METAPOISON: LEARNING TO CRAFT POISON

W. Ronny Huang,* Jonas Geiping,*
Liam Fowl,^ Tom Goldstein

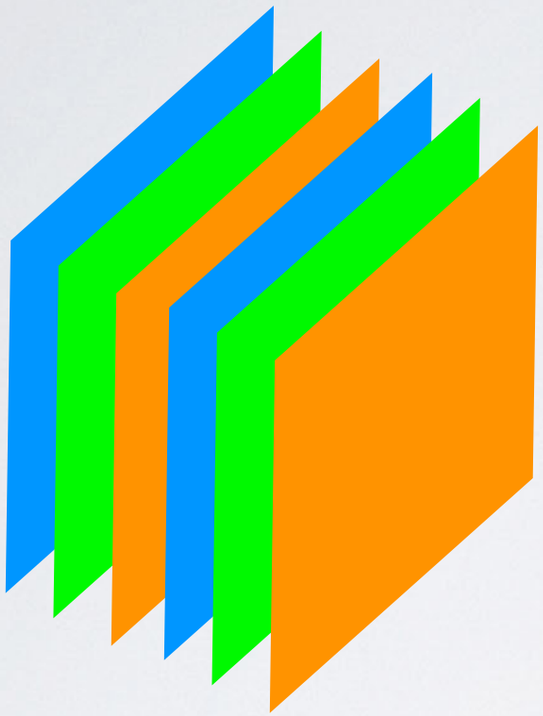
*Equal Contribution
^Speaker

University of Maryland

NeurIPS MetaLearn 2019

DATA POISONING

Training data



Testing example

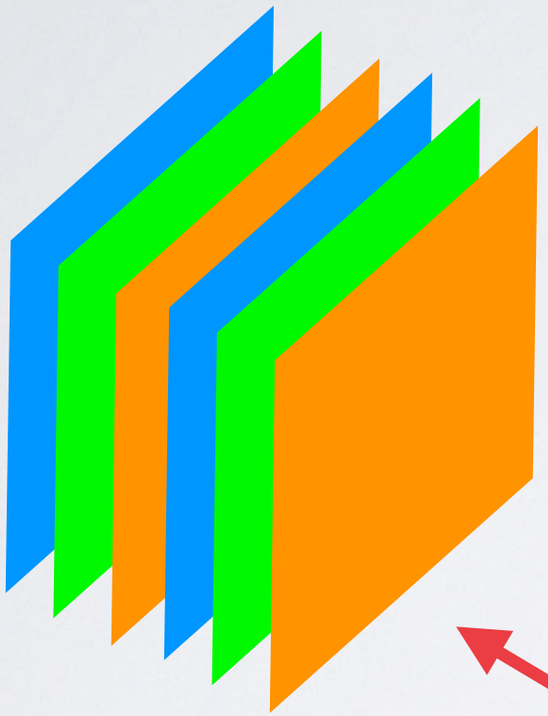
Plane

Frog

Base

DATA POISONING

Training data



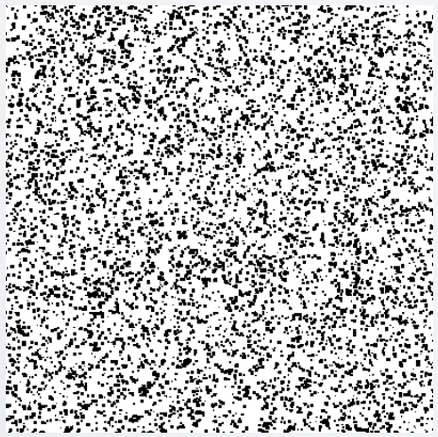
Testing example

Plane

Frog

Base

+

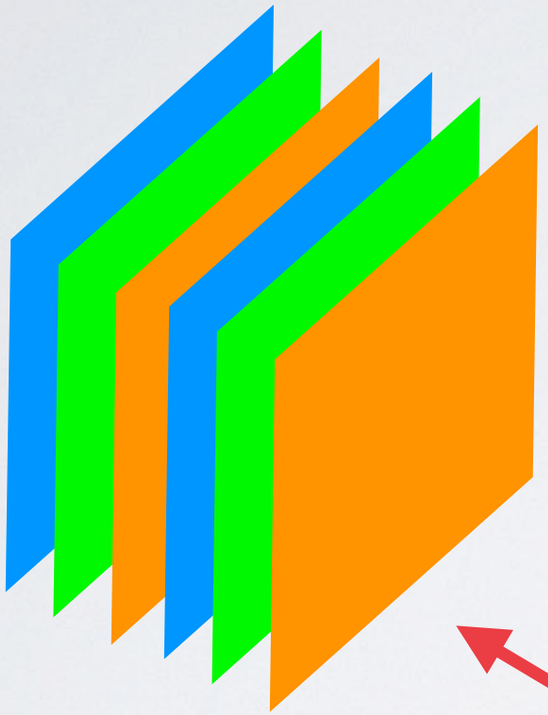


=

Poison!

