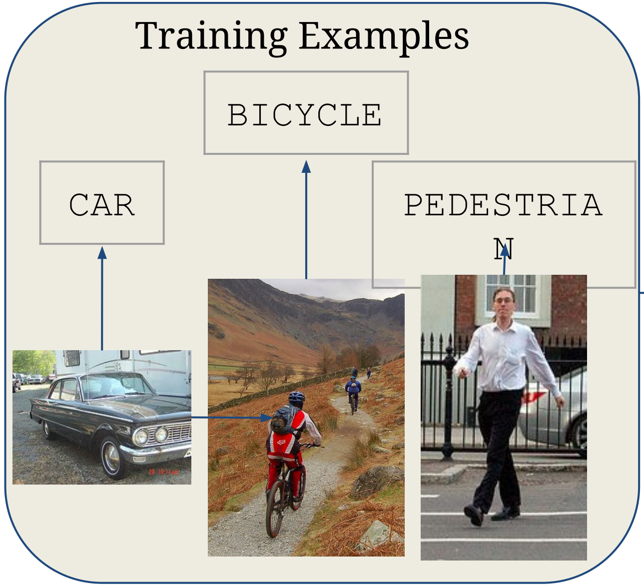
Physical Adversarial Examples

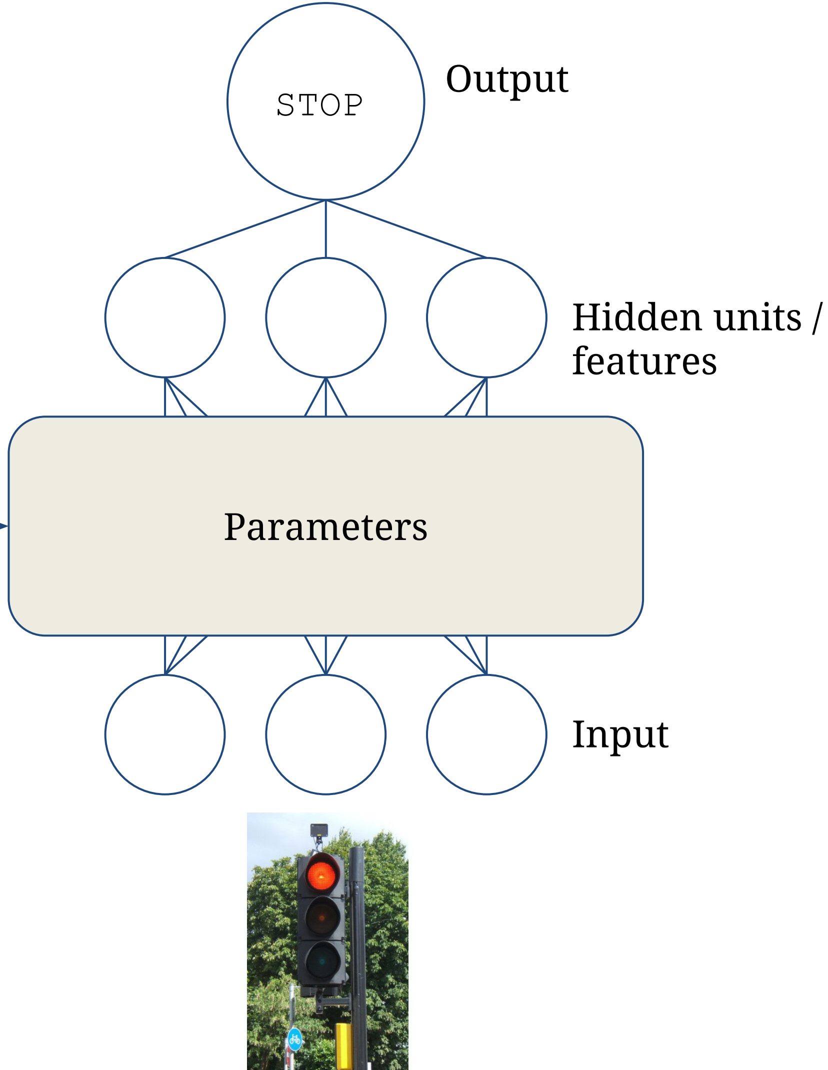
