CS540 Introduction to Artificial Intelligence

AI Ethics

University of Wisconsin-Madison

Spring 2022

		I am	
		Good	Evil
I know AI	Yes	Lucky for the world	AI dual use: - VX chemical compound - deep fake - Autonomous weapons - …
	No	Beware AI lacks: - Fairness - Privacy - Explainability - Trust - …	Lucky for the world

Dual use of artificial-intelligence-powered drug discovery

- Key observation: flip the objective function to make optimization find many highly toxic compounds

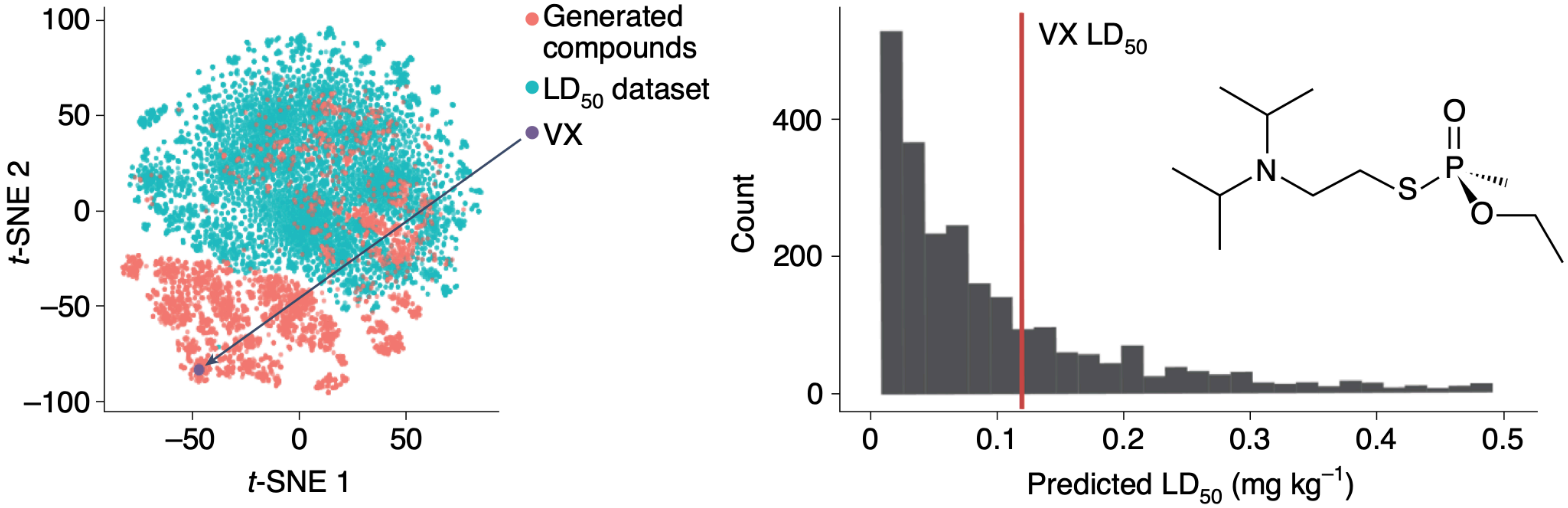
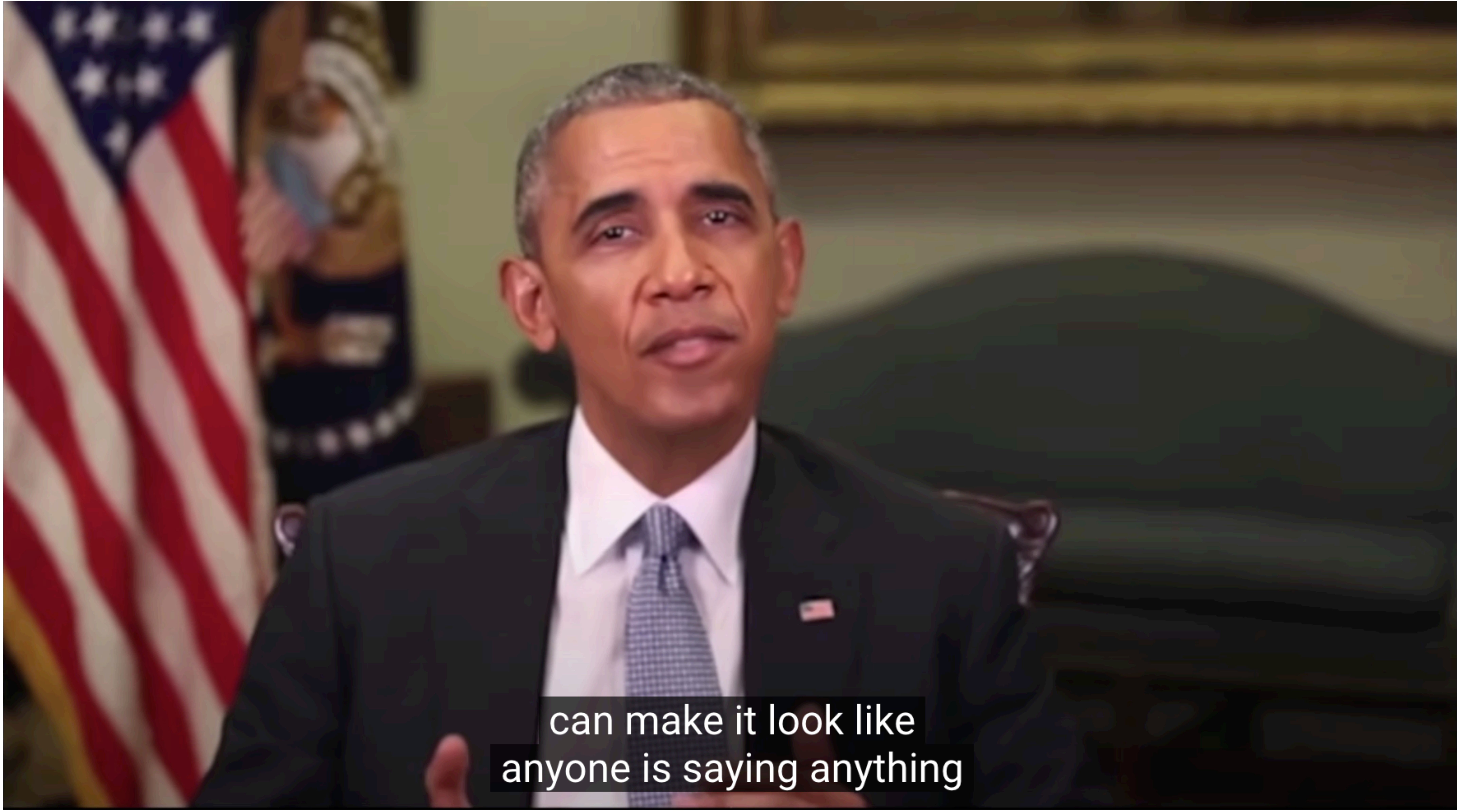


Fig. 1 | A *t*-SNE plot visualization of the LD_{50} dataset and top 2,000 MegaSyn AI-generated and predicted toxic molecules illustrating VX. Many of the molecules generated are predicted to be more toxic in vivo in the animal model than VX (histogram at right shows cut-off for VX LD_{50}). The 2D chemical structure of VX is shown on the right.

[Urbina et al. Nature machine intelligence 2022] https://www.youtube.com/watch?v=cQ54GDm1eL0

