

CSE 599 Empirical Foundations of Machine Learning

University of Washington, Autumn 2021

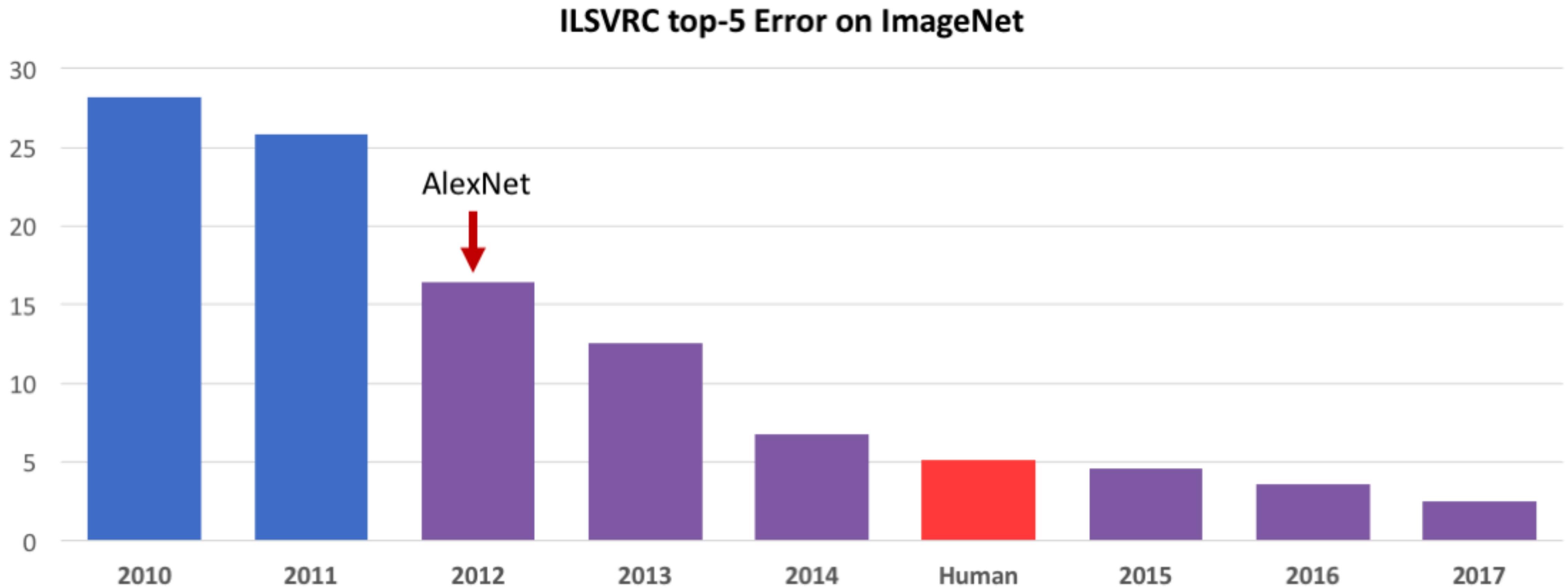
Goals for today

Give examples of research on the “other” (non-model) half of machine learning

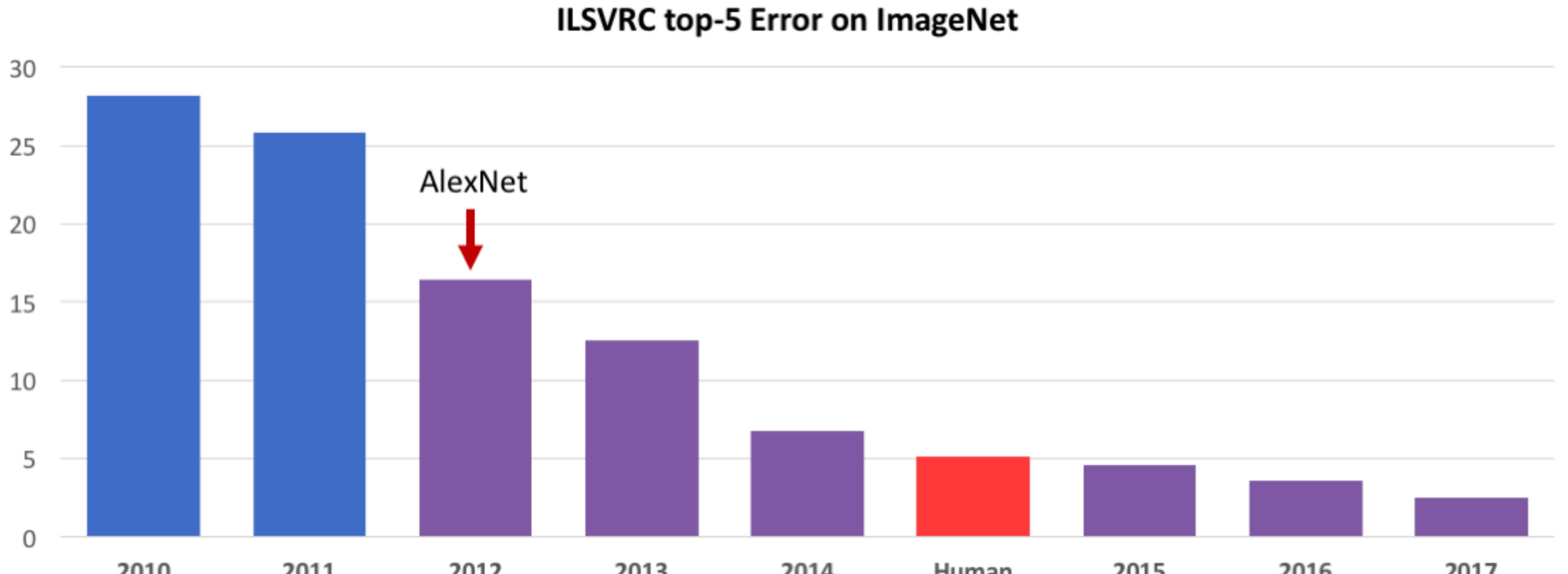
Inspiration for course projects

Course project timeline, expectations, etc.

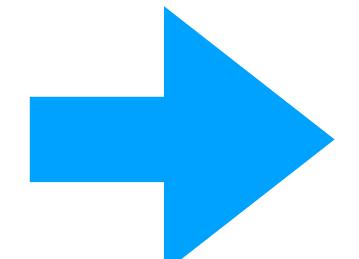
Recap from last lecture



Recap from last lecture



Large improvement, new methods



Tremendous impact on machine learning

Contrast to classical algorithms

Classical algorithms

Precisely defined general problems
(e.g., shortest path on directed graphs)

Algorithm is provably correct

Algorithm compared in terms of time complexity, space complexity, etc.

Empirical machine learning

Problems defined by specific datasets
(e.g., ImageNet, SQuAD, etc.)

Accuracy measured on a test set

Algorithm compared on specific benchmark results

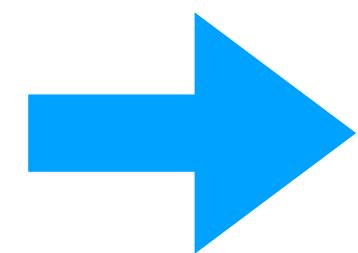
Contrast to classical algorithms

Classical algorithms

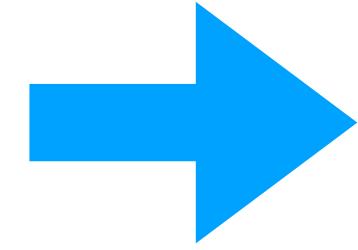
Precisely defined general problems
(e.g., shortest path on directed graphs)

Algorithm is provably correct

Algorithm compared in terms of time complexity, space complexity, etc.



Validity of empirical results is crucial



Need conceptual frameworks to organize datasets & benchmarks

Empirical machine learning

Problems defined by specific datasets
(e.g., ImageNet, SQuAD, etc.)

Accuracy measured on a test set

Algorithm compared on specific benchmark results

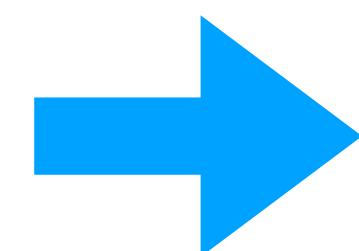
Contrast to classical algorithms

Classical algorithms

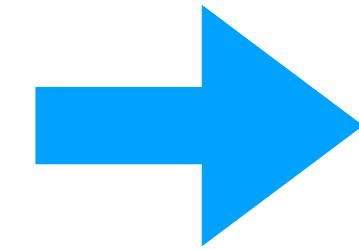
Precisely defined general problems
(e.g., shortest path on directed graphs)

Algorithm is provably correct

Algorithm compared in terms of time complexity, space complexity, etc.



Validity of empirical results is crucial



Need conceptual frameworks to organize datasets & benchmarks

(Note: learning theory offers provable guarantees, but not for specific algorithms such as the latest network architecture, or for specific datasets.)

Empirical machine learning

Problems defined by specific datasets
(e.g., ImageNet, SQuAD, etc.)

Accuracy measured on a test set

Algorithm compared on specific benchmark results

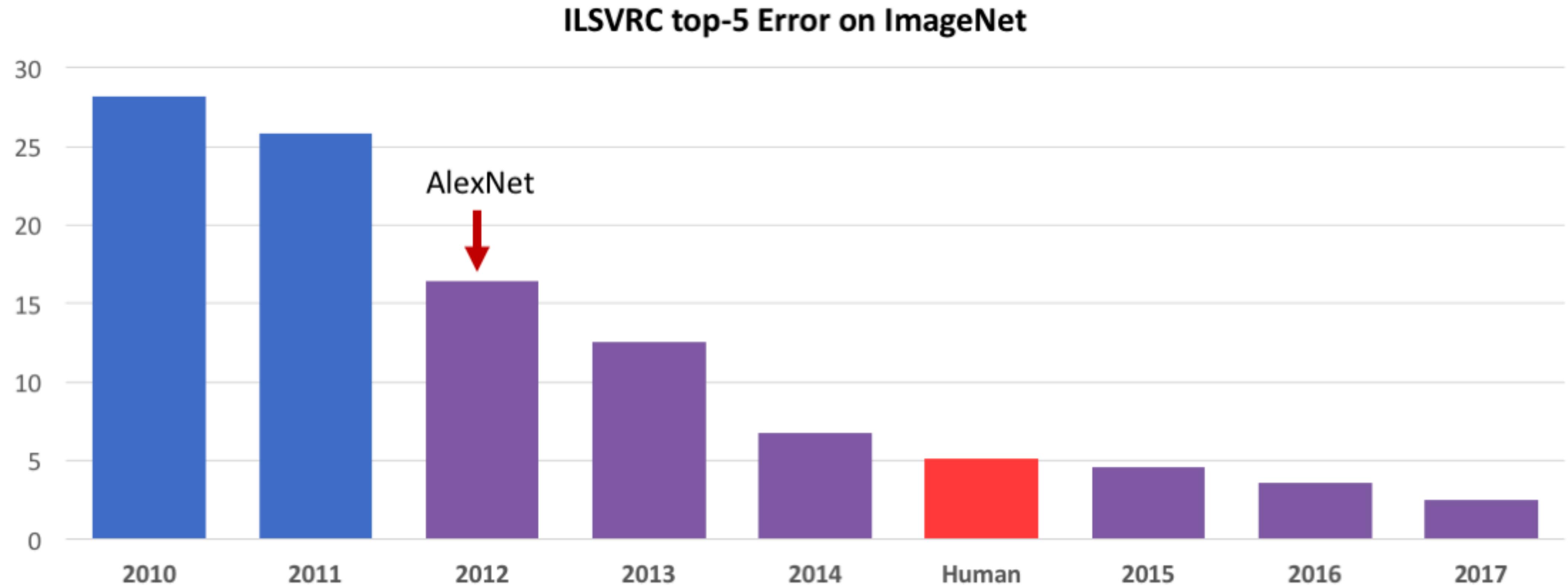
1. How reliable are ML benchmark results?
2. Do benchmark results transfer across learning problems?
3. Do benchmark results transfer across test distributions?

1. How reliable are ML benchmark results? (Internal validity)
2. Do benchmark results transfer across learning problems? (External validity)
3. Do benchmark results transfer across test distributions? (External validity)

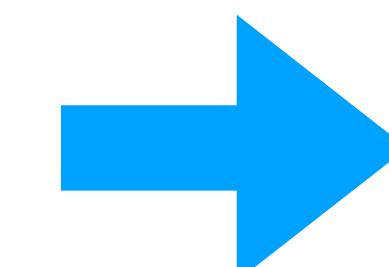
1. How reliable are ML benchmark results? (Internal validity)
2. Do benchmark results transfer across learning problems? (External validity)
3. Do benchmark results transfer across test distributions? (External validity)
4. Course projects

1. How reliable are ML benchmark results? (Internal validity)
2. Do benchmark results transfer across learning problems? (External validity)
3. Do benchmark results transfer across test distributions? (External validity)
4. Course projects

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**.

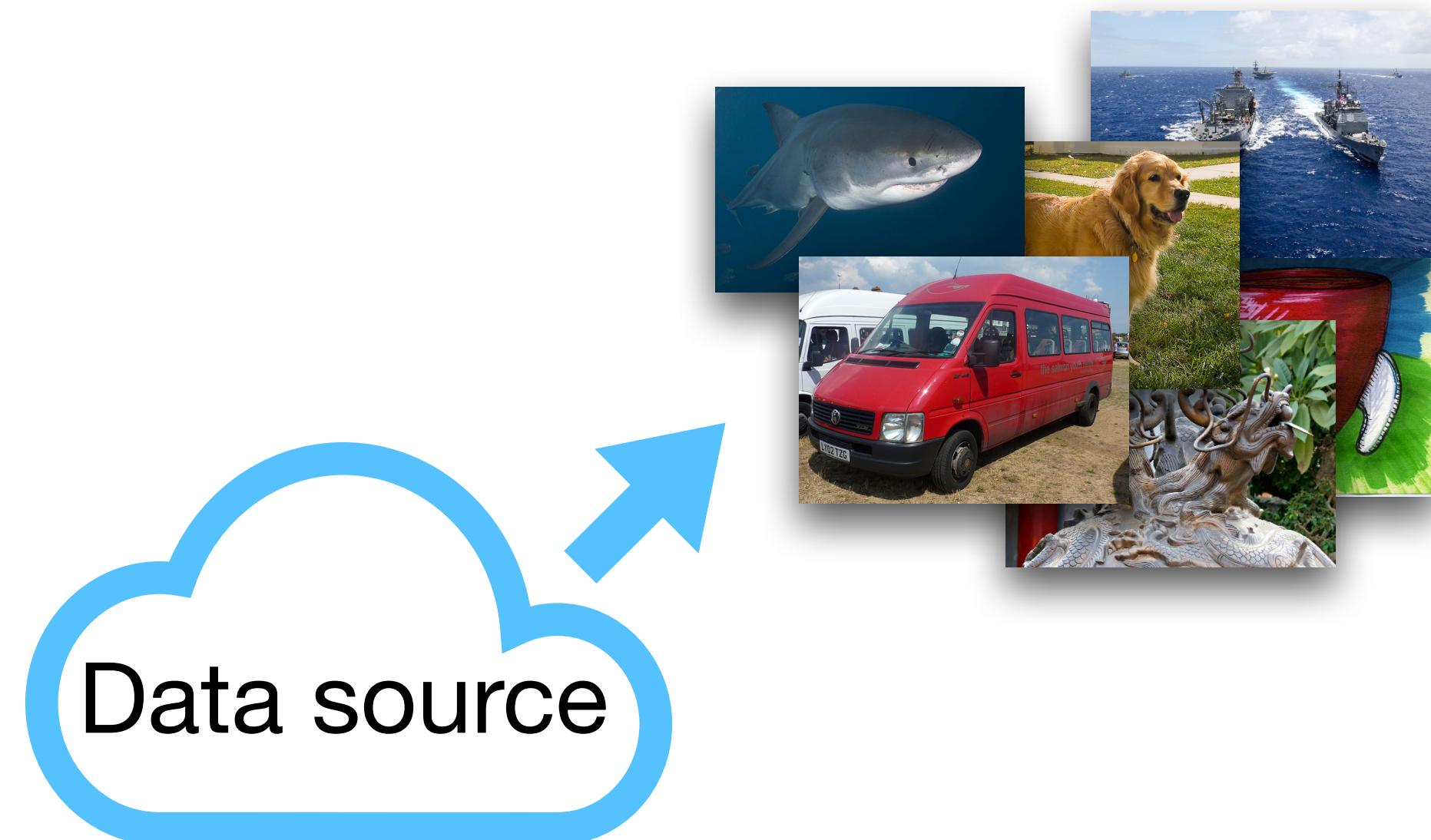
Generalization

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



Generalization

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



Generalization

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



Generalization

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



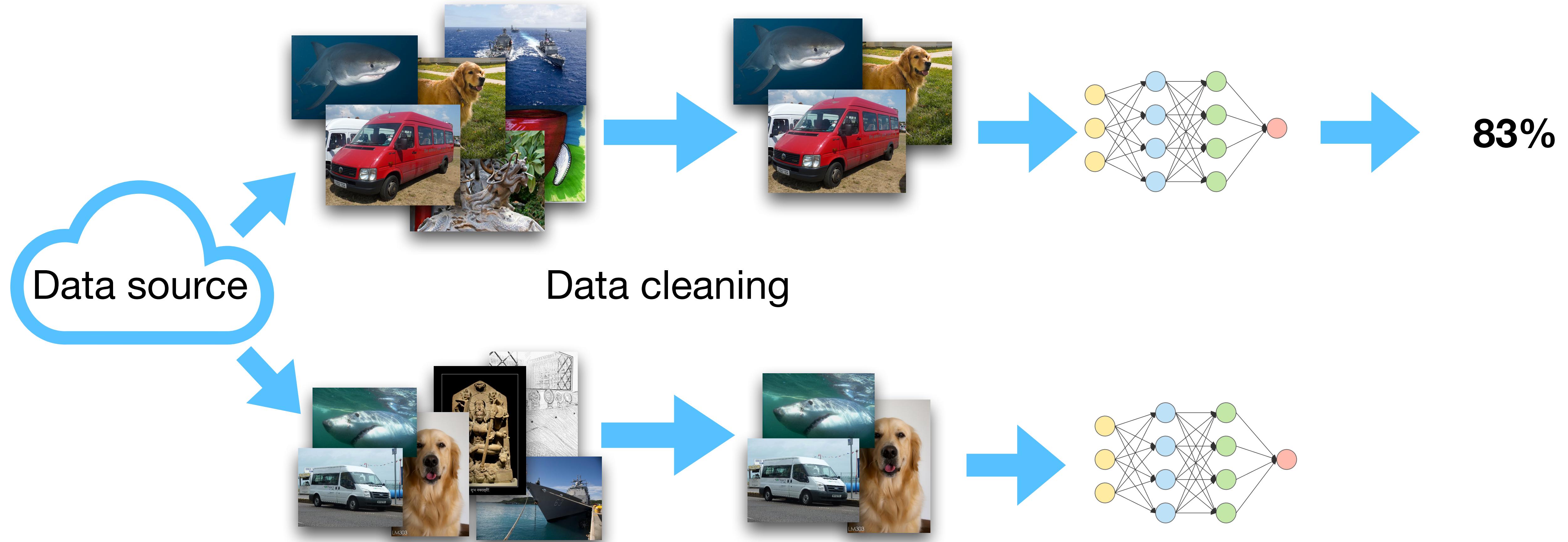
Generalization

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



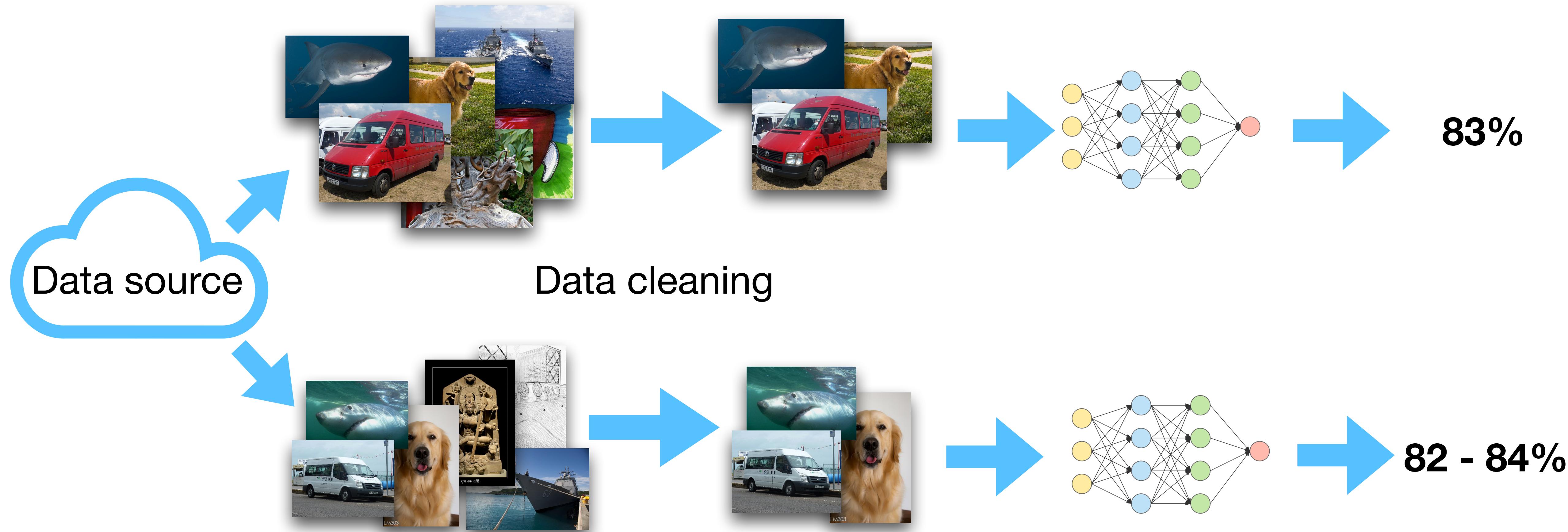
Generalization

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



Generalization

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



Generalization, more formally

How can we reliably measure generalization?

Ideal ML workflow: holdout method

1. Collect data

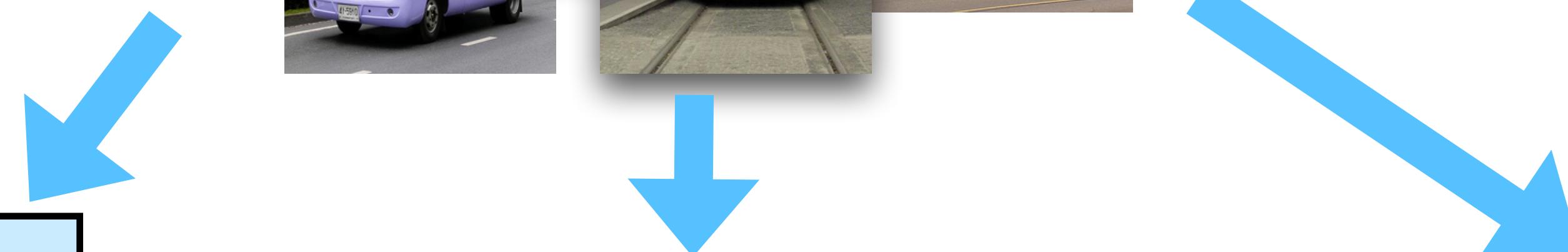
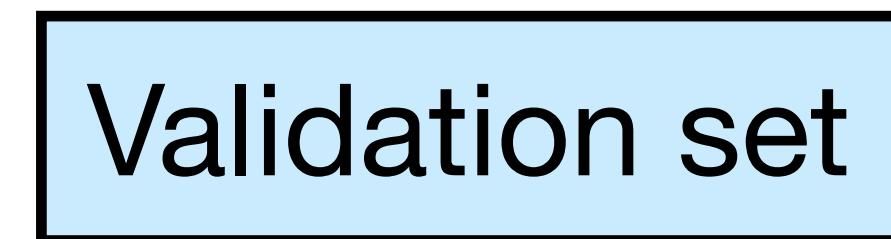
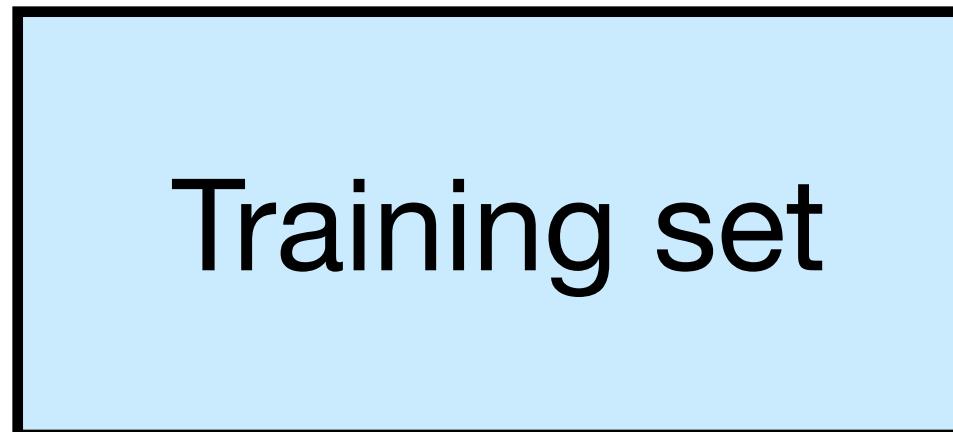


Ideal ML workflow: holdout method

1. Collect data



2. Split data

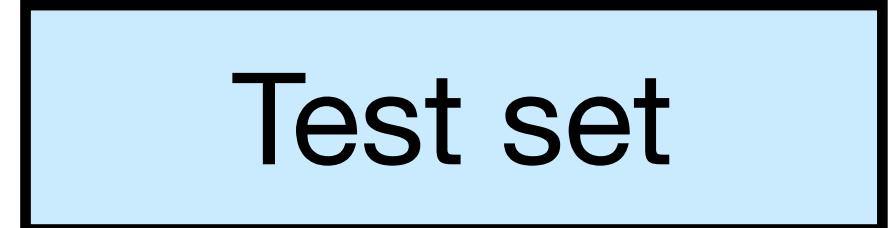
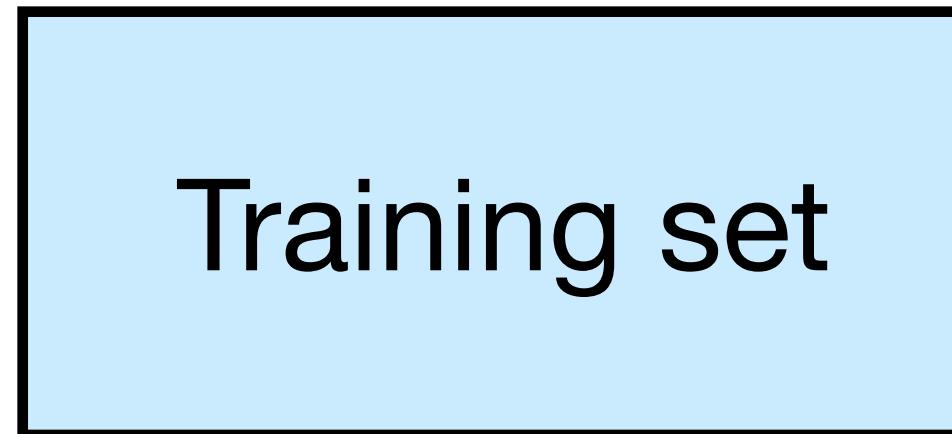


Ideal ML workflow: holdout method

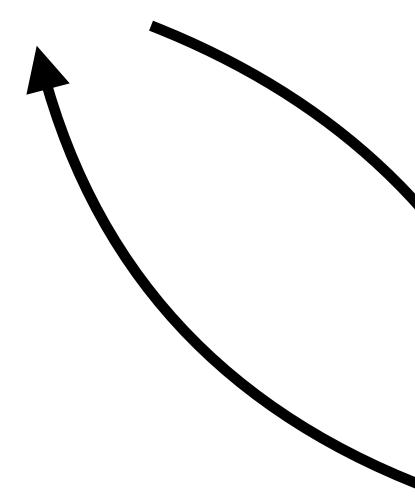
1. Collect data



2. Split data



3. Train and
tune model

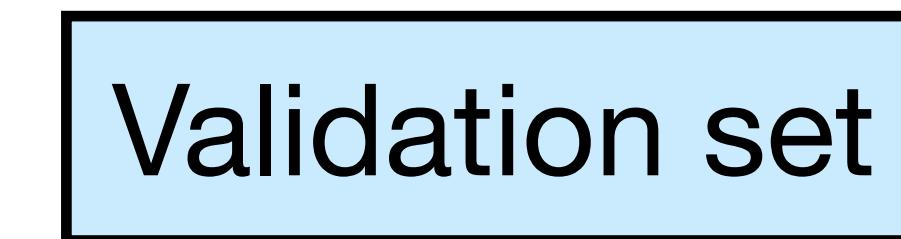
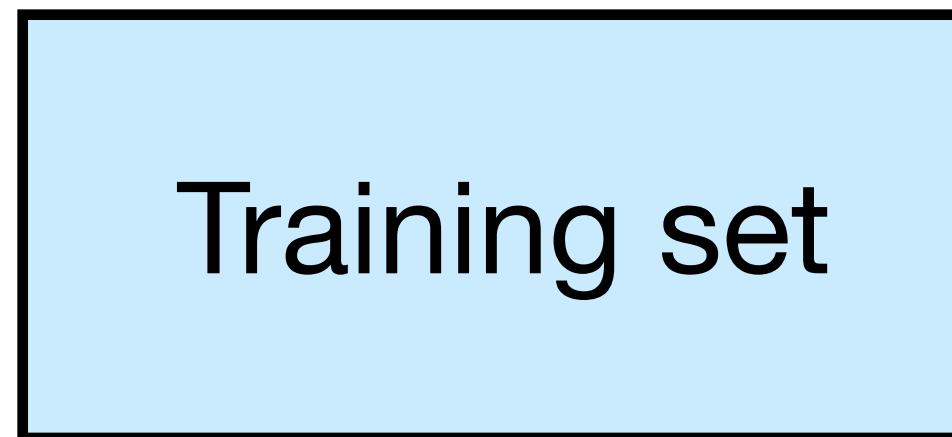


Ideal ML workflow: holdout method

1. Collect data



2. Split data



3. Train and
tune model

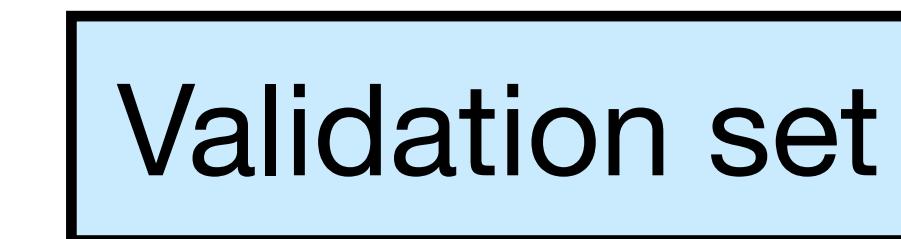
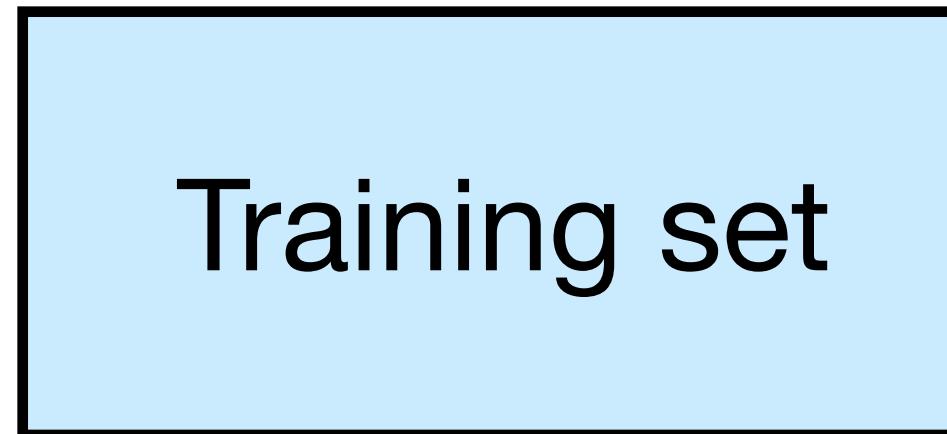


Ideal ML workflow: holdout method

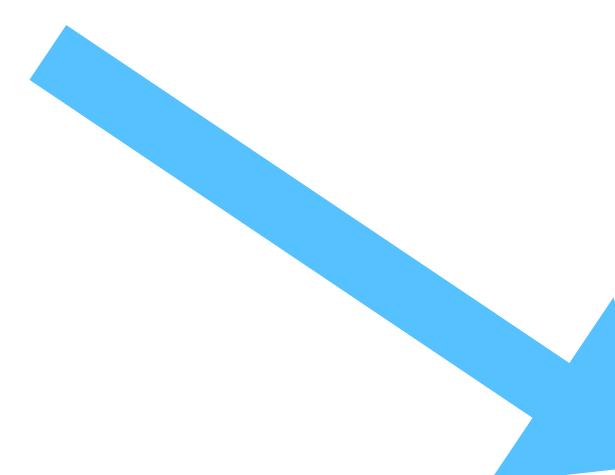
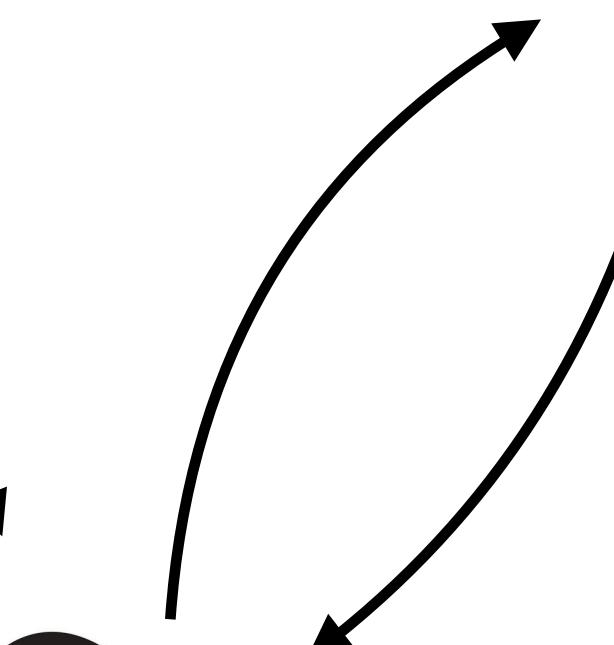
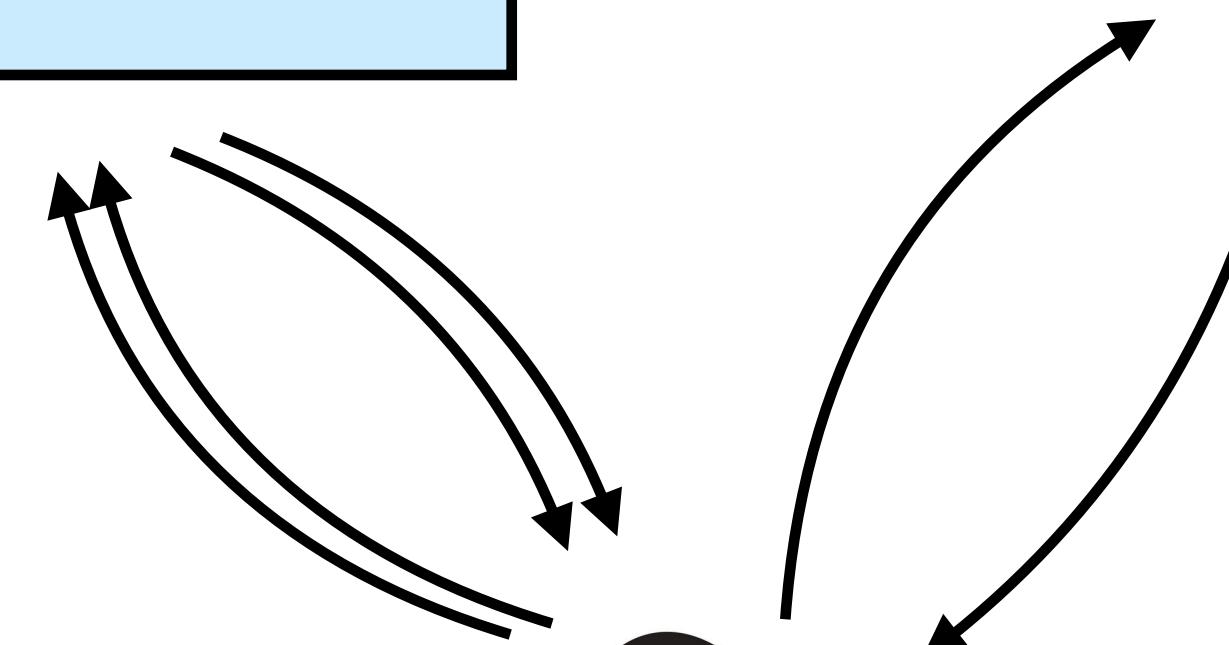
1. Collect data



2. Split data



3. Train and
tune model

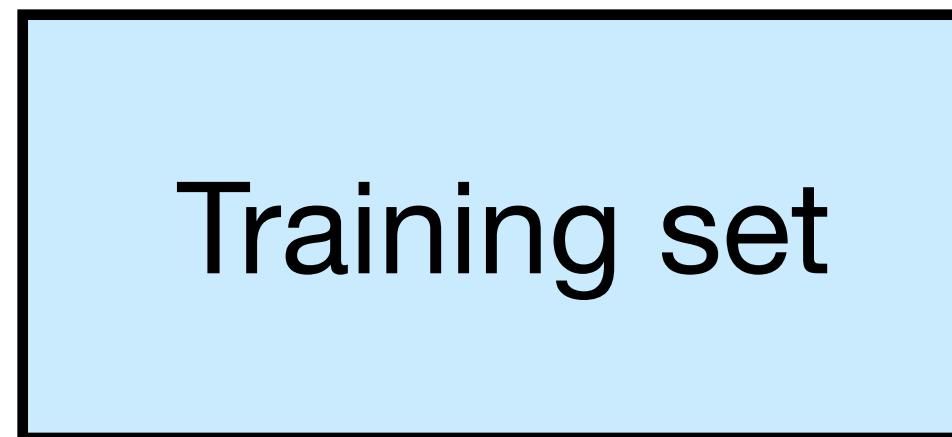


Ideal ML workflow: holdout method

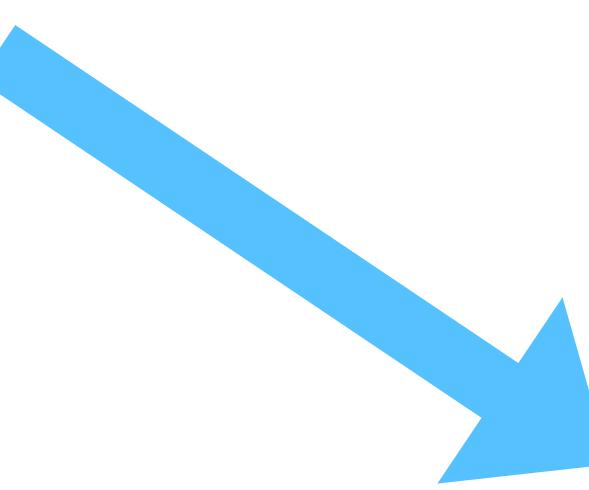
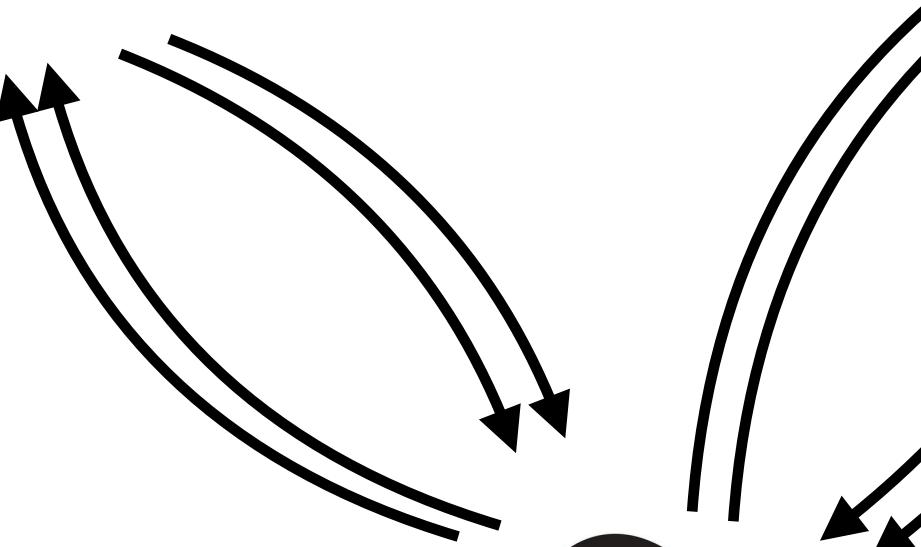
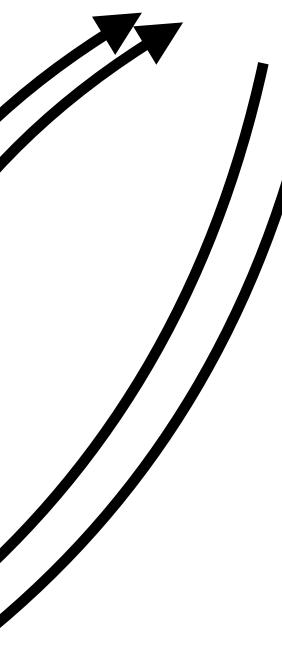
1. Collect data



2. Split data



3. Train and
tune model

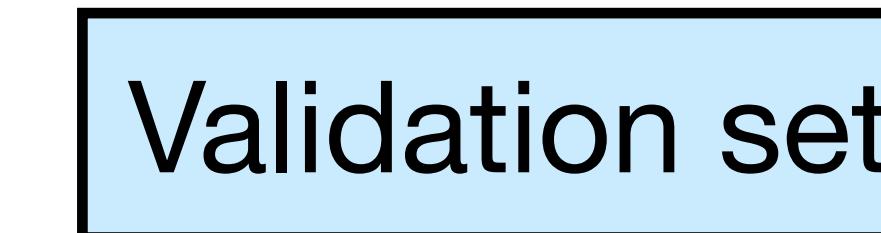
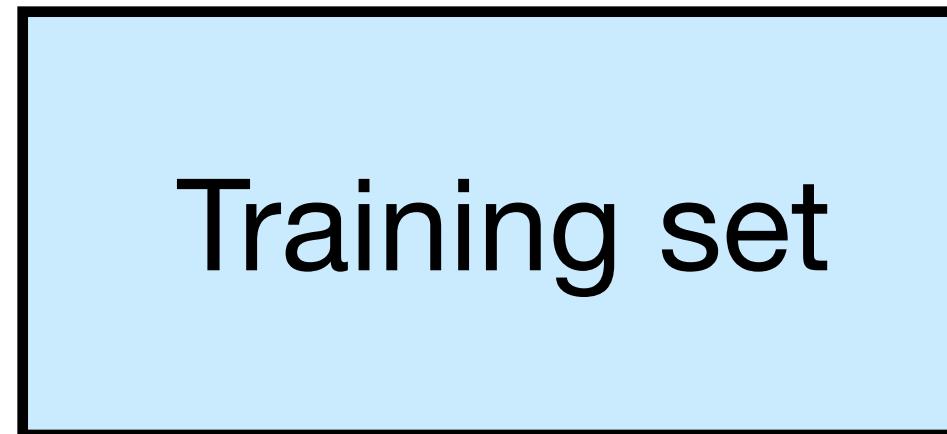


Ideal ML workflow: holdout method

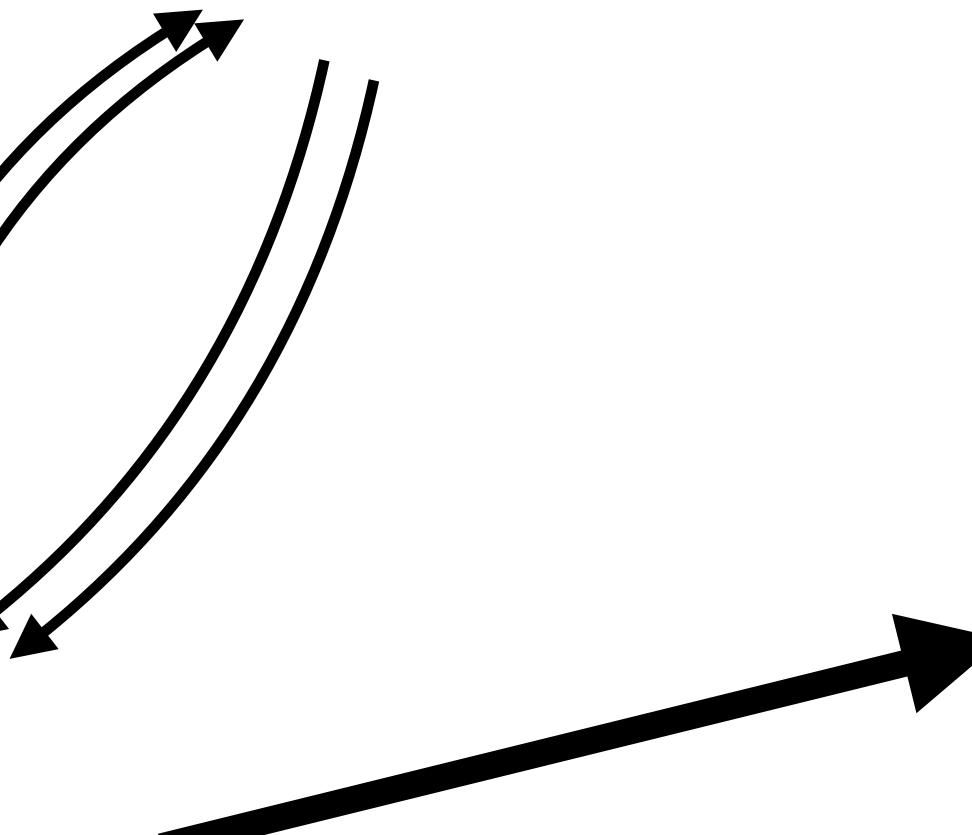
1. Collect data



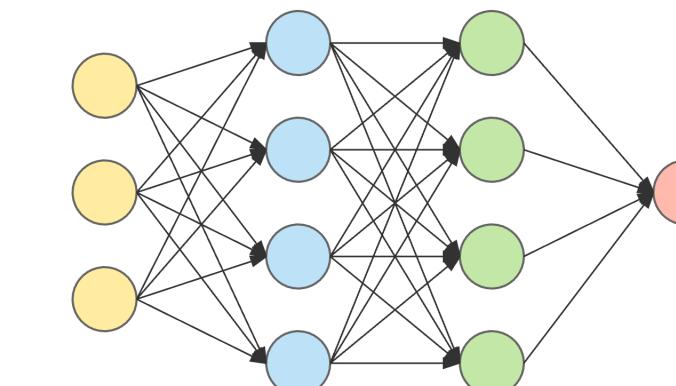
2. Split data



3. Train and
tune model



4. Compute final test accuracy

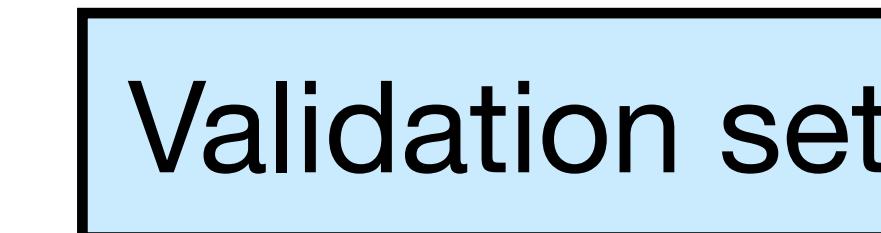
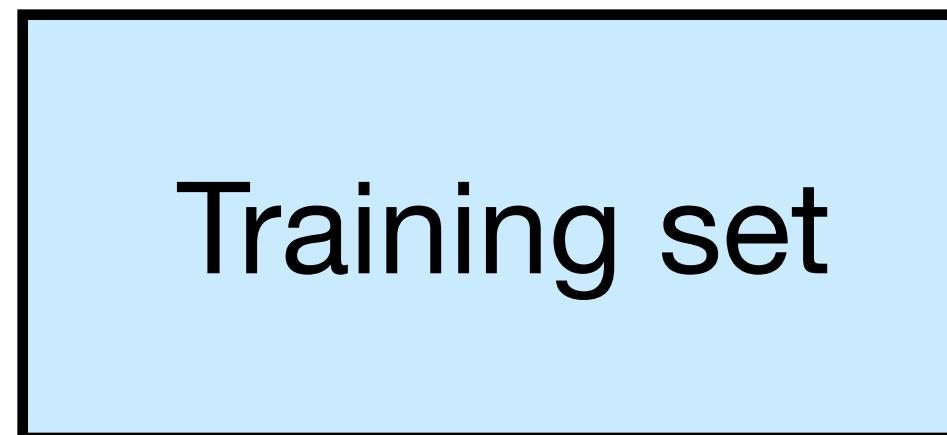


Ideal ML workflow: holdout method

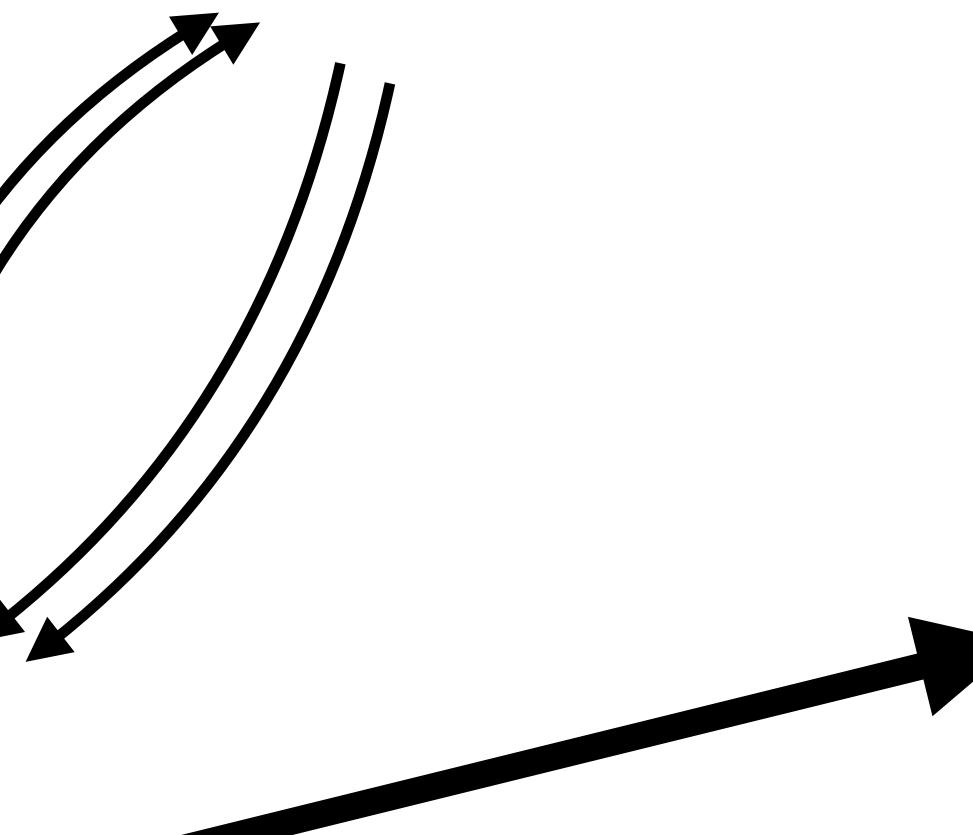
1. Collect data



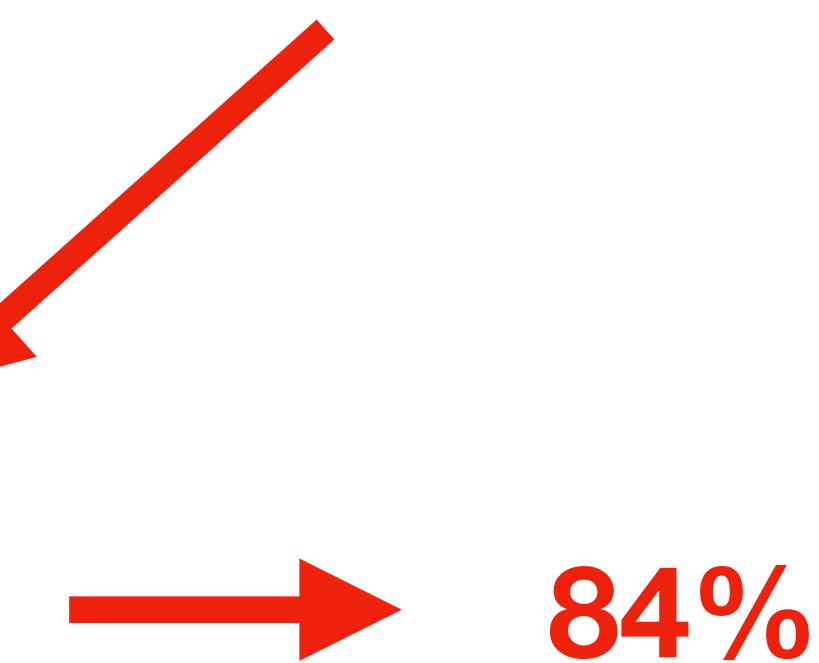
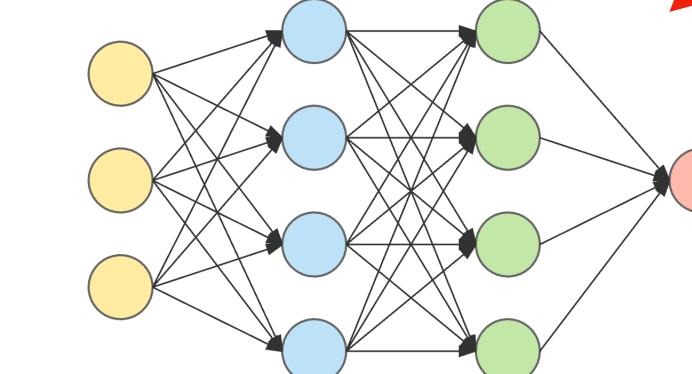
2. Split data



3. Train and
tune model

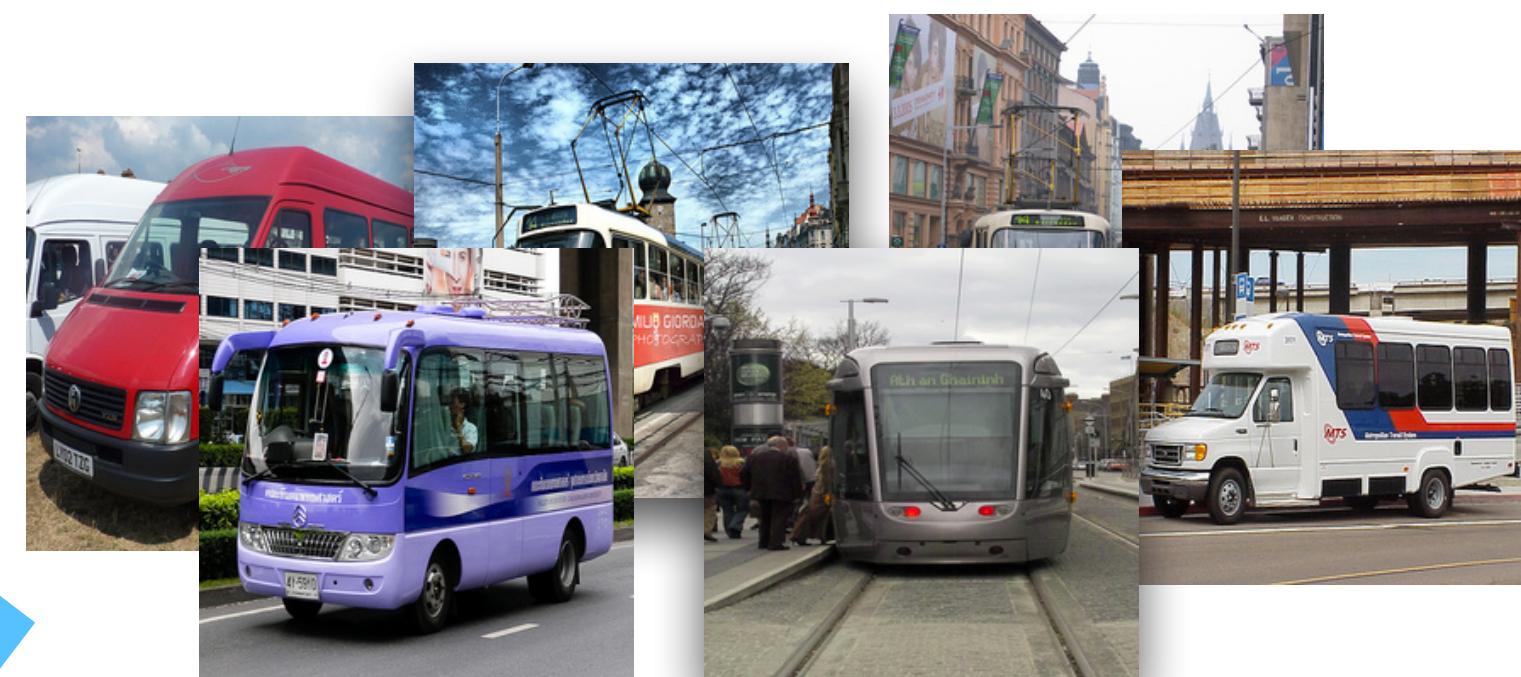


4. Compute final test accuracy

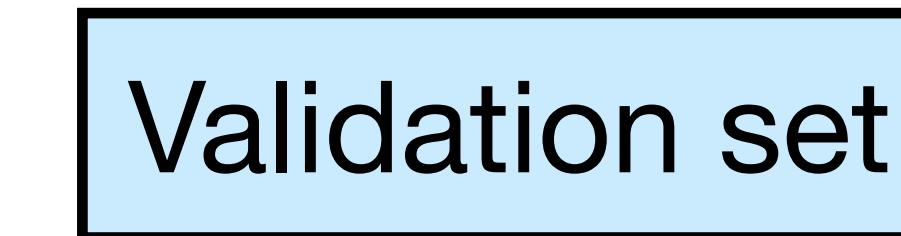
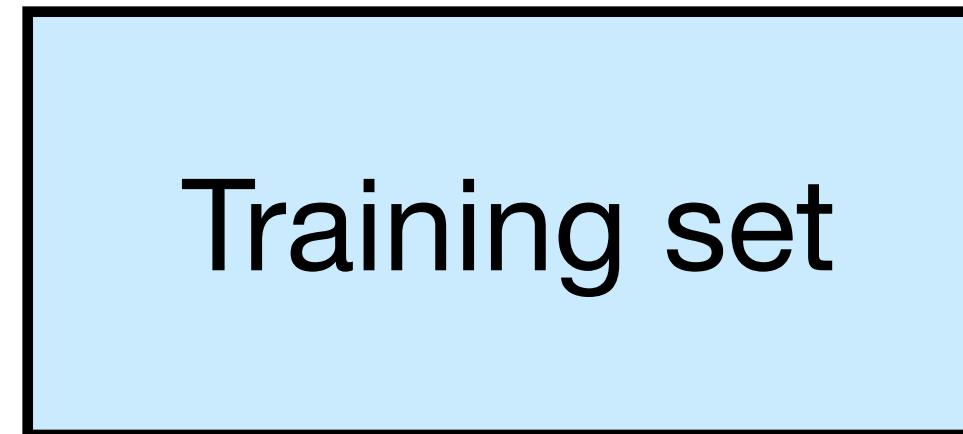


