Defense Against the Dark Arts: Machine Learning Security and Privacy

Ian Goodfellow, Staff Research Scientist, Google Brain

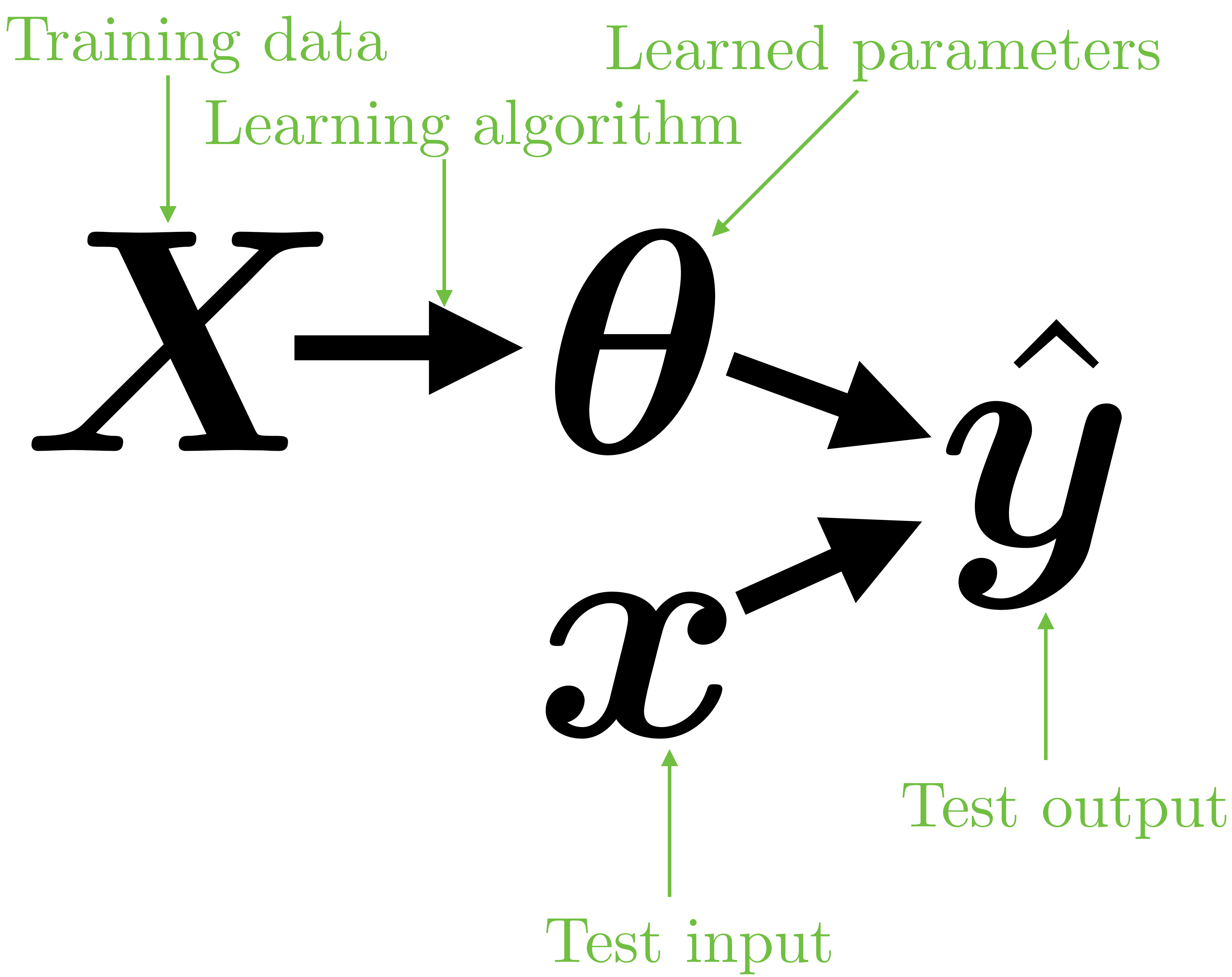
BayLearn 2017

An overview of a field

- This presentation summarizes the work of many people, not just my own / my collaborators
- Please check out the slides and view this link of extensive references
- The presentation focuses on the concepts, not the history or the inventors

(Goodfellow 2017)

Machine learning pipeline



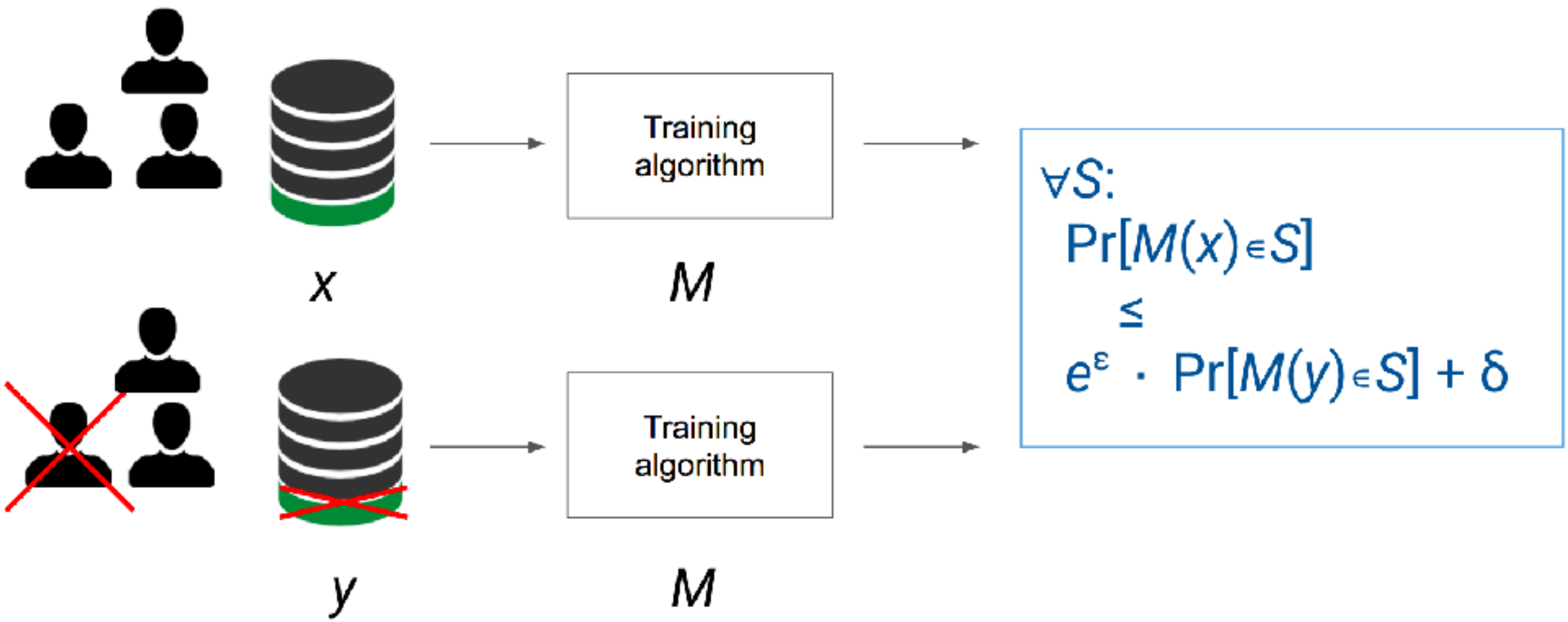
(Goodfellow 2017)

Privacy of training data

$$\boldsymbol{X} \rightarrow \;\vert\; \boldsymbol{\theta} \rightarrow \hat{\boldsymbol{X}}$$

(Goodfellow 2017)

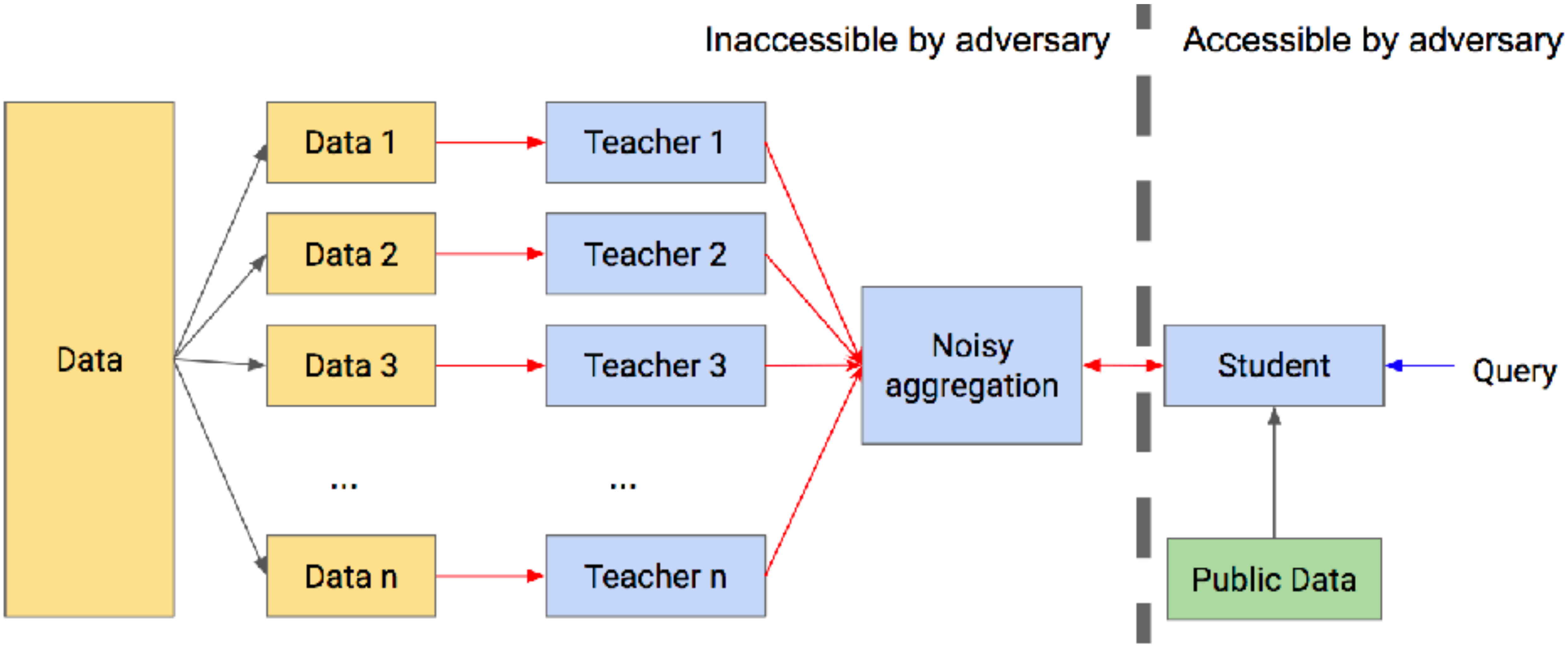
Defining (ε, δ)-Differential Privacy



$$\forall S: \Pr[M(x) \in S] \leq e^{\varepsilon} \cdot \Pr[M(y) \in S] + \delta$$

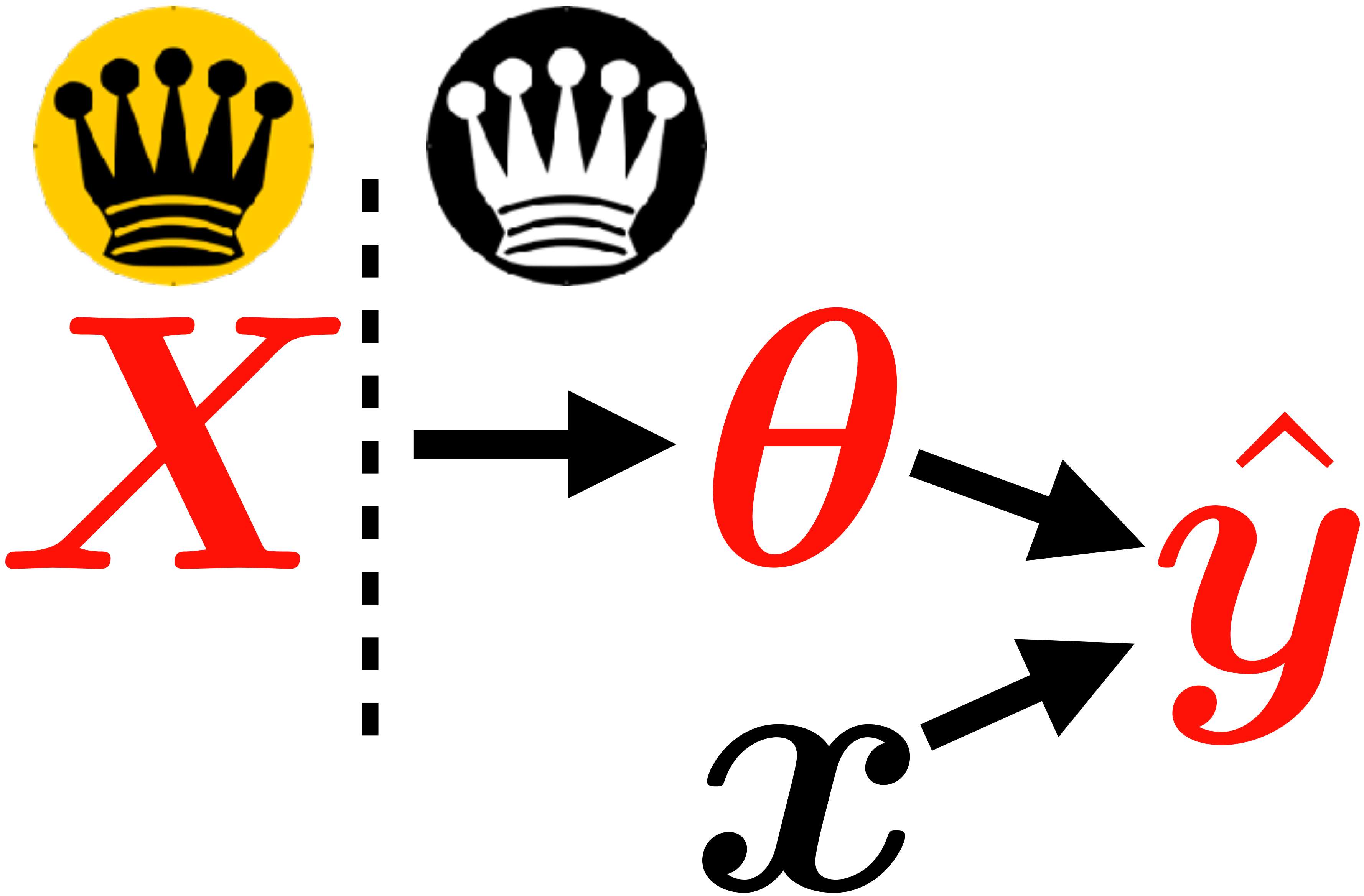
(Abadi 2017)

