

# SECURE & INTERPRETABLE AI



**Polo Chau**

Associate Professor

Associate Director, MS Analytics

Associate Director of Corporate Relations, ML Center

Georgia Tech



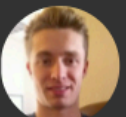
Fred  
CSE PhD



Nilaksh  
CSE PhD



Haekyu  
CS PhD



Scott  
ML PhD



Jay  
ML PhD



Austin  
ML PhD



Rahul  
CS PhD



Anmol  
MS CSE



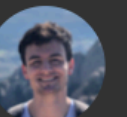
Bob  
CS Undergrad



Jonathan  
CS Undergrad



Will  
CS Undergrad



Rob  
CS Undergrad



Omar  
CS Undergrad



Frank  
CS Undergrad



Jon  
CS Undergrad



Robert  
CS Undergrad



Dongkyu  
Post-Doc.

# Polo Club of Data Science

# AI

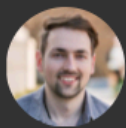
# +

# HI

ARTIFICIAL  
INTELLIGENCE

HUMAN  
INTELLIGENCE

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



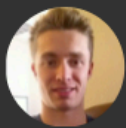
Fred  
CSE PhD



Nilaksh  
CSE PhD



Haekyu  
CS PhD



Scott  
ML PhD



Jay  
ML PhD



Austin  
ML PhD



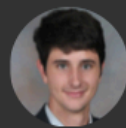
Rahul  
CS PhD



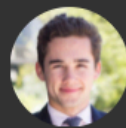
Anmol  
MS CSE



Bob  
CS Undergrad



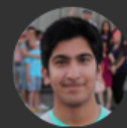
Jonathan  
CS Undergrad



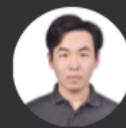
Will  
CS Undergrad



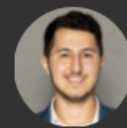
Rob  
CS Undergrad



Omar  
CS Undergrad



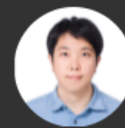
Frank  
CS Undergrad



Jon  
CS Undergrad

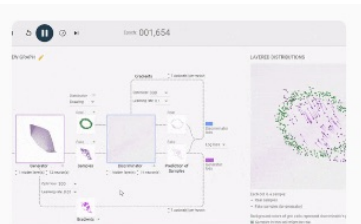


Robert  
CS Undergrad



Dongkyu  
Post-Doc.

## Human-Centered AI



**GAN Lab**  
Playing with Generative Adversarial Networks in Browser

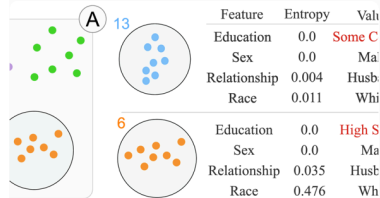
Google



**ActiVis**

Visual Exploration of Facebook Deep Neural Network Models

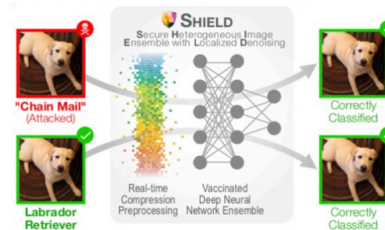
Deployed Facebook



**Discovering Intersectional Bias**

Discovery of Intersectional Bias in Machine Learning Using Automatic Subgroup Generation

## ML Security & Fraud



**SHIELD**

Fast, practical defense for deep learning

Audience Appreciation Award, Runner-up



**ShapeShifter**

1st Targeted Physical Attack on Faster R-CNN Object Detector



**MARCO**

Fake Review Detection  
SDM'14 Best Student Paper

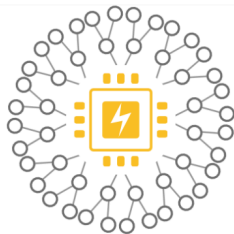
## Large Graph Mining & Visualization



**VIGOR**

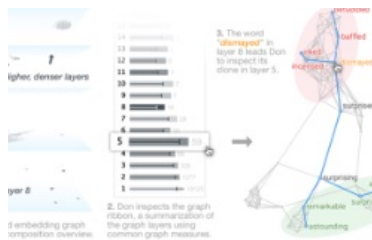
Interactive Visual Exploration of Graph Query Results

Symantec



**M-Flash**

Billion-Scale Graph Computation by Bimodal Block Processing



**Atlas**

Local Graph Exploration in a Global Context

## Social Good & Health



**DeepPop**

Deep Learning on Satellite Imagery for Population Estimation

Microsoft AI for Earth



**Firebird**

Predicting Fire Risk in Atlanta

KDD'16 Best Student Paper, runner-up

Deployed Atlanta Fire Rescue Department



**mHealth Visual Discovery Dashboard**

Making Sense of Mobile Health Data

# Current Research Thrusts

**Secure**<sub>AI</sub>

**Interpretable**<sub>AI</sub>

Why focus on them?  
How are they related?



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

An aerial photograph of a street scene. A white car is visible on the road, moving north. The surrounding area includes buildings, trees, and a sidewalk.

**Secure**<sub>AI</sub>

New York Times, 2018

The self-driving Uber  
was traveling north at  
about 40 m.p.h.