Ideal ML workflow: holdout method

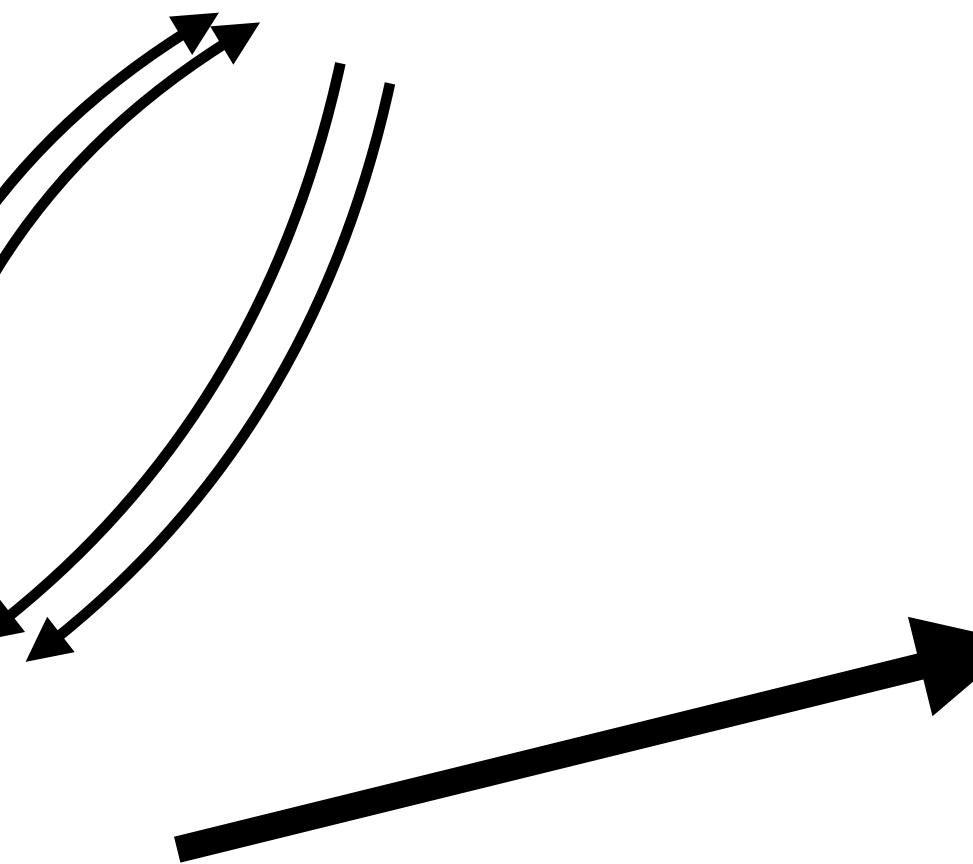
1. Collect data



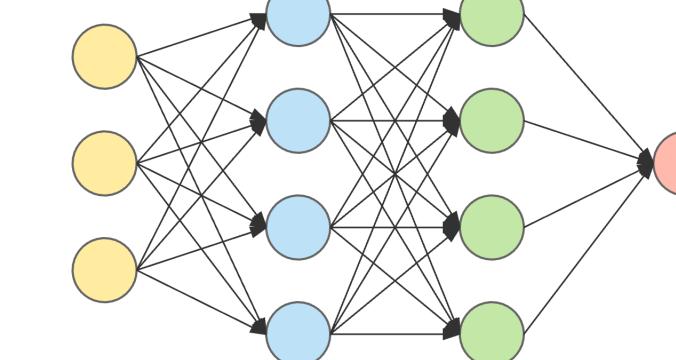
2. Split data



3. Train and tune model



4. Compute final test accuracy

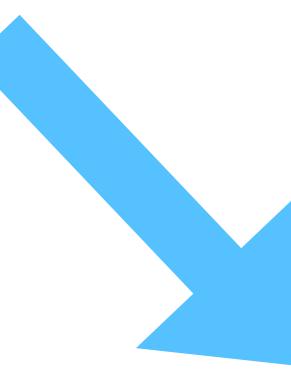
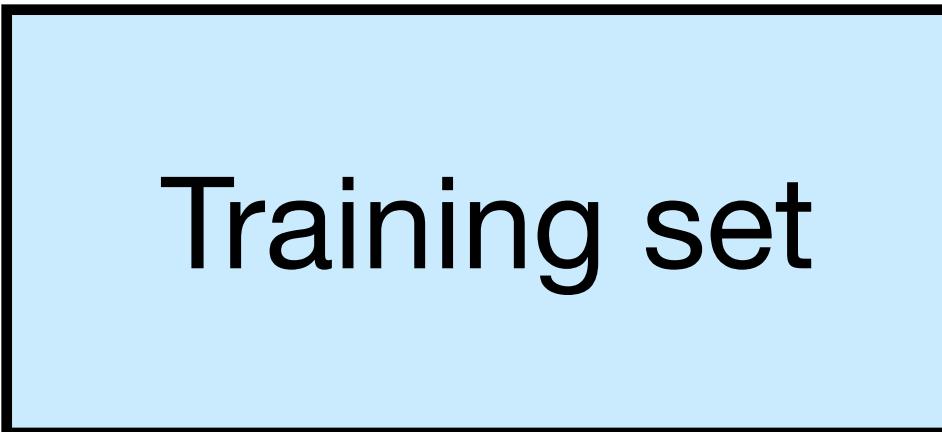
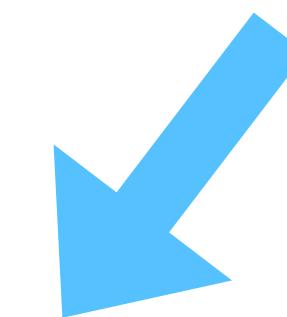


→ 84%



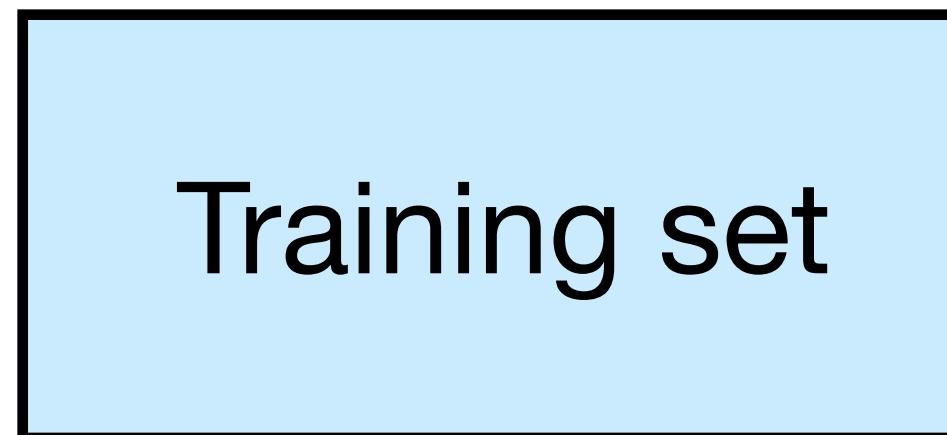
Typical ML workflow: hold-out re-use

1. Download data
(fixed split)

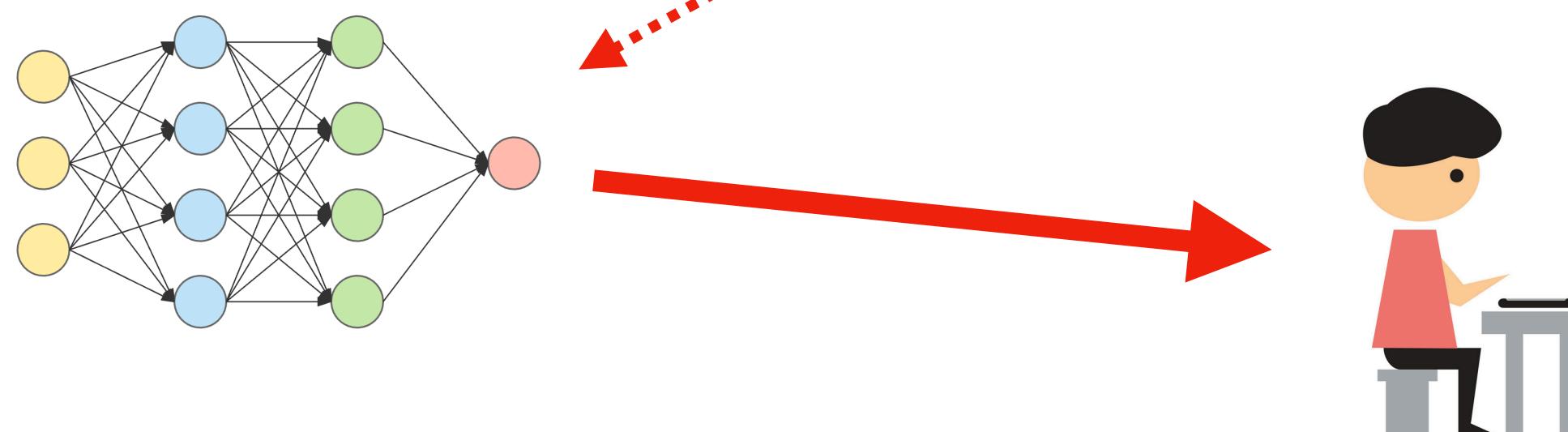


Typical ML workflow: hold-out re-use

1. Download data
(fixed split)

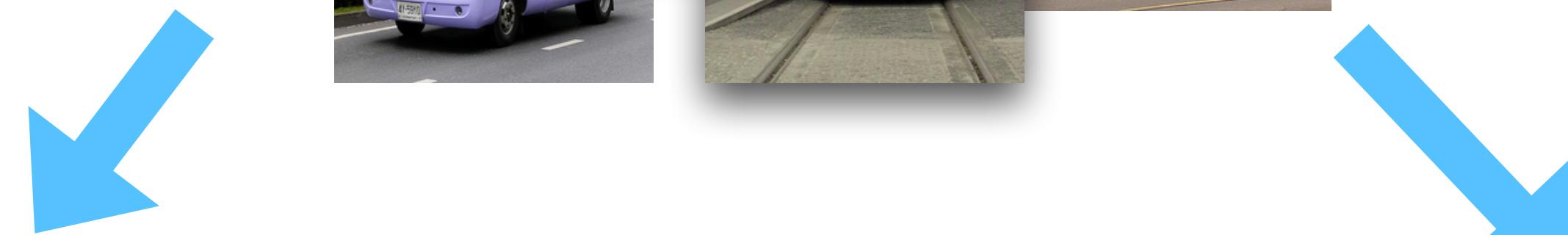


2. Download model



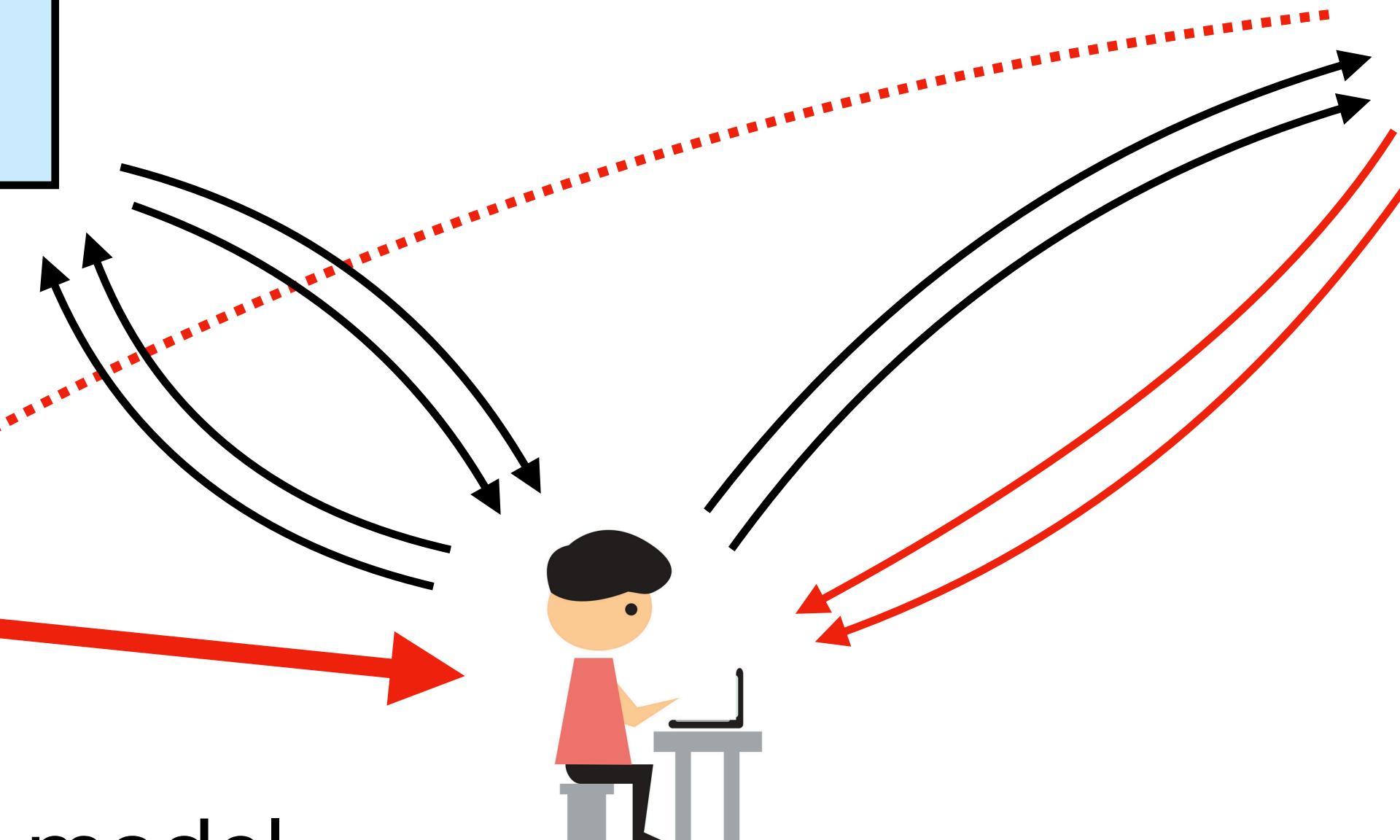
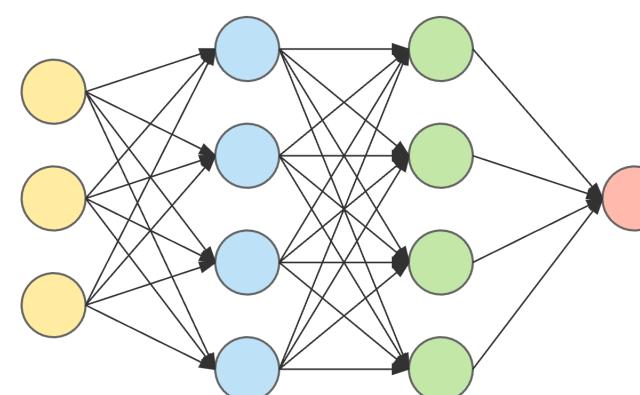
Typical ML workflow: hold-out re-use

1. Download data
(fixed split)



Training set

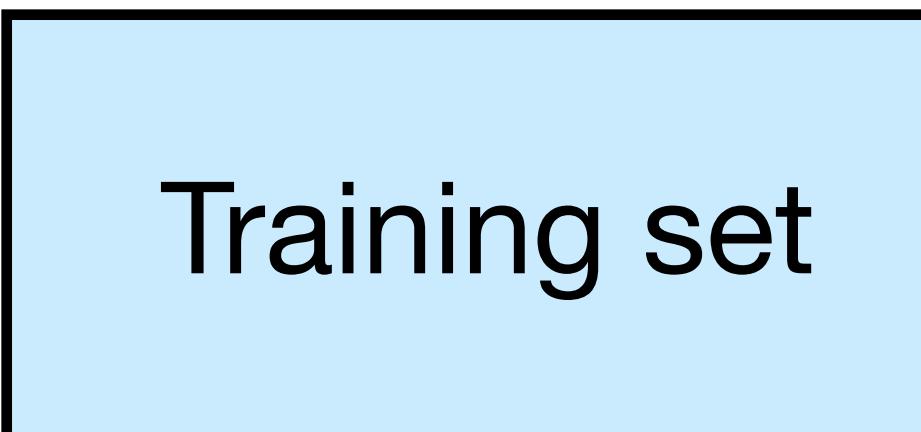
2. Download model



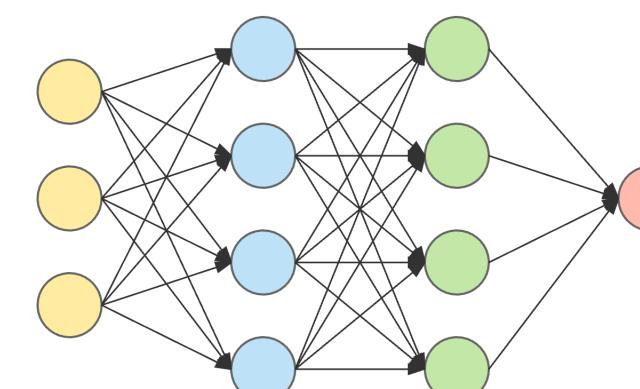
3. Train and tune model

Typical ML workflow: hold-out re-use

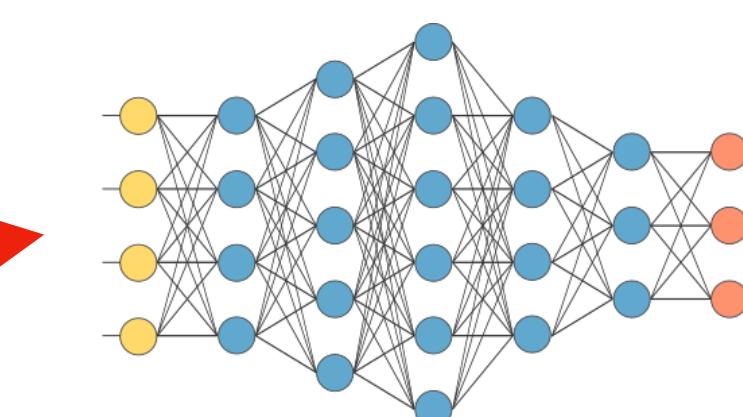
1. Download data
(fixed split)



2. Download model



3. Train and tune model

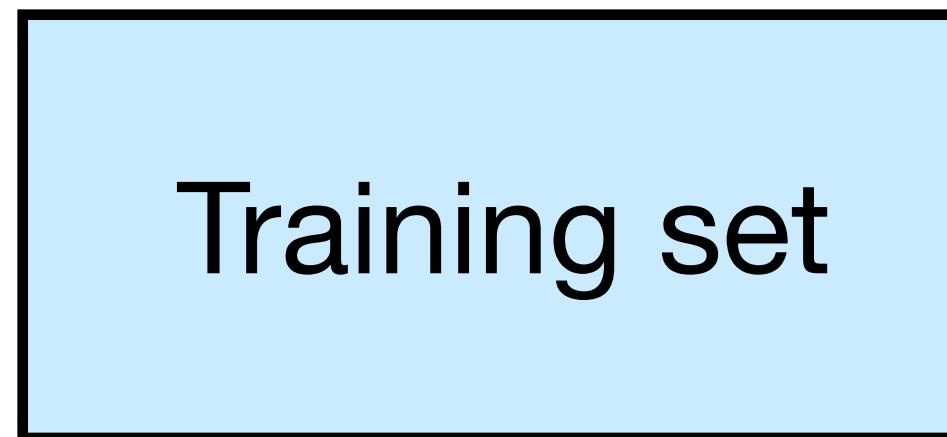


90%

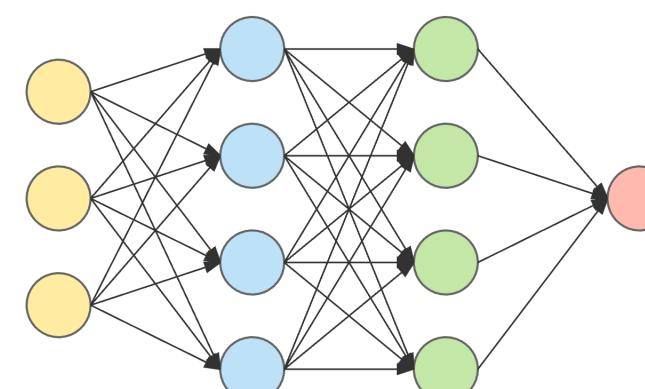
4. Compute final test accuracy

Typical ML workflow: hold-out re-use

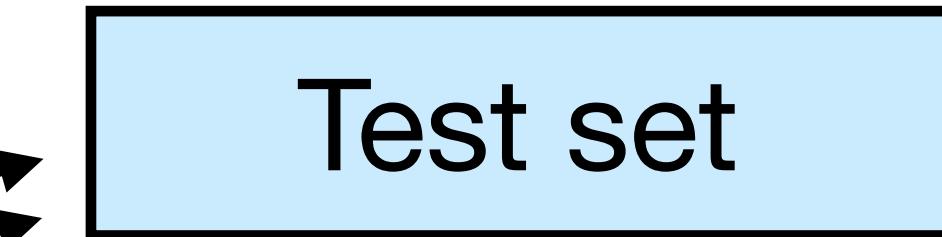
1. Download data
(fixed split)



2. Download model



3. Train and tune model



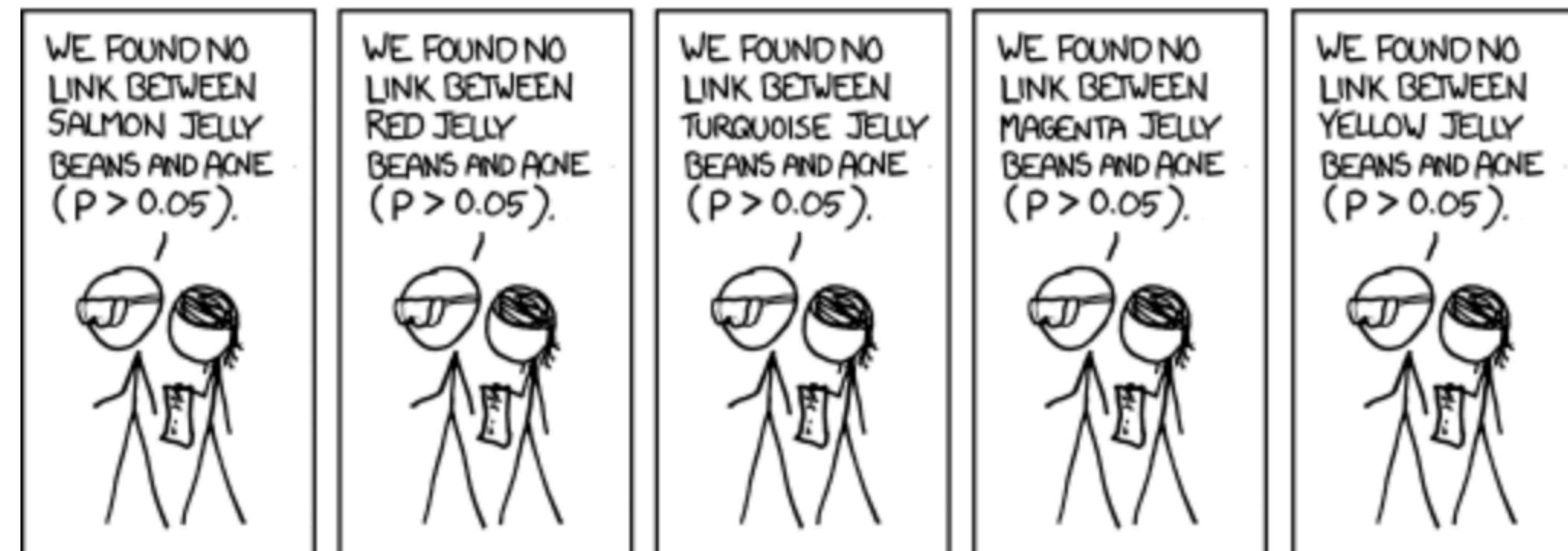
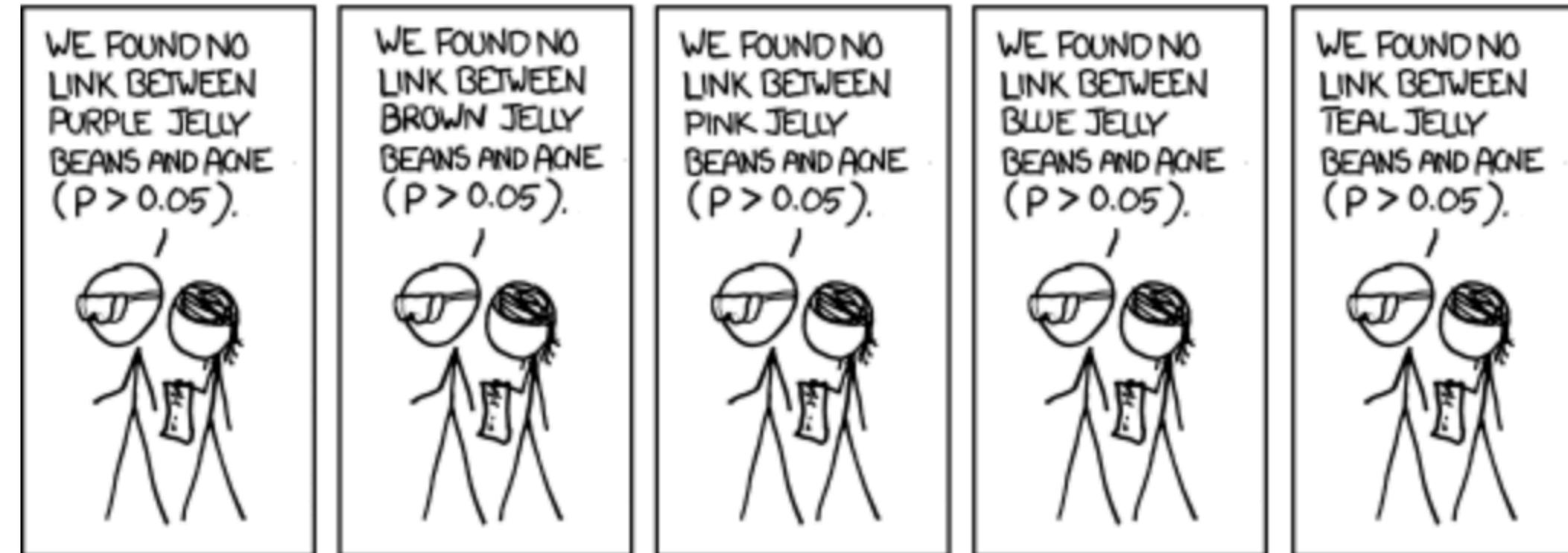
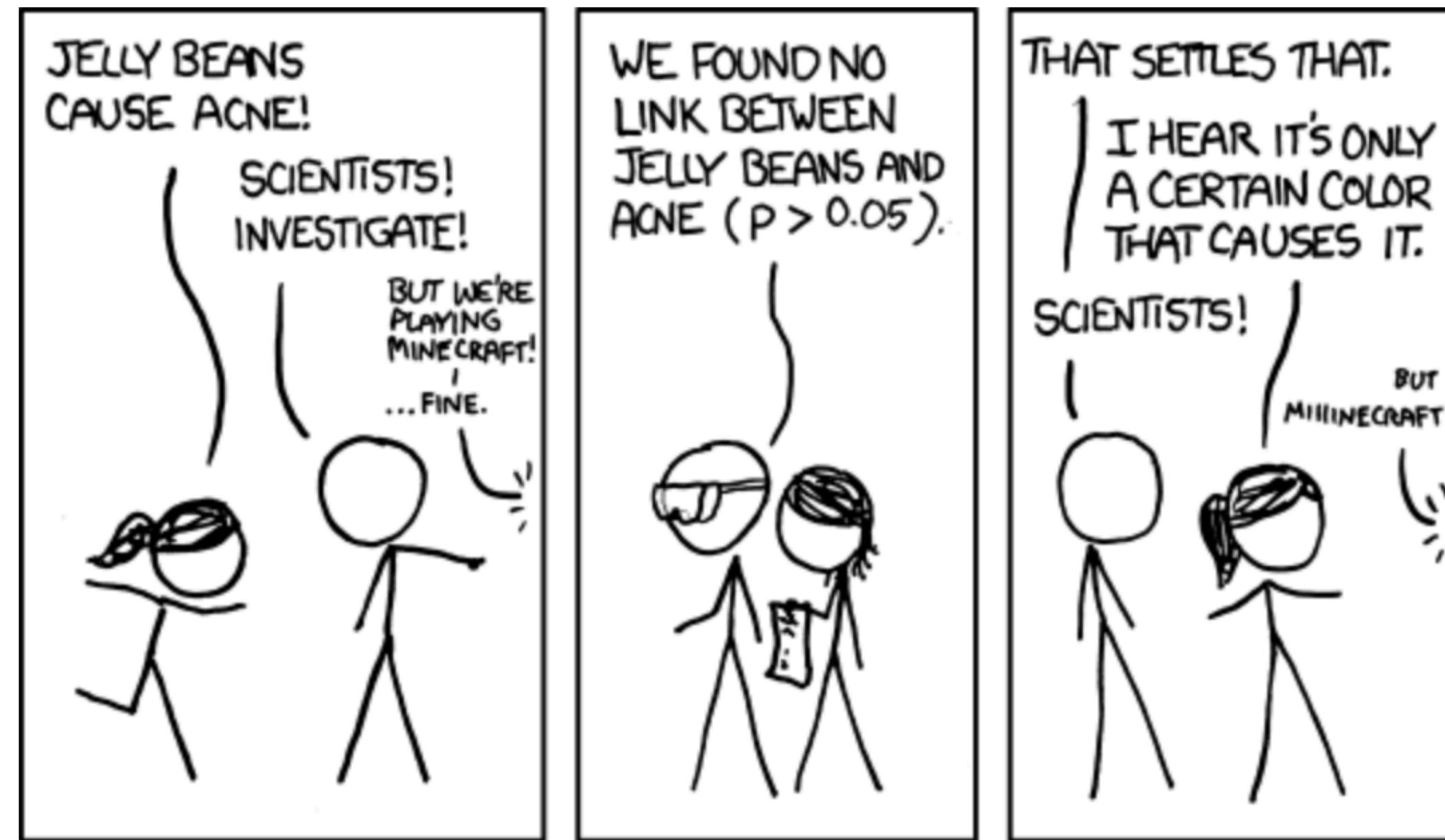
4. Compute final test accuracy

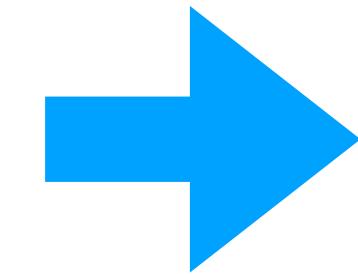
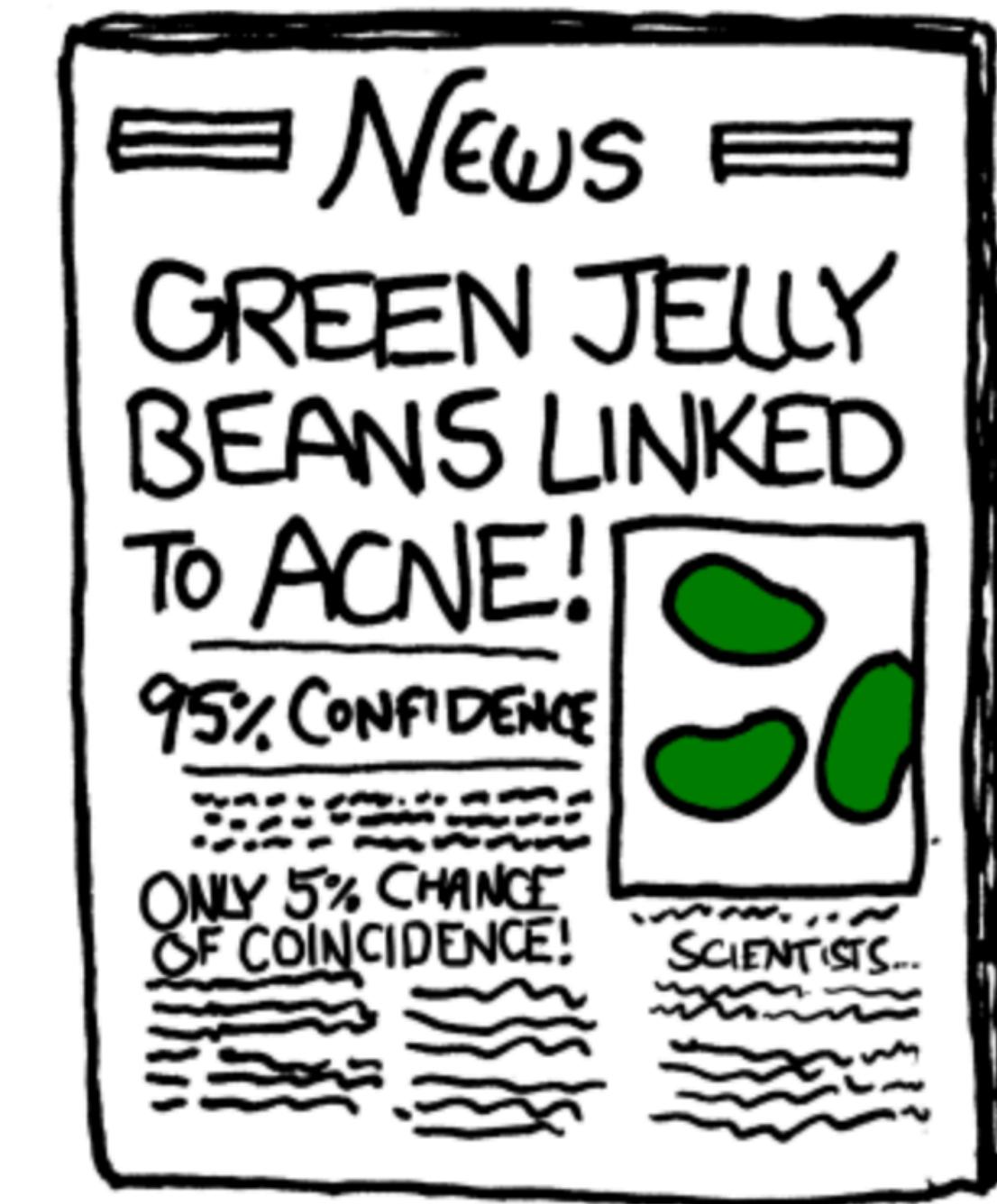
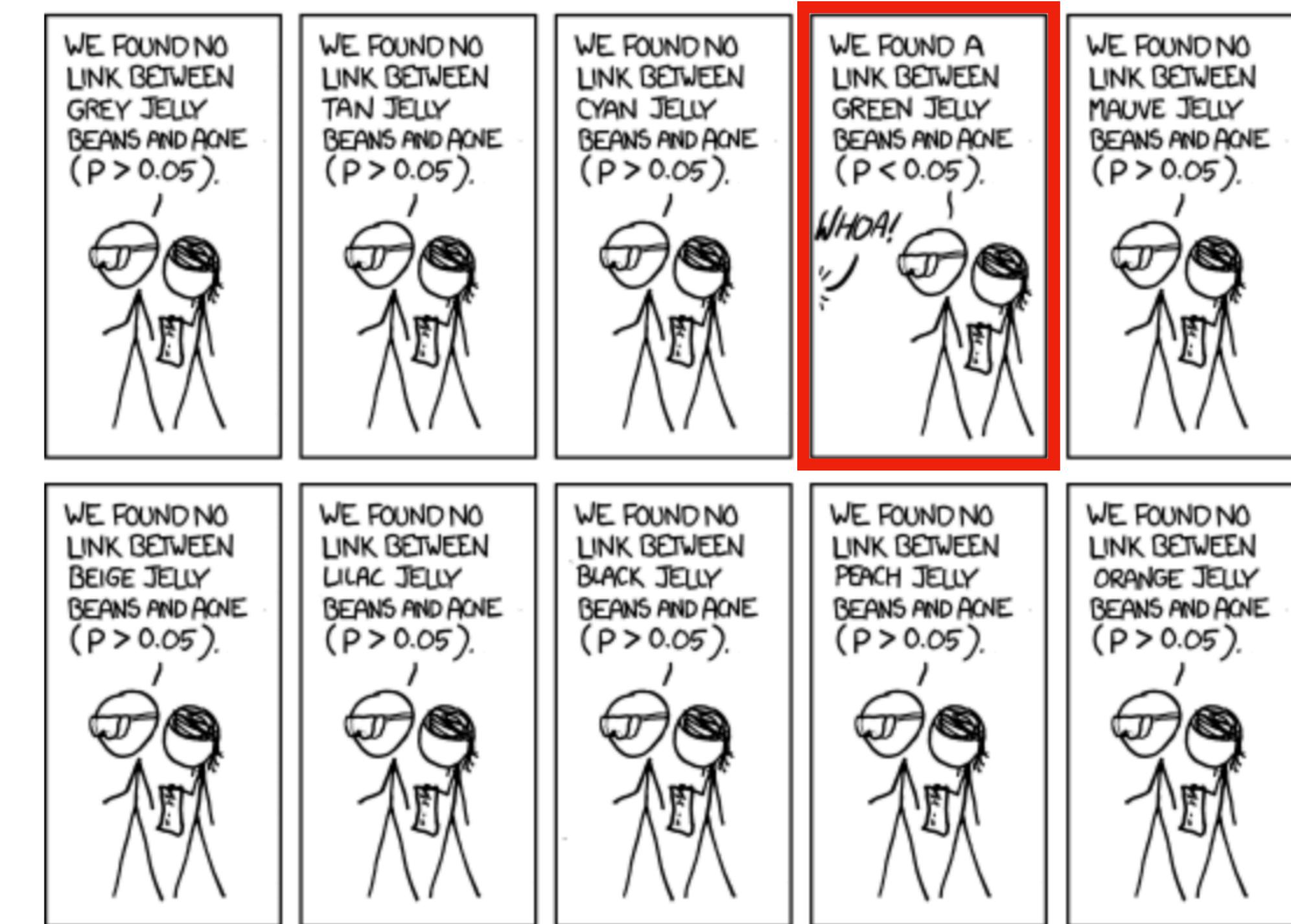
?

90%

Multiple hypothesis testing

Each model we test on the “hold-out” is a hypothesis.





If you test multiple hypotheses without corrections, statistical guarantees can become meaningless.

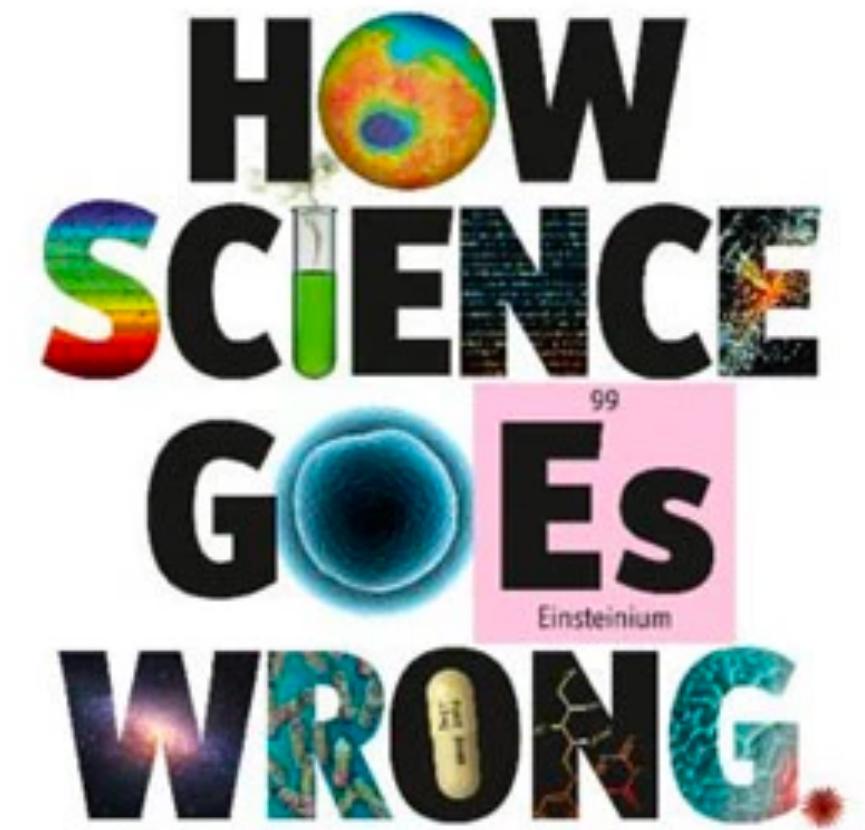
Replication Crisis in the Sciences

The screenshot shows a news article from the 'nature' journal. The header includes the 'nature' logo and 'International weekly journal of science'. Below the header is a navigation bar with links to Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and For Advertisers. A breadcrumb navigation shows the path: News & Comment > News > 2019 > May > Article. The main title of the article is 'Over half of psychology studies fail reproducibility test'. Below the title is a sub-headline: 'Largest replication study to date casts doubt on many published positive results.' The author is listed as 'Monya Baker' and the date is '27 August 2015'.

The screenshot shows a research article from 'Science'. The header includes the 'Science' logo and navigation links for Contents, News, Careers, and Journals. Below the header is a 'SHARE' section with icons for Facebook, Twitter, LinkedIn, and Email. The article title is 'Estimating the reproducibility of psychological science' by the Open Science Collaboration. It includes a note: '+ See all authors and affiliations'. Below the title is the publication information: 'Science 28 Aug 2015; Vol. 349, Issue 6251, aac4716 DOI: 10.1126/science.aac4716'. At the bottom of the article section are links for Article, Figures & Data, Info & Metrics, eLetters, and PDF.



The screenshot shows a news article titled 'Psychology's Replication Crisis Is Running Out of Excuses' by Ed Yong. The article discusses a study finding that only half of psychology studies can be repeated. The author's name is Ed Yong and the date is November 19, 2018.



Replication Crisis in the Sciences

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue

News & Comment > News > 2019 > May > Article

NATURE | NEWS

Over half of psychology studies test

Largest replication study to date casts doubt on many psychological findings

Monya Baker

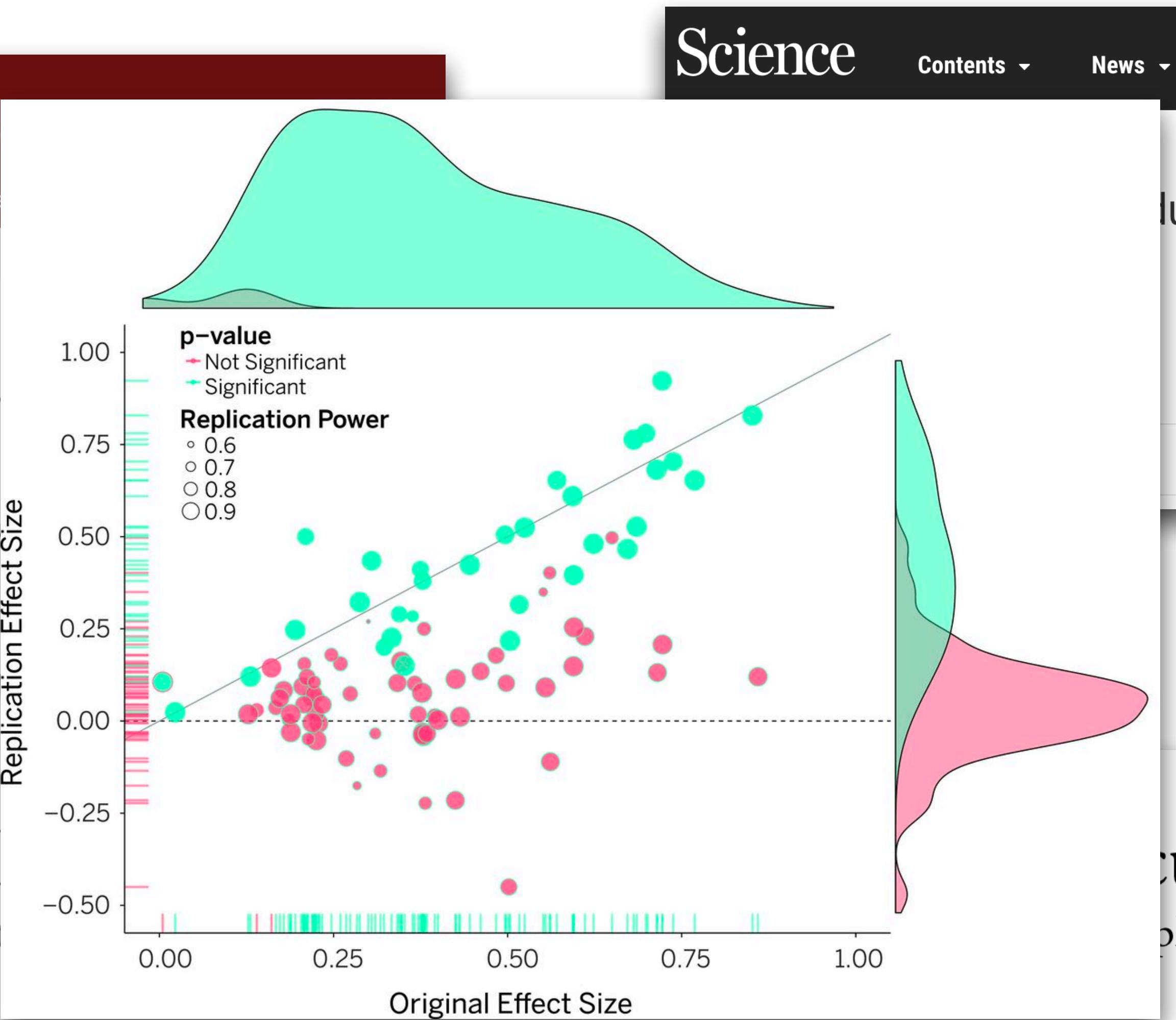
27 August 2015

SCIENCE

Psychology's Replication Crisis

Another big project has found that many findings fall flat.

ED YONG NOVEMBER 19, 2018



Science

Contents ▾ News ▾ Careers ▾ Journals ▾

lucibility of psychological science

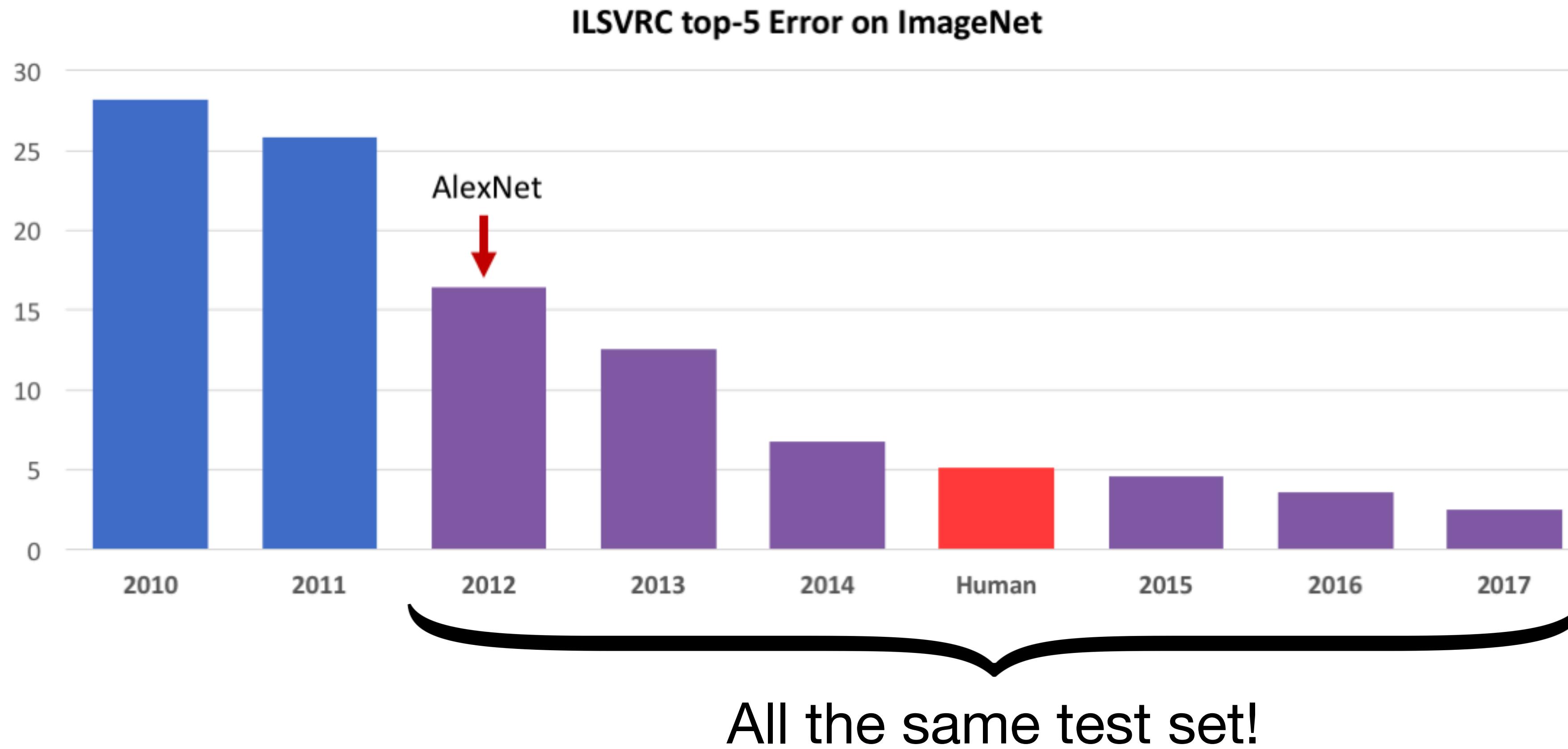
Info & Metrics eLetters PDF

The Economist OCTOBER 19TH-25TH 2013

Britain's angry white men
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

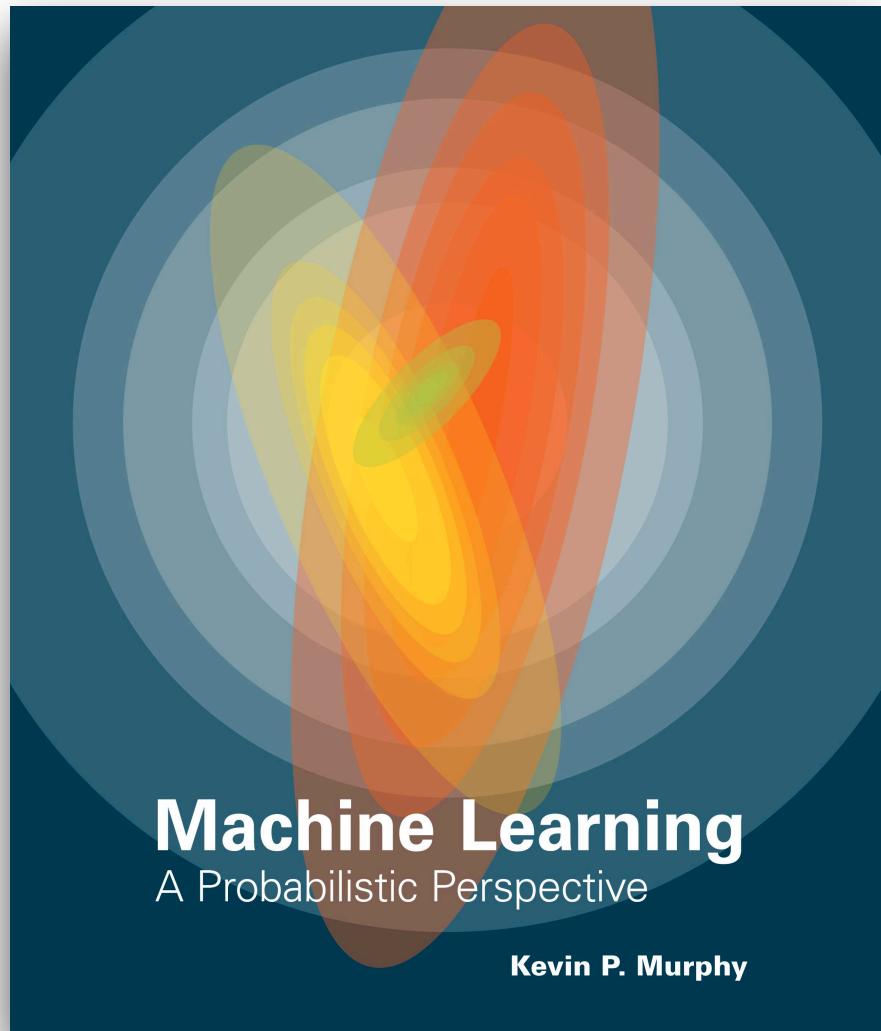
HOW SCIENCE GOES WRONG

Real Cause for Concern

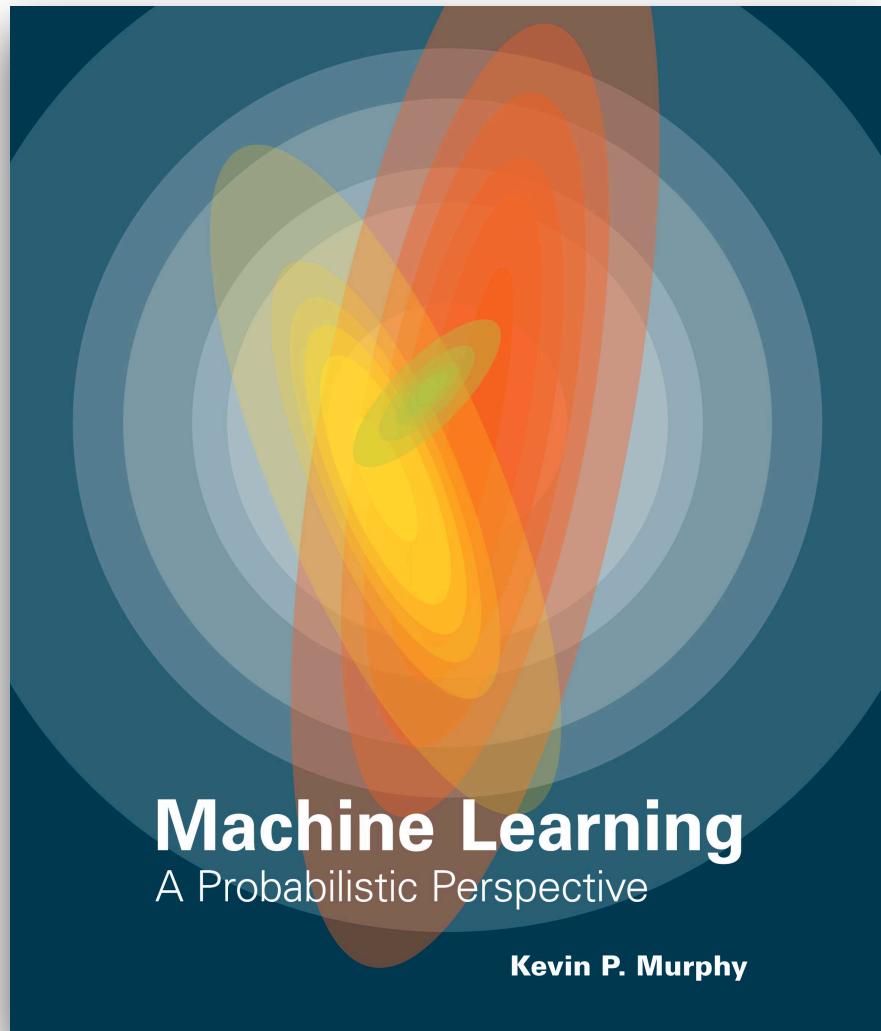


Also true for **CIFAR-10**: fixed, public train / test split since 2008.

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



[...] 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.



[...] 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.

Int J Comput Vis (2010) 88: 303–338
DOI 10.1007/s11263-009-0275-4

The PASCAL Visual Object Classes (VOC) Challenge

Mark Everingham · Luc Van Gool ·
Christopher K. I. Williams · John Winn ·
Andrew Zisserman

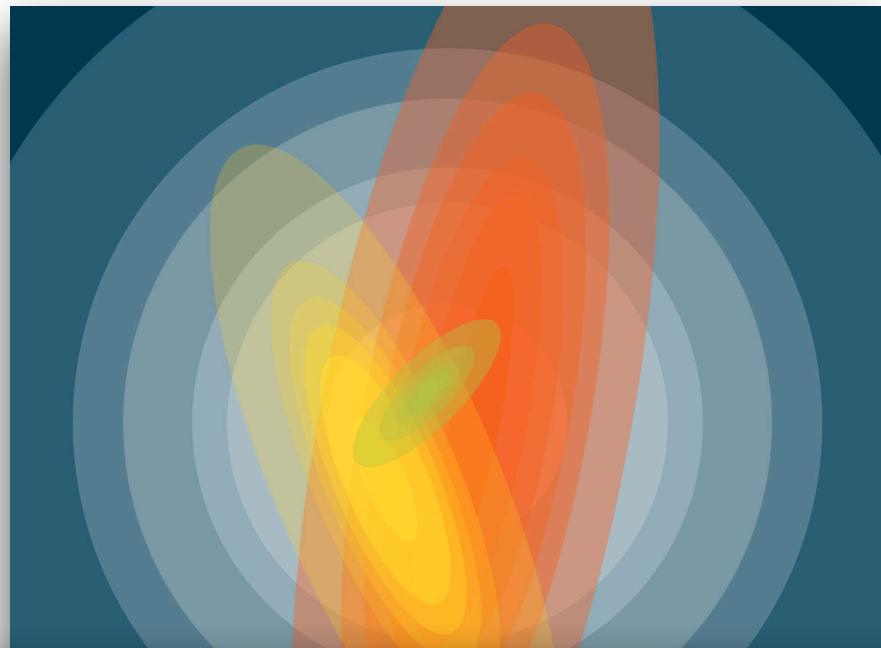
Received: 30 July 2008 / Accepted: 16 July 2009 / Published online: 9 September 2009
© Springer Science+Business Media, LLC 2009

Abstract The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. Organised annu-

1 Introduction

The PASCAL¹ Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available *dataset* of images and annotation, together with standardised eval-

[...] new test data is required each year in order to avoid participants [...] overfitting on the test set.



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

Received: 30 July 2008 / Accepted: 16 July 2009 / Published online: 9 September 2009
© Springer Science+Business Media, LLC 2009

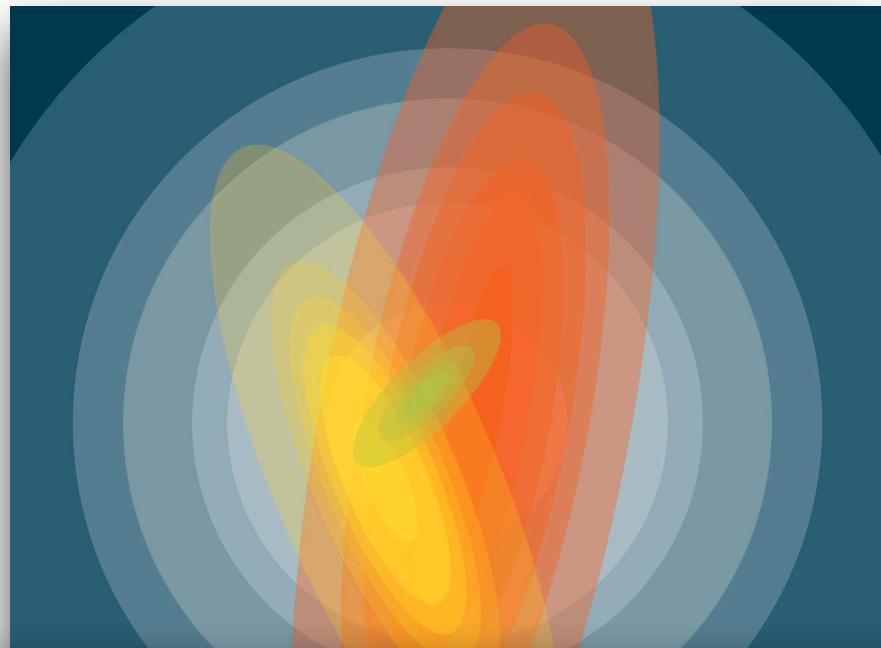
Abstract The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. Organised annu-

1 Introduction

The PASCAL¹ Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available *dataset* of images and annotation, together with standardised eval-

[...] 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 'den rules' of machine learning research.

[...] new test data is required each year in order to avoid participants [...] overfitting on the test set.



