GeekPwn 极棒

GEEKPWN CYBERSECURITY CONTEST
U.S. 2018 LAUNCH

COMPUTER HISTORY MUSEUM
MOUNTAIN VIEW, CALIFORNIA
NOV.13th , 2017

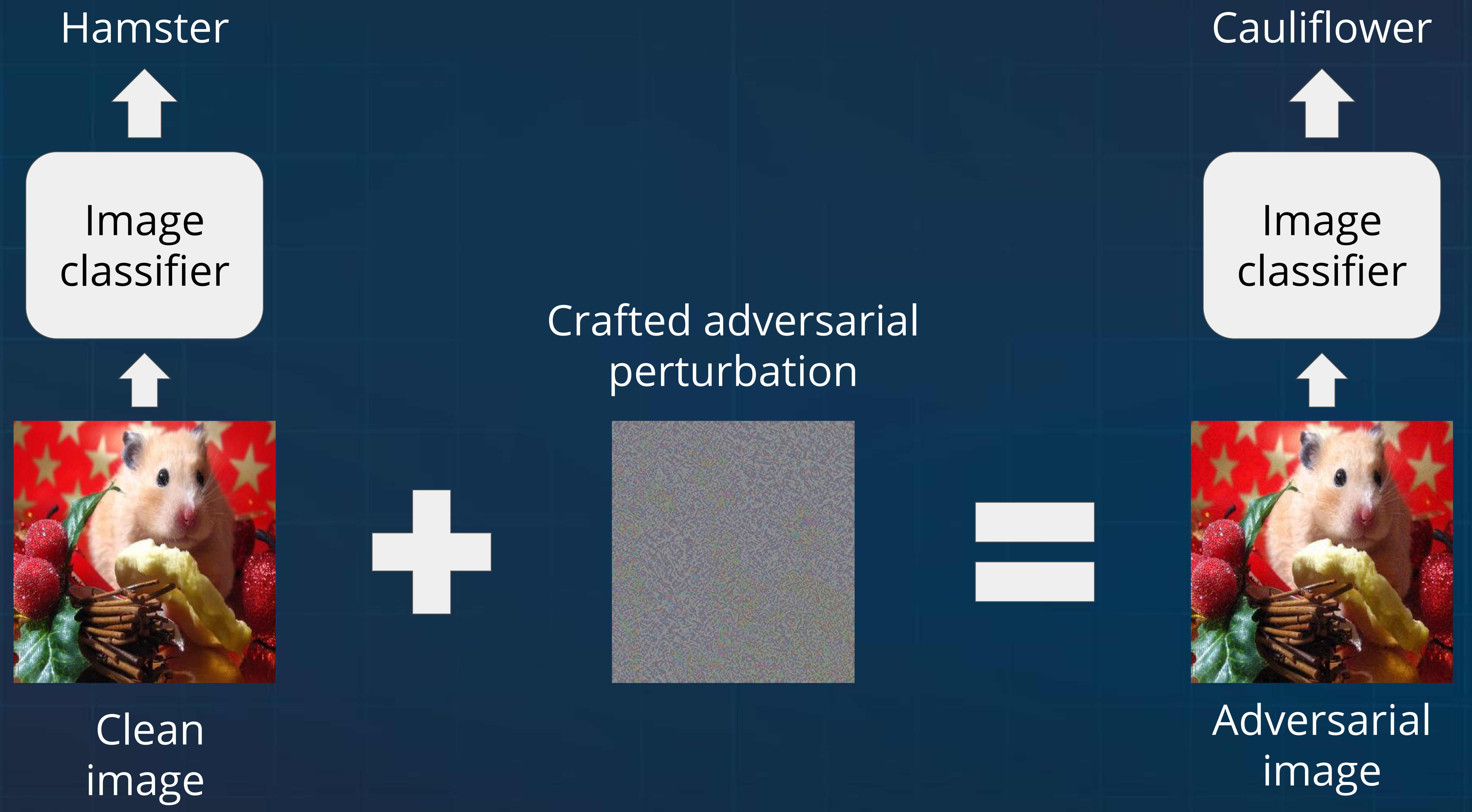
The recent advancement of adversarial machine learning