Alex Kurakin Ian Goodfellow

Google OpenAI

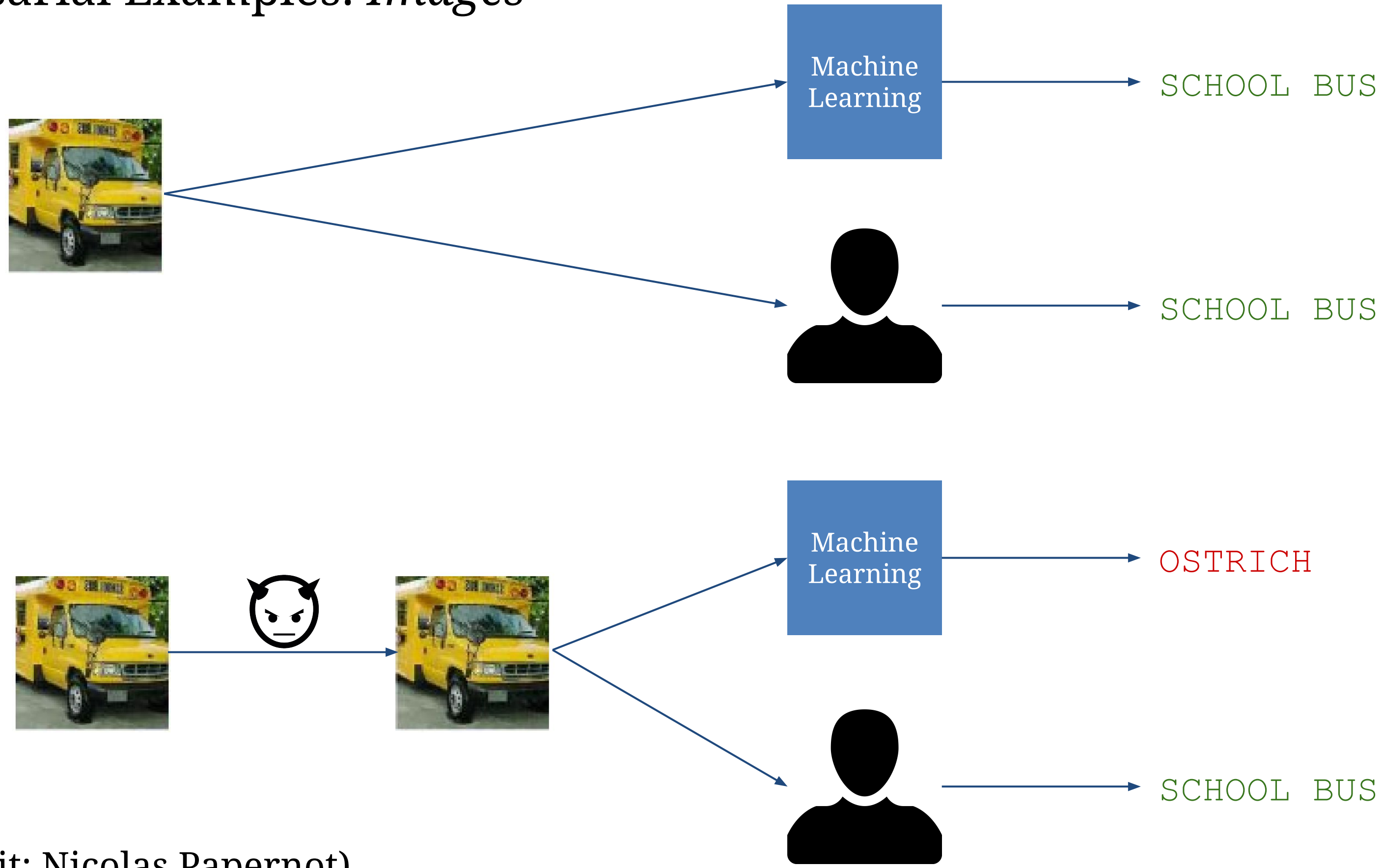
Machine Learning



ImageNet (Russakovsky et al 2015)