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

Received: 30 July 2008 / Accepted: 16 July 2009 / Published online: 9 September 2009
© Springer Science+Business Media, LLC 2009

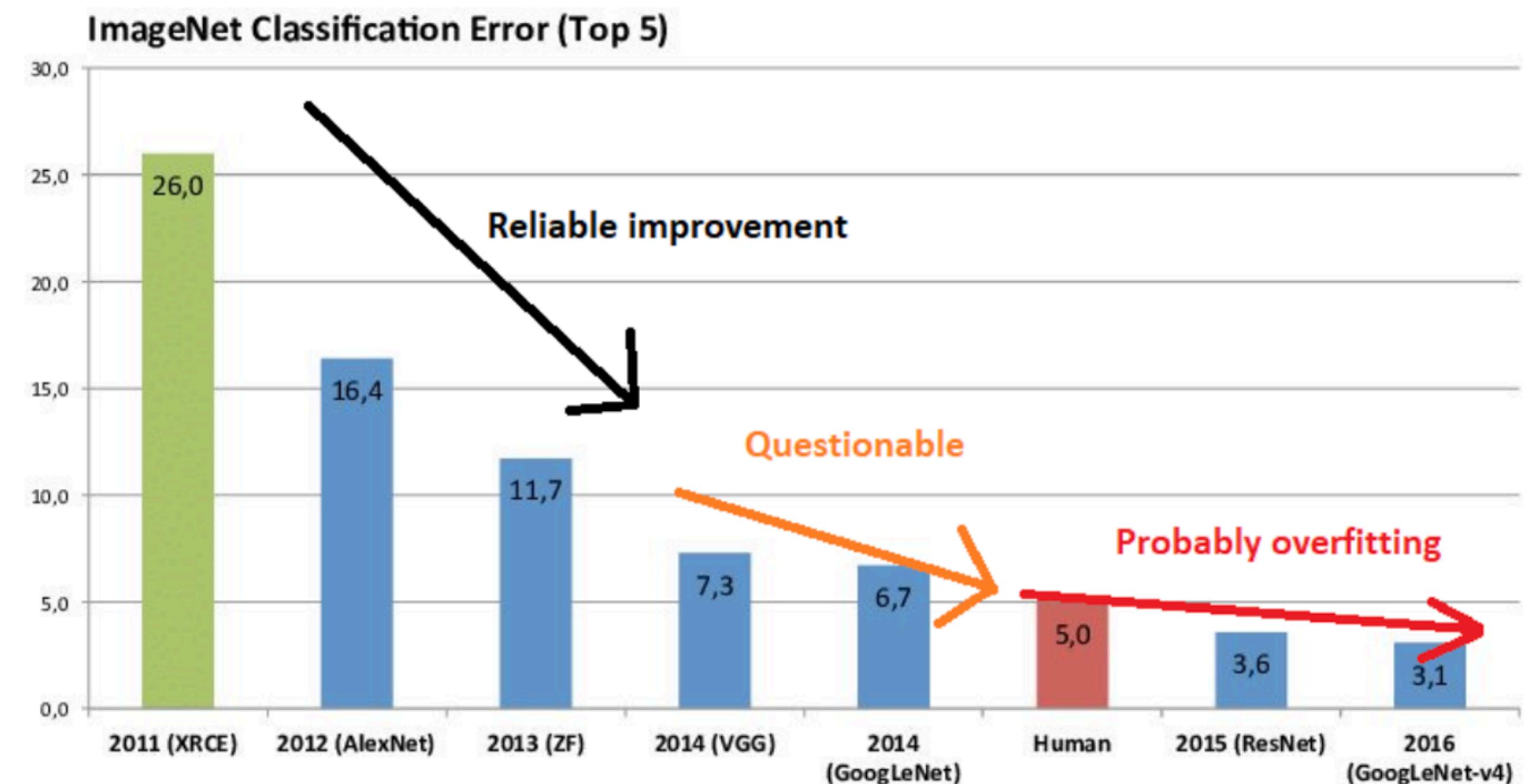
Abstract The PASCAL Visual Object Classes (VOC) challenge is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. Organised annu-

1 Introduction

The PASCAL¹ Visual Object Classes (VOC) Challenge consists of two components: (i) a publicly available *dataset* of images and annotation, together with standardised eval-

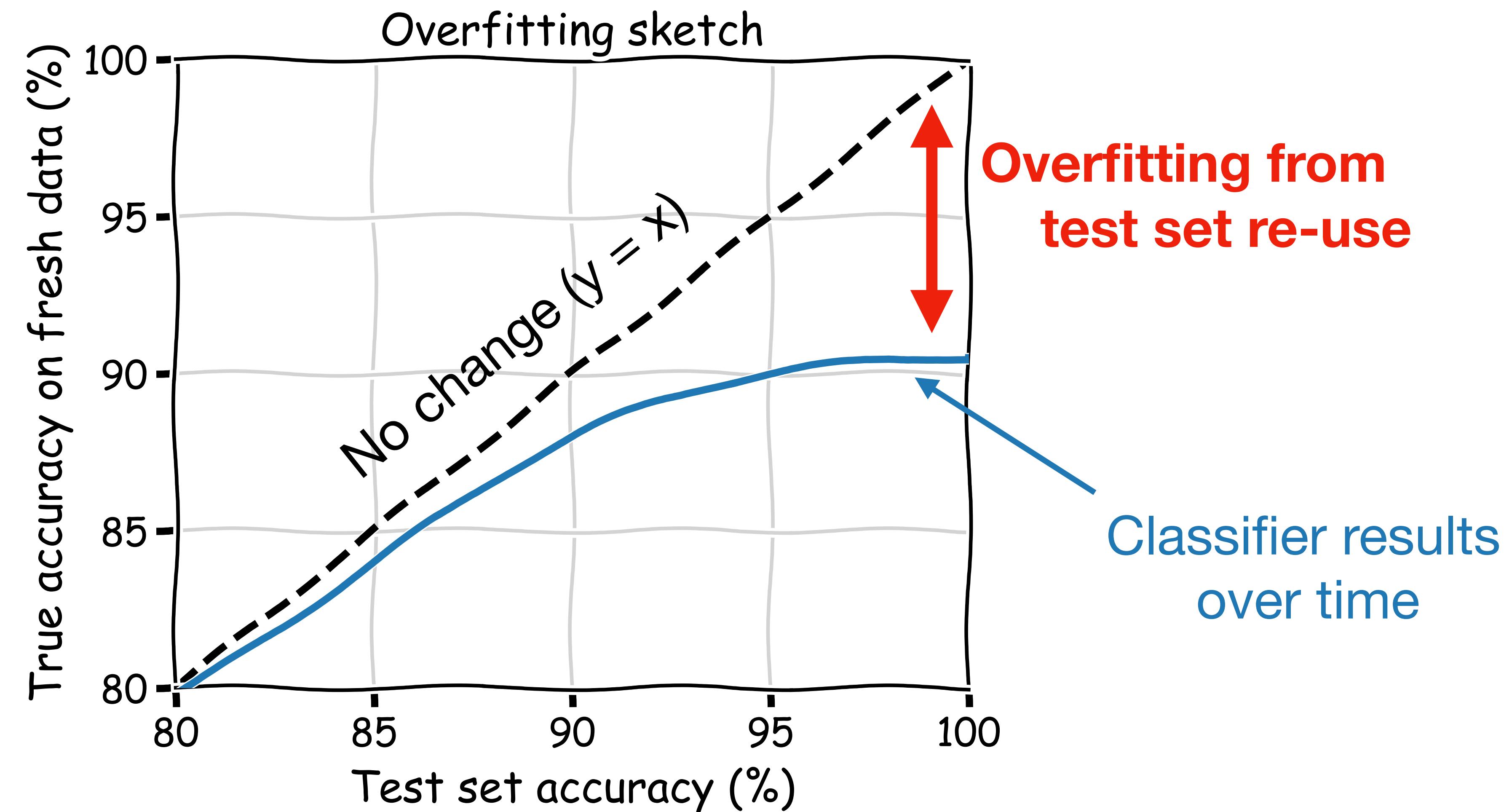
[...] we should not use [the test set] for model fitting or model selection, otherwise we will get an unrealistically optimistic estimate of performance of our method. This is ‘den rules’ of machine learning research.

AI competitions don’t produce useful models



Danger with Test Set Re-Use: Overfitting

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



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**.

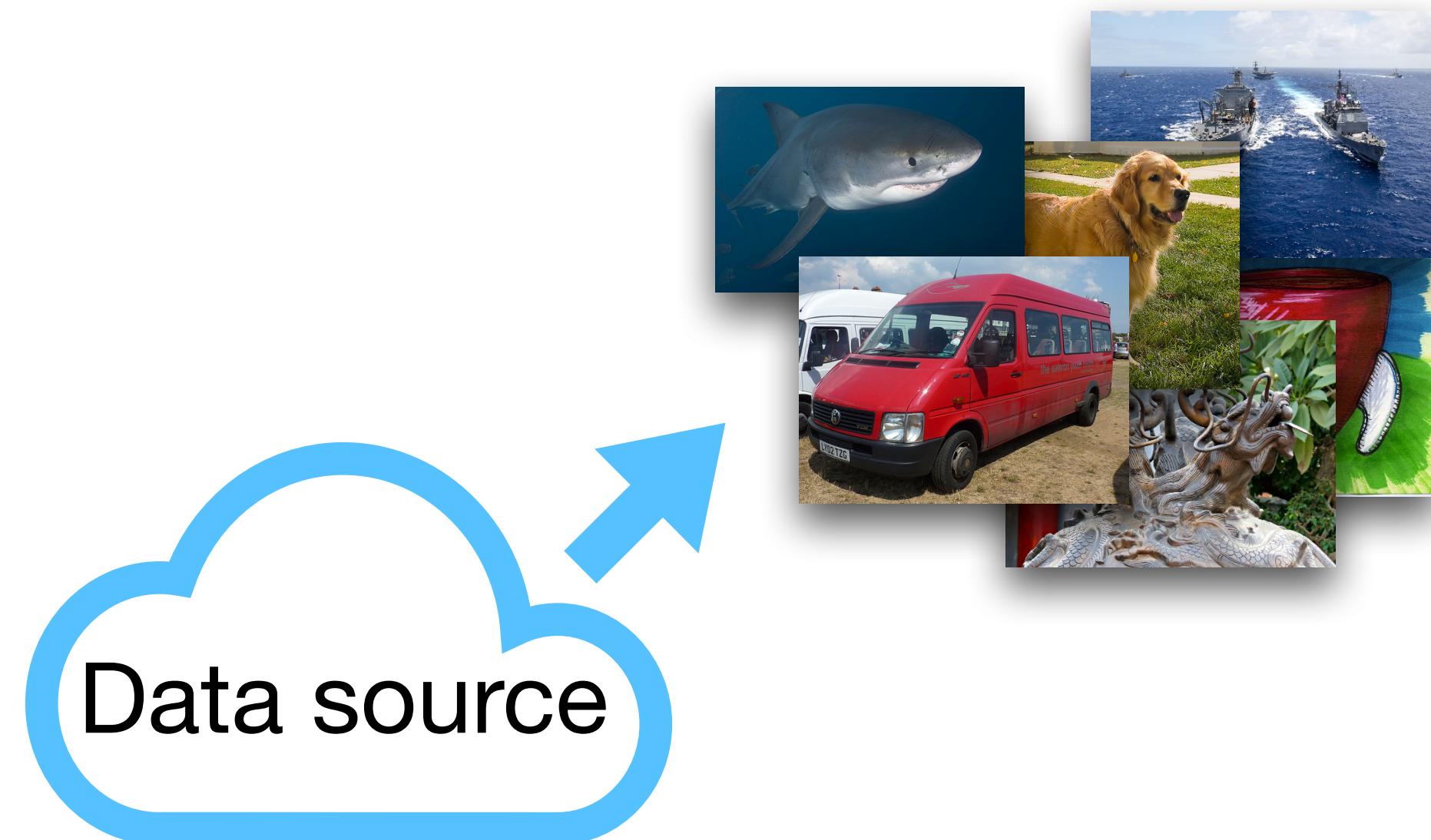
Generalization

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



Generalization

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



Generalization

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



Generalization

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



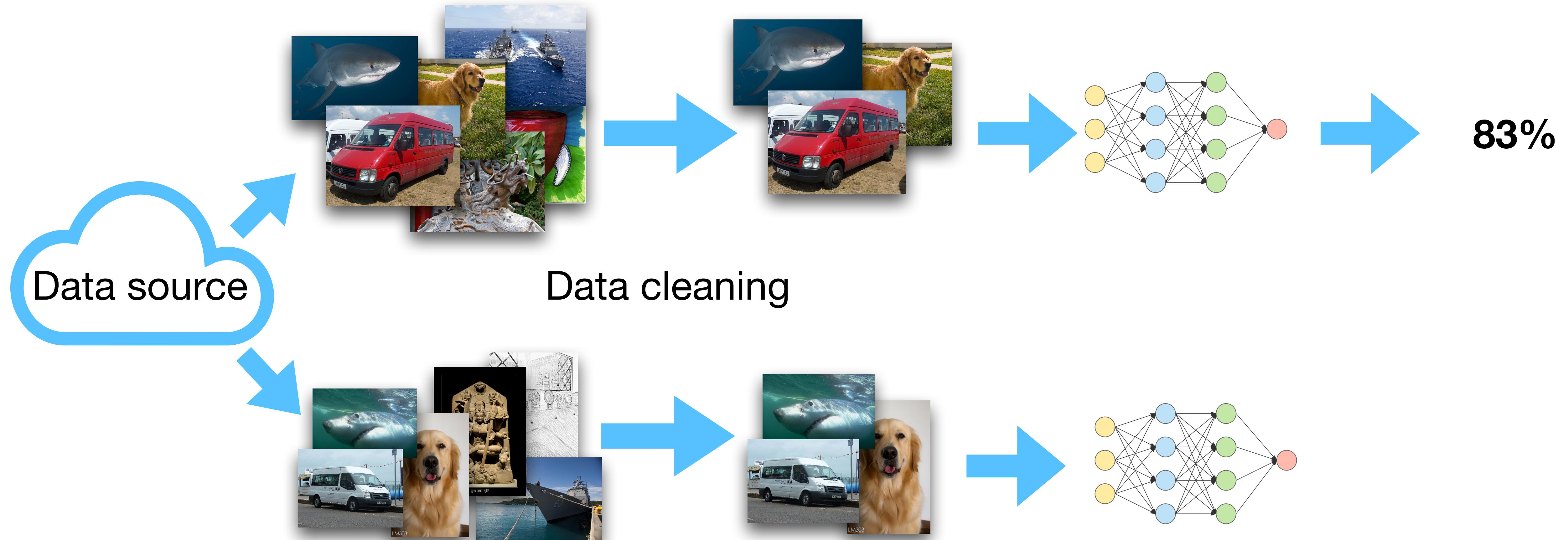
Generalization

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



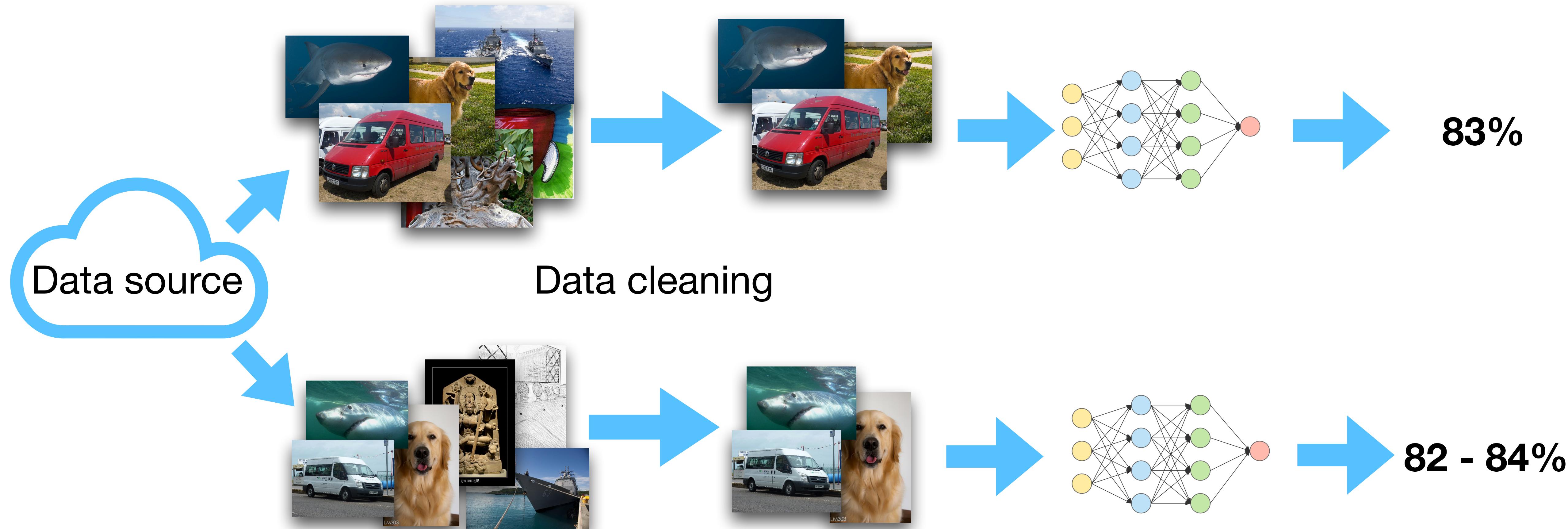
Generalization

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



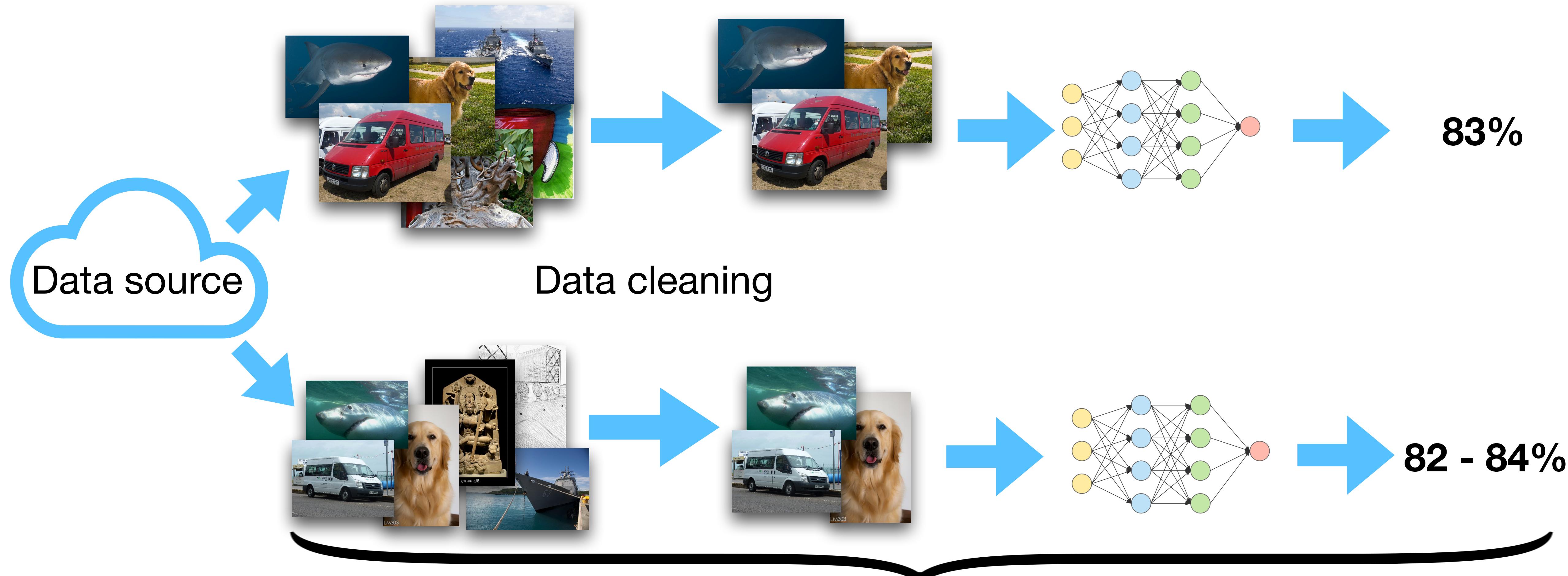
Generalization

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



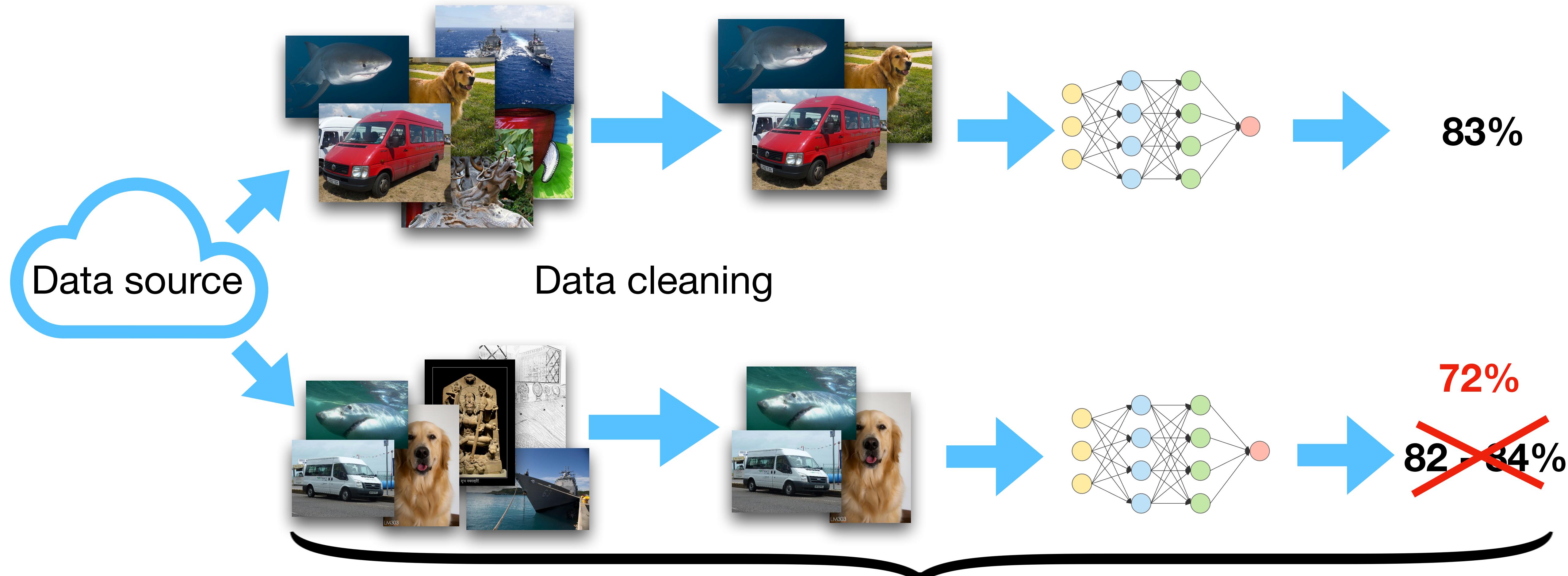
Generalization

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



Generalization

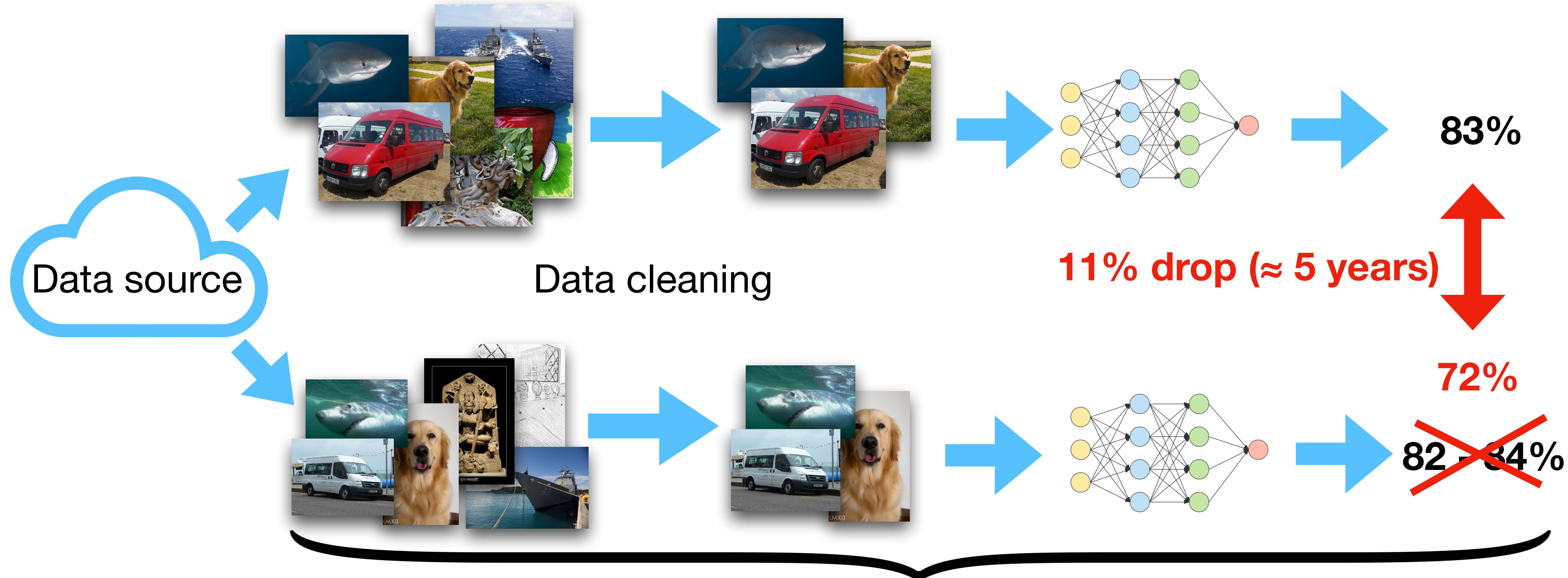
At least, the classifiers should perform similarly well on new data from the **same source**.



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

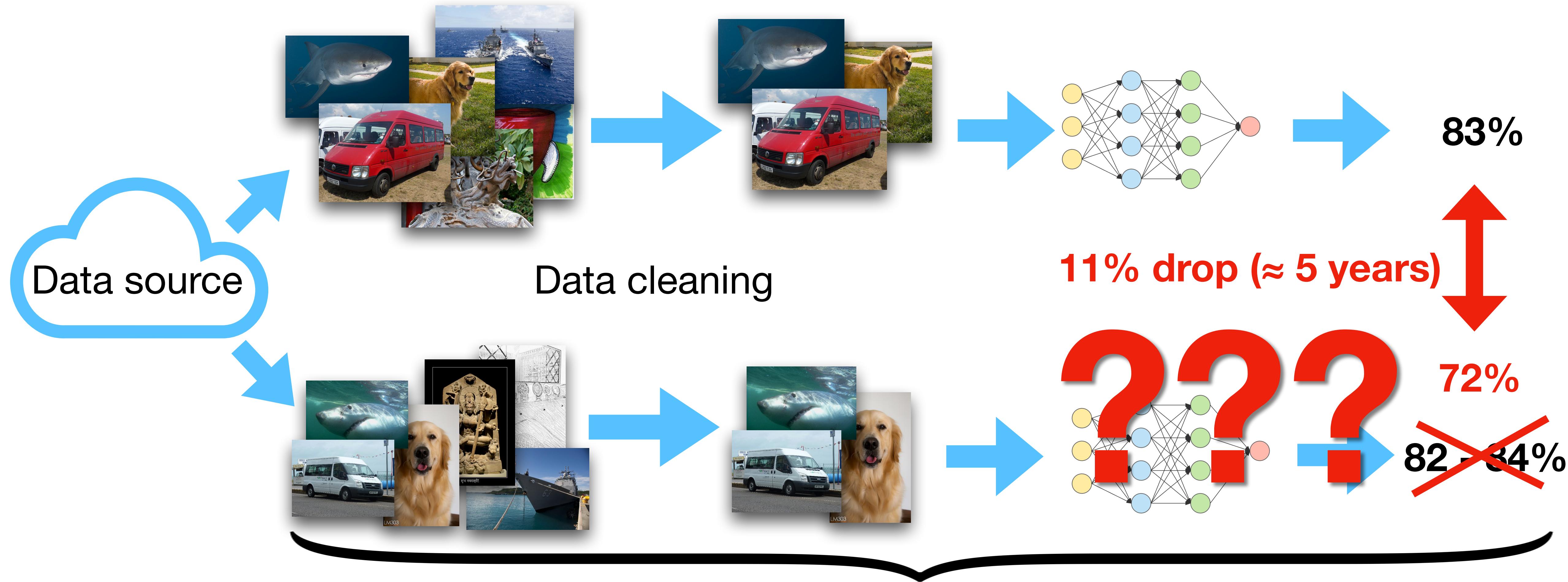
Generalization

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



Generalization

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



Overfitting

Overfitting

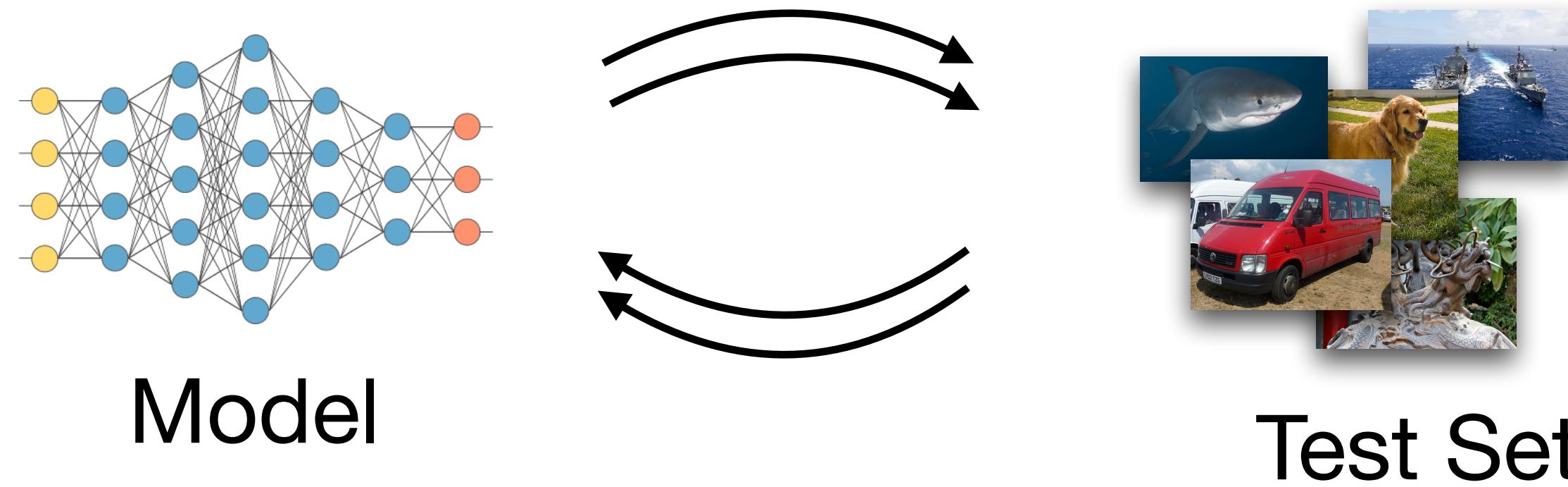
?

Three Forms of Overfitting

1. Test error \geq training error

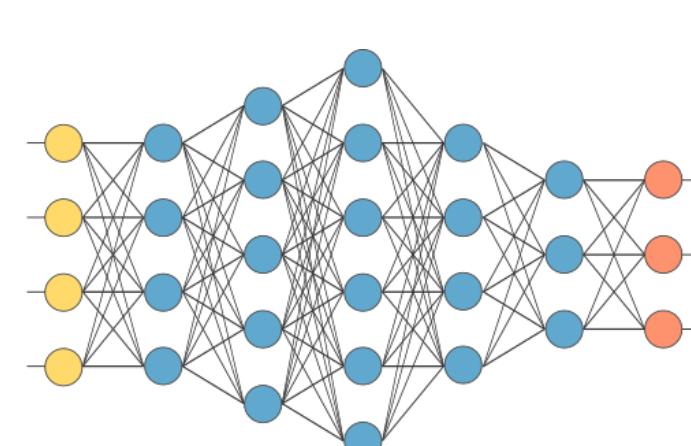
Three Forms of Overfitting

1. Test error \geq training error
2. Overfitting through test set re-use

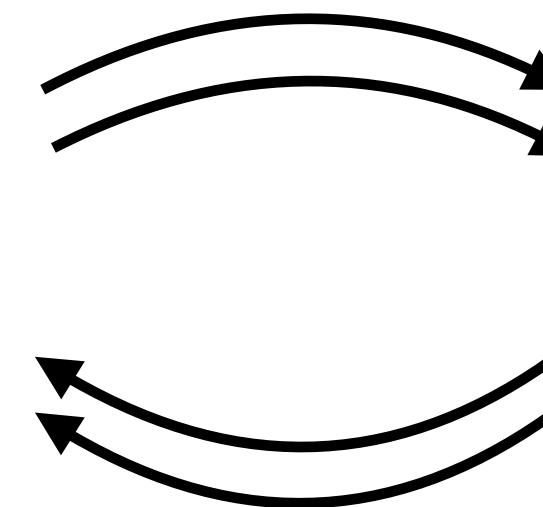


Three Forms of Overfitting

1. Test error \geq training error
2. Overfitting through test set re-use



Model

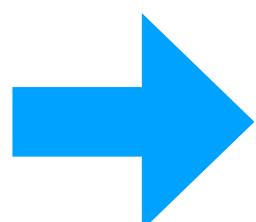


Test Set

3. Distribution shift



Original Test Set



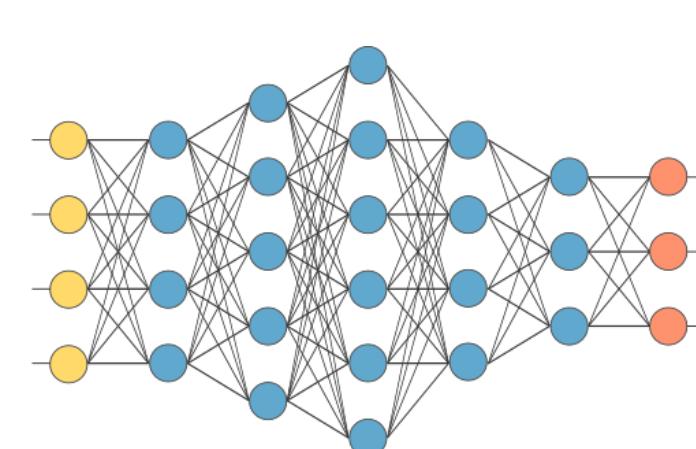
New Test Set

Three Forms of Overfitting

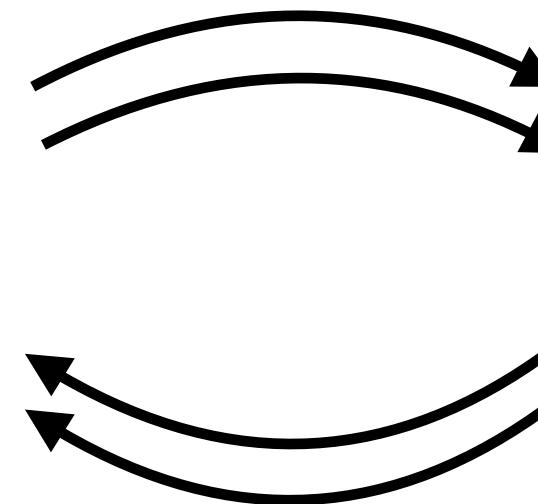
1. Test error \geq training error

Three Forms of Overfitting

1. Test error \geq training error
2. Overfitting through test set re-use



Model

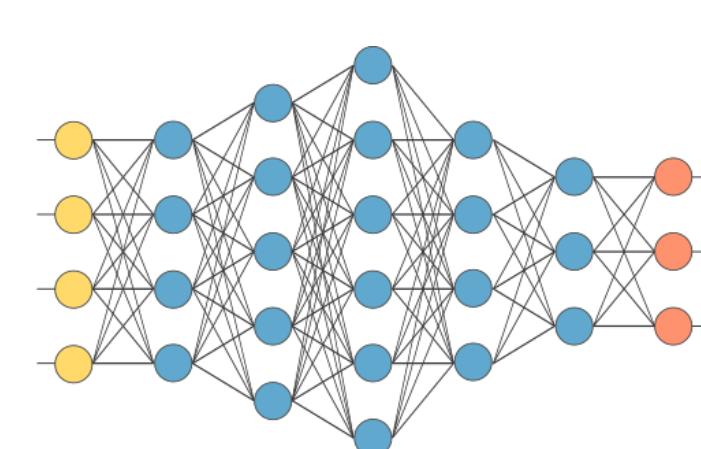


Test Set

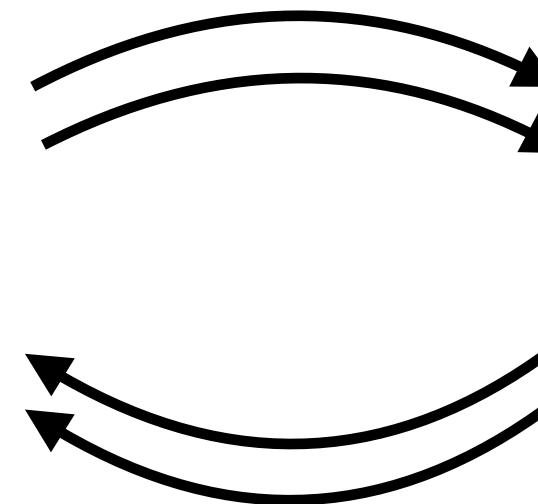
Three Forms of Overfitting

1. Test error \geq training error

2. Overfitting through test set re-use

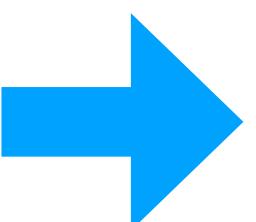


Model



Test Set

3. Distribution shift



Original Test Set

New Test Set

Two Possible Causes

New test accuracy

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

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

Two Possible Causes

New test accuracy

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

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

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

Two Possible Causes

New test accuracy

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

$\approx 11\%$

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

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

Two Possible Causes

$$\widehat{\text{acc}}_S(f) - \widehat{\text{acc}}_{S'}(f) = \widehat{\text{acc}}_S(f) - \text{acc}_D(f)$$

New test accuracy

Overfitting through test set re-use

$\approx 11\%$

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

A black bracket is placed under the term $\widehat{\text{acc}}_S(f) - \widehat{\text{acc}}_{S'}(f)$. Below this bracket, the text $\approx 11\%$ is centered. A black arrow points from the text "New test accuracy" to the term $\widehat{\text{acc}}_{S'}(f)$. A red arrow points from the text "Overfitting through test set re-use" to the term $\text{acc}_D(f)$.

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

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

Two Possible Causes

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

New test accuracy

Overfitting through test set re-use

Distribution shift

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

The diagram illustrates the decomposition of the difference in test accuracy between two sets, S and S'. The total difference is shown as a black bracket under the equation, labeled "New test accuracy". This bracket is further divided into two components: "Overfitting through test set re-use" (in red) and "Distribution shift" (in blue). The overfitting component is the difference between the model's performance on the original test set S and its performance on a different dataset D. The distribution shift component is the difference between the model's performance on dataset D and its performance on a shifted version of dataset D, D'. A note at the bottom specifies that S is drawn from D.

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

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

Two Possible Causes

$$\widehat{\text{acc}}_S(f) - \widehat{\text{acc}}_{S'}(f) = \widehat{\text{acc}}_S(f) - \text{acc}_D(f) + \text{acc}_D(f) - \text{acc}_{D'}(f) + \text{acc}_{D'}(f) - \widehat{\text{acc}}_{S'}(f)$$

New test accuracy

Overfitting through test set re-use

Distribution shift

Generalization error ($\approx 1\%$)

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

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

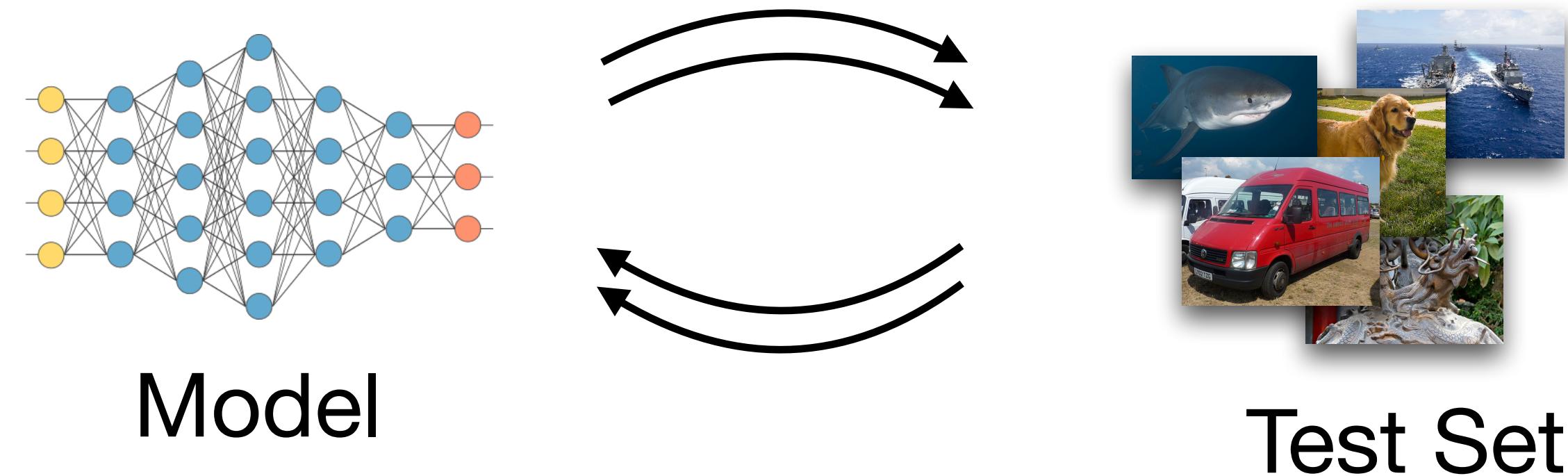
$\text{acc}_D(f) = \mathbb{E}_{(x,y) \sim D} \mathbb{1}[f(x) = y] \quad (\text{S is drawn from D})$

$\approx 11\%$

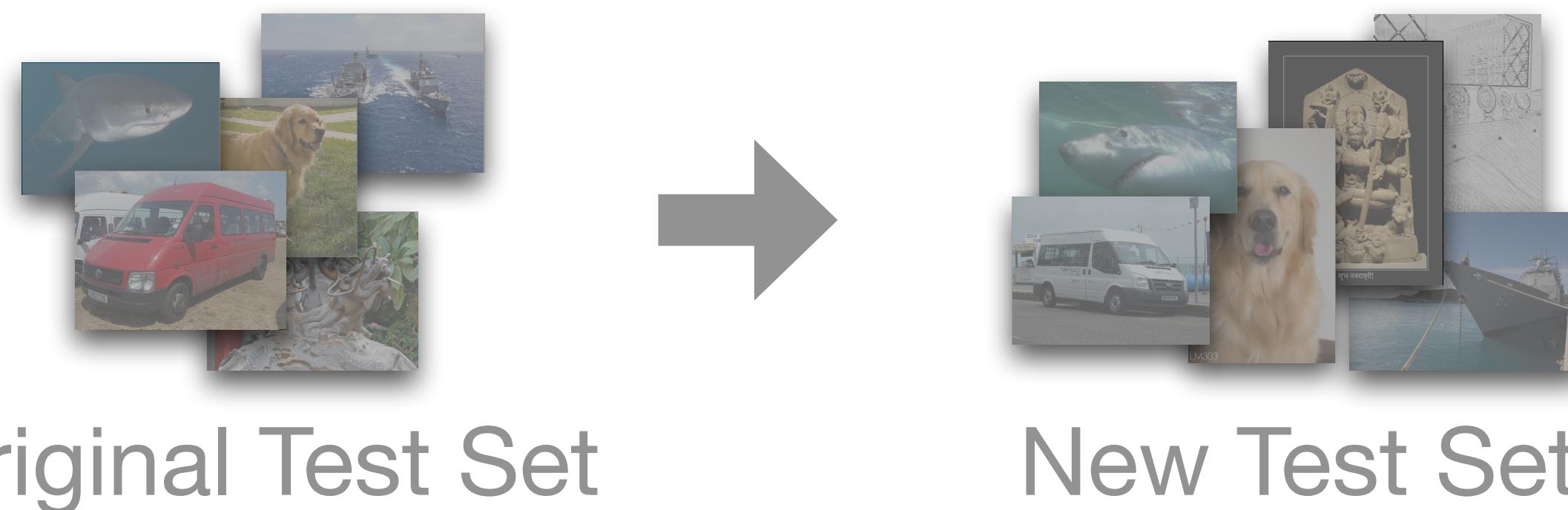
The diagram illustrates the decomposition of the difference in test accuracy between two datasets, S and S' . It starts with the equation $\widehat{\text{acc}}_S(f) - \widehat{\text{acc}}_{S'}(f) = \widehat{\text{acc}}_S(f) - \text{acc}_D(f) + \text{acc}_D(f) - \text{acc}_{D'}(f) + \text{acc}_{D'}(f) - \widehat{\text{acc}}_{S'}(f)$. The first term, $\widehat{\text{acc}}_S(f) - \text{acc}_D(f)$, is labeled "Overfitting through test set re-use" in red. The second term, $\text{acc}_D(f) - \text{acc}_{D'}(f)$, is labeled "Distribution shift" in blue. The third term, $\text{acc}_{D'}(f) - \widehat{\text{acc}}_{S'}(f)$, is labeled "Generalization error ($\approx 1\%$)" in green. A bracket under the first term is labeled "New test accuracy". A bracket under the entire equation is labeled "Original test accuracy (orig. test set S, new S')". A bracket under the last term is labeled "Approximately 11%". Arrows point from the labels to their corresponding terms in the equation.

Three Forms of Overfitting

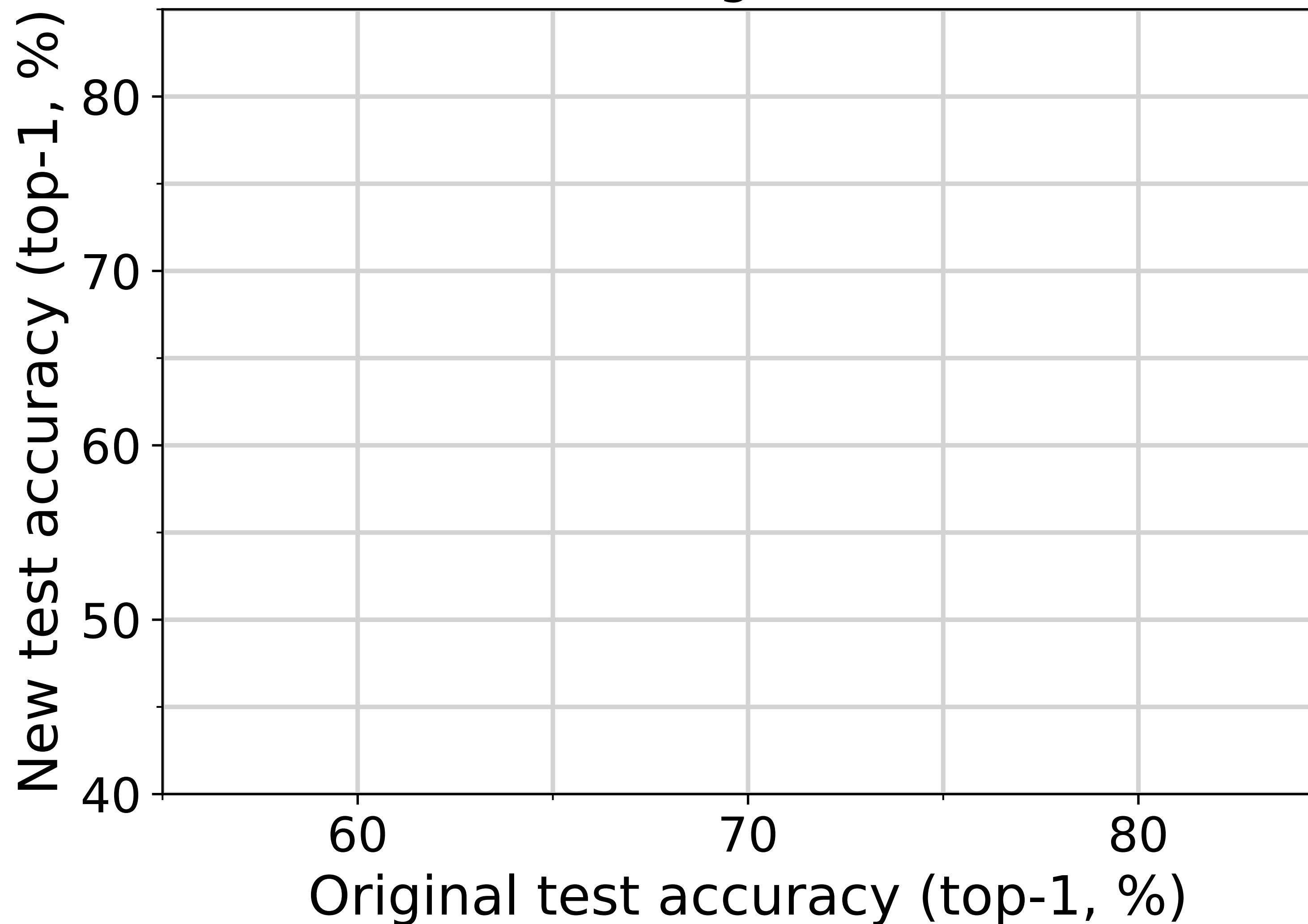
1. Test error \geq training error
2. Overfitting through test set re-use



