# CSE 493 / 599 Advanced Machine Learning

University of Washington, Spring 2023

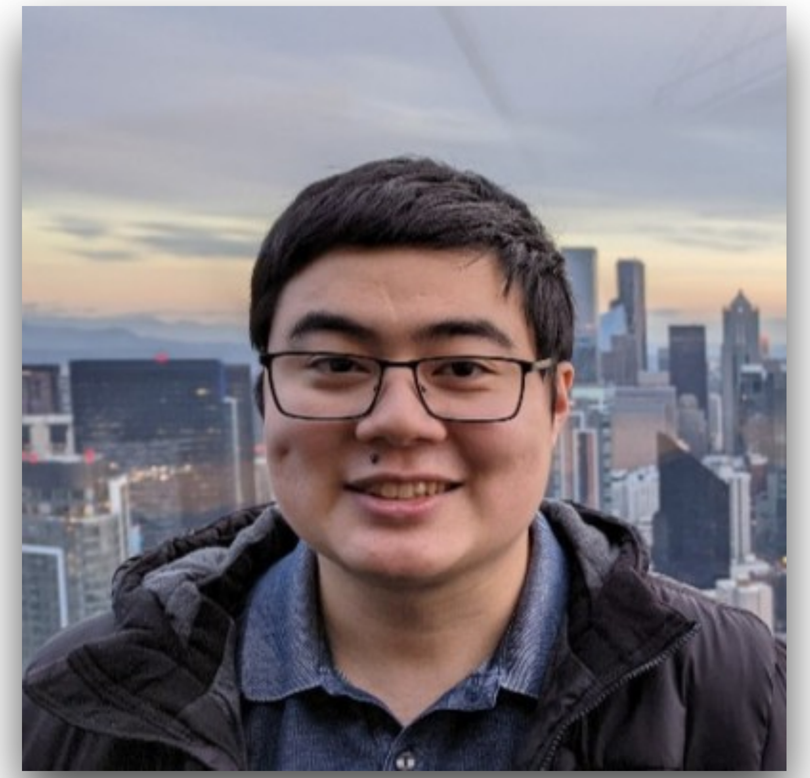
Normally class starts at 10 am, today 10:05 so people can find the room.

# Welcome!

# Introduction

Ludwig Schmidt

Tim Dettmers



Jonathan Hayase

Gabriel Ilharco

Mitchell Wortsman

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/advancedml-sp23/

- Announcements
- Material (schedule, slides, lecture notes, etc.)
- Links (**Ed**)

**Course staff mailing list:** multi_cse493s_sp23@uw.edu

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

Likely **no recordings** (room not set up for lecture recordings).

**Ask questions any time!**

# Grading

Exact details still TBD, most likely:

**Two homeworks**

- One for the theory-oriented part of the class
  (Released end of next week, due three weeks later)
- One for the experiment-oriented part


**Course project**, for instance:

- Re-implementing a paper
- New idea on top of an existing code base
- Summarizing a line of theoretical work
- Original research

1. Logistics

2. Background & motivation

3. Course outline

# Goal for the class

**Learning outcome:** foundations for graduate research in machine learning.

Advanced ML is going to fill a **gap** in our ML classes at UW CSE:

**Introduction to machine learning (446 / 546)**

- Overview of existing methods and how to apply them
- Less emphasis on theory or developing new state-of-the-art methods

**Graduate classes** (deep learning, reinforcement learning, interactive learning, etc.)

- More specialized to certain research directions
- Often assume basic background in learning theory (or would like to)

# What are the foundations for ML research?

**Traditionally,** ML foundations emphasize the **mathematical theory** in the field.

**High-level approach:**

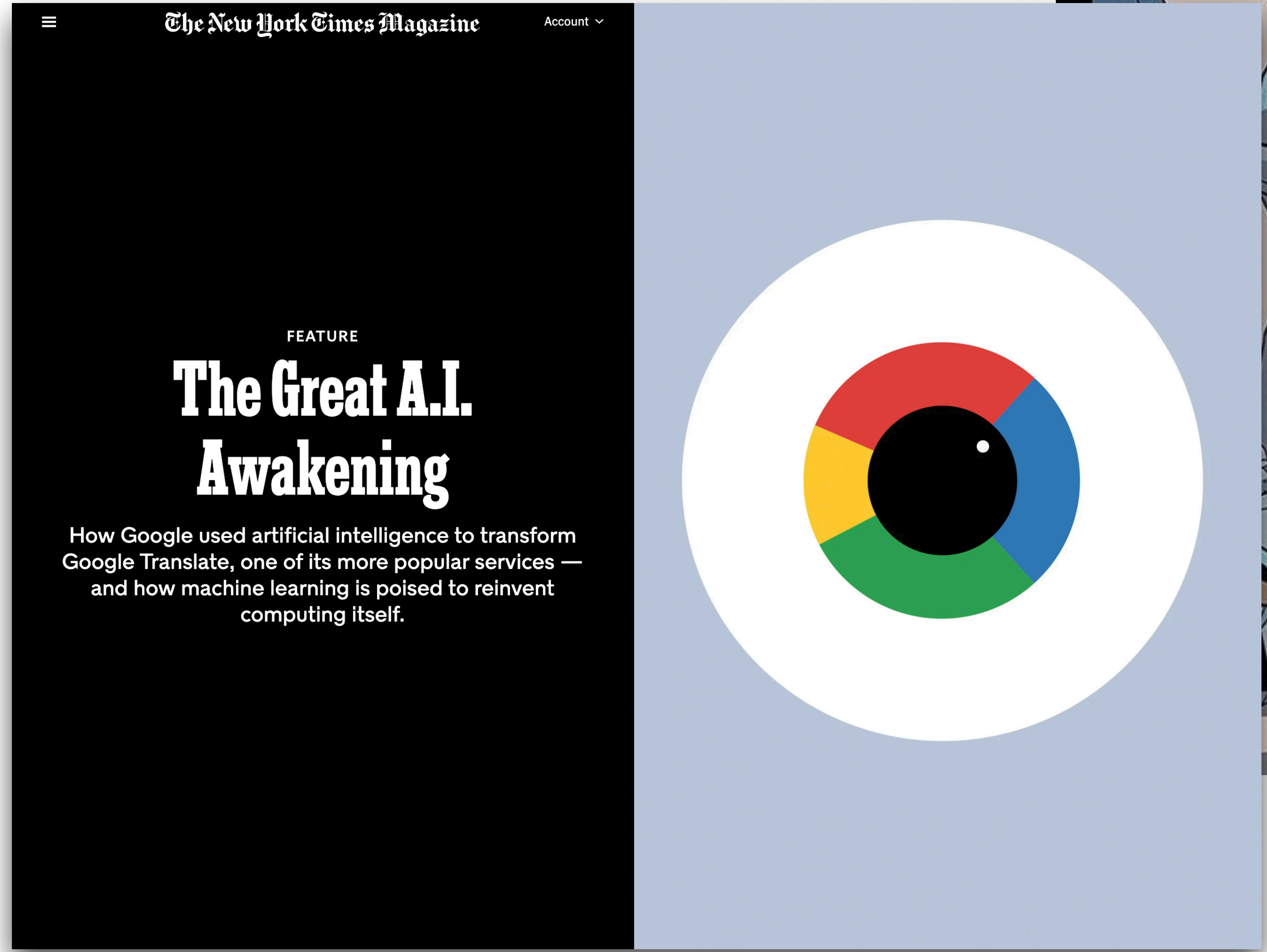
1. Formally define a learning problem
2. Propose algorithms to solve the problem
3. Analyze their running time and sample complexity

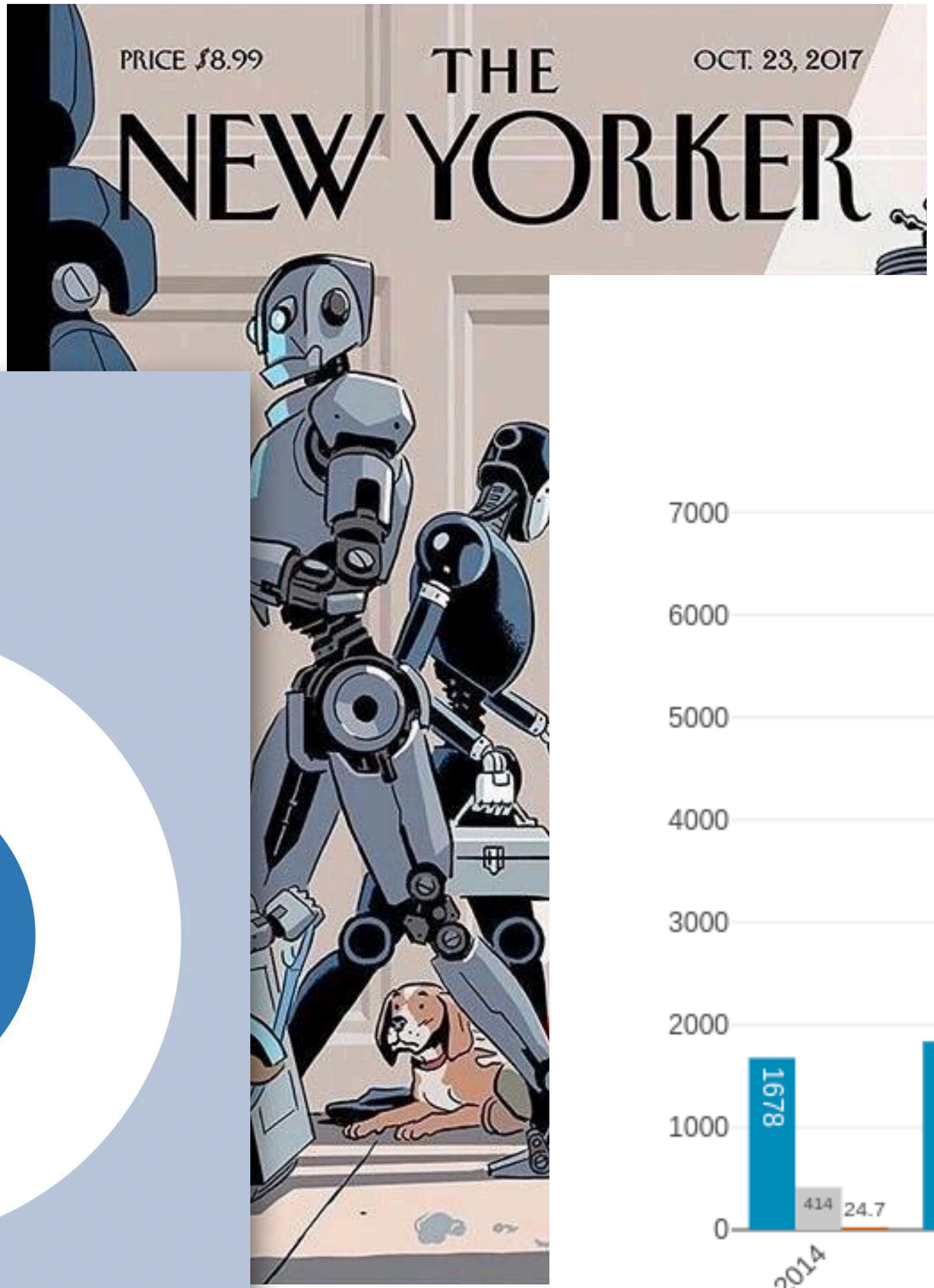
Theoretical development based on deductive reasoning (theorems & proofs), similar to a classical algorithms class (sorting, shortest path, etc.)

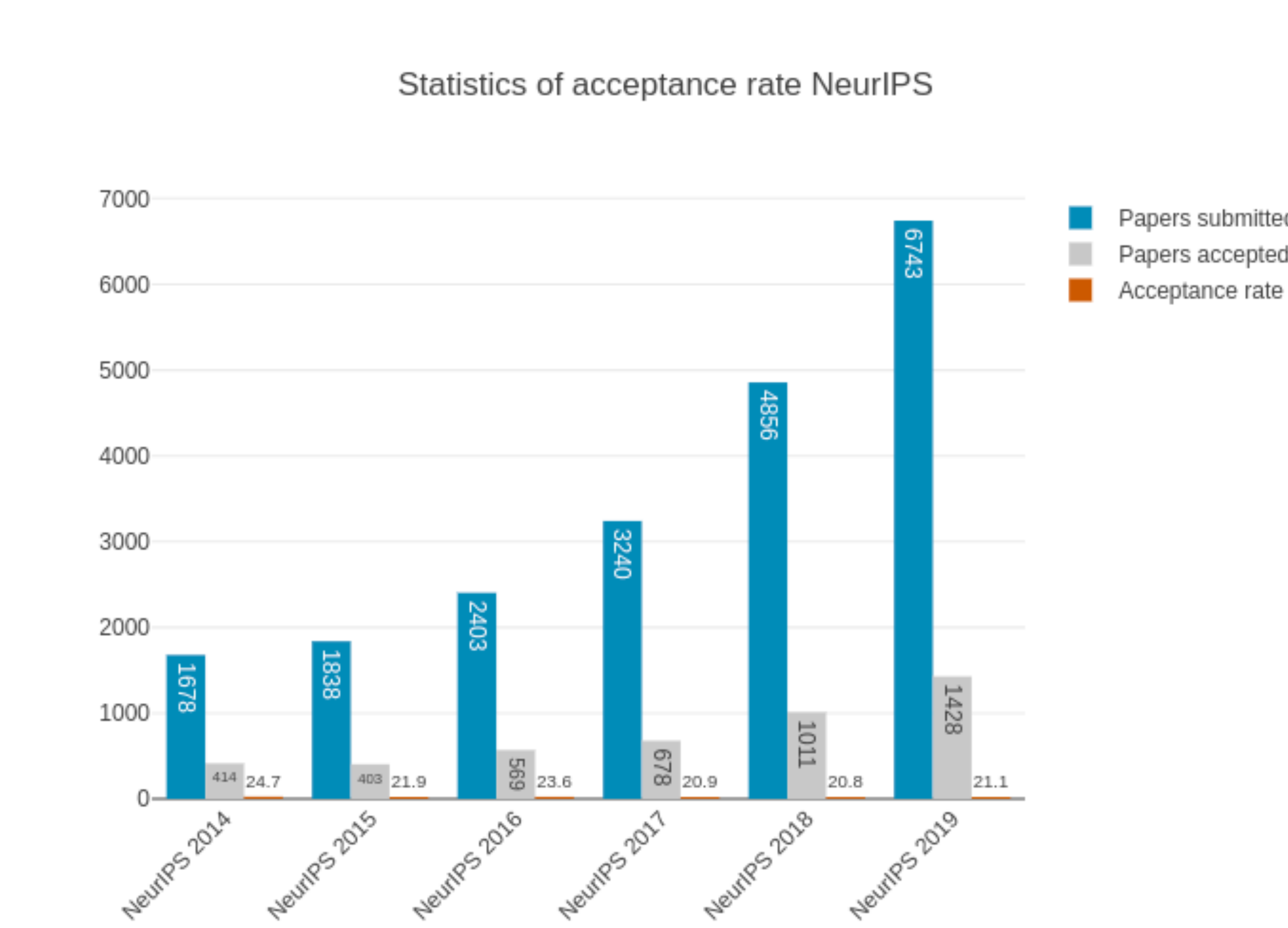
**Over the past 10 years,** a lot of progress on the AI-side of ML has been driven by **experiments,** usually with little or no mathematical theory.

We will cover **both** theoretical and empirical foundations.

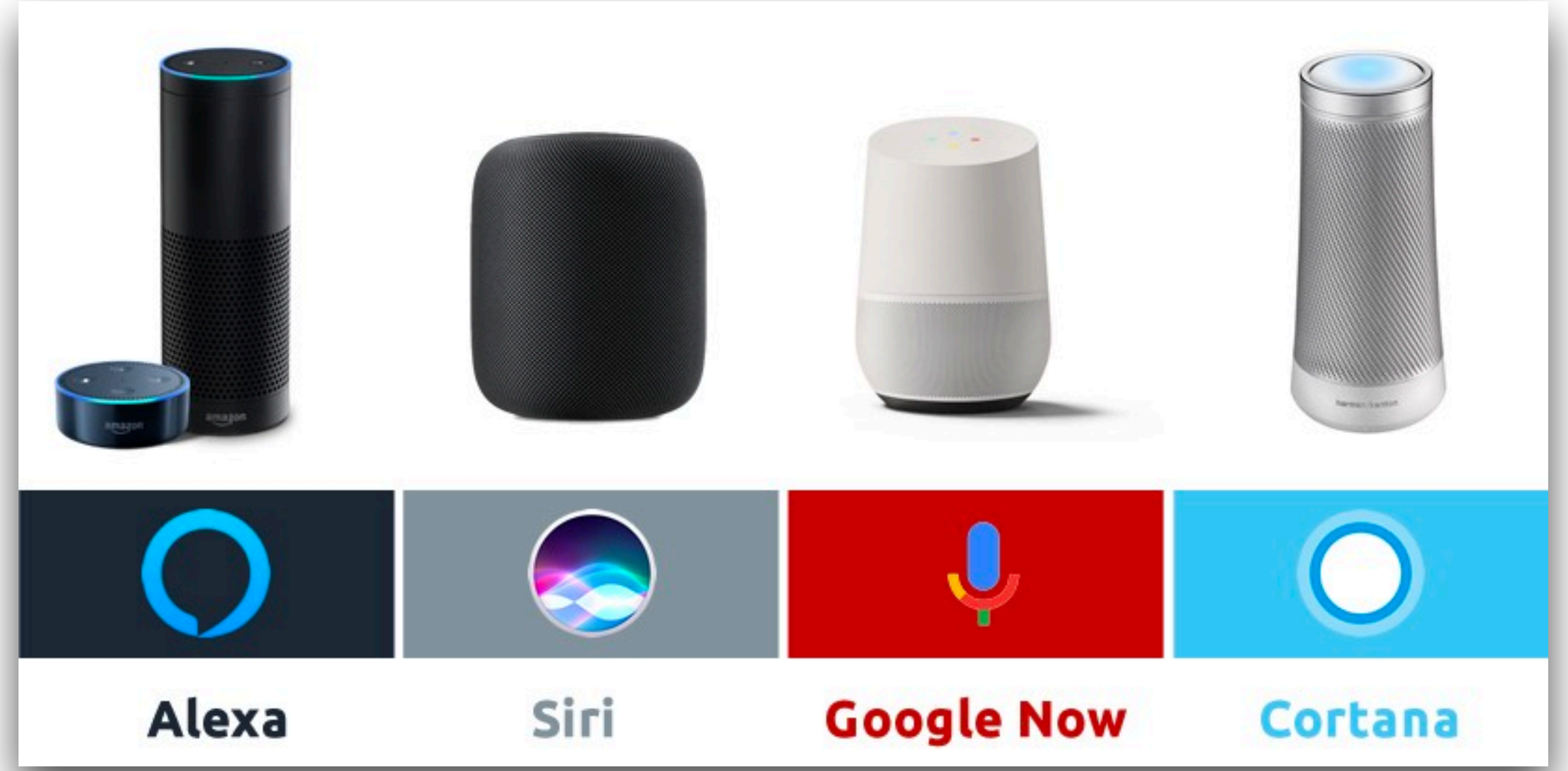
# Explosive Growth in ML



Self-driving cars




Voice assistants

Content moderation

Chatbots

# 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]

# ImageNet

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

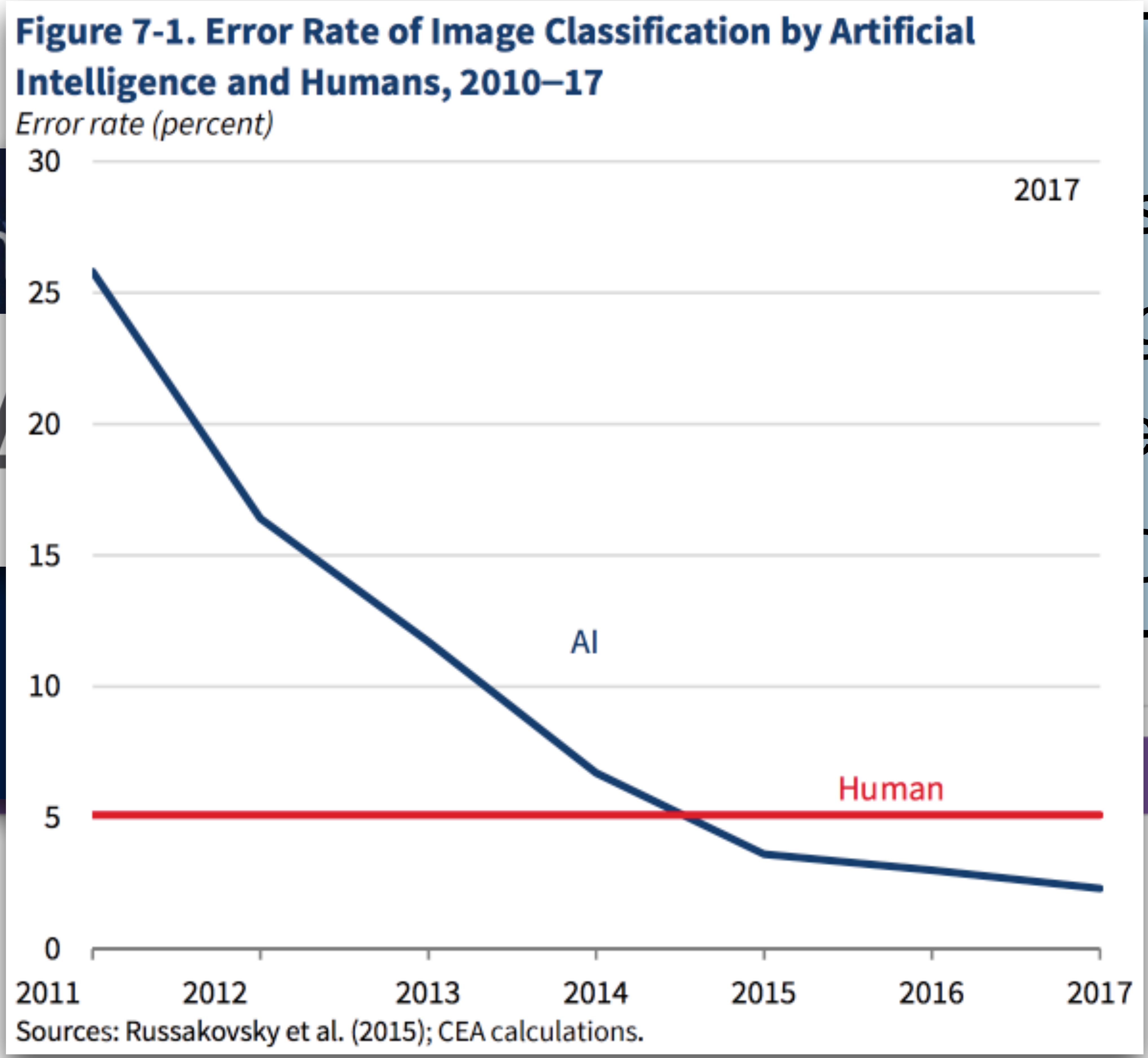
Golden retriever


Great white shark

Minibus

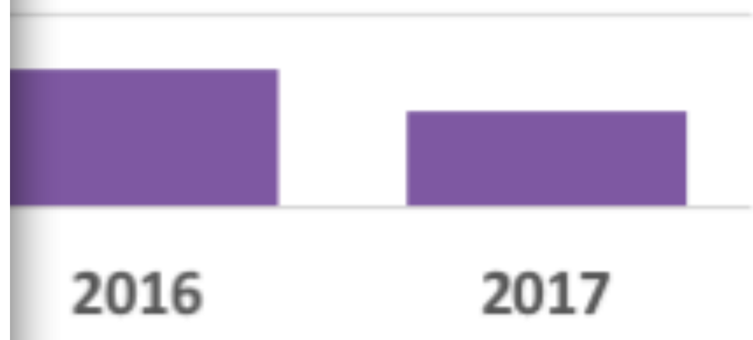
# ImageNet

st decade:


*that the following*
***st impactful paper in***
*g and computer vision*
*ears.”*

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

- dog, domestic dog, Canis familiaris
  - canine, canid
    - carnivore
      - placental, placental mammal, eutherian, eutherian mammal
        - mammal
          - vertebrate, craniate
            - chordate
              - animal, animate being, beast, brute, creature, fauna
                - ...

