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

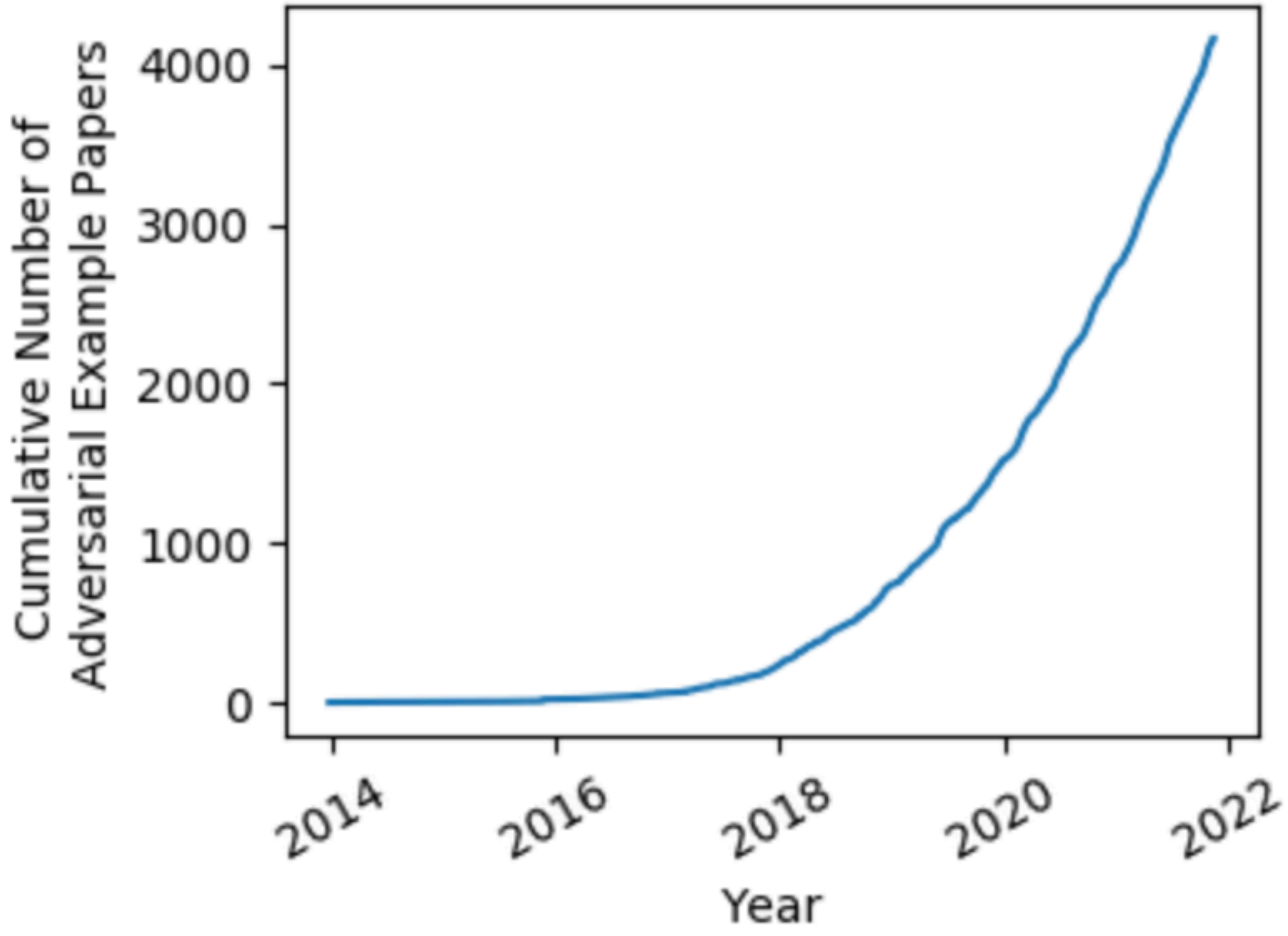
Google Research
Verified email at google.com
Machine learning Computer Vision Artificial Intelligence Automated Reasoning

