CS760 Machine Learning
Ethics and Trust in AI

Kirthi Kandasamy
University of Wisconsin-Madison

May 3, 2023

Slides originally created by Sharon Li

Announcements

- **Course evaluations - due on Friday**
  - **37** have already filled out. Thank you!

- **Grades for HW5 and HW6 will be released this week.**

Artificial Intelligence in Society

Outline

- Bias and Fairness
- Fake Content
- Adversarial robustness
- Privacy

Bias and Fairness

Example 1: Skin color bias in face recognition

https://www.nytimes.com/2020/11/11/movies/coded-bias-review.html

Example 2: Gender Bias in GPT-3

- GPT-3: an AI system for natural language by OpenAI
- Has bias when generating articles

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

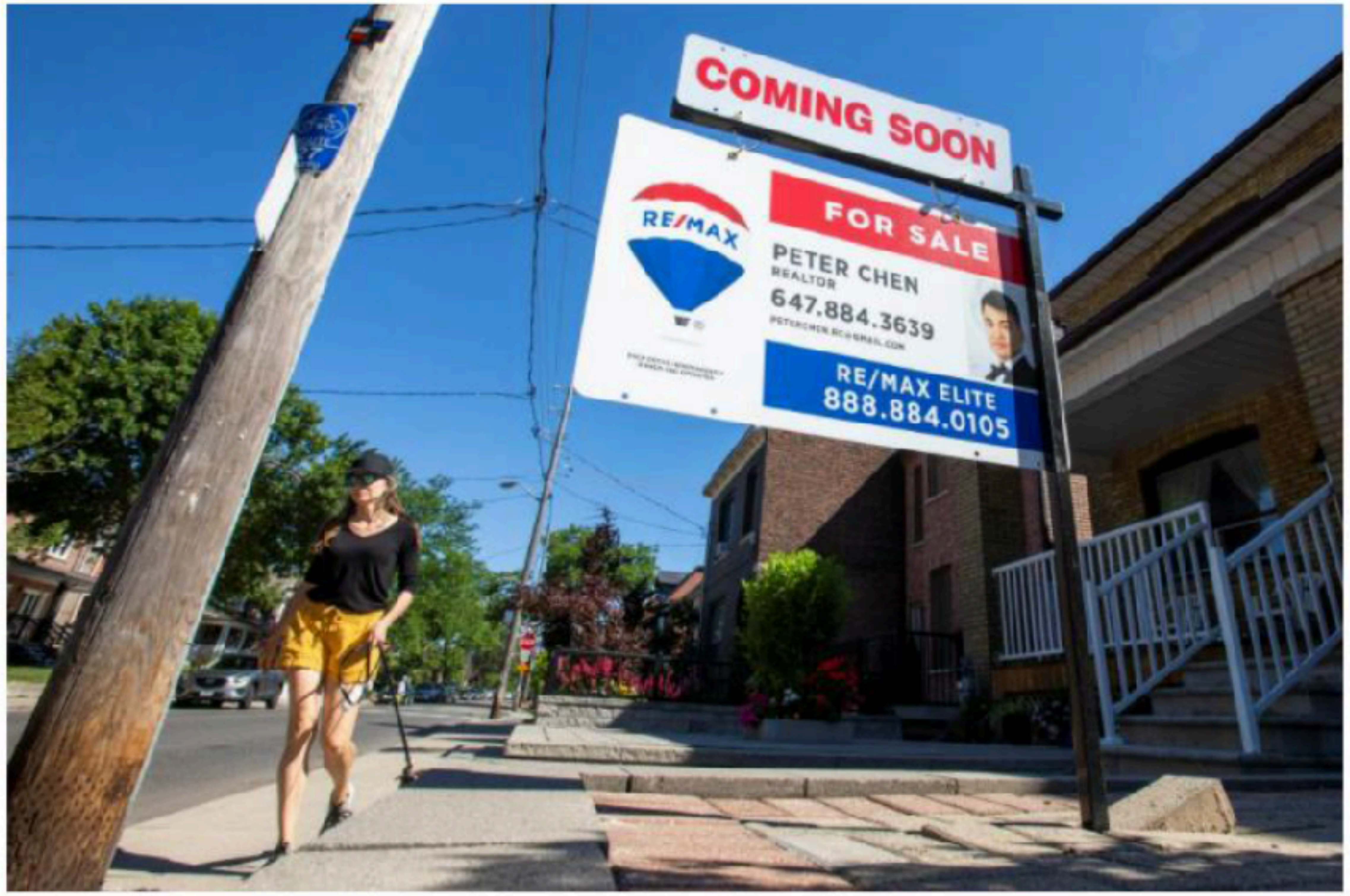
https://arxiv.org/pdf/2005.14165.pdf

Real world consequences

Our test used Amazon Rekognition to compare images of members of Congress with a database of mugshots. The results included 28 incorrect matches.

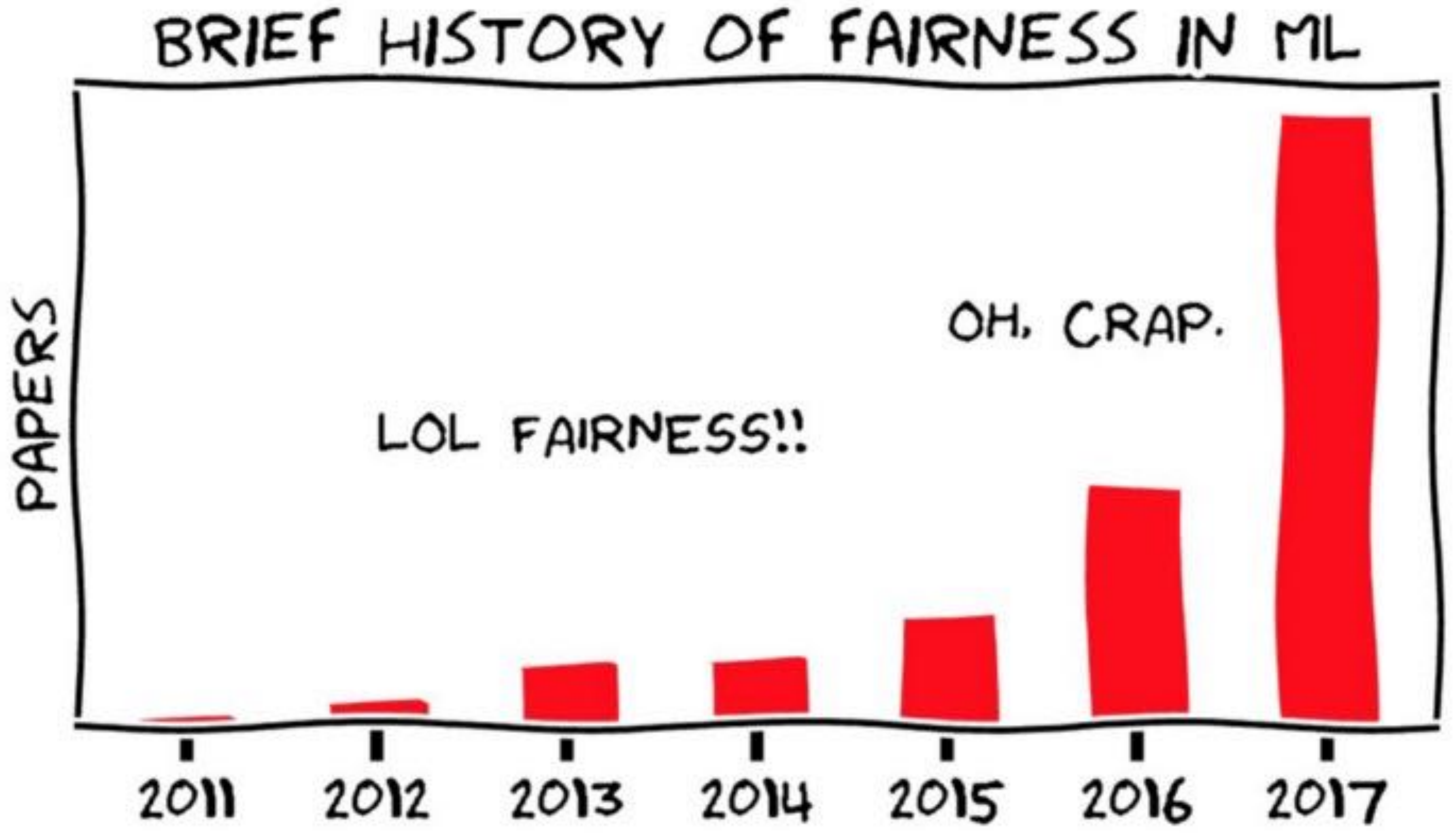
The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28



Credit scoring models can be between 5 and 10 percent less accurate for lower-income and minority homebuyers, new research shows. | Carlos Osorio

https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit



https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Where is the bias from?

- Main reason: the data for training the system are biased
    - Face recognition: training data has few faces of minority people
    - GPT-3: training data (internet text) has the gender bias

Machine learning systems inherit the bias from the training data.

