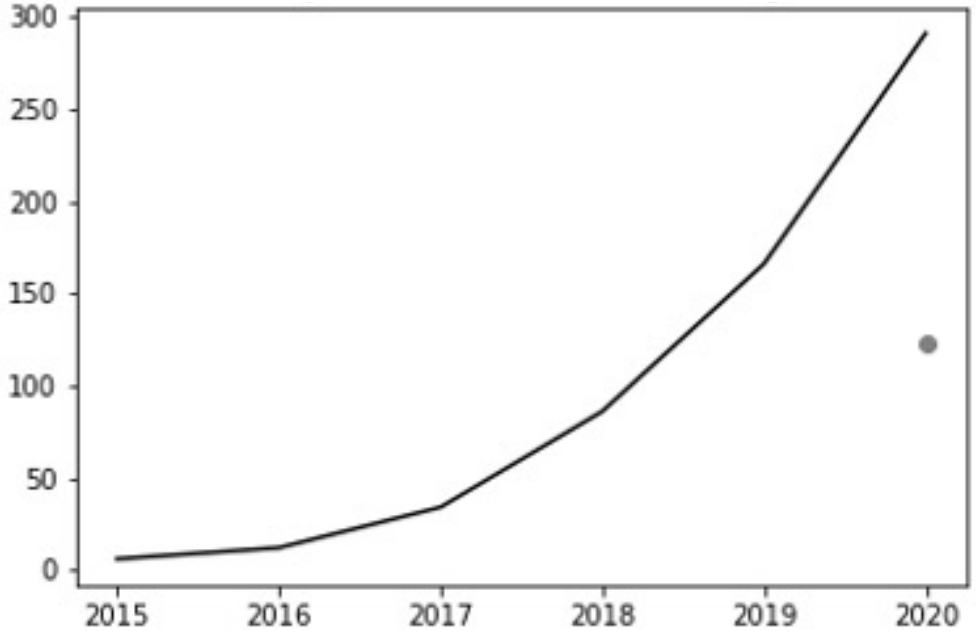
Machine learning security

Danila Yu. Emelyanov

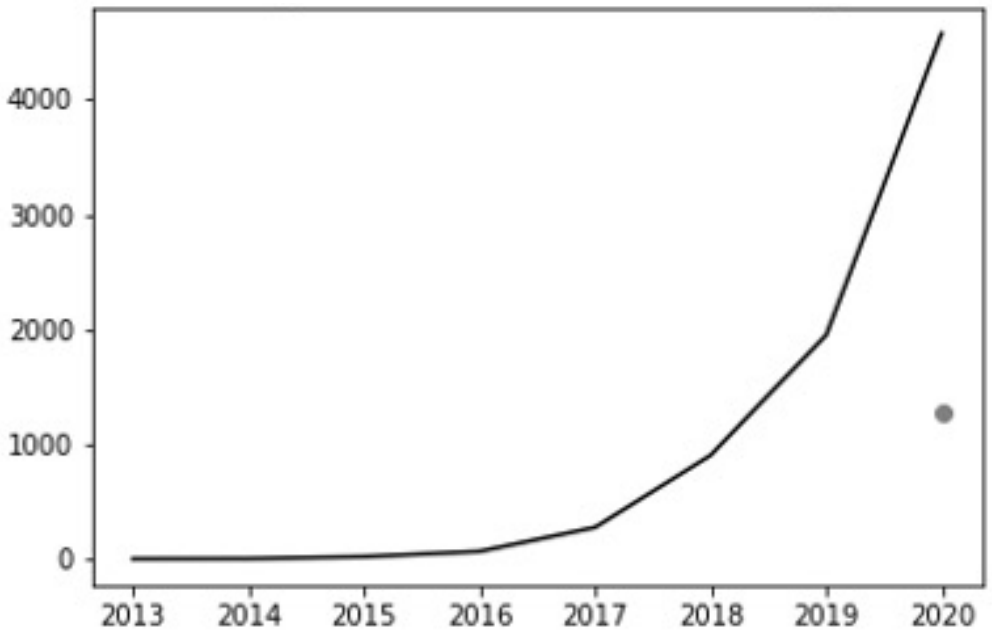
Lead mathematician, JSC “NPK Kryptonite”

Different search requests results count on arxiv.org

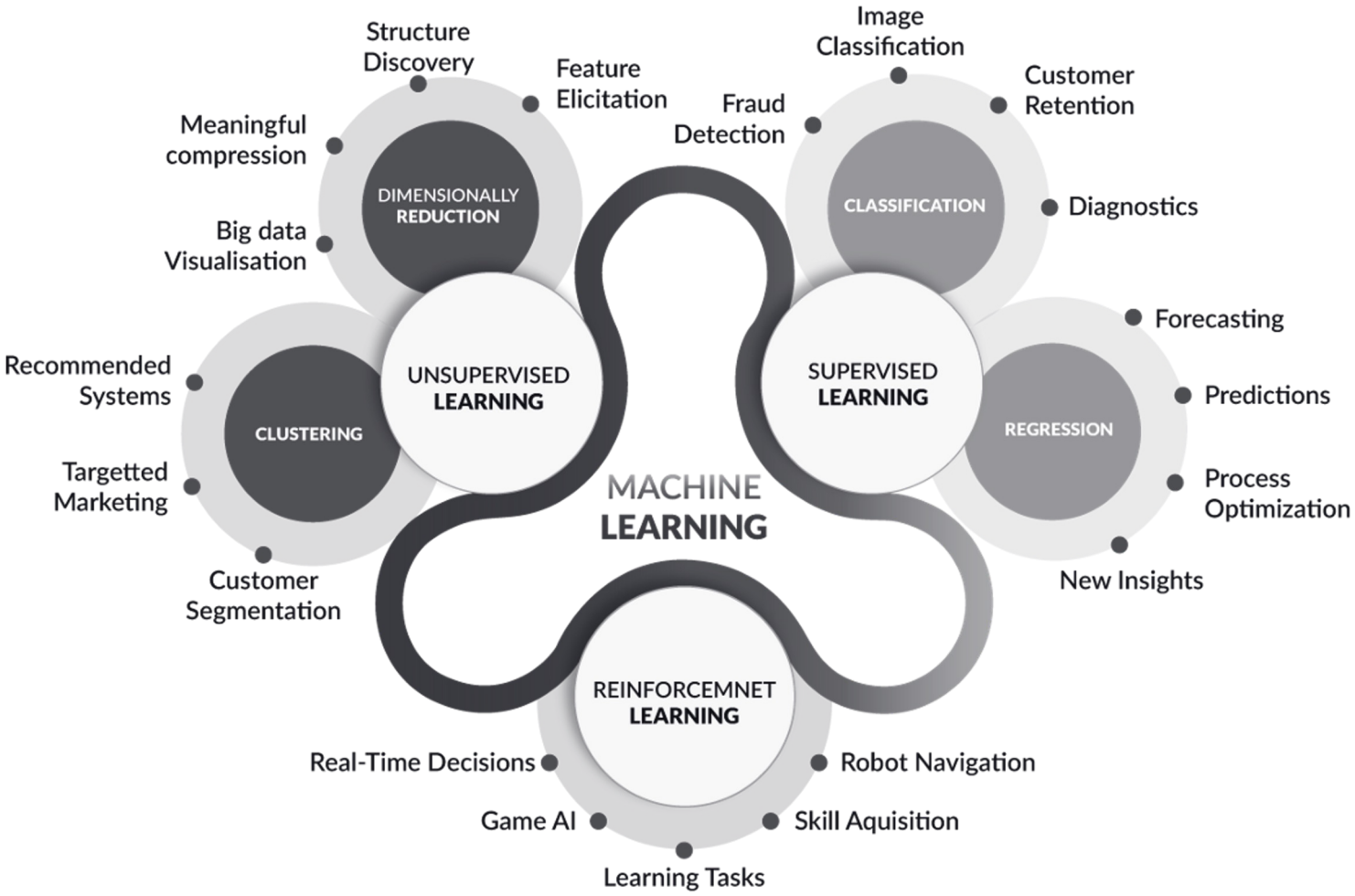
”threat machine learning”



”adversarial machine learning”