Alexey Kurakin

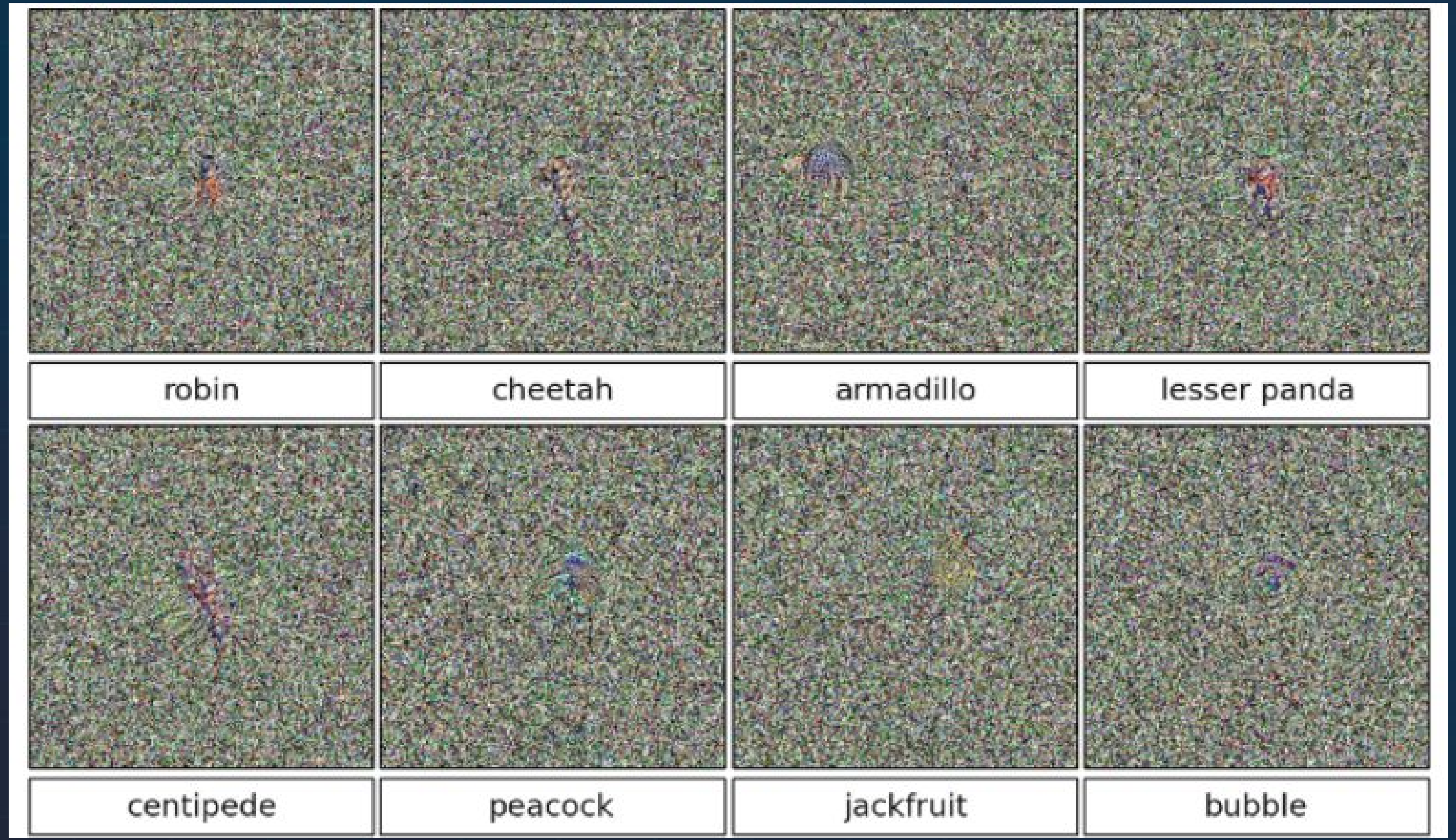
The Google Brain Team

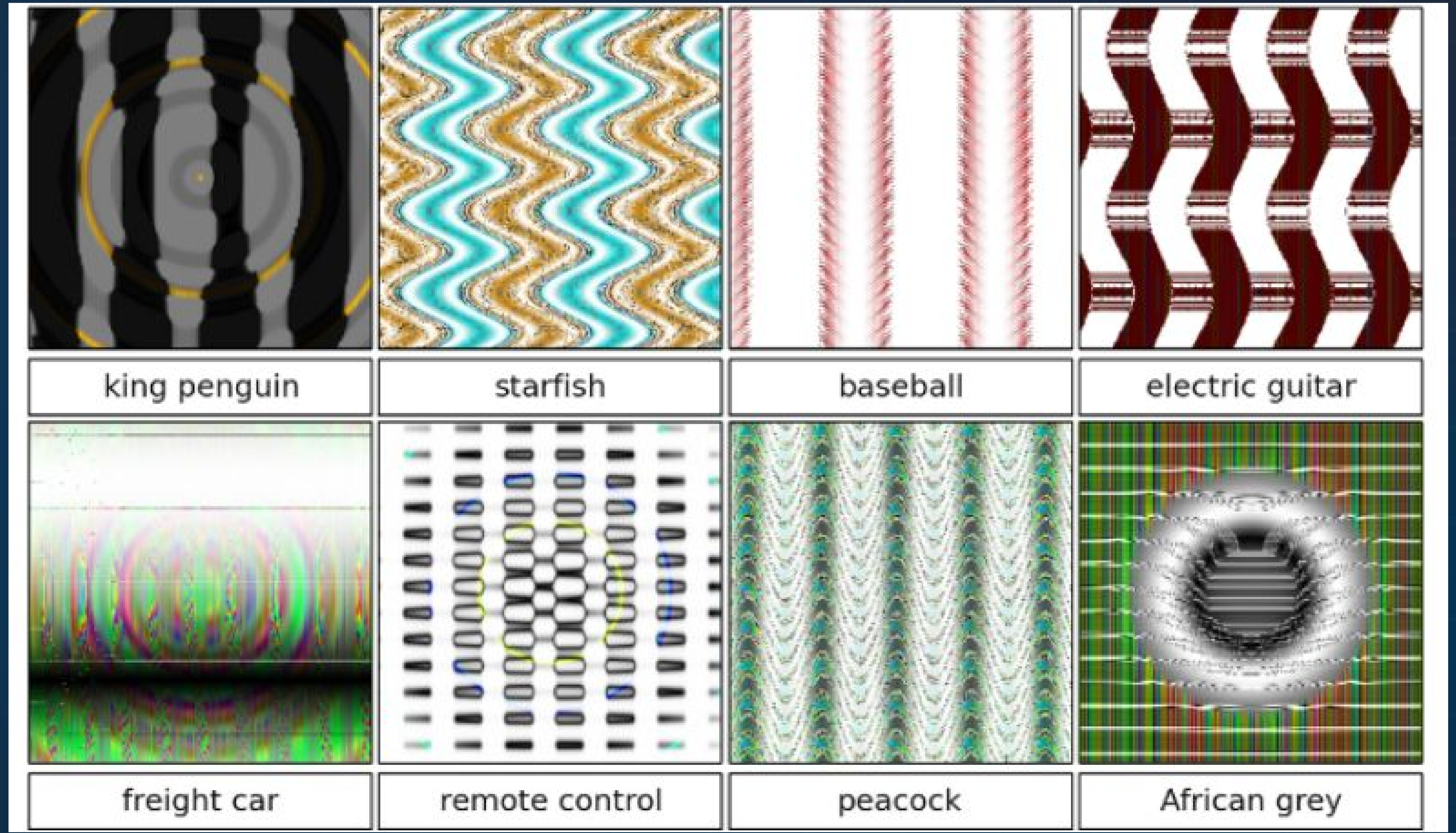
What are adversarial examples?
Why they are important?

What are adversarial examples?



Related problem - garbage class examples





(Nguyen, Yosinski, Clune, 2015)

Why adversarial examples are important?

- Small changes in input -> change of classification label
  - Lack of robustness of machine learning
- Security problem
  - You can fool computer, human won't notice a problem

