Poisoning attacks on computer vision models

Tom Goldstein

UNIVERSITY OF MARYLAND

WHAT IS POISONING?

Train-time attacks:
adversary controls training data

Does this *actually* happen?

Scraping images from the web

Harvesting system inputs (spam detector)

Bad actors/inside agents

COOL STUFF I WON'T TALK ABOUT

Regression

"Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," Jagielski et al. 2018

Label flipping

"Poisoning attacks against support vector machines," Biggio et al., 2021

"Efficient label contamination attacks against black-box learning models," Zhang et al., 2017

Cryptography / P-tampering

"Blockwise p-tampering attacks on cryptographic primitives, extractors, and learners," Mahloujifar and Mahmoody.

Federated learning

"Data poisoning attacks against federated learning systems," Tolpegin 2020

"Analyzing federated learning through an adversarial lens," Bhagoji 2019

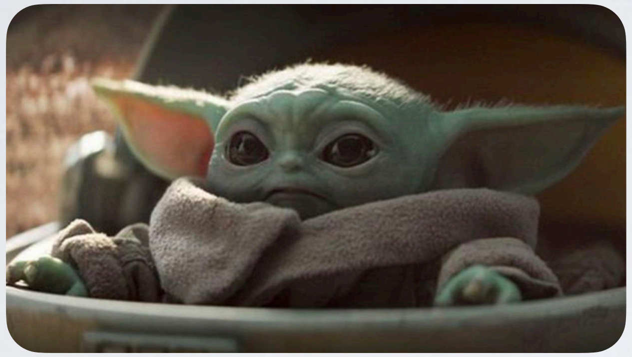
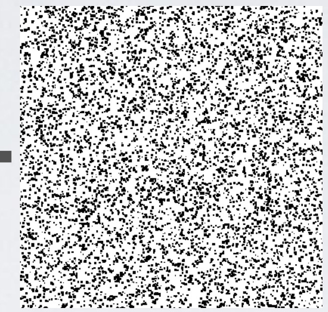
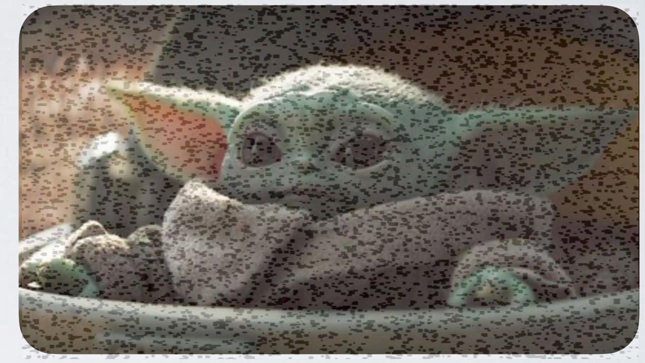
"Data poisoning attacks on federated machine learning," Sun 2020

Overview paper

"Dataset Security for Machine Learning: Data poisoning, Backdoor Attacks, and Defenses"

STUFF I WILL TALK ABOUT

Training-only attacks

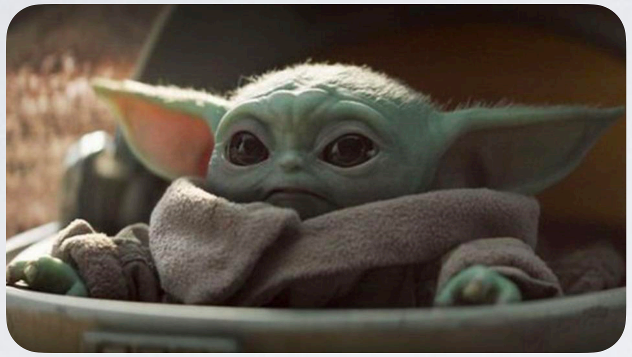
Train  +  = 

Test  → Adversarial label “Boba Fett”

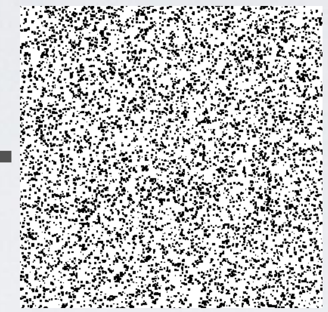
STUFF I WILL TALK ABOUT

Training-only attacks

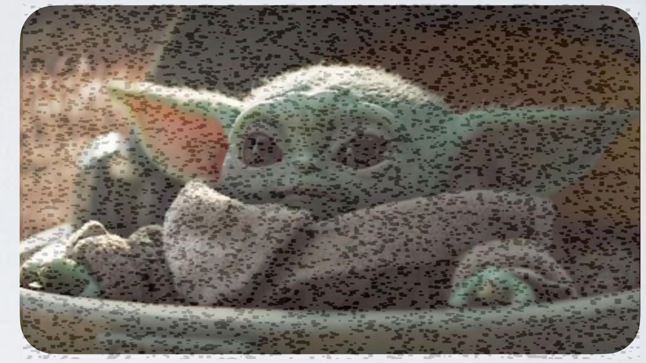
Train



+



=



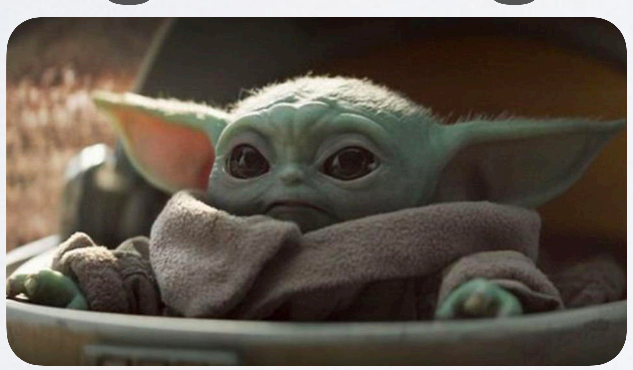
Test

Adversarial label
"Boba Fett"

Training-testing attacks

"Backdoors/trojans"

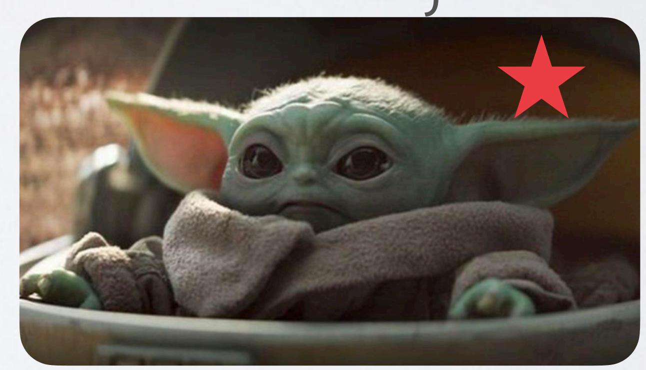
Train



+

"trigger"

=



Test

Adversarial label
"frog"

CLEAN-LABEL + TARGETED

Clean label: poisons are labeled "correctly"

This makes attacks hard to detect by auditing.

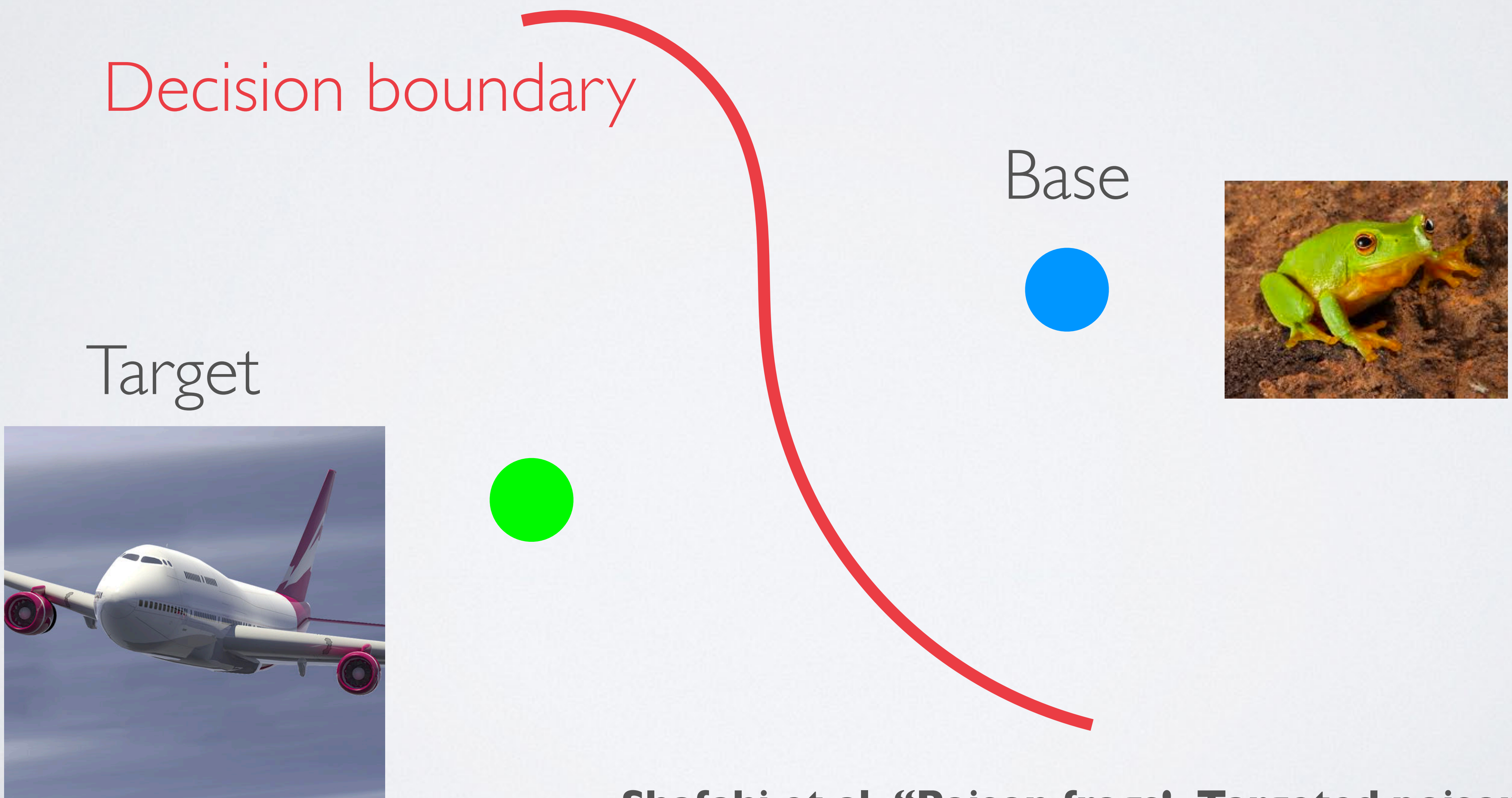
Targeted: Performance only changes on selected target

This makes attacks hard to detect by testing.

