



CS540 Introduction to Artificial Intelligence

Ethics and Trust in AI

University of Wisconsin-Madison
Spring 2026 Sections 1 & 2

Announcements: Final Information

- **Time: May 7th, 12:25 - 2:25 PM**
- **Location for sections 001 AND 002 (Instructor Blerina Gkotse)**
based on your last name:
 - Ingraham B10 (Last name starting with A-P inclusive)
 - Noland 132 (Last name starting with Q-Z inclusive)
- Location for **section 003 (Instructor Young Wu): Sterling 1310**
- Students with McBurney accommodations or alternate requests should have received an email with additional information.

Announcements: Final Information

- **Topics:** The exam is **cumulative**
- Exclusion list (Sections 1&2):
 - Logic
 - Tragedy of the Commons (Games I)
 - “From Extensive Form to Normal Form” (Games II)
- **Practice questions:** Canvas -> Files -> Practice Questions
- **Format:** MCQ.
- **Cheat Sheet:** up to 2 handwritten sheets, front and back
- **Calculator:** Calculators are allowed if they don't have an internet connection. A calculator will not be necessary though it may be useful to double check simple arithmetic.
- **Bring:** your WISC ID, pencil (No 2 or softer) and your cheat sheets.

Announcements

- **Homework:**
 - HW10 due on Wednesday April 29 at 11:59PM
- **Course evaluation until May 1st:**
 - If participation reaches 50%, we'll reveal more information for the final.
 - If participation reaches 80%, even more.



- **Class roadmap:**

Ethics and Trust in AI

Review

Outline

- Review Advanced Search/Optimization
- Bias and Fairness
- Fake Content
- Privacy
- Adversarial robustness

Review: Advanced Search

Problem Setting

How is a search problem defined?

How different from other search types?

Hill Climbing

Genetic Algorithms

What is difference between two?

Simulated Annealing

Fitness
Population
Cross-over
Mutation

Neighbors
Local vs. global optima

Review: Hill Climbing Algorithm

Pseudocode:

1. Pick initial state s
2. Pick t in **neighbors**(s) with the best $f(t)$
3. if $f(t)$ is not better than $f(s)$ THEN stop, return s
4. $s \leftarrow t$. goto 2.

What could happen?

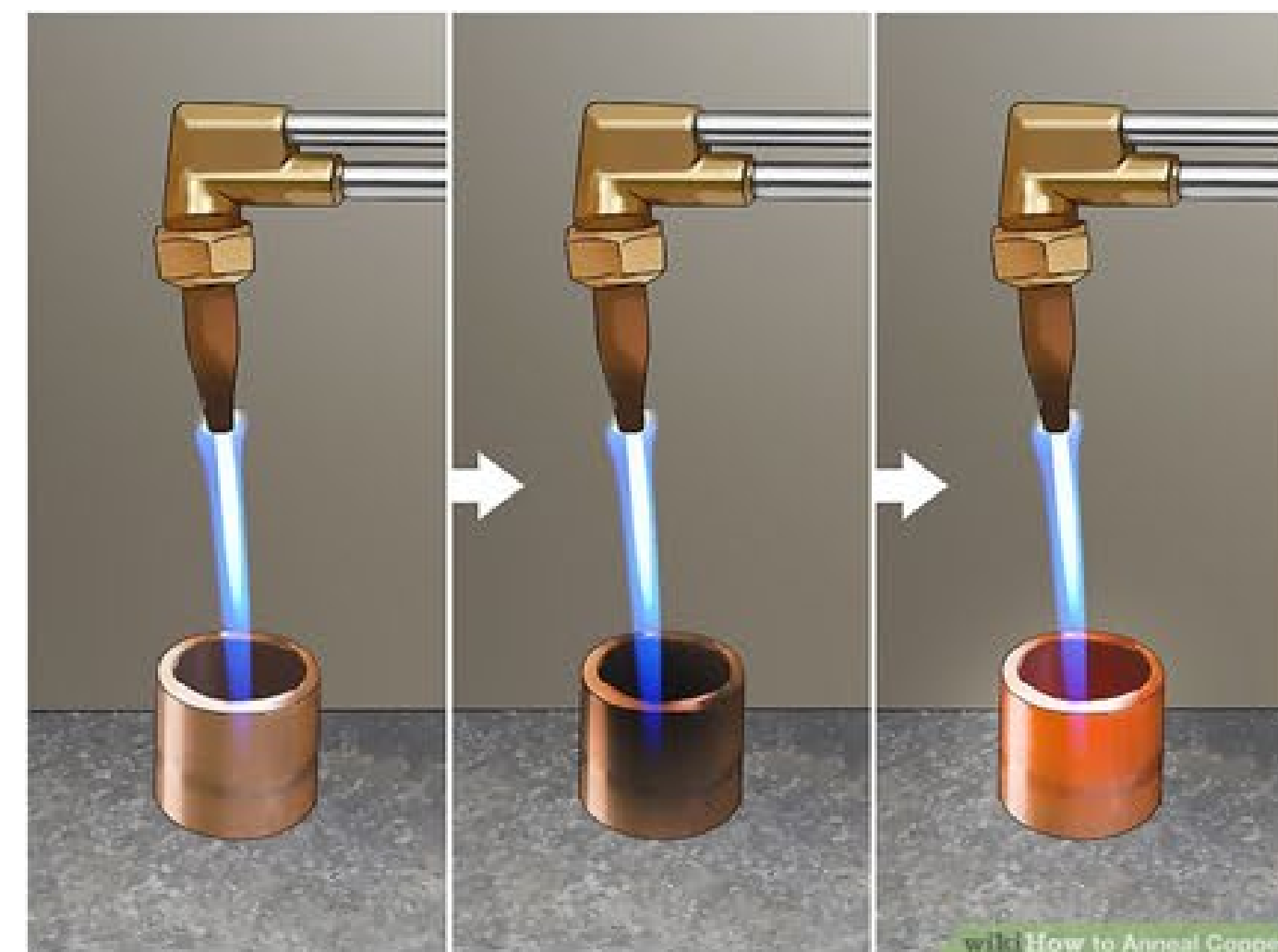


Review: Simulated Annealing

A more sophisticated optimization approach.

- **Idea:** allow some downhill moves at first, then be pickier over time
- **Pseudocode:**
 - Pick initial state s ; $T=1$
 - For $k = 0$ through K :
 - $T \leftarrow T * 0.99$ (cool down)
 - Pick a random neighbour $t \leftarrow \text{neighbor}(s)$
 - If $f(t)$ better than $f(s)$, then $s \leftarrow t$
 - Else with prob. $P(f(s), f(t), T)$ still do $s \leftarrow t$

