training on the test set and other heresies

Benjamin Recht
University of California, Berkeley

Old ML Conventional Wisdom

- Good prediction balances bias and variance.
- You should not perfectly fit your training data as some in-sample errors can reduce out-of-sample error.
- High capacity models don't generalize.
- Optimizing to high precision harms generalization.
- Nonconvex optimization is hard in machine learning.

None of these are true.

Generalization in Machine Learning

Given: i.i.d. sample $S = \{z_1, \ldots, z_n\}$ from dist D

Goal: Find a good predictor function f

$$R[f] = \mathbb{E}_z \text{loss}(f; z)$$

Population risk
(test error)

unknown!

$$R_S[f] = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f; z_i)$$

Empirical risk
(training error)

Minimize using SGD!

Generalization error: $R[f] - R_S[f]$

How much empirical risk underestimates population risk

We can optimize R_S...

When is it a good proxy for R?

Fundamental Theorem of Machine Learning

$$R[f] = (R[f] - R_S[f]) + R_S[f]$$

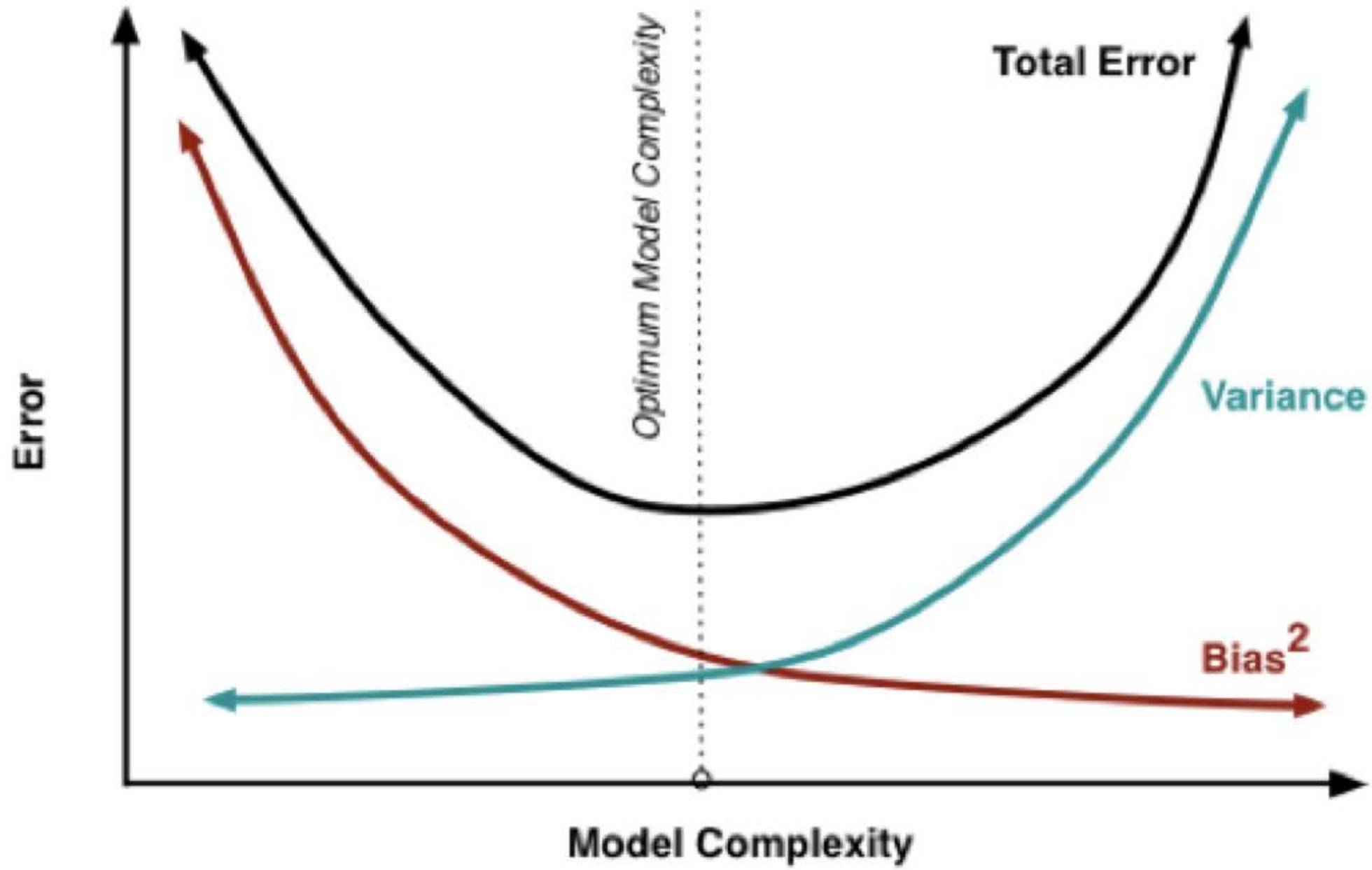
population risk | generalization error | training error

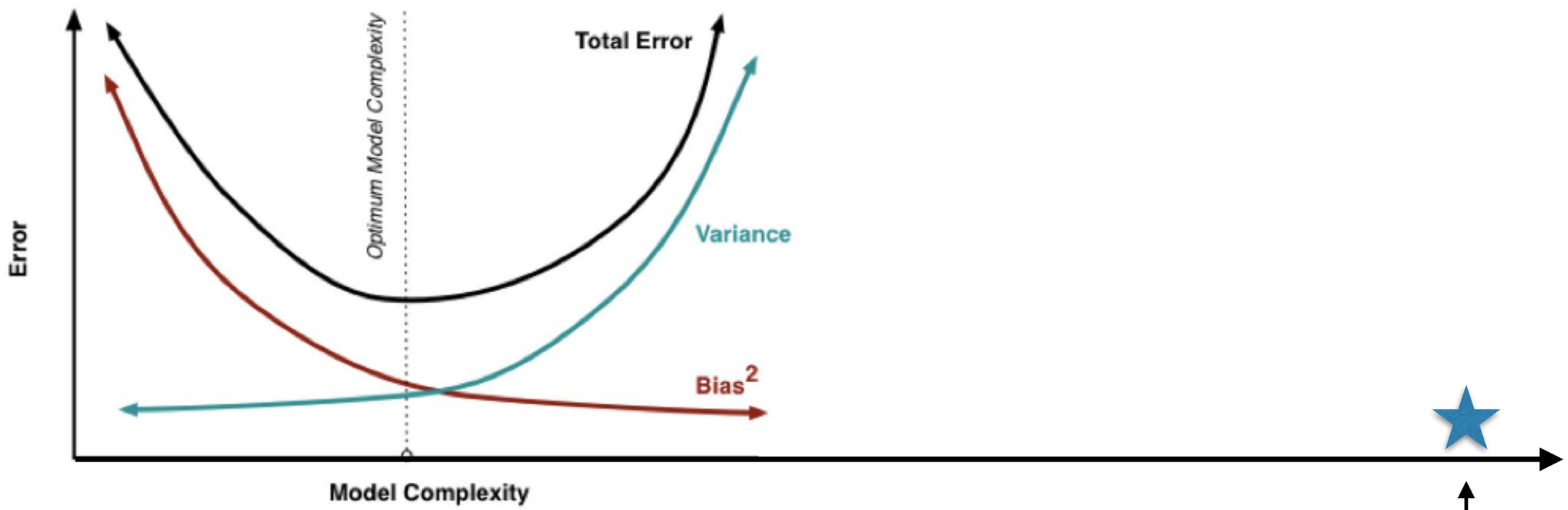
- small training error implies risk ≅ generalization error
- zero training error *does not imply* overfitting

$$R[f] = (R[f] - R[f_{\mathcal{H}}]) \quad \text{error vs best in class}$$

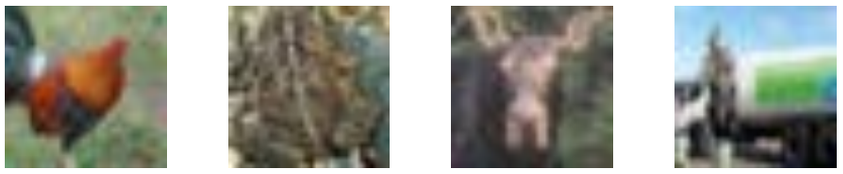
$$+ (R[f_{\mathcal{H}}] - R[f_\star]) \quad \text{approximation error}$$

$$+ R[f_\star] \quad \text{irreducible error}$$


Optimum Model Complexity
Total Error
Variance
Bias^2
Error
Model Complexity


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

Deep models



CIFAR10

n=50,000
d=3,072
k=10

What happens when I turn off the regularizers?

Model	parameters	p/n	Train ***loss***	Test ***error***
CudaConvNet	145,578	2.9	0	23%
CudaConvNet (with regularization)	145,578	2.9	0.34	18%
MicroInception	1,649,402	33	0	14%
ResNet	2,401,440	48	0	13%

Zhang, Bengio, Hardt, R., Vinyals, 2017

Machine Learning's Open Dirty Secrets

Given: i.i.d. sample $T = \{z_1,\ldots,z_n\}$ and $H = \{z'_1,\ldots,z'_m\}$ from dist D

Goal: Find a good predictor function f

Myth

$$R_T[f] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{loss}(f; z_i)$$

(training error)

Minimize using SGD!

$$R_H[f] = \frac{1}{m}\sum_{j=1}^{m} \mathrm{loss}(f; z'_j)$$

(test/holdout error)

Only look once!!!

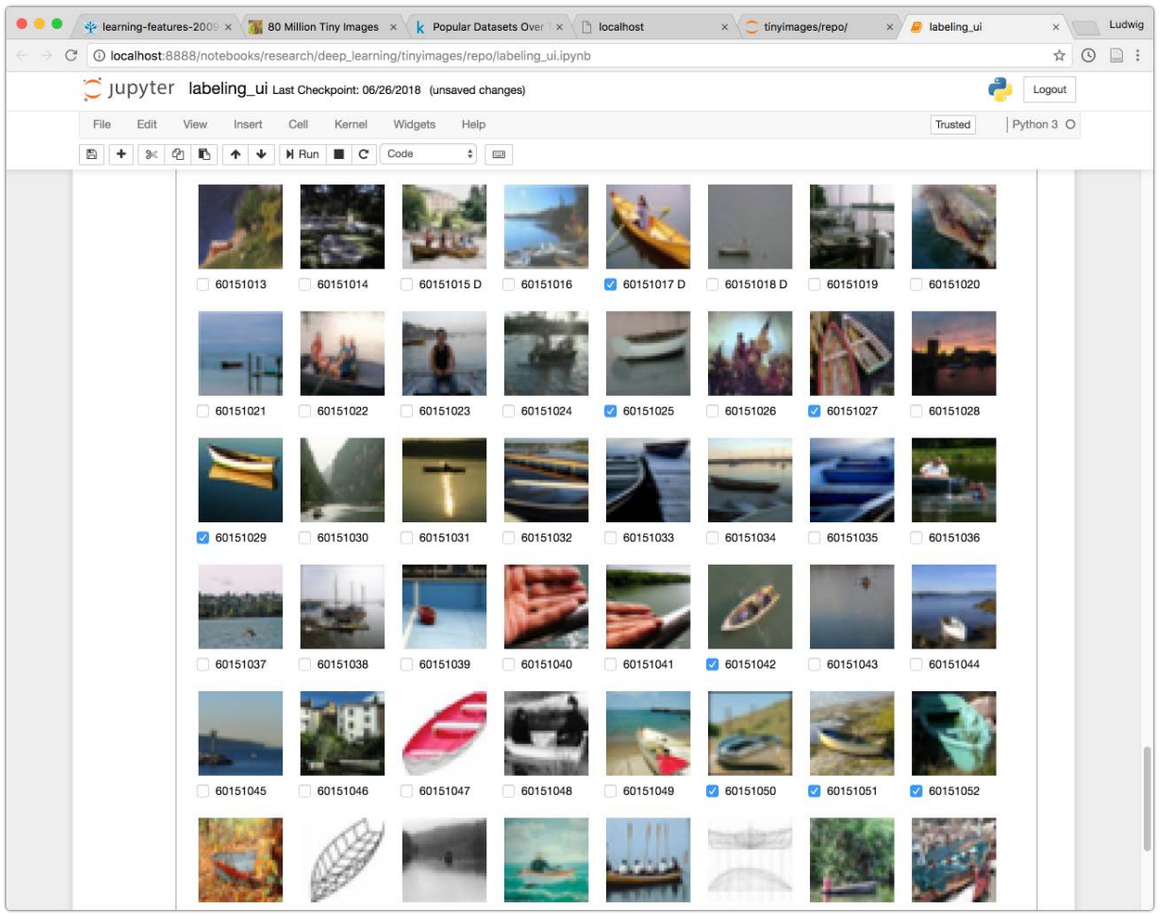
Reality

$$\begin{aligned} \text{minimize} \quad & R_H[f] \\ \text{subject to} \quad & R_T[f] \leq \epsilon \\ & f \in \mathcal{F} \end{aligned}$$

$$\mathcal{F} = \left\{ \begin{array}{c} \text{functions computable} \\ \text{before the heat death} \\ \text{of the universe} \end{array} \right\}$$

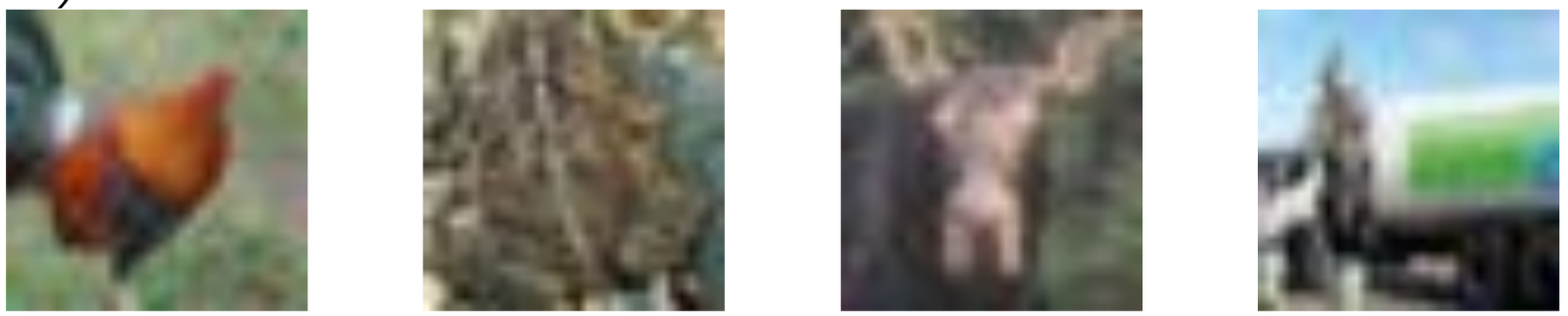
Tiny Images Dataset

- 79,302,017 images
- 32x32x3
- 400GB
- 75,062 non-abstract nouns (WordNet)
- Collected by [Torralba, Fergus, Freeman'08]
- Collected via queries to image search engines



CIFAR10

- 10 classes: airplane, car, bird, cat, deer, dog, frog, horse, ship, truck.
- 32 x 32 color images
- Used to prototype models for imagenet.
- (It is not true that something that is good on one is good on the other.)



CIFAR-10 Creation Process

Detailed description in [Krizhevsky'09]:

1. Find relevant keywords for each class from WordNet
   (e.g., "tabby_cat", "tabby", "domestic_cat", etc. for class "cat")
2. Present candidate images from TinyImages to student labelers
3. CIFAR-10 researchers remove unsuitable images
4. Remove near-duplicates
5. Randomly split into class-balanced train and test sets

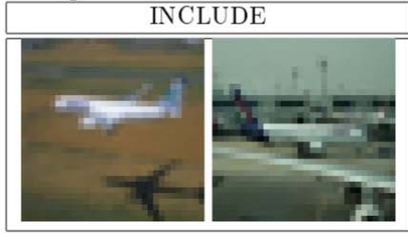
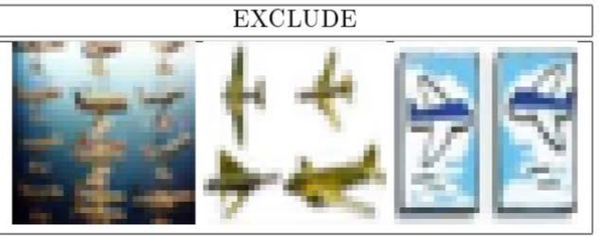
Labeler instruction sheet

Criteria for deciding whether to include an image

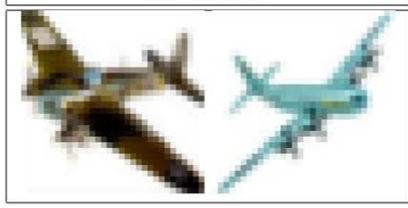
1. The main test is: Would you be quite likely to say the category name if asked to give a single basic category to describe the main object in the image?

2. It's worse to include one that shouldn't be included than to exclude one. False positives are worse than false negatives.