WORDNET

# 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)

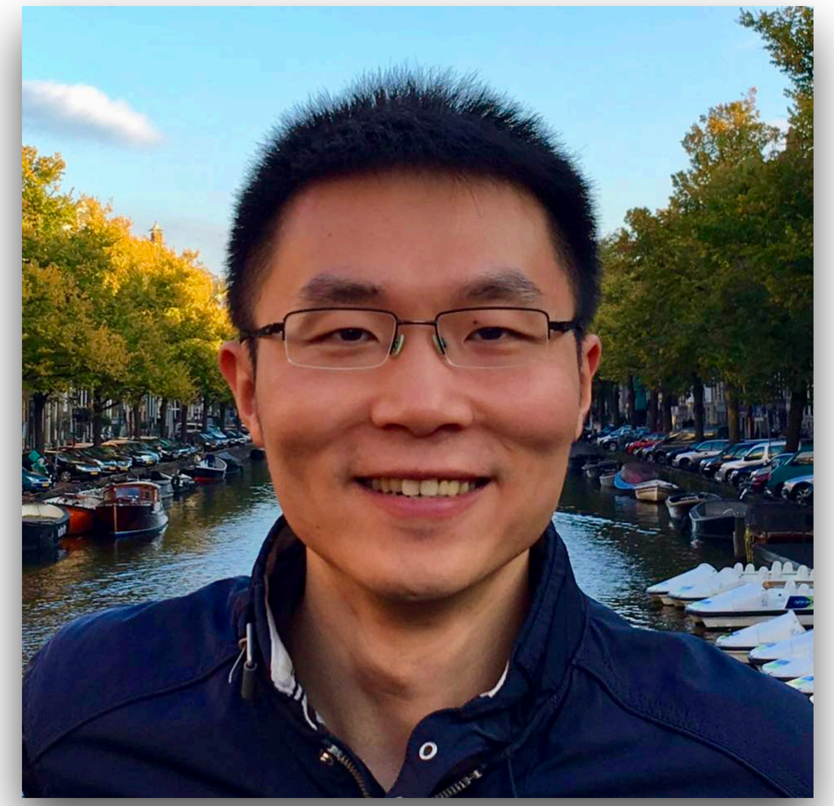
**(Planned) ImageNet is 1000x larger!**

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)

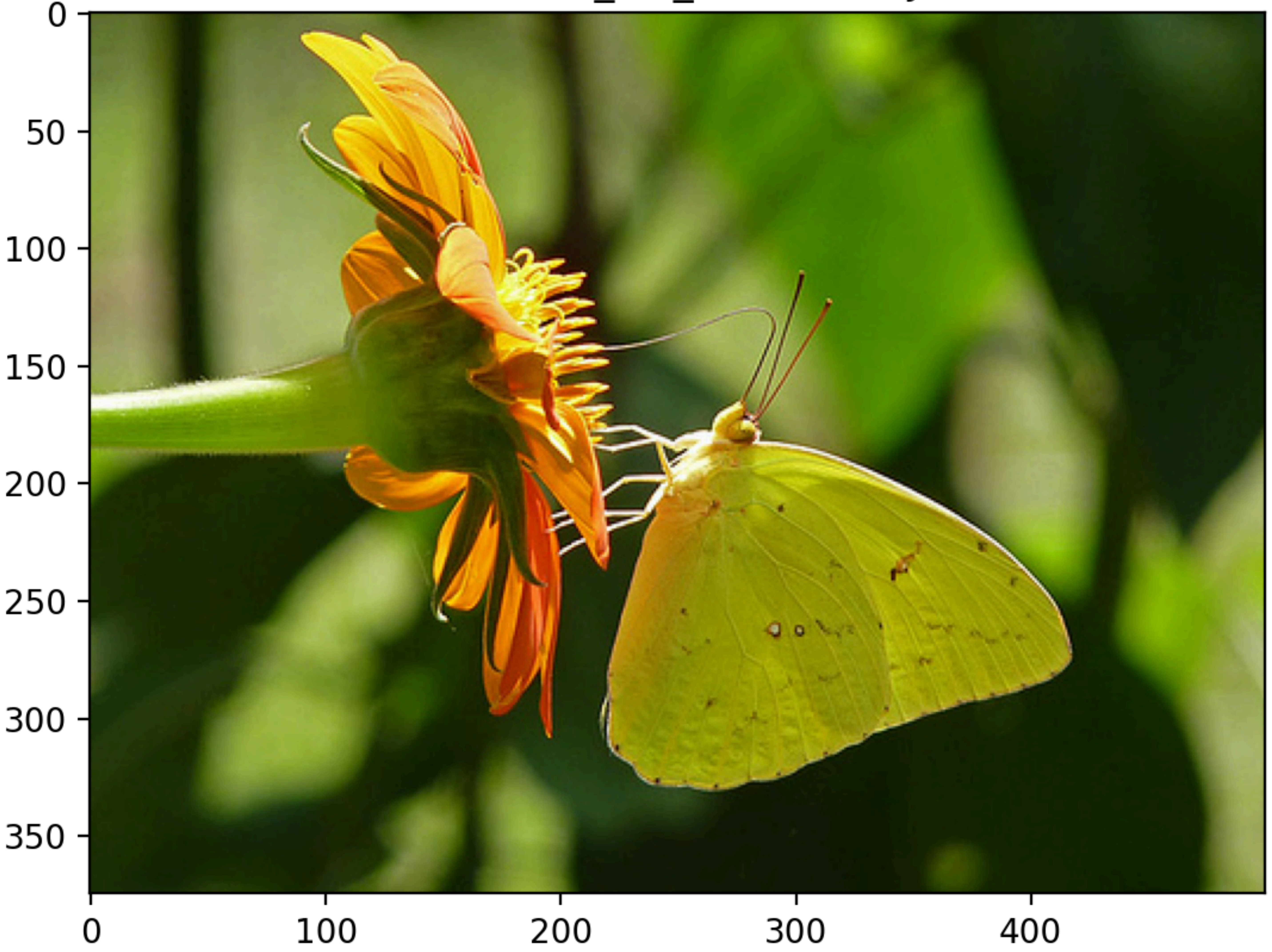
+ lots of **clever** task design

+ lots of **hard** work

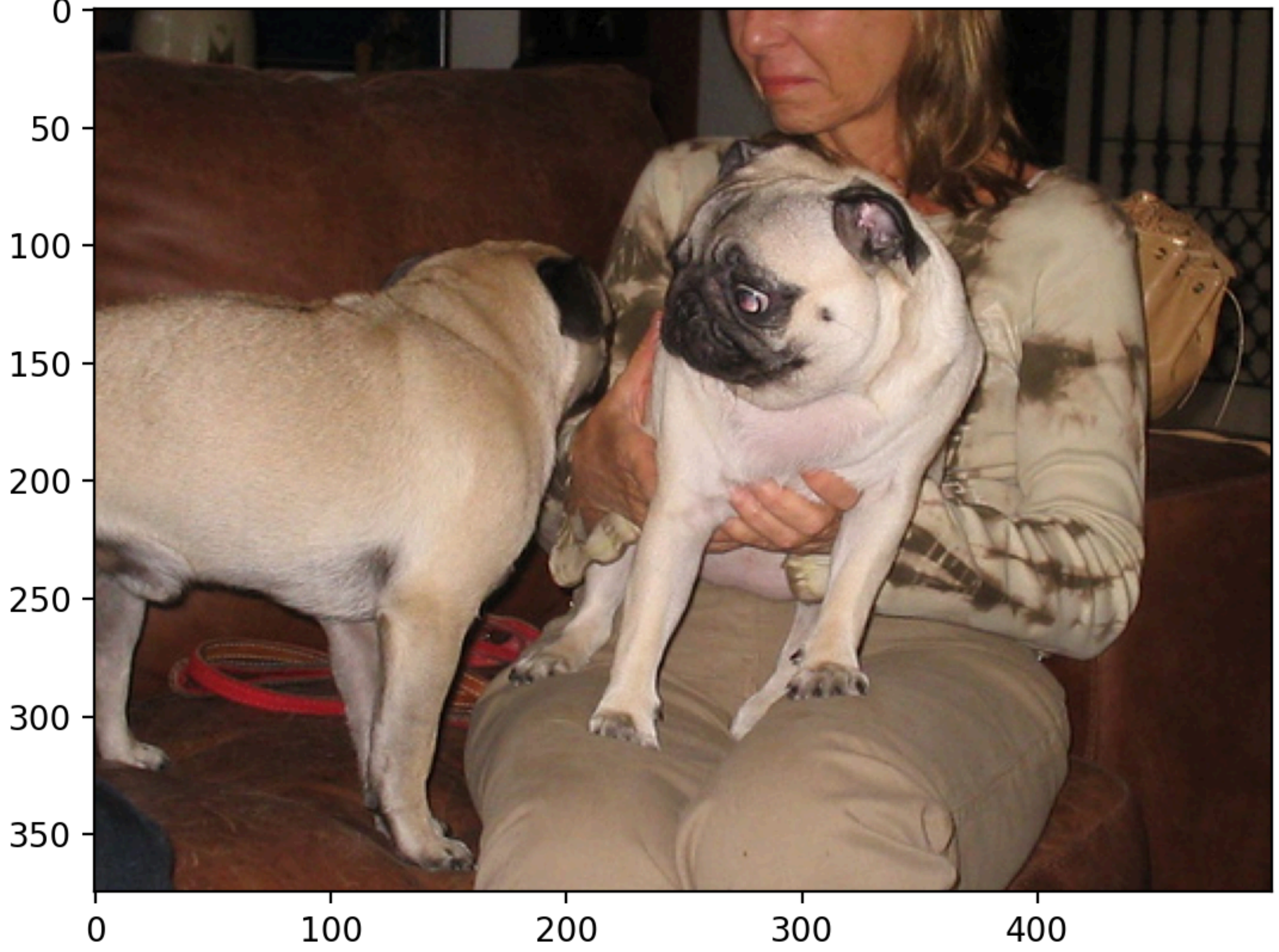
[Deng, Dong, Socher, Li, Li, Fei-Fei'09]

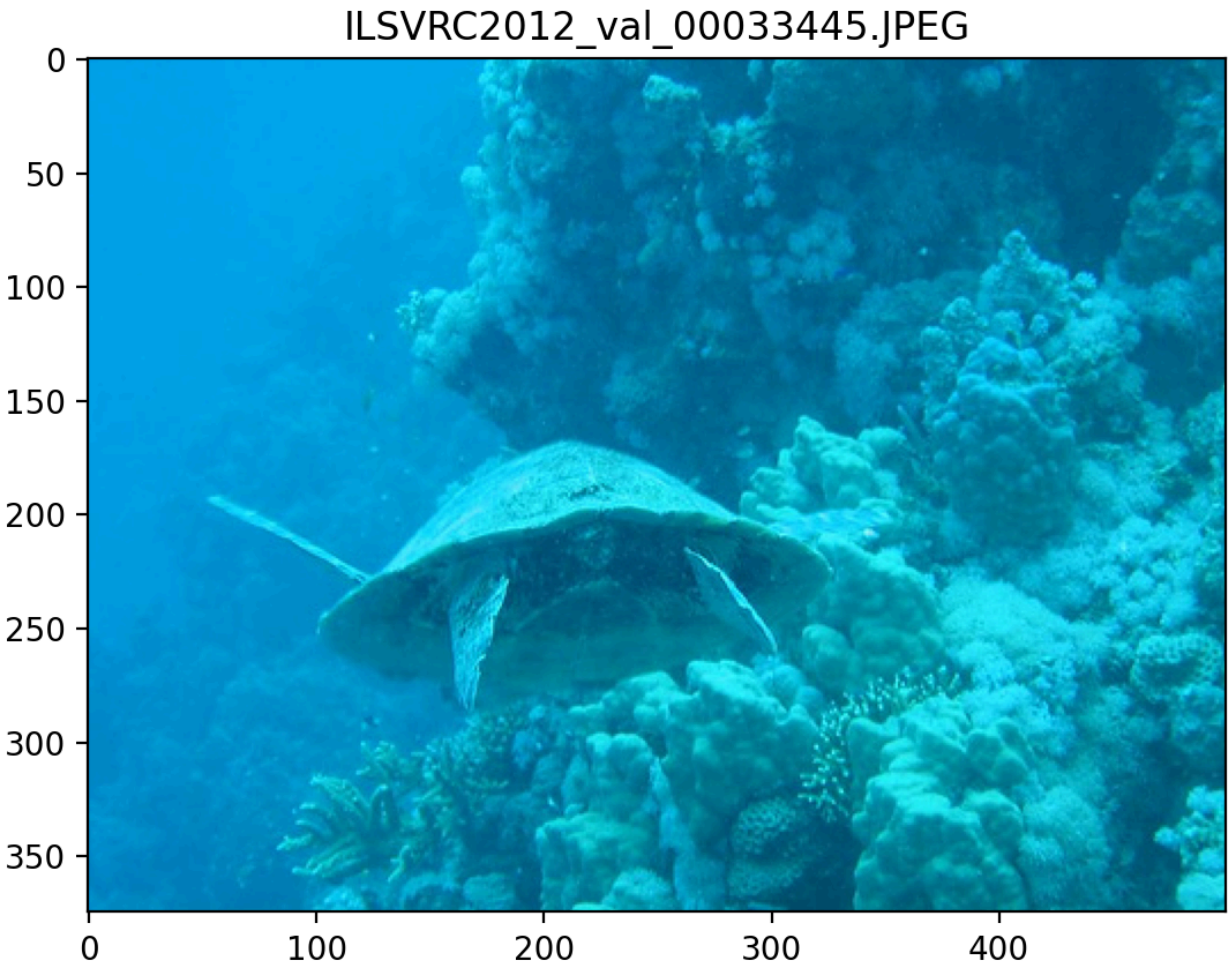
ILSVRC2012_val_00000293.JPEG

ILSVRC2012_val_00025559.JPEG



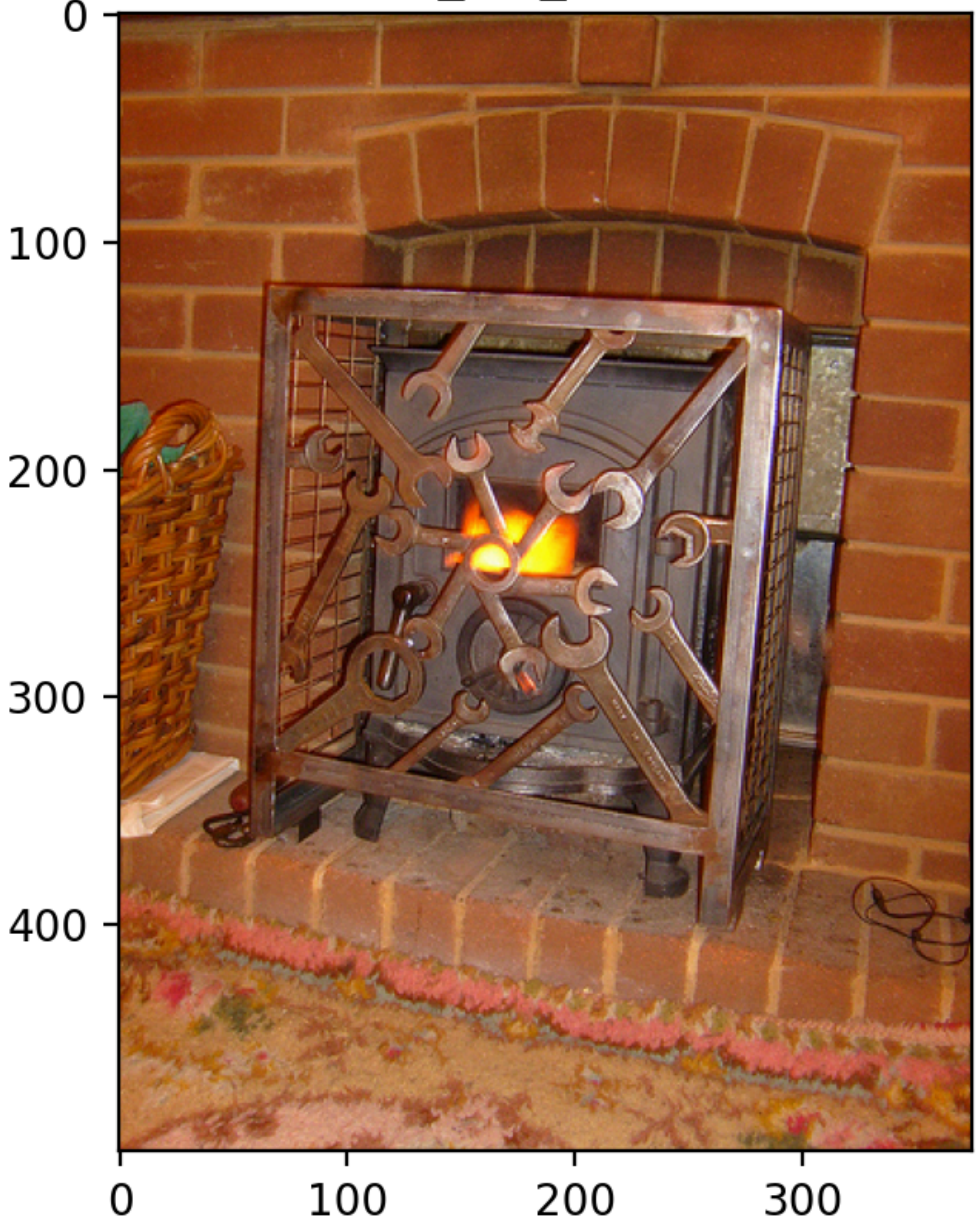
ILSVRC2012_val_00047583.JPEG

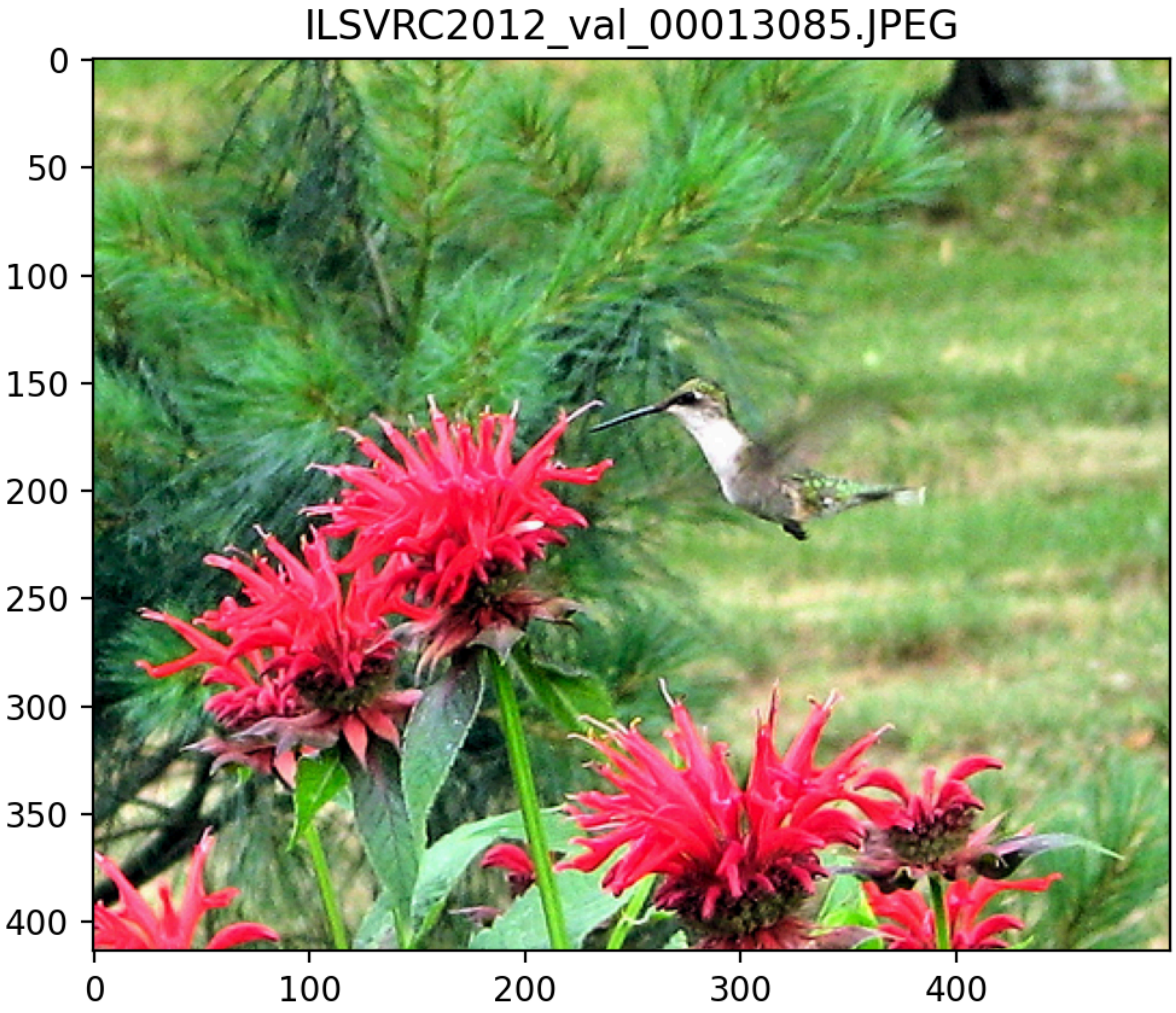




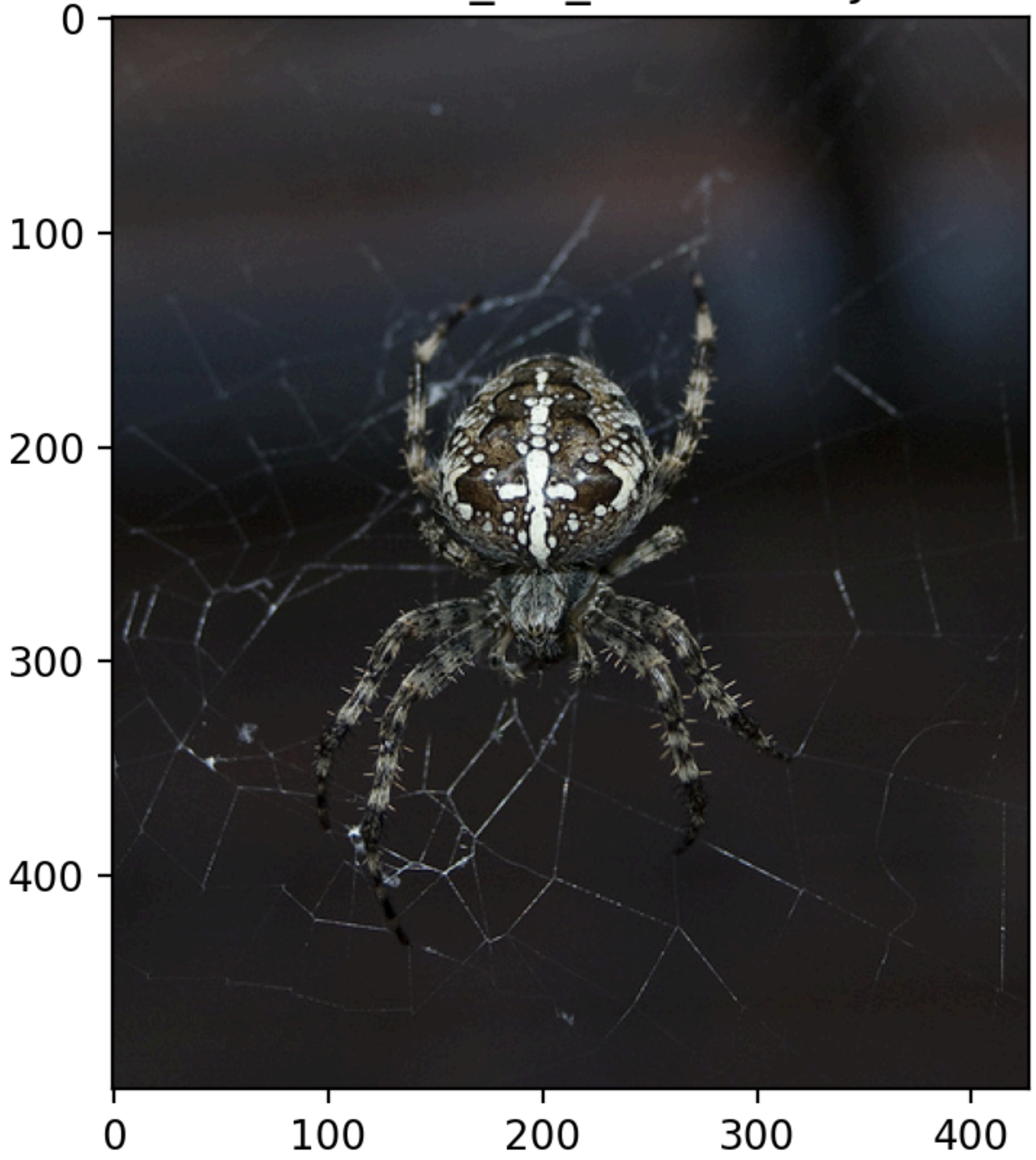
ILSVRC2012_val_00033445.JPEG
0
50
100
150
200
250
300
350
0
100
200
300
400

