TRANSFERABLE CLEAN-LABEL POISONING ATTACKS ON DEEP NEURAL NETS

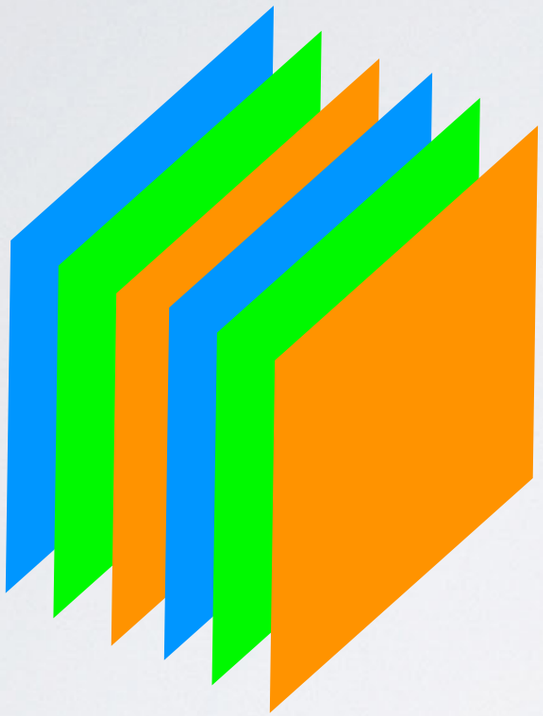
Chen Zhu*, W. Ronny Huang*^, Ali Shafahi,
Hengduo Li, Gavin Taylor, Christoph Studer,
Tom Goldstein

* equal contribution
^ presenter

Cornell University

WHAT IS POISONING?

Training data



Base

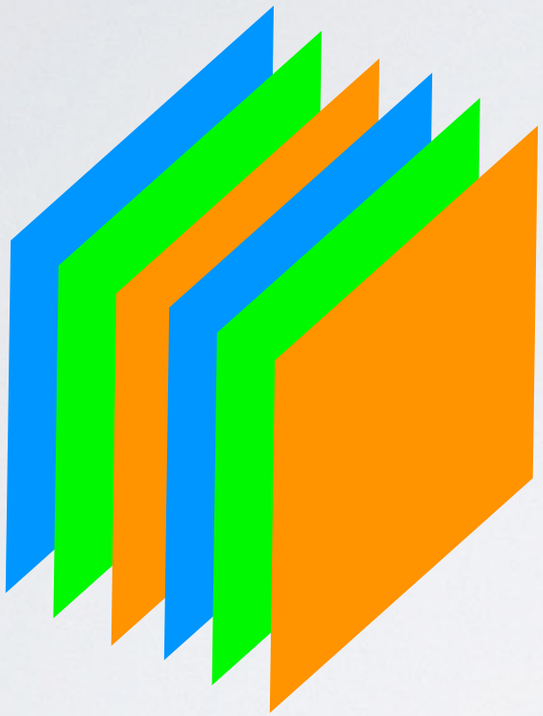
Testing example

Plane

Frog

WHAT IS POISONING?

Training data



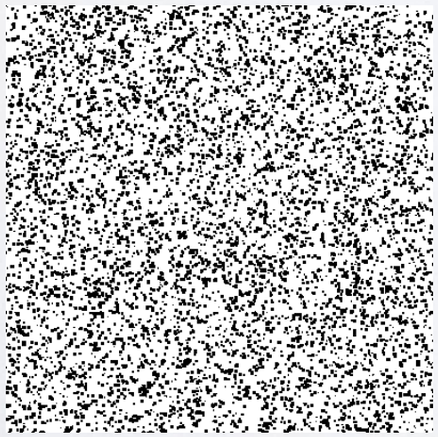
Testing example

Plane

Frog

Base

+

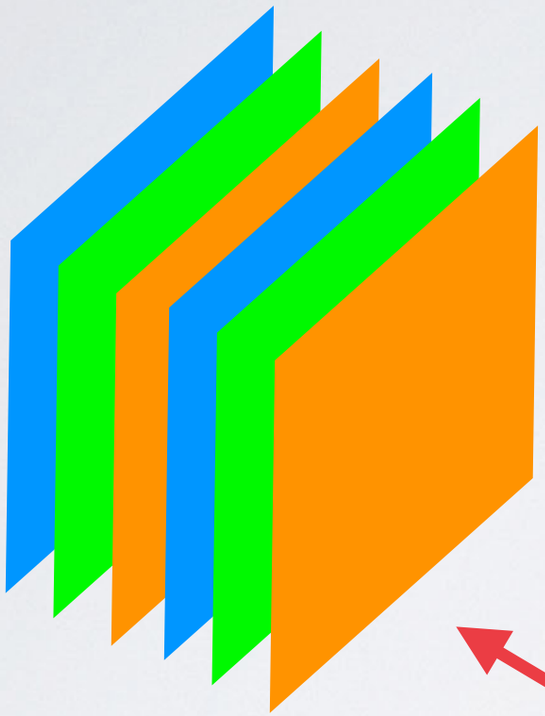


=

Poison!