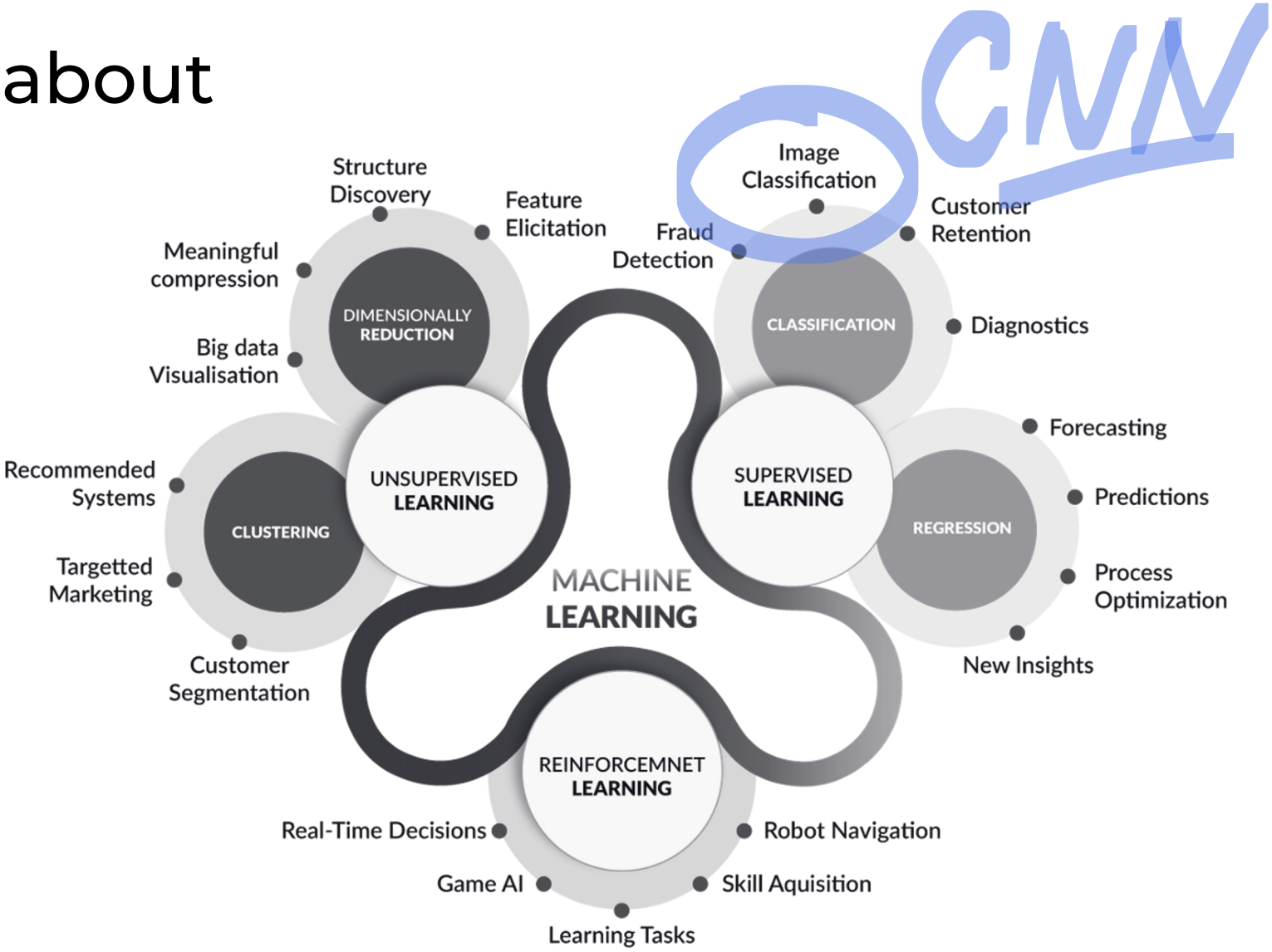


Let’s talk about

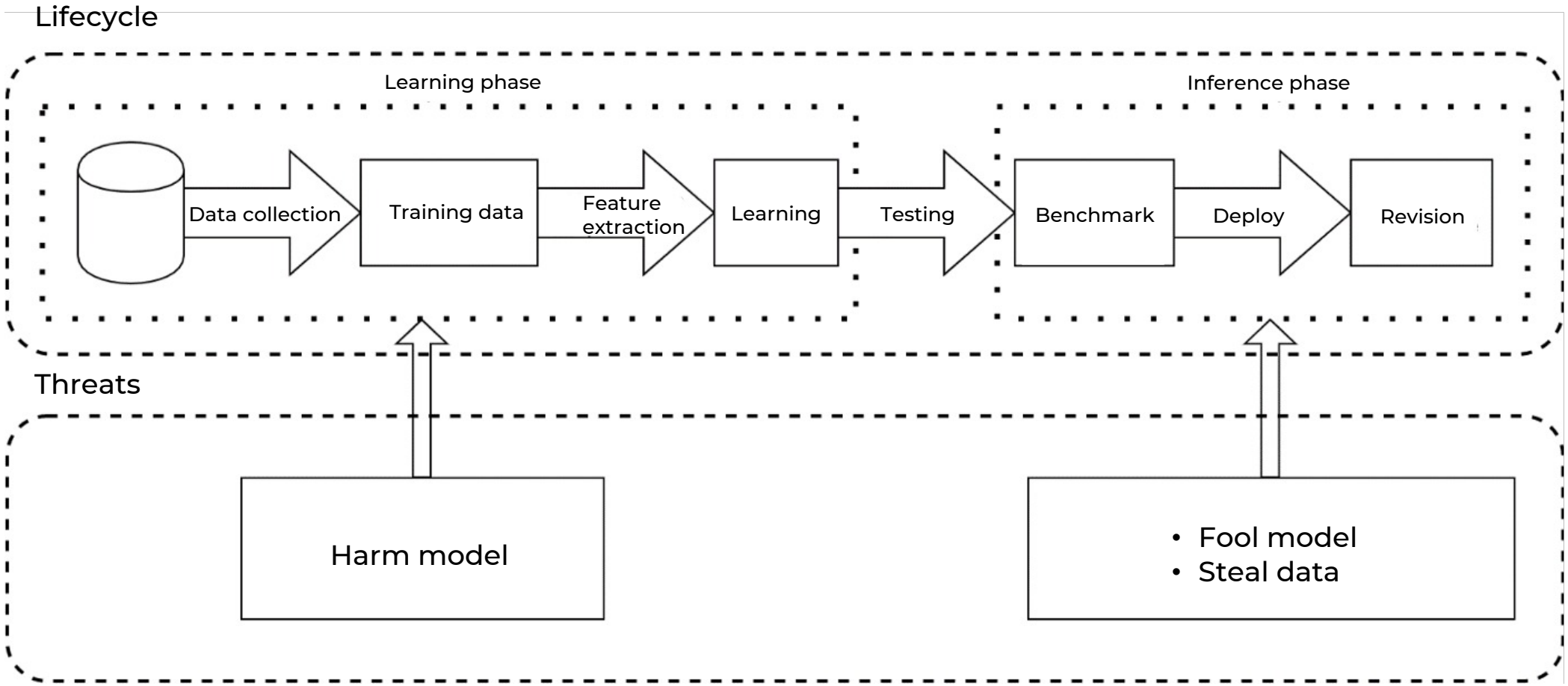


Let’s talk about

CNN-based supervised learning model for image classification

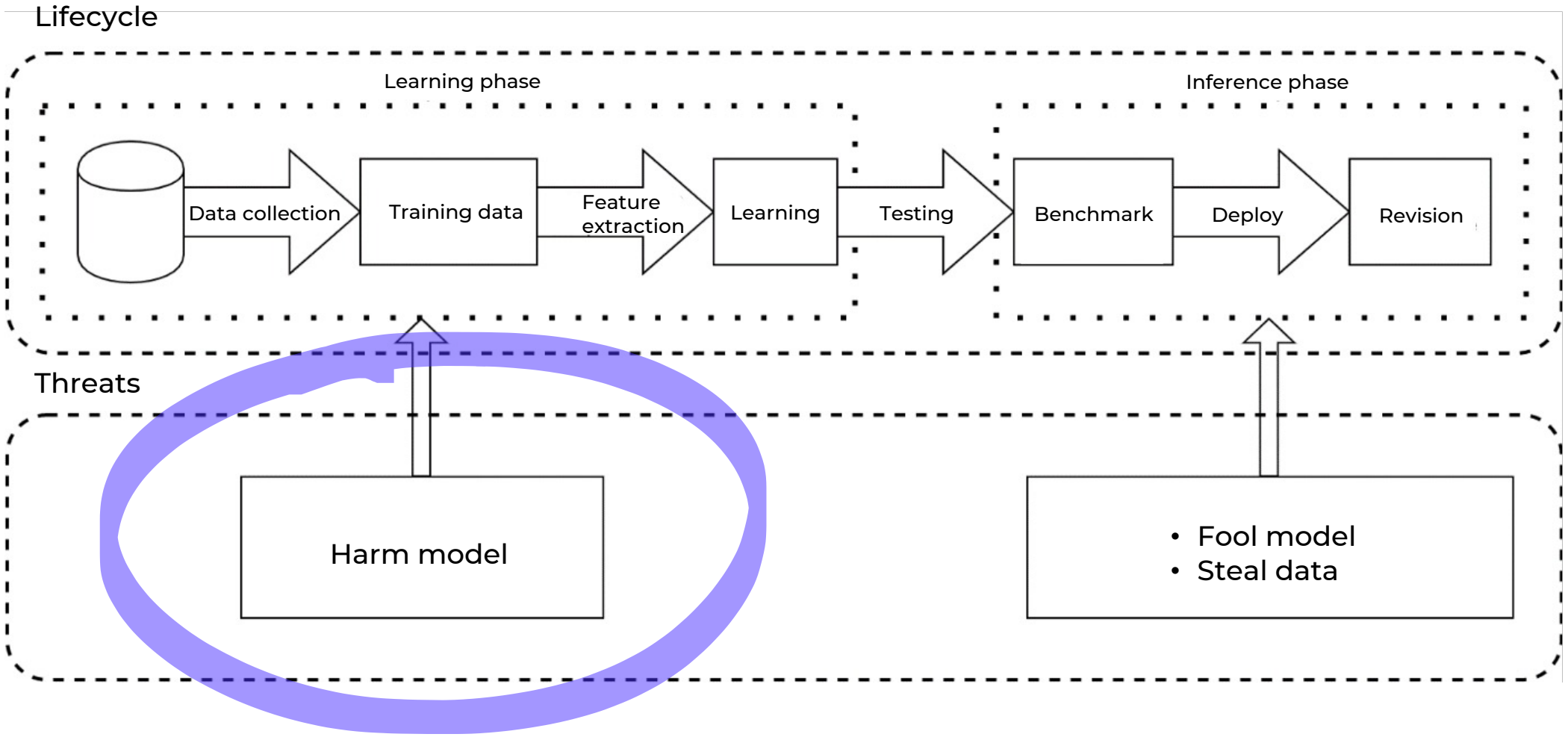


Threat model



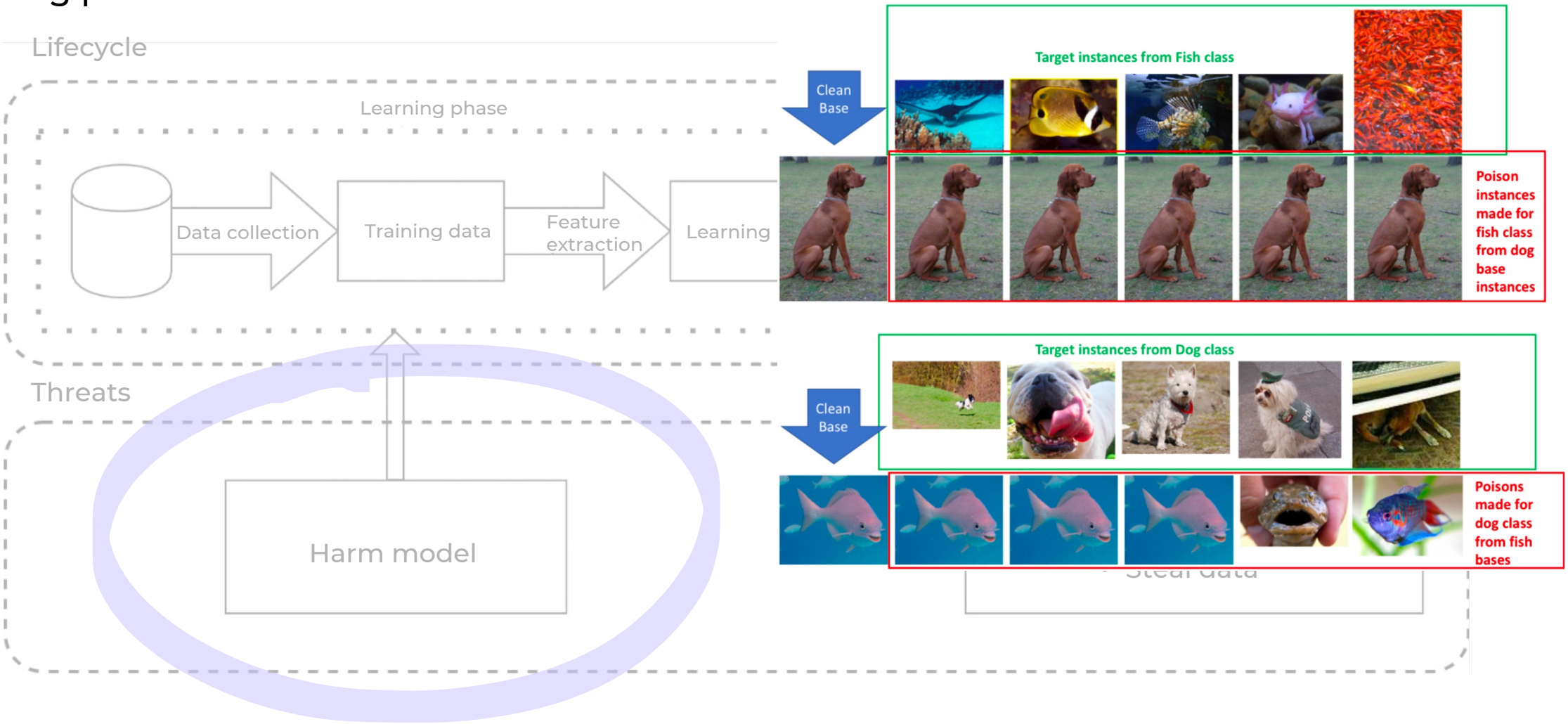
Threat model

Training phase



Threat model

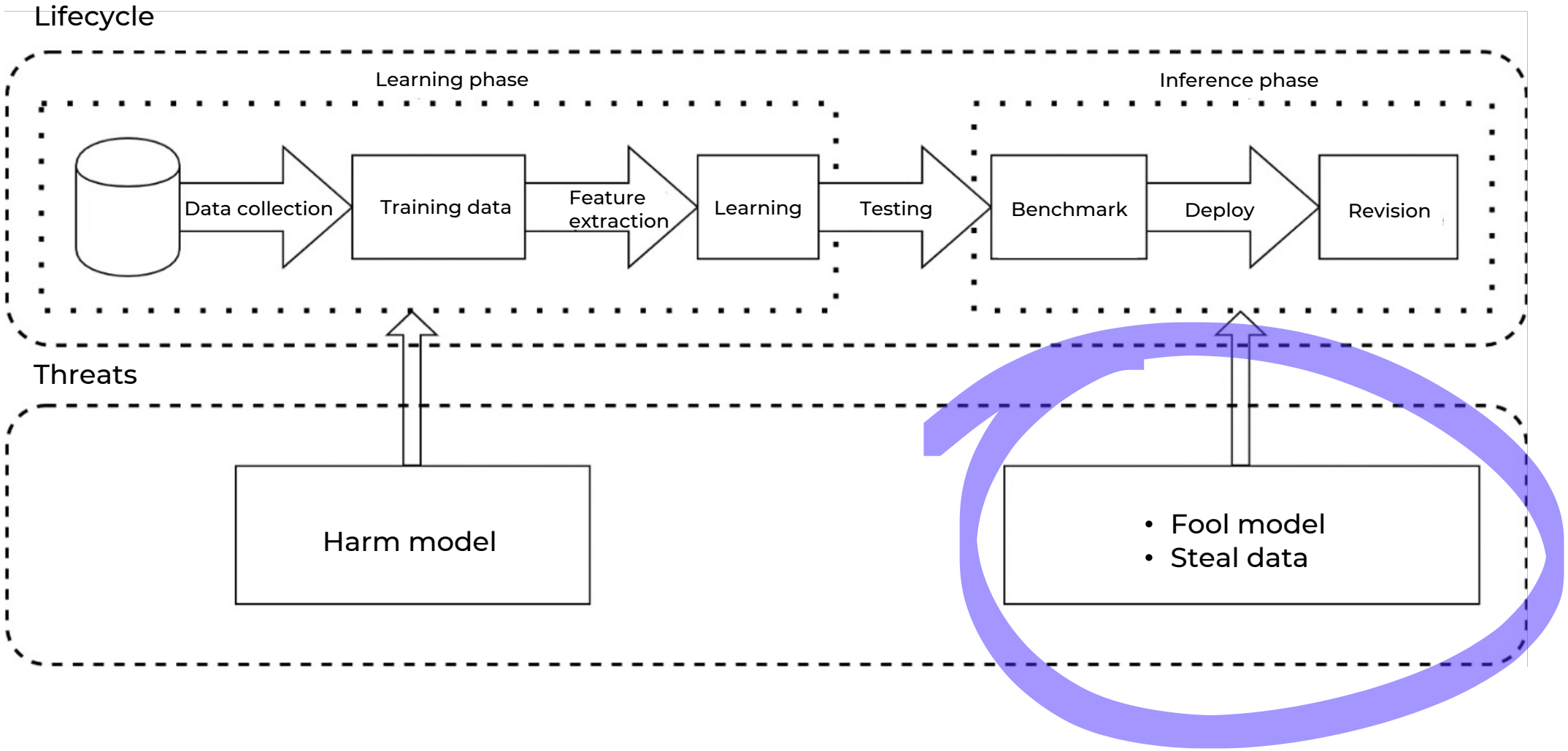
Training phase



arXiv:1804.00792