DATA POISONING

Training data



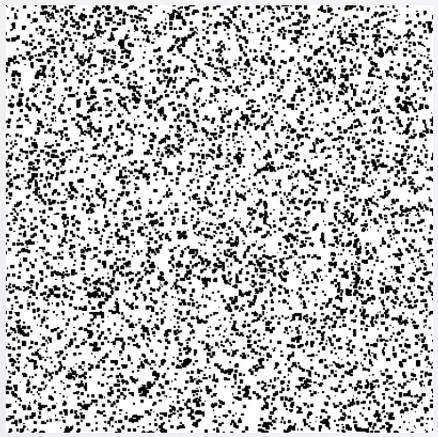
Testing example

Plane

Frog

Base

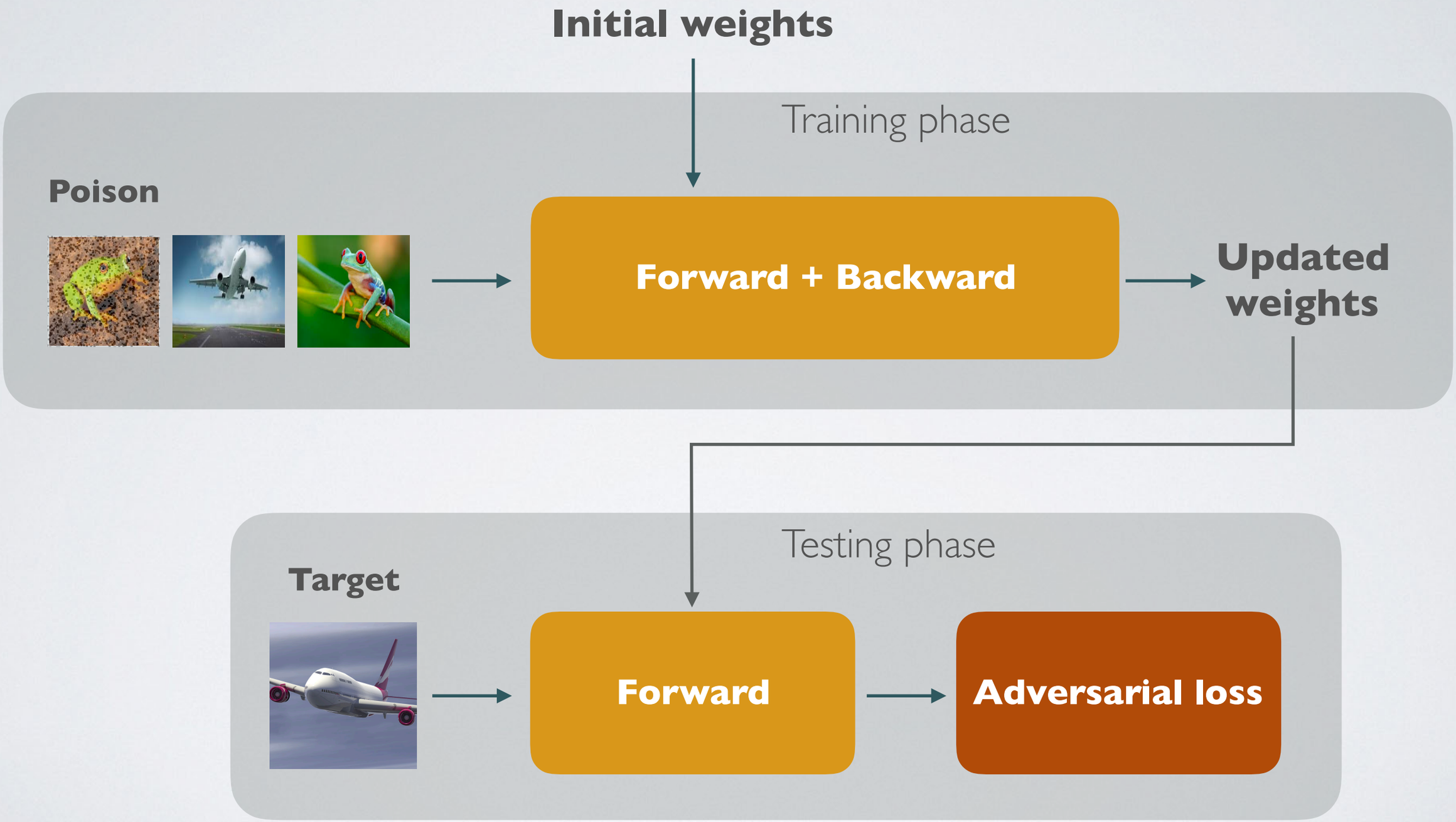
+



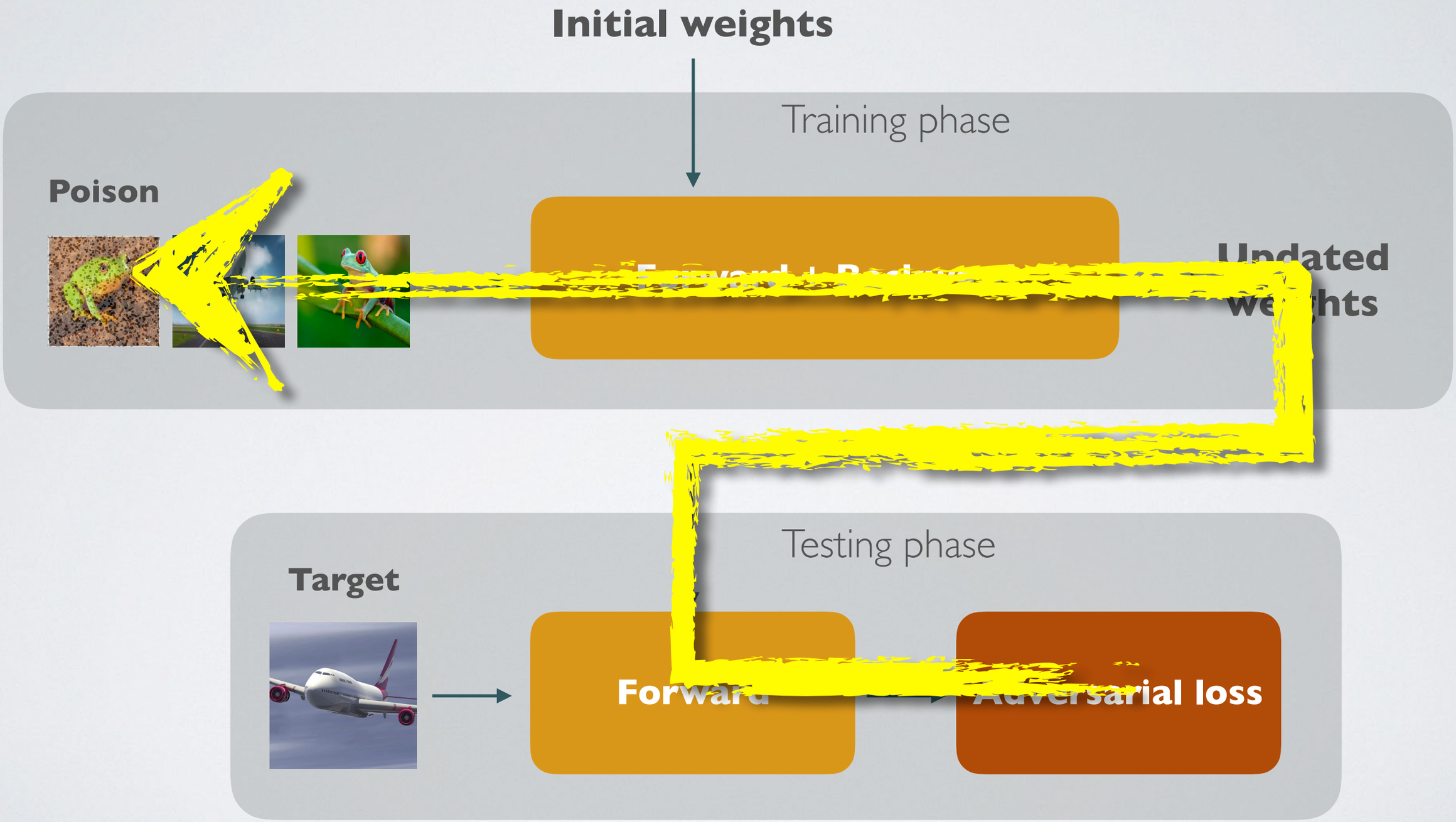
=

Poison!

