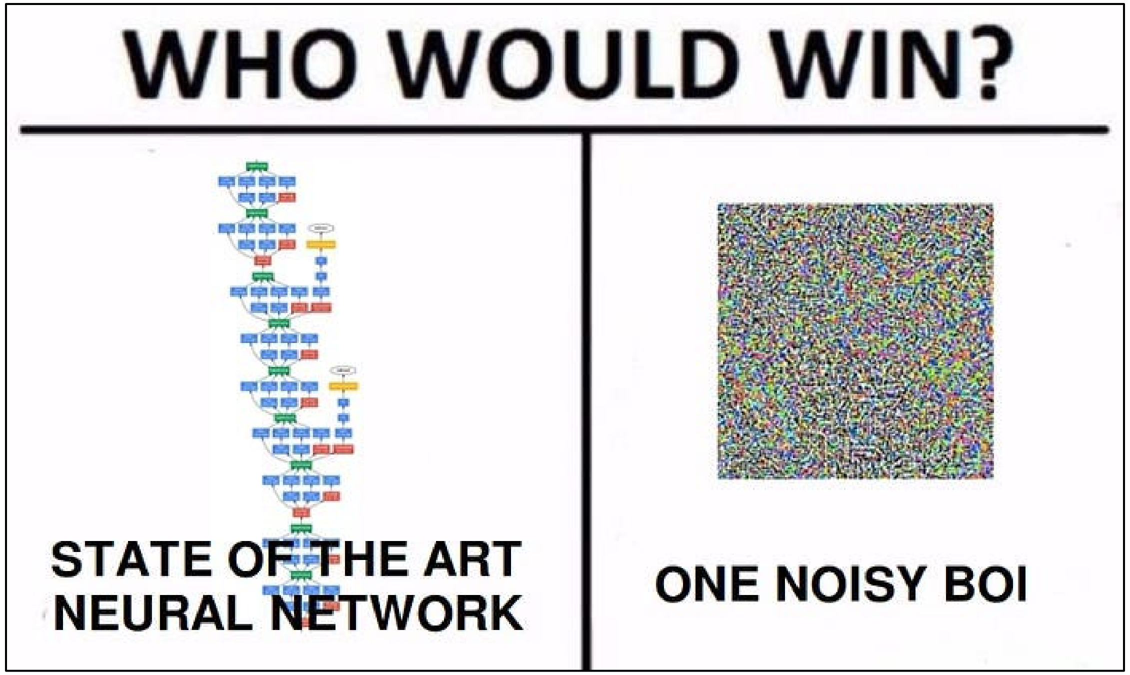
CMSC 475 / 675 Neural Networks

Lecture 15

Robustness of Neural Networks



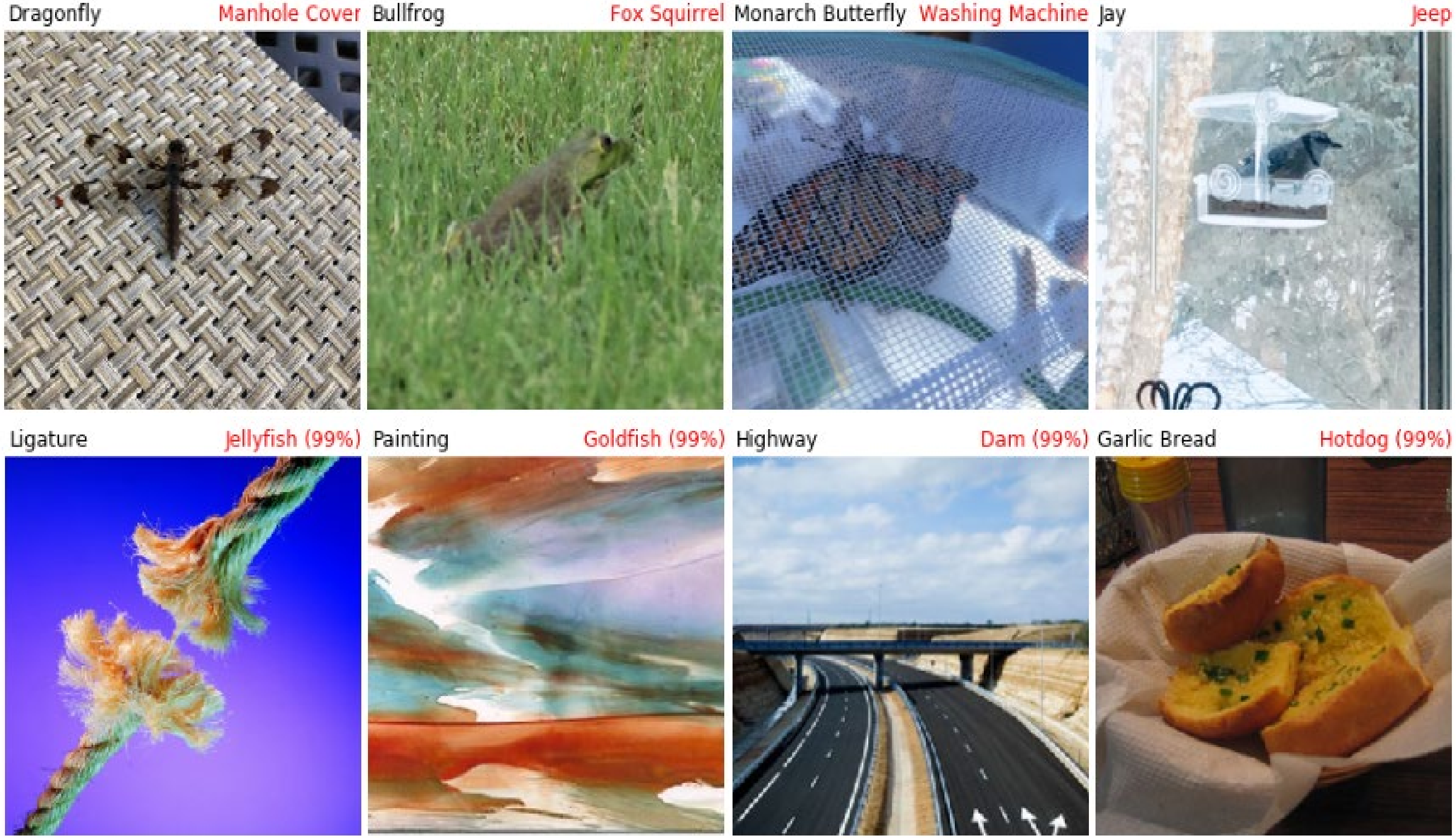
UMBC

Chihuahua … or Muffin … ?

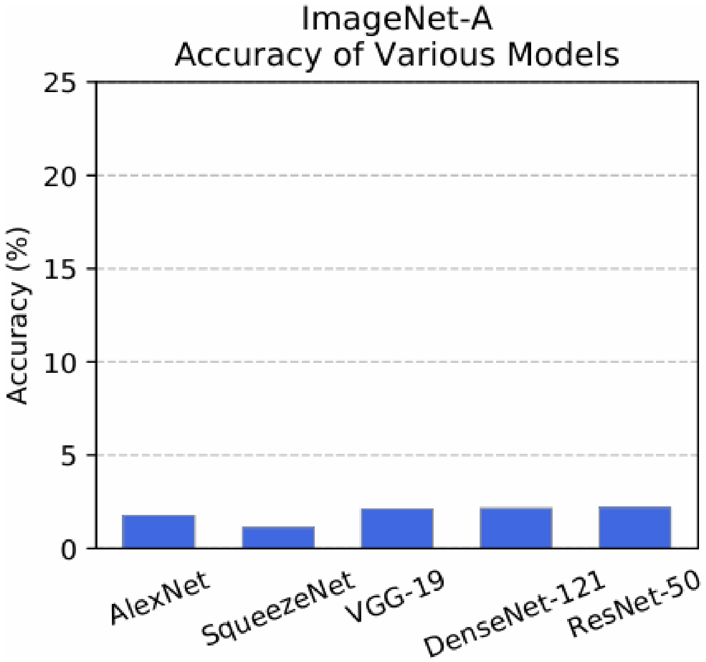
Chihuahua ... or Muffin ... ?

Mop ...
or Sheepdog ... ?

NNs get fooled ... *very easily* ...



"manually constructed by graduate students over several months"



source: Hendrycks et al. CVPR 2021 https://arxiv.org/abs/1907.07174 Data: https://github.com/hendrycks/natural-adv-examples

Machine Learning: The Success Story

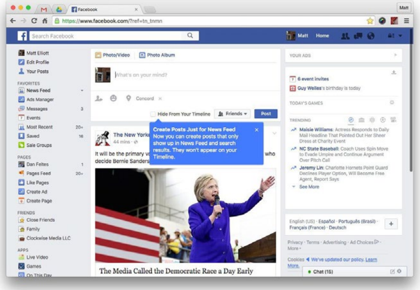
IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

Andrew Ng @AndrewYNg
Follow

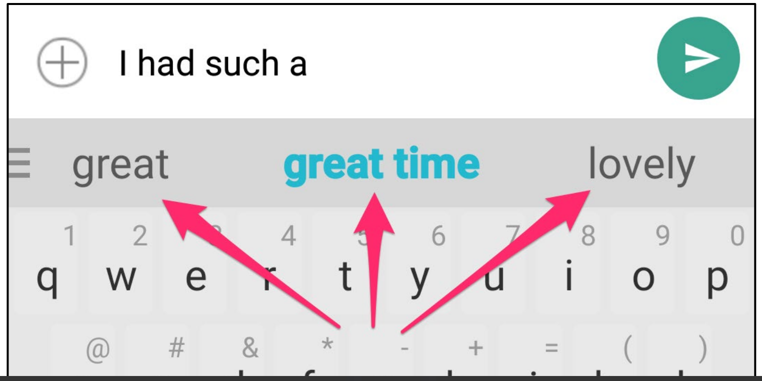
"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

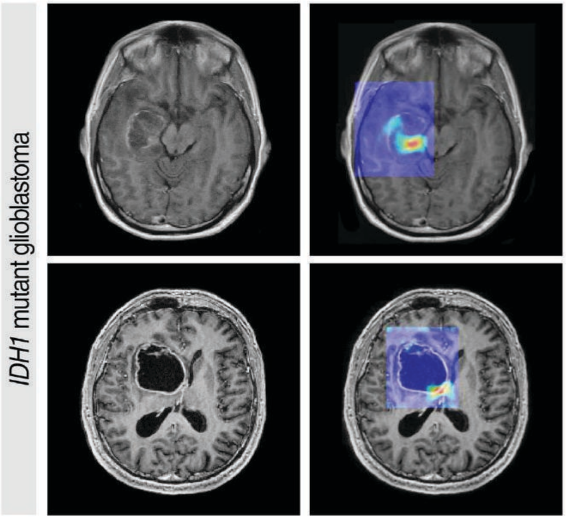
2016: The Year That Deep Learning Took Over t

