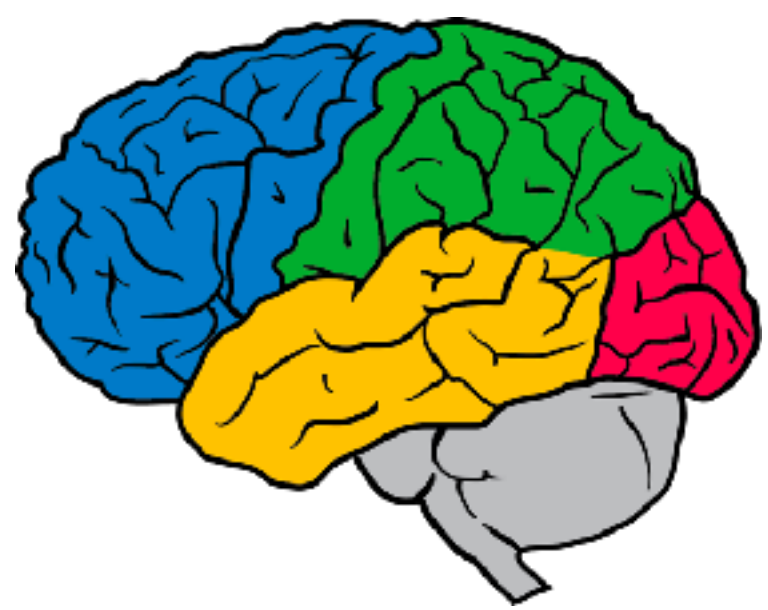
Defending Against Adversarial Examples

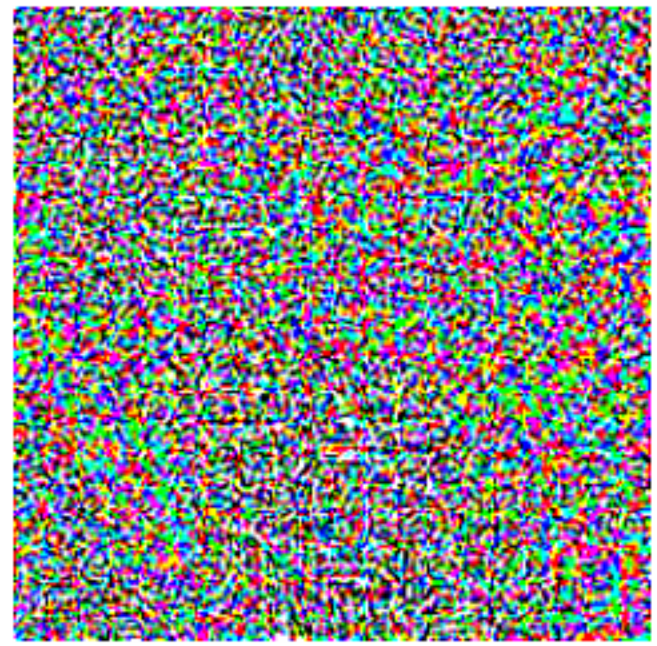
Ian Goodfellow, Staff Research Scientist, Google Brain

NIPS 2017 Workshop on Machine Learning and Security



Adversarial Examples

$+ .007 \times$



$=$

Timeline:

"Adversarial Classification" Dalvi et al 2004: fool spam filter

"Evasion Attacks Against Machine Learning at Test Time"