# How a Self-Driving Uber Killed a Pedestrian in Arizona

# AI Security Problems Are Everywhere

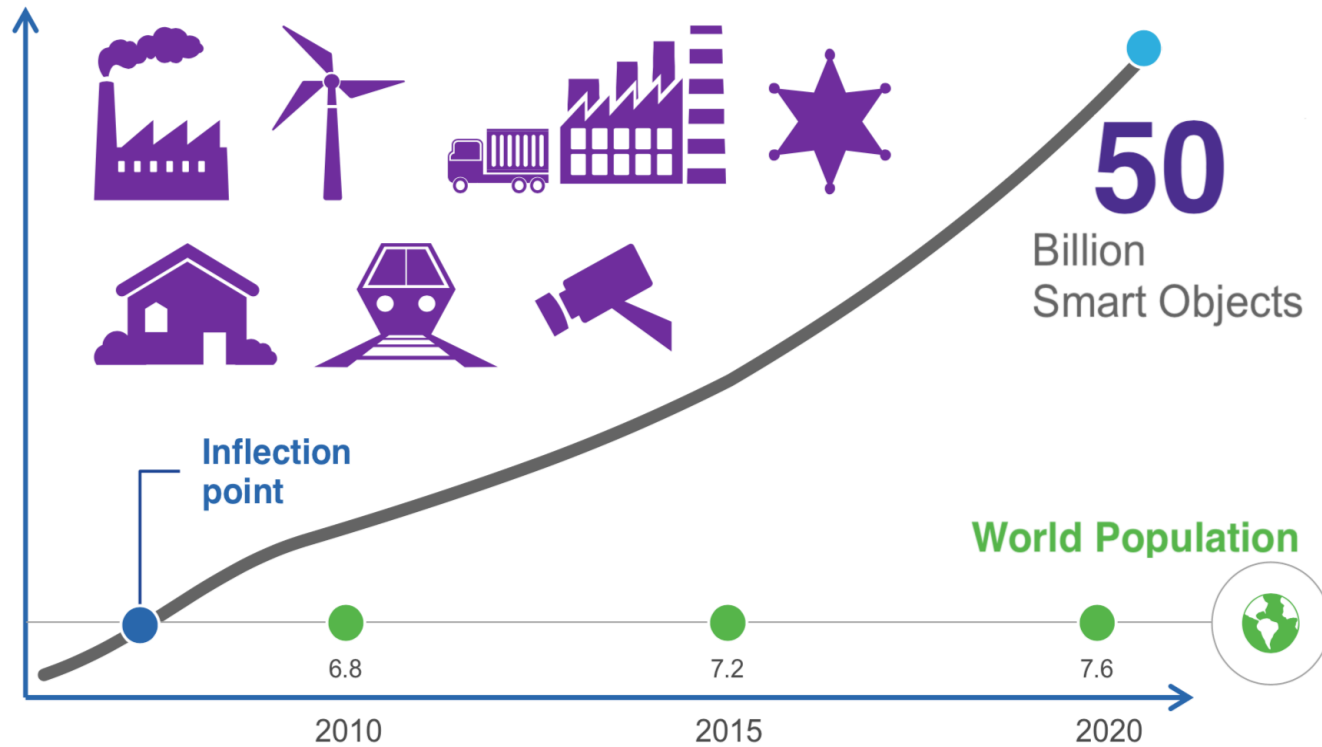


"THE TOASTER HAS BEEN HACKED INTO THINKING IT'S A BLENDER."



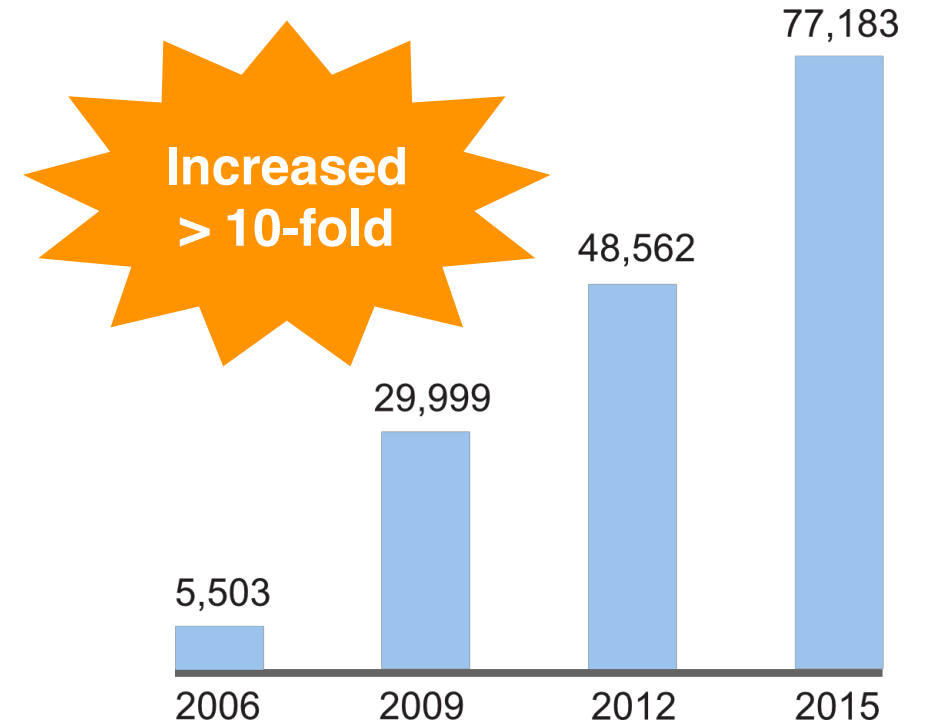
Smart toaster does exist!

# AI Security is becoming increasingly important



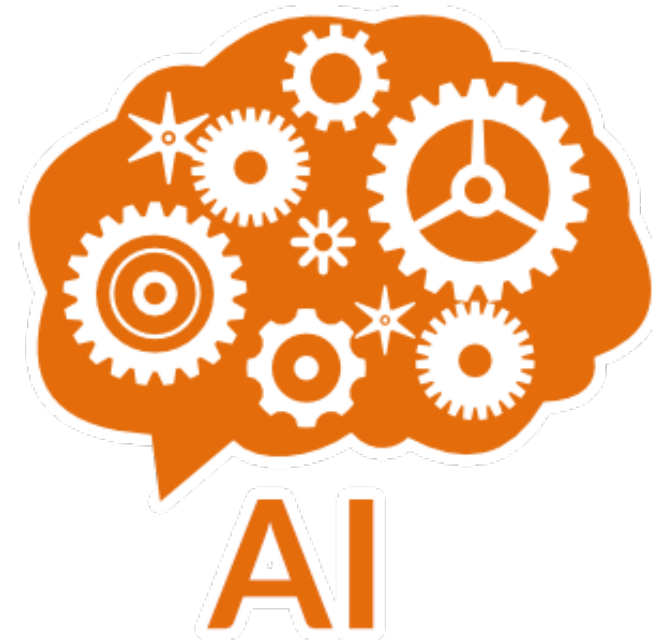
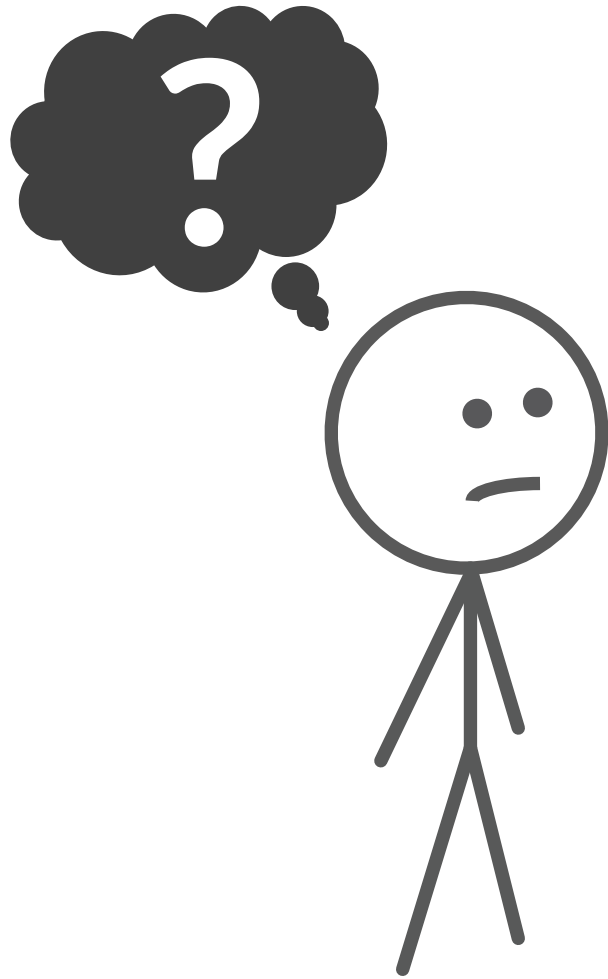
Source: Cisco