Sources of bias in datasets

- Spurious correlation
    - e.g. the relationship between “man” and “computer programmers” was found to be highly similar to that between “woman” and “homemaker” (Bolukbasi et al. 2016)
- Sample size disparity
    - If the training data coming from the minority group is much less than those coming from the majority group, it is less likely to model the minority group well.
- Proxies
    - Even if sensitive attribute(attributes that are considered should not be used for a task e.g. race/gender) is not used for training a ML system, there can always be other features that are proxies of the sensitive attribute(e.g. neighborhood).

How to mitigate bias?

- **Removing bias from data**
  - Collect representative data from minority groups
  - Remove bias associations (GPT-3: remove the sentences with the gender-biased association)
- **Designing fair learning methods**
  - Add fairness constraints to the optimization problem for learning

Ignore all irrelevant and protected attributes

Group fairness

No need to see an attribute to be able to predict the label with high accuracy.

[Sagawa et al. 2019]

Group fairness (a.k.a demographic parity)

Equalize two groups **S**, **T** at the level of outcomes

Pr[outcome o | **S**] = Pr[outcome o | **T**]

"Fraction of people in S getting job offers is the same as in T."

GDRO [Sagawa et al. 2019]

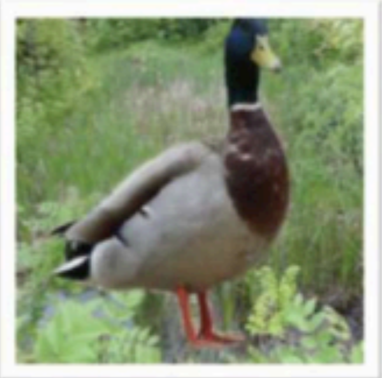
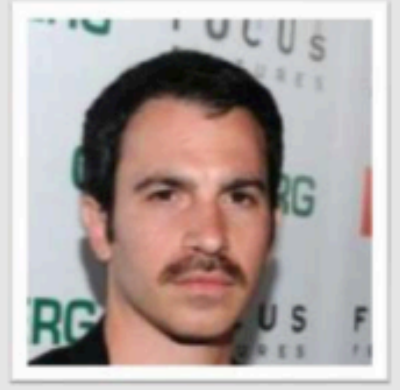
Group Distributionally Robust Optimization

- ERM: $$\hat{\theta}_{\text{ERM}} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\hat{P}}[\ell(\theta; (x,y))]$$
- DRO: $$\hat{\theta}_{\text{DRO}} := \arg\min_{\theta \in \Theta} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim\hat{P}_g}[\ell(\theta; (x,y))] \right\}$$

Minimize the empirical worst-group risk

GDRO [Sagawa et al. 2019]

Group Distributionally Robust Optimization

	Common training examples		Test examples
Waterbirds	y: waterbird a: water background 	y: landbird a: land background 	y: waterbird a: land background 
CelebA	y: blond hair a: female 	y: dark hair a: male 	y: blond hair a: male 
MultiNLI	y: contradiction a: has negation (P) The economy could be still better. (H) The economy has never been better.	y: entailment a: no negation (P) Read for Slate's take on Jackson's findings. (H) Slate had an opinion on Jackson's findings.	y: entailment a: has negation (P) There was silence for a moment. (H) There was a short period of time where no one spoke.

GDRO [Sagawa et al. 2019]

Group Distributionally Robust Optimization

		Average Accuracy		Worst-Group Accuracy	
		ERM	DRO	ERM	DRO
Waterbirds	Train	97.6	99.1	35.7	97.5
	Test	95.7	96.6	21.3	84.6
CelebA	Train	95.7	95.0	40.4	93.4
	Test	95.8	93.5	37.8	86.7

ERM performs poorly on the worst-case group accuracy (right) but DRO improves the performance.

Group fairness can be manipulated by bad actors

Malicious vendor wants to sell a high-fee exclusive credit card **only** to people who have purple skin, not people with green skin

- Target 500 high income people with purple skin
- Target 500 low income people with green skin

Yet, group fairness between purple and green skin

Individual Fairness

Treat *Similar* Individuals *Similarly*

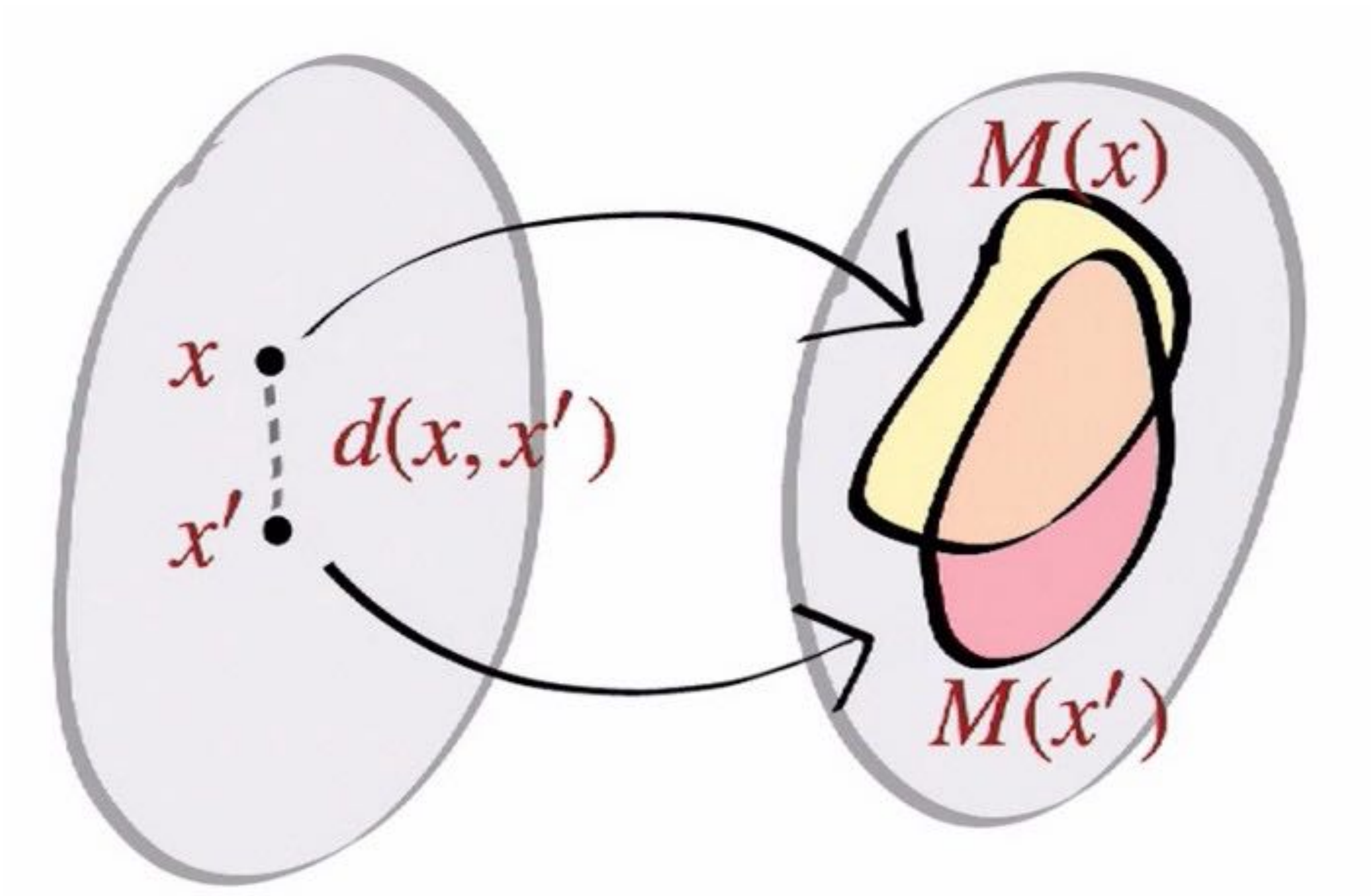
Similar for the purpose of the classification task