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE



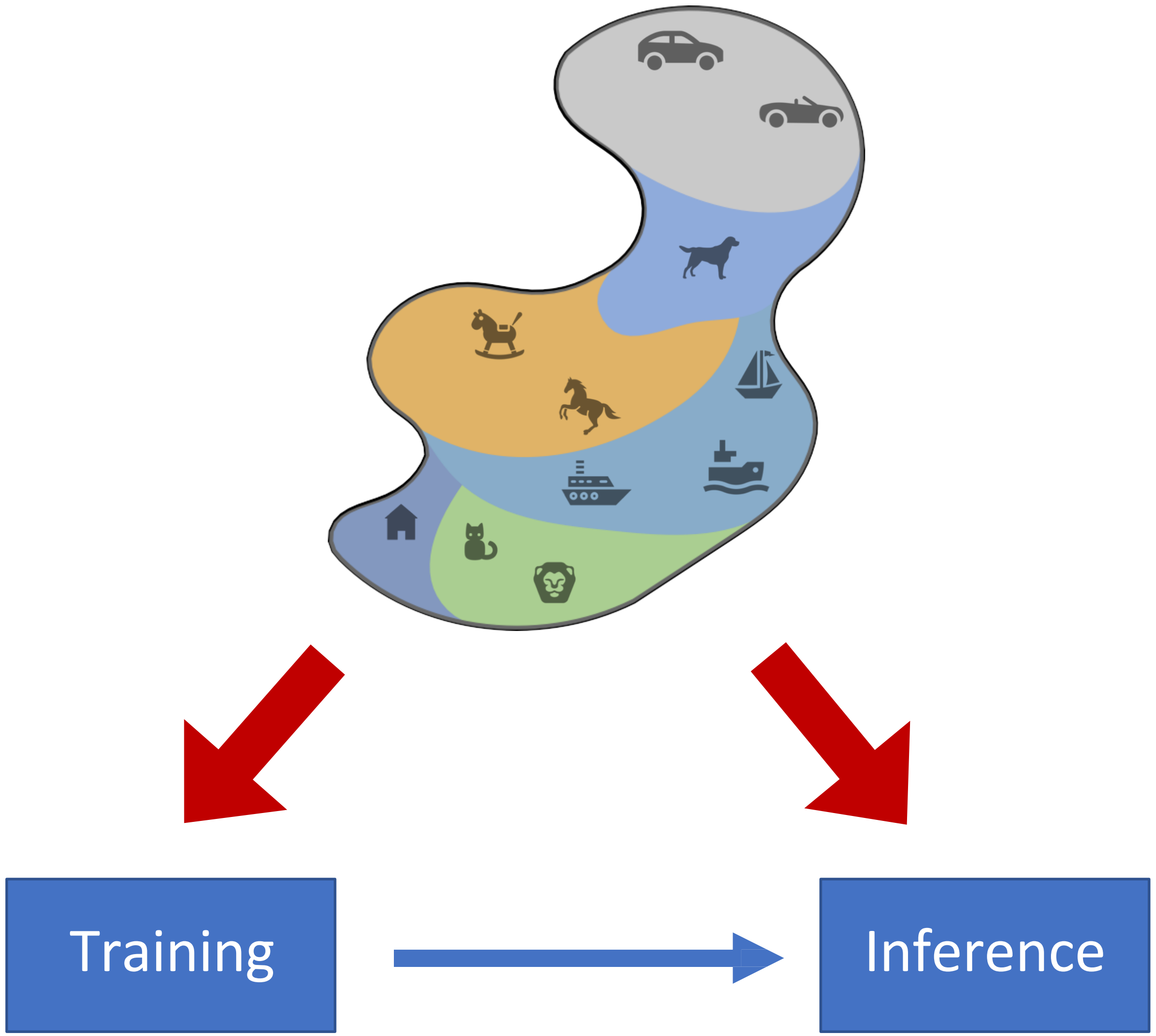
Models that learn from data are embedded in our lives





But …

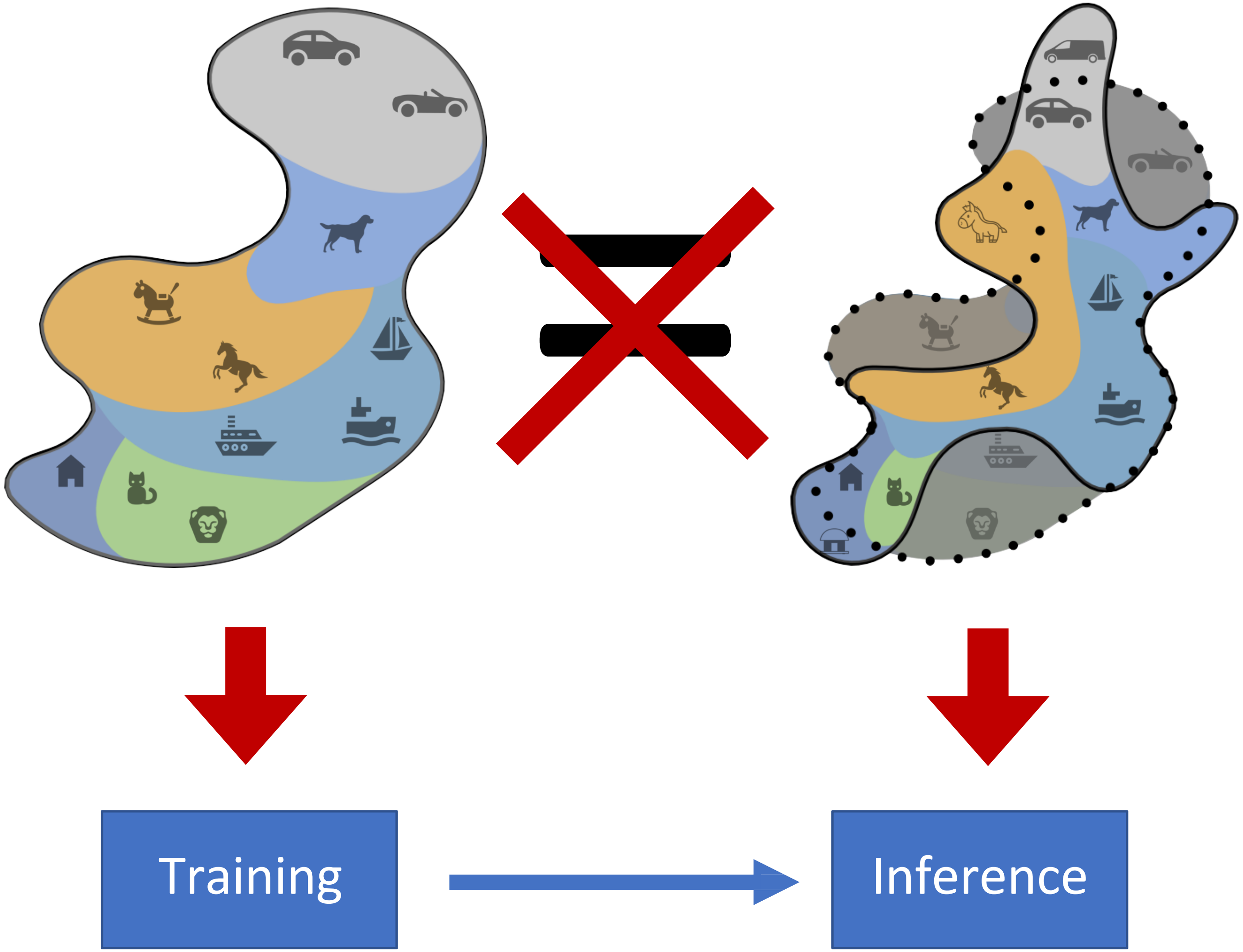
A Limitation of the (Supervised) ML Framework



Measure of performance:
Fraction of mistakes during testing

But: In reality, the distributions we **use** ML on are NOT the ones we **train** it on

A Limitation of the (Supervised) ML Framework



Measure of performance:
Fraction of mistakes during testing

But: In reality, the distributions we **use** ML on are NOT the ones we **train** it on

What can go wrong?

Standard *i.i.d.* Assumption in Machine Learning

Training Data Distribution

=

Test Data Distribution

"Independent and Identically Distributed"

Models learn useful patterns

Standard *i.i.d.* Assumption in Machine Learning

Training Data Distribution $\neq$ Test Data Distribution

IID Assumption collapses in real-world "in-the-wild" settings

Model performance deteriorates

Are Neural Networks Robust?

Findings from Previous Work (in computer vision, but similar results in other domains too)

Poses can fool Image Classifiers

school bus 1.0 garbage truck 0.99 punching bag 1.0 snowplow 0.92

Alcorn et al. CVPR 2019

Poses can fool Image Classifiers

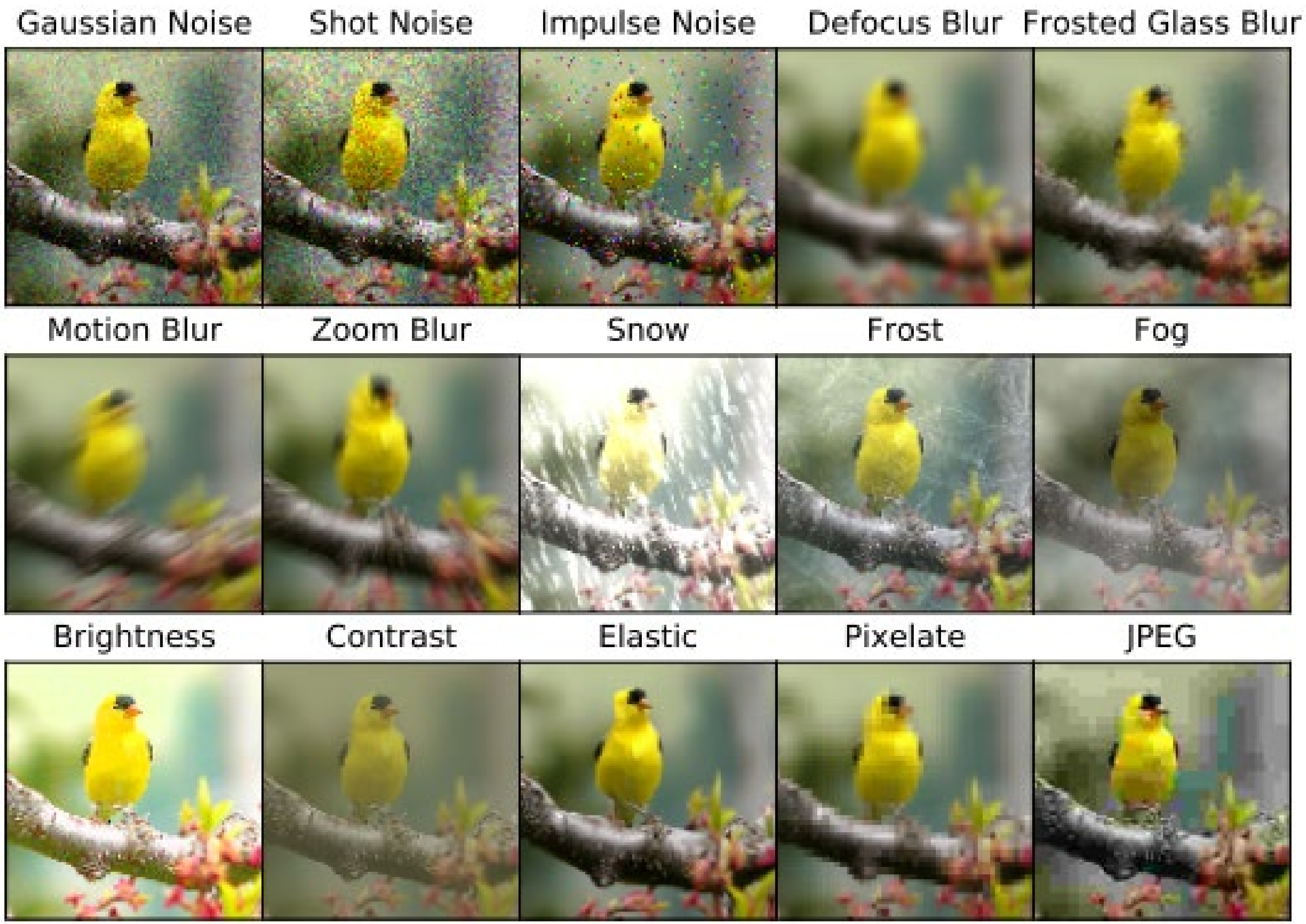
- **Goal:** correctly classify previously unseen test images.

