CSCI1470

Deep Learning

Randall Balestriero

Recap

Recap

- Model-Free vs Model-Based
- On-Policy vs Off-Policy
- Value-Based vs Policy-Based (Actor-Critic)
- Reconstruction-Based vs JEPAs

“panda”

“panda”

?

“panda”

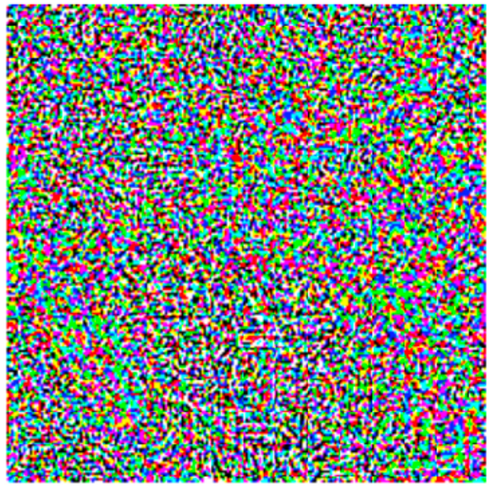
“gibbon”

“panda”

“gibbon”

Why?

$+ .007 \times$



$=$

“panda”

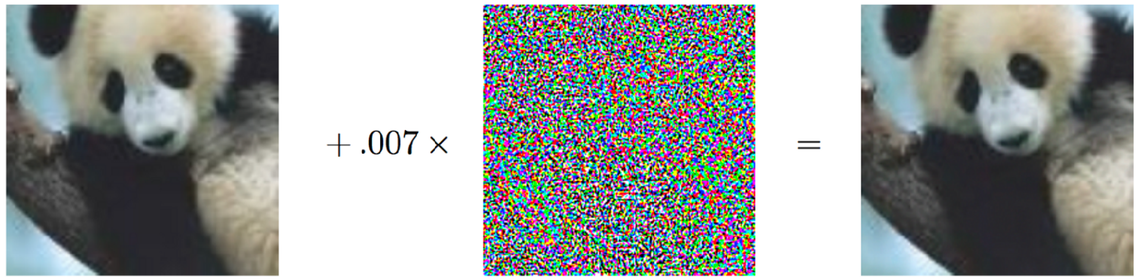
57.7% confidence

noise

“gibbon”

99.3% confidence

How to find that “noise”?



“panda”