Private Aggregation of Teacher Ensembles

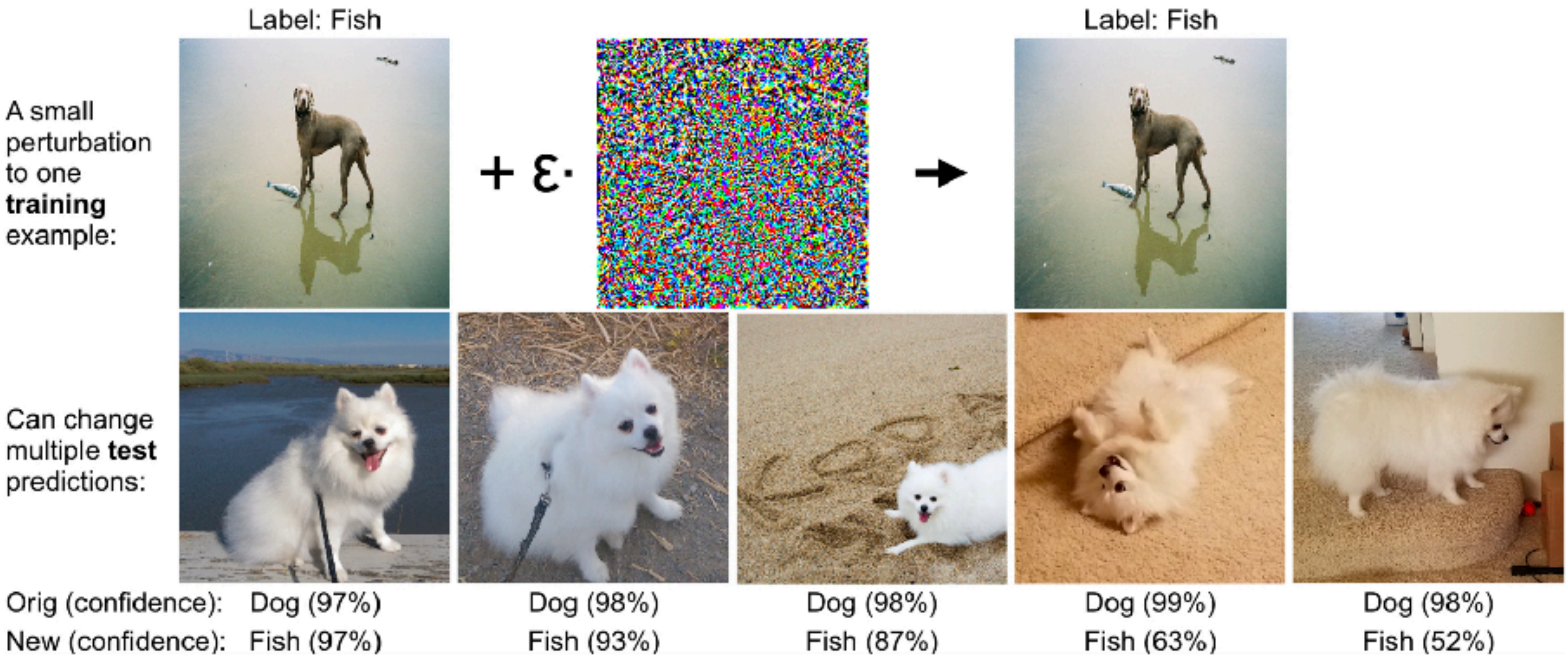


(Papernot et al 2016)

Training Set Poisoning

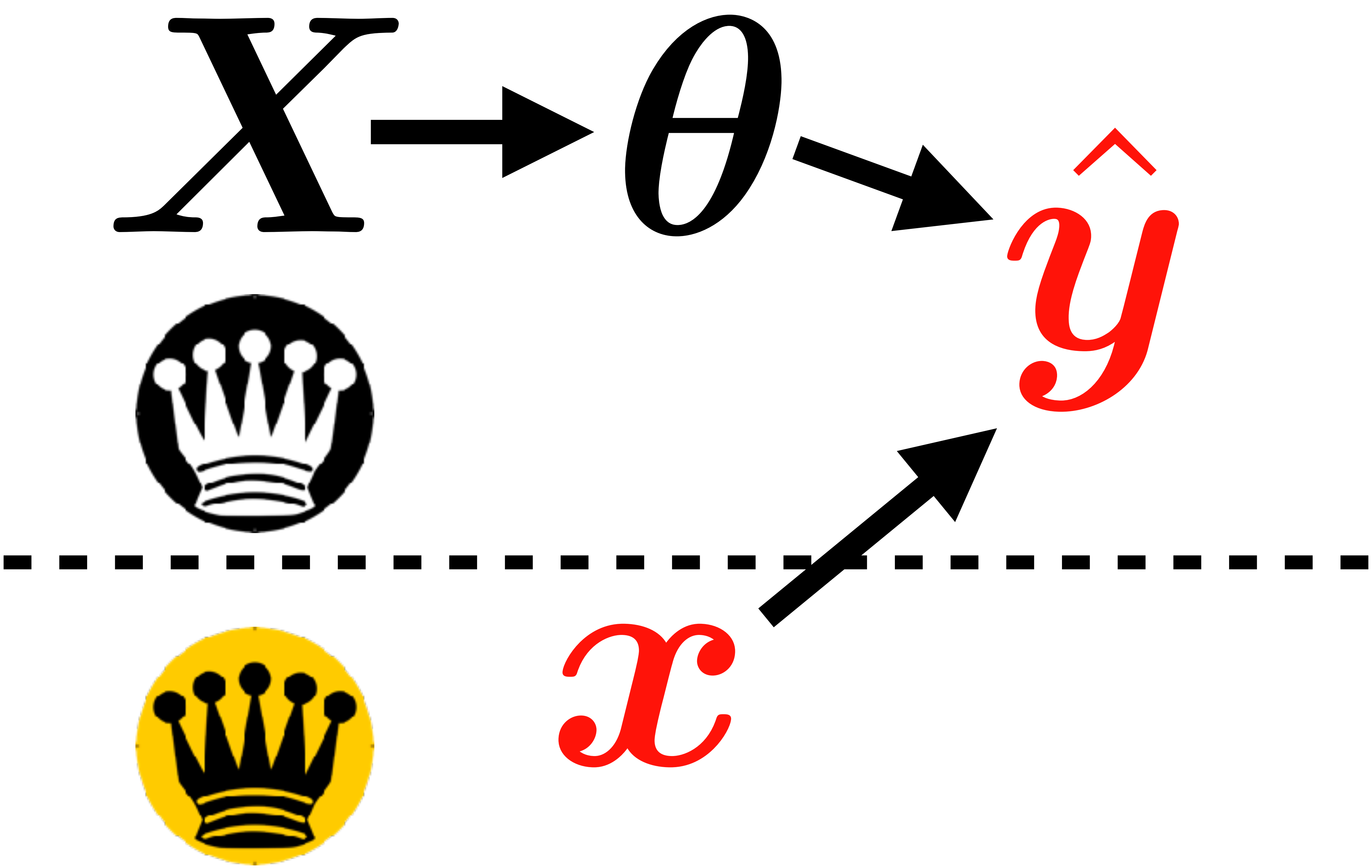


ImageNet poisoning



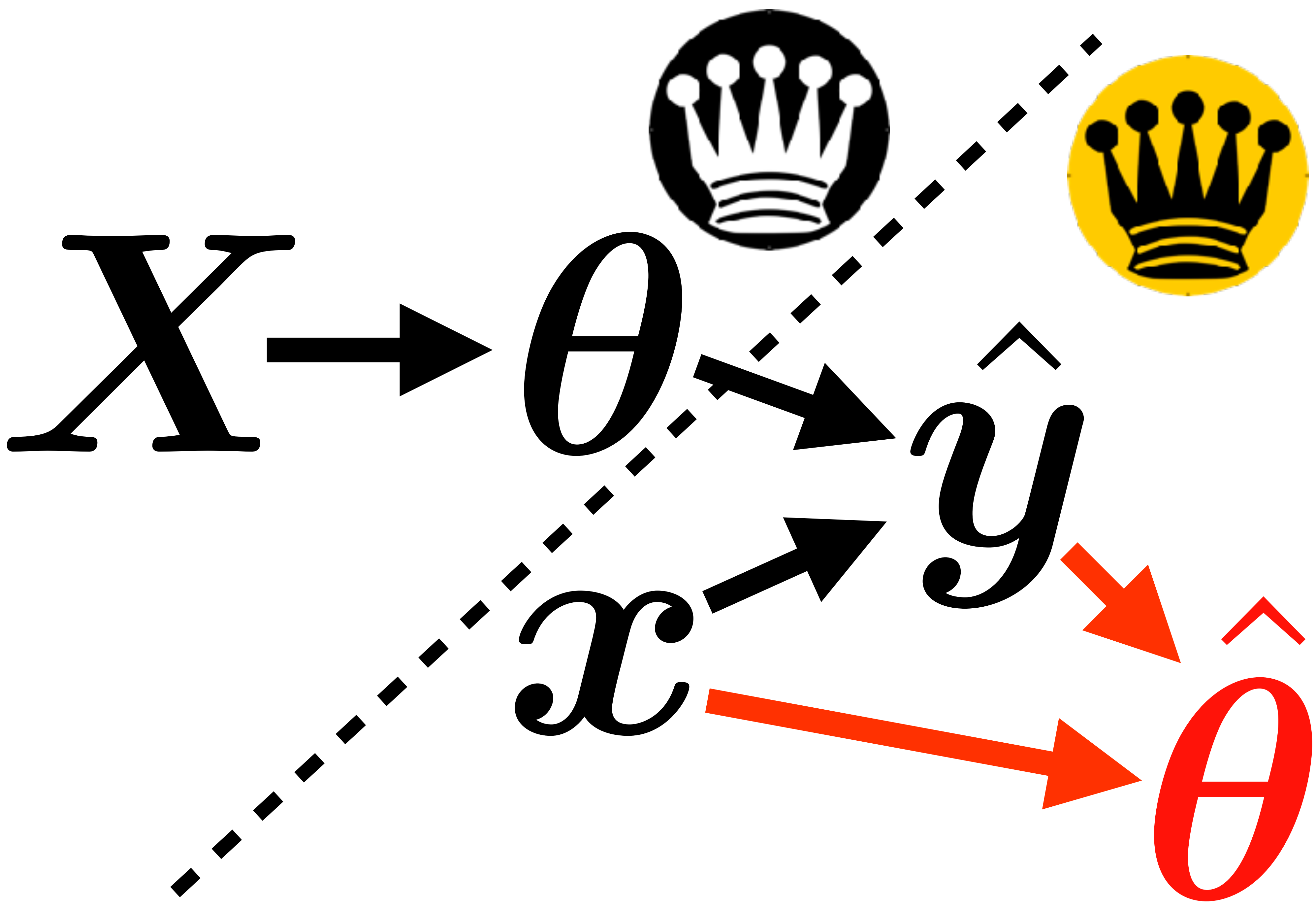
(Koh and Liang 2017)

Adversarial examples



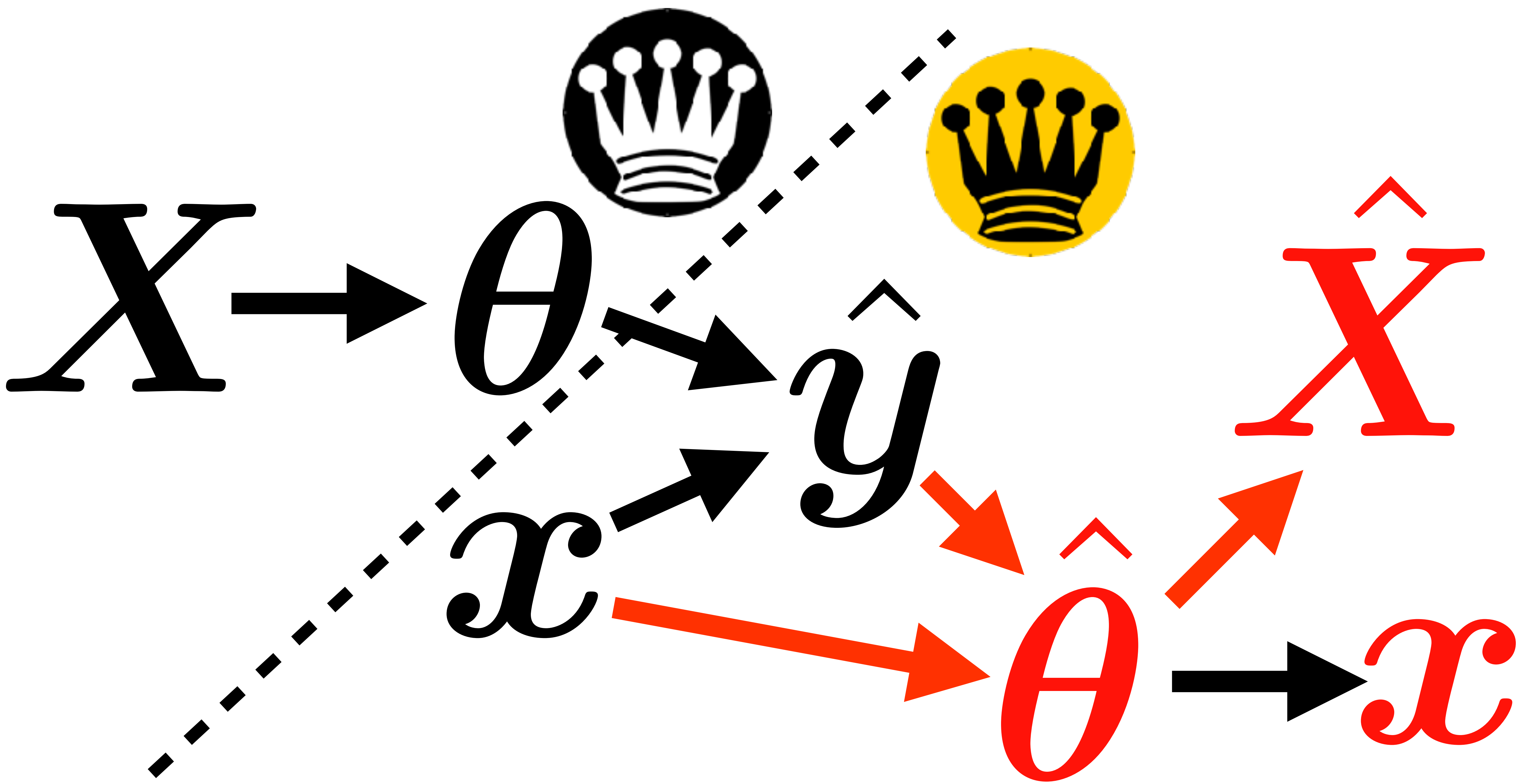
(Goodfellow 2017)