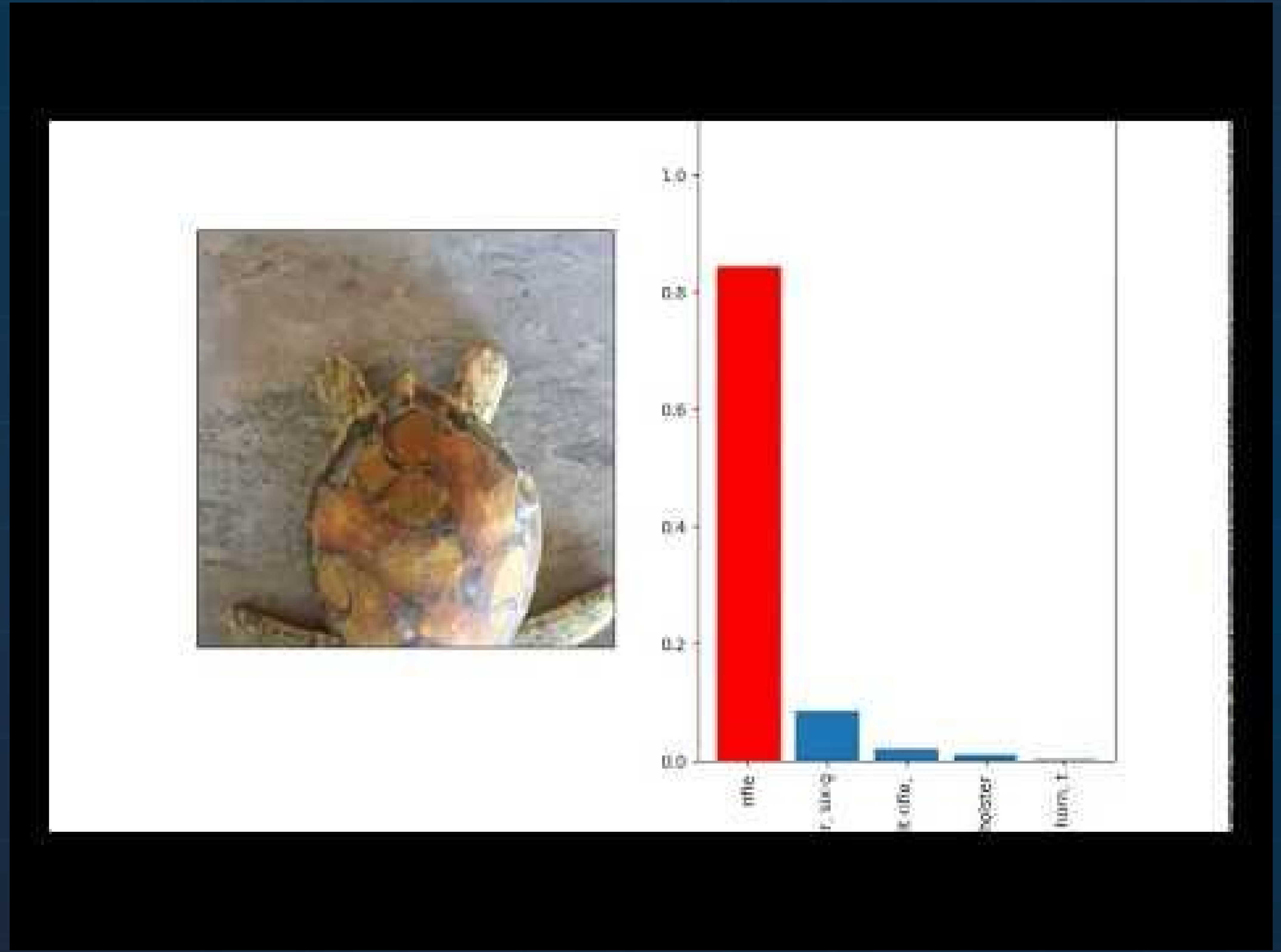
Clean image

Adversarial image

Why adversarial examples are important?

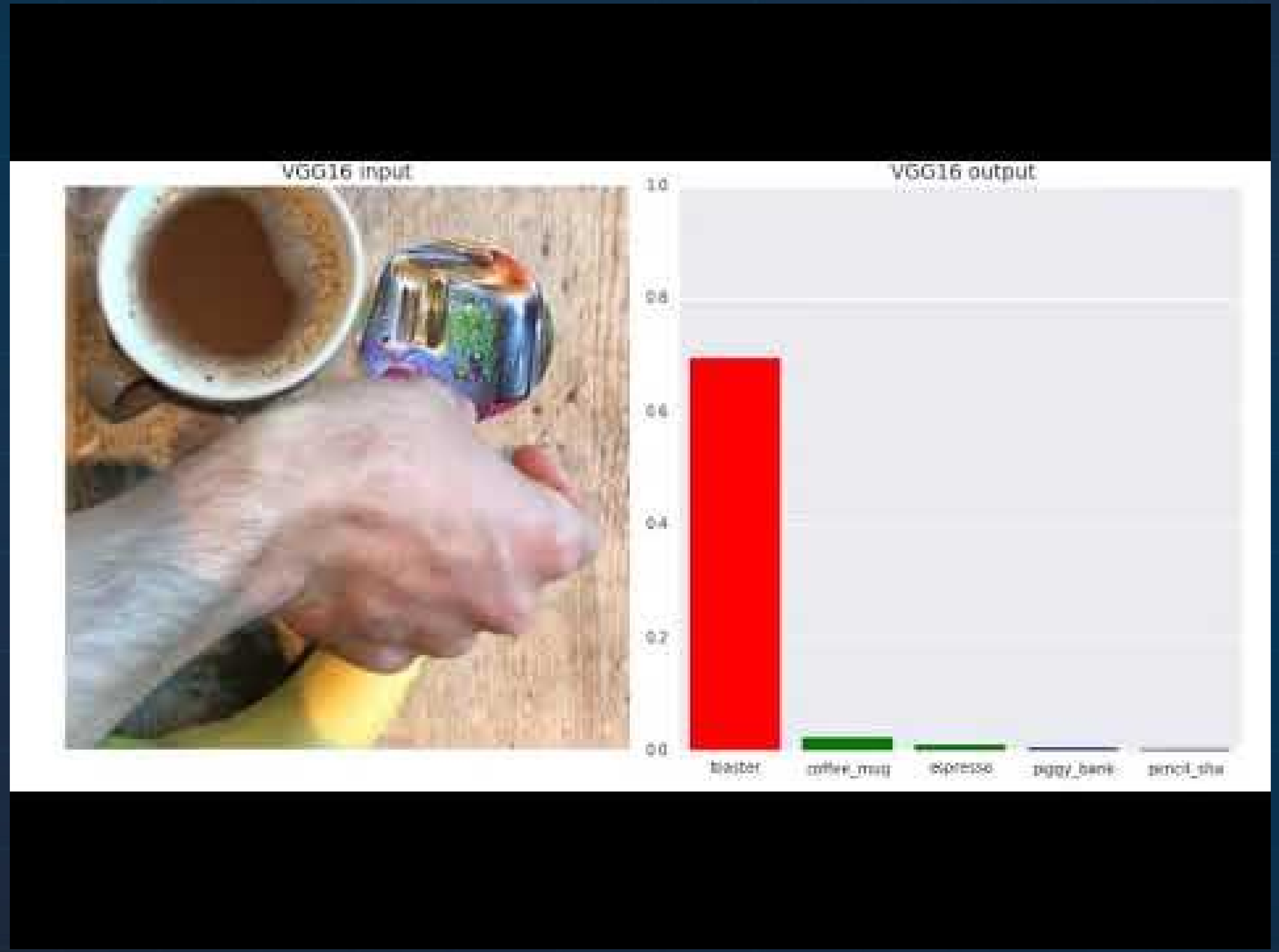
(Kurakin, Goodfellow, Bengio, 2016)

Why adversarial examples are important?