Attacks on transfer learning

COLLISION ATTACK

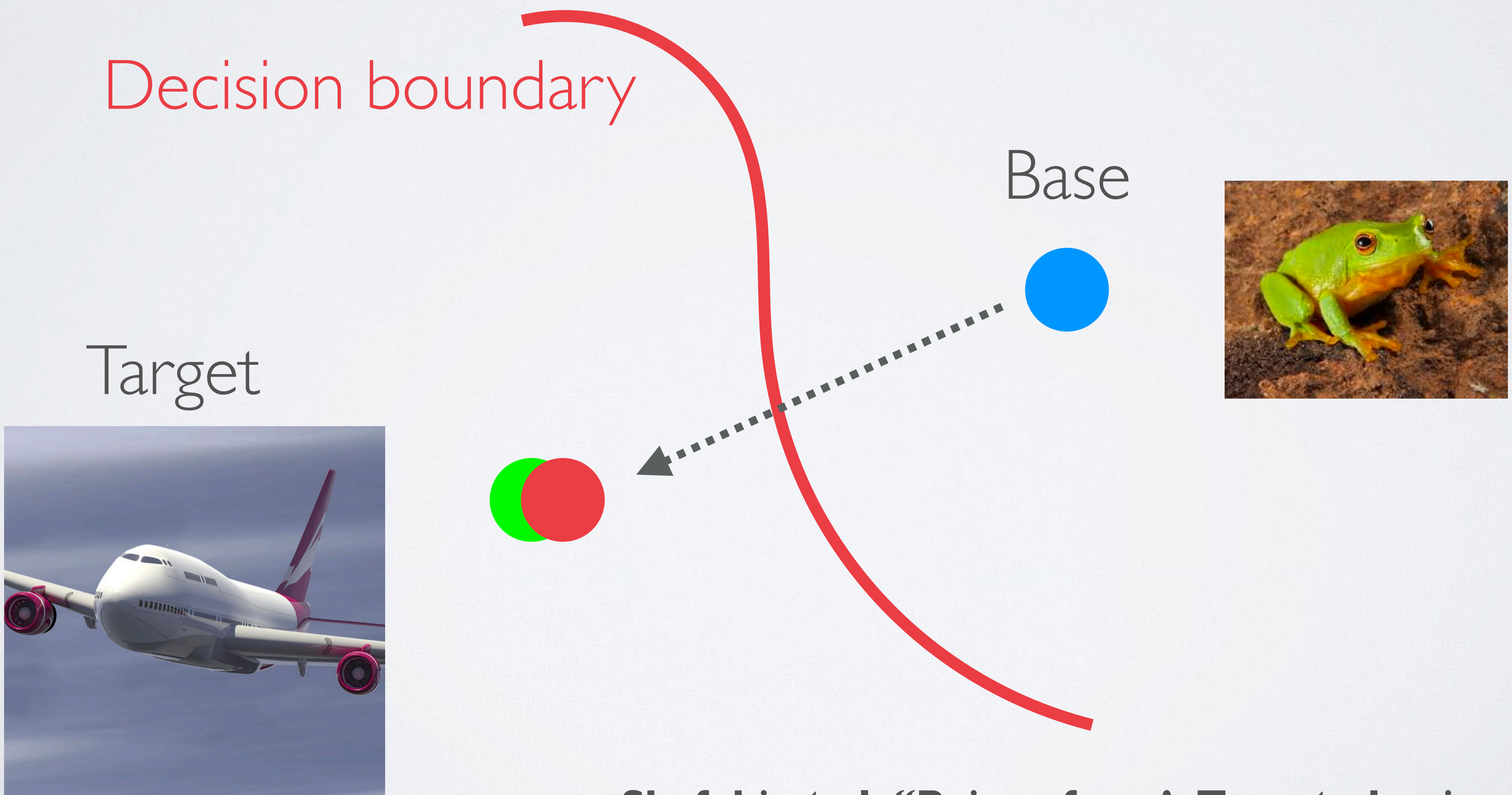
$$\mathbf{p} = \underset{\forall \mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|^2 + \beta \|\mathbf{x} - \mathbf{b}\|^2 \qquad (1)$$



Shafahi et al. “Poison frogs! Targeted poisoning attacks…”

COLLISION ATTACK

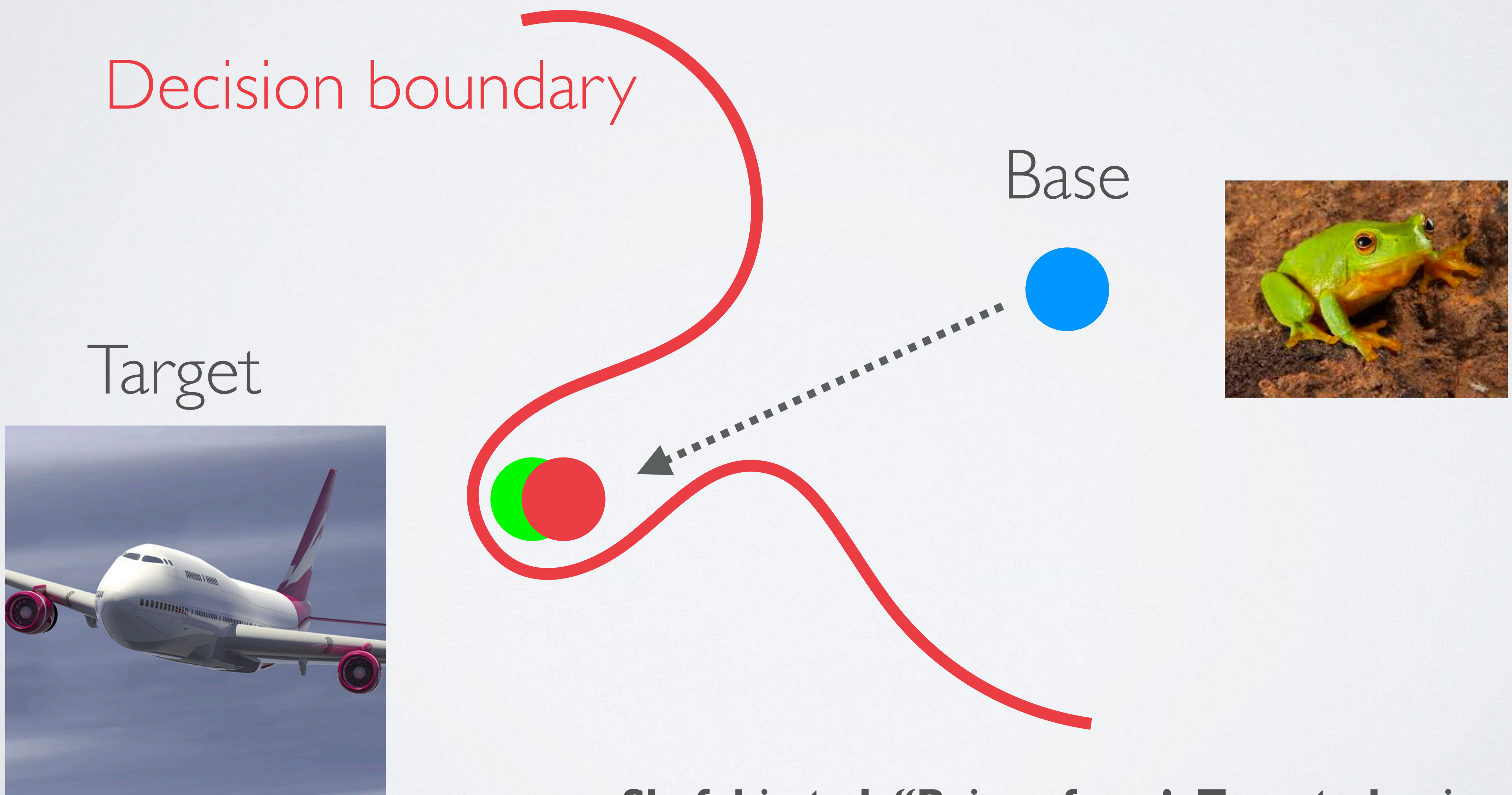
$$\mathbf{p} = \underset{\forall \mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|^2 + \beta \|\mathbf{x} - \mathbf{b}\|^2 \qquad (1)$$



Shafahi et al. “Poison frogs! Targeted poisoning attacks…”

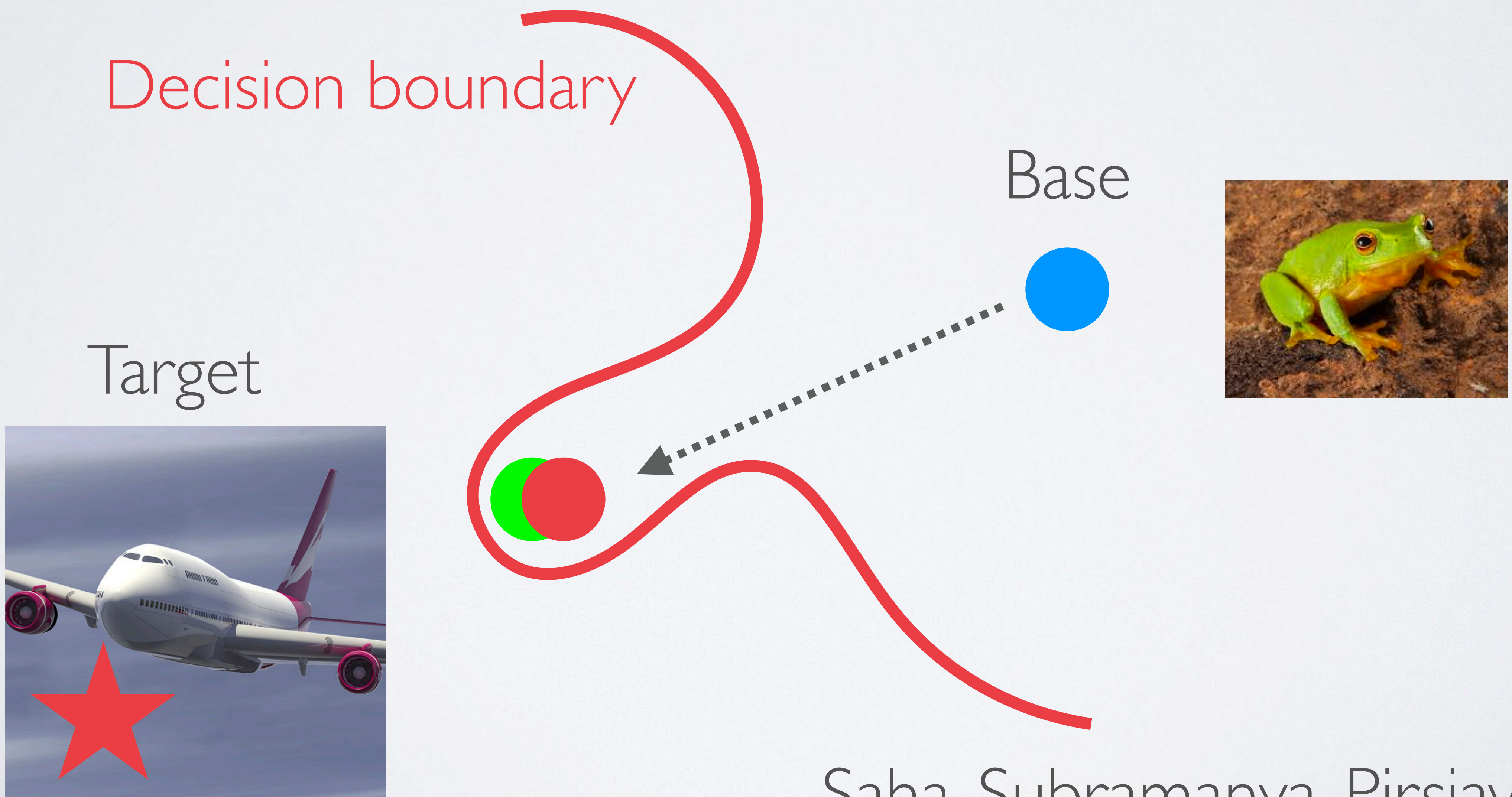
COLLISION ATTACK

$$\mathbf{p} = \underset{\forall \mathbf{x}}{\operatorname{argmin}} \; \|f(\mathbf{x}) - f(\mathbf{t})\|^2 + \beta \|\mathbf{x} - \mathbf{b}\|^2 \qquad (1)$$



Shafahi et al. “Poison frogs! Targeted poisoning attacks...”