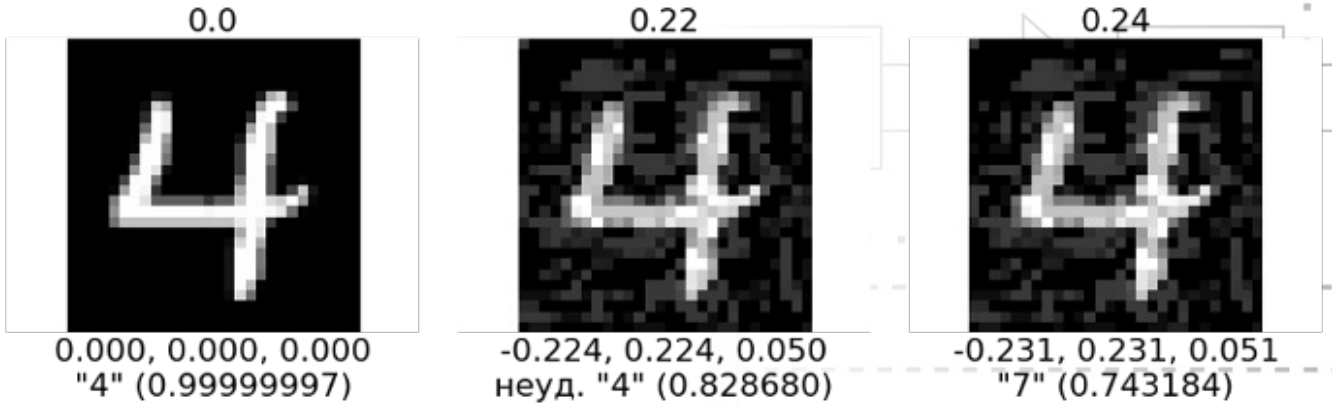
Threat model

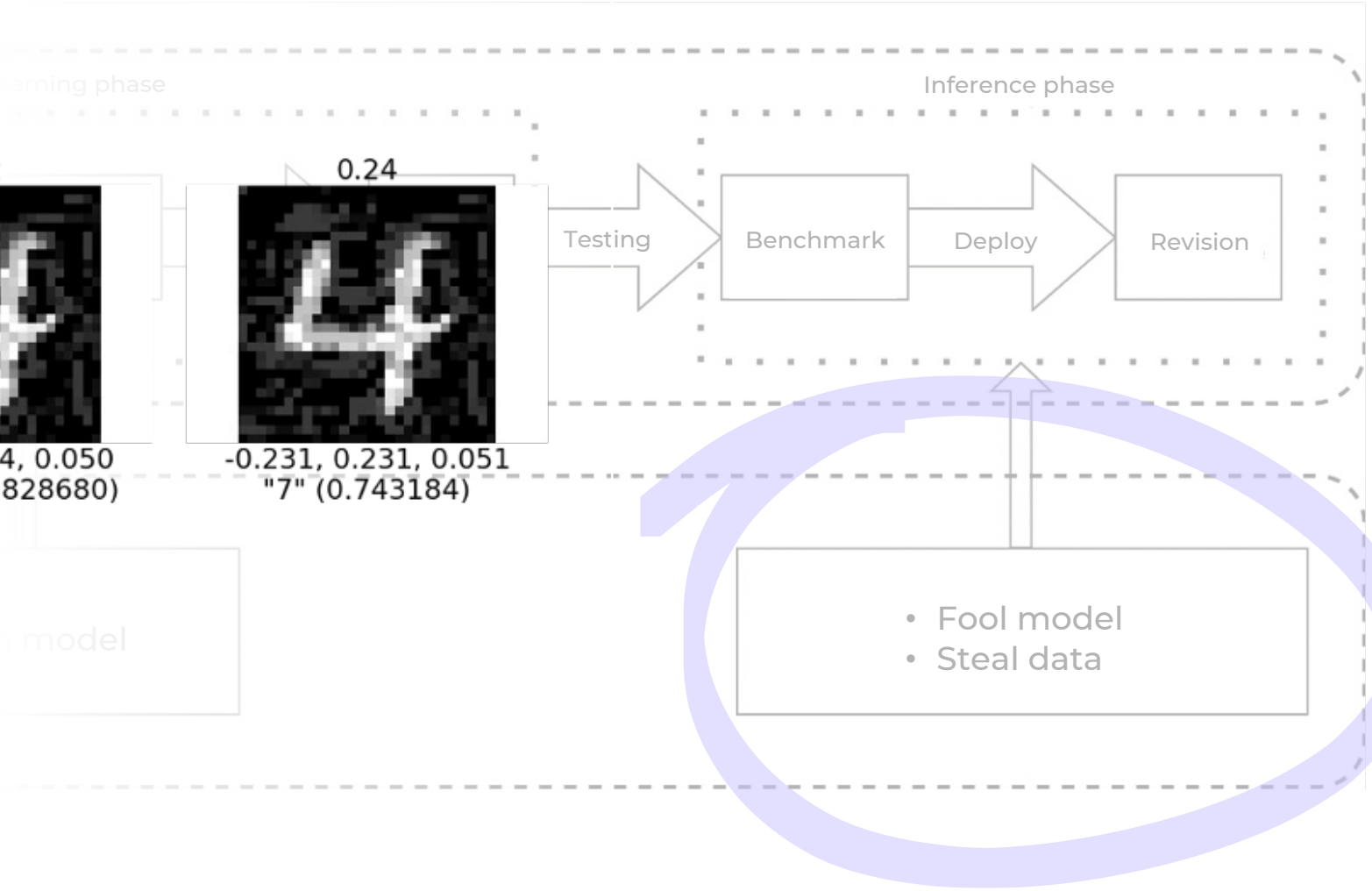
Inference phase



Adversarial examples

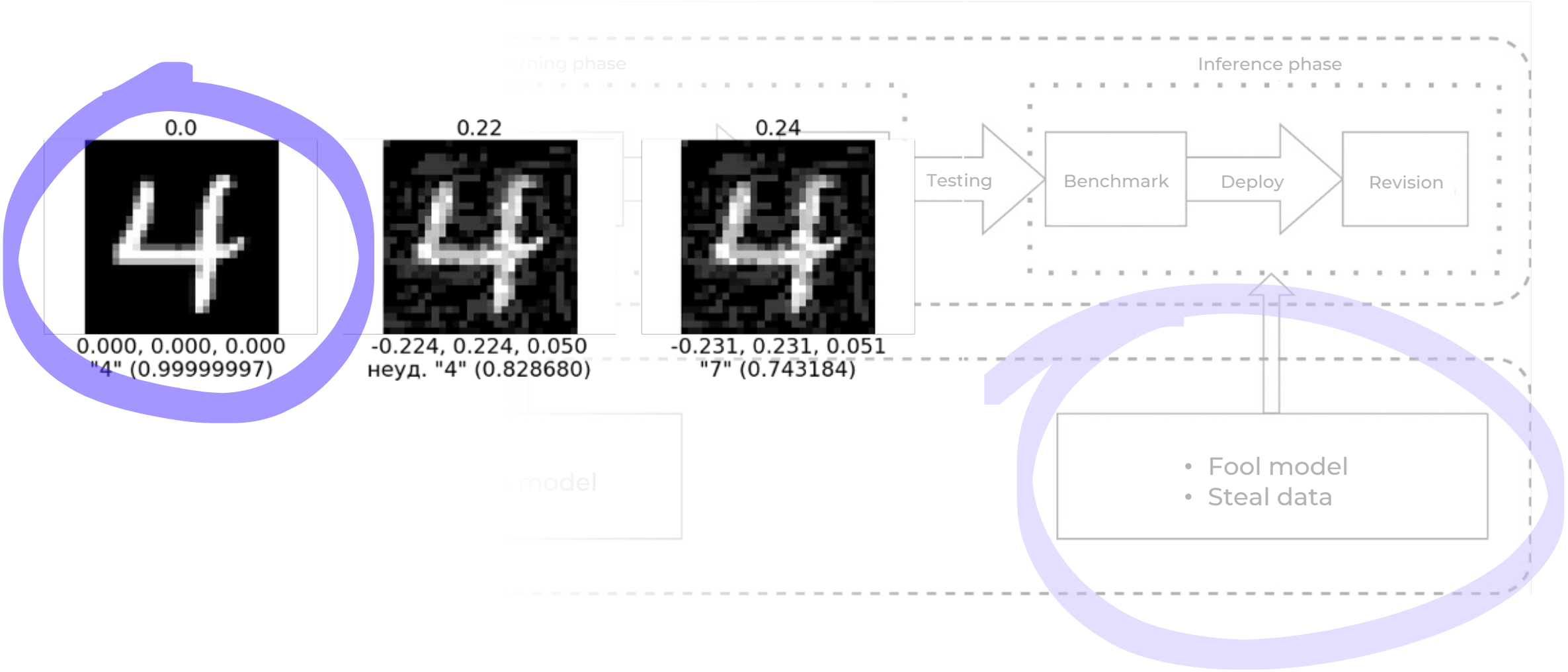
Perturbations





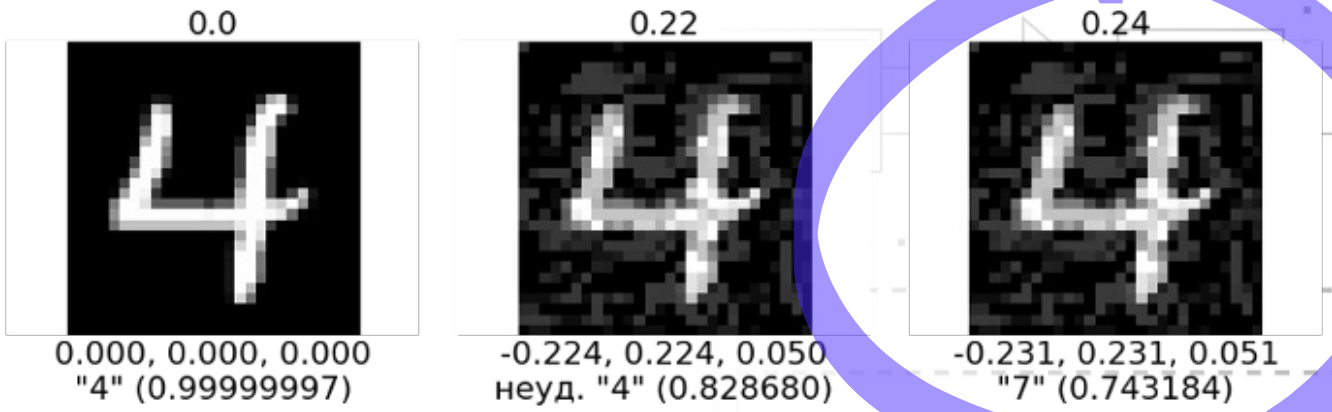
Adversarial examples

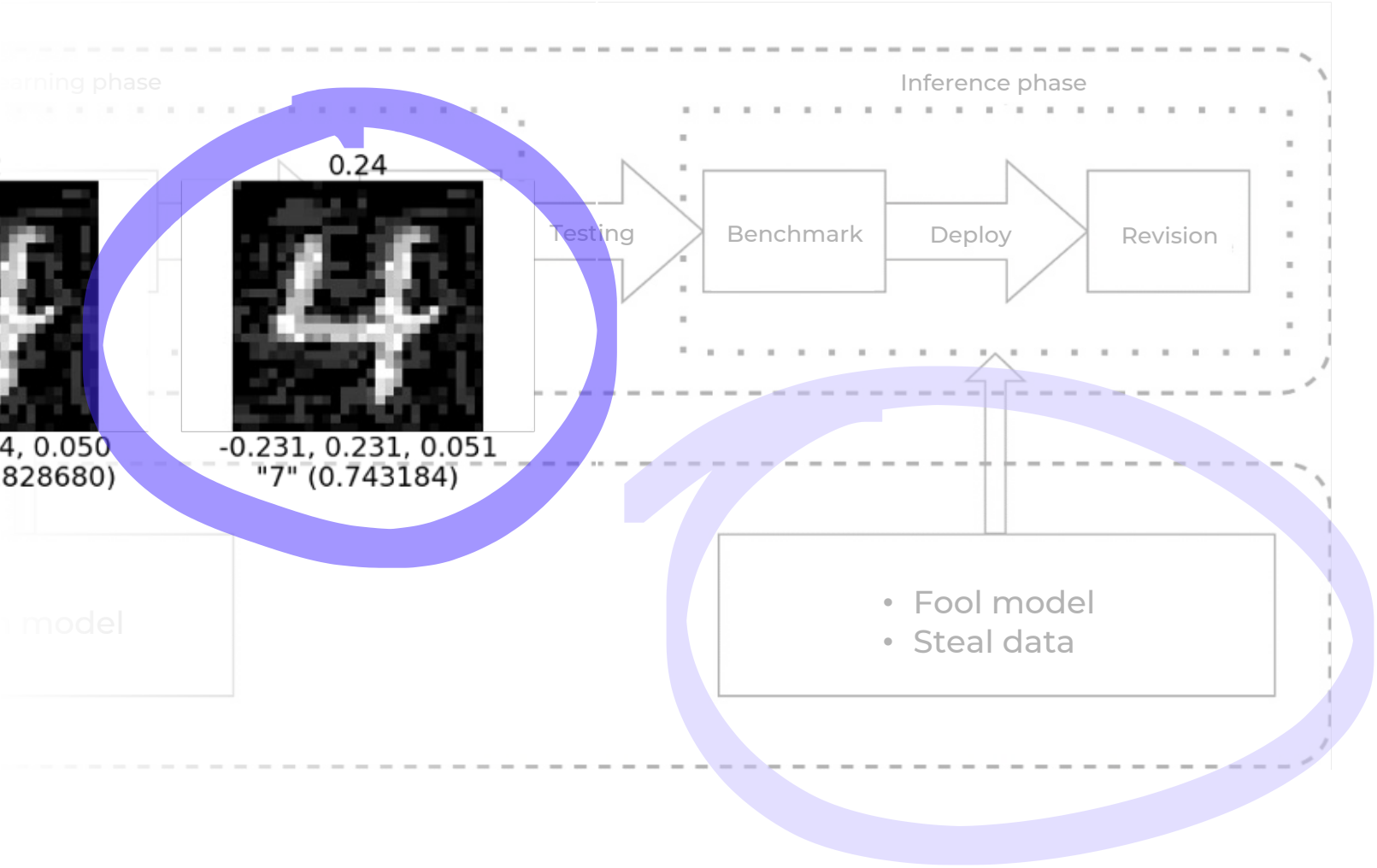
Perturbations



Adversarial examples

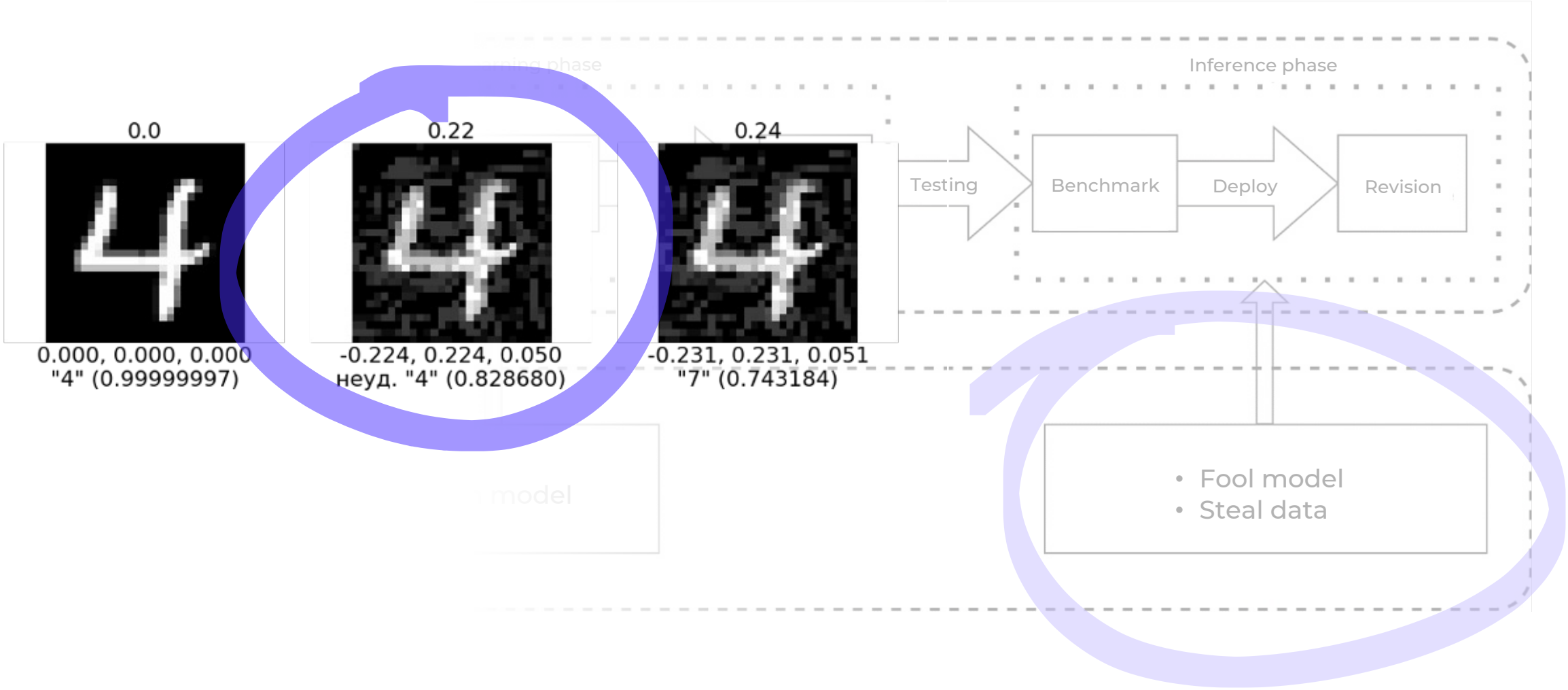
Perturbations





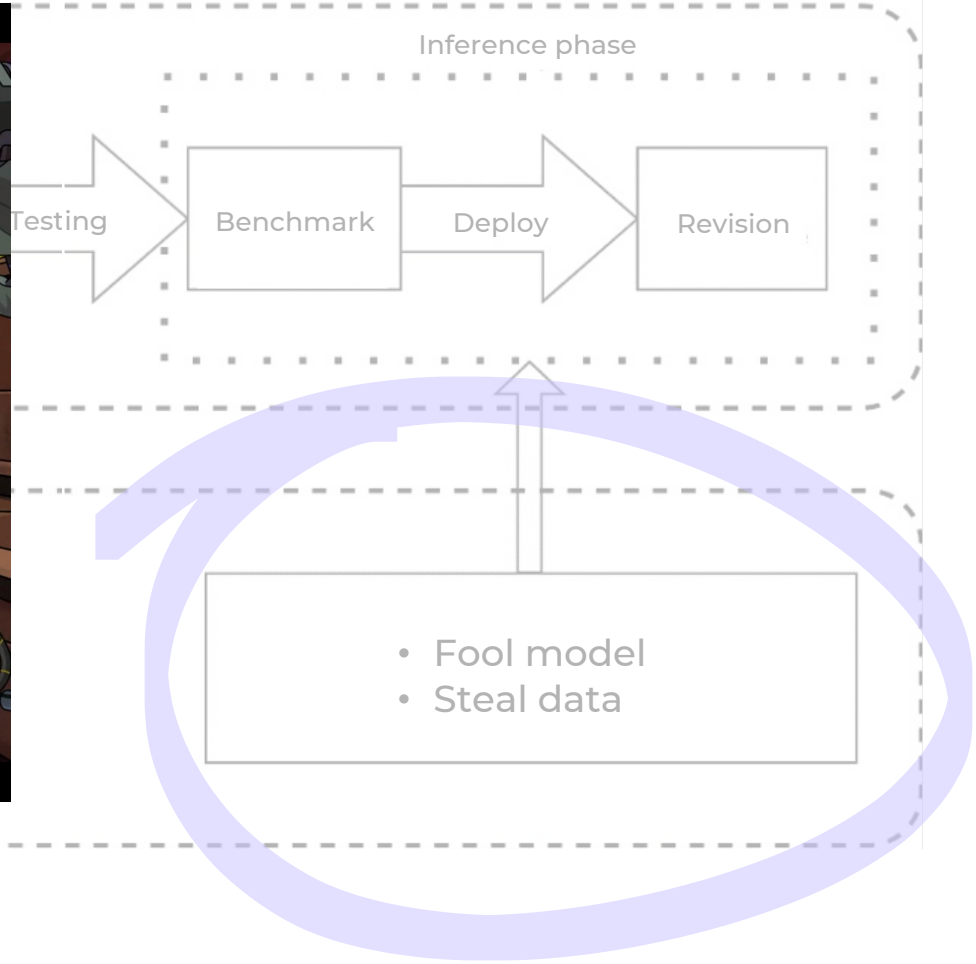