ILSVRC2012_val_00027126.JPEG

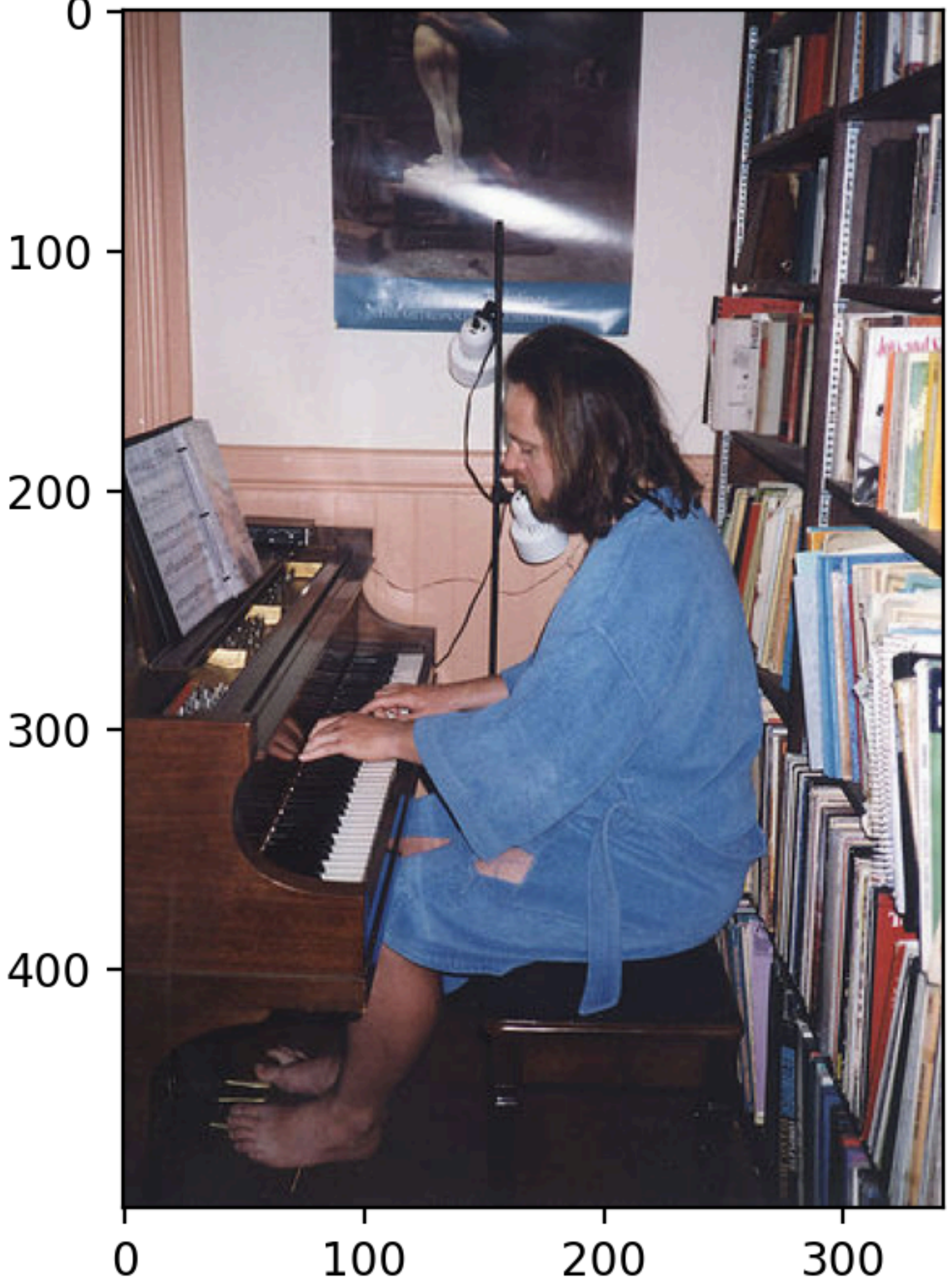



ILSVRC2012_val_00013085.JPEG
0
50
100
150
200
250
300
350
400
0
100
200
300
400

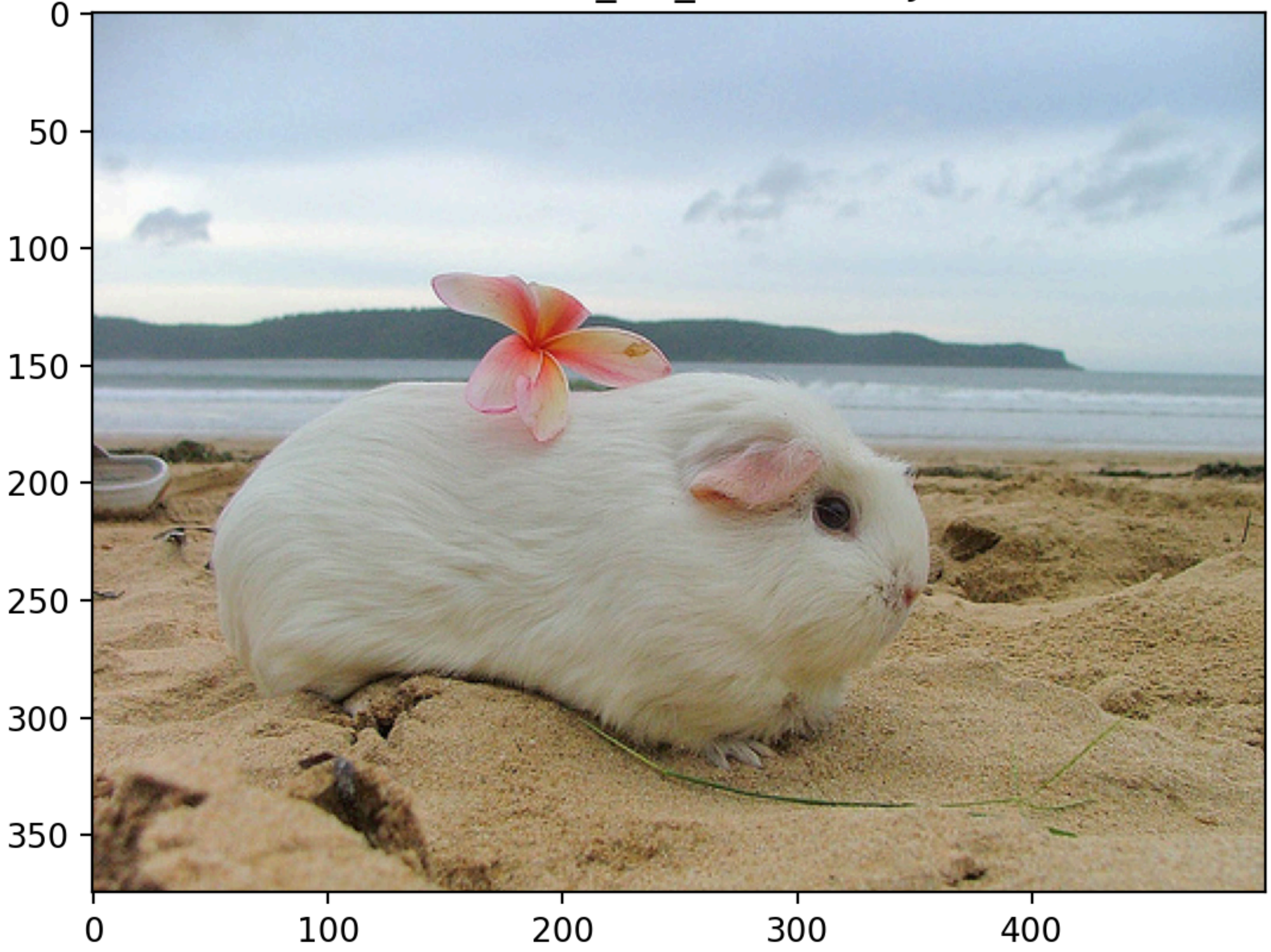
ILSVRC2012_val_00035593.JPEG



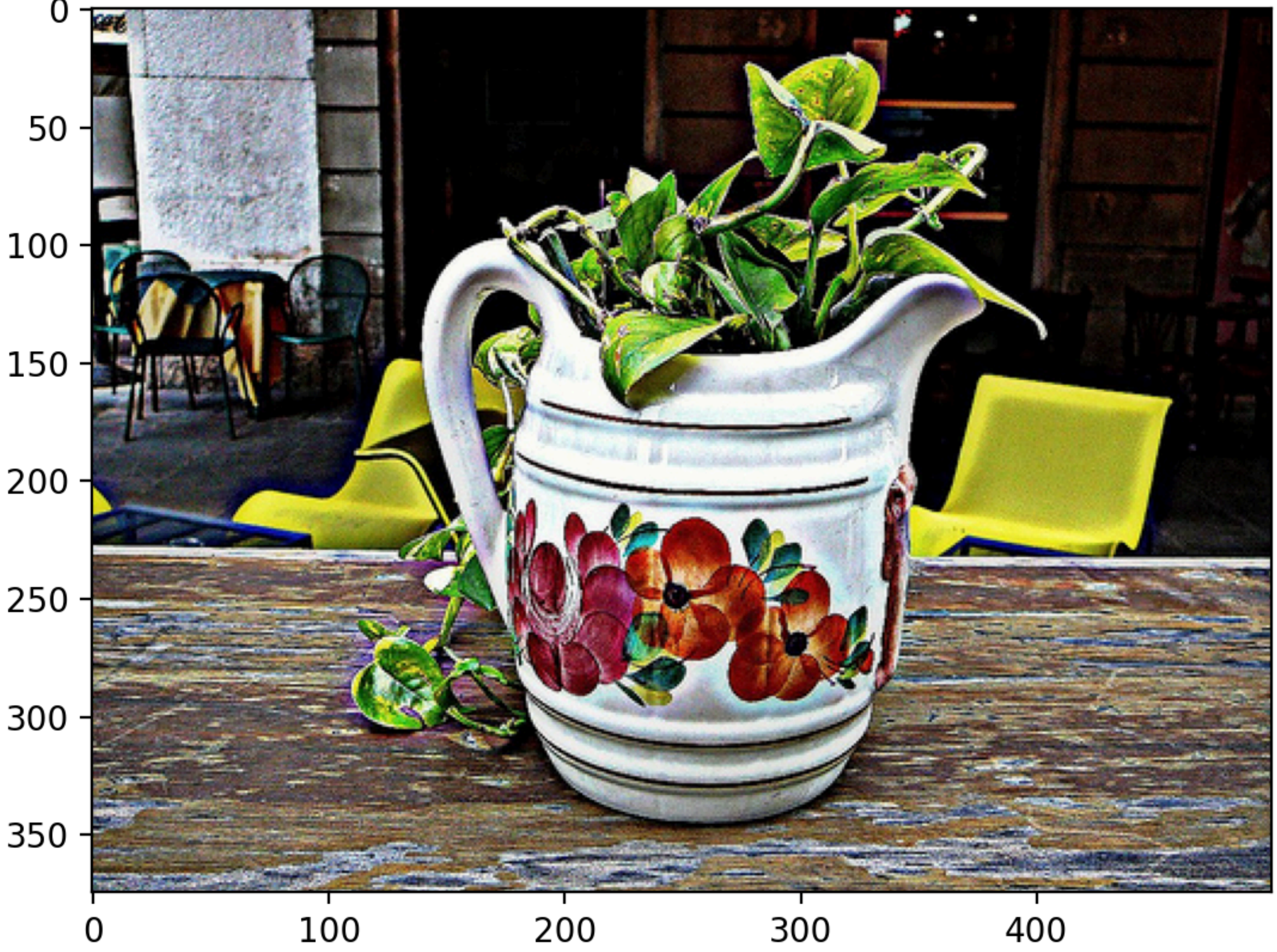
ILSVRC2012_val_00012694.JPEG



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)

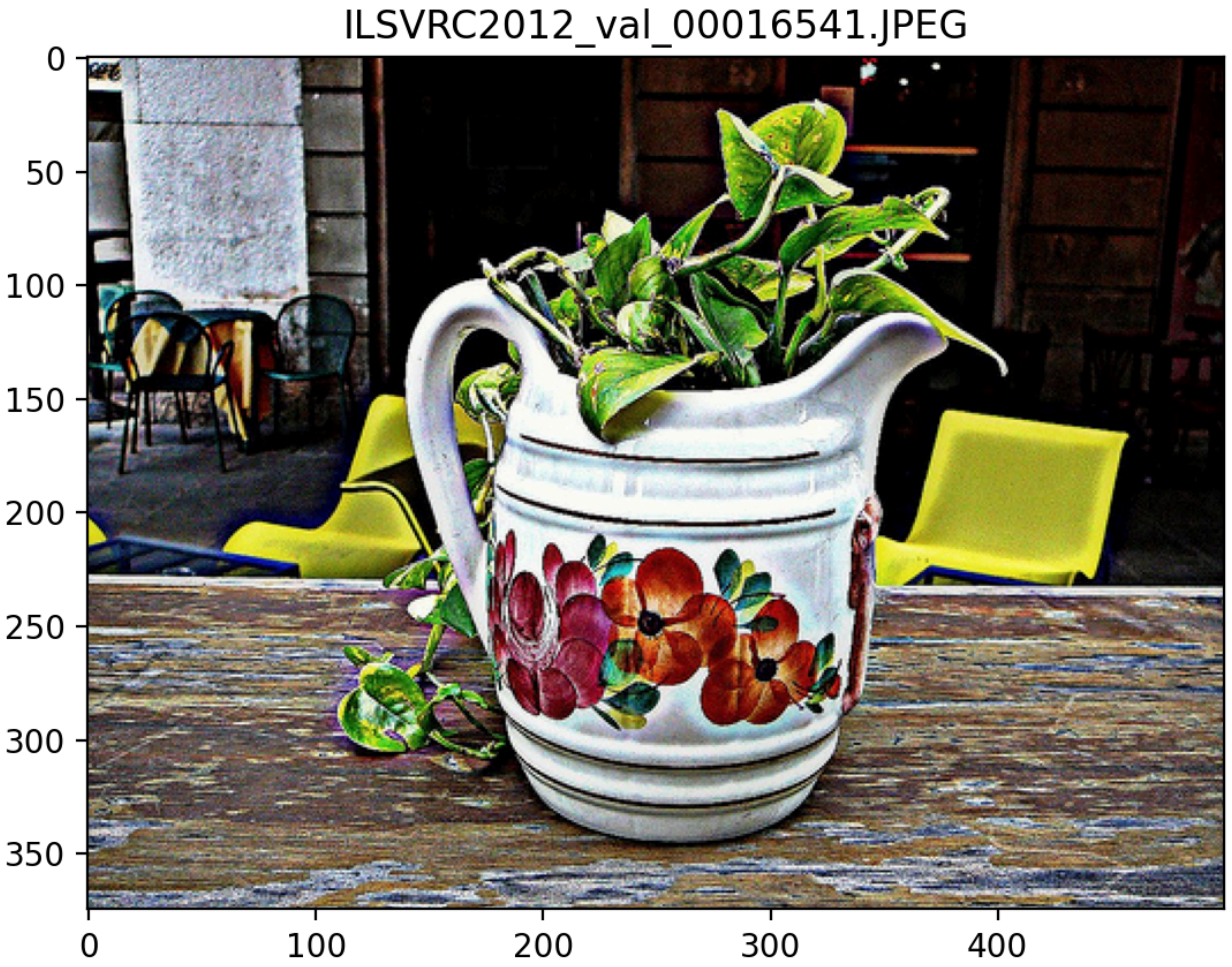
**Validation set:** 50,000 images for 1,000 classes (exactly class-balanced)

**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



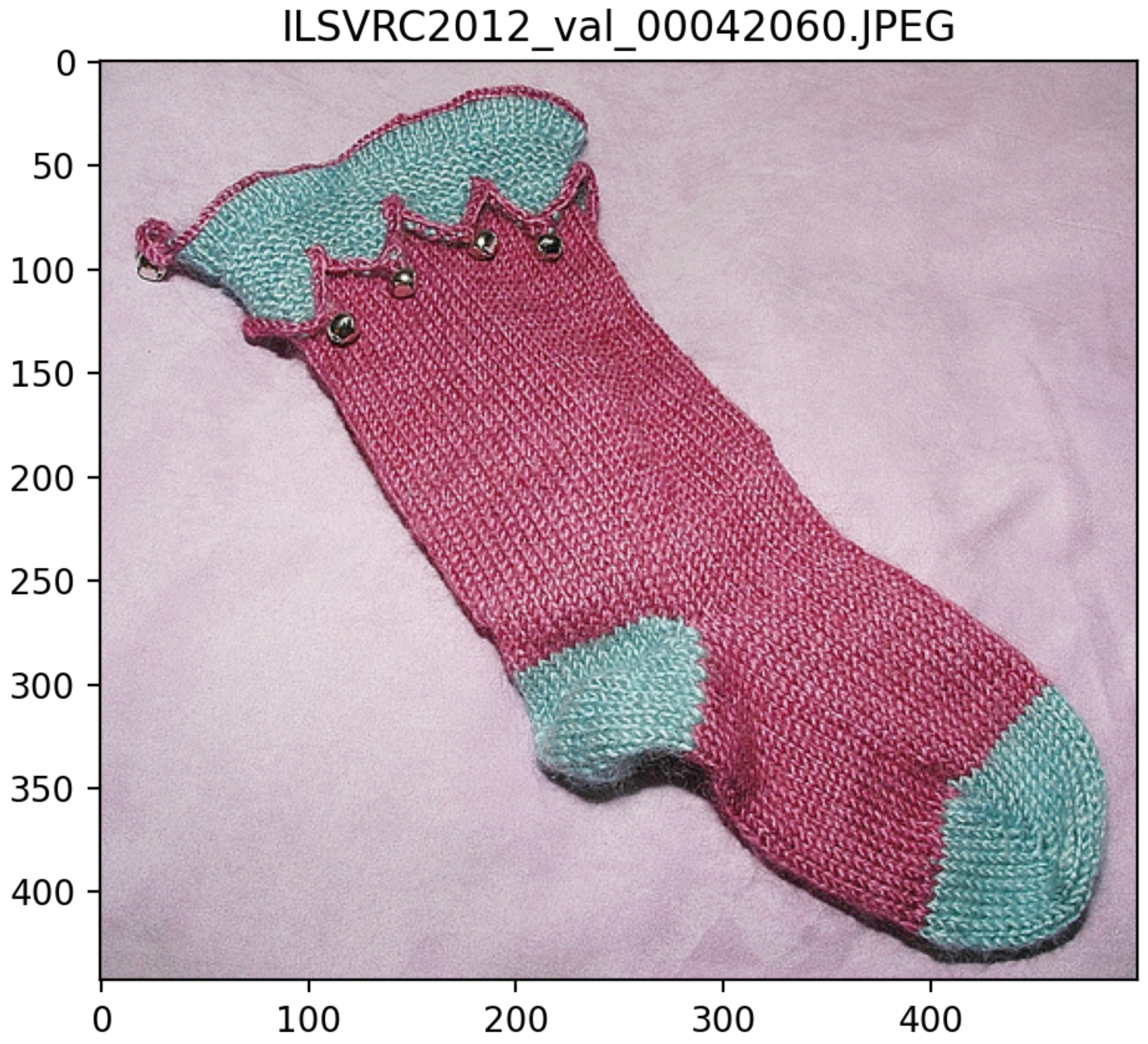

n03950228 pitcher, ewer

WordNet ID (wnid)

