

Introduction to Adversarial Machine Learning

Bibek Poudel

Sections

- Origin story
- Optimization problem
- Attacks
- Defenses
- Theories

Origin story



Christian Szegedy

Google Research

Verified email at google.com

[Machine learning](#) [Computer Vision](#) [Artificial Intelligence](#) [Automated Reasoning](#)



Christian Szegedy

[Google Research](#)

Verified email at google.com

[Machine learning](#) [Computer Vision](#) [Artificial Intelligence](#) [Automated Reasoning](#)





Christian Szegedy

Google Research

Verified email at google.com

Machine learning Computer Vision Artificial Intelligence Automated Reasoning



7



?



?



?



3



Christian Szegedy

Google Research

Verified email at google.com

Machine learning Computer Vision Artificial Intelligence Automated Reasoning

- Can I craft an optimization problem?





Christian Szegedy

Google Research

Verified email at google.com

Machine learning Computer Vision Artificial Intelligence Automated Reasoning

- Can I craft an optimization problem?
 - ▶ Looks like a 7 to human eye
 - ▶ But a model thinks its a 3





Christian Szegedy

Google Research

Verified email at google.com

Machine learning Computer Vision Artificial Intelligence Automated Reasoning

- “Intriguing properties of neural networks”
 - ▶ ICLR 2014, ~ 9000 citations
 - ▶ Birth of Adversarial Machine Learning (AML)



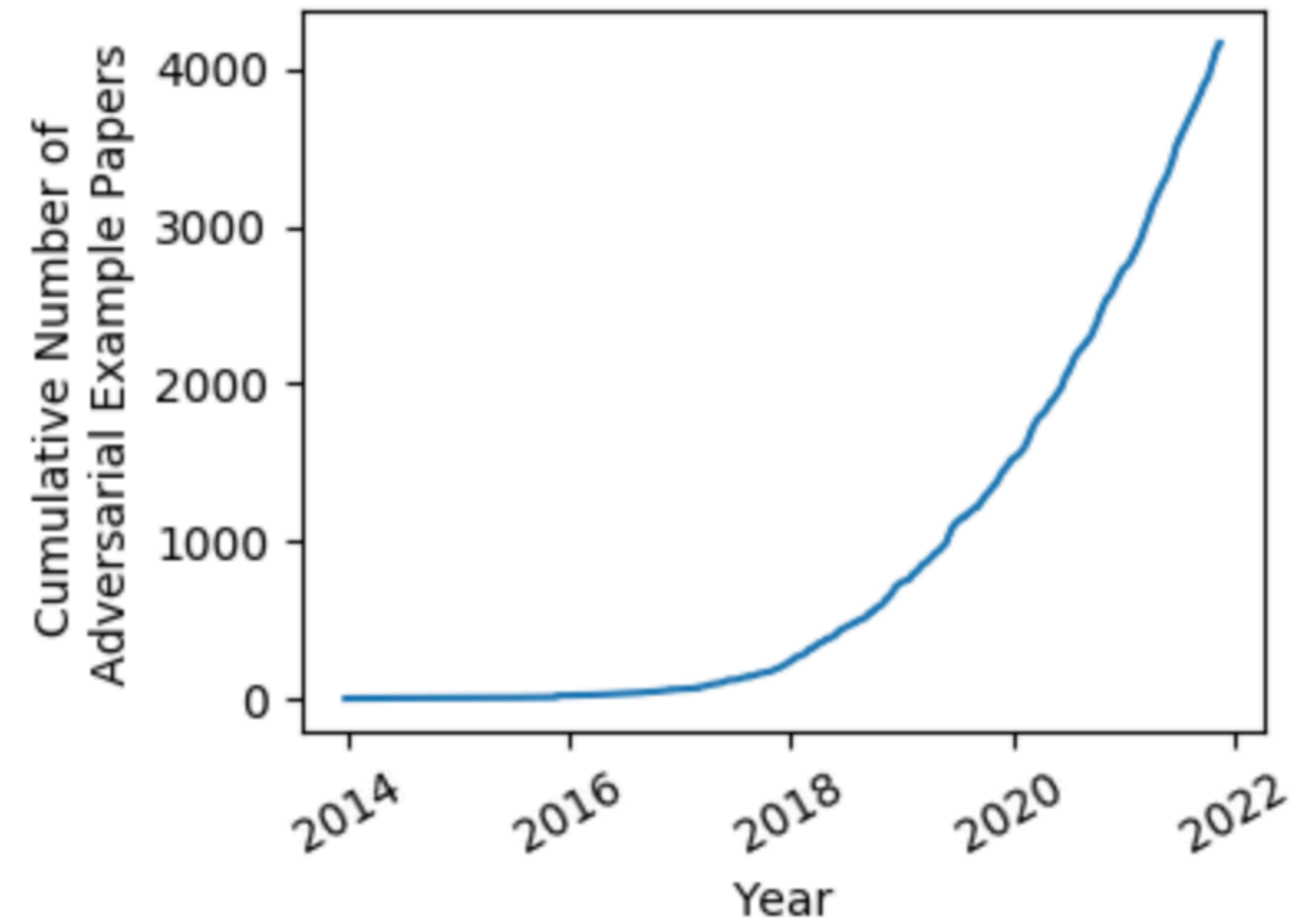
Christian Szegedy

[Google Research](#)

Verified email at google.com

[Machine learning](#) [Computer Vision](#) [Artificial Intelligence](#) [Automated Reasoning](#)