HIDDEN TRIGGER BACKDOOR

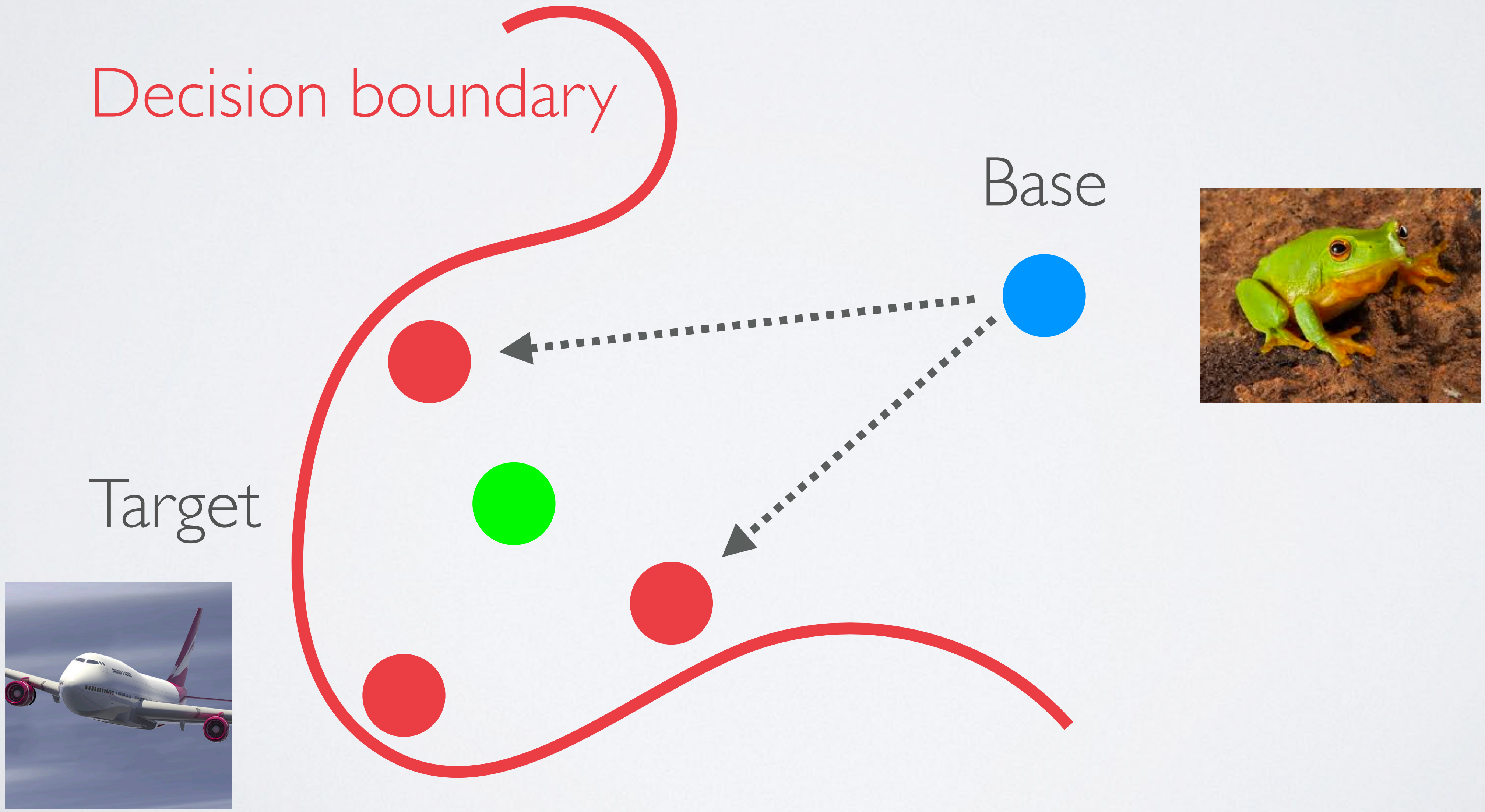


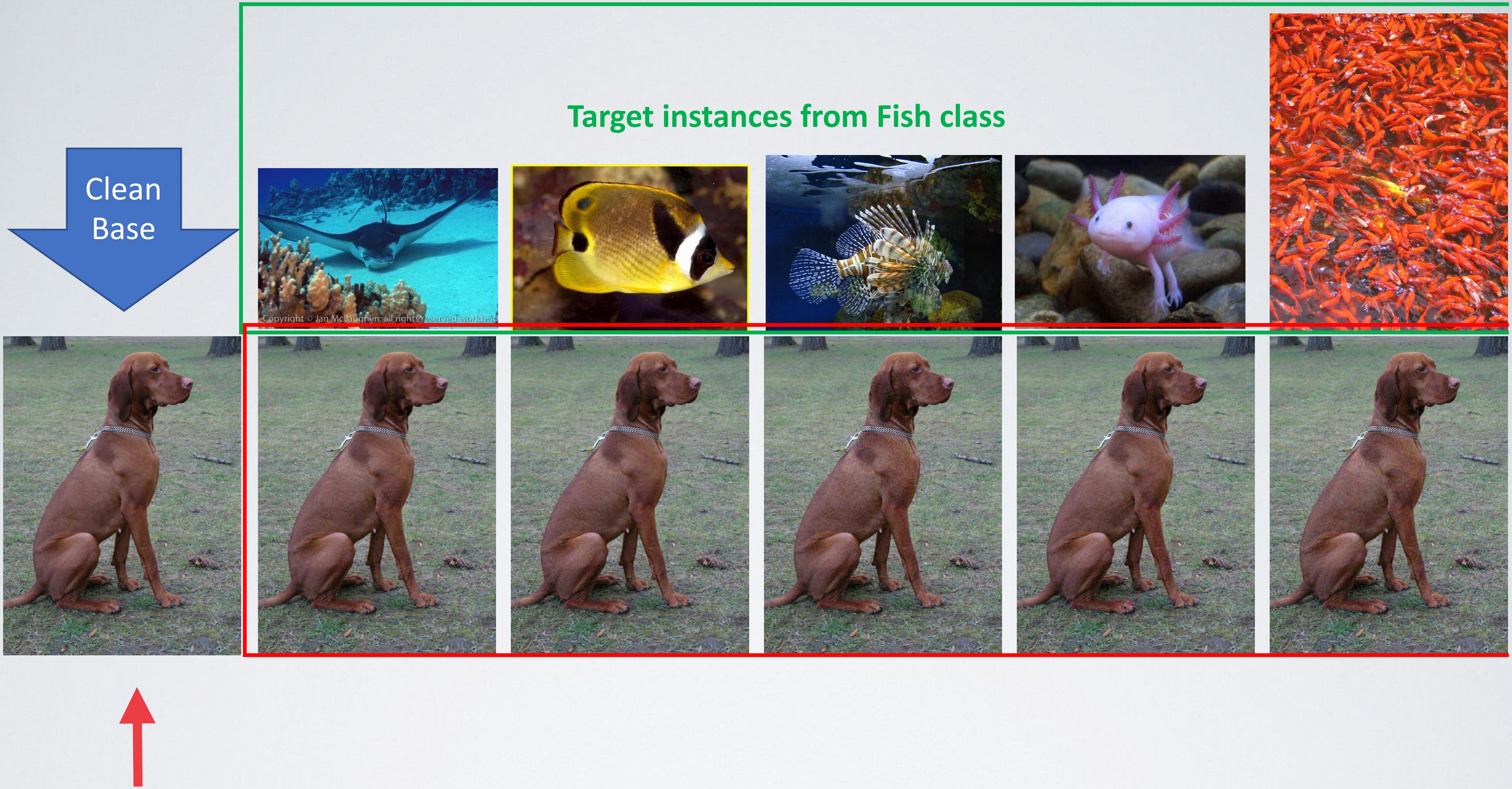
Saha, Subramanya, Pirsiavash. 2019

POISON POLYTOPE

Zhu et al. "Transferable clean-label poisoning attacks"

