Do ImageNet Classifiers Generalize to ImageNet?

Guest Lecture in B9145: Reliable Statistical Learning

Ludwig Schmidt

UC Berkeley → Toyota Research → UW

One Theoretician's Perspective on Empirical ML

Goals for today:

1. Get an overview of **progress on the empirical side** of machine learning.
2. Understand how the **benchmarking paradigm** creates reliable empirical knowledge about machine learning.
3. Identify **limitations** of current machine learning methods.
4. Learn to **connect** theoretical & empirical perspectives and discuss the role of theory in contemporary machine learning.

Different flavor compared to previous lectures: focus on **experiments.**

Please ask questions!

1. Empirical progress in machine learning: benchmarks

2. What can we learn from ML benchmarks?

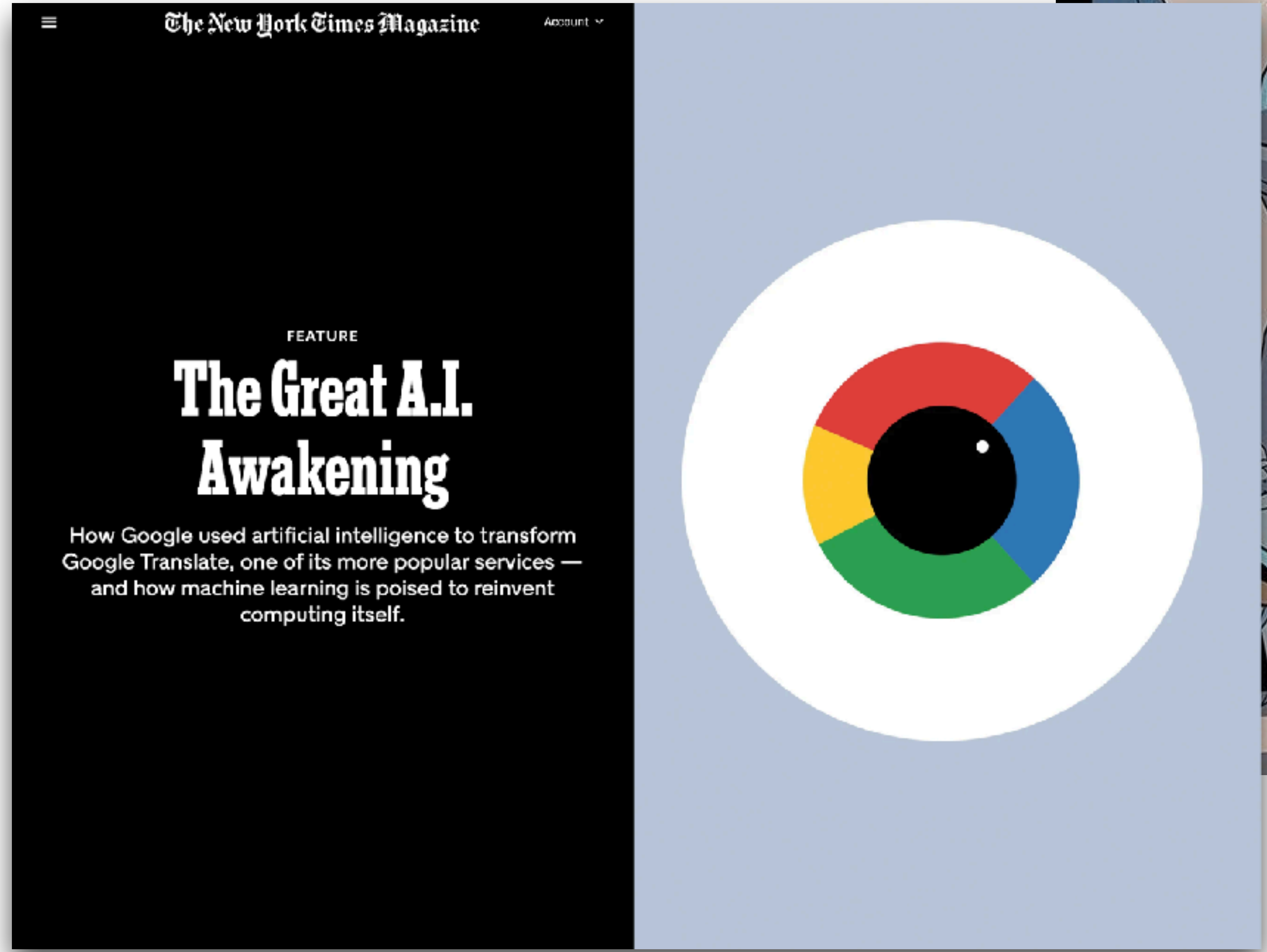
3. Limitations of current ML methods

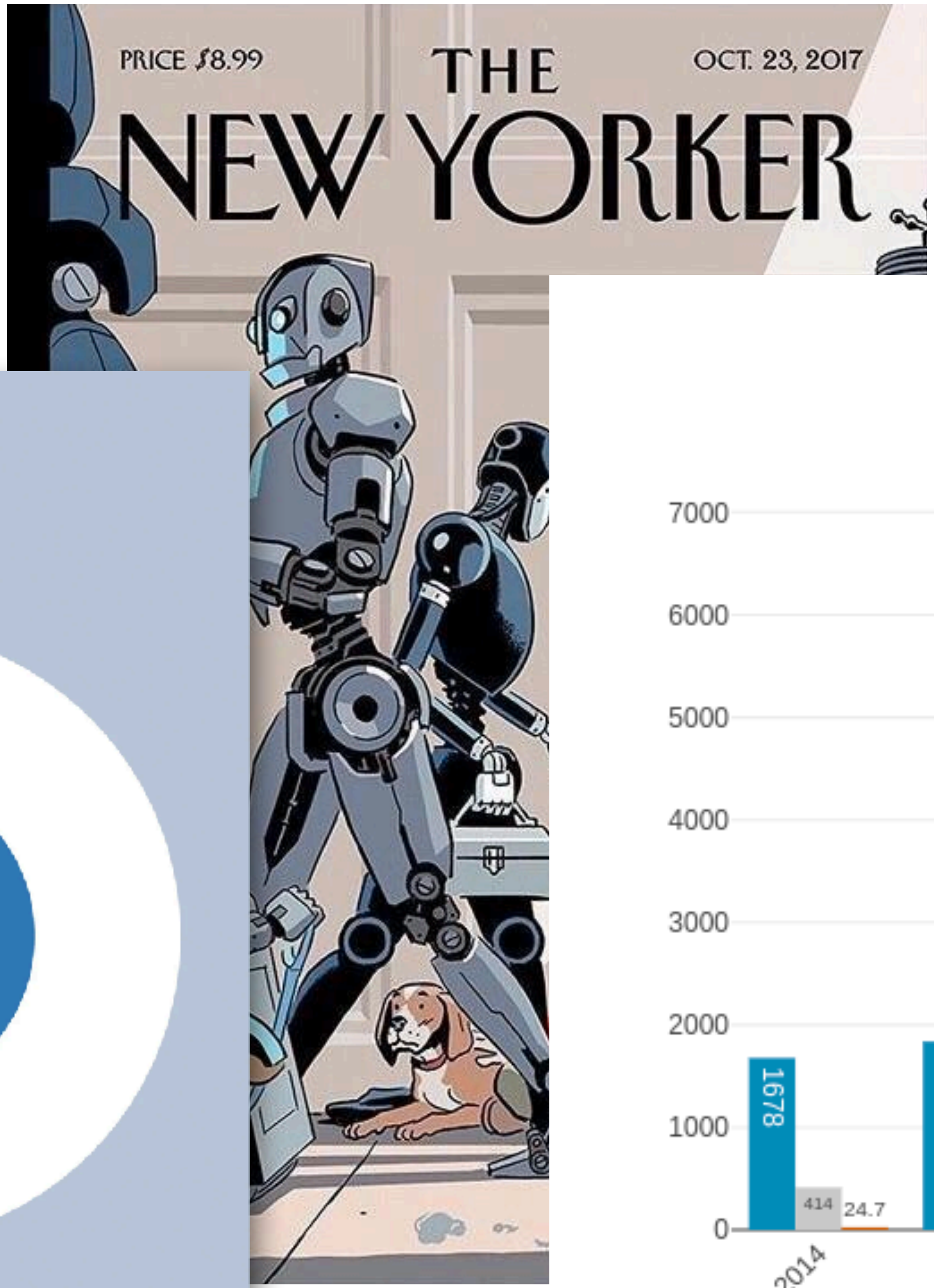
1. Empirical progress in machine learning: benchmarks

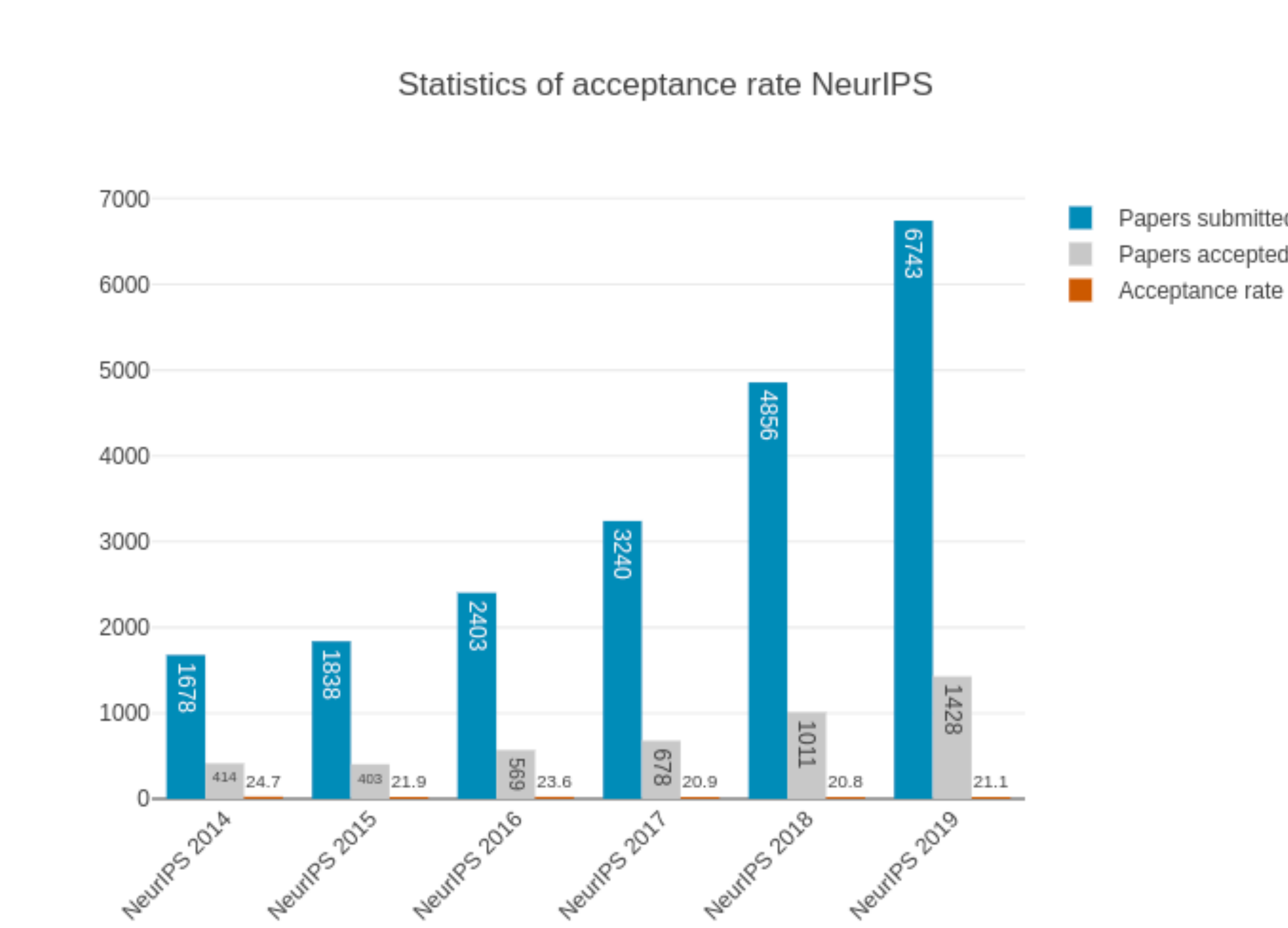
2. What can we learn from ML benchmarks?

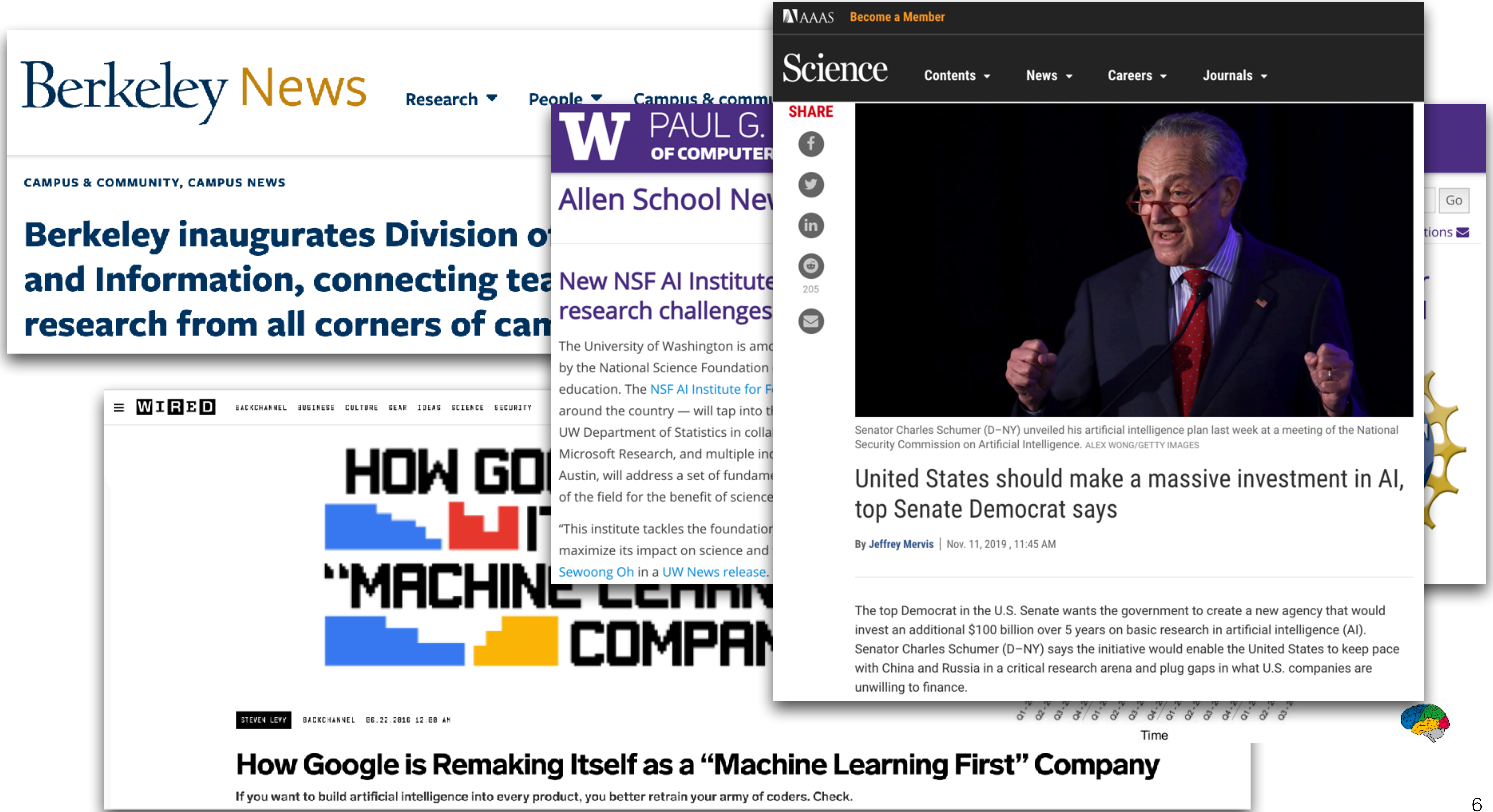
3. Limitations of current ML methods

Explosive Growth in ML









Berkeley News
Research ▾ People ▾ Campus & comm
CAMPUS & COMMUNITY, CAMPUS NEWS
Berkeley inaugurates Division o
and Information, connecting tea
research from all corners of can

W PAUL G.
OF COMPUTER
Allen School Ne
New NSF AI Institute
research challenges
The University of Washington is amo
by the National Science Foundation
education. The NSF AI Institute for F
around the country — will tap into t
UW Department of Statistics in colla
Microsoft Research, and multiple in
Austin, will address a set of fundam
of the field for the benefit of science
"This institute tackles the foundatio
maximize its impact on science and
Sewoong Oh in a UW News release.

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.

Go
tions

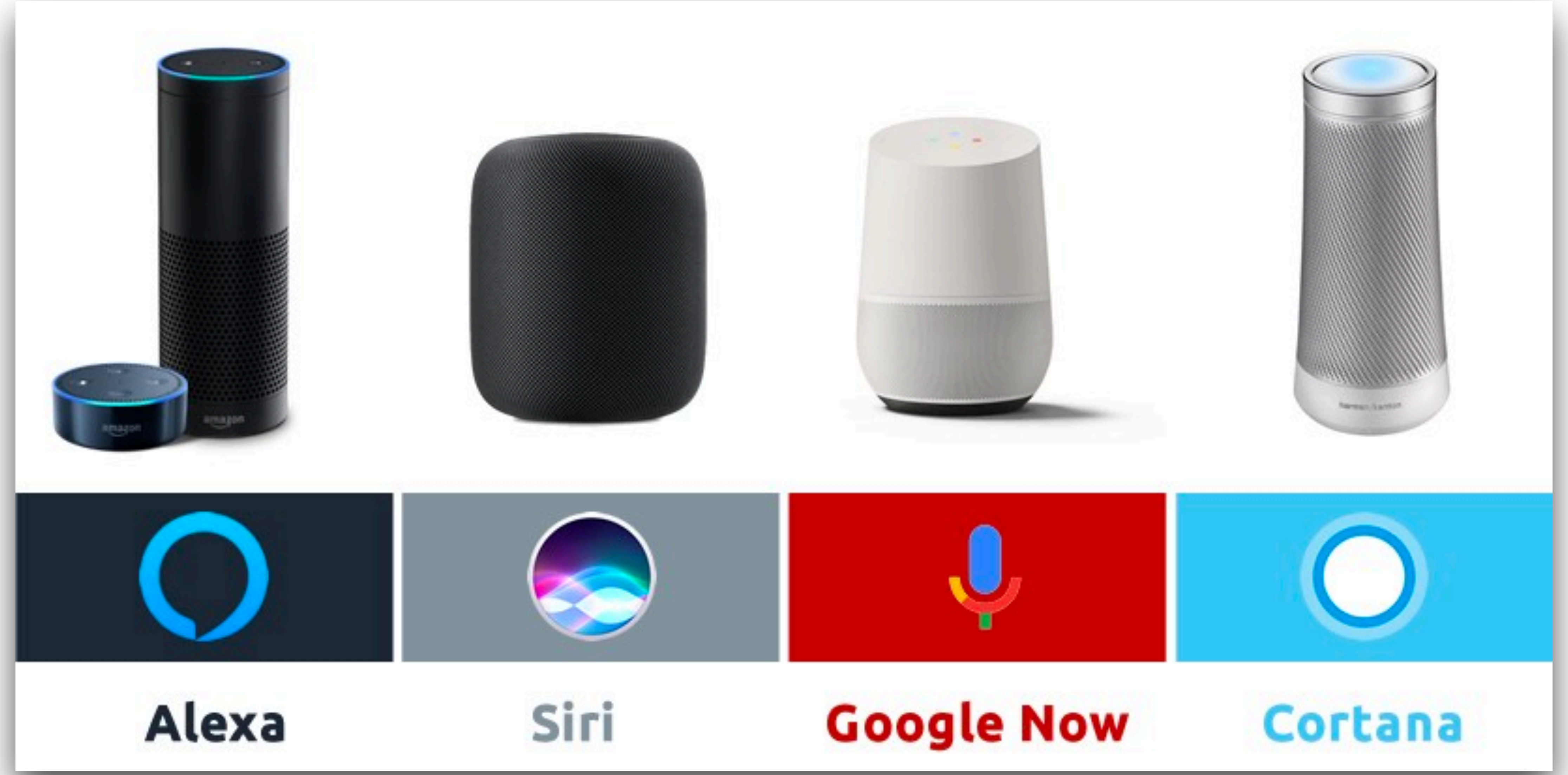
WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY
HOW GO
"MACHINE
COMPA
STEVEN LEVY BACKCHANNEL 06.22.2016 12:00 AM
How Google is Remaking Itself as a "Machine Learning First" Company
If you want to build artificial intelligence into every product, you better retrain your army of coders. Check.

Time

Self-driving cars

Games

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]

ImageNet

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

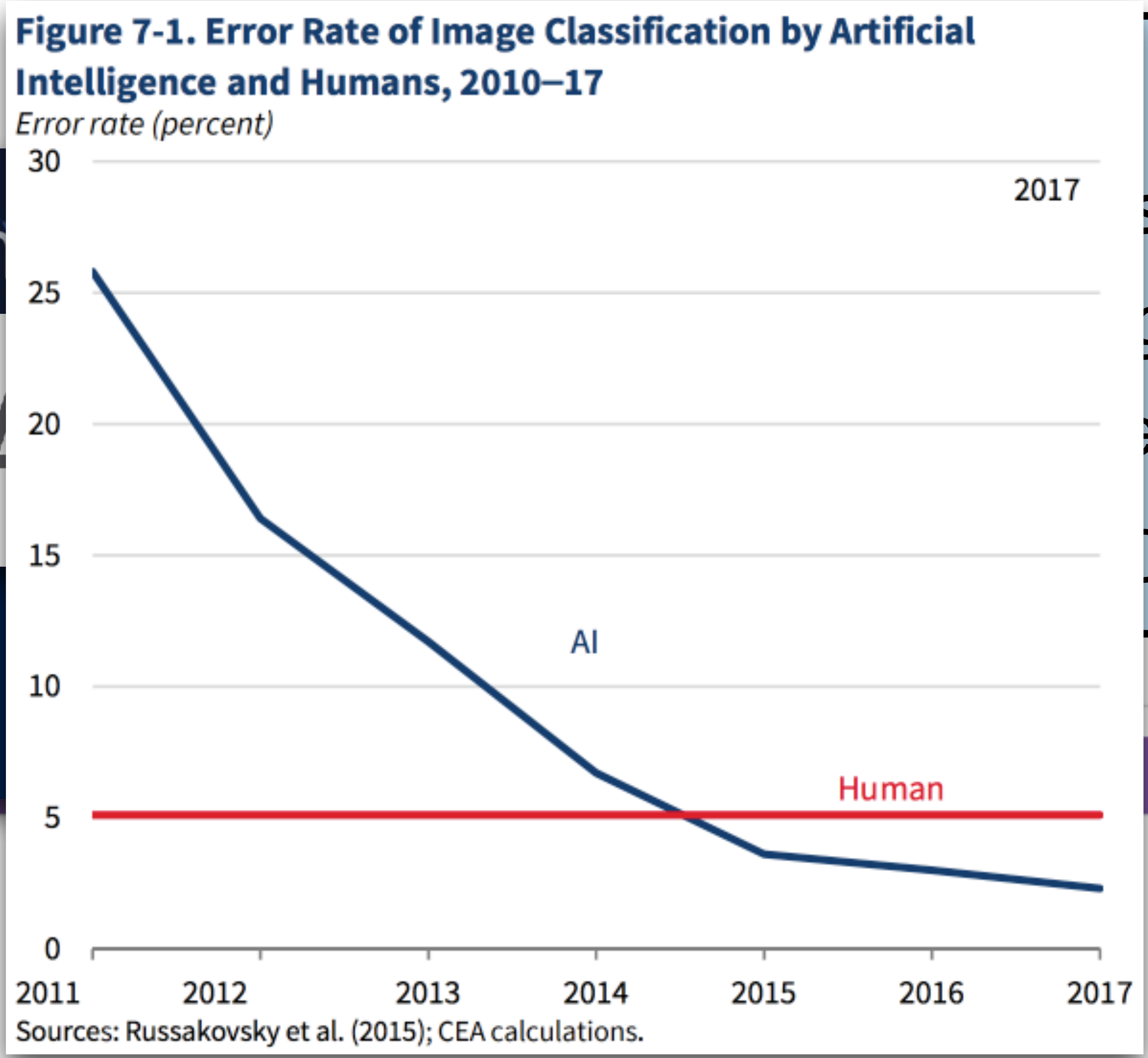
Golden retriever

Great white shark

Minibus

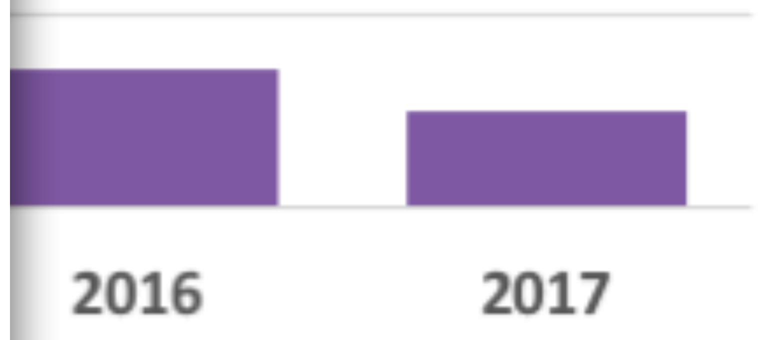
ImageNet

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)

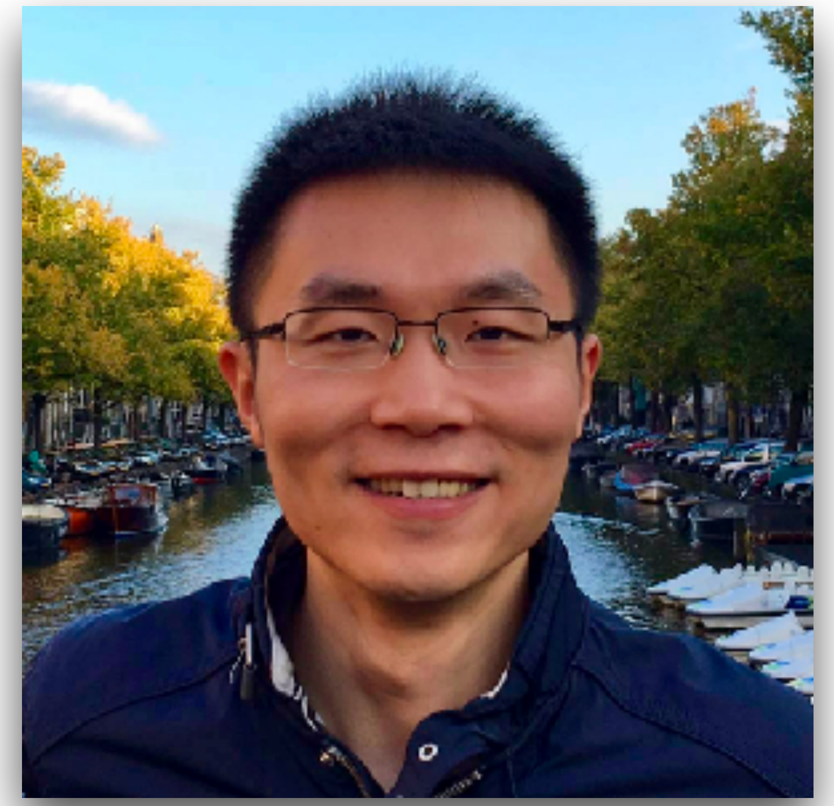
(Planned) ImageNet is 1000x larger!

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)

flickr

How do you label them?

Internet! (Crowdsourcing platforms)

+ lots of **clever** task design

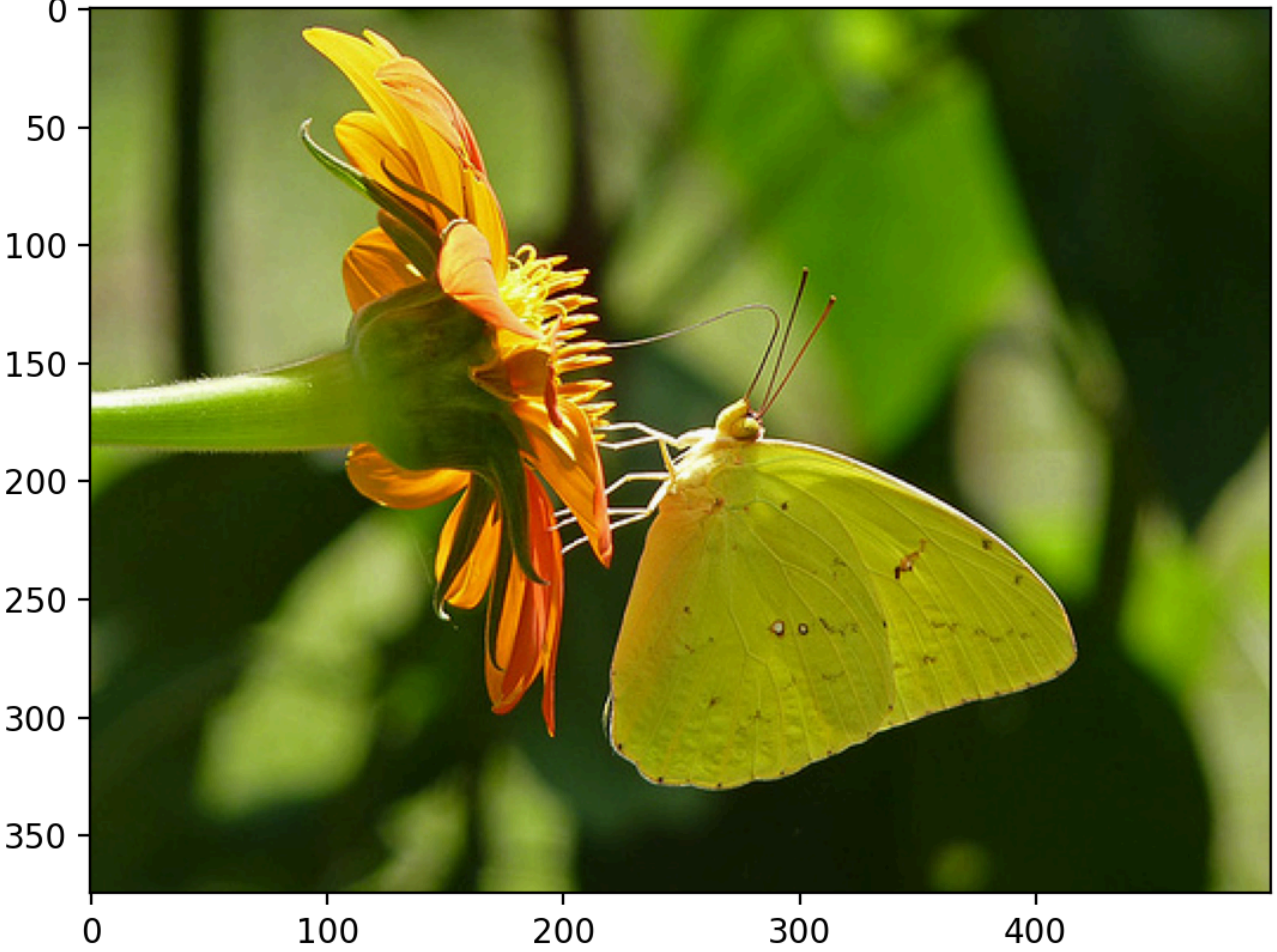
+ lots of **hard** work

amazon
beta
mechanical turk

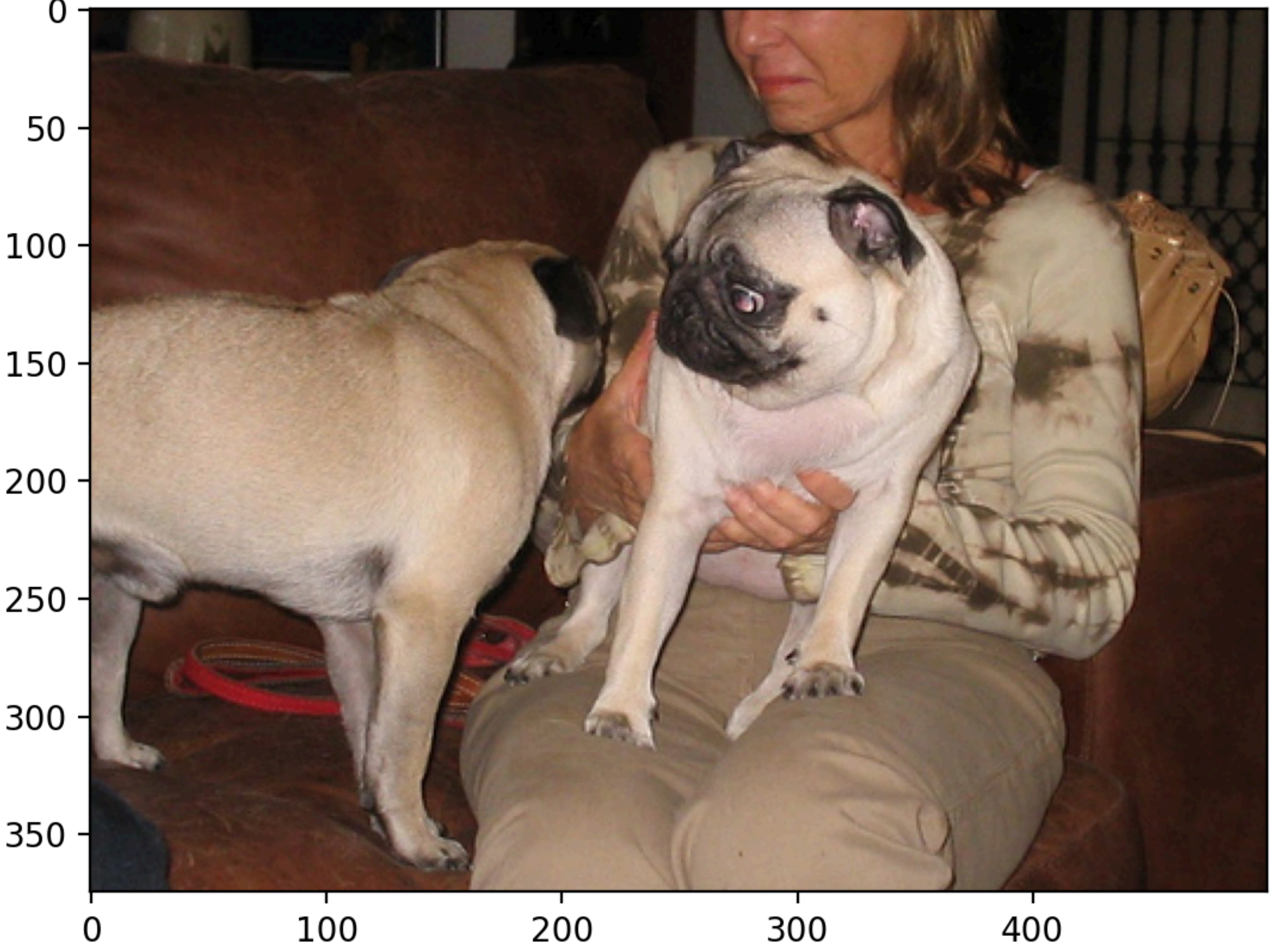
[Deng, Dong, Socher, Li, Li, Fei-Fei'09]

ILSVRC2012_val_00000293.JPEG

ILSVRC2012_val_00025559.JPEG

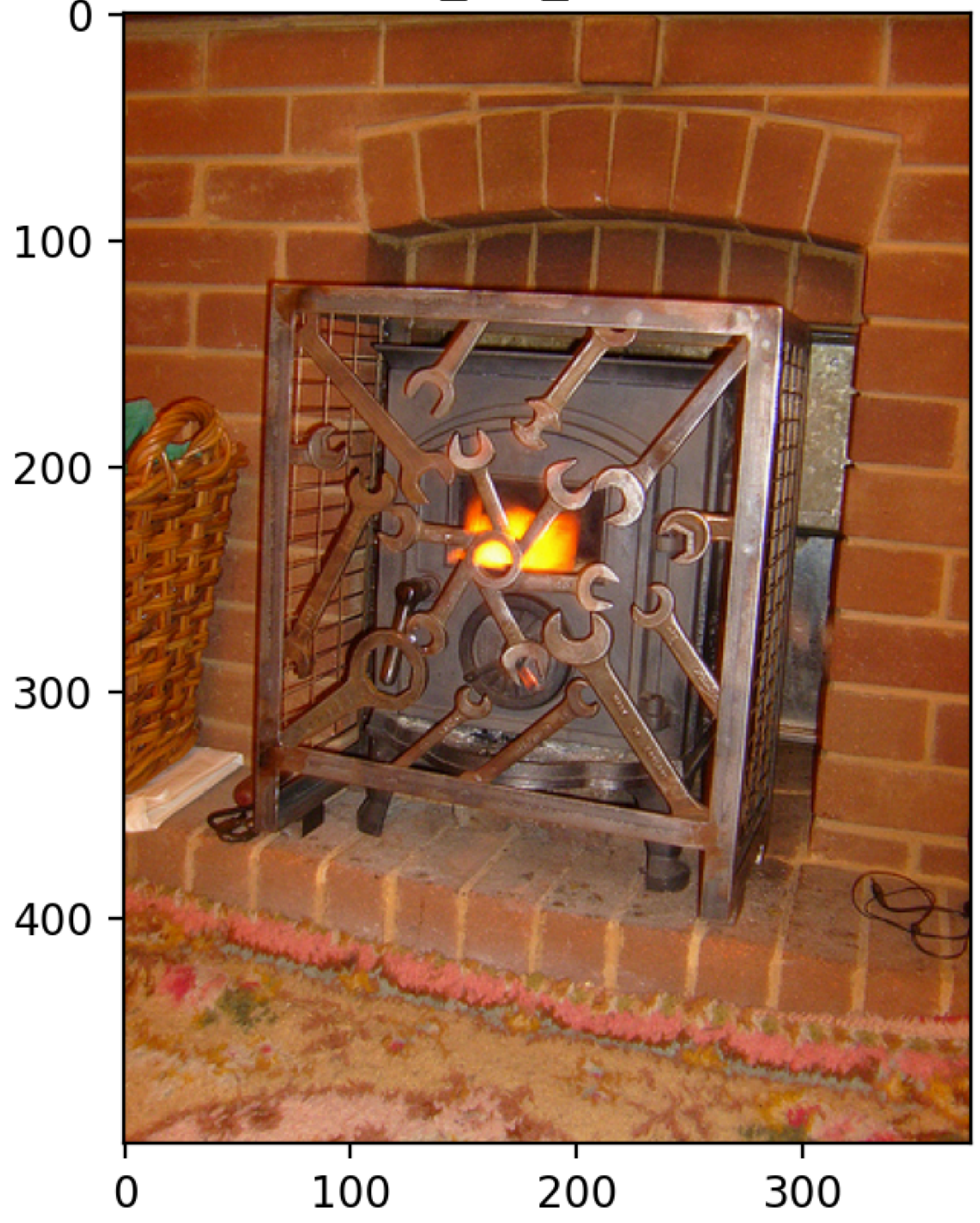


ILSVRC2012_val_00047583.JPEG



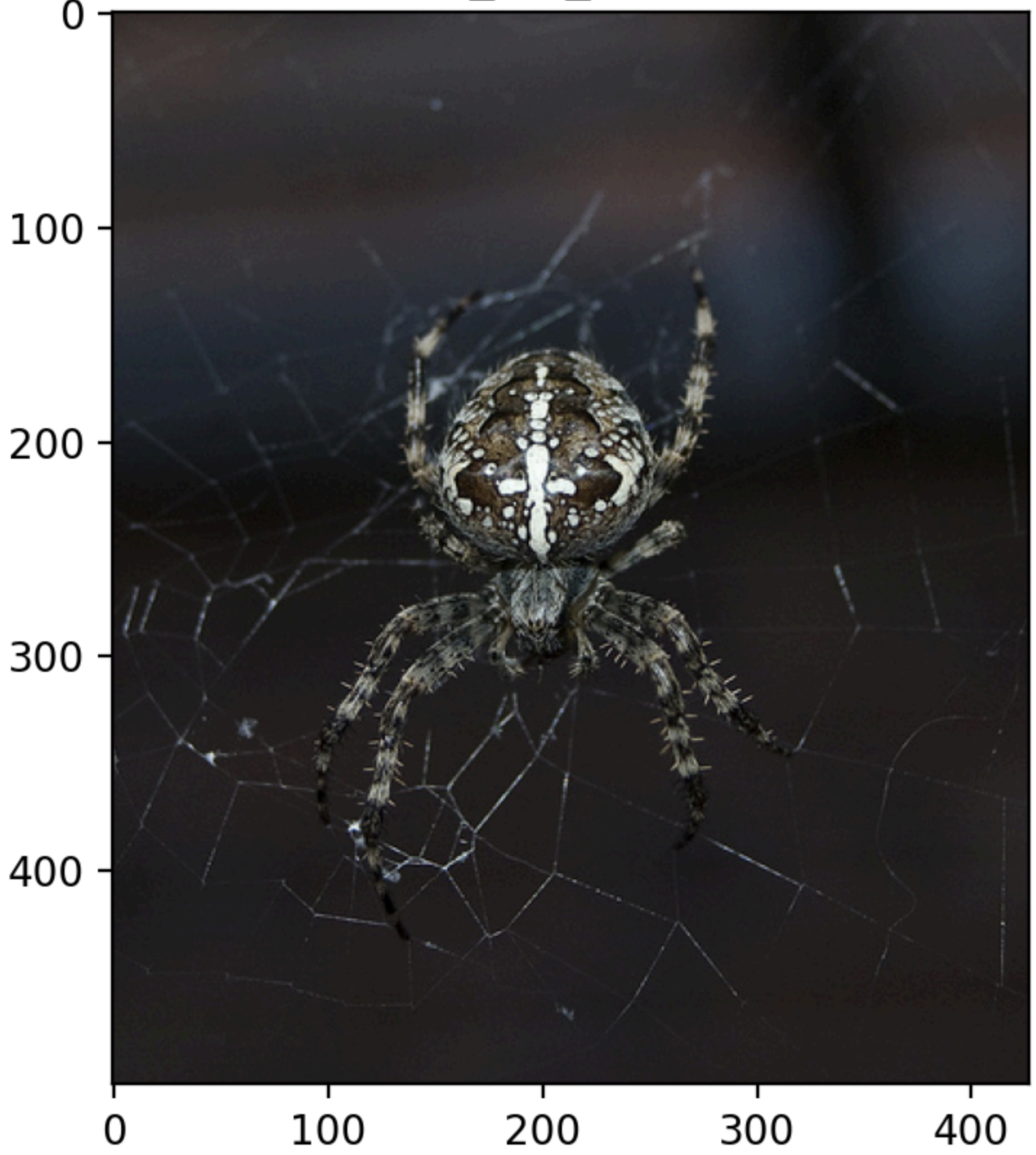
ILSVRC2012_val_00033445.JPEG

ILSVRC2012_val_00027126.JPEG

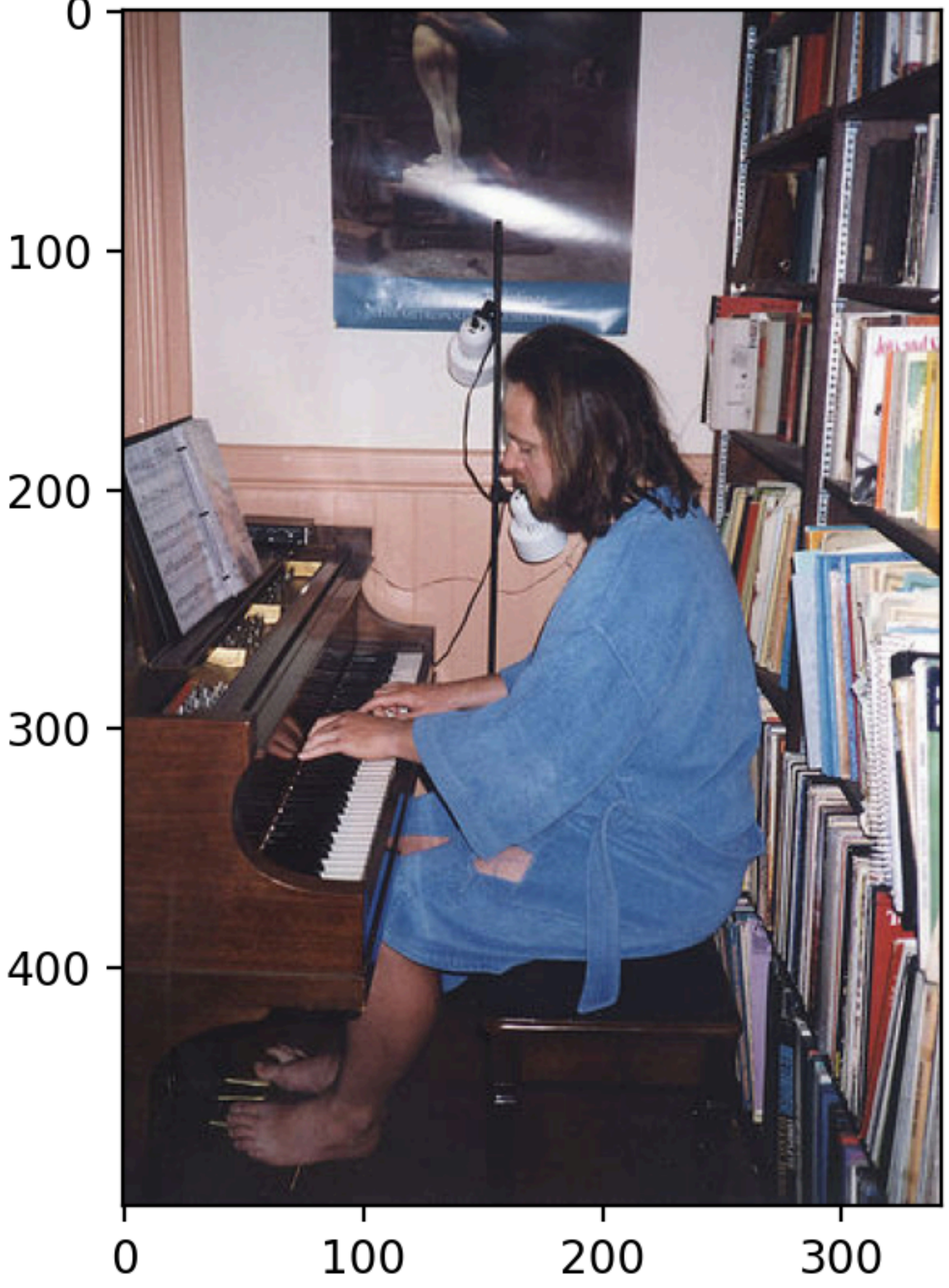


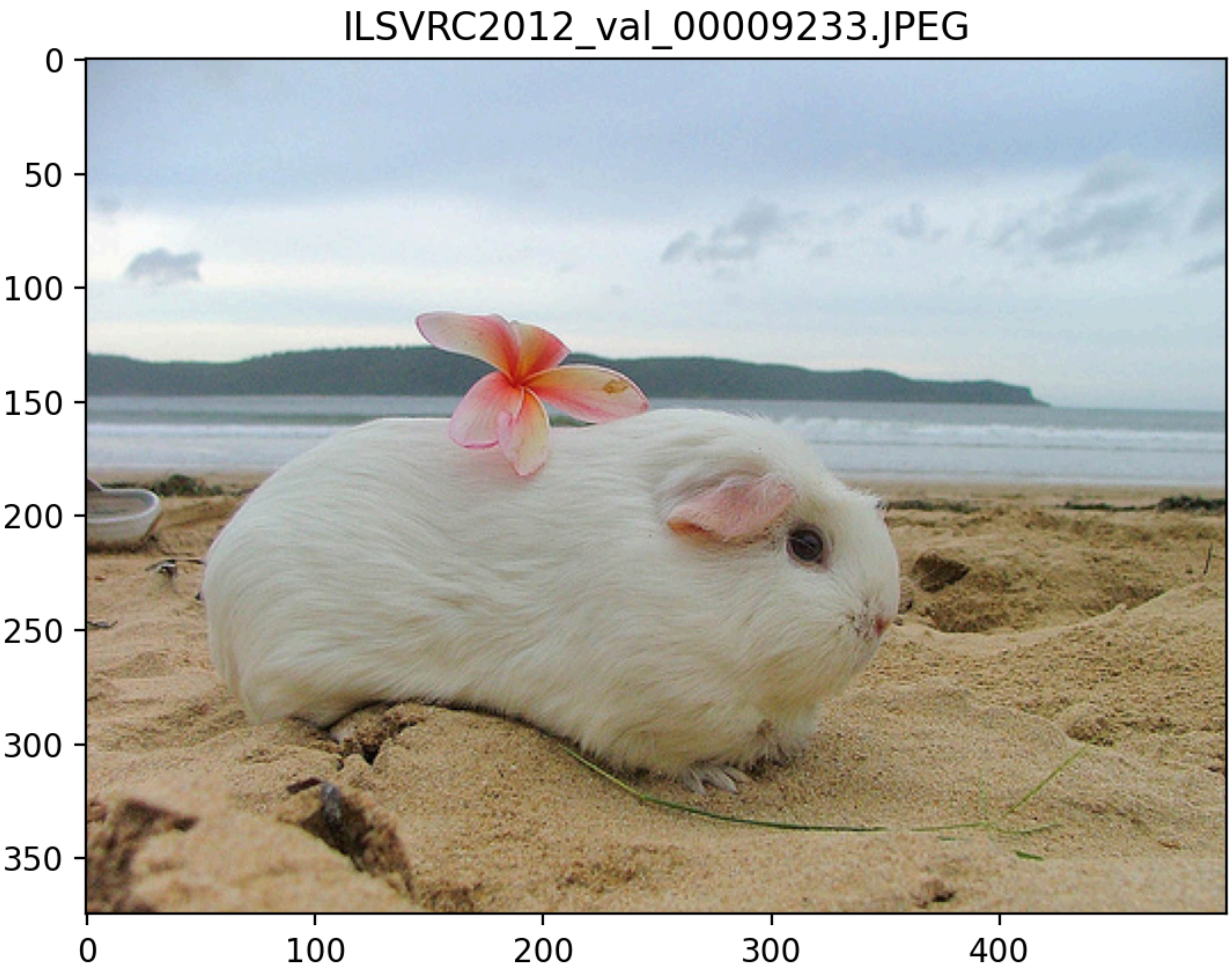
ILSVRC2012_val_00013085.JPEG

ILSVRC2012_val_00035593.JPEG

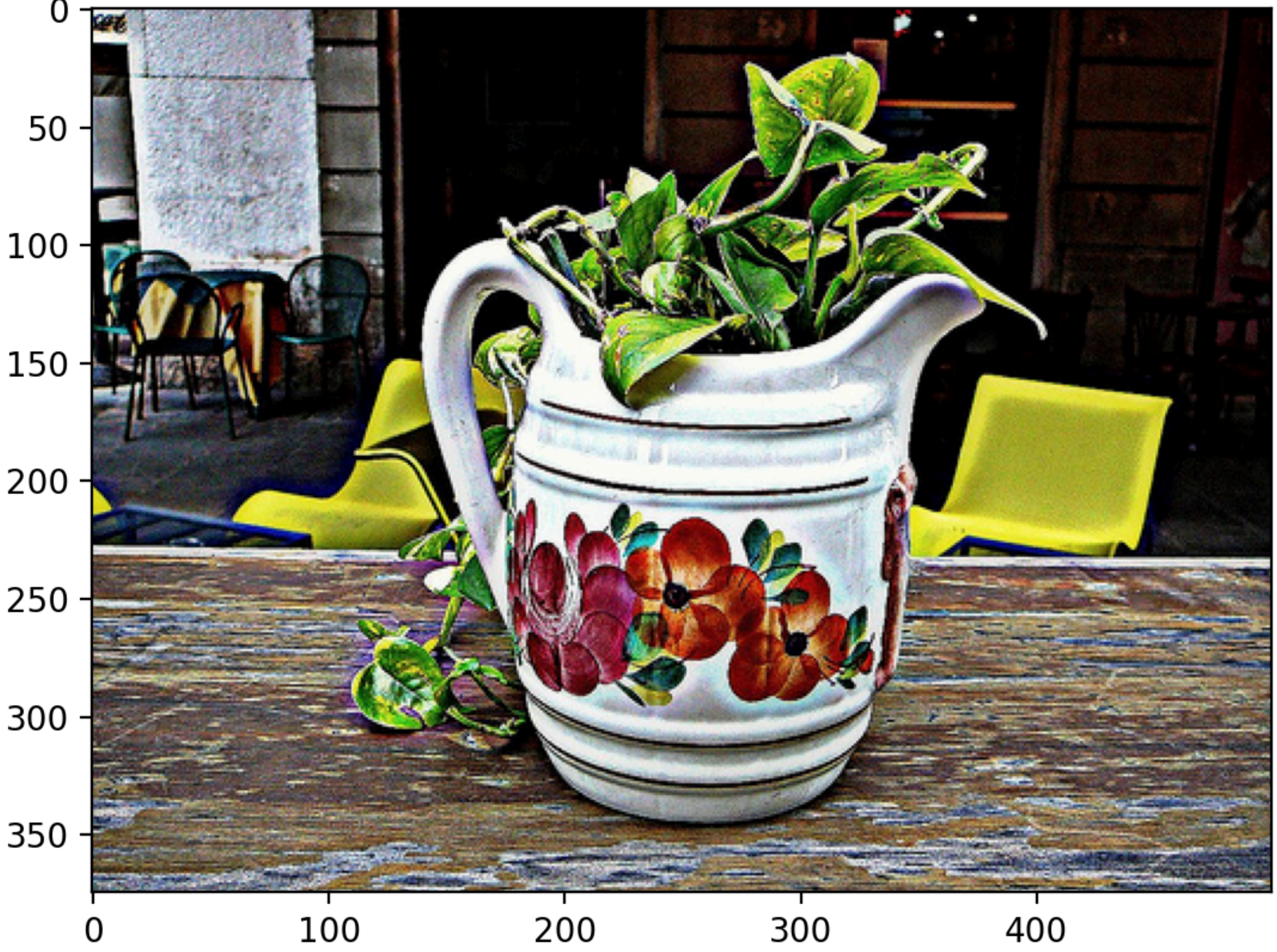


ILSVRC2012_val_00012694.JPEG




ILSVRC2012_val_00009233.JPEG
0
50
100
150
200
250
300
350
0
100
200
300
400

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)

