tejasgokhale.com

Lecture 1: Introduction

CMSC 491/691 Robust Machine Learning

Lecture 1: Introduction

CMSC 491/691 Robust Machine Learning

UMBC

Welcome to the class!

This is an interactive class.

Don't just consume – participate!

This is a research class.

Don't just "learn" – critique, ideate, design, evaluate!

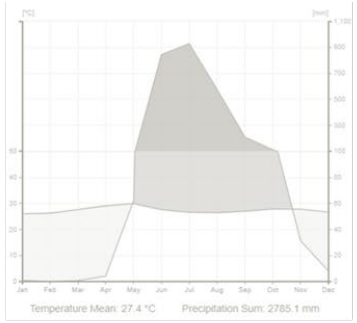
tejasgokhale.com

Tejas Gokhale

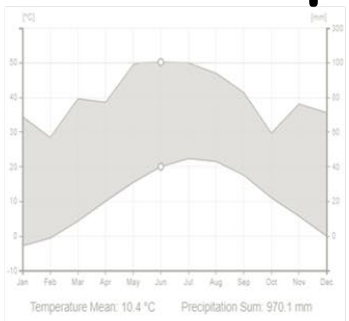
~ TAY + JUSS

~ GO + CLAY

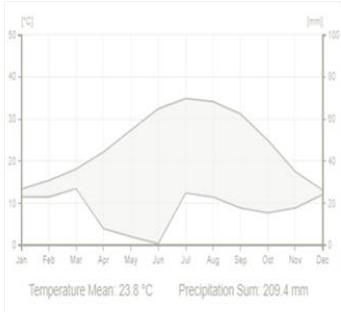
Assistant Professor
Computer Science
University of Maryland, Baltimore County



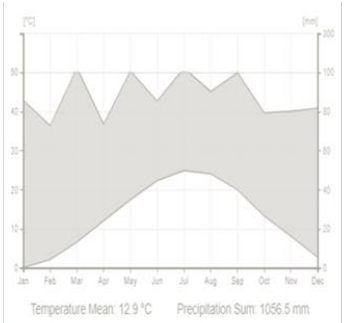
2011—2015
B.E. (Honours)
BITS Pilani
(Goa)



2016-2018:
M.S.
Carnegie Mellon
University



2018—2023:
Ph.D.
Arizona State
University



2023—present
Assistant Professor
University of Maryland
Baltimore County

UMBC

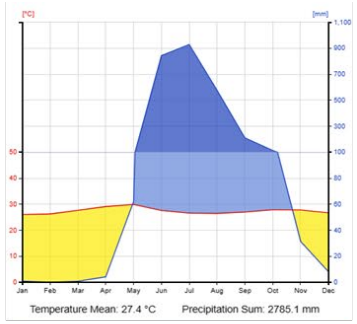
tejasgokhale.com

Tejas Gokhale

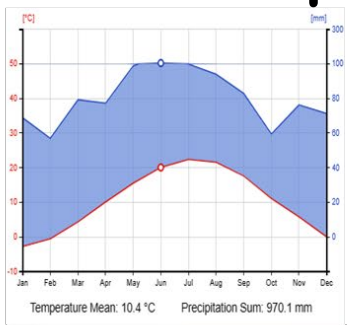
~ TAY + JUSS

~ GO + CLAY

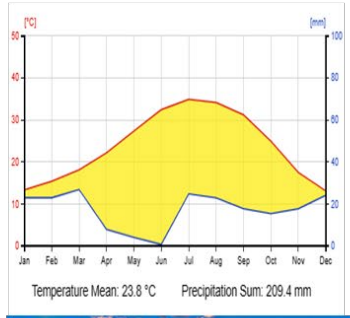
Assistant Professor
Computer Science
University of Maryland, Baltimore County



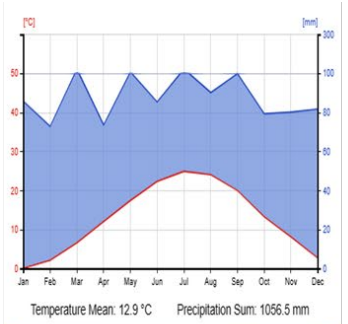
2011—2015
B.E. (Honours)
BITS Pilani
(Goa)



2016-2018:
M.S.
Carnegie Mellon
University



2018—2023:
Ph.D.
Arizona State
University



2023—present
Assistant Professor
University of Maryland
Baltimore County

UMBC

Tejas Gokhale, Assistant Professor

Cognitive Vision Group, ITE 368

Research Areas

- Computer Vision
- Vision & Language
- Image Generation
- Robustness & Reliability
- Active Perception

Current Projects

- Domain Adaptation/Generalization
- Quantifying Visual Typicality
- Visual Semantics and Pragmatics
- Continual Concept Learning
- Generative Model Evaluation
- Multimodal Active Inference

Highlights / Activities

- AAAI 2024: New Faculty Highlights Invited Talk
- ECCV'24 & WACV'24: Tutorials on Reliability of Generative Models
- Area Chair: NeurIPS, ACL ARR, EMNLP, NAACL, WACV
- UMBC PPR Seminar




www.tejasgokhale.com

UMBC

Course Staff

Instructor: Tejas Gokhale

Assistant Professor, CSEE

OH:

Wed 2:30 – 3:30 PM ITE 214

gokhale@umbc.edu

TA/Grader ...

Course Website

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

Quick Introductions

Course Website

(1) Name

(2) Major (e.g. CS)

(3) Level (BS/MS/PhD)

(4) Why are you taking this class?

(5) Something you did this summer

Class Structure: Overview

- [Wed 08/28]: Today, Class Logistics and Introduction
- [Mon 09/02]: Labor Day, NO CLASS
- [Wed 09/04]: Machine Learning Review
- Every Week after that:
  - MON: [TEJAS] Overview of a Robustness Challenge
  - WED: [STUDENTS] Quiz; Paper Presentations; Class Discussion

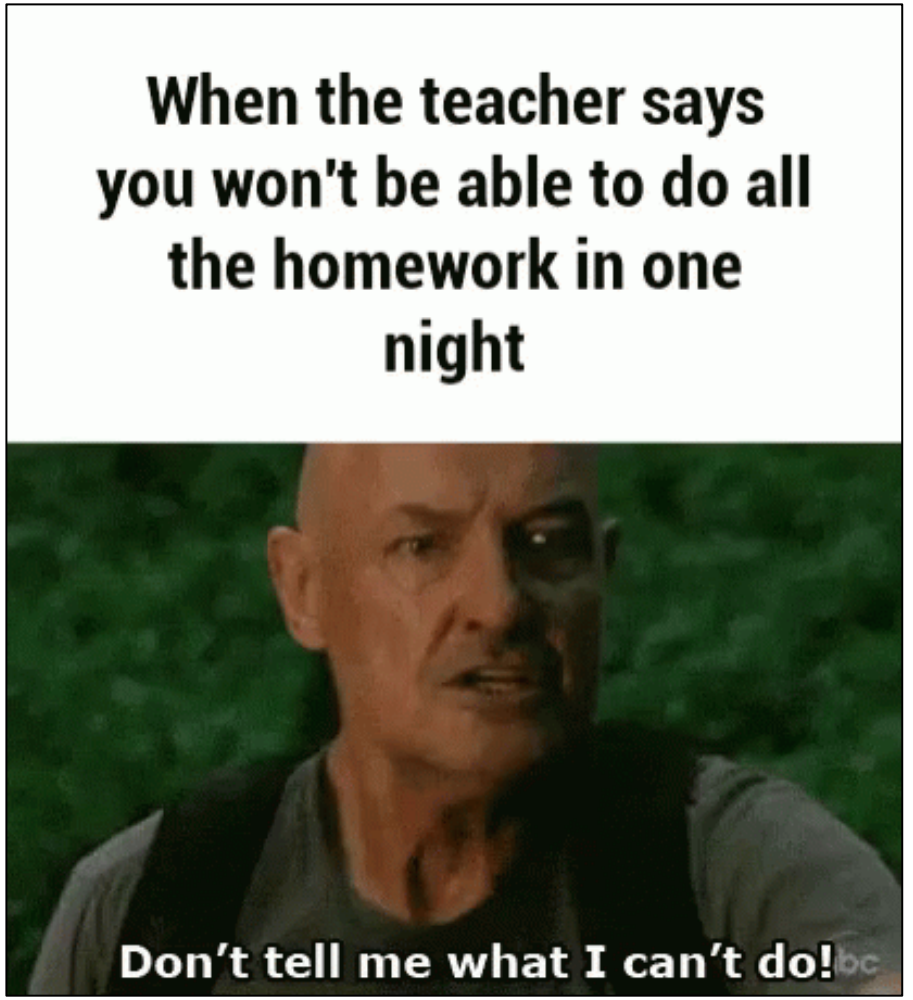
We have 12 Robustness Topics

Class Structure: Tentative Schedule

We have 12 Robustness Topics

	A	B	C	D	E	F	G	H	I
1	Date	Day	Topic	Notes		Date	Day	Topic	Notes
2	26-Aug	M				28-Aug	W	Intro	
3	2-Sep	M	LABOR DAY			4-Sep	W	ML Refresher	
4	9-Sep	M	Domain Adaptation			11-Sep	W	Presentations	
5	16-Sep	M	Domain Generalization			18-Sep	W	Presentations	
6	23-Sep	M	OOD Detection			25-Sep	W	Presentations	
7	30-Sep	M	Adversarial Attacks, Backdoor Attacks			2-Oct	W	Presentations	
8	7-Oct	M	Uncertainty and Calibration			9-Oct	W	Presentations	
9	14-Oct	M	Online/Continual Learning			16-Oct	W	Presentations	
10	21-Oct	M	Self-Supervised/Unsupervised Learning			23-Oct	W	Presentations	
11	28-Oct	M	Test-Time Learning, Adaptation			30-Oct	W	Presentations	
12	4-Nov	M	Machine Unlearning, Model Editing			6-Nov	W	Presentations	
13	11-Nov	M	Interpretability, Explanability, Compositionality			13-Nov	W	Presentations	
14	18-Nov	M	ML Evaluation: Metrics and Benchmarks			20-Nov	W	Presentations	
15	25-Nov	M	Robustness Tradeoffs			27-Nov	W	no class (but take home assignment?)	
16	2-Dec	M	Project Presentations			4-Dec	W	Project Presentations	
17	9-Dec	M	Invited Talk / "The Last Lecture"						

Class Structure: Grading



Class Structure: Grading

We have 12 Robustness Topics

- Paper Presentations 20% pick two topics (10% each)
- Survey Papers 20% pick two topics (≠ your presentation topic)
- Quizzes (each Wed) 25% best 10 out of 12 (2.5% each)
- Project (group of 3)* 35%
  - Proposal (5%), Midterm Update Video (5%), Final Presentation (10%), Final Report (15%)
  - ***If you're a PhD/MS thesis student, you can opt to work alone***
    - ***But you need my APPROVAL!***
- ***Extra Credit*** ***max 10%*** (opportunities will be announced periodically)

Class Structure: Deadlines & Late Submission

- **Paper Presentations:** In class (sign up sheet will be shared)
  - If you are scheduled to present, but can't make it, send an email to me and the TA to request date-change
  - Permissible reasons: unforeseen events (illness, injury, emergency), travel to academic conferences, interviews
  - Bottom Line: I reserve the right to approve or deny date-change requests

- **Survey Papers:** Due "next Wednesday 2359 UMBC time". See example below.
  - If Lecture on "Domain Adaptation" is on Monday, Sept 09.
  - Then Survey paper on "Domain Adaptation" is due on Wednesday, Sept 18, 2359 UMBC time.
  - Late submissions: 10% deducted for each late day.

- **Quizzes:** In class. 12 quizzes. Your highest 10 grades will be chosen.
  - Bottom Line: if you miss a quiz, you miss a quiz.

- **Project:** Each milestone has fixed deadline.
  - Late Submissions: 10% deducted for each late day FOR ALL GROUP MEMBERS

Academic Integrity

I take academic integrity very seriously. You should too.
If you're unsure about something, ask us.
You are at a top-tier (R1) research university.
Use this privilege to learn. A good grade will follow.
Don't throw away this opportunity by cheating.

- Presentations, Survey Papers, Quizzes must be done independently.
- Do not plagiarize. Consequences will not be pleasant.
- Do not use "AI" assistants for any part of any assignment. Consequences will not be pleasant.
- Familiarize yourself with UMBC policy on plagiarism and other forms of cheating: https://academicconduct.umbc.edu/resources-for-students/
- Read the syllabus for consequences of academic integrity violations.

Academic Integrity

I take academic integrity very seriously. You should too.
If you're unsure about something, ask us.
You are at a top-tier (R1) research university.
Use this privilege to learn. A good grade will follow.
Don't throw away this opportunity by cheating.

3.6 Good Practices

If the integrity of your work in this course is challenged, you are responsible for demonstrating proof that the work submitted is your own. A good starting point is to enable versioning/tracking in Google Docs, Word, Pages, or other software so that your writing activities/progress during the semester can be logged if necessary. Keeping copies of research notes, scribbles, and related material may be helpful, too.

3.7 Viva or Oral Defense of Flagged Submissions

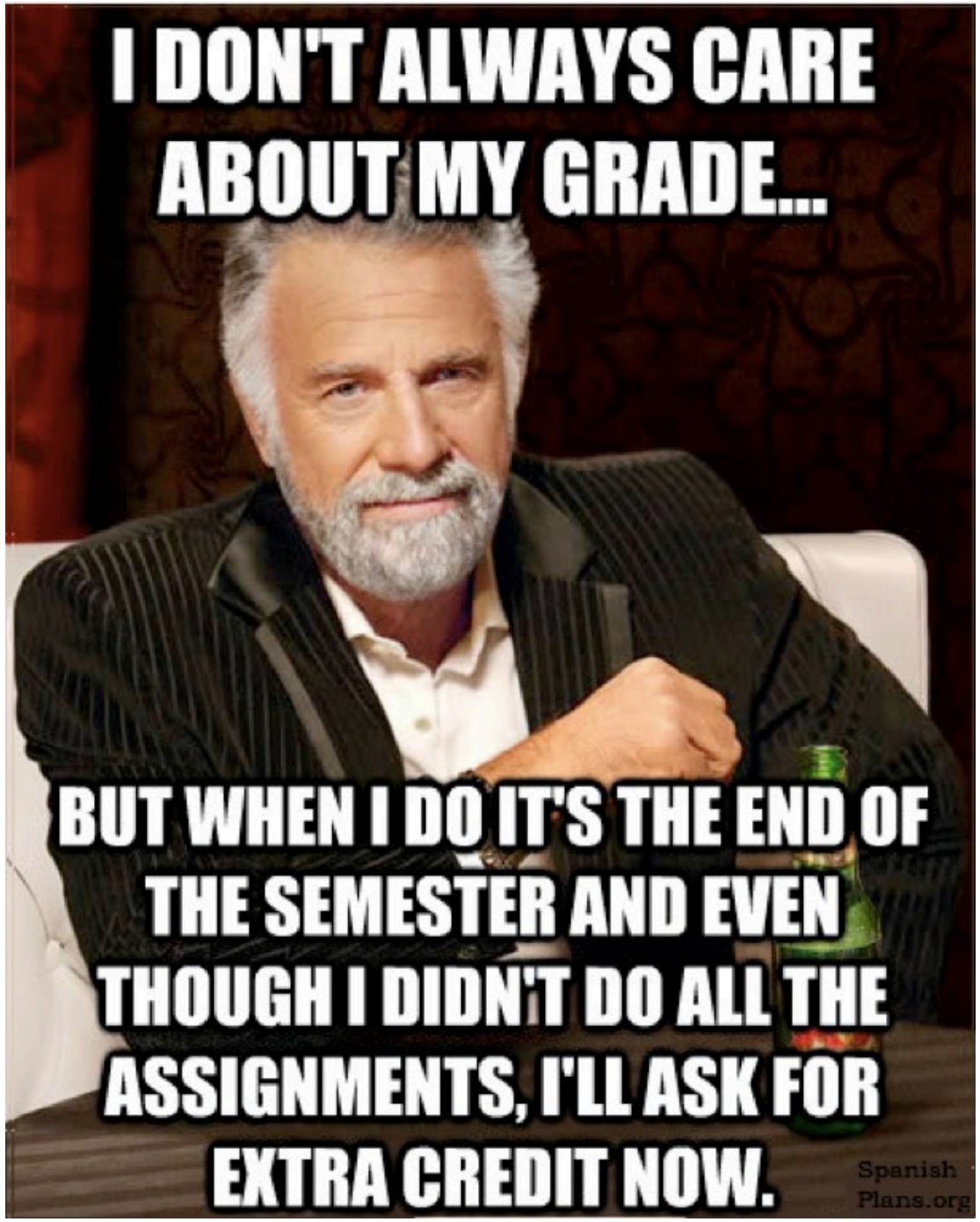
To ensure academic and professional integrity, I reserve the right to hold a one-on-one oral viva (defense) of submissions deemed questionable, to determine your knowledge and mastery of the topic/resources versus the material submitted. Failing that viva will result in an 'F' on the assignment and an Academic Integrity violation report filed with the Graduate School.