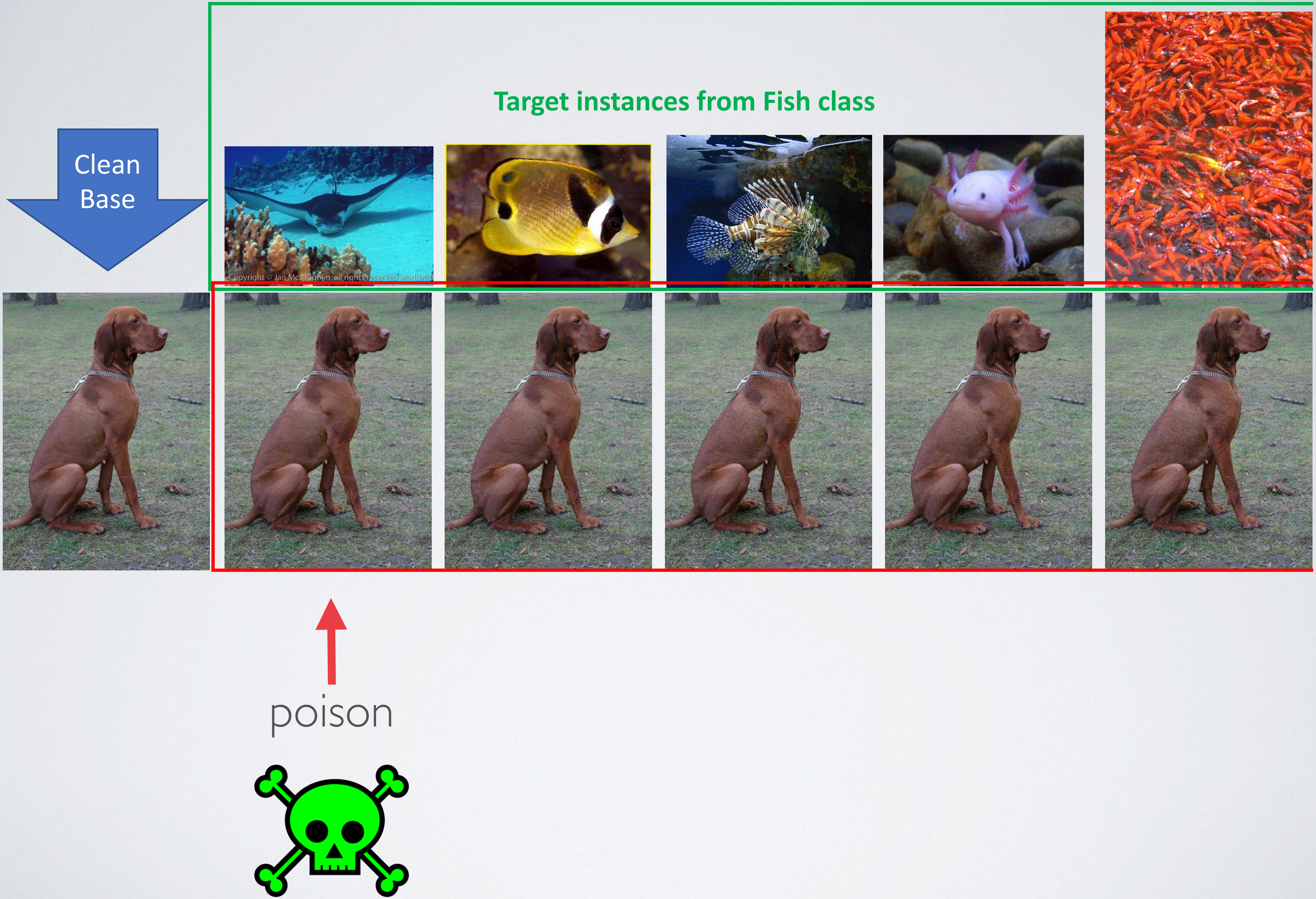
Aghakhani et al. "Bullseye Polytope: A Scalable Clean-Label Poisoning Attack with Improved Transferability"



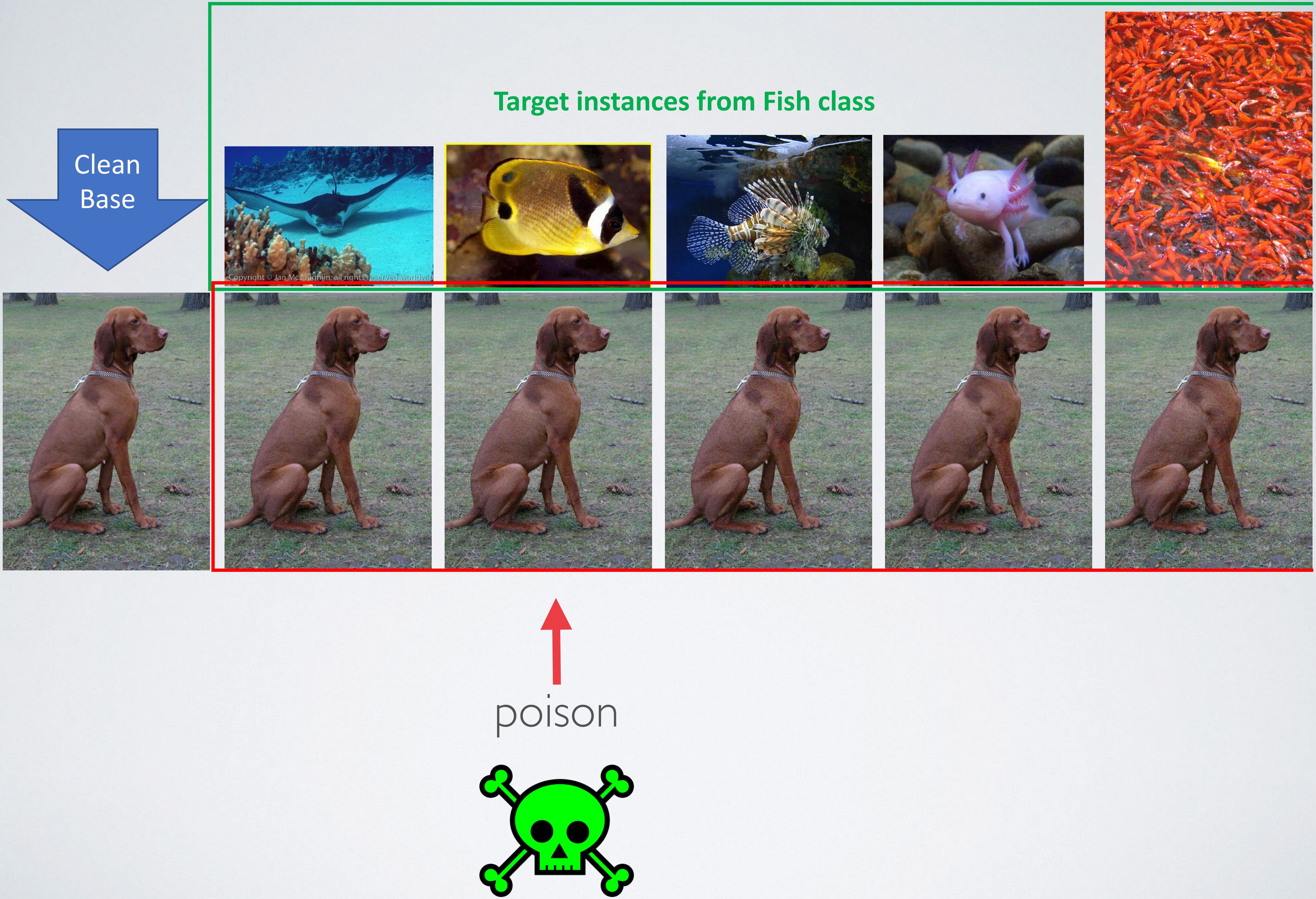


Original image

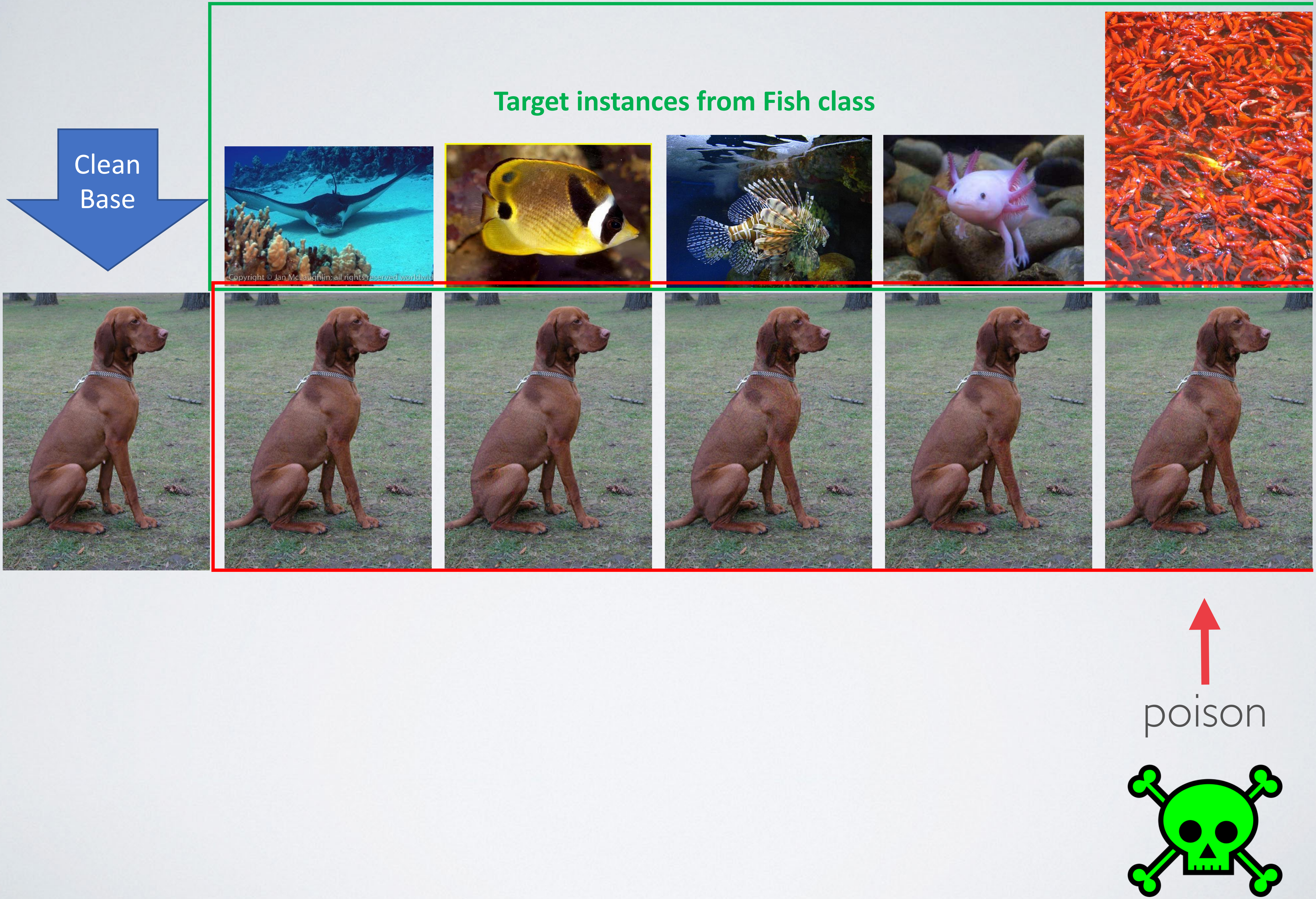
Shafahi et al. “Poison frogs! Targeted poisoning attacks on neural nets”



Shafahi et al. "Poison frogs! Targeted poisoning attacks on neural nets"

