CS540 Introduction to Artificial Intelligence

Ethics and Trust in AI

Yudong Chen
University of Wisconsin-Madison

December 14, 2021

Slides created by Sharon Li [modified by Yudong Chen]

Reminders

- Please complete course evaluation

- Final exam: Dec 20, 2:45-4:45pm CT, online
  - Final is cumulative

Artificial Intelligence in Society

https://techvidvan.com/tutorials/artificial-intelligence-applications/

Outline

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

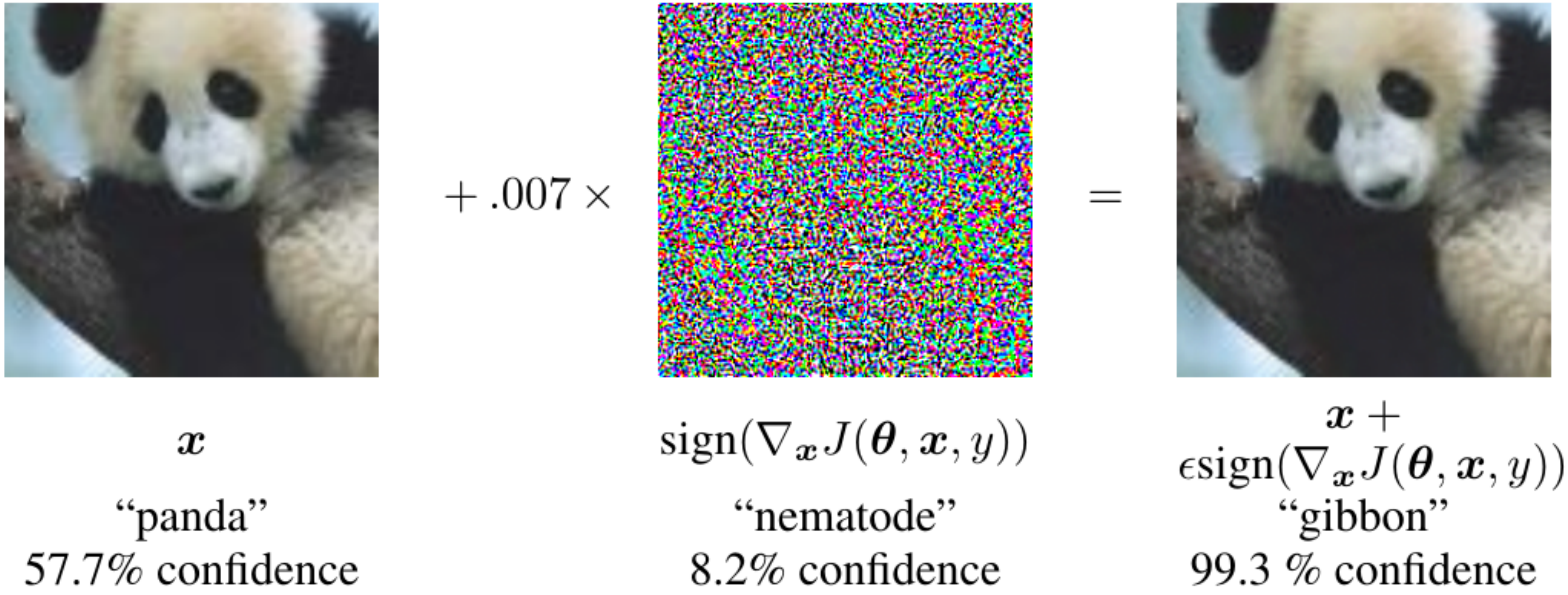
Adversarial Robustness

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



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

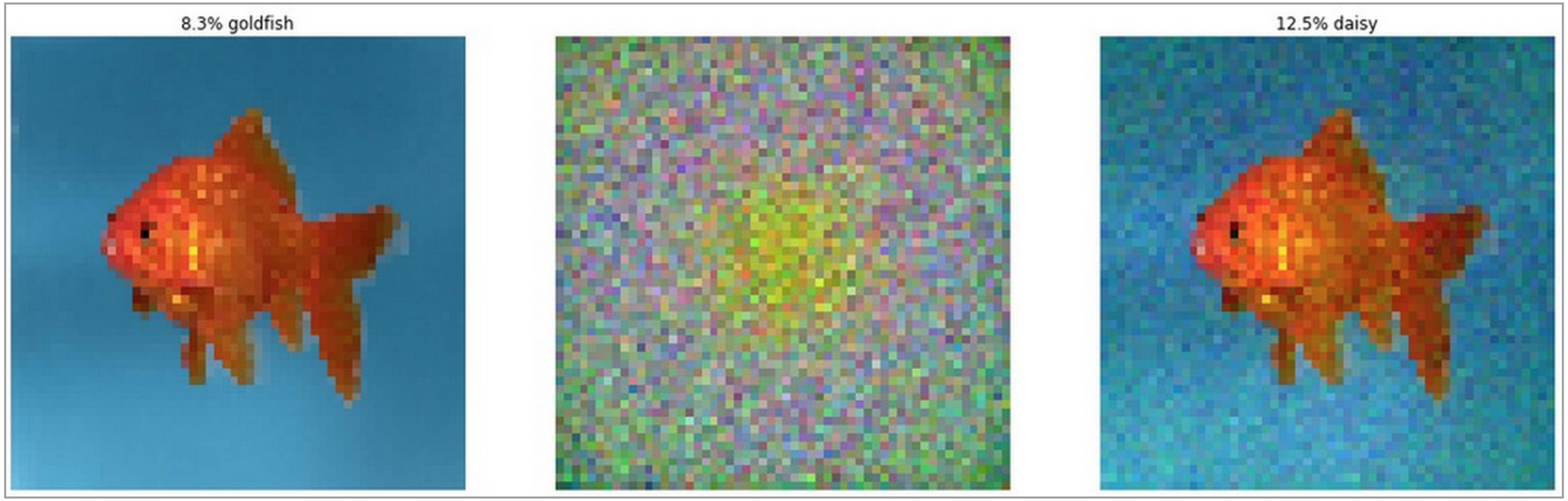
$$||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_{\infty} \leq \epsilon$$

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

Not just for neural nets

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

Adversarial Examples
Linear Models of ImageNet



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

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]

