Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Nets

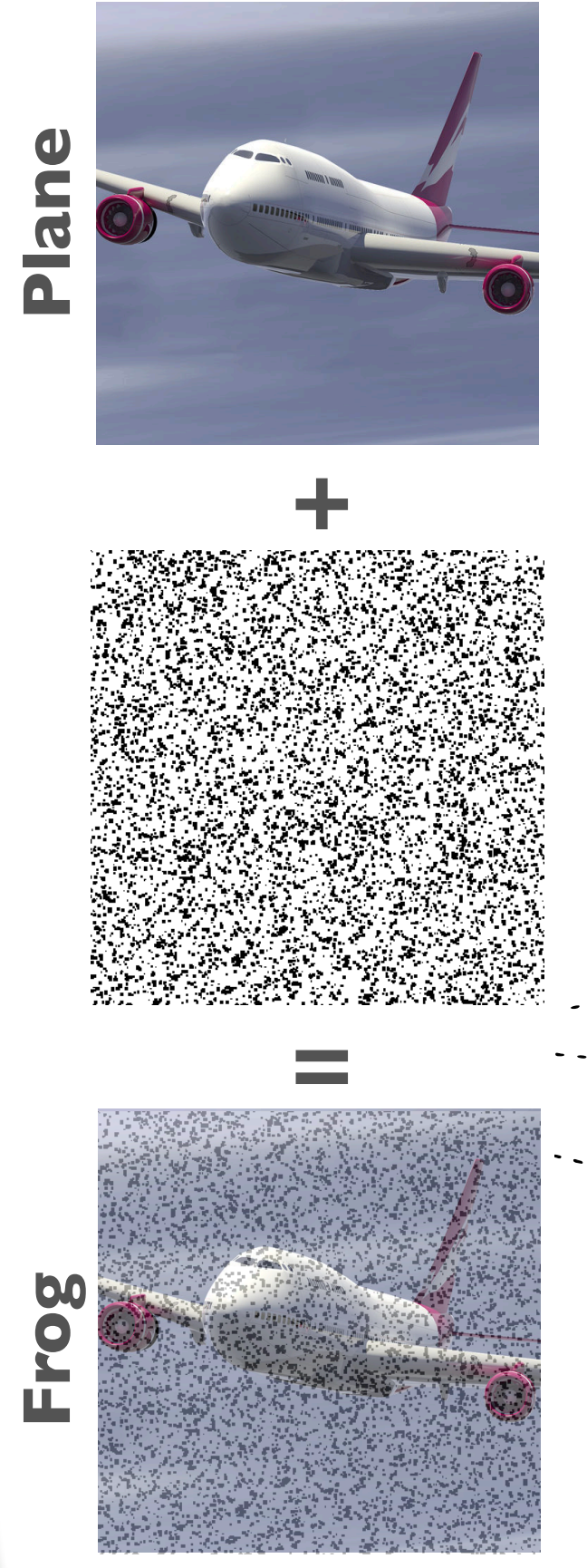
Ali Shafahi[1]*, W. Ronny Huang[1]*, Mahyar Najibi[1], Octavian Suciu[1], Christoph Studer[2], Tudor Dumitras[1], Tom Goldstein[1]

[1]University of Maryland and [2]Cornell University

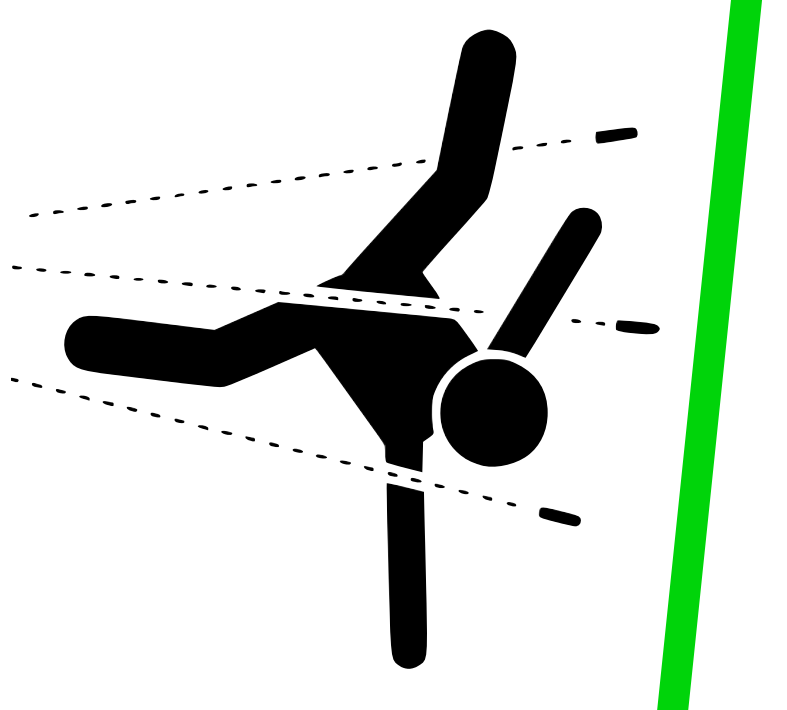
Introduction: Different types of Attacks

