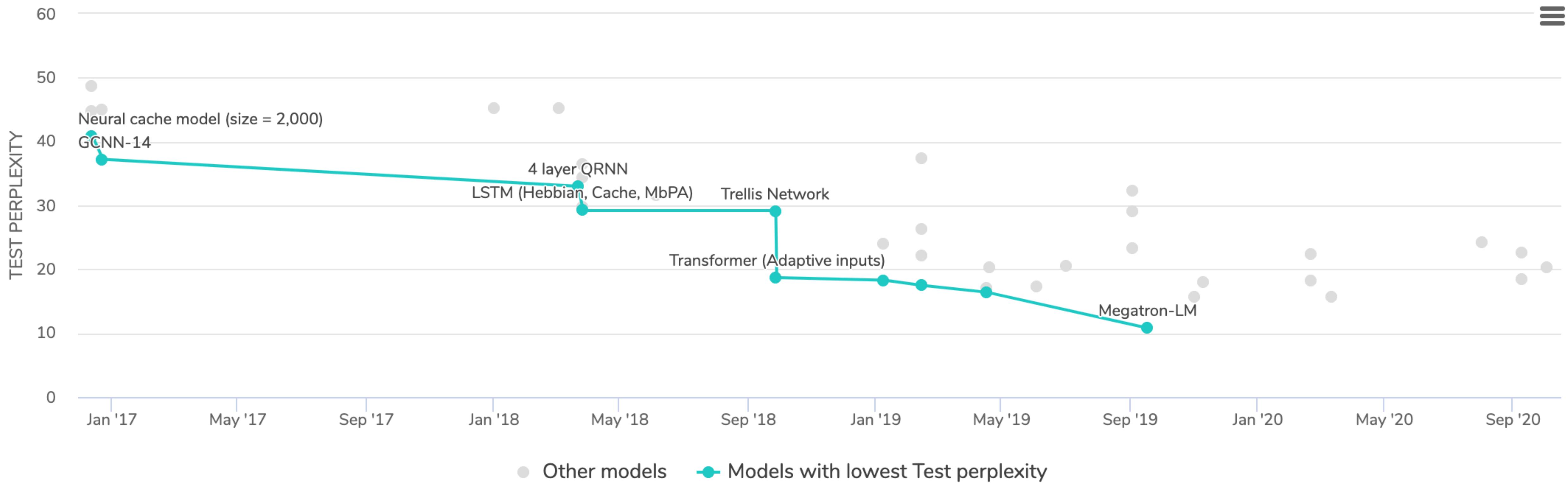
What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud



Language Modeling (WikiText-103)



Key points

Field largely guided by **benchmarks**

Small number of **key datasets** for each task (image classification, detection, etc.)

Algorithmic / model innovations justified by improvements on benchmarks

Algorithmic innovations usually tested on **multiple datasets**

Little to no **mathematical theory**

Substantial **progress** on a wide range of benchmarks

Culture shift

2000 - 2010

- Support vector machines & kernels
- Boosting
- Matrix factorization and tensor methods
- Compressed sensing / high-dim stats
- Convex optimization

Empirical progress usually goes
hand in hand with theoretical results

2010 - 2020

- Convolutional neural networks
- Recurrent neural networks
- Transformers (NLP)
- Network architecture improvements
- Zoo of different architectures

Empirical progress usually comes
without mathematical theory

Culture shift

2000 - 2010

Empirical progress usually goes hand in hand with theoretical results

Emphasis on **provable guarantees**

Optimization problems often **convex**

No specialized hardware

2010 - 2020

Empirical progress usually comes without mathematical theory

Emphasis on **benchmarks**

Non-convexity is fine

Large-scale purely experimental work

Still major caveats with benchmarks

Excitement about experimental results, rapid growth in machine learning

But: even results on datasets like ImageNet remained controversial until recently.

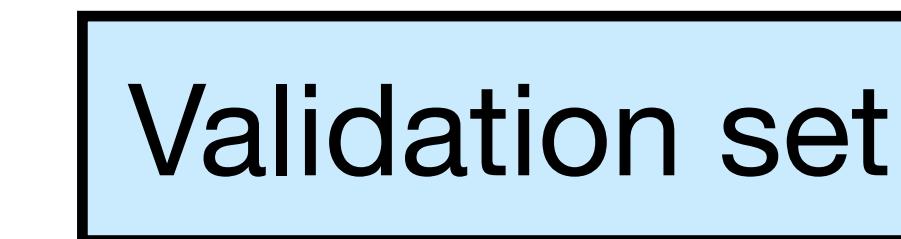
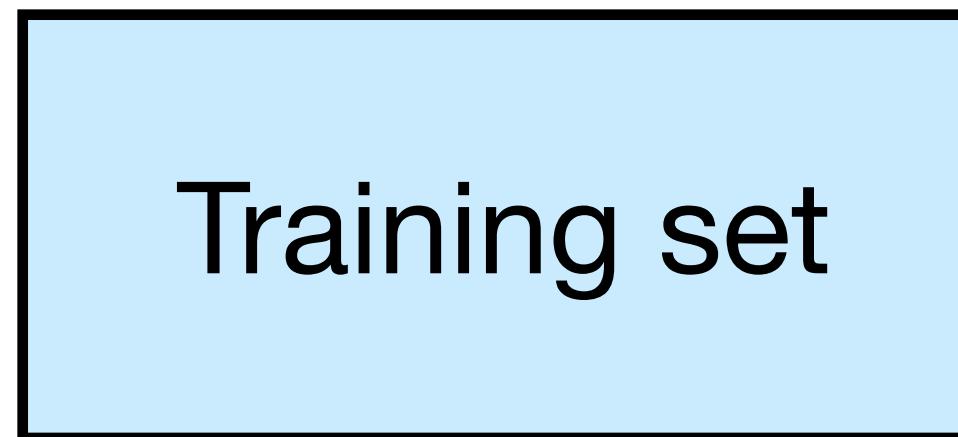
One common criticism: **overfitting from test set re-use**