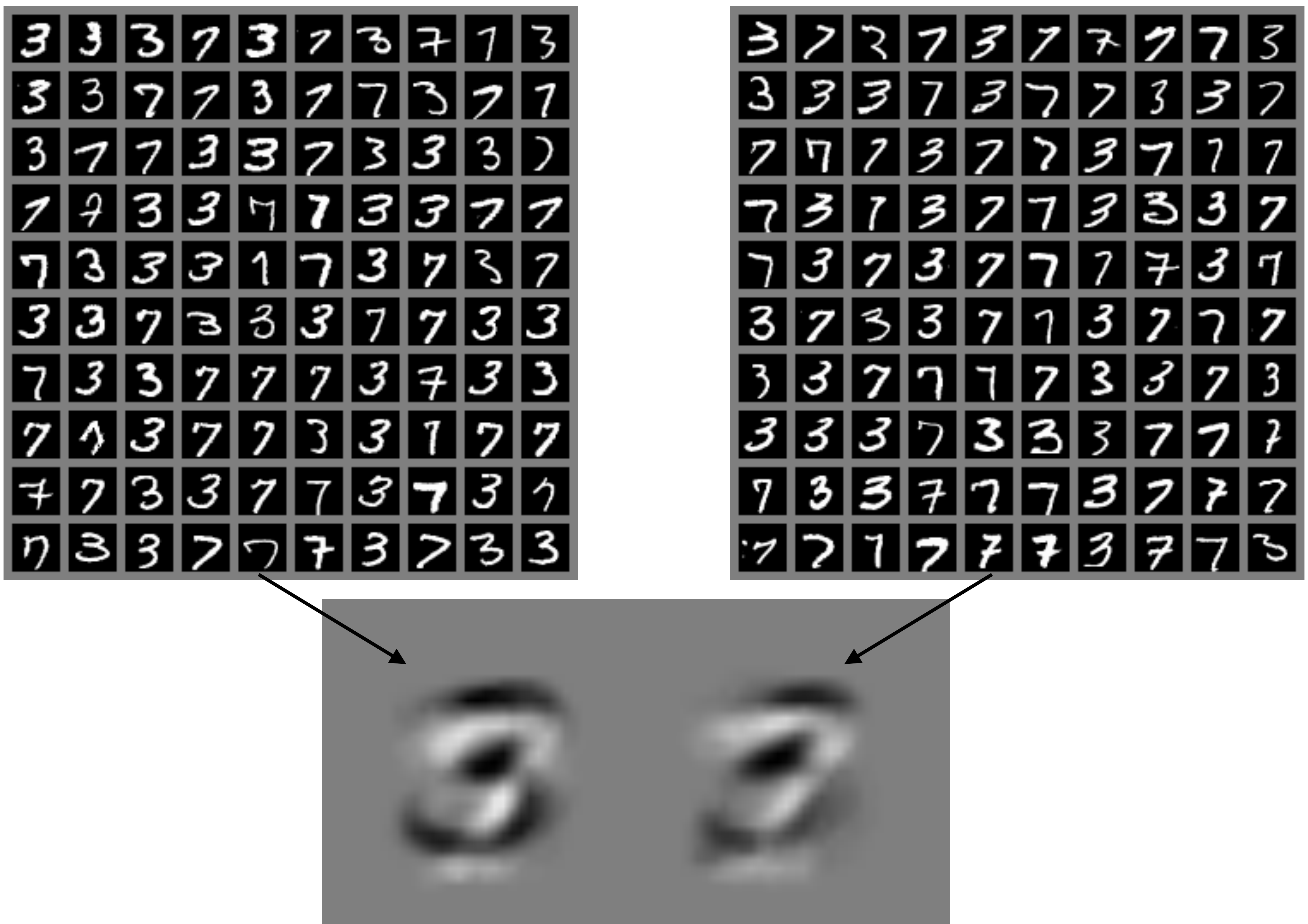
Biggio 2013: fool neural nets

Szegedy et al 2013: fool ImageNet classifiers imperceptibly

Goodfellow et al 2014: cheap, closed form attack

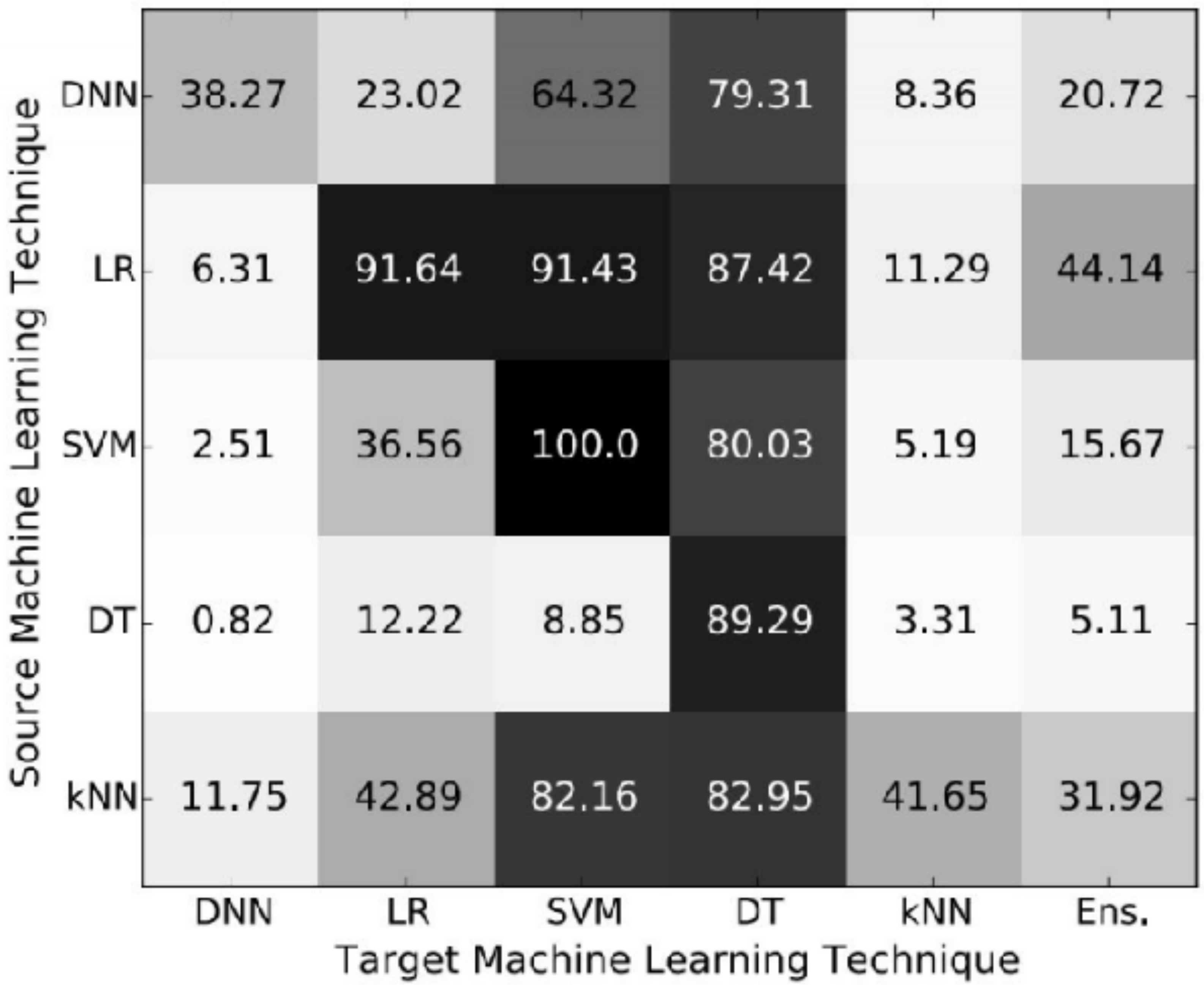
(Goodfellow 2017)

Cross-model, cross-dataset generalization



(Goodfellow 2017)

Cross-technique transferability



(Papernot 2016)

Enhancing Transfer With Ensembles

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
-ResNet-152	17.17	0%	0%	0%	0%	0%
-ResNet-101	17.25	0%	1%	0%	0%	0%
-ResNet-50	17.25	0%	0%	2%	0%	0%
-VGG-16	17.80	0%	0%	0%	6%	0%
-GoogLeNet	17.41	0%	0%	0%	0%	5%

Table 4: Accuracy of non-targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell (i, j) corresponds to the accuracy of the attack generated using four models except model i (row) when evaluated over model j (column). In each row, the minus sign "−" indicates that the model of the row is not used when generating the attacks. Results of top-5 accuracy can be found in the appendix (Table 14).

(Liu et al, 2016)

Transferability Attack

Target model with unknown weights, machine learning algorithm, training set; maybe non-differentiable

Train your own model

Substitute model mimicking target model with known, differentiable function

Adversarial crafting against substitute

Adversarial examples

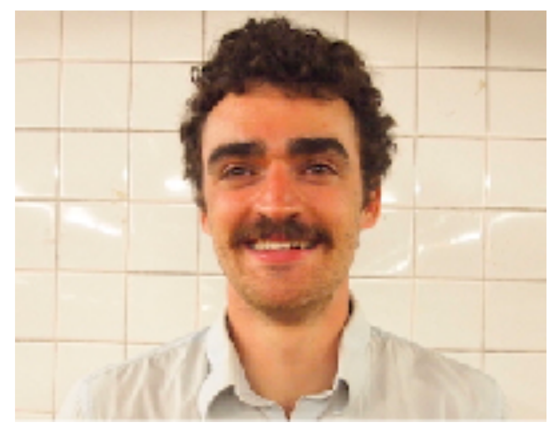
Deploy adversarial examples against the target; transferability property results in them succeeding

(Szegedy 2013, Papernot 2016)

