poloclub.github.io

SECURE & INTERPRETABLE AI

Polo Chau

Associate Professor
Associate Director, MS Analytics
Associate Director of Corporate Relations, ML Center
Georgia Tech

Fred	Nilaksh	Haekyu	Scott	Jay	Austin	Rahul	Anmol	Bob	Jonathan	Will	Rob	Omar	Frank	Jon	Robert	Dongkyu
CSE PhD	CSE PhD	CS PhD	ML PhD	ML PhD	ML PhD	CS PhD	MS CSE	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	Post-Doc.

Polo Club of Data Science

AI + **HI**

ARTIFICIAL INTELLIGENCE

HUMAN INTELLIGENCE

Scalable interactive tools to make sense of complex large-scale datasets and models

Fred	Nilaksh	Haekyu	Scott	Jay	Austin	Rahul	Anmol	Bob	Jonathan	Will	Rob	Omar	Frank	Jon	Robert	Dongkyu
CSE PhD	CSE PhD	CS PhD	ML PhD	ML PhD	ML PhD	CS PhD	MS CSE	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	Post-Doc.

Current Research Thrusts

Secure AI

Interpretable AI

Why focus on them?
How are they related?

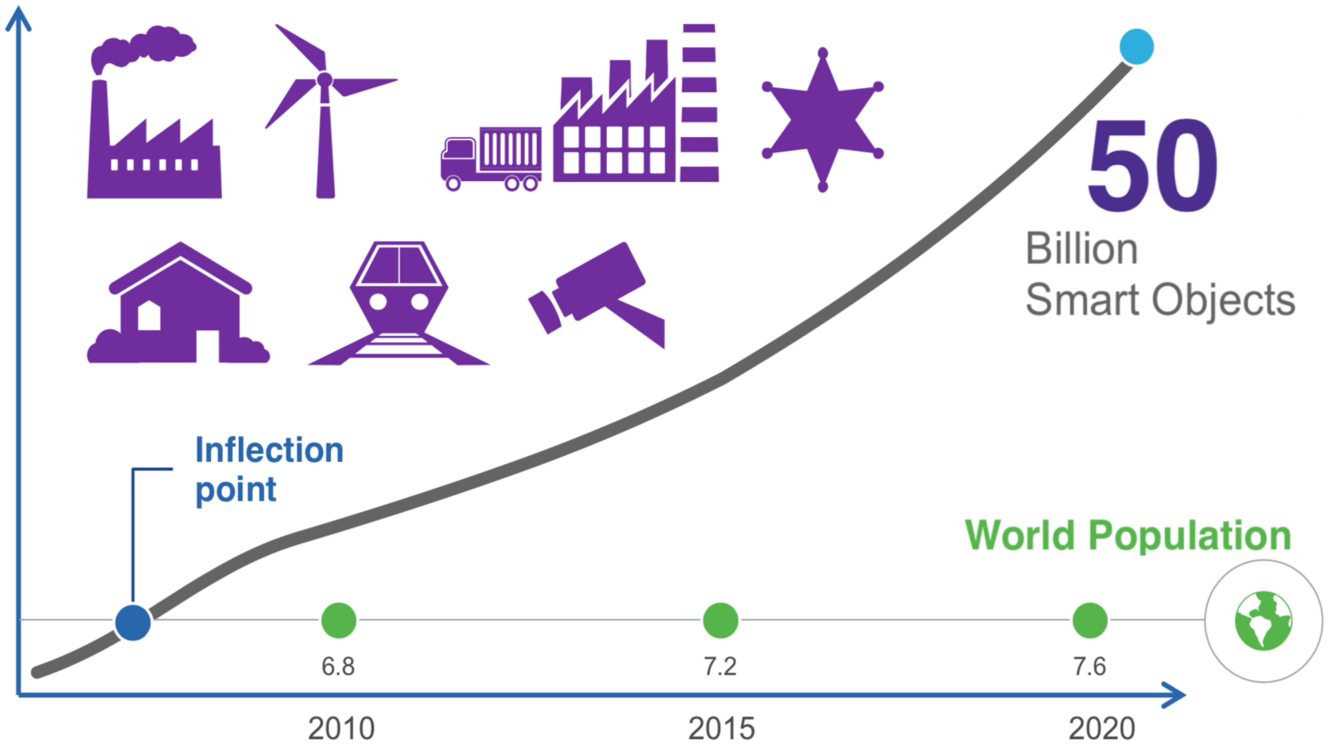
AI now used in safety-critical applications.
Important to study threats & countermeasures.

How a Self-Driving Uber Killed a Pedestrian in Arizona

AI Security Problems Are Everywhere

Smart toaster does exist!

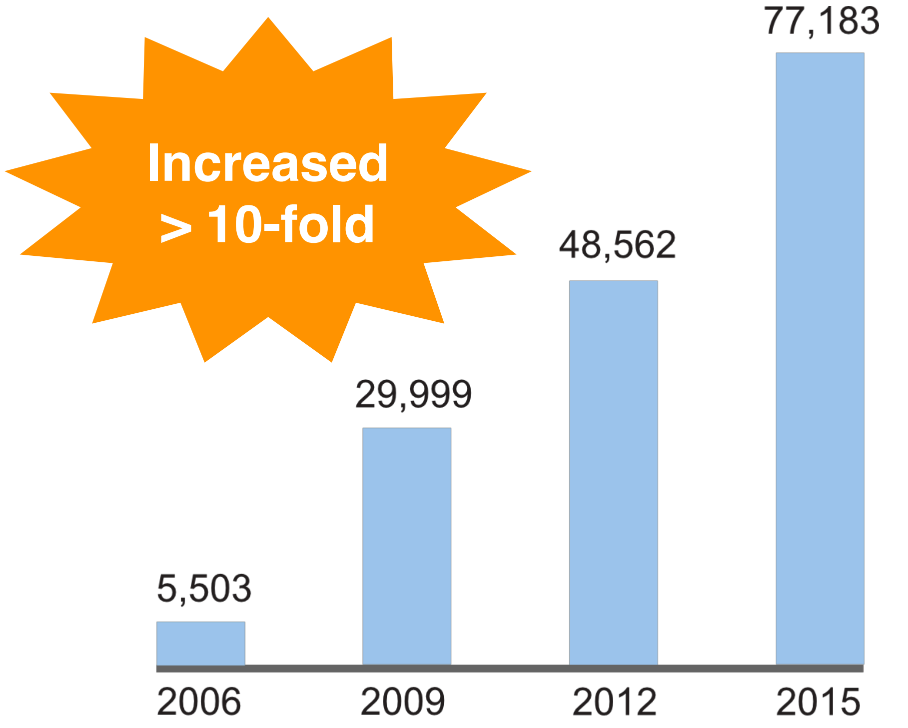
AI Security is becoming increasingly important