Synonym set

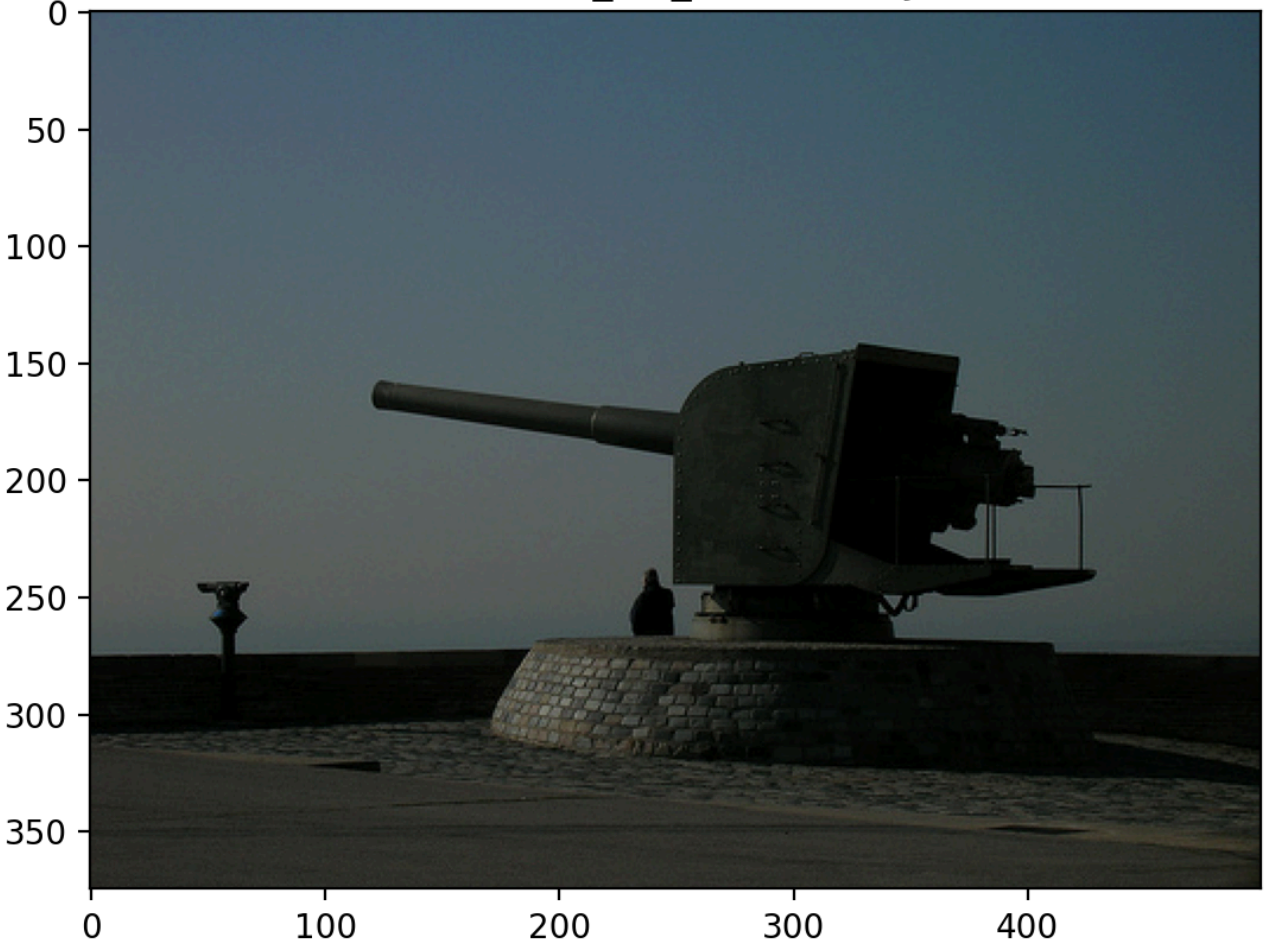
ILSVRC2012_val_00007151.JPEG


n02488702 colobus, colobus monkey

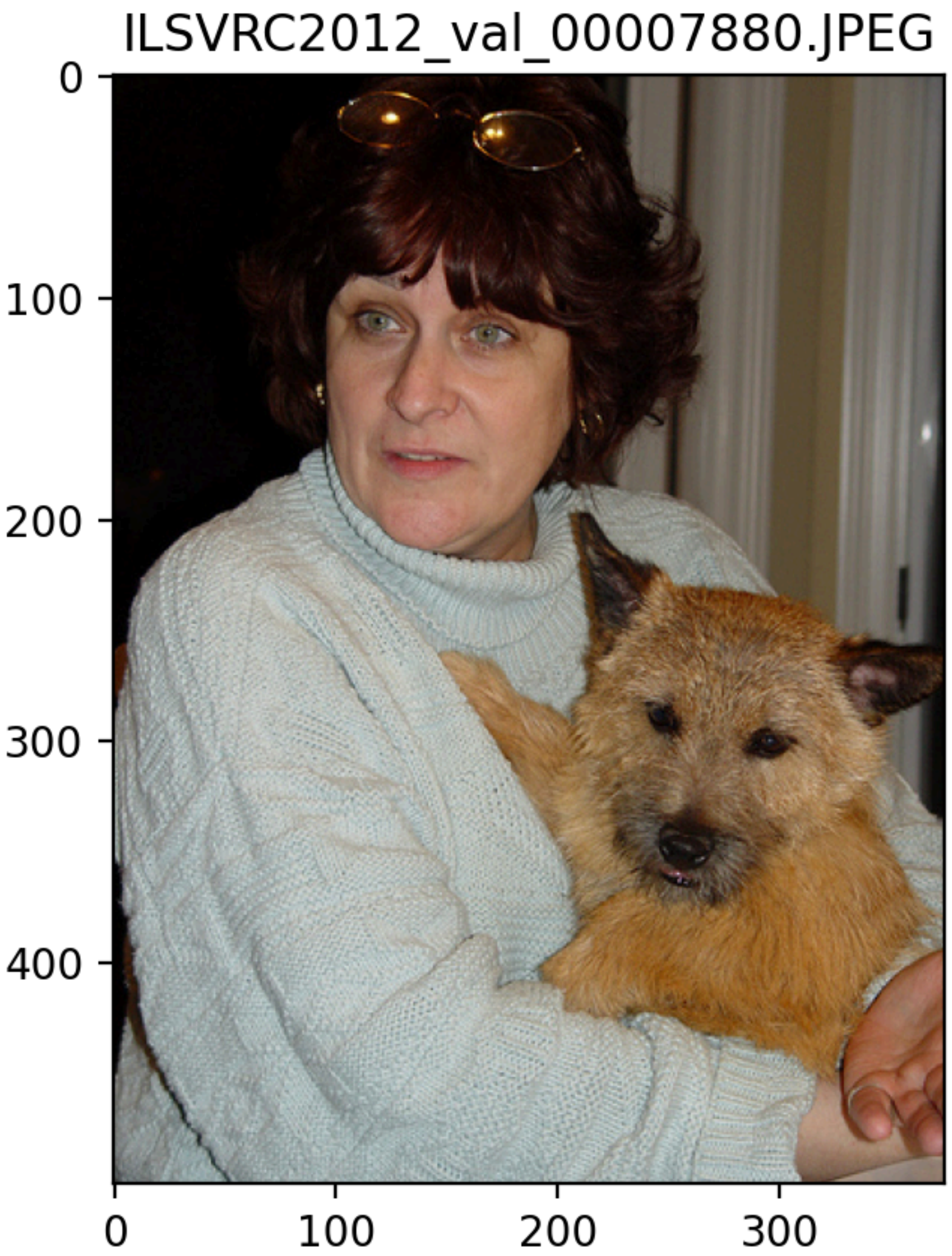



n03026506 Christmas stocking

ILSVRC2012_val_00001902.JPEG

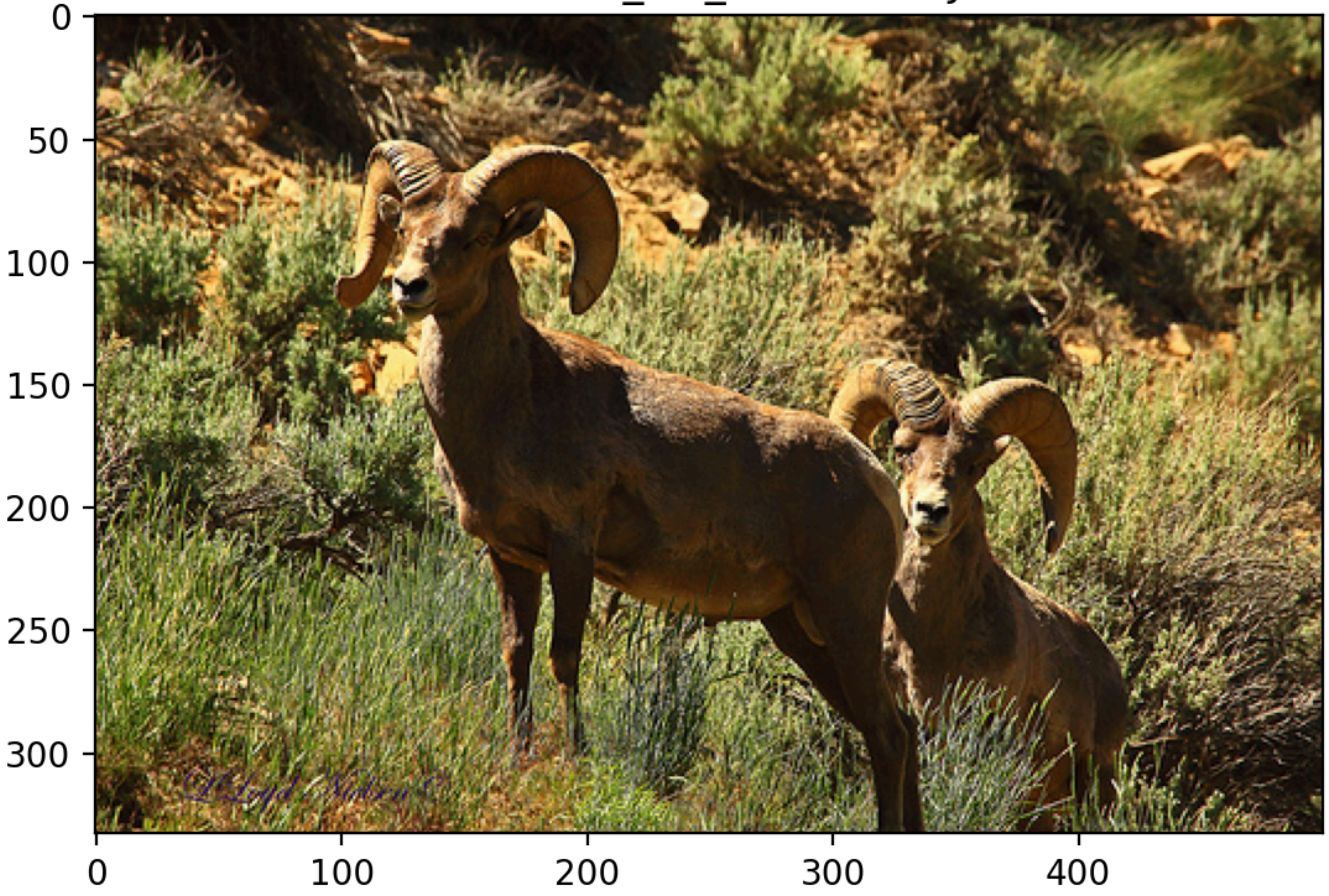



n02950826 cannon



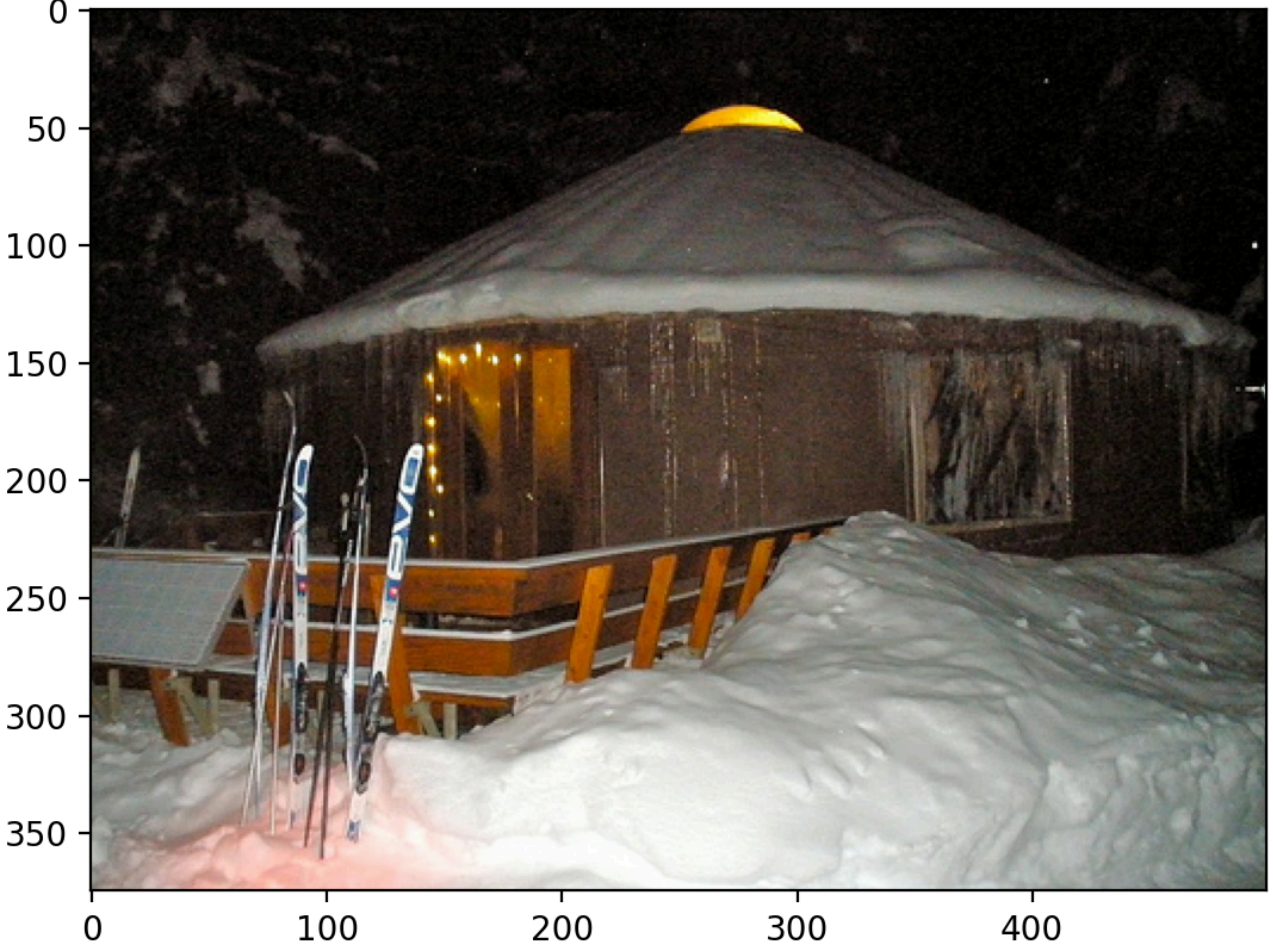

n02094258 Norwich terrier

ILSVRC2012_val_00016391.JPEG

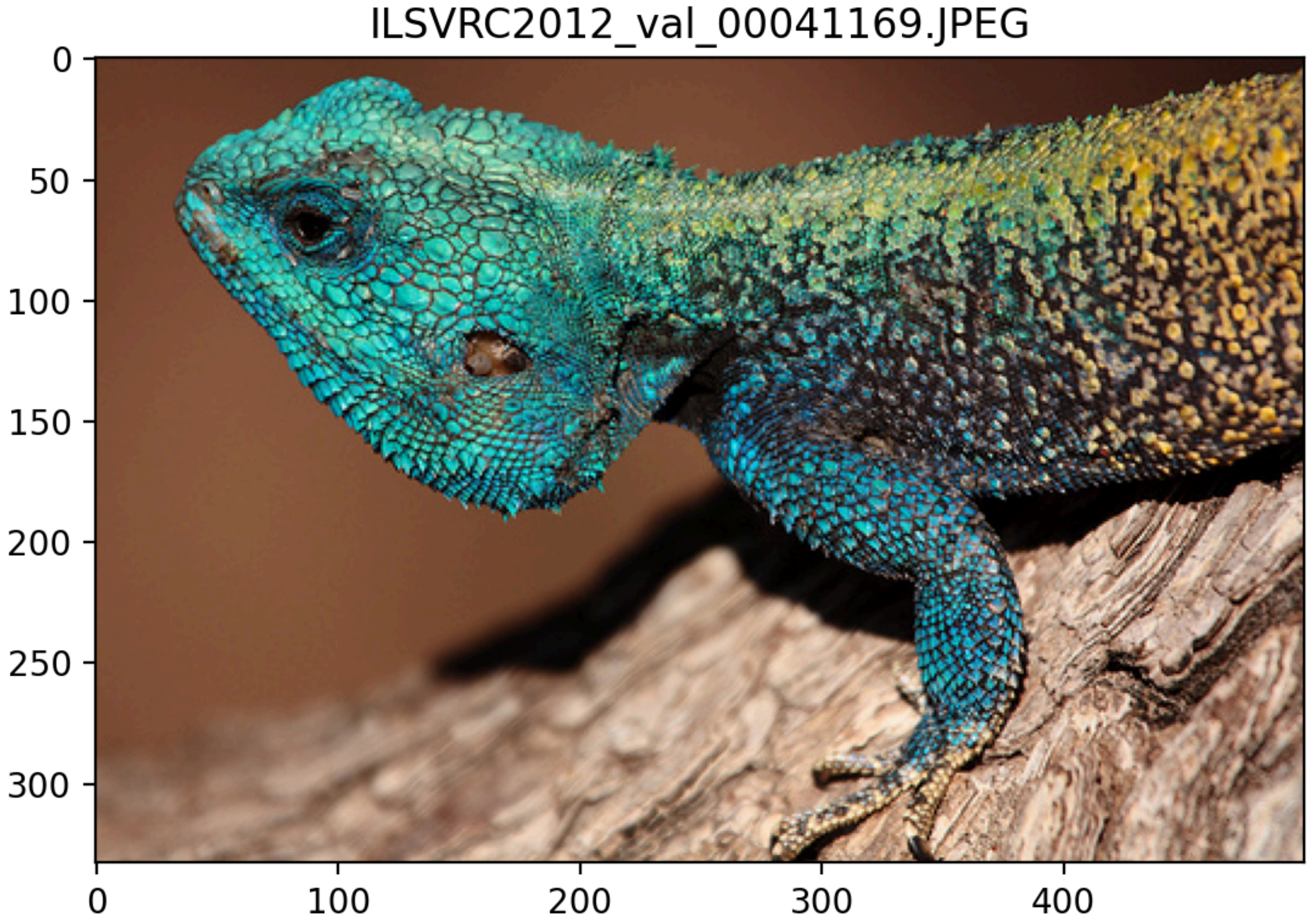



n02412080 ram, tup

ILSVRC2012_val_00020151.JPEG



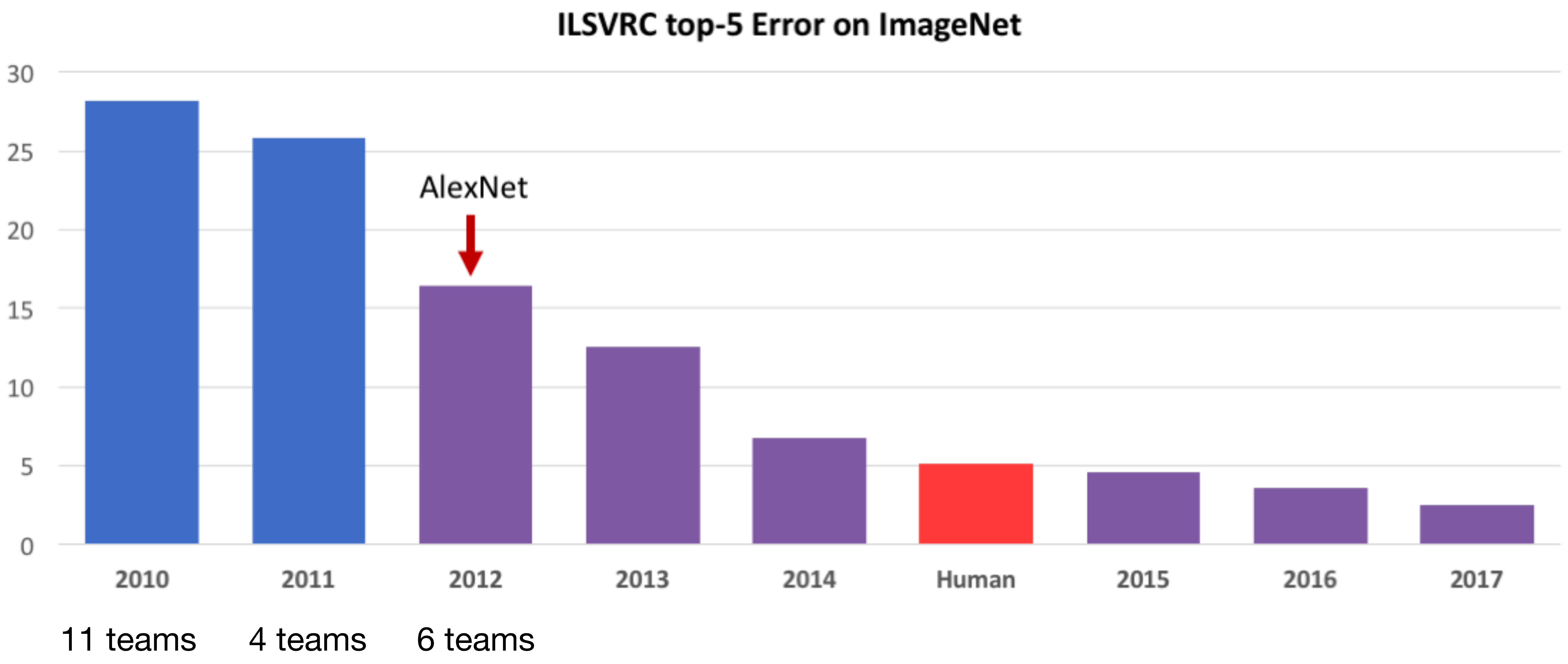

n04613696 yurt




n01687978 agama


n02134418 sloth bear, Melursus ursinus, Ursus ursinus


n04591713     wine bottle


ILSVRC top-5 Error on ImageNet
30
25
20
15
10
5
0
AlexNet
2010
2011
2012
2013
2014
Human
2015
2016
2017
11 teams
4 teams
6 teams

# 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

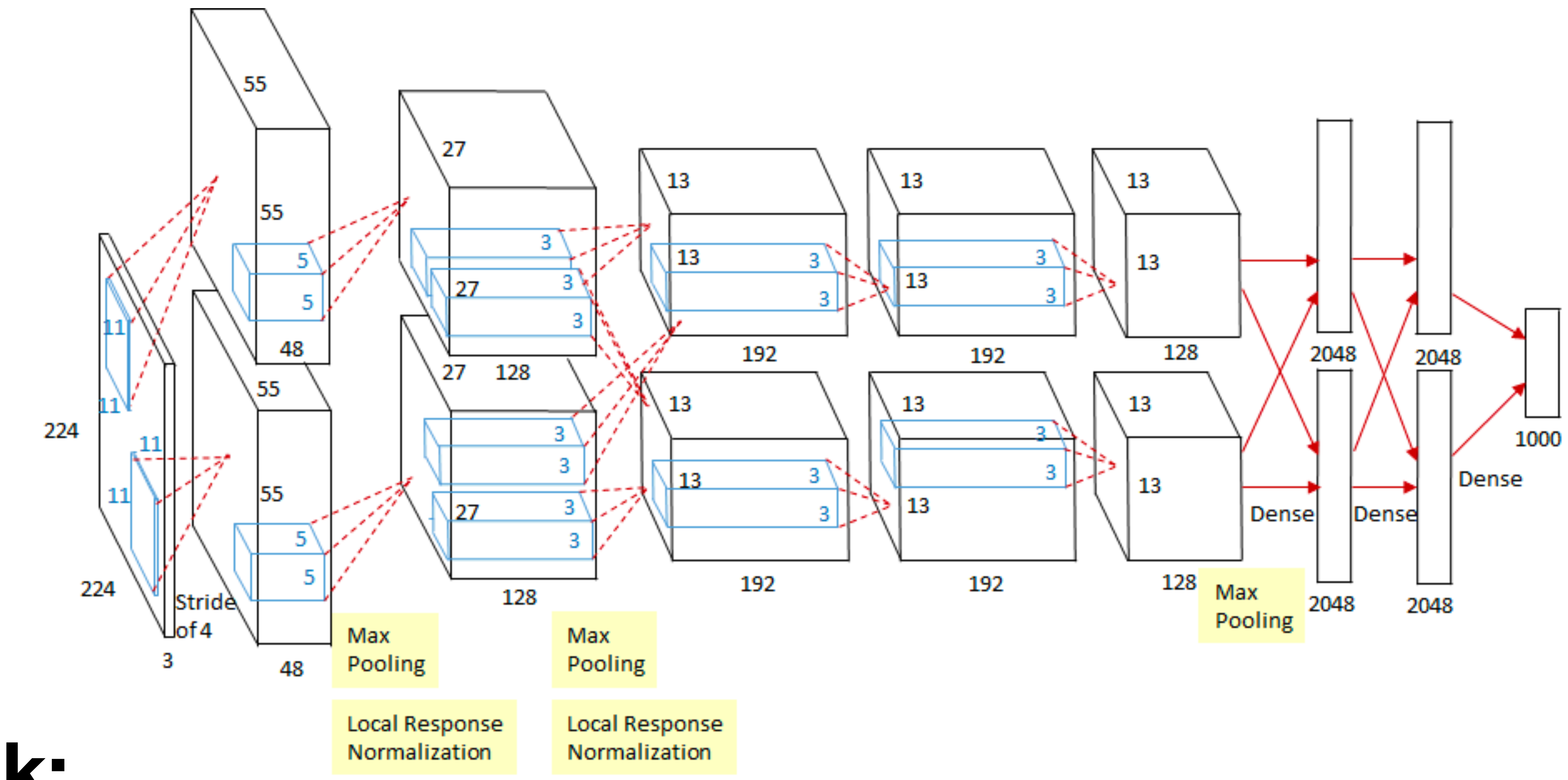
### Abstract

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



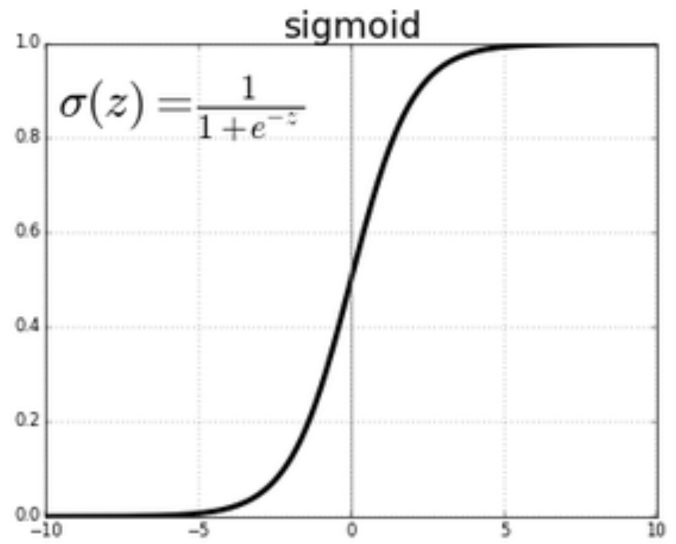

**Basic building block:**

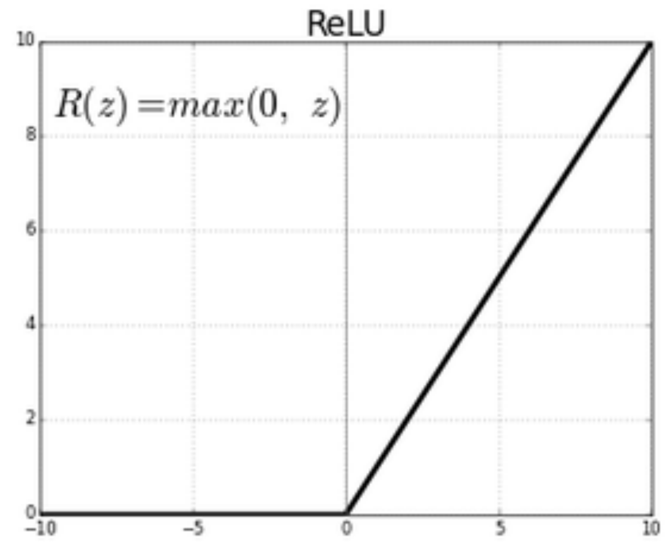
Structured, learnable linear layer followed by a simple element-wise non-linearity

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

# AlexNet Ingredients

**ReLU (rectified linear unit) non-linearity**



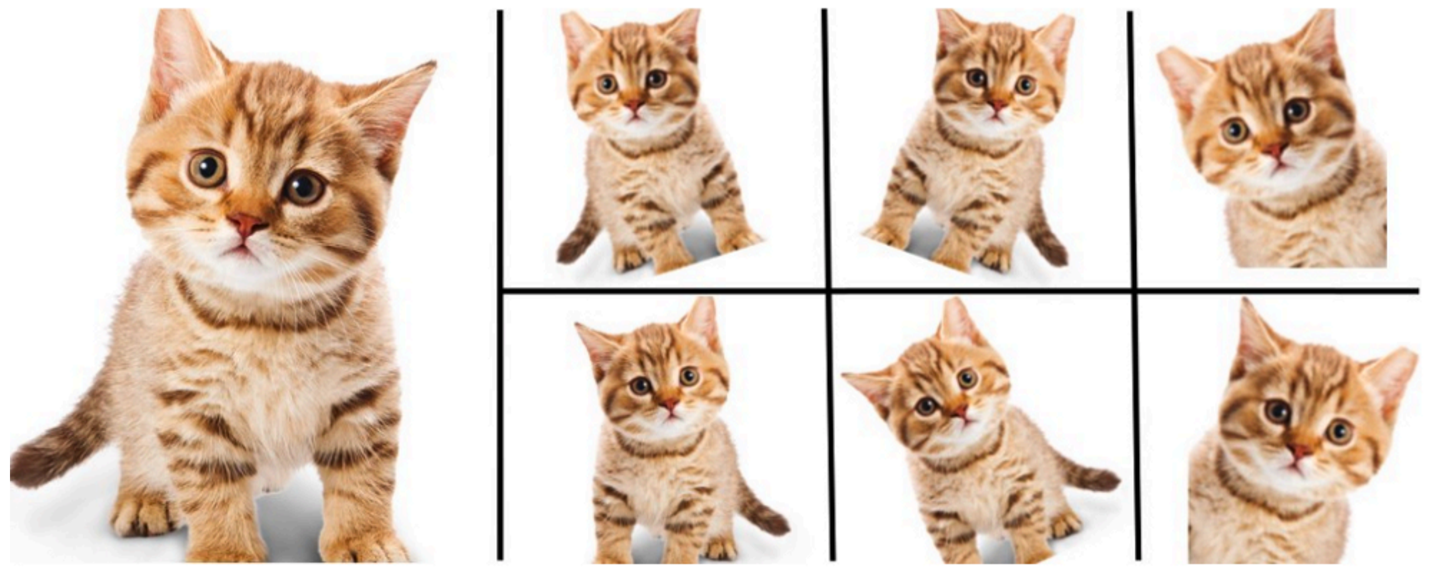

Local response normalization

**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 |
|---|---|---|
| *Sparse coding [2]* | *47.1%* | *28.2%* |
| *SIFT + FVs [24]* | *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) |
|---|---|---|---|
| *SIFT + FVs [7]* | — | — | *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
- 130,000 citations, Turing award, transformation of computer science

???

# 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 — Oct 14, 2012

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 is 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 t

a

w

f

d

a

**Geoffrey Hinton** +31 — Oct 15, 2012

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.

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

d at it and failed to do nearly as well. I also think

scredit a system that gets about 84% correct by

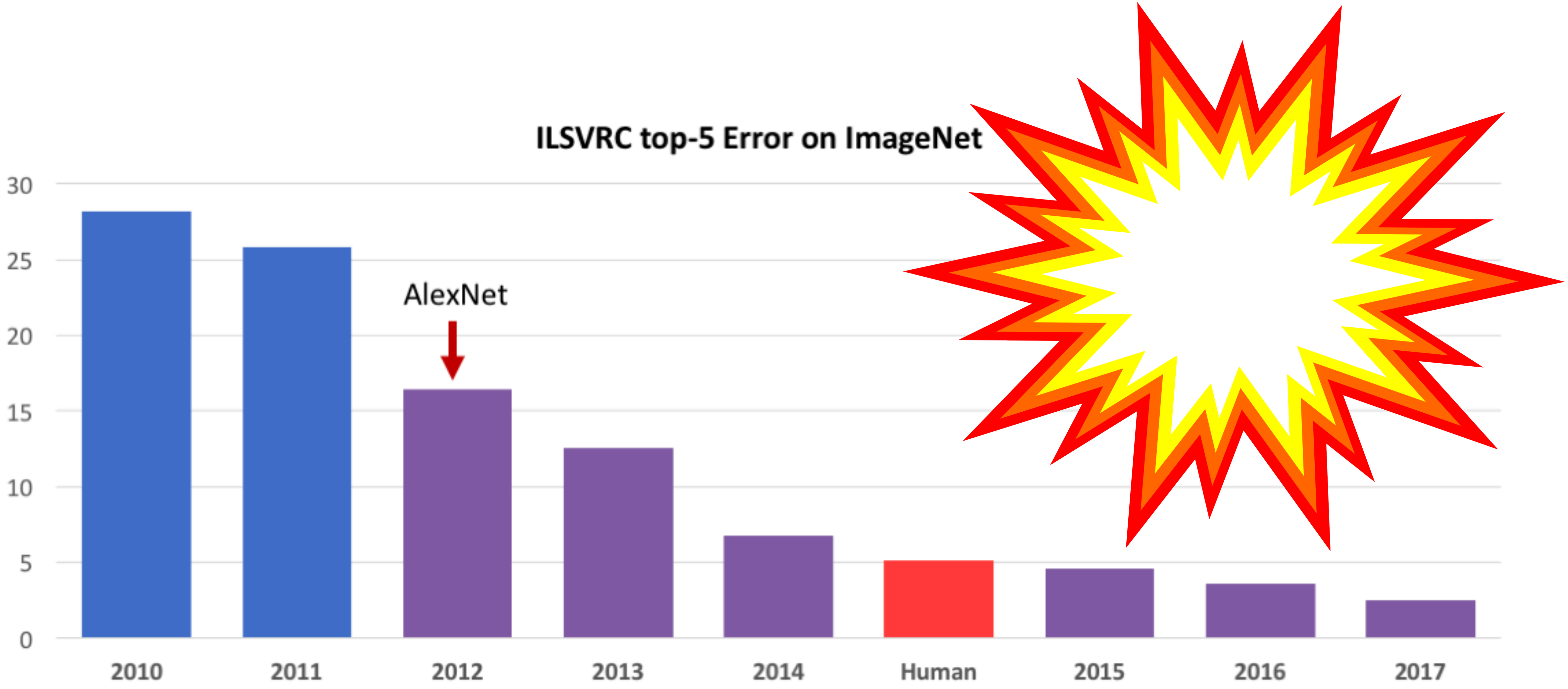
d get 0.5% correct by chance is a bit desperate.

**Yann LeCun** +16 — Oct 16, 2012

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.

NOTE: Alyosha is a great scientist.

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




11 teams 4 teams 6 teams **24 teams** **32 teams** **68 teams** **84 teams** **28 teams**

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)

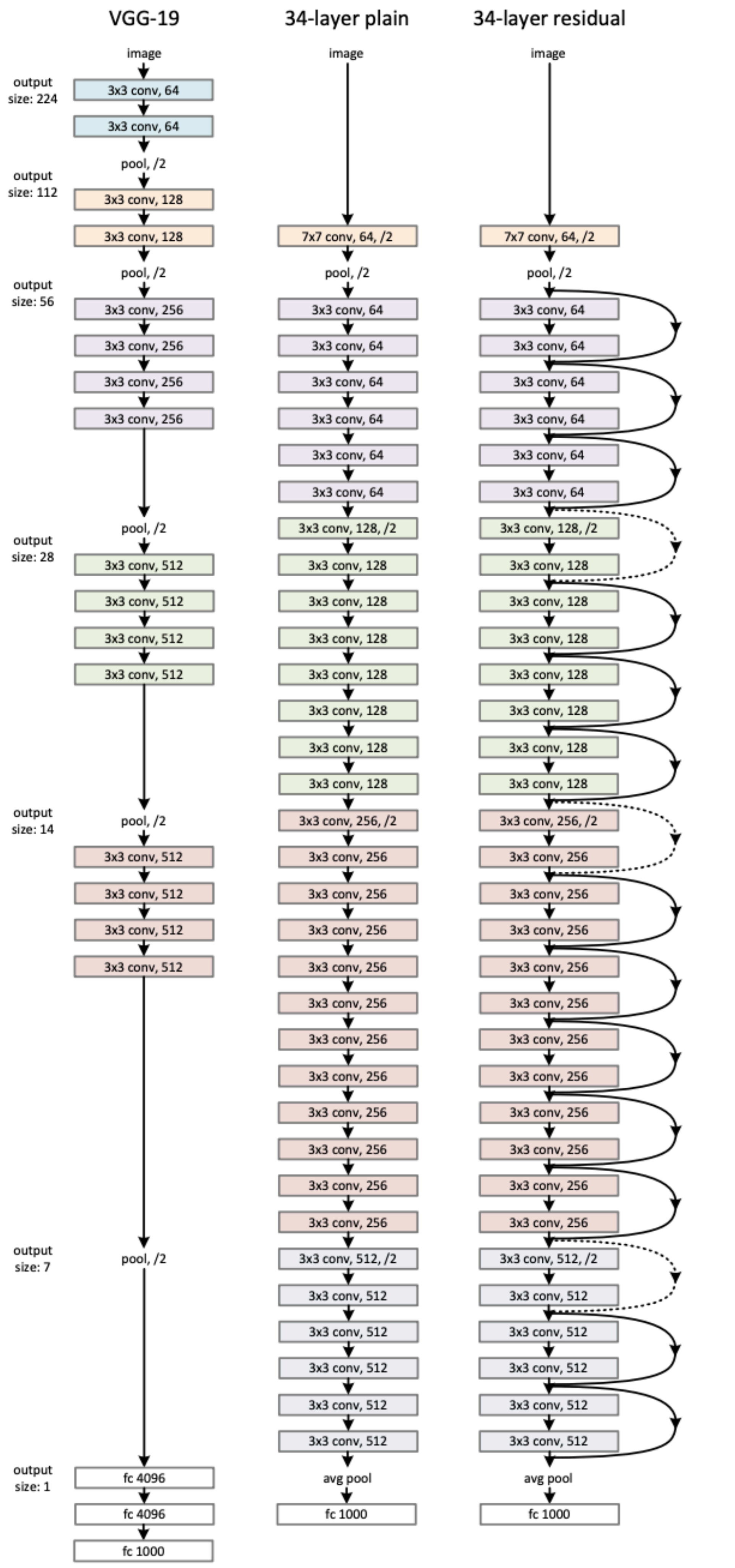
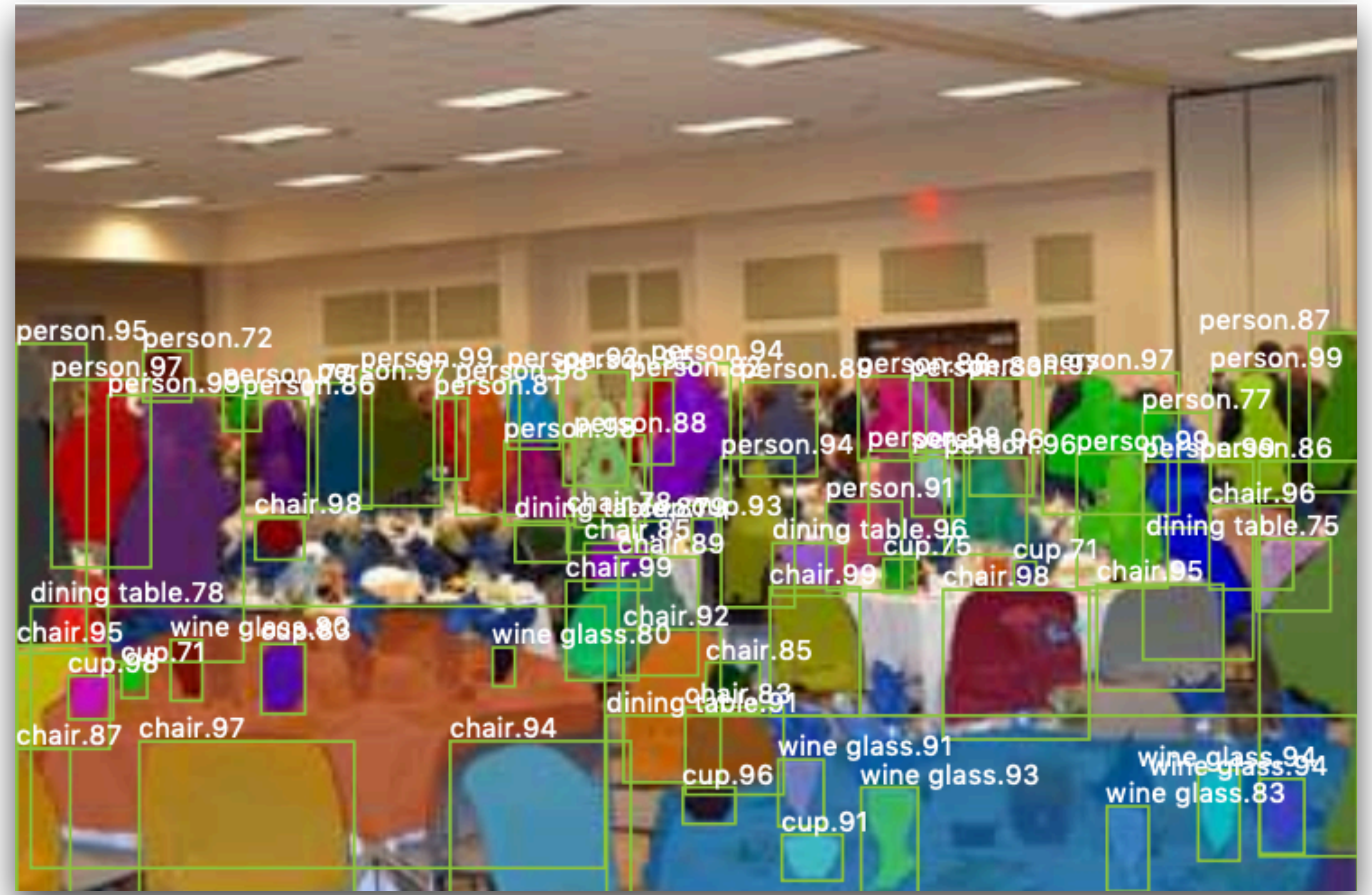