Adversarial examples

Perturbations



Adversarial examples

Watermarks

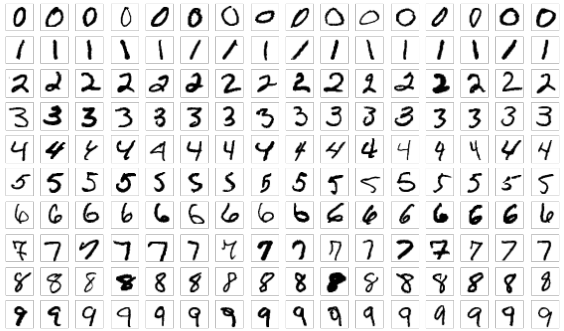


youtube.com/watch?v=OLqMKYJYjDA

Summary

We shall deal with

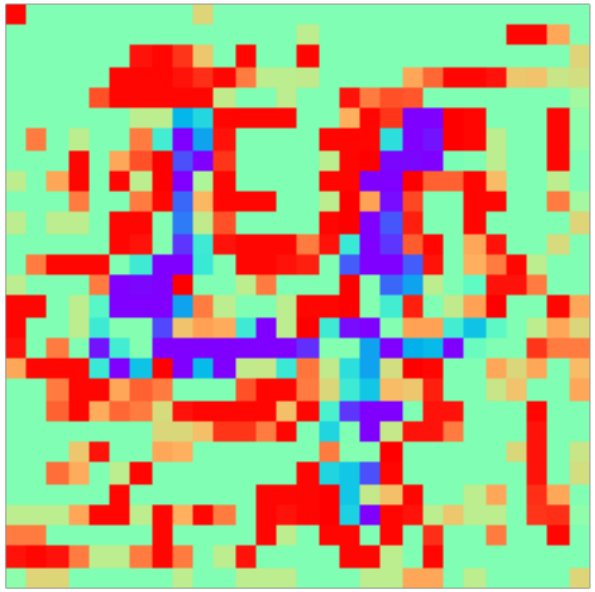
MNIST, PyTorch



PyTorch

Adversarial examples ≡ perturbations

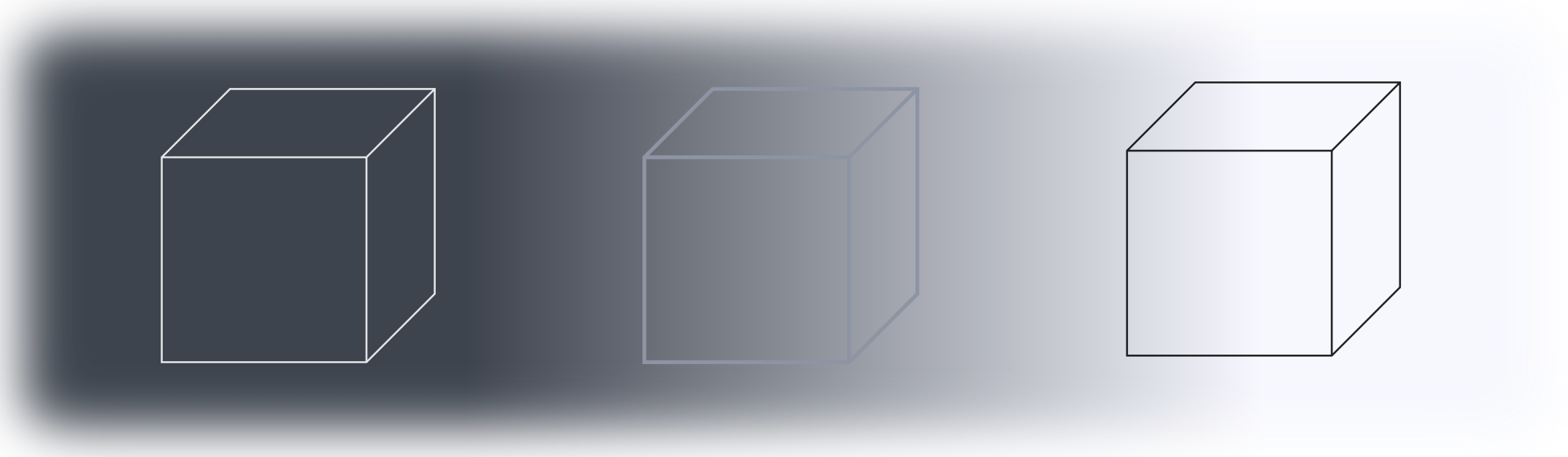
-0.231, 0.231, 0.051
“7” (0.743184)



-0.231, 0.231, 0.051
“7” (0.743184)