Ideal ML Workflow

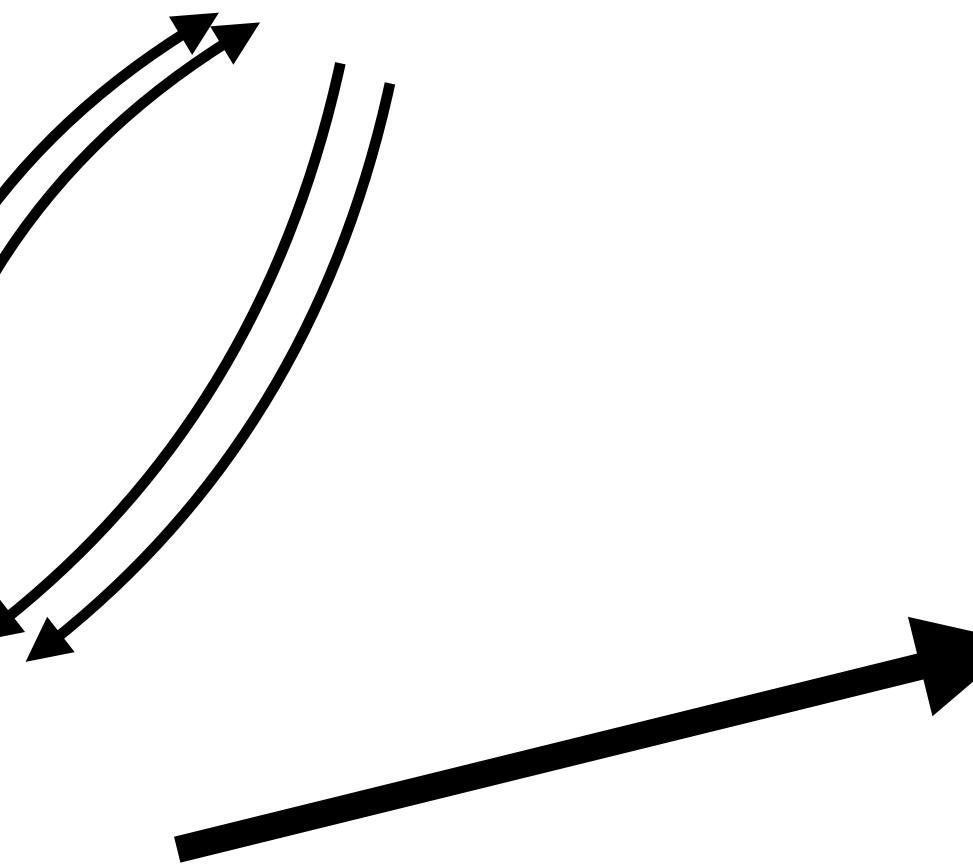
1. Collect data



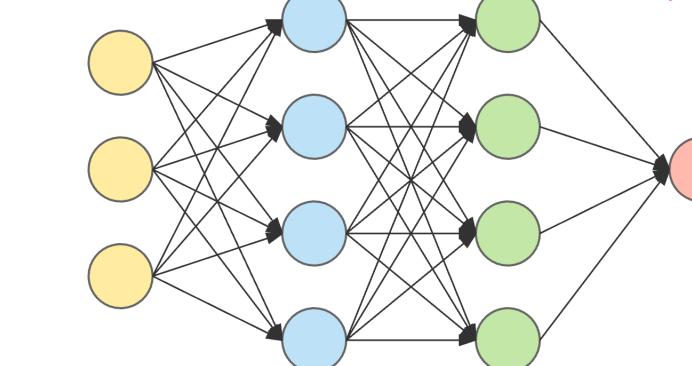
2. Split data



3. Train and tune model



4. Compute final test accuracy

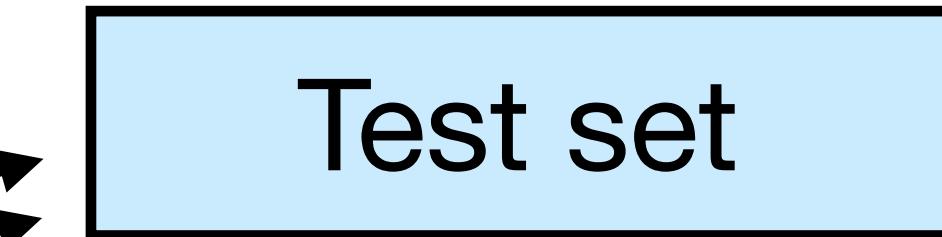
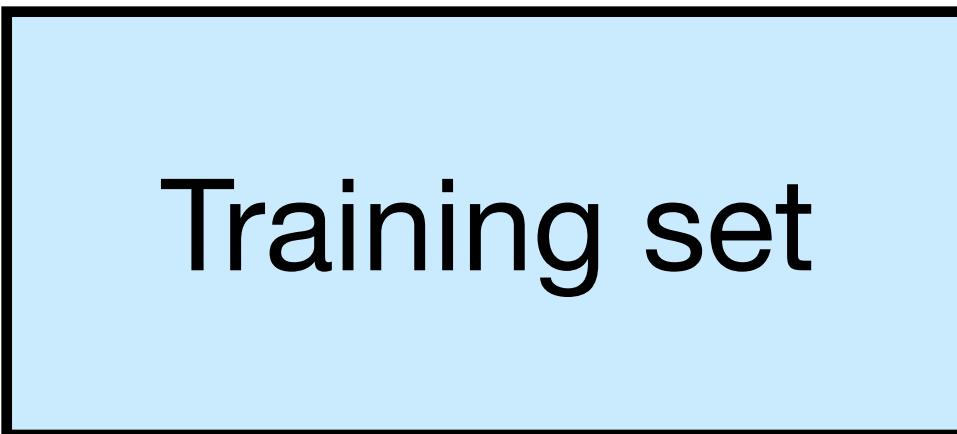
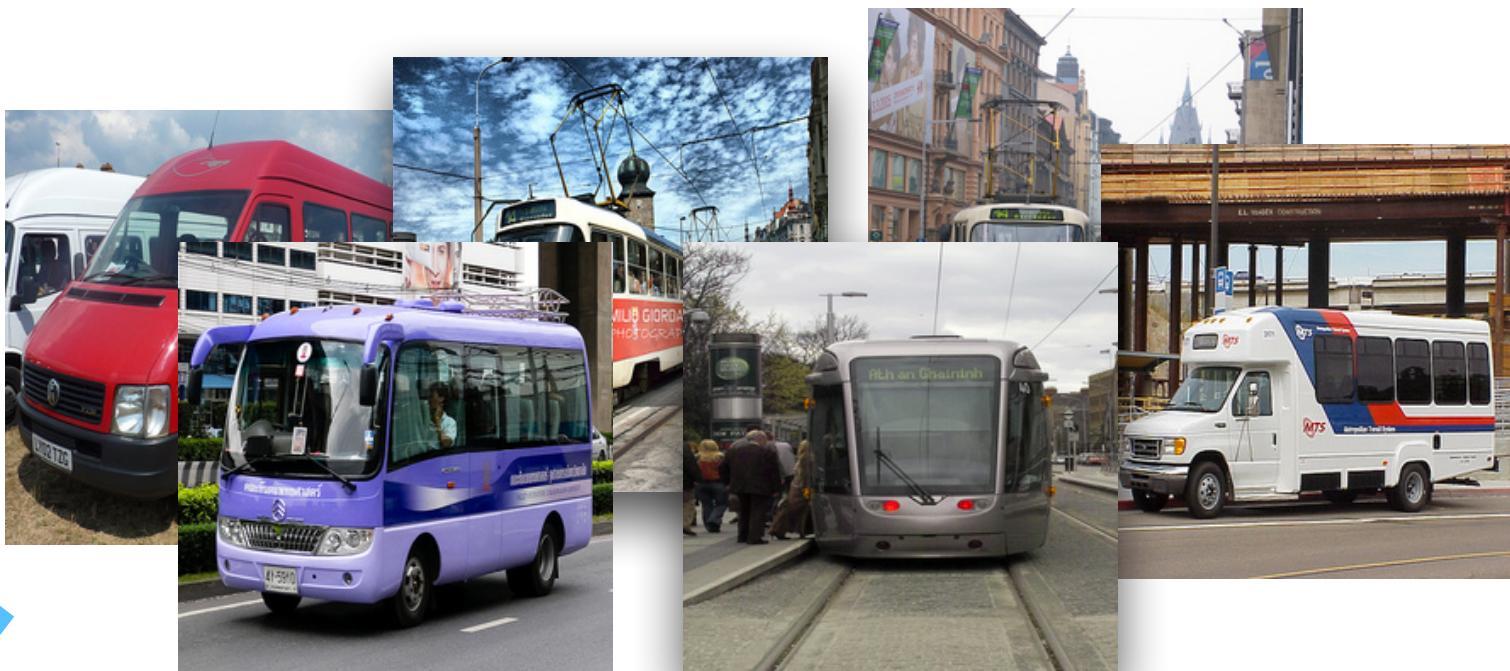


84%
→

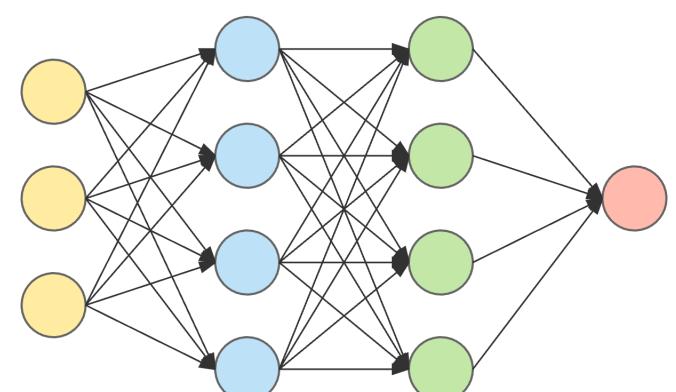


Typical ML Workflow

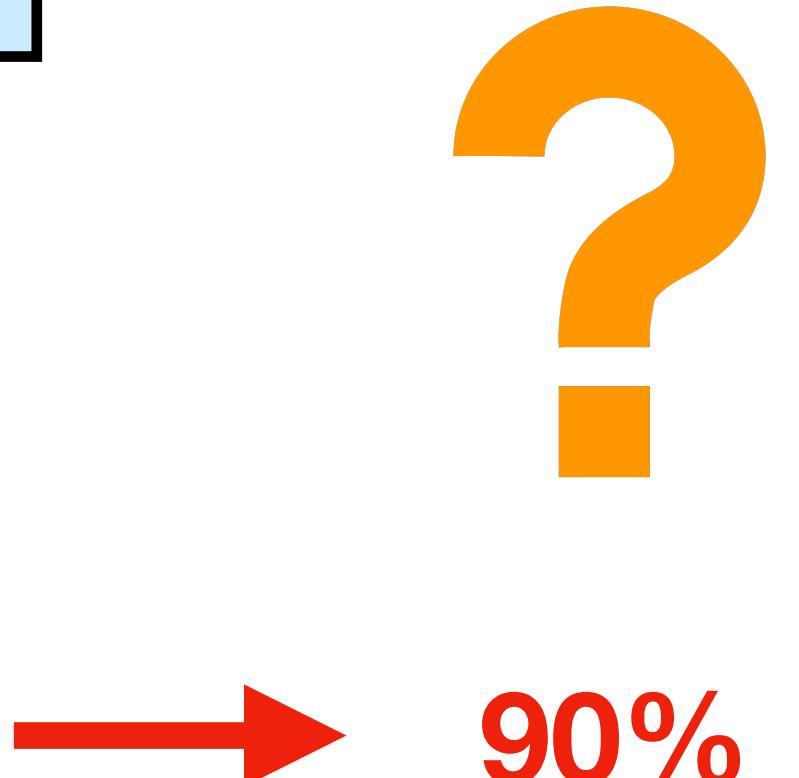
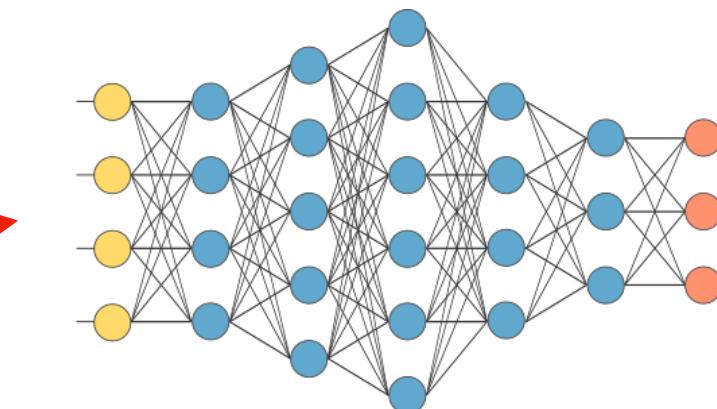
1. Download data
(fixed split)