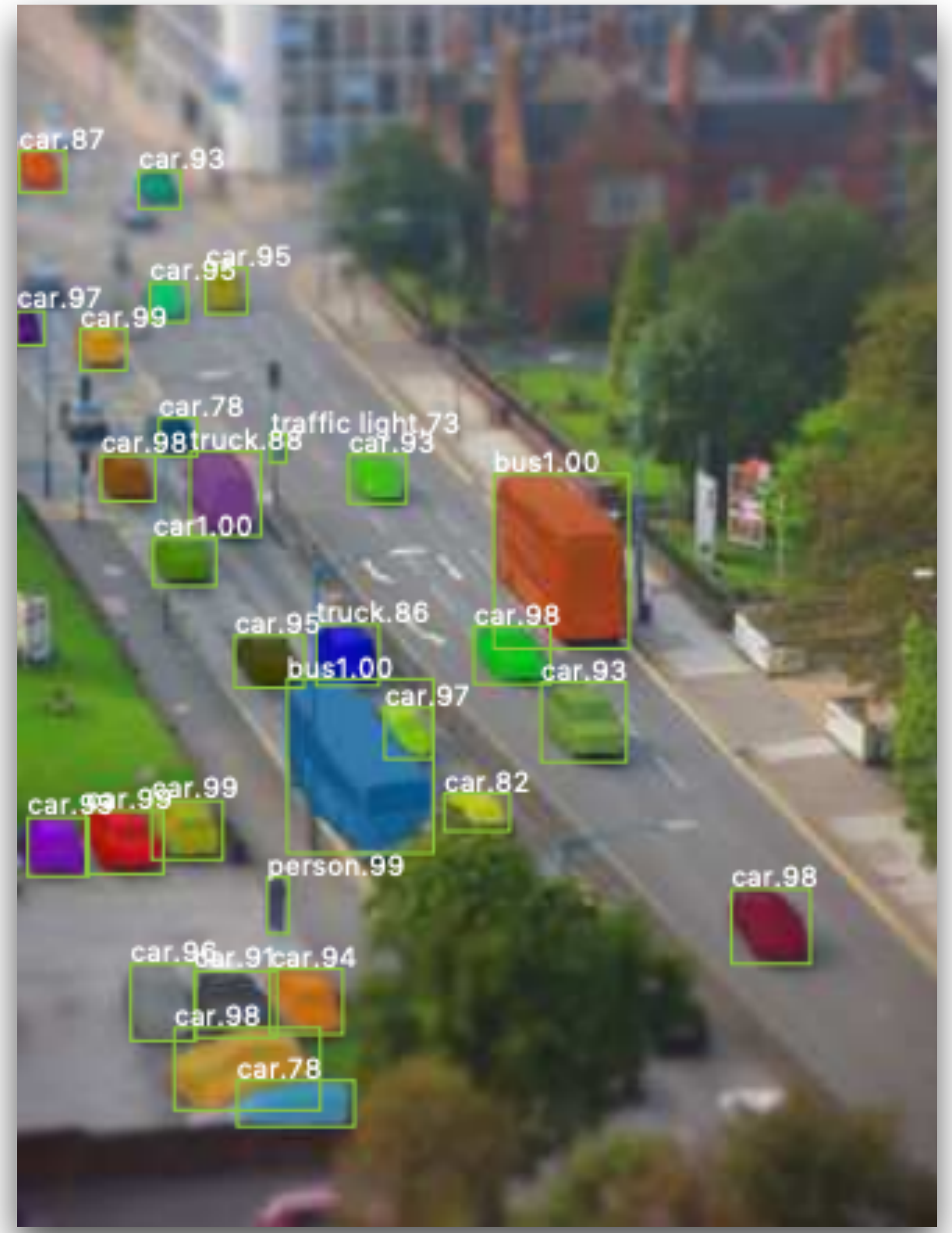

Figure 3. Example network architectures for ImageNet. **Left**: the VGG-19 model [41] (19.6 billion FLOPs) as a reference. **Middle**: a plain network with 34 parameter layers (3.6 billion FLOPs). **Right**: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase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

## Karl Marx

FOLLOW

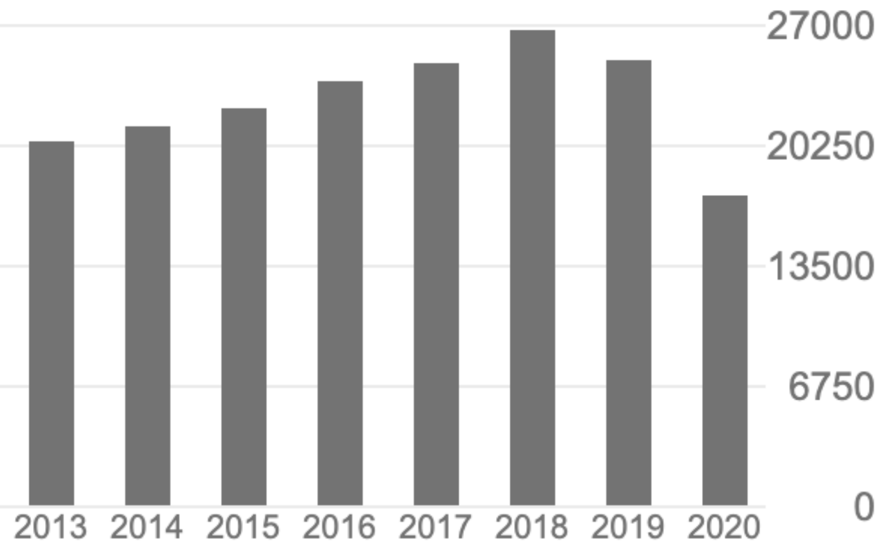
British National Library
Verified email at tsn.at

Kapitalismuskritiker Marxist Religionskritiker Philosophie Soziologie

| TITLE | CITED BY | YEAR |
|---|---|---|
| Le capital<br>K Marx<br>Librairie du progrès | 38580 | 1875 |
| Capital: volume I<br>K Marx<br>Penguin UK | 19350 * | 2004 |
| The communist manifesto<br>K Marx, F Engels<br>Penguin | 11661 | 2002 |
| The german ideology<br>K Marx, F Engels<br>International Publishers Co | 11652 | 1970 |
| Grundrisse: Foundations of the critique of political economy<br>K Marx<br>Penguin UK | 11326 | 2005 |
| A ideologia alemã: crítica da mais recente filosofia alemã em seus representantes Feuerbach, B. Bauer e Stirner, e do socialismo alemão em seus diferentes profetas<br>K Marx, F Engels<br>Boitempo editorial | 8366 | 2015 |
| Das kapital<br>K Marx<br>e-artnow | 7511 | 2018 |

Cited by VIEW ALL

| | All | Since 2015 |
|---|---|---|
| Citations | $\boxed{381827}$ | 142067 |
| h-index | 213 | 134 |
| i10-index | 1431 | 902 |



# Historical Comparison - Revolutions

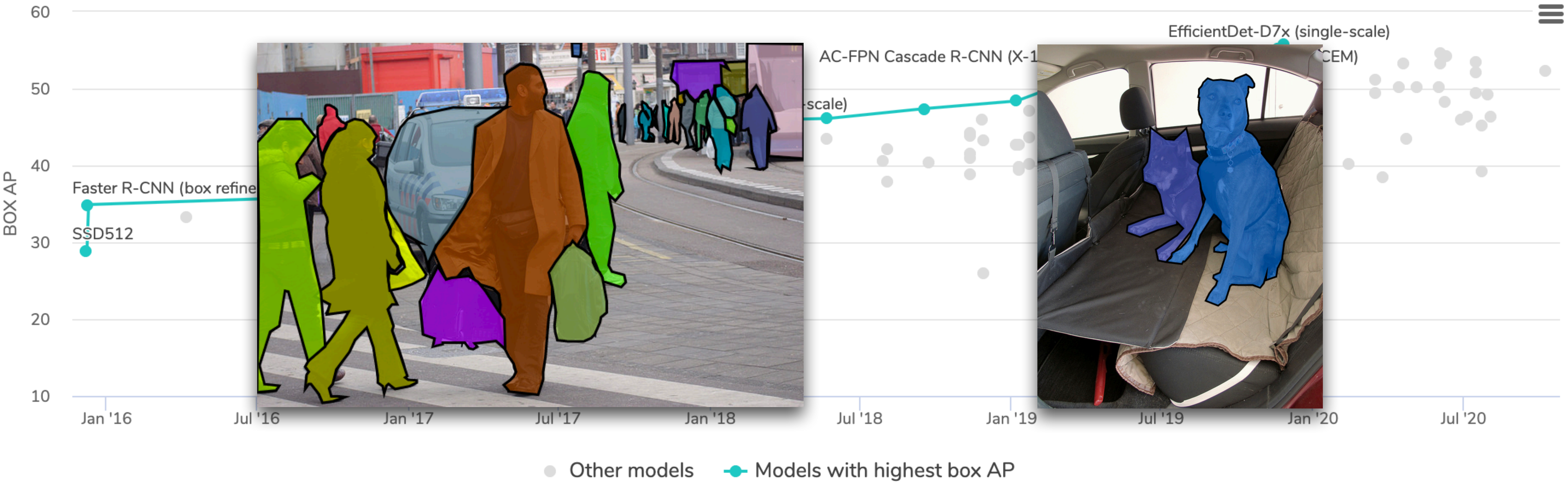

10k more than Marx!

**CAVEAT: DO NOT MEASURE SCIENCE BY CITATION COUNT**

# Similar Performance Trends for Many Other Datasets

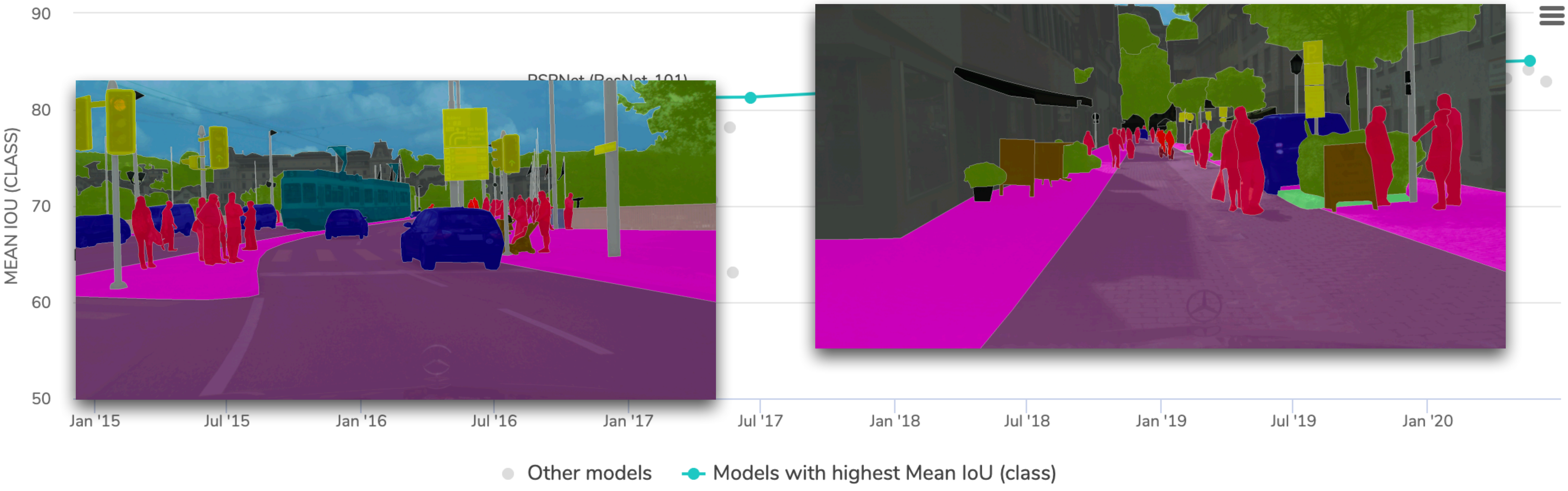
Object detection (PASCAL VOC)

# Object Detection (MS COCO)

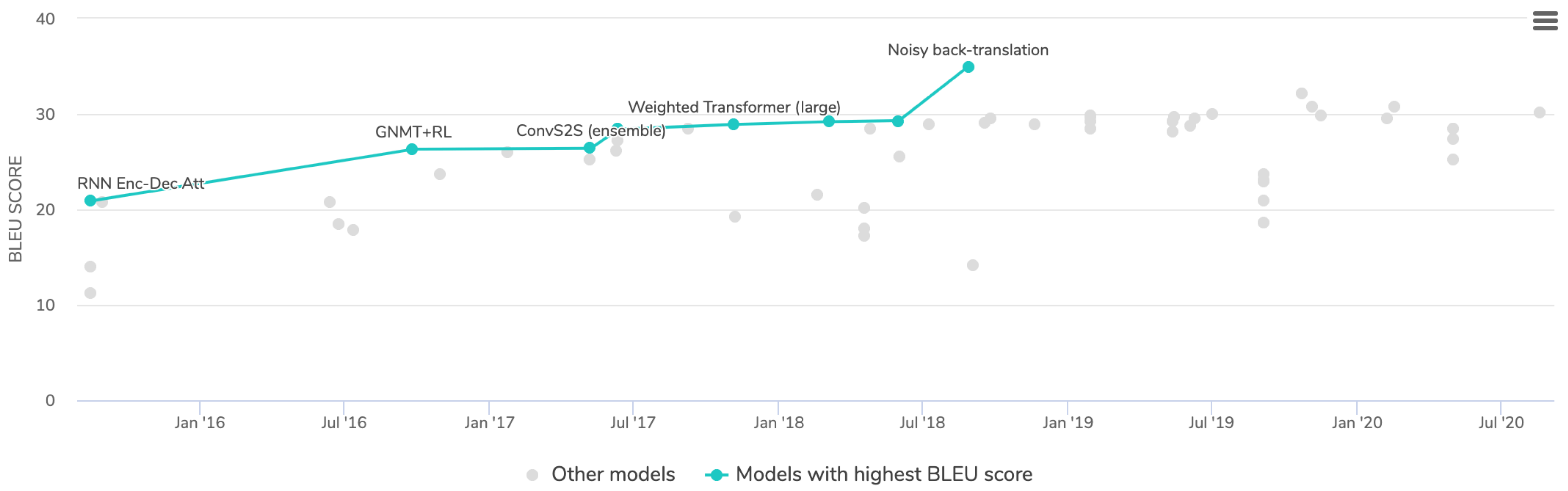



https://paperswithcode.com/sota

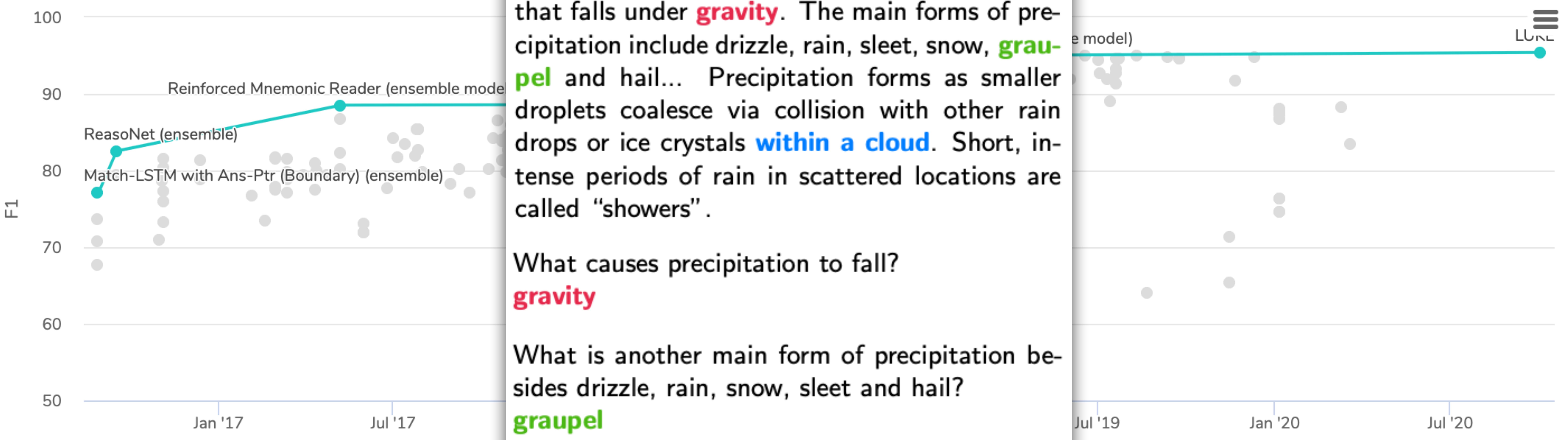
# Semantic Segmentation (Cityscapes)



# Machine Translation (WMT EN-DE)



# Question Answering (SQuAD 1.1)



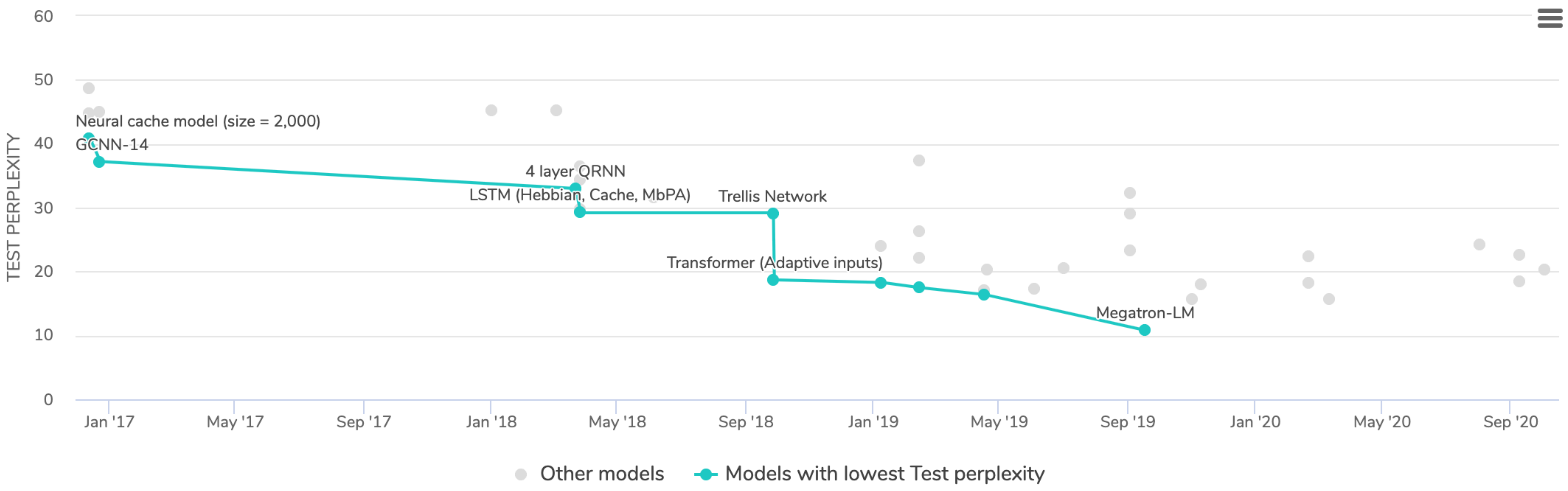

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or 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 | 2010 - 2020 |
|---|---|
| Empirical progress usually goes hand in hand with theoretical results | Empirical progress usually comes without mathematical theory |
| Emphasis on **provable guarantees** | Emphasis on **benchmarks** |
| Optimization problems often **convex** | **Non-convexity** is fine |
| **No** specialized hardware | **Large-scale** purely experimental work |

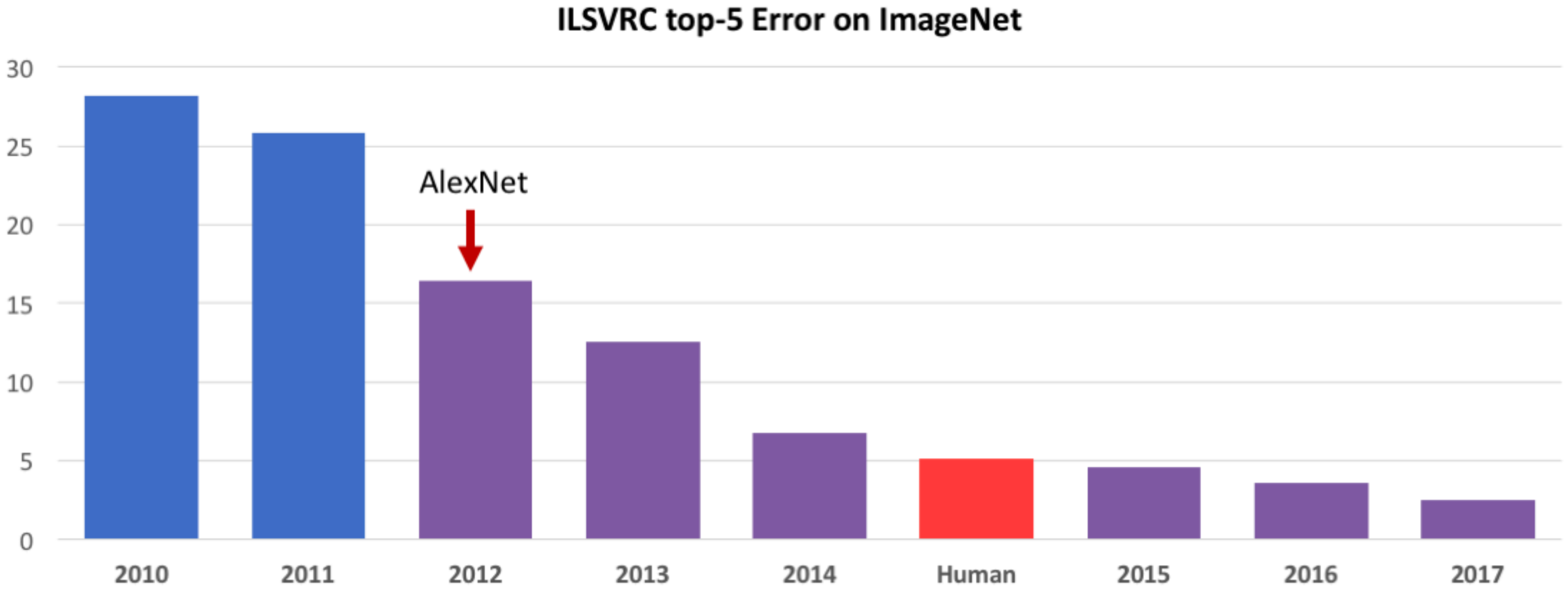
# A caveat with ML benchmarks

Excitement about experimental results, rapid growth in machine learning

But: even results on datasets like ImageNet remained controversial until about 2019.

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

# What are we Measuring with a Benchmark?



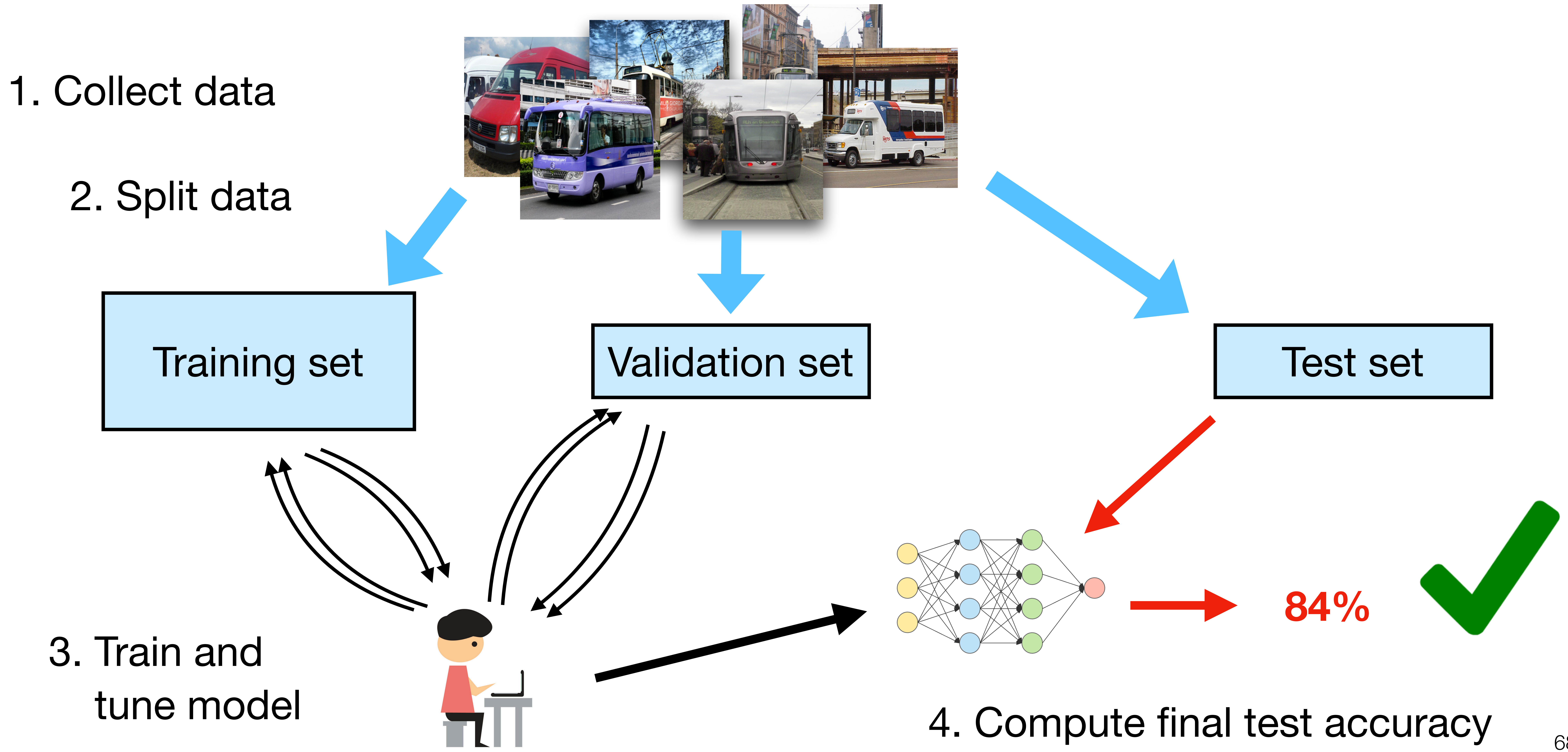

**There is nothing special about the 100k images in the ImageNet test set.**

→ What do we really care about?

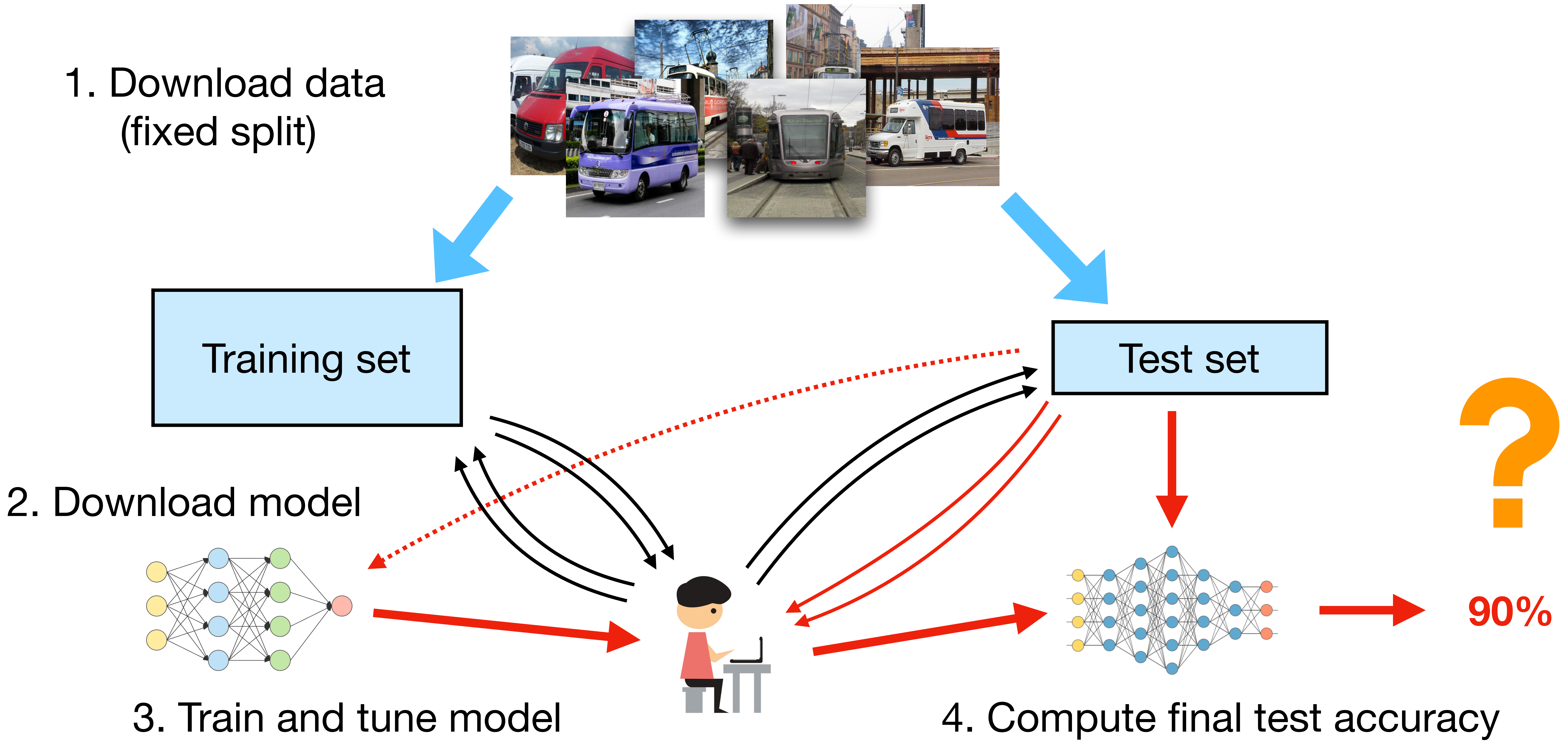
# Generalization

At least, the classifiers should perform similarly well on new data from the **same source**.

