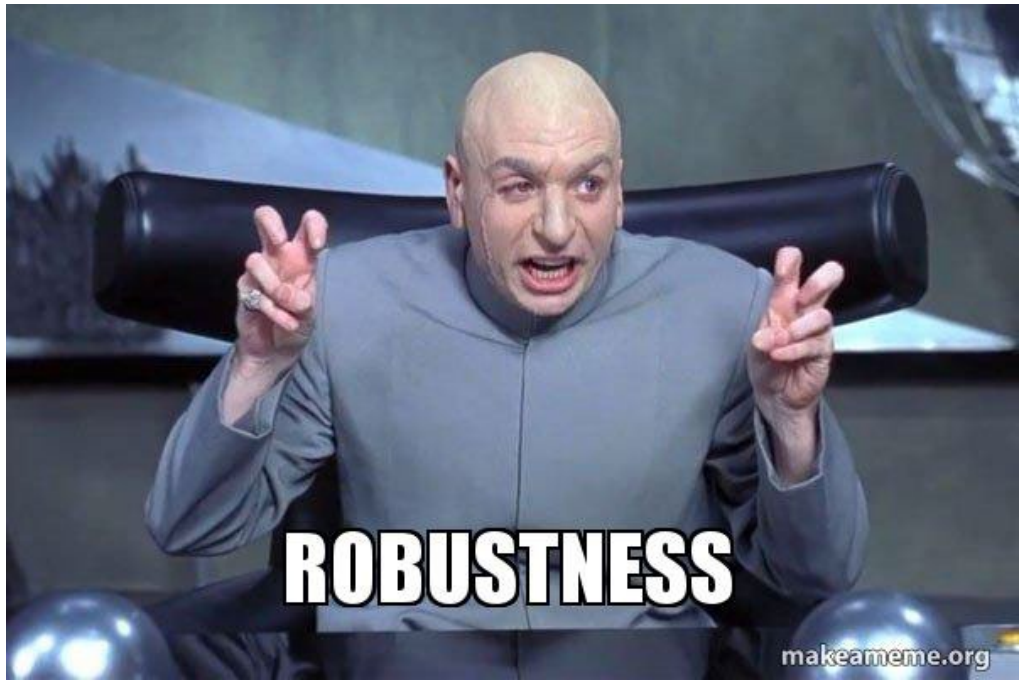
Lecture 20b

Robustness in Computer Vision



Machine Learning: The Success Story

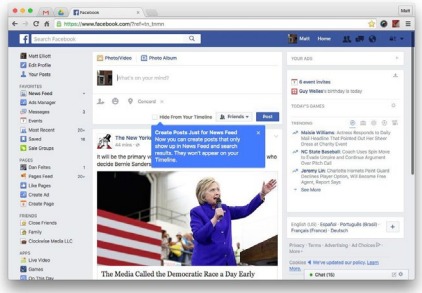
IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

Andrew Ng
@AndrewYNg
Follow

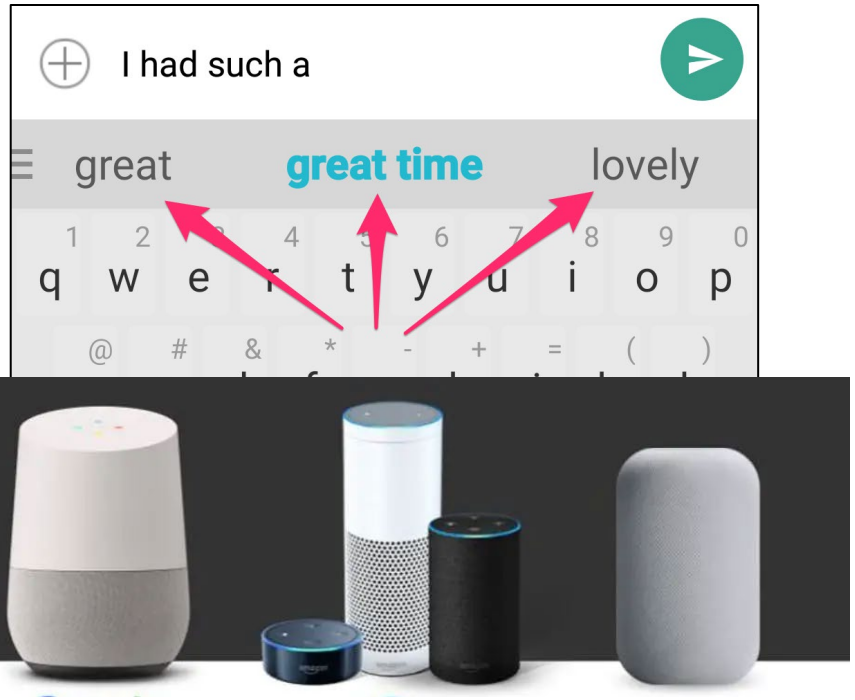
"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

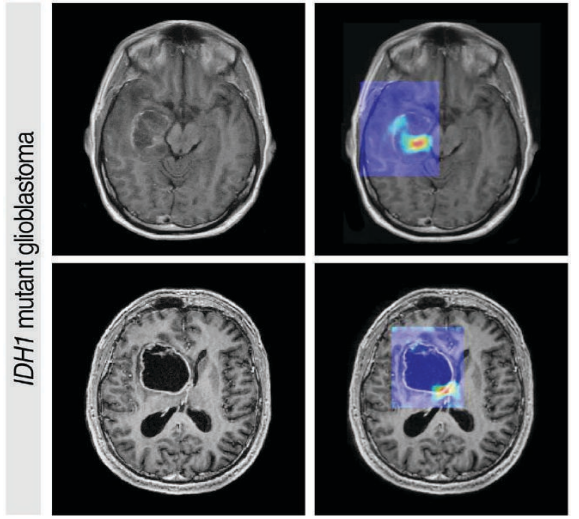
2016: The Year That Deep Learning Took Over t

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE



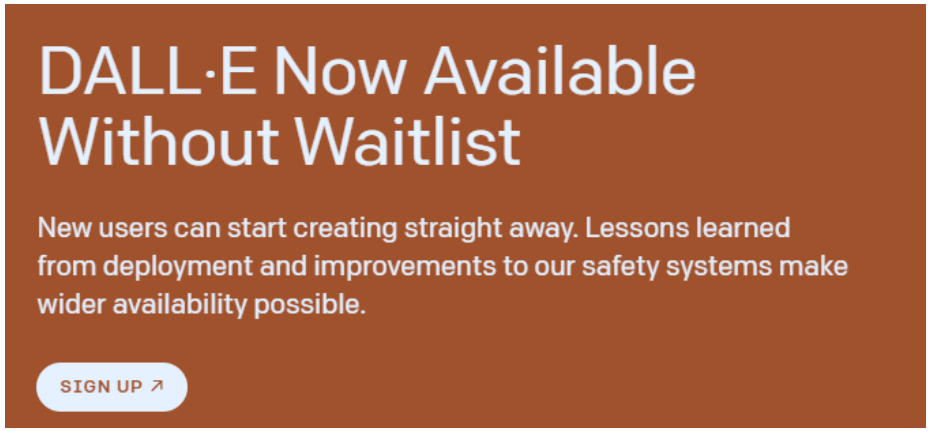
Models that learn from data are embedded in our lives

