Lecture 21: Adversarial Networks

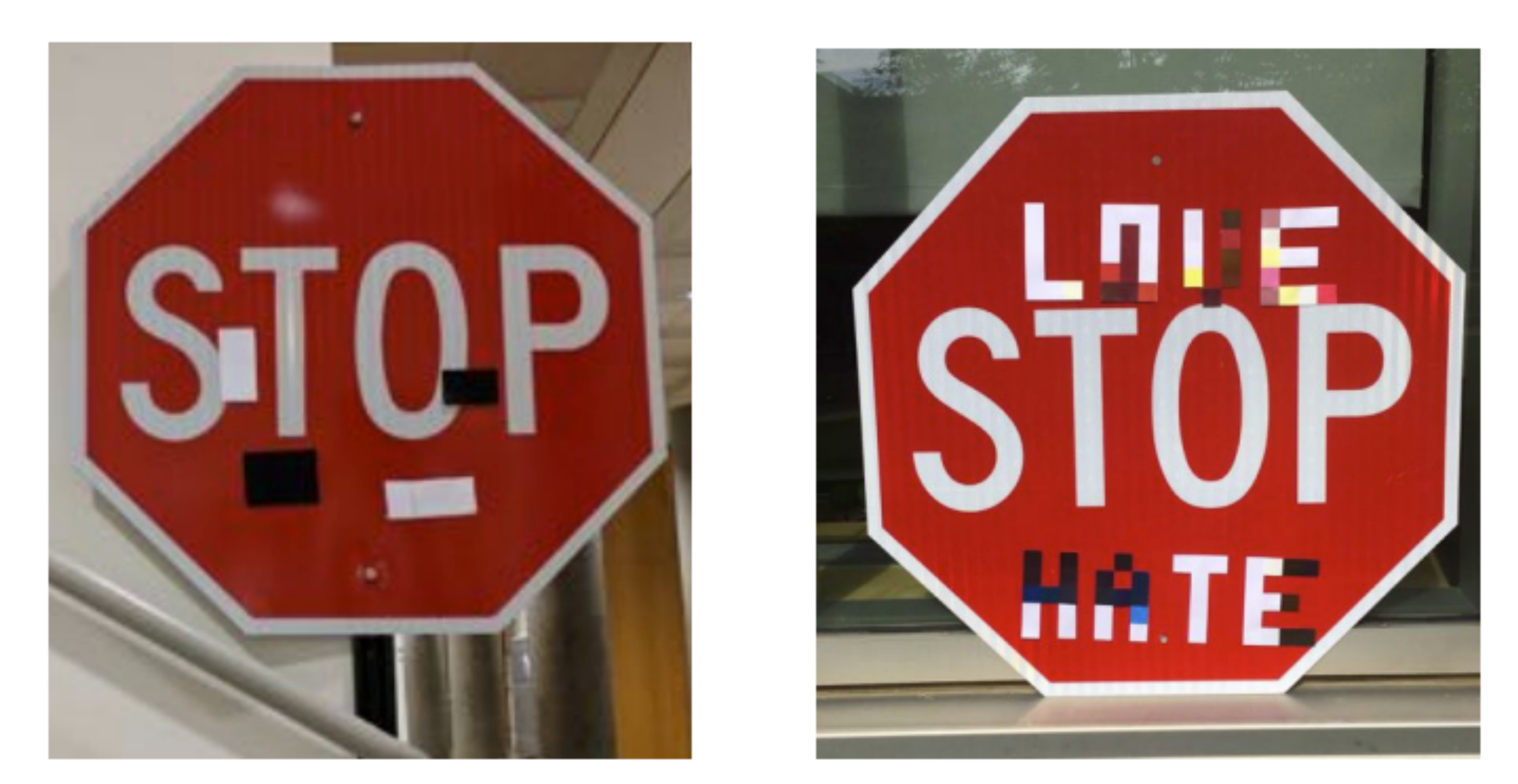
CS109B Data Science 2

Pavlos Protopapas and Mark Glickman

How vulnerable are Neural Networks?

Uses of Neural Networks

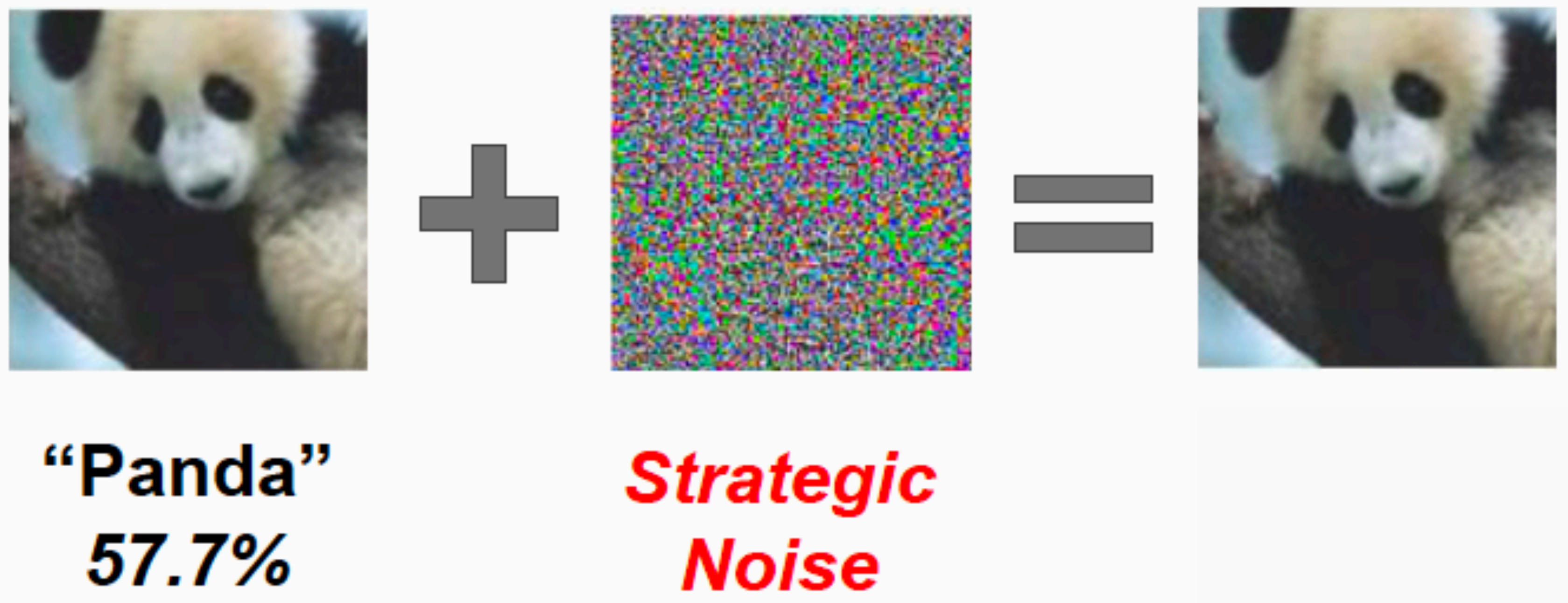
How vulnerable are Neural Networks?