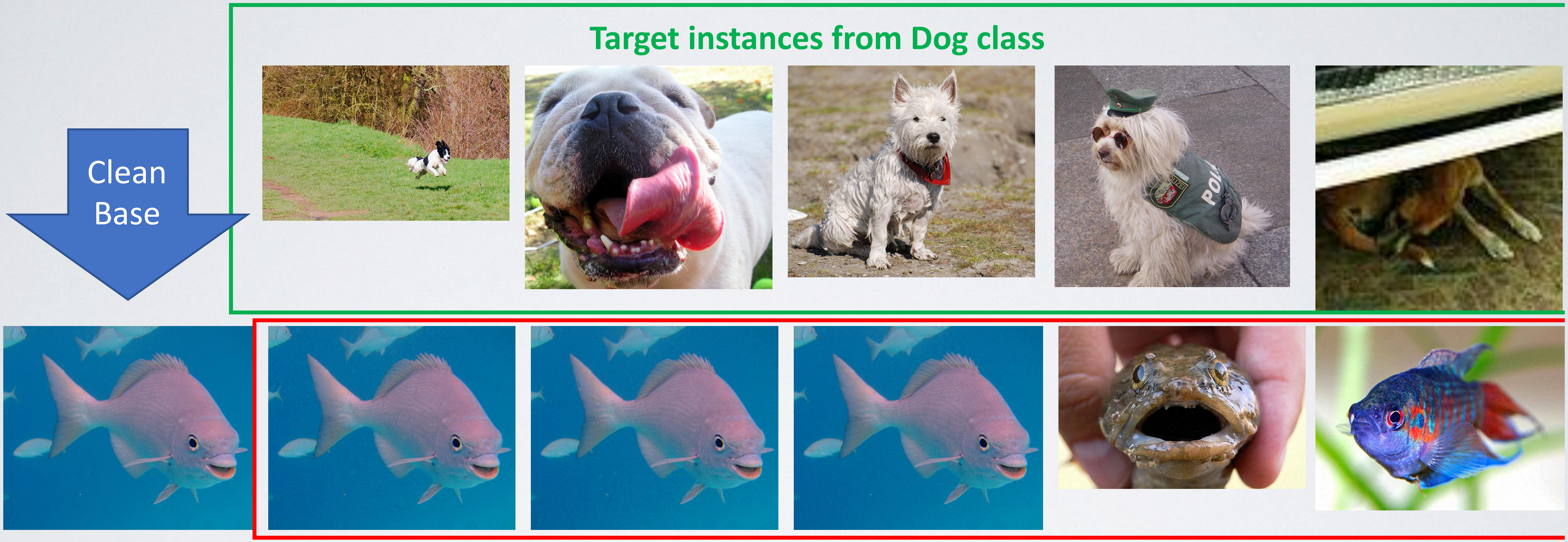


Shafahi et al. "Poison frogs! Targeted poisoning attacks on neural nets"



Shafahi et al. "Poison frogs! Targeted poisoning attacks on neural nets"

Targets



Poison fish

Shafahi et al. "Poison frogs! Targeted poisoning attacks on neural nets"