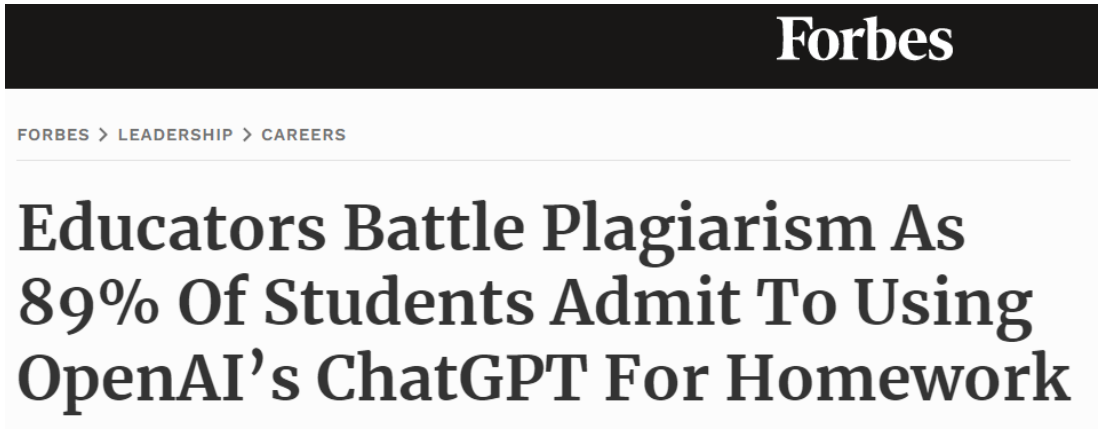


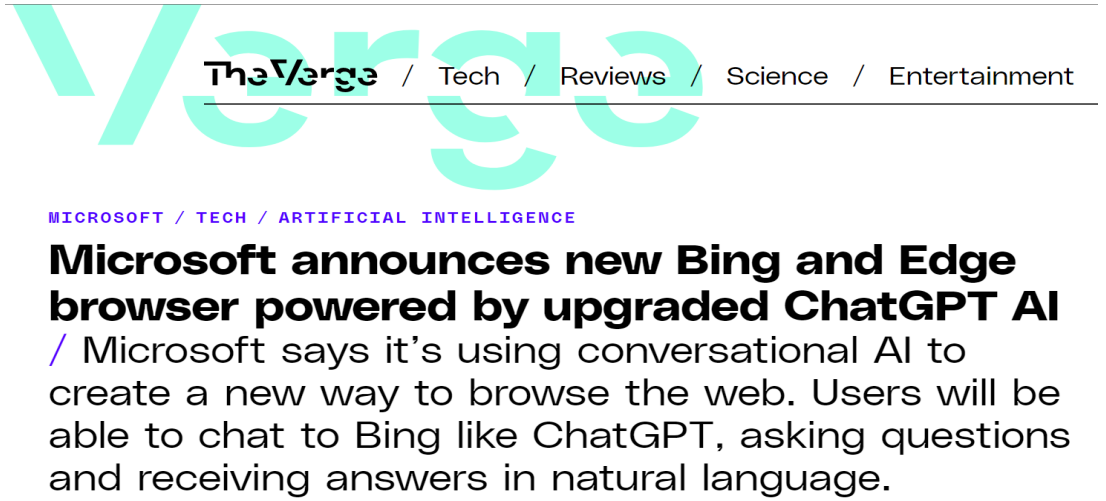


Models that learn from data are embedded in our lives

Recent advances have been rapidly adopted by common, non-expert users

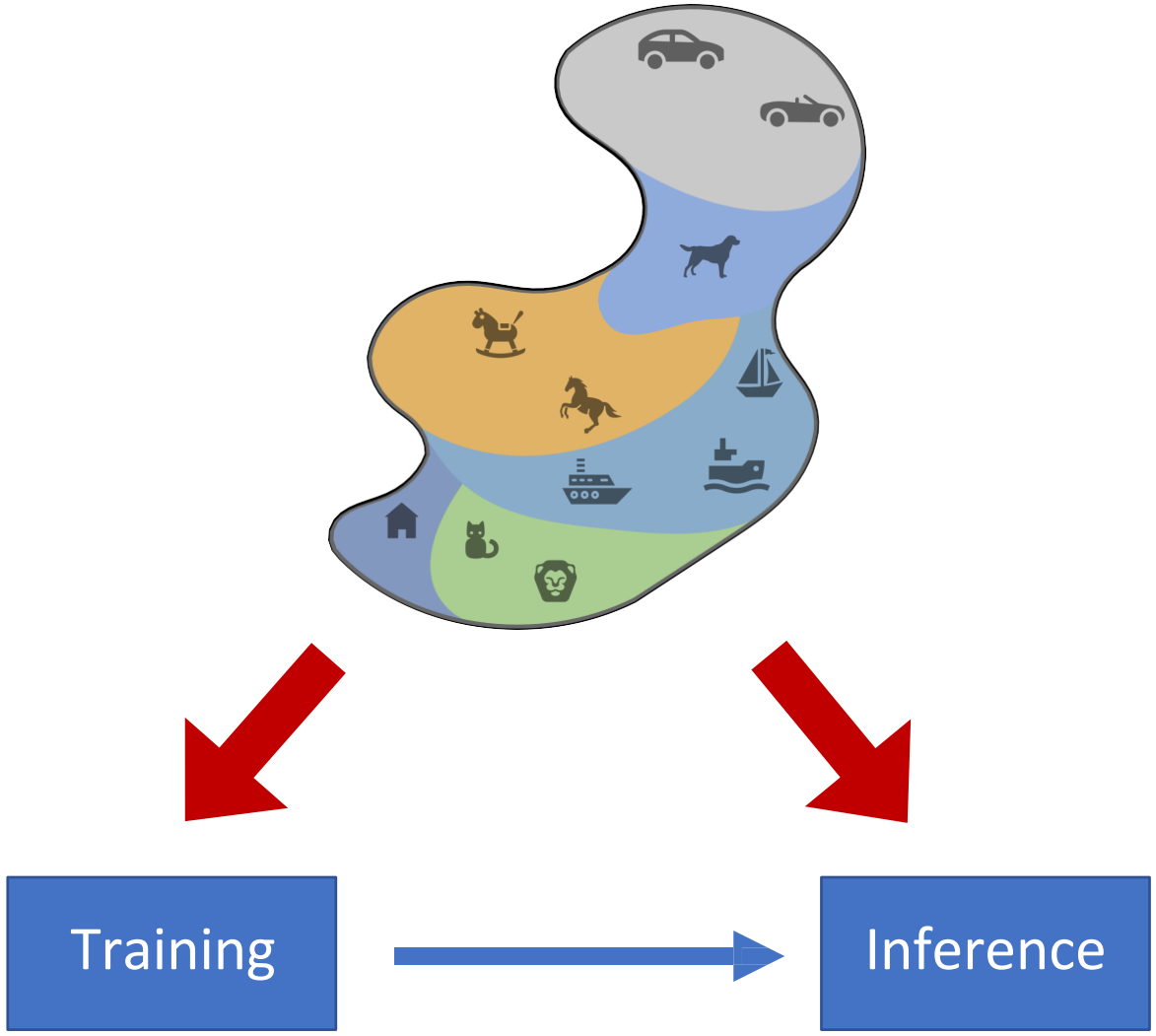






But …

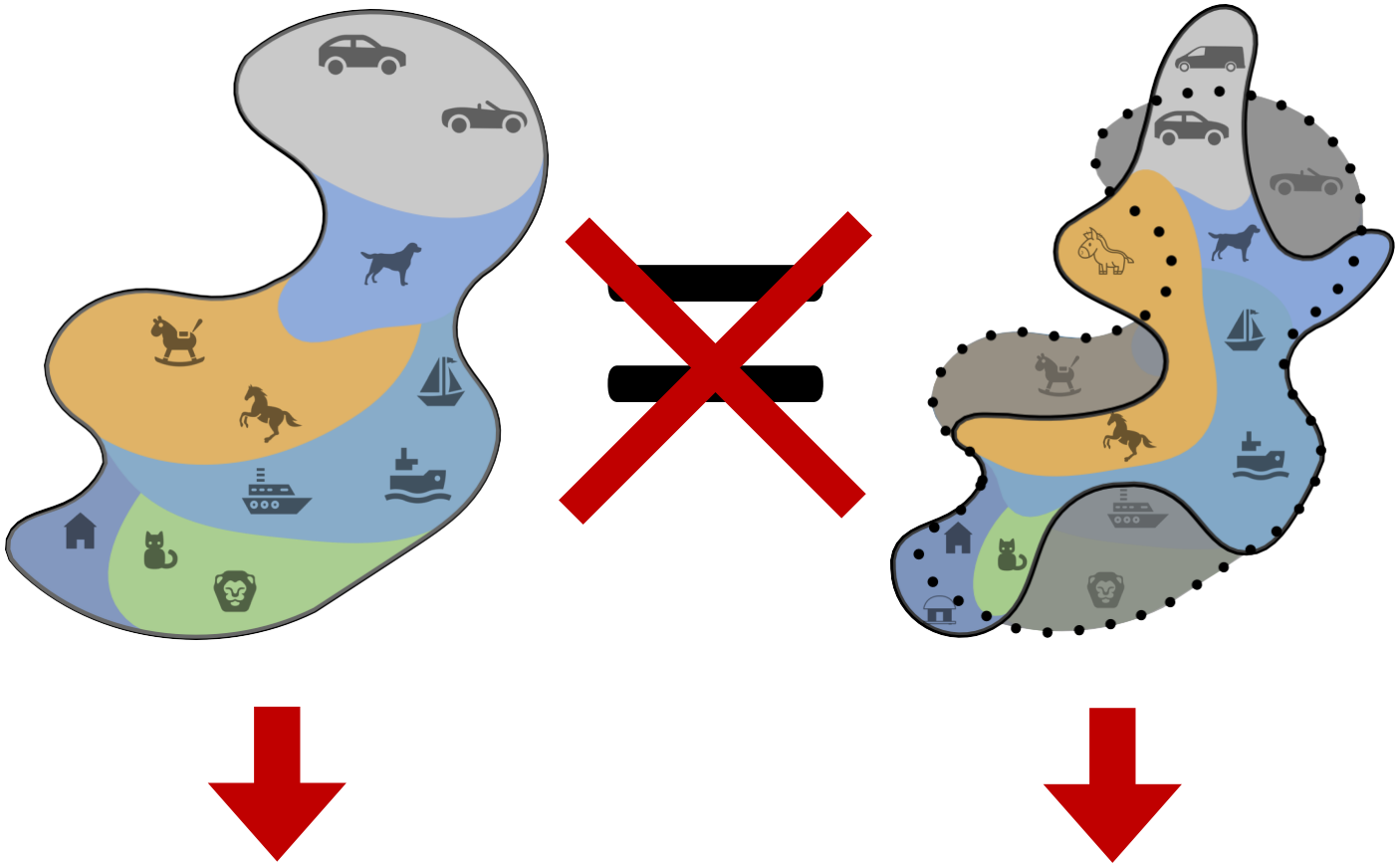
A Limitation of the (Supervised) ML Framework



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

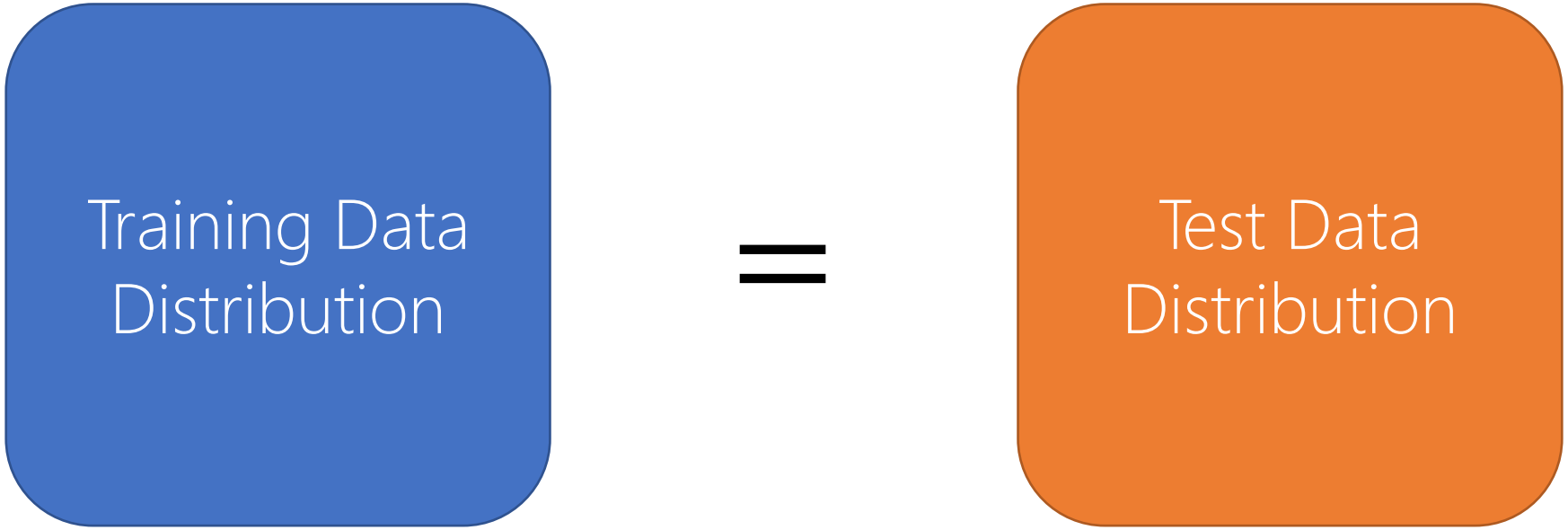


Measure of performance:
Fraction of mistakes during testing

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

What can go wrong?