Academic Integrity

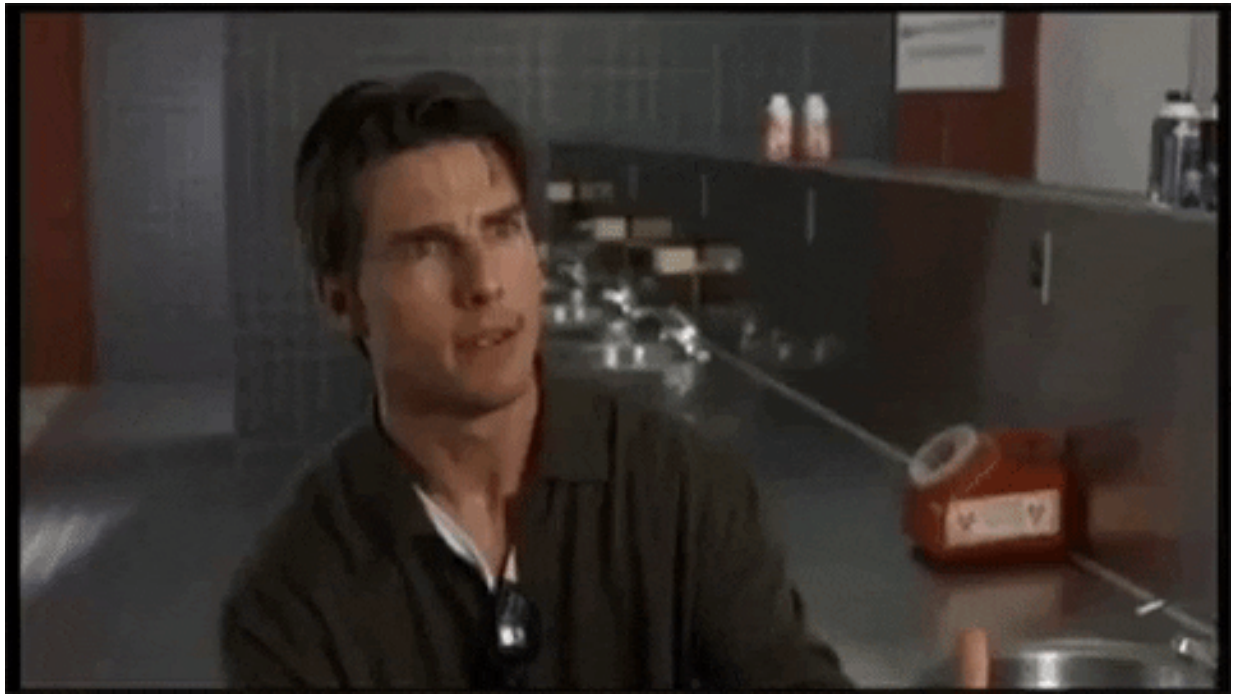
I take academic integrity very seriously. You should too.
If you're unsure about something, ask us.
You are at a top-tier (R1) research university.
Use this privilege to learn. A good grade will follow.
Don't throw away this opportunity by cheating.

3.8 Penalties

Academic misconduct could result in disciplinary action that may include, but is not limited to, suspension or dismissal. The **absolute minimum penalty** for a first offense of academic dishonesty in this course is a grade of zero on the assignment and a one-letter-grade reduction in the final class grade. However, depending on the nature of the offense, the penalty may be more severe, including but not limited to an `F` for the course, suspension, or expulsion. The minimum penalty for a second offense of academic dishonesty is an `F` for the course without possibility of dropping, but may be more severe.



Seek Help Early!



Help us help you.

491 (Undergrad) vs 691 (Grad)

This is a "Research" class – anyone with the right mindset (+ an intro ML class) is ready

Main difference: projects (scope and novelty)

- Projects will be graded in terms of "relative growth"
  - ***Some may have previous research experience***
  - ***Some are taking this class to get research experience***
- Grad projects:
  - Original & unique research hypothesis with a potential for publication
- Undergrads projects can be:
  - Original & unique research hypothesis with a potential for publication
  - Working on an idea that we provide (i.e. you get to skip "ideation")
  - Innovative applications or combination of existing work

FAQ: Can I join your research lab?

- Joining (See FAQ on my website)
  - Take this class and talk to me during office hours about your interests

- Will I get paid for research ?
  - Depends (I currently only have funding for PhD students)
  - Undergraduate Research & Prestigious Scholarships - UMBC
  - CWIT Scholars – Center for Women in Technology – UMBC
  - You can also do research for credit (e.g. CMSC 299, 499, 698, 699)

Other Questions?

Robust Machine Learning
Lecture 0: Introduction

Machine Learning: The Success Story

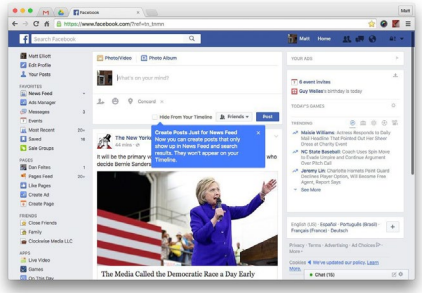
IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

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



Machine Learning: The Success Story

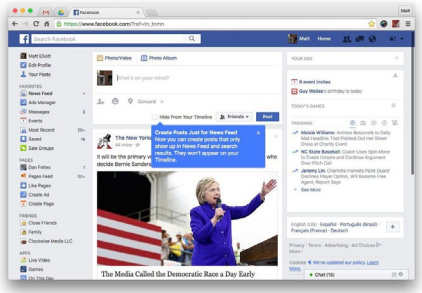
IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

Andrew Ng
@AndrewYNg
Follow

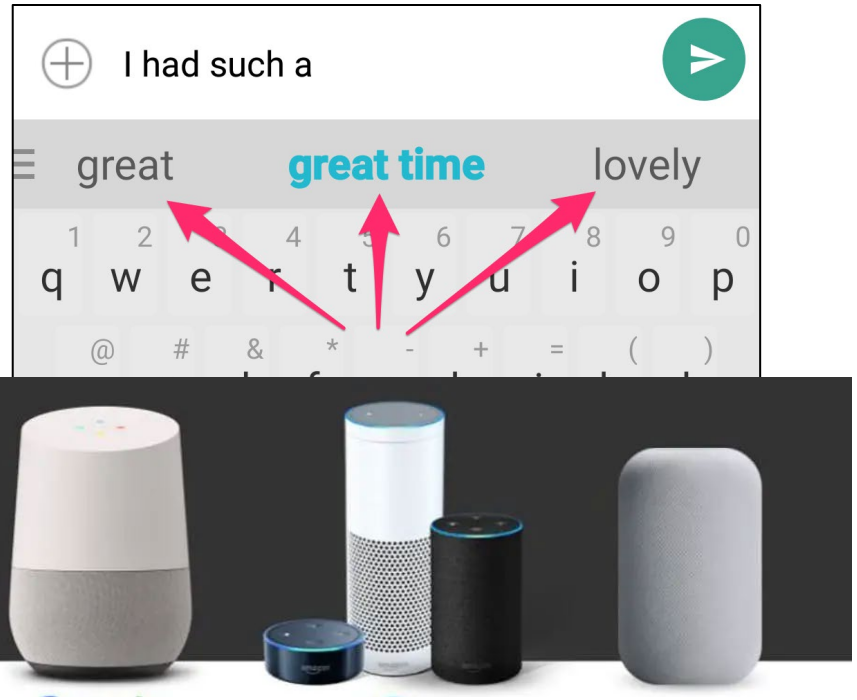
"AI is the new electricity!" Electricity transformed countless industries; AI will now do the same.

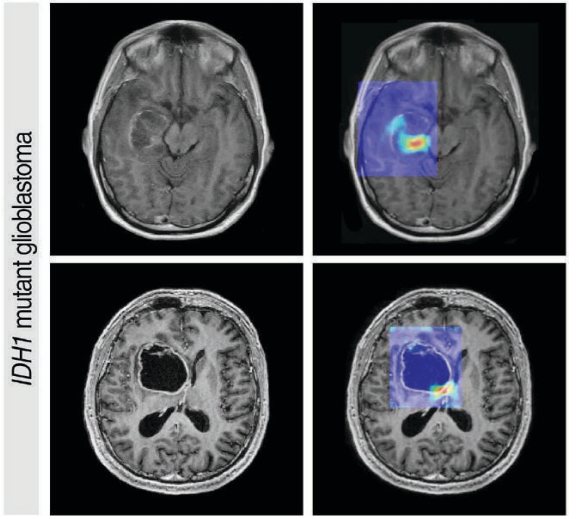
2016: The Year That Deep Learning Took Over t

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE



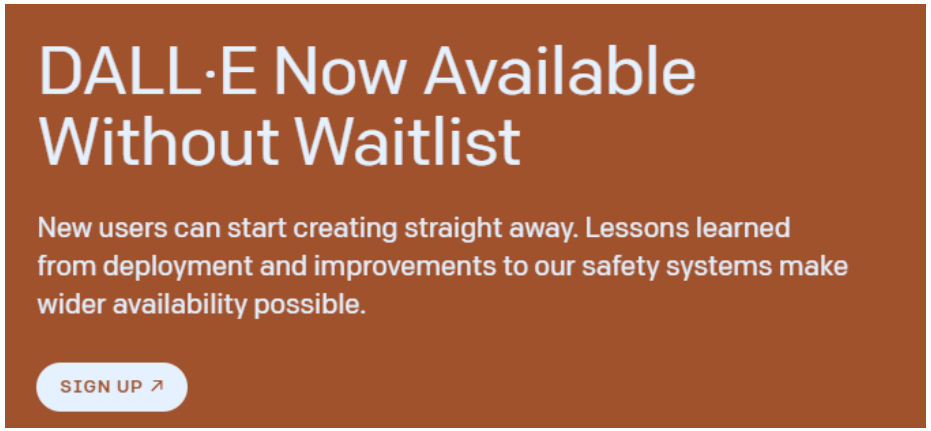
Models that learn from data are embedded in our lives

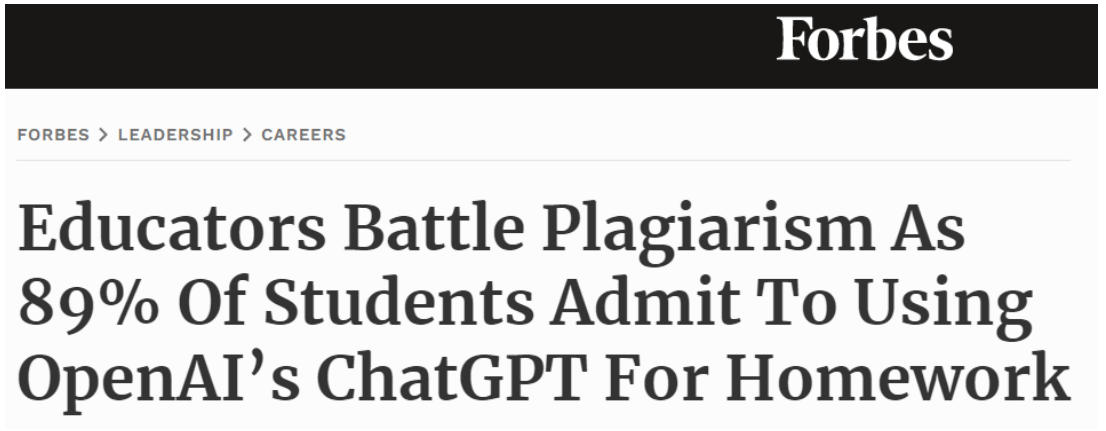


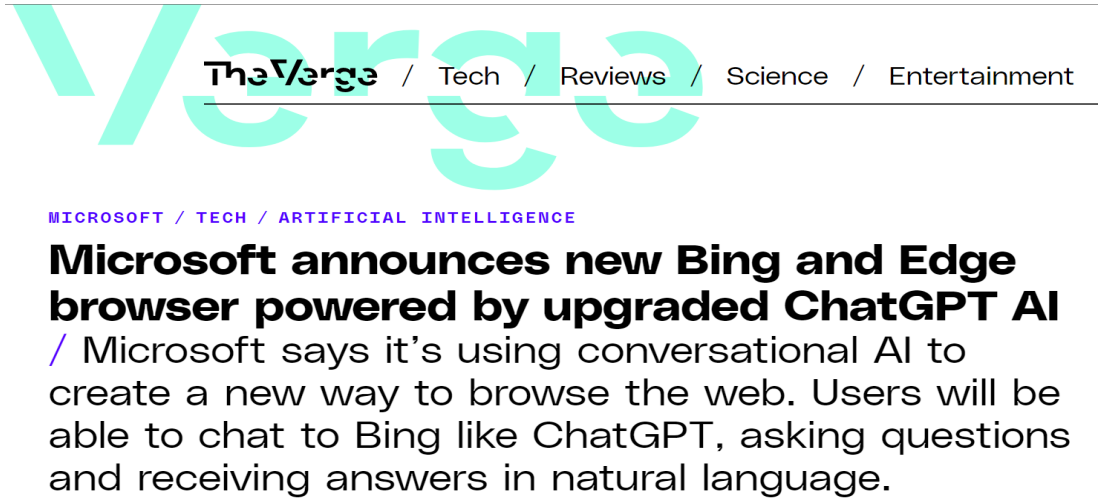


Models that learn from data are embedded in our lives

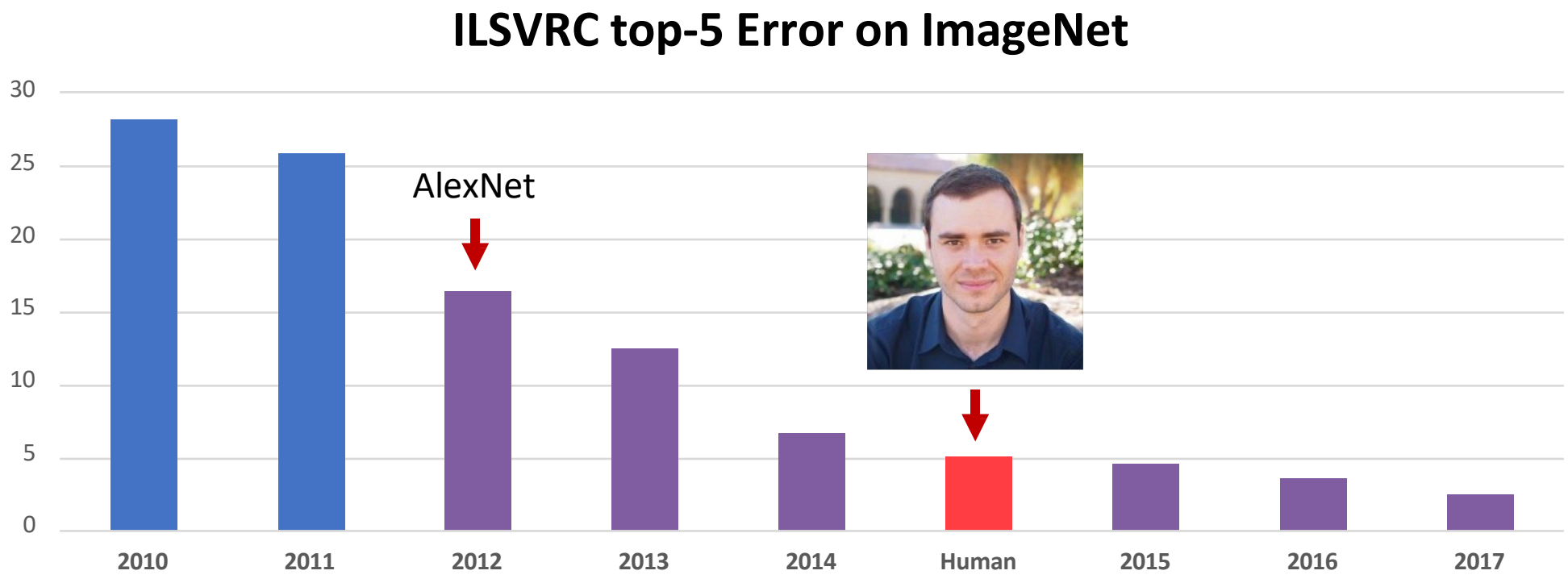
Recent advances have been rapidly adopted by common, non-expert users





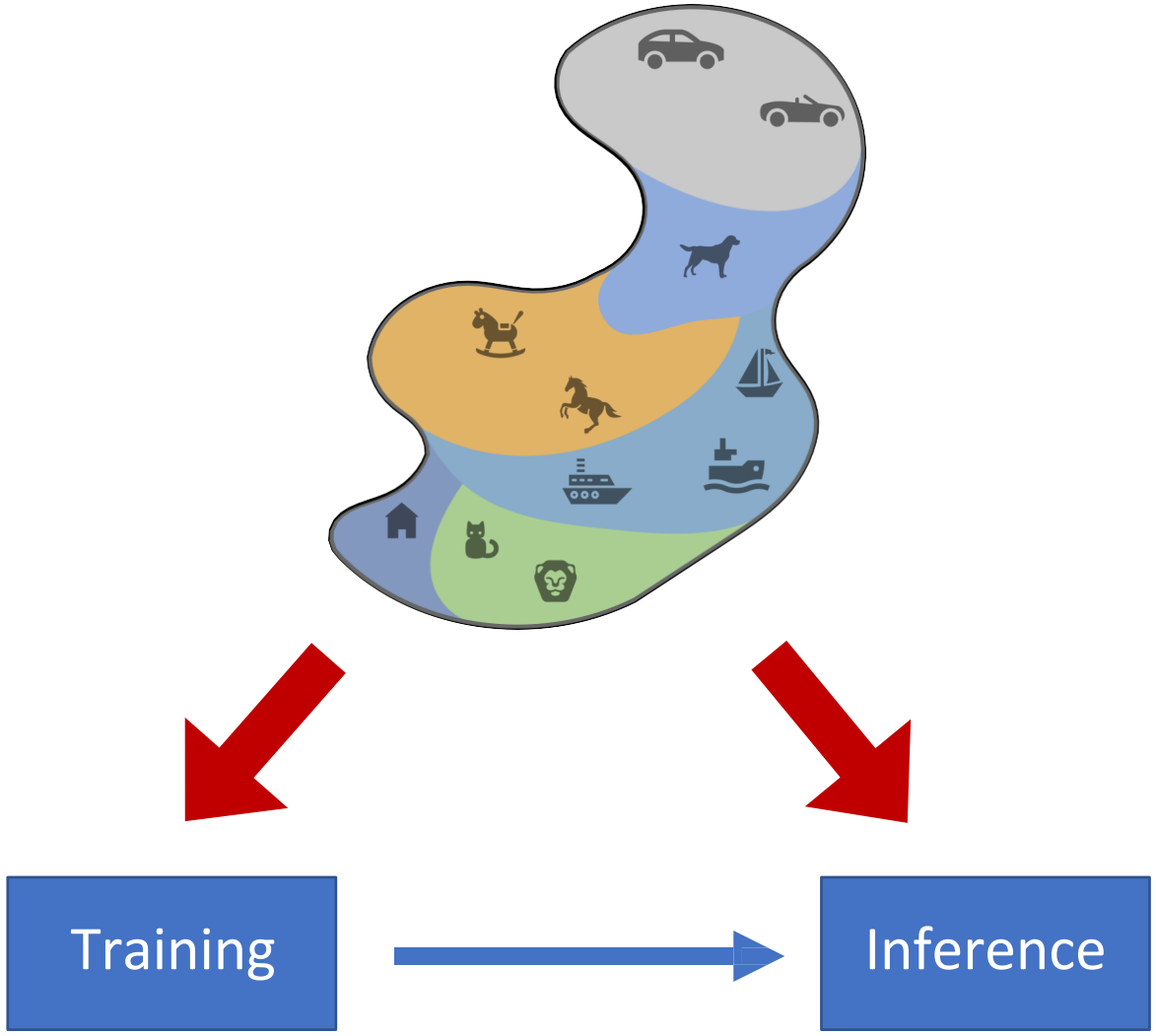


ImageNet: An ML Home Run



But what do these results *really* mean?