- Recent interest in *AML*



Adversarial examples in action

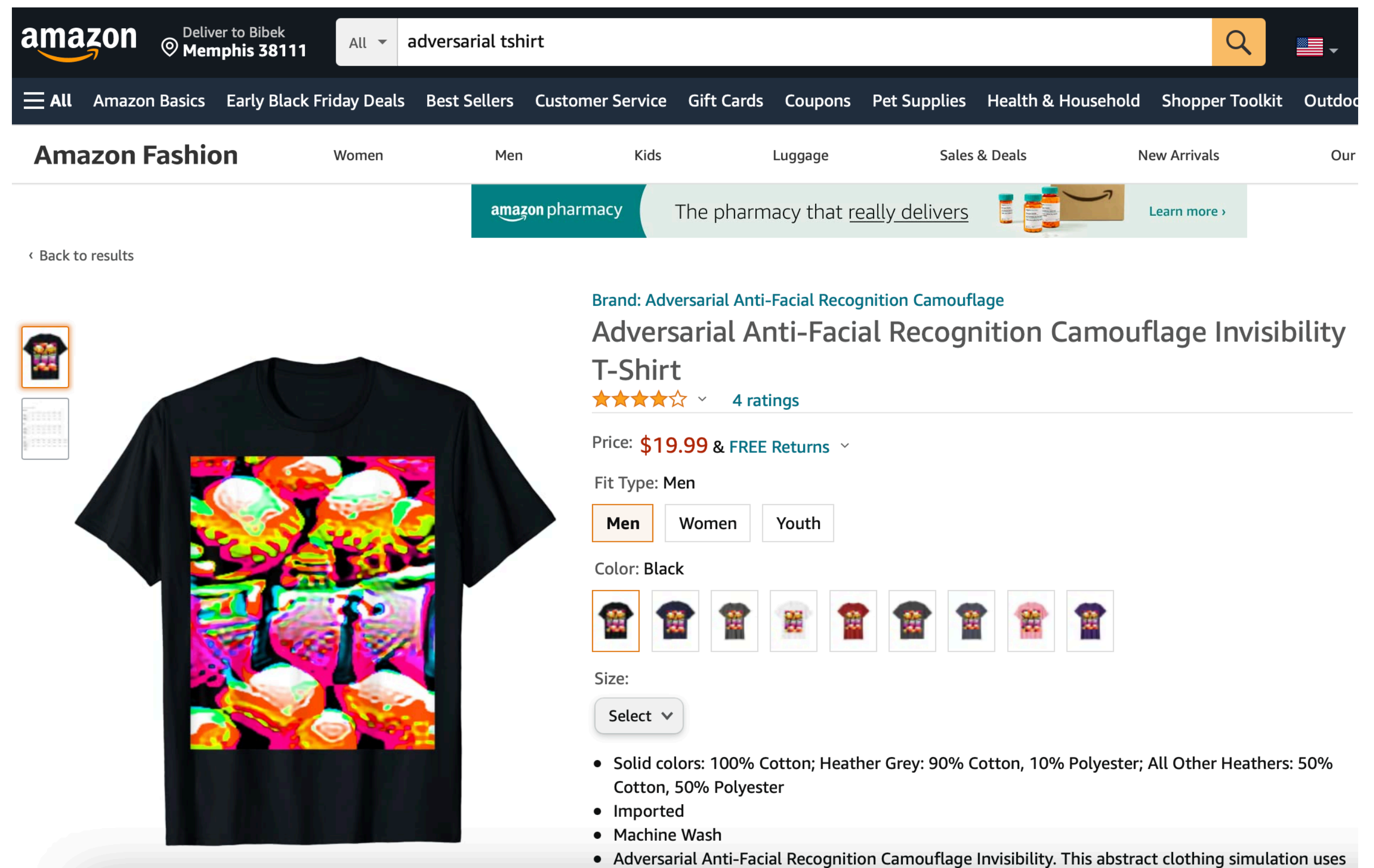
Adversarial examples in action

- Autonomous driving and traffic signs
- Video



Adversarial examples in action

- Surveillance, facial recognition
- Video



The screenshot shows an Amazon product page for a black t-shirt with a colorful, abstract camouflage pattern. The pattern is designed to be adversarial, meaning it is intended to cause facial recognition software to fail. The product title is "Adversarial Anti-Facial Recognition Camouflage Invisibility T-Shirt". The price is \$19.99, and it includes free returns. The product has 4 ratings. The page also shows navigation options like "Amazon Fashion", "Women", "Men", "Kids", "Luggage", "Sales & Deals", "New Arrivals", and "Our".

amazon Deliver to Bibek Memphis 38111 All adversarial tshirt

All Amazon Basics Early Black Friday Deals Best Sellers Customer Service Gift Cards Coupons Pet Supplies Health & Household Shopper Toolkit Outdoor

Amazon Fashion Women Men Kids Luggage Sales & Deals New Arrivals Our

amazon pharmacy The pharmacy that really delivers Learn more

Back to results

Brand: Adversarial Anti-Facial Recognition Camouflage

Adversarial Anti-Facial Recognition Camouflage Invisibility T-Shirt

★★★★☆ 4 ratings

Price: \$19.99 & FREE Returns

Fit Type: Men

Men Women Youth

Color: Black

Size: Select

- Solid colors: 100% Cotton; Heather Grey: 90% Cotton, 10% Polyester; All Other Heathers: 50% Cotton, 50% Polyester
- Imported
- Machine Wash
- Adversarial Anti-Facial Recognition Camouflage Invisibility. This abstract clothing simulation uses

Adversarial examples in action

- Reinforcement learning
- Video

Optimization problem

The optimization problem

- “Lp norm” distance metric



Image 1



Image 2

The optimization problem

- “Lp norm” distance metric

37	128	64
18	220	59
100	50	33

Image 1

38	128	64
18	99	59
100	50	33

Image 2

The optimization problem

- “Lp norm” distance metric
 - ▶ L0 distance = 2

37	128	64
18	220	59
100	50	33

Image 1

38	128	64
18	99	59
100	50	33

Image 2

The optimization problem

- “Lp norm” distance metric

▶ L1 distance = $|37 - 38| + |220 - 99|$

37	128	64
18	220	59
100	50	33

Image 1

38	128	64
18	99	59
100	50	33

Image 2

The optimization problem

- “Lp norm” distance metric
 - ▶ L2 distance = $(37 - 38)^2 + (220 - 99)^2$

37	128	64
18	220	59
100	50	33

Image 1

38	128	64
18	99	59
100	50	33

Image 2

The optimization problem

- “Lp norm” distance metric
 - ▶ L^∞ distance = $(220 - 99)$, max difference

37	128	64
18	220	59
100	50	33

Image 1

38	128	64
18	99	59
100	50	33

Image 2

The optimization problem

- Objective + constraints

The optimization problem

- Objective + constraints

$$\text{minimize } D(x, x + \delta_x)$$

The optimization problem

- Objective + constraints

minimize $D(x, x + \delta_x)$

subject to:

$$f(x) \neq f(x + \delta_x)$$

$$x + \delta_x \in [0, 1]^n$$

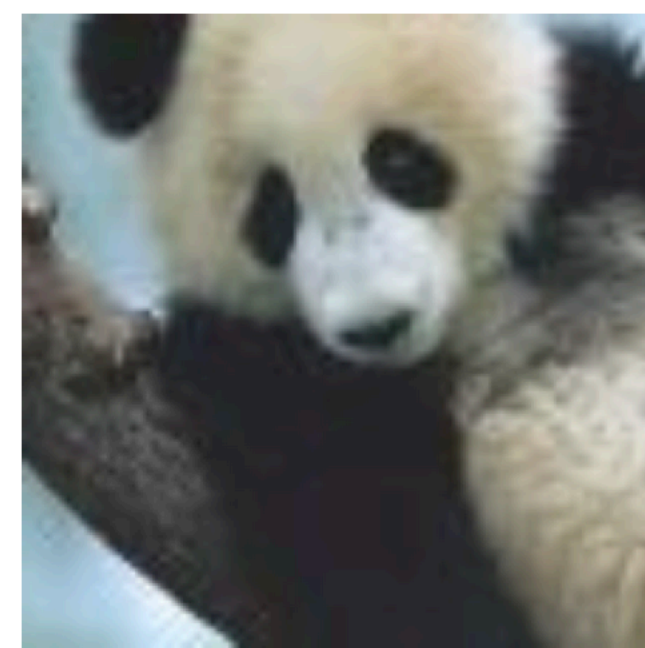
Attacks

Attacks

- Fast Gradient Sign Method (FGSM)
- “Explaining and harnessing adversarial examples”, Goodfellow et. al. 2015

