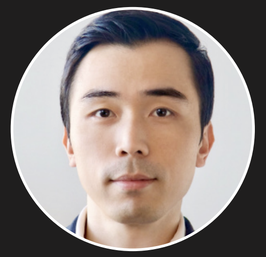
Understanding, Fortifying and Democratizing AI Security

Nilaksh Das
nilakshdas.com

Georgia Tech

PhD Thesis Dissertation — April 13, 2022

Thesis Committee



PhD Advisor

Dr. Duen Horng Chau

Georgia Tech

Data Science & Adversarial ML

Dr. Wenke Lee

Georgia Tech

Network Security & Privacy

Dr. Xu Chu

Georgia Tech

Data Management Systems & ML

Dr. Srijan Kumar

Georgia Tech

Robust & Adversarial ML

Dr. Ponnurangam Kumaraguru ("PK")

IIIT-Hyderabad

Privacy & Security in Online Social Media

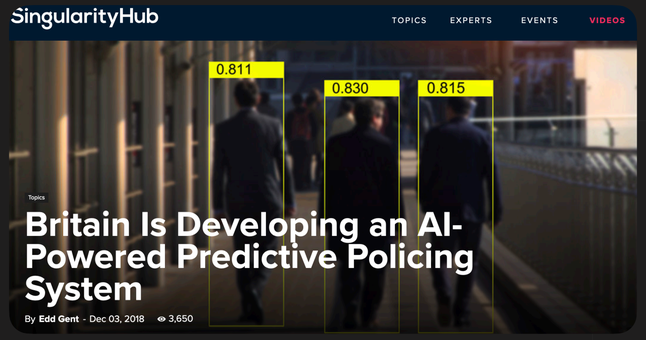
Dr. Oliver Brdiczka

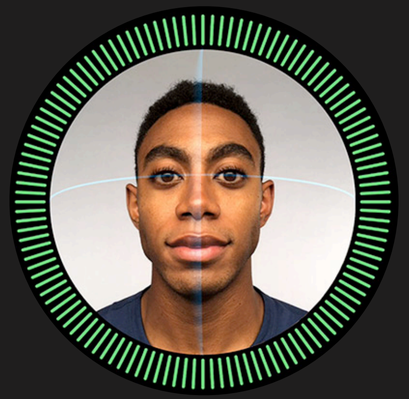
Adobe

AI & Human-Computer Interaction

AI is becoming ubiquitous

- Smart voice assistants
- Biometric authentication
- Social media newsfeeds
- Autonomous vehicles
- High-frequency trading
- Healthcare
- Public safety and policing
- Massive data collection





AI Security is becoming critical

Smart voice assistants	Malicious commands
Biometric authentication	Unauthorized access
Social media newsfeeds	Misinformation and emotion hacking
Autonomous vehicles	Cause fatal accidents
High-frequency trading	Manipulate markets
Healthcare	Life-threatening
Public safety and policing	Social havoc
Massive data collection	Steal training data

AI Security is becoming critical

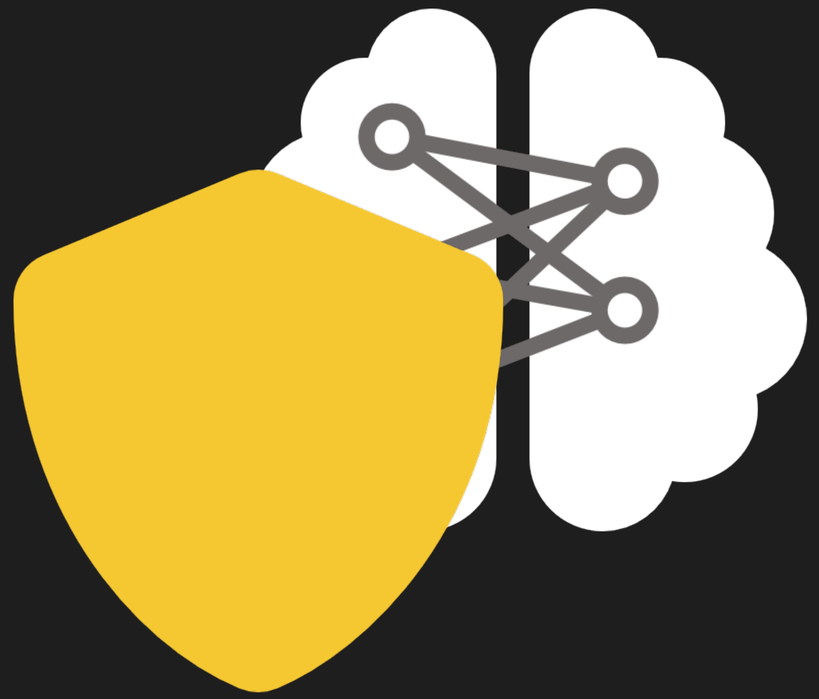
Smart voice assistants	Malicious commands
Biometric authentication	Unauthorized access
Social media newsfeeds	Misinformation and emotion hacking
Autonomous vehicles	Cause fatal accidents
High-frequency trading	Manipulate markets
Healthcare	Life-threatening
Public safety and policing	Social havoc
Massive data collection	Steal training data

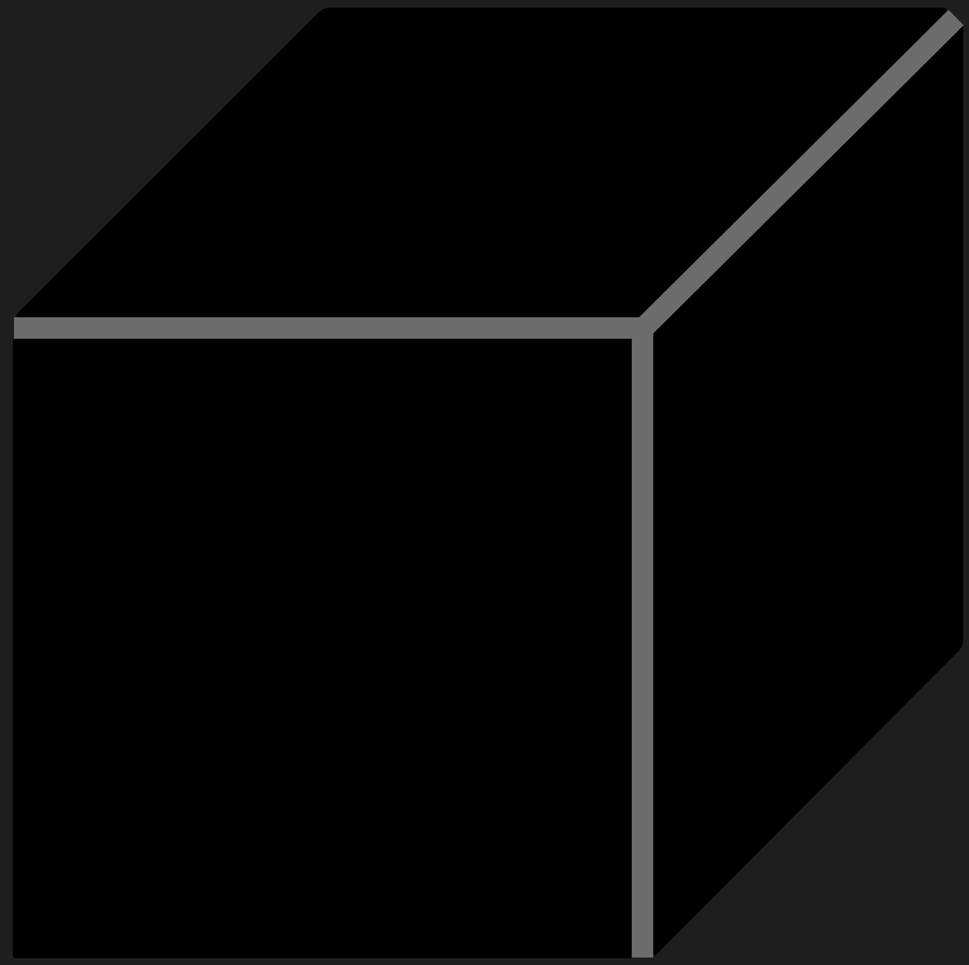
AI Security is becoming critical

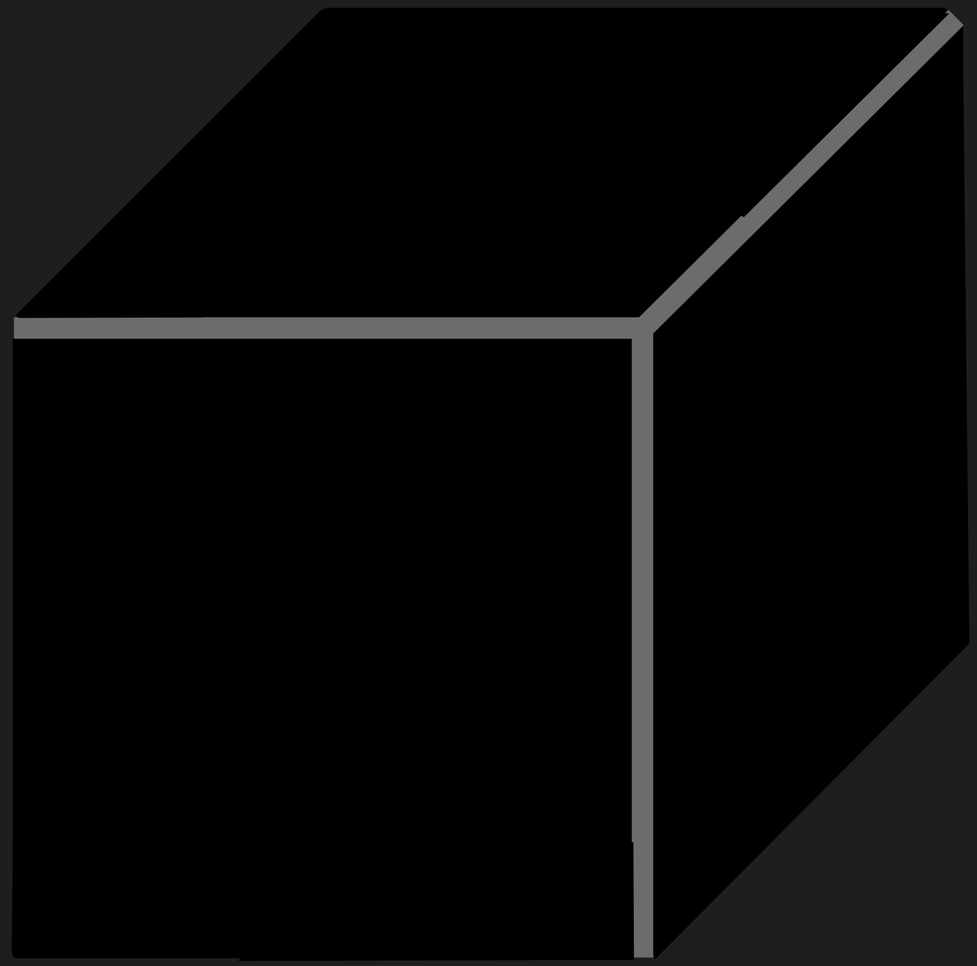
Smart voice assistants	Malicious commands
Biometric authentication	Unauthorized access
Social media newsfeeds	Misinformation and emotion hacking
Autonomous vehicles	Cause fatal accidents
High-frequency trading	Manipulate markets
Healthcare	Life-threatening
Public safety and policing	Social havoc
Massive data collection	Steal training data

AI Security is becoming critical

Smart voice assistants	Malicious commands
Biometric authentication	Unauthorized access
Social media newsfeeds	Misinformation and emotion hacking
Autonomous vehicles	Cause fatal accidents
High-frequency trading	Manipulate markets
Healthcare	Life-threatening
Public safety and policing	Social havoc
Massive data collection	Steal training data







Research Mission

Enable everyone to understand and fortify AI security

Research Mission

Enable everyone to understand and fortify AI security

Research Challenge I

How to UNDERSTAND AI vulnerabilities?

How can we intuitively interpret the core reason for AI vulnerabilities?

Research Challenge II

How to FORTIFY AI security?

How can we effectively mitigate the harm from adversarial examples?

Research Challenge III

How to ENABLE everyone to test AI attacks & defenses?

How can we make AI security techniques more accessible?

UNDERSTAND **FORTIFY** **ENABLE**

"Fortify", "Understand" and "Enable" icons are created using Font Awesome (fontawesome.com/license/free)

Research Thrust I

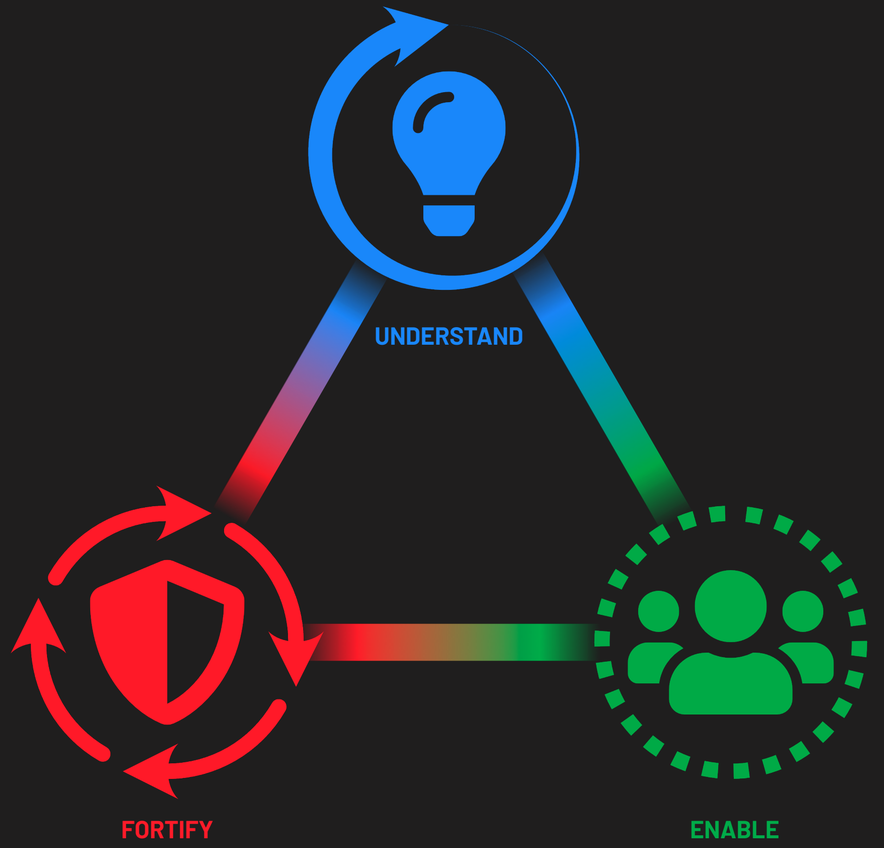
Exposing AI Vulnerabilities through Visualization & Interpretable Representations.

Research Thrust II

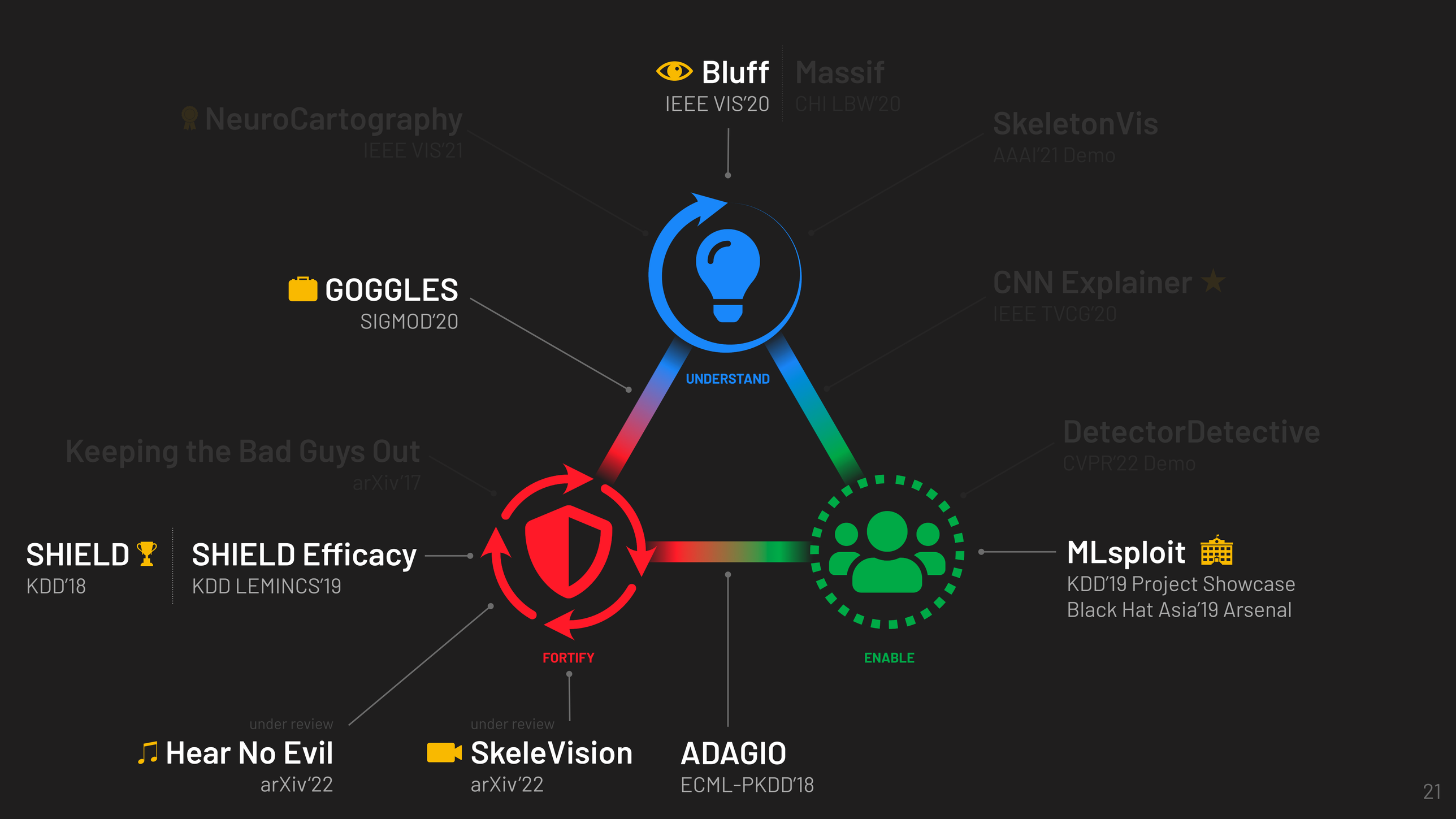
Mitigating Adversarial Examples Across Modalities & Tasks.

Research Thrust III

Democratizing AI Security Research & Pedagogy with Scalable Interactive Experimentation.


UNDERSTAND
FORTIFY
ENABLE


Bluff
IEEE VIS'20
Massif
CHI LBW'20
NeuroCartography
IEEE VIS'21
SkeletonVis
AAAI'21 Demo
GOGGLES
SIGMOD'20
CNN Explainer ★
IEEE TVCG'20
UNDERSTAND
Keeping the Bad Guys Out
arXiv'17
DetectorDetective
CVPR'22 Demo
SHIELD
KDD'18
SHIELD Efficacy
KDD LEMINCS'19
MLsploit
KDD'19 Project Showcase
Black Hat Asia'19 Arsenal
FORTIFY
ENABLE
under review
Hear No Evil
arXiv'22
under review
SkeleVision
arXiv'22
ADAGIO
ECML-PKDD'18


Bluff
IEEE VIS'20
Massif
CHI LBW'20
NeuroCartography
IEEE VIS'21
SkeletonVis
AAAI'21 Demo
GOGGLES
SIGMOD'20
CNN Explainer
IEEE TVCG'20
UNDERSTAND
Keeping the Bad Guys Out
arXiv'17
DetectorDetective
CVPR'22 Demo
SHIELD
KDD'18
SHIELD Efficacy
KDD LEMINCS'19
MLsploit
KDD'19 Project Showcase
Black Hat Asia'19 Arsenal
FORTIFY
ENABLE
under review
Hear No Evil
arXiv'22
under review
SkeleVision
arXiv'22
ADAGIO
ECML-PKDD'18

Thesis Statement

Overcome the threat of **adversarial attacks** by ***enabling*** everyone to:

1. **understand** AI vulnerabilities through **intuitive interpretation**
2. **fortify** AI applications with **robust mitigation** of adversarial examples
3. **easily test** AI security techniques with **interactive experimentation**

Part I

Understand
AI Vulnerabilities

GOGGLES SIGMOD 2020

Bluff IEEE VIS 2020

Part II

Fortify
AI Security

SHIELD KDD 2018

SkeleVision arXiv 2022 (under review)

Hear No Evil arXiv 2022 (under review)

Part III

Enable
Use of AI Security

ADAGIO ECML-PKDD 2018

MLsploit KDD Showcase 2019

Part I

Understand
AI Vulnerabilities

GOGGLES SIGMOD 2020

Bluff IEEE VIS 2020

Part II

Fortify
AI Security

SHIELD KDD 2018

SkeleVision arXiv 2022 (under review)

Hear No Evil arXiv 2022 (under review)

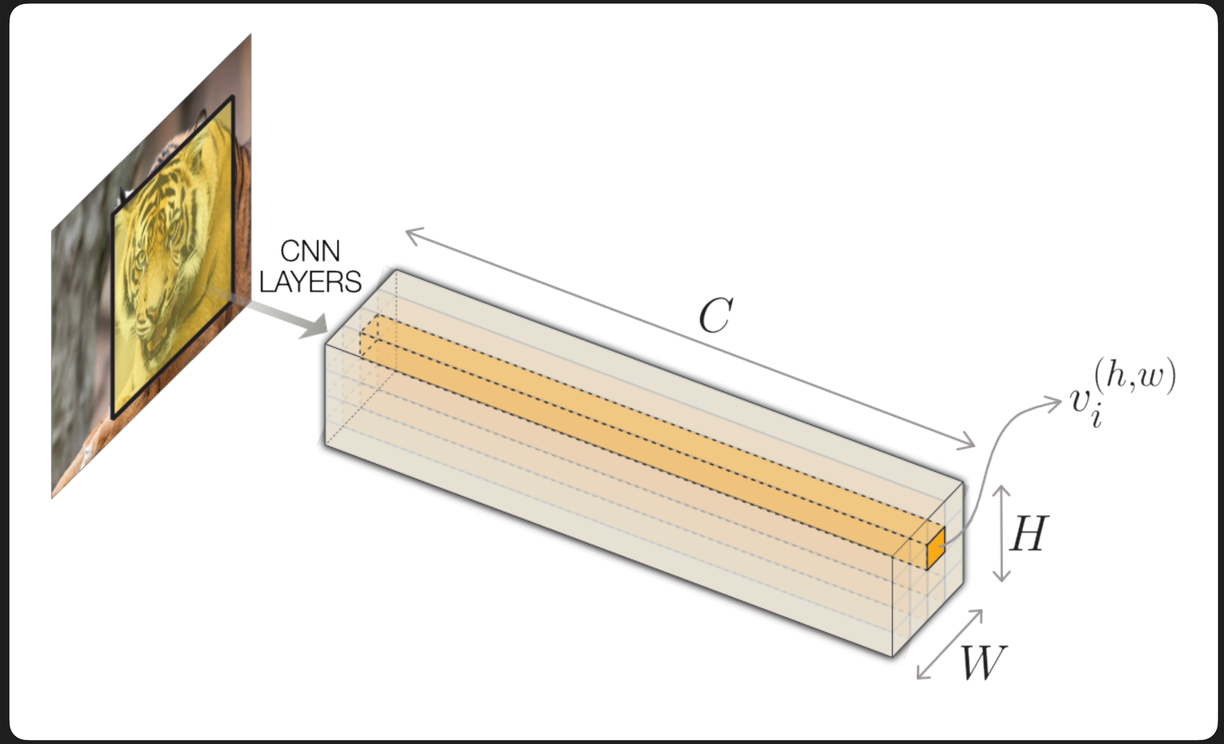
Part III

Enable
Use of AI Security

ADAGIO ECML-PKDD 2018

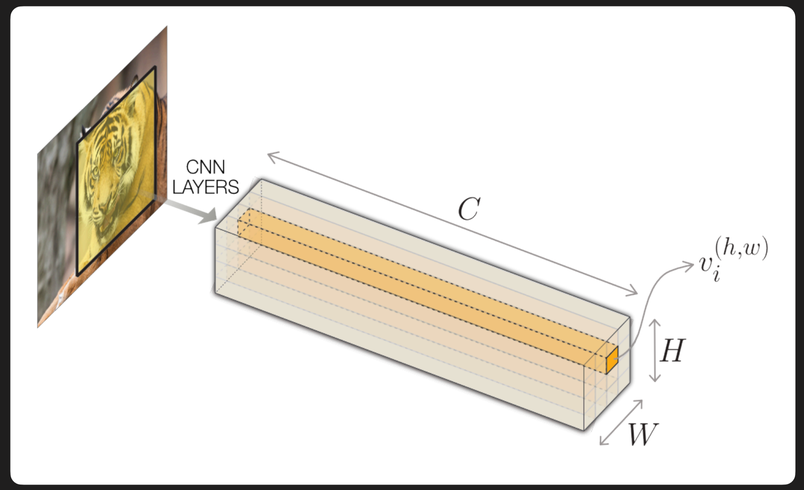
MLsploit KDD Showcase 2019

Which region of an image is the most predictive?



