Recent Advances in Adversarial Machine Learning

Nicholas Carlini
Google Research

Recent Advances in Adversarial (Examples in) Machine Learning

Nicholas Carlini
Google Research

The Year is **2014**

Someone tells you they have a new algorithm to generate human faces

The Year is **2014**

"more results of how this helps on real tasks or real datasets"

"the theoretical work is primitive, and the experiments are pretty basic."

The Year is **2017**

Someone tells you they have a new algorithm to generate human faces

The Year is **2017**

The Year is **2013**

Someone tells you they have discovered
a flaw in the robustness of neural networks

The Year is **2013**

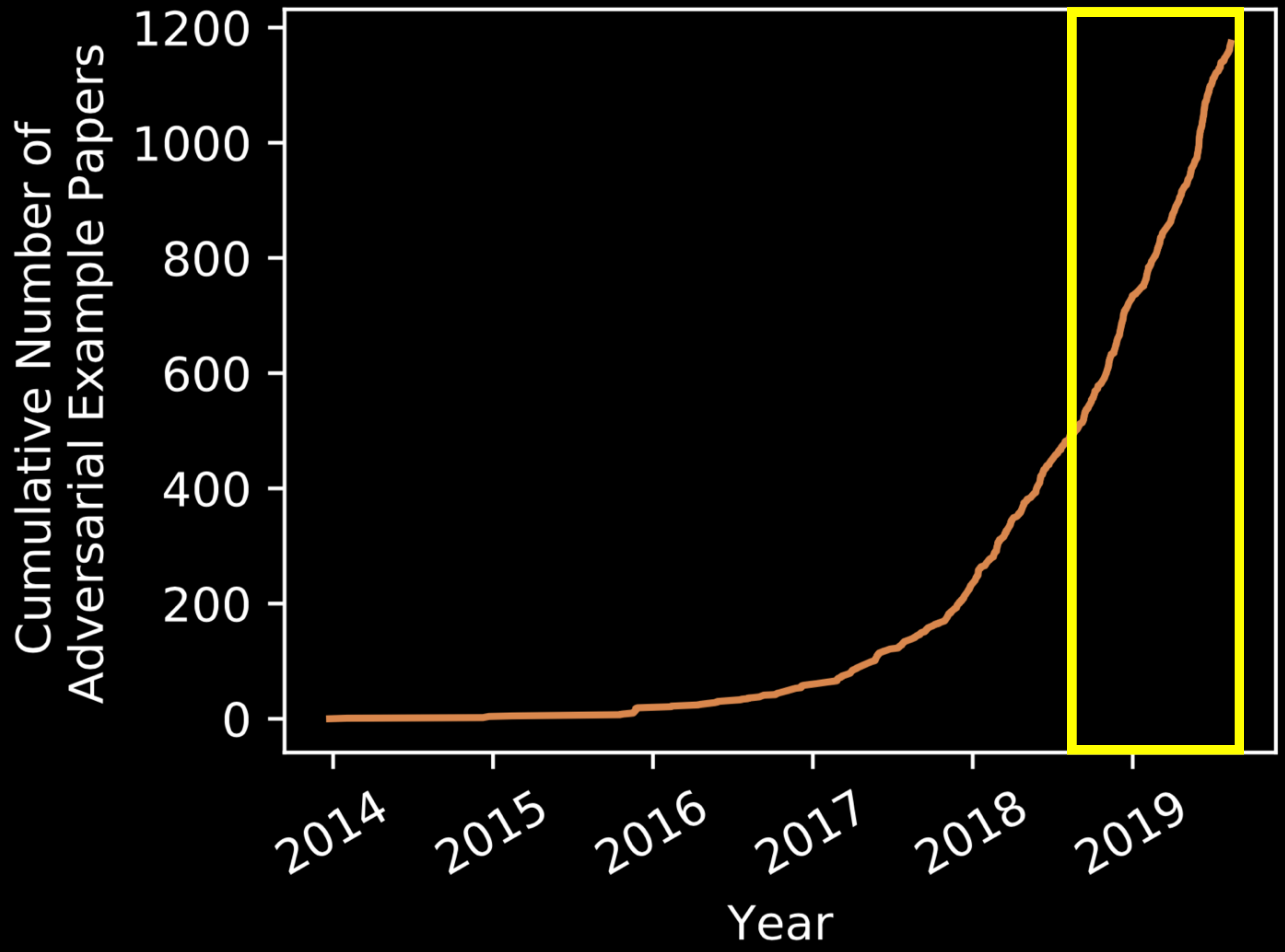
The Year is **2019**

Someone tells you they have discovered
a flaw in the robustness of neural networks

The Year is **2019**

3 years:

6 years:


Cumulative Number of
Adversarial Example Papers
1200
1000
800
600
400
200
0
2014
2015
2016
2017
2018
2019
Year

Background:
Adversarial Examples

88% **tabby cat**

88% **tabby cat**

adversarial
perturbation

adversarial
perturbation

88% **tabby cat**

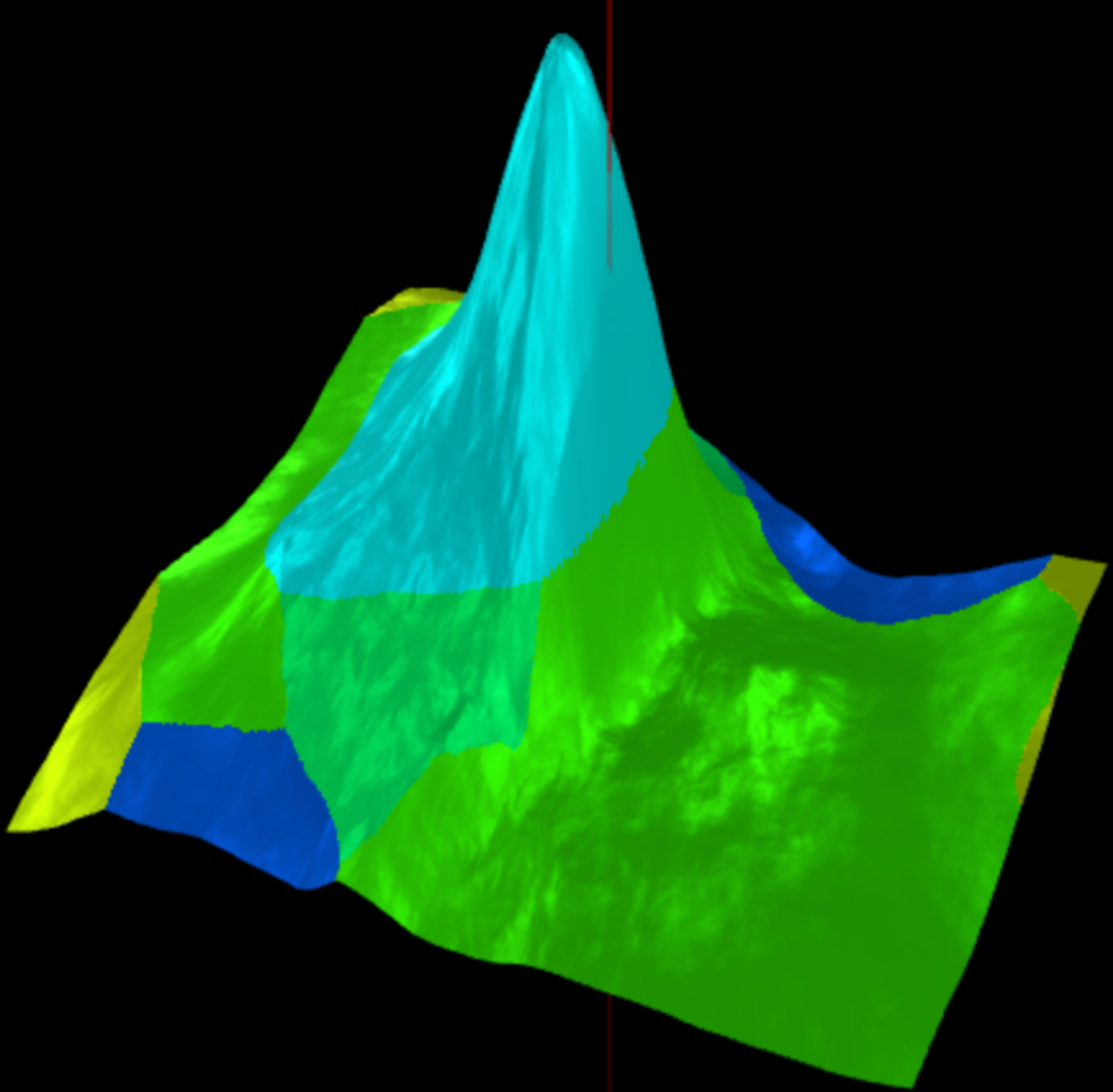
adversarial
perturbation

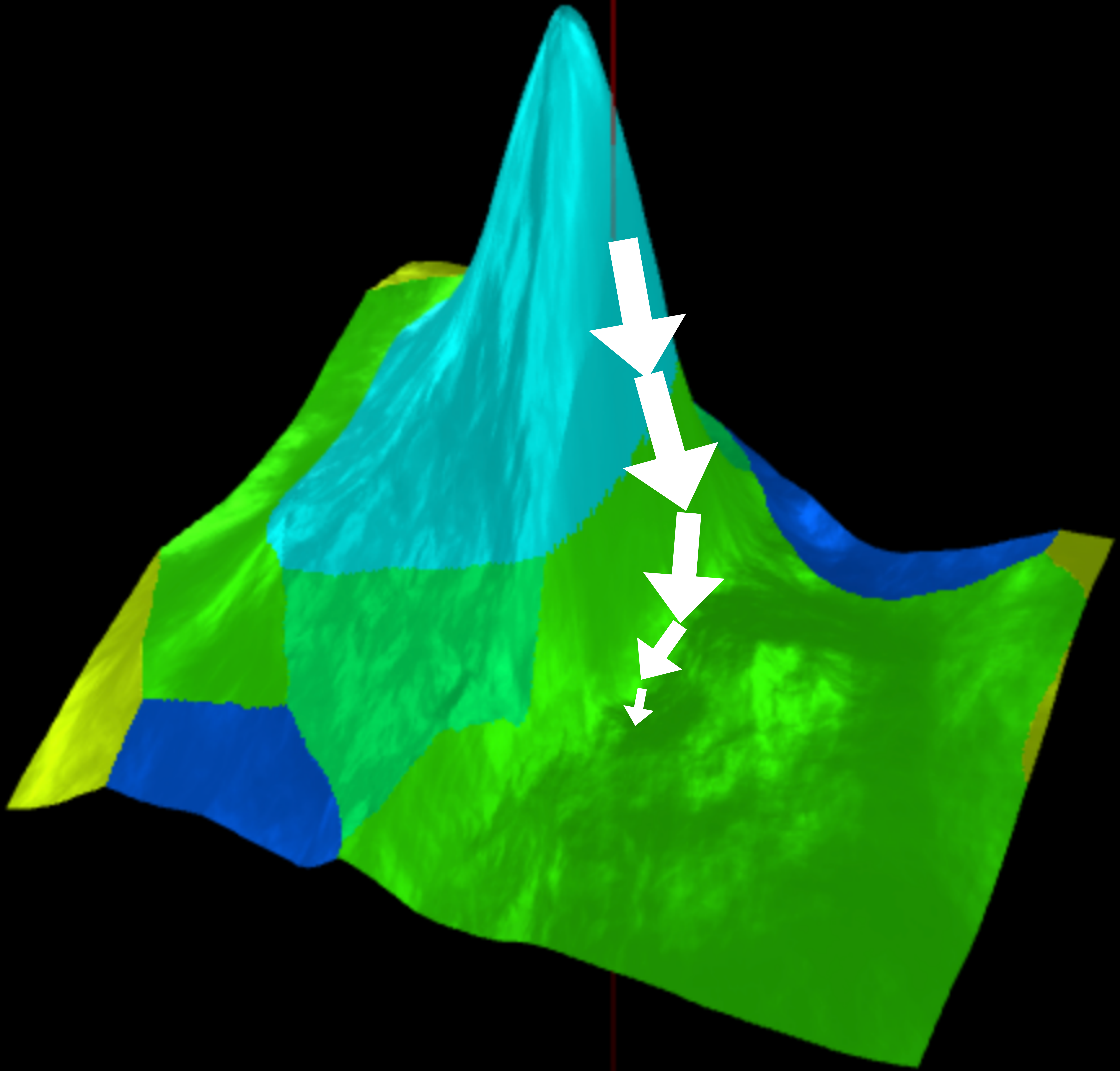
88% tabby cat

99% guacamole


Dog
Truck
Random Direction
Random Direction


Random Direction
Truck
Adversarial Direction
Dog
Airplane





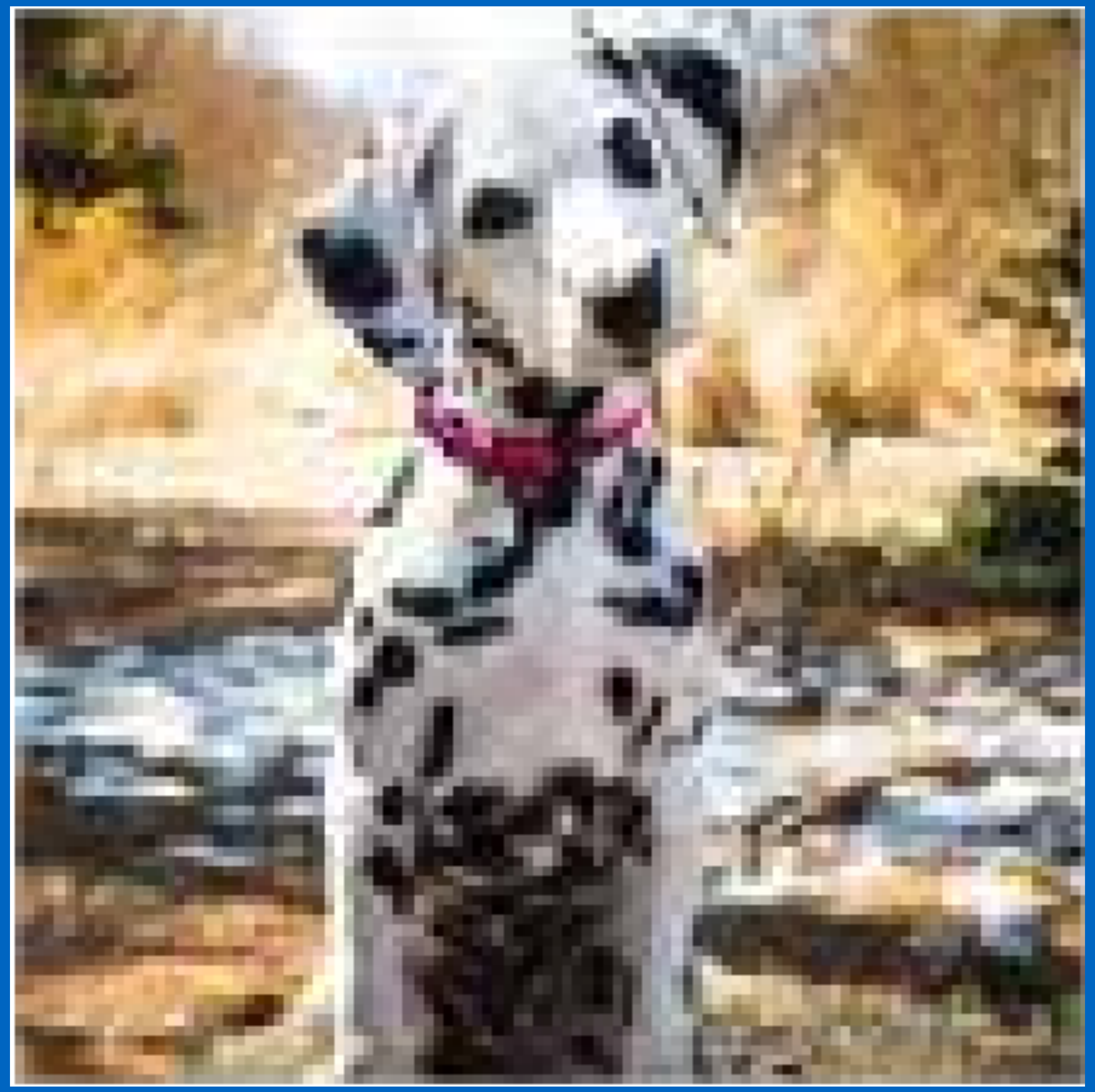
Recent advances in …
Generating Adversarial Examples

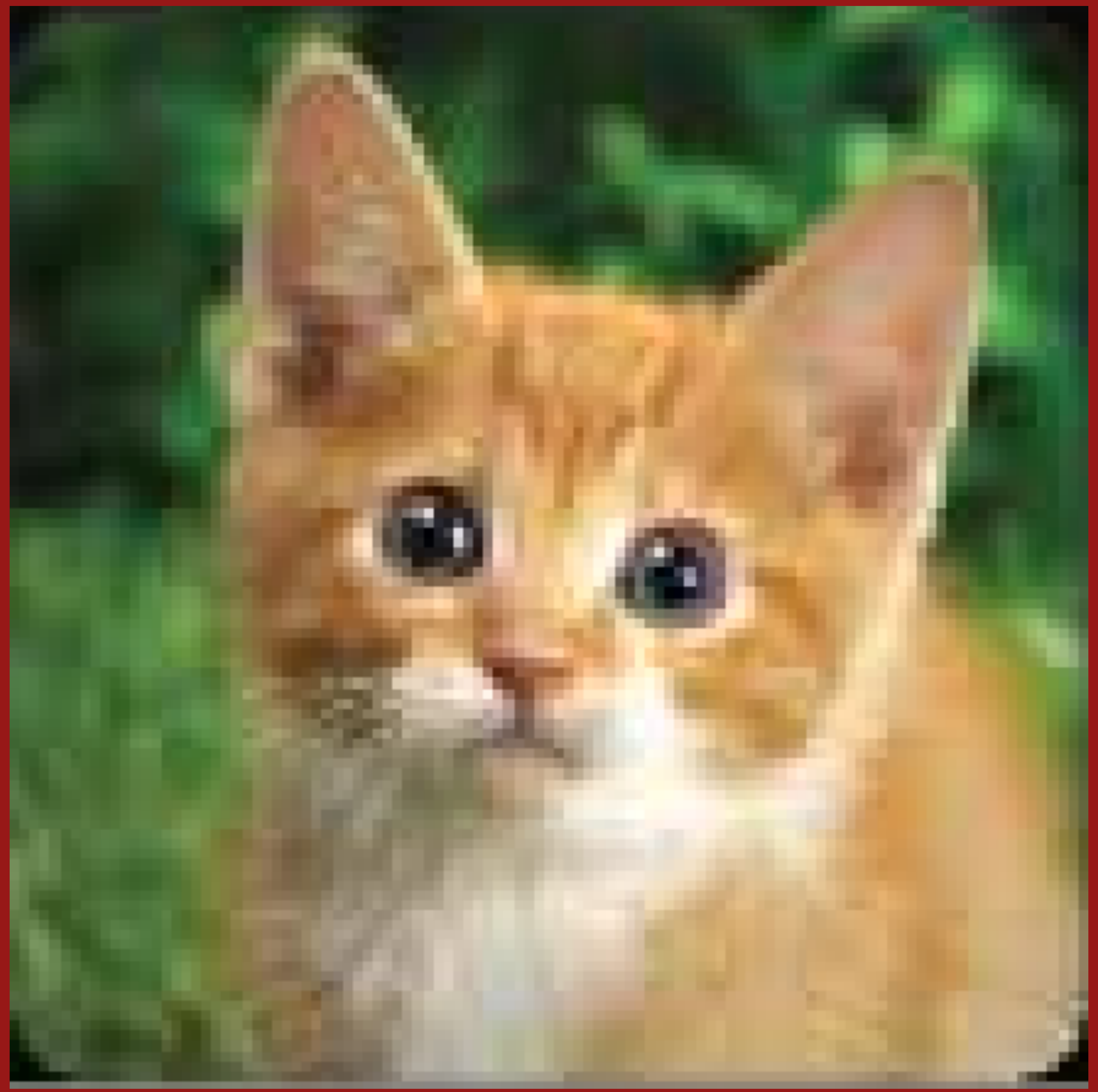
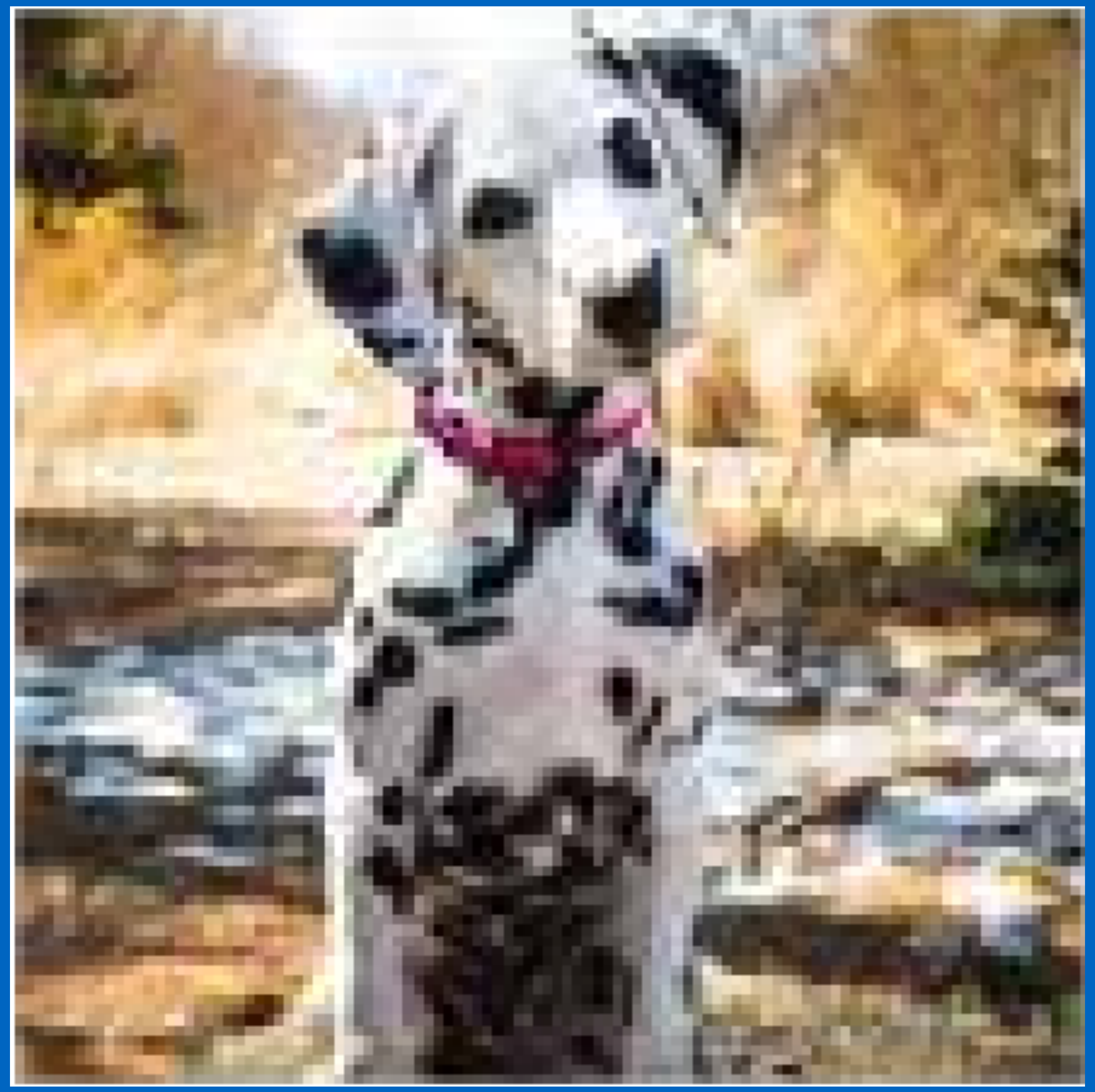
Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models

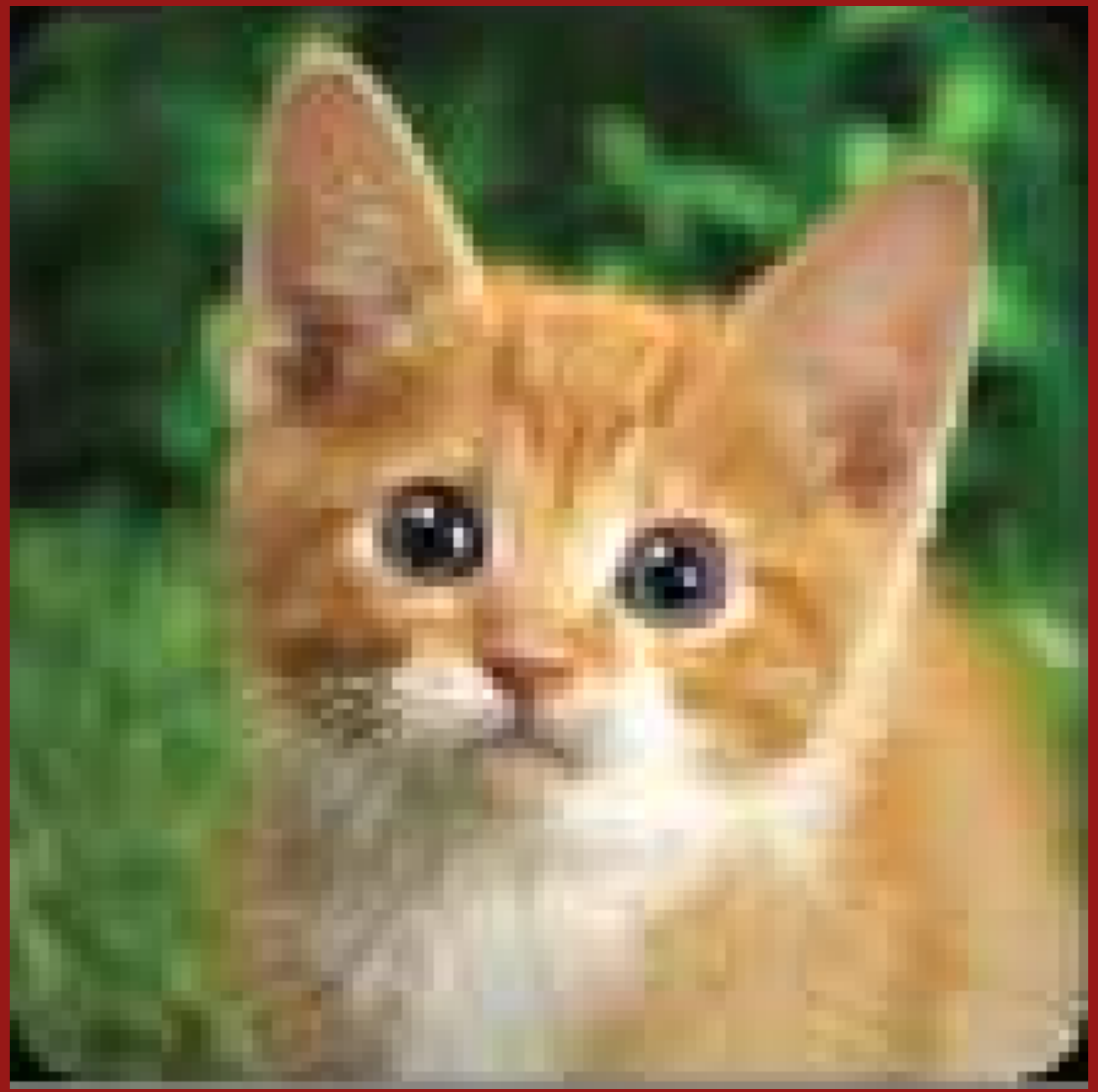
Wieland Brendel*, Jonas Rauber* & Matthias Bethge
Werner Reichardt Centre for Integrative Neuroscience,
Eberhard Karls University Tübingen, Germany
{wieland,jonas,matthias}@bethgelab.org

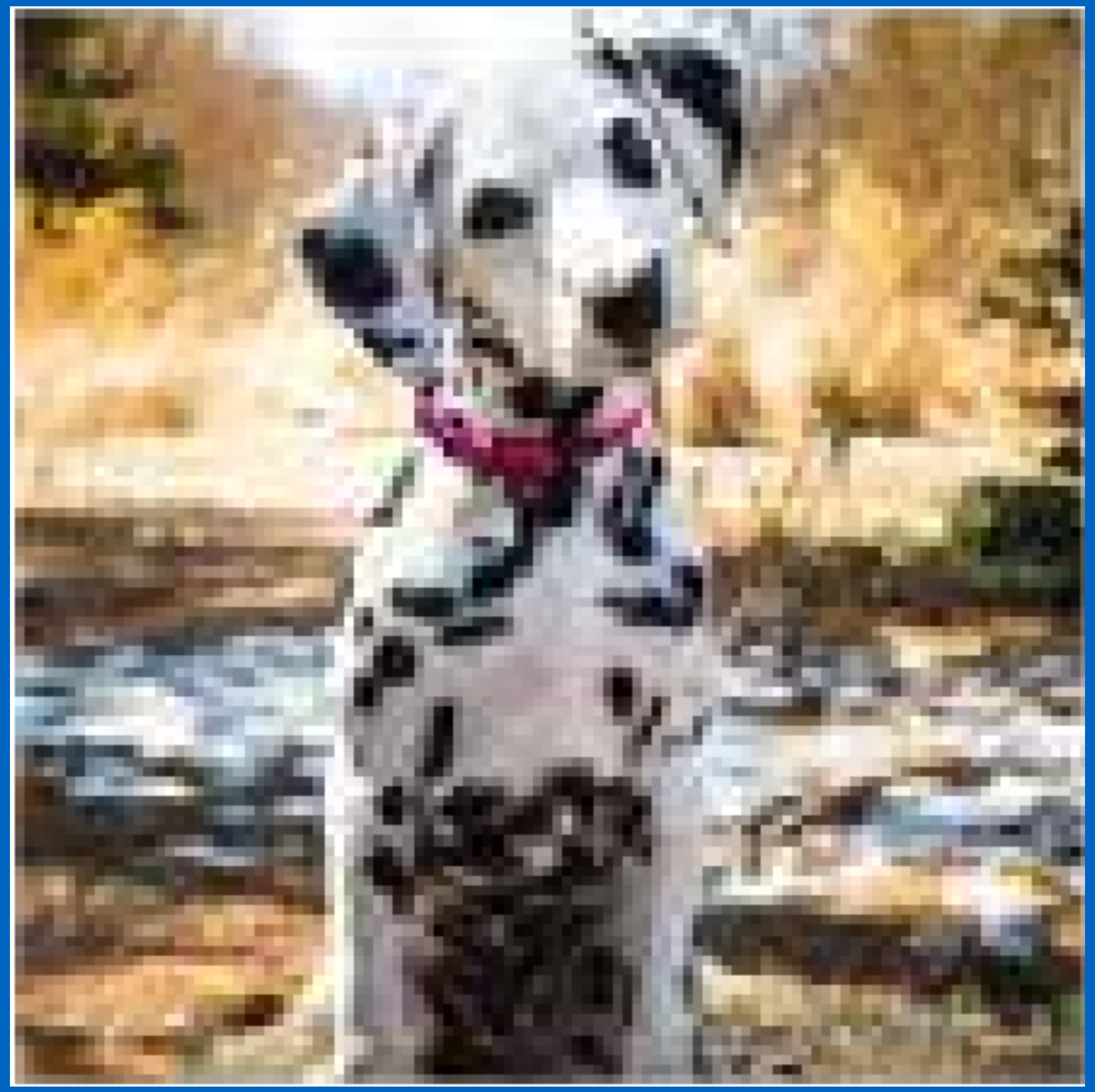
Threat Model:

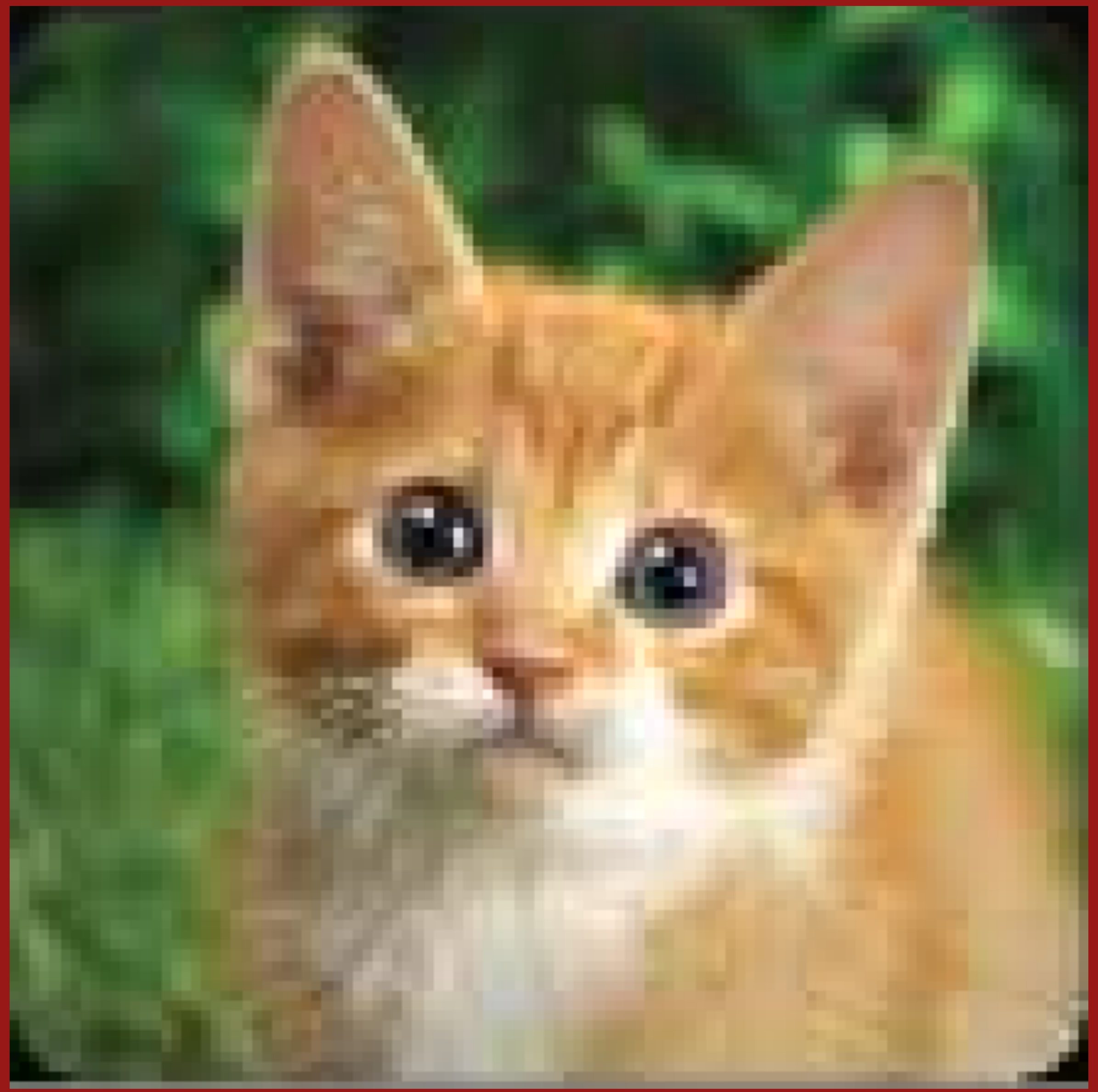
- Black Box
- Hard Label
- Query Access

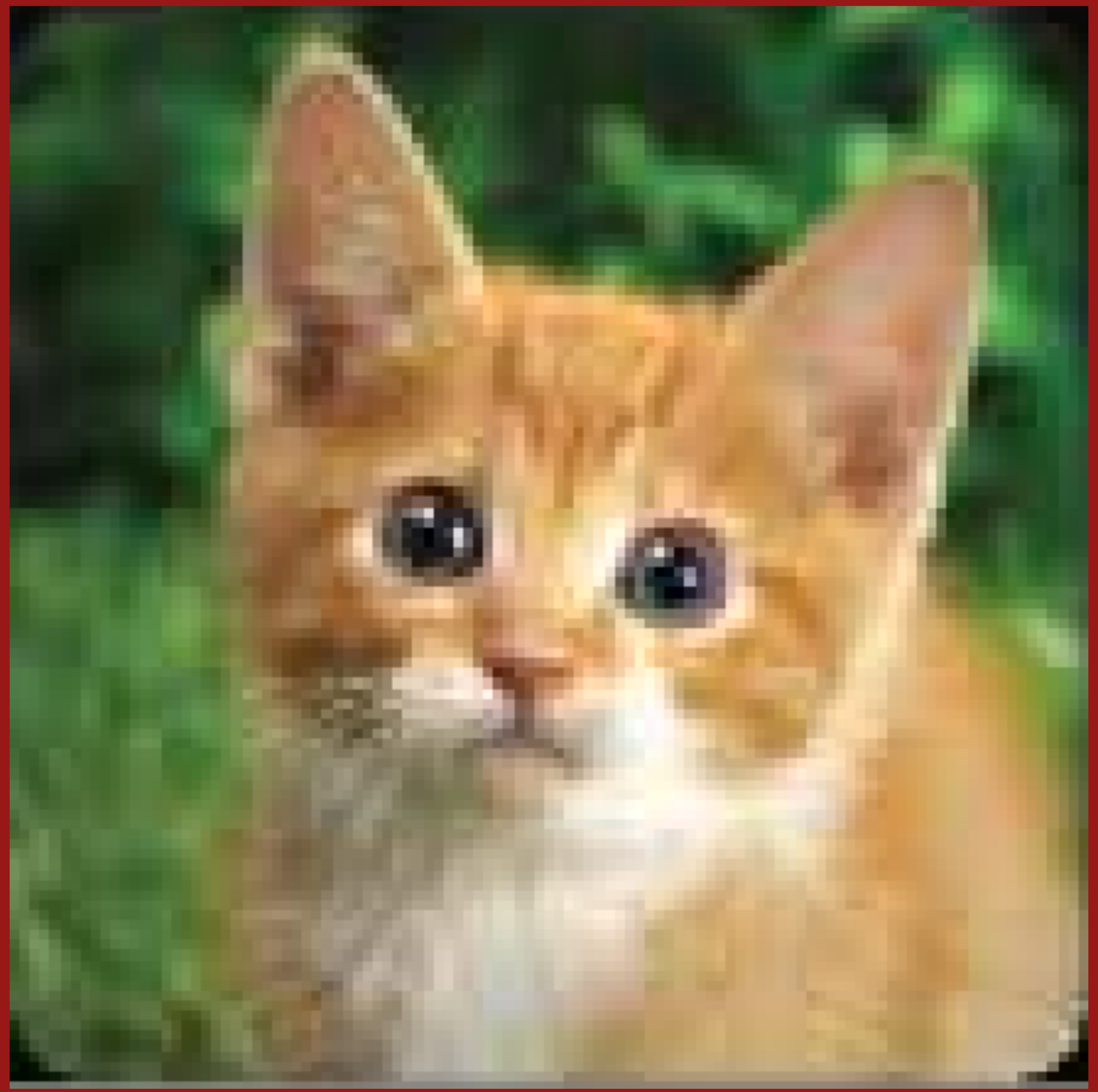



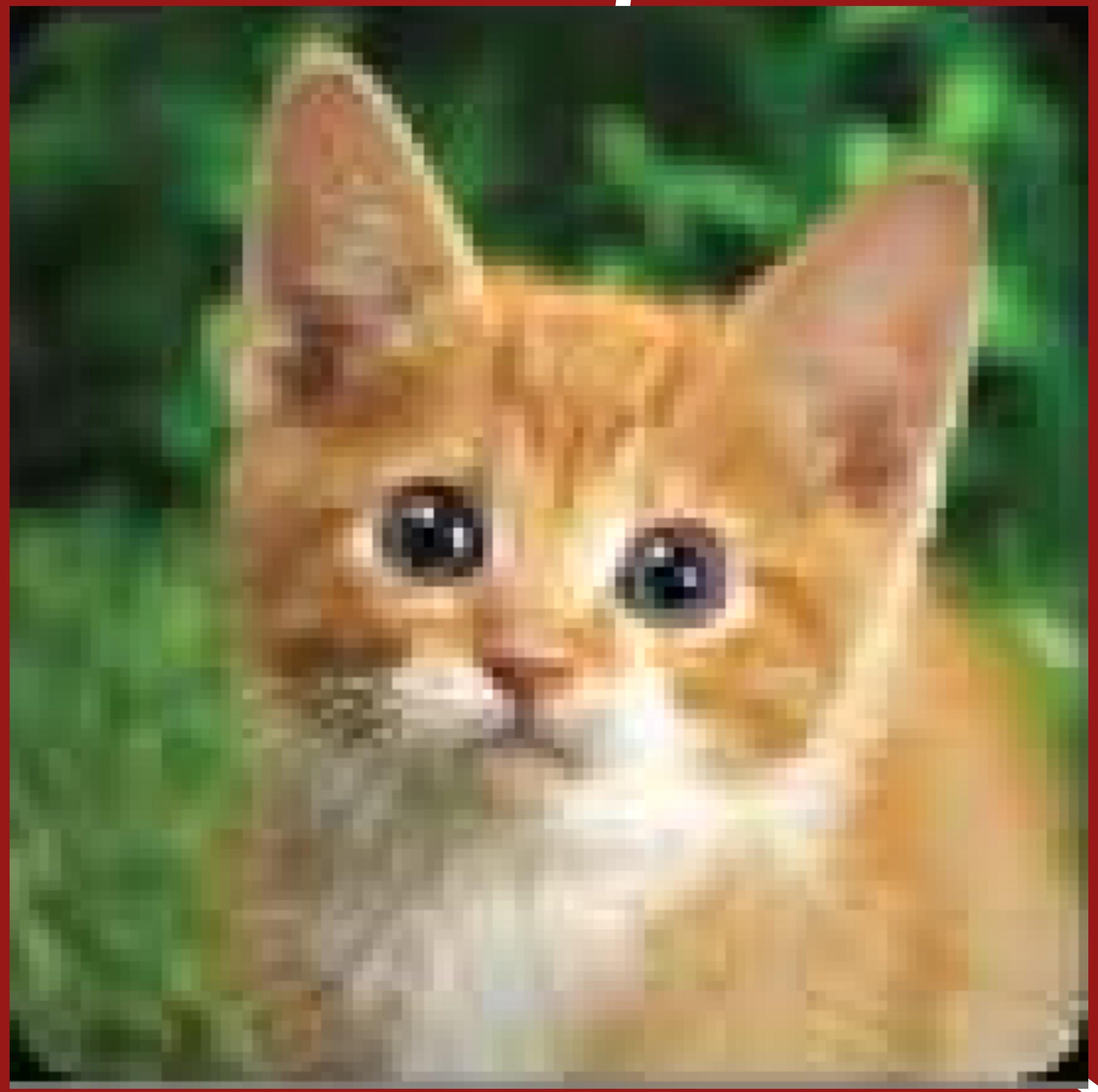



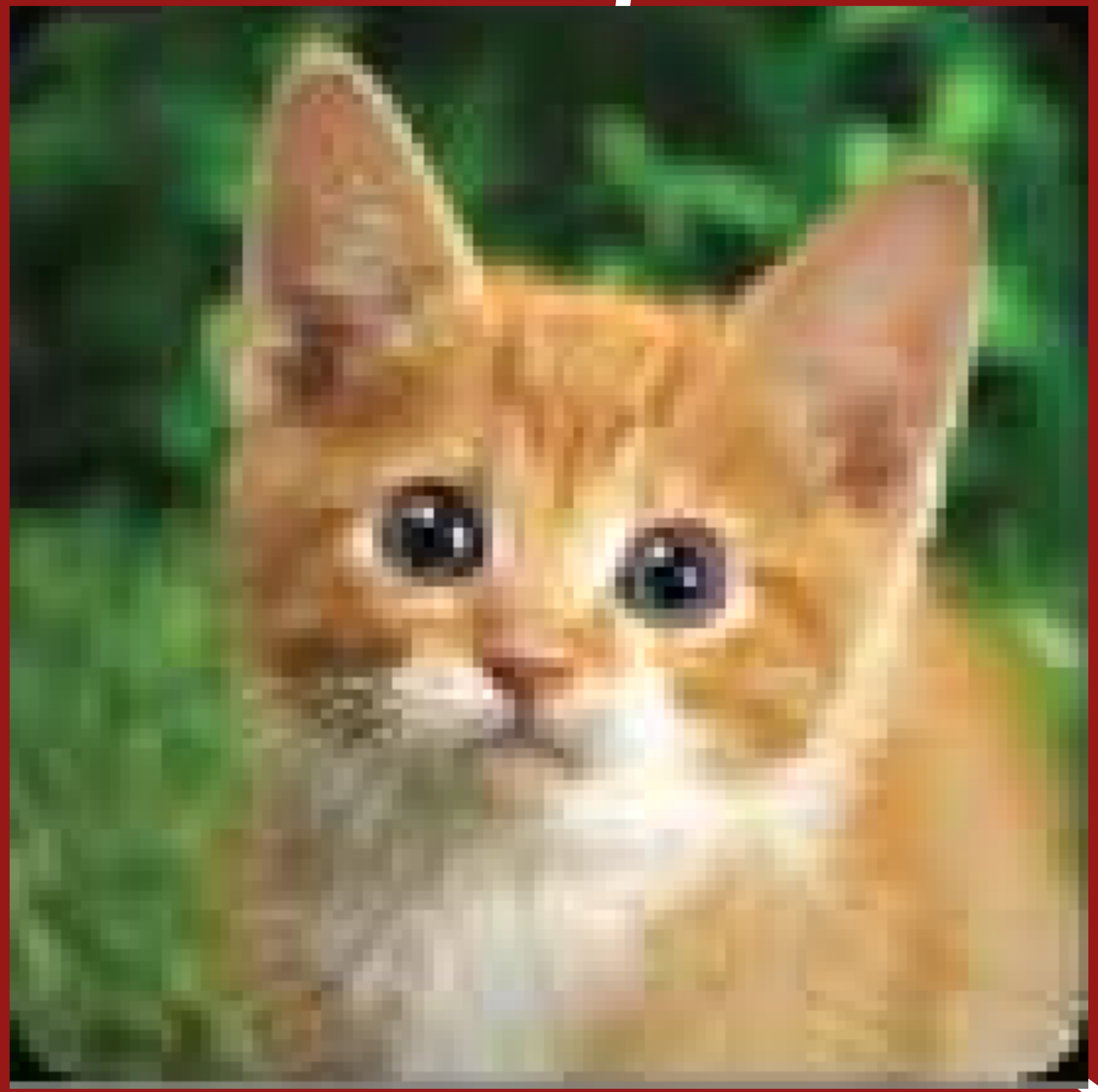


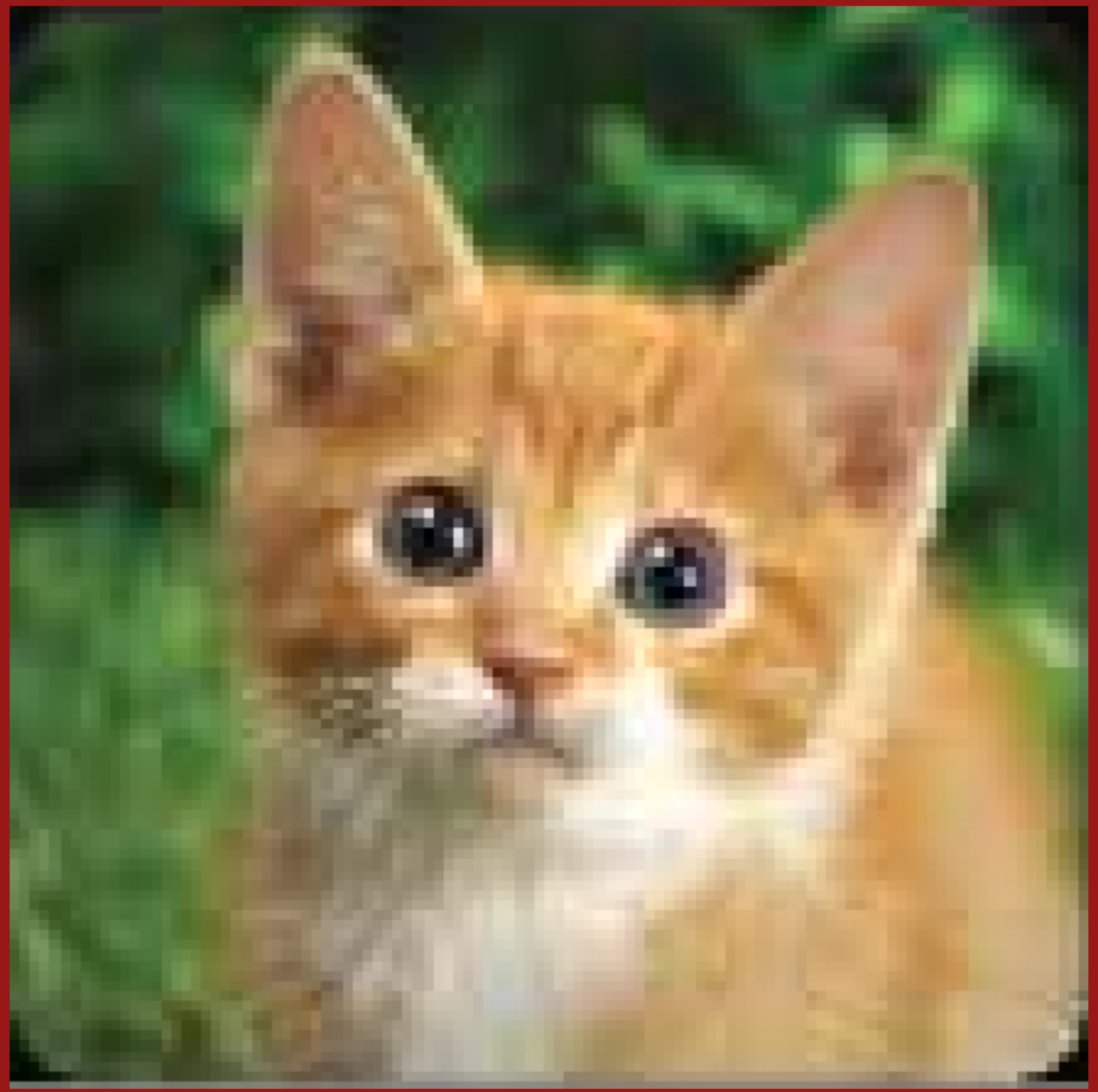
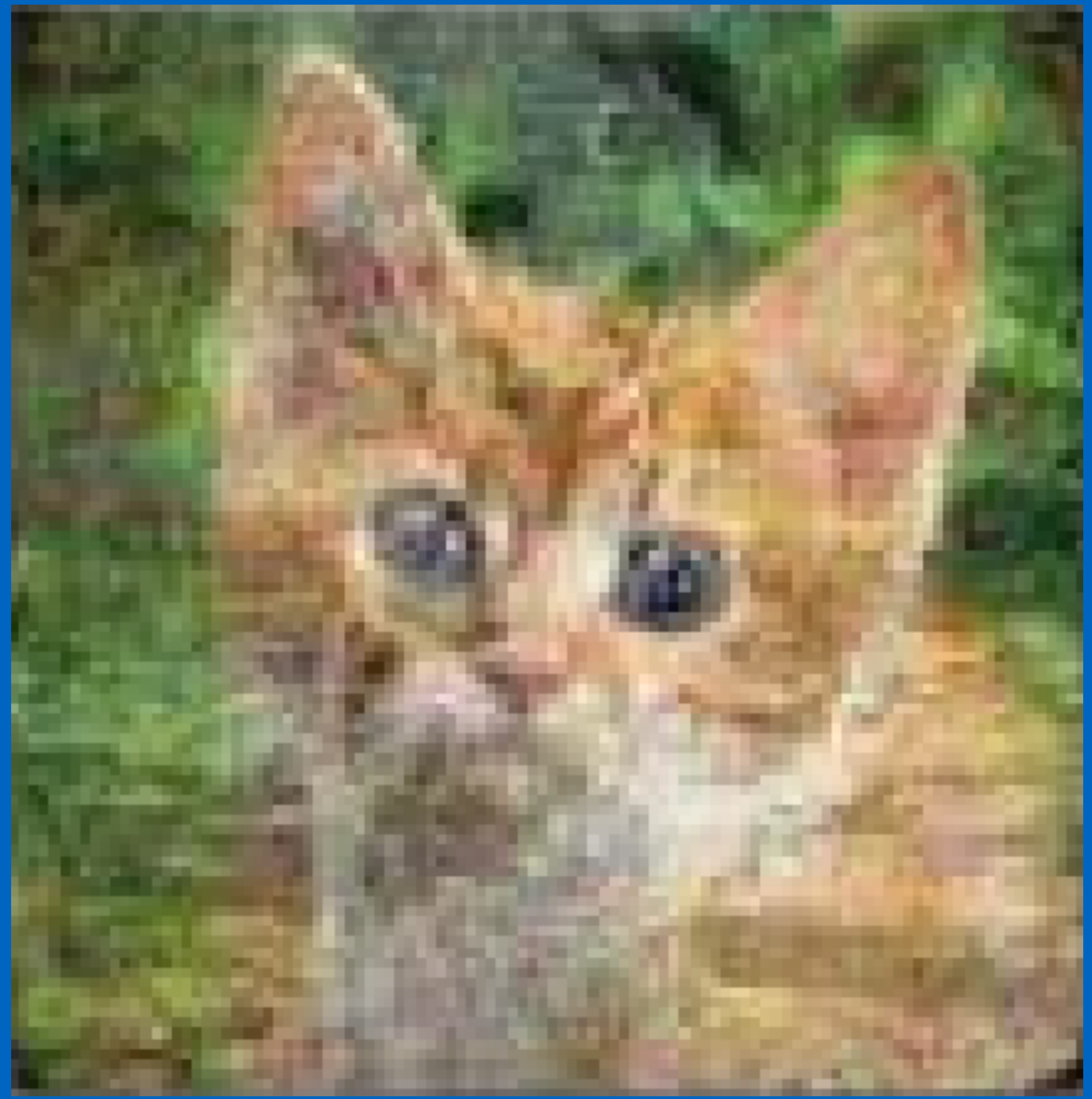