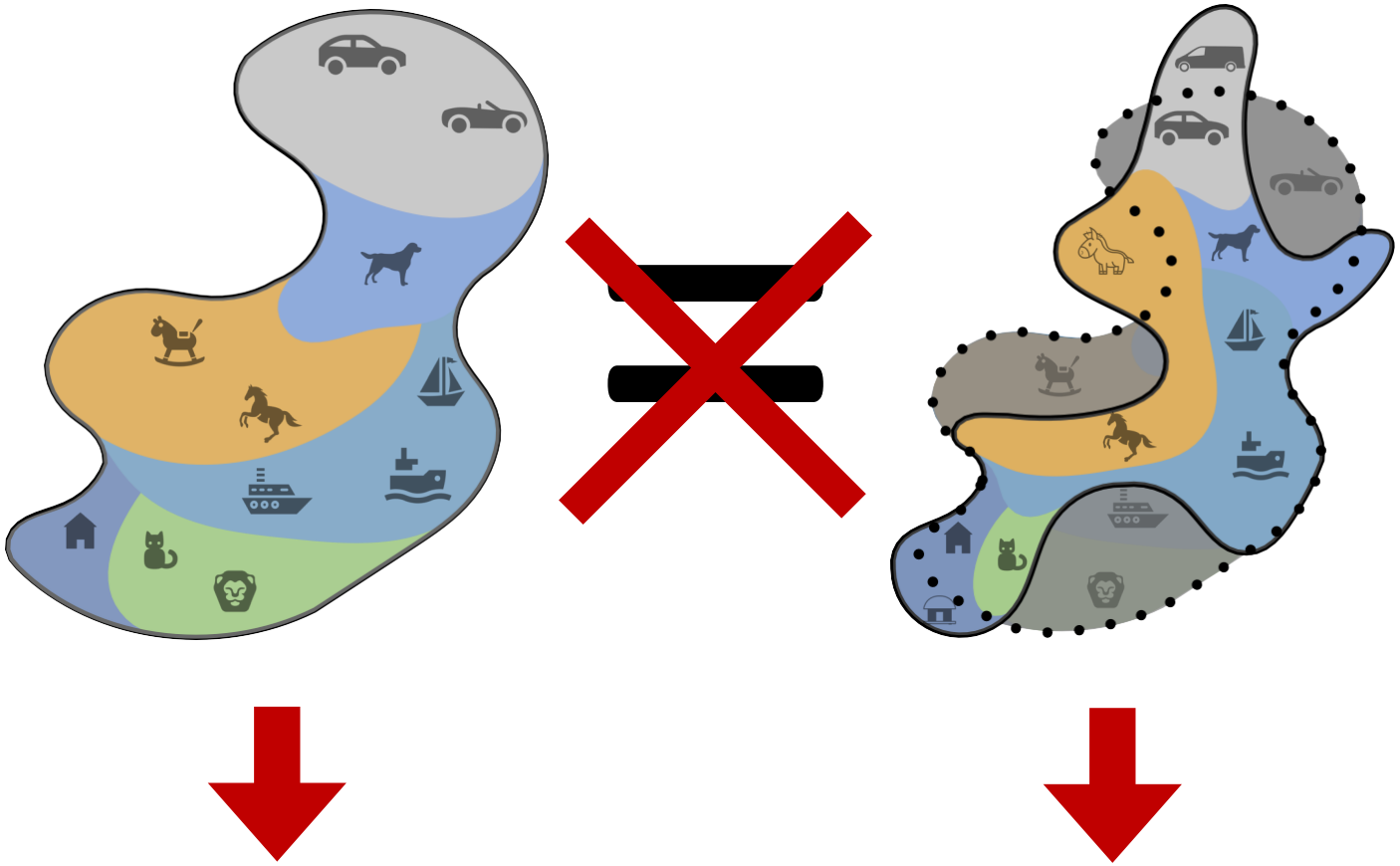
A Limitation of the (Supervised) ML Framework



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

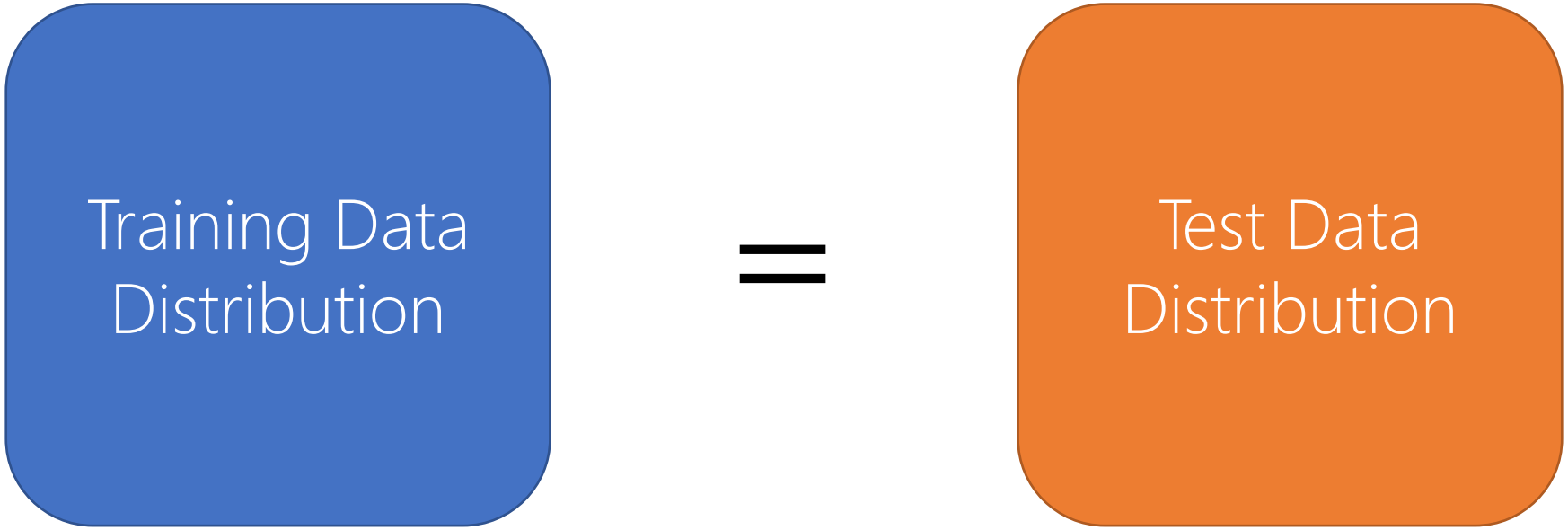


Measure of performance:
Fraction of mistakes during testing

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

What can go wrong?

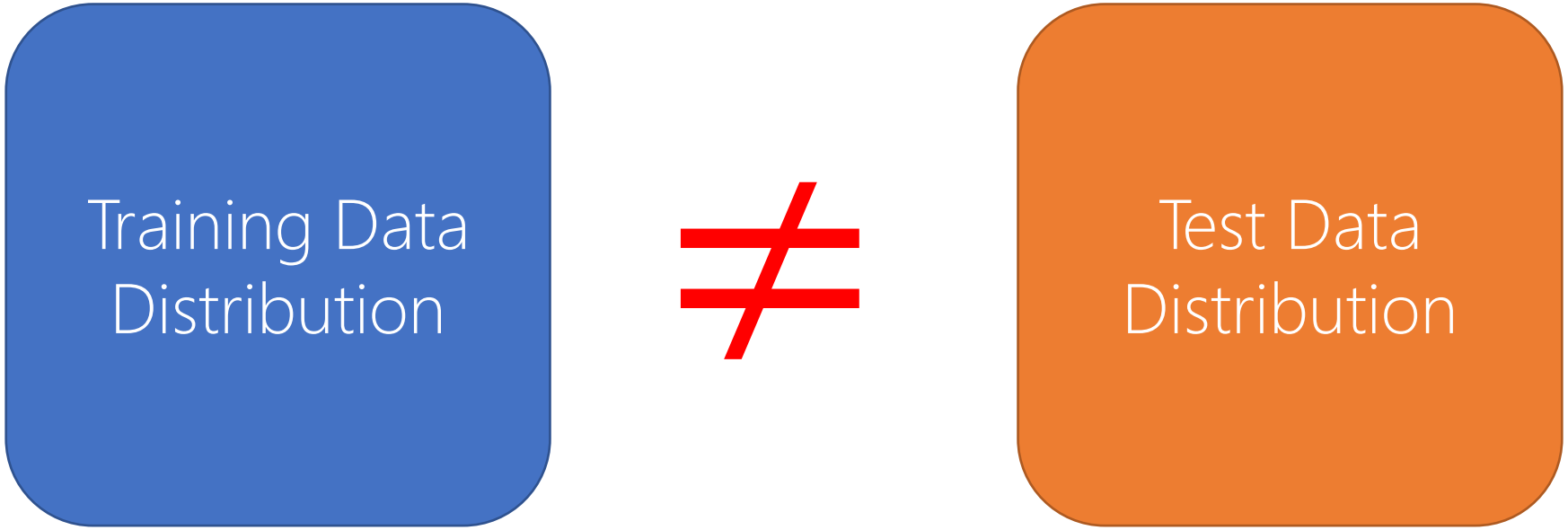
Standard *i.i.d.* Assumption in Machine Learning



"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 in Computer Vision

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

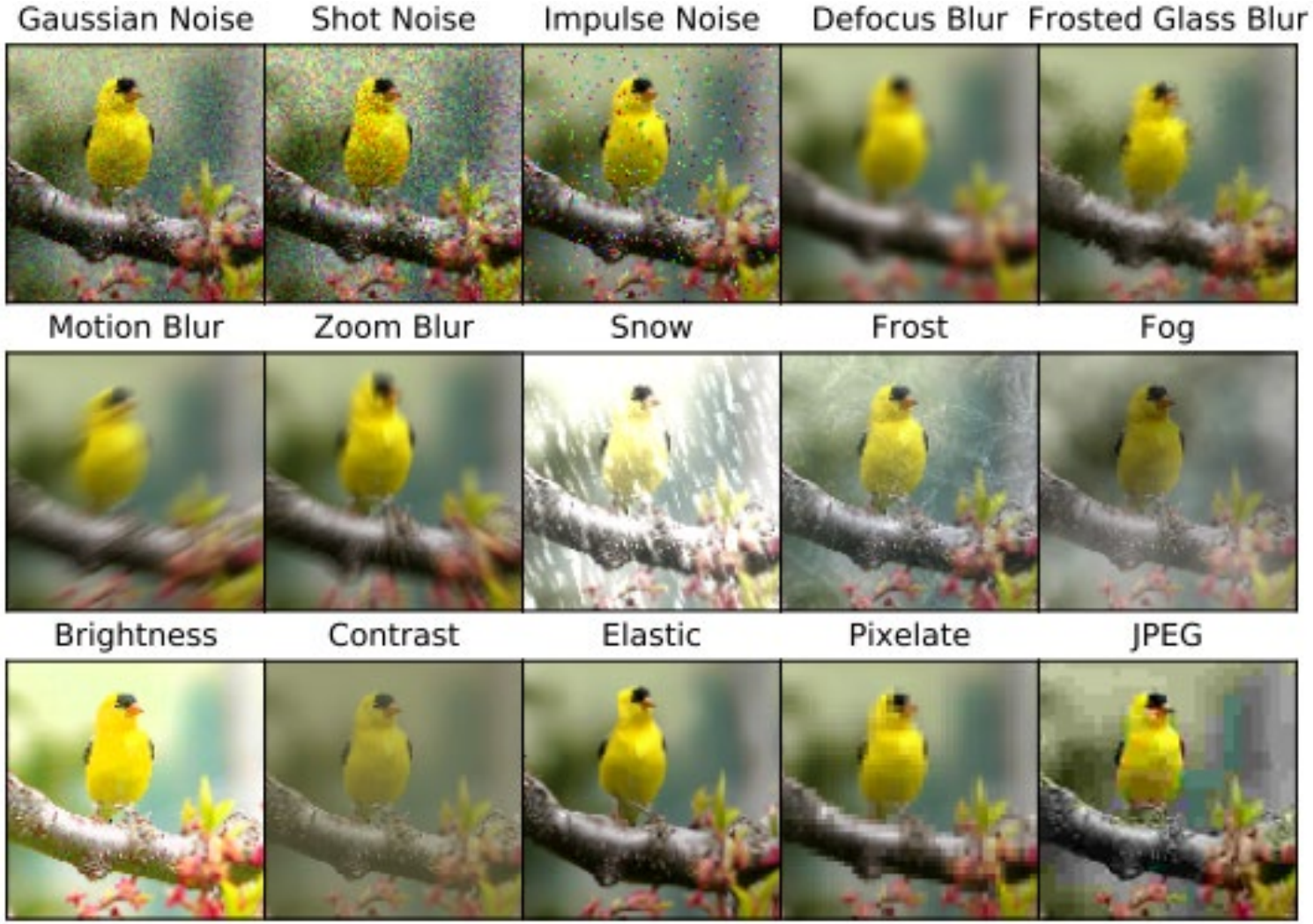
Poses can fool Image Classifiers

- **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
    - ...