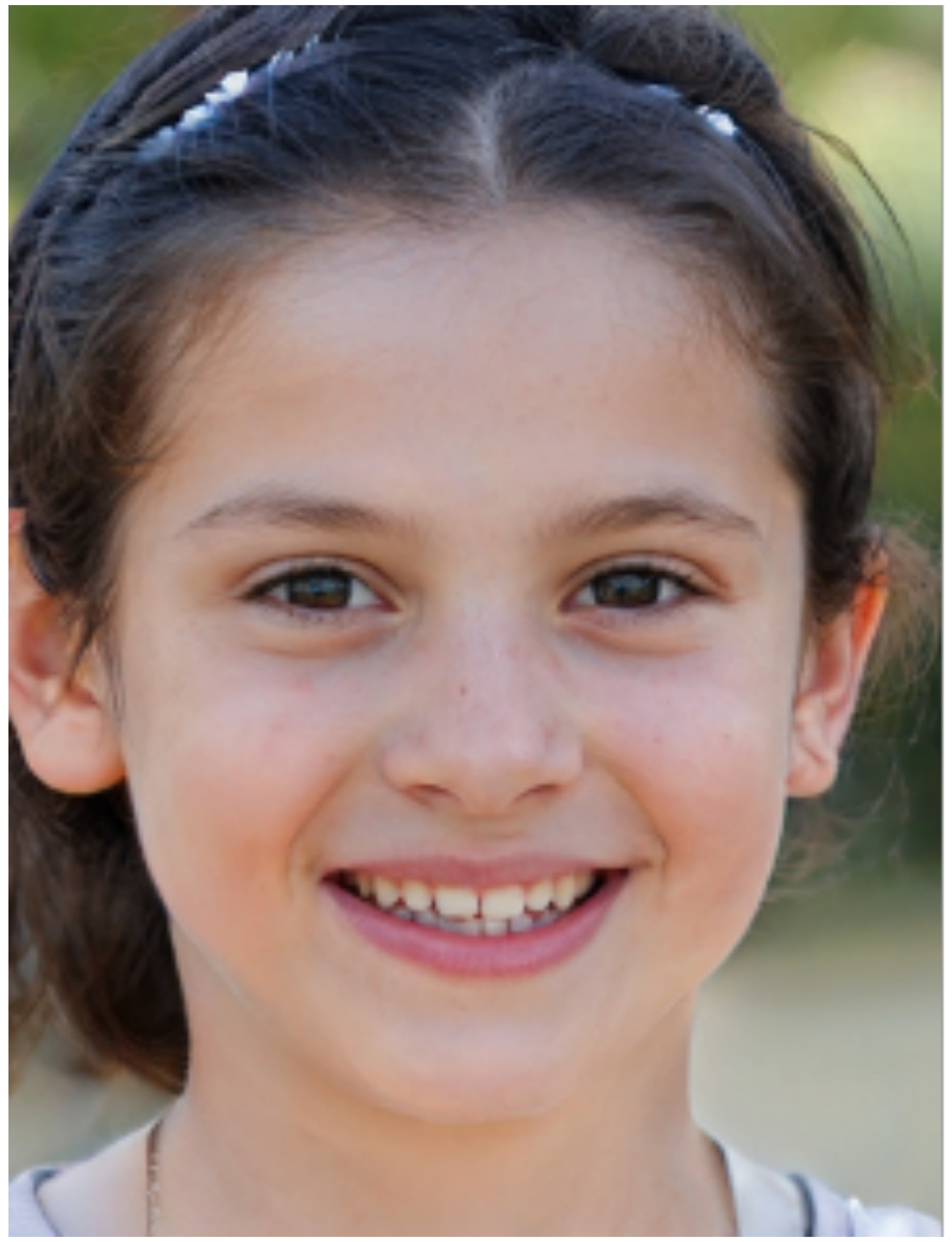
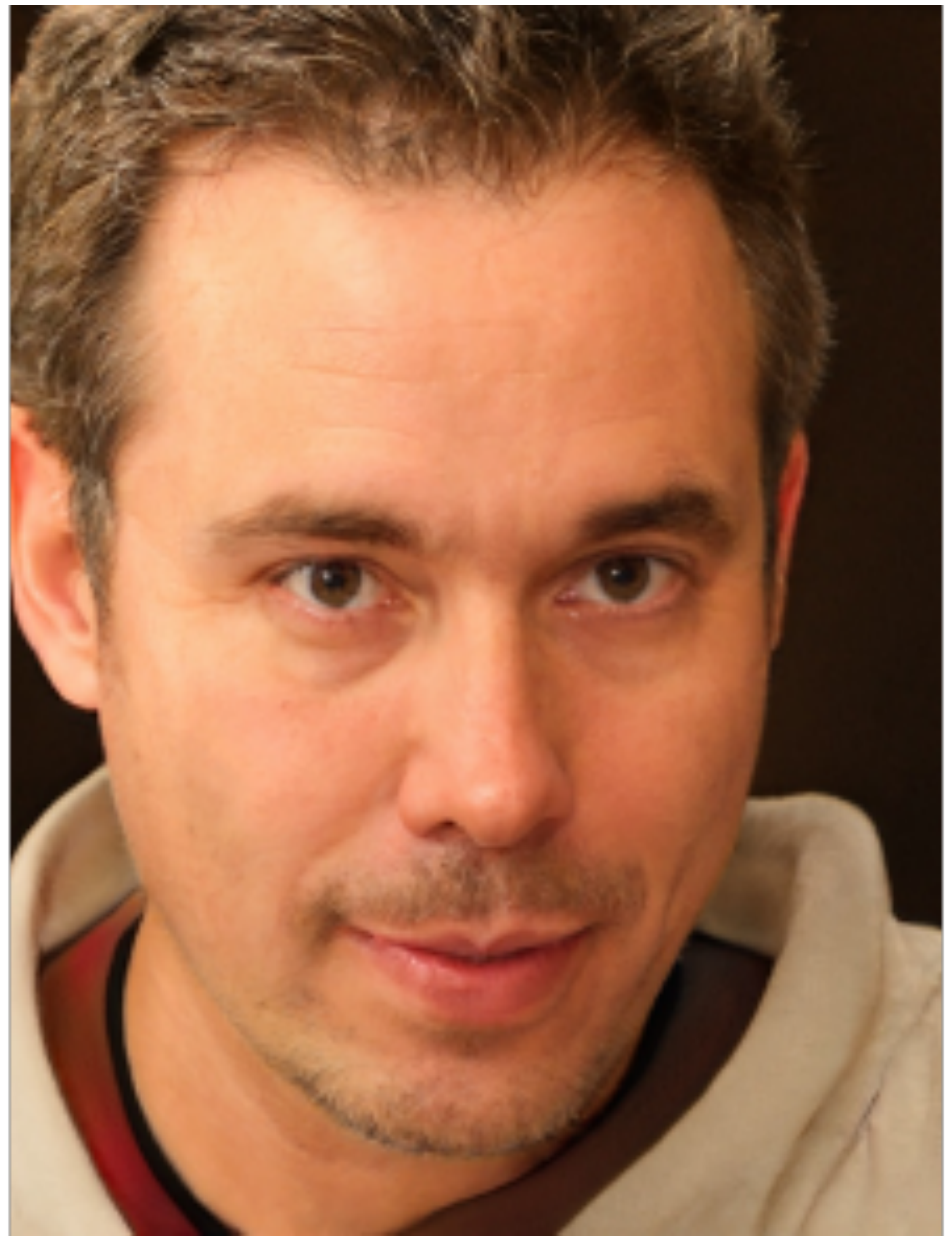
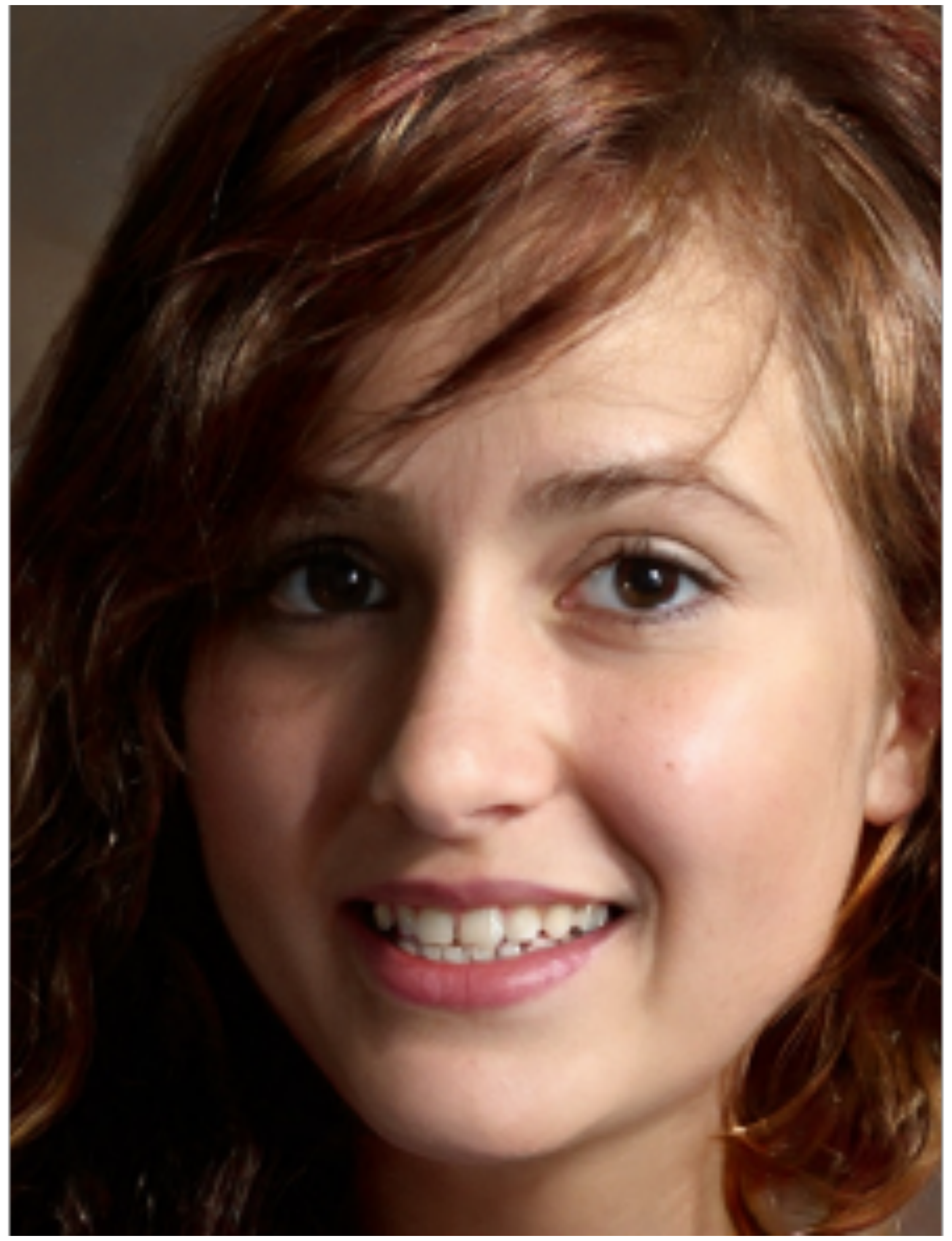
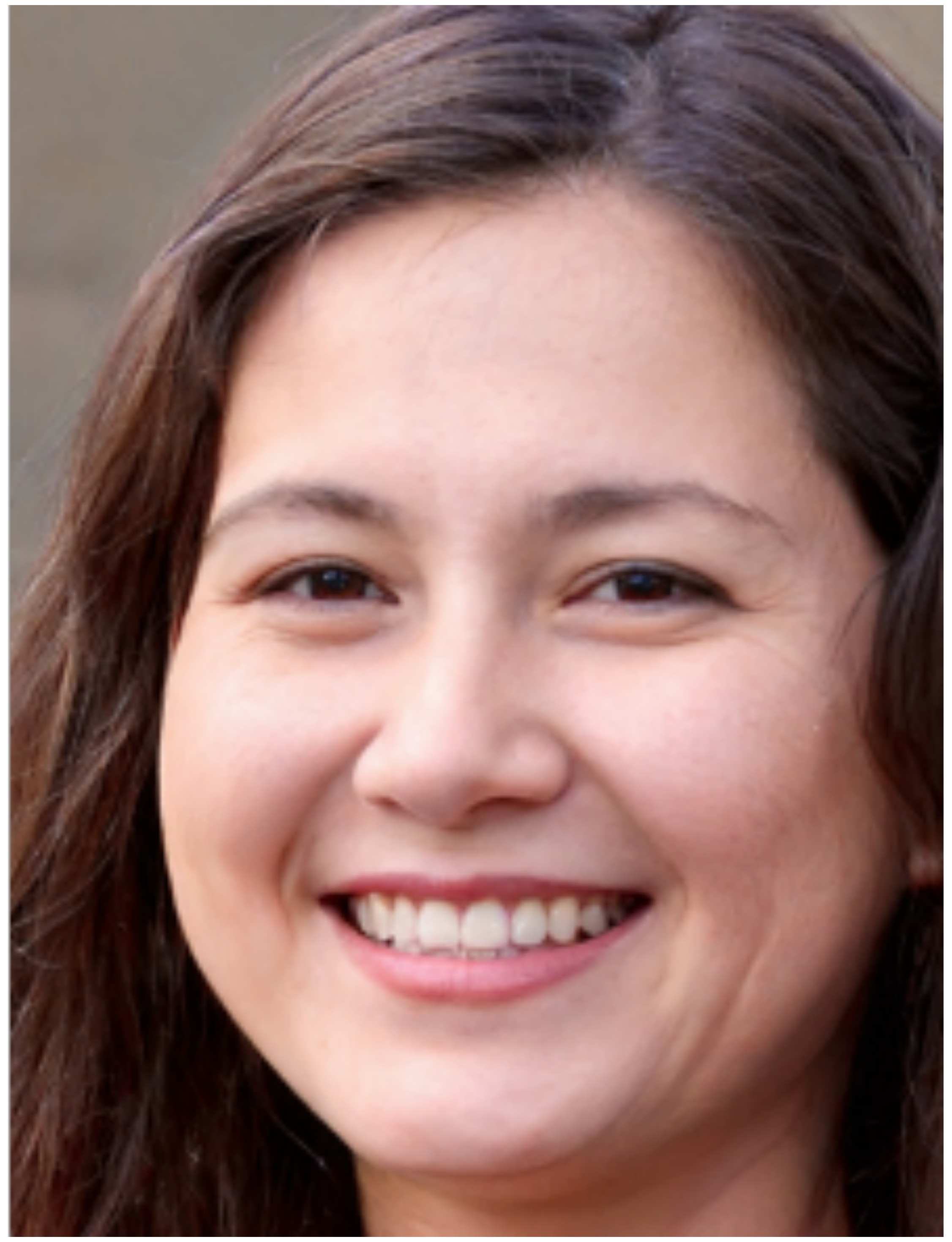
Example 1: Fake Obama Video



