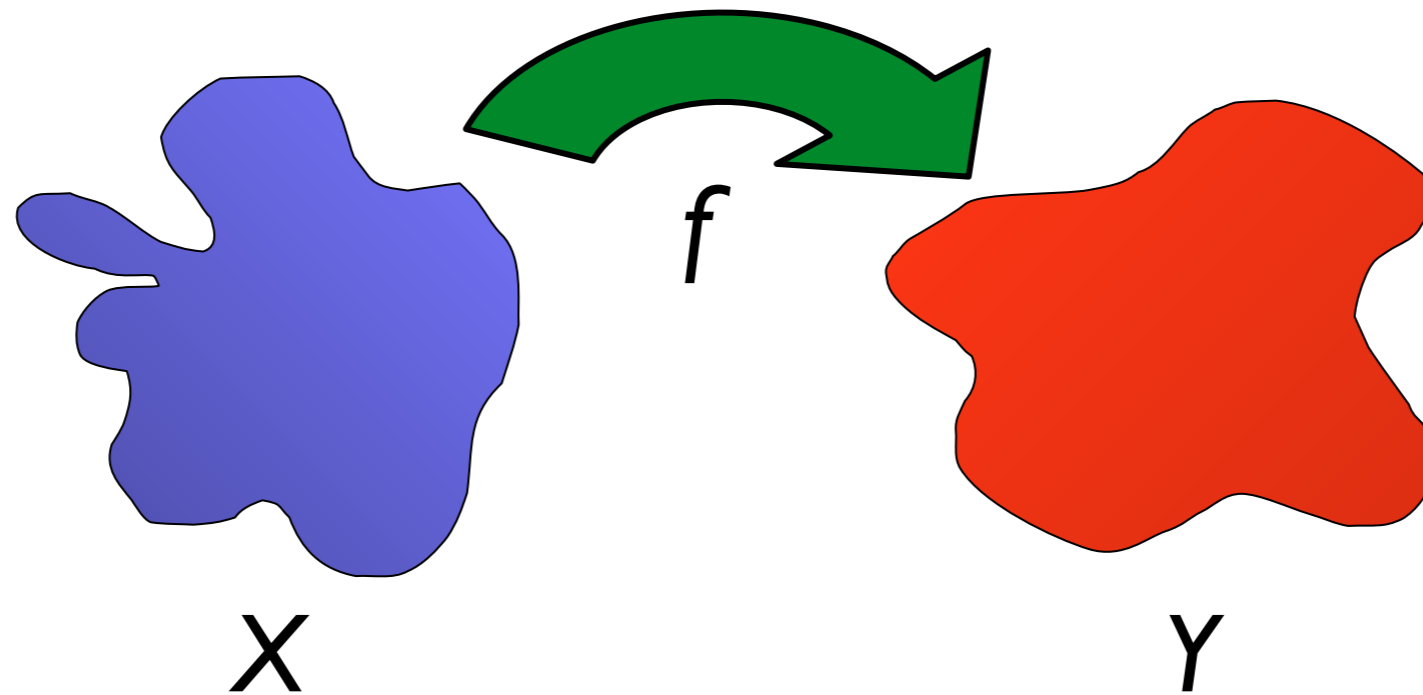


training on the test set and other heresies

Benjamin Recht
University of California, Berkeley

A narrow view of machine learning

the study of *prediction* from *examples*



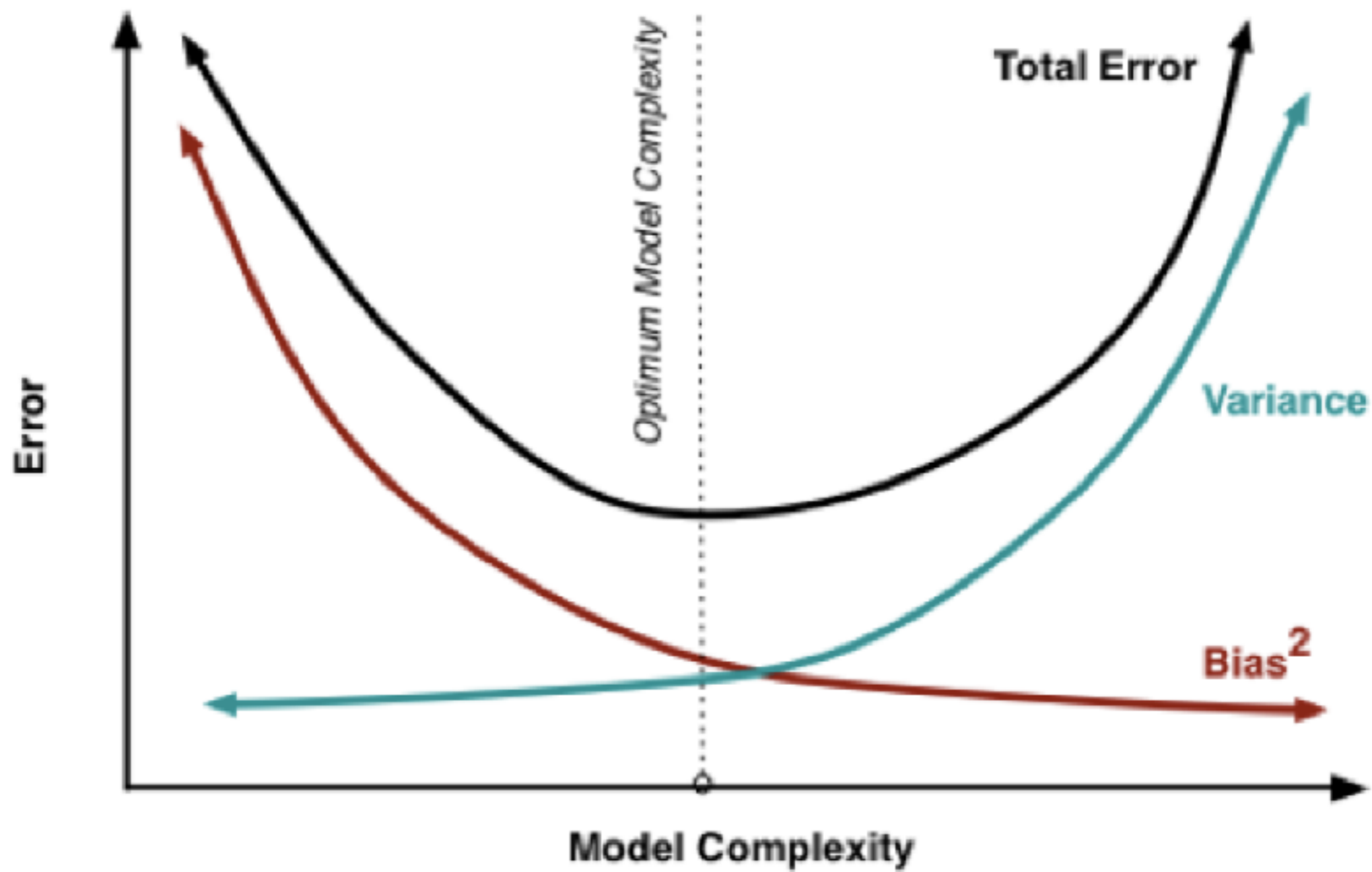
Estimate f from observations
 $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$

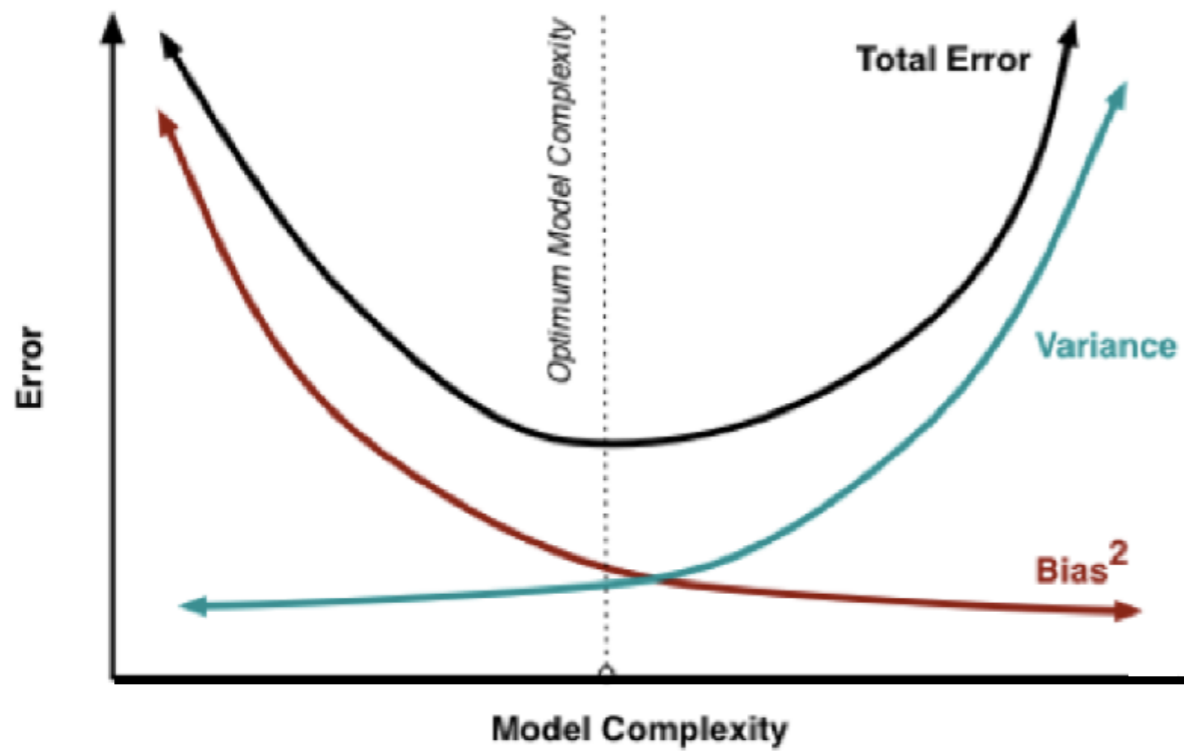
Hope that this also works on new examples.