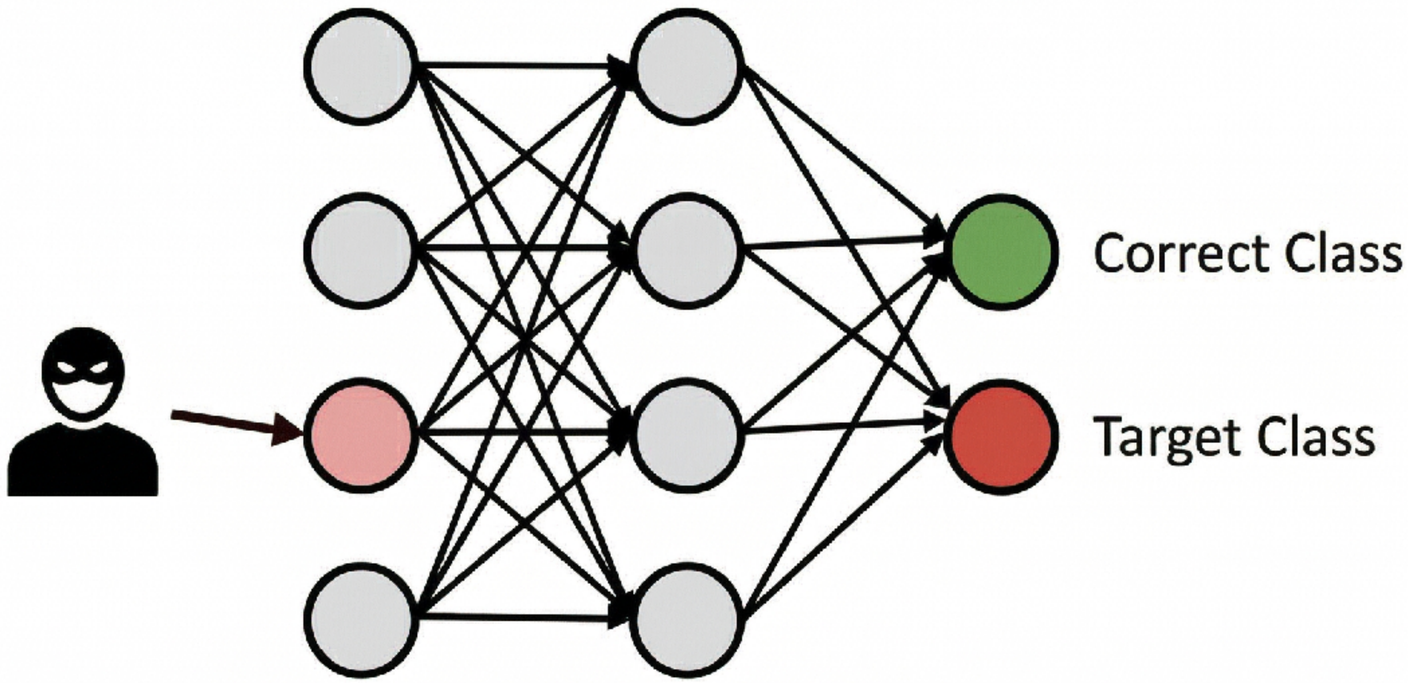
57.7% confidence

noise

“gibbon”