WHAT IS POISONING?

Training data



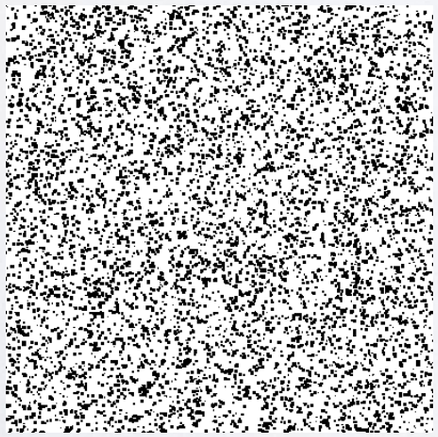
Testing example

Plane

Frog

Base

\+



=

Poison!

WHITE BOX CASE

Victim network is <u>known</u>

COLLISION ATTACK

Feature extractor

$$\arg\min_x \|f(x) - f(t)\|^2 + \|x - b\|^2$$

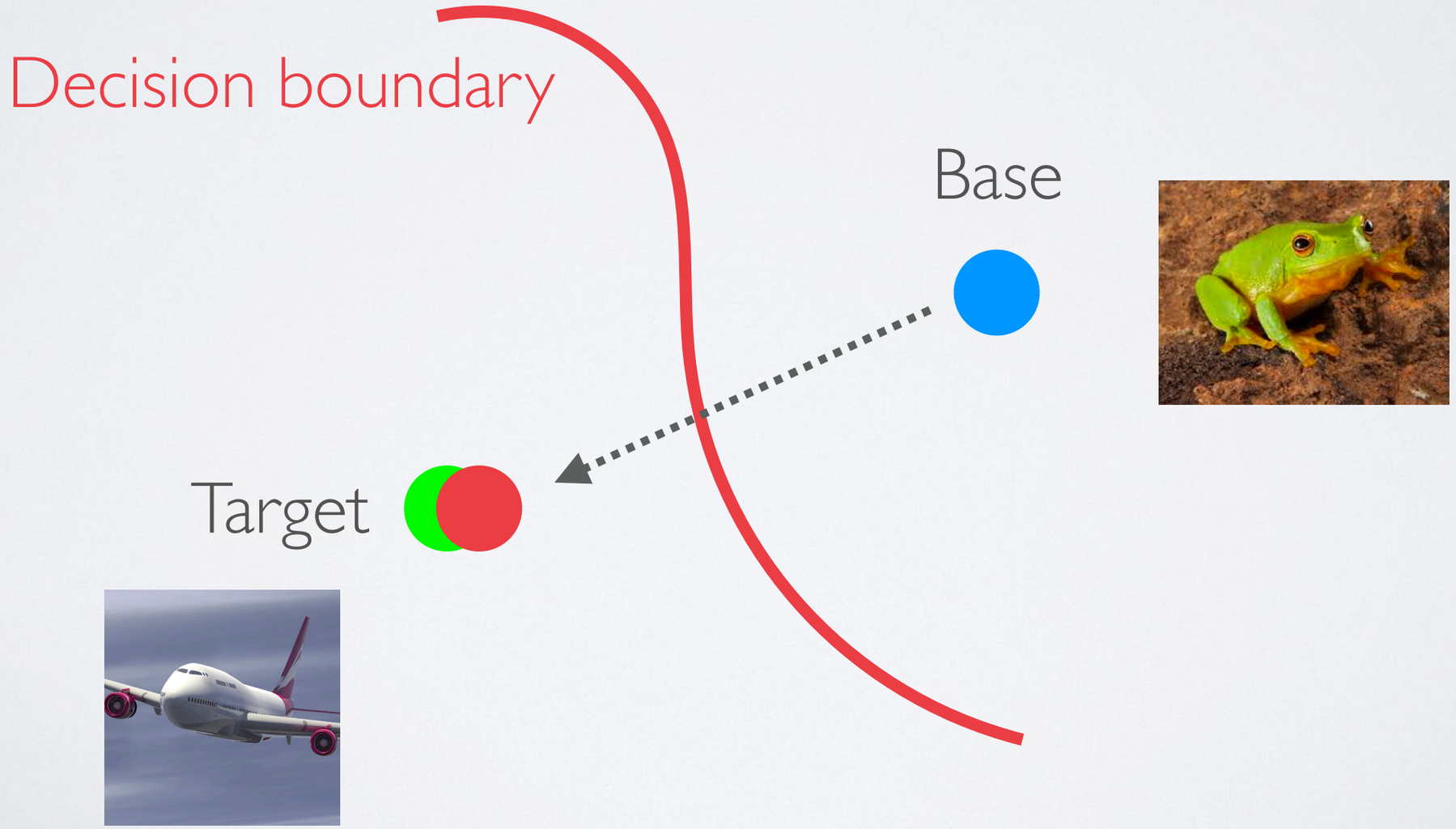
Decision boundary

Base

Target

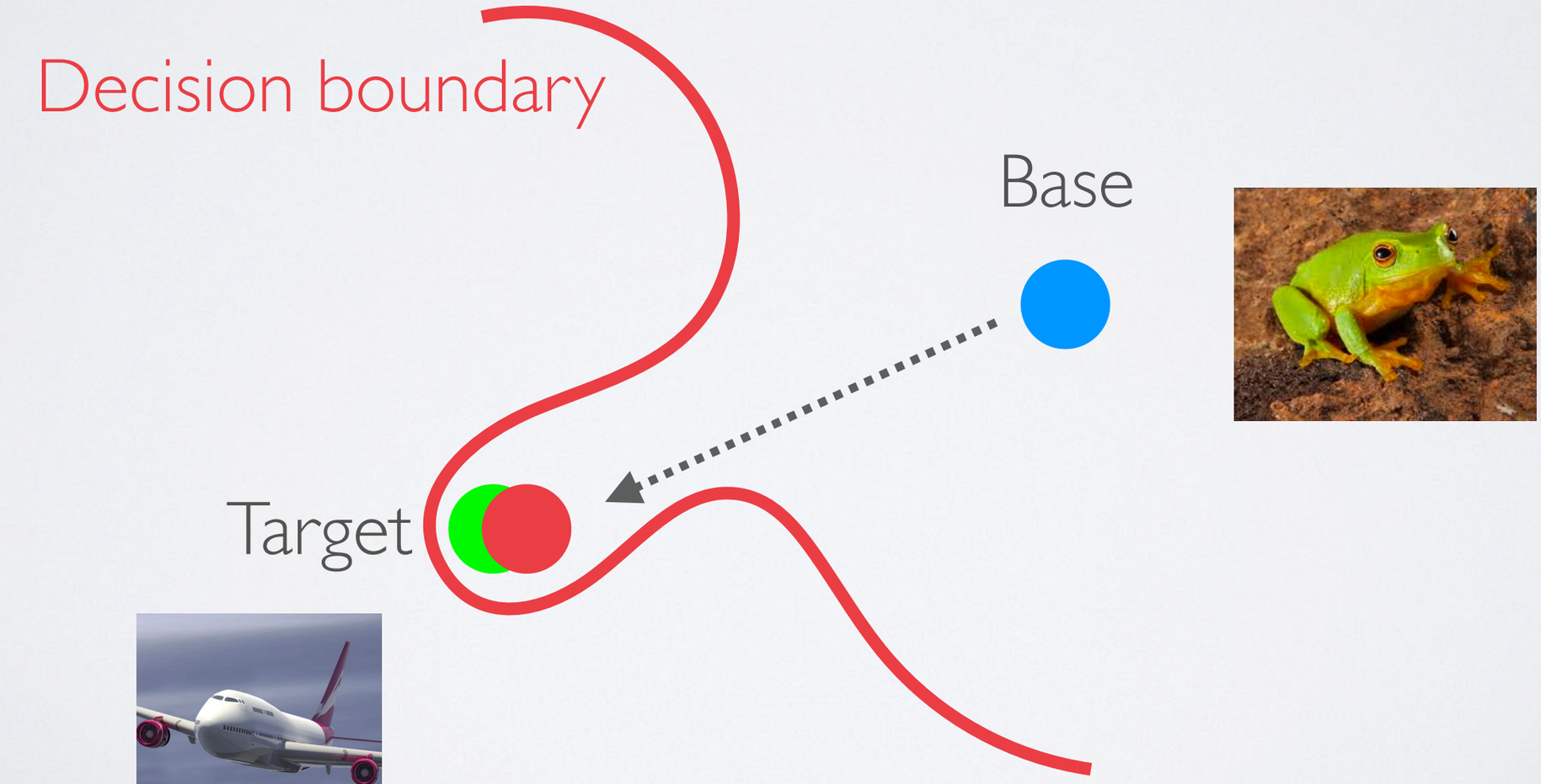
COLLISION ATTACK

$$\arg\min_x \|f(x) - f(t)\|^2 + \|x - b\|^2$$



COLLISION ATTACK

$$\arg\min_x \|f(x) - f(t)\|^2 + \|x - b\|^2$$