Old ML Conventional Wisdom

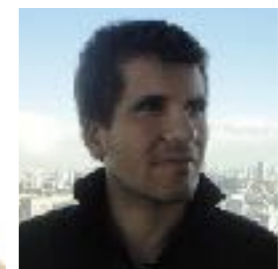
- Good prediction balances bias and variance.
- You should not perfectly fit your training data as some in-sample errors can reduce out-of-sample error.
- High capacity models don't generalize.
- Optimizing to high precision harms generalization.
- Nonconvex optimization is hard in machine learning.

None of these are true.

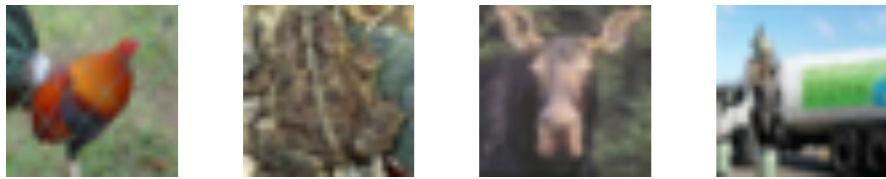




★
↑
Deep
models



Zhang, Bengio, Hardt, R., Vinyals



CIFAR10

$n=50,000$

$d=3,072$

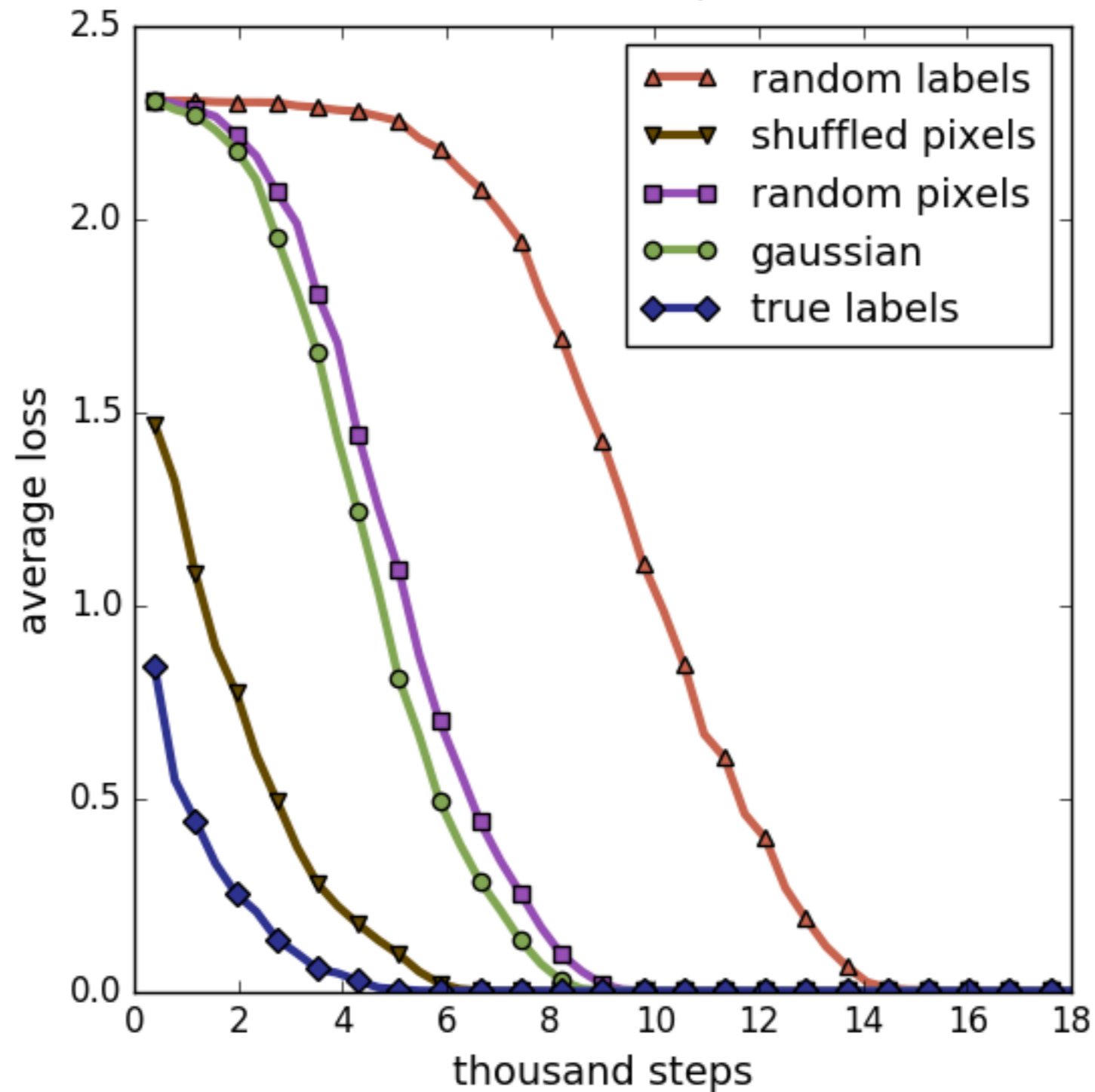
$k=10$

What happens when I turn off the regularizers?

<u>Model</u>	<u>parameters</u>	<u>p/n</u>	Train <u>loss</u>	Test <u>error</u>
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%