2. Download model



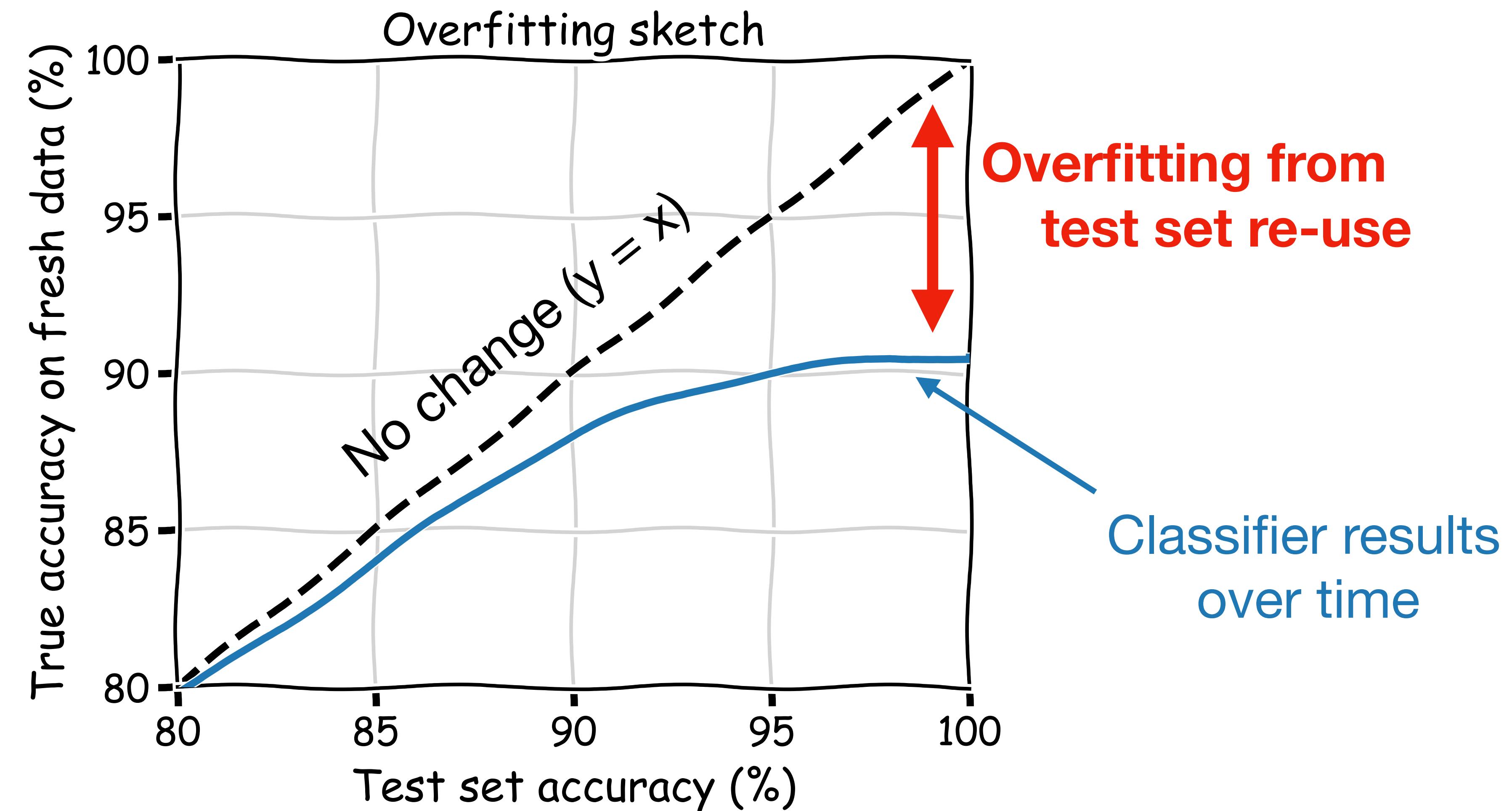
3. Train and tune model



4. Compute final test accuracy

Danger with Test Set Re-Use: Overfitting

Maybe we are just incrementally fitting to more and more random noise.

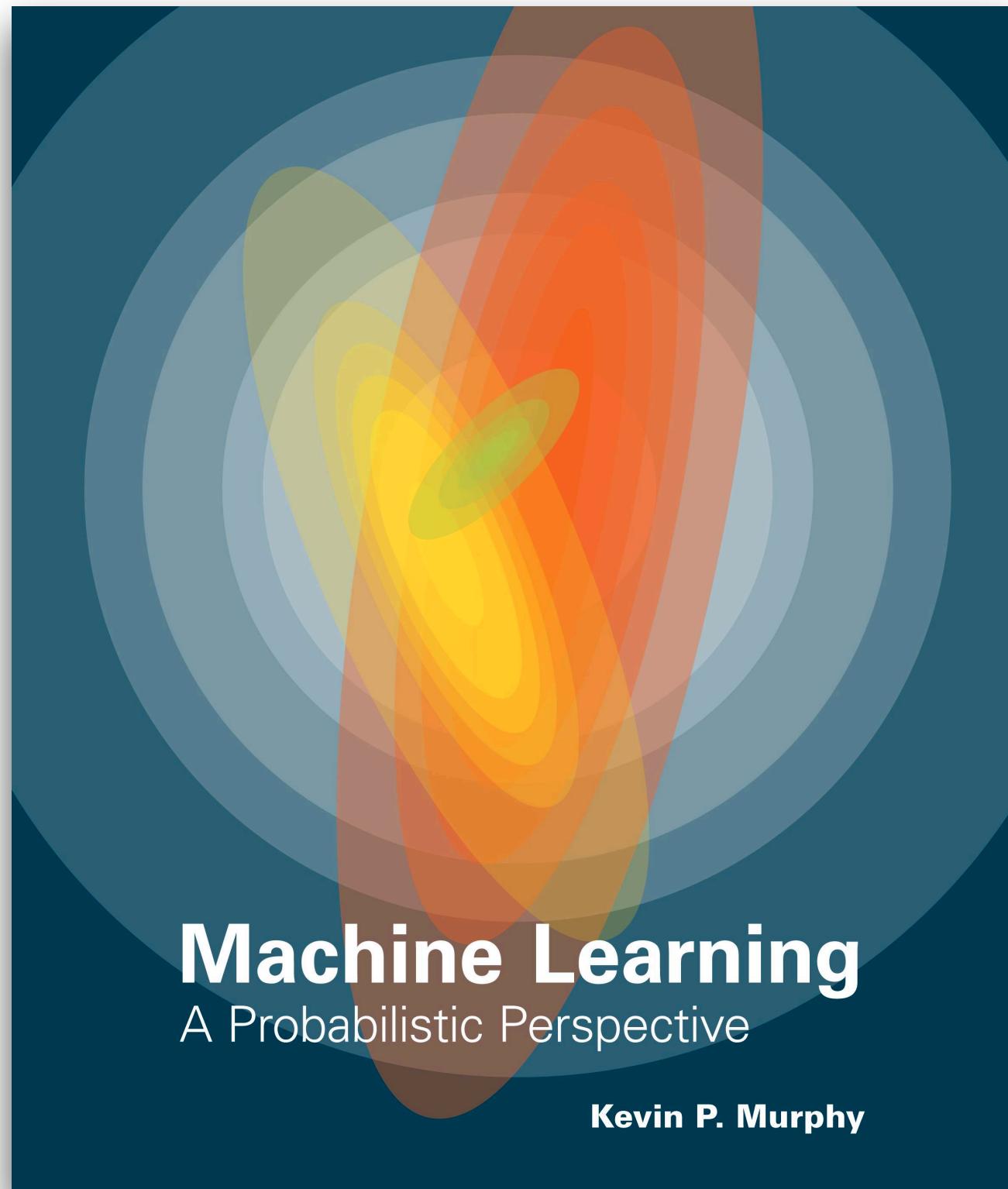


To be clear: We now know that there is no evidence of overfitting through test set re-use on many contemporary ML benchmarks (e.g., ImageNet)

However, the community was majorly confused about this.

We can learn from this story.

Textbooks



Chapter 1:

[...] we should not use [the test set] for model fitting or model selection, otherwise we will get an unrealistically optimistic estimate of performance of our method. This is one of the “golden rules” of machine learning research.

Slides from a Stanford NLP Class

Training models and pots of data

- The **train**, **tune**, **dev**, and **test** sets need to be completely distinct
- It is invalid to test on material you have trained on
 - You will get a falsely good performance. We usually overfit on train
- You need an independent tuning set
 - The hyperparameters won't be set right if tune is same as train
- If you keep running on the same evaluation set, you begin to overfit to that evaluation set
 - Effectively you are “training” on the evaluation set ... you are learning things that do and don’t work on that particular eval set and using the info
- To get a valid measure of system performance you need another untrained on, **independent** test set ... hence dev2 and final test

Research Papers, e.g., PASCAL VOC

*“Withholding the annotation of the test data until completion of the challenge played a significant part in **preventing over-fitting** of the parameters of classification or detection methods. In the VOC2005 challenge, test annotation was released and this led to some “**optimistic**” reported results, where a number of parameter settings had been run on the test set, and only the best reported. This danger emerges in any evaluation initiative where ground truth is publicly available.”*

+ several more mentions of “danger of overfitting” in the various PASCAL papers.

(Note: I searched for a while, there is not a single documented case of overfitting through test set re-use on PASCAL VOC. Alyosha helped with this.)

Context: a group had just released a new test set for MNIST

Invented CNNs, won a Turing award



Yann LeCun
@ylecun

MNIST reborn, restored and expanded.
Now with an extra 50,000 training samples.

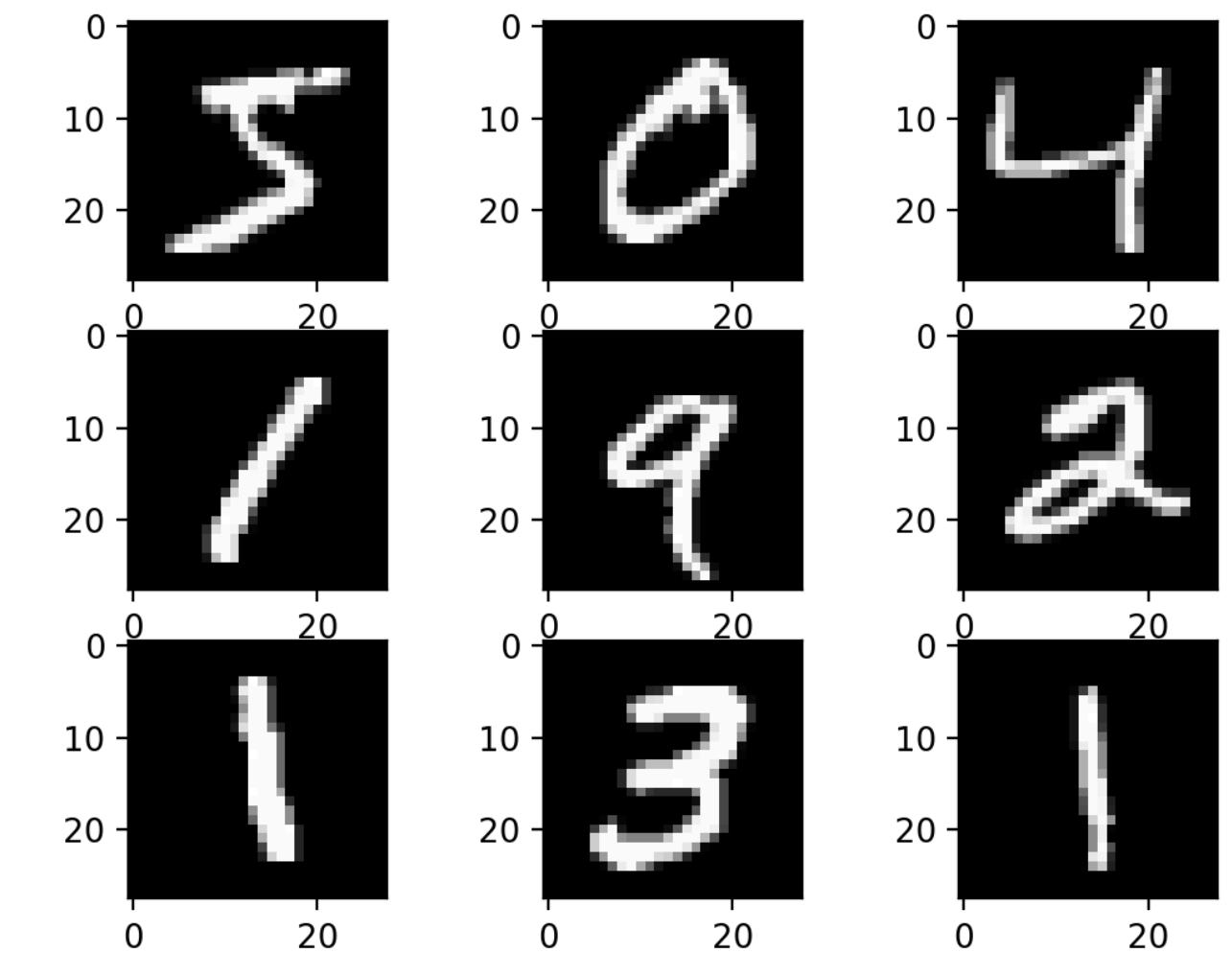
If you used the original MNIST test set more than a few times, chances are your models overfit the test set