BLACK BOX CASE

Victim network is <u>unknown</u>

BLACK BOX ATTACK

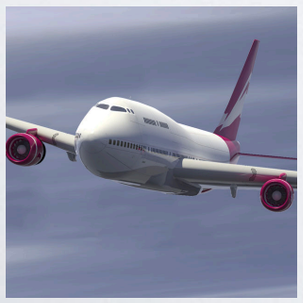
Guess the model

$$\arg\min_x \|f_{guess}(x) - f_{guess}(t)\|^2 + \|x - b\|^2$$

Decision boundary

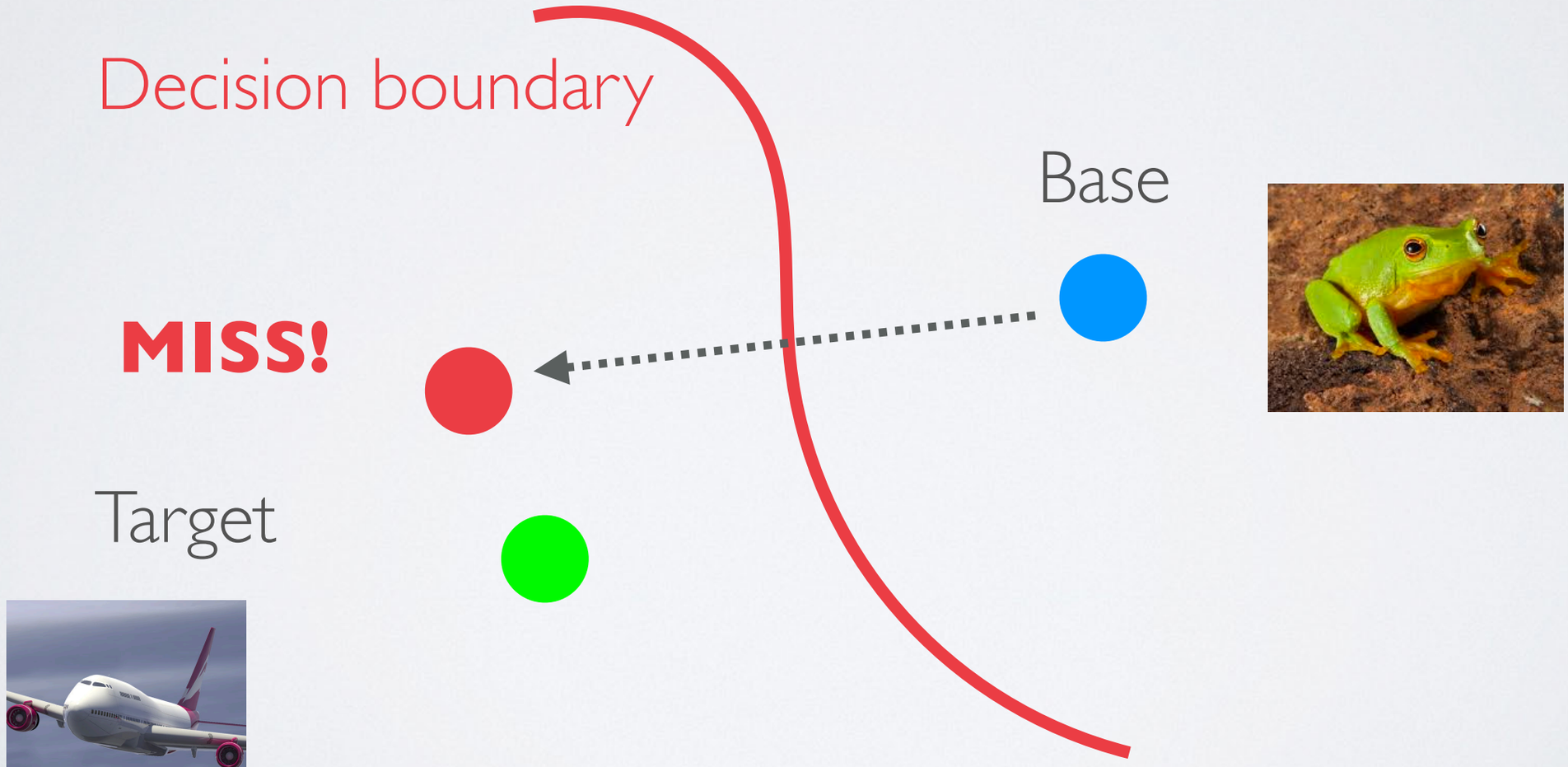
Base

Target



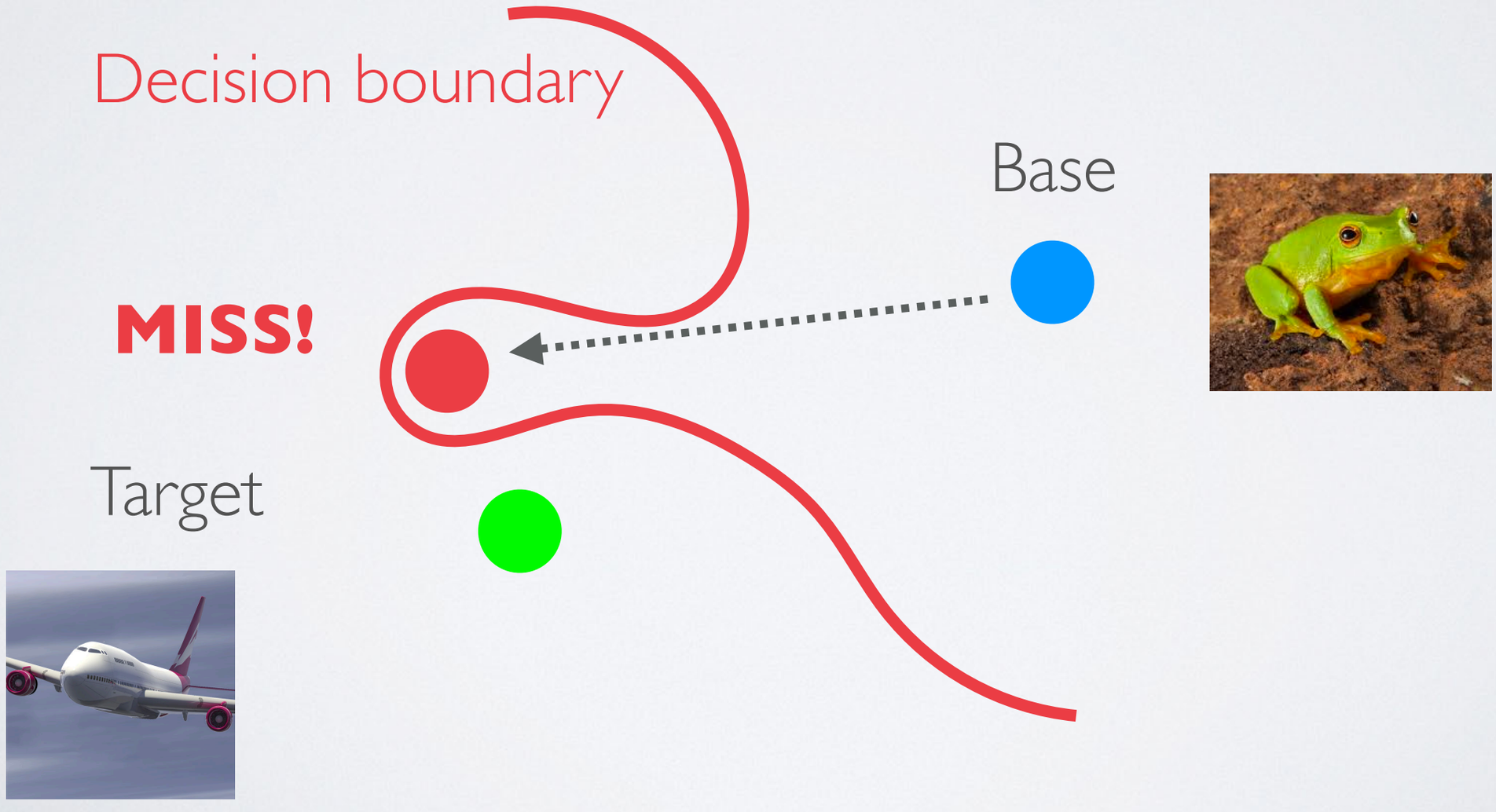
BLACK BOX ATTACK

$$\arg\min_x \|f_{guess}(x) - f_{guess}(t)\|^2 + \|x - b\|^2$$



BLACK BOX ATTACK

$$\arg\min_x \|f_{guess}(x) - f_{guess}(t)\|^2 + \|x - b\|^2$$