99.3% confidence

Adversarial Examples in AI



The Core Concept

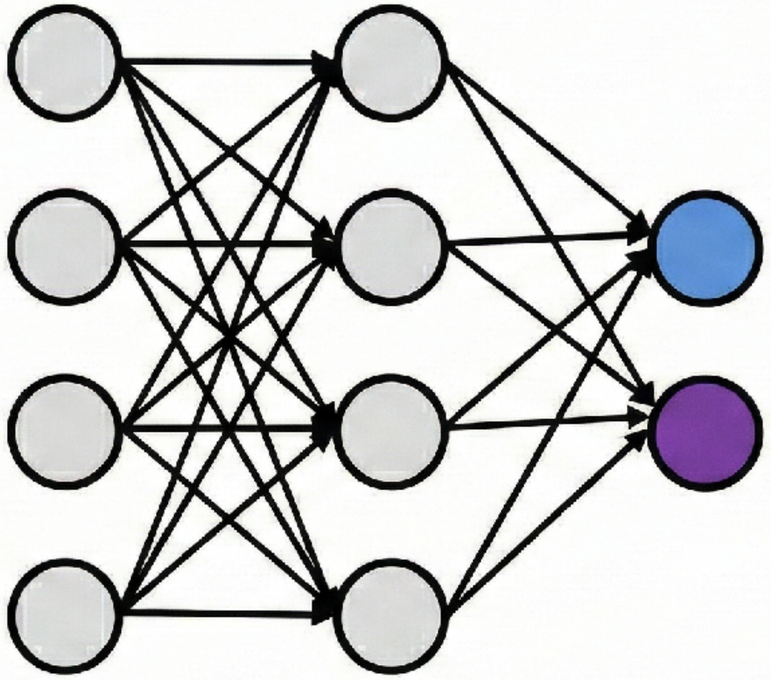
Classification
Confidence High

Original input

+

Adversarial

noise

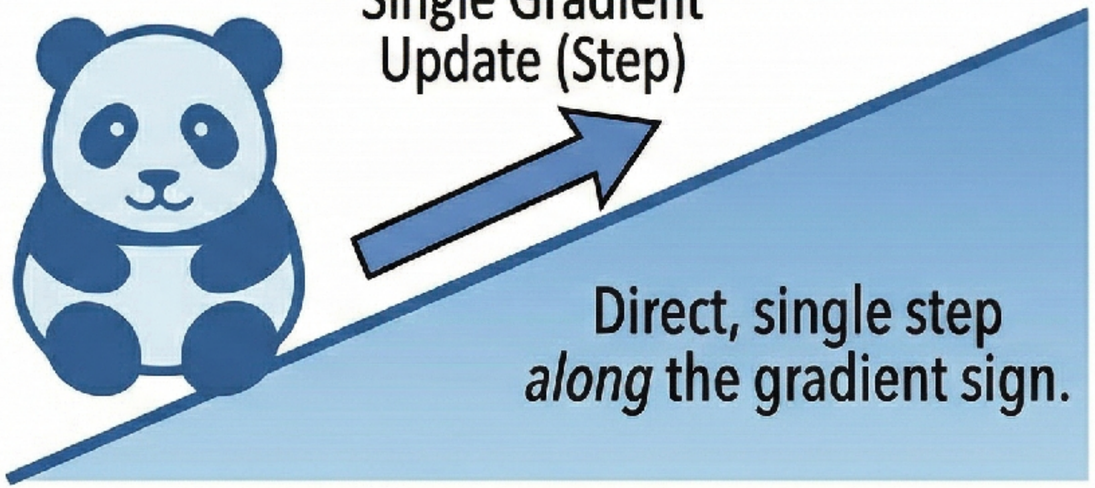


Misclassified: gibbon
(confidence is High)

adversarial example

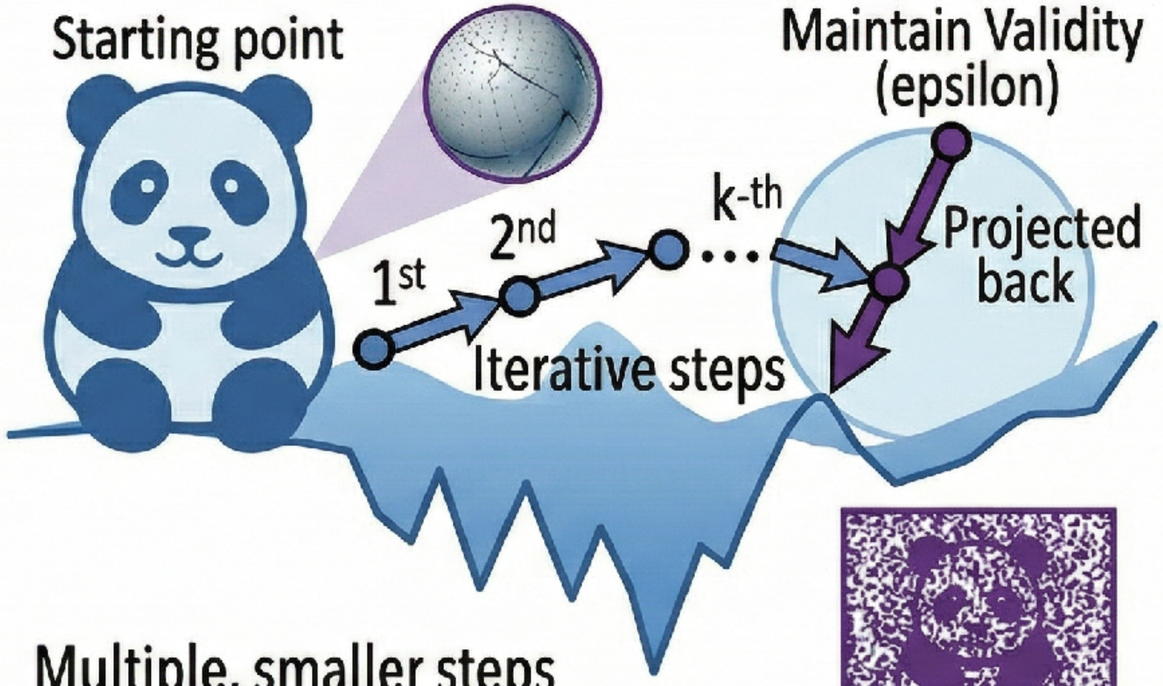
Comparing Adversarial Attacks: FGSM vs. PGD (A Visual Guide).