Model theft



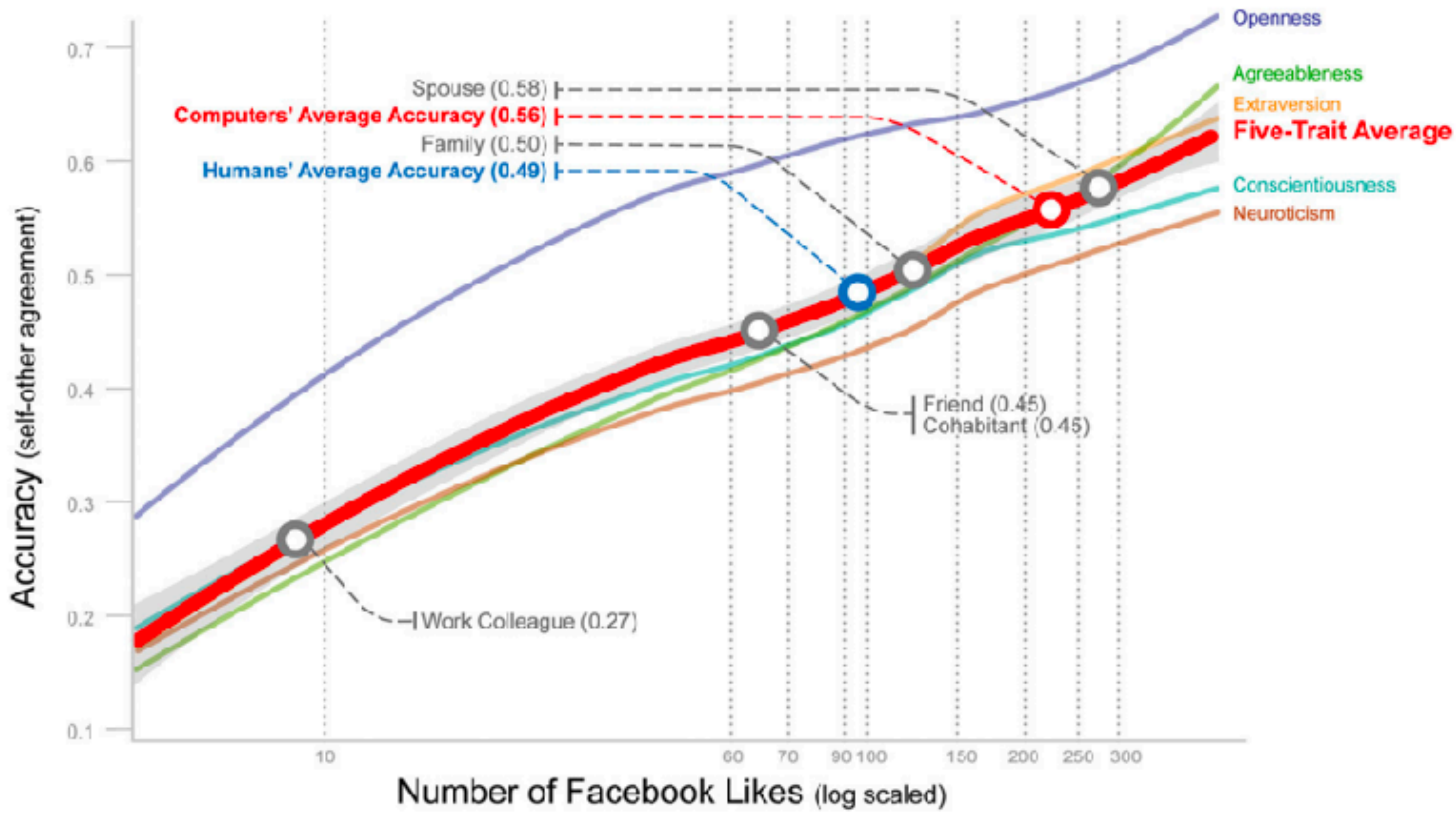
(Goodfellow 2017)

Model theft^{++}



(Goodfellow 2017)

Advanced models can infer private information



(Youyou et al 2014)

Automated Crowdturfing

Temperature	Generated Review Text
0.1	I love this place! I have been here a few times and have never been disappointed. The service is always great and the food is always great. The staff is always friendly and the food is always great. I will definitely be back and try some of their other food and service.
0.5	I love this place. I have been going here for years and it is a great place to hang out with friends and family. I love the food and service. I have never had a bad experience when I am there.
0.7	My family and I are huge fans of this place. The staff is super nice and the food is great. The chicken is very good and the garlic sauce is perfect. Ice cream topped with fruit is delicious too. Highly recommended!
1.0	I had the grilled veggie burger with fries!!!! Ohhhh and taste. Omgggg! Very flavorful! It was so delicious that I didn't spell it!!

(Yao et al 2017)

Fake News

www.futureoffakenews.com

(Goodfellow 2017)

Machine learning for password guessing

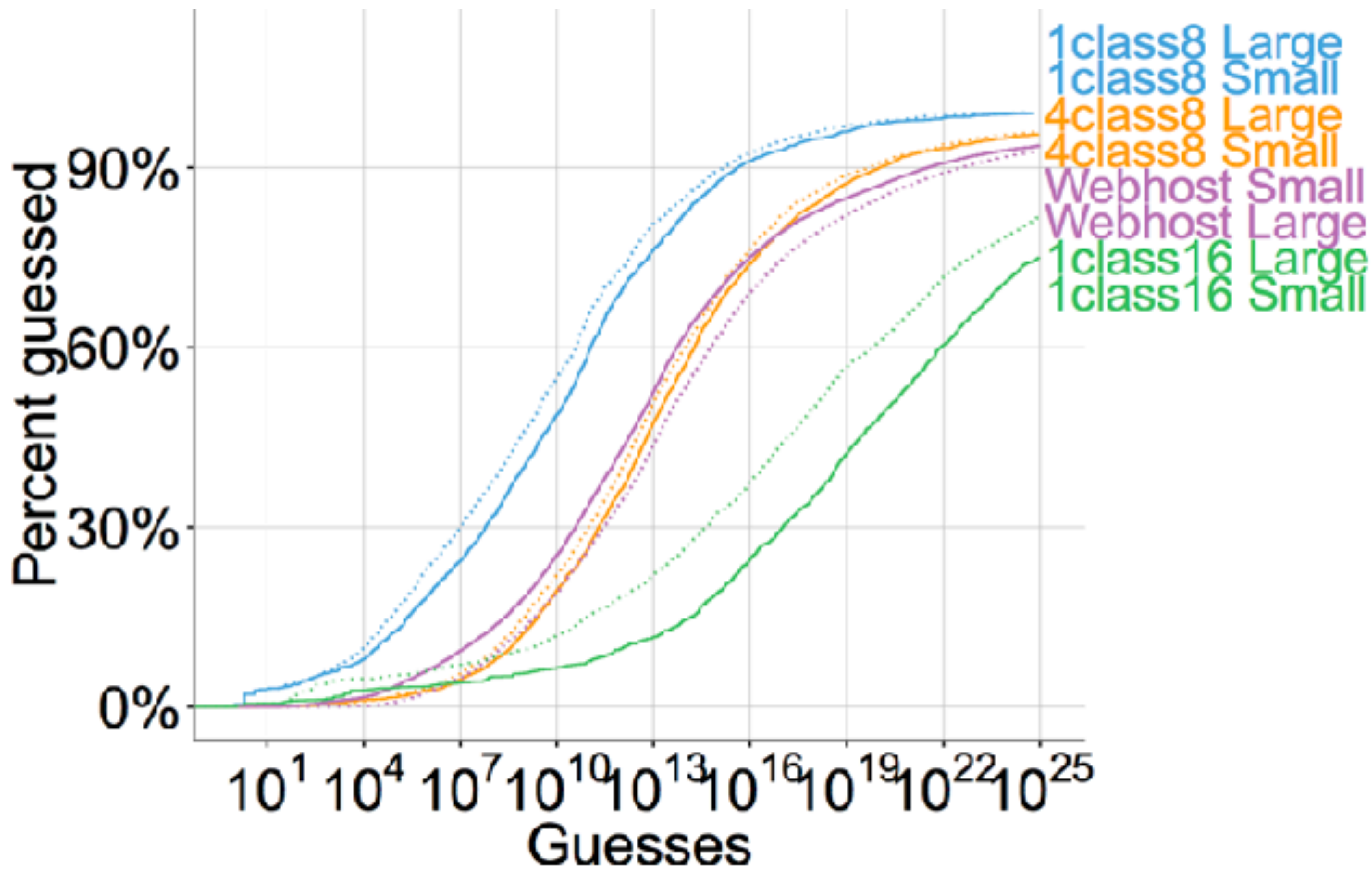


Figure 3: **Neural network size and password guessability.** Dotted lines are large networks; solid lines are small networks.

(Melicher et al 2016)

AI for geopolitics?

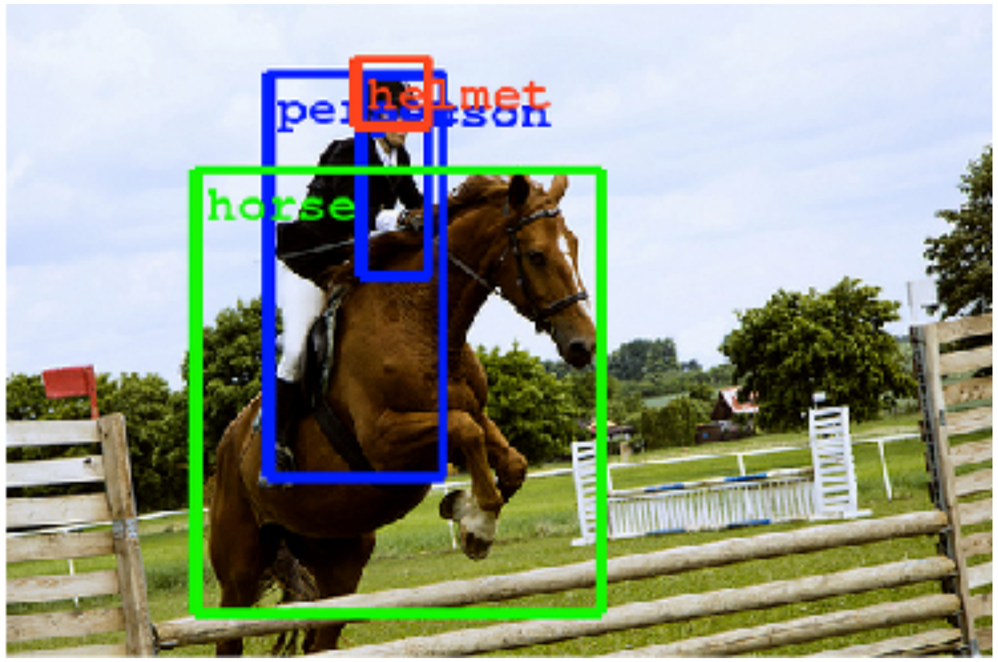
"Artificial intelligence is the future, not only for Russia, but for all humankind," said Putin, reports *RT*. "It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world."

THE VERGE

(Goodfellow 2017)

Deep Dive on Adversarial Examples

Since 2013, deep neural networks have matched human performance at...



(Szegedy et al, 2014)

...recognizing objects and faces....

(Taigmen et al, 2013)

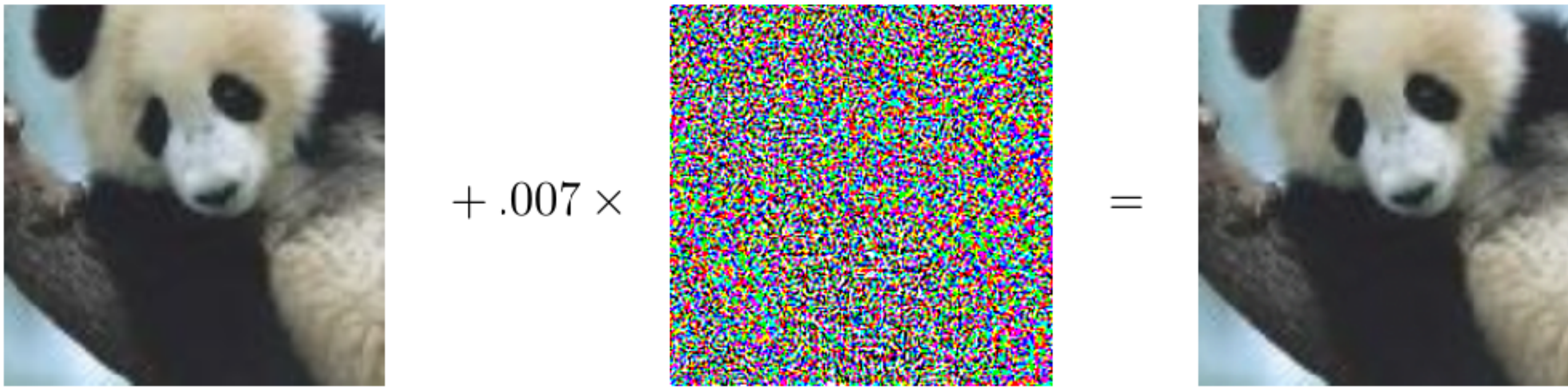
(Goodfellow et al, 2013)

...solving CAPTCHAS and reading addresses...

(Goodfellow et al, 2013)

and other tasks...