Source: Cisco

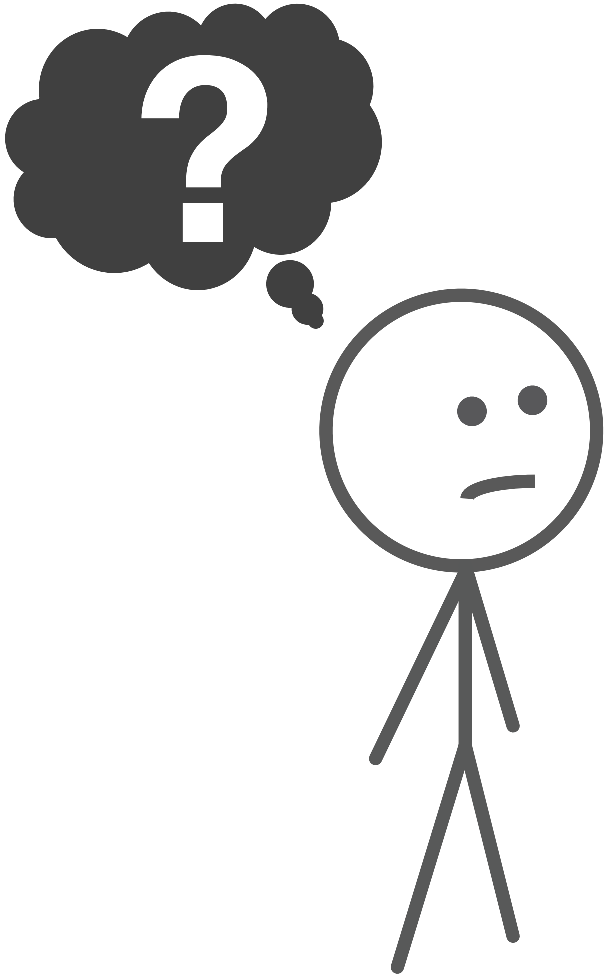
incidents

reported by U.S. federal agencies

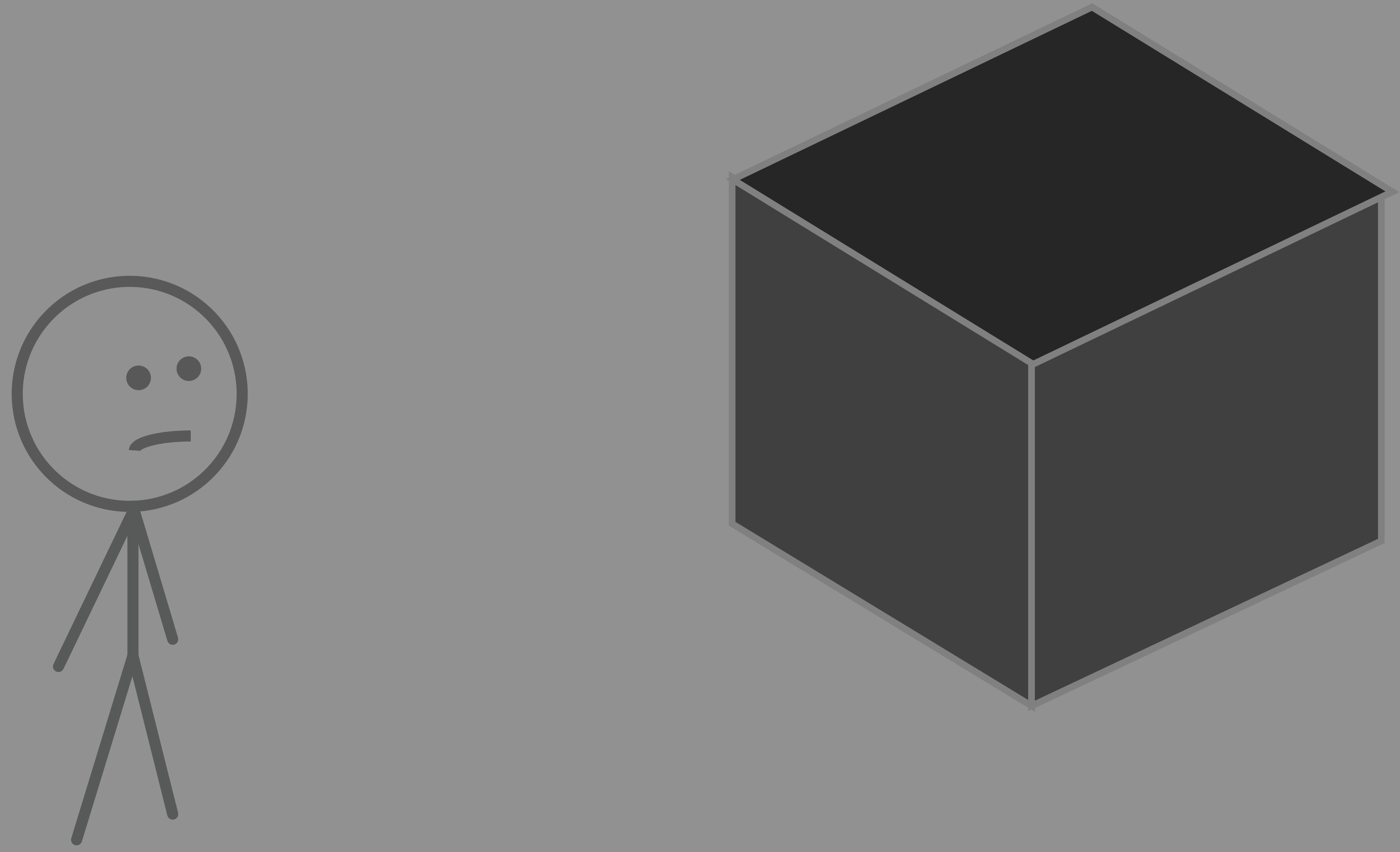


Source: US Department of Homeland Security

How do we know if a defense for AI is working?



AI models often used as black-box



Interpretable AI

Interpretable AI

Via scalable, interactive, usable interfaces to help people understand complex, large-scale ML systems.

