RSAConference2018

San Francisco | April 16 – 20 | Moscone Center

MATTERS NOW

SESSION ID:

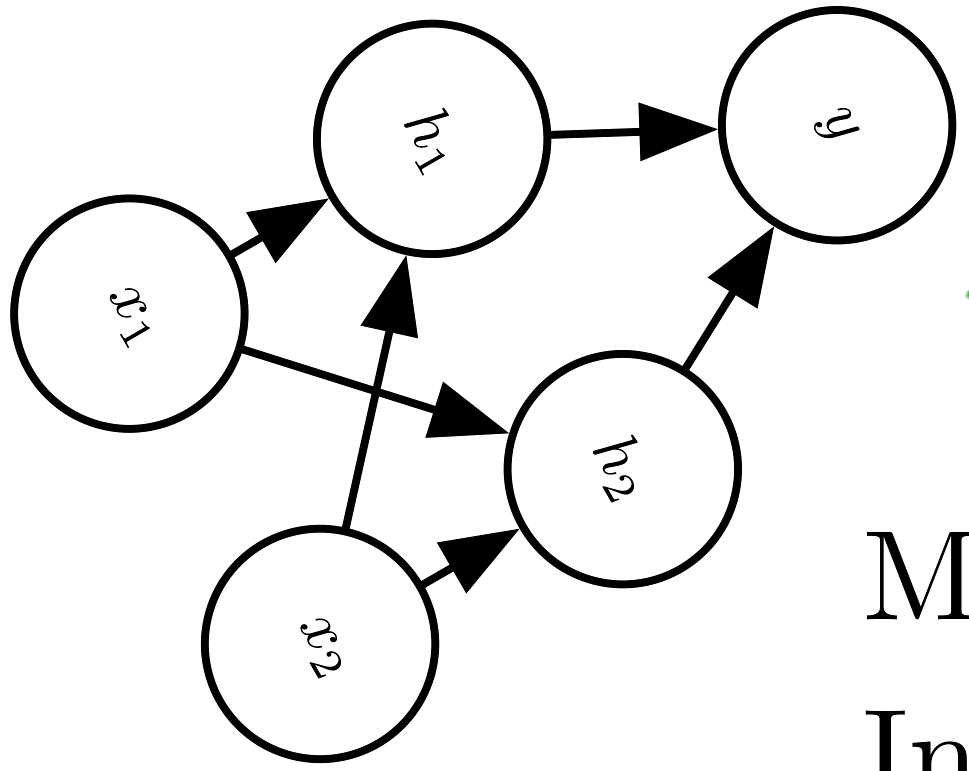
SECURITY AND PRIVACY OF MACHINE LEARNING

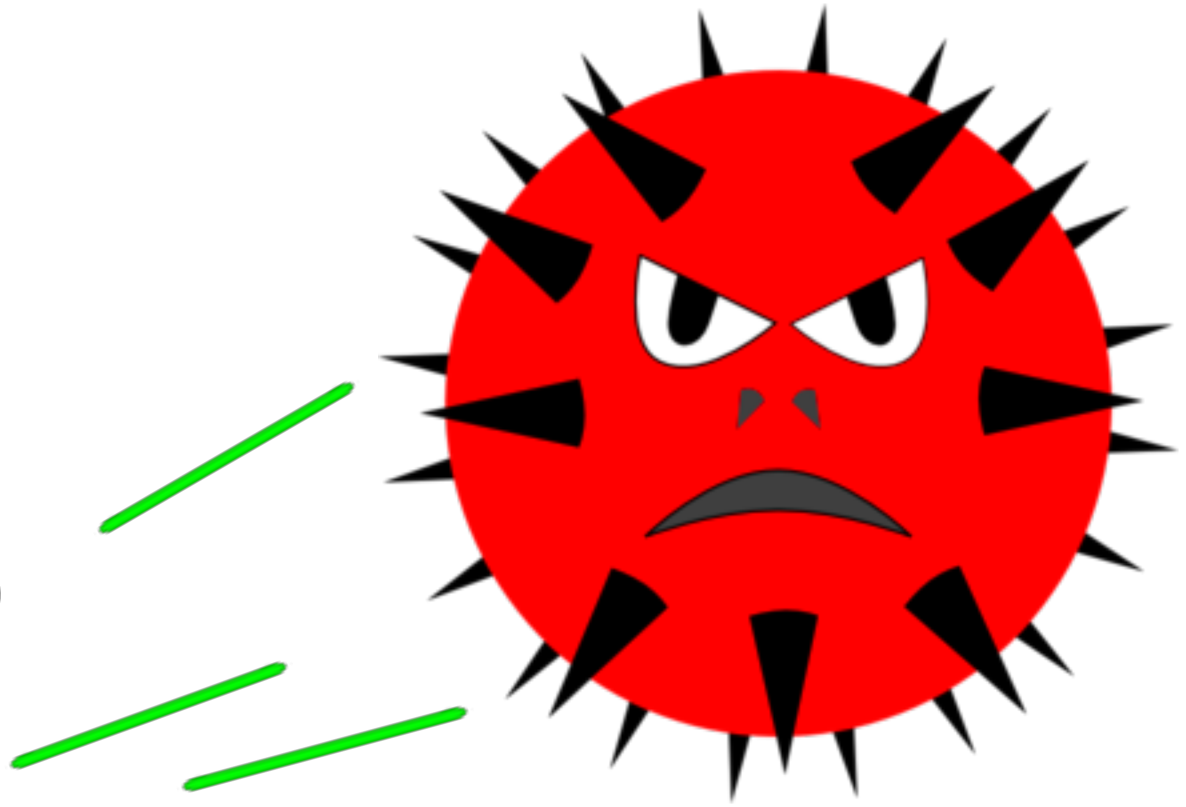
Ian Goodfellow

Staff Research Scientist
Google Brain
@goodfellow_ian

Machine Learning and Security

Machine Learning for Security



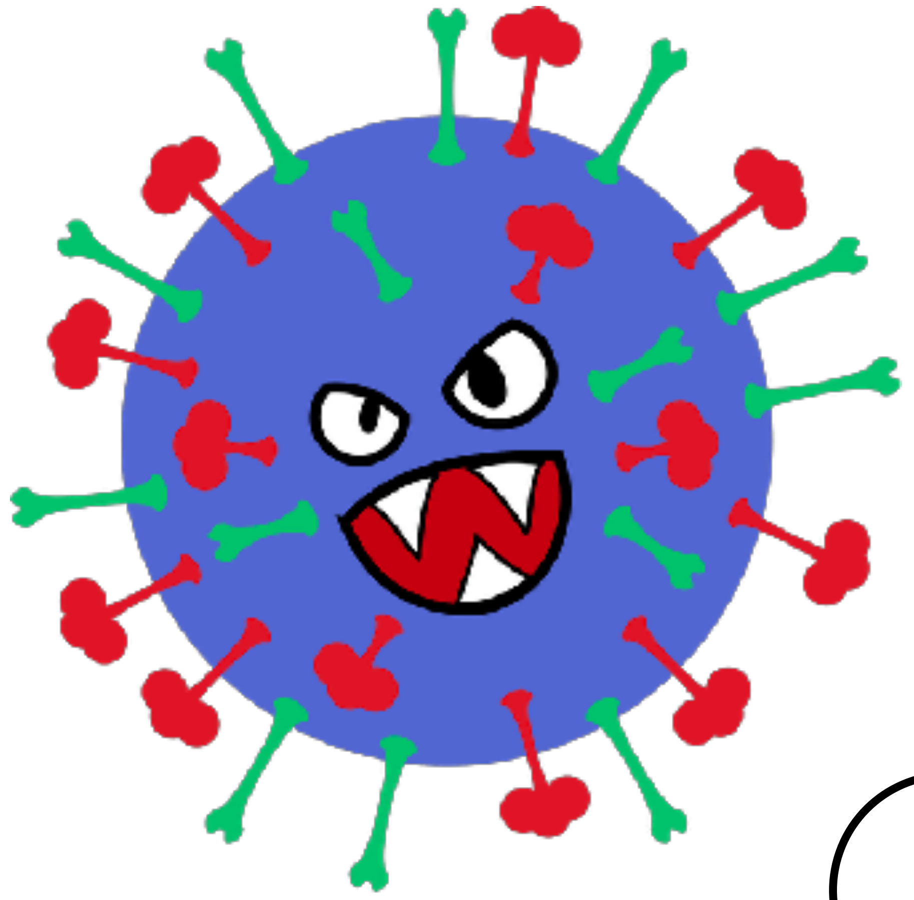


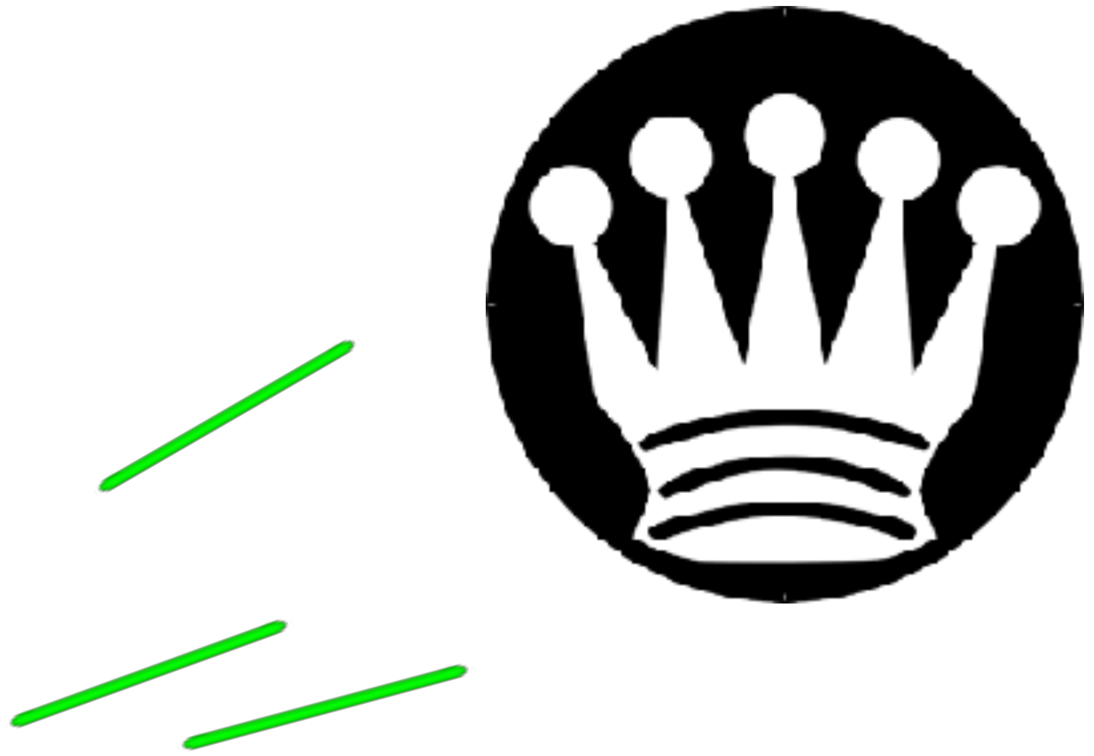
Malware detection

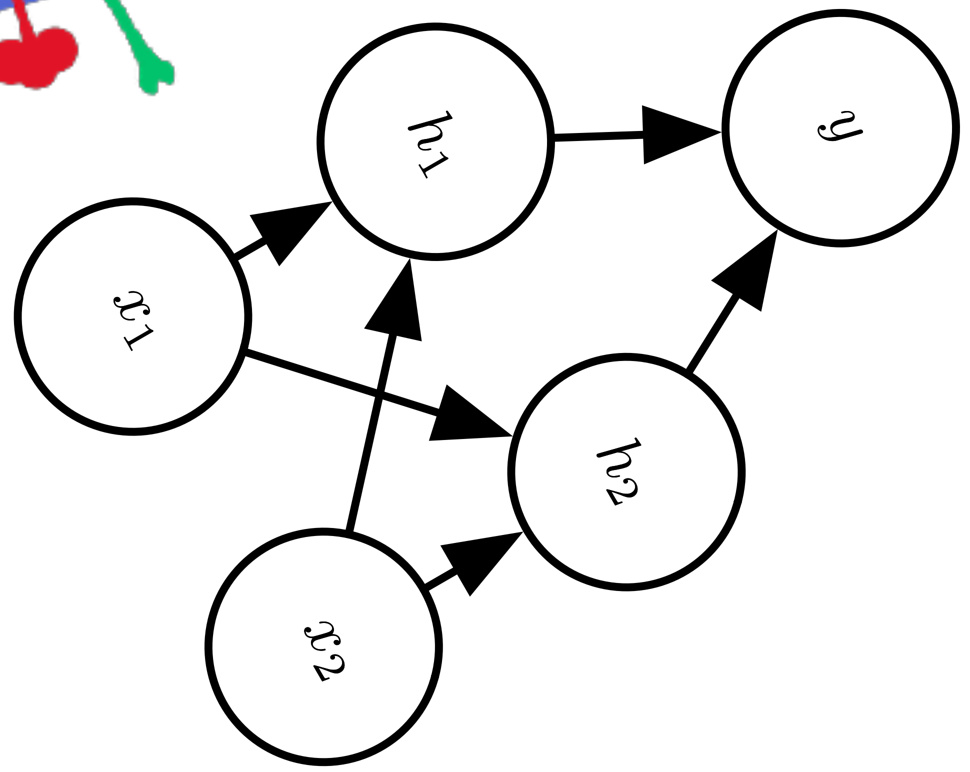
Intrusion detection