Similar distribution over outcomes

Formalize Individual Fairness

$M : x \to \Delta(O)$ Maps each individual example to a distribution of outcomes

$D(M(x), M(x')) \leq d(x, x')$ Where *d* and *D* are two distance functions



Fake Content

Example 1: Fake face Images by GAN

- Which are real/fake? https://thispersondoesnotexist.com/




https://www.youtube.com/watch?v=cQ54GDm1eL0

Example 2: Fake Obama Video

Example 3: fiction Generated by GPT-3

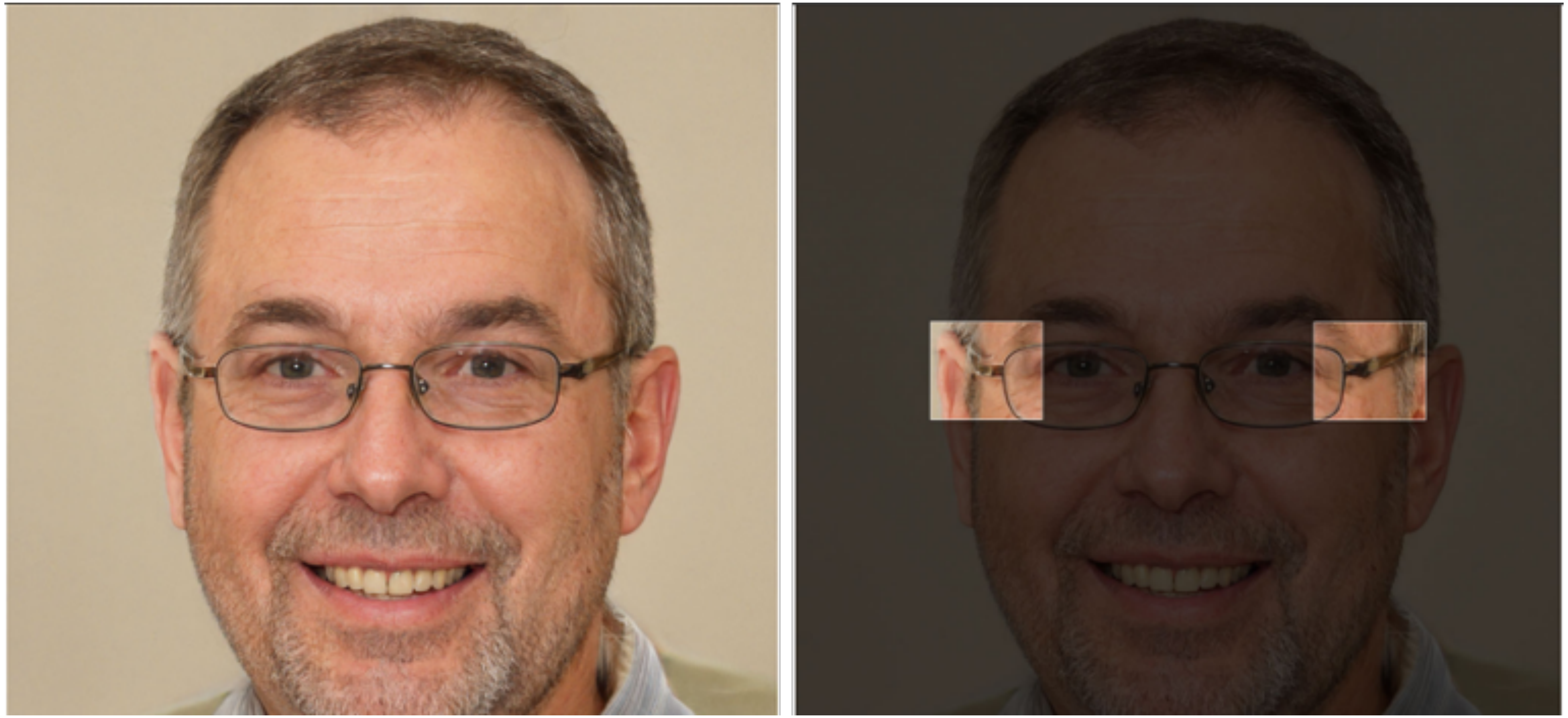
- Completing a prompt from "Harry Potter and the Methods of Rationality":

> "… If there were any other monster that could defeat you as easily as that one, then you would have died of it long ago. That monster is stupidity. And that is why, my young apprentices, you must never, never, NEVER use the Killing Curse on anything without a brain!"
> Professor Quirrell was now leaning on Harry's desk.
> Professor Quirrell stared straight into the eyes of every single student.
> "The Killing Curse is too good for something without a brain. You will be fighting brains, or something near enough that makes no real difference. You will not be fighting trolls. You will not be fighting Dementors. The Killing Curse is no tool for anything less than the third most perfect killing machine in all Nature. If you are not prepared to use it against a mountain troll, then you are not prepared to use it at all. Now. Pay attention to yourselves as I cast a simple spell. Listen to your own thoughts as I tell you how stupid you are."
> Professor Quirrell started pointing his wand at the ceiling.
> …"

https://www.gwern.net/GPT-3#harry-potter-and-the-methods-of-rationality

Detecting Fake Content

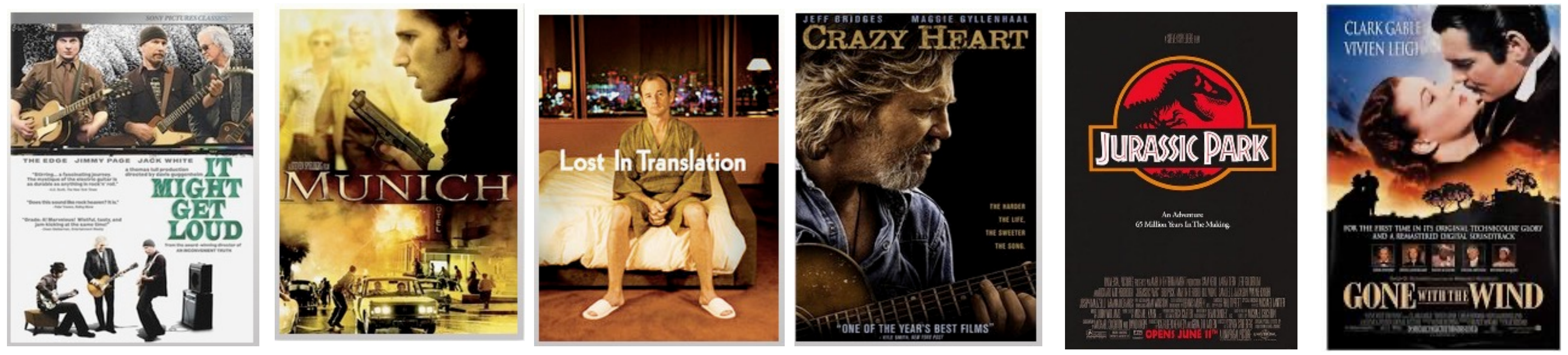
Fake photos/videos can have drawbacks.



Privacy

Example 1: Netflix Prize Competition

- Netflix Dataset: 480189 users x 17770 movies



	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6
Tom	5	?	?	1	3	?
George	?	?	3	1	2	5
Susan	4	3	1	?	5	1
Beth	4	3	?	2	4	2

- The data was released by Netflix in 2006
  - replaced individual names with random numbers
  - moved around personal details, etc

Example 1: Netflix Prize Competition

- Arvind Narayanan and Vitaly Shmatikov compared the data with the non-anonymous IMDb users' movie ratings
- Very little information from the database was needed to identify the subscriber
  - simply knowing data about only two movies a user has reviewed allows for 68% re-identification success