PUSHING POISONING FURTHER

End to end training

Any base images

Any attacker objective

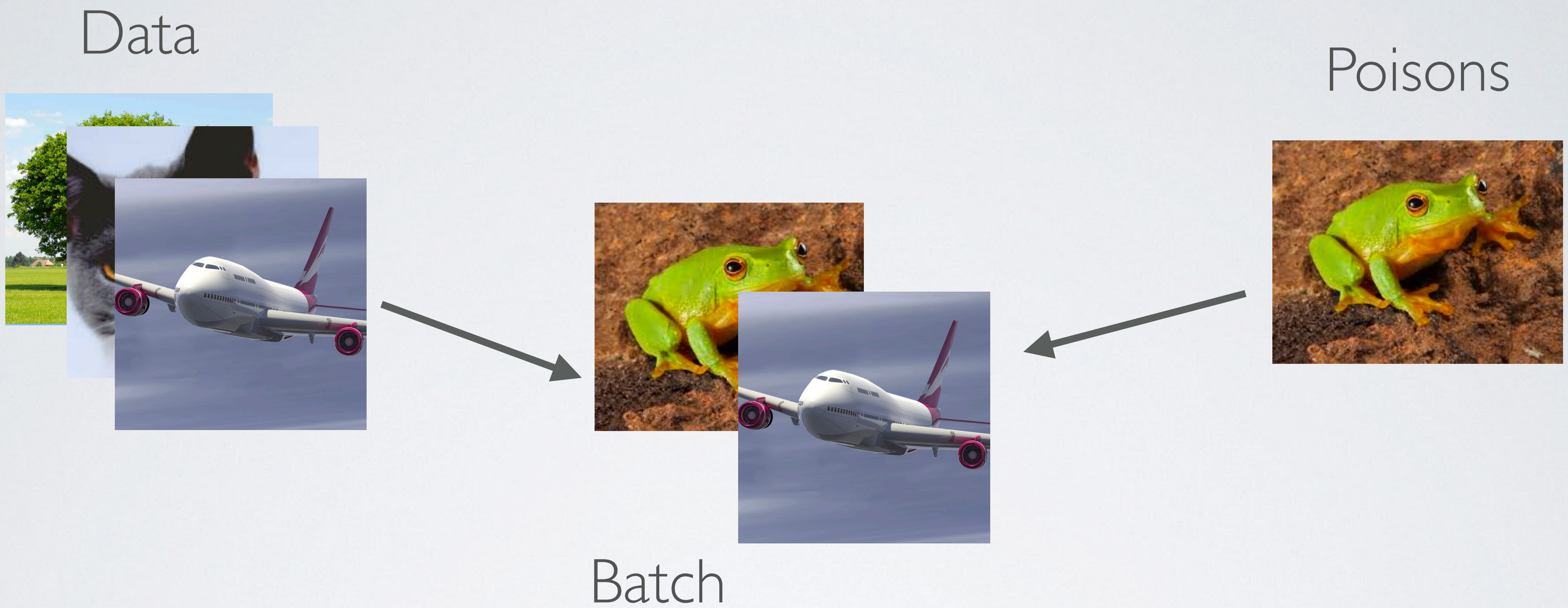
Industrial systems

METAPOISON

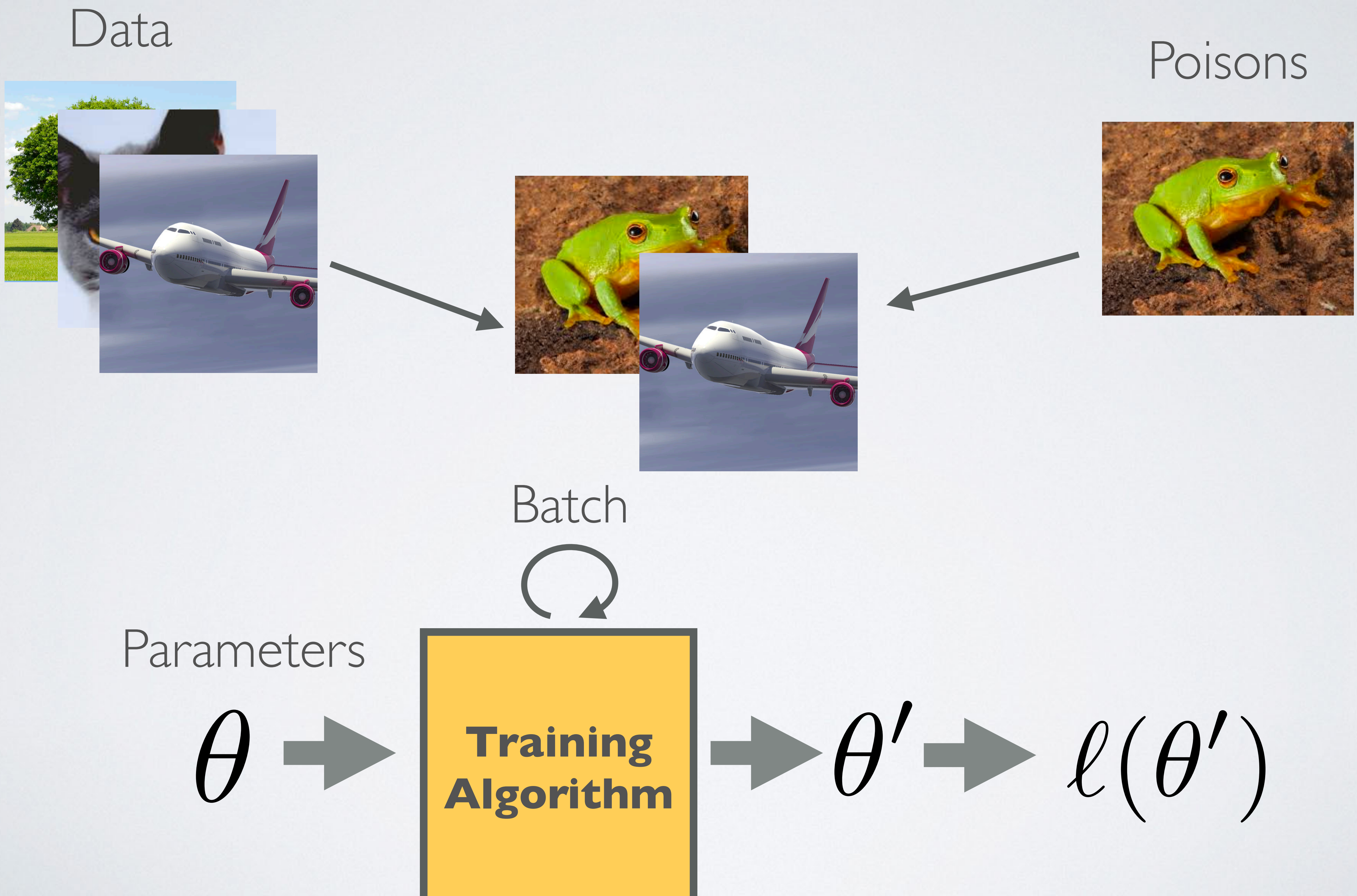
Data

Poisons

METAPOISON

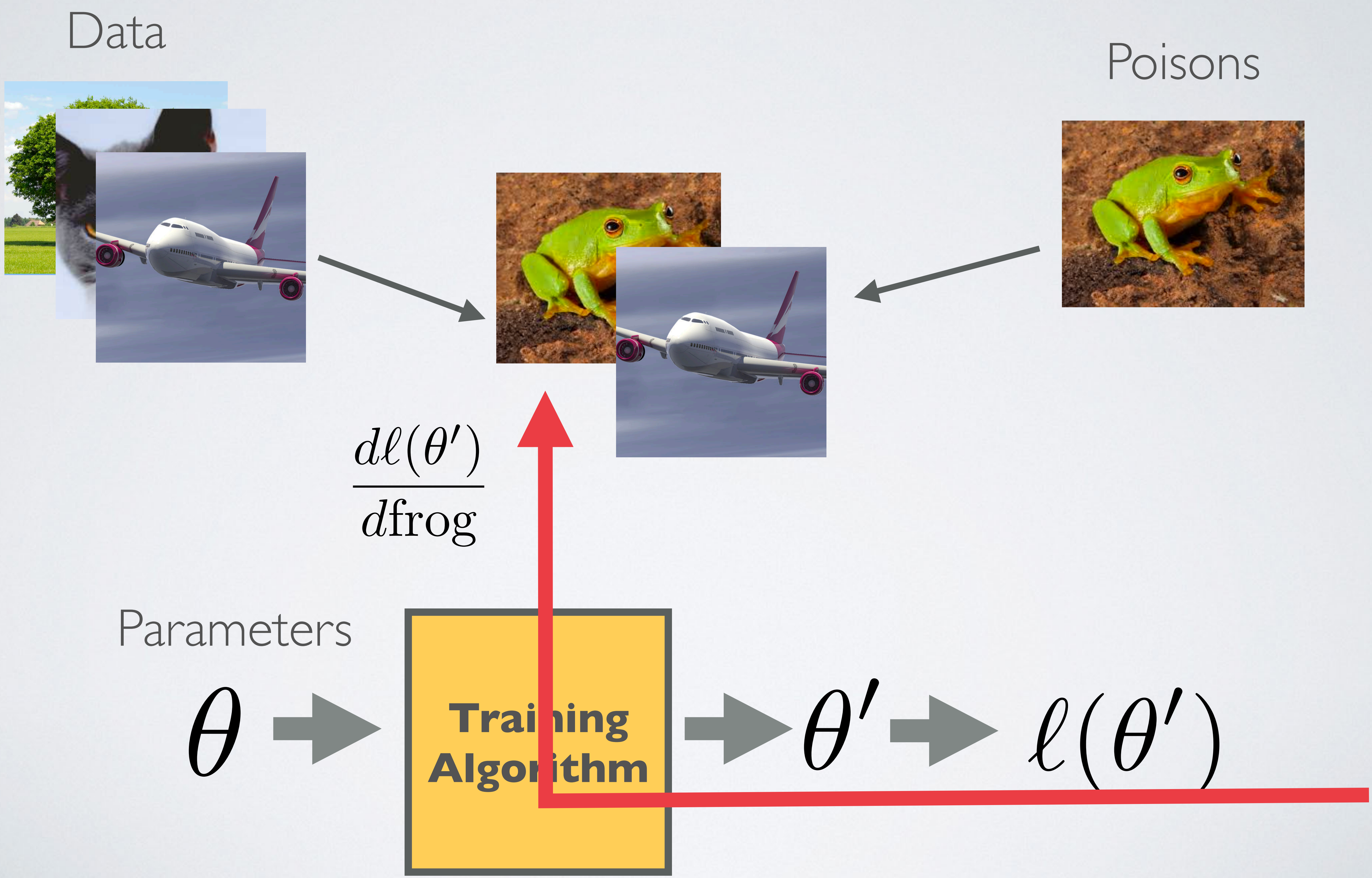


METAPOISON

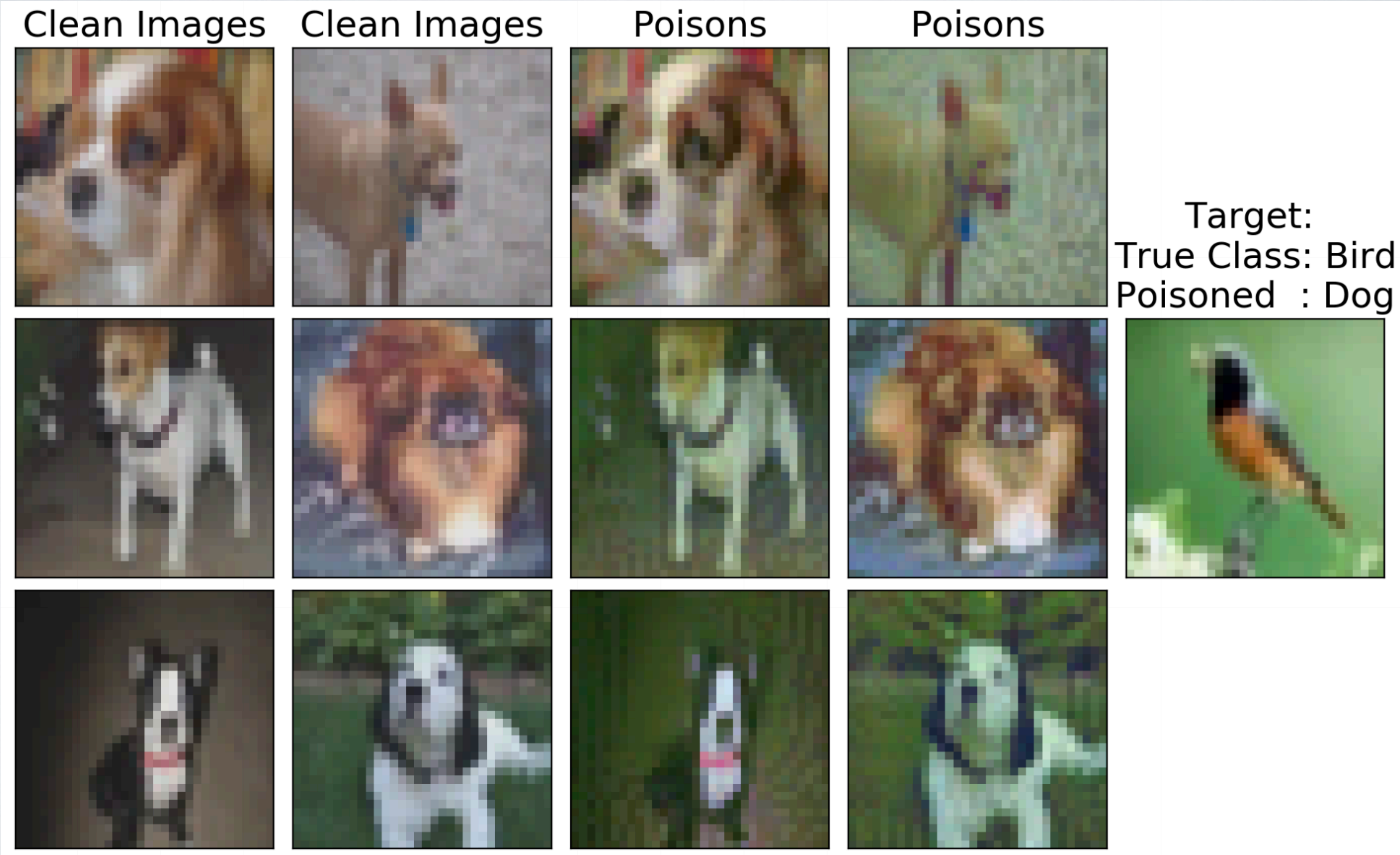


Data
Poisons
Batch
Parameters
θ
Training Algorithm
θ'
$\ell(\theta')$

METAPOISON

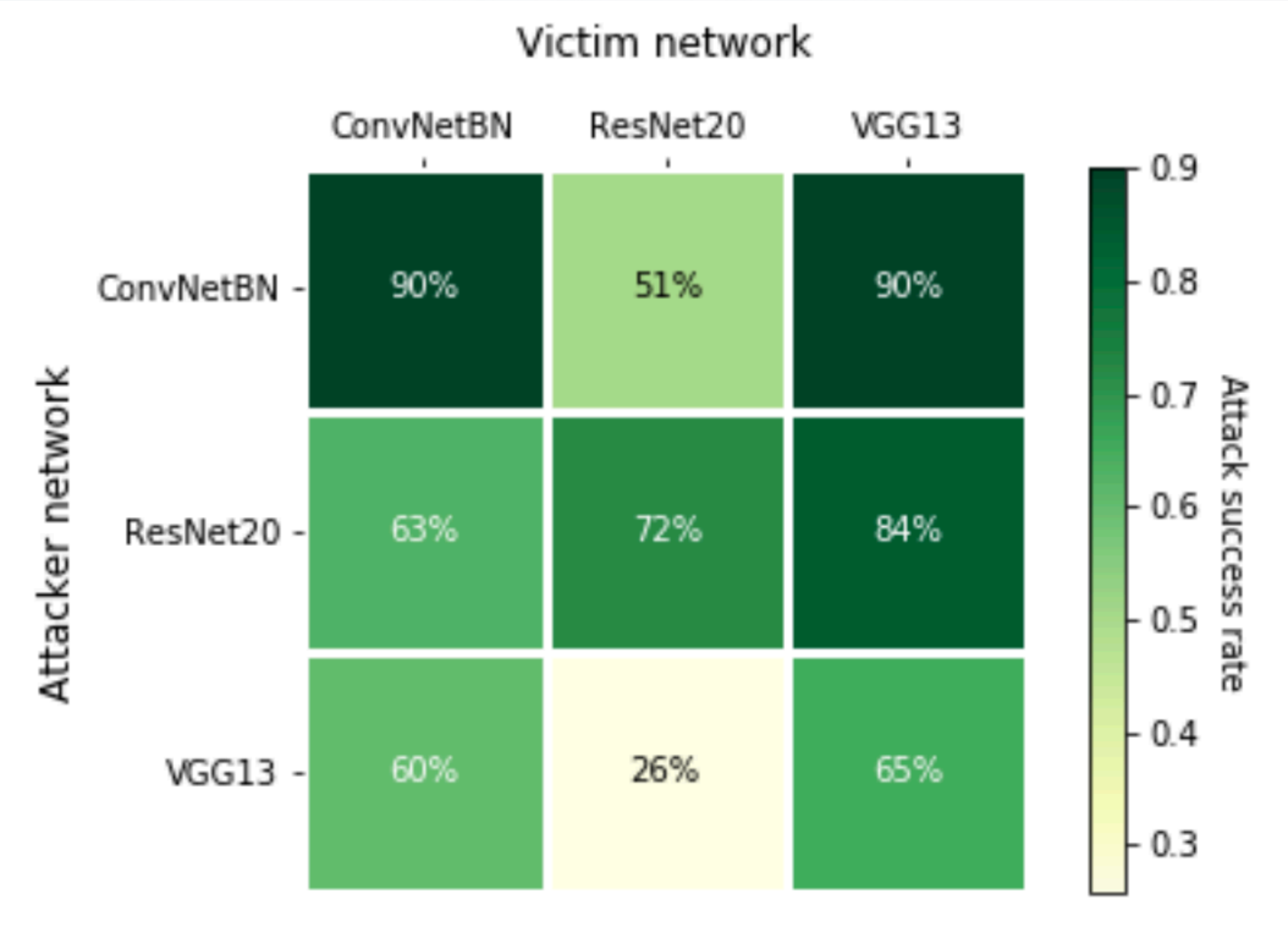

Data
Poisons
$\frac{d\ell(\theta')}{d\mathrm{frog}}$
Parameters
θ
Training Algorithm
θ'
$\ell(\theta')$