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)

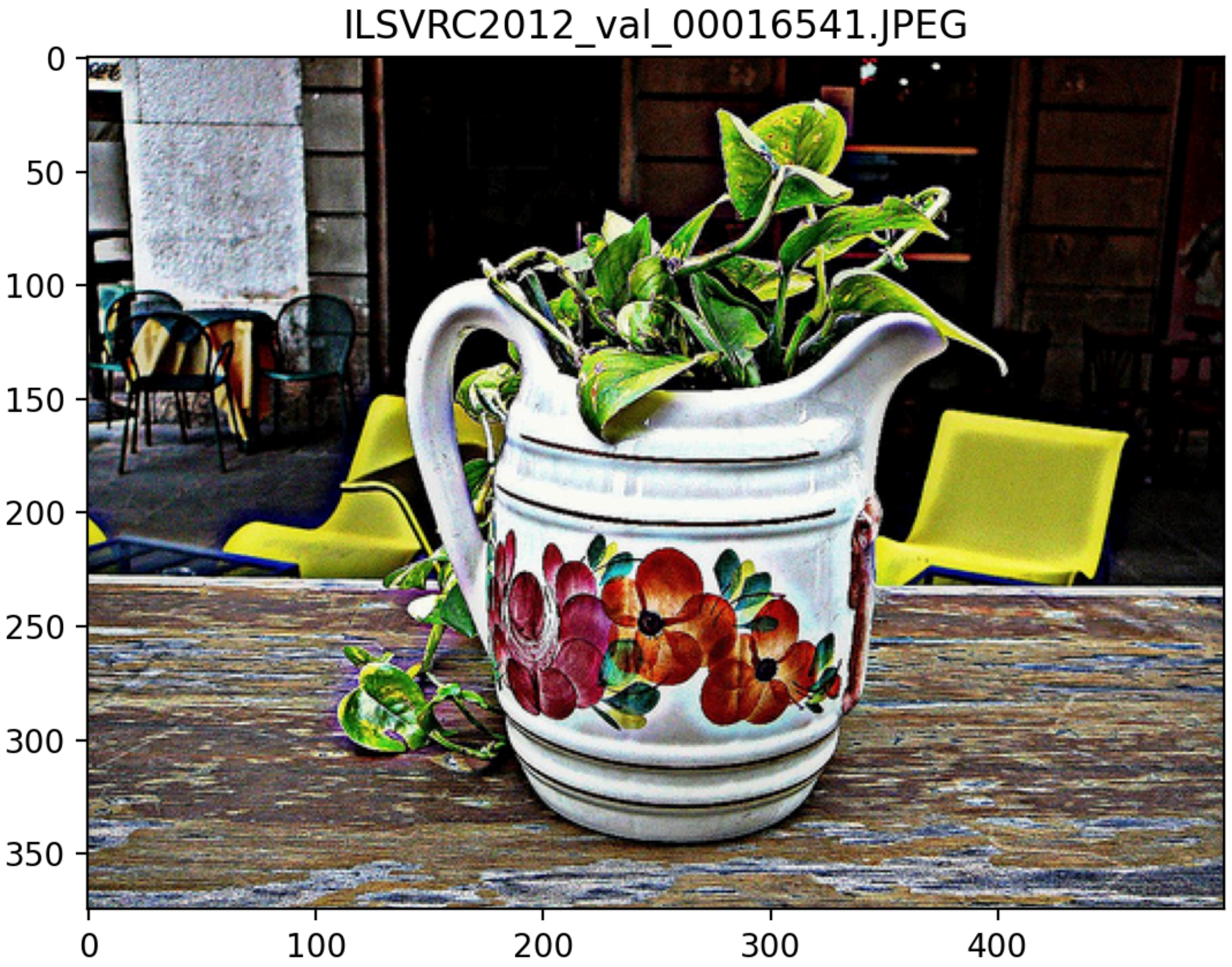
Validation set: 50,000 images for 1,000 classes (exactly class-balanced)

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

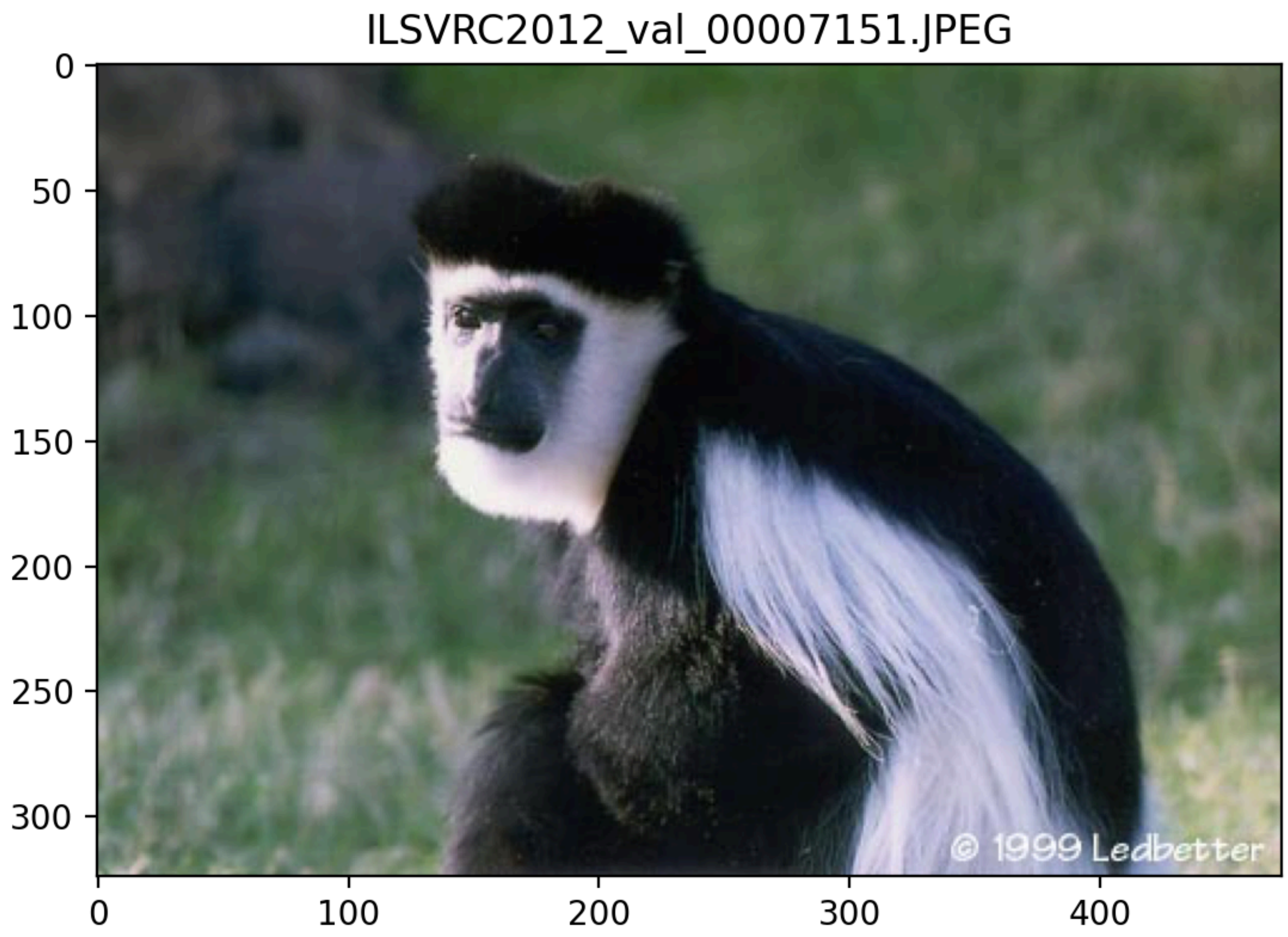
Why? Sometimes multiple labels per image, sometimes unclear class boundaries.
+ task is already hard enough



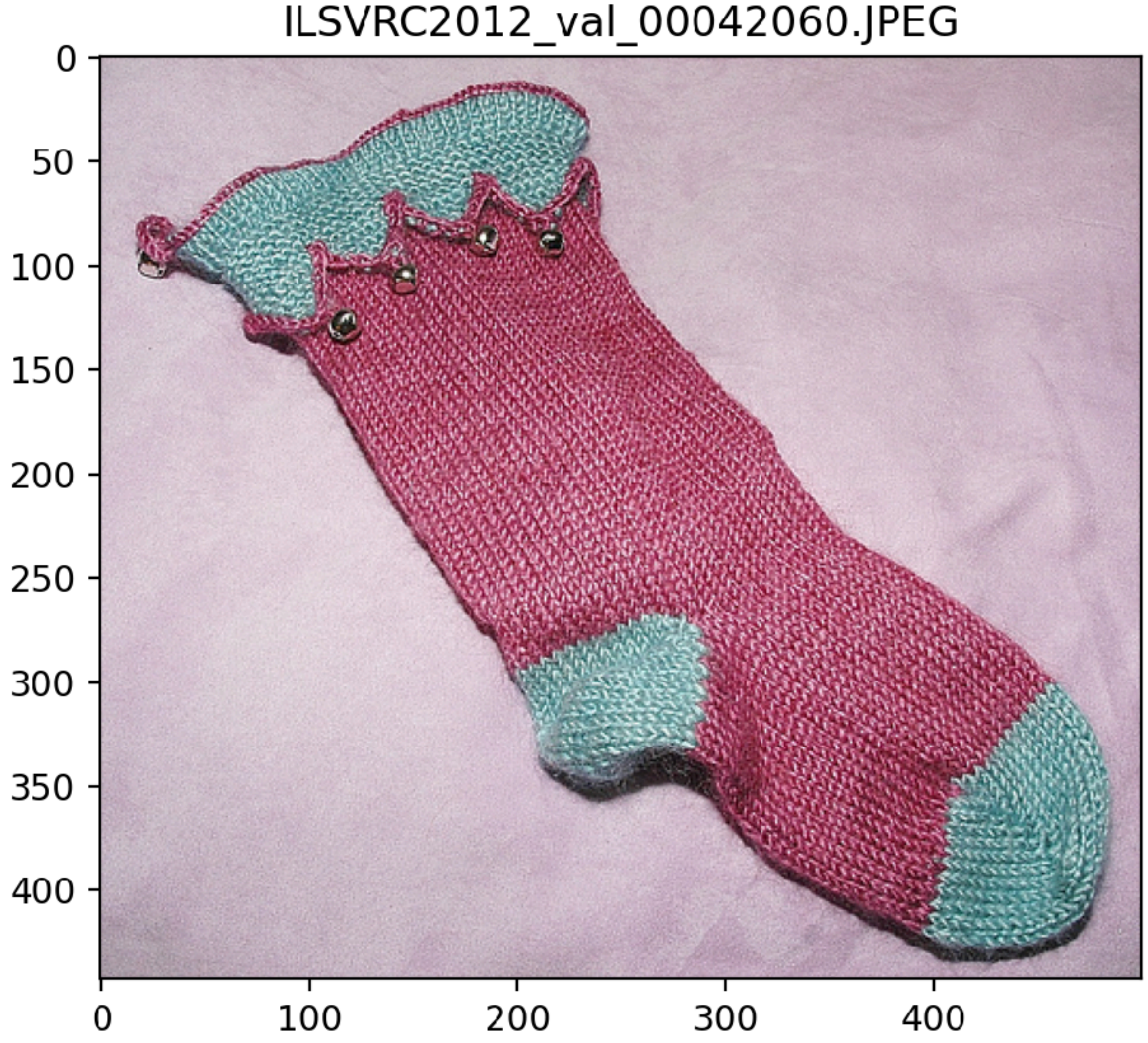
n03950228 pitcher, ewer

WordNet ID (wnid)

Synonym set

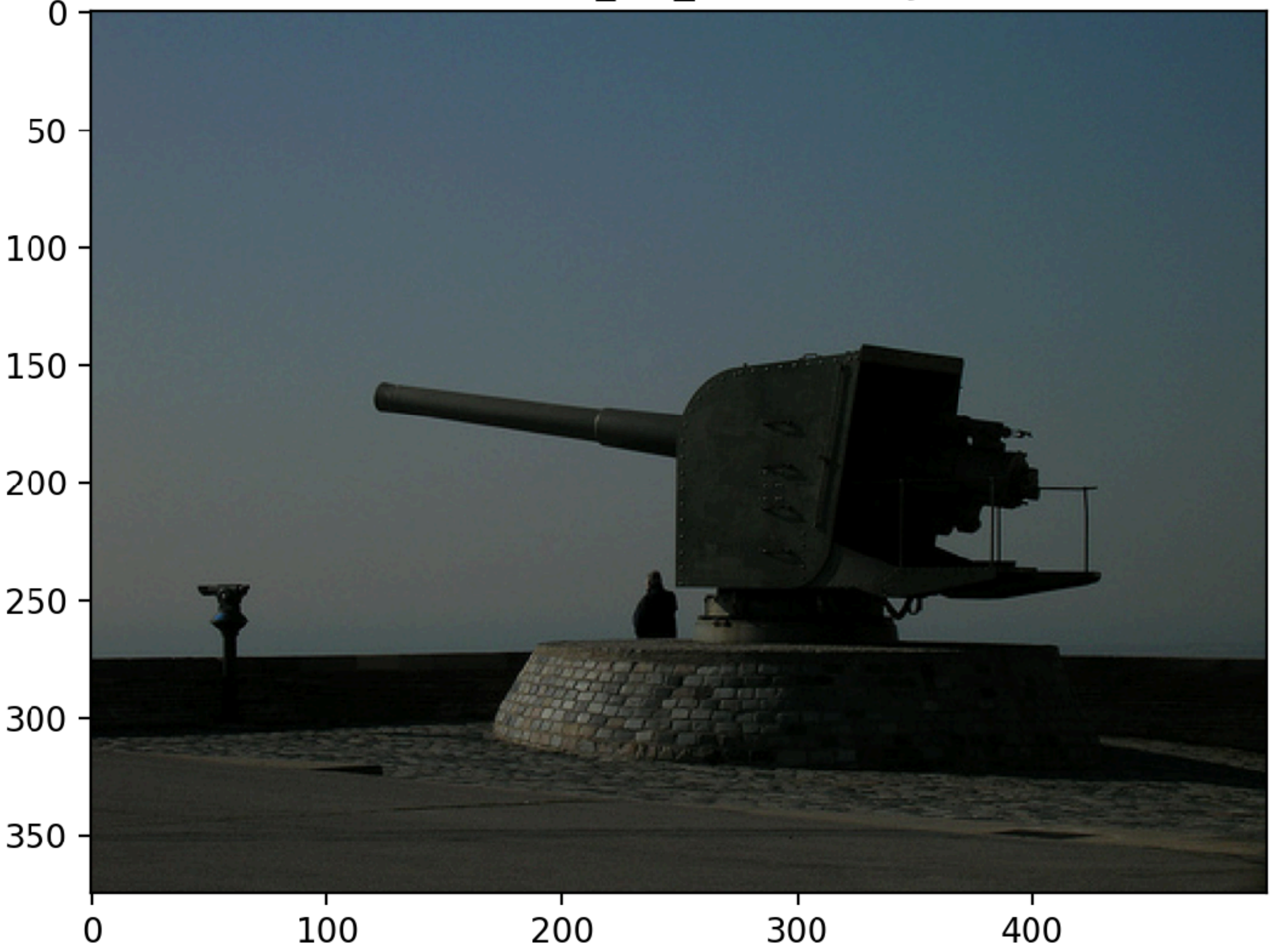


n02488702 colobus, colobus monkey

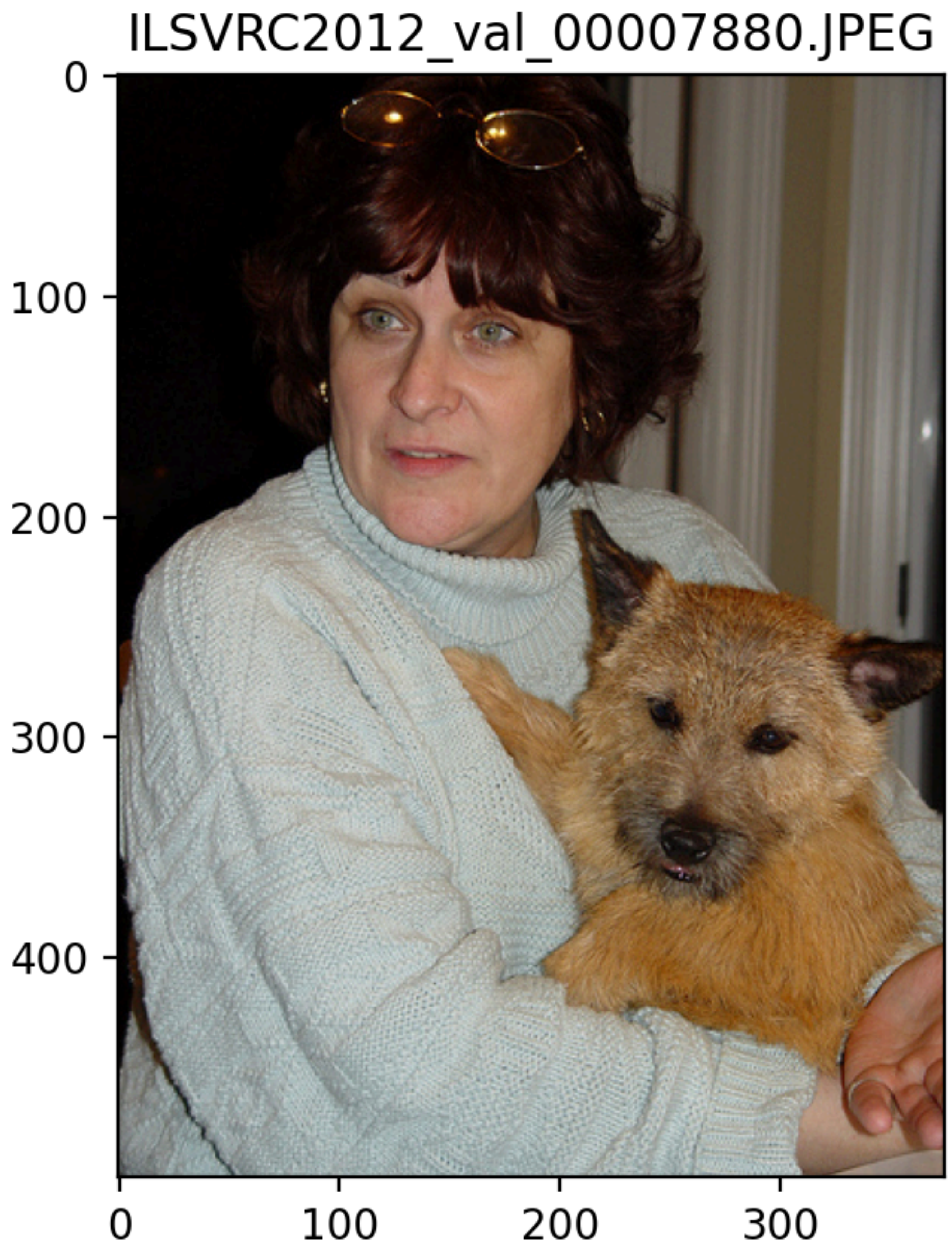


n03026506 Christmas stocking

ILSVRC2012_val_00001902.JPEG

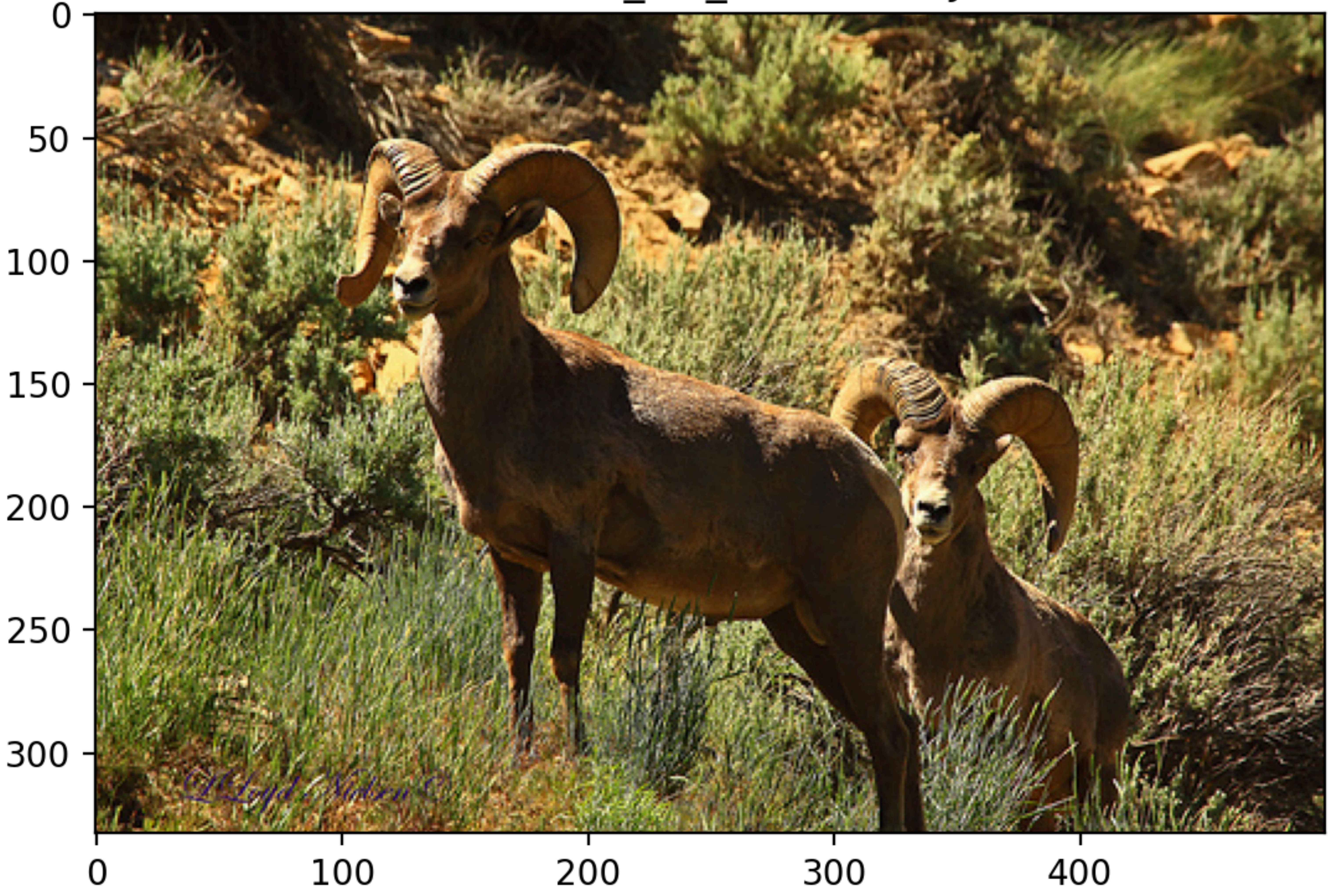


n02950826 cannon



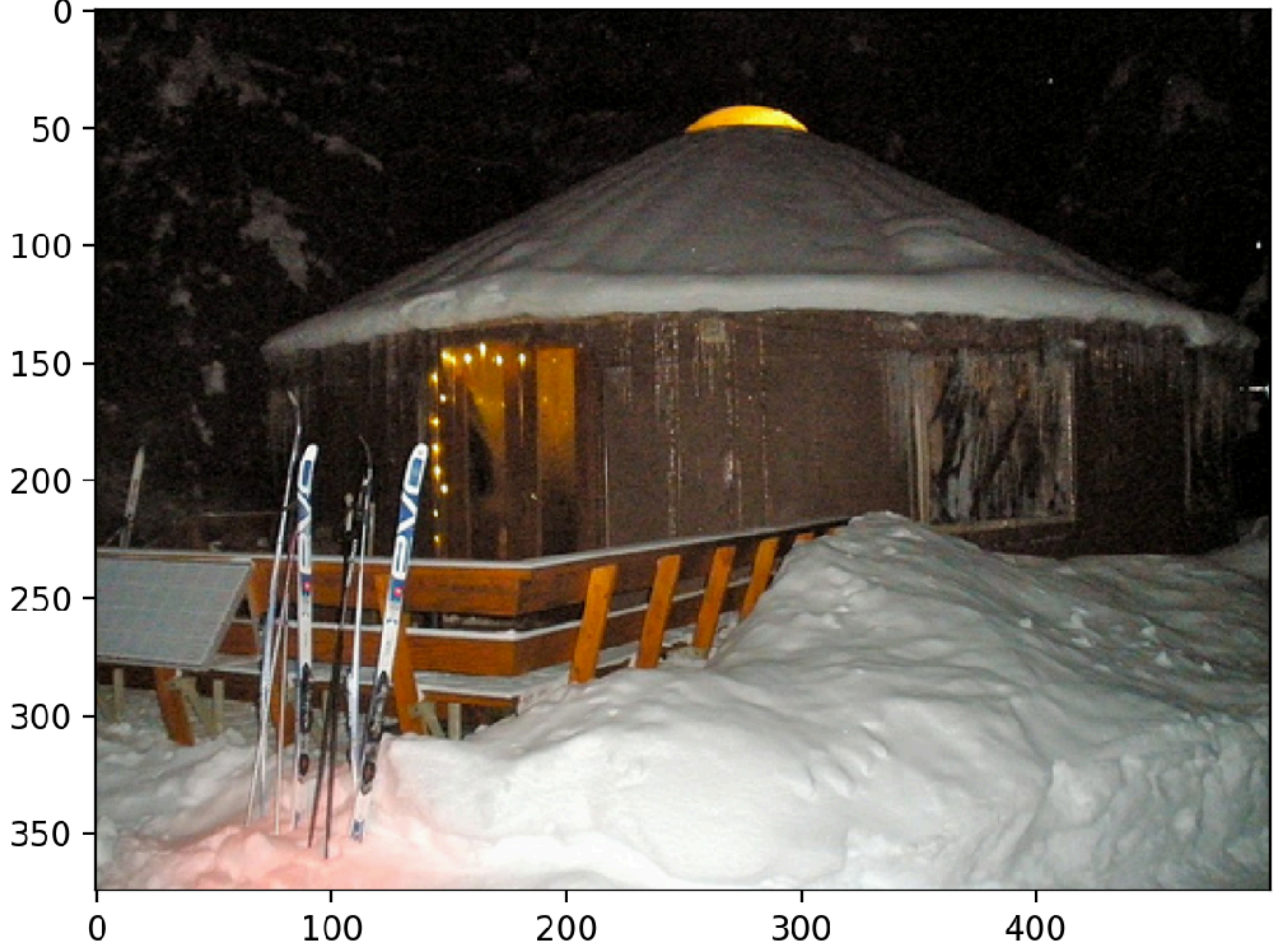
n02094258 Norwich terrier

ILSVRC2012_val_00016391.JPEG



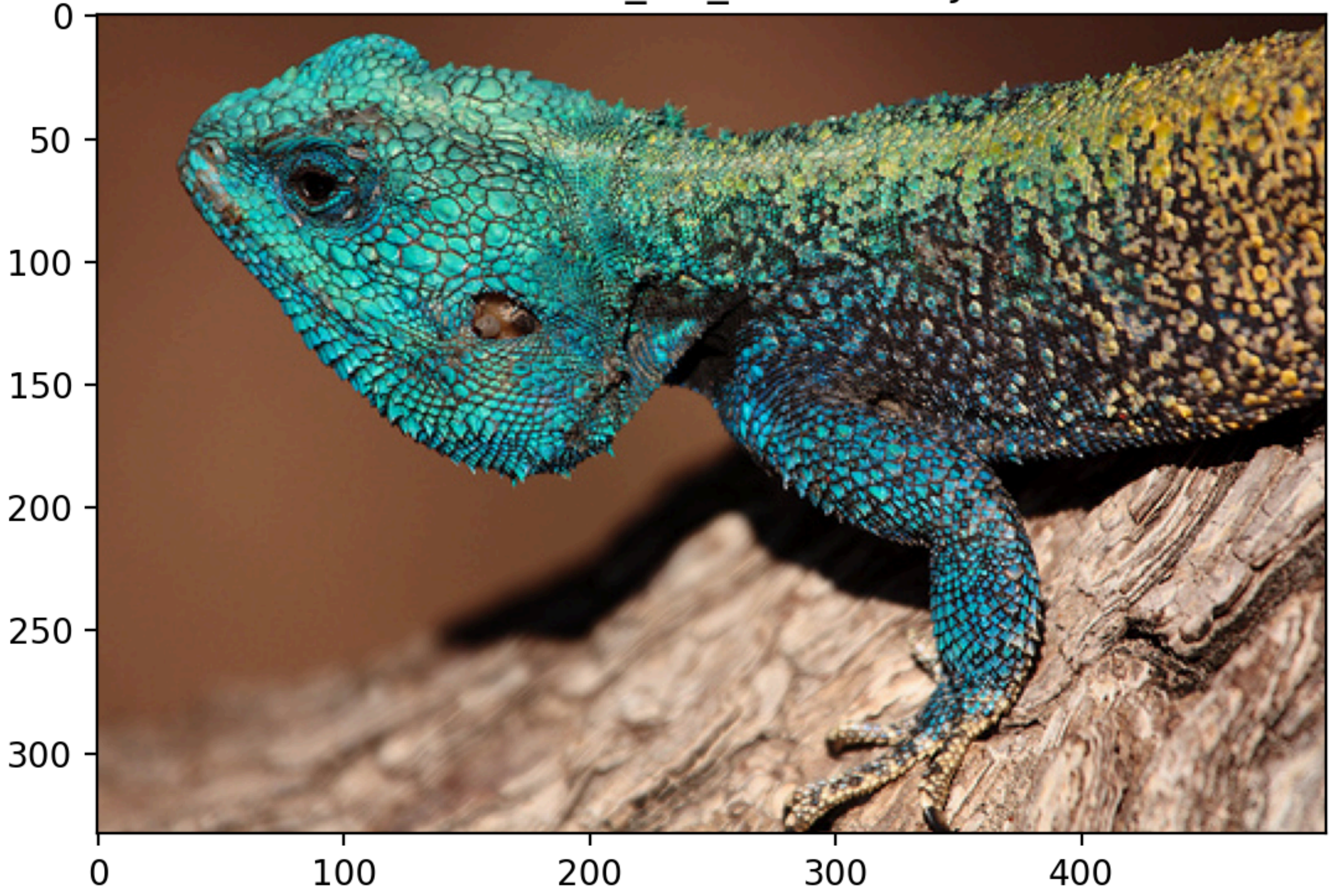
n02412080 ram, tup

ILSVRC2012_val_00020151.JPEG



n04613696 yurt

ILSVRC2012_val_00041169.JPEG



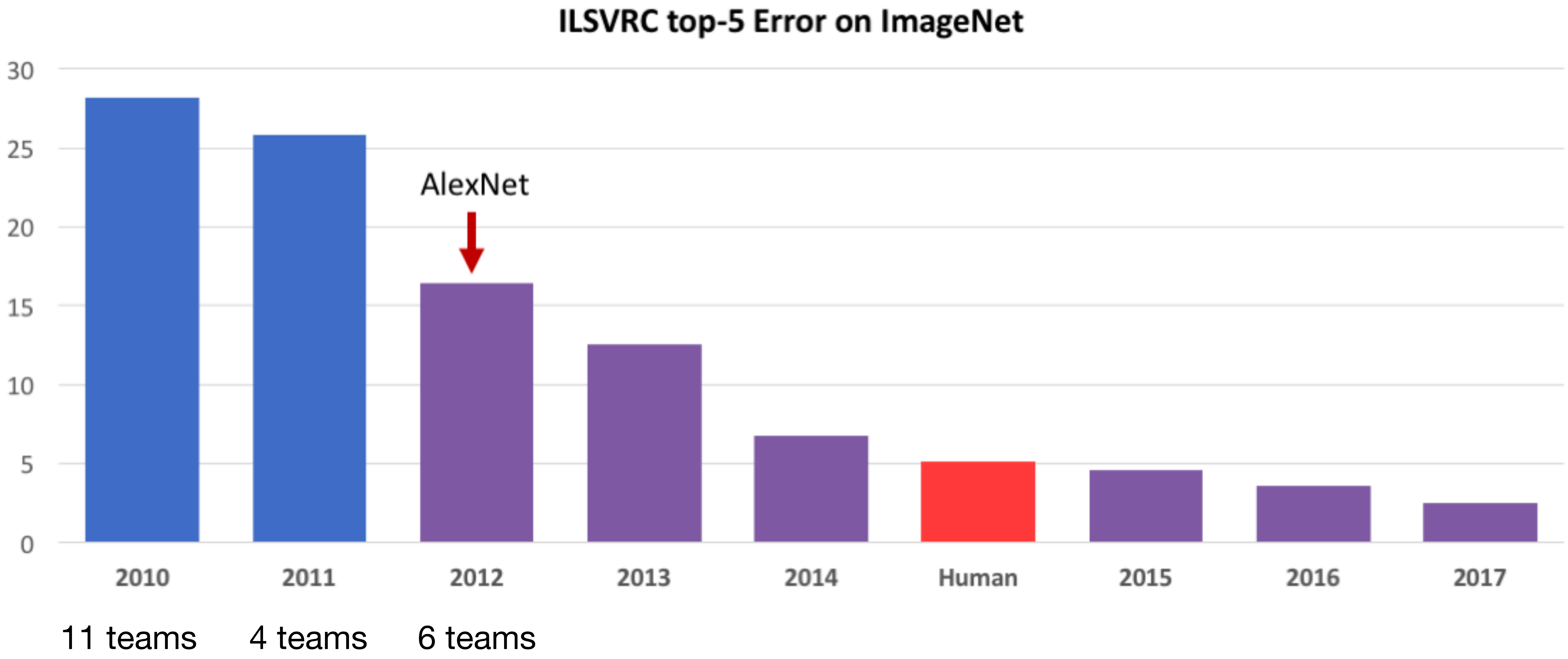
n01687978 agama

n02134418 sloth bear, Melursus ursinus, Ursus ursinus

n04591713 wine bottle

OK, now we have trained Hong

➡ Test time!


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

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 th

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.

AlexNet

ImageNet Classification with Deep Convolutional Neural Networks

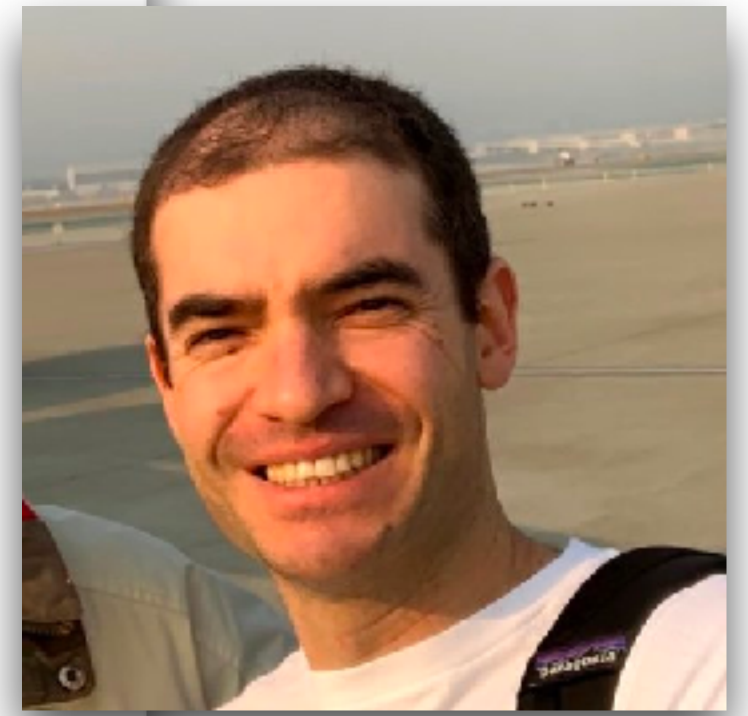
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

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

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

Abstract

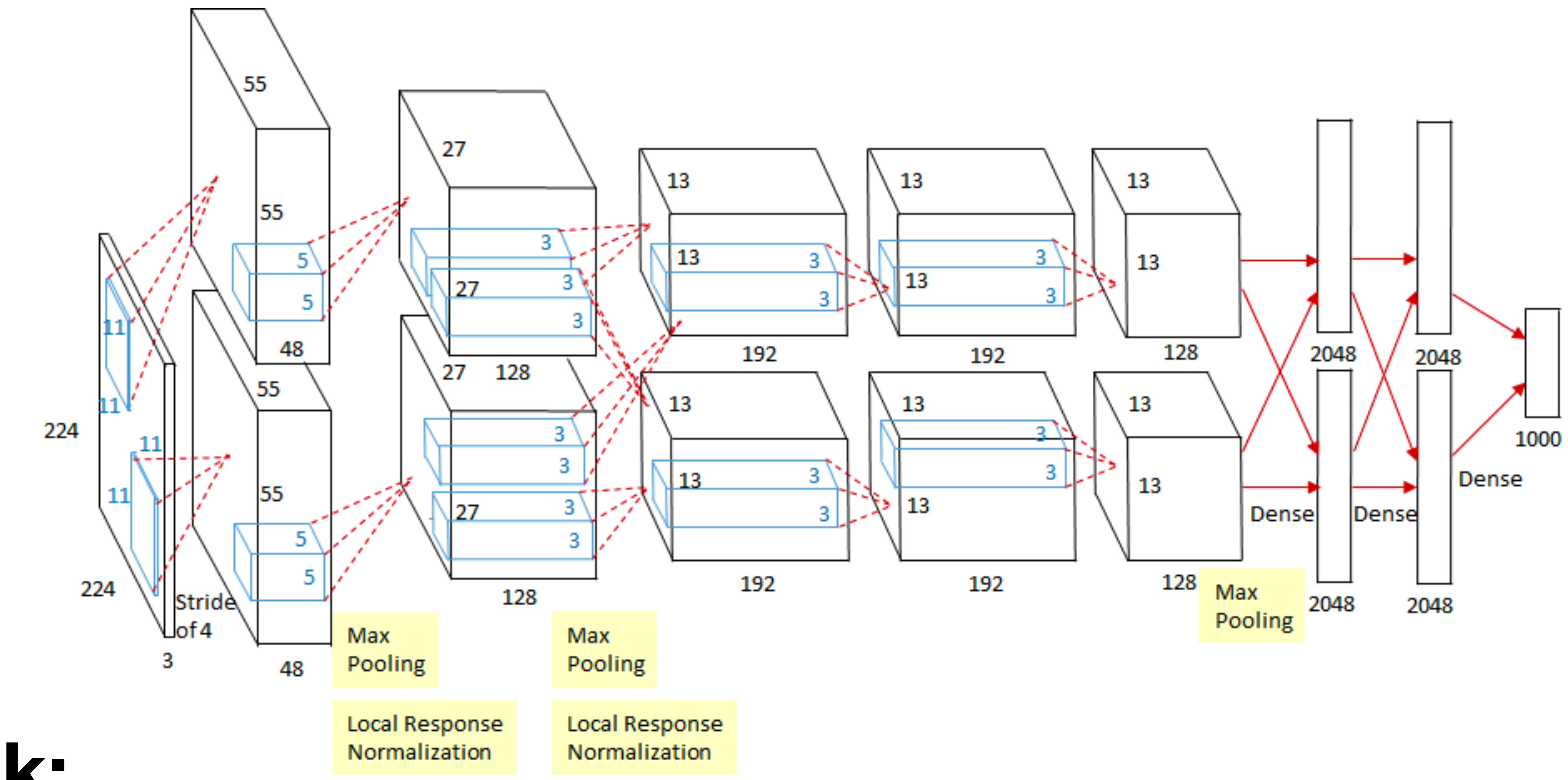
We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. 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

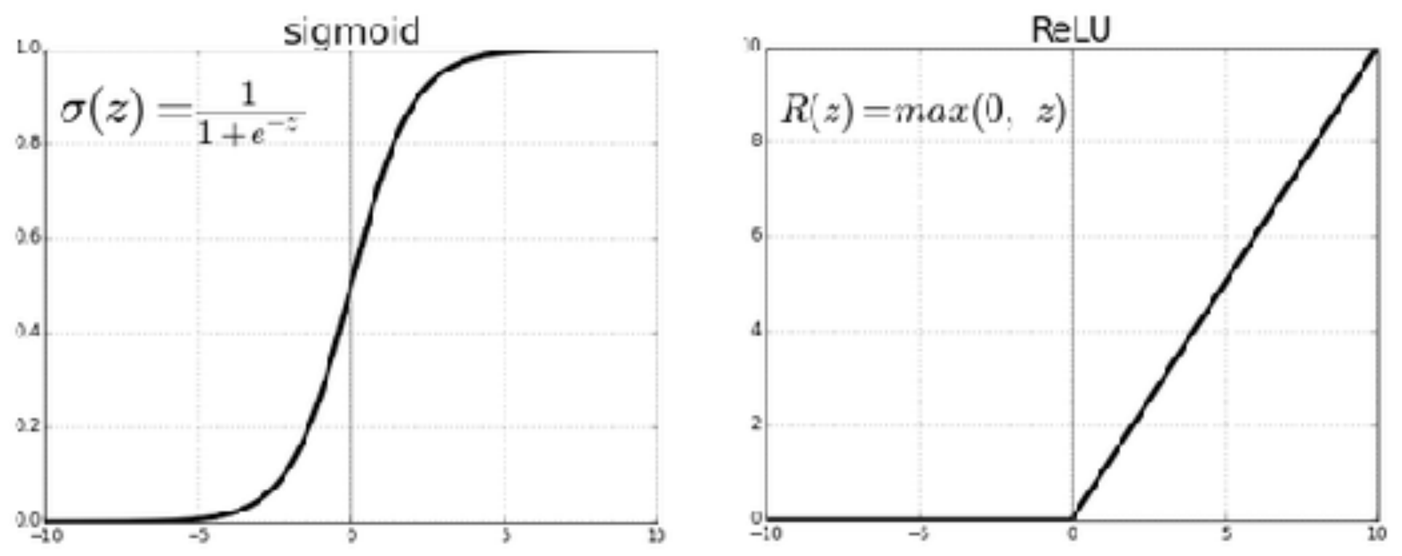


Basic building block:

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

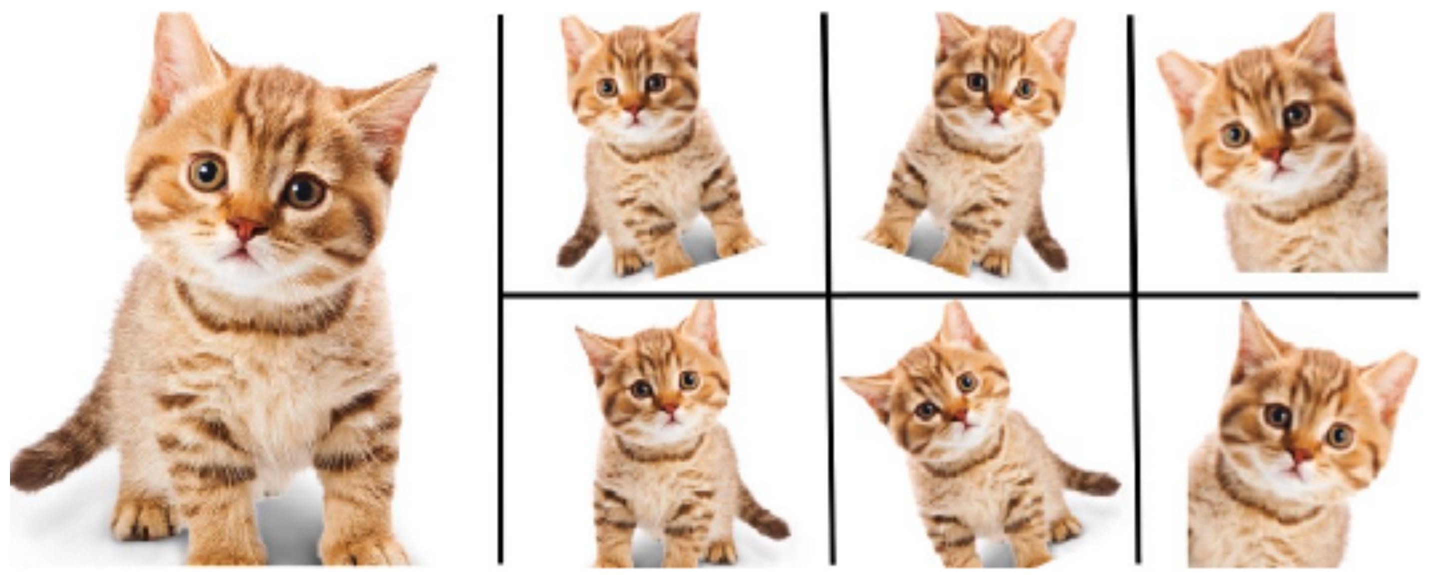


ReLU (rectified linear unit) non-linearity

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%**

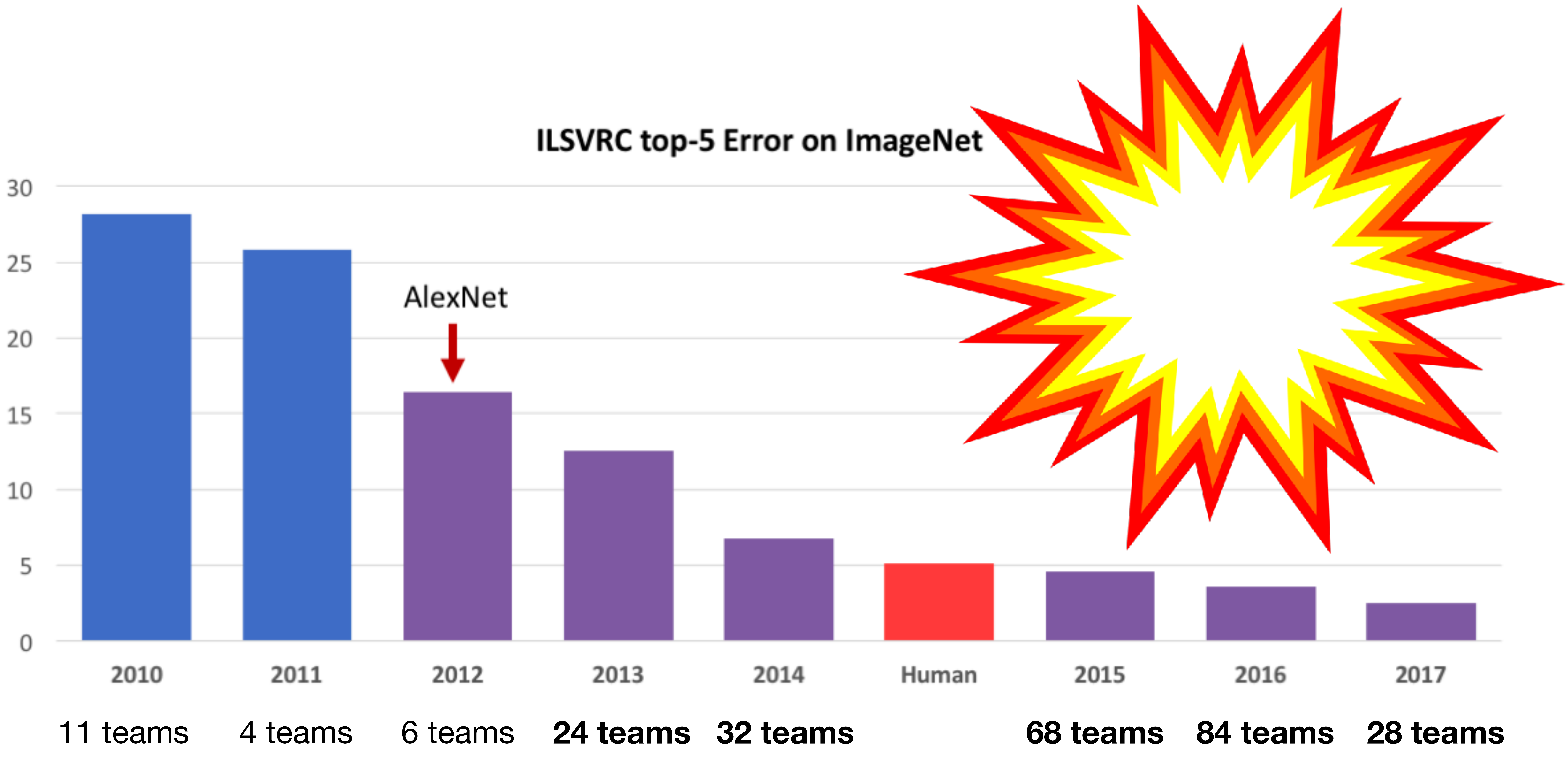
Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

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
- 74,000 citations, Turing award, transformation of computer science

???