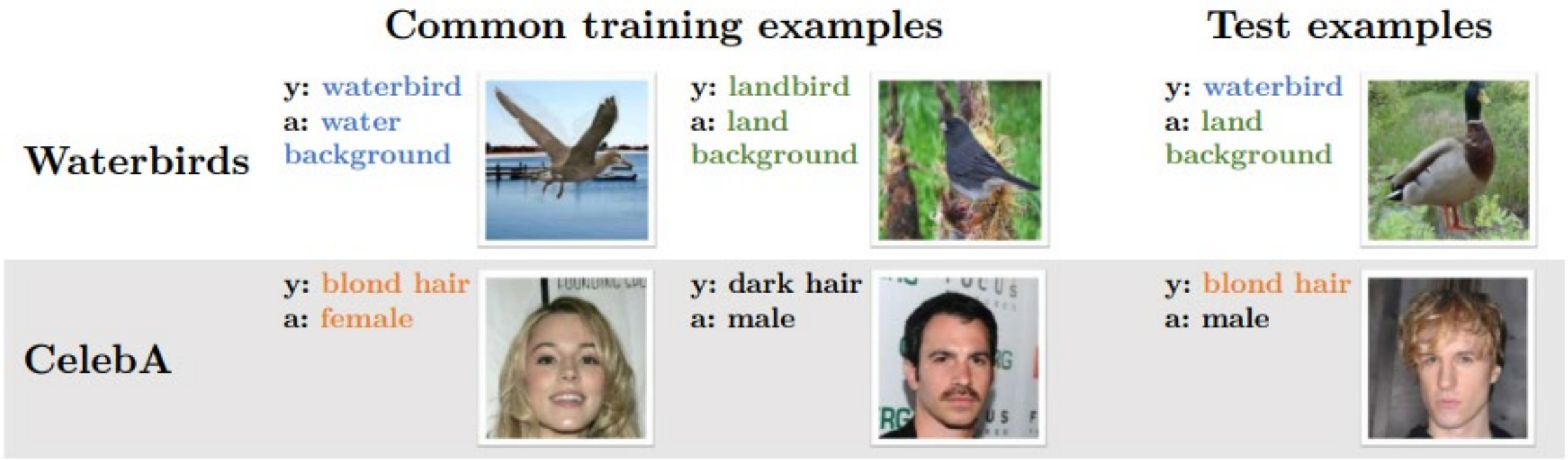
Alcorn et al. CVPR 2019

Natural Corruptions affect accuracy



Hendrycks et al. ICLR 2019

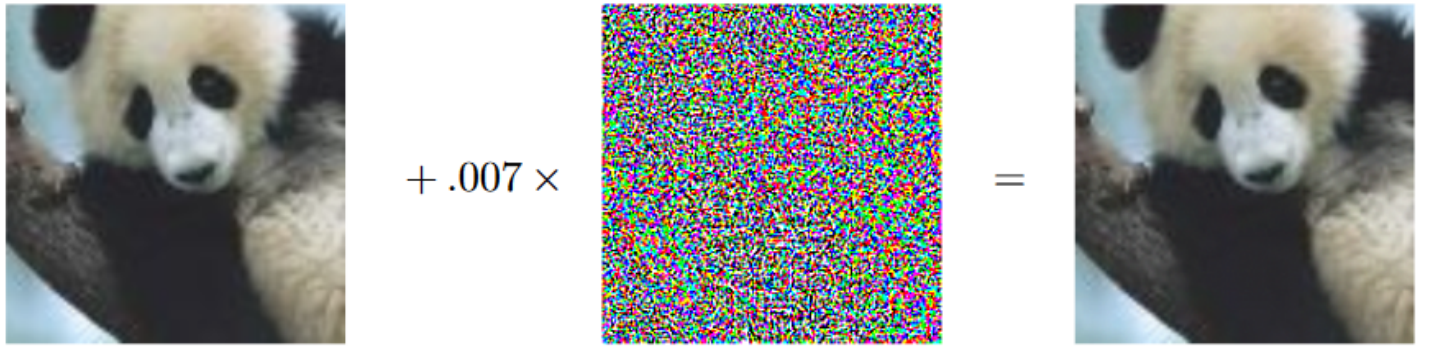
Spurious Correlations / Biased Datasets



Sagawa et al. ICLR 2020

Adversarial Attacks on Image Classifiers

- 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

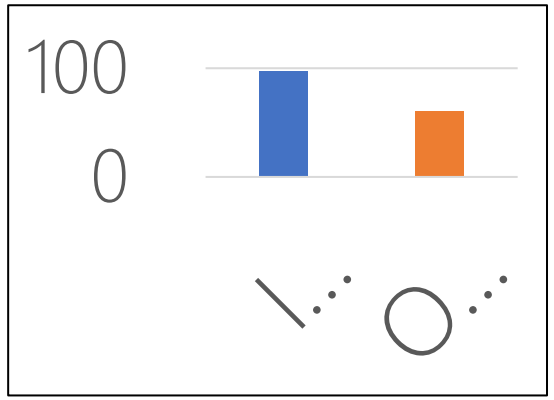
Lack of Diverse Data hurts Reliability

Mallard
Water Background

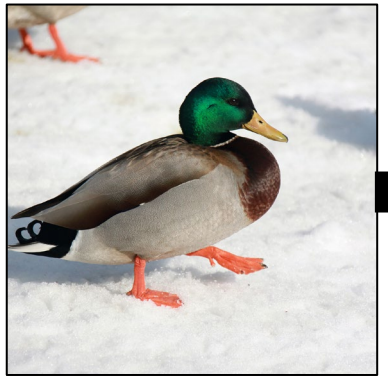
Penguin
Snow Background

Training

ML Model



Mallard
Snow Background



Test

ML Model

"Penguin"

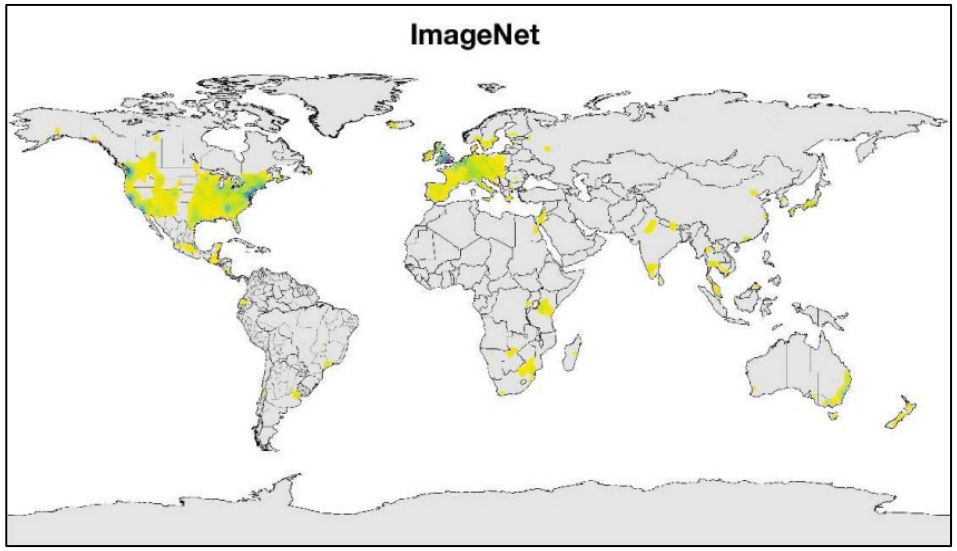
Lack of Diverse Data hurts Reliability

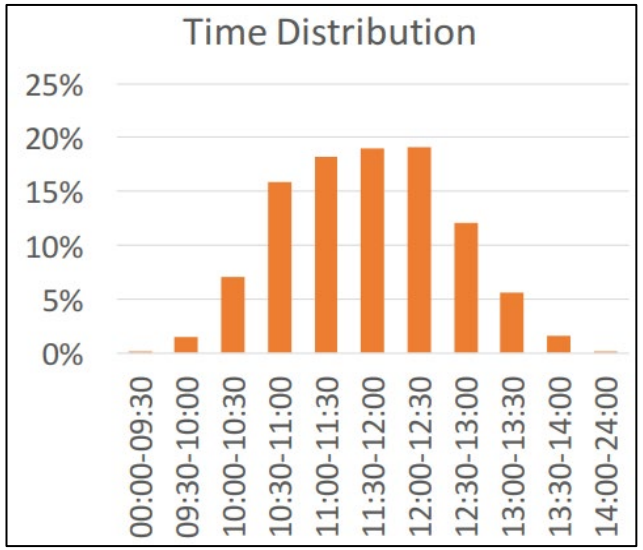
Mallard
Snow Background

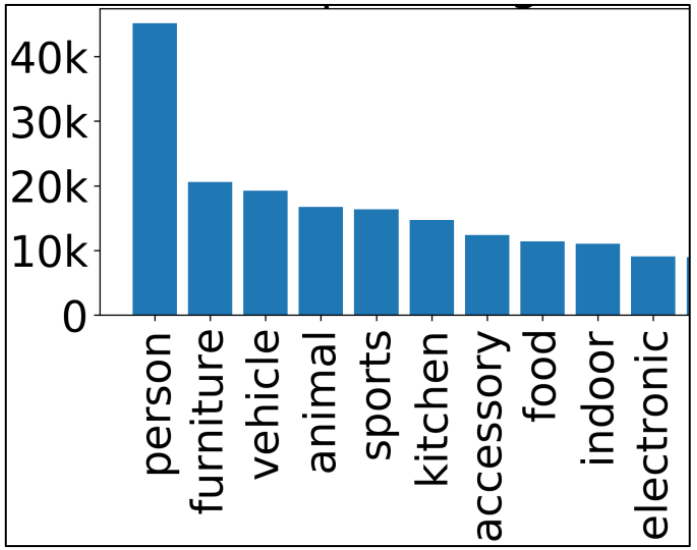
Test

ML Model

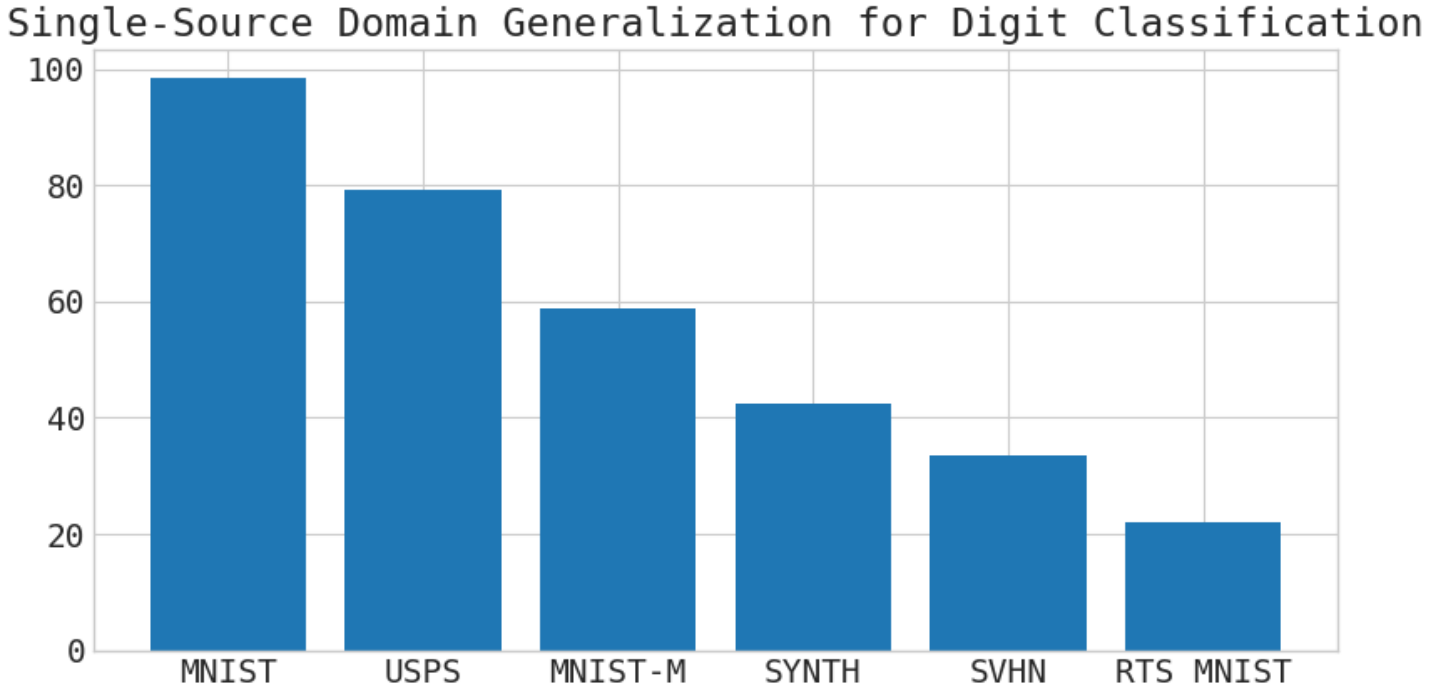
"Pengui n"







Domain Shift is a Nuisance

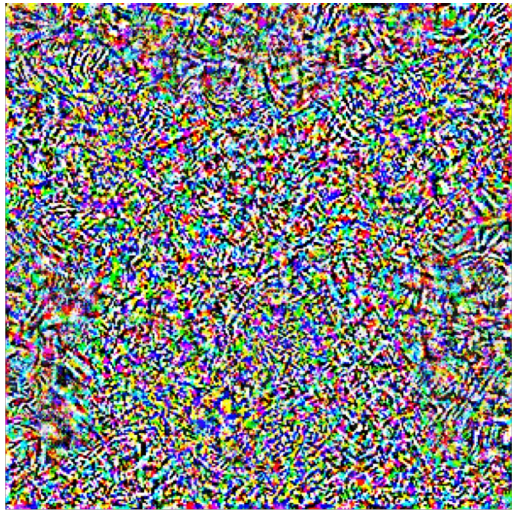


ML Predictions Are (Mostly) Accurate but Brittle

“pig” (91%)

+ 0.005 x

noise (NOT random)



=

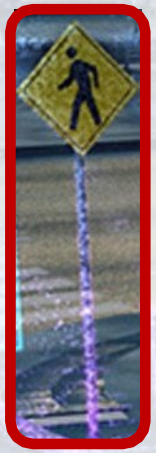
“airliner” (99%)

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

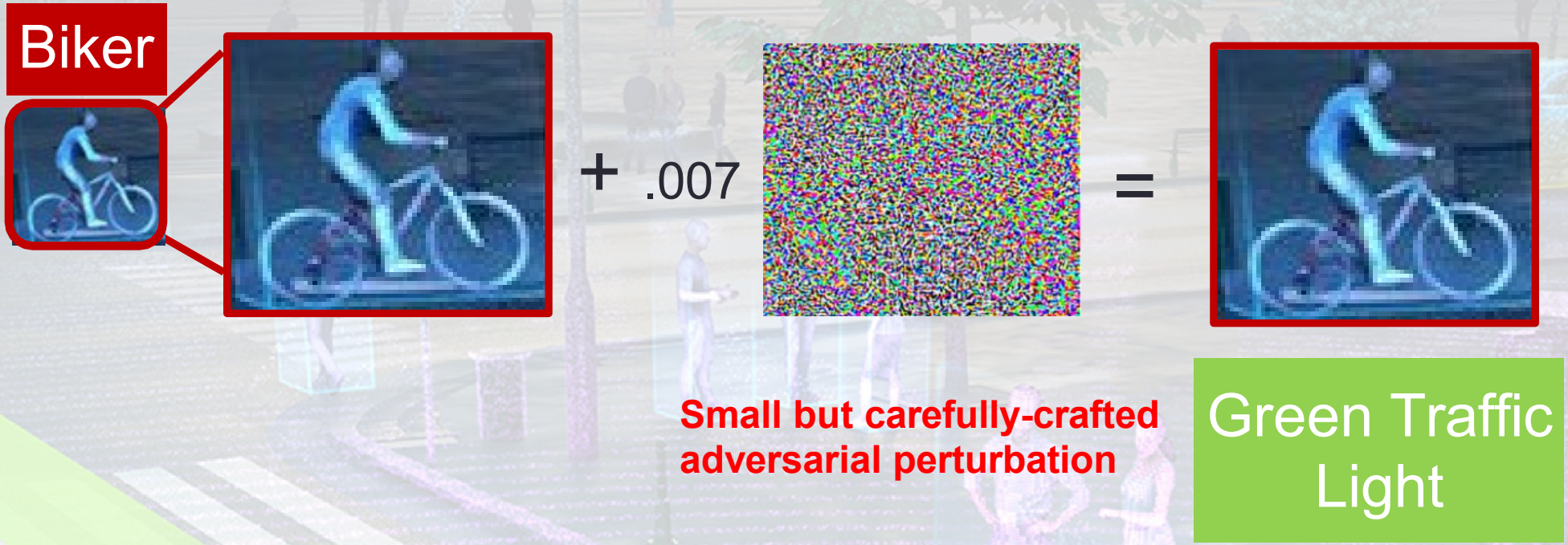
But also: [Dalvi Domingos Mausam Sanghai Verma 2004][Lowd Meek 2005]
[Globerson Roweis 2006][Kolcz Teo 2009][Barreno Nelson Rubinstein Joseph Tygar 2010]
[Biggio Fumera Roli 2010][Biggio Fumera Roli 2014][Srndic Laskov 2013]