- “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



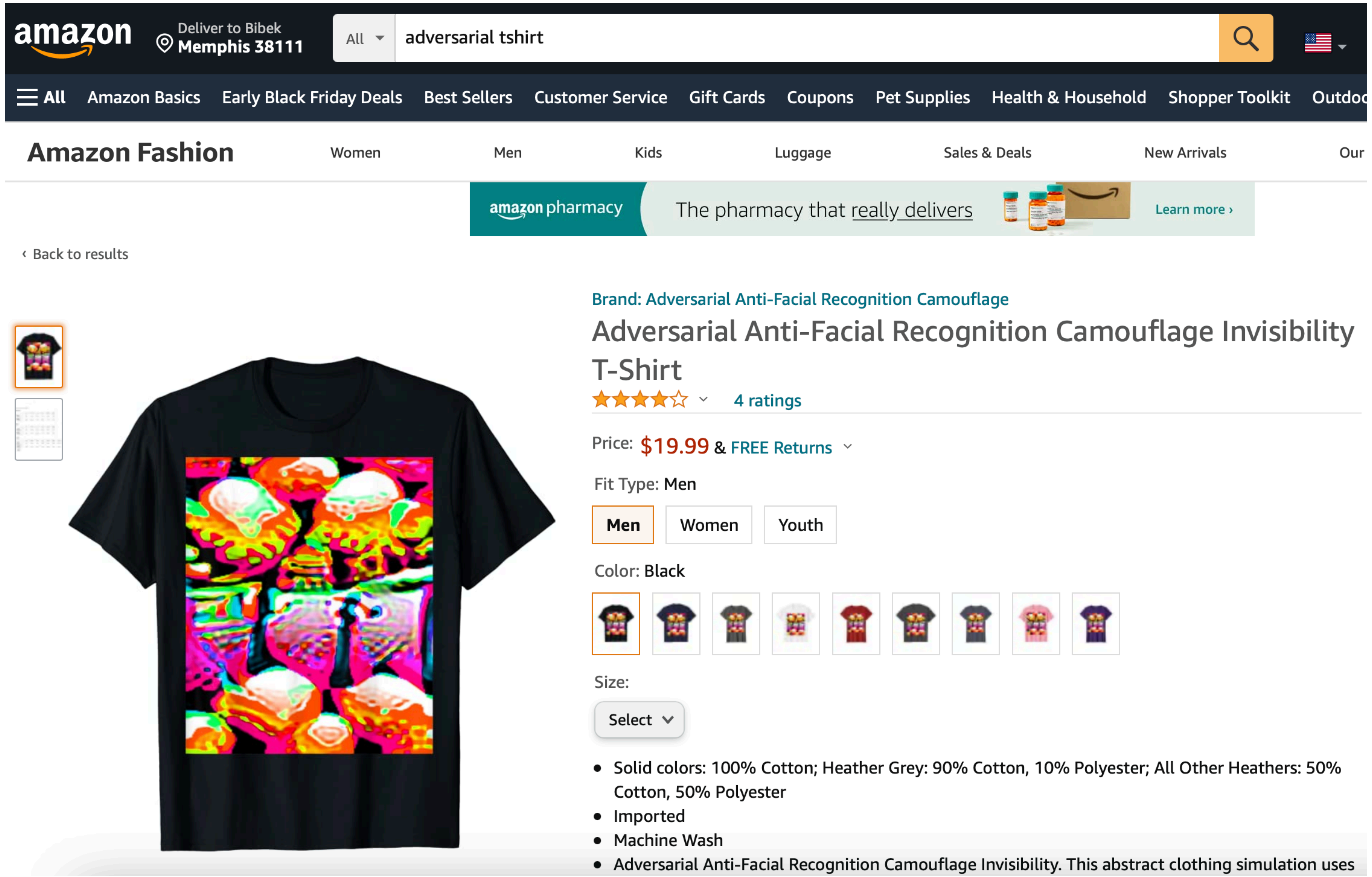
Adversarial examples in action

Adversarial examples in action

- Autonomous driving and traffic signs
- Video

Adversarial examples in action

- Surveillance, facial recognition
- Video



Adversarial examples in action

- Reinforcement learning
- Video

Optimization problem

The optimization problem

- "Lp norm" distance metric

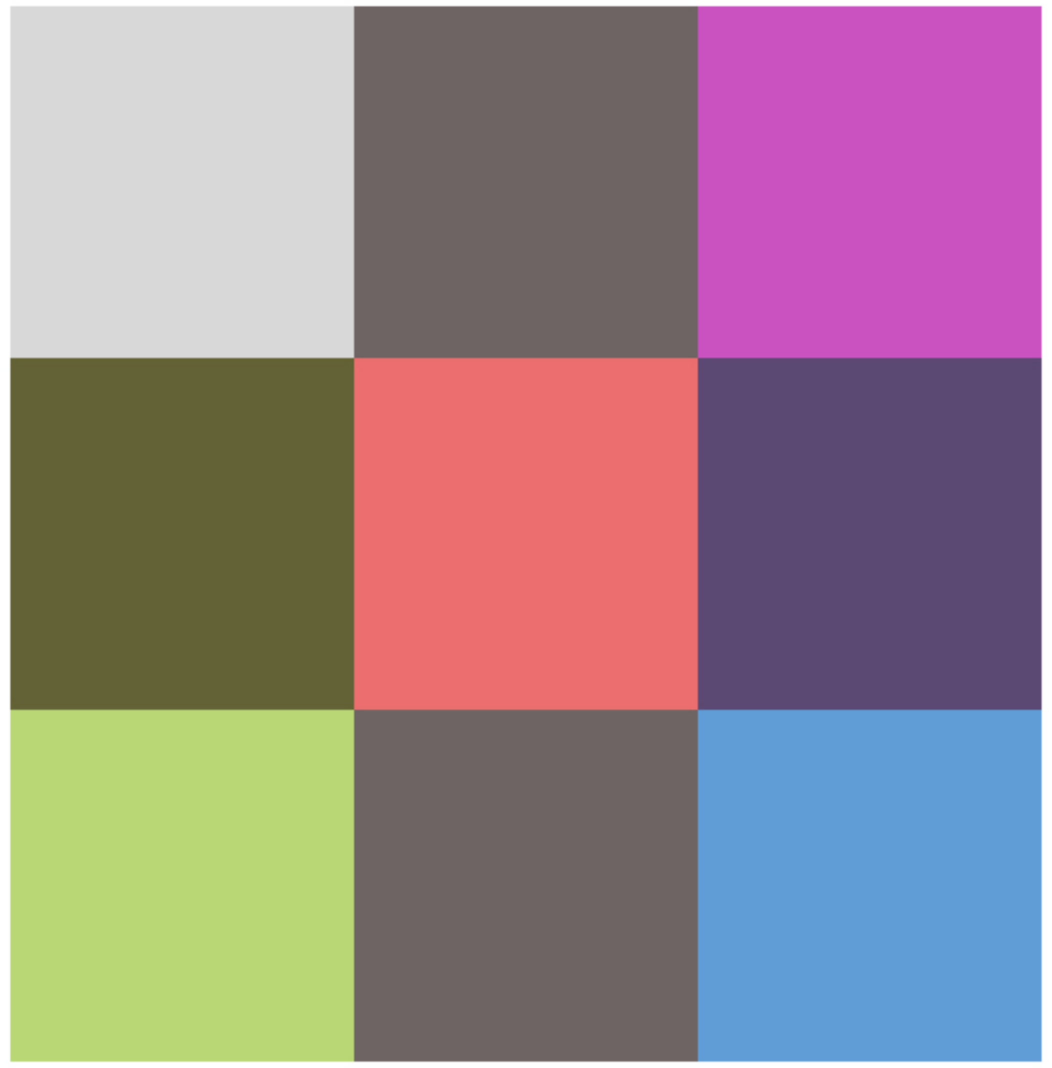


Image 1

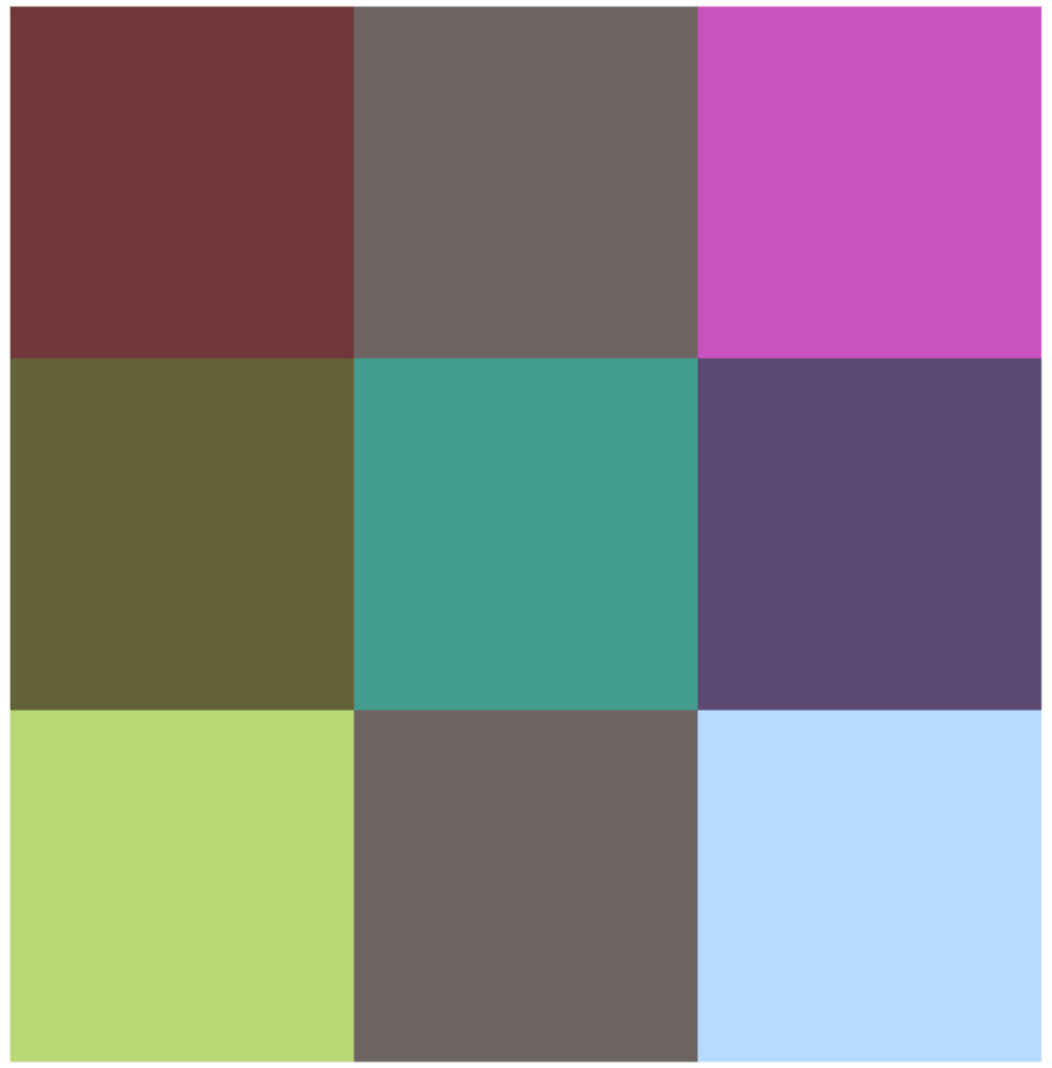


Image 2

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

$$
\begin{aligned}
&\text{minimize } D(x,\ x+\delta_x) \\
&\text{subject to:} \\
&\qquad f(x) \neq f(x+\delta_x) \\
&\qquad x+\delta_x \in [0,1]^n
\end{aligned}
$$

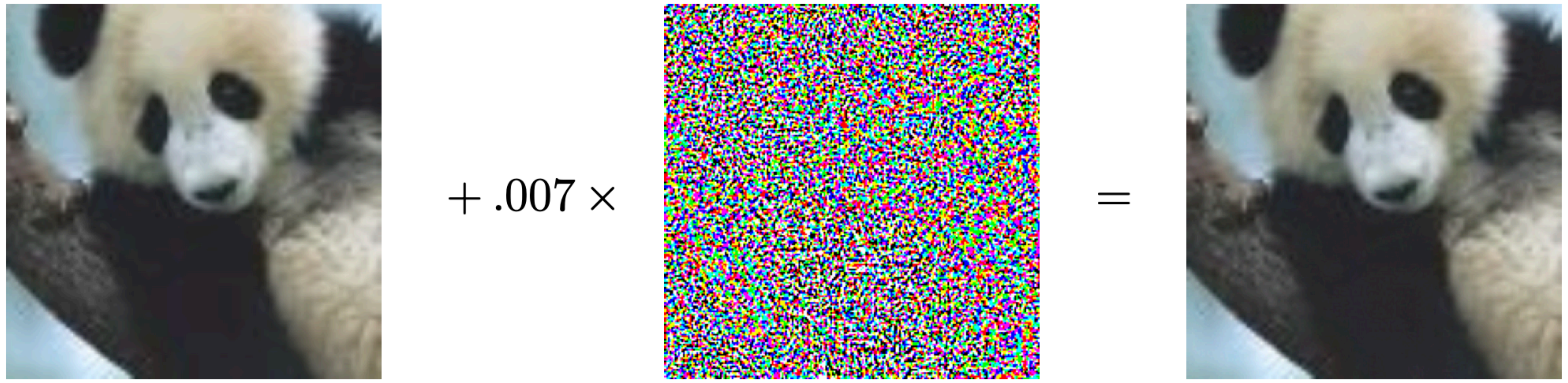
Attacks

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



$\boldsymbol{x}$ "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

Attacks

- Fast Gradient Sign Method (FGSM)

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

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

Attacks

- Fast Gradient Sign Method (FGSM)

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

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \ \mathrm{sign}(\nabla_x J(\underbrace{\theta}_{\text{Model parameters}}, \underbrace{x}_{\text{Input image}}, \underbrace{y}_{\text{Target label}}))$$