# incidents  
reported by U.S. federal agencies

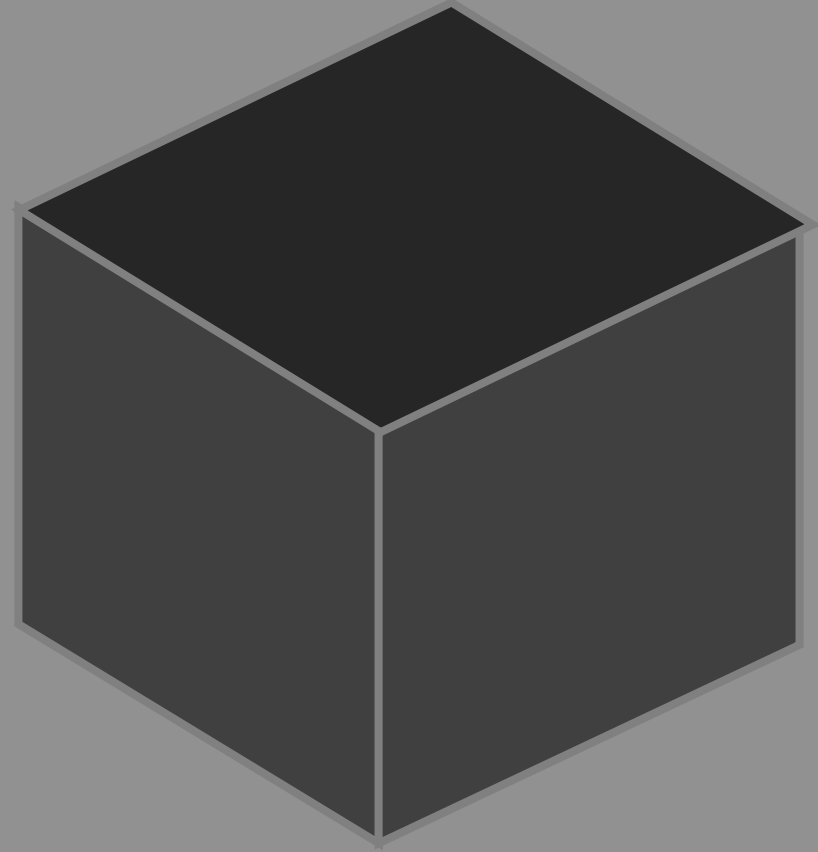
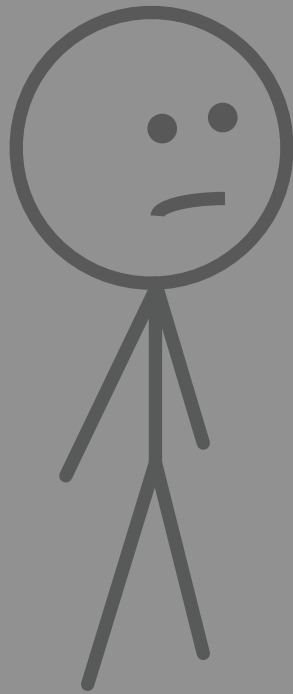


Source: US Department of Homeland Security

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

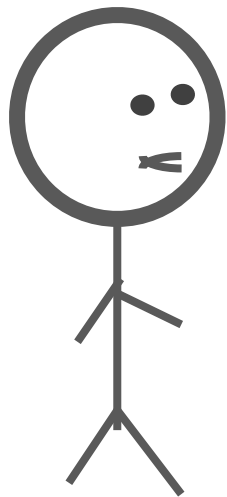


# AI models often used as black-box





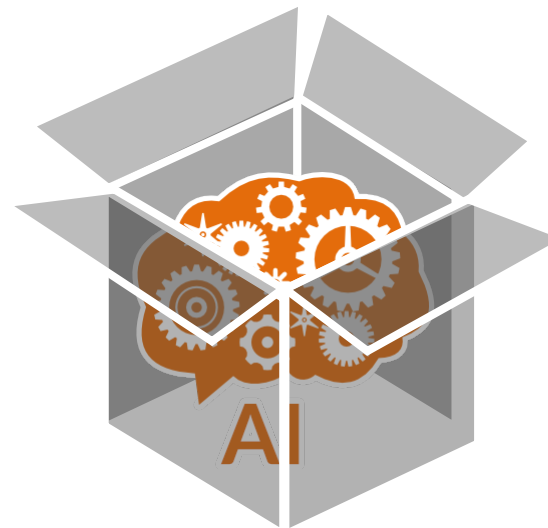
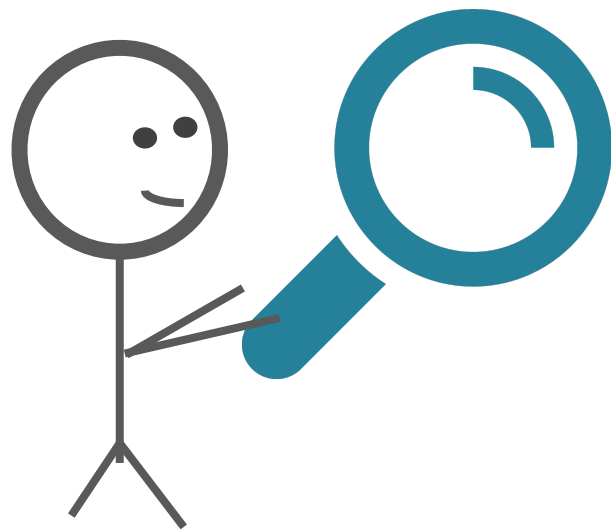
# Interpretable<sub>AI</sub>