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)

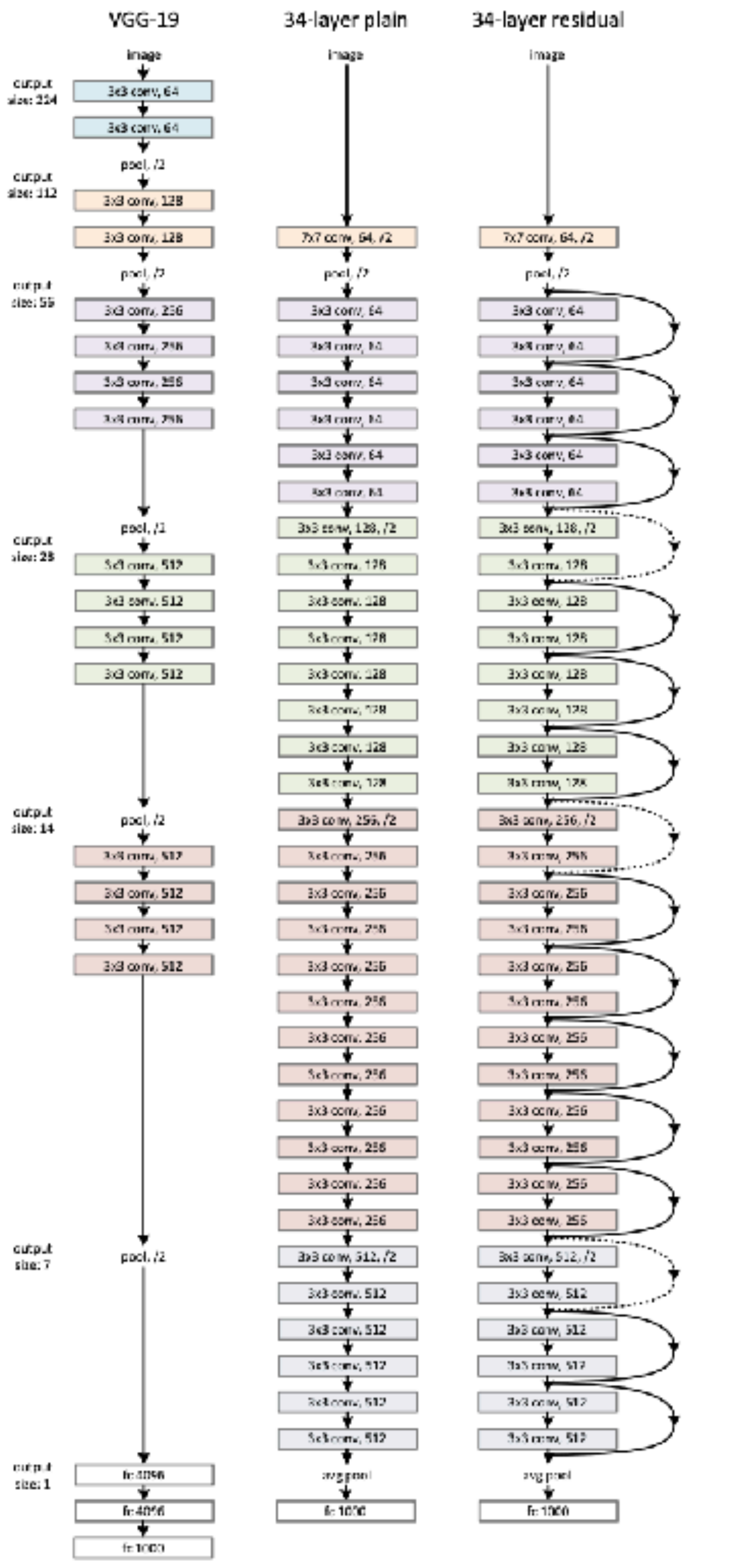
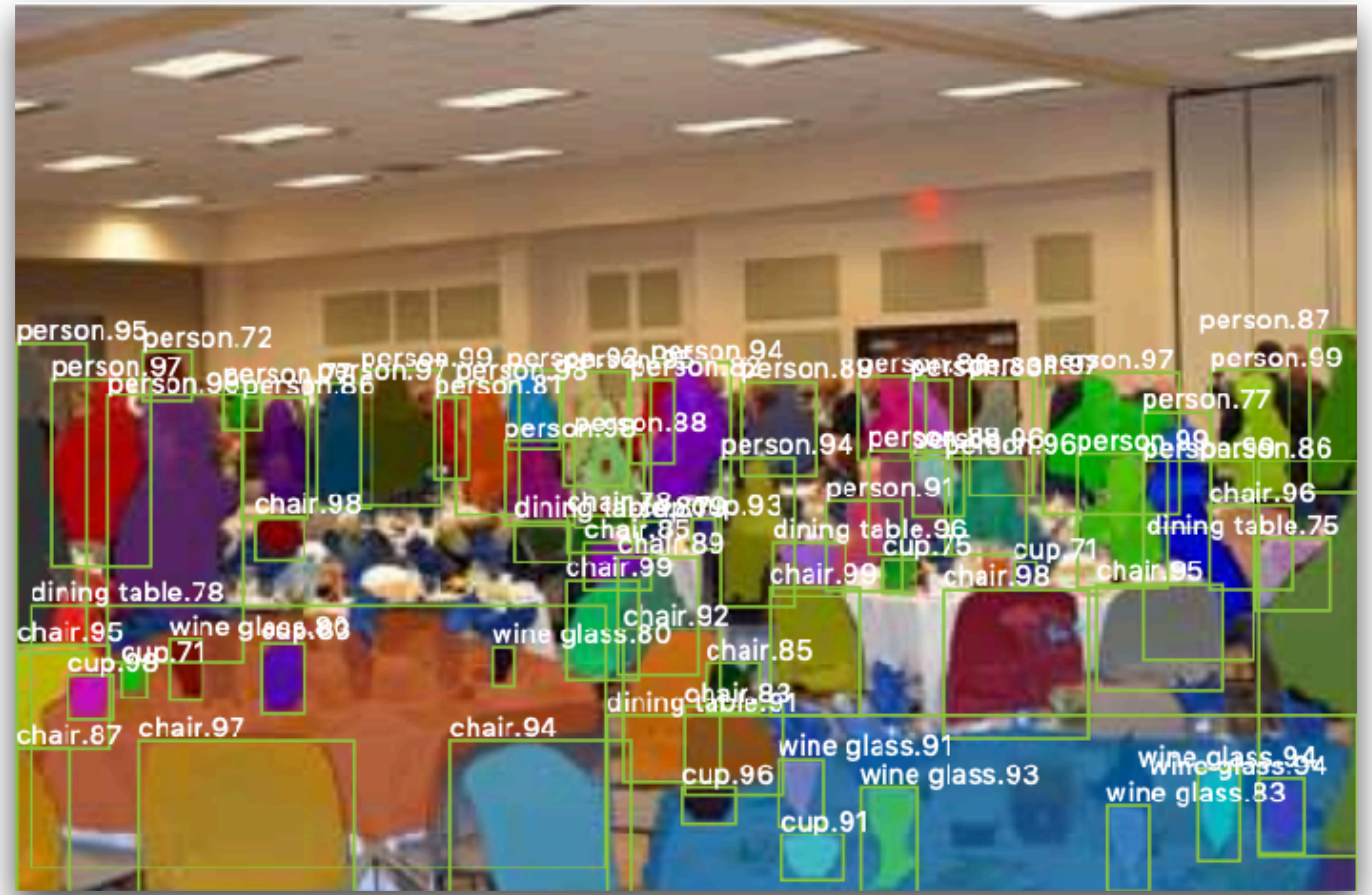


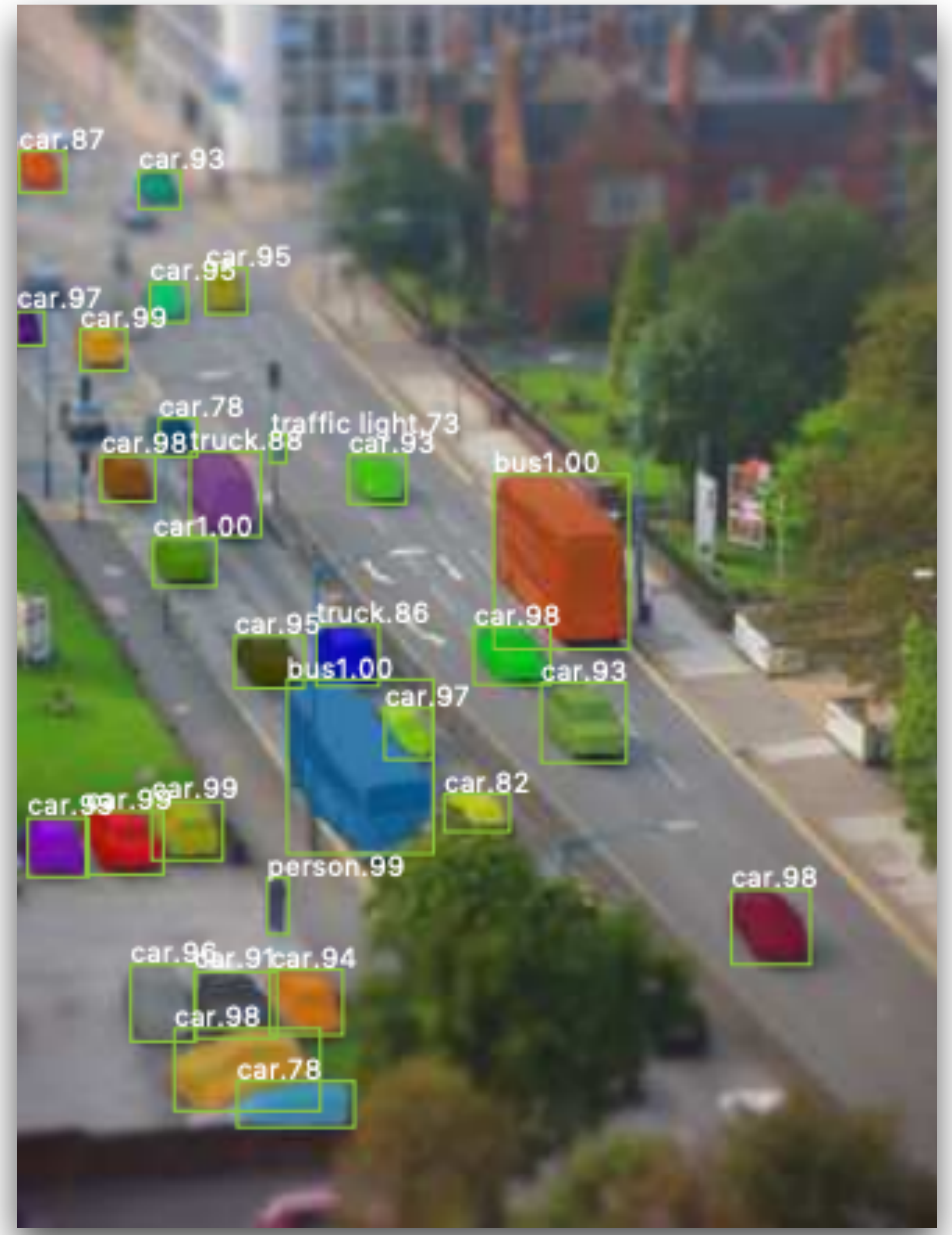
Figure 3. Example network architectures for ImageNet. **Left**: the VGG-19 model [41] (19.6 billion FLOPs) as a reference. **Middle**: a plain network with 34 parameter layers (3.6 billion FLOPs). **Right**: a residual network with 34 parameter layers (3.6 billion FLOPs). The dotted shortcuts increase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

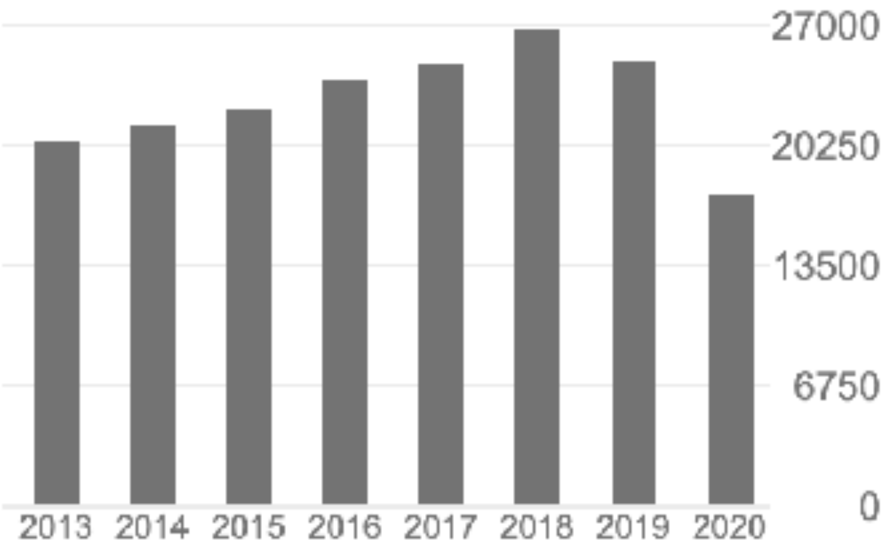
British National Library
Verified email at tsn.at
Kapitalismuskritiker Marxist Religionskritiker Philosophie Soziologie

FOLLOW

TITLE	CITED BY	YEAR
Le capital K Marx Librairie du progrès	38580	1875
Capital: volume I K Marx Penguin UK	19350 *	2004
The communist manifesto K Marx, F Engels Penguin	11661	2002
The german ideology K Marx, F Engels International Publishers Co	11652	1970
Grundrisse: Foundations of the critique of political economy K Marx 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 K Marx, F Engels Boitempo editorial	8366	2015
Das kapital K Marx 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

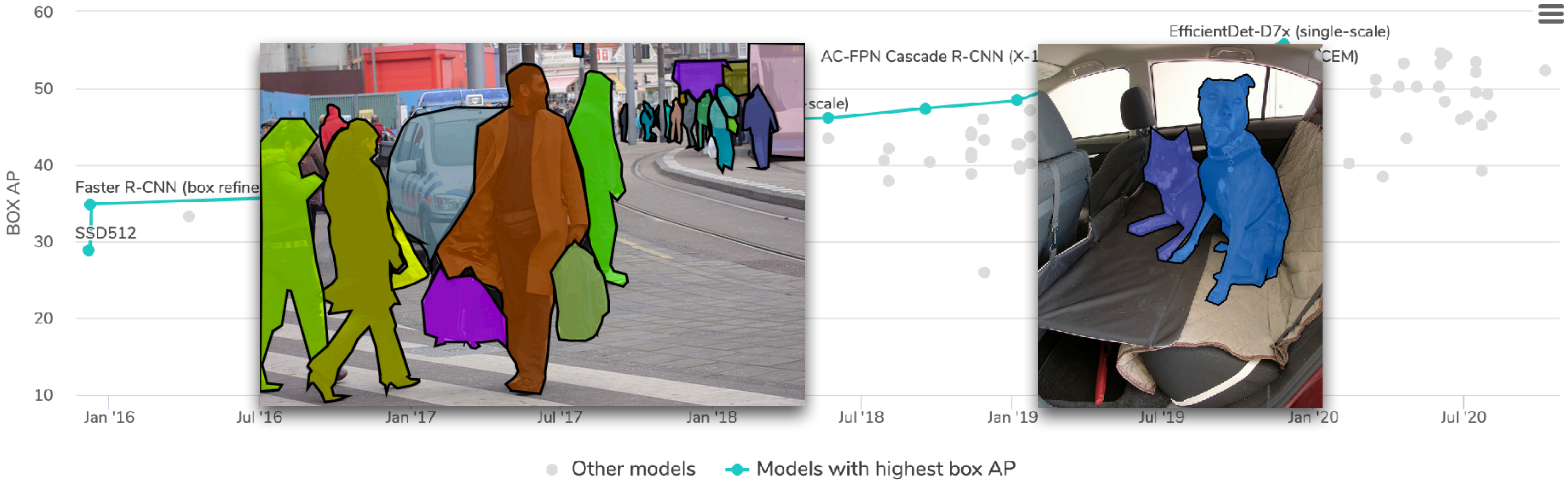
10k more than Marx!

CAVEAT: DO NOT MEASURE SCIENCE BY CITATION COUNT

Similar Performance Trends for Many Other Datasets

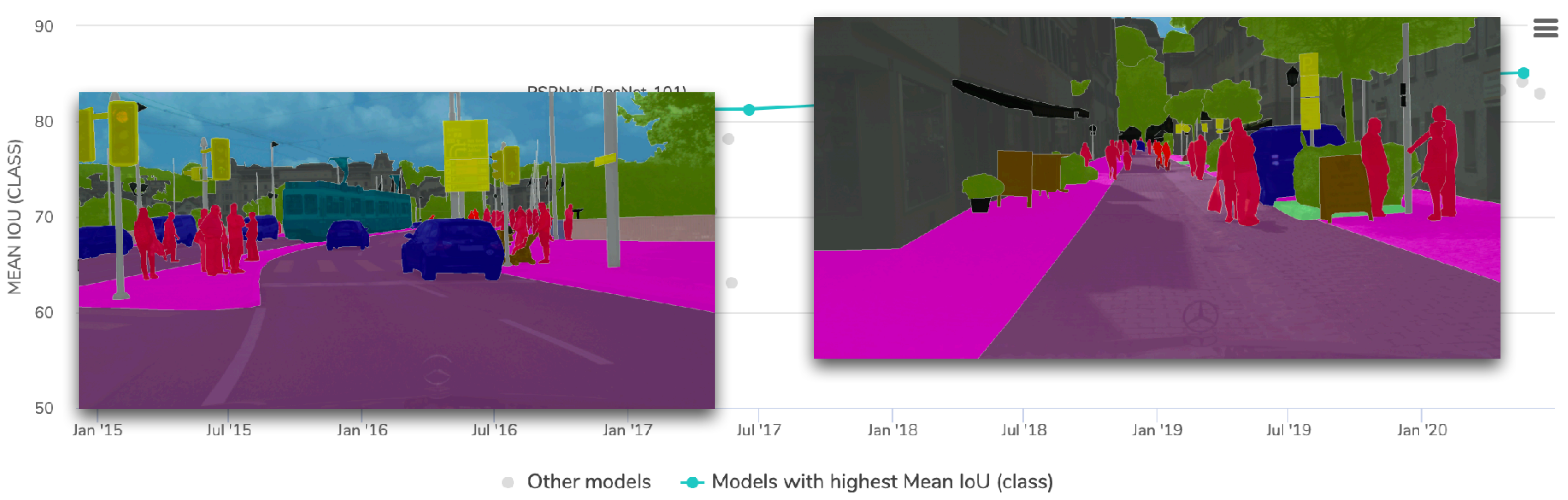
Object detection (PASCAL VOC)

Object Detection (MS COCO)

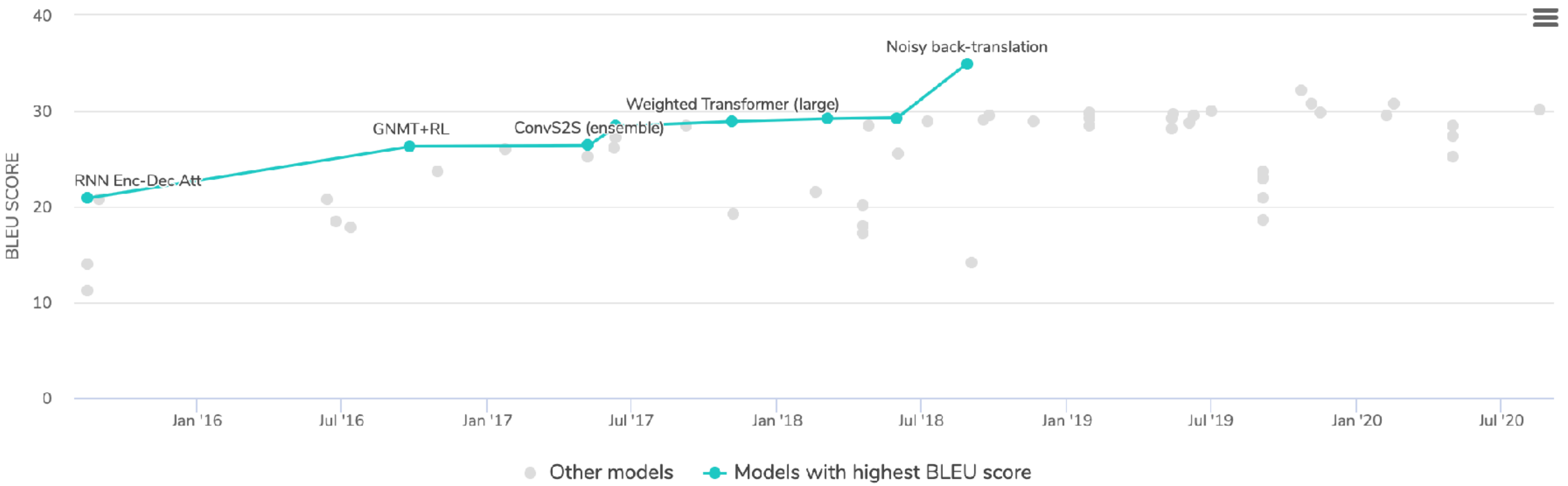


https://paperswithcode.com/sota

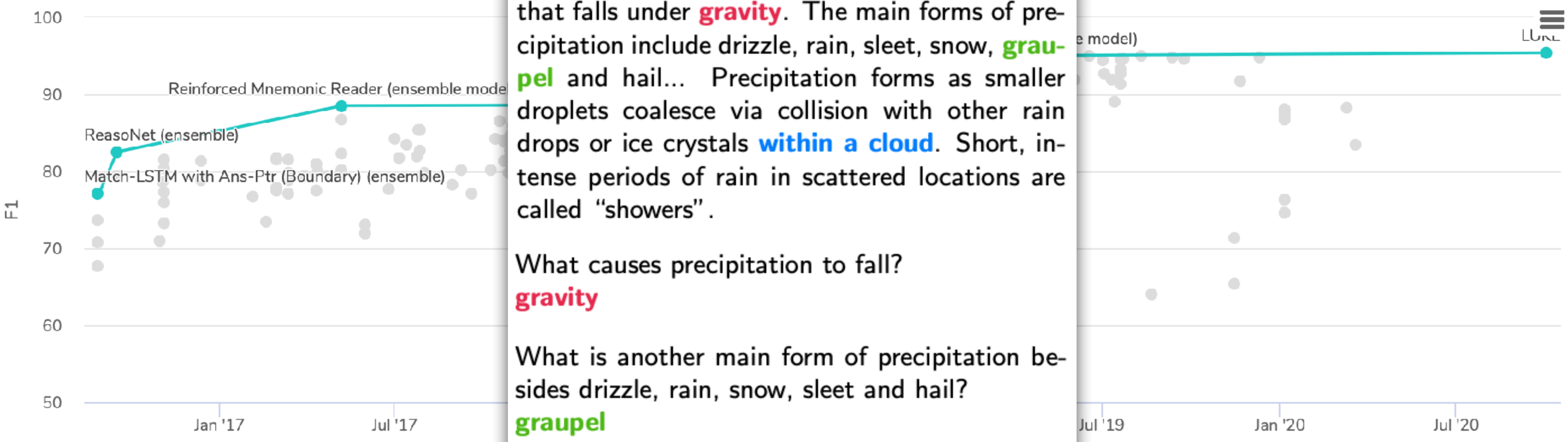
Semantic Segmentation (Cityscapes)



Machine Translation (WMT EN-DE)



Question Answering (SQuAD 1.1)



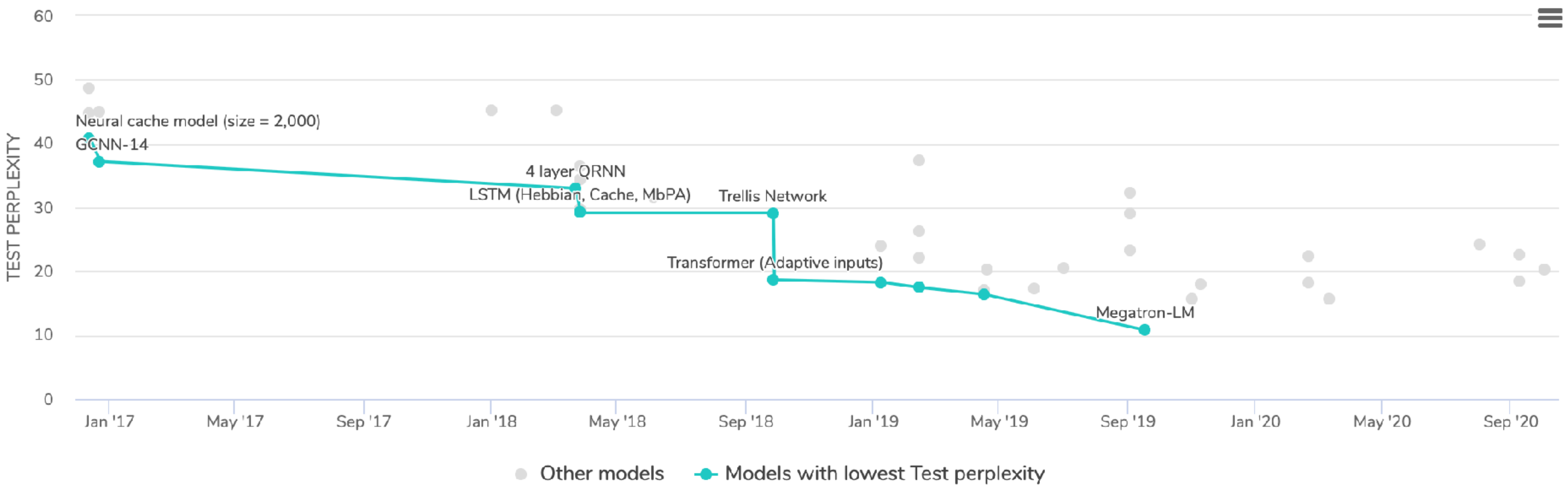
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or 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

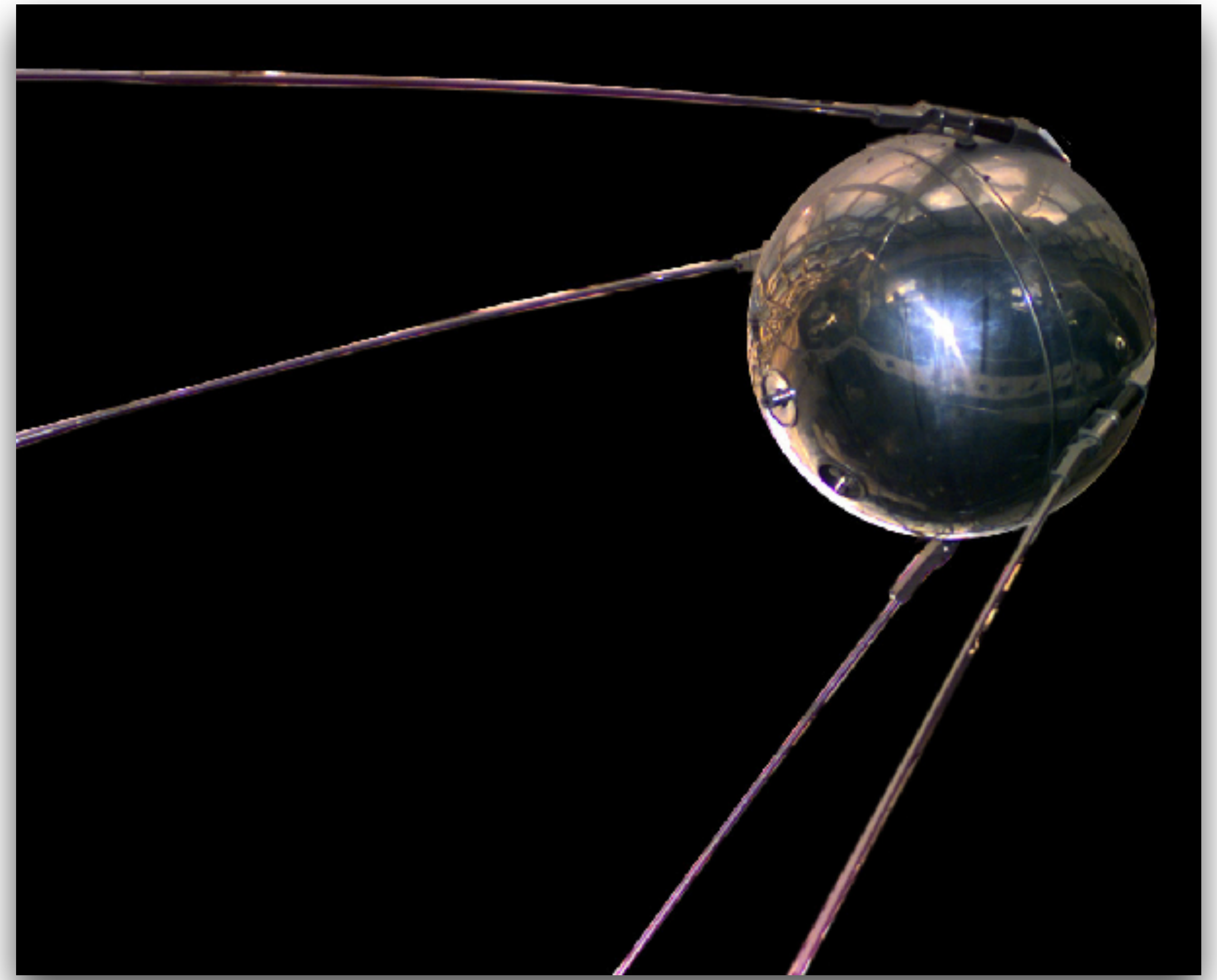
Culture shift

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

History of Benchmarking in ML

1960s: large investments in science and technology

(Result of Sputnik, etc.)



Speech recognition and translation get a lot of attention,

are glamorous fields, and attract funding.

But **results are lacking**

John R. Pierce (1910 - 2002)

Director of research at Bell Labs

Co-invented **pulse code modulation**, managed the team that invented the **transistor** (and invented the name), led development of first commercial **communications satellite**, etc.

Did not like AI and wrote about it in the ALPAC report and “Whither Speech Recognition?”

ALPAC Report (1964 - 1966)

Automatic **L**anguage **P**rocessing **A**dvisory **C**ommittee: 7 researchers led by Pierce

Established by the US government to evaluate potential of machine translation for various government agencies (mostly defense / science focused (Russian journals)).

Negative conclusions for machine translation, recommends investment in computational linguistics instead

➡ No government funding for machine translation for 10 - 20 years

“Whither Speech Recognition?” (1969)

Again John Pierce, this time a single-author short 1.5 page letter to the Journal of the Acoustical Society of America

Very critical of speech recognition research

> *“We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn’t attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses* ***deceit and offers glamour****.”*

➡ No funding for speech recognition for 10 - 20 years

Quote from “Whither Speech Recognition?”

*Most recognizers behave, not like scientists, but like **mad inventors** or **untrustworthy engineers**. The typical recognizer gets it into his head that he can solve “the problem.” The basis for this is either individual inspiration (the “mad inventor” source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach). . . .*

*The typical recognizer . . . builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. **No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment.***

Quote from “Whither Speech Recognition?”

*It is clear that **glamor and any deceit** in the field of speech recognition blind the takers of funds as much as they blind the givers of funds. What particular considerations have led to this enthusiasm? [...]*

Turing asked, On what basis can we say that a machine thinks? His perfectly rational answer was that if, in conversing with a machine, we cannot tell whether it is a human being or a machine, then we can scarcely deny that the machine thinks. [...]

*We should consider, however, that **in deception, studied and artful deceit is apt to succeed better and more quickly than science**.*

Bringing Funding for Translation and Speech Recognition Back

Two people were key in resuming government funding for speech and translation in the mid to late 80s:

Fred Jelinek: research manager at IBM

Charles Wayne: program manager at DARPA

Key idea: make evaluations “glamour and deceit”-proof

Fred Jelinek

PhD in information theory (Fano)

Led IBM's effort on the "**general dictation problem**" from 1972 to 1980

Advocate for comparing the **quantitative performance** of alternative algorithms on **test sets**, using fixed and automatically calculated **evaluation metrics**.

Also strongly in favor of **sharing datasets, evaluation metric, algorithms**, etc.

Same approach for machine translation and other problems in his group.

"*Every time I fire a linguist, the performance of the speech recognizer goes up.*"

Charles Wayne

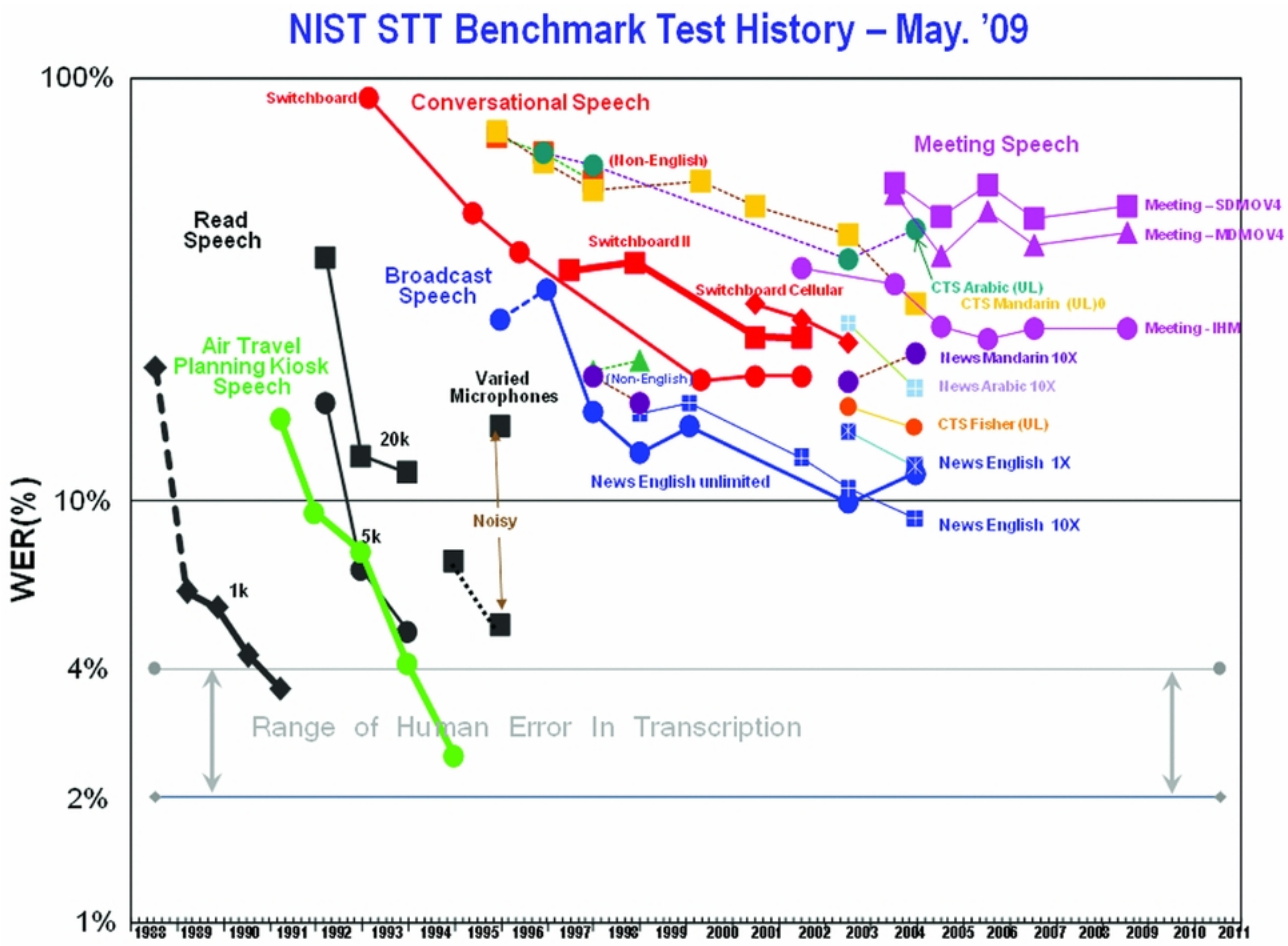
DARPA program manager responsible for funding restart in 1986

Key idea: emphasize evaluation. **Well-defined objective** evaluation, applied by a **neutral agent** (NIST) on **shared datasets** (often Linguistic Data Consortium)

Initially both Pierce-style engineers and speech researchers were skeptical, but the approach was successful

“Glamour and deceit”-proof, funders could measure progress

Speech Recognition Benchmarks



Also in 1987:

David Aha creates the **UCI dataset repository**

ML community adopts benchmark paradigm

Summary

Progress on key benchmarks, especially ImageNet

Empirically motivated methods **outperform** theoretically grounded methods

Shift towards **benchmark-driven research** in machine learning over the past 10 years

1. Empirical progress in machine learning: benchmarks

2. What can we learn from ML benchmarks?

3. Limitations of current ML methods

Caveats with Benchmarks

A: Are new methods really better? What about the methods we already had?

B: Are we just overfitting to the benchmark test sets?

C: Do we have progress beyond the immediate benchmark?

If we don't have proofs any more, our experiments better be rock-solid!

Caveats with Benchmarks

A: Are new methods really better? What about the methods we already had?

B: Are we just overfitting to the benchmark test sets?

C: Do we have progress beyond the immediate benchmark?

What about Kernels?

Lots of insightful theory, Gaussian kernel SVM was / is competitive on many tasks

Could we have "solved" ImageNet with kernels?

Ben Recht

Counterfactuals here are hard

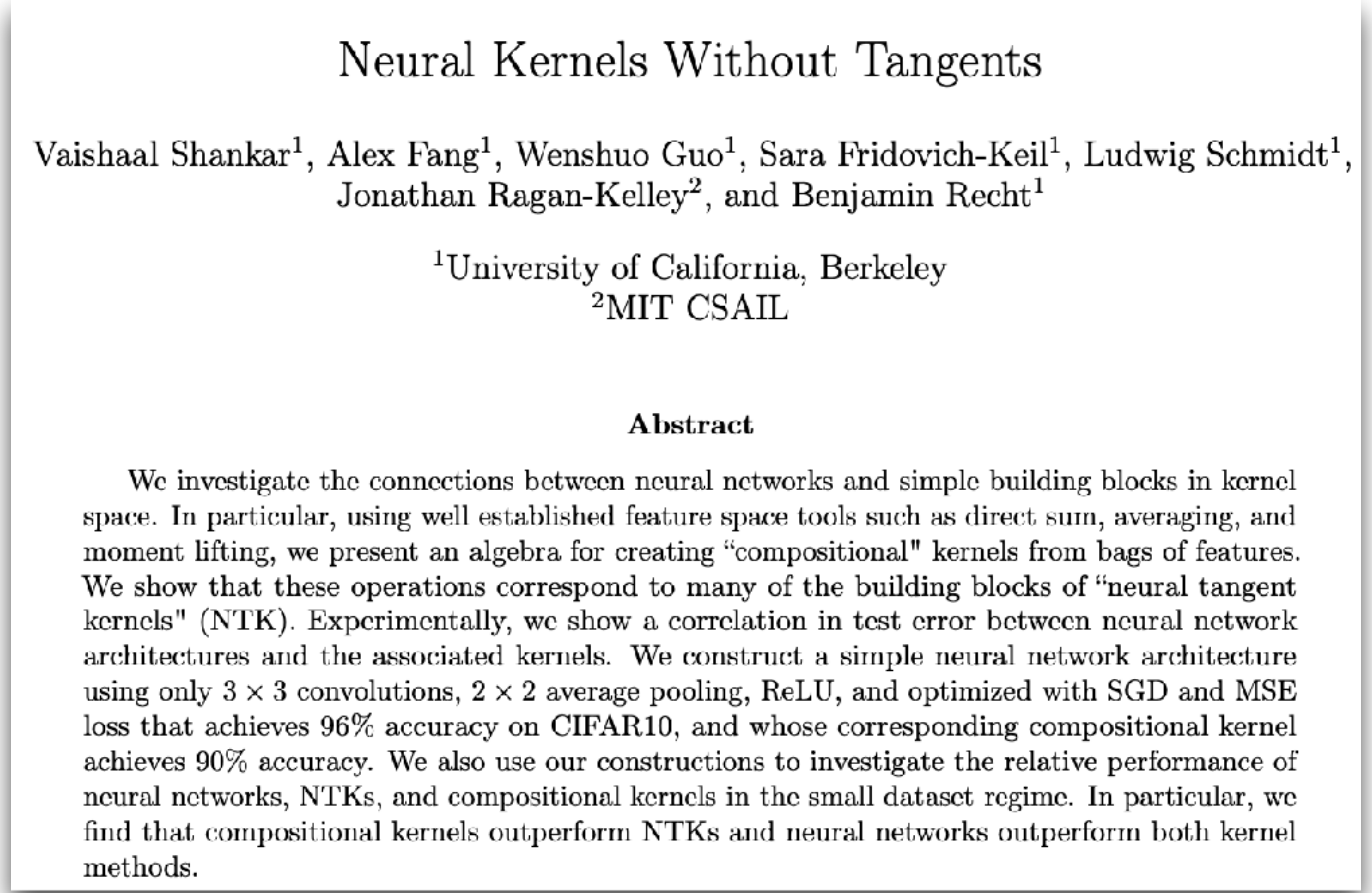
- Deep learning requires lots of engineering
- Major community effort

Ben and Vaishaal worked on this for multiple years

Vaishaal Shankar

90% accuracy on CIFAR-10

AlexNet had 89% in 2012



Neural Kernels Without Tangents

Vaishaal Shankar[1], Alex Fang[1], Wenshuo Guo[1], Sara Fridovich-Keil[1], Ludwig Schmidt[1], Jonathan Ragan-Kelley[2], and Benjamin Recht[1]

[1]University of California, Berkeley
[2]MIT CSAIL

Abstract

We investigate the connections between neural networks and simple building blocks in kernel space. In particular, using well established feature space tools such as direct sum, averaging, and moment lifting, we present an algebra for creating "compositional" kernels from bags of features. We show that these operations correspond to many of the building blocks of "neural tangent kernels" (NTK). Experimentally, we show a correlation in test error between neural network architectures and the associated kernels. We construct a simple neural network architecture using only 3×3 convolutions, 2×2 average pooling, ReLU, and optimized with SGD and MSE loss that achieves 96% accuracy on CIFAR10, and whose corresponding compositional kernel achieves 90% accuracy. We also use our constructions to investigate the relative performance of neural networks, NTKs, and compositional kernels in the small dataset regime. In particular, we find that compositional kernels outperform NTKs and neural networks outperform both kernel methods.

Kernel is CNN-inspired

87% with two-layer kernels

Computationally expensive

100x more than a CNN (but unfair)

No published results on ImageNet

Currently best kernel on CIFAR-10

Better than any NTK!

At least we know beating CNNs with kernels is not easy.

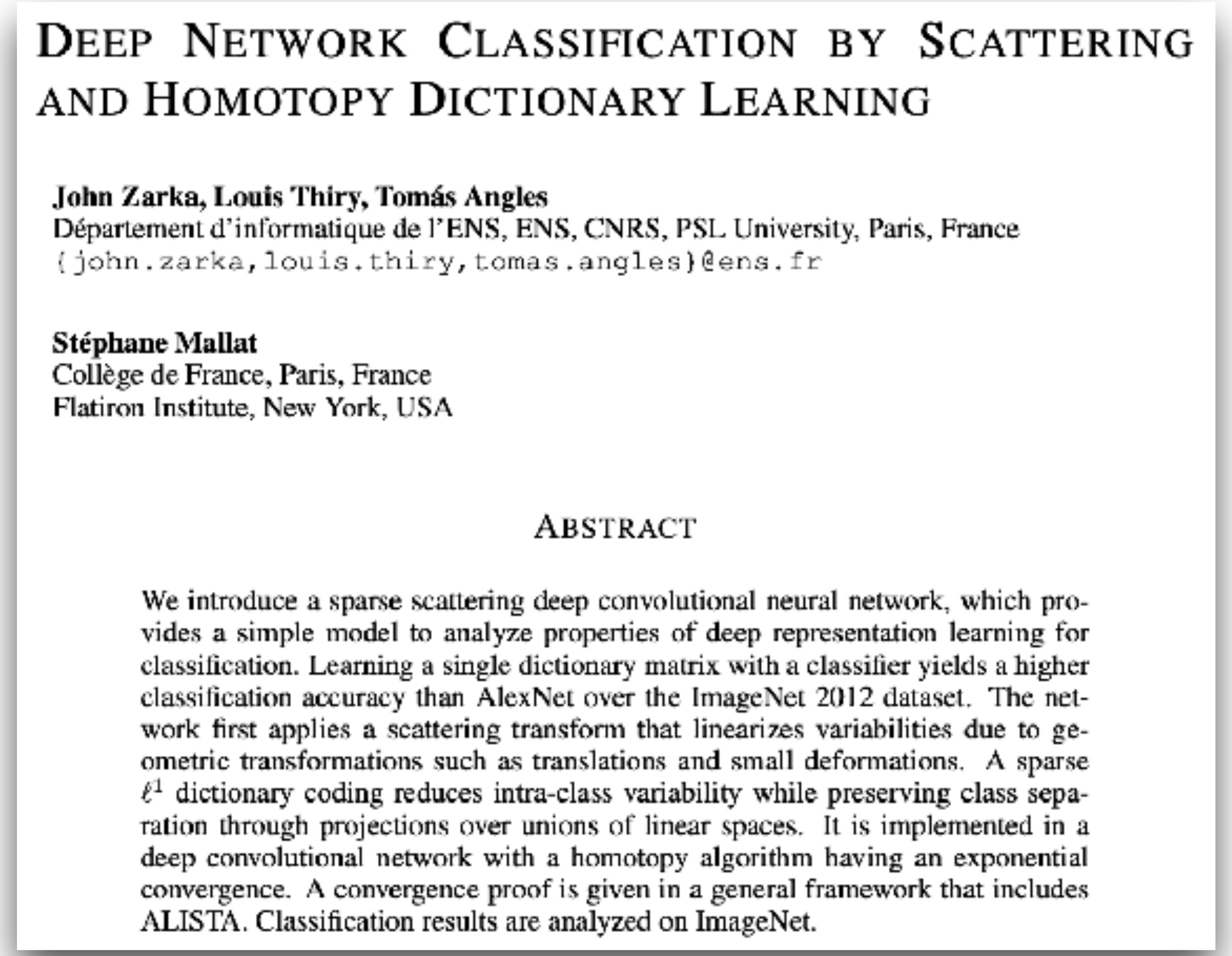
What about Wavelets?

Another image representation. Very active in signal processing in the 90s.

Multi-layer variant: **scattering transform** (2013)

Stephane Mallat

Also multiple years of work, currently culminating in:



DEEP NETWORK CLASSIFICATION BY SCATTERING AND HOMOTOPY DICTIONARY LEARNING

John Zarka, Louis Thiry, Tomás Angles
Département d'informatique de l'ENS, ENS, CNRS, PSL University, Paris, France
{john.zarka,louis.thiry,tomas.angles}@ens.fr

Stéphane Mallat
Collège de France, Paris, France
Flatiron Institute, New York, USA

ABSTRACT

We introduce a sparse scattering deep convolutional neural network, which provides a simple model to analyze properties of deep representation learning for classification. Learning a single dictionary matrix with a classifier yields a higher classification accuracy than AlexNet over the ImageNet 2012 dataset. The network first applies a scattering transform that linearizes variabilities due to geometric transformations such as translations and small deformations. A sparse ℓ^1 dictionary coding reduces intra-class variability while preserving class separation through projections over unions of linear spaces. It is implemented in a deep convolutional network with a homotopy algorithm having an exponential convergence. A convergence proof is given in a general framework that includes ALISTA. Classification results are analyzed on ImageNet.

Surpasses AlexNet-performance by **6 percentage points** (pp) in 2020.

In the meantime, CNN accuracy has improved by **32 pp**.

Joan Bruna

ImageNet & Co are solid so far

But: Not Everything Neural is Good!