3. If there is more than one object that is roughly equally prominent, reject even if they are all of the right class.

INCLUDE	EXCLUDE
	

4. If it is a line drawing or cartoon, reject. You can accept fairly photorealistic drawings that have internal texture.

INCLUDE	EXCLUDE
	

5. Do not reject just because the viewpoint is unusual or the object is partially occluded (provided you think you might have assigned the right label without priming). We want ones with unusual

CIFAR-10 State of the Art

Year	Model	Test accuracy
2009	Raw pixels	37.3%
2009	RBM	64.8%
2011	Random features	79.6%
2012	AlexNet	88.5%
2014	VGG	92.8%
2015	ResNet	93.5%
2016	Wide ResNet	95.9%
2017	Shake Shake	97.1%

Can match this with "shallow" learning.

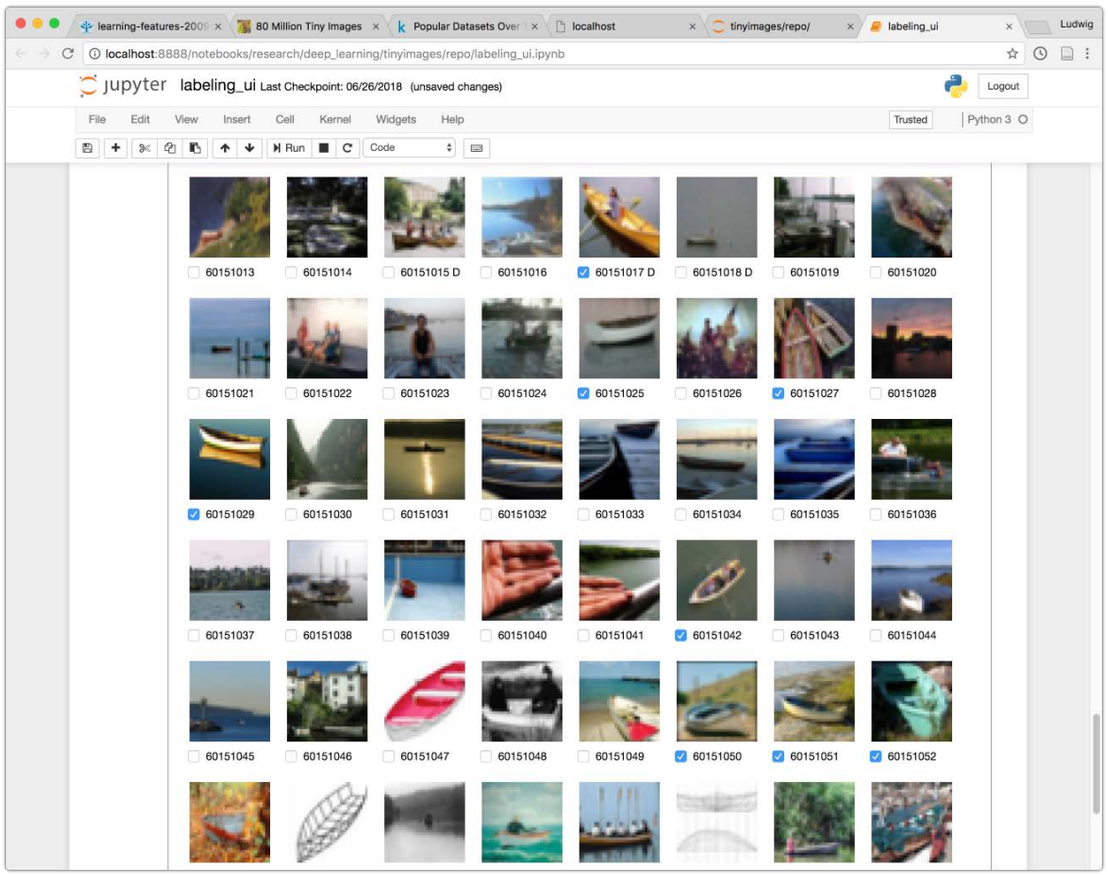
Deeeeep networks

Is this overfitting?

Building a *New Test Set*

CIFAR-10 is a subset of the **Tiny Images** dataset

- Collected by [Torralba, Fergus, Freeman'08]
- 80 million images
- Organized into 75,000 keywords (WordNet)
- Collected via queries to image search engines



Can we get an i.i.d. resampling?

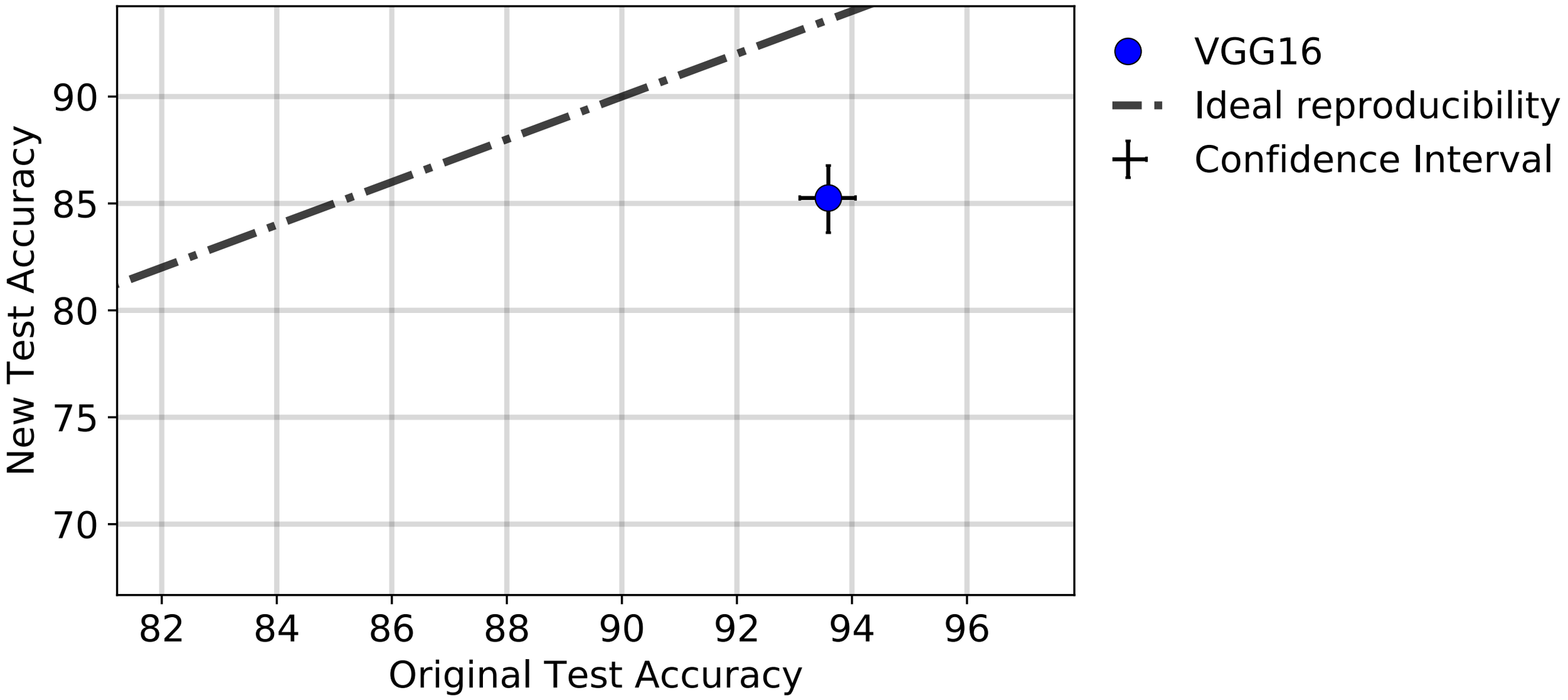
Roelofs, Schmidt, Shankar, R. 2018

CIFAR 10.1 Creation Process

Target size: 2,000 new images
(➡ confidence intervals for accuracy ±1%)

1. Determine keyword-level distribution (top 25 keywords per class)
2. Select suitable images for each keyword from Tiny Images ("student" labeler)
3. Sub-select images for each keyword from Tiny Images ("researcher" labeler)
4. Select a random sample from the candidate images
5. Remove near-duplicates and repeat until convergence

No classifier was evaluated on the data until the new test set was complete.

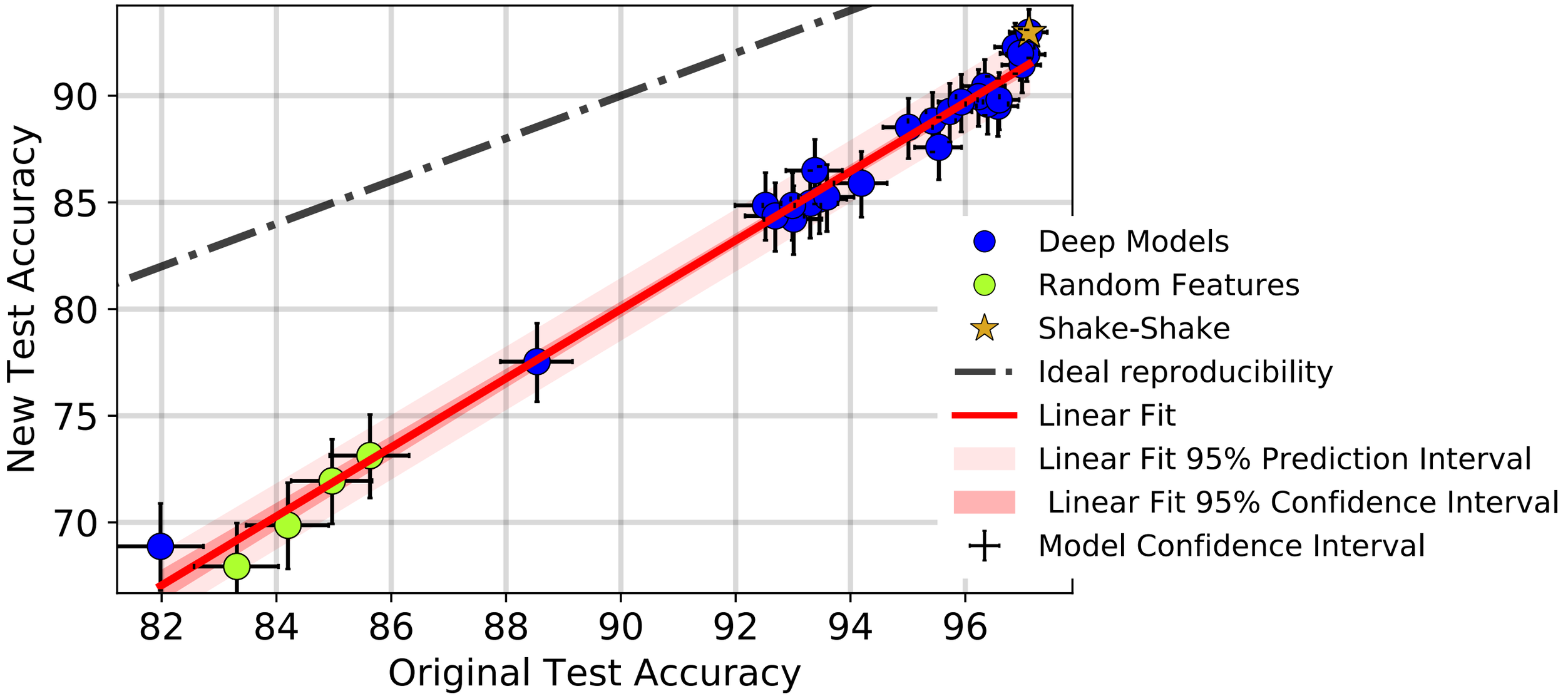


VGG16: 93.6% (original) ➡ 85.3% (new) 8% drop

CIFAR-10 State of the Art

Year	Model	Test accuracy
2009	Raw pixels	37.3%
2009	RBM	64.8%
2011	Random features	79.6%
2012	AlexNet	88.5%
2014	VGG	92.8%
2015	ResNet	93.5%
2016	Wide ResNet	95.9%
2017	Shake Shake	97.1%

Deeeeep networks



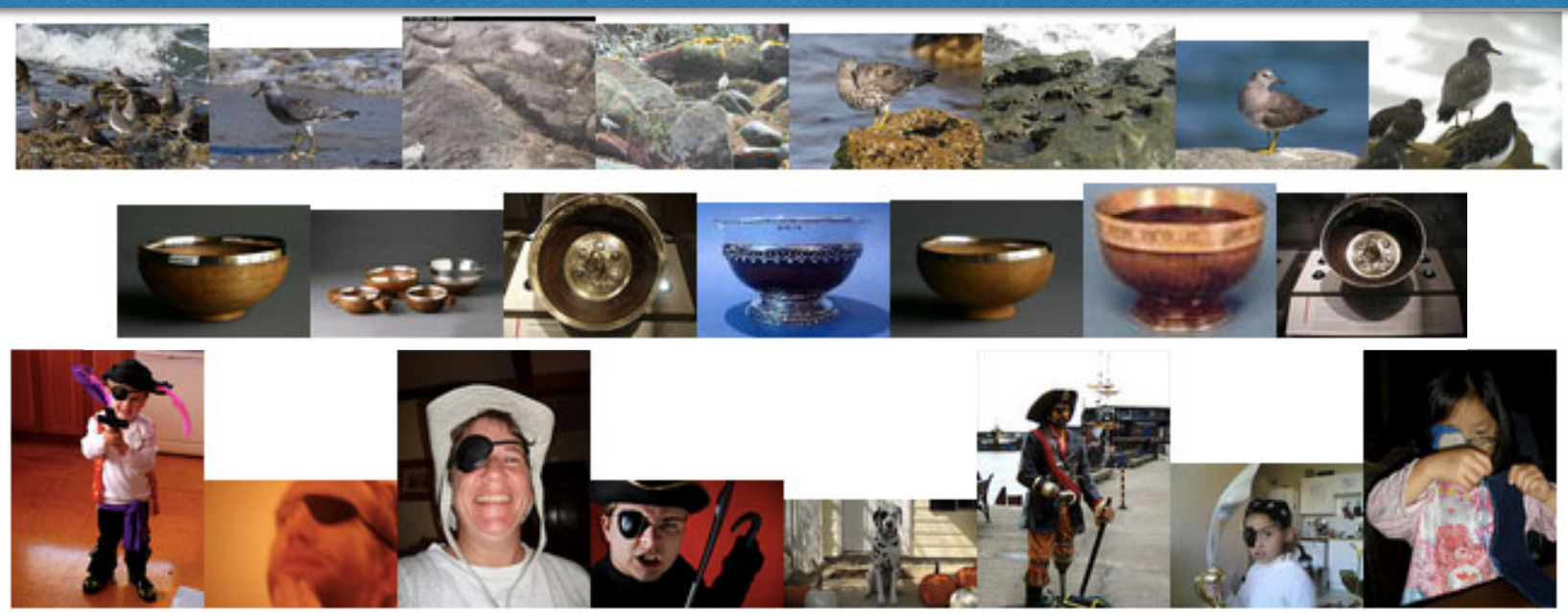
VGG16:	93.6% (original)	➡	85.3% (new)	8% drop
Random Features:	85.6% (original)	➡	73.1% (new)	12% drop
Shake-Shake:	97.1% (original)	➡	93.0% (new)	4% drop




IMAGENET

- Introduced in [Deng, Dong, Socher, Li, Li, Fei-Fei'09]
- organized according to the "WordNet hierarchy"
- 1.2 million training images, 50k validation images
- RGB color images with around 500 x 400 pixels
- 1,000 classes (about 150 dog breeds)

Can we get an i.i.d. resampling of imagenet too?



This research study is being conducted by Ben Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar from UC Berkeley. For questions about this study, please contact ludwig@berkeley.edu and roelofs@cs.berkeley.edu. In this study, we will ask you to indicate whether given images belong to a certain object category. Occasionally, the images may contain disturbing or adult content. We would like to remind you that participation in our study is voluntary and that you can withdraw from the study at any time.