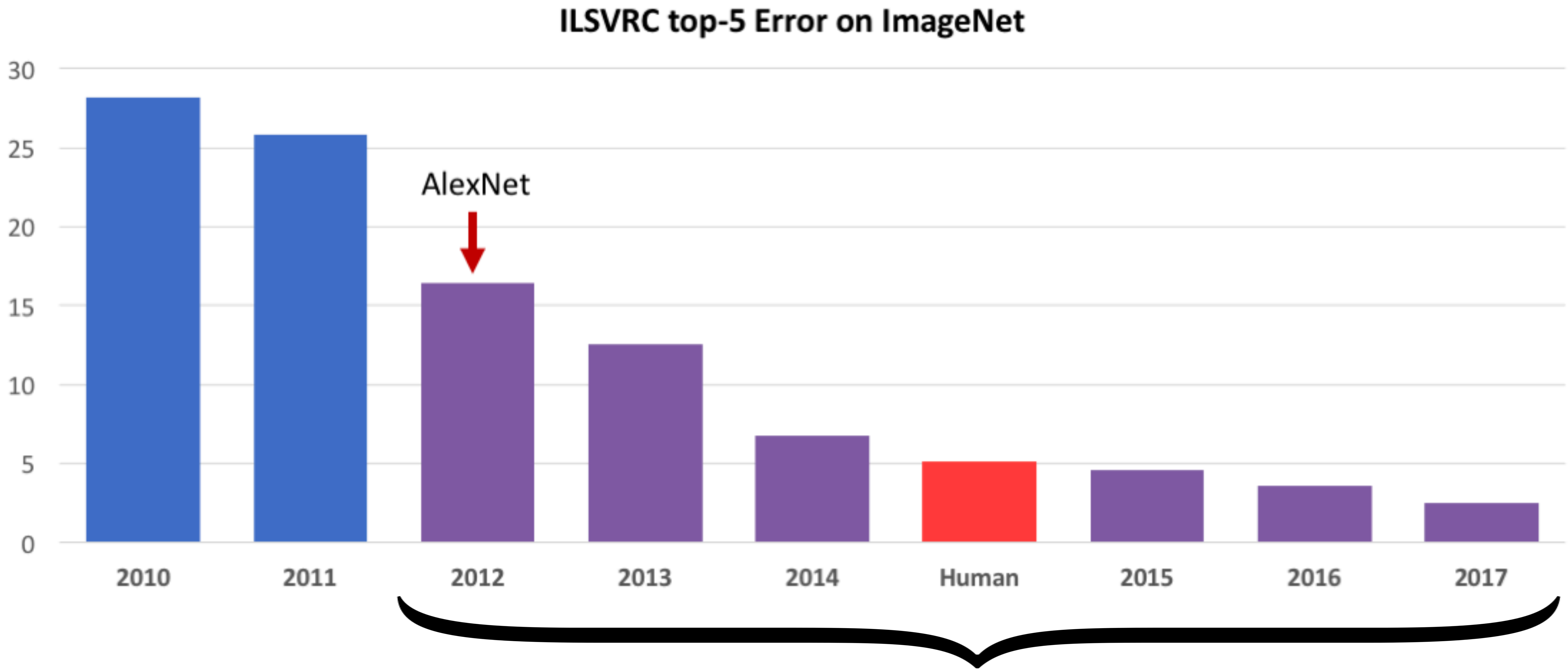
# Ideal ML Workflow



# Typical ML Workflow



# Real Cause for Concern



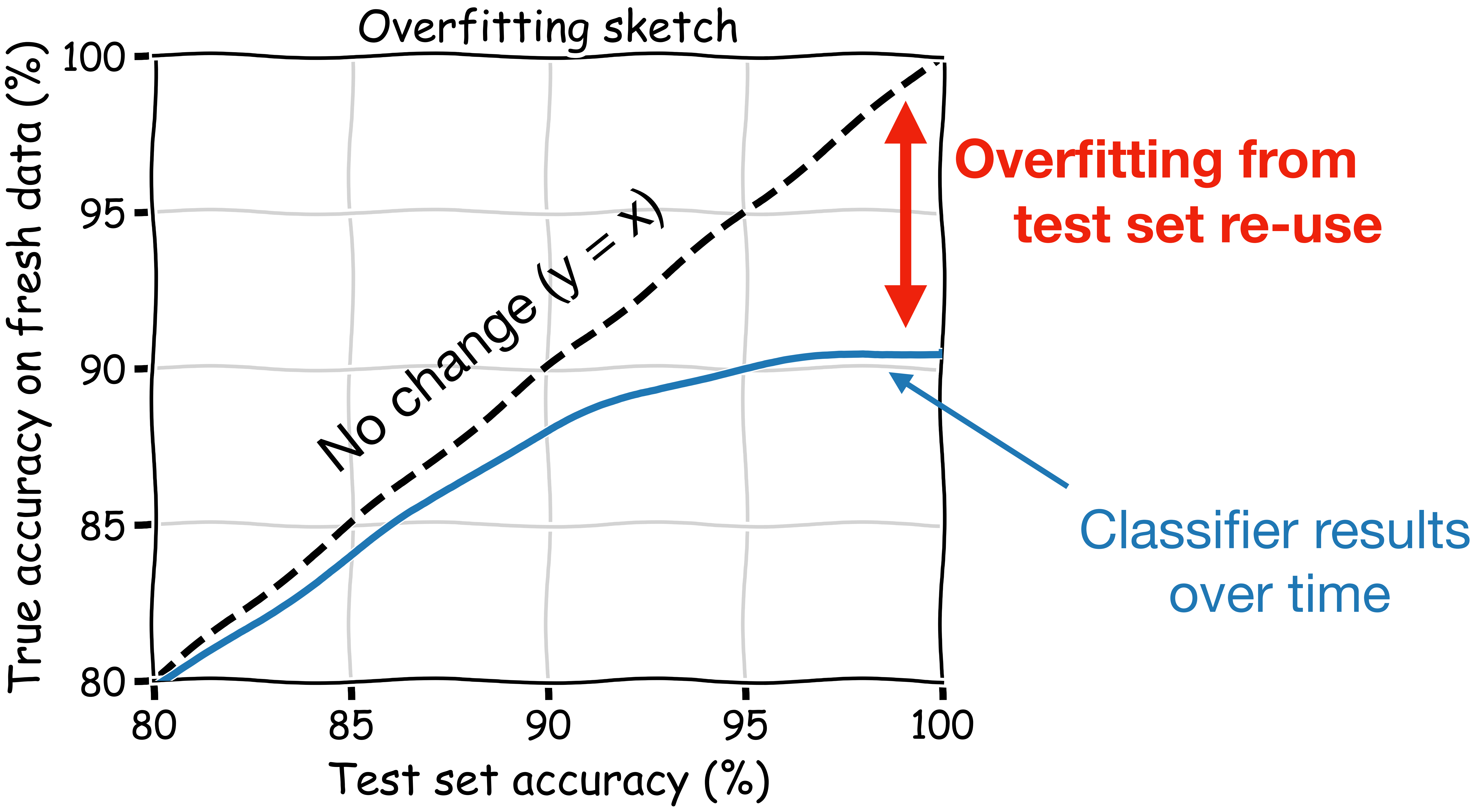

All the same test set!

Also true for **CIFAR-10**: fixed, public train / test split since 2008.

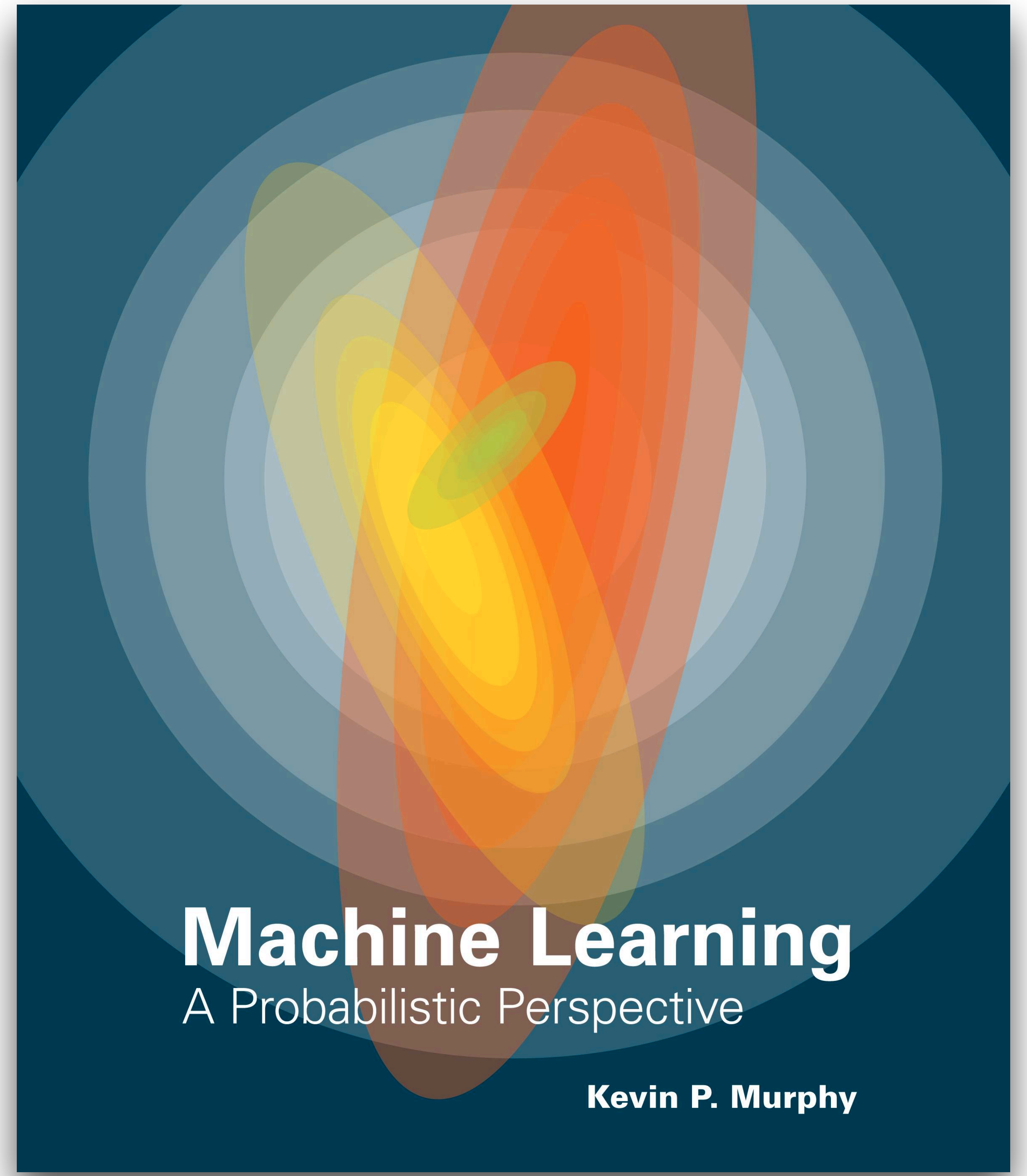
→ Numbers looked good, but there was uncertainty around them.

# Danger with Test Set Re-Use: Overfitting

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



# 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

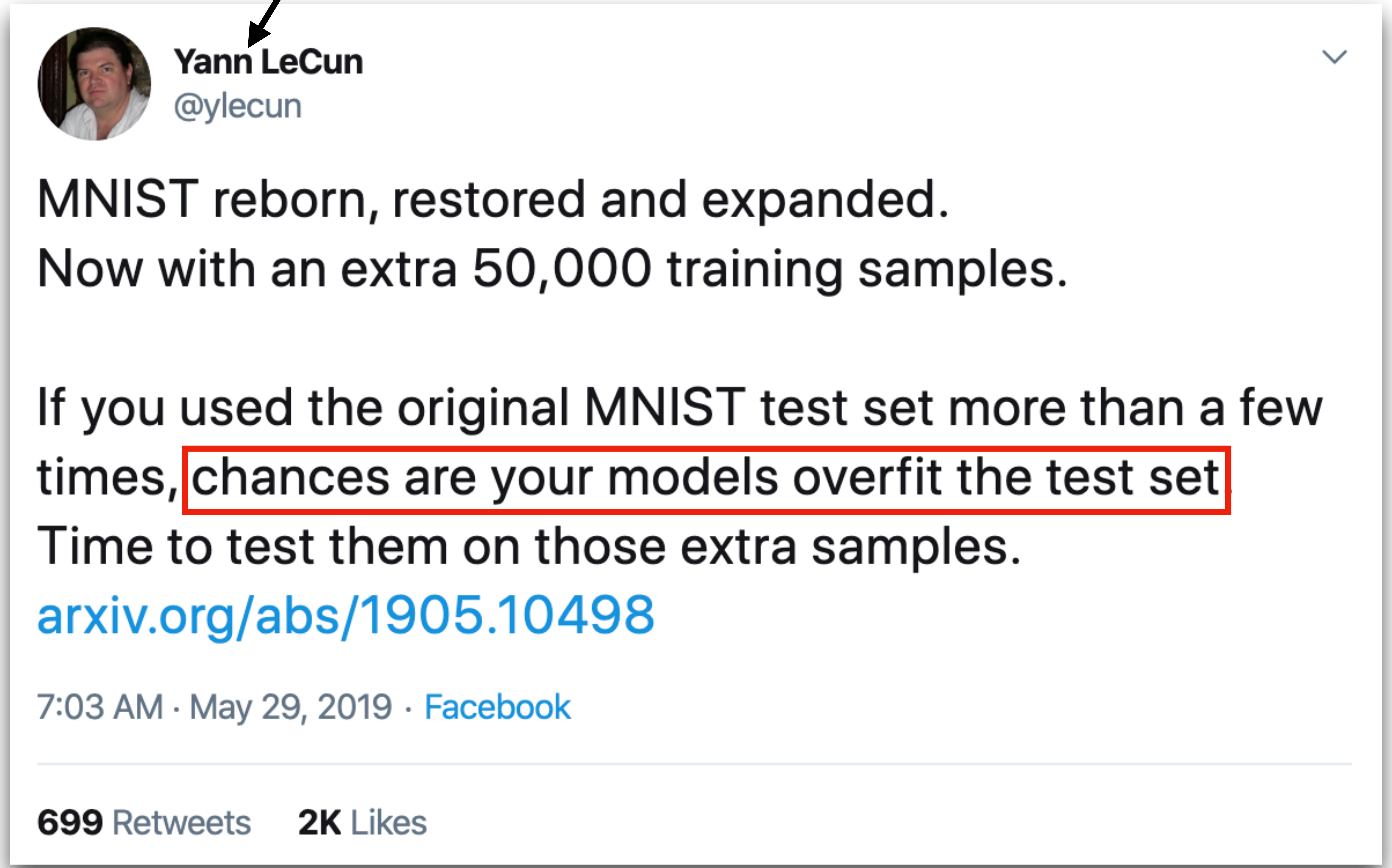
*"Withholding the annotation of the test data until completion of the challenge played a significant part in **preventing over-fitting** of the parameters of classification or detection methods. In the VOC2005 challenge, test annotation was released and this led to some **"optimistic" reported results, 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



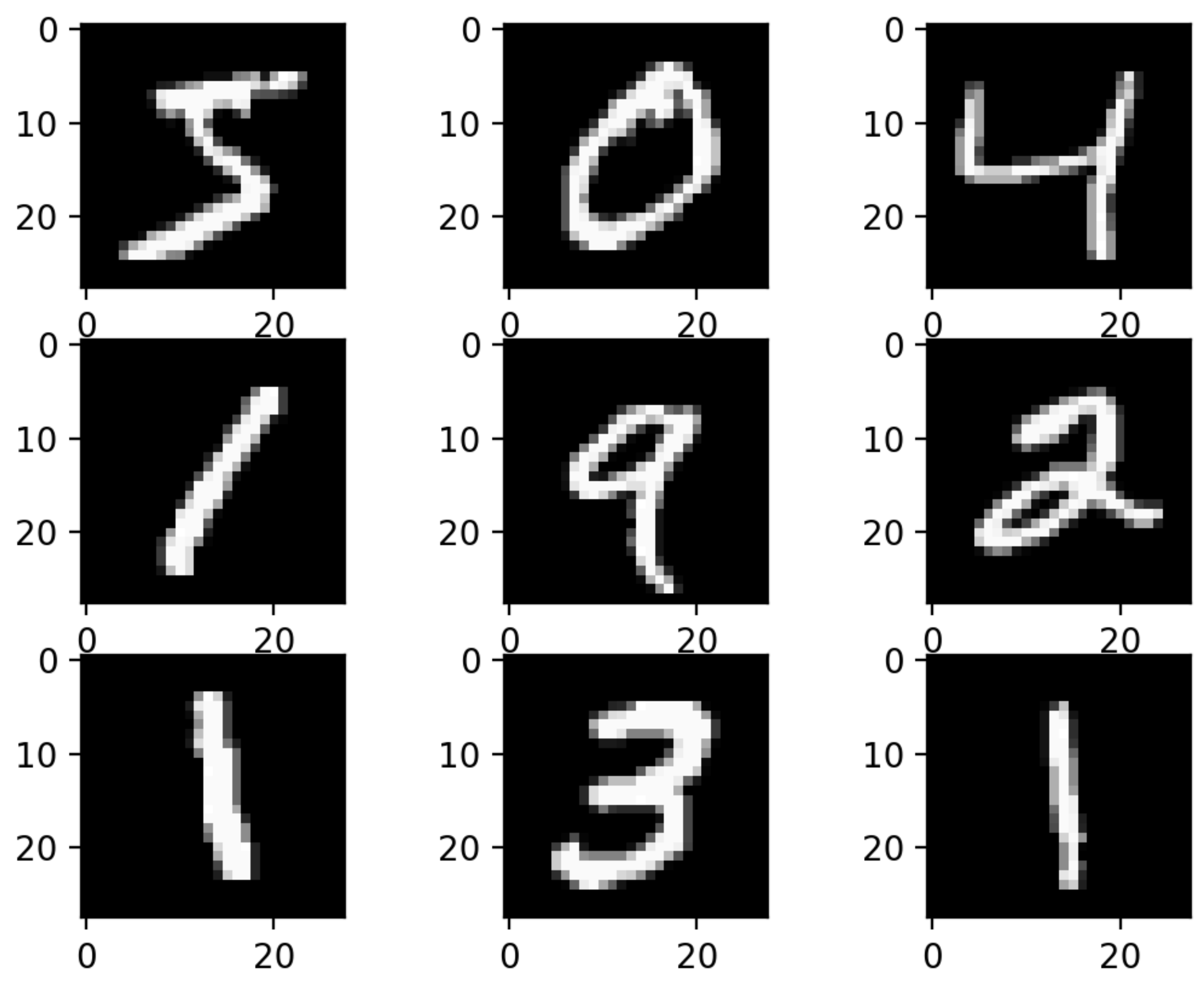
Yann LeCun
@ylecun

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

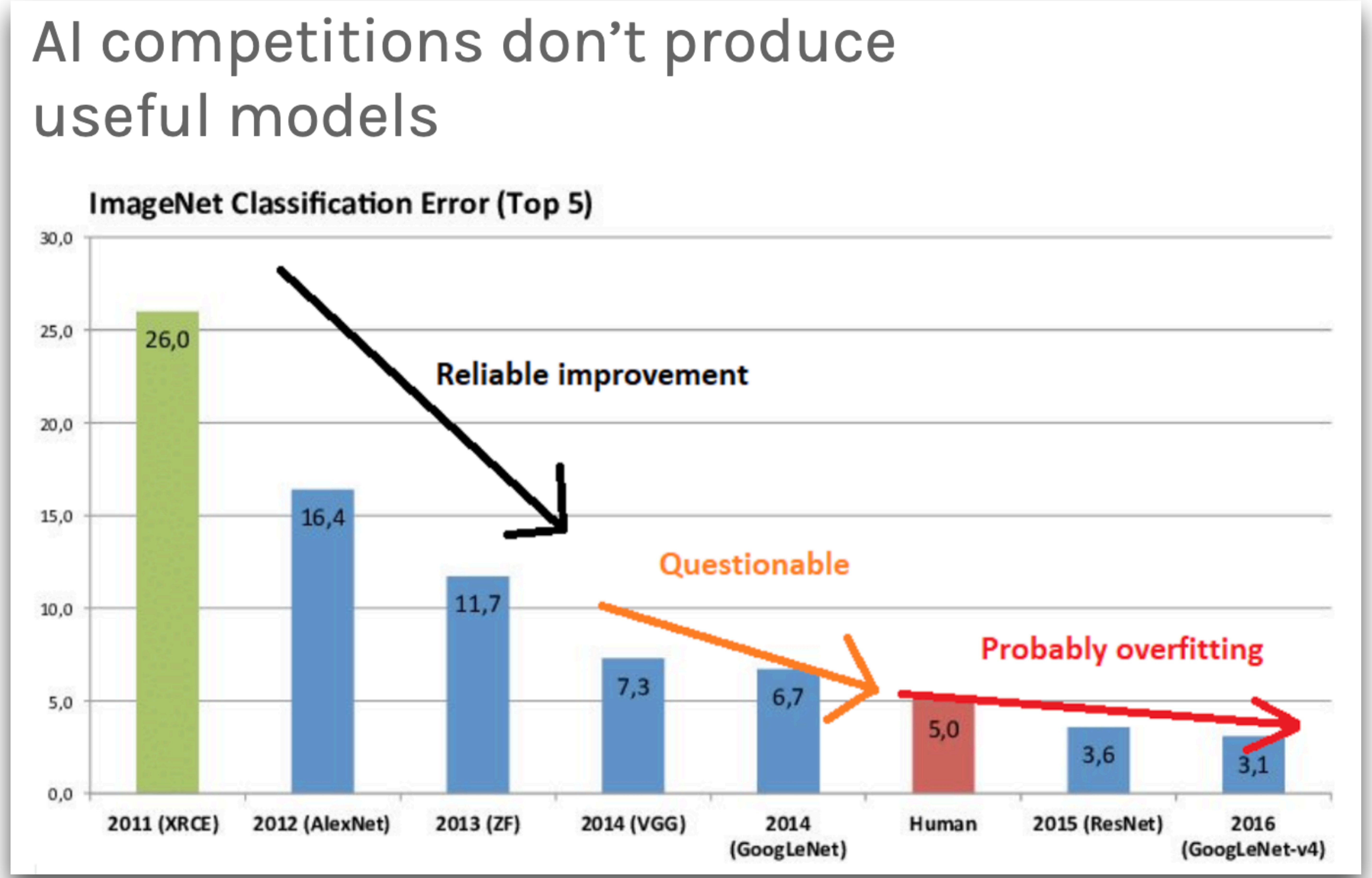
10 classes

Released in 1998

Oldest widely used dataset

Now considered “easy”

https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-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?*

# Testing 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**

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the 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.

# Generalization

At least, the classifiers should perform similarly well on new data from the **same source**.


Our experiment: sample a new ImageNet test set *nearly* i.i.d.