Thermometer Encoding: One Hot Way to Resist Adversarial Examples


Jacob Buckman*

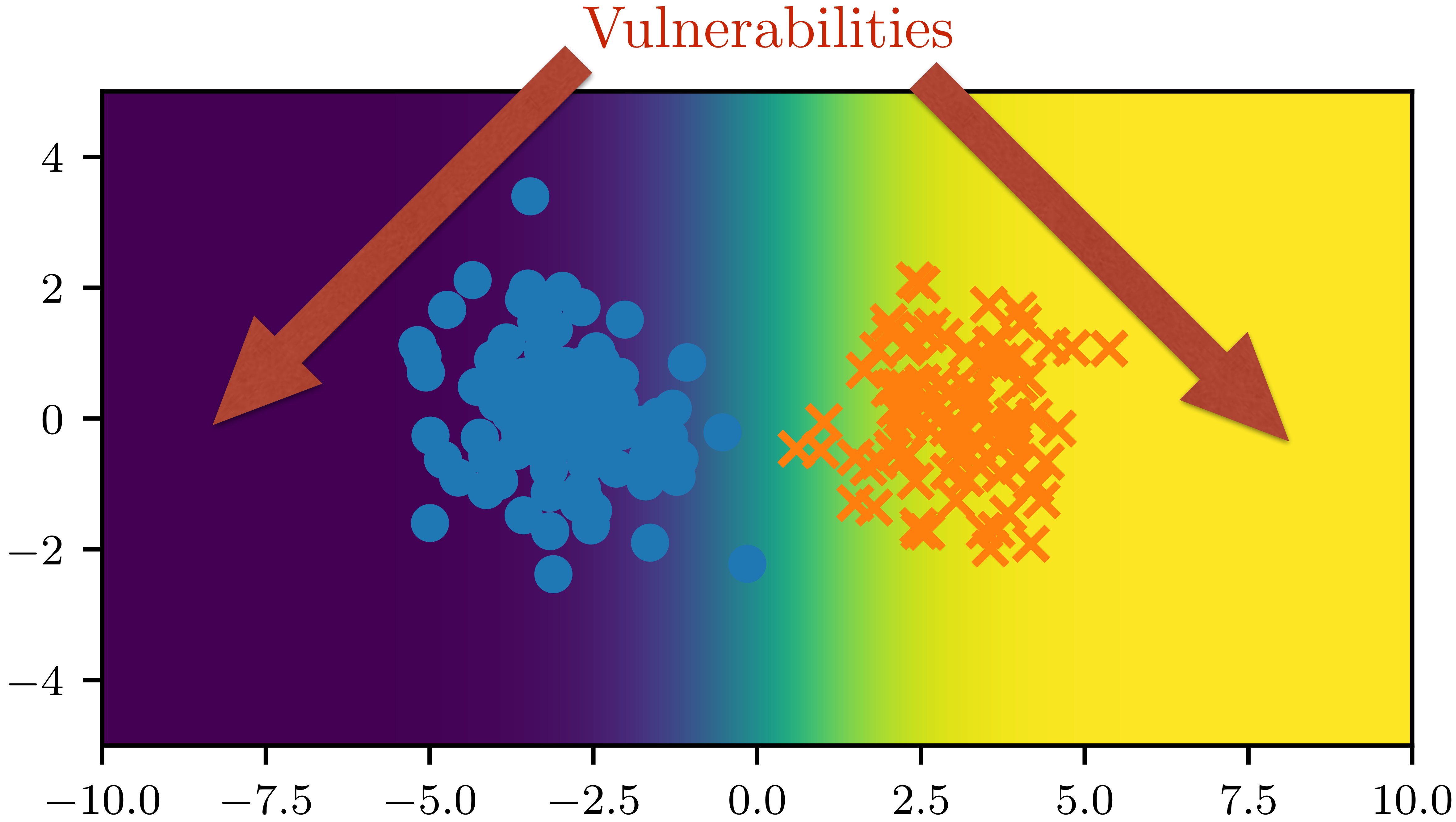

Aurko Roy*


Colin Raffel


Ian Goodfellow

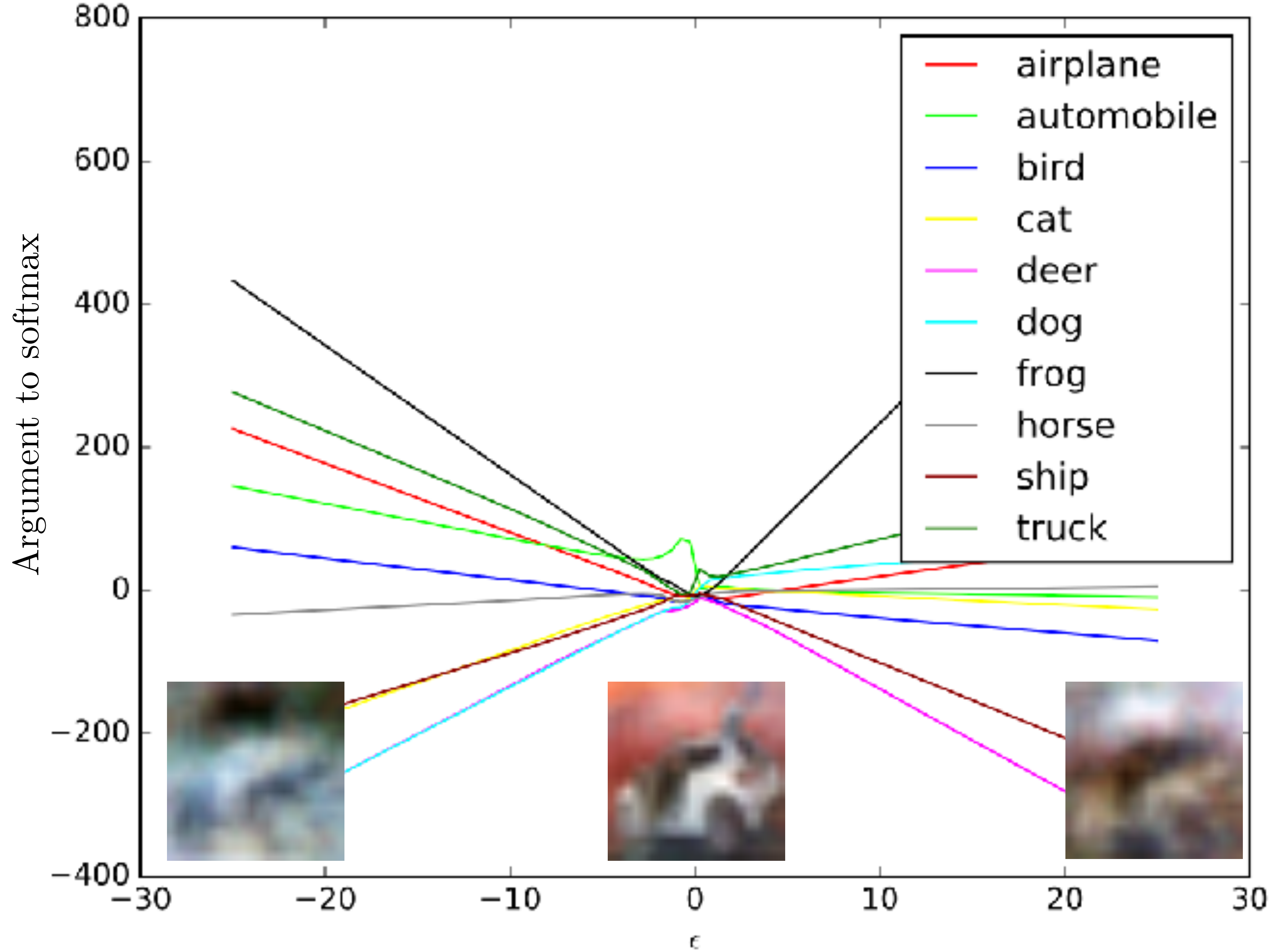
*joint first author

Linear Extrapolation



(Goodfellow 2017)