Biker

Pedestrian Sign



Persons



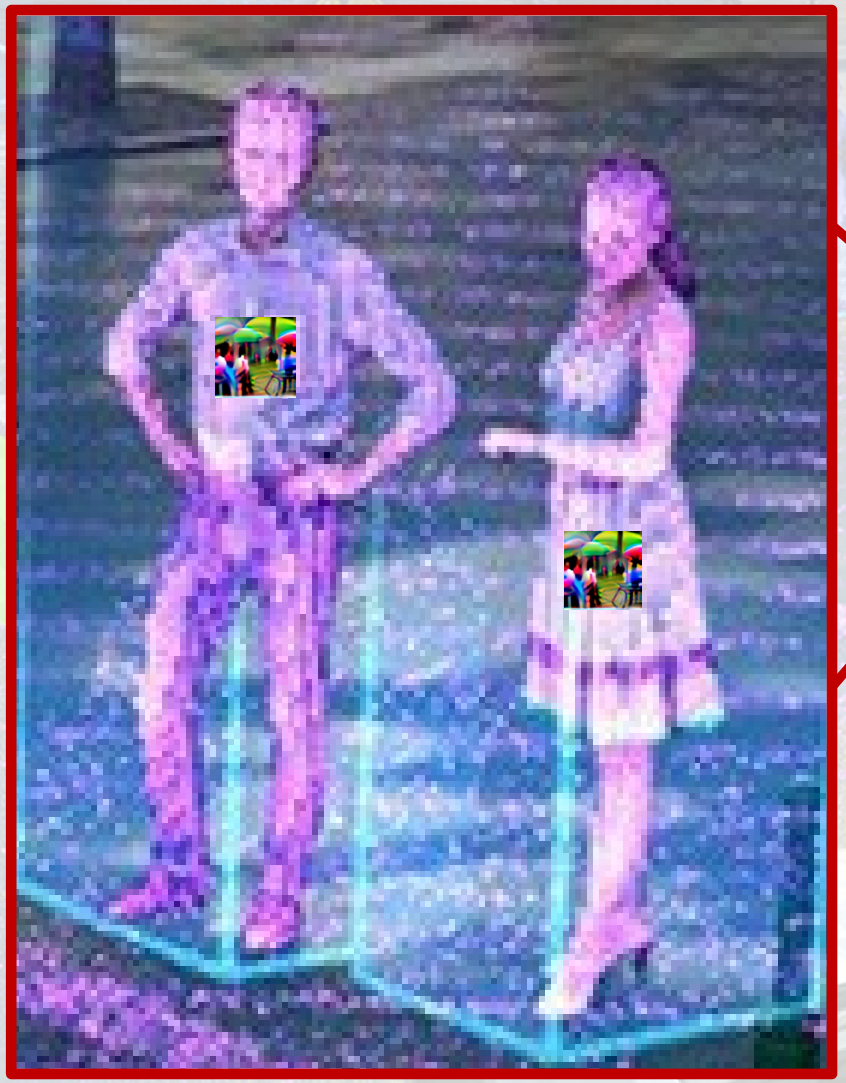
Adversarial Perturbation Attack

Pedestrian Sign

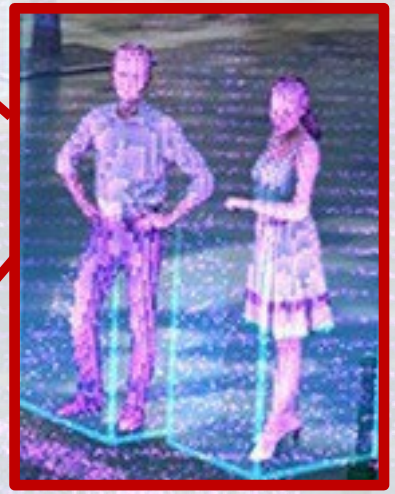
Speed Limit 45 Sign

Adversarial Rotation Attack

No Person



Persons



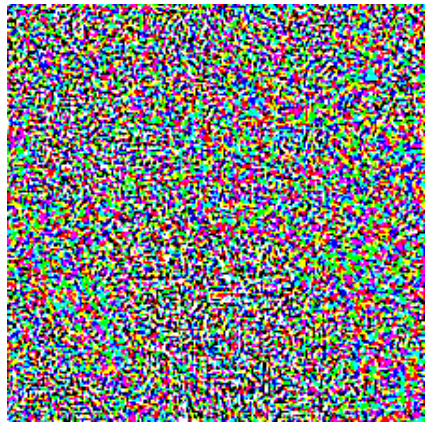
Adversarial Patch Attack

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

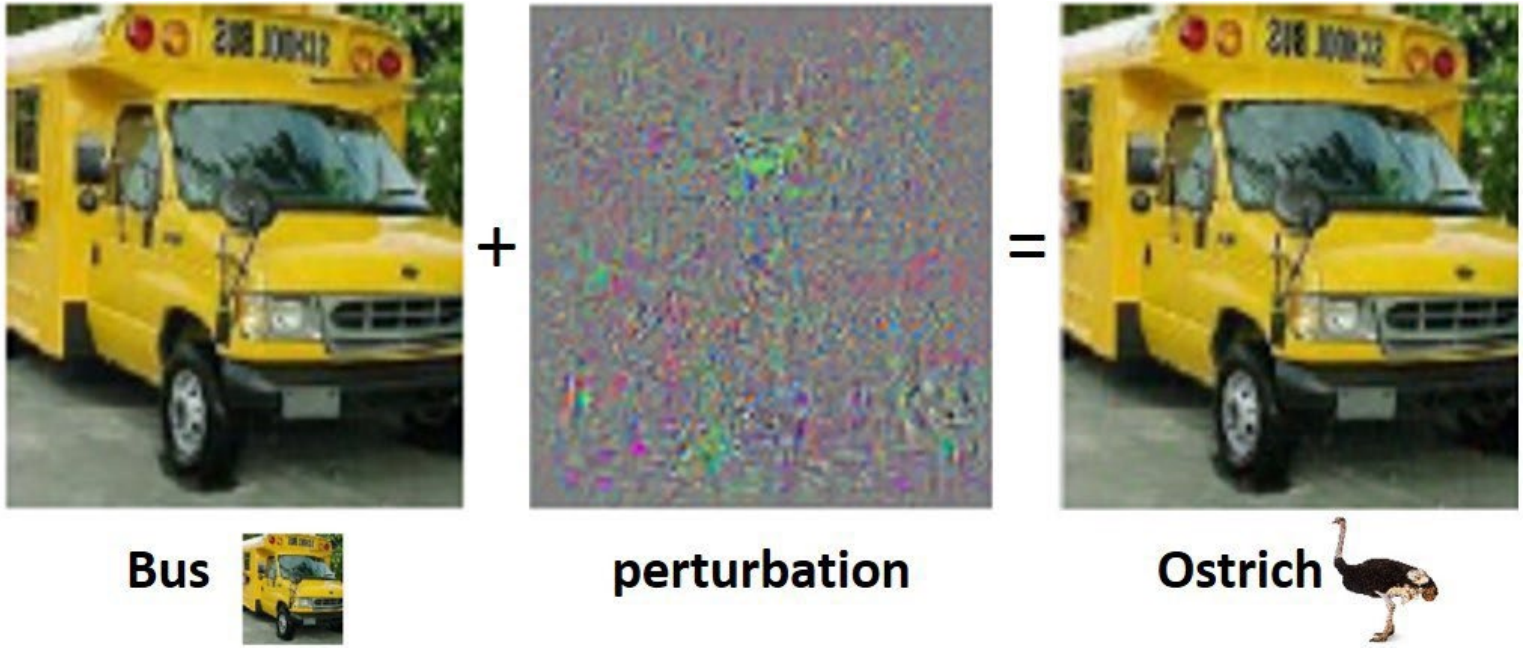
Classified as gibbon
99.3% confidence

Gibbon

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

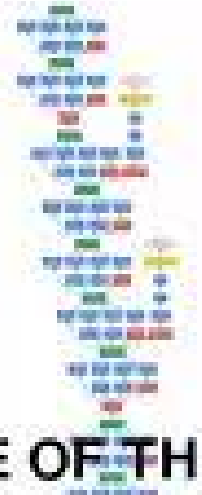
Adversarial Examples

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

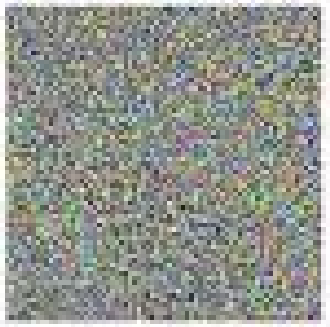


Picture from: Szagedy et al. (2014) – Intriguing Properties of Neural Networks

WHO WOULD WIN?



STATE OF THE ART NEURAL NETWORK

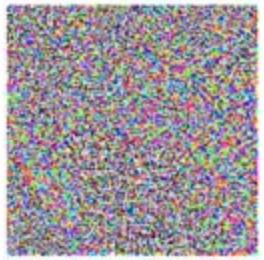


ONE NOISY BOI

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



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

"nematode"
8.2% confidence

$=$

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

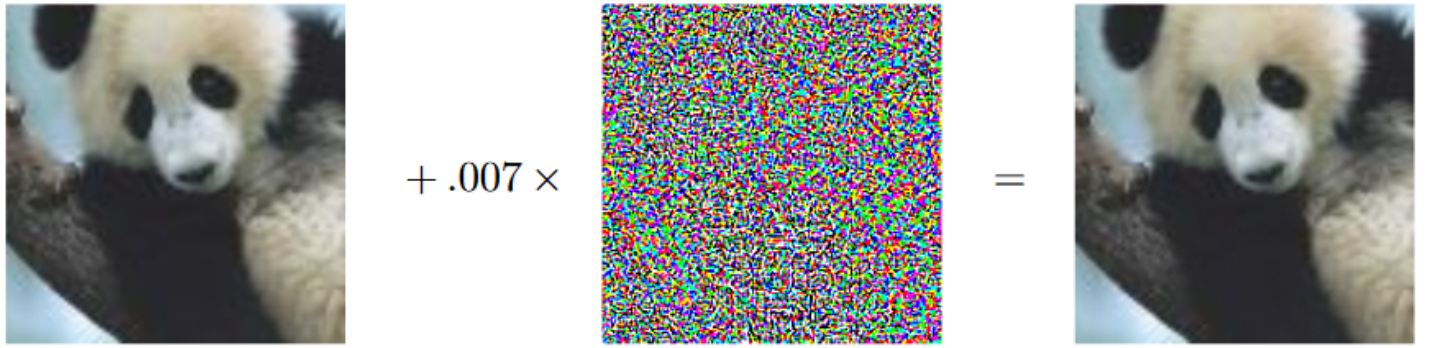
"gibbon"
99.3 % confidence

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

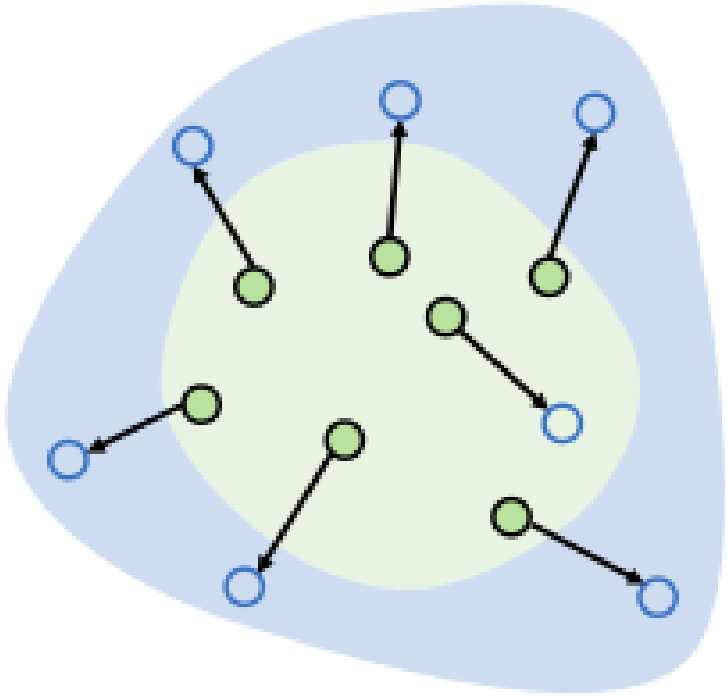
min—max optimization

maximization: find adversarial images

minimization: train classifier to correctly classify such images

norm-bounded perturbations

==> robustness within the norm-ball

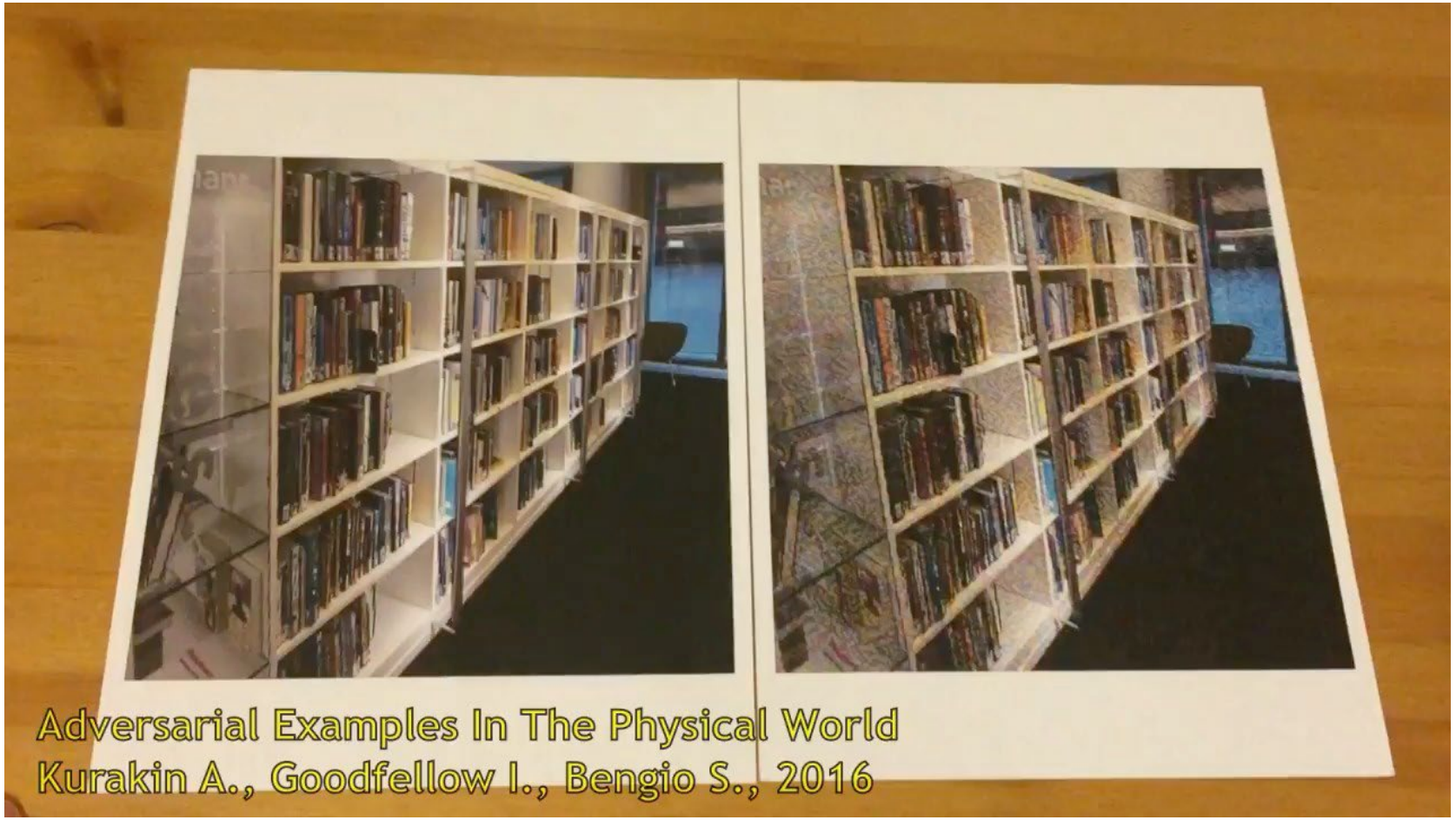


$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

Madry et al. ICLR 2018

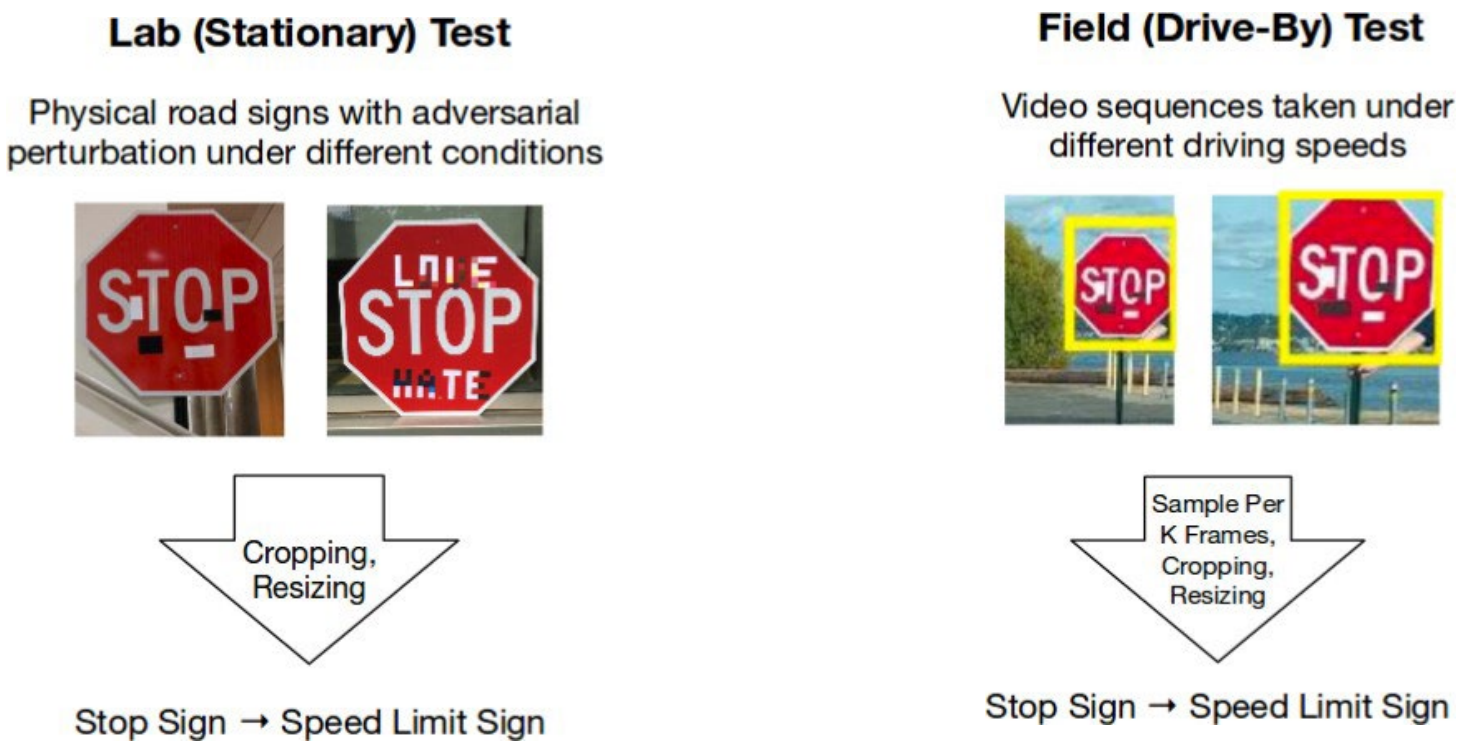
Physical-World Attack: Printed Adversarial Images