Which of these images contain at least one object of type

bow

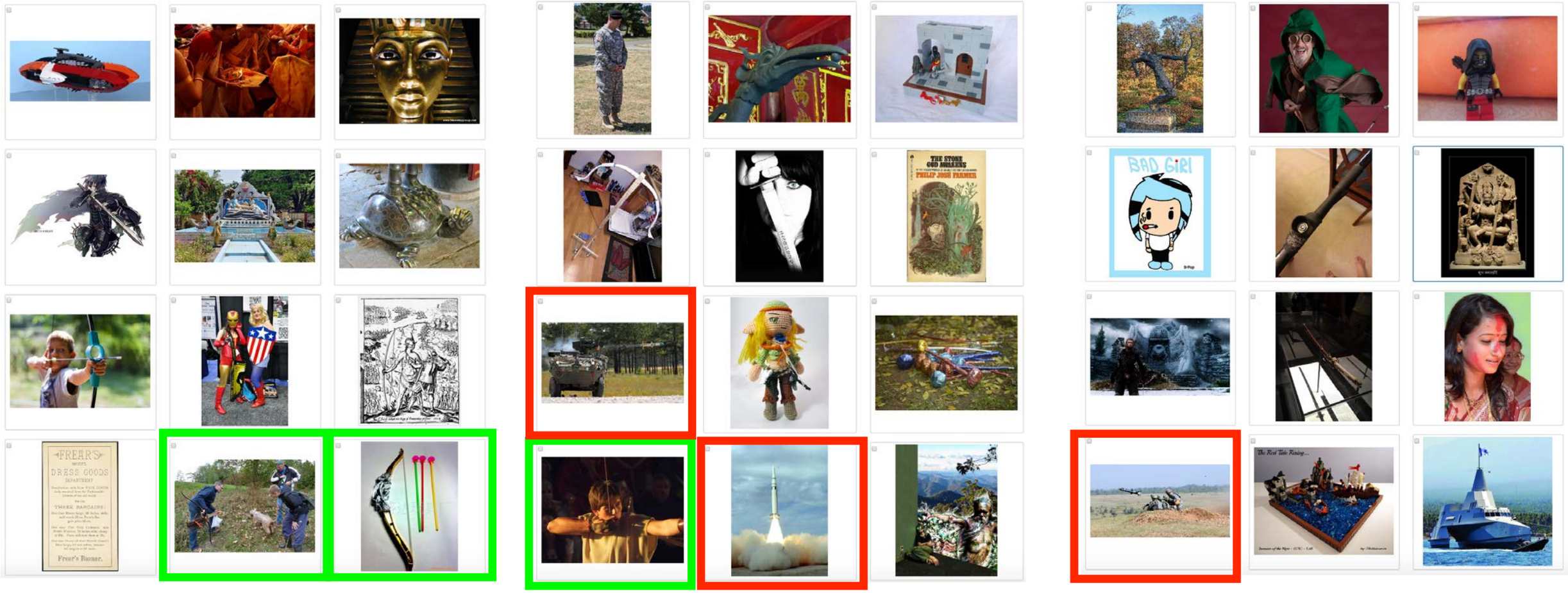
Definition: a weapon for shooting arrows, composed of a curved piece of resilient wood with a taut cord to propel the arrow

Task:
For each of the following images, check the box next to an image if it contains at least one object of type *bow*. Select an image if it contains the object regardless of occlusions, other objects, and clutter or text in the scene. Only select images that are photographs (**no drawings or paintings**).

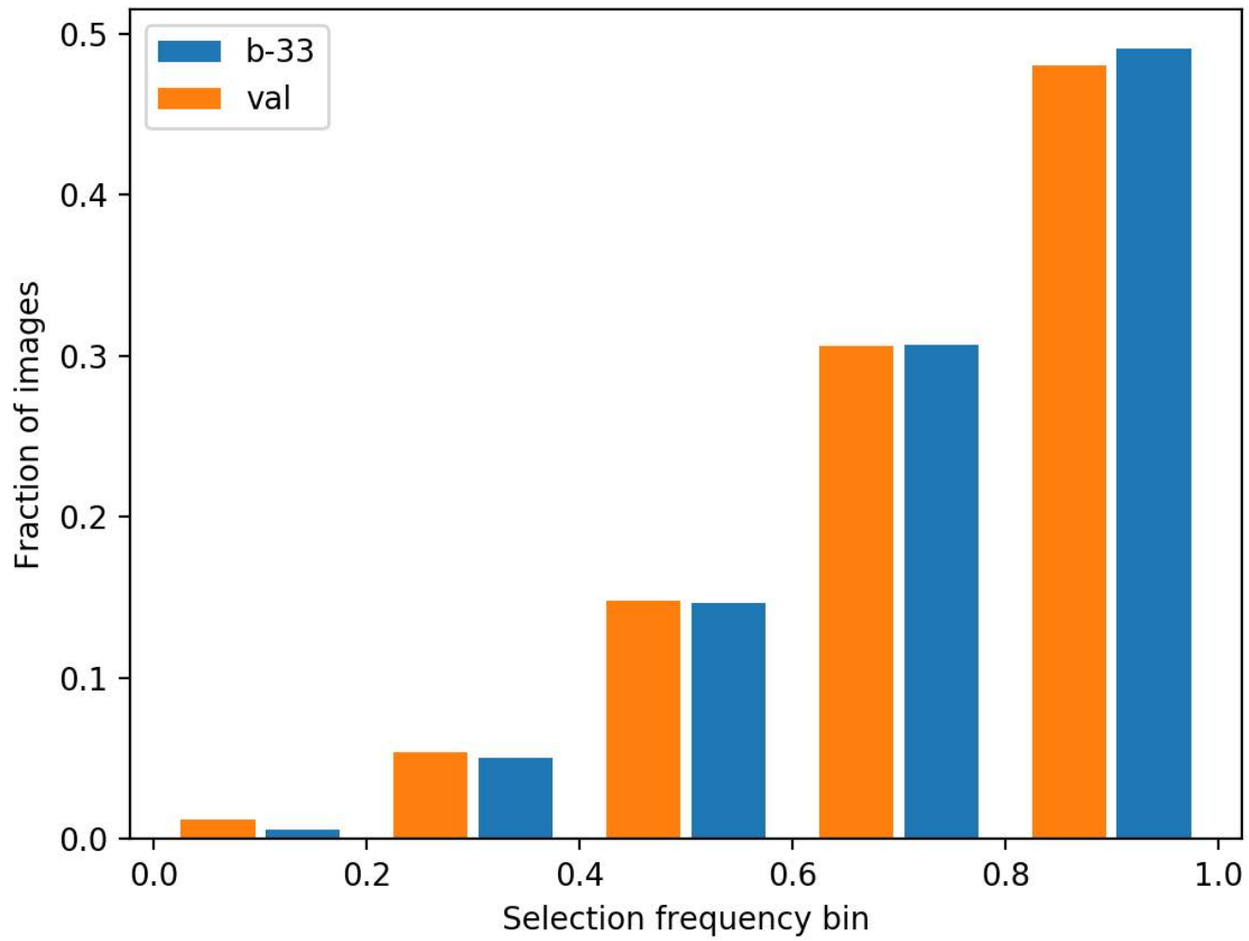
Please make accurate selections!

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/Bow_and_arrow

If it is impossible to complete a HIT due to missing data or other problems, please return the HIT.



Submit


b-33
val
Fraction of images
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Selection frequency bin

1.0

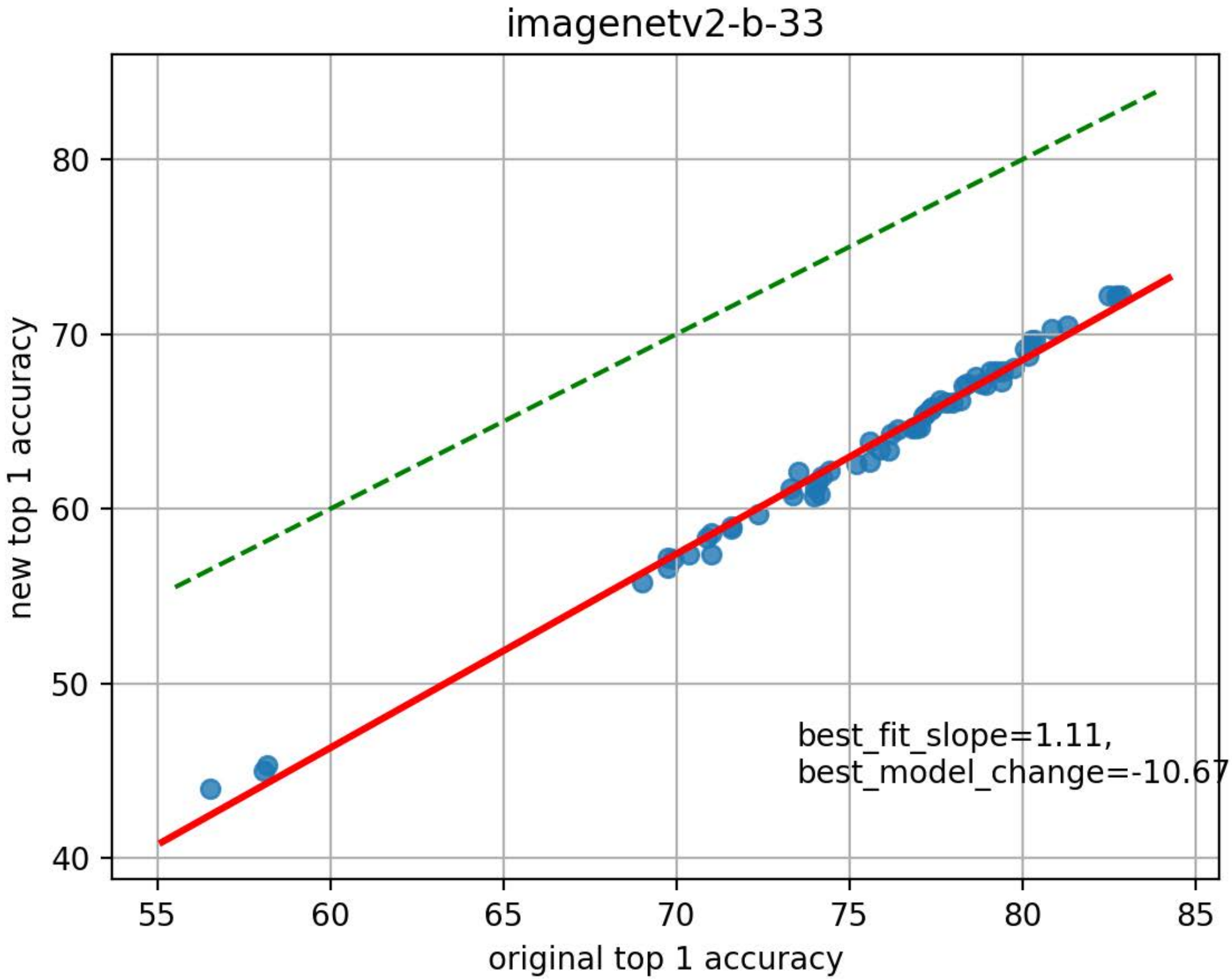
0.7


Elite Archers
~LCC

0.5


50¢
RIBBON AND BOW

0.2


imagenetv2-b-33
new top 1 accuracy
original top 1 accuracy
80
70
60
50
40
55
60
65
70
75
80
85
best_fit_slope=1.11,
best_model_change=-10.67

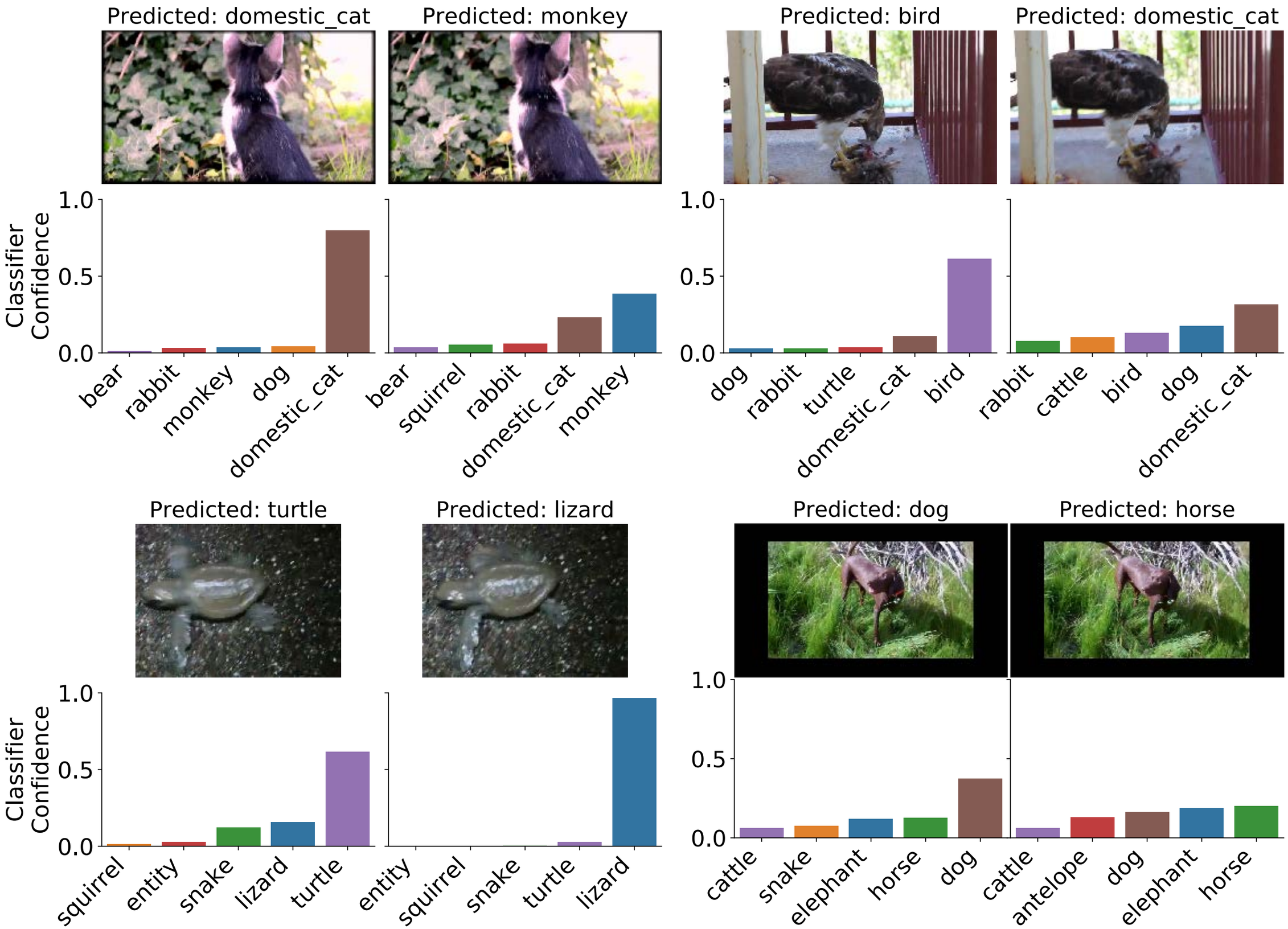
Can we find more evidence of classification fragility?

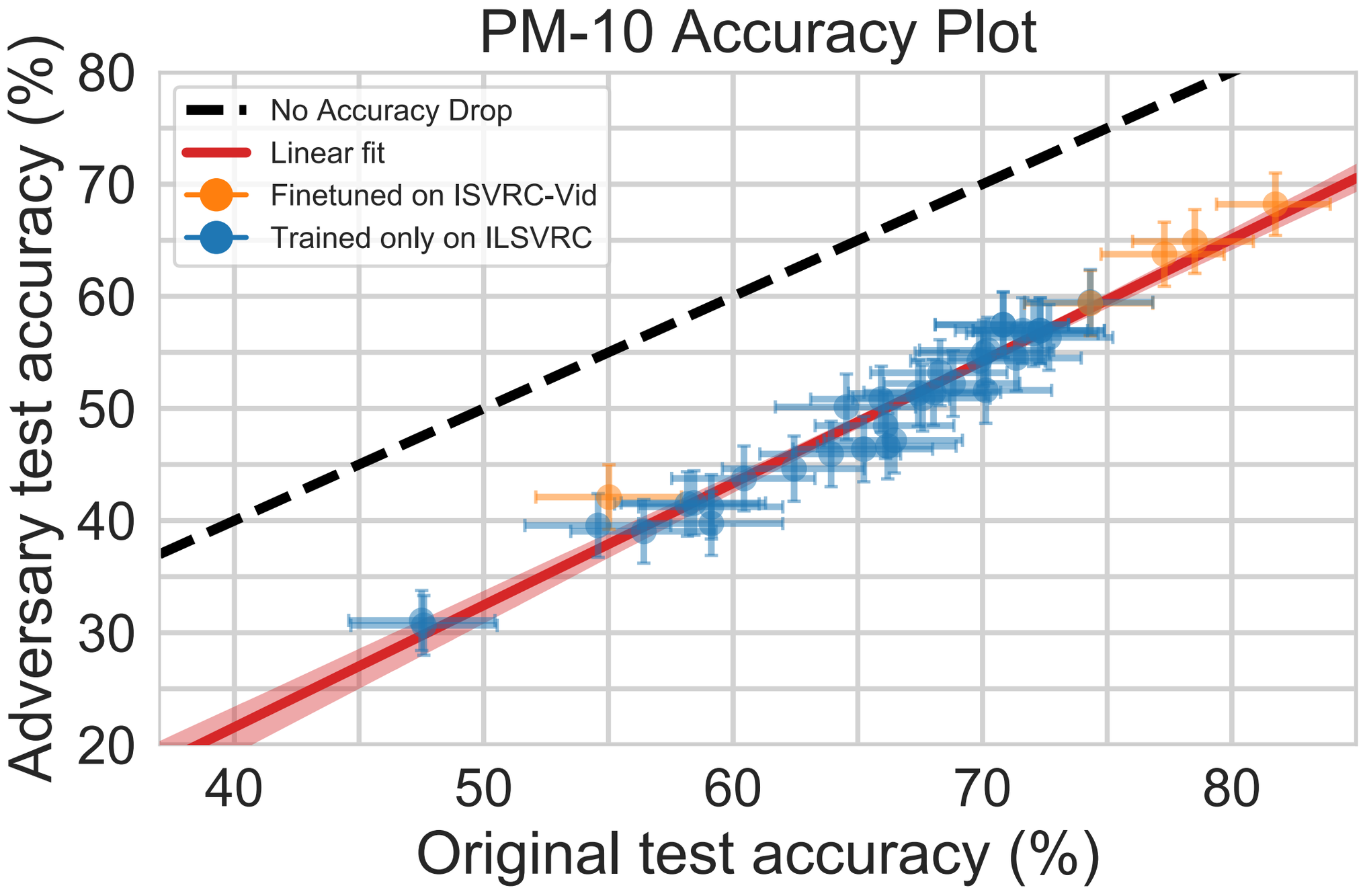
IMAGENET ILSVRC Video

- Video dataset curated by ImageNet team in 2015
  - Original designed as a detection/tracking task
  - 4000 training videos (presented as 1M jpeg frames)
  - 1314 validation videos (500k frames)
  - 30 classes (correspond to 288 ImageNet classes)