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \,\mathrm{sign}(\nabla_x \overbrace{J(\underbrace{\theta}_{\text{Model parameters}}, \underbrace{x}_{\text{Input image}}, \underbrace{y}_{\text{Target label}})}^{\text{Loss value}})$$

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \; \mathrm{sign}(\overbrace{\nabla_x}^{\text{Gradient w.r.t. input}} \overbrace{J(\underbrace{\theta}_{\text{Model parameters}}, \underbrace{x}_{\text{Input image}}, \underbrace{y}_{\text{Target label}})}^{\text{Loss value}})$$

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \epsilon \, \underbrace{\mathrm{sign}}_{\text{Just take the sign}}\Big(\overbrace{\nabla_x}^{\text{Gradient w.r.t. input}} \overbrace{J(\underbrace{\theta}_{\text{Model parameters}}, \underbrace{x}_{\text{Input image}}, \underbrace{y}_{\text{Target label}})}^{\text{Loss value}}\Big)$$

Attacks

- Fast Gradient Sign Method (FGSM)

$$\delta = \underbrace{\epsilon}_{\text{Scale}} \, \underbrace{\mathrm{sign}}_{\text{Just take the sign}}\Big(\overbrace{\nabla_x}^{\text{Gradient w.r.t. input}} \overbrace{J(\underbrace{\theta}_{\text{Model parameters}}, \underbrace{x}_{\text{Input image}}, \underbrace{y}_{\text{Target label}})}^{\text{Loss value}}\Big)$$