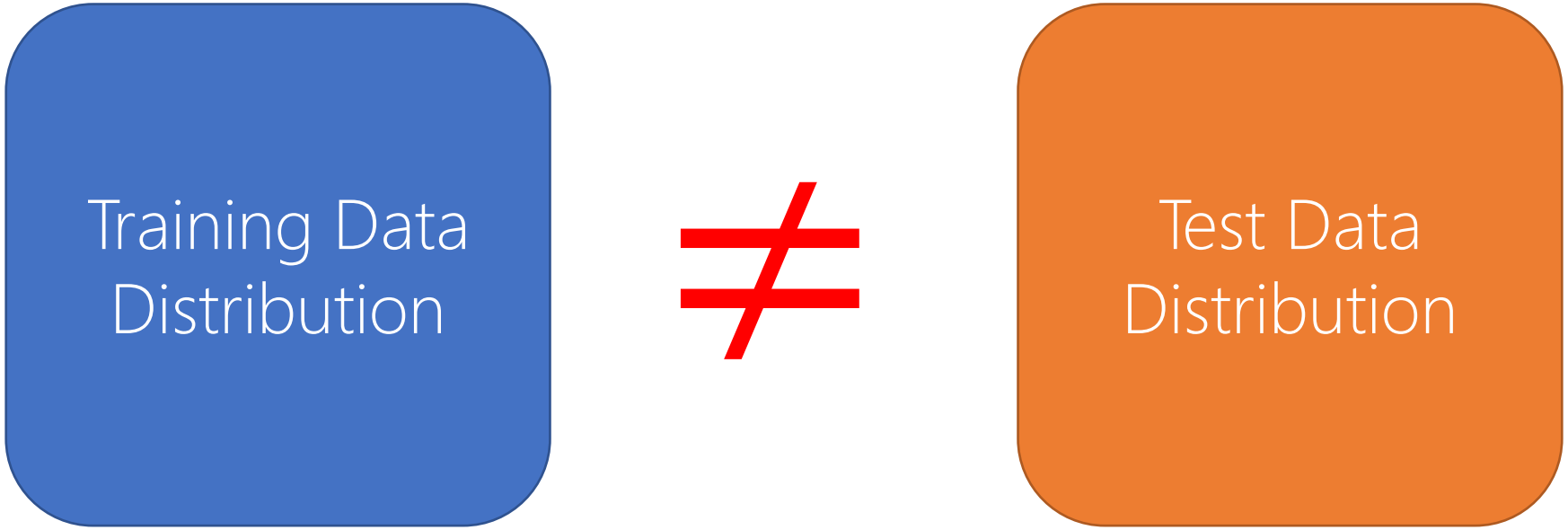
Standard *i.i.d.* Assumption in Machine Learning



"Independent and Identically Distributed"

Models learn useful patterns

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



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

Model performance deteriorates

Robustness and Generalization

Findings from Previous Work

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

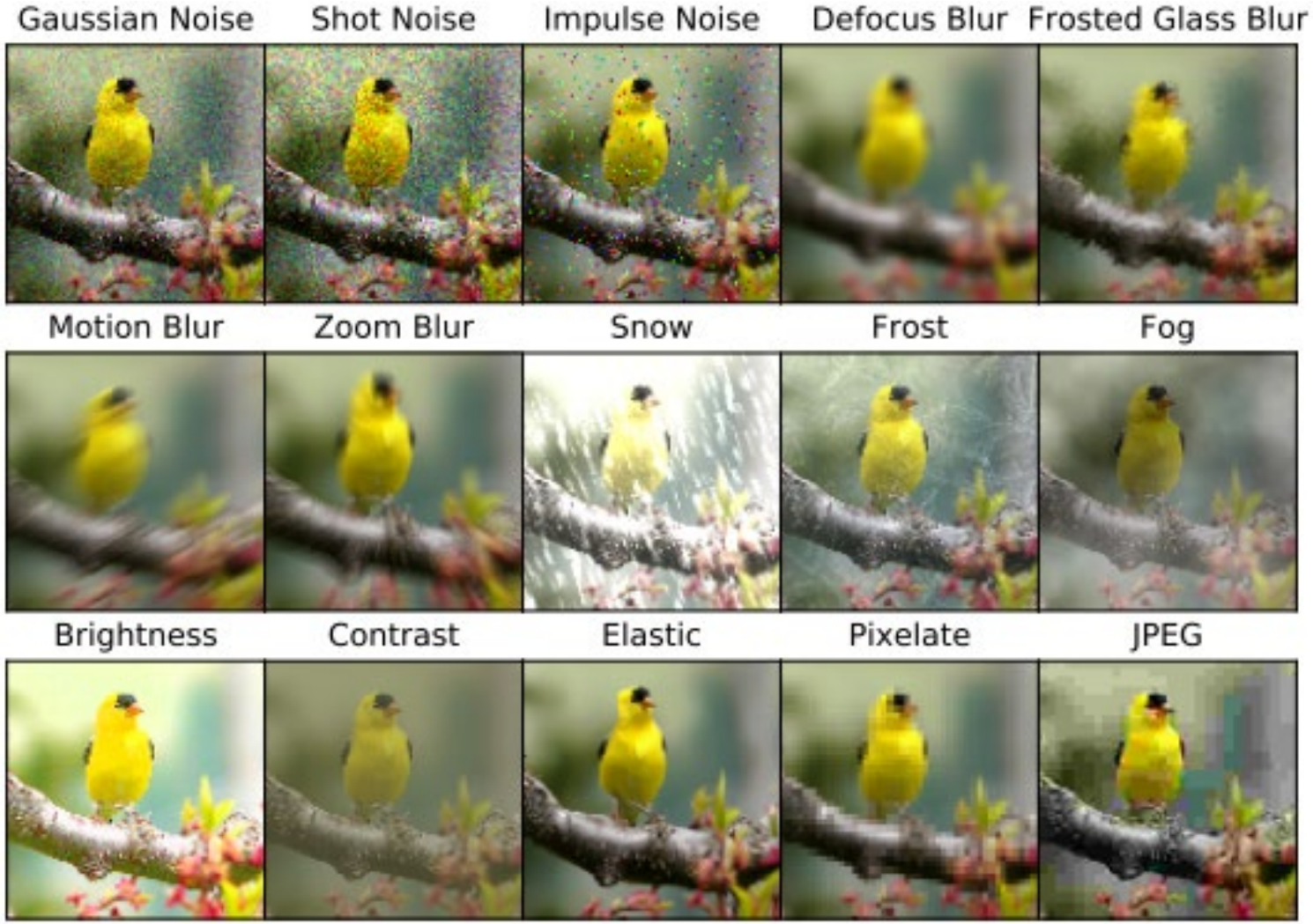
Poses can fool Image Classifiers

- **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
    - ...

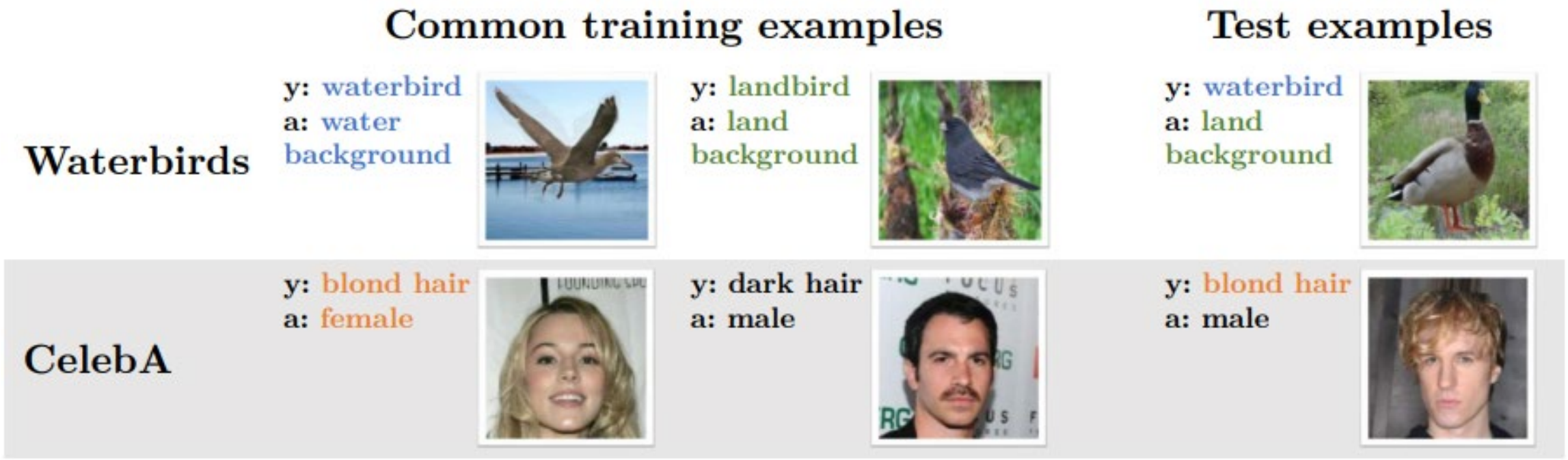
Alcorn et al. CVPR 2019

Natural Corruptions affect accuracy



Hendrycks et al. ICLR 2019

Spurious Correlations / Biased Datasets



Sagawa et al. ICLR 2020

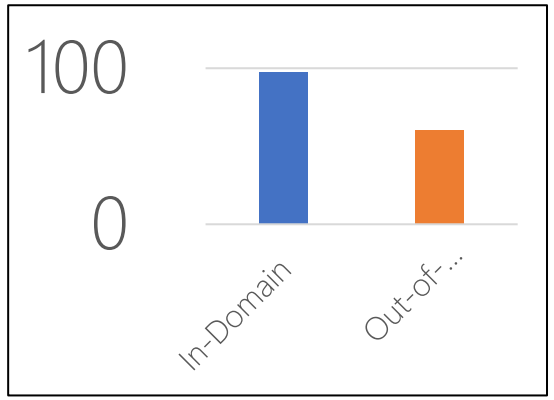
Lack of Diverse Data hurts Reliability

Mallard
Water Background

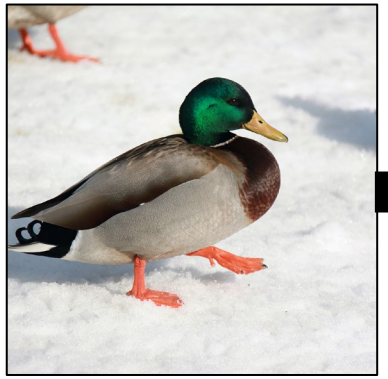
Penguin
Snow Background

Training

ML Model



Mallard
Snow Background



Test

ML Model

"Penguin"