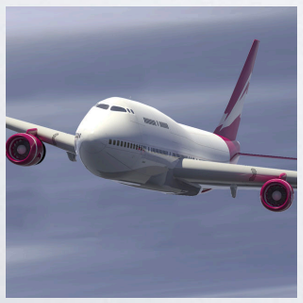


CONVEX POLYTOPE ATTACK

Decision boundary

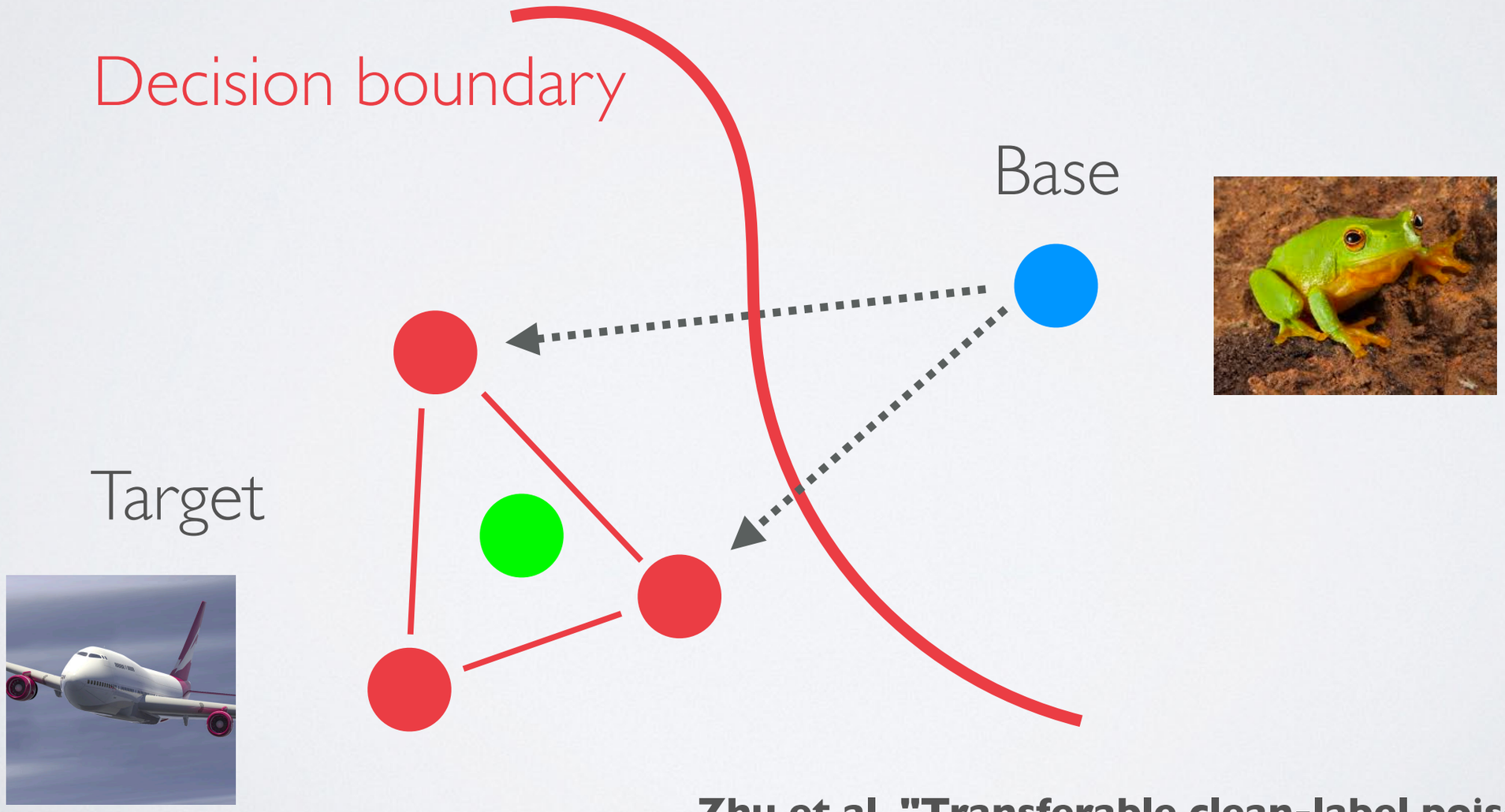
Base

Target



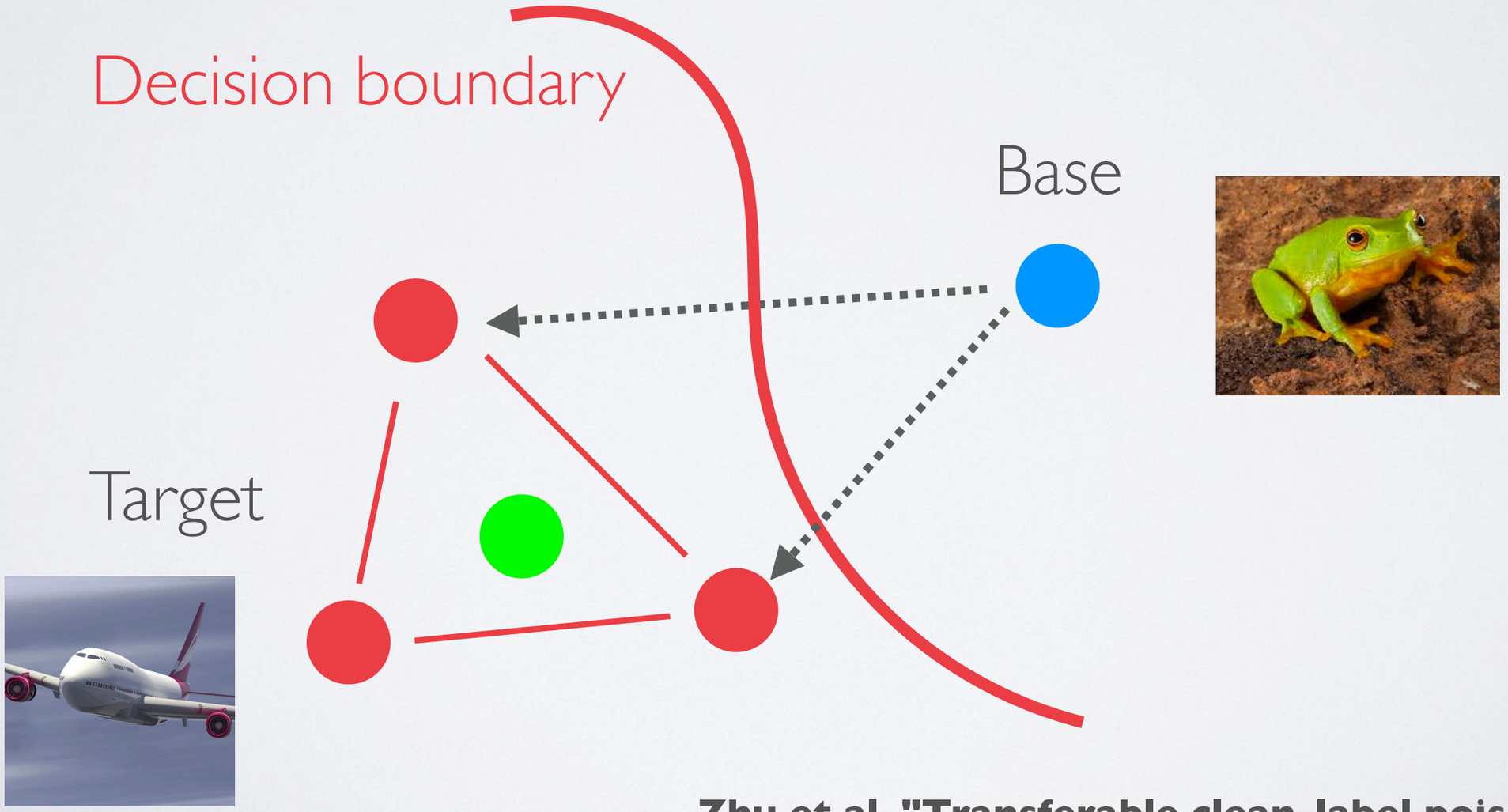
Zhu et al. "Transferable clean-label poisoning attacks”