Adversarial Examples in the Physical World

Fooling Image Recognition with Adversarial Examples

https://www.youtube.com/watch?v=piYnd_wYlT8

Defense: Adversarial Training

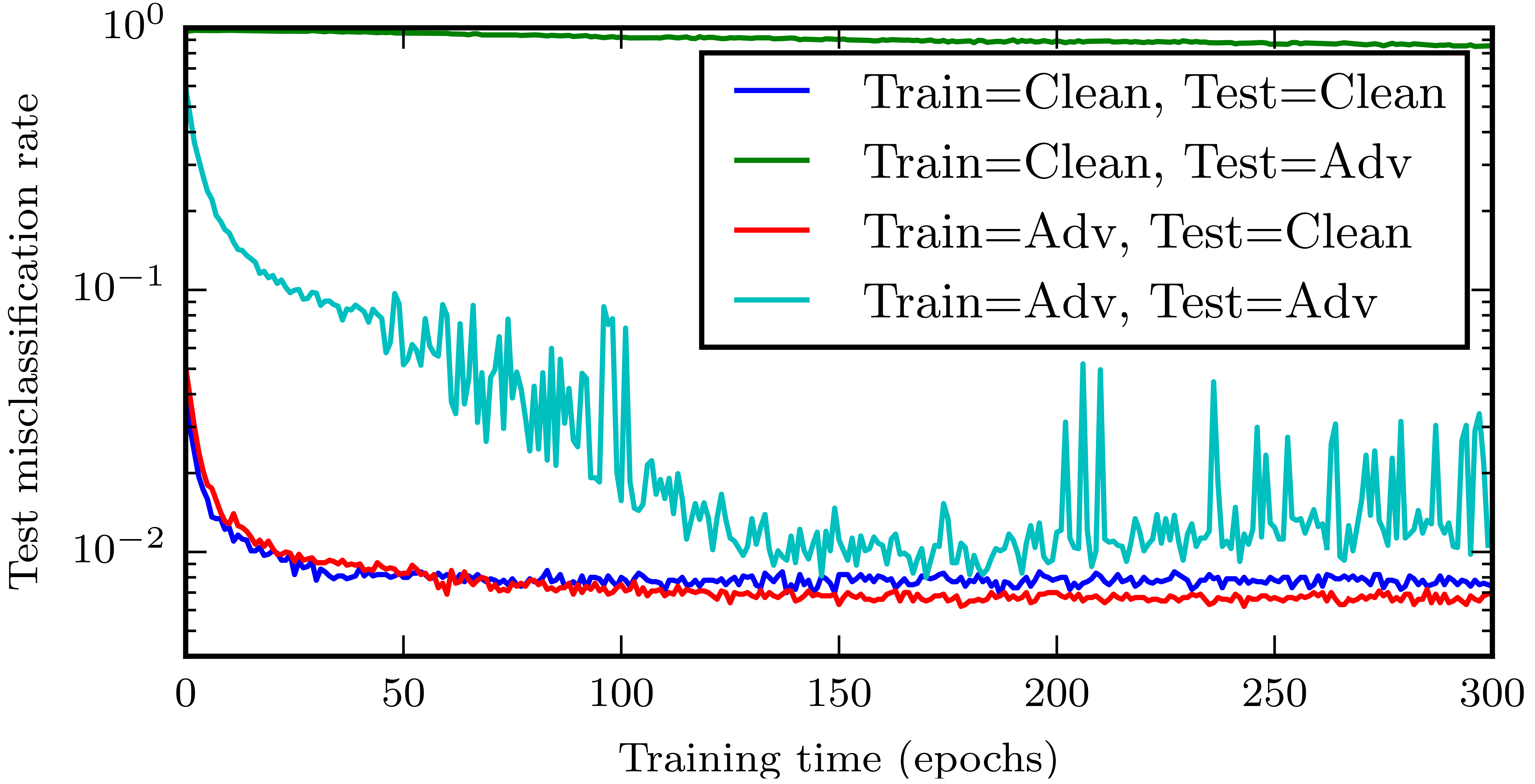
Labeled as bird

Add adversarial perturbation **during training**

Still has same label (bird)

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

Training on Adversarial Examples



(Goodfellow 2016)

Fake Content

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

Example 1: Fake Obama Video

Example 2: Fake face Images by GAN

- Which are real/fake?




Example 2: Fake face Images by GAN

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

Example 3: fiction Generated by GPT-3

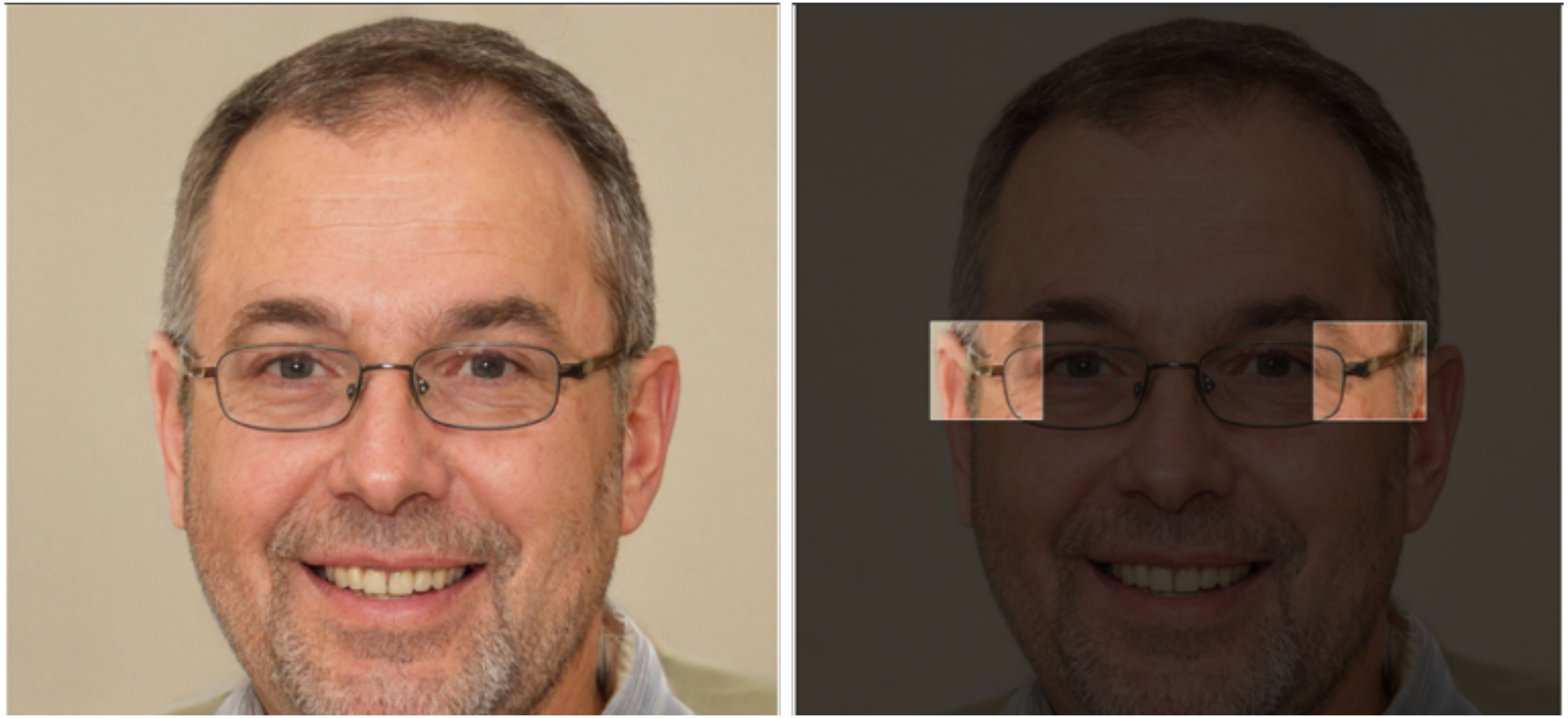
- Completing a prompt from “Harry Potter and the Methods of Rationality”:

> “… 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.