Time to test them on those extra samples.

arxiv.org/abs/1905.10498

7:03 AM · May 29, 2019 · Facebook

699 Retweets 2K Likes



MNIST: digit classification

60k train, 10k test

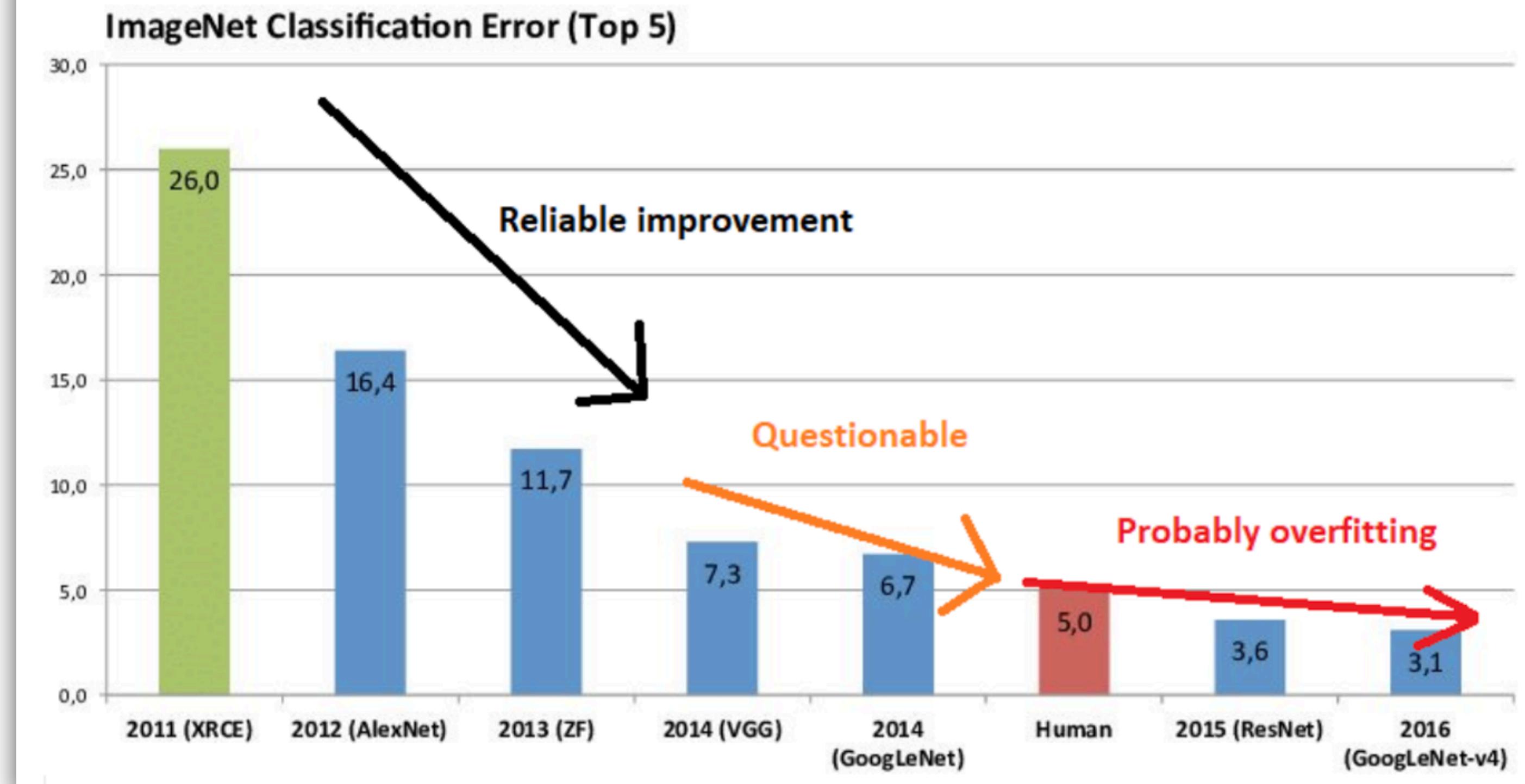
10 classes

Released in 1998

Oldest widely used dataset

Now considered “easy”

AI competitions don't produce useful models



I can't really estimate the numbers, but knowing what we know about multiple testing does anyone really believe the SOTA rush in the mid 2010s was anything but crowdsourced overfitting?

We tested for Overfitting

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht
UC Berkeley

Rebecca Roelofs
UC Berkeley

Ludwig Schmidt
UC Berkeley

Vaishaal Shankar
UC Berkeley

Abstract

and ImageNet datasets. Both benefit from re-use, raising the danger of overfitting. By examining original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

Outcome: There is actually no overfitting from test set re-use at all on ImageNet.

Meta-outcome: A lot of people were really confused about this.

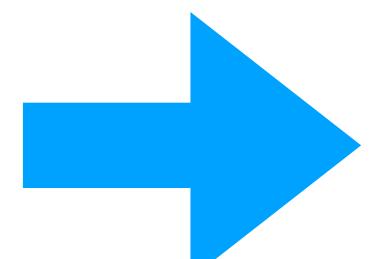
AlexNet Results

Model	Top-1	Top-5
<i>Sparse coding [2]</i>	47.1%	28.2%
<i>SIFT + FVs [24]</i>	45.7%	25.7%
CNN	37.5%	17.0%

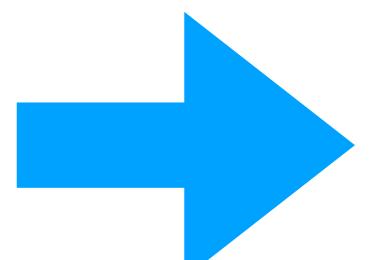
Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

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

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were “pre-trained” to classify the entire ImageNet 2011 Fall release. See Section 6 for details.



About 9 percentage points improvement over previous state-of-the art



88,000 citations, Turing award, transformation of computer science

???

An analogy to complexity theory

P vs NP is one of the core problems in theoretical computer science - why?

Quick complexity recap

A lot of important computational problems are in either P or NP.

P: set of problems solvable in polynomial time

(Sorting, shortest paths, linear programming, matrix multiplication, etc.)

NP: set of problems solvable in polynomial time on a non-deterministic Turing machine

(Satisfiability, traveling salesman problem, vertex cover, etc.)

NP-Completeness

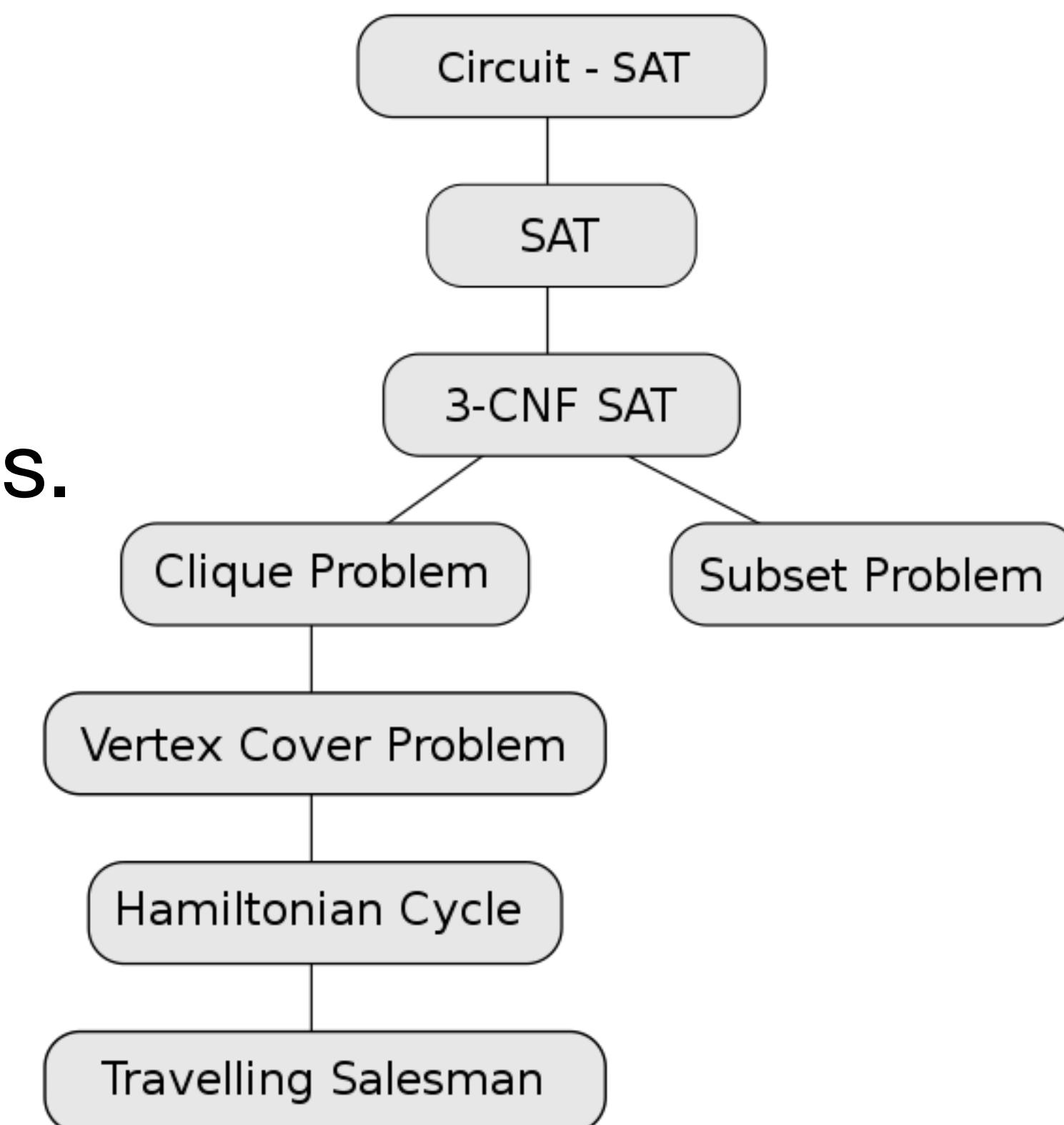
A key property of many important problems in NP: they are **NP-complete**.