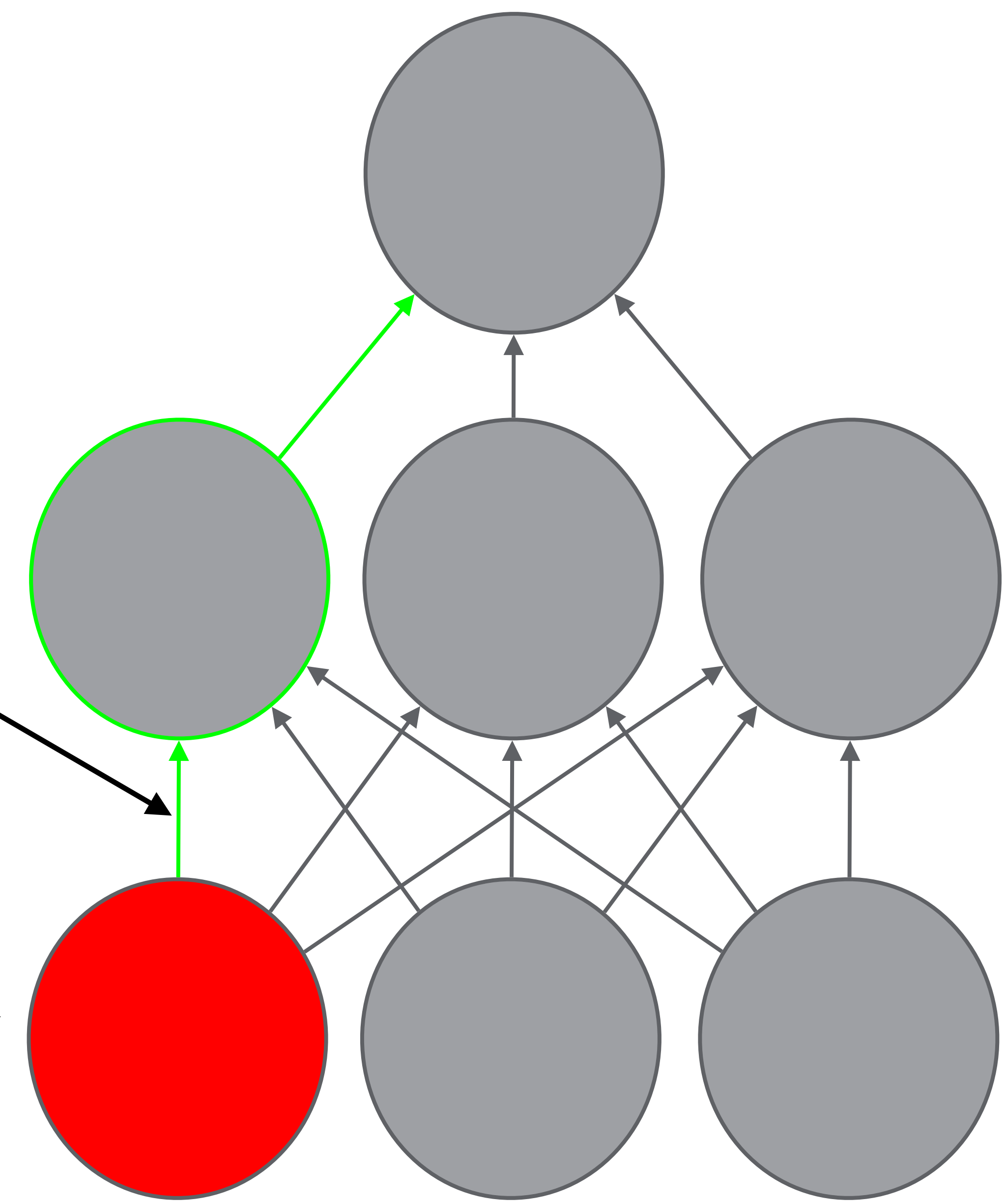
Neural nets are “too linear”



Plot from “Explaining and Harnessing Adversarial Examples”, Goodfellow et al, 2014

Difficult to train extremely nonlinear hidden layers

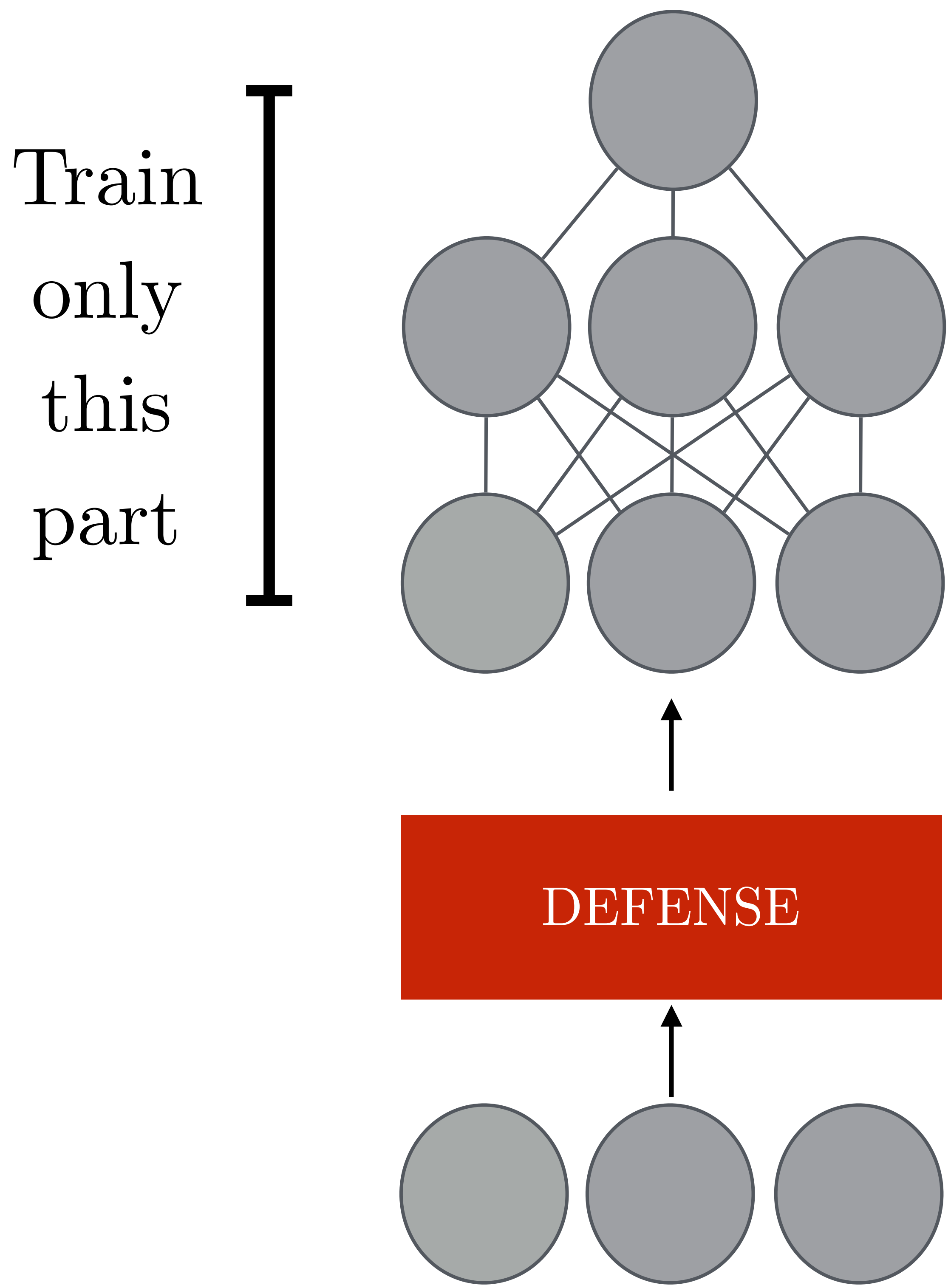
To train:
changing this weight needs to have a large, predictable effect