Black box, gray box, white box

Adversary knowledge about the model

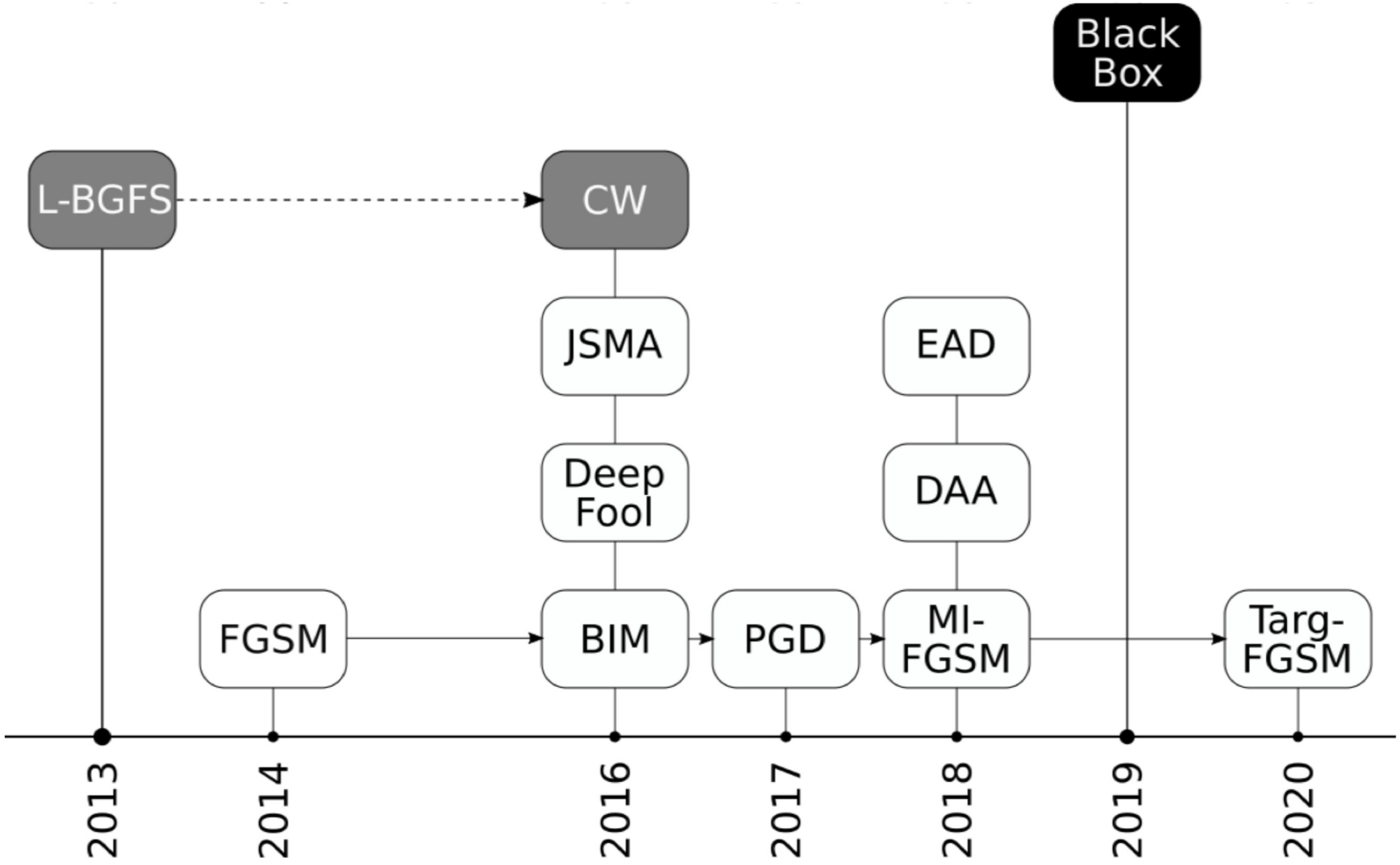


Nothing

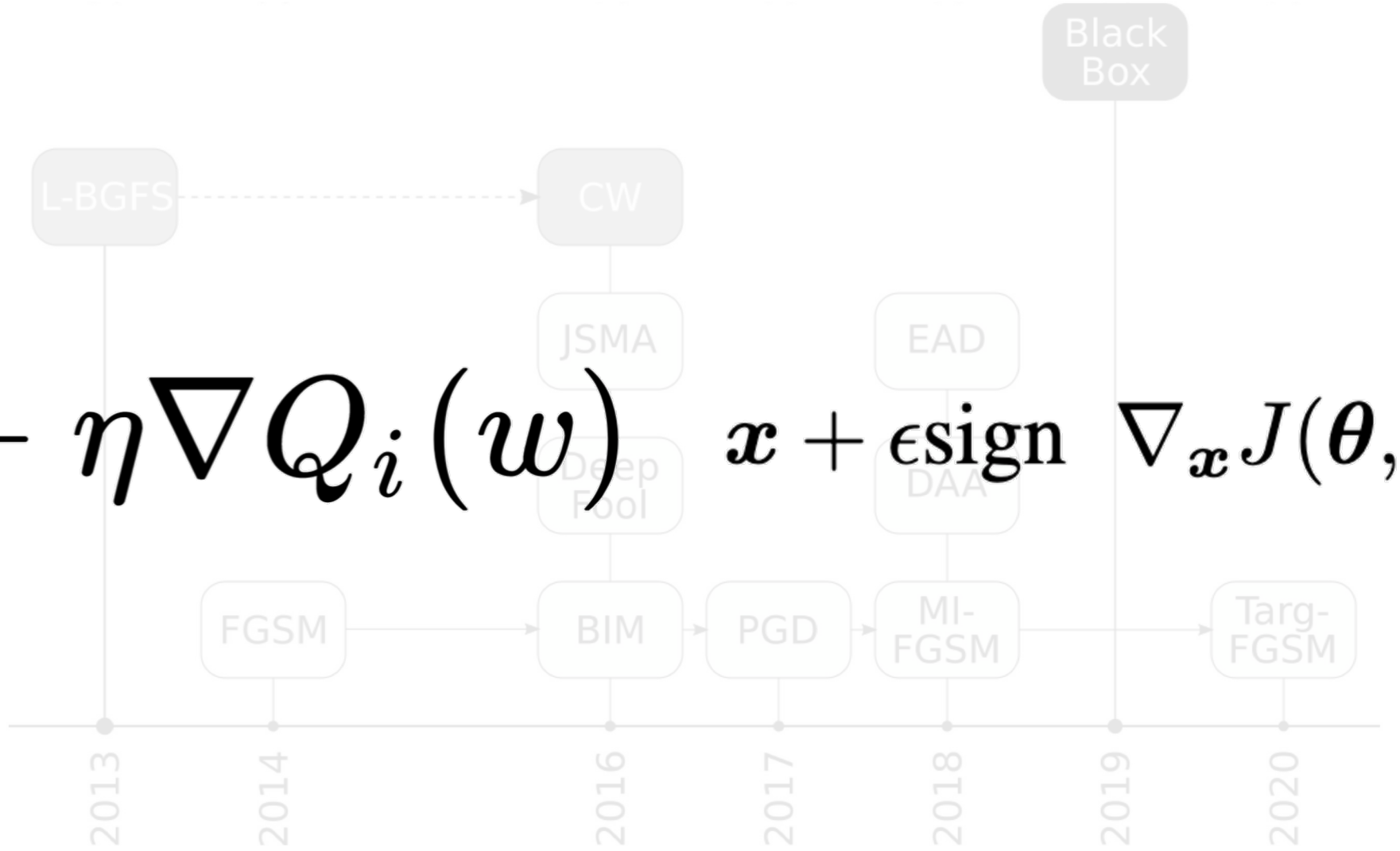
Architecture, loss

Everything

Adversarial examples generating methods



Adversarial examples generating philosophy



$$w - \eta \nabla Q_i(w)$$

$$x + \epsilon \text{sign} \ \nabla_x J(\theta, x, y)$$

wikipedia.org/wiki/Stochastic_gradient_descent, arxiv.org/abs/1412.6572

Frameworks

List selected for testing

AdvBox

ART

Foolbox

DeepRobust

Frameworks

AdvBox

AdvBox

ART

Foolbox

DeepRobust

- https://github.com/advboxes/AdvBox
- 2020
- Based on Foolbox v1
- It’s alive!

Frameworks

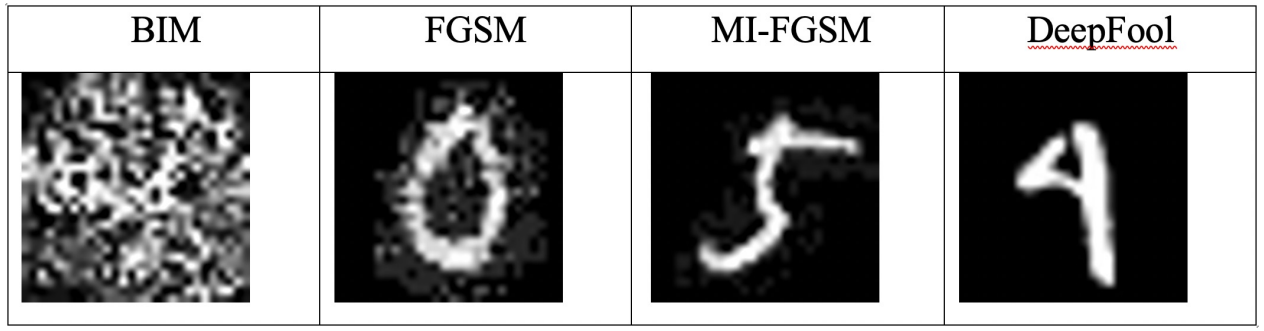
AdvBox

AdvBox

ART

Foolbox

DeepRobust



Frameworks

ART

AdvBox

ART

Foolbox

DeepRobust

Adversarial Robustness Toolbox

- https://github.com/Trusted-AI/adversarial-robustness-toolbox
- 2020
- Not bad

Frameworks

ART

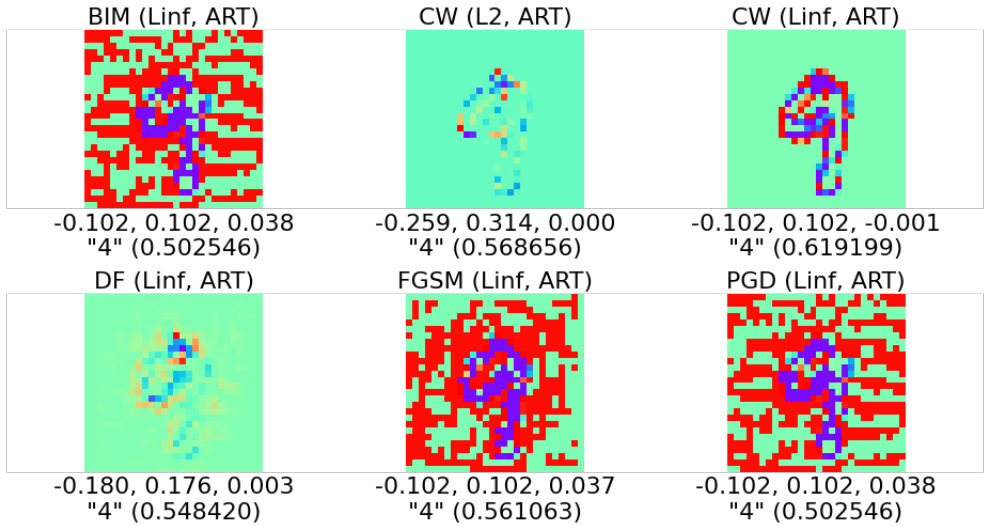
AdvBox

ART

Foolbox

DeepRobust

Adversarial Robustness Toolbox



Frameworks

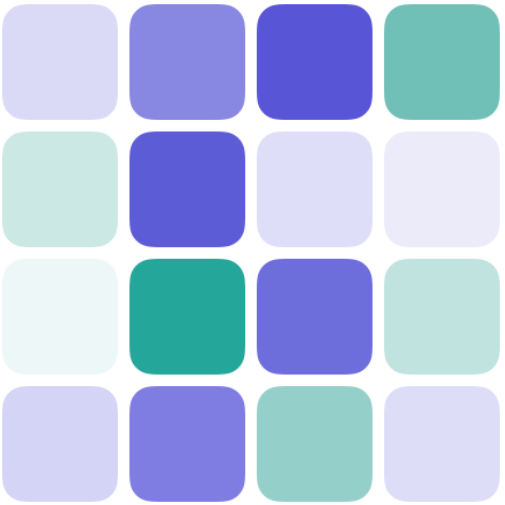
Foolbox

AdvBox

ART

Foolbox

DeepRobust



- https://github.com/bethgelab/foolbox
- 2017
- Epic!

Frameworks

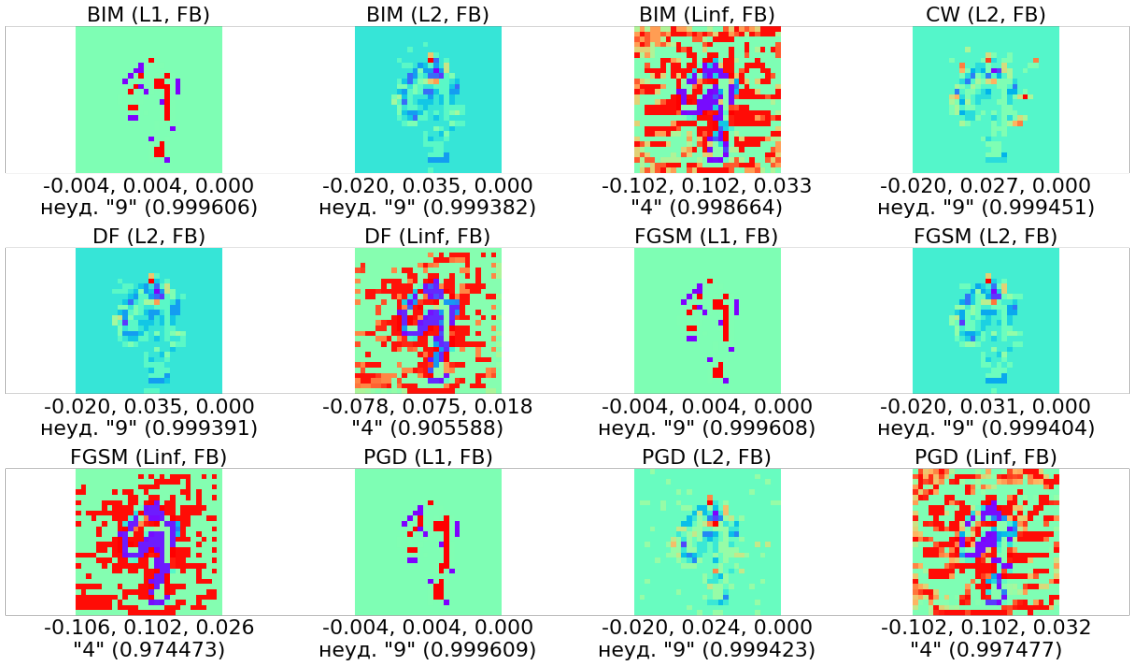
Foolbox

AdvBox

ART

Foolbox

DeepRobust



Frameworks

DeepRobust

AdvBox

ART

Foolbox

DeepRobust

- https://github.com/DSE-MSU/DeepRobust
- 2020
- not touching that one with a twenty-foot pole