Q1-1:

In class, we’ve seen a video of Obama. Which is true about the video?

A. It’s a video of BBC interview.
B. It’s a private video of Obama leaked by hackers.
C. It’s a fake video.

Q1-2:

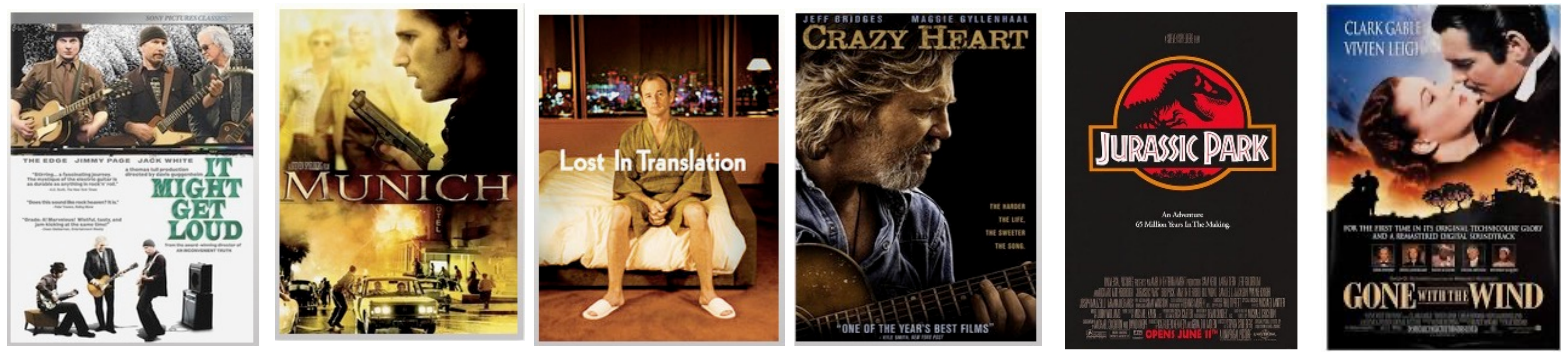
Which of the following is right?

A. Fake images can have drawbacks, so a person can detect a fake image easily.
B. Fake image detection is hard but not impossible.
C. Fake things make life happier so we should generate as many as possible.