- Statistical ML operates with the “i.i.d.” assumption
- But real-world test inputs are *often* NOT i.i.d. !!!
  - Poses can fool classifiers
    - Rotation
    - Translation
    - Scale
    - Occlusion
    - …

school bus 1.0 garbage truck 0.99 punching bag 1.0 snowplow 0.92

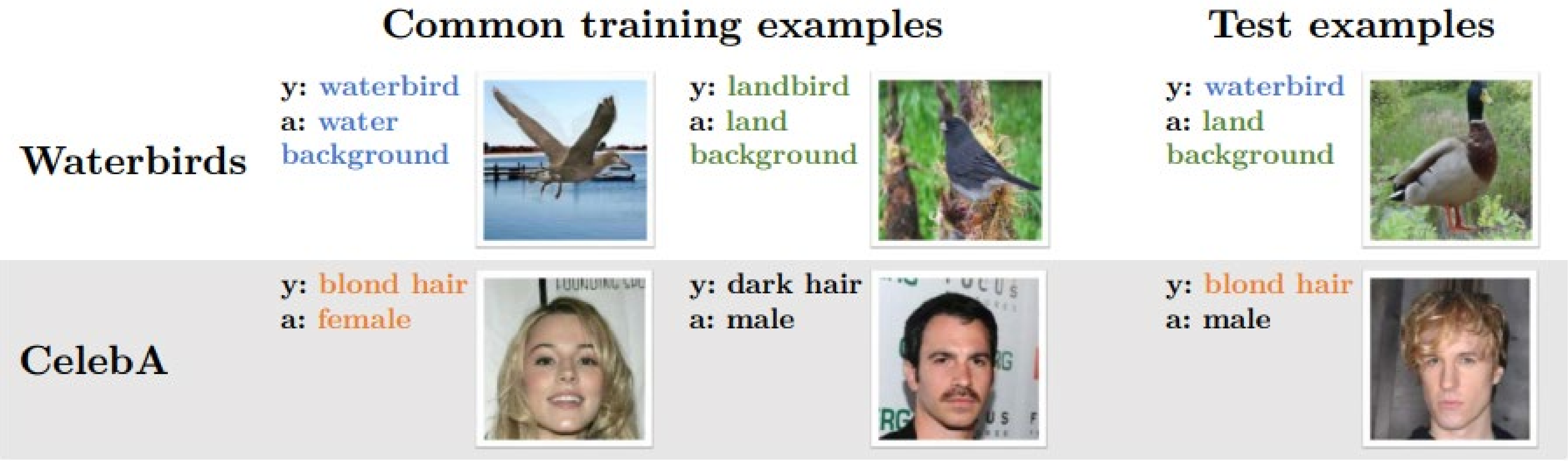
Alcorn et al. CVPR 2019

Natural Corruptions affect accuracy



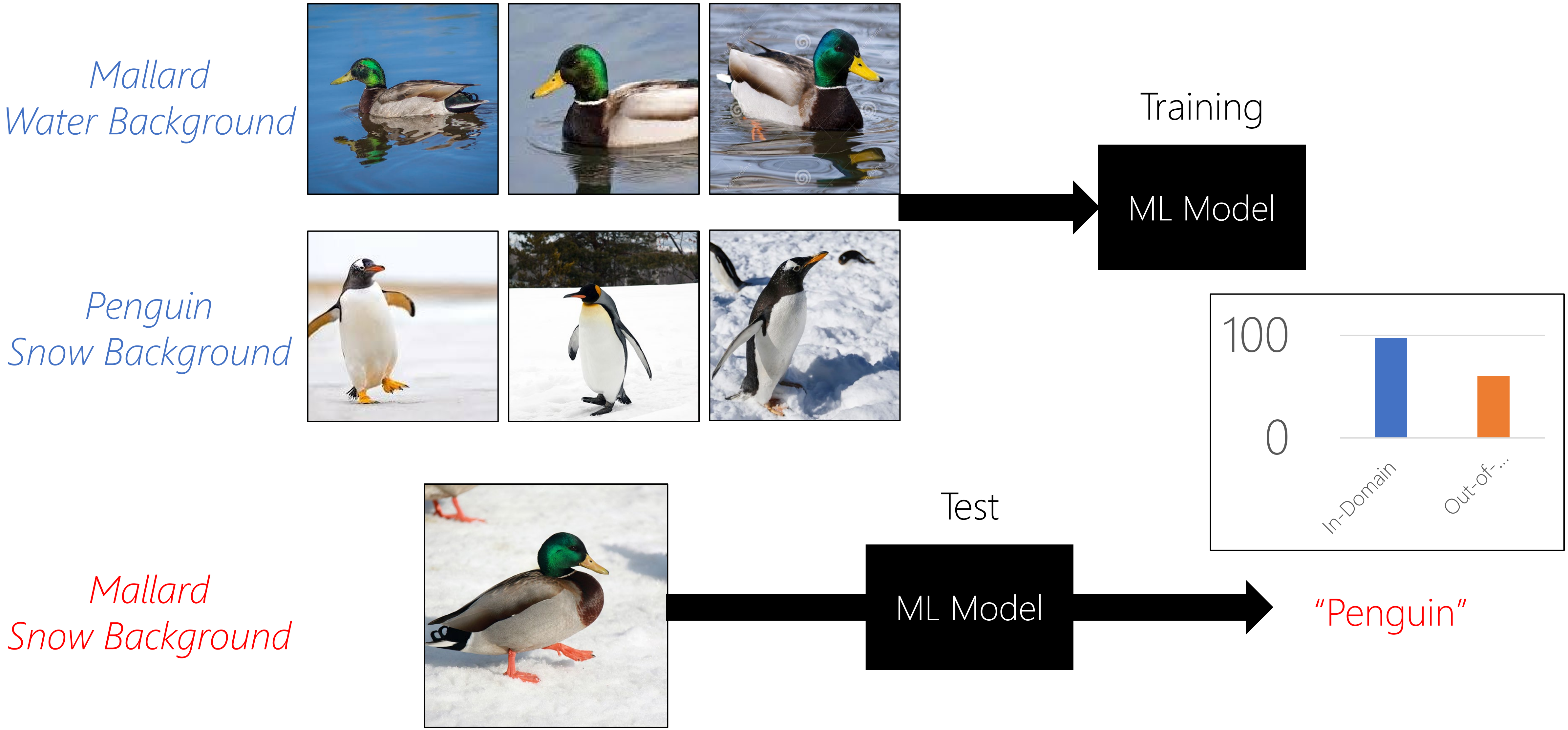
Hendrycks et al. ICLR 2019

Spurious Correlations / Biased Datasets

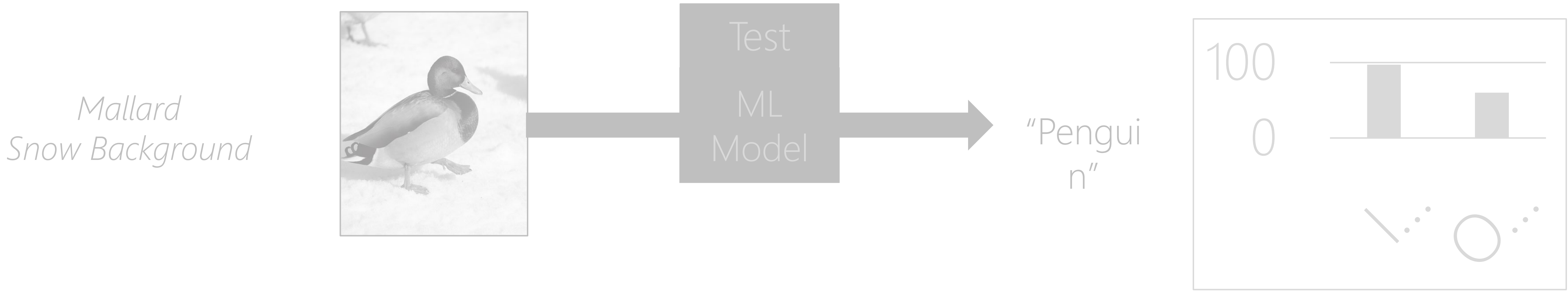


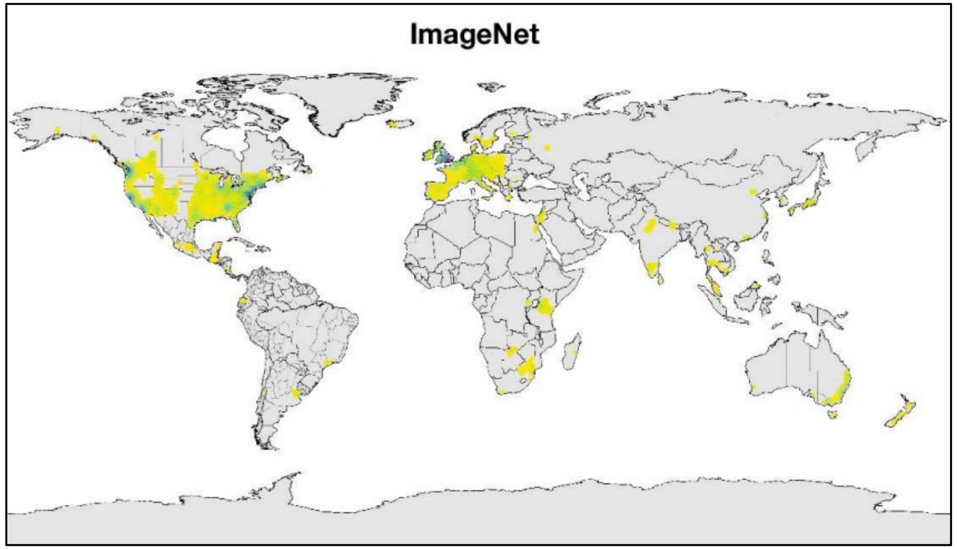