n=50,000
d=3,072
k=10
p=1,649,402

MicroInception

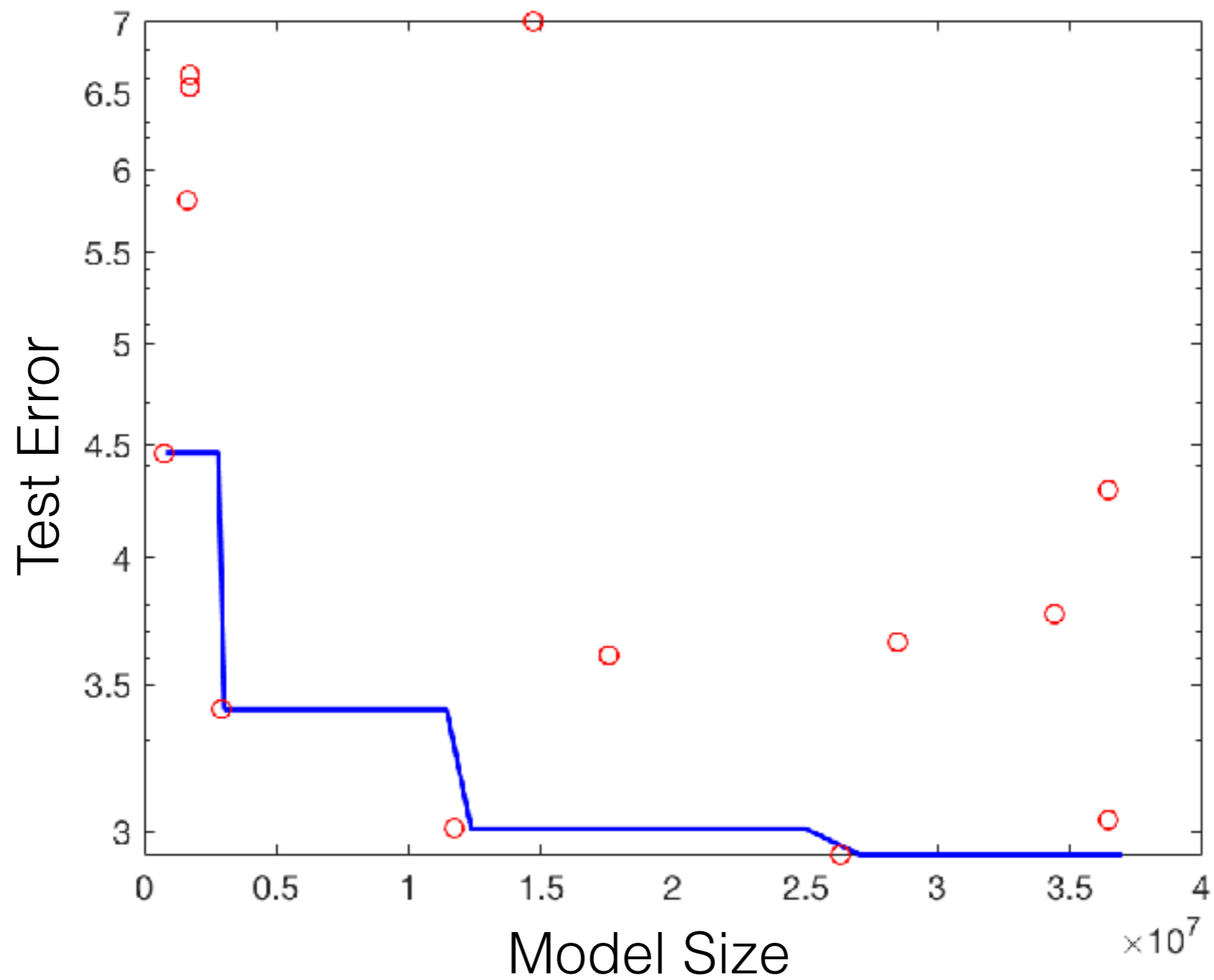


Neural Nets on CIFAR10

$n=50,000$

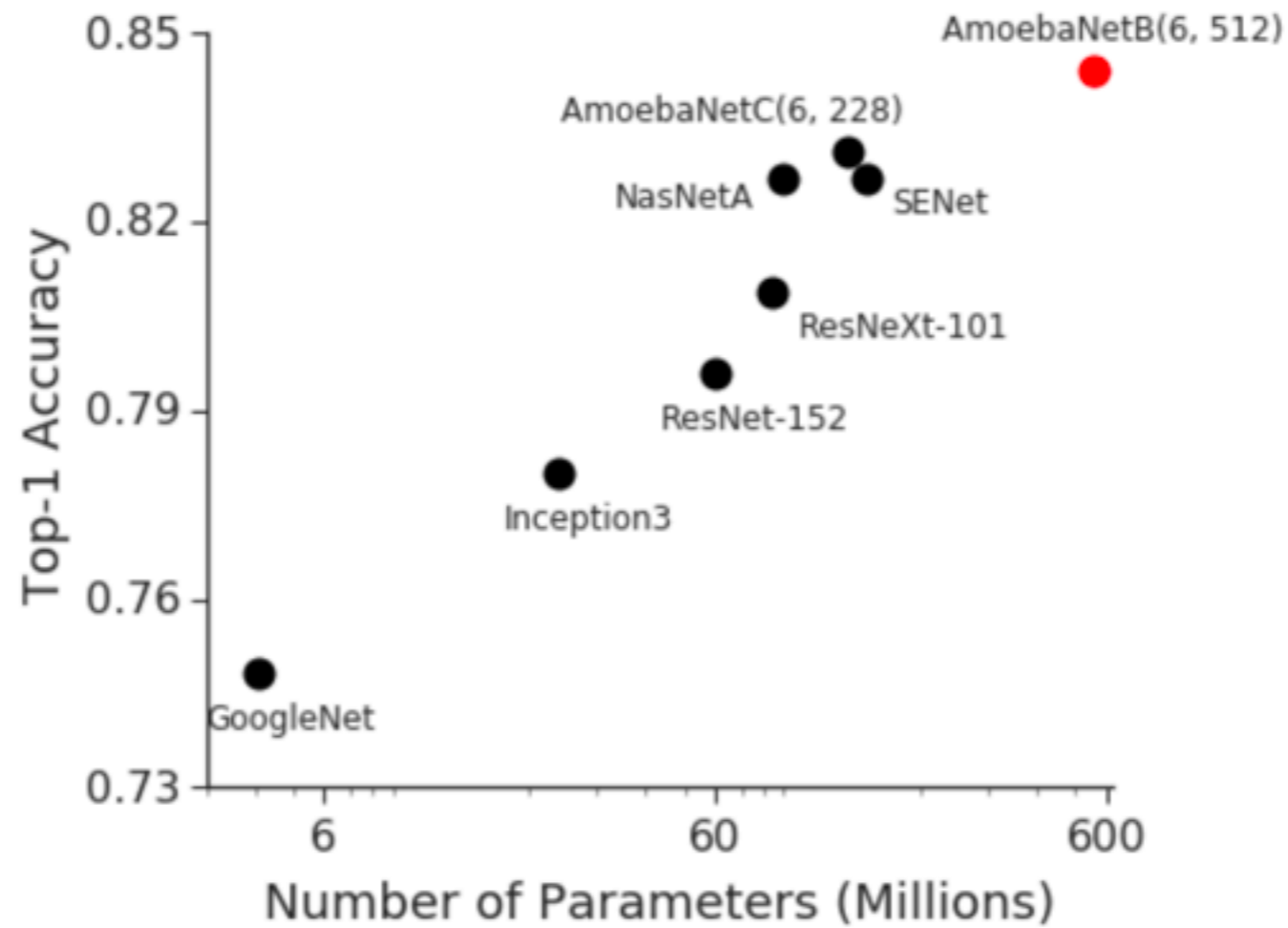
$d=3,072$

$k=10$



Cherry picked deep models for Imagenet

$n = 1.3M$
 $d = 150528$
 $k = 1000$

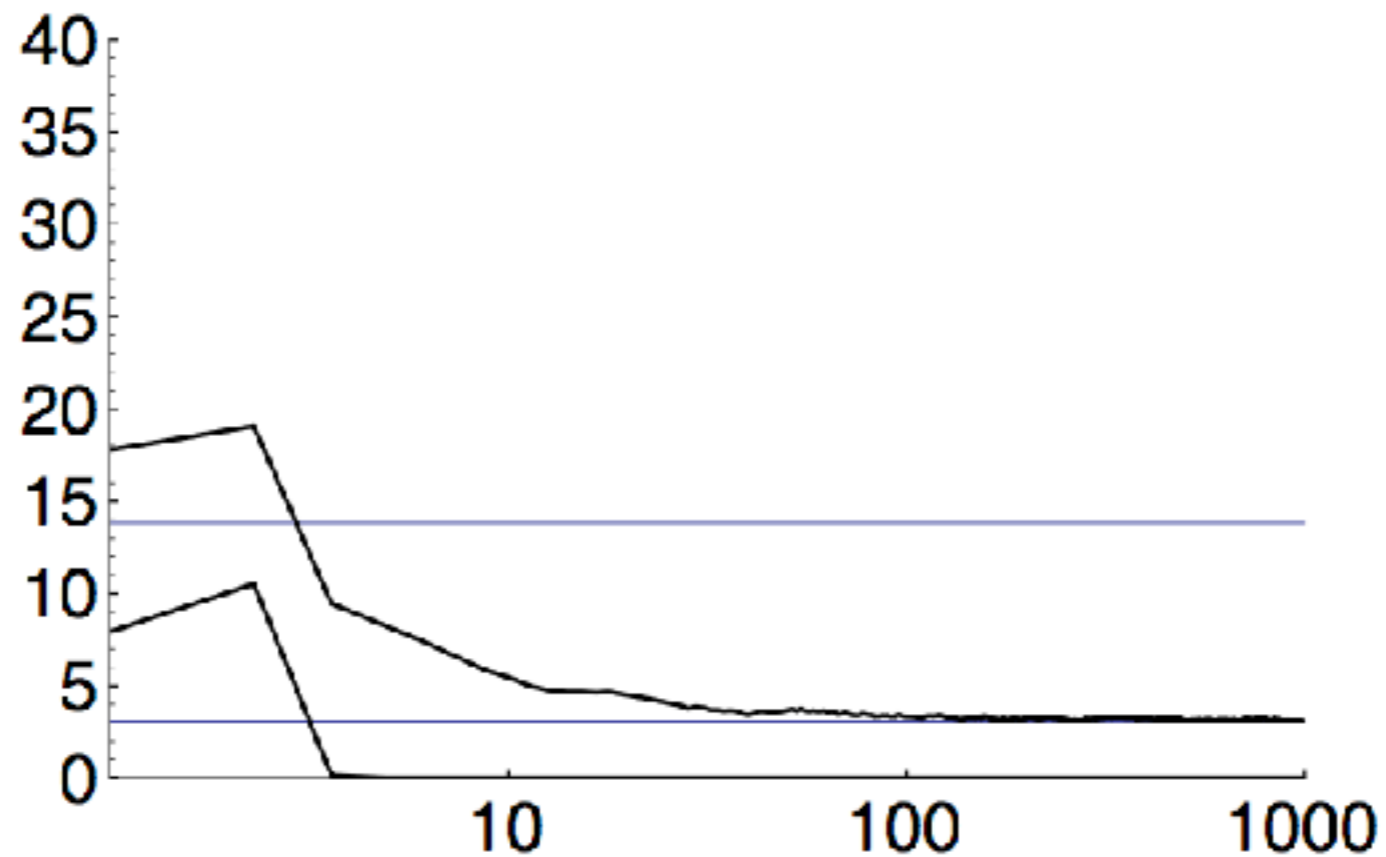


Boosting on UCI Letter data set.