1. FGSM (Fast Gradient Sign Method)



Gibbon
(High Confidence)

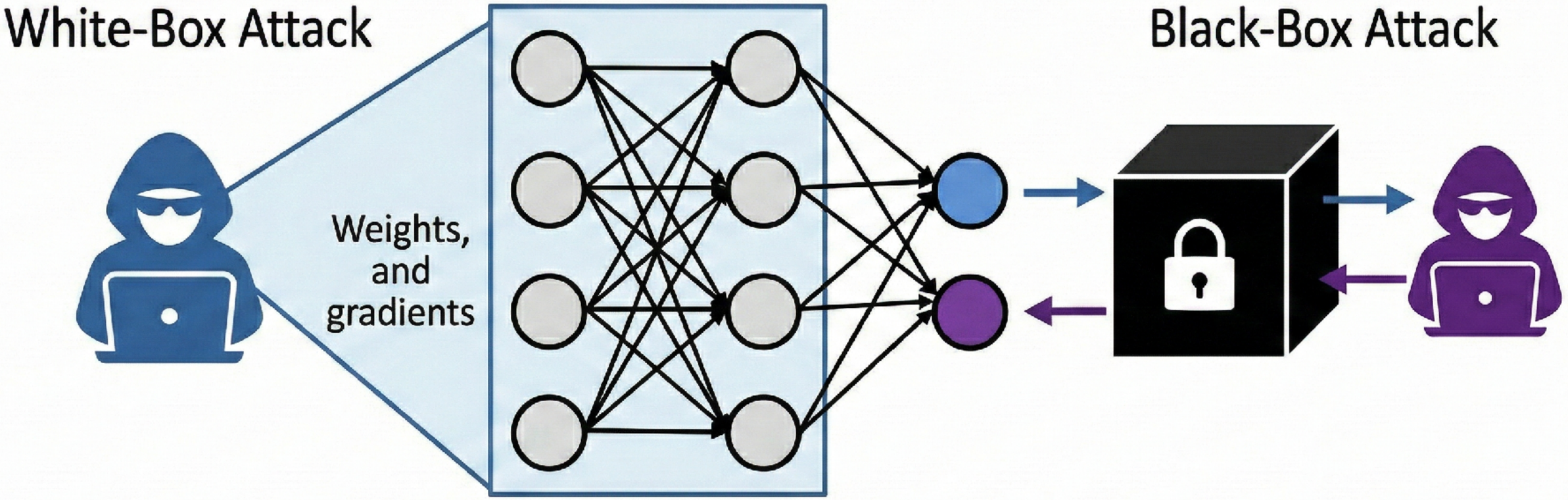
2. PGD (Projected Gradient Descent)



Multiple, smaller steps
guided by the gradient landscape.