Sagawa et al. ICLR 2020

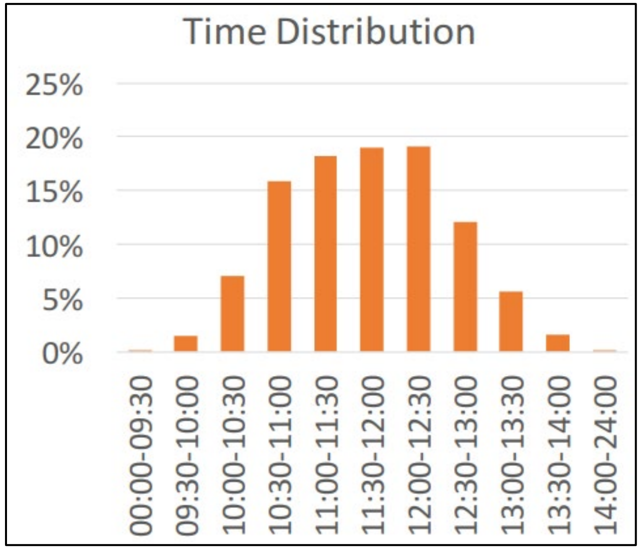
Lack of Diverse Data hurts Reliability

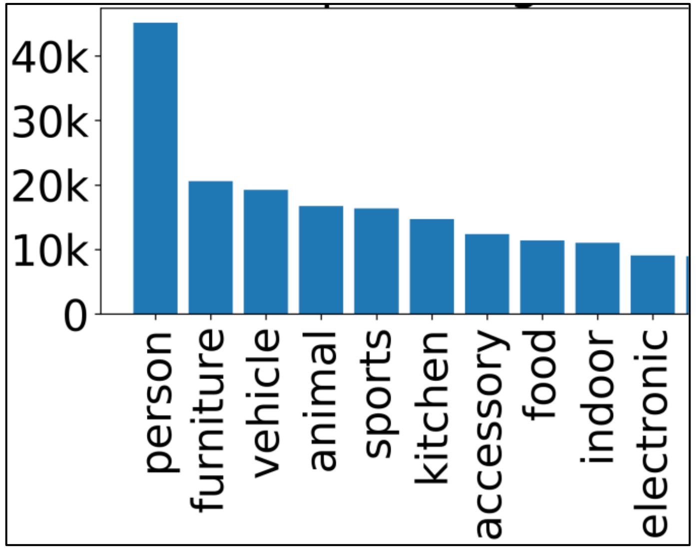


Lack of Diverse Data hurts Reliability

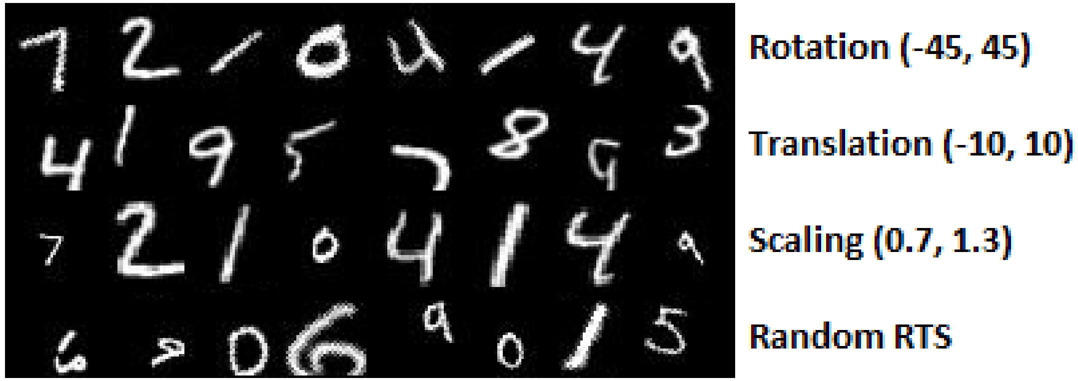


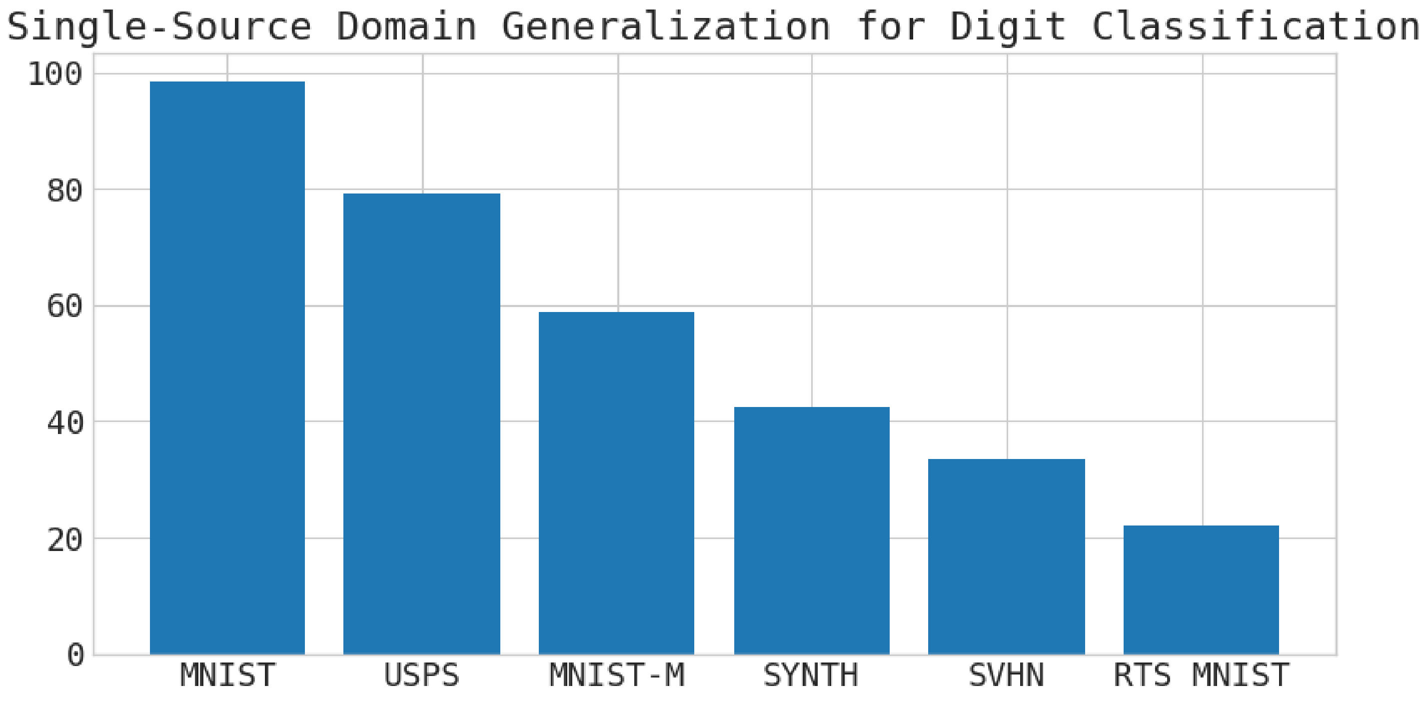






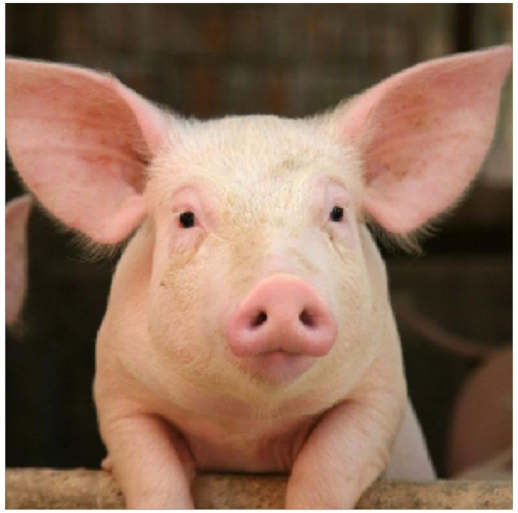
Domain Shift is a Nuisance





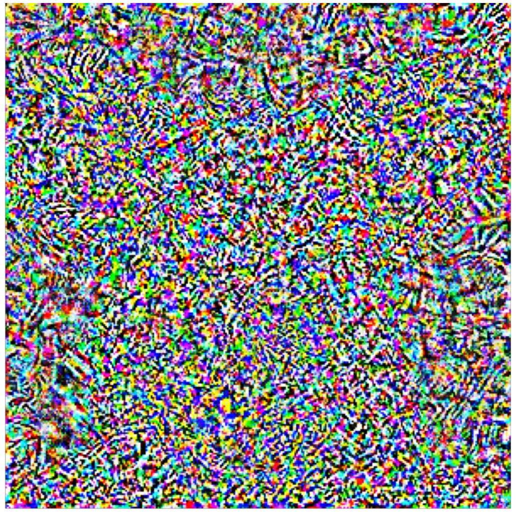
ML Predictions Are (Mostly) Accurate but Brittle

“pig” (91%)



+ 0.005 x

noise (NOT random)



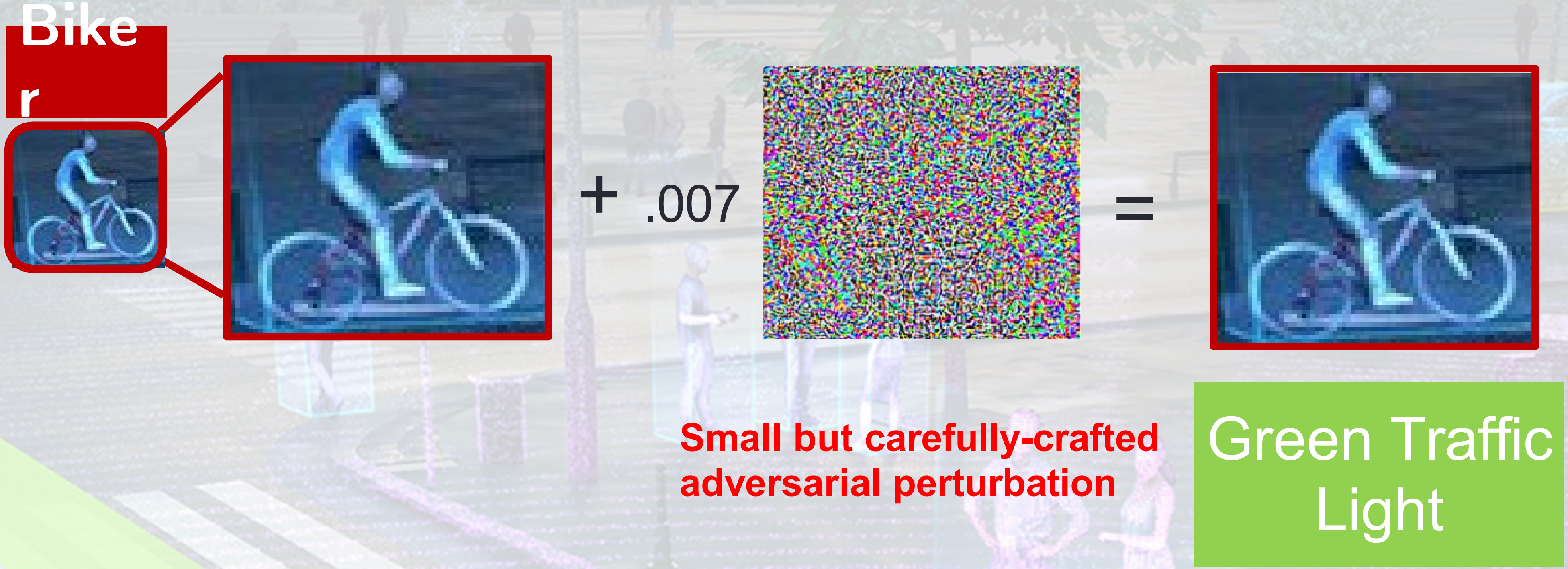
=

“airliner” (99%)