To defend:
changing this input needs to have a small or unpredictable effect

(Goodfellow 2017)

Idea: edit only the input layer



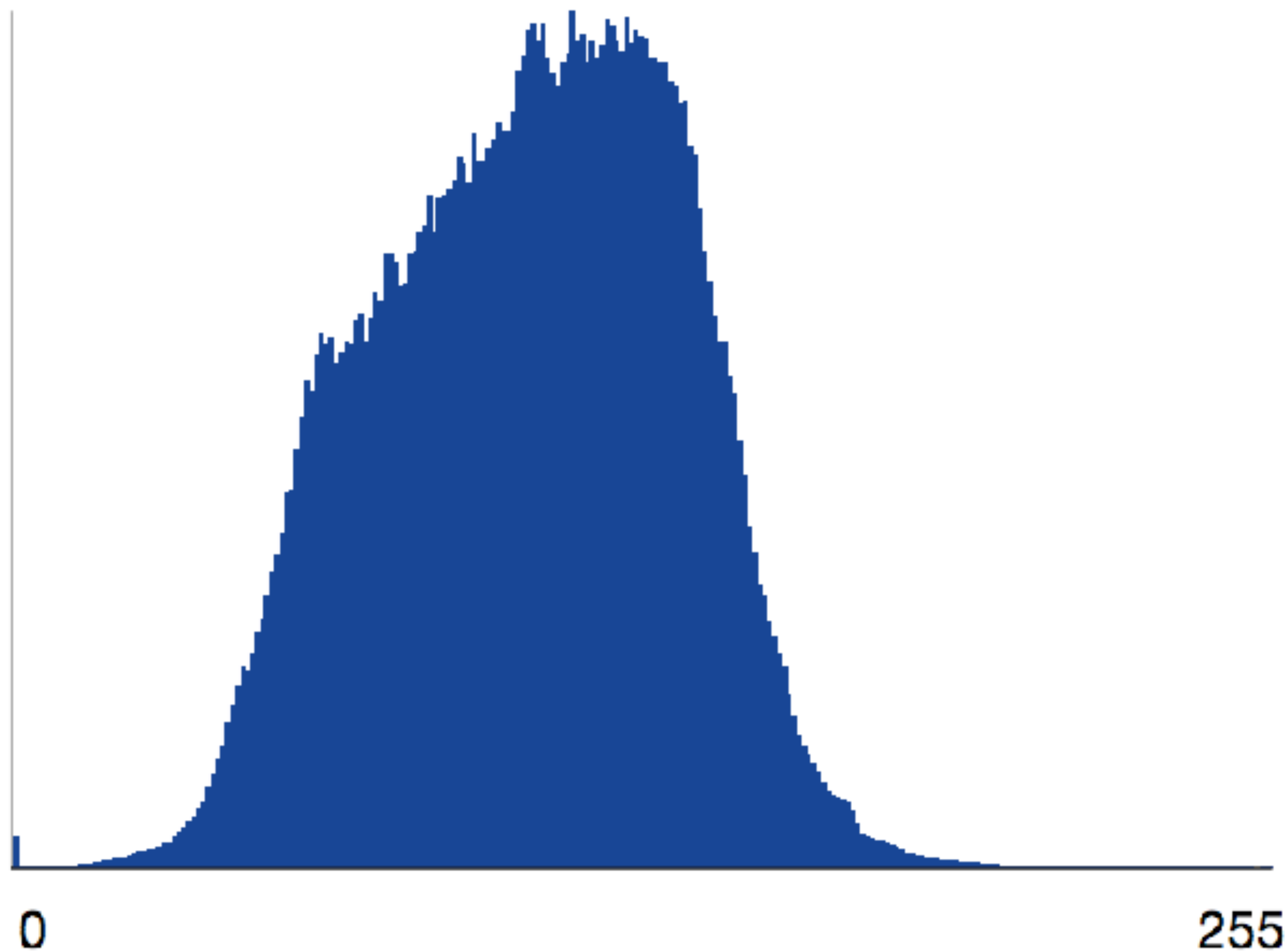
(Goodfellow 2017)

Real-valued	Quantized
0.13	0.15
0.66	0.65
0.92	0.95

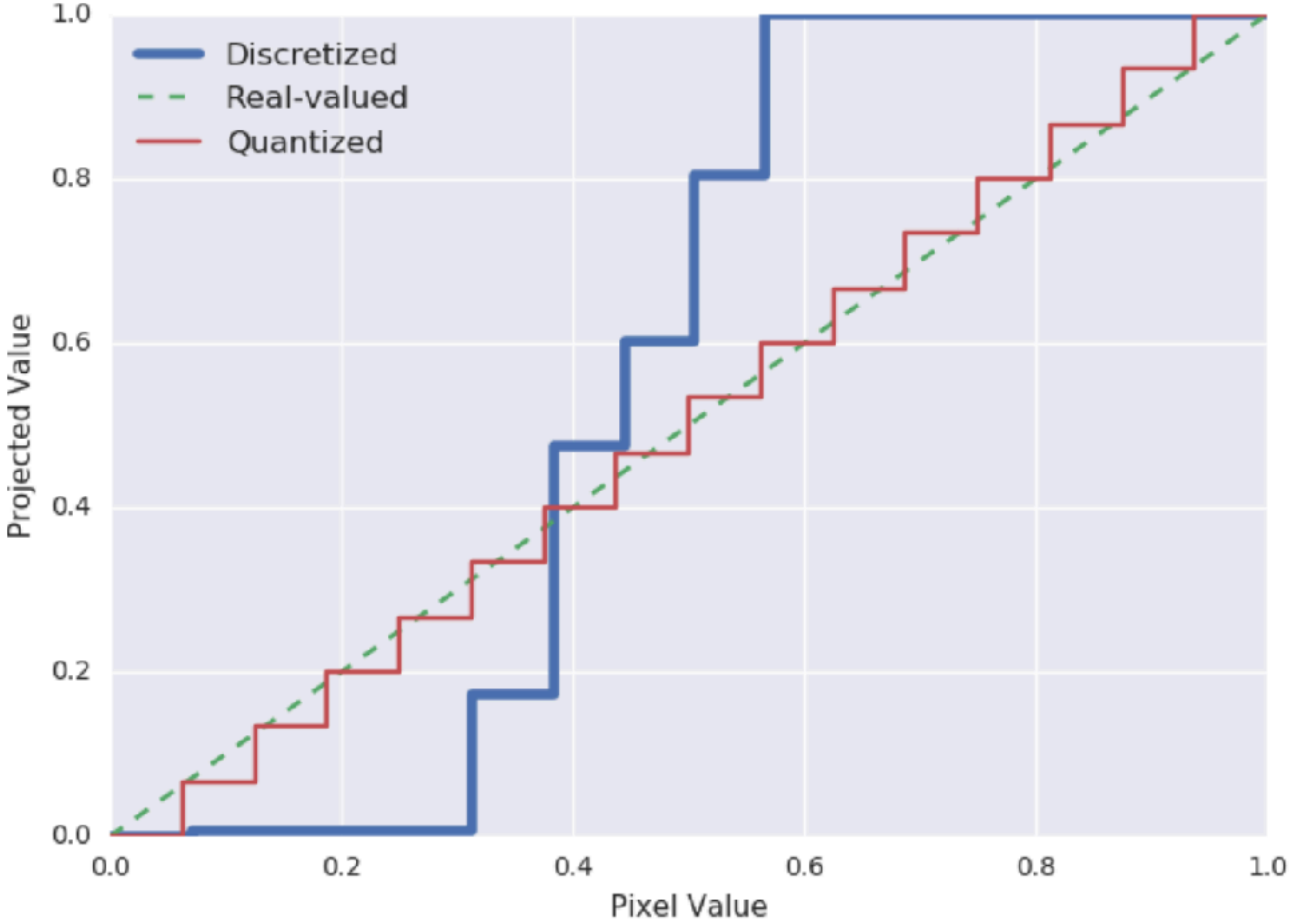
Discretized (one-hot)	Discretized (thermometer)
[0100000000]	[0111111111]
[0000001000]	[0000001111]
[0000000001]	[0000000001]