Adversarial Examples



Timeline:

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

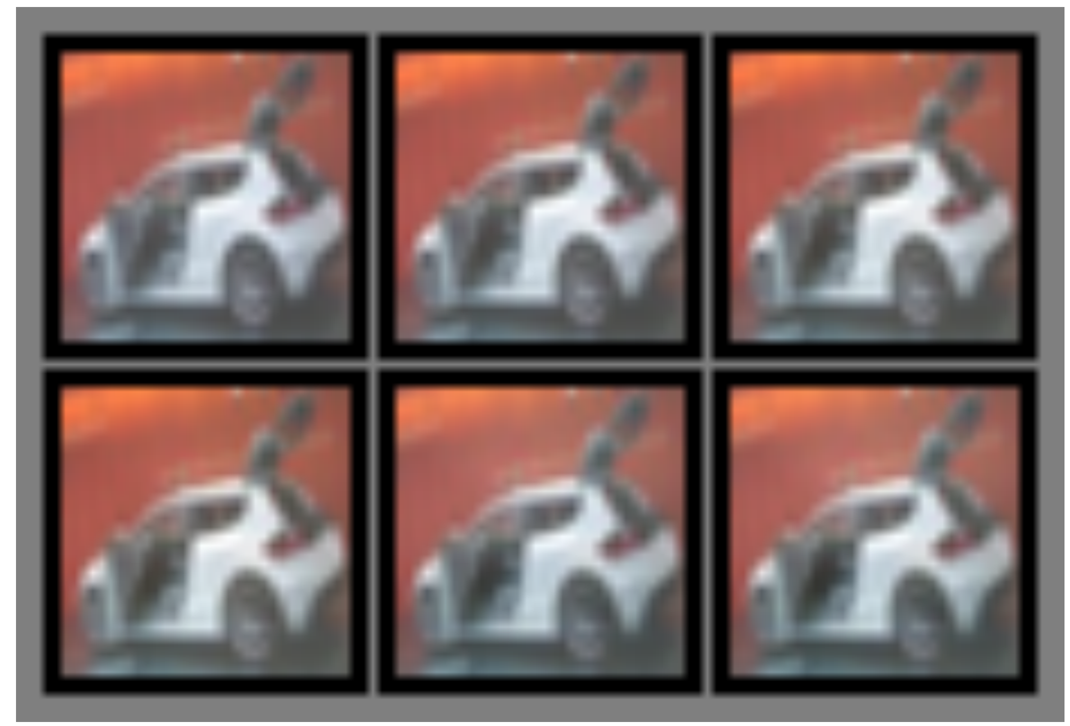
"Evasion Attacks Against Machine Learning at Test Time"

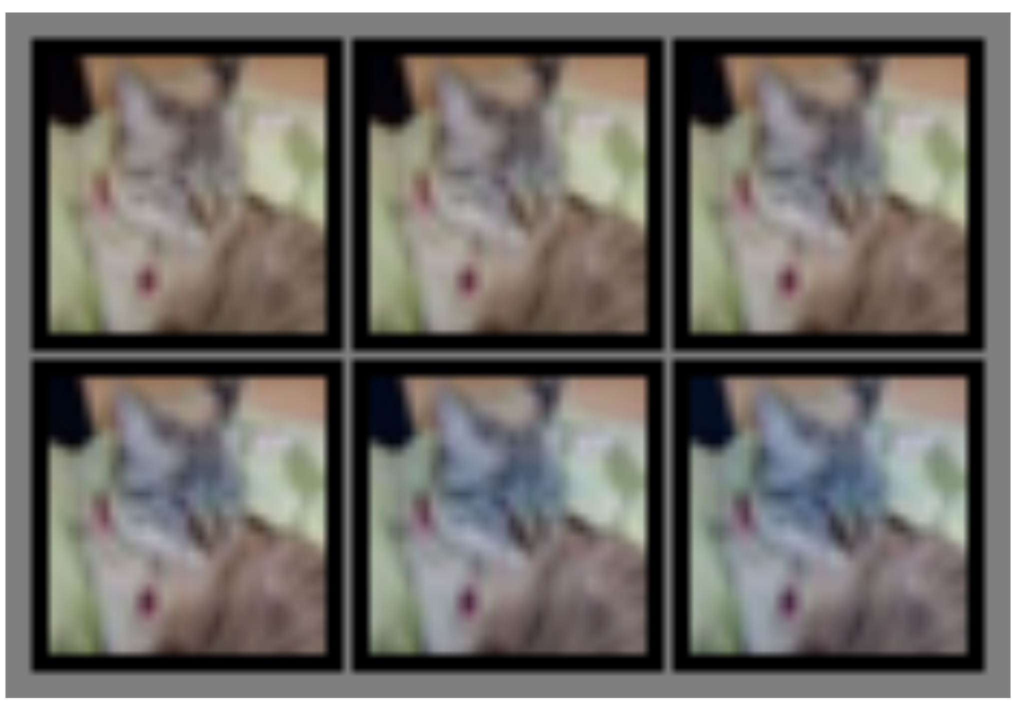
Biggio 2013: fool neural nets

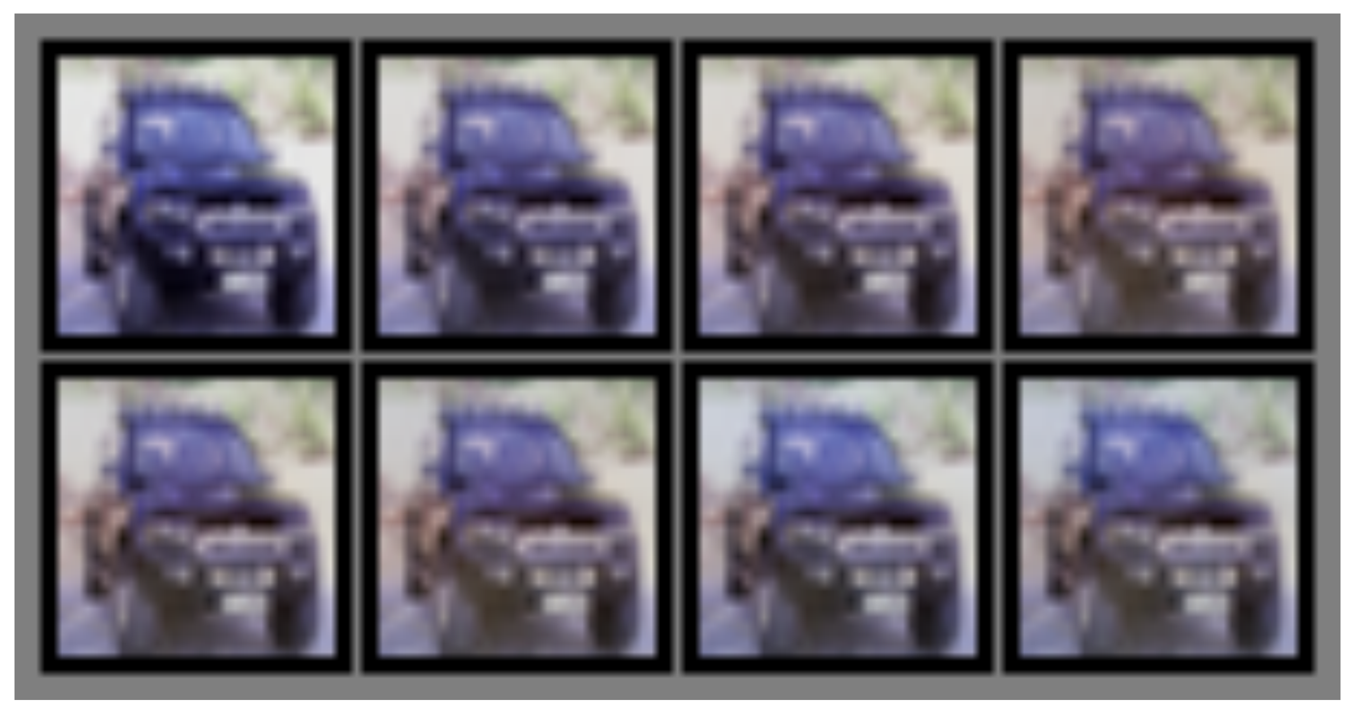
Szegedy et al 2013: fool ImageNet classifiers imperceptibly

Goodfellow et al 2014: cheap, closed form attack

Turning Objects into “Airplanes”

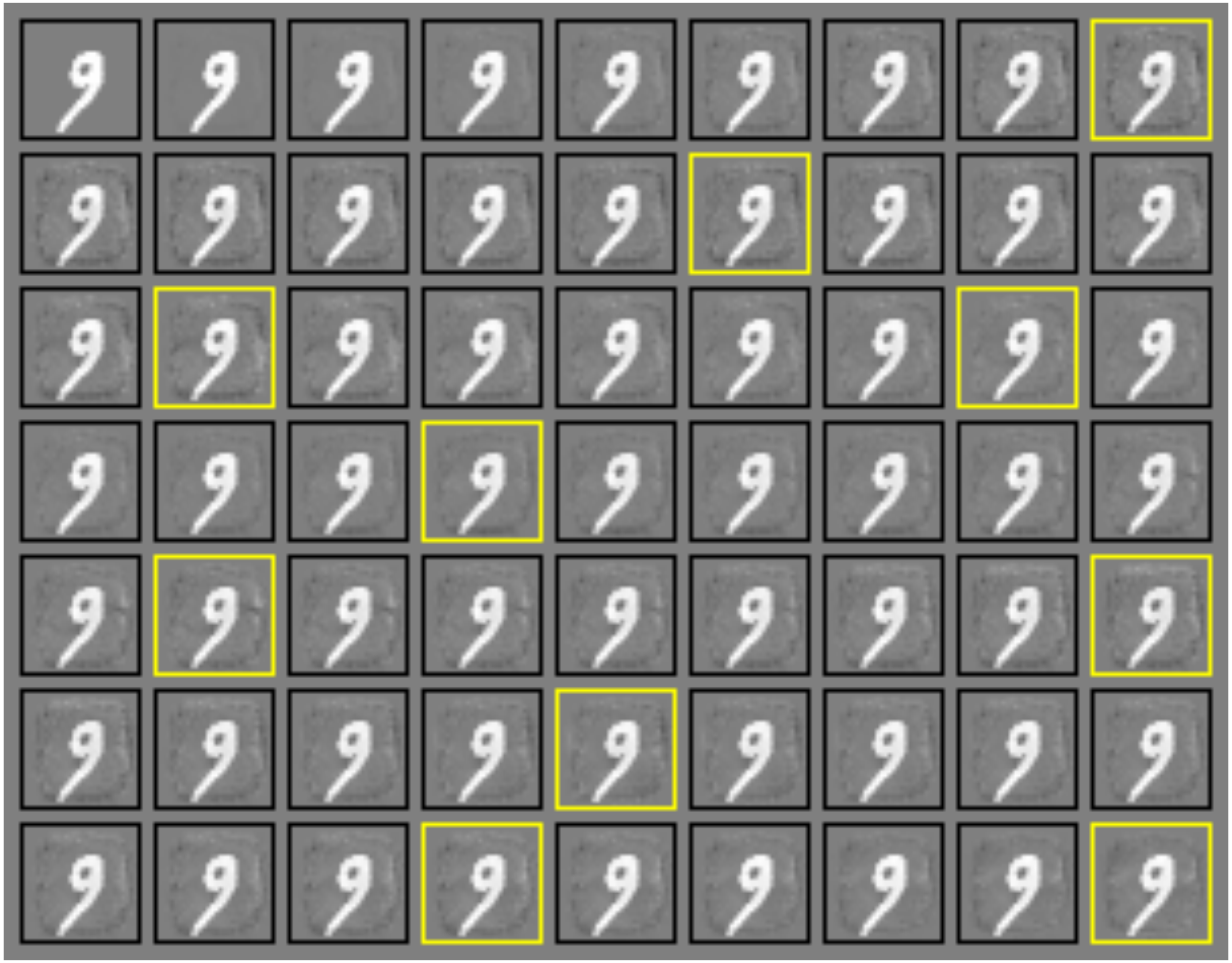






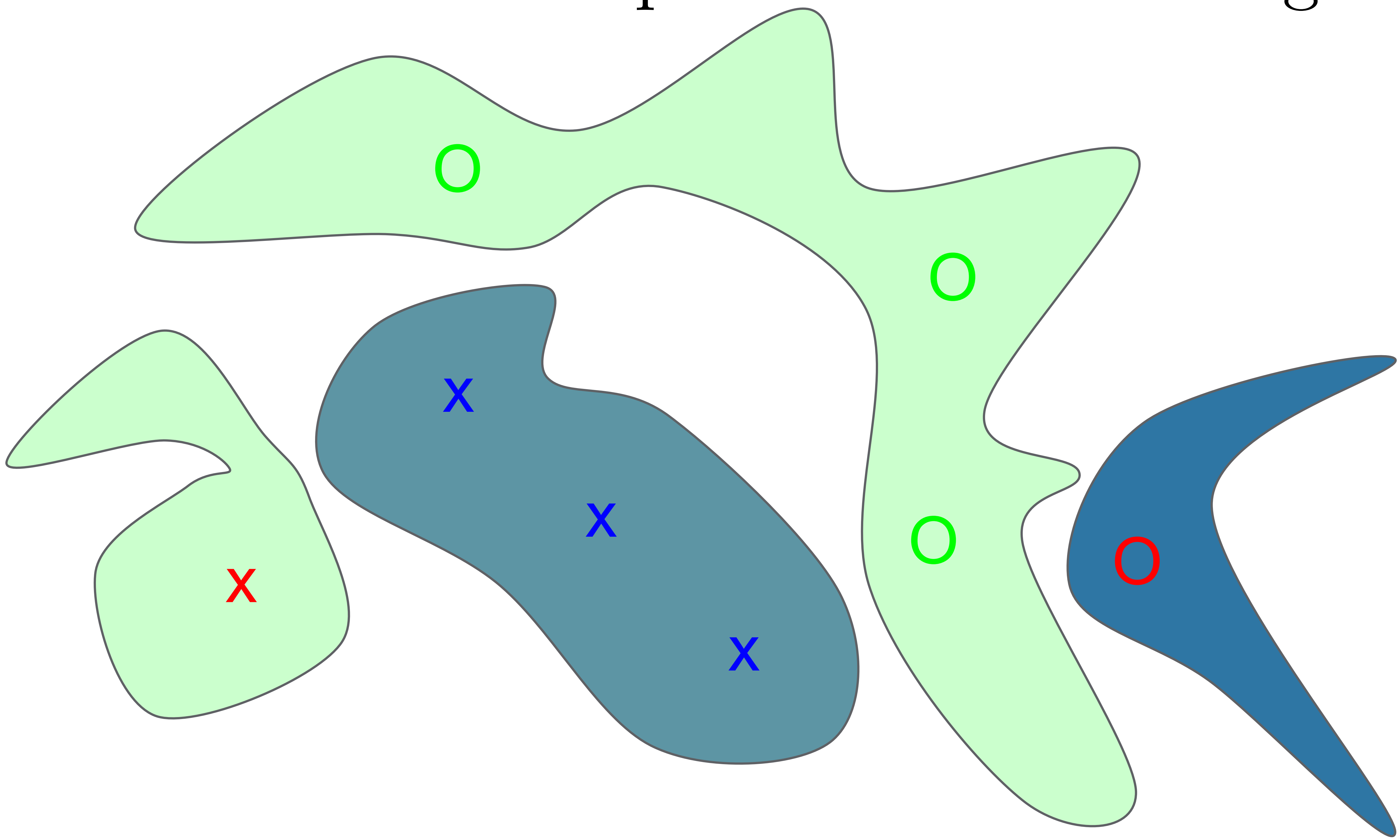
(Goodfellow 2017)