# Interpretable<sub>AI</sub>

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



## Secure AI

**ShapeShifter** First attack fooling object detectors

**SHIELD** Real-time Defense

**REST** Energy efficient, noise-robust sleep tracking

## Interpretable AI

**Summit** Scalable interpretation for deep learning

**GAN Lab & CNN Explainer** Interactive learning

# ShapeShifter

ECML-PKDD 2018

First Targeted *Physical*  
Adversarial Attack  
for Object Detection



**Shang-Tse  
Chen**

Georgia Tech



**Cory  
Cornelius**

Intel



**Jason  
Martin**

Intel



**Polo  
Chau**

Georgia Tech



Stop Sign → Person

Real Stop Sign

car: 89%



car: 89%



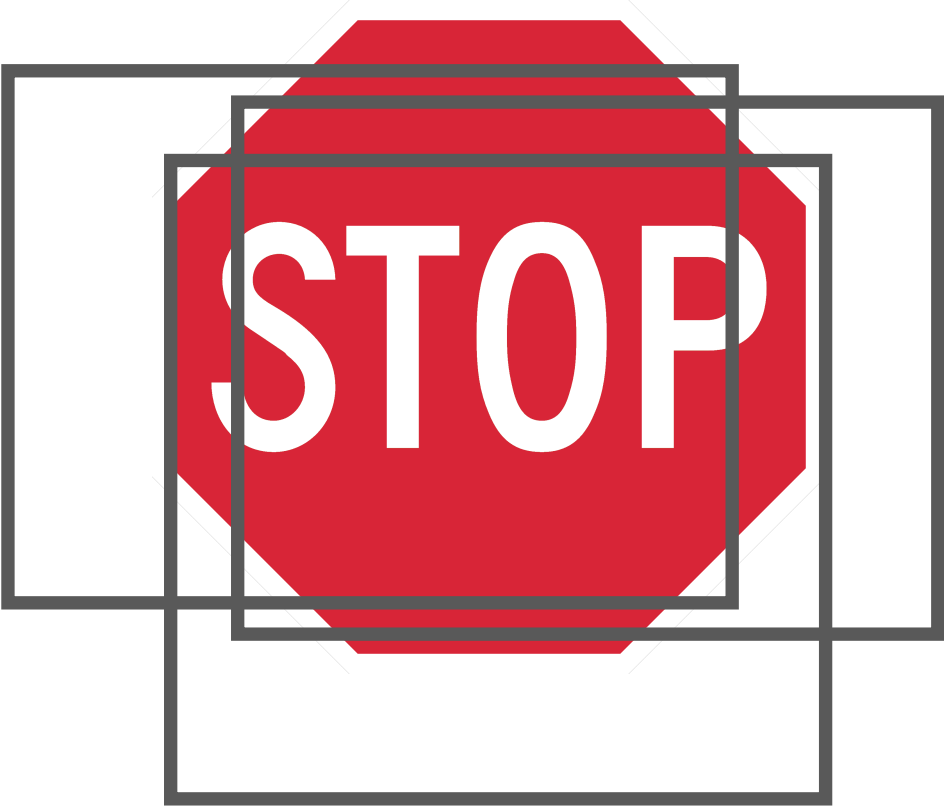
stop sign: 60%



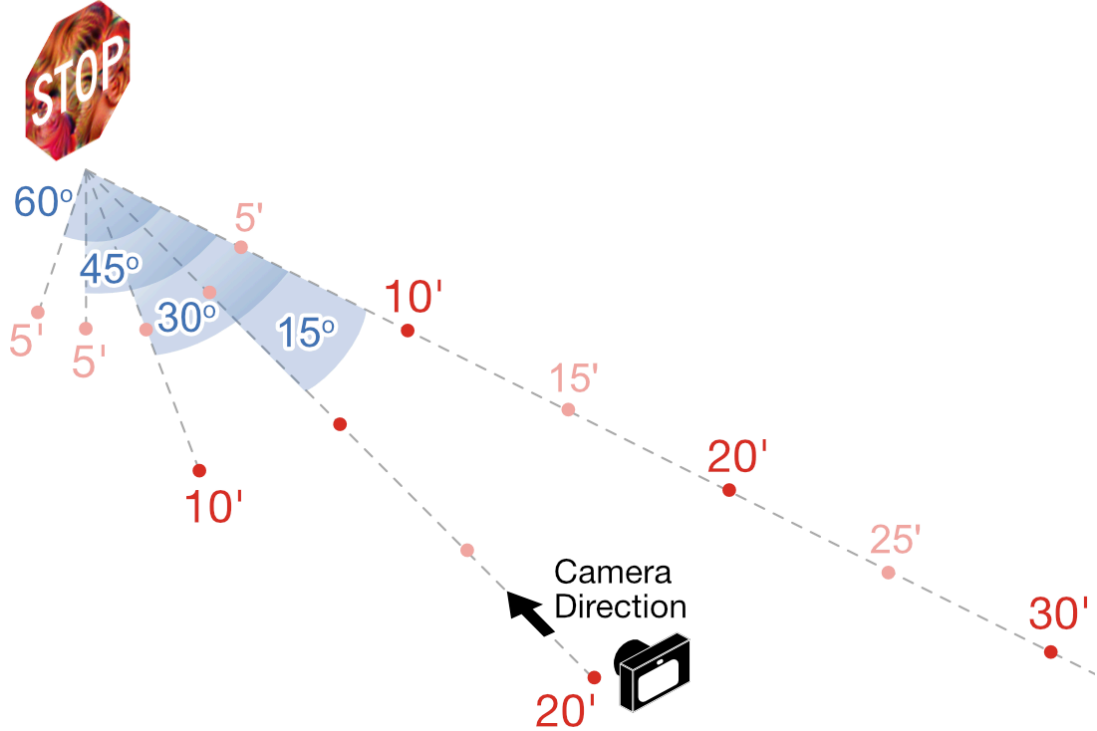
Printed Adversarial Stop Sign

# Challenges of **Physically Attacking** **Faster R-CNN**

1. Multiple region proposals



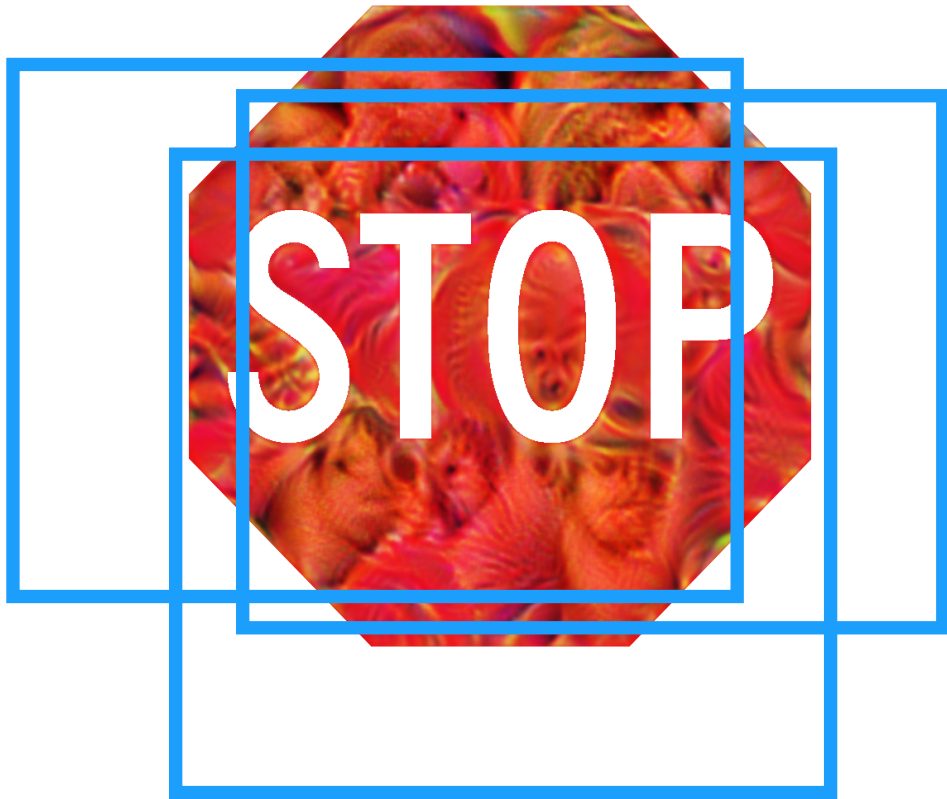
2. Distances, angles, lightings





# 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)



State of the art: few physical attacks

Graffiti:



(Evtimov et al., UC Berkeley, 2017)

Patch:

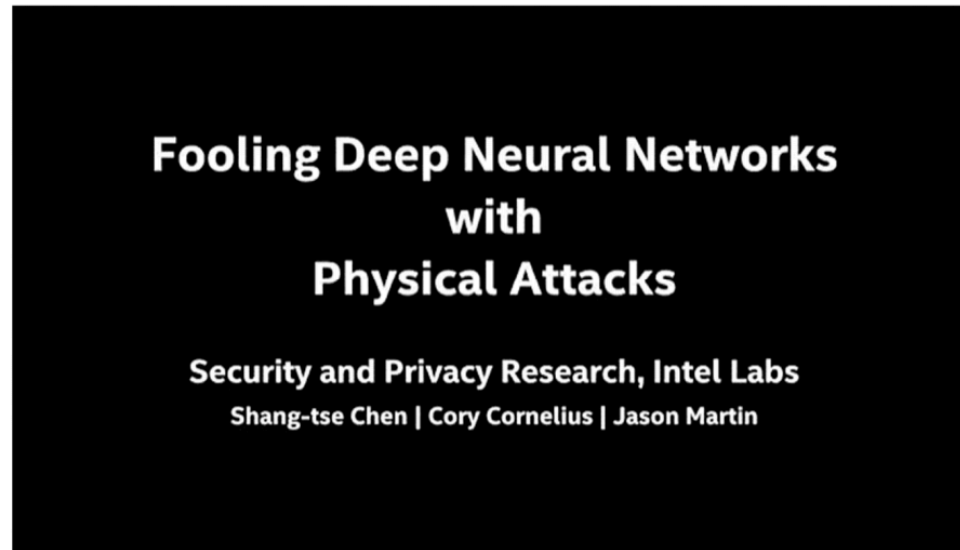


(Brown et al., Google, 2017)