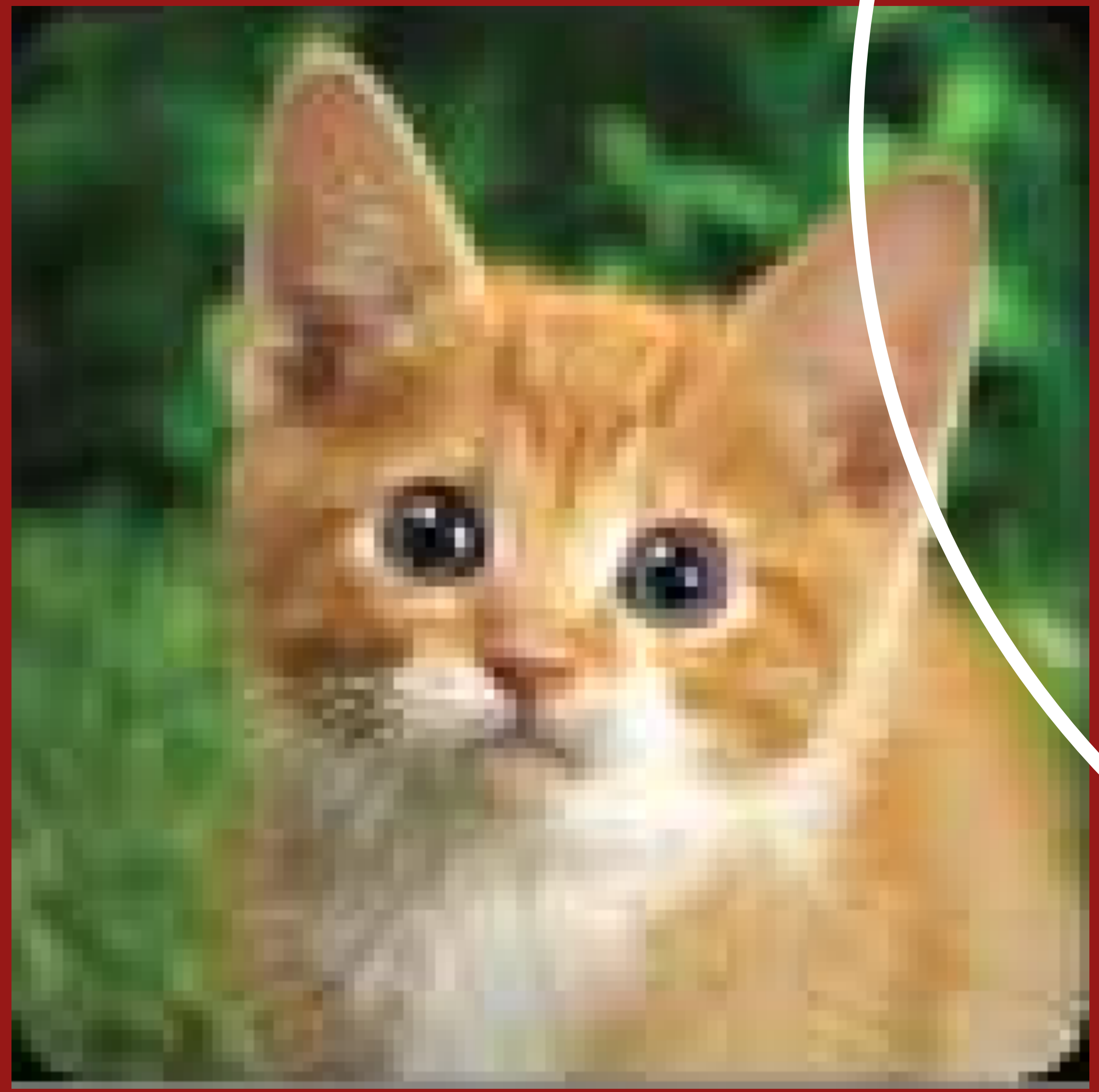


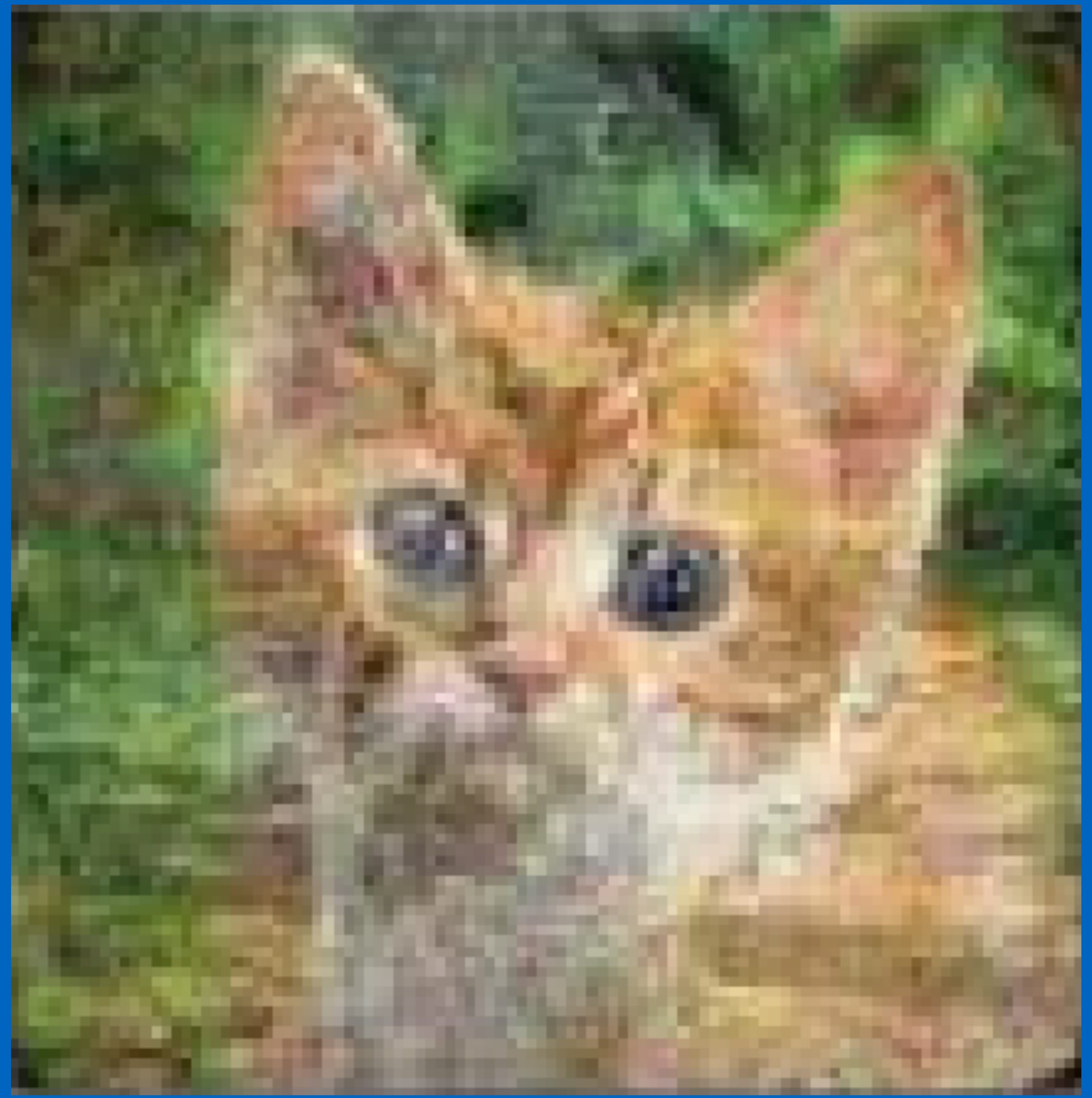


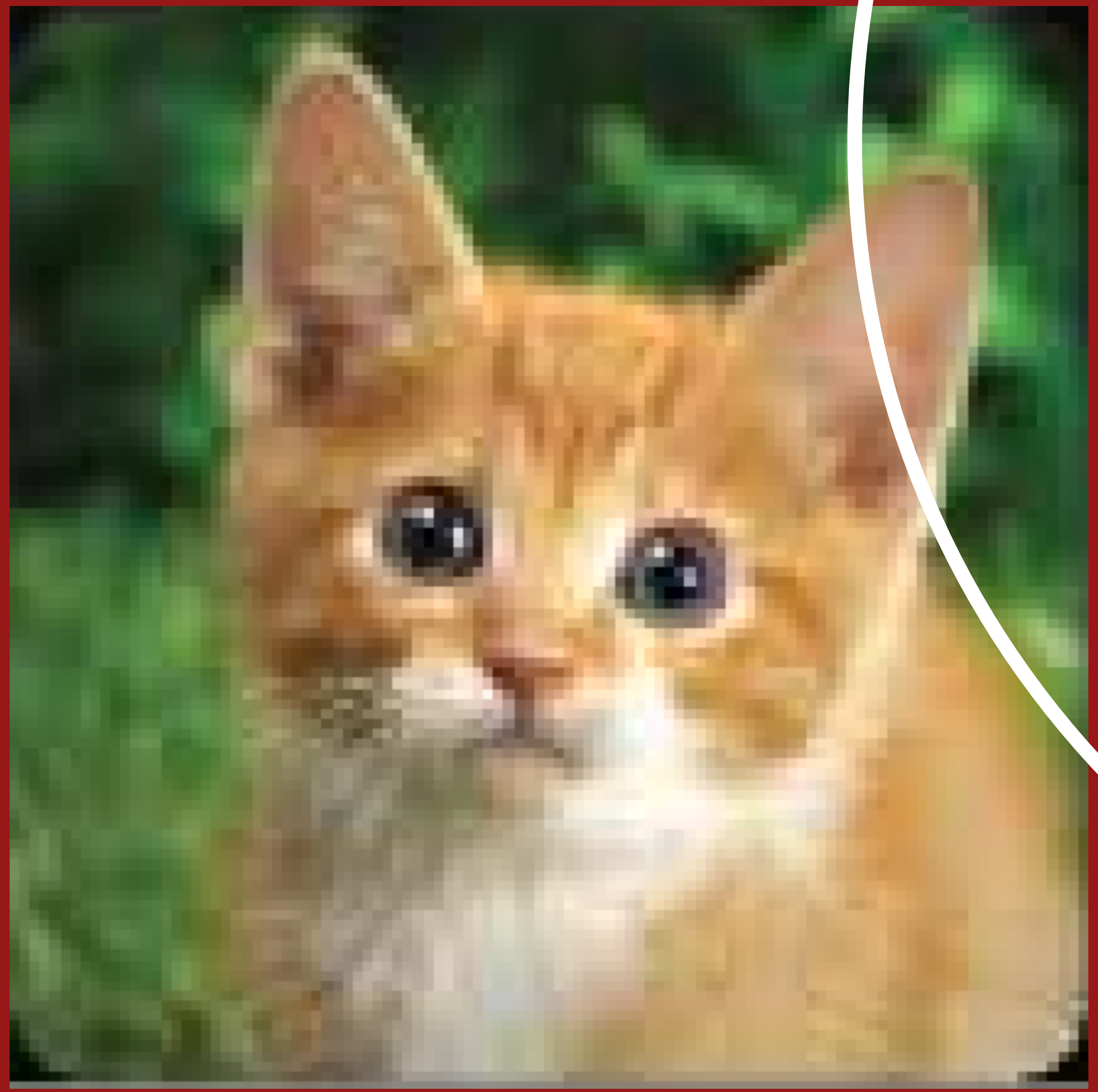


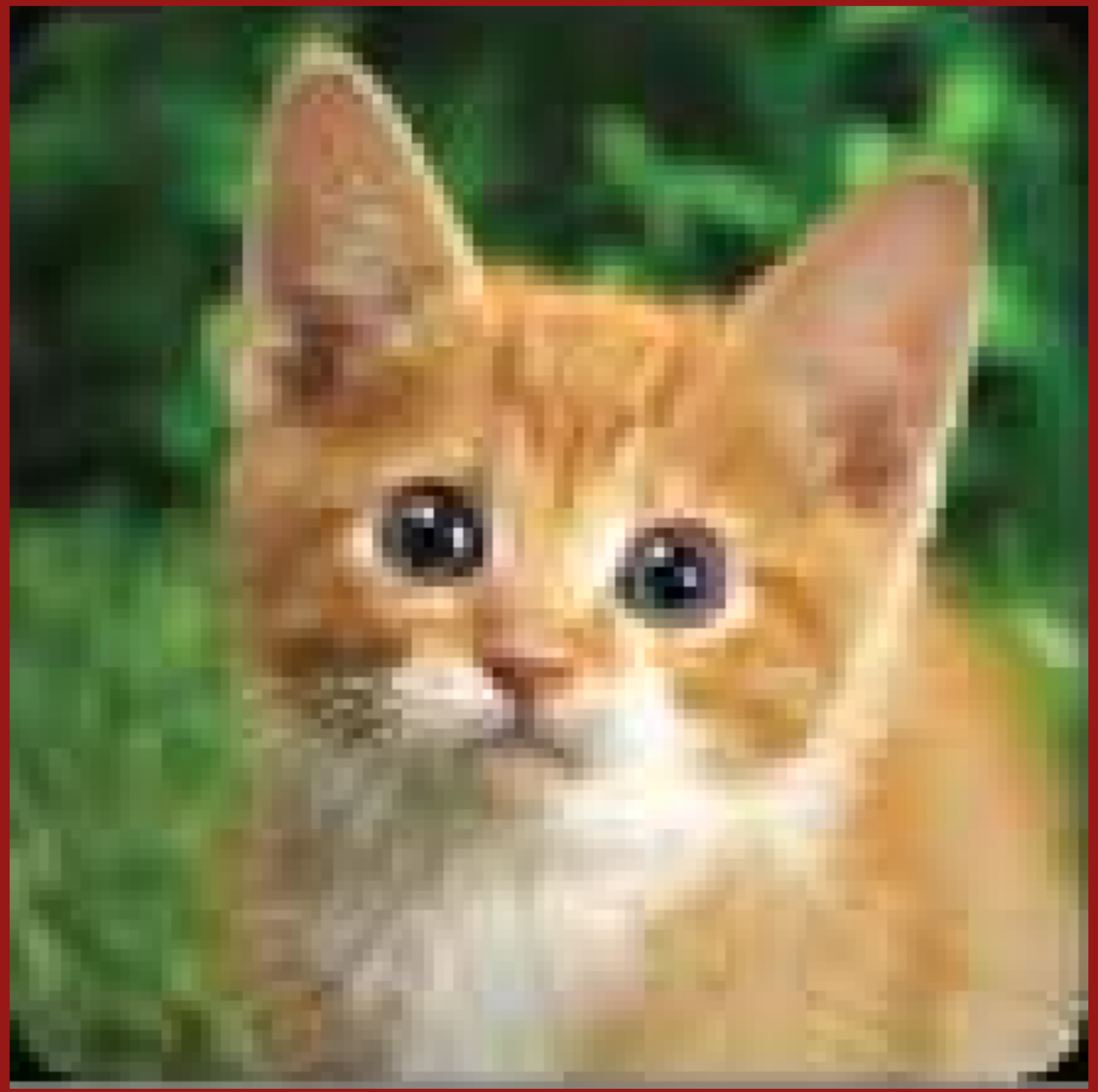


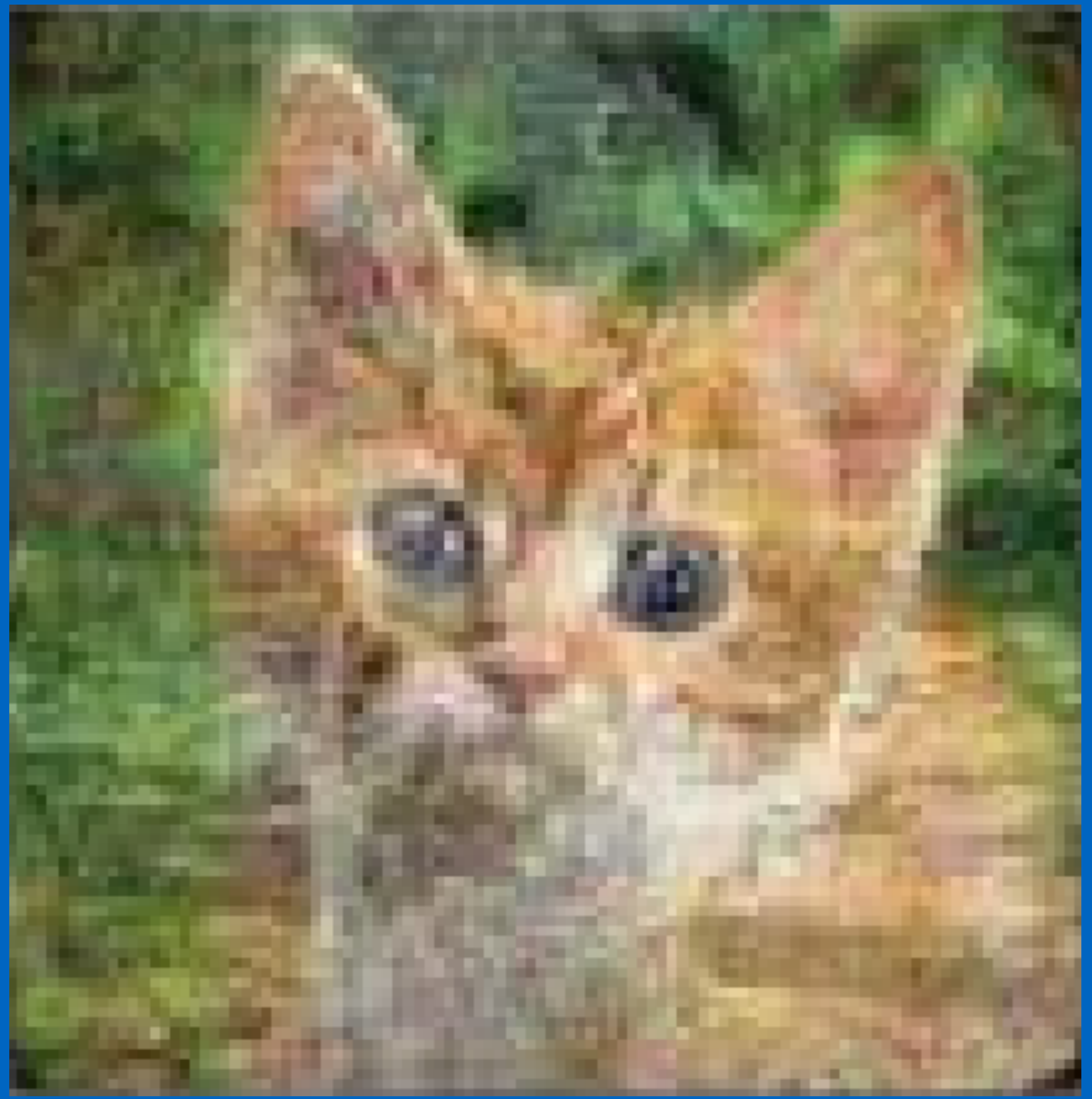



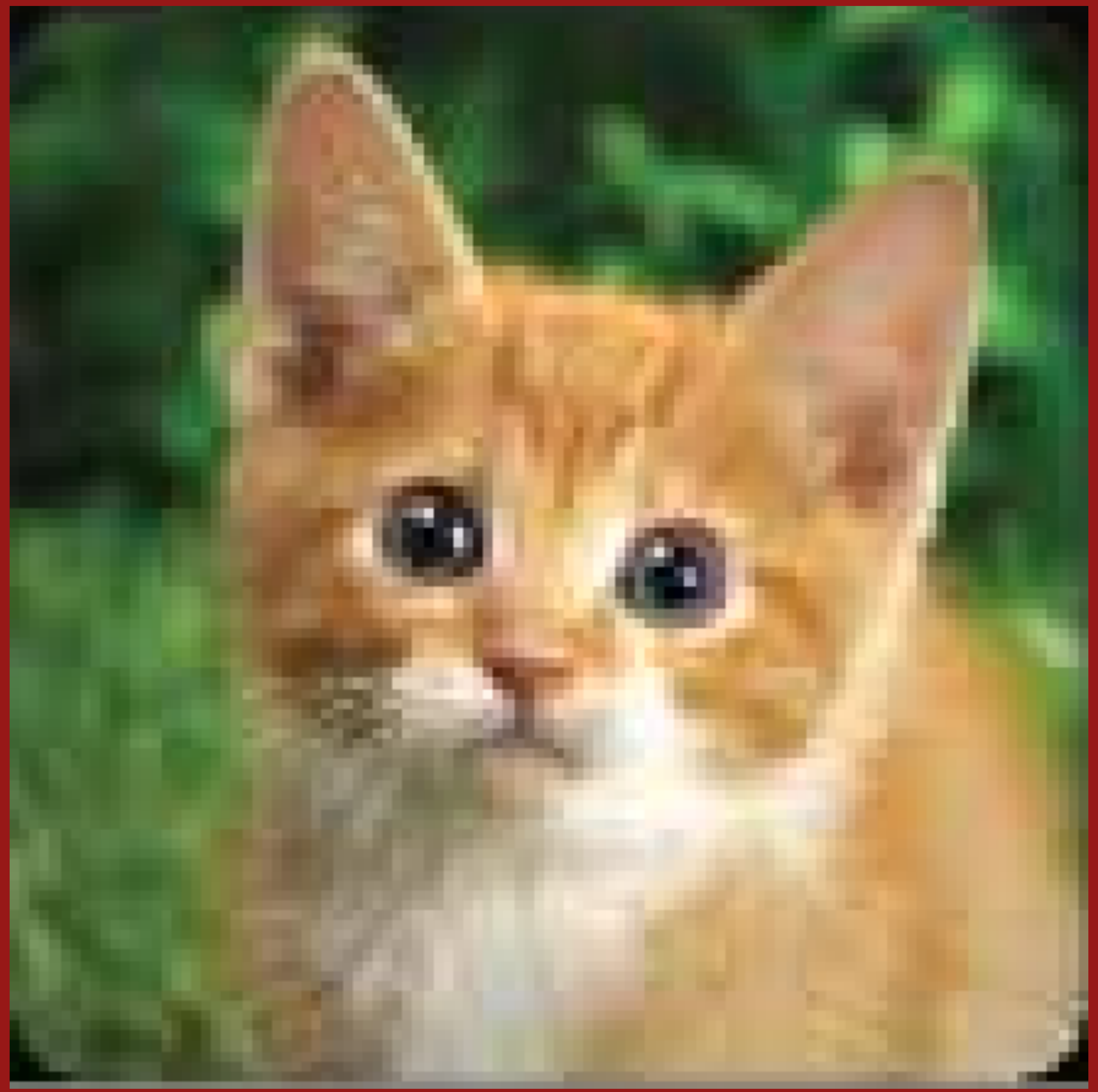


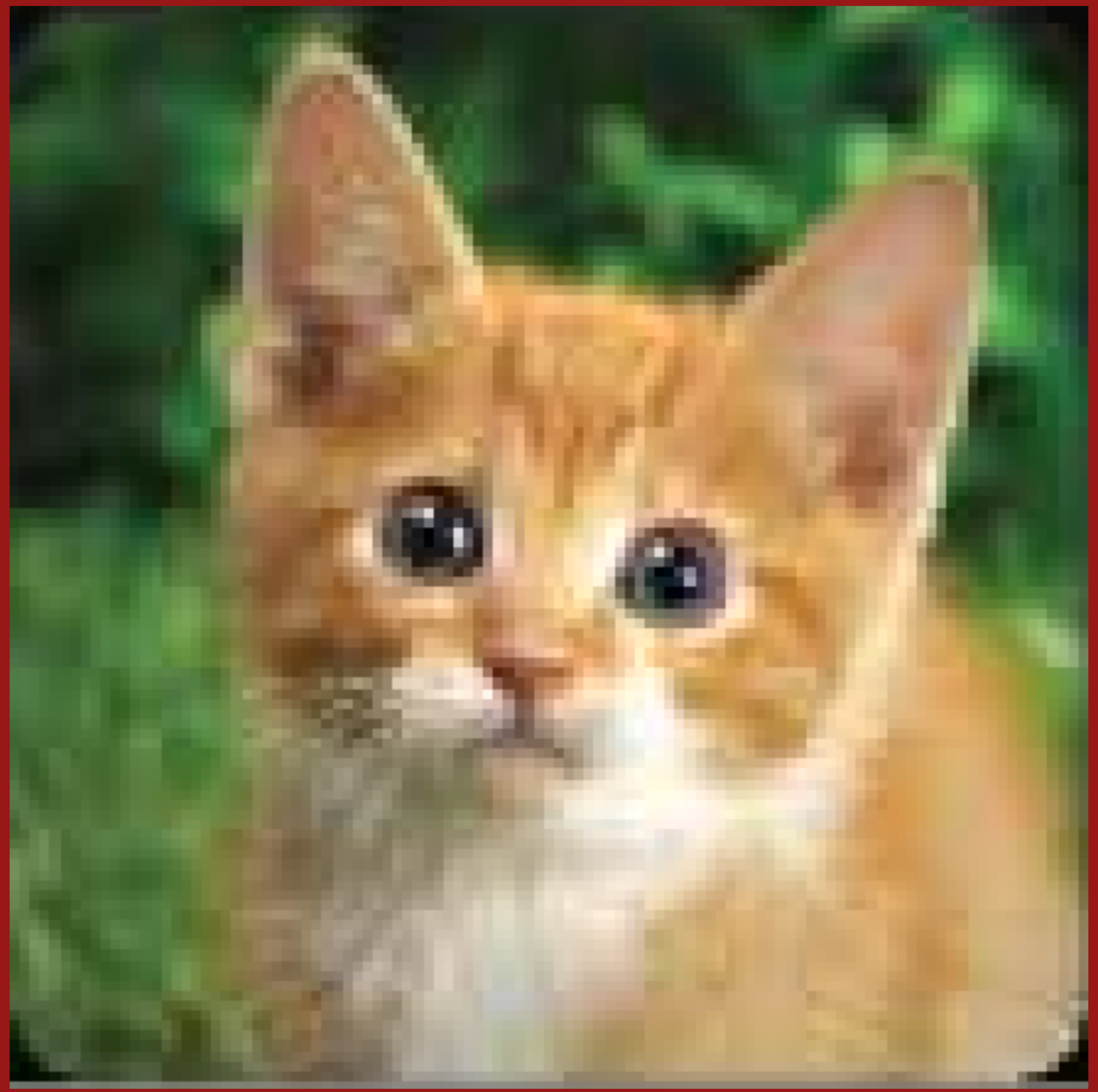


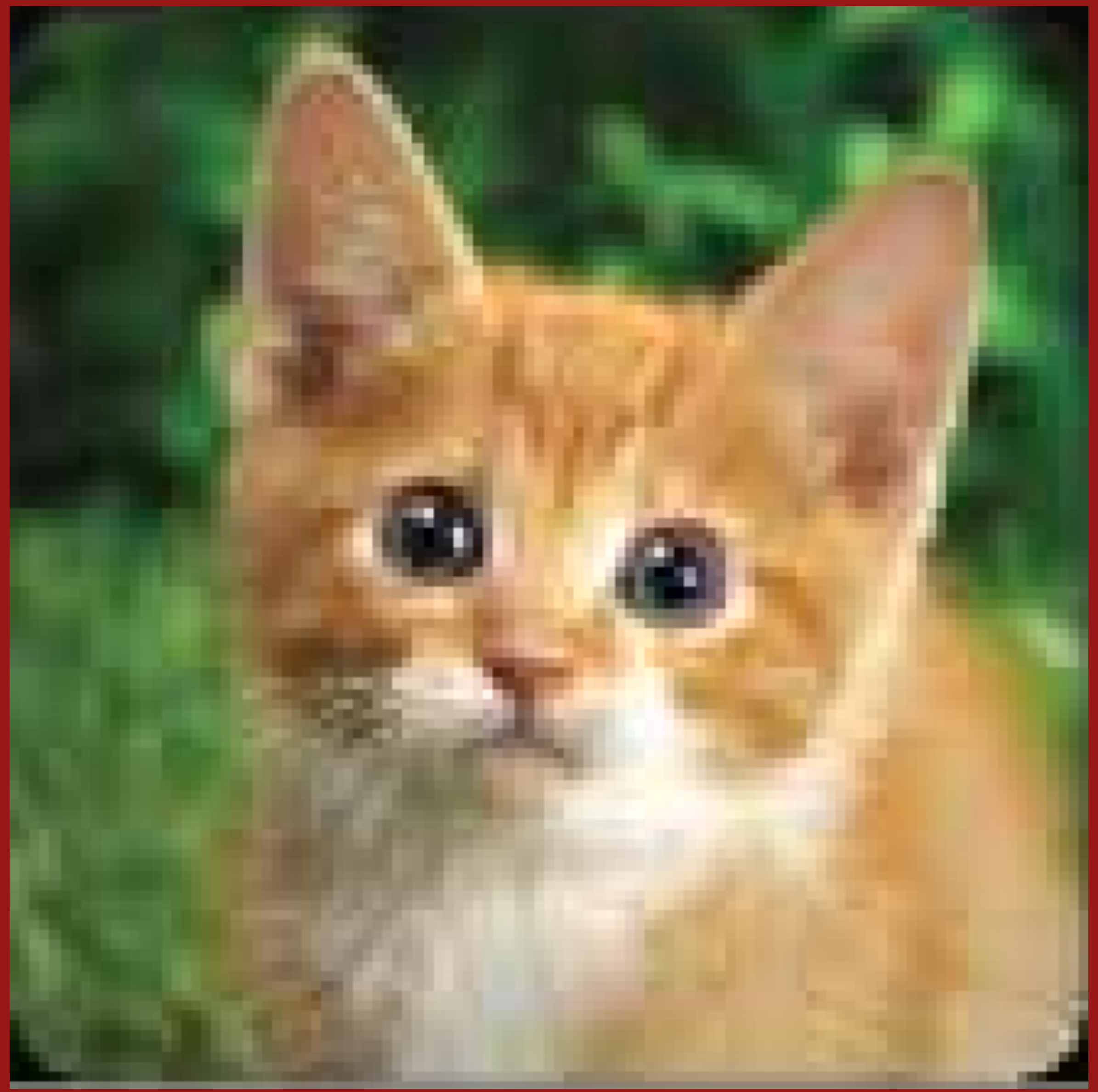


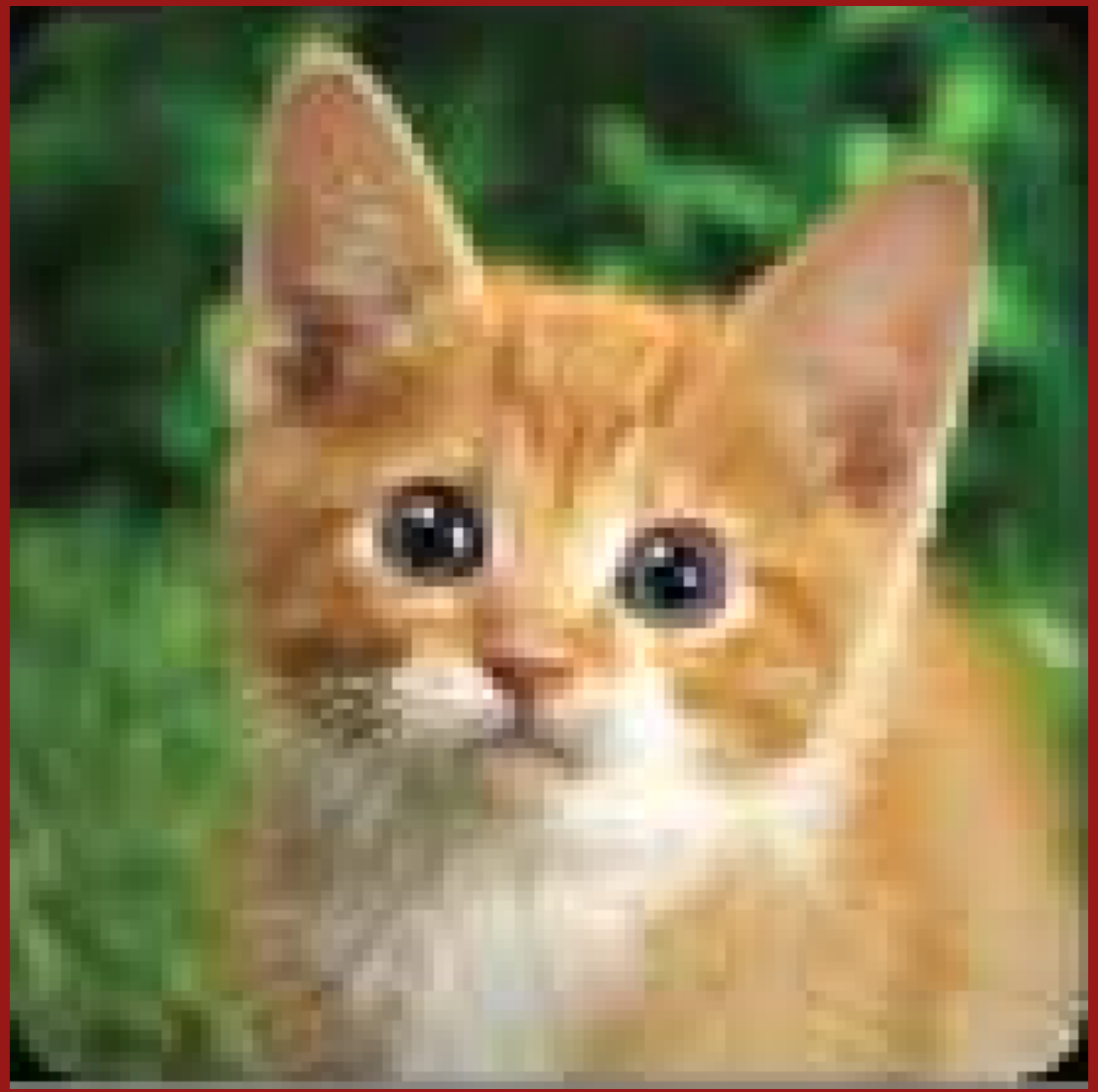


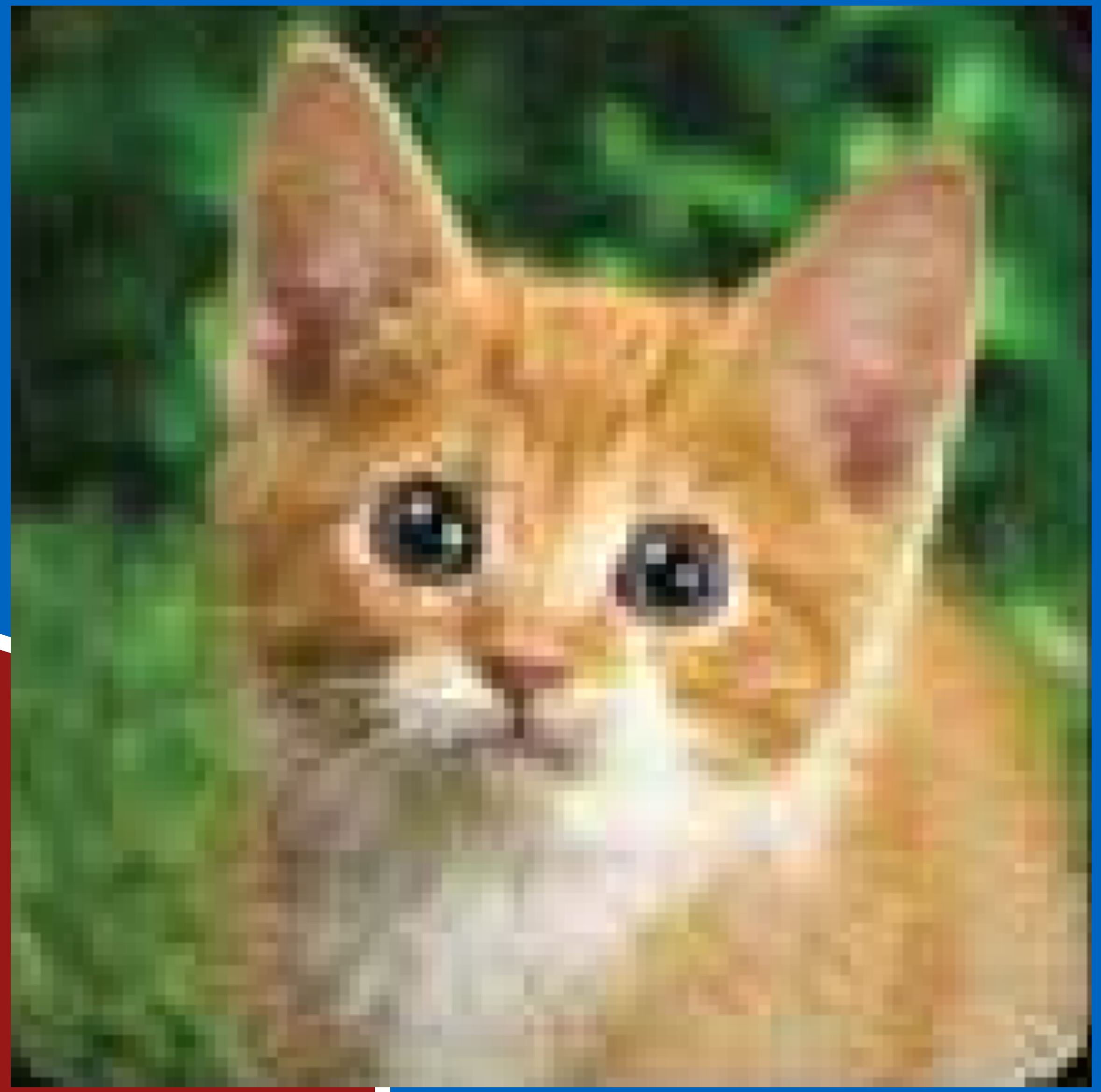










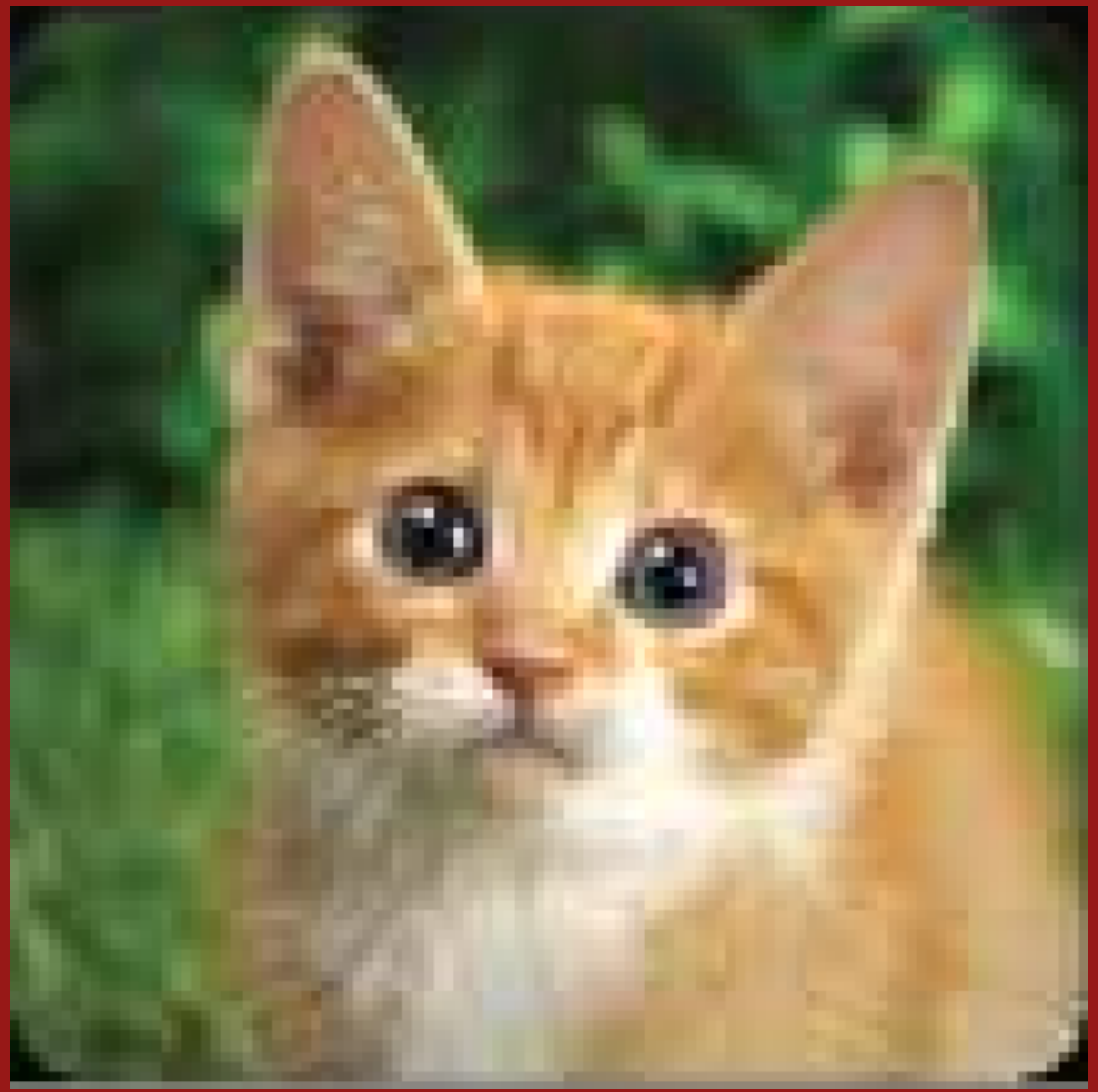


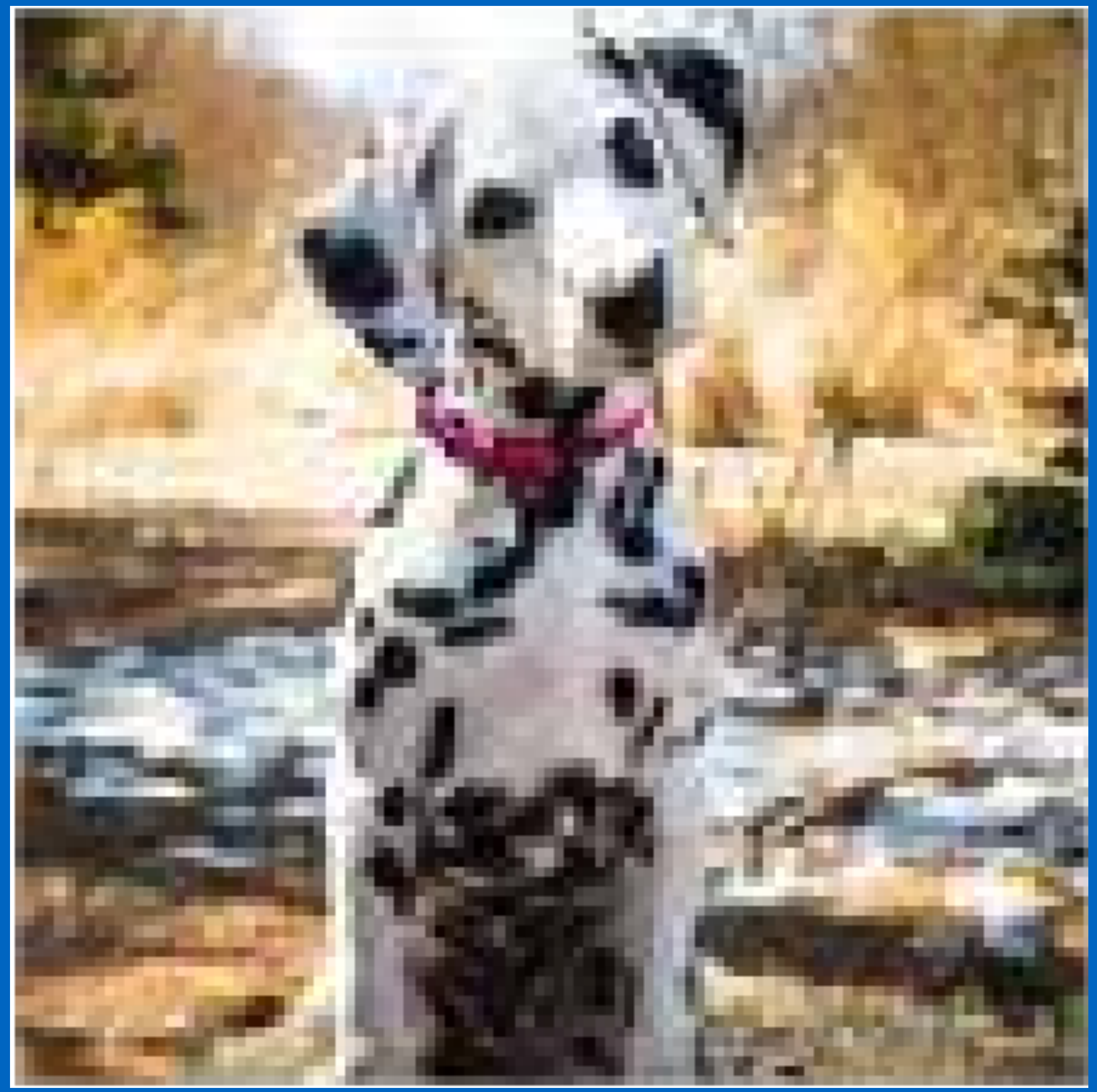
A geometry-inspired decision-based attack

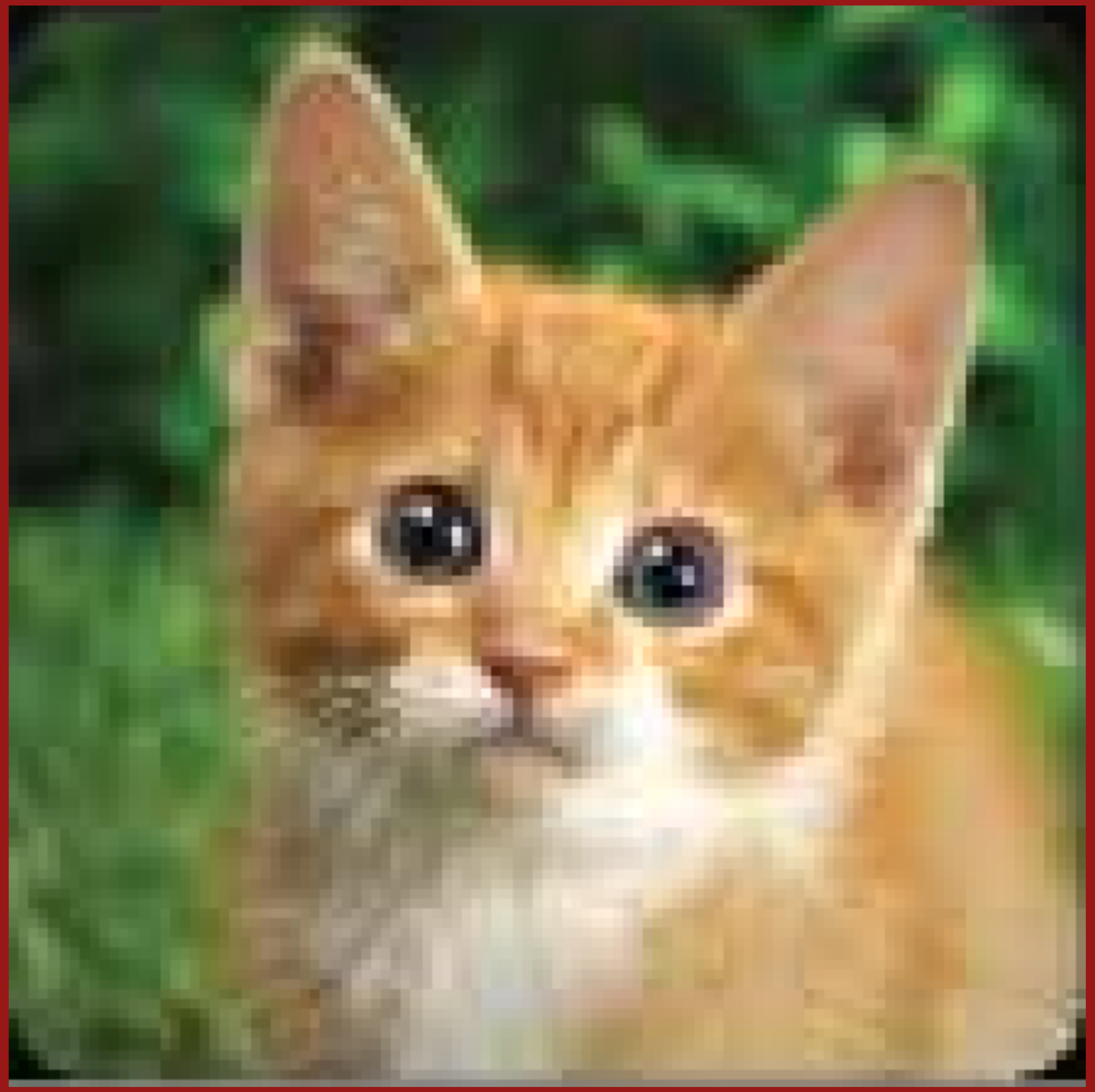
Yujia Liu*
University of Science and Technology of China
Hefei, China
yjcaihon@mail.ustc.edu.cn

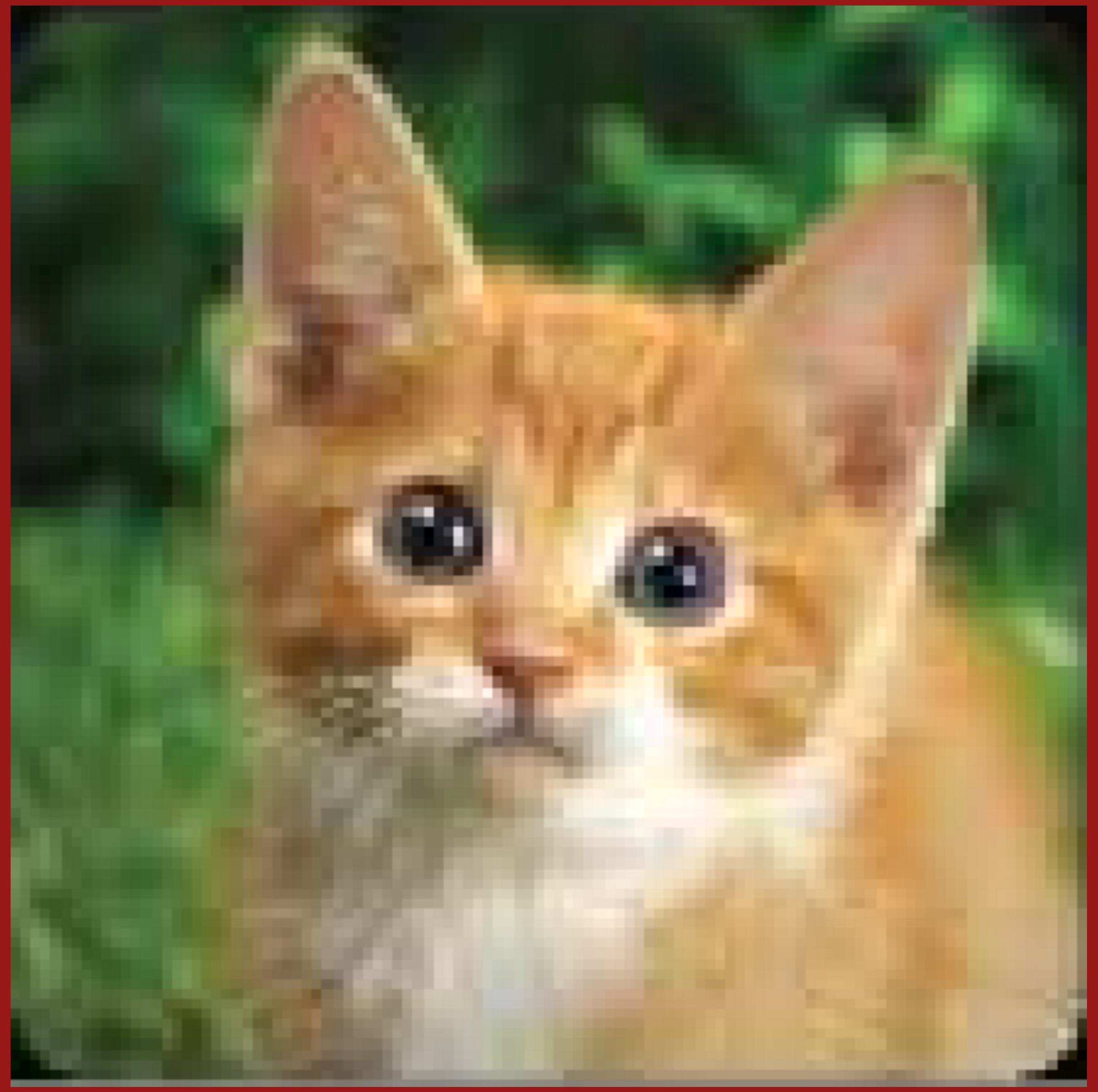
Seyed-Mohsen Moosavi-Dezfooli
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
seyed.moosavi@epfl.ch

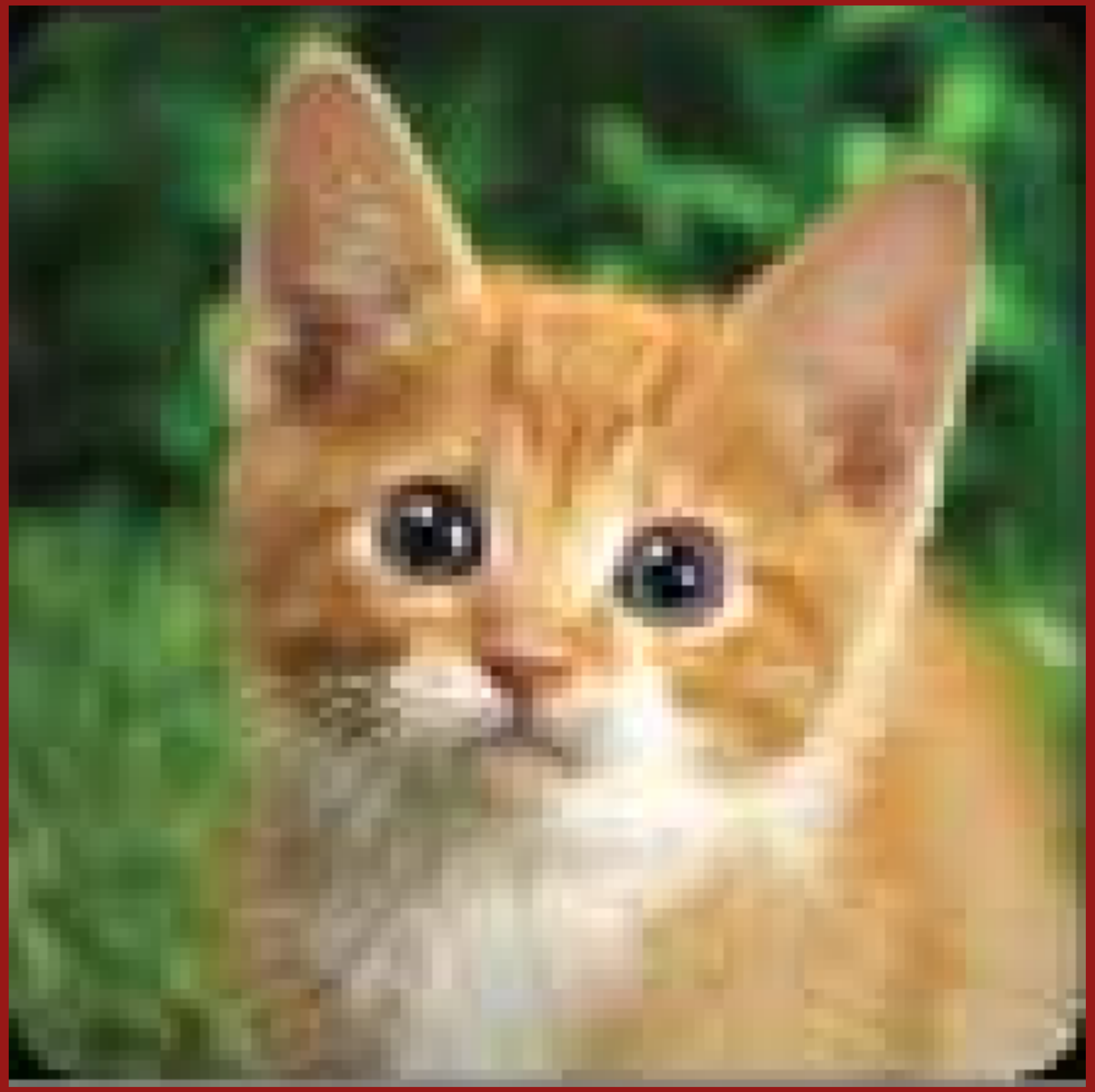
Pascal Frossard
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
pascal.frossard@epfl.ch