[Szegedy Zaremba Sutskever Bruna Erhan Goodfellow Fergus 2013]
[Biggio Corona Maiorca Nelson Srndic Laskov Giacinto Roli 2013]

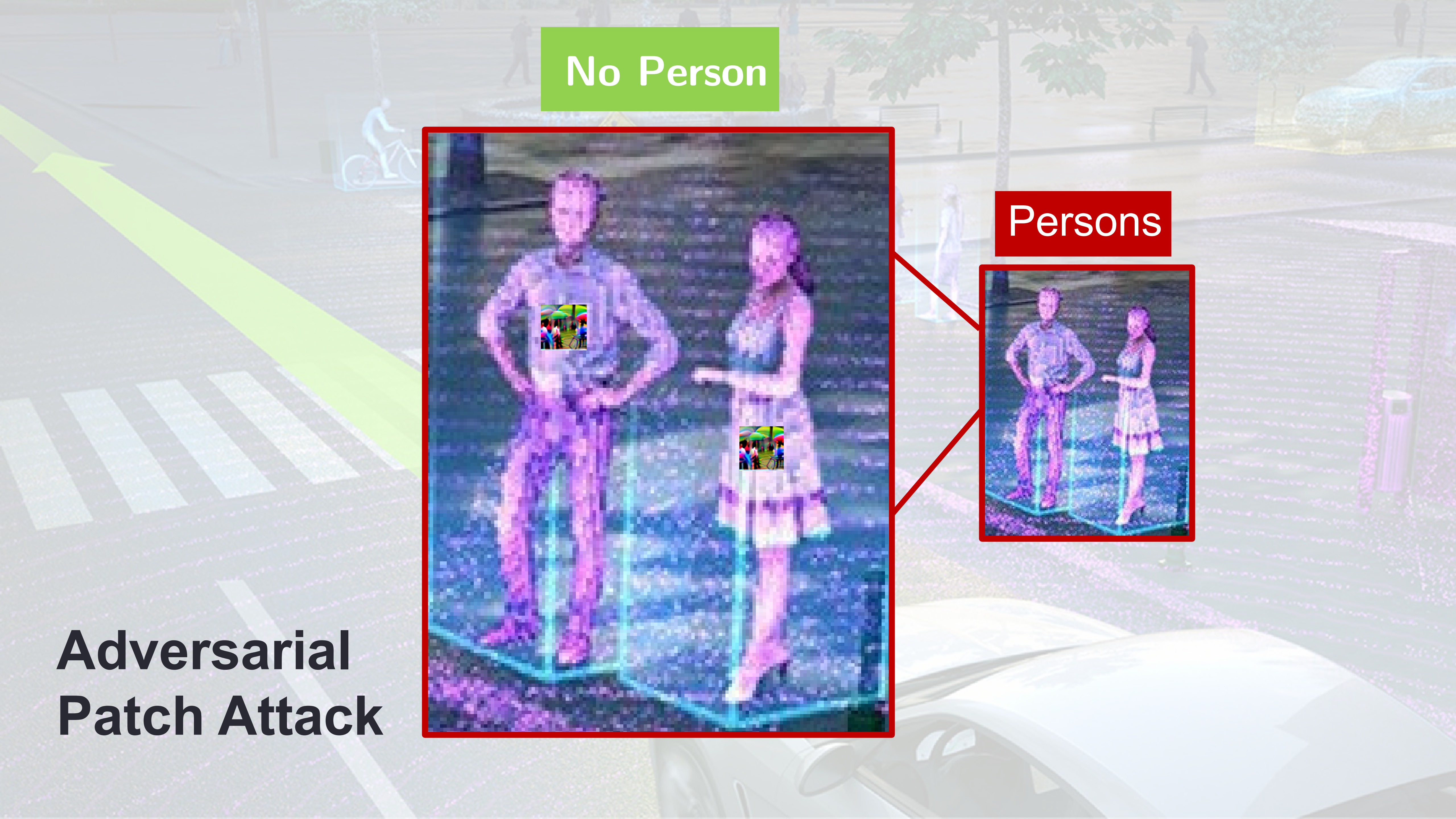
But also: [Dalvi Domingos Mausam Sanghai Verma 2004][Lowd Meek 2005]
[Globerson Roweis 2006][Kolcz Teo 2009][Barreno Nelson Rubinstein Joseph Tygar 2010]
[Biggio Fumera Roli 2010][Biggio Fumera Roli 2014][Srndic Laskov 2013]


Biker
Pedestrian Sign
Persons



Adversarial Perturbation Attack


Pedestrian Sign
Speed Limit 45 Sign
Adversarial Rotation Attack


No Person
Persons
Adversarial
Patch Attack

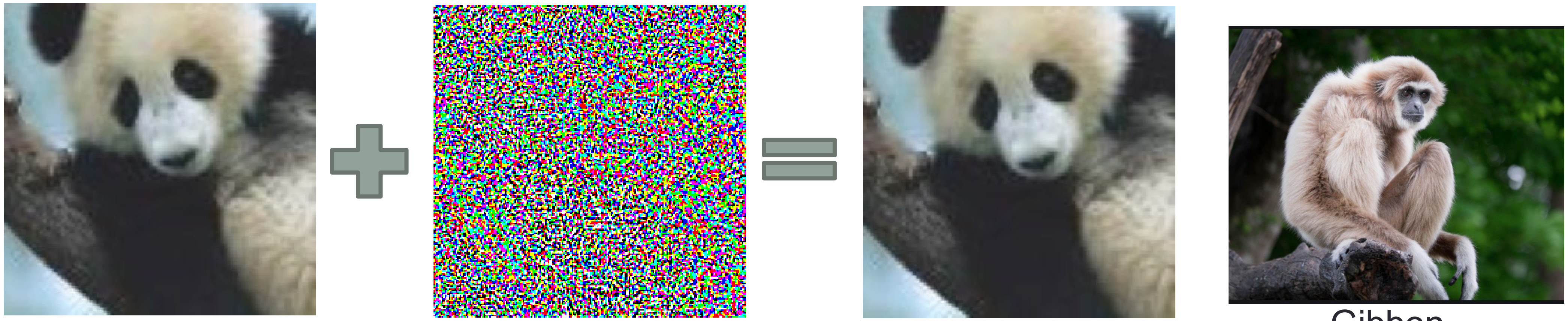
Adversarial Examples

- In 2014, one of the seminal papers of Goodfellow et al. shows that an adversarial image of a panda can fool the ML model to output "gibbon"

- This paper was influential in building community interest in "adversarial ML"