...

Security against Machine Learning







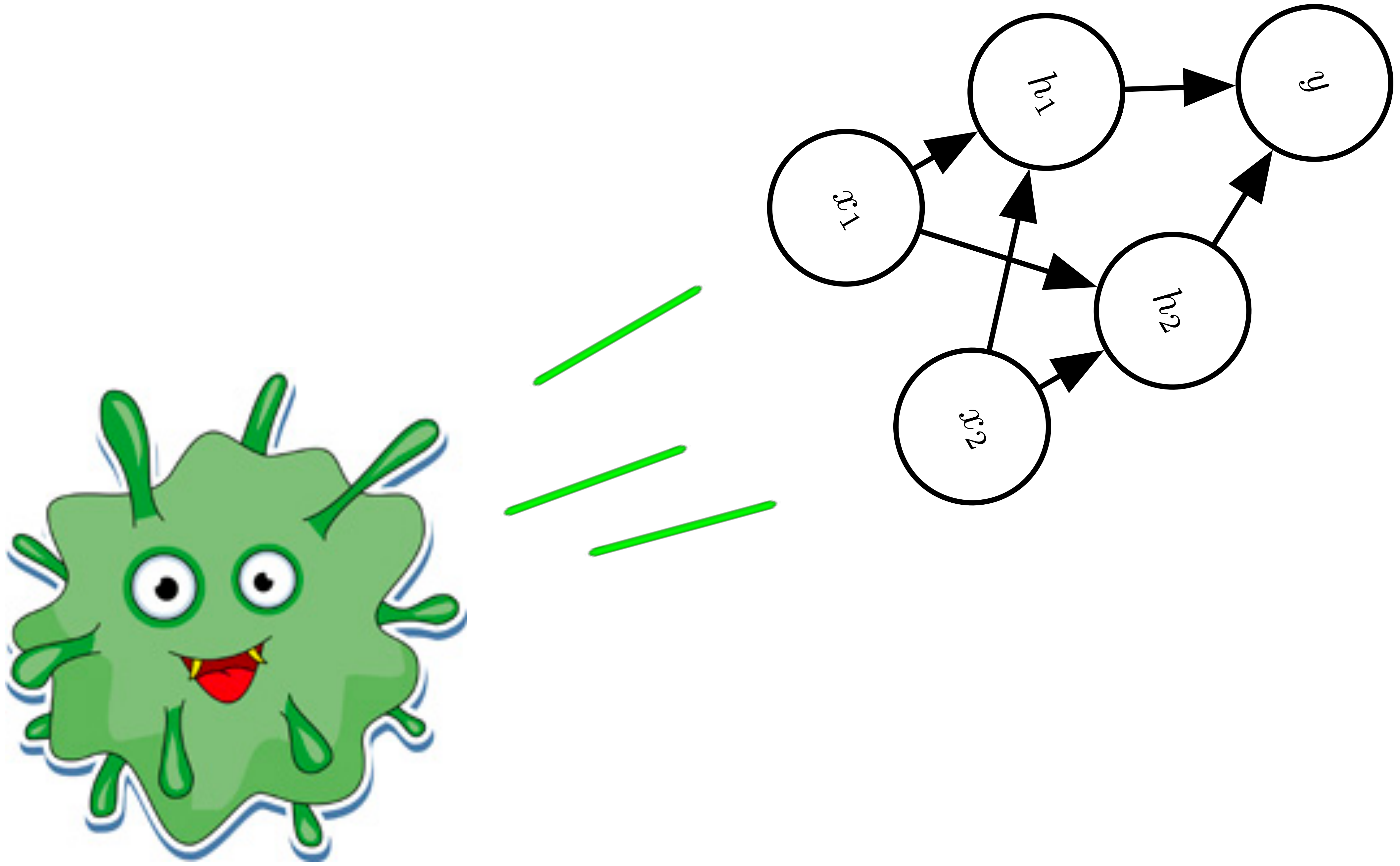
Password guessing

Fake reviews

...

(Goodfellow 2018)

Security of Machine Learning



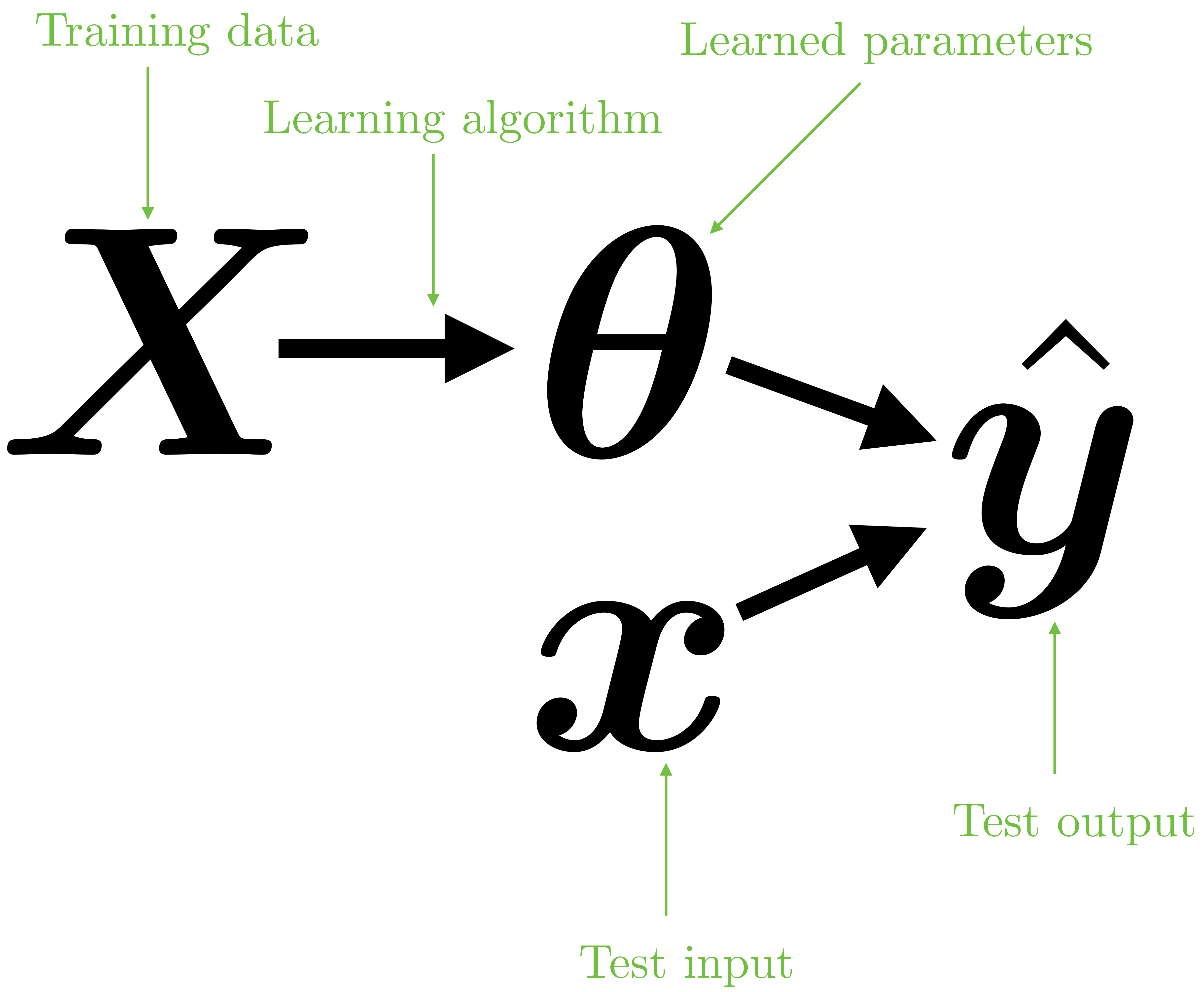
(Goodfellow 2018)

An overview of a field

- This presentation summarizes the work of many people, not just my own / my collaborators
- Download the slides for this link to extensive references
- The presentation focuses on the *concepts*, not the history or the inventors

(Goodfellow 2018)