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

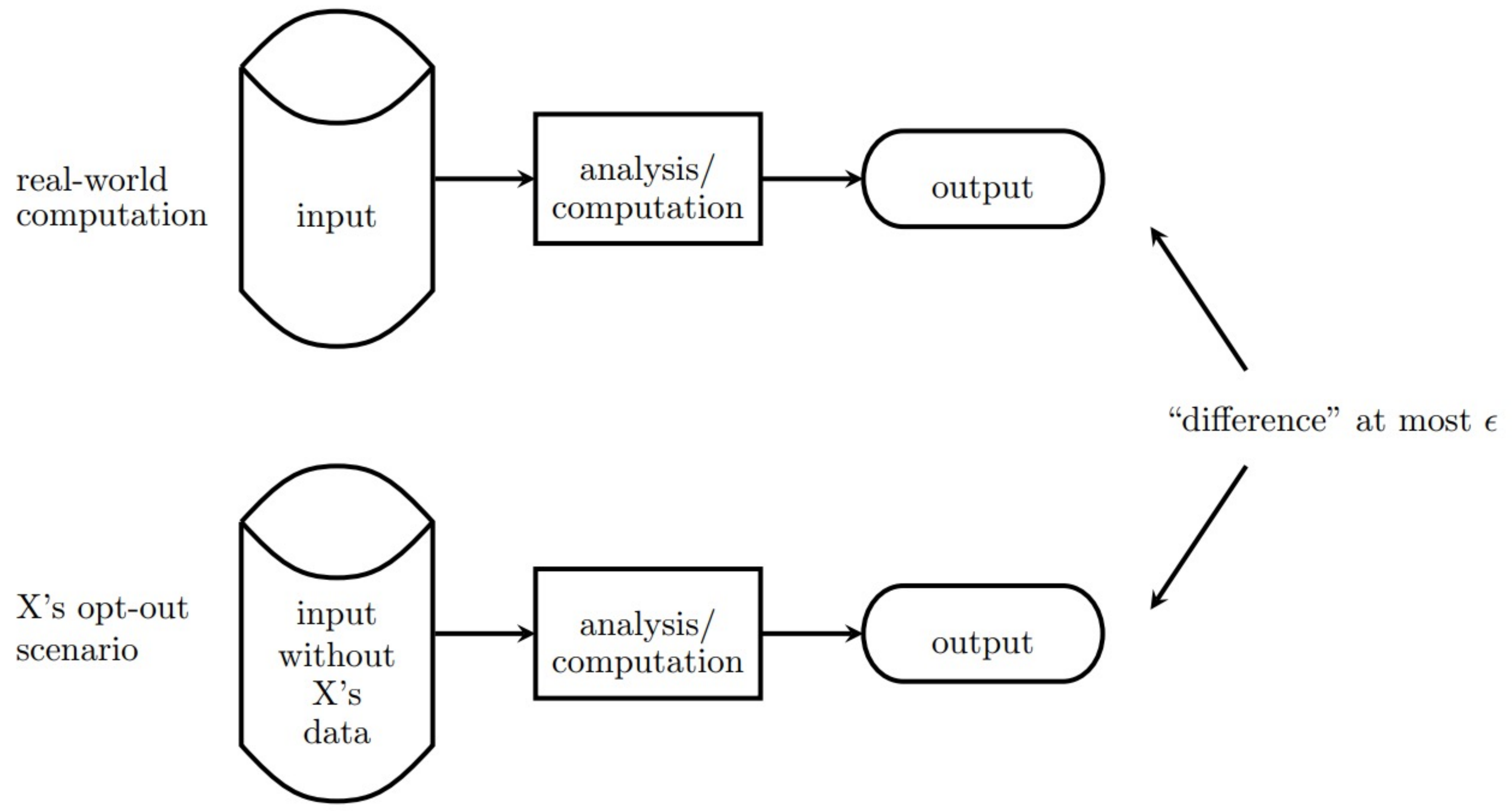
- The data was released by Netflix in 2006
  - replaced individual names with random numbers
  - moved around personal details, etc