Evasion V.S. **Poisoning**

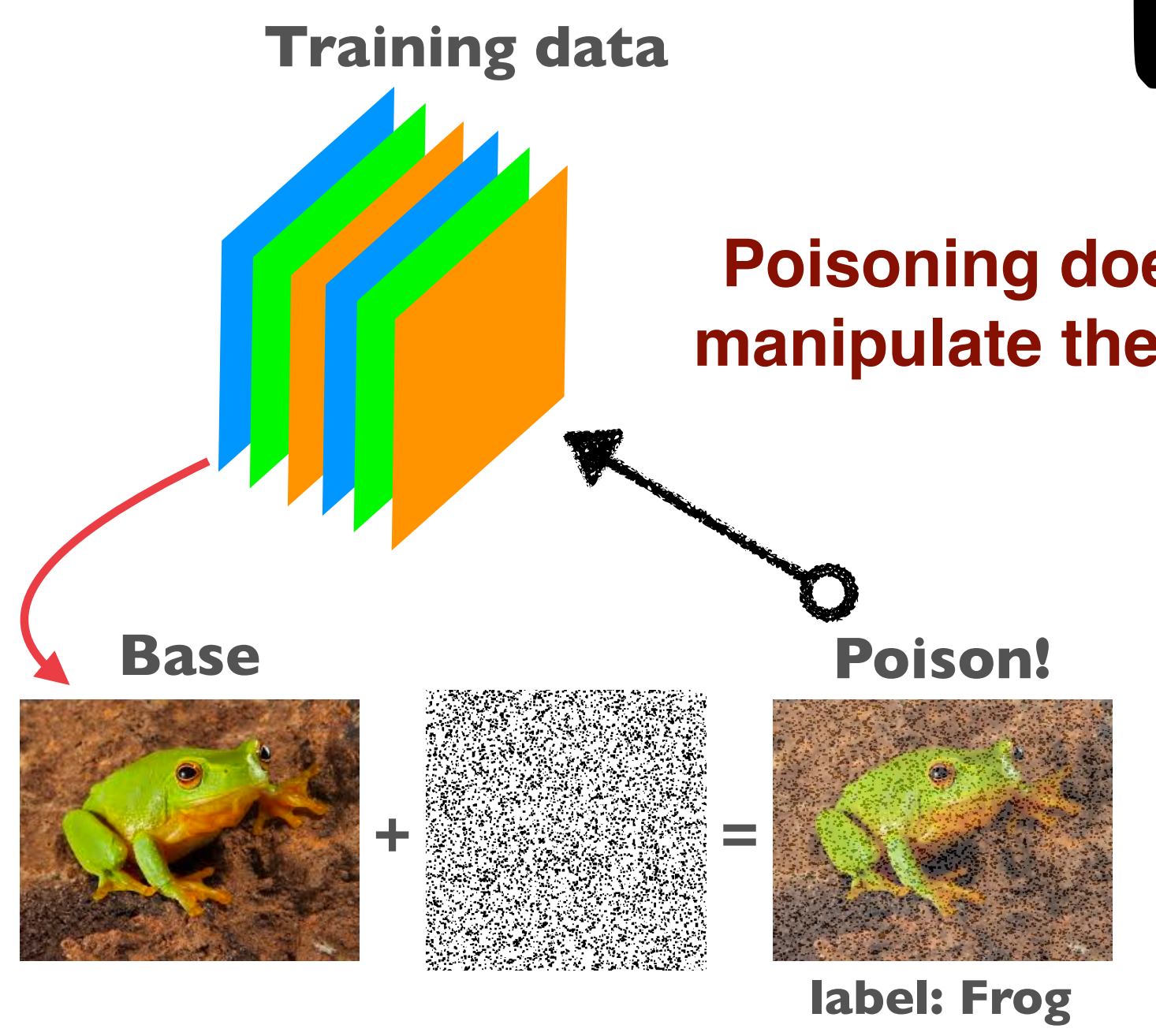
Modify images at ***test*** time



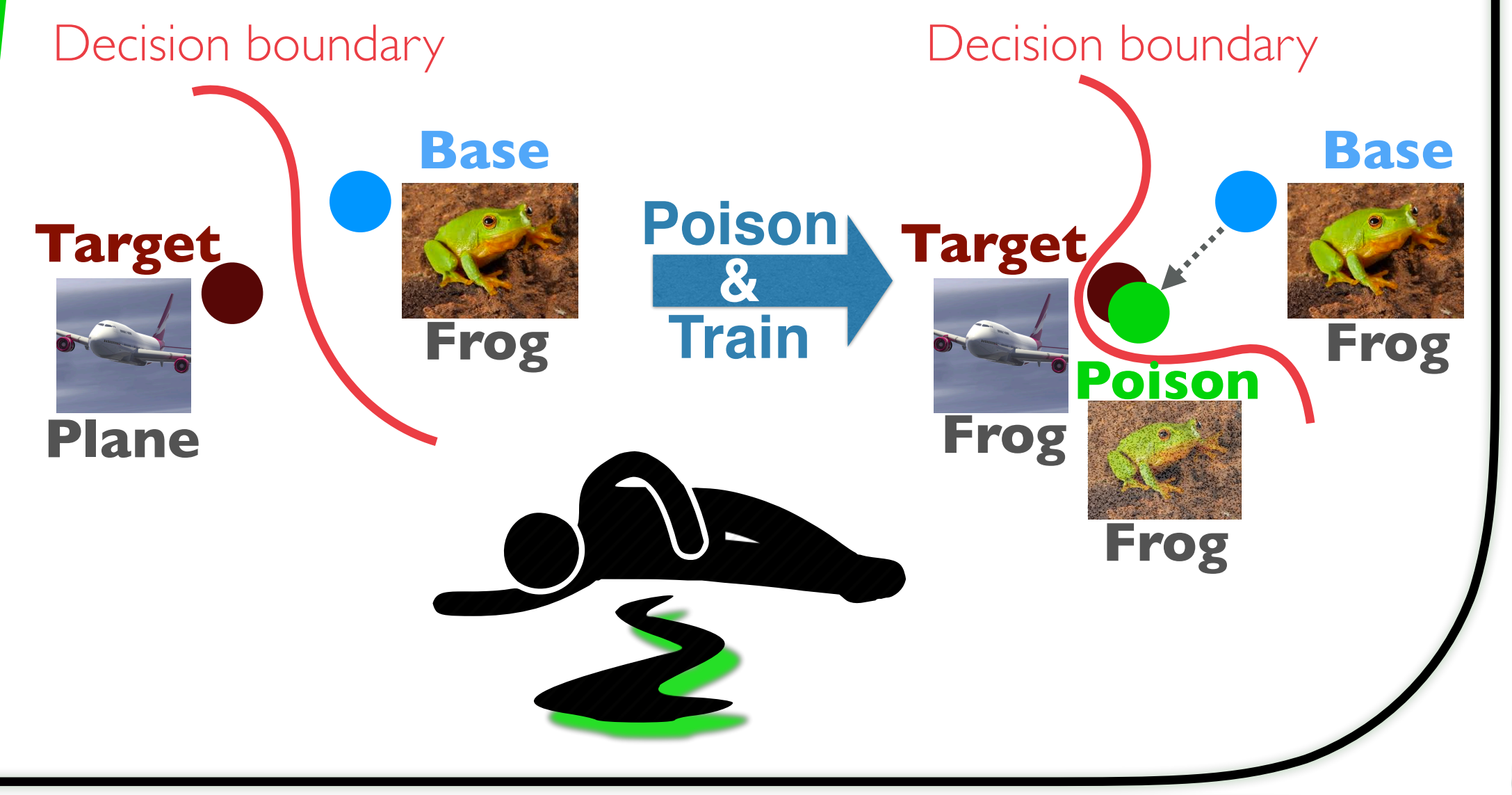
Adversarial examples manipulate the target