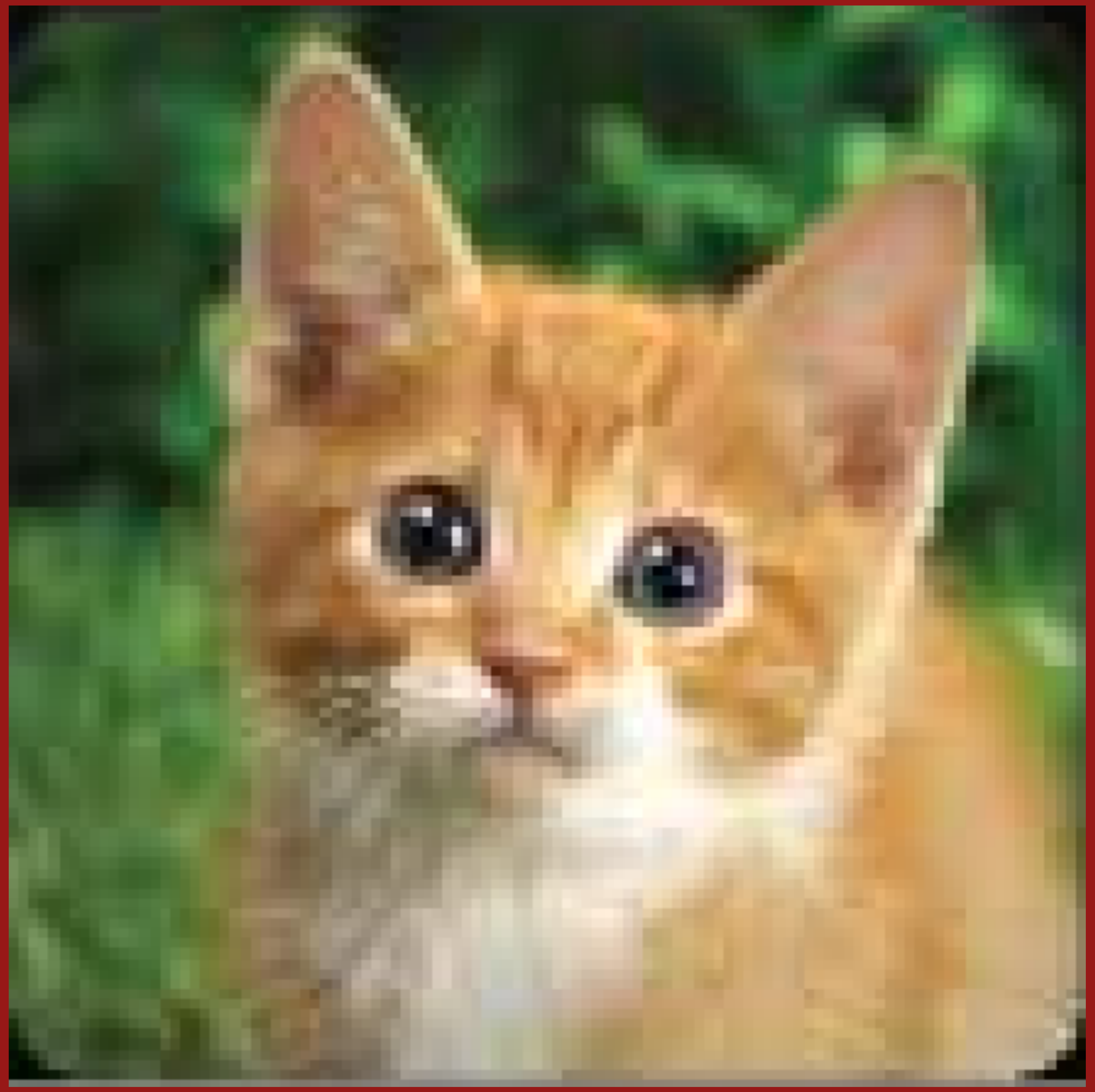


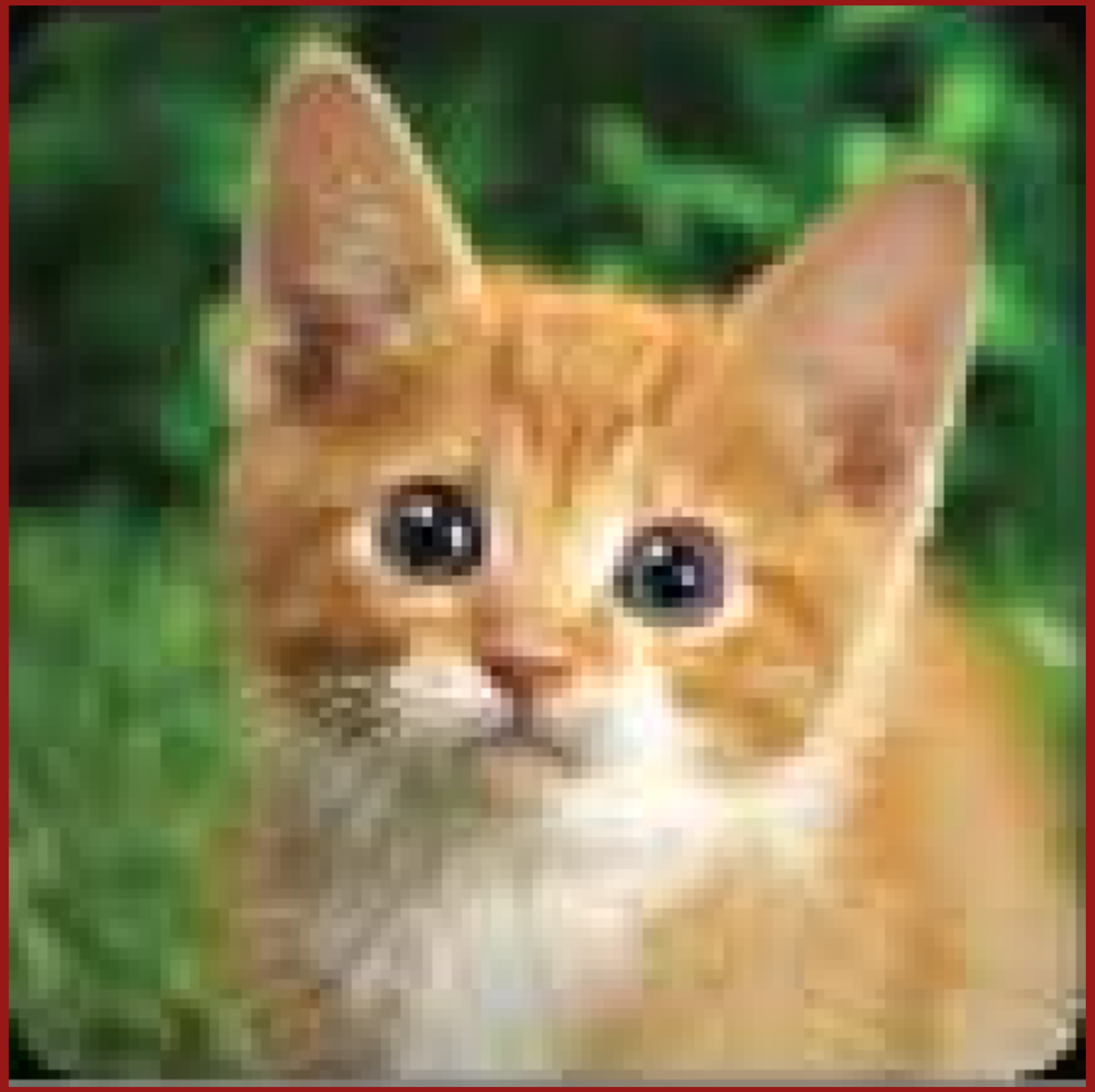


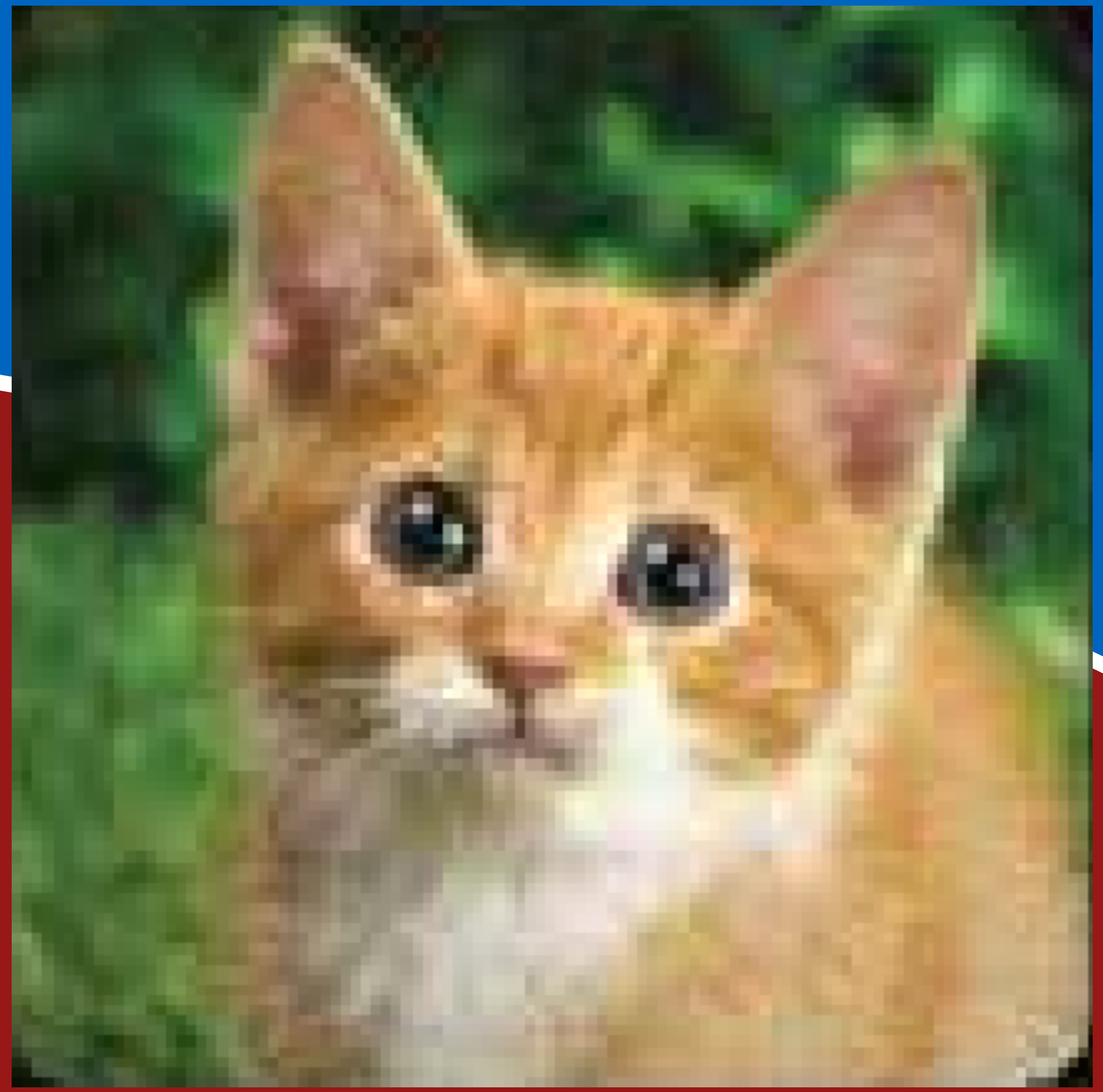


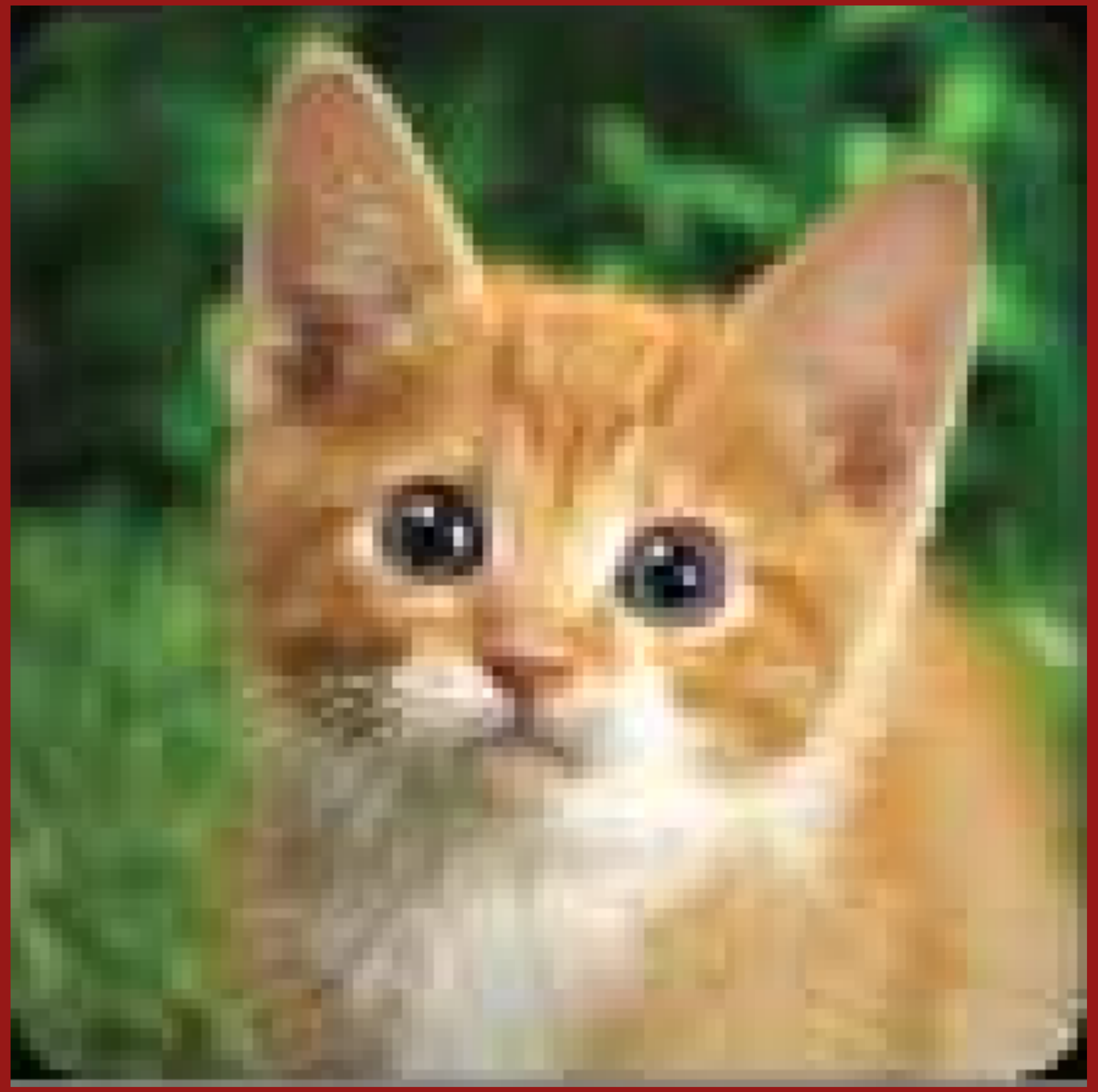






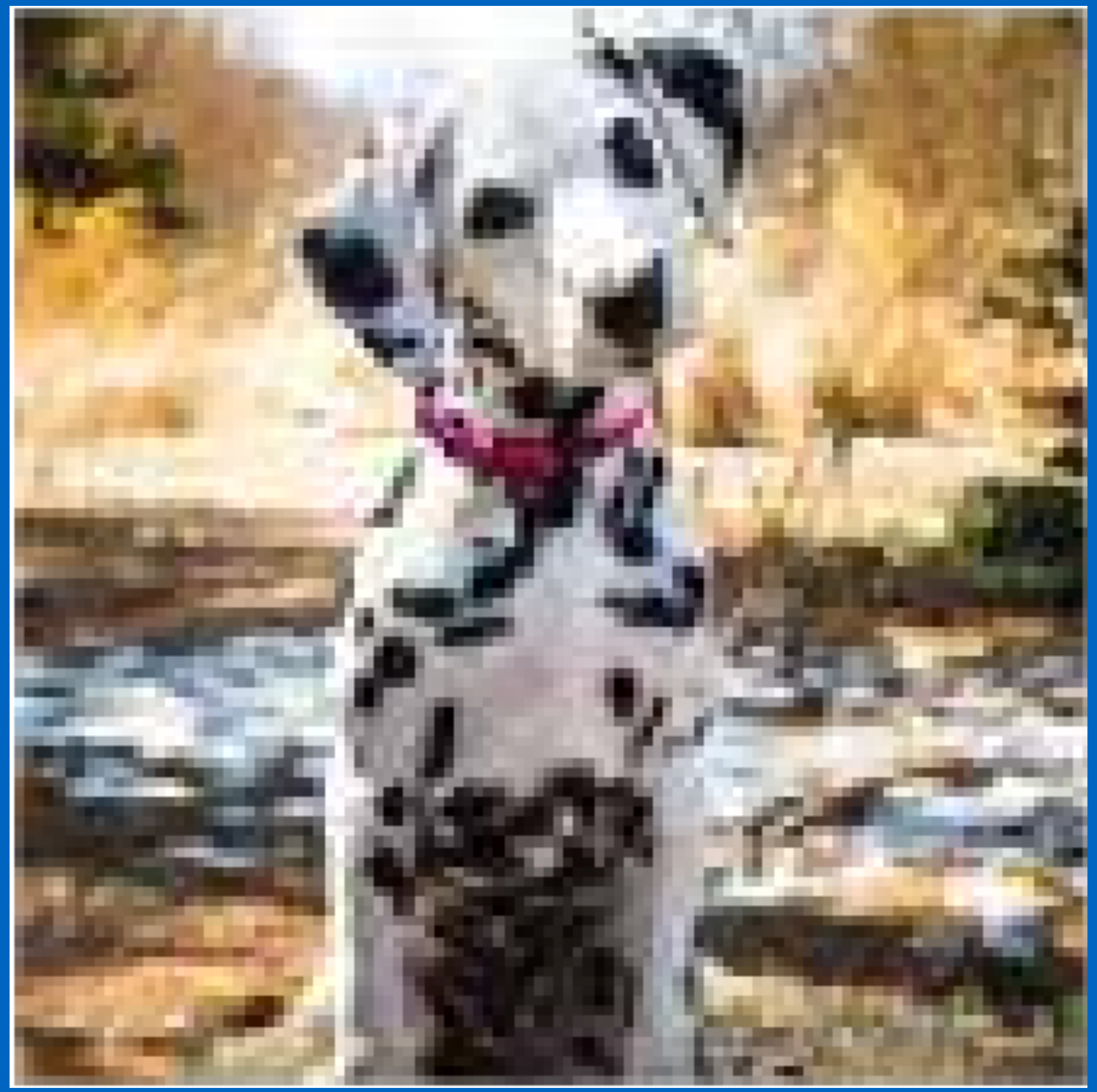


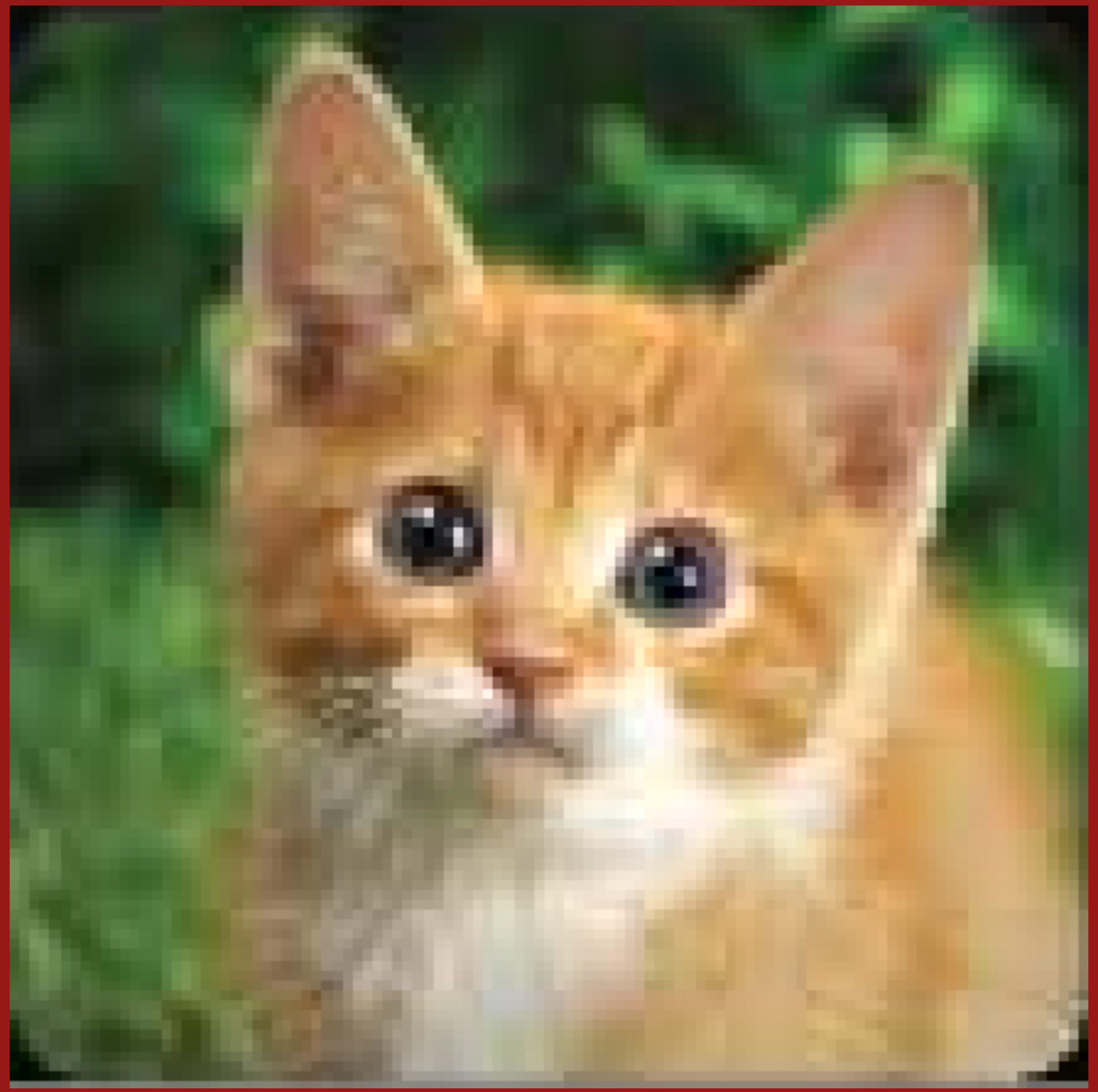


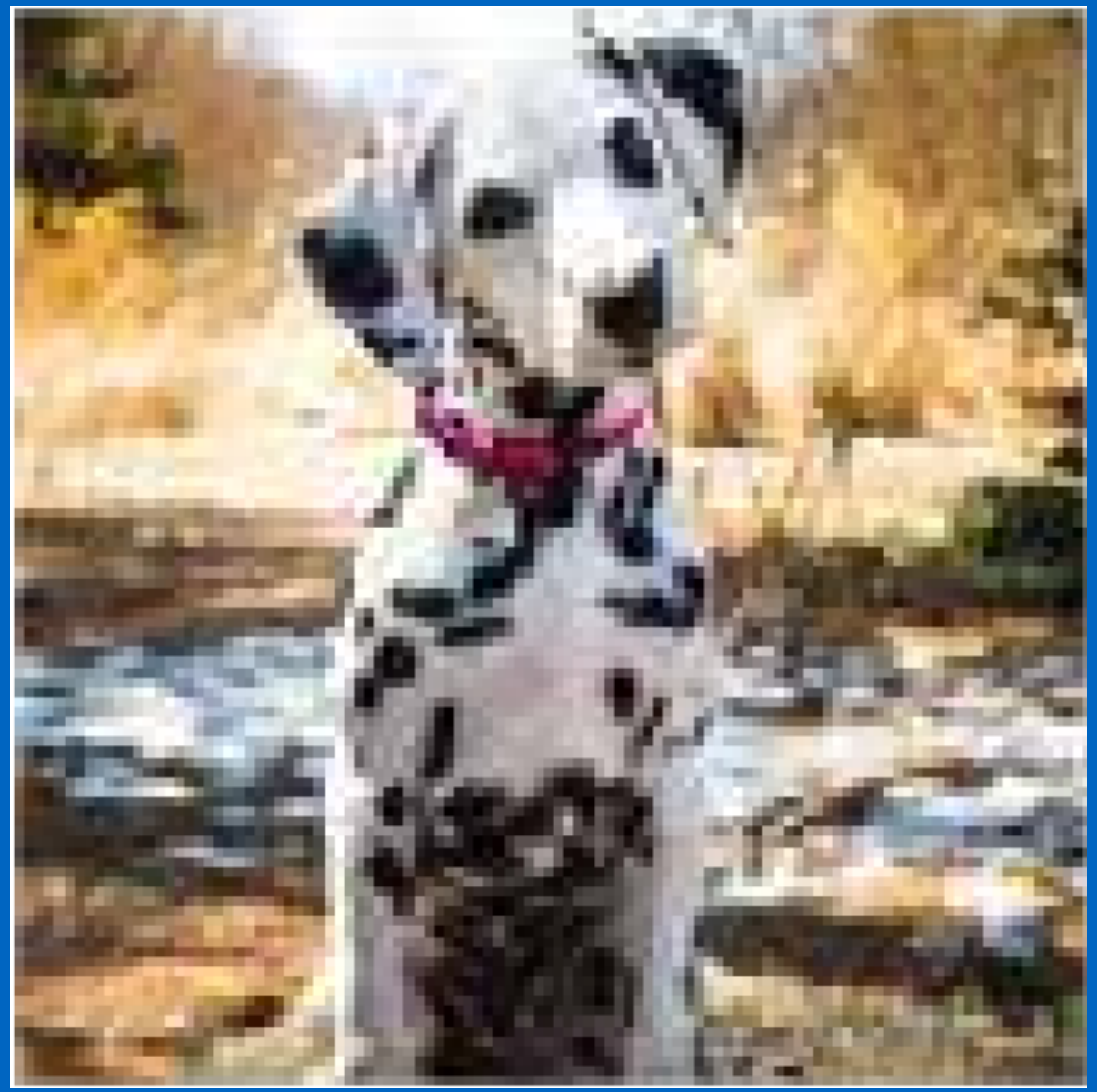


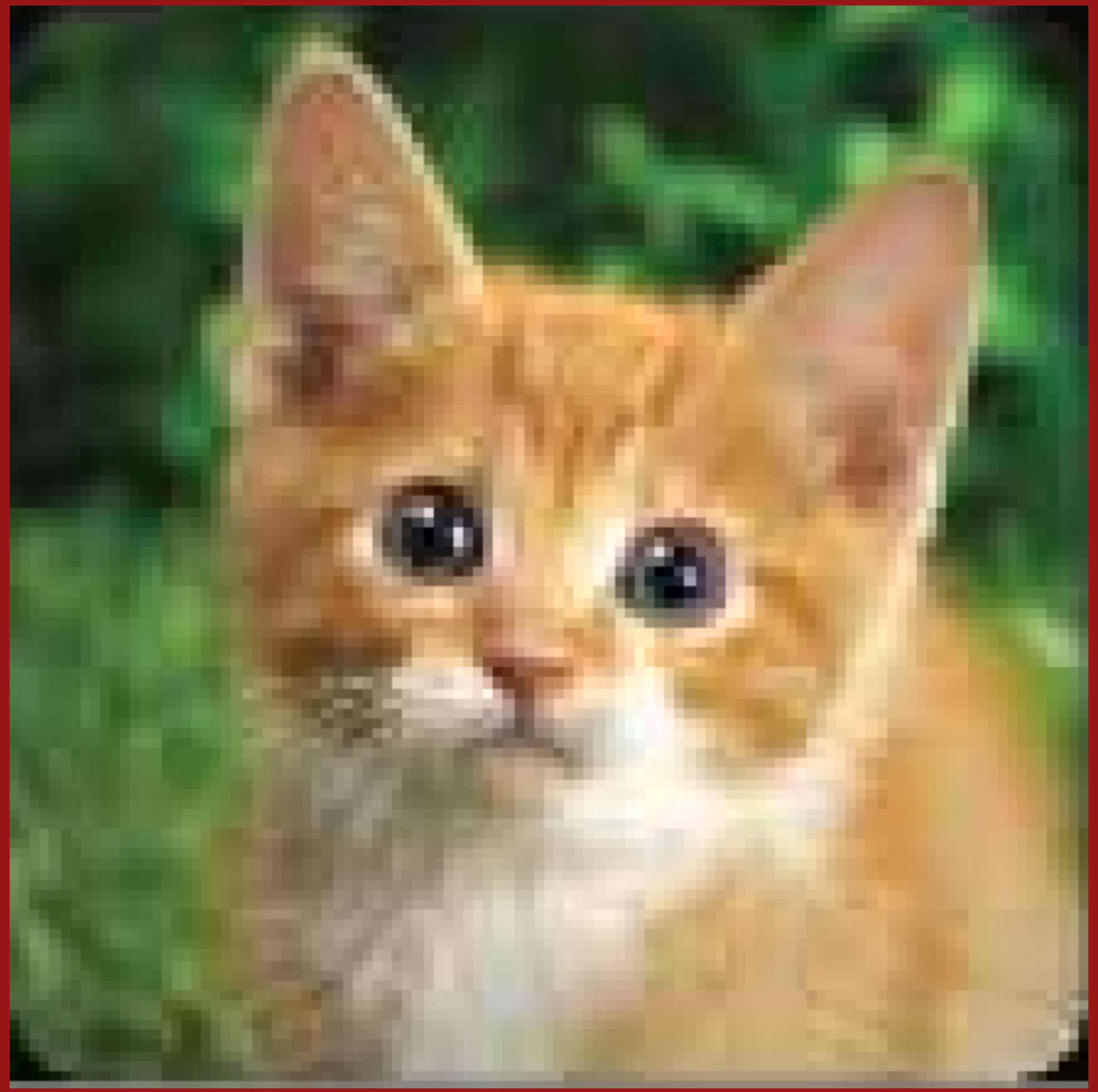
Excessive Invariance Causes Adversarial Vulnerability

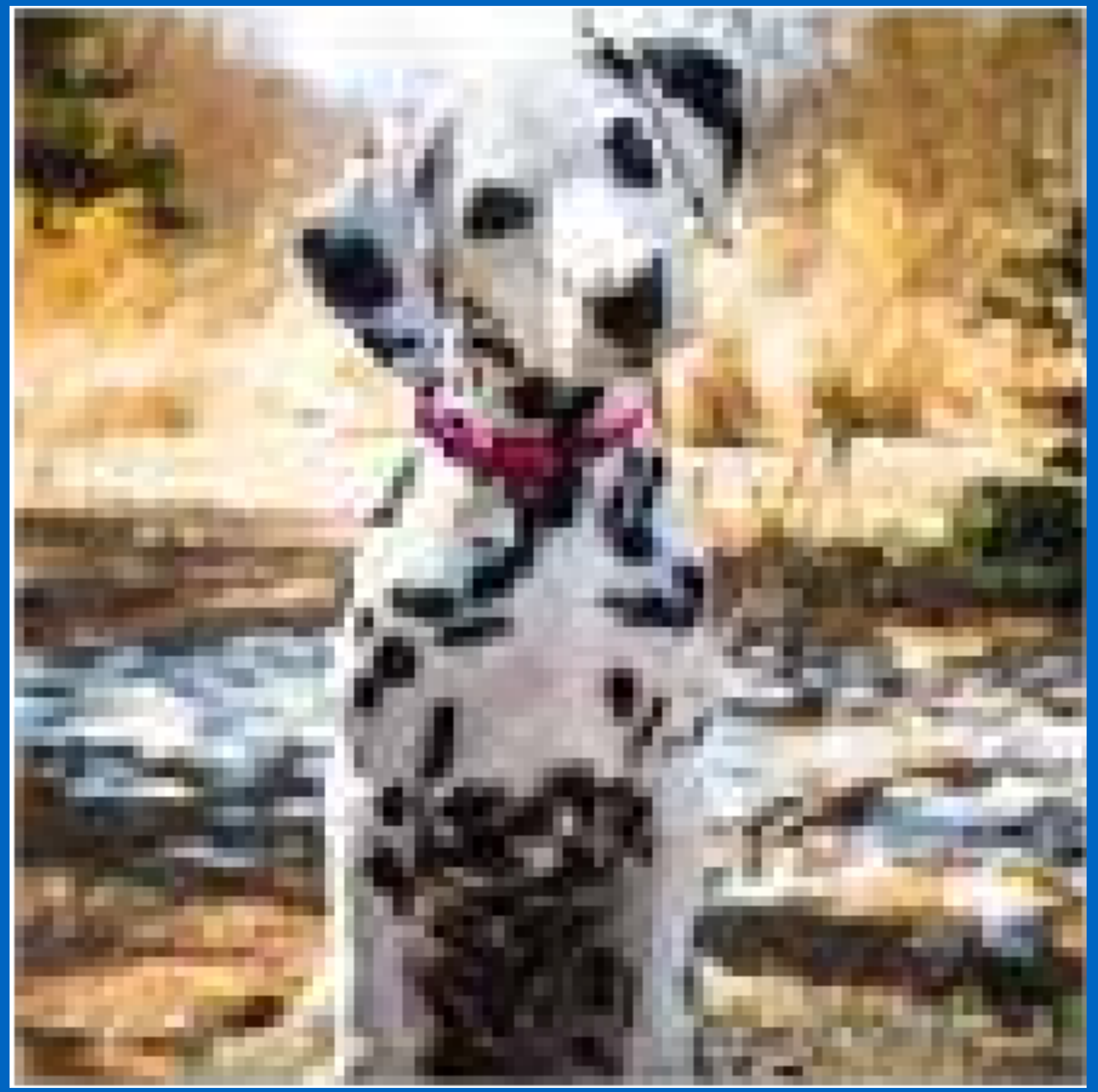
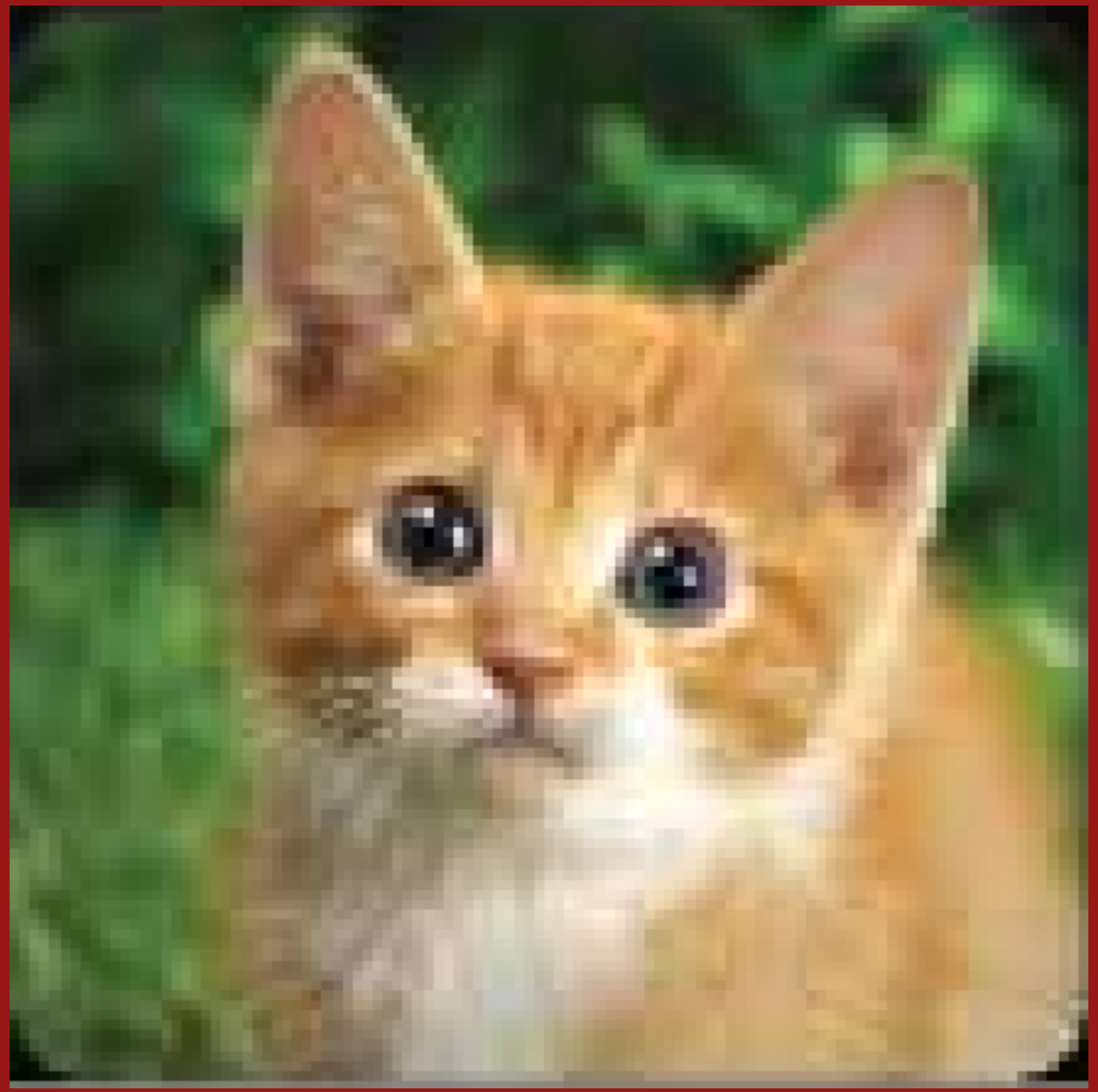
Jörn-Henrik Jacobsen [1*], Jens Behrmann[1,2], Richard Zemel[1], Matthias Bethge[3]
[1]Vector Institute and University of Toronto
[2]University of Bremen, Center for Industrial Mathematics
[3]University of Tübingen

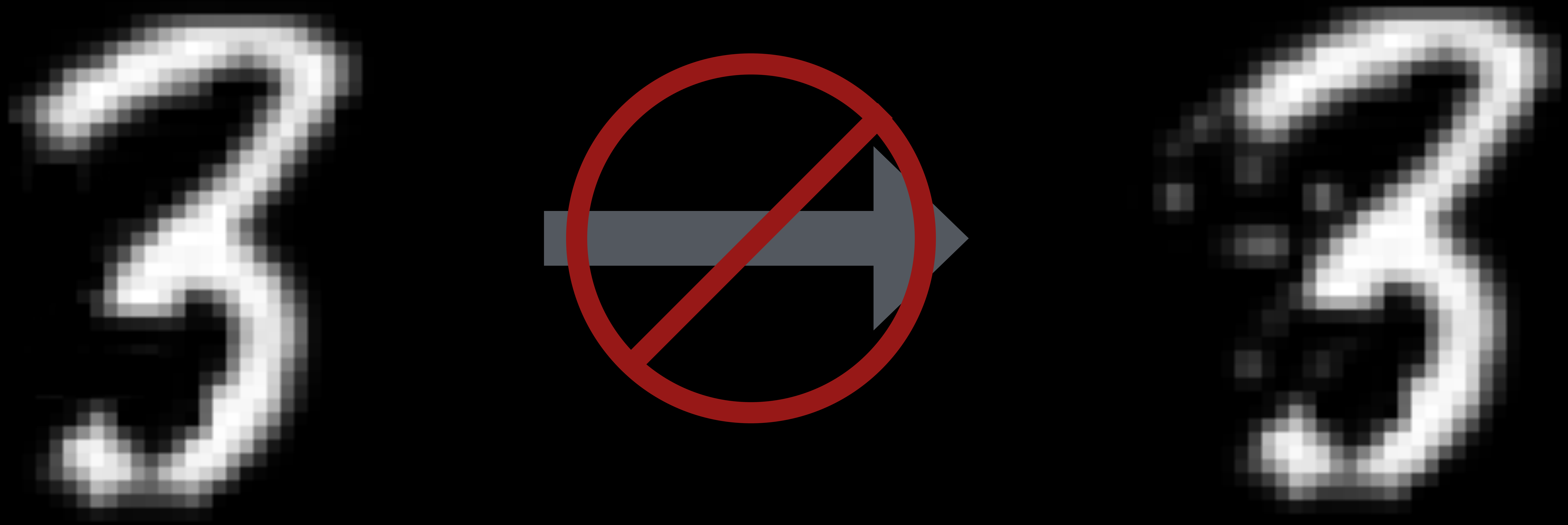


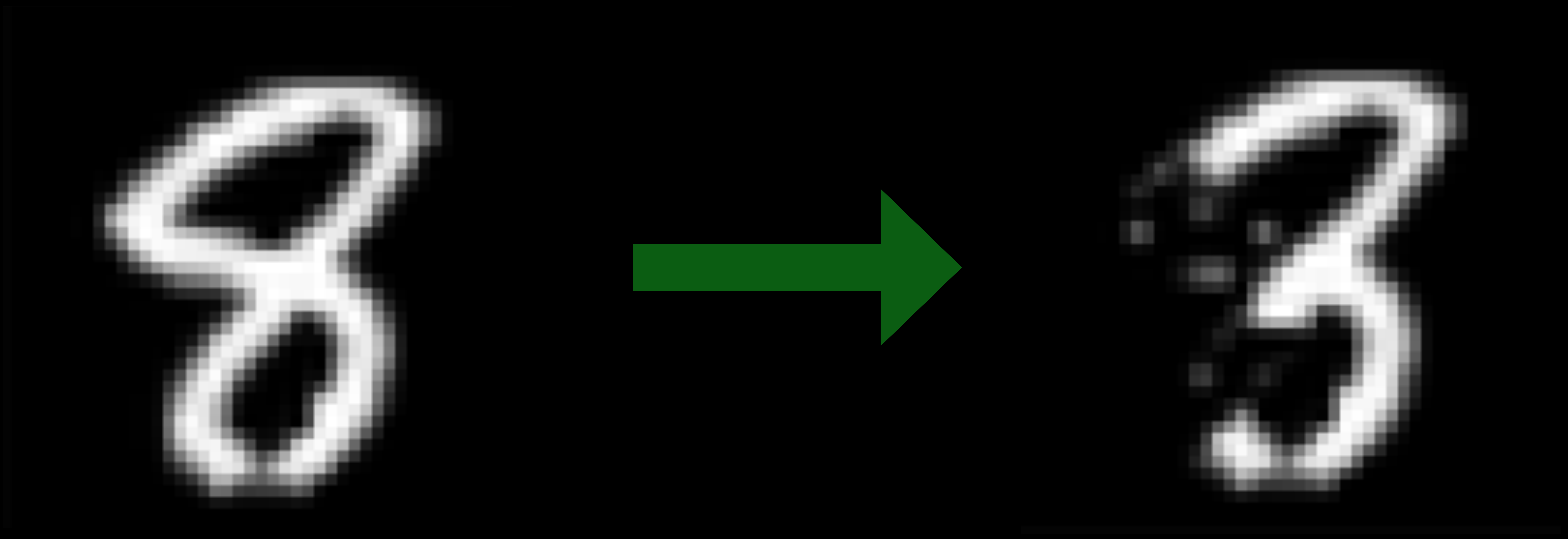


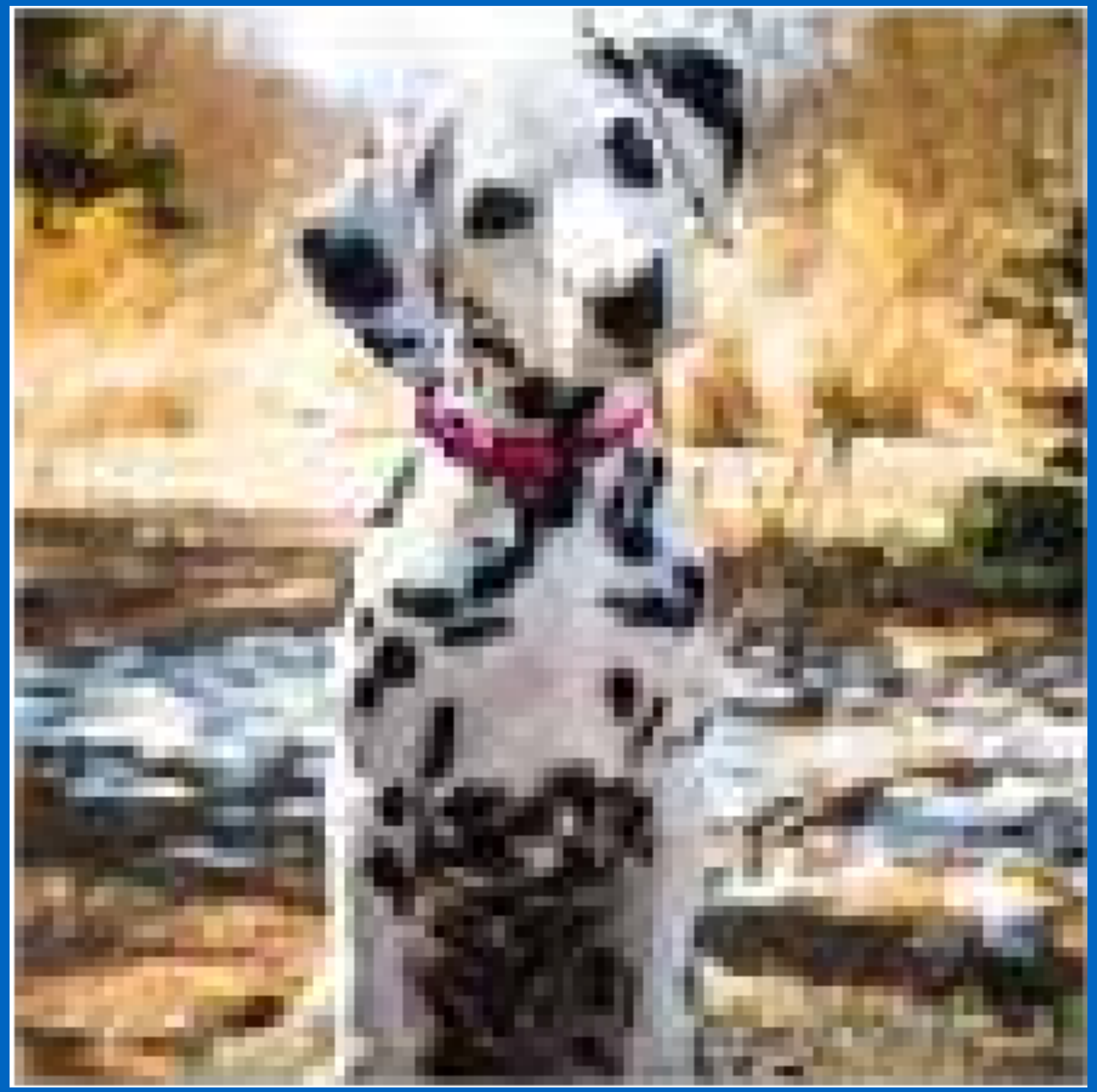
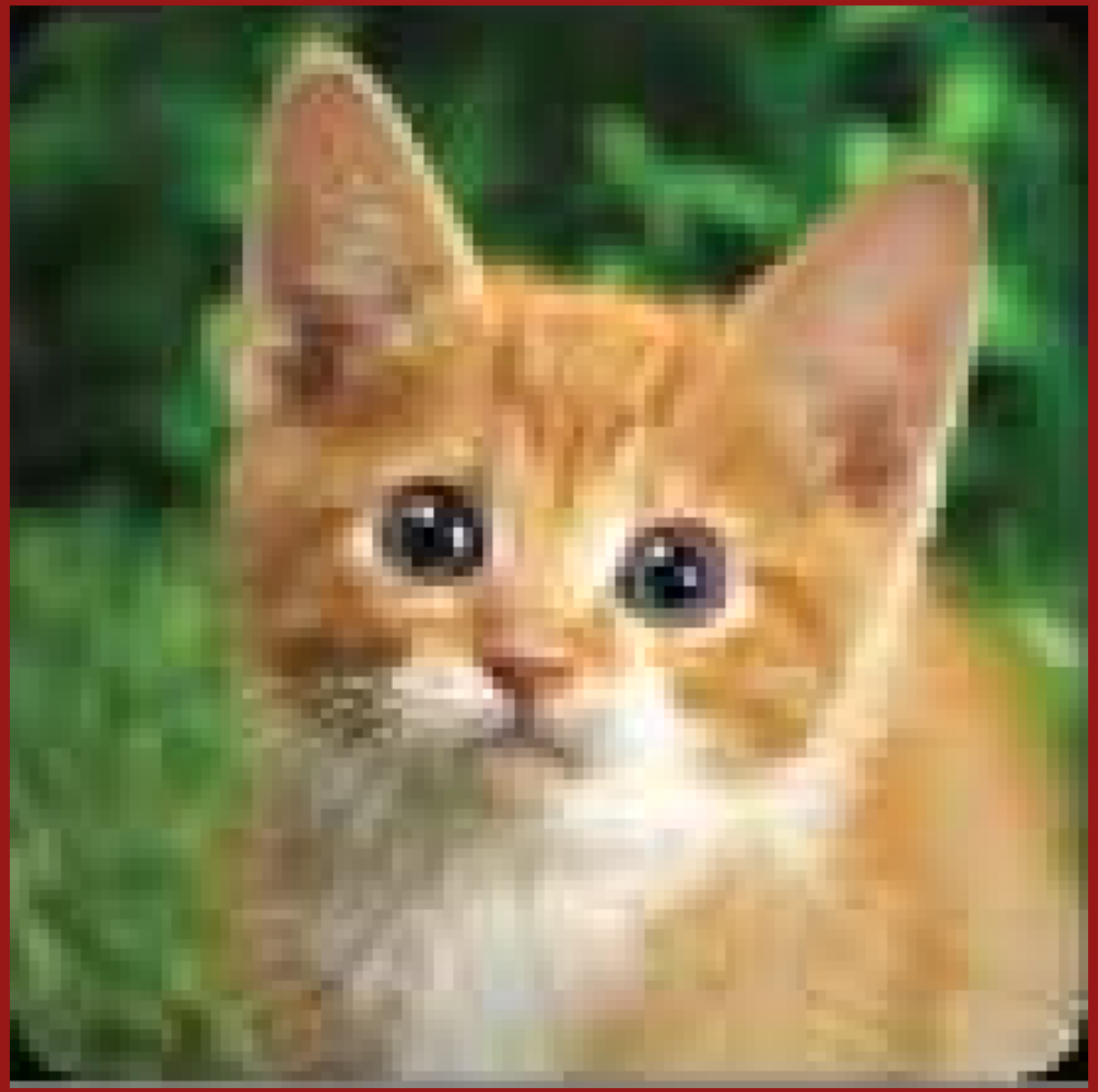


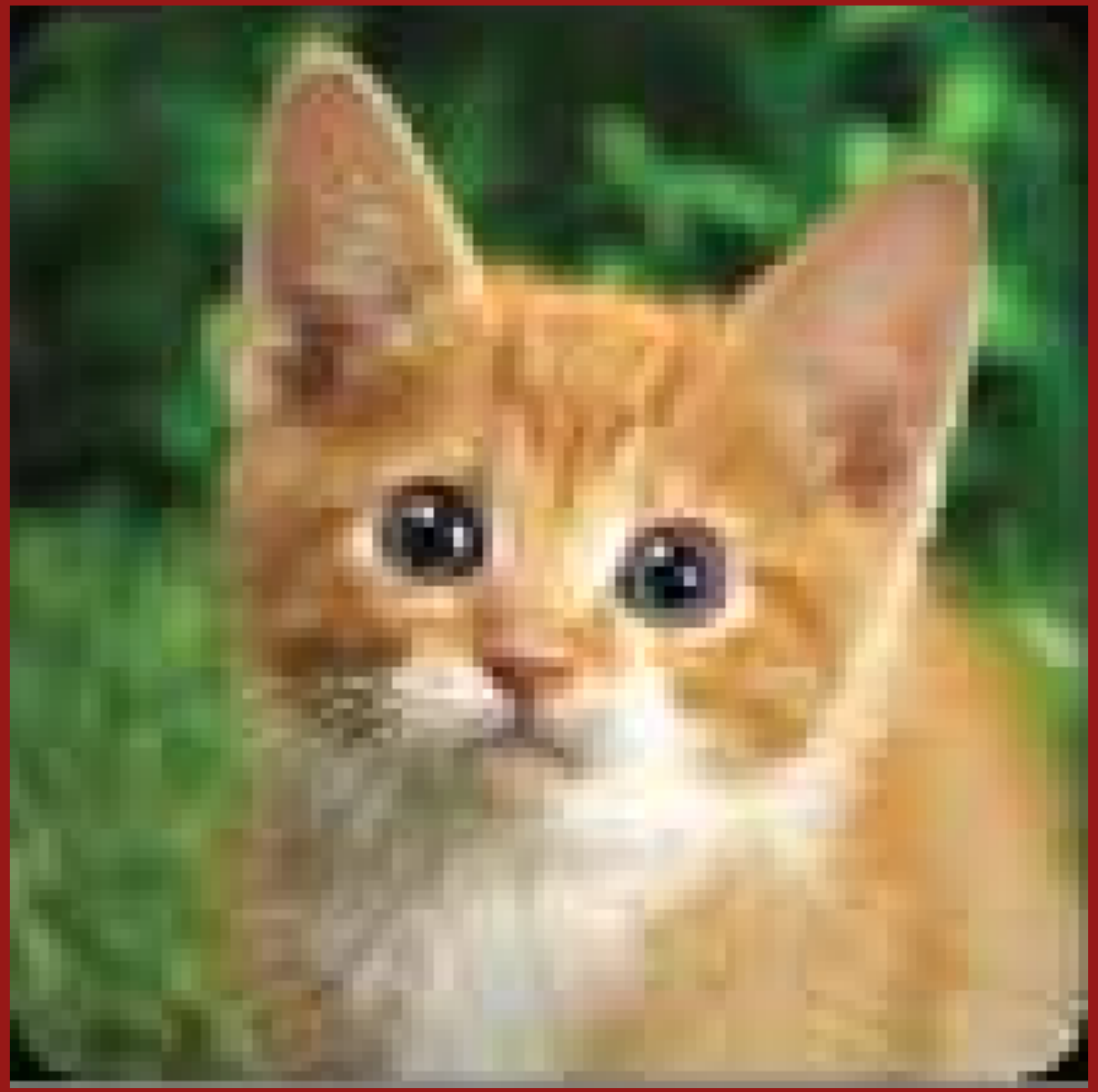


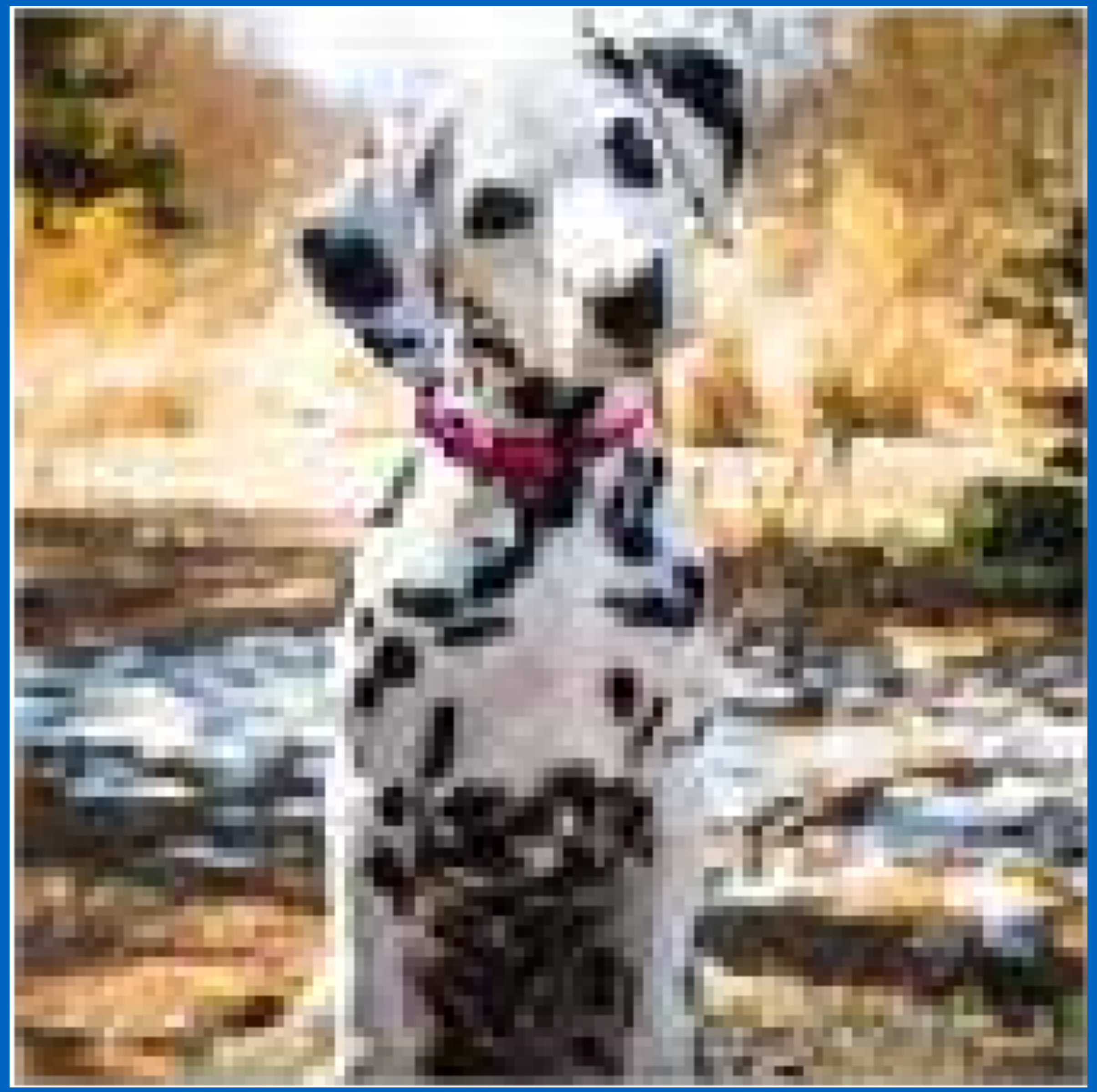







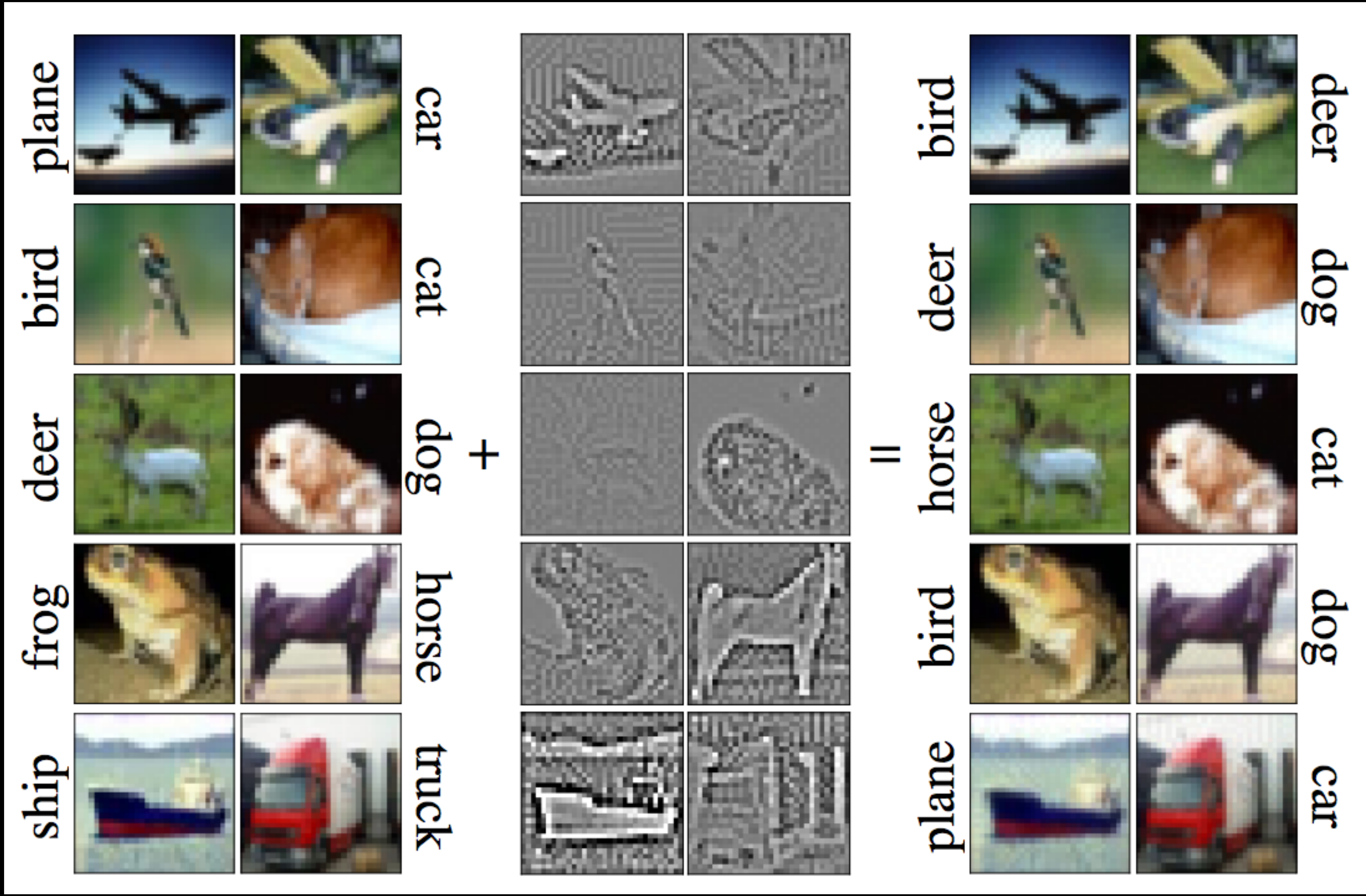







Wasserstein Adversarial Examples via Projected Sinkhorn Iterations

Eric Wong [1] Frank R. Schmidt [2] J. Zico Kolter [3 4]


plane
car
bird
cat
deer
dog
frog
horse
ship
truck
+
=
bird
deer
deer
dog
horse
cat
bird
dog
plane
car