PM-K metric

- Plus-Minus-K frames metric
- Treat each video as a **set** of "similar images"
- Pick a frame index compute prediction at frame
- Now in a window of K frames from the frame look for a misclassified frame.
- Note: for a 30 FPS video 10 frames is 0.3s


Predicted: domestic_cat
Predicted: monkey
Predicted: bird
Predicted: domestic_cat
Classifier Confidence
1.0
0.5
0.0
bear
rabbit
monkey
dog
domestic_cat
bear
squirrel
rabbit
domestic_cat
monkey
dog
rabbit
turtle
domestic_cat
bird
rabbit
cattle
bird
dog
domestic_cat
Predicted: turtle
Predicted: lizard
Predicted: dog
Predicted: horse
Classifier Confidence
1.0
0.5
0.0
squirrel
entity
snake
lizard
turtle
entity
squirrel
snake
turtle
lizard
cattle
snake
elephant
horse
dog
cattle
antelope
dog
elephant
horse


PM-10 Accuracy Plot
No Accuracy Drop
Linear fit
Finetuned on ISVRC-Vid
Trained only on ILSVRC
Adversary test accuracy (%)
80
70
60
50
40
30
20
40
50
60
70
80
Original test accuracy (%)

18 Active Competitions

Two Sigma: Using News to Predict Stock Movements
Use news analytics to predict stock price performance
Featured · Kernels Competition · 2 months to go · news agencies, time series, finance, m...
$100,000
2,927 teams

Jigsaw Unintended Bias in Toxicity Classification
Detect toxicity across a diverse range of conversations
Featured · Kernels Competition · 2 months to go · biases, nlp, text data
$65,000
1,612 teams

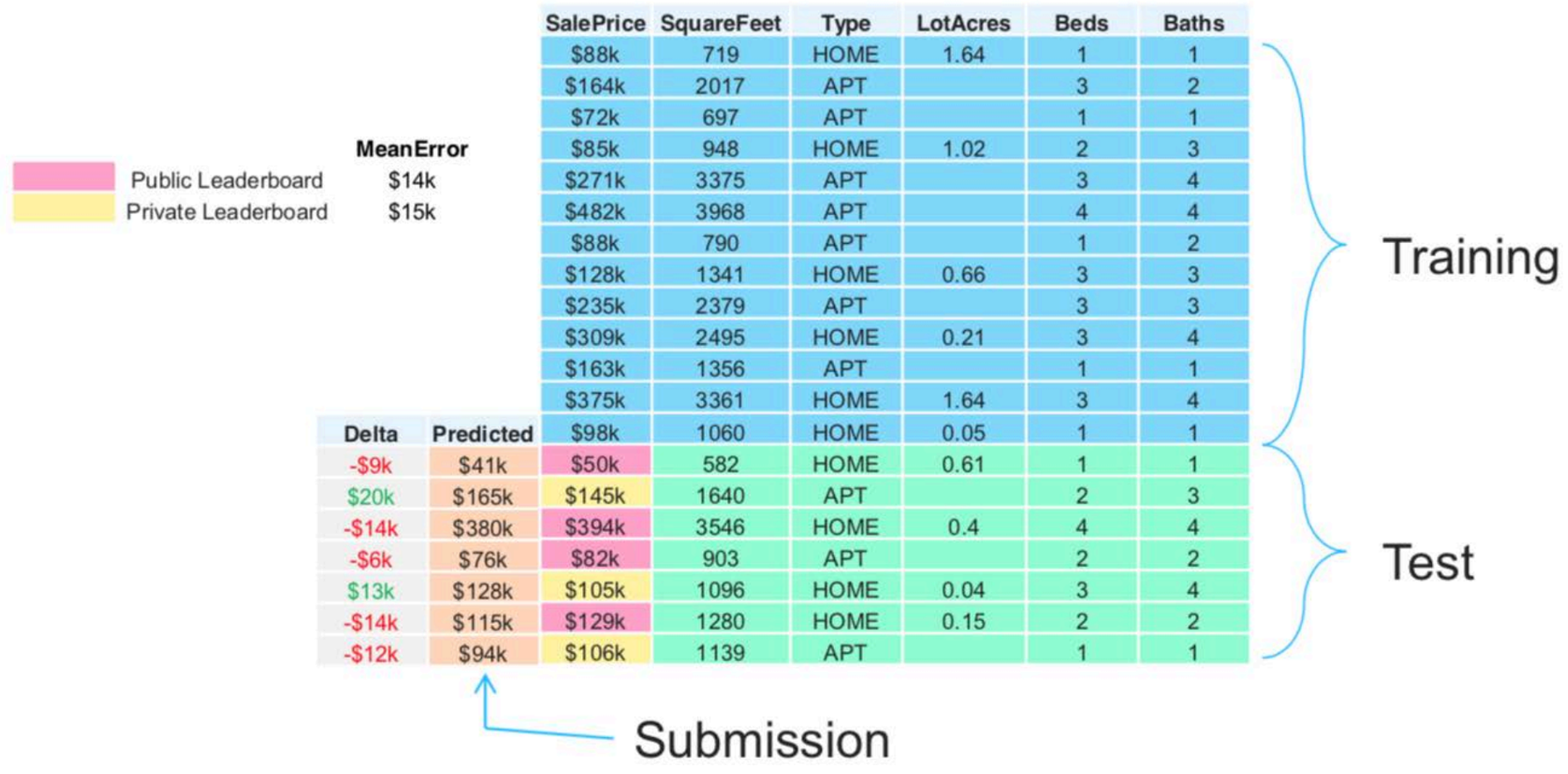
LANL Earthquake Prediction
Can you predict upcoming laboratory earthquakes?
Research · a month to go · earth sciences, physics, signal processing
$50,000
3,544 teams

Google Landmark Recognition 2019
Label famous (and not-so-famous) landmarks in images
Research · a month to go
$25,000
112 teams

Google Landmark Retrieval 2019
Given an image, can you find all of the same landmarks in a dataset?
Research · a month to go
$25,000
91 teams

Freesound Audio Tagging 2019
Automatically recognize sounds and apply tags of varying natures
Research · Kernels Competition · a month to go · sound technology, audio data
$5,000
508 teams

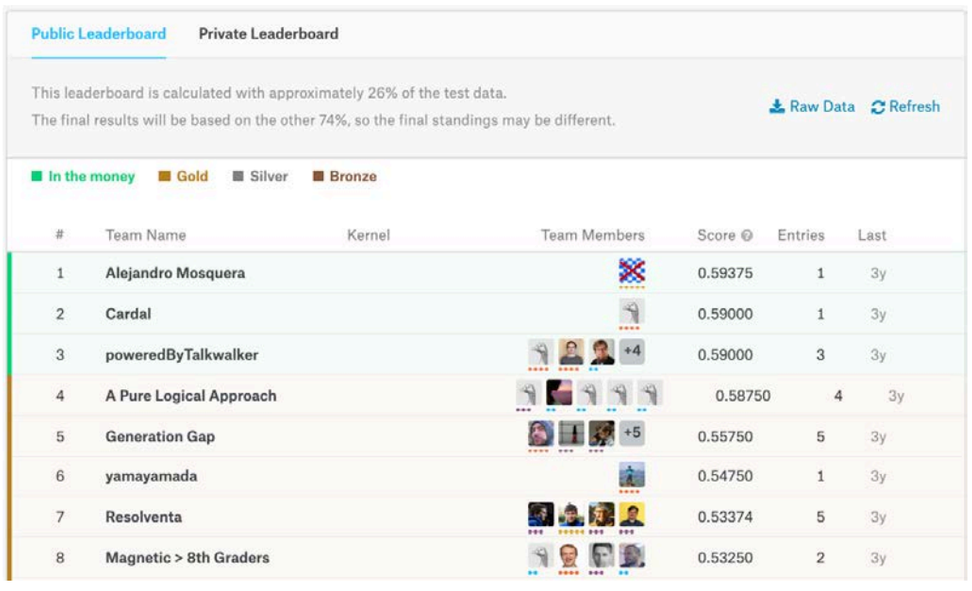
Kaggle computes two scores, one for the private leaderboard and one for the public leaderboard



[Ben Hamner, Kaggle at MLConf SF, 2015]

Participant immediately sees score on public leaderboard

At the end, participant with the best score on the private leaderboard wins



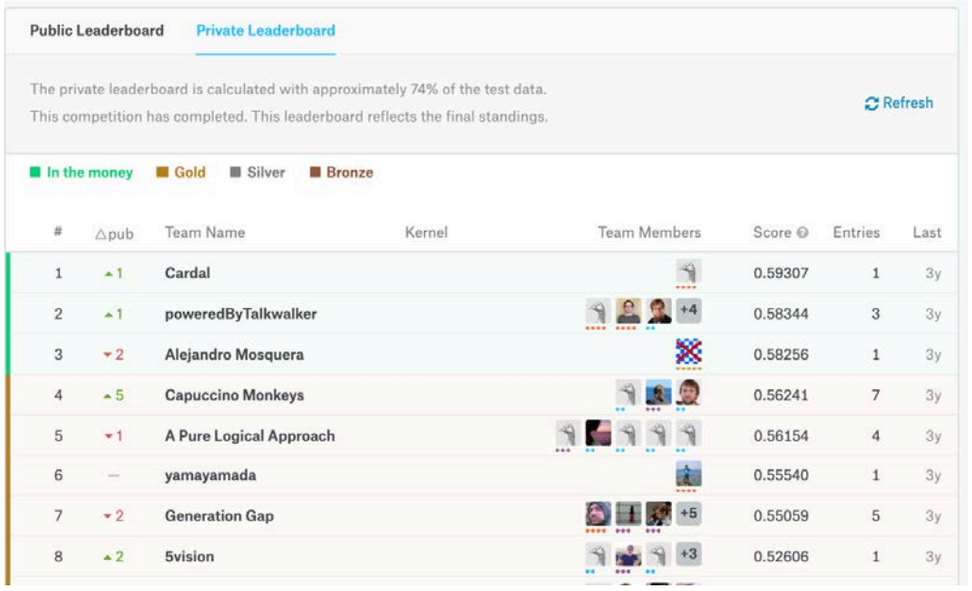
Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 26% of the test data.
The final results will be based on the other 74%, so the final standings may be different.

Raw Data | Refresh

In the money | Gold | Silver | Bronze

#	Team Name	Kernel	Team Members	Score	Entries	Last
1	Alejandro Mosquera			0.59375	1	3y
2	Cardal			0.59000	1	3y
3	poweredByTalkwalker		+4	0.59000	3	3y
4	A Pure Logical Approach			0.58750	4	3y
5	Generation Gap		+5	0.55750	5	3y
6	yamayamada			0.54750	1	3y
7	Resolventa			0.53374	5	3y
8	Magnetic > 8th Graders			0.53250	2	3y



Public Leaderboard | Private Leaderboard

The private leaderboard is calculated with approximately 74% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Refresh

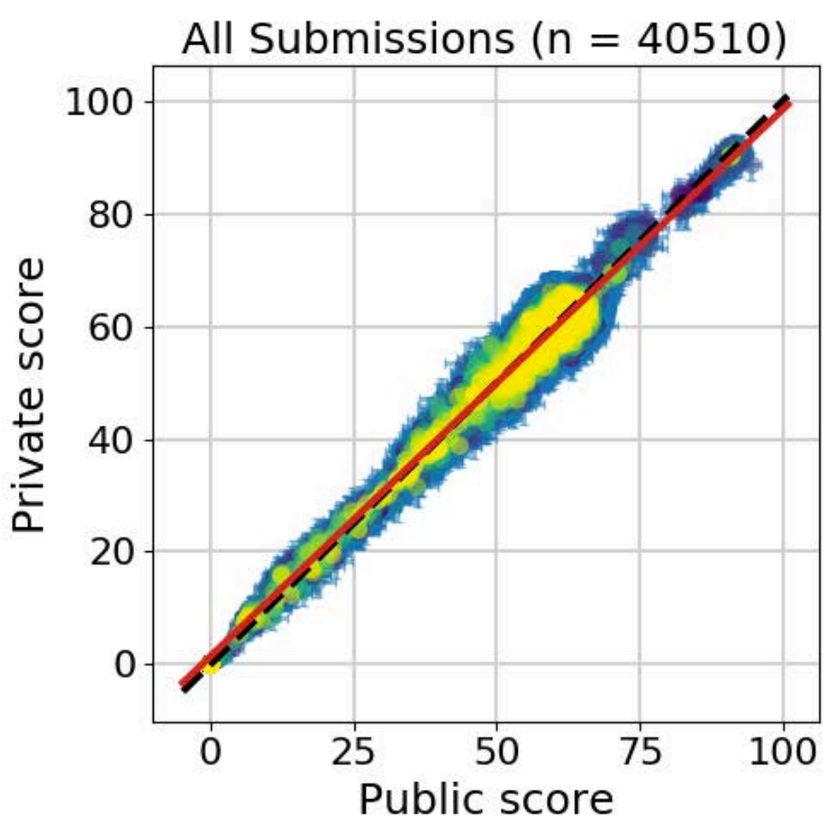
In the money | Gold | Silver | Bronze

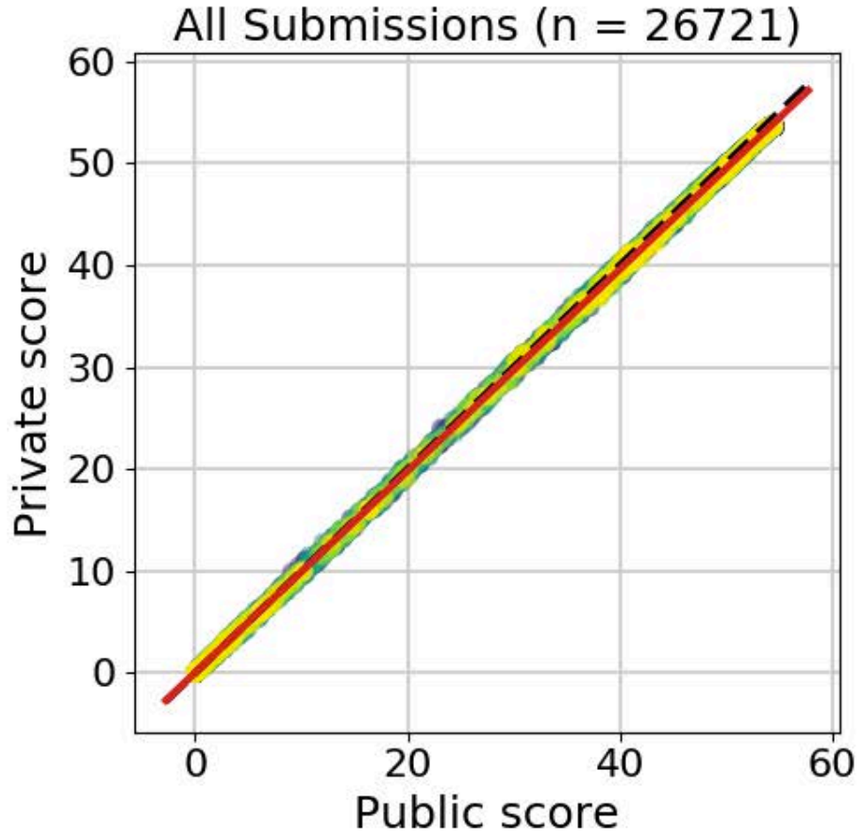
#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	Cardal			0.59307	1	3y
2	▲1	poweredByTalkwalker		+4	0.58344	3	3y
3	▼2	Alejandro Mosquera			0.58256	1	3y
4	▲5	Capuccino Monkeys			0.56241	7	3y
5	▼1	A Pure Logical Approach			0.56154	4	3y
6	—	yamayamada			0.55540	1	3y
7	▼2	Generation Gap		+5	0.55059	5	3y
8	▲2	5vision		+3	0.52606	1	3y