LEARNING TO CRAFT

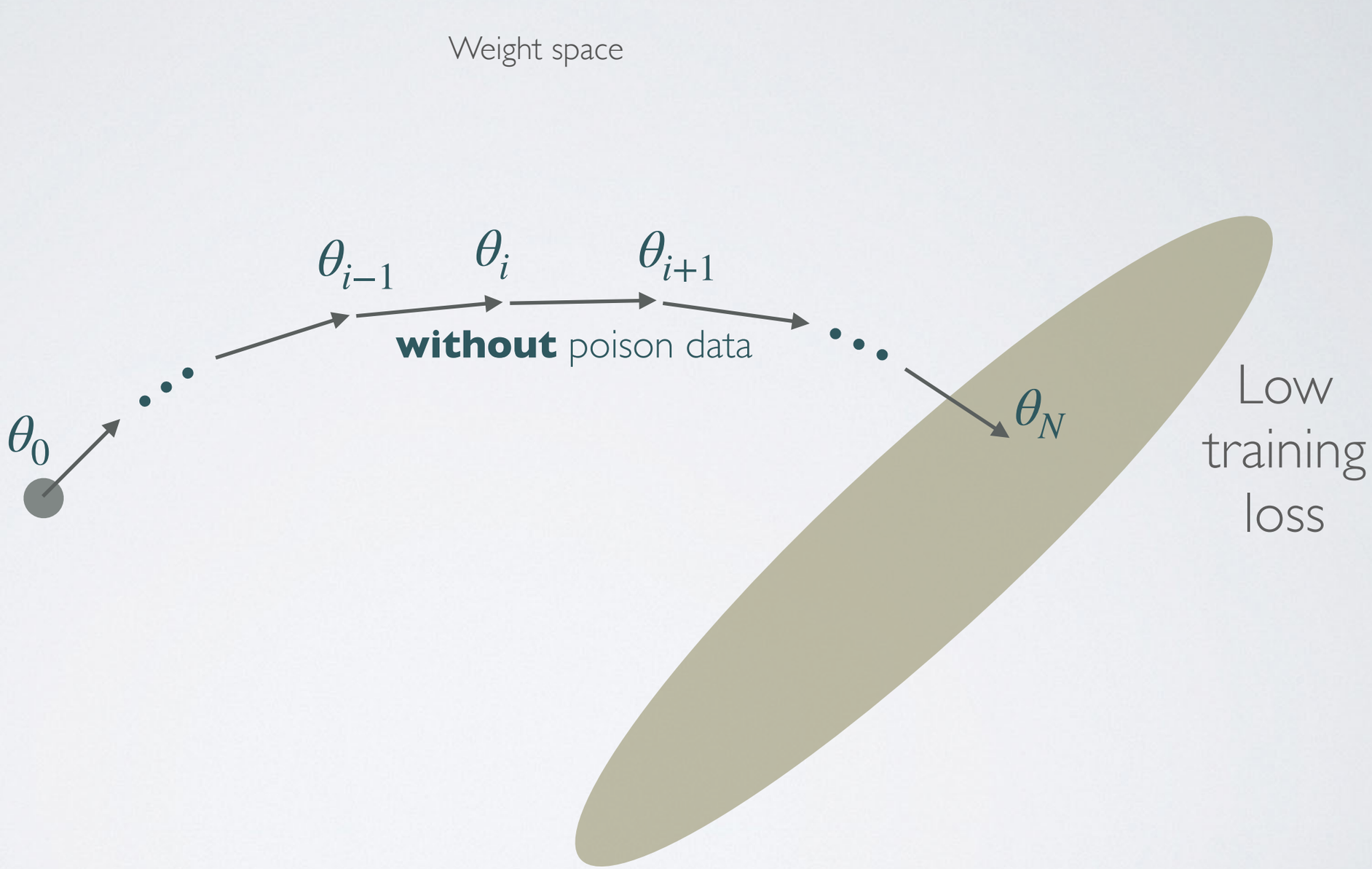


LEARNING TO CRAFT