Recent advances in …

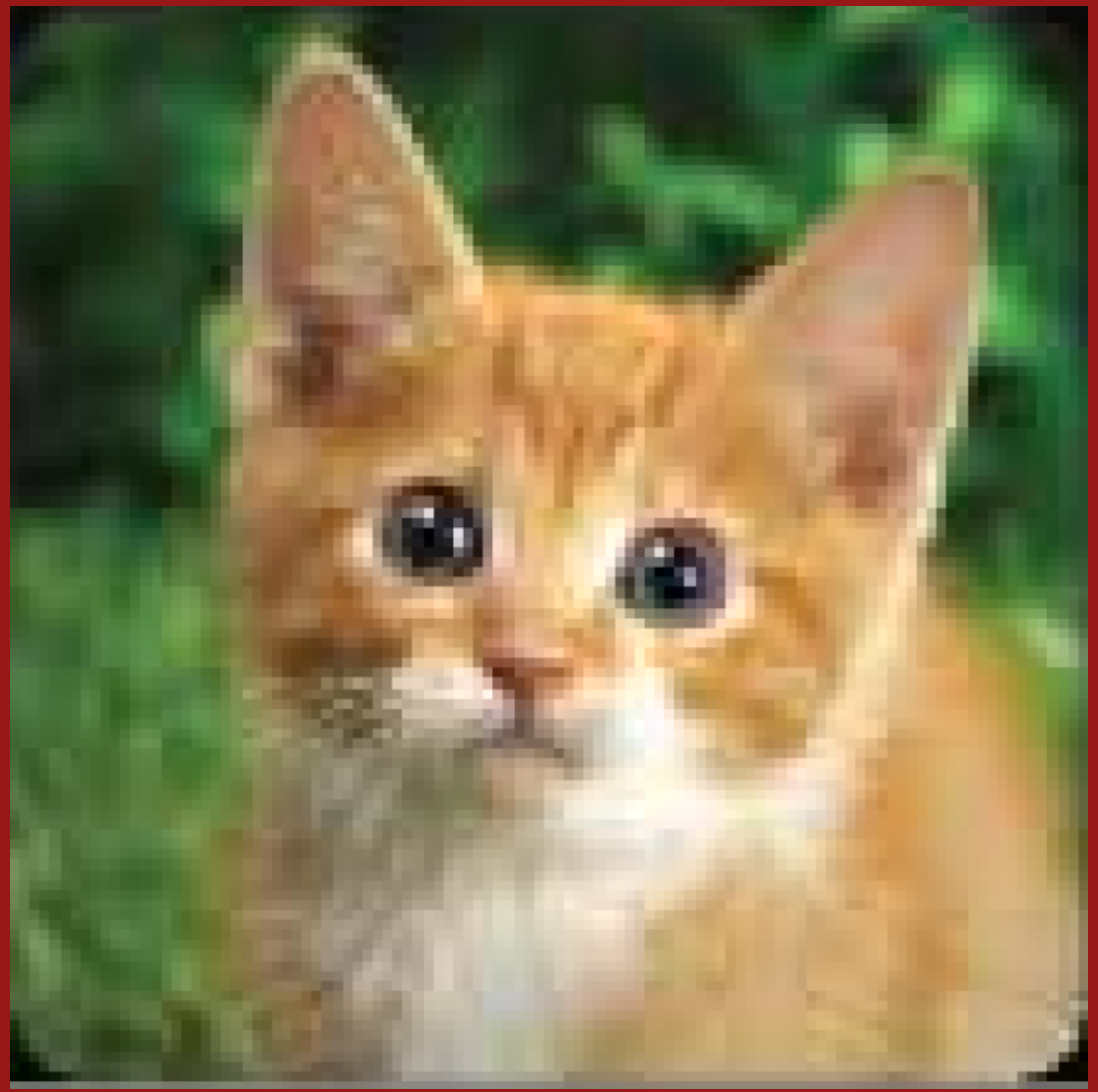
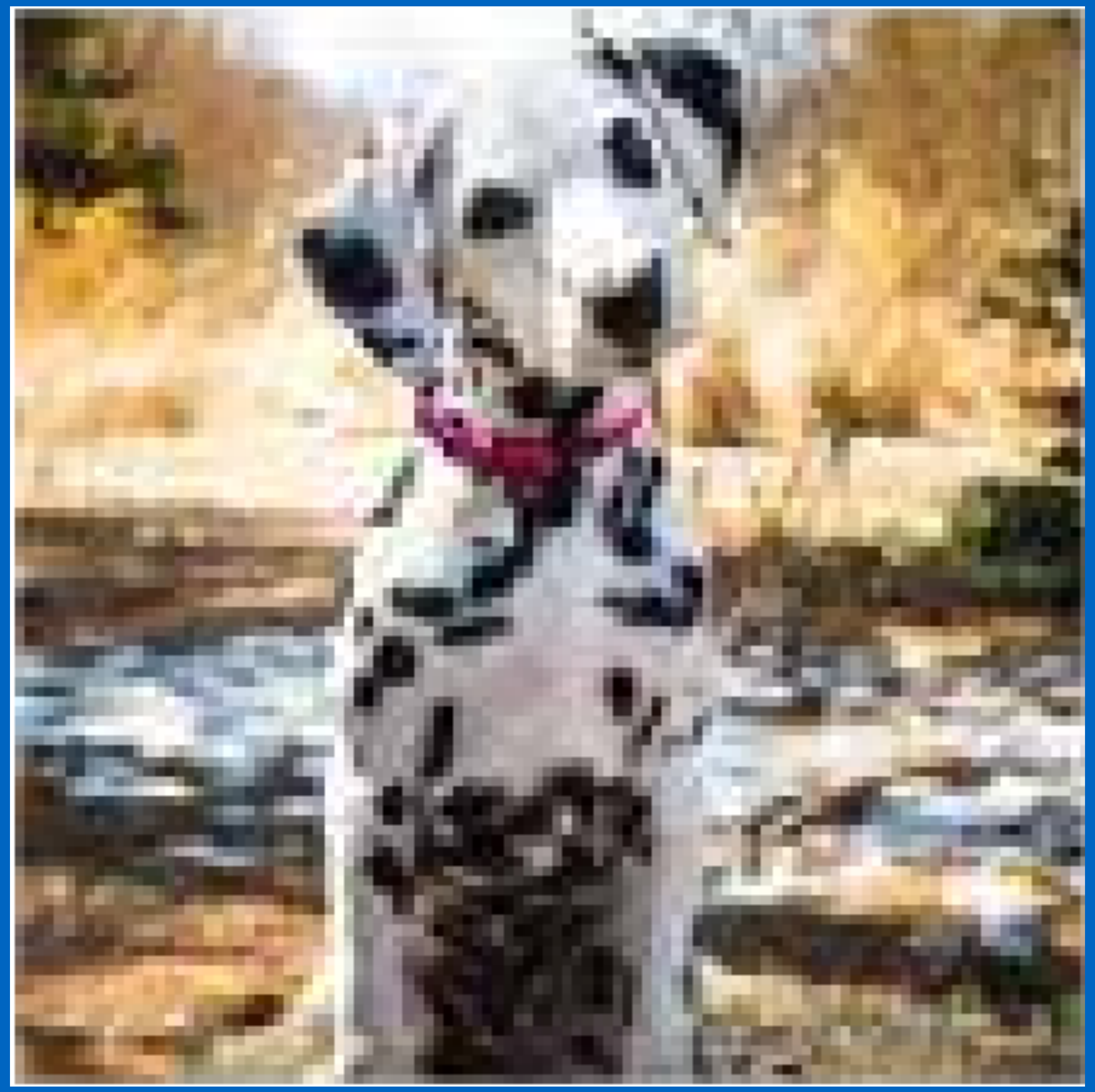
Defending Against

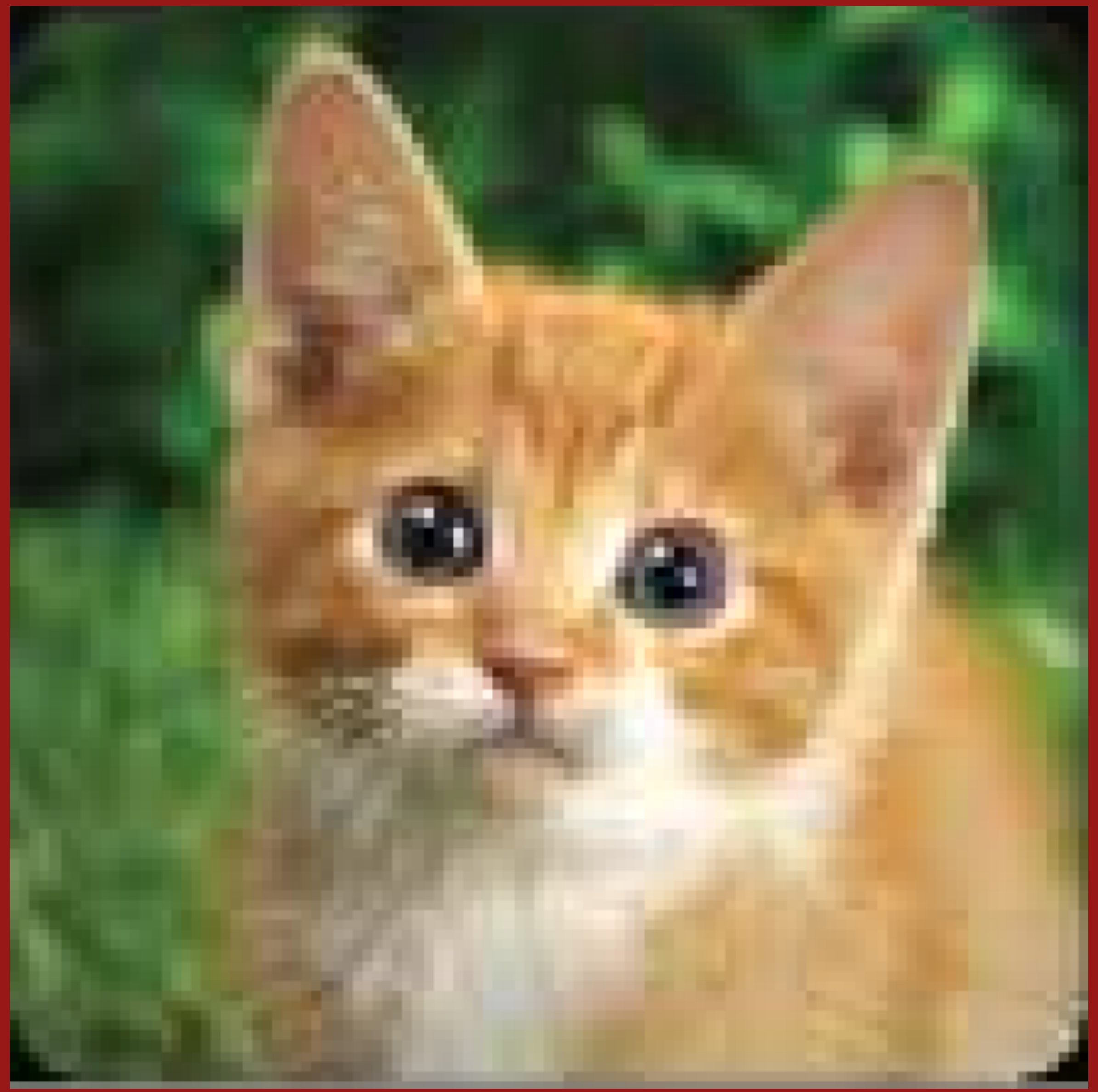
Adversarial Examples

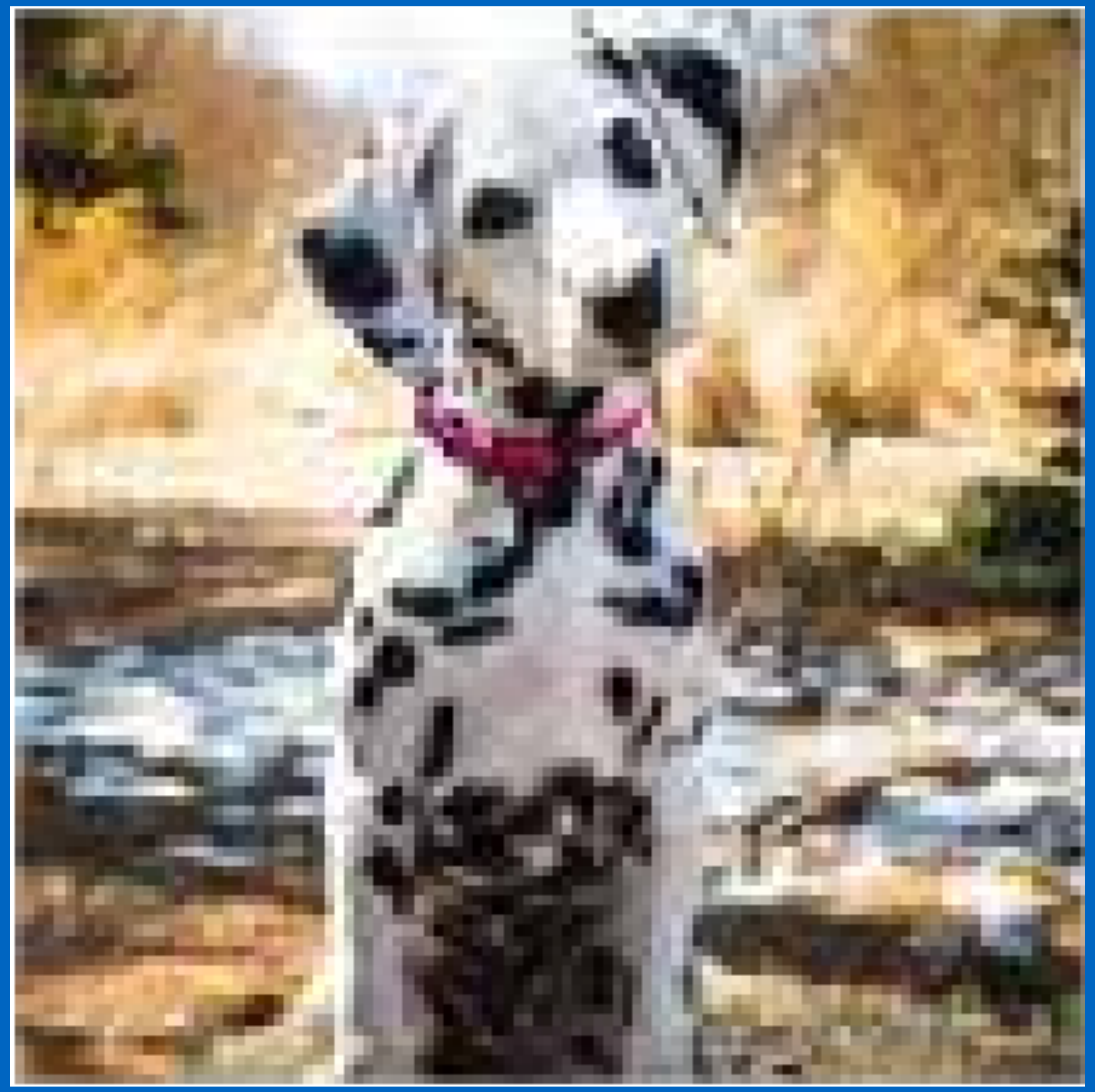
Defenses I *don't* believe will be effective

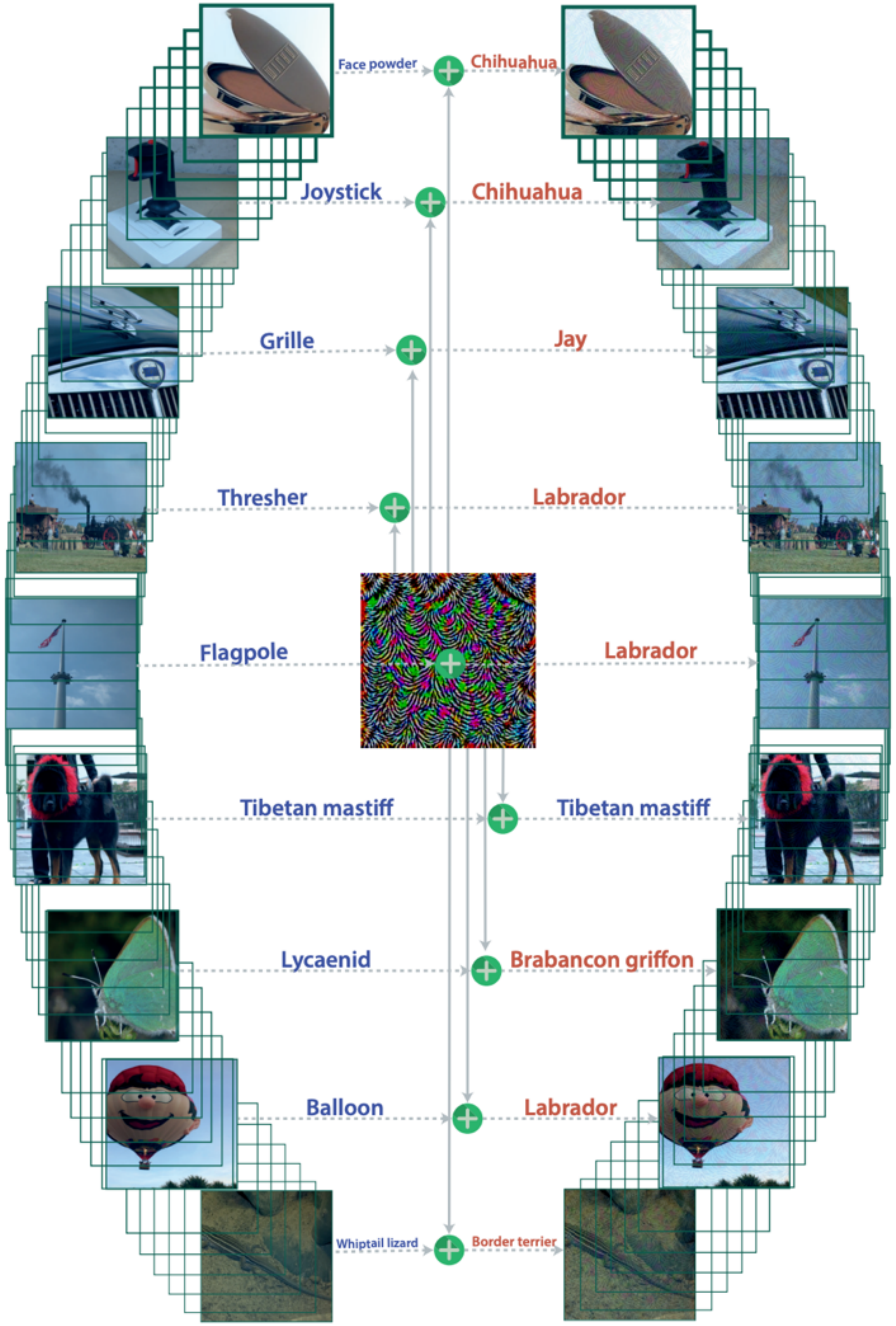
... a bit more background

Transferability

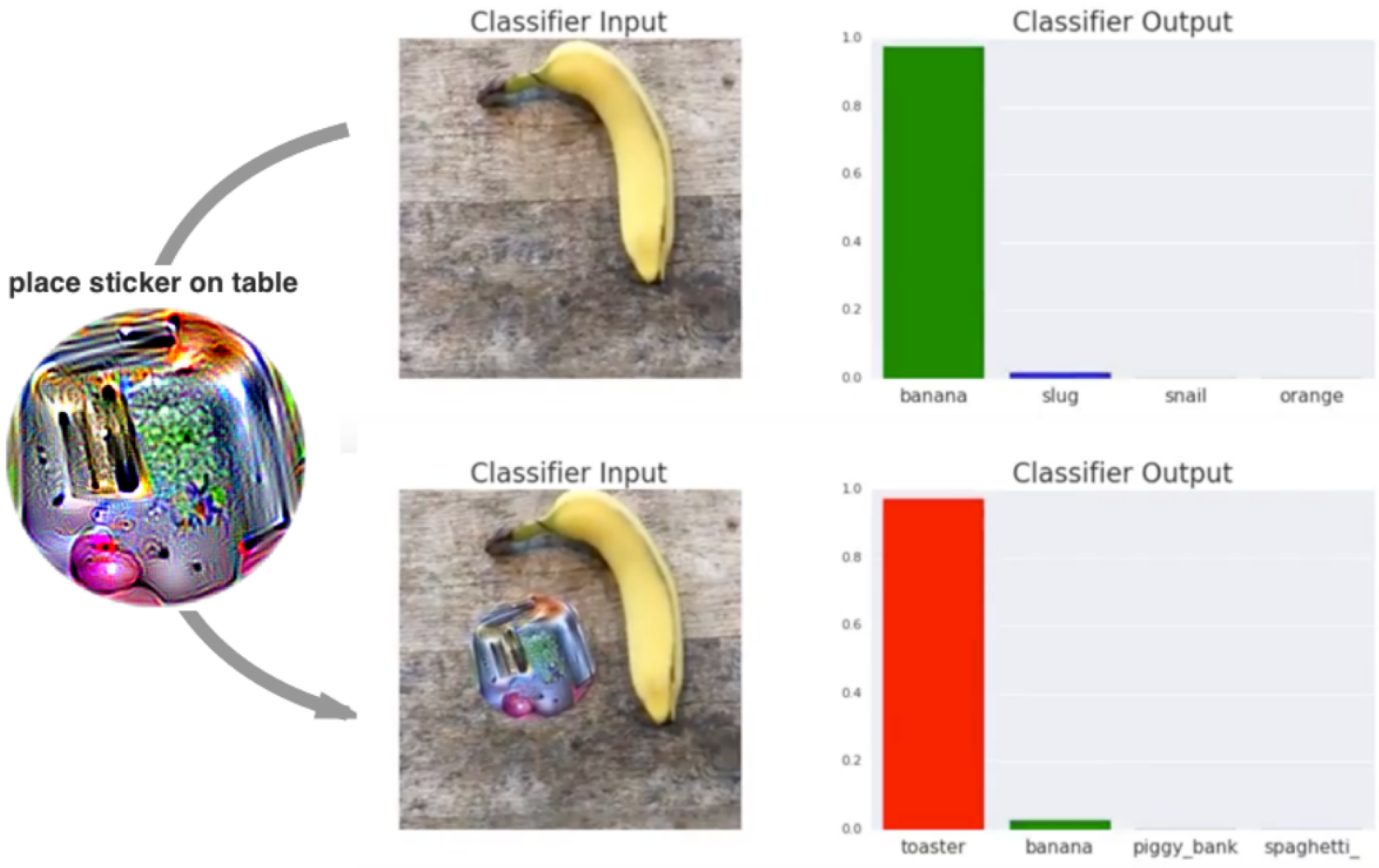







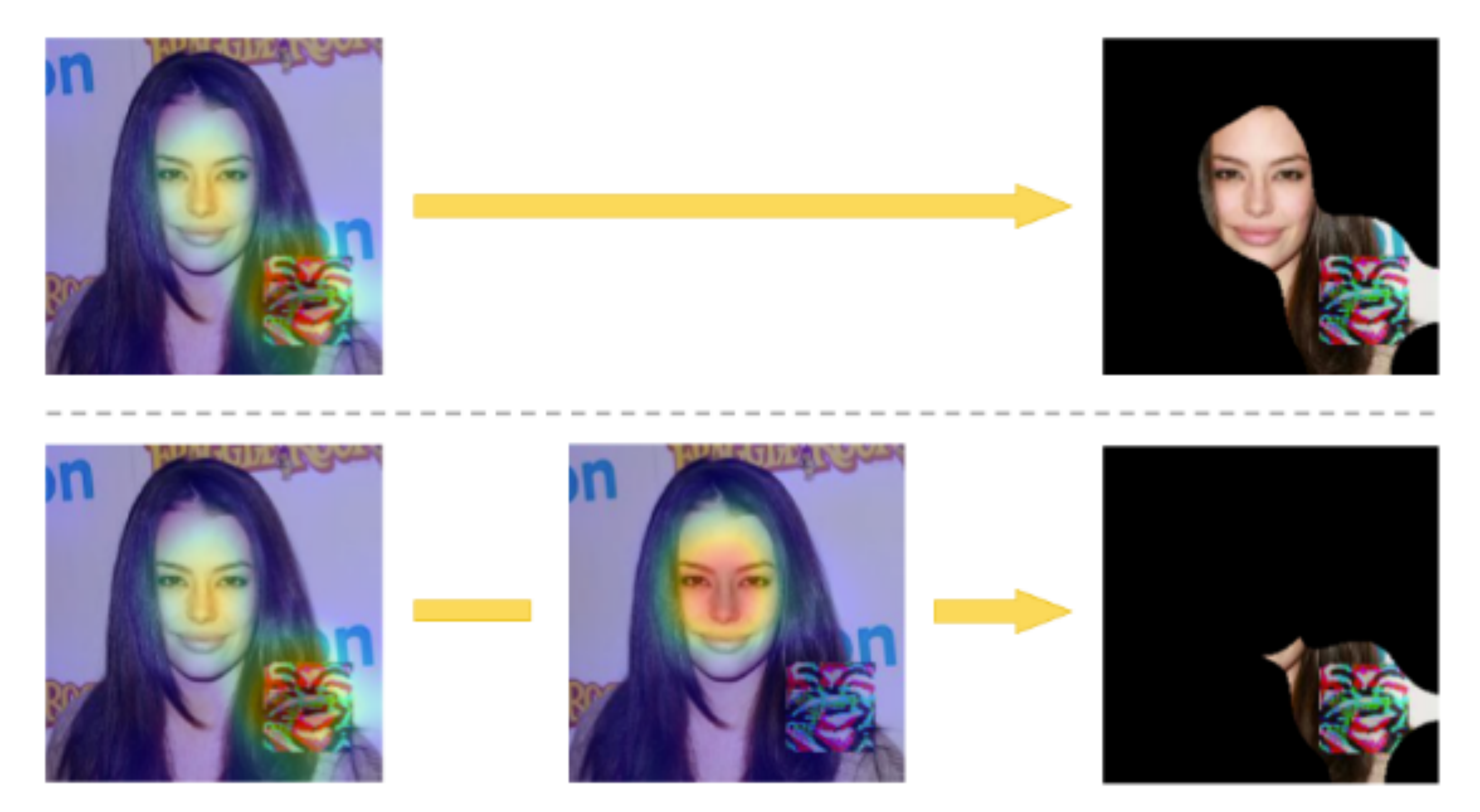


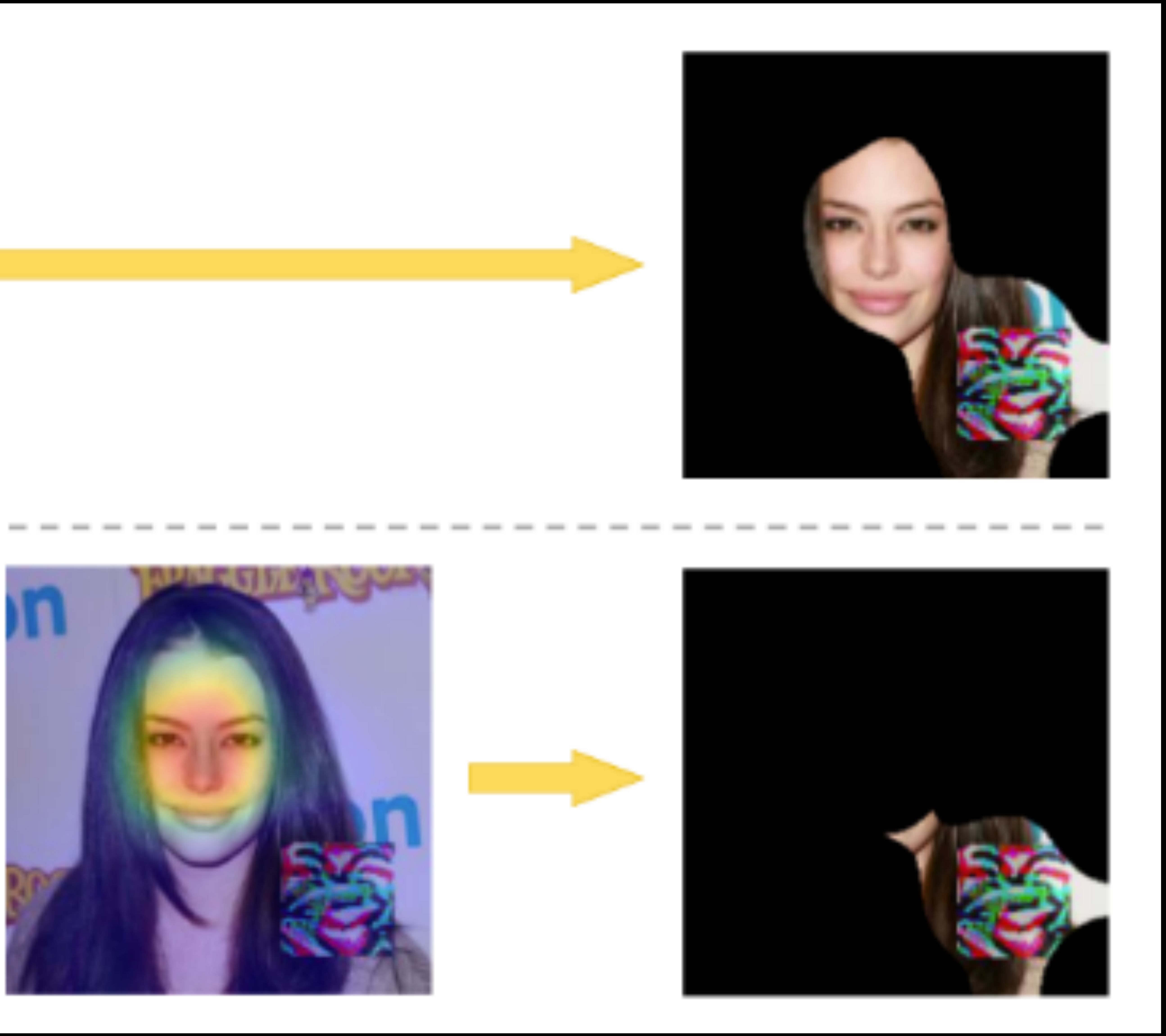
Face powder
Chihuahua
Joystick
Chihuahua
Grille
Jay
Thresher
Labrador
Flagpole
Labrador
Tibetan mastiff
Tibetan mastiff
Lycaenid
Brabancon griffon
Balloon
Labrador
Whiptail lizard
Border terrier


place sticker on table
Classifier Input
Classifier Output
1.0
0.8
0.6
0.4
0.2
0.0
banana
slug
snail
orange
Classifier Input
Classifier Output
1.0
0.8
0.6
0.4
0.2
0.0
toaster
banana
piggy_bank
spaghetti_

SentiNet: Detecting Physical Attacks Against Deep Learning Systems

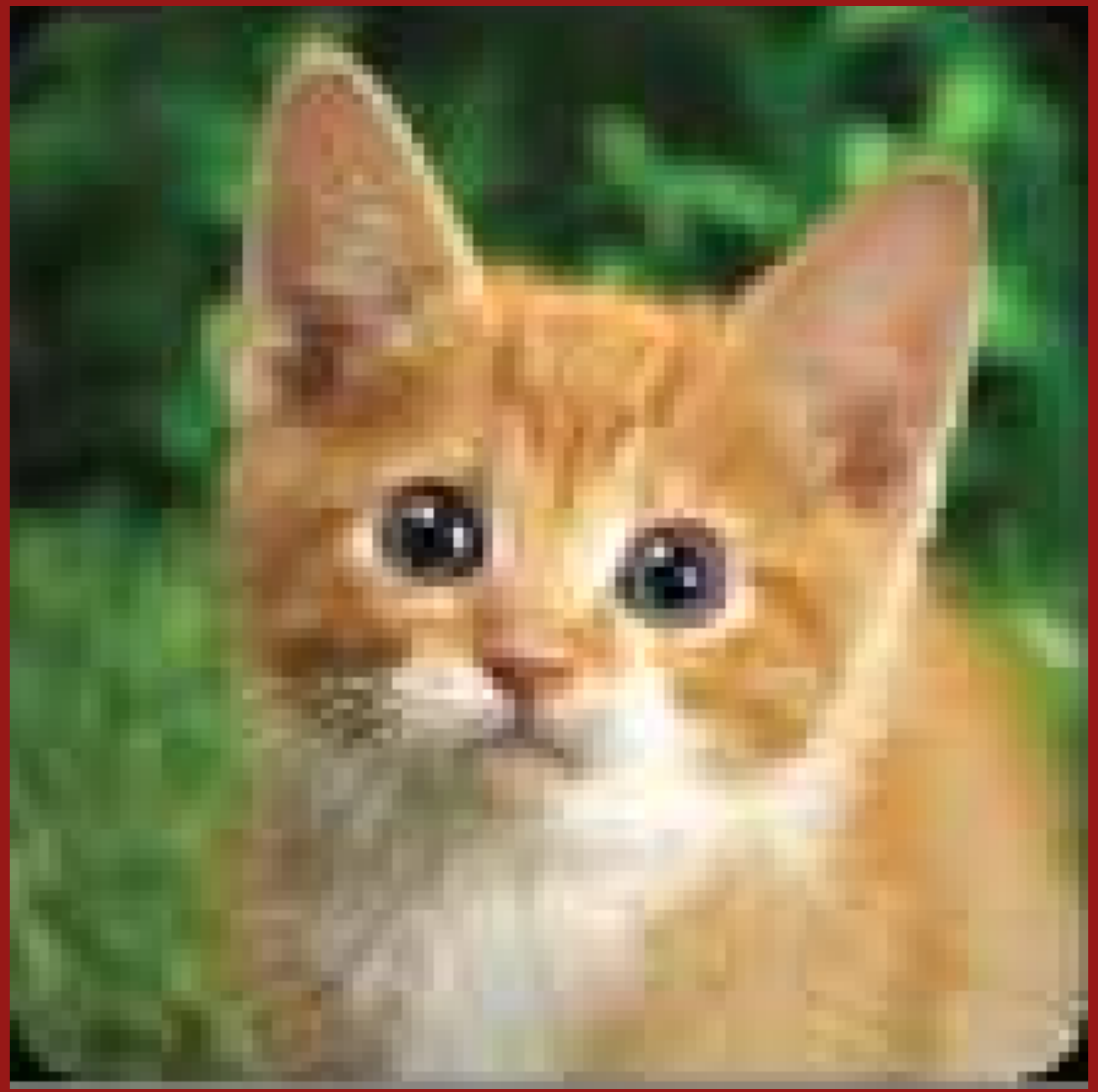
Edward Chou[1] Florian Tramèr[1] Giancarlo Pellegrino[1,2] Dan Boneh[1]