3D Printed Objects:



(Athalye et al., MIT, 2017)



(Intel / GTECH 2018)

- All physical attacks to date are White Box
- No current consideration of resource constraints

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



# 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



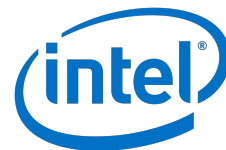
Li  
Chen



Michael  
Kounavis



Polo  
Chau



# Adversarial Machine Learning Landscape



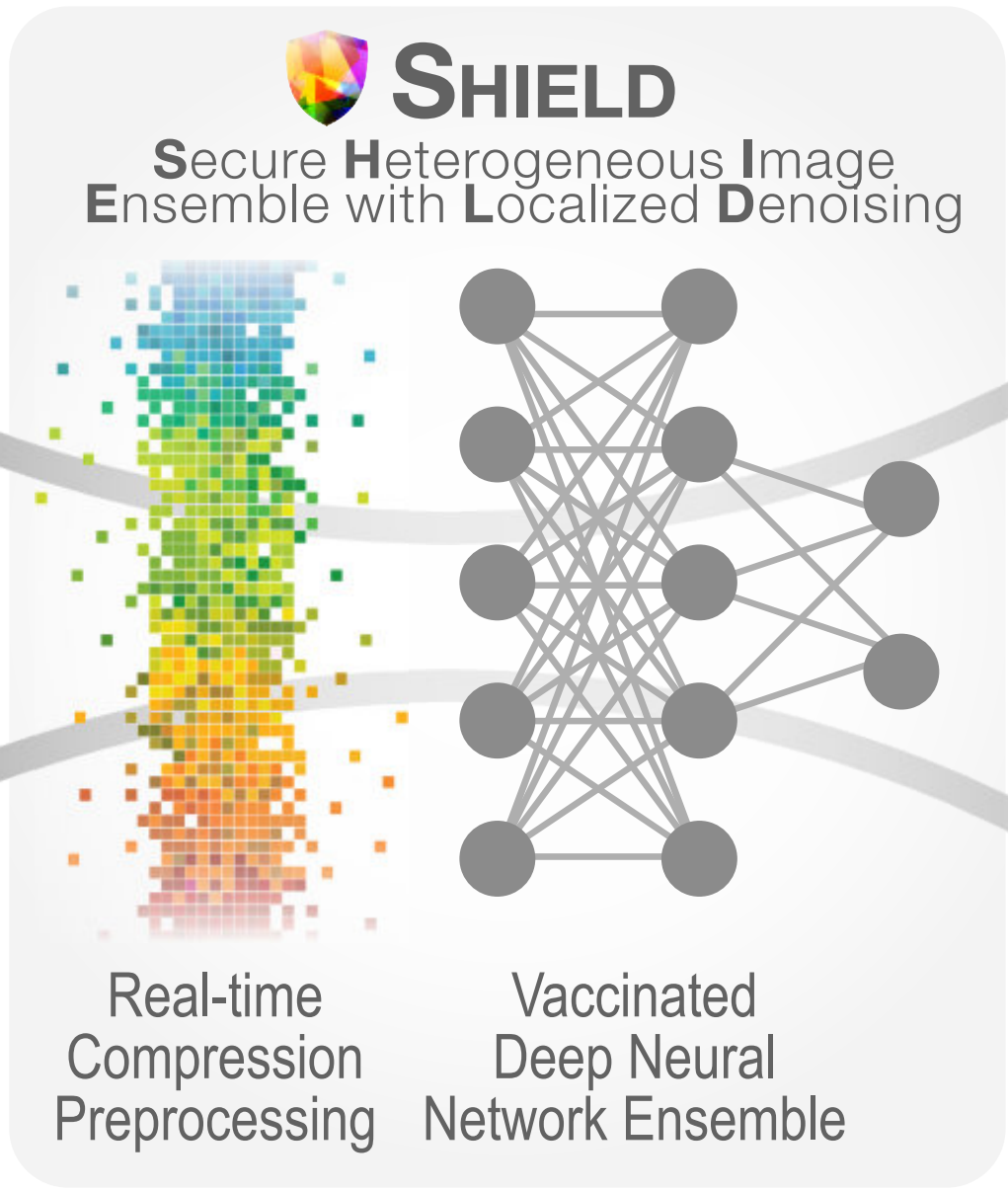
Our Focus:  
**Fast & Practical**  
(digital)



**"Chain Mail"**  
(Attacked)



**Labrador  
Retriever**



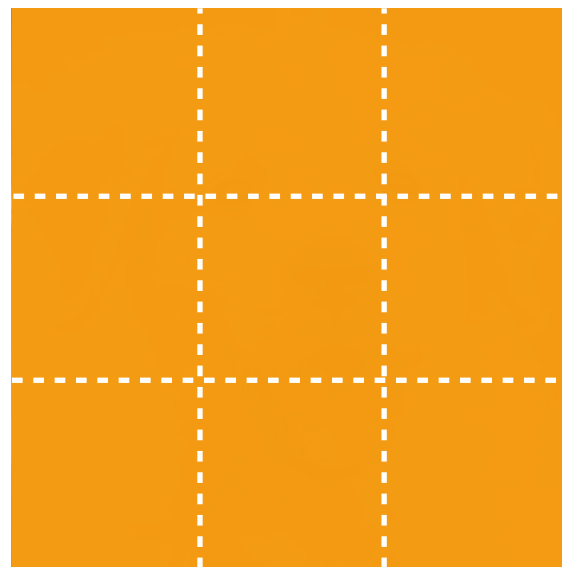
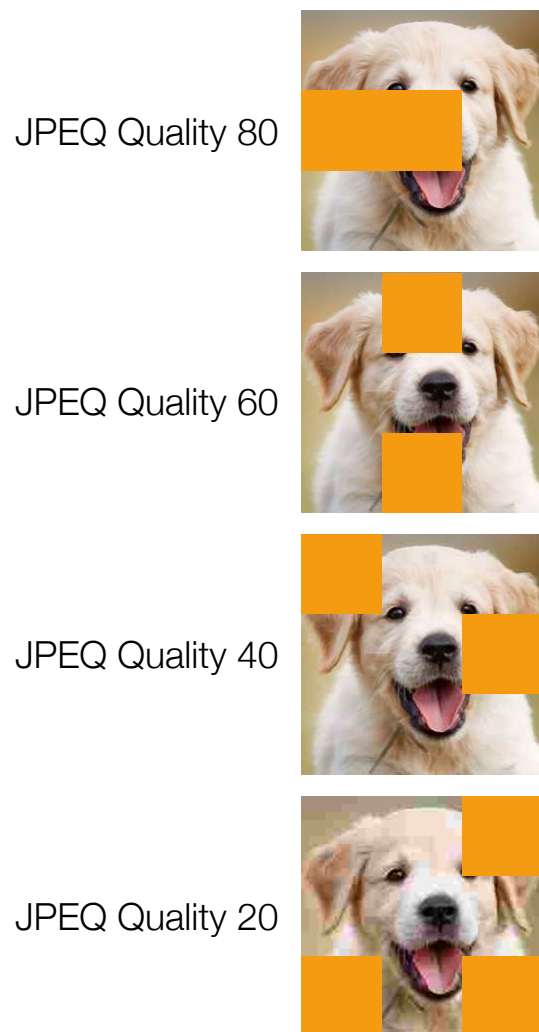
Correctly  
Classified