CONVEX POLYTOPE ATTACK



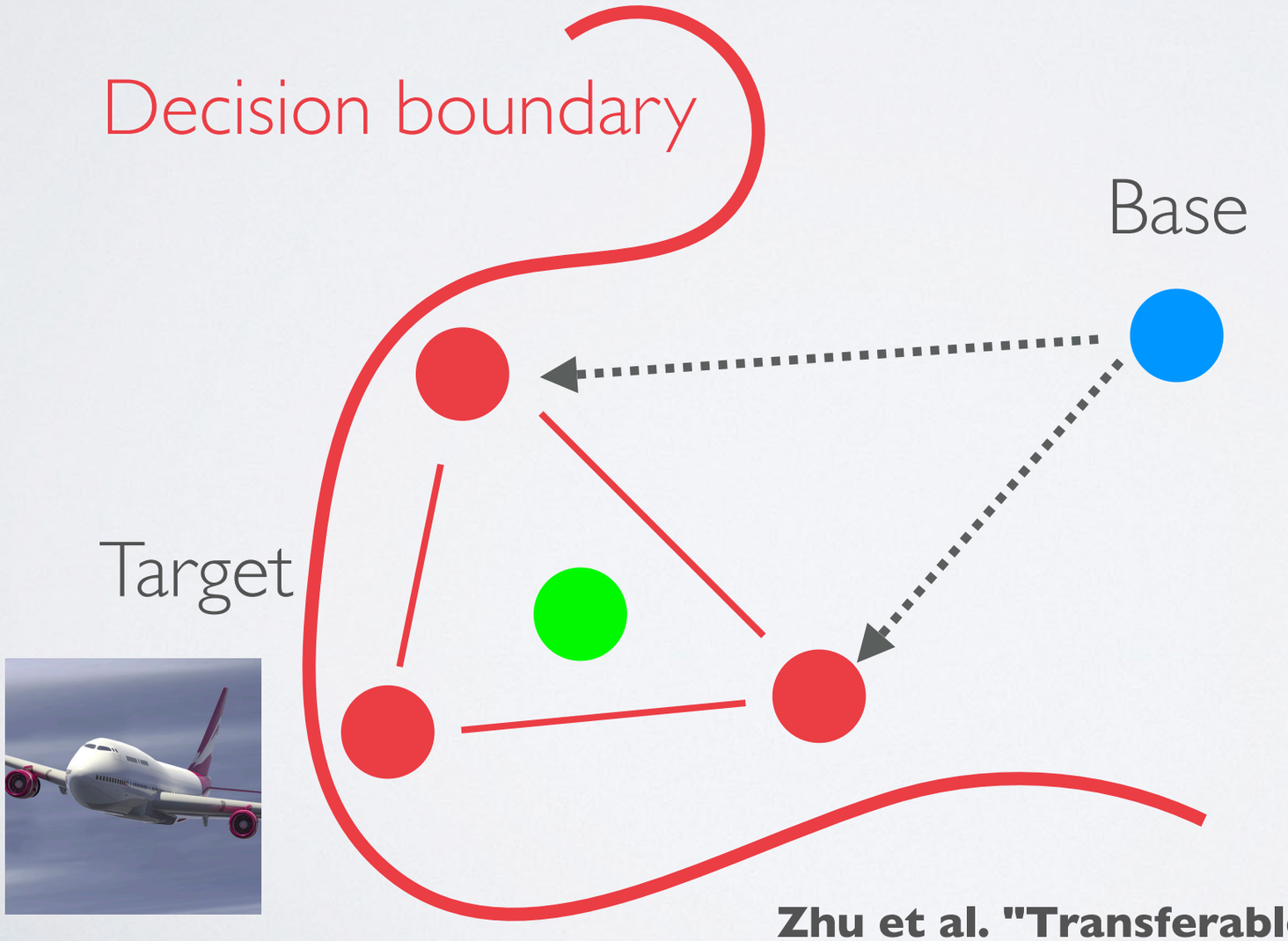
Zhu et al. "Transferable clean-label poisoning attacks”

CONVEX POLYTOPE ATTACK



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

CONVEX POLYTOPE ATTACK



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

POISON POLYTOPE

Target (fish)