- If you can solve a single NP-complete problem in polynomial time, you can solve **all** problems in NP in polynomial time.

This is formally established via **reductions** between problems.

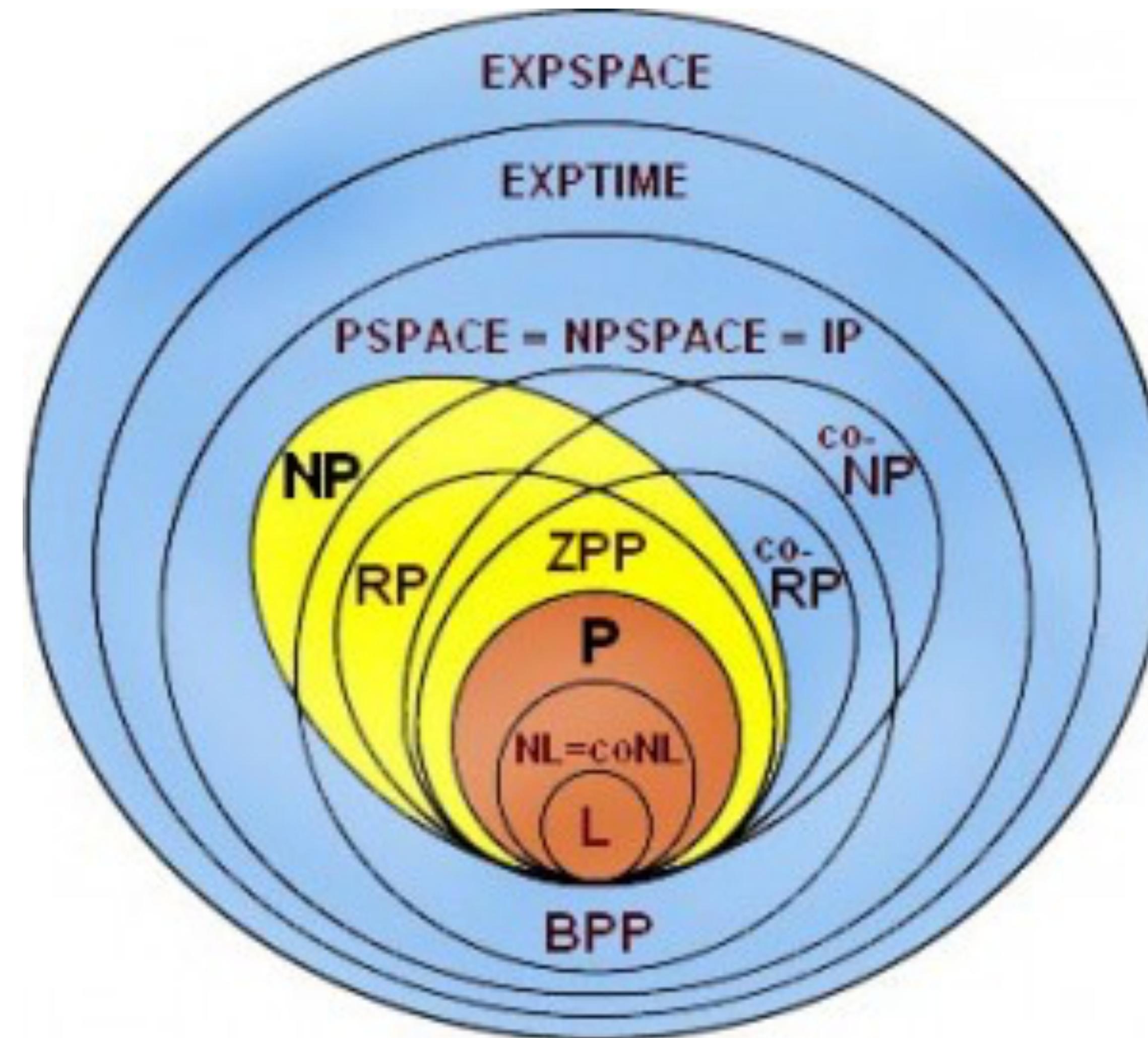
By now there are **thousands** of NP-complete problems.

- All of them have the same computational hardness, up to polynomial factors in the running time.

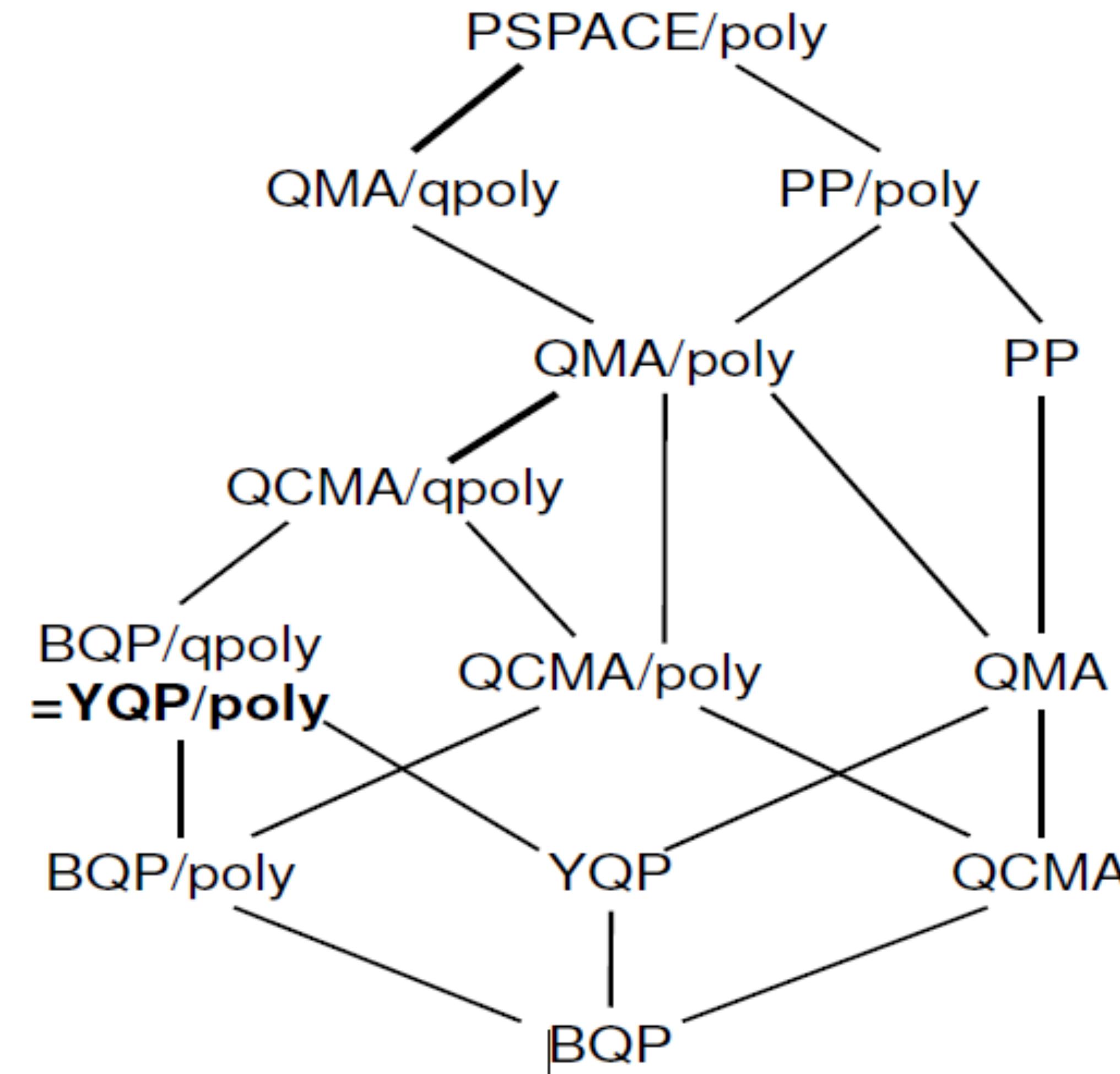


Big open question (P vs NP): is there a poly-time algorithm for any of these problems?