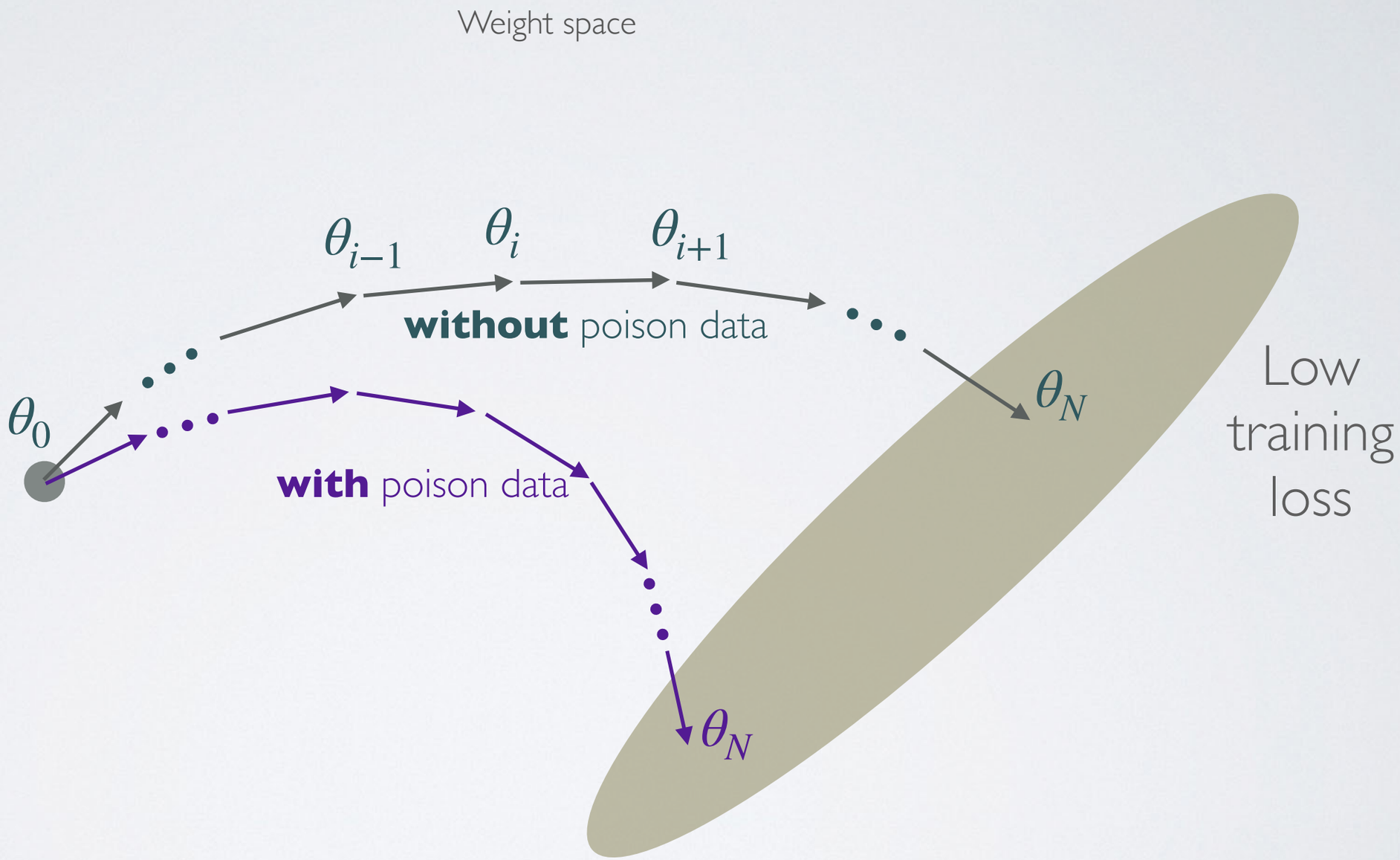
Backprop to the poison!