Attacks

- Fast Gradient Sign Method (FGSM)
- “Explaining and harnessing adversarial examples”, Goodfellow et. al. 2015


 \mathbf{x}

“panda”

57.7% confidence

+ .007 ×


 $\text{sign}(\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y))$

“nematode”

8.2% confidence

=


 $\mathbf{x} +$
 $\epsilon \text{sign}(\nabla_{\mathbf{x}} J(\boldsymbol{\theta}, \mathbf{x}, y))$

“gibbon”

99.3 % confidence

Attacks

- Fast Gradient Sign Method (FGSM)

$$x_{adv} = x + \delta$$

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

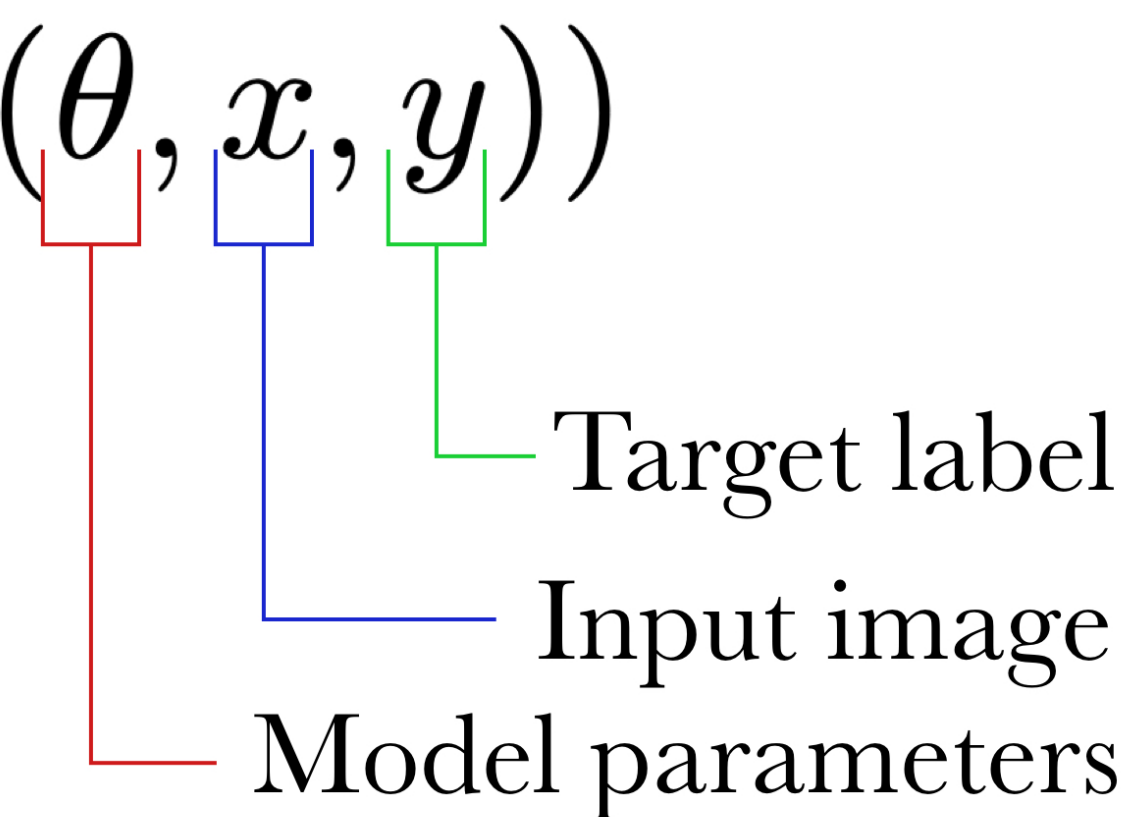
Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$


Target label

Input image

Model parameters

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

Loss value

Target label

Input image

Model parameters

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

Gradient w.r.t. input

Loss value

Target label

Input image

Model parameters

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

Gradient w.r.t. input

Loss value

Just take the sign

Target label

Input image

Model parameters

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y))$$

Diagram illustrating the components of the FGSM equation:

- ϵ : Scale
- sign : Just take the sign
- ∇_x : Gradient w.r.t. input
- θ : Model parameters
- x : Input image
- y : Target label
- J : Loss value

Attacks

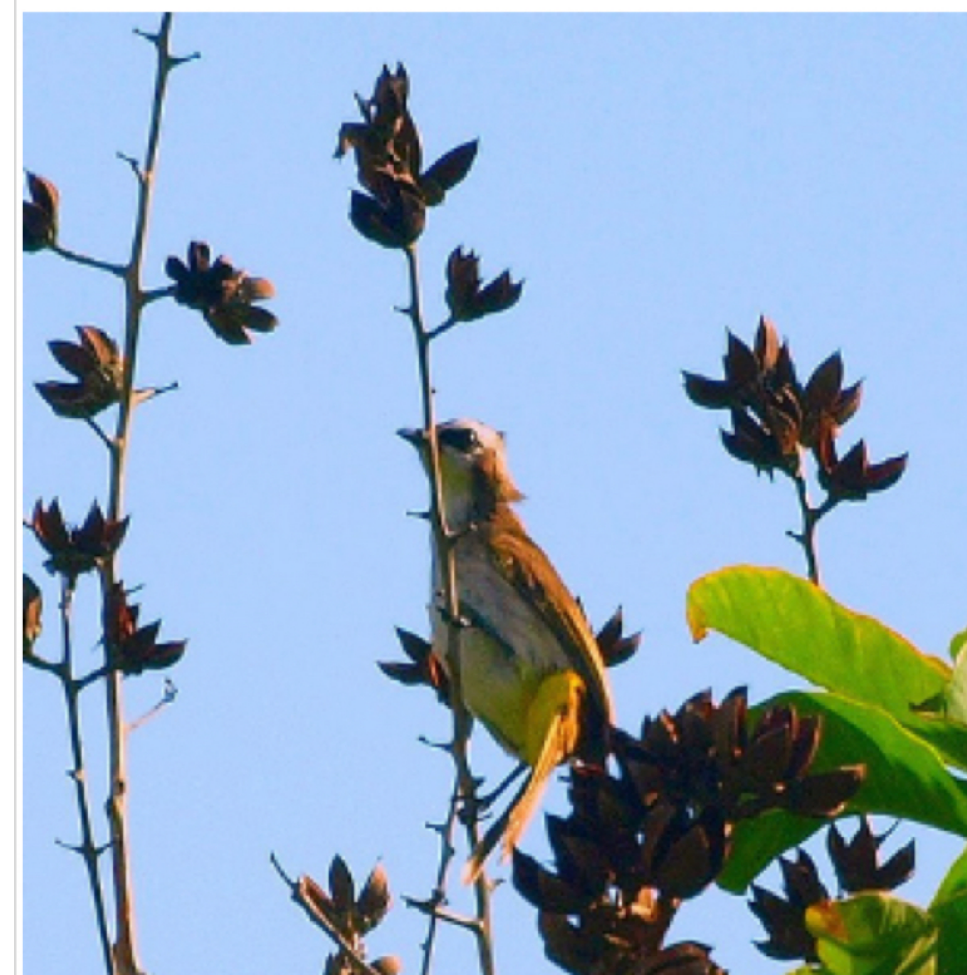
- Projected Gradient Descent (PGD)