Complexity theory beyond P vs NP



Complexity theory beyond P vs NP



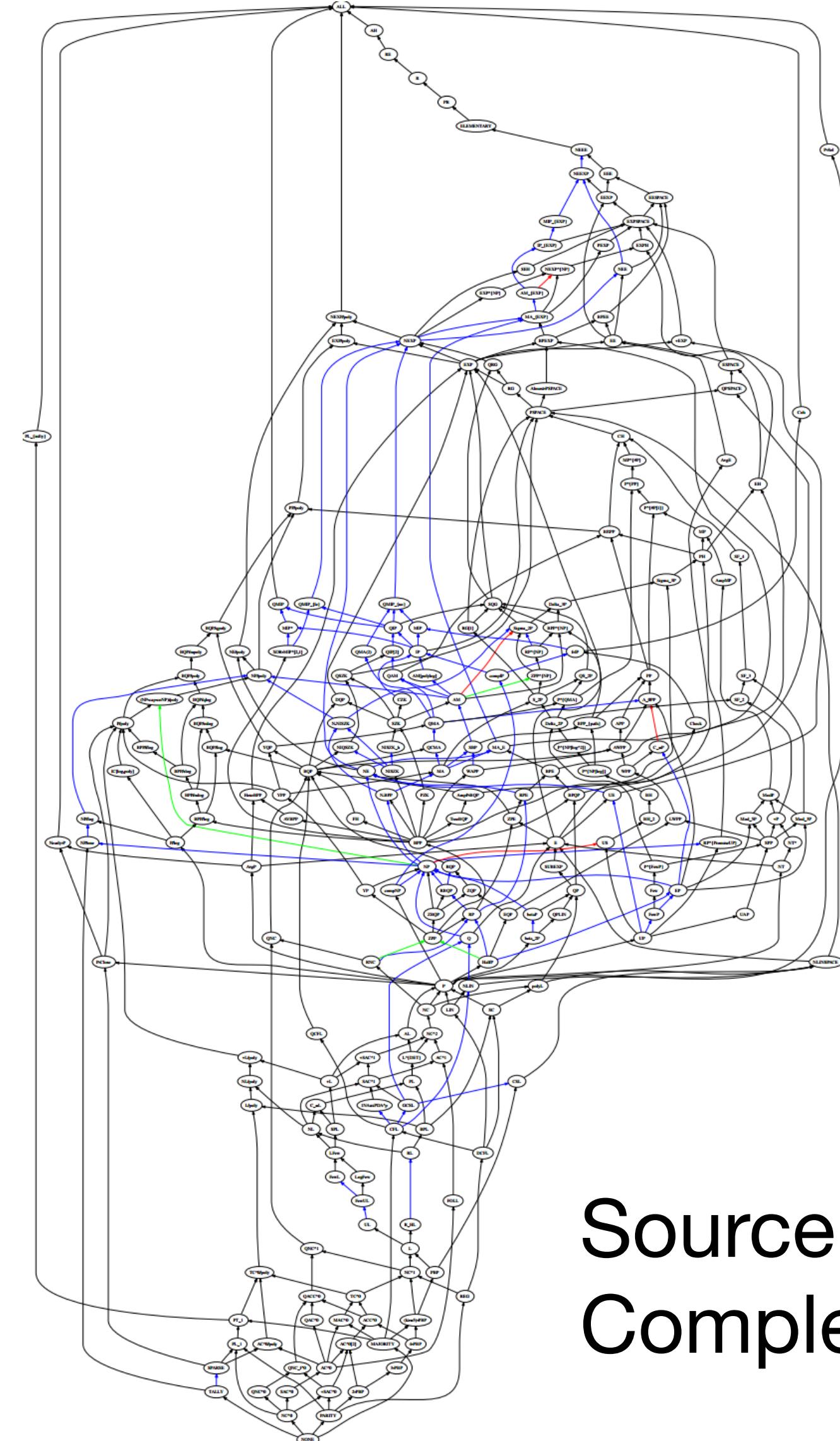
Complexity theory beyond P vs NP

Complexity theory has built a rich hierarchy of computational problems.

→ Many advantages, e.g., can quickly put a new computational problem in context.

Similar story in optimization: linear programs, quadratic programs, semi-definite programs, etc.

How does a similar problem hierarchy for data distributions and tasks in machine learning (across vision, NLP, etc.) look?



Source:
Complexity Zoo

Not all is well: failures of benchmarks

Different field: recommender systems

On the Difficulty of Evaluating Baselines

A Study on Recommender Systems

Steffen Rendle*

srendle@google.com

Li Zhang*

liqzhang@google.com

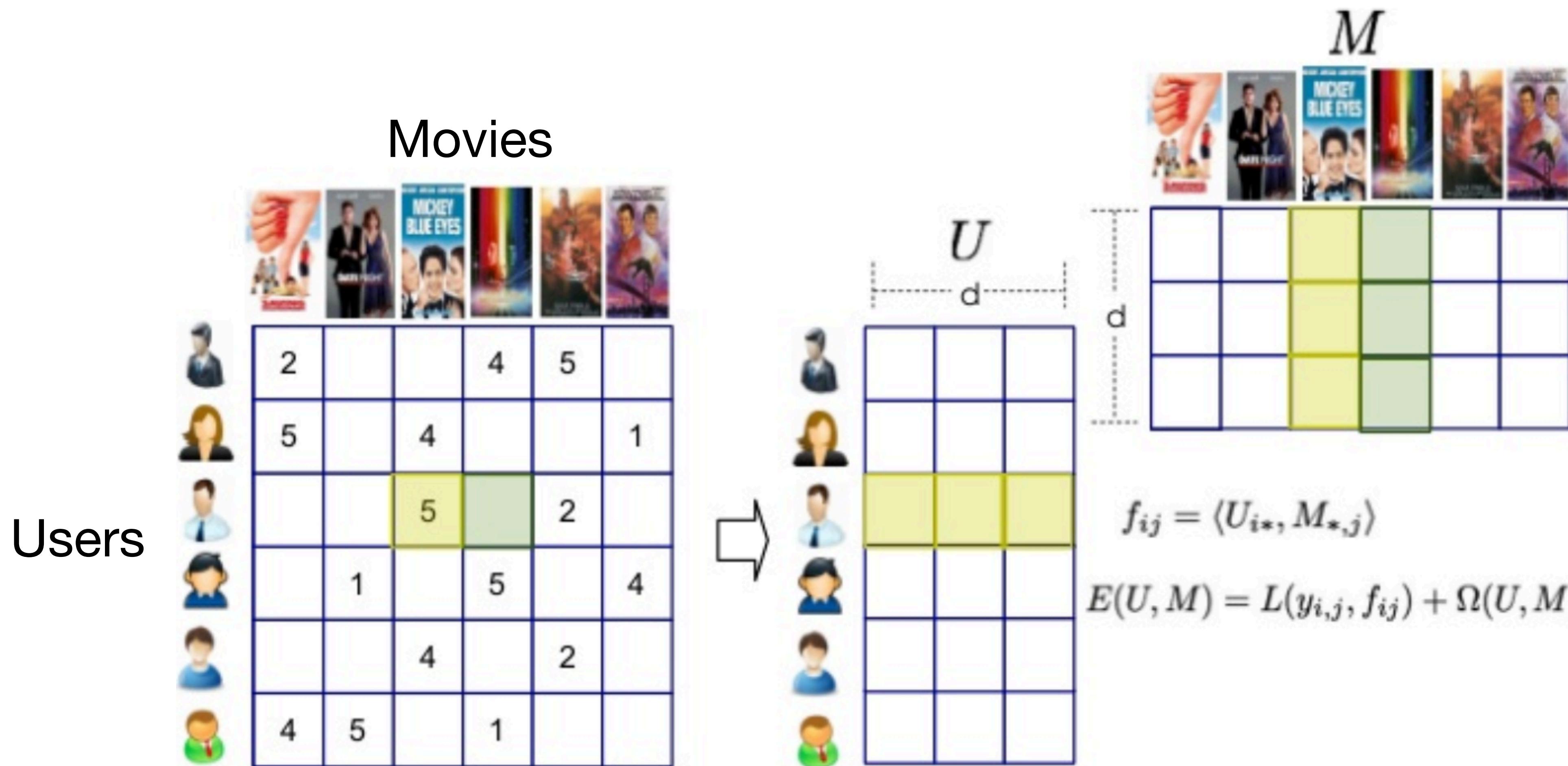
Yehuda Koren†

yehuda@google.com

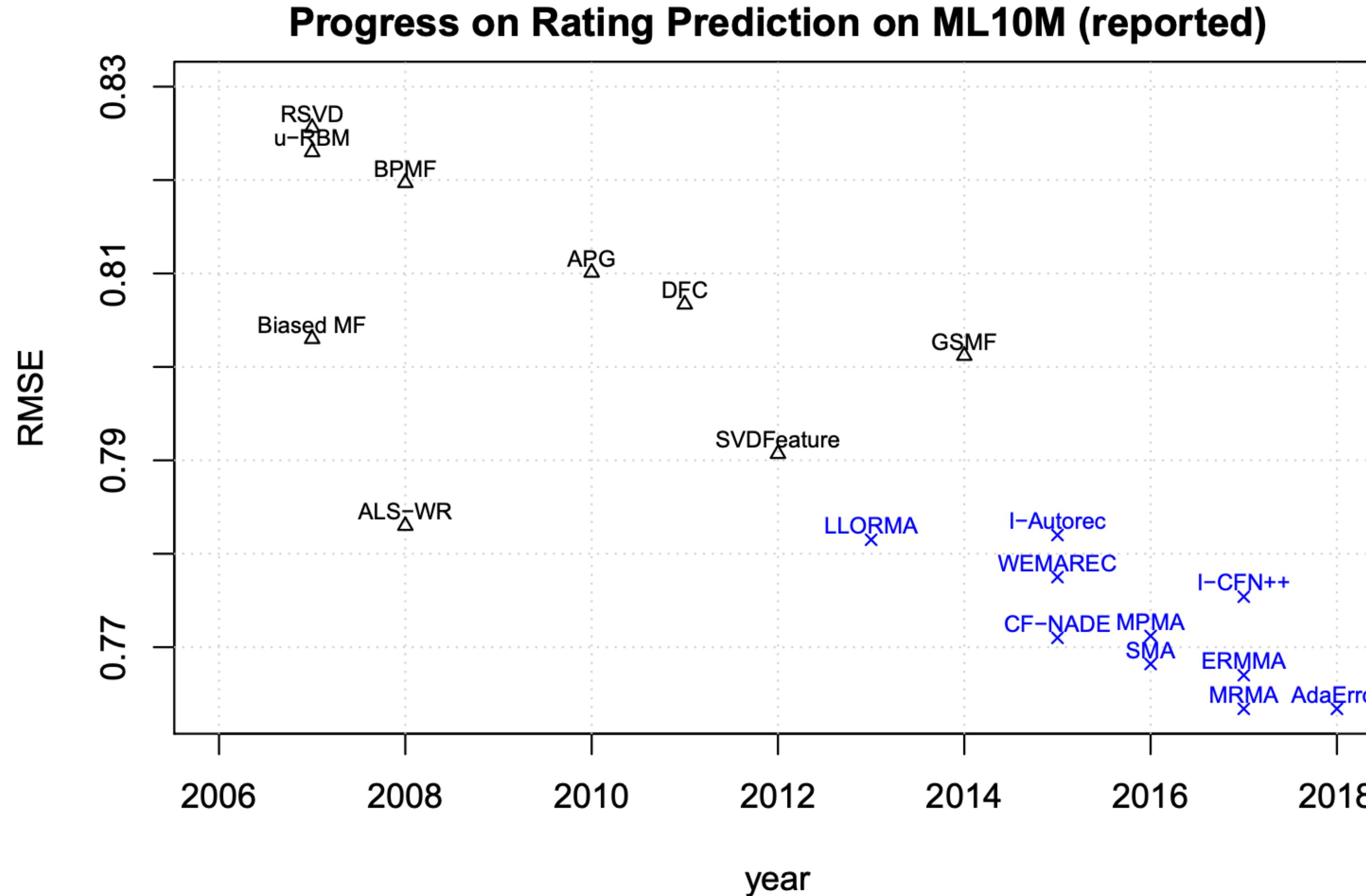
Abstract

Numerical evaluations with comparisons to baselines play a central role when judging research in recommender systems. In this paper, we show that running baselines properly is difficult. We demonstrate this issue on two extensively studied datasets. First, we show that results for baselines that have been used in numerous publications over the past five years for the MovieLens 10M benchmark are suboptimal. With a careful setup of a vanilla matrix factorization baseline, we are not only able to improve upon the reported results for this baseline but even outperform the reported results of any newly proposed method. Secondly, we recap the tremendous effort that was required by the community to obtain high quality results for simple methods on the Netflix Prize. Our results indicate that empirical findings in research papers are questionable unless they were obtained on standardized benchmarks where baselines have been tuned extensively by the research community.