Original image

Adversarial image



Gibbon

Classified as panda
57.7% confidence

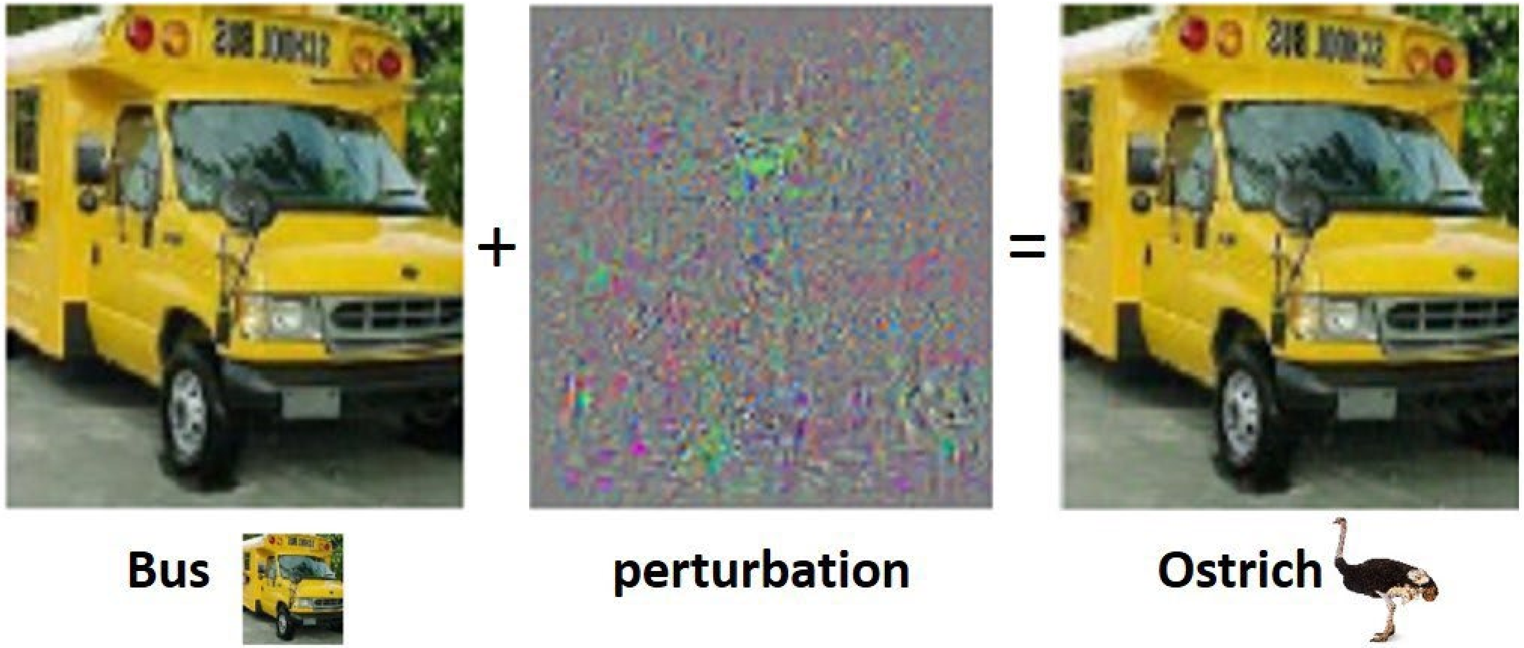
Small adversarial noise

Classified as gibbon
99.3% confidence

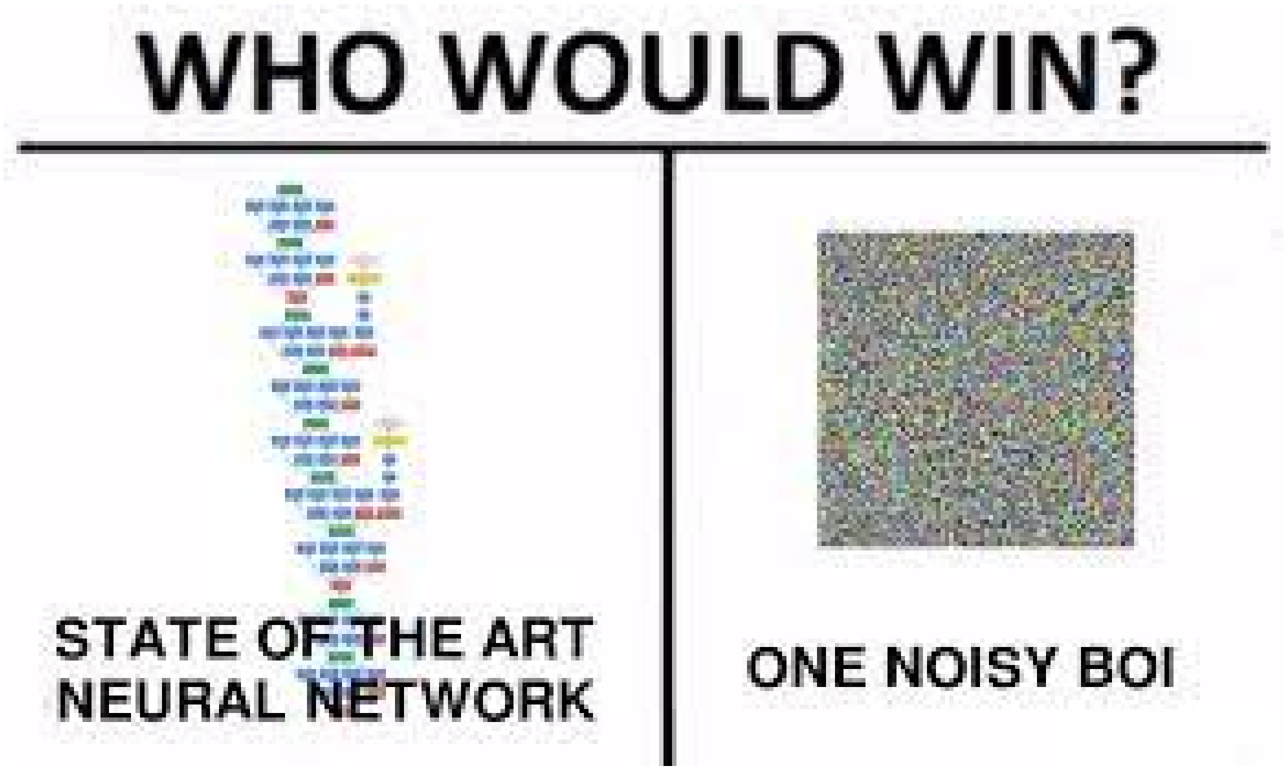
Picture from: Goodfellow et al. (2014) – Explaining and Harnessing Adversarial Examples

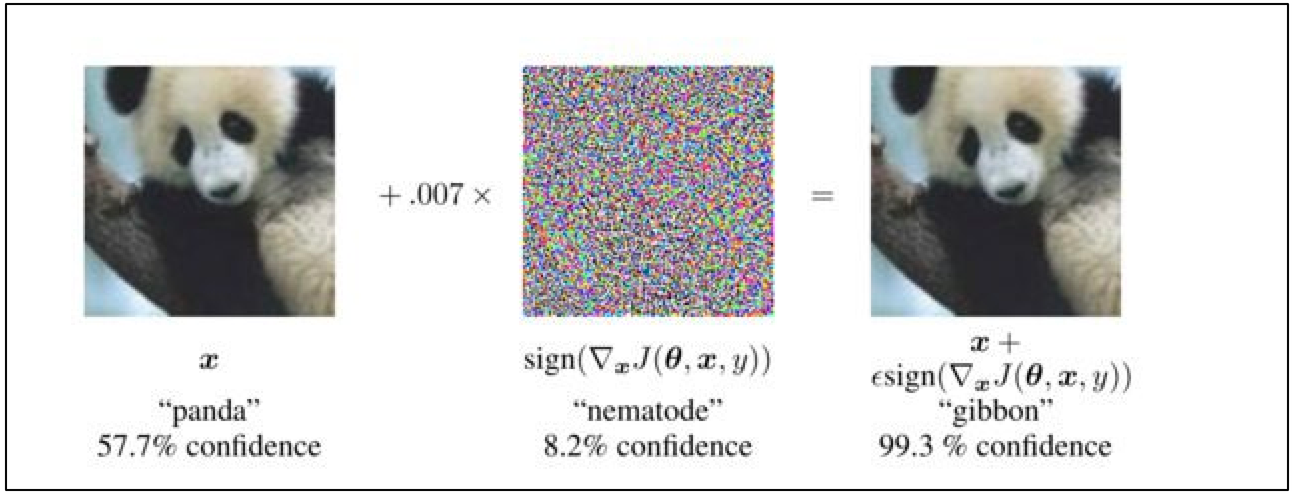
Adversarial Examples

- Similar example, from Szegedy et al. (2014)



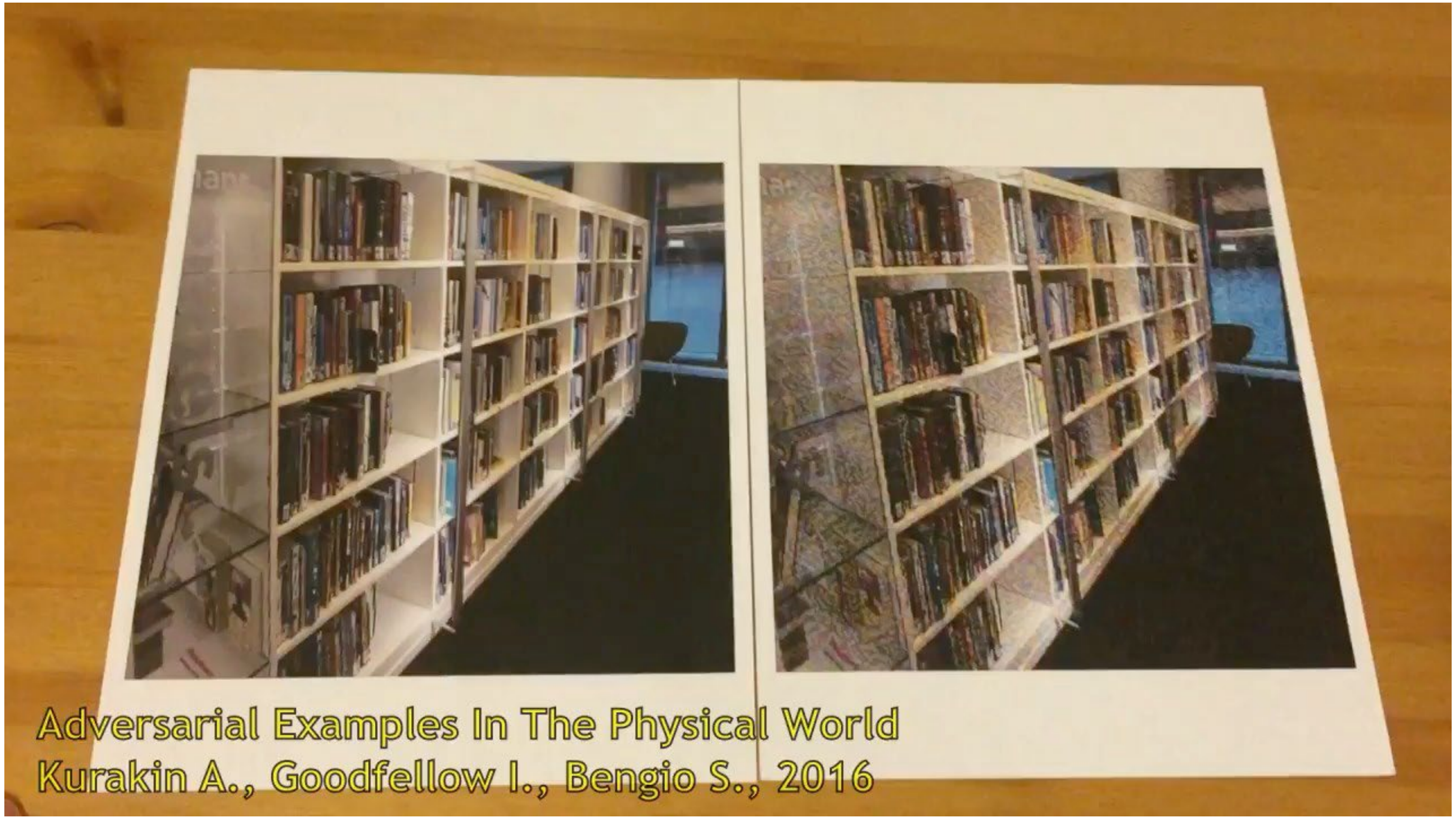
Picture from: Szagedy et al. (2014) – Intriguing Properties of Neural Networks


WHO WOULD WIN?
STATE OF THE ART NEURAL NETWORK
ONE NOISY BOI


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

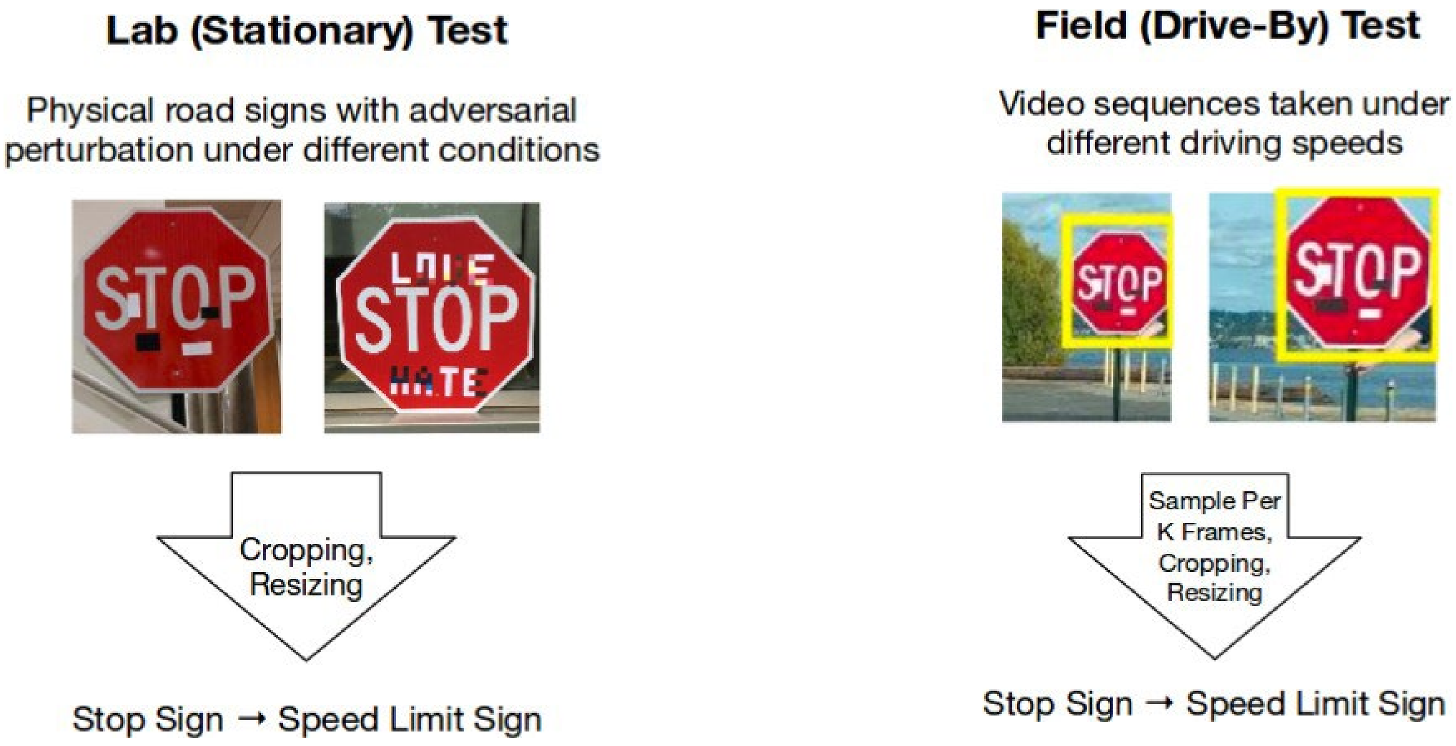
Physical-World Attack: Printed Adversarial Images