- 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

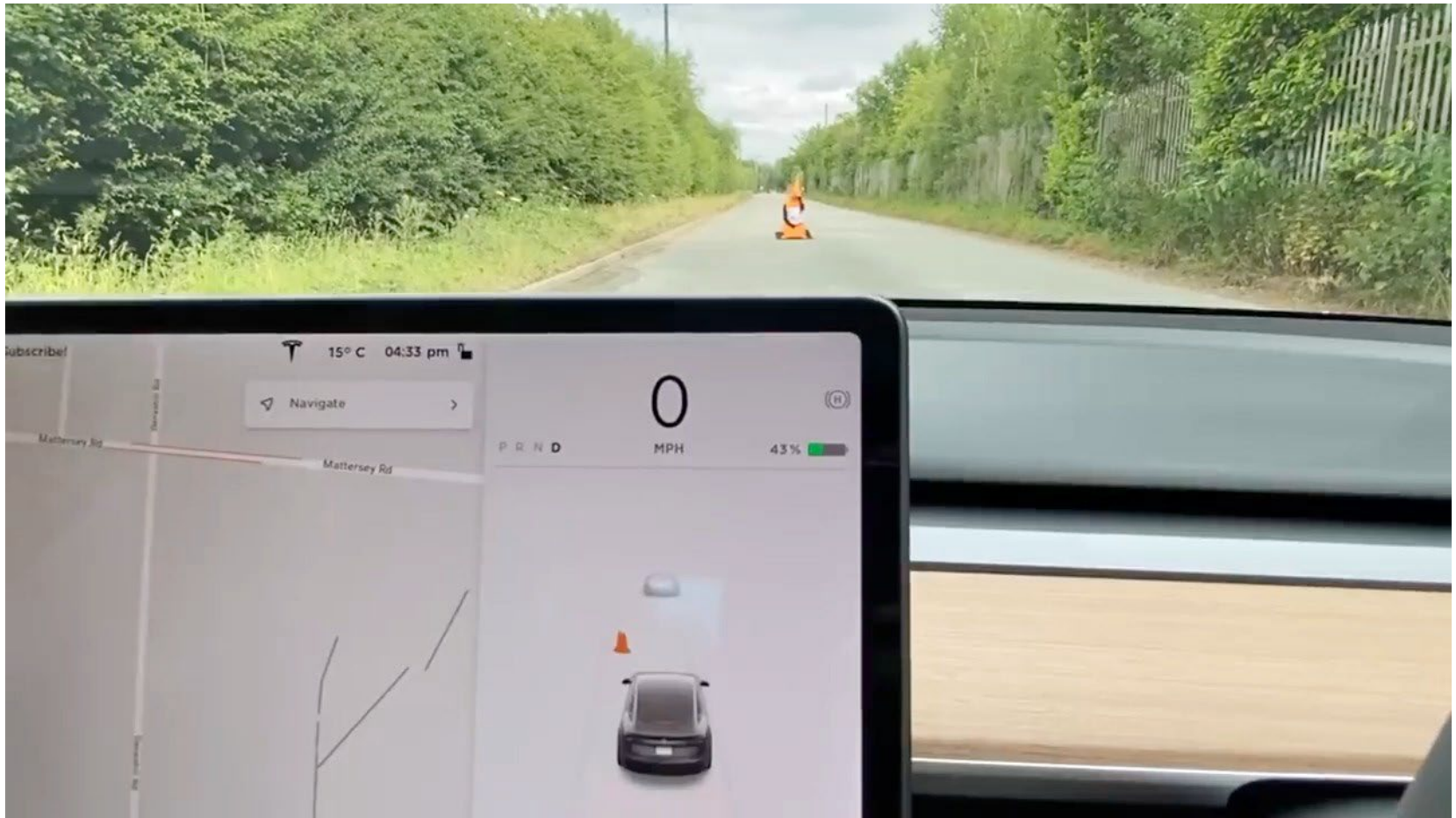
Physical-World Attack: Adversarial Patch

- 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

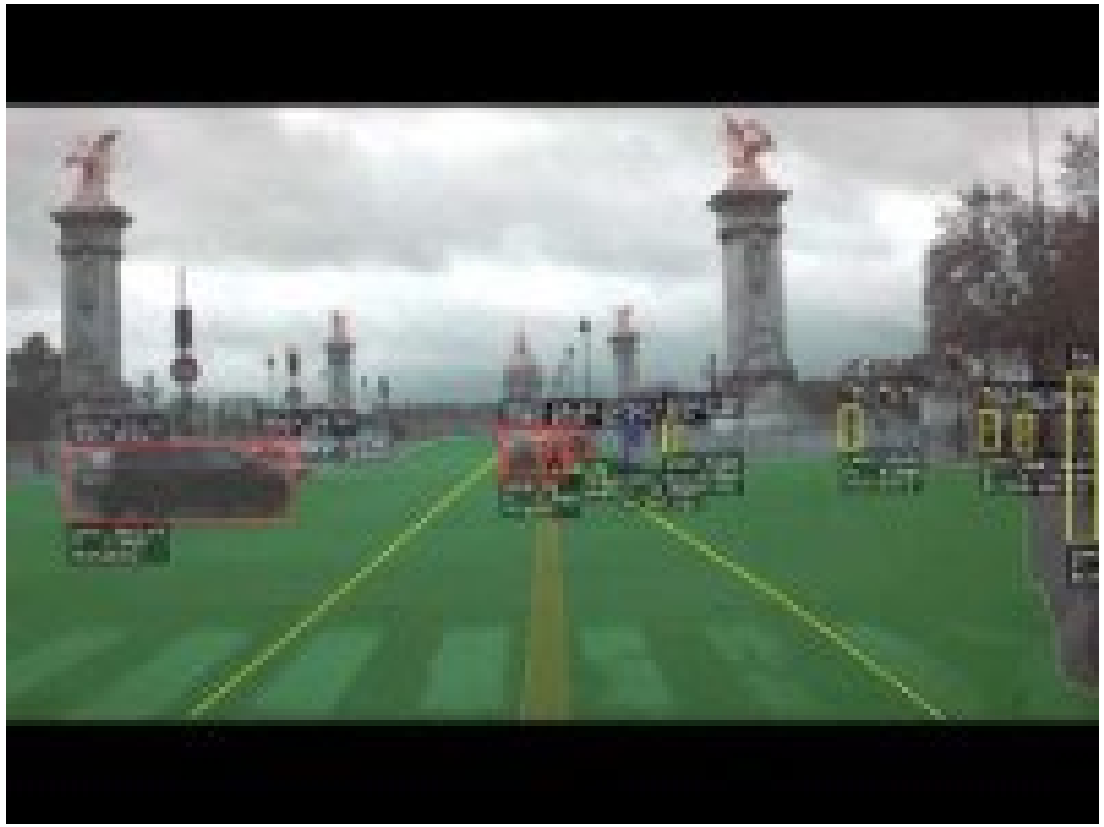
- Non-scientific example: a Tesla owner checks if the car can distinguish a person wearing a cover-up from a traffic cone



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.

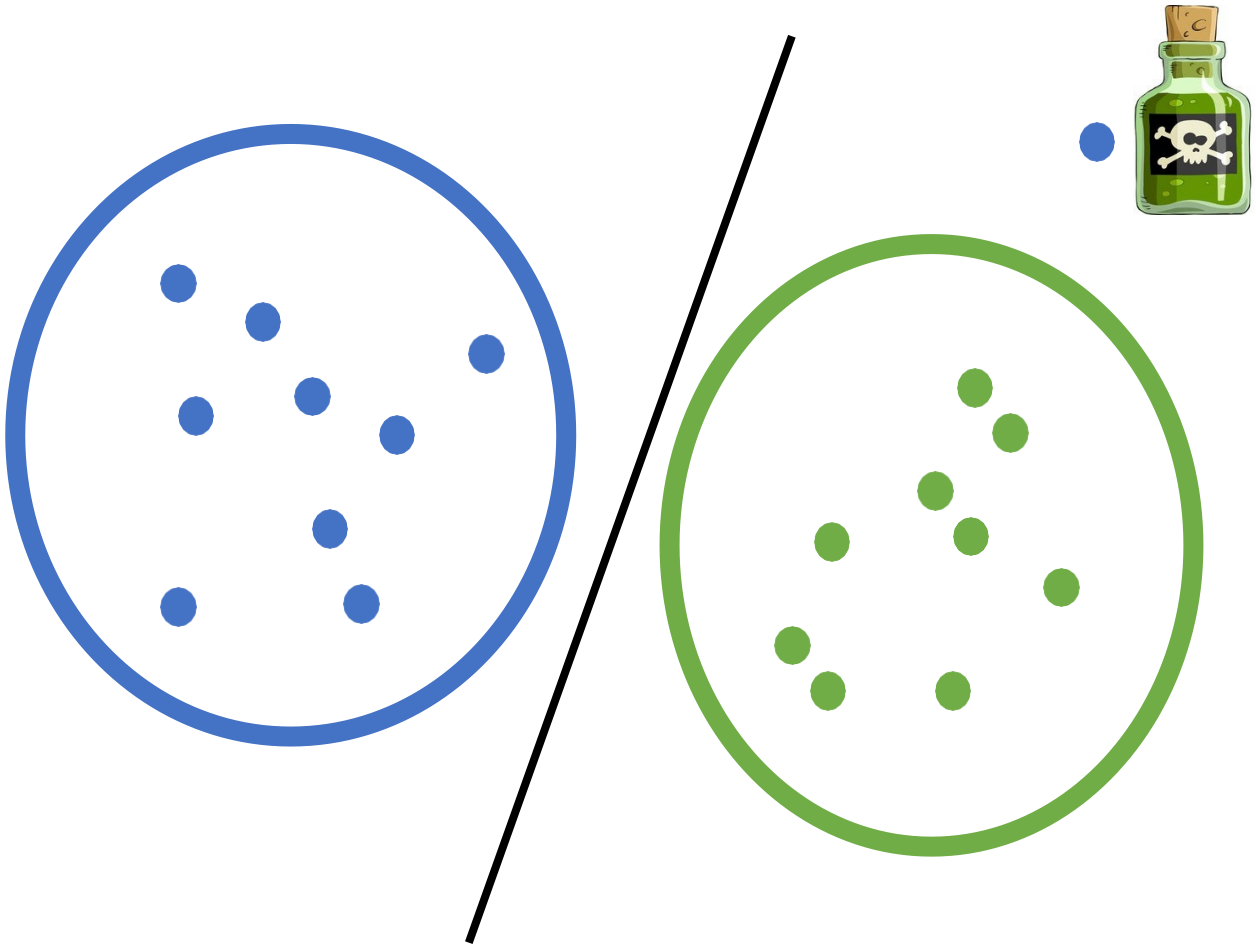
npr

TECHNOLOGY

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

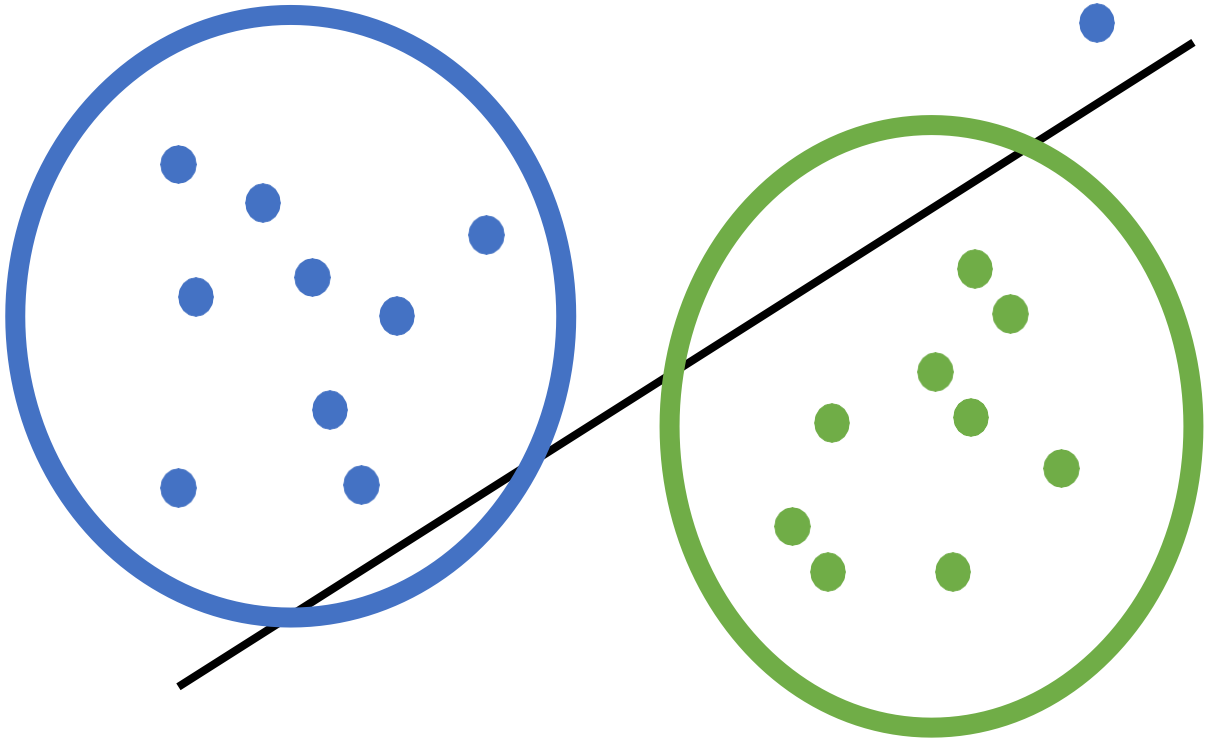
Data Poisoning

Goal: Maintain training accuracy but hamper generalization



Data Poisoning

Goal: Maintain training accuracy but hamper generalization



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

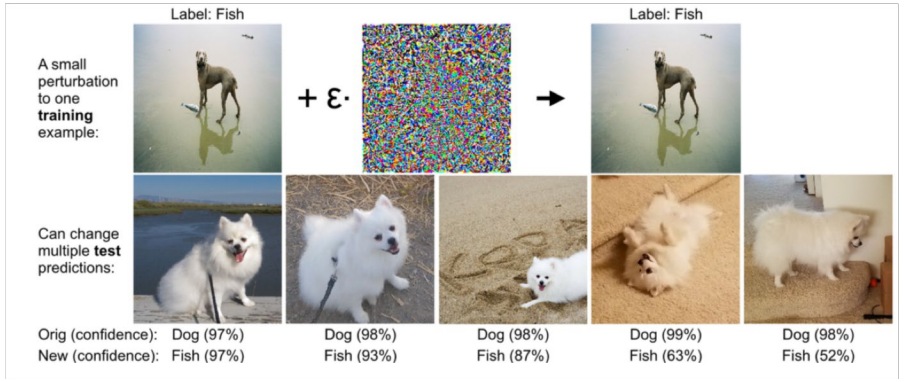
→ **But:** seems less so in deep learning

→ **Reason:** Memorization?

Data Poisoning

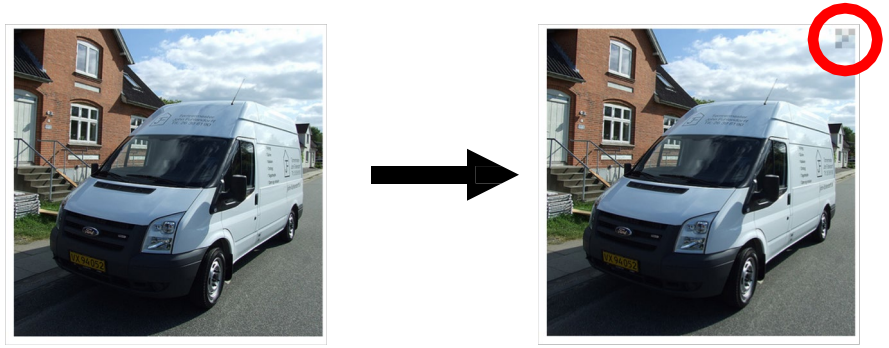
classification of **specific** inputs

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]**)

You don’t even need poisonous samples.