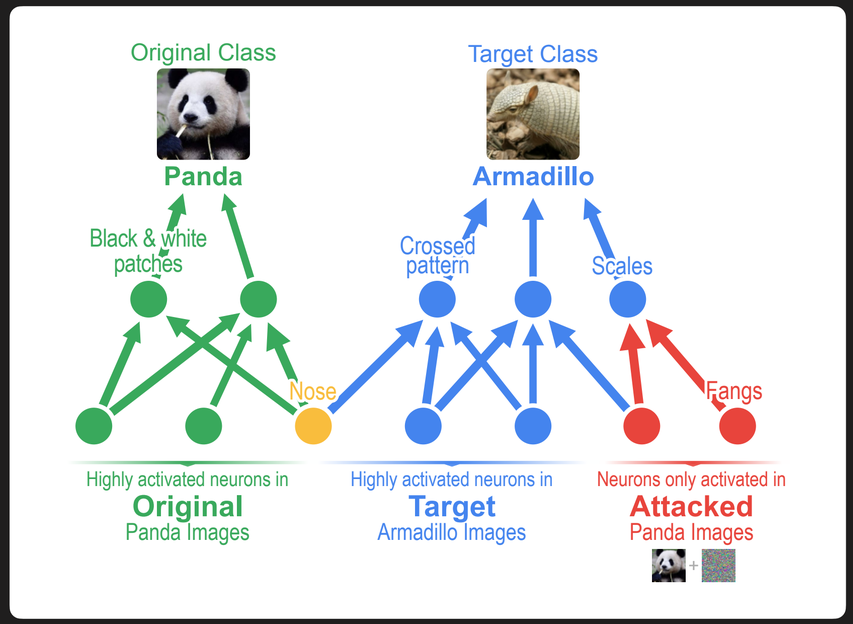
GOGGLES

Data Programming



GOGGLES

Adversarial ML Interpretation



Bluff

GOGGLES

Automatic Image Labeling with *Affinity Coding*

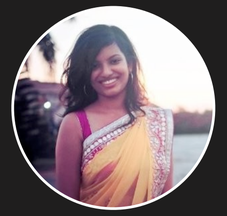
SIGMOD 2020

Open-sourced at *github.com/chu-data-lab/GOGGLES*

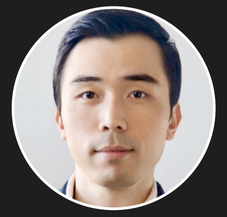
Nilaksh Das
Georgia Tech

Sanya Chaba
Georgia Tech

Renzhi Wu
Georgia Tech

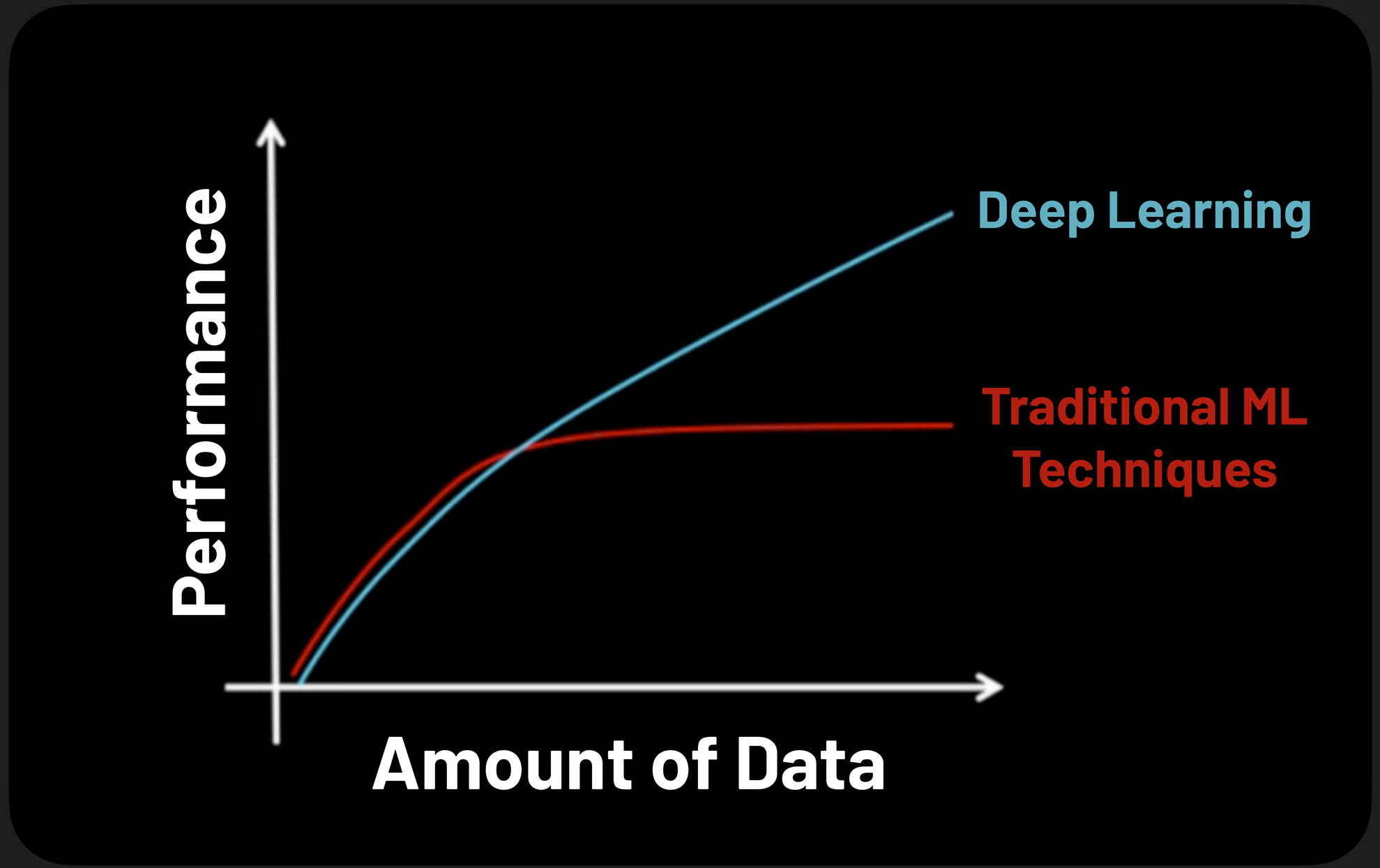


Sakshi Gandhi
Georgia Tech



Polo Chau
Georgia Tech

Xu Chu
Georgia Tech

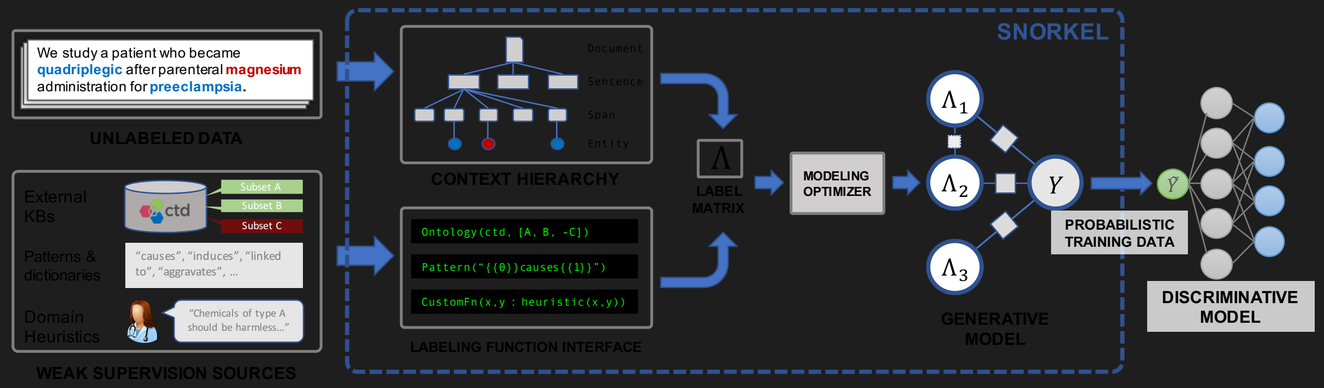

Performance
Deep Learning
Traditional ML
Techniques
Amount of Data

Collecting massive labeled datasets is

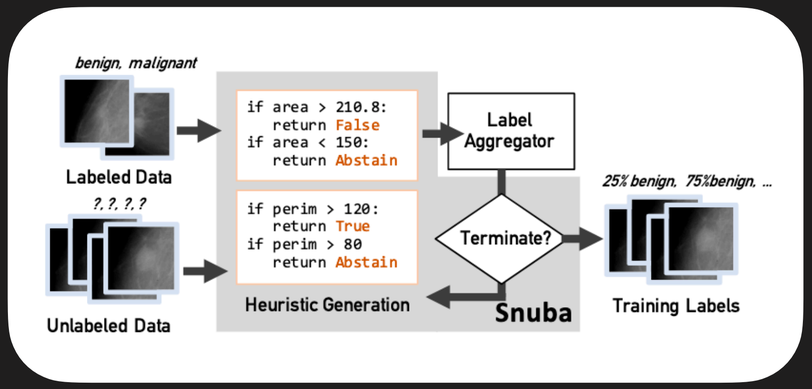
COSTLY

Data Programming

Snorkel



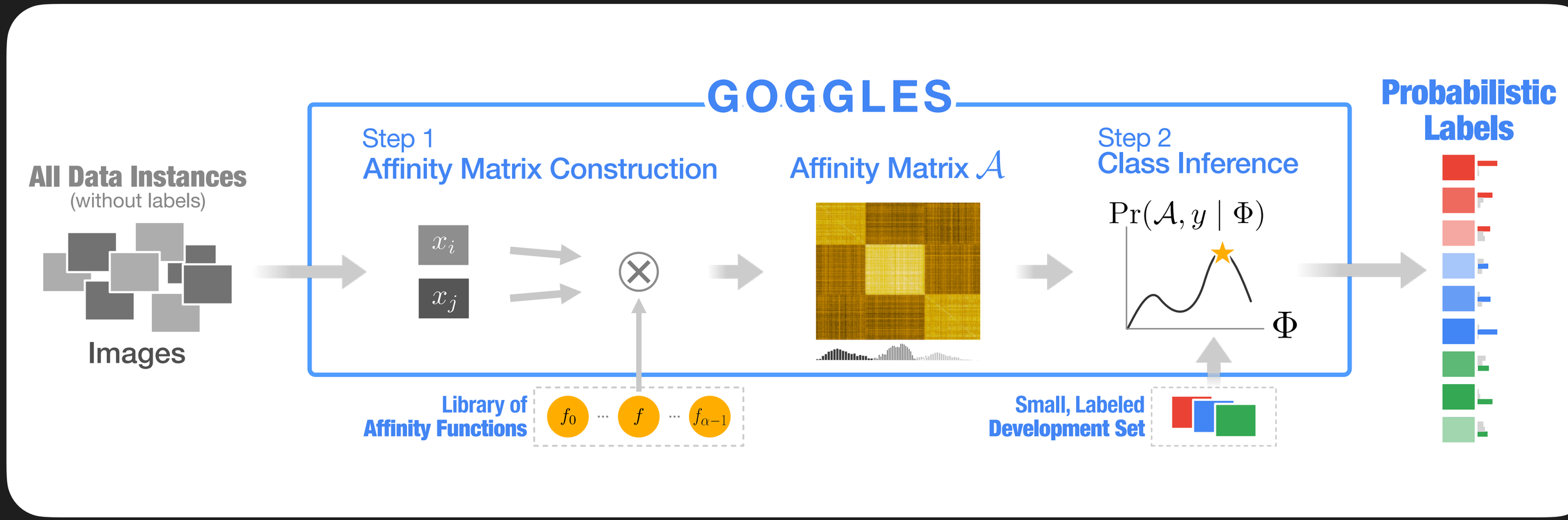
Snuba



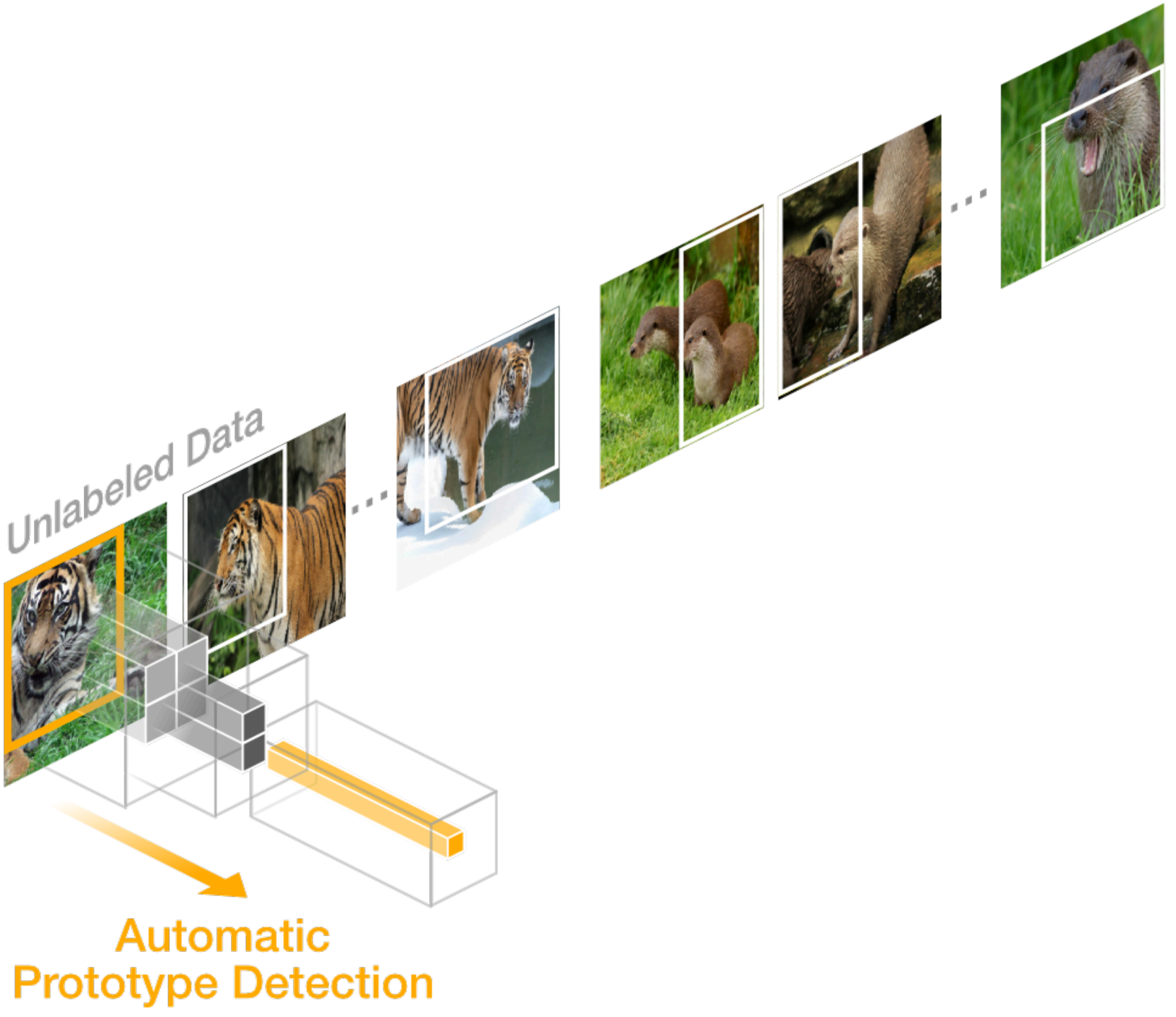
Affinity Coding

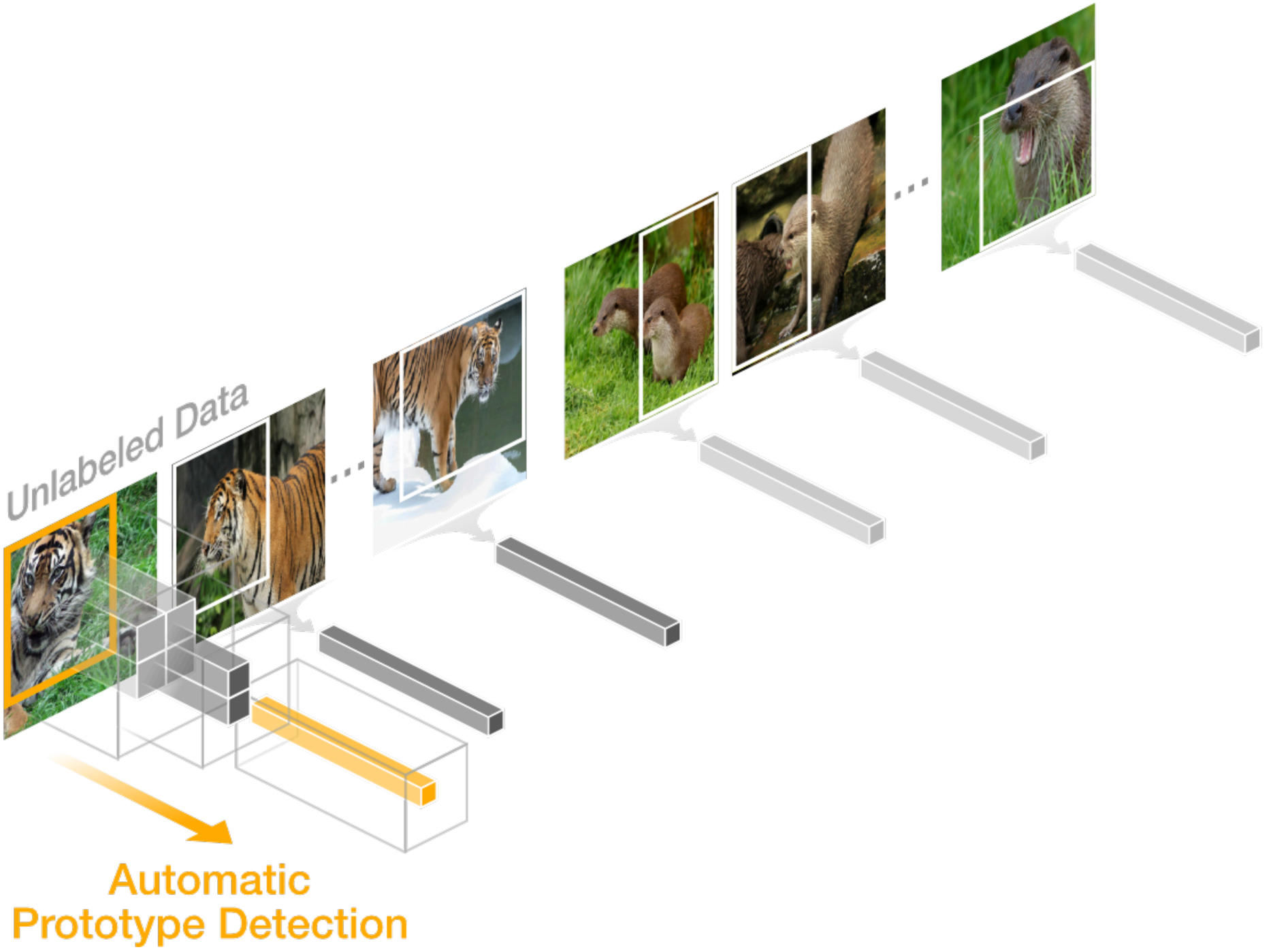
Affinity Coding

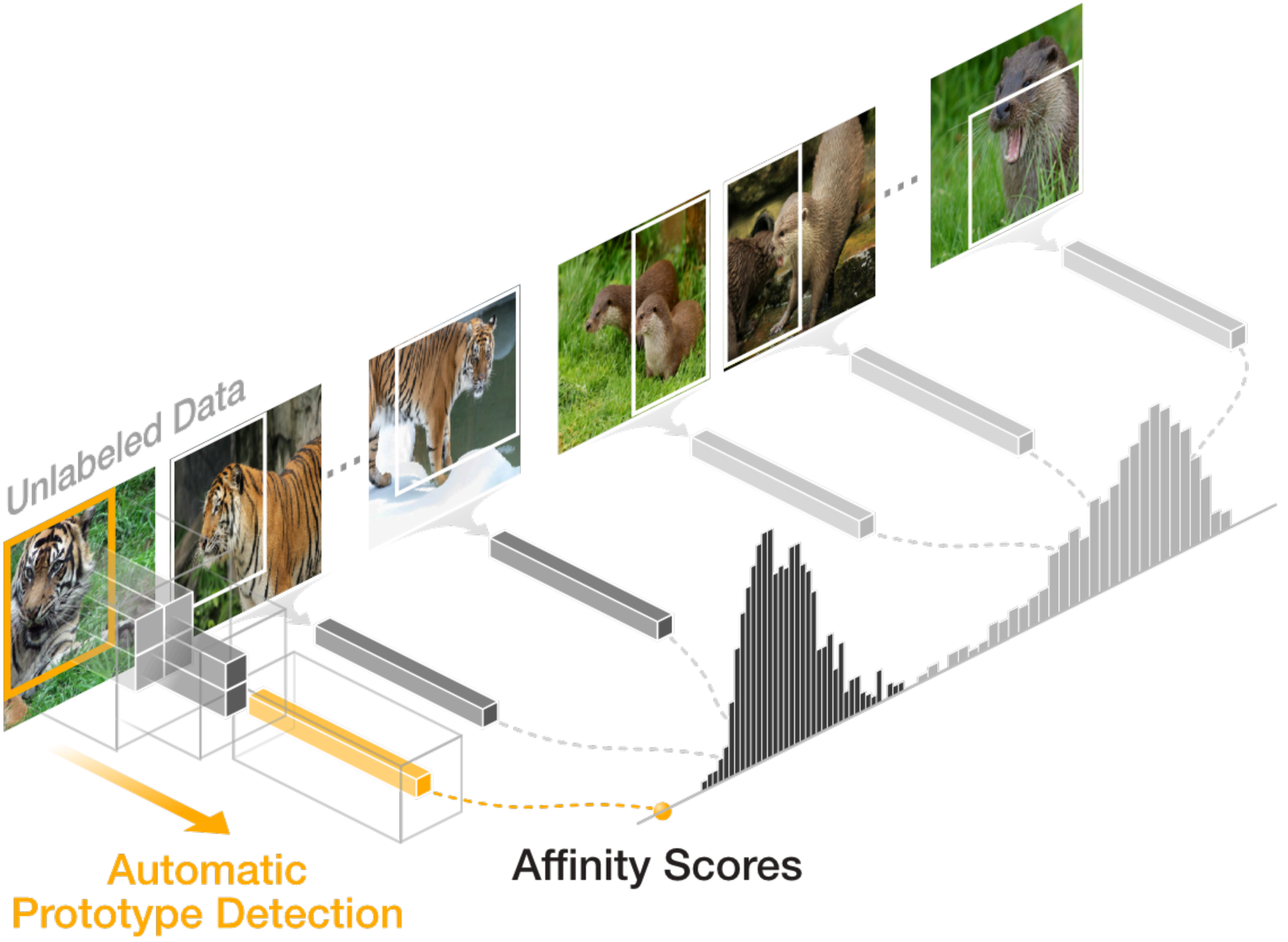
with GOGGLES

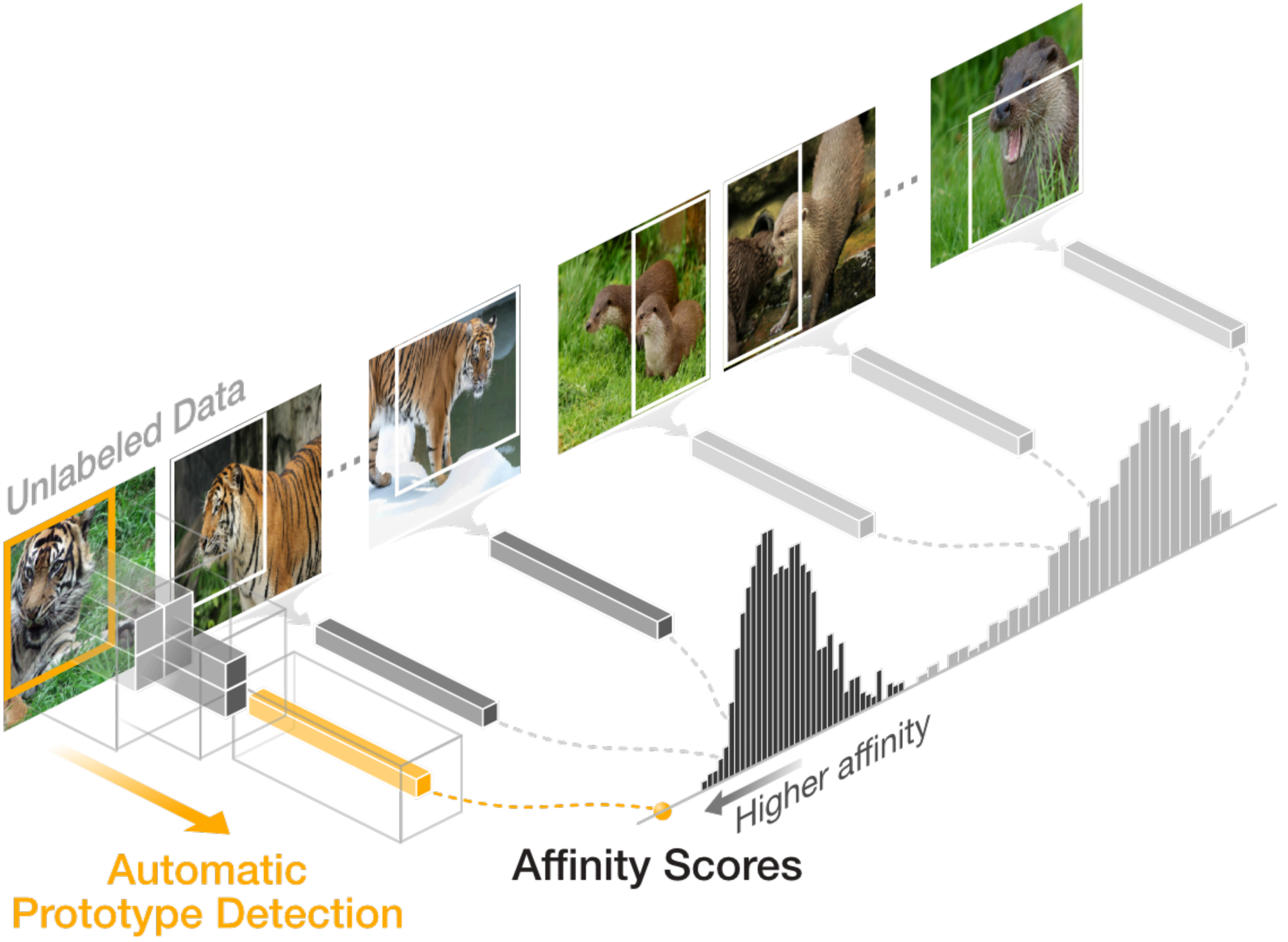

GOGGLES
All Data Instances
(without labels)
Images
Step 1
Affinity Matrix Construction
x_i
x_j
Affinity Matrix $\mathcal{A}$
Step 2
Class Inference
$\Pr(\mathcal{A}, y \mid \Phi)$
Φ
Probabilistic Labels
Library of
Affinity Functions
f_0 ... f ... $f_{\alpha-1}$
Small, Labeled
Development Set

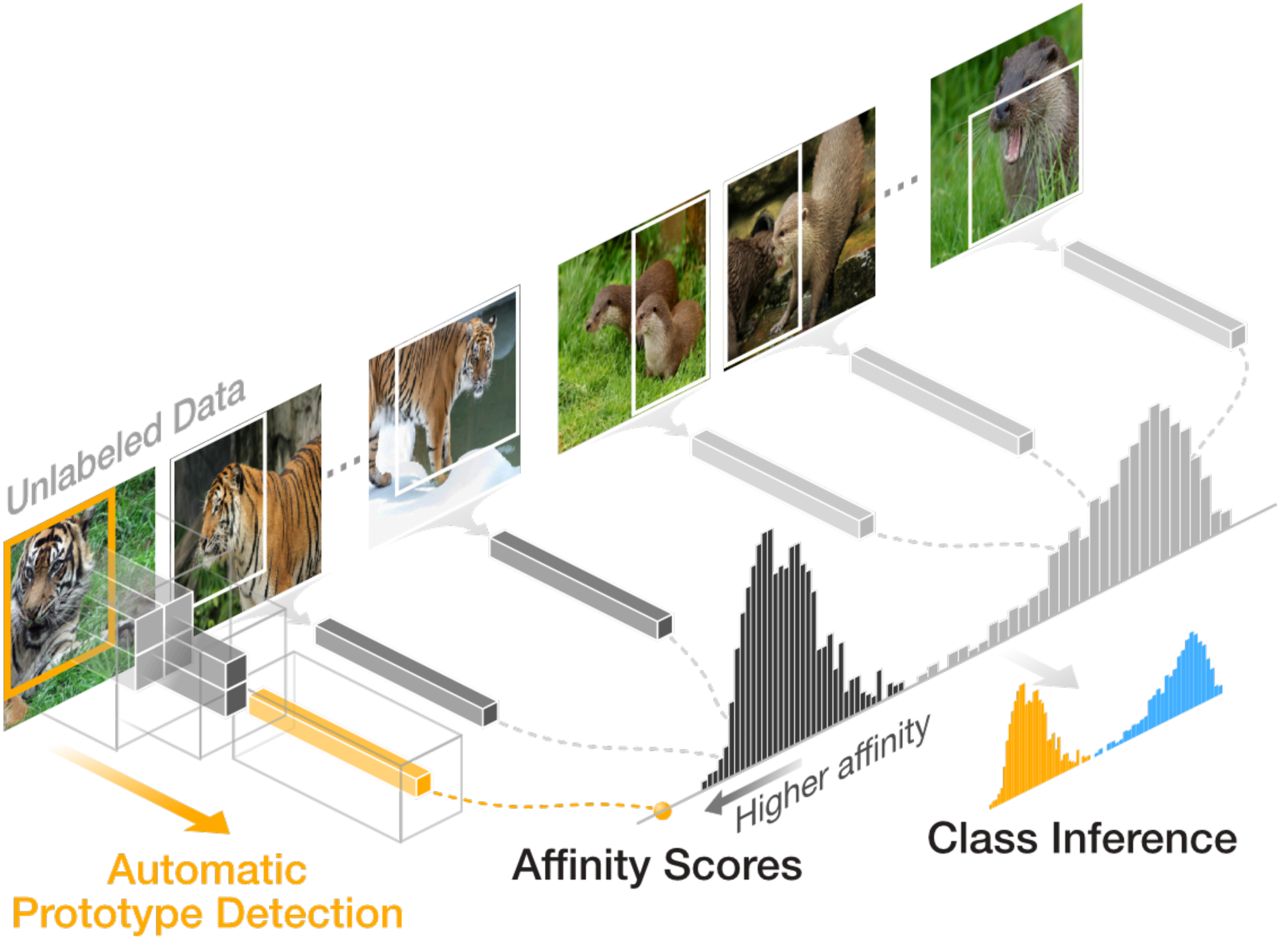

Unlabeled Data

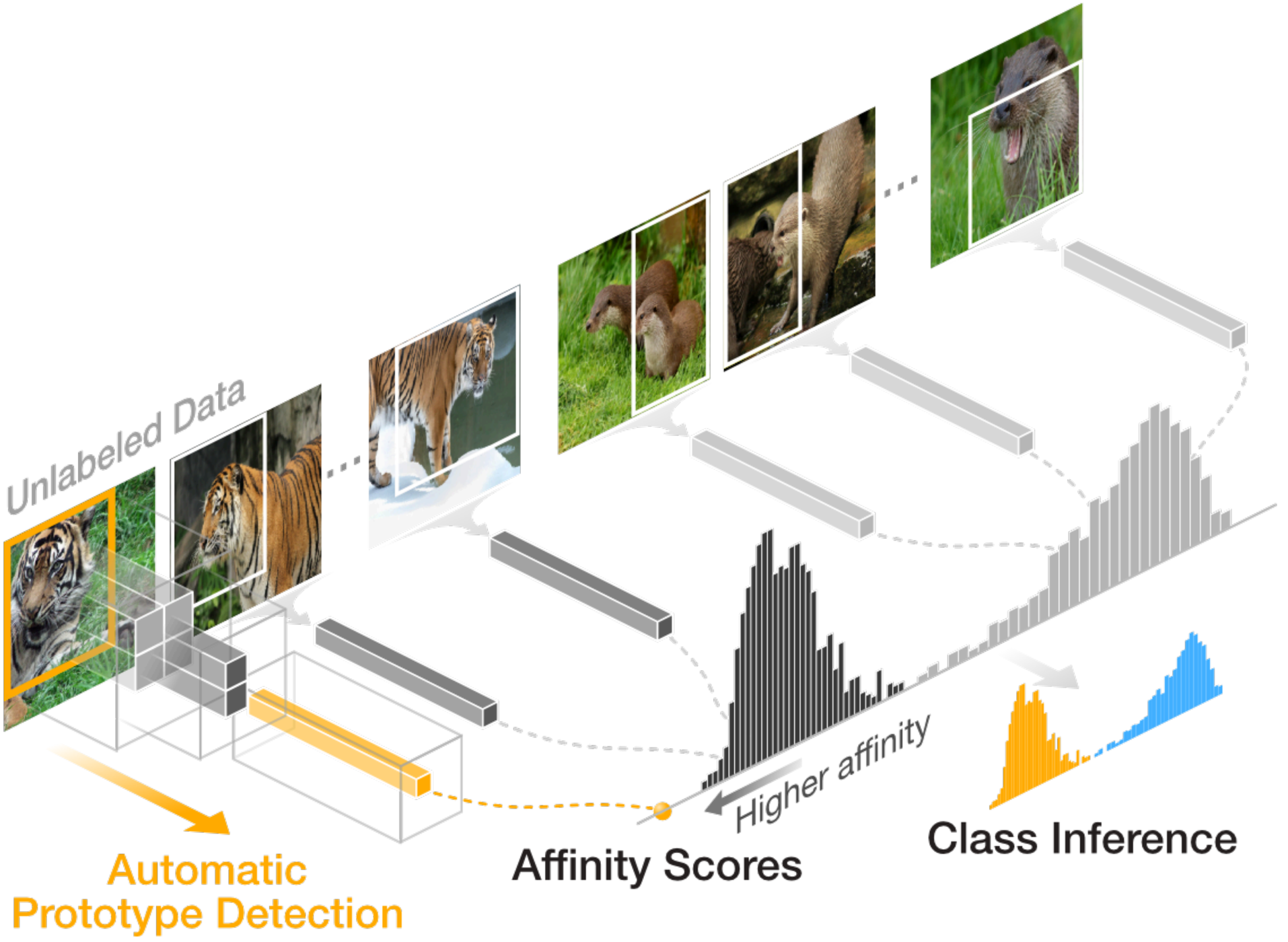

Unlabeled Data
Automatic
Prototype Detection


Unlabeled Data
Automatic
Prototype Detection


Unlabeled Data
Automatic
Prototype Detection
Affinity Scores


Unlabeled Data
Automatic
Prototype Detection
Affinity Scores
Higher affinity


Unlabeled Data
Automatic
Prototype Detection
Affinity Scores
Higher affinity
Class Inference


Unlabeled Data
Automatic
Prototype Detection
Affinity Scores
Higher affinity
Class Inference

VGG16

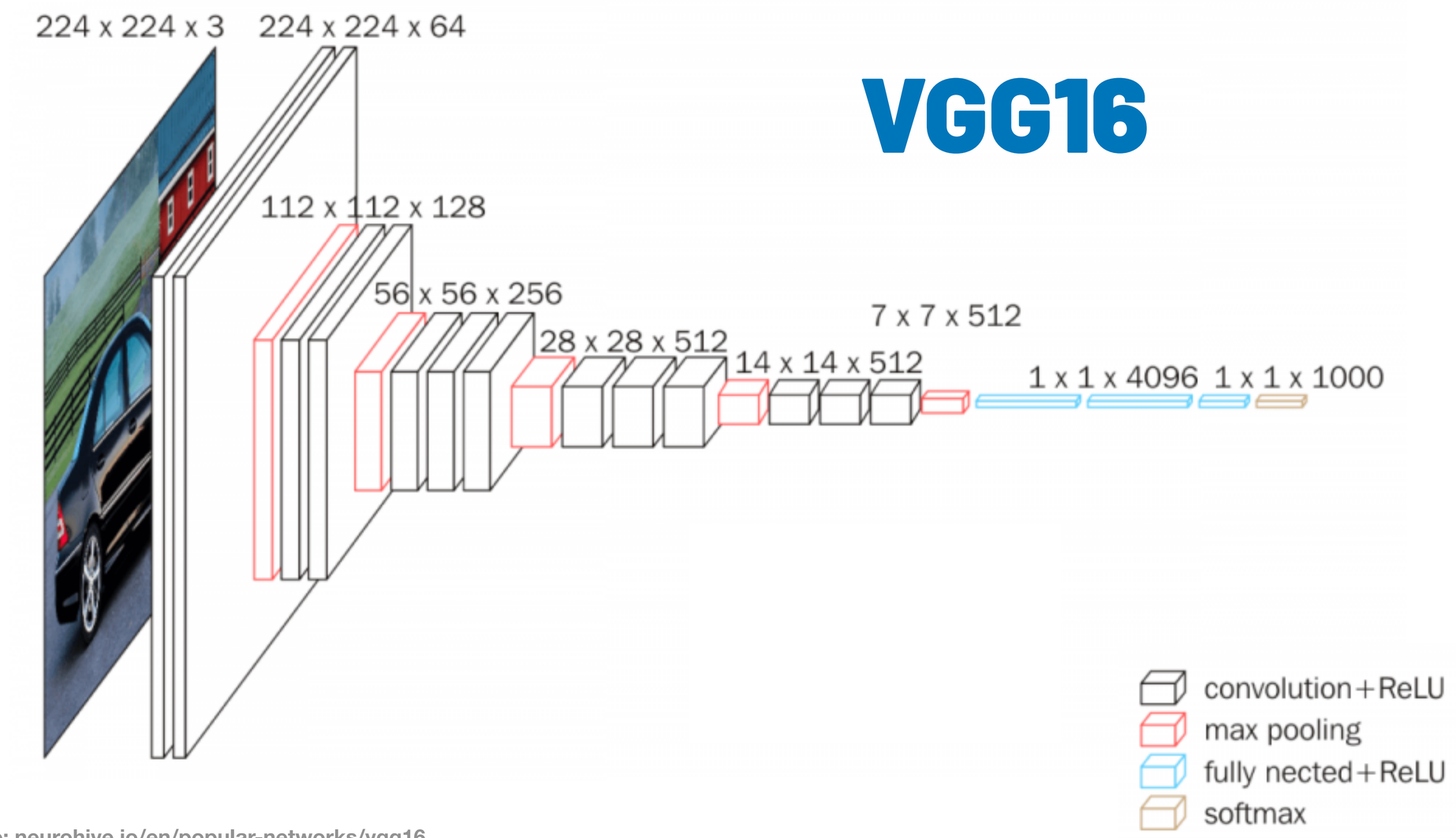
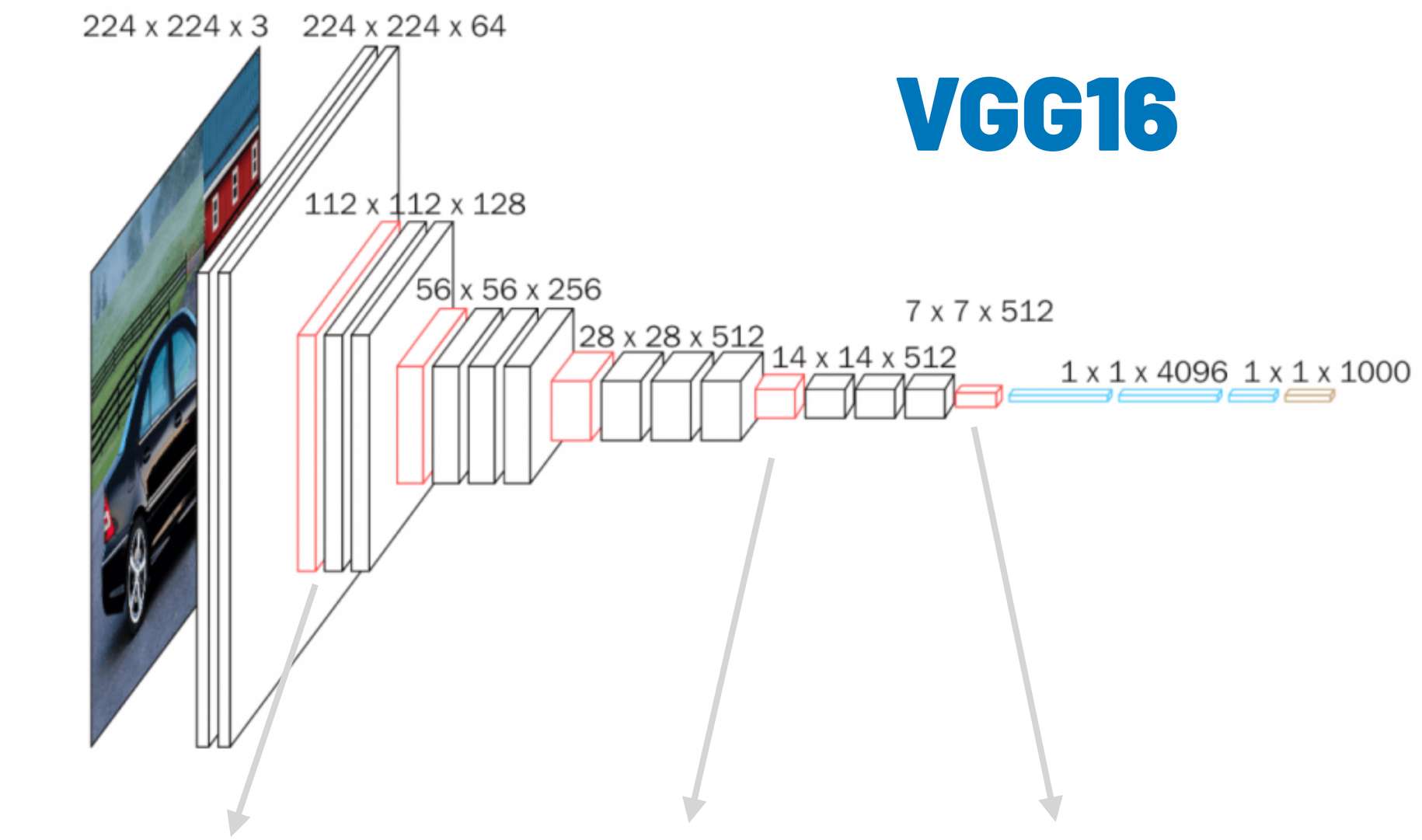
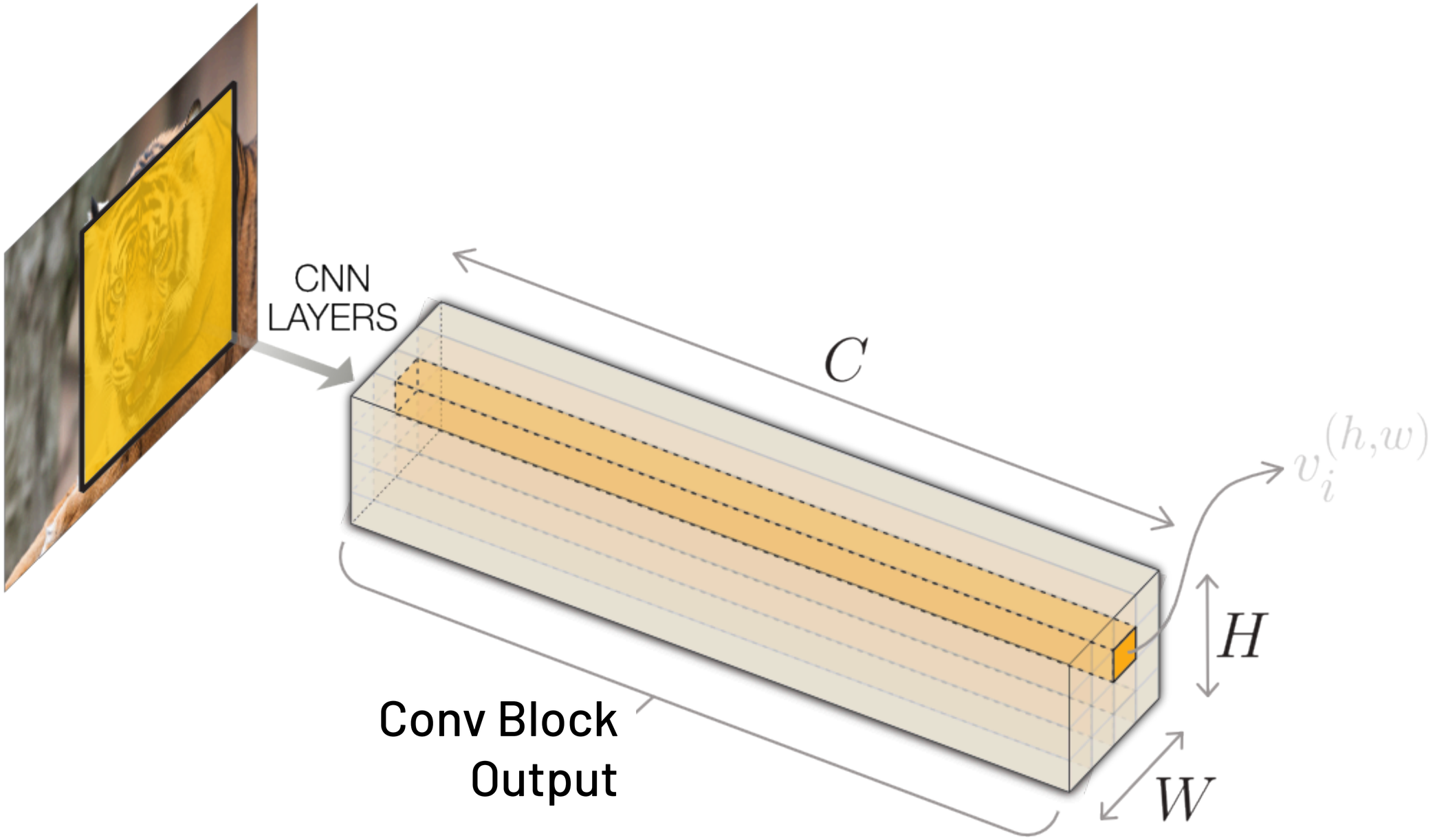


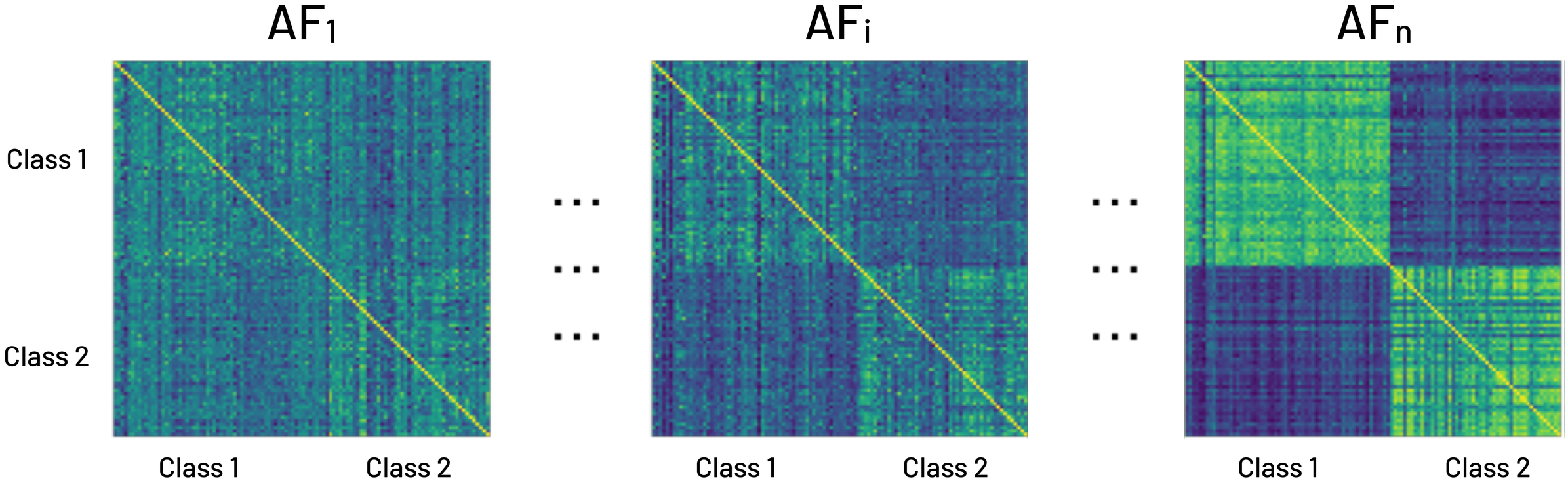
Image: neurohive.io/en/popular-networks/vgg16


224 x 224 x 3
224 x 224 x 64
112 x 112 x 128
56 x 56 x 256
28 x 28 x 512
14 x 14 x 512
7 x 7 x 512
1 x 1 x 4096
1 x 1 x 1000
VGG16
Affinity Function1
...
AFi
...
AFn


CNN
LAYERS
C
$v_i^{(h,w)}$
H
W
Conv Block
Output

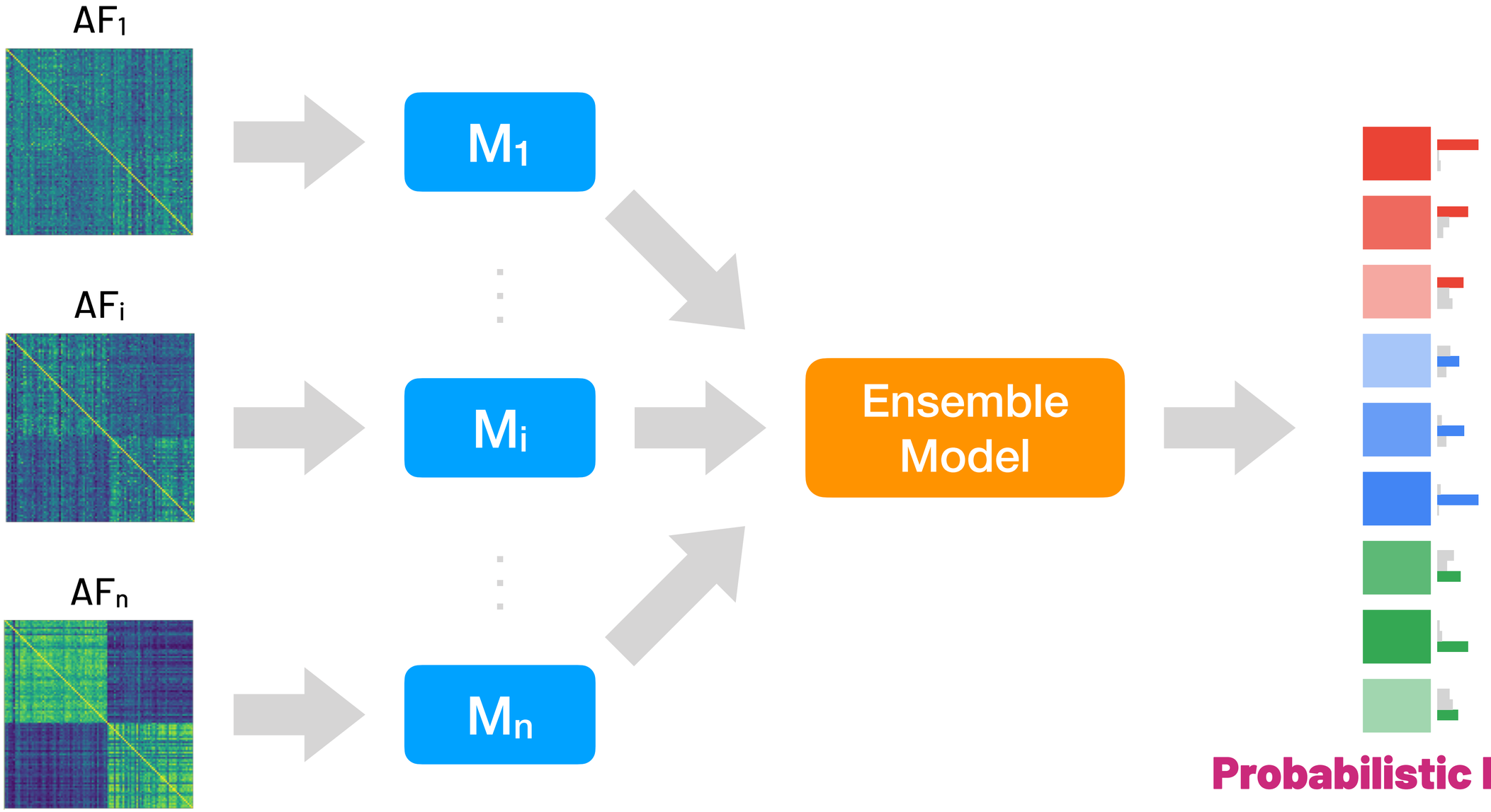
Affinity Matrices

from different Affinity Functions



Class Inference

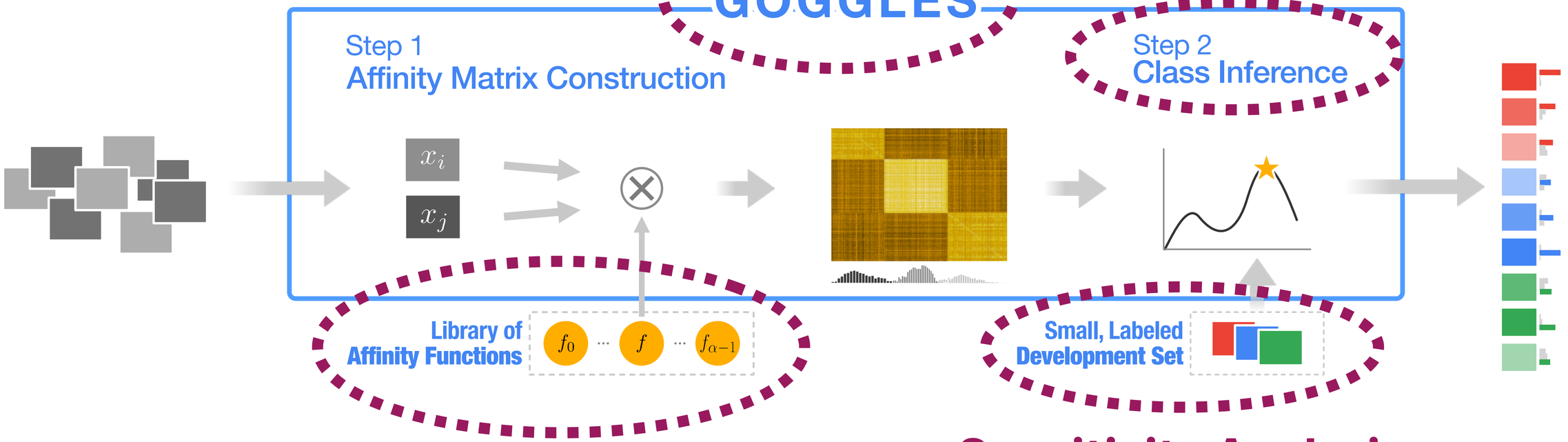
with Hierarchical Generative Model



Exhaustive **Evaluation**

GOGGLES vs **Snorkel** vs **Snuba**

GOGGLES vs **K-means** vs **GMM** vs **Spectral**



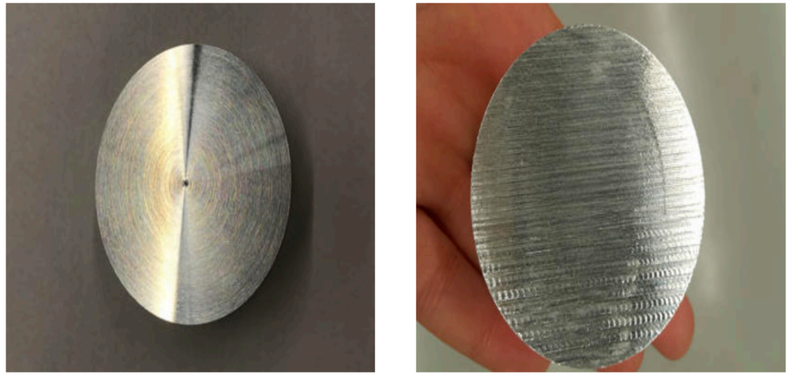
GOGGLES vs **HoG** vs **Logits**

Sensitivity Analysis

Diverse **Datasets**

CUB (Birds)