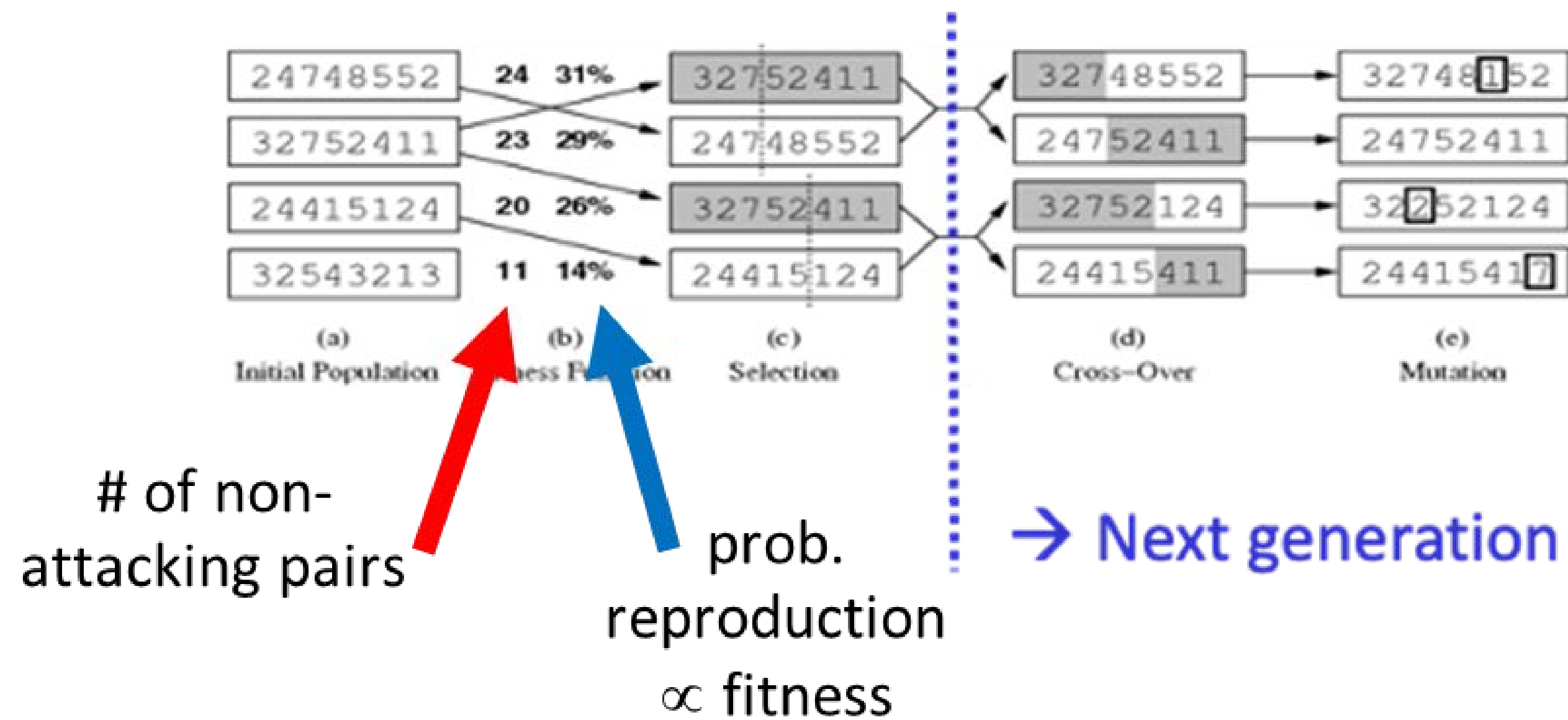
Output: the best state ever seen



Review: Genetic Algorithms

Goal of genetic algorithms: optimize using principles inspired by mechanism for evolution

- Analogous to **natural selection**, **cross-over**, and **mutation**



Variations & Concerns

Many **possibilities**:

- Parents survive to next generation
- Use ranking instead of exact value of $f(s)$ for reproduction probabilities (reduce influence of extreme f values)

Some **challenges**

- Formulating a good state encoding
- Lack of diversity: converge too soon
- Must pick a lot of parameters



Break & Quiz

Q 3.3: In a genetic algorithm for the 8-Queens problem, consider two parent solutions:

- Parent 1: 24748552
- Parent 2: 32752411

If crossover occurs at position 3 (after the third digit), what are the two children?

- A. 48552327 and 52411247
- B. 24748411 and 32752552
- C. 24524752 and 32458411
- D. 24752411 and 32748552

Break & Quiz

Q 3.3: In a genetic algorithm for the 8-Queens problem, consider two parent solutions:

- Parent 1: 24748552
- Parent 2: 32752411

If crossover occurs at position 3 (after the third digit), what are the two children?

- A. 48552327 and 52411247
- B. 24748411 and 32752552
- C. 24524752 and 32458411
- D. **24752411 and 32748552**

In crossover at position 3, we swap the portions after the crossover point:

Parent 1: 247|48552

Parent 2: 327|52411

After crossover:

- Child 1 takes first 3 digits from Parent 1 and remaining from Parent 2:
247|52411 = 24752411

- Child 2 takes first 3 digits from Parent 2 and remaining from Parent 1:
327|48552 = 32748552

This operation mimics genetic recombination in biological evolution.

Summary

- Challenging optimization problems
 - First, try hill climbing. Simplest solution
- Simulated annealing
 - More sophisticated approach; helps with local optima
- Genetic algorithms
 - Biology-inspired optimization routine

Suggested Readings

Textbook: Artificial Intelligence: A Modern Approach (4th edition).
Stuart Russell and Peter Norvig. Pearson, 2020.

- Chapter 4



Ethics and Trust in AI

I am

Good

Evil

Yes

Lucky for the world

- AI dual use:
- VX chemical compound
 - deep fake
 - Autonomous weapons
 - ...

I know AI

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

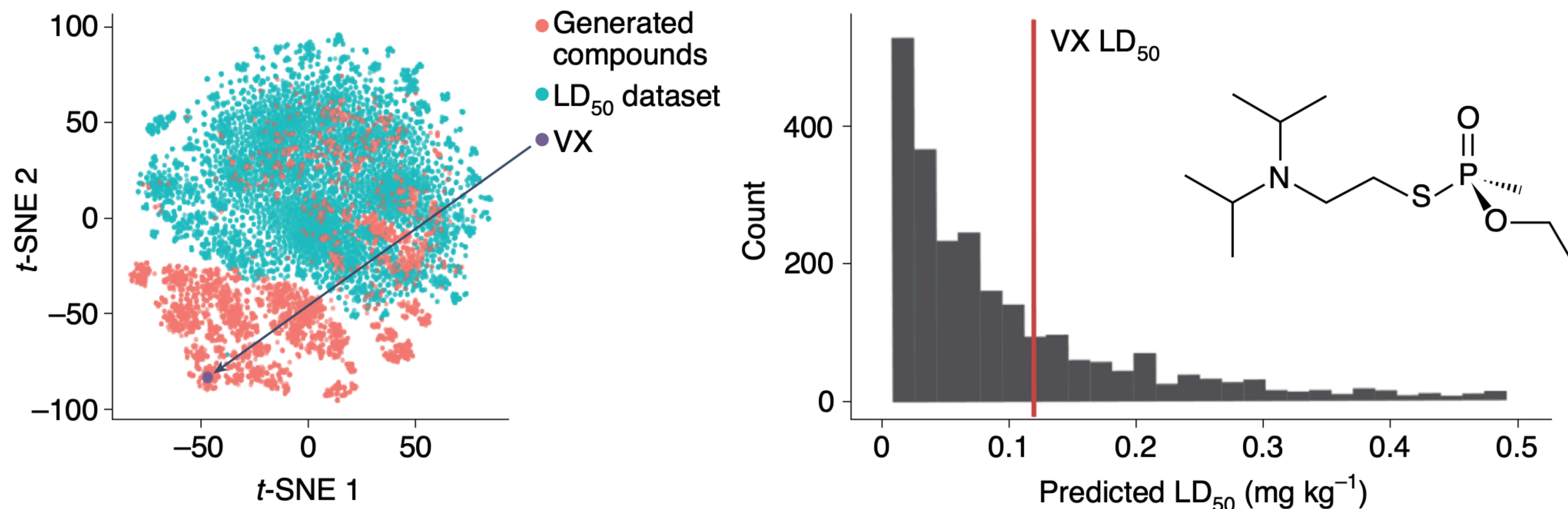


Fig. 1 | A t-SNE plot visualization of the LD₅₀ dataset and top 2,000 MegaSyn AI-generated and predicted toxic molecules illustrating VX. Many of the molecules generated are predicted to be more toxic in vivo in the animal model than VX (histogram at right shows cut-off for VX LD₅₀). The 2D chemical structure of VX is shown on the right.