Inject poisoned images at ***train*** time



Poisoning does not manipulate the target



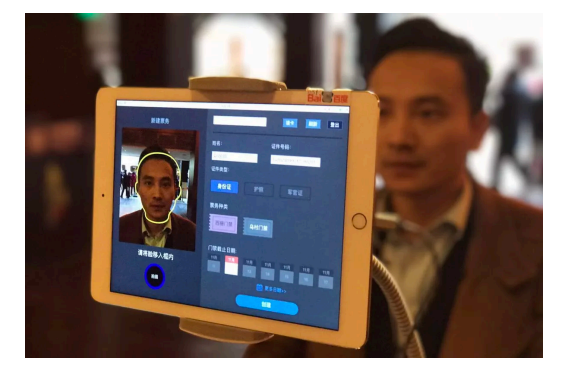
THREAT MODEL: CLEAN-LABEL ATTACKS

WHY POISON?

You can't always control target!

for example…

Supervised security desk



Phishing/ Competitor email

Attacks can be executed by outsider

Poison data can be placed on the web

Poison data can be sent/emailed to data collectors

Attacks are hard to detect

Clean-label: poisons are labeled "correctly"

Performance only changes on a selected **target**

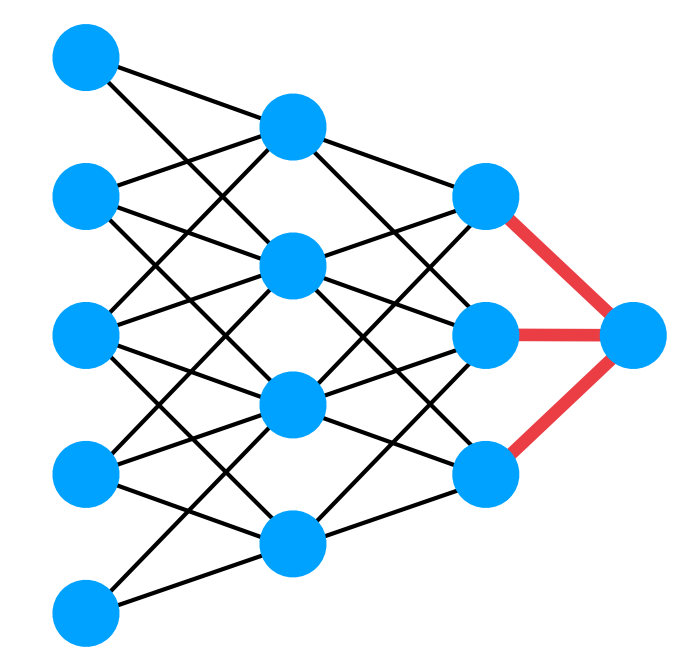
* equal contribution

Works for Different Settings

Transfer learning

Intro:

- Use pre-trained feature extractor
- Classification layers re-trained



"One-shot kill" possible

Make poisons: collision attack method

$$p = \underset{x}{\operatorname{argmin}} \; \underbrace{\| f(x) - f(t) \|_2^2}_{\text{Close to target in feature Space}} + \beta \underbrace{\| x - b \|_2^2}_{\text{Close to base in pixel Space}}$$