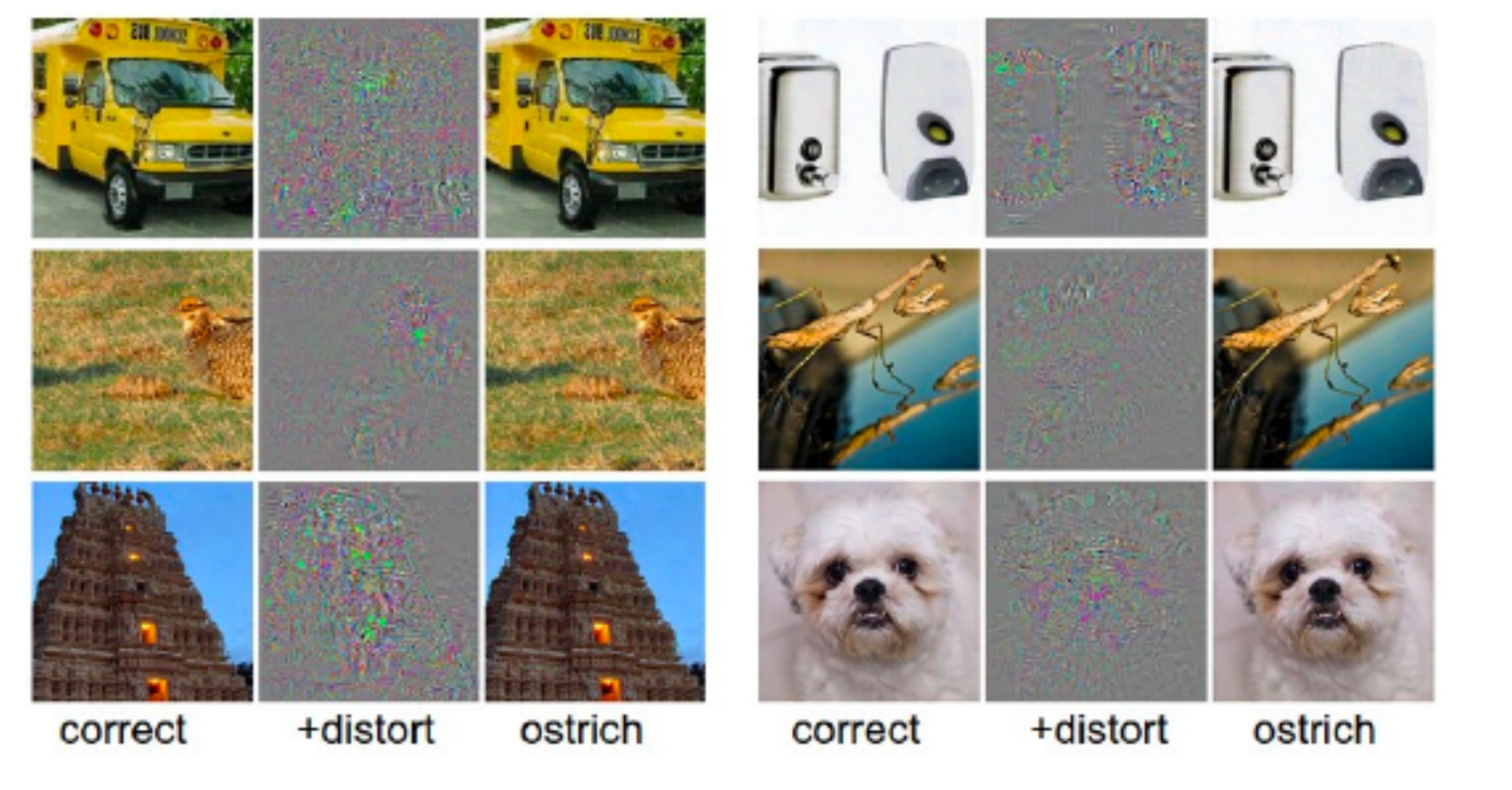
Explaining Adversarial Examples

[Goodfellow et. al ‘15]

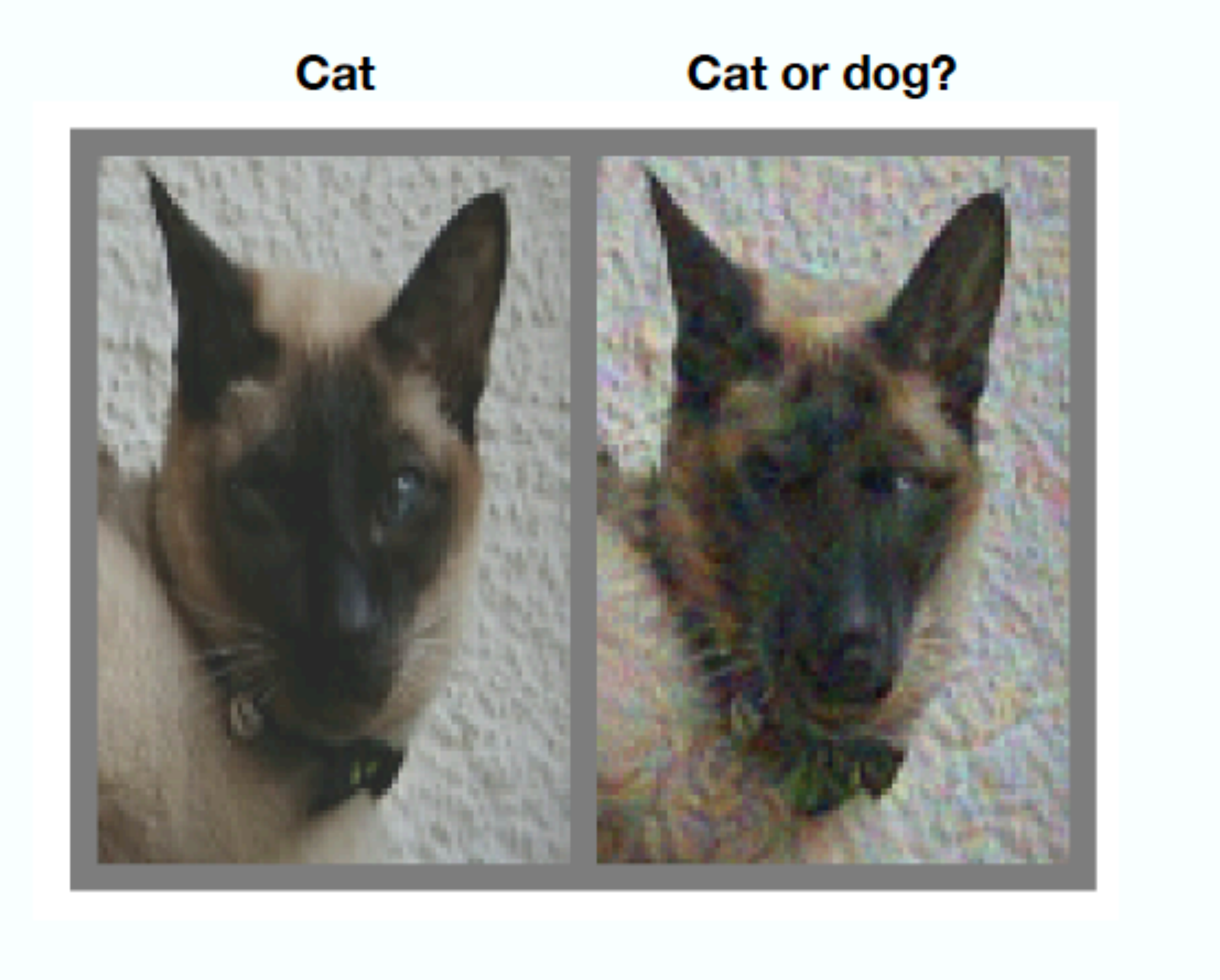
1. Robust attacks with FGSM
2. Robust defense with Adversarial Training