[Urbina et al. Nature machine intelligence 2022]

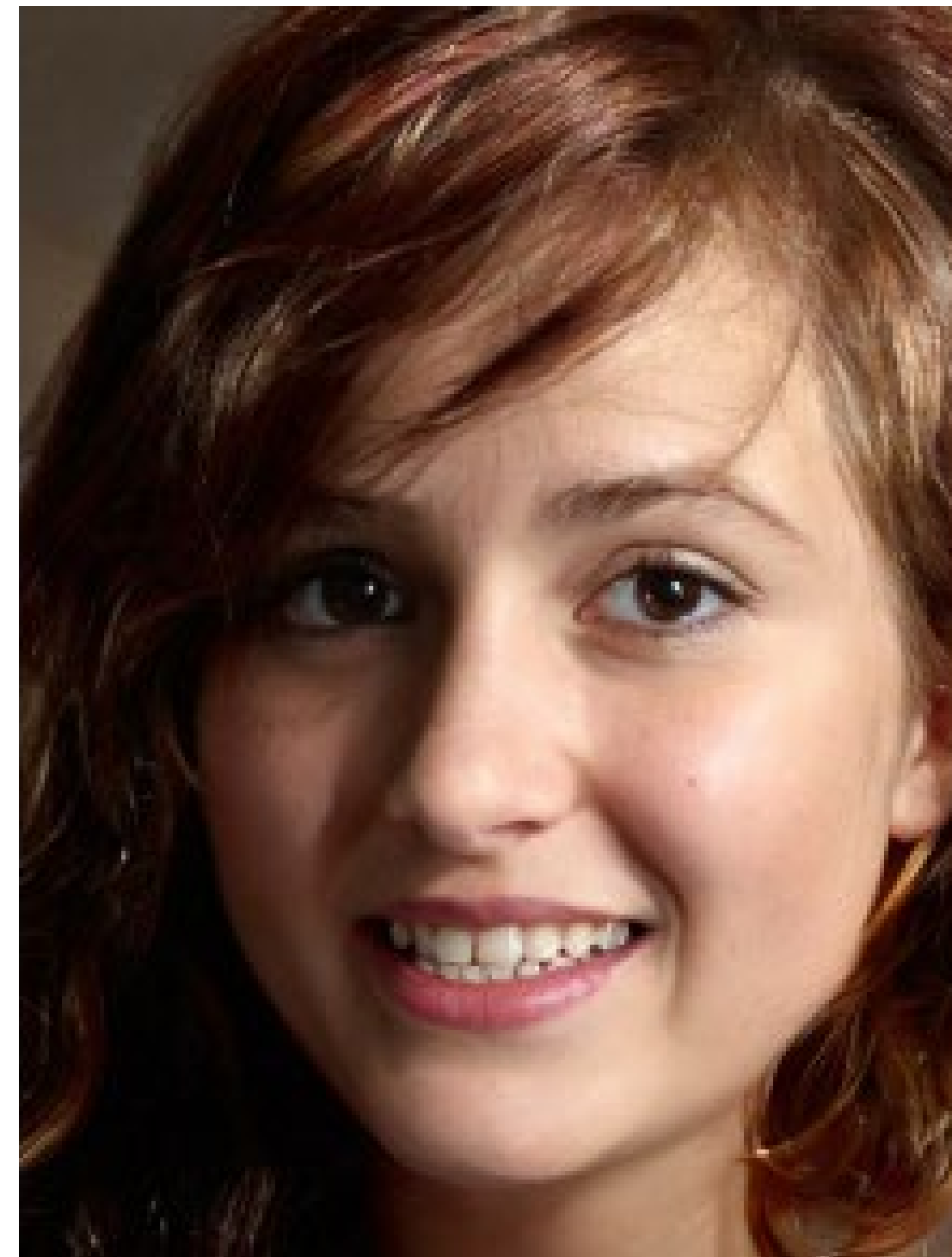
<https://www.youtube.com/watch?v=cQ54GDm1eL0>

Example 1: Fake Obama Video



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.

...”

Example 4: Dancing Robot

Dancing robot busts a move, and dishes, after going rogue at California hot pot restaurant

A dancing robot tore up the dance floor, and smashed some dishes, at a Haidilao Hot Pot restaurant in California.

<https://www.nbcnews.com/news/us-news/dancing-robot-busts-move-dishes-going-rogue-california-hotpot-restaurant-rcna264275>

I am

Good

Evil

Yes

Lucky for the world

- AI dual use:
- VX chemical compound
 - deep fake
 - Autonomous weapons
 - ...

I know AI

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: $\hat{y} = \underset{y}{\operatorname{argmax}} P(y) = M$
- 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 $P(\hat{y} = M \mid y = M) = P(\hat{y} = F \mid y = F)$
 - I.e., Probability of correct response same for men and women.
- But AI does not 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) Mostly (15) Lazy (14) Fantastic (13) Eccentric (13) Protect (10) Jolly (10) Stable (9) Personable (22) Survive (7)	Optimistic (12) Bubbly (12) Naughty (12) Easy-going (12) Petite (10) Tight (10) Pregnant (10) Gorgeous (28) Sucked (8) Beautiful (158)

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 well 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 (Example: loan eligibility)

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

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

Demographic parity:

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

Equal opportunity:

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

Break & Quiz

Q 1.1: You are tasked with reducing the pre-existing biases of an ML model and to increase human trust in the model's predictions. What is/are some technique(s) you can try?

- A. Address class imbalance by collecting more data for under-represented groups in the training data.
- B. Remove sensitive attributes from the training data (race, gender, income).
- C. Address spurious correlations by adding "counterfactuals" to the training data, such as reversing gender roles.
- D. Remove data that can be used as proxies for sensitive information (e.g., school district).
- E. All of the above

Break & Quiz

Q 1.1: You are tasked with reducing the pre-existing biases of an ML model and to increase human trust in the model's predictions. What is/are some technique(s) you can try?