- Not only adversarial examples in the digital world, but printed adversarial images can also fool machine learning models



Physical-World Attack: Adversarial STOP Sign

- An example of manipulating a STOP sign with adversarial patches
  - **Methodology:** carefully design a patch and attach it to the STOP sign
  - Cause the DL model of a self-driving car to misclassify it as a Speed Limit 45 sign
    - The authors achieved 100% attack success in lab test, and 85% in field test



Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

Physical-World Attack: Adversarial STOP Sign

- More examples of lab test for STOP signs with a target class Speed Limit 45

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

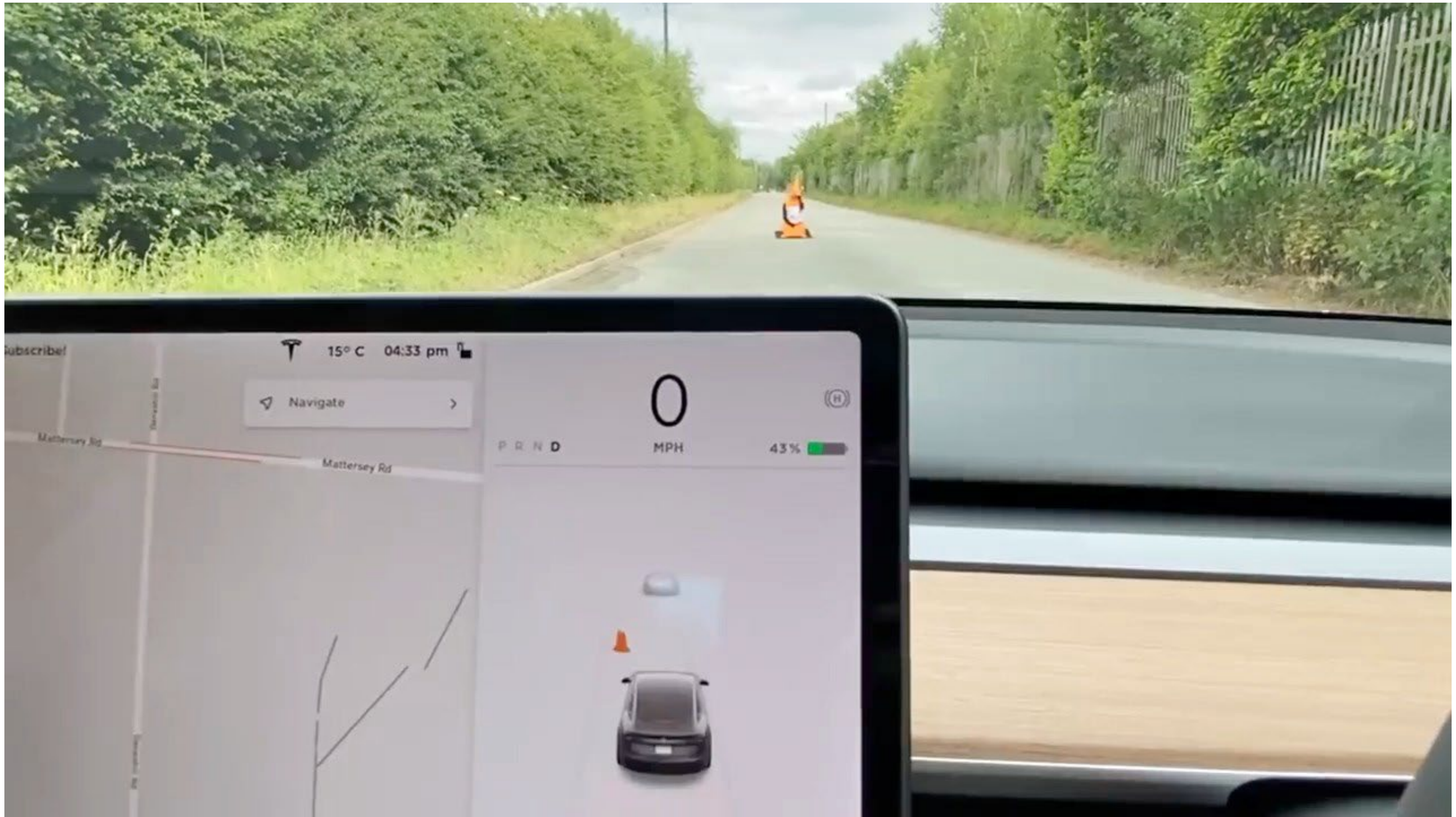
Physical-World Attack: Adversarial Patch

- Not only adversarial patch can fool a classifier, but also a SOTA detector
- An example of a person wearing an adversarial patch who cannot be detected by a YOLOv2 model
  - This can be used by intruders to get past security cameras

Thys et al. (2019) - Fooling automated surveillance cameras: adversarial patches to attack person detection

Physical-World Attack: Attack Tesla Autopilot

- A Tesla owner checks if the car can distinguish a person wearing a cover-up from a traffic cone



Physical-World Attack: Attack Waymo Autopilot

BUSINESS

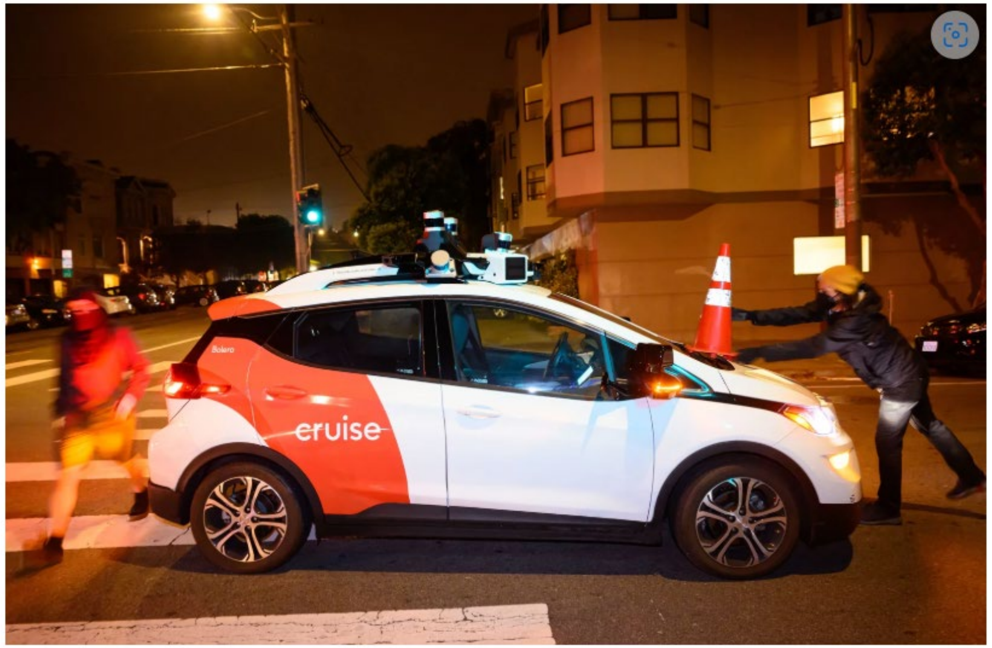
Armed with traffic cones, protesters are immobilizing driverless cars

AUGUST 26, 2023 · 7:01 AM ET

By Dara Kerr

A Waymo driverless car gets coned in San Francisco.

Dara Kerr/NPR



Members of Safe Street Rebel place a cone on a self-driving Cruise car in San Francisco.

Josh Edelson/AFP via Getty Images

Two people dressed in dark colors and wearing masks dart into a busy street on a hill in San Francisco. One of them hauls a big orange traffic cone. They sprint toward a driverless car and quickly set the cone on the hood.

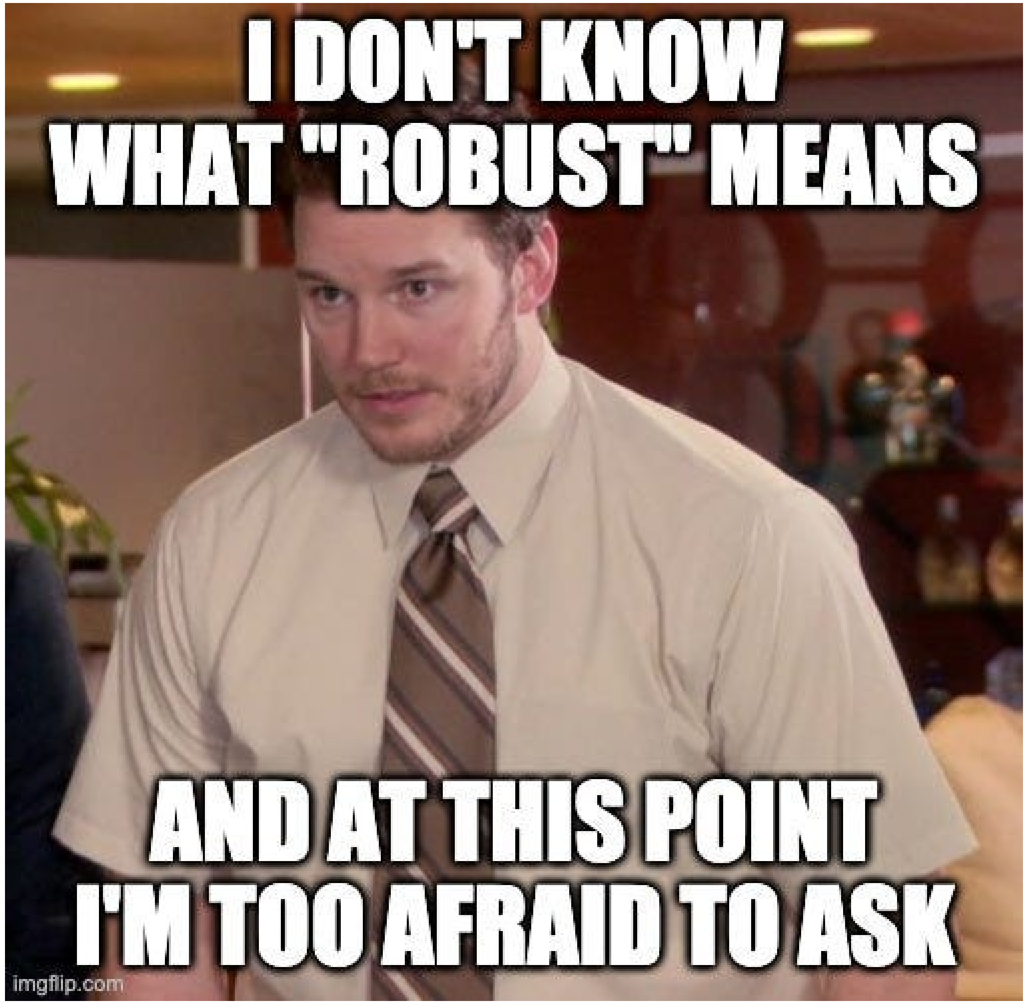
The vehicle's side lights burst on and start flashing orange. And then, it sits there immobile.