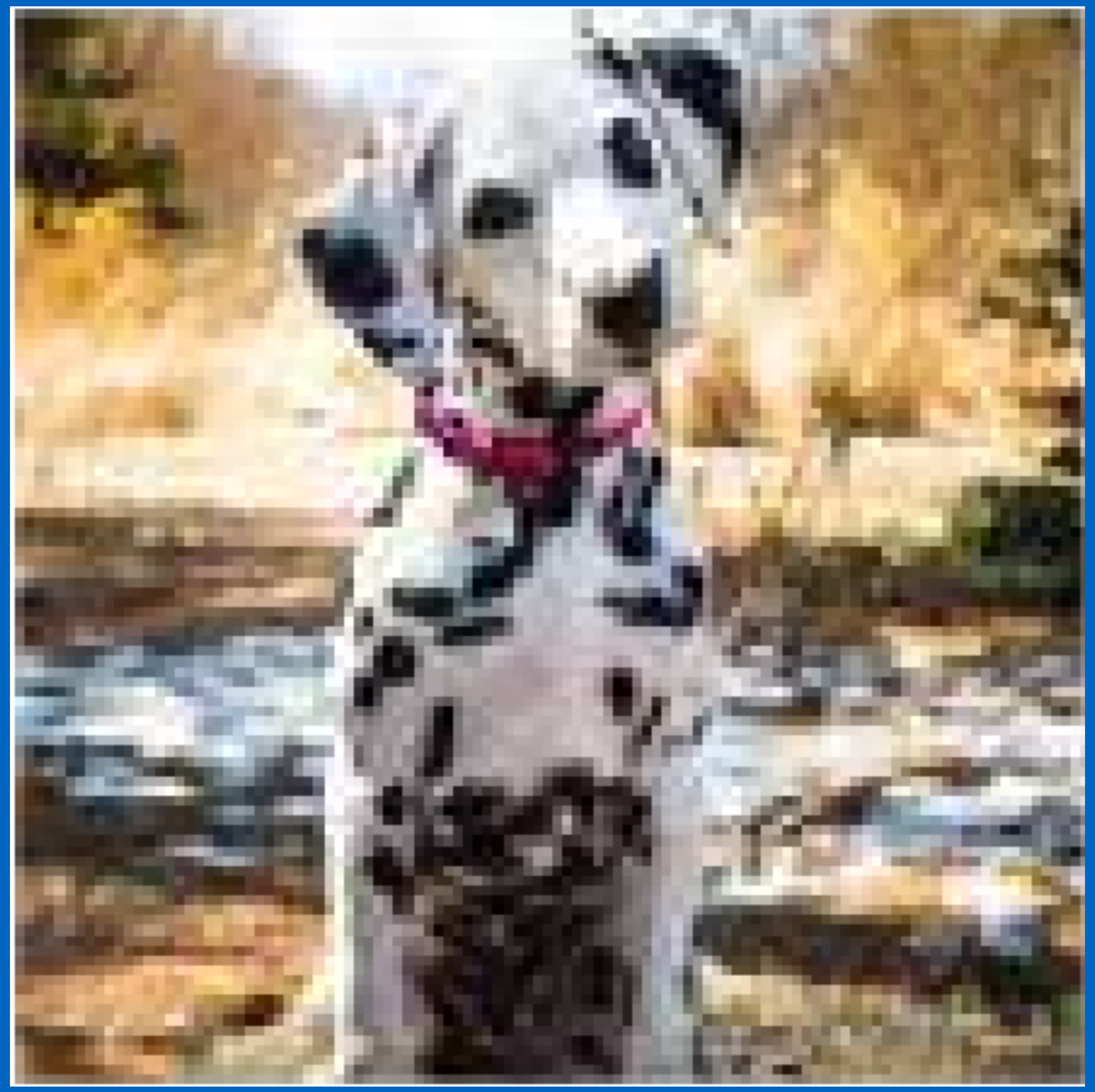


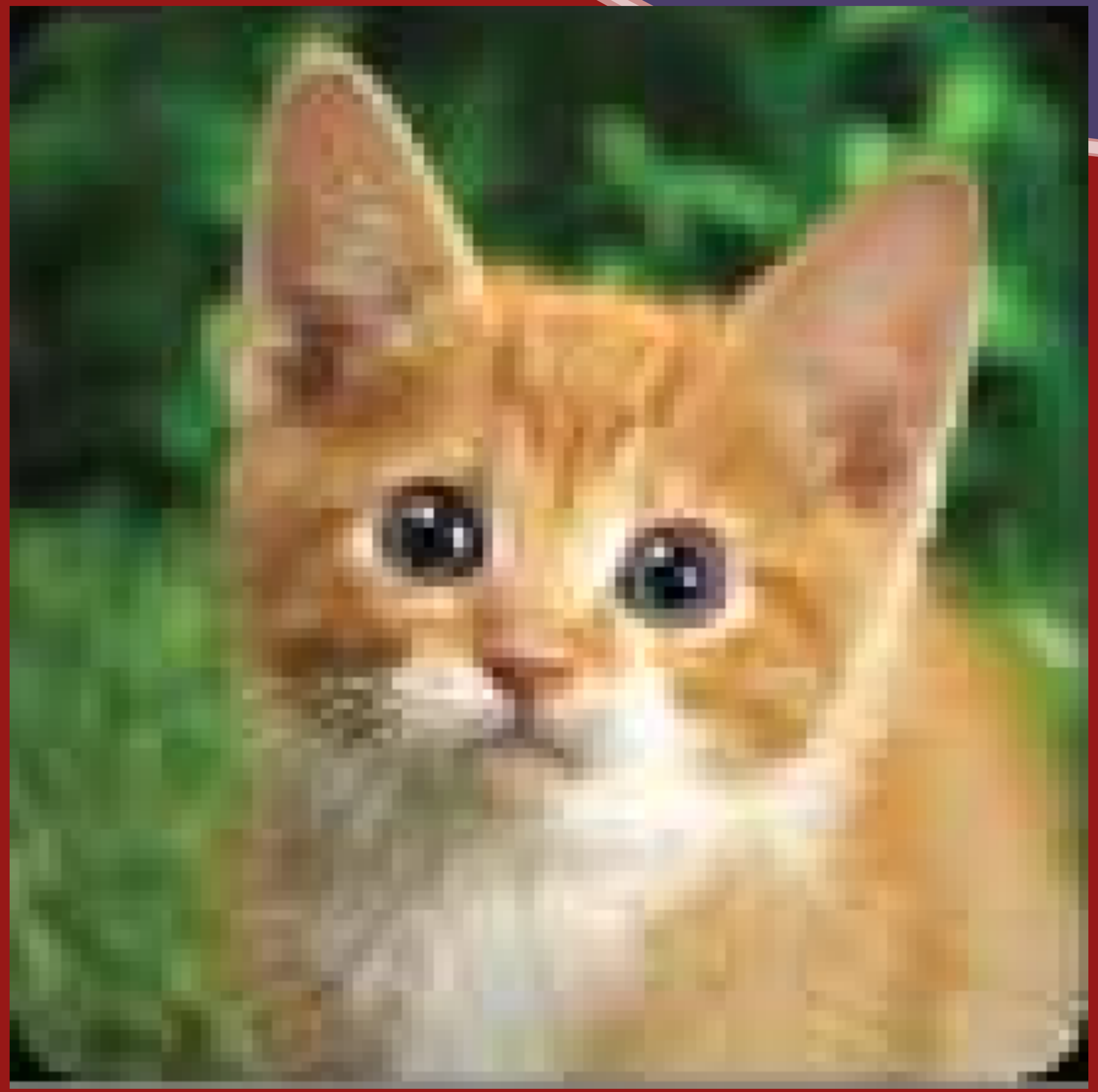


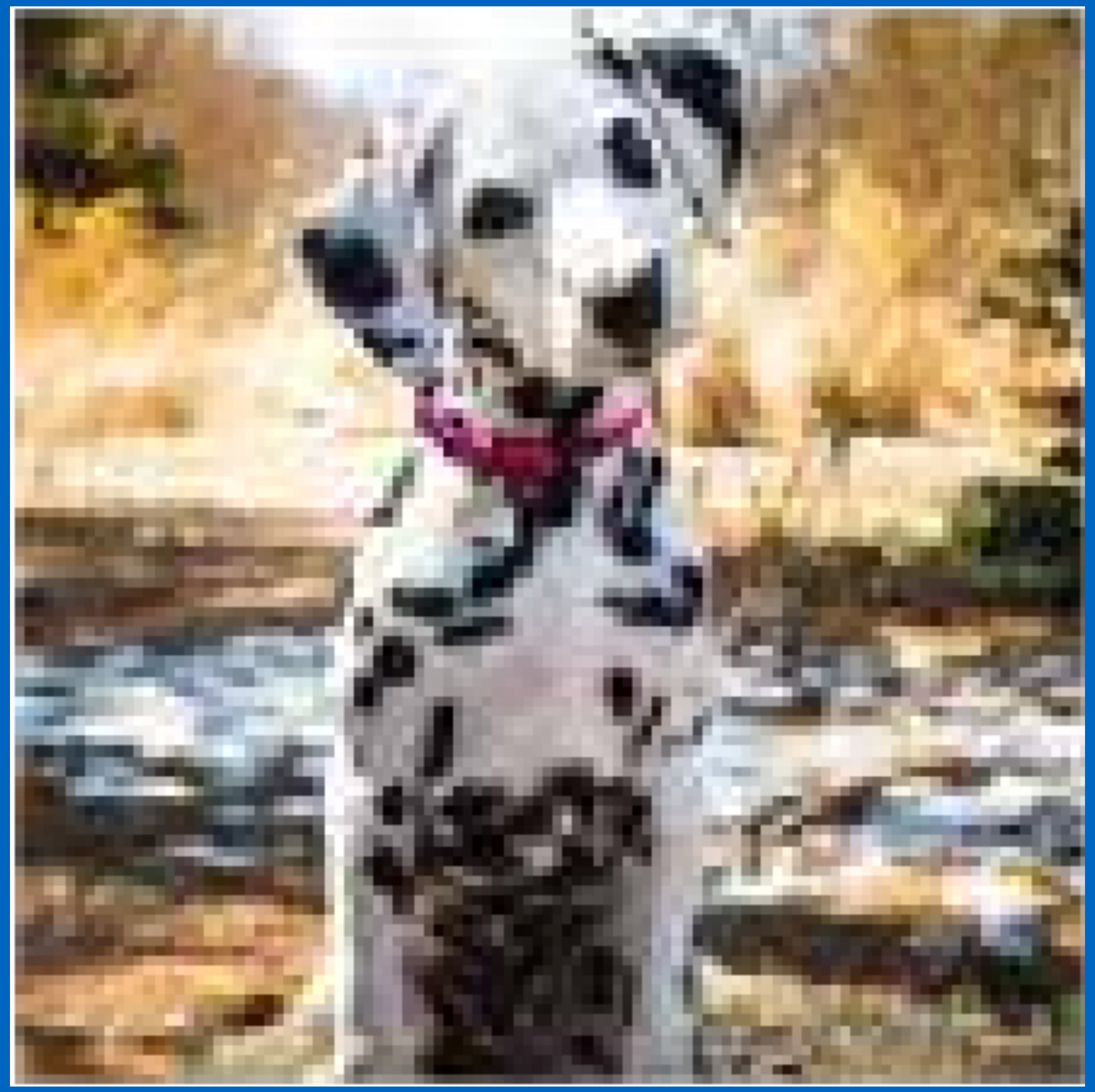
Sitatapatra: Blocking the Transfer of Adversarial Samples

Ilia Shumailov [*1] Xitong Gao [*2] Yiren Zhao [*1] Robert Mullins [1] Ross Anderson [1] Cheng-Zhong Xu [2]







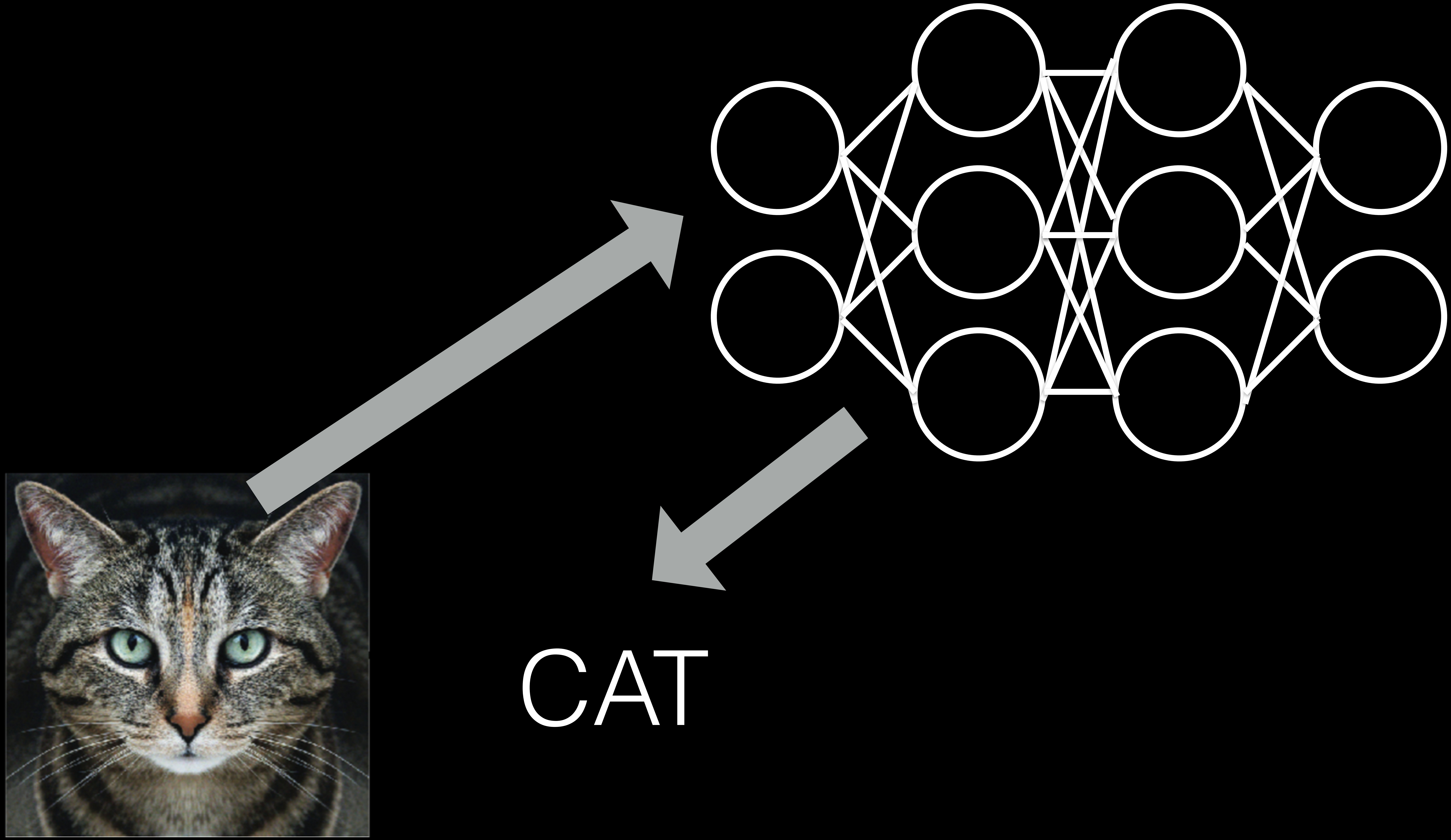


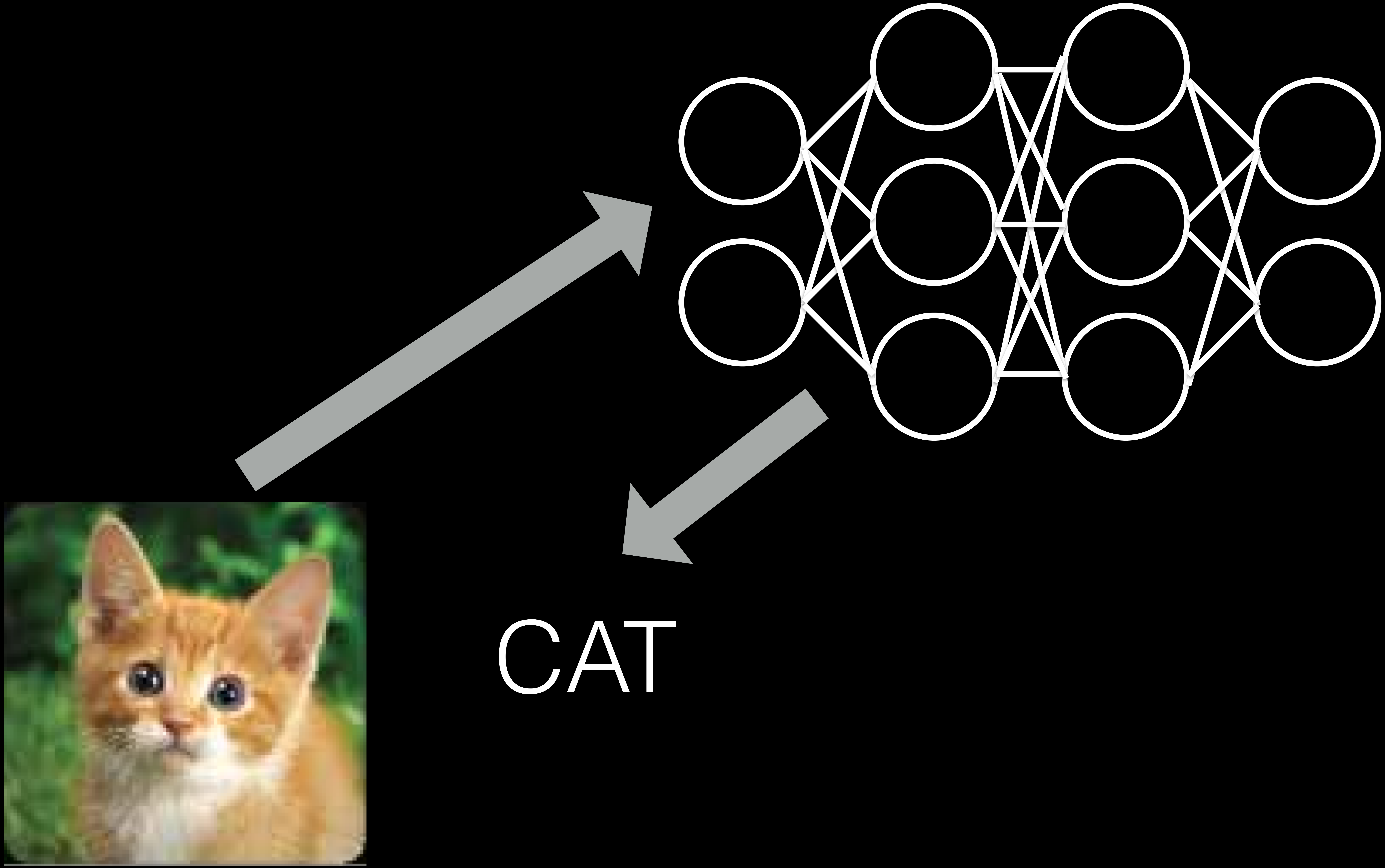
Stateful Detection of Black-Box Adversarial Attacks

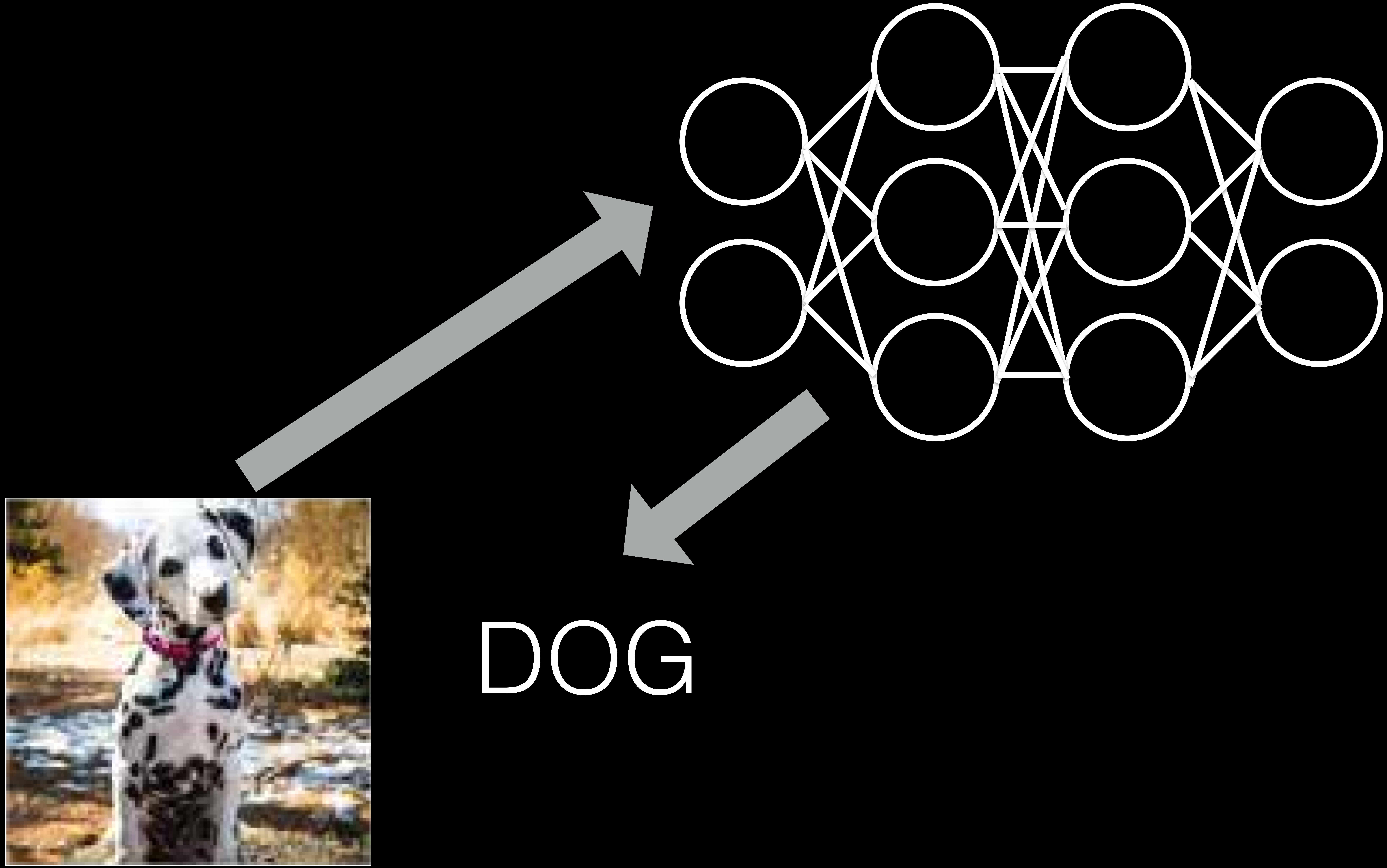
Steven Chen
University of California, Berkeley

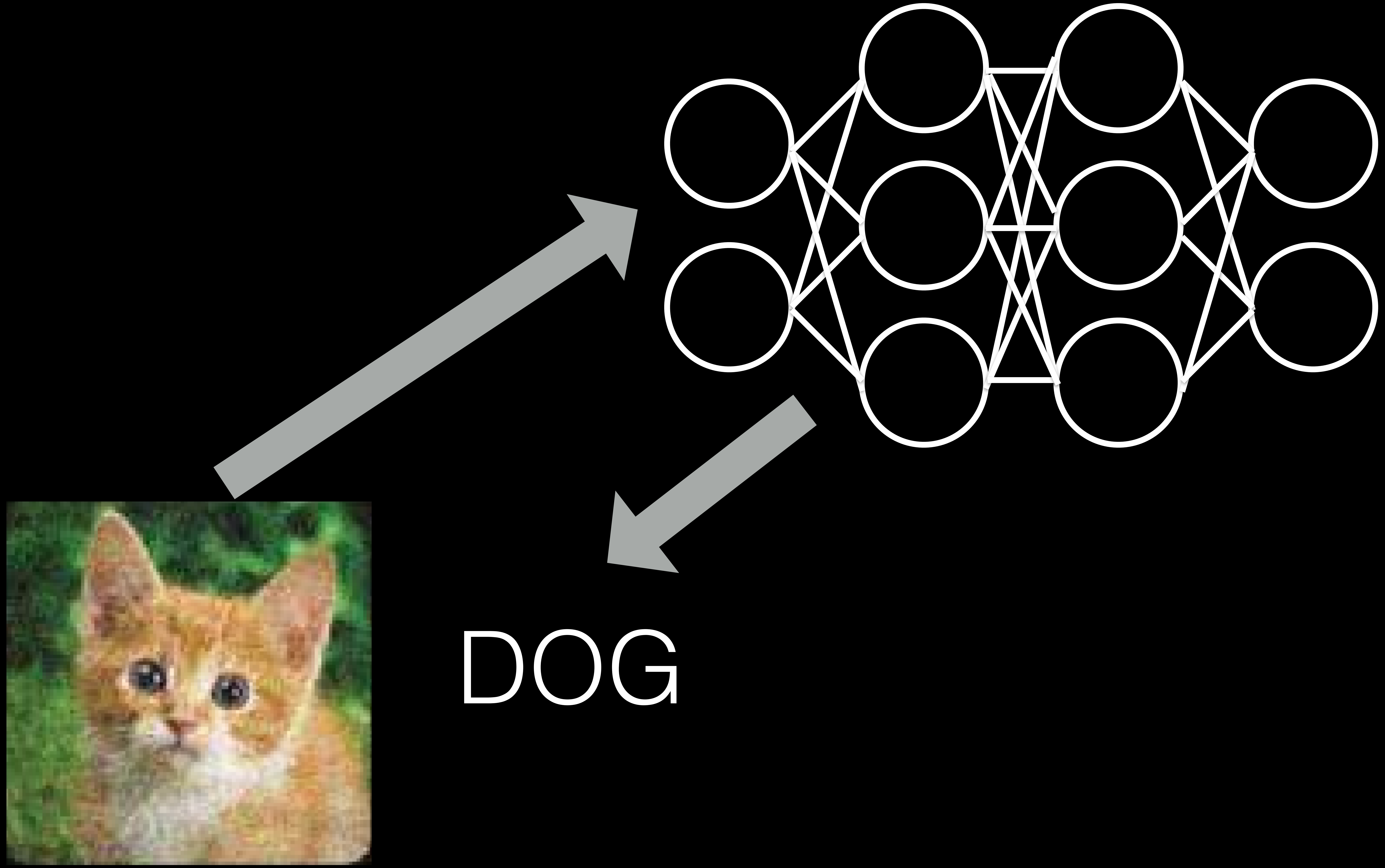
Nicholas Carlini
Google Research

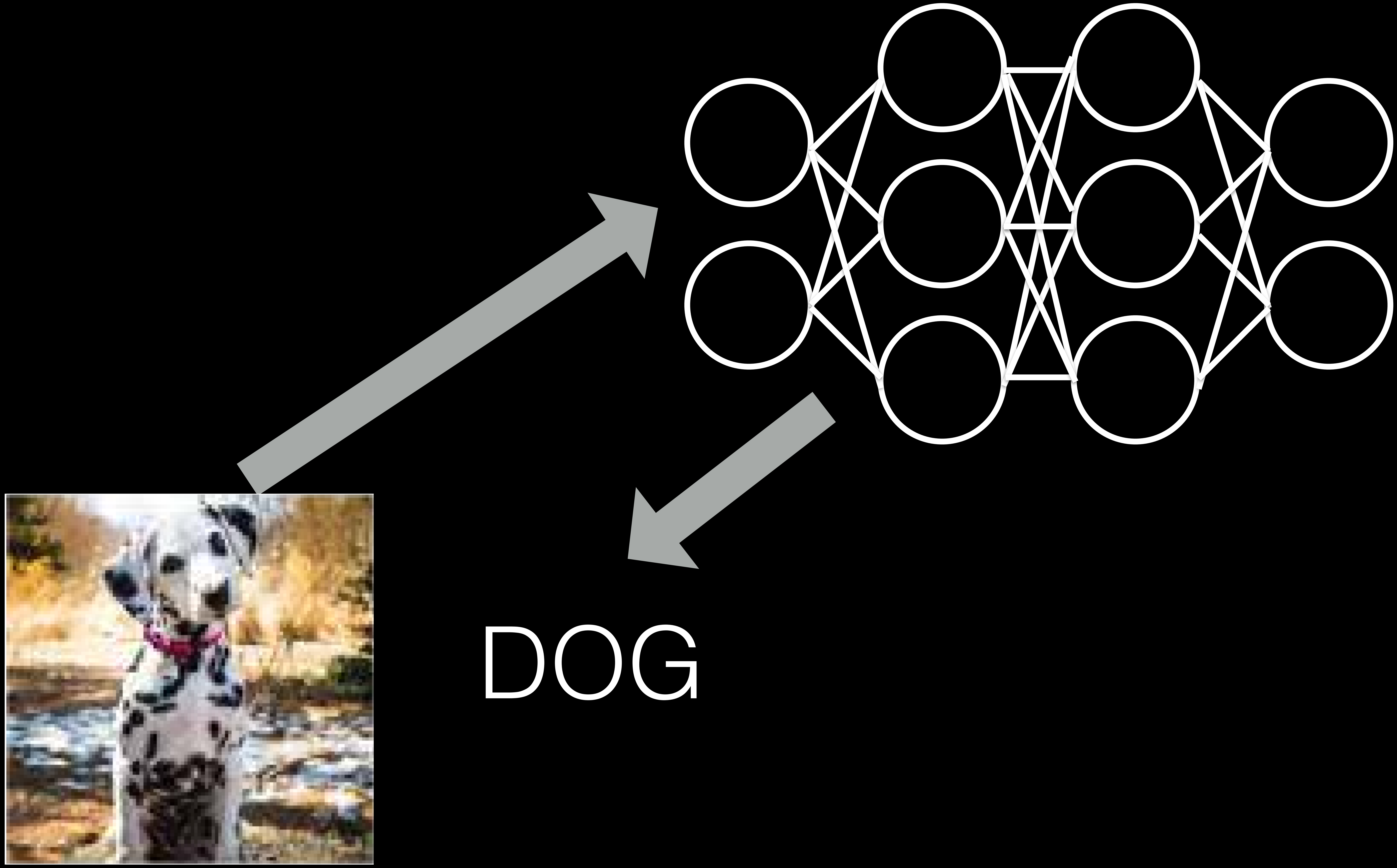
David Wagner
University of California, Berkeley