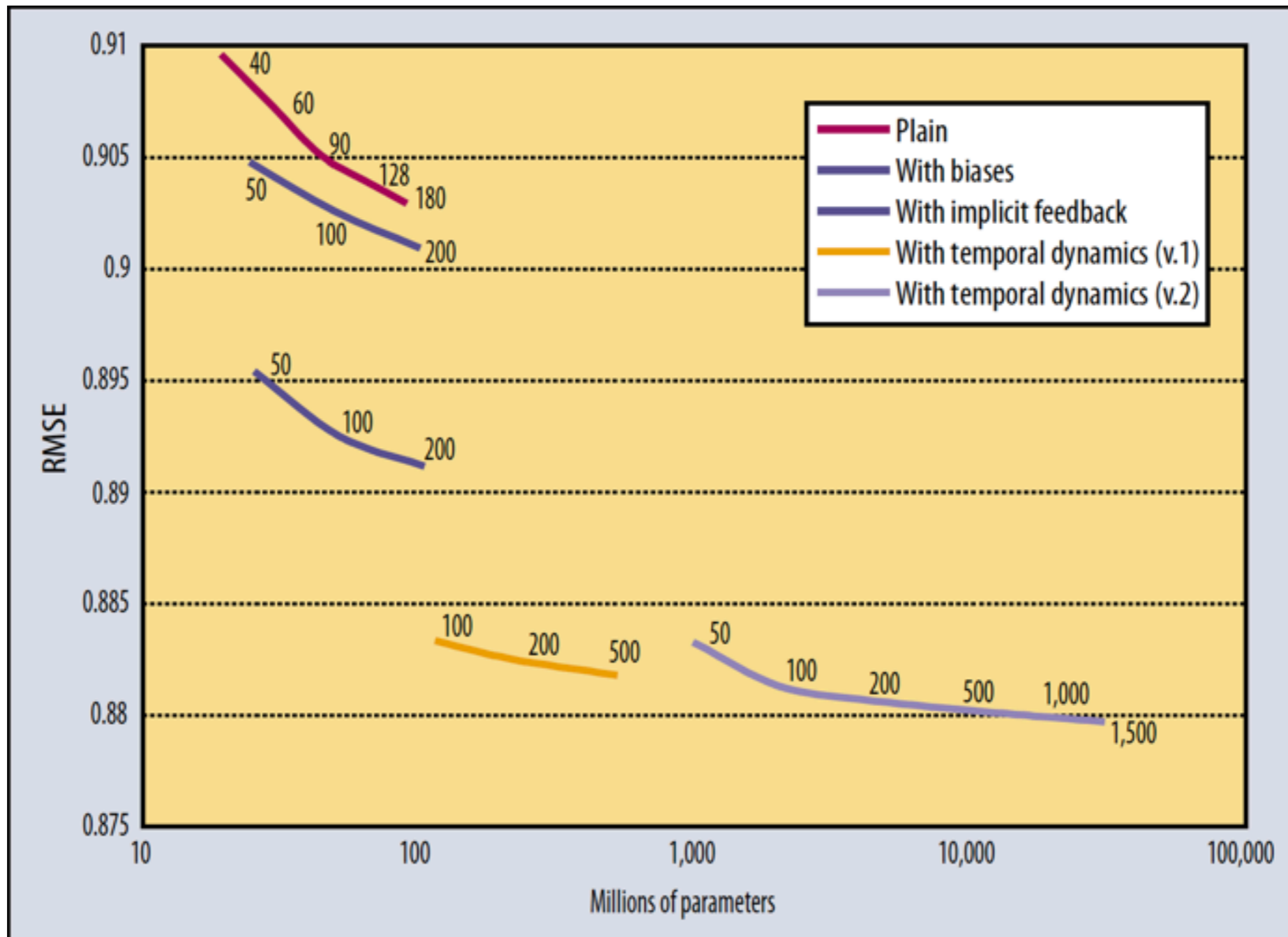
$n=16,000$

$d=16$

$k=26$



Performance on Netflix Prize



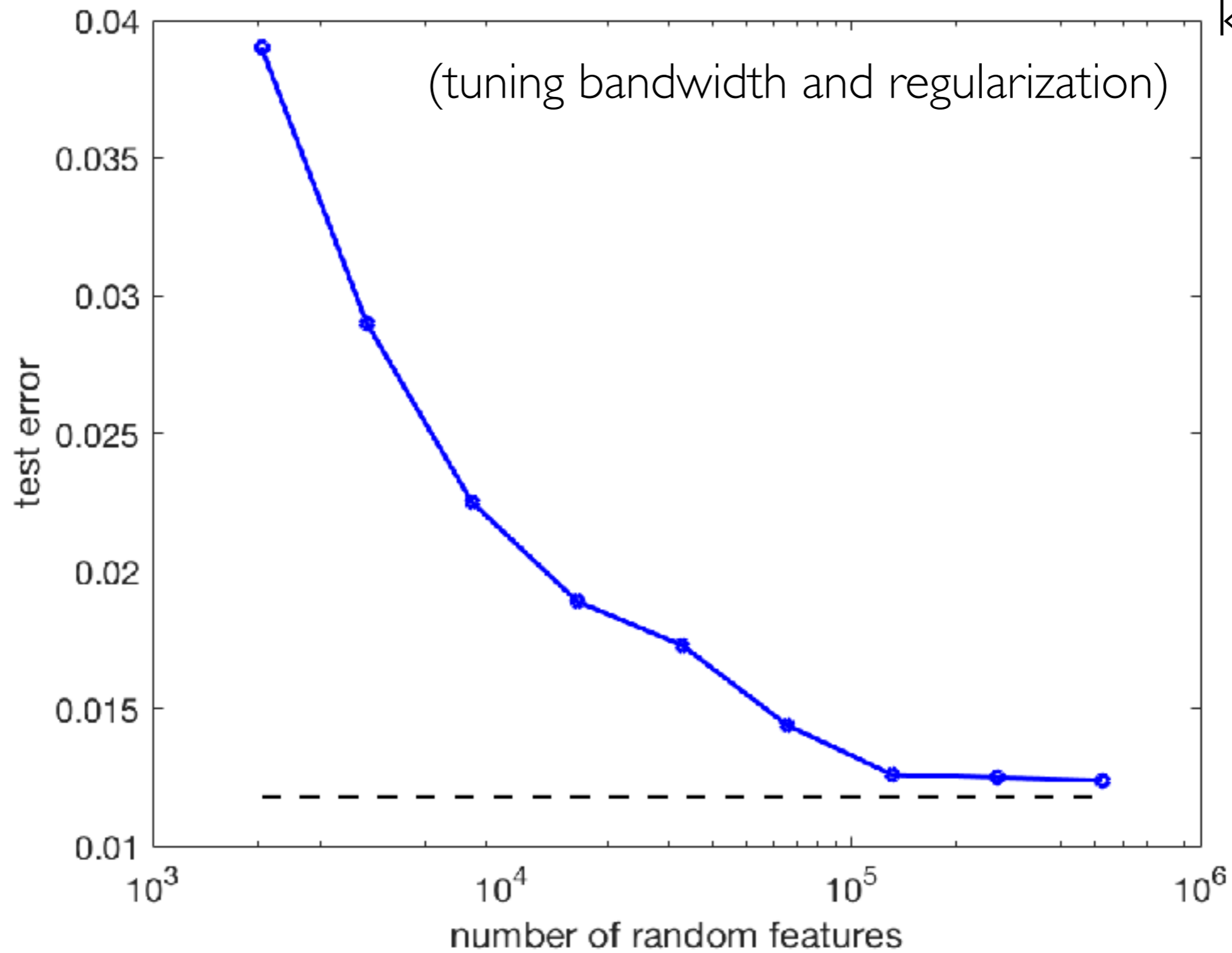
R. Bell and Y. Koren, CACM, 2009

MNIST: Cosine random features

$n=60,000$

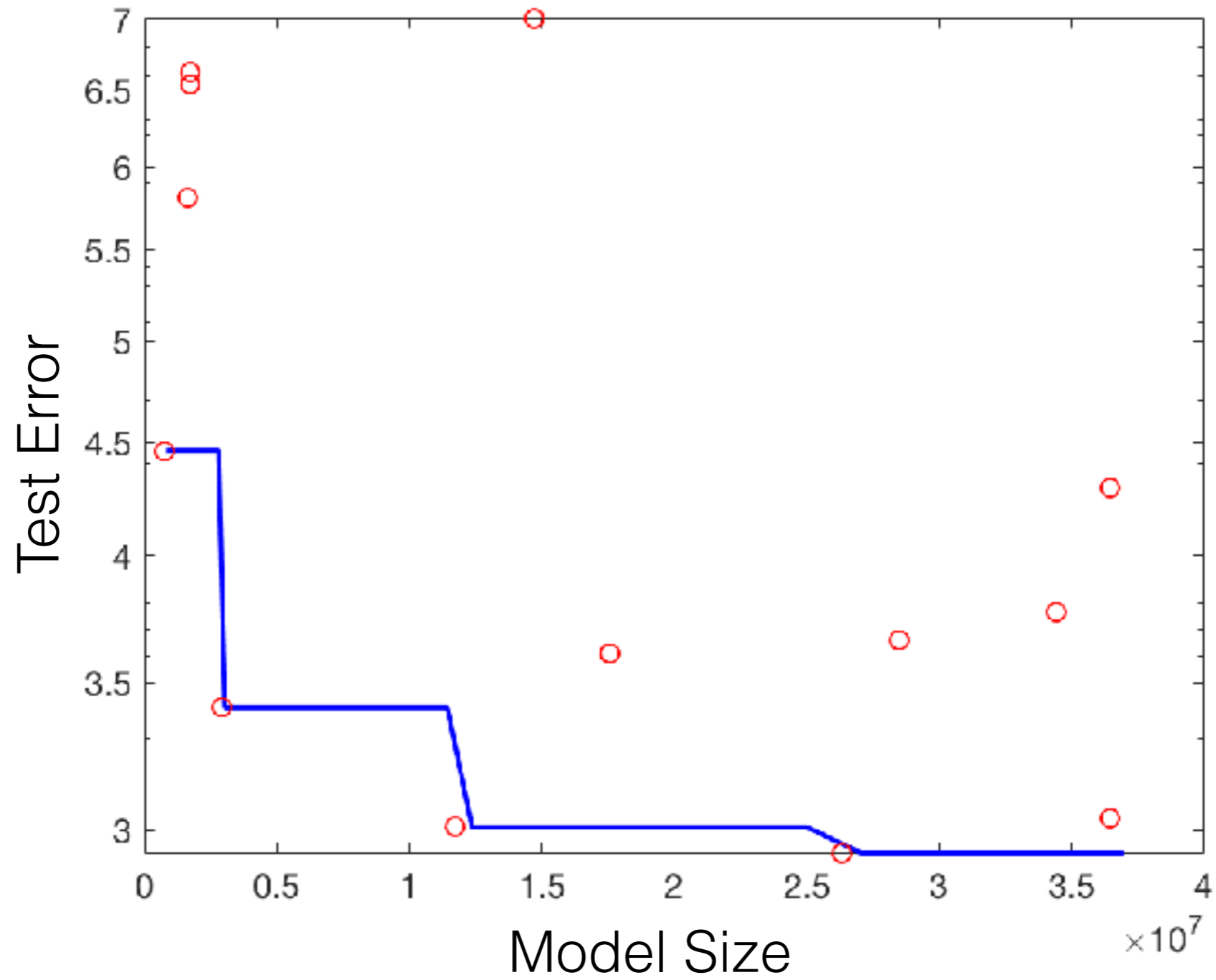
$d=784$

$k=10$



Neural Nets on CIFAR10

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



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.