Machine Learning Pipeline



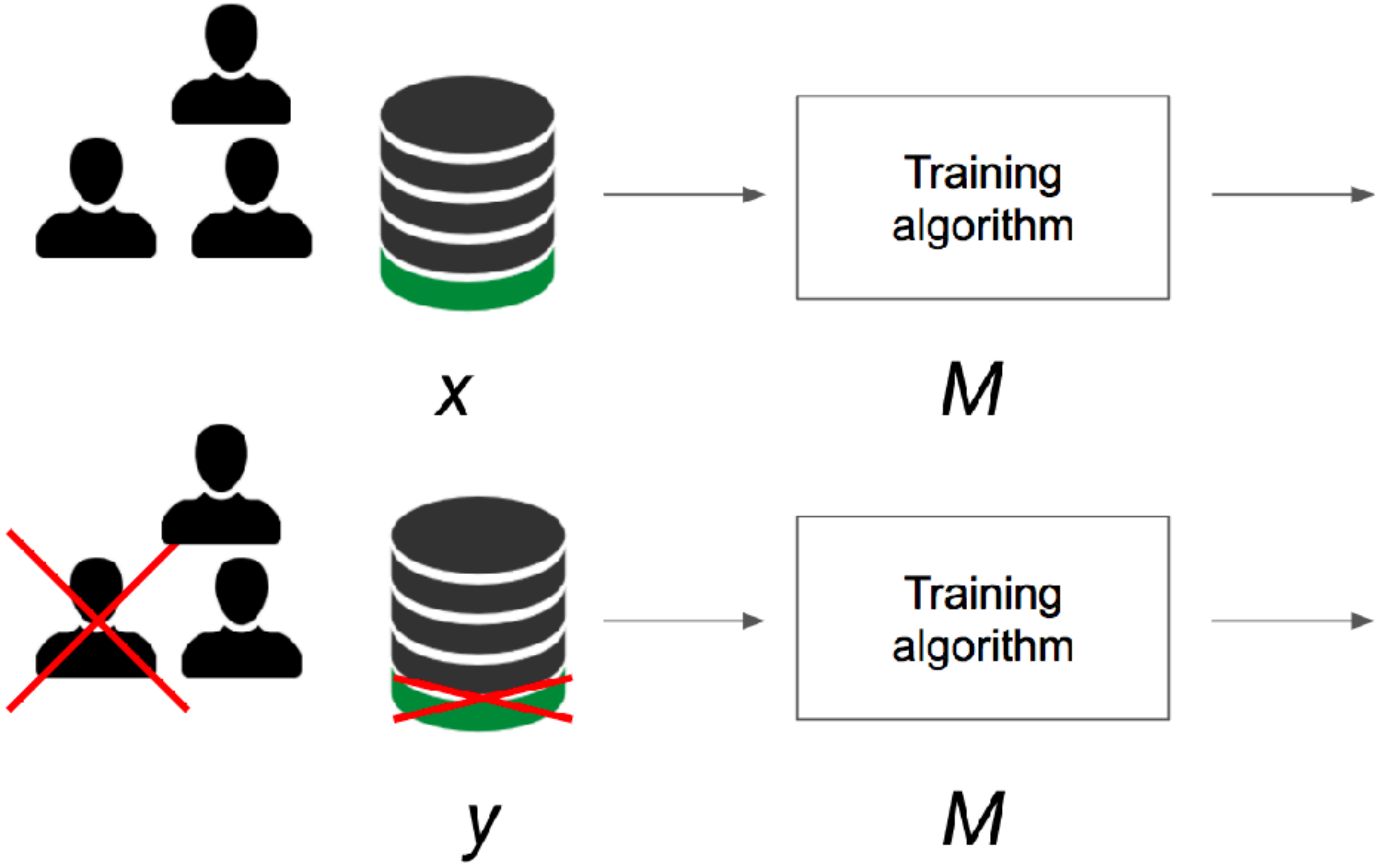
(Goodfellow 2018)

Privacy of Training Data

$$X \rightarrow \theta \rightarrow \hat{X}$$

(Goodfellow 2018)

Defining (ε, δ)-Differential Privacy

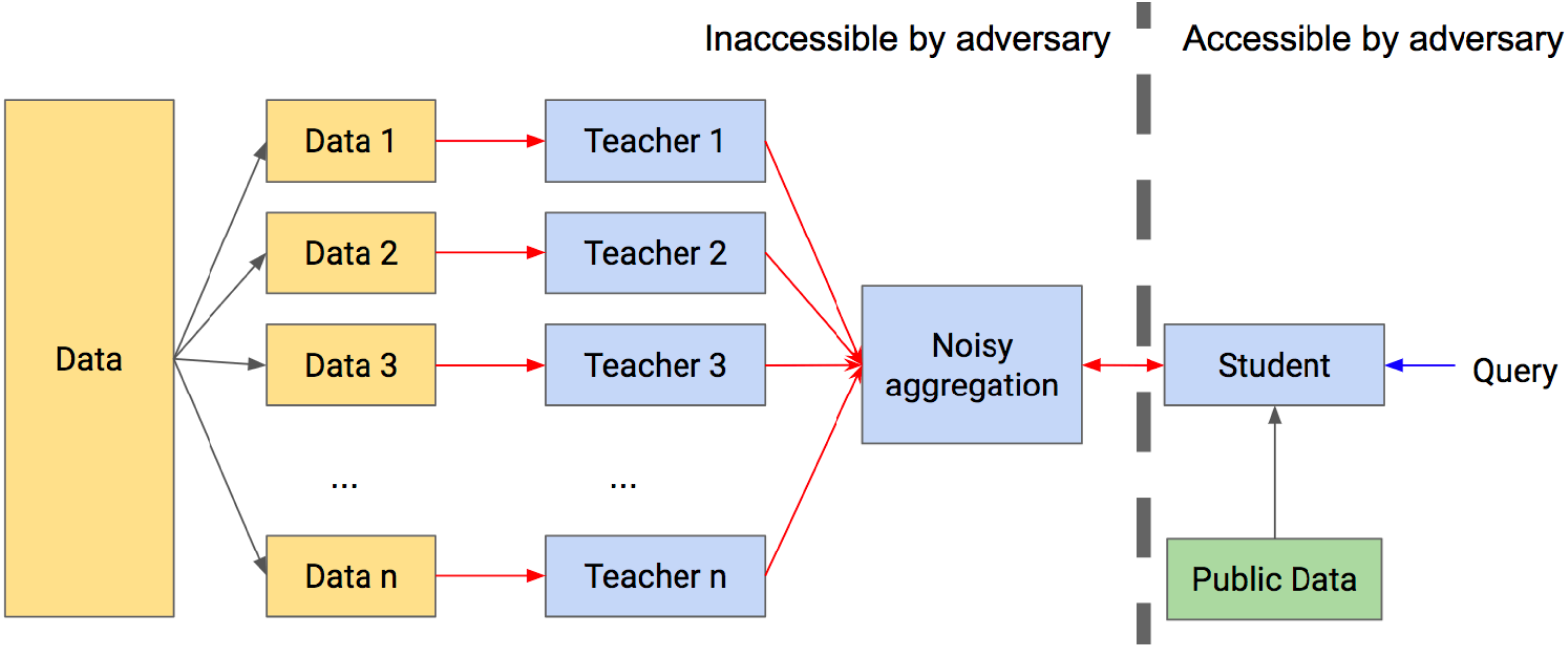


$$\forall S:\ \Pr[M(x) \in S] \le e^{\varepsilon} \cdot \Pr[M(y) \in S] + \delta$$

(Abadi 2017)

(Goodfellow 2018)

Private Aggregation of Teacher Ensembles



(Papernot et al 2016)

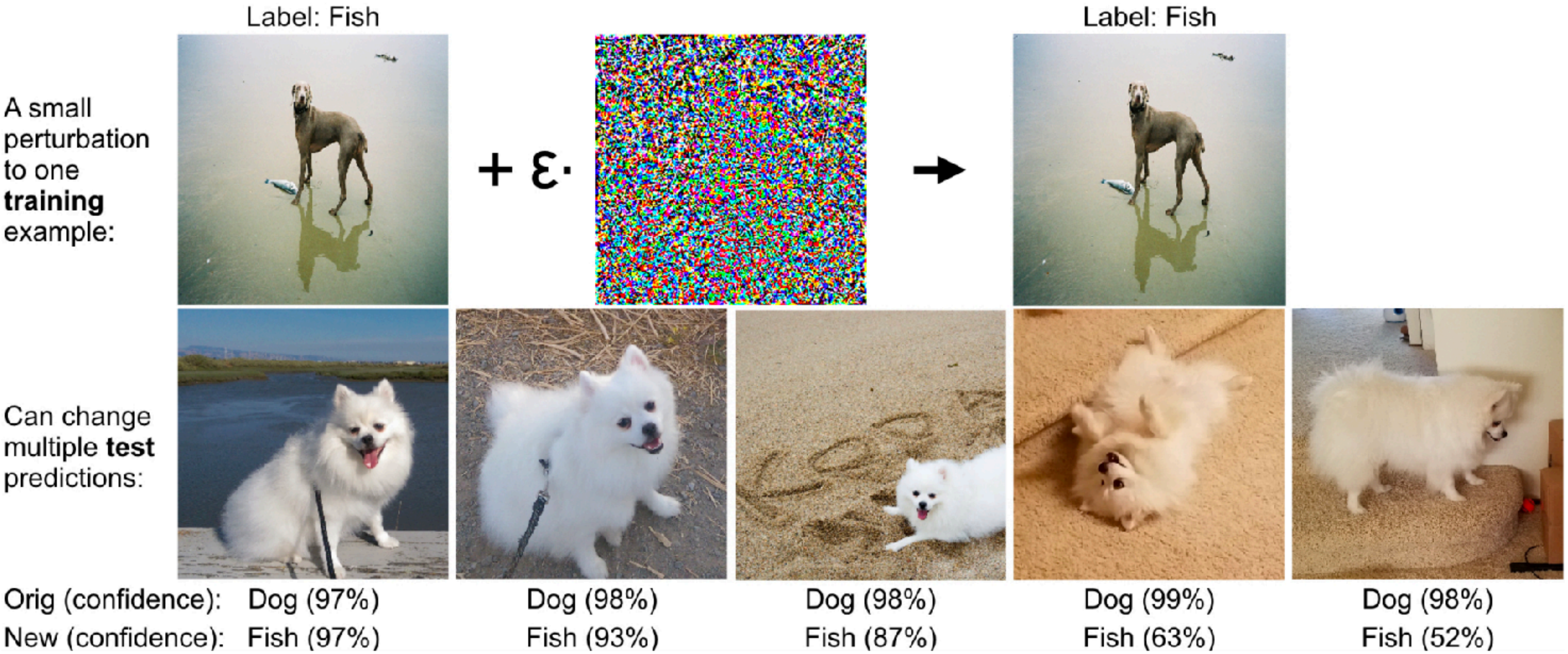
(Goodfellow 2018)