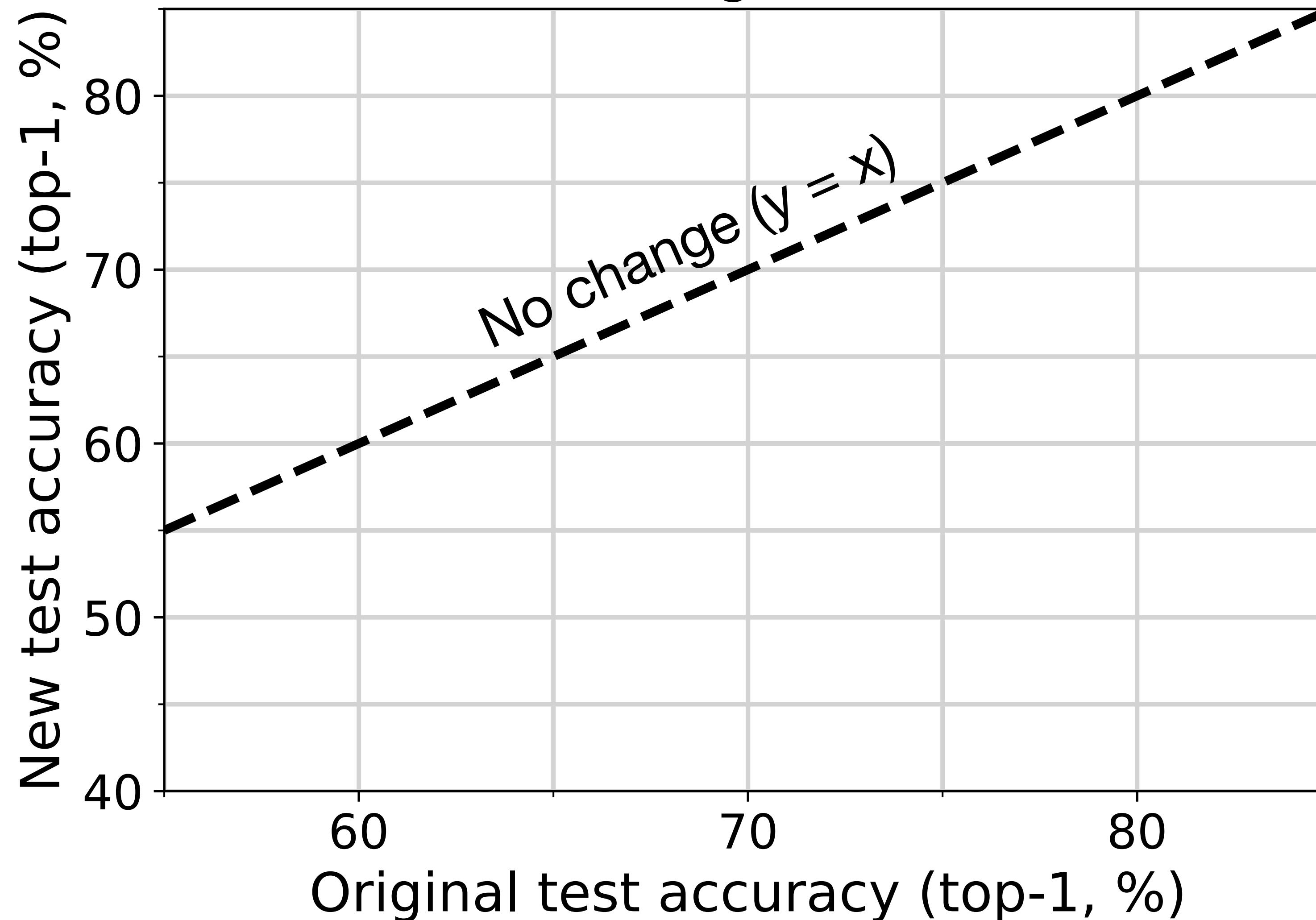
3. Distribution shift



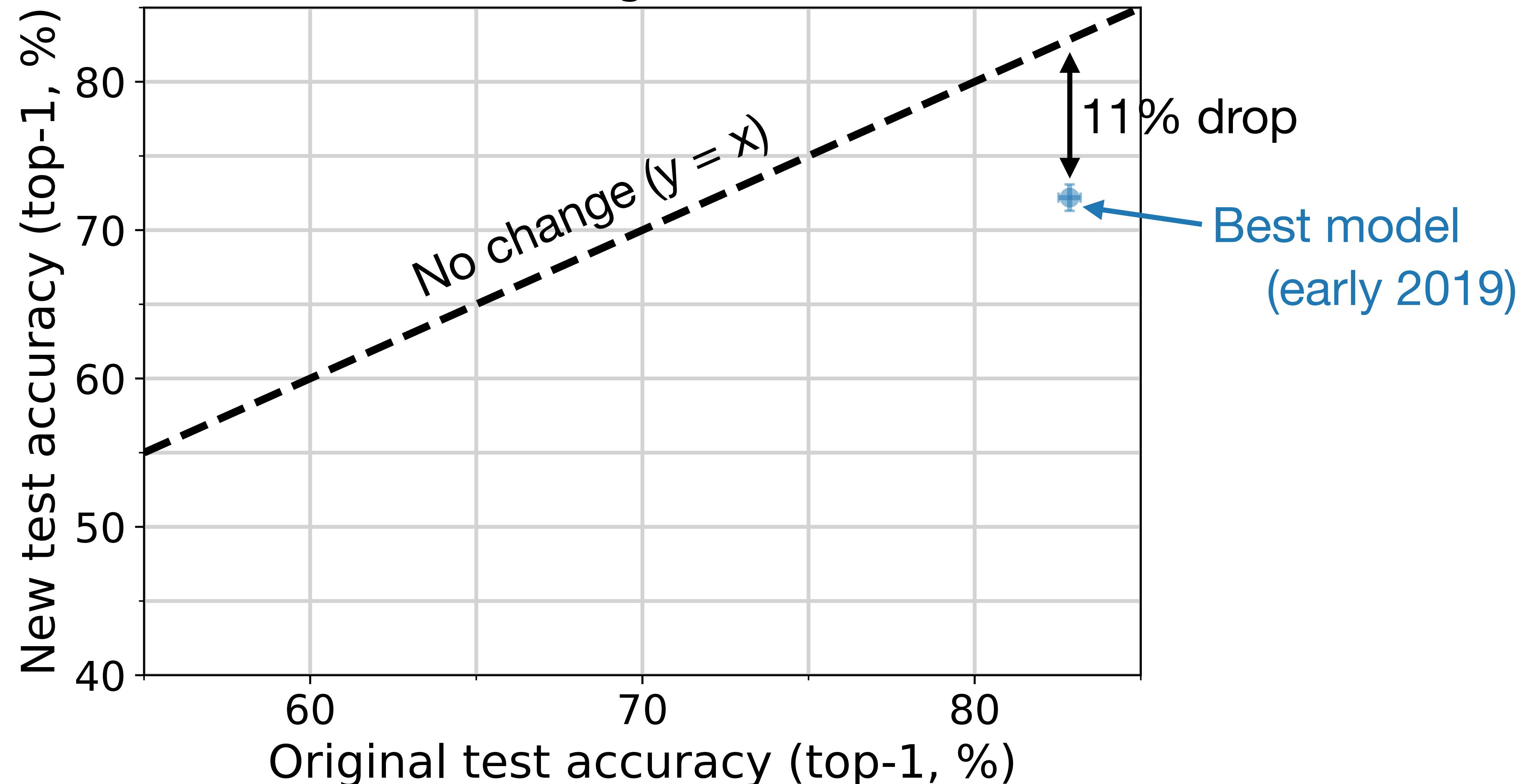
ImageNet



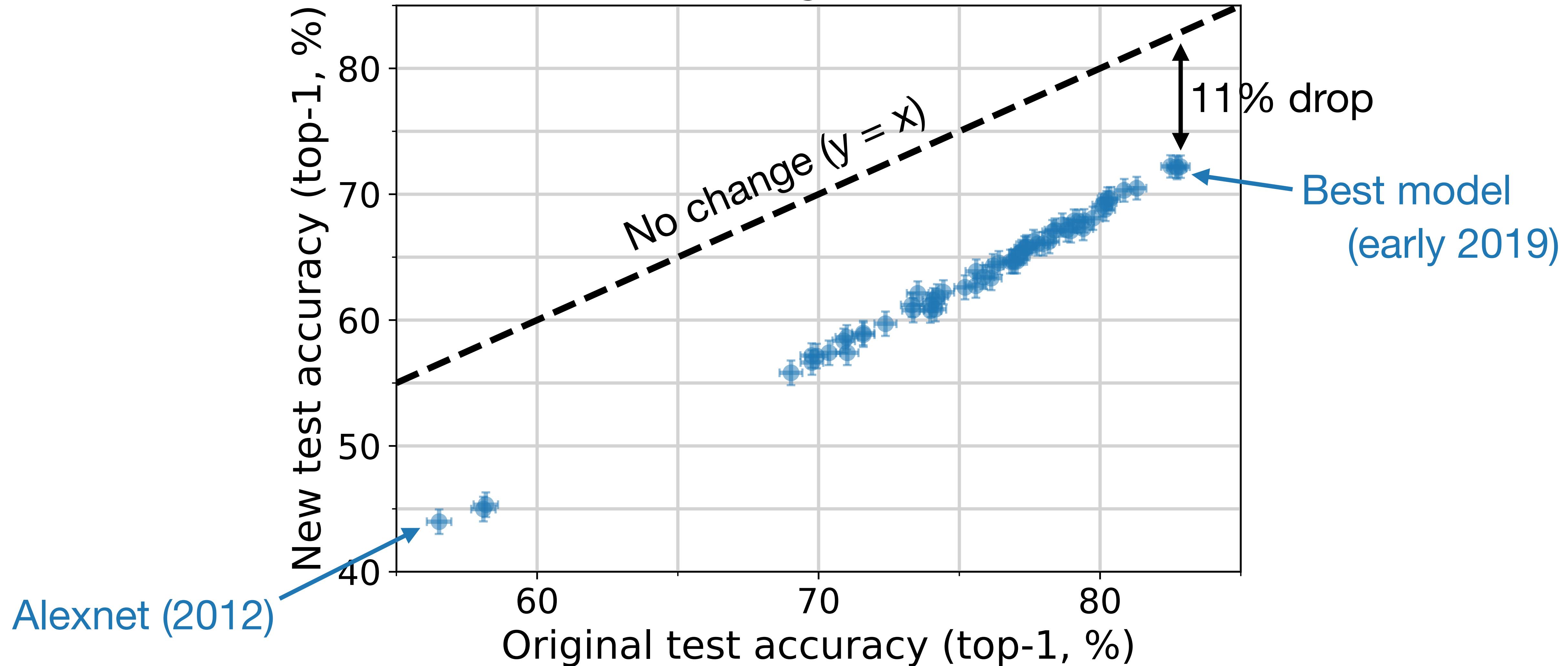
ImageNet



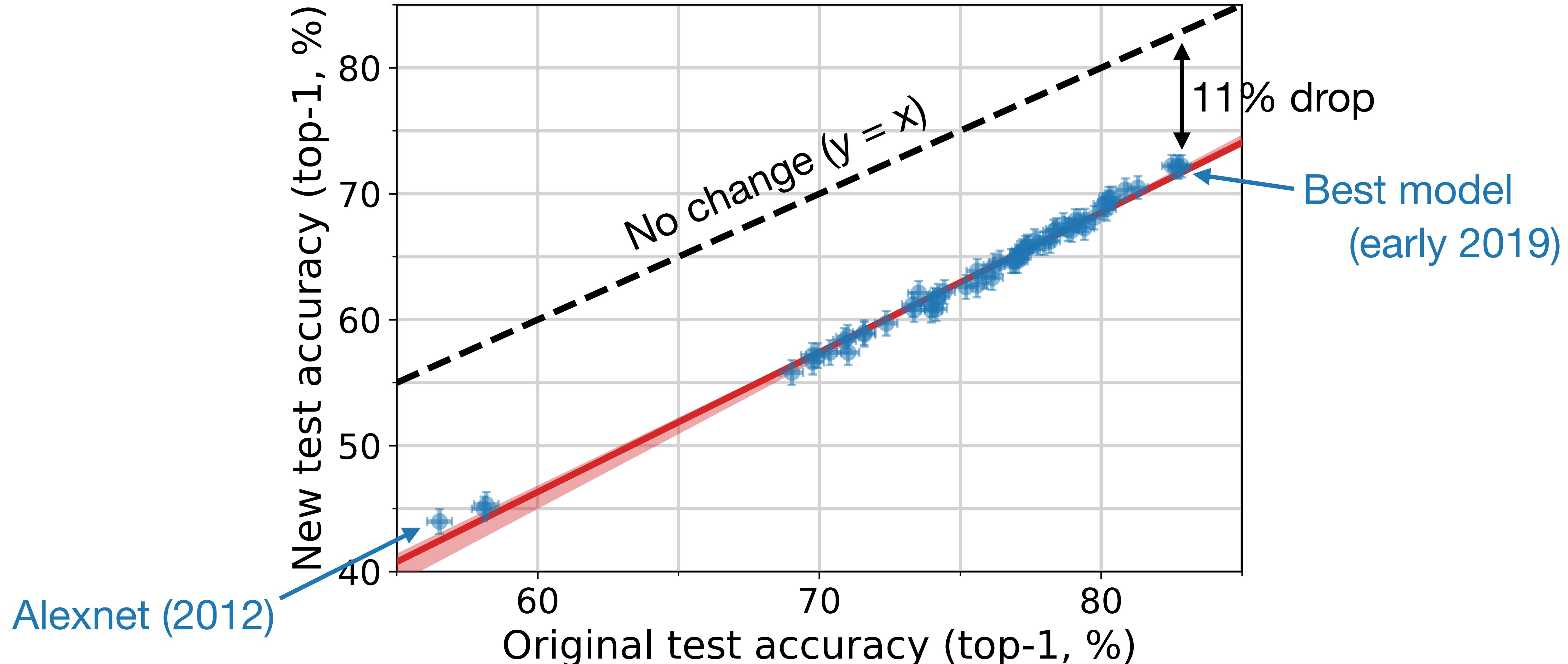
ImageNet



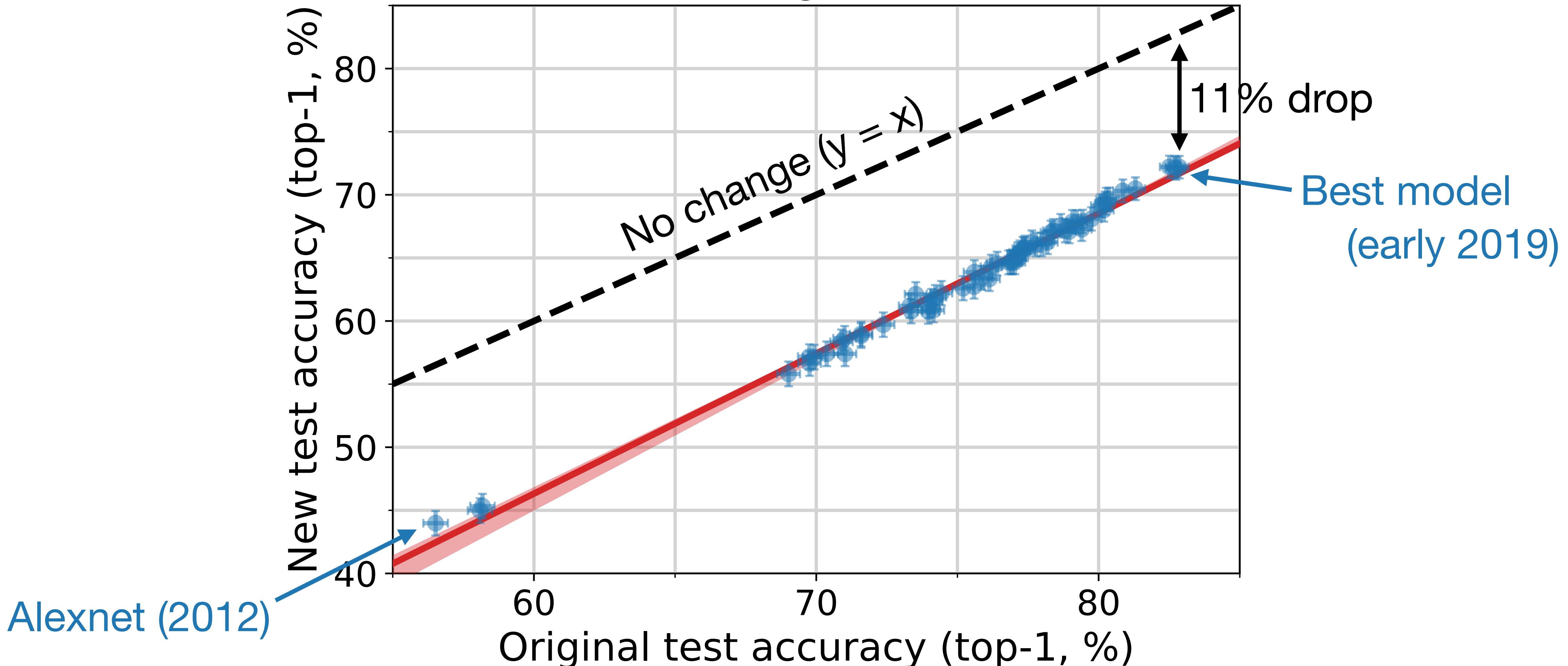
ImageNet



ImageNet

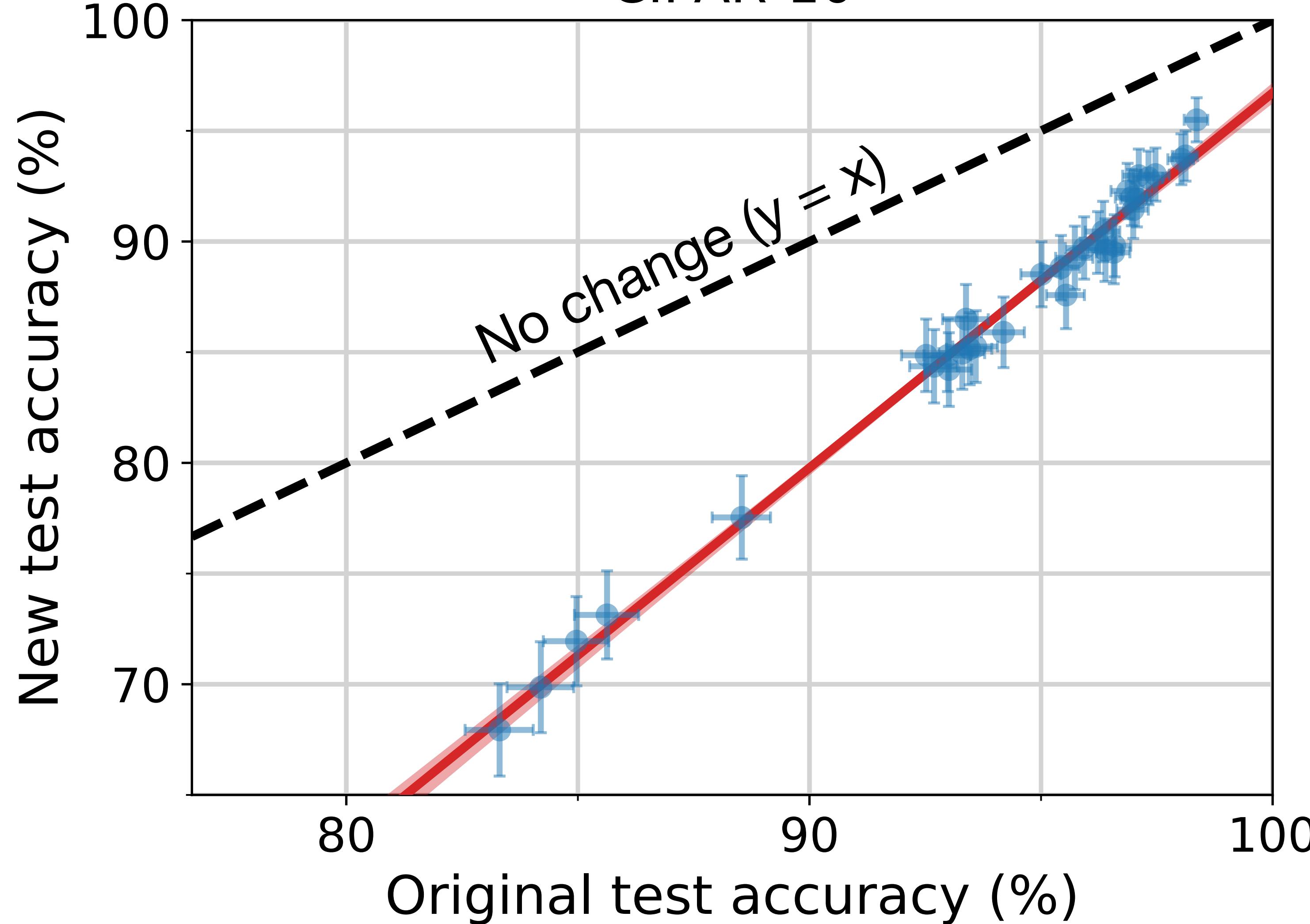


ImageNet

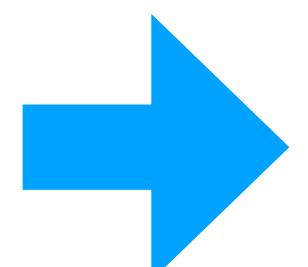
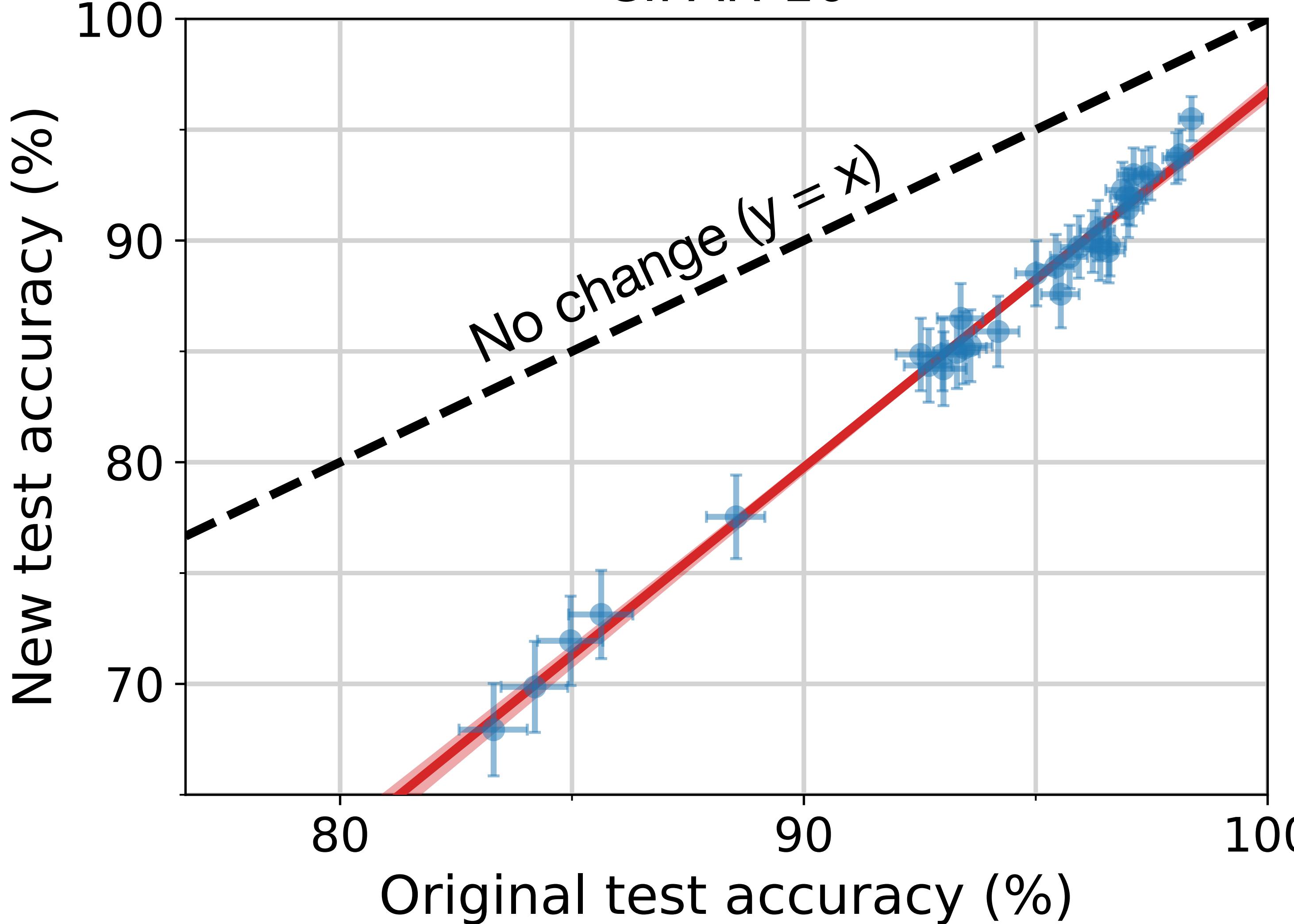


- The best models on the original test set stay the best models on the new test set.
- All models see a substantial drop in accuracy.

CIFAR-10



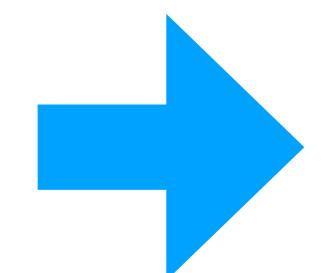
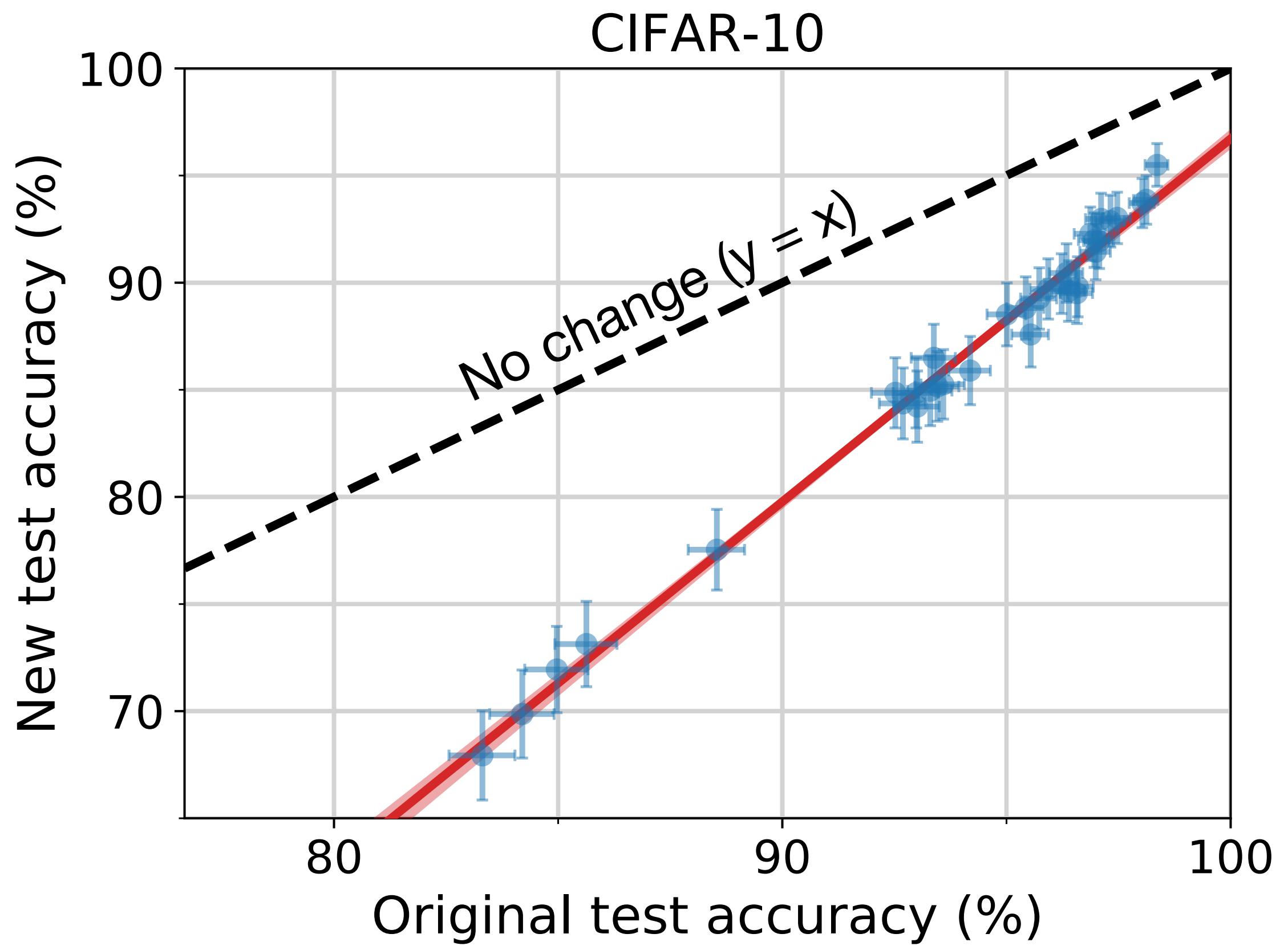
CIFAR-10



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

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

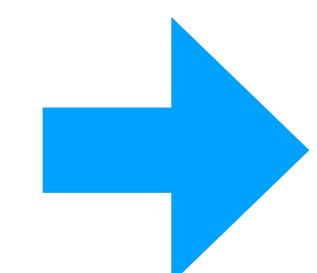
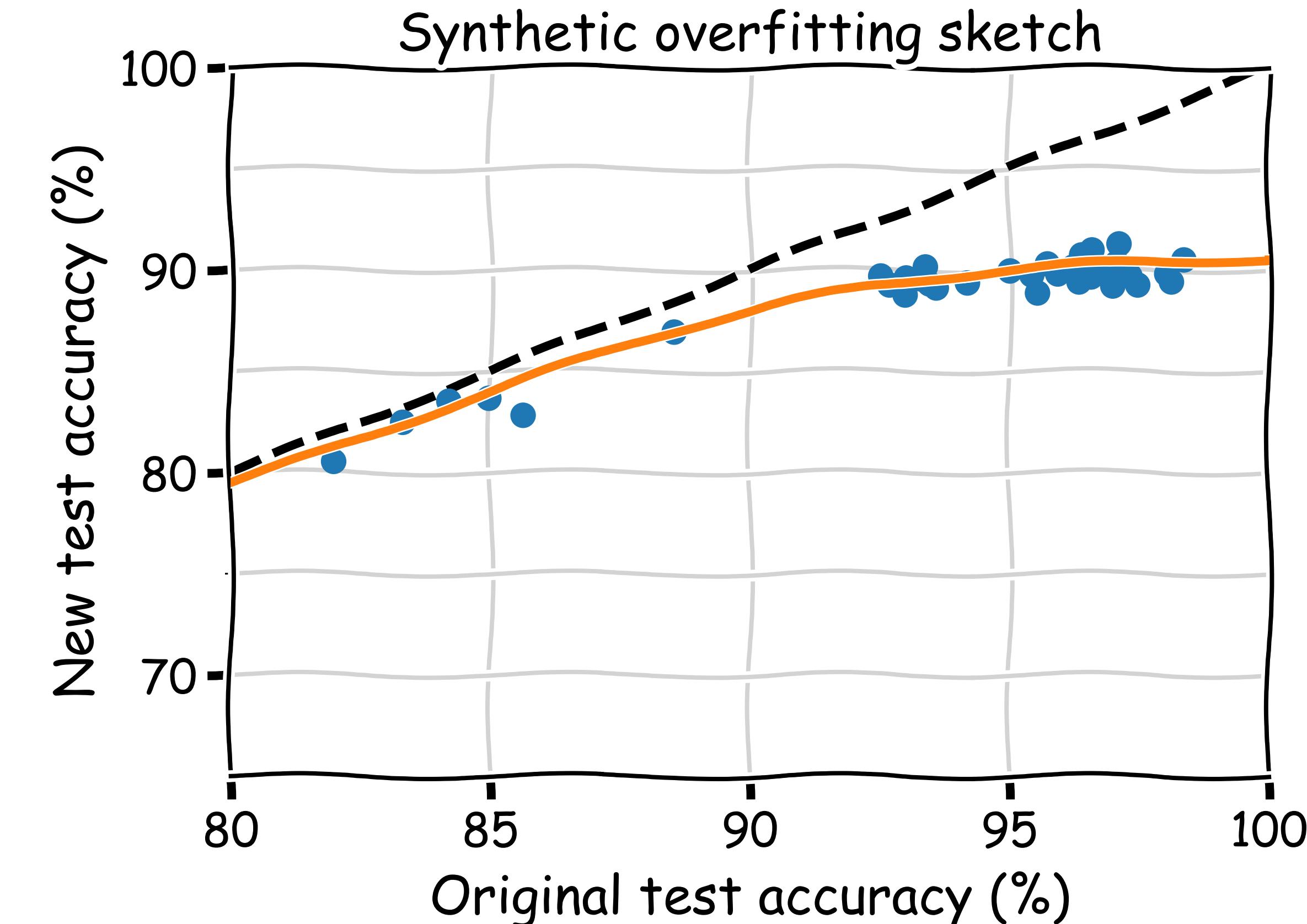
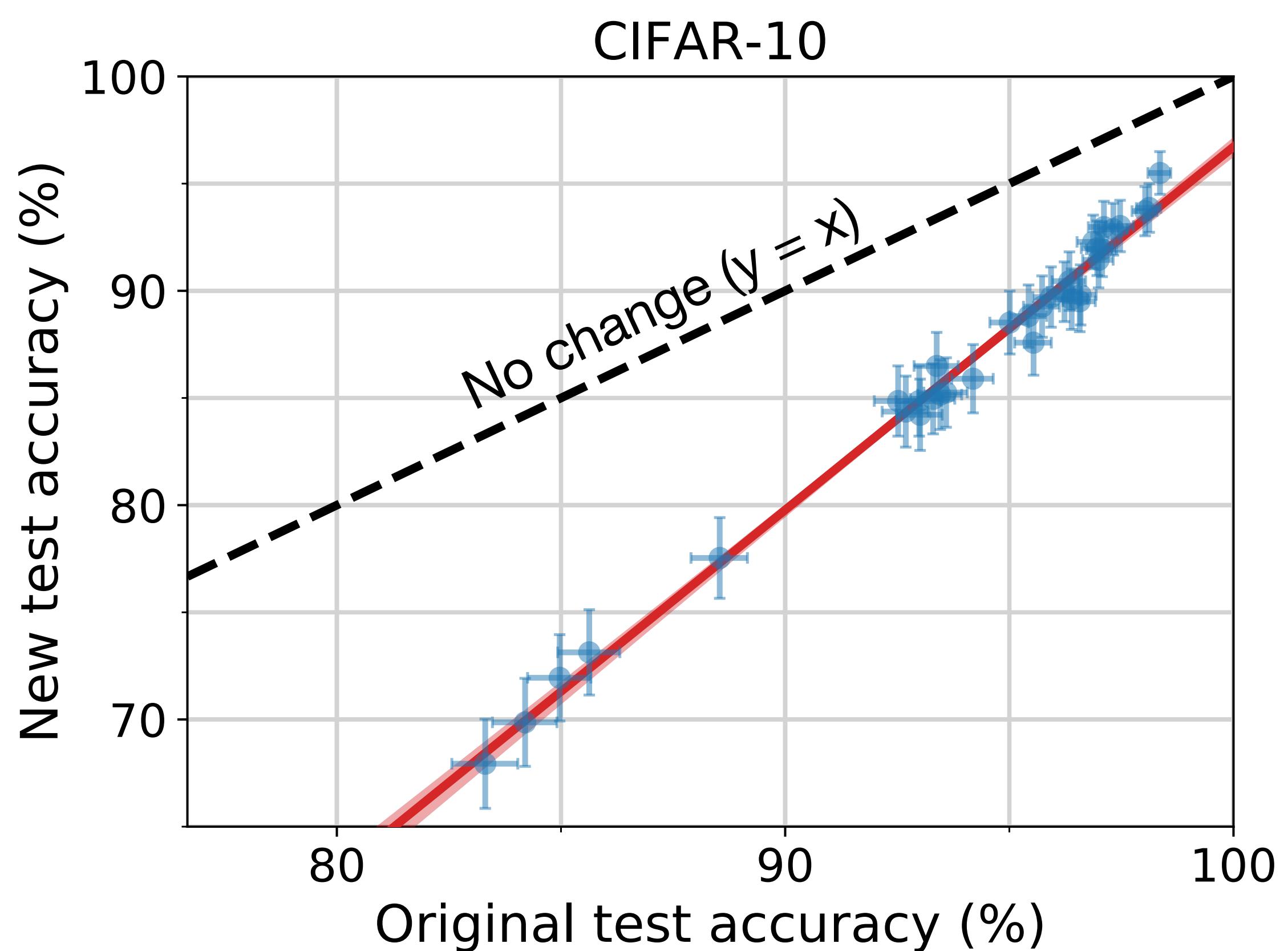
AutoAugment vs. ResNet: 10.3% difference on CIFAR-10.1



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

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

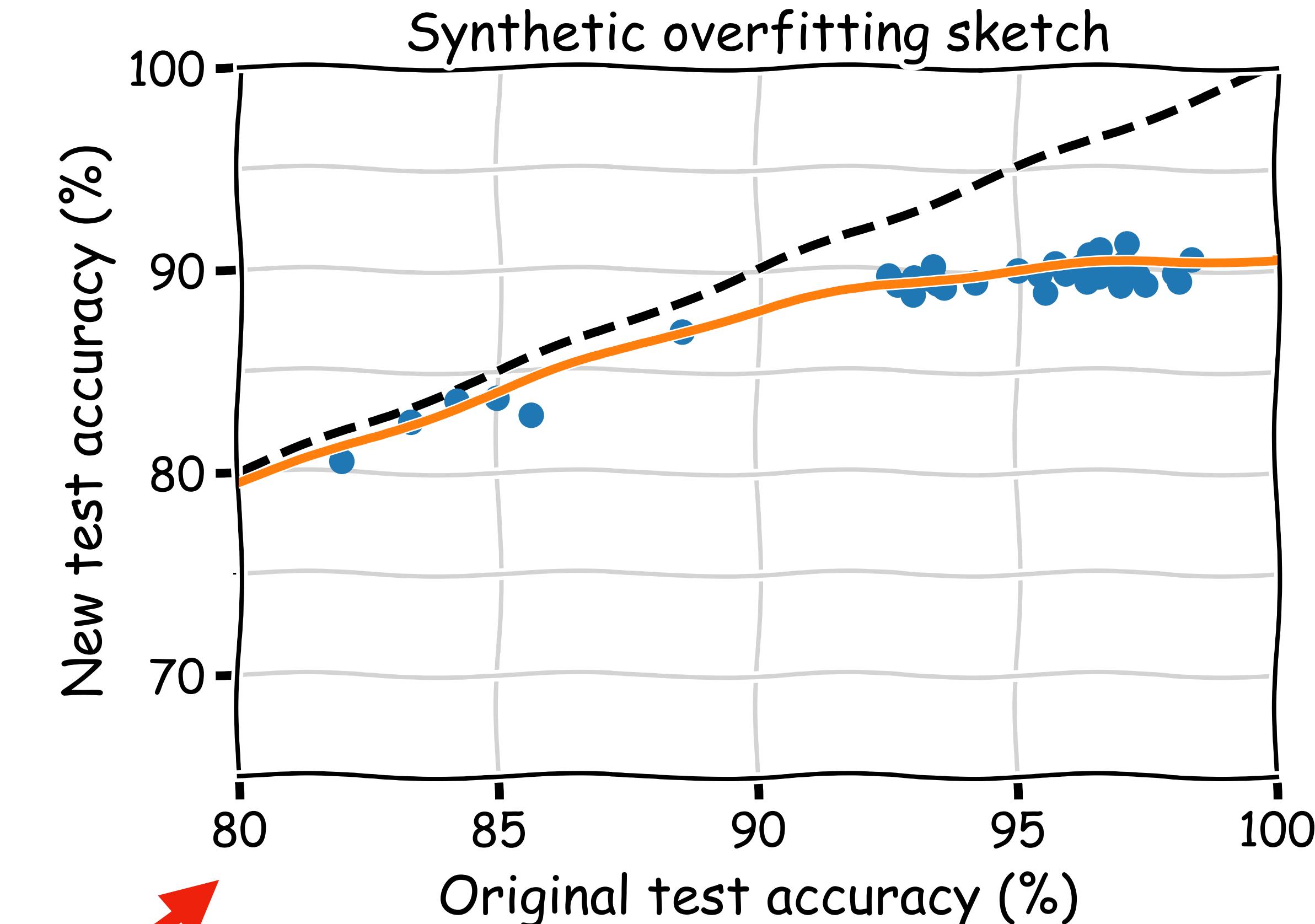
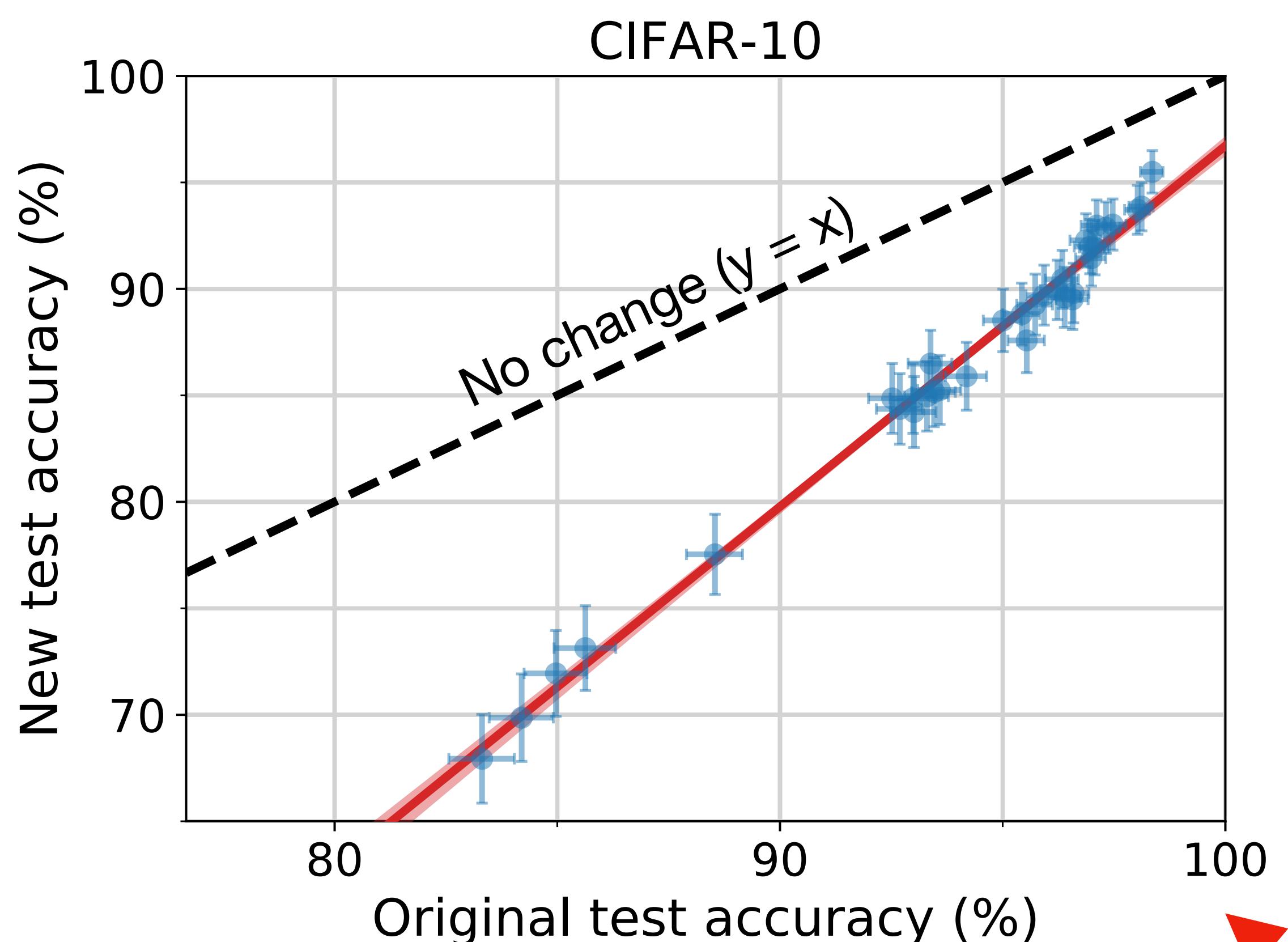
AutoAugment vs. ResNet: 10.3% difference on CIFAR-10.1



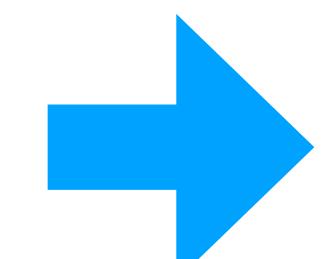
Later models see a **smaller** drop in accuracy.

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

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



Exact opposite!



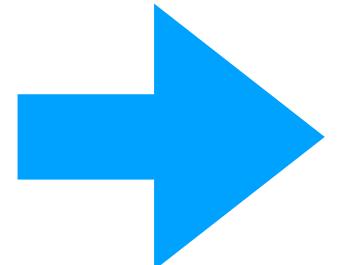
Later models see a **smaller** drop in accuracy.

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!

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



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



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.

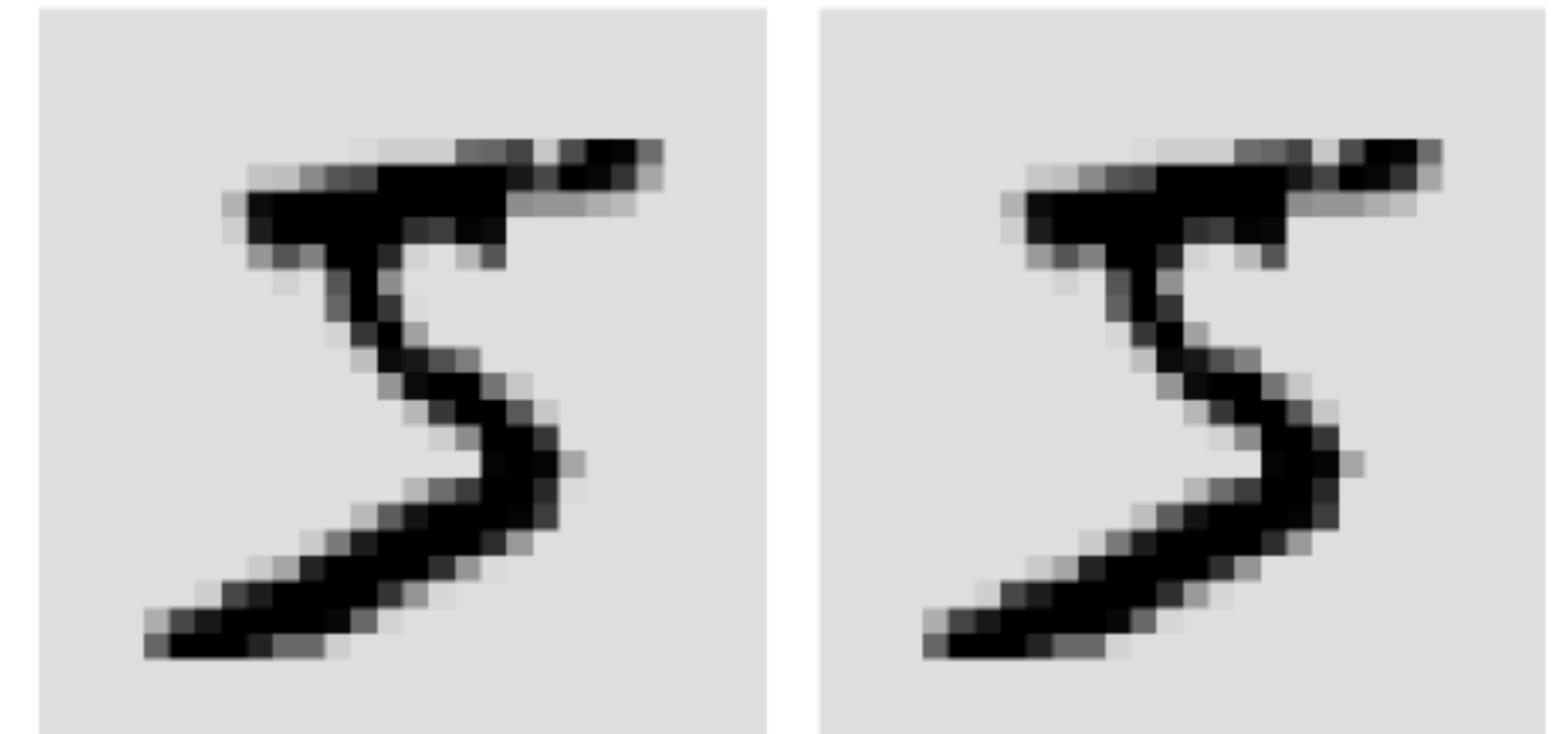
Overfitting Is Surprisingly Absent

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

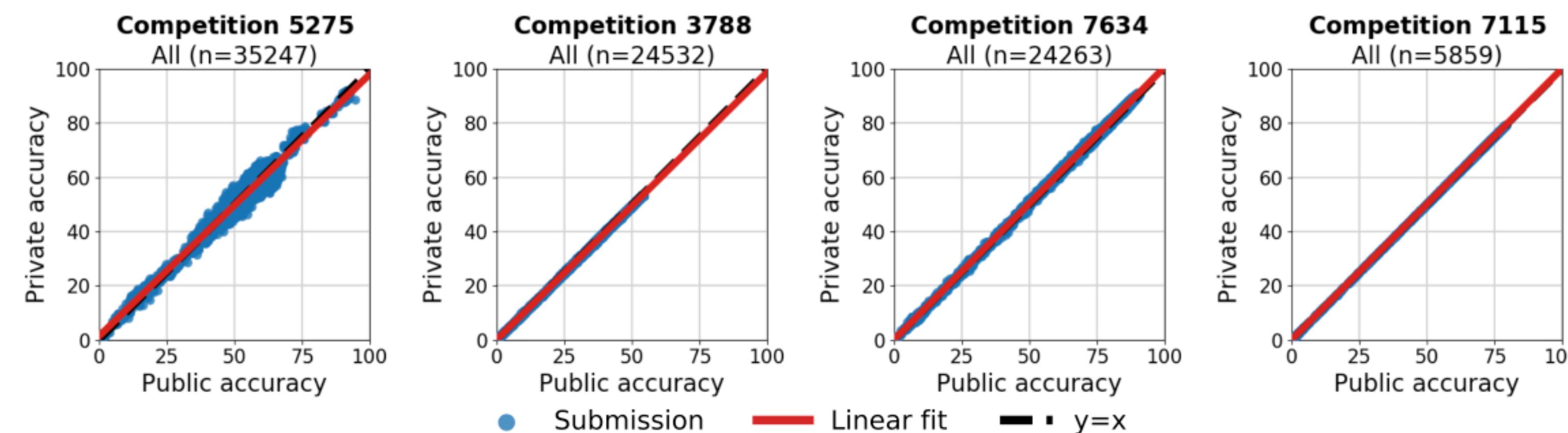
→ Relative ordering preserved. Progress is real!

MNIST: similar conclusions in [\[Yadav, Bottou'19\]](#)

no overfitting after 20+ years of MNIST



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

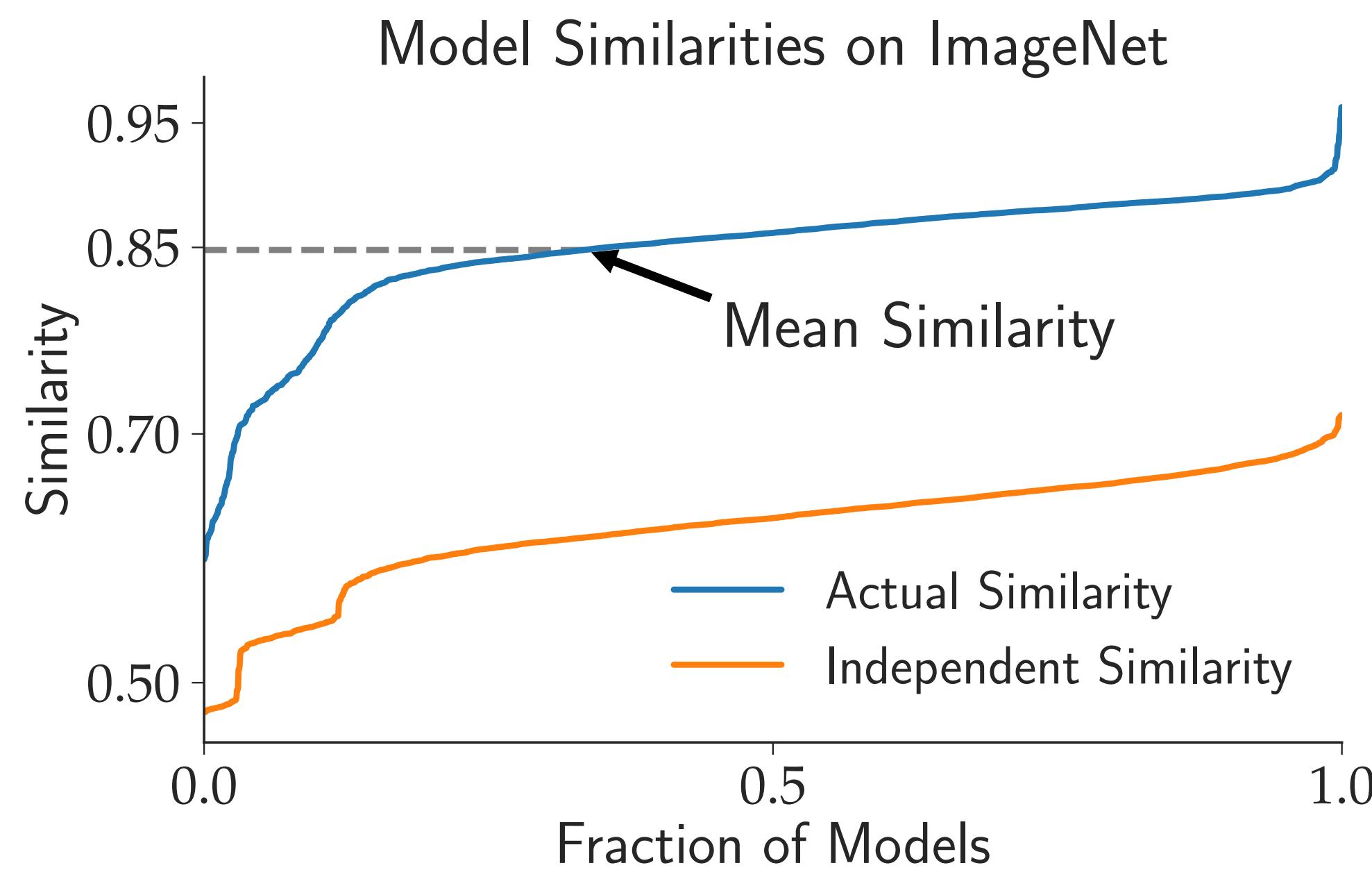


Why Does Test Set Re-use Not Lead to Overfitting?

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.

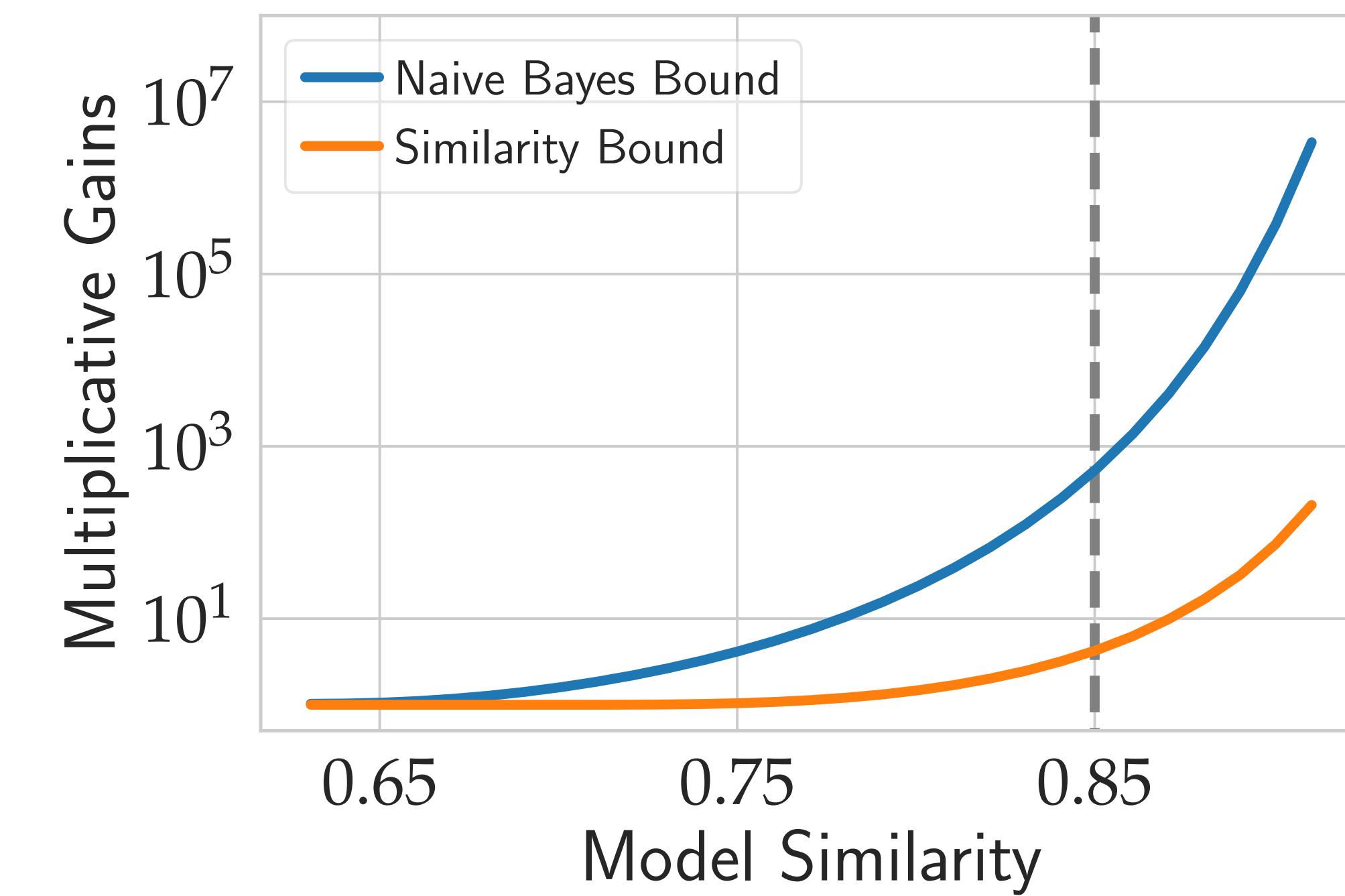
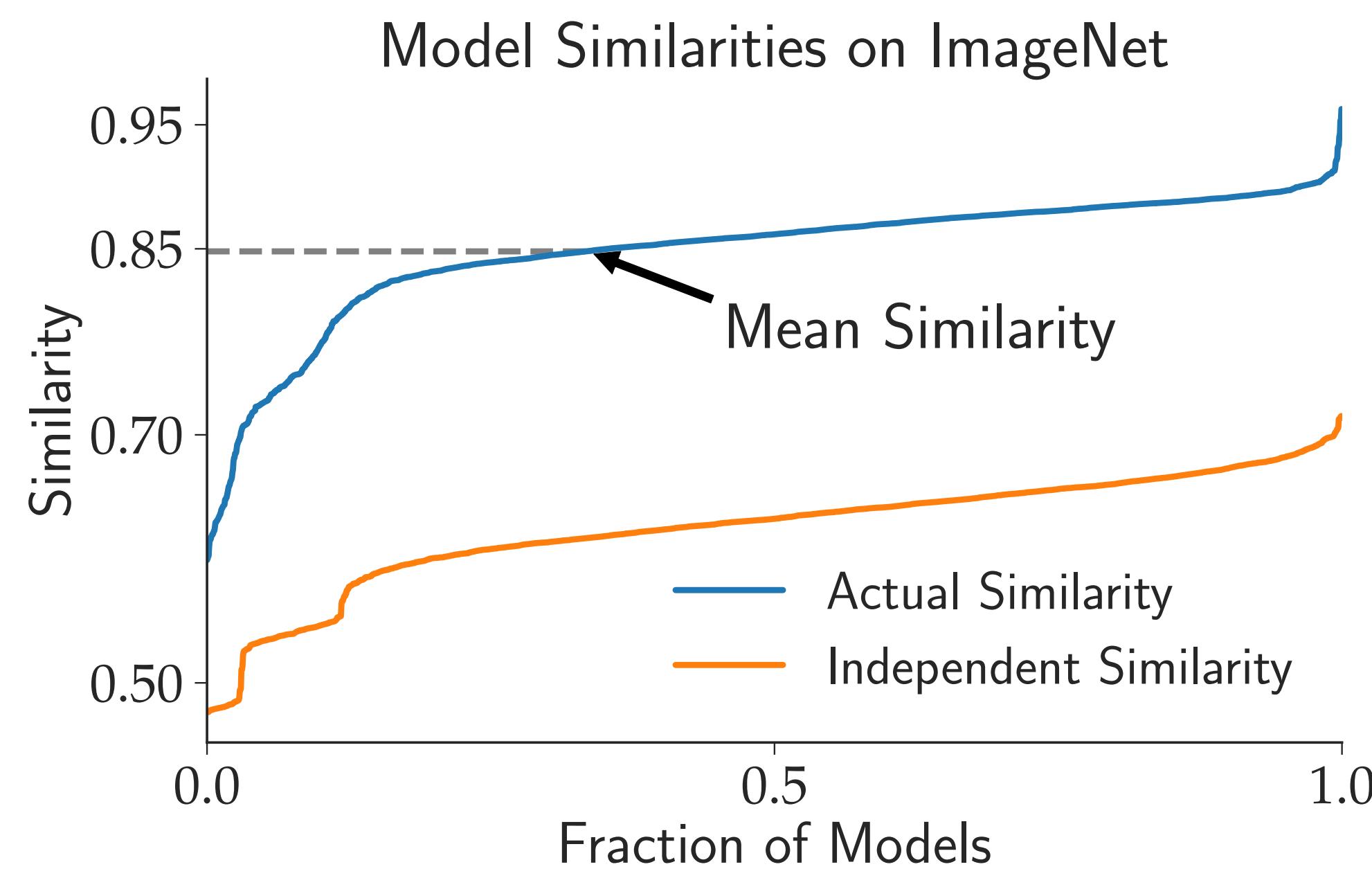


Why Does Test Set Re-use Not Lead to Overfitting?

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.

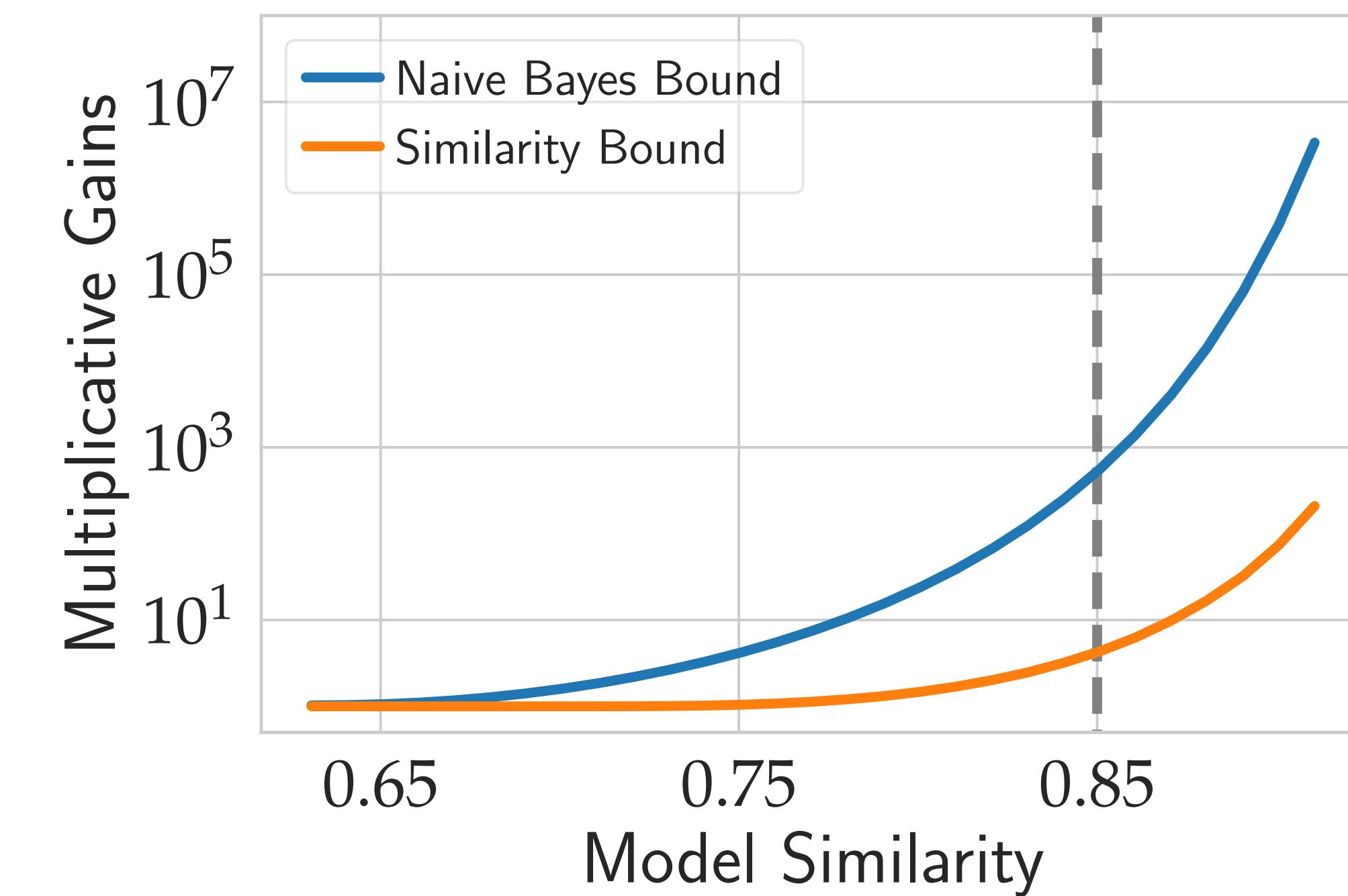
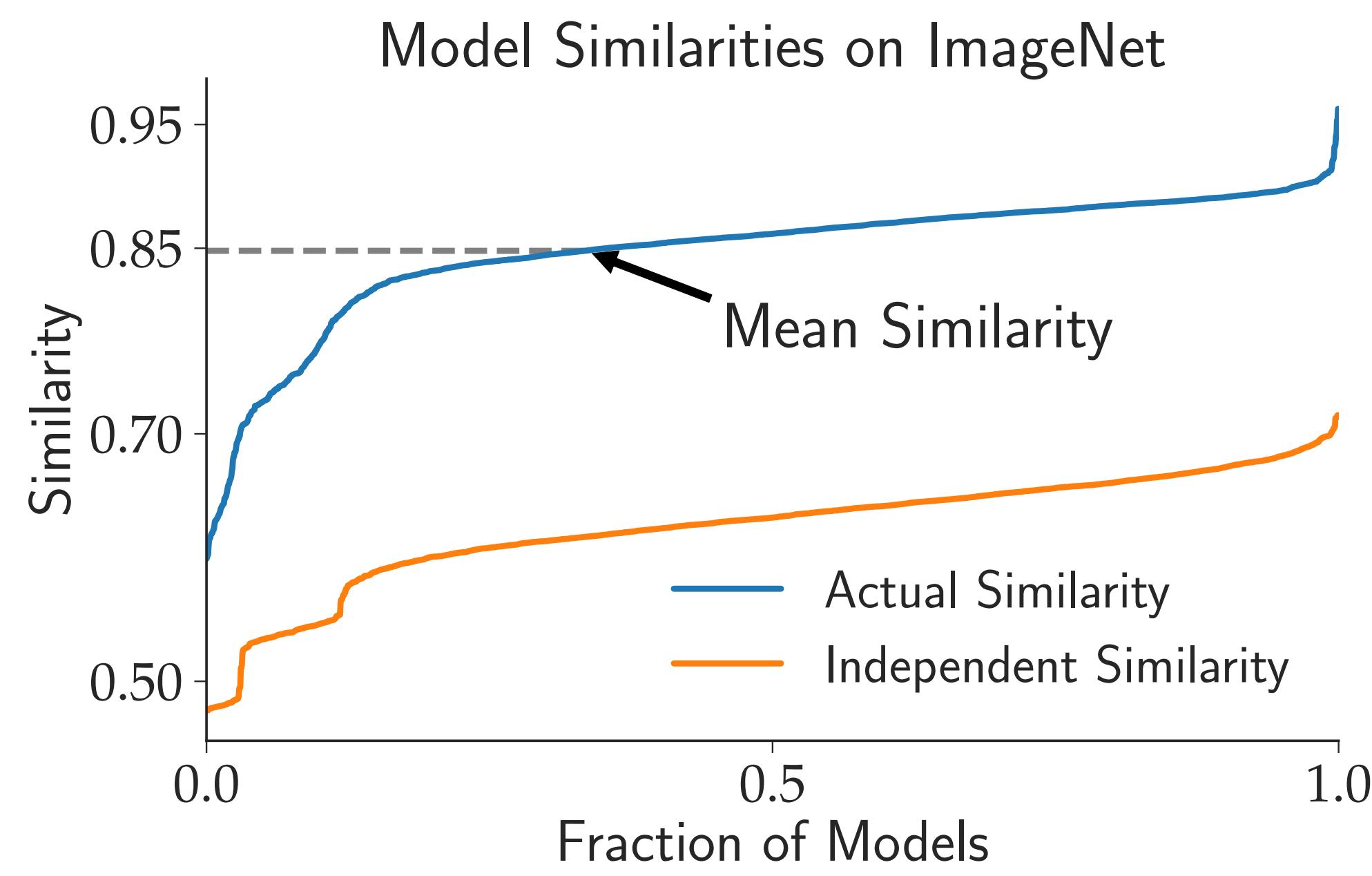


Why Does Test Set Re-use Not Lead to Overfitting?