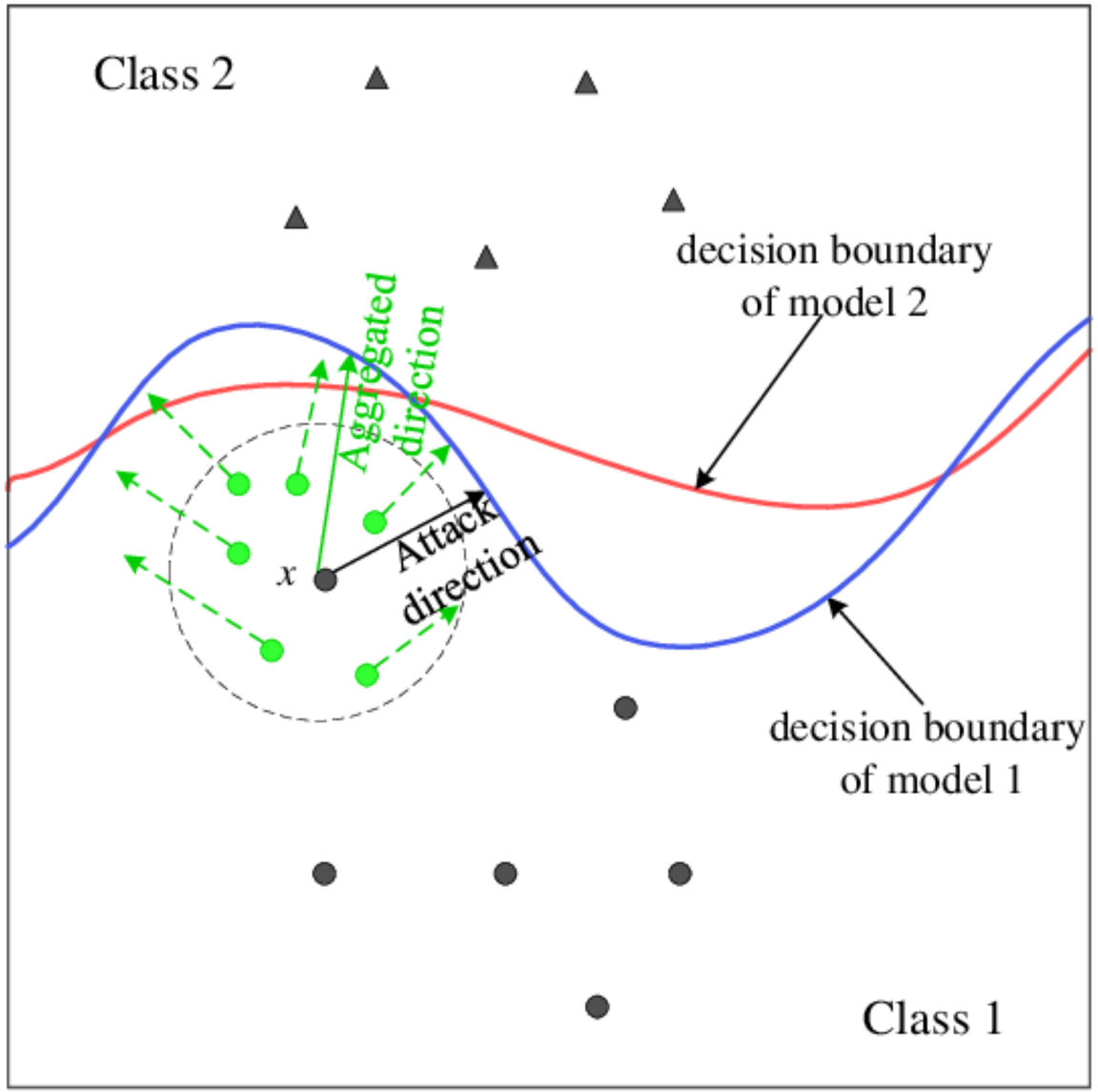
Incorrect Class
(Max Confidence)