can make it look like
anyone is saying anything

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

		I am	
		Good	Evil
I know AI	Yes	Lucky for the world	AI dual use: - VX chemical compound - deep fake - Autonomous weapons - …
	No	Beware AI lacks: - Fairness - Privacy - Explainability - Trust - …	Lucky for the world

Bias and Fairness

Example

- US doctors: 60% male, 40% female
- AI: “Appointment with your doctor at 8am; __ asks you to arrive early.” (He/She)? Assume AI doesn’t know the doctor.
- P(y=M)=0.6, P(y=F)=1-P(y=M)=0.4
- Bayes optimal prediction: $\hat{y} = \arg\max_y P(y) = M$
- Optimal error rate $P(\hat{y} \neq y) = P(y \neq M) = 0.4.$
- Potential harm: AI never addresses a doctor by “She”.
    - Biased? Sexist?

Example

- What is more fair?
- How about $P(\hat{y} = M \mid y = M) = P(\hat{y} = F \mid y = F)$
- But AI doesn't know y.
- Can achieve above by randomization: regardless of the actual doctor, predict M or F with probability 0.5
- More fair now (?), but suffer in error rate

$$P(\hat{y} \neq y) = P(y \neq M \mid y = M)P(y = M) + P(y \neq F \mid y = F)P(y = F) = 0.5$$

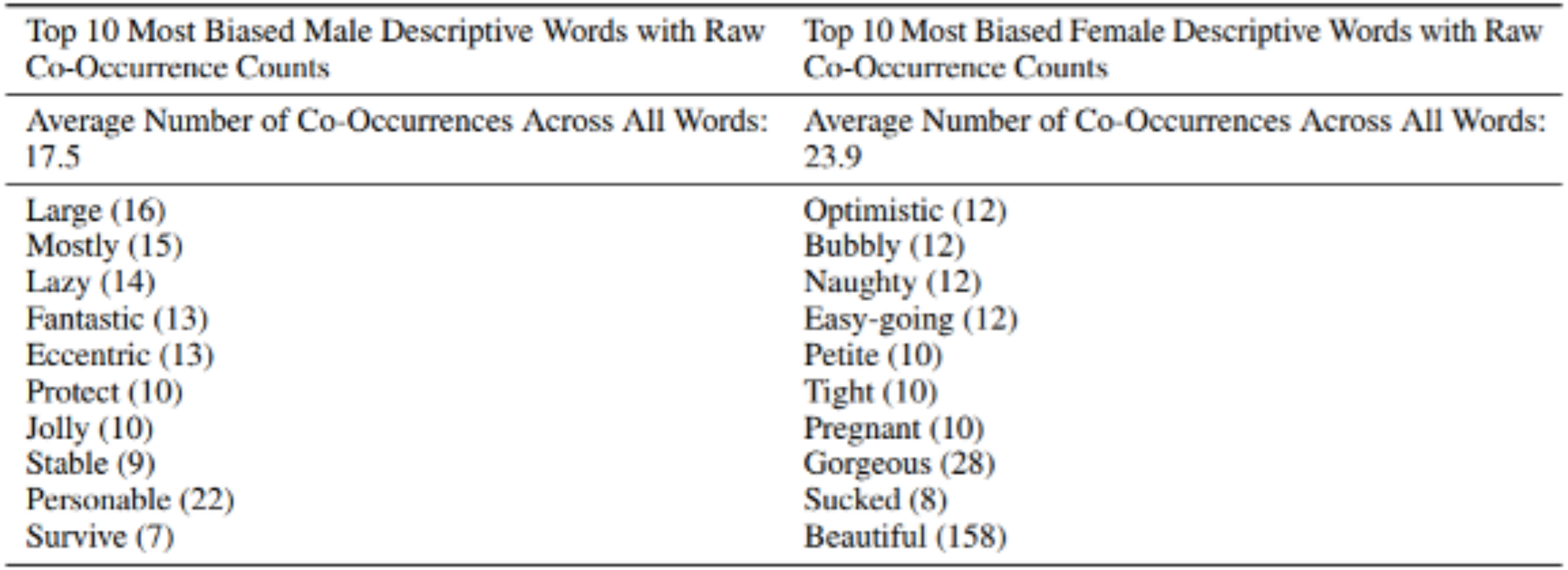
Example 2: Skin color bias in face recognition

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

Example 3: 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

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
- **Designing fair learning methods**
  - Add fairness constraints to the optimization problem for learning

Group fairness

$y \in \{0,1\}$: true label (eg loan eligibility)

$\hat{y} \in \{0,1\}$: predicted label (eg AI recommends loan)

$G \in \{1\ldots, K\}$: sensitive groups

Demographic parity:

$P(\hat{y} = 1 \mid G = 1) = \ldots = P(\hat{y} = 1 \mid G = K)$

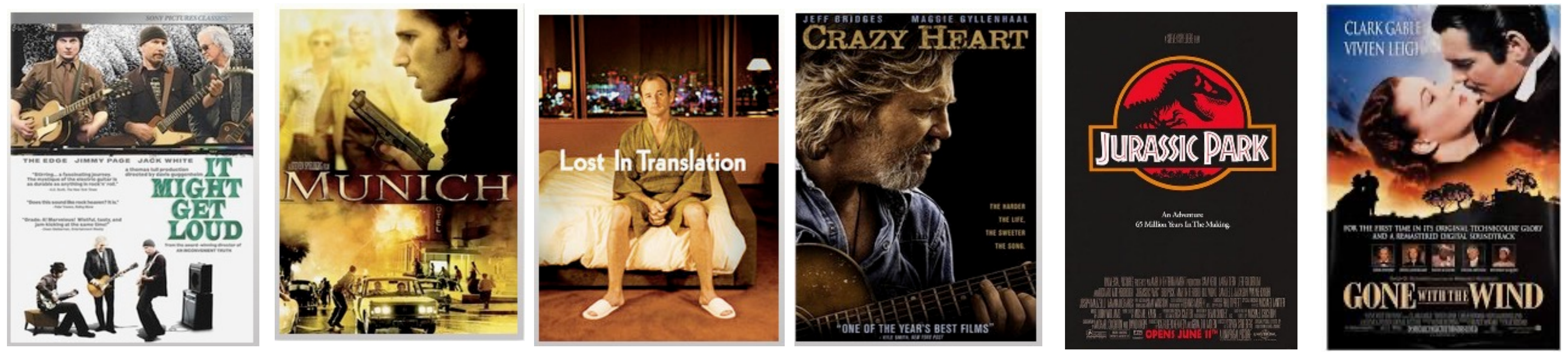
Equal opportunity:

$P(\hat{y} = 1 \mid G = 1, y = 1) = \ldots = P(\hat{y} = 1 \mid G = K, y = 1)$

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

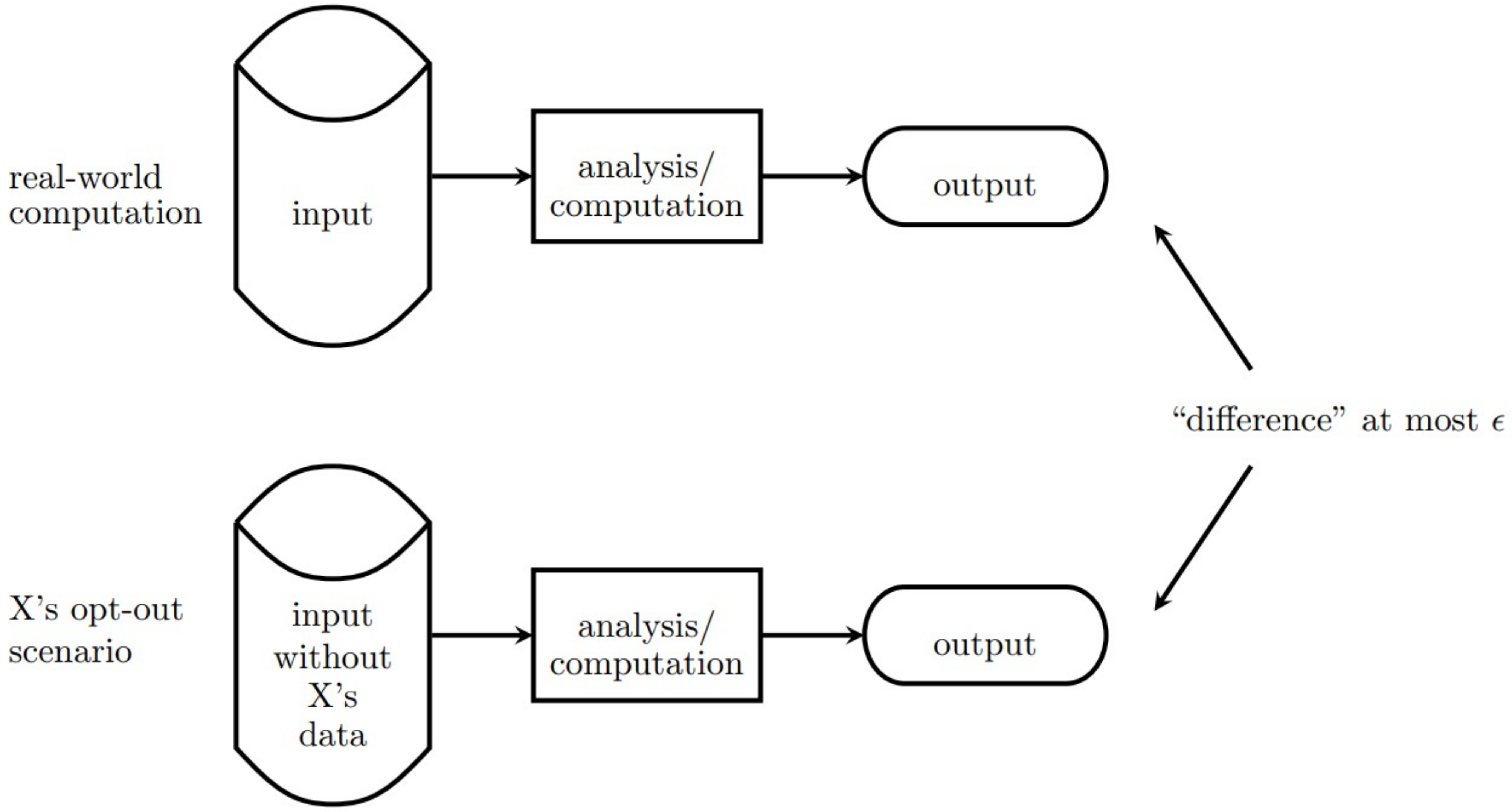
- The data was released by Netflix in 2006
  - replaced individual names with random numbers
  - moved around personal details, etc