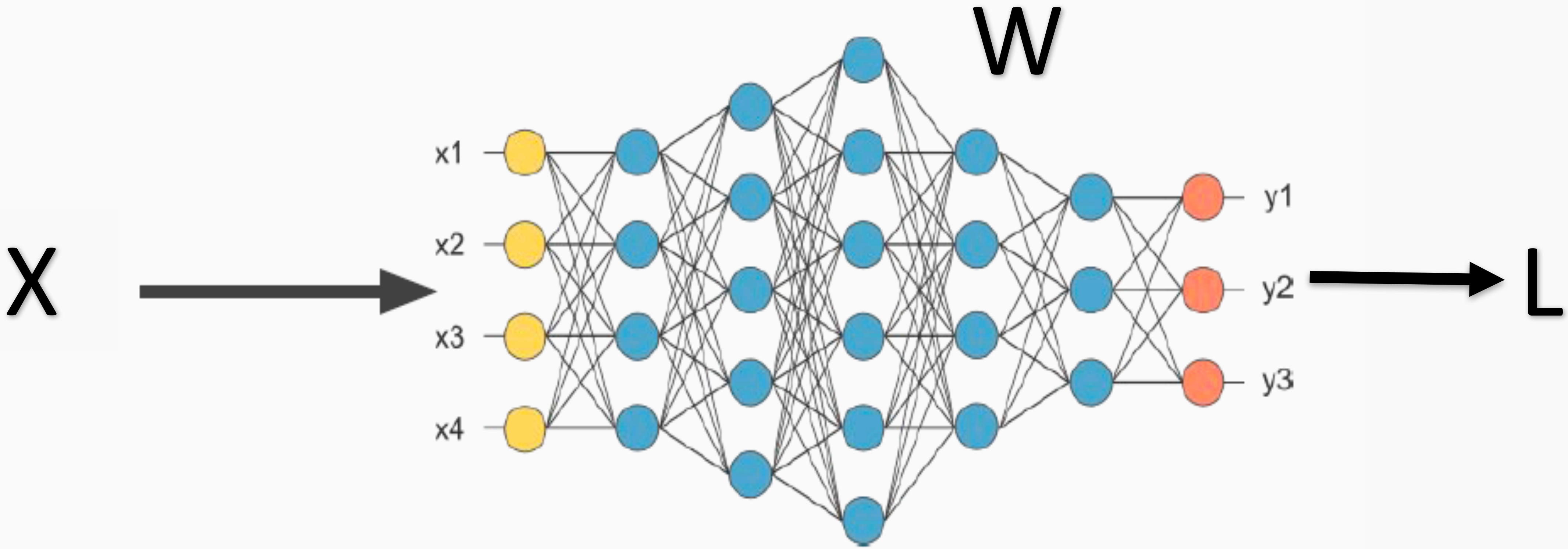
Explaining Adversarial Examples



Some of these adversarial examples can even fool humans:

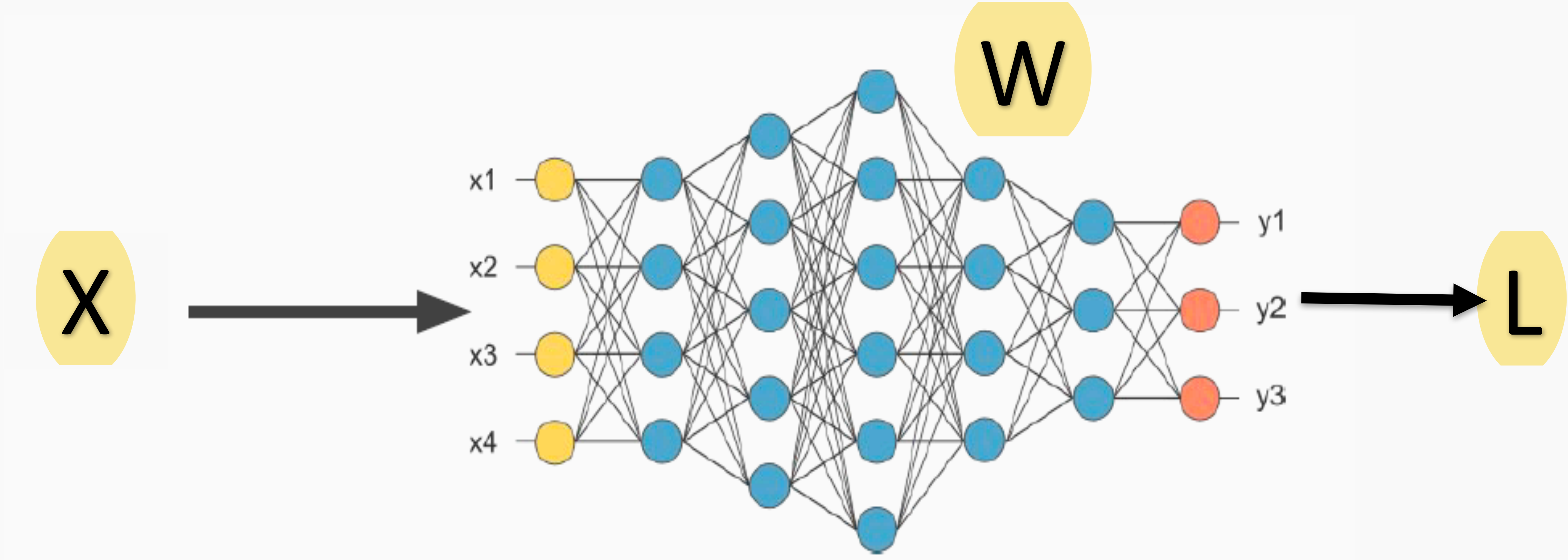


Attacking with **F**ast **G**radient **S**ign **M**ethod (FGSM)



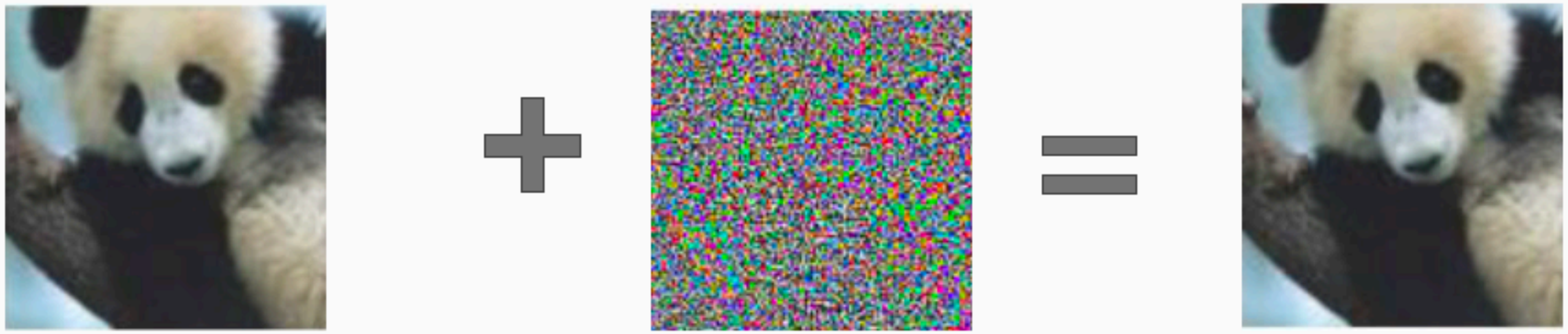
$$x + \lambda \cdot \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{L}) \Rightarrow \mathrm{x}^*$$

Attacking with **F**ast **G**radient **S**ign **M**ethod (FGSM)

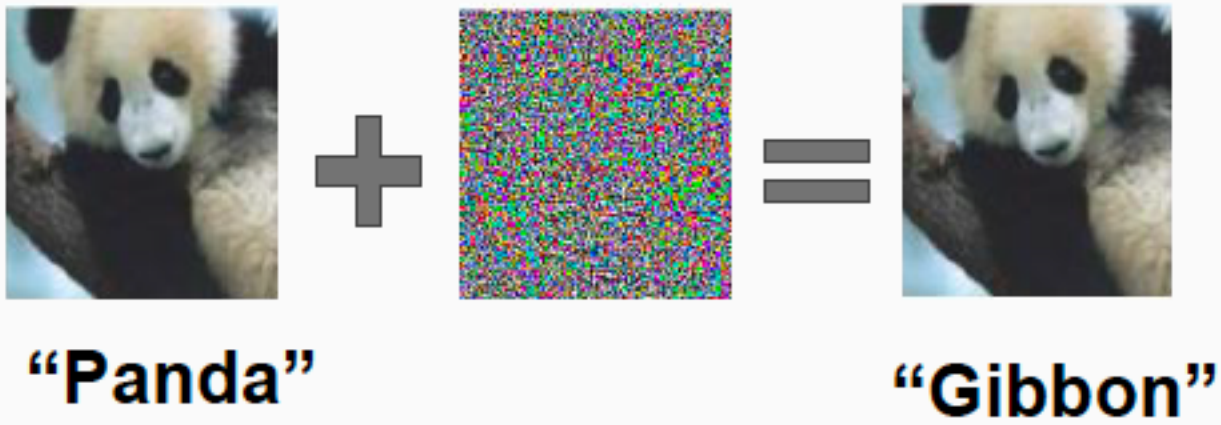


$$x + \lambda \cdot \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{L}) \Rightarrow \mathrm{x}^*$$

$$x + \lambda \cdot \mathrm{sign}(\nabla_{\mathrm{x}}\mathrm{L}) \Rightarrow \mathrm{x}^*$$