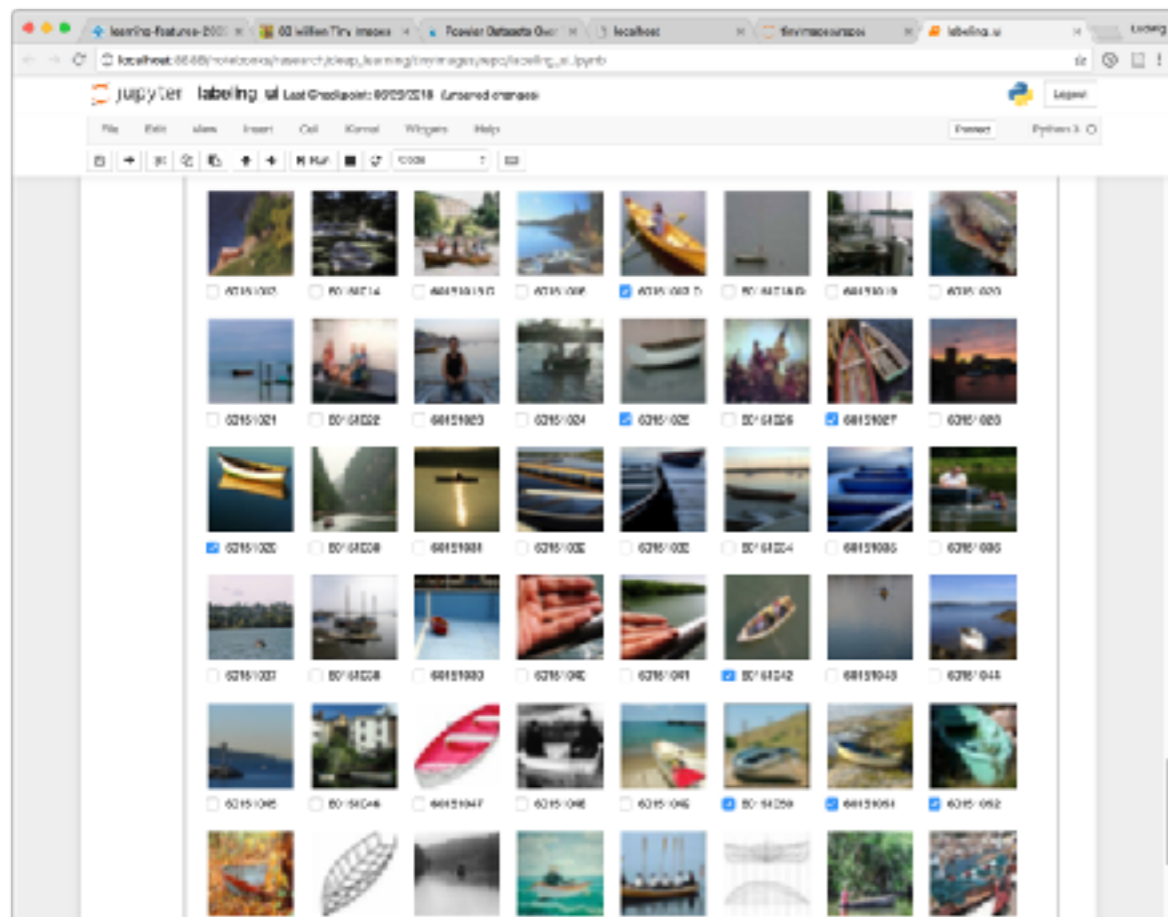
Deeeep networks

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



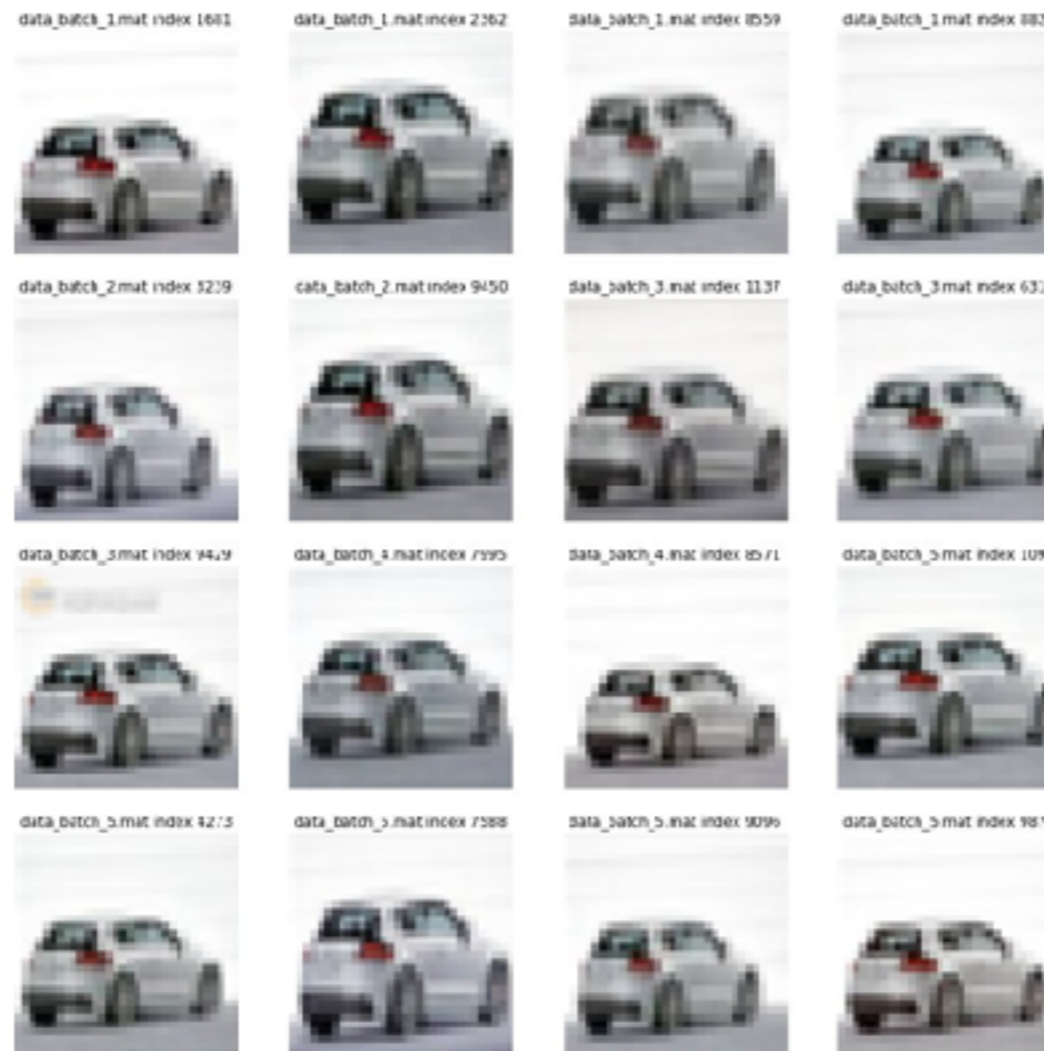
Can we get an i.i.d. resampling?

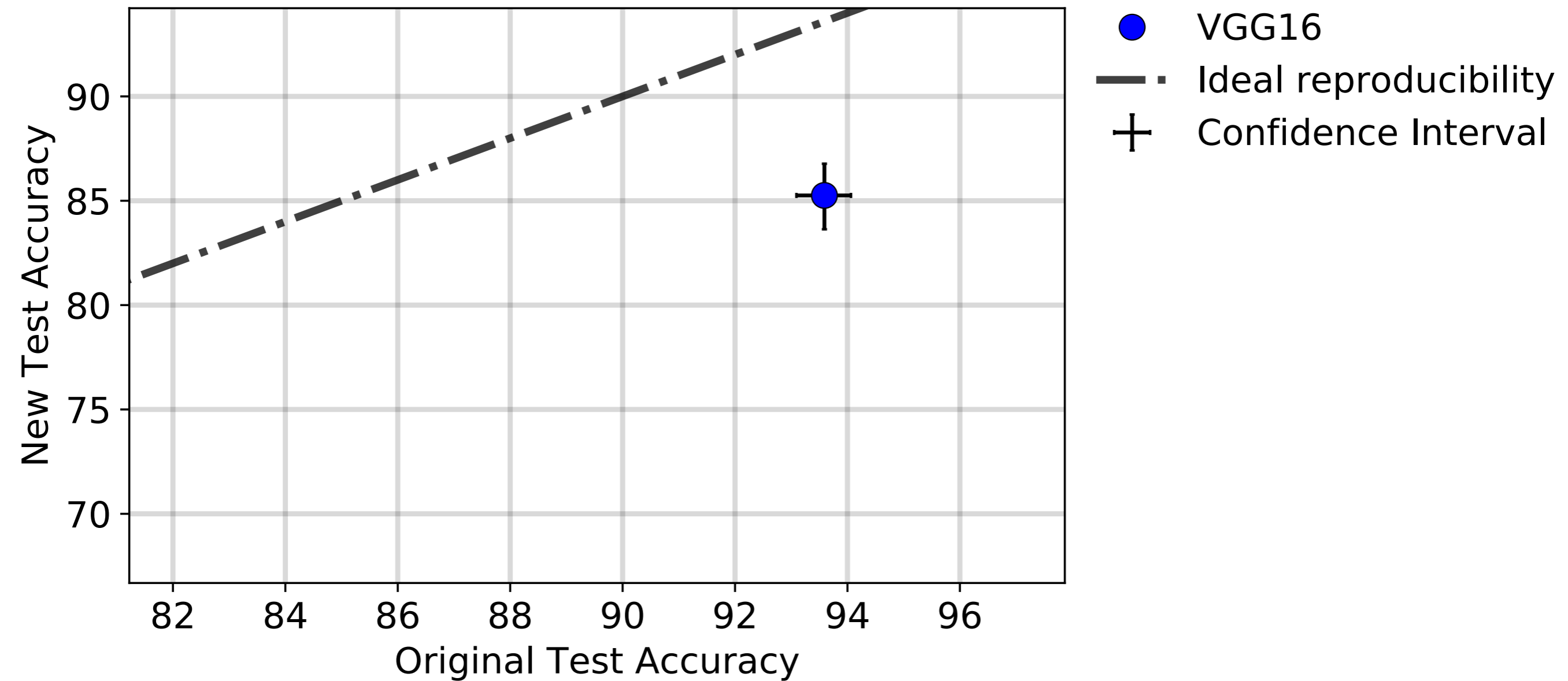


Roelofs, Schmidt, Shankar, R. 2018

Near-Duplicates in CIFAR-10

At least 8% of the original CIFAR-10 test set has a near-duplicate in the training set.