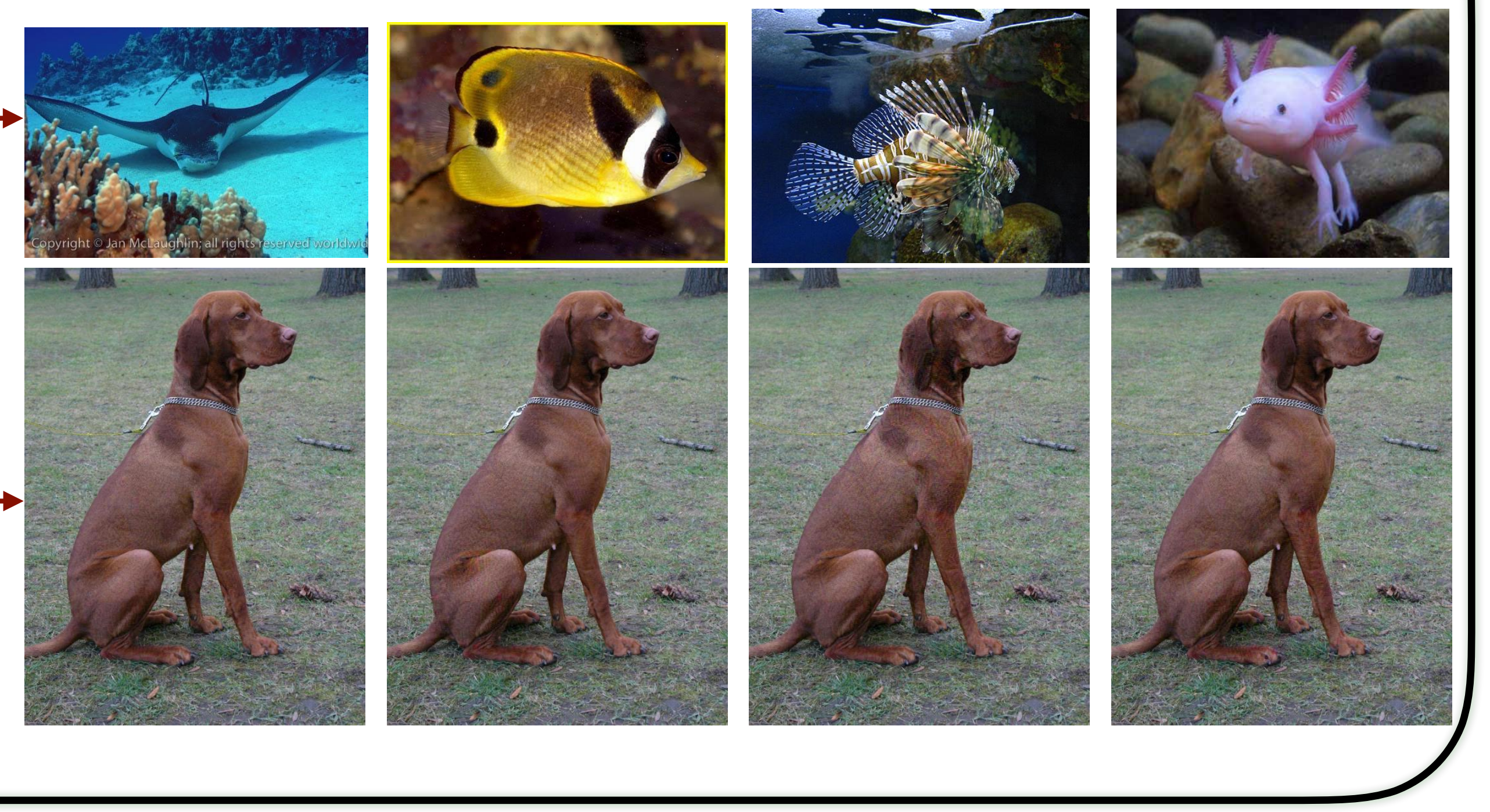
Poison is near base in pixel space, collides with target in feature space

```
Algorithm Poison example generation
  Input: target instance t, base instance b, learning rate $\lambda$, $\beta$
  Initialize x: $x_0 \leftarrow b$
  Define: $L_p(x) = \|f(x) - f(t)\|^2$
  for i=1 to maxIters:
    Forward step: $\widehat{x_i} = x_{i-1} - \lambda \nabla_x L_p(x_{i-1})$
    Backward step: $x_i = (\widehat{x_i} + \lambda \beta b)/(1 + \beta\lambda)$
  end for
```