Correctly  
Classified



# SHIELD leverages JPEG compression

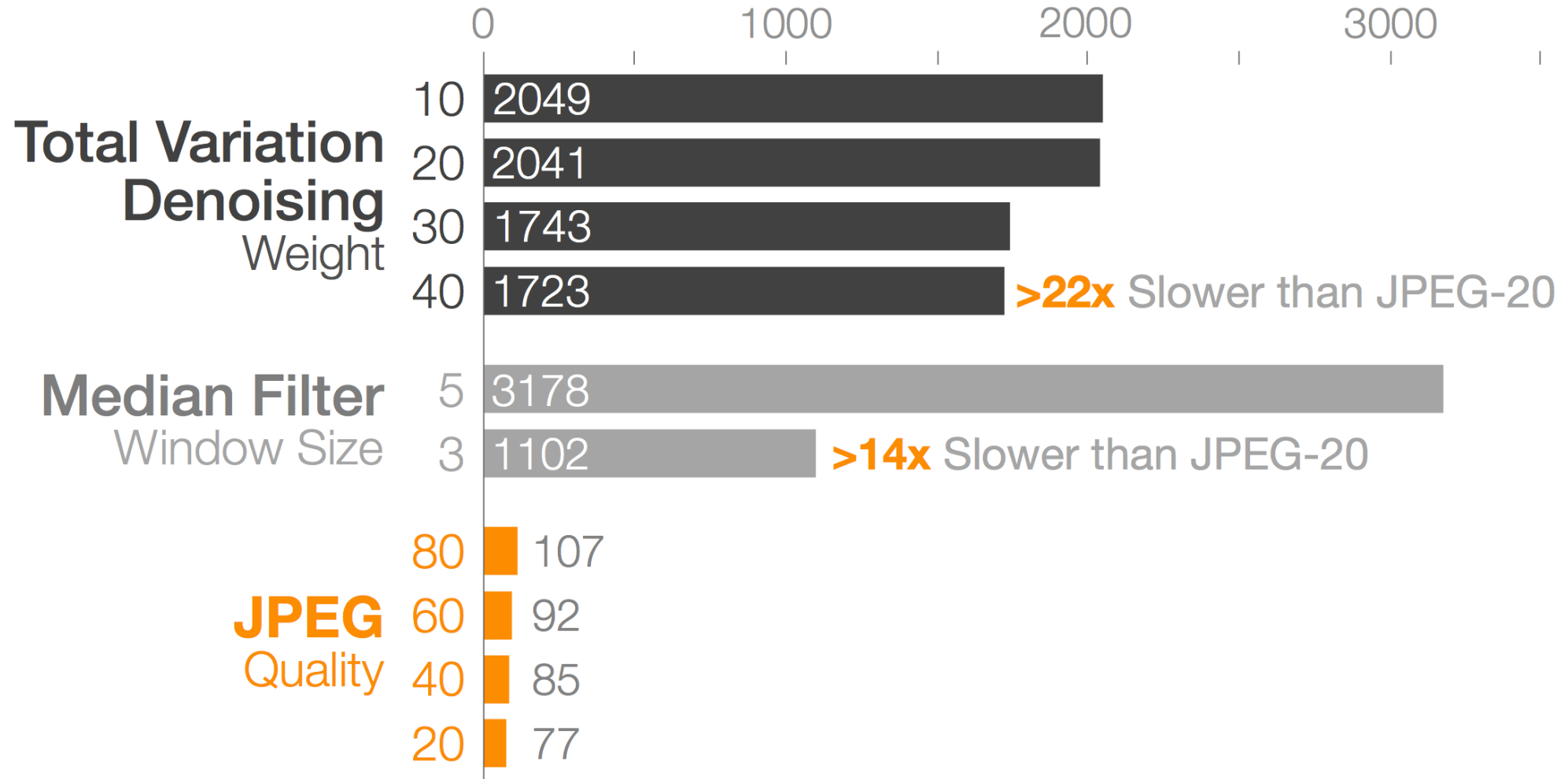


SHIELD's **SLQ** applies JPEG compression of a random quality to each 8 x 8 block of the image

\* larger blocks shown for presentation

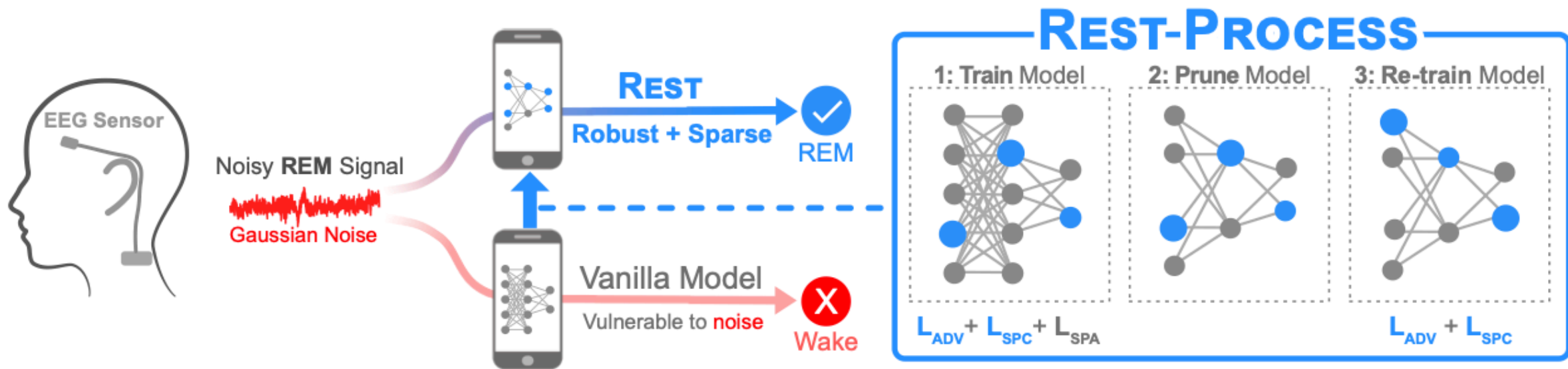
# Defense Runtime Comparison

(in seconds; shorter is better)



tested on 50,000 images from the ImageNet validation set

# REST: energy-efficient, noise-robust sleep tracking



## Efficiency Measurements

Energy Usage (J)	1143	SOTA (state-of-the-art)
	123	<b>Rest is 9x more efficient</b>
Inference Time (s)	355	SOTA
	57	<b>Rest is 6x faster</b>