GTSRB (Traffic Signs)



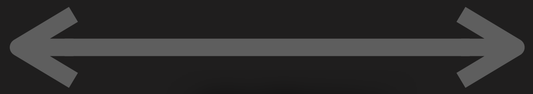
Surface (Industrial Metal Finish)

TB-Xray (Tuberculosis)

PN-Xray (Pneumonia)

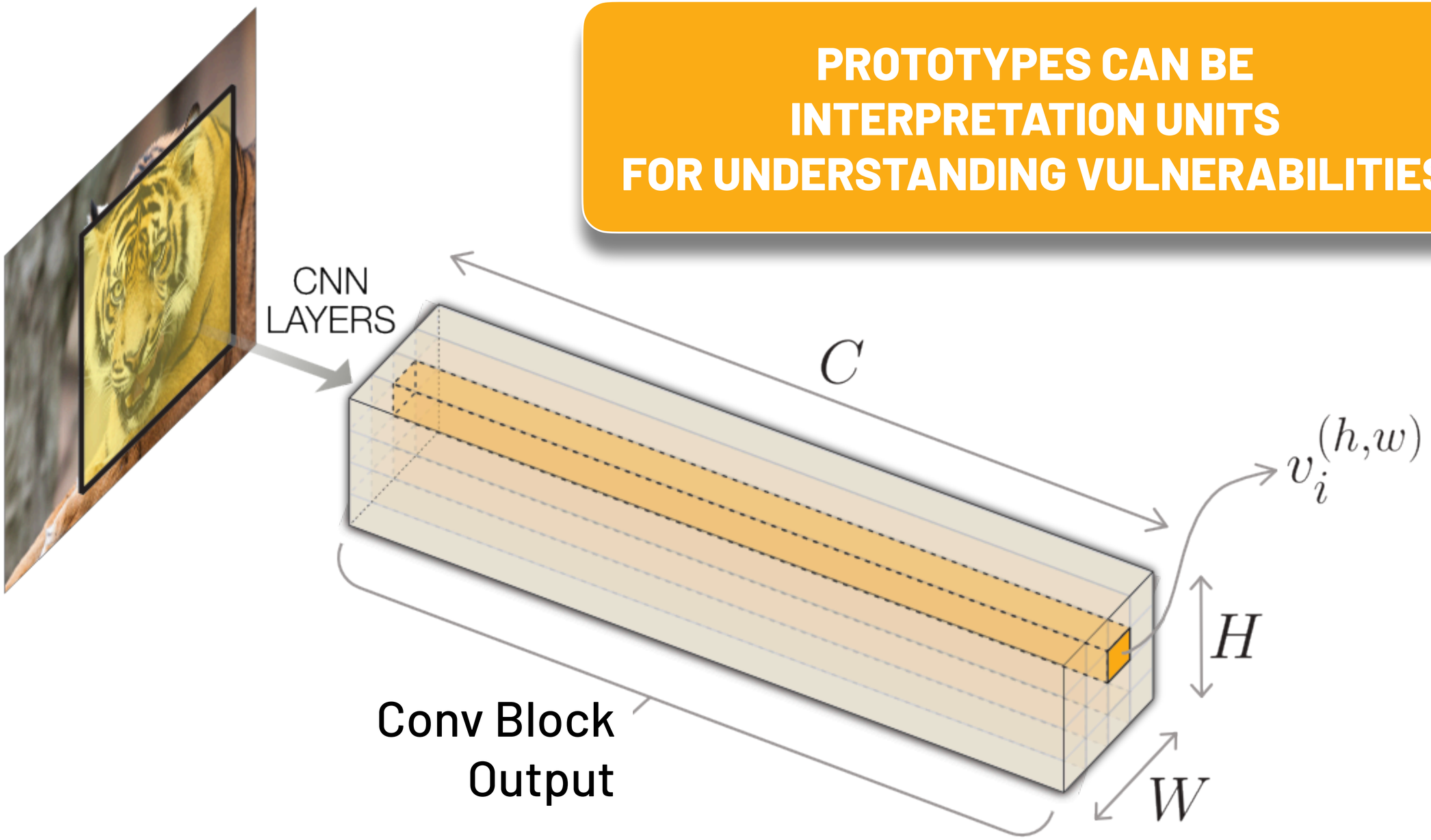
Evaluating **Labeling Accuracy** on **Training Set**

Dataset	GOGGLES (our results)	Data Programming		Representation		Class Inference Baselines		
		Snorkel	Snuba	HoG	Logits	K-Means	GMM	Spectral
CUB	97.83	89.17	58.83	62.93	96.35	98.67	97.62	72.08
GTSRB	70.51	-	62.74	75.48	64.77	70.74	69.64	62.40
Surface	89.18	-	57.86	85.82	54.08	69.08	69.14	60.82
TB-Xray	76.89	-	59.47	69.13	67.16	76.33	76.70	75.00
PN-Xray	74.39	-	55.50	53.11	71.18	50.66	68.66	75.90
Average	**81.76**	-	58.88	69.30	70.71	73.09	76.35	69.24



~23%

PROTOTYPES CAN BE INTERPRETATION UNITS FOR UNDERSTANDING VULNERABILITIES



Bluff

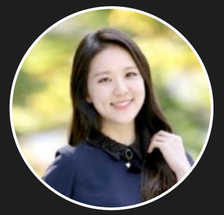
IEEE VIS 2020

Interactively Deciphering Adversarial Attacks On Deep Neural Networks

Live Demo at *poloclub.github.io/bluff*

Nilaksh Das*
Georgia Tech

*equal contribution



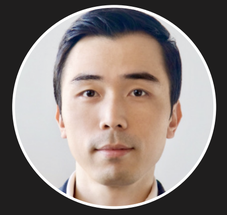
Haekyu Park*
Georgia Tech

Zijie J. Wang
Georgia Tech

Fred Hohman
Georgia Tech

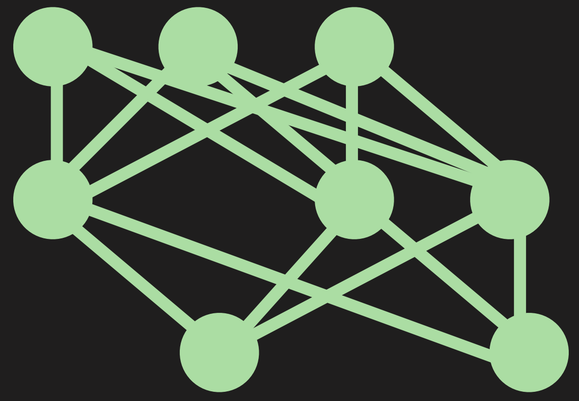
Robert Firstman
Georgia Tech

Emily Rogers
GT Research Institute

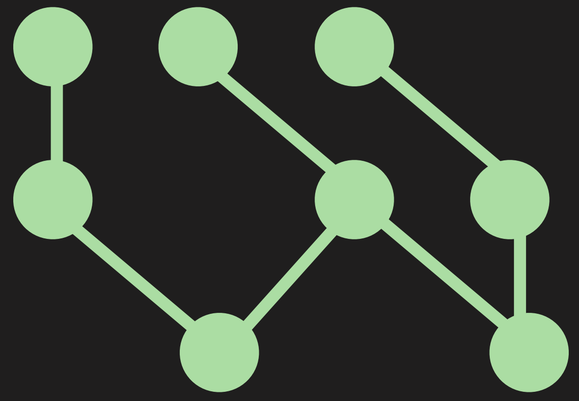


Polo Chau
Georgia Tech

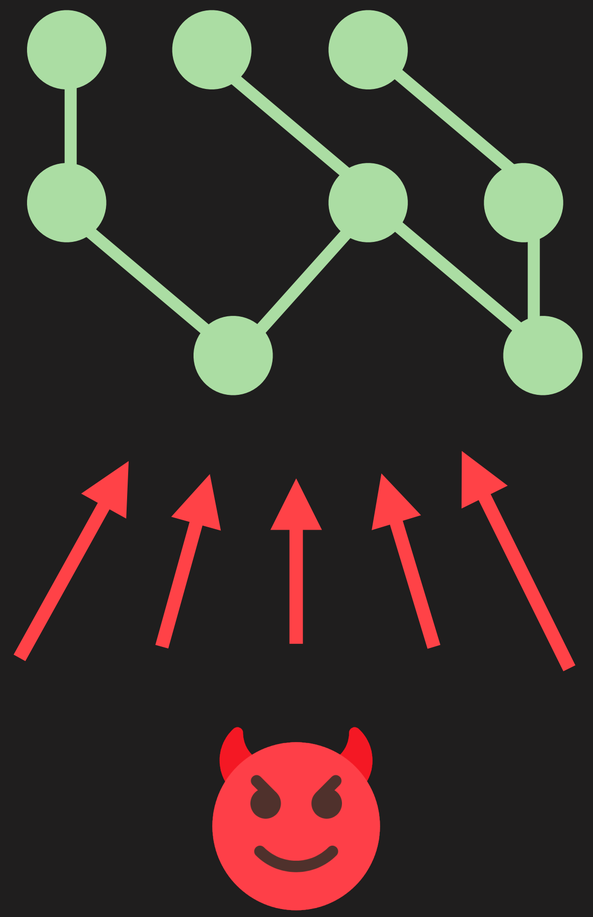
Understanding Adversarial Attacks



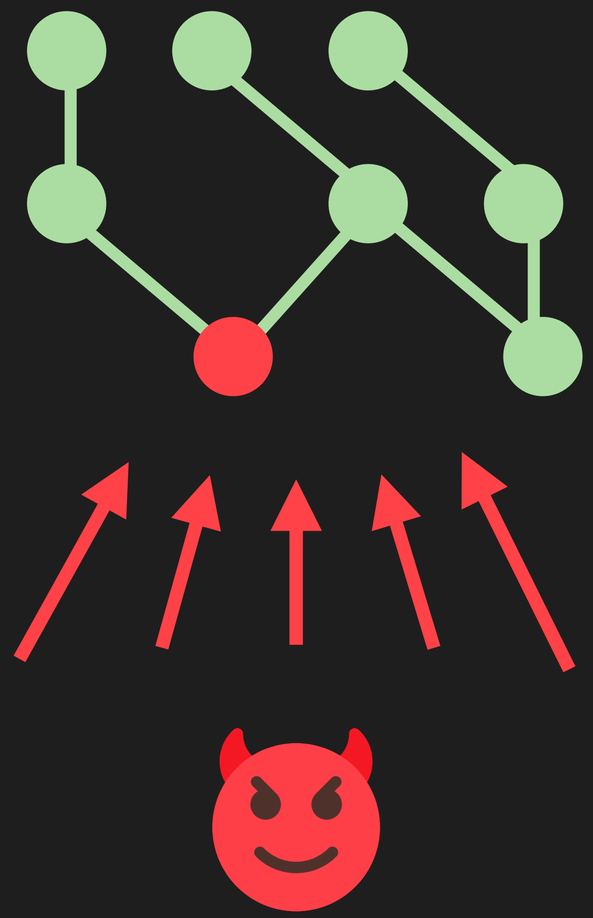
Understanding Adversarial Attacks



Understanding Adversarial Attacks

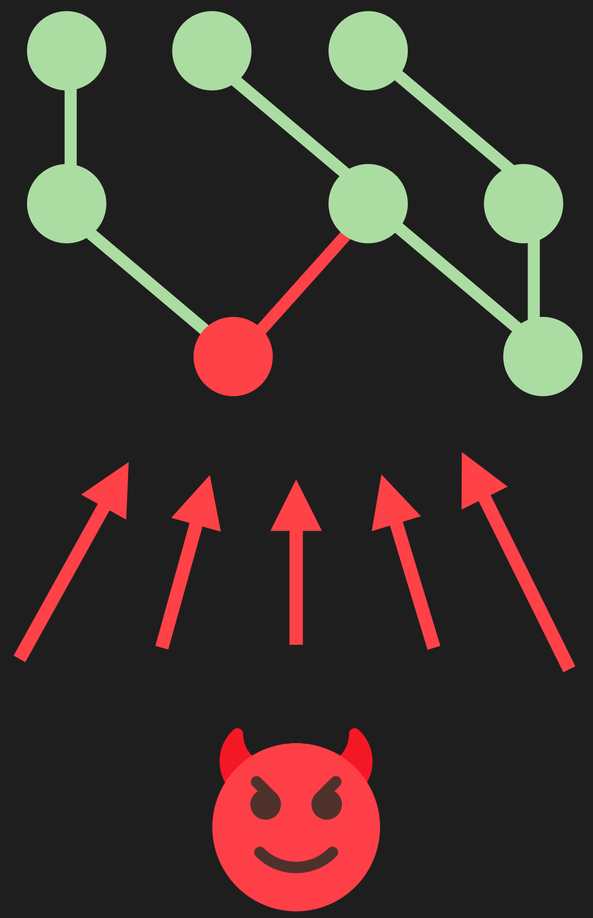


Understanding Adversarial Attacks



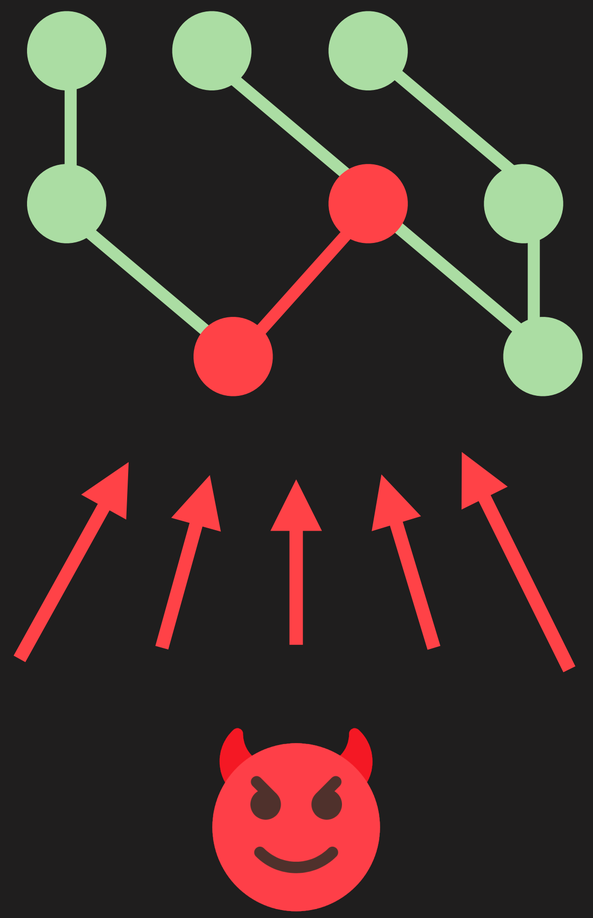
How adversarial attacks
permeate a model?

Understanding Adversarial Attacks



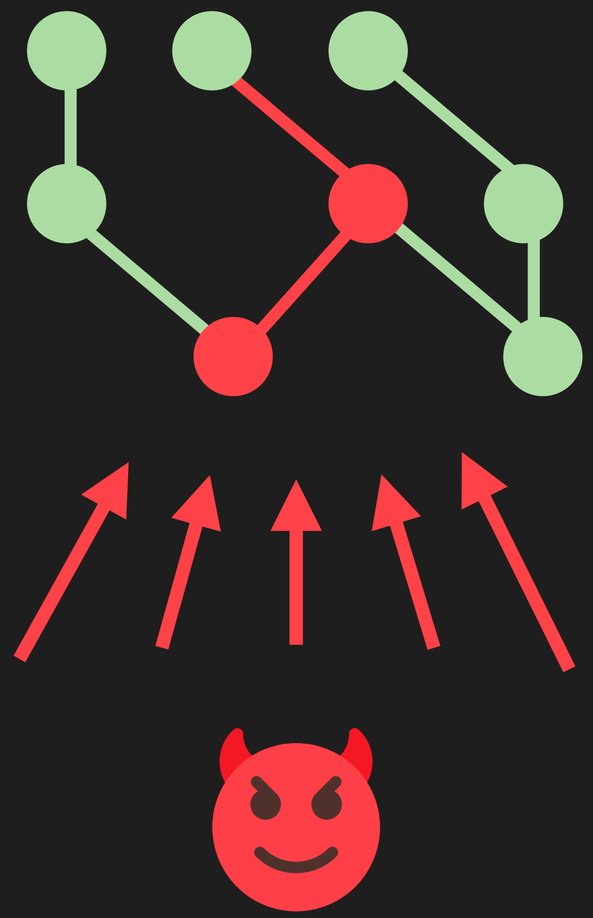
How adversarial attacks
permeate a model?

Understanding Adversarial Attacks



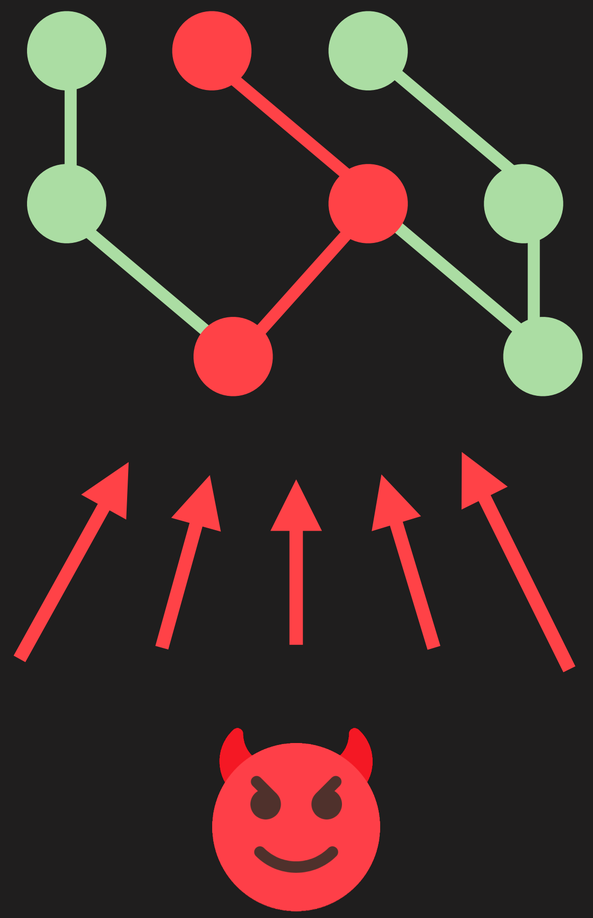
How adversarial attacks
permeate a model?

Understanding Adversarial Attacks



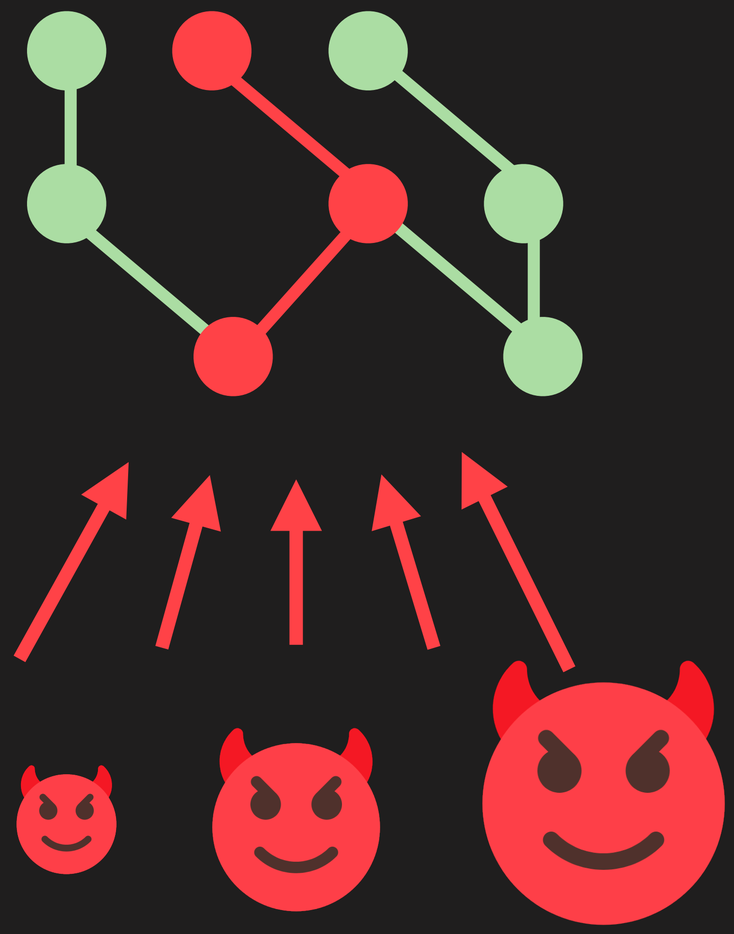
How adversarial attacks
permeate a model?

Understanding Adversarial Attacks



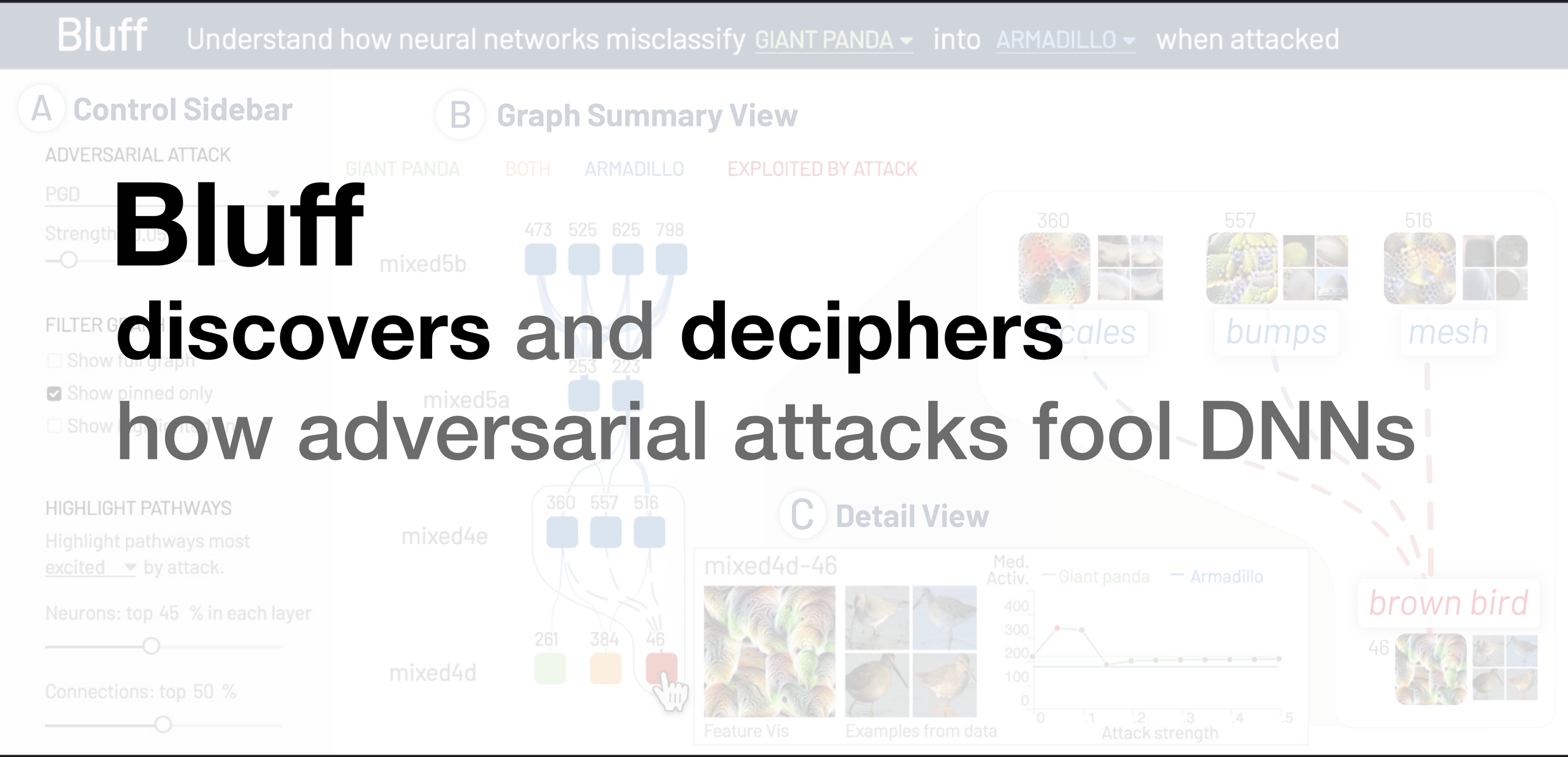
How adversarial attacks
permeate a model?

Understanding Adversarial Attacks



How adversarial attacks permeate a model?

Variation in attack patterns with varying attack strengths?

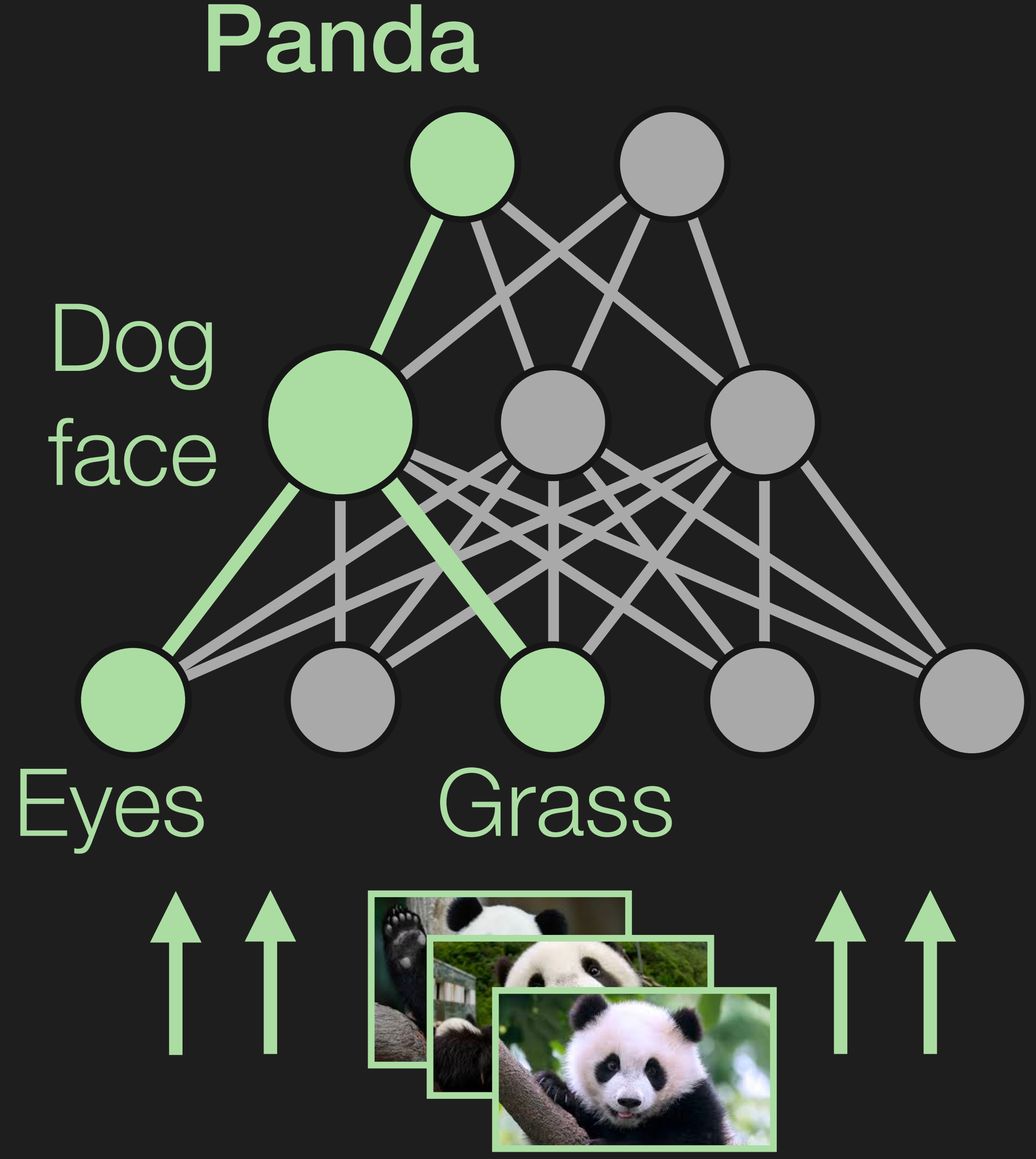


Bluff

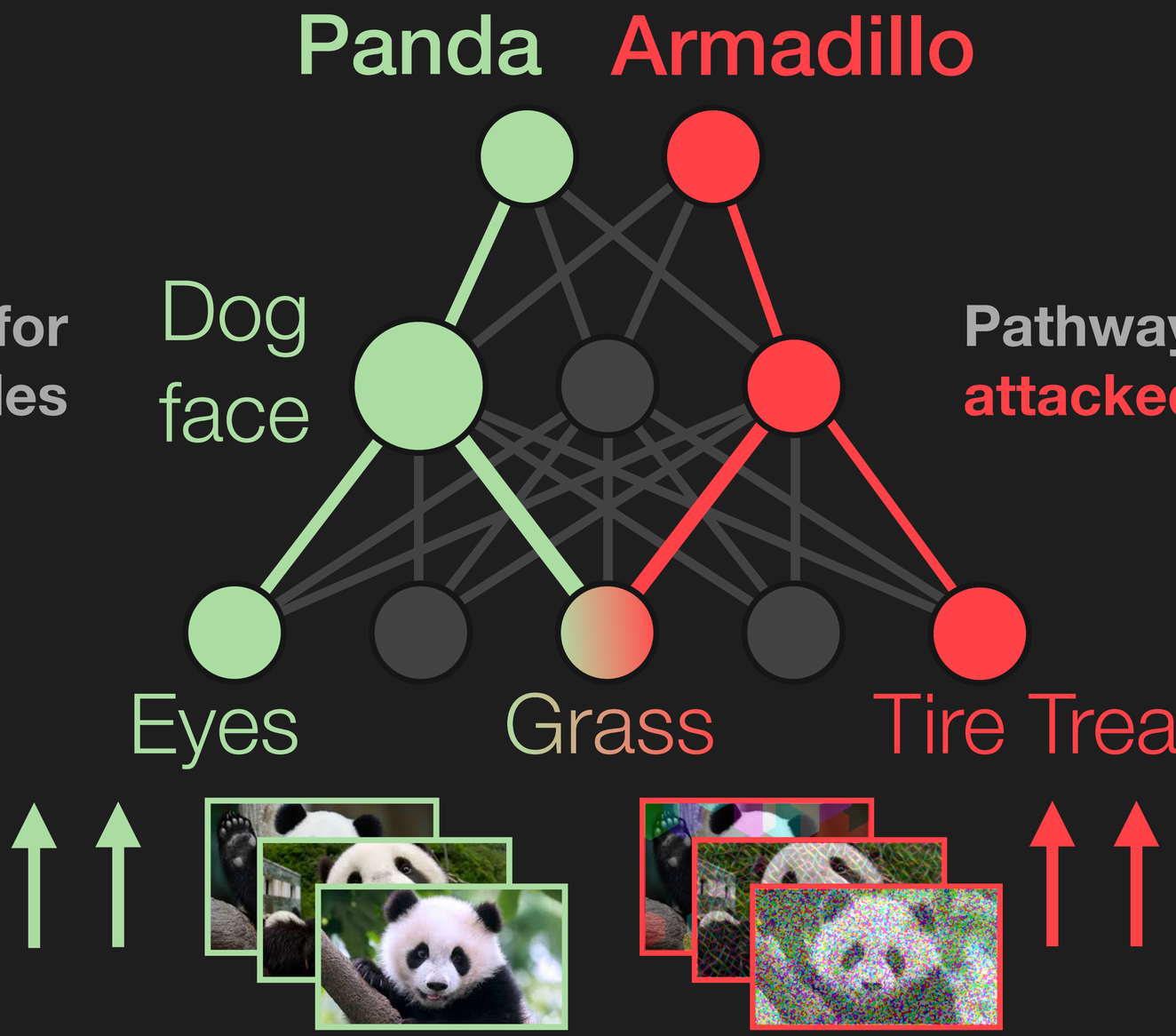
discovers and deciphers

how adversarial attacks fool DNNs

Activation Pathways — most **activated** pathways



Activation Pathways — most **activated** pathways



Pathways for benign examples

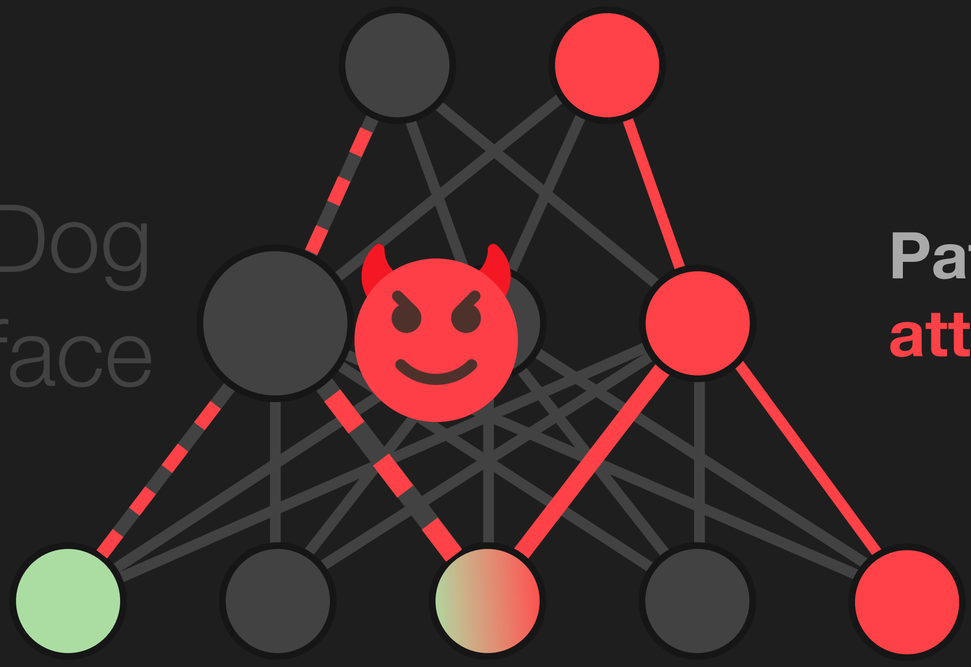
Pathways for attacked examples

Activation Pathways — most **inhibited** pathways

Panda Armadillo

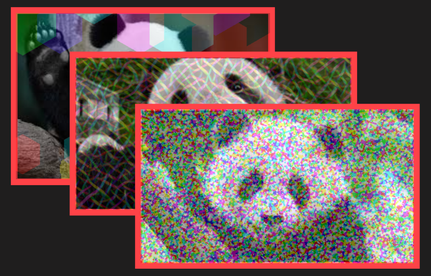
Pathways for benign examples

Dog face

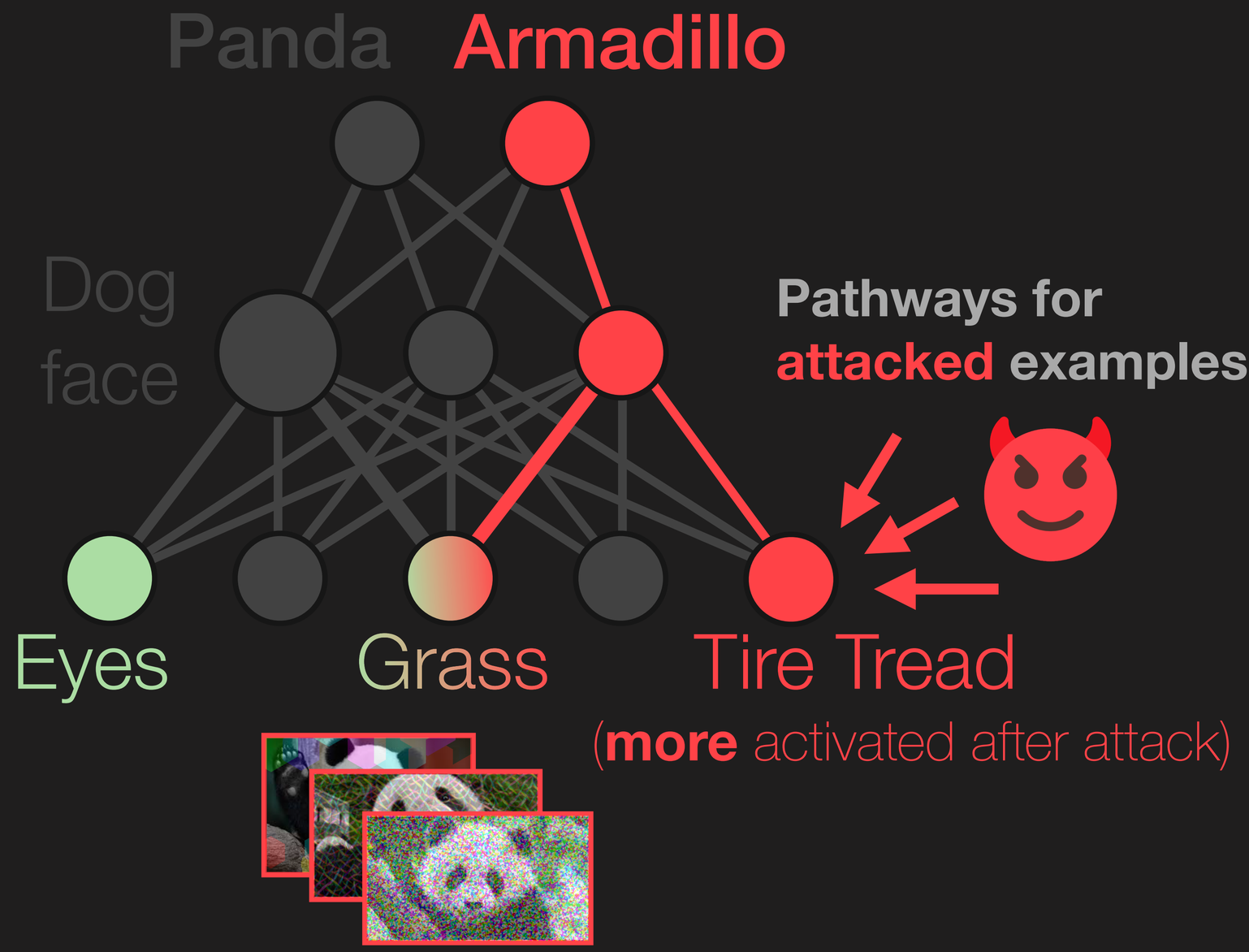


Pathways for attacked examples

Eyes Grass Tire Tread



Activation Pathways — most **excited** pathways



Bluff

Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

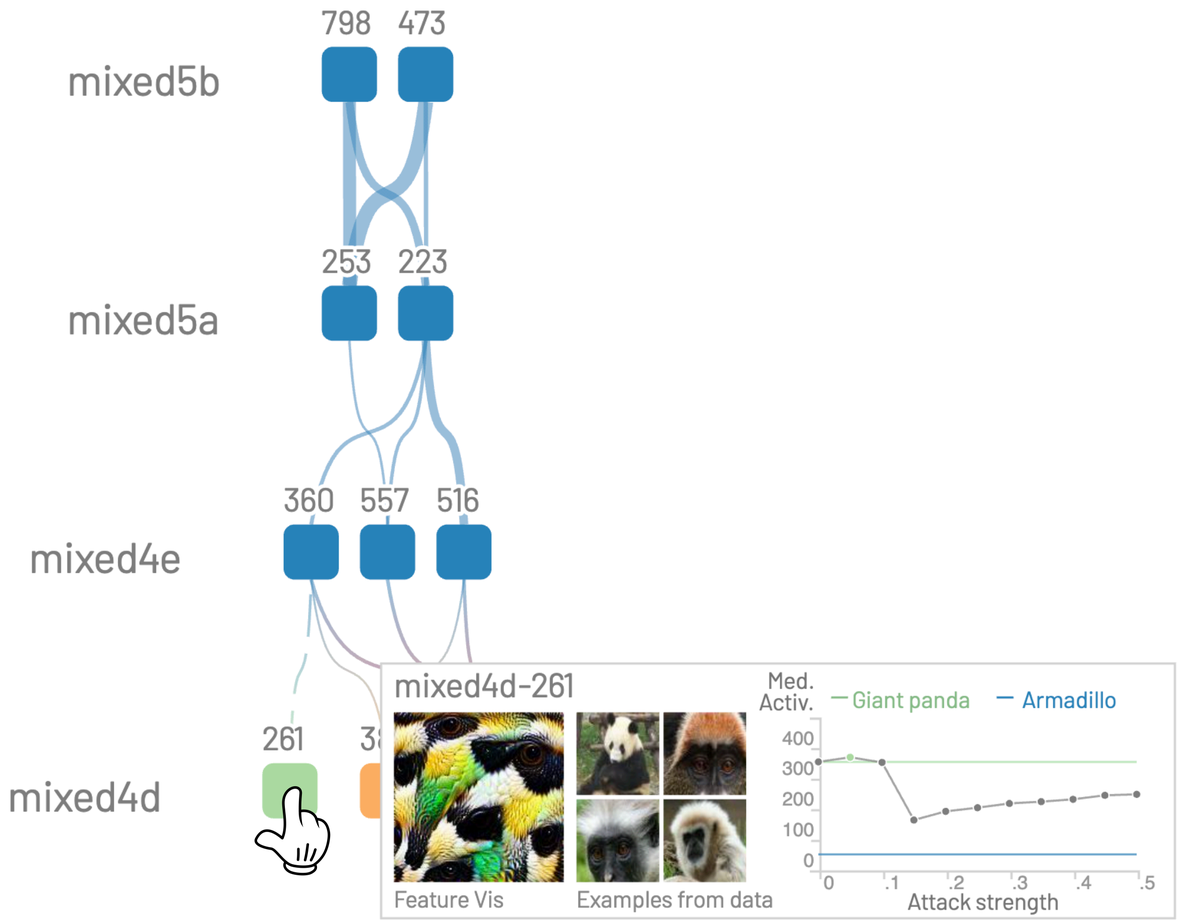
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- N/A
- Stronger
- Weaker
- Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