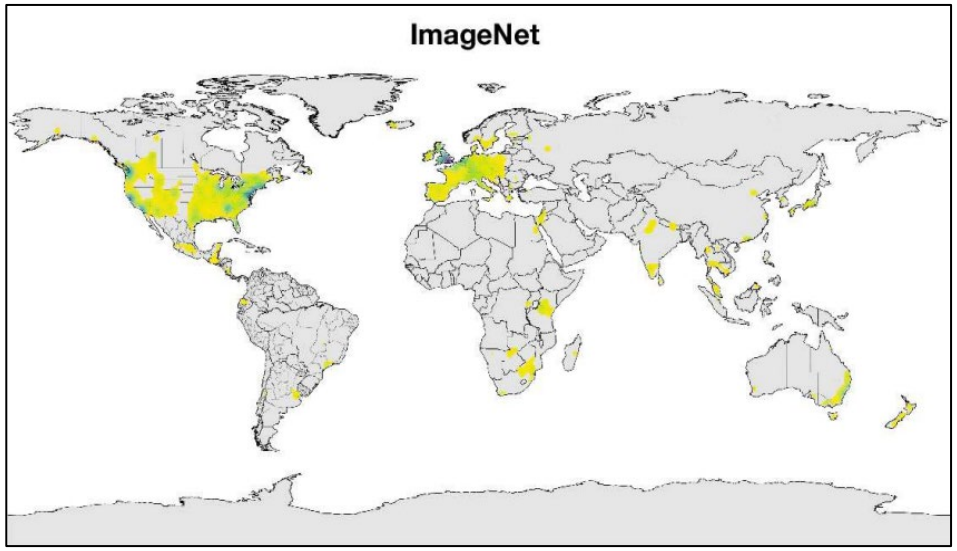
Lack of Diverse Data hurts Reliability

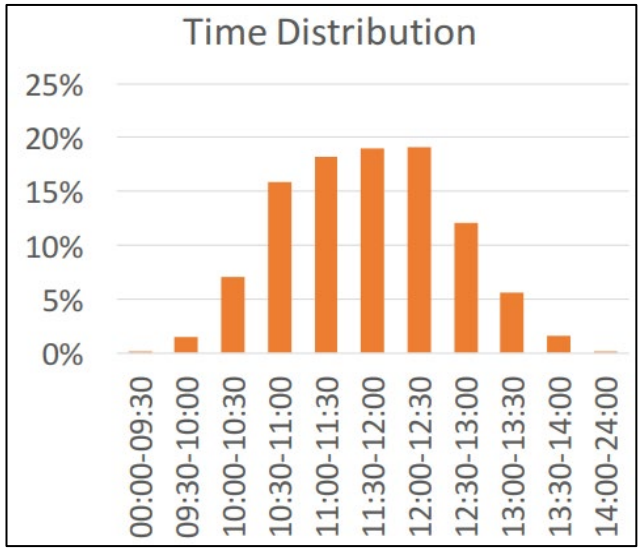
Mallard
Snow Background

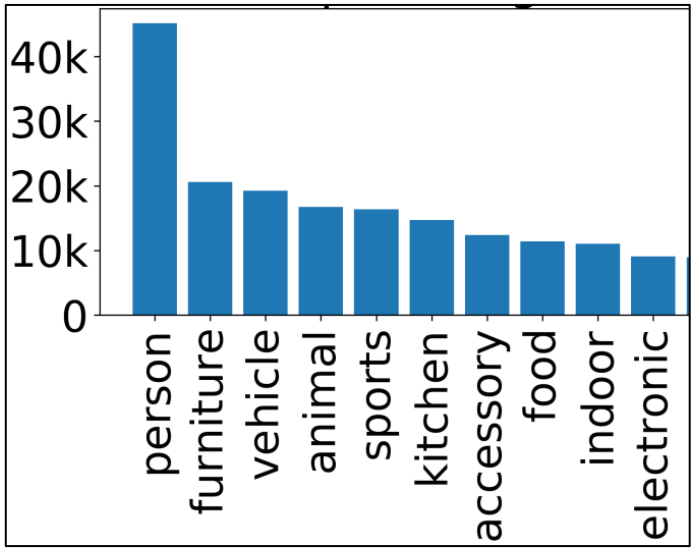
Test

ML Model

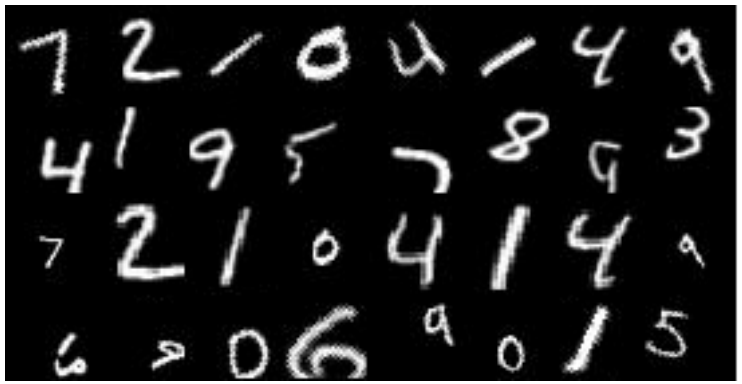
"Pengui n"

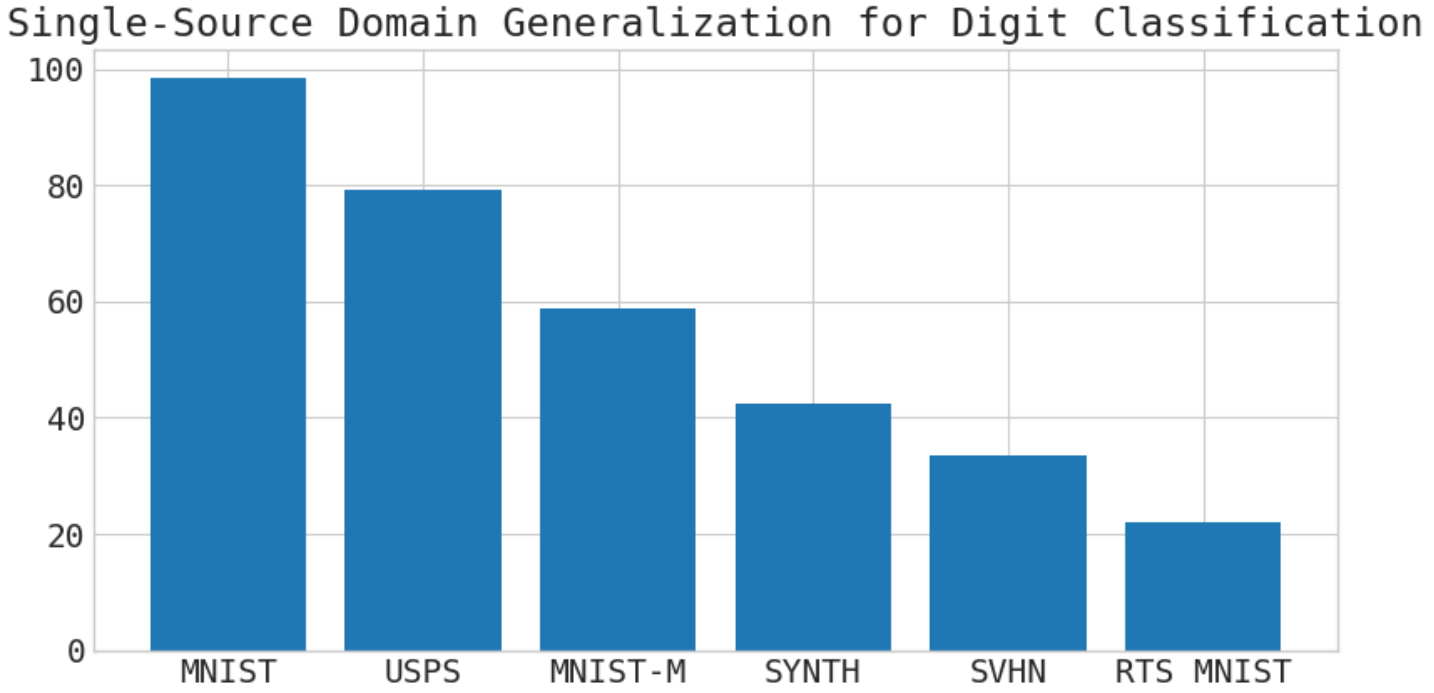






Domain Shift is a Nuisance



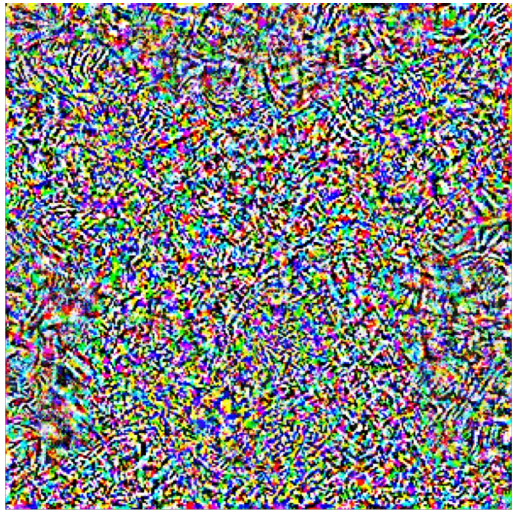


ML Predictions Are (Mostly) Accurate but Brittle

“pig” (91%)