Secure AI

ShapeShifter First attack fooling object detectors

SHIELD & Adagio Real-time defense

Interpretable AI

Summit Scalable interpretation for deep learning

GAN Lab & CNN Explainer Interactive learning

Surveys AI guidelines, visual analytics for deep learning

ShapeShifter

ECML-PKDD 2018

First Targeted *Physical* Adversarial Attack for Object Detection

Shang-Tse Chen
National Taiwan University

Cory Cornelius
Intel

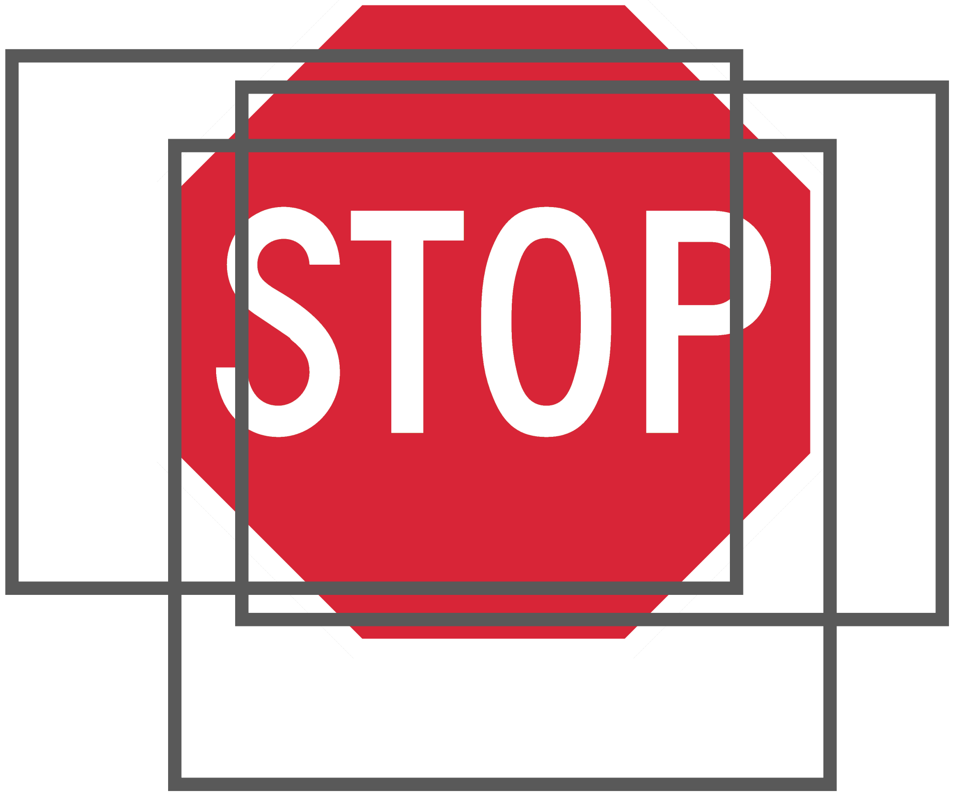
Jason Martin
Intel

Polo Chau
Georgia Tech

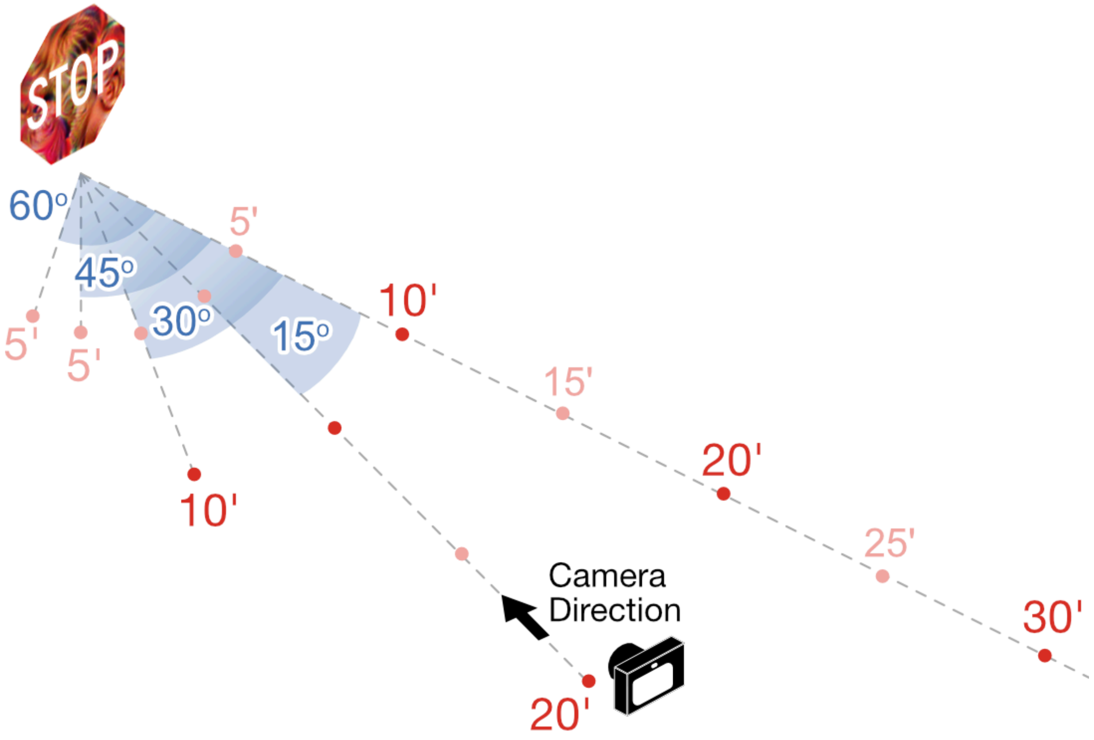
Stop Sign → Person

Challenges of Physically Attacking Faster R-CNN

1. Multiple region proposals

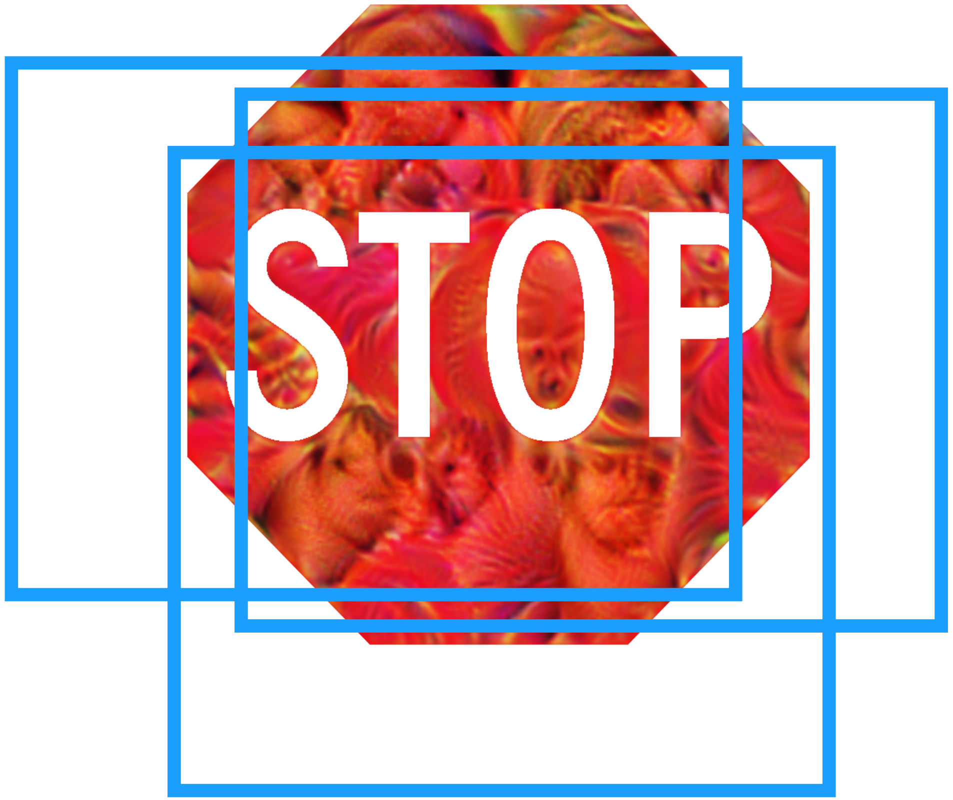


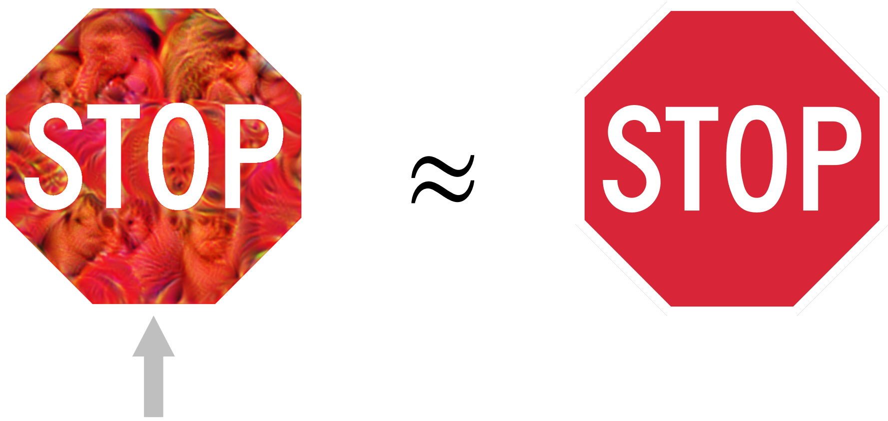
2. Distances, angles, lightings



Our Solution: **Fool Multiple Region Proposals**

Minimize: sum of classification losses + deviation loss





Only perturb **RED** area
Human eye is less sensitive
to changes in darker color

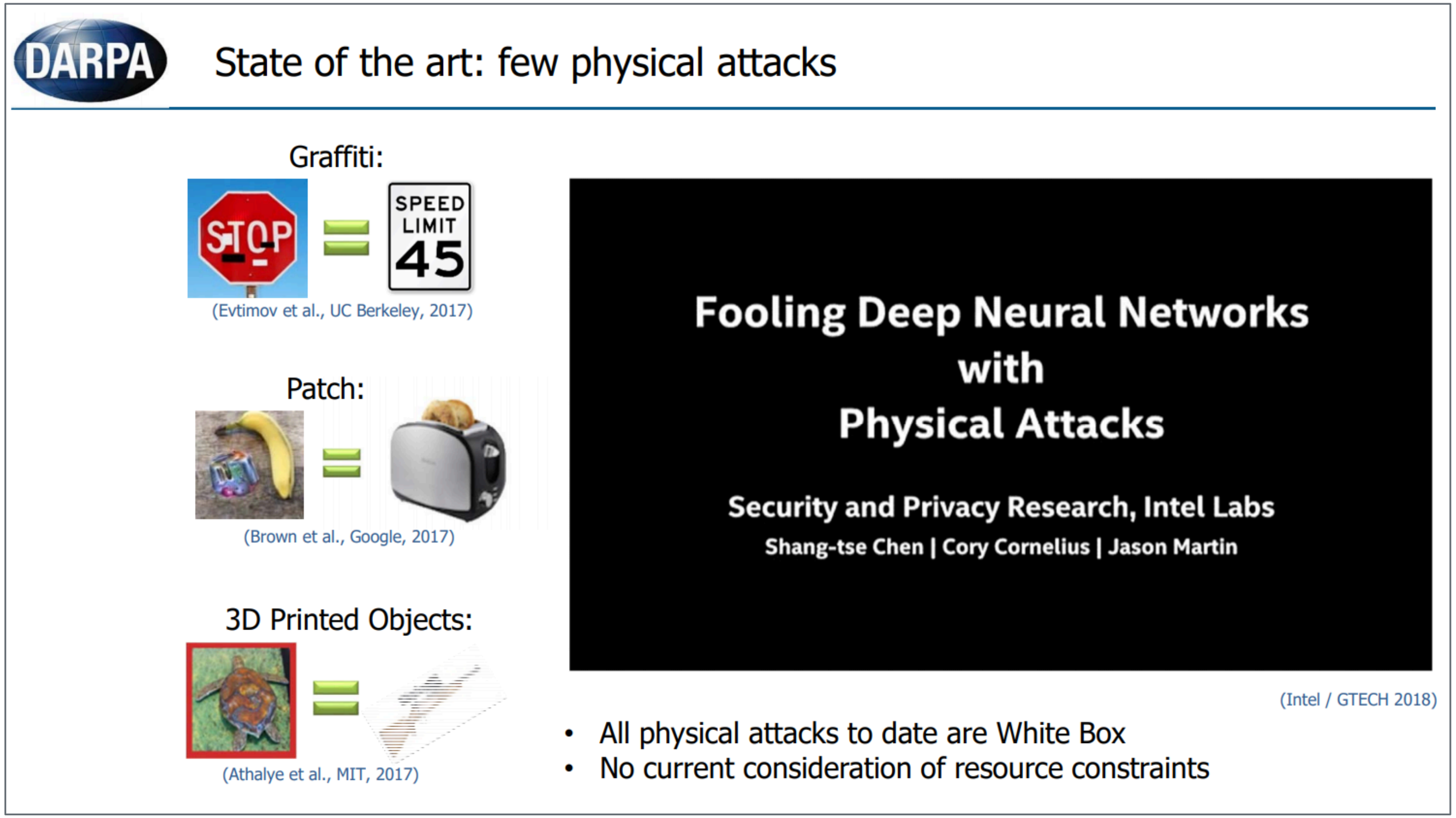
Our Solution: Robust to Real-World Distortions

Adapt Expectation over Transformation [Athalye et al, ICML'18]

Optimize over different backgrounds, scales, rotations, lightings

Untargeted Attack

ShapeShifter Motivates
DARPA Program GARD (Defense for AI)



Highlights **ShapeShifter** as the state-of-the-art physical attack

https://www.darpa.mil/attachments/GARD_ProposersDay.pdf

SHIELD Fast, Practical Defense for Image Classification

KDD’18 Audience Appreciation Award (runner-up)
KDD’19 LEMINCS

[Open-sourced]