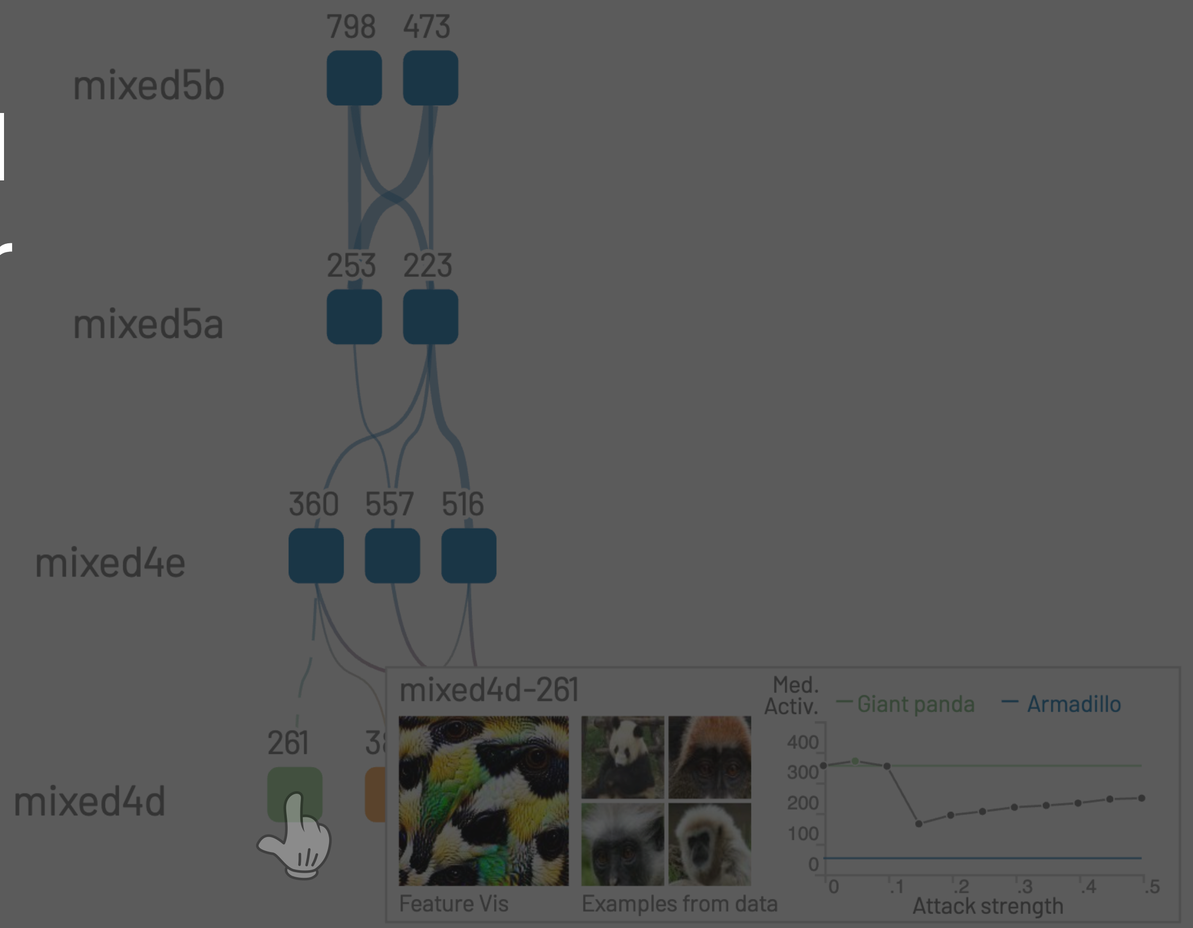
Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK

Control sidebar



Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

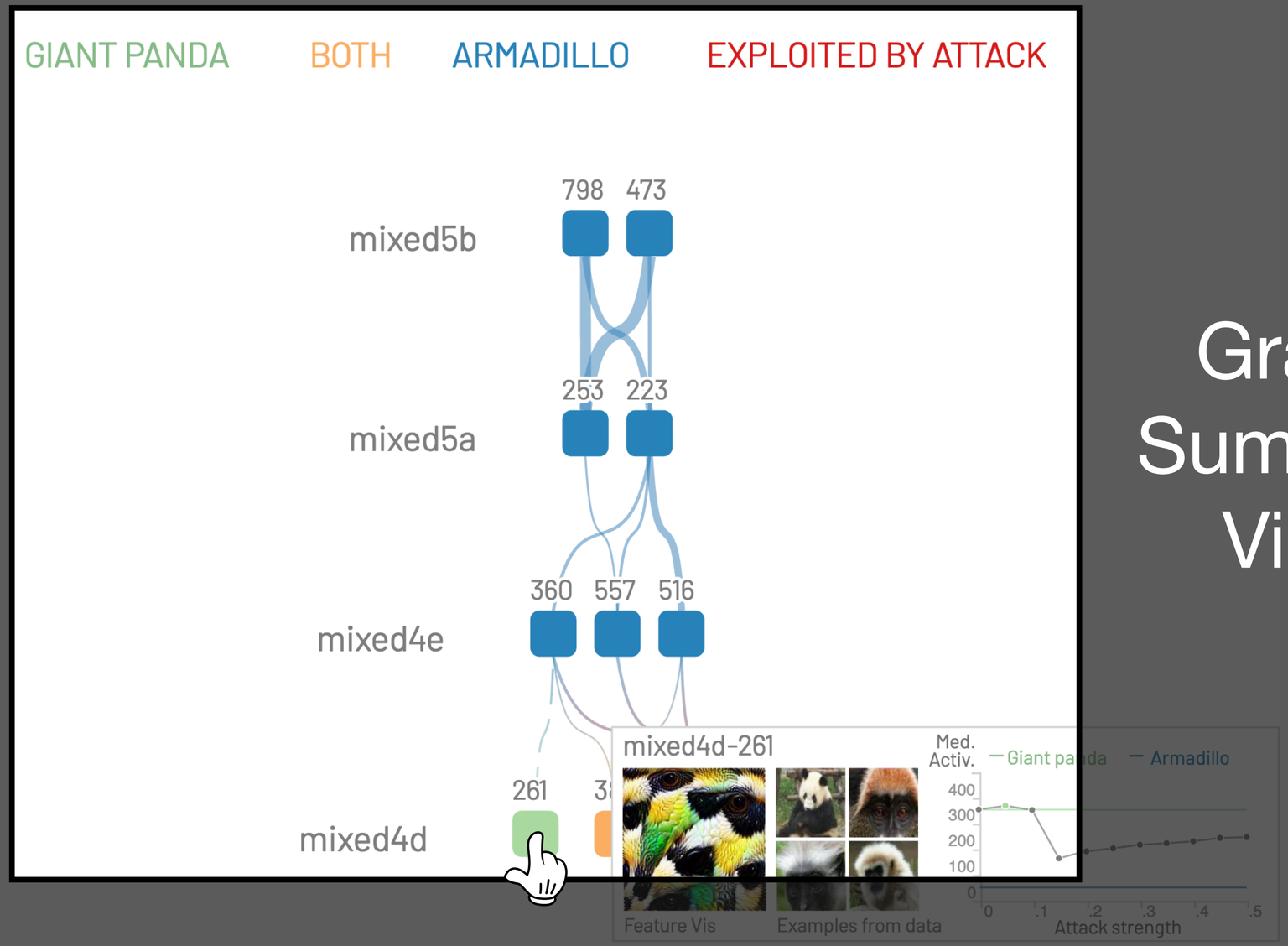
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger



Graph Summary View

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK

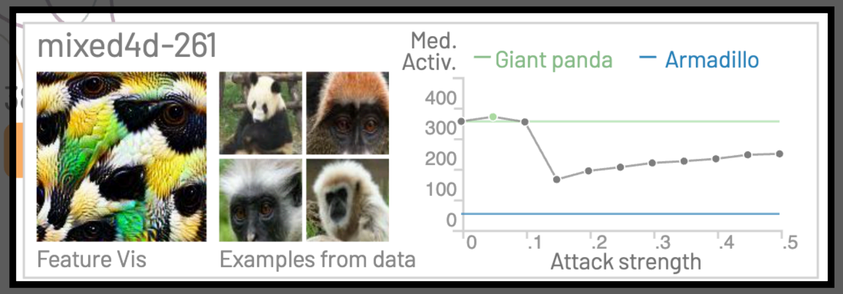
mixed5b 798 473

mixed5a 253 223

mixed4e 360 557 516

mixed4d 261

Detail View



Bluff Understand how neural networks

misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

NDA BOTH ARMADILLO

FILTER GRAPH

Show full graph

Show pinned only

Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger

Panda

PGD Attack

Armadillo

Bluff

Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

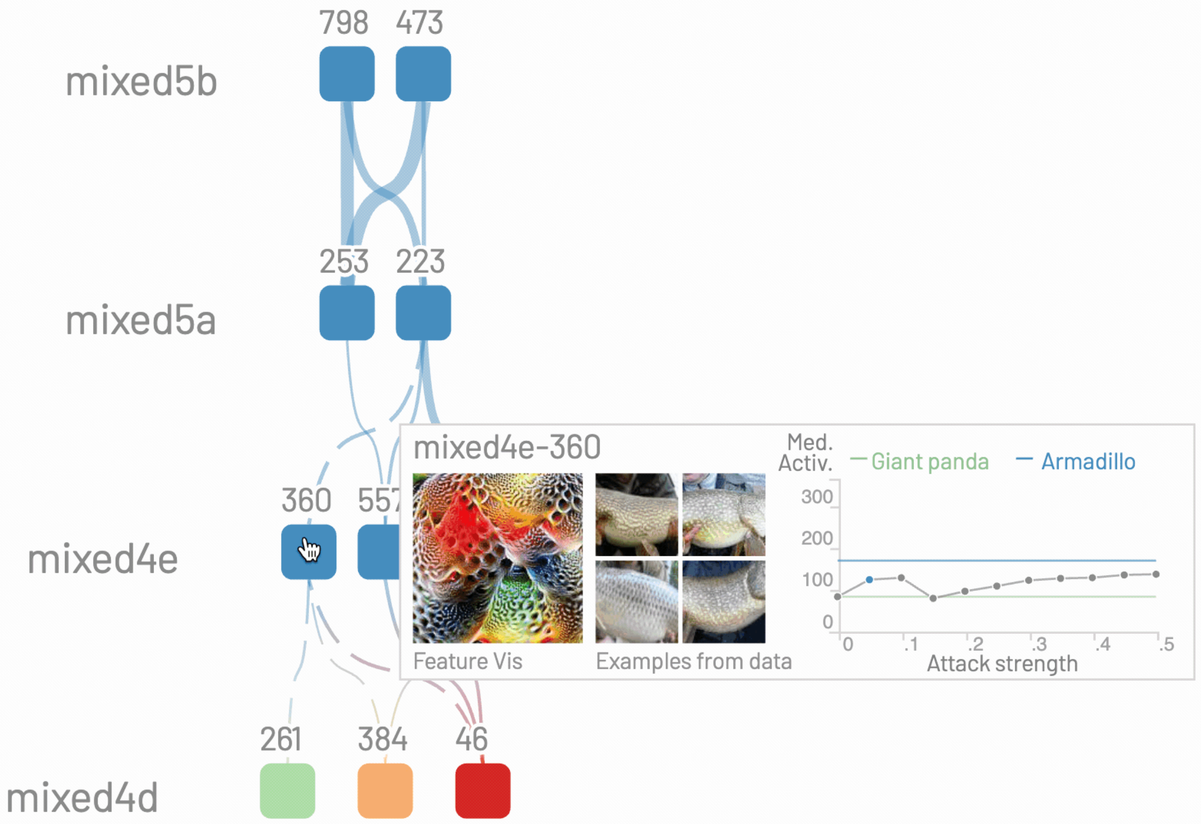
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- N/A
- Stronger
- Weaker
- Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



Bluff

Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

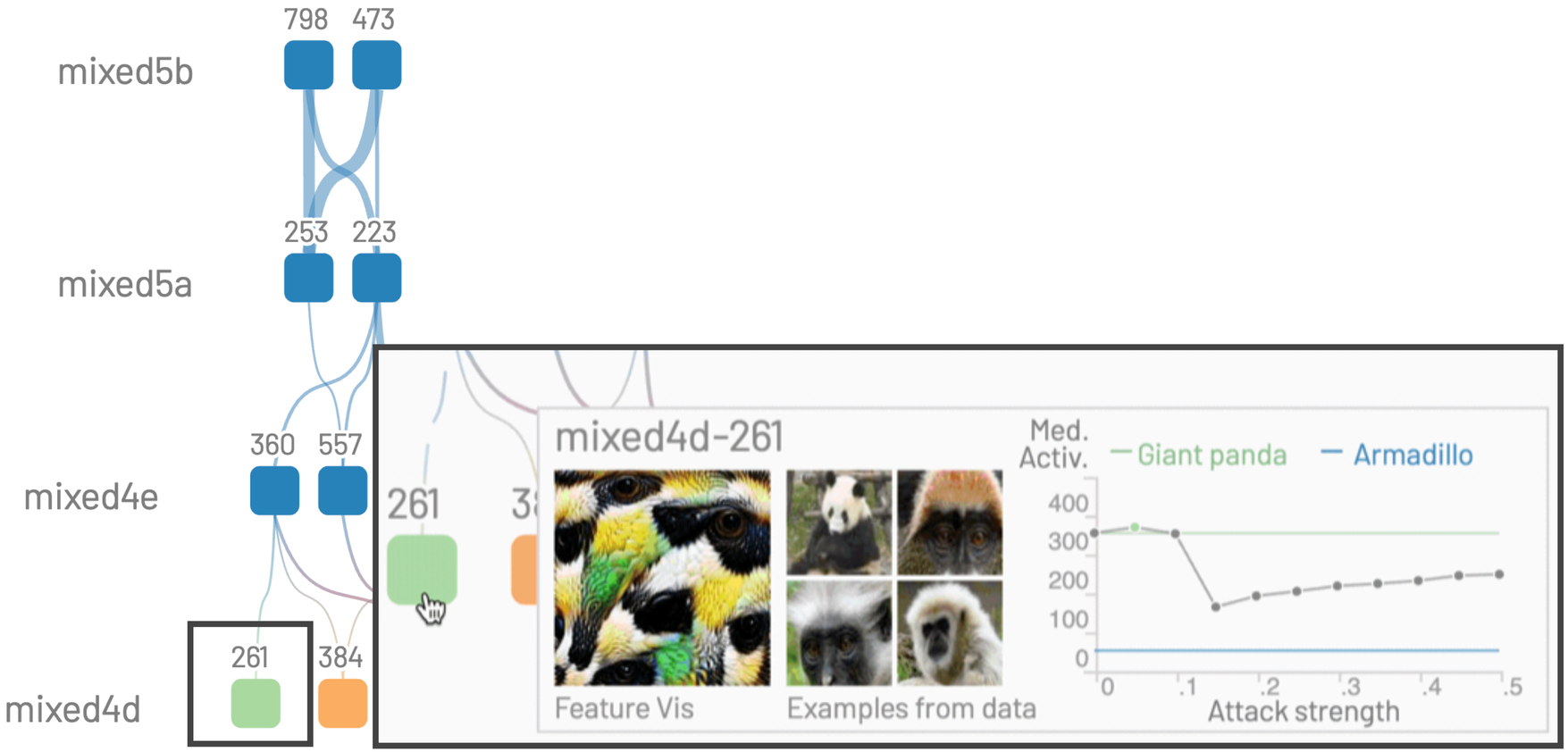
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- [ ] N/A
- [ ] Stronger
- [ ] Weaker
- [ ] Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- Show full graph
- Show pinned only
- Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK

mixed5b 798 473

mixed5a 253 223

mixed4e 360 557 516 261

mixed4d 261 384 46

Feature Visualization

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- Show full graph
- Show pinned only
- Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- N/A
- Stronger
- Weaker
- Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK

Example patches

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- Show full graph
- Show pinned only
- Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK

mixed5b 798 473

mixed5a 253 223

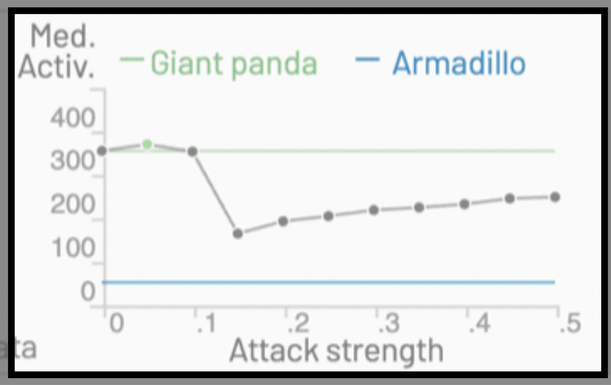
mixed4e 360 557 516 261

mixed4d 261 384 46

mixed4d-261

Feature Vis Examples from data

Median Activation across attack strength



Bluff

Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

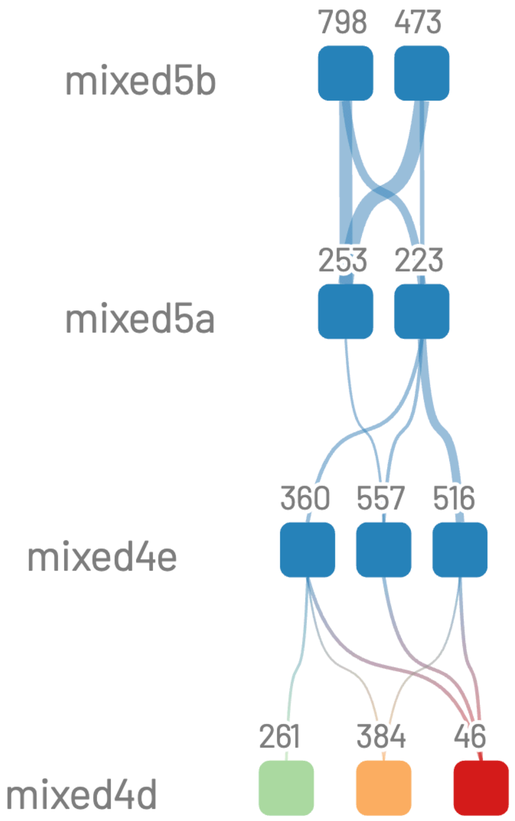
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- N/A
- Stronger
- Weaker
- Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- Show full graph
- Show pinned only
- Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

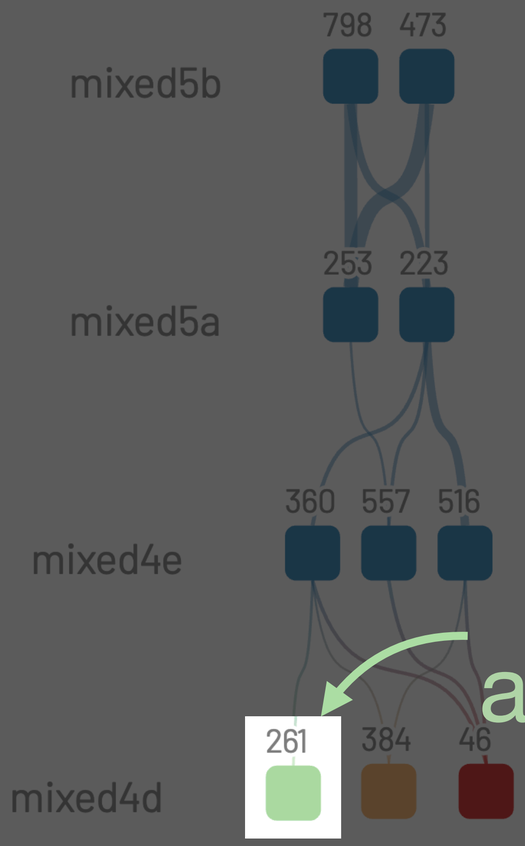
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- N/A
- Stronger
- Weaker
- Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



highly activated by

(**Original** Class)

Benign Giant Panda Images

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- Show full graph
- Show pinned only
- Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

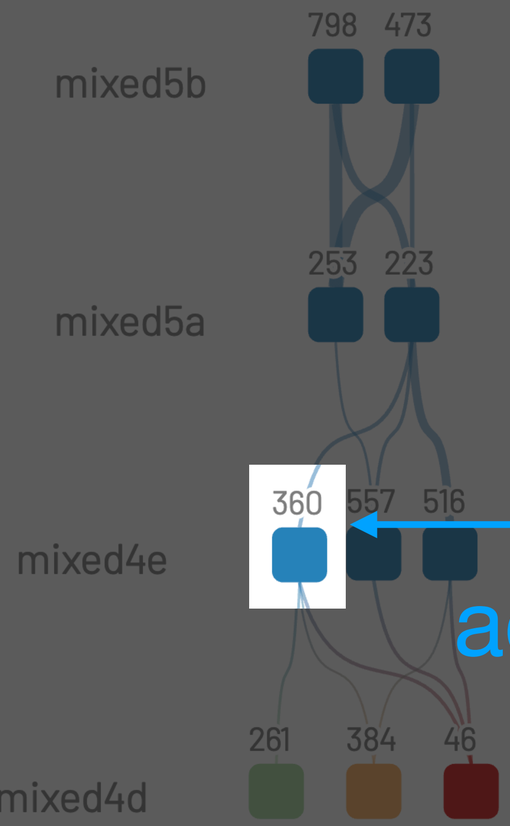
Weaker (inner): 0.05

Only show edges most excited by weaker attack

N/A Stronger

Weaker Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



highly activated by

(**Target** Class)

Benign Armadillo Images

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- Show full graph
- Show pinned only
- Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- N/A
- Stronger
- Weaker
- Weaker + Stronger

GIANT PANDA BOTH ARMADILLO

(**Original** Class)

(**Target** Class)

Benign Giant Panda Images

Benign Armadillo Images

mixed5b 798 473

mixed5a 253 223

mixed4e 360 557 516

mixed4d 261 384 46

highly activated by **Both** class images

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

Stronger (outer): 0.45

Weaker (inner): 0.05

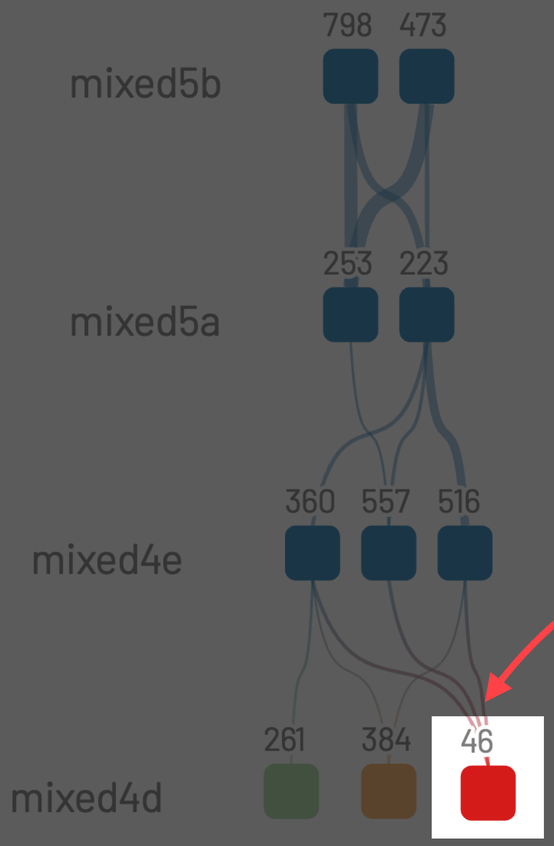
Only show edges most excited by weaker attack

N/A Stronger

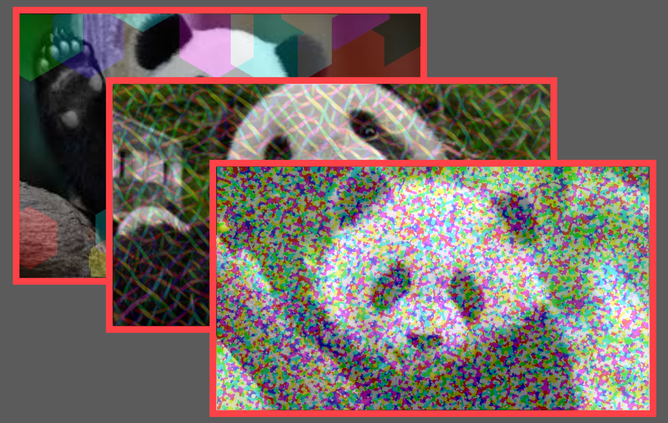
Weaker Weaker + Stronger

GIANT PANDA BOTH ARMADILLO

EXPLOITED BY ATTACK



(**Attacked** images)



highly activated by Adversarial Giant Panda Images

Bluff

Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

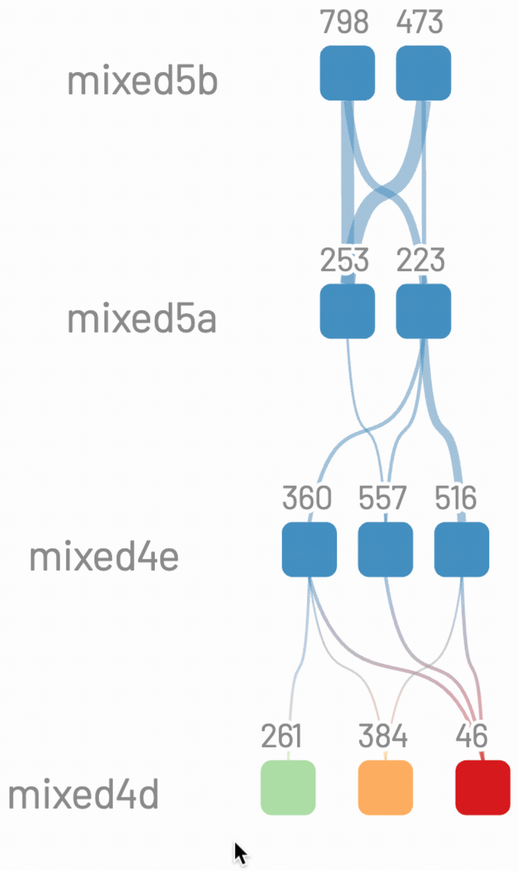
Stronger (outer): 0.45

Weaker (inner): 0.05

Only show edges most excited by weaker attack

- [ ] N/A
- [ ] Stronger
- [ ] Weaker
- [ ] Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK



Bluff

Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK

PGD

Strength: 0.05

FILTER GRAPH

- [ ] Show full graph
- [x] Show pinned only
- [ ] Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

COMPARE ATTACKS

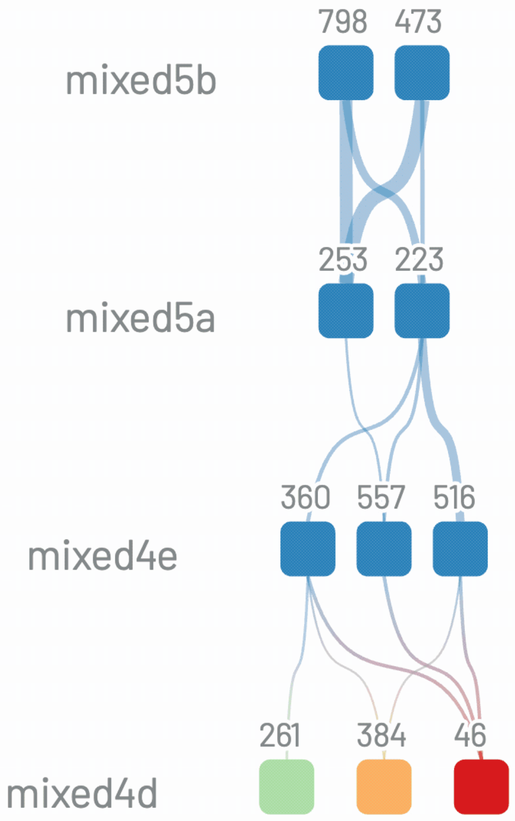
Stronger (outer): 0.45

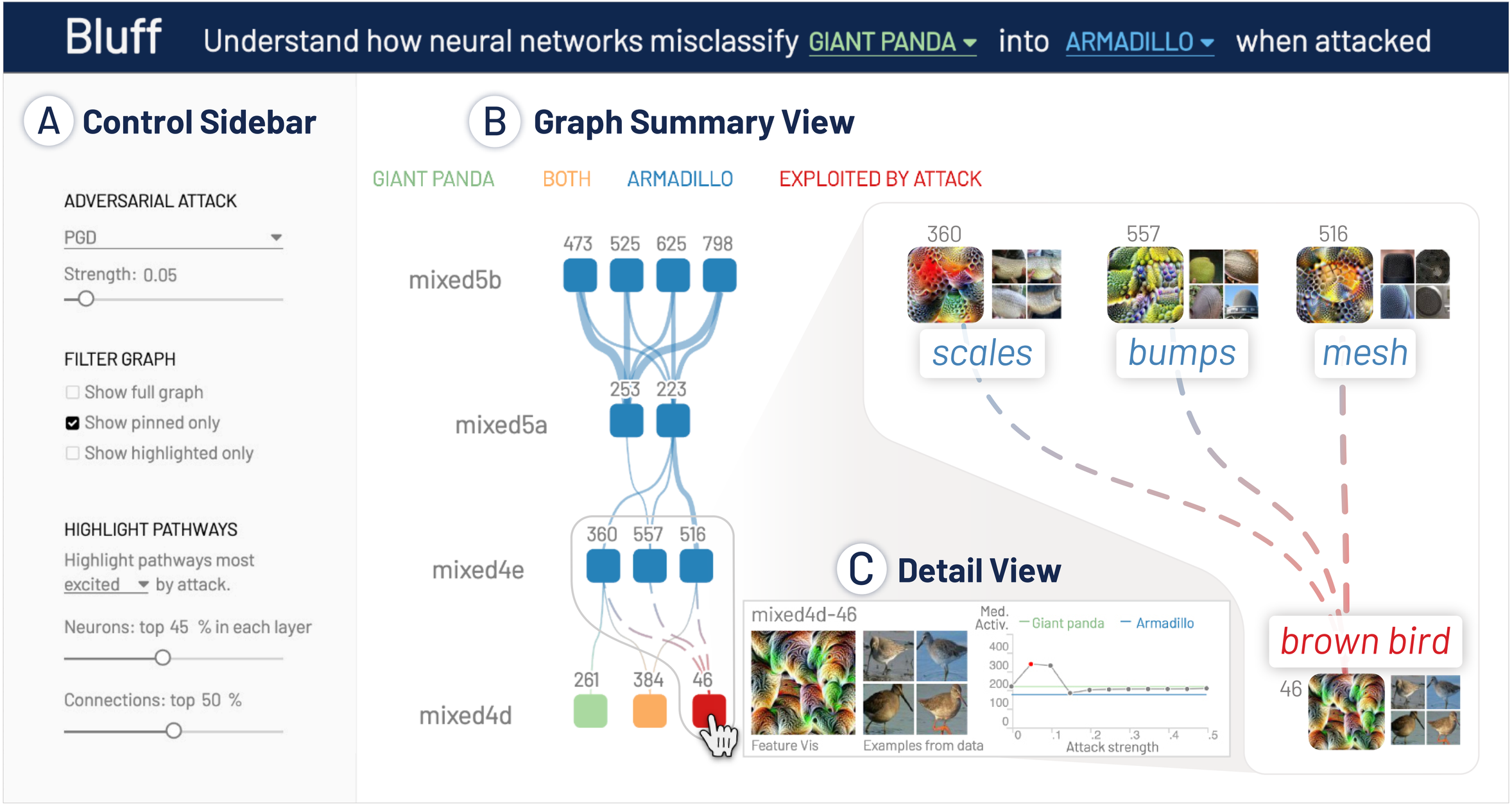
Weaker (inner): 0.05

Only show edges most excited by weaker attack

- [ ] N/A
- [ ] Stronger
- [ ] Weaker
- [ ] Weaker + Stronger

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK




Bluff
Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked
A Control Sidebar
ADVERSARIAL ATTACK
PGD
Strength: 0.05
FILTER GRAPH
Show full graph
Show pinned only
Show highlighted only
HIGHLIGHT PATHWAYS
Highlight pathways most excited by attack.
Neurons: top 45 % in each layer
Connections: top 50 %
B Graph Summary View
GIANT PANDA
BOTH
ARMADILLO
EXPLOITED BY ATTACK
mixed5b
473 525 625 798
mixed5a
253 223
mixed4e
360 557 516
mixed4d
261 384 46
360
557
516
scales
bumps
mesh
brown bird
46
C Detail View
mixed4d-46
Feature Vis
Examples from data
Med. Activ.
Giant panda
Armadillo
400
300
200
100
0
0 .1 .2 .3 .4 .5
Attack strength

Part I

Understand AI Vulnerabilities

GOGGLES SIGMOD 2020

Bluff IEEE VIS 2020

Part II

Fortify AI Security

SHIELD KDD 2018

SkeleVision arXiv 2022 (under review)

Hear No Evil arXiv 2022 (under review)

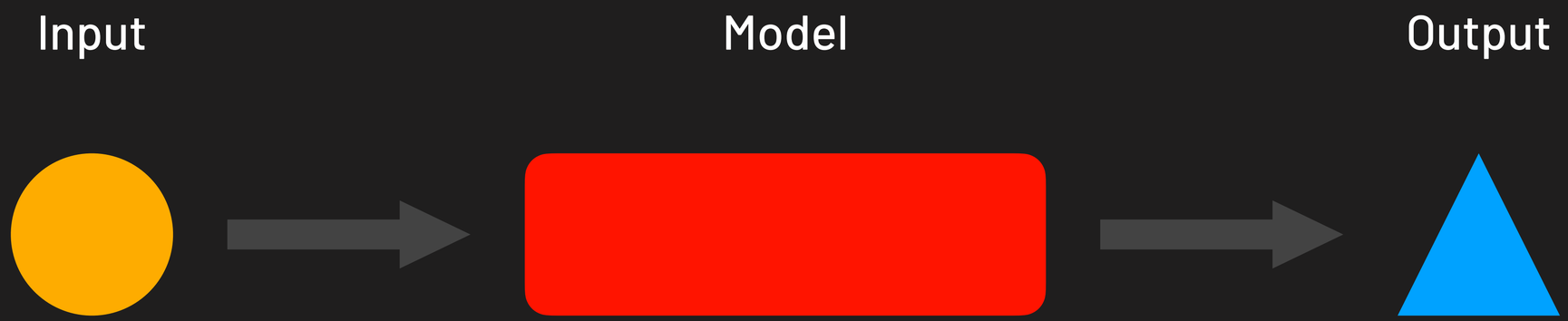
Part III

Enable Use of AI Security

ADAGIO ECML-PKDD 2018

MLsploit KDD Showcase 2019

Machine Learning Pipeline



Automatic Speech Recognition Machine Learning Pipeline

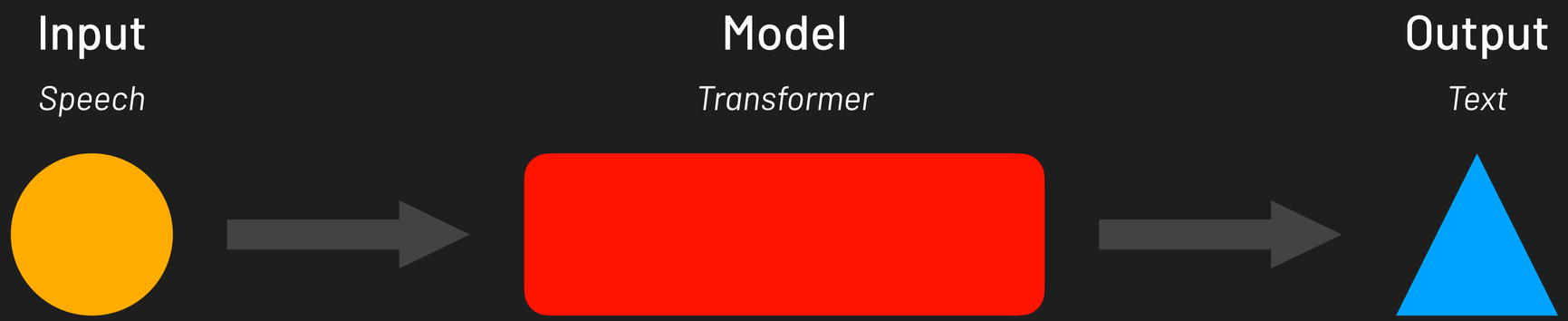
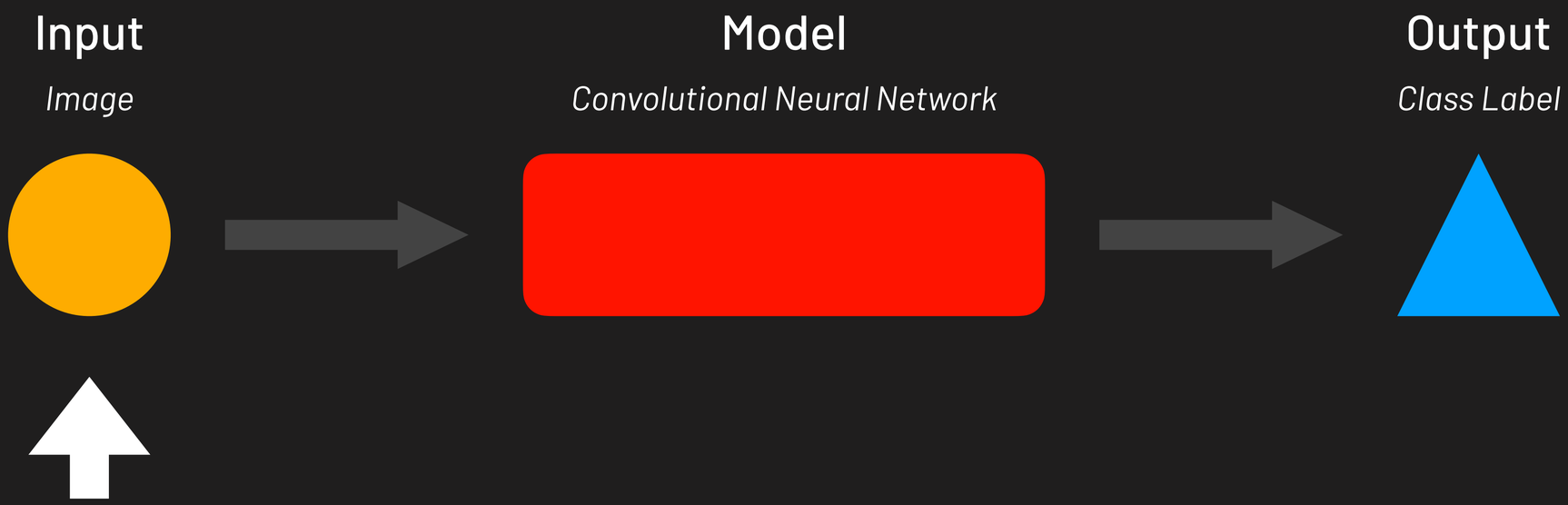