Attacks

- Projected Gradient Descent (PGD)
 - ▶ Add random noise + take multiple smaller FGSM steps
 - ▶ Iterative

Attacks

- Projected Gradient Descent (PGD)
 - ▶ Add random noise + take multiple smaller FGSM steps
 - ▶ Iterative



Input



FGSM



PGD

Attacks

- One pixel attack



SHIP
CAR(99.7%)



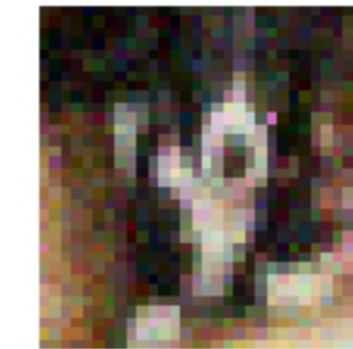
HORSE
FROG(99.9%)



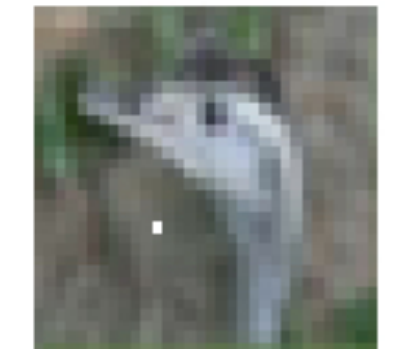
DEER
AIRPLANE(85.3%)



HORSE
DOG(70.7%)



DOG
CAT(75.5%)



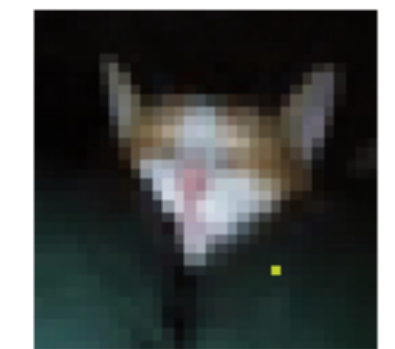
BIRD
FROG(86.5%)



CAR
AIRPLANE(82.4%)



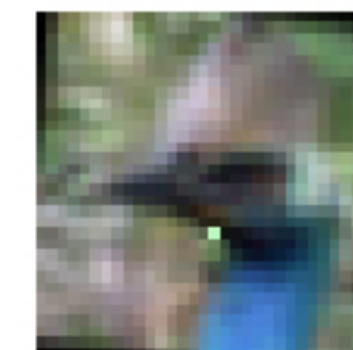
DEER
DOG(86.4%)



CAT
BIRD(66.2%)



DEER
AIRPLANE(49.8%)



BIRD
FROG(88.8%)

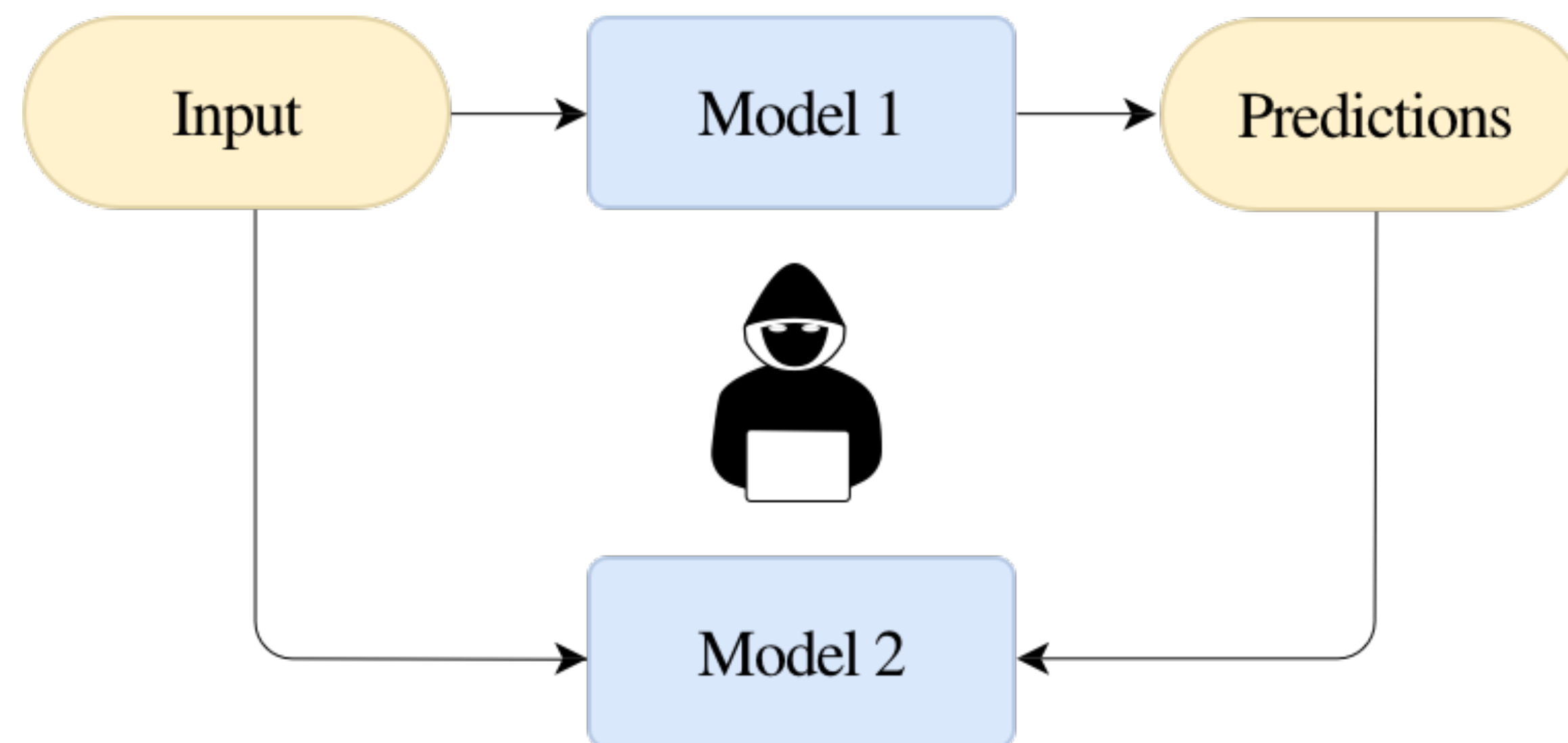


SHIP
AIRPLANE(88.2%)