POISONED TRAINING DYNAMICS



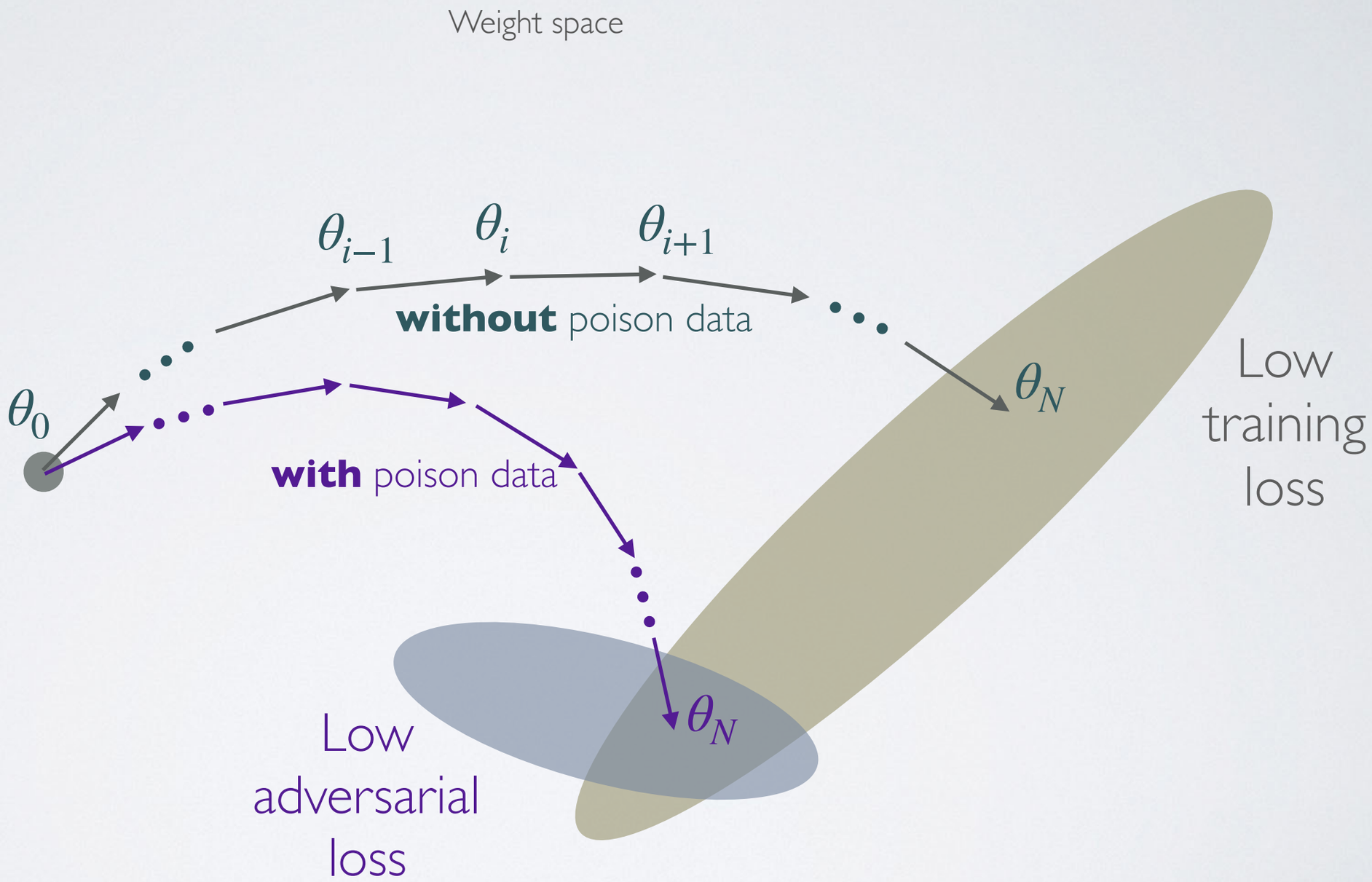
NeurIPS Metalearn 19 (spotlight) **Huang***, Geiping*, Fowl, Taylor, Goldstein, “MetaPoison: Learning to...”