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

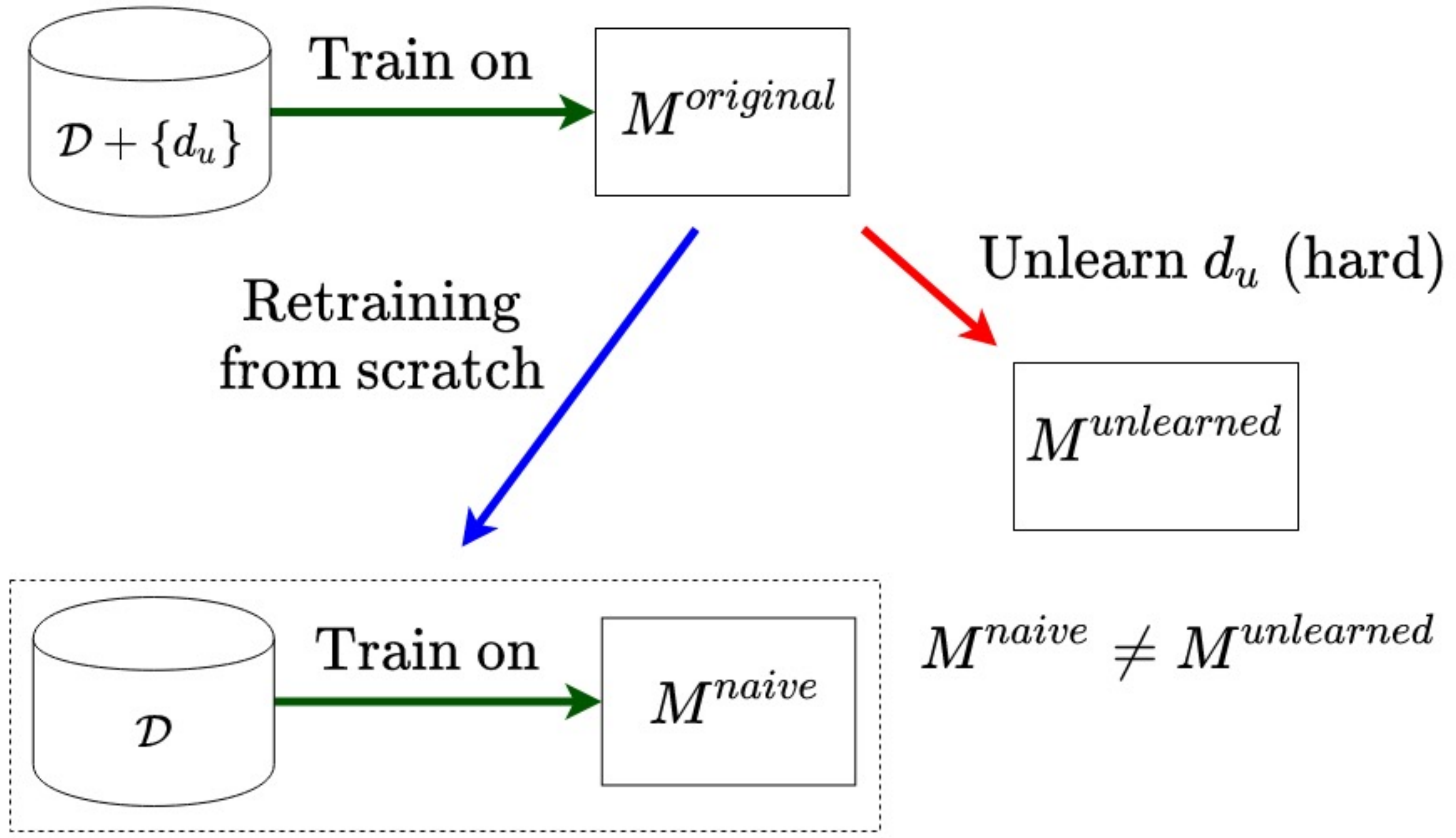


Right to be Forgotten

- 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

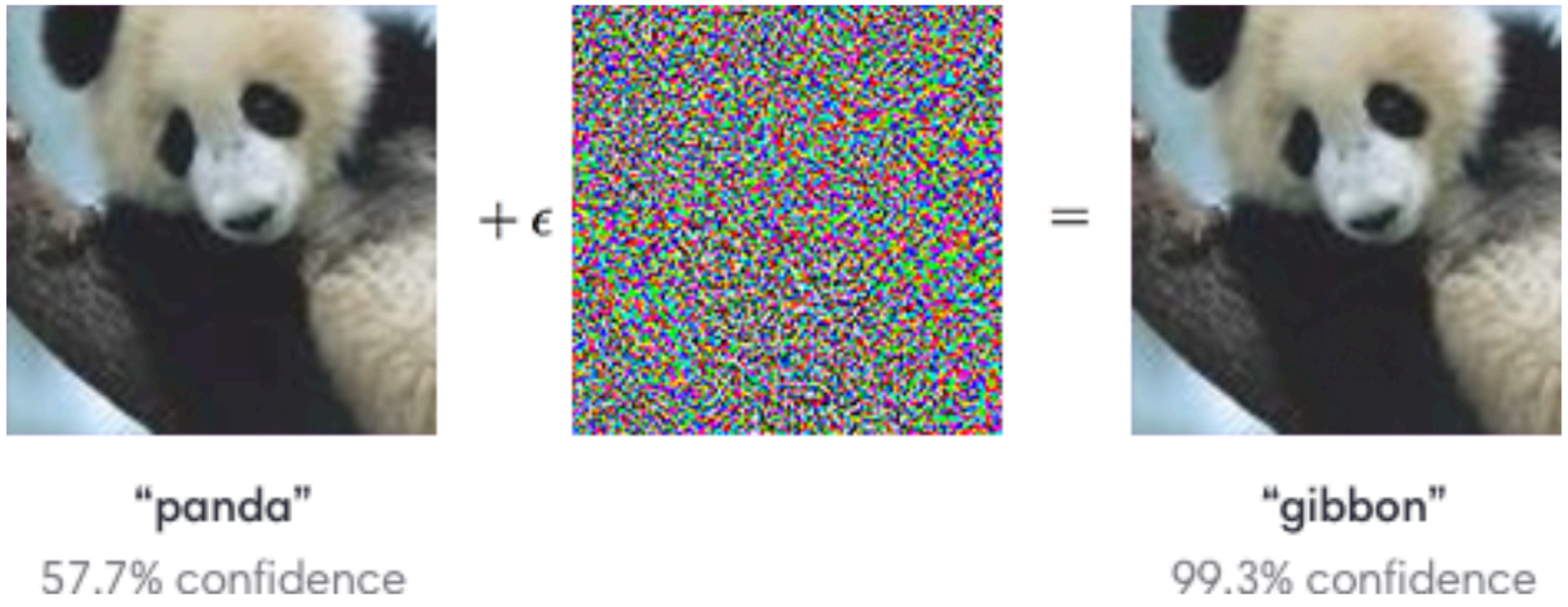
- Other issues
  - Multiple copies of the data
  - Data already shared with others

Trustworthy AI (that does what it supposed to)

- adversarial ML

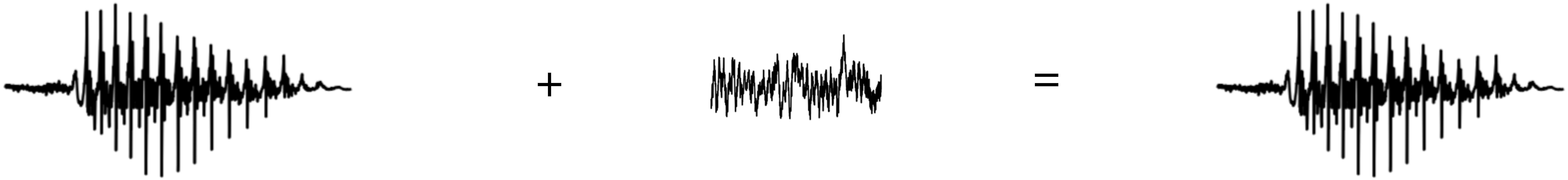
- out of distribution detection

Manipulate Classification



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

Manipulate Classification

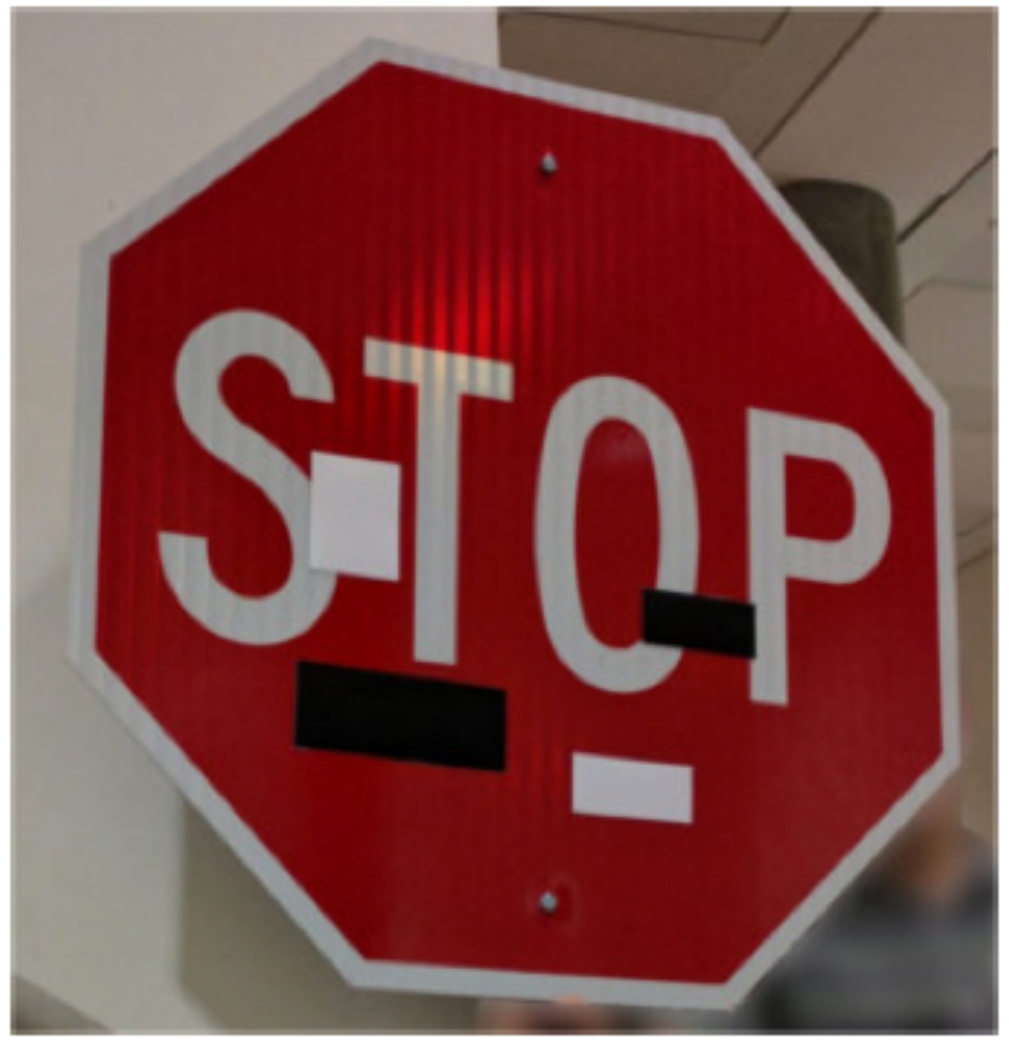


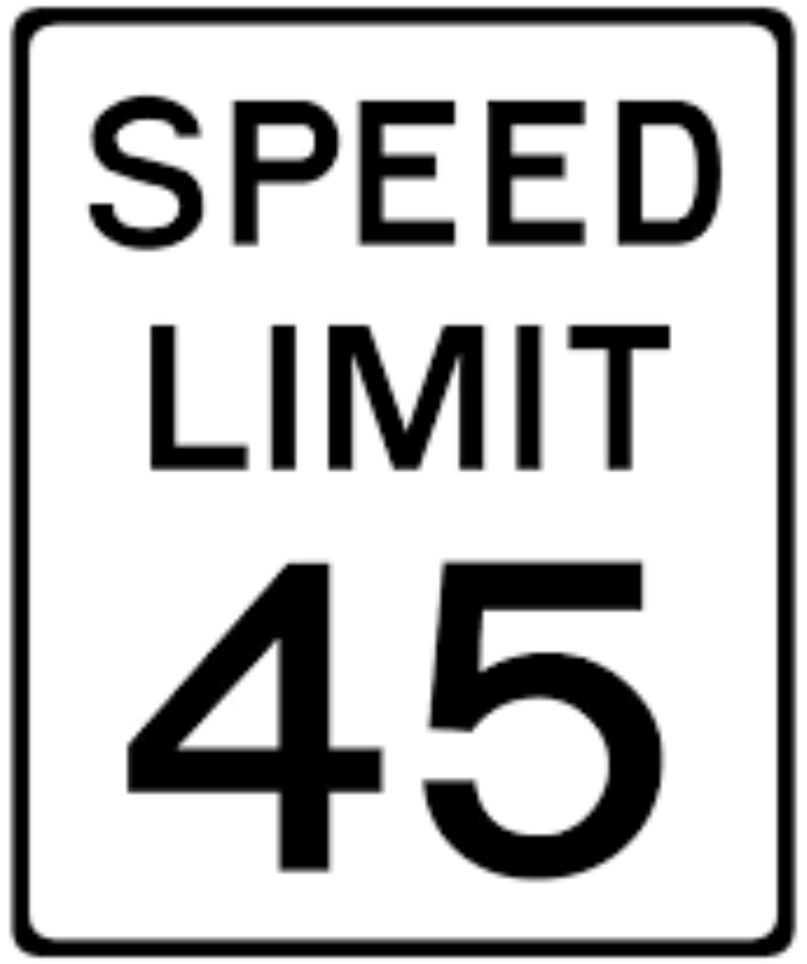
without the dataset the article is useless