(Goodfellow 2017)

Observation: PixelRNN shows one-hot codes work

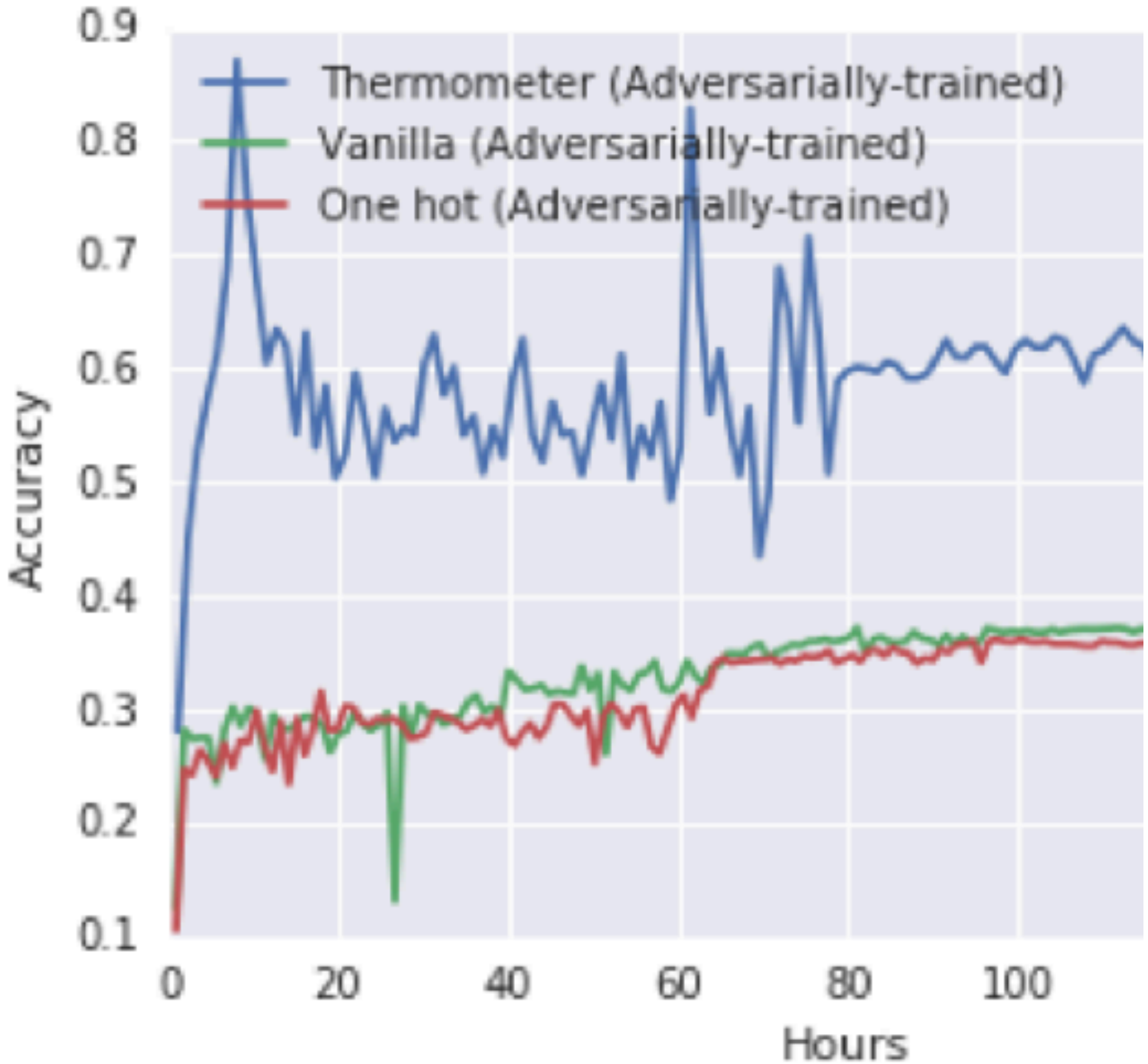


Plot from "Pixel Recurrent Neural Networks", van den Oord et al, 2016



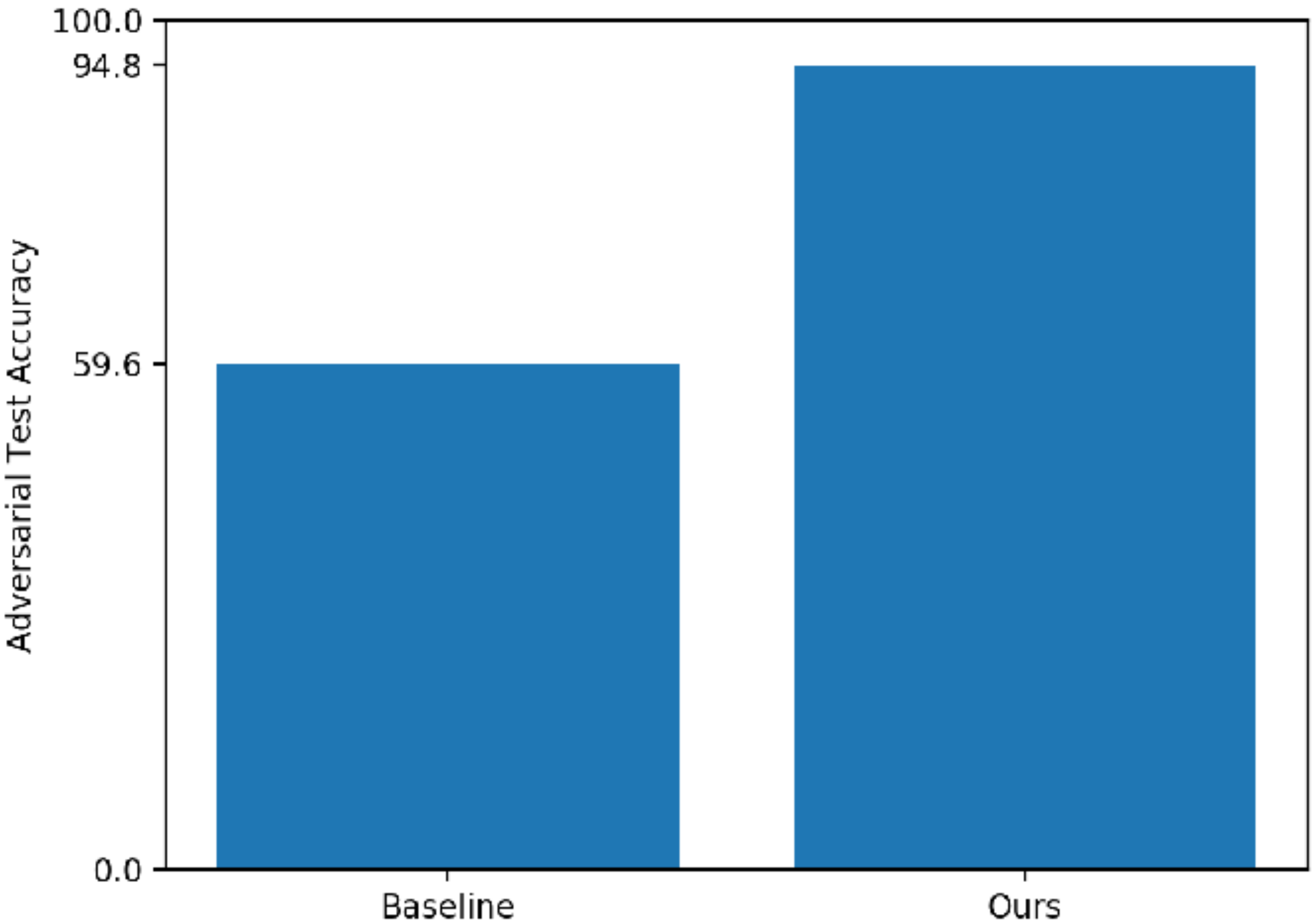
Discretized
Real-valued
Quantized
1.0
0.8
0.6
0.4
0.2
0.0
Projected Value
0.0
0.2
0.4
0.6
0.8
1.0
Pixel Value