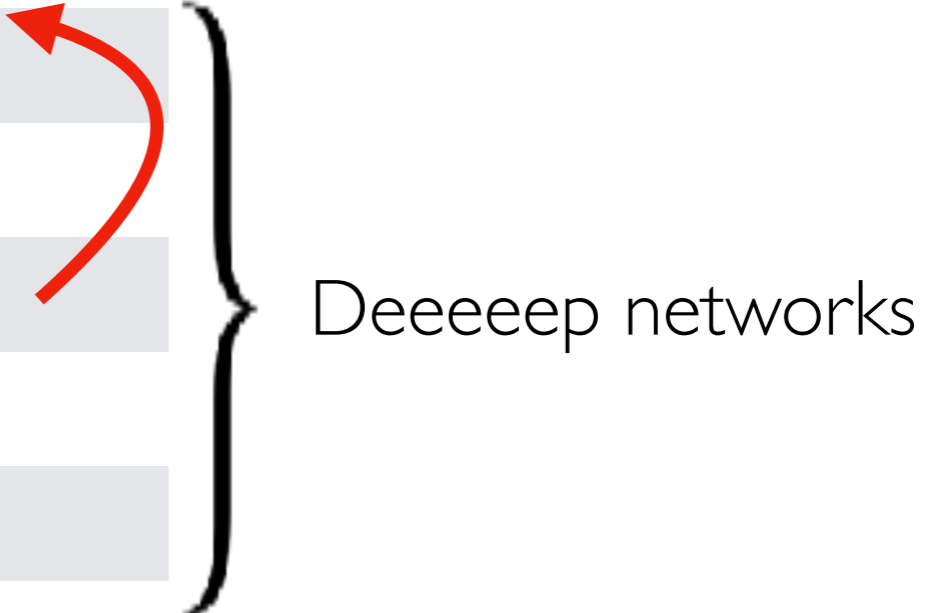




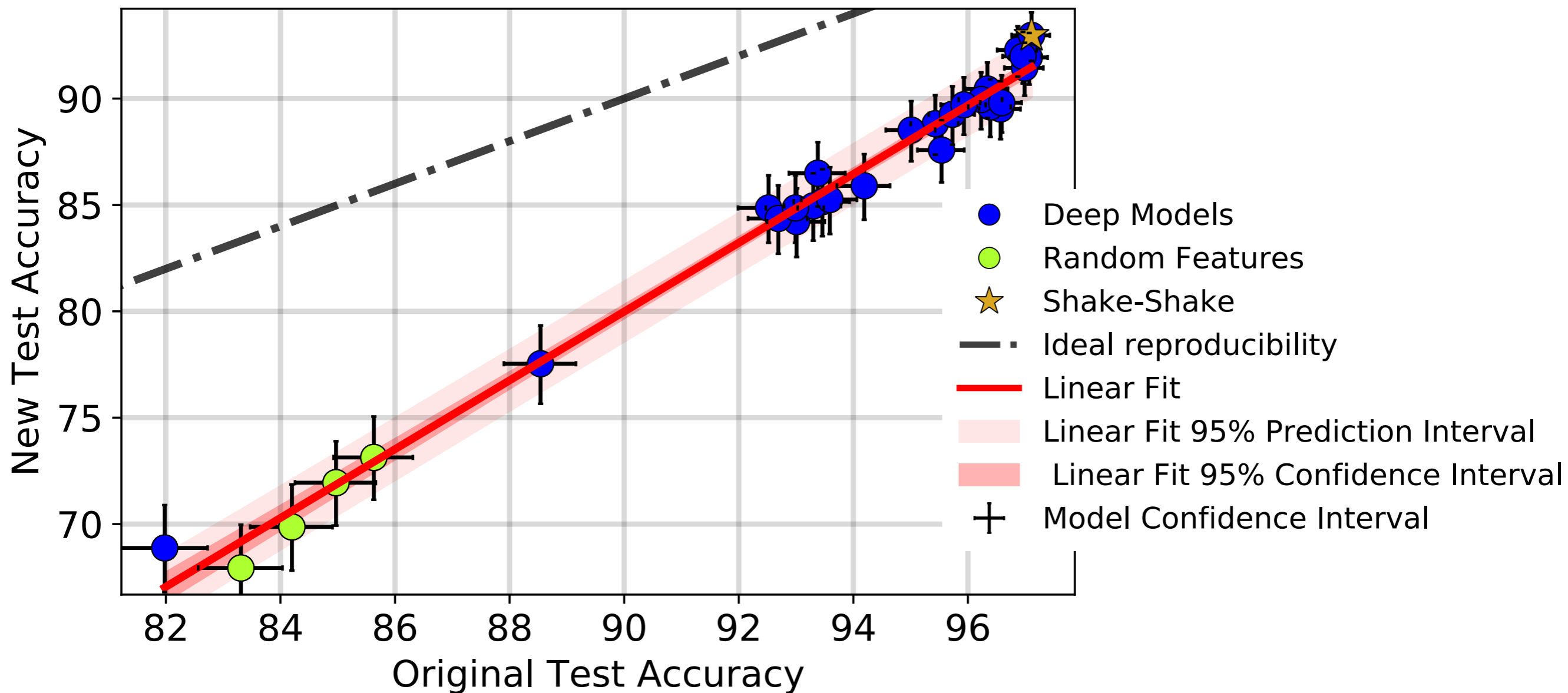
VGG16: 93.6% (original) ➔ 85.3% (new) 8% drop

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 Vaishal 