Image Classification
Machine Learning Pipeline



SHIELD

Fast, Practical Defense for Image Classification

KDD 2018
KDD 2019 LEMINCS

- KDD'18 Audience Appreciation Award (runner-up)
- Tech-transferred to Intel Labs
- Open-sourced at *github.com/poloclub/jpeg-defense*

Nilaksh Das
Georgia Tech

Madhuri Shanbhogue
Georgia Tech

Shang-Tse Chen
Georgia Tech

Fred Hohman
Georgia Tech

Siwei Li
Georgia Tech

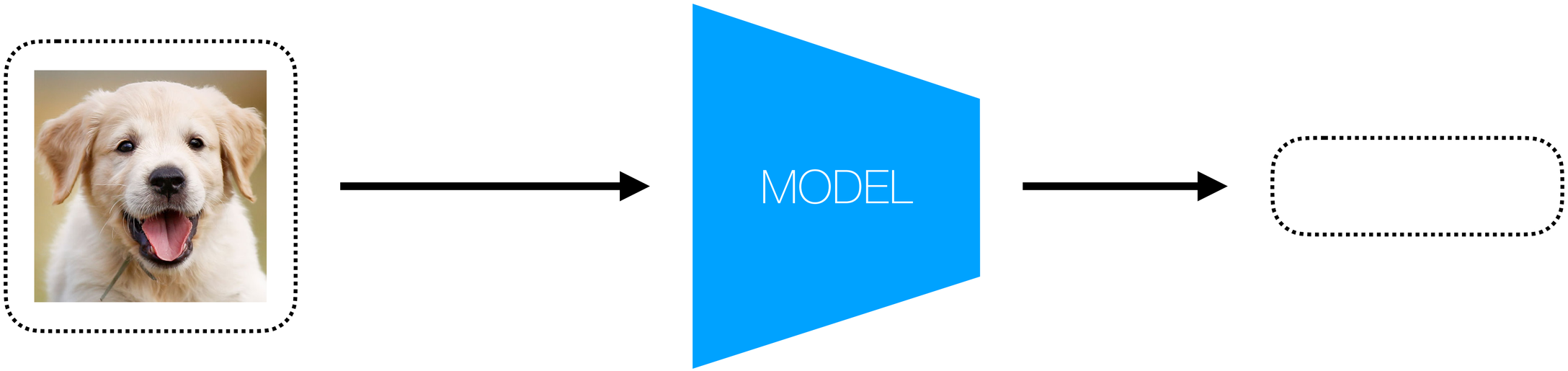
Cory Cornelius
Intel Labs

Li Chen
Intel Labs

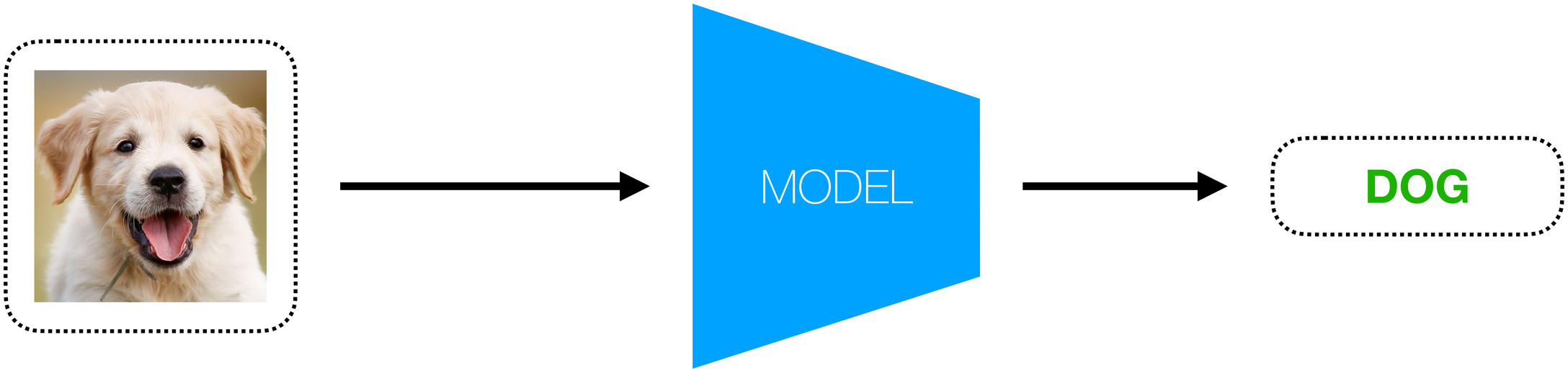
Michael Kounavis
Intel Labs

Polo Chau
Georgia Tech

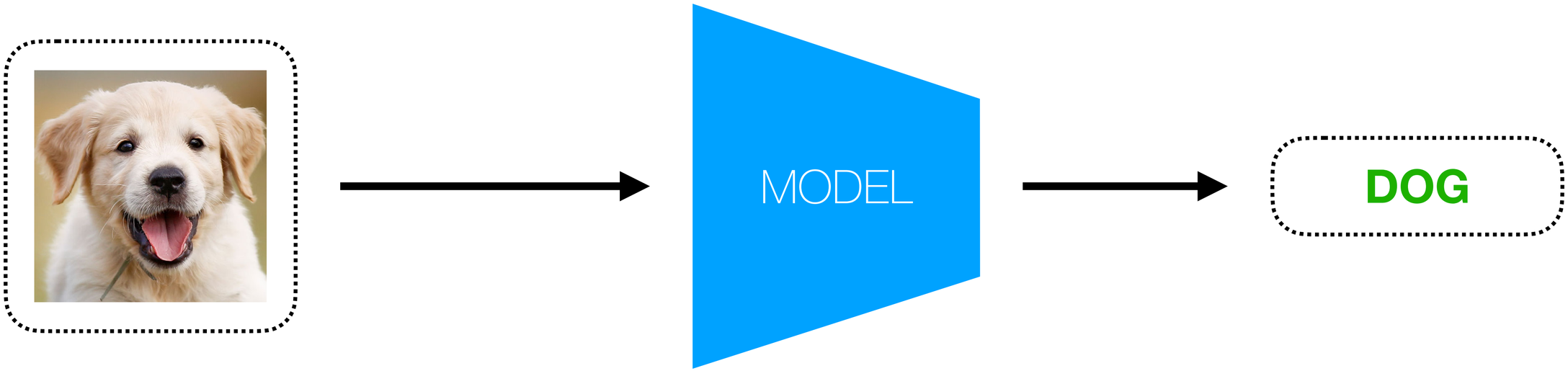
Deep Learning for Image Classification



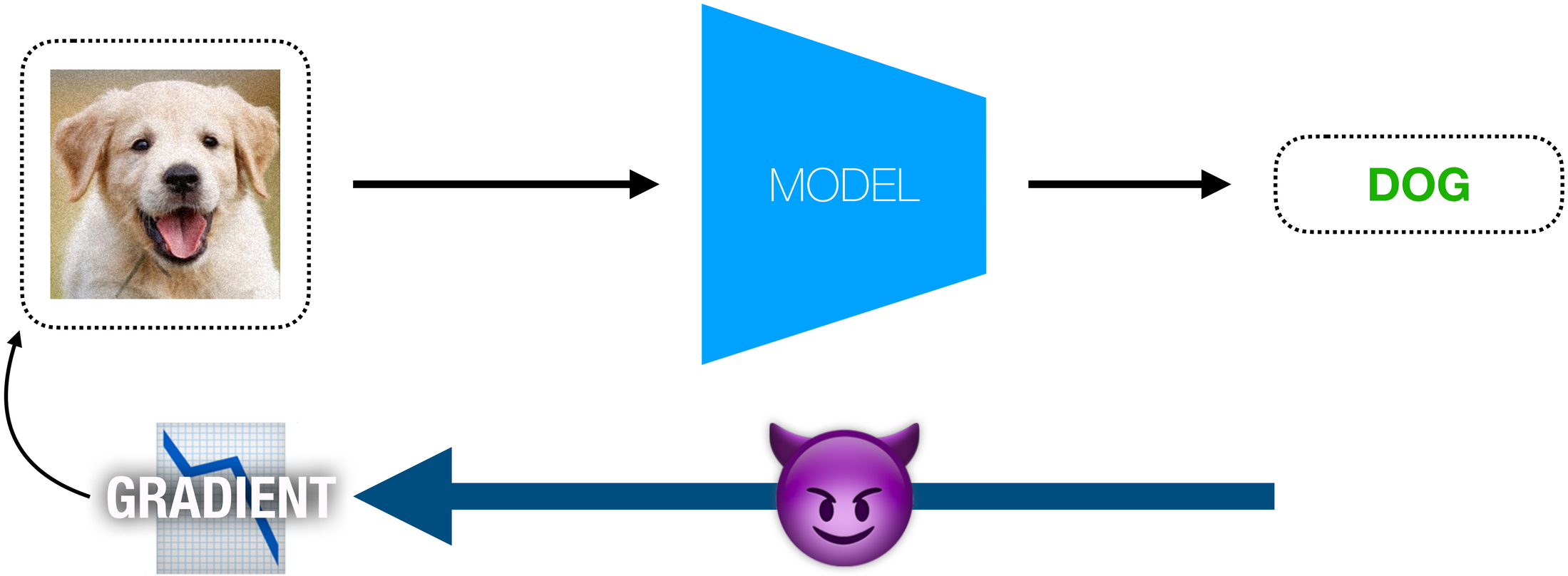
Deep Learning for Image Classification



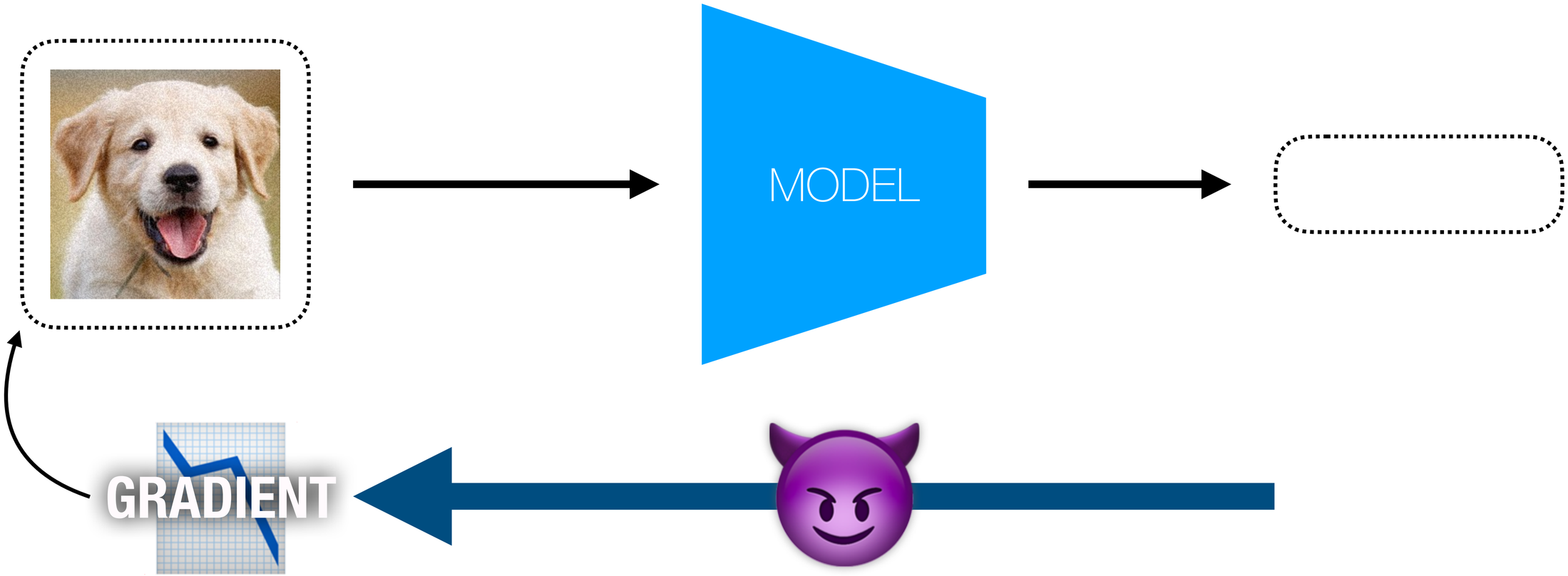
Adversarial Attack on Deep Learning



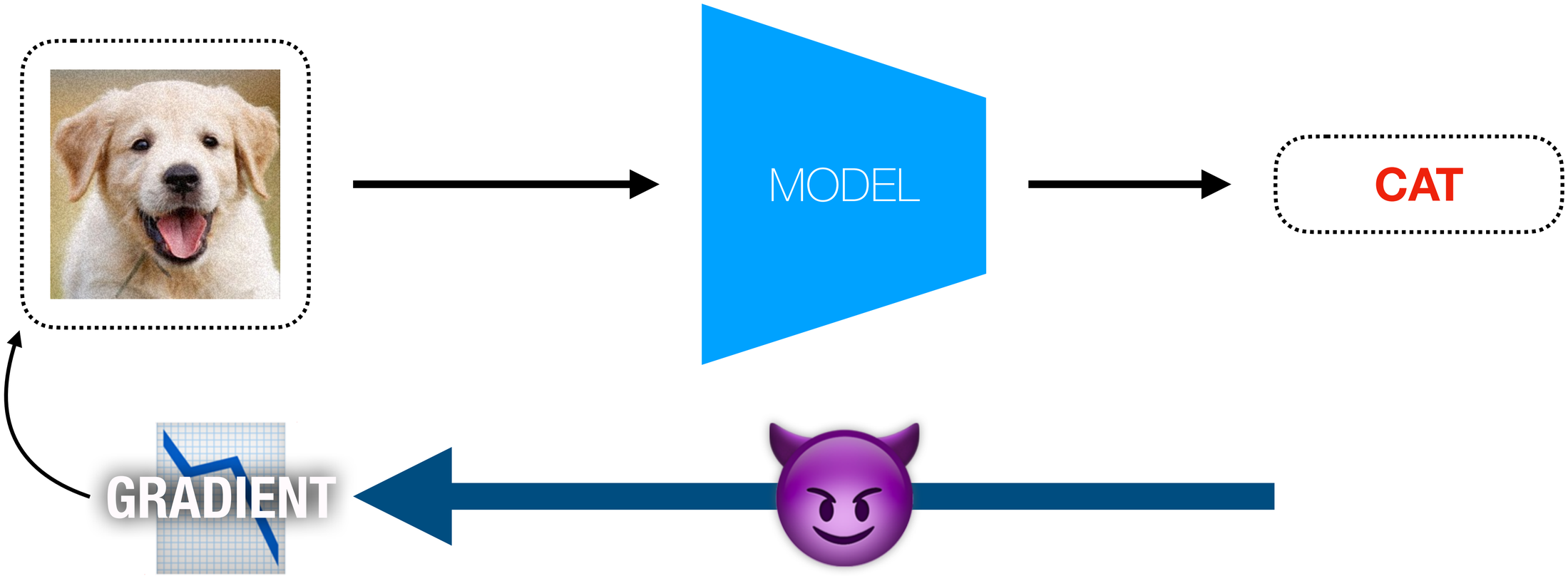
Adversarial Attack on Deep Learning



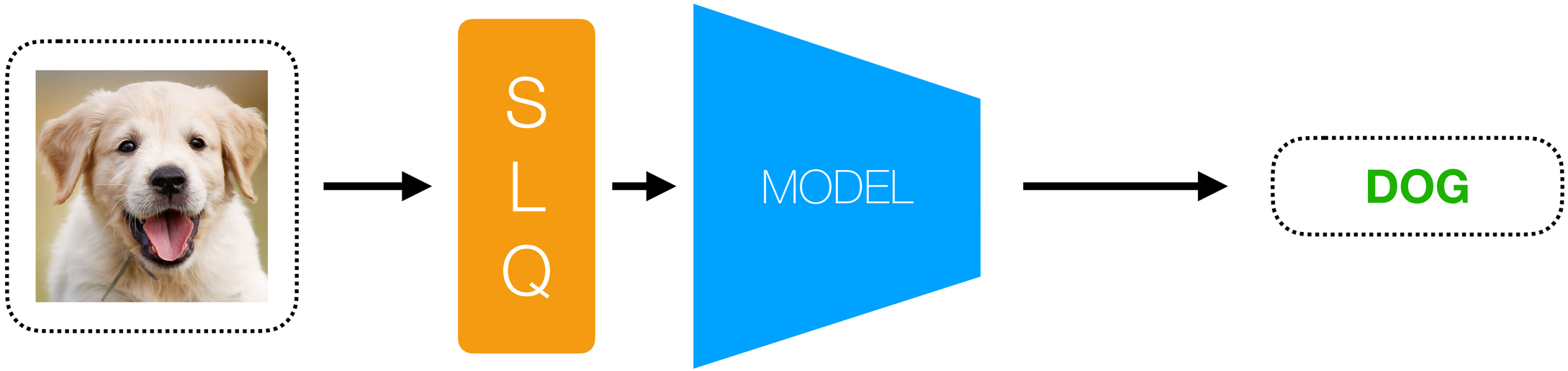
Adversarial Attack on Deep Learning



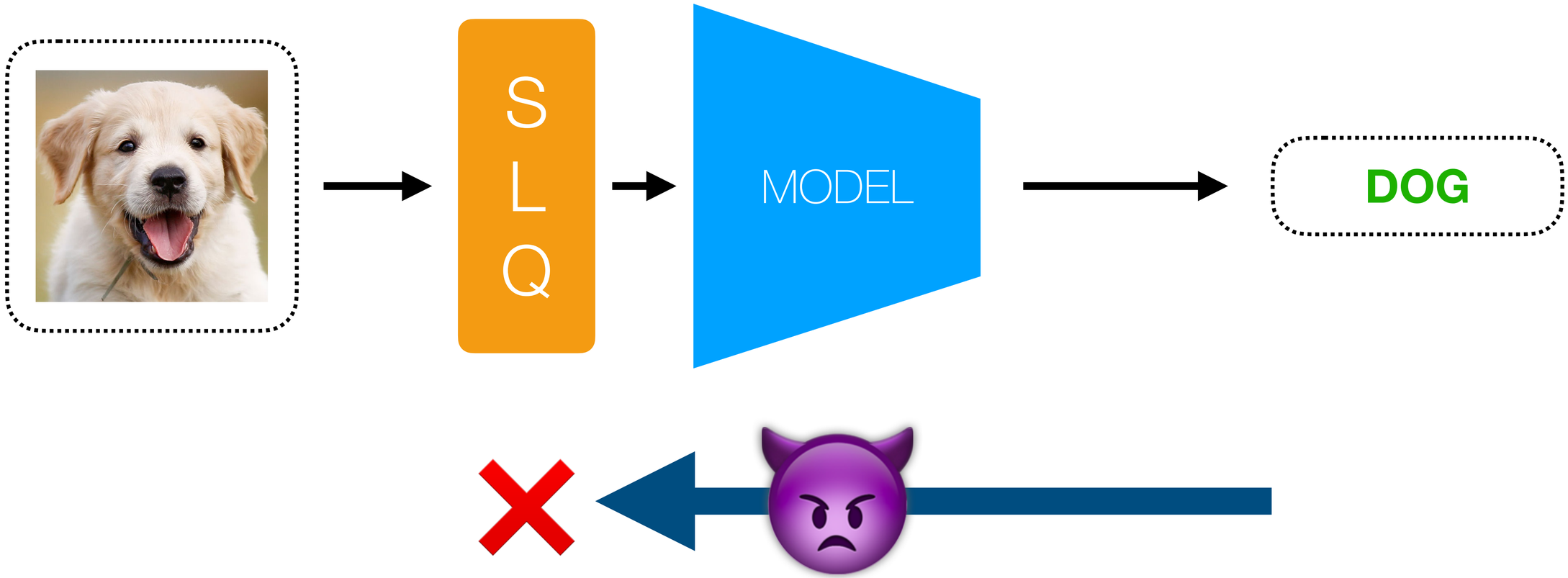
Adversarial Attack on Deep Learning



Stochastic Local Quantization (SLQ)



Stochastic Local Quantization (SLQ)



SLQ leverages JPEG compression

SLQ leverages JPEG compression

JPEG Quality 80

JPEG Quality 60

JPEG Quality 40

JPEG Quality 20

SLQ leverages JPEG compression

JPEG Quality 80

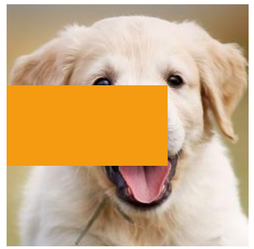
JPEG Quality 60

JPEG Quality 40

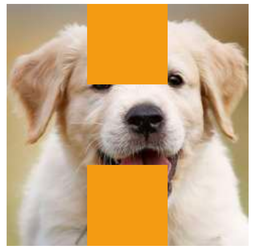
JPEG Quality 20

SLQ leverages JPEG compression with ***randomization***

JPEQ Quality 80

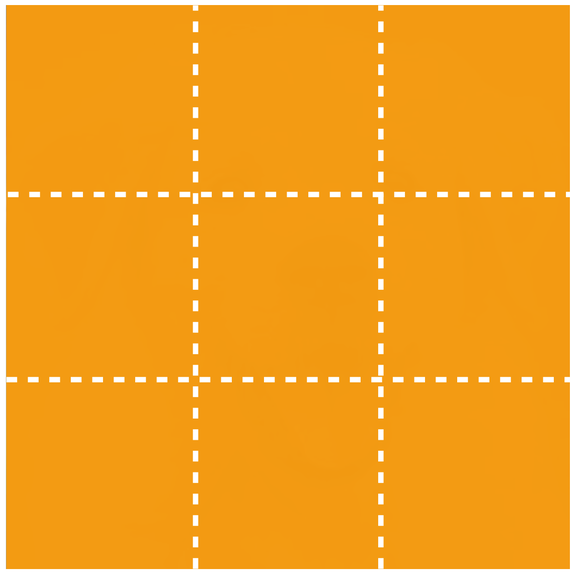


JPEQ Quality 60



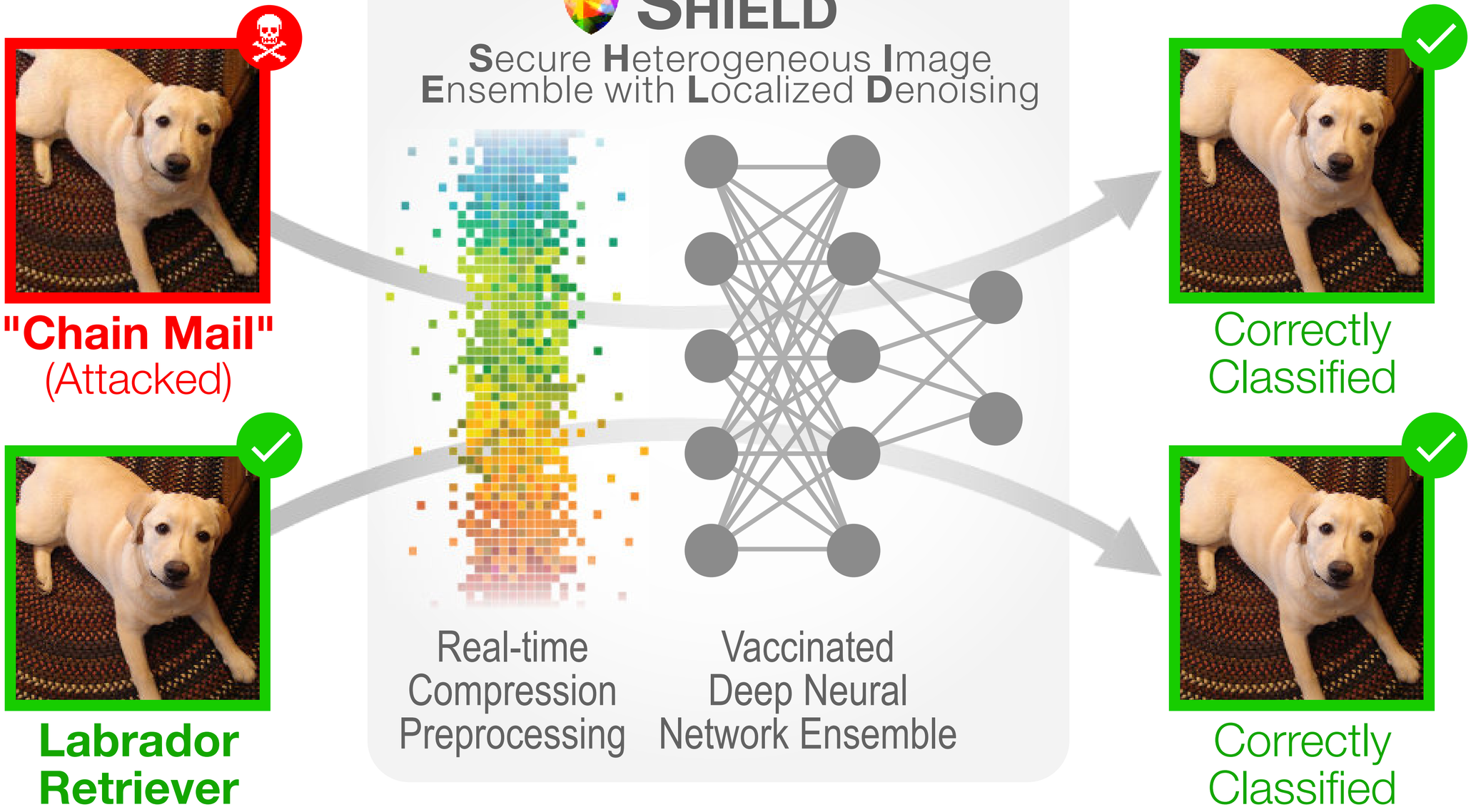
JPEQ Quality 40

JPEQ Quality 20

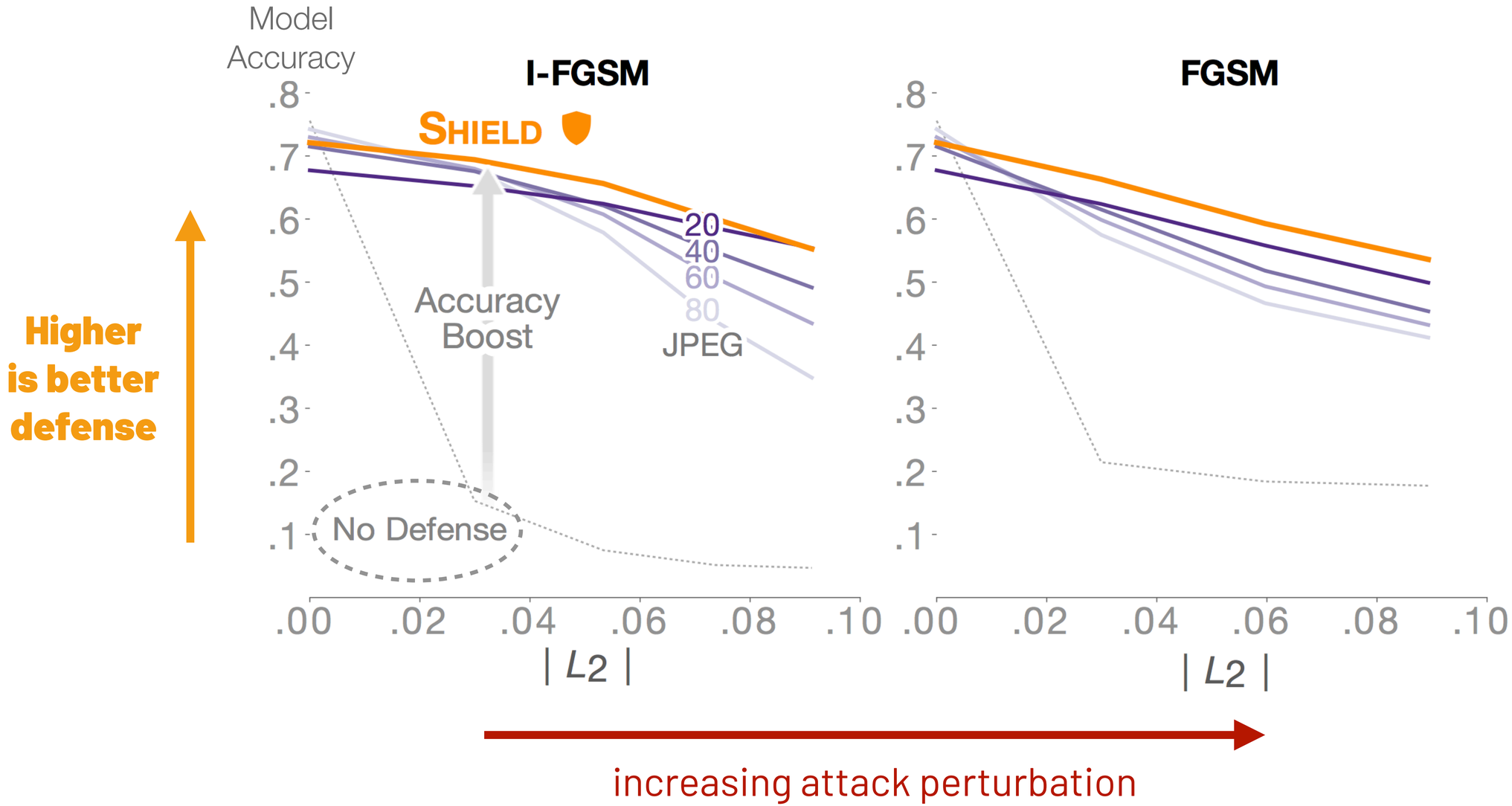


SLQ applies JPEG compression of a random quality to each 8 x 8 block of the image

* larger blocks shown for presentation

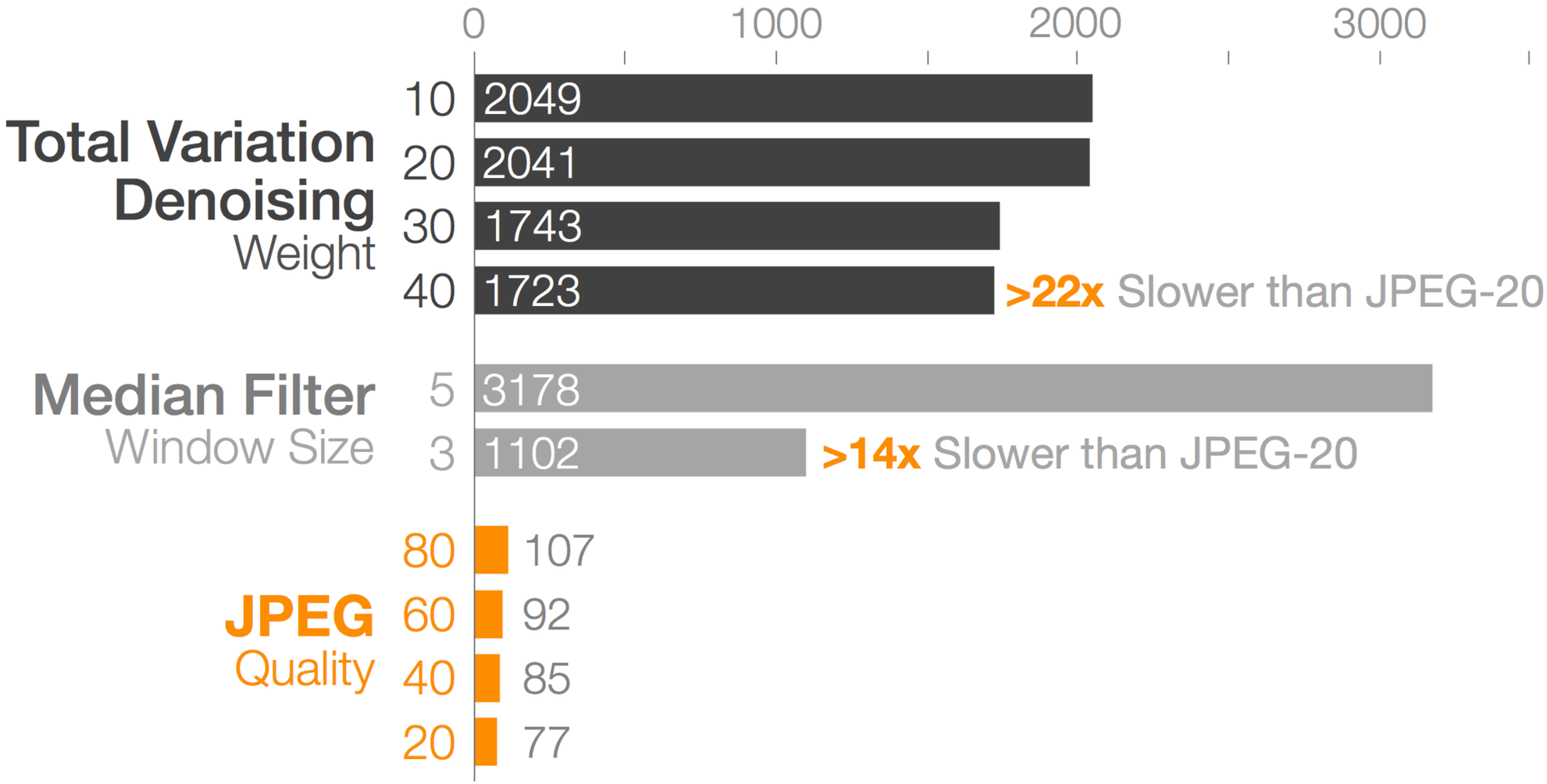

SHIELD
Secure Heterogeneous Image Ensemble with Localized Denoising
"Chain Mail"
(Attacked)
Labrador Retriever
Real-time Compression Preprocessing
Vaccinated Deep Neural Network Ensemble
Correctly Classified
Correctly Classified

Results with ResNet-50 v2 (on ImageNet validation set)



Defense Runtime Comparison

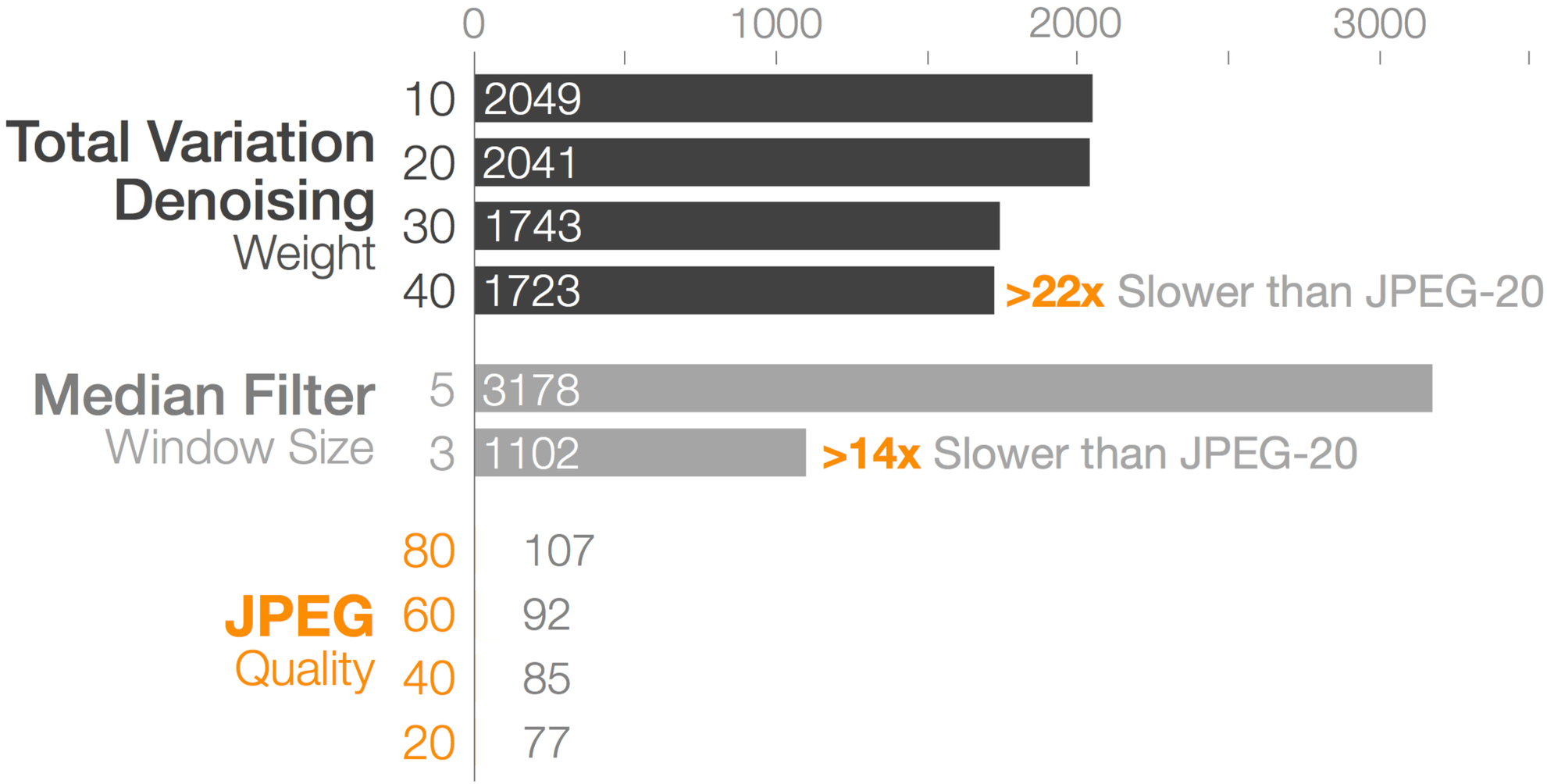
(in seconds; shorter is better)



tested on 50,000 images from the ImageNet validation set

Defense Runtime Comparison

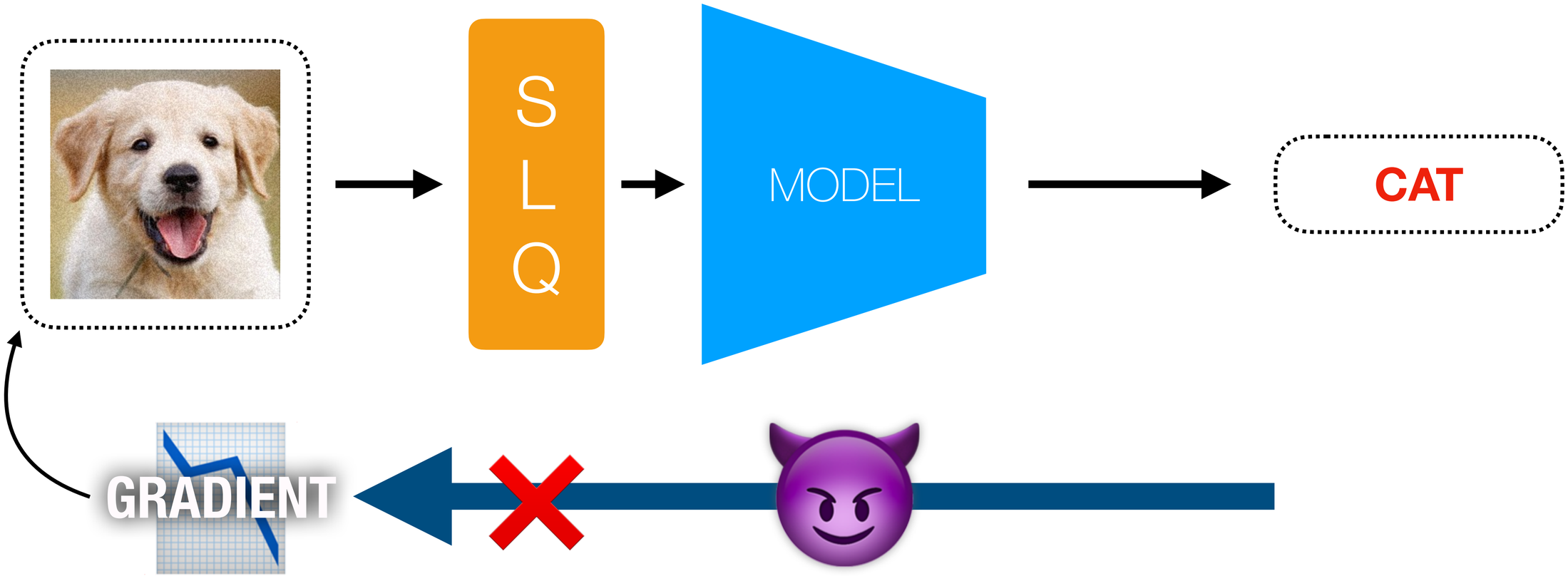
(in seconds; shorter is better)



tested on 50,000 images from the ImageNet validation set

Adaptive Attacks

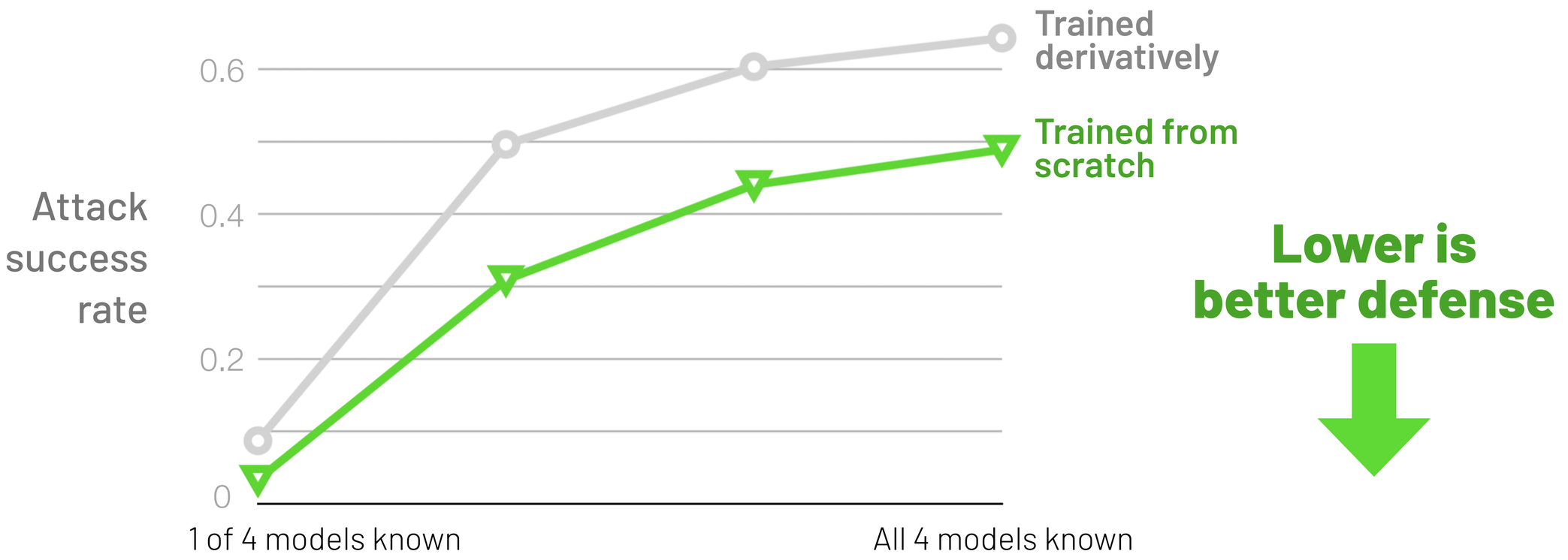
- Backward Pass Differentiable Approximation (BPDA)
- Expectation over Transformations (EoT)