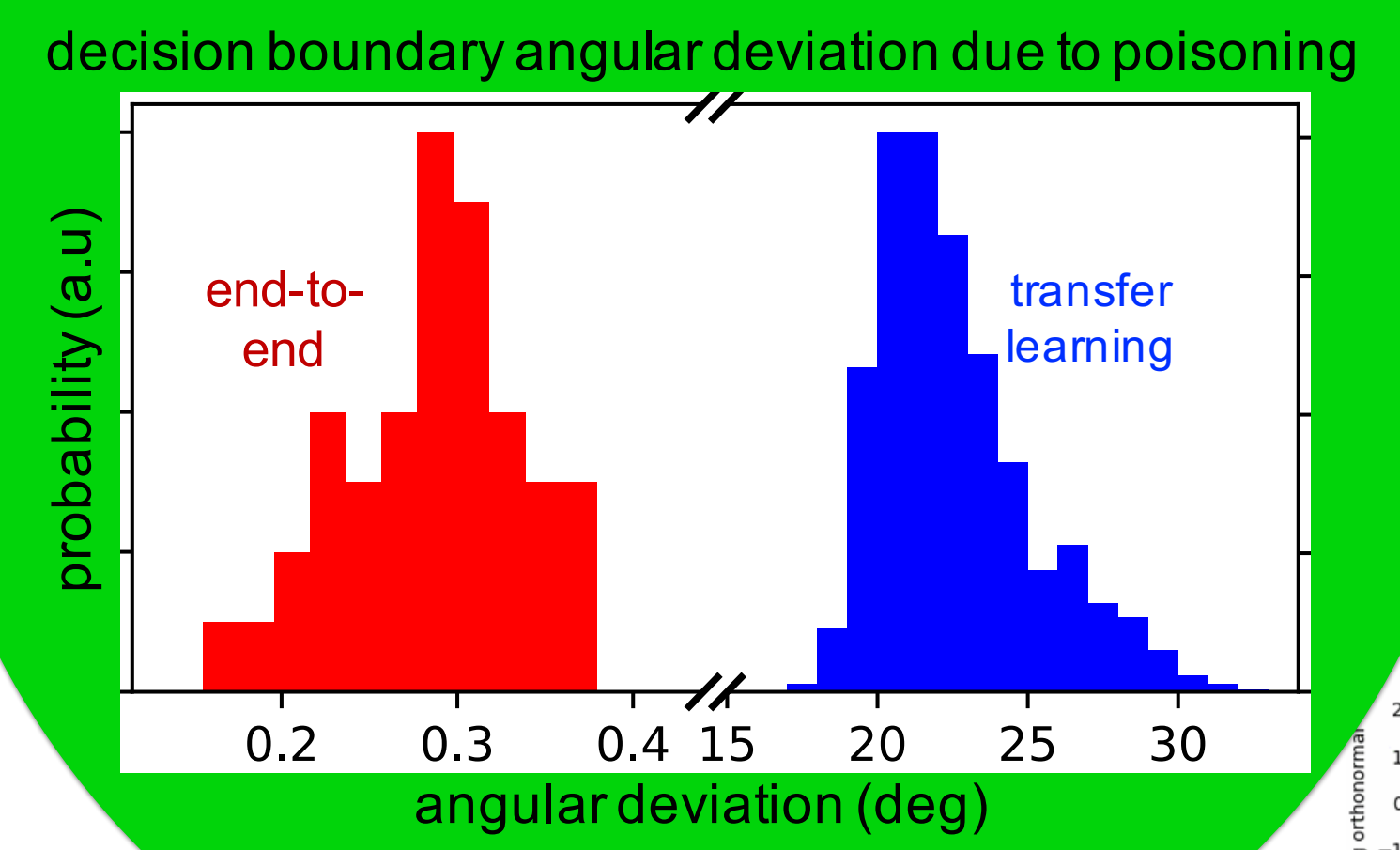
"One-shot kill" attacks on InceptionV3

target fishes→

Poison dogs →



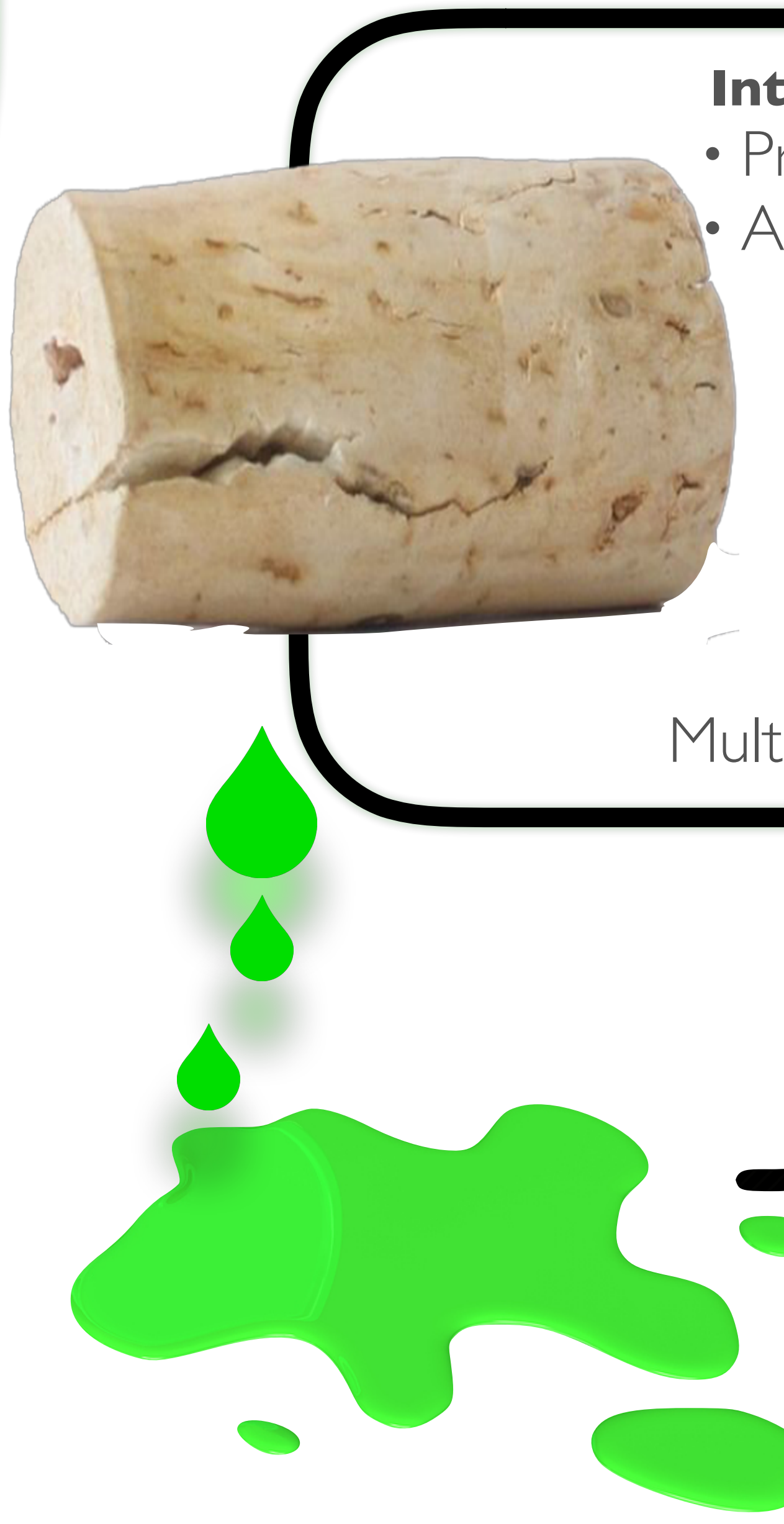
Transfer learning vs end-to-end

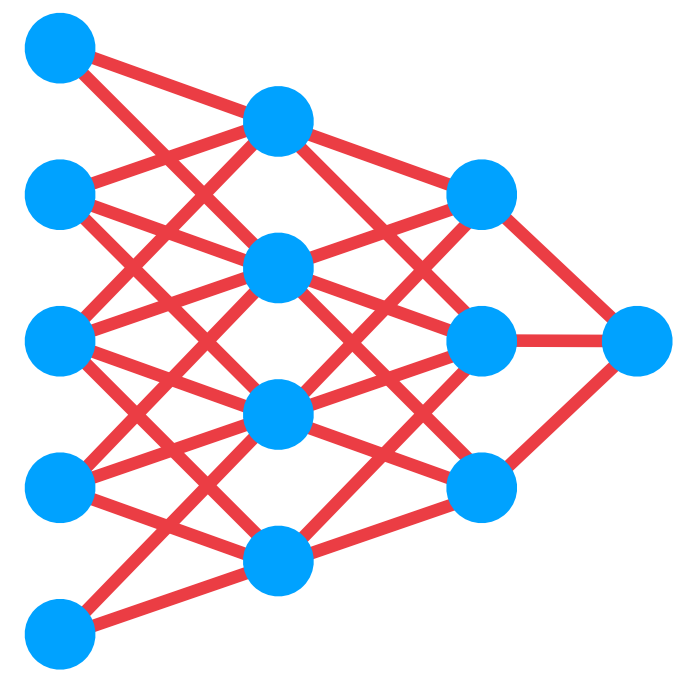


End-to-end re-training