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

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

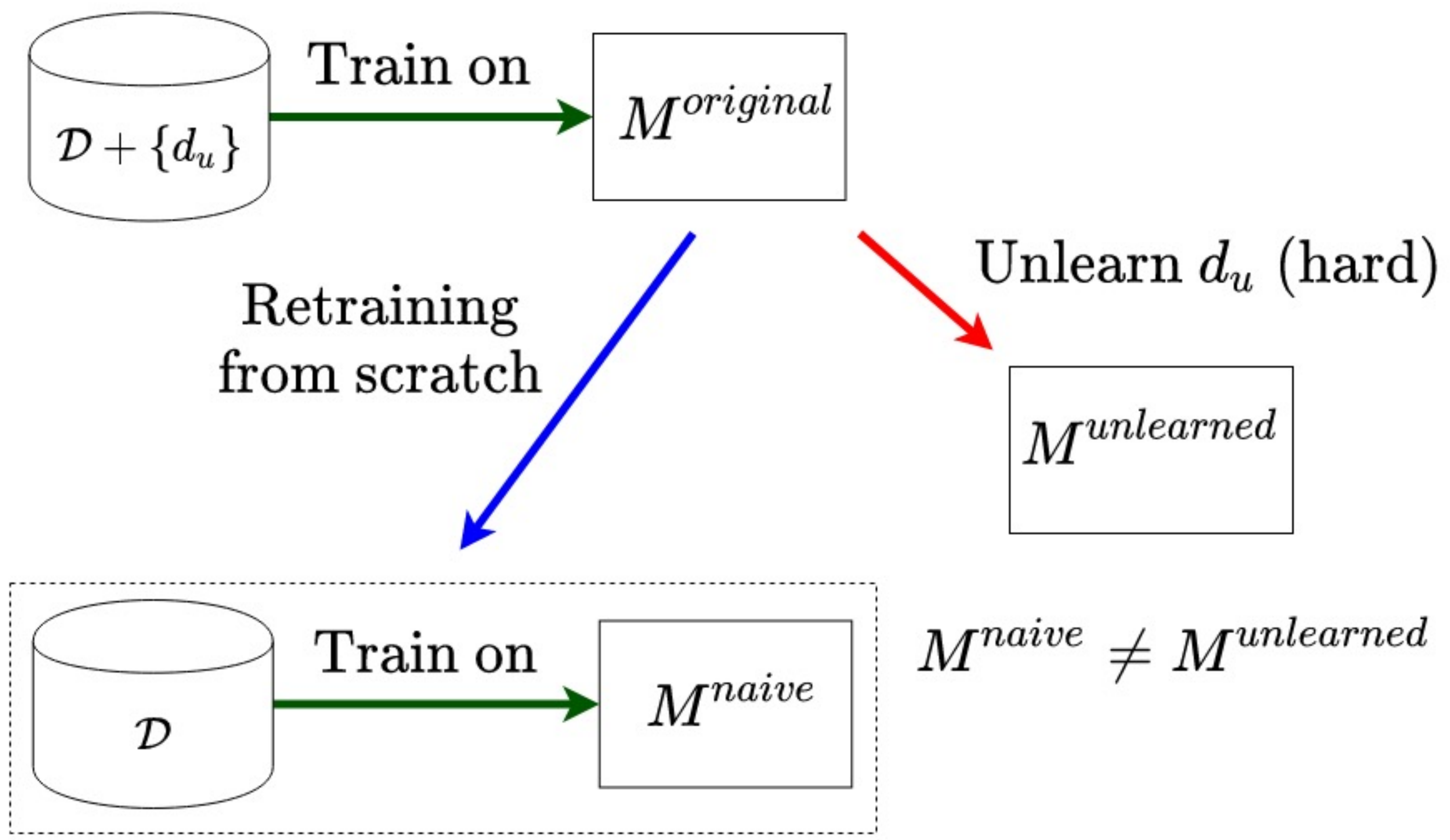


Right to be Forgotten

- The right to request that personally identifiable data be deleted
- E.g., an individual who did something foolish as a teenager doesn’t want it to appear in web searches for the name for the rest of the life

Right to be Forgotten

- What if the data has been used in training a deep network?
  - Need to unlearn



From Link

- Other issues