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

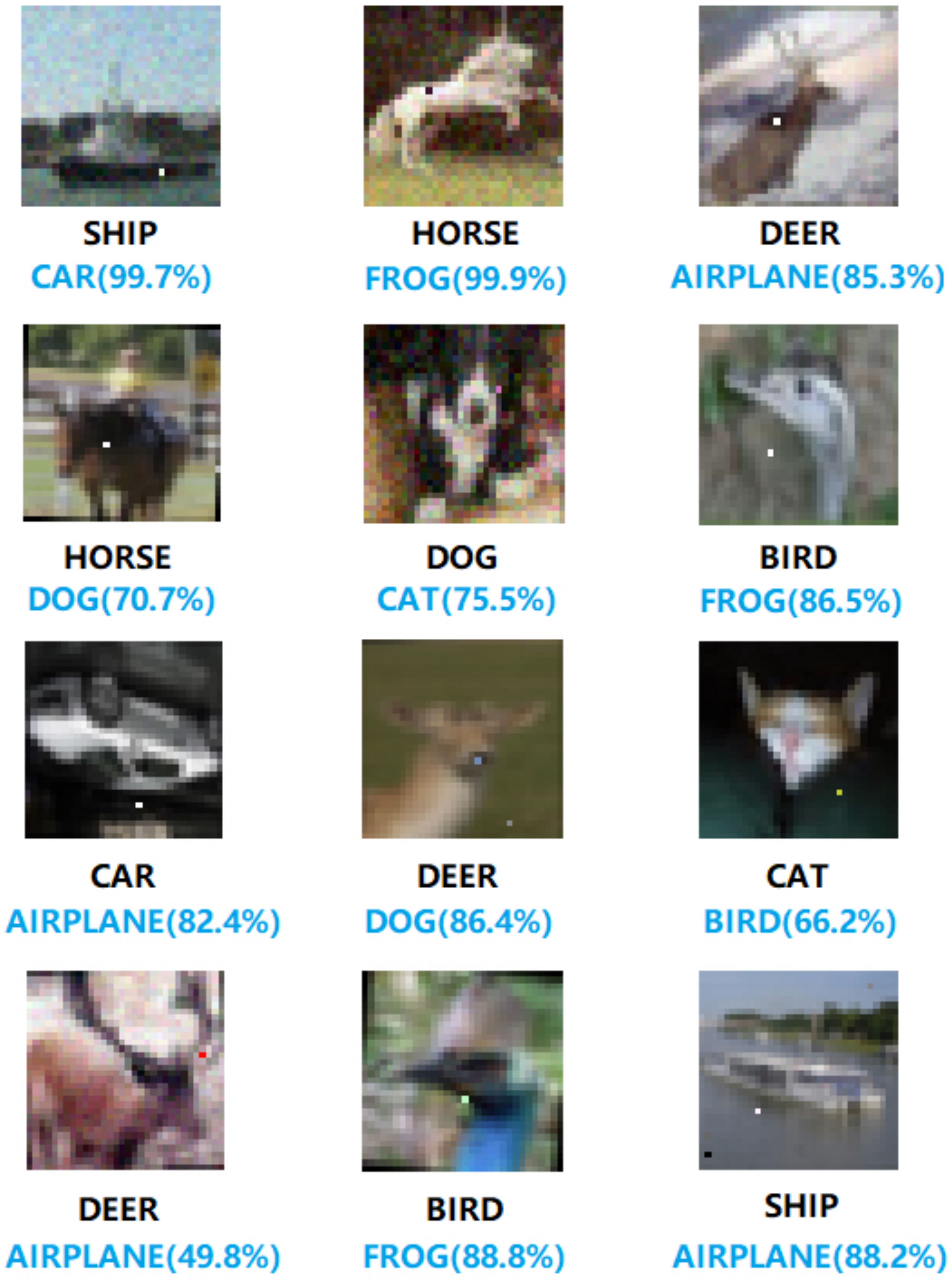
Input

FGSM

PGD

Attacks

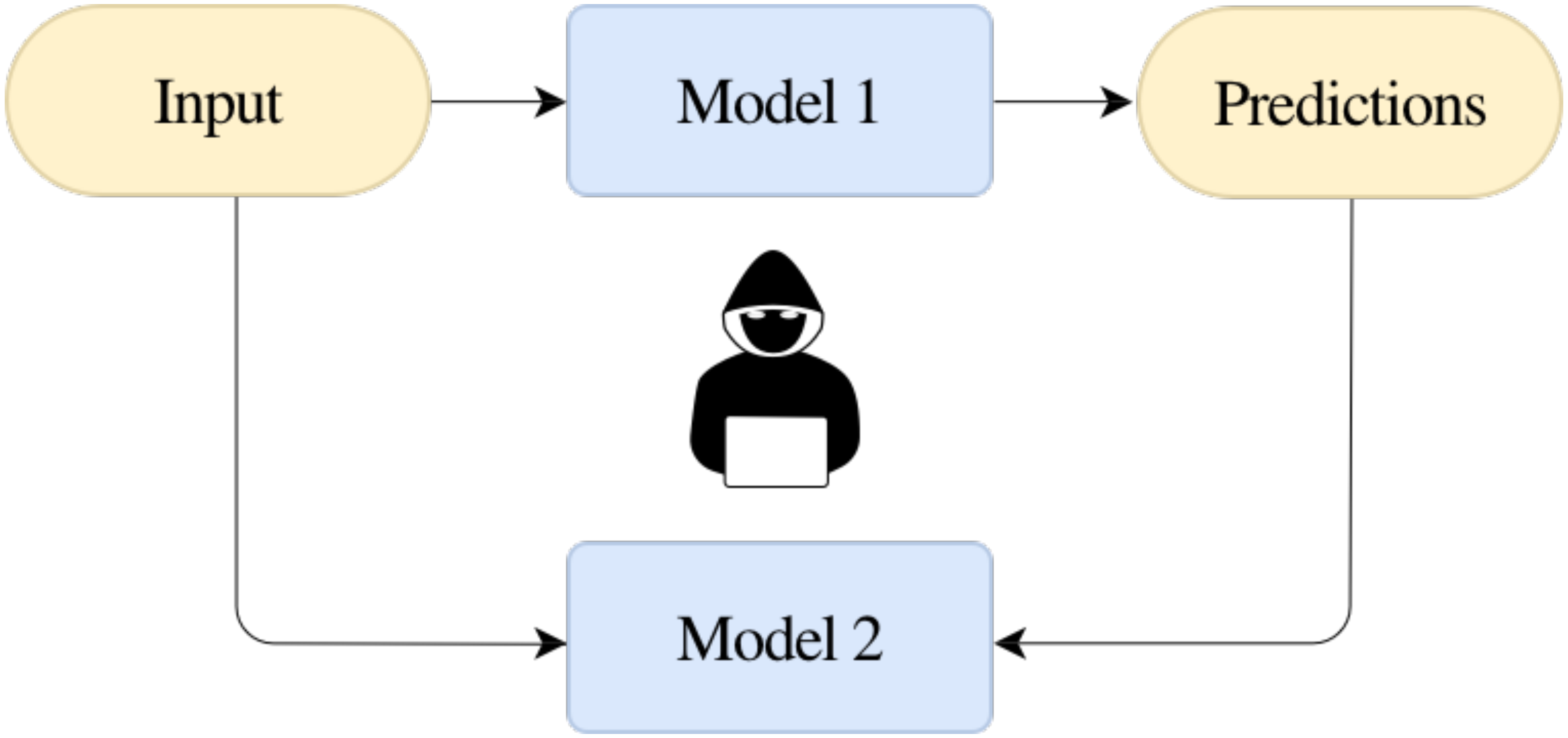
- One pixel attack



Threat models

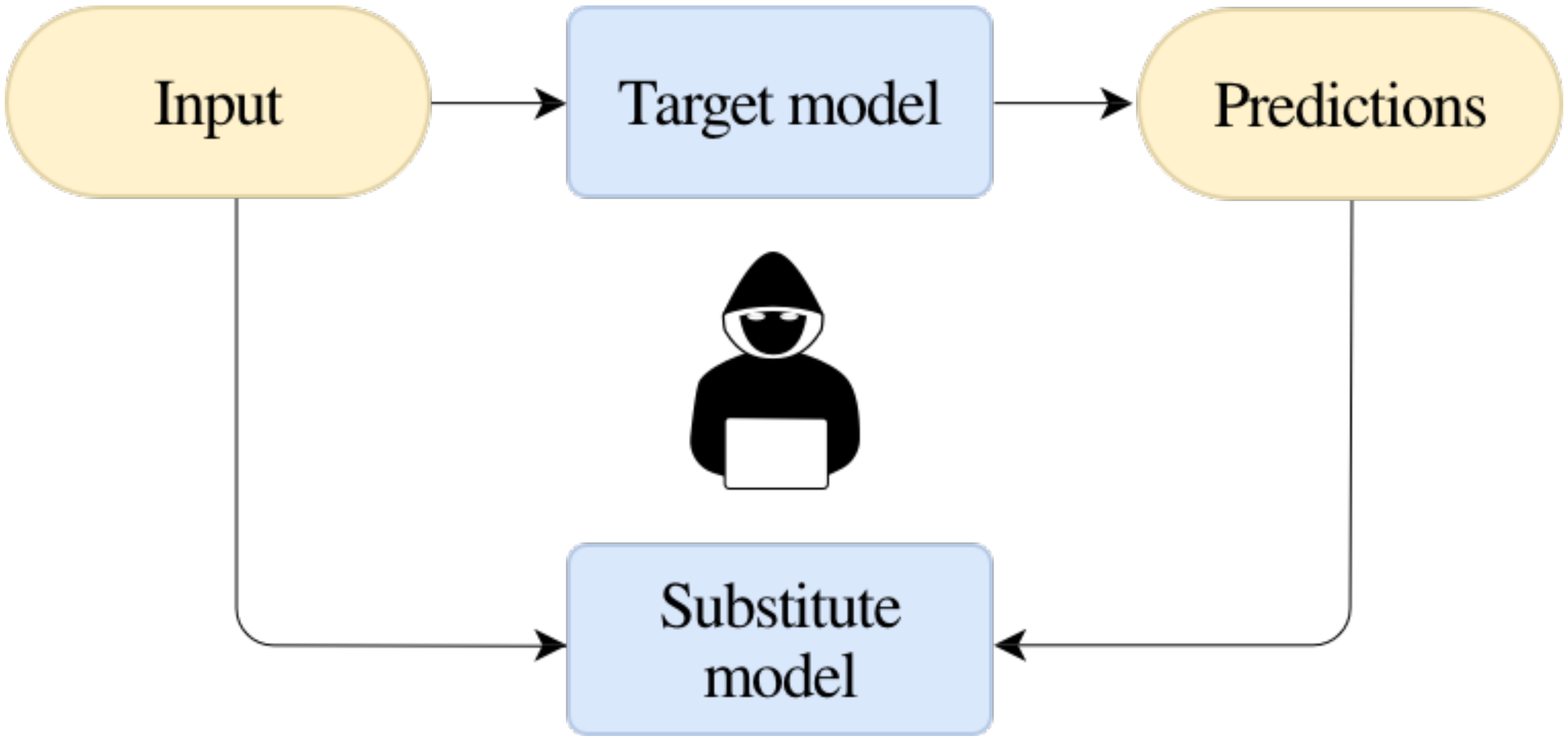
Threat models

- Black-box