Netflix Cancels Contest After Concerns Are Raised About Privacy

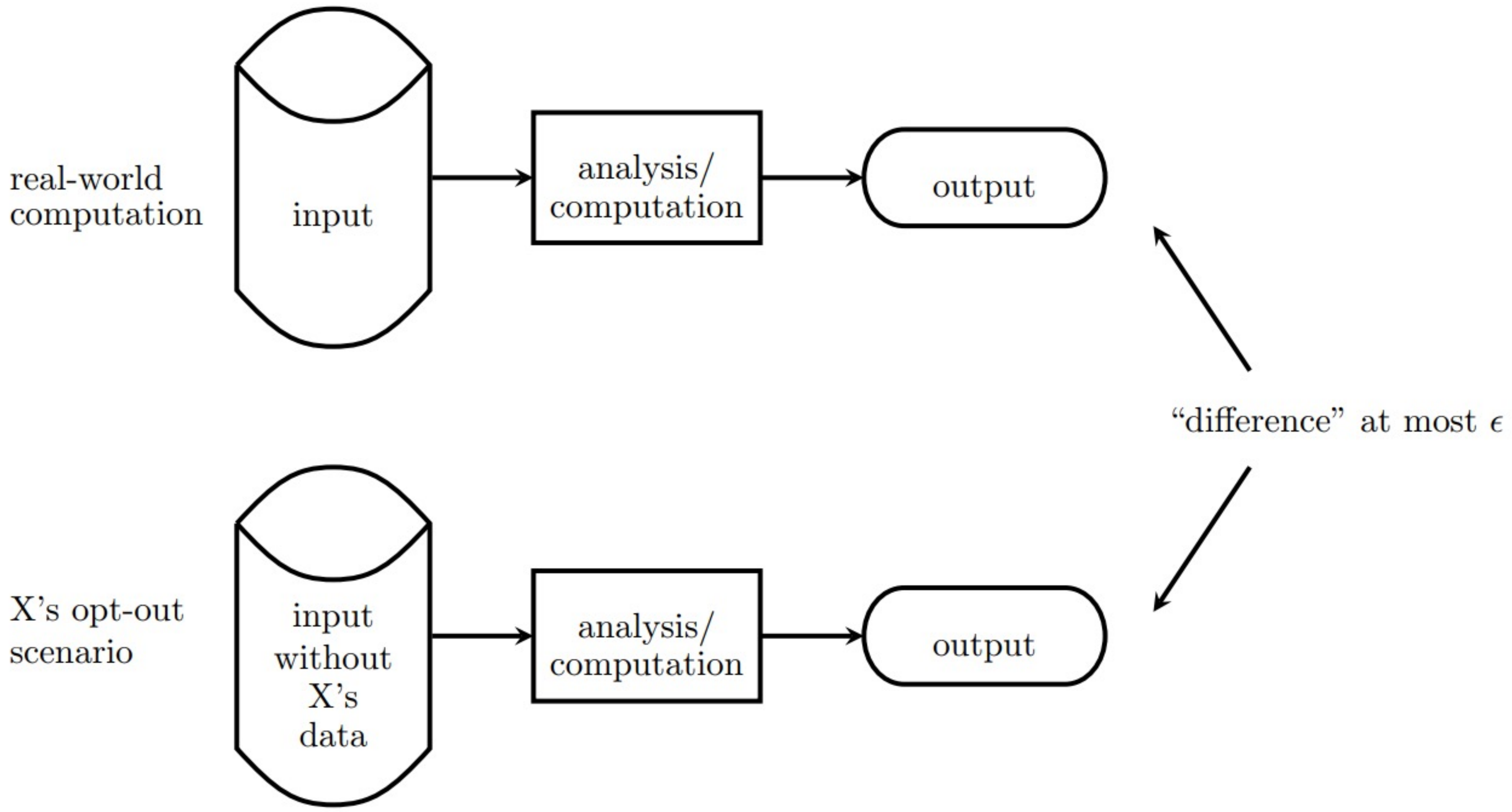
By Steve Lohr

March 12, 2010

https://www.nytimes.com/2010/03/13/technology/13netflix.html

Popular framework: Differential Privacy

- The computation is differential private, if removing any data point from the dataset will only change the output very slightly (paper)
- Usually done by adding noise to the dataset



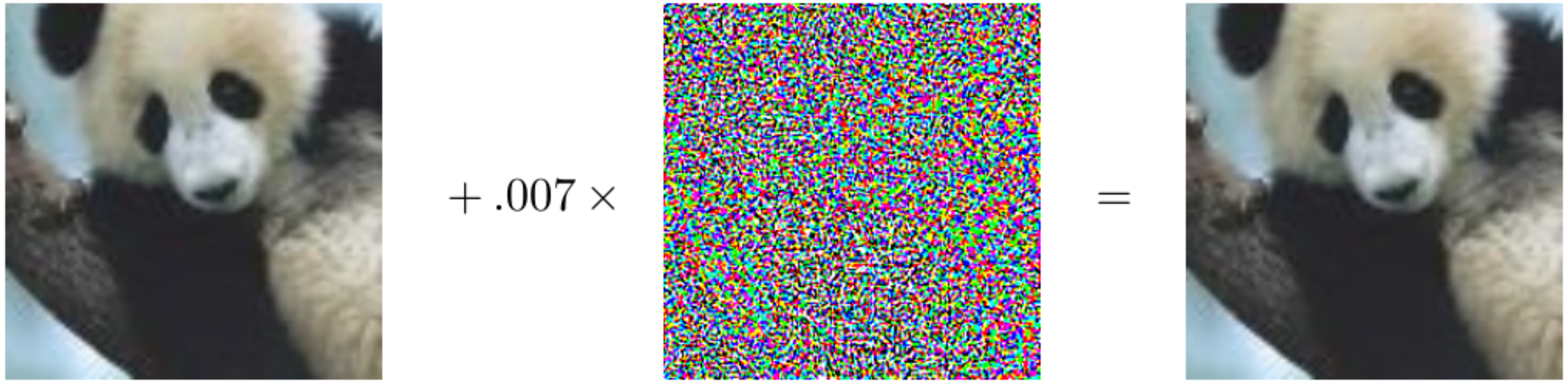
Adversarial Robustness

Adversarial Examples

“Inputs to ML models that an attacker has **intentionally** designed to cause the model to make a mistake”

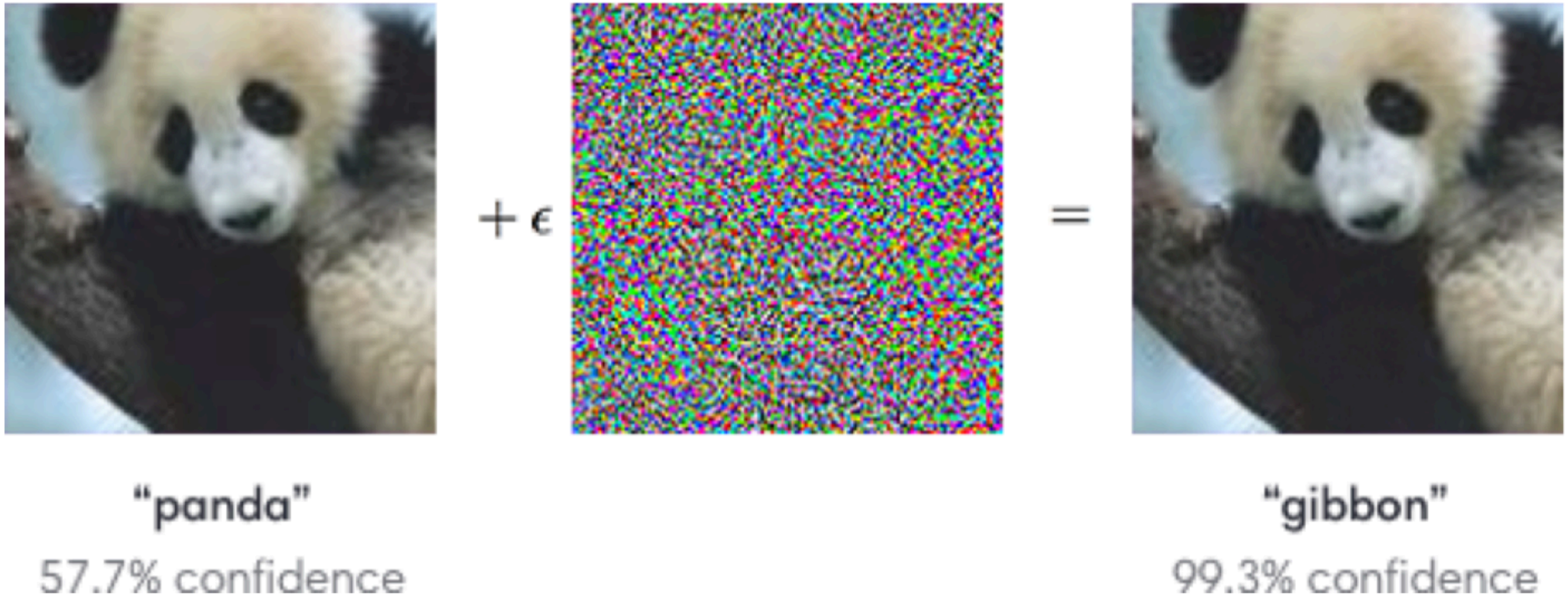
https://blog.openai.com/adversarial-example-research/

Adversarial Examples



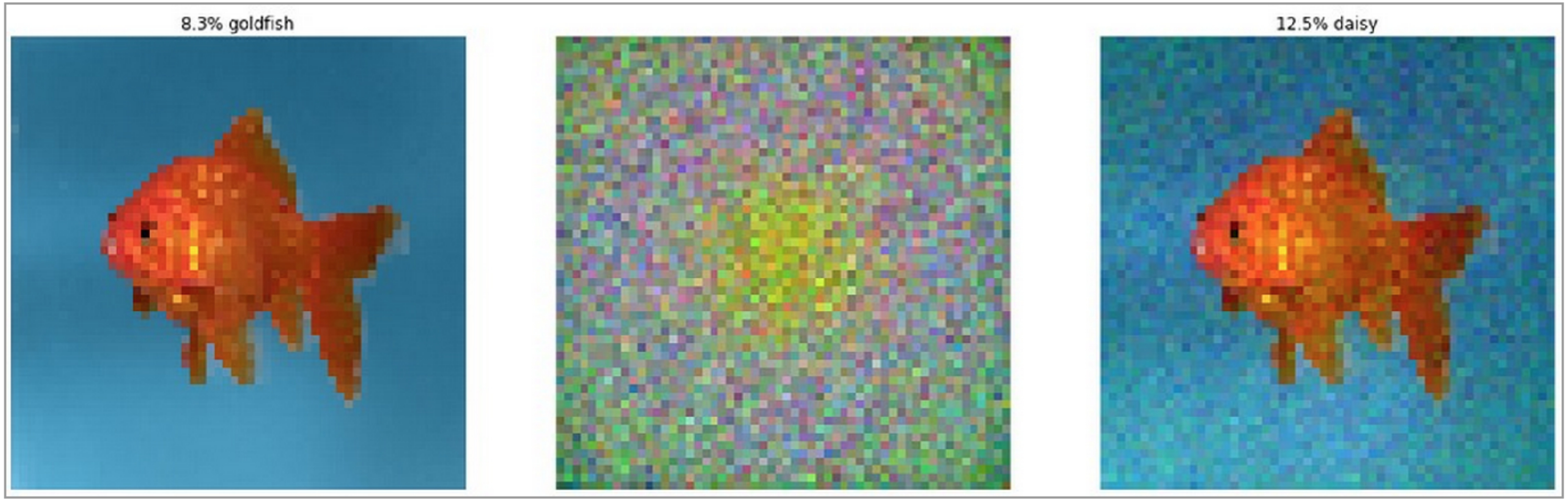
“Adversarial Classification” Dalvi et al 2004: fool spam filter
“Evasion Attacks Against Machine Learning at Test Time” Biggio 2013: fool neural nets
Szegedy et al 2013: fool ImageNet classifiers imperceptibly
Goodfellow et al 2014: cheap, closed form attack