Nilaksh Das


Madhuri Shangbogue


Shang-Tse Chen


Fred Hohman


Siwei Li


Cory Cornelius


Li Chen

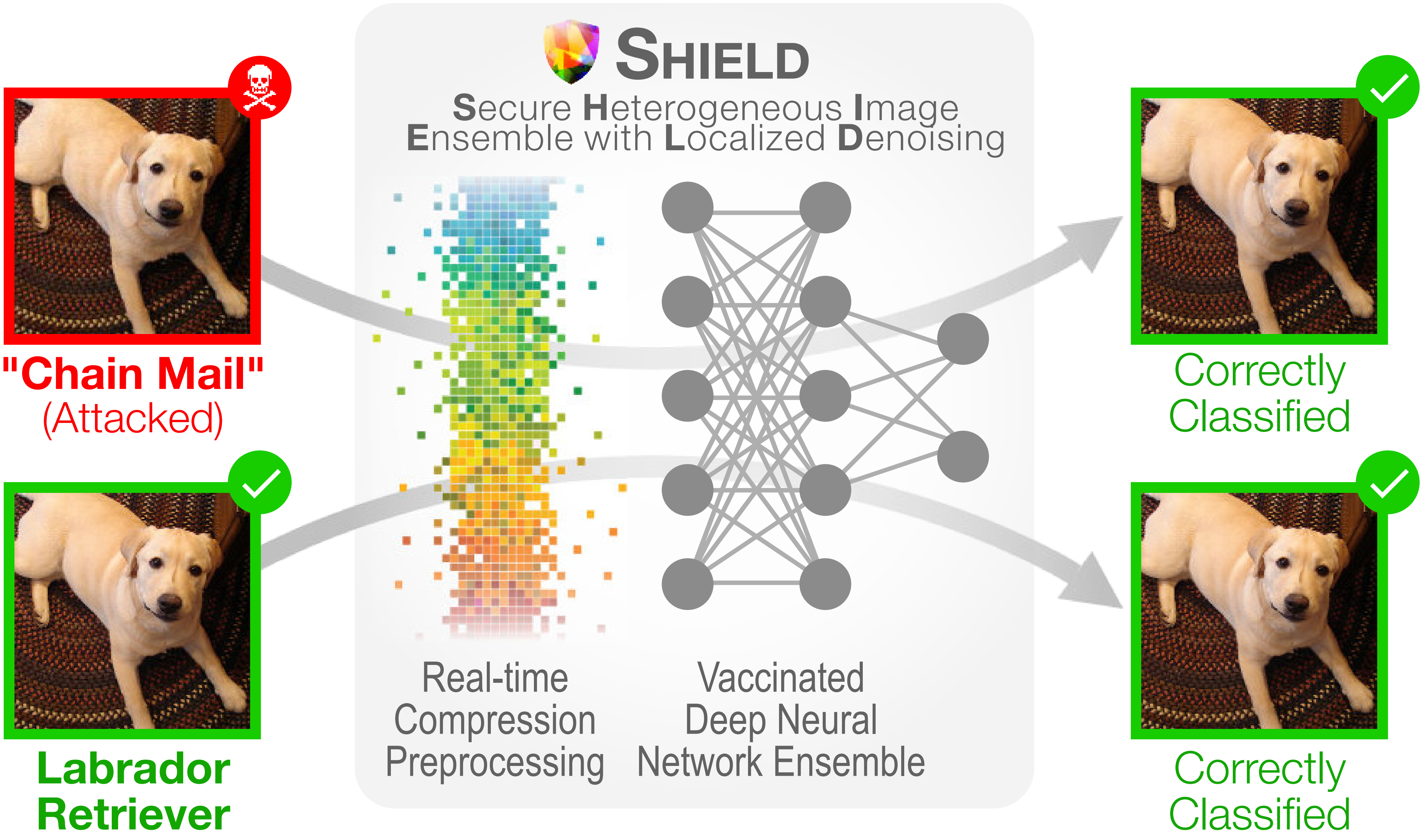

Michael Kounavis


Polo Chau

Georgia Tech

intel

Adversarial Machine Learning Landscape



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

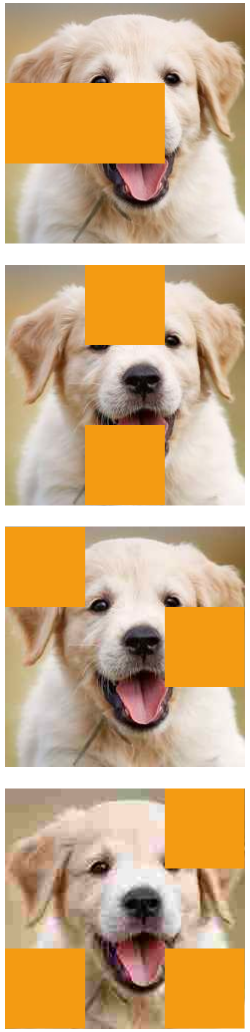
SHIELD leverages JPEG compression

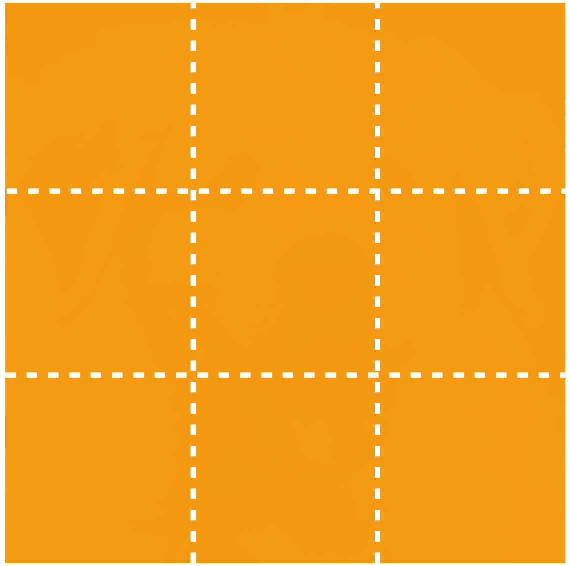
JPEQ Quality 80

JPEQ Quality 60

JPEQ Quality 40

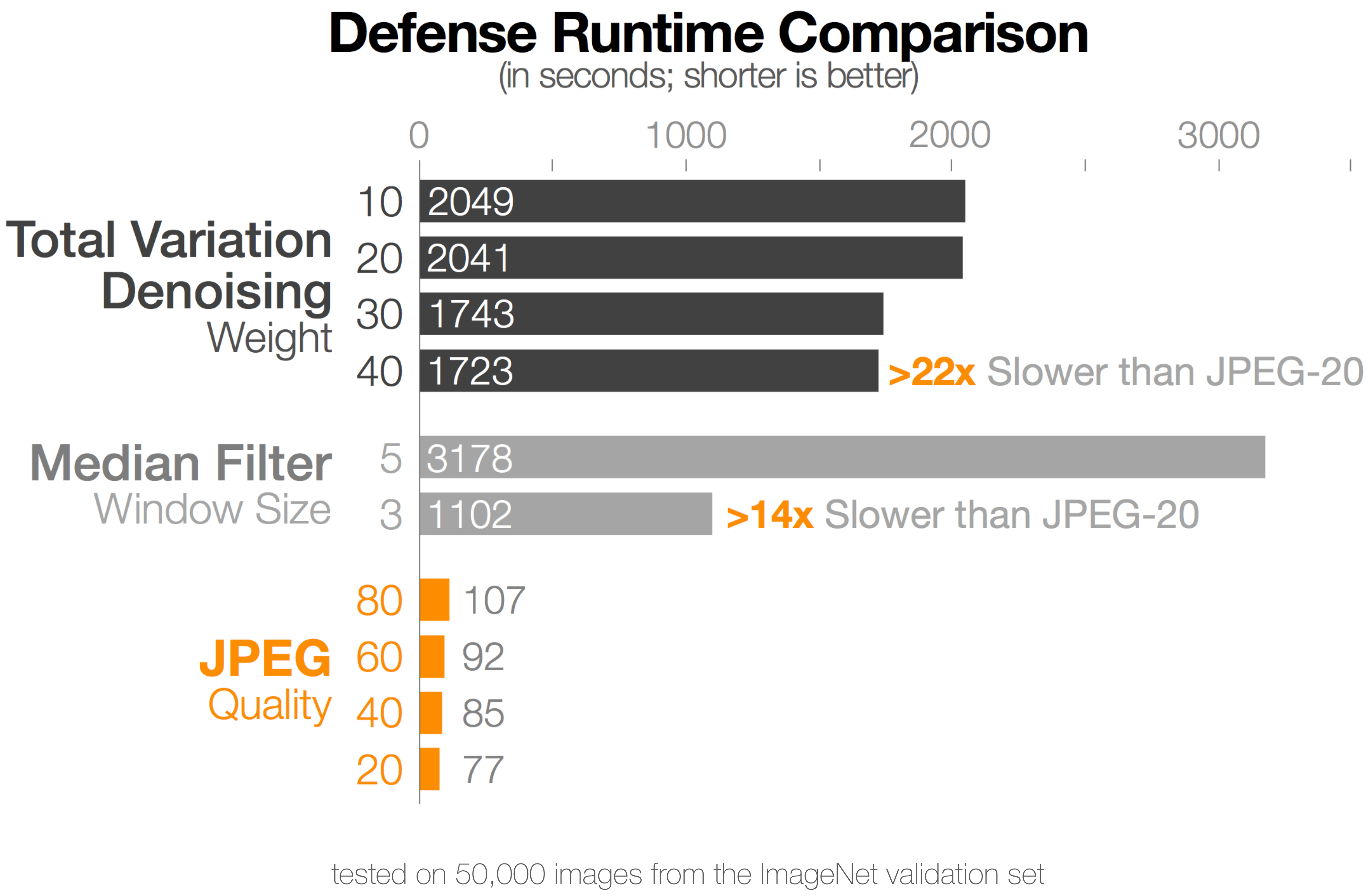
JPEQ Quality 20





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