Threat models

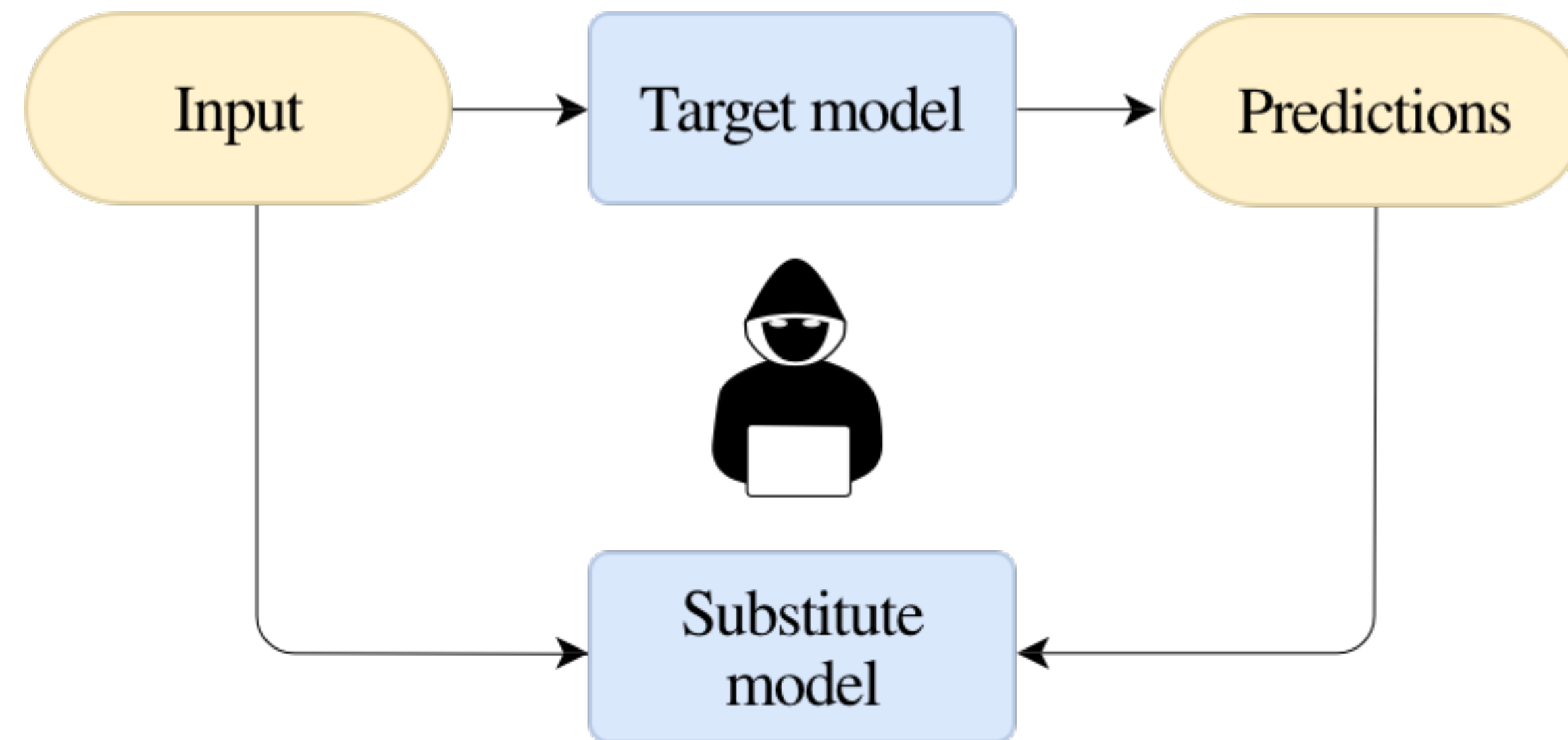
Threat models

- Black-box



Threat models

- Black-box



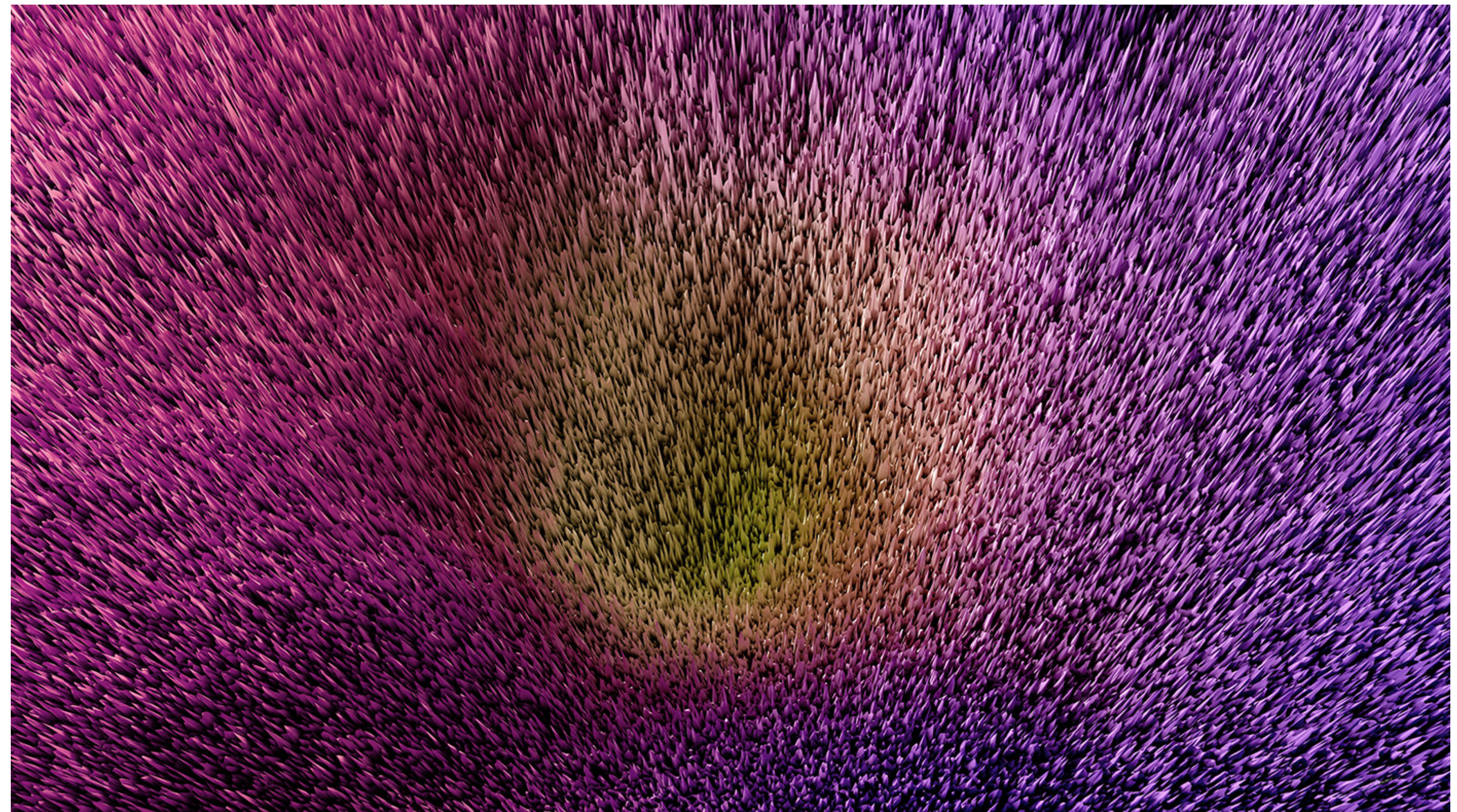
Threat models

- White-box
 - ▶ Training data, hyper-parameters, model architecture

Defenses

Defenses

- Gradient Masking
 - ▶ Hide gradient information
 - ▶ Discarded



Defenses

- Adversarial Training
 - ▶ Most successful

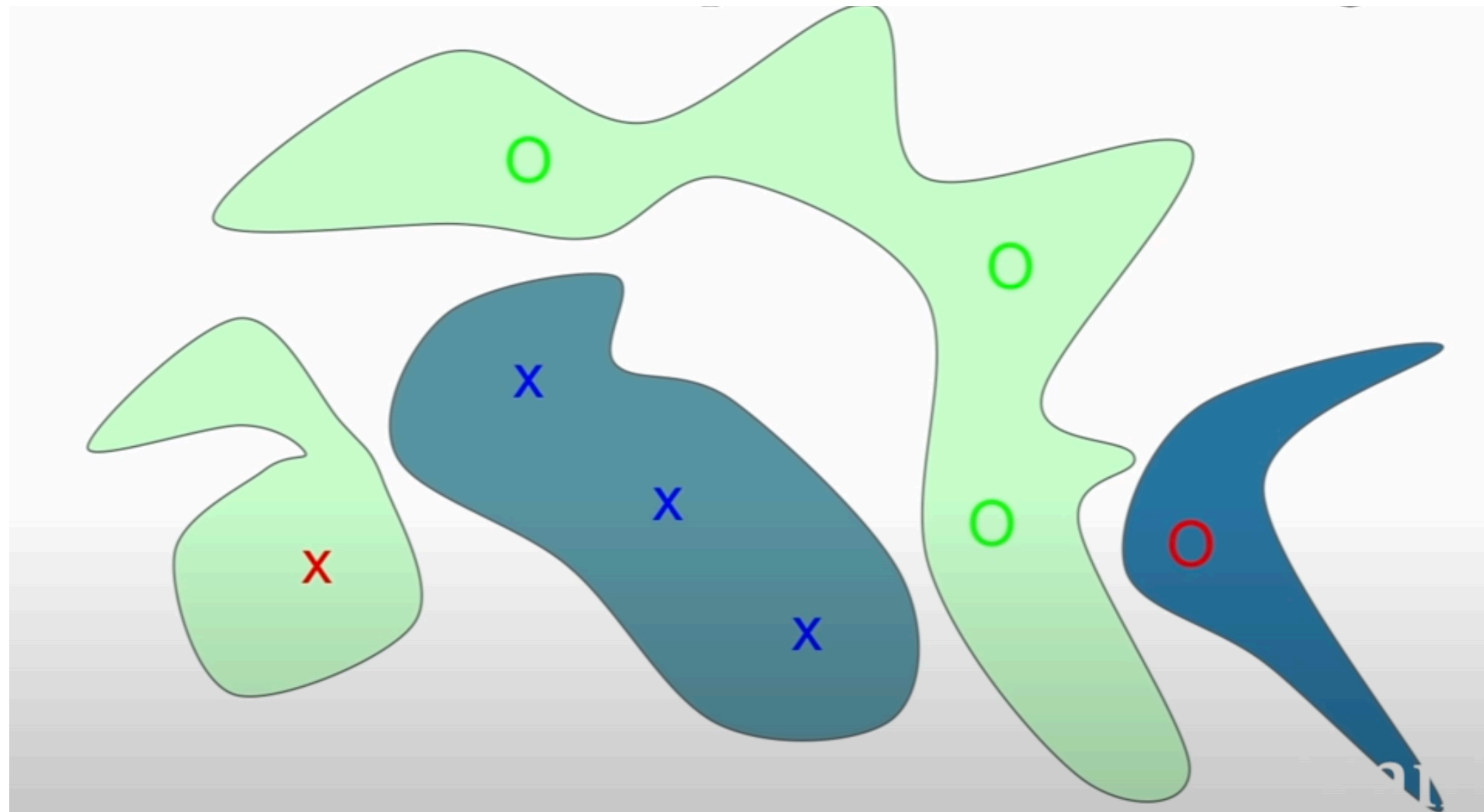
Theories

Theories