+ 0.005 x

noise (NOT random)



=

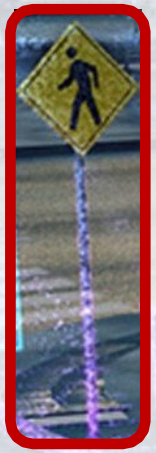
“airliner” (99%)

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

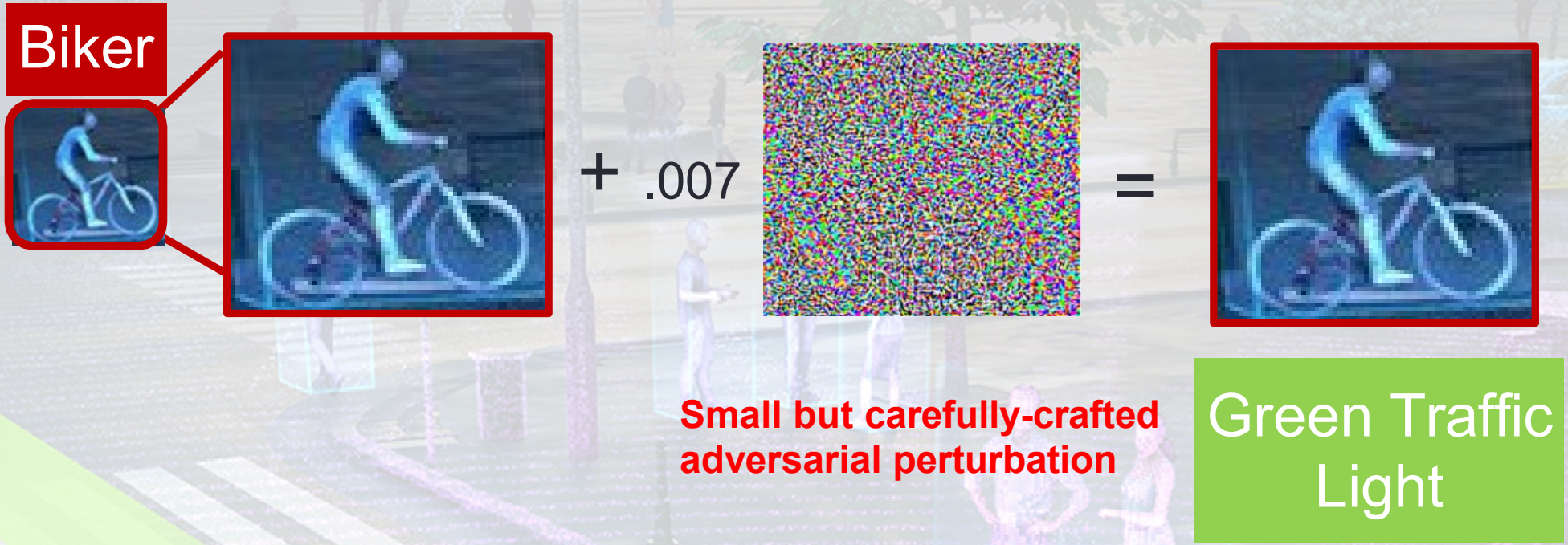
But also: [Dalvi Domingos Mausam Sanghai Verma 2004][Lowd Meek 2005]
[Globerson Roweis 2006][Kolcz Teo 2009][Barreno Nelson Rubinstein Joseph Tygar 2010]
[Biggio Fumera Roli 2010][Biggio Fumera Roli 2014][Srndic Laskov 2013]

Biker

Pedestrian Sign



Persons



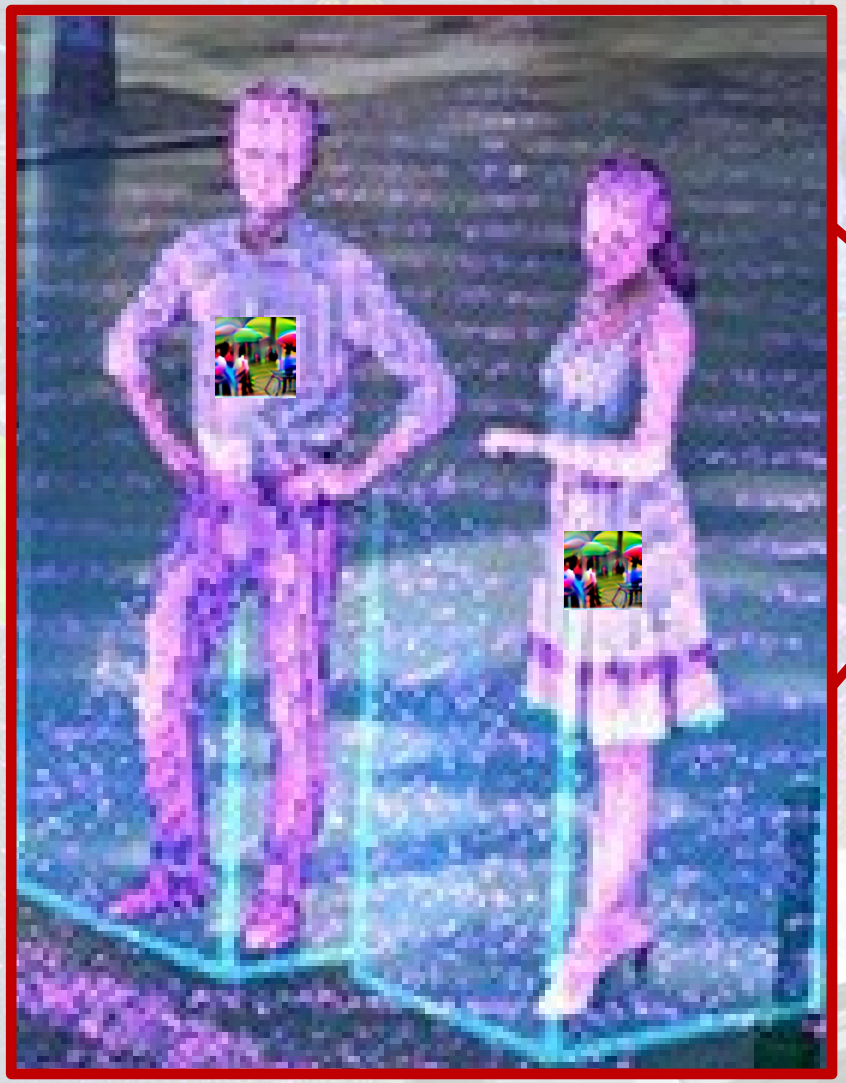
Adversarial Perturbation Attack

Pedestrian Sign

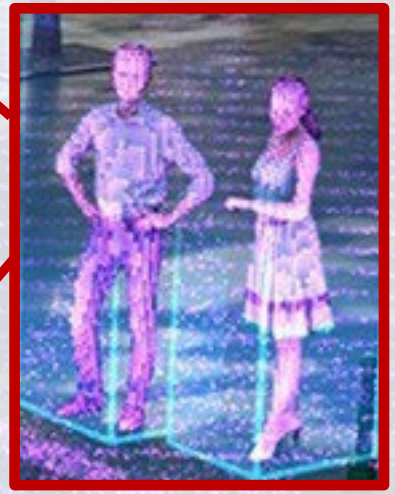
Speed Limit 45 Sign

Adversarial Rotation Attack

No Person



Persons



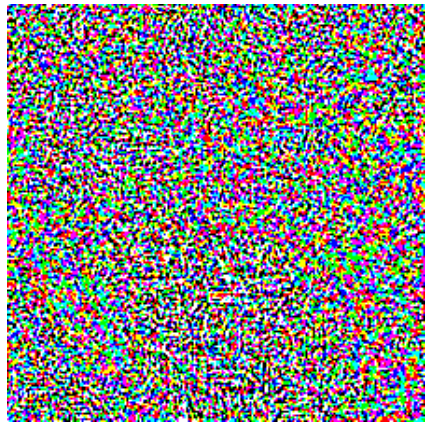
Adversarial Patch Attack

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”, which started the area of adversarial ML