Different Field: Recommender Systems

On the Difficulty of Evaluating Baselines

A Study on Recommender Systems

Steffen Rendle*
srendle@google.com

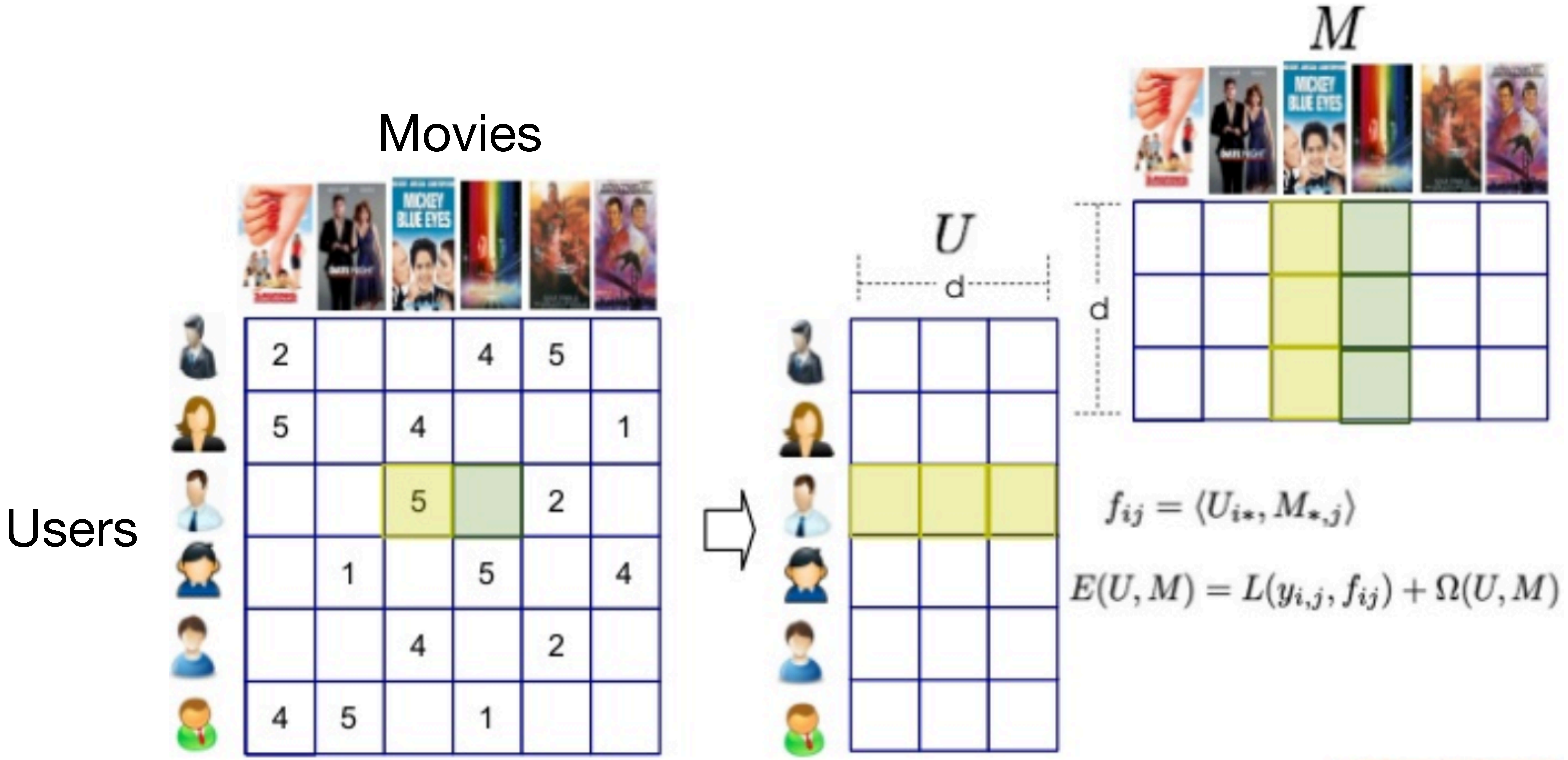
Li Zhang*
liqzhang@google.com

Yehuda Koren†
yehuda@google.com

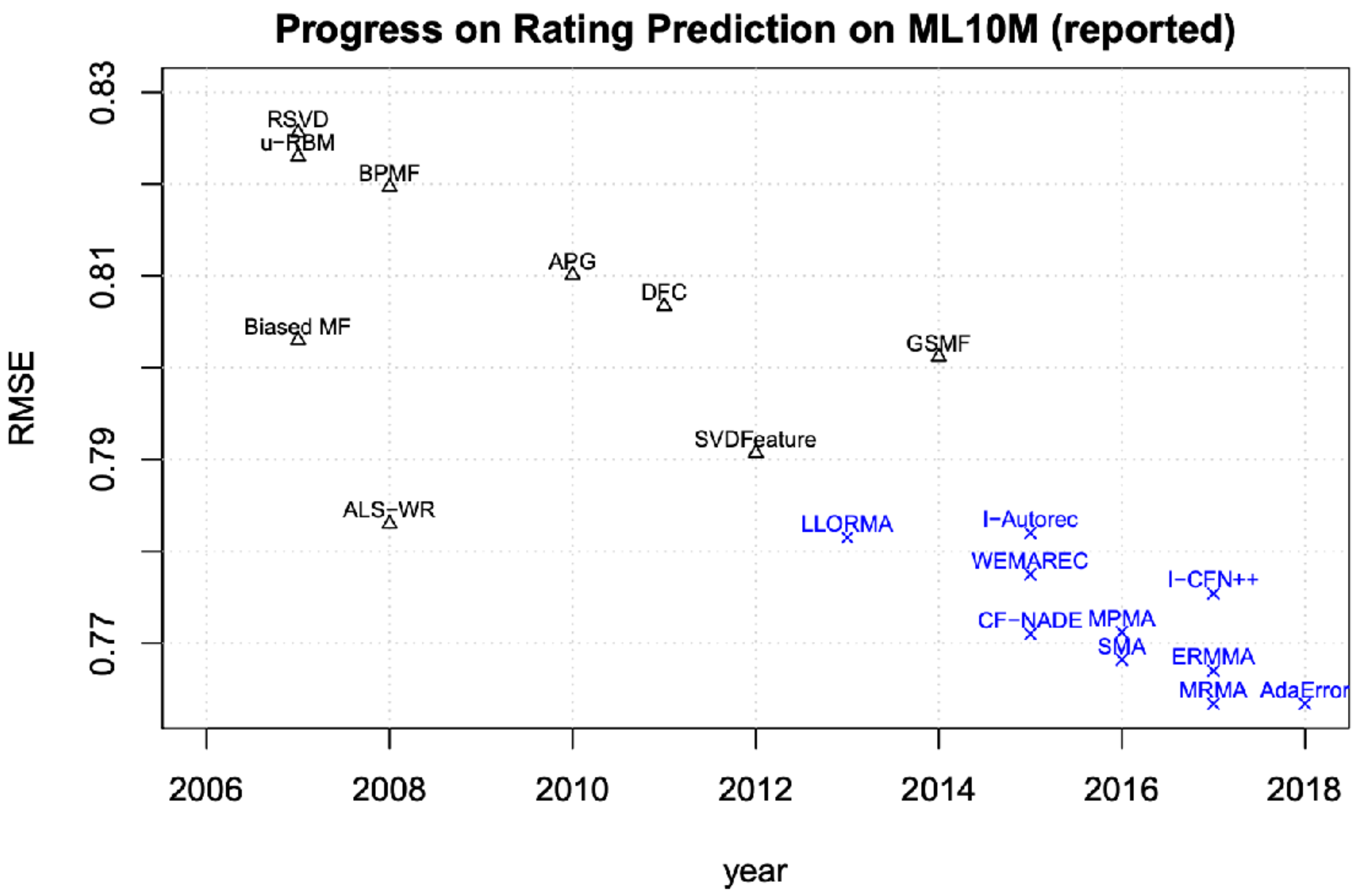
Abstract

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.

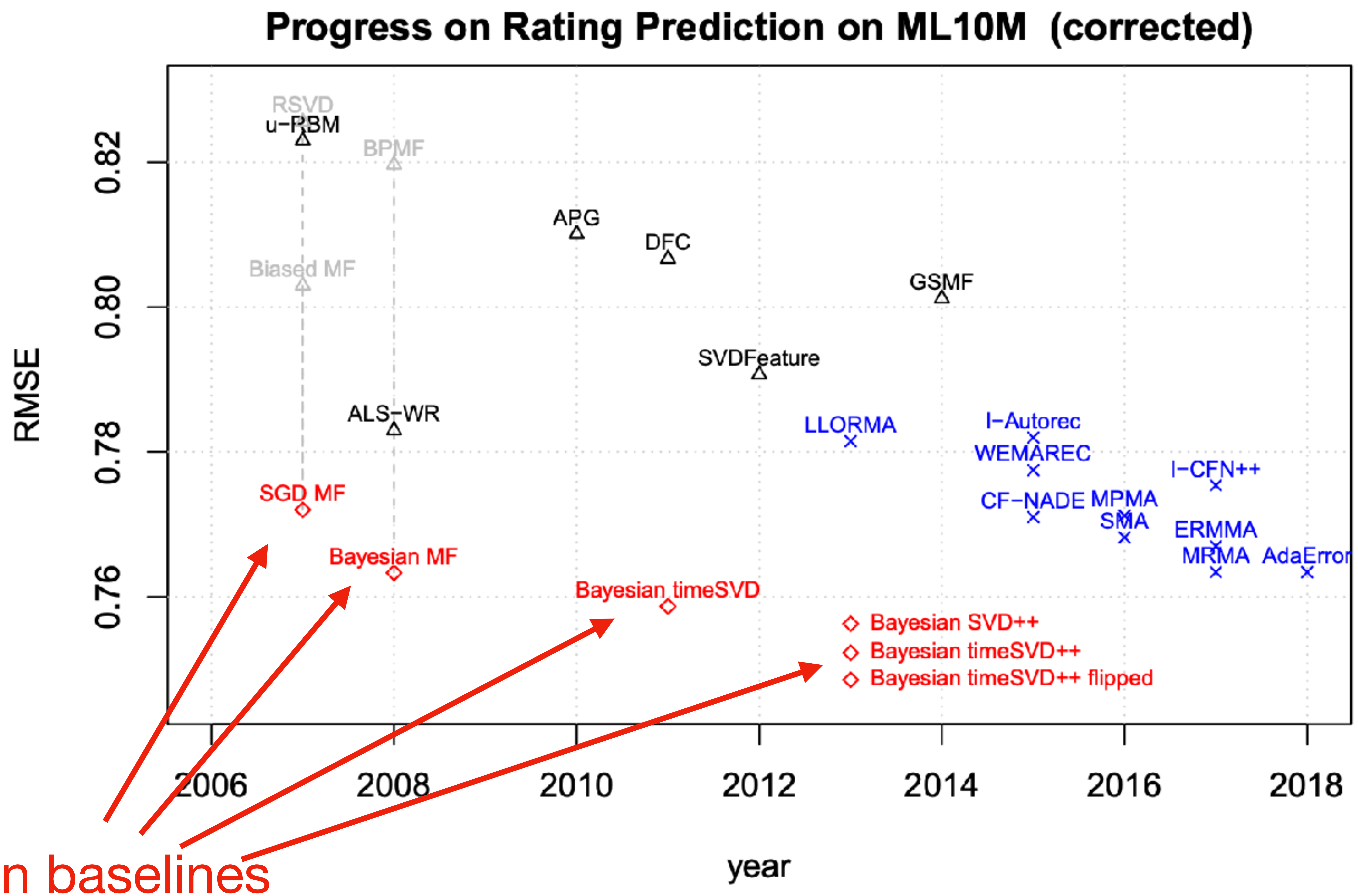
Recommender Systems & Matrix Factorization



“State of the Art”



Actual State of the Art



Known baselines

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.

Similar for major NLP benchmarks (but smaller datasets have quality problems)

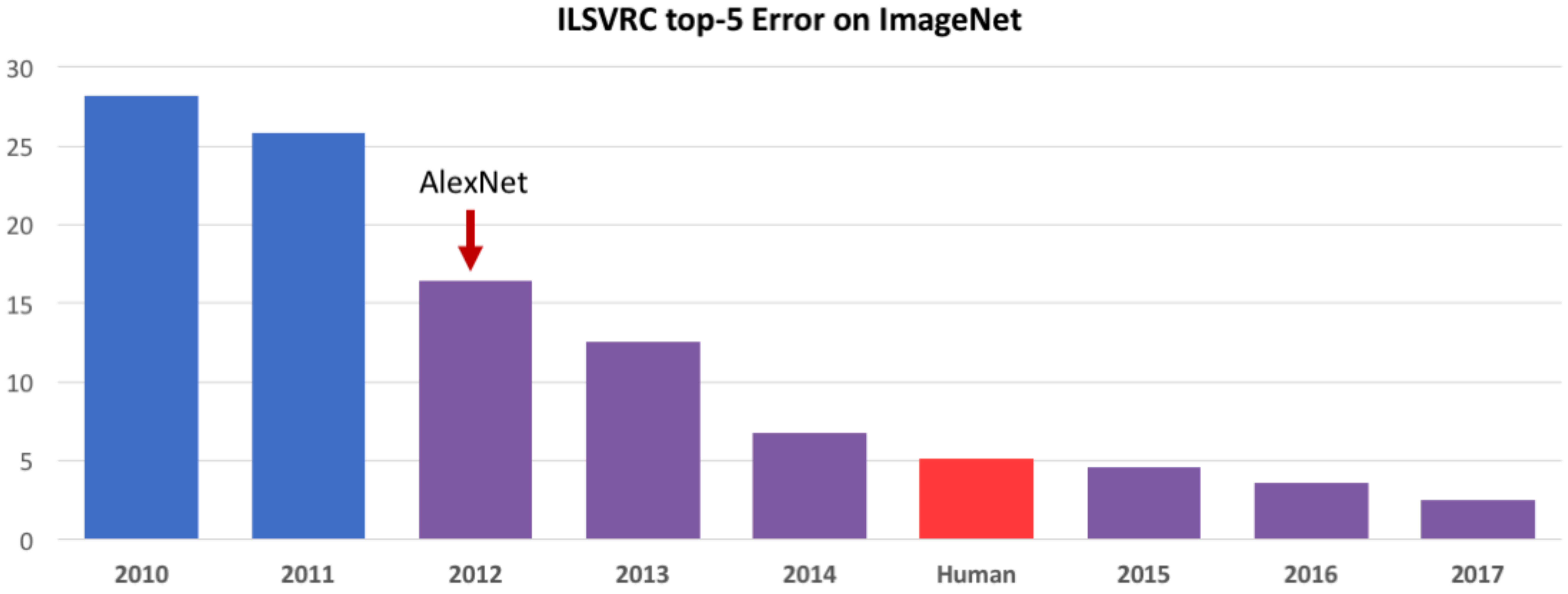
Caveats with Benchmarks

A: Are new methods really better? What about the methods we already had?

B: Are we just overfitting to the benchmark test sets?

C: Do we have progress beyond the immediate benchmark?

What are we Measuring with a Benchmark?



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

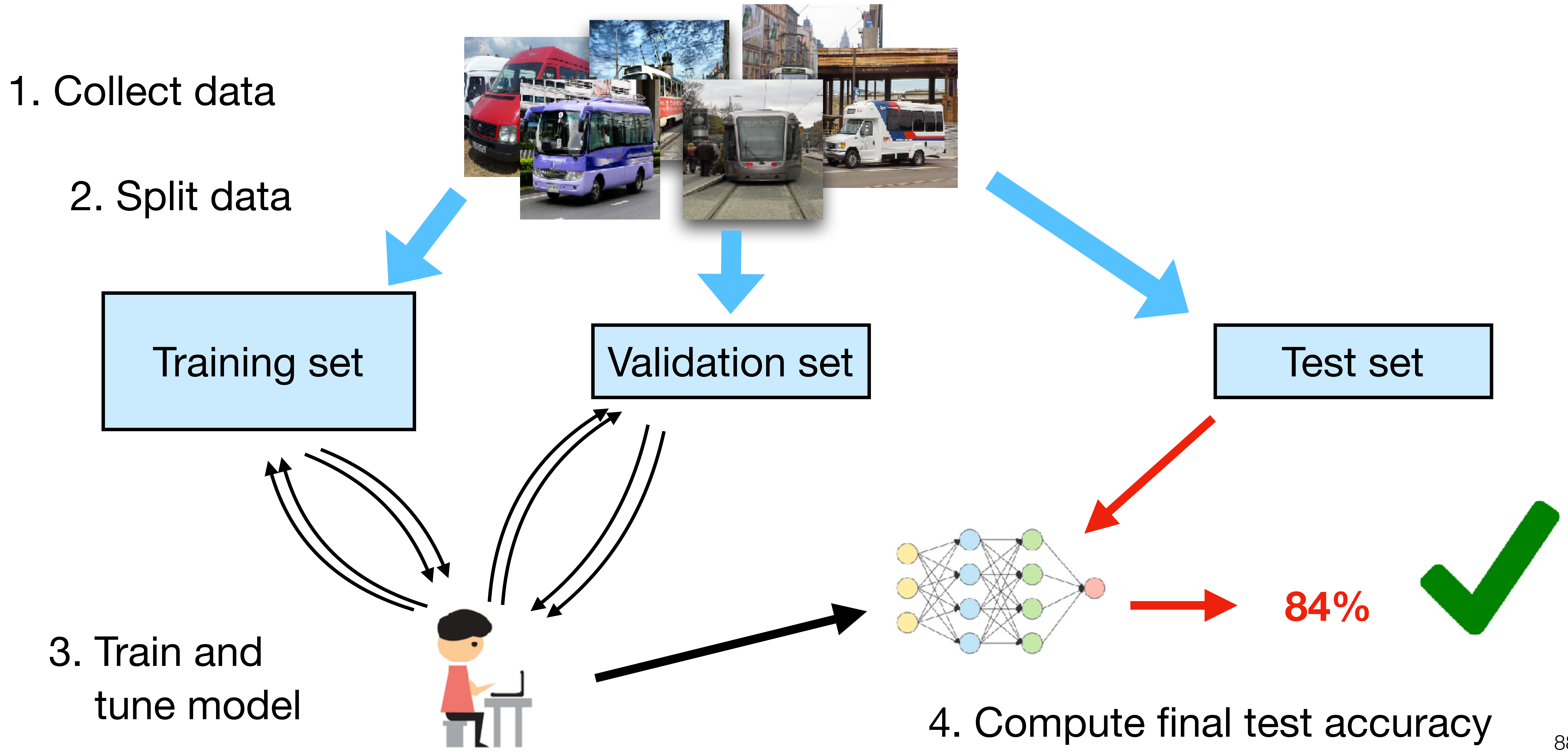
➡ What do we really care about?

Generalization

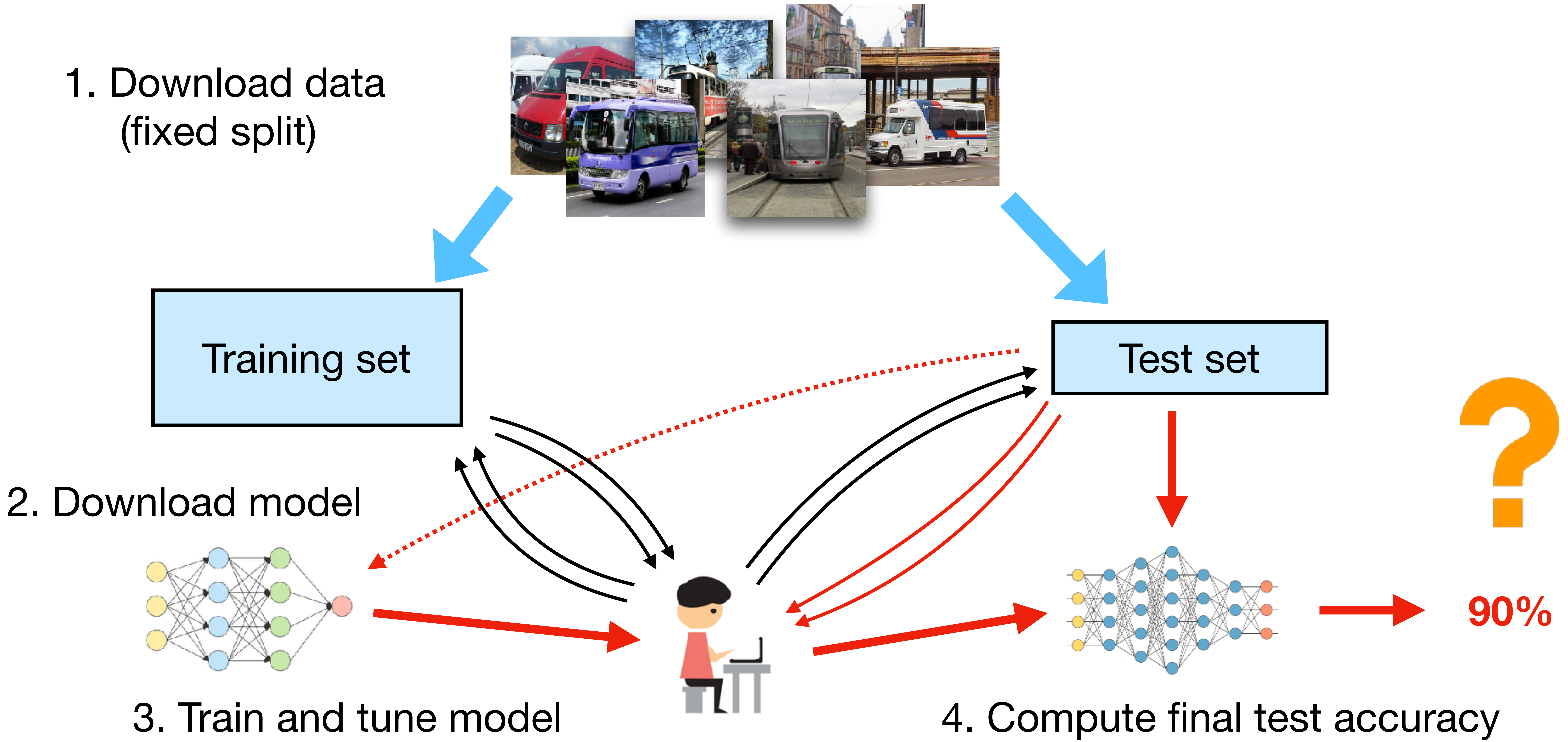
At least, the classifiers should perform similarly well on new data from the **same source**.

How can we reliably measure generalization?

Ideal ML Workflow

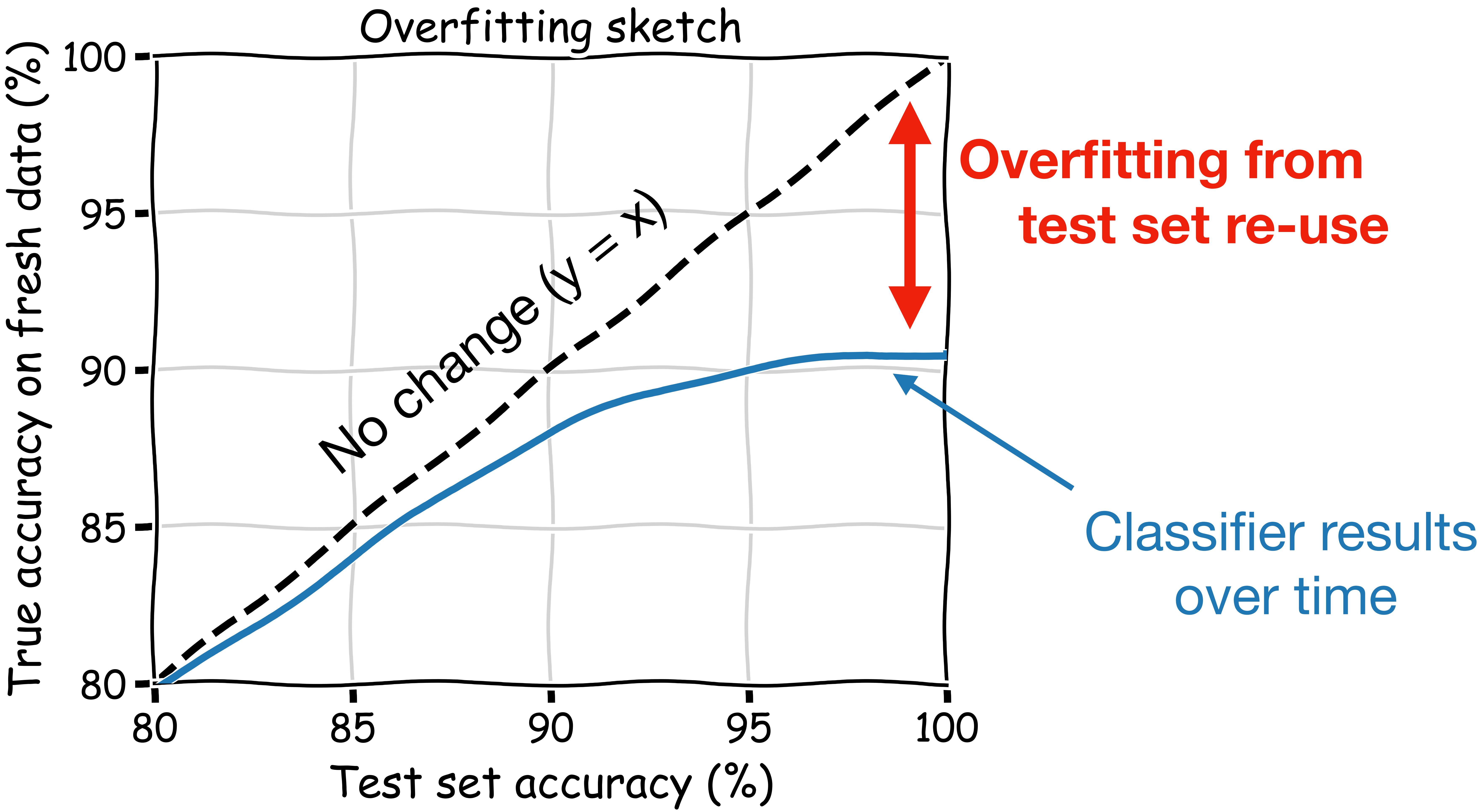


Typical ML Workflow



Danger with Test Set Re-Use: Overfitting

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

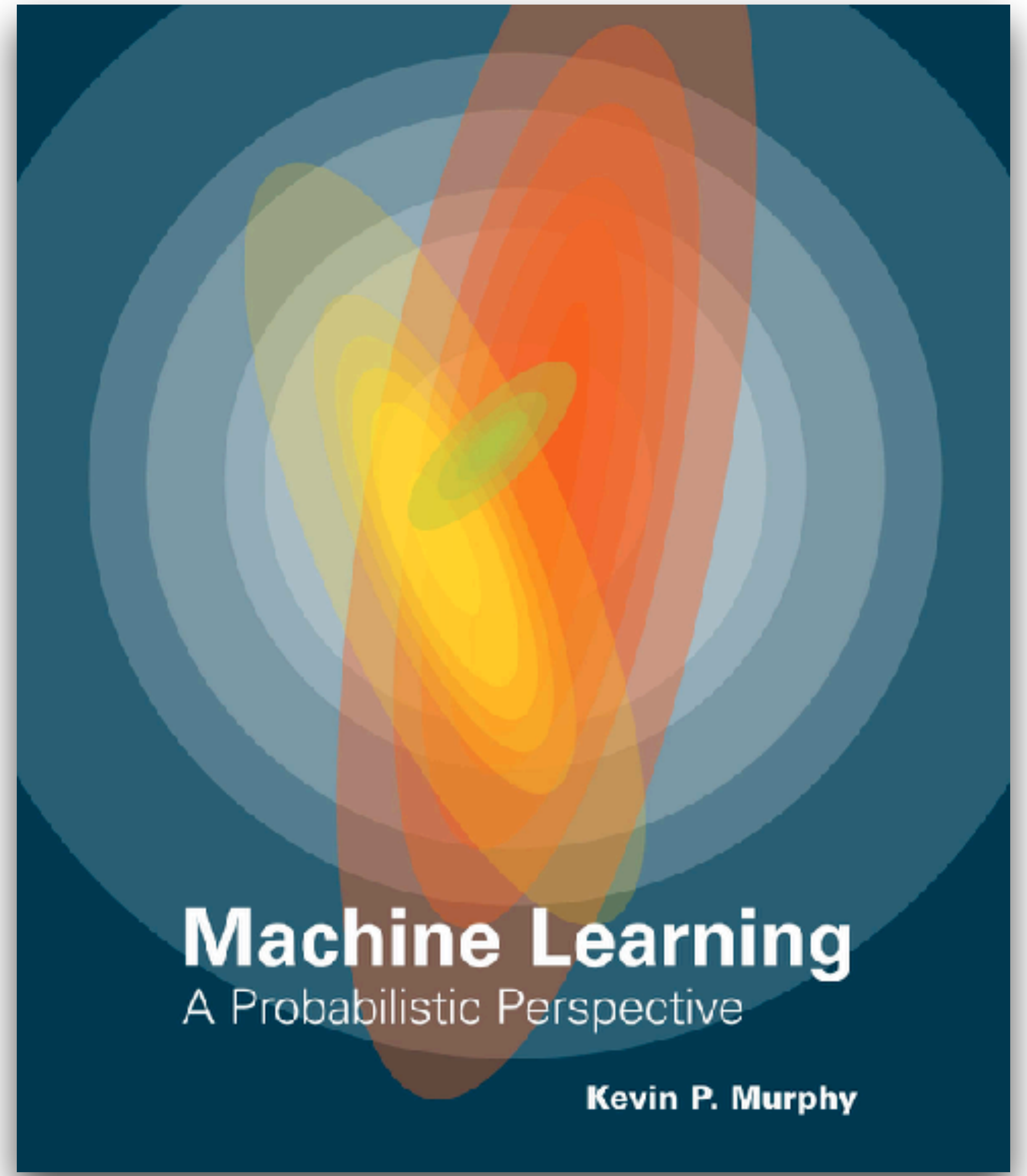


To be clear: We now know that there is no evidence of overfitting through test set re-use on many contemporary ML benchmarks (e.g., ImageNet)

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

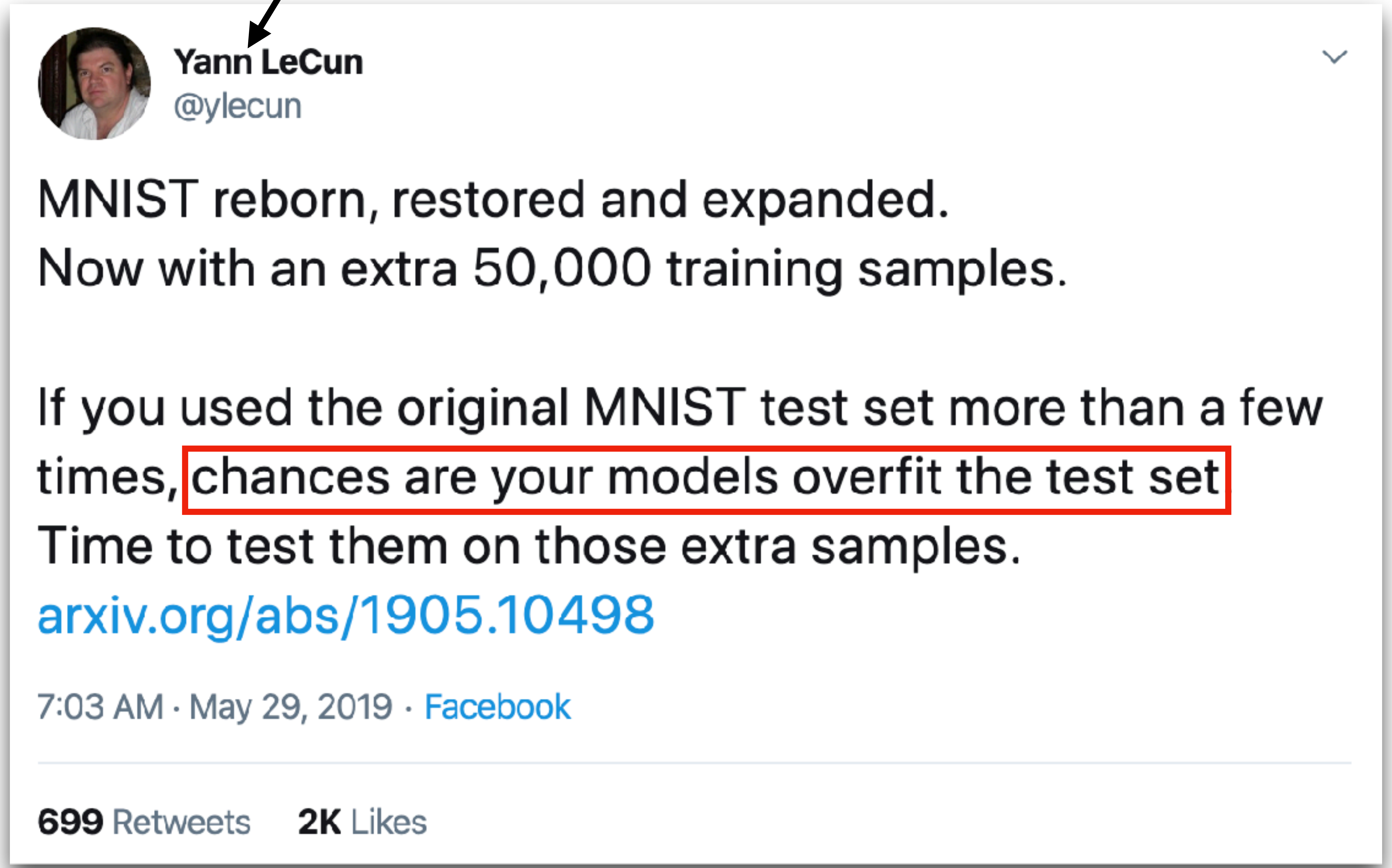
"Withholding the annotation of the test data until completion of the challenge played a significant part in ***preventing over-fitting*** *of the parameters of classification or detection methods. In the VOC2005 challenge, test annotation was released and this led to some* ***"optimistic" reported results,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



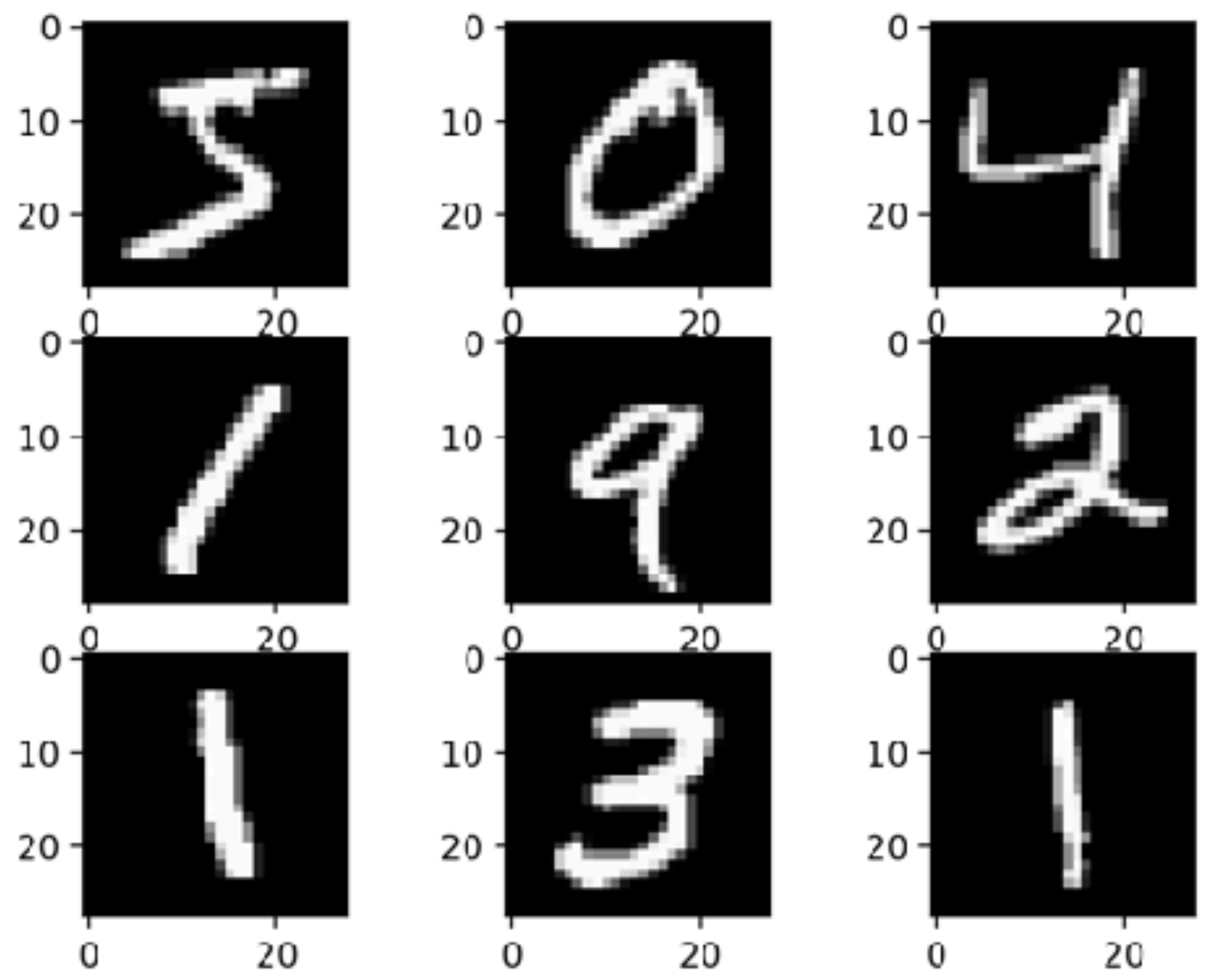
Yann LeCun
@ylecun

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

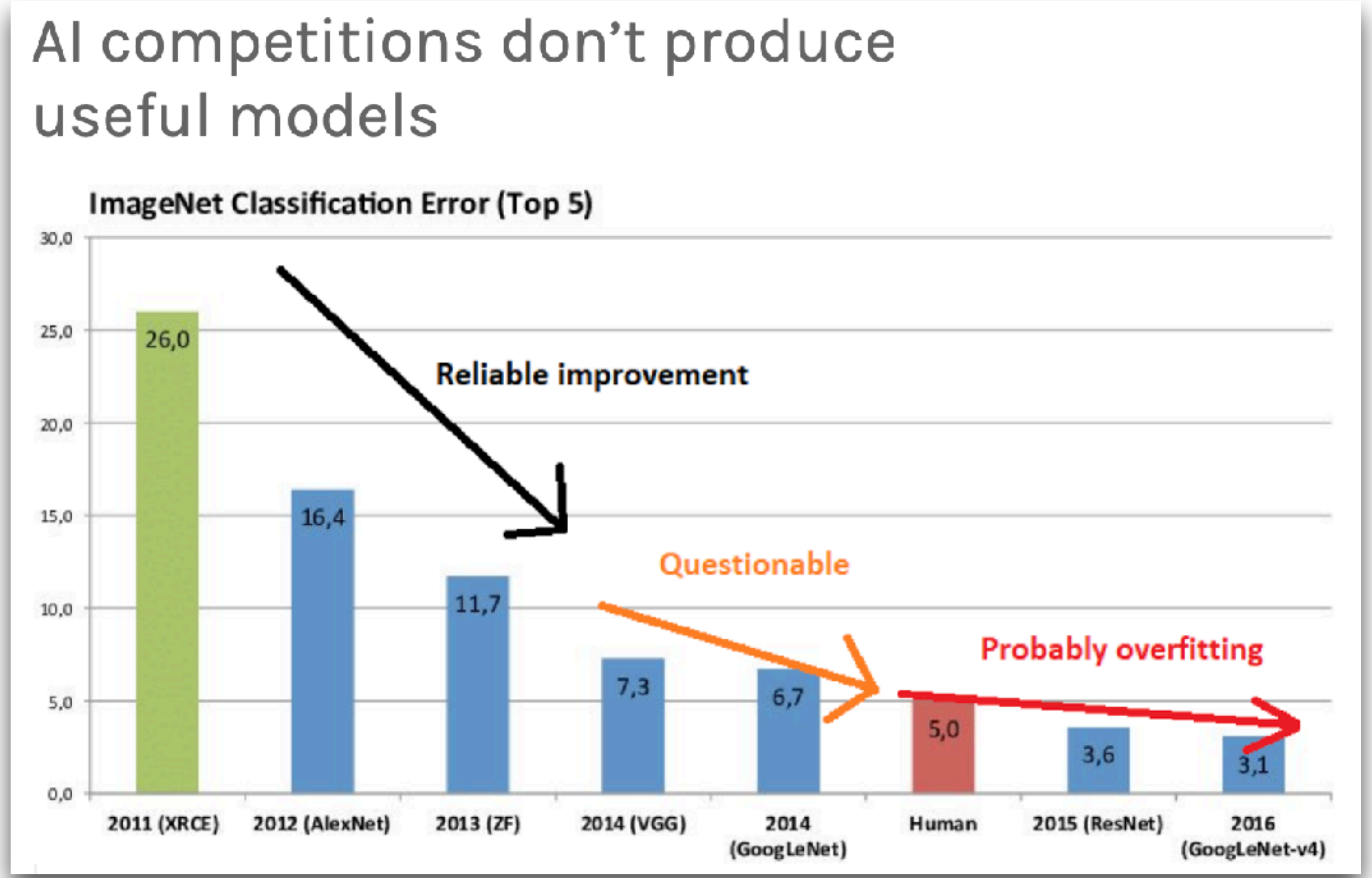
10 classes

Released in 1998

Oldest widely used dataset

Now considered “easy”

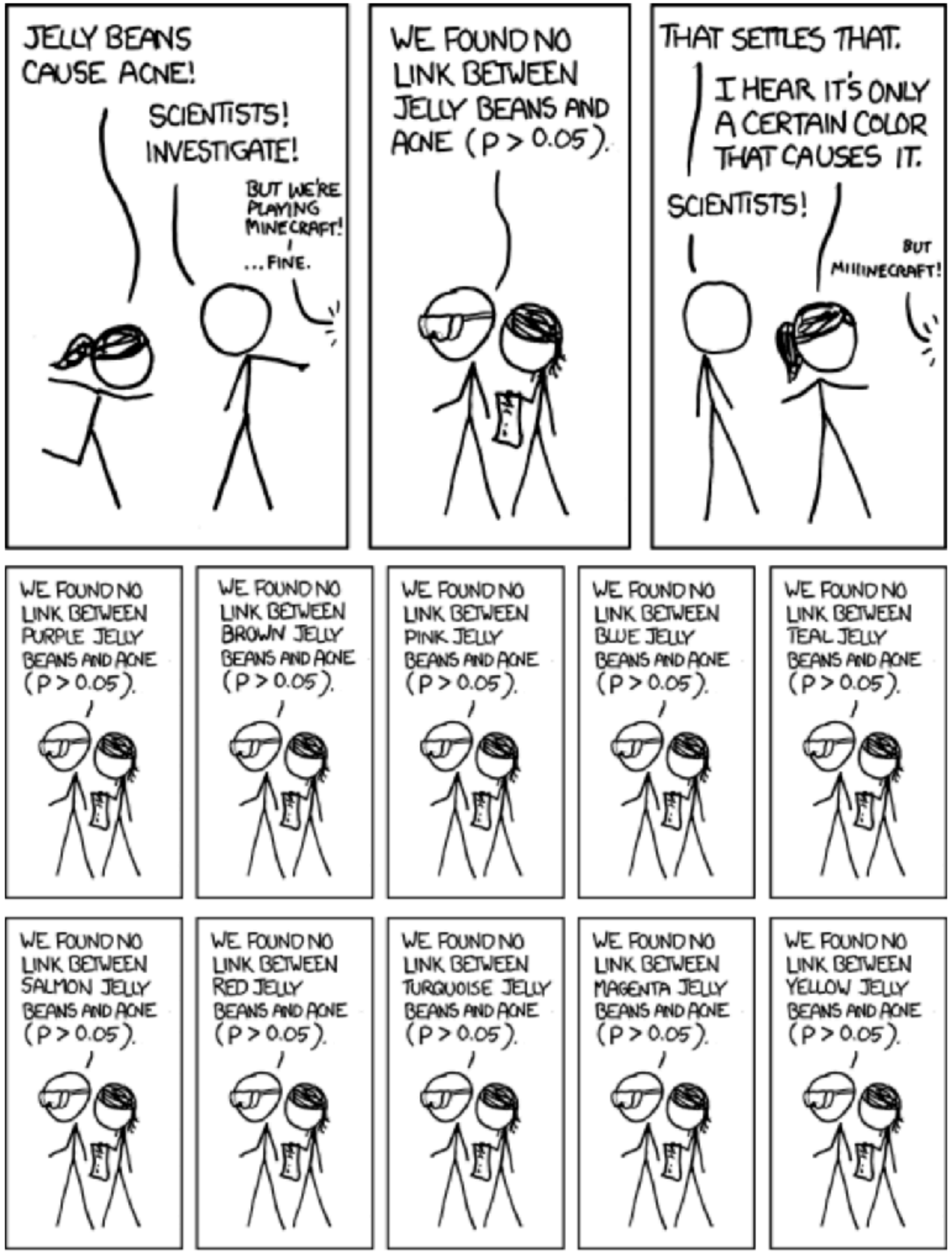
https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/

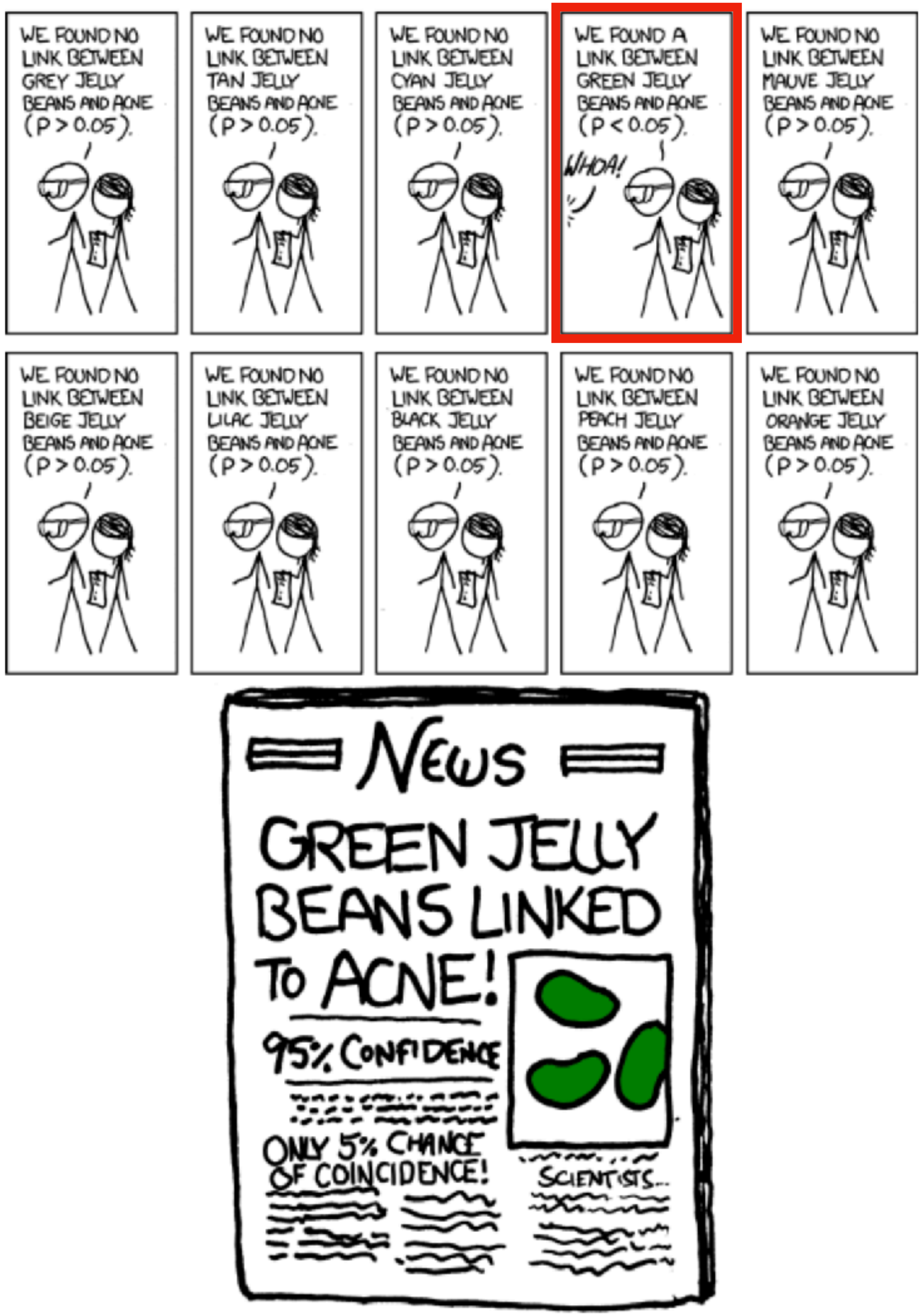


I can't really estimate the numbers, but knowing what we know about multiple testing does anyone really believe the SOTA rush in the mid 2010s was anything but crowdsourced overfitting?