"All right, looks good," one of them says after making sure no one is inside. "Let's get out of here." They hop on e-bikes and pedal off.

All it takes to render the technology-packed self-driving car inoperable is a traffic cone. If all goes according to plan, it will stay there, frozen, until someone comes and removes it.

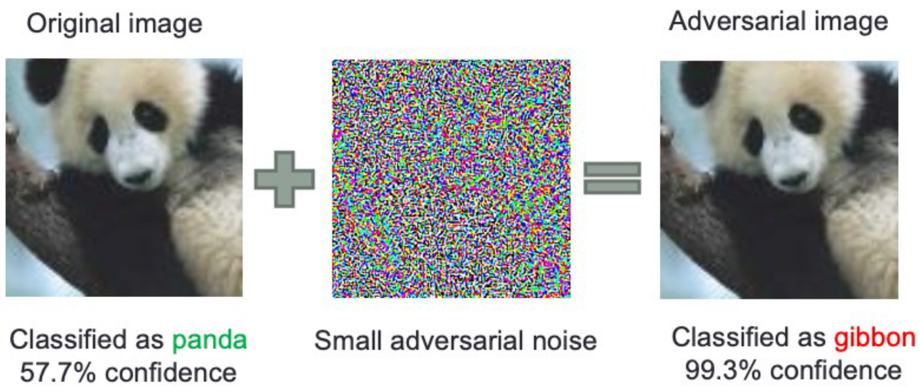

THIS IS FINE.

Robustness Definitions

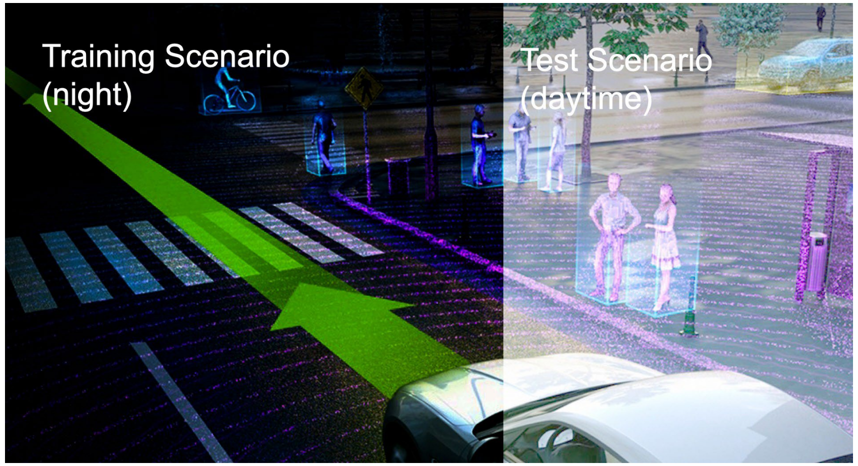


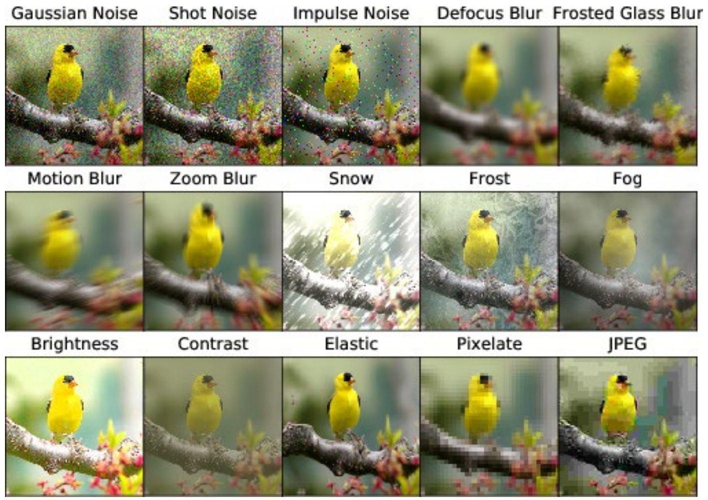
Adversarial vs. Out-of-distribution

- Each of the described attacks can further be:
  - ***Adversarial*** example (algorithmically manipulated example)



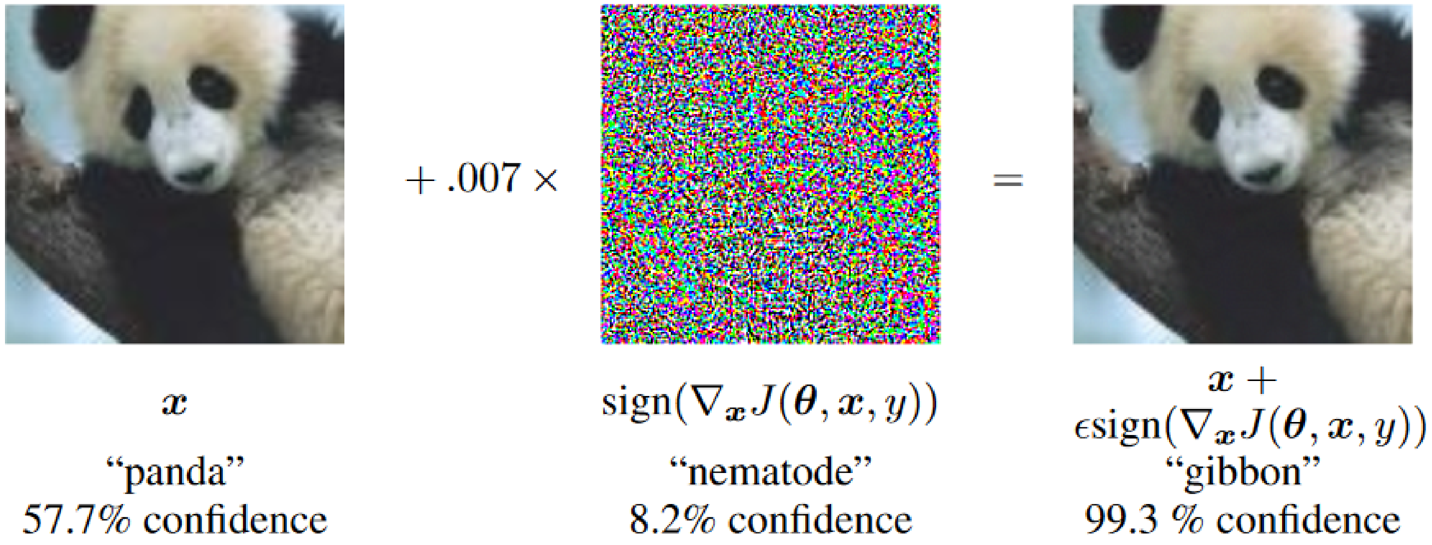
  - ***Out-of-distribution*** example (natural example)





Adversarial "Attacks"

- Algorithms that can "find" perturbations to add to images, in order to fool classifiers
- Given image $\boldsymbol{x}$, find $\boldsymbol{g}(\boldsymbol{x})$ s.t. $\boldsymbol{x} + \epsilon \boldsymbol{g}(\boldsymbol{x})$ fools classifier
- Perturbations are typically norm-bounded



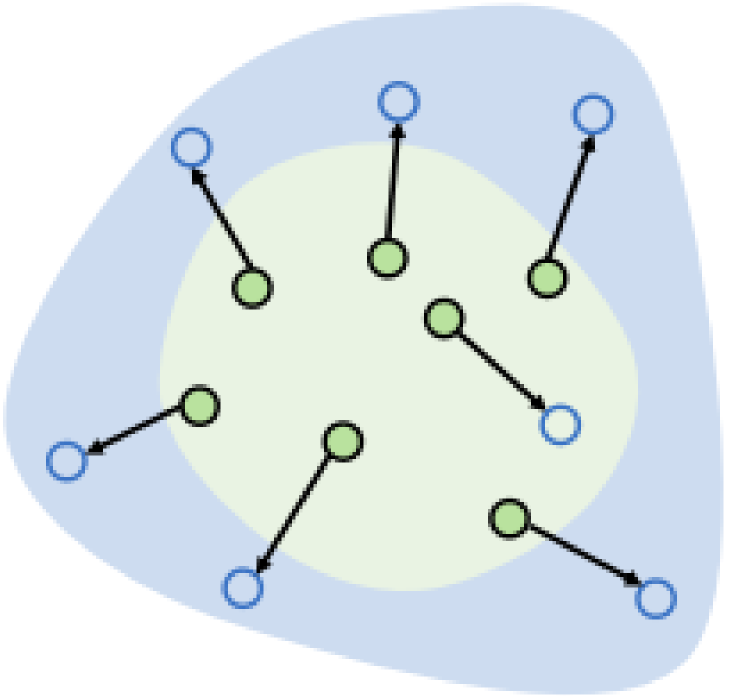
Goodfellow et al. (2014) – Explaining and Harnessing Adversarial Examples

A web demo to create your own adversarial examples

https://kennysong.github.io/adversarial.js/

Handwritten notes ...

Adversarial “Defense”



- Algorithms that make the model robust against attacks
- Leverages the concept of adversarial examples, to improve classifier robustness to such attacks
- min—max optimization
  - *maximization:* find adversarial images
  - *minimization:* train classifier to correctly classify such images
- norm-bounded perturbations: robustness within the norm-ball

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$

Madry et al. ICLR 2018. "Adversarial Training"

In Fall 2024, I taught a

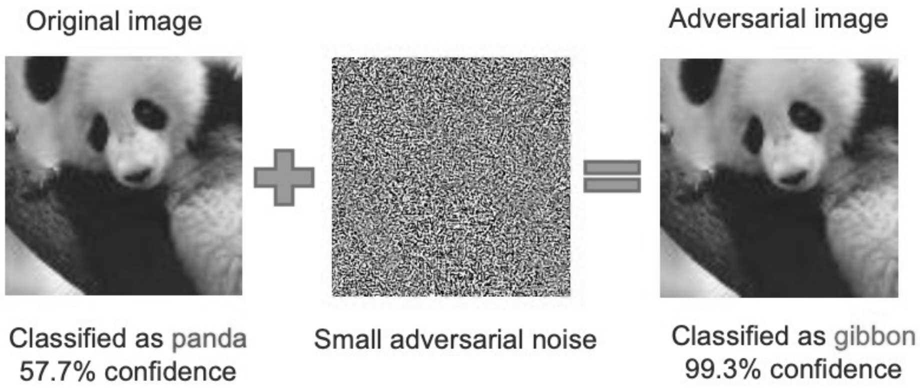
Special Topics course on “Robust ML”

Slides are available for advanced reading!

https://courses.cs.umbc.edu/graduate/691rml/

Today:

- Each of the described attacks can further be:
  - ***Adversarial*** example (algorithmically manipulated example



Next time:

- ***Out-of-distribution*** example (natural example)