Training Set Poisoning

$$\boldsymbol{X} \longrightarrow \boldsymbol{\theta} \longrightarrow \hat{\boldsymbol{y}} \longleftarrow \boldsymbol{x}$$

(Goodfellow 2018)

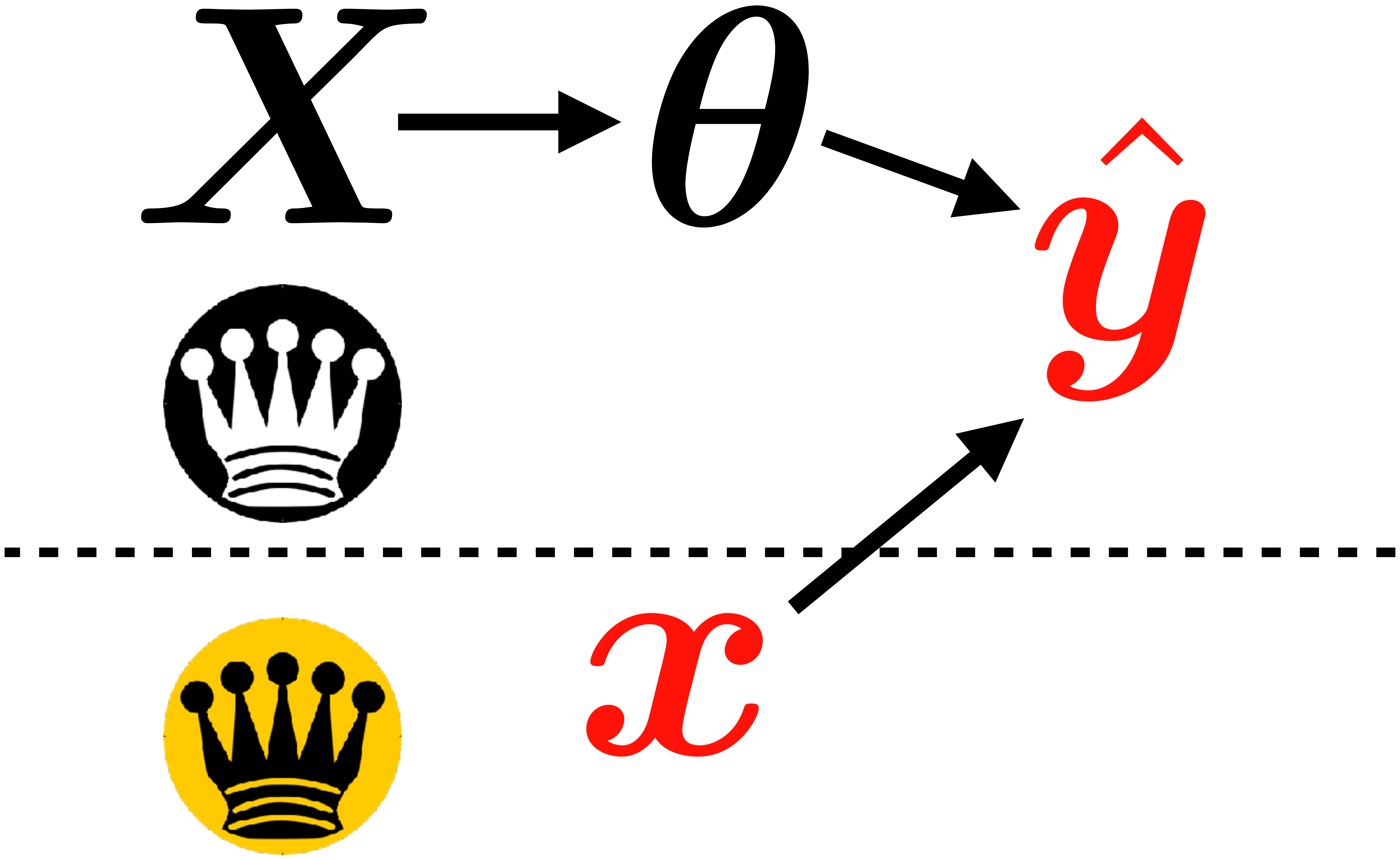
ImageNet Poisoning



(Koh and Liang 2017)

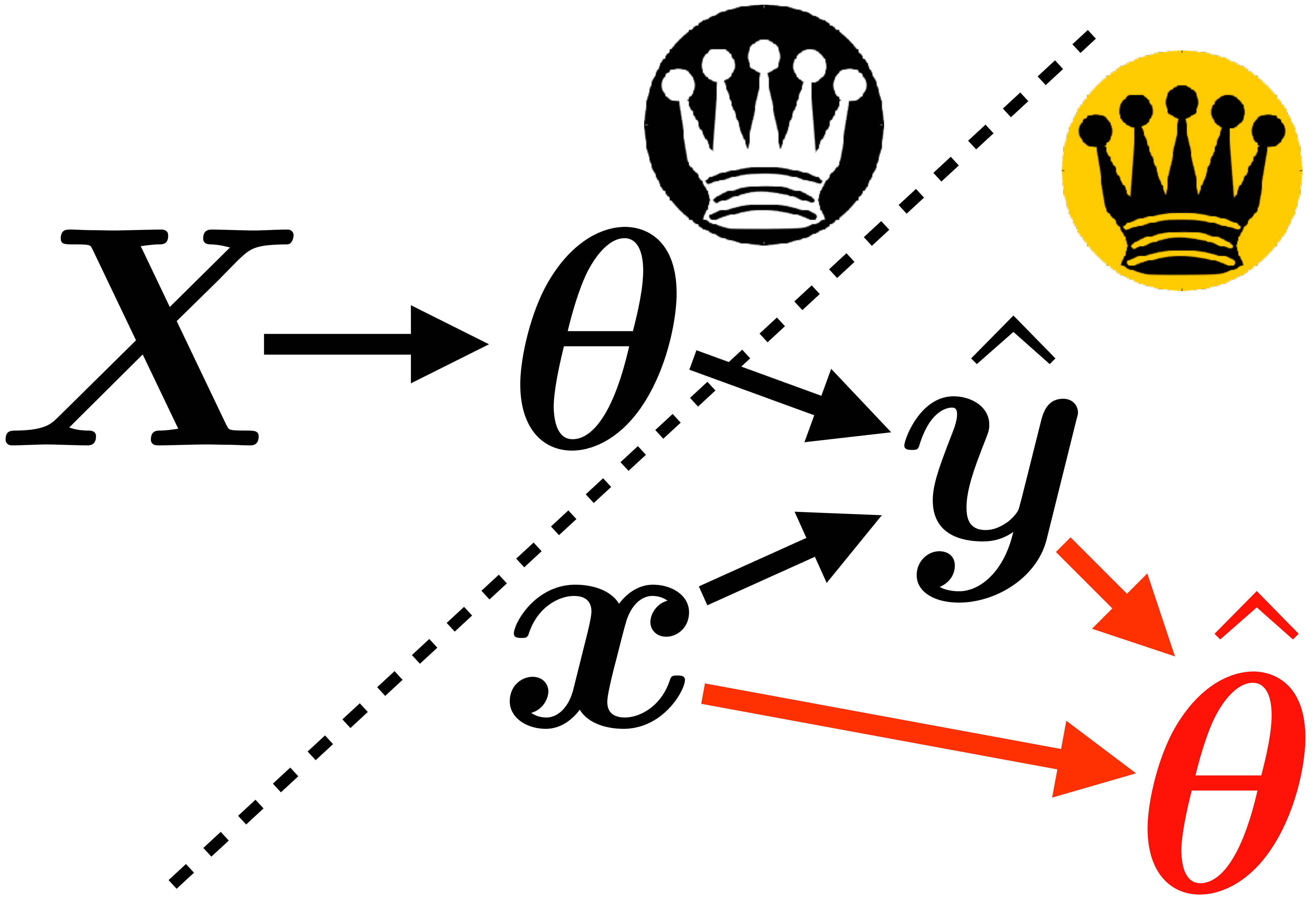
(Goodfellow 2018)

Adversarial Examples



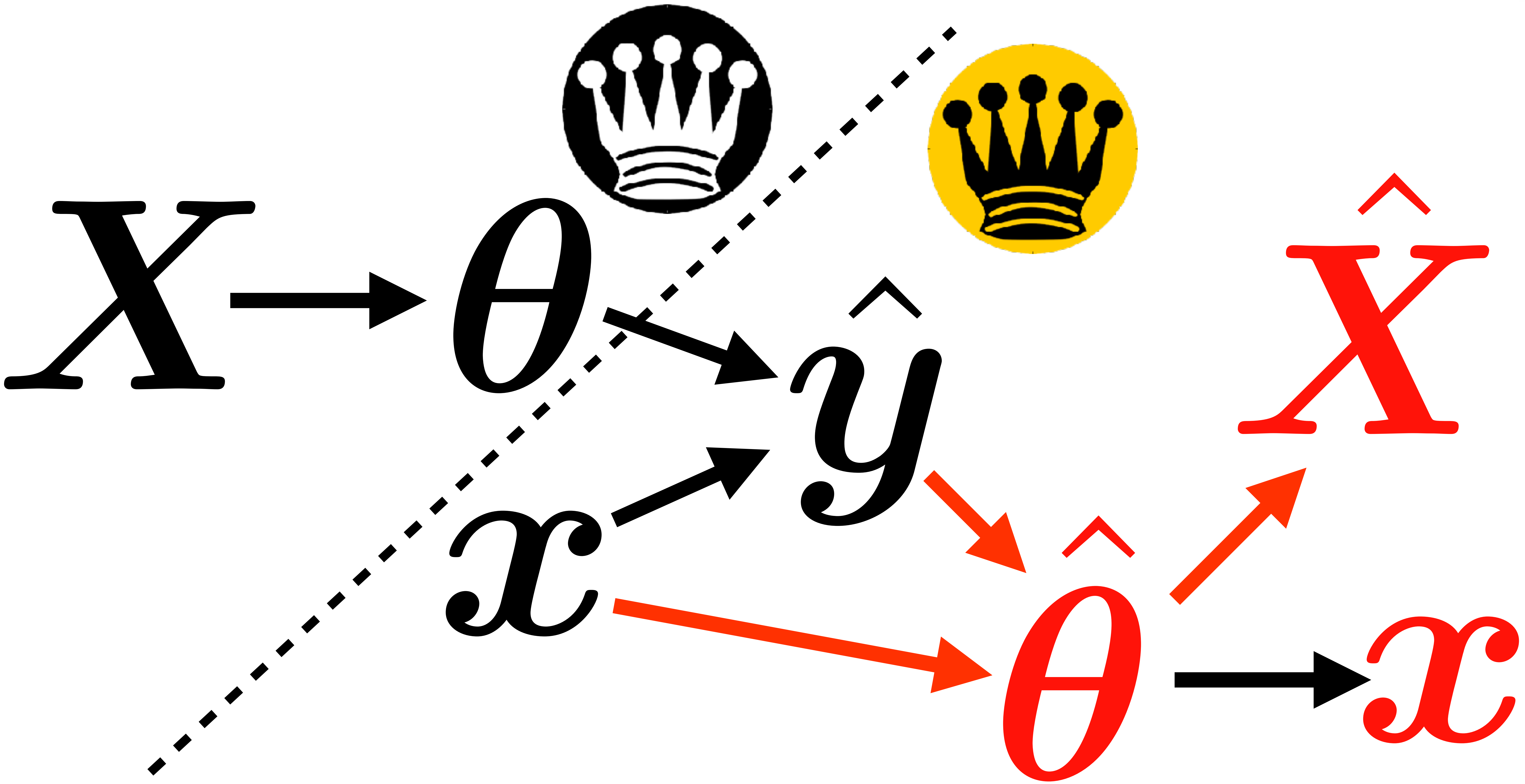
(Goodfellow 2018)