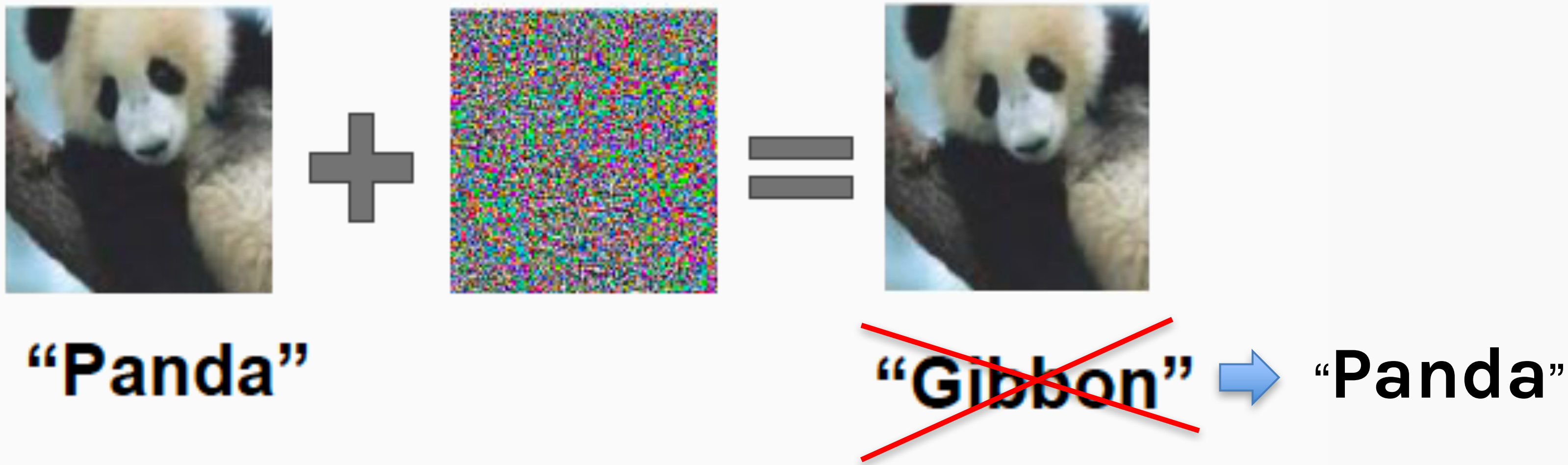


Defending with Adversarial Training



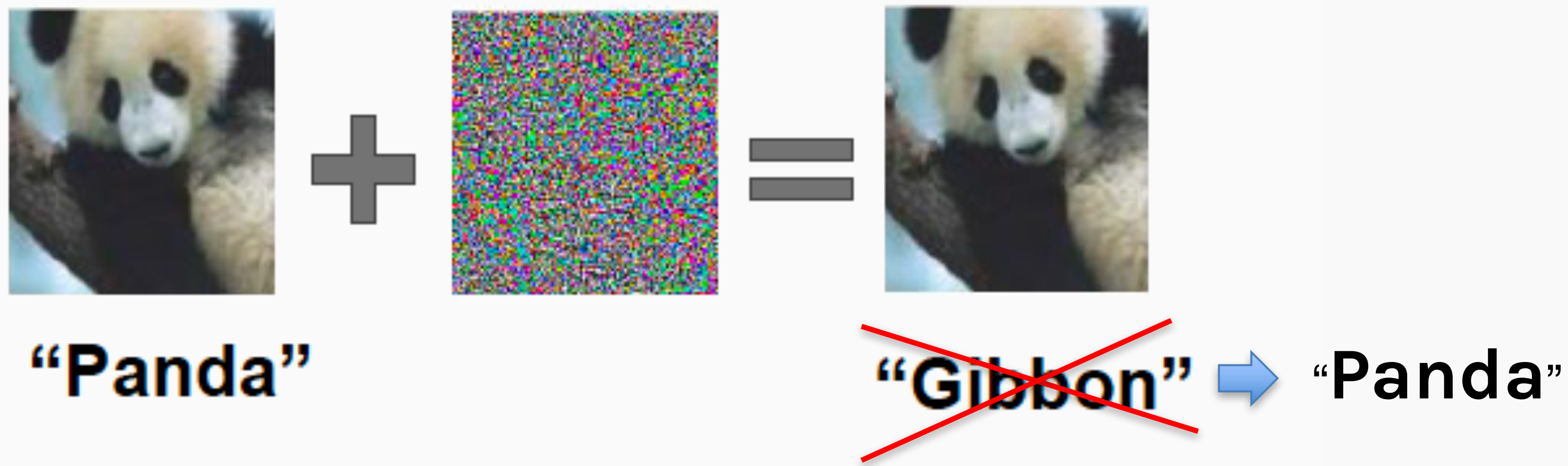
1. Generate adversarial examples
2. Adjust labels

Defending with Adversarial Training



1. Generate adversarial examples
2. Adjust labels

Defending with Adversarial Training

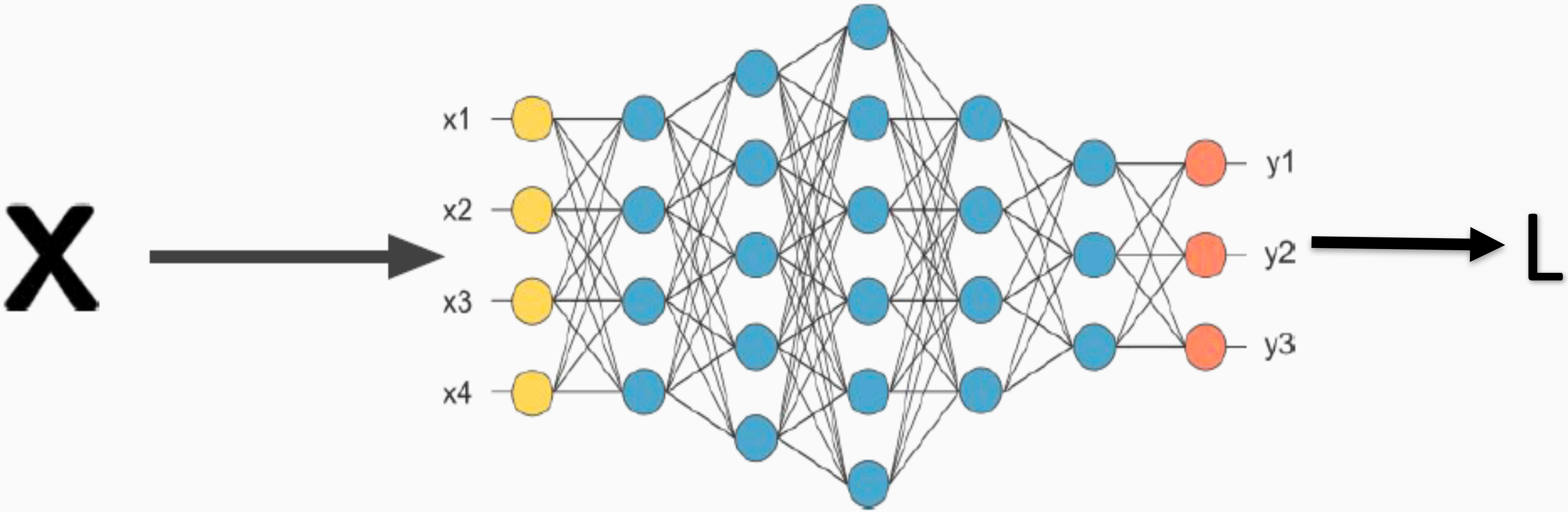


1. Generate adversarial examples
2. Adjust labels
3. Add them to the training set
4. Train new network

Attack methods post GoodFellow 2015

- FGSM [Goodfellow et. al ‘15]
- JSMA [Papernot et. al ‘16]
- C&W [Carlini + Wagner ‘16]
- Step-LL [Kurakin et. al ‘17]
- I-FGSM [Tramer et. al ‘18]