POISONED TRAINING DYNAMICS

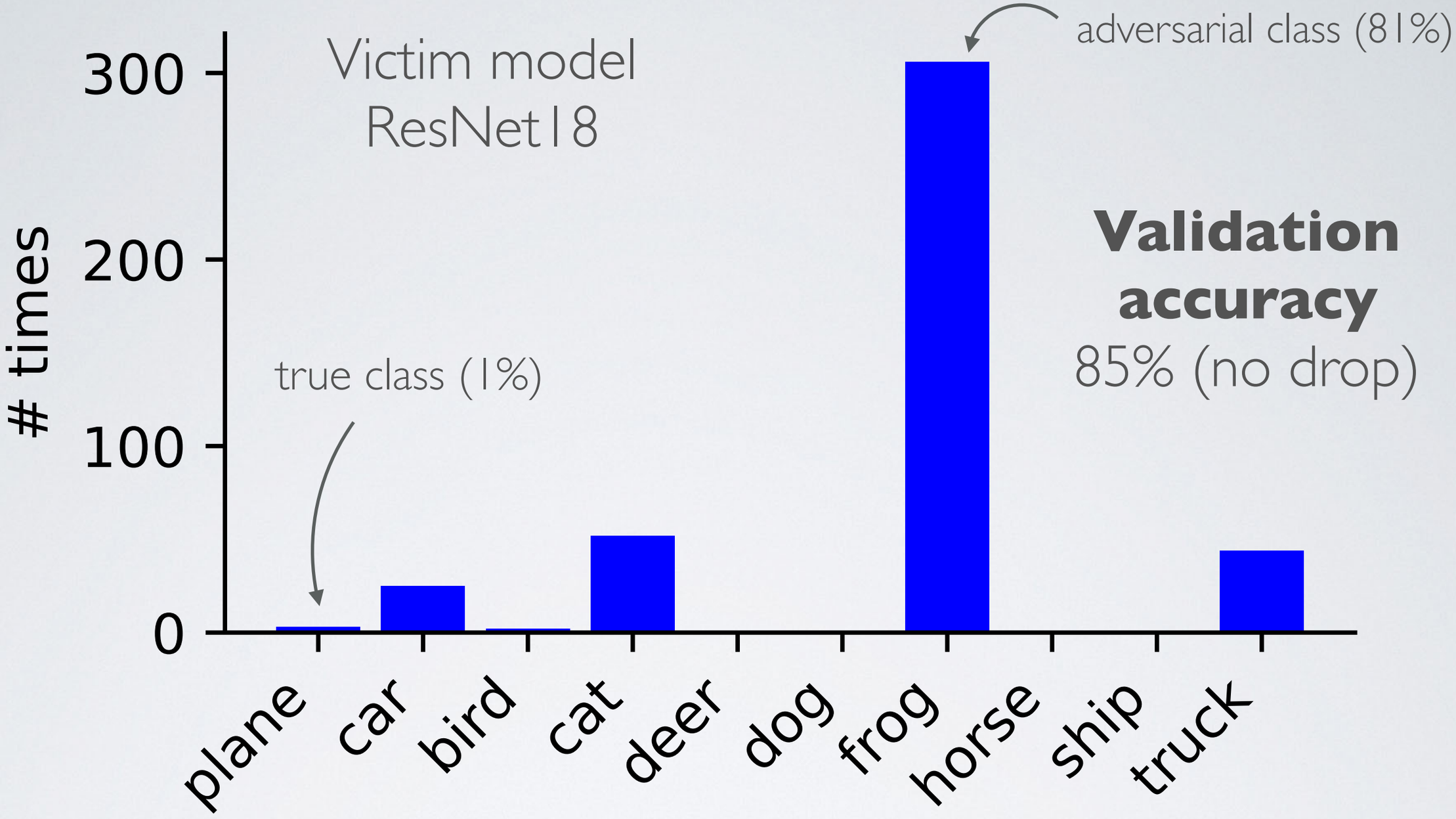


NeurIPS Metalearn 19 (spotlight) **Huang***, Geiping*, Fowl, Taylor, Goldstein, "MetaPoison: Learning to..."