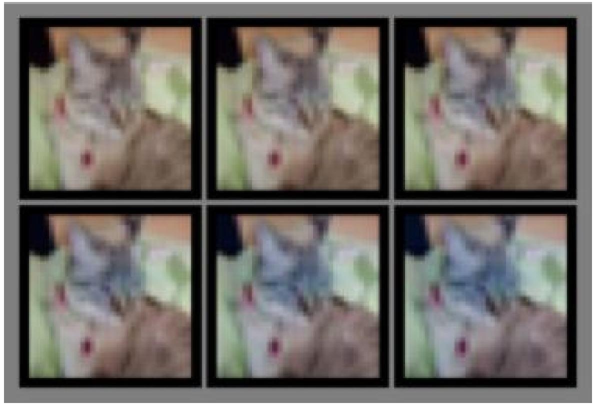
Adversarial Examples: *Images*



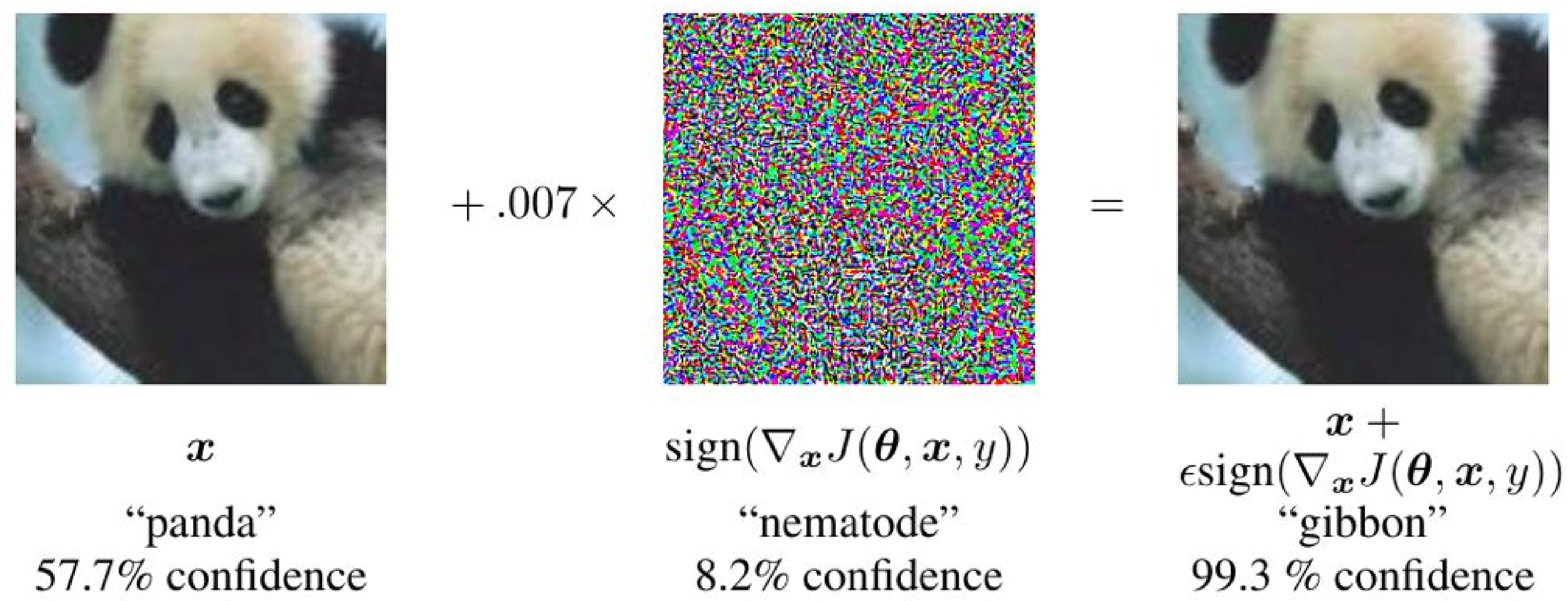
(Figure credit: Nicolas Papernot)

Turning Objects into "Airplanes"