Threat models

- Black-box



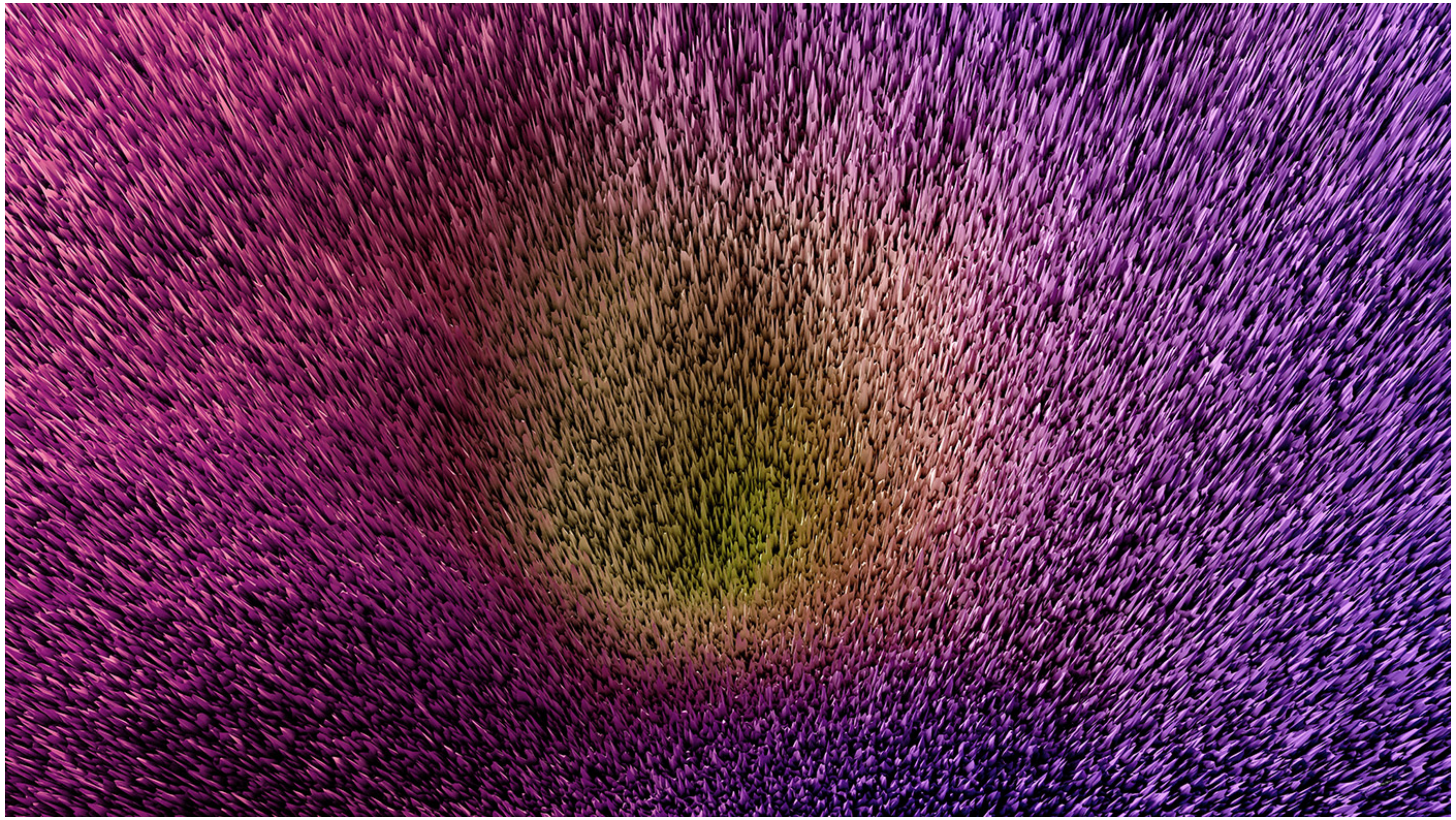
Threat models

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

Defenses

Defenses

- Gradient Masking
  - Hide gradient information
  - Discarded



Defenses

- Adversarial Training
  - Most succesful