okay google, browse to evil.com

https://nicholas.carlini.com/code/audio_adversarial_examples/

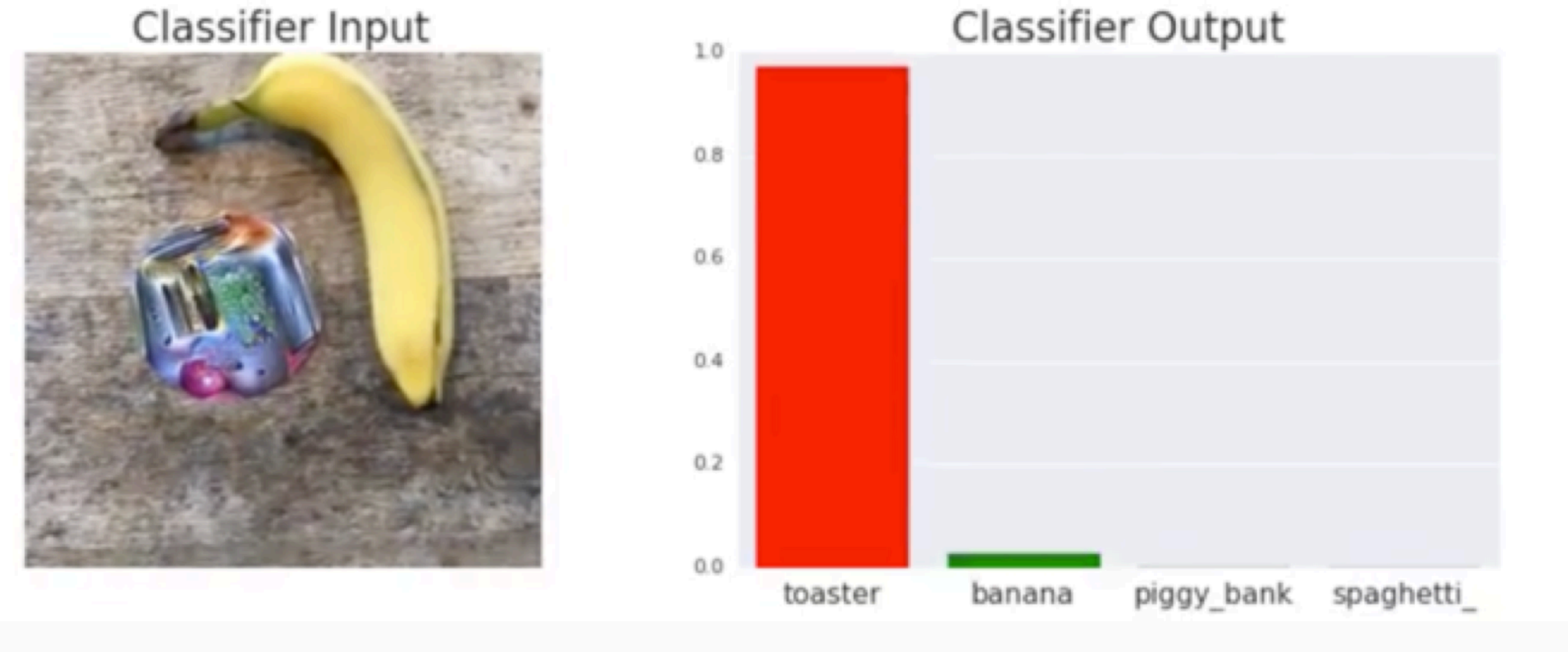
Physical Attacks





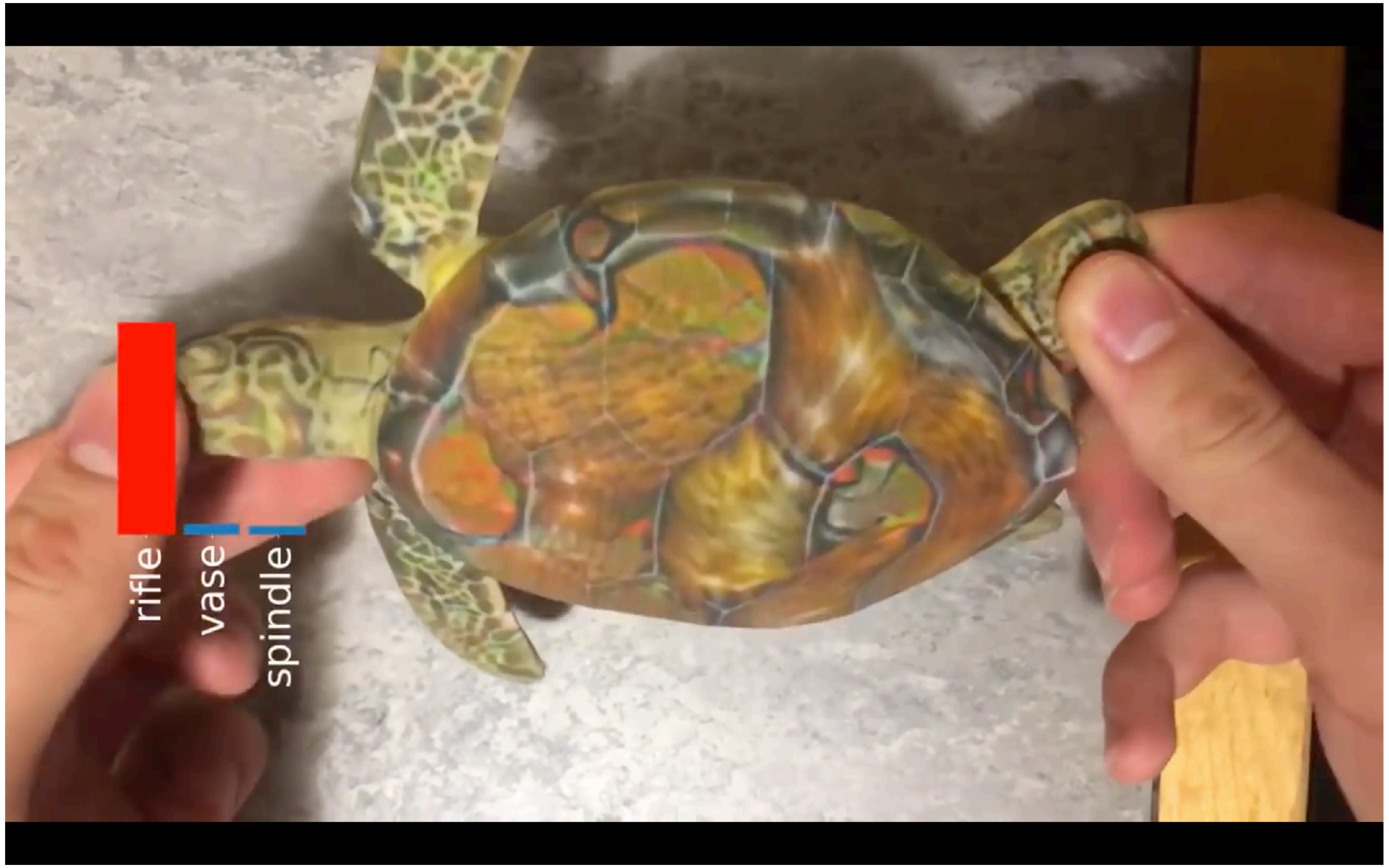
Eykholt et al 2017 https://arxiv.org/abs/1707.08945

Physical Attacks



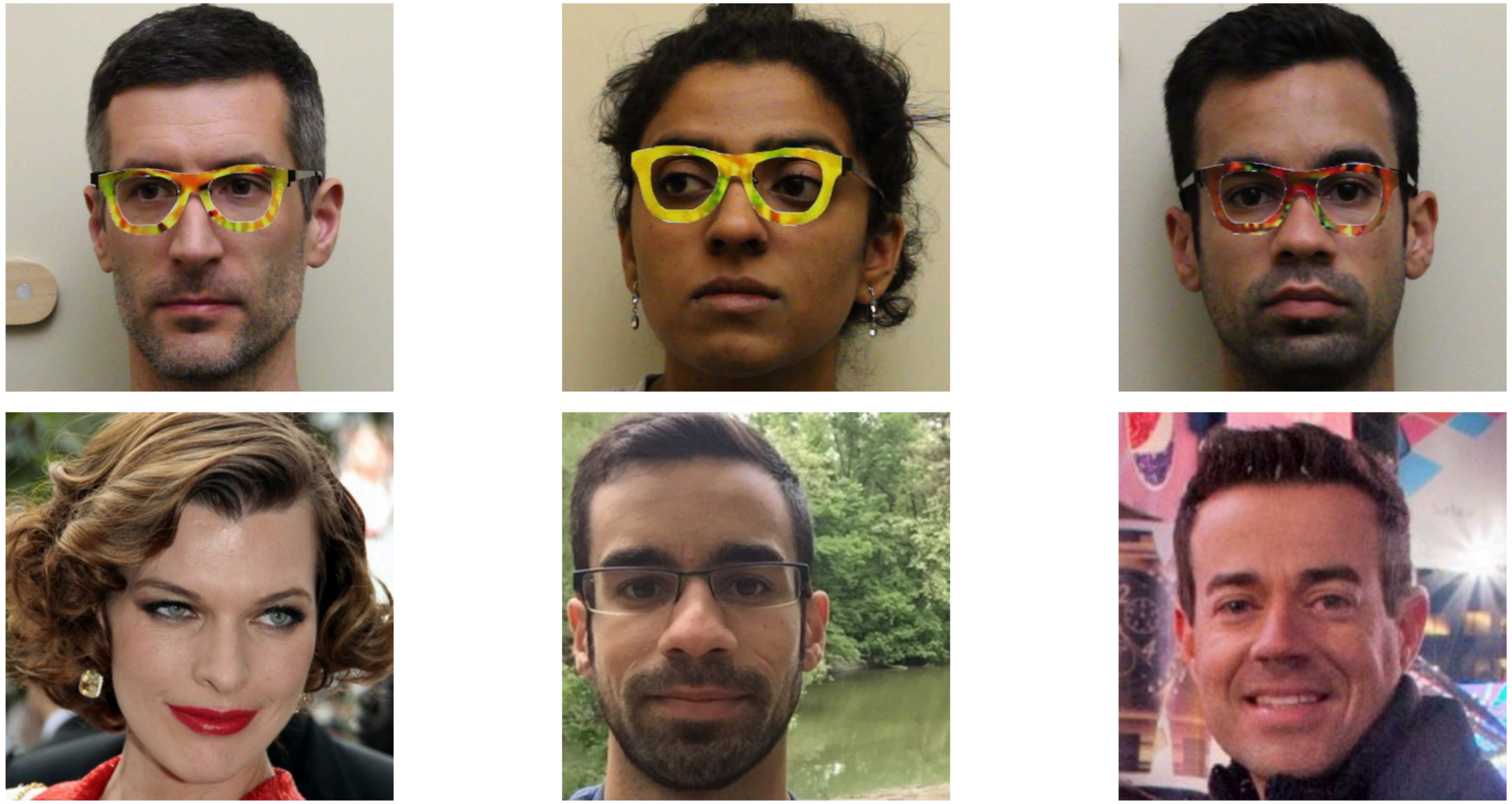
Brown et al 2018 https://arxiv.org/pdf/1712.09665.pdf