Multiple hypothesis testing

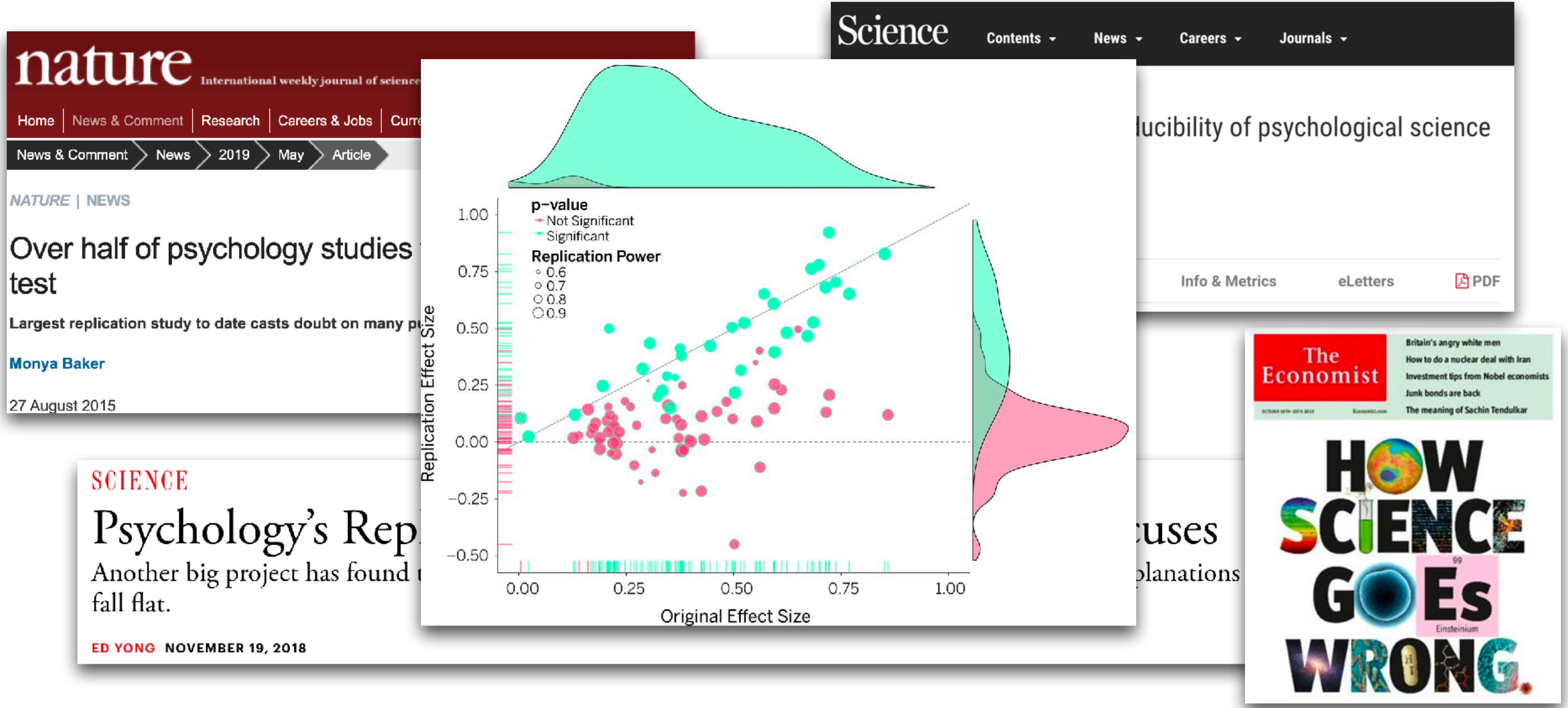
"p-hacking"



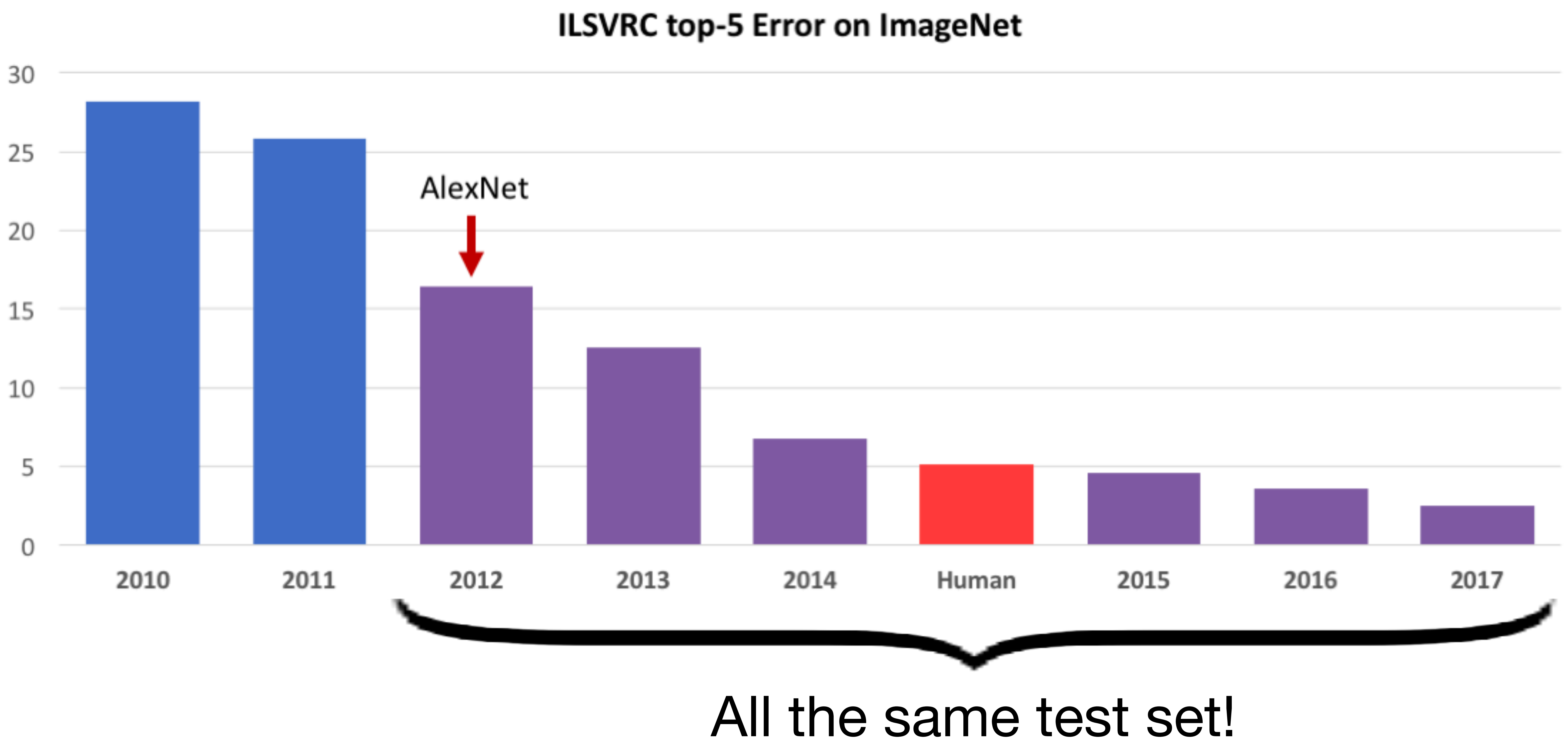


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Replication Crisis in the Sciences



Real Cause for Concern



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

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

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

stract

w test nd ImageNet datasets. Both be
th ense re ade, raising the danger of overf
re s. By iginal 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.

Generaliza

Glamor and deceit?

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.

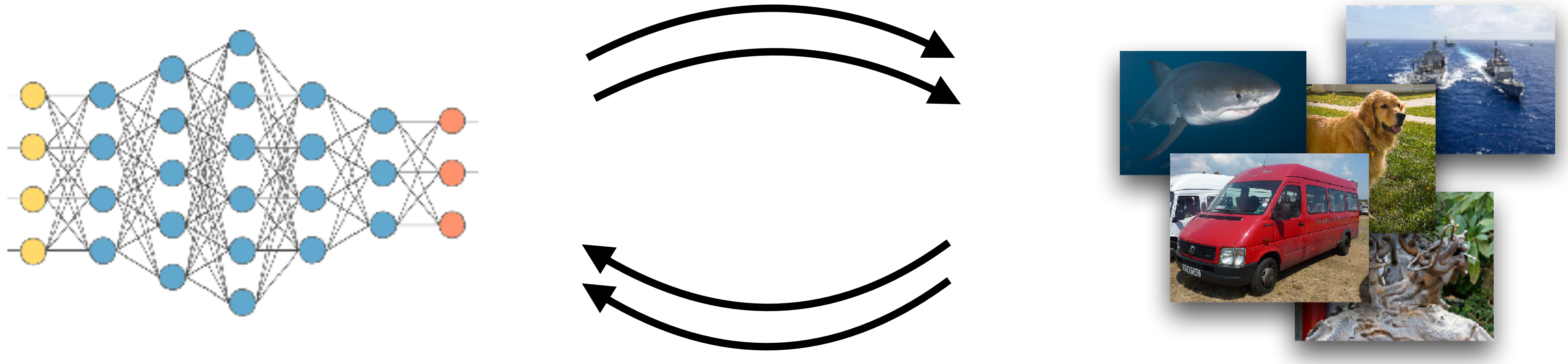
Overfitting

?

Three Forms of Overfitting

1. Test error ≥ training error

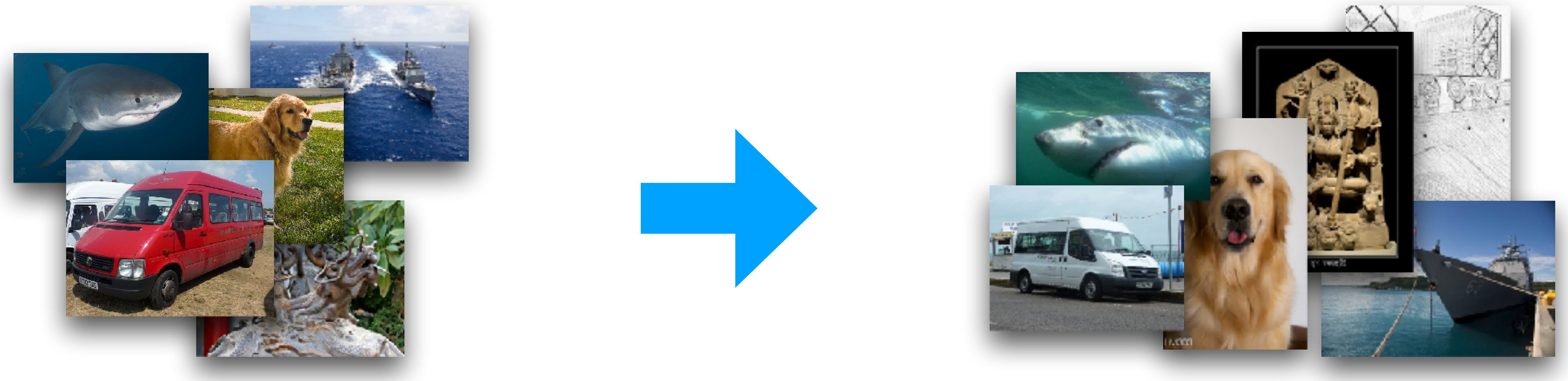
2. Overfitting through test set re-use



Model

Test Set

3. Distribution shift



Original Test Set

New Test Set

Two Possible Causes

New test accuracy

Overfitting through test set re-use

Distribution shift

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

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

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

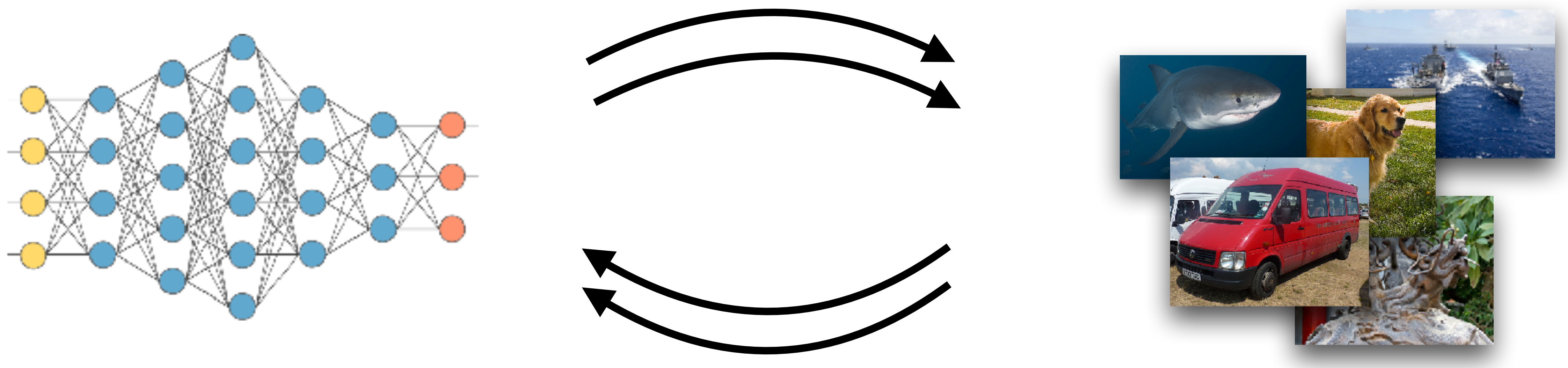
$$\text{acc}_D(f) = \mathbb{E}_{(x,y) \sim D} \mathbb{1}[f(x) = y]$$ (S is drawn from D)

Generalization error (≈ 1%)

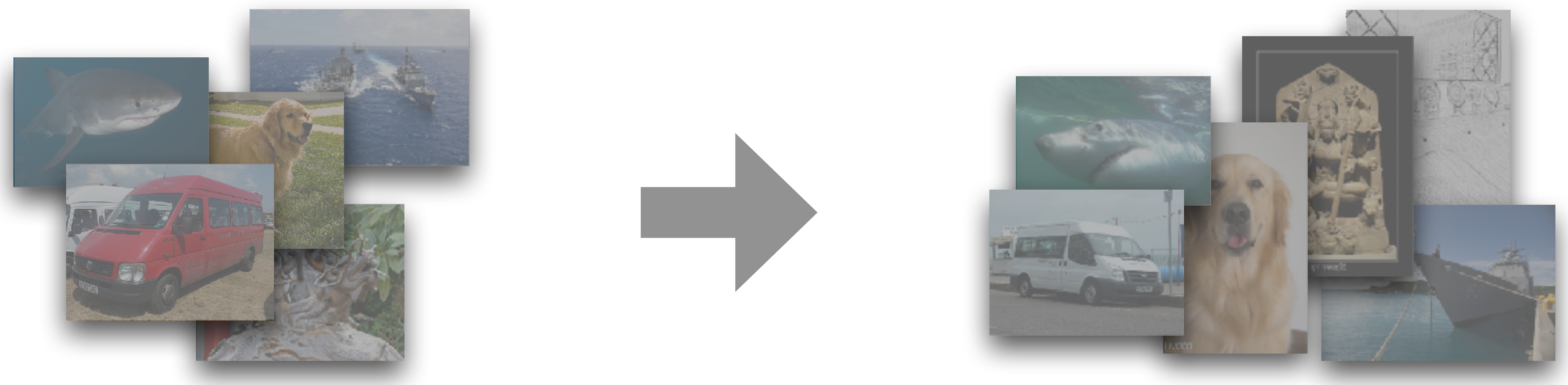
Three Forms of Overfitting

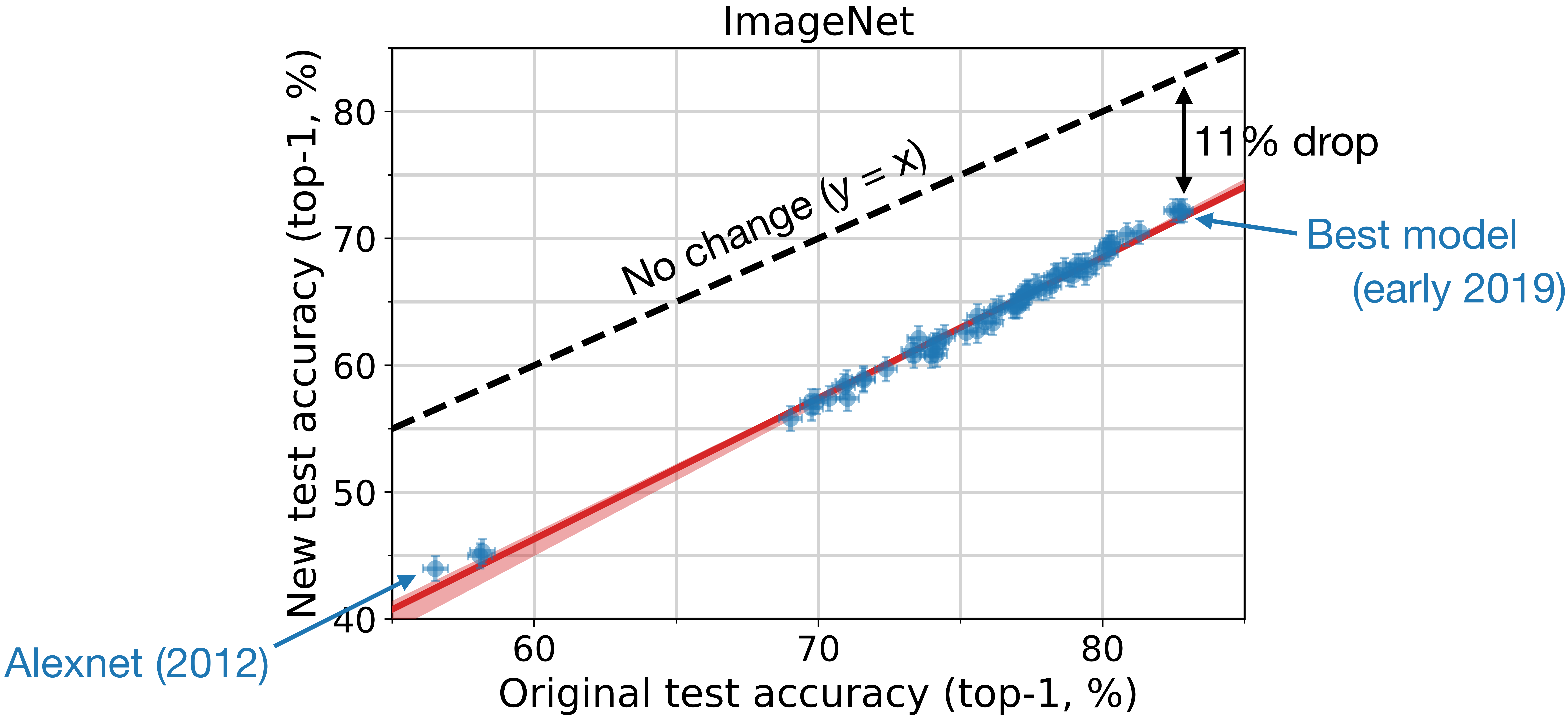
1. Test error ≥ training error

2. Overfitting through test set re-use



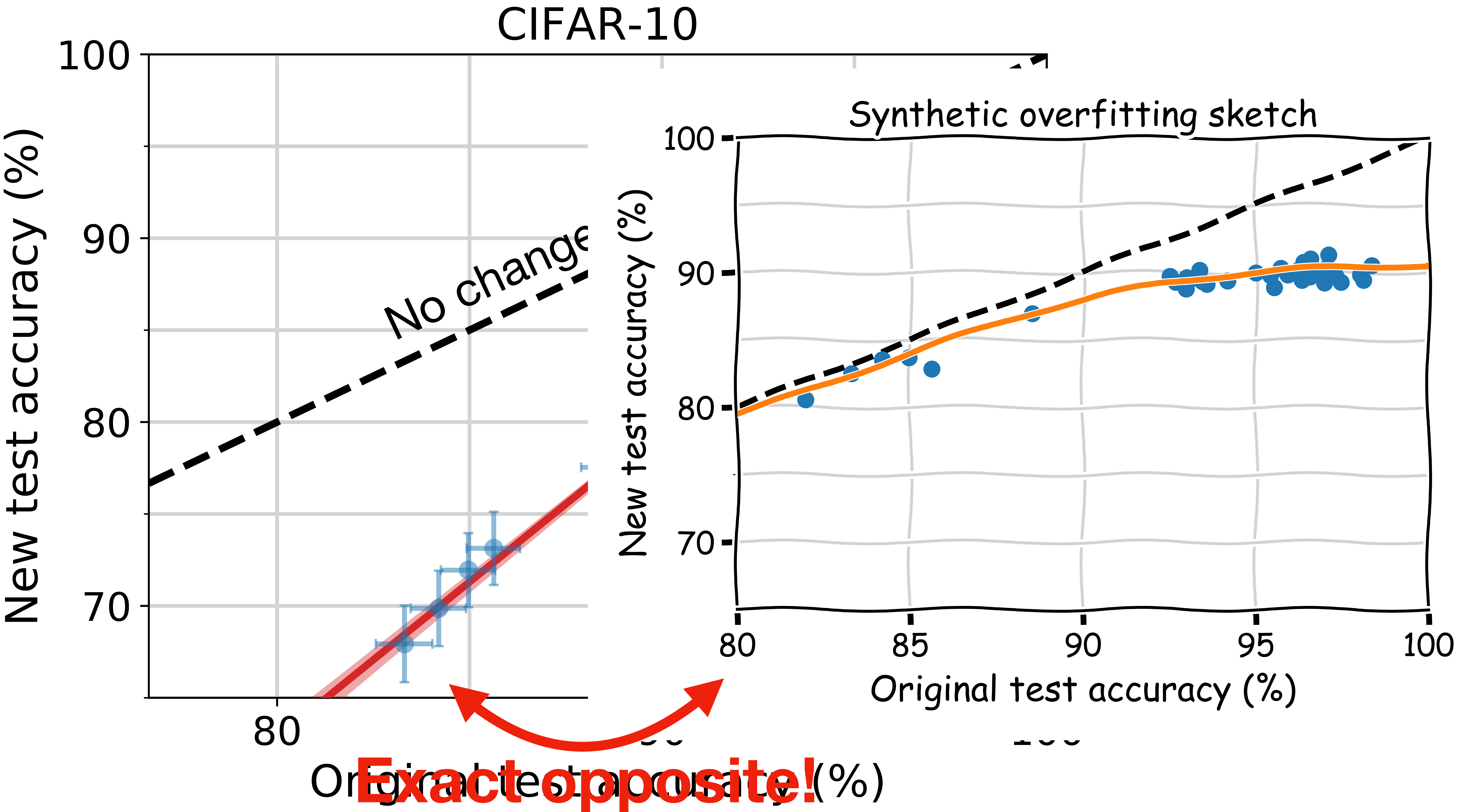
3. Distribution shift





➡ 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. [Recht, Roelofs, **Schmidt**, Shankar '19]



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.

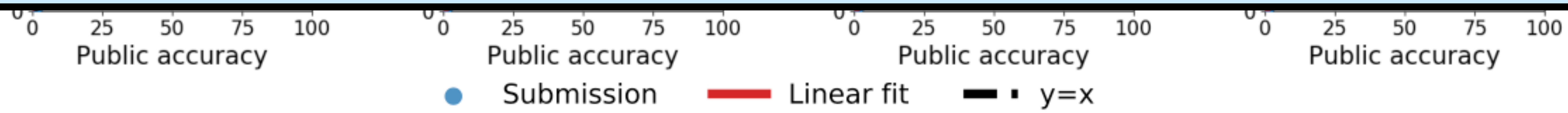
Relative ordering preserved. Progress is real!

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]

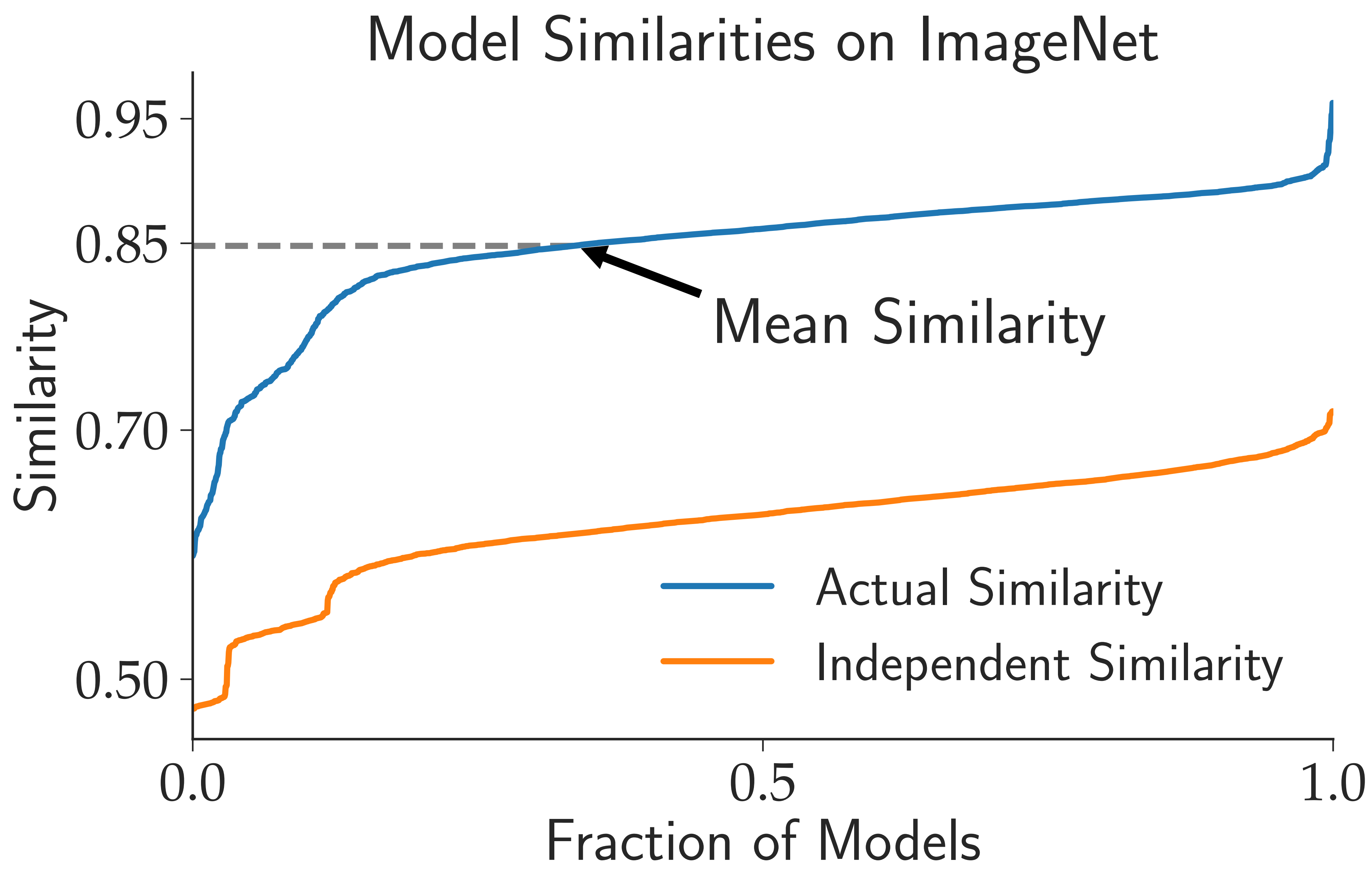
Our results unambiguously confirm the trends observed by Recht et al. [2018, 2019]: although the misclassification rates are slightly off, classifier ordering and model selection remain broadly reliable.

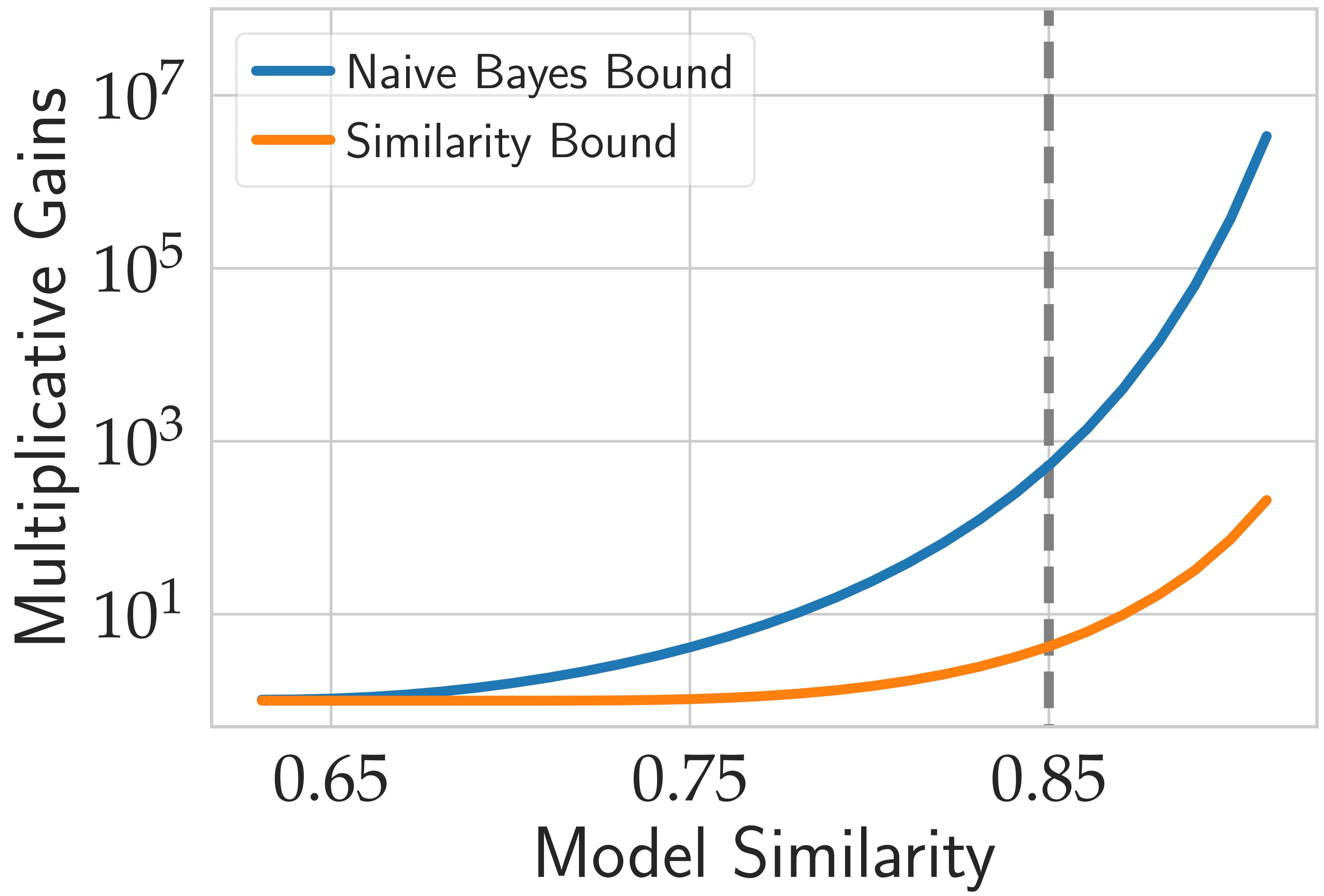


Why Does Test Set Re-use Not Lead to Overfitting?

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%)

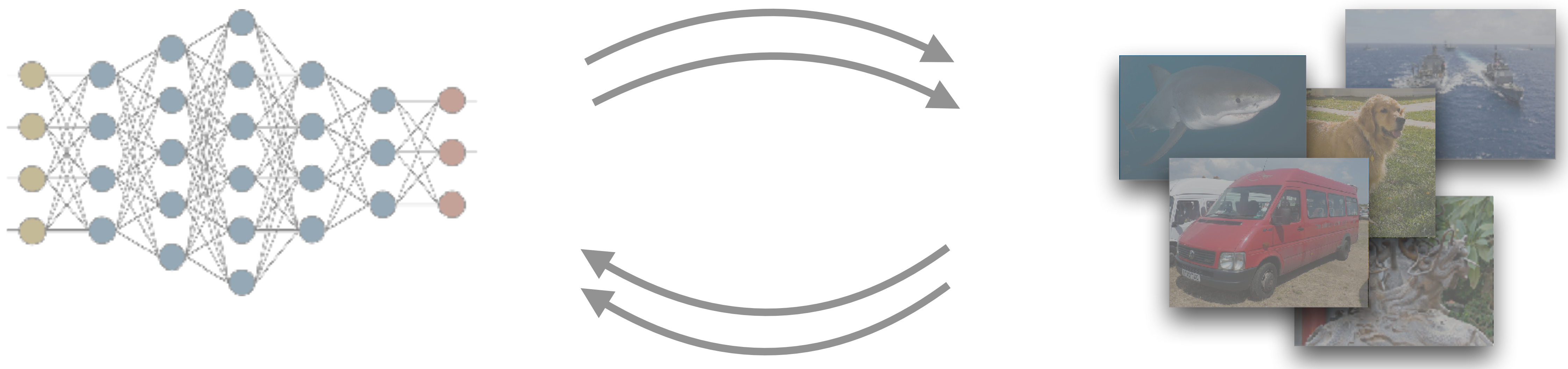
$$\widehat{\text{acc}}_S(f) = \frac{1}{|S|} \sum_{(x,y)\in S} \mathbb{1}[f(x) = y]$$

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

Three Forms of Overfitting

1. Test error ≥ training error

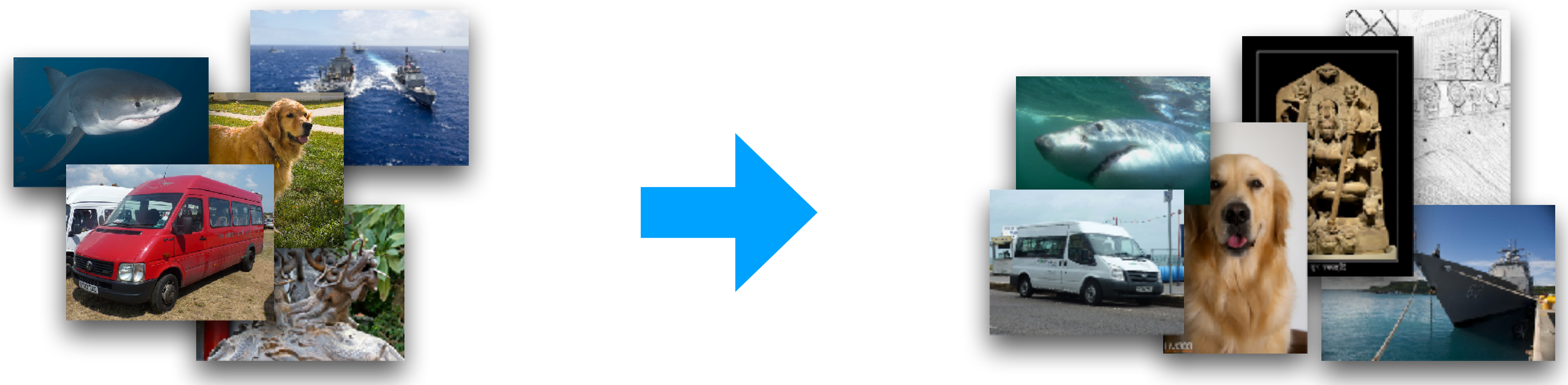
2. Overfitting through test set re-use



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**

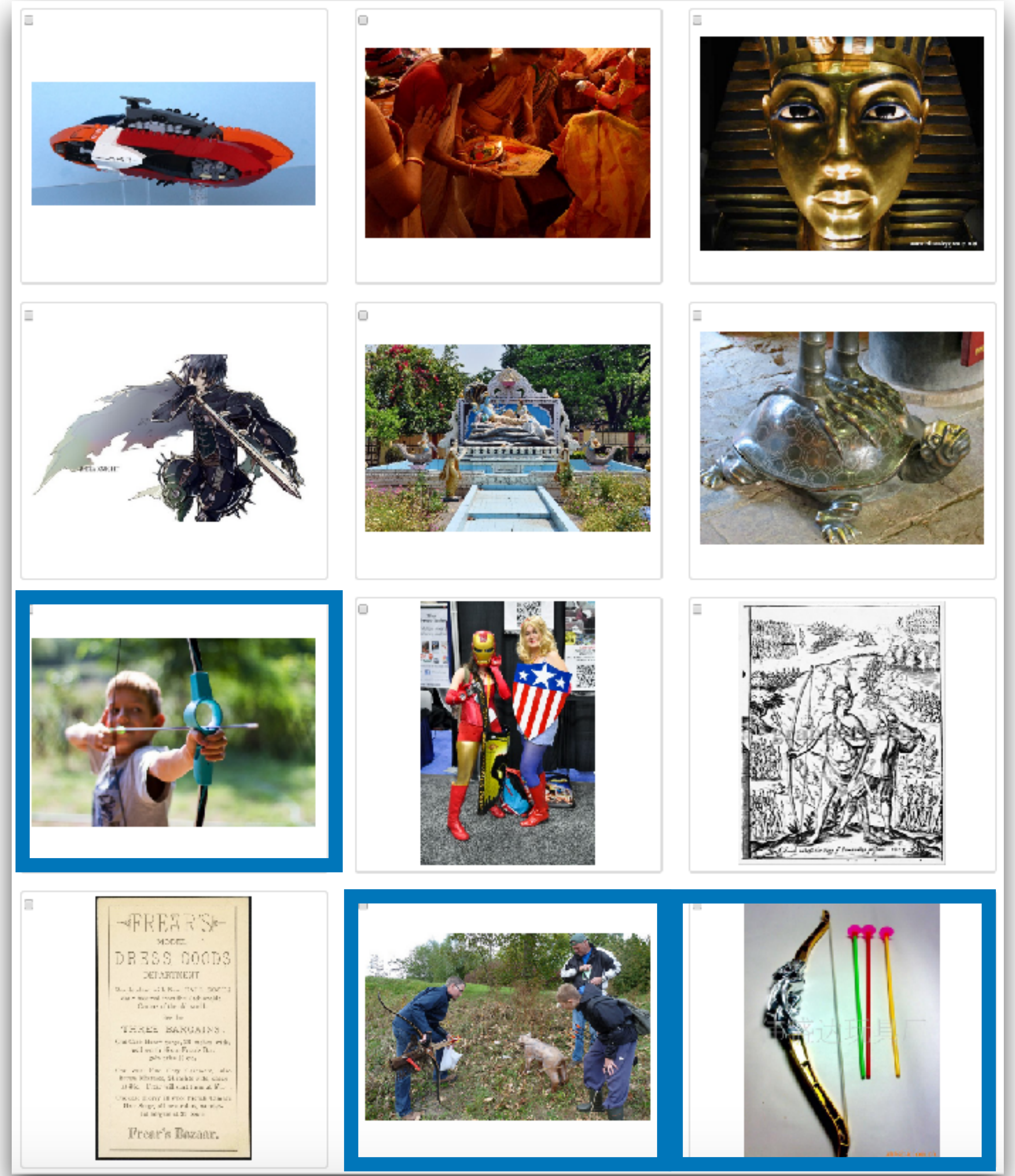
3. Show images to **MTurk** workers ← Likely source of distribution shift

4. Sample a class-balanced dataset

We replicated this process as closely as possible.

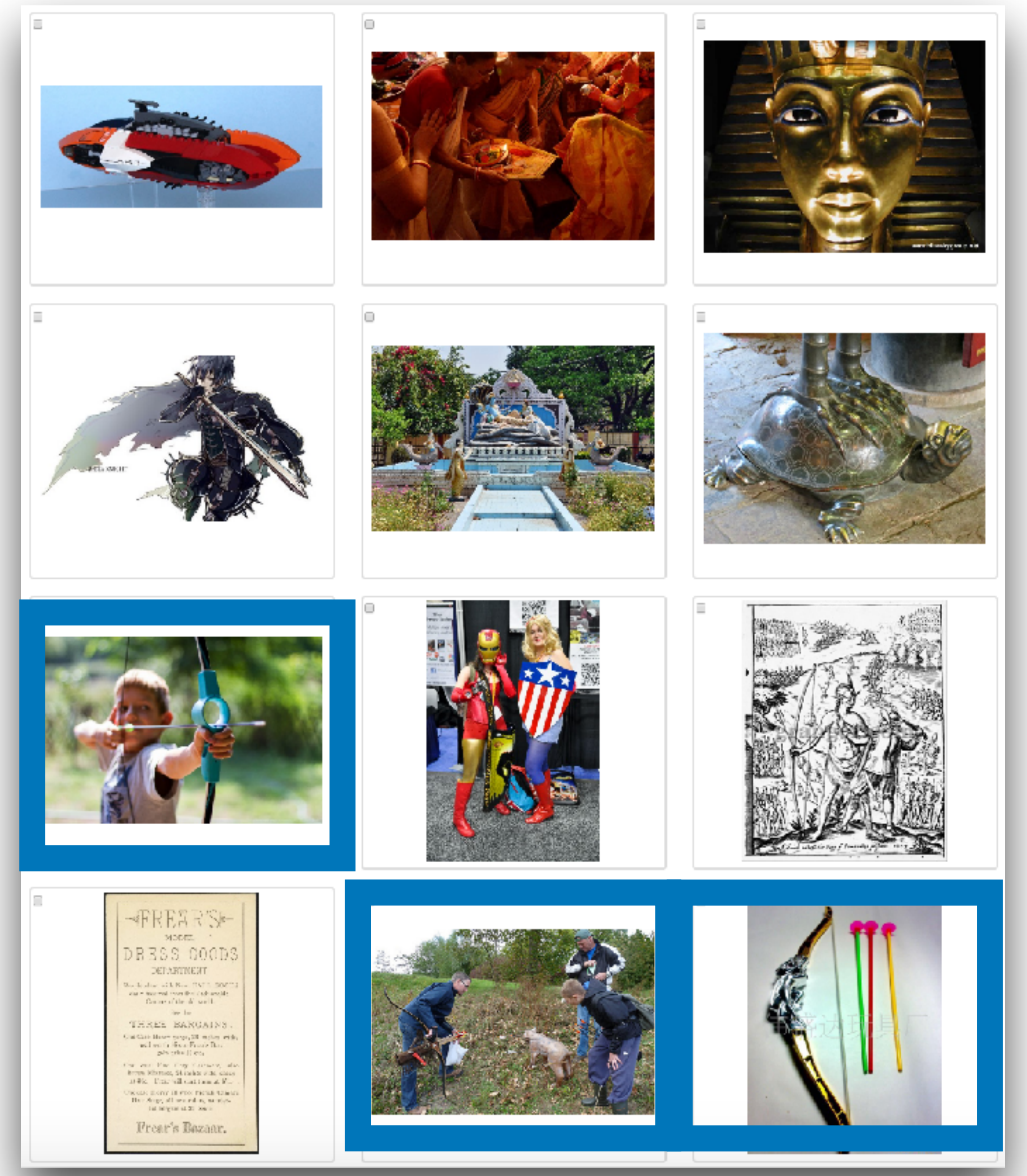
Data Cleaning With MTurk

Instructions: Select all images containing a bow.

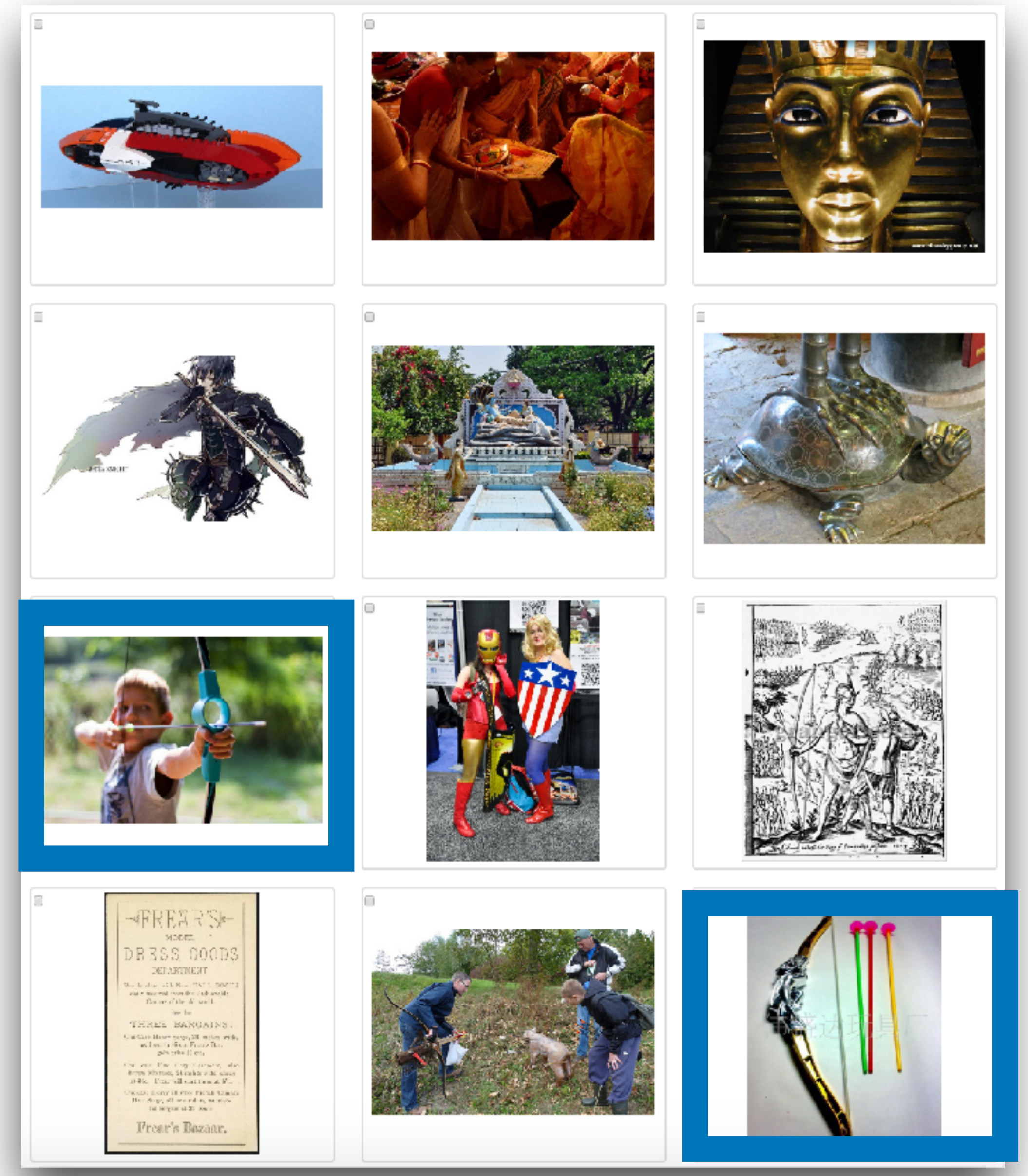


Data Cleaning With MTurk

 Worker 1

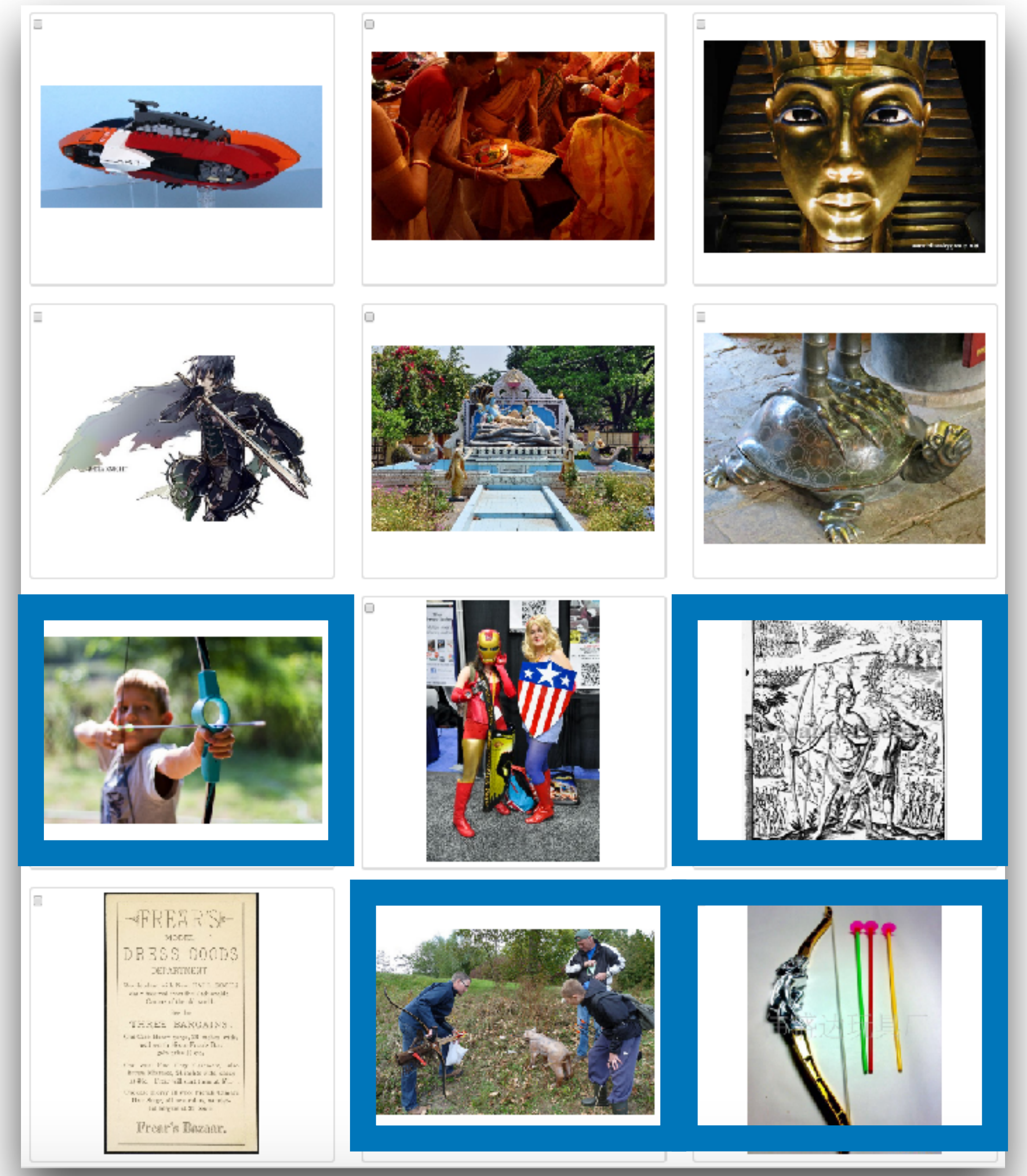


 Worker 2

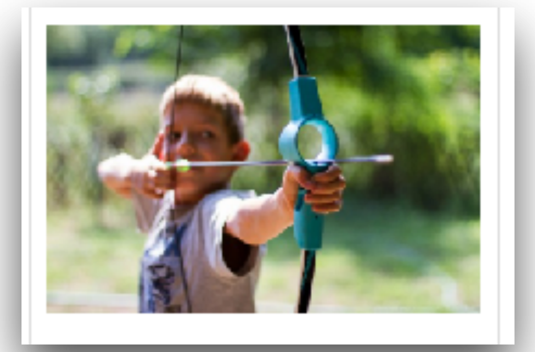


...

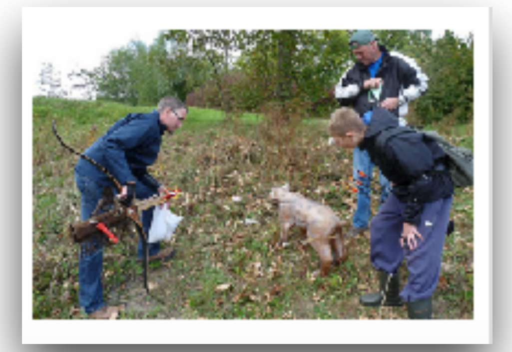
 Worker 10



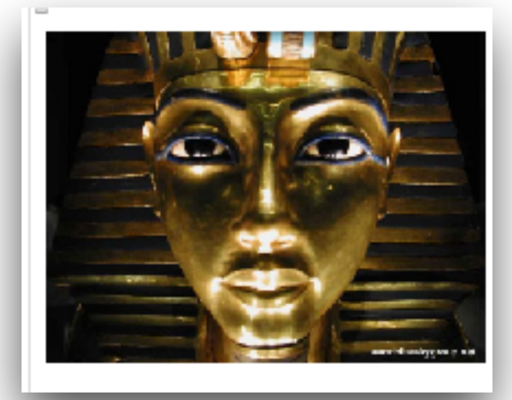
Main quantity: **selection frequency** $=$ $\dfrac{\textbf{Number of workers who selected image } i}{\textbf{Number of workers who saw image } i}$

 : 1.0

 : 1.0

 : 0.67

 : 0.33

 : 0.0

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.