* larger blocks shown for presentation


Defense Runtime Comparison
(in seconds; shorter is better)
0
1000
2000
3000
Total Variation Denoising
Weight
10
2049
20
2041
30
1743
40
1723
>22x Slower than JPEG-20
Median Filter
Window Size
5
3178
3
1102
>14x Slower than JPEG-20
JPEG
Quality
80
107
60
92
40
85
20
77
tested on 50,000 images from the ImageNet validation set

[PKDD18]

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

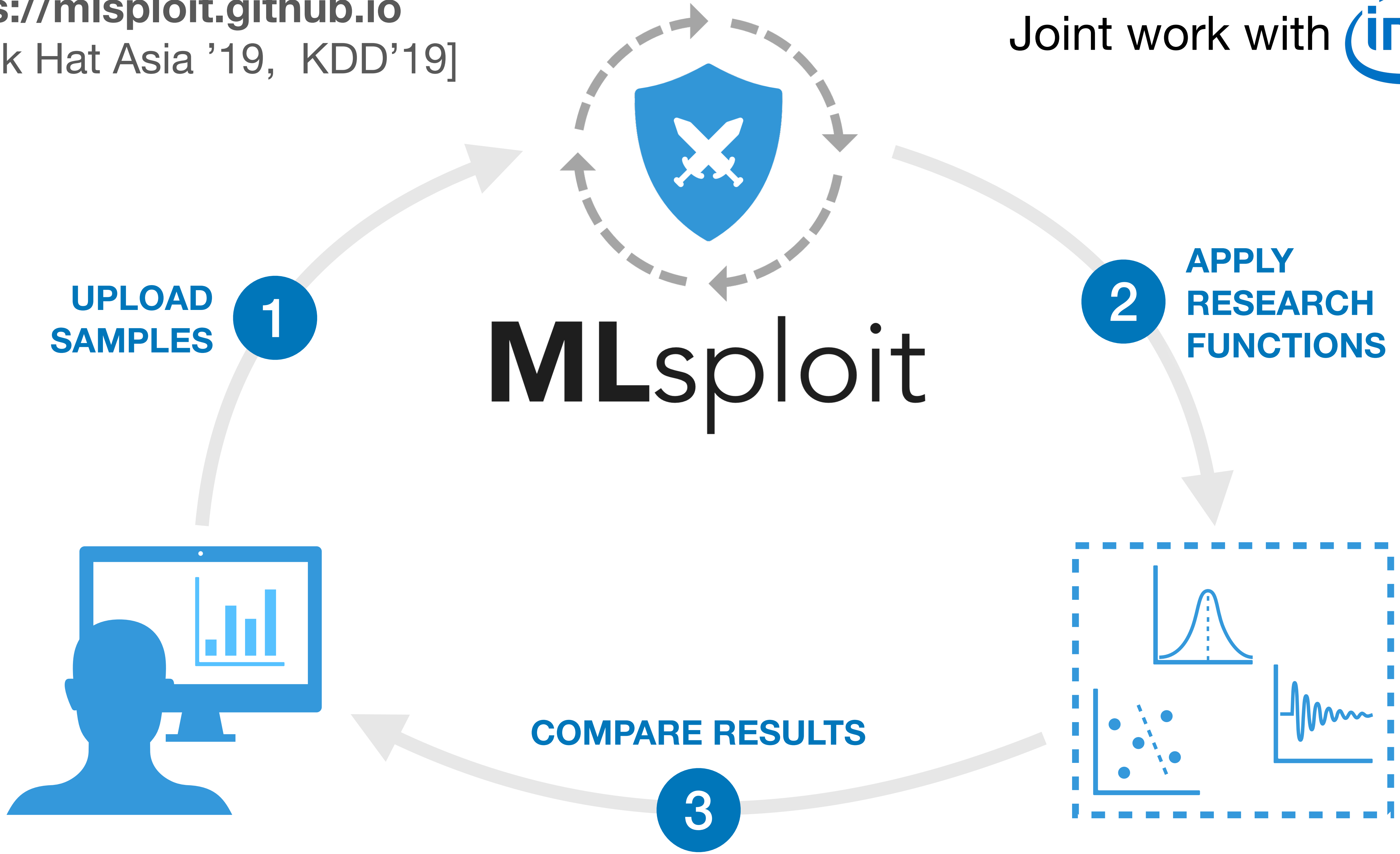
ADAGIO = **A**ttack & **D**efense for **A**udio in a **G**adget with **I**nteractive **O**perations