(Athalye, Engstrom, Ilyas, Kwok, 2017)

Why adversarial examples are important?



(Brown, Mane, Roy, Abadi, Gilmer, 2017)

Adversarial attacks
How to generate adversarial examples

White box VS black box

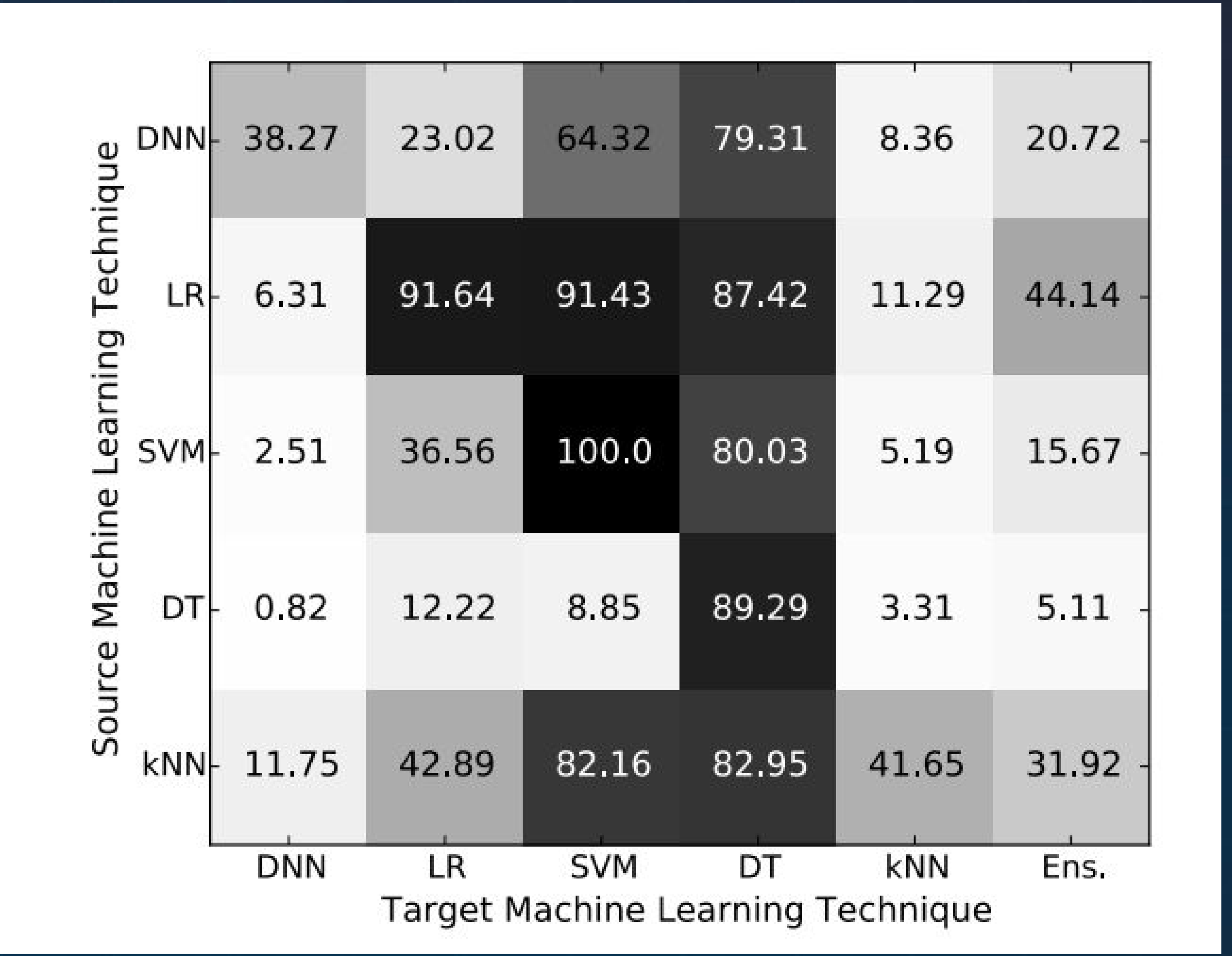
White box case:

- Everything is known about classifier (including architecture, model parameters)

Black box case:

- Model parameters are unknown

  In such case craft adversarial examples for known model and transfer to unknown



(Papernot, McDaniel, Goodfellow, 2016)

Digital VS Physical

Digital attack:

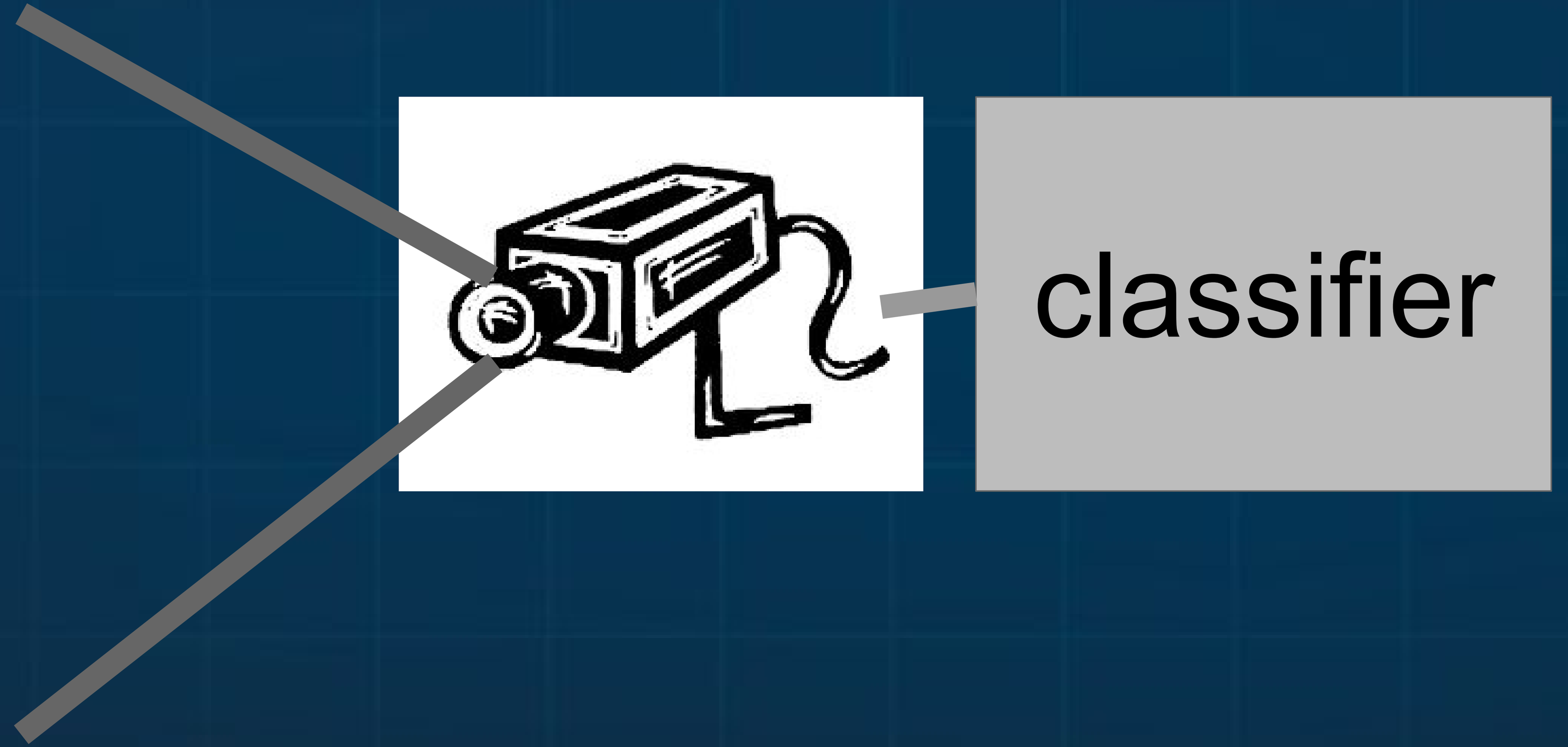
- Directly feeding numbers into classifier

```
img = plt.imread('adversarial_image.png')

with tf.Session() as sess:
  inp = tf.placeholder(tf.float32, shape=[1, height, width, 3])
  logits = model(inp)
  prediction = tf.argmax(logits, 1)
  prediction_value = sess.run(prediction, feed_dict={inp: img})
```