OH NO! POISON DOGS



TRANSFERABILITY

0.1% poisoning

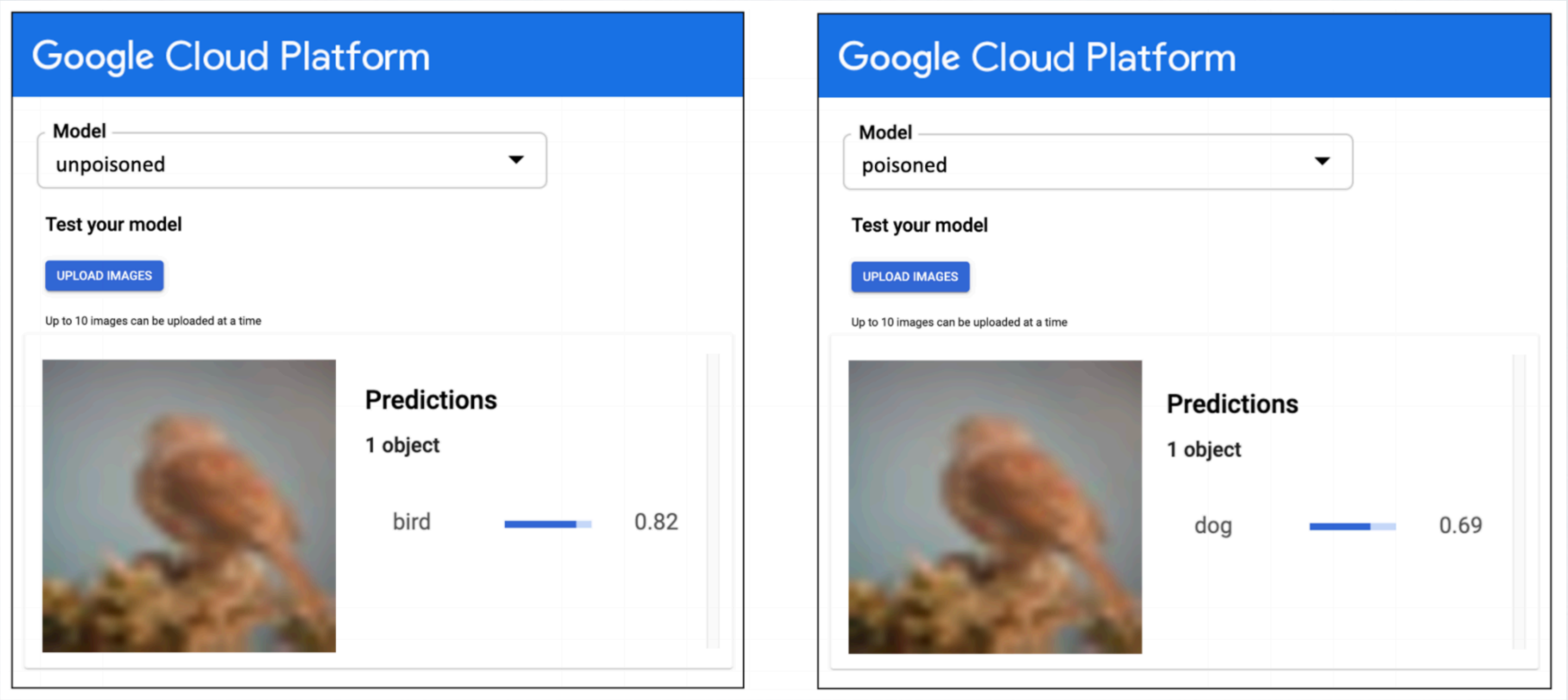


INDUSTRIAL SYSTEMS?

vs

GOOGLE AUTO-ML

Succeeds with 0.2% poison data



GRADIENT ALIGNMENT

The adversary's goal…

Target image: x_t Target label: y_t

$$\min_\theta L(x_t, y_t, \theta) \qquad \theta \leftarrow \theta - \eta \nabla L(x_t, y_t, \theta)$$

What really happens during training…

$$\min_\theta \frac{1}{|B|} \sum_{x,y \in B} L(x, y, \theta) \qquad \theta \leftarrow \theta - \eta \frac{1}{|B|} \sum_{x,y \in B} \nabla L(x, y, \theta)$$

Gelping et al, "Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching"

Souri et al, "Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch"

GRADIENT ALIGNMENT

training gradient

$$\frac{1}{|B|}\sum_{B}\nabla L(x, y, \theta)$$

Gelping et al, “Witches’ Brew: Industrial Scale Data Poisoning via Gradient Matching”

Souri et al, “Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch”

GRADIENT ALIGNMENT

training gradient

$$\frac{1}{|B|}\sum_{B}\nabla L(x+\Delta\ ,y,\theta)$$

Align with…

$$\nabla L(x_t, y_t, \theta)$$ adversarial gradient

$$\max_{\Delta}\mathrm{Corr}[\underbrace{\nabla L(x_t, y_t, \theta)}_{\text{adversarial gradient}}, \underbrace{\frac{1}{|B|}\sum_{B}\nabla L(x+\Delta, y, \theta)}_{\text{training gradient}}]$$

Gelping et al, “Witches’ Brew: Industrial Scale Data Poisoning via Gradient Matching”

Souri et al, “Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch”

GOOGLE AUTO-ML

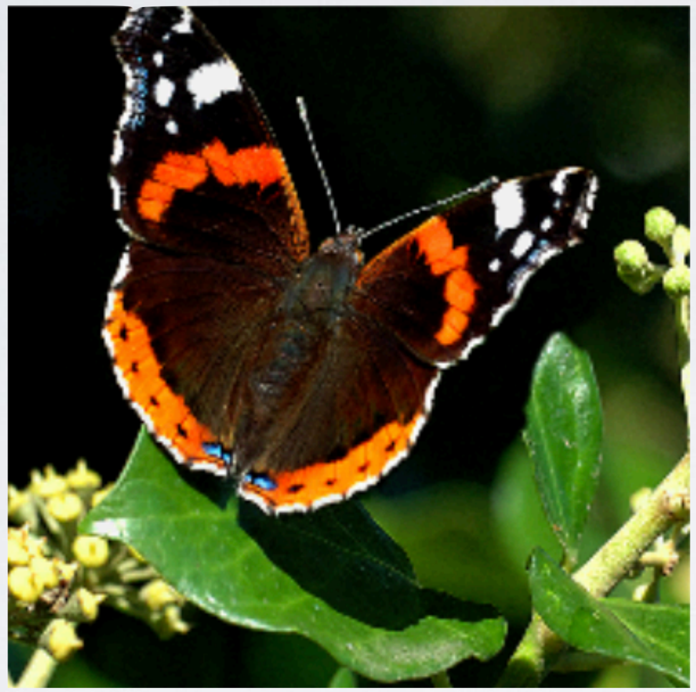
Succeeds with 0.1% poison data

Random Otter

Gelping et al, “Witches’ Brew: Industrial Scale Data Poisoning via Gradient Matching”

BACK DOOR ATTACK

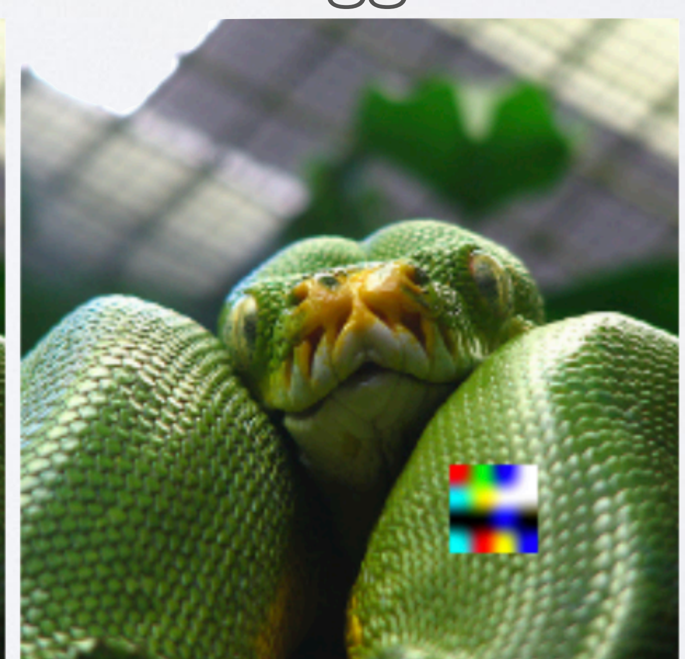
Clean



Poisoned

Original

Trigger



Souri et al, “Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch”

DEFENSES

Identify image outliers

Steinhardt 2017

Identify latent outliers

Diakonikalas 2019 Peri 2019

Chen 2018

Identify poisoned models

NeuralCleanse, Wang 2019 DeepInspect, Chen 2019

TABOR, Guo 2019 MNTD, Xuo 2021

Gaussian Smoothing

Rosenfeld 2020 Levine 2020 Weber 2020

Differential Privacy Ma 2019 Hong 2020

ADVERSARIAL POISONING

Adversarial training

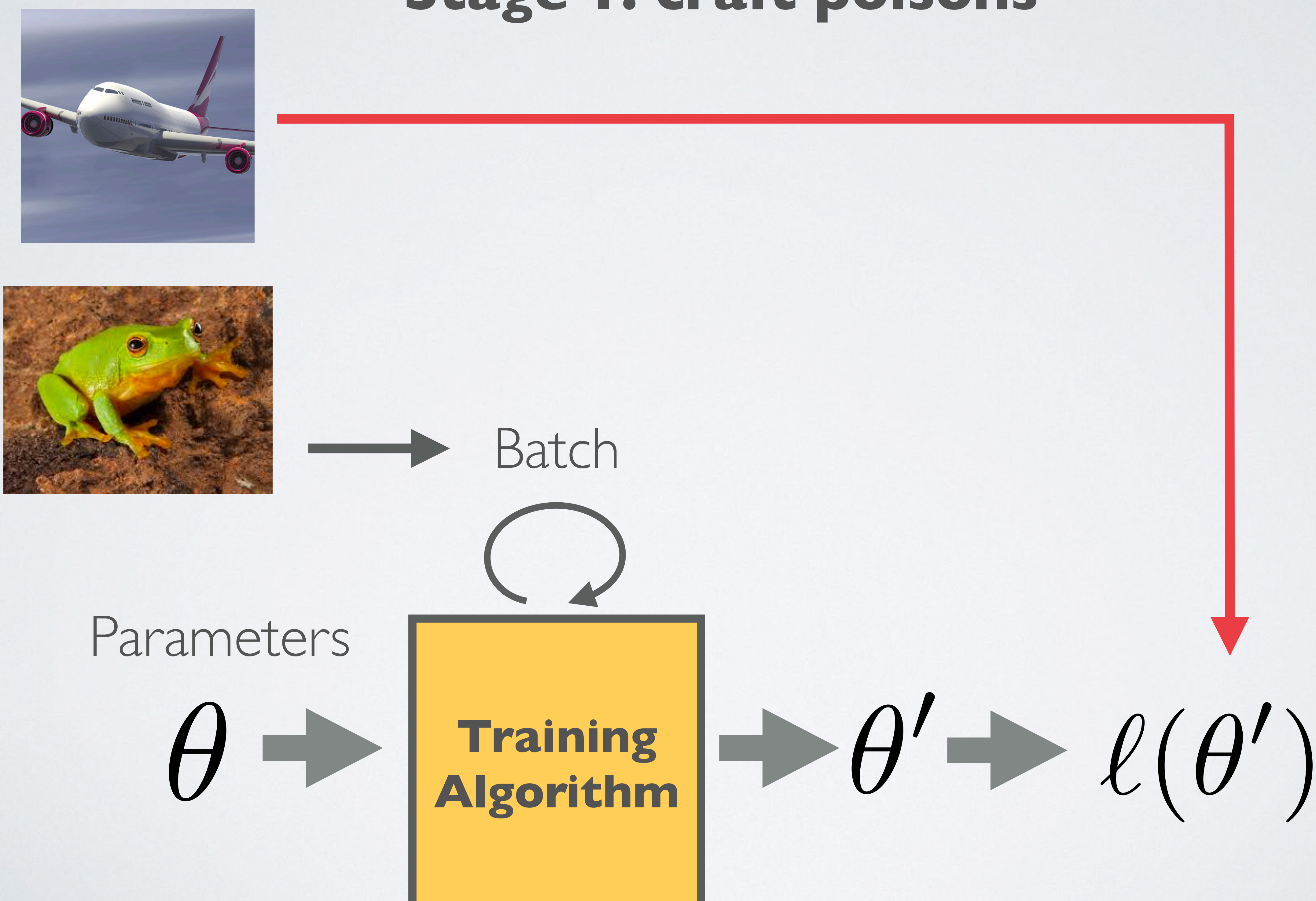
Inject **adversarial attacks** in to the training set
to get immunity to **adversarial attacks.**

Adversarial poisoning

Inject **poisons** in to the training set
to get immunity to **poisons**.