POISONED TRAINING DYNAMICS



NeurIPS Metalearn 19 (spotlight) **Huang***, Geiping*, Fowl, Taylor, Goldstein, "MetaPoison: Learning to..."



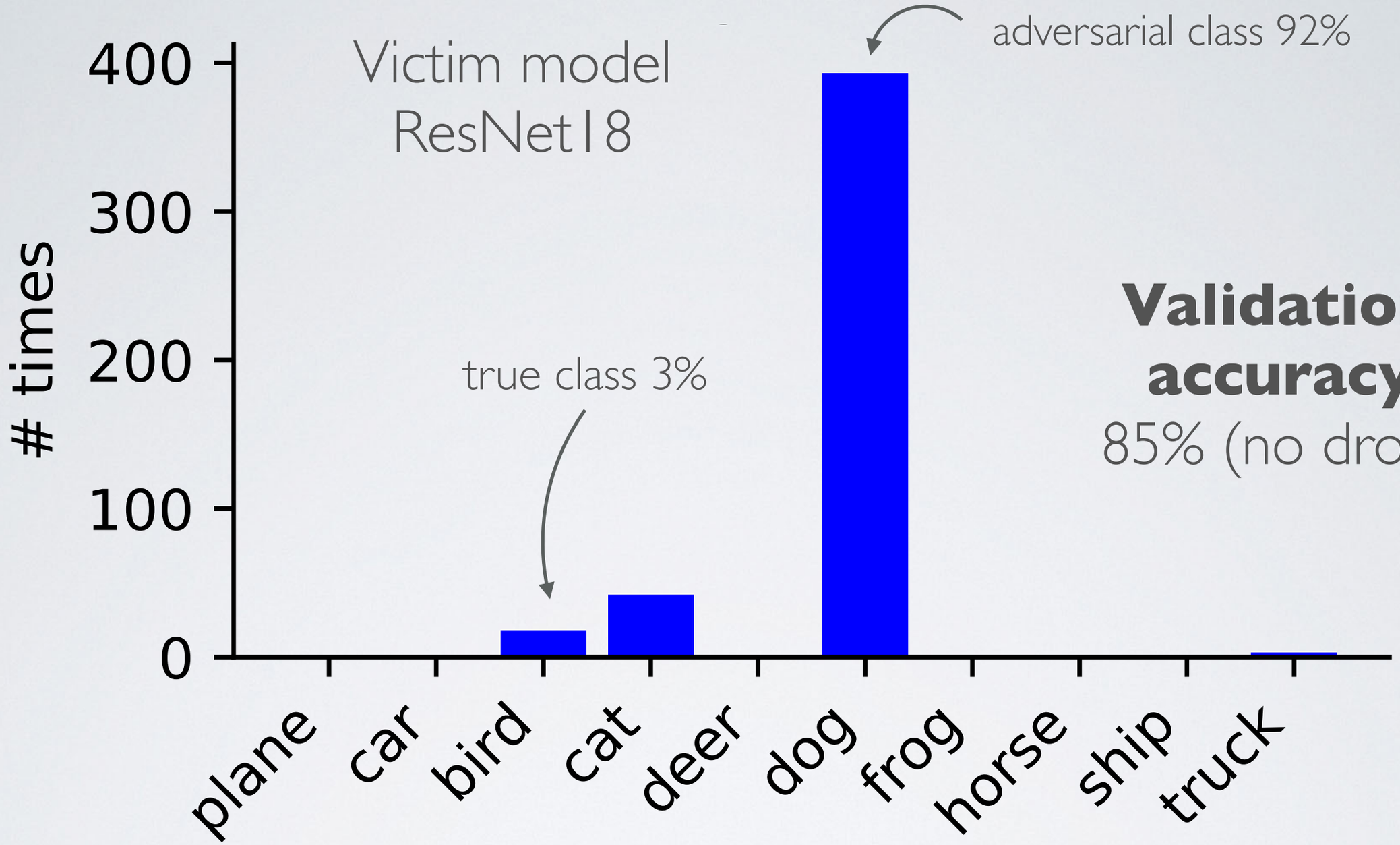
Validation accuracy
85% (no drop)

5000 poisons (10%)

cause

Target

classified as



Validation accuracy
85% (no drop)

5000 poisons (10%)

cause

Target

classified as