Kaggle Competition Analysis

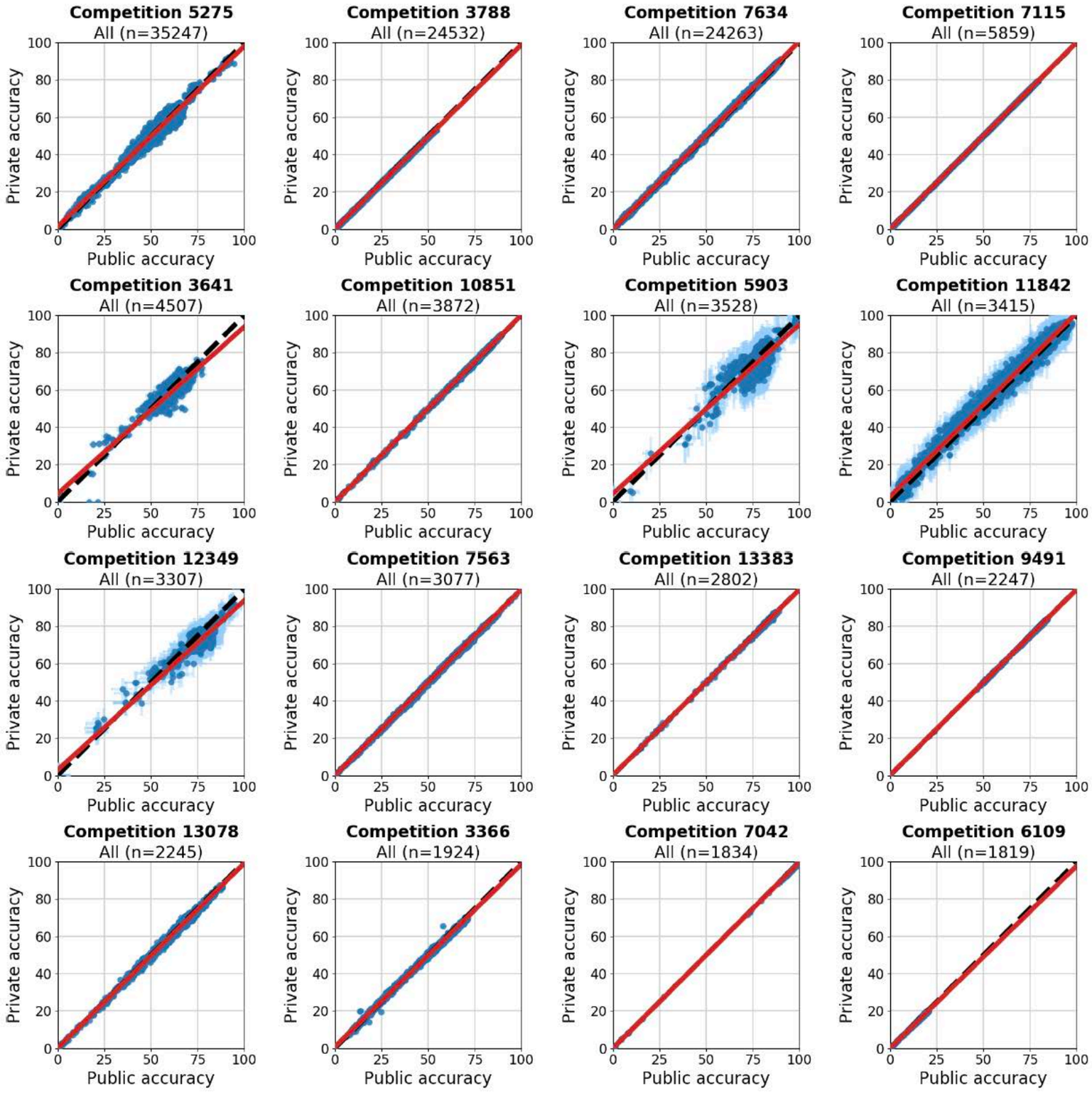
Currently analyzing competitions

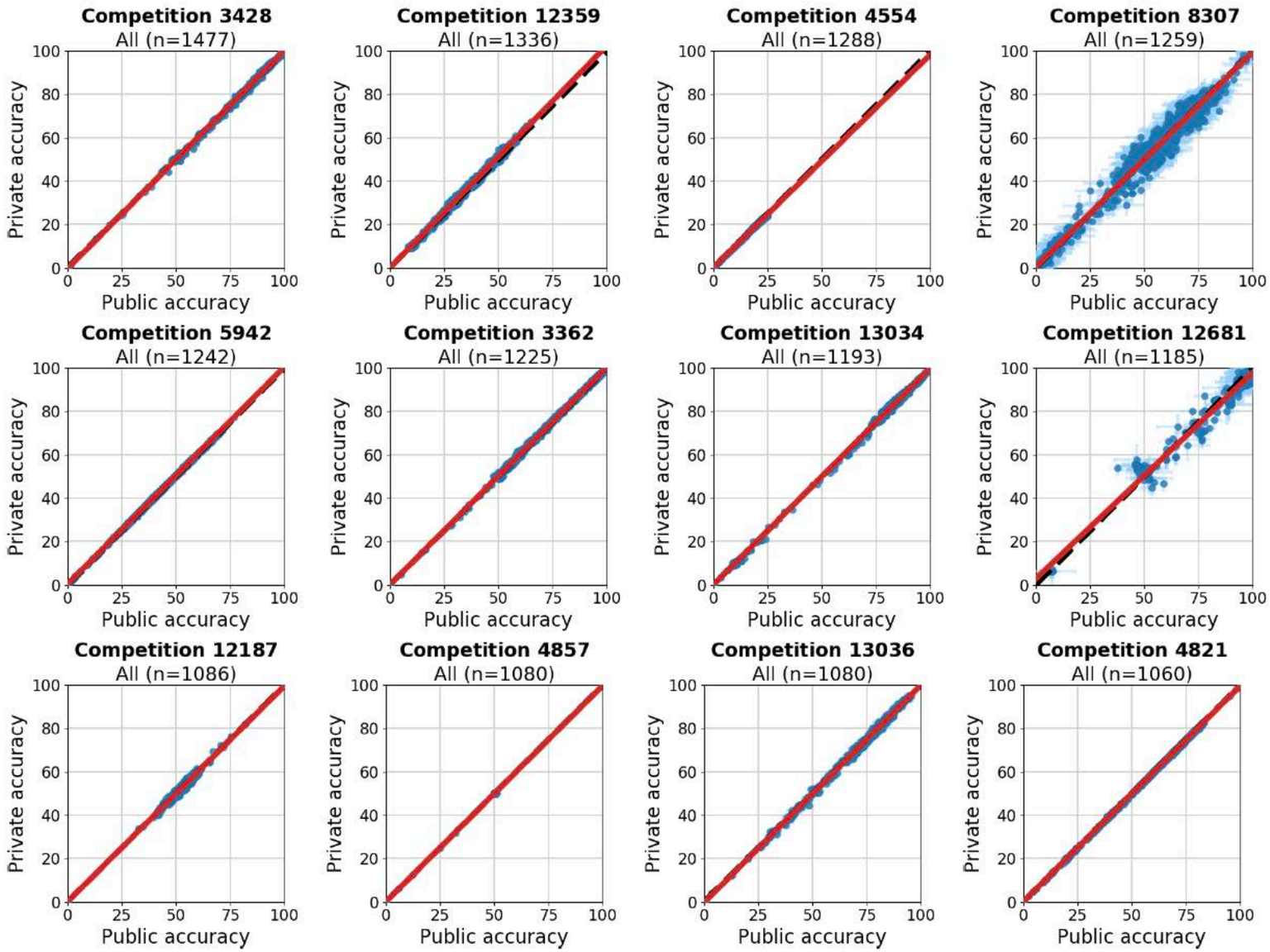
Accuracy
RMSE
AUC





no signs of overfitting here either.

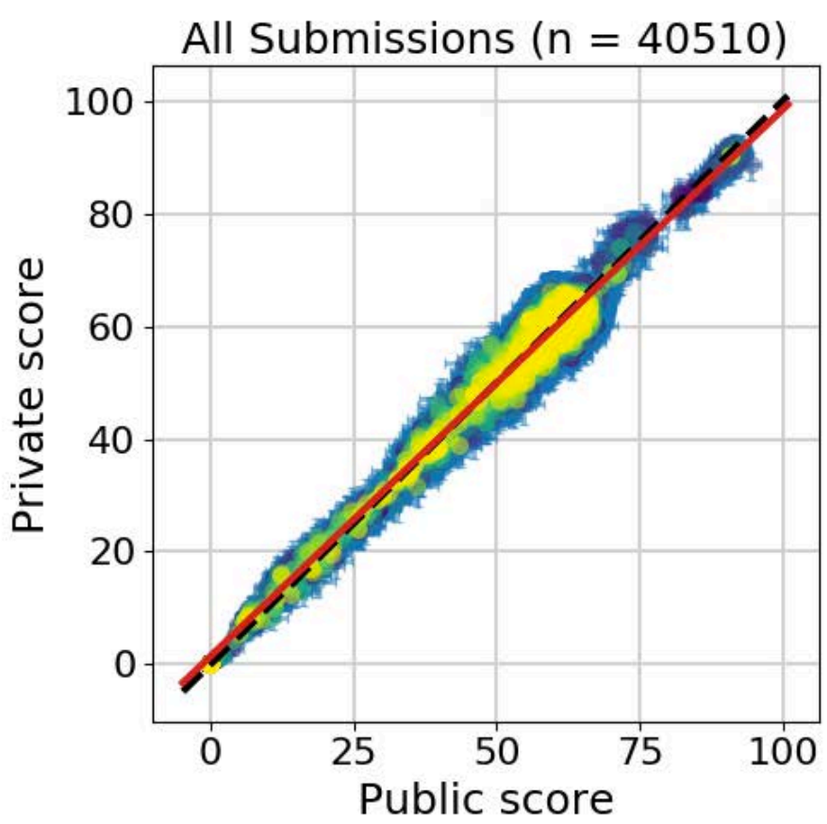

Competition 5275
All (n=35247)
Competition 3788
All (n=24532)
Competition 7634
All (n=24263)
Competition 7115
All (n=5859)
Competition 3641
All (n=4507)
Competition 10851
All (n=3872)
Competition 5903
All (n=3528)
Competition 11842
All (n=3415)
Competition 12349
All (n=3307)
Competition 7563
All (n=3077)
Competition 13383
All (n=2802)
Competition 9491
All (n=2247)
Competition 13078
All (n=2245)
Competition 3366
All (n=1924)
Competition 7042
All (n=1834)
Competition 6109
All (n=1819)
Private accuracy
Public accuracy
0
20
40
60
80
100
0
25
50
75
100

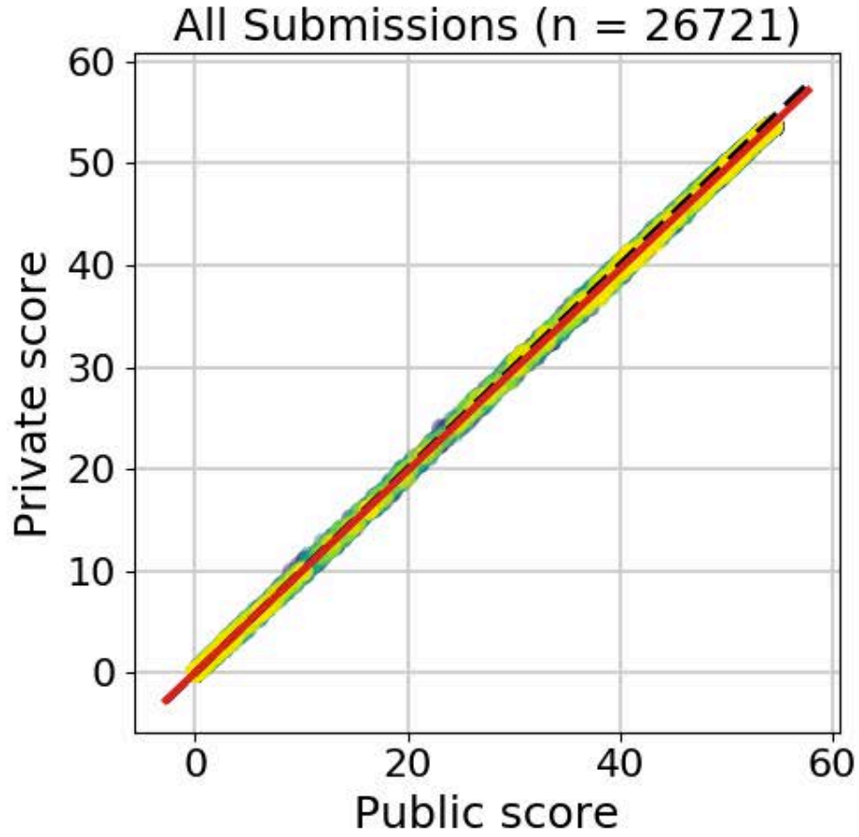

Competition 3428
All (n=1477)
Competition 12359
All (n=1336)
Competition 4554
All (n=1288)
Competition 8307
All (n=1259)
Competition 5942
All (n=1242)
Competition 3362
All (n=1225)
Competition 13034
All (n=1193)
Competition 12681
All (n=1185)
Competition 12187
All (n=1086)
Competition 4857
All (n=1080)
Competition 13036
All (n=1080)
Competition 4821
All (n=1060)
Private accuracy
Public accuracy
0
20
40
60
80
100
0
25
50
75
100

Kaggle Competition Analysis

Currently analyzing competitions

Accuracy
RMSE
AUC



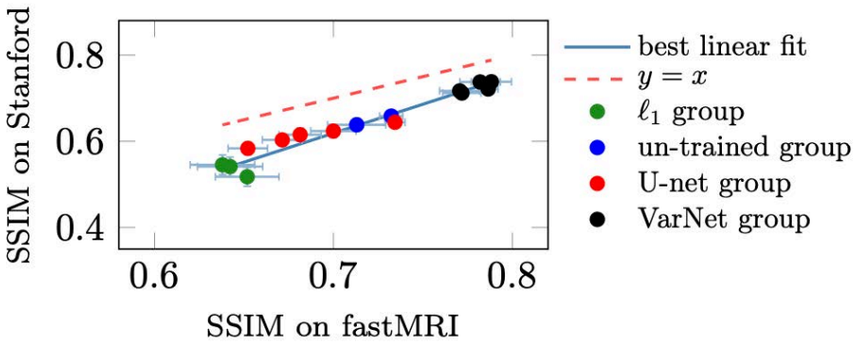


no signs of overfitting here either.