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

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

Overfitting through test set re-use

Distribution shift

Generalization error ($\approx 1\%$)

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

$$\widehat{\text{acc}}_S(f) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}[f(x) = y]$$
$$\text{acc}_D(f) = \mathbb{E}_{(x,y) \sim D} \mathbb{1}[f(x) = y] \quad (\text{S is drawn from D})$$

Two Possible Causes

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

New test accuracy

Overfitting through test set re-use ($\approx 0\%$)

Distribution shift

Generalization error ($\approx 1\%$)

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

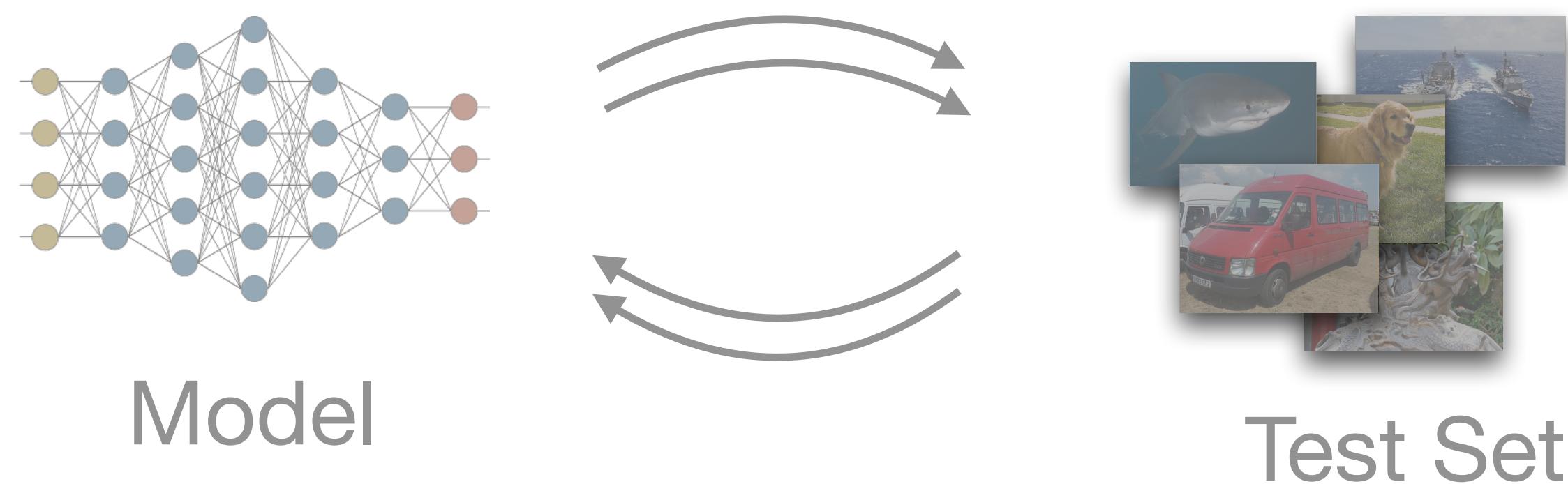
$\widehat{\text{acc}}_S(f) = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}[f(x) = y]$

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

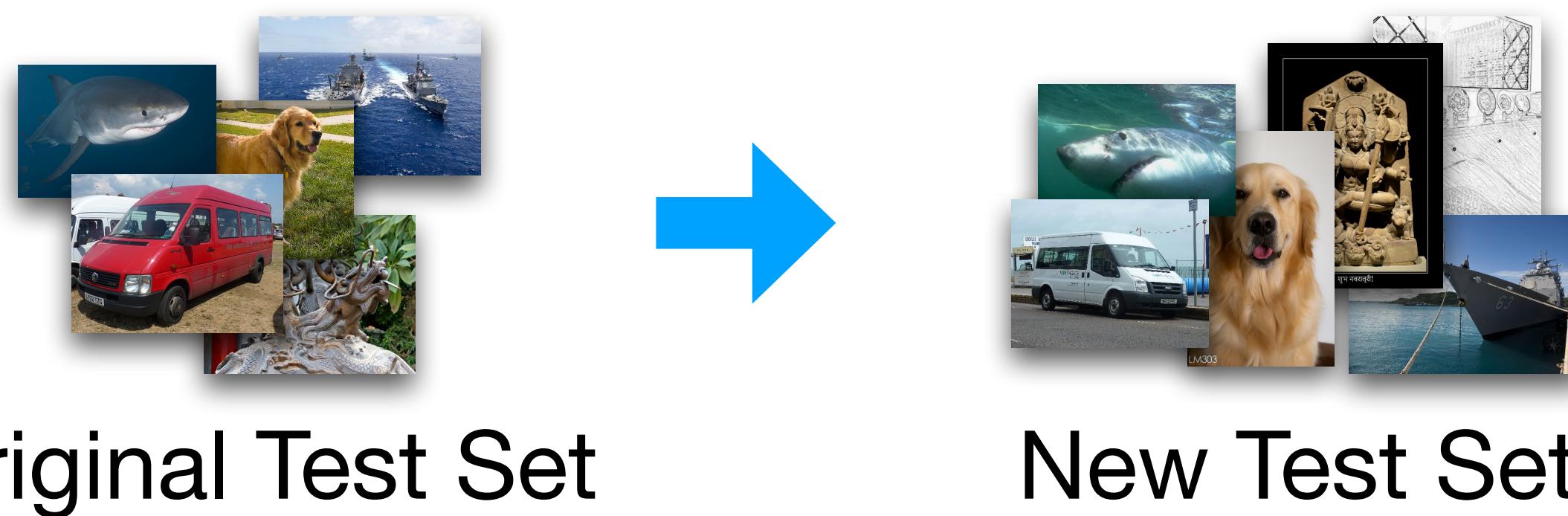
$\approx 11\%$

Three Forms of Overfitting

1. Test error \geq training error
2. Overfitting through test set re-use



3. Distribution shift



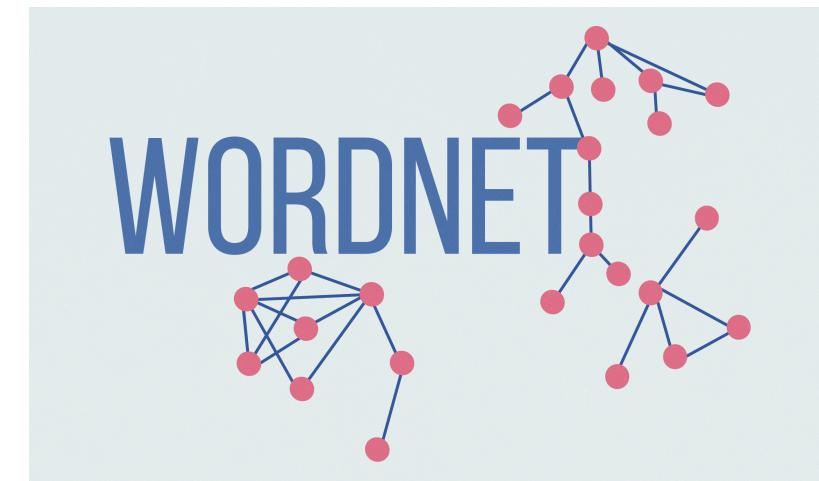
ImageNet Creation Process

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

ImageNet Creation Process

Detailed description in [\[Deng, Dong, Socher, Li, Li, Fei-Fei'09\]](#):

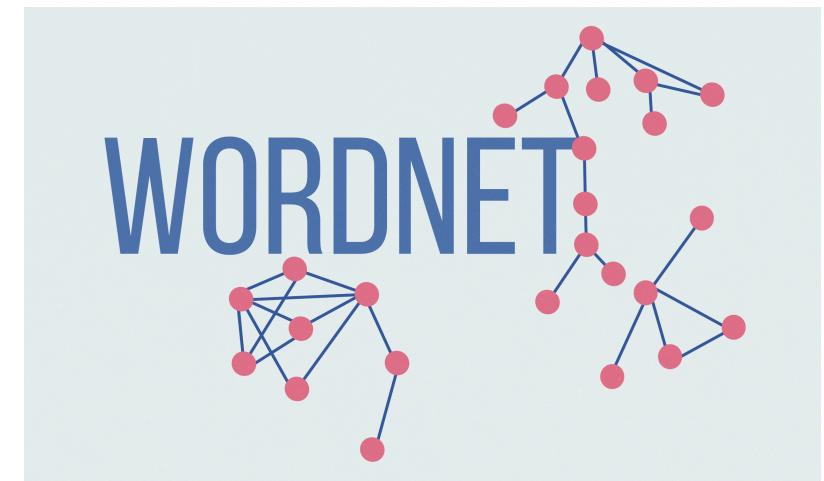
1. Find relevant search keywords for each class from **WordNet**
(e.g., “goldfish”, “Carassius auratus” for wnid “n01443537”)



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**

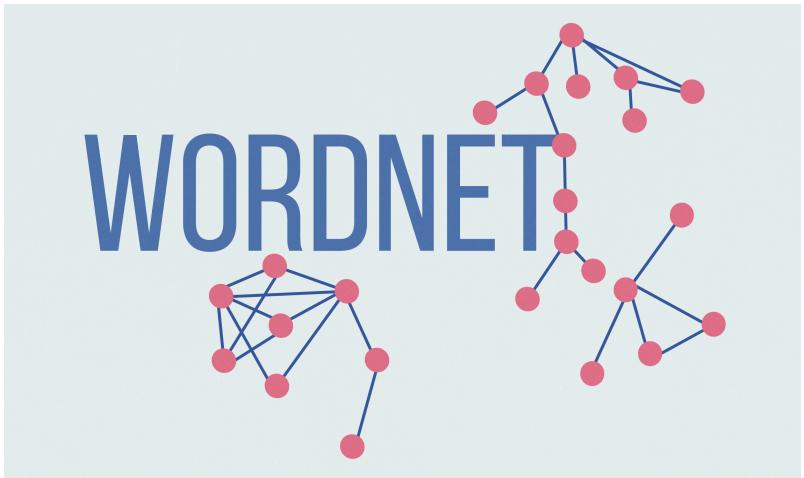


+ **flickr**

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



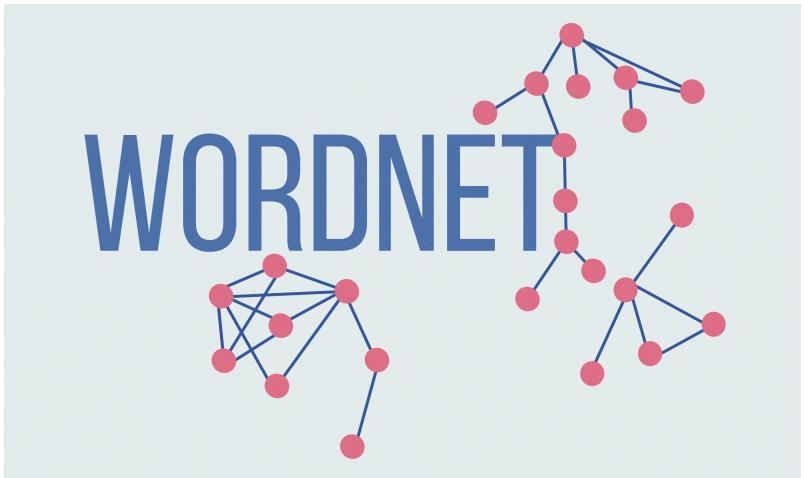
+ **flickr**

+ **amazon**
beta
mechanical turk

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
4. Sample a class-balanced dataset



+ **flickr**

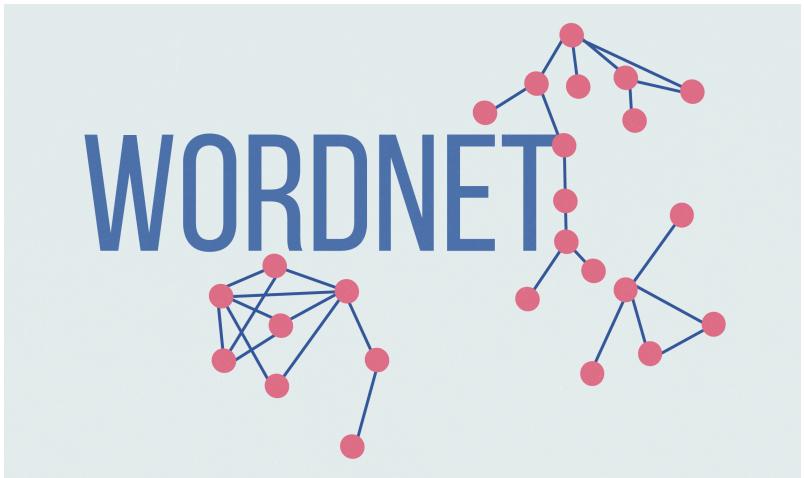
+ **amazon**
mechanical turk beta

IMAGENET

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
4. Sample a class-balanced dataset



+ **flickr**

+ **amazon**
mechanical turk beta

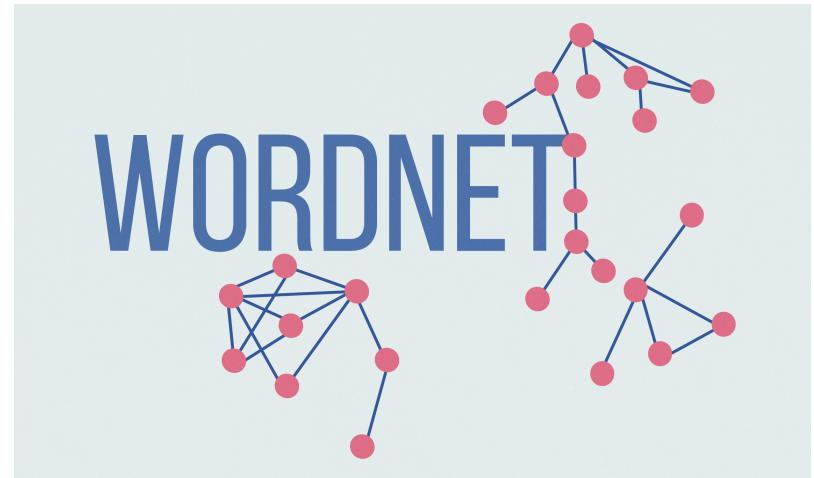
We replicated this process as closely as possible.

IMAGENET

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

+ **flickr**

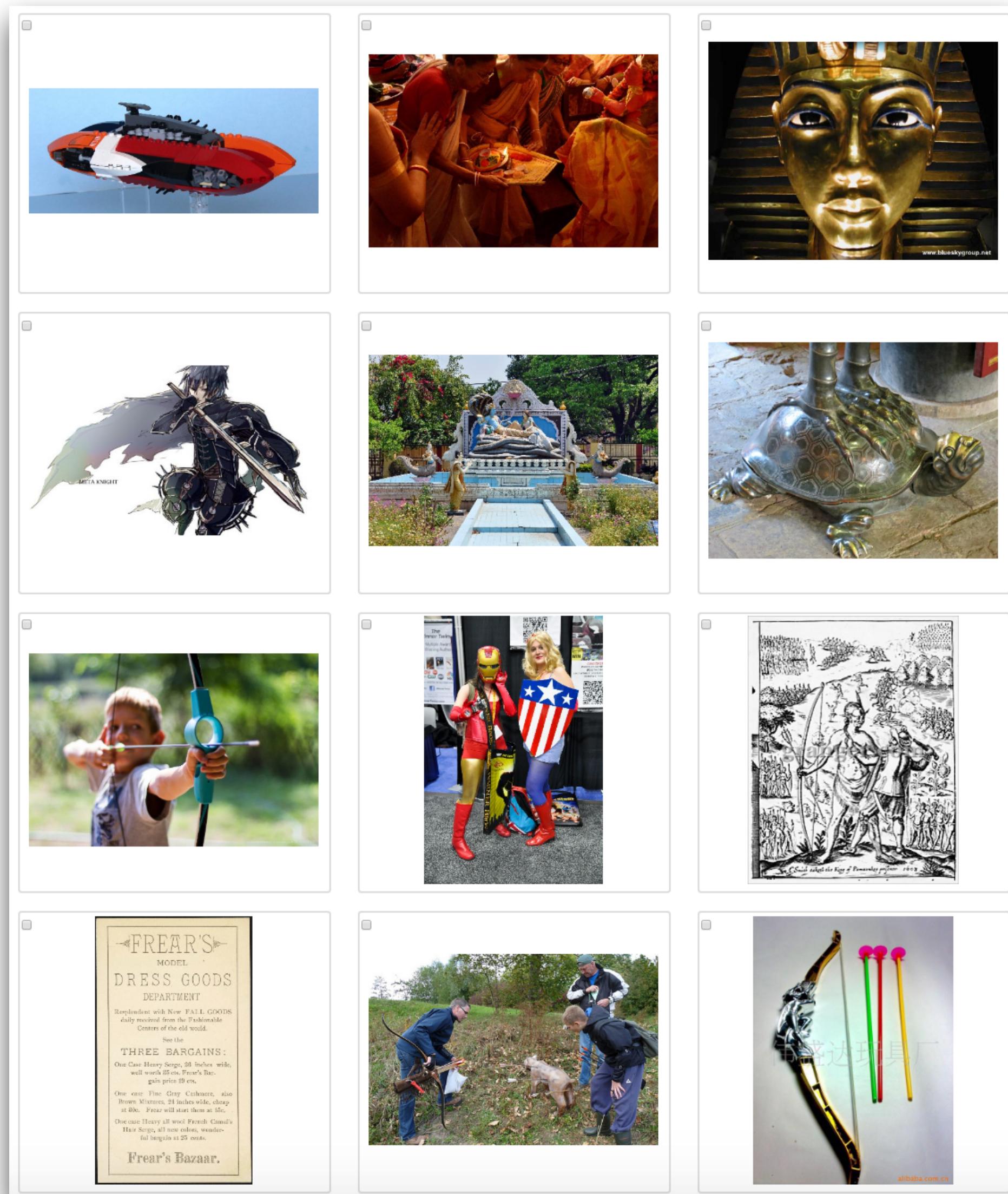
+ **amazon**
mechanical turk beta

We replicated this process as closely as possible.

IMAGENET

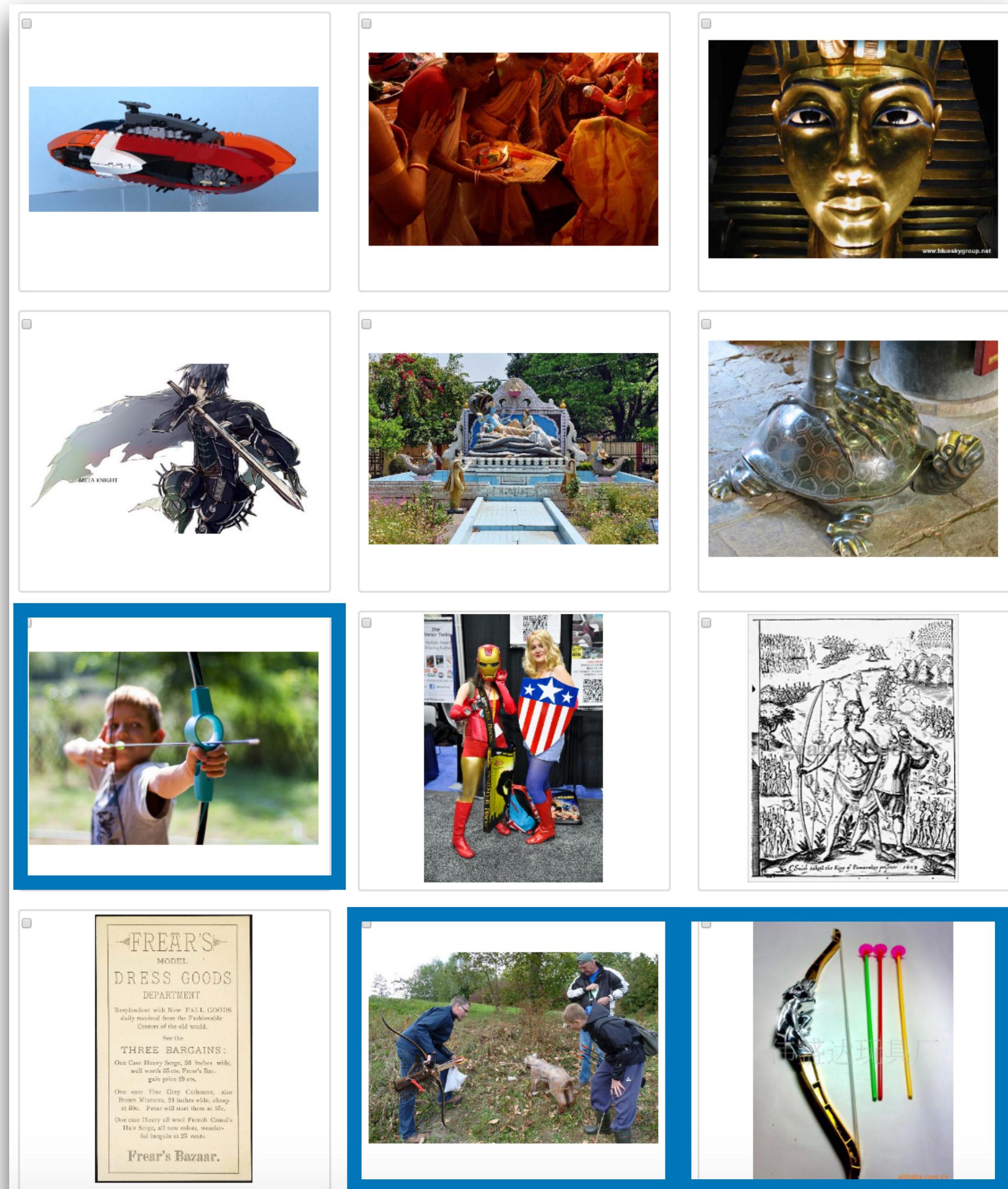
Data Cleaning With MTurk

Instructions: Select all images containing a bow.



Data Cleaning With MTurk

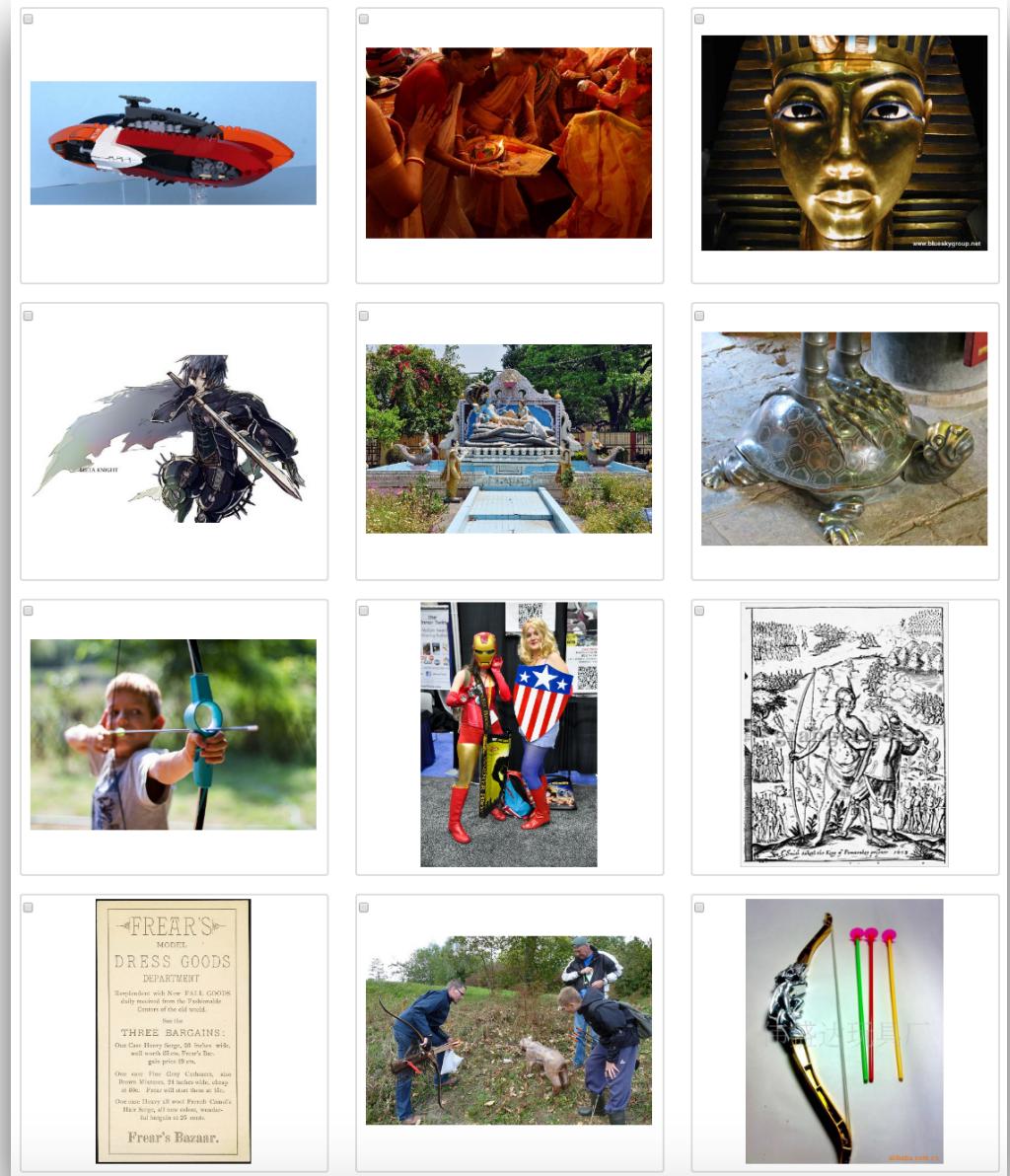
Instructions: Select all images containing a bow.



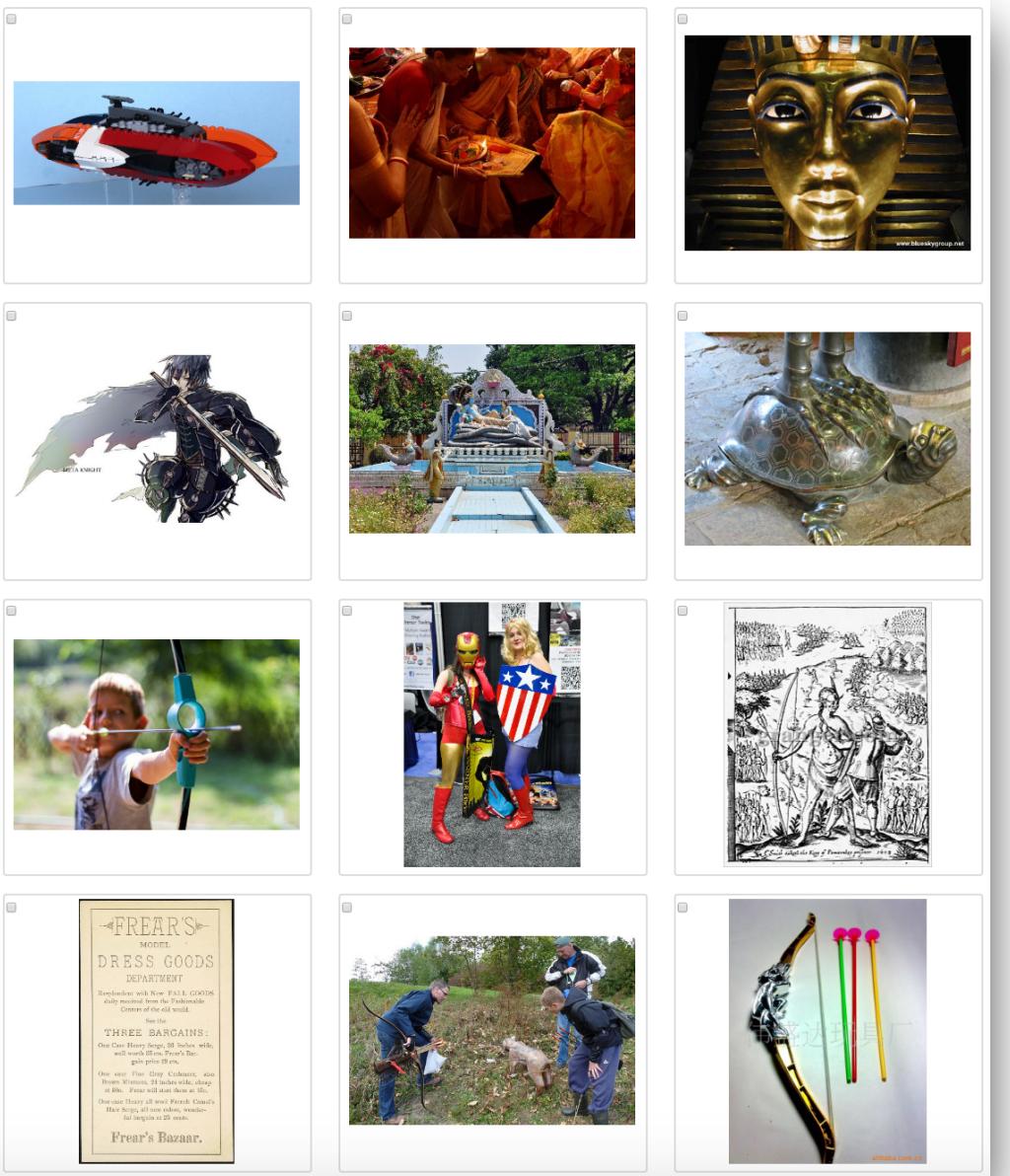
Data Cleaning With MTurk



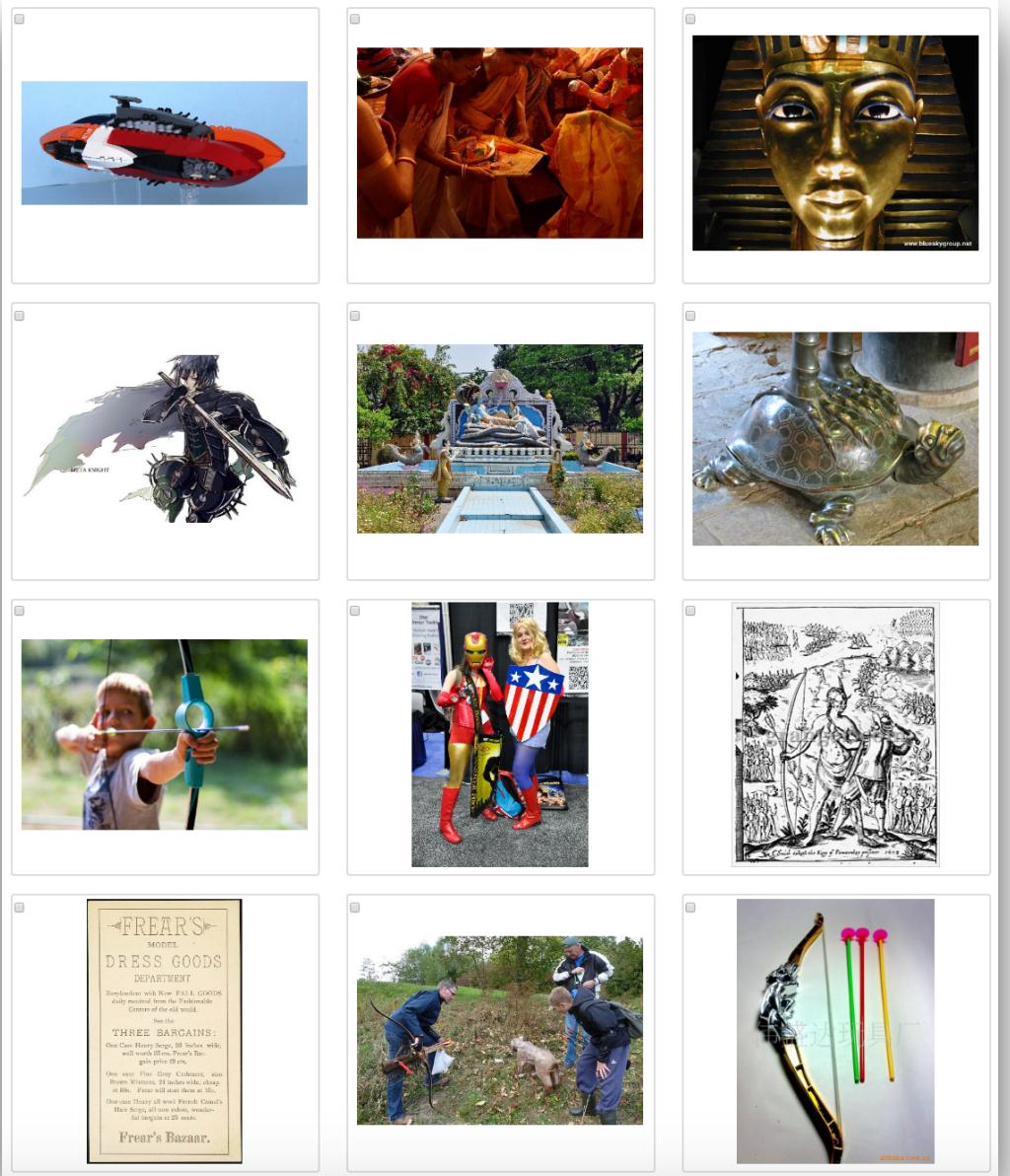
Worker 1



Worker 2



Worker 10

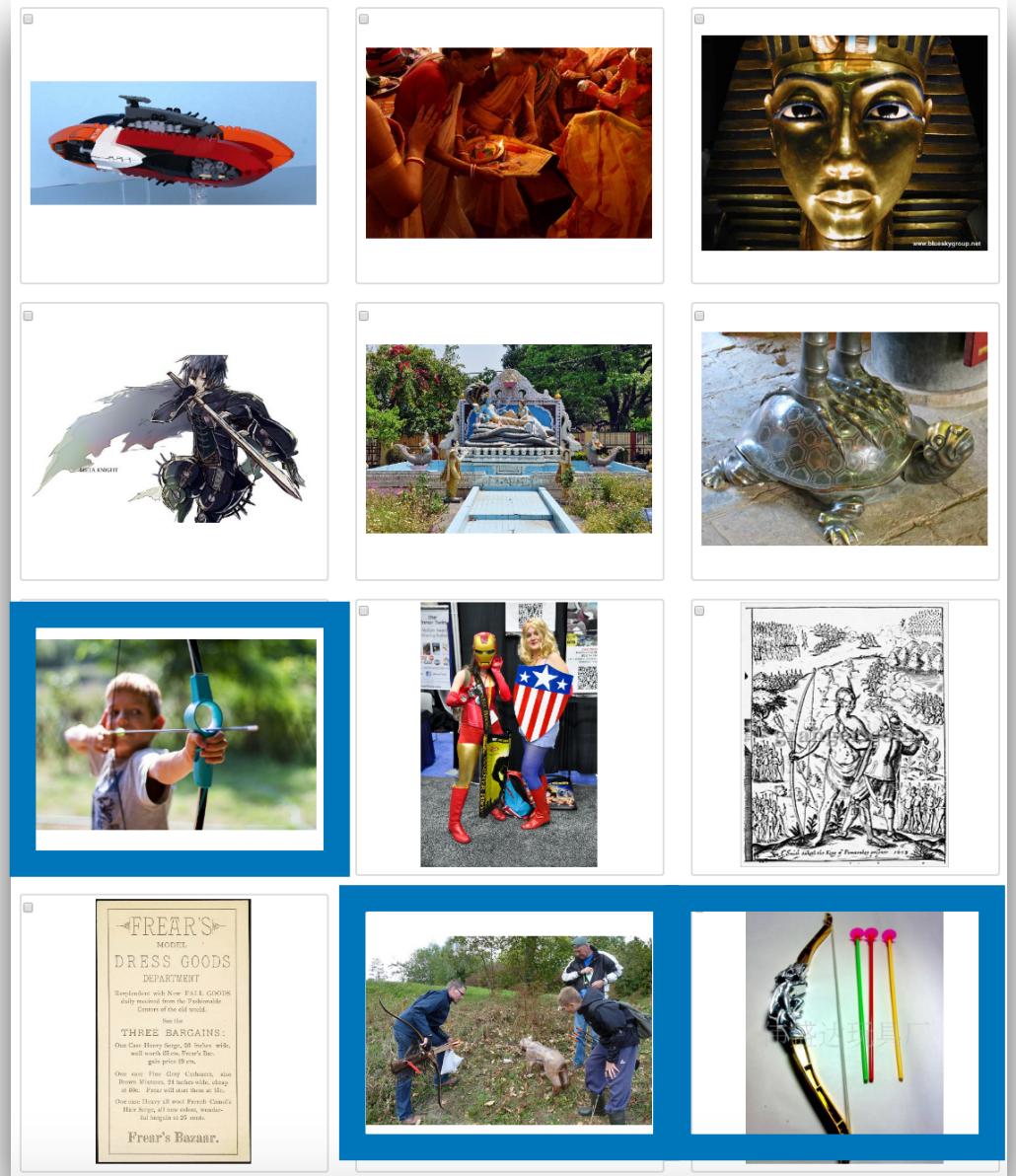


...

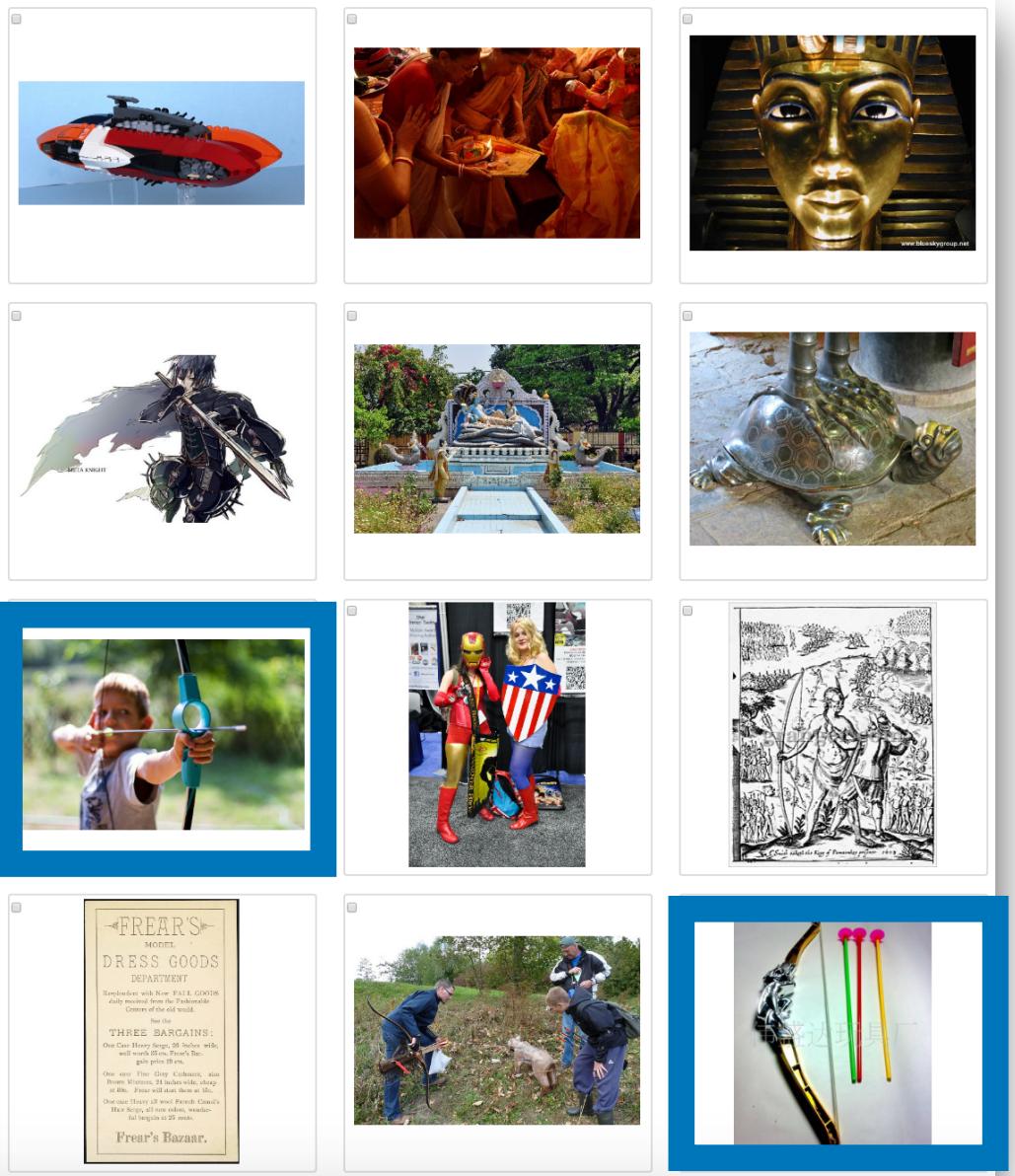
Data Cleaning With MTurk



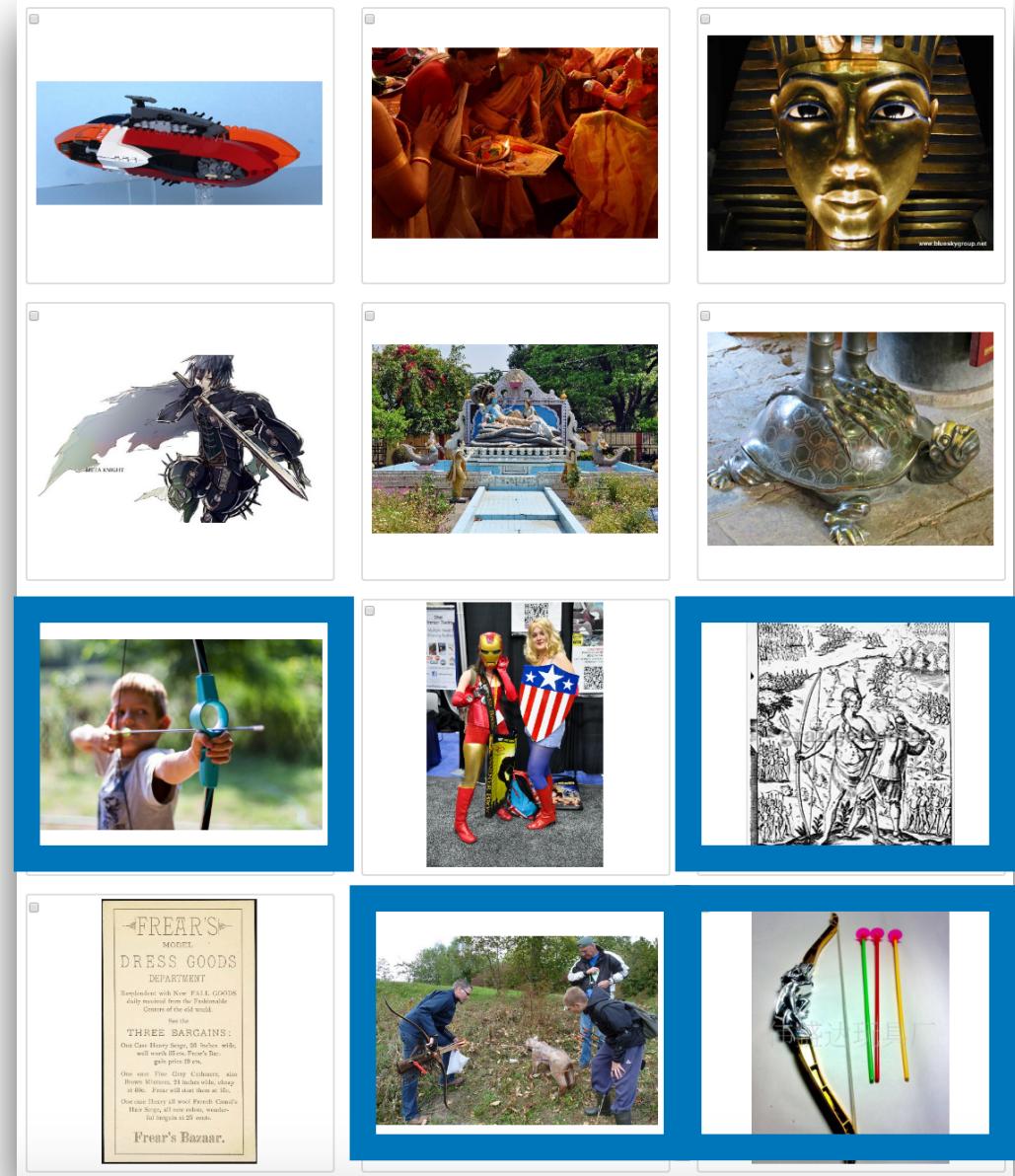
Worker 1



Worker 2



Worker 10

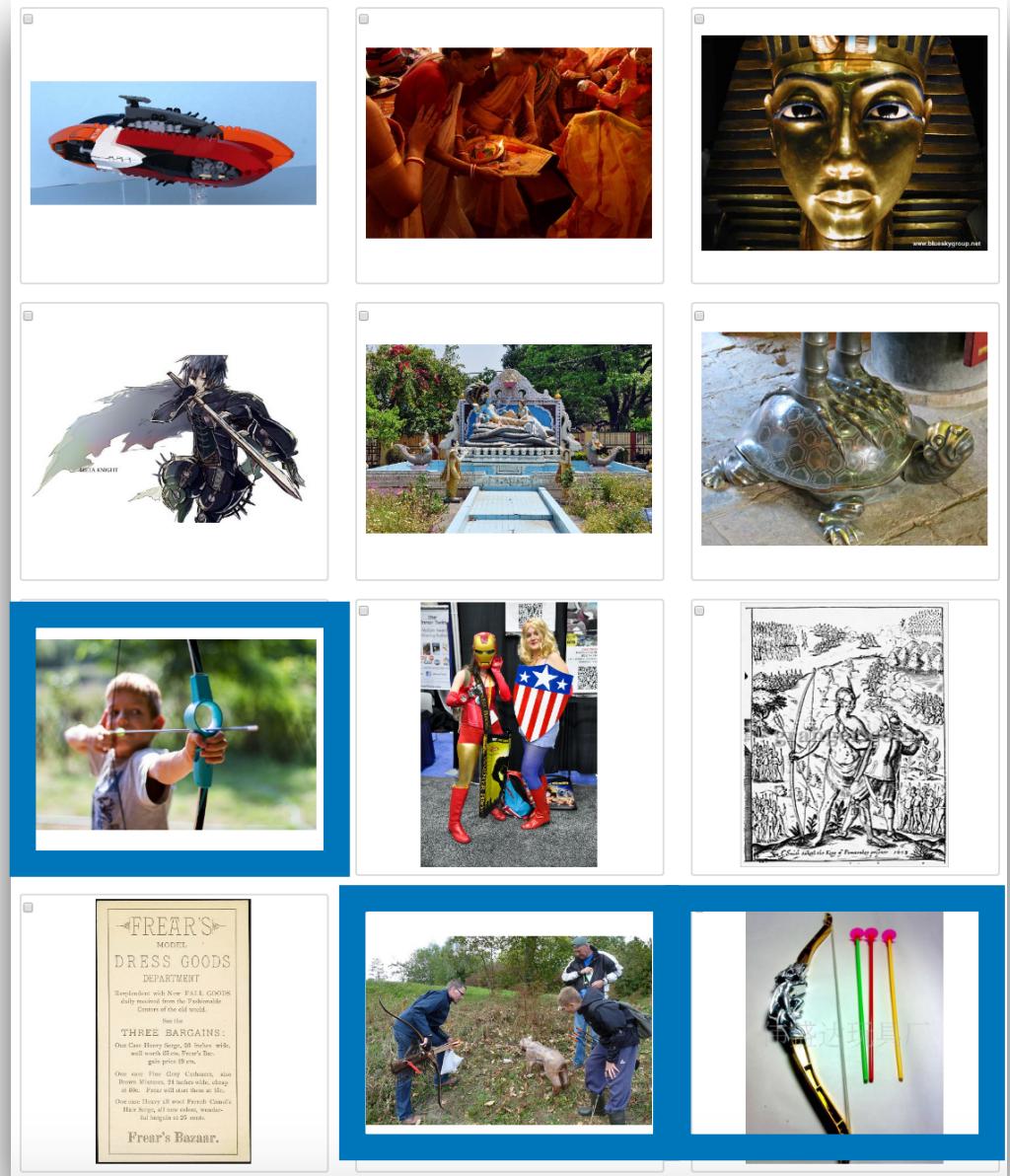


...

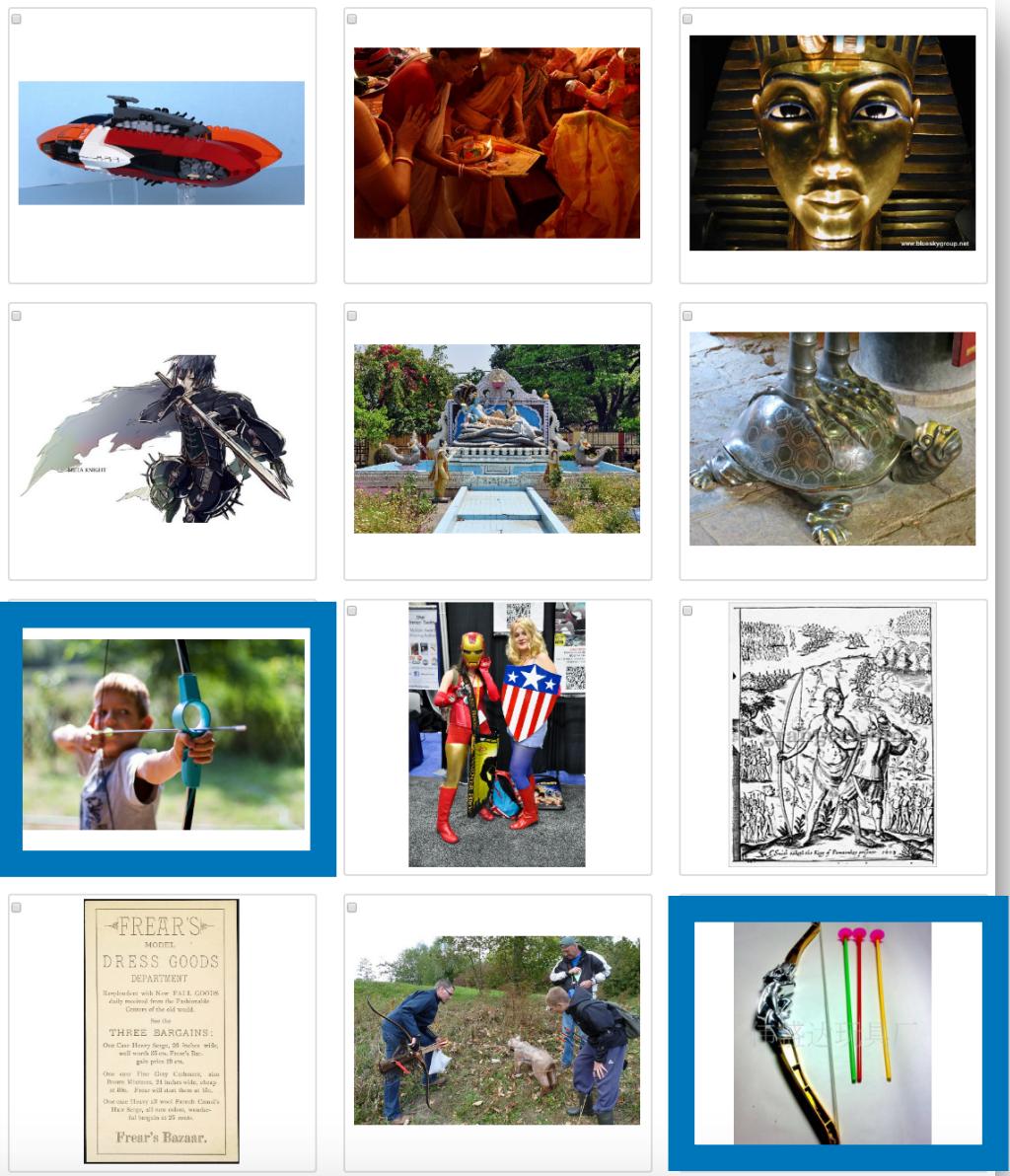
Data Cleaning With MTurk



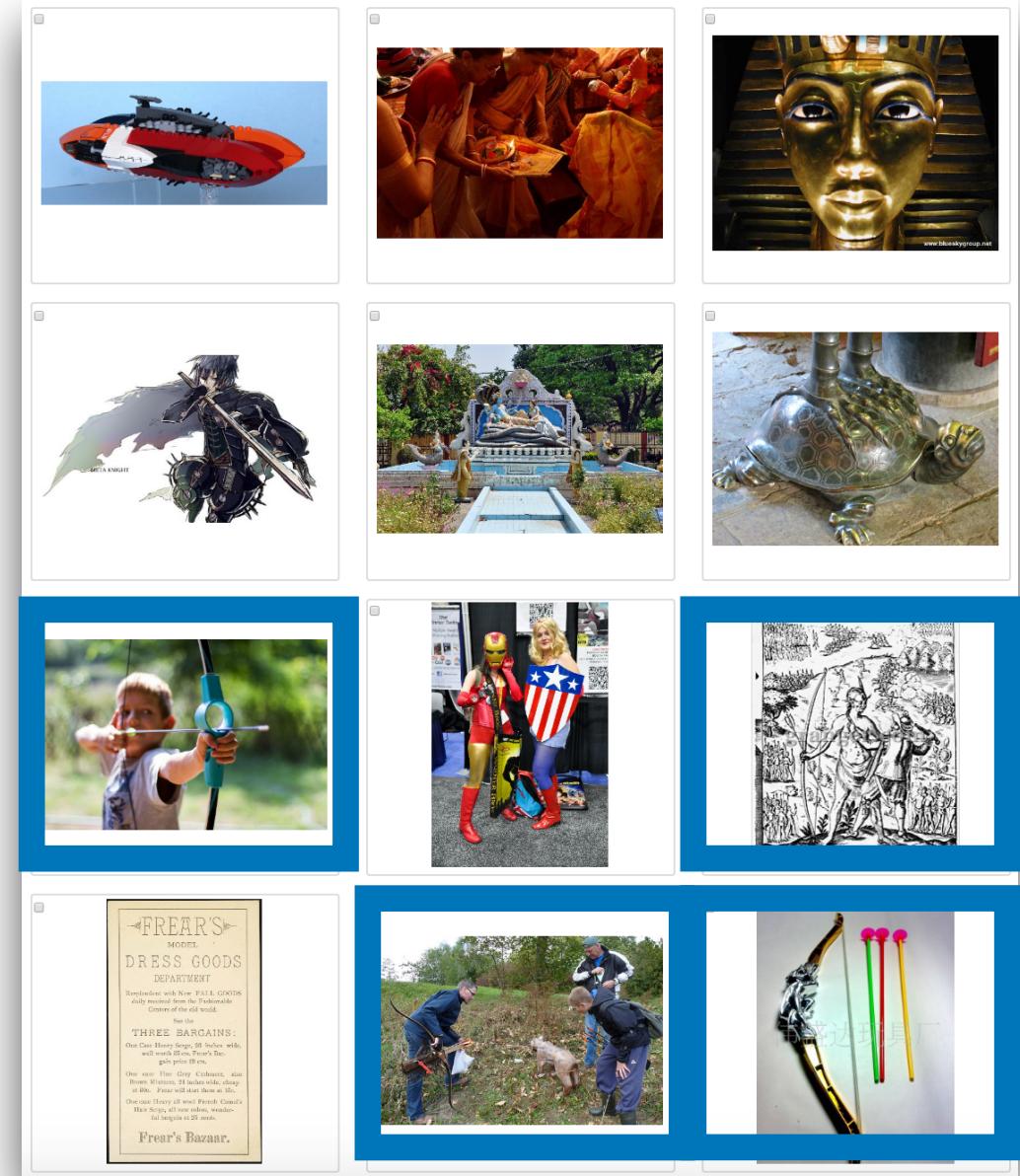
Worker 1



Worker 2



Worker 10



...

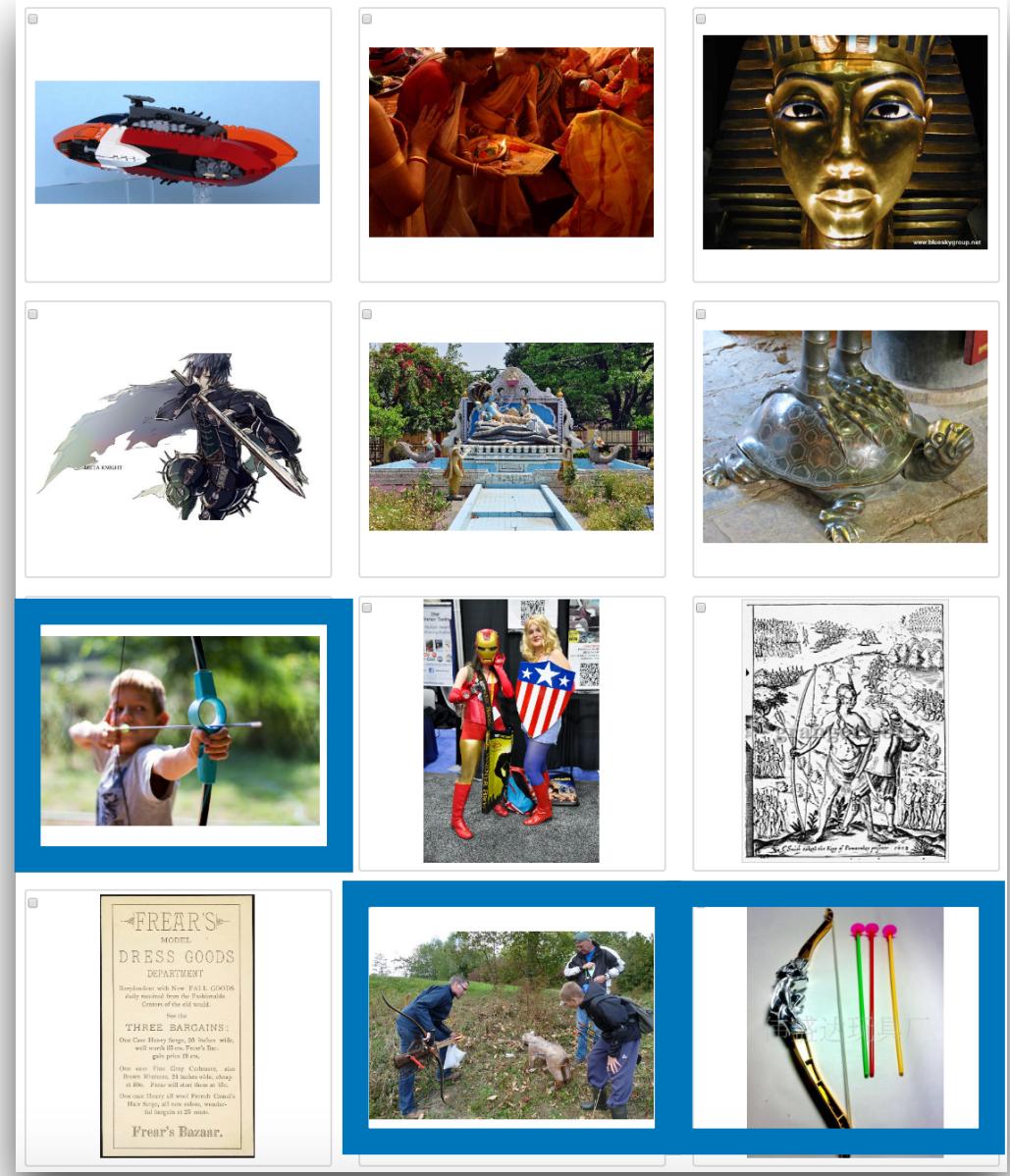
Main quantity: **selection frequency** =

$$\frac{\text{Number of workers who selected image } i}{\text{Number of workers who saw image } i}$$

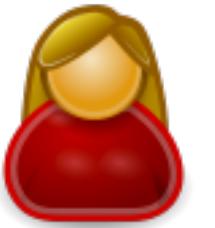
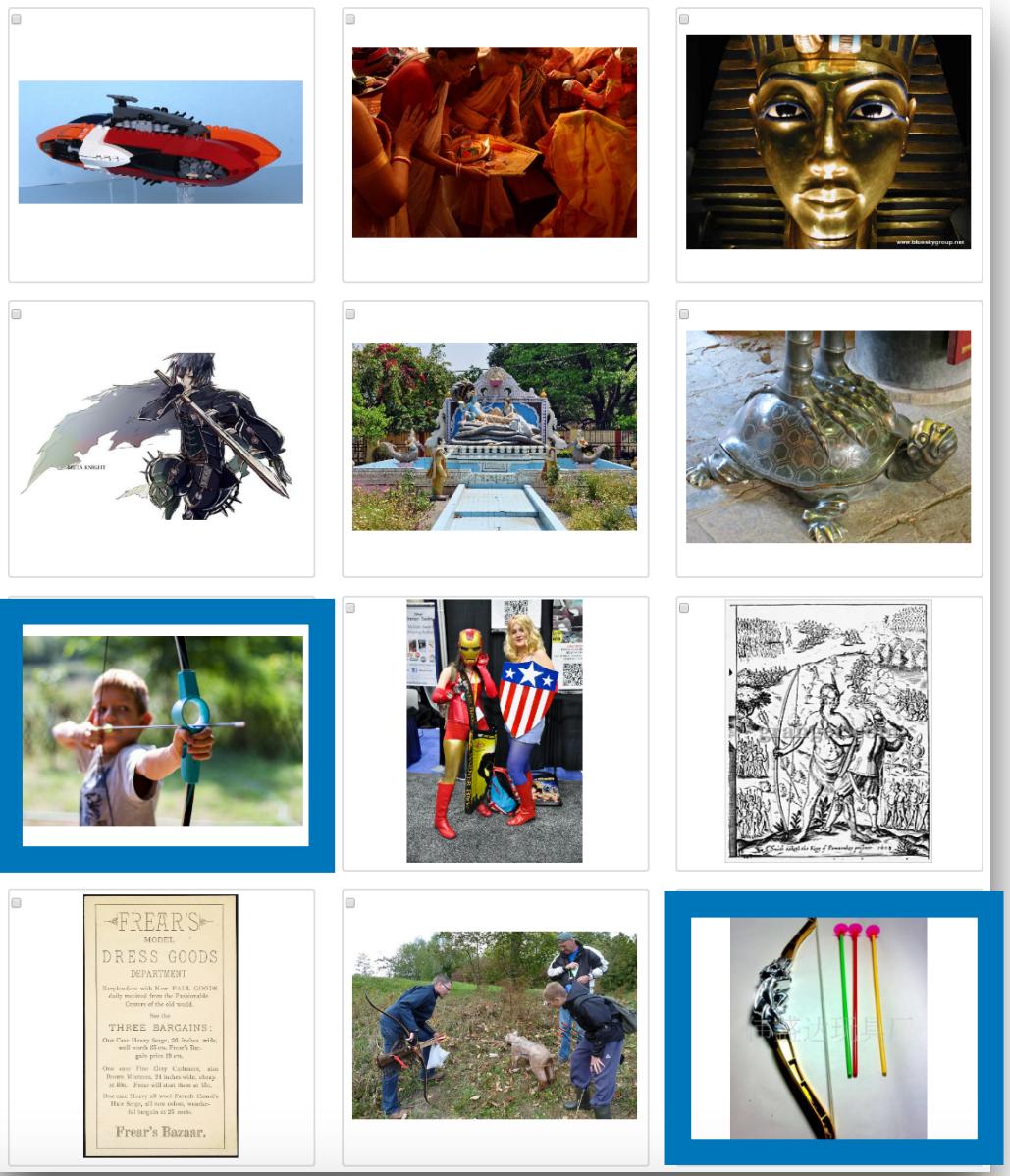
Data Cleaning With MTurk



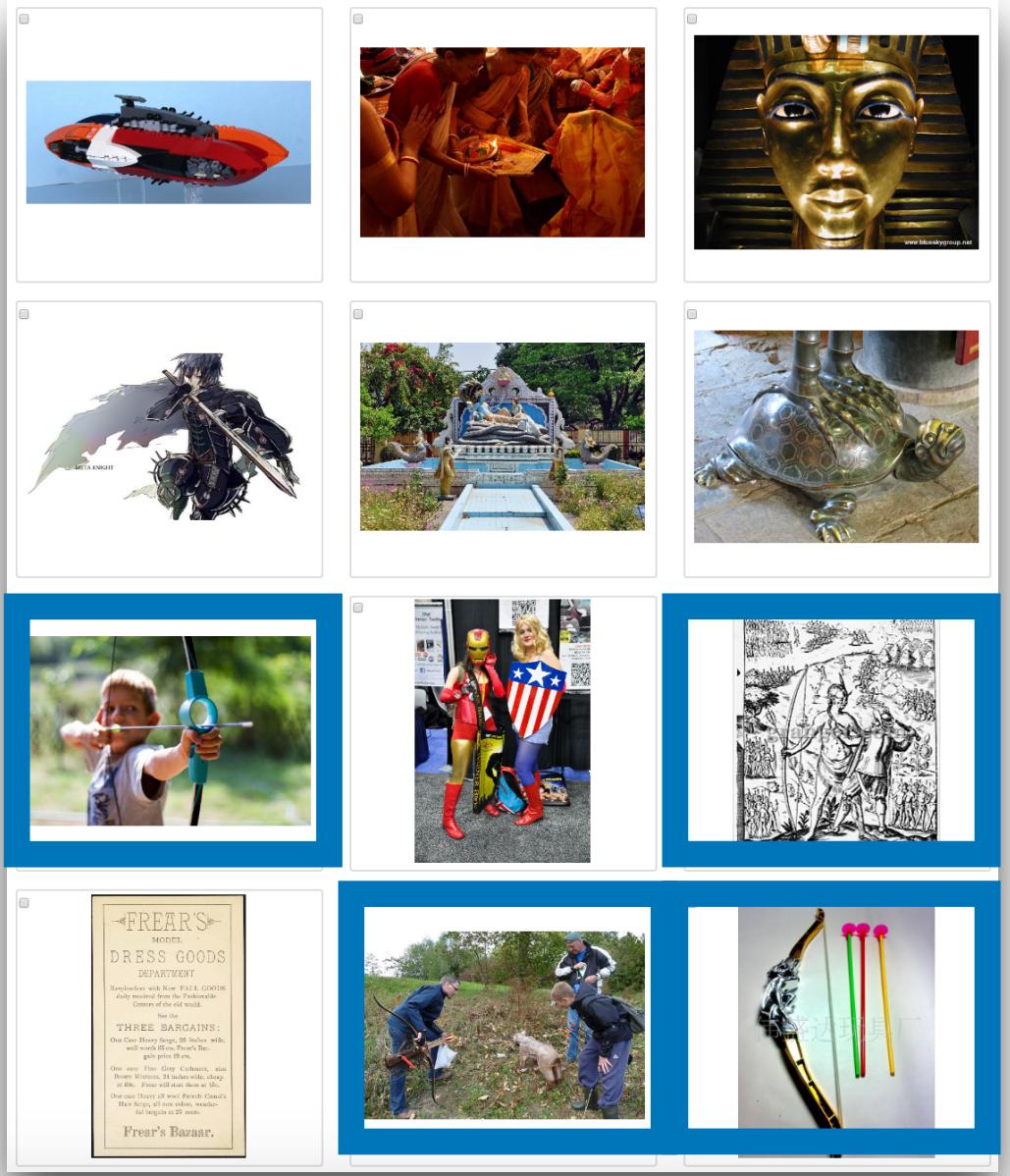
Worker 1



Worker 2



Worker 10



...