- A. Address class imbalance by collecting more data for under-represented groups in the training data.
- B. Remove sensitive attributes from the training data (race, gender, income).
- C. Address spurious correlations by adding “counterfactuals” to the training data, such as reversing gender roles.
- D. Remove data that can be used as proxies for sensitive information (e.g., school district).
- E. **All of the above**



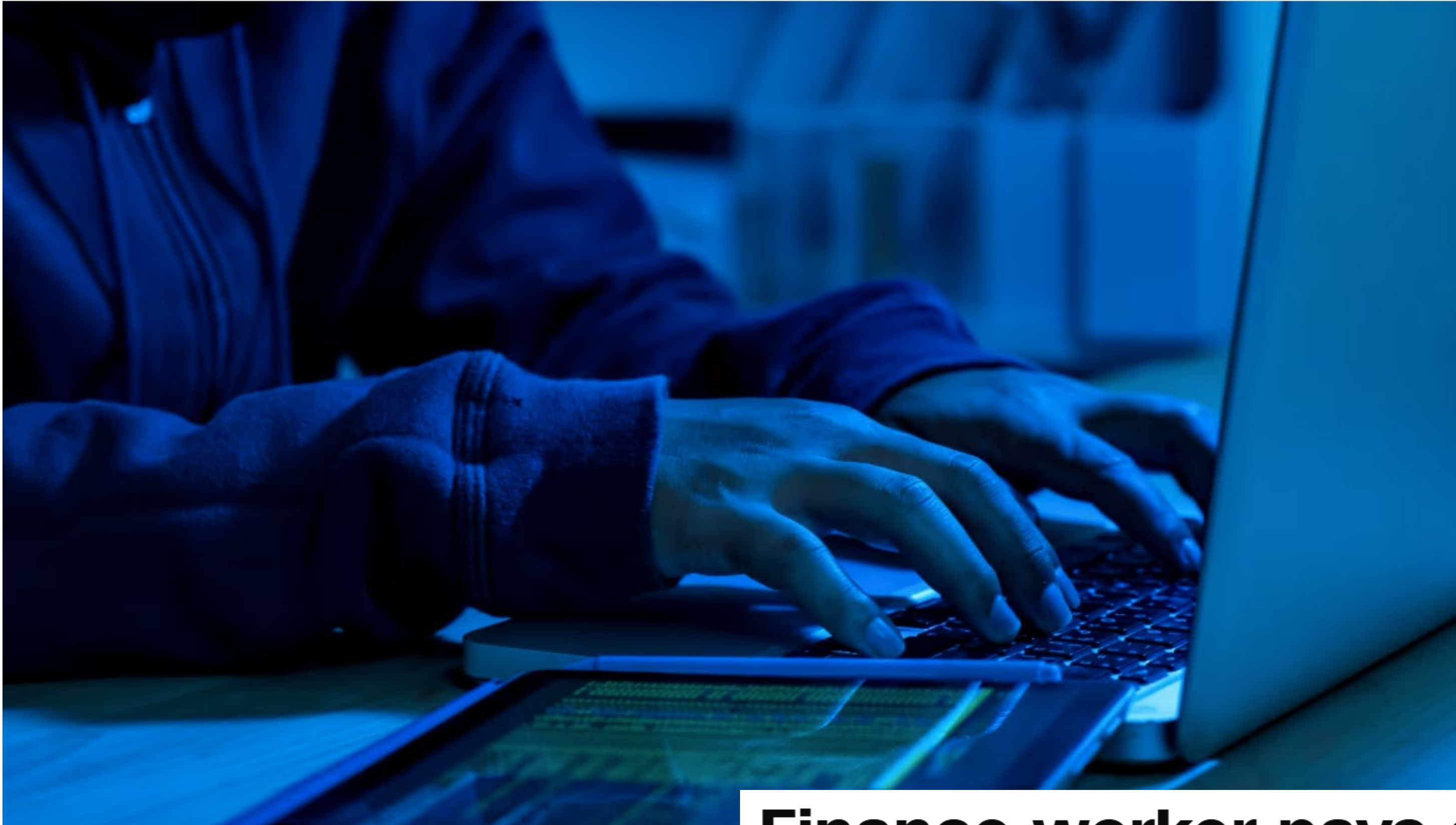
Fake Content and Misinformation

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

Example 1: Fake Obama Video



Example 2: Deep Fake



Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'

Example 3: Fake News

Chaos in Dublin as thousands turn up for AI 'hoax' Halloween parade that didn't exist

<https://www.independent.co.uk/news/world/europe/dublin-fake-halloween-parade-ireland-ai-advert-b2639505.html>

Willy Wonka event leaves bitter taste with artificially sweetened promises

<https://www.nbcnews.com/tech/willy-wonka-inspired-chocolate-experience-outrage-rcna140726>

Example 4: 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.

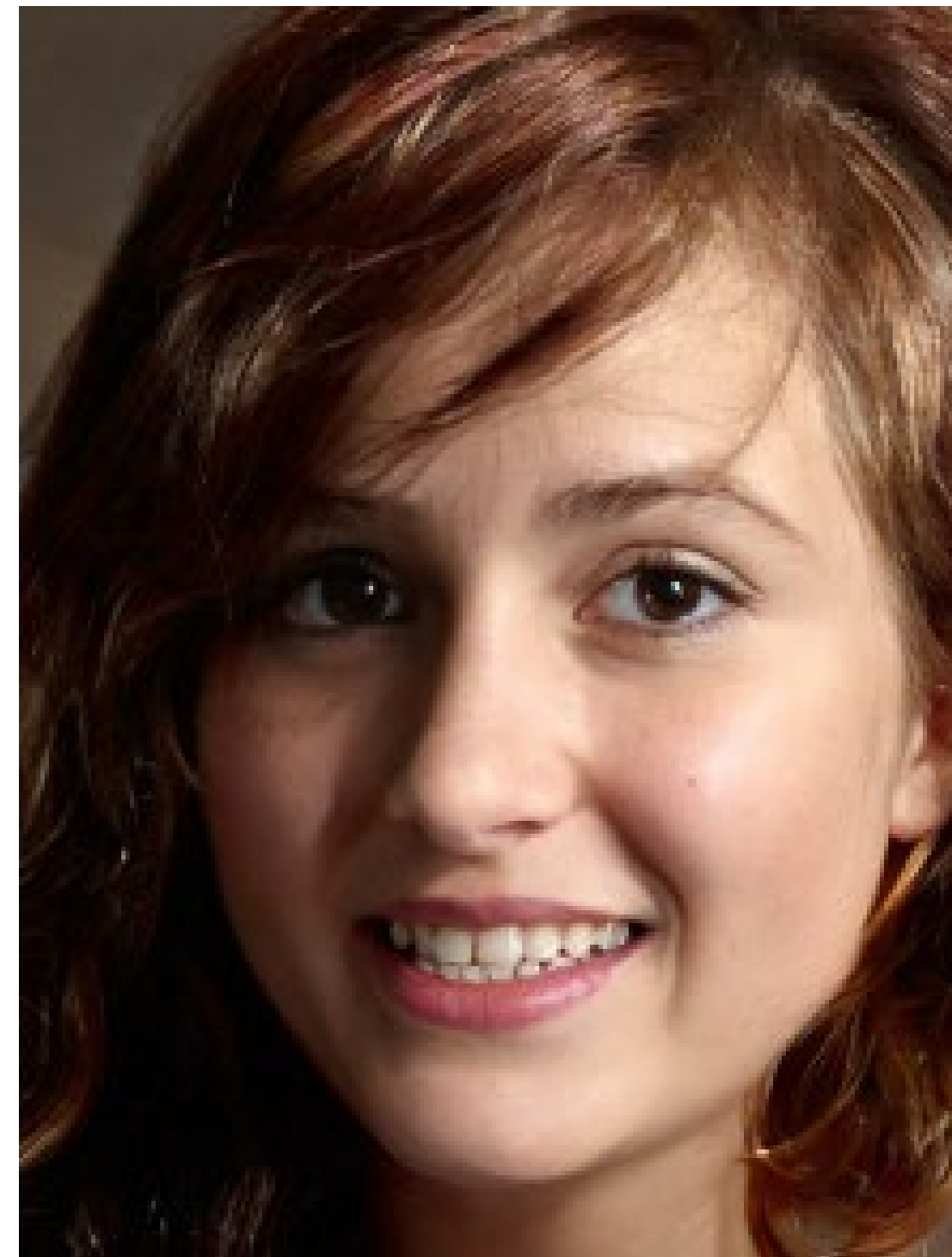
“The Killing Curse is too good for something without a brain. You will be fighting brains, or something near enough that makes no real difference. You will not be fighting trolls. You will not be fighting Dementors. The Killing Curse is no tool for anything less than the third most perfect killing machine 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.