Beyond image classification

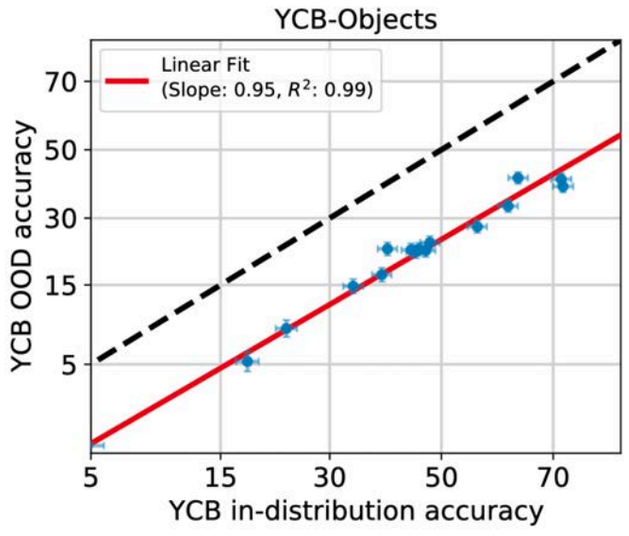
Similar phenomena appear in other computer vision problems:

MRI reconstruction



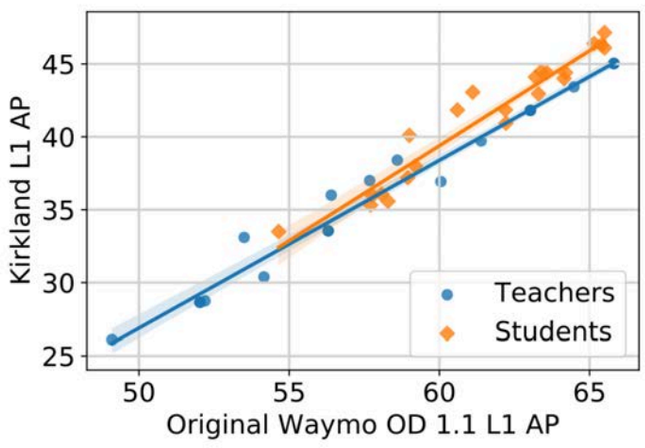
[Darestani, Chaudhari, Heckel '21]

Pose estimation



[Miller et al. '21]

Object detection

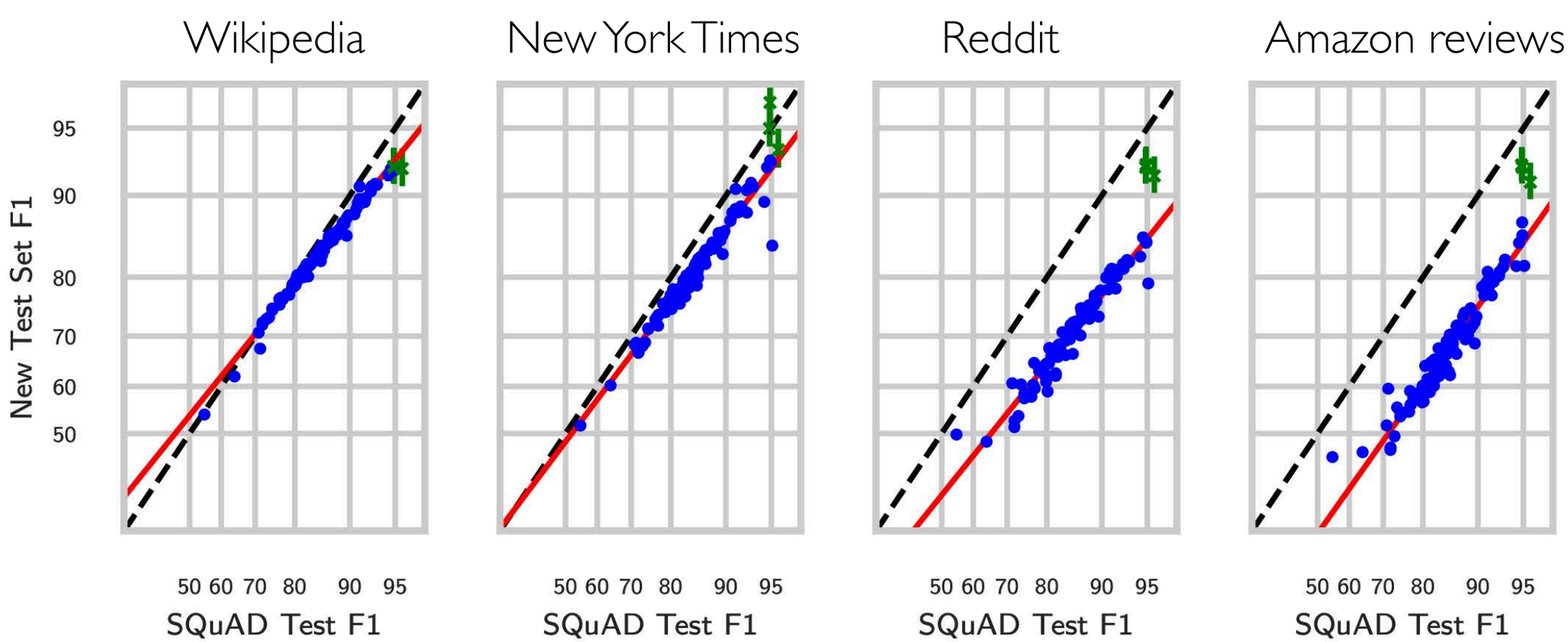


[Roelofs et al. '21]

Beyond computer vision

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

➡ Similar trends in natural language processing.



[Miller et al., '20]

Similar story in domain generalization

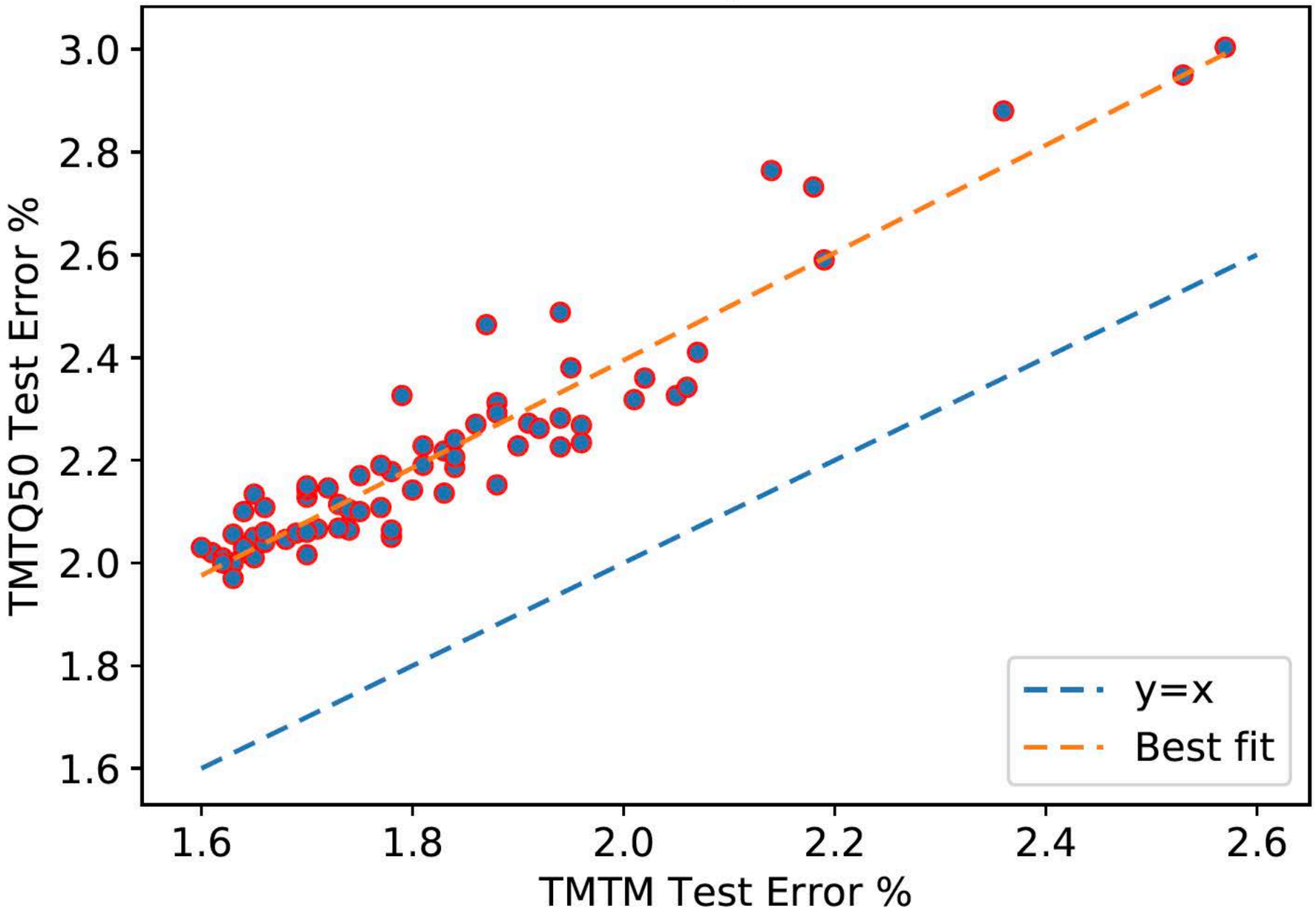
In Search of Lost Domain Generalization

Ishaan Gulrajani and David Lopez-Paz*
Facebook AI Research
igul222@gmail.com, dlp@fb.com

Abstract

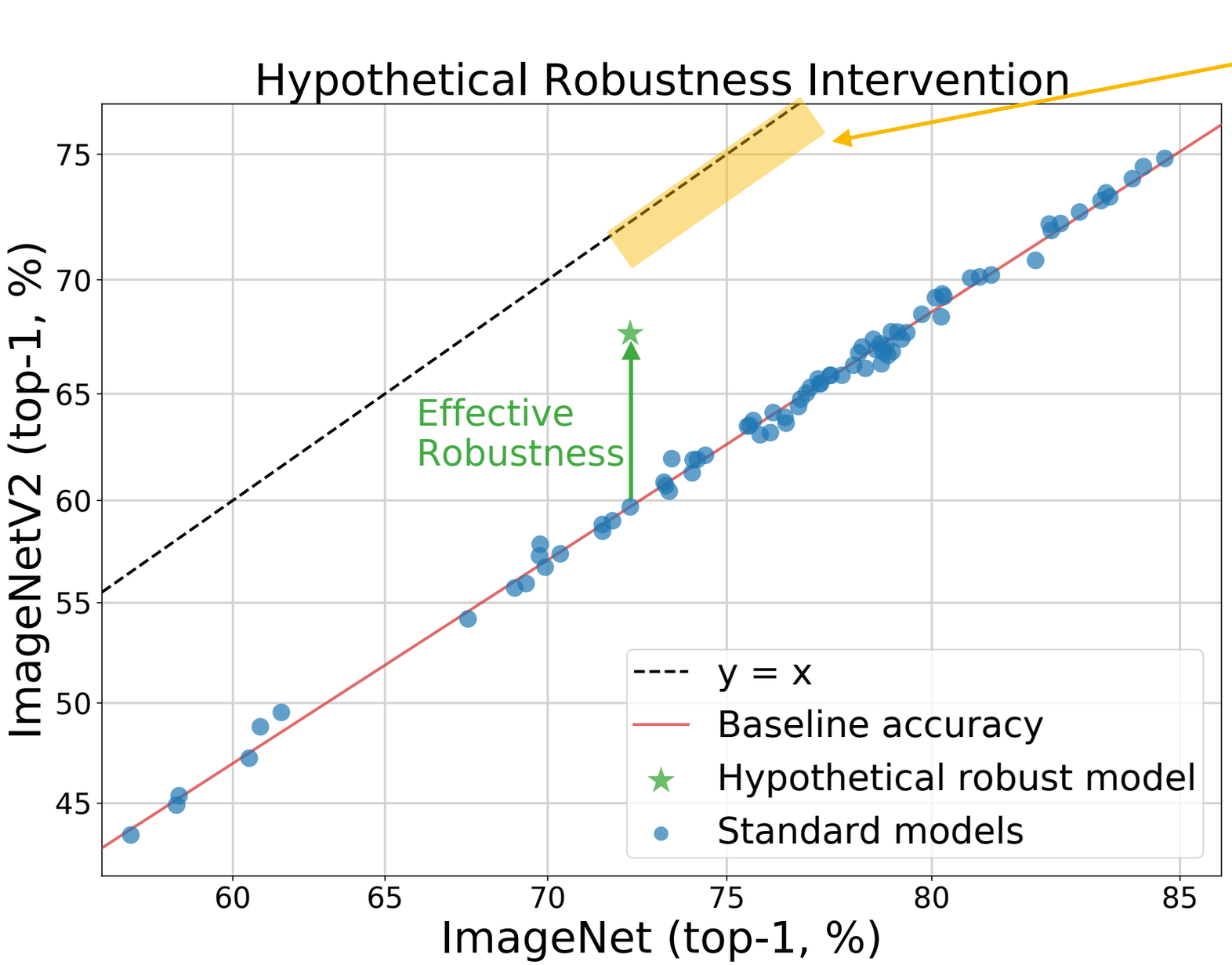
The goal of domain generalization algorithms is to predict well on distributions different from those seen during training. While a myriad of domain generalization algorithms exist, inconsistencies in experimental conditions—datasets, architectures, and model selection criteria—render fair and realistic comparisons difficult. In this paper, we are interested in understanding how useful domain generalization algorithms are in realistic settings. As a first step, we realize that model selection is non-trivial for domain generalization tasks. Contrary to prior work, we argue that domain generalization algorithms without a model selection strategy should be regarded as incomplete. Next, we implement DomainBed, a testbed for domain generalization including seven multi-domain datasets, nine baseline algorithms, and three model selection criteria. We conduct extensive experiments using DomainBed and find that, when carefully implemented, empirical risk minimization shows state-of-the-art performance across all datasets. Looking forward, we hope that the release of DomainBed, along with contributions from fellow researchers, will streamline reproducible and rigorous research in domain generalization.

MNIST: the only data set that matters



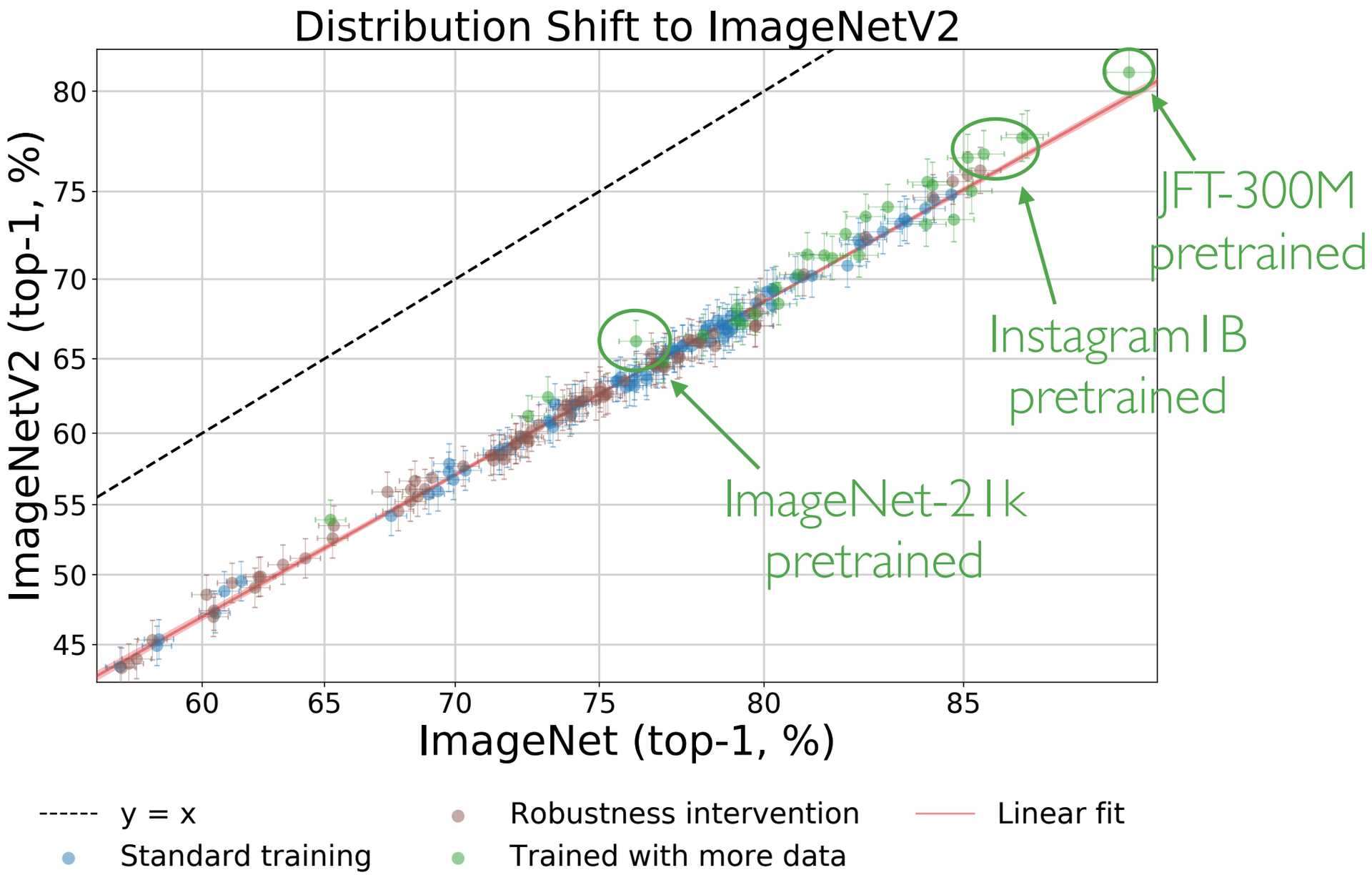
Yadav and Bottou, 2019

arXiv:1905.10498v1



[Shankar, Roelofs, Mania, Fang, Recht, Schmidt '20]

Do current robustness interventions achieve effective robustness?

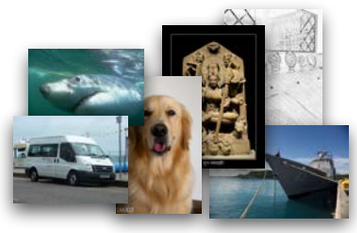


Only training on (a lot) more data gives a small amount of effective robustness.