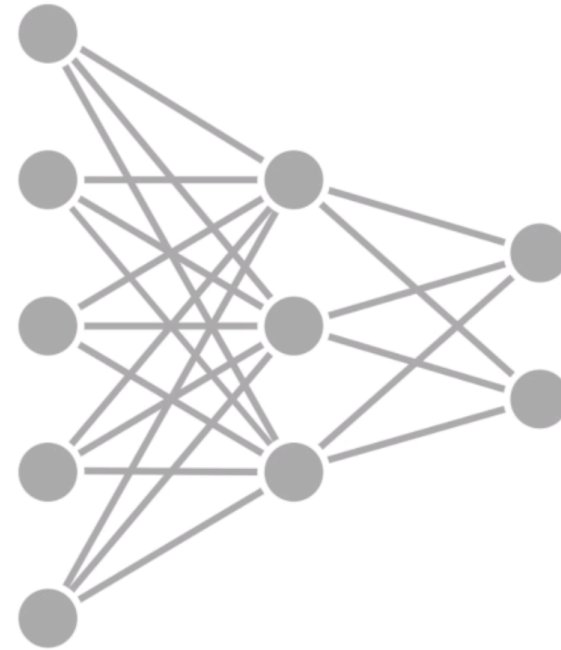
# SUMMIT

IEEE VIS 2019

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



*white wolf*

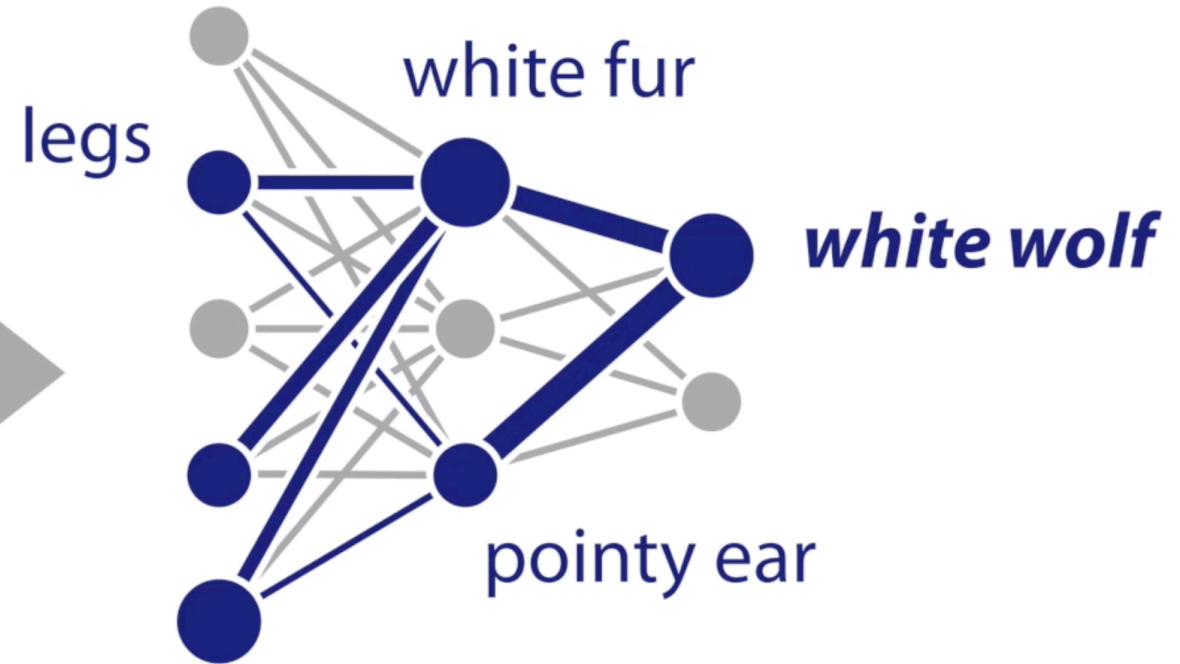


# SUMMIT

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



*white wolf*





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

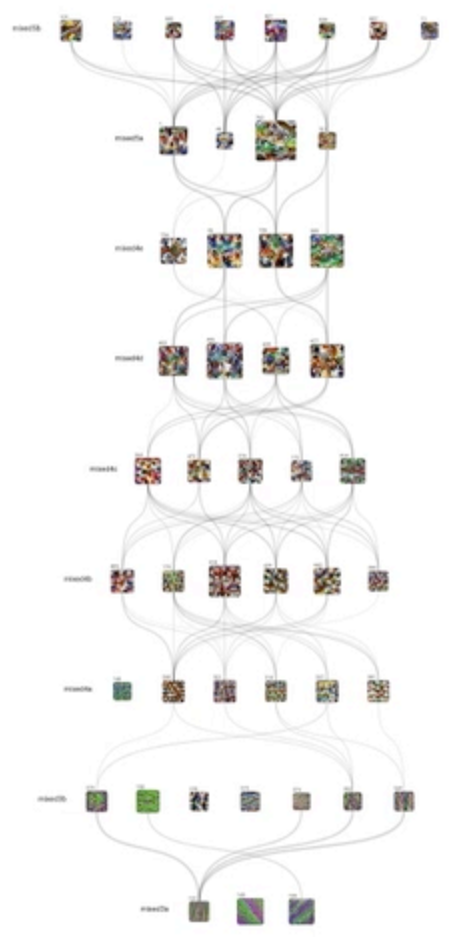
- timber wolf
- malamute ● white wolf
- pembroke
- samoyed
- shetland sheepdog ◦ arctic fox
- lesser panda
- papillon ◦ keeshond
- collie
- chow