Caveats with Benchmarks

A: Are new methods really better? What about the methods we already had?

B: Are we just overfitting to the benchmark test sets?

C: Do we have progress beyond the immediate benchmark?

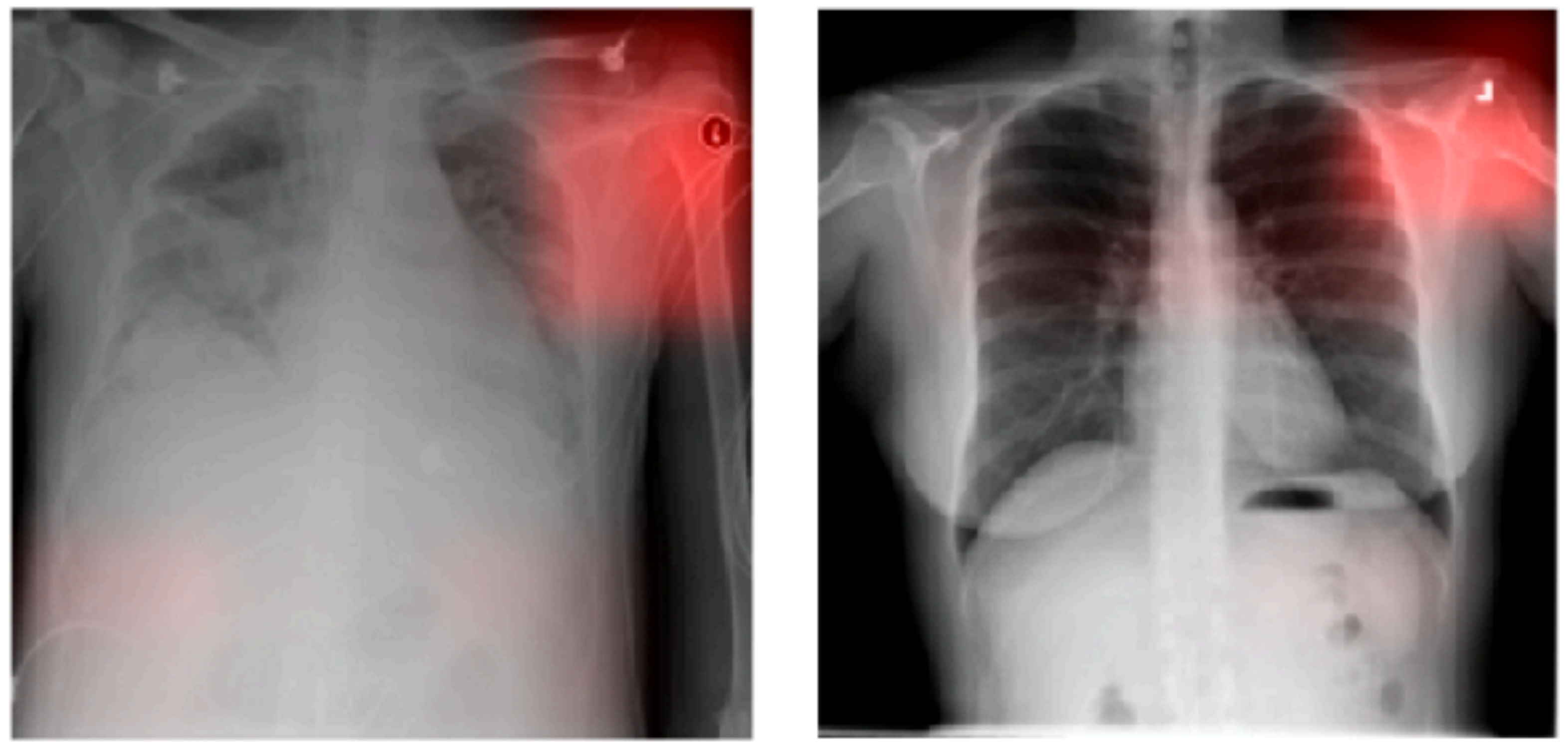
Why Focus on ImageNet?

The community has spent **a lot** of effort on ImageNet.

In the end, ImageNet is not a real problem but an experiment / **toy dataset**.

Does progress on ImageNet actually lead to **progress more broadly**?

Food-101

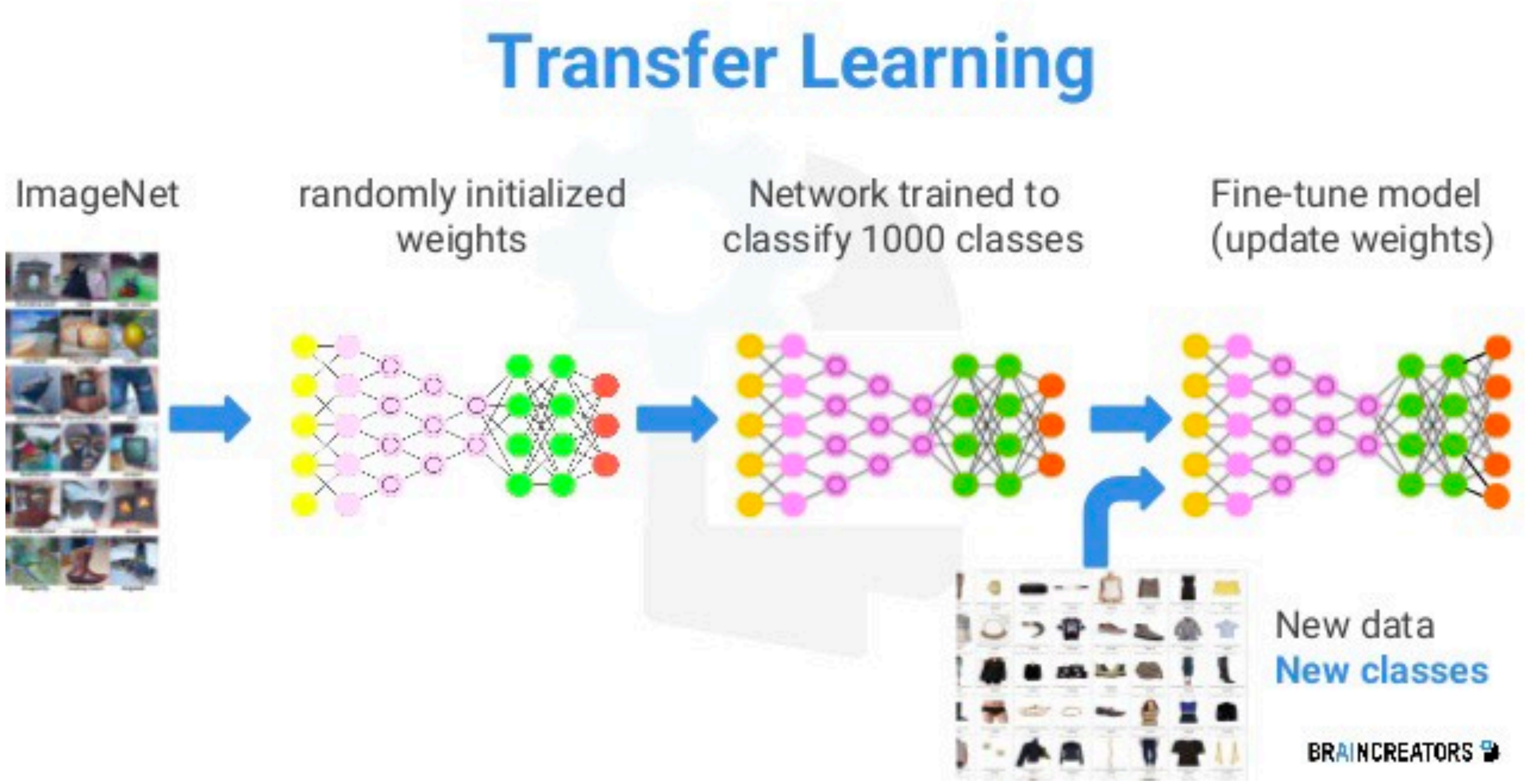


Medical imaging

Transfer Learning

Common paradigm in machine learning

Core idea: leverage a large dataset to improve performance on a small dataset



Do Better ImageNet Models Transfer Better?

Simon Kornblith*, Jonathon Shlens, and Quoc V. Le
Google Brain
{skornblith,shlens,qvl}@google.com

Abstract

Transfer learning is a cornerstone of computer vision, yet little work has been done to evaluate the relationship between architecture and transfer. An implicit hypothesis in modern computer vision research is that models that perform better on ImageNet necessarily perform better on other vision tasks. However, this hypothesis has never been systematically tested. Here, we compare the performance of 16 classification networks on 12 image classification datasets. We find that, when networks are used as fixed feature extractors or fine-tuned, there is a strong correlation between ImageNet accuracy and transfer accuracy ($r = 0.99$ and 0.96, respectively). In the former setting, we find that this relationship is very sensitive to the way in which networks are trained on ImageNet; many common forms of regularization slightly improve ImageNet accuracy but yield penultimate layer features that are much worse for transfer learning. Additionally, we find that, on two small fine-grained image classification datasets, pretraining on ImageNet provides minimal benefits, indicating the learned features from ImageNet do not transfer well to fine-grained tasks. Together, our results show that ImageNet architectures generalize well across datasets, but ImageNet features are less general than previously suggested.

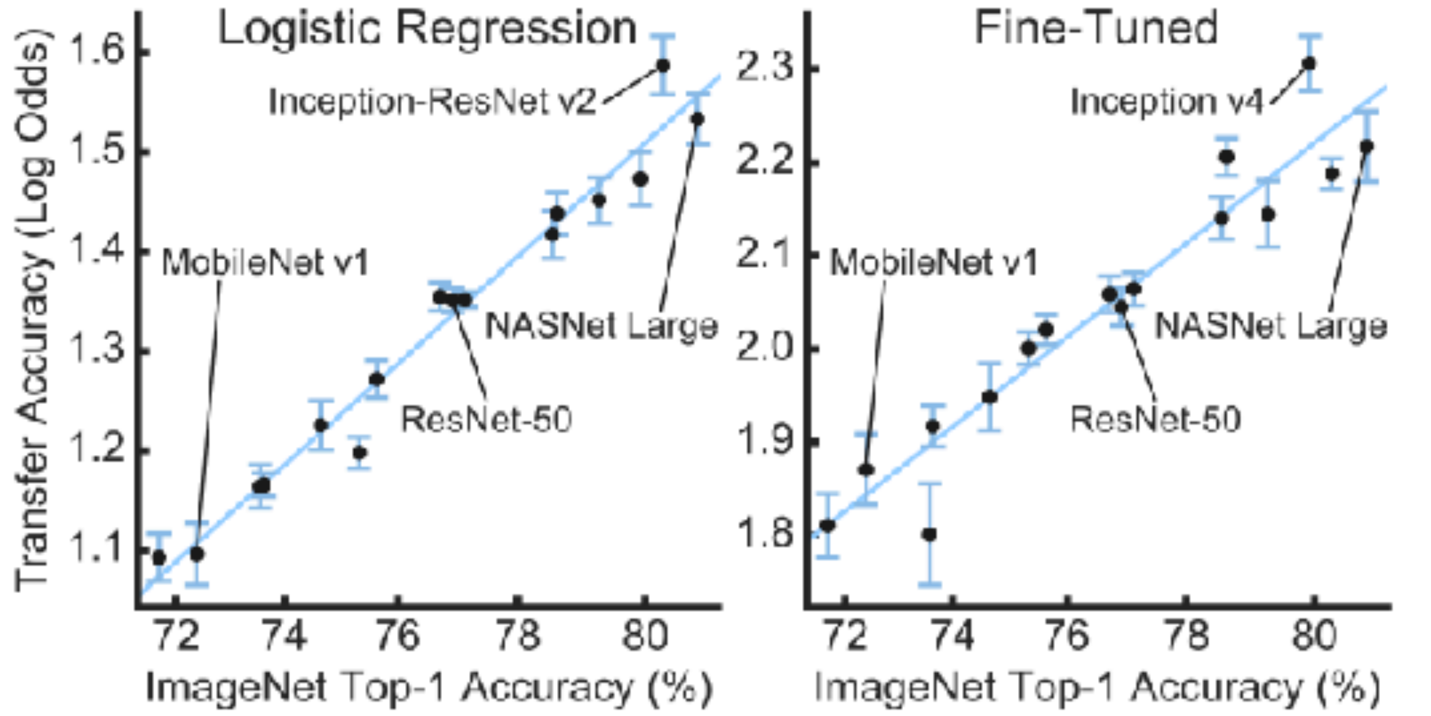


Figure 1. Transfer learning performance is highly correlated with ImageNet top-1 accuracy for fixed ImageNet features (left) and fine-tuning from ImageNet initialization (right). The 16 points in each plot represent transfer accuracy for 16 distinct CNN architectures, averaged across 12 datasets after logit transformation (see Section 3). Error bars measure variation in transfer accuracy across datasets. These plots are replicated in Figure 2 (right).

ter network architectures learn better features that can be transferred across vision-based tasks. Although previous studies have provided some evidence for these hypotheses (e.g. [6, 71, 37, 35, 31]), they have never been systematically explored across network architectures.

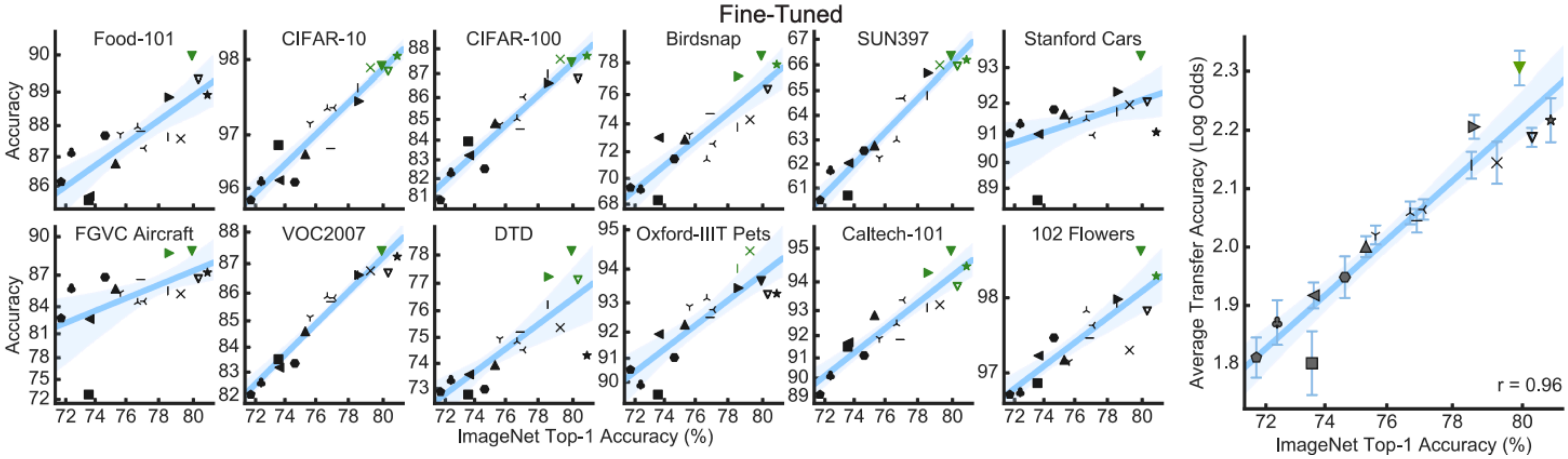
In the present work, we seek to test these hypotheses by investigating the transferability of both ImageNet features and

Datasets evaluated

Dataset	Classes	Size (train/test)	Accuracy metric
Food-101 [5]	101	75,750/25,250	top-1
CIFAR-10 [43]	10	50,000/10,000	top-1
CIFAR-100 [43]	100	50,000/10,000	top-1
Birdsnap [4]	500	47,386/2,443	top-1
SUN397 [84]	397	19,850/19,850	top-1
Stanford Cars [41]	196	8,144/8,041	top-1
FGVC Aircraft [55]	100	6,667/3,333	mean per-class
PASCAL VOC 2007 Cls. [22]	20	5,011/4,952	11-point mAP
Describable Textures (DTD) [10]	47	3,760/1,880	top-1
Oxford-IIIT Pets [61]	37	3,680/3,369	mean per-class
Caltech-101 [24]	102	3,060/6,084	mean per-class
Oxford 102 Flowers [59]	102	2,040/6,149	mean per-class

Recall ImageNet has 1.2 million training images (and 1,000 classes).

Better ImageNet Models Transfer Better



Progress on ImageNet helps on a wide range of image classification datasets.

Also transfer of techniques to other tasks (object detection, etc.)

But: This is not guaranteed. Some datasets are considered “bad” or too specialized. (Models don’t work “in the wild”)

Caveats with Benchmarks

A: Are new methods really better? What about the methods we already had?

Depends on the benchmark. Competitive, standardized benchmarks usually have good baselines.

B: Are we just overfitting to the benchmark test sets?

Not in classification tasks with at least 1,000 test examples.

C: Do we have progress beyond the immediate benchmark?

Depends on the benchmark. Several popular benchmarks promote broad progress.

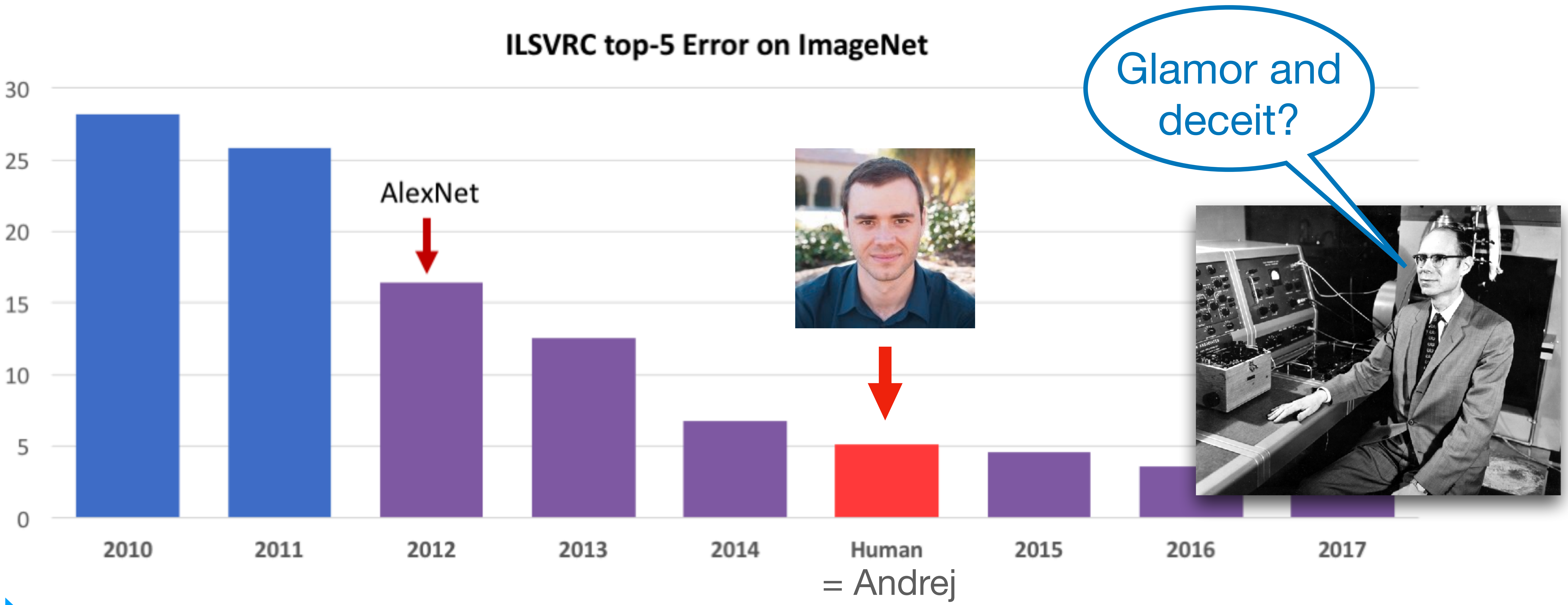
ImageNet served as a reliable indicator of progress for 10 years!

1. Empirical progress in machine learning: benchmarks

2. What can we learn from ML benchmarks?

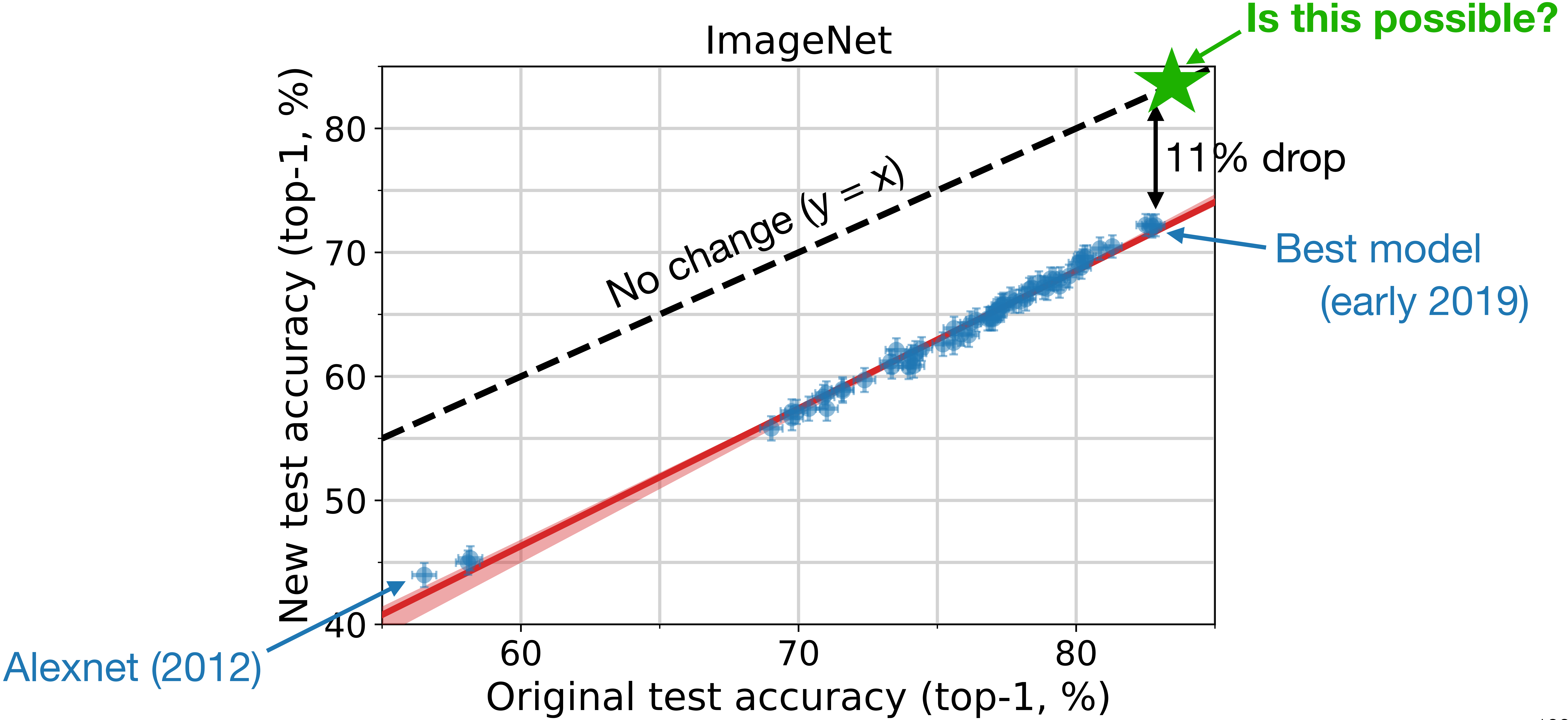
3. Limitations of current ML methods

So Far, Things are Looking Good



- What is good performance (Bayes error)?
- Can we get a more fine-grained understanding of model performance?

Also: What about ImageNetV2?



Evaluating Machine Accuracy on ImageNet

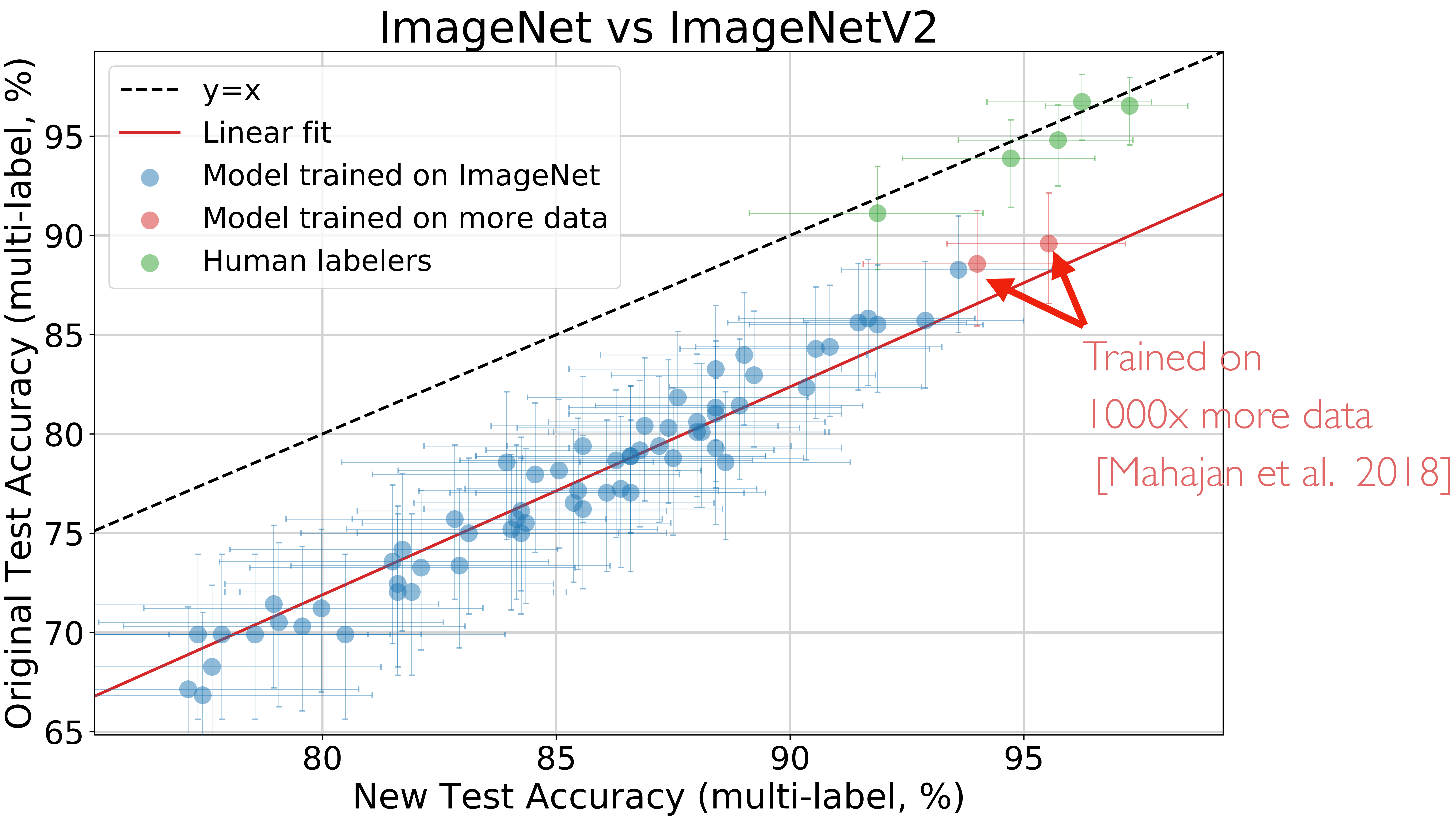
Vaishaal Shankar [* 1] **Rebecca Roelofs** [* 2] **Horia Mania** [1] **Alex Fang** [1] **Benjamin Recht** [1] **Ludwig Schmidt** [1]

Abstract

We evaluate a wide range of ImageNet models with five trained human labelers. In our year-long experiment, trained humans first annotated 40,000 images from the ImageNet and ImageNetV2 test sets with multi-class labels to enable a semantically coherent evaluation. Then we measured the classification accuracy of the five trained humans on the full task with 1,000 classes. Only the latest models from 2020 are on par with our best human labeler, and human accuracy on the 590 object classes is still 4% and 11% higher than the best model on ImageNet and ImageNetV2, respectively. Moreover, humans achieve the same accuracy on ImageNet and ImageNetV2, while all models see a consistent accuracy drop. Overall, our results show that there is still substantial room for improvement on ImageNet and direct accuracy comparisons between humans and machines may overstate machine performance.

In this paper, we contextualize progress on ImageNet by comparing a wide range of ImageNet models to five trained human labelers. Our year-long experiment consists of two parts: first, three labelers thoroughly re-annotated 40,000 test images in order to create a testbed with minimal annotation artifacts. The images are drawn from both the original ImageNet validation set and the ImageNetV2 replication study of Recht et al. (2019). Second, we measured the classification accuracy of the five trained labelers on the full 1,000-class ImageNet task. We again utilized images from both the original and the ImageNetV2 test sets. This experiment led to the following contributions:

Multi-label annotations. Our expert labels quantify multiple issues with the widely used `top-1` and `top-5` metrics on ImageNet. For instance, about 20% of images have more then one valid label, which makes `top-1` numbers overly pessimistic. To ensure a consistent annotation of all 40,000 images, we created a 400-page labeling guide describing the fine-grained class distinctions. In addition, we



- Same accuracy on ImageNet and ImageNetV2 is possible (achieved by humans)
- Humans still better than best models in early 2020 (much better than 2015)

How Should We Evaluate ImageNet?

Recall: current evaluation metrics are **top-1 and top-5 accuracy**.

These are informative in the medium accuracy regime from 2010,
but have drawbacks in the high accuracy regime in 2020.

Problem 1: images with several objects

ImageNet classes:

- Monitor
- Screen
- Table lamp
- Lamp shade
- Desk
- Computer keyboard
- Mouse
- Speaker
- Desktop computer
- maybe more …

Problem 2: subset relationships in the ImageNet class hierarchy

Tusker vs. Indian Elephant

Mushroom vs. Gyromitra

Shortcomings of Current Metrics

Top-1 Accuracy

Desk, Laptop, Monitor, etc…

Mushroom vs.Gyromitra

Paper Towel, Dock, Pier, …

Tusker vs African Elephant

Crowded Images

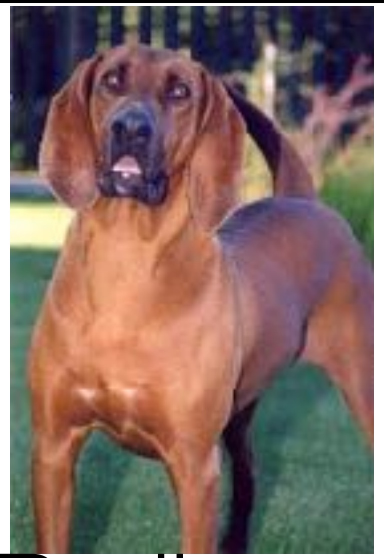
Subset Relationships

Makes the task too hard

(Multiple correct answers)

Top-5 Accuracy

Vizsla



Redbone

Chesapeake Bay

Rhodesian

Makes the task too easy

(Classes can be distinguished)

Our Approach: Multi-Label Accuracy

Each Classifier **predicts one label per image**

An image can have **multiple labels**

Prediction counts as **correct if in the label set**

Multi-label accuracy has been studied before.

We are the first to systematically collect annotations with expert labelers.

ImageNet label: Picket Fence

Our labels: Groom, Bowtie, Gown, Picket Fence

Collecting Multi-Label Annotations

n03461385 grocery store, grocery, food market, market
a marketplace where groceries are sold; "the grocery store included a meat market"
Correct Wrong Unclear Don't know Unreviewed

n07717556 butternut squash
buff-colored squash with a long usually straight neck and sweet orange flesh
Correct Wrong Unclear Don't know Unreviewed

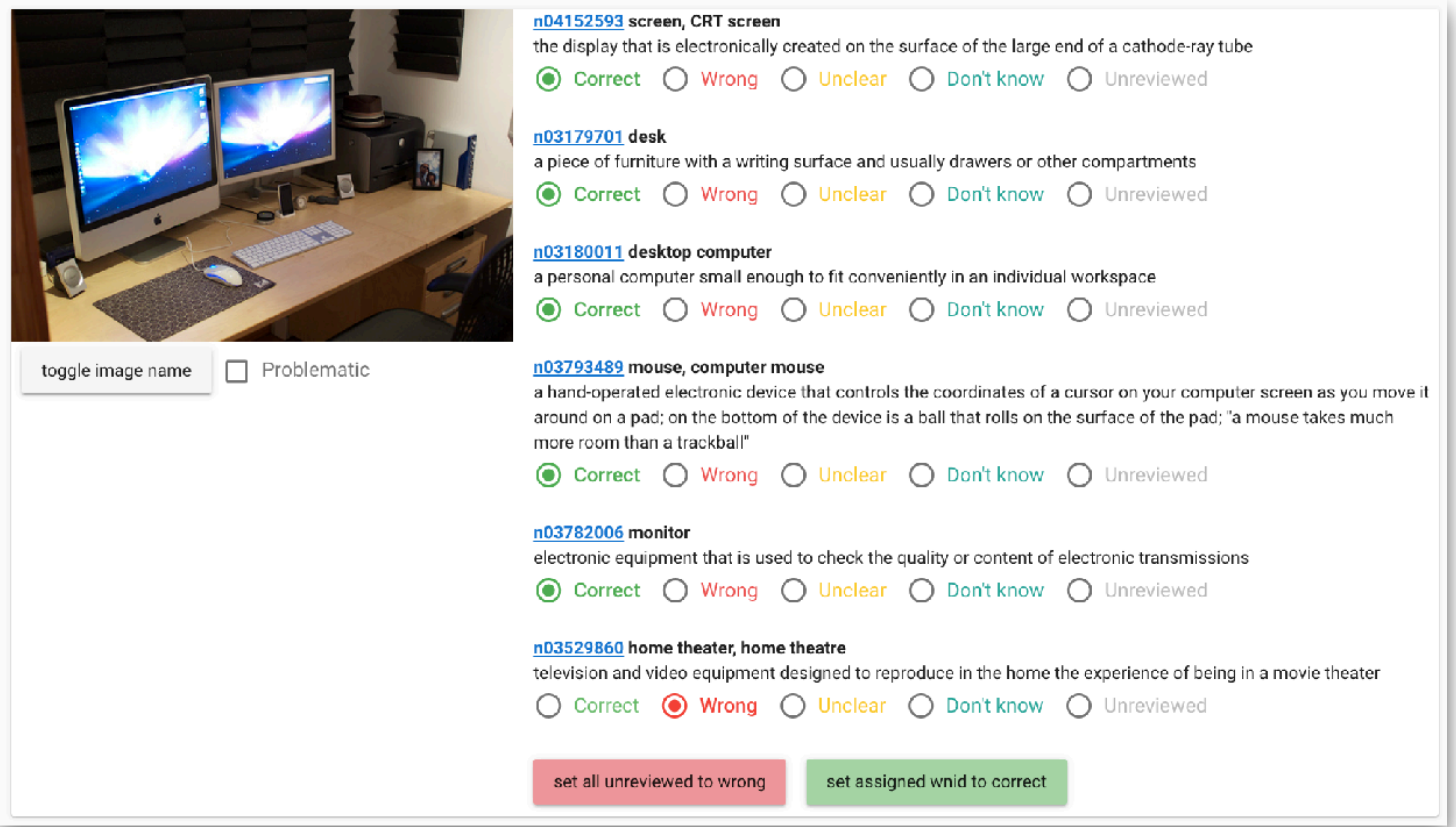
n07716906 spaghetti squash
medium-sized oval squash with flesh in the form of strings that resemble spaghetti
Correct Wrong Unclear Don't know Unreviewed

n07717410 acorn squash
small dark green or yellow ribbed squash with yellow to orange flesh
Correct Wrong Unclear Don't know Unreviewed

set all unreviewed to wrong set assigned wnid to correct

toggle image name Problematic

Collecting Multi-Label Annotations



Majority vote for contentious labels.

Collecting Multi-Label Annotations

Some classes (especially dog breeds, some monkeys, etc.) took hours of research.

French Bulldog

Head large and square. **Eyes** dark in color, wide apart, set low down in the skull, as far from the ears as possible, round in form, of moderate size, neither sunken nor bulging… (AKC.org)

Boston Terrier

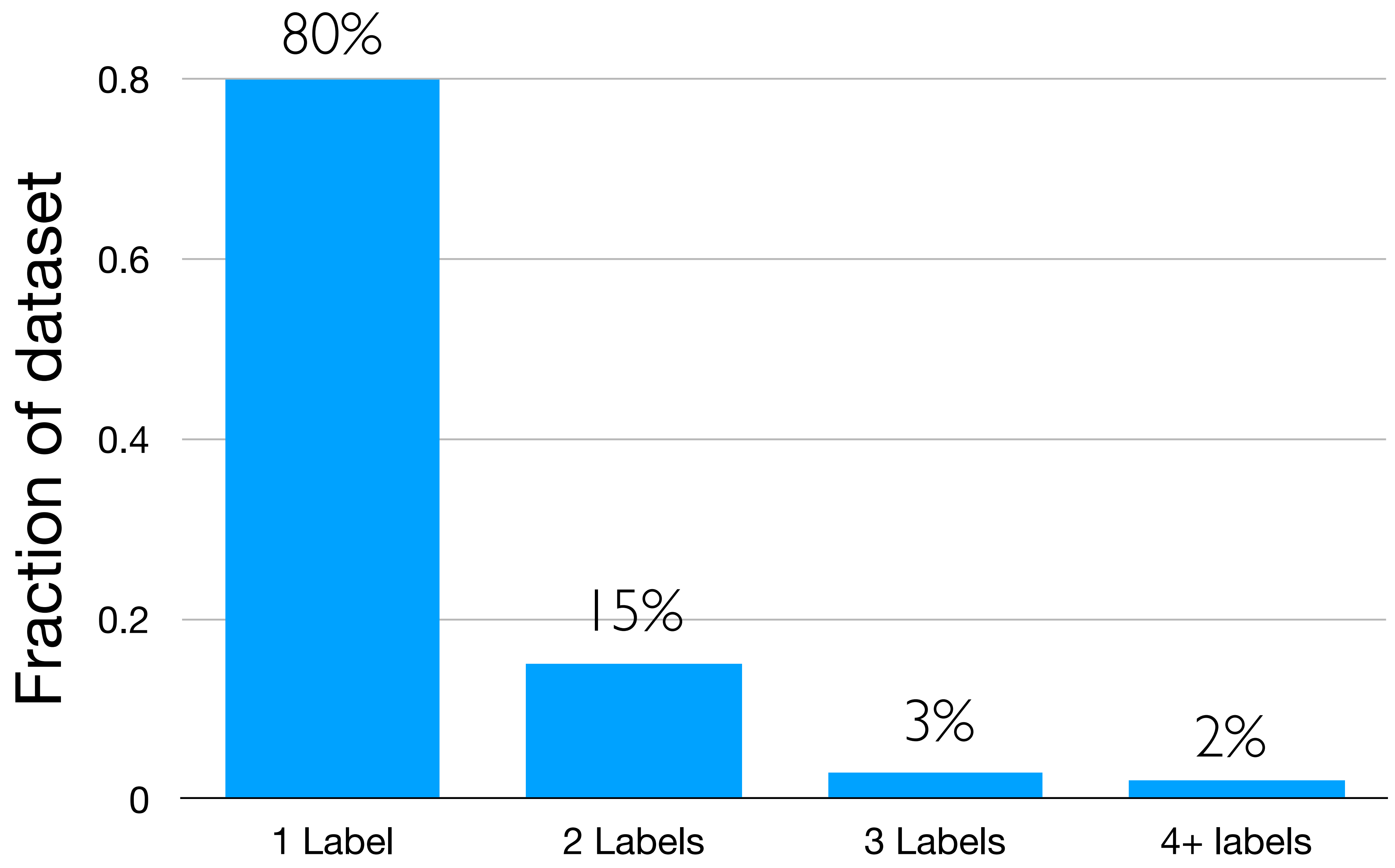
The **skull** is square, flat on top, free from wrinkles, cheeks flat, brow abrupt and the stop well defined. … The **eyes** are wide apart, large and round and dark in color… (AKC.org)

Our labeling guide is about 400 pages long (though parts of it are auto-generated).

Multi-Label Statistics

40,683 Images Annotated from ImageNet and ImageNetV2

182,597 unique model predictions reviewed.



Measuring the Accuracy of Five Humans

Phase 1: collection multi-label annotations (Becca, Ludwig, Vaishaal — 6 months)

Potential problem: We labeled the test set!

Solution: Part A: 6 month break before phase 2
(Subjectively you forget images fairly quickly, but not 100% sure)
Part B: Two expert labelers joined the project (Alex and Horia)

Phase 2: Train human labelers (2 months)

Phase 3: Evaluate human labelers (1 month)

Phase 4: Final label review (10 days)

Best model accuracy: 96%

Best Model Accuracy: 96%

Best Model Accuracy: 95%

Organisms

Objects

Best model accuracy: 90% (-6%)
Best human accuracy: 97% (+0.5%)

Accuracy difference between ImageNet and ImageNetV2

Humans still 11% better on objects!

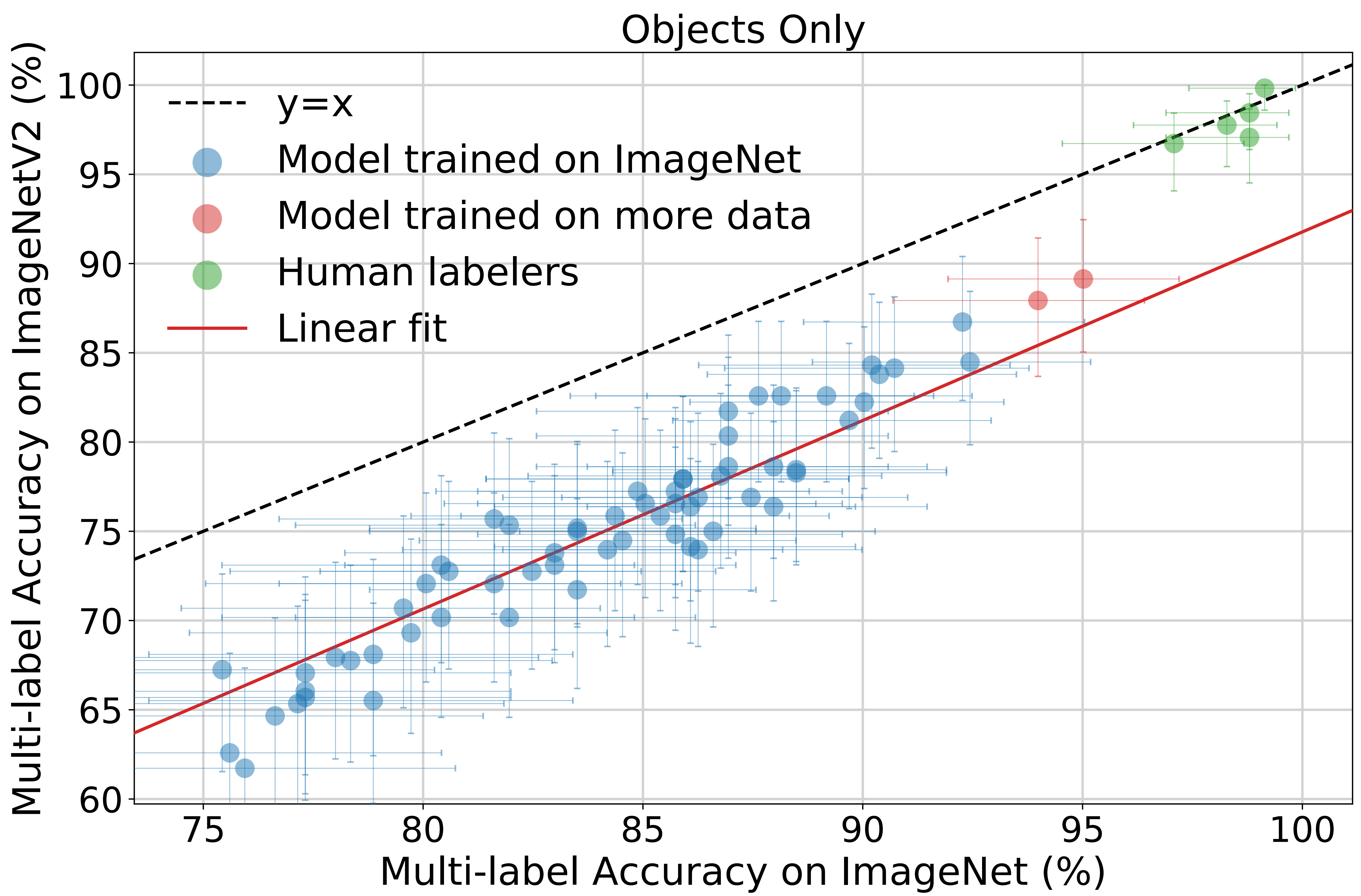
Best model accuracy: 90% (-6.3%)
Best human accuracy: 93%(+0.2%)

Organisms

Best model accuracy: 89% (-5.9%)
Best human accuracy: 99.8% (+0.7%)

Objects

ImageNetV2 Scatter Plot for Objects Only



CAVEAT:

Should we care about accuracy on 130 dog breeds?

Probably not.

➡ Likely closer to “real” relative performance on ImageNet

We worked with a judge from the American Kennel Club who has 20 years of experience: there is still room for improvement in our dog accuracies.

More Evidence

Generalisation in humans and deep neural networks

Robert Geirhos[1-3*§] **Carlos R. Medina Temme**[1*] **Jonas Rauber**[2,3*]

Heiko H. Schütt[1,4,5] **Matthias Bethge**[2,6,7*] **Felix A. Wichmann**[1,2,6,8*]

[1]Neural Information Processing Group, University of Tübingen
[2]Centre for Integrative Neuroscience, University of Tübingen
[3]International Max Planck Research School for Intelligent Systems
[4]Graduate School of Neural and Behavioural Sciences, University of Tübingen
[5]Department of Psychology, University of Potsdam
[6]Bernstein Center for Computational Neuroscience Tübingen
[7]Max Planck Institute for Biological Cybernetics
[8]Max Planck Institute for Intelligent Systems
[*]Joint first / joint senior authors
[§]To whom correspondence should be addressed: `robert.geirhos@bethgelab.org`

Synthetic Distribution Shifts

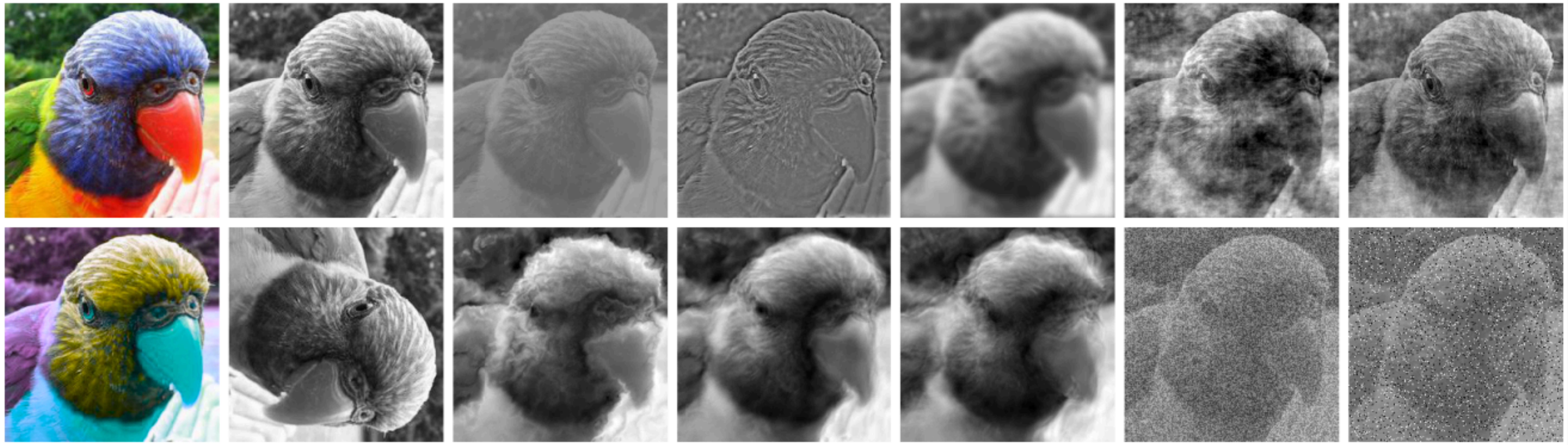
Key idea: evaluate networks and humans under a range of synthetic distribution shifts

Advantage: easy to generate

Disadvantage: not real data

➡ Still a good starting point!

Unperturbed image



Various Perturbations

Figure 2: Example stimulus image of class `bird` across all distortion types. From left to right, image manipulations are: colour (undistorted), greyscale, low contrast, high-pass, low-pass (blurring), phase noise, power equalisation. Bottom row: opponent colour, rotation, Eidolon I, II and III, additive uniform noise, salt-and-pepper noise. Example stimulus images across all used distortion levels are available in the supplementary material.

Results

Caveat: humans saw the image for only 200 ms (+ 1.5s decision time)

Caveat: 16 class version of ImageNet

→ Networks fail to generalize across distribution shifts, even if trained on all but one.