[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

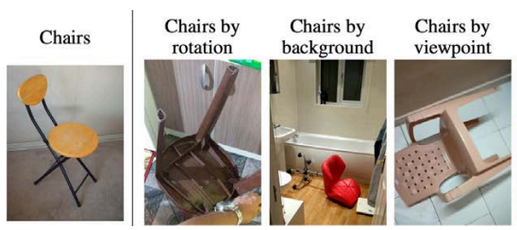
More Robustness on ImageNet

Even more test sets



ImageNetV2

[Recht, Roelofs, Schmidt, Shankar '19]



ObjectNet

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz '19]

ImageNet-Sketch

[Wang, Ge, Lipton, Xing '19]

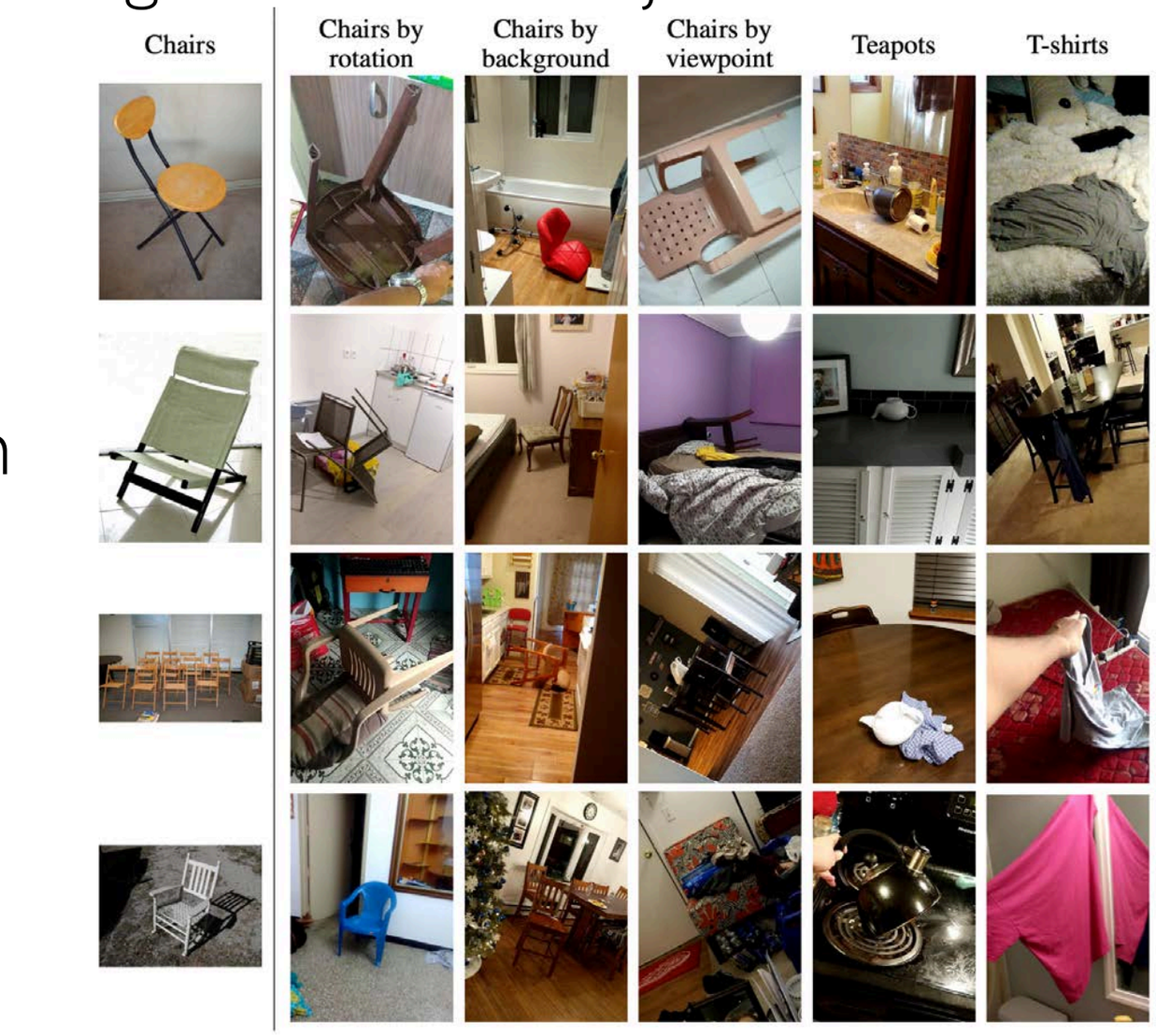
ImageNet-R

[Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo, Desai, Zhu, Parajuli, Guo, Song, Steinhardt, Gilmer '20]

ObjectNet: Objects in Unusual Positions

Mainly object-centric and clean images

(collected from Flickr)



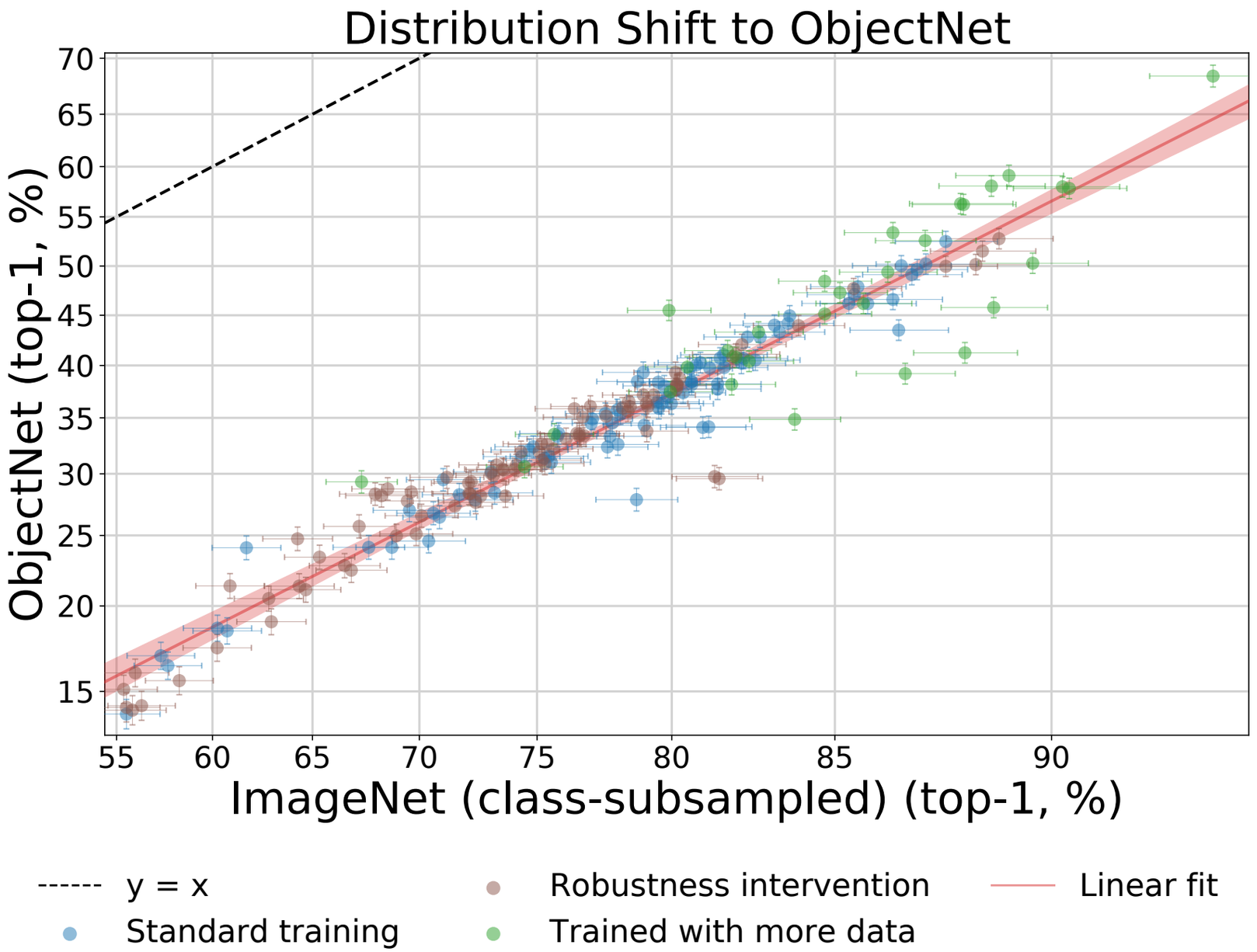
Intentionally randomized:

- object poses
- locations
- etc.

(collected via specific crowd worker annotations)

[Barbu et al.'19]

[Barbu, Mayo, Alverio, Luo, Wang, Gutfreund, Tenenbaum, Katz '19]



Same trend: only more data gives effective robustness.

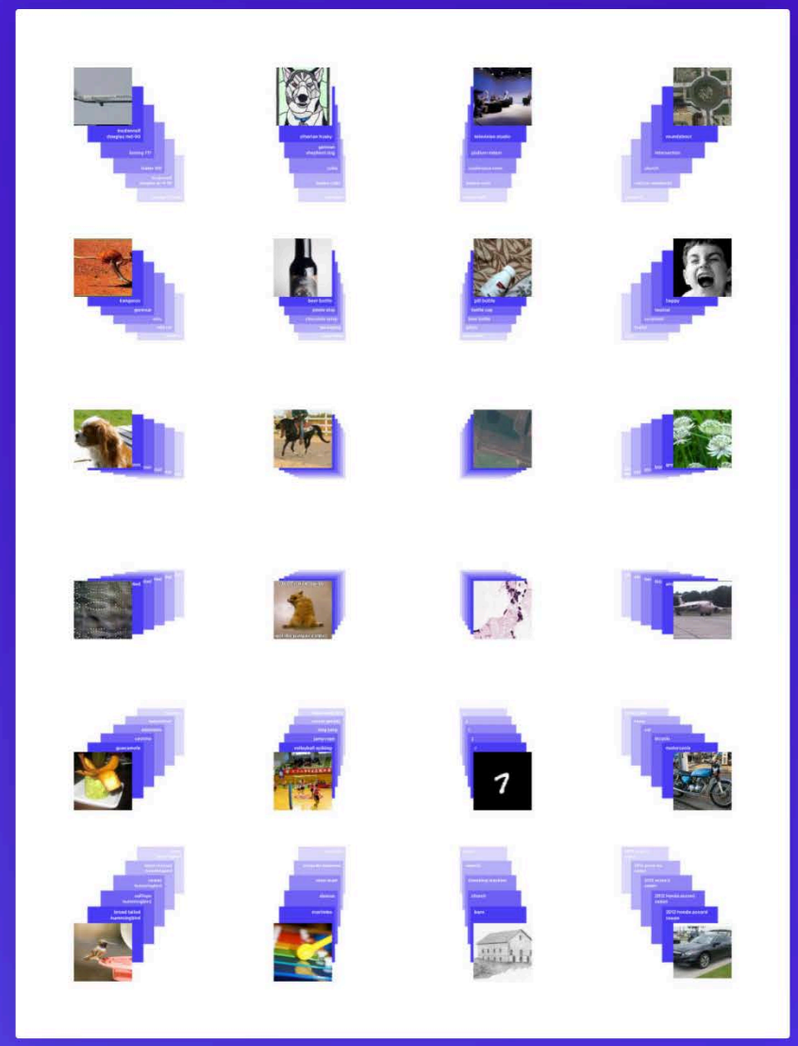
[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

Why is this happening?

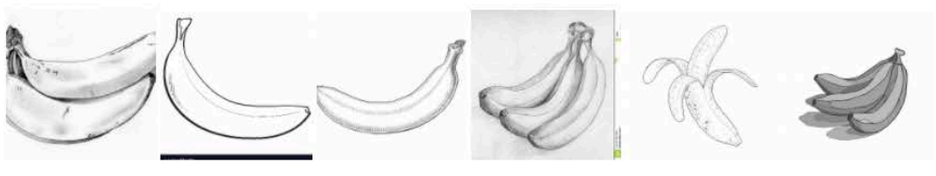
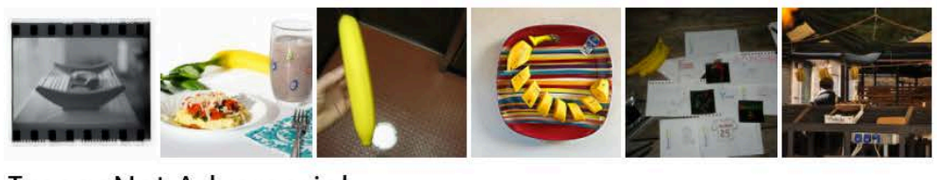
- Many hypotheses
- Ladder, Model-similarity, data robustness, dataset diversity, etc. etc.
- Great papers, you should read them!
- I'm not sure any of these answers are satisfying

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.

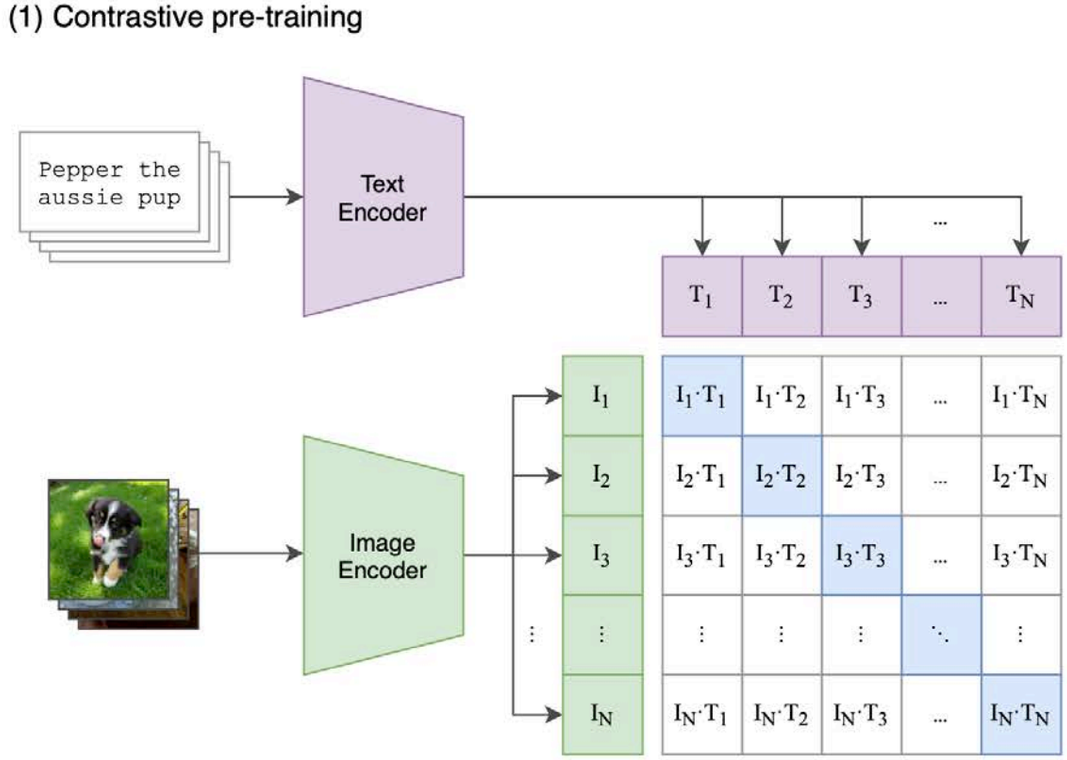


January 5, 2021
15 minute read

DATASET	IMAGENET RESNET101	CLIP VIT-L	
 ImageNet	76.2%	76.2%	
 ImageNet V2	64.3%	70.1%	+6%
 ImageNet Rendition	37.7%	88.9%	+51%
 ObjectNet	32.6%	72.3%	+40%
 ImageNet Sketch	25.2%	60.2%	+35%
 ImageNet Adversarial	2.7%	77.1%	+74%

Very large improvements in out-of-distribution robustness.

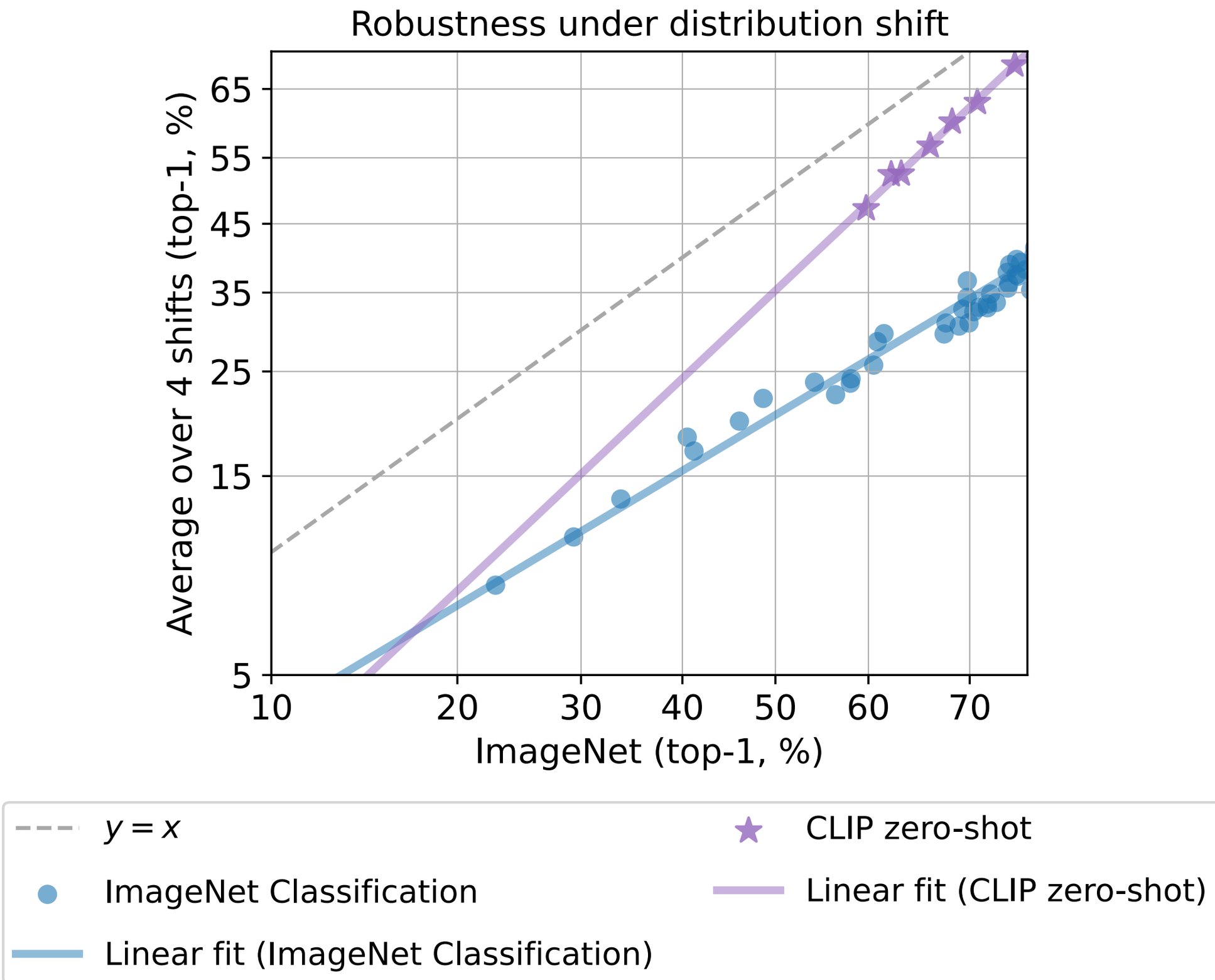
CLIP is not (explicitly) designed for robustness



Training data: 400 million images collected from the web (dataset internal to OpenAI).