Model Theft



(Goodfellow 2018)

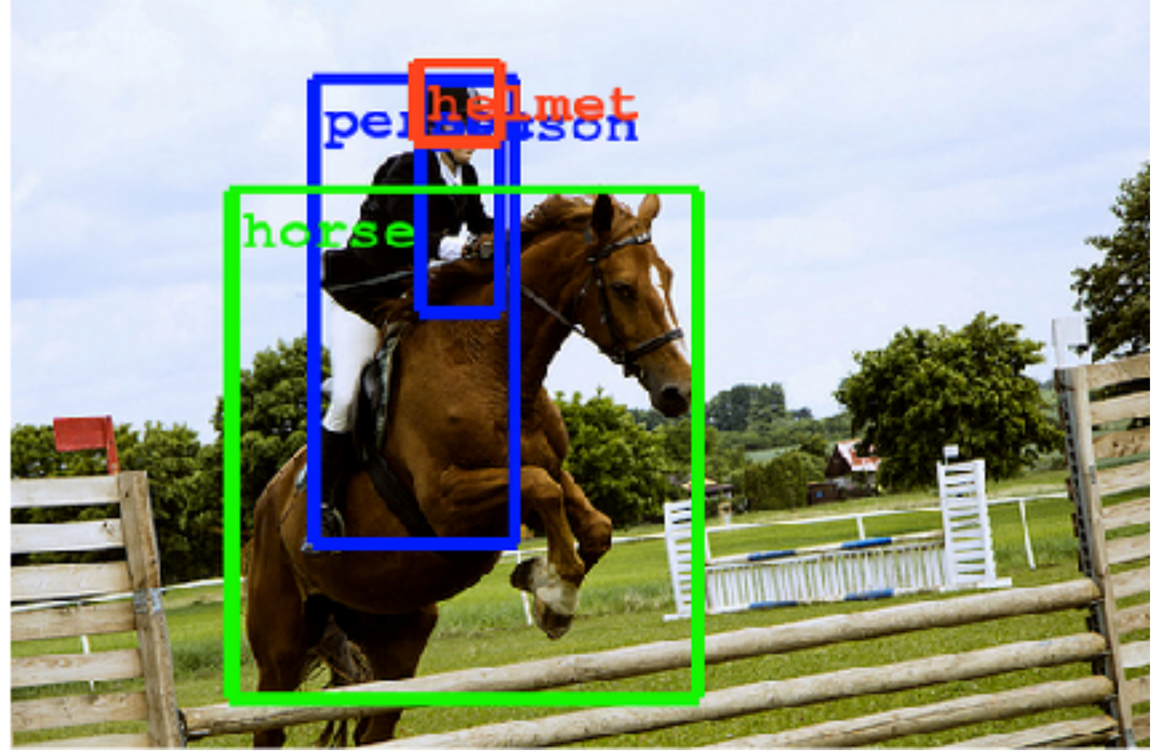
Model Theft++



(Goodfellow 2018)

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...

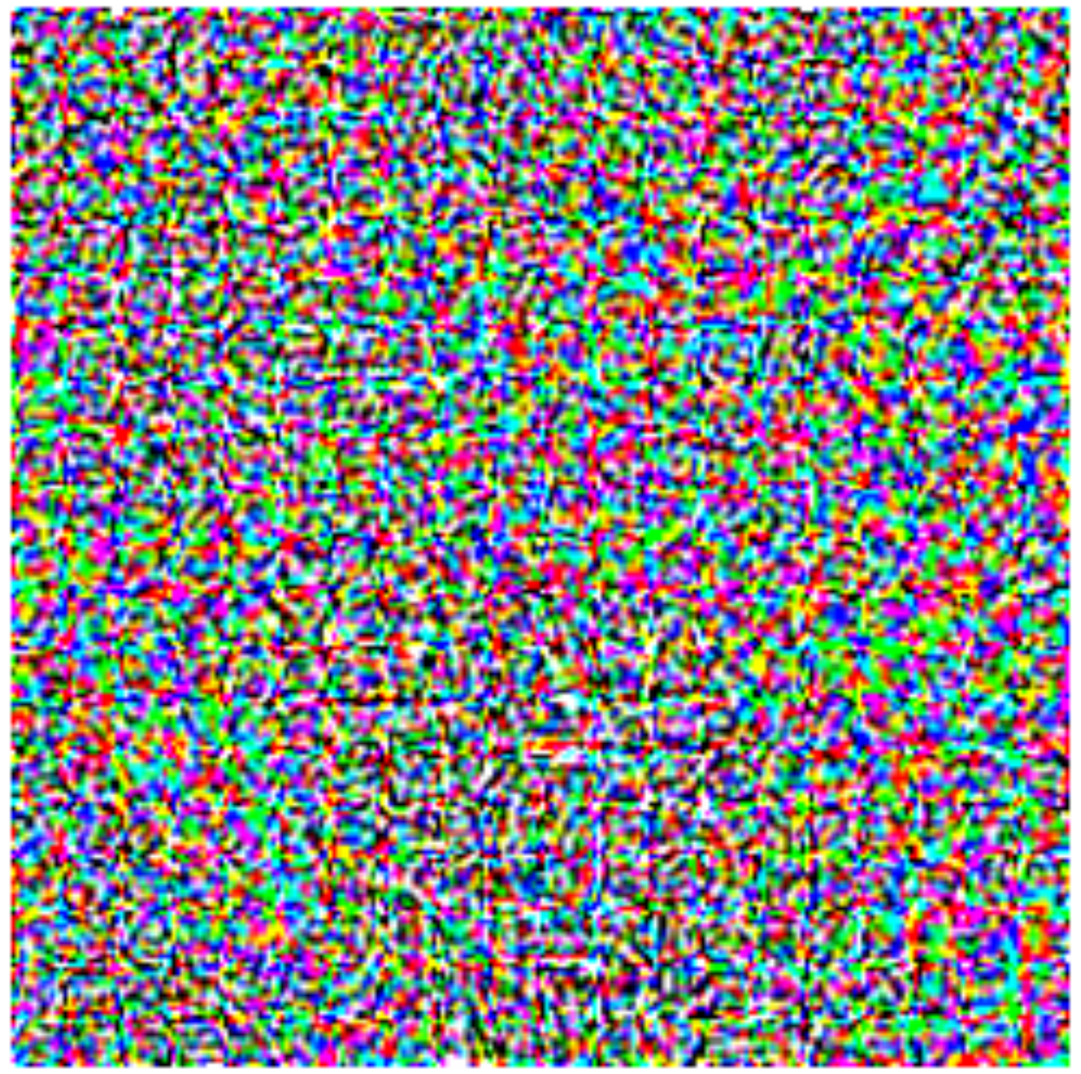
(Goodfellow et al, 2013)

and other tasks...

(Goodfellow 2018)

Adversarial Examples

$+ .007 \times$



$=$

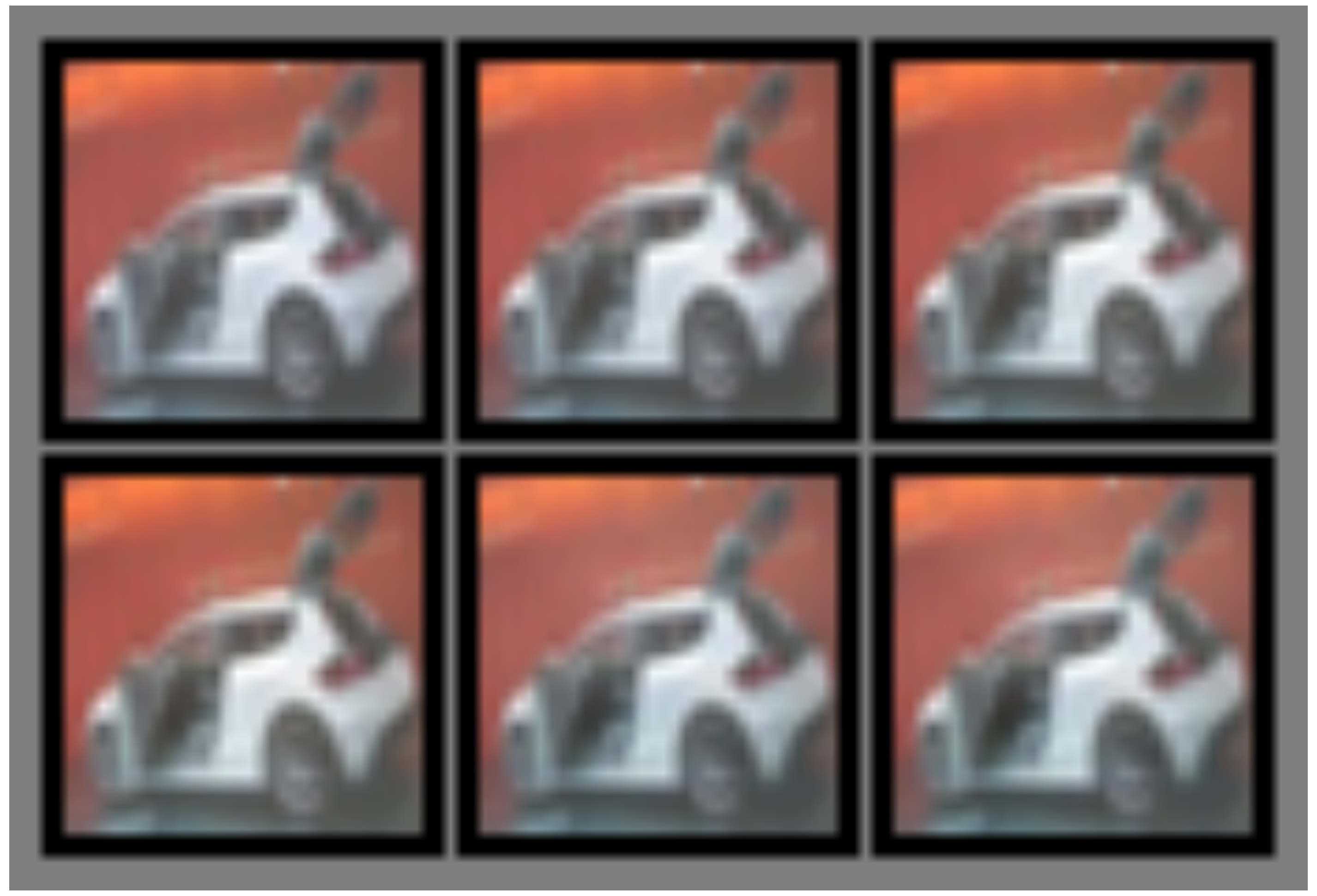
$\boldsymbol{x}$
"panda"
57.7% confidence