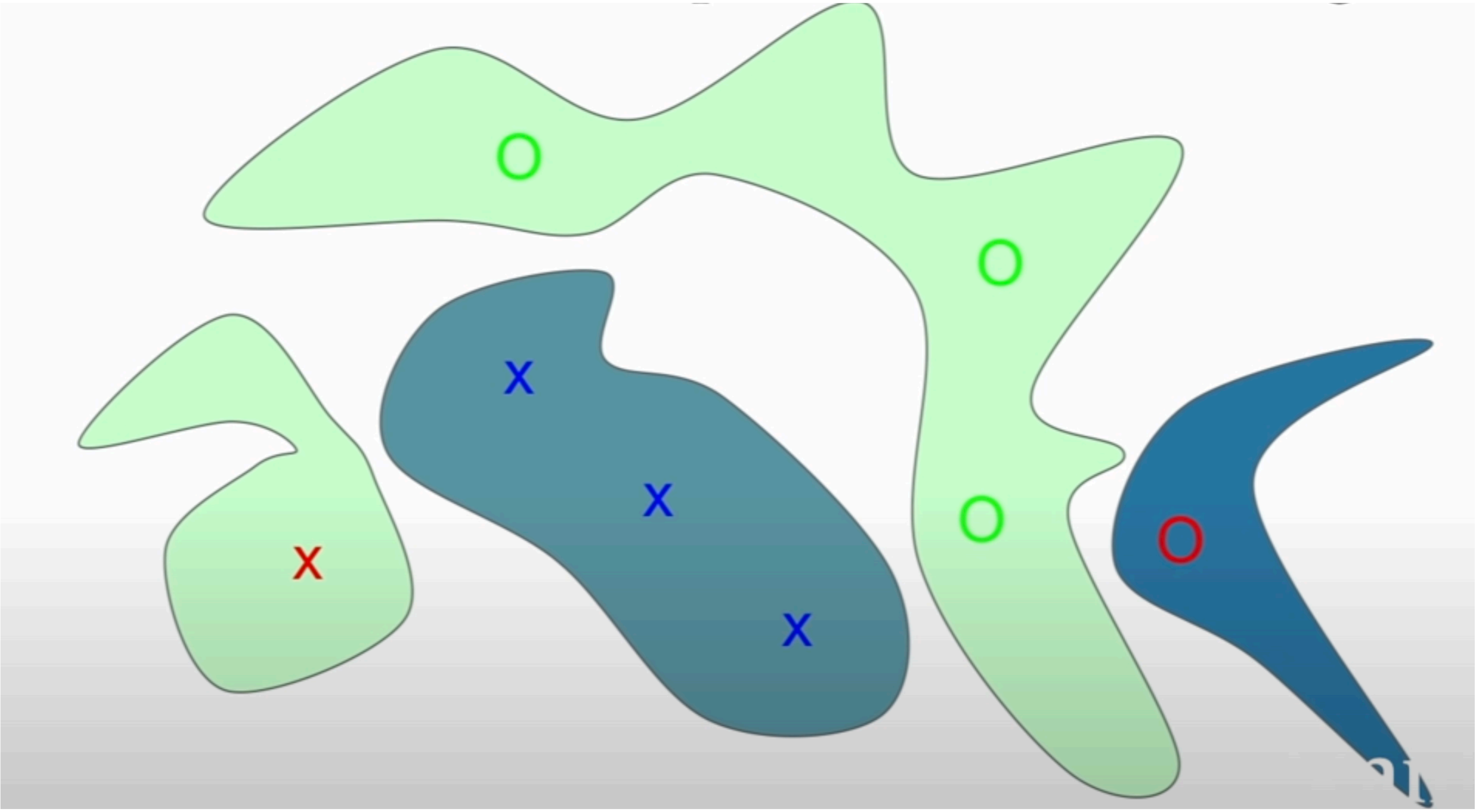
Theories

Theories

- Intuitively make sense but discarded
    - Overfitting

Theories

- Intuitively make sense but discarded
  - Overfitting

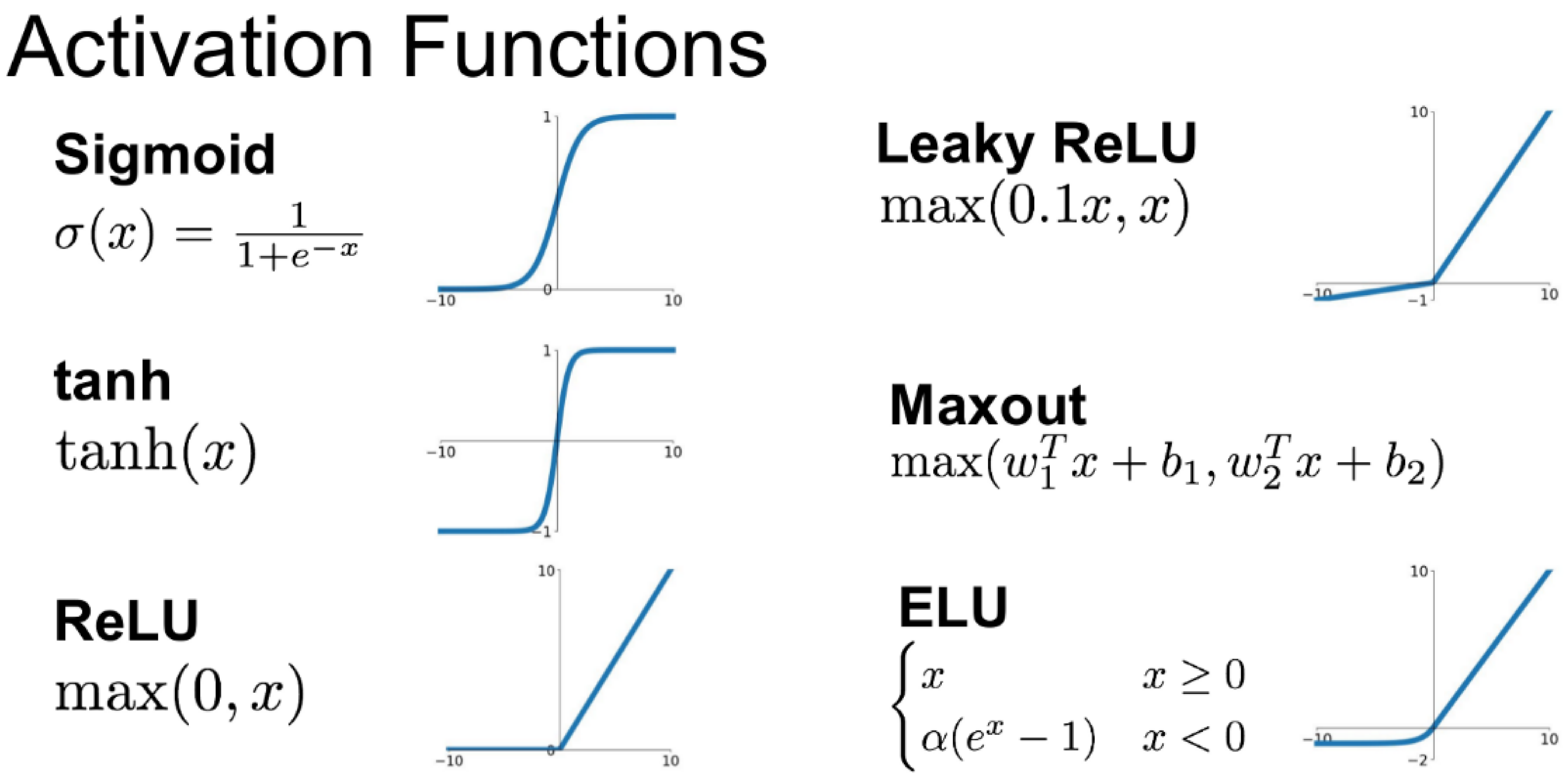


Theories

- Intuitively make sense but discarded