Attacking a Linear Model



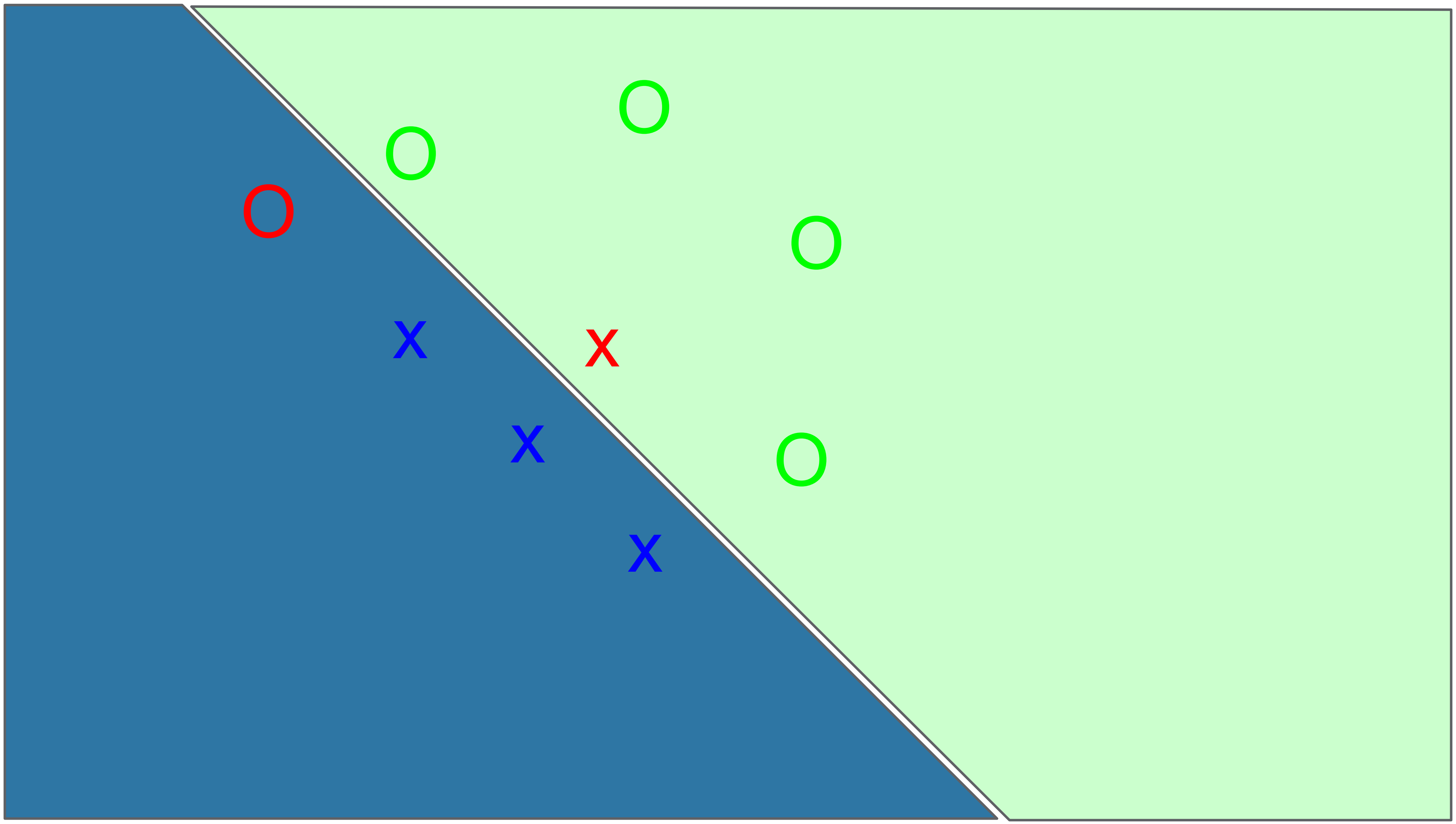
(Goodfellow 2017)

Adversarial Examples from Overfitting



(Goodfellow 2017)

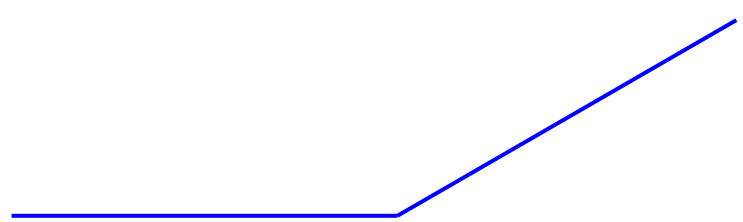
Adversarial Examples from Excessive Linearity



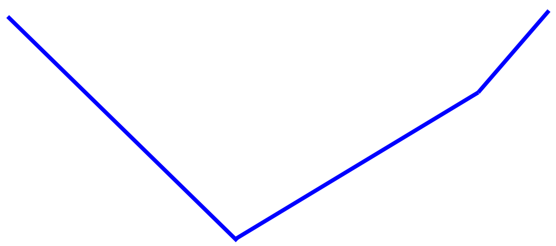
(Goodfellow 2017)

Modern deep nets are very piecewise linear

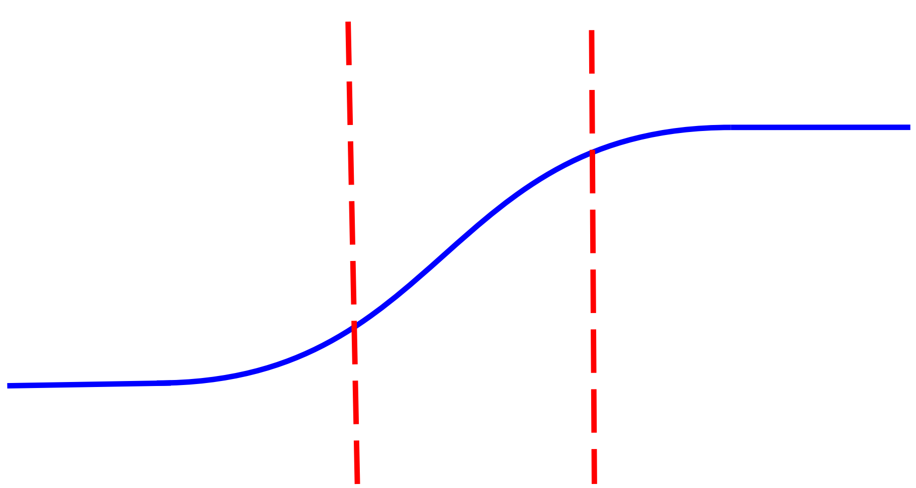
Rectified linear unit



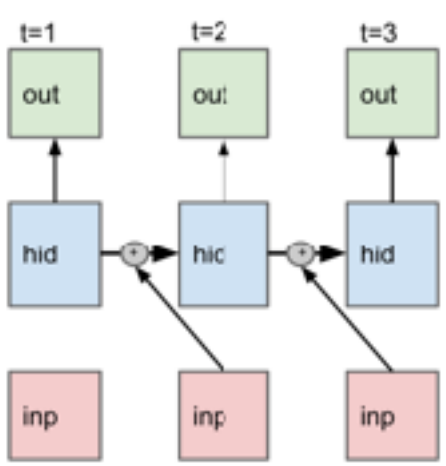
Maxout



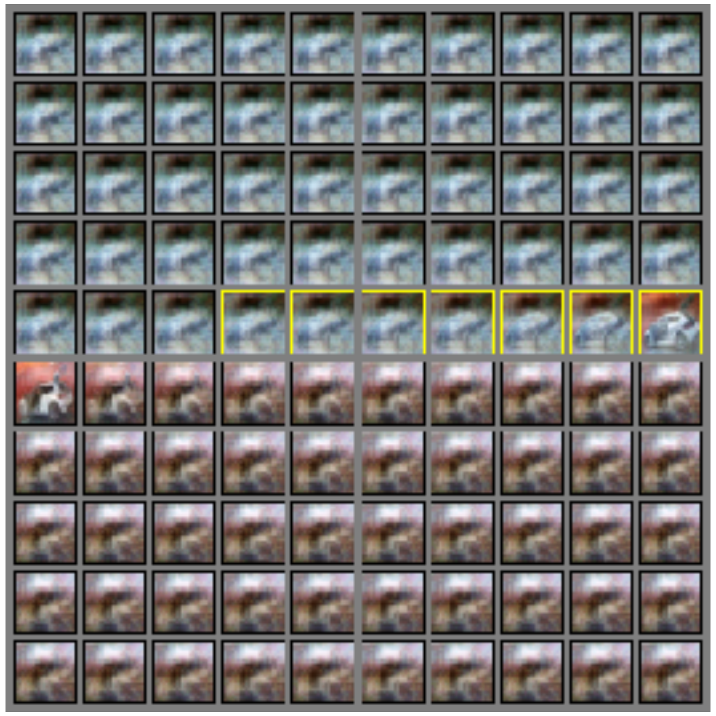
Carefully tuned sigmoid

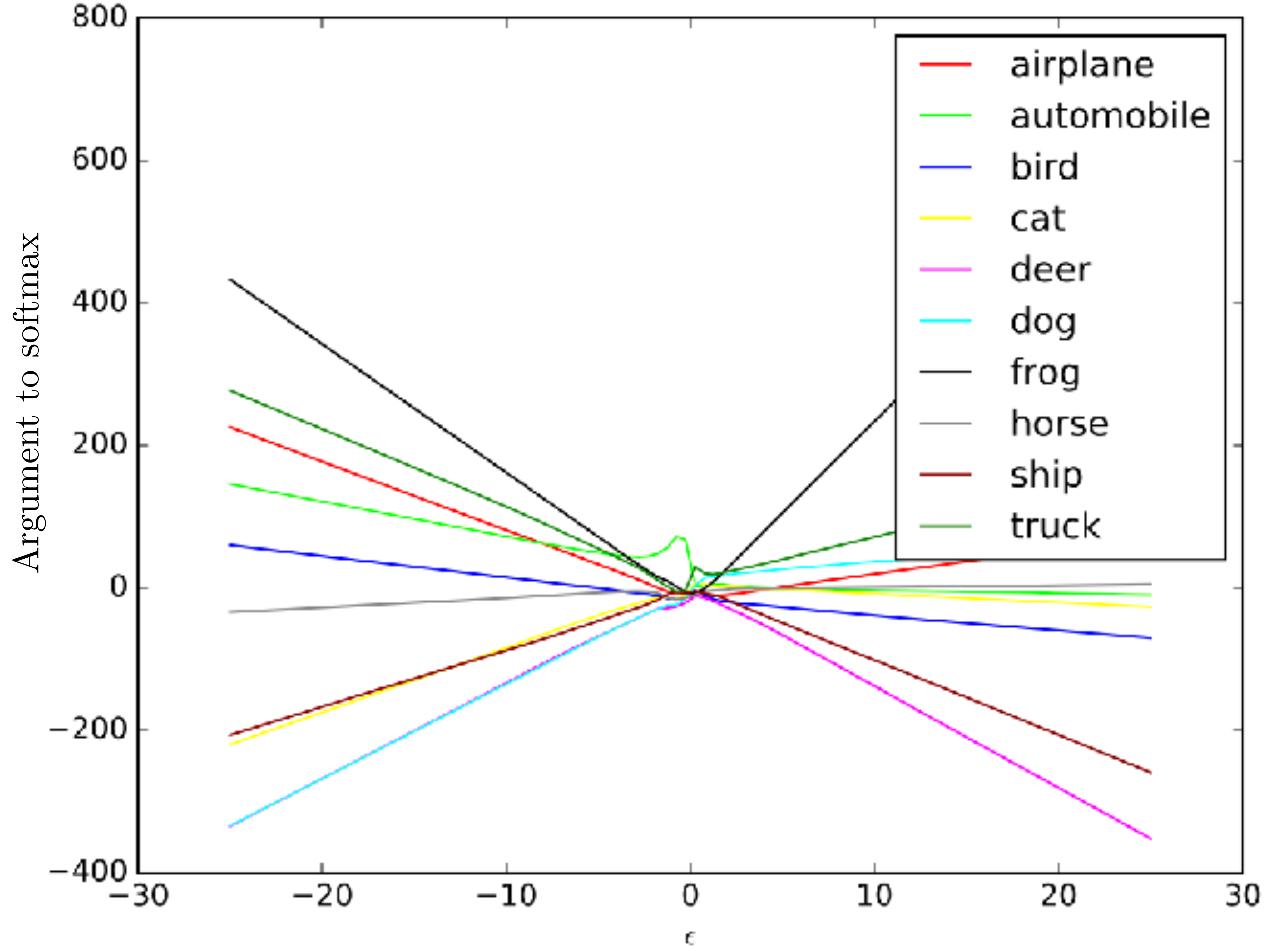


LSTM



Nearly Linear Responses in Practice

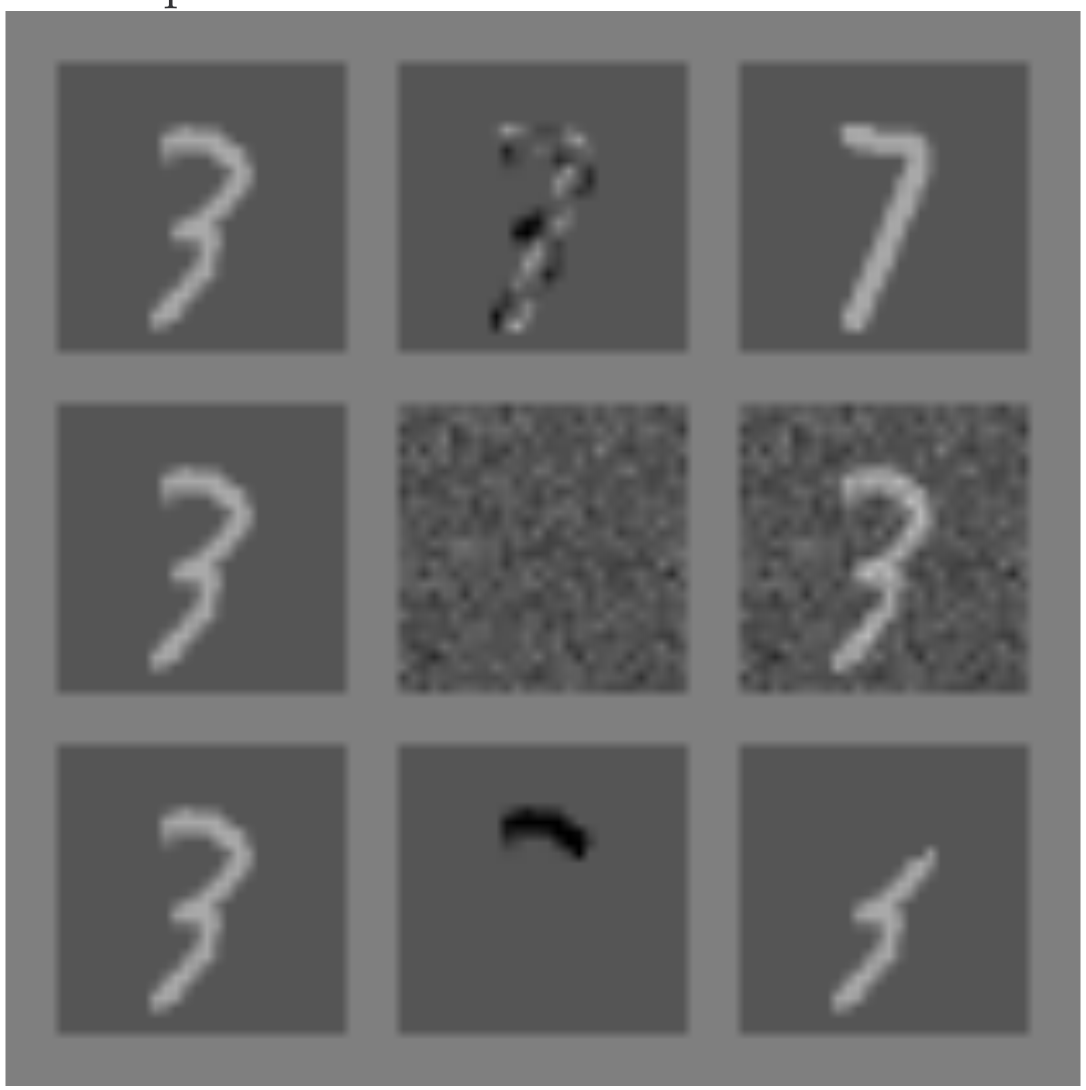




(Goodfellow 2017)

Small inter-class distances

Clean example | Perturbation | Corrupted example



Perturbation changes the true class

Random perturbation does not change the class

Perturbation changes the input to “rubbish class”

All three perturbations have L2 norm 3.96

This is actually small. We typically use 7!