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

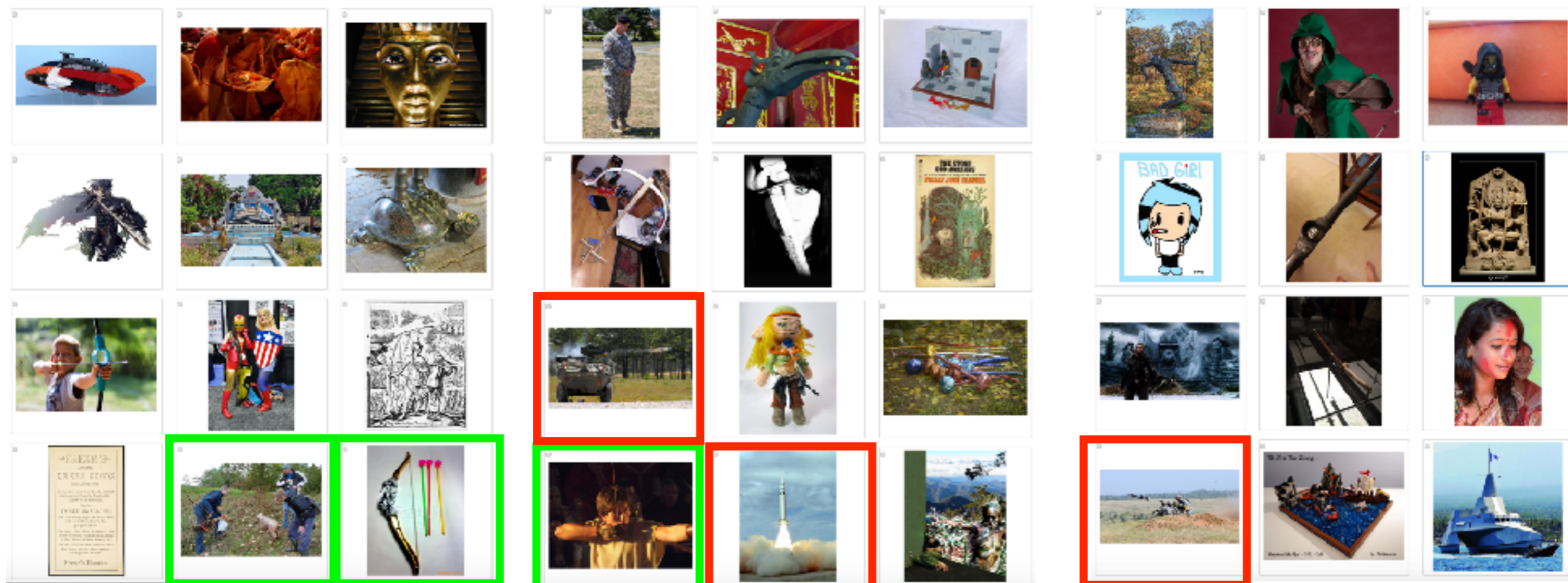
Task:

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).

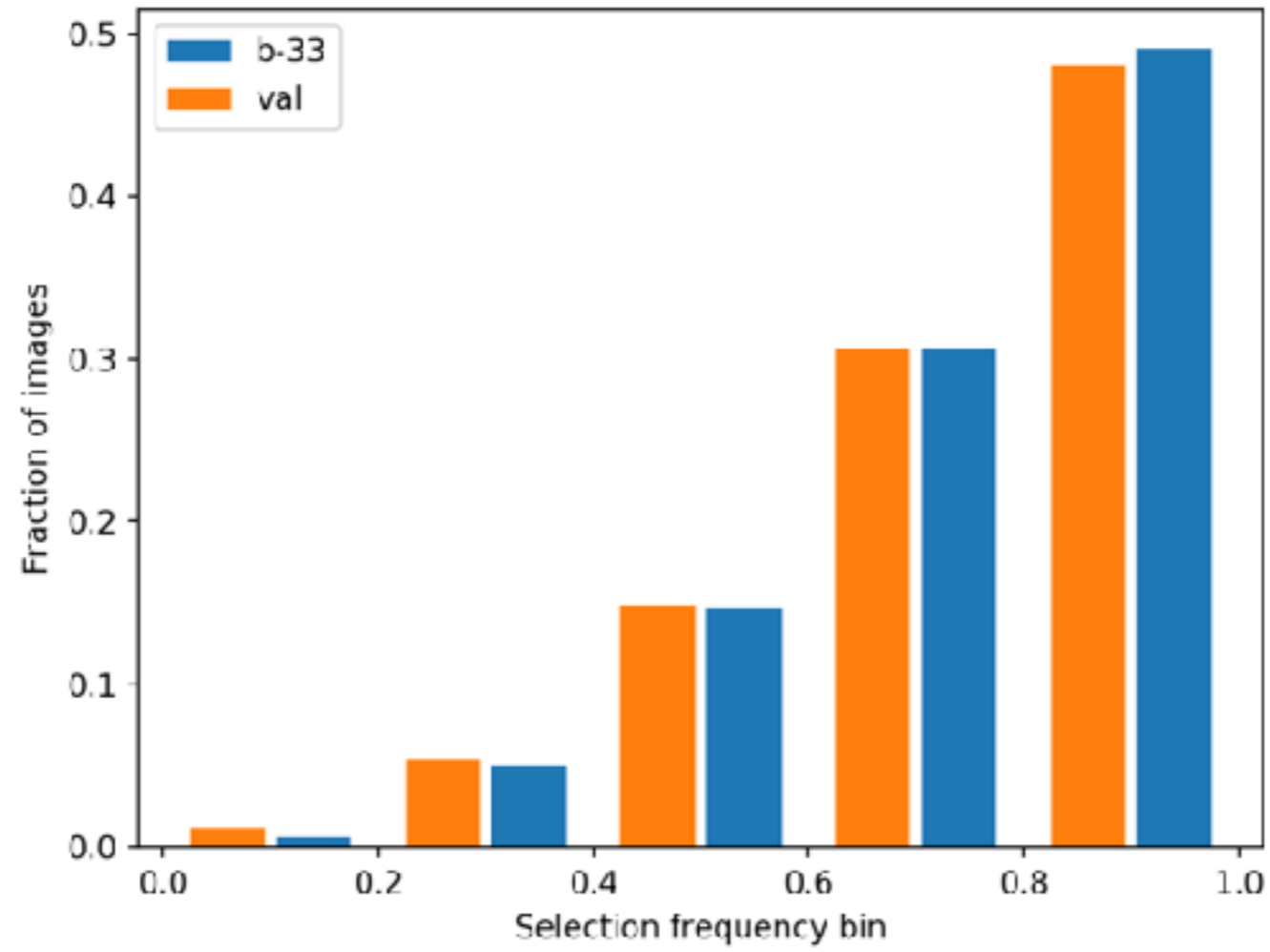
Please make accurate selections!