Main quantity: **selection frequency**

$$= \frac{\text{Number of workers who selected image } i}{\text{Number of workers who saw image } i}$$



: 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

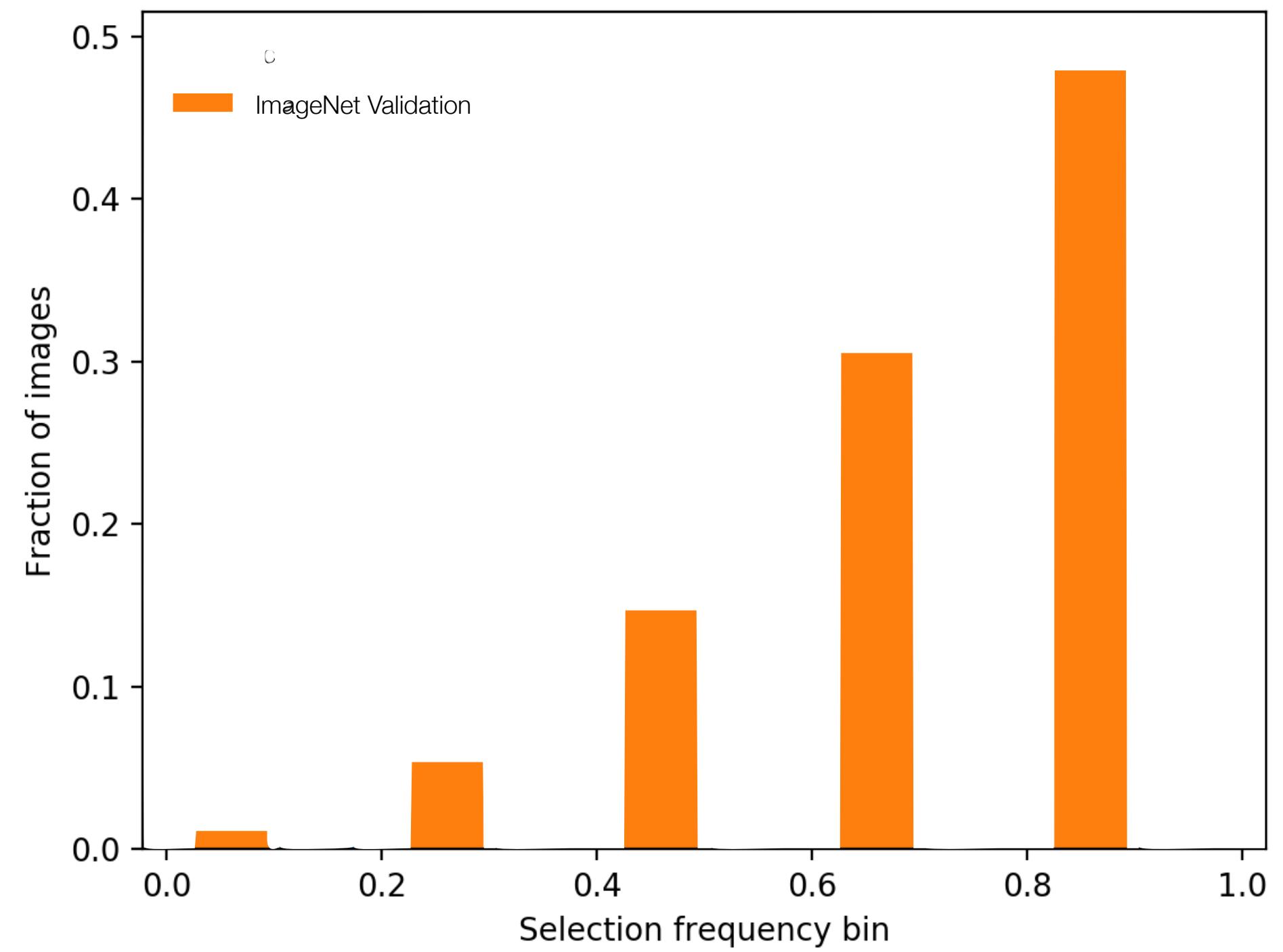
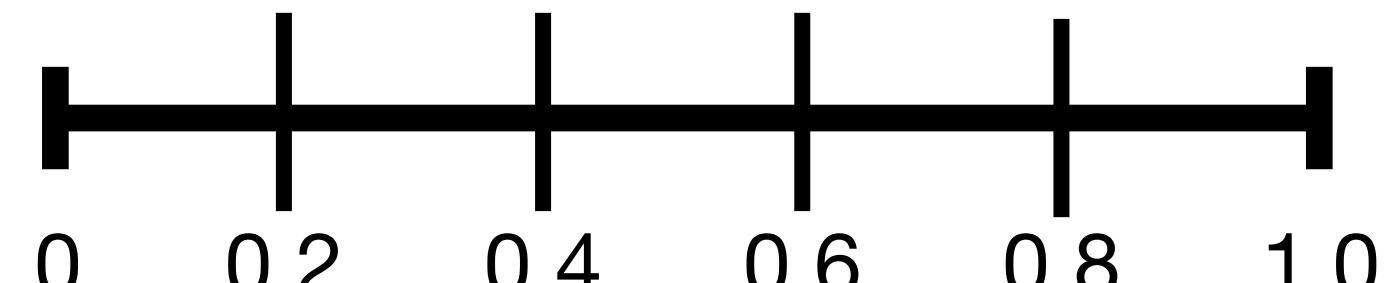
Sampling Strategy for a New Test Set

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.



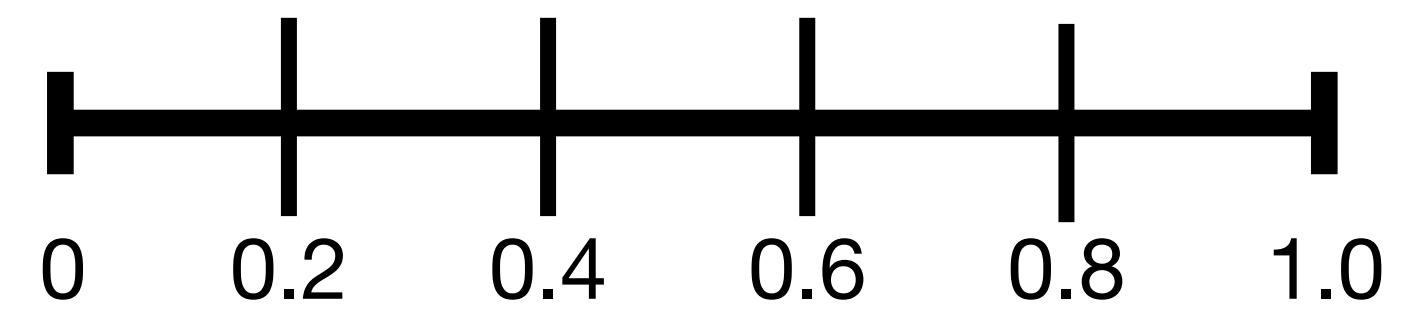
Sampling Strategy for a New Test Set

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

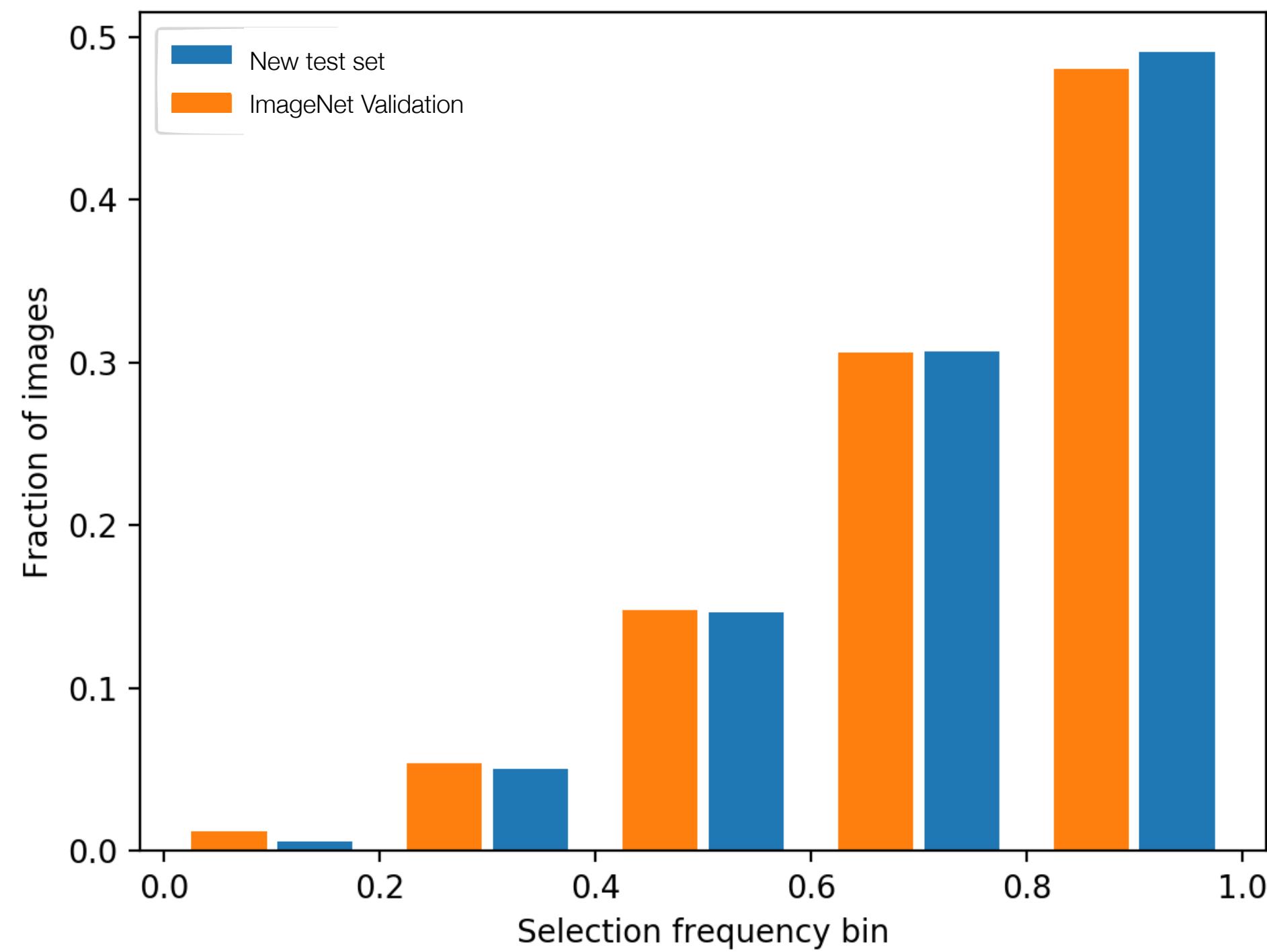
Output: representative & correct subset

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**).

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

Three New Test Sets

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

Easier: Different sampling strategy, higher selection frequencies.

Test Set	Average MTurk Selection Frequency	Average Top-1 Accuracy Change
ApproxCalibrated	0.73	- 12%
Easier	0.85	- 3%

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.

Test Set	Average MTurk Selection Frequency	Average Top-1 Accuracy Change
ApproxCalibrated	0.73	- 12%
Easier	0.85	- 3%
Easiest	0.93	+ 2%

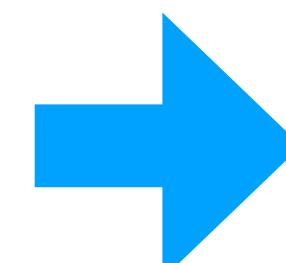
Three New Test Sets

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.

Test Set	Average MTurk Selection Frequency	Average Top-1 Accuracy Change
ApproxCalibrated	0.73	- 12%
Easier	0.85	- 3%
Easiest	0.93	+ 2%



Selection frequencies have large impact on classification accuracies.

Three New Test Sets

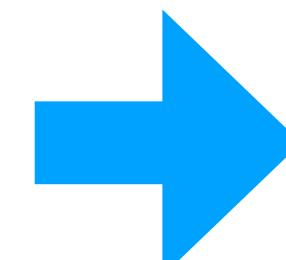
ApproxCalibrated: Selection frequencies comparable to the original test set (**0.71**).

Easier: Different sampling strategy, higher selection frequencies.

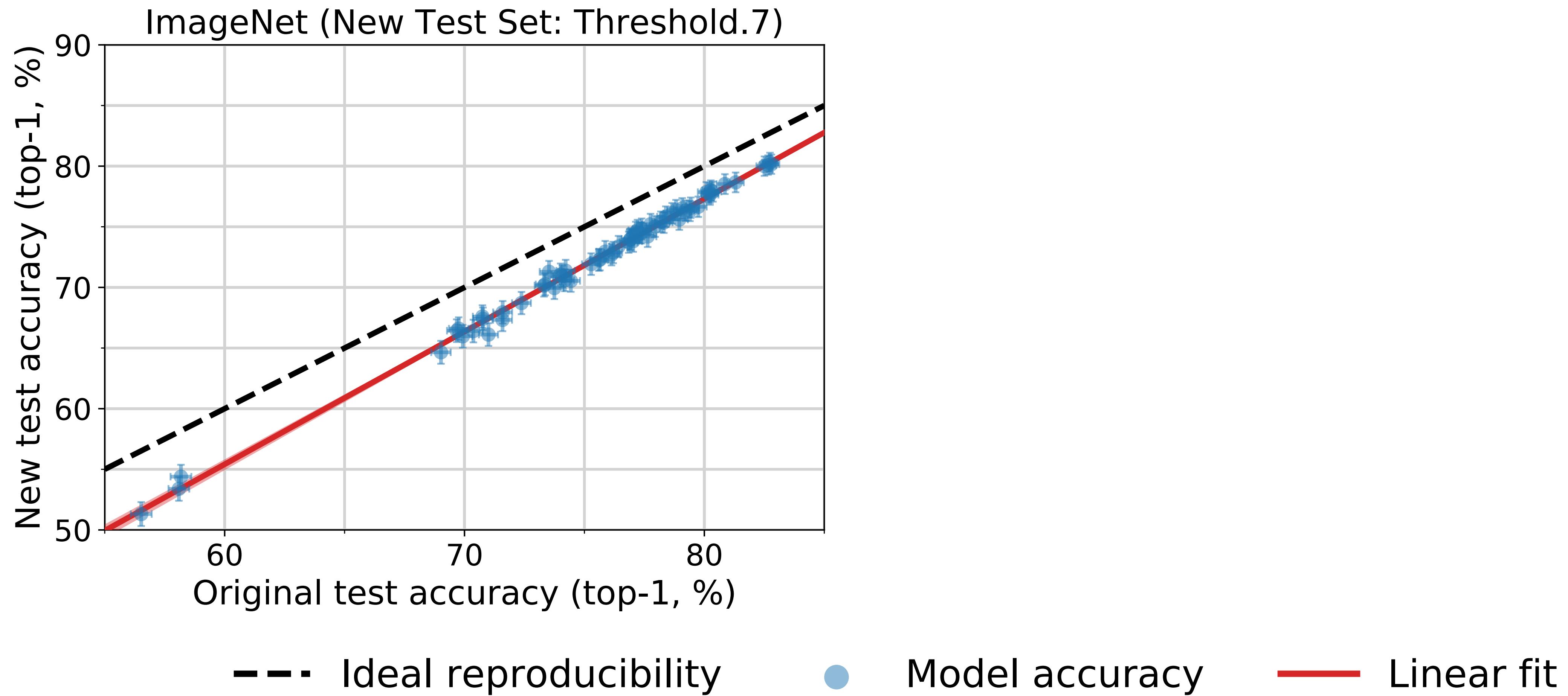
Easiest: Highest selection frequencies in our candidate pool.

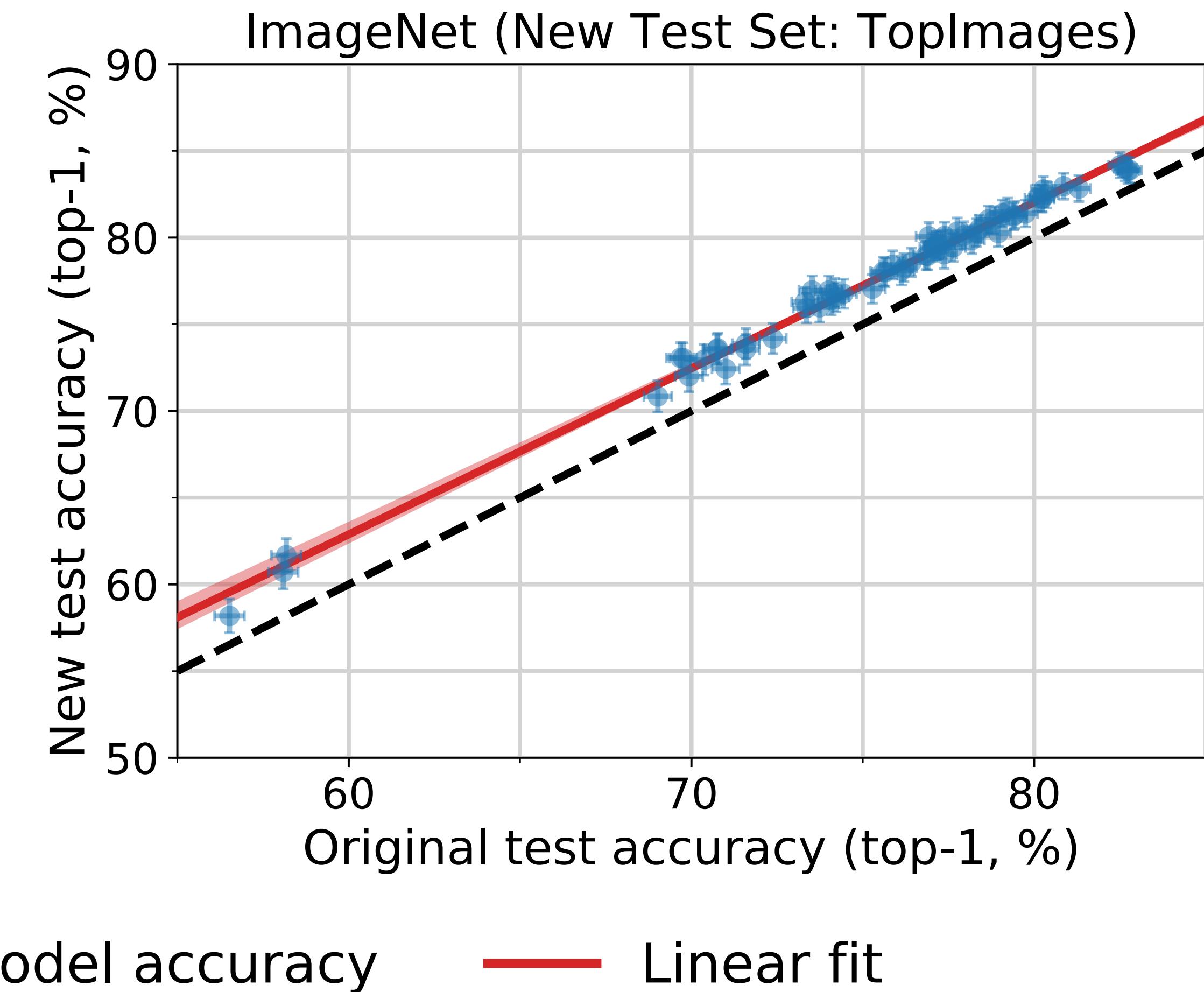
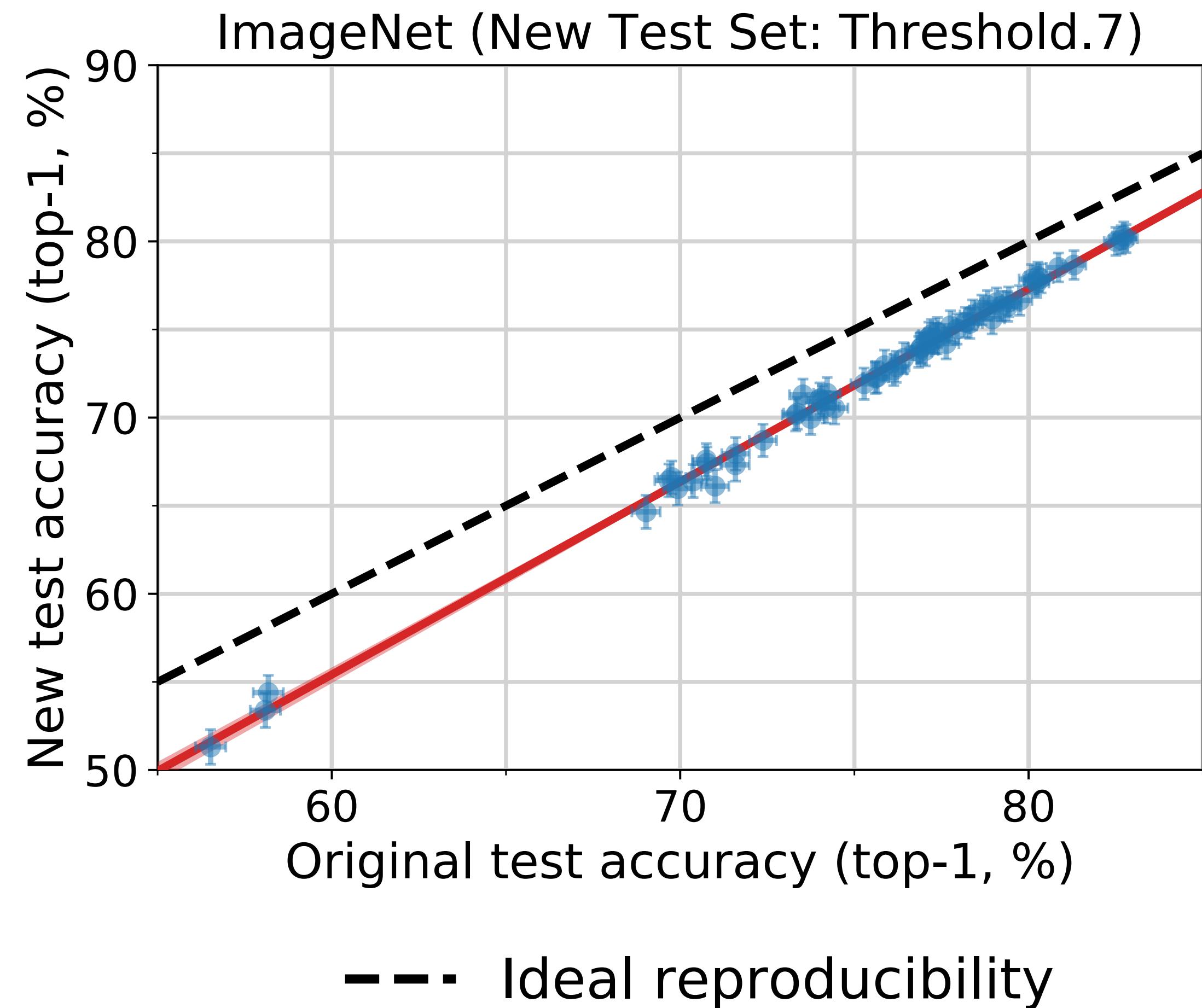
All correctly labeled!

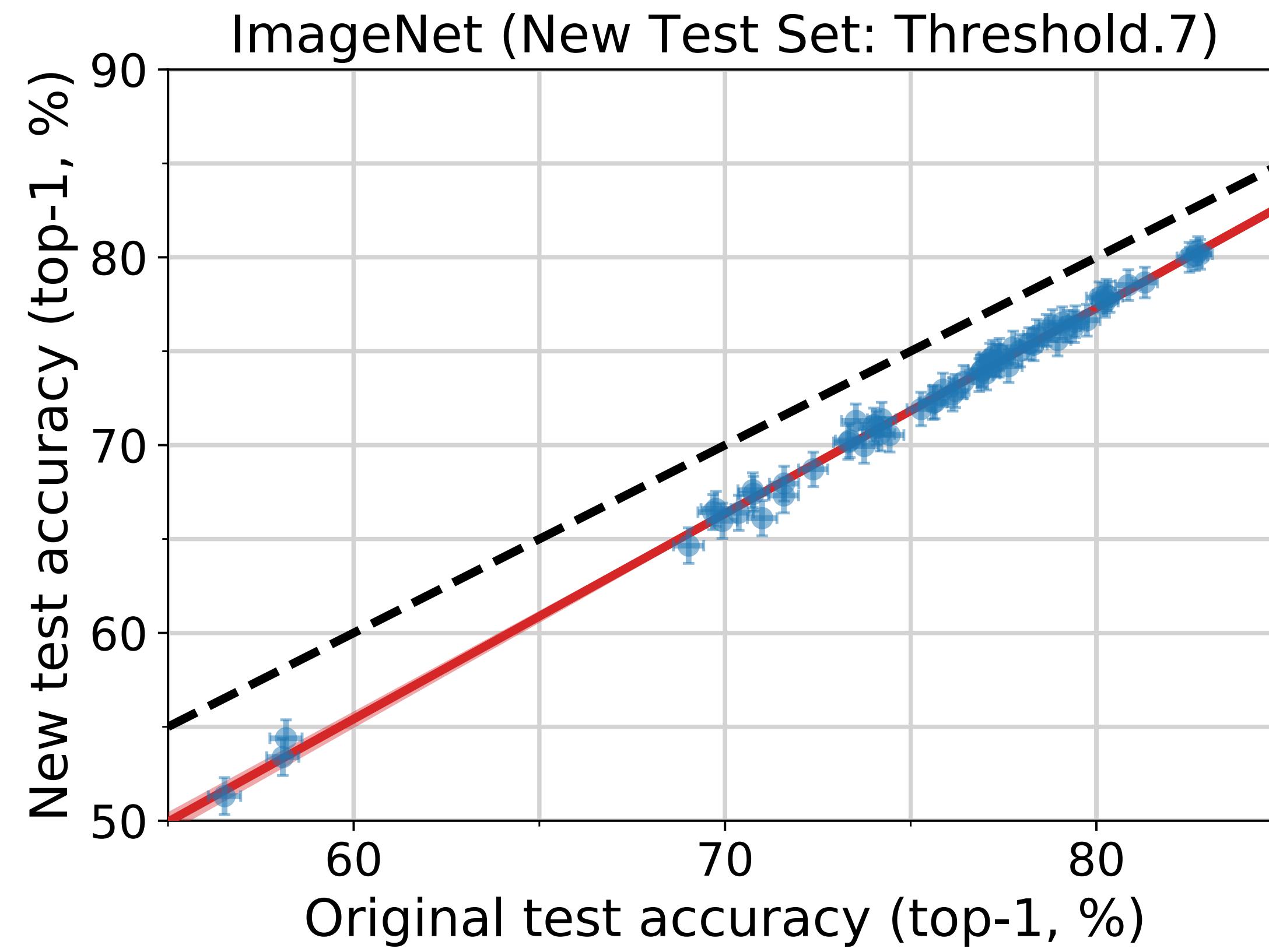
Test Set	Average MTurk Selection Frequency	Average Top-1 Accuracy Change
ApproxCalibrated	0.73	- 12%
Easier	0.85	- 3%
Easiest	0.93	+ 2%



Selection frequencies have large impact on classification accuracies.



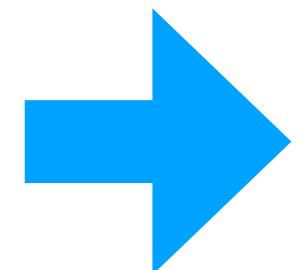
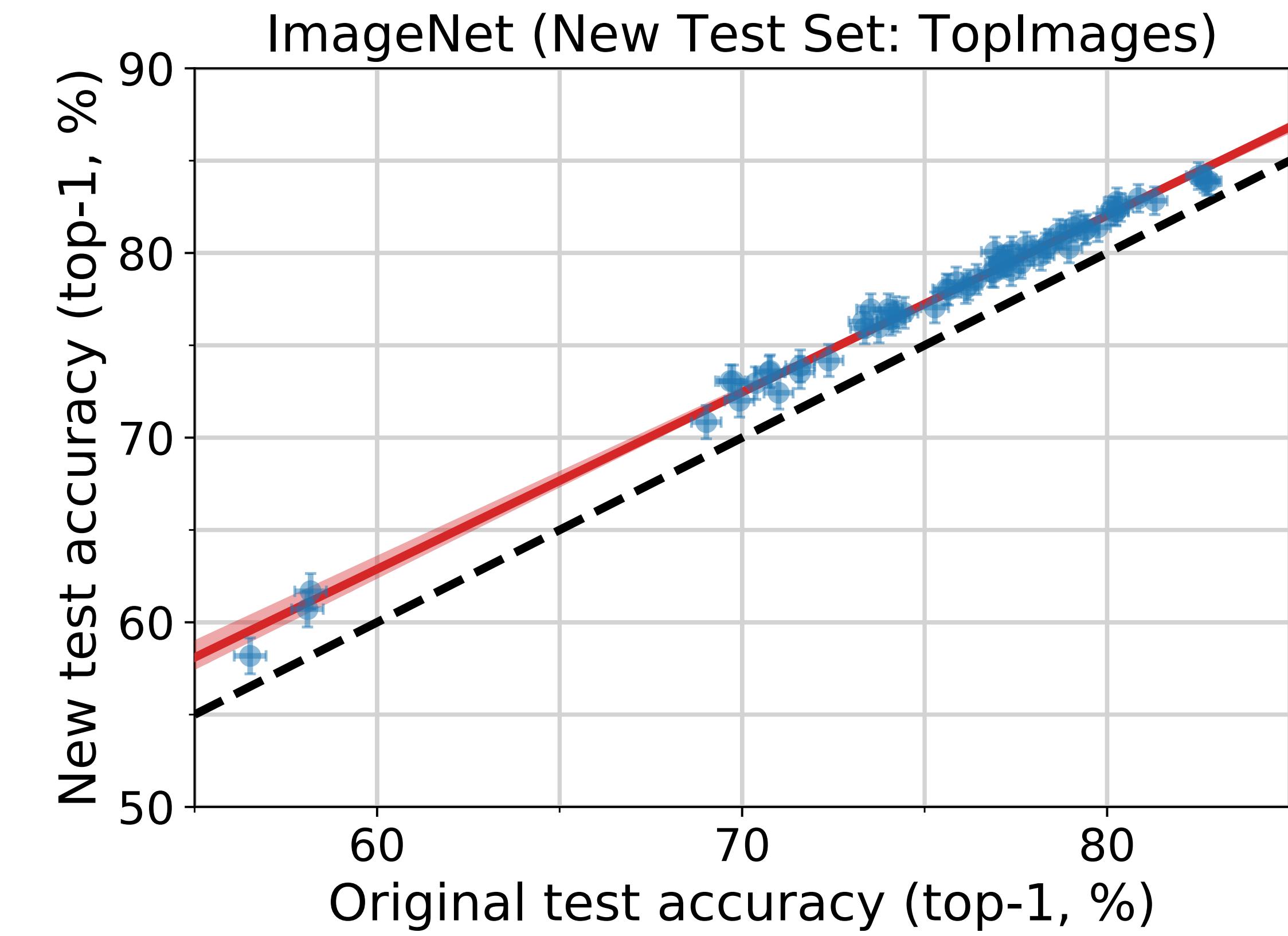




— Ideal reproducibility

Model accuracy

— Linear fit



Relative ordering is stable, absolute accuracies are brittle.

1. How reliable are ML benchmark results? (Internal validity)
2. Do benchmark results transfer across learning problems? (External validity)
3. Do benchmark results transfer across test distributions? (External validity)
4. Course projects

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

Why focus on ImageNet?

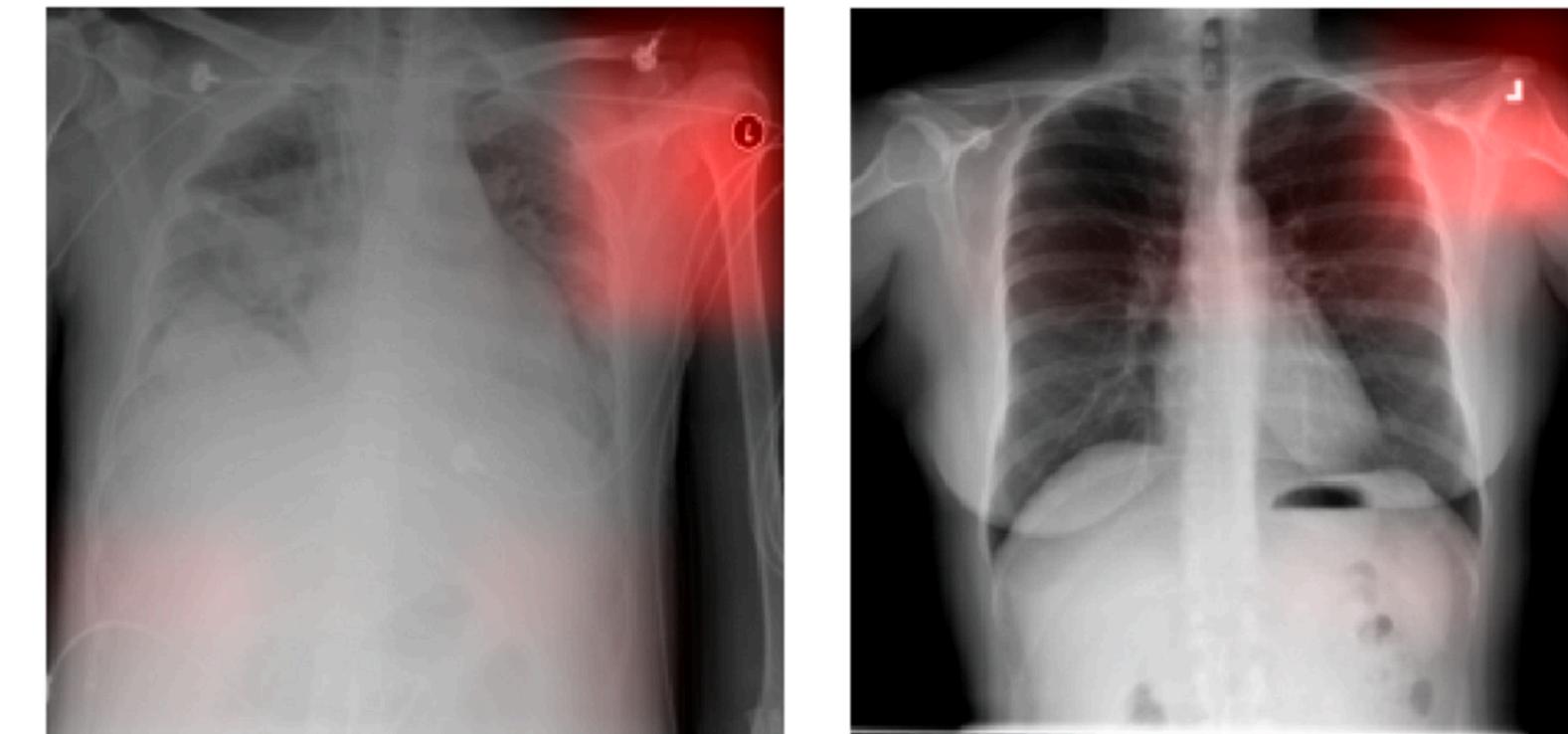
The community has spent **a lot** of effort on ImageNet.

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

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



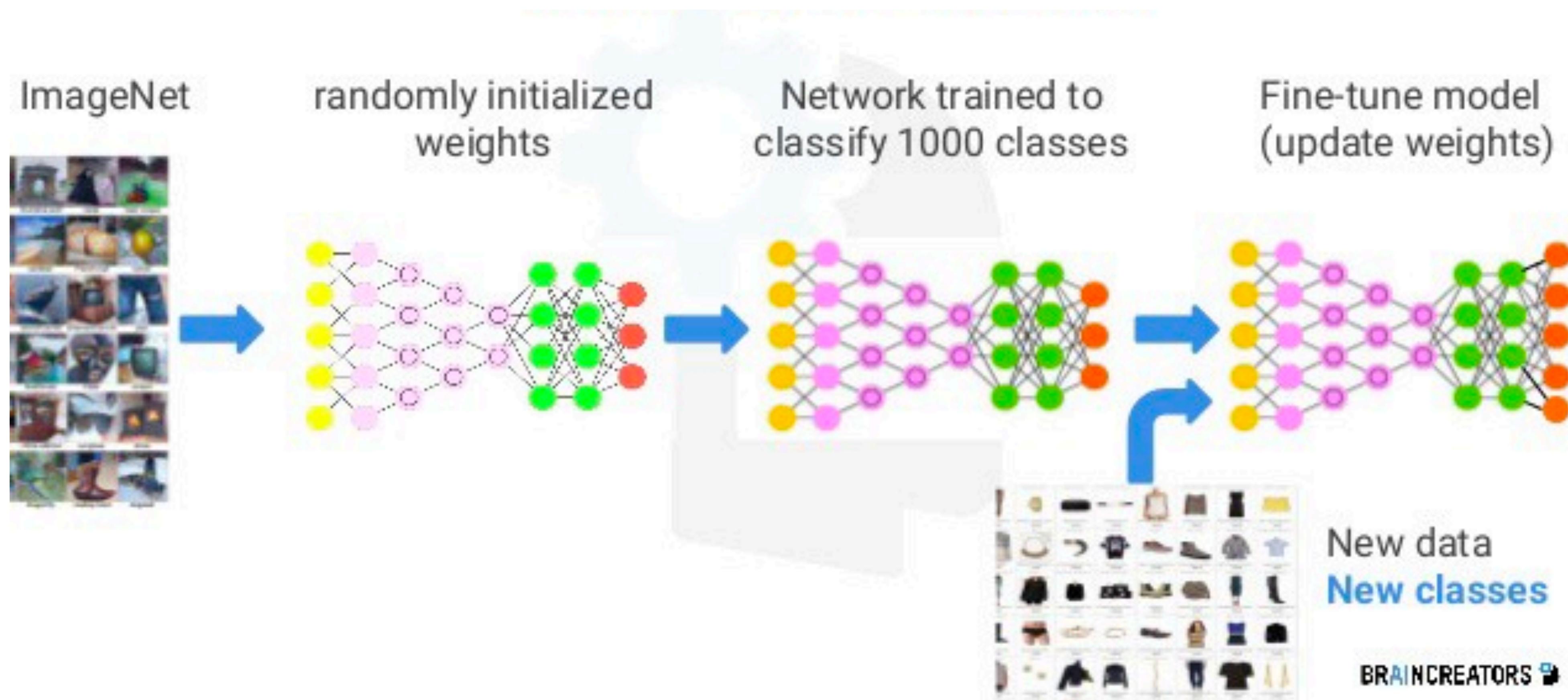
Food-101



Medical imaging

Transfer Learning

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.

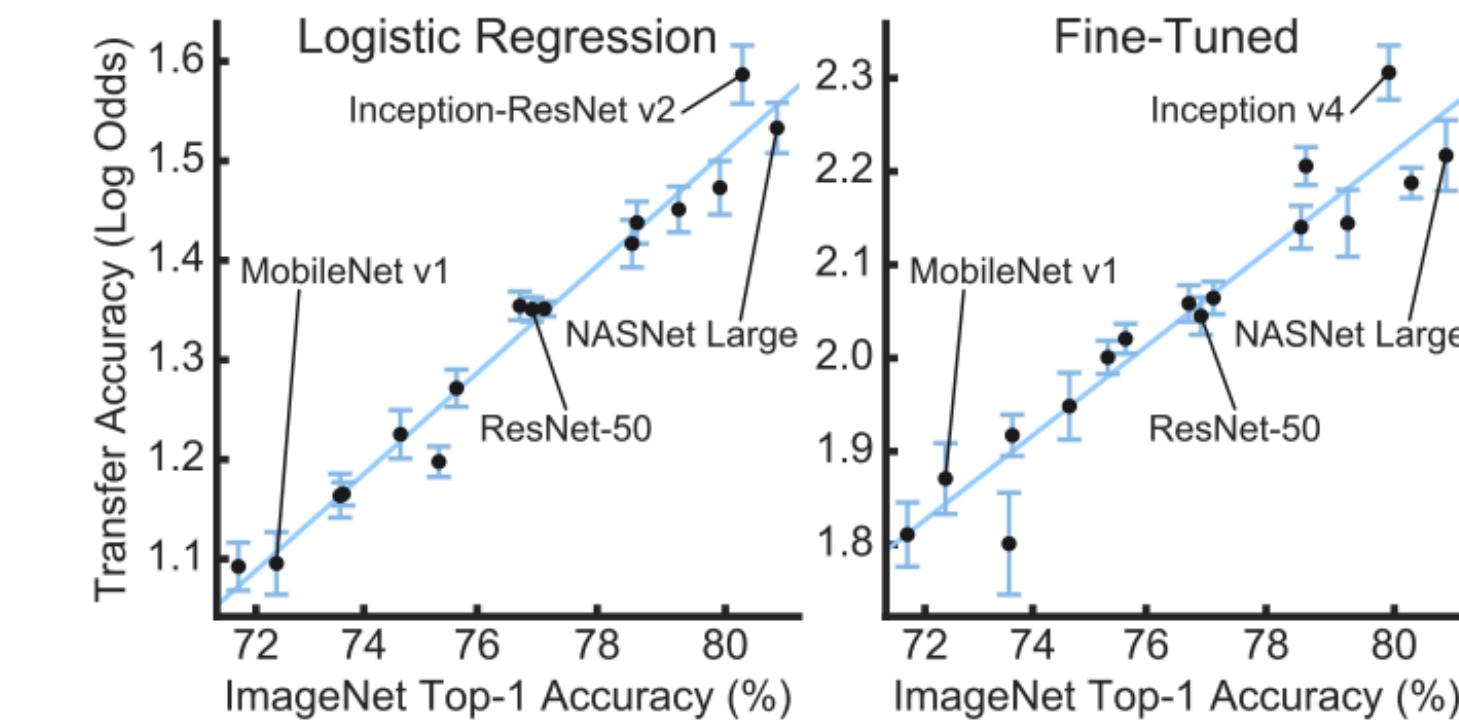


Figure 1. Transfer learning performance is highly correlated with ImageNet top-1 accuracy for fixed ImageNet features (left) and fine-tuning from ImageNet initialization (right). The 16 points in each plot represent transfer accuracy for 16 distinct CNN architectures, averaged across 12 datasets after 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.

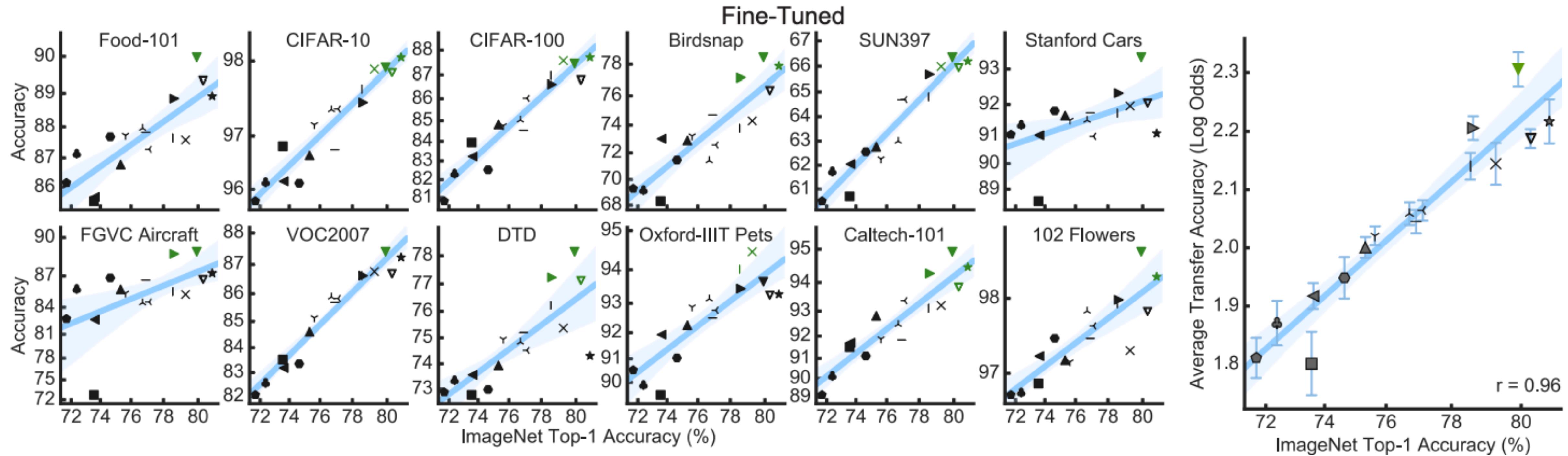
In the present work, we seek to test these hypotheses by investigating the transferability of both ImageNet features and

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).

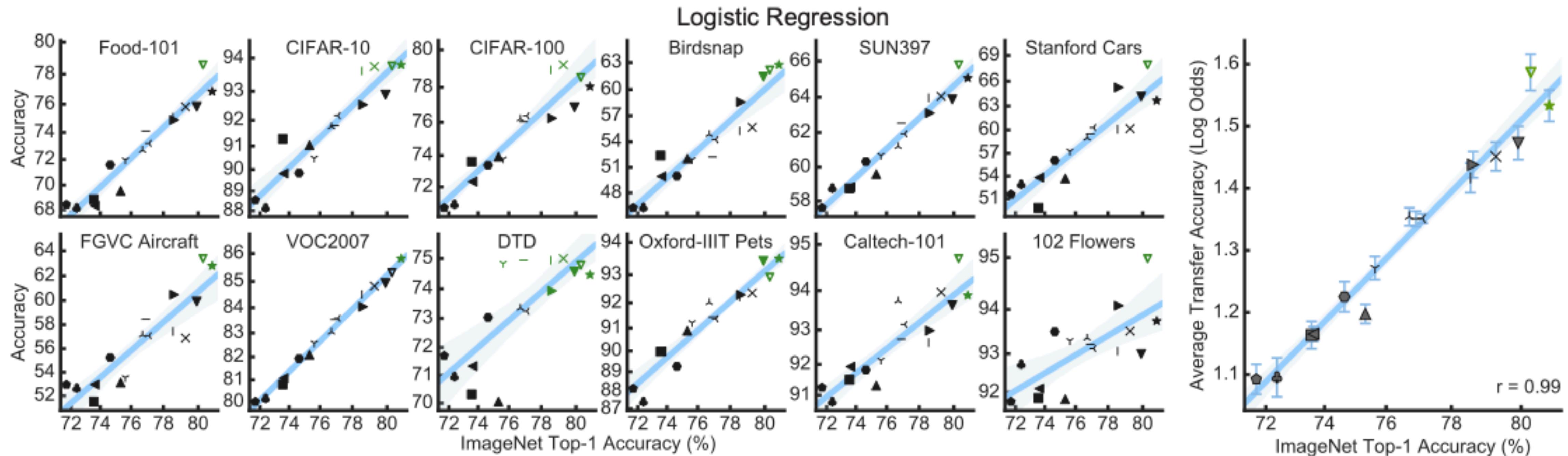
Better ImageNet Models Transfer Better



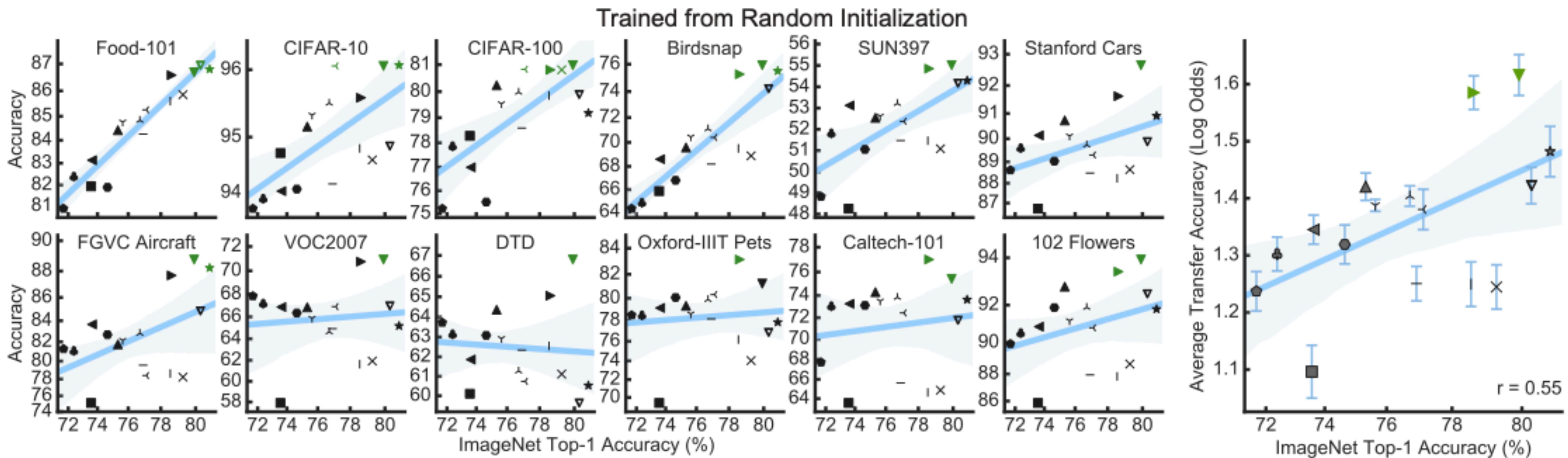
→ Progress on ImageNet helps on a wide range of image classification datasets.
Also transfer of techniques to other tasks (object detection, etc.)

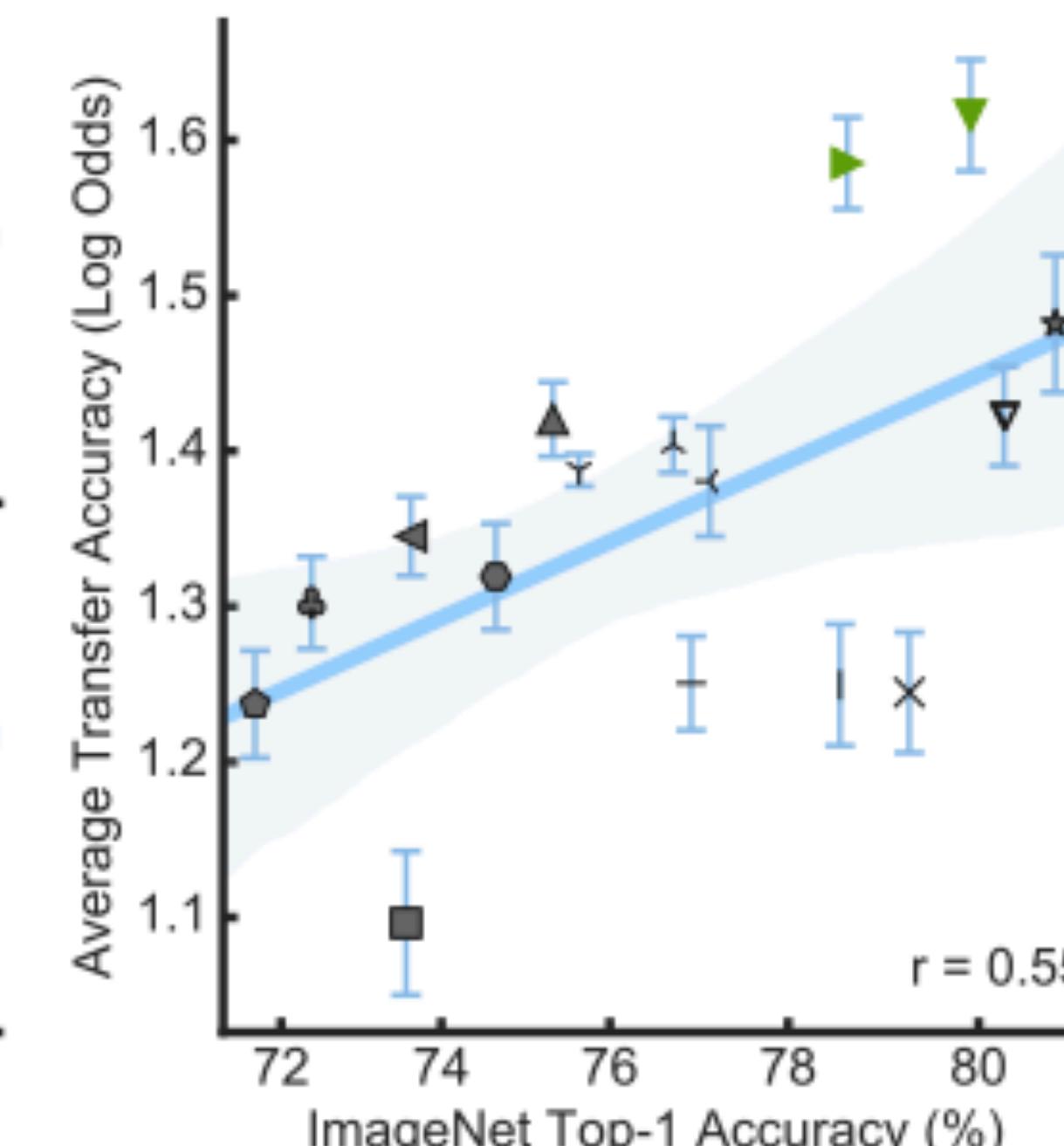
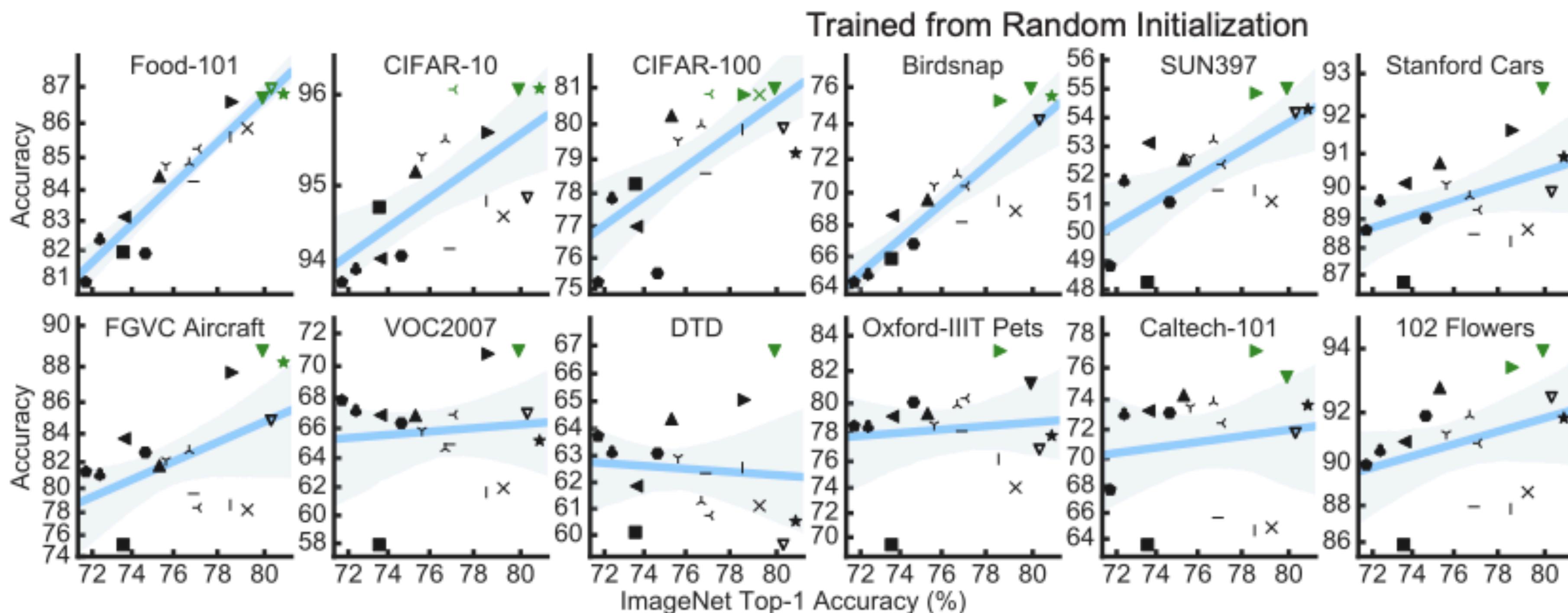
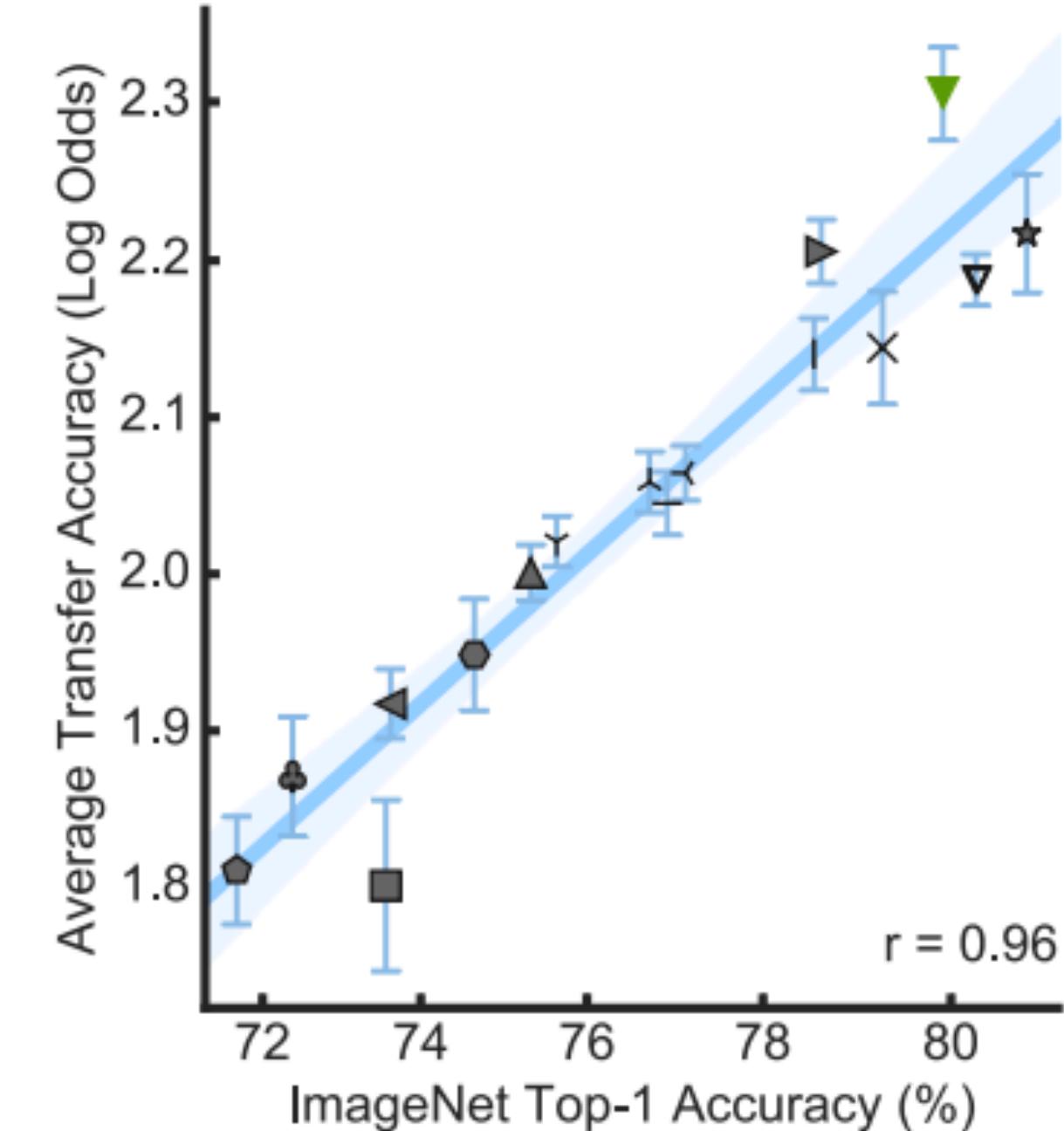
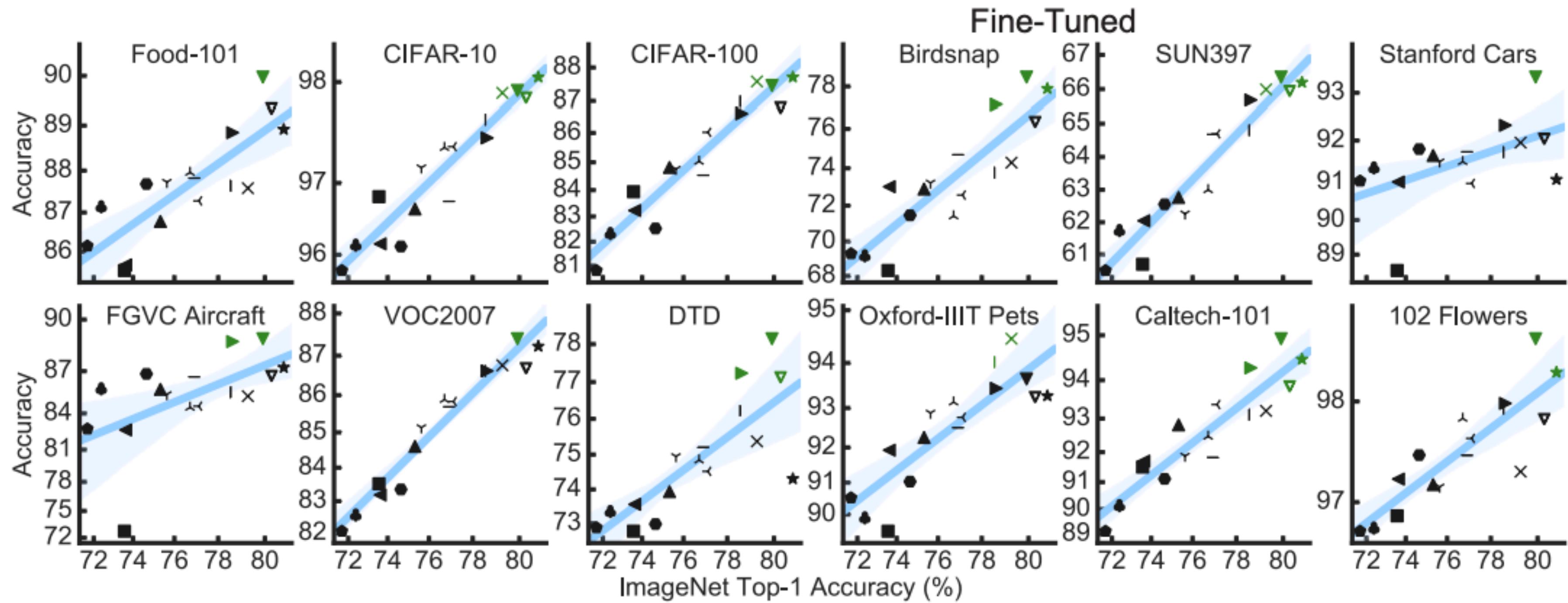
But: This is not guaranteed. Some datasets are considered “bad” or too specialized.
(Models don’t work “in the wild”)

More results from the paper



More results from the paper

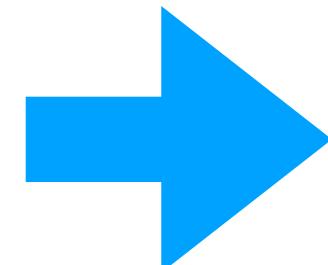




1. How reliable are ML benchmark results? (Internal validity)
2. Do benchmark results transfer across learning problems? (External validity)
3. Do benchmark results transfer across test distributions? (External validity)
4. Course projects

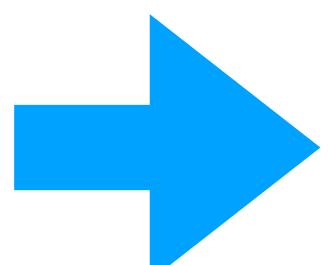
Model evaluation in ML benchmarks

Train and test sets are usually derived from a larger dataset via a random split

- 
- Train and test set are from the same distribution

Papers usually rank models by their performance on a single test set

But: when deployed “in the wild”, models usually encounter a different distribution.