(Goodfellow 2017)

The Fast Gradient Sign Method

$$J(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) \approx J(\boldsymbol{x}, \boldsymbol{\theta}) + (\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} J(\boldsymbol{x}).$$

Maximize

$$J(\boldsymbol{x}, \boldsymbol{\theta}) + (\tilde{\boldsymbol{x}} - \boldsymbol{x})^\top \nabla_{\boldsymbol{x}} J(\boldsymbol{x})$$

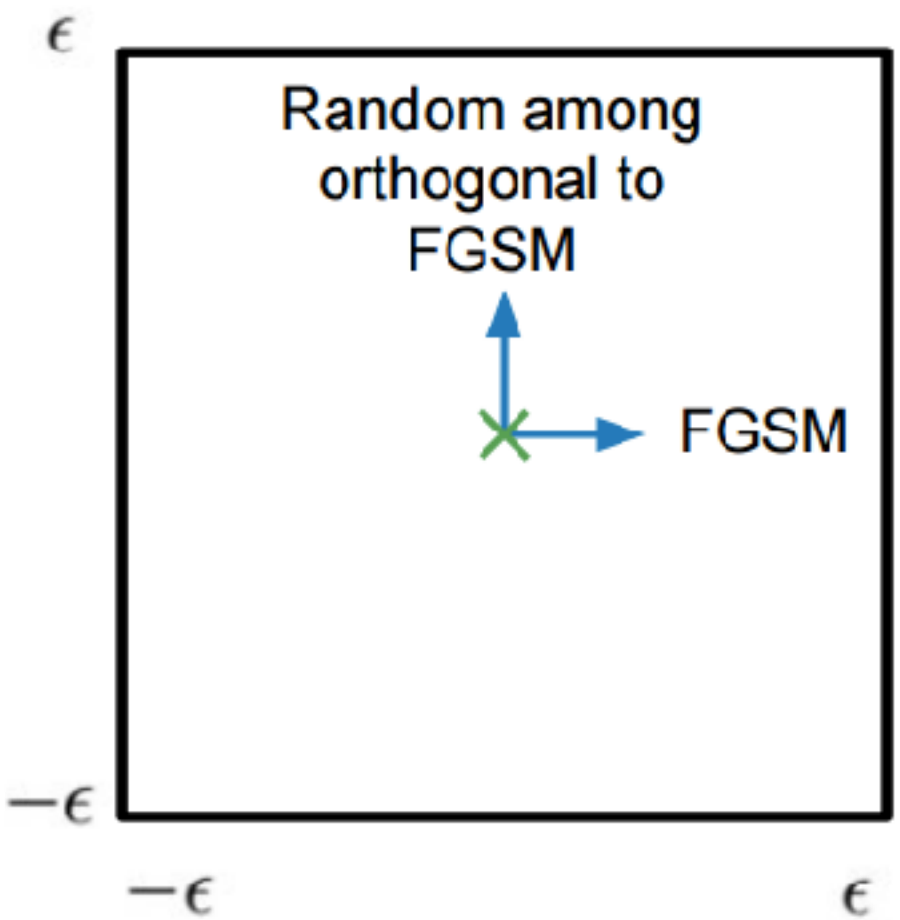
subject to

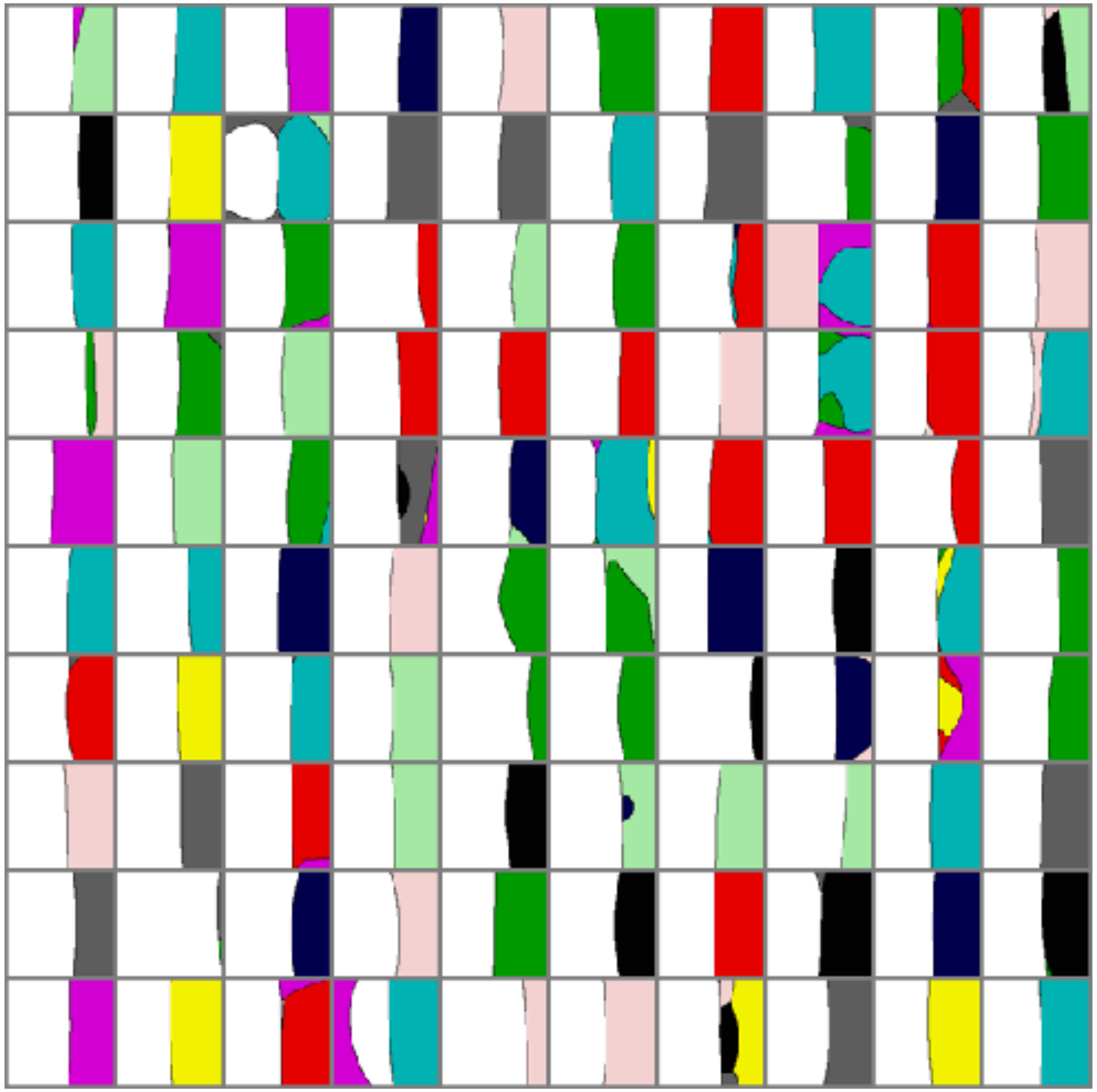
$$||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon$$

$$\Rightarrow \tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{x})\right).$$

(Goodfellow 2017)

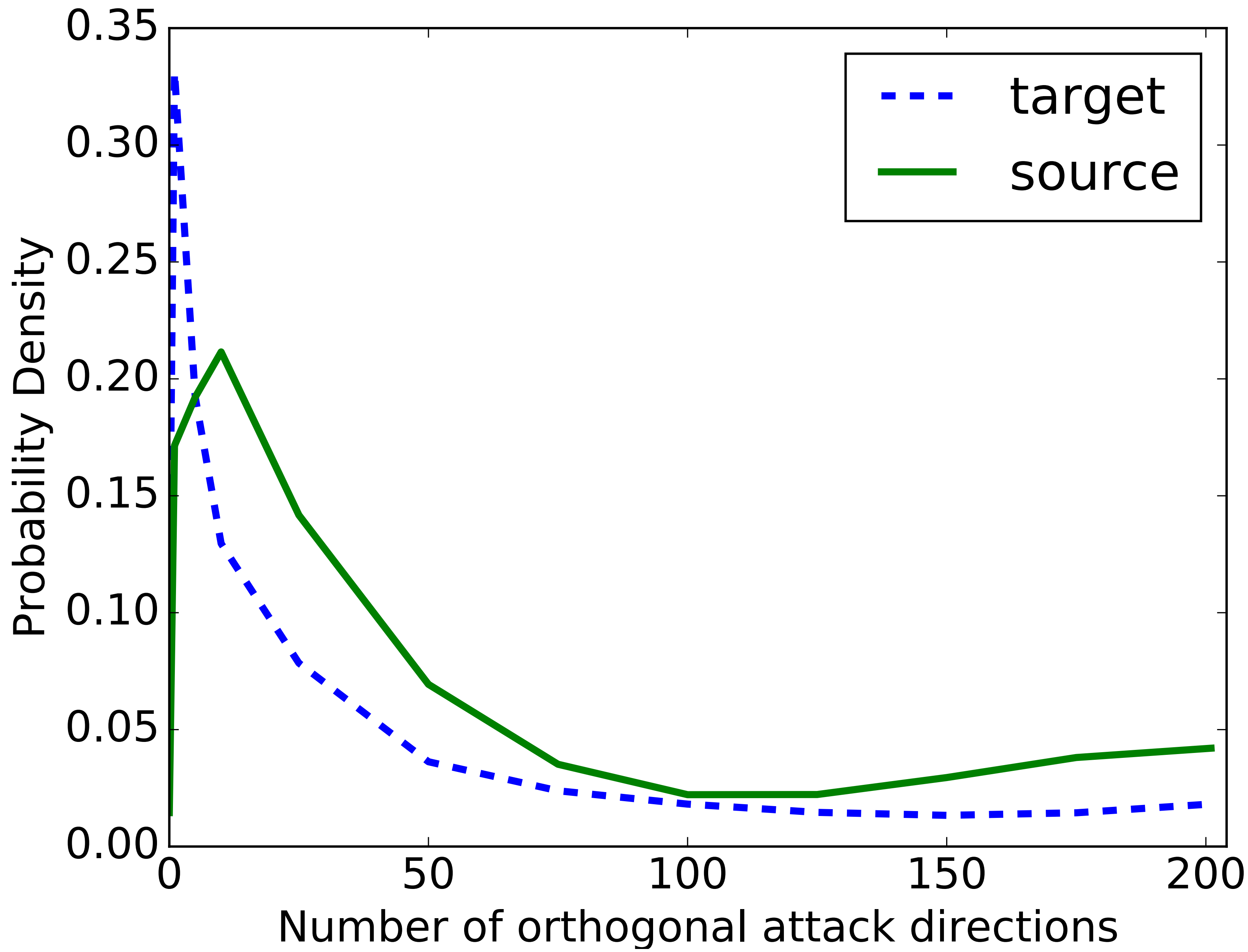
Maps of Adversarial and Random Cross-Sections





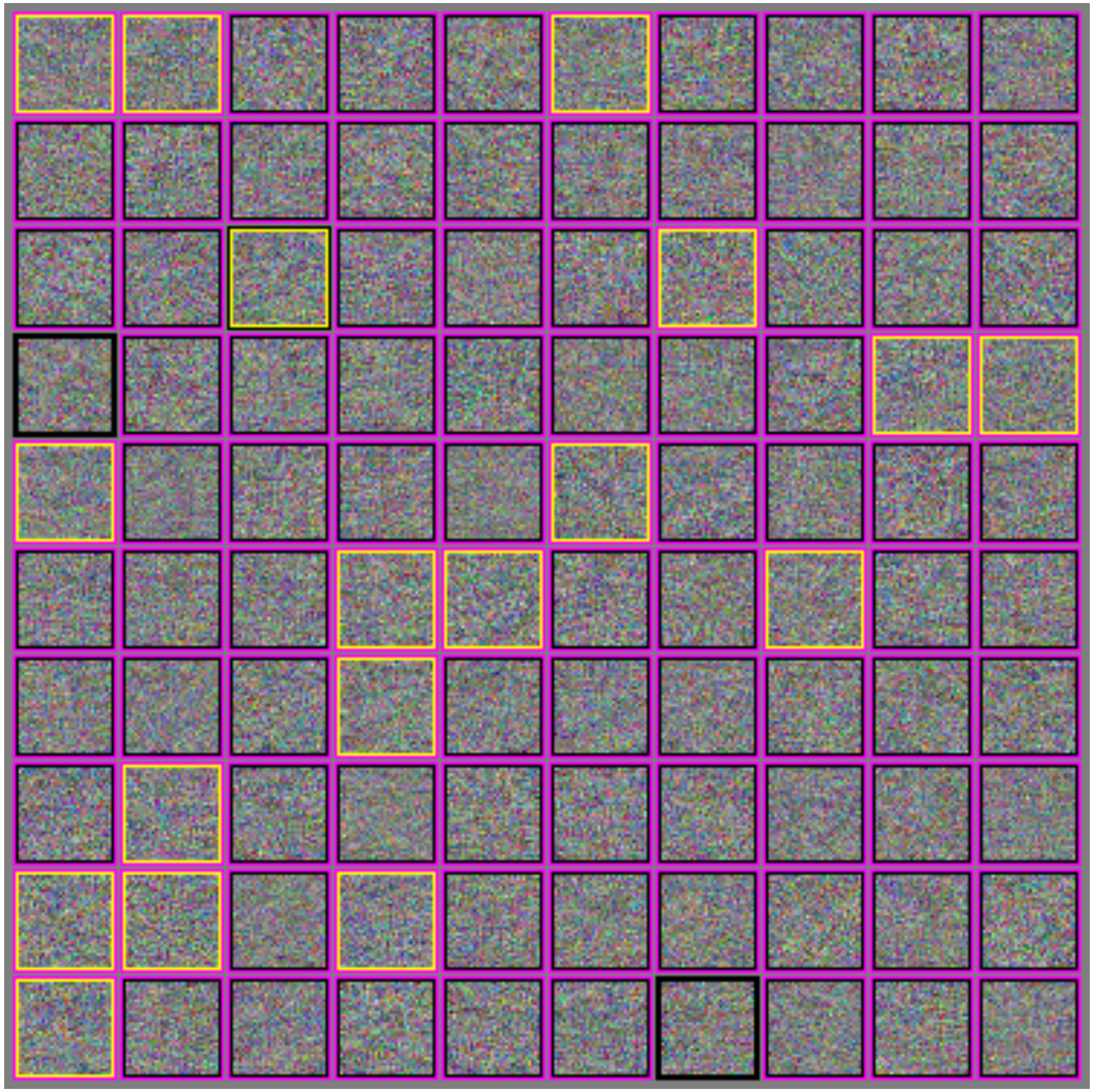
(collaboration with David Warde-Farley and Nicolas Papernot)