Attack Taxonomy



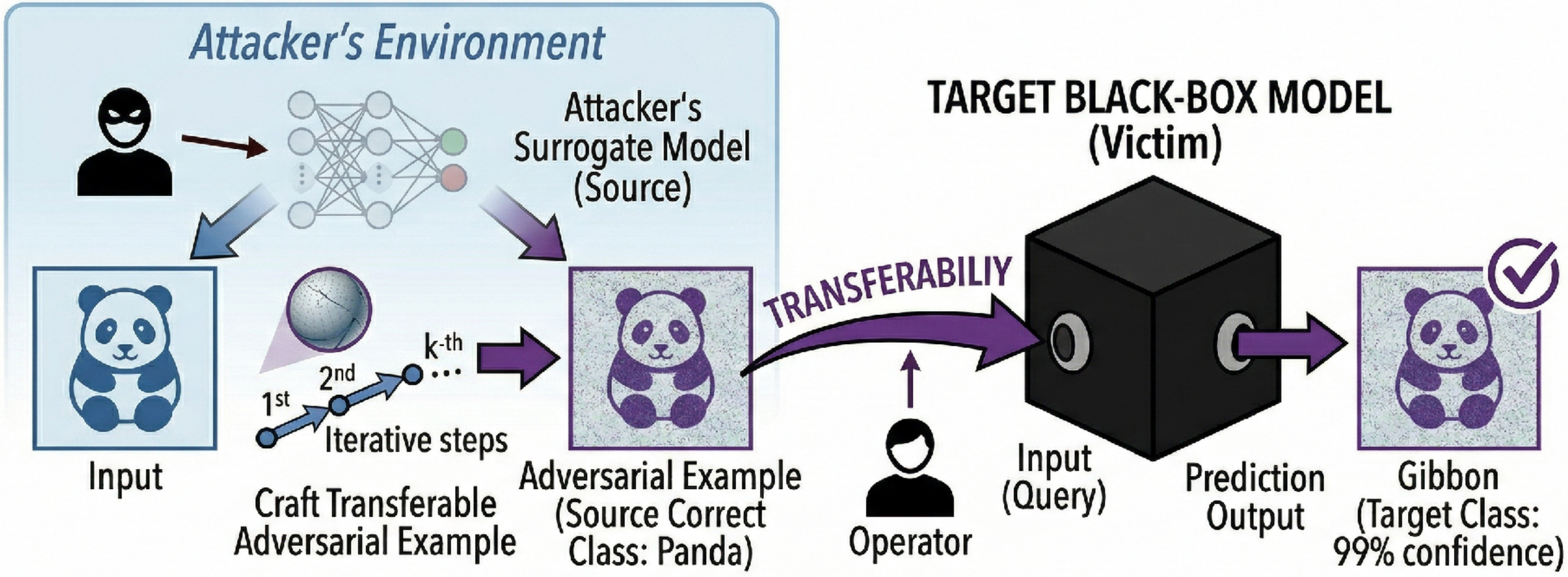
How to do a Black-Box attack?

Adversarial Attacks may Transfer!



How to do a Black-Box attack?