Clean

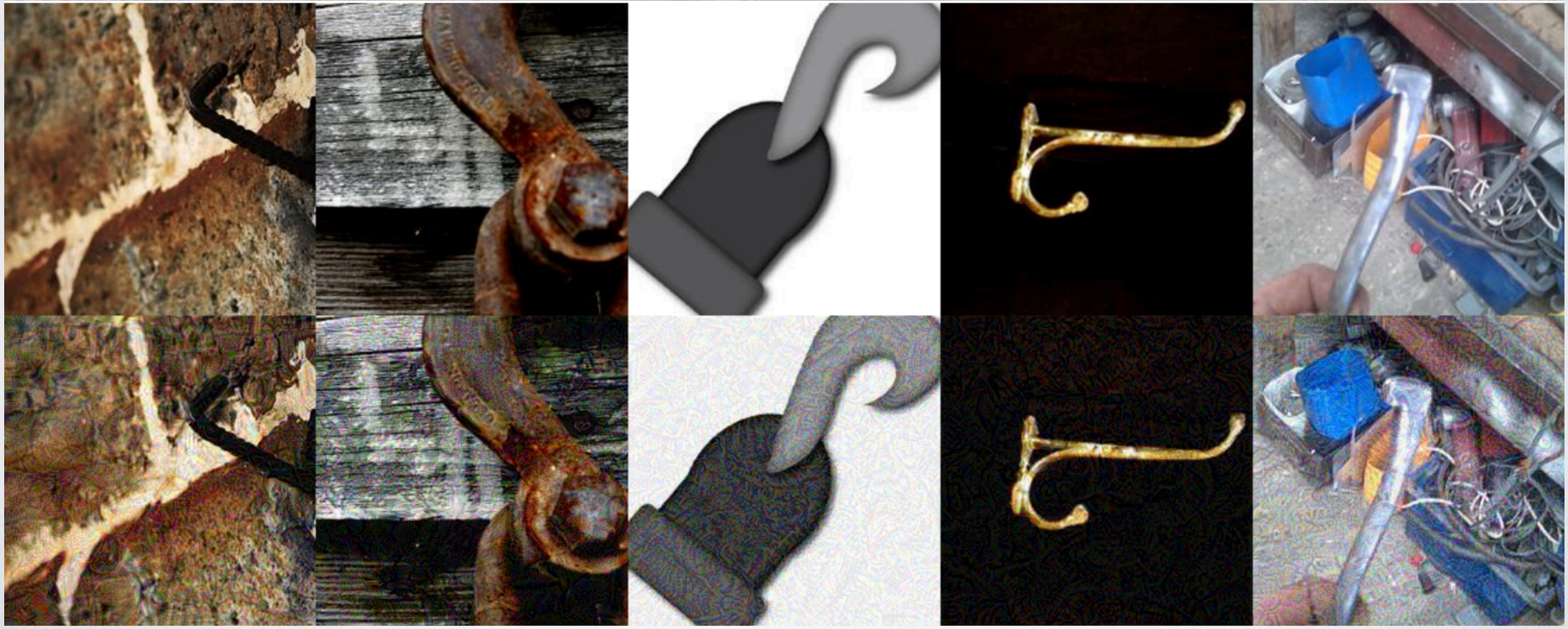
Poison

POISON POLYTOPE

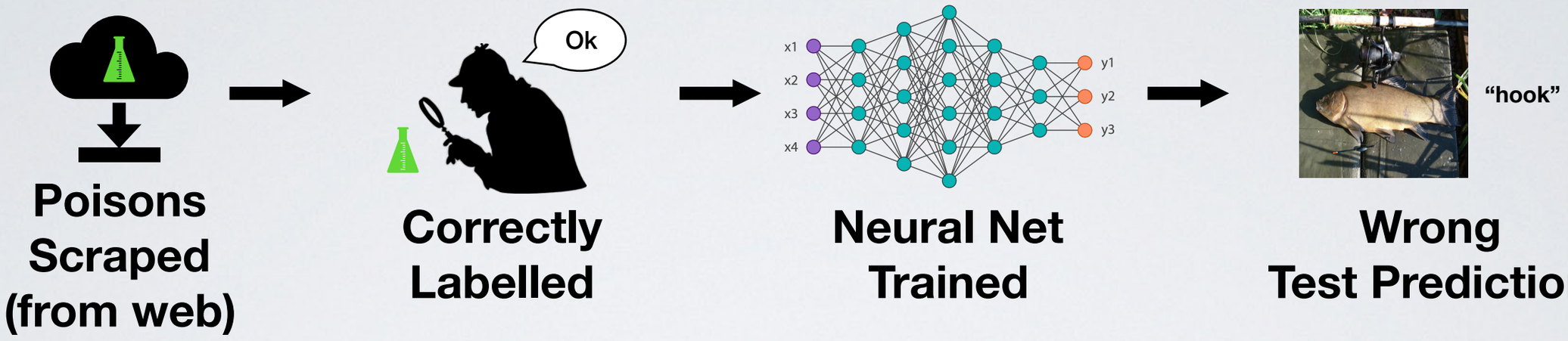
Target (fish)

Clean

Poison



COME SEE POSTER #68!



Attack success rate ~50% on unknown architectures

Works under many scenarios

- No training data overlap
- Transfer learning and end-to-end training

No drop in overall test accuracy

Link to paper & code