Estimating the Subspace Dimensionality



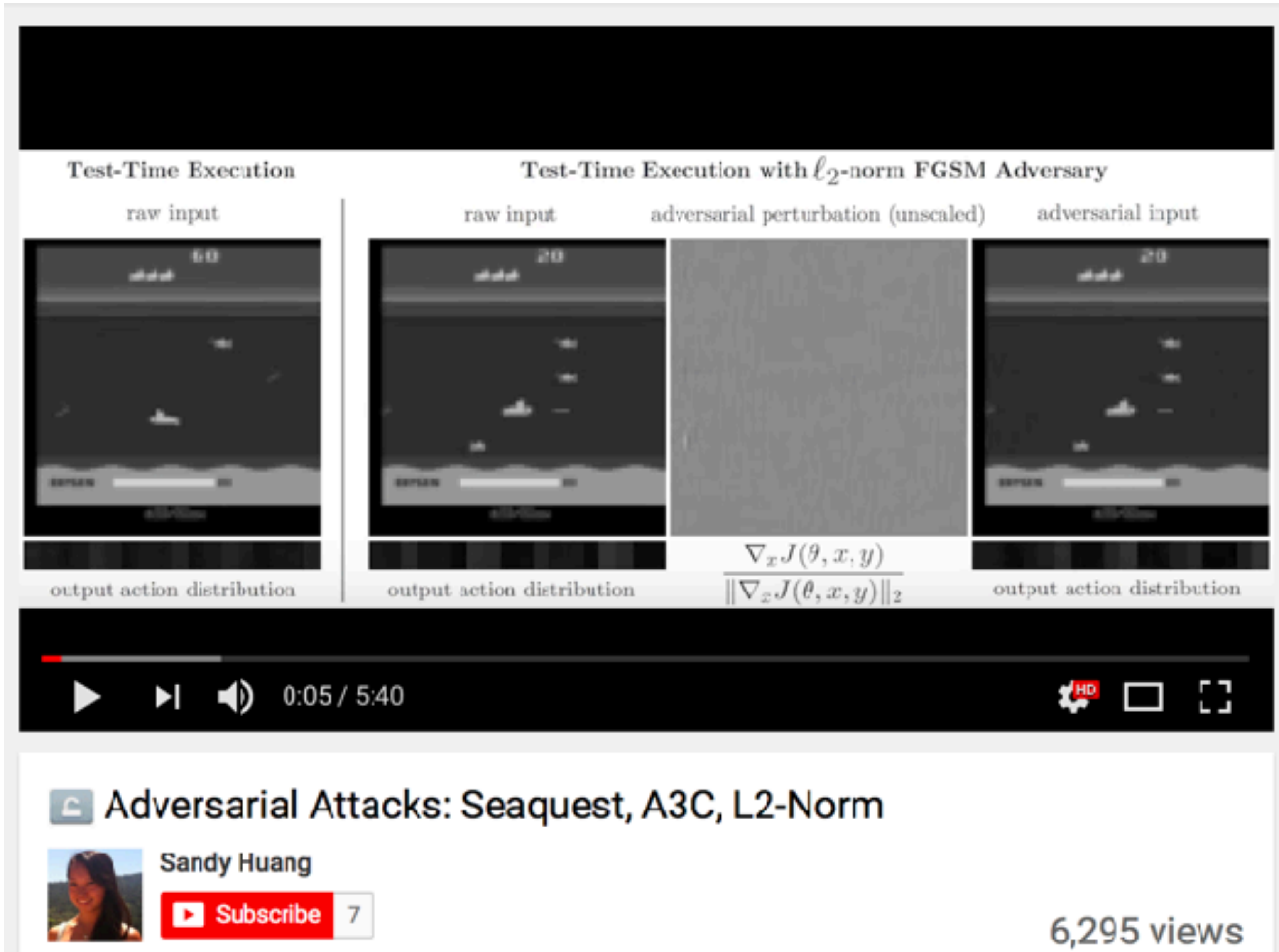
(Tramèr et al, 2017)

Wrong almost everywhere



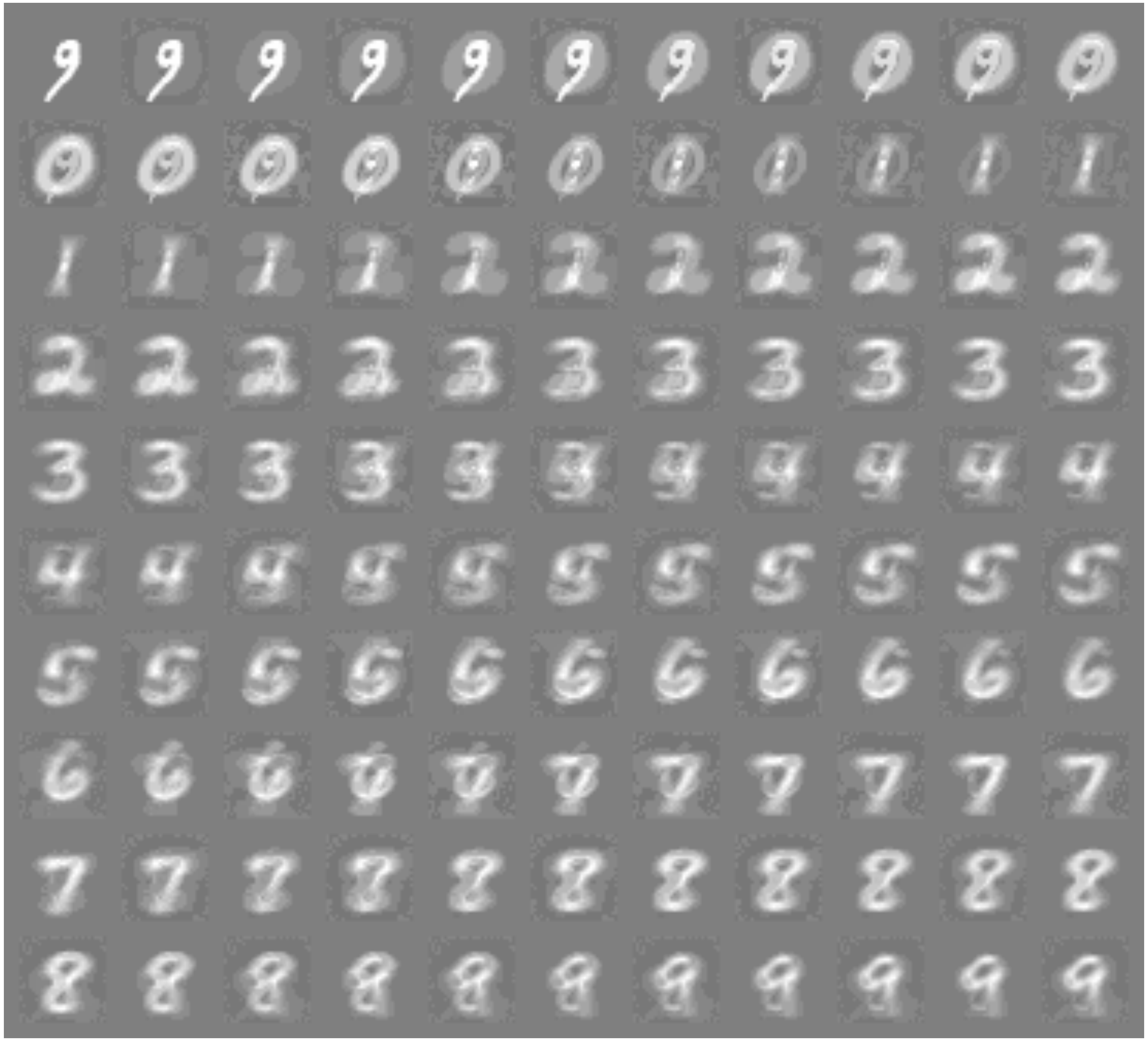
(Goodfellow 2017)

Adversarial Examples for RL



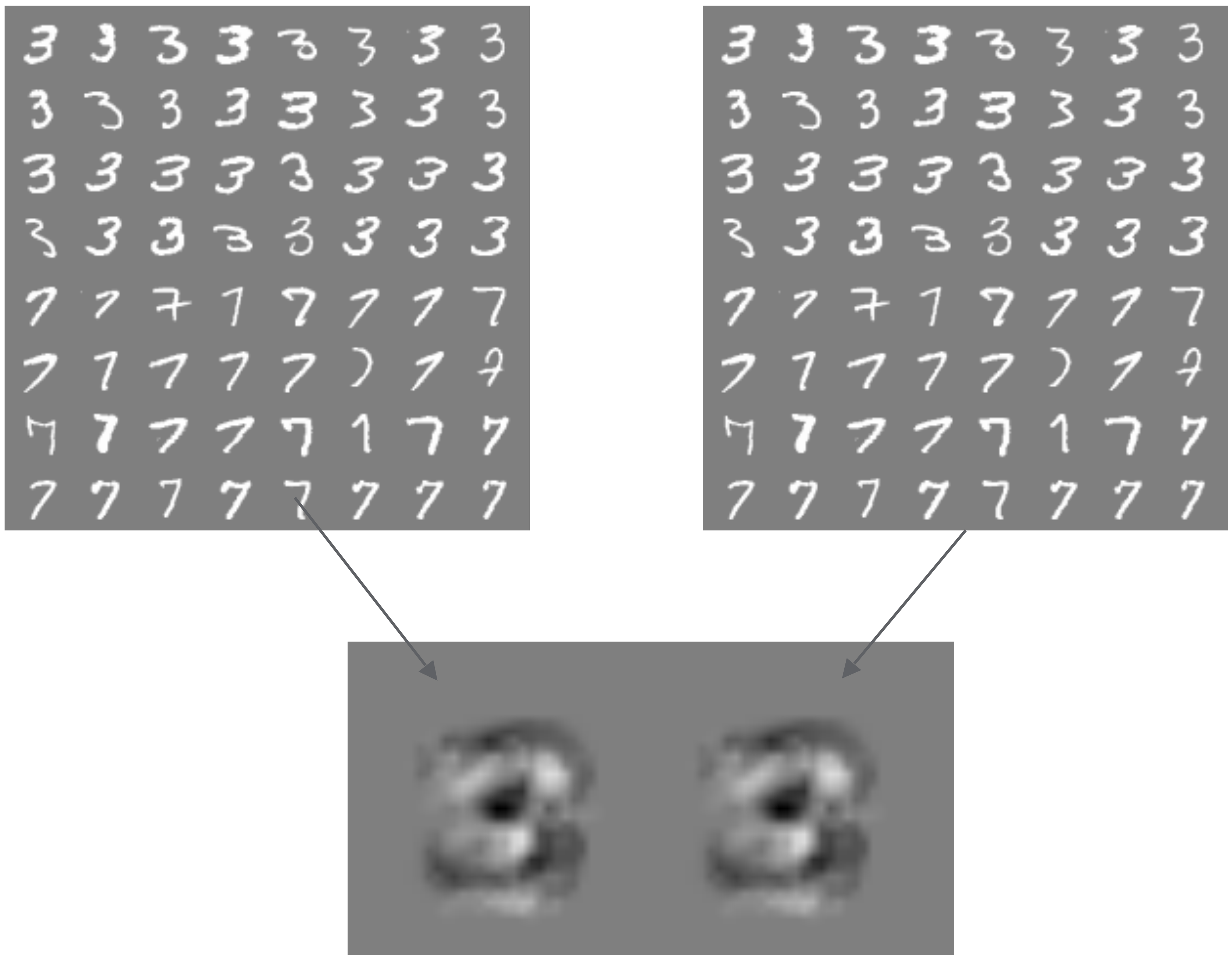
(Huang et al., 2017)

RBFs behave more intuitively



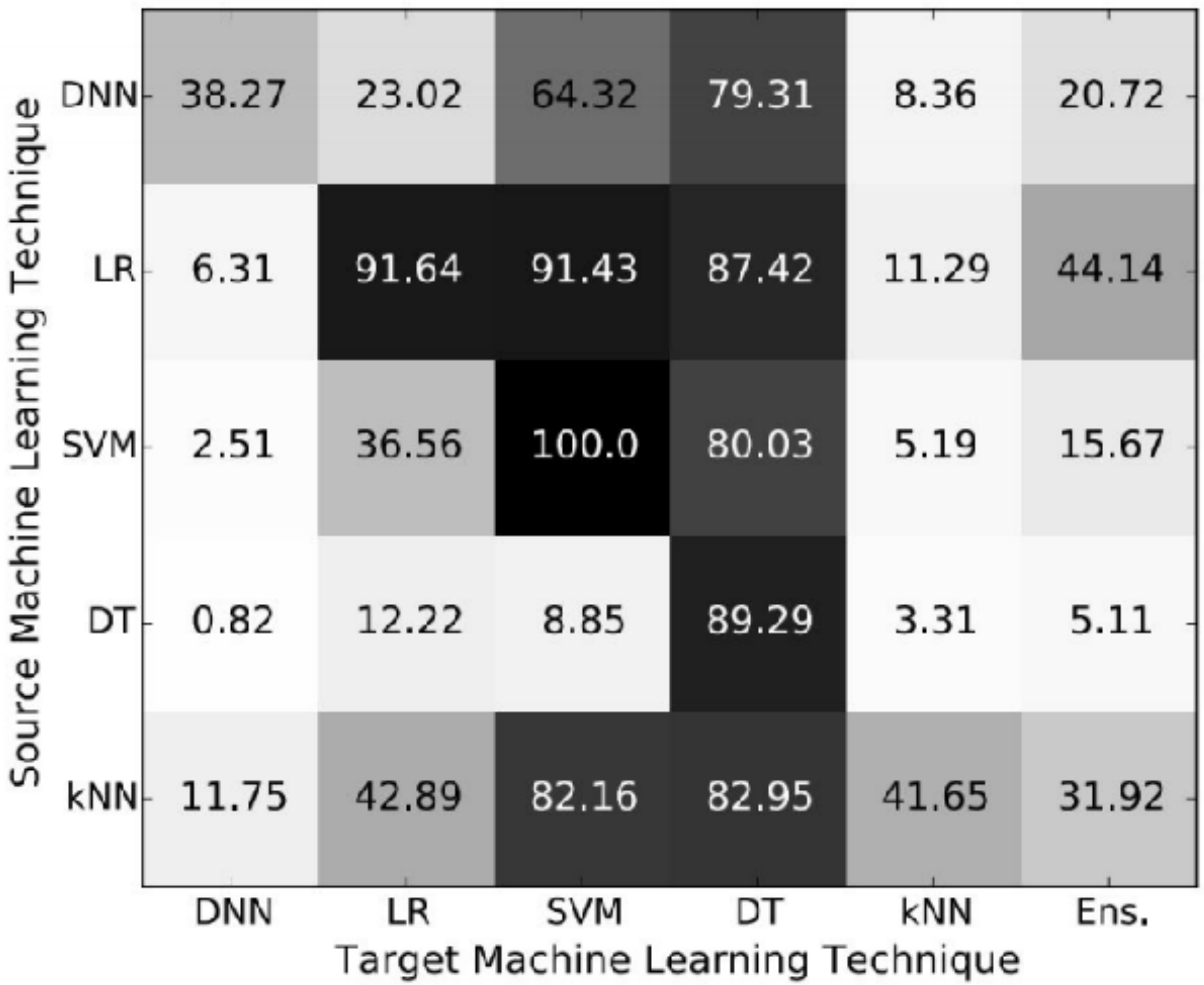
(Goodfellow 2017)