- 
- What happens on other test distributions?
 - How large is the performance drop of a model?
 - Is the rank ordering consistent?

Distribution shifts are a real problem

Distribution shifts are a real problem

February 2018:

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

Darrell Etherington @etherington 3:13 pm PST • February 7, 2018

 Comment

Tesla  had aimed to do a cross-country U.S. drive in one of its vehicles using fully autonomous driving capabilities by the end of last year. Obviously it didn't make that goal, or you'd have heard about it. Instead, Tesla CEO Elon Musk now says he anticipates being able to make the trip within three months, or six months at the long end.

Distribution shifts are a real problem

February 2018:

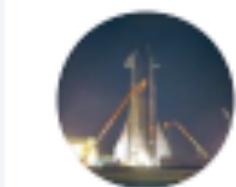
Elon Musk expects to do coast-to-coast autonomous Tesla drive in 3 to 6 months

Darrell Etherington @etherington 3:13 pm PST • February 7, 2018

Comment

Tesla had aimed to do a cross-country U.S. drive in one of its vehicles using fully autonomous driving capabilities by the end of last year. Obviously it didn't make that goal, or you'd have heard about it. Instead, Tesla CEO Elon Musk now says he anticipates being able to make the trip within three months, or six months at the long end.

July 2019:



Elon Musk 
@elonmusk

Replying to @PatrickLac007 @AngelGPonce and @Tesla

Parking lots are a remarkably hard problem. Doing an in-depth engineering review of Enhanced Summon later today.

11:42 AM · Jul 13, 2019 · Twitter for iPhone

83 Retweets 2.7K Likes



Distribution shifts are a real problem

February 2018:

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



July 2019:

Elon Musk @elonmusk

Replying to @PatrickLac007 @AngelGPonce and @Tesla

Parking lots are a remarkably hard problem. Doing an in-depth engineering review of Enhanced Summon later today.

11:42 AM · Jul 13, 2019 · Twitter for iPhone

83 Retweets 2.7K Likes

September 2019: Enhanced Summon released

Distribution shifts are a real problem

February 2018:

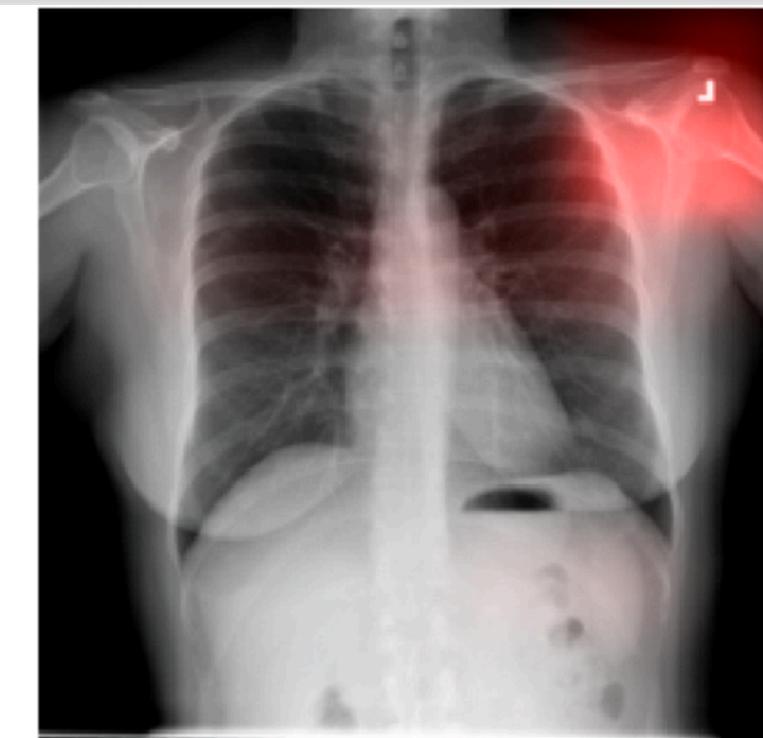
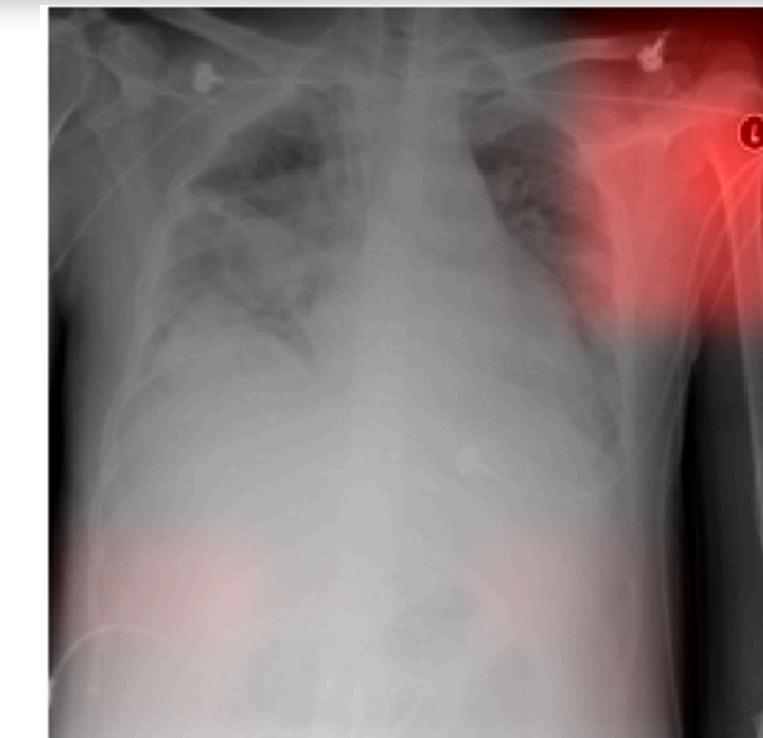
Elon Musk expects to do coast-to-coast autonomous Tesla drive in 3 to 6 months



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>



September 2019: Enhanced Summon released

Distribution shifts are a real problem

February 2018:

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

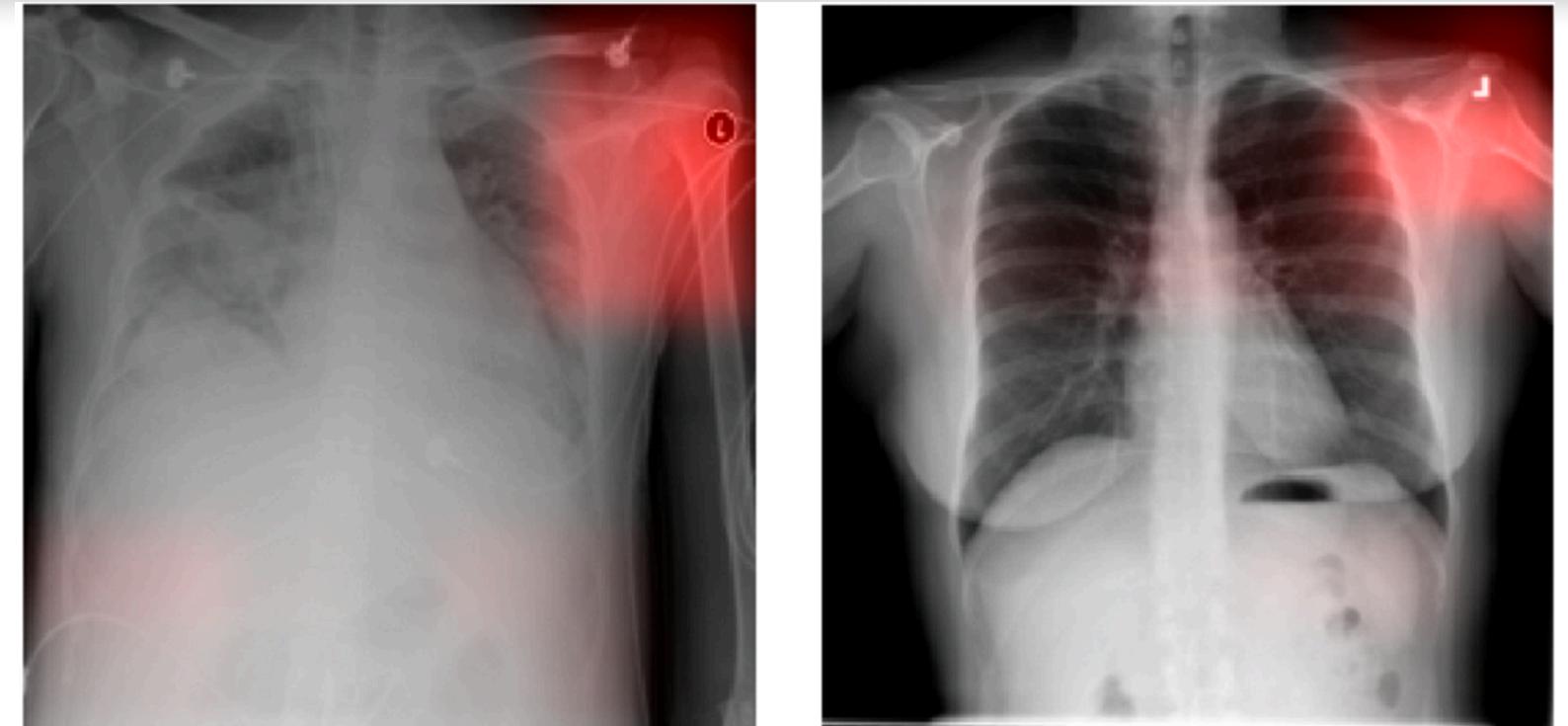


September 2019: Enhanced Summon

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

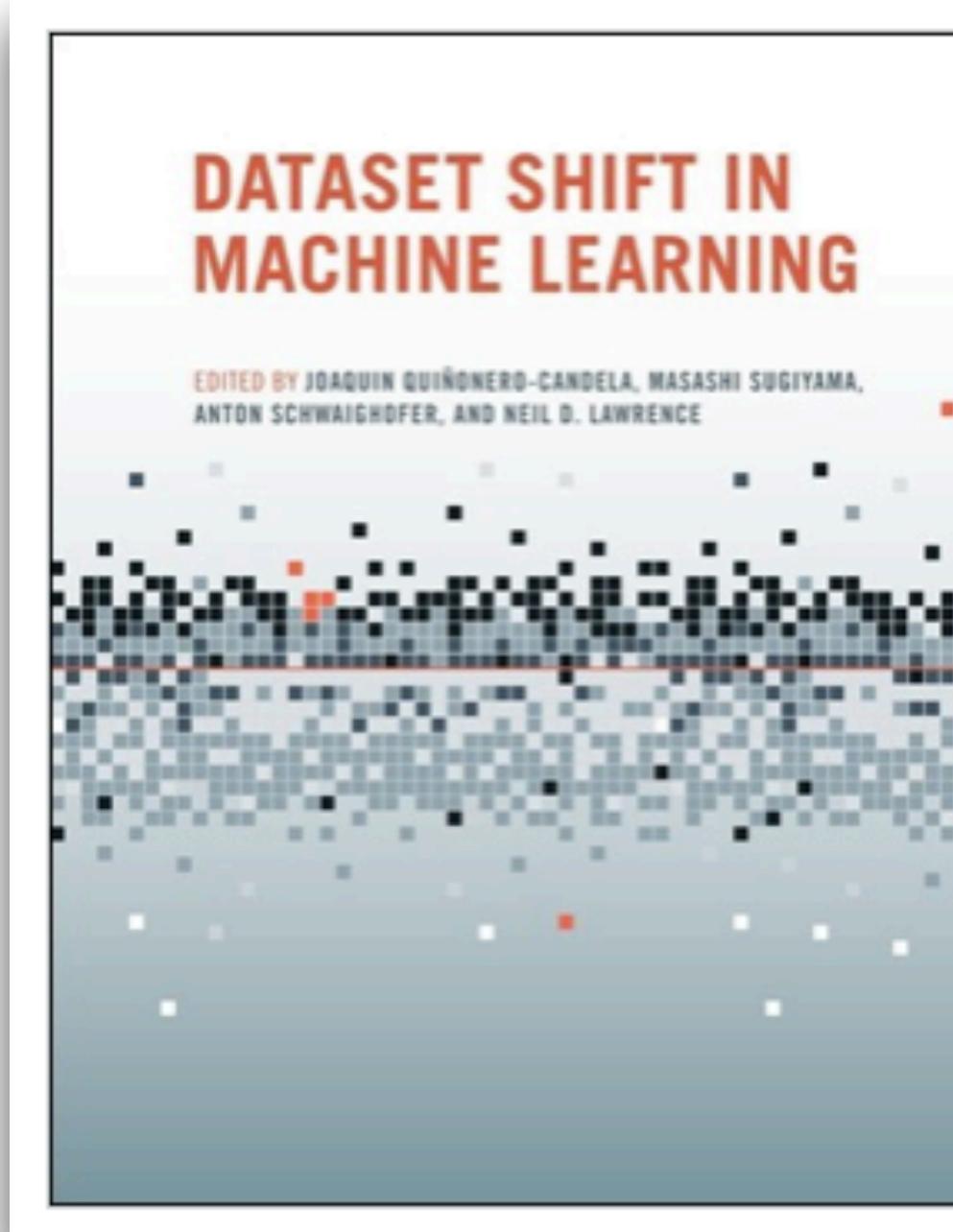
Published: November 6, 2018 • <https://doi.org/10.1371/journal.pmed.1002683>



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”.

Distribution shifts have been studied for a long time

2011



2008

But ML didn't work that well yet
even in-distribution ...

From [Neural Information Processing series](#)

Dataset Shift in Machine Learning

Edited by [Joaquin Quiñonero-Candela](#), [Sugiyama](#), [Anton Schwaighofer](#) and [Neil D. Lawrence](#)

An overview of recent efforts in the machine learning community to deal with dataset and covariate shift. This occurs when test and training inputs and outputs are drawn from different distributions.

Unbiased Look at Dataset Bias

Antonio Torralba
Massachusetts Institute of Technology
torralba@csail.mit.edu

Alexei A. Efros
Carnegie Mellon University
efros@cs.cmu.edu

Abstract

Datasets are an integral part of contemporary object recognition research. They have been the chief reason for the considerable progress in the field, not just as source of large amounts of training data, but also as means of measuring and comparing performance of competing algorithms. At the same time, datasets have often been blamed for narrowing the focus of object recognition research, reducing it to a single benchmark performance number. Indeed, some datasets, that started out as data capture efforts aimed at representing the visual world, have become closed worlds unto themselves (e.g. the Corel world, the Caltech-101 world, the PASCAL VOC world). With the focus on beating the latest benchmark numbers on the latest dataset, have we perhaps lost sight of the original purpose?

The goal of this paper is to take stock of the current state of recognition datasets. We present a comparison study using a set of popular datasets, evaluated based on a number of criteria including: relative data bias, cross-dataset generalization, effects of closed-world assumption, and sample value. The experimental results, some rather surprising, suggest directions that can improve dataset collection as well as algorithm evaluation protocols. But more broadly, the hope is to stimulate discussion in the community regarding this very important, but largely neglected issue.

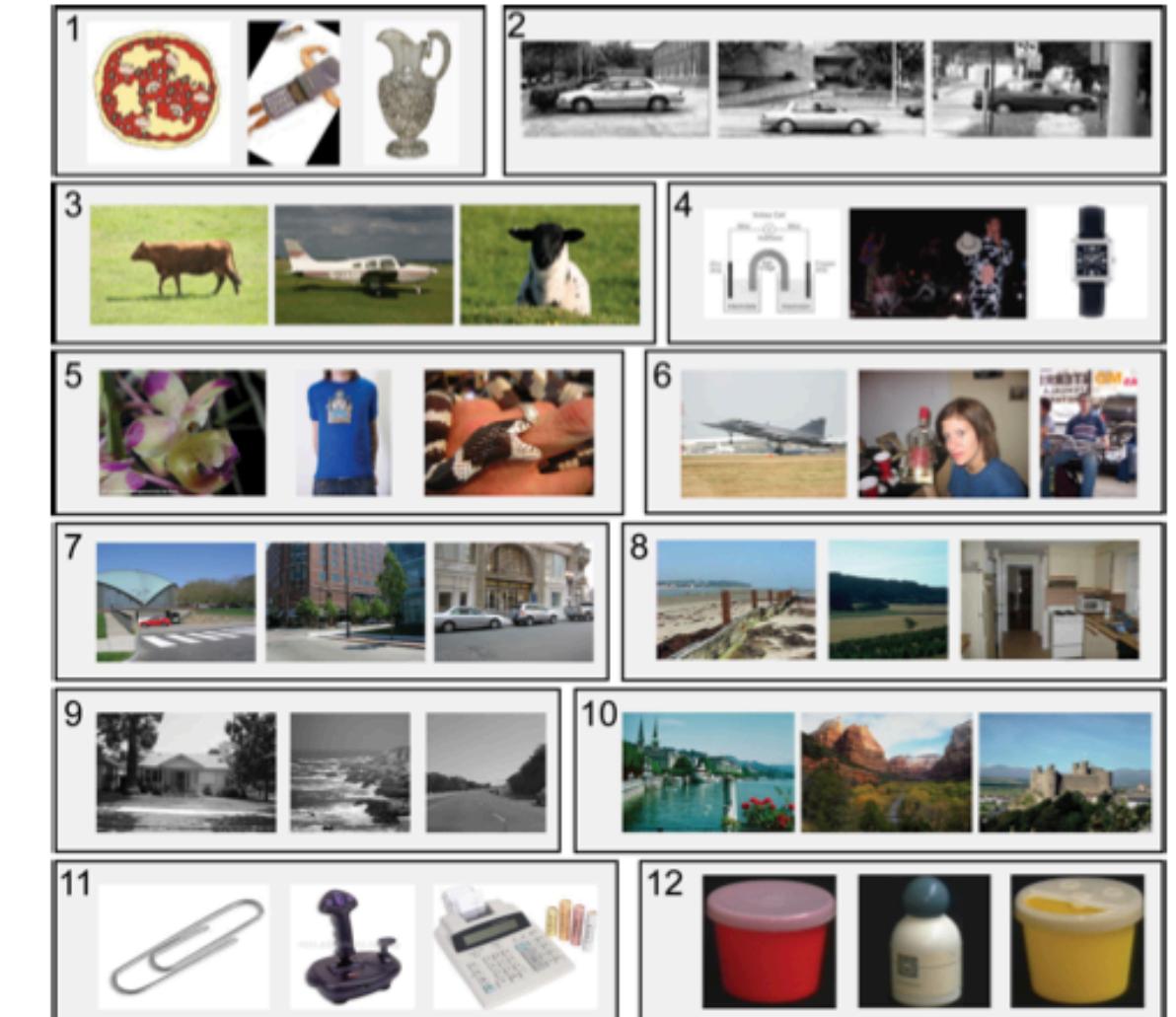


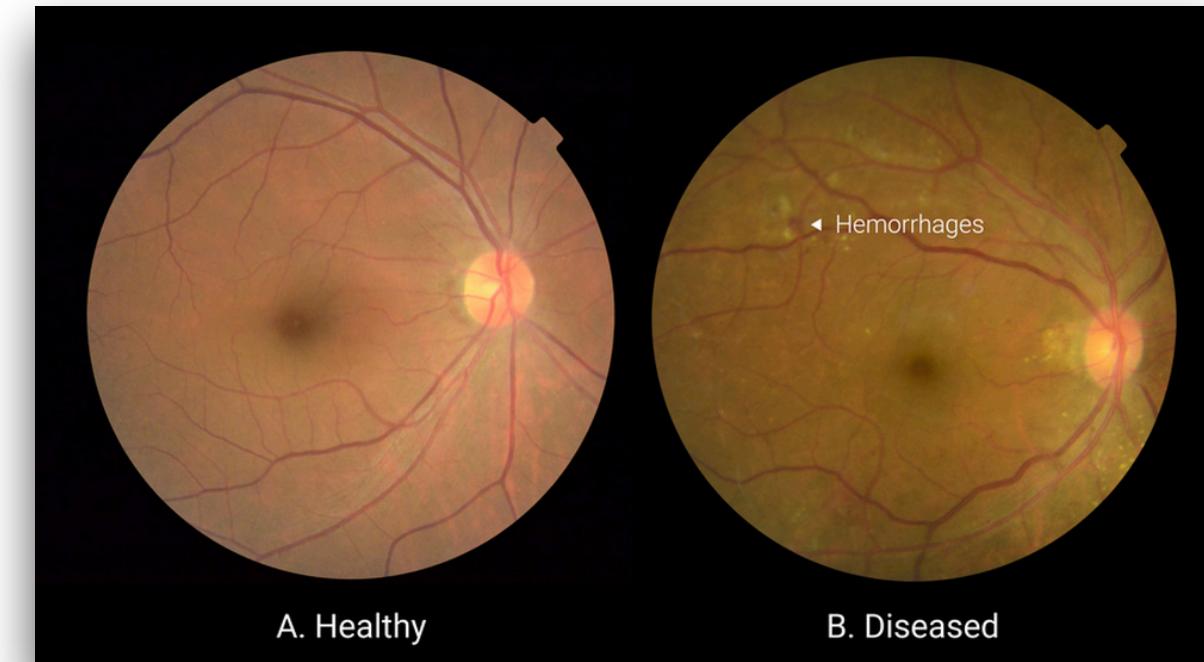
Figure 1. Name That Dataset: Given three images from twelve popular object recognition datasets, can you match the images with the dataset? (answer key below)

anyone who has worked in object and scene recognition (in

Safety-Critical Applications of ML



Transportation



Health care



Robotics

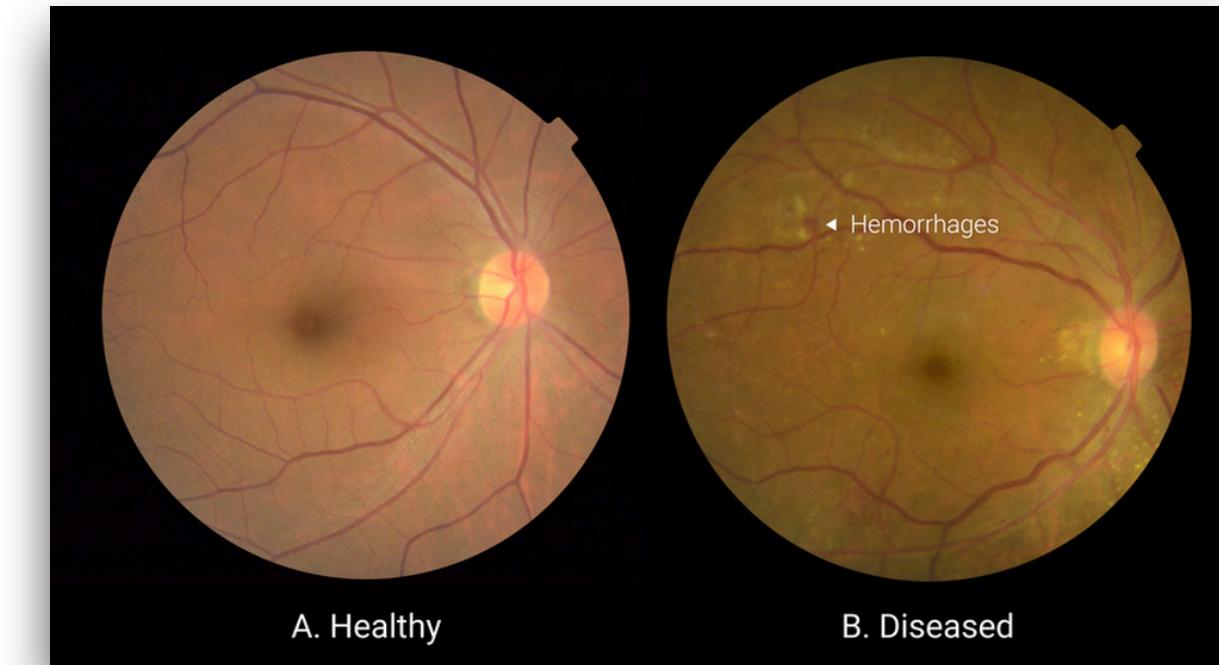


Content moderation

Safety-Critical Applications of ML



Transportation



Health care



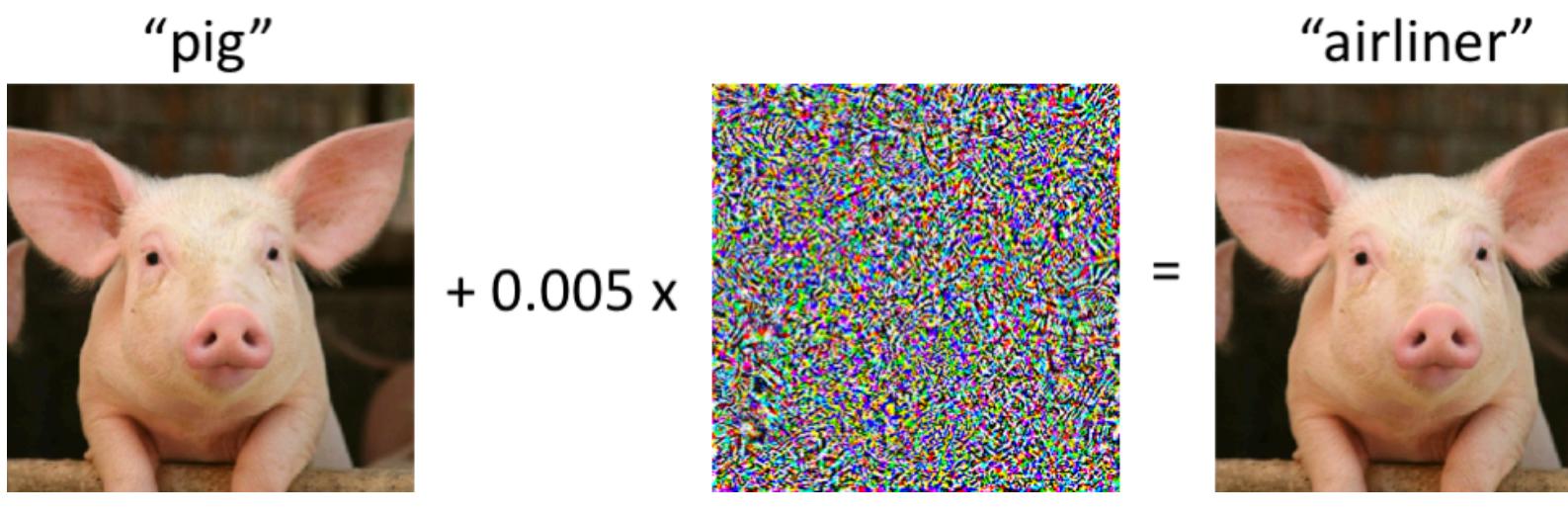
Robotics



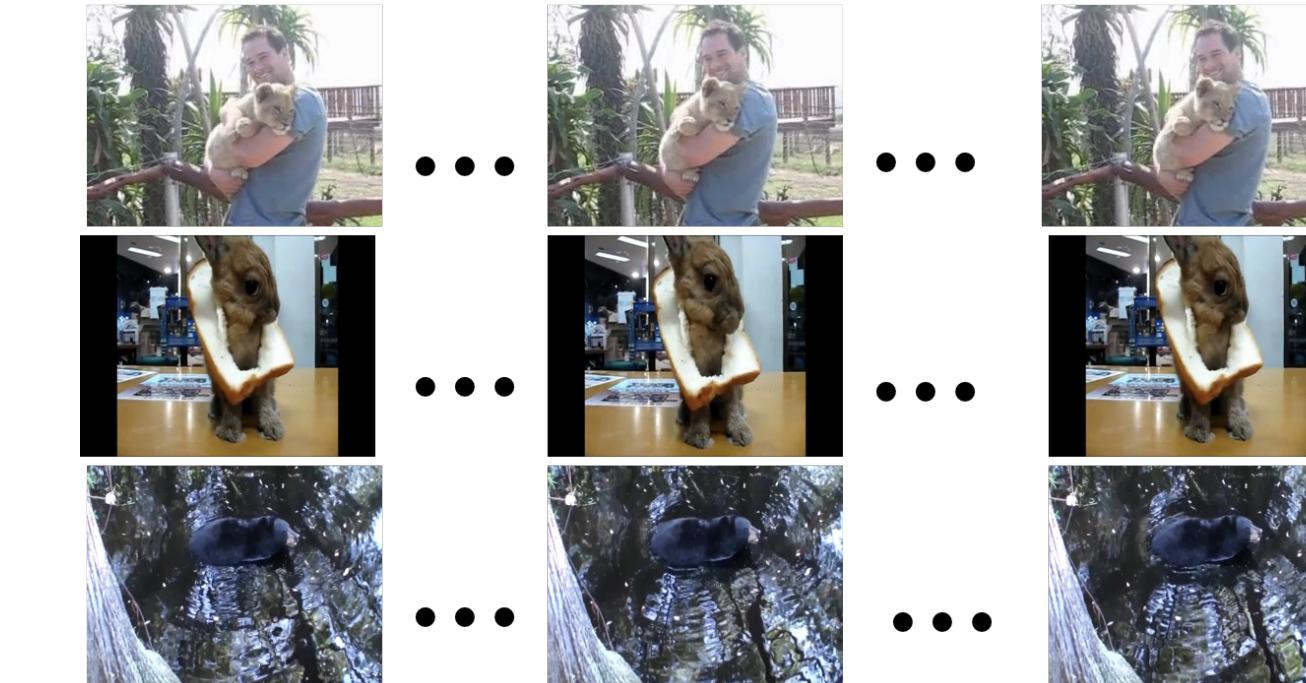
Content moderation

→ Need reliable machine learning

Robustness Notions in Image Classification



Adversarial examples



Video perturbations

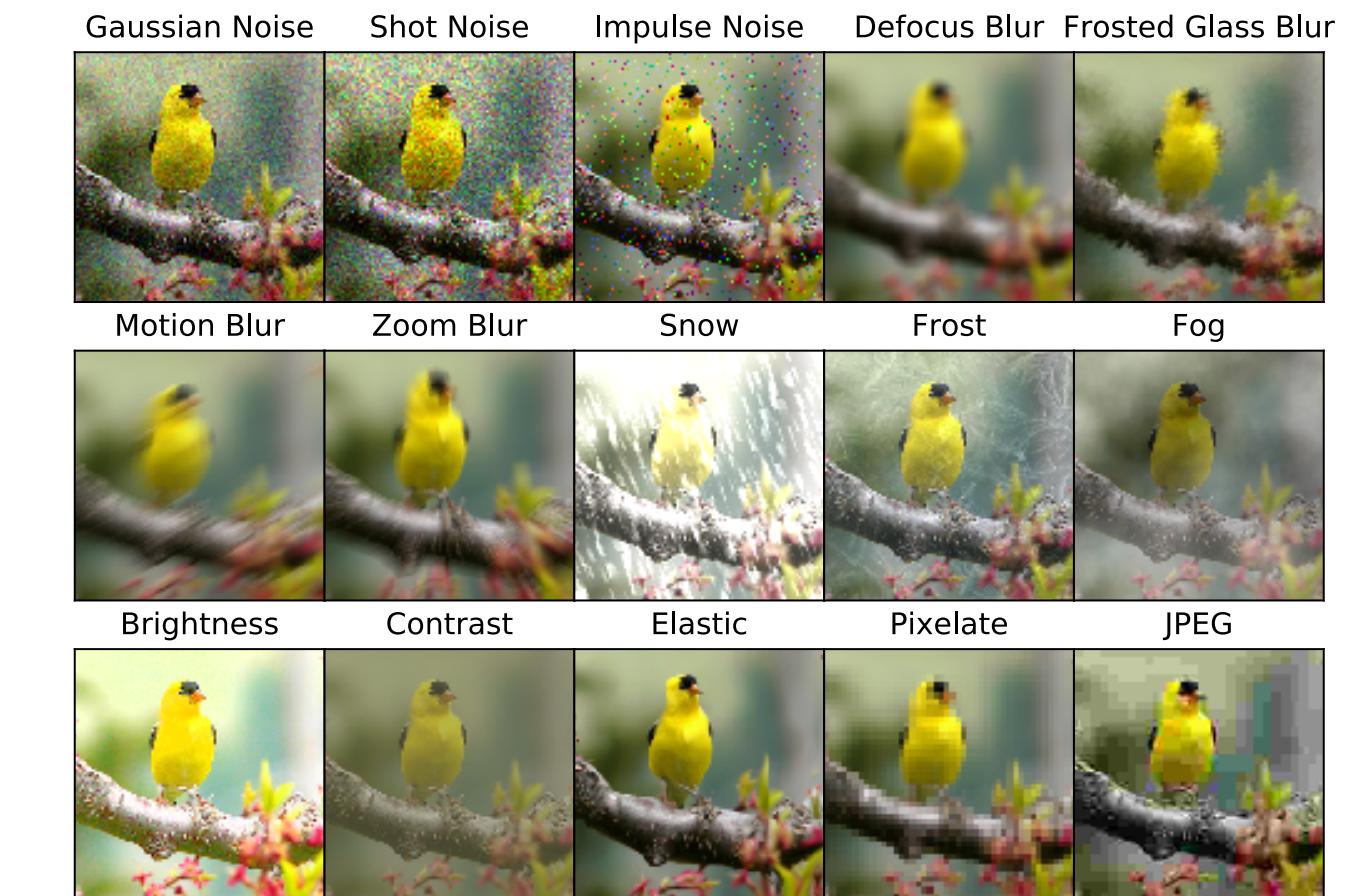
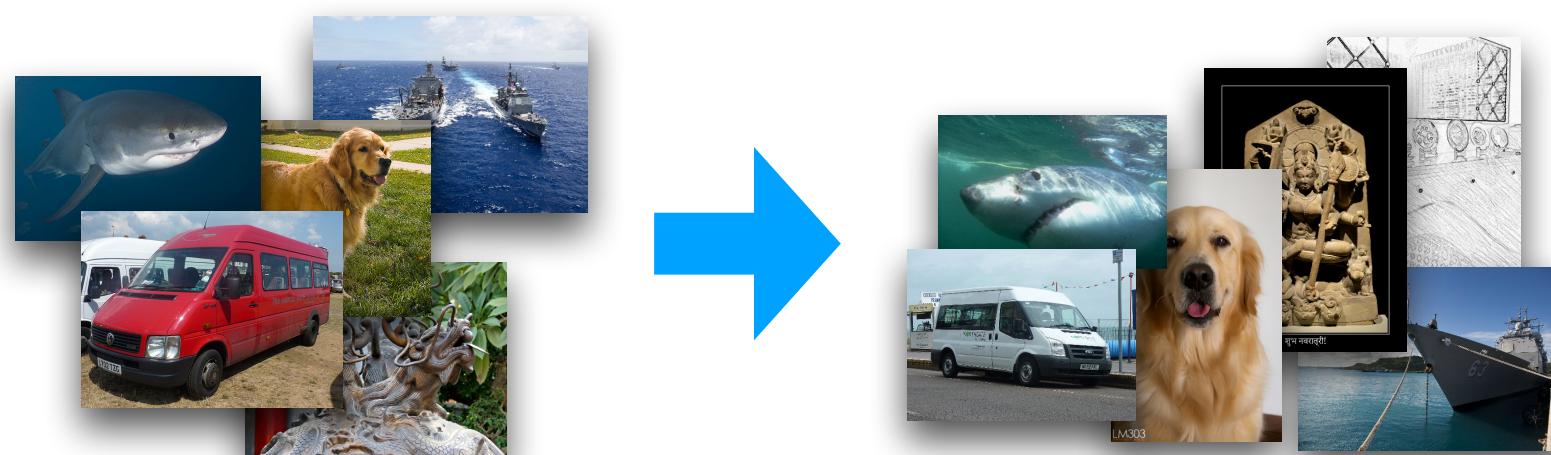


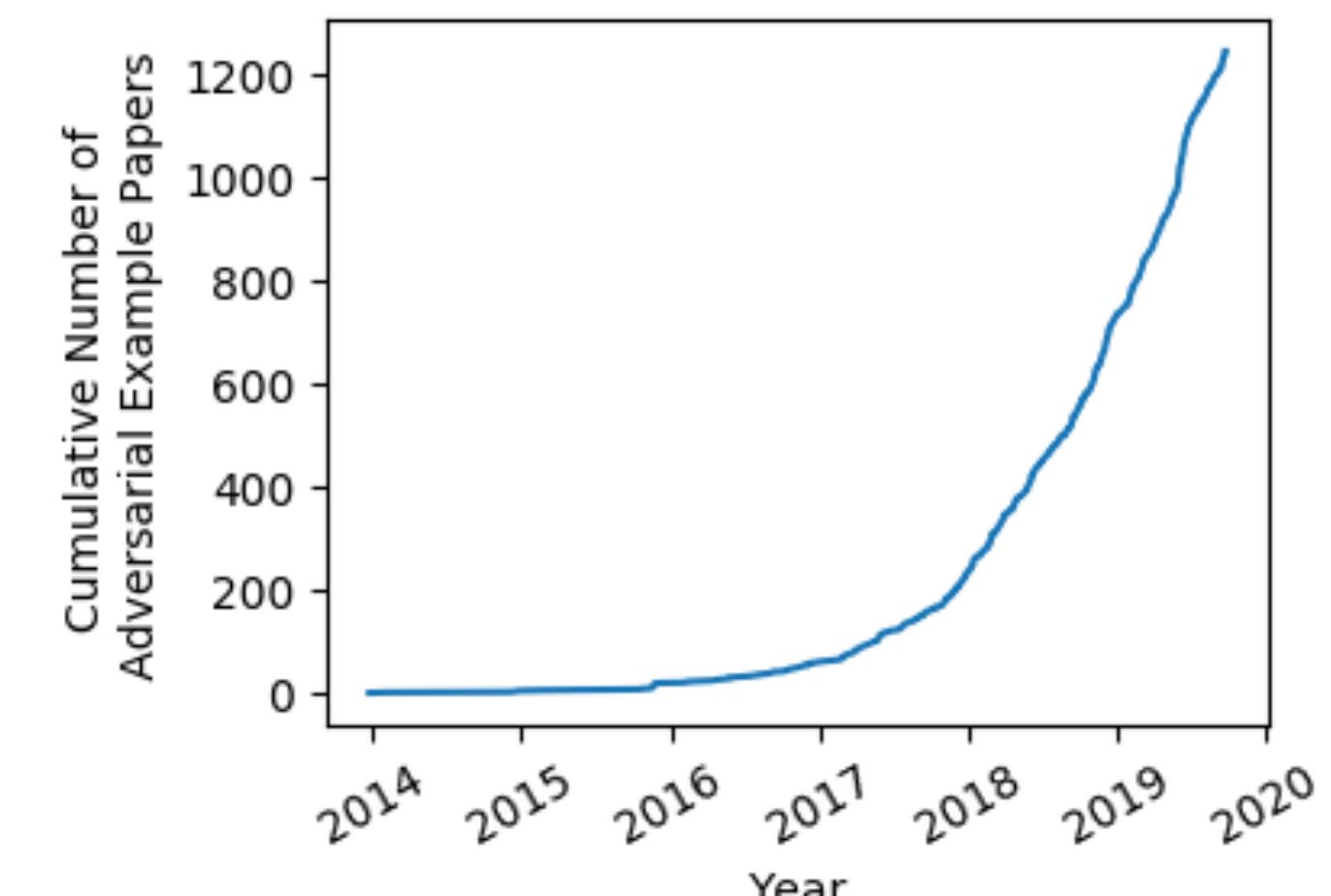
Image corruptions



Dataset shift



Geometric transformations



[Carlini '19]

Measuring Robustness to Natural Distribution Shifts in Image Classification

Rohan Taori
UC Berkeley

Achal Dave
CMU

Vaishaal Shankar
UC Berkeley

Nicholas Carlini
Google Brain

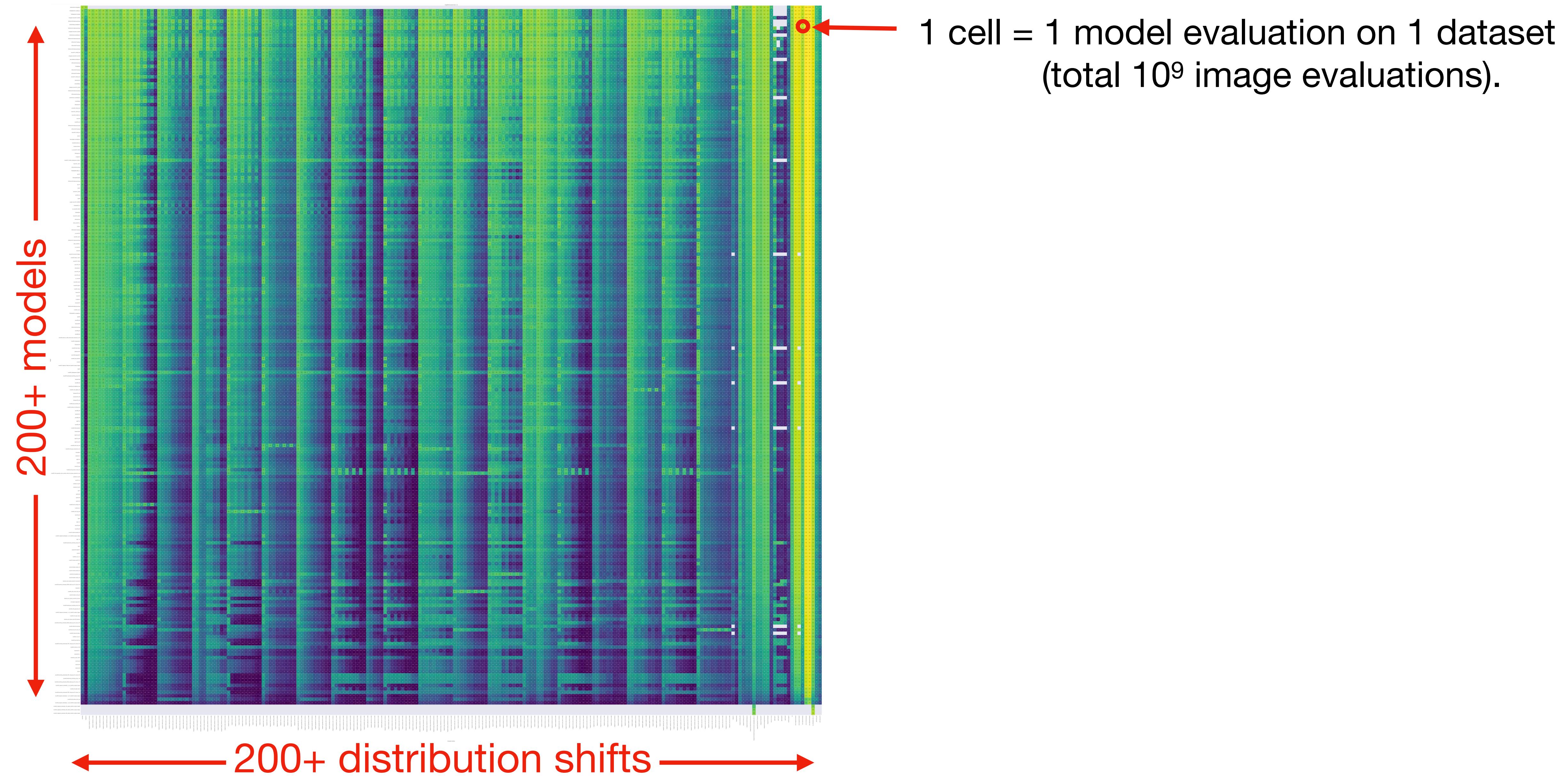
Benjamin Recht
UC Berkeley

Ludwig Schmidt
UC Berkeley

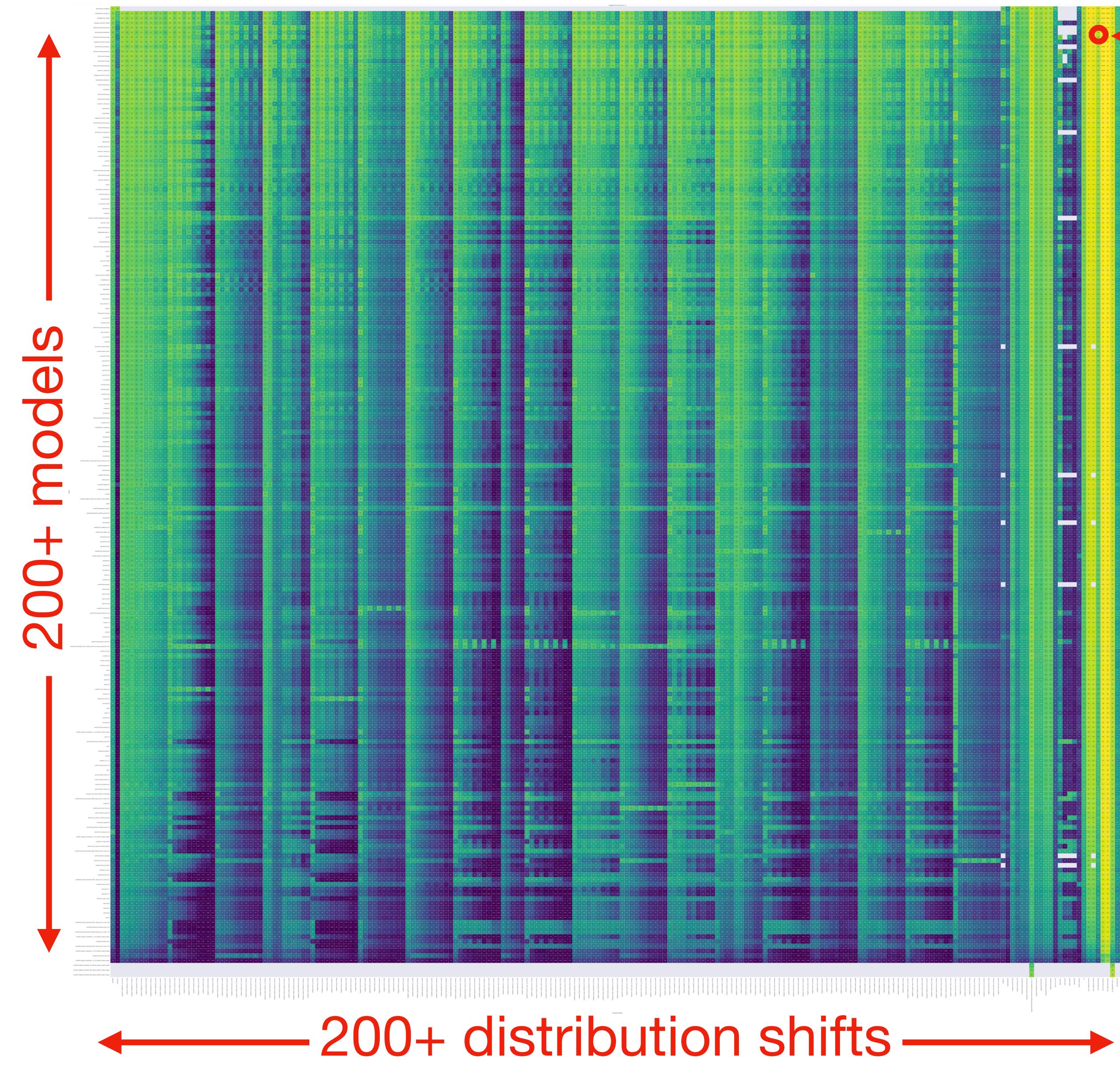
Abstract

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/>.

Our Testbed



Our Testbed

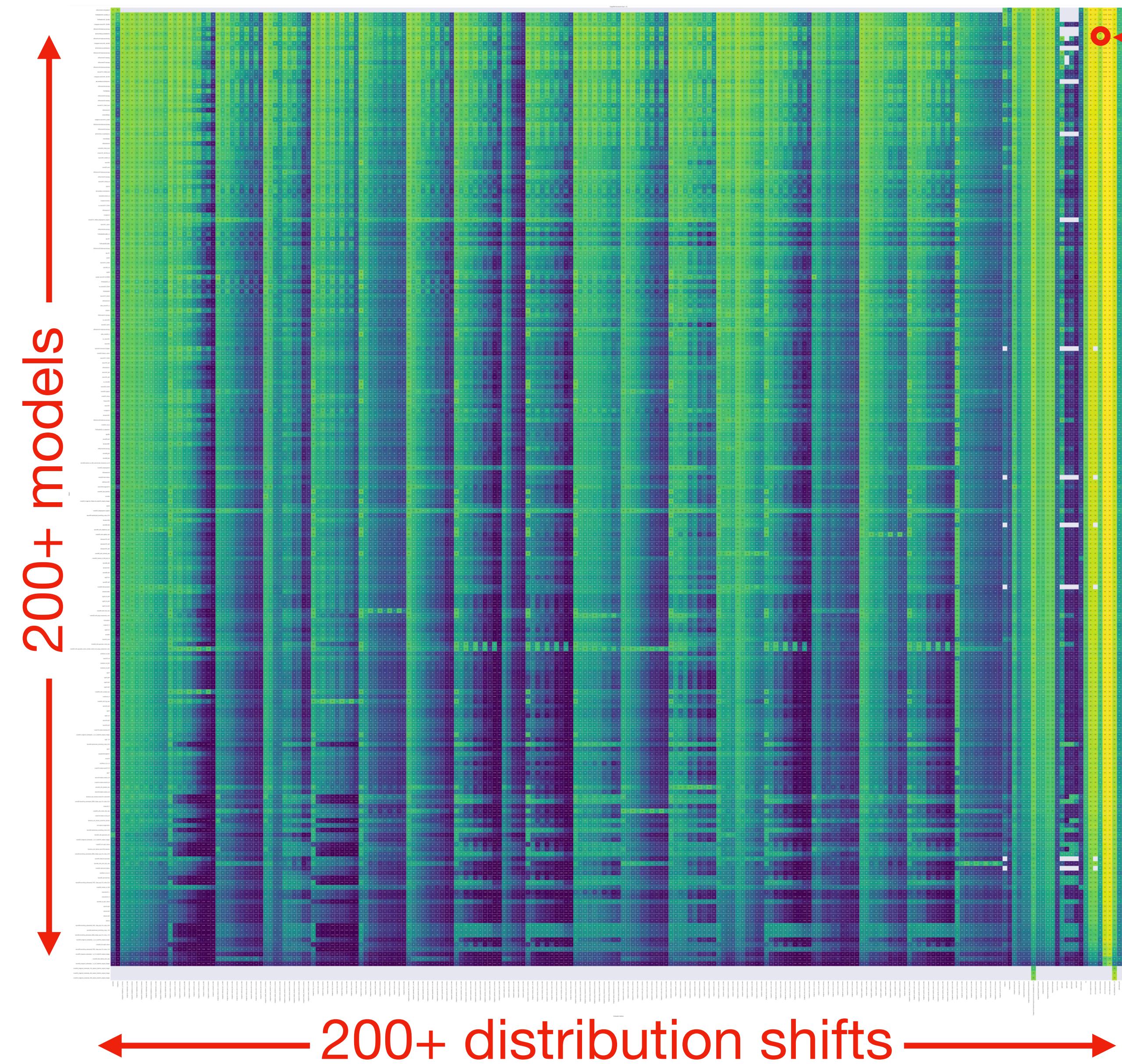


1 cell = 1 model evaluation on 1 dataset
(total 10^9 image evaluations).

Models:

- “Standard” models (just ImageNet acc.)
- Robust models (adversarially robust models, models with special data augmentation, etc.)
- Models trained on more data

Our Testbed



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

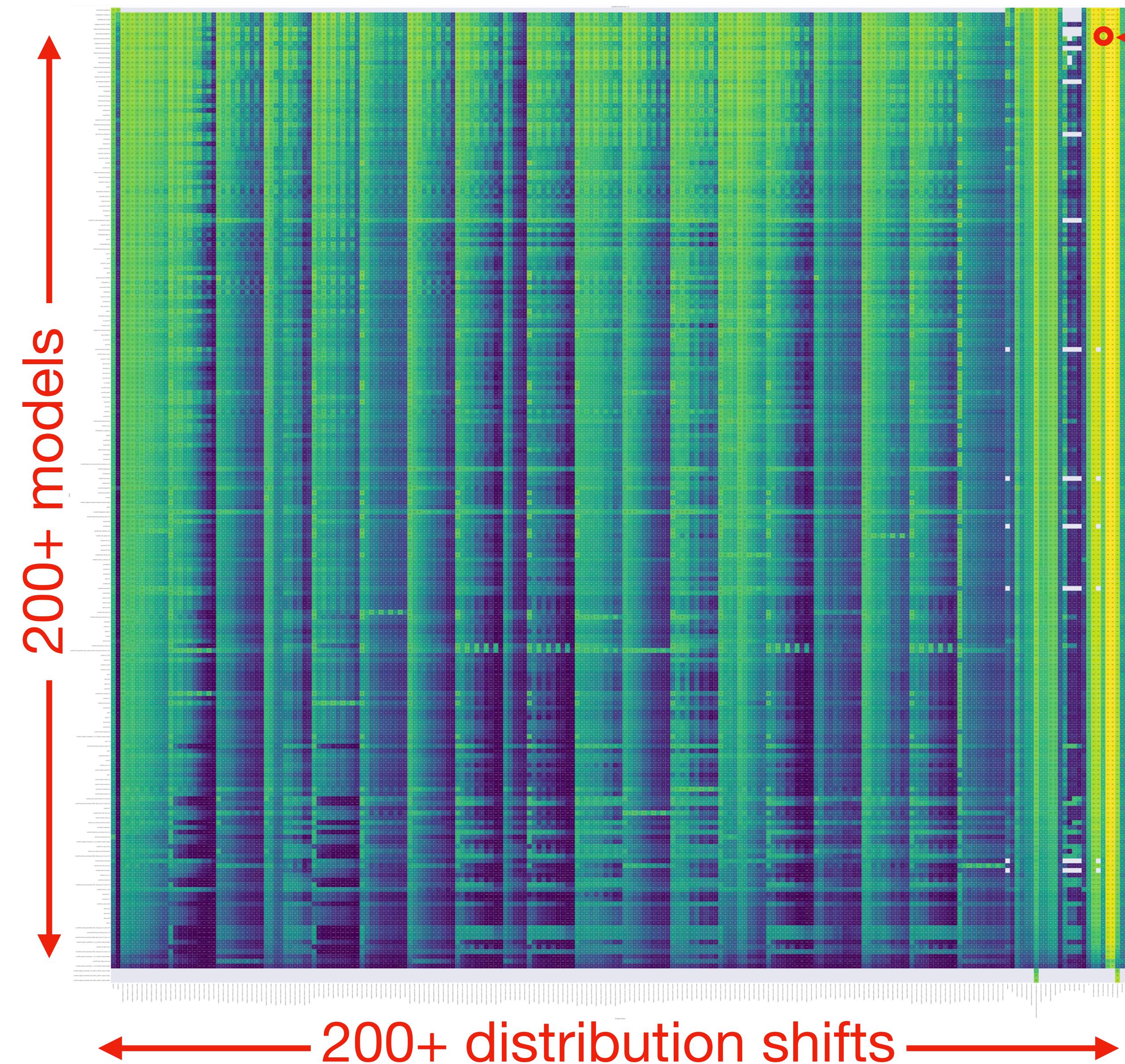
Models:

- “Standard” models (just ImageNet acc.)
- Robust models (adversarially robust models, models with special data augmentation, etc.)
- Models trained on more data

Natural distribution shifts:

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

Our Testbed



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

Models:

- “Standard” models (just ImageNet acc.)
- Robust models (adversarially robust models, models with special data augmentation, etc.)
- Models trained on more data

Natural distribution shifts:

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

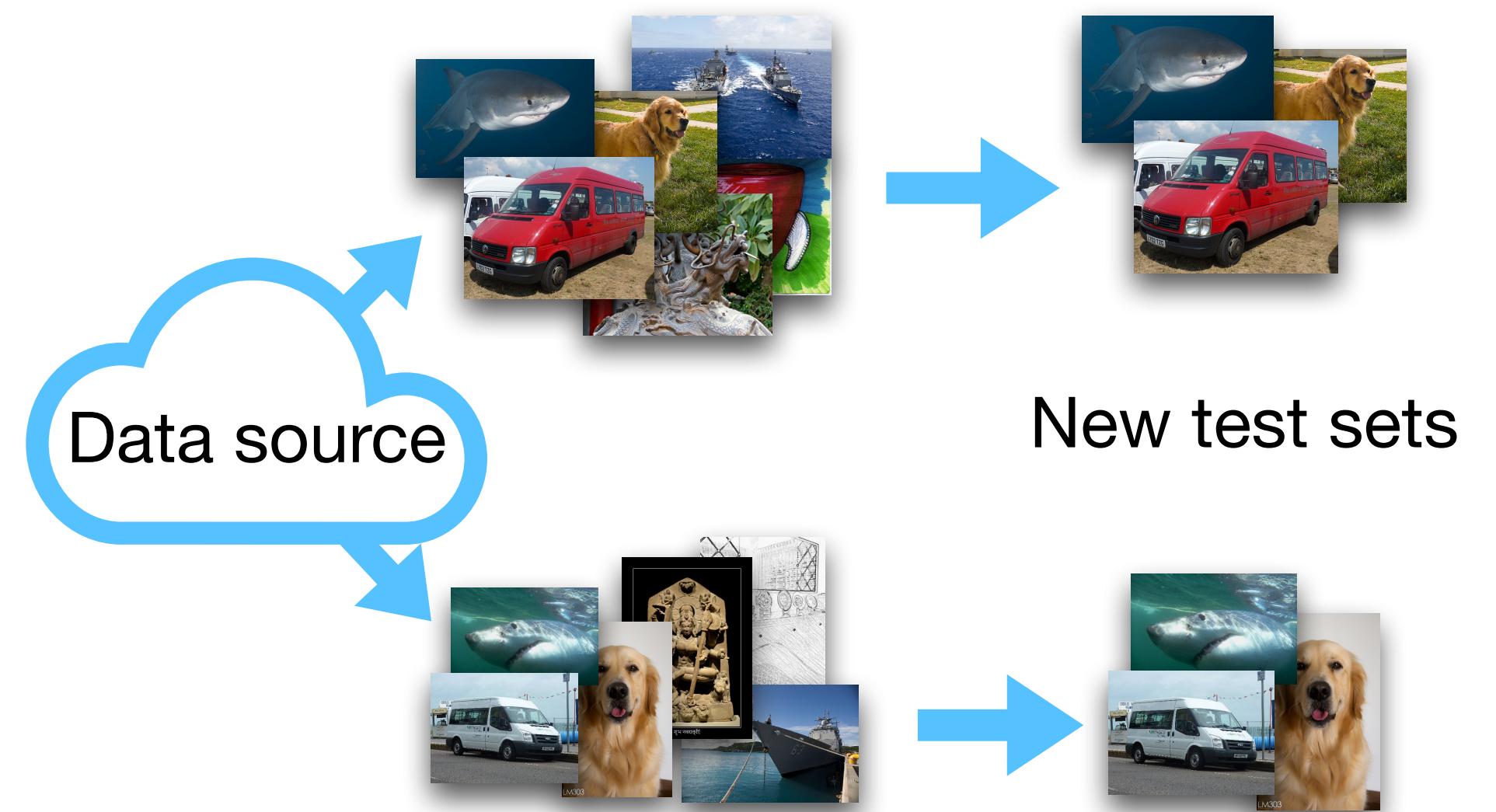
Synthetic distribution shifts:

- L_p-attacks
- Image corruptions

Goal: Classify Robustness Notions

We classify test-time robustness along two axes:

1. **Adversarial vs. benign**: does the input change depend on a trained model?