Manipulate Classification



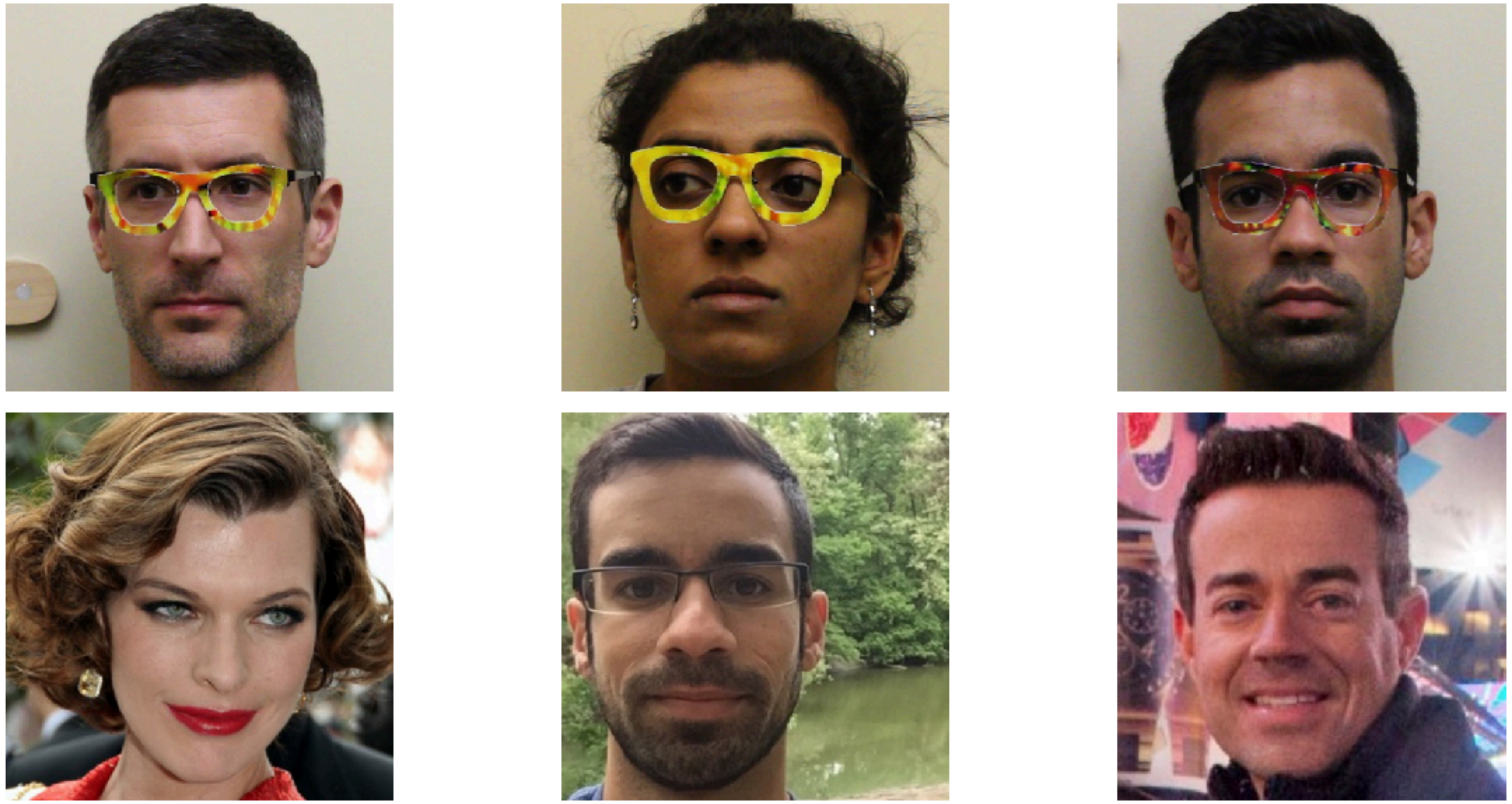
https://openai.com/blog/adversarial-example-research/

Adversarial Examples
Linear Models of ImageNet



(Andrej Karpathy, “Breaking Linear Classifiers on ImageNet”)

Physical Attacks

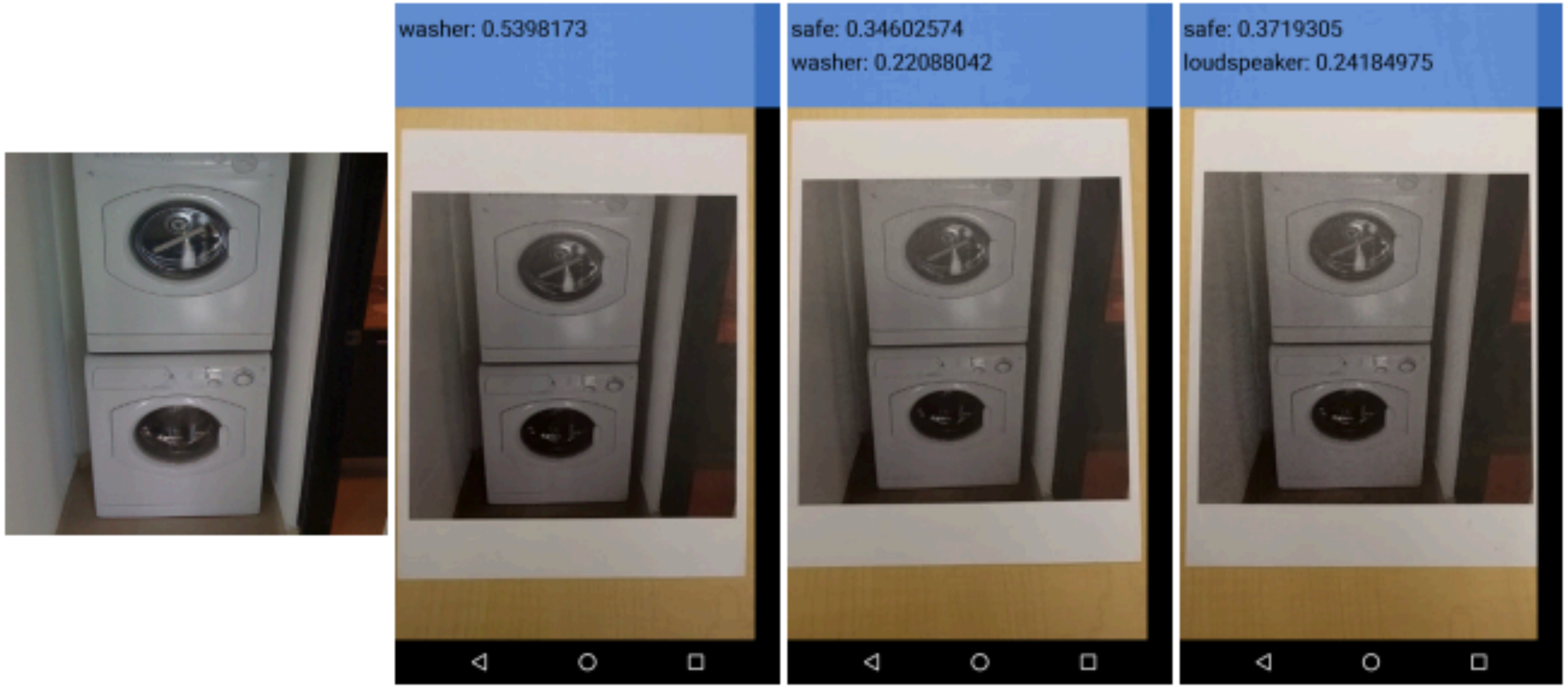


Sharif et al 2016 https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Physical Attacks

Eykholt et al 2017 https://arxiv.org/abs/1707.08945

Physical Attacks



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

(Kurakin et al, 2016)