Re-ordering training batches of clean data → failures

Mole Recruitment: Poisoning of Image Classifiers via Selective Batch Sampling

Ethan Wisdom Tejas Gokhale Chaowei Xiao Yezhou Yang

Arizona State University

{ewisdom, tgokhale, cxiao17, yz.yang}@asu.edu

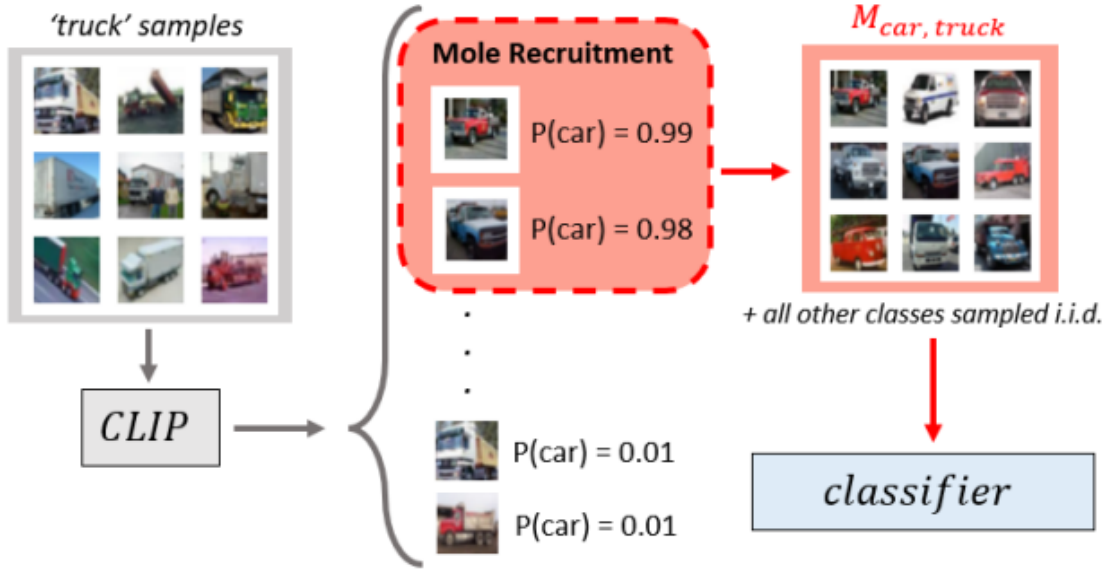


Figure 1. Using an external model such as CLIP, the distribution of the confounding class (truck) samples is examined with respect to their softmax probabilities of the attacked class (car). The samples with the highest probability are used to train an image classifier, which confounds the classifier’s ability to distinguish between the attacked and confounding classes.

Are we doomed?

Are we doomed?

(Is ML inherently not reliable?)

Are we doomed?

(Is ML inherently not reliable?)

We need to re-think how we do ML

THE WHITE HOUSE

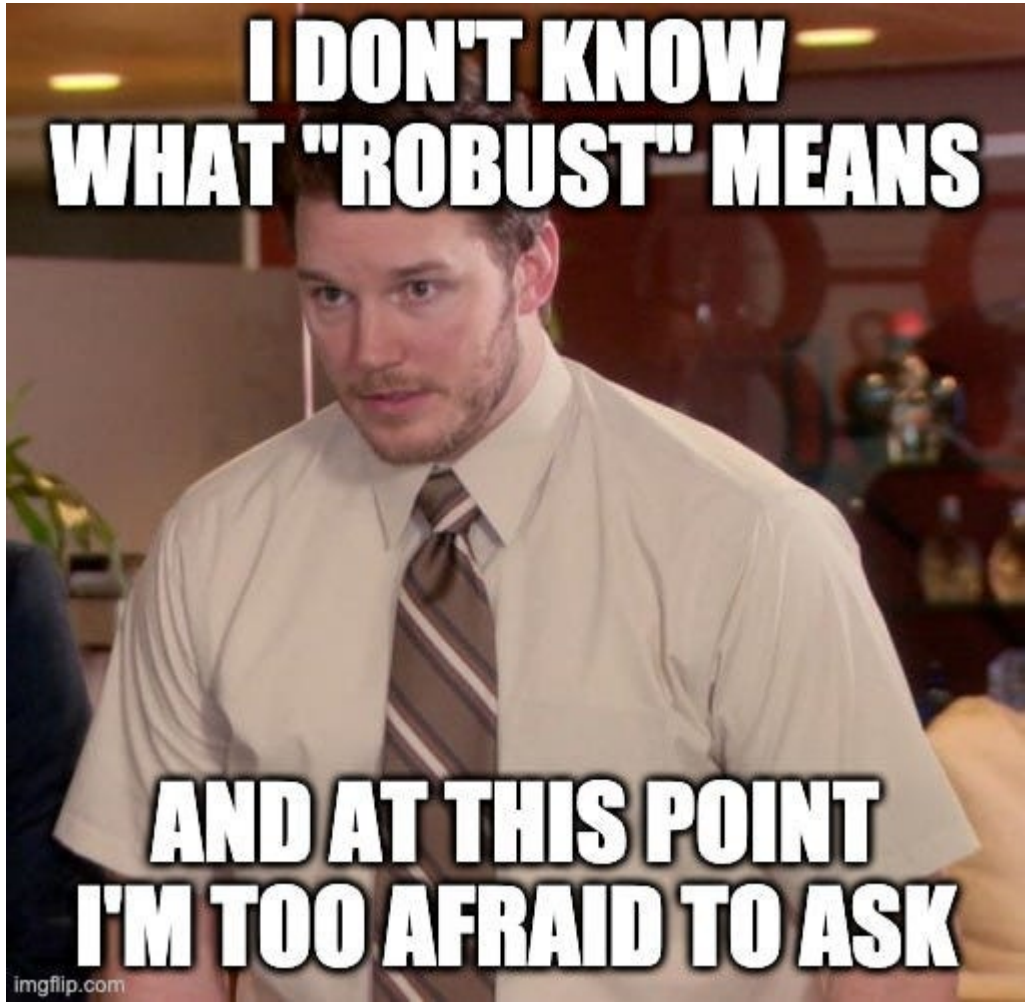
OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

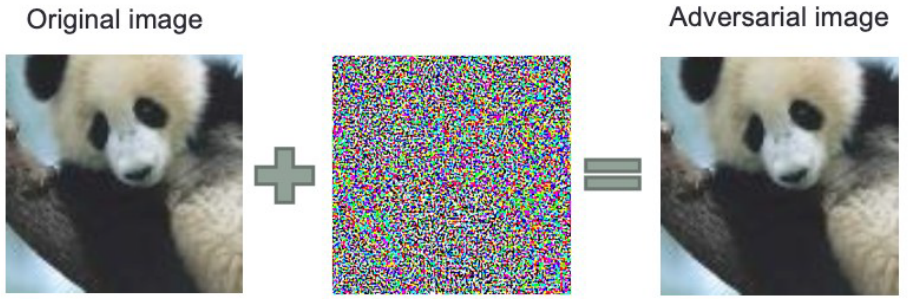
Robustness Definitions

(or the lack thereof)

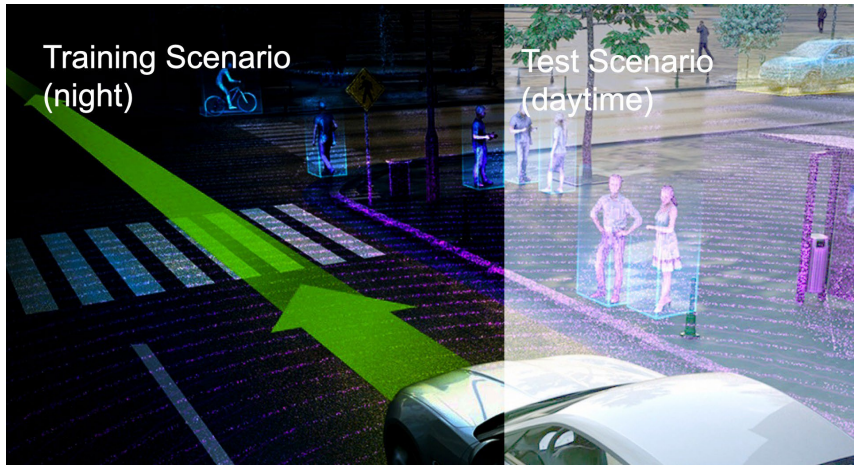


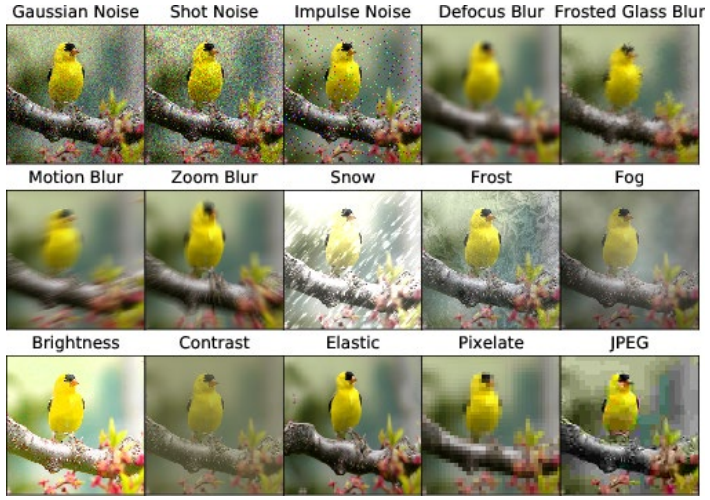
Adversarial vs. Out-of-distribution

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



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





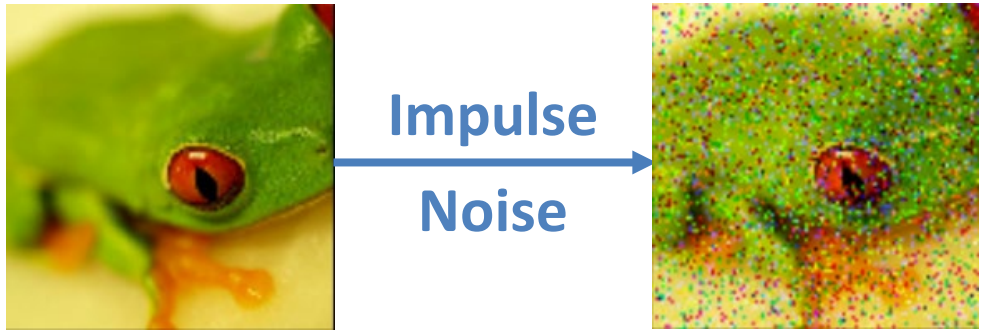
Robustness

Predictor f

$f(x) = y$ for $x \in \mathcal{X}$

Robustness to small perturbations:

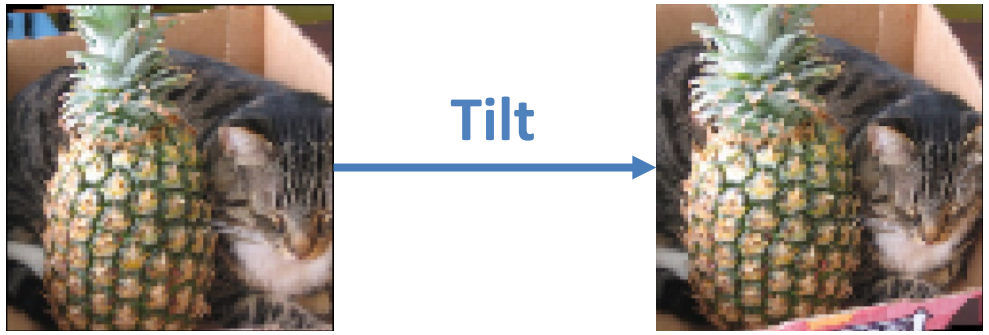
$f(x + \Delta x) = ?$ s.t. $\left|\left|\Delta x\right|\right|_p \leq \epsilon$



Robustness to small transformations:

$f\big(g(x)\big) = ?$ s.t. $\left|\left|g(x) - x\right|\right|_p \leq \rho$

and $g: \mathcal{X} \to \mathcal{X}$



Examples from "ImageNet-C" and "ImageNet-P" datasets – Hendrycks et al. ICLR 2019

Generalization

IID Generalization

Training and Testing inputs sampled independently and identically from a common underlying distribution $\mathcal{X}$

$$x_{train} \sim \mathcal{X}$$
$$x_{test} \sim \mathcal{X}$$

OOD Generalization *aka* Domain Generalization

Training and Testing inputs samples from non-identical distributions $\mathcal{X}_{source}$ and $\mathcal{X}_{target}$

Thus test images are “out-of-distribution” for the classifier that has seen training images.

$$x_{train} \sim \mathcal{X}_{source}$$
$$x_{test} \sim \mathcal{X}_{target}$$

Some of Tejas' Research

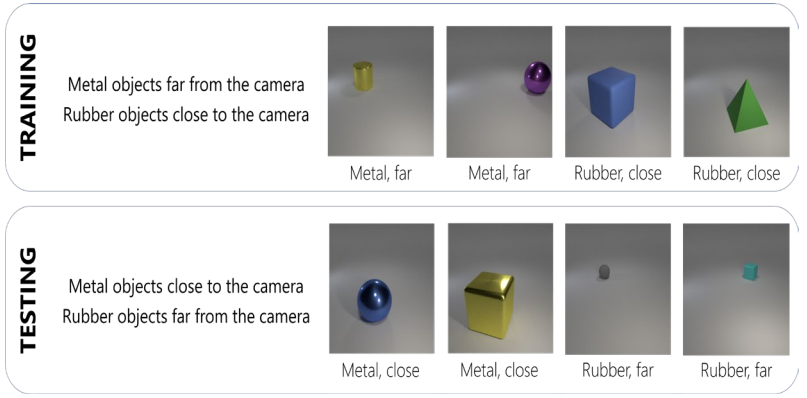
(or why I care about Robust Machine Learning)

Perception & Reasoning with Robustness

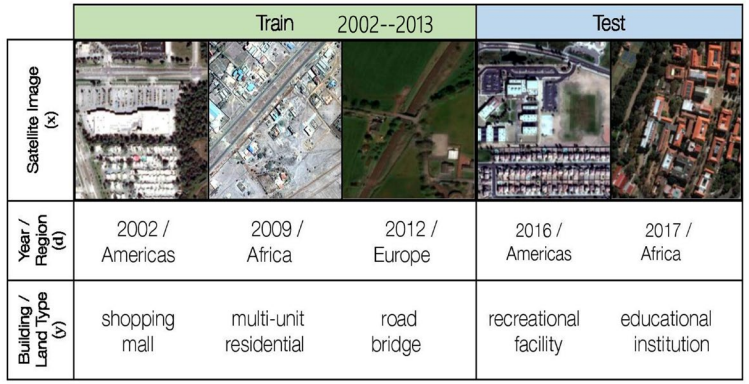
Robust Image Recognition

Dealing with Style Shift

Dealing with Attribute-Level Shift

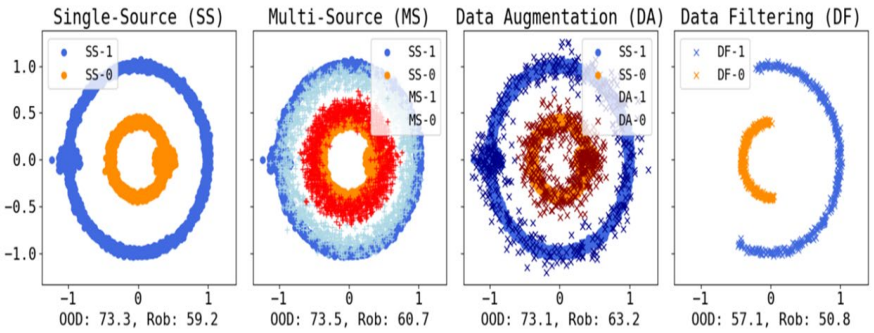


Robust Overhead image recognition

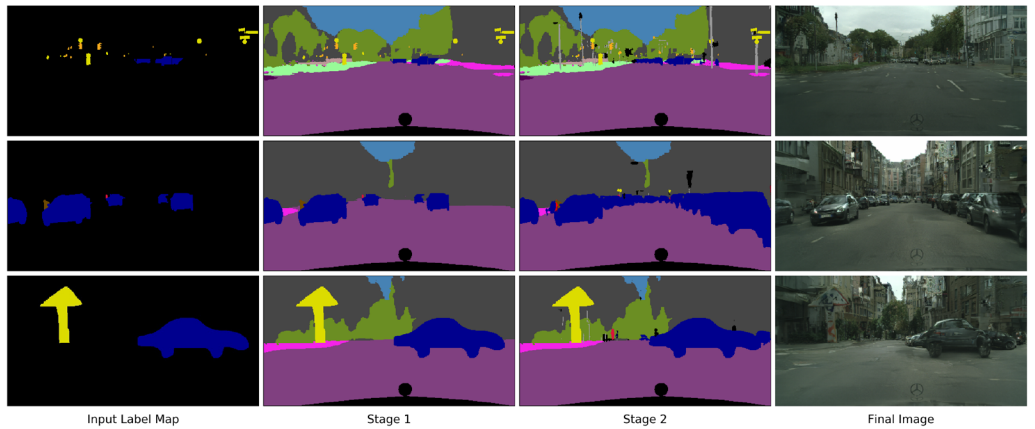


Gokhale AAAI'21;
Gokhale ACL'22;
Gokhale WACV'23;
Cheng ICCV '23;
Wisdom arxiv 2023;
Kulkarni CVPR-W'21

Effects of multiple data sources on OOD and adversarial robustness



Scene Completion for Missing Sensor/Modality



Perception & **Reasoning** with Robustness

Robust Visual Reasoning (Visual QA, Video Captioning, V&L Inference)

V&L Robustness: Logical, Semantic, Spatial
(use additional knowledge sources and sensors)

Is the fork **NOT** on the plate?

yes 94.785%
no 5.215%

Negation

Is the fork on the plate **AND** is the food made of eggs?

no 97.855%
yes 2.144%

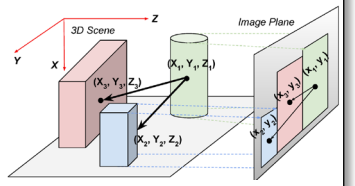
Conjunction

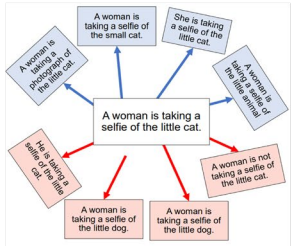
Is the fork on the plate **OR** is the food made of eggs?

no 32.221%
scrambled 17.040%

Disjunction

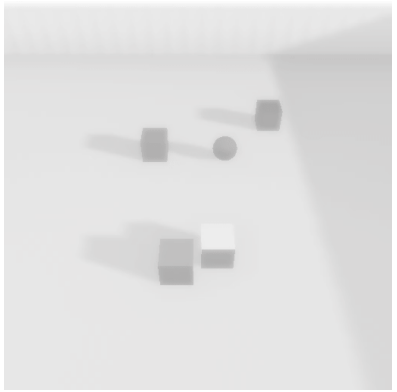
Question	Answer
Is that a giraffe or an elephant?	Giraffe
Who is feeding the giraffe *behind* the man?	Lady
Is there a fence *near* the animal *behind* the man?	Yes
On which side of the image is the man?	*Right*
Is the giraffe *behind* the man?	Yes





f → $\hat{y}_e$

Understanding Agent Actions in Videos with Commonsense, Counterfactual and Physics-Based Reasoning

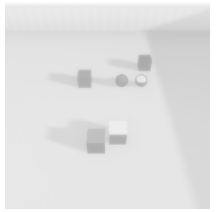


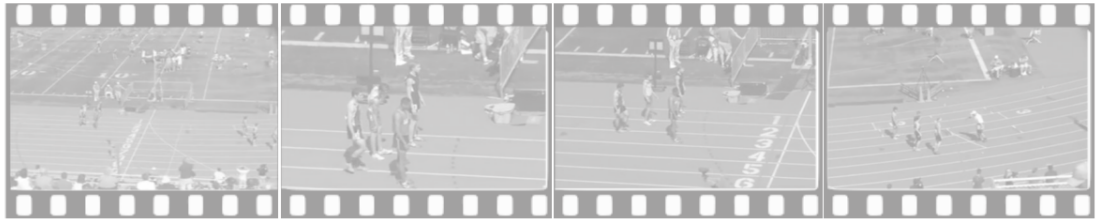
Counterfactual Question
What will happen if the yellow cube is **removed** ?