  - Multiple copies of the data
  - Data already shared with others

Q2-1:

Which of the following is correct about privacy?

A. Privacy is a great concern in current big data era.
B. Big tech companies can always protect individual privacy well enough.
C. Both of above.

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

Where is the bias from?

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

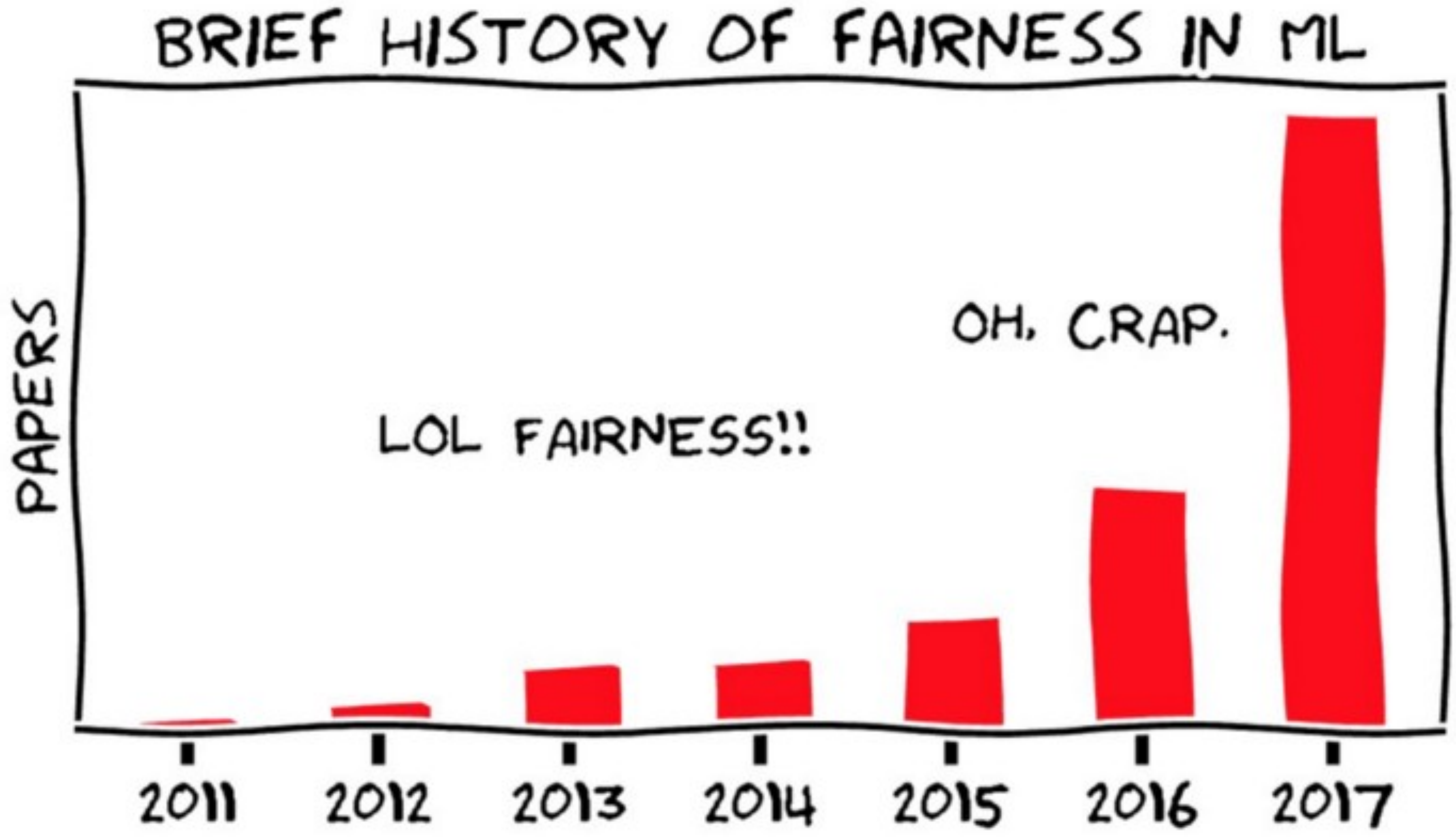
Machine learning systems inherit the bias from the training data.