Evaluation condition / Model

Evaluation condition	human observers	A1	A2	A3	A4	A5	A6	A7	A8	A9	B1	B2	B3	B4	B5	B6	B7	B8	B9	C1	C2
colour	88.5	96.7	90.6	50.0	83.1	86.1	84.2	90.8	10.4	8.1	97.9	95.4	72.3	93.0	91.1	92.4	94.9	10.2	11.2	95.5	95.9
greyscale	86.6	87.8	95.6	94.1	86.2	93.2	87.8	90.5	10.3	9.8	94.0	96.8	96.2	93.3	95.7	94.3	90.9	11.4	12.8	94.8	95.1
contrast (5%)	47.6	13.1	14.2	89.4	19.6	39.8	17.1	10.2	28.6	29.0	46.3	51.7	95.1	50.5	79.1	59.4	45.2	34.6	37.9	90.9	88.2
low−pass (std=7)	48.5	18.9	16.1	16.4	78.4	11.9	16.0	9.8	6.9	6.6	16.0	18.6	14.4	87.2	20.5	13.8	13.5	7.1	9.3	74.7	74.9
high−pass (std=0.7)	49.8	21.1	24.7	29.9	11.7	92.6	27.7	8.3	10.4	20.6	25.1	22.8	29.2	25.0	94.3	27.5	28.3	18.9	19.8	91.4	90.7
phase noise (90°)	57.4	23.3	28.3	31.2	27.0	46.6	81.4	24.4	7.4	8.9	30.8	31.4	30.6	31.4	43.4	87.4	24.1	7.8	7.6	82.9	82.6
rotation (90°)	78.5	36.5	43.3	39.9	31.8	40.4	37.7	89.0	8.5	8.0	38.5	41.9	40.3	35.2	40.1	40.5	89.0	8.3	8.8	80.1	80.5
salt−and−pepper noise (0.2)	NA	6.1	6.4	5.8	7.9	6.2	6.2	6.4	79.4	6.2	6.2	6.1	6.3	5.4	5.8	5.7	6.2	89.6	6.2	78.6	13.6
uniform noise (0.35)	45.6	6.2	7.3	6.9	9.0	7.3	6.2	6.0	10.2	80.3	84.6	83.3	85.0	84.6	83.7	82.5	83.8	85.4	89.8	11.0	71.5

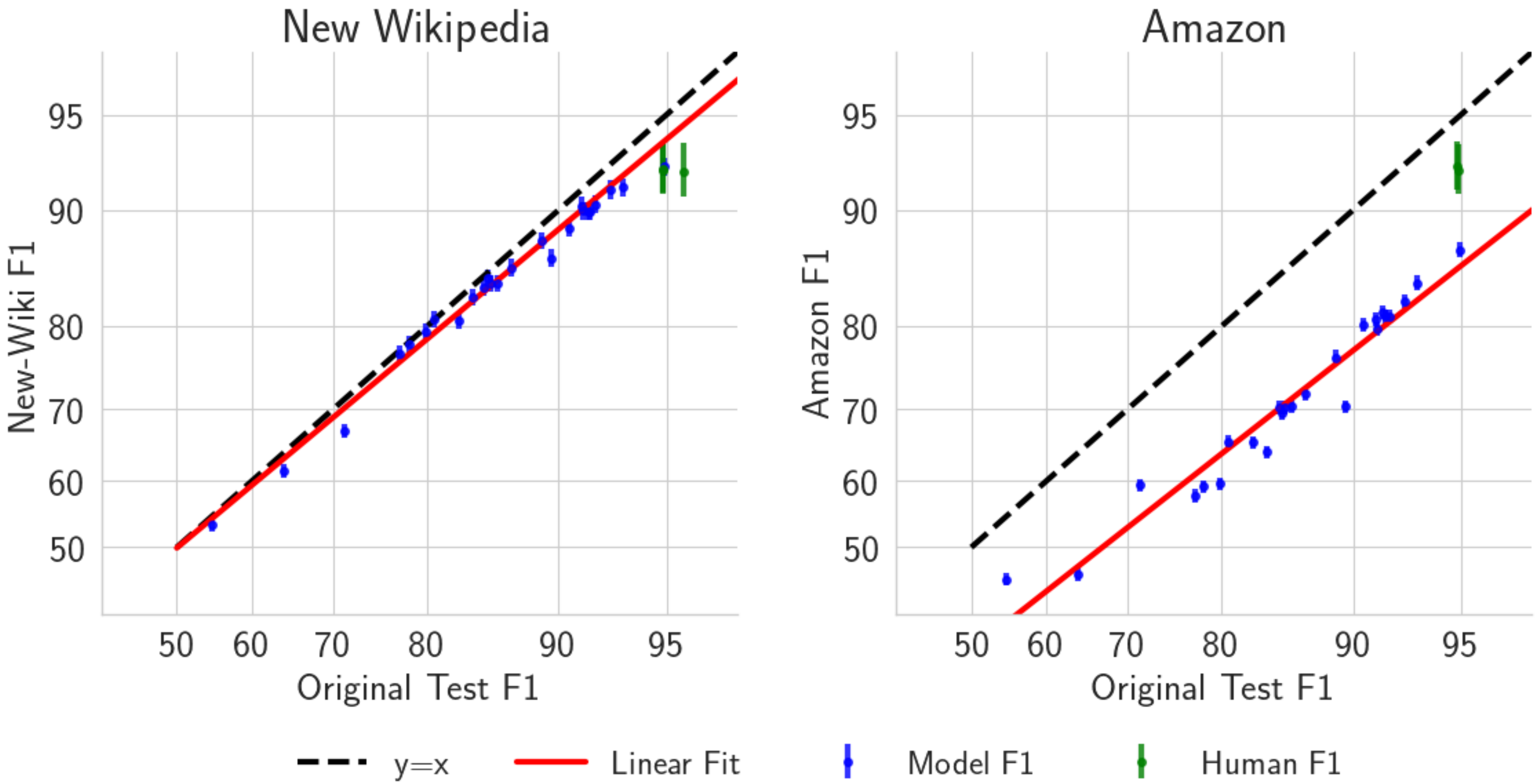
☐ = manipulation included in training data

Figure 4: Classification accuracy (in percent) for networks with potentially distorted training data. Rows show different test conditions at an intermediate difficulty (exact condition indicated in brackets, units as in Figure 3). Columns correspond to differently trained networks (leftmost column: human observers for comparison; no human data available for salt-and-pepper noise). All of the networks were trained from scratch on (a potentially manipulated version of) 16-class-ImageNet. Manipulations included in the training data are indicated by a red rectangle; additionally 'greyscale' is underlined if it was part of the training data because a certain distortion encompasses greyscale images at full contrast. Models **A1 to A9**: ResNet-50 trained on a single distortion (100 epochs). Models **B1 to B9**: ResNet-50 trained on uniform noise plus one other distortion (200 epochs). Models **C1 & C2**: ResNet-50 trained on all but one distortion (200 epochs). Chance performance is at $\frac{1}{16} = 6.25\%$ accuracy.

Beyond Image Classification

SQuAD (Stanford Question Answering Dataset): question answering on paragraphs

➡ Similar trends in natural language processing. [Miller, Krauth, Recht, **Schmidt** ’20]



Distribution Shifts Are a Real Problem

Even in a carefully-controlled reproducibility experiment.

February 2018:

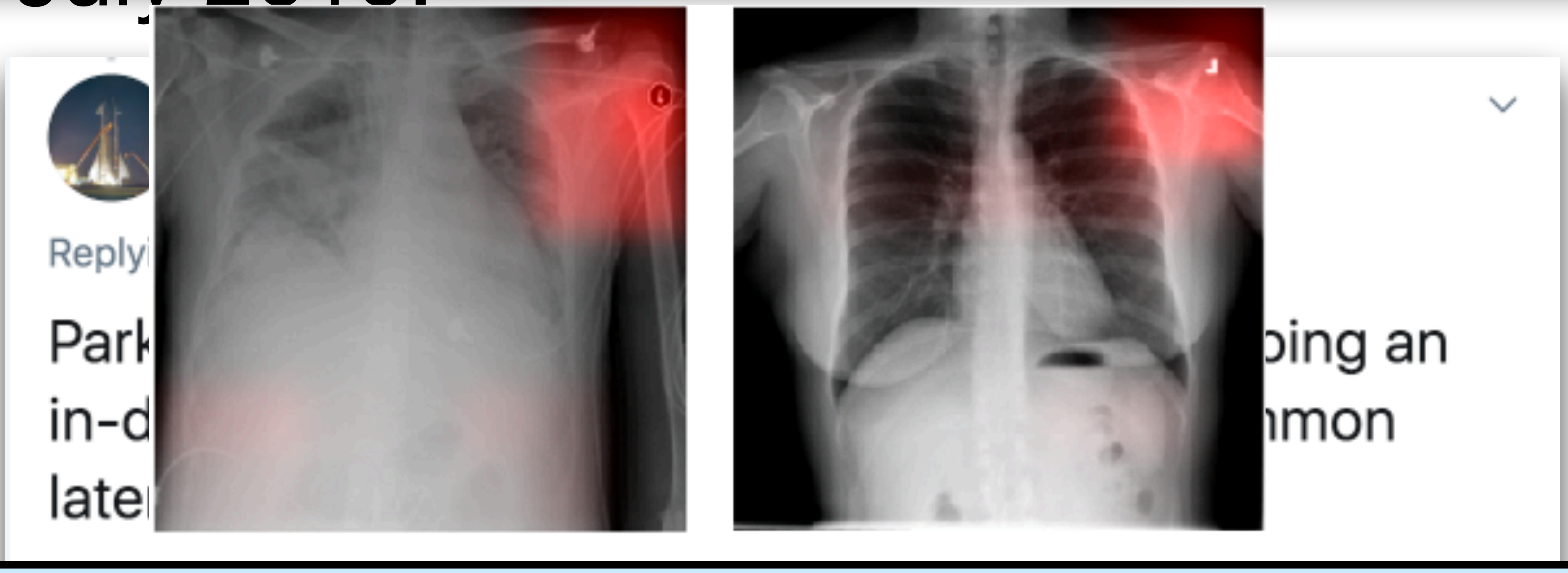
Elon Musk expects to do coast-to-coast autonomous Tesla drive in 3 to 6 months

September 2019: Enhanced Summon

Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, Eric Karl Oermann

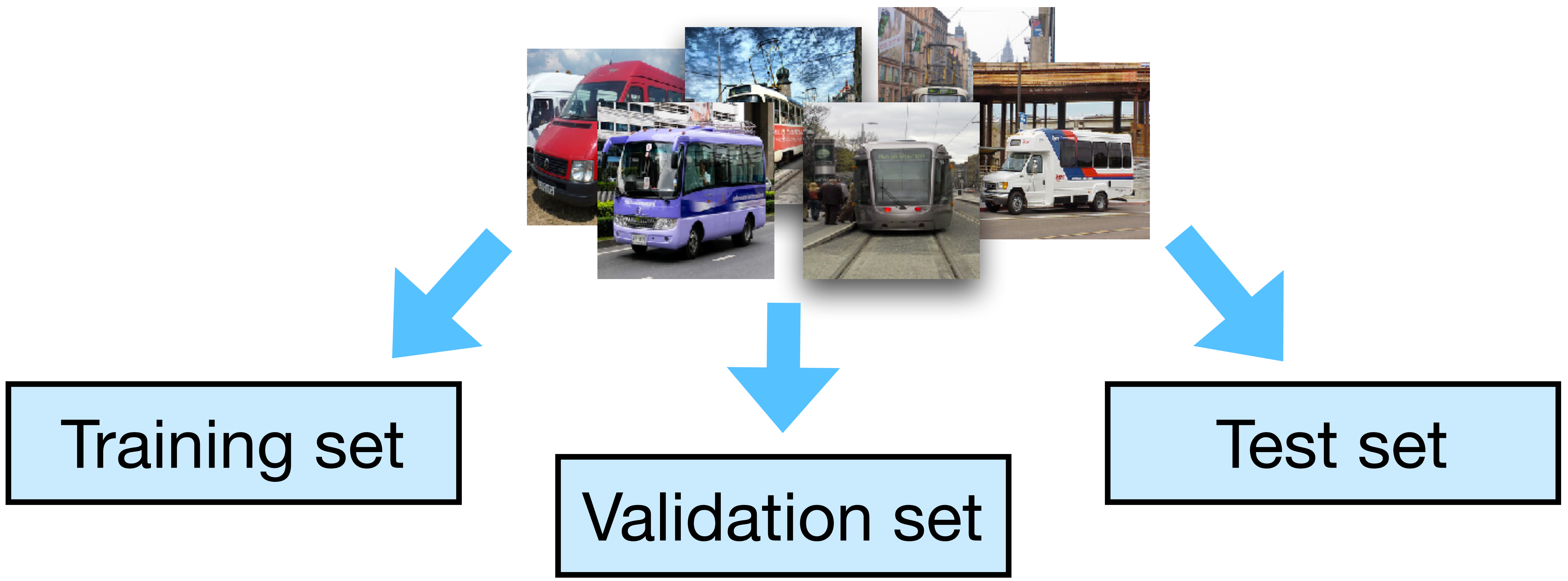
Published: November 6, 2018 • https://doi.org/10.1371/journal.pmed.1002683

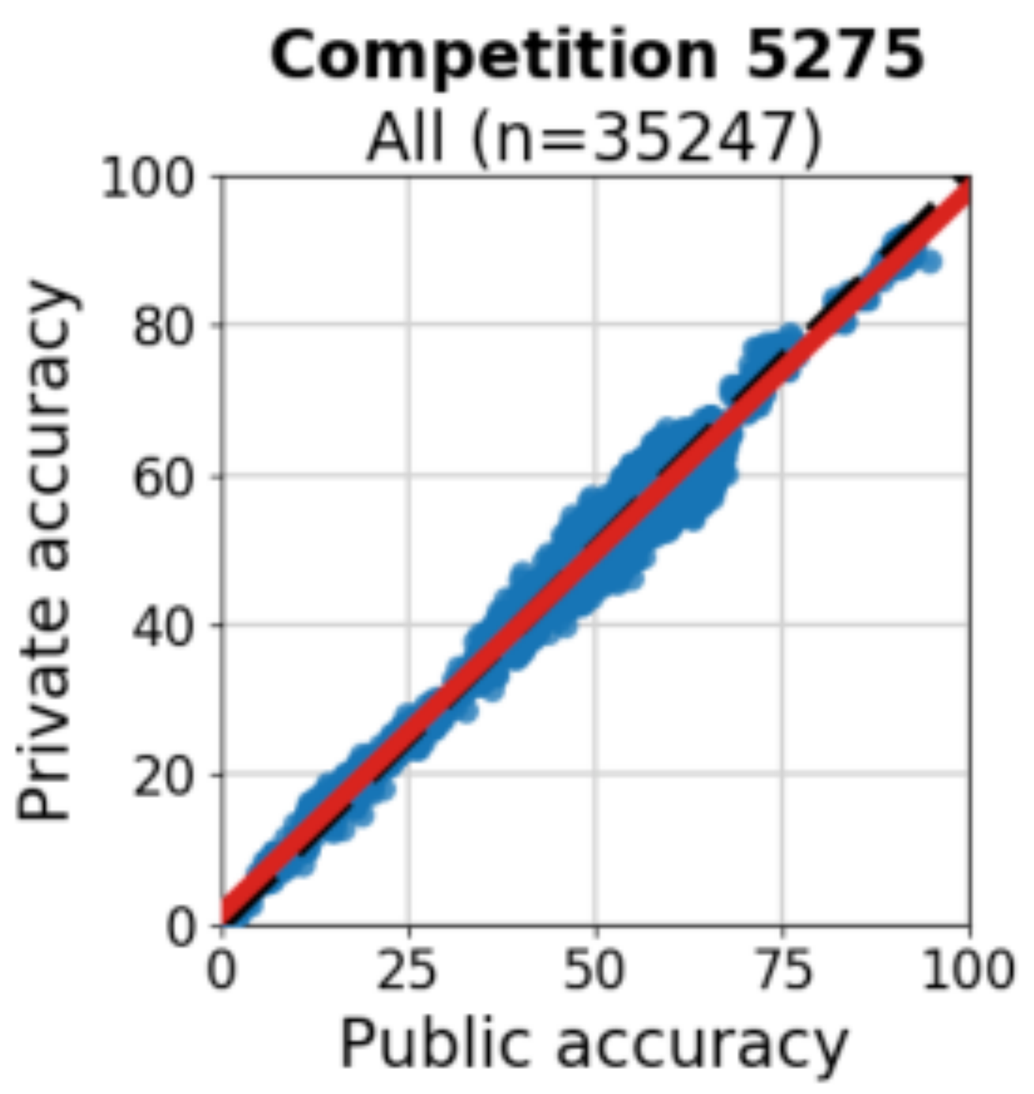


Even in the absence of recognized confounders, we would caution, following Recht and colleagues, that "current accuracy numbers are brittle and susceptible to even minute natural variations in the data distribution".

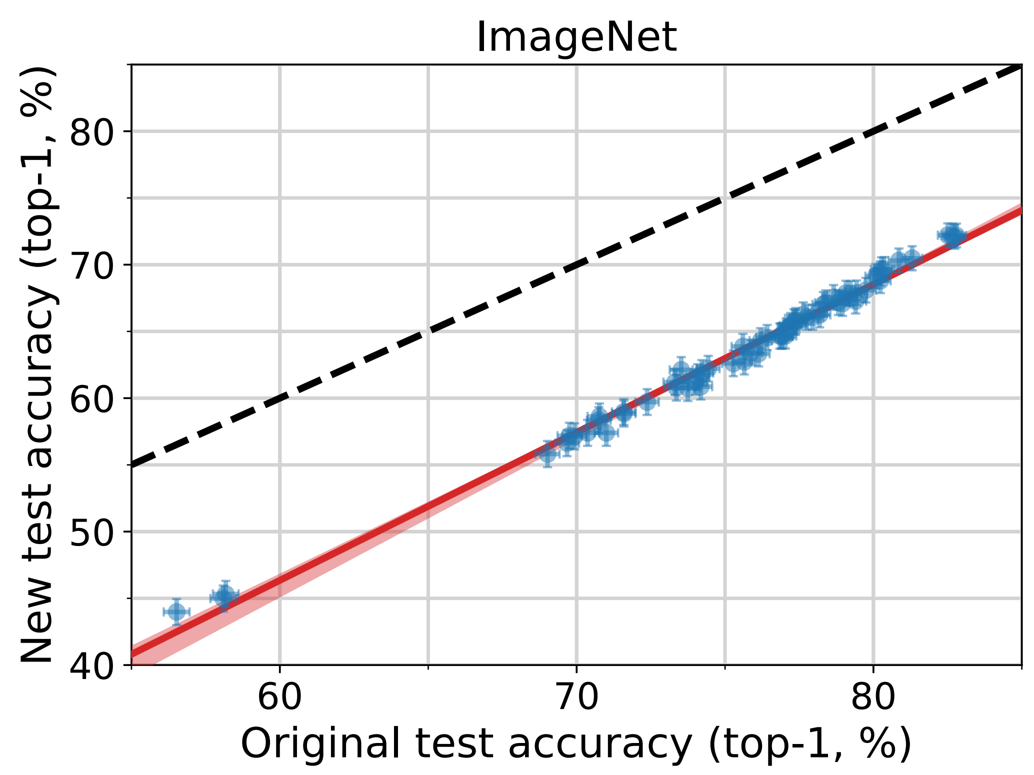
Implications for Evaluating ML

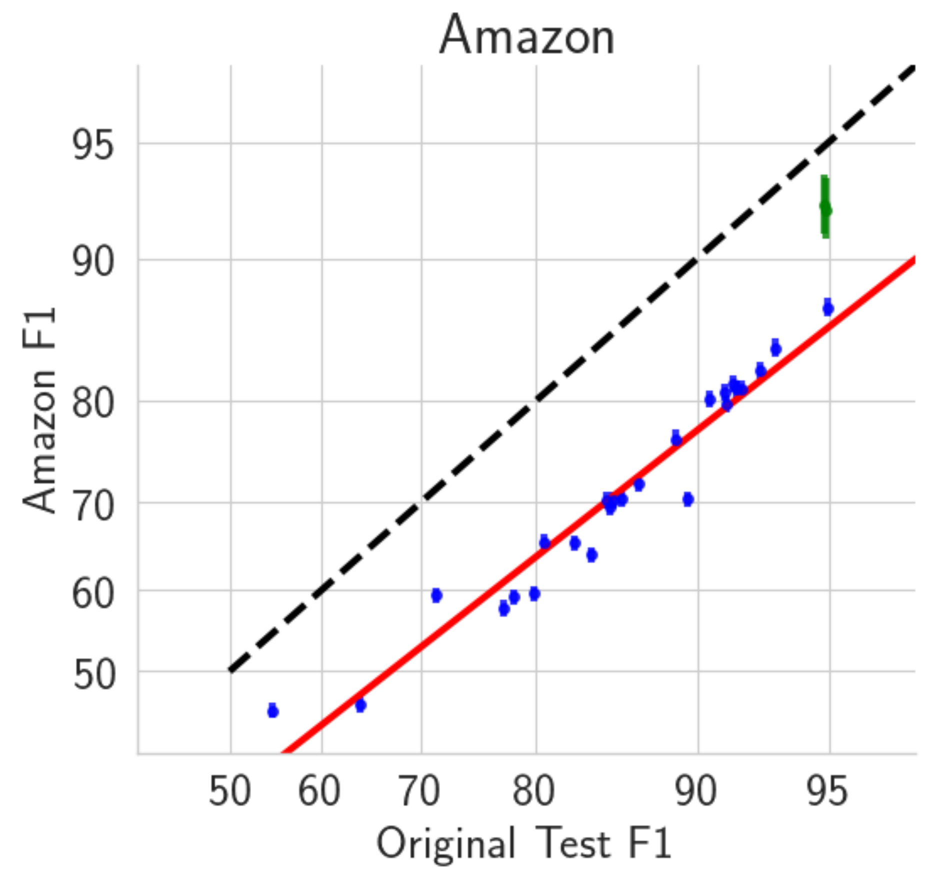
Need to go beyond i.i.d. data splits to measure robustness.





Instead: measure performance with test sets from different distributions.





First Attempt at Broader Evaluation

Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori
UC Berkeley

Achal Dave
CMU

Vaishaal Shankar
UC Berkeley

Nicholas Carlini
Google Brain

Benjamin Recht
UC Berkeley

Ludwig Schmidt
UC Berkeley

Abstract

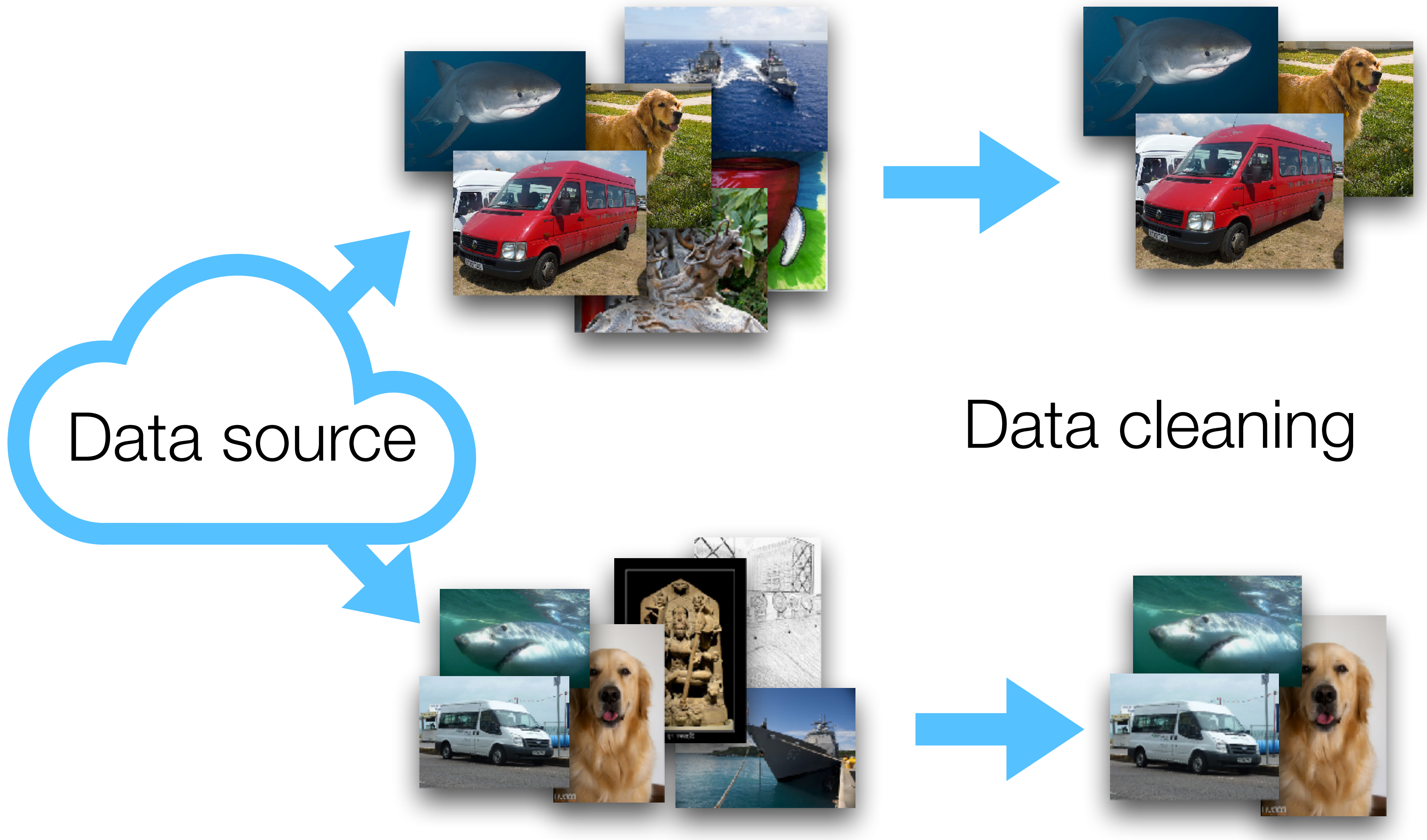
We study how robust current ImageNet models are to distribution shifts arising from natural variations in datasets. Most research on robustness focuses on synthetic image perturbations (noise, simulated weather artifacts, adversarial examples, etc.), which leaves open how robustness on synthetic distribution shift relates to distribution shift arising in real data. Informed by an evaluation of 204 ImageNet models in 213 different test conditions, we find that there is often little to no transfer of robustness from current synthetic to natural distribution shift. Moreover, most current techniques provide no robustness to the natural distribution shifts in our testbed. The main exception is training on larger and more diverse datasets, which in multiple cases increases robustness, but is still far from closing the performance gaps. Our results indicate that distribution shifts arising in real data are currently an open research problem. We provide our testbed and data as a resource for future work at https://modestyachts.github.io/imagenet-testbed/.

Synthetic vs Natural

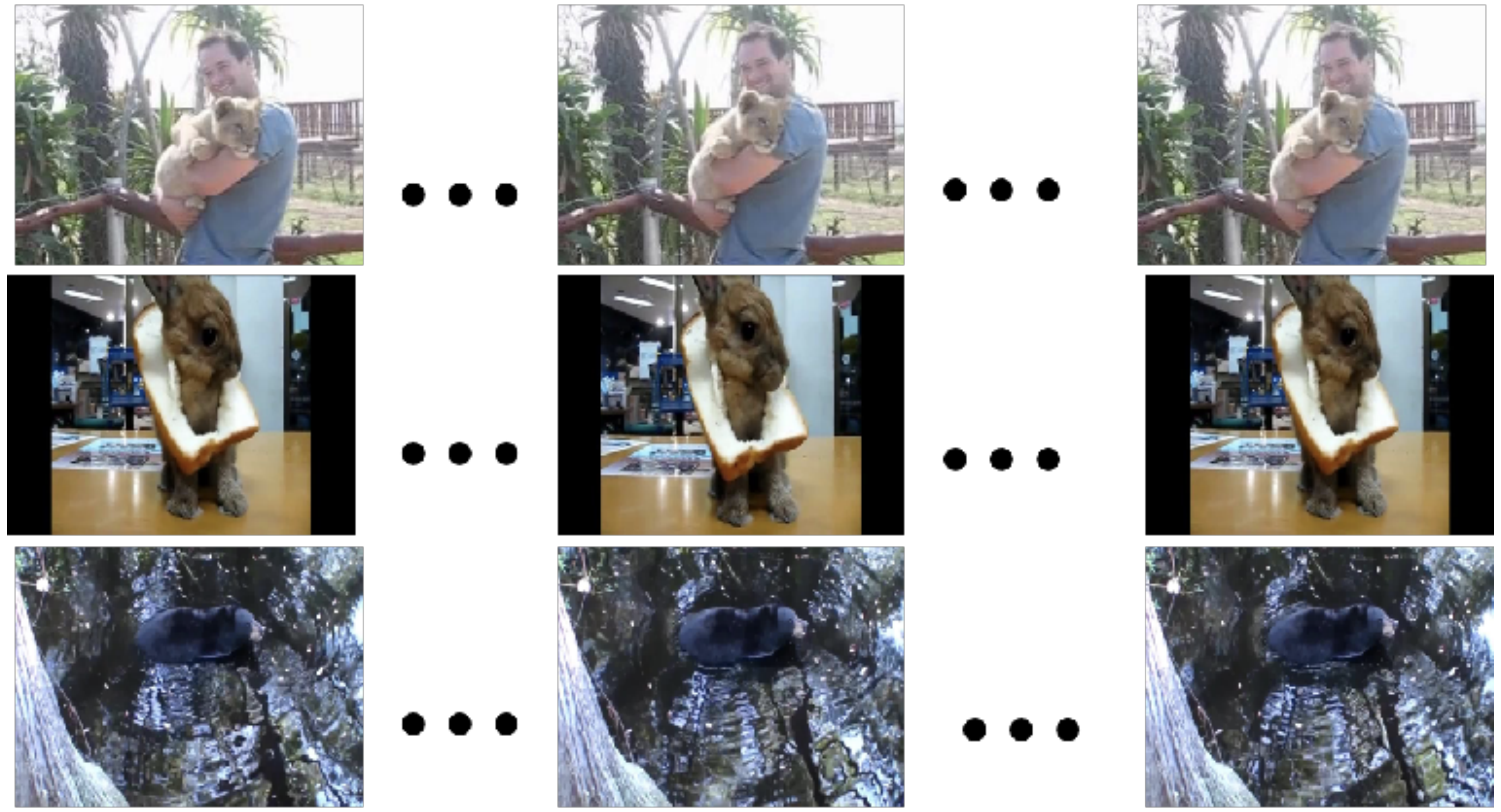
Synthetic: computer-generated perturbations of a real dataset

D = {  + f() =  }

Natural: images as they were recorded



ImageNetV2



ImageNet-Vid-Robust

[Shankar, Dave, Roelofs, Ramanan, Recht, Schmidt '19]

Overview

→ Are current vision models robust to natural distribution shift?

1. Define what it means to be robust to distribution shift.

2. Evaluate 200+ models on 200+ distribution shifts.

3. Results on 3 “flavors” of natural distribution shifts.

Overview

Are current vision models robust to natural distribution shift?

1. Define what it means to be robust to distribution shift.

2. Evaluate 200+ models on 200+ distribution shifts.

3. Results on 3 “flavors” of natural distribution shifts.

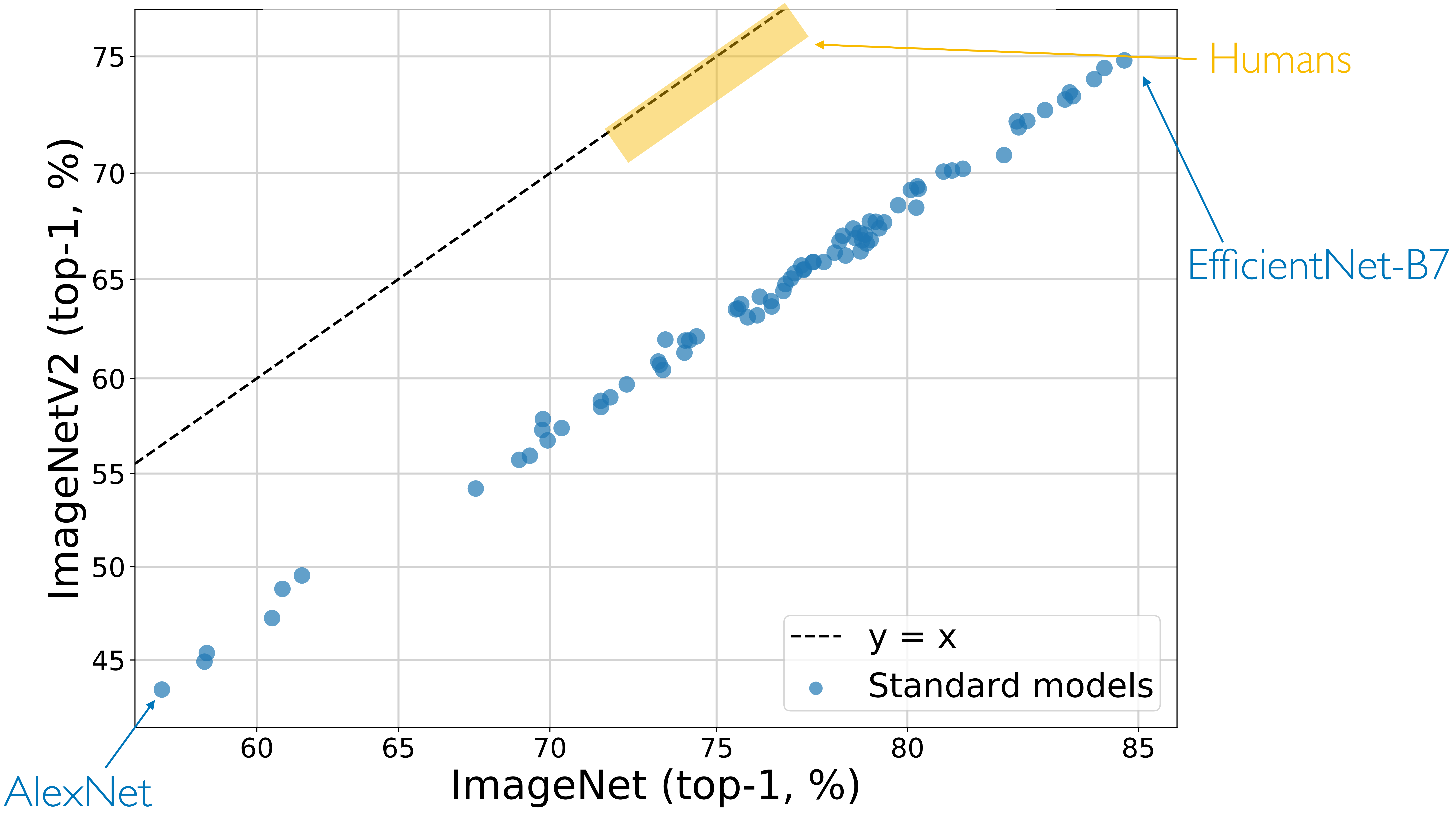
Hypothetical Models

	In-distribution (Source) Accuracy	Out-of-distribution (Target) Accuracy
Model A	80%	75%
Model B	90%	77%

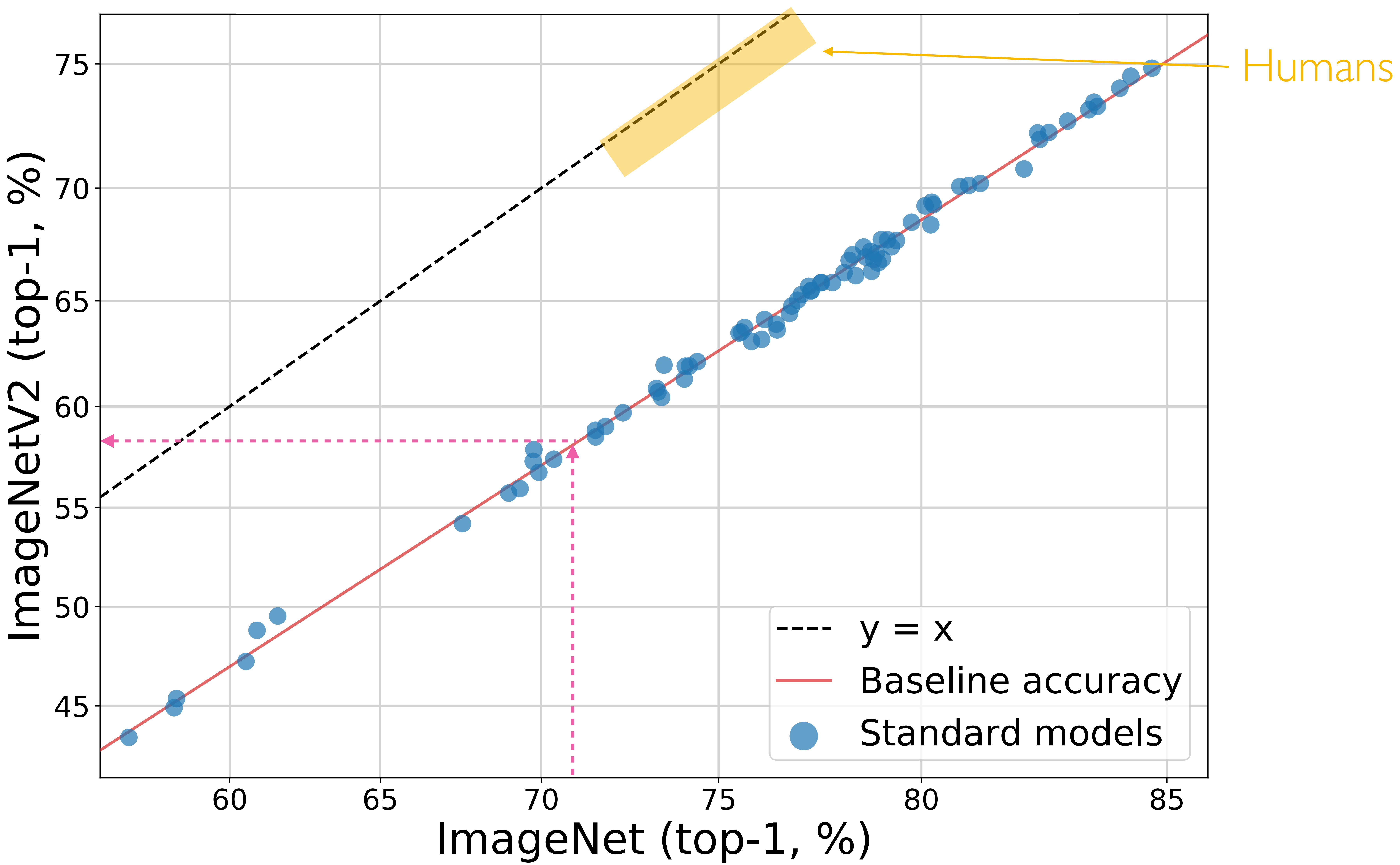
Hypothetical Models

	In-distribution (Source) Accuracy	Out-of-distribution (Target) Accuracy	Accuracy Drop
Model A	80%	75%	5%
Model B	90%	77%	13%

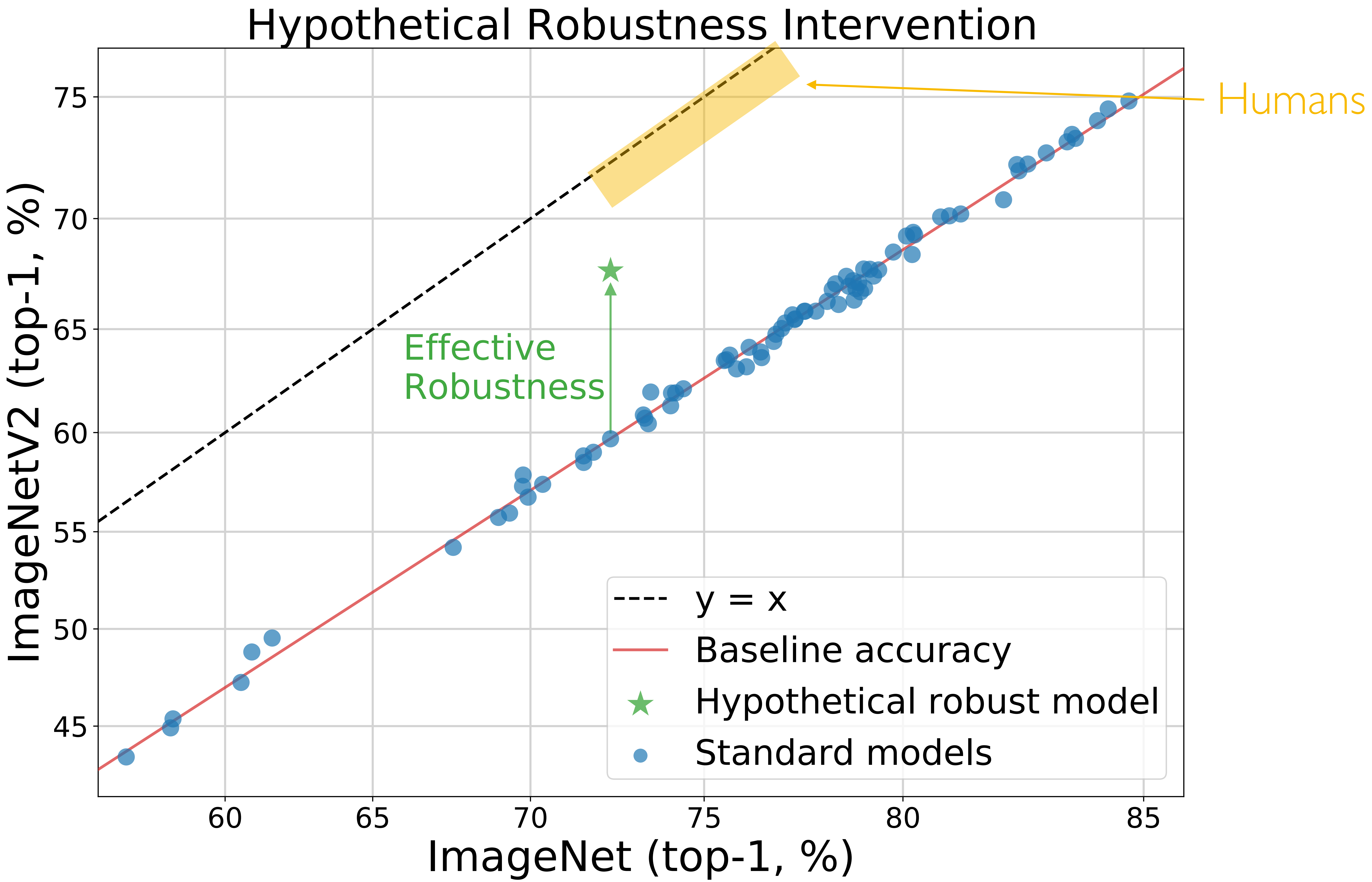
How do we compare models with different in-distribution accuracy?



Humans
EfficientNet-B7
AlexNet
ImageNetV2 (top-1, %)
ImageNet (top-1, %)
y = x
Standard models
75
70
65
60
55
50
45
60
65
70
75
80
85



Humans
y = x
Baseline accuracy
Standard models
ImageNet (top-1, %)
ImageNetV2 (top-1, %)
60
65
70
75
80
85
45
50
55
60
65
70
75



Do **any** current models achieve effective robustness?

Overview

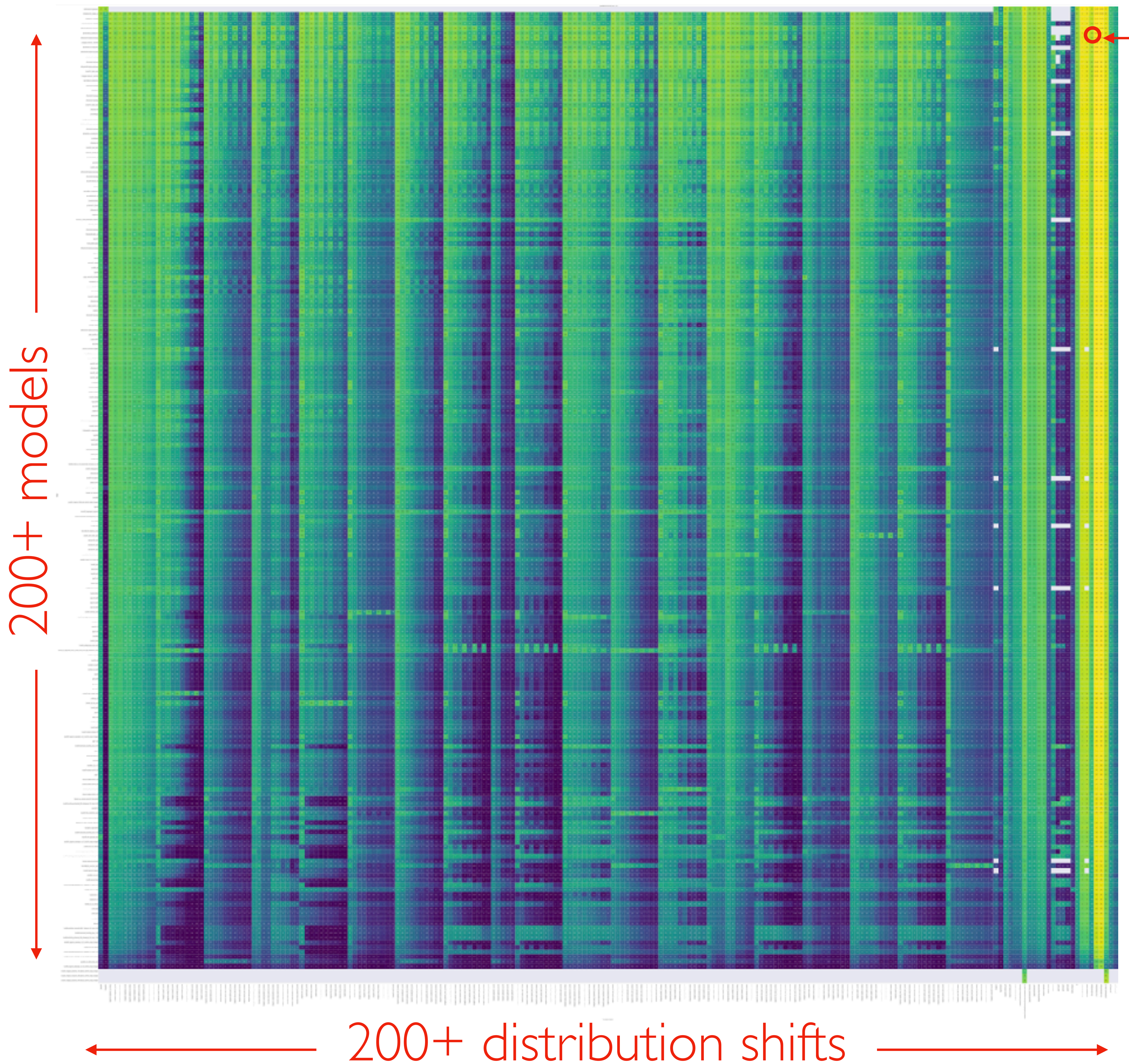
Are current vision models robust to natural distribution shift?

1. Define what it means to be robust to distribution shift.

2. Evaluate 200+ models on 200+ distribution shifts.

3. Results on 3 “flavors” of natural distribution shifts.

Our Testbed



1 cell = 1 model evaluation on 1 dataset (total 10^9 model evaluations).