Original image

Classified as panda
57.7% confidence



Small adversarial noise

Adversarial image

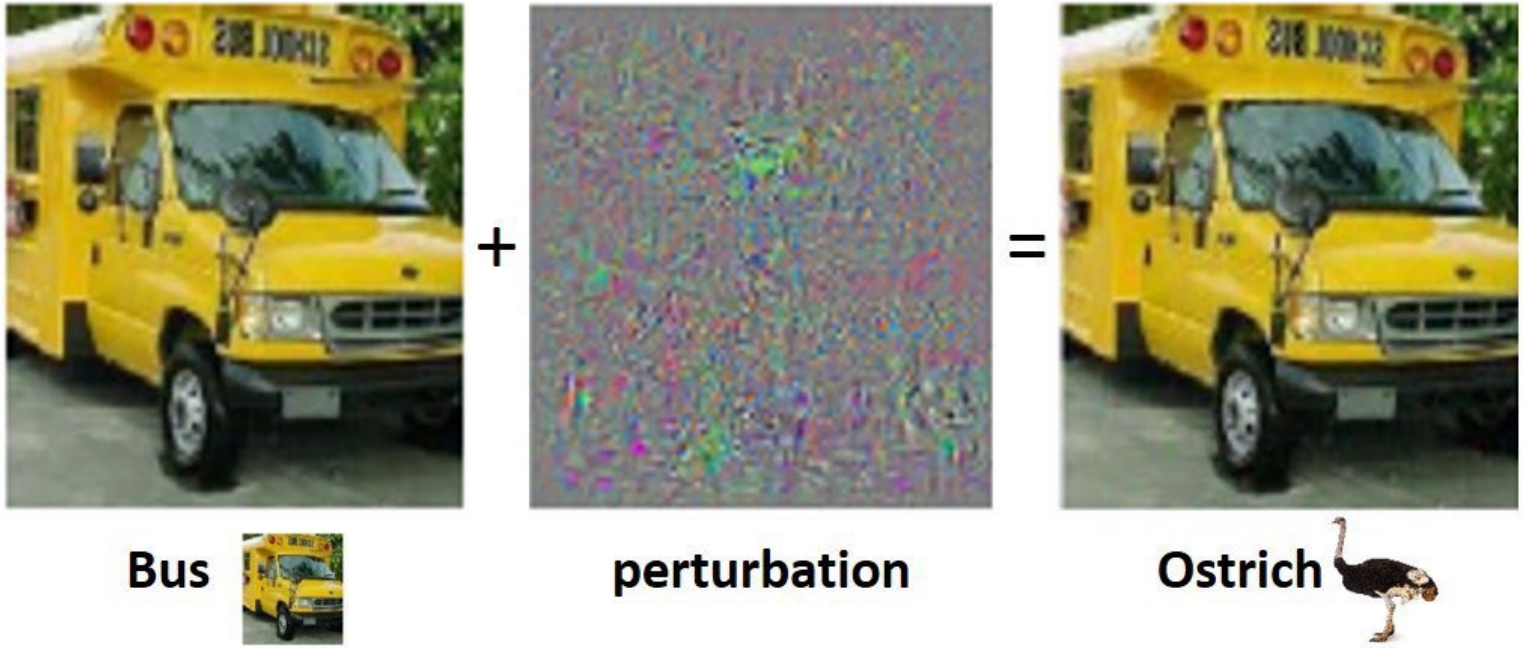
Classified as gibbon
99.3% confidence

Gibbon

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

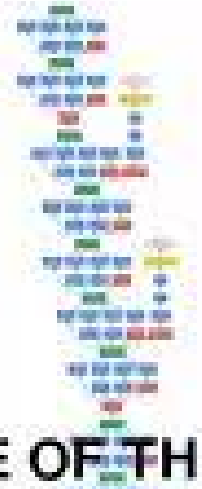
Adversarial Examples

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

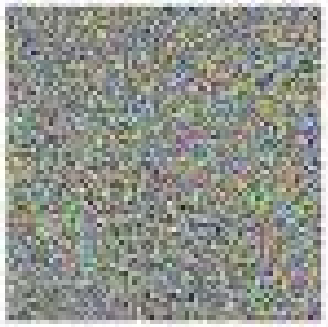


Picture from: Szagedy et al. (2014) – Intriguing Properties of Neural Networks

WHO WOULD WIN?



STATE OF THE ART NEURAL NETWORK

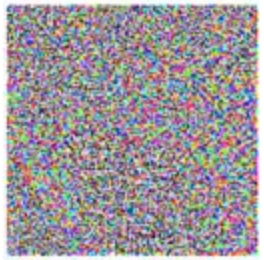


ONE NOISY BOI

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

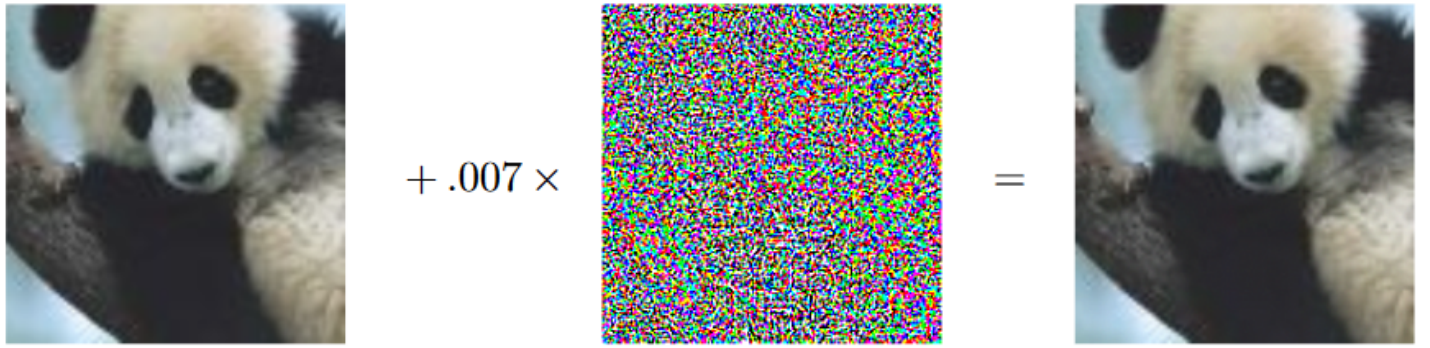
"gibbon"
99.3 % confidence

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. $\boldsymbol{x} + \boldsymbol{\epsilon g}(\boldsymbol{x})$ fools classifier

Perturbations are typically norm-bounded



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Goodfellow et al. ICLR 2015

Adversarial Training

Leverages the concept of adversarial examples, in order to improve classifier robustness to such attacks