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



1.0



0.7

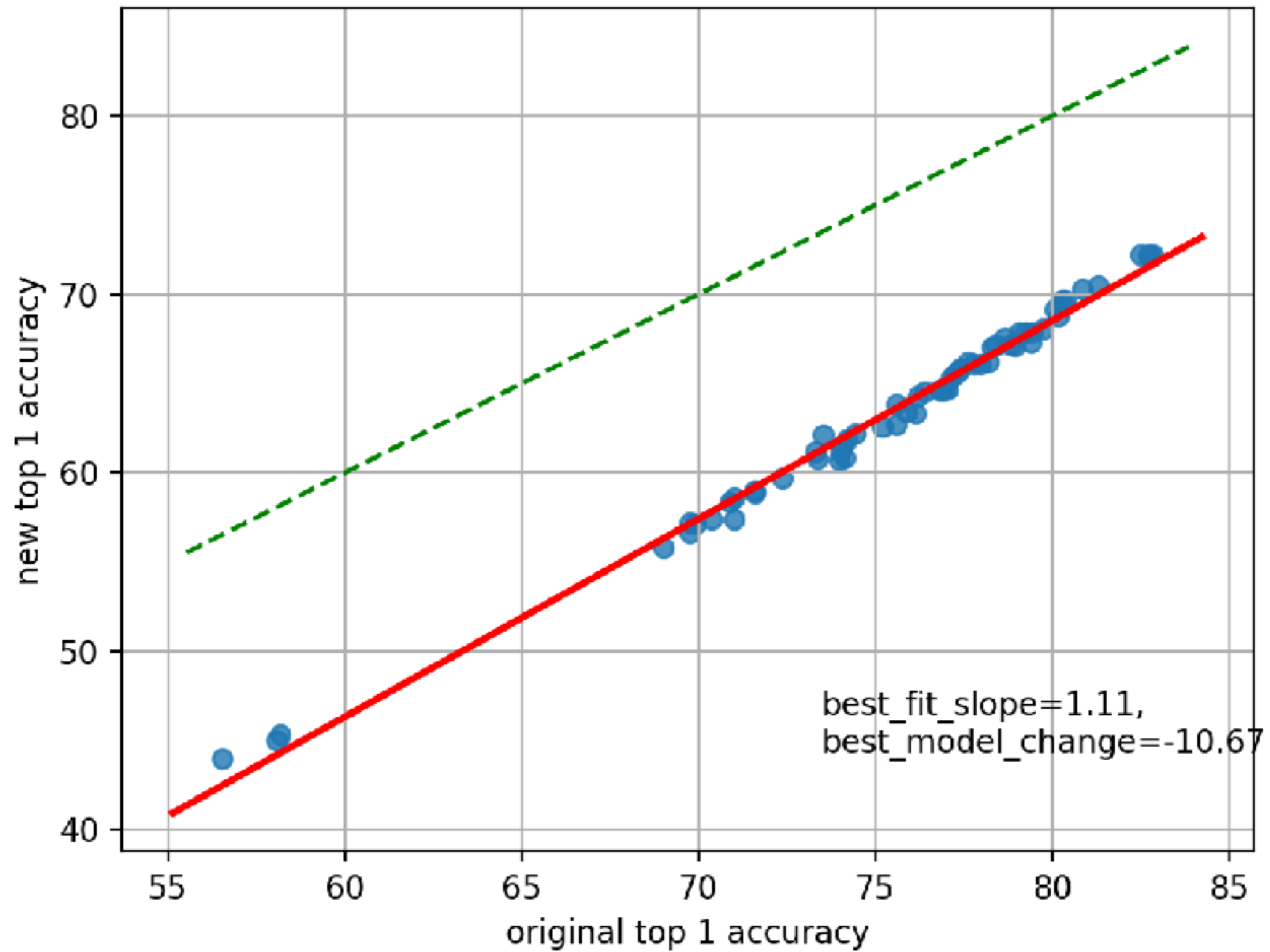


0.5



0.2

imagenetv2-b-33



What we have *always* seen

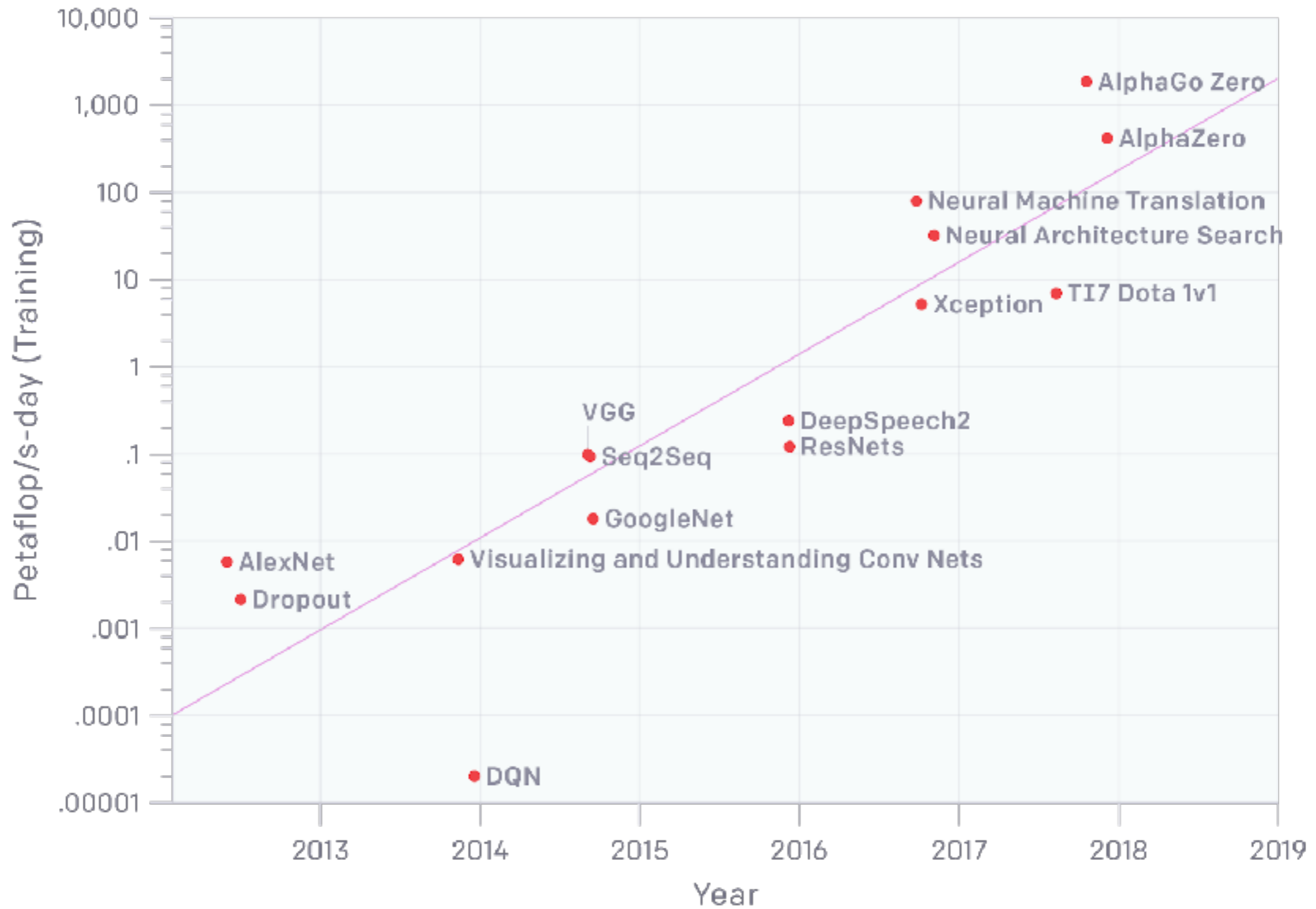
- Interpolating your training data is fine.
- Training on your test set is fine.
- Making models huge doesn't hurt.
- Making models huge doesn't help much.

Maybe we're just running ERM on the test set and the hypothesis space is ERM-ish solutions on the train set?

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*.

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



G. Brockman

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.

References

- “Understanding Deep Learning Requires Rethinking Generalization.” C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. In ICLR 2017.
- “Do CIFAR-10 Classifiers Generalize to CIFAR-10?” B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. 2018. [arXiv:1806.00451](https://arxiv.org/abs/1806.00451)