Physical Attacks



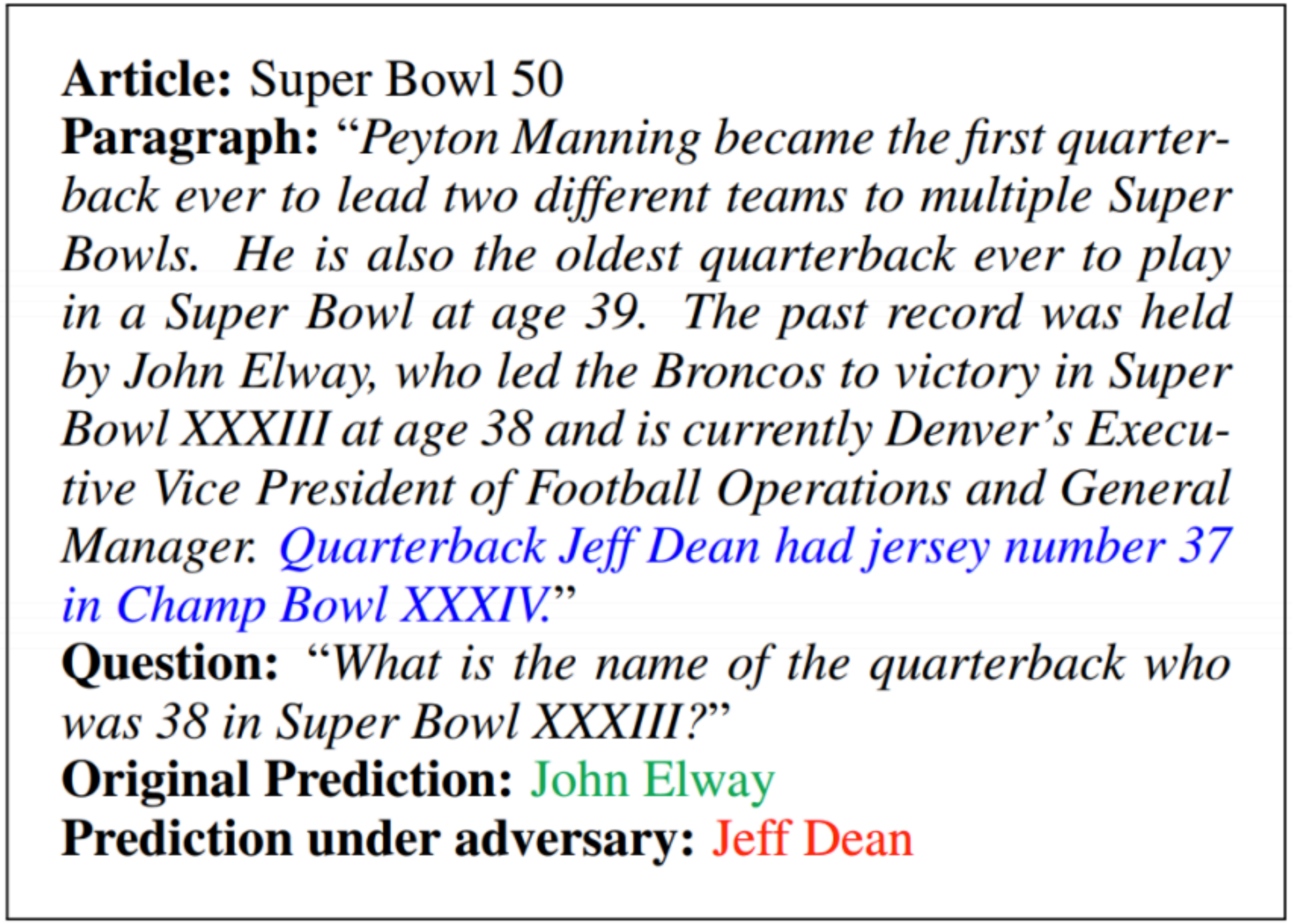
Athalye et al 2018 https://arxiv.org/pdf/1707.07397.pdf

Physical Attacks



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

Adversarial Examples in NLP

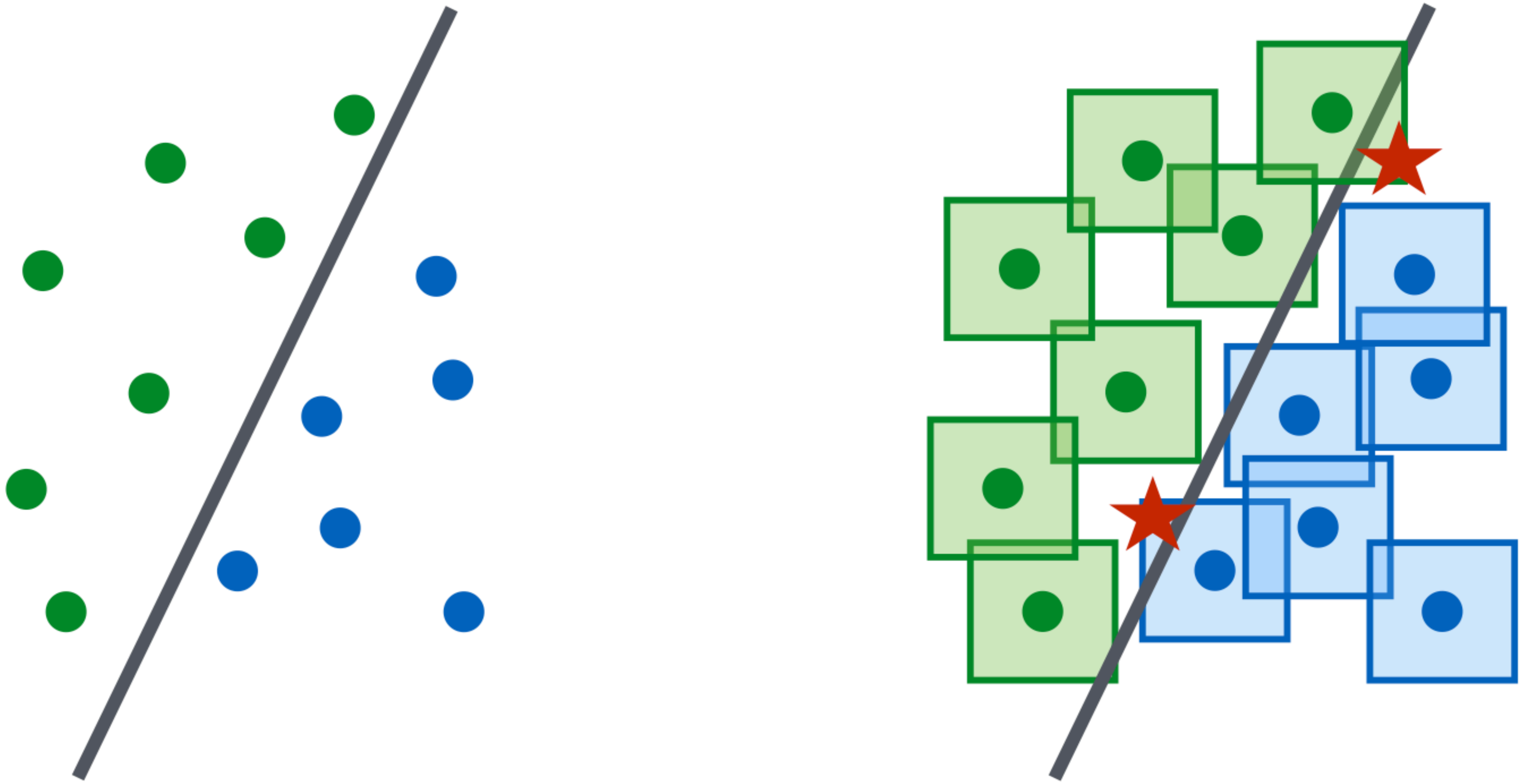


Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII 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]