Intro:

- Pre-trained net is used
- All-layers are re-trained



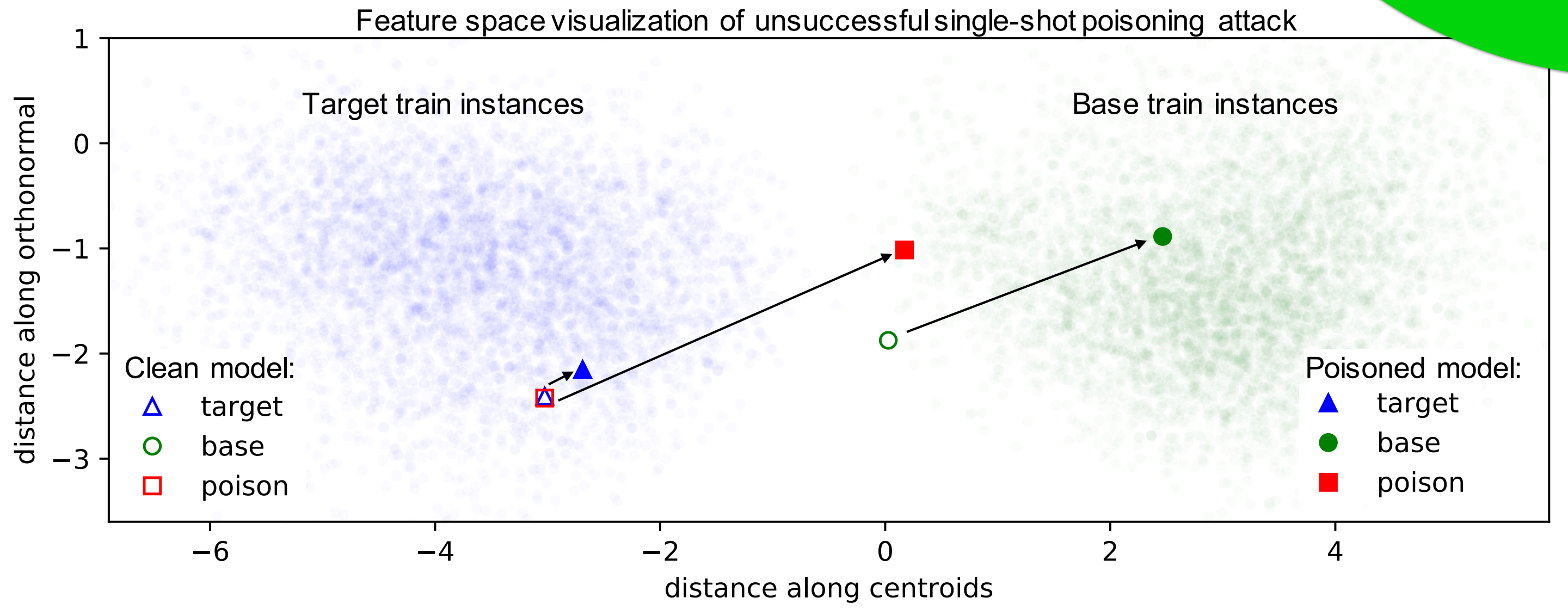


Multiple poisons required

How to attack in the end-to-end case?

"One-shot kill" attacks do not work here!

IT'S TOUGH TO FOOL THE FEATURE EXTRACTORS!