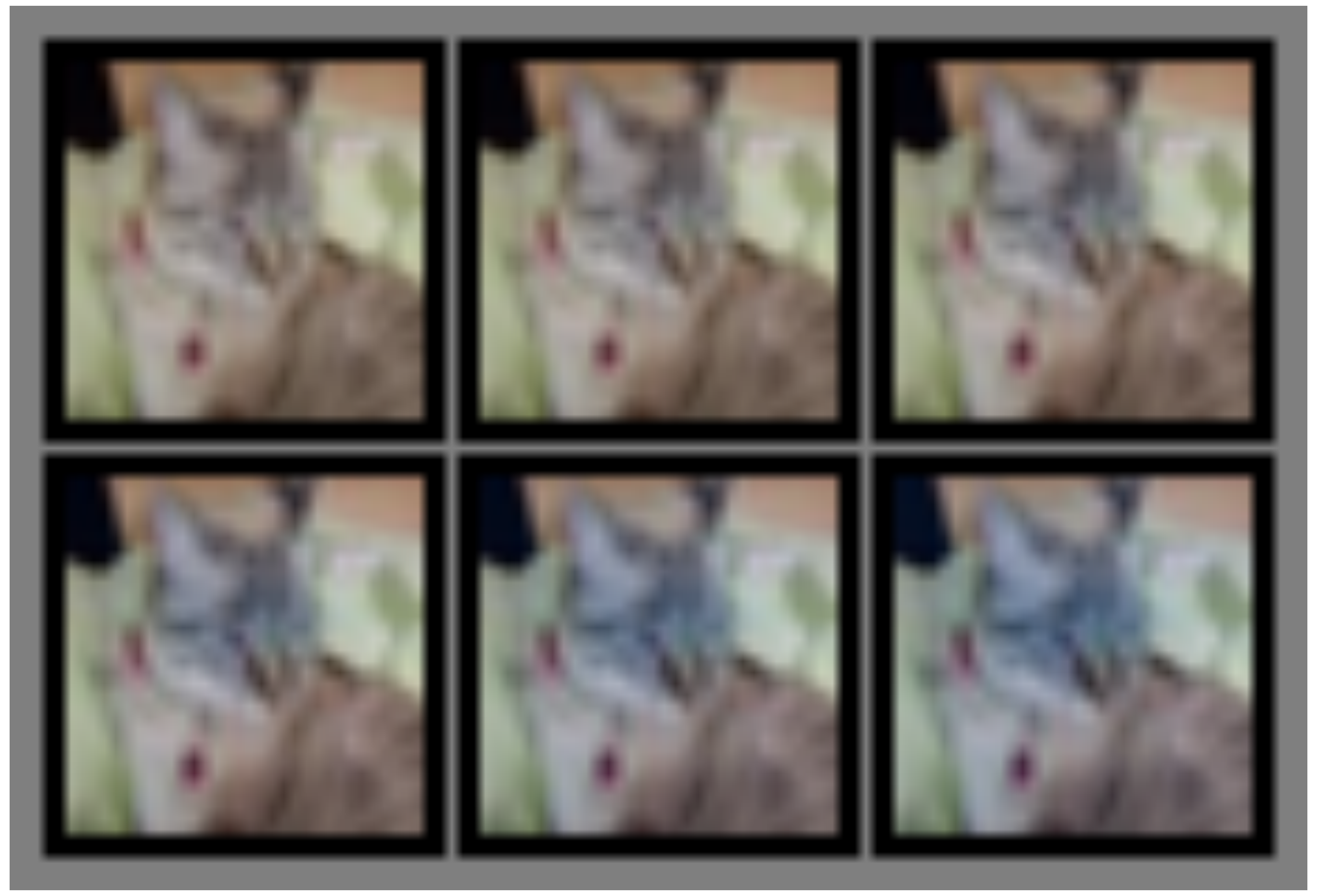
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

(Goodfellow 2018)

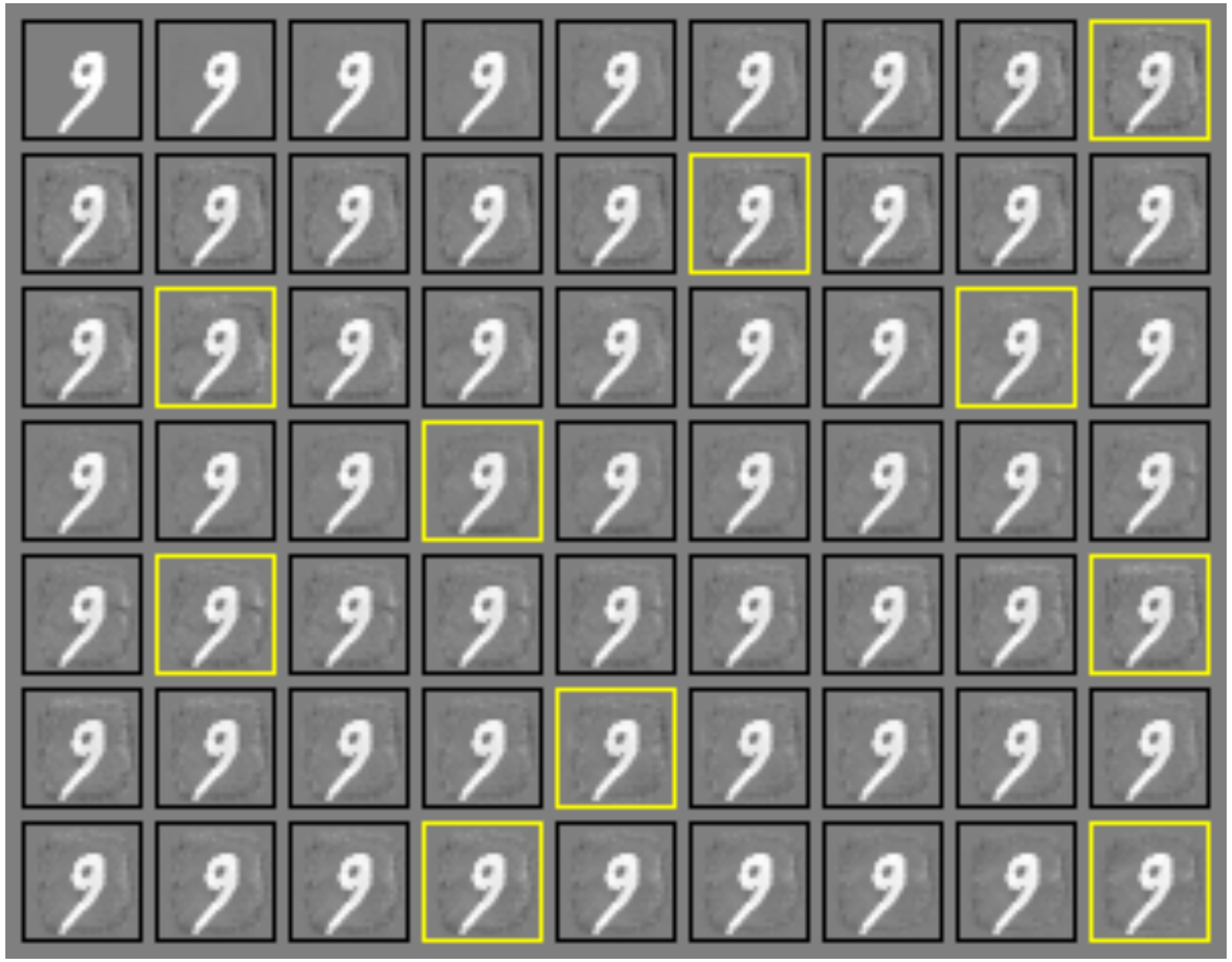
Turning objects into airplanes





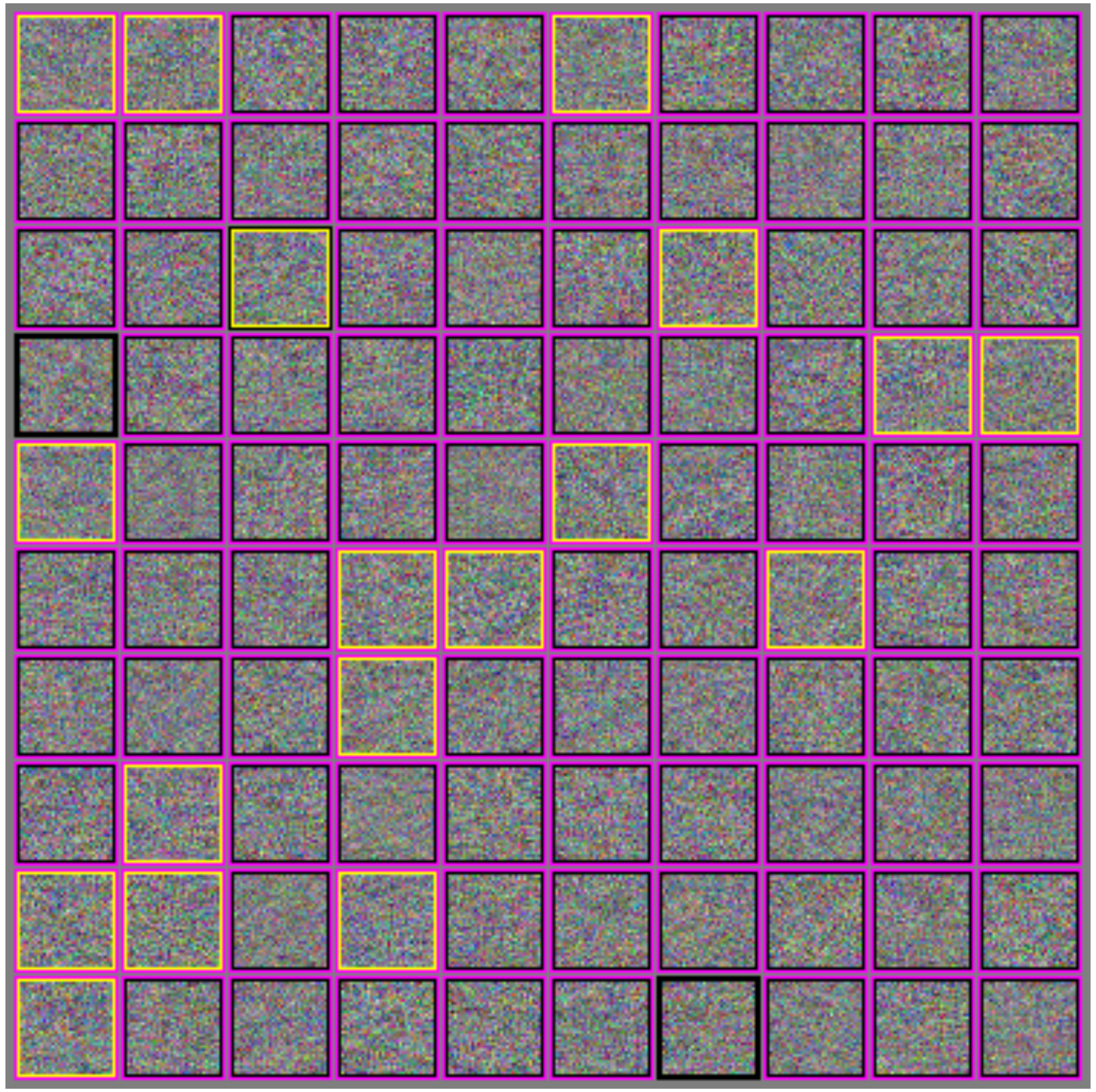
(Goodfellow 2018)