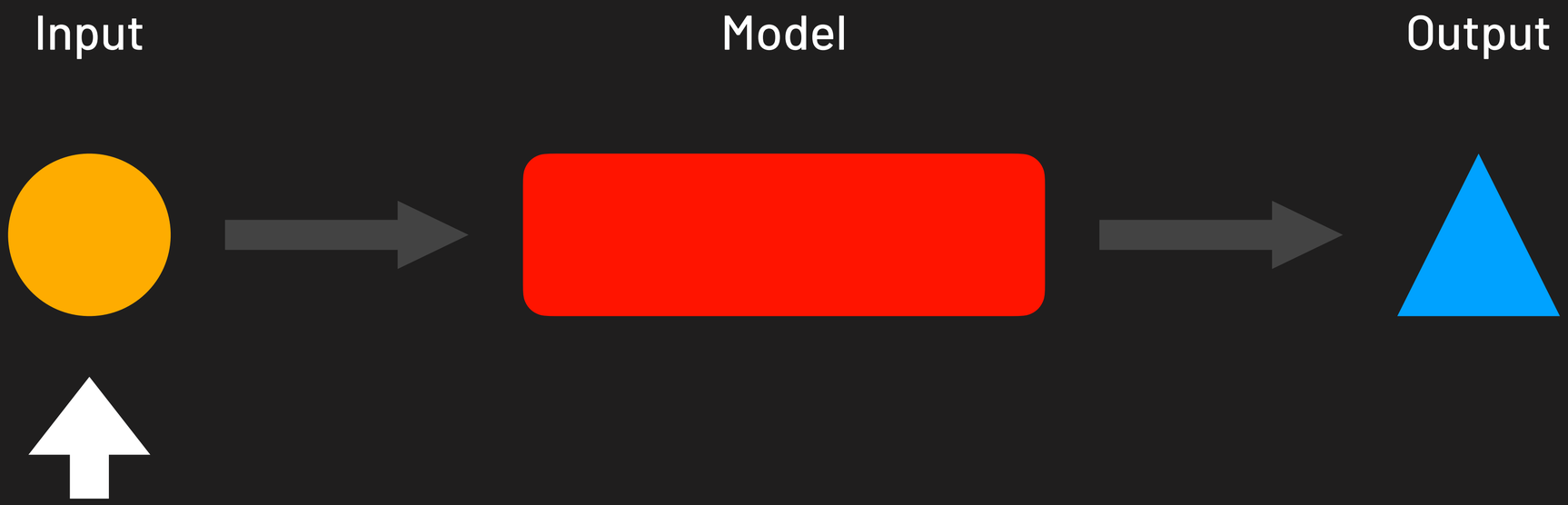
Extending to Adaptive Attacks

SHIELD ensemble with ***less correlated*** model weights are ***more robust*** to targeted **adaptive attacks**

[KDD'19 LEMINCS]



Machine Learning Pipeline



SkeleVision

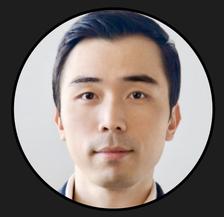
Adversarial Resiliency of Person Tracking with Multi-Task Learning

arXiv 2022
(under review)

- Robust inference in ***video*** domain
- Open-sourced at *github.com/nilakshdas/SkeleVision*

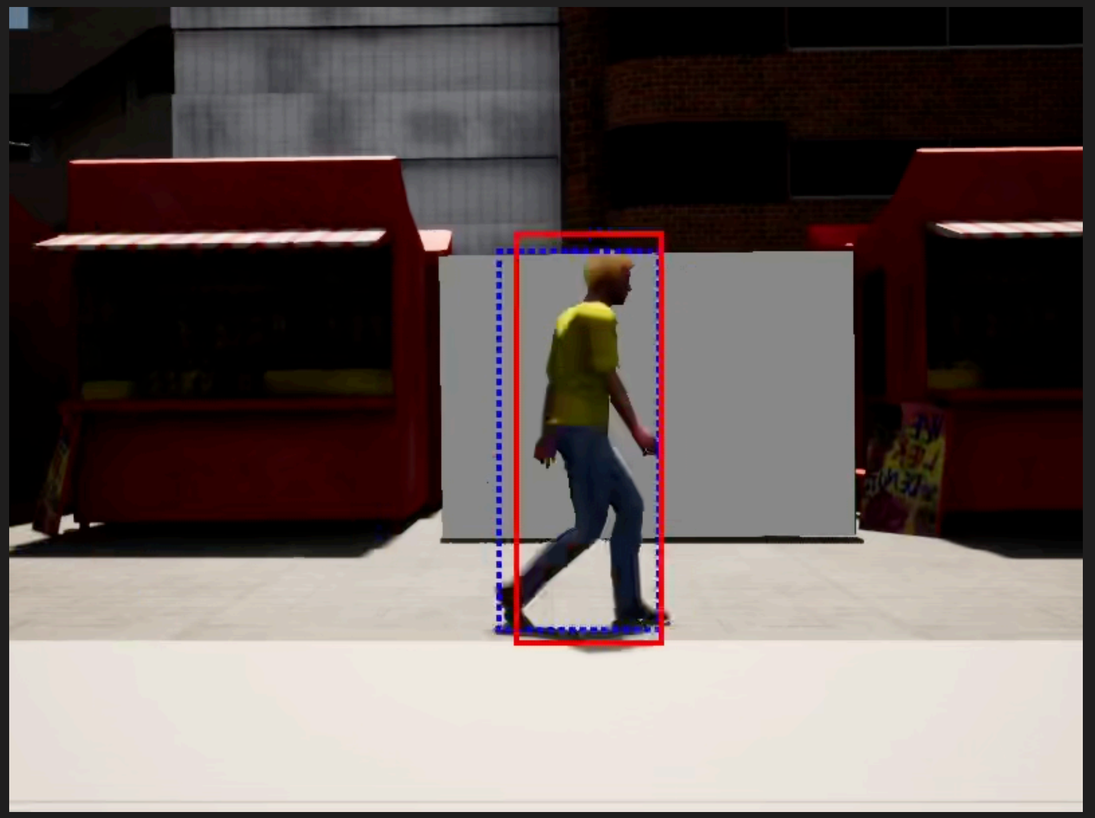
Nilaksh Das

Sheng-Yun Peng

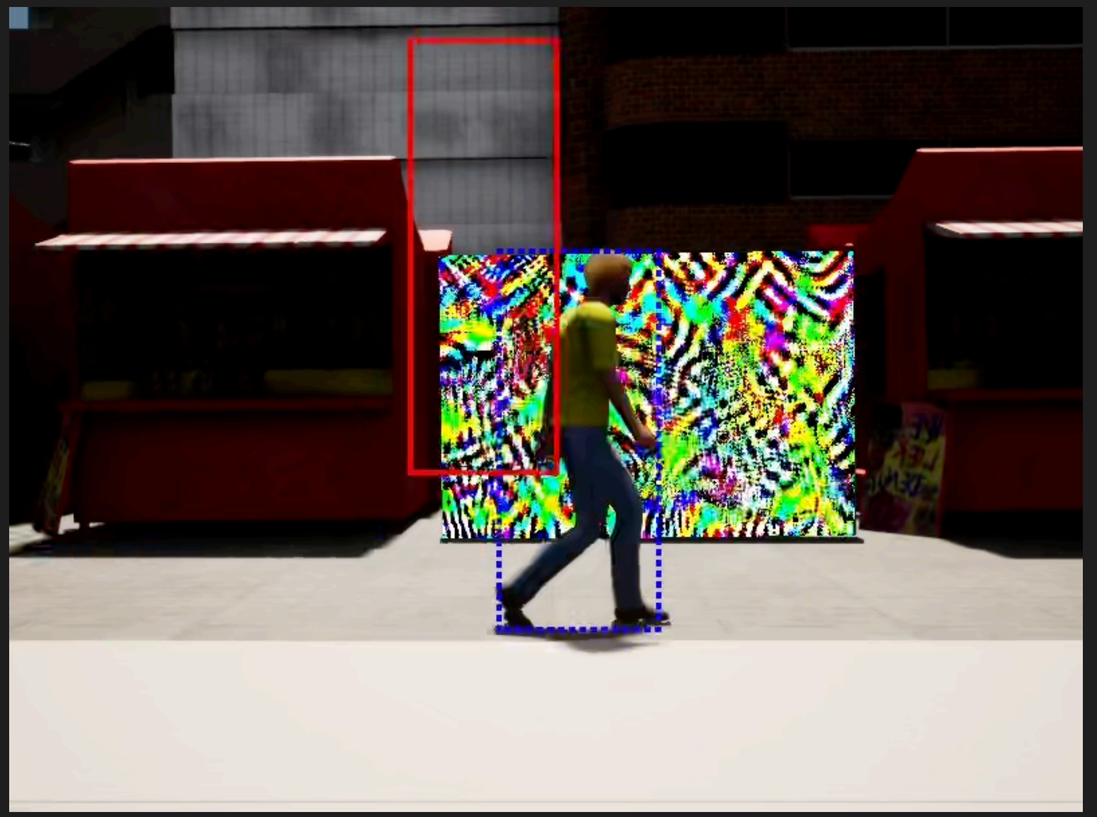


Polo Chau

Person Tracking



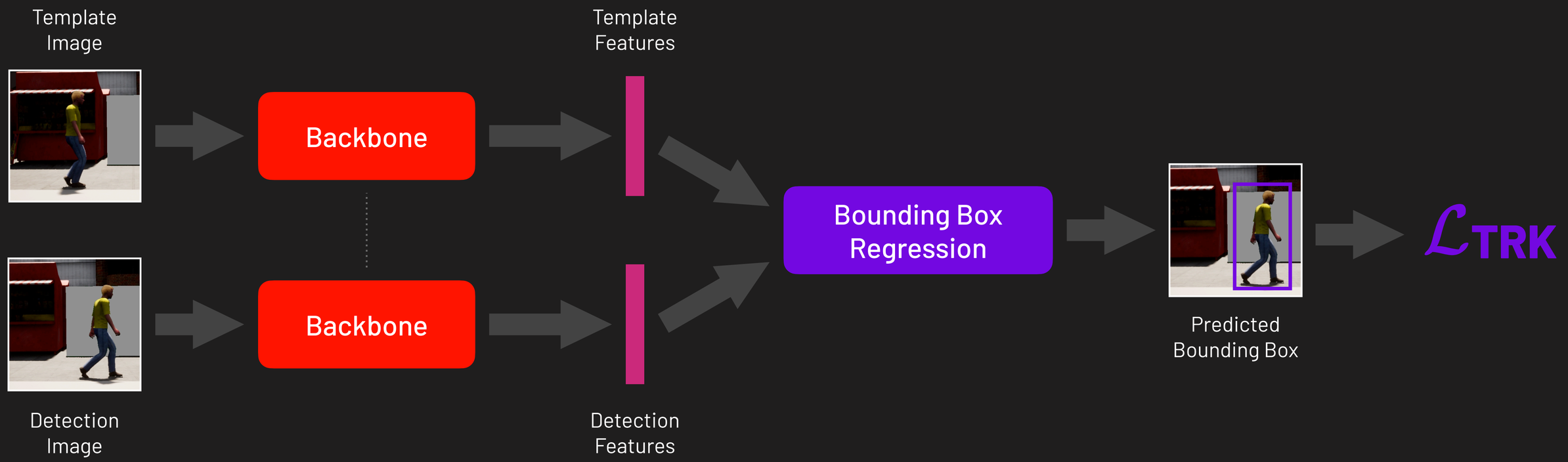
Adversarial Attack on Person Tracking



Ground Truth
Predicted

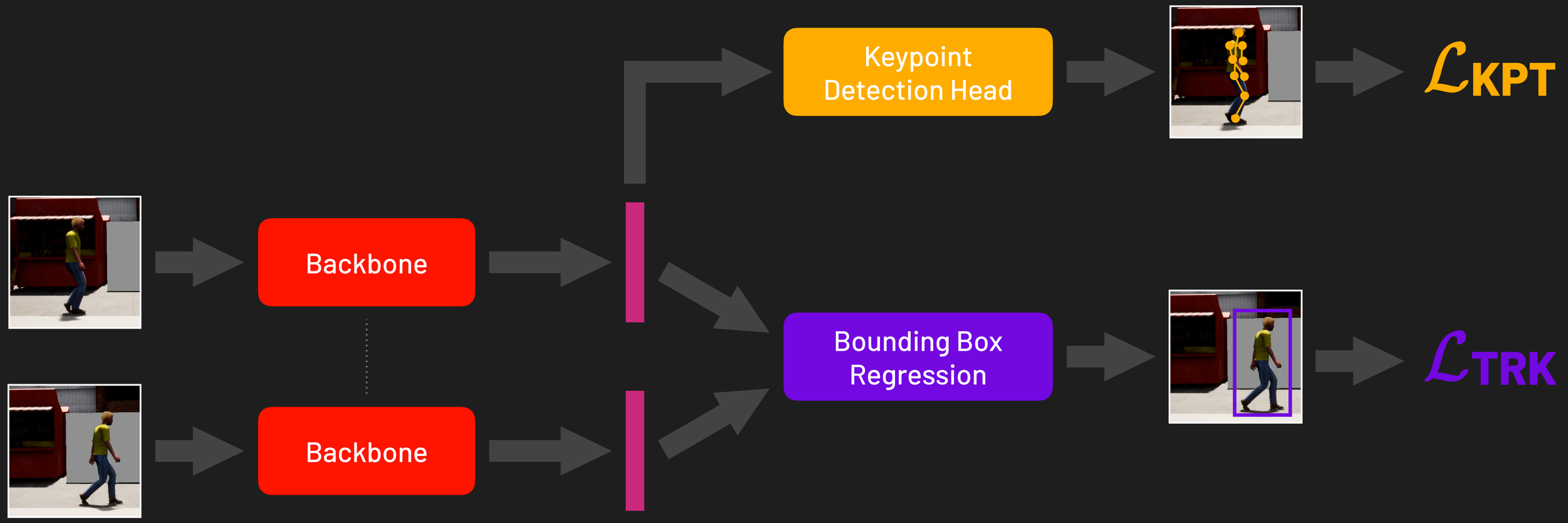
Physically Realizable

Perceptually Unbounded

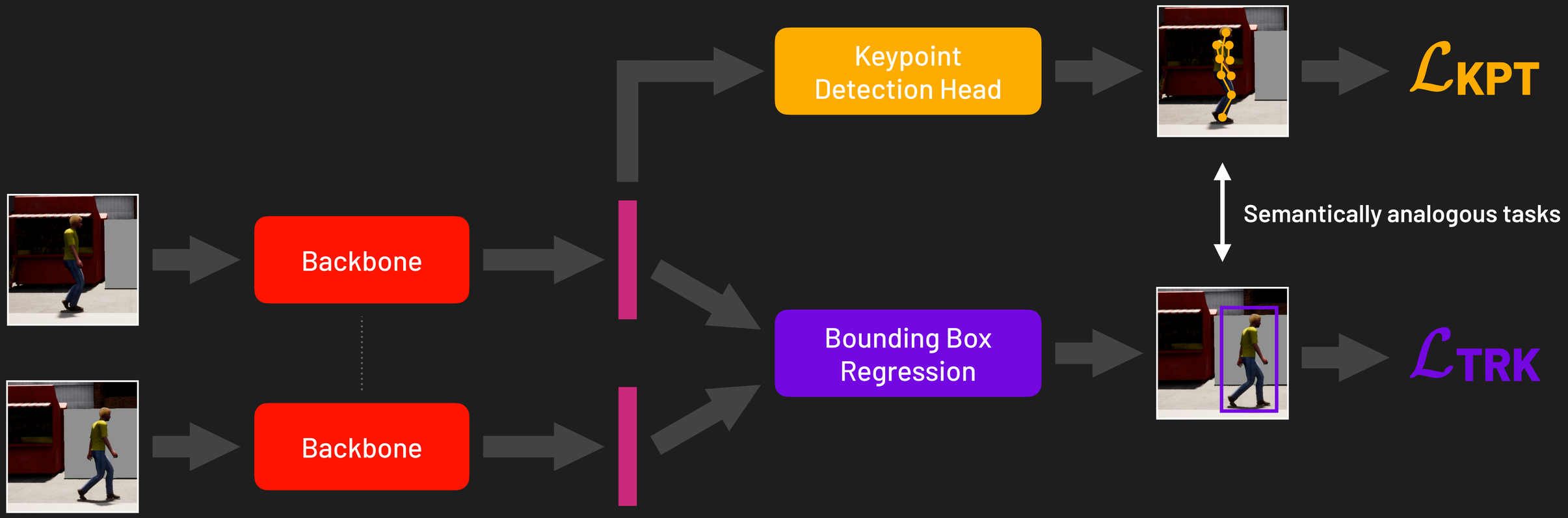


SiamRPN Model

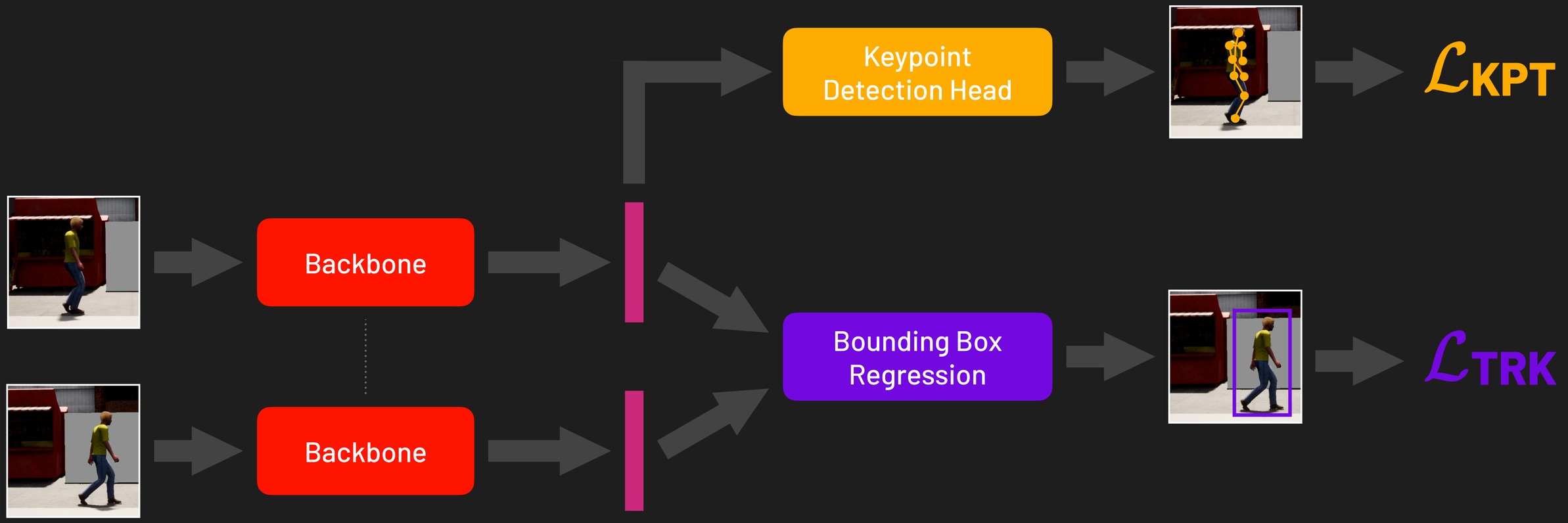
High Performance Visual Tracking with Siamese Region Proposal Network
B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu
CVPR 2018



Multi-Task Learning with Tracking + Keypoint Detection



Multi-Task Learning with Tracking + Keypoint Detection



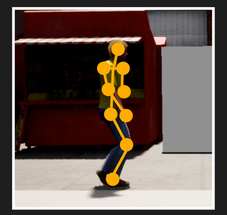
$$\mathcal{L}_{MTL} = \mathcal{L}_{TRK} + \lambda_K \mathcal{L}_{KPT}$$

Backbone

Backbone

Keypoint Detection Head

Bounding Box Regression



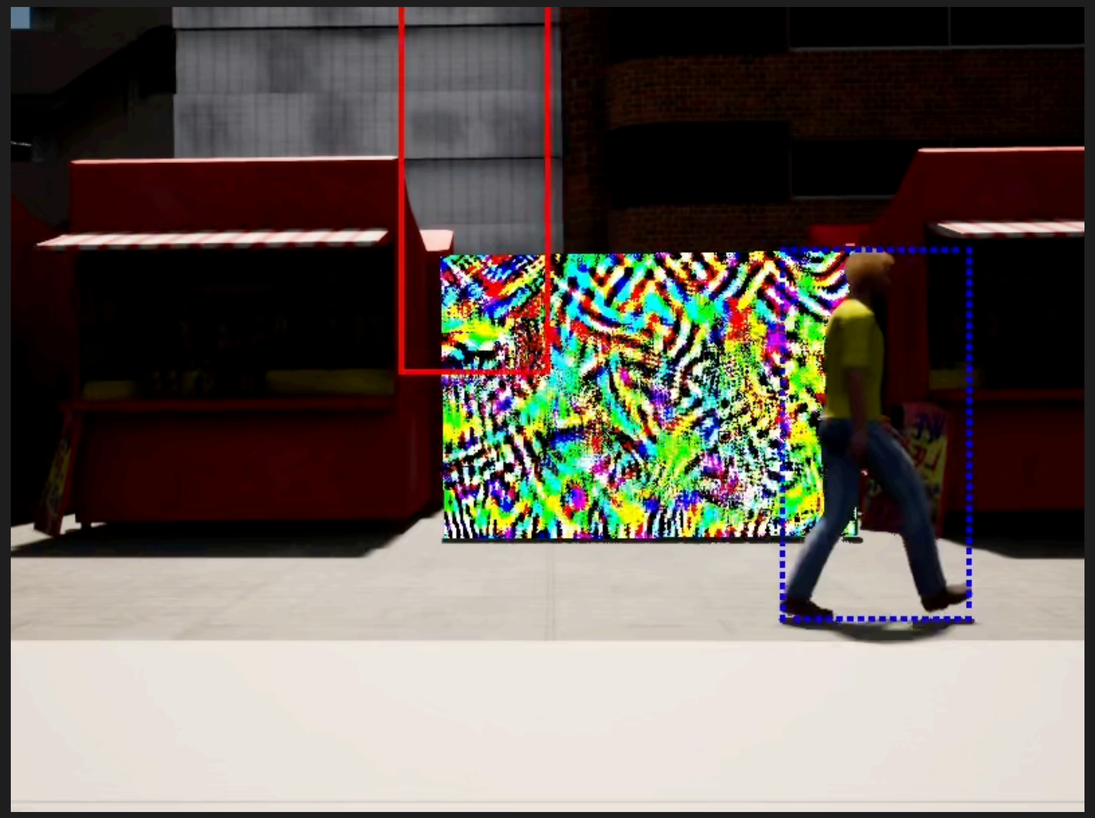
$\mathcal{L}_{KPT}$

$\mathcal{L}_{TRK}$

Induce robust features

$$\mathcal{L}_{MTL} = \mathcal{L}_{TRK} + \lambda_K \mathcal{L}_{KPT}$$

Single-Task Learning (STL) Tracker

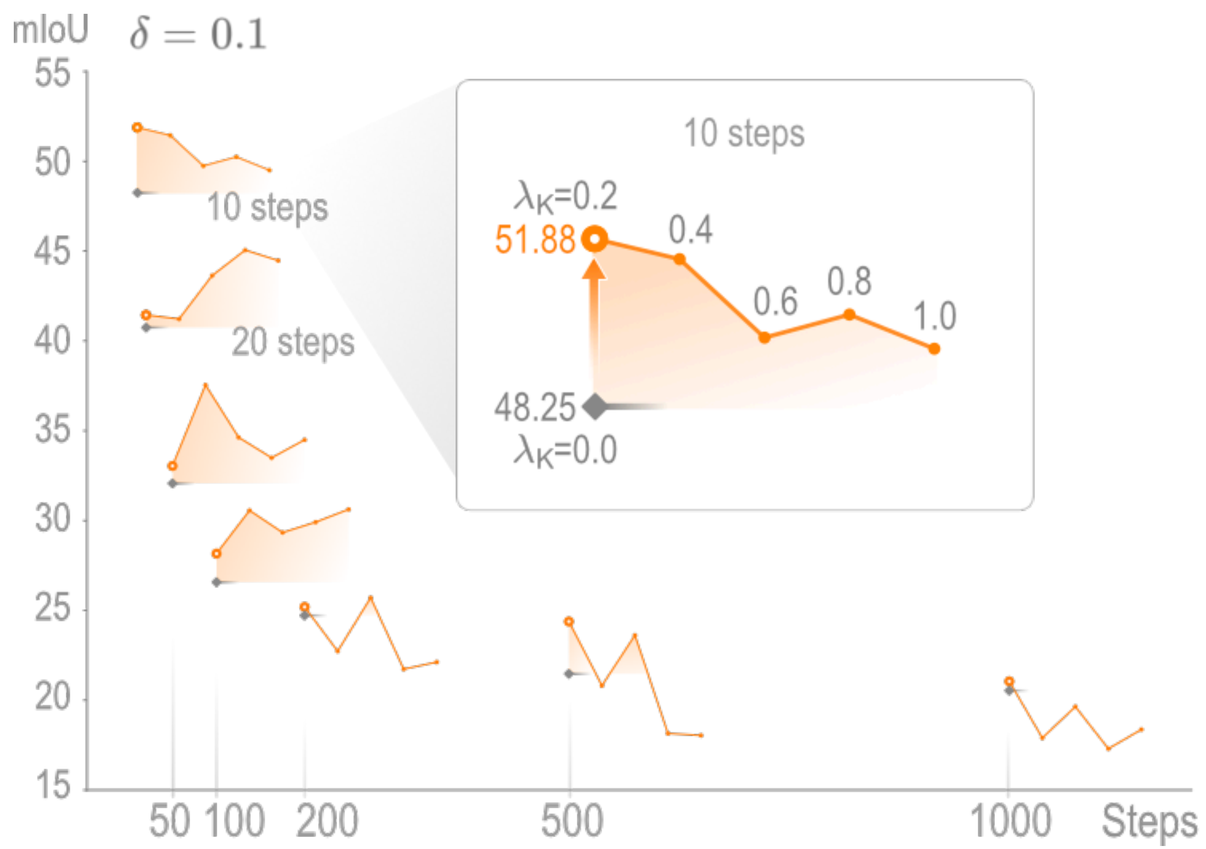


STL Tracker loses target

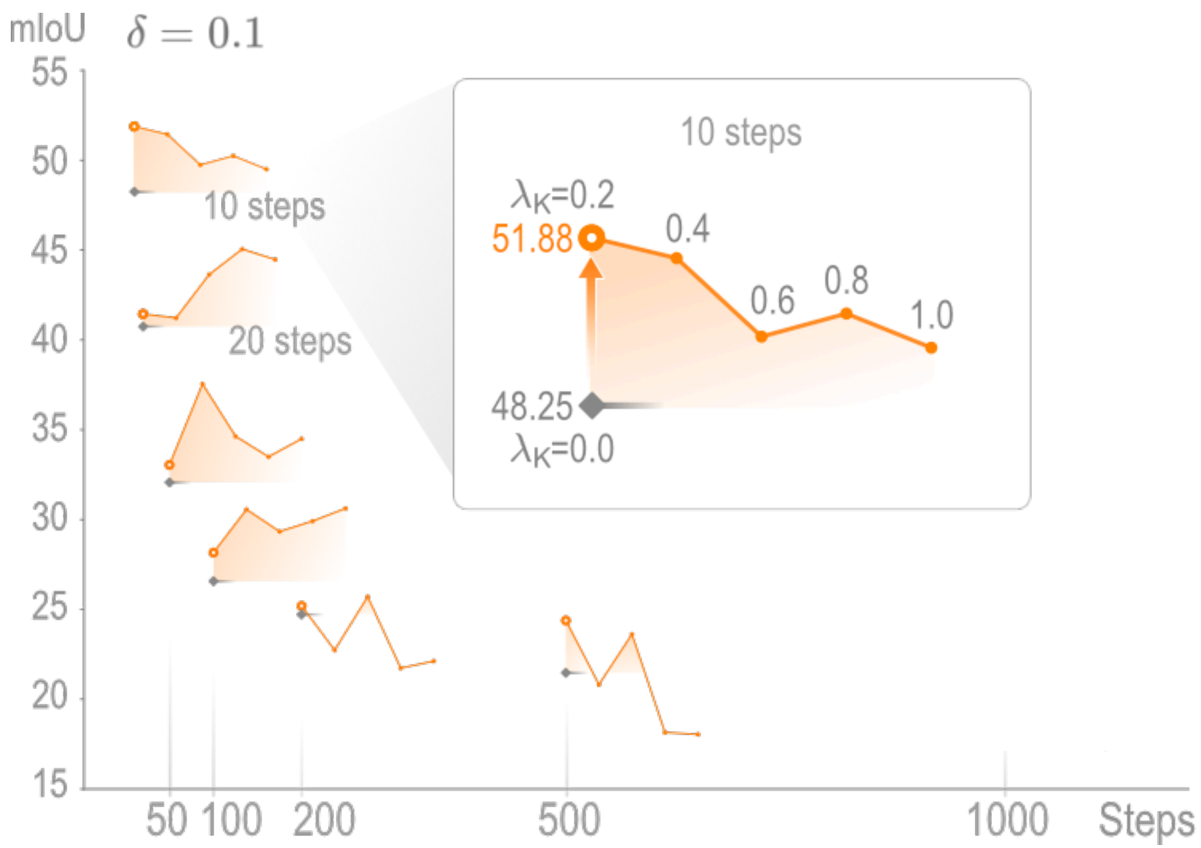
Multi-Task Learning (MTL) Tracker

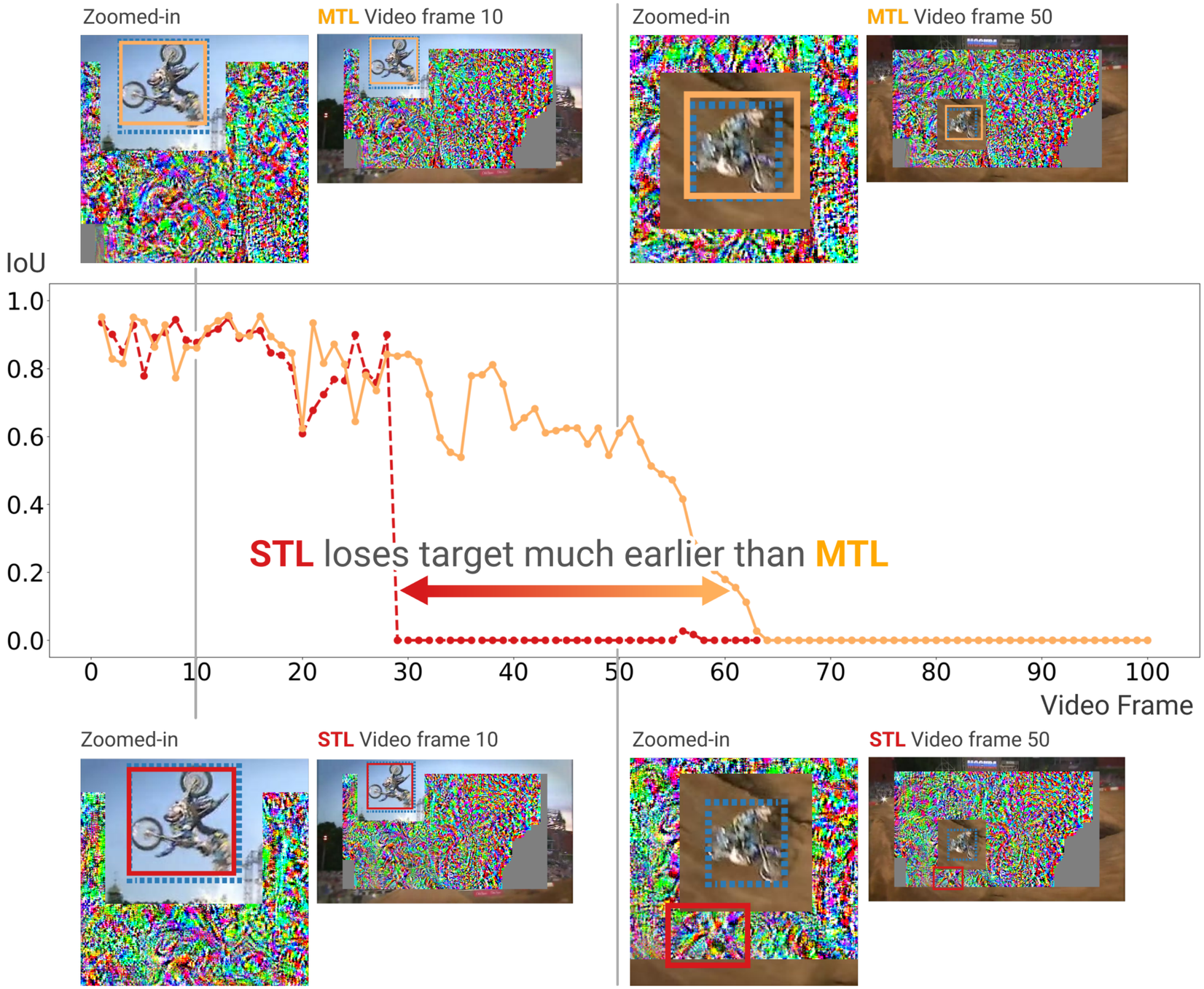
MTL Tracker recovers target

Mean Intersection-over-Union for ARMORY-CARLA dataset



Mean Intersection-over-Union for ARMORY-CARLA dataset




Zoomed-in
MTL Video frame 10
Zoomed-in
MTL Video frame 50
IoU
1.0
0.8
0.6
0.4
0.2
0.0
STL loses target much earlier than MTL
0
10
20
30
40
50
60
70
80
90
100
Video Frame
Zoomed-in
STL Video frame 10
Zoomed-in
STL Video frame 50

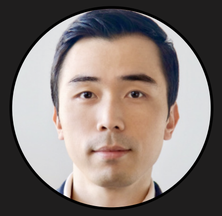
Hear No Evil

Towards Adversarial Robustness of Automatic Speech Recognition via Multi-Task Learning

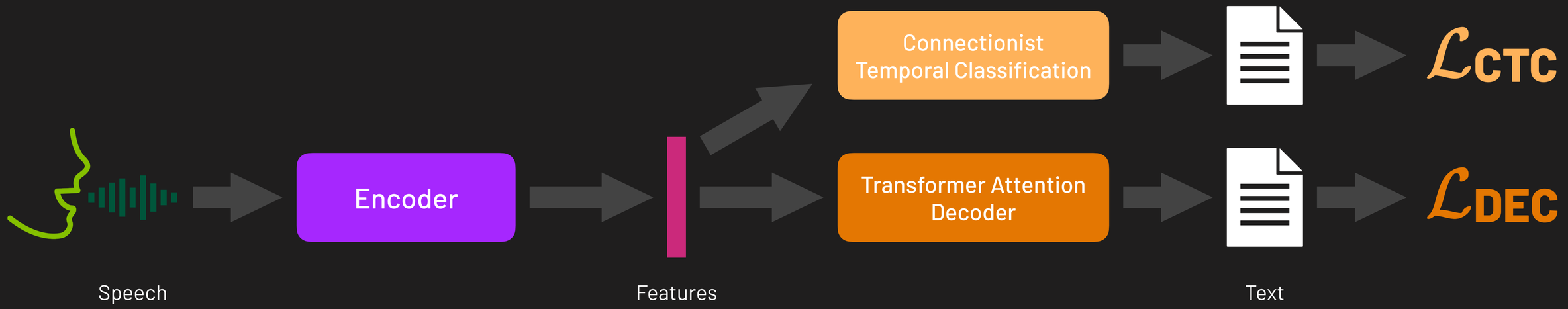
arXiv 2022
(under review)