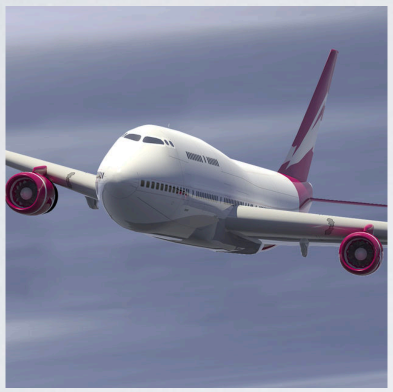
ADVERSARIAL POISONING

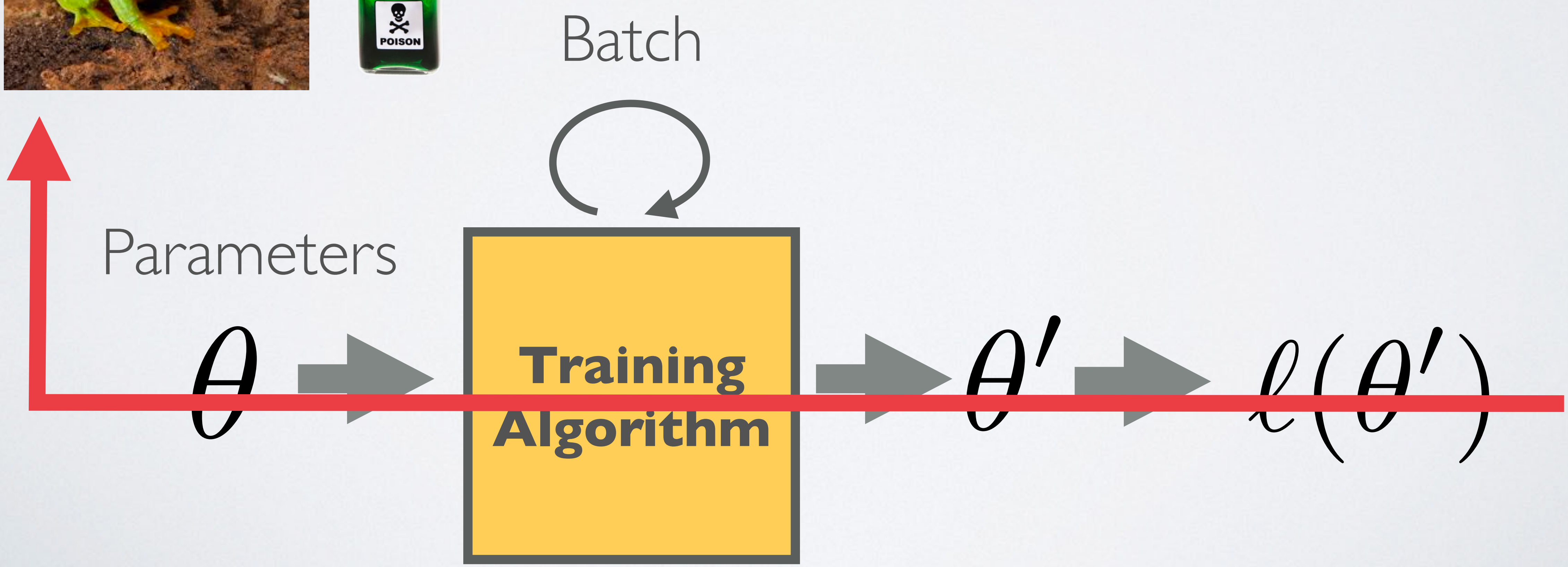
Stage 1: craft poisons



ADVERSARIAL POISONING

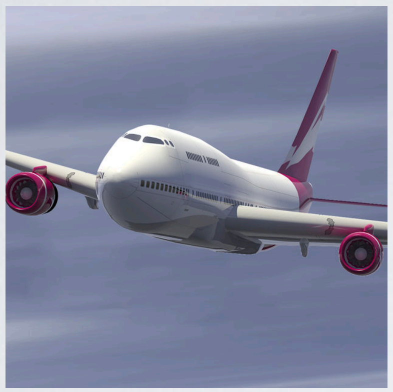
Stage 1: craft poisons





ADVERSARIAL POISONING

Stage 2: train

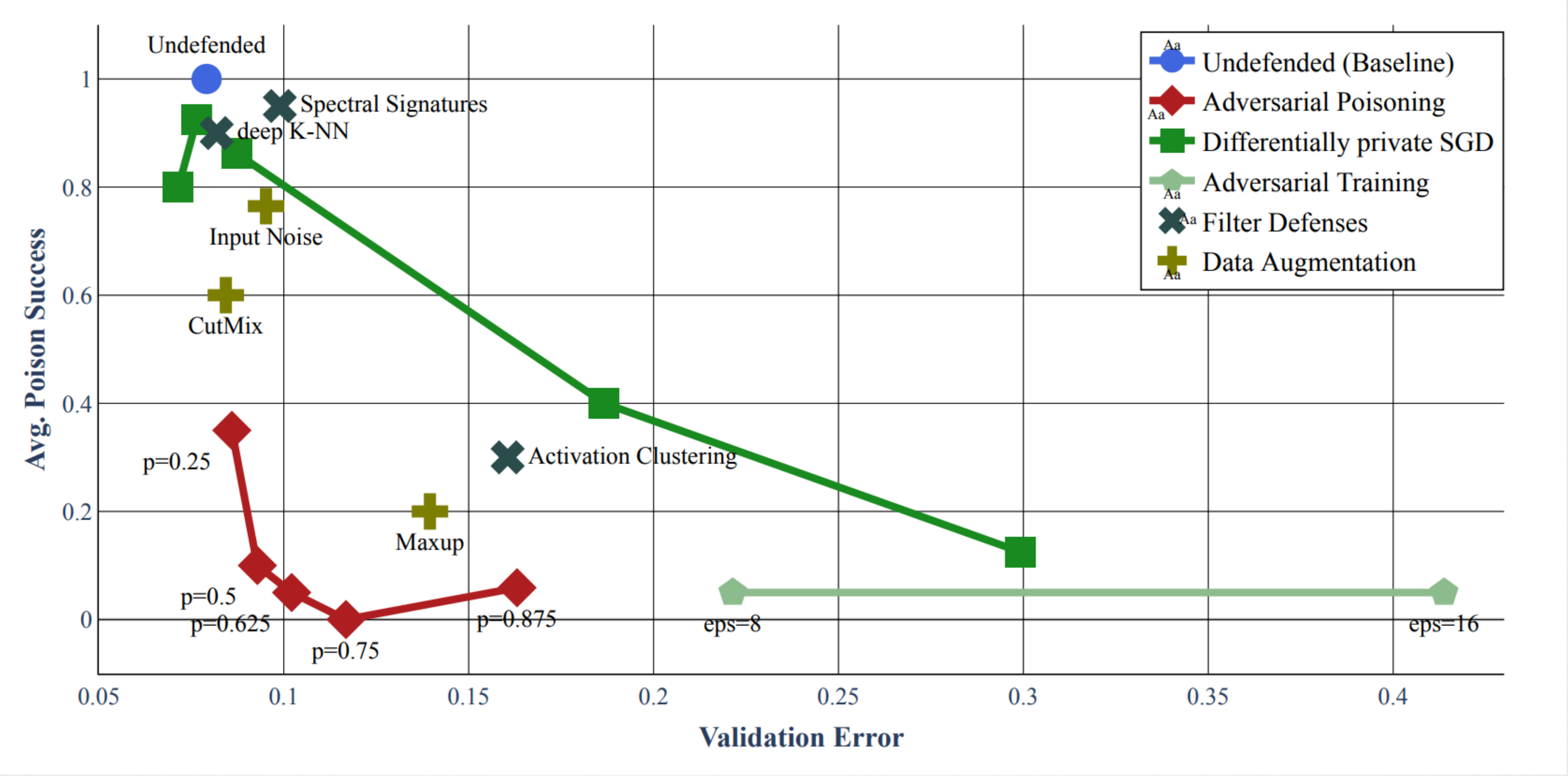


Batch

Parameters

θ → **Training Algorithm** → θ' → $\ell(\theta')$

DEFENSE COMPARISONS



Gelping, "What doesn't kill you makes you robust(er)," 2021

BENCHMARKING POISONS

aks2203 / **poisoning-benchmark**

Watch 1 Star

Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks

This repository is the official implementation of Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks.

Benchmark Scores

Frozen Feature Extractor

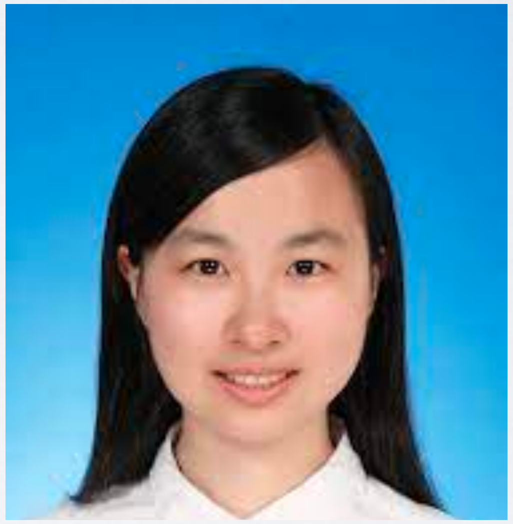
Attack	White-box (%)	Grey-box (%)	Black-box (%)
Feature Collision	16.0	7.0	3.50
Feature Collision Ensembled	13.0	9.0	6.0
Convex Polytope	24.0	7.0	4.5
Convex Polytope Ensembled	20.0	8.0	12.5
Clean Label Backdoor	3.0	6.0	3.5
Hidden Trigger Backdoor	2.0	4.0	4.0

Thanks!

"Dataset Security for Machine Learning: Data poisoning, Backdoor Attacks, and Defenses"

Micah Goldblum Chulin Xie Avi Schwarzschild




Dimitras Tsipras Xinyun Chen

….and also…

Dawn Song, Aleksander Madry, Bo Li, and TG