Attacking a linear model



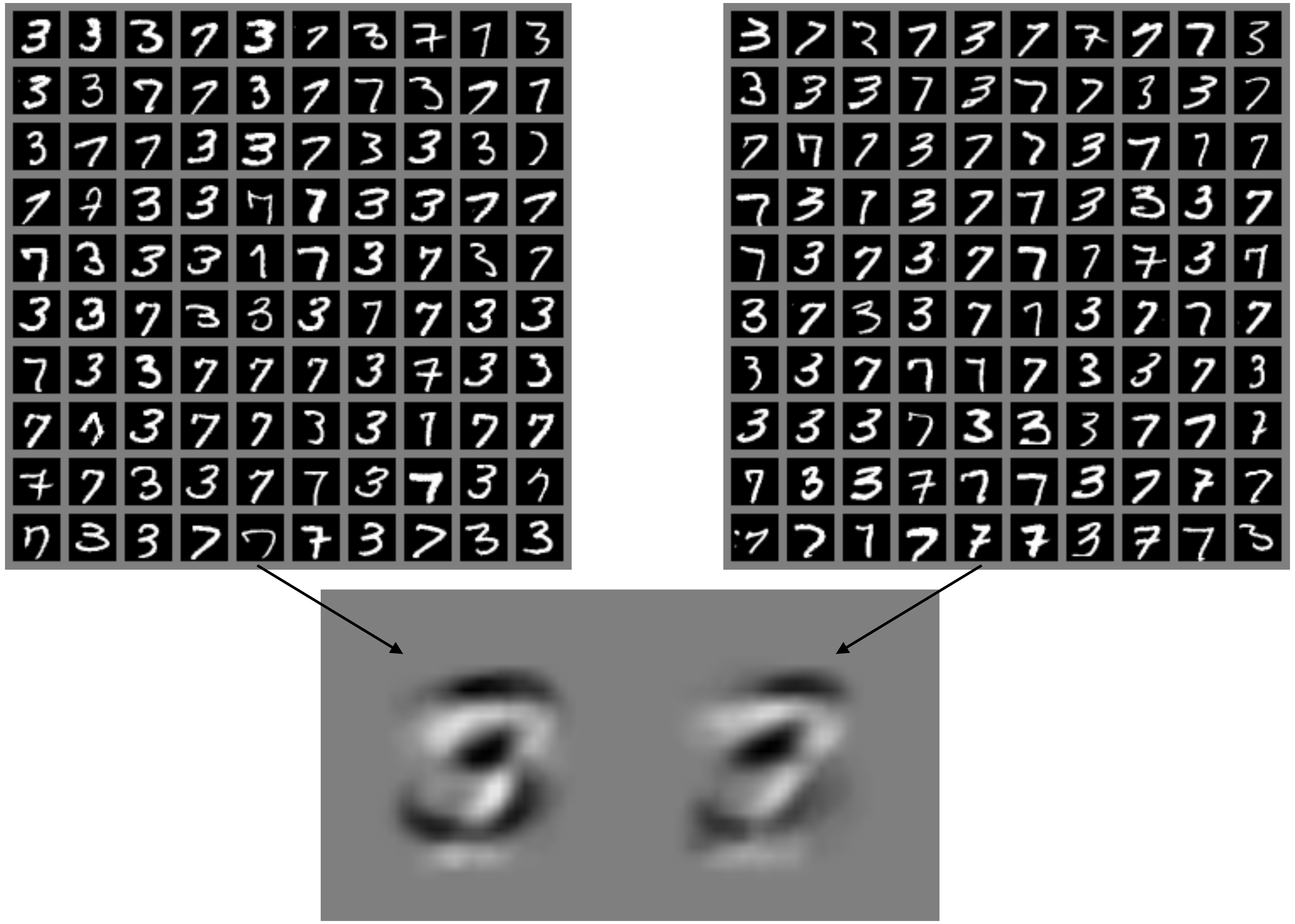
(Goodfellow 2018)

Wrong almost everywhere



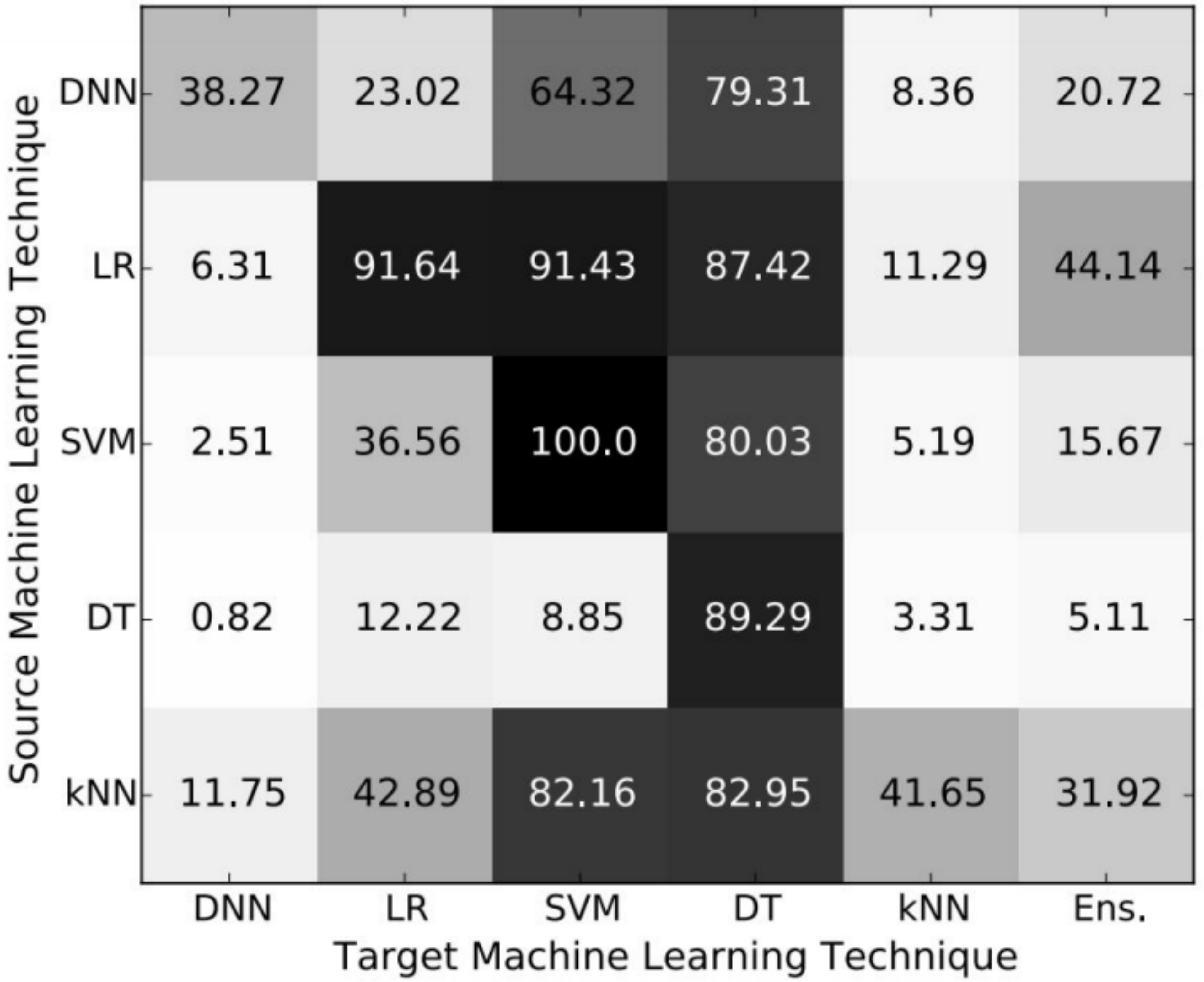
(Goodfellow 2018)

Cross-model, cross-dataset transfer



(Goodfellow 2018)

Transfer across learning algorithms



(Papernot 2016)

Transfer 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 2018)

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)