Cross-model, cross-dataset generalization



(Goodfellow 2017)

Cross-technique transferability



(Papernot 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

(Goodfellow 2017)

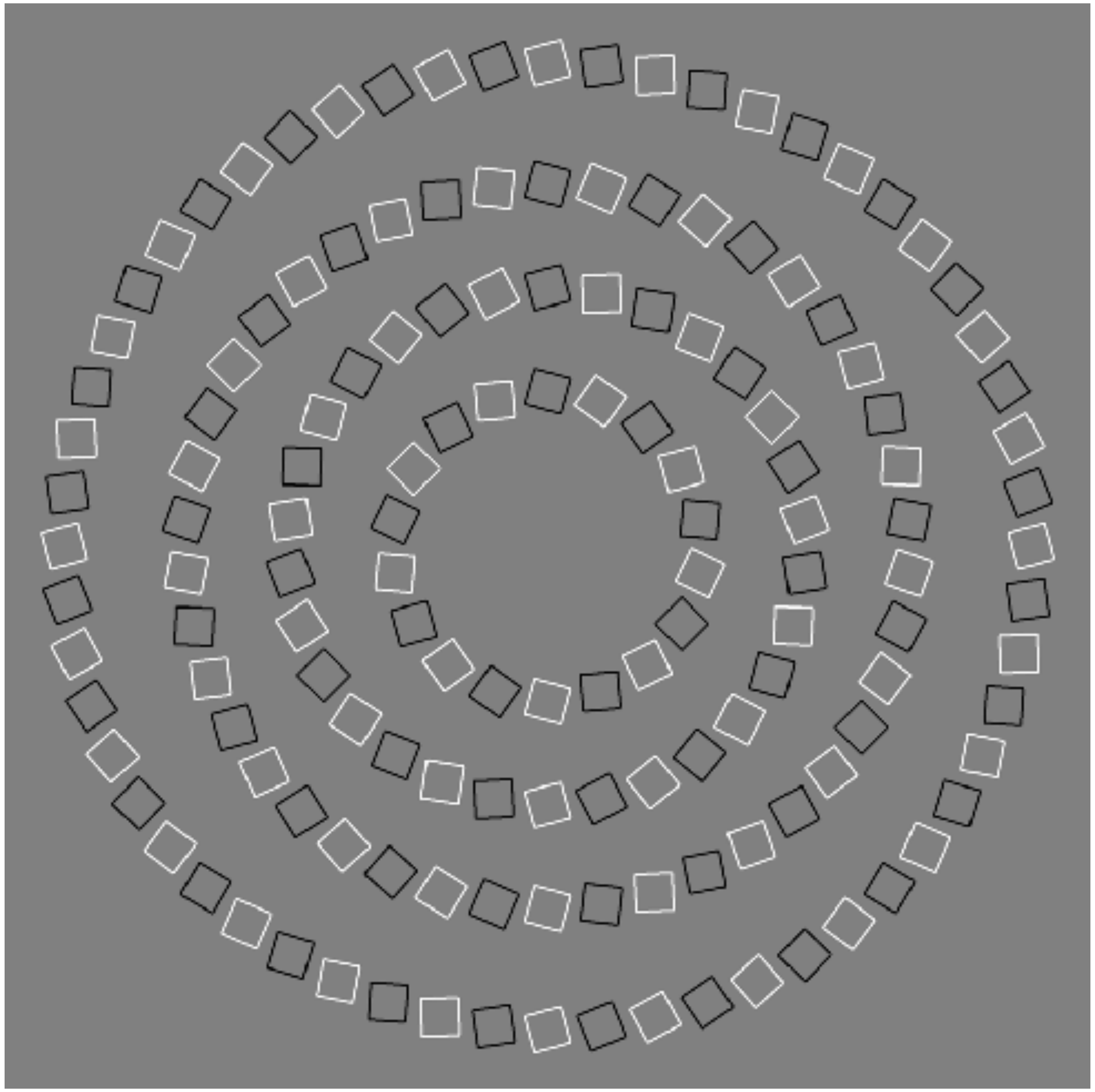
Enhancing Transfer With Ensembles

	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)

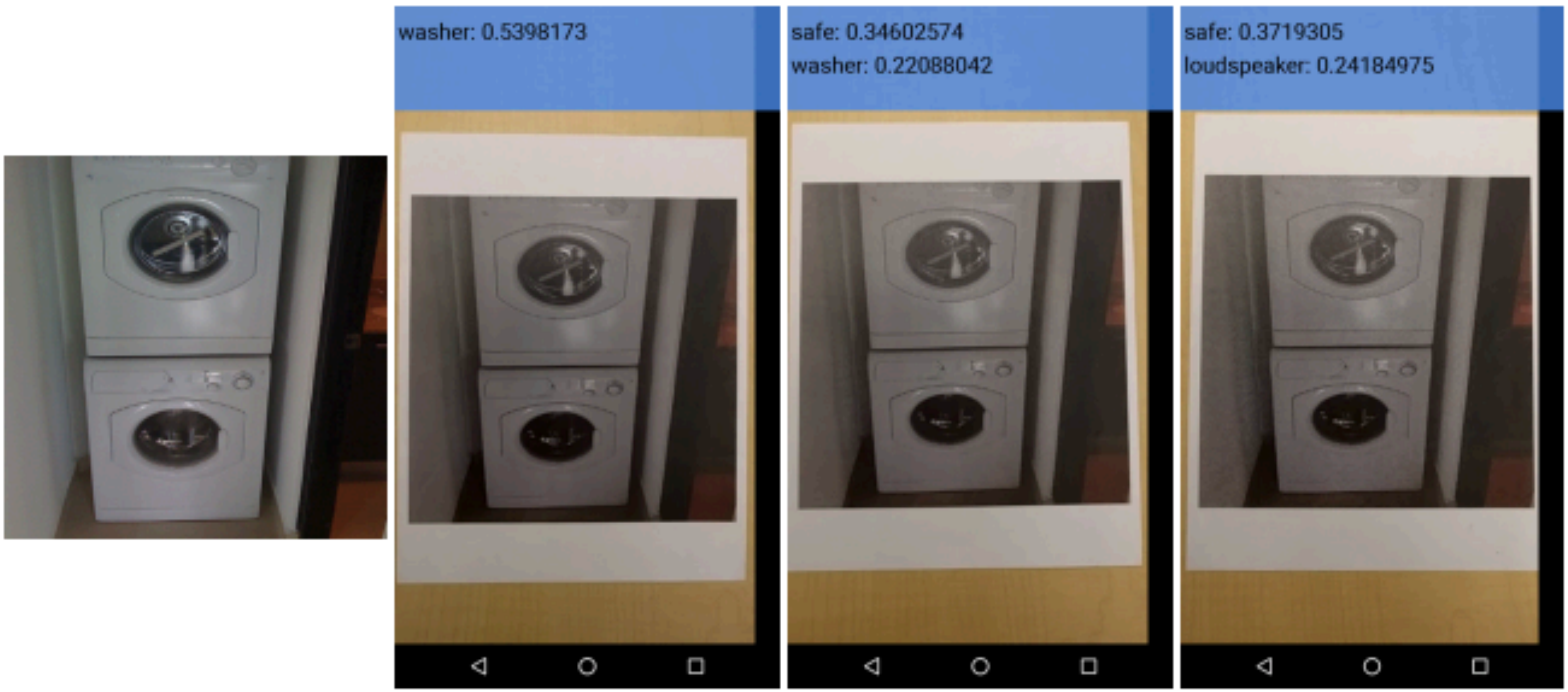
Adversarial Examples in the Human Brain



These are concentric circles, not intertwined spirals.

(Pinna and Gregory, 2002)

Adversarial Examples in the Physical World



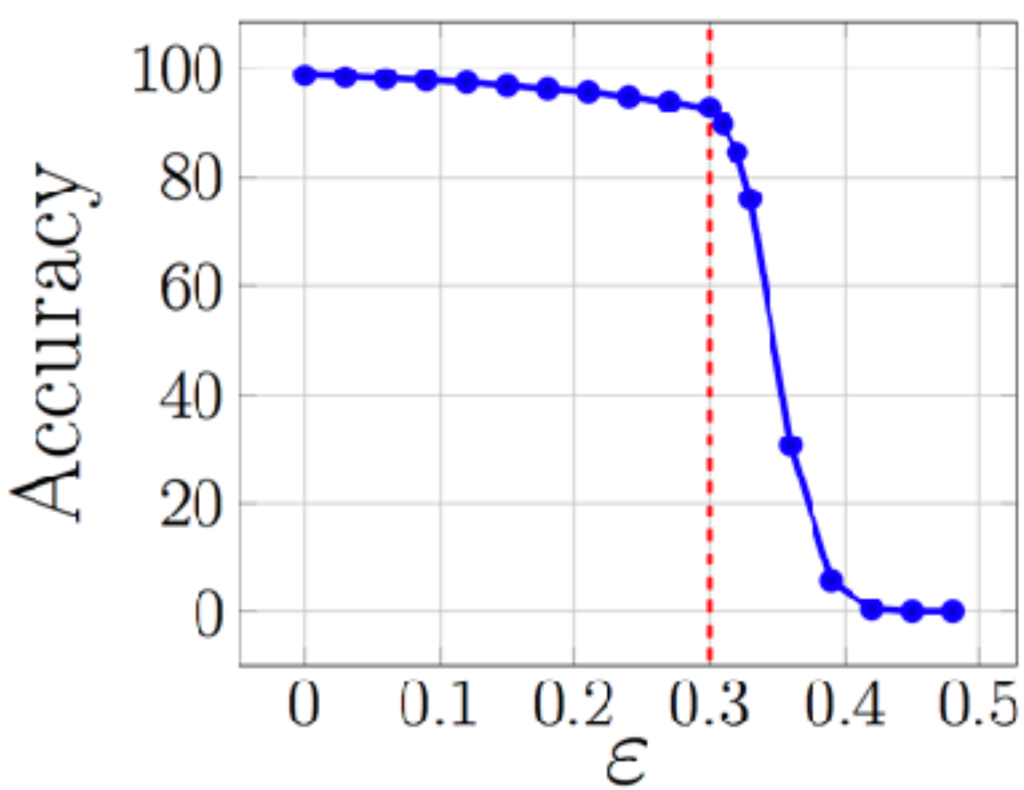
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

(Kurakin et al, 2016)

Training on Adversarial Examples

Success on MNIST?

- Open challenge to break model trained on adversarial perturbations initialized with noise
- Even strong, iterative white-box attacks can't get more than 12% error so far
- Larger datasets remain challenging



(Madry et al 2017)

Verification

- Given a seemingly robust model, can we prove that no adversarial examples exist near a given point?
- Yes, but hard to scale to large models (Huang et al 2016, Katz et al 2017)
- What about adversarial near test points that we don't know to examine ahead of time?

Competition

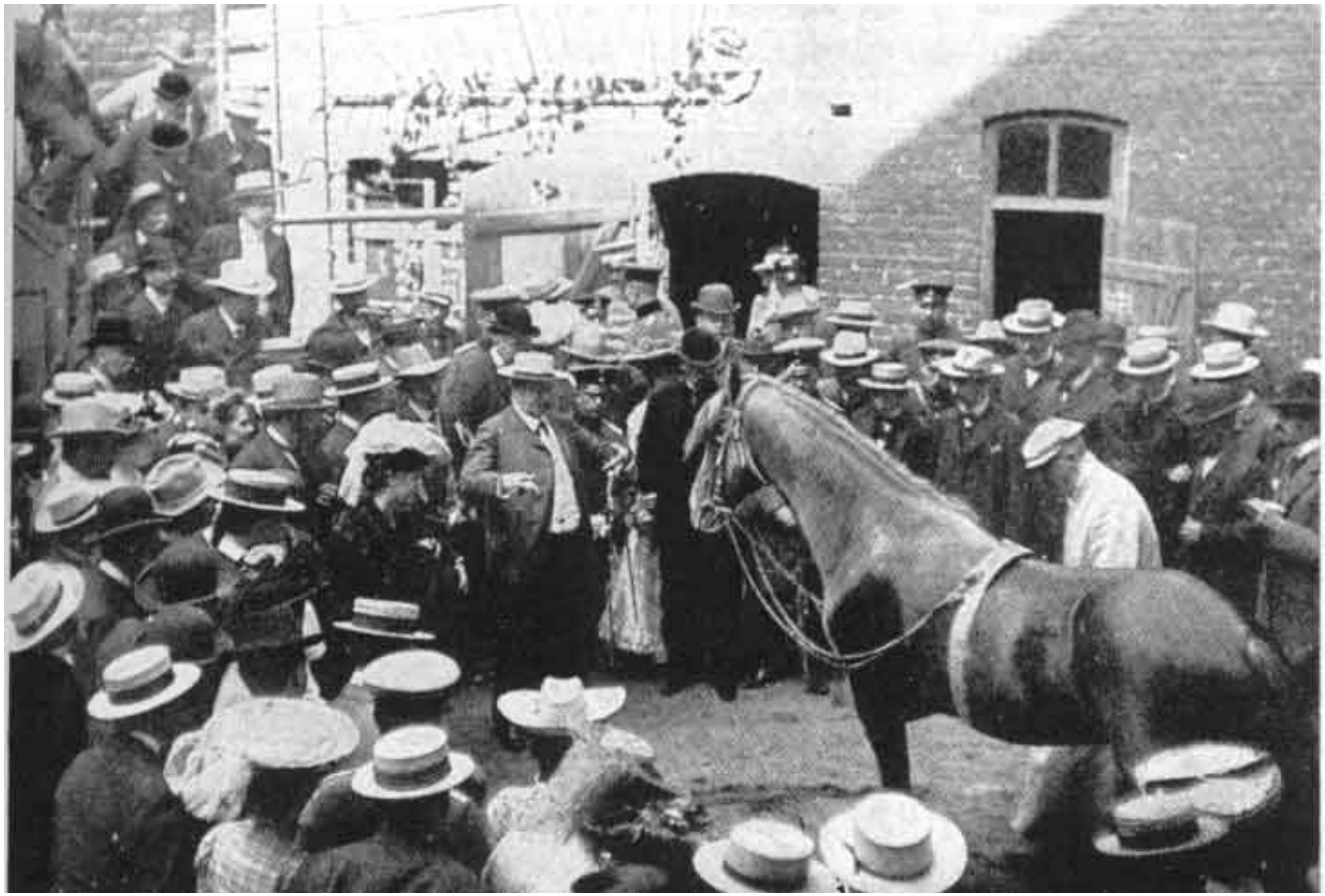
AI Fight Club Could Help Save Us from a Future of Super-Smart Cyberattacks

MIT Technology Review

Best defense so far on ImageNet:
Ensemble adversarial training,
Tramèr et al 2017.

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

Clever Hans



("Clever Hans,
Clever
Algorithms,"
Bob Sturm)

Get involved!

https://github.com/tensorflow/cleverhans

cleverhans

Check out Justin Gilmer's
BayLearn poster on Adversarial
Sphere