What causes bias in ML?

- 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 perfectly the minority group.
- 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



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

Fairness through Blindness

Fairness through Blindness

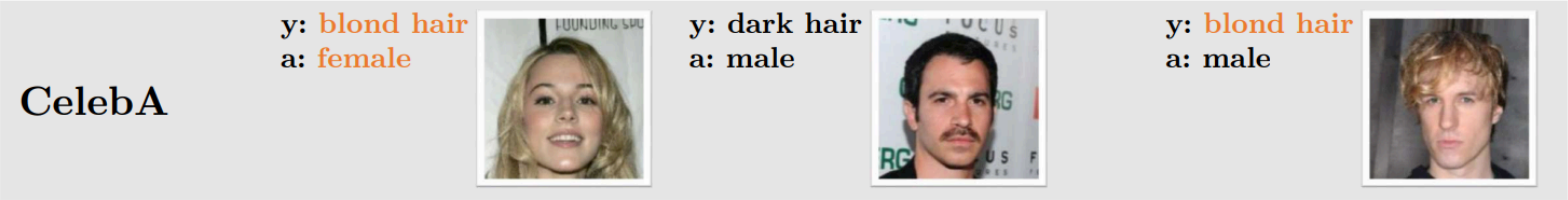
Ignore all irrelevant/protected attributes

Group fairness

Ignore irrelevant/protected attributes ***a*** when trying to predict the label ***y***.

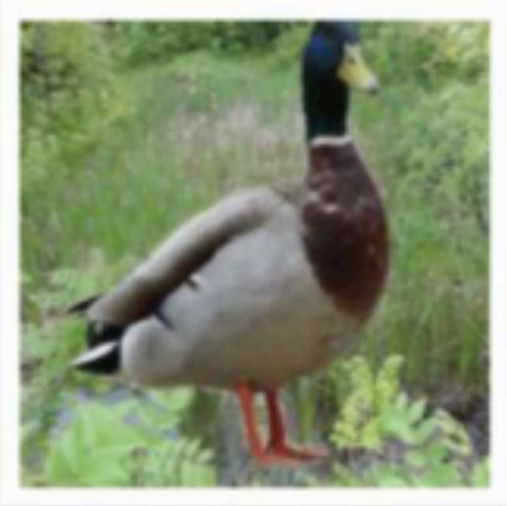
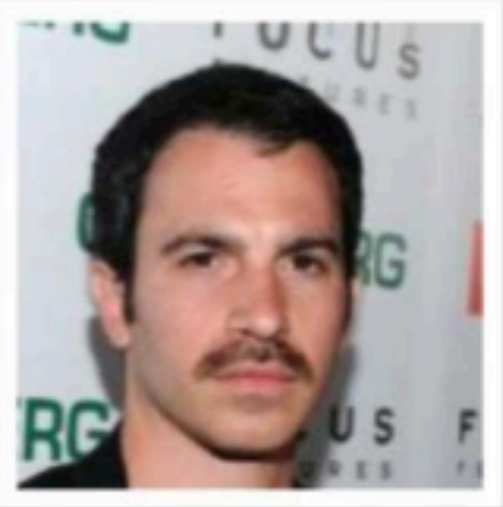
Common Training Examples

Test Example



[Sagawa et al. 2019]

Group fairness

	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.	

Statistical Parity (Group Fairness)

Equalize two groups g_1, g_2 at the level of outcomes

$\Pr[\text{outcome } o \mid g_1] = \Pr[\text{outcome } o \mid g_2]$

"Hair color prediction should not depend on being in g_1 or g_2."