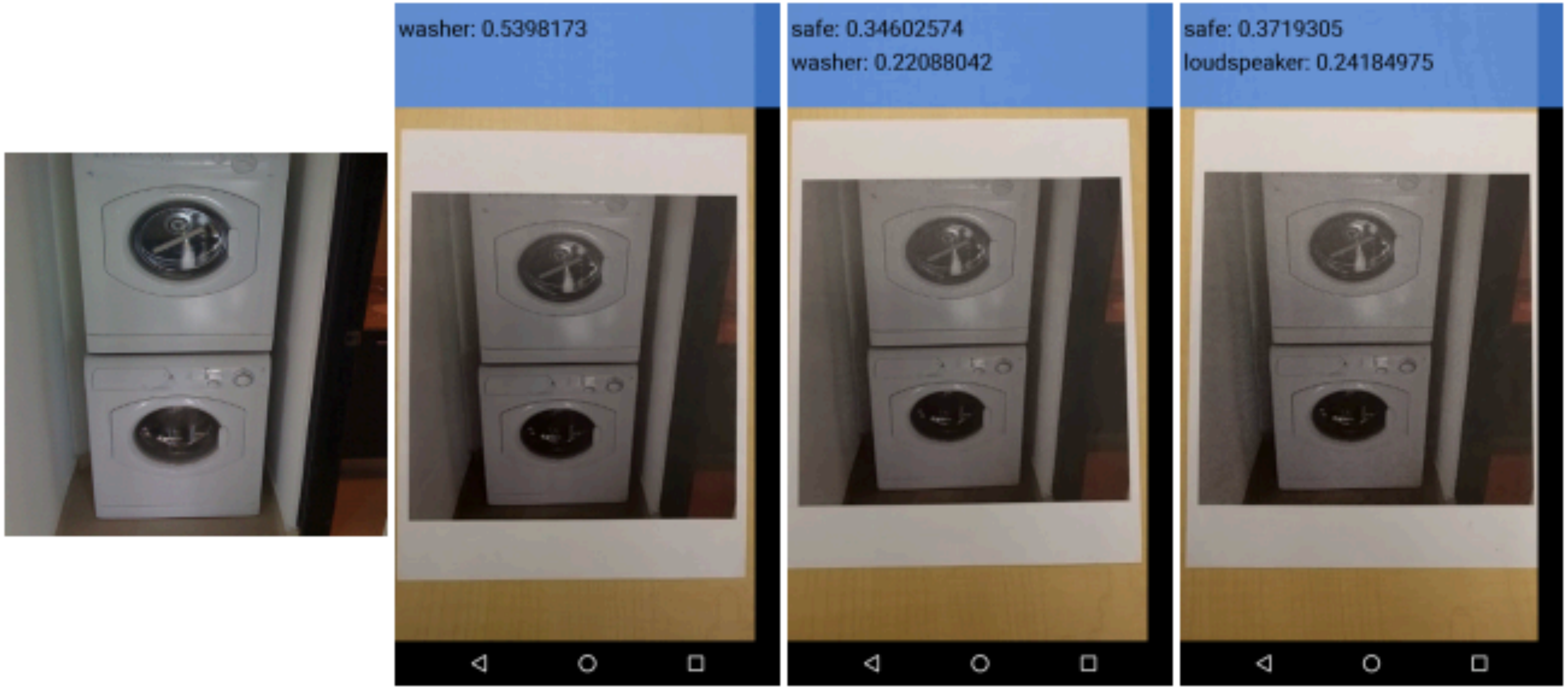
(Goodfellow 2018)

Transfer to the Human Brain

(Elsayed et al, 2018)

(Goodfellow 2018)

Transfer to the Physical World

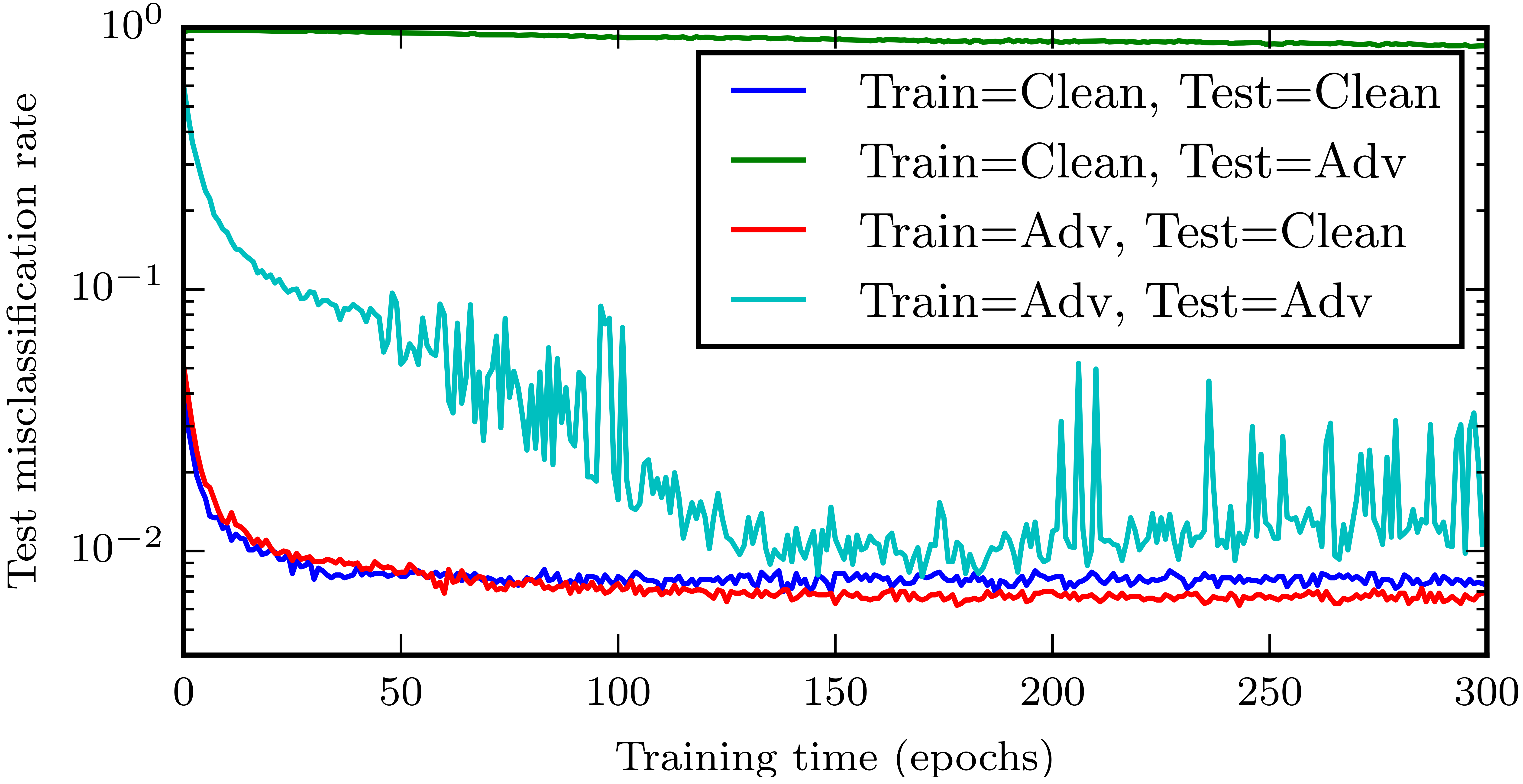


(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

(Kurakin et al, 2016)

(Goodfellow 2018)

Adversarial Training



(Goodfellow 2018)

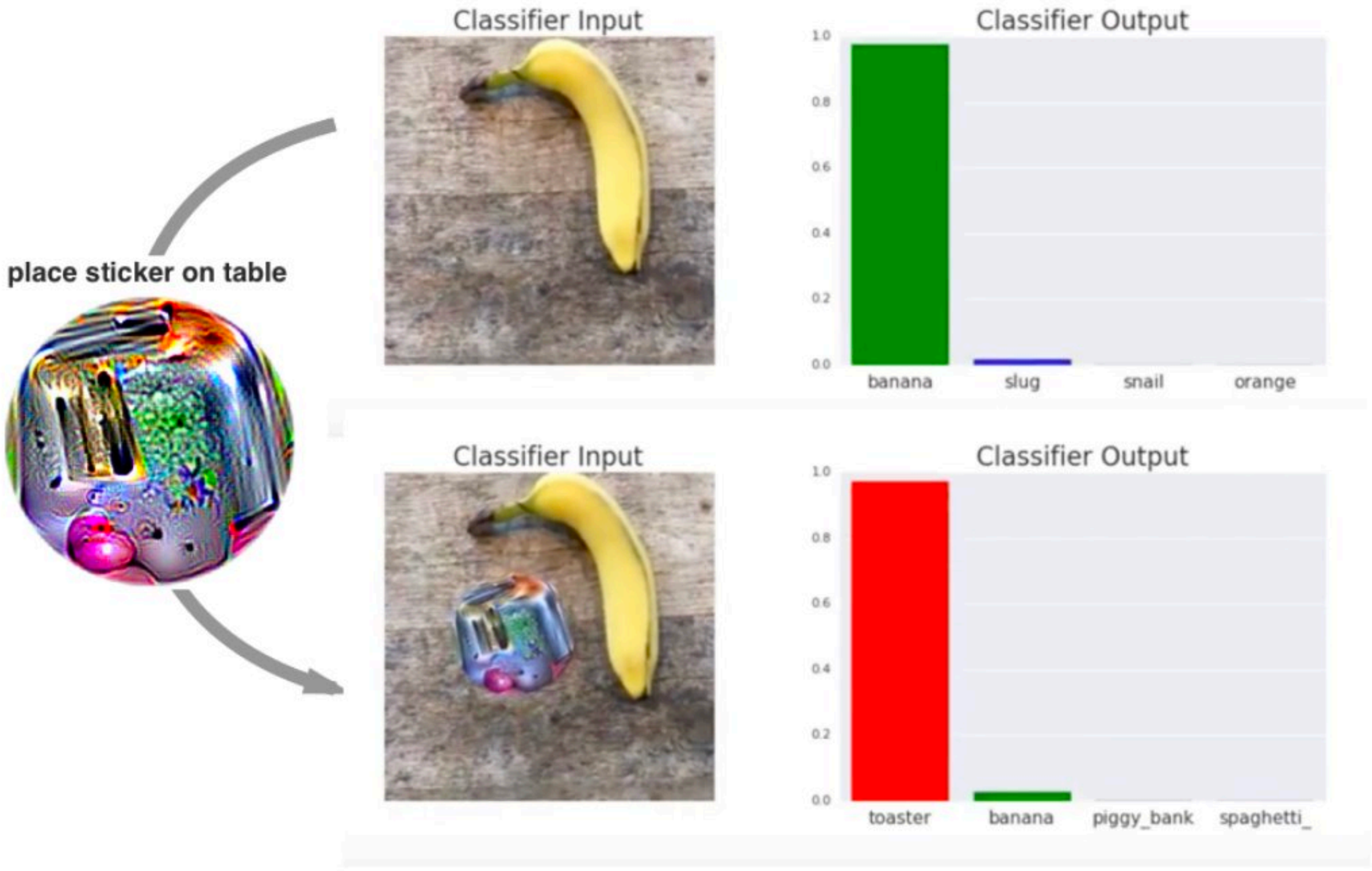
Adversarial Training vs Certified Defenses

- Adversarial Training:
  - Train on adversarial examples
  - This *minimizes a lower bound* on the true worst-case error
  - Achieves a high amount of (empirically tested) robustness on small to medium datasets
- Certified defenses
  - Minimize an *upper bound* on true worst-case error
  - Robustness is guaranteed, but amount of robustness is small
  - Verification of models that weren't trained to be easy to verify is hard

(Goodfellow 2018)

Limitations of defenses

- Even certified defenses so far assume unrealistic threat model
  - Typical model: attacker can change input within some norm ball
- Real attacks will be stranger, hard to characterize ahead of time



(Brown et al., 2017)

(Goodfellow 2018)

Clever Hans

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

(Goodfellow 2018)

Get involved!

https://github.com/tensorflow/cleverhans

cleverhans

(Goodfellow 2018)

Apply What You Have Learned

- Publishing an ML model or a prediction API?
  - Is the training data sensitive? -> train with differential privacy
- Consider how an attacker could cause damage by fooling your model
  - Current defenses are not practical
  - Rely on situations with no incentive to cause harm / limited amount of potential harm

(Goodfellow 2018)