Compute: Trained on 250 - 600 GPUs for up to 18 days.

Model: ResNets and ViTs with up to 300M parameters.


Robustness under distribution shift
Average over 4 shifts (top-1, %)
65
55
45
35
25
15
5
10
20
30
40
50
60
70
ImageNet (top-1, %)
$y = x$
ImageNet Classification
Linear fit (ImageNet Classification)
CLIP zero-shot
Linear fit (CLIP zero-shot)

What we have *always* seen

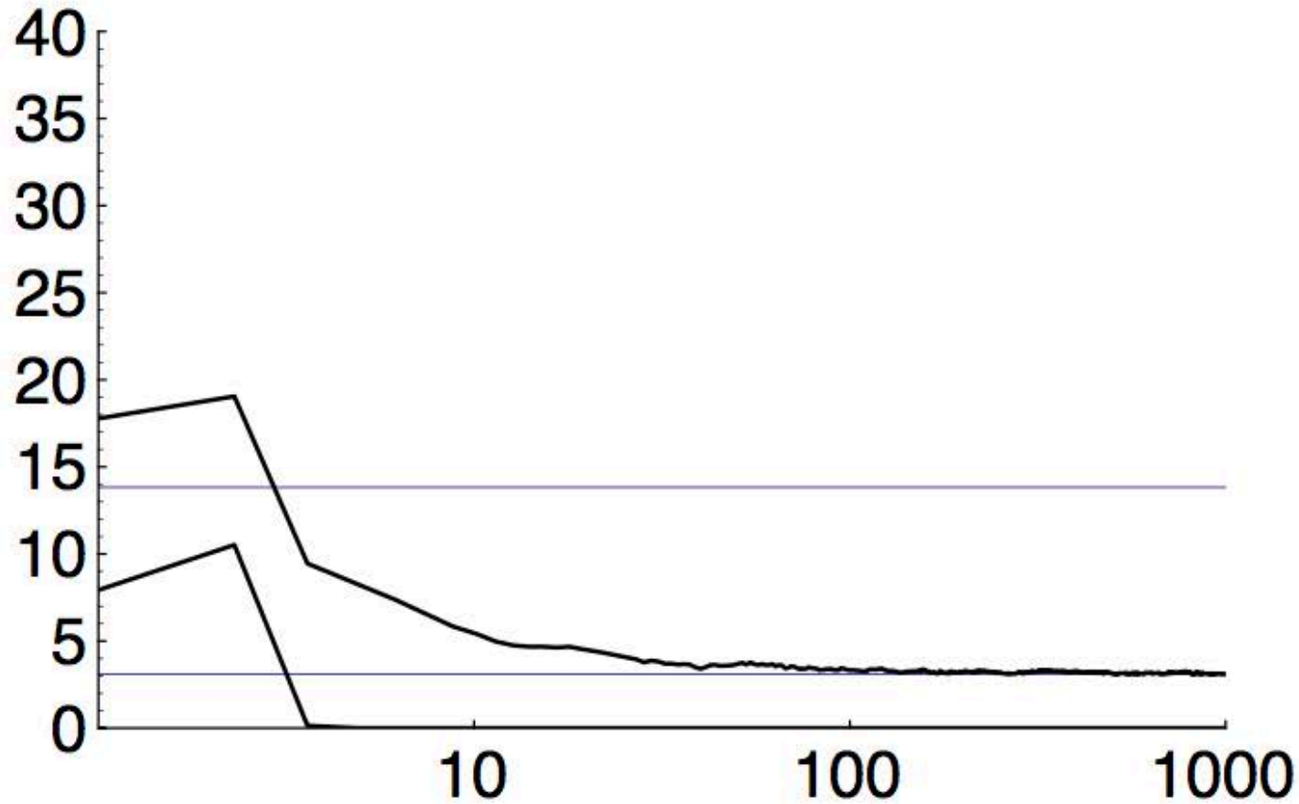
- Interpolating your training data is fine.
- Training on your test set is fine.
- Making models huge doesn't hurt.
- Making models huge doesn't help much.

We have to reorient how we talk about ML before we figure out a better way forward.

- Diminishing returns means wasting resources.
- Distribution shift is real and dangerous.

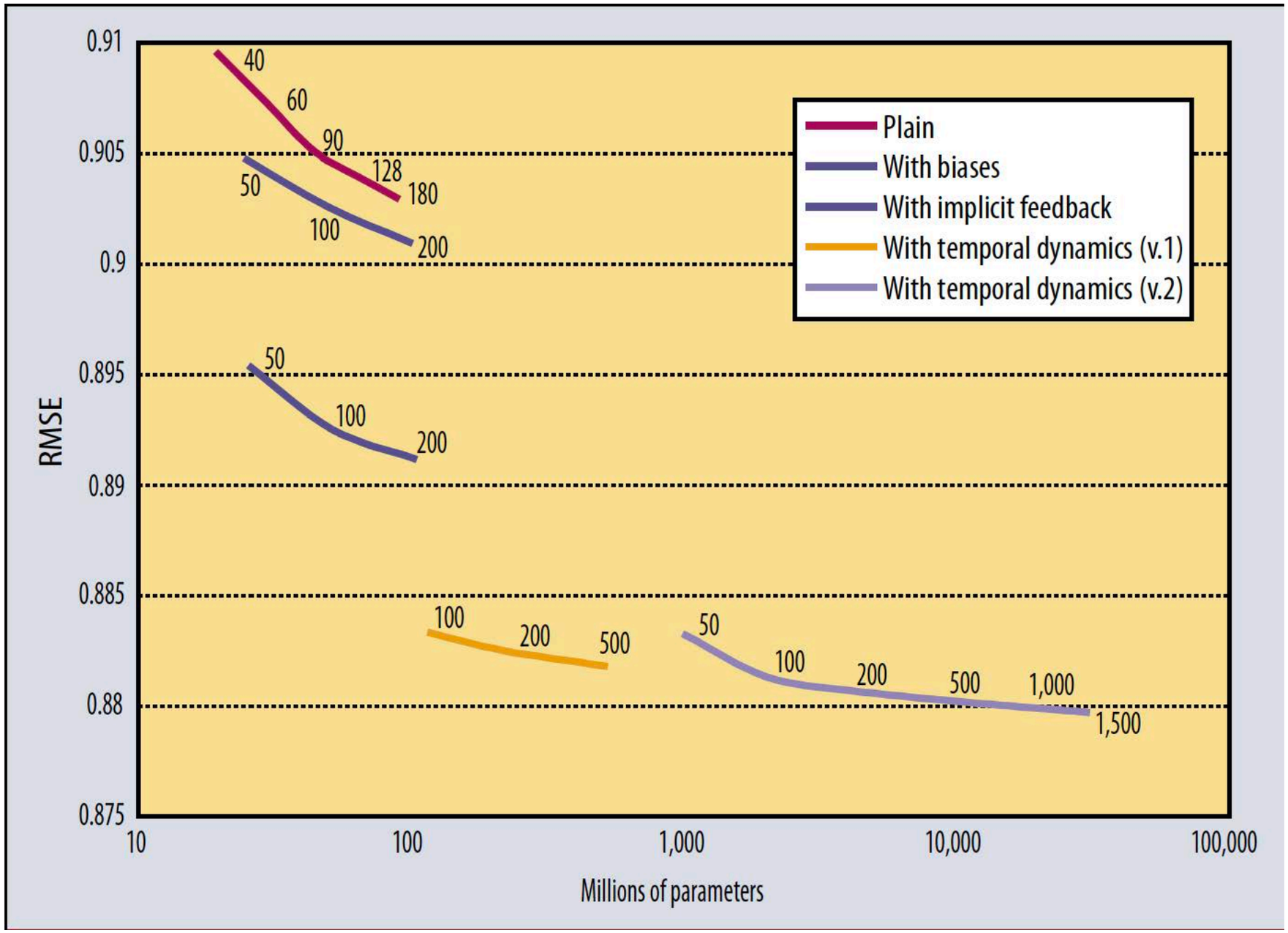
Boosting on UCI Letter data set.

n=16,000
d=16
k=26



P. Bartlett, N(eur)IPS Tutorial, 1998

Performance on Netflix Prize



R. Bell and Y. Koren, CACM, 2009

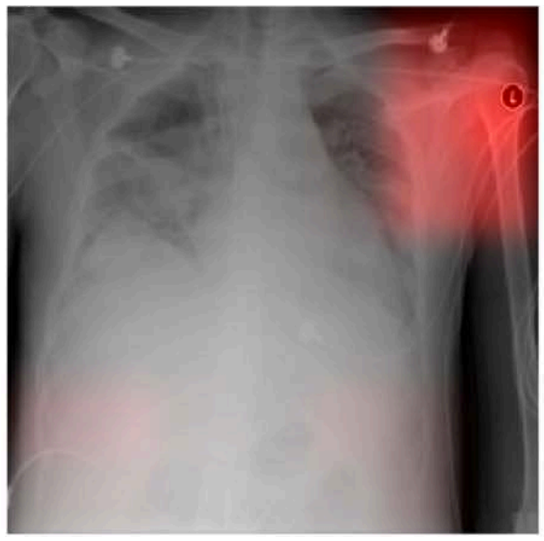
Distribution Shift is Dangerous

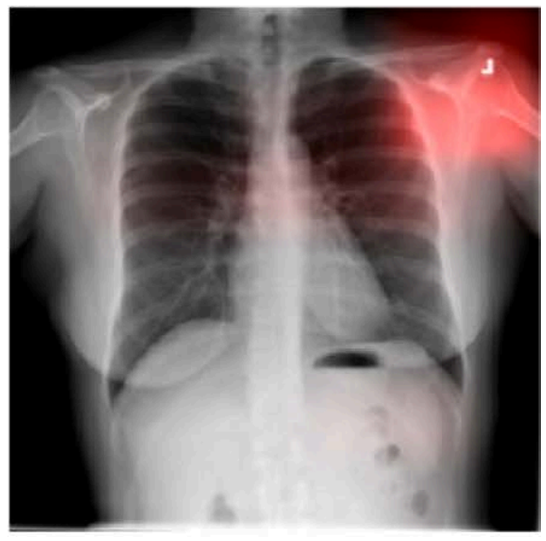
Even in a carefully-controlled reproducibility experiment.

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".

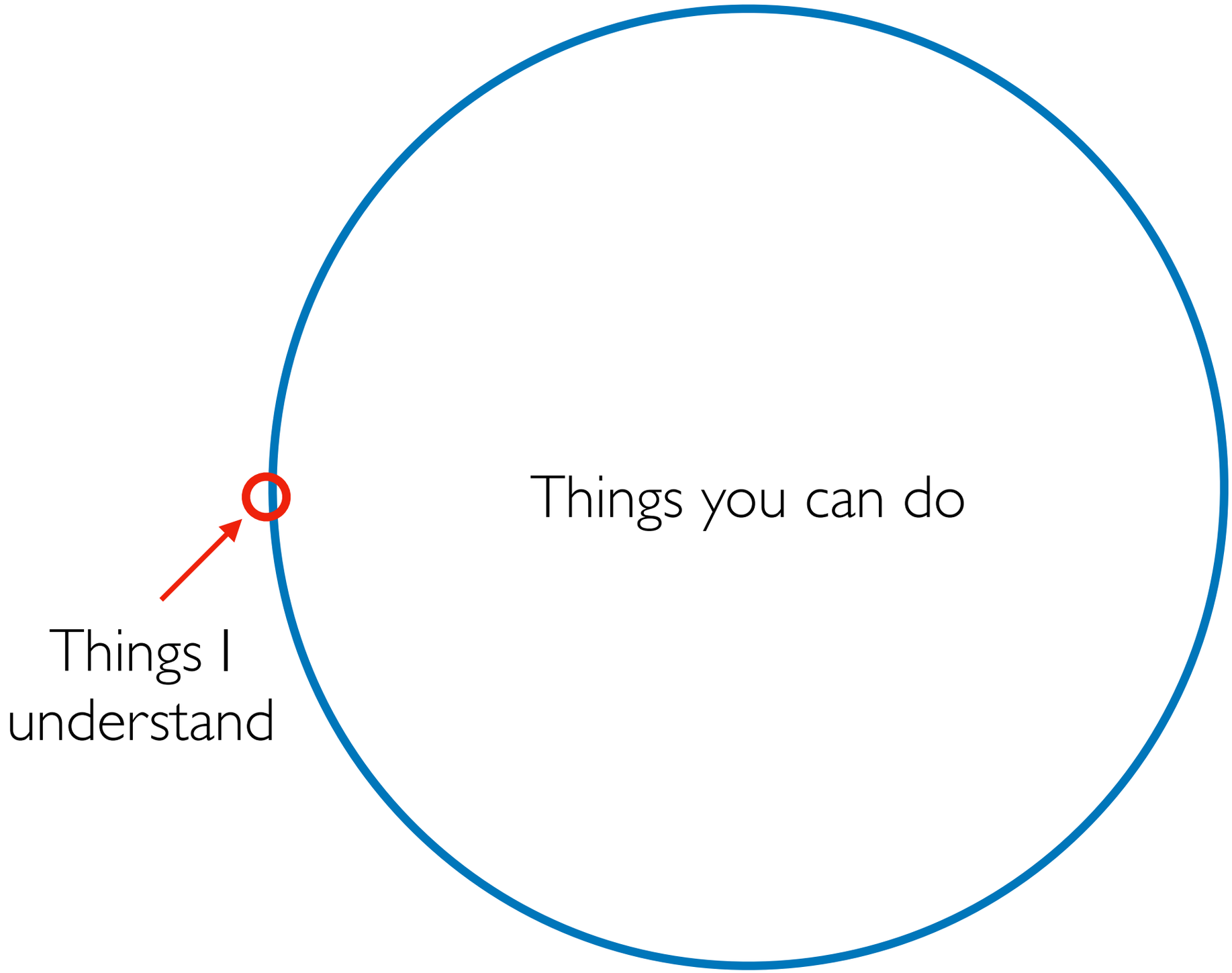
Distribution Shift is Dangerous

What we have *always* seen

- Interpolating your training data is fine.
- Training on your test set is fine.
- Making models huge doesn't hurt.
- Making models huge doesn't help much.

We have to reorient how we talk about ML before we figure out a better way forward.

- Diminishing returns means wasting resources.
- Distribution shift is real and dangerous.


Things you can do
Things I
understand

Very partial list of references

- "Understanding deep learning requires rethinking generalization." C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. ICLR 2017.
- "Do ImageNet Classifiers Generalize to ImageNet?" B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. ICML 2019.
- "A Meta Analysis of Overfitting in Machine Learning." R. Roelofs, S. Fridovich-Keil, J. Miller, M. Hardt, B. Recht, and L. Schmidt. NeurIPS 2019.
- "Cold Case: the Lost MNIST Digits." C. Yadav and L. Bottou. NeurIPS 2019.
- "Evaluating Machine Accuracy on ImageNet." V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt. ICML 2020.
- "The Effect of Natural Distribution Shift on Question Answering Models." J. Miller, K. Krauth, B. Recht, and L. Schmidt. ICML 2020.
- "Measuring Robustness to Natural Distribution Shifts in Image Classification." R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. NeurIPS 2020.
- "Do Image Classifiers Generalize Across Time?" V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt. ICCV 2021.
- "Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)." A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, L. Schmidt. ICML 2022.