ADAGIO

https://mlsploit.github.io

[Black Hat Asia ’19, KDD’19]

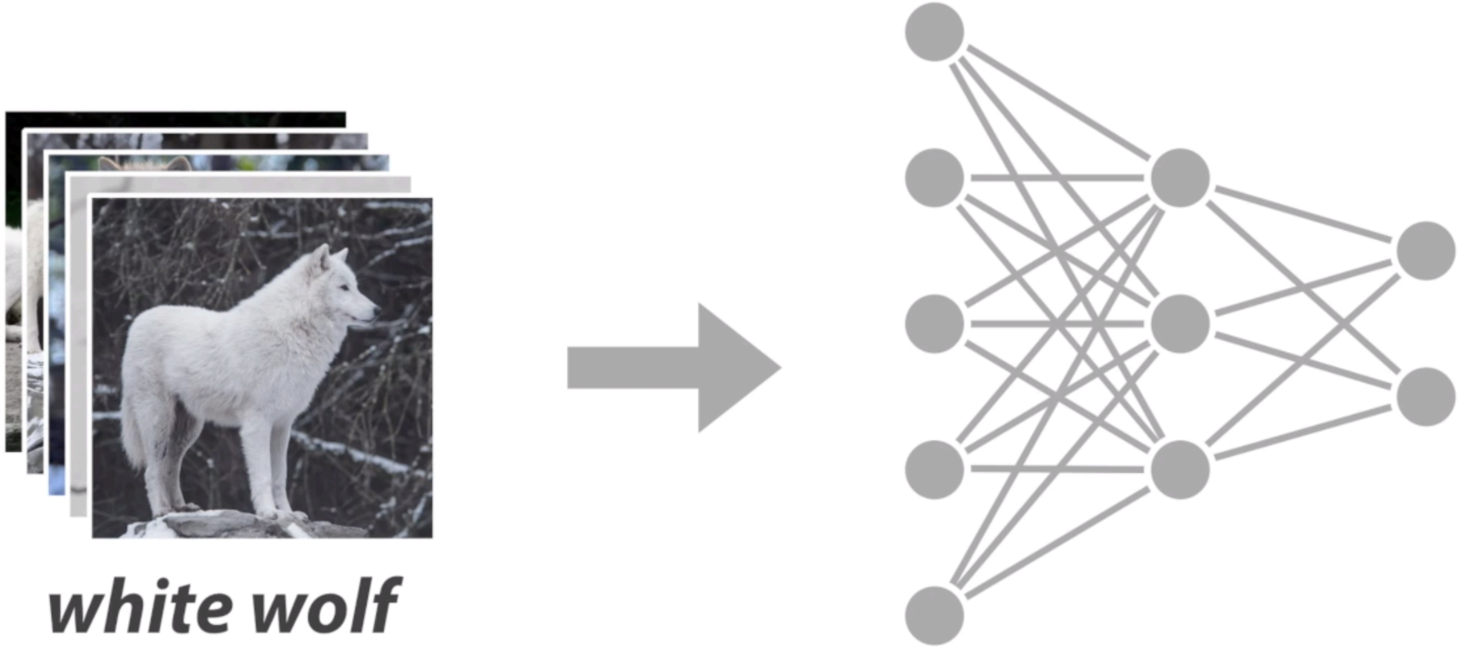
Joint work with intel



SUMMIT

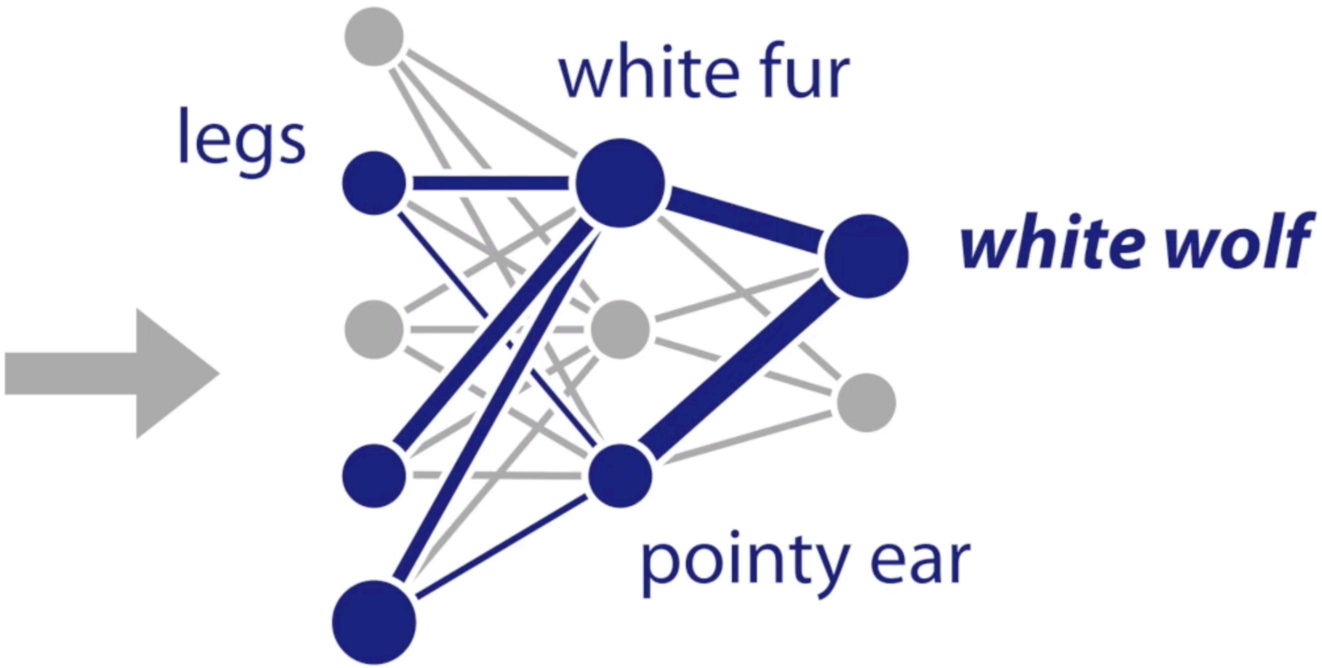
IEEE VIS 2019

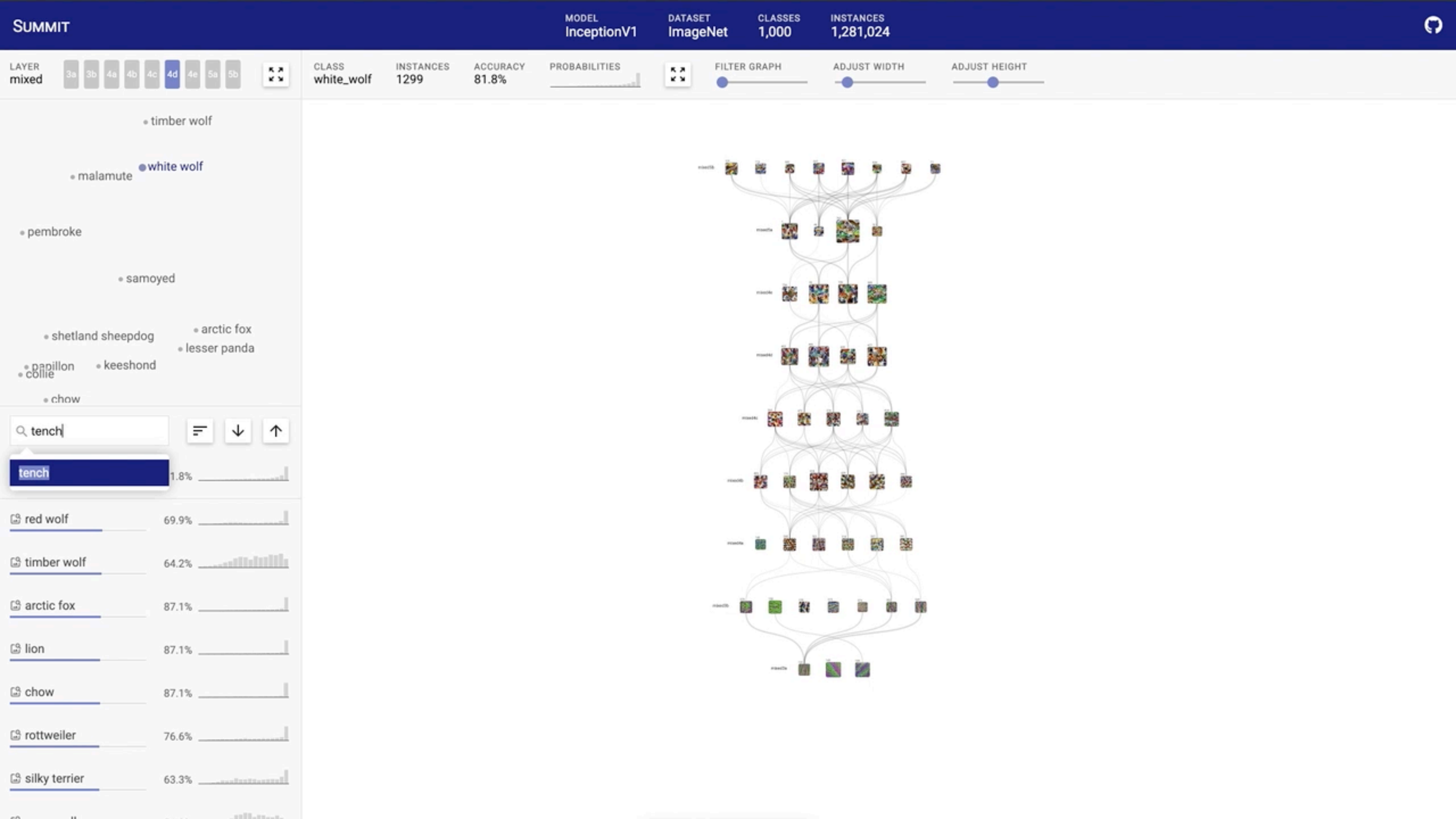
Scalably summarize and **interactively visualize** neural network feature representations for millions of images

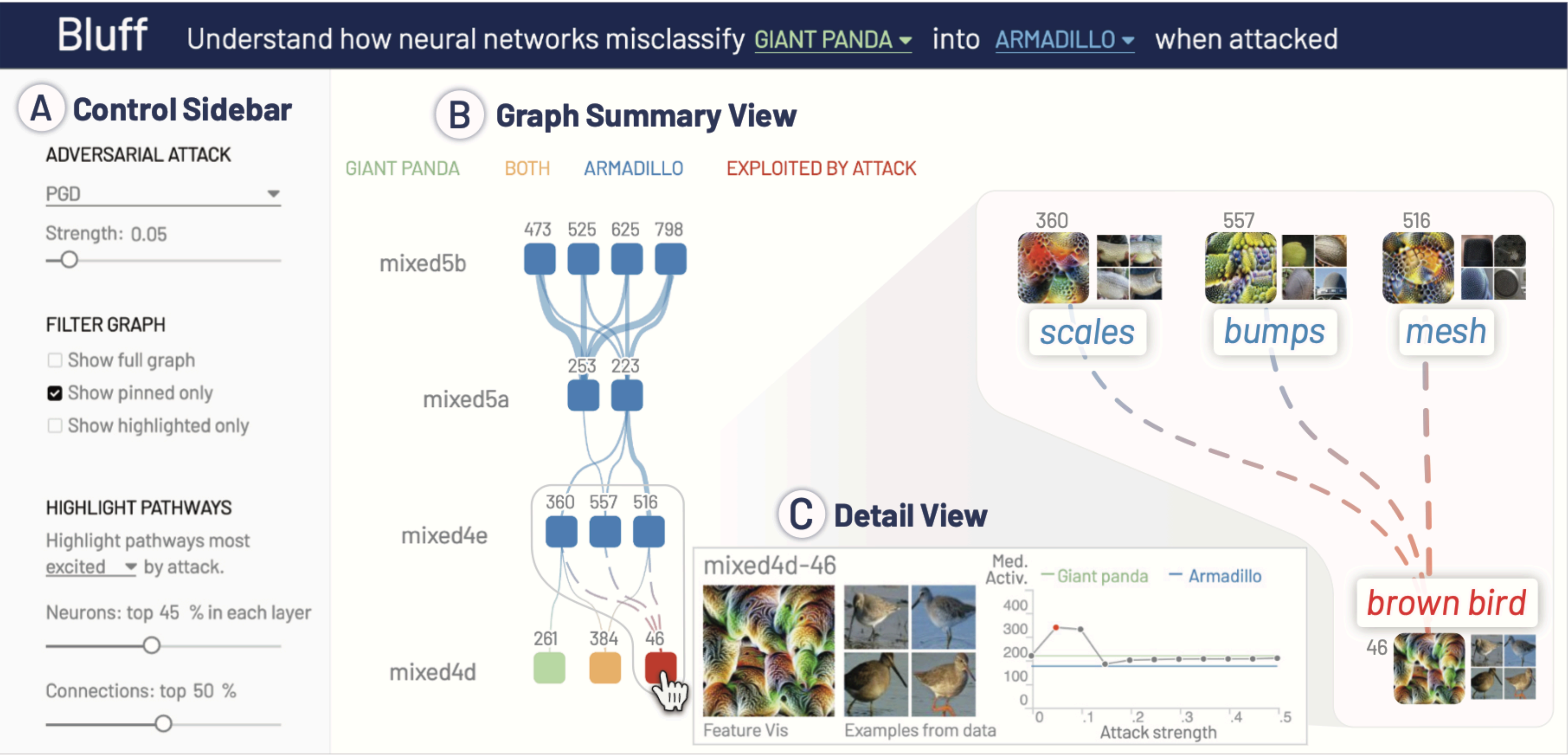


SUMMIT

Scalably summarize and **interactively visualize** neural network feature representations for millions of images




SUMMIT
MODEL InceptionV1
DATASET ImageNet
CLASSES 1,000
INSTANCES 1,281,024
LAYER mixed
3a 3b 4a 4b 4c 4d 4e 5a 5b
CLASS white_wolf
INSTANCES 1299
ACCURACY 81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
timber wolf
white wolf
malamute
pembroke
samoyed
shetland sheepdog
arctic fox
lesser panda
papillon
keeshond
collie
chow
tench
tench
1.8%
red wolf 69.9%
timber wolf 64.2%
arctic fox 87.1%
lion 87.1%
chow 87.1%
rottweiler 76.6%
silky terrier 63.3%