Adversarial Examples in NLP

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

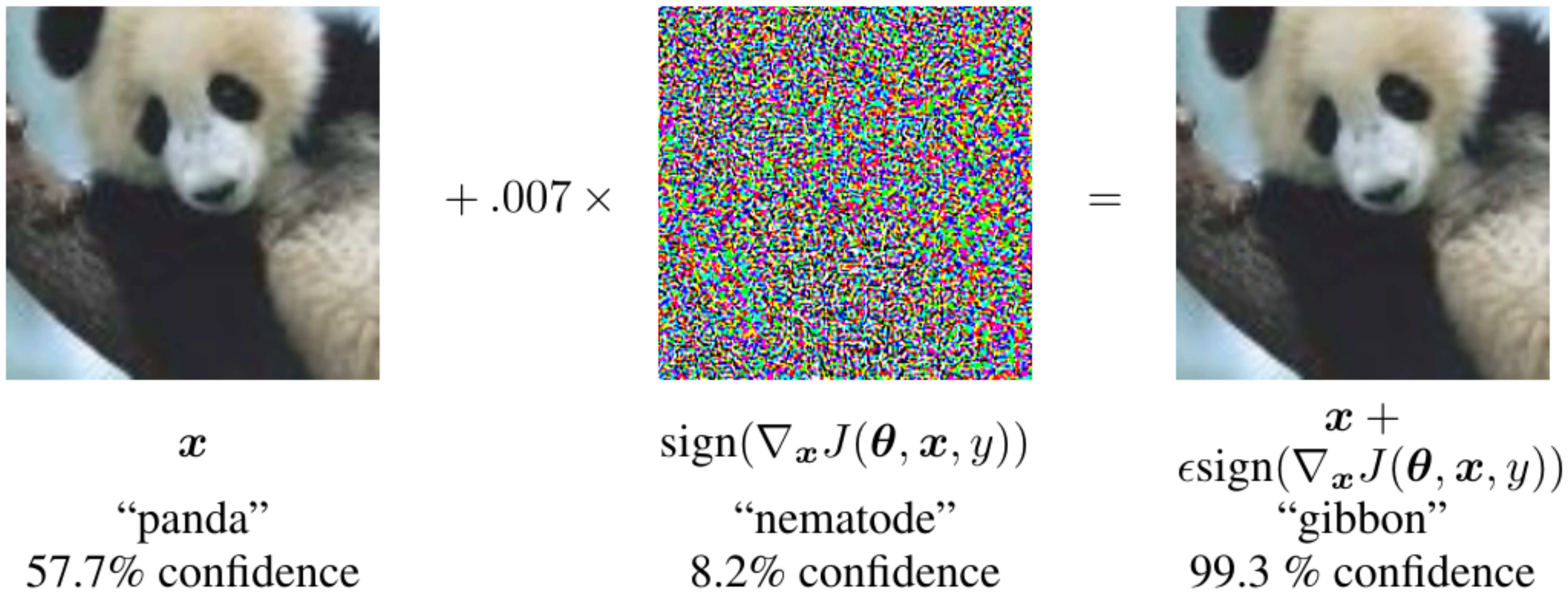
[Jia and Liang, 2017]

Not just for neural networks

- Linear models
  - Logistic loss
  - Softmax loss
- Decision trees
- Nearest neighbors

Generating Adversarial Examples

Simple approach: Fast Gradient Sign Method (FGSM) [Goodfellow et. al 2014]

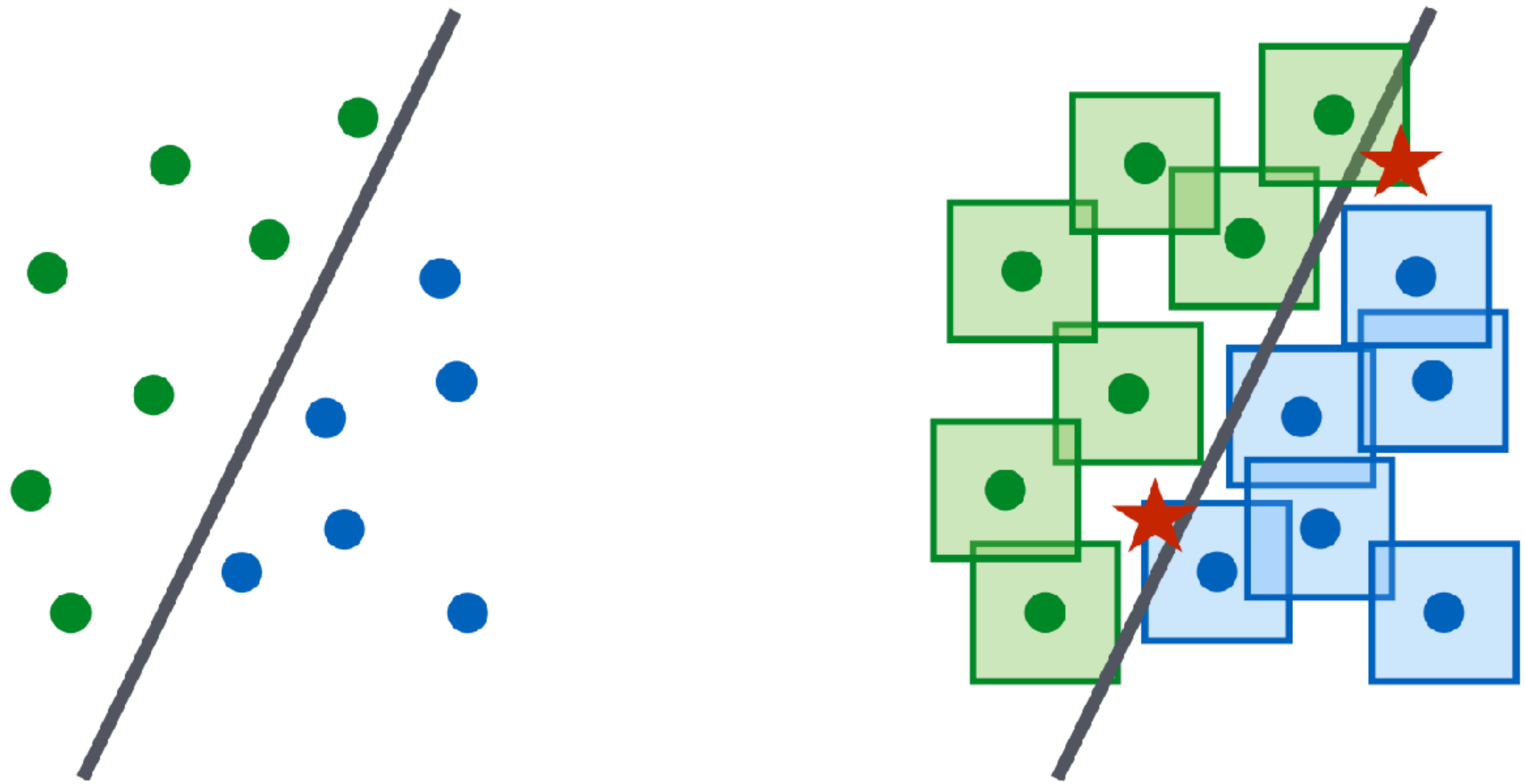


$$||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \epsilon$$

$$\Rightarrow \tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon\text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{x})\right).$$

Test-time Attack

$$\max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$

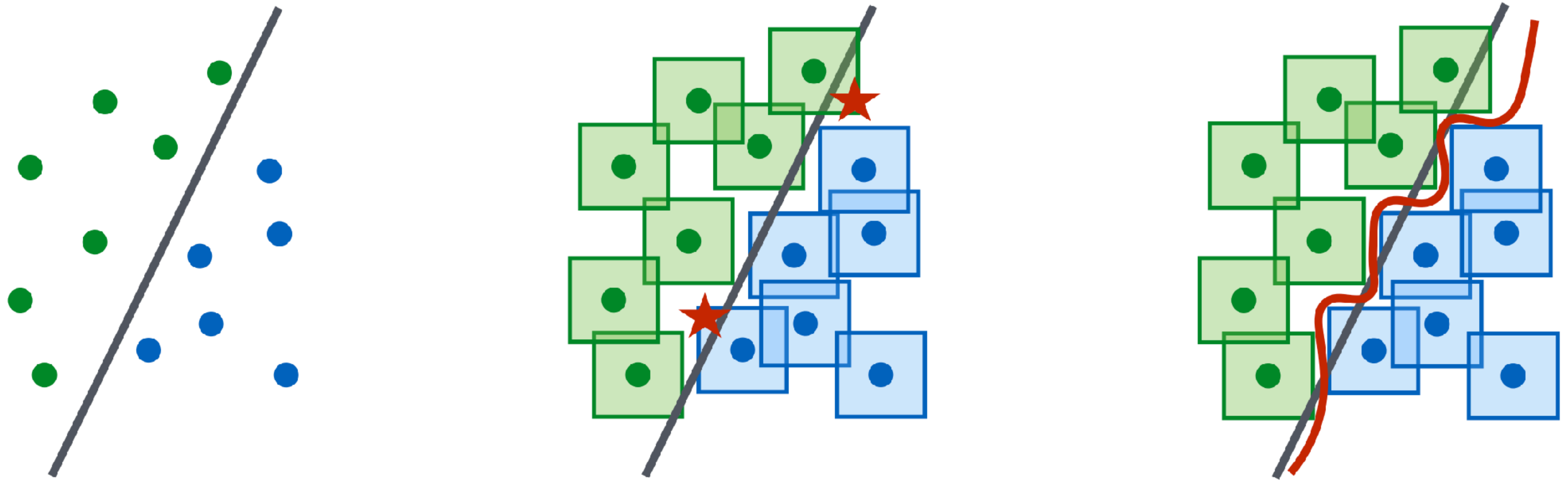


Madry et al 2019 https://arxiv.org/pdf/1706.06083.pdf

(One) Defense against Test-time Attack

Adversarial Training

$$\min_{\theta} \mathbb{E}_D \max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$