  - Excessive linearity

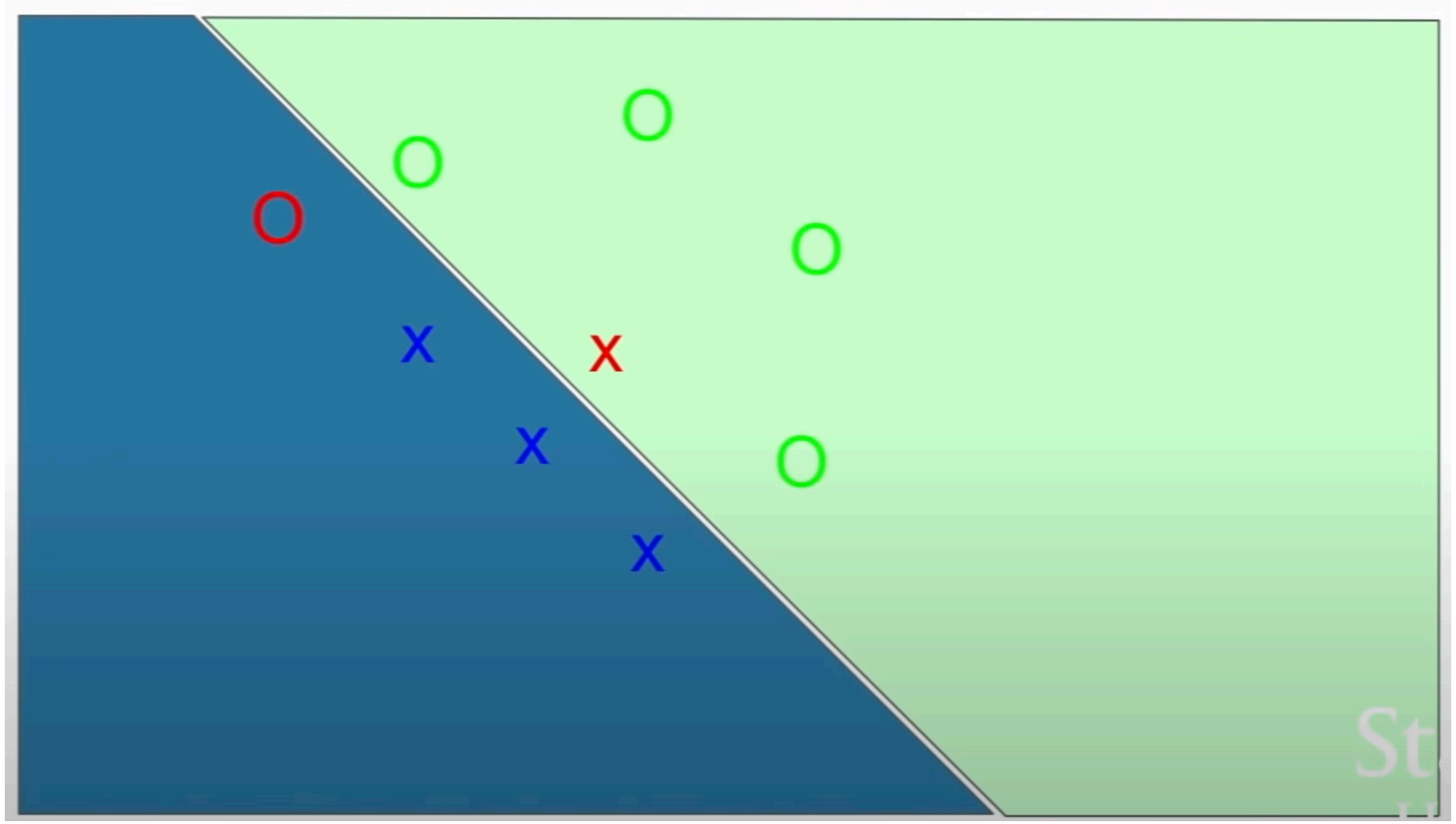
Theories

- Intuitively make sense but discarded
  - Excessive linearity



Theories

- Intuitively make sense but discarded
  - Excessive linearity

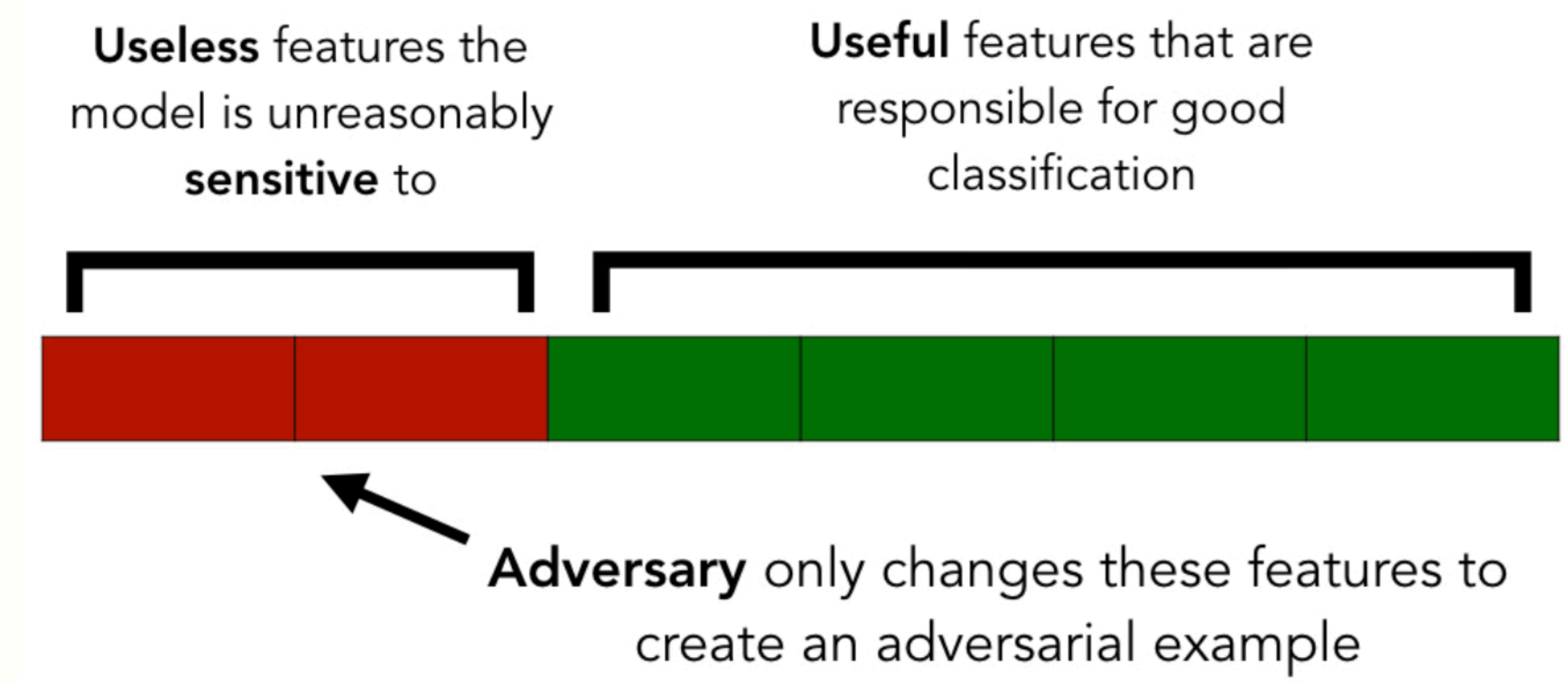


Theories