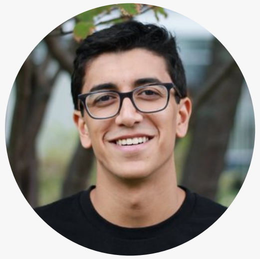


Bluff
Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked
A Control Sidebar
ADVERSARIAL ATTACK
PGD
Strength: 0.05
FILTER GRAPH
Show full graph
Show pinned only
Show highlighted only
HIGHLIGHT PATHWAYS
Highlight pathways most excited by attack.
Neurons: top 45 % in each 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 scales
557 bumps
516 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

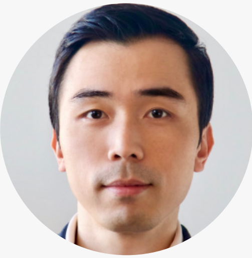
GAN Lab

Understanding Complex Deep Generative Models using Interactive Visual Experimentation

Minsuk Kahng
Georgia Tech



Nikhil Thorat
Google



Polo Chau
Georgia Tech

Fernanda Viégas
Google

Martin Wattenberg
Google

Georgia Tech

Google AI
PAIR | People + AI Research Initiative

Generative Adversarial Networks (GANs)

"the most interesting idea in the last 10 years in ML"

- Yann LeCun

Face images generated by BEGAN [Berthelot et al., 2017]

Why GANs are hard?