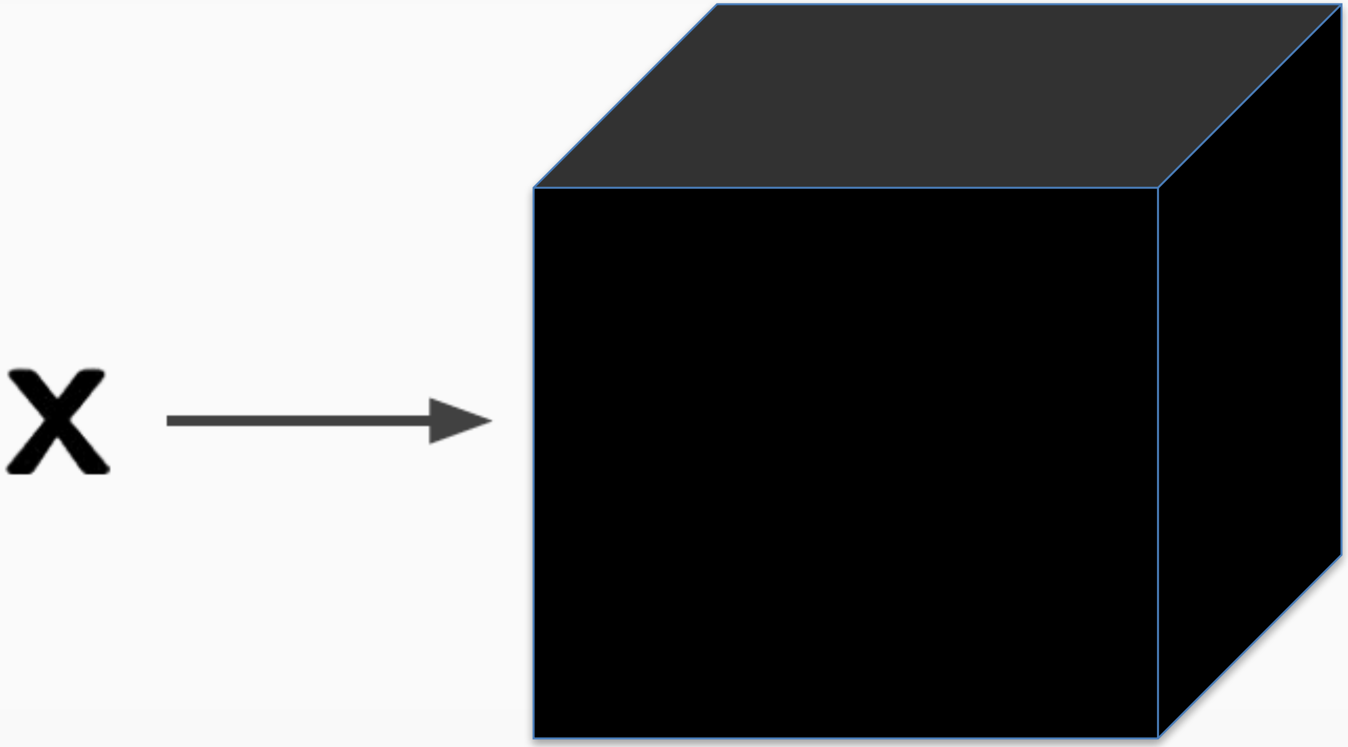
White box attacks



$$x + \lambda \cdot \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{L}) \Rightarrow \mathrm{x}^{*}$$

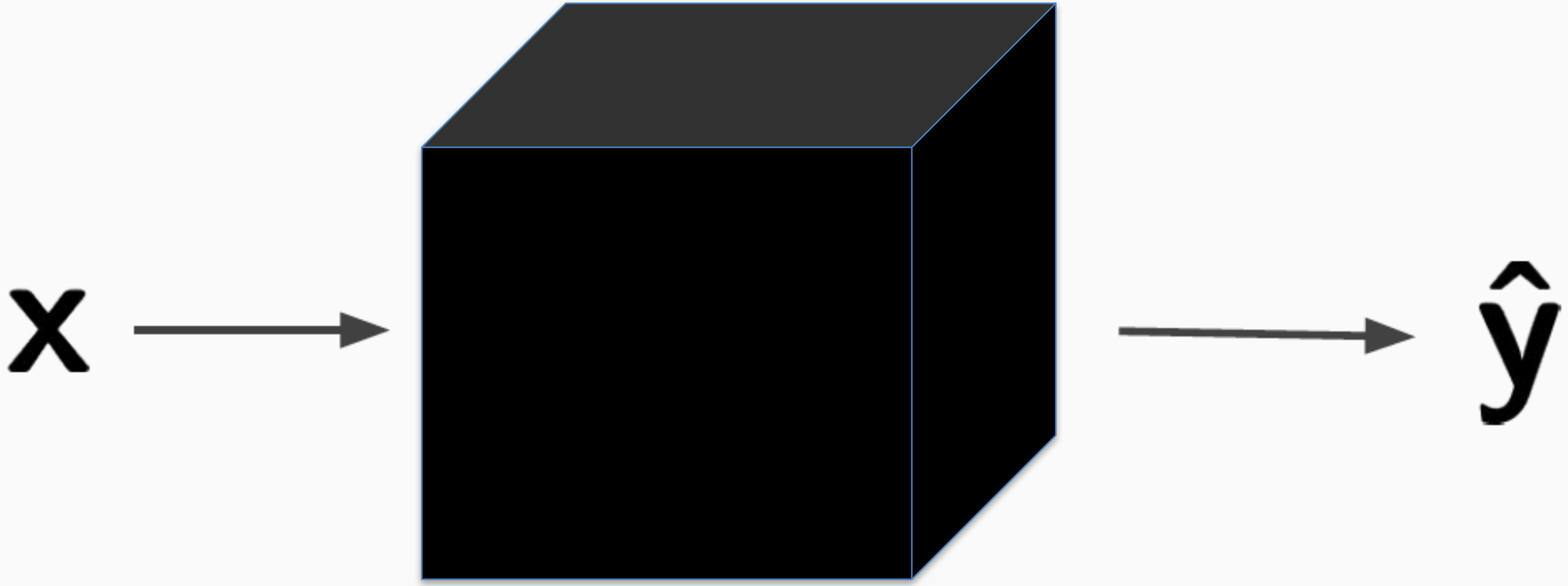
“Black Box” Attacks

“Black Box” Attacks [Papernot et. al ‘17]