- Intuitively make sense but discarded
  - Adversarial examples are bugs

Theories

- Intuitively make sense but discarded
  - Adversarial examples are bugs



Theories

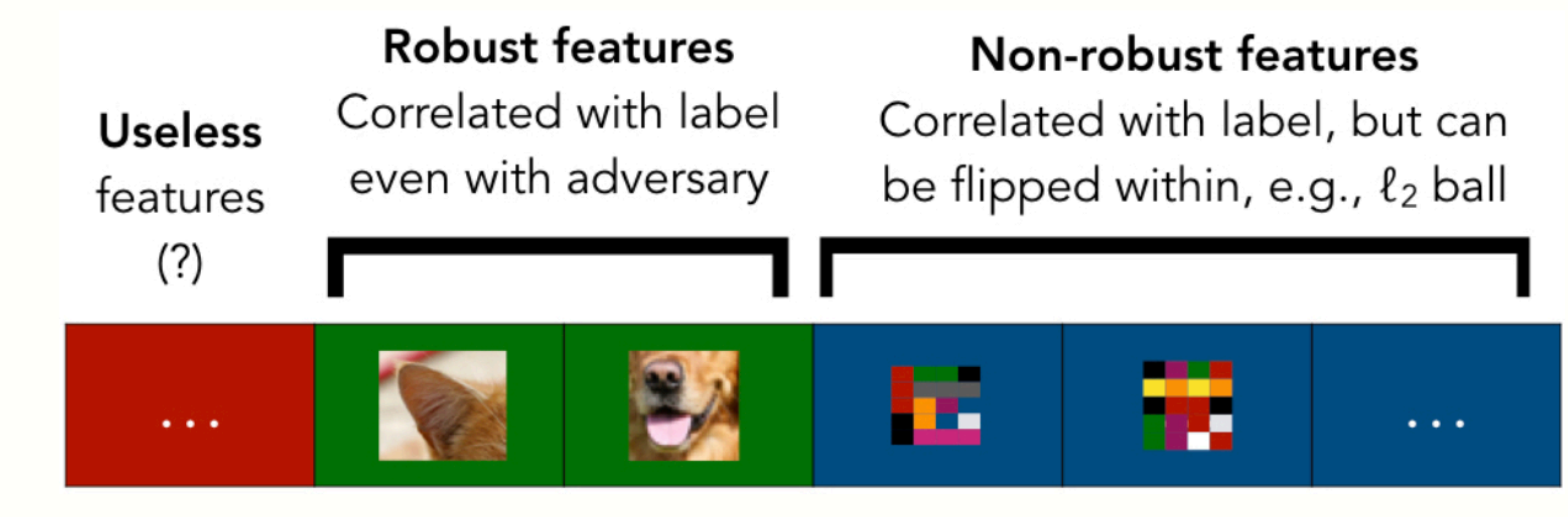
- Widely accepted

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