A GAN uses two *competing* neural networks

Generator
synthesizes outputs

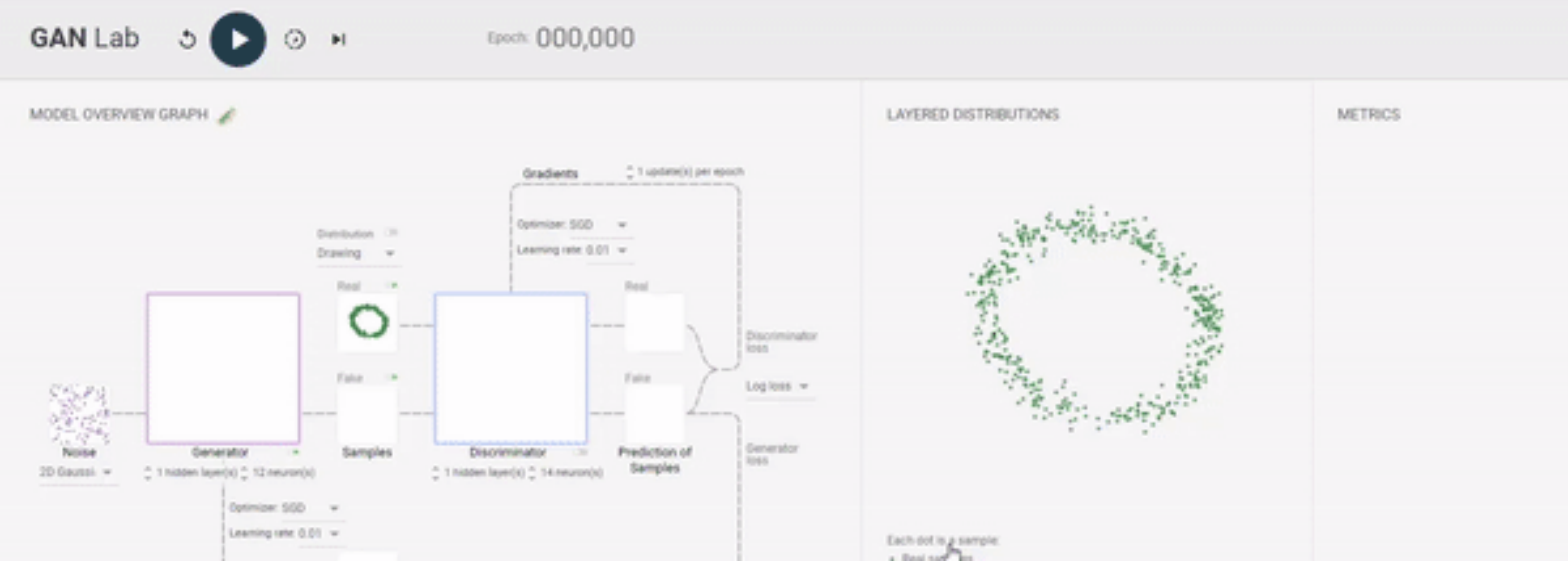
Discriminator
spots fake

Counterfeiter
makes fake bills

Police
spots fake bills

GAN Lab is Live! Try at **bit.ly/gan-lab**

1.9K Likes 800+ Retweets 30K visitors, 135 countries

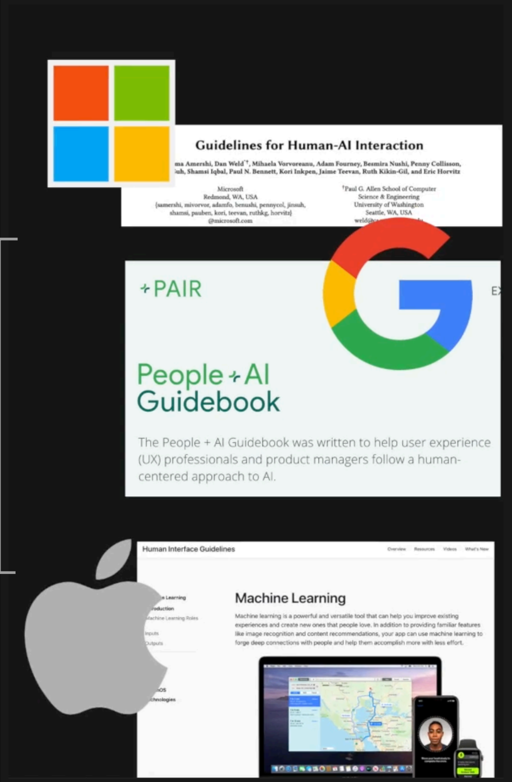


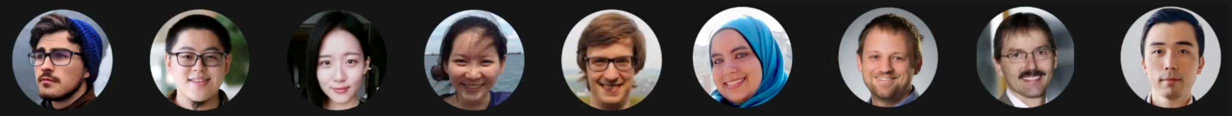
CNN Explainer also went viral! Try at **bit.ly/cnn-explainer**

5.3K GitHub Stars 700 Likes 36K visitors, 151 countries

A Comparative Analysis of

Industry Human-AI Interaction Guidelines





Austin P Wright | Zijie J. Wang | Haekyu Park | Grace Guo | Fabian Sperrle | Mennatallah El-Assady | Alex Endert | Daniel Keim | Polo Chau

TREX Workshop @ IEEE VIS 2020

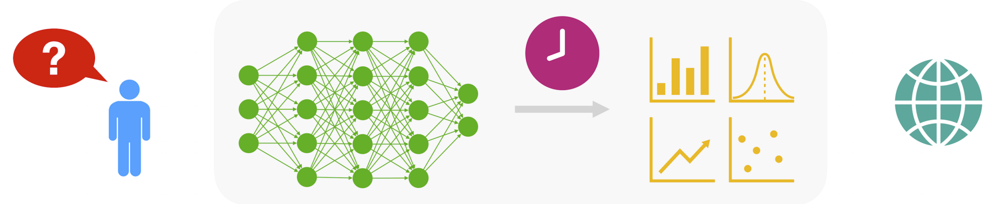
An Interrogative Survey for the Next Frontiers

Visual Analytics in Deep Learning

§4 WHY
Why would one want to use visualization in deep learning?
Interpretability & Explainability
Debugging & Improving Models
Comparing & Selecting Models
Teaching Deep Learning Concepts

§6 WHAT
What data, features, and relationships in deep learning can be visualized?
Computational Graph & Network Architecture
Learned Model Parameters
Individual Computational Units
Neurons In High-dimensional Space
Aggregated Information

§8 WHEN
When in the deep learning process is visualization used?
During Training
After Training



§5 WHO
Who would use and benefit from visualizing deep learning?
Model Developers & Builders
Model Users
Non-experts

§7 HOW
How can we visualize deep learning data, features, and relationships?
Node-link Diagrams for Network Architecture
Dimensionality Reduction & Scatter Plots
Line Charts for Temporal Metrics
Instance-based Analysis & Exploration
Interactive Experimentation
Algorithms for Attribution & Feature Visualization

§9 WHERE
Where has deep learning visualization been used?
Application Domains & Models
A Vibrant Research Community

Fred Hohman | **Minsuk** Kahng | **Robert** Pienta | **Polo** Chau

IEEE VIS 2019

poloclub.github.io

SECURE & INTERPRETABLE AI

Thanks!

Polo Chau

Associate Professor
Associate Director, MS Analytics
Associate Director of Corporate Relations, ML Center
Georgia Tech

Fred	Nilaksh	Haekyu	Scott	Jay	Austin	Rahul	Anmol	Bob	Jonathan	Will	Rob	Omar	Frank	Jon	Robert	Dongkyu
CSE PhD	CSE PhD	CS PhD	ML PhD	ML PhD	ML PhD	CS PhD	MS CSE	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad	Post-Doc.