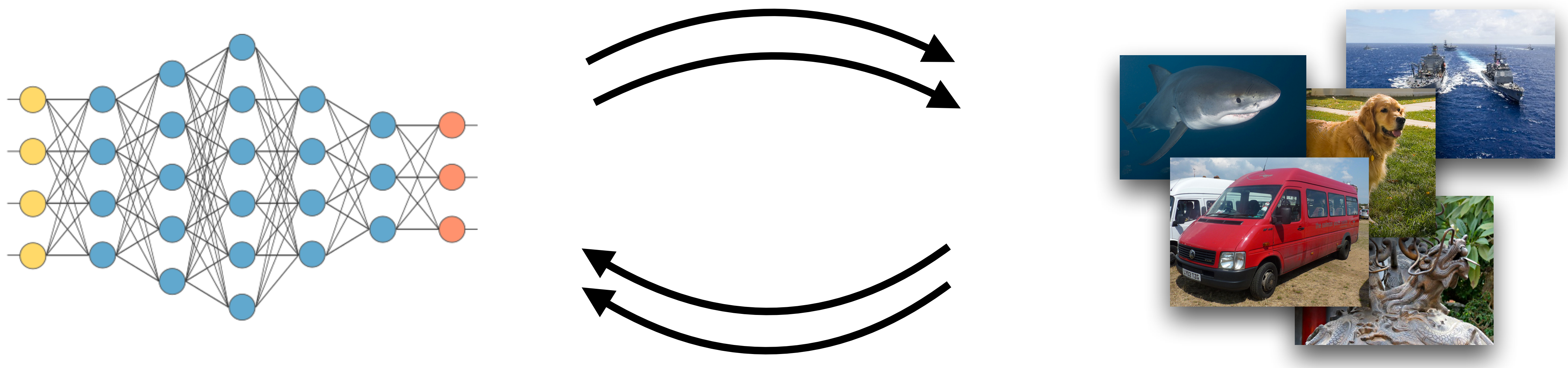
# Overfitting

# ?

# Three Forms of Overfitting

1. Test error ≥ training error

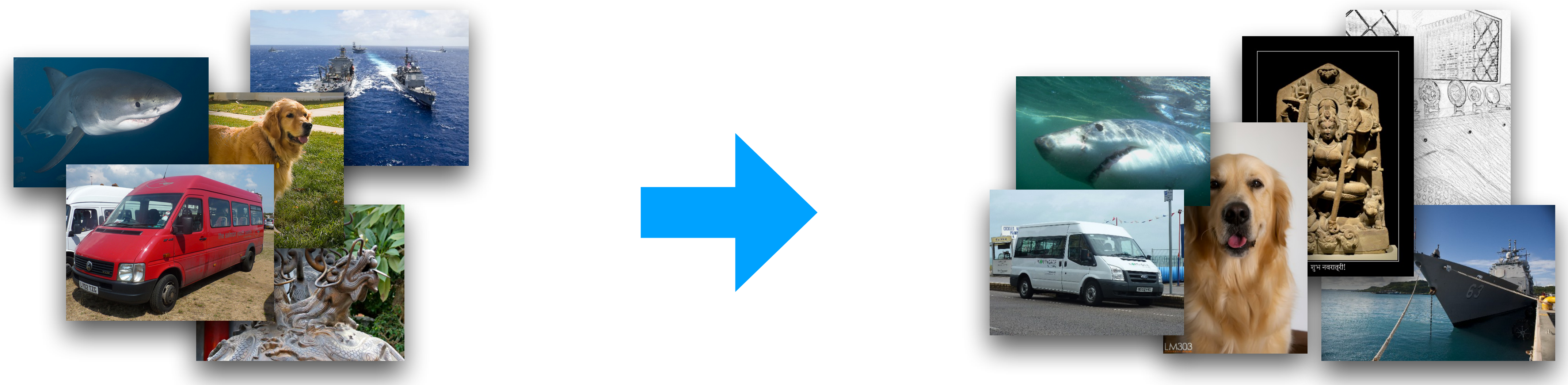
2. Overfitting through test set re-use



Model

Test Set

3. Distribution shift



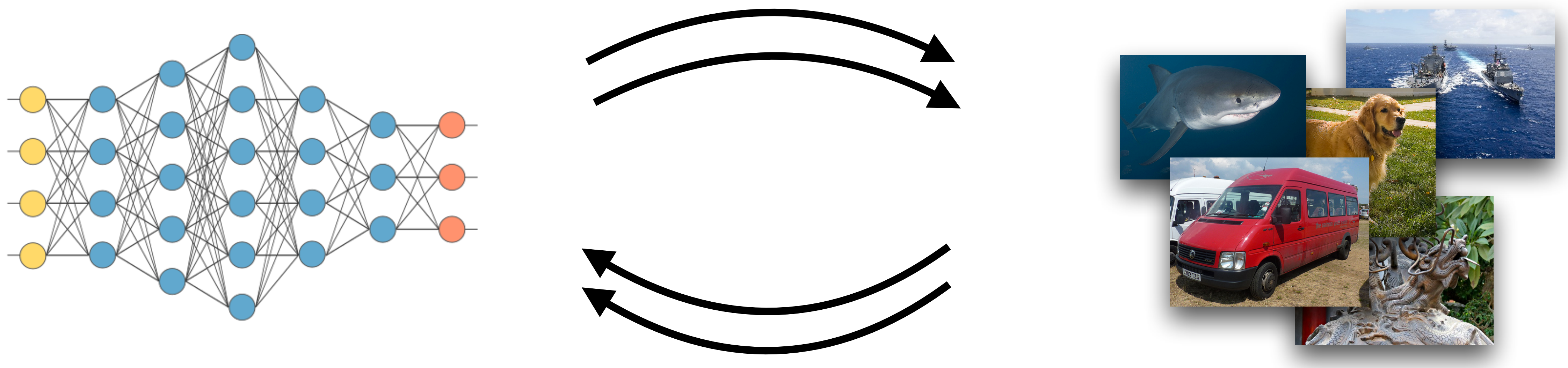
Original Test Set

New Test Set

# Three Forms of Overfitting

1. Test error ≥ training error

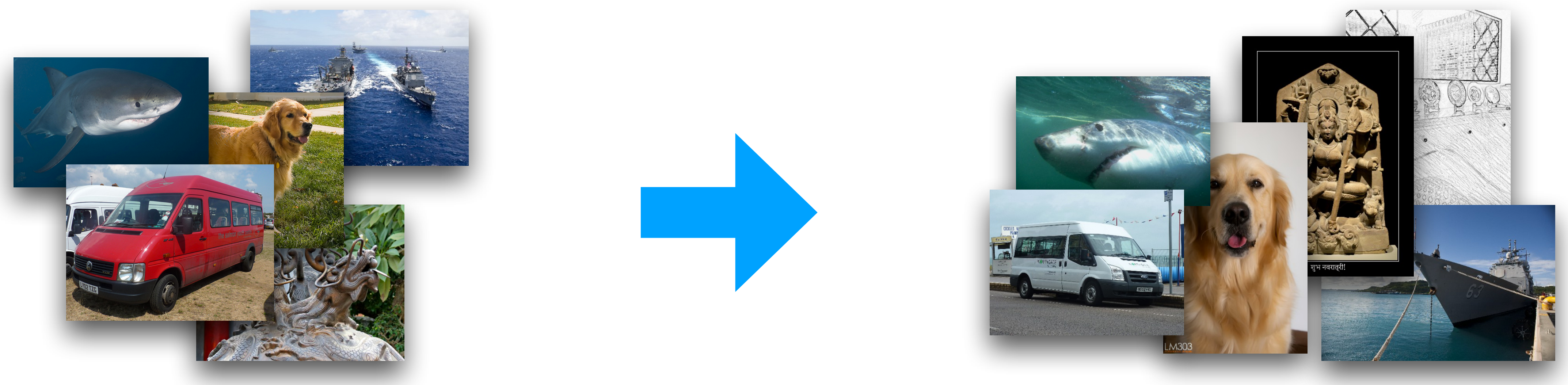
2. Overfitting through test set re-use



Model

Test Set

3. Distribution shift



Original Test Set

New Test Set

# Two Possible Causes

New test accuracy

Overfitting through test set re-use

Distribution shift

$$\underbrace{\widehat{\mathrm{acc}}_S(f) - \widehat{\mathrm{acc}}_{S'}(f)}_{\approx 11\%} =$$

Generalization error (≈ 1%)

Original test accuracy (orig. test set S, new S')

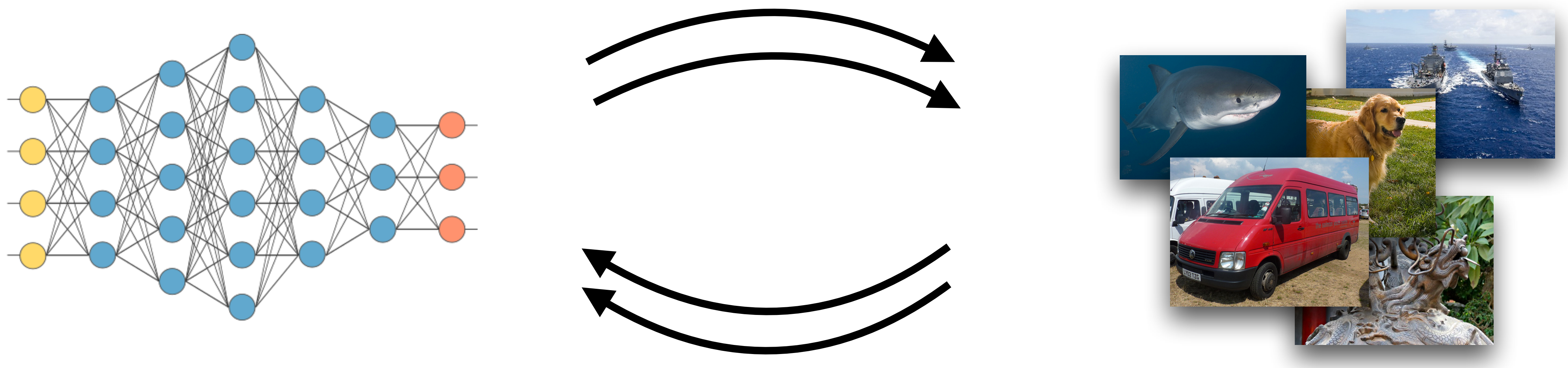
$$\widehat{\mathrm{acc}}_S(f) = \frac{1}{|S|} \sum_{(x,y)\in S} \mathbb{1}[f(x) = y]$$

$$\mathrm{acc}_D(f) = \mathbb{E}_{(x,y)\sim D}\, \mathbb{1}[f(x) = y]$$ (S is drawn from D)

# Three Forms of Overfitting

1. Test error ≥ training error

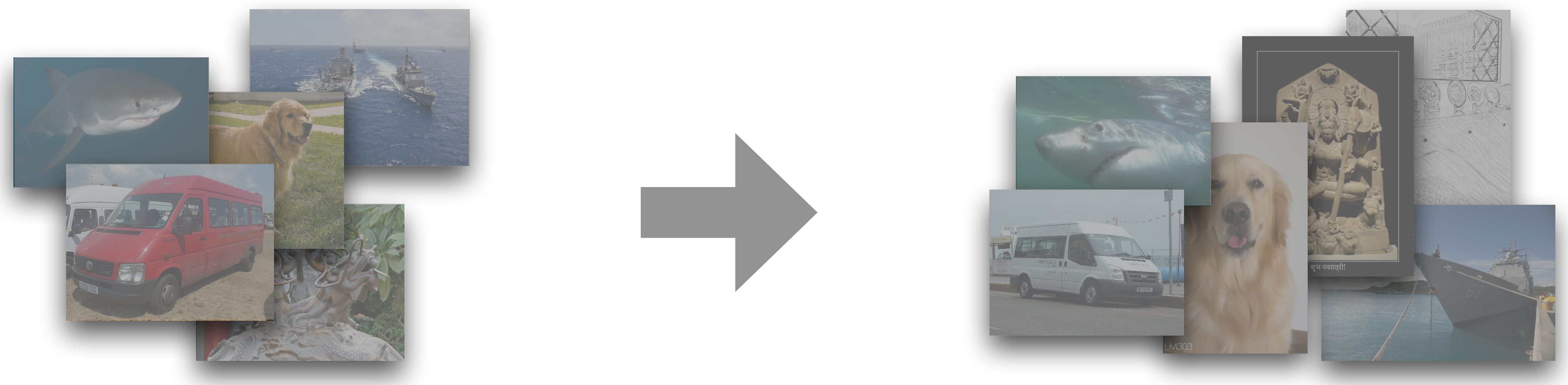
**2. Overfitting through test set re-use**



Model

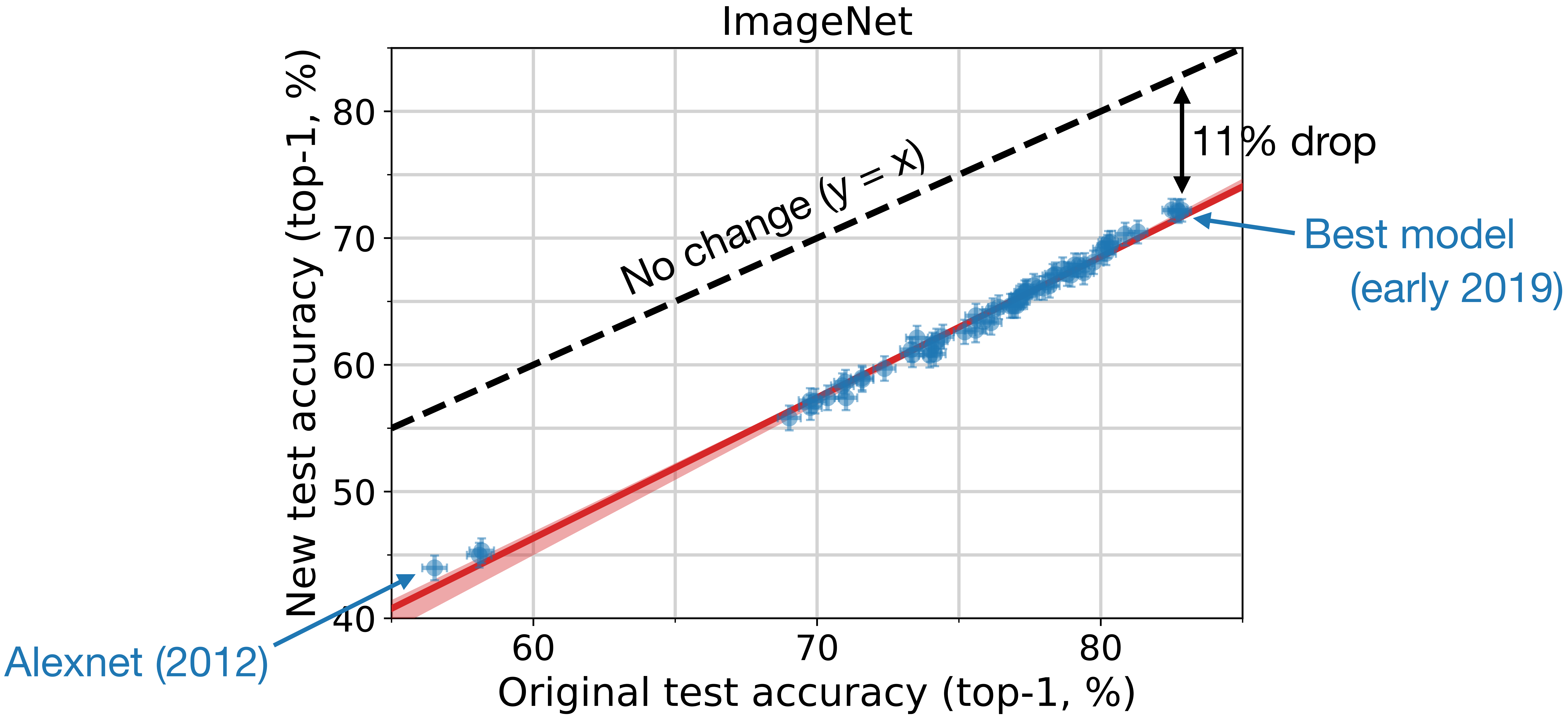
Test Set

3. Distribution shift

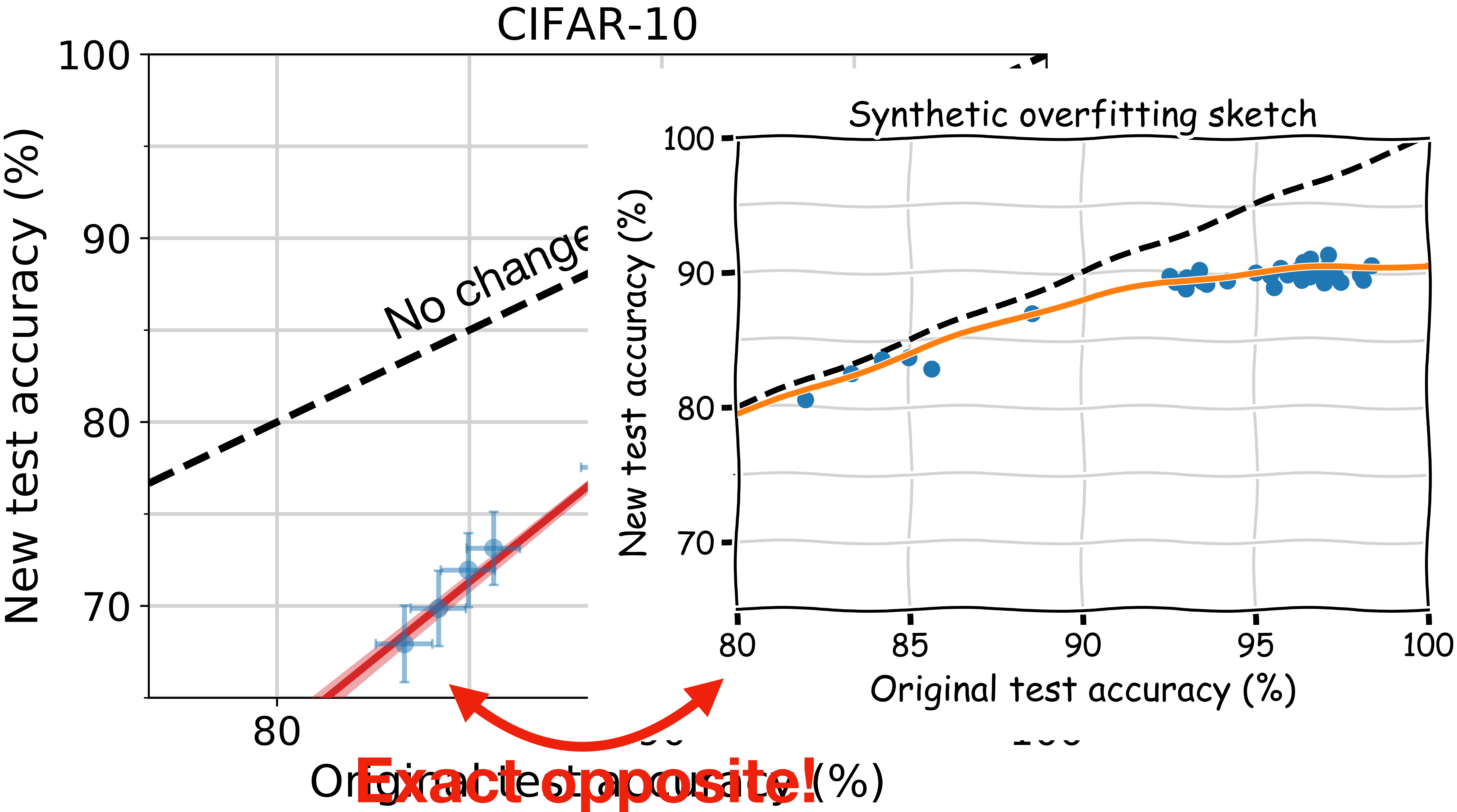


Original Test Set

New Test Set




- The best models on the original test set stay the best models on the new test set.
- All models see a substantial drop in accuracy.




Later models see a **smaller** drop in accuracy.

AutoAugment vs. ResNet: 4.9% difference on CIFAR-10

AutoAugment vs. ResNet: 10.3% difference on CIFAR-10.1

# Overfitting Is Surprisingly Absent

No overfitting despite 10 years of test set re-use on CIFAR-10 and ImageNet.

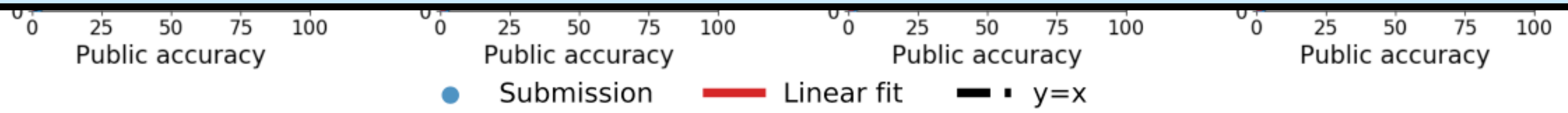
→ Relative ordering preserved. Progress is real!

**MNIST:** similar conclusions in [Yadav, Bottou'19]

no overfitting after 20+ years of MNIST

**Kaggle:** Meta-analysis of 120 ML competitions [Roelofs, Fridovich-Keil, Miller, Shankar, Hardt, Recht, **Schmidt** '19]

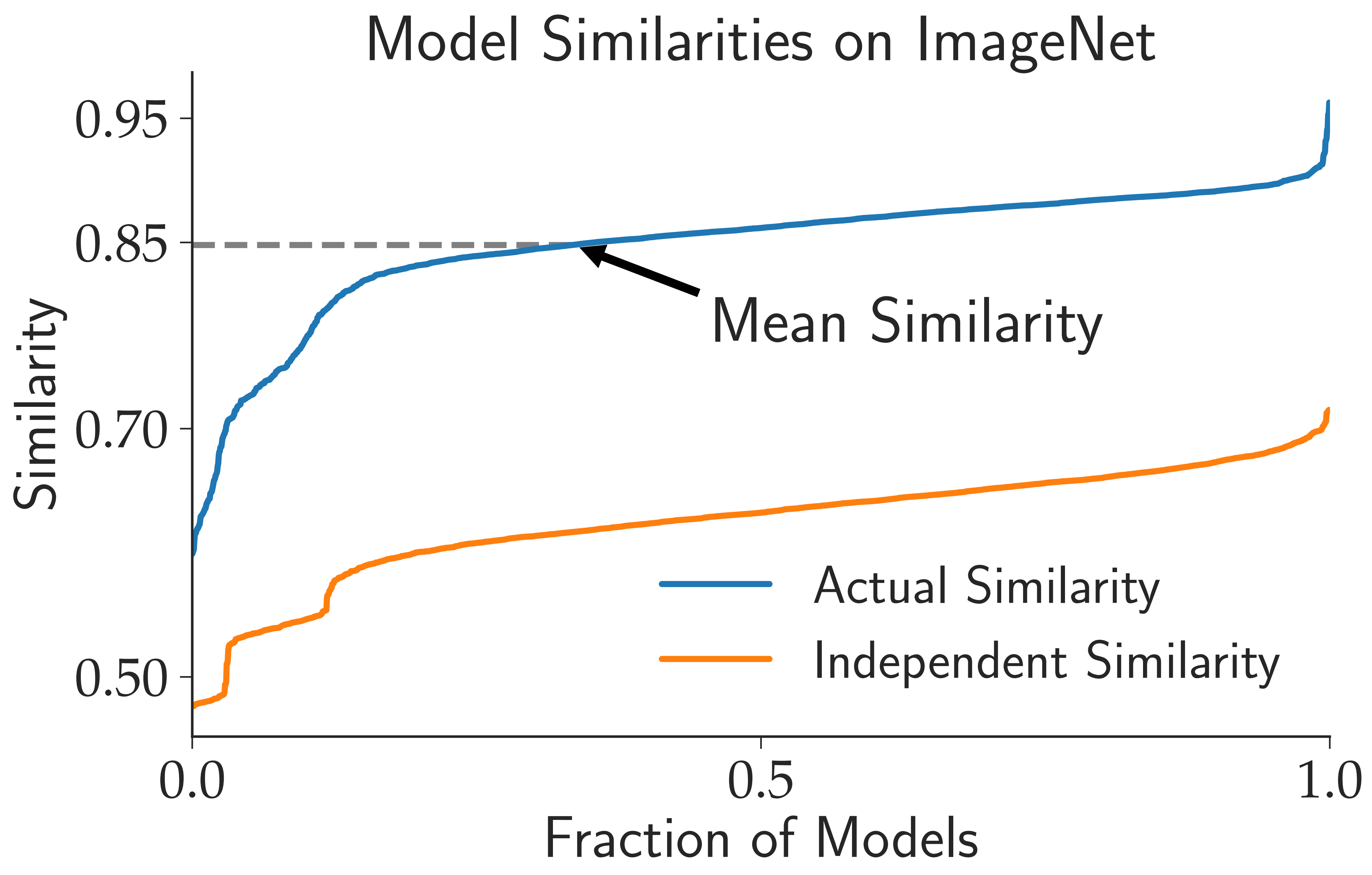
*Our results unambiguously confirm the trends observed by Recht et al. [2018, 2019]: although the misclassification rates are slightly off, classifier ordering and model selection remain broadly reliable.*

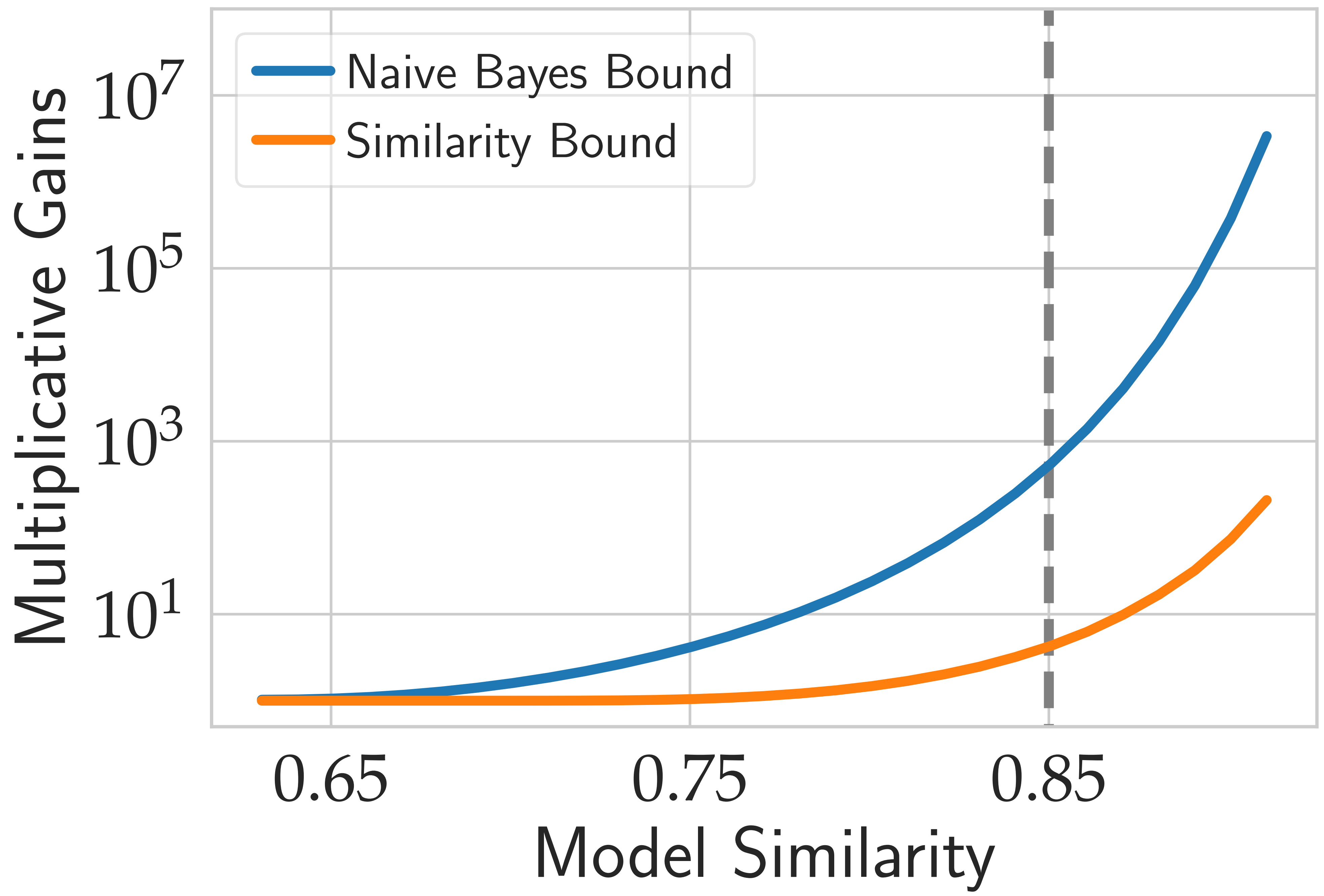


# Why Does Test Set Re-use Not Lead to Overfitting?

One mechanism: model similarity mitigates test set re-use.
[Mania, Miller, **Schmidt**, Hardt, Recht'19]

Similarity of two models $f_i$ and $f_j$: agreement of 0-1 loss on the data distribution.




Likely only a partial explanation (see Moritz Hardt's keynote at COLT 2019).