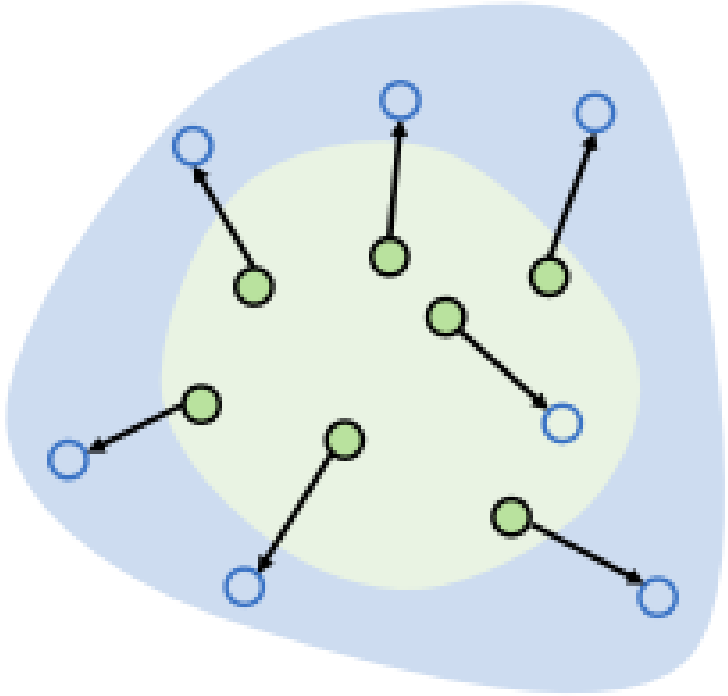
min—max optimization

maximization: find adversarial images

minimization: train classifier to correctly classify such images

norm-bounded perturbations

==> robustness within the norm-ball

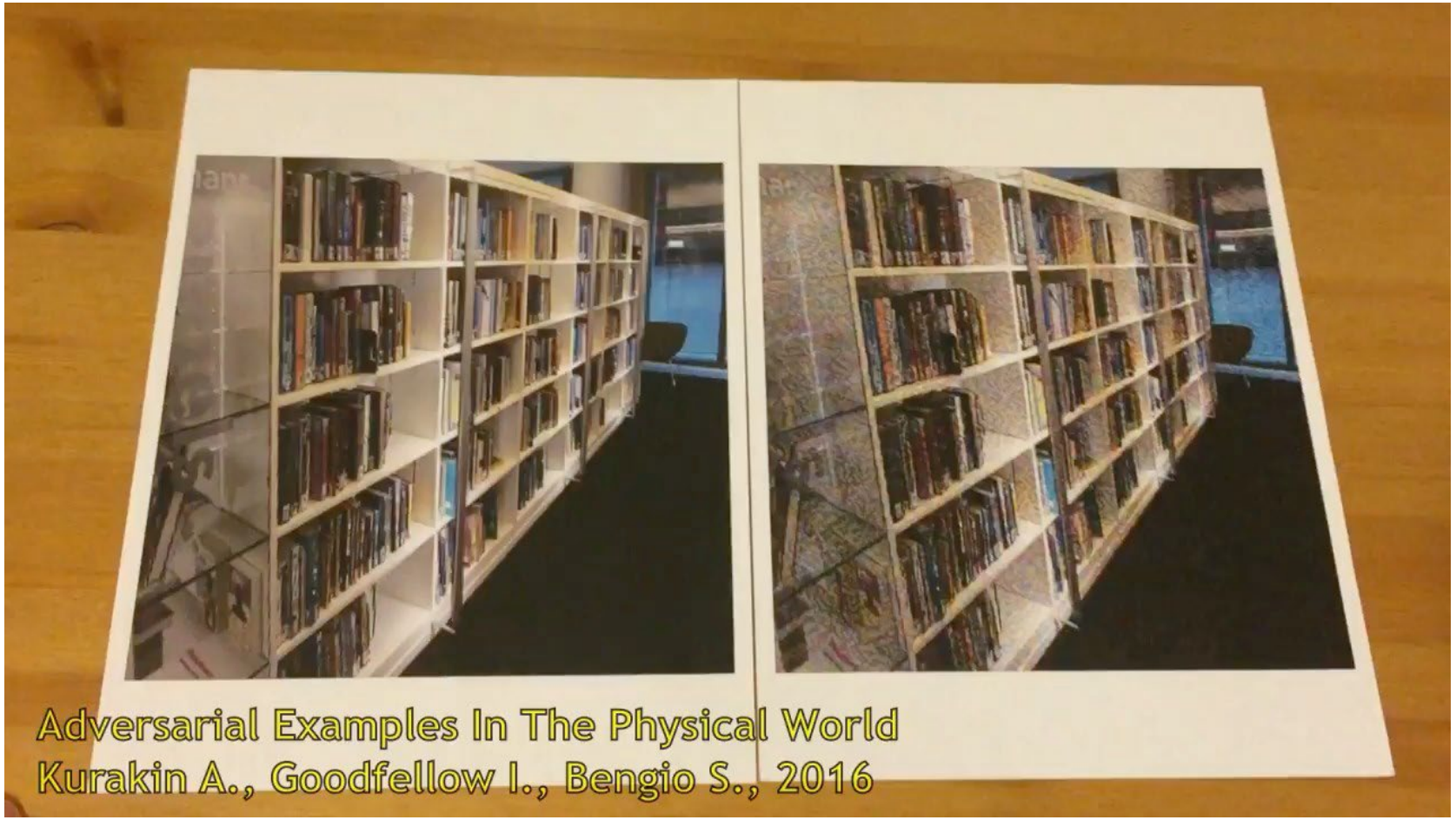


$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

Madry et al. ICLR 2018

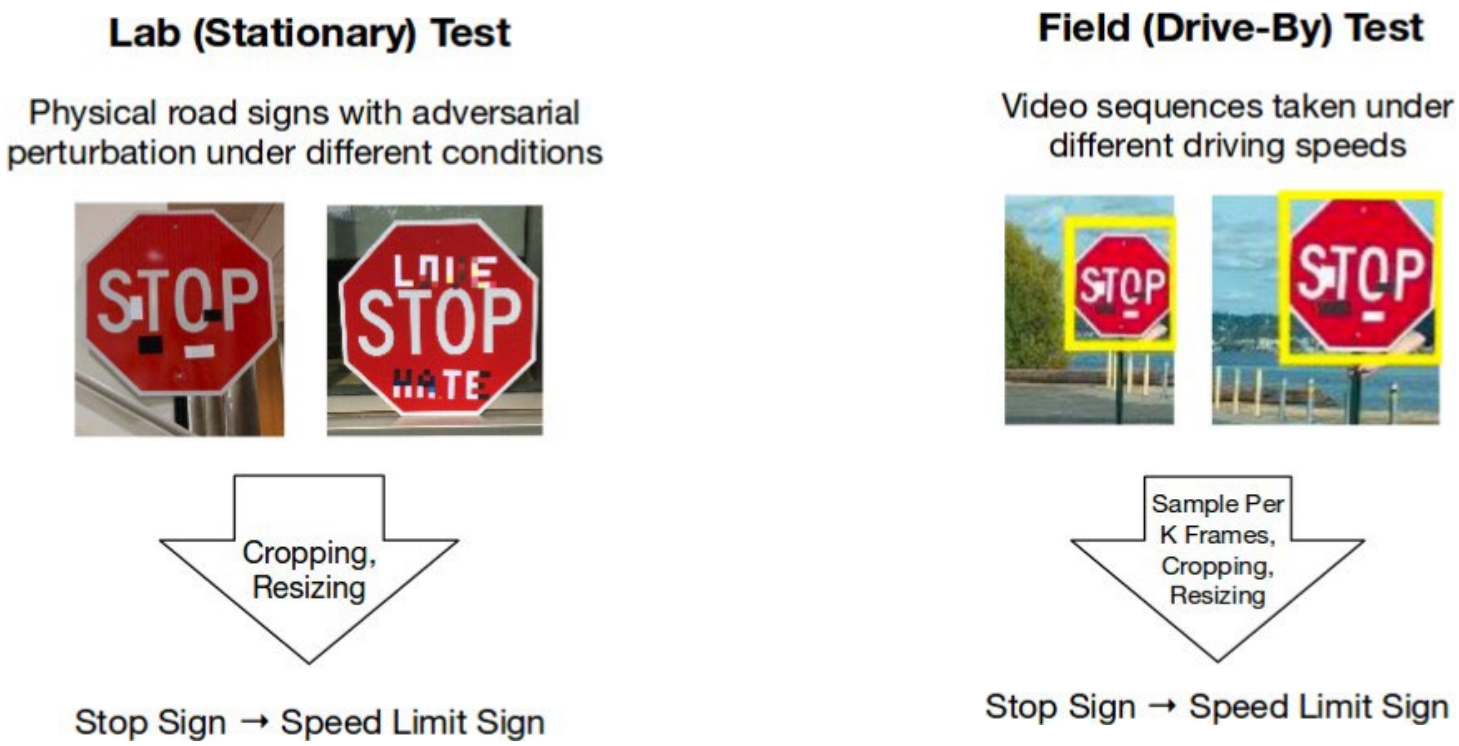
Physical-World Attack: Printed Adversarial Images

- 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

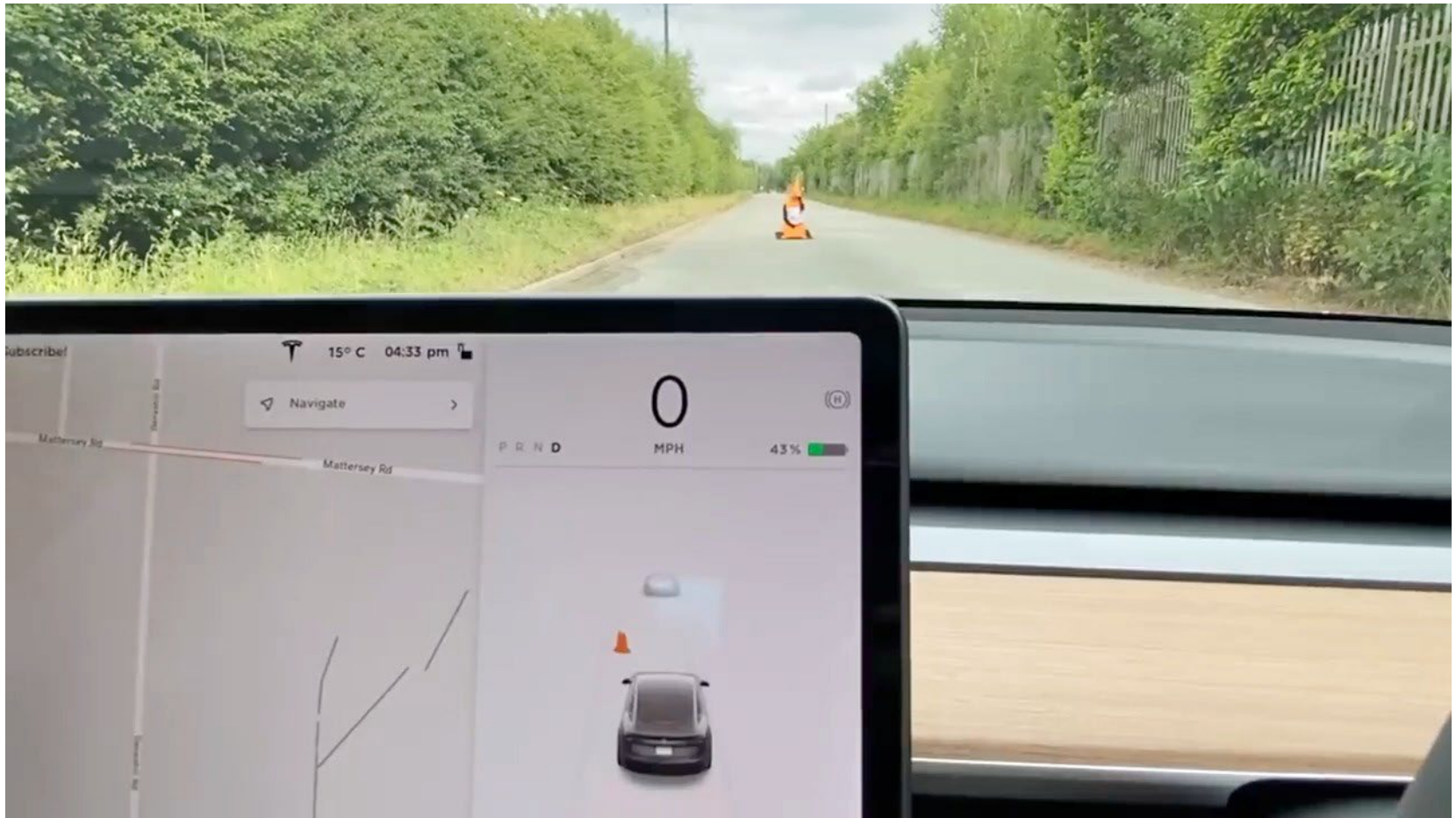
Physical-World Attack: Adversarial Patch

- 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 System

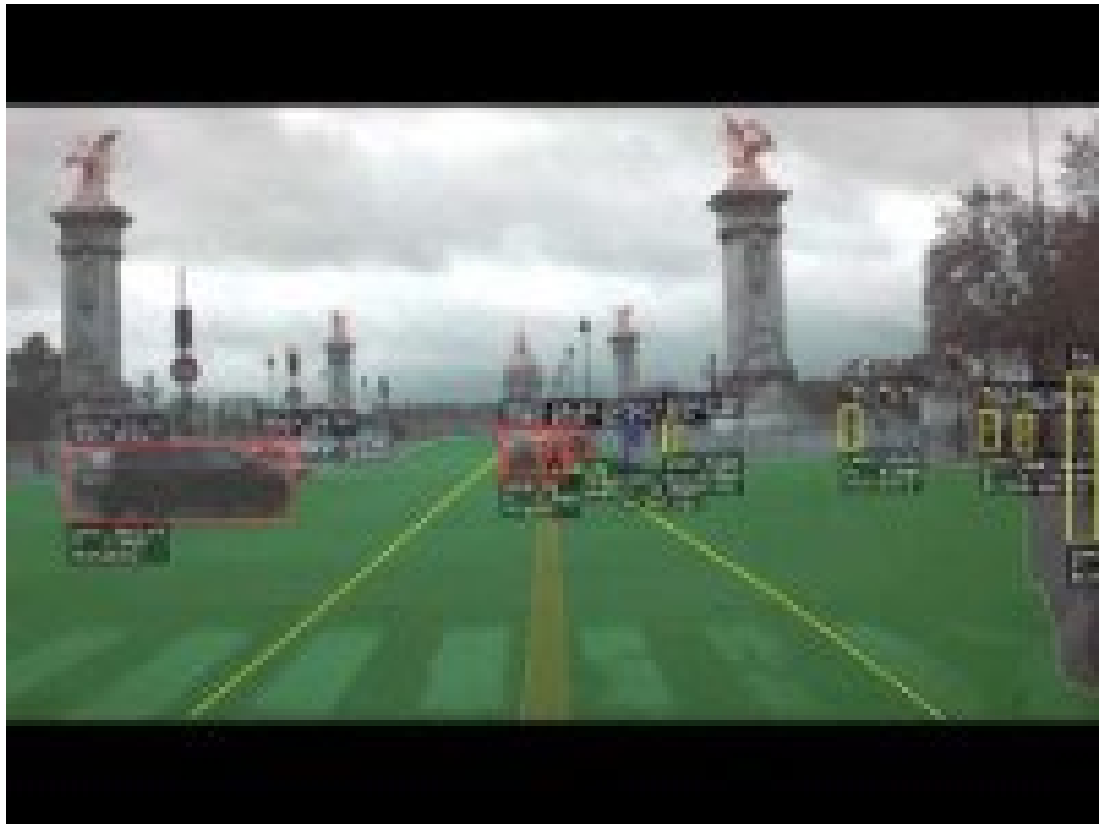
- Non-scientific example: a Tesla owner checks if the car can distinguish a person wearing a cover-up from a traffic cone



Why should we care?