Frameworks

DeepRobust

AdvBox

ART

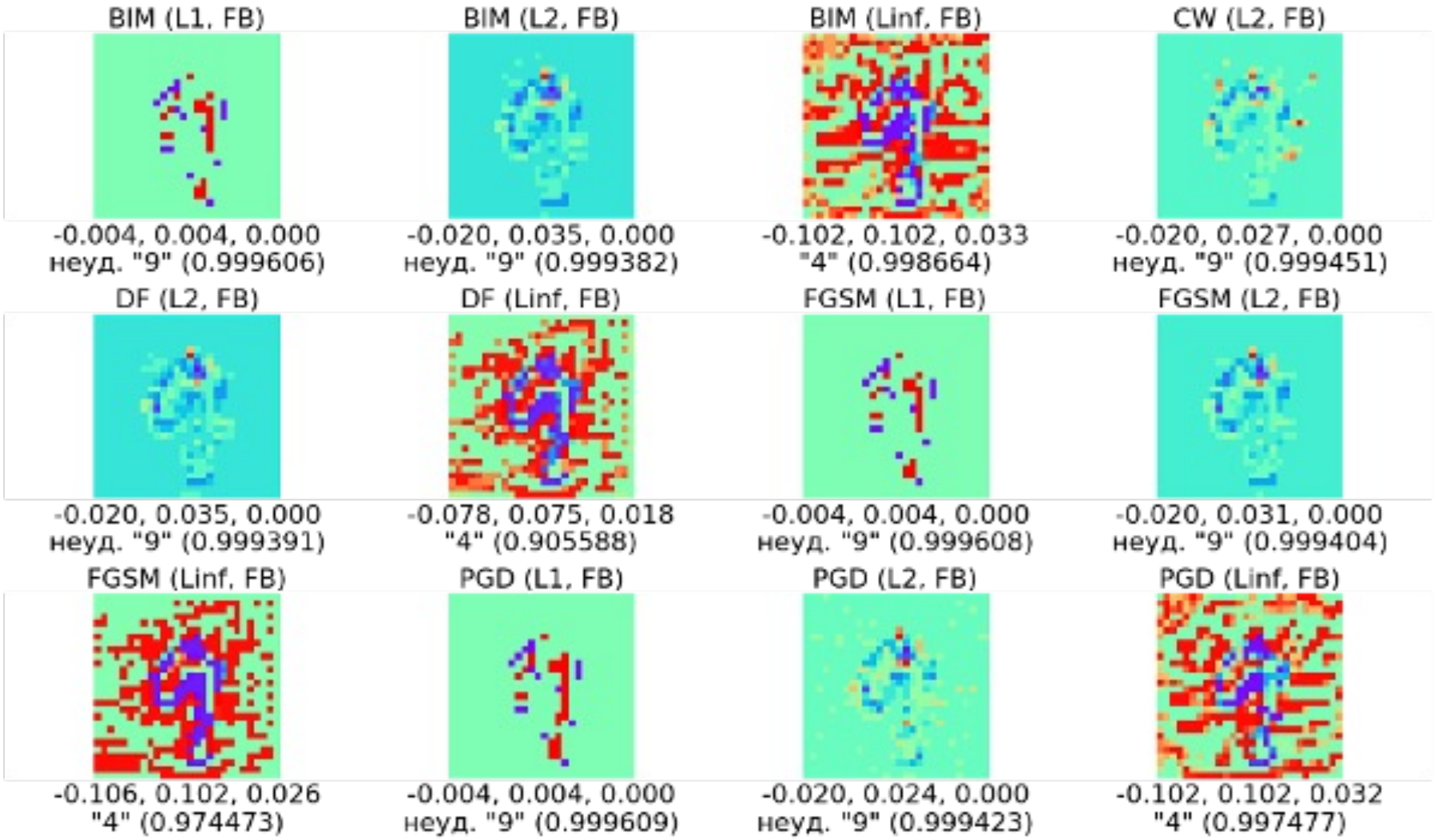
Foolbox

DeepRobust

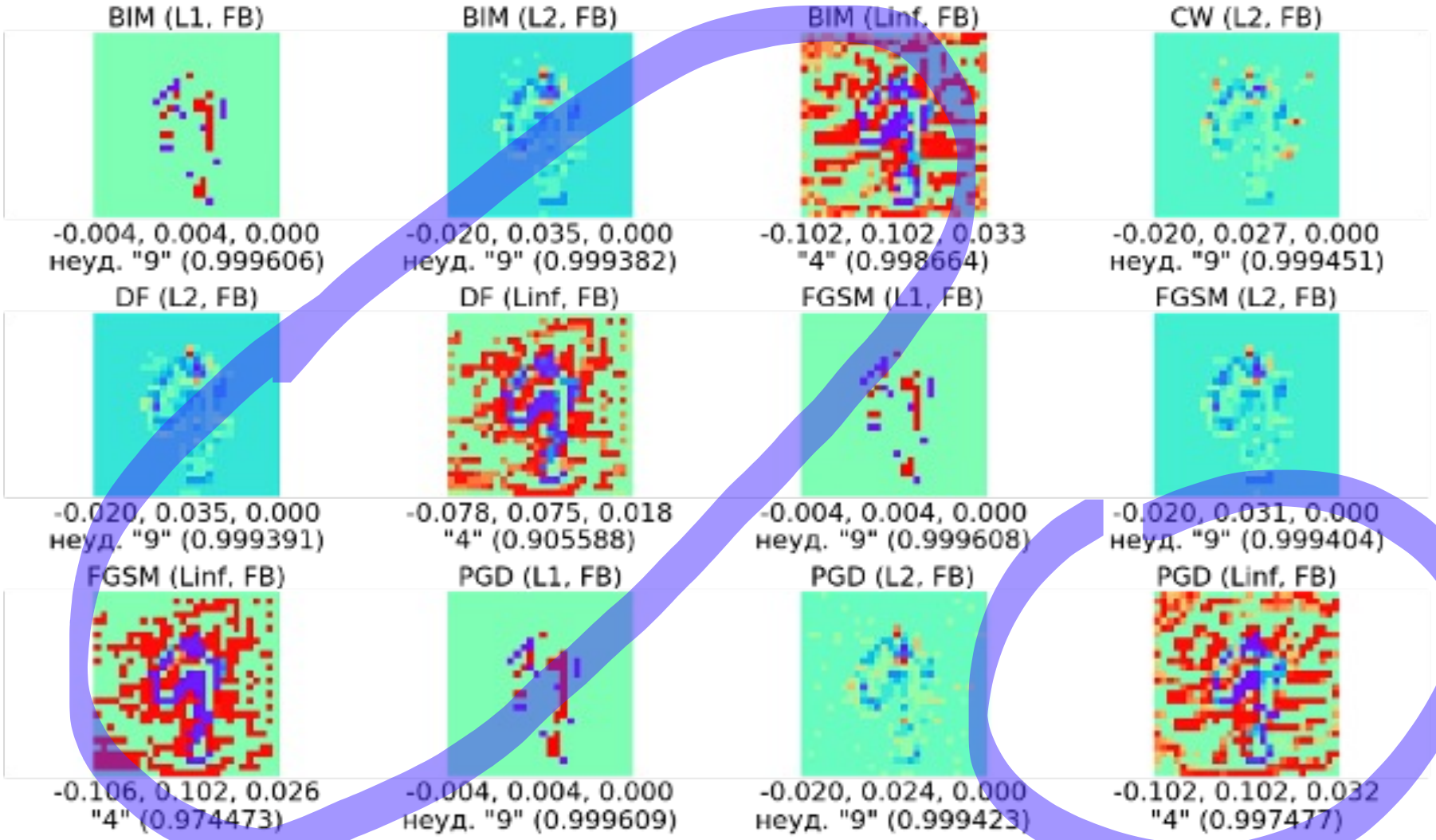
*

*As of 2020

Adversarial examples for class “9” obtained by Foolbox

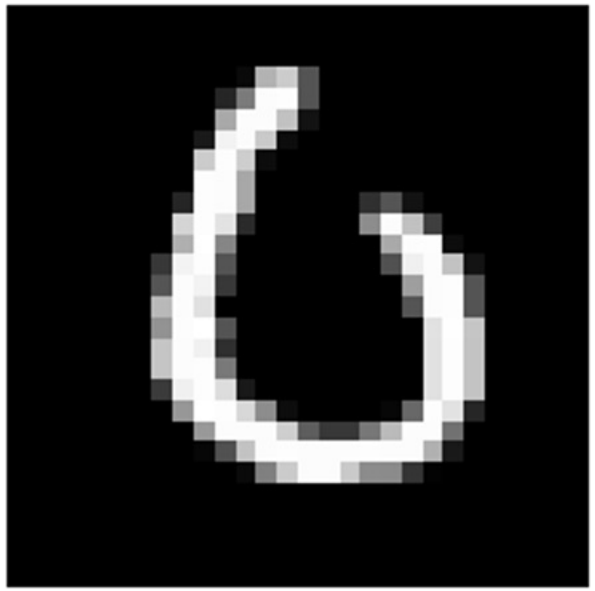


Adversarial examples for class “9” obtained by Foolbox

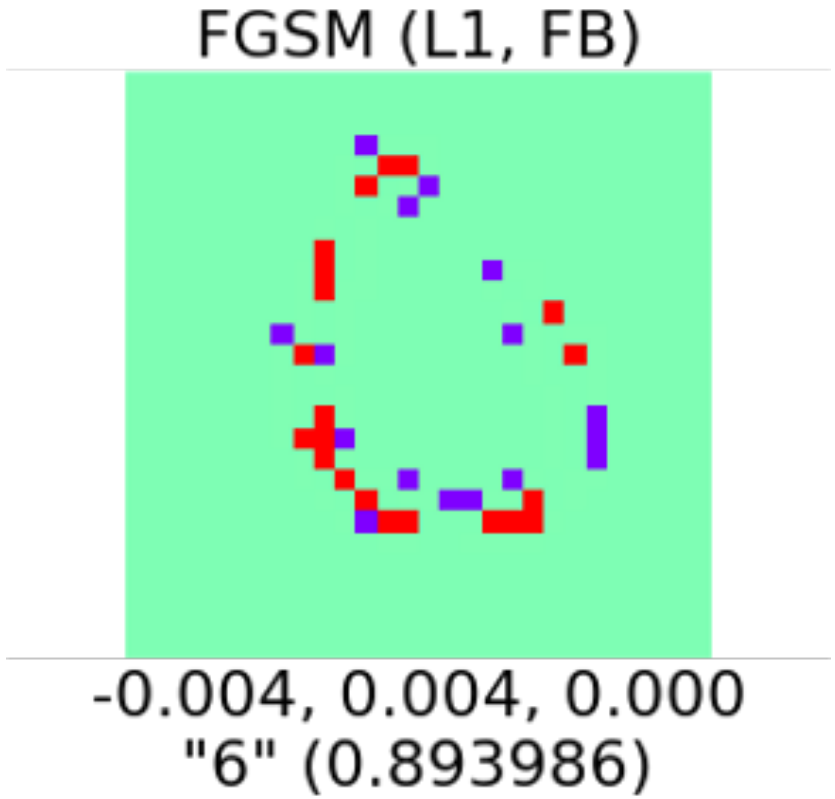


Perturbation value

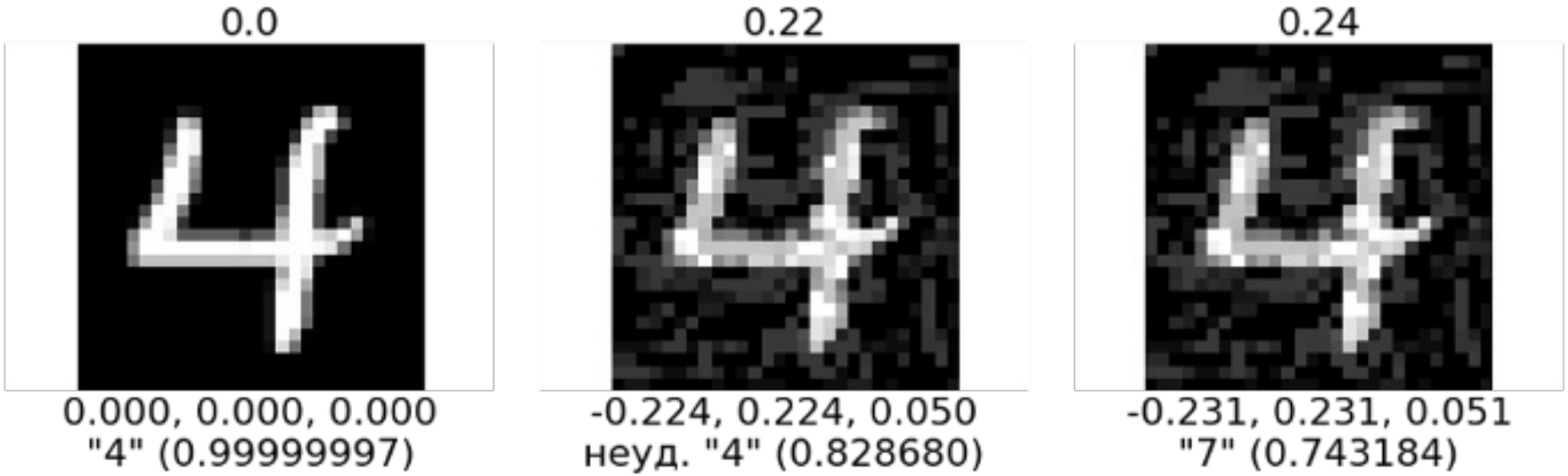