- Intuitively make sense but discarded
 - ▶ Overfitting

Theories

- Intuitively make sense but discarded
 - ▶ Overfitting



Theories

- Intuitively make sense but discarded
 - ▶ Excessive linearity

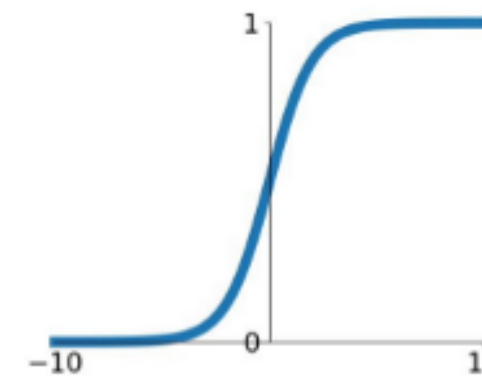
Theories

- Intuitively make sense but discarded
 - ▶ Excessive linearity

Activation Functions

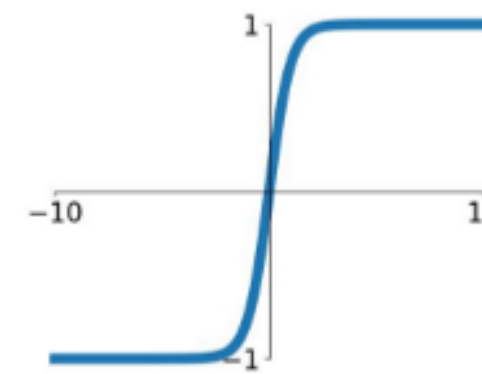
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