Physical attack:

- Classifier perceives world through sensor (e.g. camera)



White box adversarial attack - L-BFGS

$$minimize\|r\|_2 s.t.$$

$$F(x+r) = L$$

$$(x+r) \in [0,1]^m$$

Where:

- F(•) - neural network or another ML classifier
- x - clean image
- L - desired target class
- r - adversarial perturbation, which makes x+r adversarial image

Method is very computationally demanding and slow

(Szegedy et al, 2014)

White box adversarial attack - FGSM

$$X_{adv} = X + \epsilon sign(\nabla_x J(X, Y_{true}))$$

Where:

- J(x,y) - cross entropy loss of prediction for input sample x and label y
- x - clean image with true label Y_{true}
- ϵ - method parameter, size of perturbation

(Goodfellow, Shlens, Szegedy, 2014)

White box adversarial attack - Iterative FGSM

$$X_0^{adv} = X, \quad X_{N+1}^{adv} = Clip(X_N^{adv} + \alpha sign(\nabla_X J(X_N^{adv}, Y_{true})))$$

Where:

- J(x,y) - cross entropy loss of prediction for input sample x and label y
- x - clean image with true label Y_{true}
- α - method parameter, size of one step

(Kurakin, Goodfellow, Bengio, 2016)

White box adversarial attack - Iterative FGSM

$$X_0^{adv} = X, \quad X_{N+1}^{adv} = Clip(X_N^{adv} - \alpha sign\left(\nabla_X J(X_N^{adv}, Y_{target})\right))$$

Where:

- J(x,y) - cross entropy loss of prediction for input sample x and label y
- x - clean image, Y_{target} - desired target class
- α - method parameter, size of one step

(Kurakin, Goodfellow, Bengio, 2016)

White box adversarial attack - C&W

$$\|X_{adv} - X\| + c\Big(\max_{i \neq t}\big(Logits(X_{adv})_i\big) - Logits(X_{adv})_t\Big)^{+} \rightarrow min$$

Where:

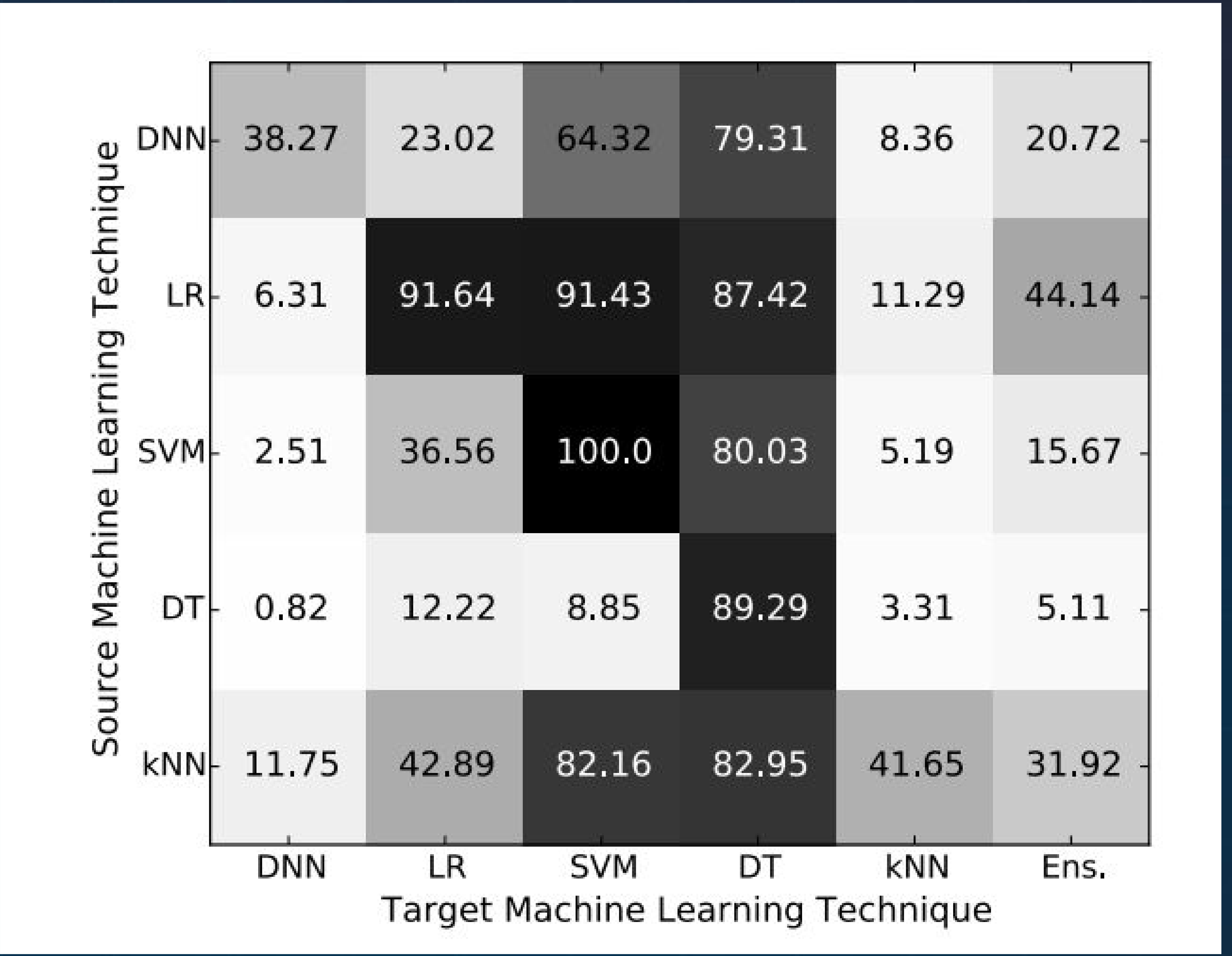
- Logits(•) - logits of the network
- X - clean image, X_{adv} - adversarial image
- t - desired target class
- $(z)^{+} = \max(0, z)$