# Two Possible Causes

New test accuracy

Overfitting through test set re-use **(≈ 0%)**

Distribution shift

$$\underbrace{\widehat{\text{acc}}_S(f) - \widehat{\text{acc}}_{S'}(f)}_{\approx 11\%} = \cancel{\widehat{\text{acc}}_S(f) - \text{acc}_D(f)}$$
$$+ \text{acc}_D(f) - \text{acc}_{D'}(f)$$
$$+ \text{acc}_{D'}(f) - \widehat{\text{acc}}_{S'}(f)$$

Original test accuracy (orig. test set S, new S')

Generalization error (≈ 1%)

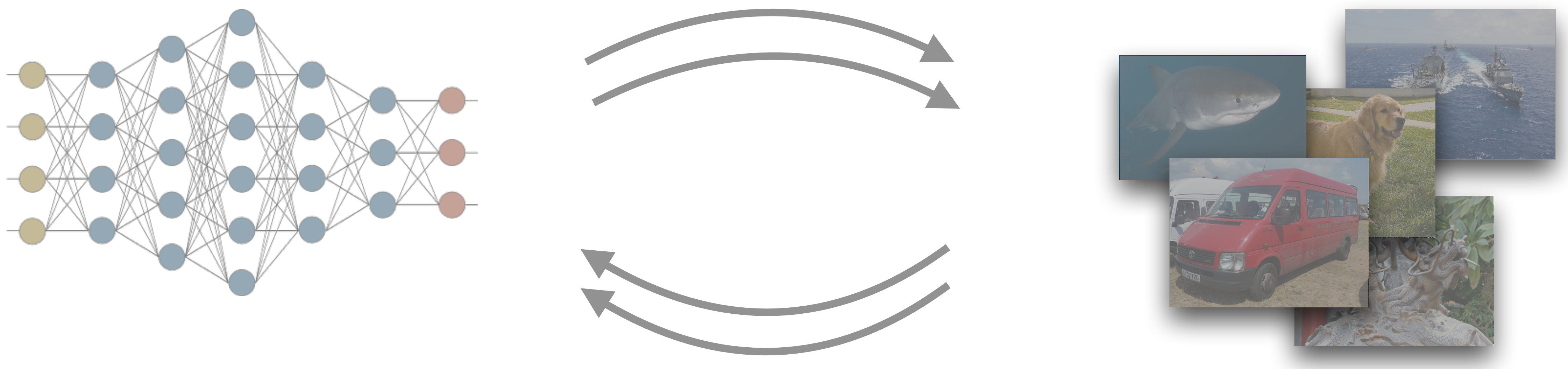
$$\widehat{\text{acc}}_S(f) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}[f(x) = y]$$

$$\text{acc}_D(f) = \mathbb{E}_{(x,y)\sim D}\, \mathbb{1}[f(x) = y]$$ (S is drawn from D)

# Three Forms of Overfitting

1. Test error ≥ training error

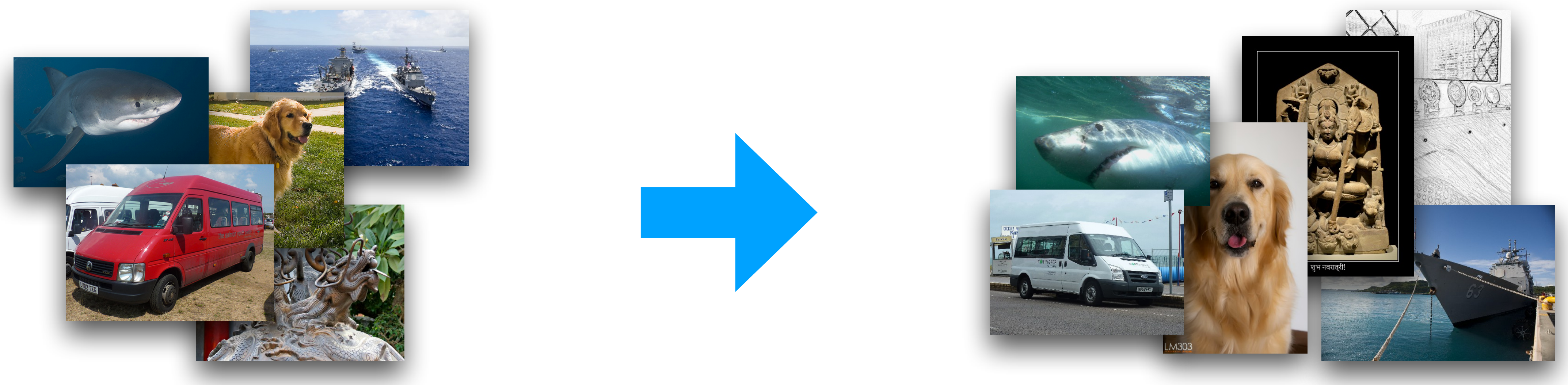
2. Overfitting through test set re-use



Model

Test Set

**3. Distribution shift**



Original Test Set

New Test Set

# ImageNet Creation Process

**Detailed description** in [Deng, Dong, Socher, Li, Li, Fei-Fei'09]:

1. Find relevant search keywords for each class from **WordNet** (e.g., "goldfish", "Carassius auratus" for wnid "n01443537")

2. Search for images on **Flickr**

3. Show images to **MTurk** workers

Likely source of distribution shift

4. Sample a class-balanced dataset

WORDNET

+ flickr

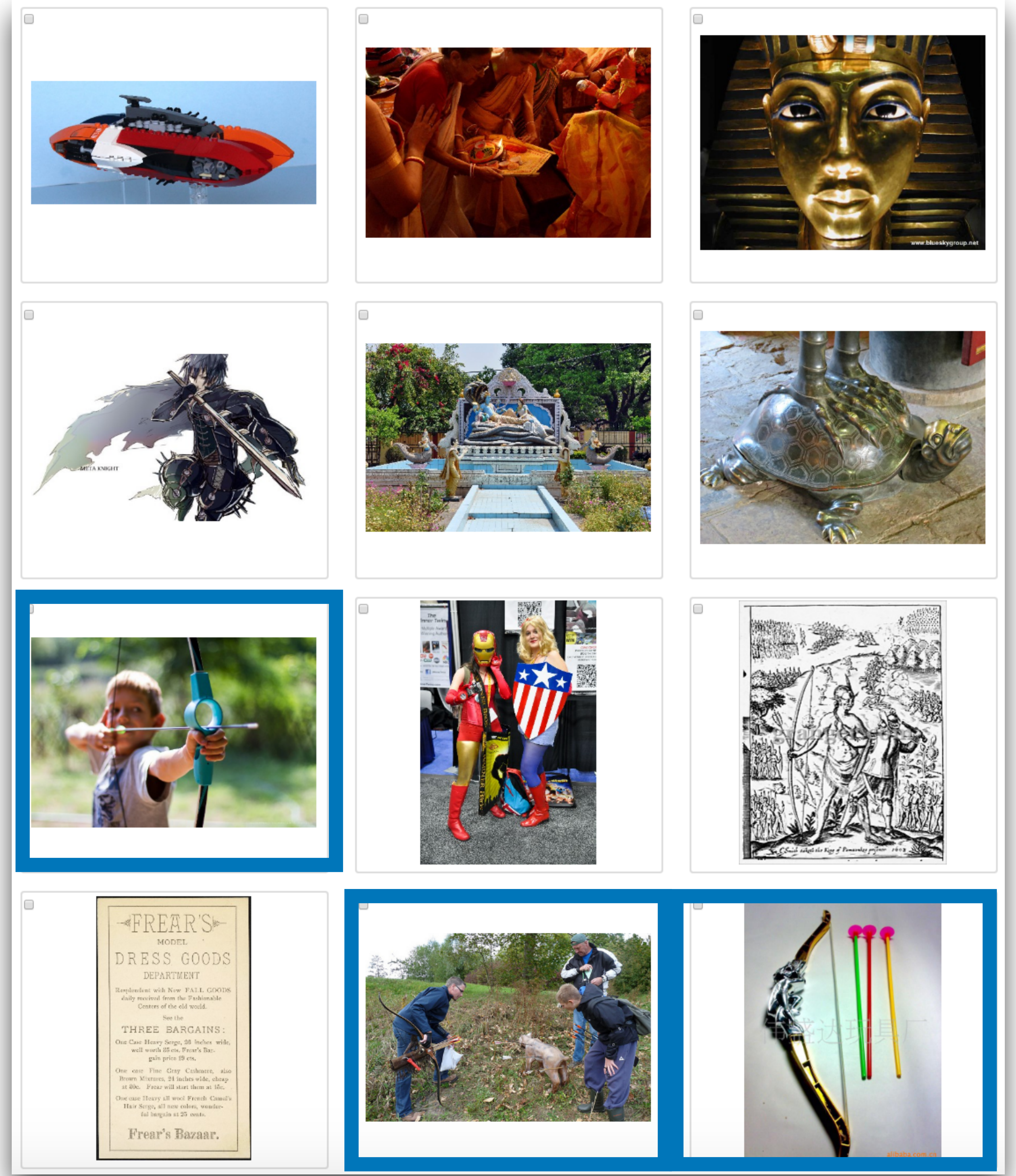
+ amazon mechanical turk beta

IMAGENET

**We replicated this process as closely as possible.**

# Data Cleaning With MTurk

**Instructions:** Select all images containing a bow.

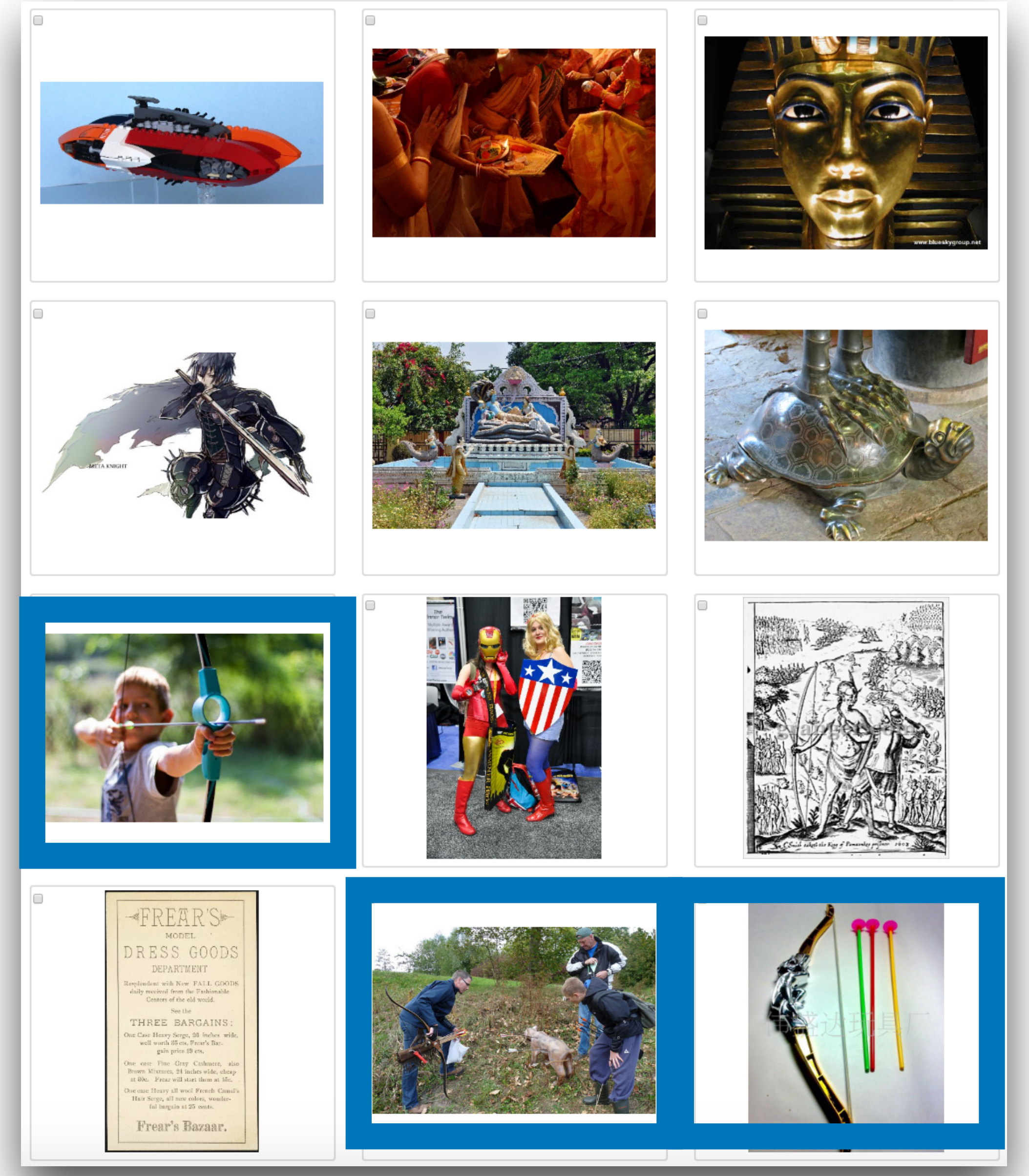


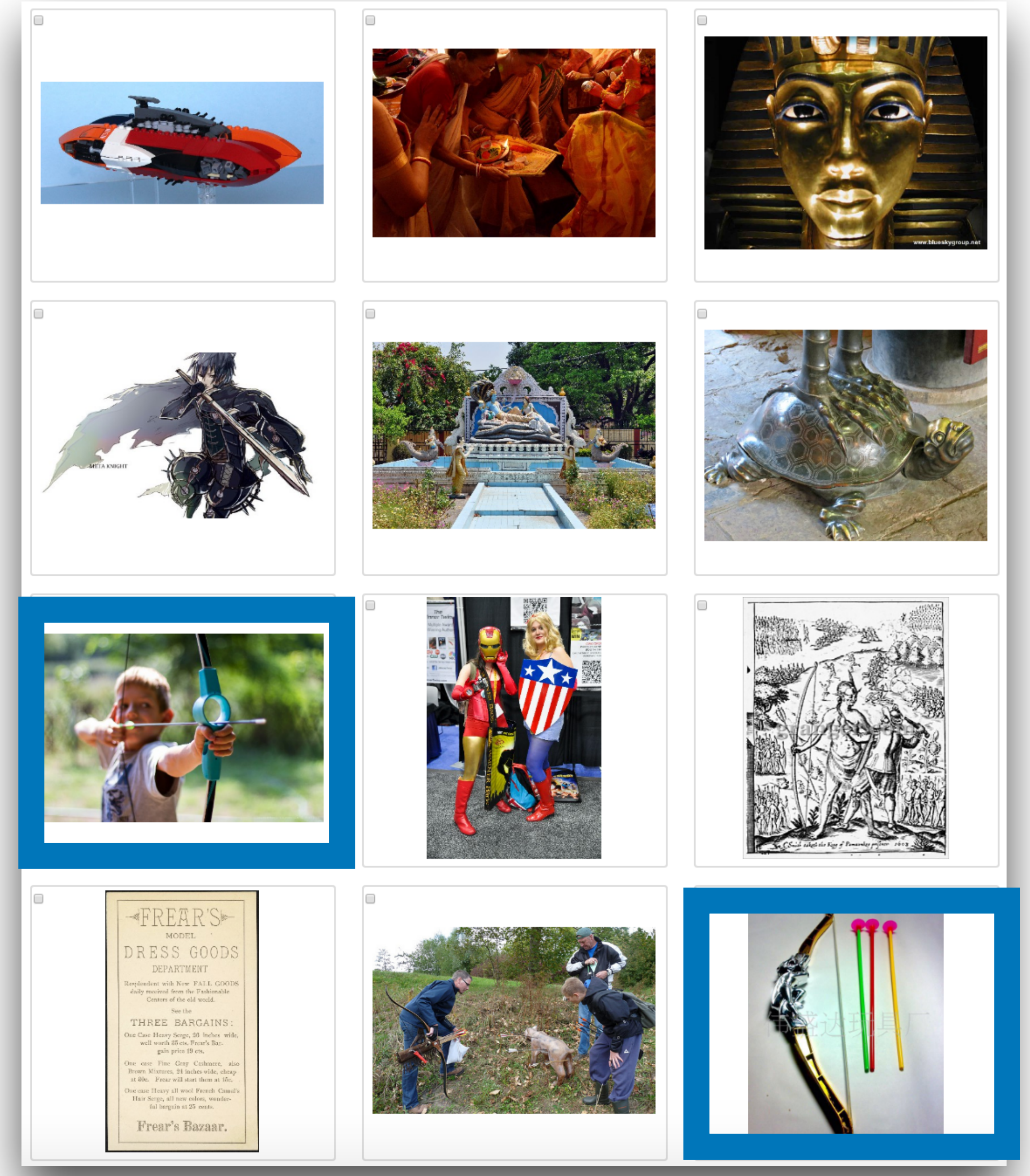
# Data Cleaning With MTurk

Worker 1

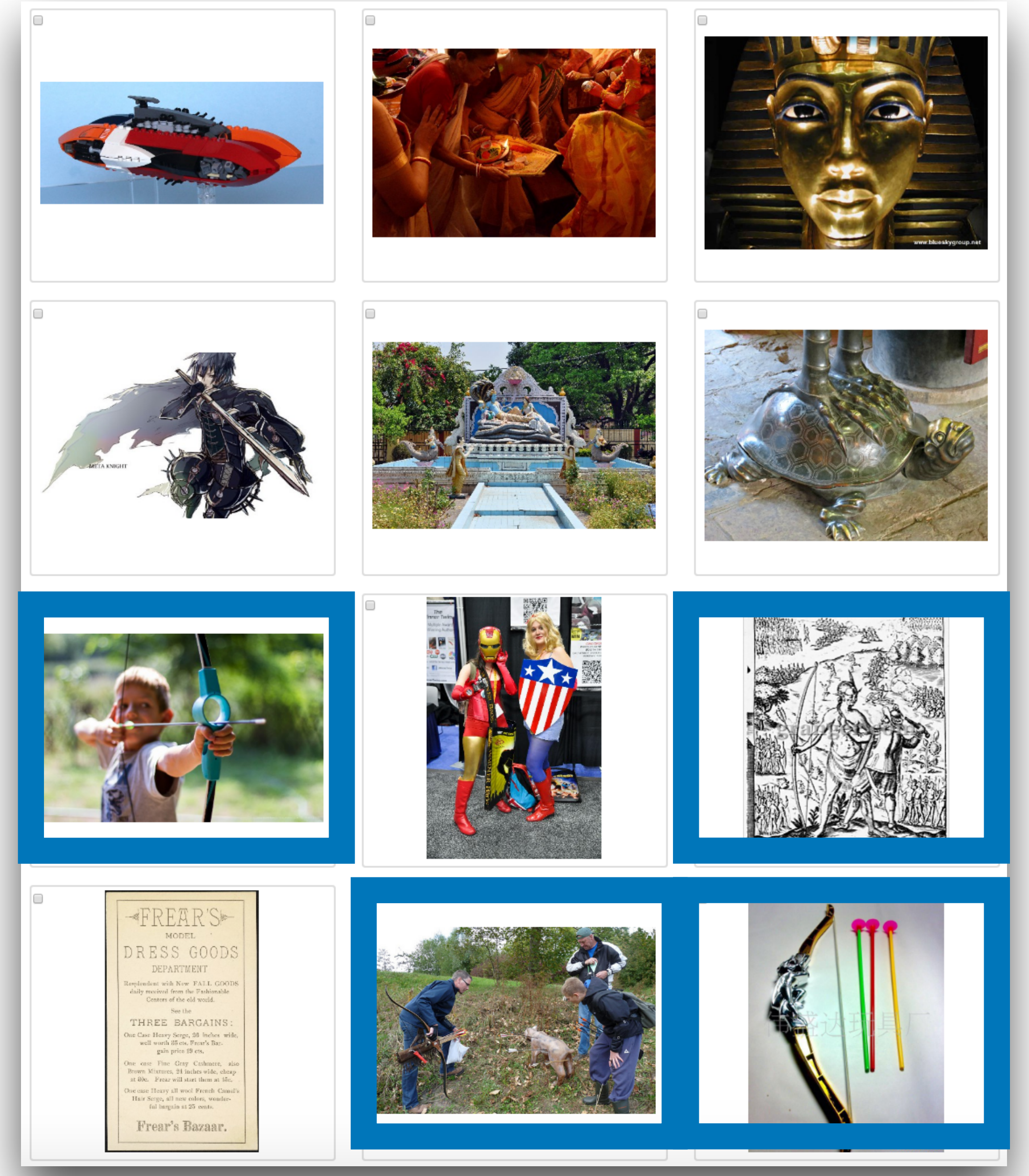
Worker 2

Worker 10

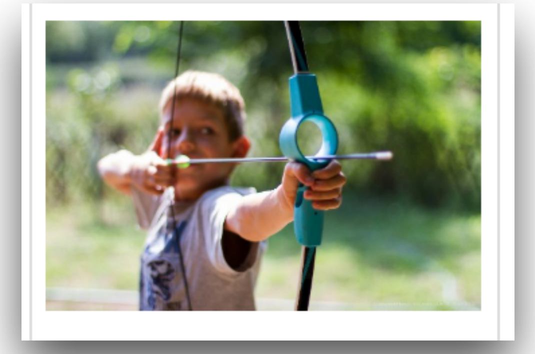




...



Main quantity: **selection frequency** $= \dfrac{\textbf{Number of workers who selected image } i}{\textbf{Number of workers who saw image } i}$

 : 1.0

 : 1.0

 : 0.67

 : 0.33

 : 0.0

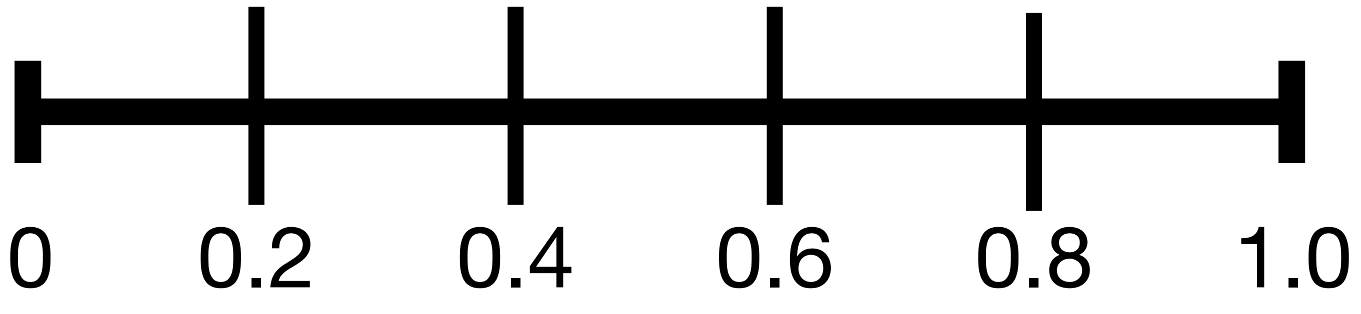
# Sampling Strategy for a New Test Set

**Input:** Selection frequencies from MTurk
(= fraction of workers selecting the image)

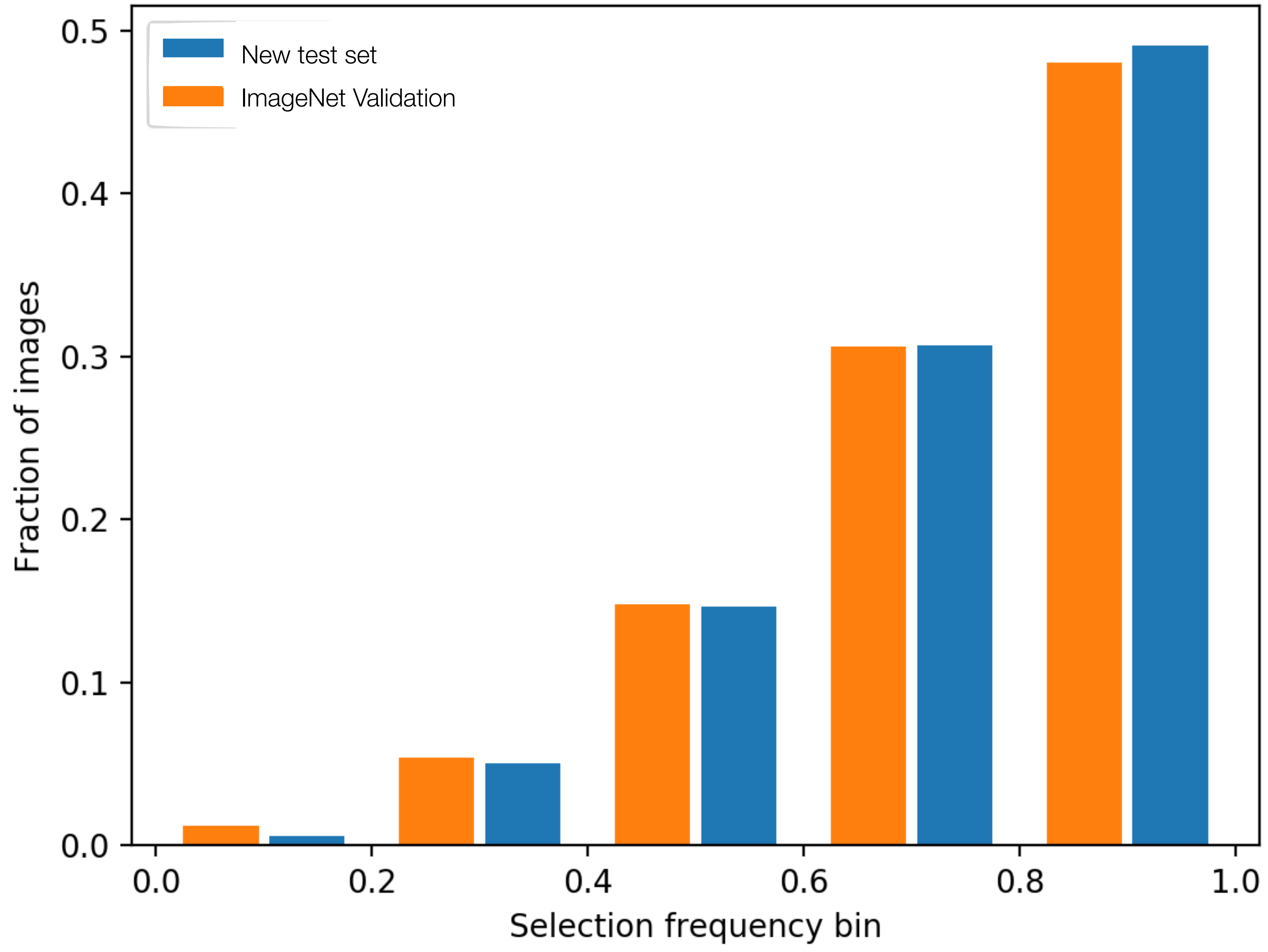
**Output:** representative & correct subset

**Our approach:**

1. Bin the existing validation images by selection frequency.




2. Sample images from our candidate pool to match the selection frequency distribution.



# Three New Test Sets

**ApproxCalibrated:** Selection frequencies comparable to the original test set (**0.71**).

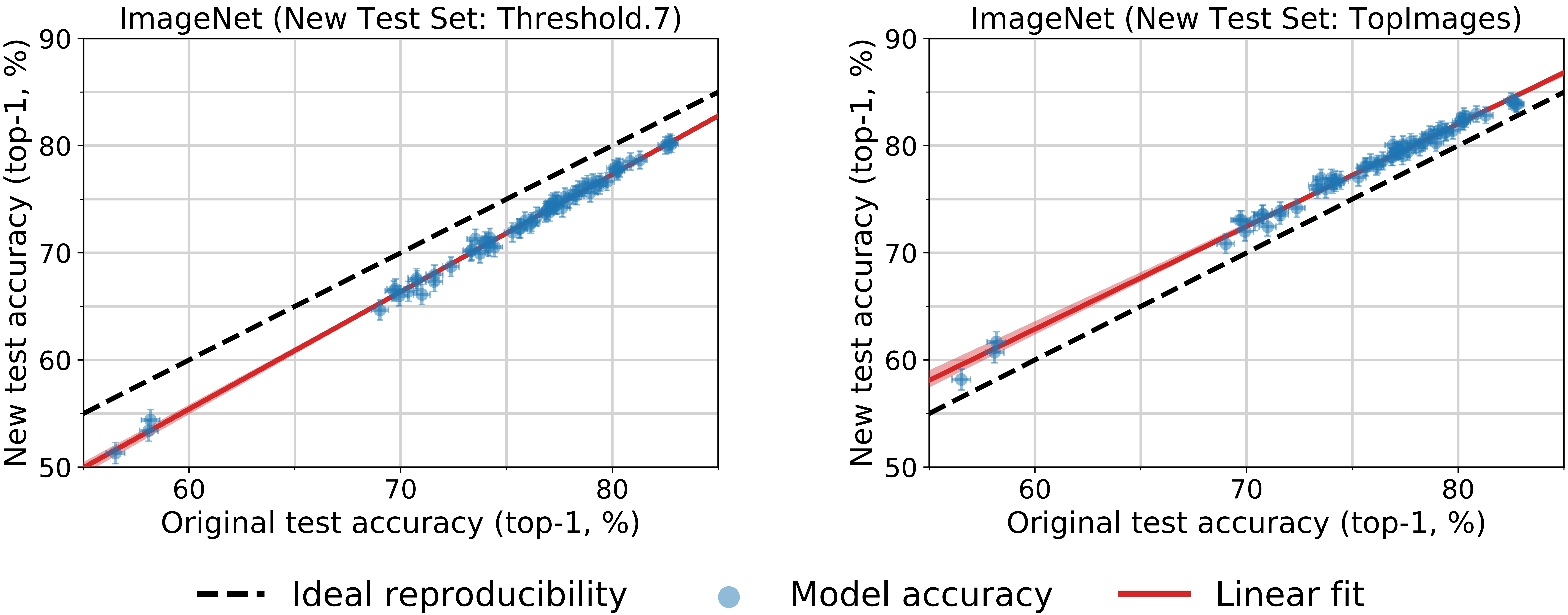
**Easier:** Different sampling strategy, higher selection frequencies.

**Easiest:** Highest selection frequencies in our candidate pool.

All correctly labeled!

| Test Set | Average MTurk Selection Frequency | Average Top-1 Accuracy Change |
|---|---|---|
| **ApproxCalibrated** | **0.73** | - 12% |

➡ Selection frequencies have large impact on classification accuracies.




Relative ordering is stable, absolute accuracies are brittle.

1. Logistics

2. Background & motivation

3. Course outline

# Course Outline

Two parts:

1. Theoretical foundations (7 lectures)

   Guiding principles: **generalization** and **empirical risk minimization**

   We will look at both **statistical** aspects (generalization bounds)

   and **algorithmic** aspects (optimization algorithms)

2. Empirical foundations (12 lectures)

   Goal: understand the ingredients for **large language models,** specifically GPT-3.

   Model architecture, language modeling, scaling laws, evaluations, efficiency, etc.

   **Also:** multimodal models, fine-tuning (RLHF), datasets, generative models.

Thanks!

Questions?