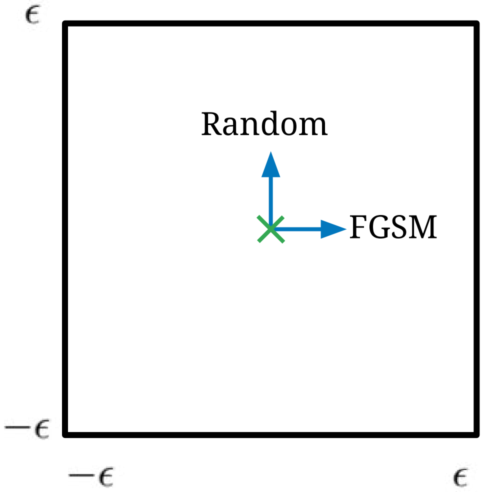


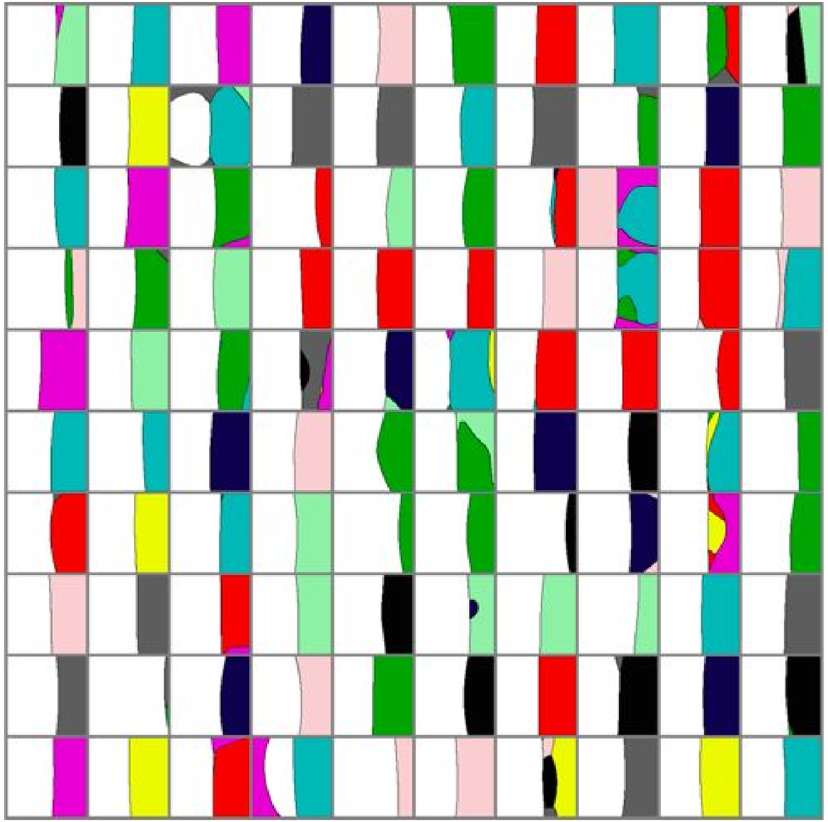



Fast Gradient Sign Method (FGSM)