“Adversarial example” for human



“0” 0.895

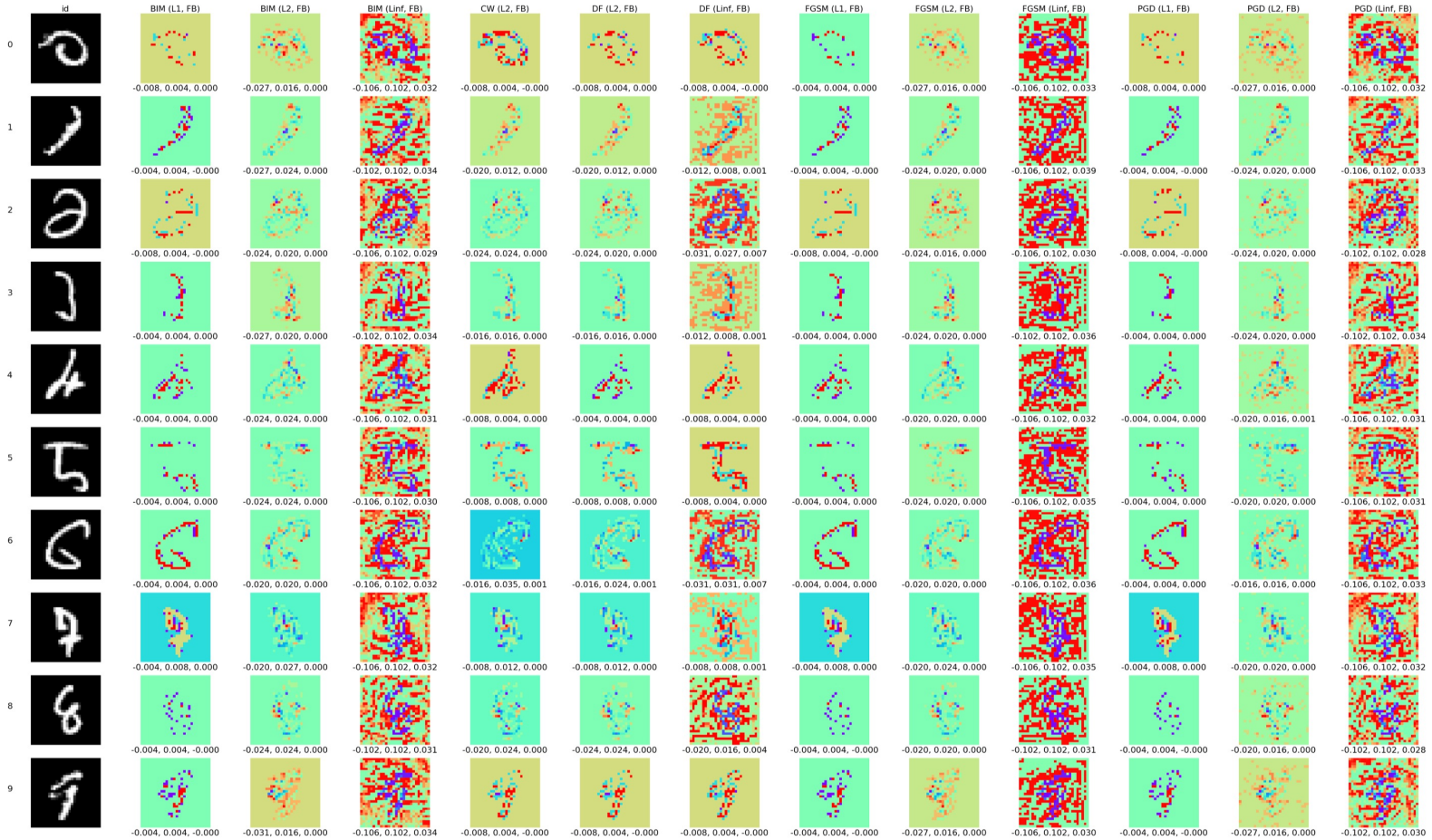


Perturbation value



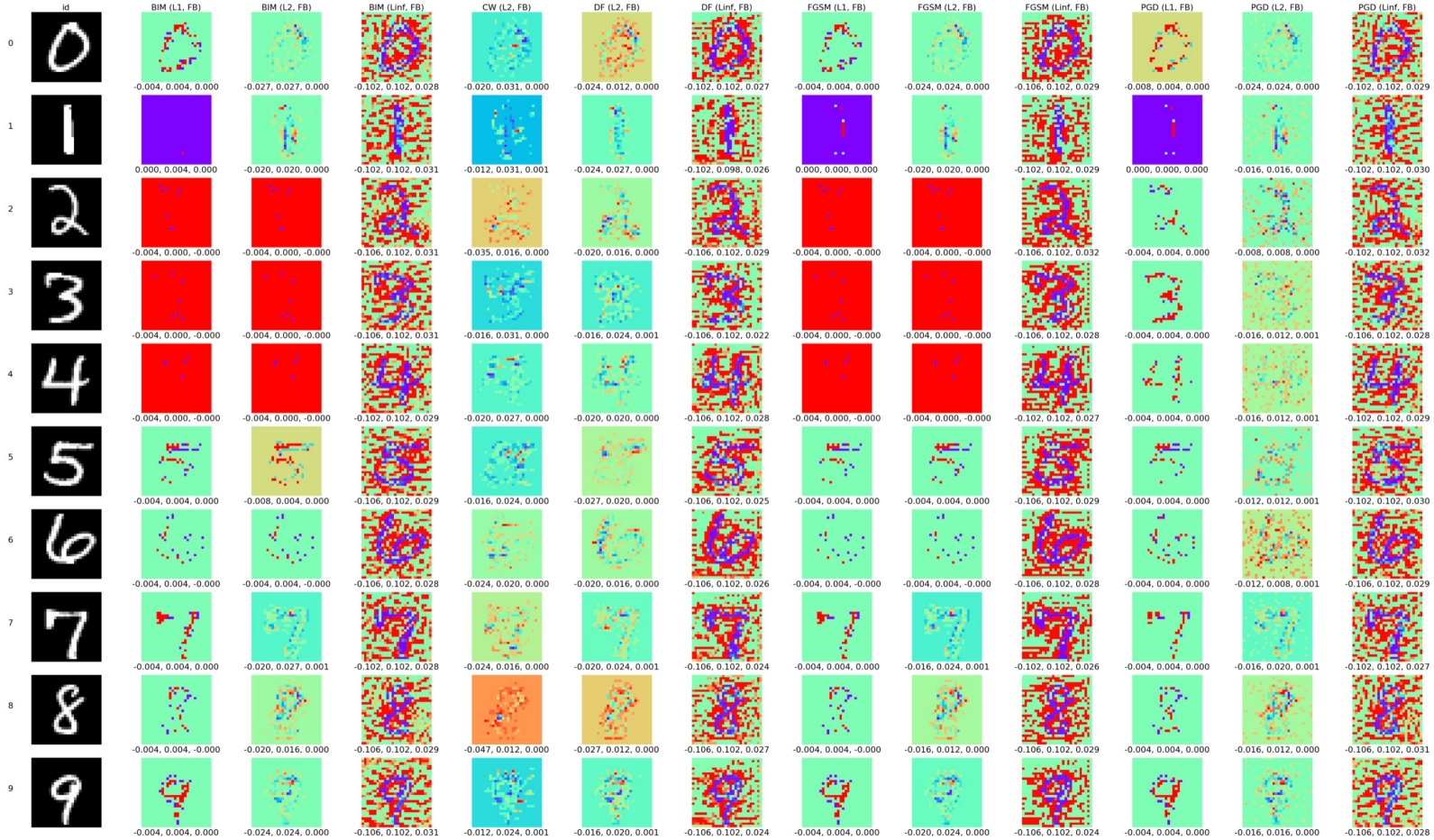
Perturbation value

Bad examples



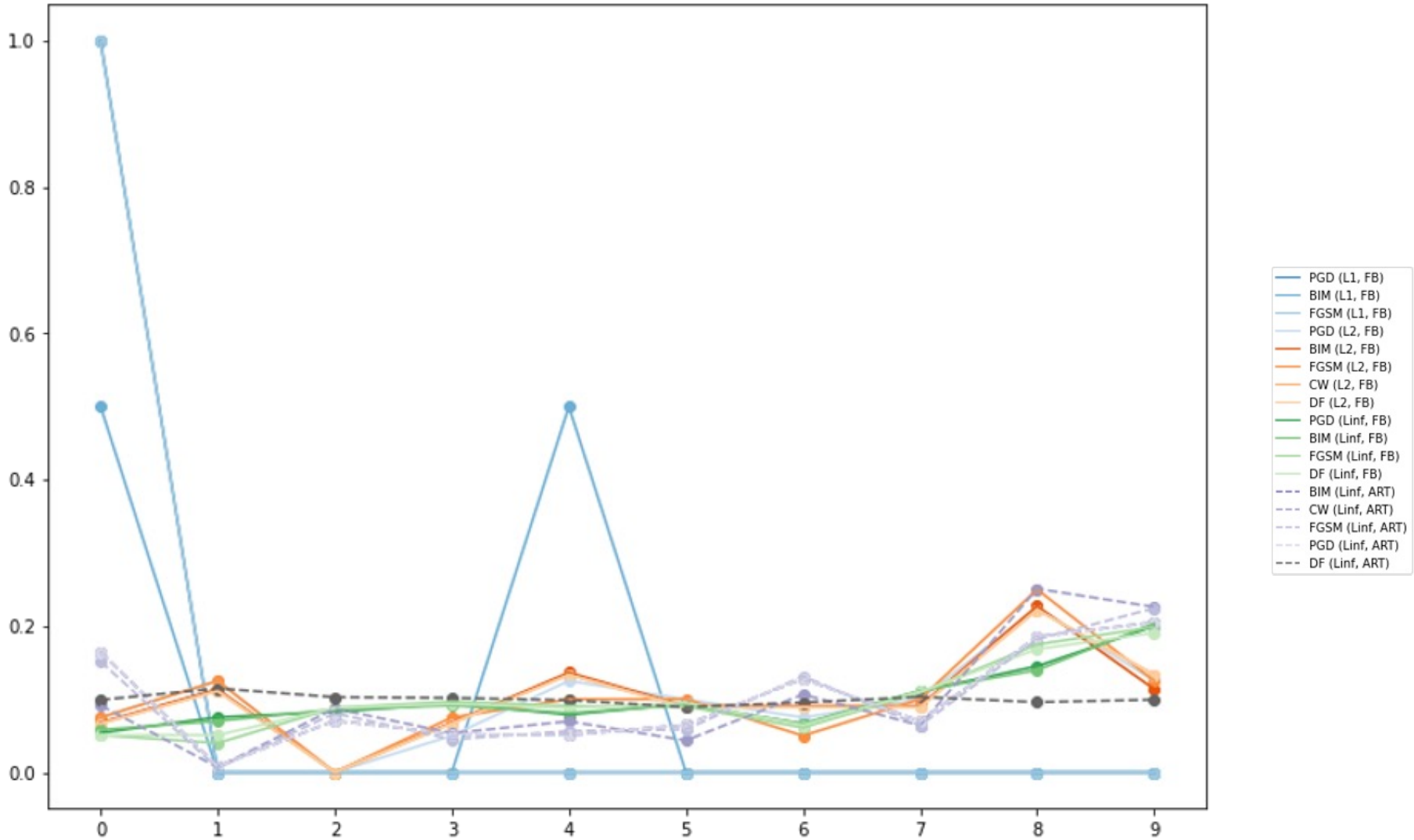
Perturbation value

Good examples



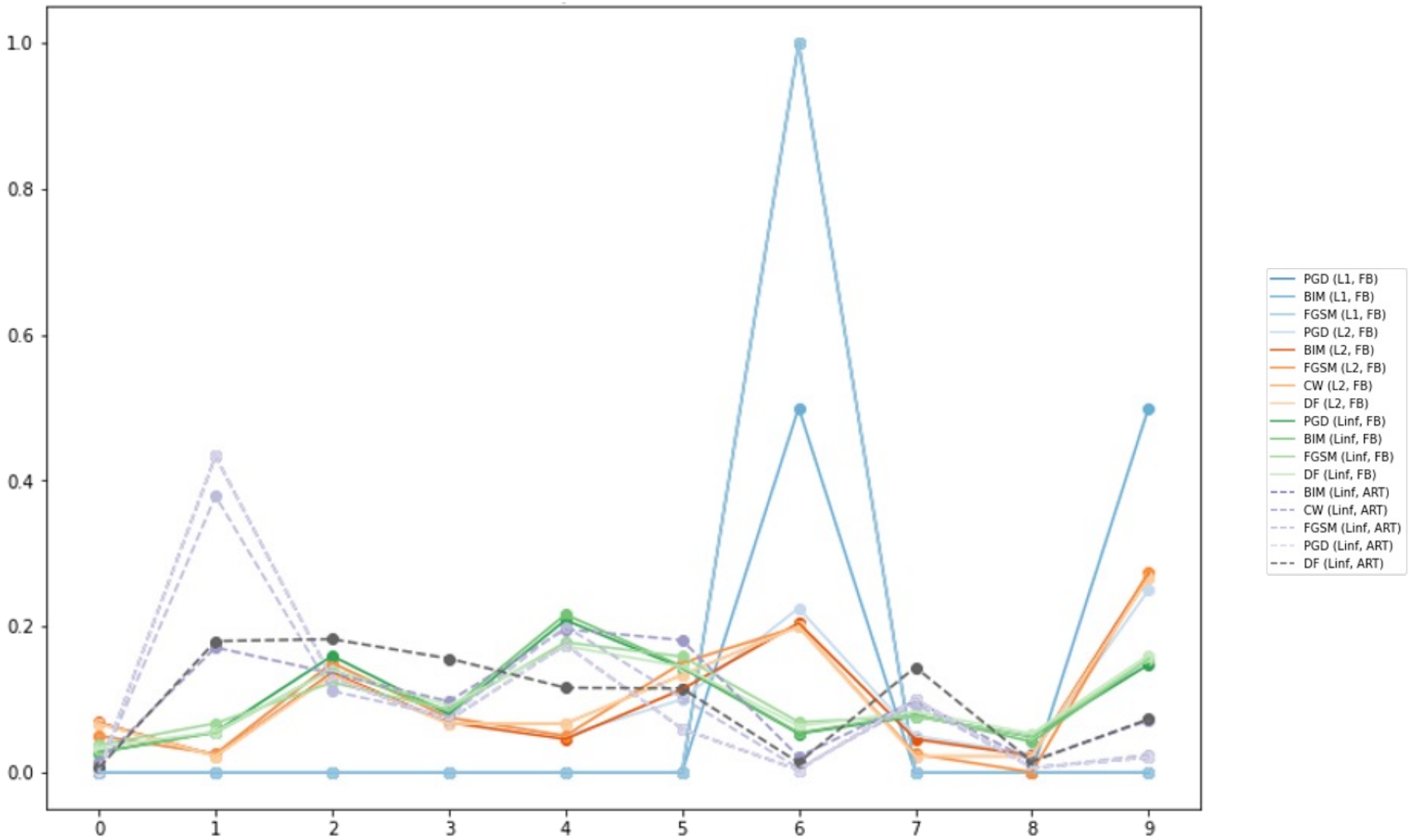
Attack performance vs label

Adversarial examples sources chart

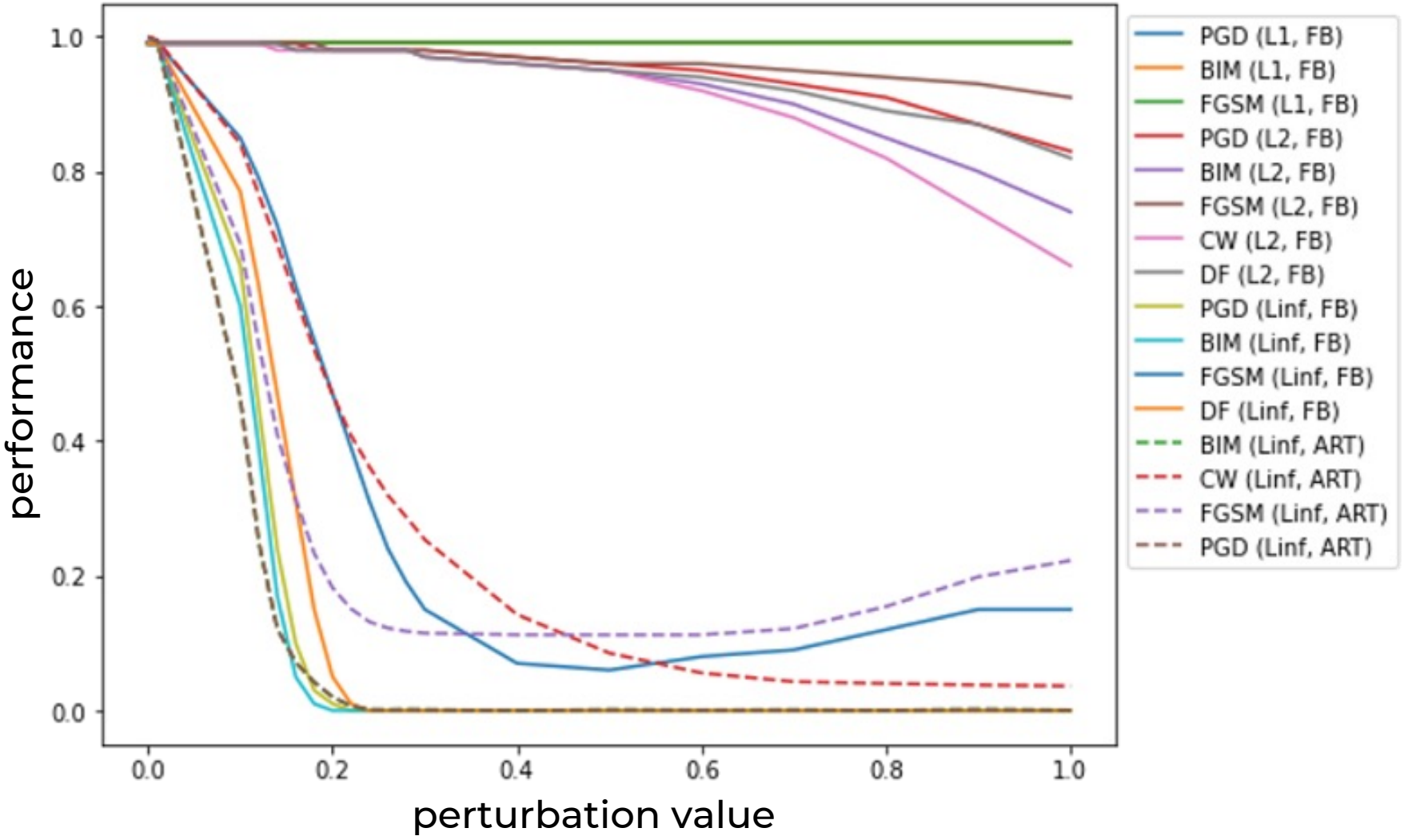