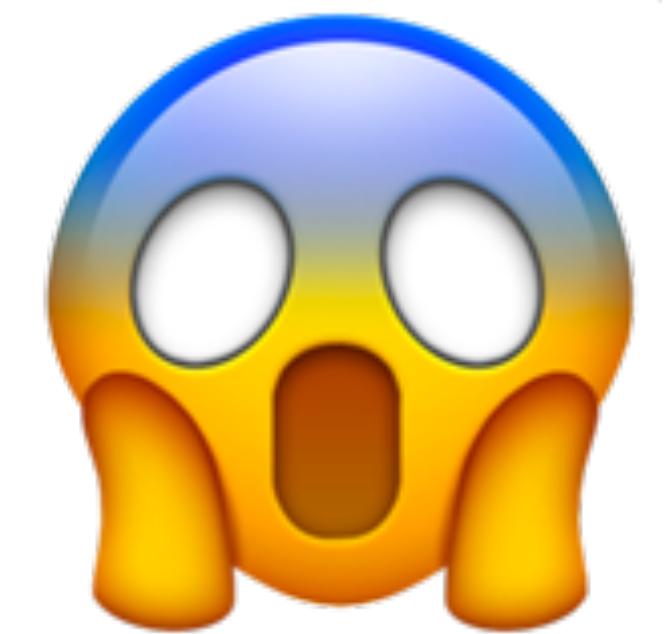
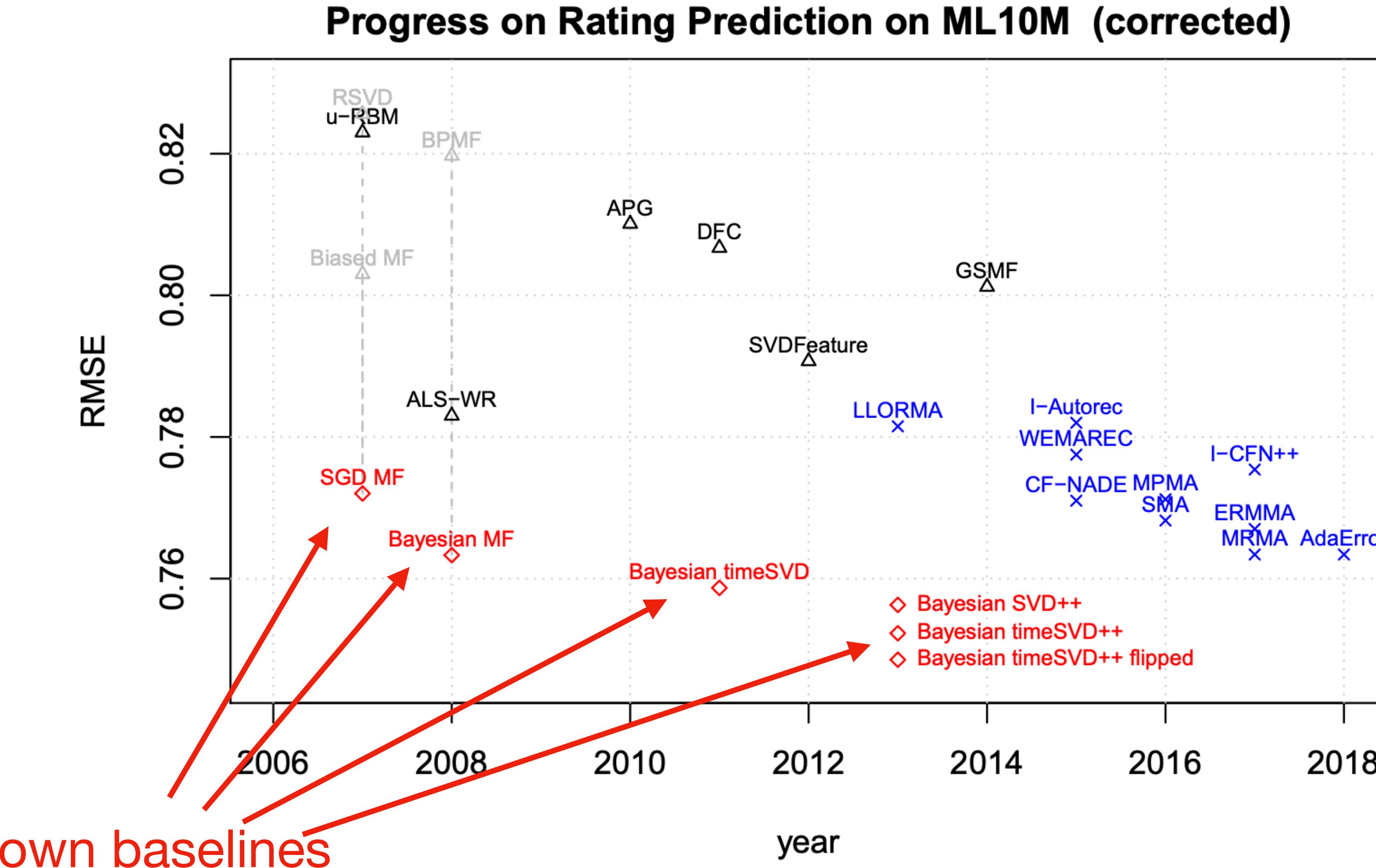
Recommender Systems & Matrix Factorization



“State of the Art”



Actual State of the Art

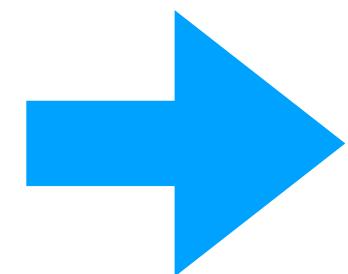


Danger with Empirical Evaluations

Difficulty of properly running baselines

Variations in tasks (exact dataset, evaluation metric, etc.)

Incentives around baselines



Standardized, competitive benchmarks address these points

Standard computer vision benchmarks (CIFAR-10, ImageNet, COCO) are so competitive that missed baselines seem unlikely by now.

What makes a good ML evaluation?

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts , Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | [Cite this article](#)

55k Accesses | **38** Citations | **1066** Altmetric | [Metrics](#)

Abstract

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

Questions

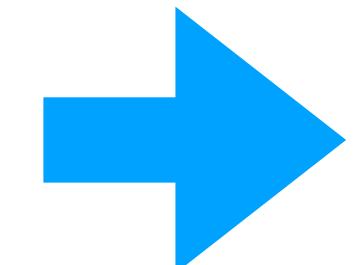
How reliable are performance measurements on ML benchmarks?

Why does progress on ImageNet lead to progress on many other tasks and datasets?

What tasks and datasets does ImageNet progress **not** help on?

How well do models with 90% top-1 accuracy on ImageNet really work?

What is the role of ImageNet in this story? What makes a good ML dataset?



What kind of answers am I looking for?

Why empirical foundations?

It's interesting! Lots of progress over the past years, still not well understood.

People expect more: reliability, fairness, security, etc.

Are the investments in ML going to the right problems?

Not all is well: many papers with failed evaluations, etc.

It leads to better methods!



Alec Radford

OpenAI

Verified email at openai.com

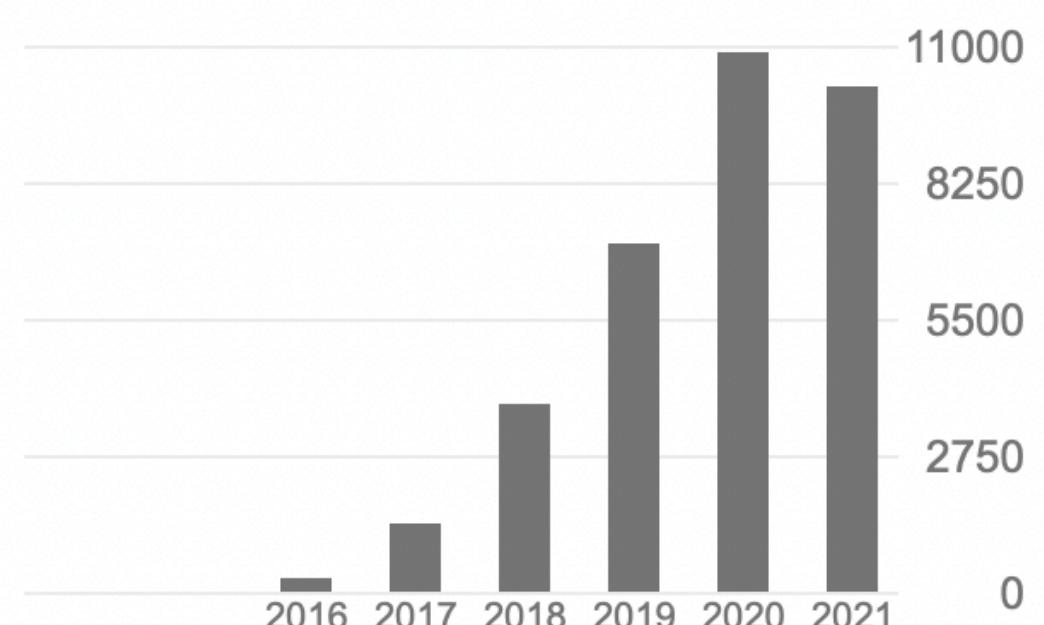
[Deep Learning](#) [Machine Learning](#)

FOLLOW

Cited by

[All](#) [Since 2016](#)

	All	Since 2016
Citations	34065	33801
h-index	25	25
i10-index	27	27



TITLE

CITED BY

YEAR

[Unsupervised representation learning with deep convolutional generative adversarial networks](#)

10311

2015

A Radford, L Metz, S Chintala
arXiv preprint arXiv:1511.06434

[Improved techniques for training gans](#)

5659

2016

T Salimans, I Goodfellow, W Zaremba, V Cheung, A Radford, X Chen
Advances in neural information processing systems 29, 2234-2242

[Proximal policy optimization algorithms](#)

5371

2017

J Schulman, F Wolski, P Dhariwal, A Radford, O Klimov
arXiv preprint arXiv:1707.06347

[Language Models are Unsupervised Multitask Learners](#)

4463 *

2019

A Radford, J Wu, R Child, D Luan, D Amodei, I Sutskever
Technical report, OpenAi

[Improving language understanding by generative pre-training](#)

2938 *

2018

A Radford, K Narasimhan, T Salimans, I Sutskever

[Language models are few-shot learners](#)

1847

2020

TB Brown, B Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, ...
arXiv preprint arXiv:2005.14165

[Openai baselines](#)

710

2017

P Dhariwal, C Hesse, O Klimov, A Nichol, M Plappert, A Radford, ...