Considered **<u>one of the strongest</u>** white box attacks.

(Carlini, Wagner, 2016)

Black box attacks

1. Train your own classifier on similar problem
2. Construct adversarial examples for your classifiers
3. Use them to attack unknown classifier



(Papernot, McDaniel, Goodfellow, 2016)

Defenses against adversarial examples

Defenses - input preprocessing

Transformation

(ex.: blur)

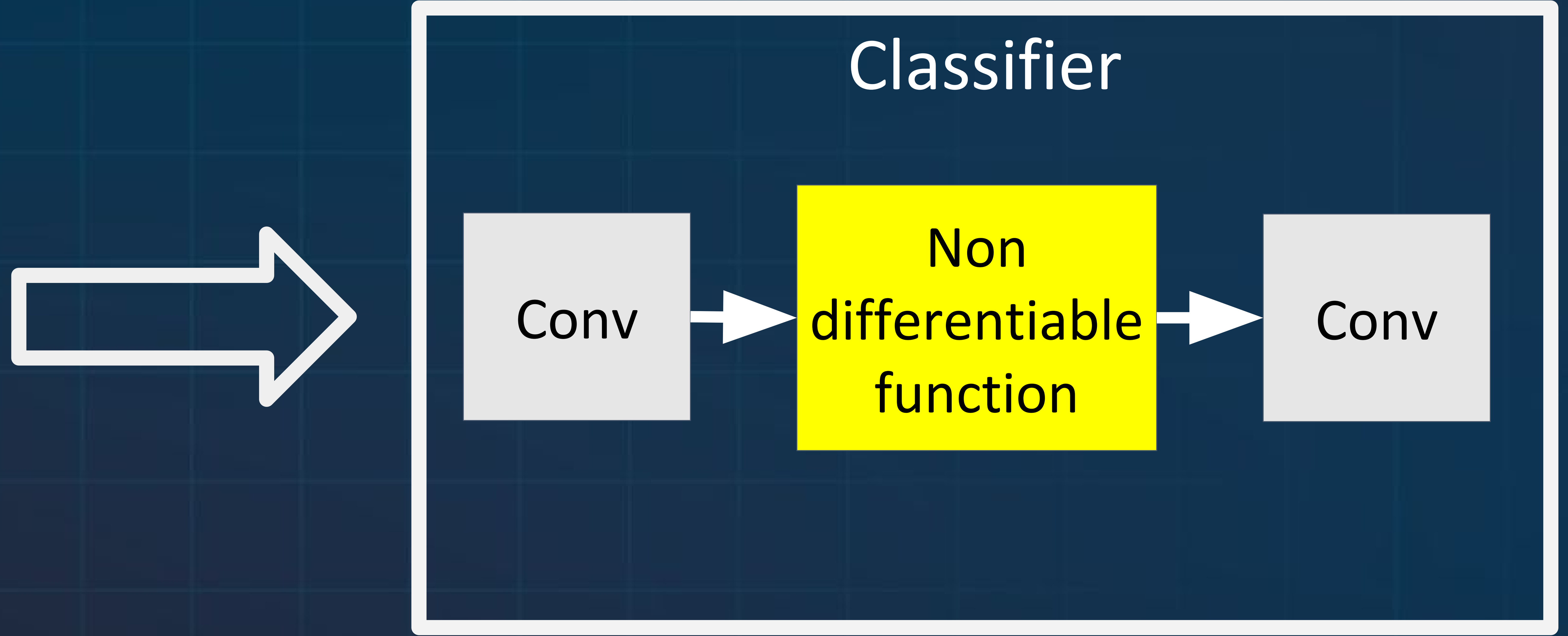
Classifier

Problems:

- May degrade quality on clean images
- Broken when attacker is aware of transformation

(multiple work by multiple authors)

Defenses - gradient masking



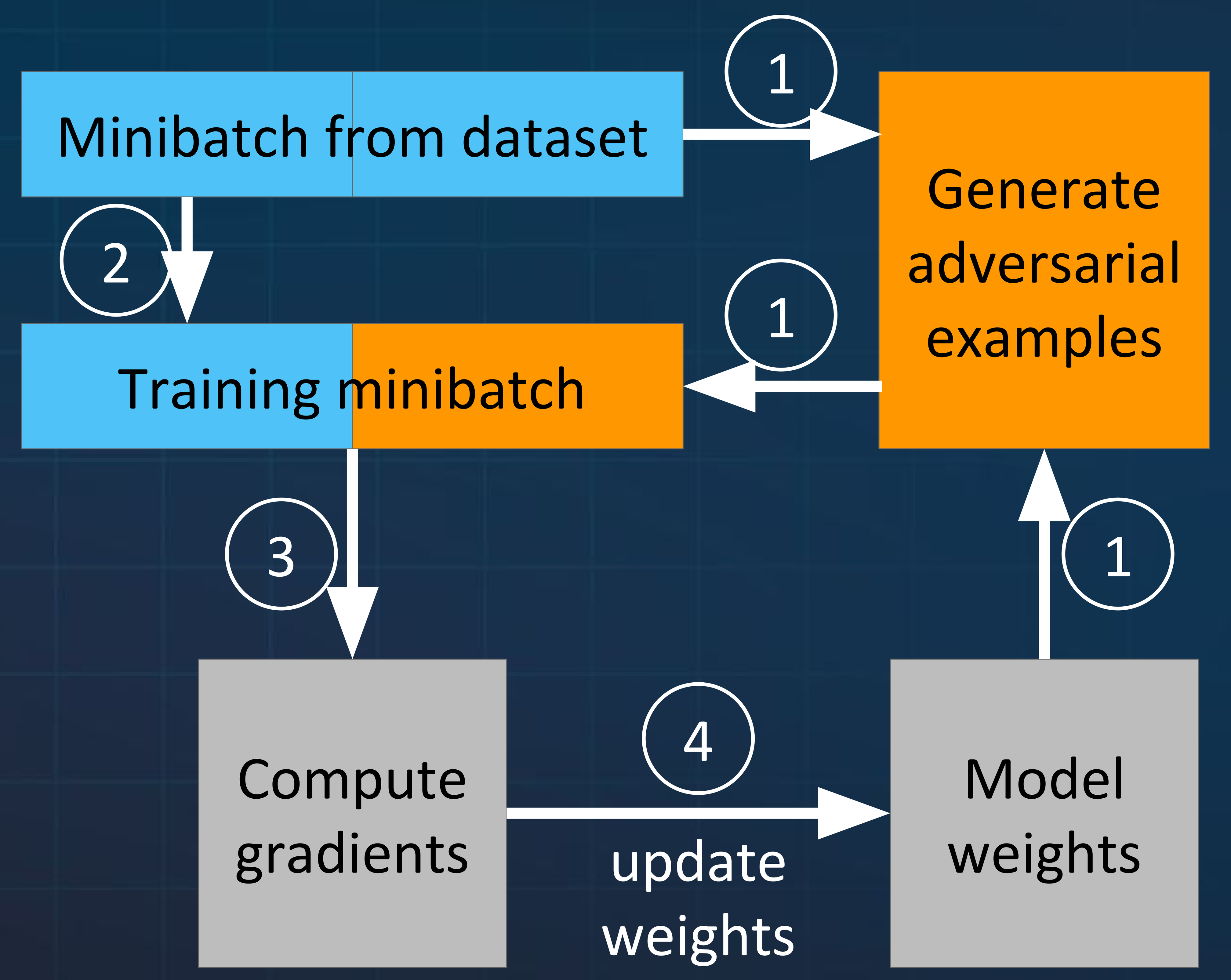
Similar to previous, but with non differentiable transformation