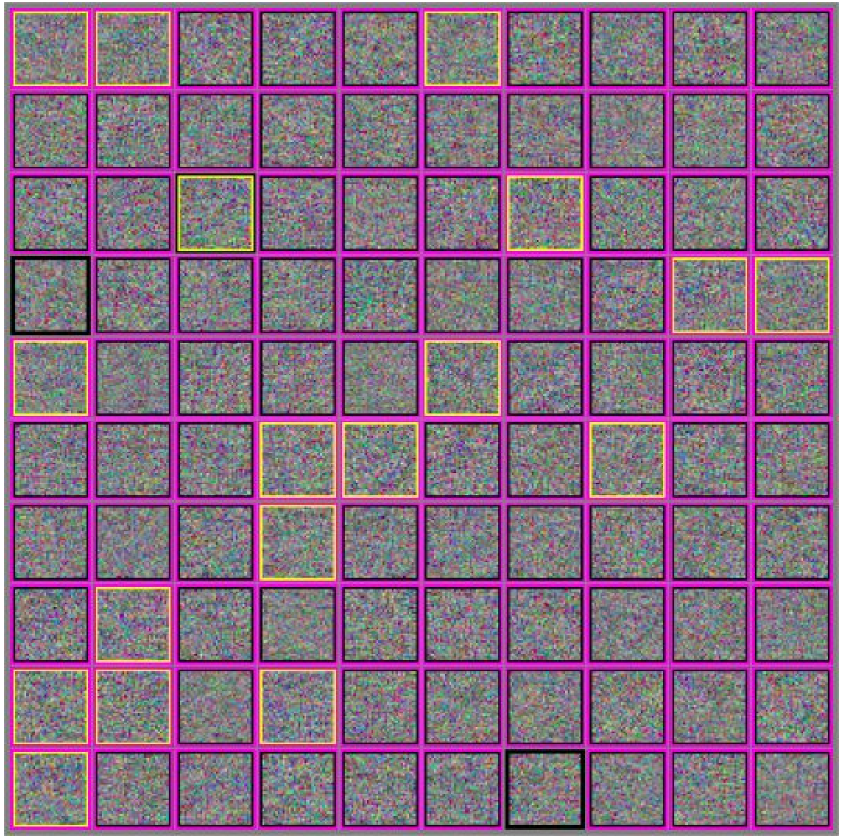


Maps of Adversarial Examples

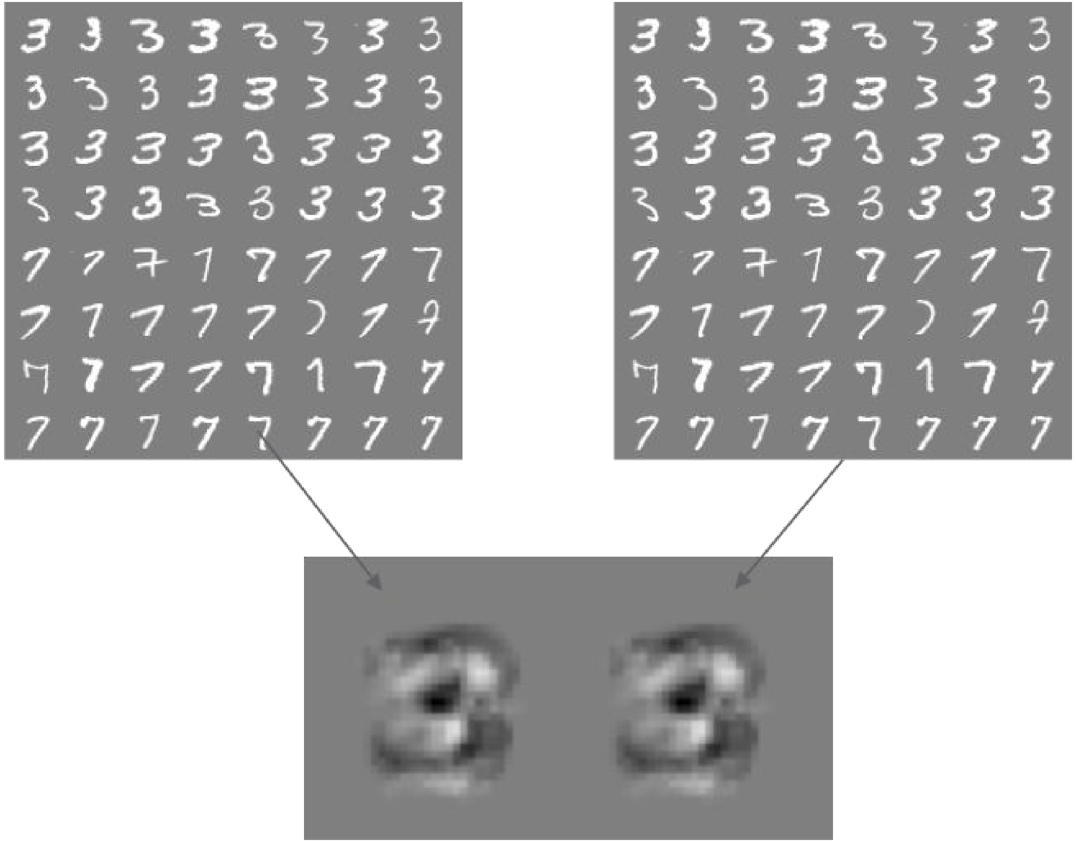




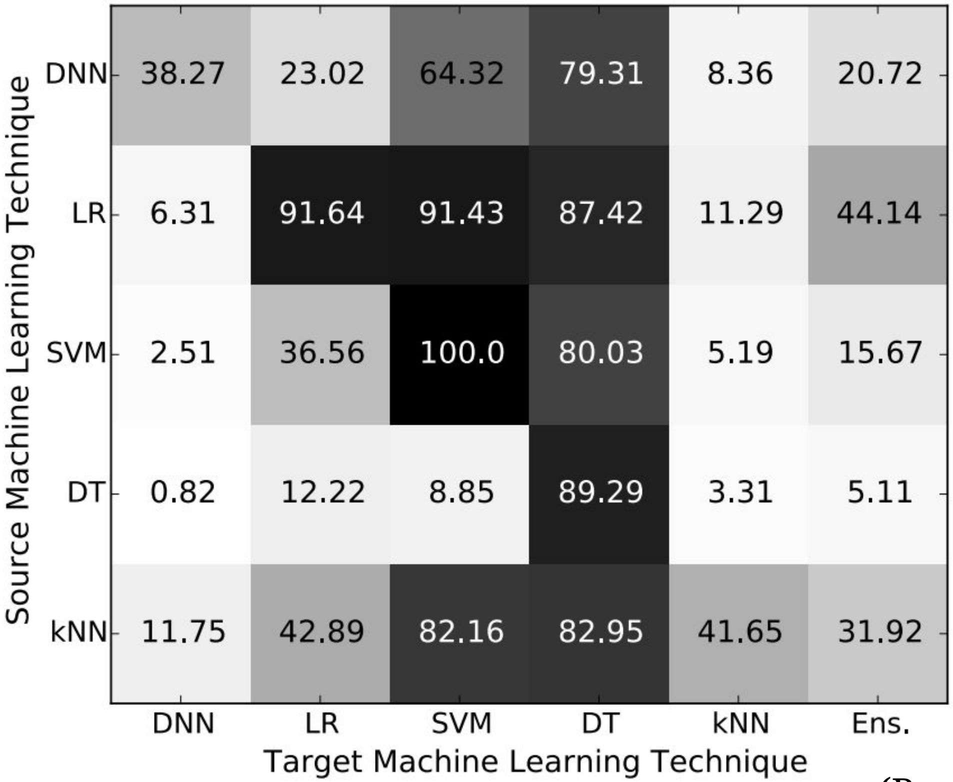
Almost all inputs are misclassified



Generalization across training sets

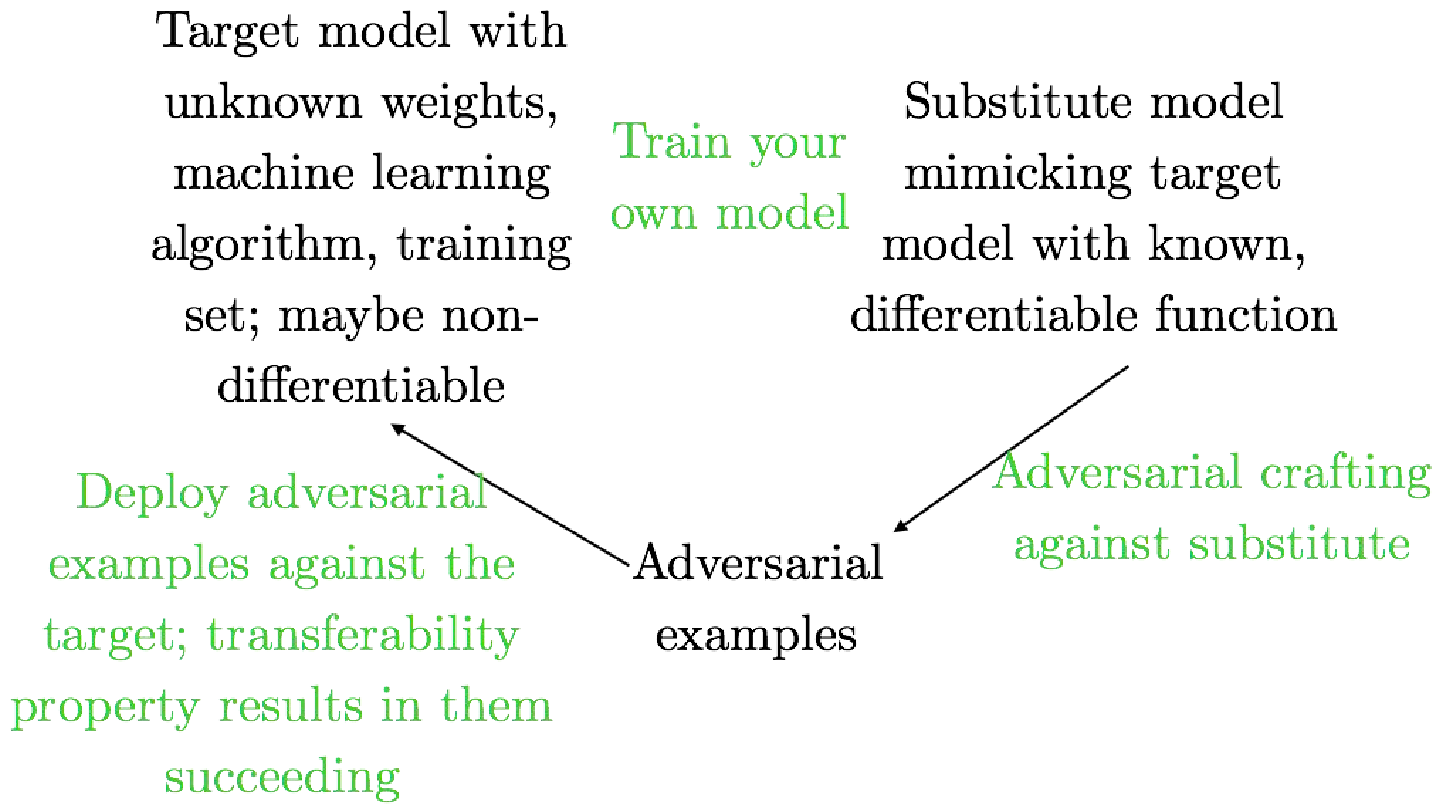


Cross-Technique Transferability



(Papernot et al 2016)

Transferability attack



Results on Real-World Remote Systems

All remote classifiers are trained on the MNIST dataset (10 classes, 60,000 training samples)

Remote Platform	ML technique	Number of queries	Adversarial examples misclassified (after querying)
MetaMind	Deep Learning	6,400	84.24%
amazon web services™	Linear Regression	800	96.19%
Google Cloud Platform	Unknown	2,000	97.72%

(Papernot et al 2016)

Adversarial examples in the physical world?

- Question: Can we build adversarial examples in the physical world?
- Let's try the following:
    - Generate and print picture of adversarial example
    - Take a photo of this picture (with cellphone camera)
    - Crop+warp picture from the photo to make it 299x299 input to Imagenet inception
    - Classify this image
- Would the adversarial image remain misclassified after this transformation?
- If we succeed with "photo" then we potentially can alter real-world objects to mislead deep-net classifiers

Adversarial examples in the physical world?