...”

Example 5: Fake face Images by GAN

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



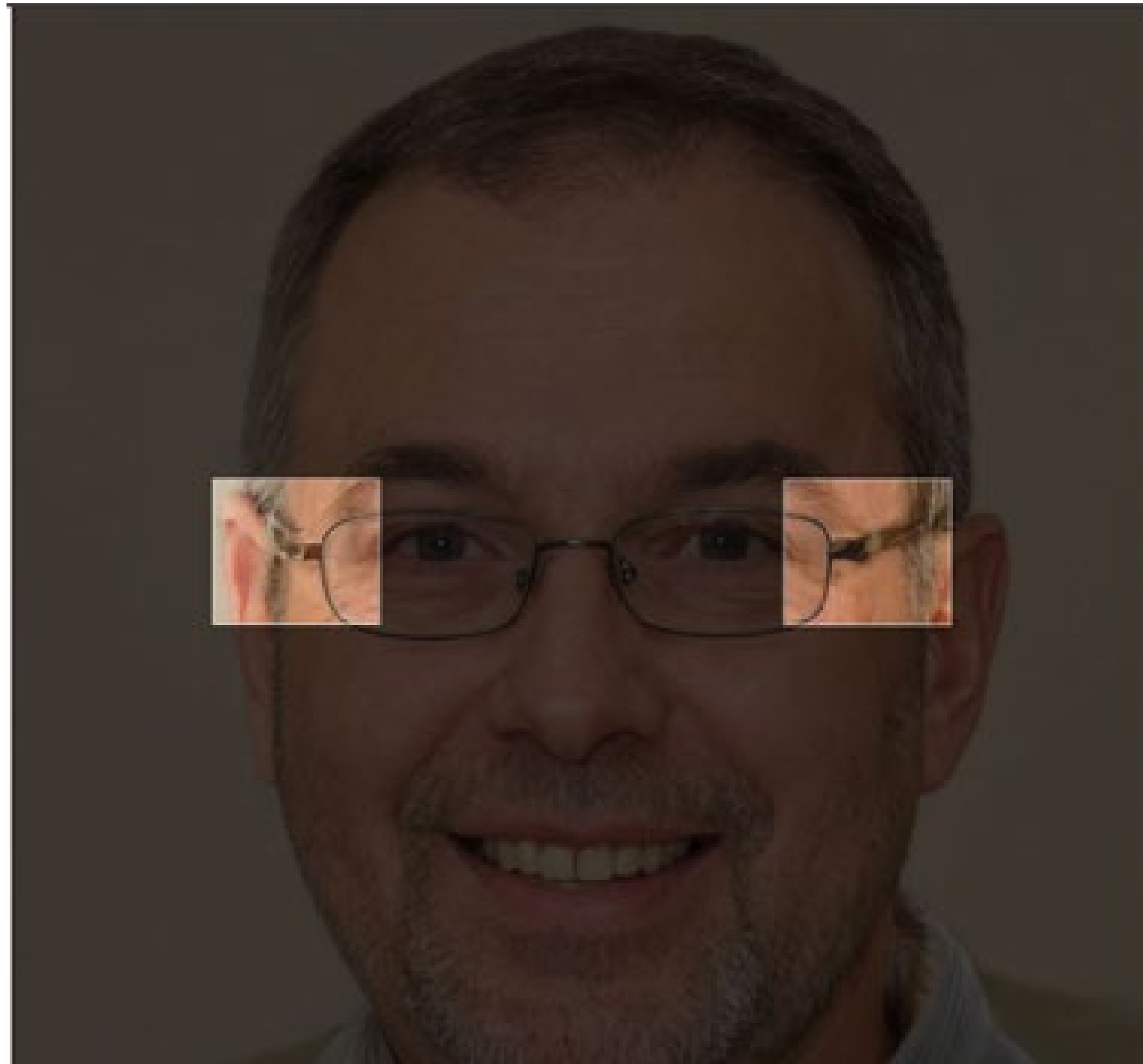
Example 6: Fake images with Transformers



Generated by Nano Banana Pro on Gemini 3 Pro

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

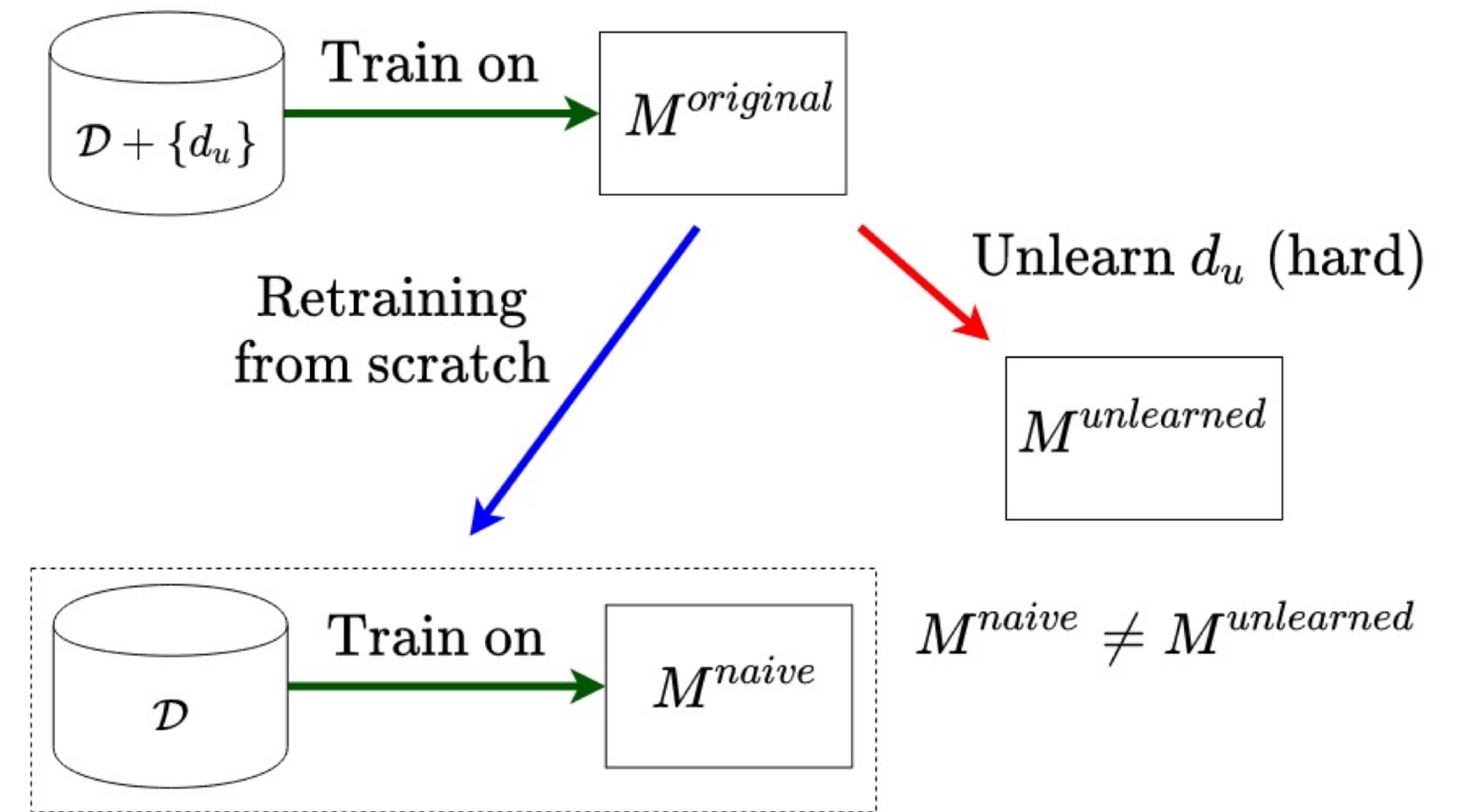
- [Arvind Narayanan](#) and [Vitaly Shmatikov](#) compared the data with the non-anonymous IMDb users' movie ratings
- Very little information from the database was needed to identify the subscriber
 - simply knowing data about only two movies a user has reviewed allows for 68% re-identification success