(A) Purple Cube will collide with brown cube

Planning Question
How can the collision between yellow and purple cube be stopped?

(A) **Add** teal sphere to the right of purple sphere





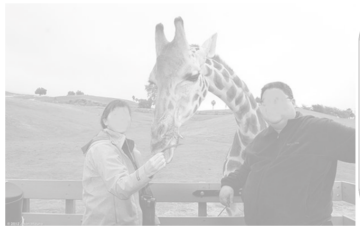
Conventional Caption	Group of runners get prepared to run a race.
Commonsense-Enriched Caption	In order to win a medal, a group of runners get prepared to run a race. As a result they are congratulated at the finish line. They are athletic.
Commonsense Question Answering	What happens next to the runners? { Are congratulated at the finish line / become tired

Gokhale ECCV '20; Gokhale EMNLP'20; Gokhale ACL'21;
Fang EMNLP'20; Banerjee ICCV'21; Patel EMNLP'22

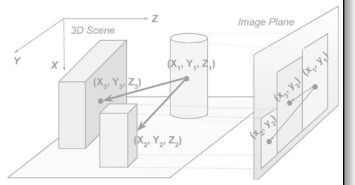
Perception & **Reasoning** with Robustness

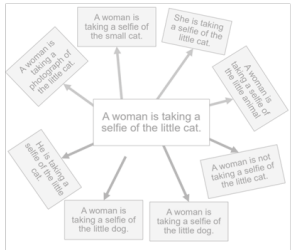
Robust Visual Reasoning (Visual QA, Video Captioning, V&L Inference)

V&L Robustness: Logical, Semantic, Spatial
(use additional knowledge sources and sensors)



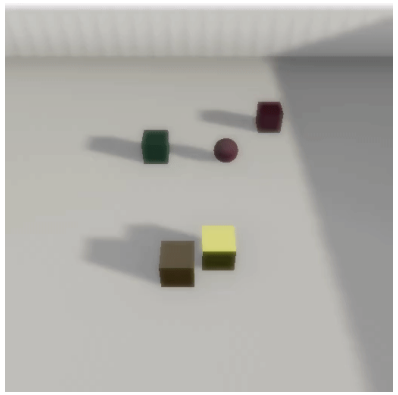
Question	Answer
Is that a giraffe or an elephant?	Giraffe
Who is feeding the giraffe *behind* the man?	Lady
Is there a fence *near* the animal *behind* the man?	Yes
On which side of the image is the man?	*Right*
Is the giraffe *behind* the man?	Yes





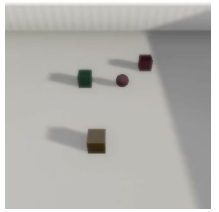
f → $\hat{y}_e$

Understanding Agent Actions in Videos with Commonsense, Counterfactual and Physics-Based Reasoning



Counterfactual Question
What will happen if the yellow cube is **removed** ?

(A) Purple Cube will collide with brown cube



Planning Question
How can the collision between yellow and purple cube be stopped?

(A) **Add** teal sphere to the right of purple sphere

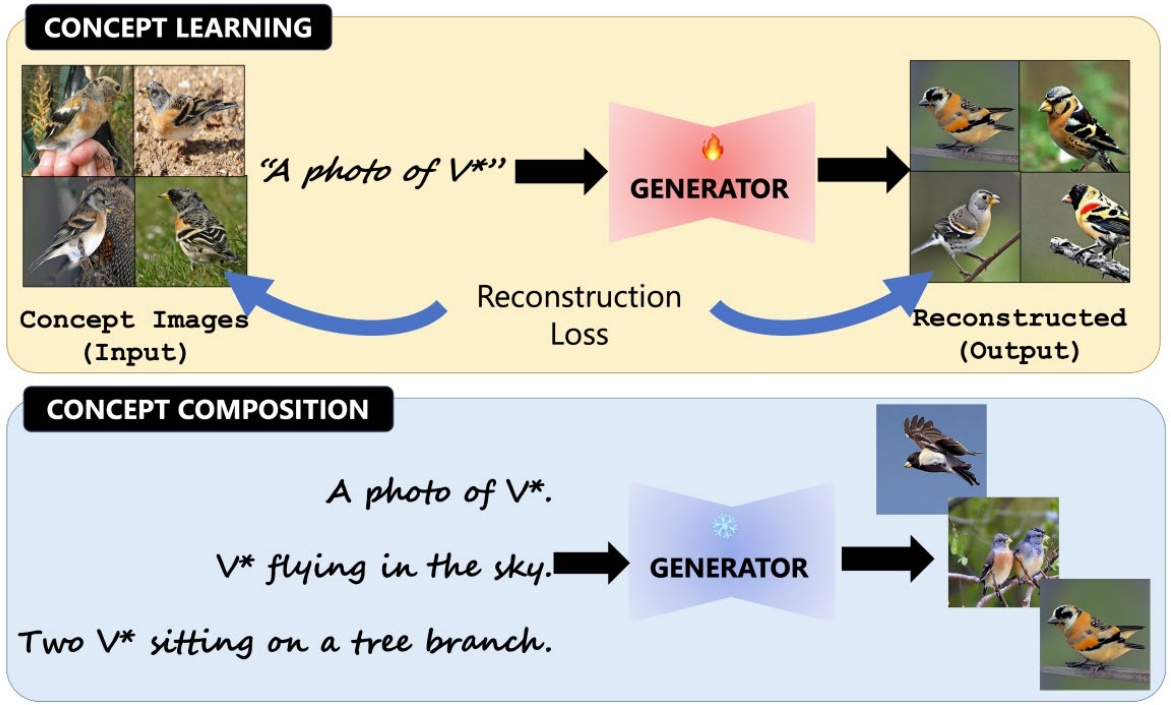
Conventional Caption	Group of runners get prepared to run a race.
Commonsense-Enriched Caption	In order to win a medal, a group of runners get prepared to run a race. As a result they are congratulated at the finish line. They are athletic.
Commonsense Question Answering	What happens next to the runners? { Are congratulated at the finish line / become tired

Gokhale ECCV '20; Gokhale EMNLP'20; Gokhale ACL'21;
Fang EMNLP'20; Banerjee ICCV'21; Patel EMNLP'22

Concept Learning/Distillation in Text-to-Image Generation

Few-Shot Concept Leaning in Text-to-Image Models

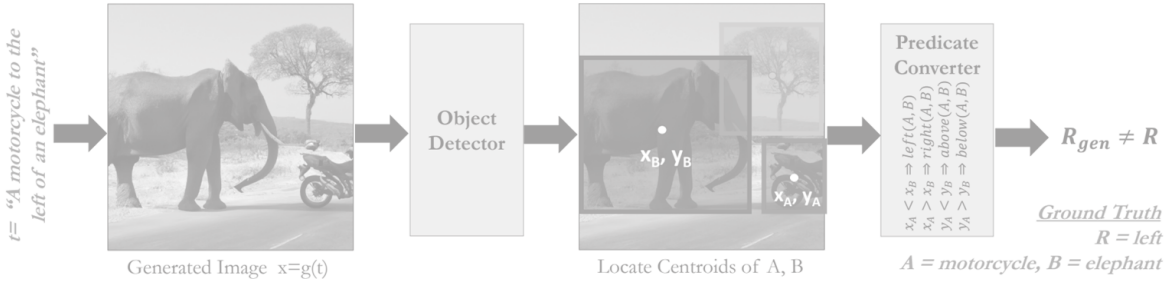
- learn common visual concepts from a few images
- assign semantic meaning (in latent space)
- Reproduce the concept (novel view synthesis)



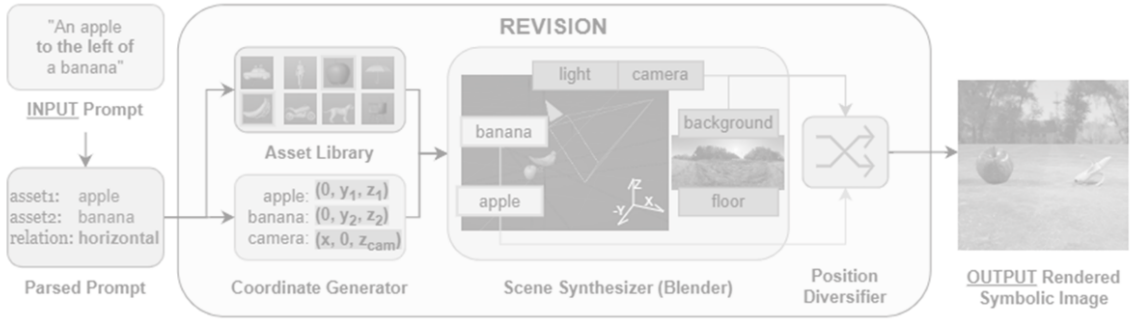
Patel AAAI '24; Gokhale Tech Rep 2022; Chatterjee ECCV '24; Chatterjee ECCV '24

Spatial Reasoning in Text-to-Image Models

- Evaluation Dataset (SR2D) and Metrics (VISOR)



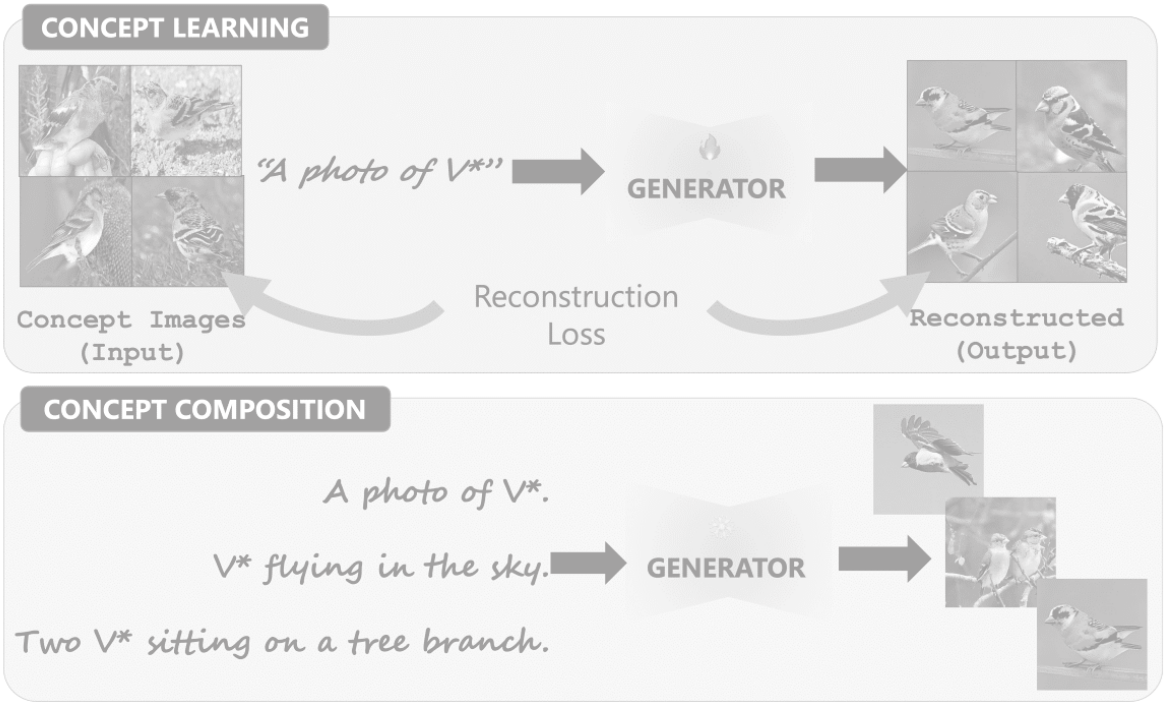
- Improving T2I spatial reasoning with
  1. Recaptioning
     (how can a very small number of specialized "spatial" captions" help finetune T2I models?)
  2. guidance from graphics engines



Concept Learning/Distillation in Text-to-Image Generation

Few-Shot Concept Leaning in Text-to-Image Models

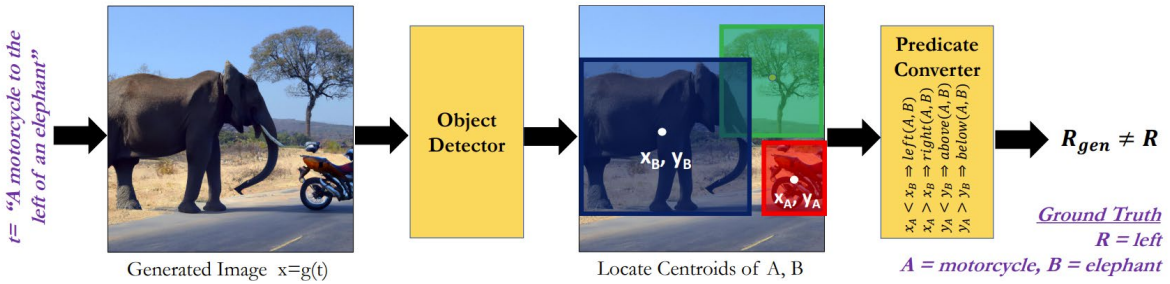
- learn common visual concepts from a few images
- assign semantic meaning (in latent space)
- Reproduce the concept (novel view synthesis)



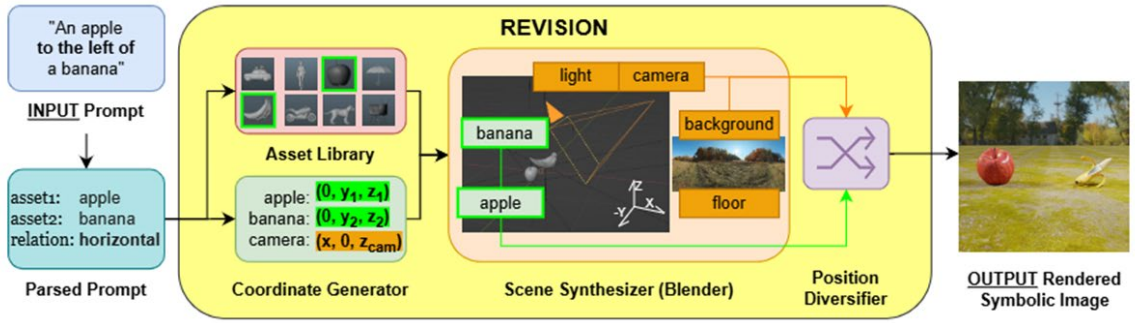
Patel AAAI '24; Gokhale Tech Rep 2022; Chatterjee ECCV '24; Chatterjee ECCV '24

Spatial Reasoning in Text-to-Image Models

- Evaluation Dataset (SR2D) and Metrics (VISOR)



- Improving T2I spatial reasoning with
  1. Recaptioning
     (how can a very small number of specialized "spatial" captions" help finetune T2I models?)
  2. guidance from graphics engines



Next Time:

Machine Learning Review

- Machine Learning Training and Inference Pipeline
- Neural Networks / Deep Learning
- ML methods for visual recognition