Robust inference in ***audio*** domain

Nilaksh Das

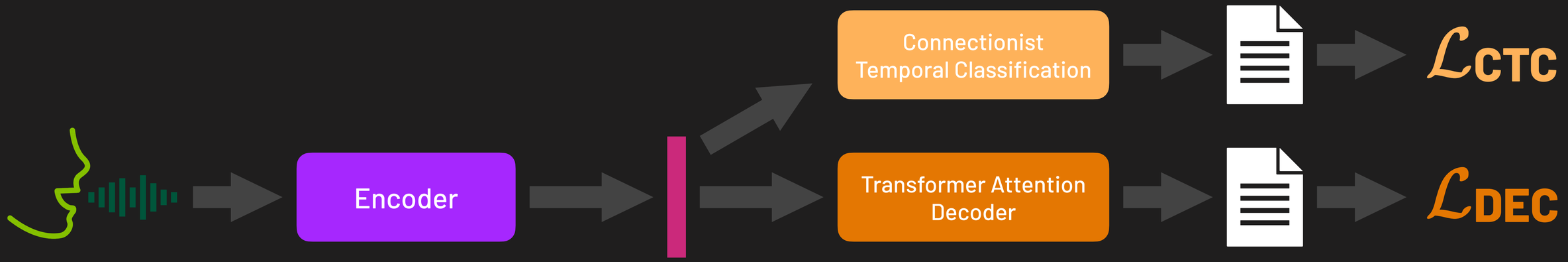


Polo Chau

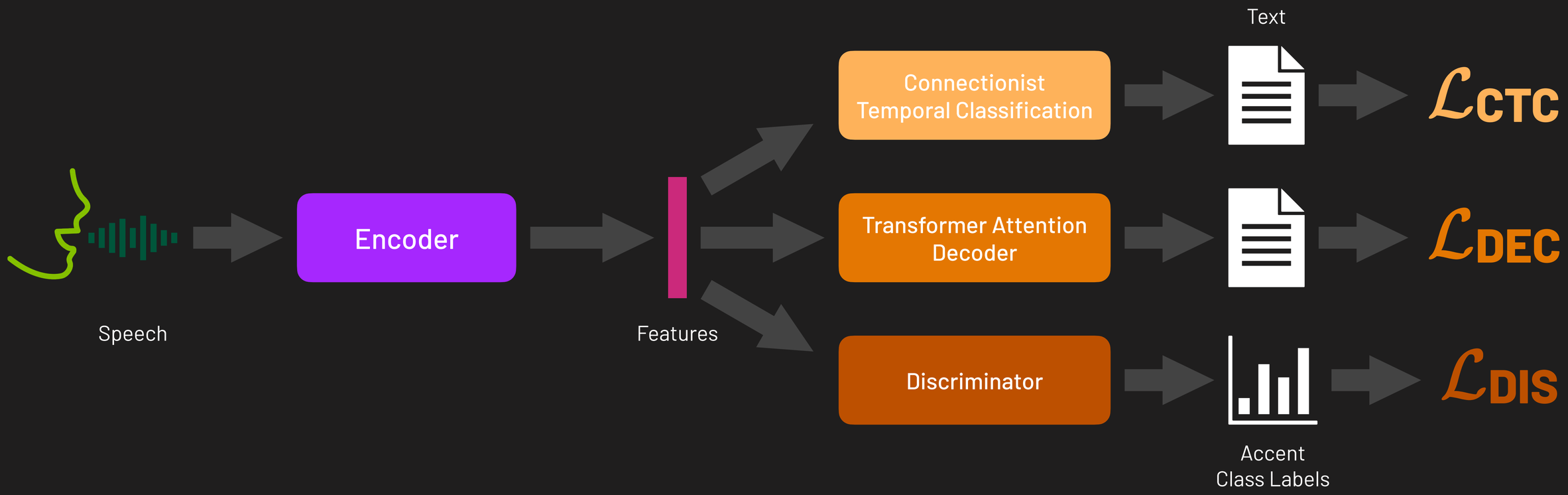


Hybrid ASR Model

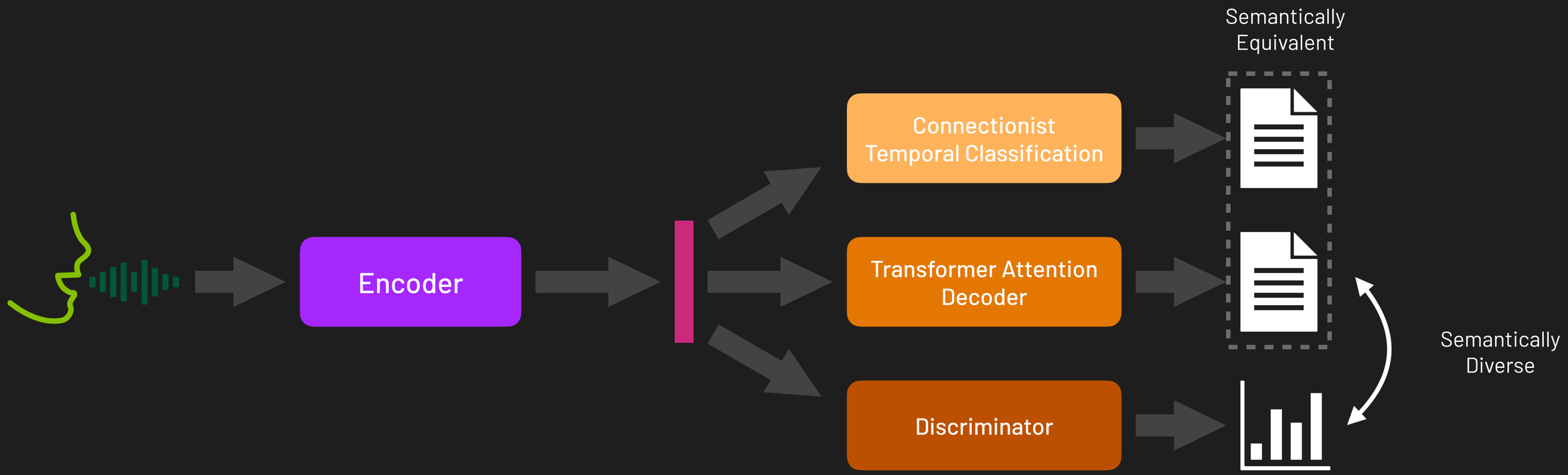
Hybrid CTC/Attention Architecture for End-to-End Speech Recognition
S. Watanabe, T. Hori, S. Kim, J. R. Hershey and T. Hayashi
IEEE Journal of Selected Topics in Signal Processing, vol. 11

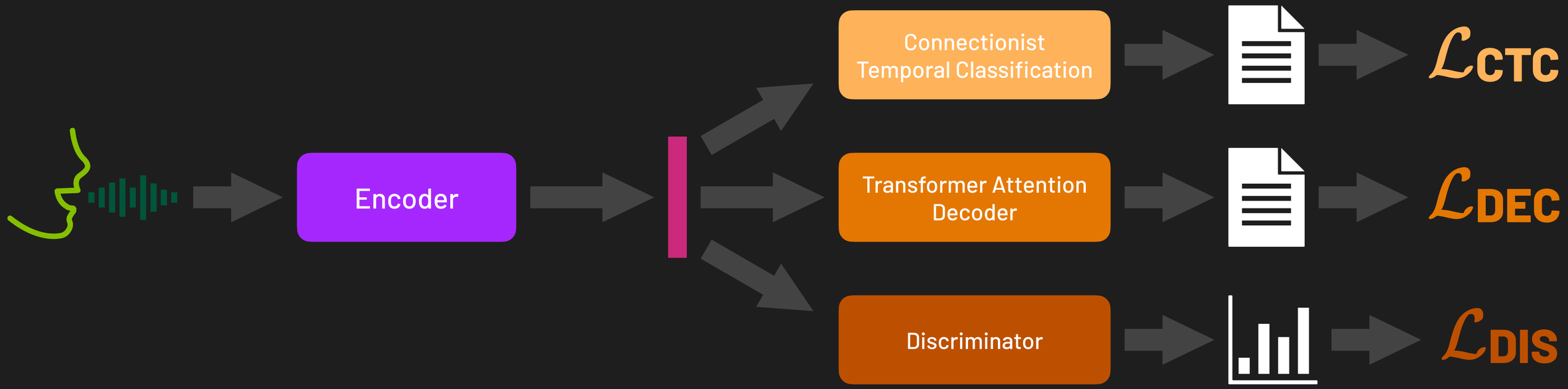


$$\mathcal{L}_{ASR} = \lambda_C \mathcal{L}_{CTC} + (1-\lambda_C) \mathcal{L}_{DEC}$$

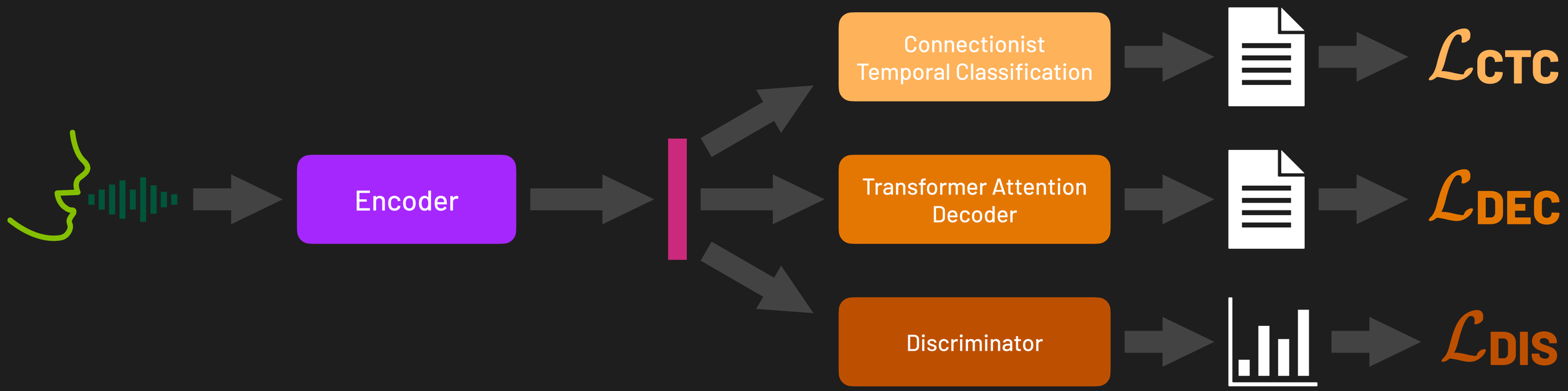


Multi-Task Learning with ASR + Accent Classification


Semantically
Equivalent
Connectionist
Temporal Classification
Encoder
Transformer Attention
Decoder
Discriminator
Semantically
Diverse

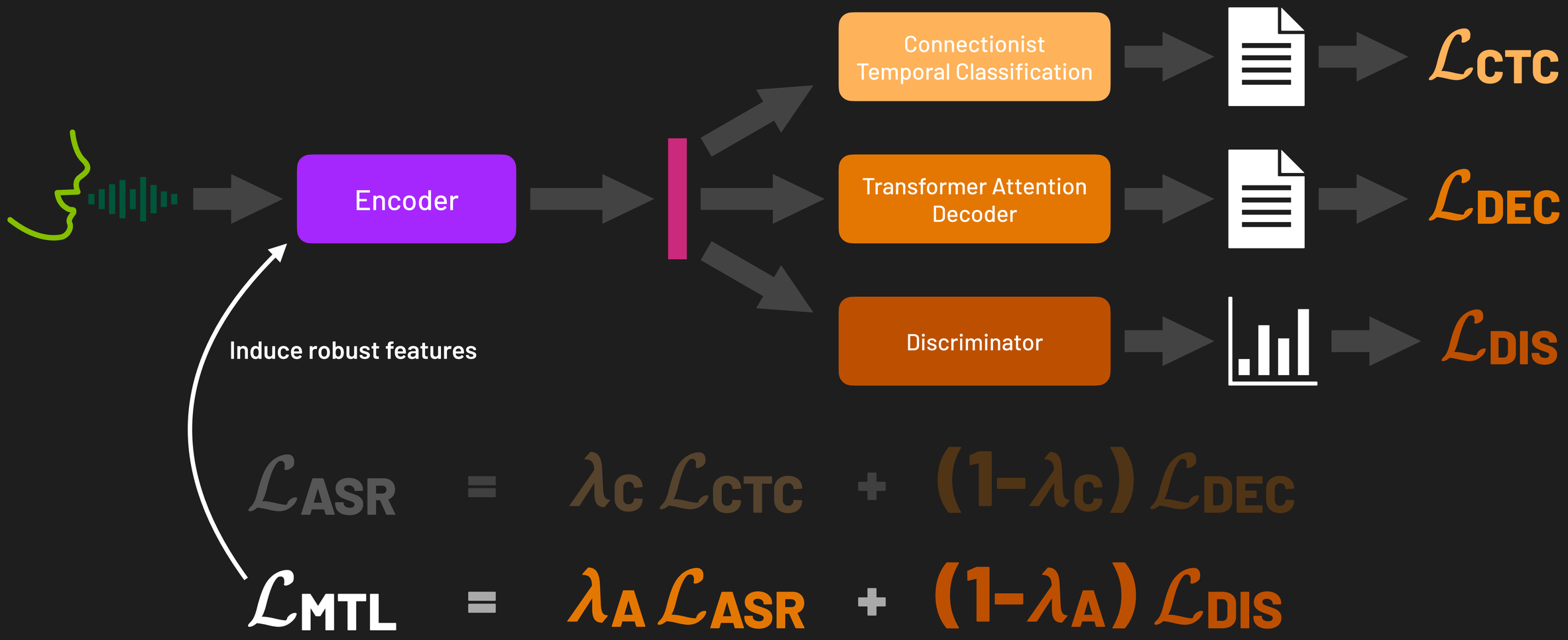


$$\mathcal{L}_{ASR} = \lambda_C \mathcal{L}_{CTC} + (1-\lambda_C) \mathcal{L}_{DEC}$$

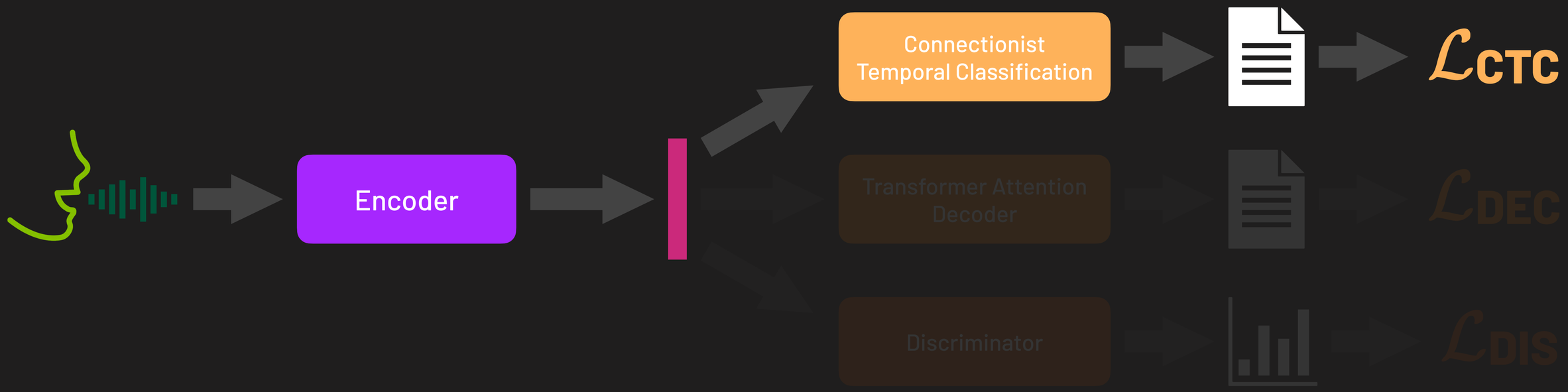


$$\mathcal{L}_{ASR} = \lambda_C \mathcal{L}_{CTC} + (1-\lambda_C) \mathcal{L}_{DEC}$$

$$\mathcal{L}_{MTL} = \lambda_A \mathcal{L}_{ASR} + (1-\lambda_A) \mathcal{L}_{DIS}$$


Encoder
Connectionist Temporal Classification
Transformer Attention Decoder
Discriminator
$\mathcal{L}_{CTC}$
$\mathcal{L}_{DEC}$
$\mathcal{L}_{DIS}$
Induce robust features
$\mathcal{L}_{ASR} = \lambda_C \mathcal{L}_{CTC} + (1-\lambda_C) \mathcal{L}_{DEC}$
$\mathcal{L}_{MTL} = \lambda_A \mathcal{L}_{ASR} + (1-\lambda_A) \mathcal{L}_{DIS}$

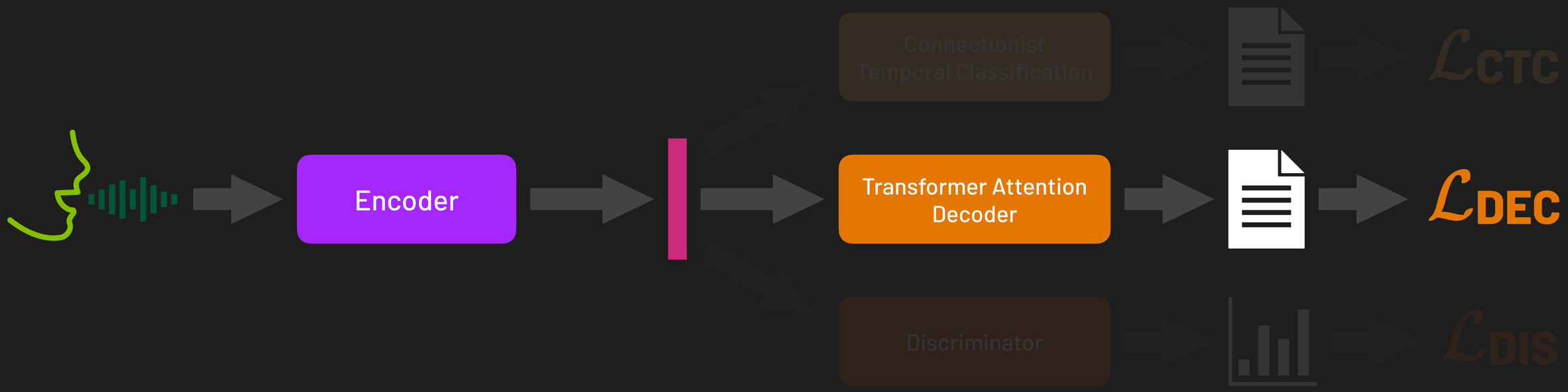
Single-Task Learning
with CTC Loss



$$\mathcal{L}_{ASR} = \lambda_C \mathcal{L}_{CTC} + (1-\lambda_C) \mathcal{L}_{DEC} \qquad \lambda_C = 1.0$$

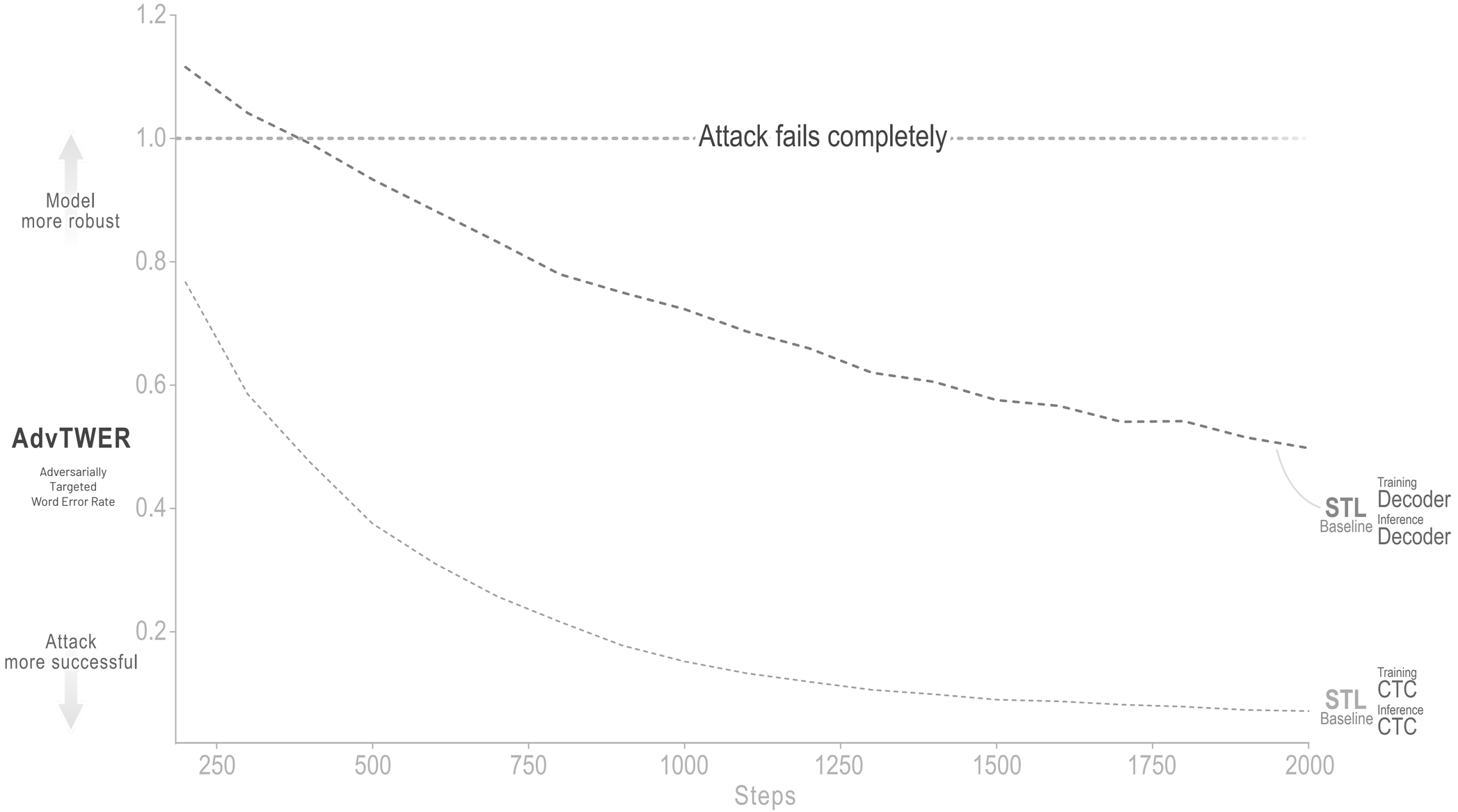
$$\mathcal{L}_{MTL} = \lambda_A \mathcal{L}_{ASR} + (1-\lambda_A) \mathcal{L}_{DIS} \qquad \lambda_A = 0.0$$

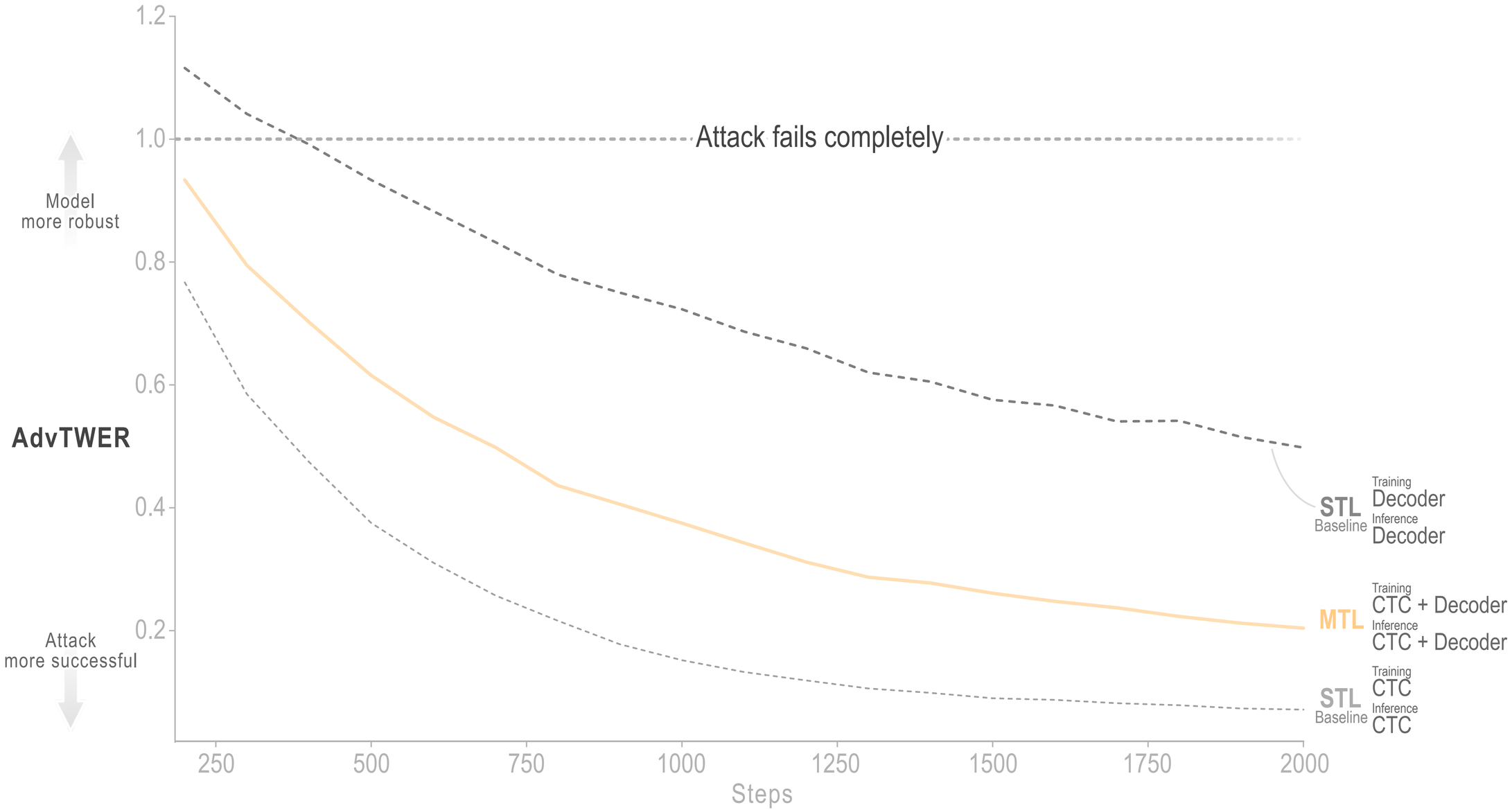
Single-Task Learning
with Decoder Loss

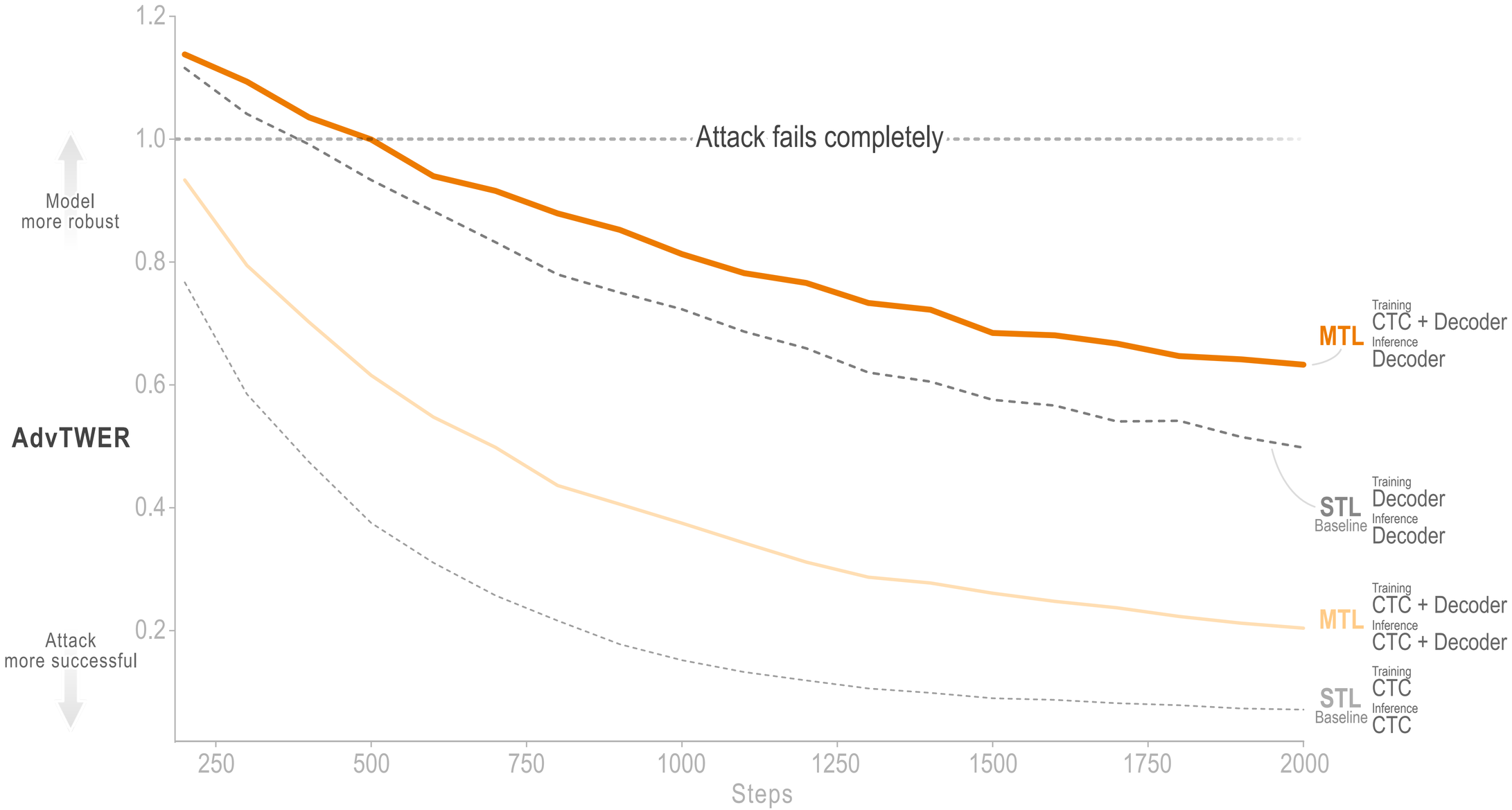


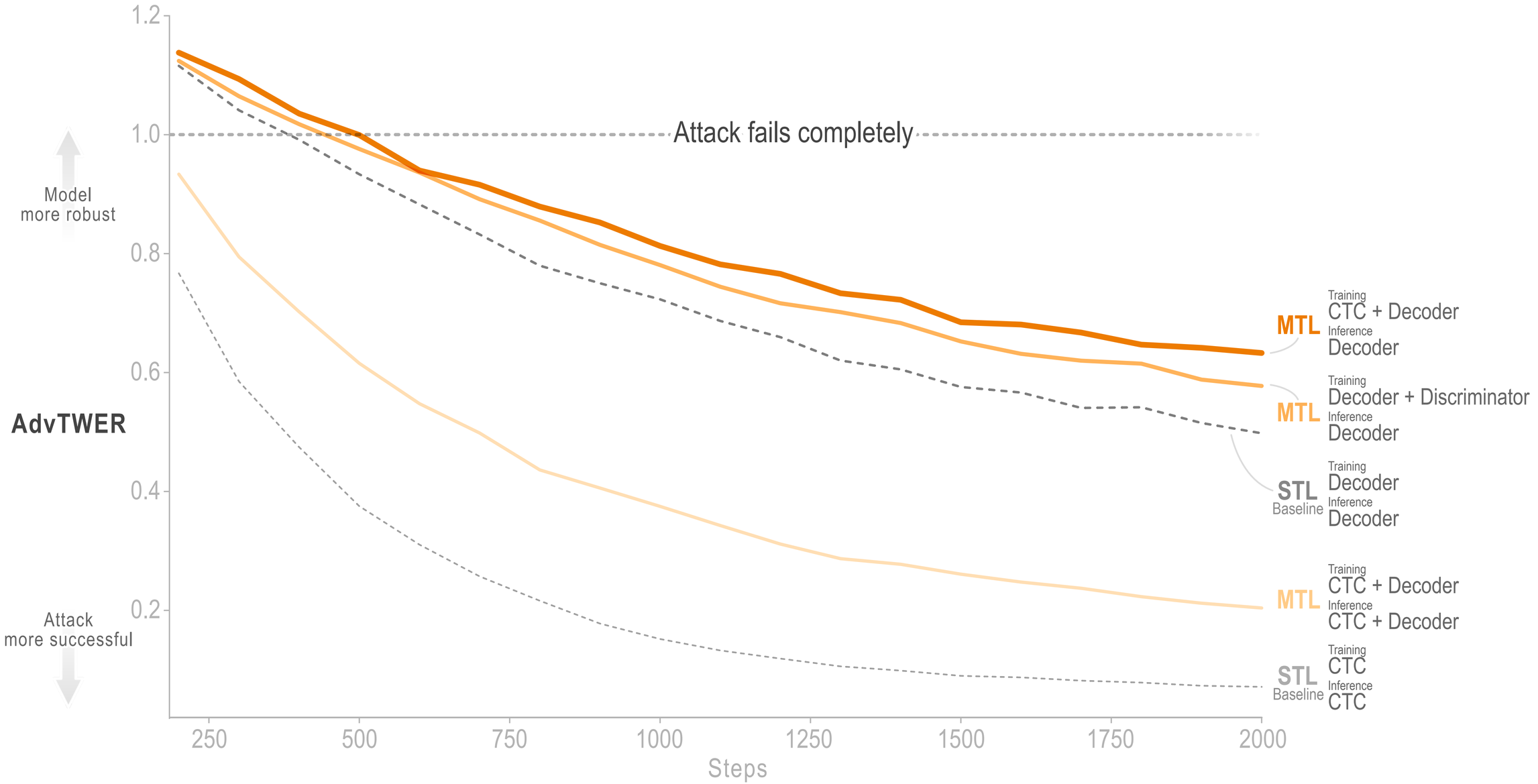
$$\mathcal{L}_{ASR} = \lambda_C \mathcal{L}_{CTC} + (1-\lambda_C) \mathcal{L}_{DEC} \qquad \lambda_C = 0.0$$

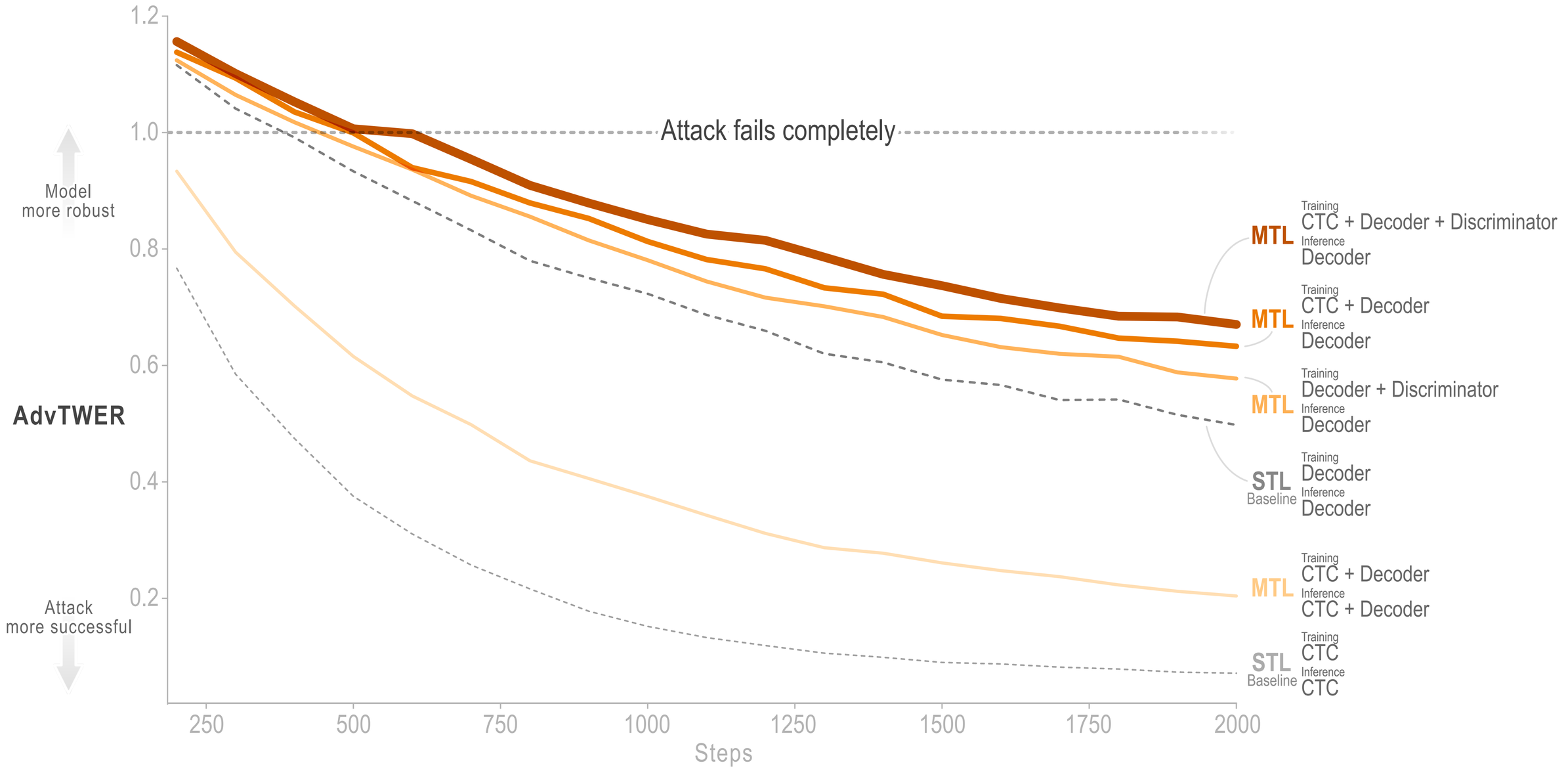
$$\mathcal{L}_{MTL} = \lambda_A \mathcal{L}_{ASR} + (1-\lambda_A) \mathcal{L}_{DIS} \qquad \lambda_A = 0.0$$

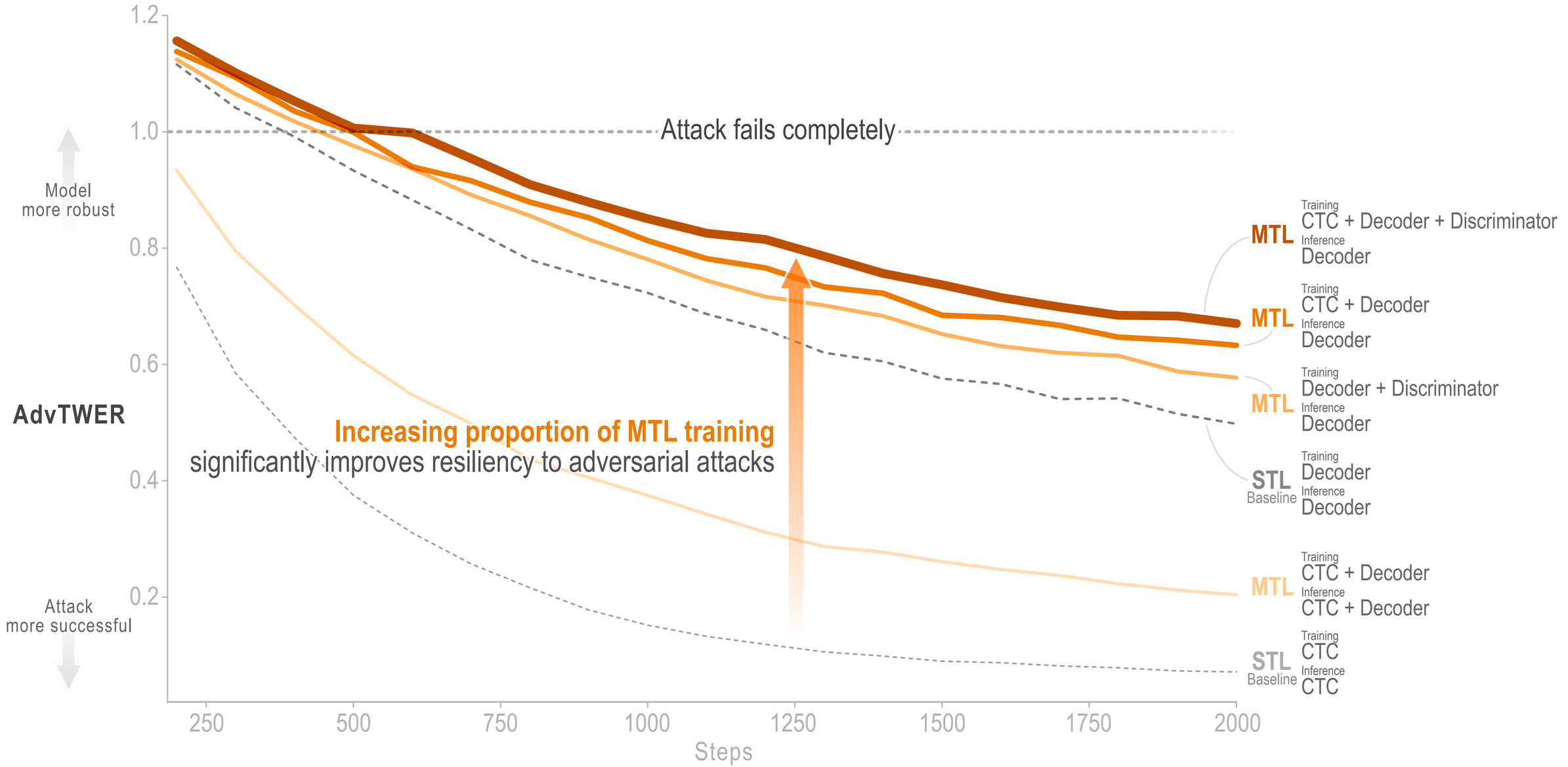

1.2
1.0
0.8
0.6
0.4
0.2
Attack fails completely
Model
more robust
AdvTWER
Adversarially
Targeted
Word Error Rate
Attack
more successful
STL
Baseline
Training
Decoder
Inference
Decoder
STL
Baseline
Training
CTC
Inference
CTC
250
500
750
1000
1250
1500
1750
2000
Steps