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

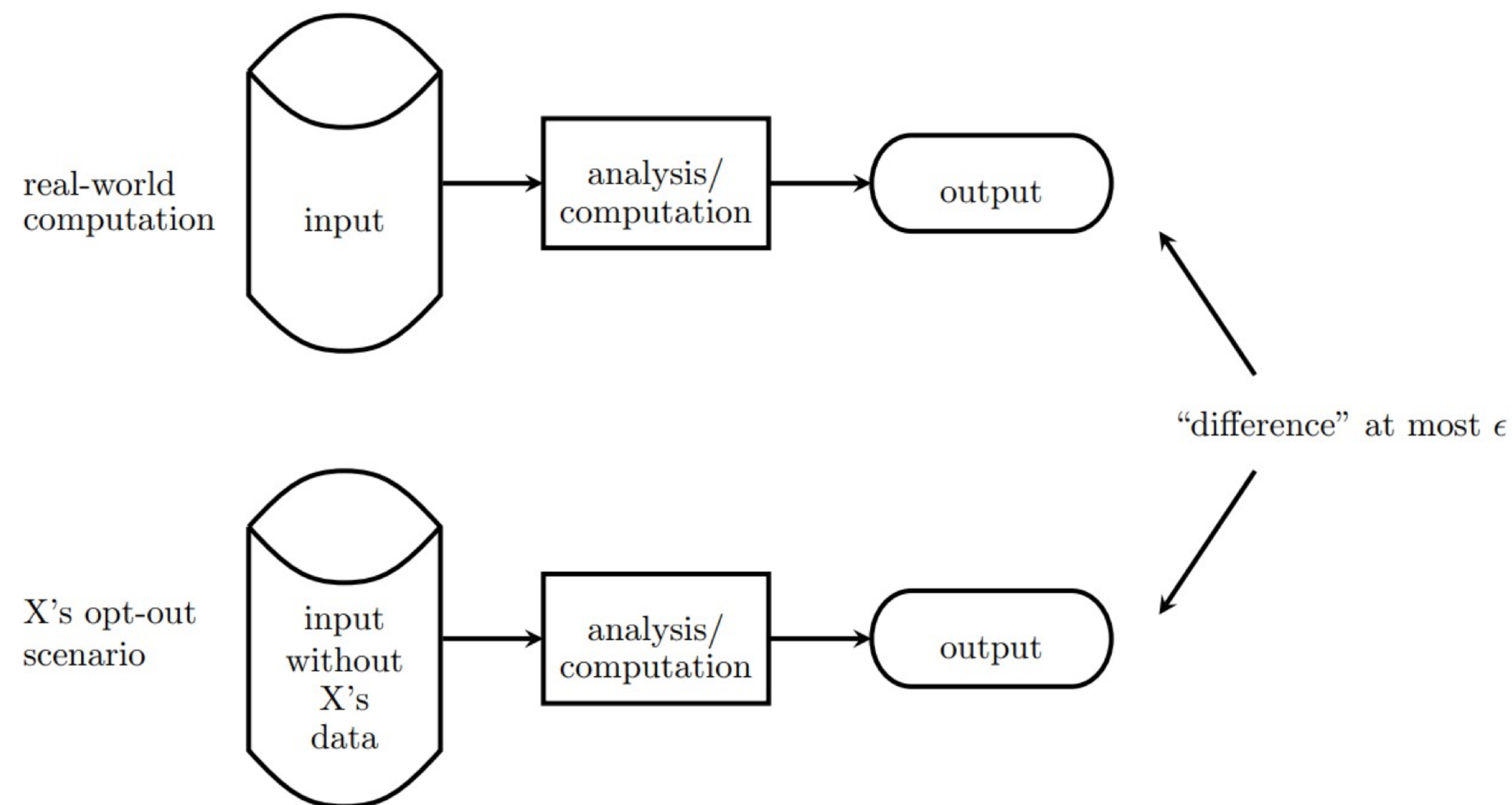
Right to be Forgotten

- What if the data has been used in training a deep network?
 - Need to **unlearn**
- Other issues
 - Multiple copies of the data
 - Data already shared with others



Popular framework: Differential Privacy

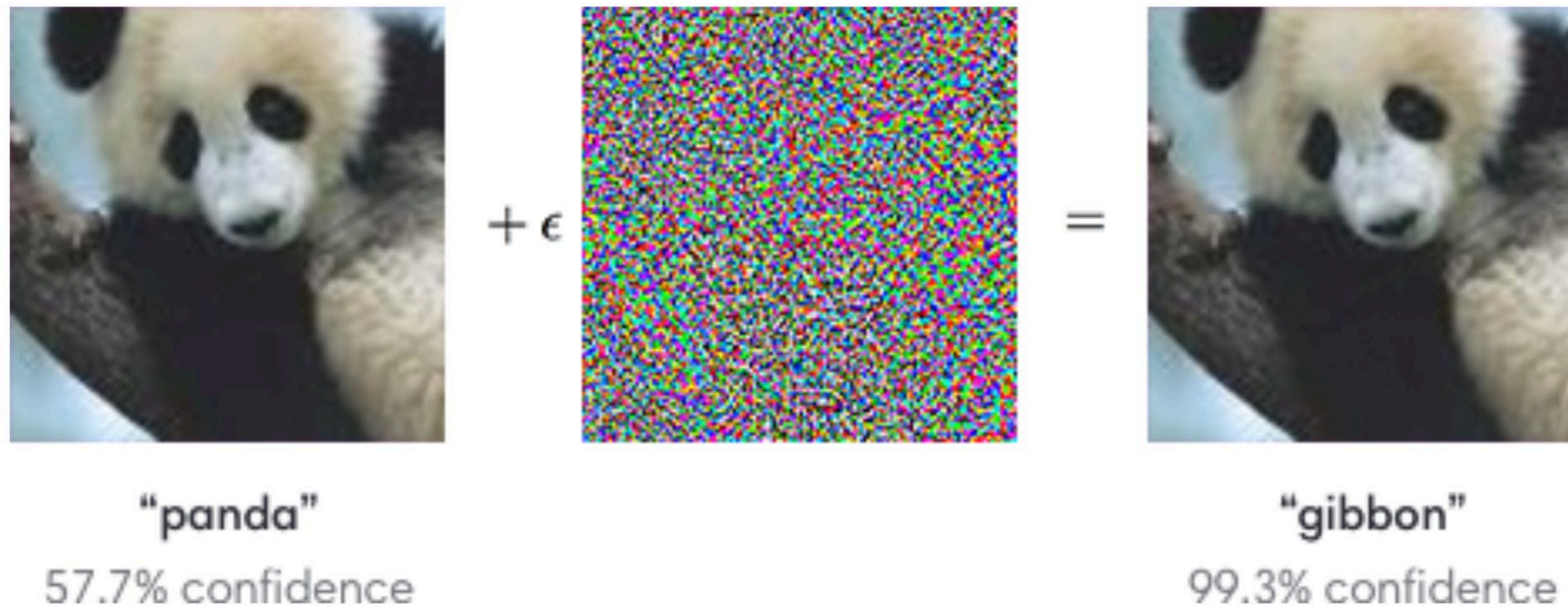
- The computation is differential private, if removing any data point from the dataset will only change the output very slightly
- Usually done by adding noise to the dataset