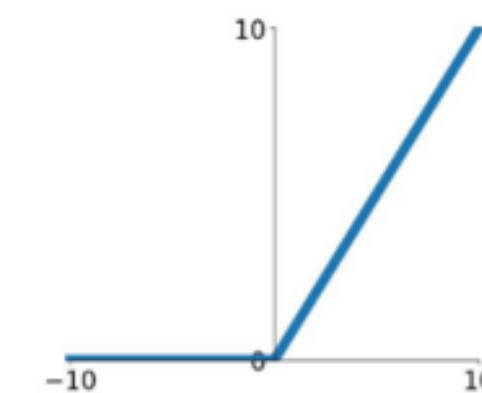
tanh

$$\tanh(x)$$



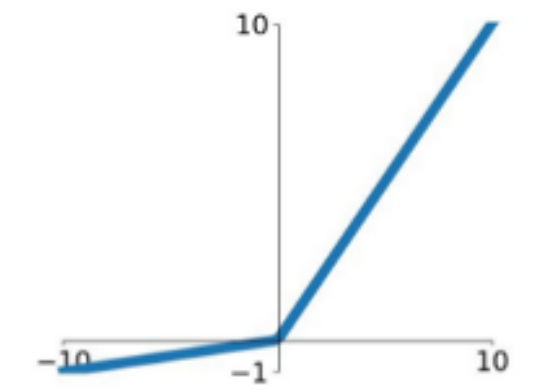
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

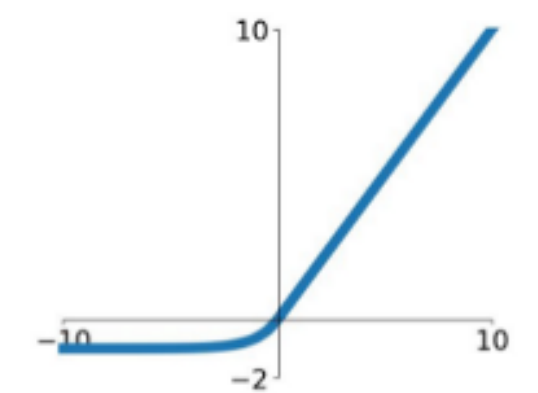


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

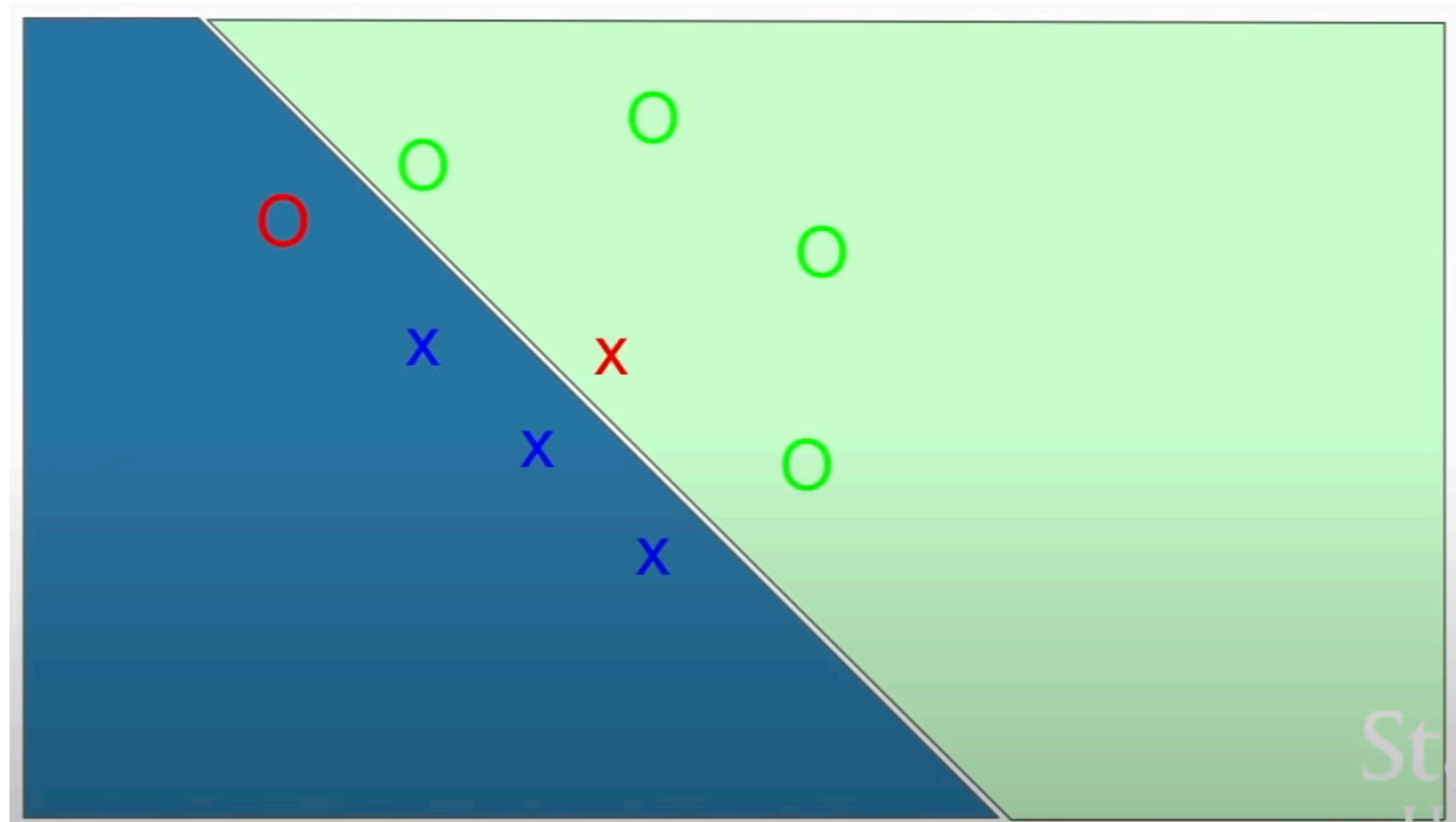
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Theories