Goal: Classify Robustness Notions

We classify test-time robustness along two axes:

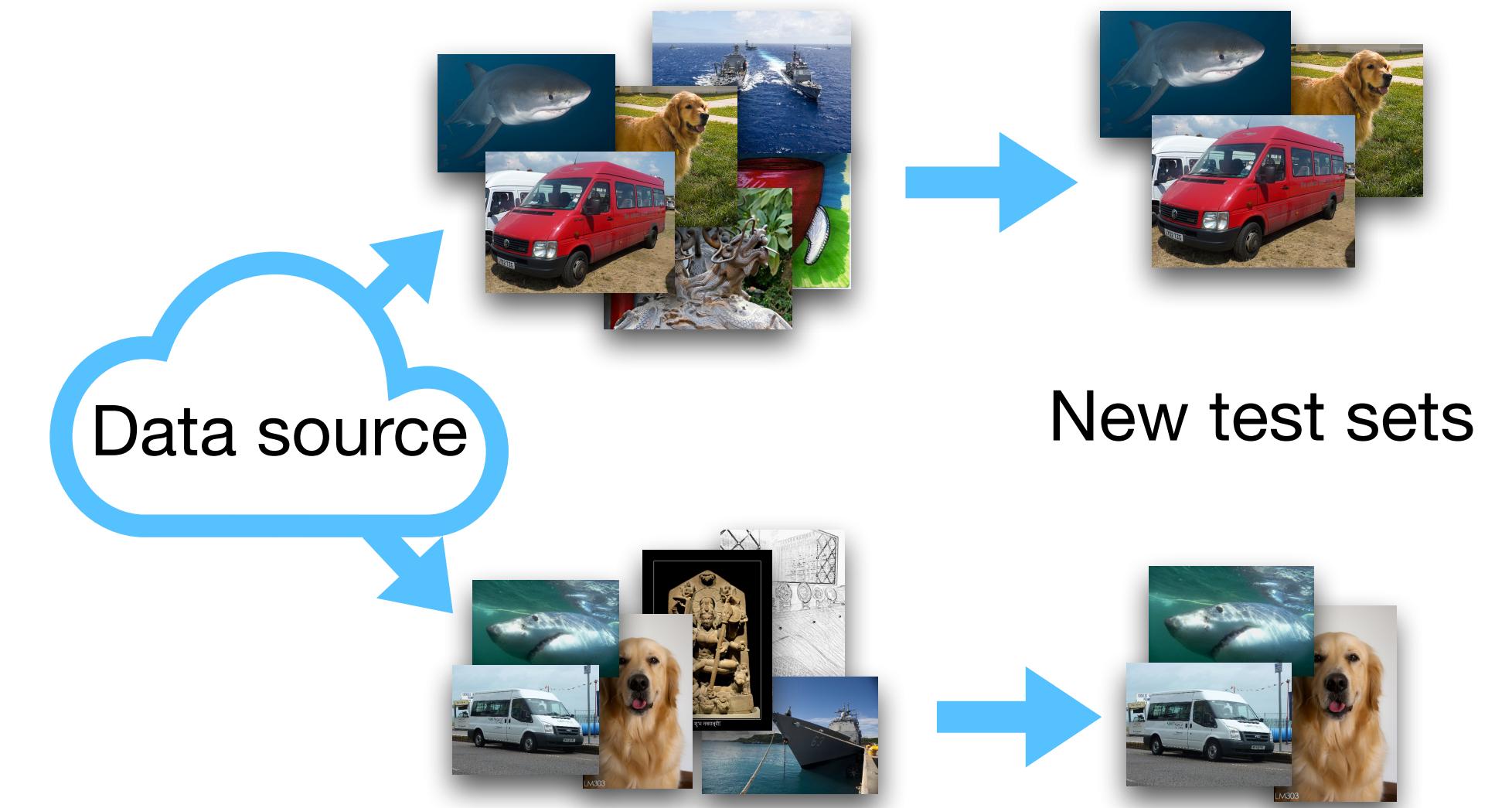
1. **Adversarial vs. benign**: does the input change depend on a trained model?
2. **Synthetic vs. natural**

Synthetic: computer-generated perturbations of a real dataset

$$D = \{ \text{[Image of a bird]} + f(\text{[Image of a bird]}) = \text{[Image of a noisy bird]} \}$$

Gaussian noise, contrast changes, adversarial examples, etc.

Natural: images as they were recorded



New, unperturbed images.

Quantifying Robustness

Often in-distribution (“standard”) accuracy acts as a **confounder**.

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

Quantifying Robustness

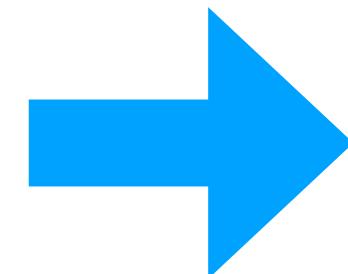
Often in-distribution (“standard”) accuracy acts as a **confounder**.

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

Quantifying Robustness

Often in-distribution (“standard”) accuracy acts as a **confounder**.

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

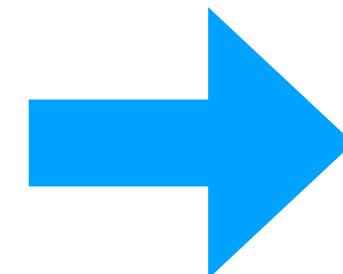


How do we compare models with different in-distribution accuracy?

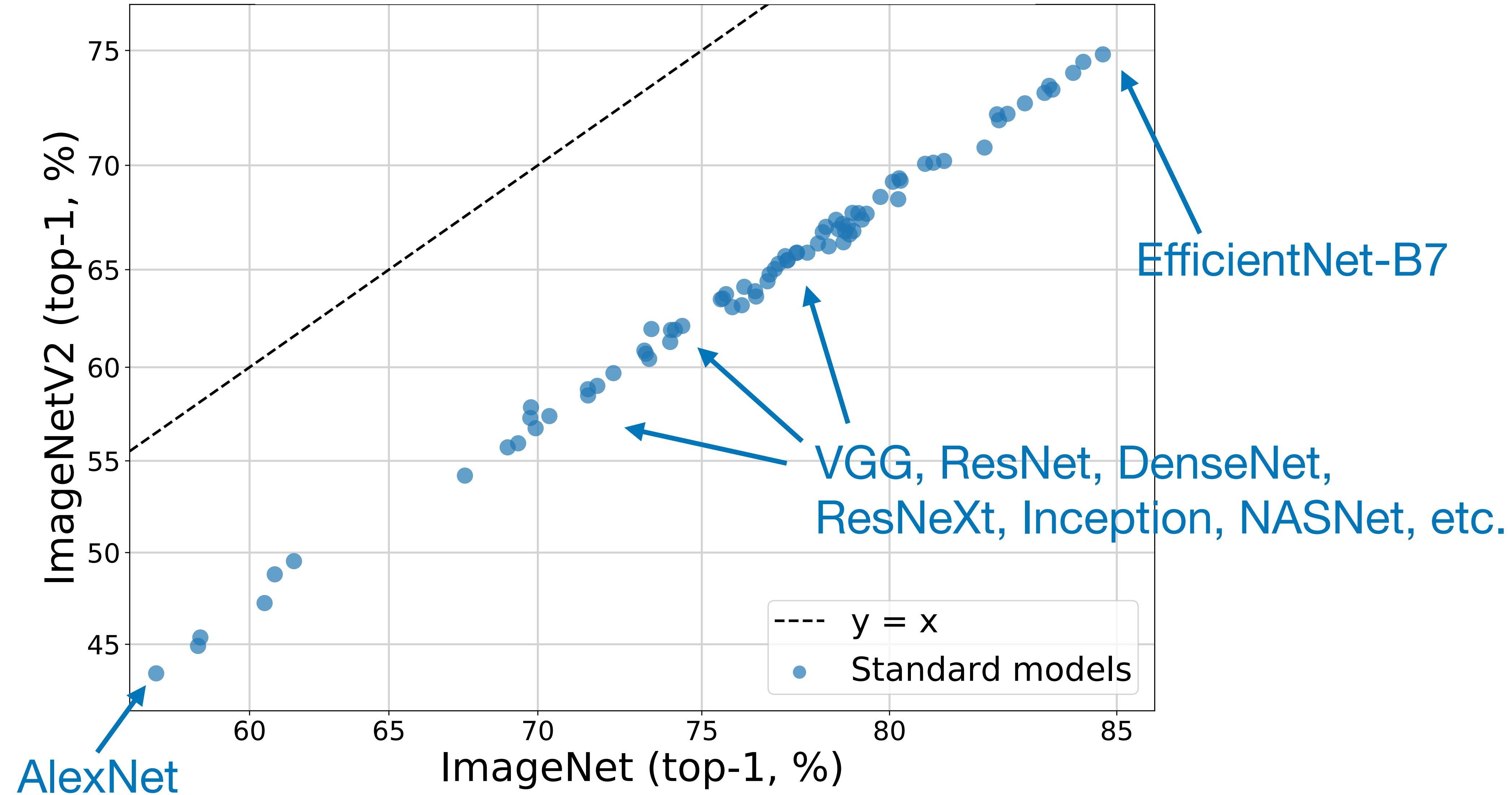
Quantifying Robustness

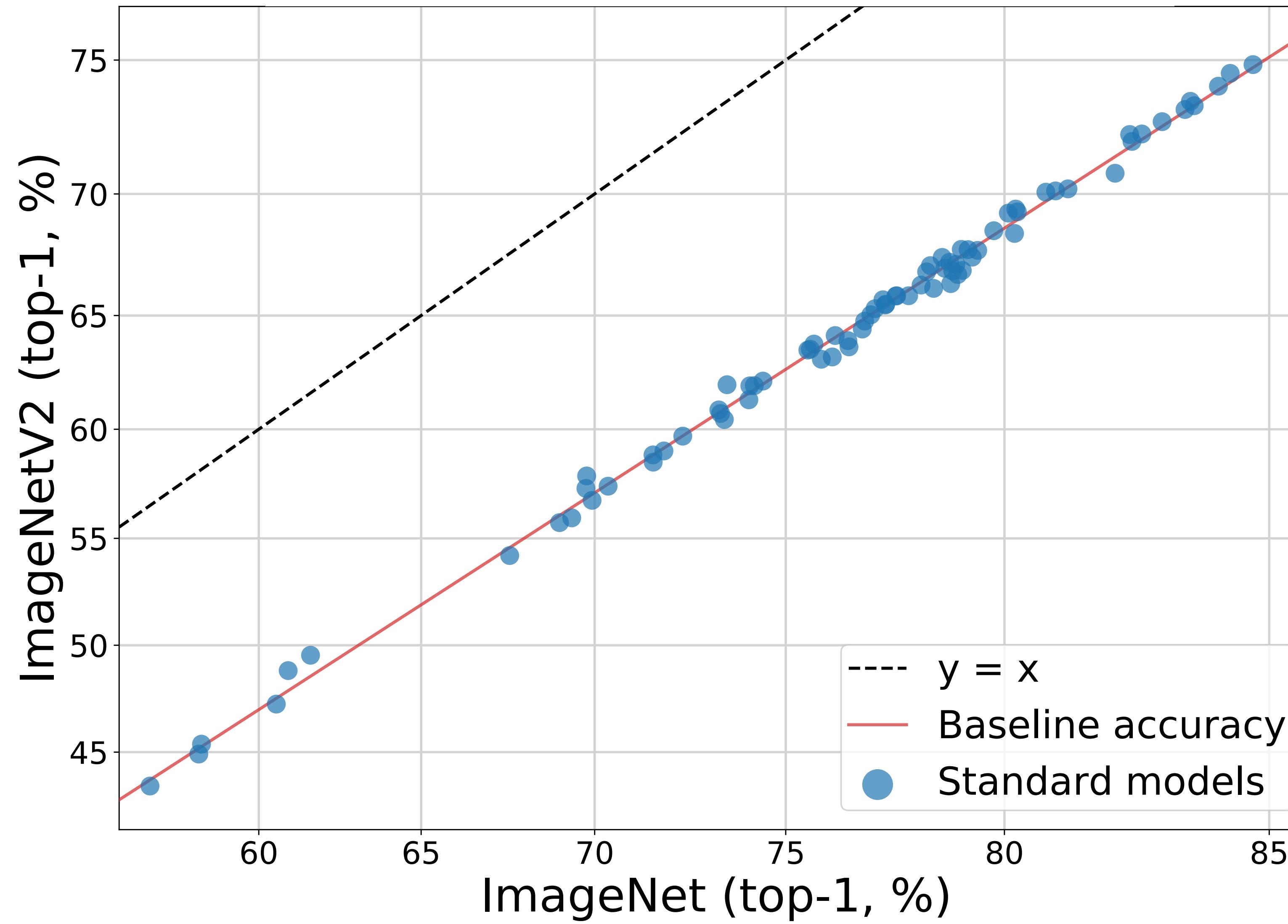
Often in-distribution (“standard”) accuracy acts as a **confounder**.

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

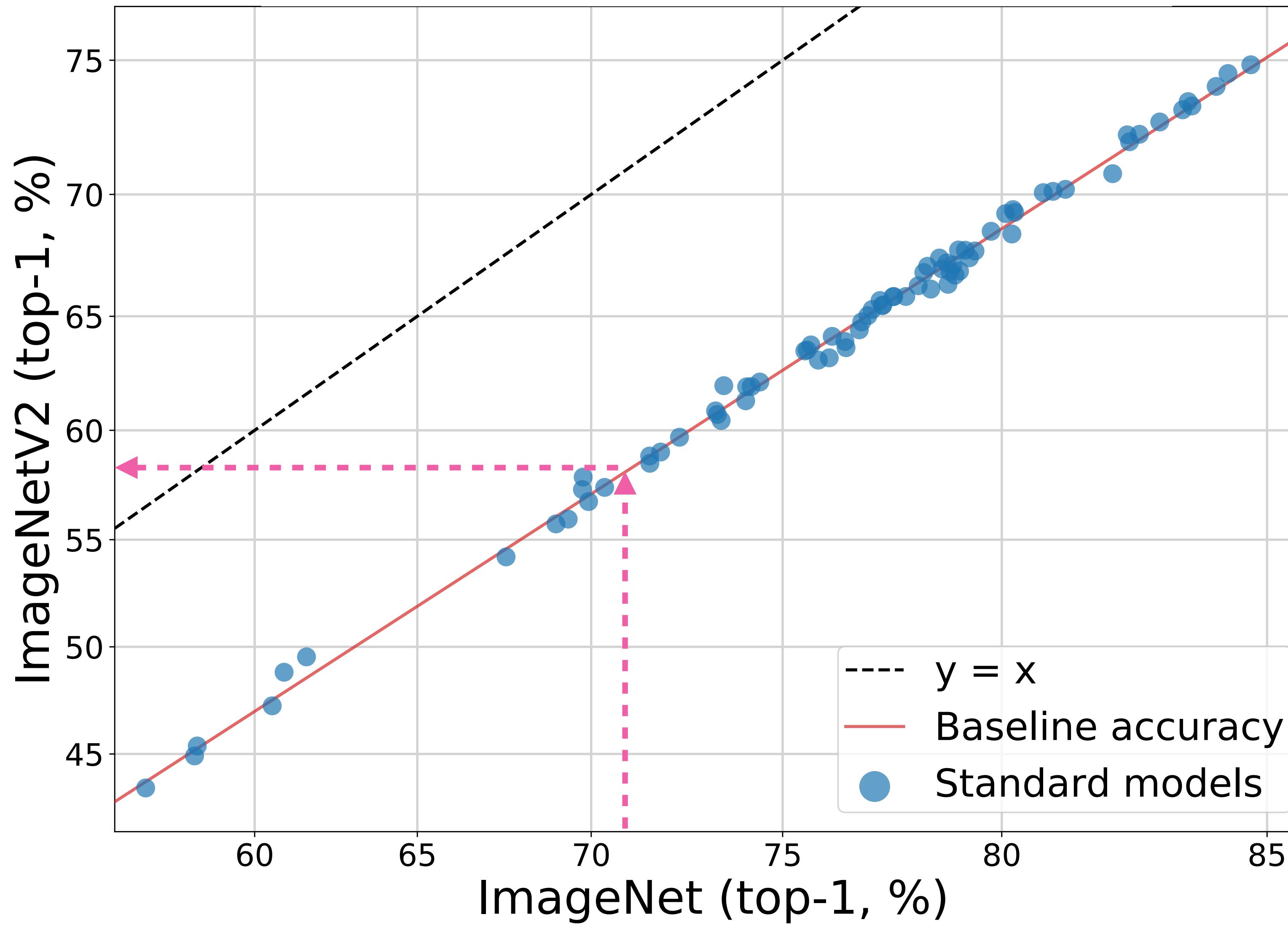


How do we compare models with different in-distribution accuracy?



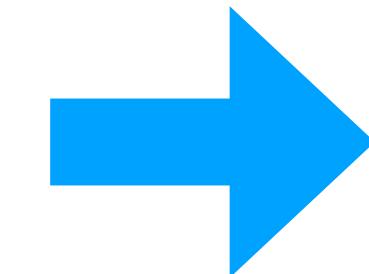
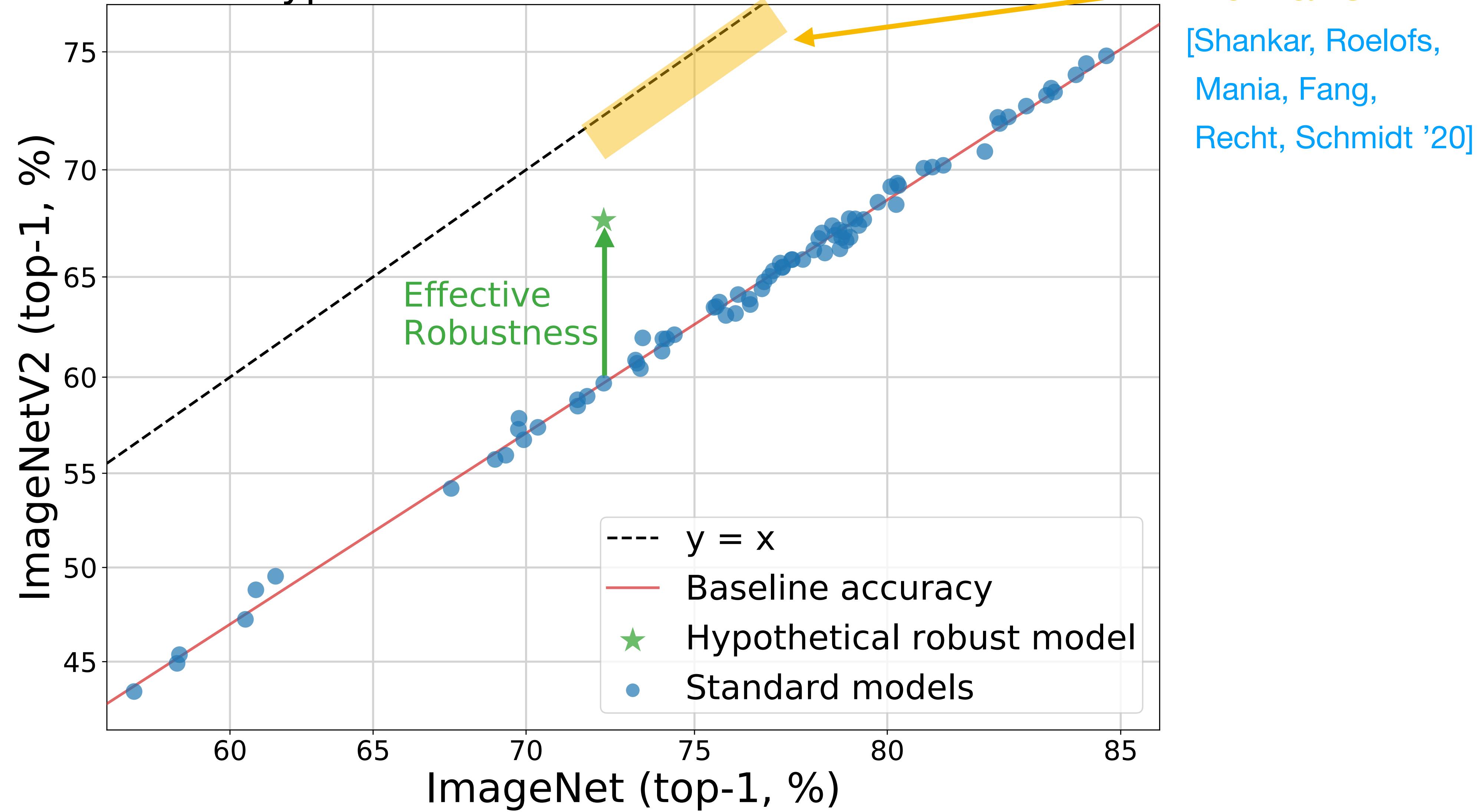


Expected out-of-distribution accuracy



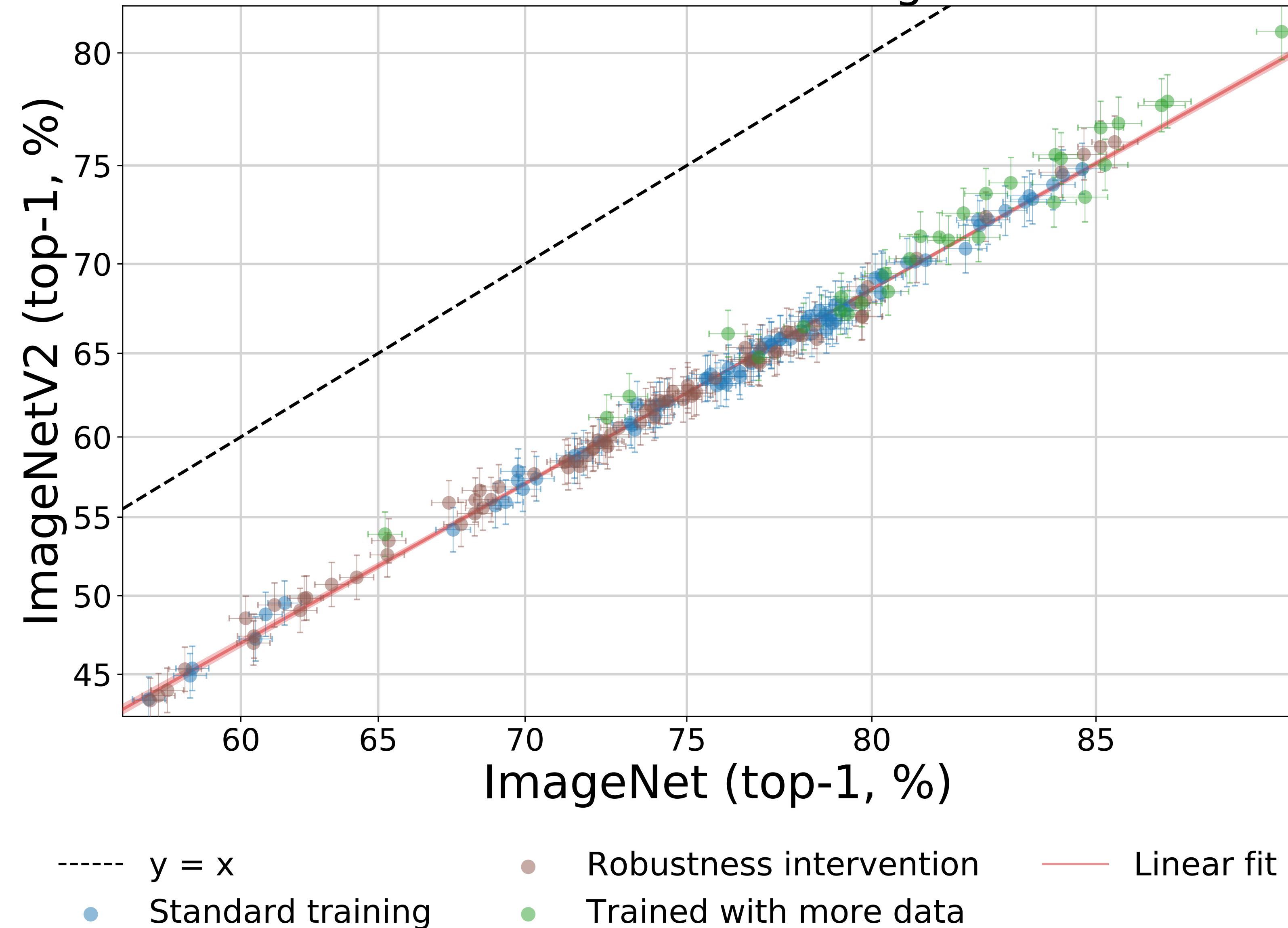
In-distribution accuracy

Hypothetical Robustness Intervention



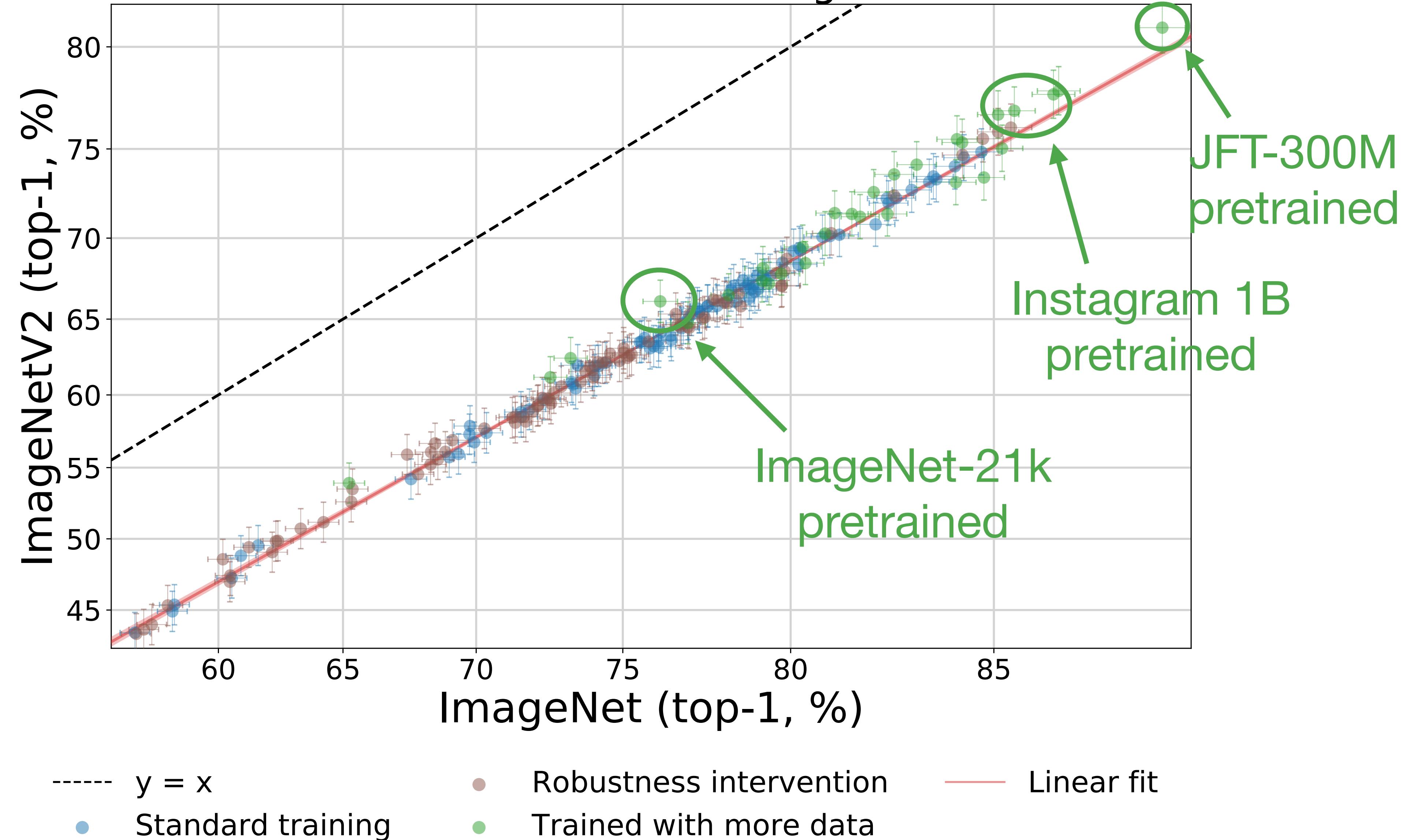
Do current models achieve effective robustness?

Distribution Shift to ImageNetV2



Training on (a lot) **more data** gives a small amount of effective robustness.

Distribution Shift to ImageNetV2



Training on (a lot) **more data** gives a small amount of effective robustness.

ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models

Andrei Barbu*
MIT, CSAIL & CBMM

David Mayo*
MIT, CSAIL & CBMM

Julian Alverio
MIT, CSAIL

William Luo
MIT, CSAIL

Christopher Wang
MIT, CSAIL

Dan Gutfreund
MIT-IBM Watson AI

Joshua Tenenbaum
MIT, BCS & CBMM

Boris Katz
MIT, CSAIL & CBMM

Abstract

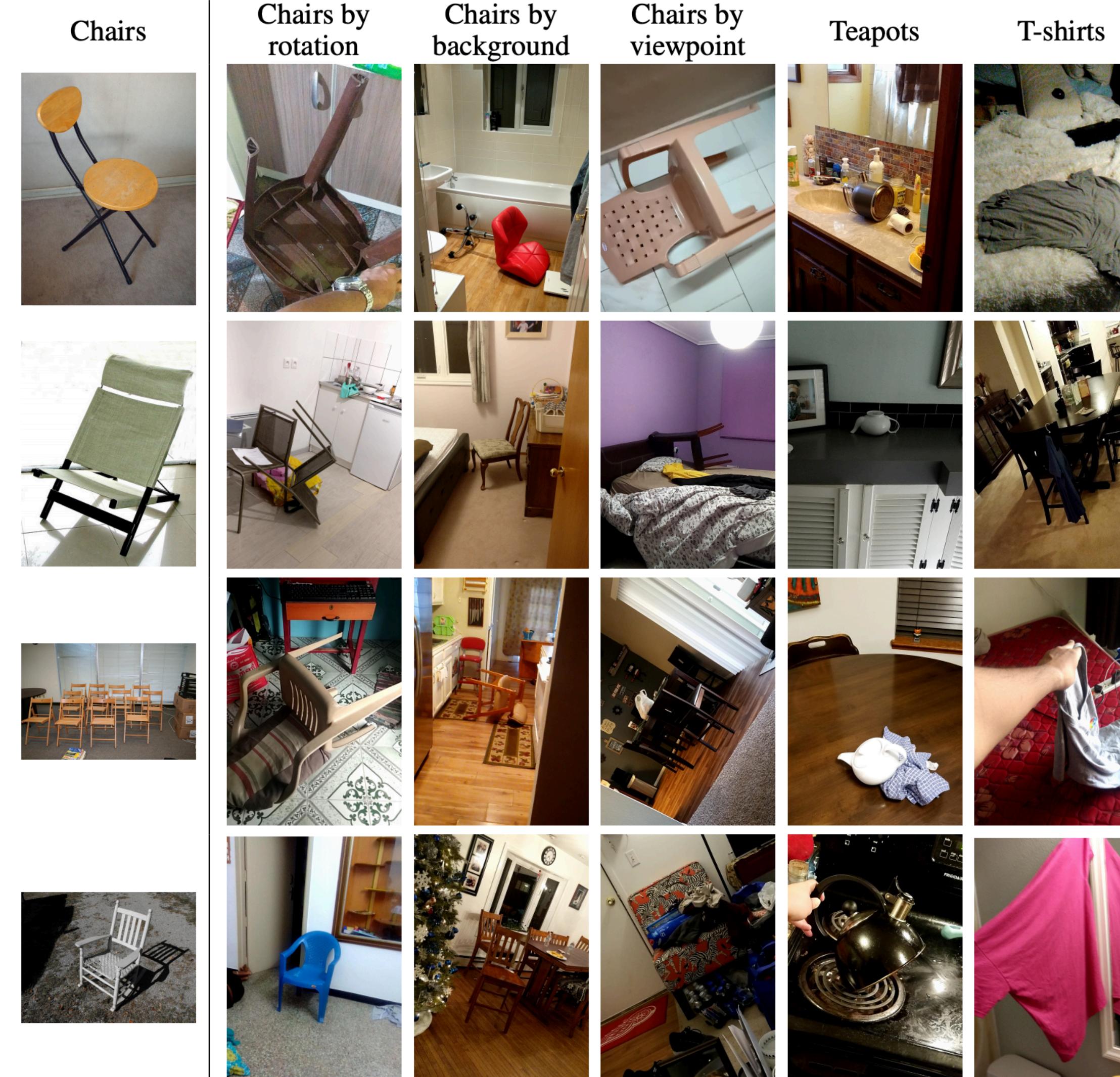
We collect a large real-world test set, ObjectNet, for object recognition with controls where object backgrounds, rotations, and imaging viewpoints are random. Most

Idea: Put Objects in Unusual Positions

ImageNet

Mainly object-centric and clean images

(collected from Flickr)



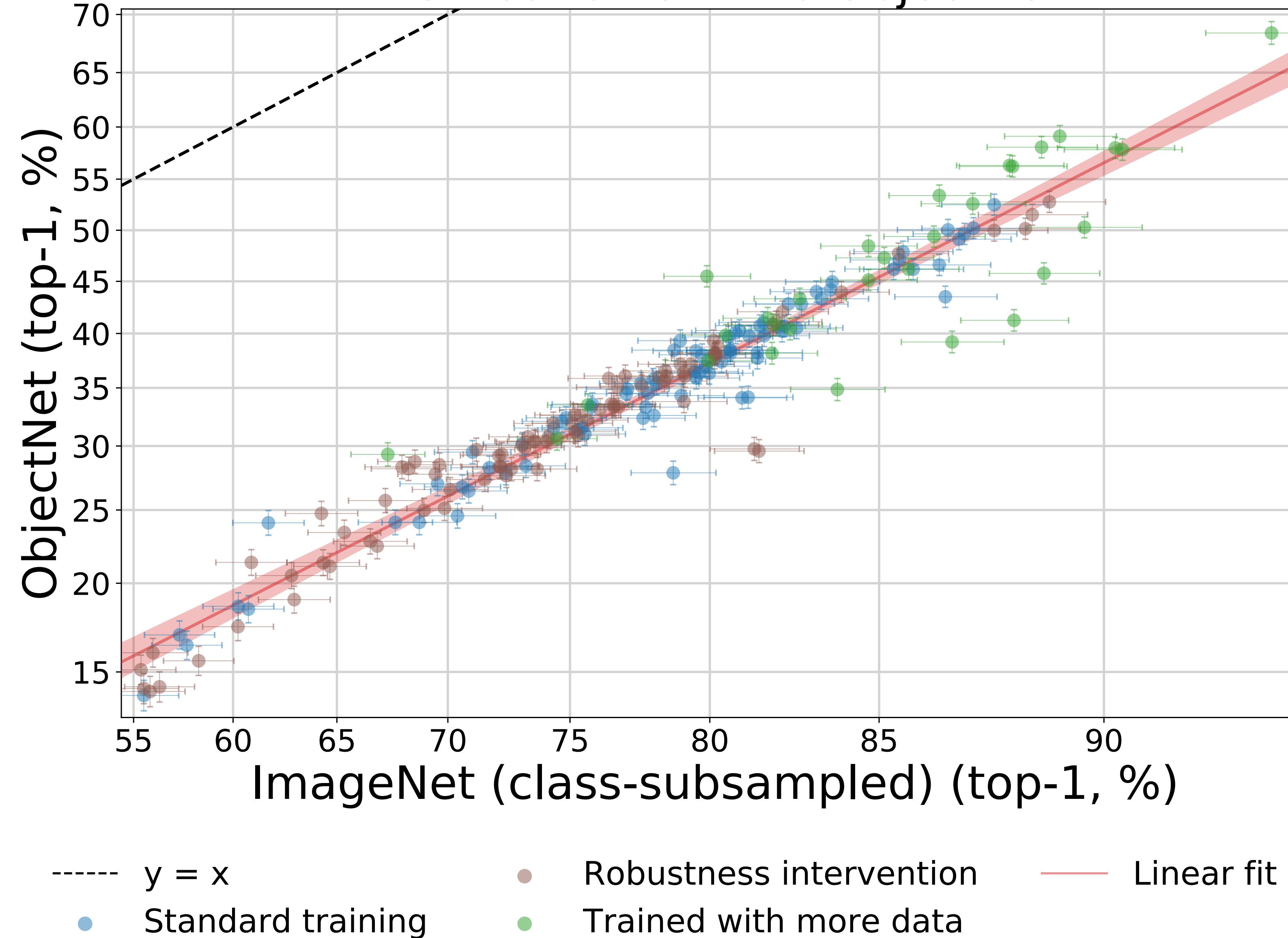
ObjectNet

Intentionally randomized:

- object poses
- locations
- etc.

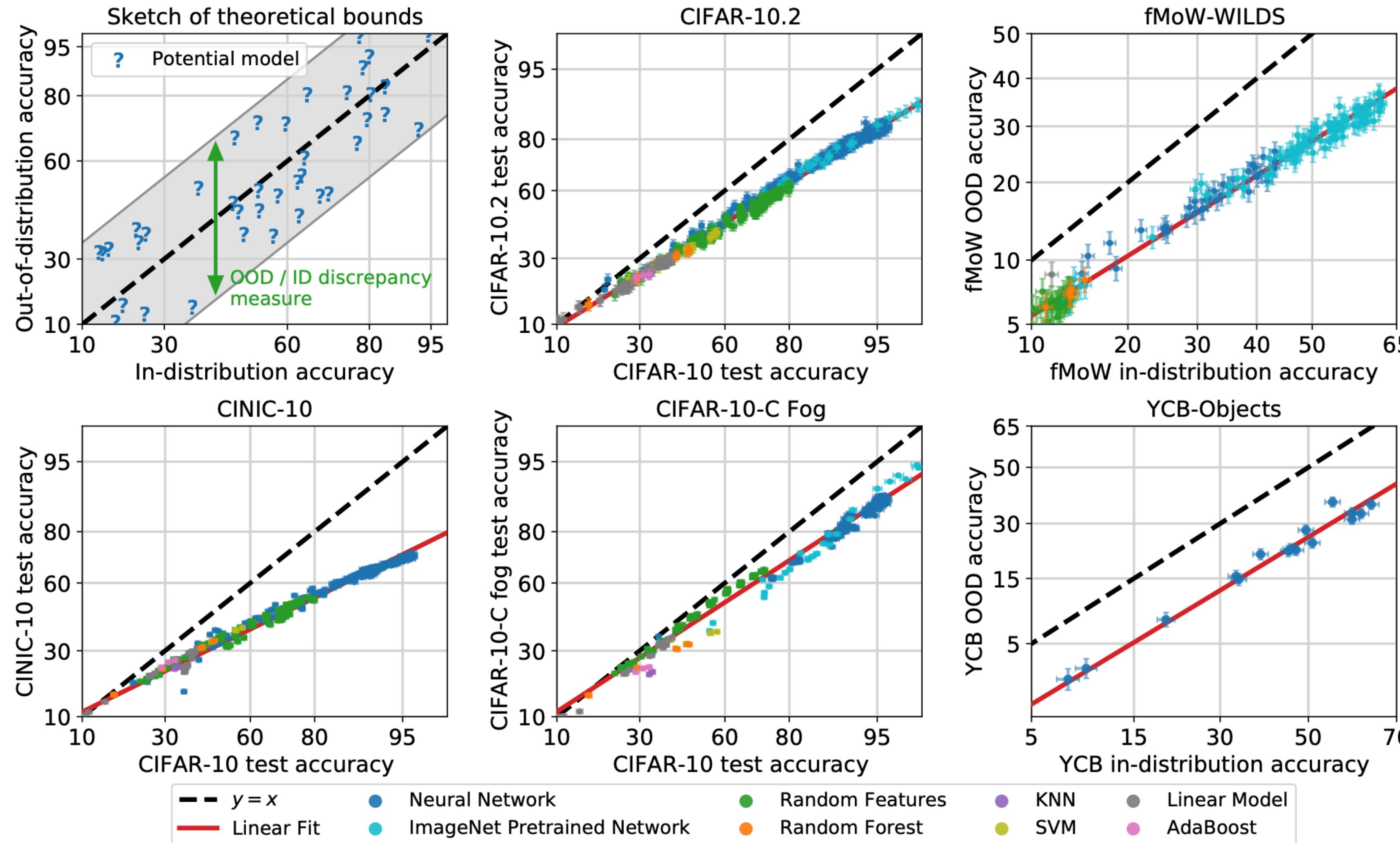
(collected via specific crowd worker annotations)

Distribution Shift to ObjectNet



Same trend: only **more data** gives effective robustness (but sometimes actually worse!)

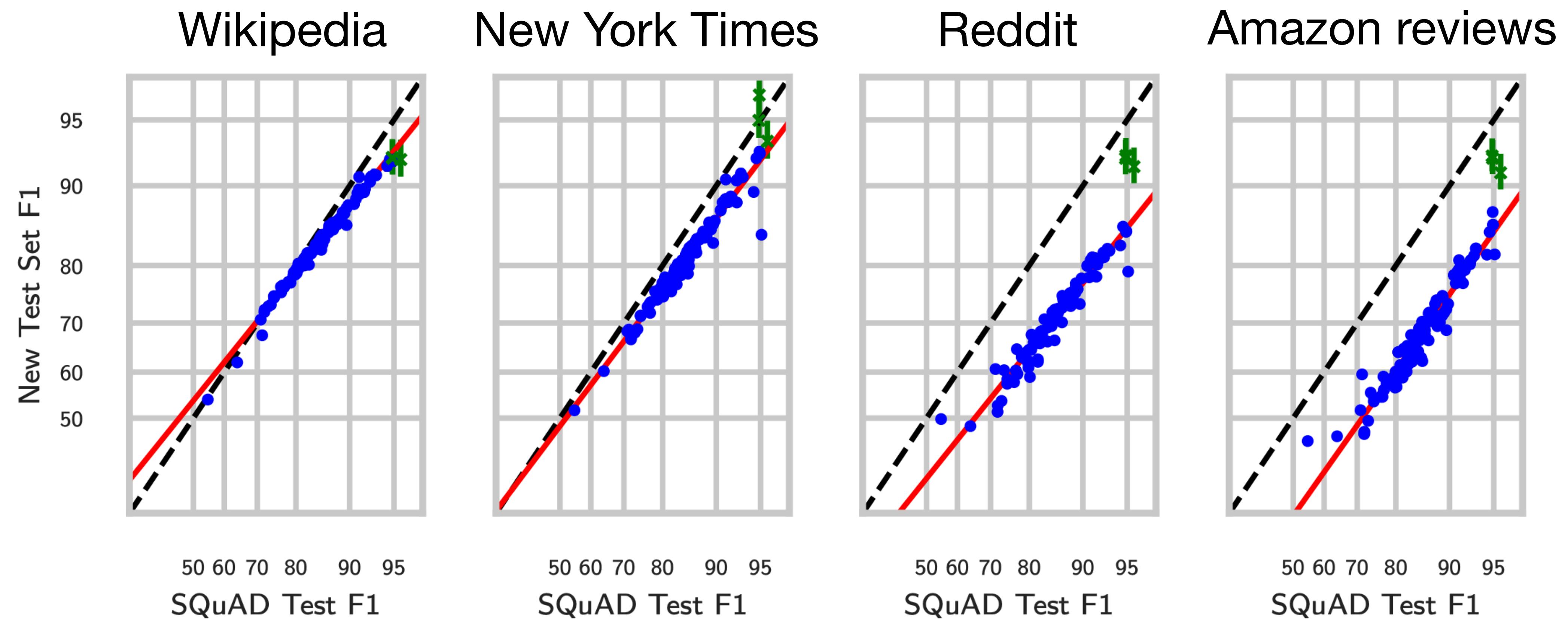
Linear trends beyond ImageNet



Beyond Image Classification

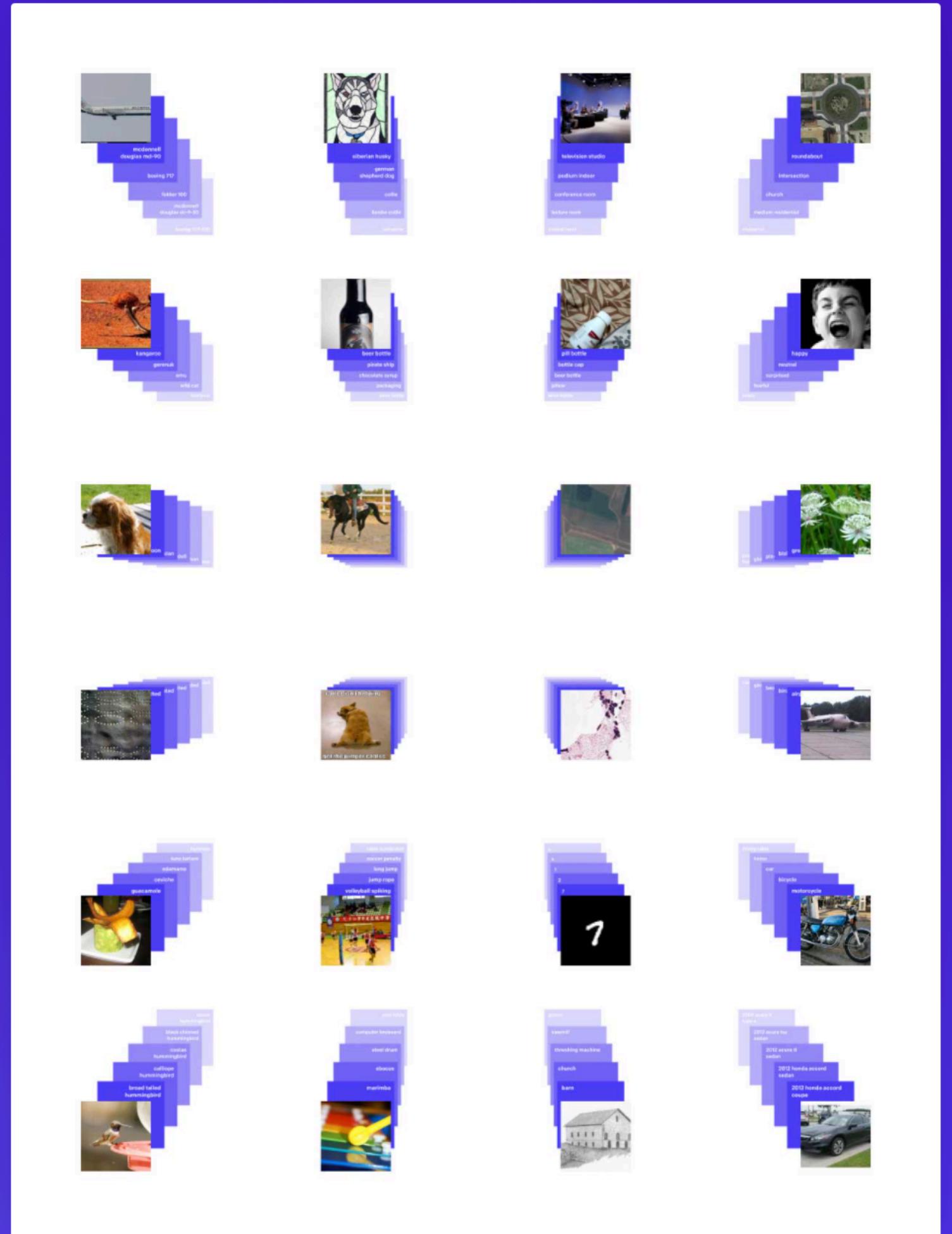
SQuAD (Stanford Question Answering Dataset): question answering on paragraphs

→ Similar trends in natural language processing. [Miller, Krauth, Recht, Schmidt '20]



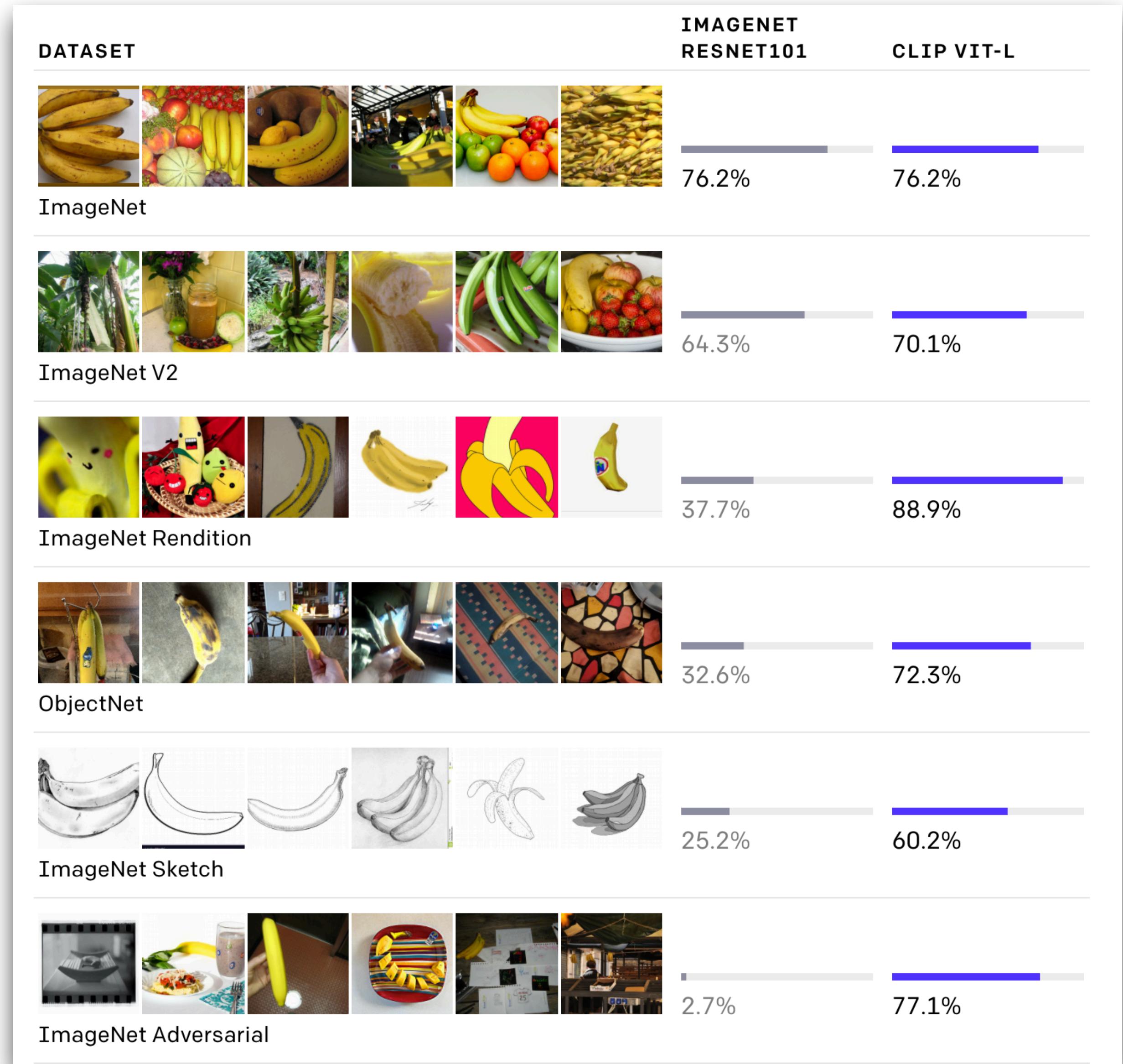
CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.



January 5, 2021

15 minute read



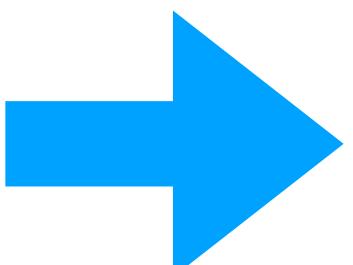
+6%

+51%

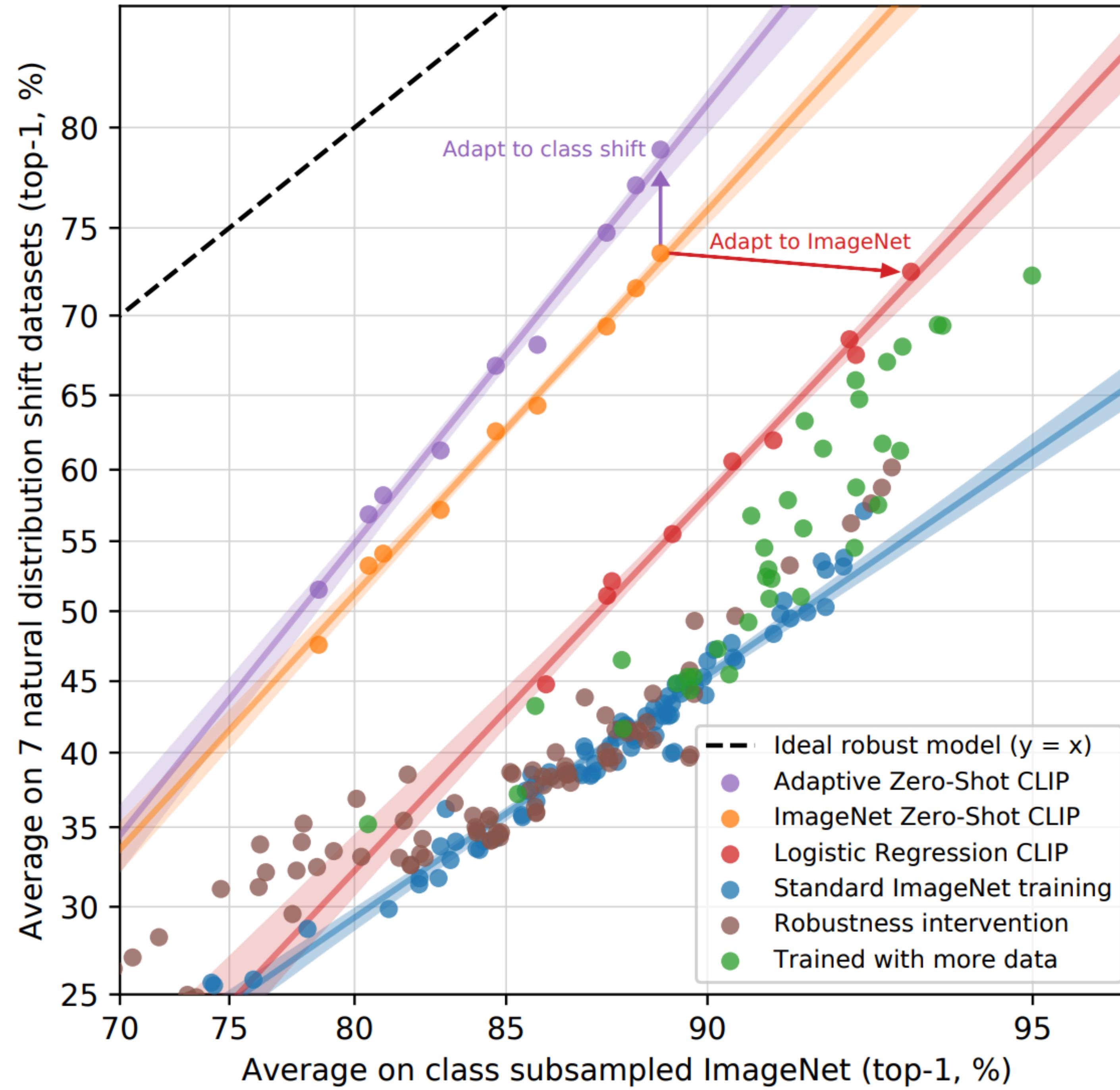
+40%

+35%

+74%



Very large improvements in out-of-distribution robustness.



[Radford, Kim,
Hallacy, Ramesh,
Goh, Agarwal, Sastry,
Askell, Mishkin,
Clark, Krueger,
Sutskever '21]

1. How reliable are ML benchmark results? (Internal validity)
2. Do benchmark results transfer across learning problems? (External validity)
3. Do benchmark results transfer across test distributions? (External validity)
4. Course projects

Project content

In principle, anything broadly related to the course is welcome (ongoing research OK).

Talk to me if you have a specific idea in mind (please send me a message).

In the following, we'll sketch out a default template for projects that will be interesting across multiple domains.

Short version:

Project content

In principle, anything broadly related to the course is welcome (ongoing research OK).

Talk to me if you have a specific idea in mind (please send me a message).

In the following, we'll sketch out a default template for projects that will be interesting across multiple domains.

Short version:

Scatter plots

Why scatter plots?

A lot of questions in this course revolve around what ML evaluations **mean**.

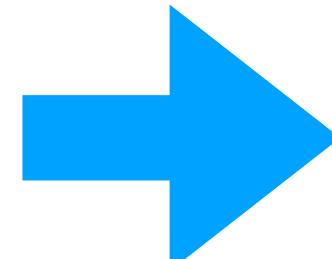
How do measurements acquire **meaning**?

Some measurements are **directly** relevant, e.g., performance in an application

Other measurements acquire meaning via **connections** to other quantities.

(E.g., most physical quantities like mass, etc.)

A simple but very useful way to understand the connection between two quantities is to run experiments, measure both, and plot them



Scatter plots

Two types of experiments

Transfer of performance across datasets

Pick two or more datasets / benchmarks in your domain of expertise
(e.g., pairs of datasets that seem more or less related)

Build a testbed with a range of different methods
Different architectures, pre-training datasets, optimizers, etc.
The testbed should cover a large performance range on the source dataset.

Evaluate the testbed on the source dataset and the other datasets
(With fine-tuning, training from scratch, etc.)

Make a scatter plot and interpret the results
Are there consistent trends? What do they tell us about the datasets?

Two types of experiments

Performance under **distribution shift**

Pick one dataset in your domain of expertise and create one or more dist. shifts
(e.g., from other datasets, structured splits of datasets, new data, etc.)

Build a testbed with a range of different methods

Different architectures, pre-training datasets, optimizers, etc.

The testbed should cover a large performance range on the standard test set

Evaluate the testbed on the standard test set and the other test sets

No fine-tuning / further training!

Make a scatter plot and interpret the results

Are there consistent trends? What do they tell us about the datasets?

Project Logistics

Group size 1 - 3 people

Project proposals due Thursday next week (October 14)

Talk to us before then if you have questions

Proposals should be **1 - 2 pages** of text describing the experiment (datasets used, models in the testbed, etc.) and contain **sketches of the key plots**.

One class (likely October 19): **short presentations** what everyone is planning to do.

Last two classes of the quarter: **project presentations** with results

Final deliverable: project report + GitHub repository (due date likely Dec 11).

Project Logistics

Group size 1 - 3 people

Project proposals due Thursday next week (October 14)

Talk to us before then if you have questions

Proposals should be **1 - 2 pages** of text describing the experiment (datasets used, models in the testbed, etc.) and contain **sketches of the key plots**.

One class (likely October 19): **short presentations** what everyone is planning to do.

Last two classes of the quarter: **project presentations** with results

Final deliverable: project report + GitHub repository (due date likely Dec 11).

Questions?