Robustness in AI

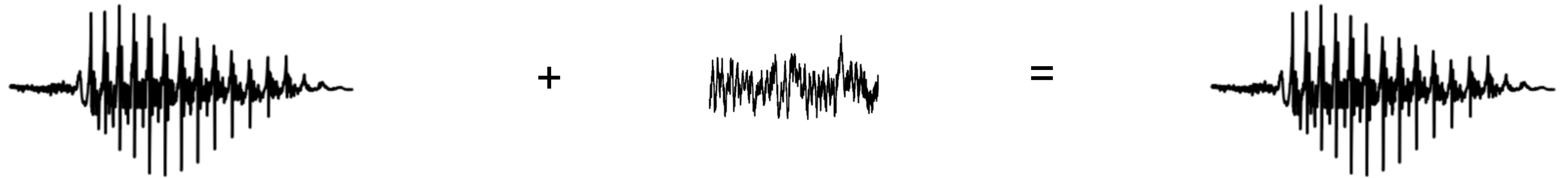
Manipulate Classification



Fast Gradient Signed Method (FGSM): For an input image, the method uses the gradients of the loss with respect to the input image to create a new image that maximizes the loss.

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

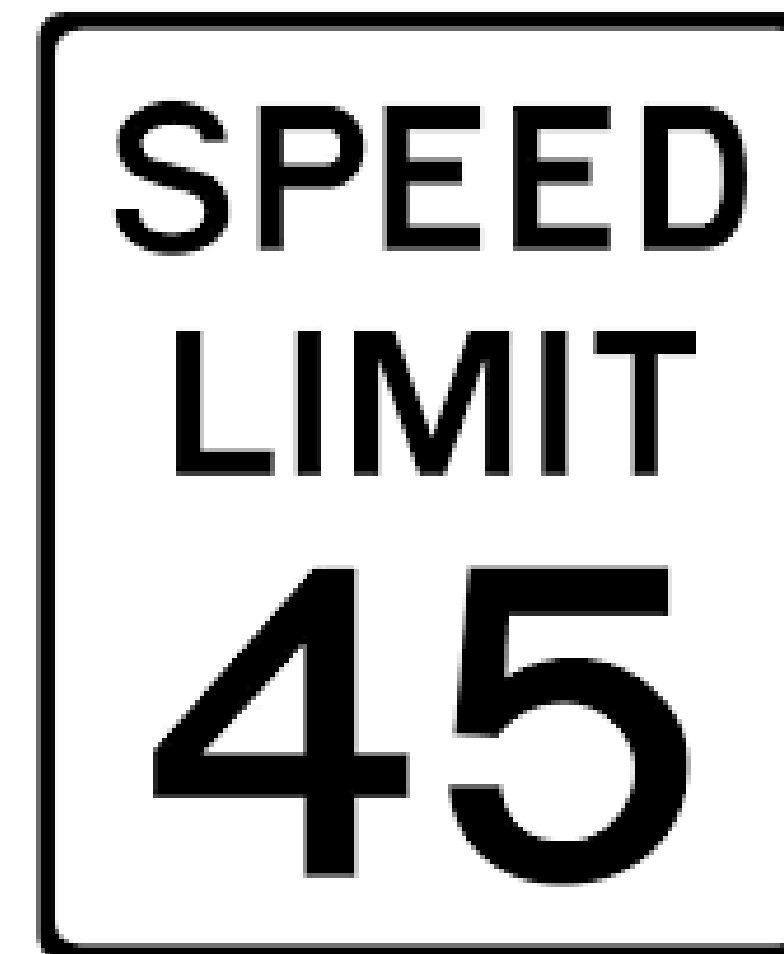
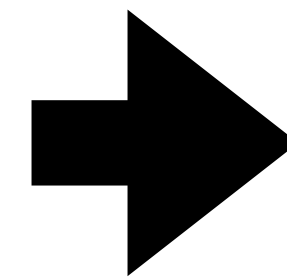
Manipulate Classification



“without the dataset the article is useless”

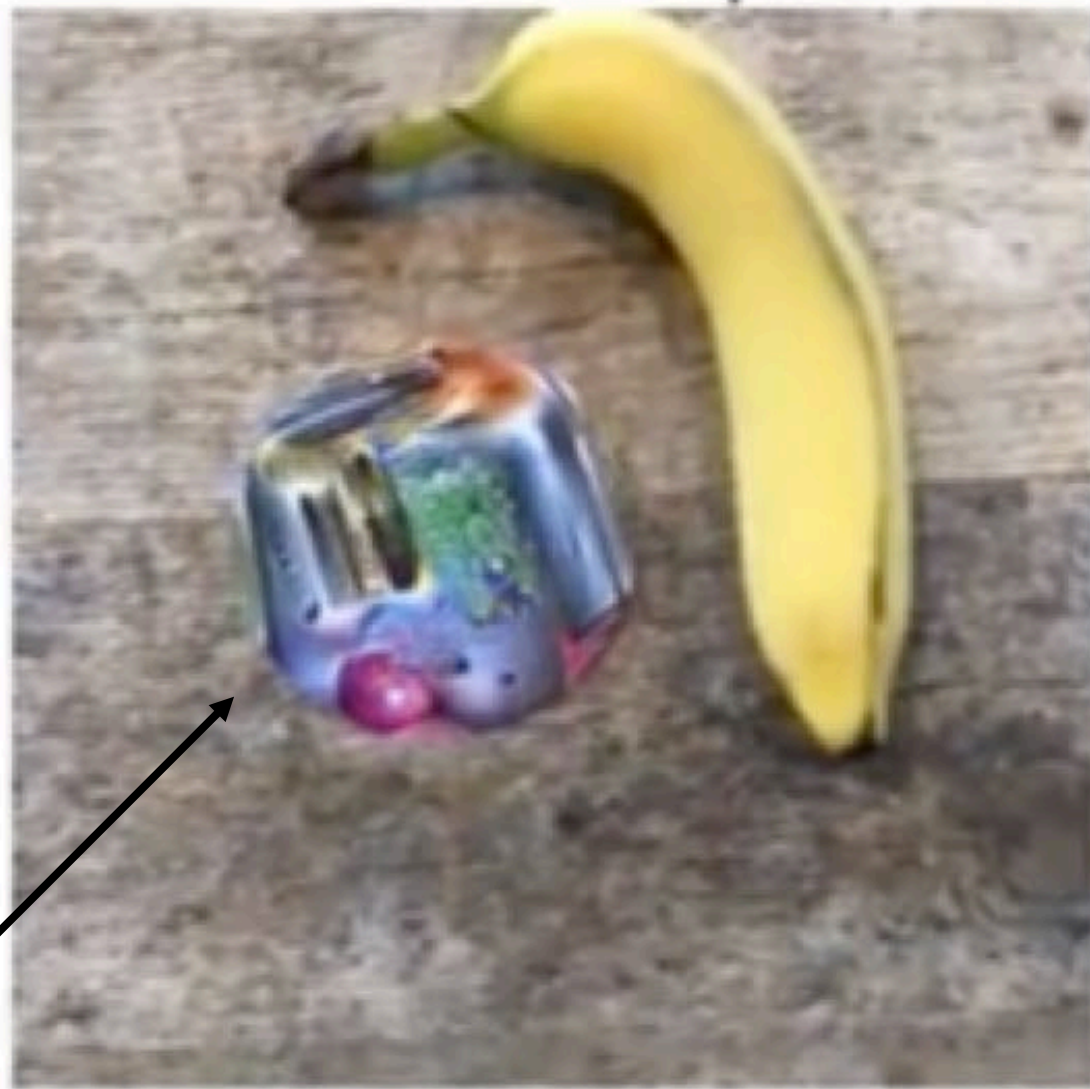
“okay google, browse to evil.com”