Madry et al 2019 https://arxiv.org/pdf/1706.06083.pdf

Defense: Adversarial Training

Labeled as bird

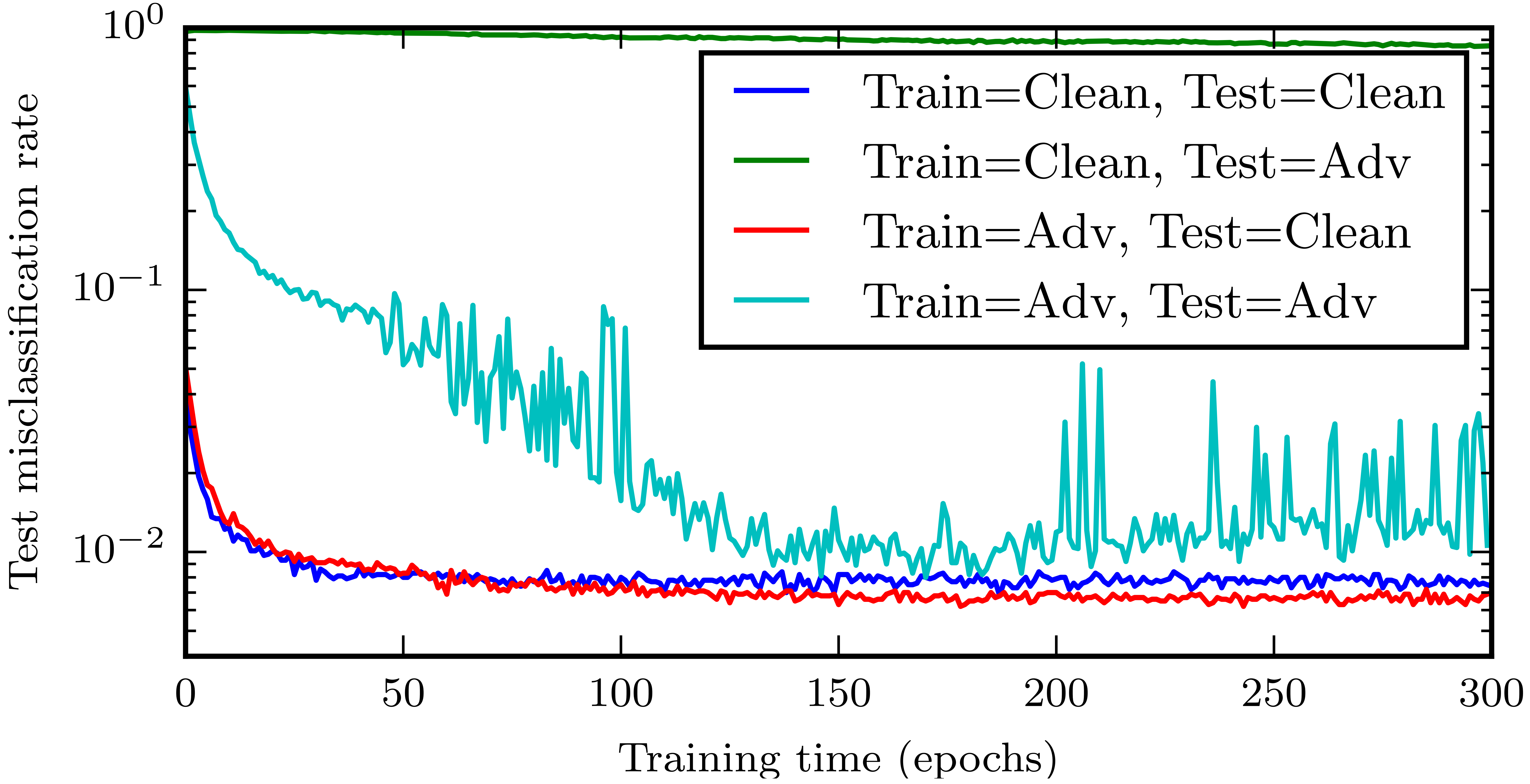
Decrease probability of bird class

Still has same label (bird)

Defense: Adversarial Training

Adversarial training can be viewed as **augmenting** the training data with adversarial examples.

Training on Adversarial Examples

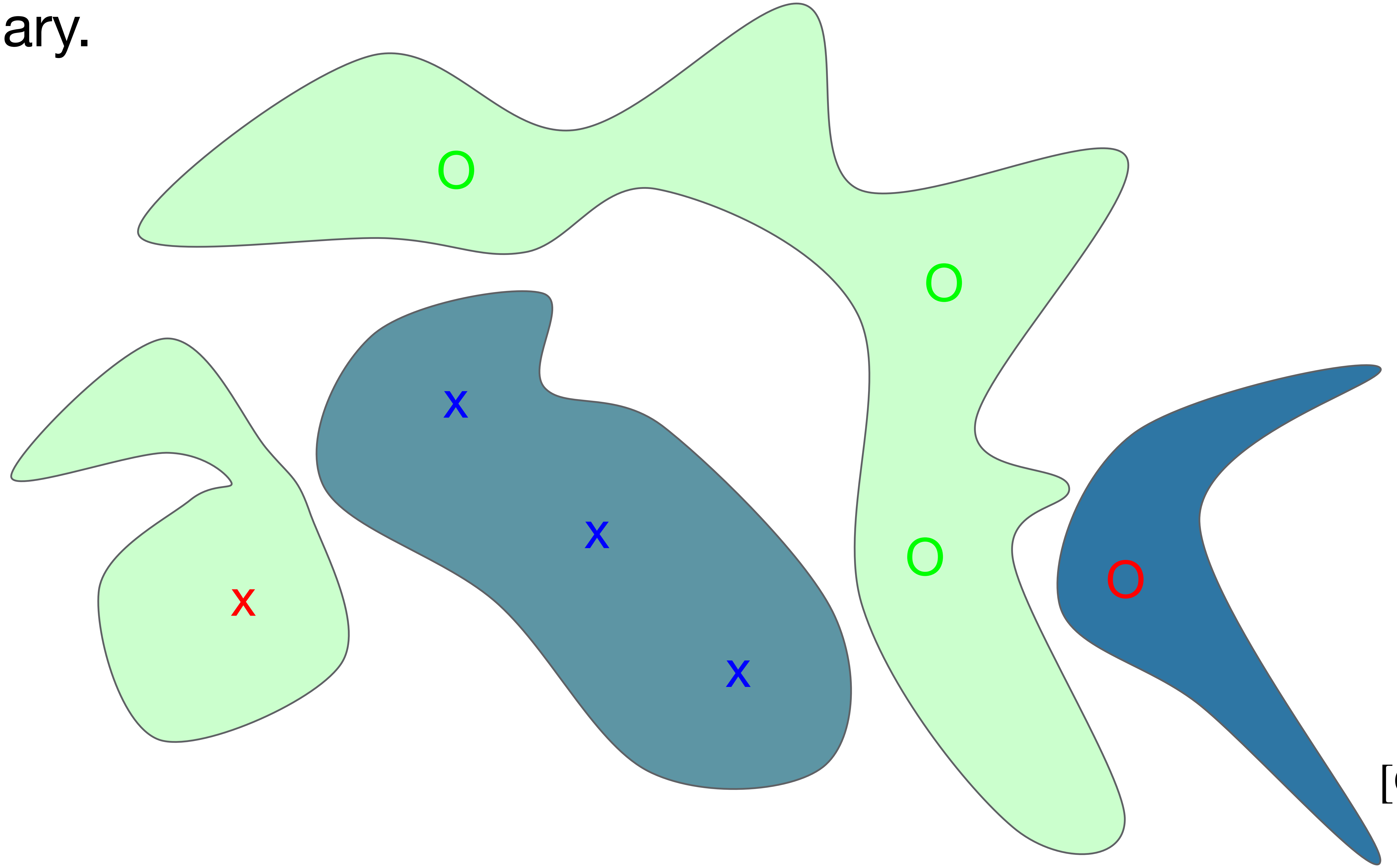


(Goodfellow 2016)

Why ML models are prone to adversary?

Conjecture 1: Overfitting.

Natural images are within the correct regions but are also sufficiently close to the decision boundary.