Fast Improvement Early in Learning



(Goodfellow 2017)

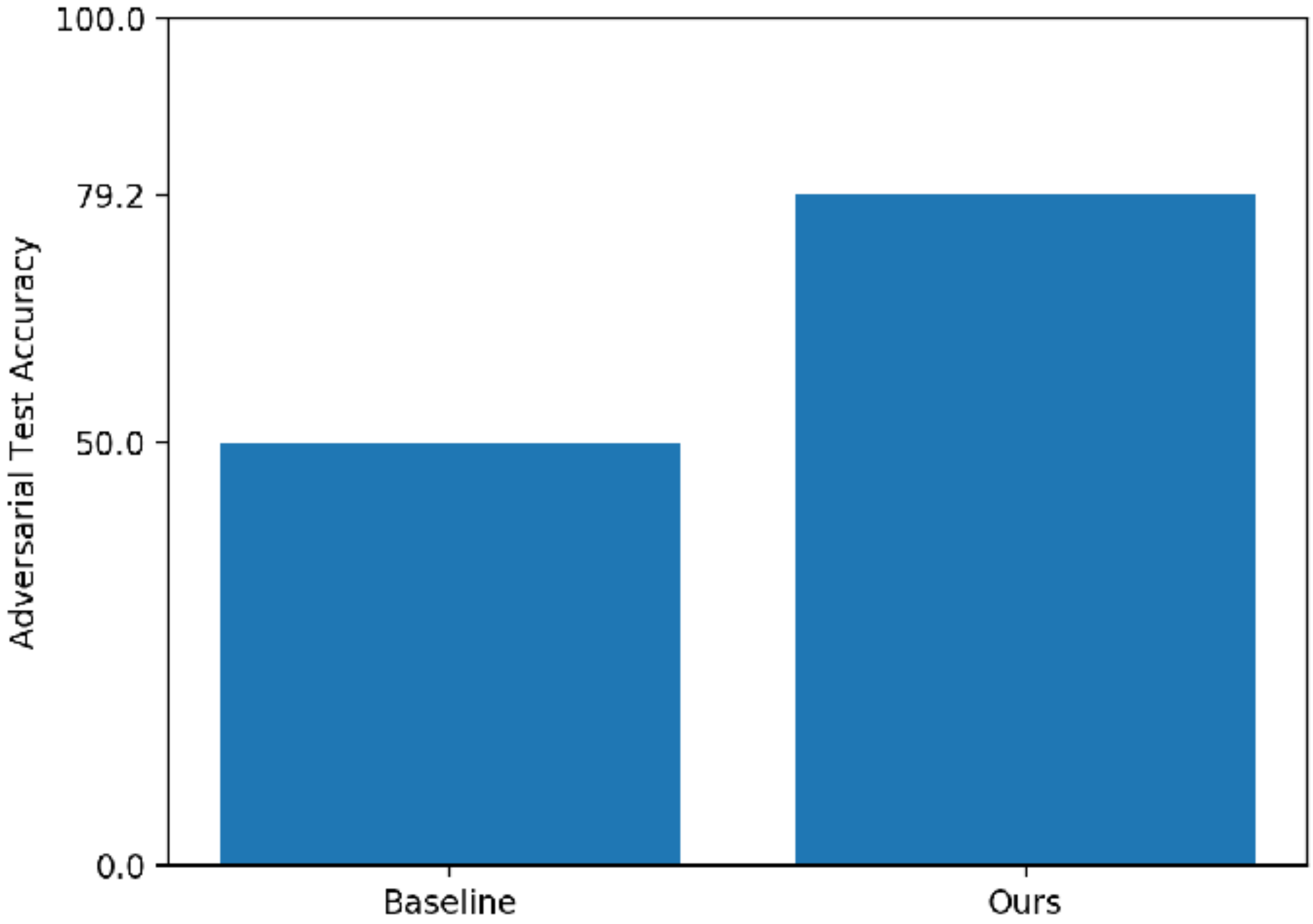
Large improvements on SVHN direct ("white box") attacks



5 years ago, this would have been SOTA on *clean* data

(Goodfellow 2017)

Large Improvements against CIFAR-10 direct ("white box") attacks



6 years ago, this would have been SOTA on *clean* data

(Goodfellow 2017)

Other results

- Improvement on CIFAR-100
    - (Still very broken)
- Improvement on MNIST
    - Please quit caring about MNIST

Caveats

- Slight drop in accuracy on clean examples
- Only small improvement on black-box transfer-based adversarial examples