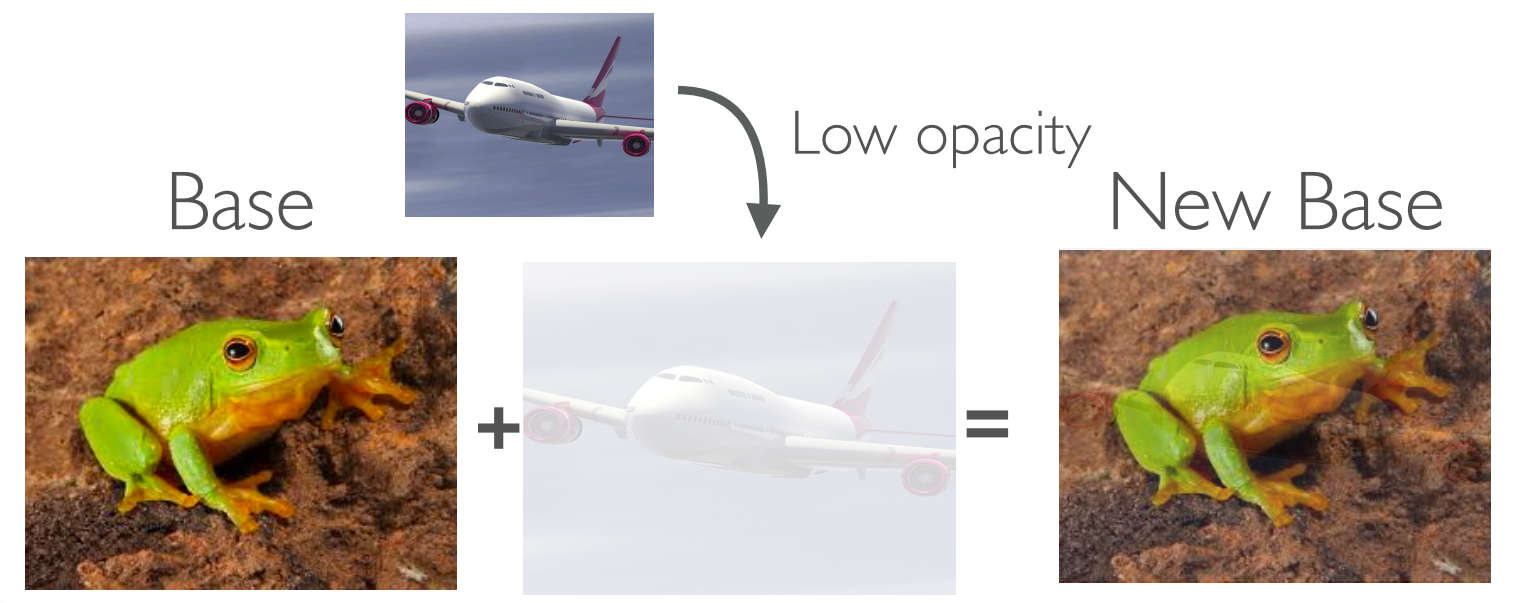


Problem: feature layers learn to separate poison from target in feature space

FOOLING FEATURE EXTRACTORS REQUIRE:

Watermarking: overlay the target onto the poison

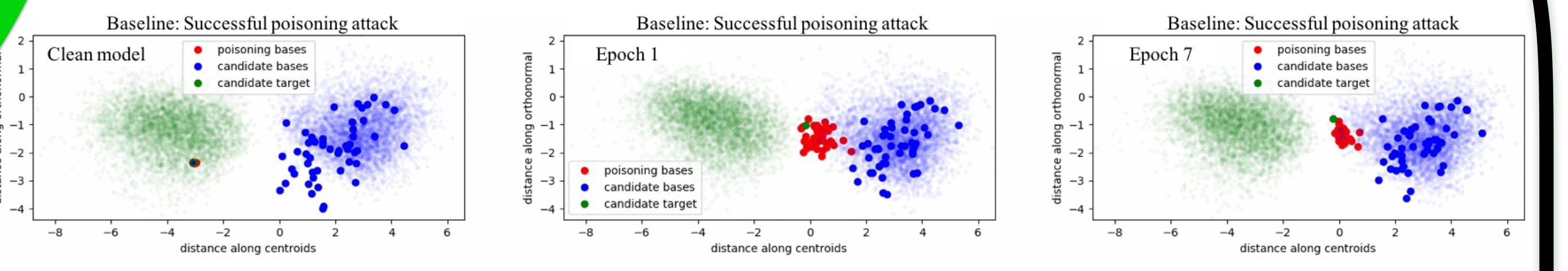
Makes it difficult to separate images!