- Question: Can we build adversarial examples in the physical world?

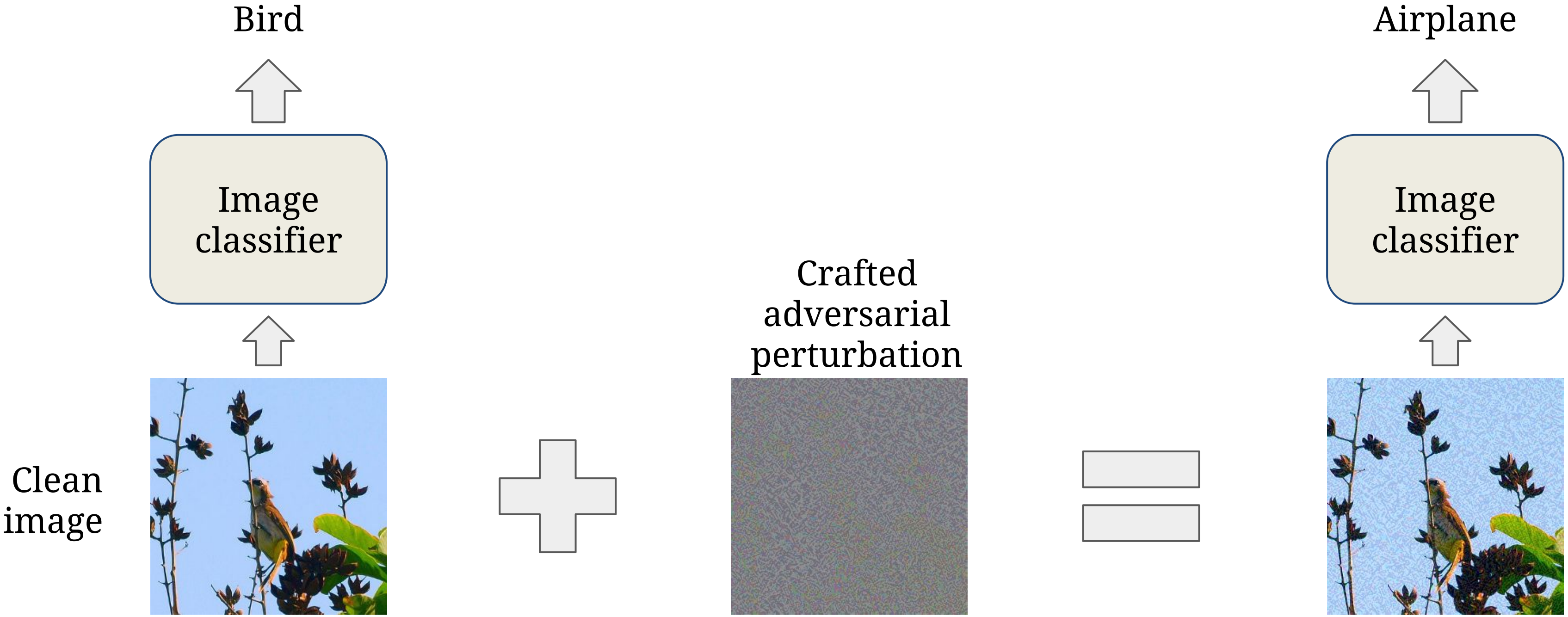
- Let’s try the following:
  - Generate and print picture of adversarial example
  - Take a photo of this picture (with cellphone camera)
  - Crop+warp picture from the photo to make it 299x299 input to Imagenet inception
  - Classify this image

- Would the adversarial image remain misclassified after this transformation?

- If we succeed with “photo” then we potentially can alter real-world objects to mislead deep-net classifiers