Problem:

- Can use transferability to attack the model

let's make this impossible → $X_{adv} = X + \epsilon sign(\nabla_x J(X, Y_{true}))$

Defenses - adversarial training



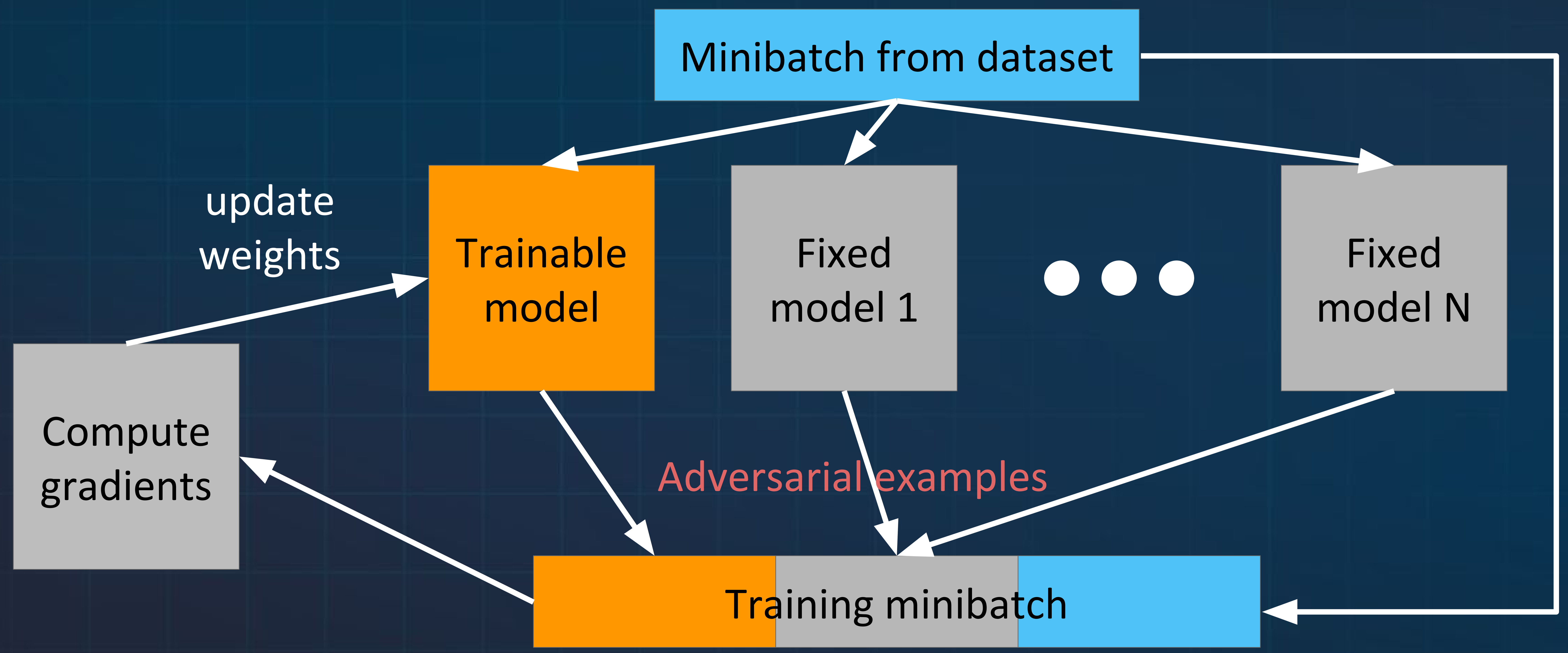
Idea: let's inject adversarial examples into training set

1. Compute adversarial examples using current model weights
2. Compose mixed minibatch of clean and adversarial examples
3. Use mixed minibatch to compute model gradients
4. Update model weights

Problems:

- may be harder to train
- model may learn to mask gradients

Defenses - ensemble adversarial training



To solve “gradient masking” problem of adversarial training let’s ensemble a few models.