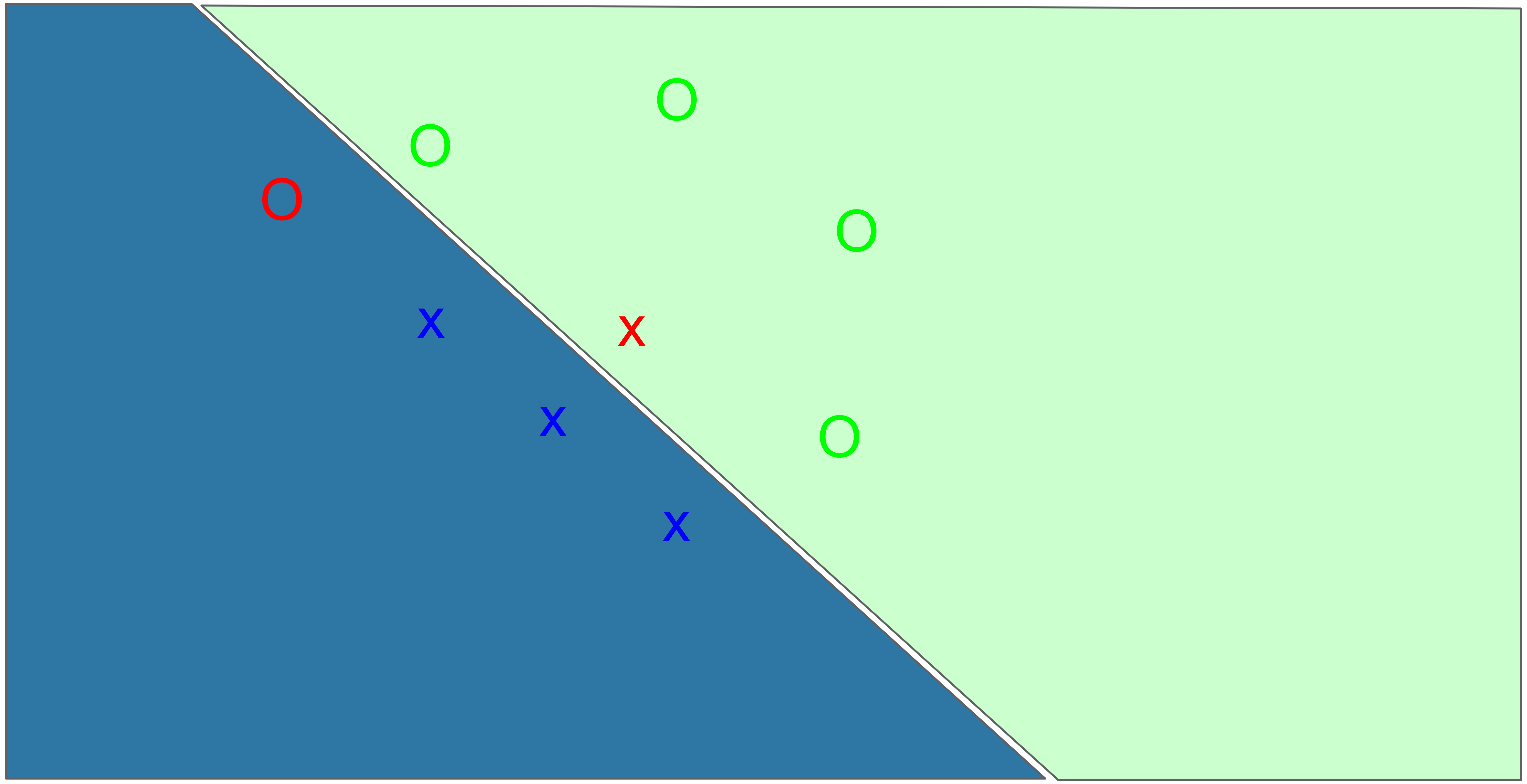


[Goodfellow 2016]

Why ML models are prone to adversary?

Conjecture 2: Excessive Linearity.
Decision boundary for most ML models are (near-) piecewise linear.

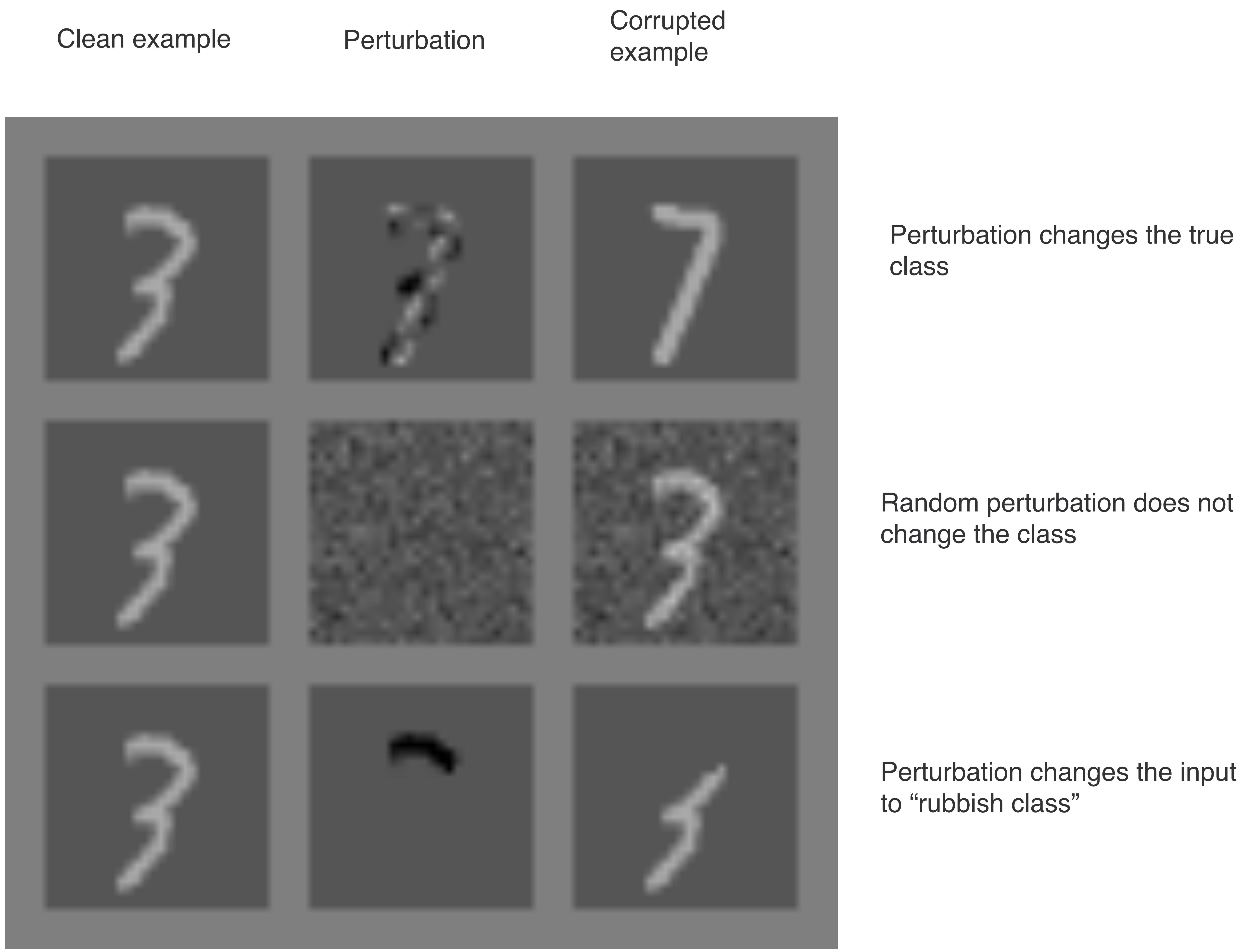
In high dimension, a linear hyperplane is prone to perturbation.



[Goodfellow 2016]

Why ML models are prone to adversary?

Conjecture 3: Small inter-class distances.



All three perturbations have L2 norm 3.96

[Goodfellow 2016]

Summary of Topics in Ethics and Trust in AI

- Bias and Fairness
- Fake Content
- Adversarial robustness
- Privacy

- **Other topics we have not covered**
  - Environmental impact of large ML models
  - “Very advanced” AI: job displacement, use by bad actors

JUSTICE, HEALTH & DEMOCRACY
IMPACT INITIATIVE

How AI Fails Us

Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, E. Glen Weyl

December 1, 2021

MIT Technology Review

SIGN IN SUBSCRIBE

ARTIFICIAL INTELLIGENCE

Geoffrey Hinton tells us why he's now scared of the tech he helped build

"I have suddenly switched my views on whether these things are going to be more intelligent than us."

Acknowledgement:

Some of the slides in these lectures have been adapted/borrowed from materials developed by Anthony Glitter, Yingu Liang, Hanxiao Liu: http://www.cs.cmu.edu/~hanxiaol/slides/adversarial.pdf, Ian Goodfellow