→ People suffer consequences because of use in real-world systems

→ Safety, security, trust in the systems that we engineer



Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

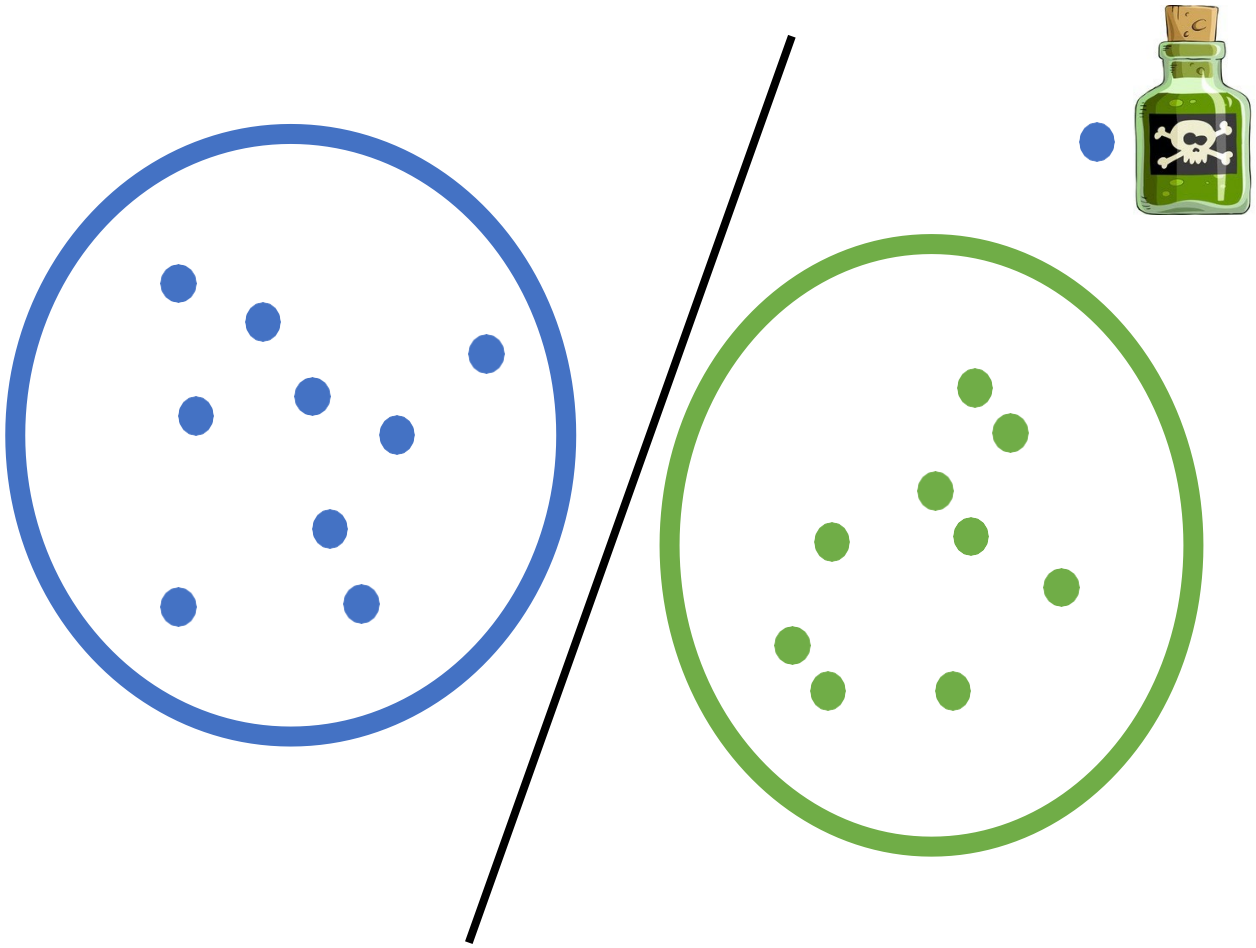
npr

TECHNOLOGY

Feds Say Self-Driving Uber SUV Did Not Recognize Jaywalking Pedestrian In Fatal Crash

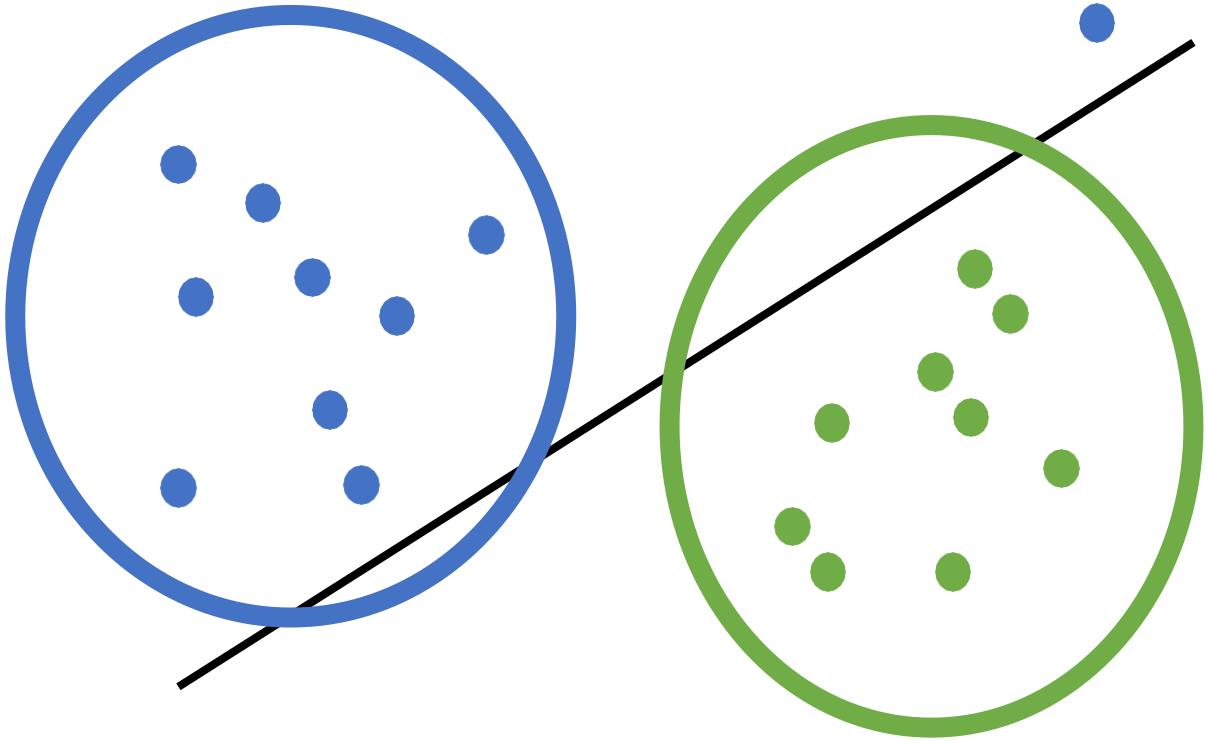
Data Poisoning

Goal: Maintain training accuracy but hamper generalization



Data Poisoning

Goal: Maintain training accuracy but hamper generalization



→ Fundamental problem
in “classic” ML (robust statistics)

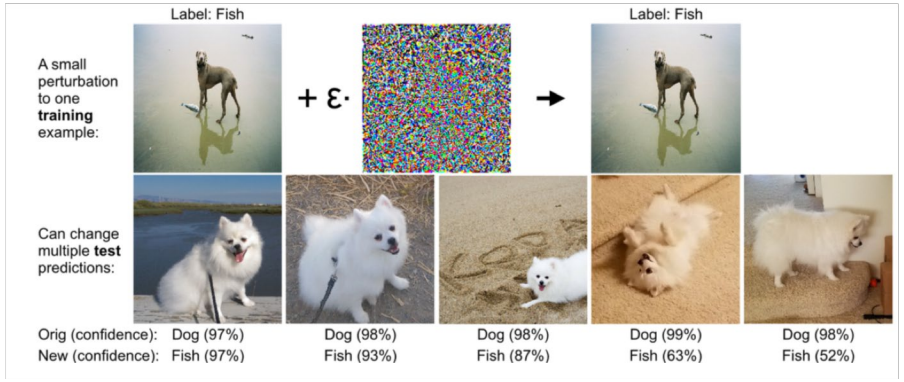
→ **But:** seems less so in deep learning

→ **Reason:** Memorization?

Data Poisoning

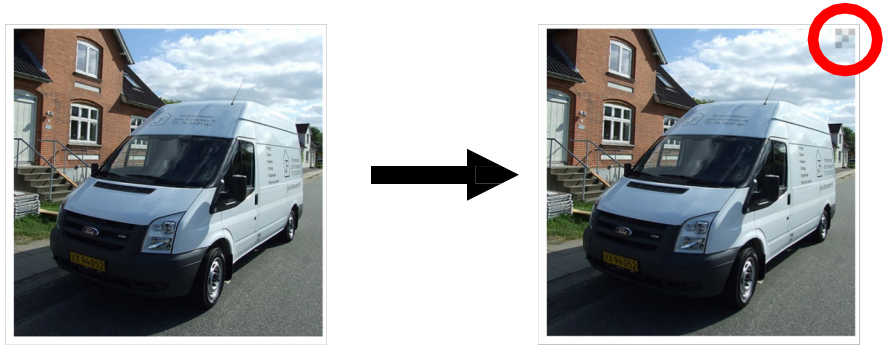
classification of **specific** inputs

Goal: Maintain training accuracy but hamper ~~generalization~~



[Koh Liang 2017]: Can manipulate **many** predictions with a **single** "poisoned" input

But: This gets (much) worse



"van" "dog"

[Gu Dolan-Gavitt Garg 2017][Turner Tsipras M 2018]: Can plant an **undetectable backdoor** that gives an almost **total** control over the model

(**To learn more about backdoor attacks:** See poster #148 on Wed **[Tran Li M 2018]**)


THIS IS FINE.

We look at robustness math and methods
in detail in CMSC 475/675 Neural Networks …

I've also taught a seminar class on "Robust ML"
Slides: https://courses.cs.umbc.edu/graduate/691rml/