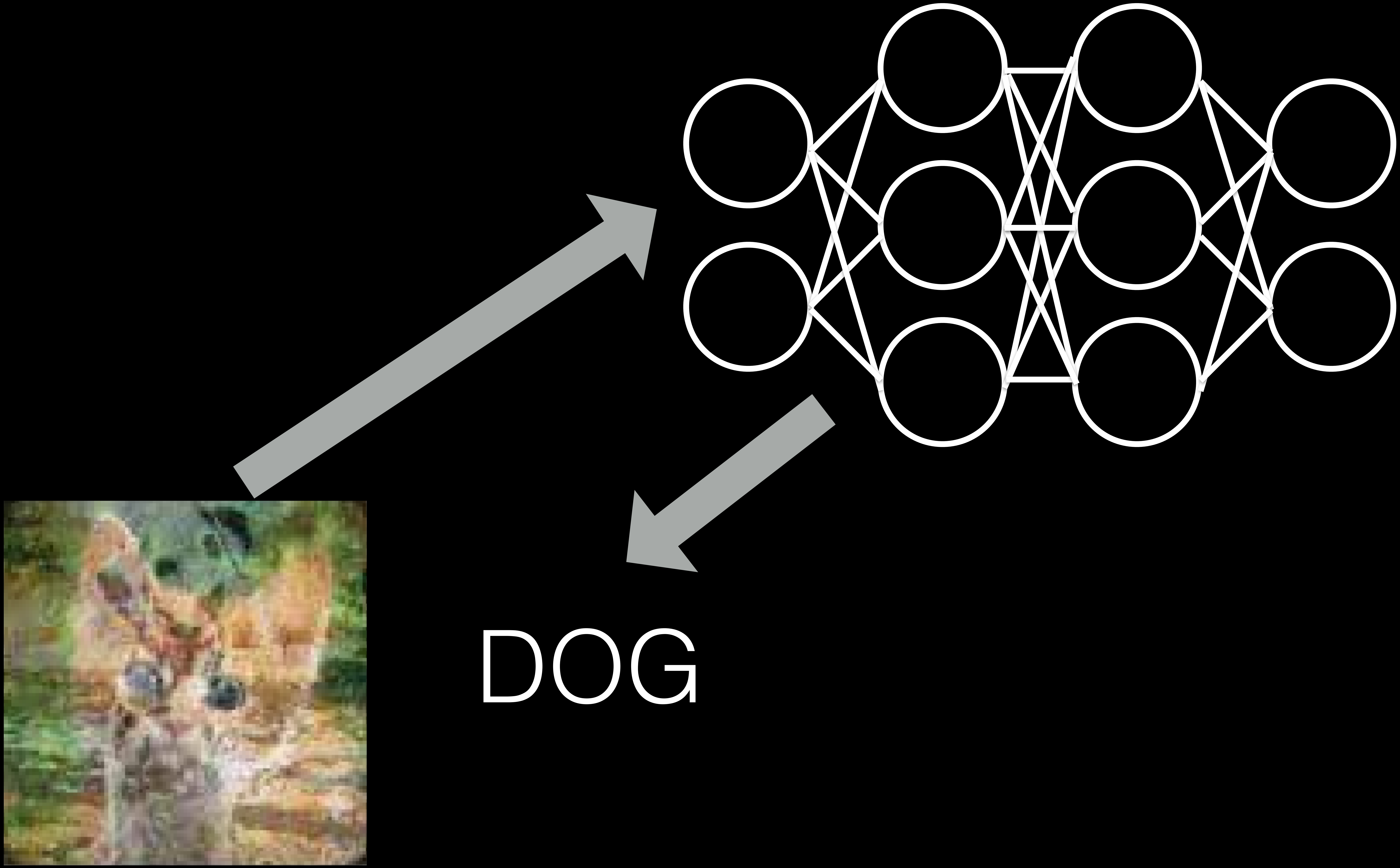

CAT

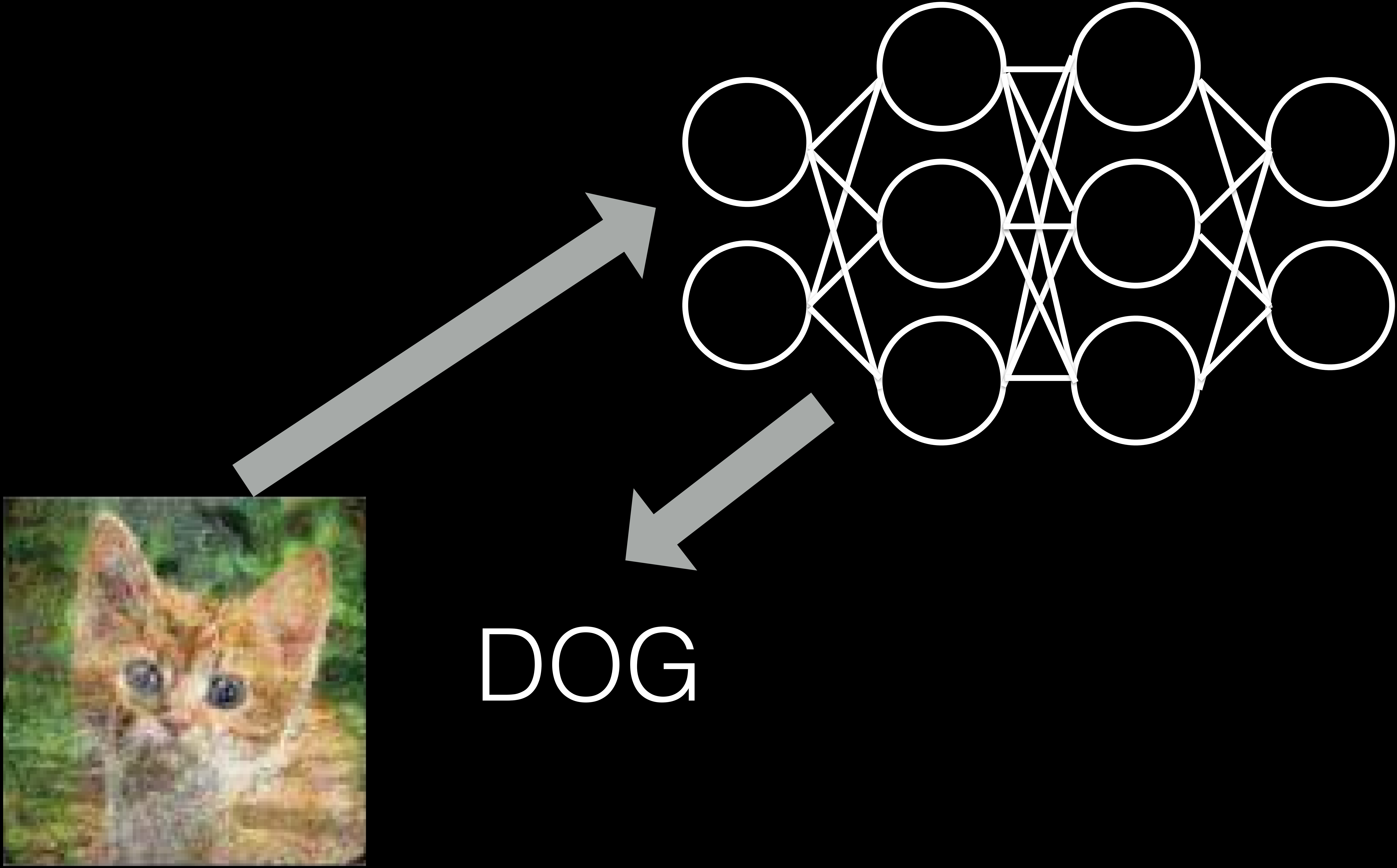

CAT

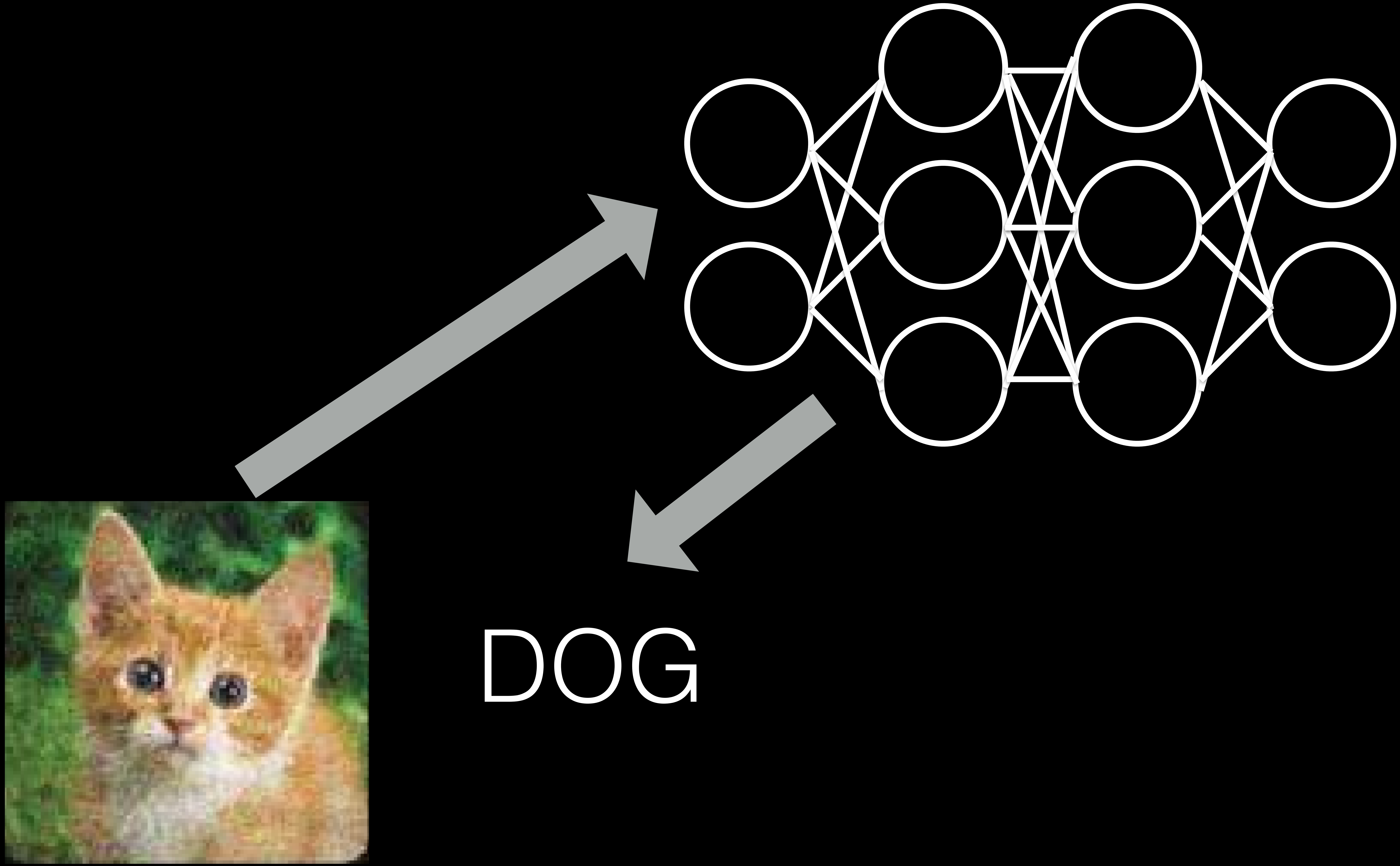

DOG

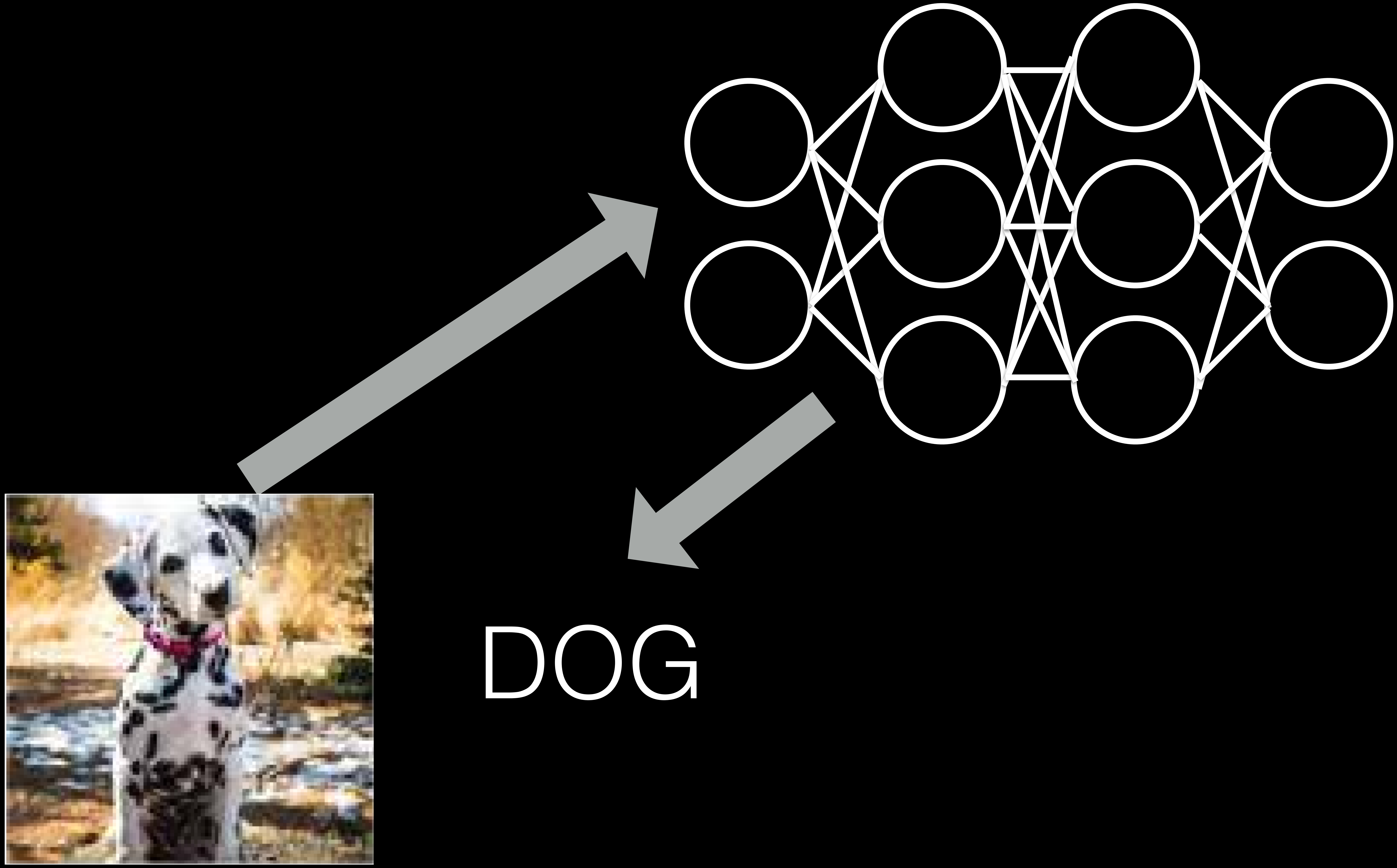

DOG

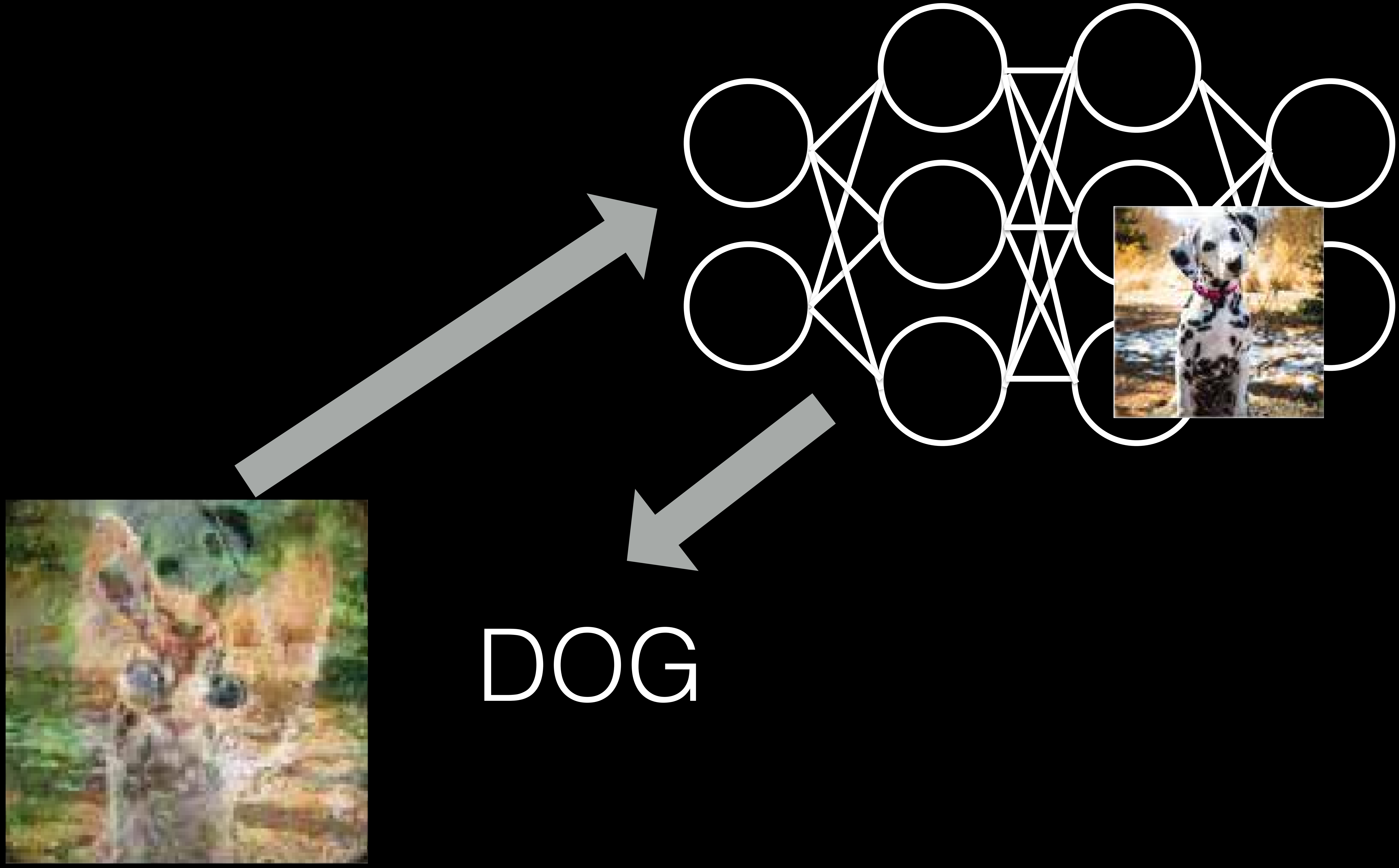

DOG

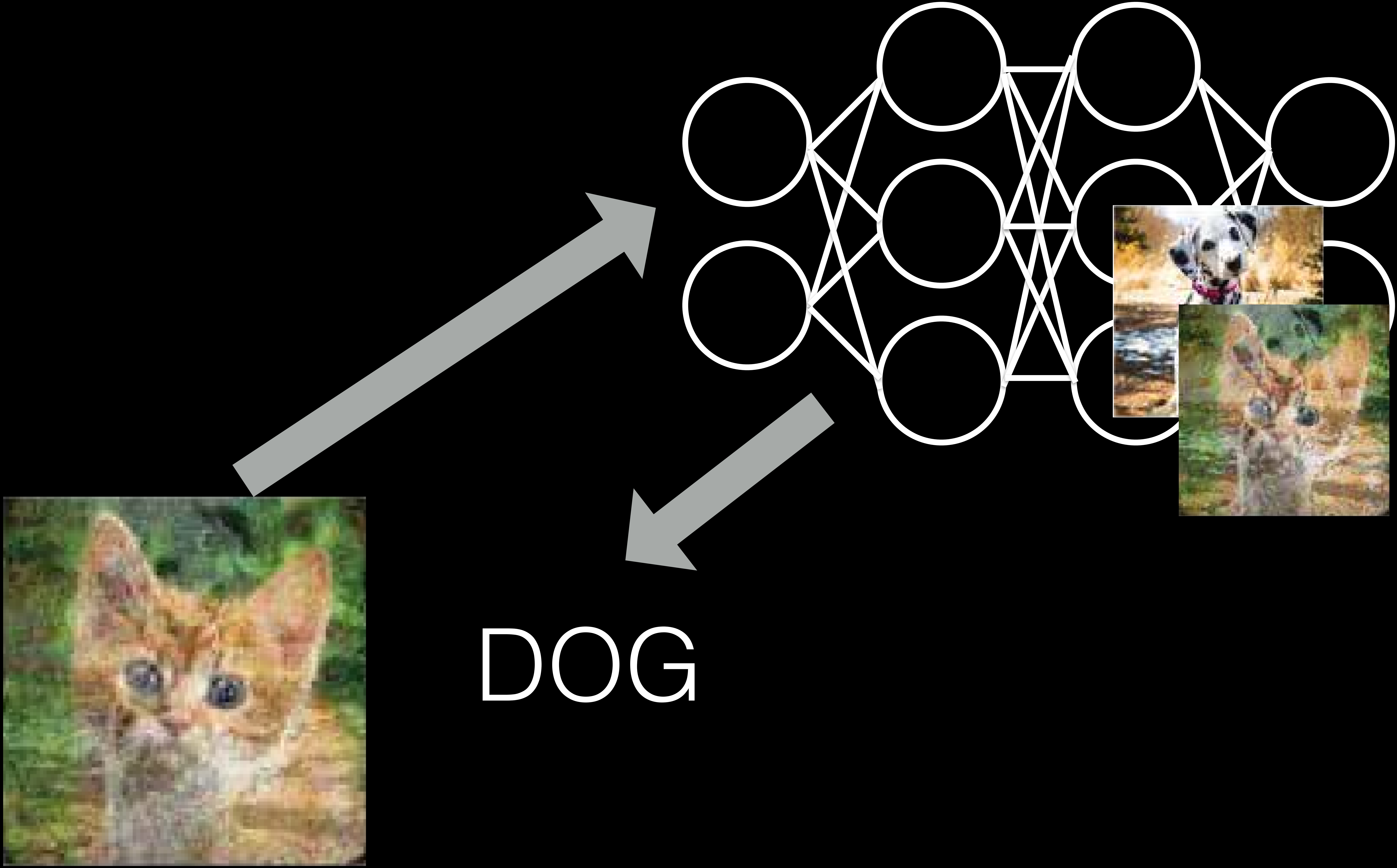

DOG

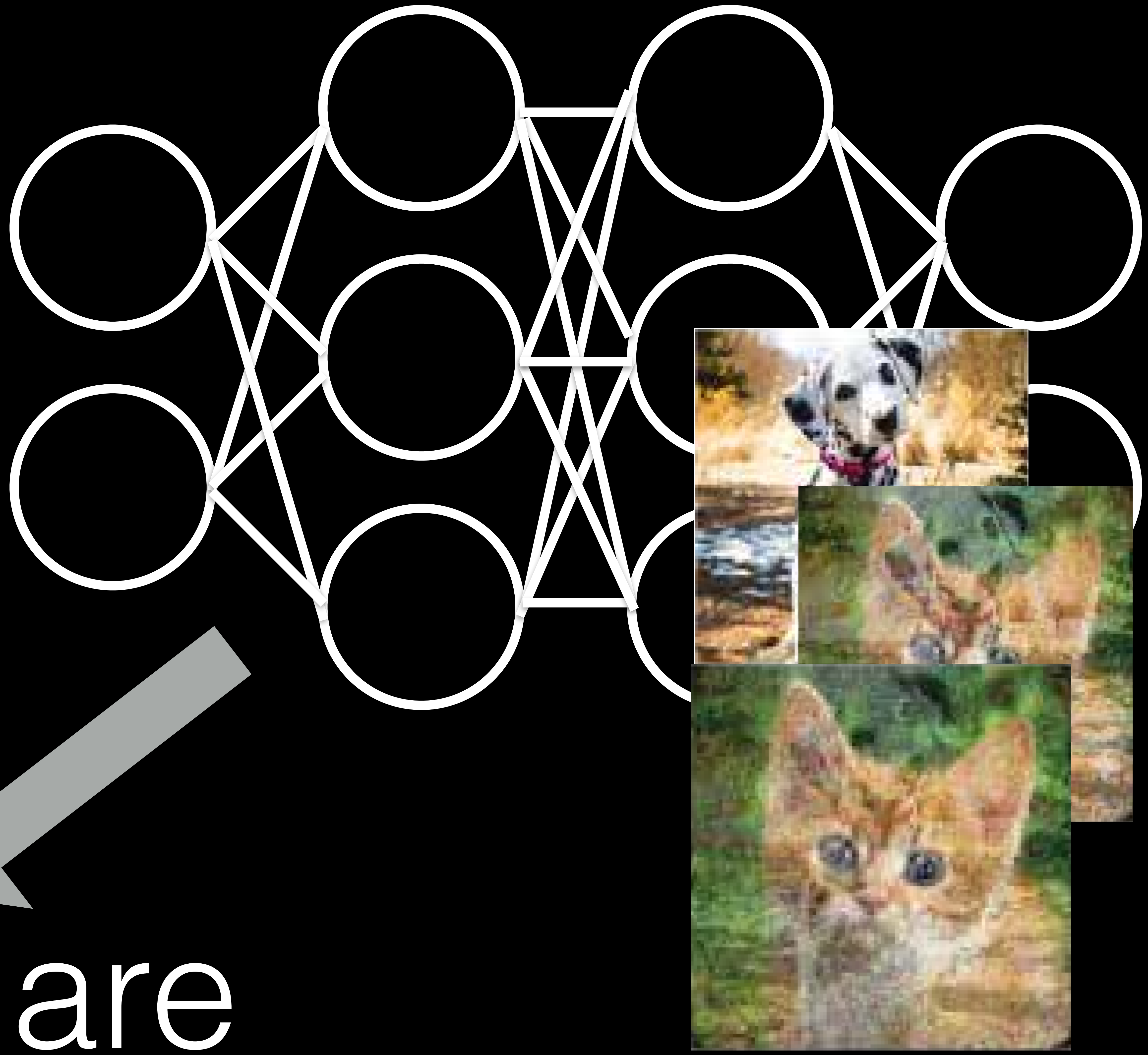

DOG


DOG


DOG


DOG


DOG



You are
being evil

Defenses I *do*
believe will be effective

Adversarially Robust Generalization Requires More Data

Ludwig Schmidt
MIT

Shibani Santurkar
MIT

Dimitris Tsipras
MIT

Kunal Talwar
Google Brain

Aleksander Mądry
MIT

Adversarially Robust Generalization Just Requires More Unlabeled Data

Unlabeled Data Improves Adversarial Robustness

Yai
Stanfo
yairc@

Are Labels Required for Improving Adversarial Robustness?

Jonathan Uesato* **Jean-Baptiste Alayrac*** **Po-Sen Huang***

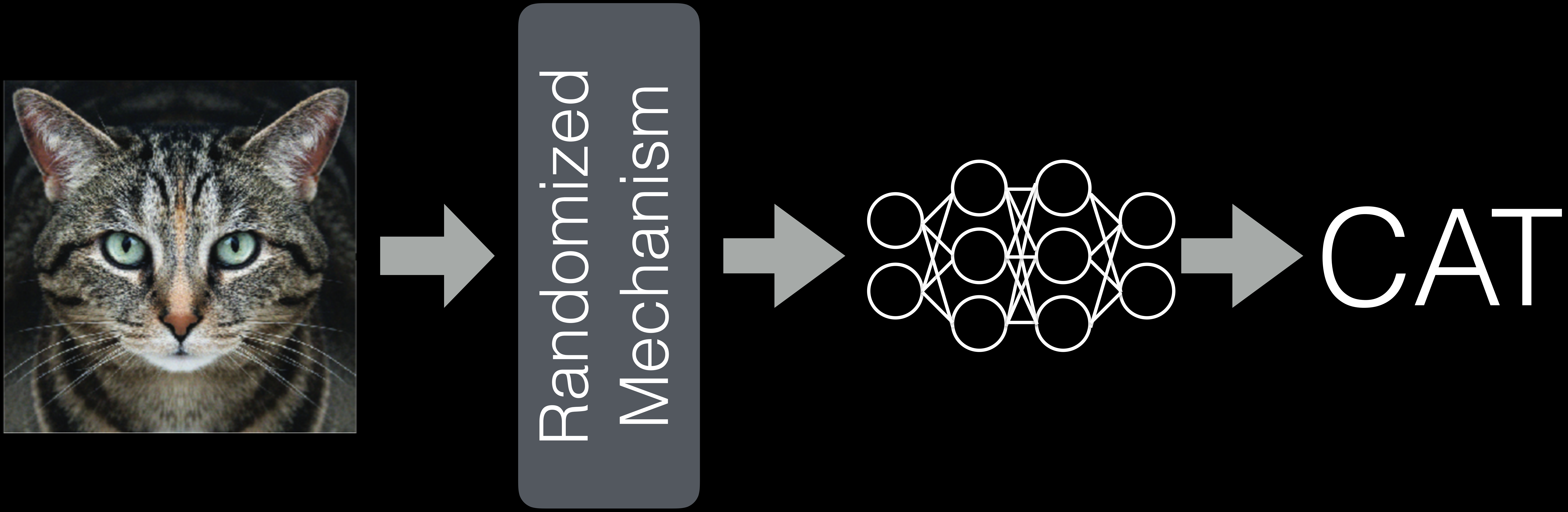
Robert Stanforth **Alhussein Fawzi** **Pushmeet Kohli**

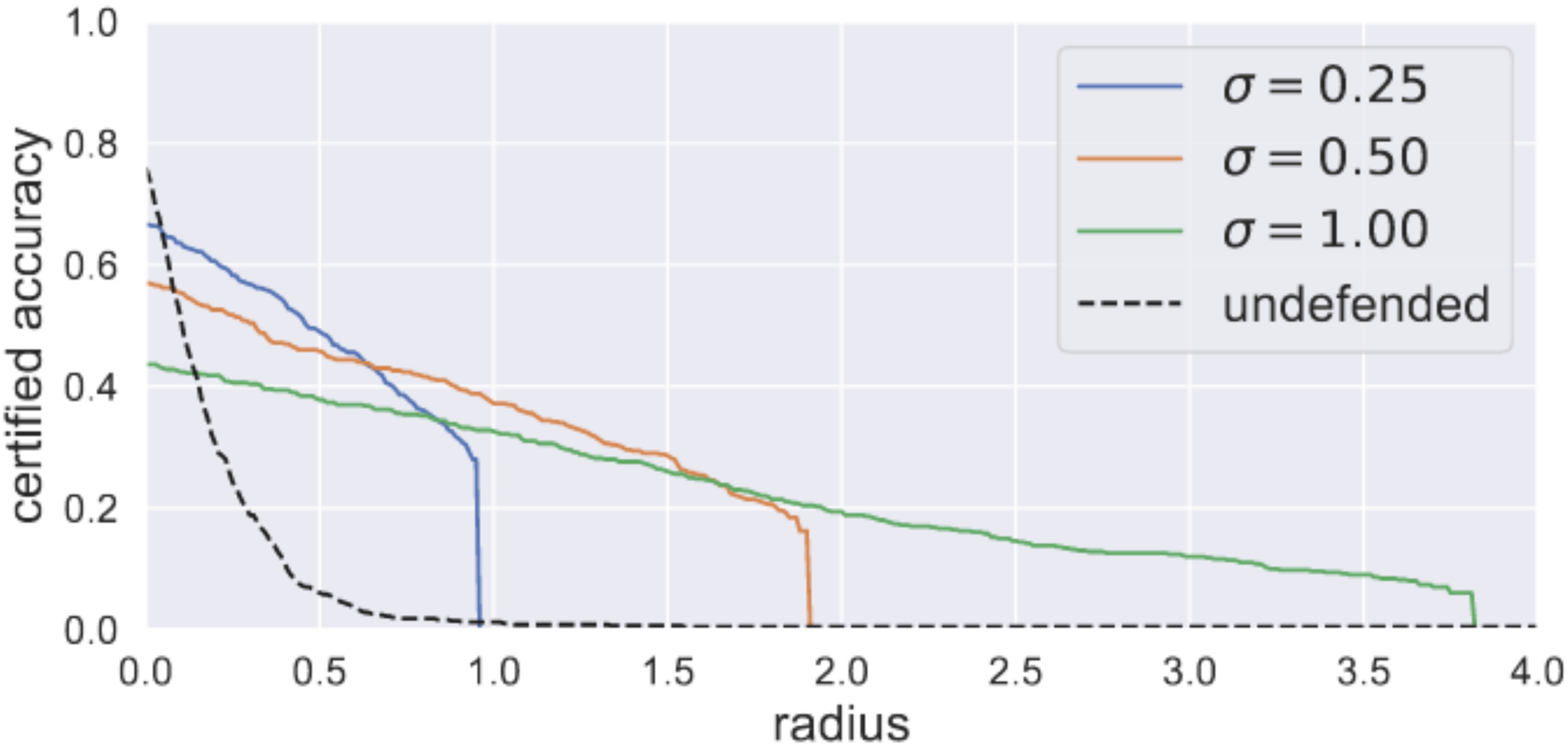
Certified Robustness to Adversarial Examples with Differential Privacy

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana
Columbia University

Certified Adversarial Robustness via Randomized Smoothing

Jeremy Cohen [1] Elan Rosenfeld [1] J. Zico Kolter [1 2]


Randomized Mechanism
CAT

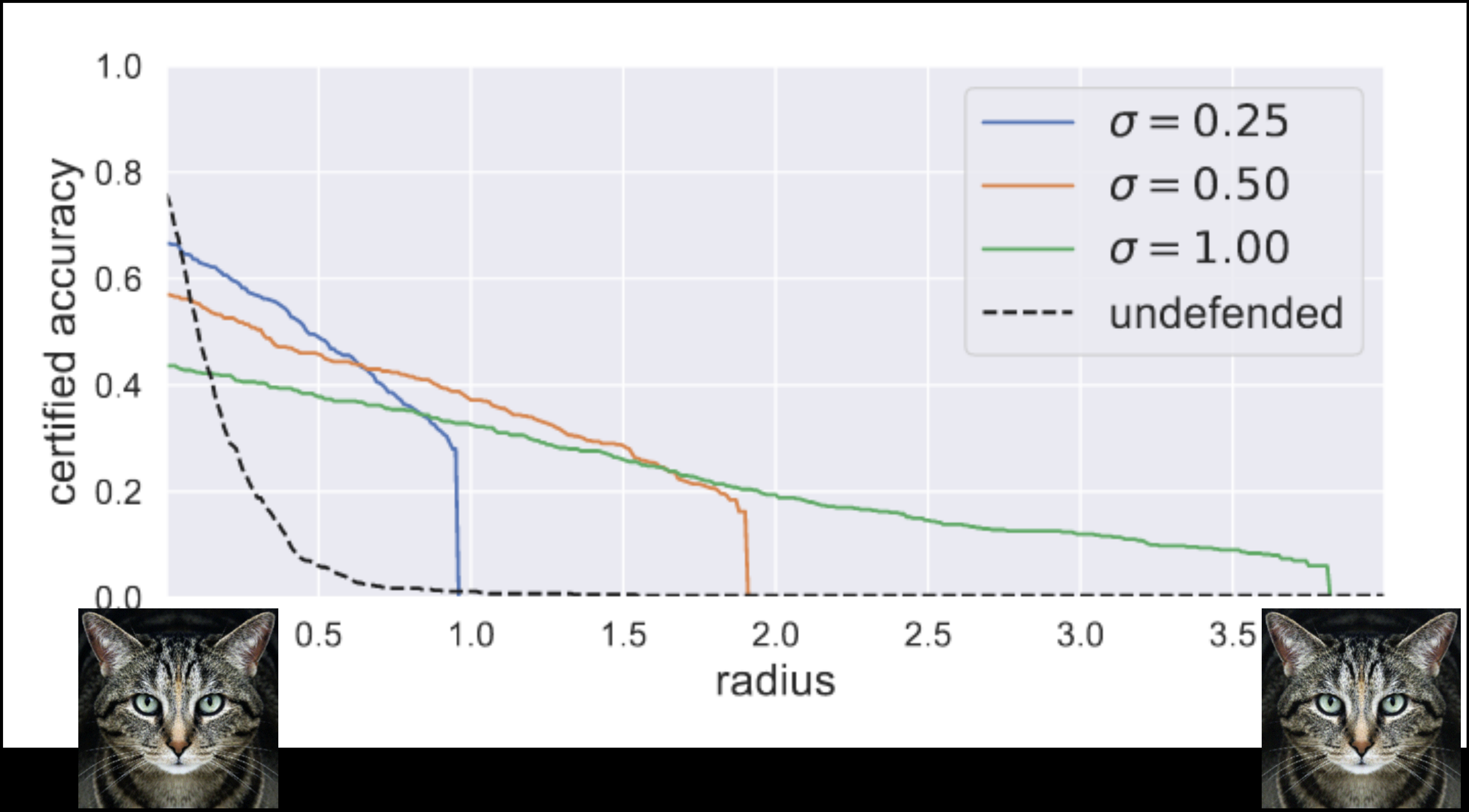

1.0
0.8
0.6
0.4
0.2
0.0
certified accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
radius
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended

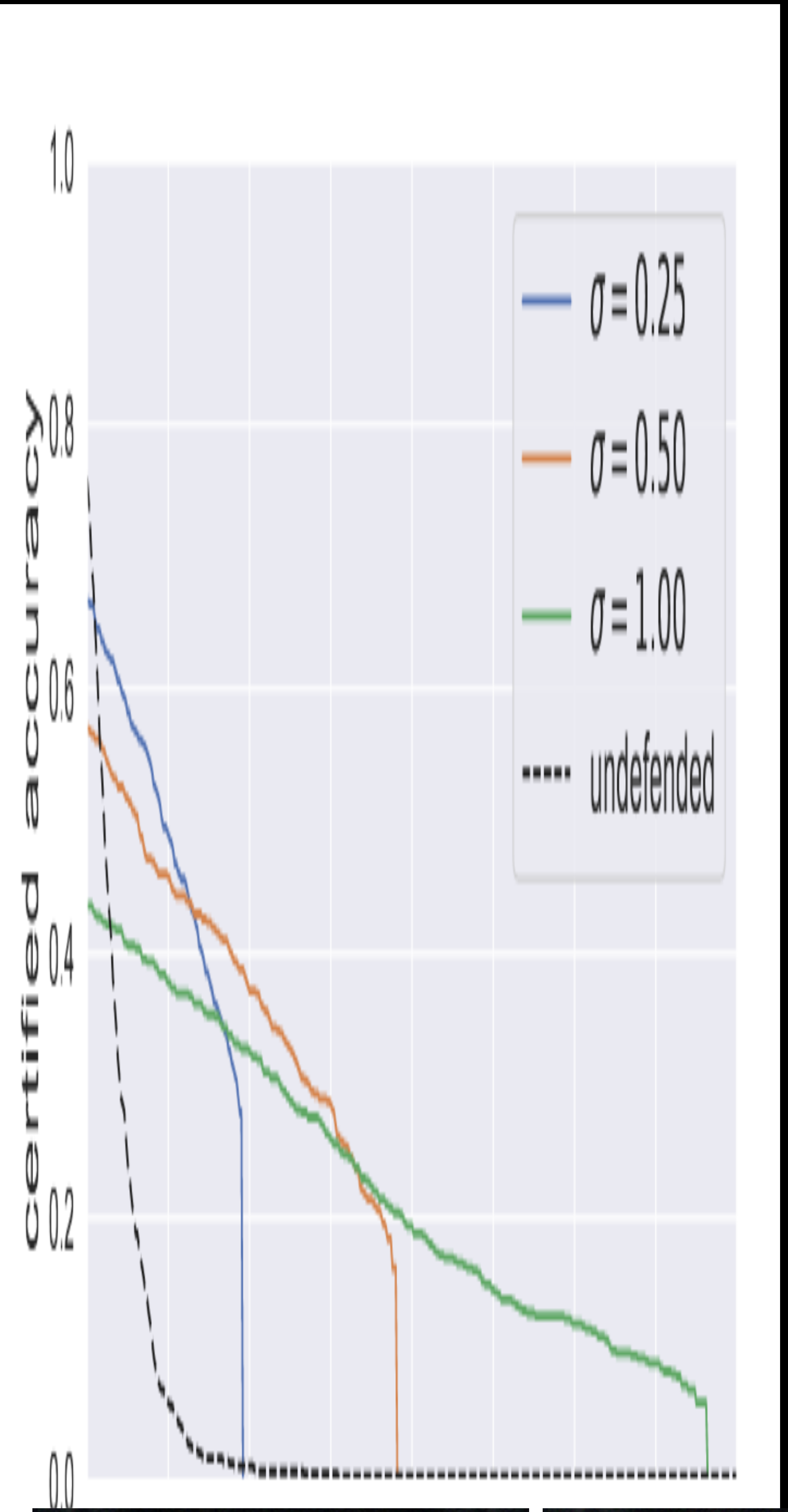
Original

L_2 distortion: 4

Original

L_2 distortion: 10


certified accuracy
radius
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended


certified accuracy
1.0
0.8
0.6
0.4
0.2
0.0
$\sigma = 0.25$
$\sigma = 0.50$
$\sigma = 1.00$
undefended

$L_2 = 75$

Original

L_2 distortion: 75

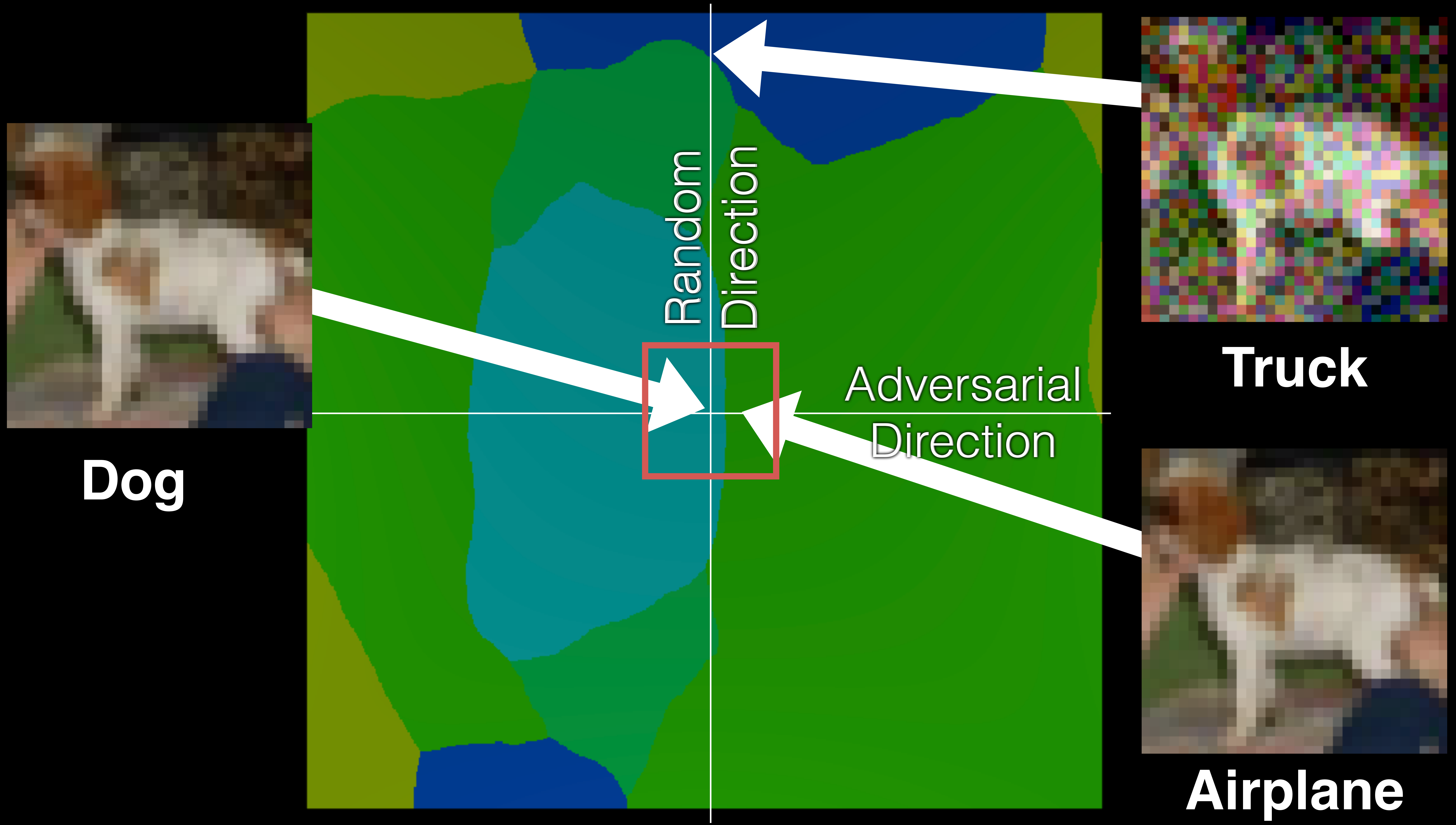
L_2 distortion: 75

Recent advances in …

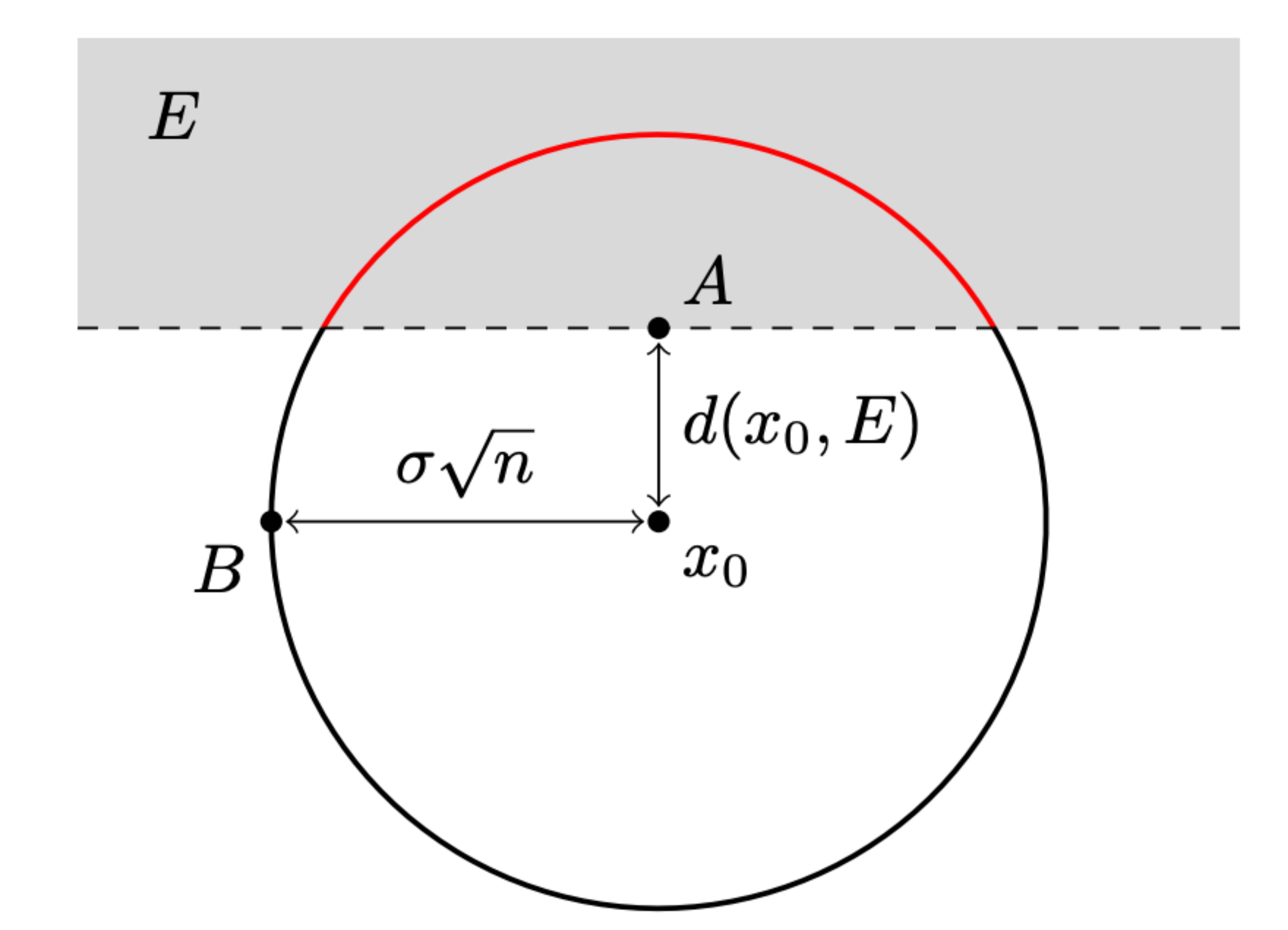
Why Adversarial Examples Exist

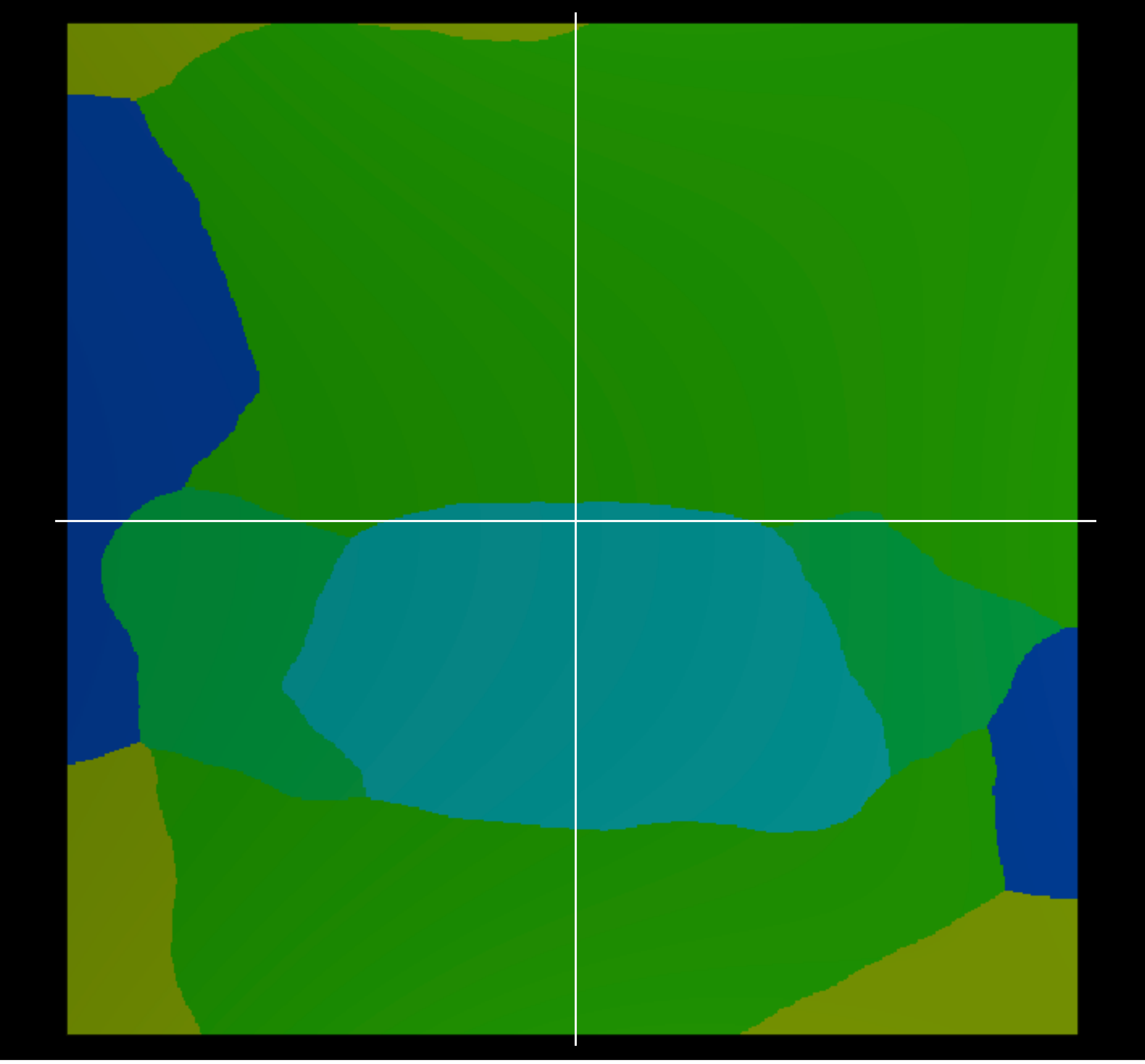
Adversarial Examples Are a Natural Consequence of Test Error in Noise

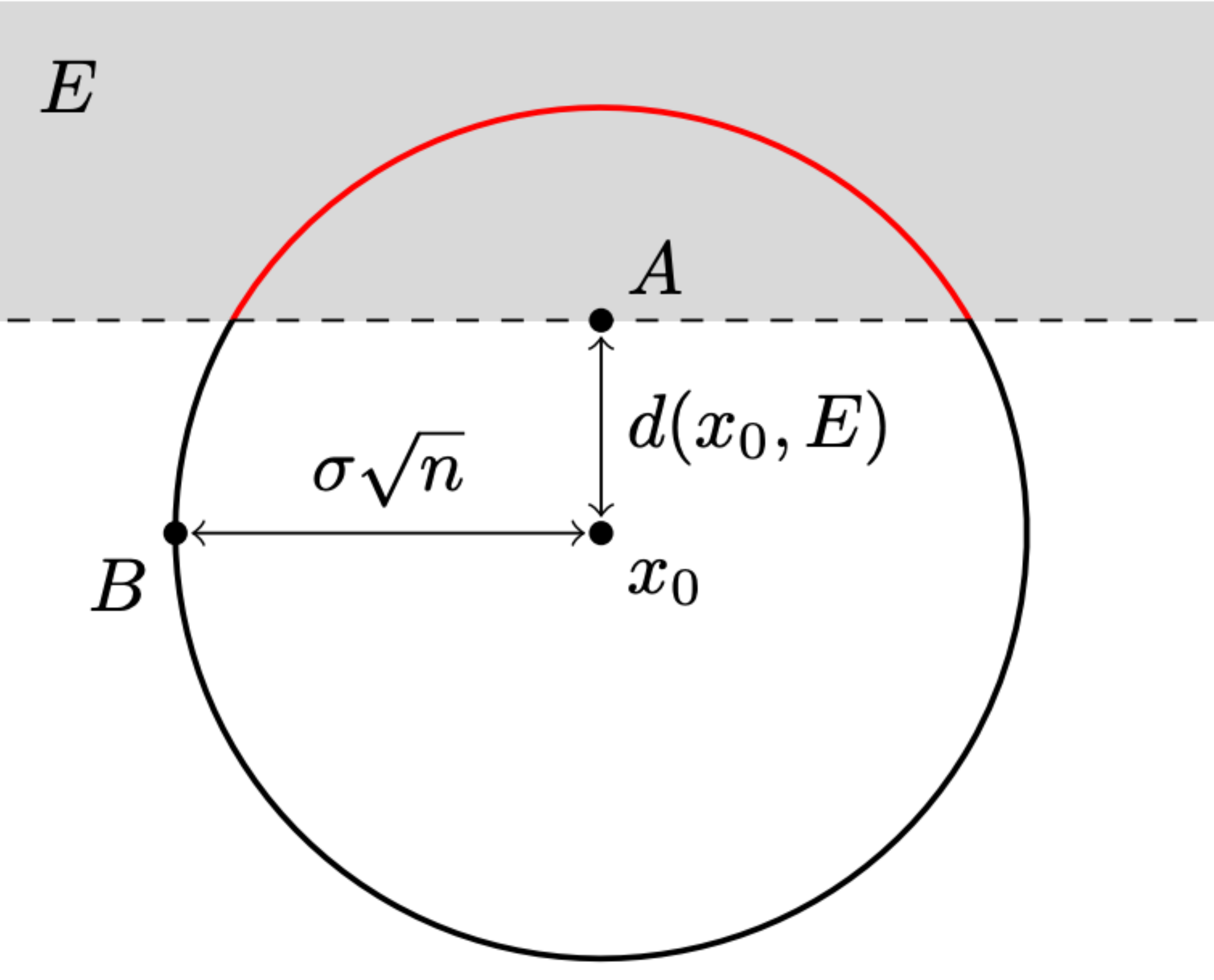
Nicolas Ford [*,1,2] Justin Gilmer [*,1] Nicholas Carlini [1] Ekin D. Cubuk [1]