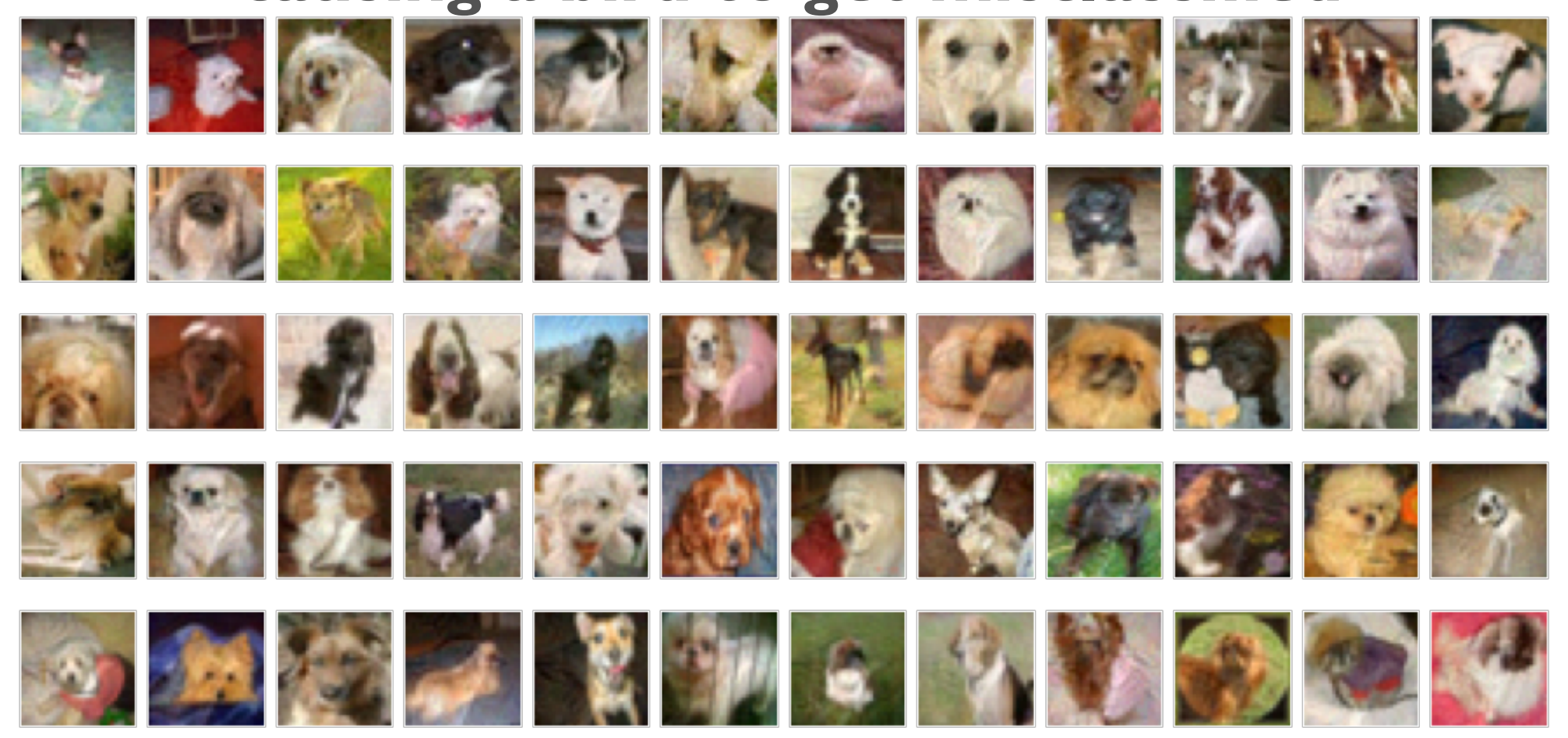
&

Using multiple Poisons

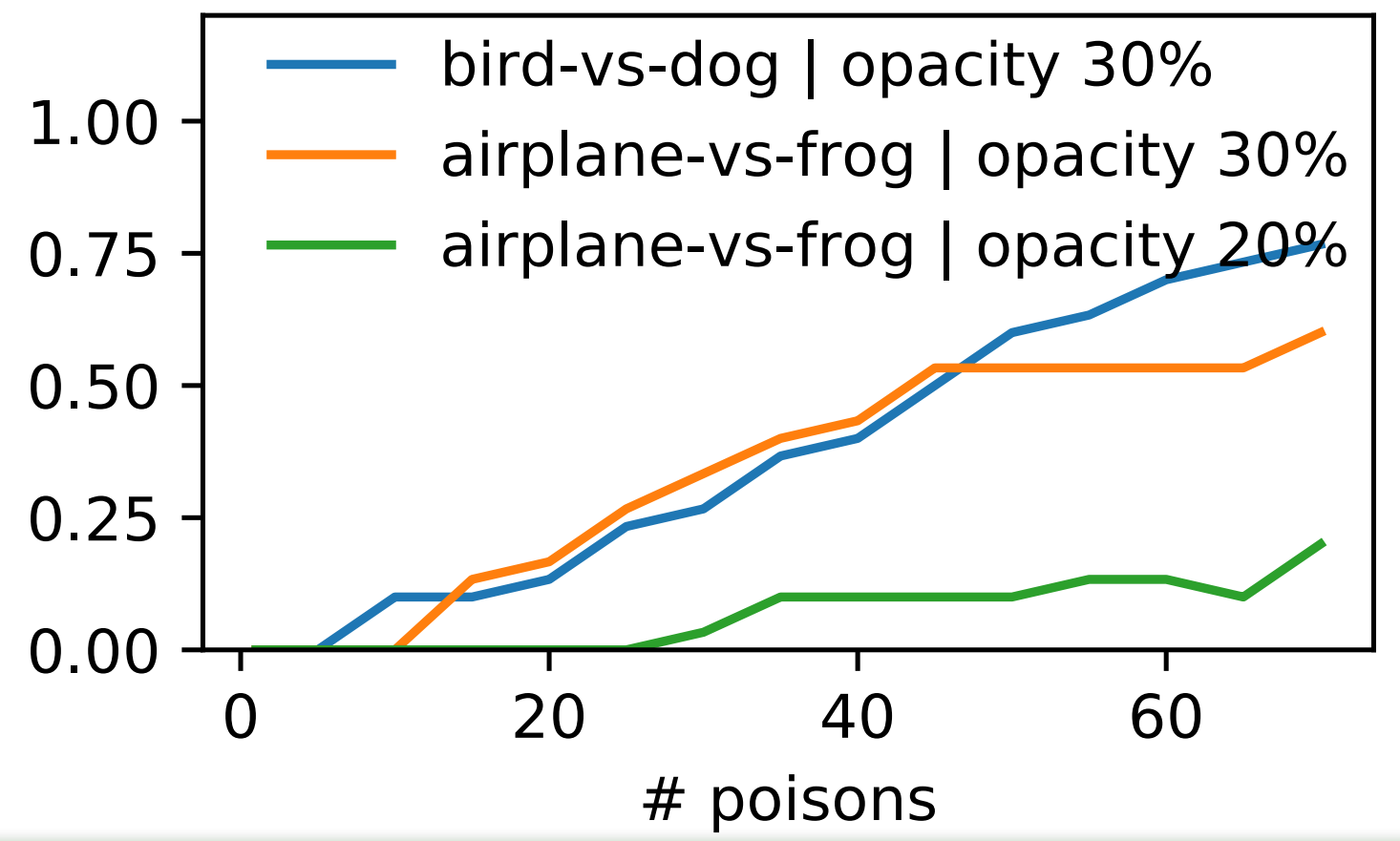
Successful poisoning in end-to-end: sucking out the target



Successful poisoning in end-to-end: 60 poison dogs! causing a bird to get misclassified



Success rate depends on # of poisons and opacity of target


bird-vs-dog | opacity 30%
airplane-vs-frog | opacity 30%
airplane-vs-frog | opacity 20%
success rate
1.00
0.75
0.50
0.25
0.00
0
20
40
60
# poisons