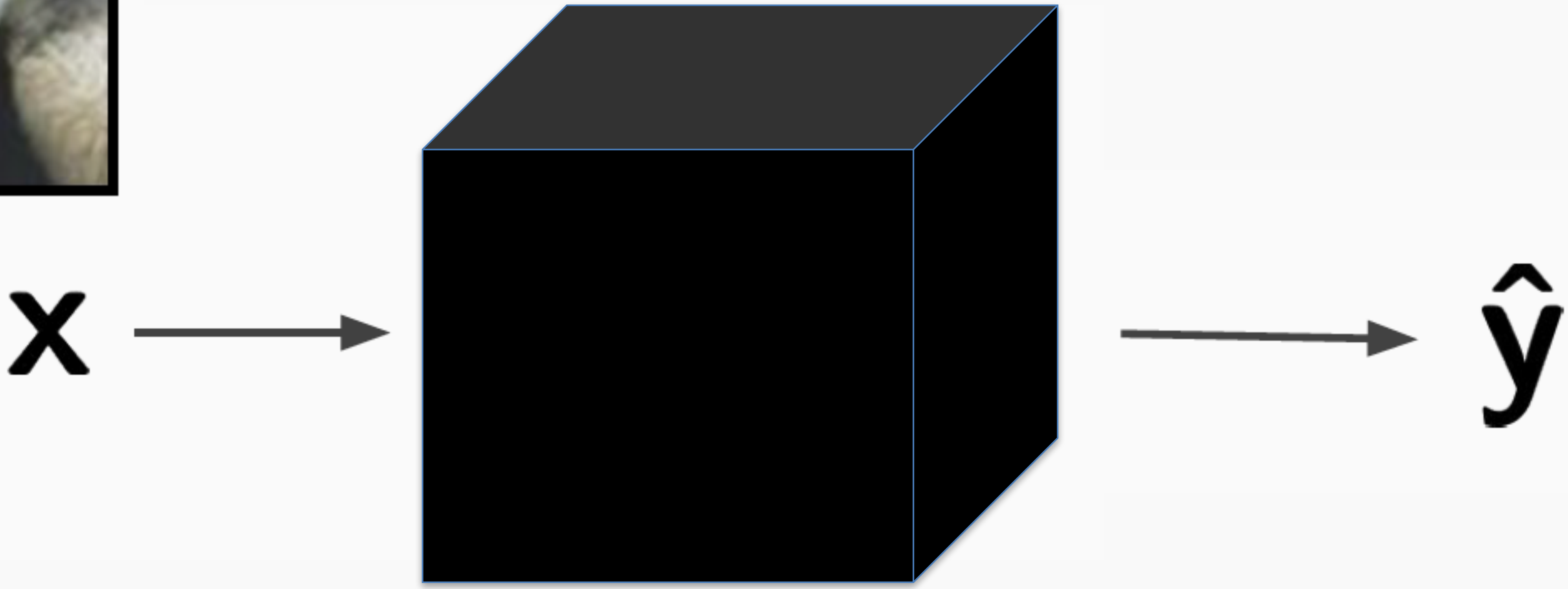
“Black Box” Attacks

Examine inputs and outputs of the model

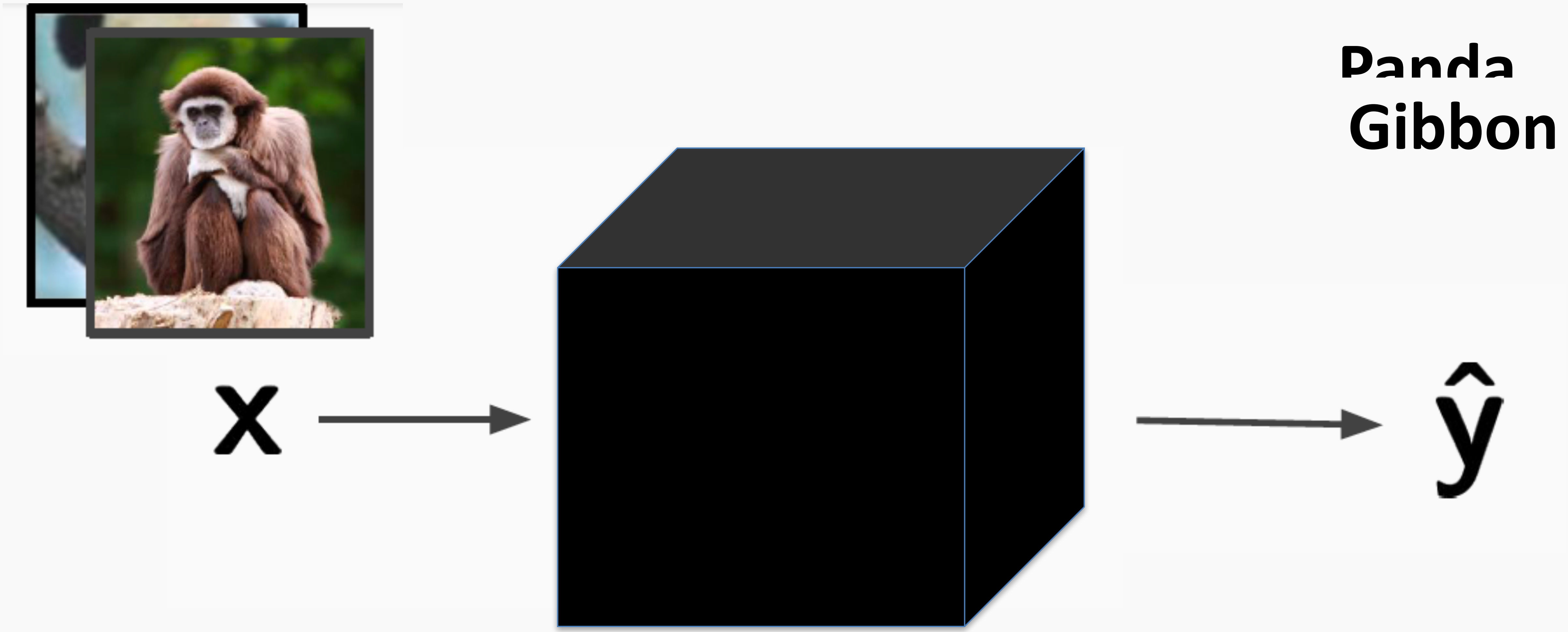


“Black Box” Attacks

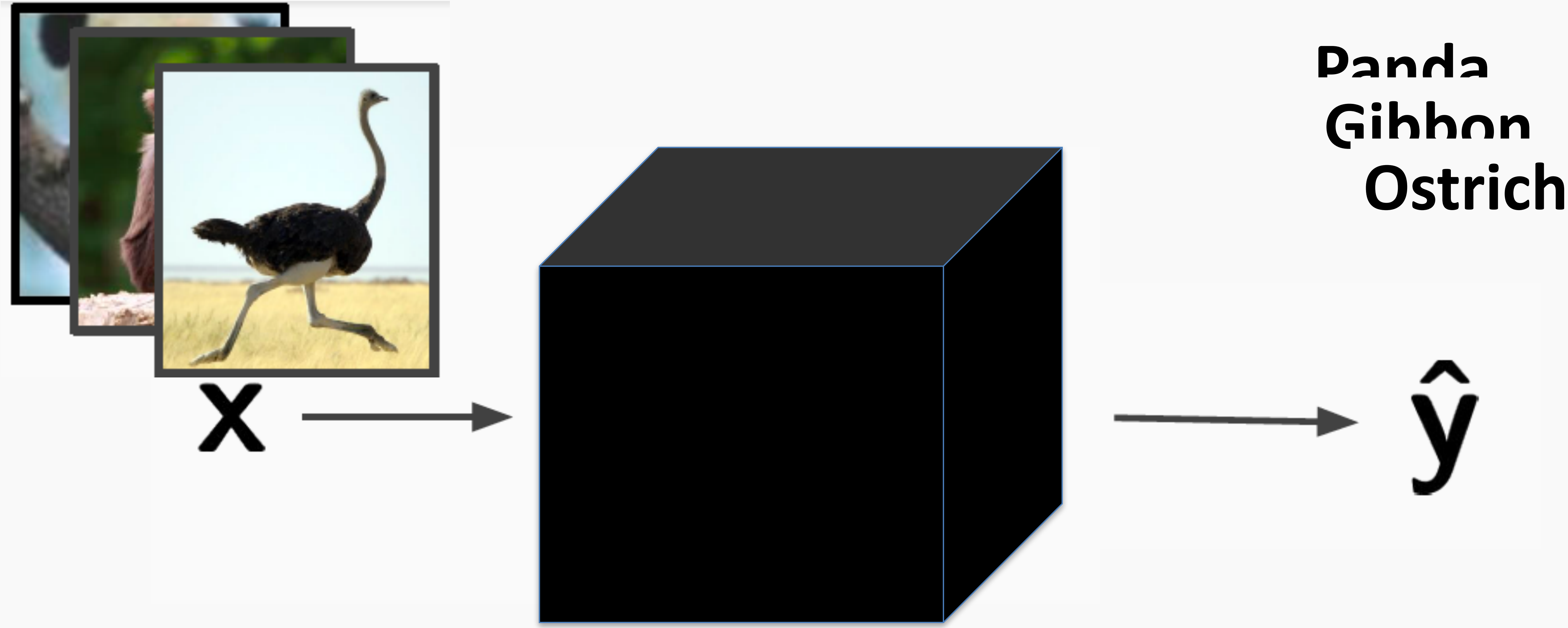
Panda



“Black Box” Attacks



“Black Box” Attacks

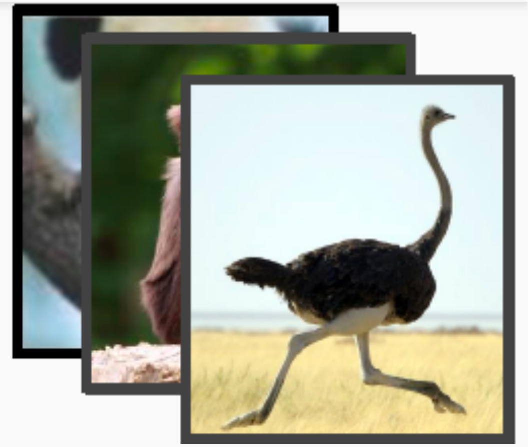


“Black Box” Attacks

Train a model that performs the same as the black box

“Black Box” Attacks