Attack performance vs label

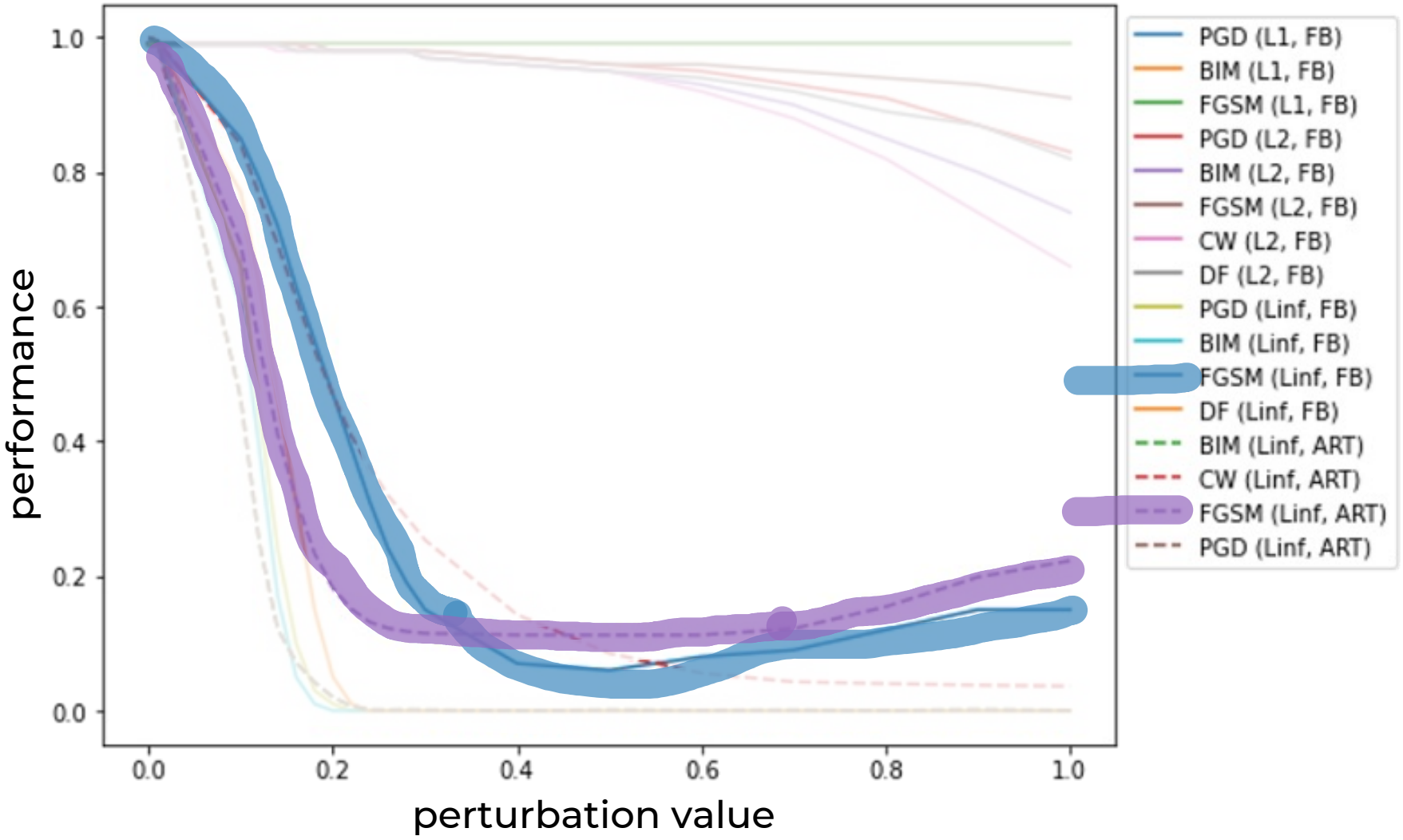
Adversarial examples targets chart



Benchmark of different methods implemented in ART and Foolbox



Benchmark of different methods implemented in ART and Foolbox

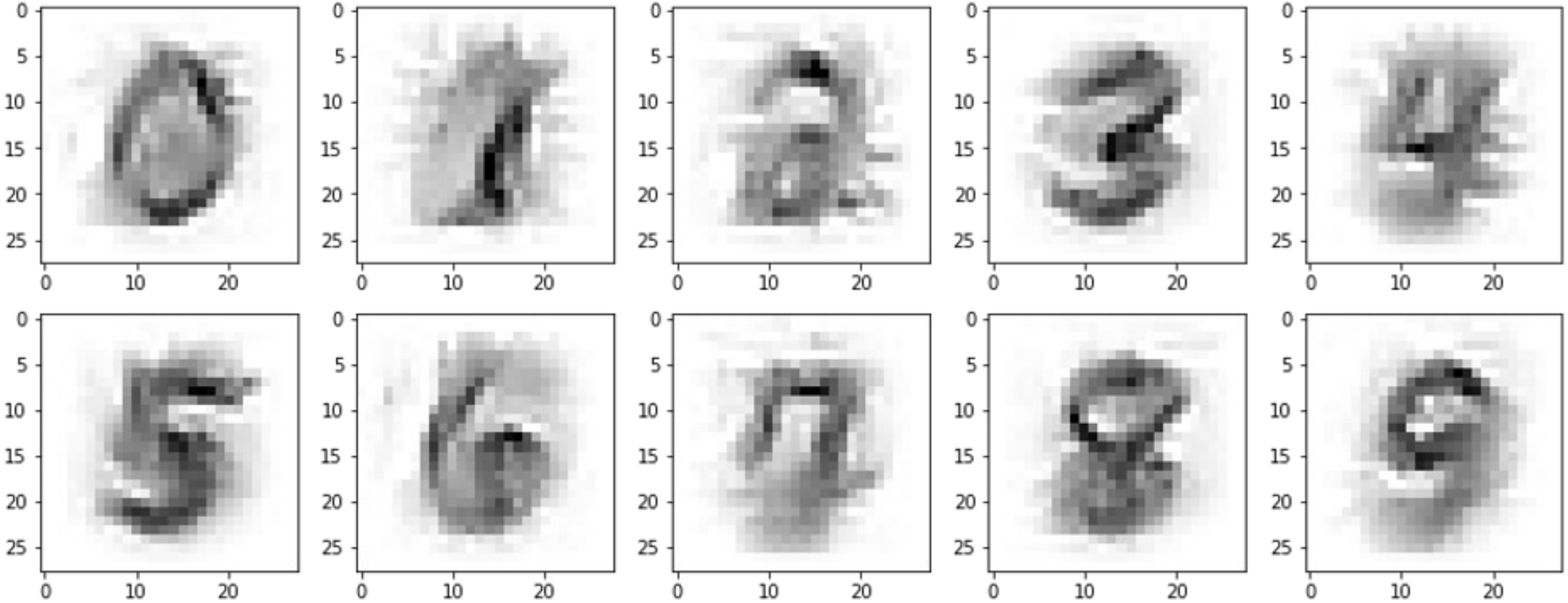


Beyond the scope

Black-box attacks

Beyond the scope

Data extraction



Thanks!

КРИПТОНИТ

Danila Yu. Emelyanov

d.emelyanov@kryptonite.ru