Co-authors

- Ilya Sutskever
Co-Founder and Chief Scientist ... [>](#)
- Soumith Chintala
Facebook AI Research [>](#)
- Luke Metz
Google Brain [>](#)
- Ian Goodfellow [>](#)
- Wojciech Zaremba
Co-Founder of OpenAI, Head of ... [>](#)



Kaiming He

Research Scientist, Facebook AI Research (FAIR)

Verified email at fb.com - [Homepage](#)

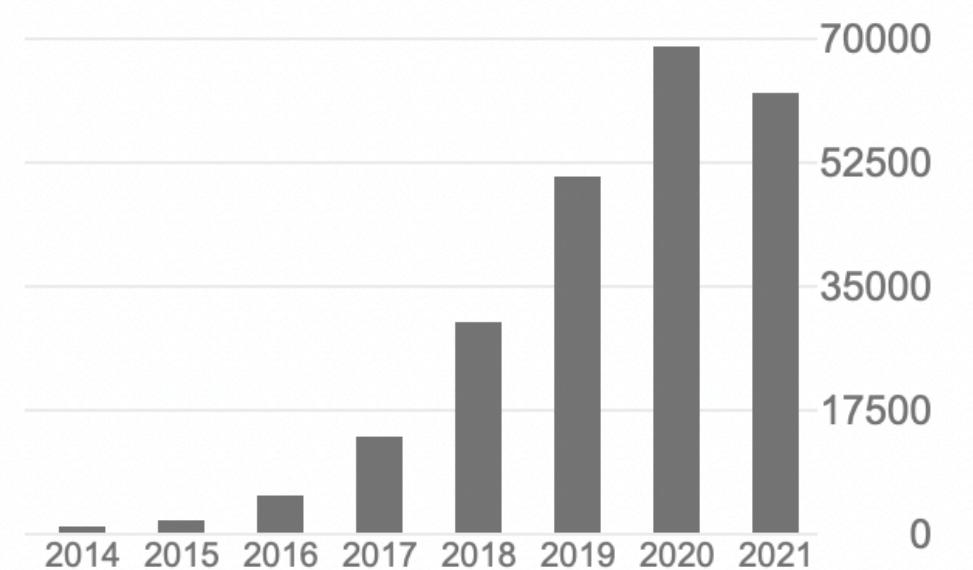
Computer Vision Machine Learning

FOLLOWING

Cited by

All Since 2016

Citations	238421	231788
h-index	58	57
i10-index	66	66



TITLE

CITED BY

YEAR

[Deep Residual Learning for Image Recognition](#)

K He, X Zhang, S Ren, J Sun

Computer Vision and Pattern Recognition (CVPR), 2016

91640 2016

[Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks](#)

S Ren, K He, R Girshick, J Sun

Neural Information Processing Systems (NIPS), 2015

34429 2015

[Mask R-CNN](#)

K He, G Gkioxari, P Dollár, R Girshick

International Conference on Computer Vision (ICCV), 2017

14490 2017

[Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification](#)

K He, X Zhang, S Ren, J Sun

International Conference on Computer Vision (ICCV), 2015

13284 2015

Public access

[VIEW ALL](#)

0 articles

[10 articles](#)

not available

[available](#)

Based on funding mandates

[Focal Loss for Dense Object Detection](#)

TY Lin, P Goyal, R Girshick, K He, P Dollár

International Conference on Computer Vision (ICCV), 2017

9302 2017

[Feature Pyramid Networks for Object Detection](#)

TY Lin, P Dollár, R Girshick, K He, B Hariharan, S Belongie
Computer Vision and Pattern Recognition (CVPR), 2017

8881 2017

Co-authors

[VIEW ALL](#)



Jian Sun
Chief Scientist/Managing Directo... >

[Learning a Deep Convolutional Network for Image Super-Resolution](#)

C Dong, CC Loy, K He, X Tang

European Conference on Computer Vision (ECCV), 2014

7813 * 2014



Ross Girshick

Research Scientist, Facebook AI Research (FAIR)

Verified email at eecs.berkeley.edu - [Homepage](#)

computer vision machine learning

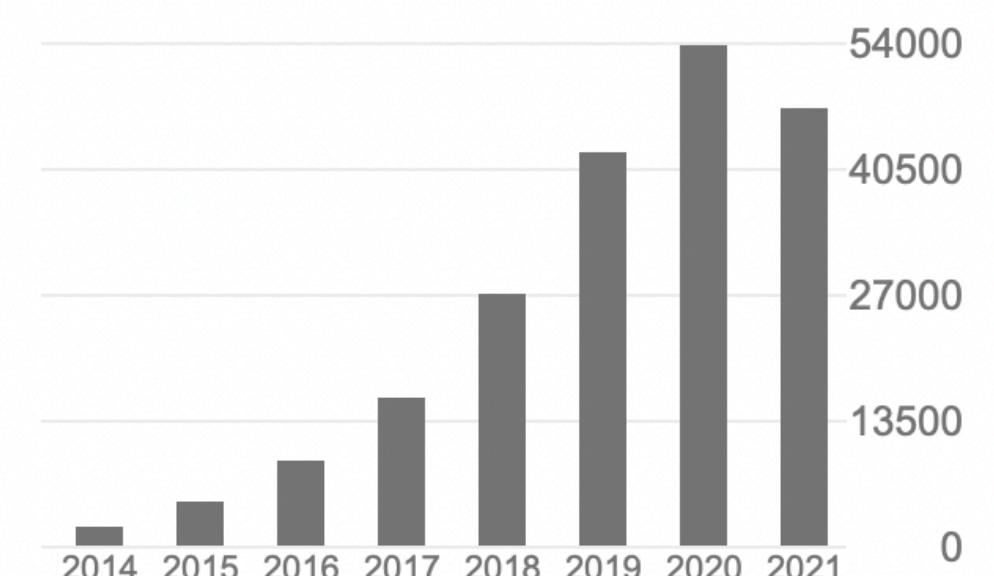
FOLLOW

Cited by

[VIEW ALL](#)

All Since 2016

	Citations	209043	196196
	h-index	73	72
	i10-index	110	107



TITLE

CITED BY

YEAR

[Faster R-CNN: Towards real-time object detection with region proposal networks](#)

34254

2015

S Ren, K He, R Girshick, J Sun

Advances in neural information processing systems, 91-99

[Rich feature hierarchies for accurate object detection and semantic segmentation](#)

21344

2014

R Girshick, J Donahue, T Darrell, J Malik

Proceedings of the IEEE conference on computer vision and pattern ...

[You only look once: Unified, real-time object detection](#)

18571

2016

J Redmon, S Divvala, R Girshick, A Farhadi

Proceedings of the IEEE conference on computer vision and pattern ...

[Microsoft coco: Common objects in context](#)

18536

2014

TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, ...

European conference on computer vision, 740-755

[Fast R-CNN](#)

16576

2015

R Girshick

Proceedings of the IEEE International Conference on Computer Vision, 1440-1448

[Caffe: Convolutional architecture for fast feature embedding](#)

15792

2014

Y Jia, E Shelhamer, J Donahue, S Karayev, J Long, R Girshick, ...

Proceedings of the 22nd ACM international conference on Multimedia, 675-678

[Mask R-CNN](#)

14051

2017

K He, G Gkioxari, P Dollár, R Girshick

arXiv preprint arXiv:1703.06870

Public access

[VIEW ALL](#)

0 articles

13 articles

not available

available

Based on funding mandates

Co-authors

[VIEW ALL](#)



Kaiming He

Research Scientist, Facebook AI...



Caveats

This class takes a technical perspective on ML.

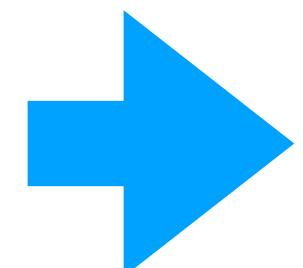
A narrow technical focus can obscure ethical questions.



But: research on ethical questions in machine learning needs solid foundations, too.

A course on empirical foundations of ML is largely new.

(It still rests on decades of work in other fields - research validity is not new.)



We're going to figure some things out as we go through the quarter.

1. Logistics

2. Background & motivation

3. Course outline

Course Outline

Main parts of the class

1. Fundamentals: applied stats, causality, a bit of philosophy of science (5 lectures)
2. Paper discussions: both “classical” and recent papers (5 lectures)
3. Guest speakers (Alec Radford , Nicholas Carlini, and more) (3 lectures)
4. Student project presentations: initial overview and final presentations (3 lectures)
5. Practical tooling for empirical ML (favorite Python packages, etc.) (1 lecture)

Grading & project

Grading: 20% participation in class discussions, 80% research project.

Project

Theme: broadly around datasets, evaluation, robustness

Can be research you are already doing

Team size 1 - 3

Proposals due at the beginning of the 4th lecture (October 12)

Next lecture: some inspiration

Thanks!

Questions?