Ensemble Adversarial Training

Florian Tramèr

Alexey Kurakin

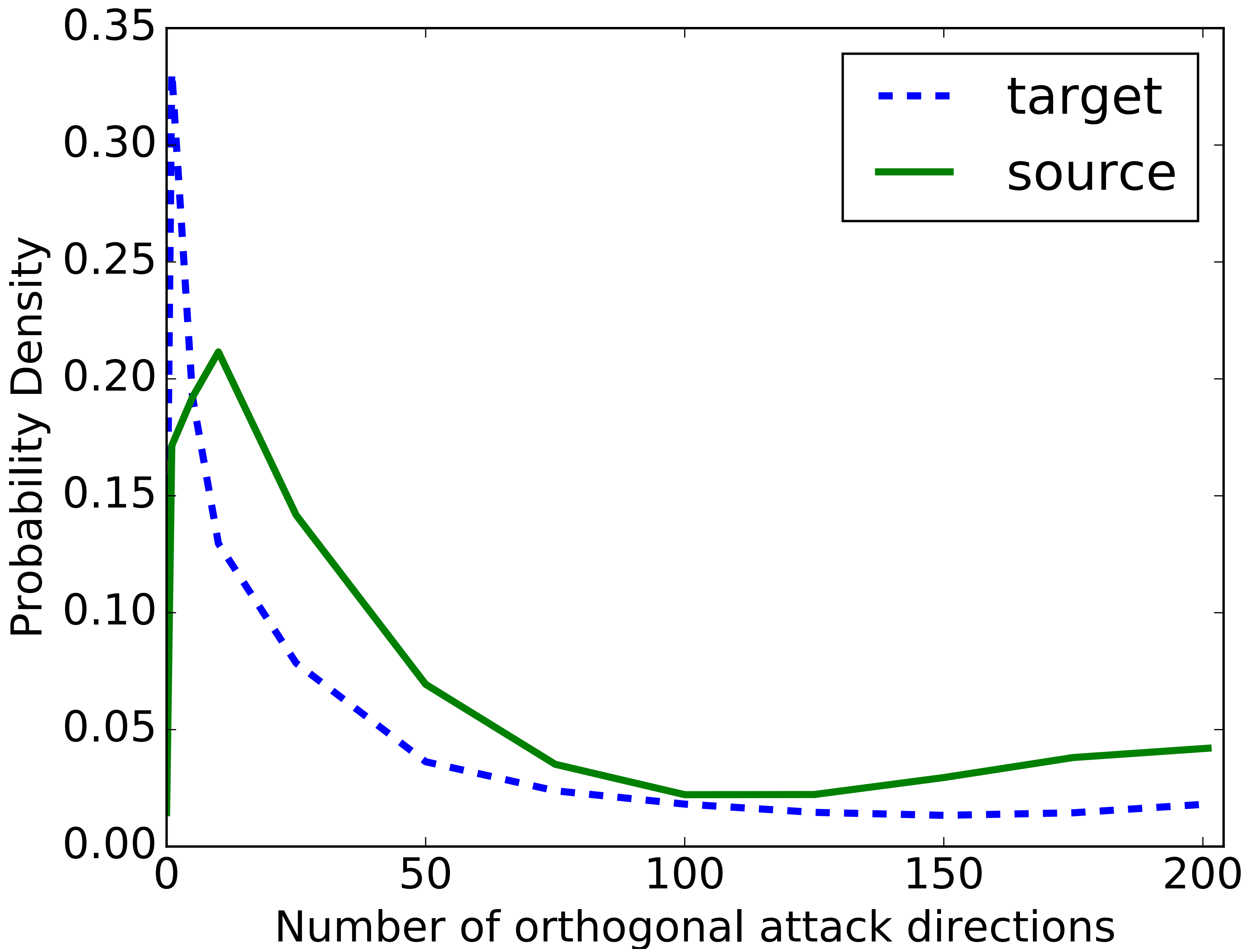
Nicolas Papernot

Ian Goodfellow

Dan Boneh

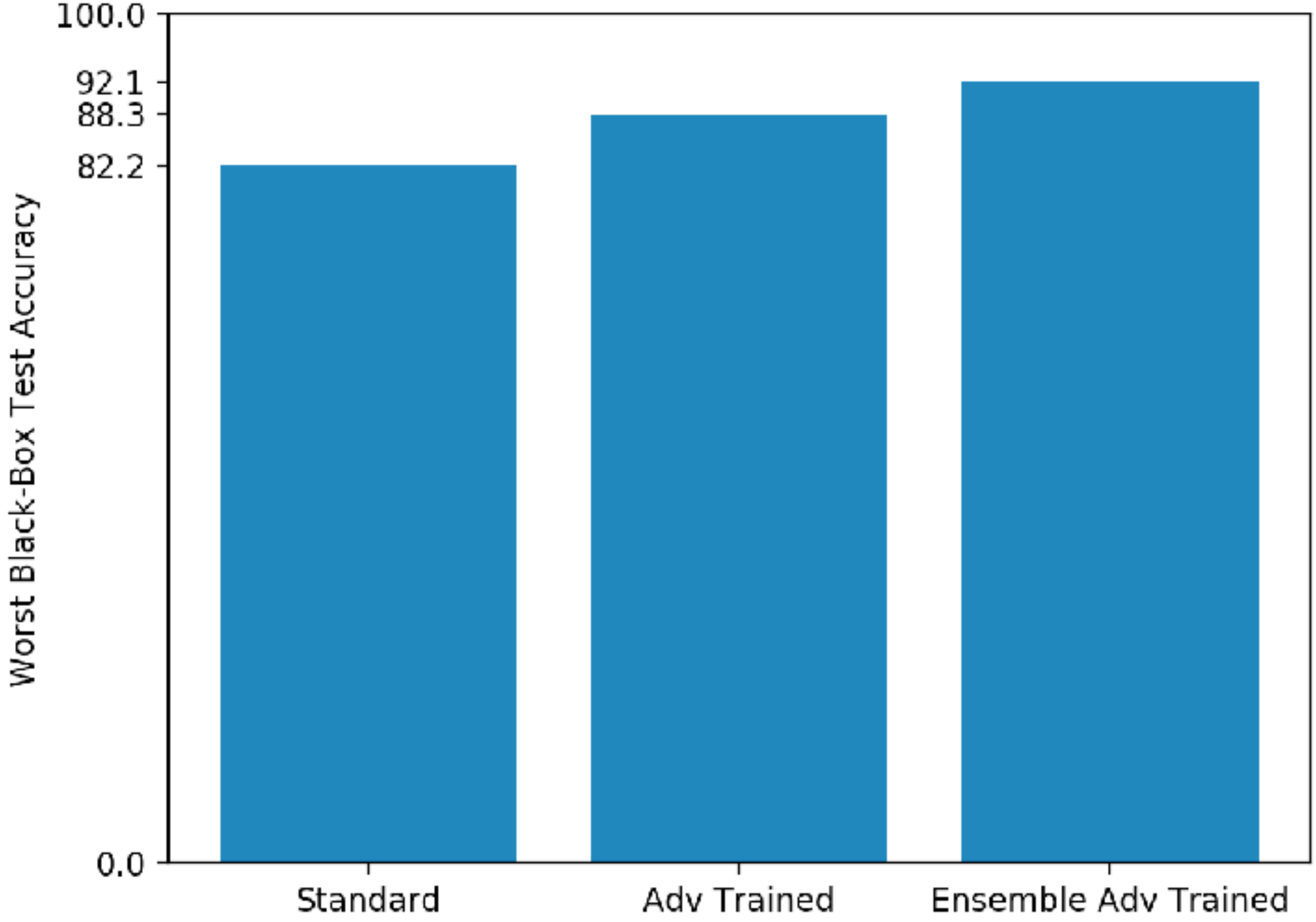
Patrick McDaniel

Estimating the Subspace Dimensionality



(Tramèr et al, 2017)

Transfer Attacks Against Inception ResNet v2 on ImageNet



(Goodfellow 2017)

Competition

AI Fight Club Could Help Save Us from a Future of Super-Smart Cyberattacks

MIT Technology Review

Best defense so far on ImageNet:
Ensemble adversarial training.

Used as at least part of all top 10 entries in dev round 3

(Goodfellow 2017)

Get involved!

https://github.com/tensorflow/cleverhans

cleverhans