(Tramer et al, 2017)

Adversarial competition

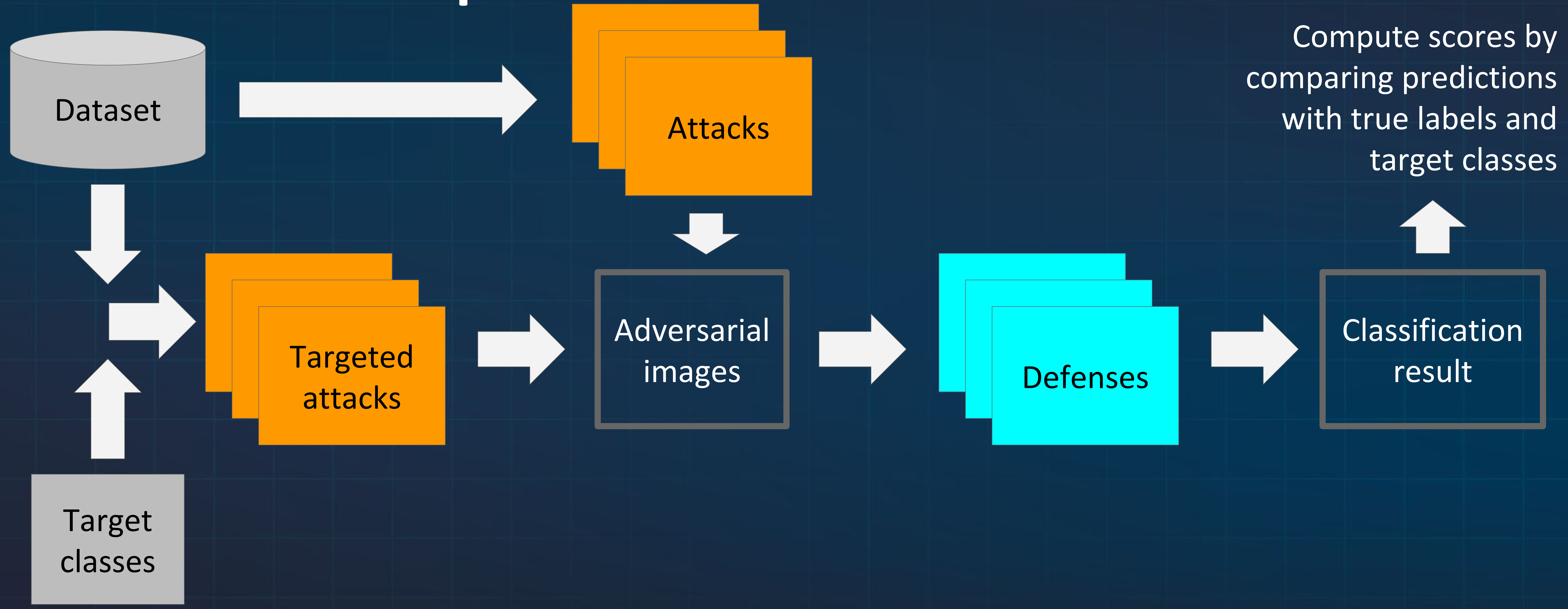
Adversarial competition

Three tracks / sub-competitions:

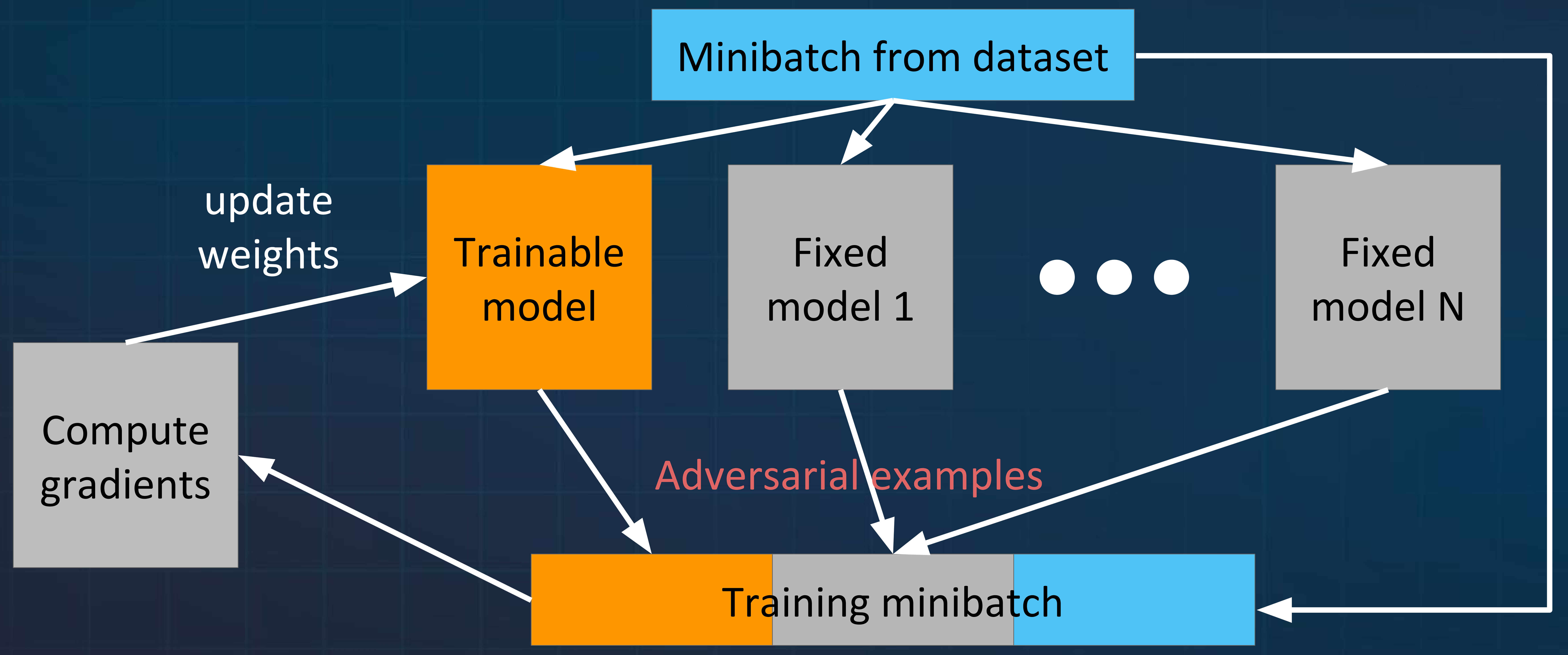
- Adversarial attacks - algorithms which try to confuse classifier
  - Input: image
  - Output: adversarial image
- Adversarial targeted attacks - algorithms which try to confuse classifiers in a very specific way
  - Input: image and target class
  - Output: adversarial image
- Adversarial defenses - classifiers which robust to adversarial examples
  - Input: image
  - Output: classification label

Attacks are not aware of defenses, so this is simulation of black box scenario. (Kurakin, Goodfellow, Bengio, 2017)

Adversarial competition



Adversarial competition - best defenses



- Top defenses are showing >90% accuracy on adversarial examples
- Most of the top defenses are using “ensemble adversarial training” as a part of the model.

Summary and conclusion

Summary

- Adversarial examples could be used to fool machine learning models
- Adversarial attacks:
  - White box VS black box
  - Digital VS physical
- Defenses against adversarial examples:
  - A lot of defenses works only if attacked does not know about them
  - One of potentially universal defenses - adversarial training
    However adversarial training does not work in all cases
- Adversarial competition
  - Showed that ensemble adversarial training could be pretty good defense against black box attack

Q & A