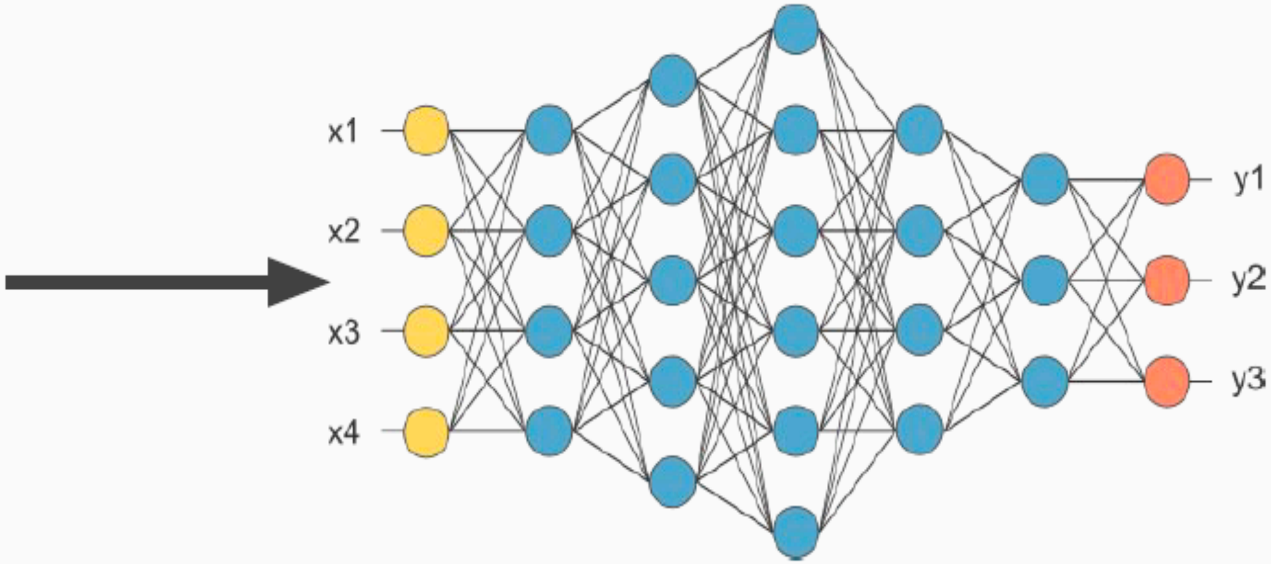
Train a model that performs the same as the black box



Panda

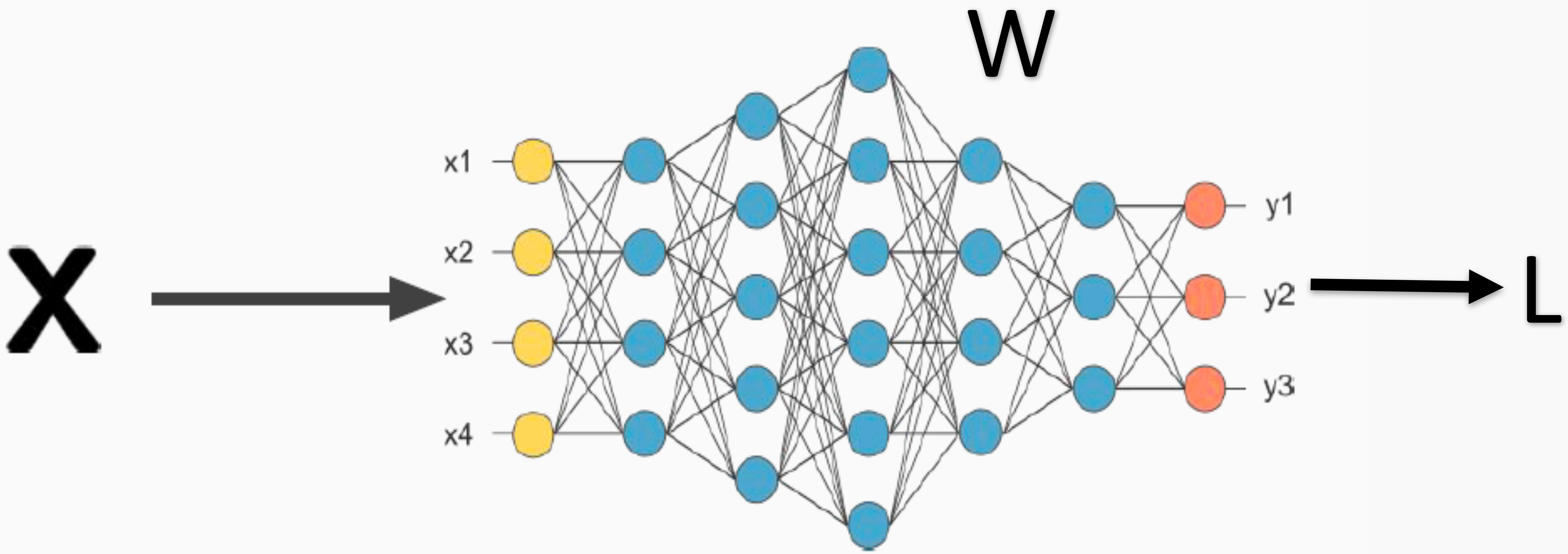
Gibbon

Ostrich



“Black Box” Attacks

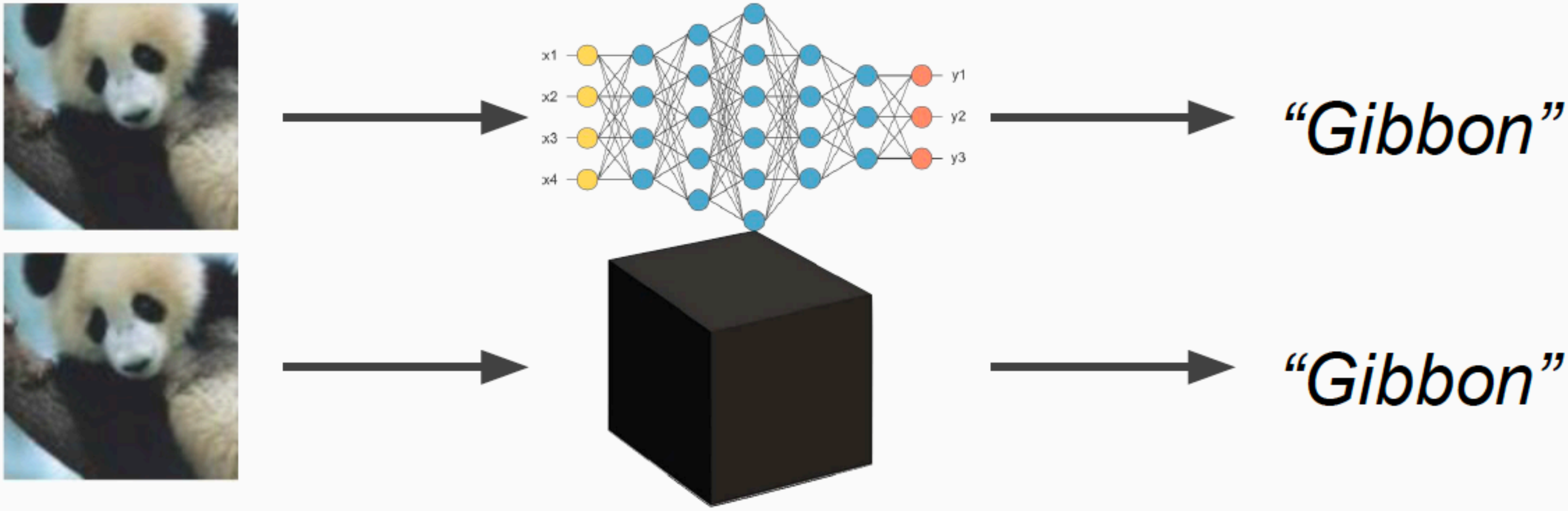
Now attack the model you just trained with “white” box attack



$$x + \lambda \cdot \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{L}) \Rightarrow \mathrm{x}^*$$