1.2
1.0
0.8
0.6
0.4
0.2
Attack fails completely
Model more robust
Attack more successful
AdvTWER
Steps
250
500
750
1000
1250
1500
1750
2000
STL Baseline
Training Decoder
Inference Decoder
MTL
Training CTC + Decoder
Inference CTC + Decoder
STL Baseline
Training CTC
Inference CTC

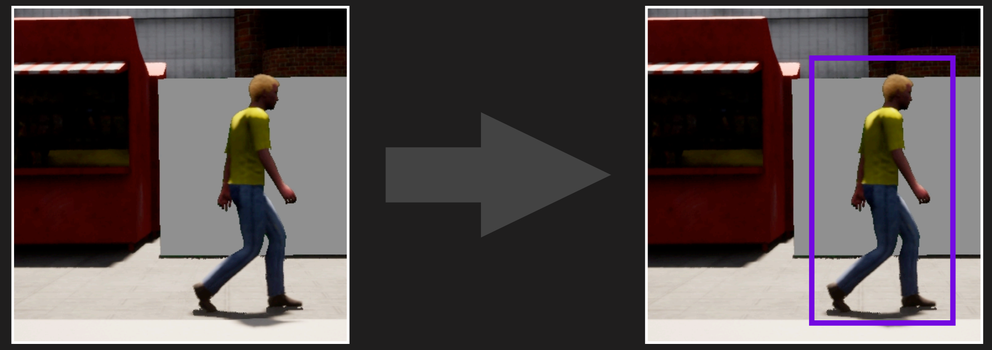

1.2
1.0
0.8
0.6
0.4
0.2
AdvTWER
Model more robust
Attack more successful
Attack fails completely
250
500
750
1000
1250
1500
1750
2000
Steps
MTL
Training
CTC + Decoder
Inference
Decoder
STL
Baseline
Training
Decoder
Inference
Decoder
MTL
Training
CTC + Decoder
Inference
CTC + Decoder
STL
Baseline
Training
CTC
Inference
CTC


1.2
1.0
0.8
0.6
0.4
0.2
Attack fails completely
Model more robust
Attack more successful
AdvTWER
250
500
750
1000
1250
1500
1750
2000
Steps
MTL
Training
CTC + Decoder
Inference
Decoder
MTL
Training
Decoder + Discriminator
Inference
Decoder
STL
Baseline
Training
Decoder
Inference
Decoder
MTL
Training
CTC + Decoder
Inference
CTC + Decoder
STL
Baseline
Training
CTC
Inference
CTC


1.2
1.0
0.8
0.6
0.4
0.2
Attack fails completely
Model more robust
Attack more successful
AdvTWER
250
500
750
1000
1250
1500
1750
2000
Steps
MTL
Training
CTC + Decoder + Discriminator
Inference
Decoder
MTL
Training
CTC + Decoder
Inference
Decoder
MTL
Training
Decoder + Discriminator
Inference
Decoder
STL
Baseline
Training
Decoder
Inference
Decoder
MTL
Training
CTC + Decoder
Inference
CTC + Decoder
STL
Baseline
Training
CTC
Inference
CTC


1.2
1.0
0.8
0.6
0.4
0.2
Attack fails completely
Model more robust
Attack more successful
AdvTWER
Increasing proportion of MTL training
significantly improves resiliency to adversarial attacks
MTL
Training
CTC + Decoder + Discriminator
Inference
Decoder
MTL
Training
CTC + Decoder
Inference
Decoder
MTL
Training
Decoder + Discriminator
Inference
Decoder
STL
Baseline
Training
Decoder
Inference
Decoder
MTL
Training
CTC + Decoder
Inference
CTC + Decoder
STL
Baseline
Training
CTC
Inference
CTC
250
500
750
1000
1250
1500
1750
2000
Steps

Person Tracking
(video domain)



Automatic Speech Recognition
(audio domain)

Multi-Task Learning

makes DNNs harder to attack ***across AI tasks***

Part I

Understand AI Vulnerabilities

GOGGLES SIGMOD 2020

Bluff IEEE VIS 2020

Part II

Fortify AI Security

SHIELD KDD 2018

SkeleVision arXiv 2022 (under review)

Hear No Evil arXiv 2022 (under review)

Part III

Enable Use of AI Security

ADAGIO ECML-PKDD 2018

MLsploit KDD Showcase 2019

ADAGIO

ECML-KDD 2018

Enabling *Interactive Experimentation* with Attacks and Defenses for Audio

First user interface for security testing with user's examples

Nilaksh Das
Georgia Tech

Madhuri Shanbhogue
Georgia Tech

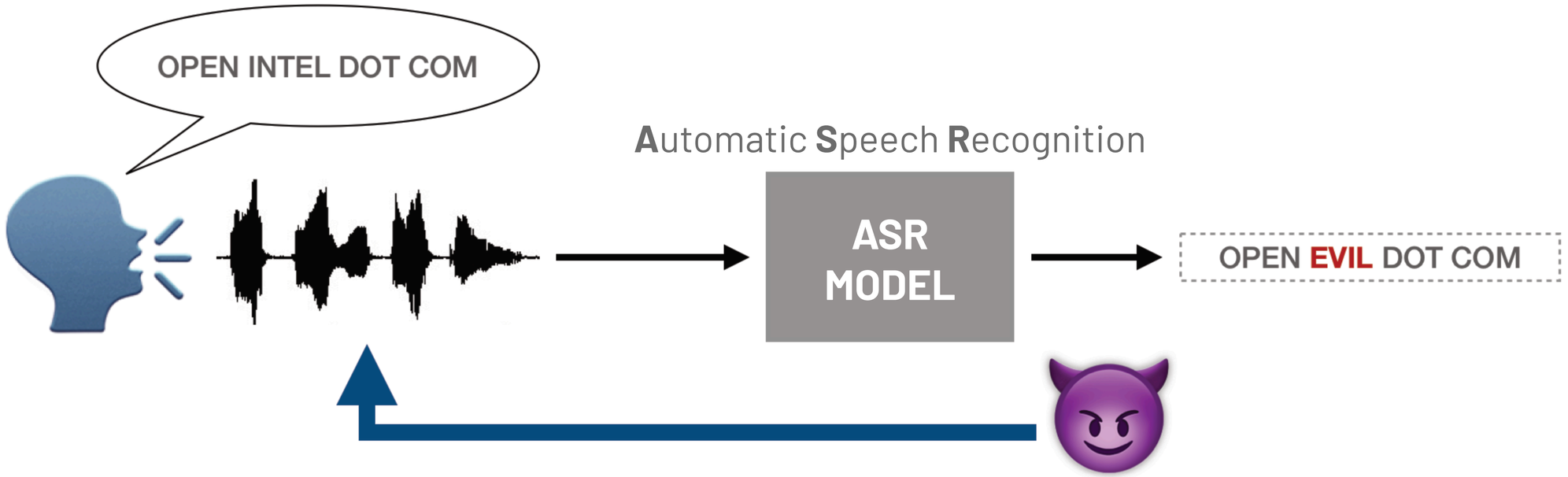
Shang-Tse Chen
Georgia Tech

Li Chen
Intel Labs

Michael Kounavis
Intel Labs

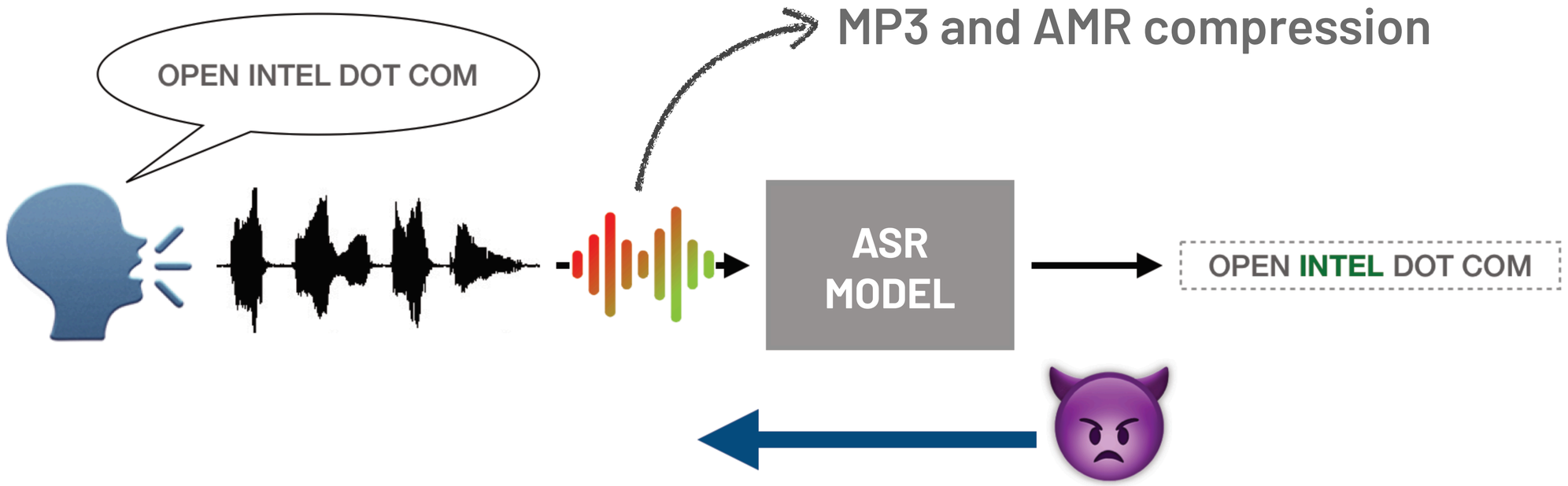
Polo Chau
Georgia Tech

Adversarial Attack on Speech-to-Text



An adversary uses backpropagation to attack the model.

Adversarial Attack on Speech-to-Text



ADAGIO incorporates compression as defense,
which blocks the gradient to the attacker.

Defense	WER (no attack)	WER (with attack)	Targeted attack success rate
None	0.369	1.287	92.45%
AMR	0.488	0.666	**0.00%**
MP3	0.400	0.780	**0.00%**

AMR & MP3 recover WER

Targeted attack is eliminated

Word Error Rate (WER) and the targeted attack success rate on the DeepSpeech ASR model (lower is better for both)

ADAGIO

Interactive Experimentation with Adversarial Attack & Defense for Audio

- 🗣️ Upload your own audio sample
- ⚔️ Perform audio adversarial attack
- 🔰 Apply compression to defend
- ▶️ Play audio, listen for differences

ADAGIO = **A**ttack & **D**efense for **A**udio in a **G**adget with **I**nteractive **O**perations

ADAGIO

MLsploit

Interactive Experimentation with Adversarial ML Research

KDD 2019 Project Showcase
Black Hat Asia, Arsenal 2019

- Employed in teaching at Georgia Tech
- Tech-transferred to Intel Labs
- Open-sourced at *github.com/mlsploit*

Nilaksh Das
Georgia Tech

Georgia Tech intel

Contributors from

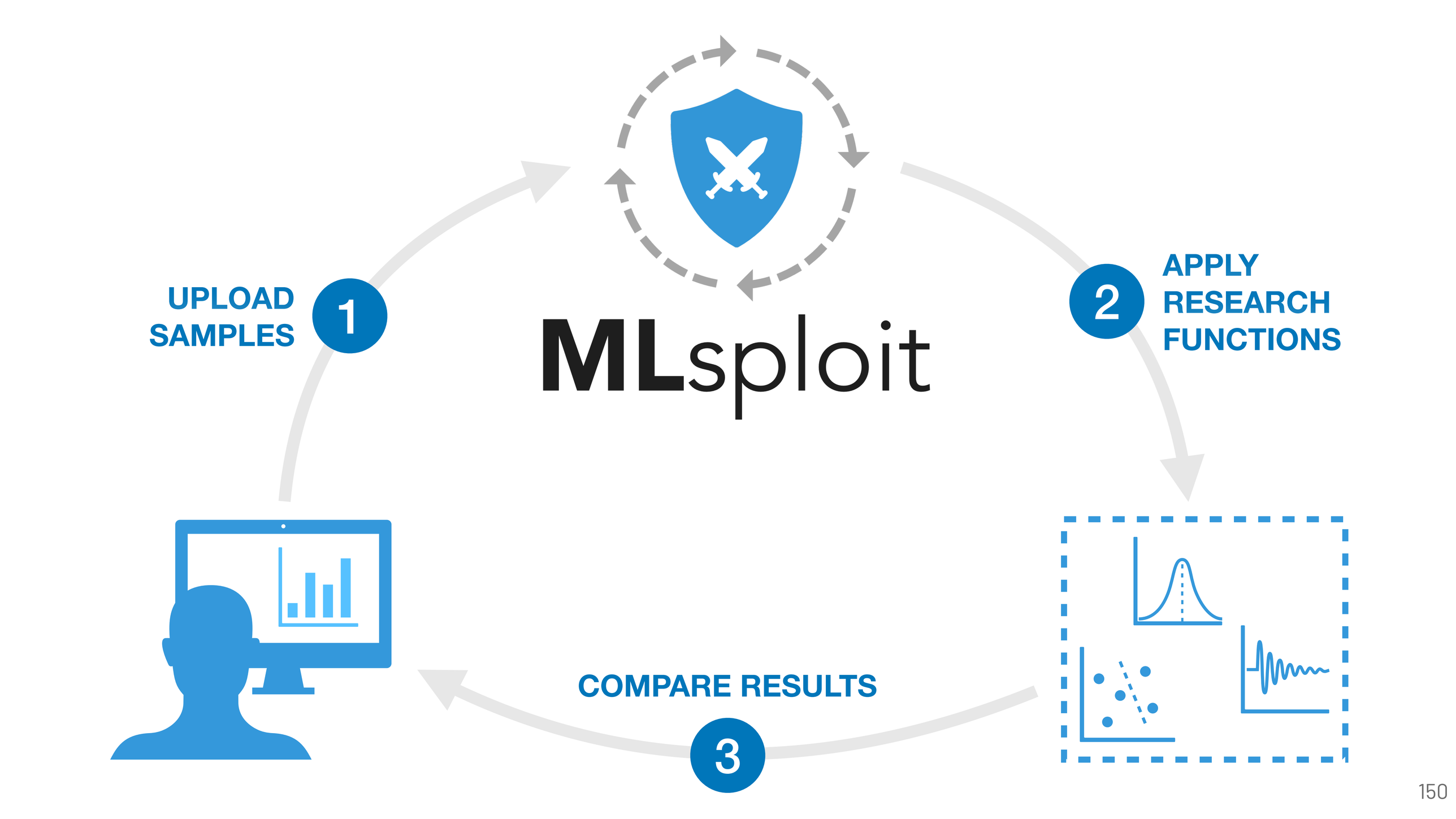
Intel Science and Technology Center for Adversary-Resilient Security Analytics:

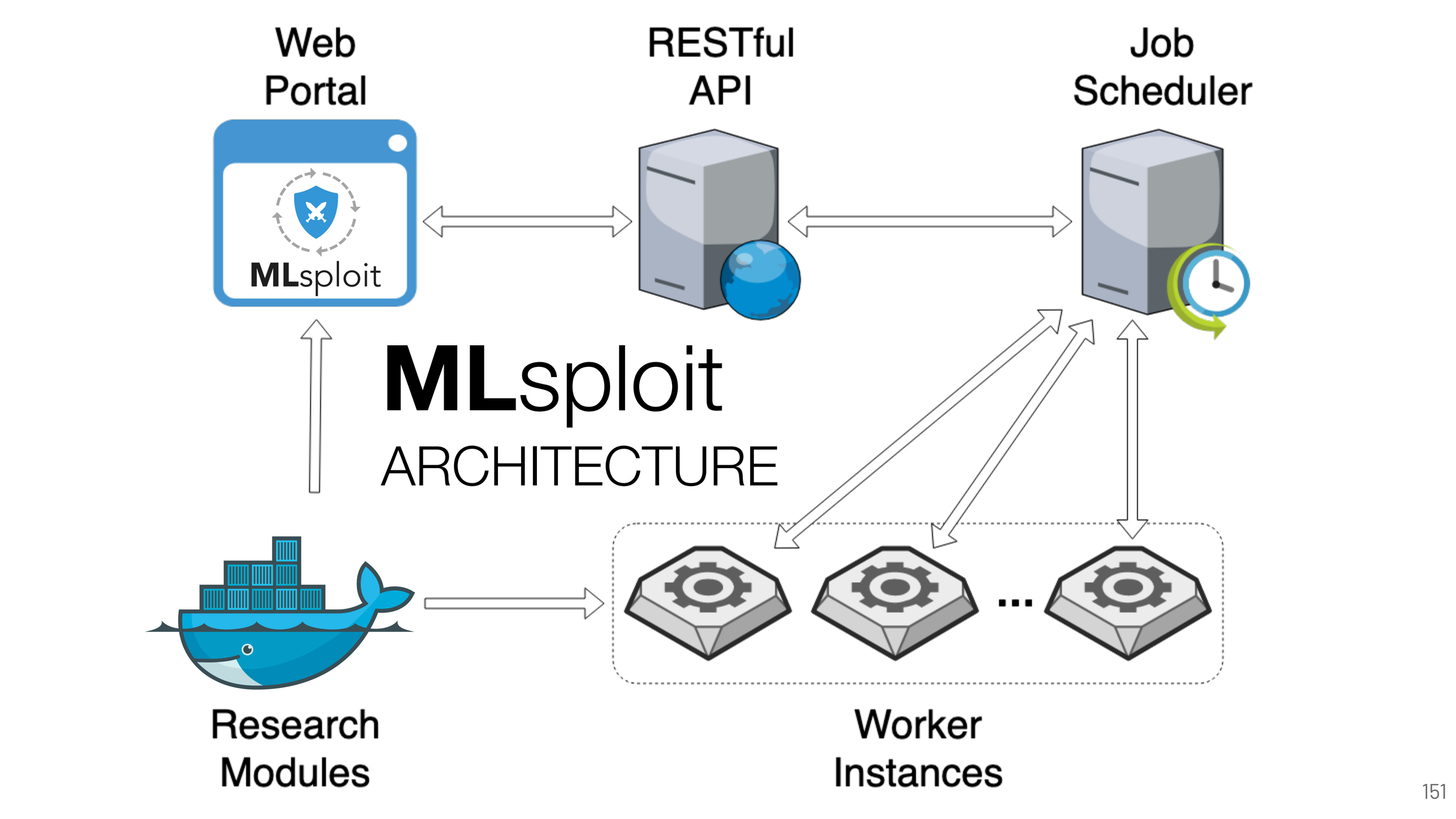
Siwei Li, Chanil Jeon, Jinho Jung*, Shang-Tse Chen*, Carter Yagemann*, Evan Downing*, Haekyu Park, Evan Yang, Li Chen, Michael Kounavis, Ravi Sahita, David Durham, Scott Buck, Polo Chau, Taesoo Kim, Wenke Lee

*equal contribution

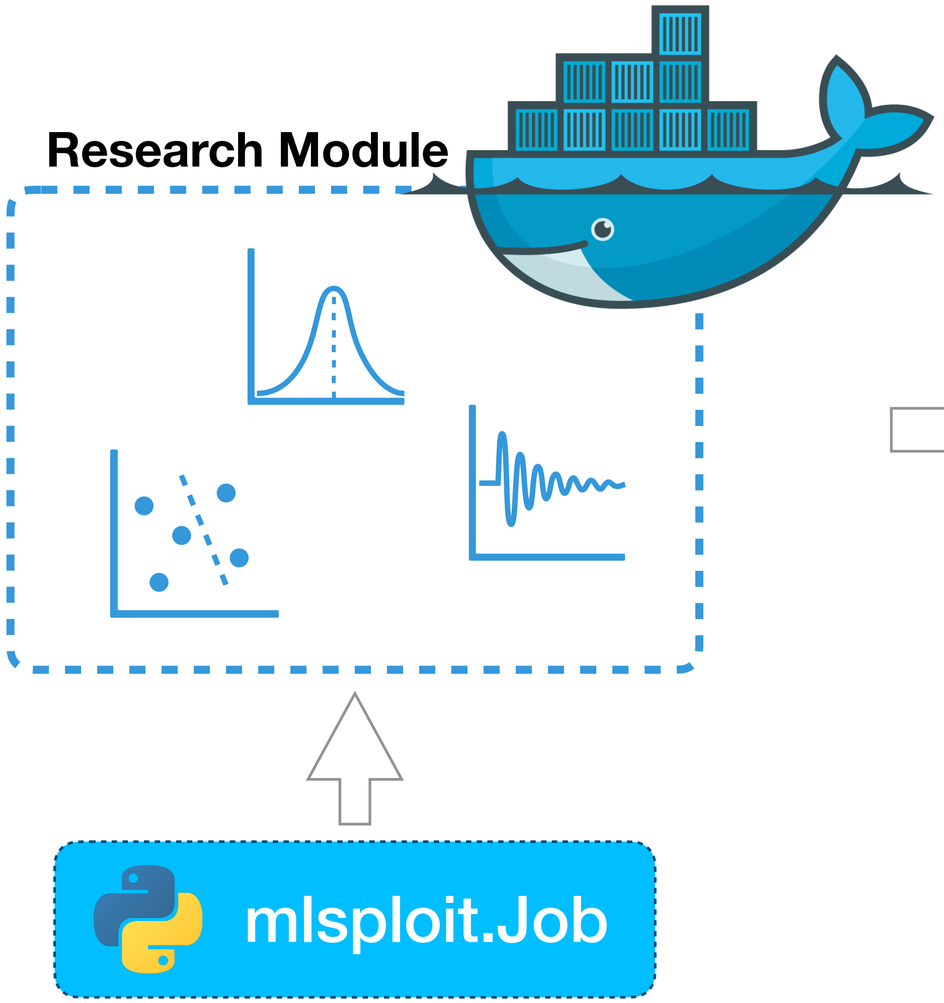
MLsploit

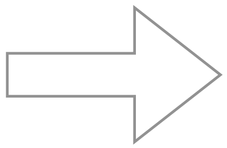
- ★ **Research modules** for adversarial ML
  - ✳ Enables **comparison** of attacks and defenses
- ★ **Interactive experimentation** with ML research
- ★ Researchers can **easily integrate** novel research into an intuitive and seamless **user interface**


MLsploit
UPLOAD SAMPLES
1
2
APPLY RESEARCH FUNCTIONS
COMPARE RESULTS
3


Web
Portal
RESTful
API
Job
Scheduler
MLsploit
MLsploit
ARCHITECTURE
...
Research
Modules
Worker
Instances

EASY INTEGRATION OF RESEARCH





MLsploit

Researchers

MLsploit

Students

Practitioners

MLsploit

Experiment with AI Security in your browser!

Research Modules

SHIELD
Fast, Practical Defense for Deep Learning

Foolbox
Fool neural networks!

AVPASS
Android Malware Detection Bypass

Barnum
Deep Learning Software Anomaly Detection

ELF
ELF File Malware Detection and Bypassing

Network
Network Intrusion Detection and Evasion

Attack and Defend

Attack

Classify

MLsploit

Experiment with AI Security in your browser!

Research Modules

SHIELD

Fast, Practical Defense for Deep Learning

Foolbox

Fool neural networks!

AVPASS

Android Malware Detection Bypass

Barnum

Deep Learning Software Anomaly Detection

ELF

ELF File Malware Detection and Bypassing

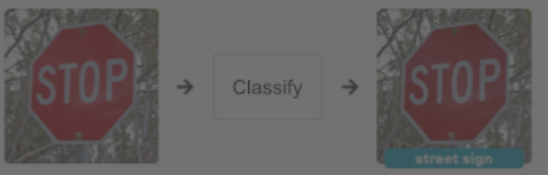
Network

Network Intrusion Detection and Evasion

Attack and Defend

Attack

Classify



MLsploit Experiment with AI Security in your browser!

Research Modules

SHIELD
Fast, Practical Defense for Deep Learning

Foolbox
Fool neural networks!

Barnum
Deep Learning Software Anomaly Detection

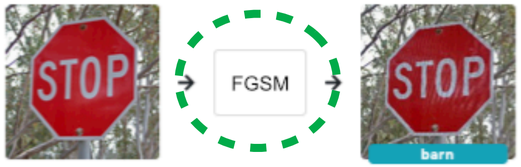
ELF
ELF File Malware Detection and Bypassing

Network
Network Intrusion Detection and Evasion

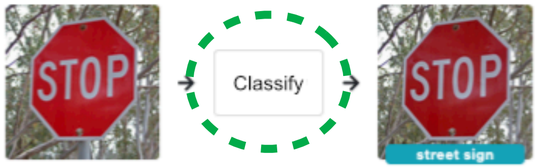
users can create multiple pipelines with different research functions

Attack and Defend

Attack



Classify



MLsploit Experiment with AI Security in your browser!

Research Modules

SHIELD
Fast, Practical Defense for Deep Learning

Foolbox
Fool neural networks!

AVPASS
Android Malware Detection Bypass

Barnum
Deep Learning Software Anomaly Detection

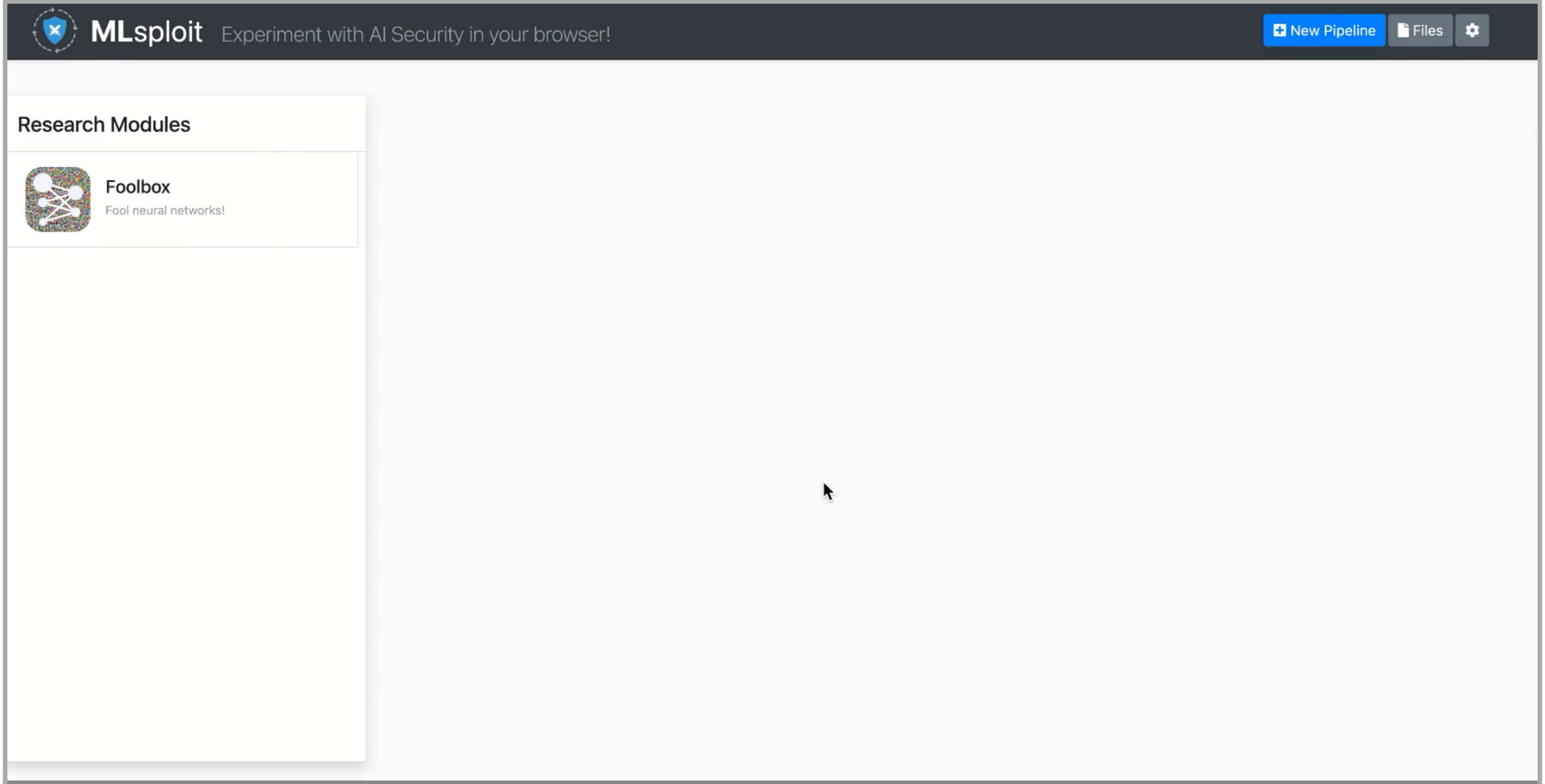
ELF
ELF File Malware Detection and Bypassing

Attack and Defend

Attack

Classify

... and experiment with their own samples!



Expand the "**Foolbox**" Research Module
to see various research functions

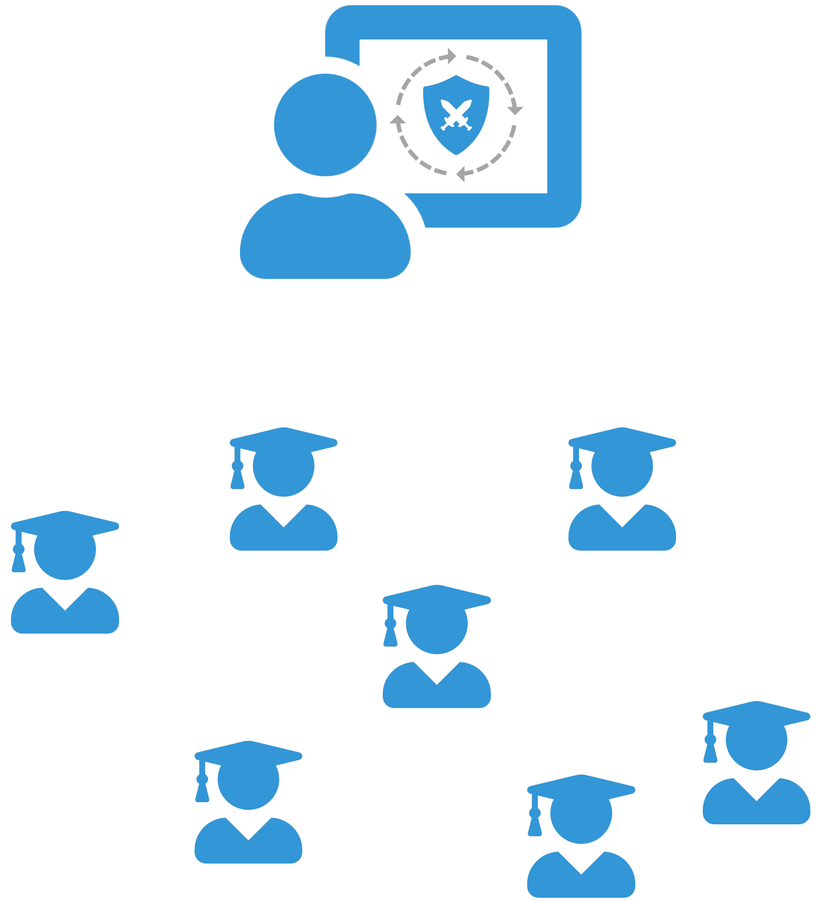
IMPACT

MLsploit employed at **GT**

to teach students how to

train, *evade* and *defend*

DNNs for malware detection



Part I

Understand
AI Vulnerabilities

GOGGLES SIGMOD 2020

Bluff IEEE VIS 2020

Part II

Fortify
AI Security

SHIELD KDD 2018

SkeleVision arXiv 2022 (under review)

Hear No Evil arXiv 2022 (under review)

Part III

Enable
Use of AI Security

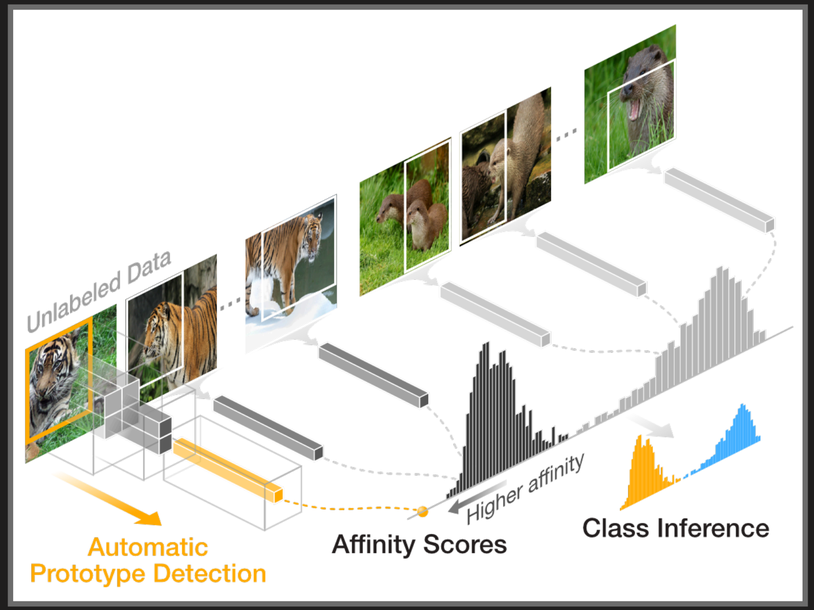
ADAGIO ECML-PKDD 2018

MLsploit KDD Showcase 2019

Research Contributions

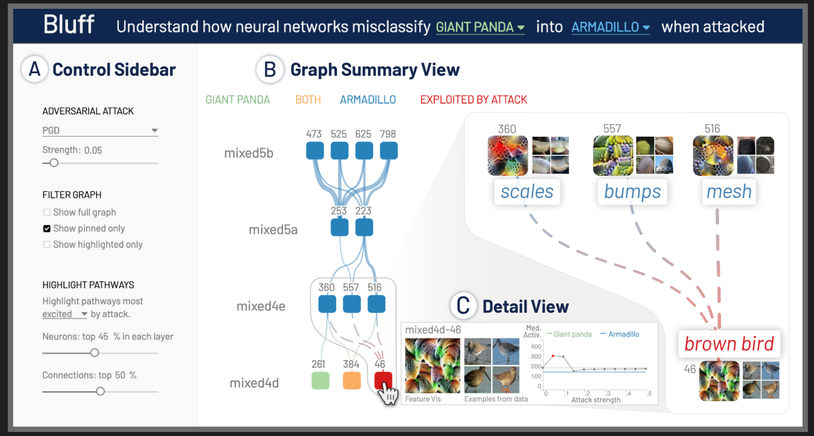
Novel principled approaches

- GOGGLES contributes novel, theoretically principled approaches to extract ***interpretable prototypes***
- weakly-supervised GOGGLES framework is only ***7% away*** from the fully-supervised baseline



Novel visualization technique for deciphering attacks

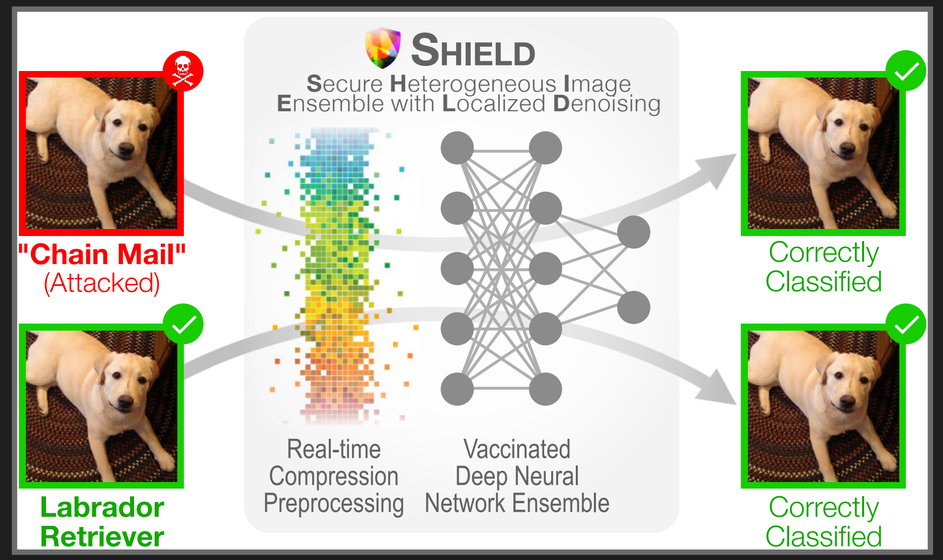
- Bluff helps in identifying the ***vulnerable neurons*** in a DNN that are non-performant and lead to misclassifications under attack



Research Contributions

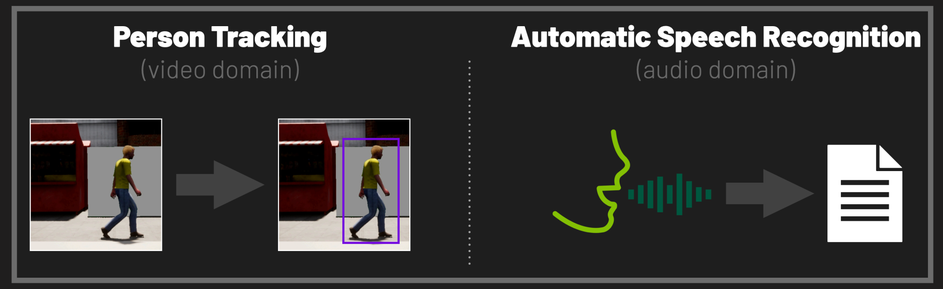
Faster, generalizable defenses

- SHIELD defense is up to ***22x faster*** than other preprocessing defenses
- ADAGIO defense removes ***all*** non-adaptive targeted adversarial attacks



Fundamentally unifying approach for training robust models across AI tasks

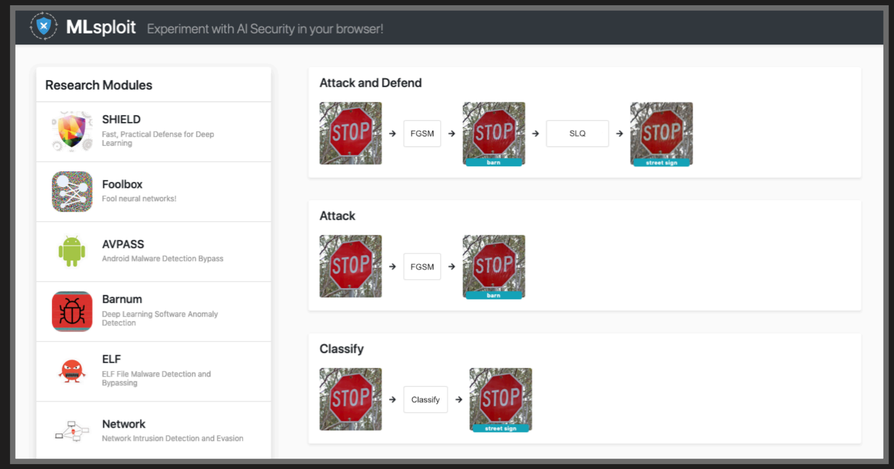
- Our in-depth study of ***Multi-Task Learning*** establishes it as a fundamentally unifying approach that helps learn robust features for resisting adversarial attacks ***across AI tasks***



Research Contributions

First scalable system for interactive Adversarial ML research

- MLsploit is the first open-source interactive system that allows in-depth security testing of AI models, and lowers barriers to entry for everyone



Understanding, Fortifying and Democratizing AI Security

Nilaksh Das
nilakshdas.com

Georgia Tech