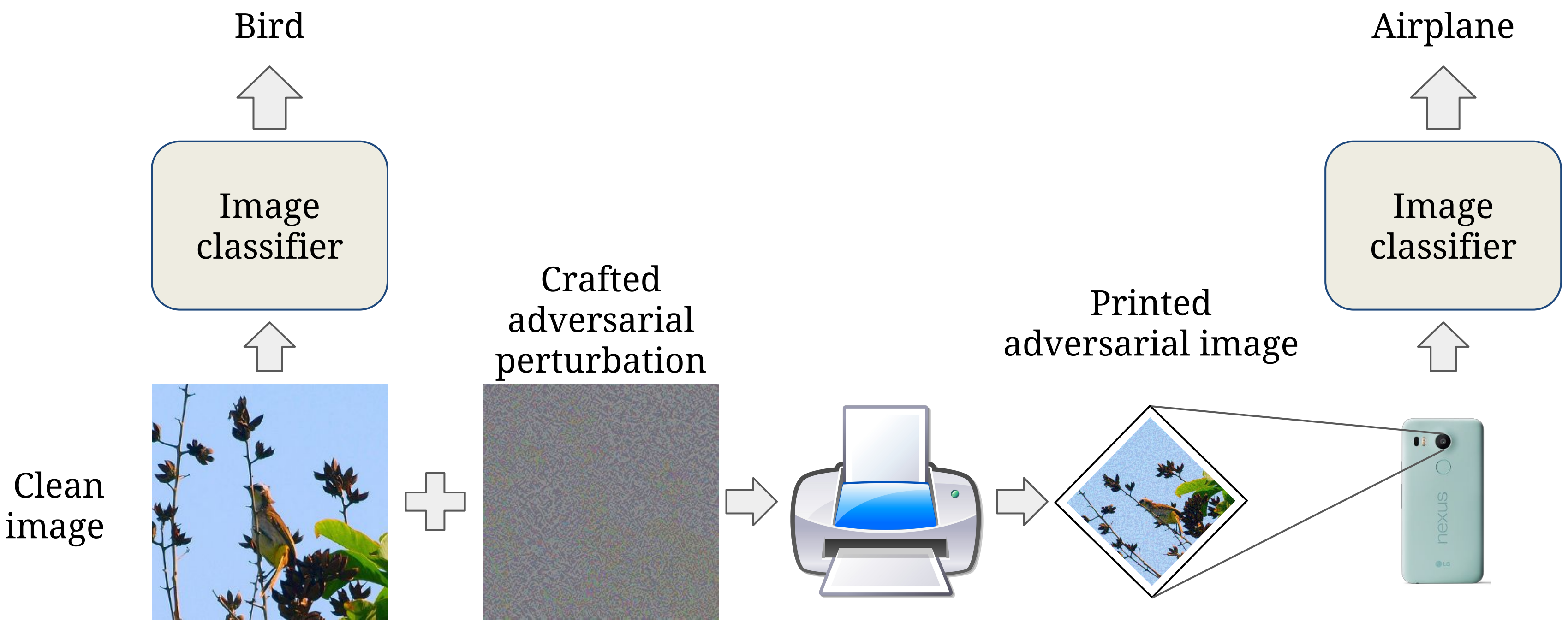
Answer: IT’S POSSIBLE

Digital adversarial examples



[Goodfellow, Shlens & Szegedy, ICLR2015]

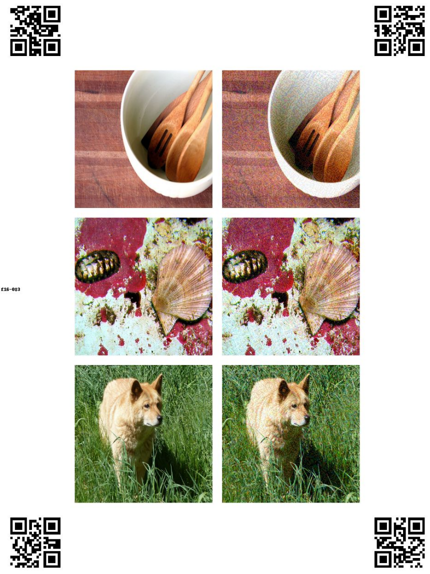
Adversarial examples in the physical world



[Kurakin & Goodfellow & Bengio, arxiv.org/abs/1607.02533]

Our experiment

1. Print pairs of normal and adversarial images



2. Take picture

3. Auto crop and classify

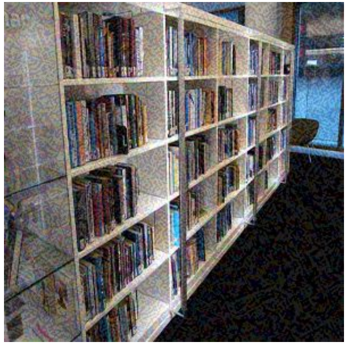
Up to 87% of images could remain misclassified!

Live demo

Library

Washer

Washer



Don’t panic! It’s not end of the ML world!

- Our experiment is a proof-of-concept set up:
  - We had full access to the model
  - 87% adversarial images rate is for only one method, which could be resisted by adversarial training. For other methods it’s much lower.
  - In many cases “adversarial” image is not so harmful: one breed of dog confused with another

- In practice:
  - Attacker doesn’t have access to model
  - You might be able to use adversarial training to defend model against some attacks
  - For other attacks, “adversarial examples in the real worlds” won’t work that well
  - It’s REALLY hard to fool your model to predict specific class

THANKS

1024