Black-Box Attack Illustration: Surrogate Model Transferability



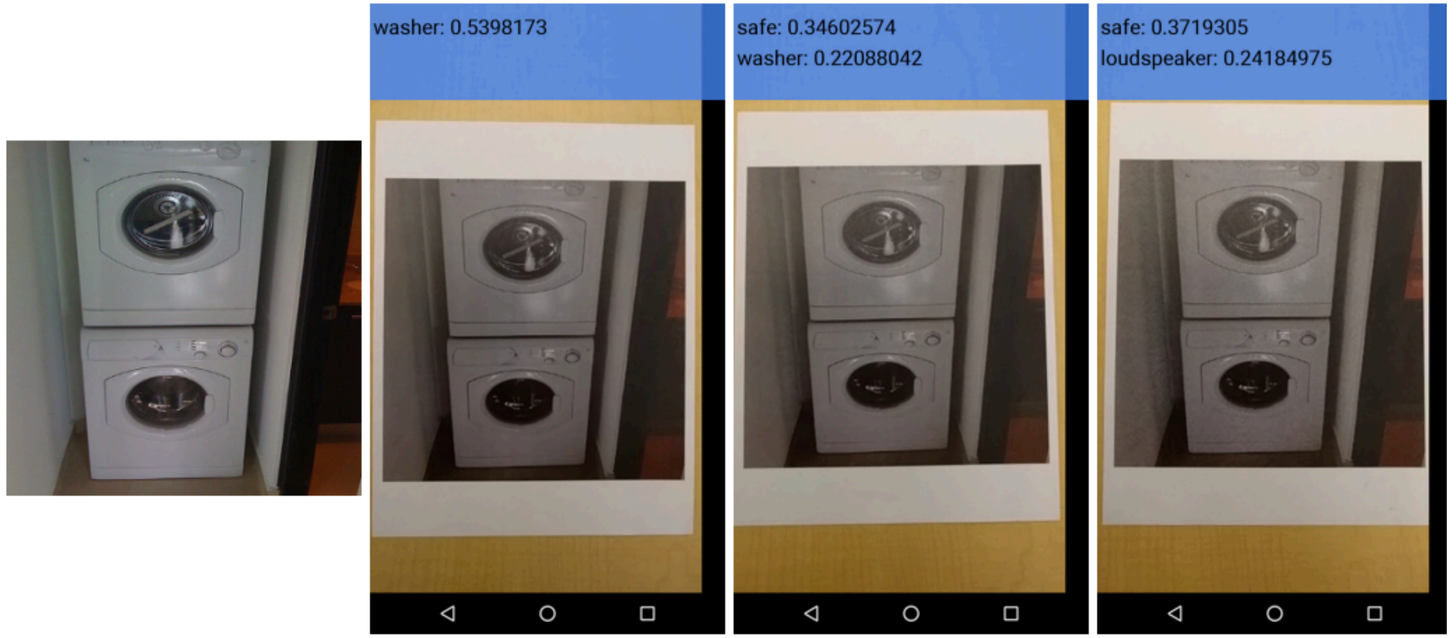
University Presentation

ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD

Alexey Kurakin
Google Brain
kurakin@google.com

Ian J. Goodfellow
OpenAI
ian@openai.com

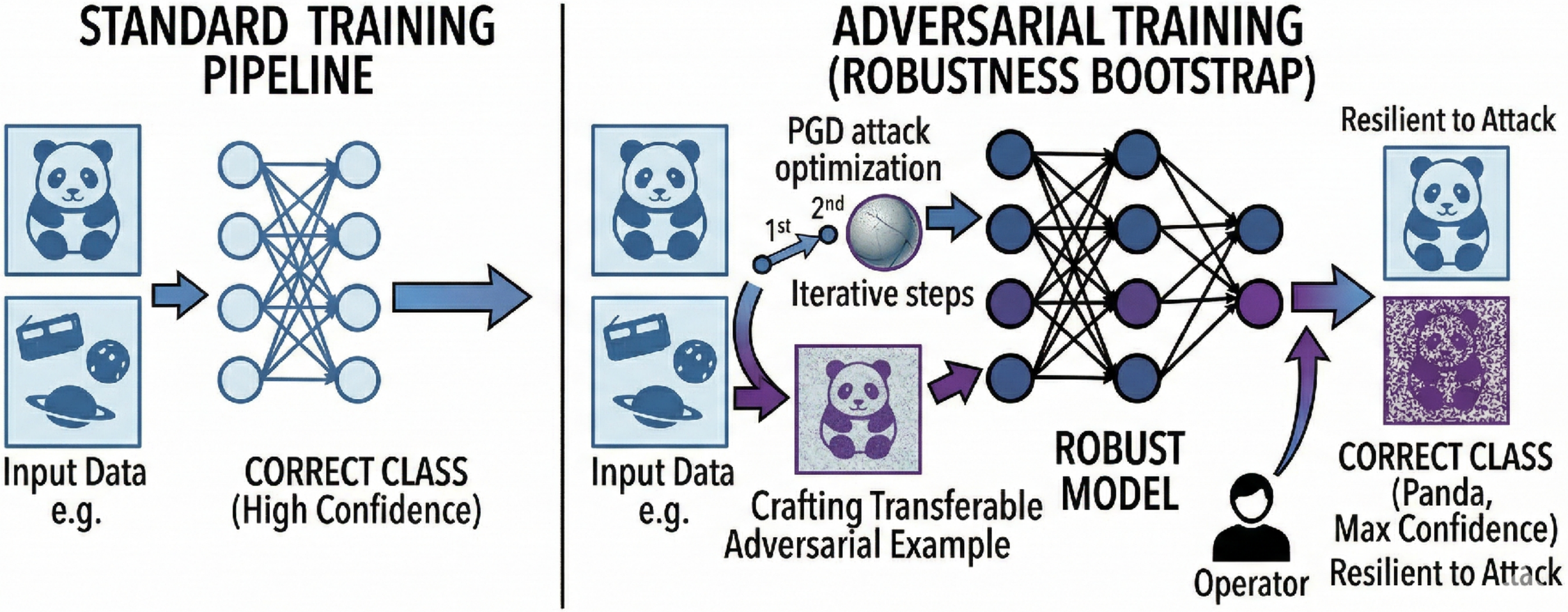
Samy Bengio
Google Brain
bengio@google.com



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

How to do protect ourselves?

ADVERSARIAL TRAINING (ROBUSTNESS BOOTSTRAP)



Thank you!
See you Friday!