🔍 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%



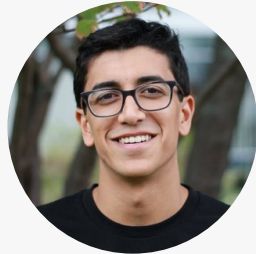


# GAN Lab

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



**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

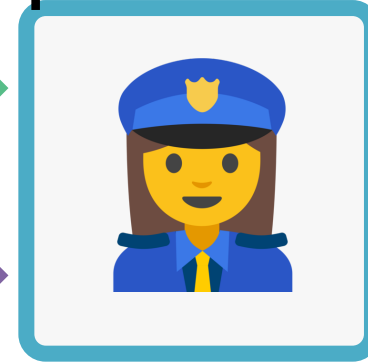
**Generator**  
synthesizes outputs



**Counterfeiter**  
makes fake bills



**Discriminator**  
spots fake

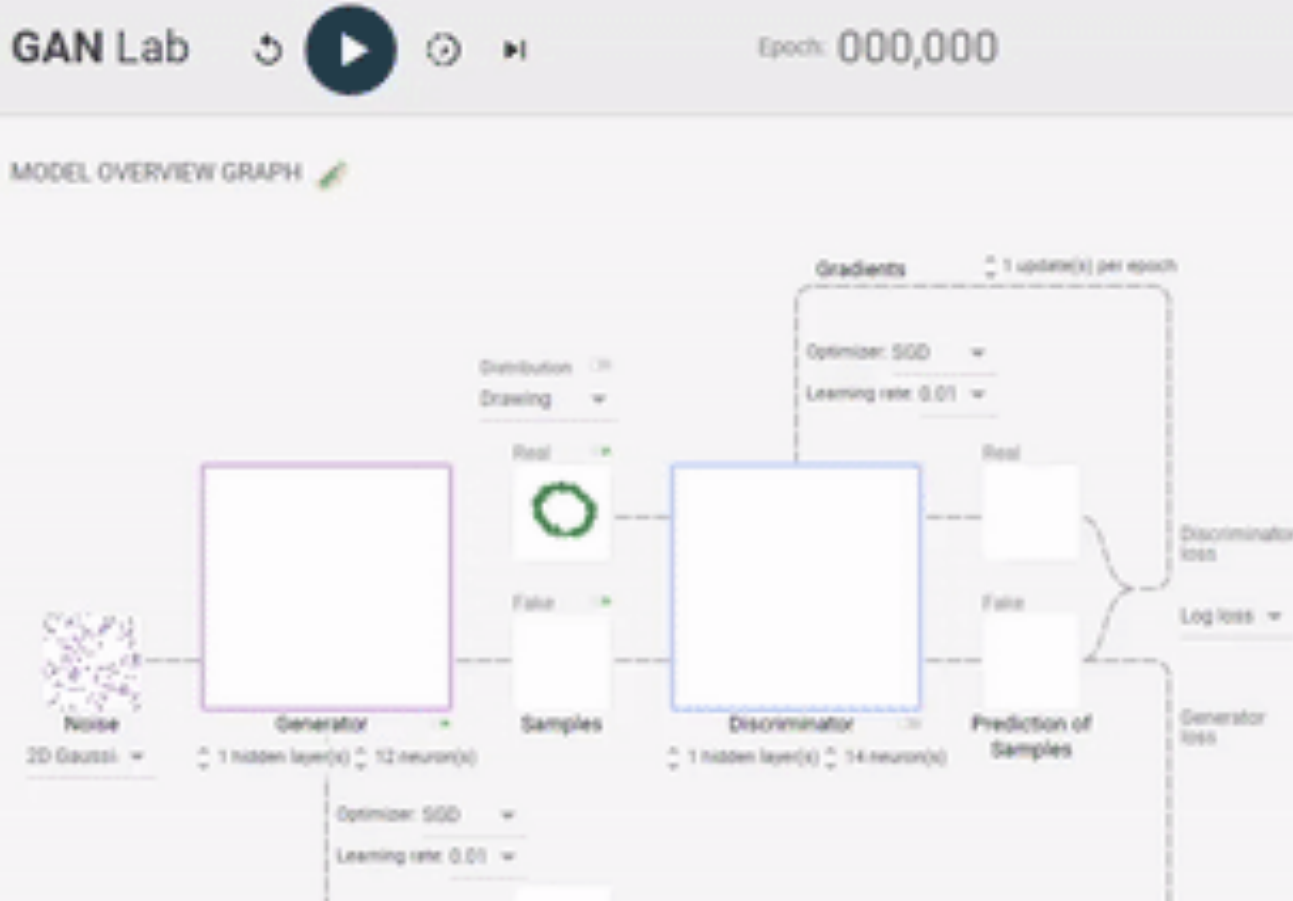


**Police**  
spots fake bills

# GAN Lab is Live! Try at [bit.ly/gan-lab](https://bit.ly/gan-lab)

30K visitors, 135 countries

♥ 1.9K Likes ↻ 800+ Retweets



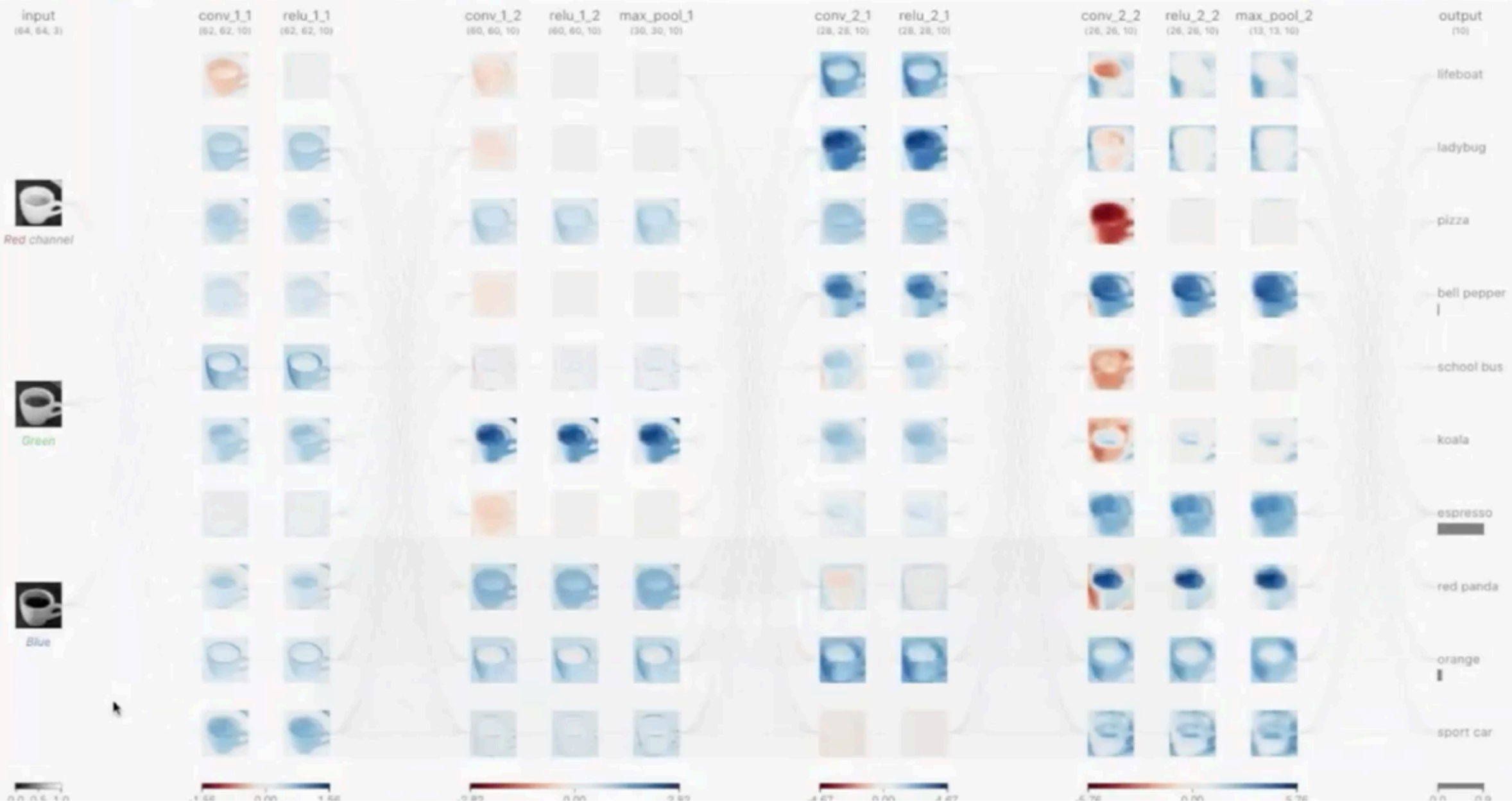
LAYERED DISTRIBUTIONS



METRICS



[Show detail](#) Unit ▼





Thanks!

# SECURE & INTERPRETABLE AI






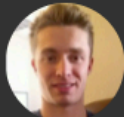






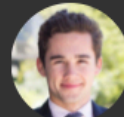

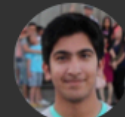




**Polo Chau**

Associate Professor

Associate Director, MS Analytics

Associate Director of Corporate Relations, ML Center

Georgia Tech

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