“Black Box” Attacks

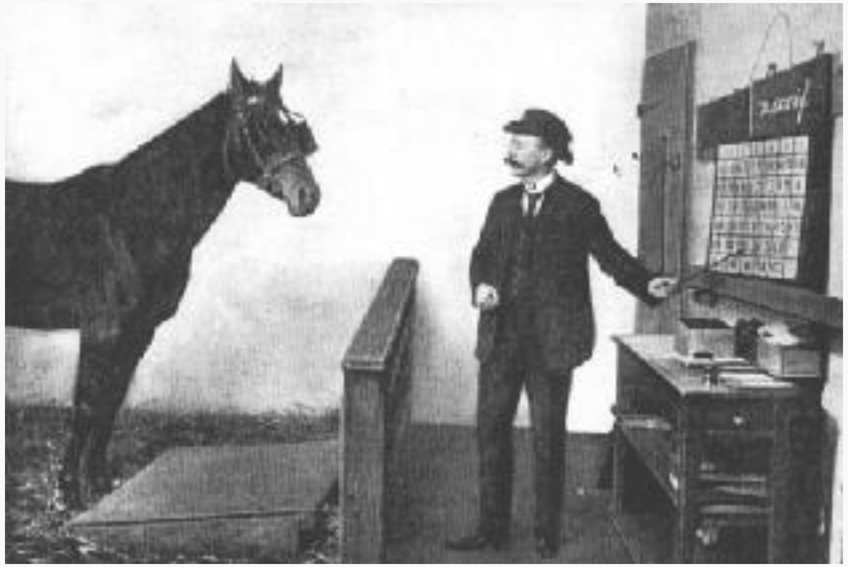
Use those adversarial examples to the “black” box



“Gibbon”

“Gibbon”

CleverHans



A Python library to benchmark machine learning systems' vulnerability to adversarial examples.

https://github.com/tensorflow/cleverhans

http://www.cleverhans.io/

More Defenses

Mixup:

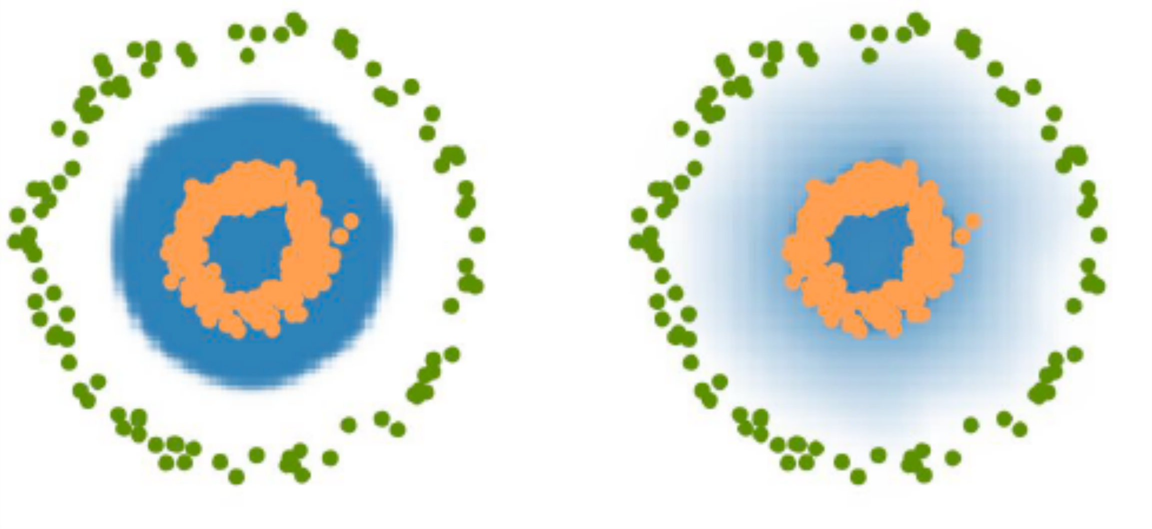
- Mix two training examples
- Augment training set

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

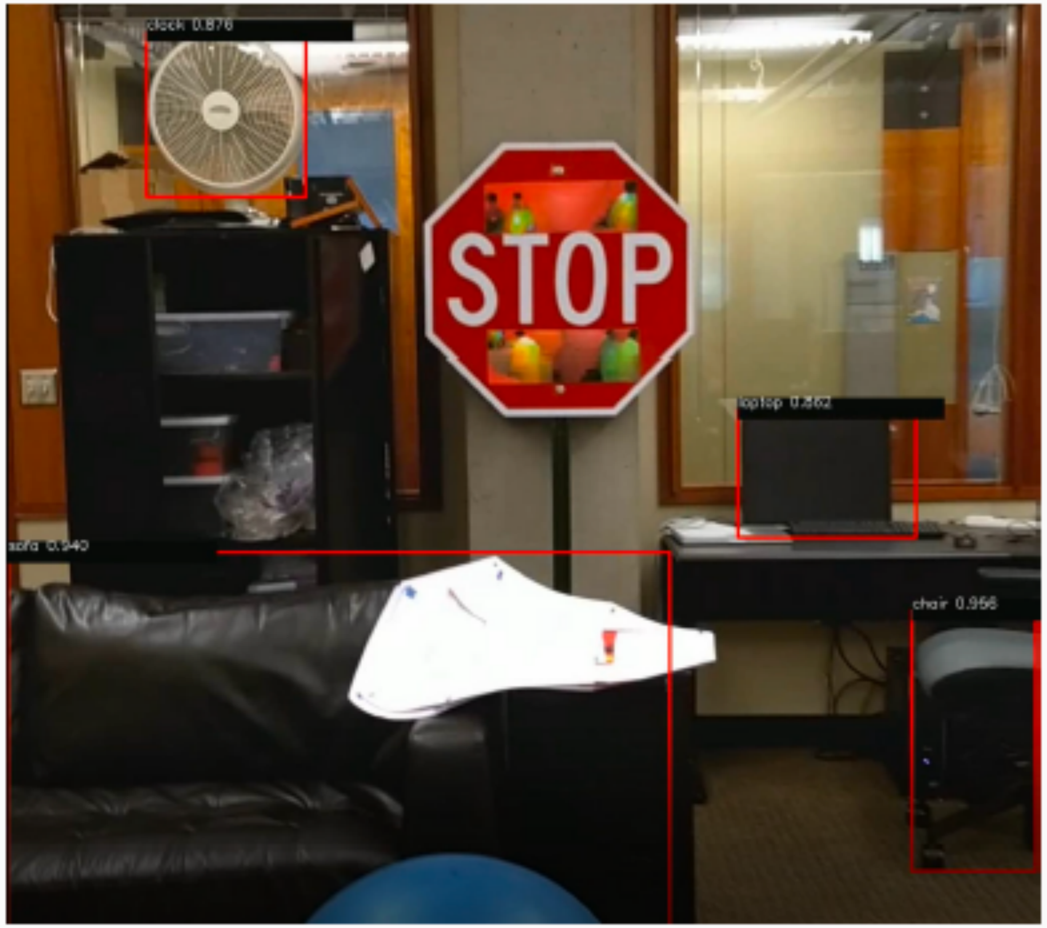
Smooth decision boundaries:

- Regularize the derivatives wrt to x



Physical attacks

- Object Detection
- Adversarial Stickers



Thank you.