Models:

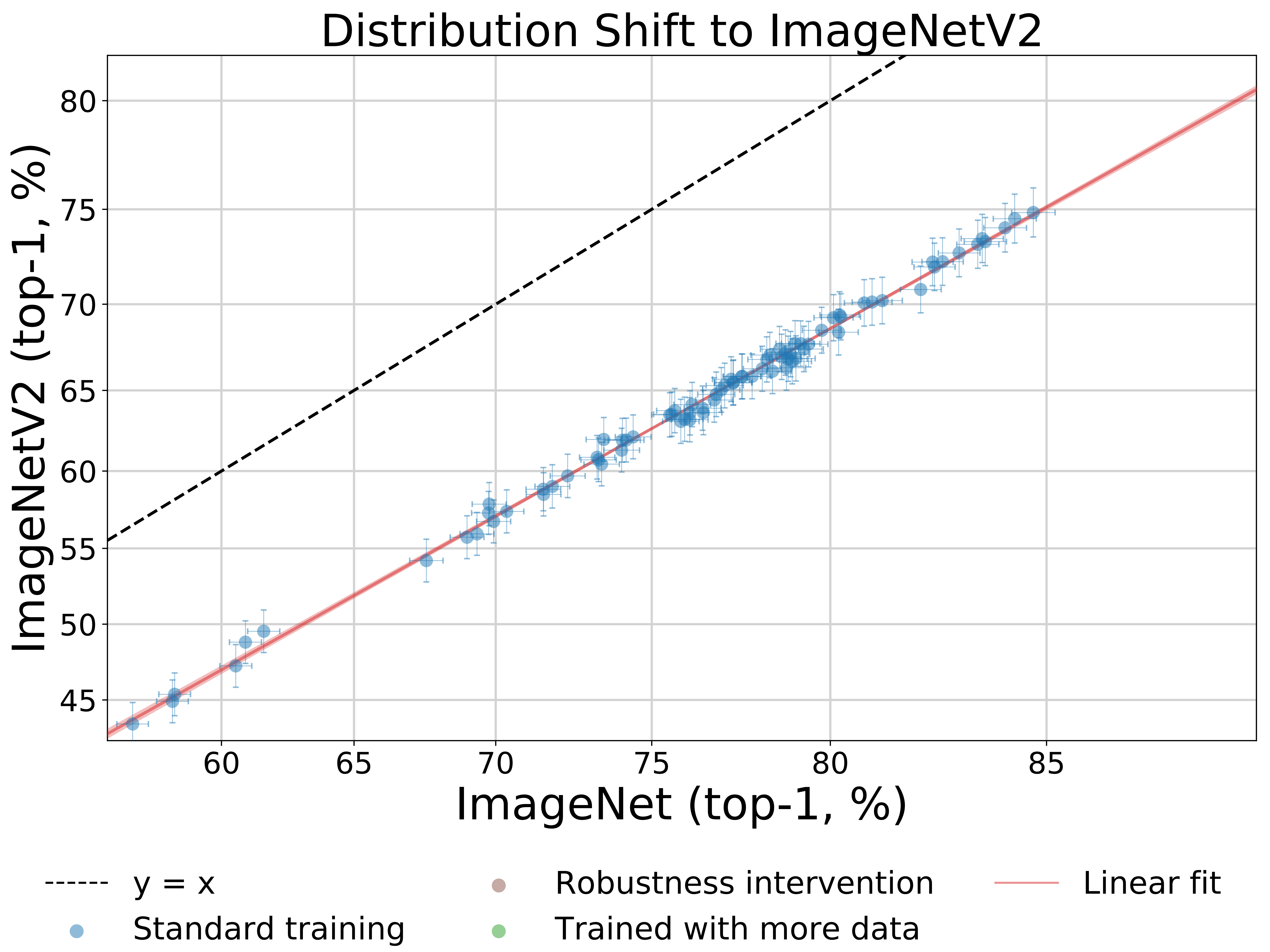
- standard models
- robust models (adversarially robust models & models with special data augmentation)
- models trained on more data

Natural distribution shifts:

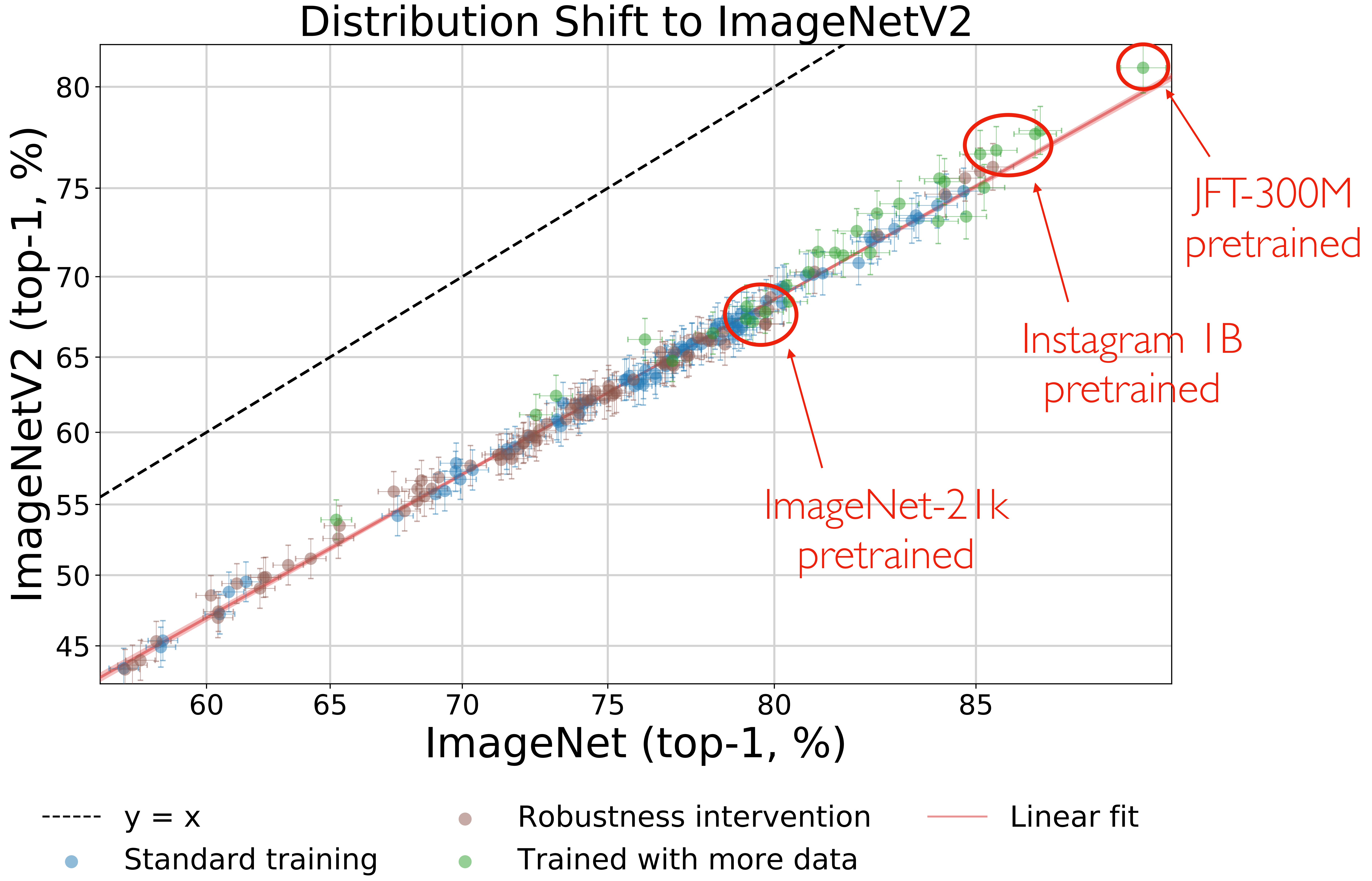
- ImageNetV2, ObjectNet, ImageNet-Vid-Anchors, YTBB-Anchors
- ImageNet-Vid-Robust, YTBB-Robust (video frames)
- ImageNet-A (adversarially filtered)

Synthetic distribution shifts:

- Lp-attacks & image corruptions

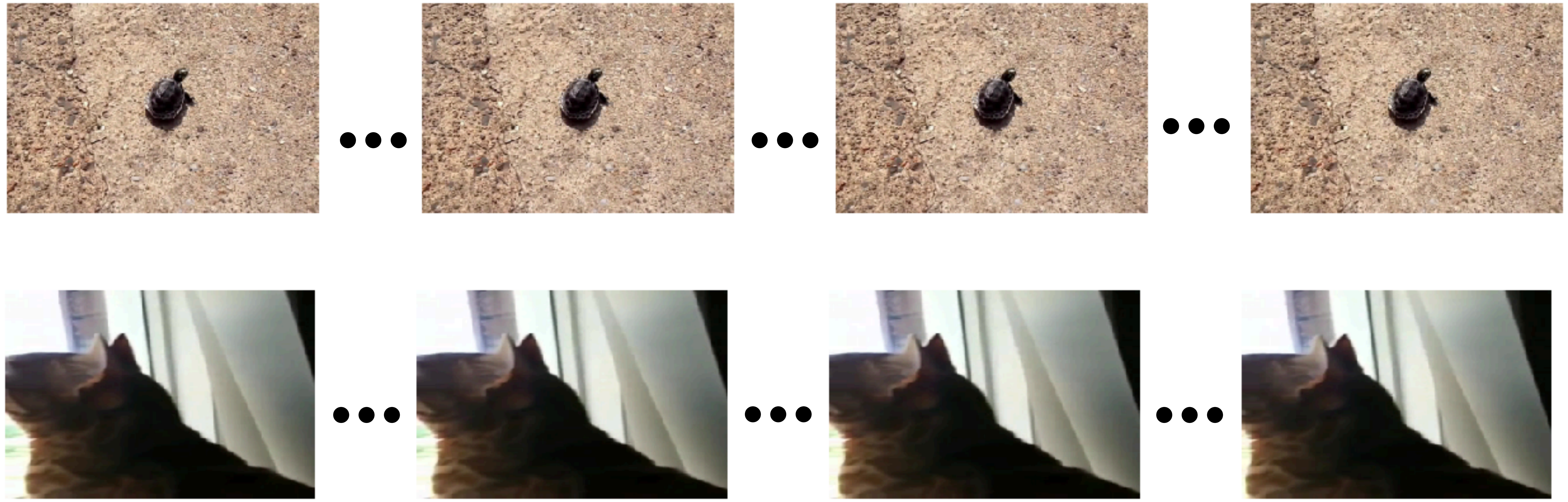


Do **any** current models achieve effective robustness?



Takeaway: Most models and robustness strategies provide no additional robustness.

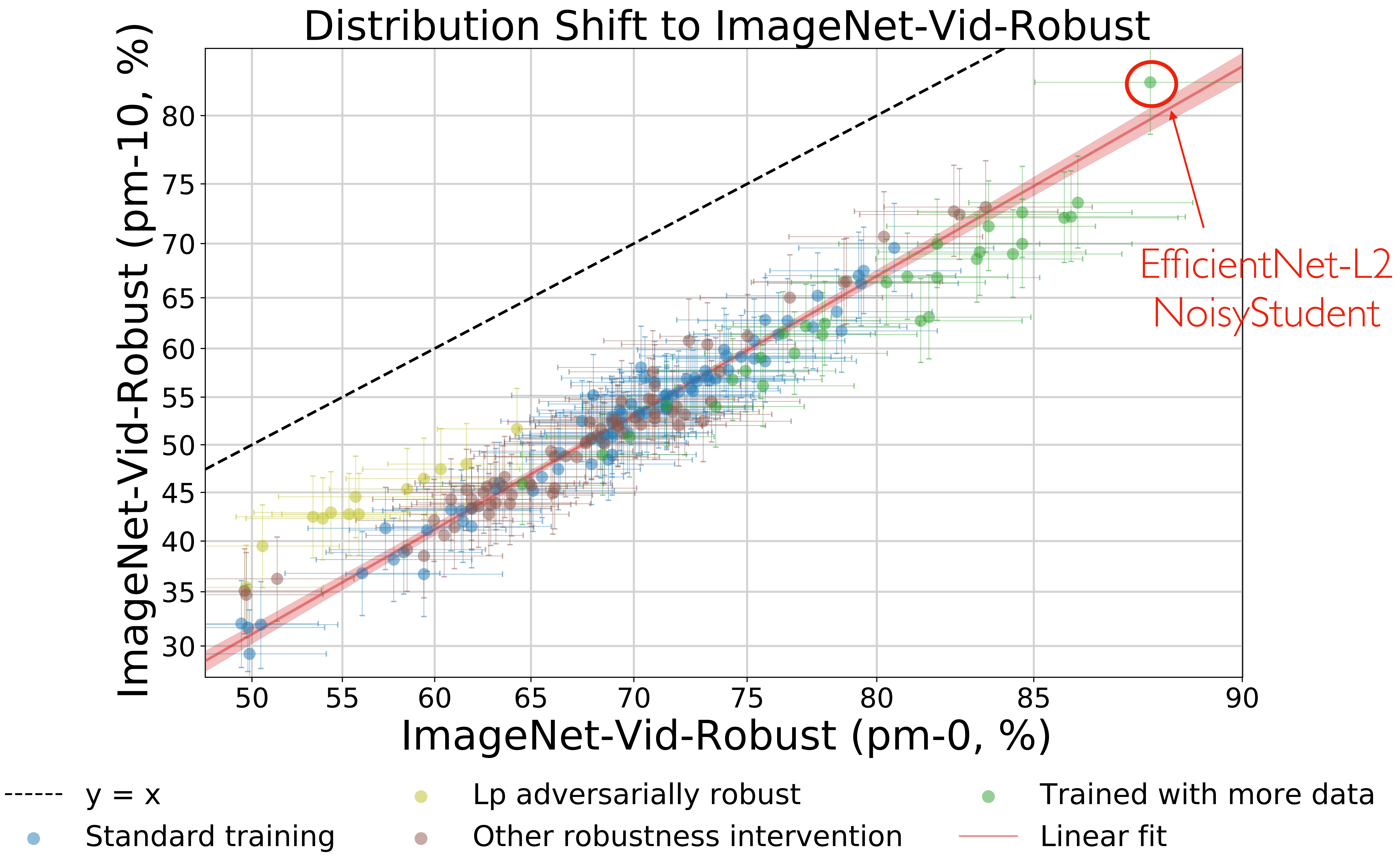
ImageNet-Vid-Robust



pm-k metric: video sequence is correctly classified only if the anchor frame and surrounding k frames (plus-minus k) are also correctly classified

pm-0: accuracy on anchor frames
pm-10: sequence is correct if anchor frame ± 10 frames are correctly classified

[Shankar, Dave, Roelofs, Ramanan, Recht, Schmidt '19]



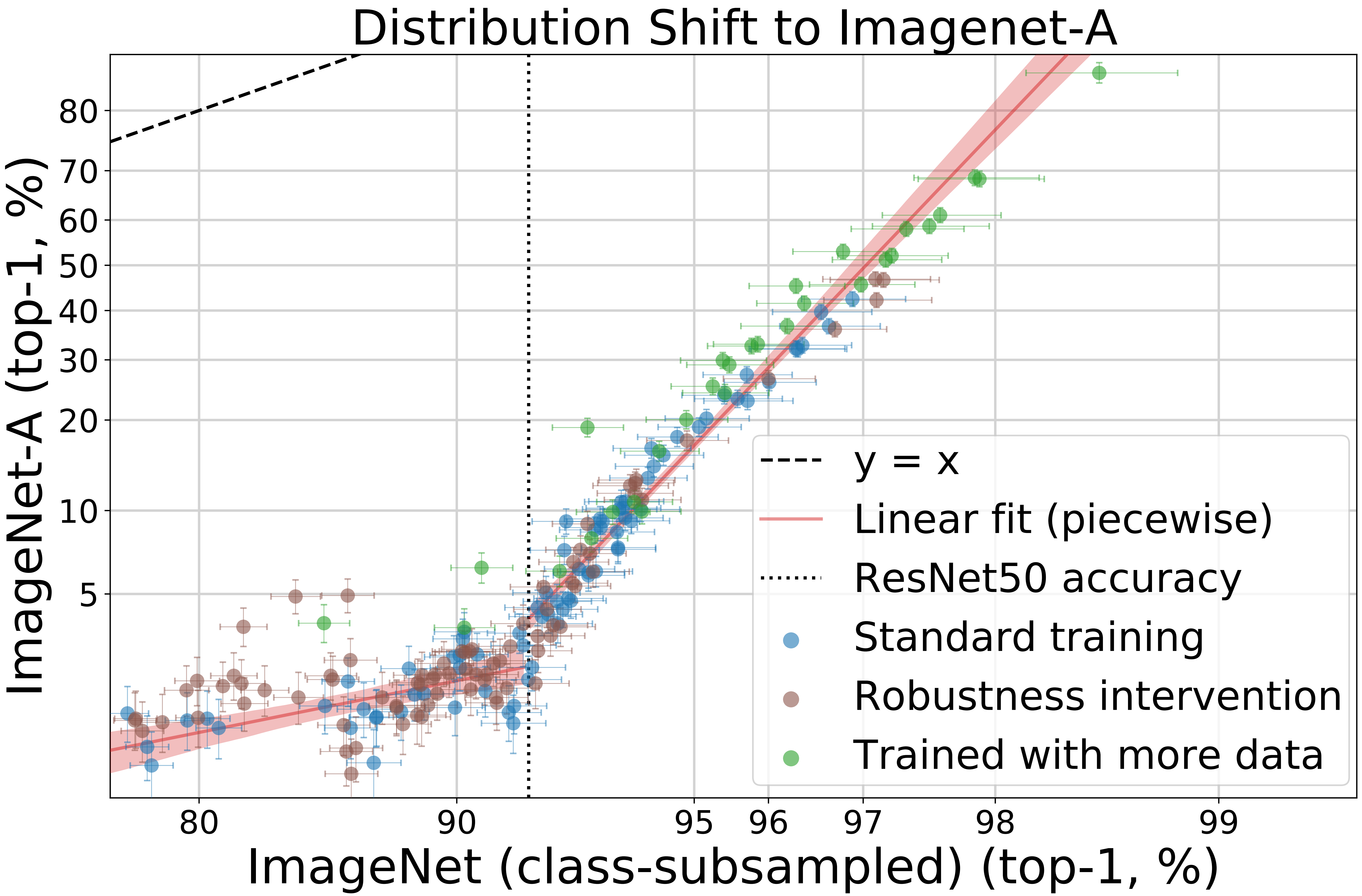
Takeaway: Adversarially robust models have effective robustness (in low-accuracy regime).

ImageNet-A
(Adversarially Filtered Shift)

1. Download a large number of labeled images from online.
2. Select only the subset that was misclassified by a ResNet-50 model.

[Hendrycks, Zhao, Basart, Steinhardt, Song '19]



Takeaway: Adversarial filtering creates a "knee" in the response curve. Initial accuracy drops are large, but higher accuracy models quickly make progress in closing the gap.

Summary

- We analyzed 200+ ImageNet models and 200+ datasets.
- We find most models & robustness strategies provide little to no effective robustness on current natural distribution shifts.
- Two concrete recommendations for researchers moving forward:
  1. Control for standard accuracy (look at effective robustness).
  2. Evaluate on natural distribution shifts.

https://tinyurl.com/imagenet-testbed

1. Empirical progress in machine learning: benchmarks

   Main paradigm: experiments, experiments, experiments

2. What can we learn from ML benchmarks?

   If done well: performance trends across a range of tasks and methods

3. Limitations of current ML methods

   Many settings going beyond i.i.d. performance

Discussion Part!

Why I Like ML Benchmarks

Opinion: Benchmarks are the only reliable framework we currently have to scale the “scientific method” to the entire ML community.

Admittedly, we often don’t learn much in terms of science (causal relationship between algorithmic interventions and performance, broad principles, etc.)

But at least methods get better and we can compare methods reliably

➡ **Falsifiable** statements about model performance (this is non-trivial)

There are certainly uninformative benchmarks (no generalizable knowledge)

Issues with ImageNet

ImageNet was **not built for what it has become** (this is **not** a fault of the authors).

Full ImageNet (21k classes) contained images for **racial slurs, “rape suspect”, etc.**

- Should not be part of a dataset
- Harmful for crowdworkers

Biased representation of humans

Three human classes: groom, scuba diver, baseball player

Many humans in images for other classes (dogs, ping pong ball, instruments, etc.)

Biased towards **affluent countries**

Humans did not provide **consent** (+ unclear licensing)

Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.

We should be specific about what datasets are for and what they aren't.

What Kind of Science is Machine Learning?

2000 - 2010

Empirical progress usually goes **hand in hand with theoretical results**

More like **physics**?

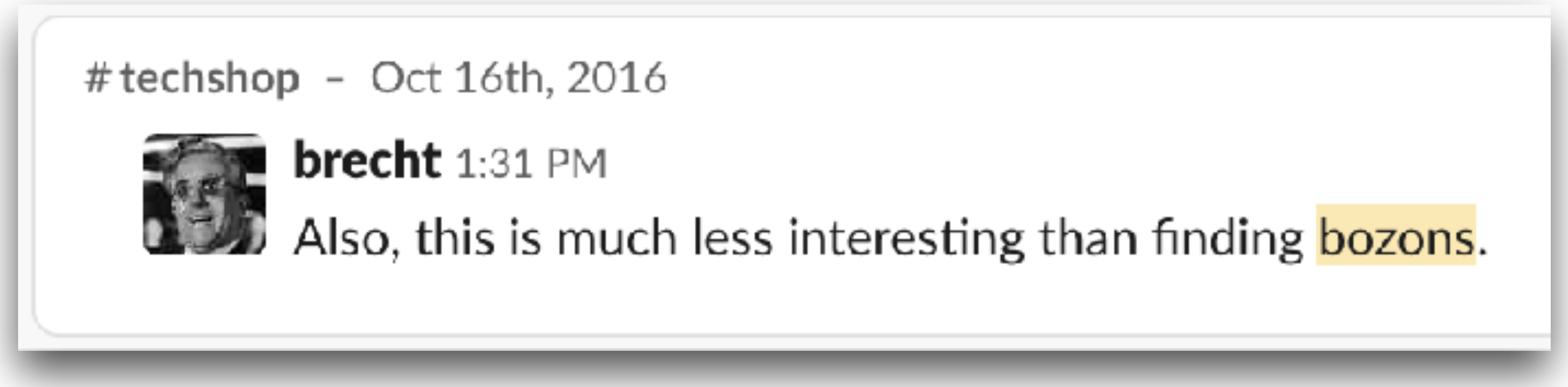
More analytical

2010 - 2020

Empirical progress usually comes **without mathematical theory**

More like **biology**?

More descriptive



Maybe comparing machine learning to a science is wrong to begin with

Is it more an **engineering discipline**? Chemical engineering? Medicine?

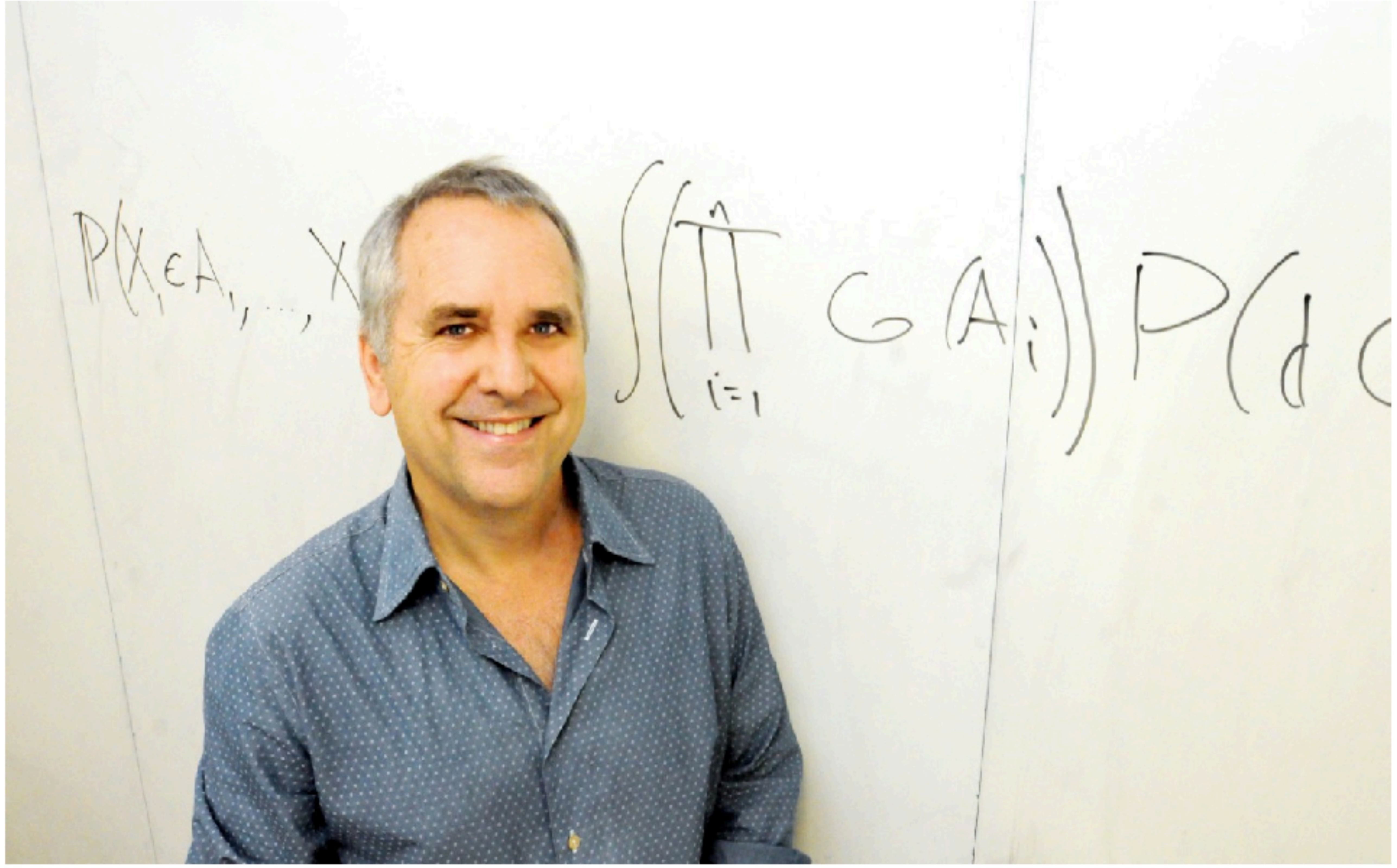


Photo credit: Peg Skorpinski

Artificial Intelligence — The Revolution Hasn't Happened Yet

Michael Jordan Apr 18, 2018 · 16 min read

95% on a Benchmark Can Be Science

The New York Times

The Road to a Coronavirus Vaccine | Vaccine Tracker | FAQ: Moderna Vaccine | FAQ: Pfizer's Vaccine | After the First Vaccine | Long-Term Safety

Early Data Show Moderna's Coronavirus Vaccine Is 94.5% Effective

Moderna is the second company to report preliminary results from a large trial testing a vaccine. But there are still months to go before it will be widely available to the public.

Moderna Therapeutics in Cambridge, Mass. Tony Luong for The New York Times

We didn't know what to expect
(Fauci said his guess was 70 - 75%)

There was / is a rigorous process to to validate the vaccine

Vaccine development went through a sequence of partially principled, partially heuristic steps

Culmination of decades of experimental work in biology (**extremely** impactful)

Will be injected into billions of people without a formal correctness proof

Role of Theory in ML

Good question! I don't see a simple answer.

Two modes for mathematical contributions in TCS:

- **Pure mathematics** (e.g., P vs NP). No need for connections to practice.
- **Theoretical physics**. *Some* empirical grounding - how much?

Divergence of practical ML from theory over the past 10 years

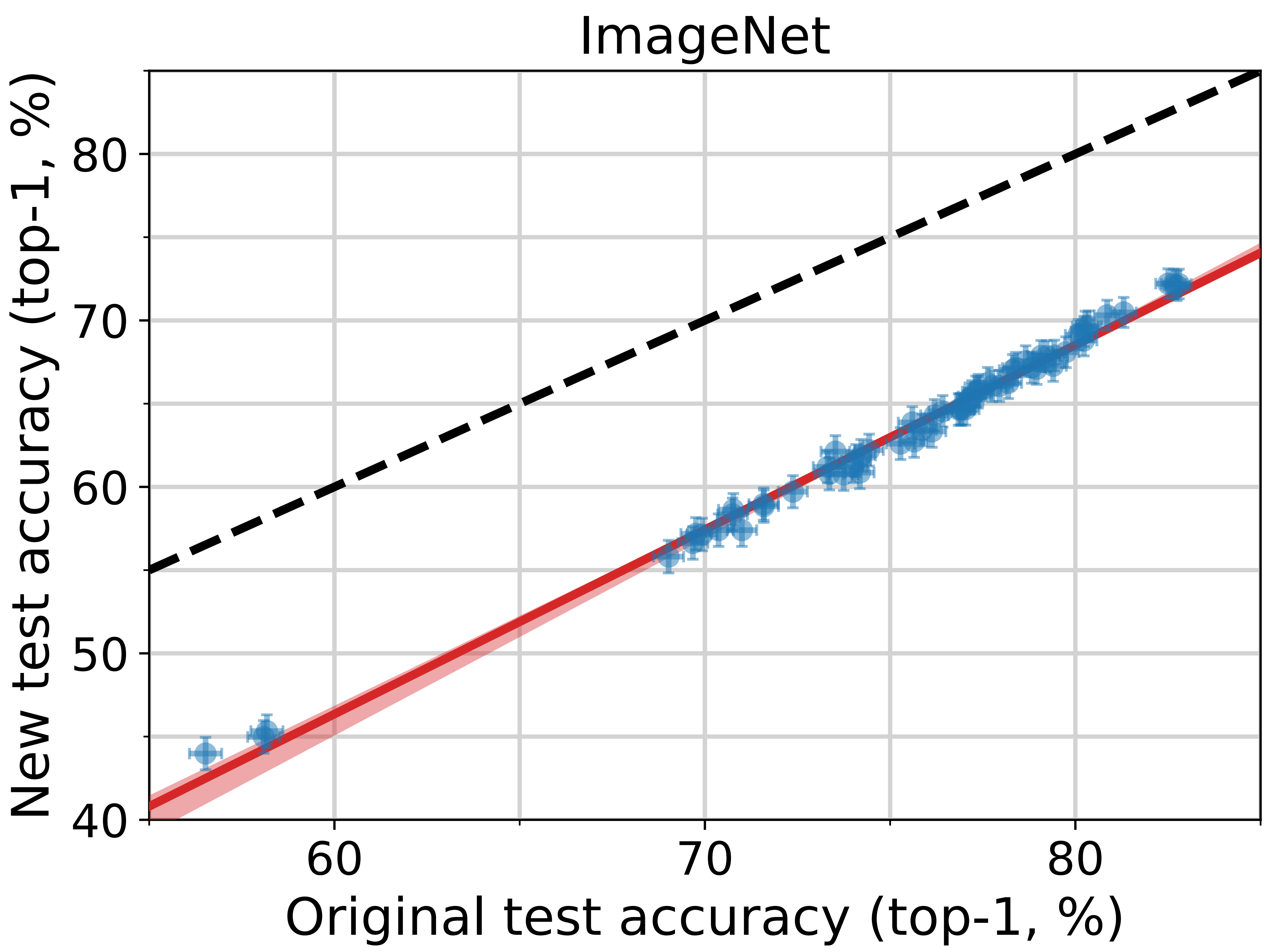
This can be an opportunity: there may be a unifying theory we haven't found yet.

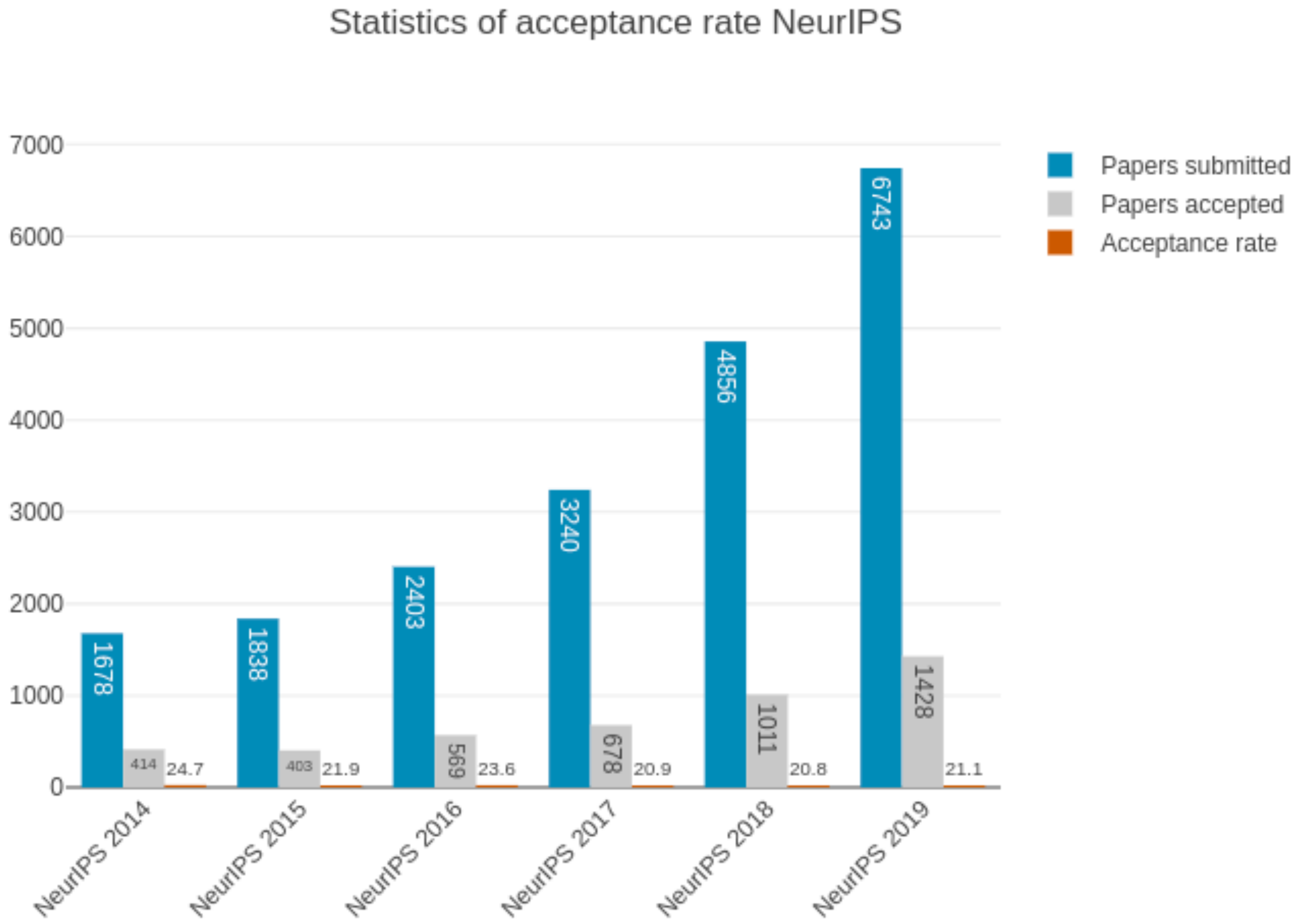
There is also the danger of losing touch with reality (c.f. criticisms of string theory).

→ On average more experiments are a good idea, but depends on the project.

Large Need for Rigor

How can we build reliable knowledge about machine learning?





Theoretically-trained researcher bring a different mindset and toolkit to empirical ML.

Future Directions

Beyond i.i.d. performance

Evaluations: what do we want our models to be robust to?

How can we make the models more reliable?

"Theory you can **plug numbers in**", e.g., for training set scaling

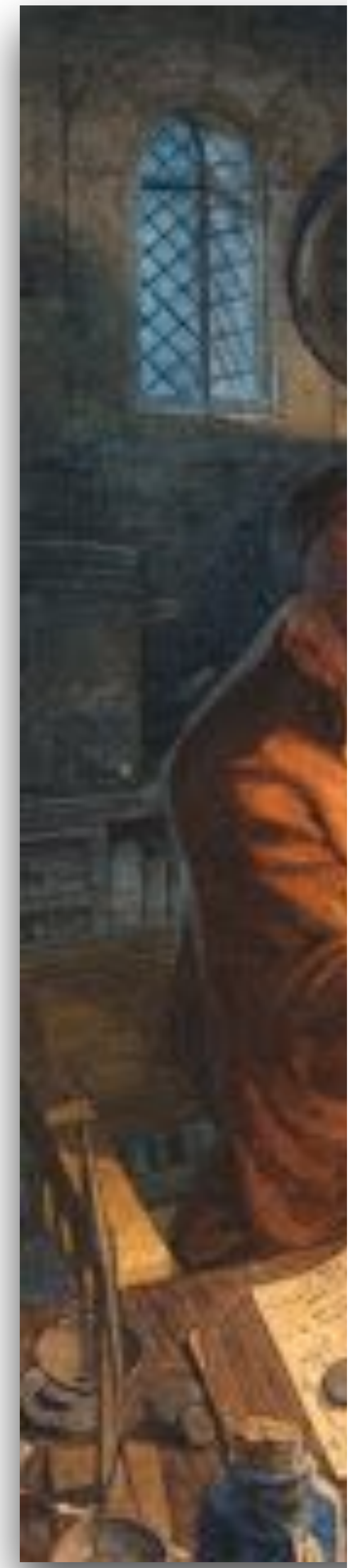
Could be extremely useful if we can reliably train on large training sets

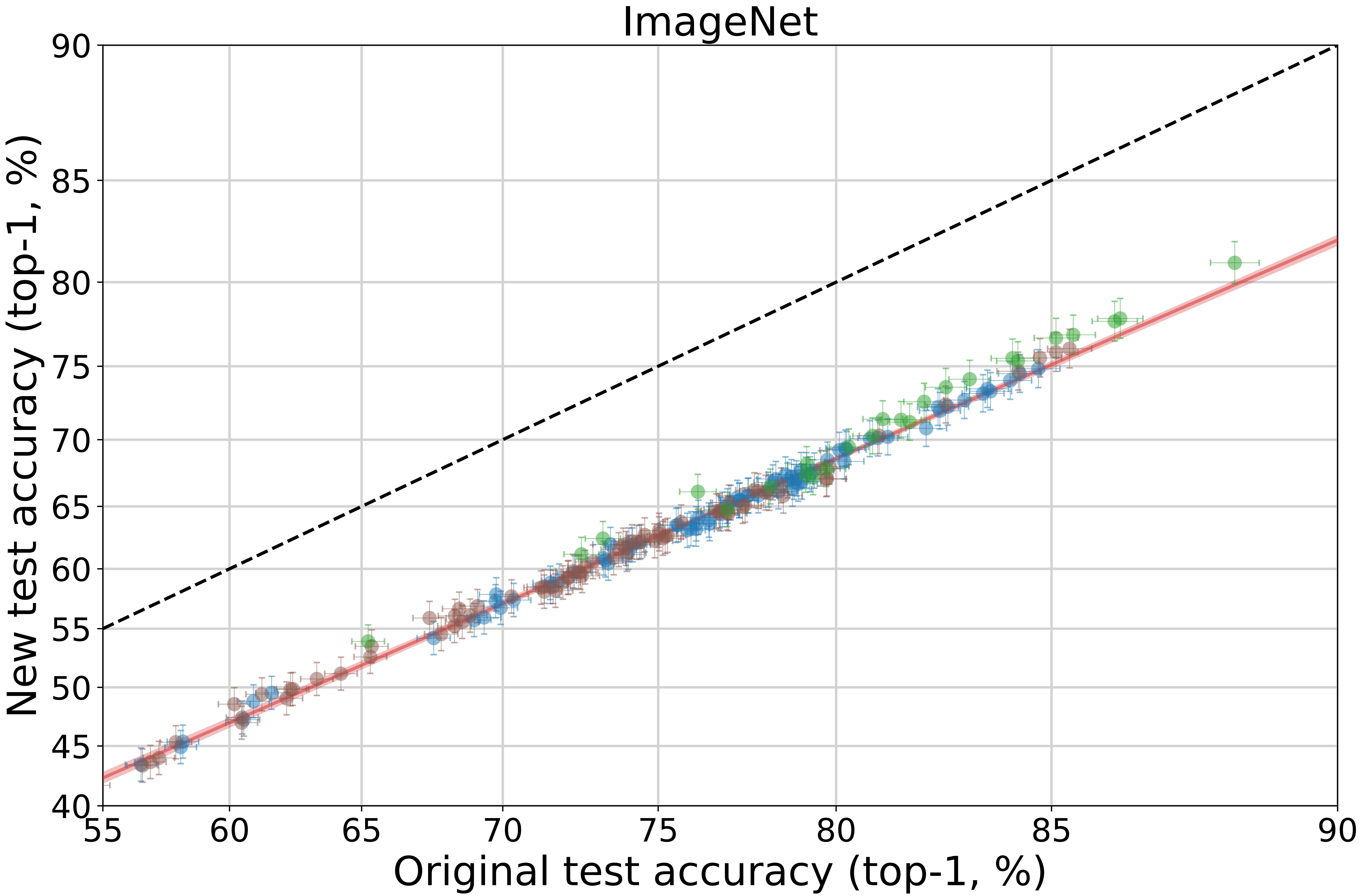
Datasets as a research topic

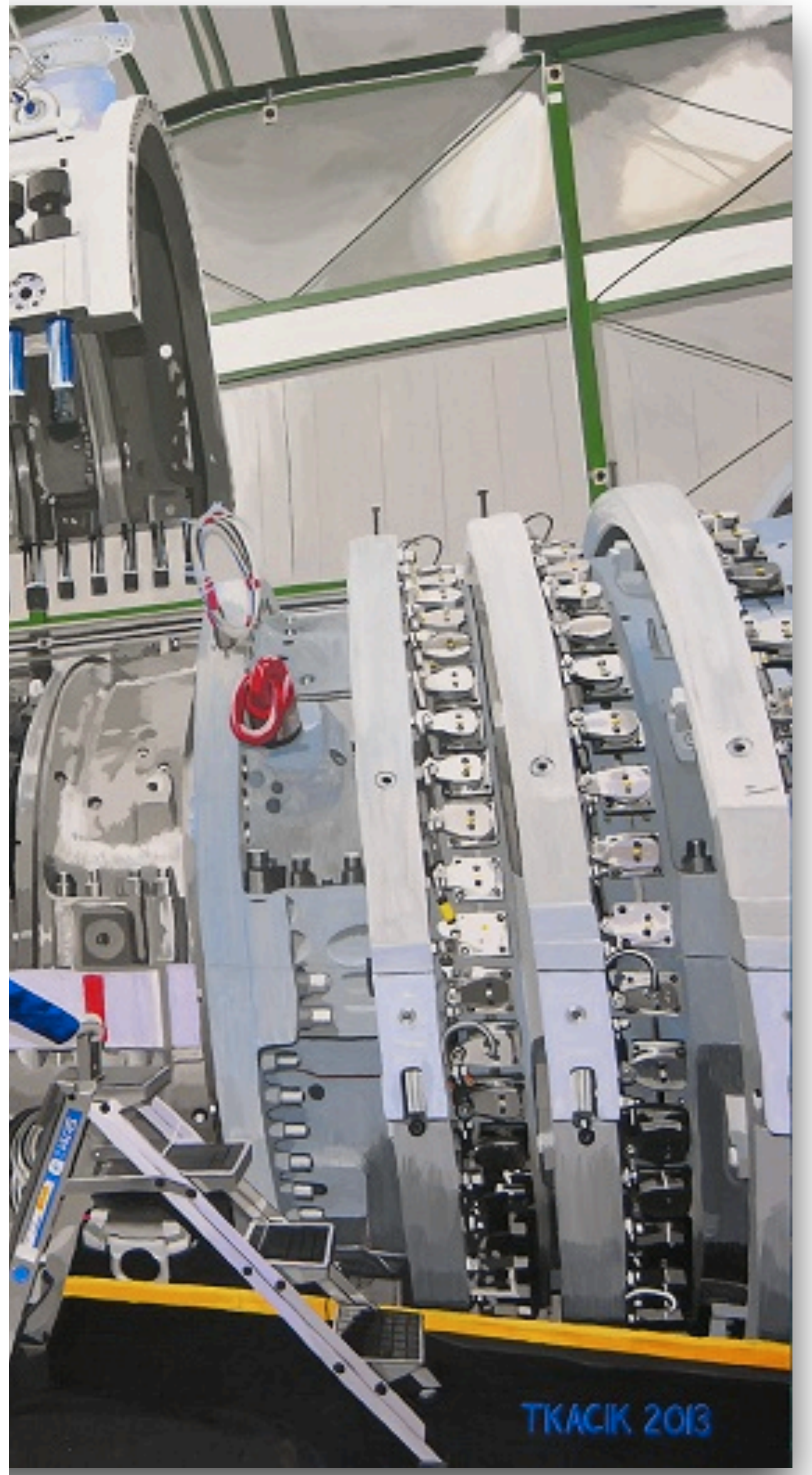
The past 10 years have focused on **model improvements**

We know relatively little about how to build **"good" datasets**

For instance, what makes **ImageNet** a "good" dataset?







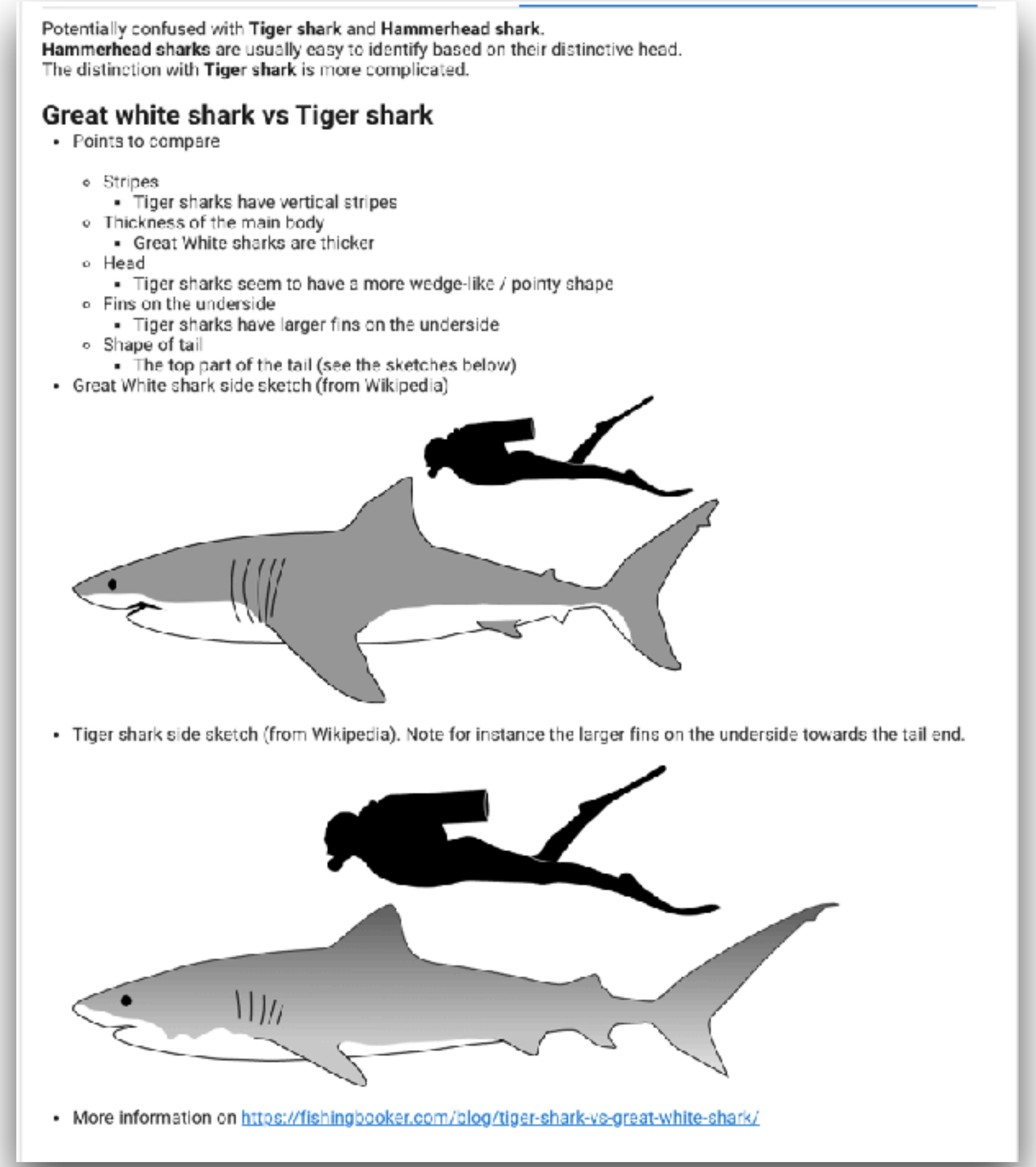
Sometim … ing discipline

Measurement is the contact of reason with nature.

Henry Margenau (1959)

Training humans for high performance

We created a **labeling guide**:

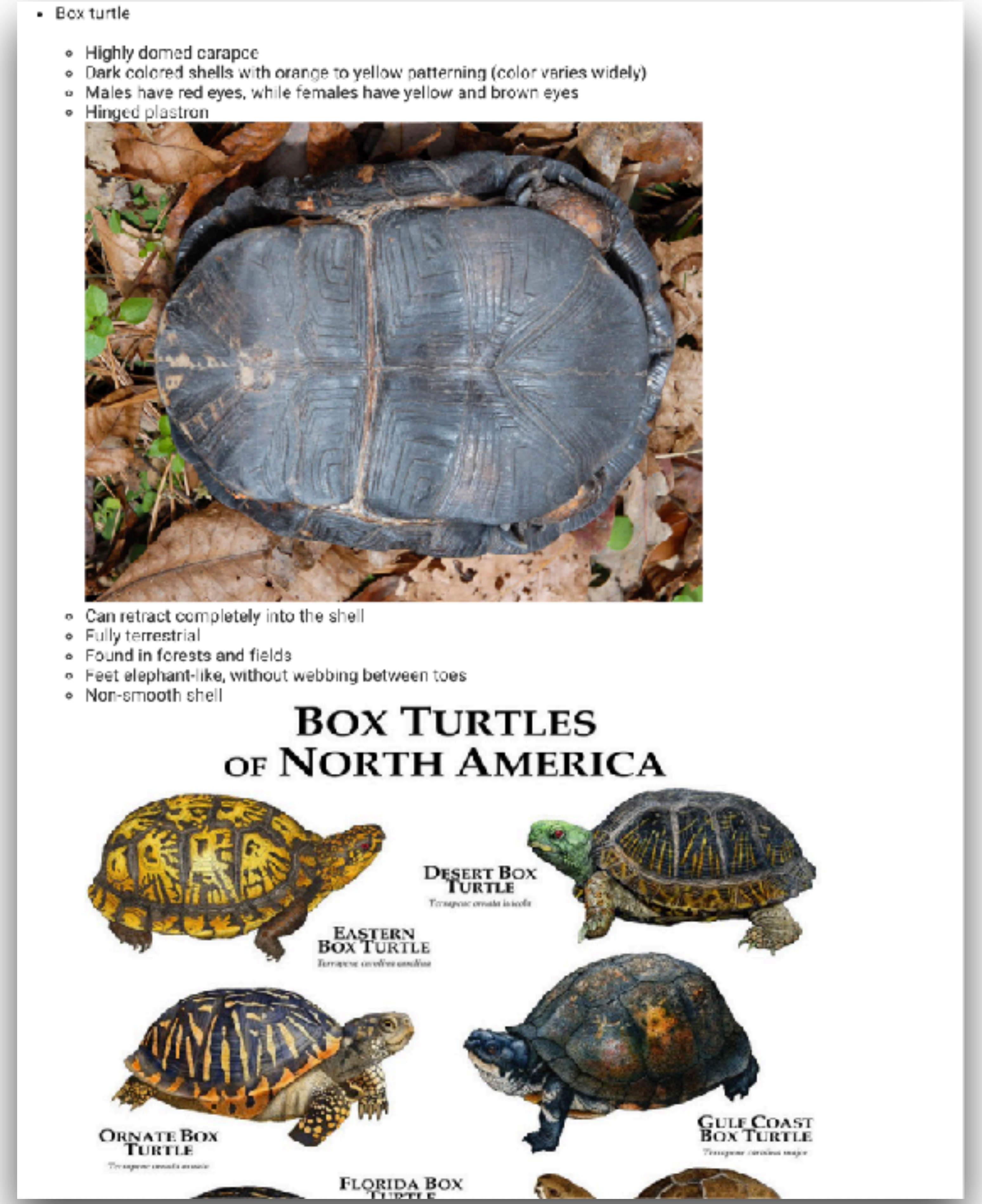


Potentially confused with Tiger shark and Hammerhead shark.
Hammerhead sharks are usually easy to identify based on their distinctive head.
The distinction with **Tiger shark** is more complicated.

Great white shark vs Tiger shark

- Points to compare
  - Stripes
    - Tiger sharks have vertical stripes
  - Thickness of the main body
    - Great White sharks are thicker
  - Head
    - Tiger sharks seem to have a more wedge-like / pointy shape
  - Fins on the underside
    - Tiger sharks have larger fins on the underside
  - Shape of tail
    - The top part of the tail (see the sketches below)
- Great White shark side sketch (from Wikipedia)
- Tiger shark side sketch (from Wikipedia). Note for instance the larger fins on the underside towards the tail end.
- More information on https://fishingbooker.com/blog/tiger-shark-vs-great-white-shark/

Sharks



- Box turtle
  - Highly domed carapce
  - Dark colored shells with orange to yellow patterning (color varies widely)
  - Males have red eyes, while females have yellow and brown eyes
  - Hinged plastron
  - Can retract completely into the shell
  - Fully terrestrial
  - Found in forests and fields
  - Feet elephant-like, without webbing between toes
  - Non-smooth shell

Turtles

Stingray vs. electric ray

- This is a hard class distinction.
- Some training images are incorrect.
- **Electric rays** tend to have a fin at the end of their tail, for instance (source biophysics.sbg.ac.at)

- The tails of **electric rays** also tend to be wider and shorter than those of a **stingray**.
- **Stingrays** look more like this (source unknown, via zazzle.com)

Stingrays