"Fraction of papers in g_1 getting accepted should be the same as in g_2."

GDRO [Sagawa et al. 2019]

Group Distributionally Robust Optimization

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

Minimize the empirical worst-group risk

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.

Individual Fairness

Treat *Similar* Individuals *Similarly*

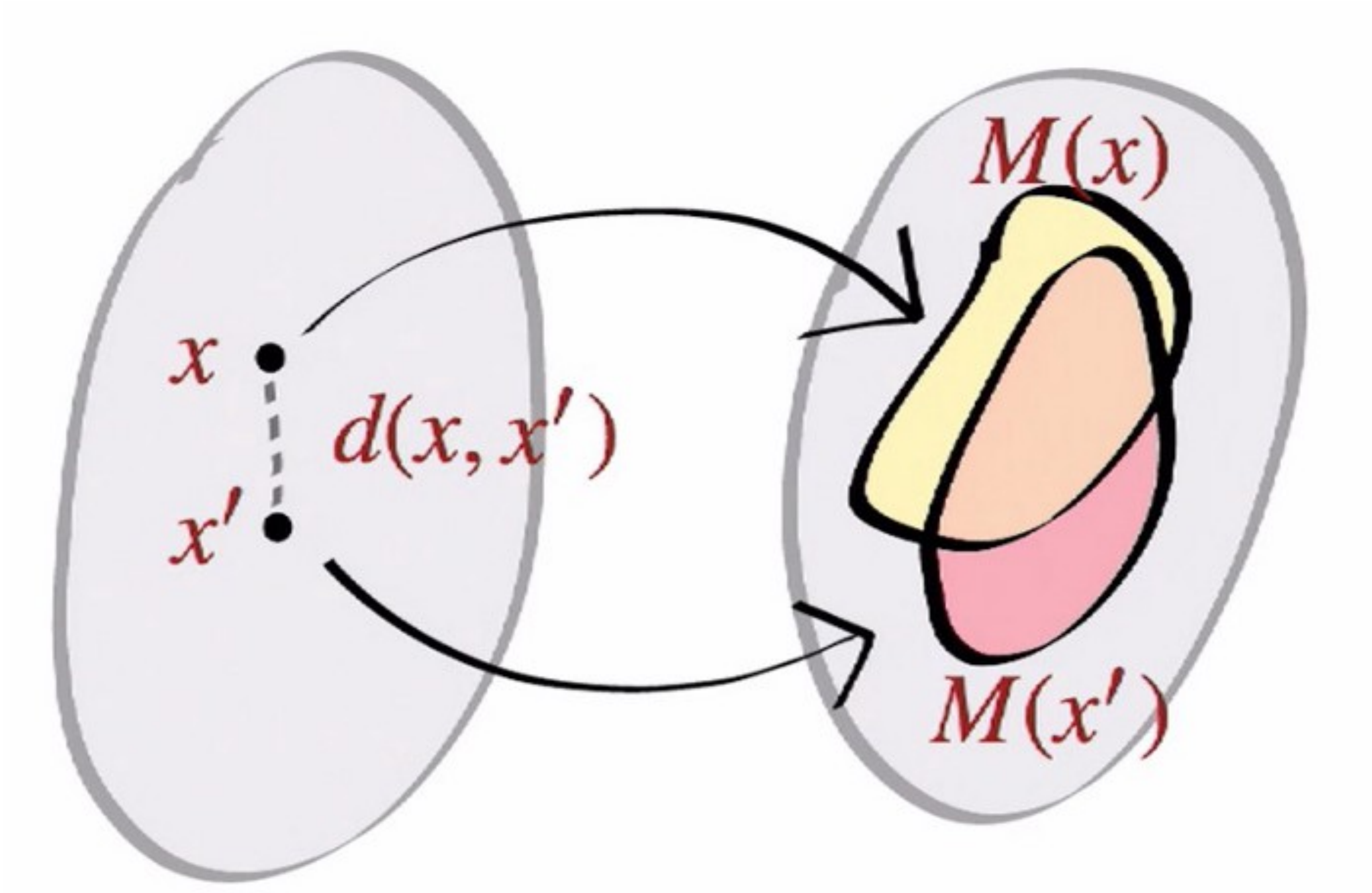
Similar for the purpose of the classification task

Similar distribution over outcomes

Formalize Individual Fairness

$M : x \rightarrow \Delta(O)$ 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



Q3-1:

What is a key reason to bias in AI:

A. Coincidence, there is no bias
B. Added by human deliberately
C. Training data are biased

Q3-2:

How can we solve the fairness problem?

A. Remove bias from training data
B. Design fair learning methods
C. Both of the above

Summary of Topics in Ethics and Trust in AI

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

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