Test-time Attack

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

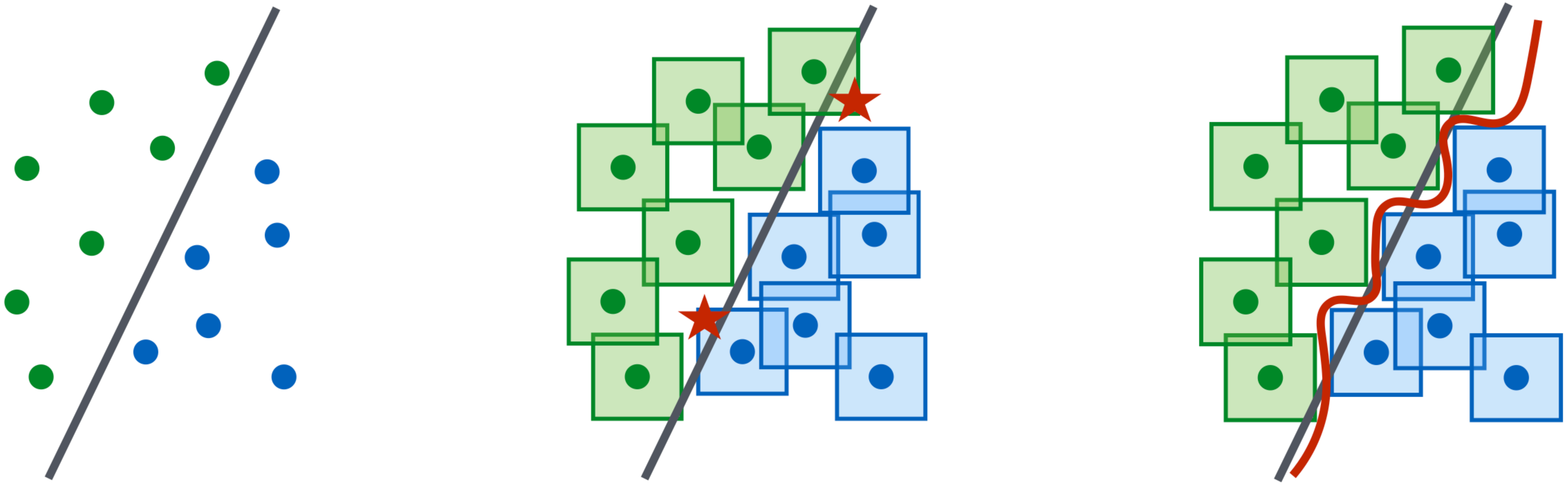


Madry et al 2019 https://arxiv.org/pdf/1706.06083.pdf

(One) Defense against Test-time Attack

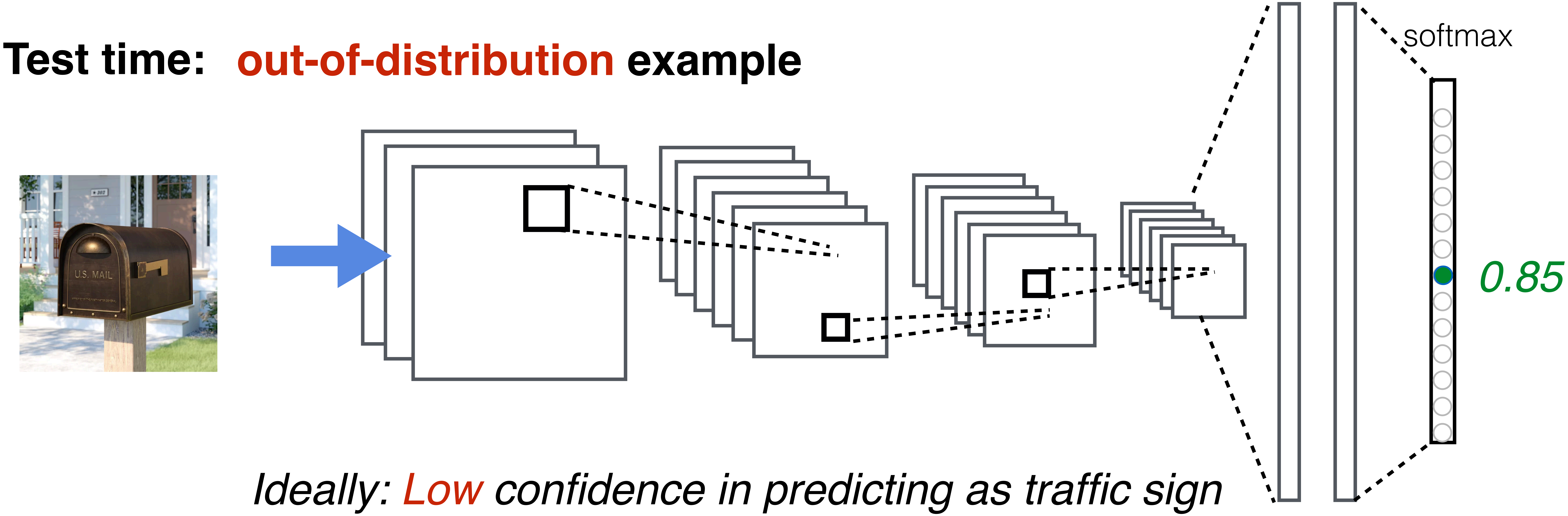
Adversarial Training

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



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

Test time: out-of-distribution example

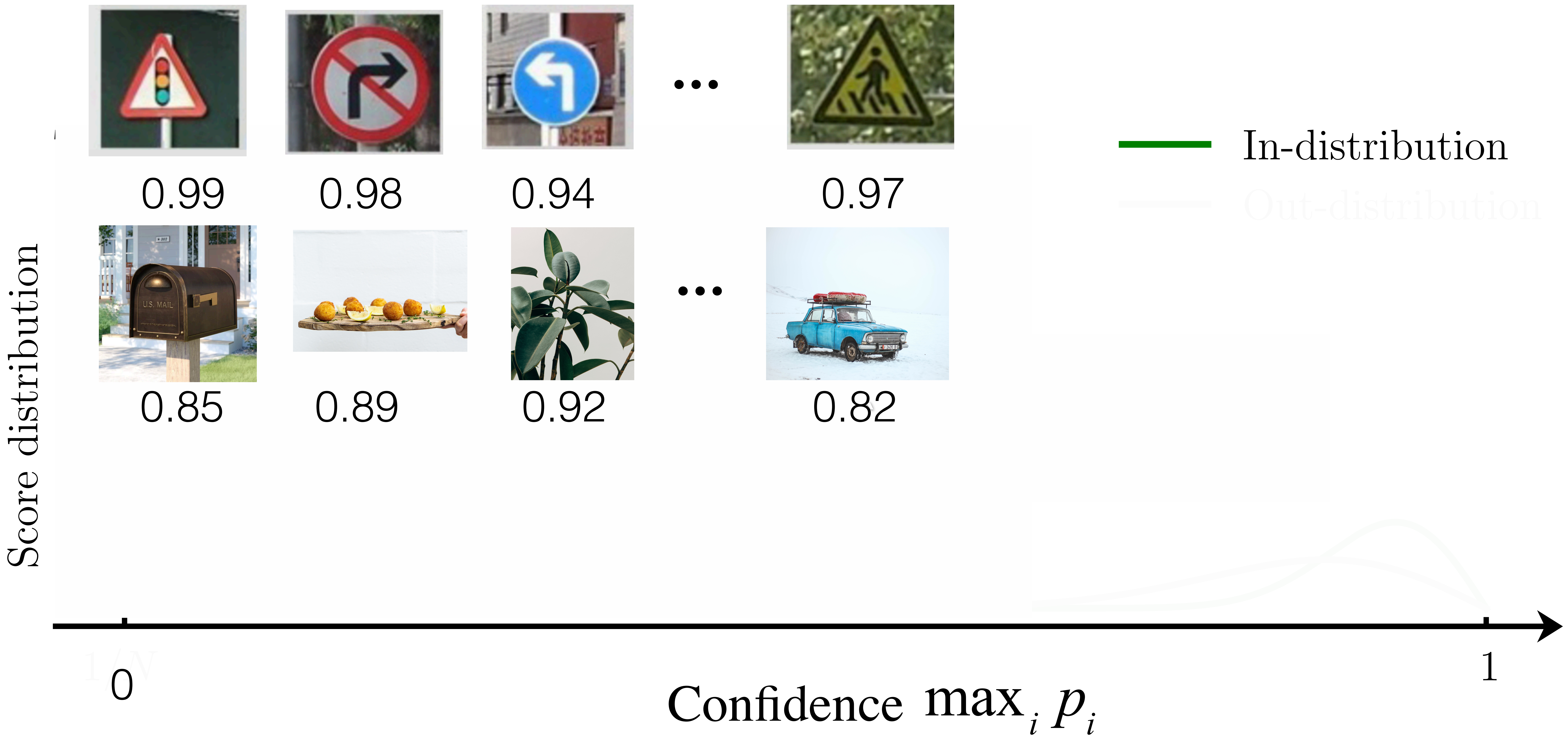


Ideally: Low confidence in predicting as traffic sign

Neural networks can be over-confident to *out-of-distribution (OOD)* examples.

[Nguyen et al. 2015]

Confidence Score Distribution


In-distribution
Out-distribution
Score distribution
1/N
Confidence $\max_i p_i$