- Intuitively make sense but discarded
 - ▶ Excessive linearity

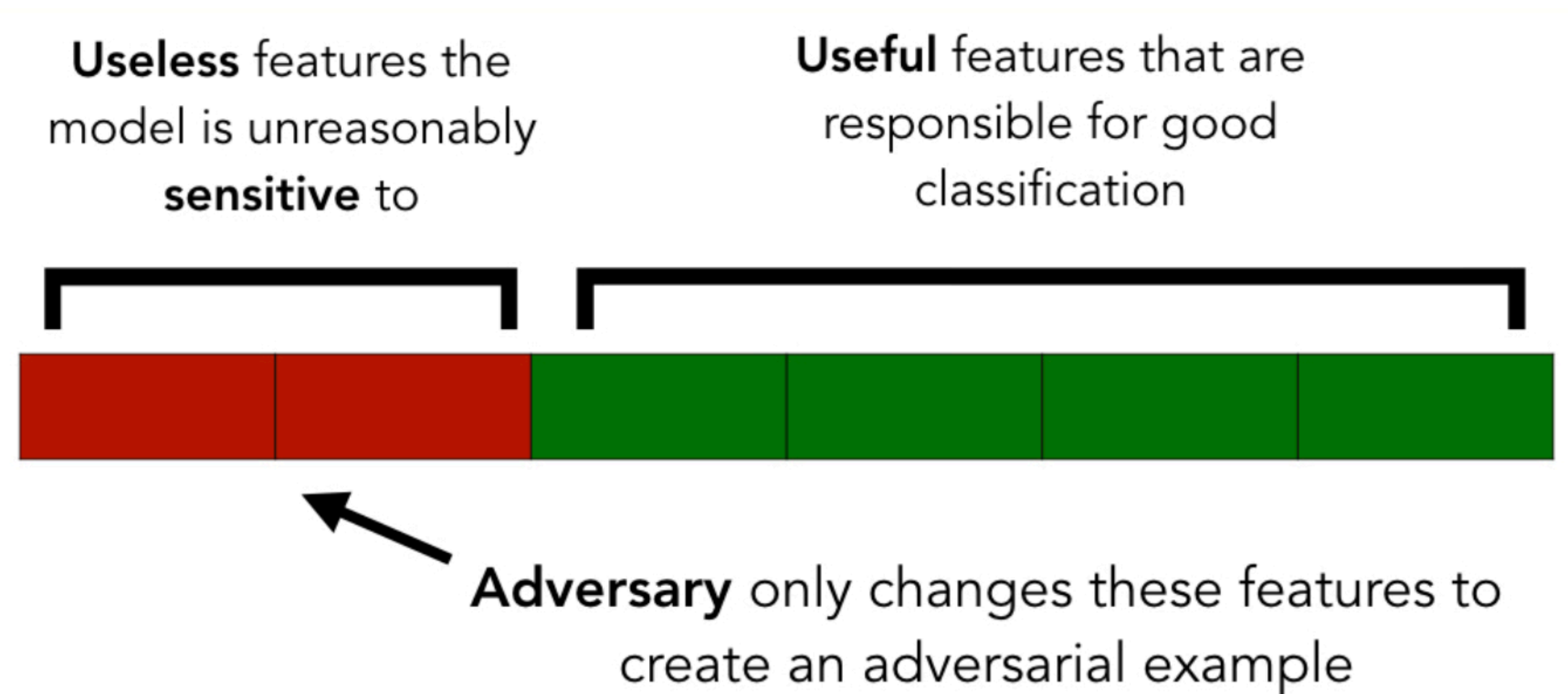


Theories

- Intuitively make sense but discarded
 - ▶ Adversarial examples are bugs

Theories

- Intuitively make sense but discarded
 - ▶ Adversarial examples are bugs



Theories

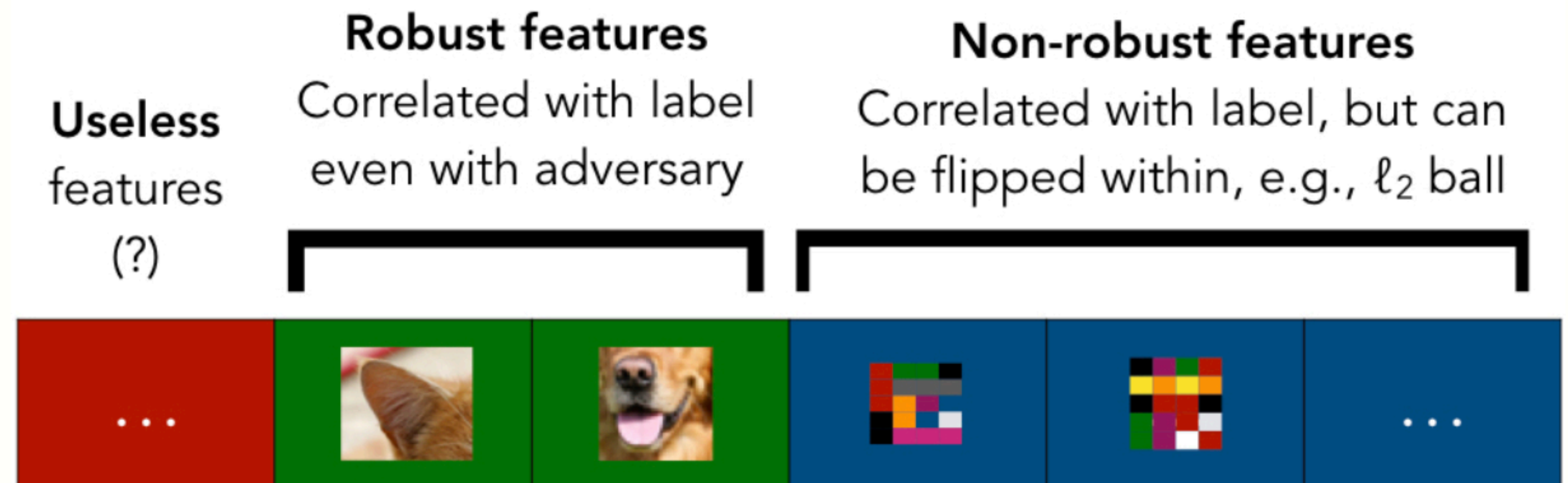
- Widely accepted

Theories

- Widely accepted
 - ▶ “Adversarial examples are not bugs, they are features” Illyas et. al. 2017

Theories

- Widely accepted
 - ▶ “Adversarial examples are not bugs, they are features” Illyas et. al. 2017



A more fundamental question

A more fundamental question

- Do our models really “learn”?

A more fundamental question

- Do our models really “learn”?
- Does the industry care about AEs? Video

Thank You!

Bibek Poudel

bpoudel@memphis.edu

But wait... there's more...

AI Camera Ruins Soccer Game For Fans After Mistaking Referee's Bald Head For Ball

69.7K
SHARES



Share on Facebook



Share on Twitter