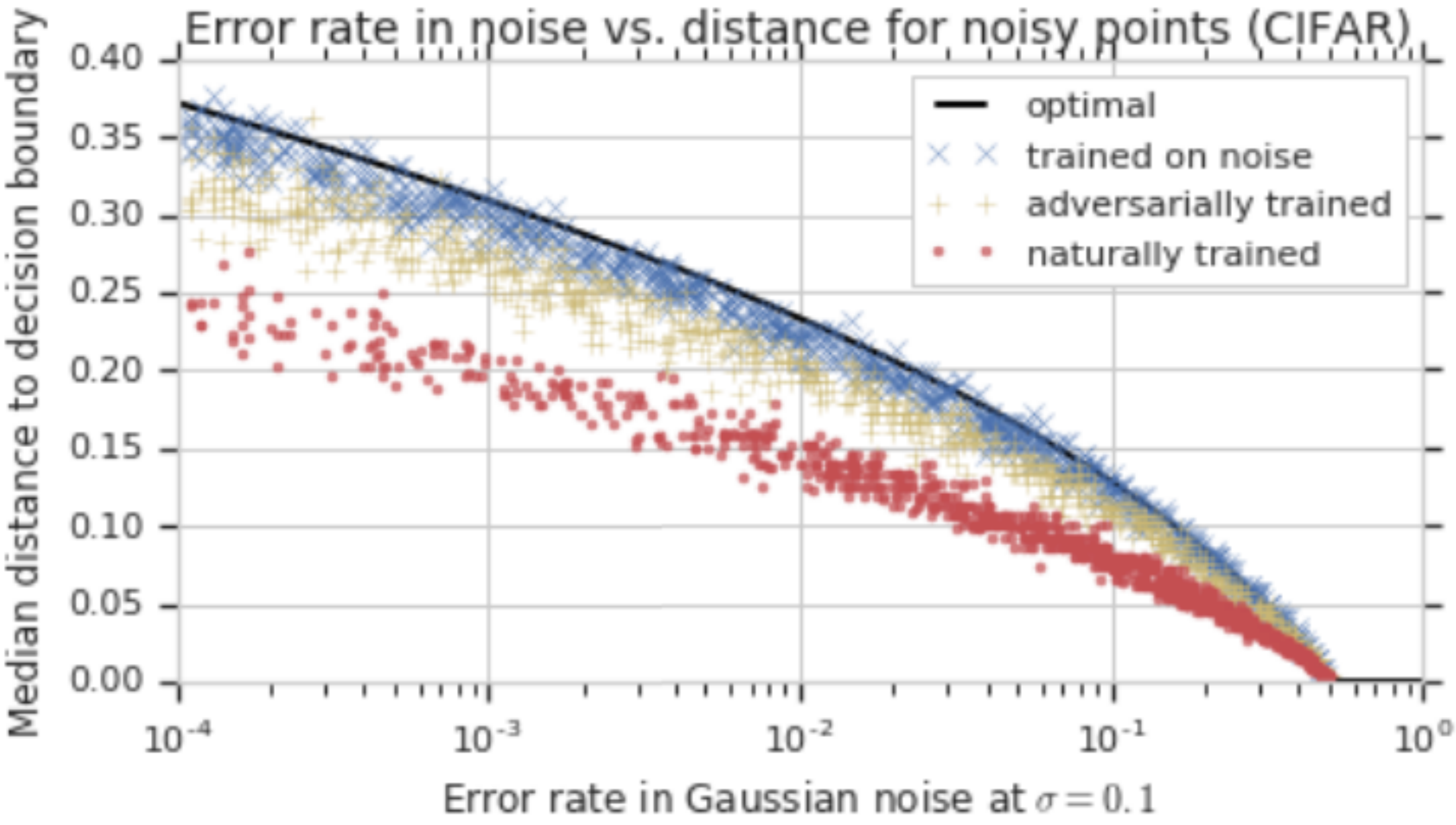


Random Direction
Adversarial Direction
Dog
Truck
Airplane


E
A
$d(x_0, E)$
$\sigma\sqrt{n}$
B
x_0




E
A
$d(x_0, E)$
$\sigma\sqrt{n}$
B
x_0



Error rate in noise vs. distance for noisy points (CIFAR)
Median distance to decision boundary
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
Error rate in Gaussian noise at $\sigma = 0.1$
optimal
trained on noise
adversarially trained
naturally trained

Adversarial Examples Are Not Bugs, They Are Features

Andrew Ilyas*
MIT
ailyas@mit.edu

Shibani Santurkar*
MIT
shibani@mit.edu

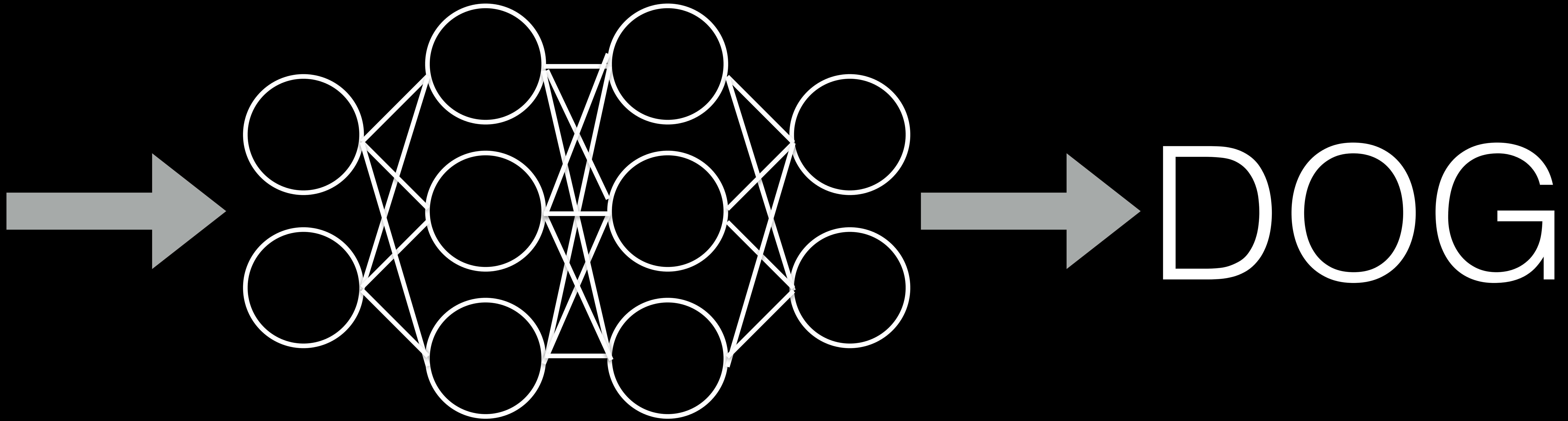
Dimitris Tsipras*
MIT
tsipras@mit.edu

Logan Engstrom*
MIT
engstrom@mit.edu

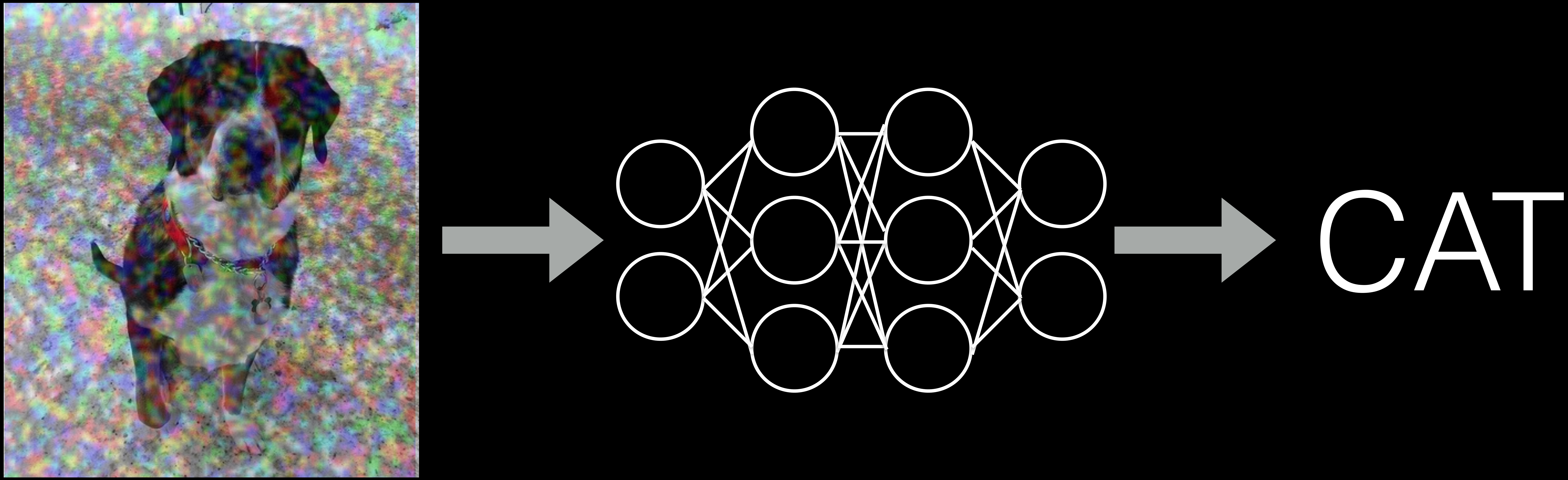
Brandon Tran
MIT
btran115@mit.edu

Aleksander Mądry
MIT
madry@mit.edu

Standard Training Dataset

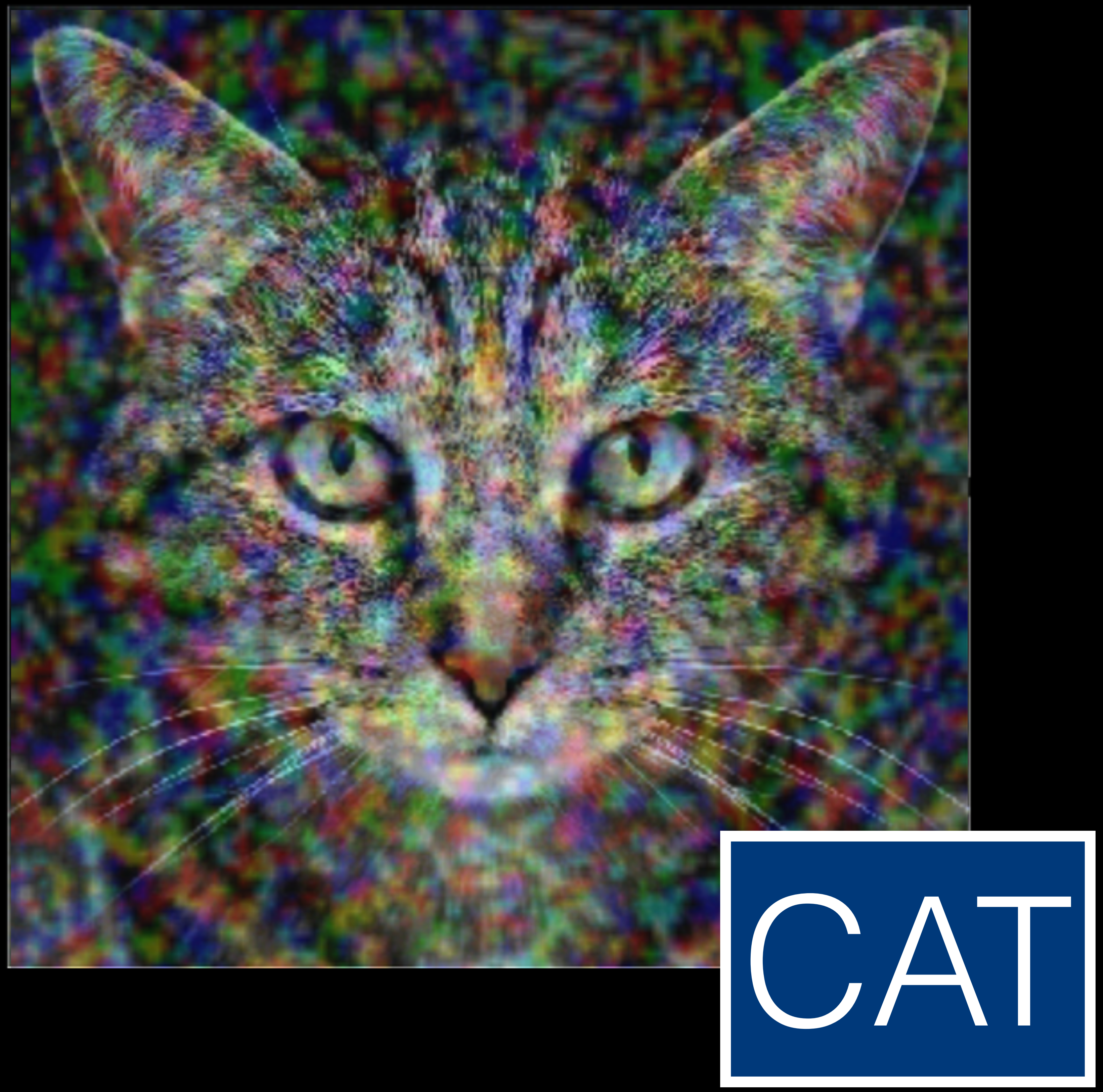


Standard Testing Setup

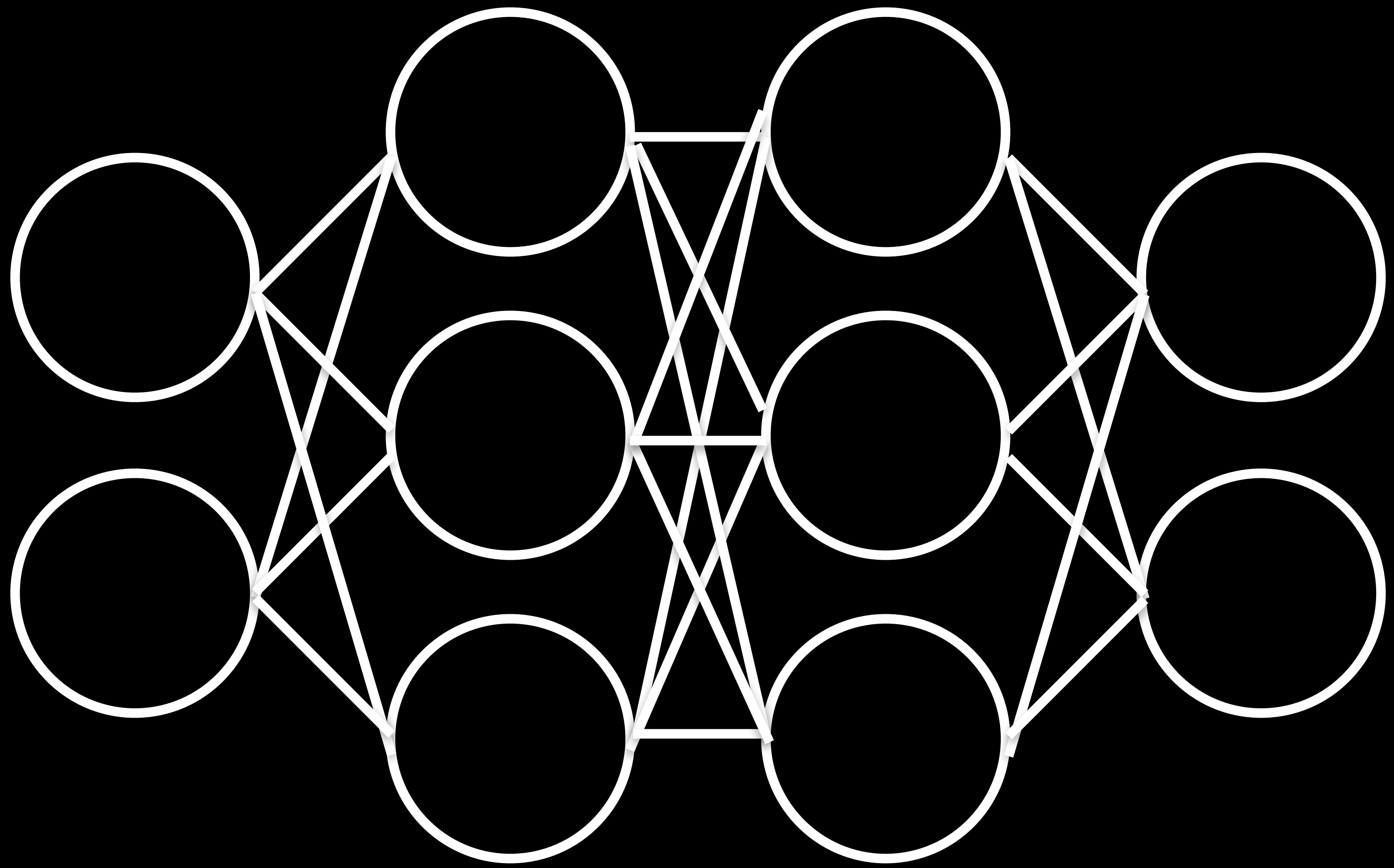


Adversarial Testing Setup

Standard Training Dataset

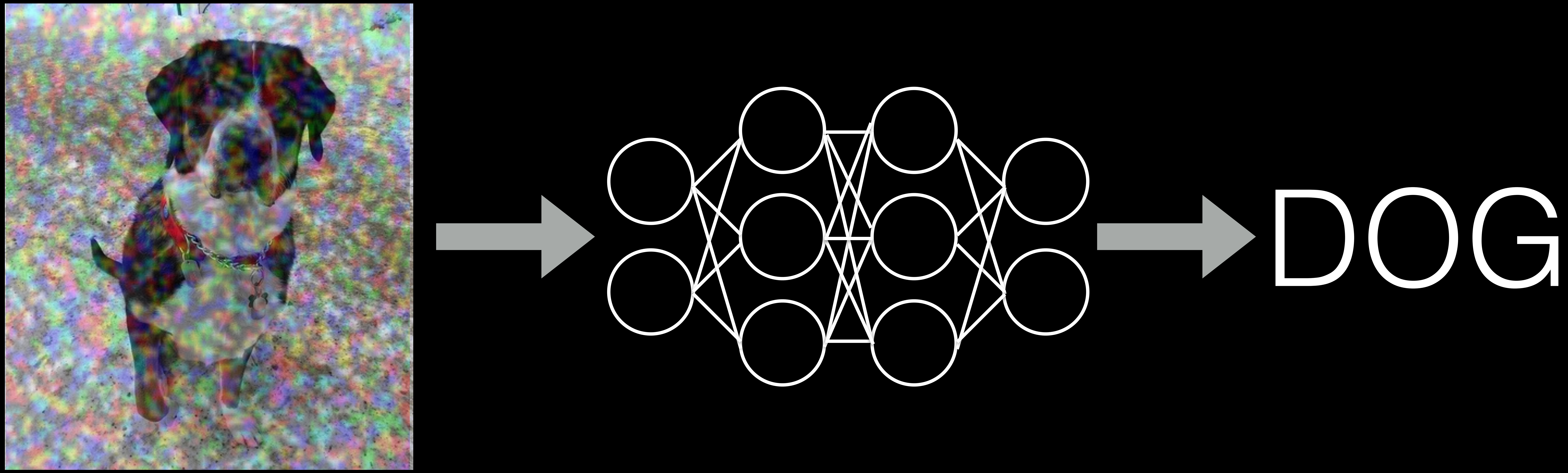


Adversarial Training Dataset



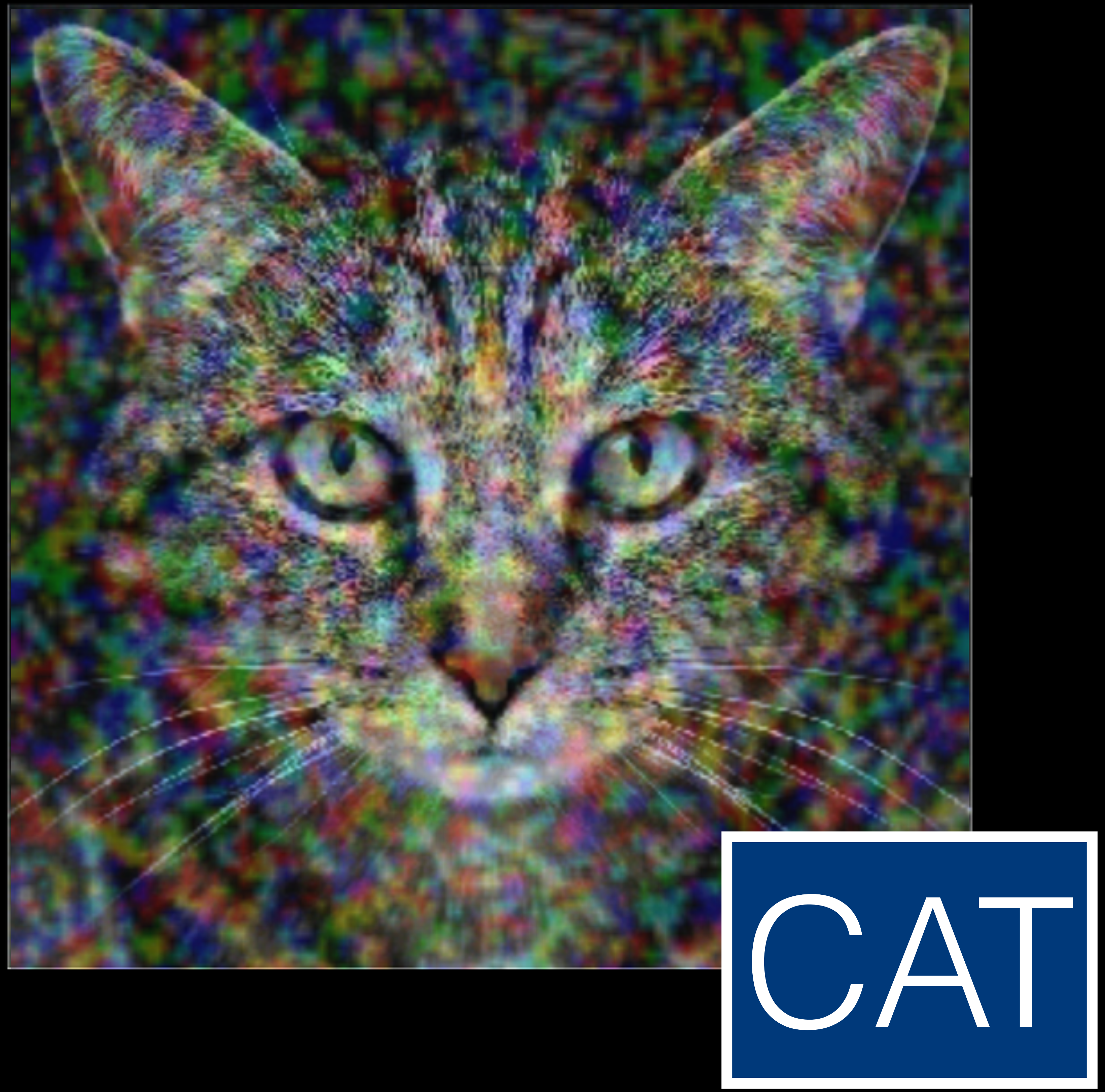
DOG

Standard Testing Setup

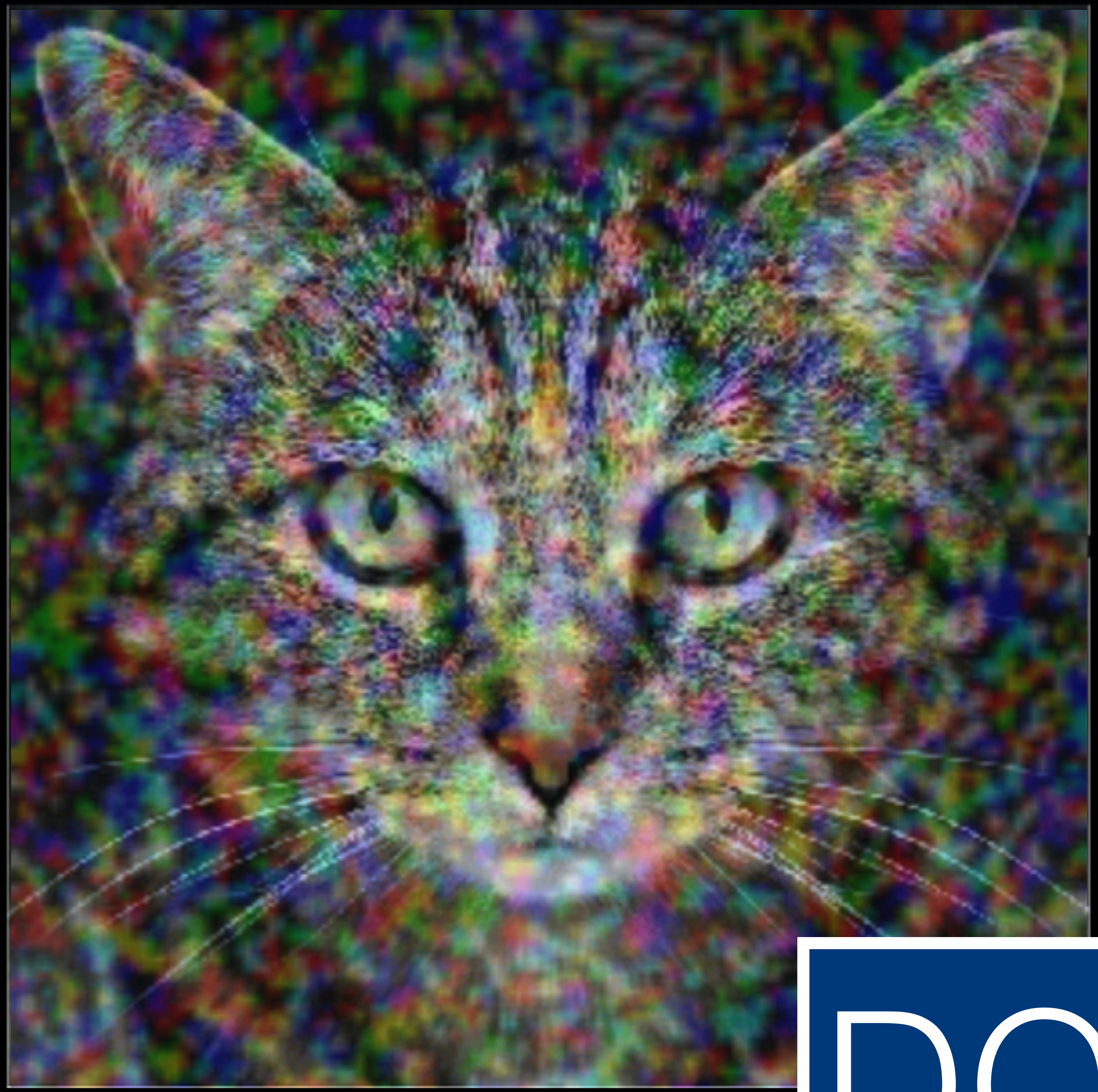


Adversarial Testing Setup

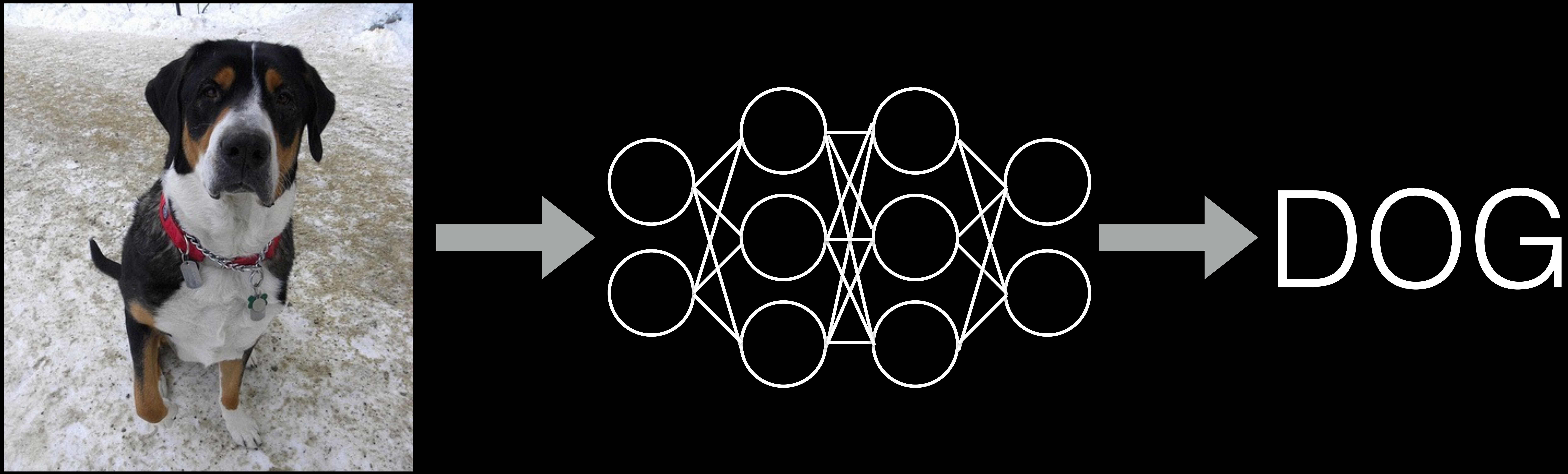
Standard Training Dataset



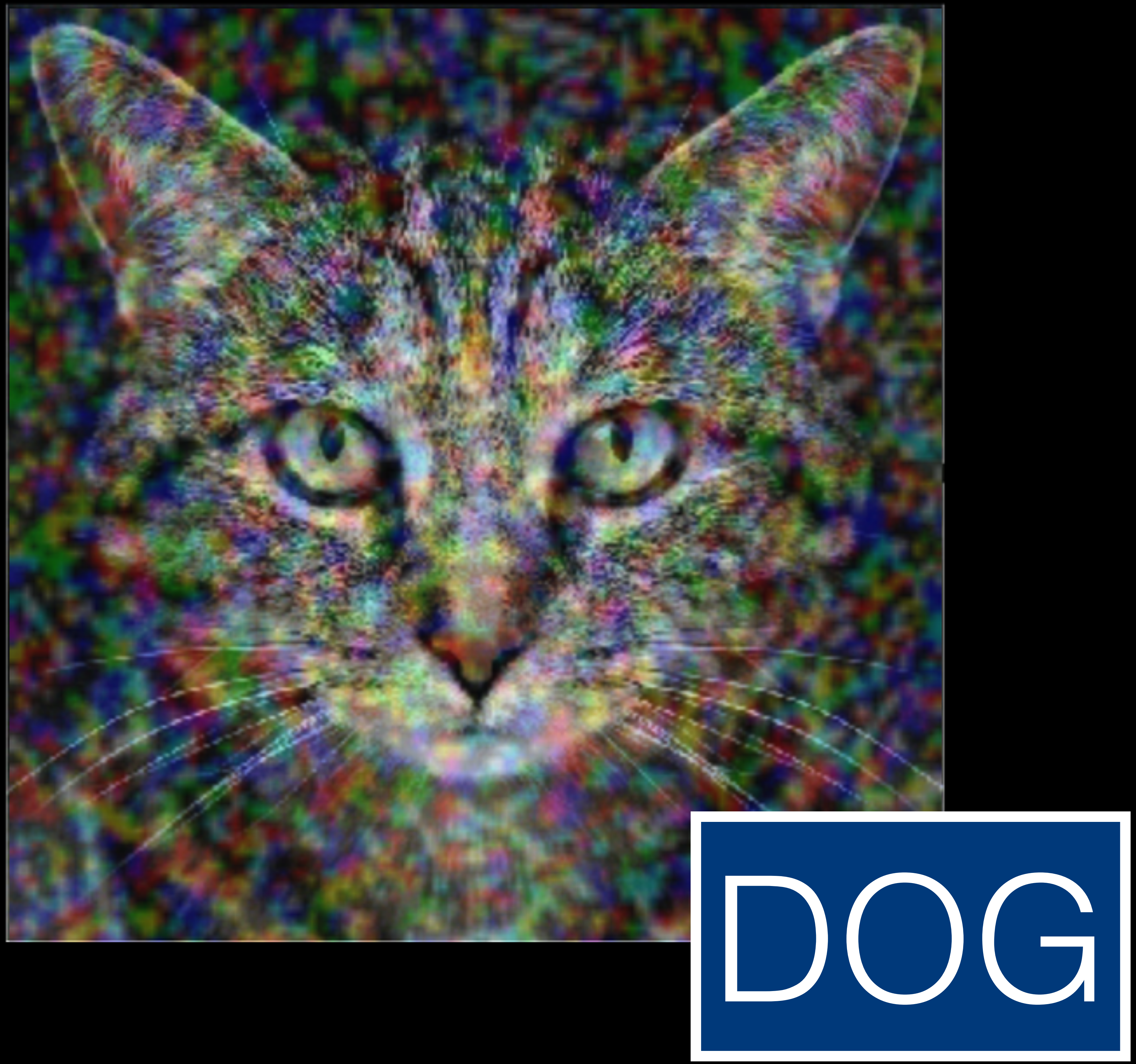
Adversarial Training Dataset



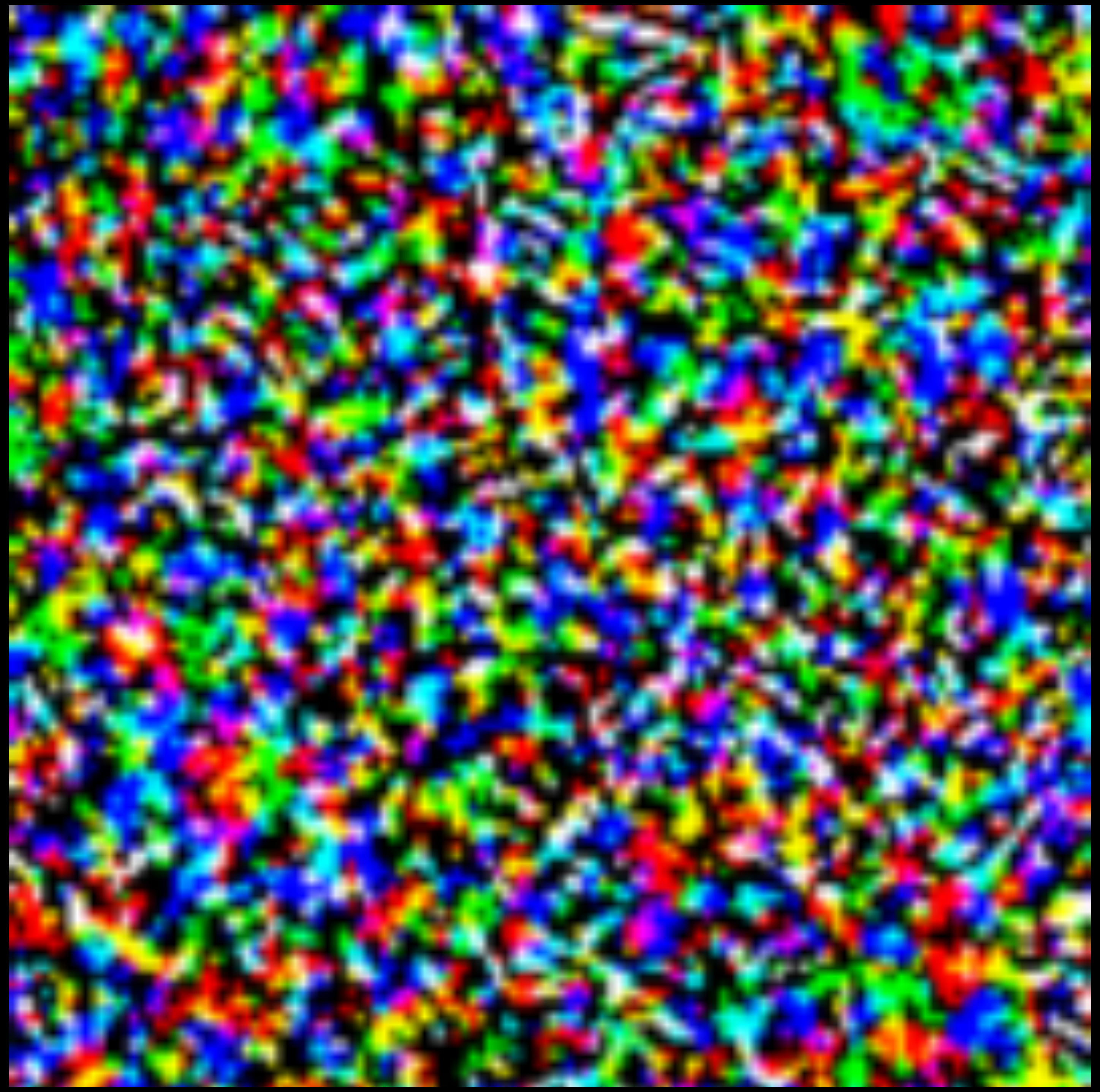
Confusing Training Dataset



Standard Testing Setup

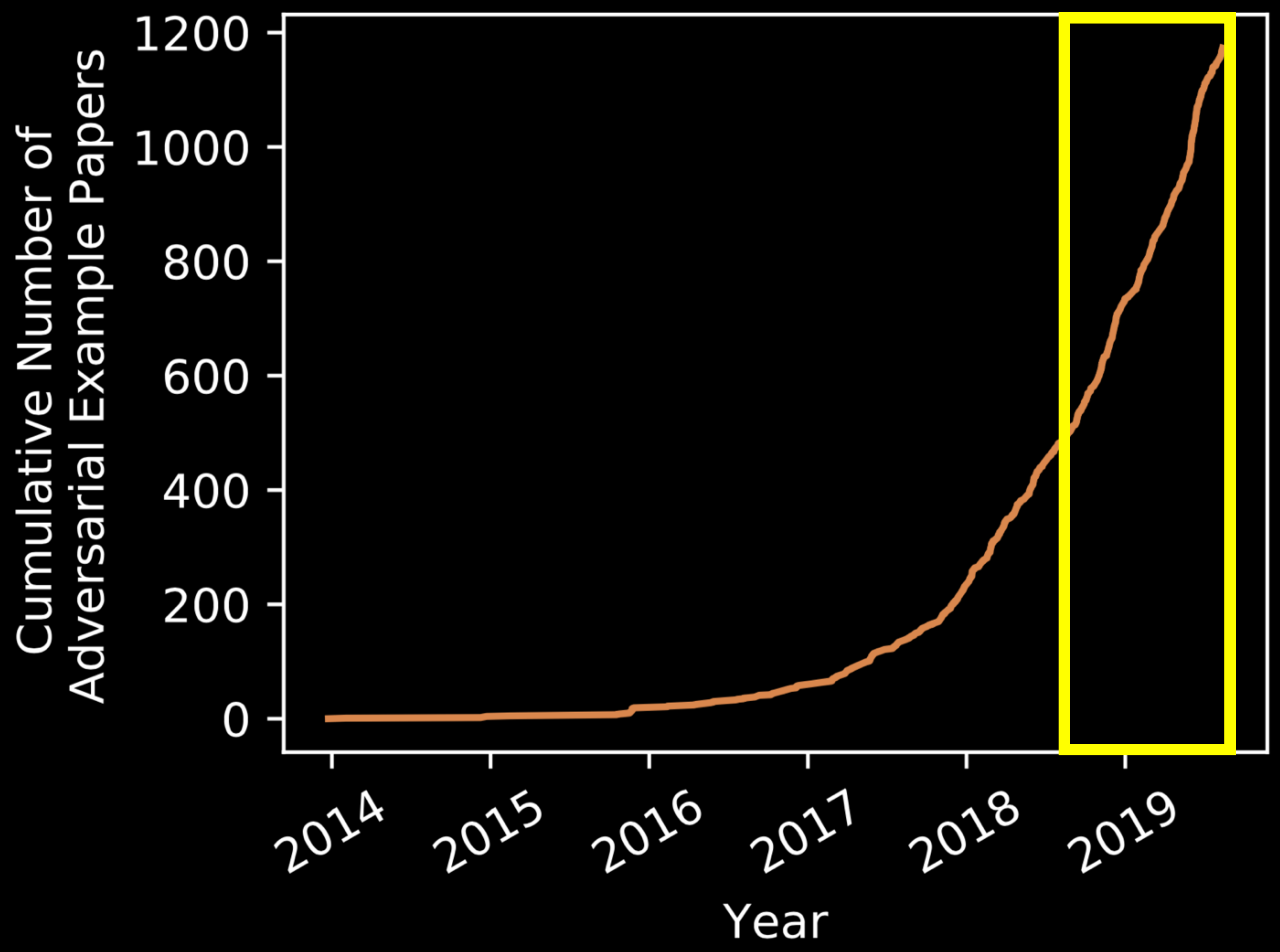


?!??!?!?? Training Dataset



Is a **well-generalizing feature** of CAT

Conclusion


Cumulative Number of
Adversarial Example Papers
1200
1000
800
600
400
200
0
2014
2015
2016
2017
2018
2019
Year

Questions?