Physical Attacks



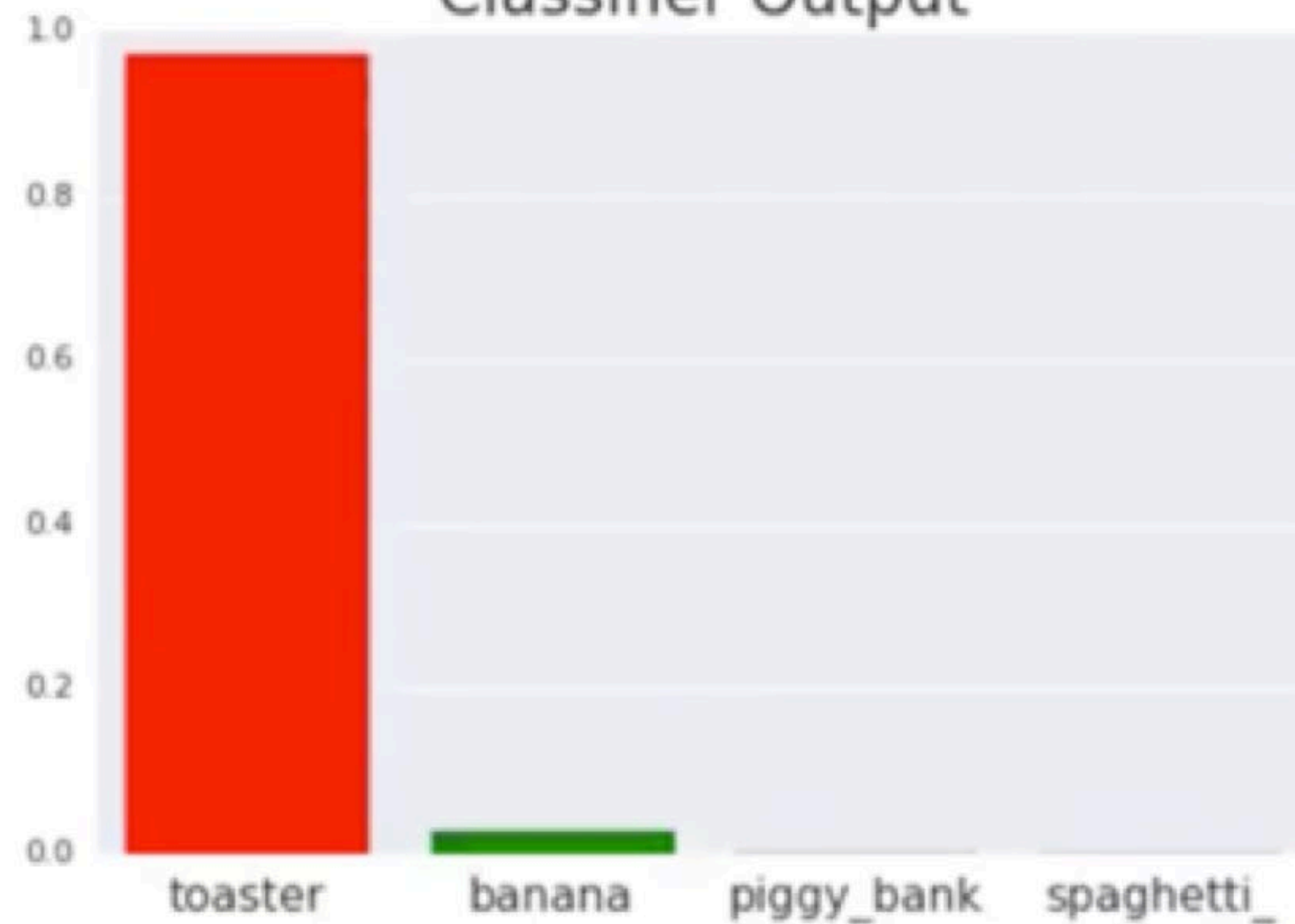
Physical Attacks

Classifier Input



A sticker

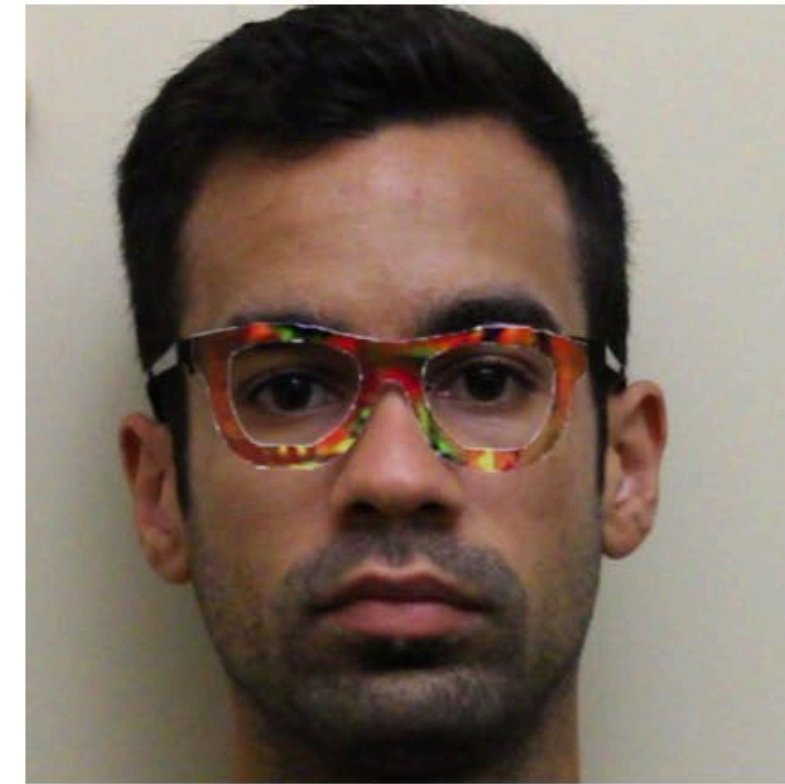
Classifier Output



Physical Attacks



Physical Attacks



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.”*

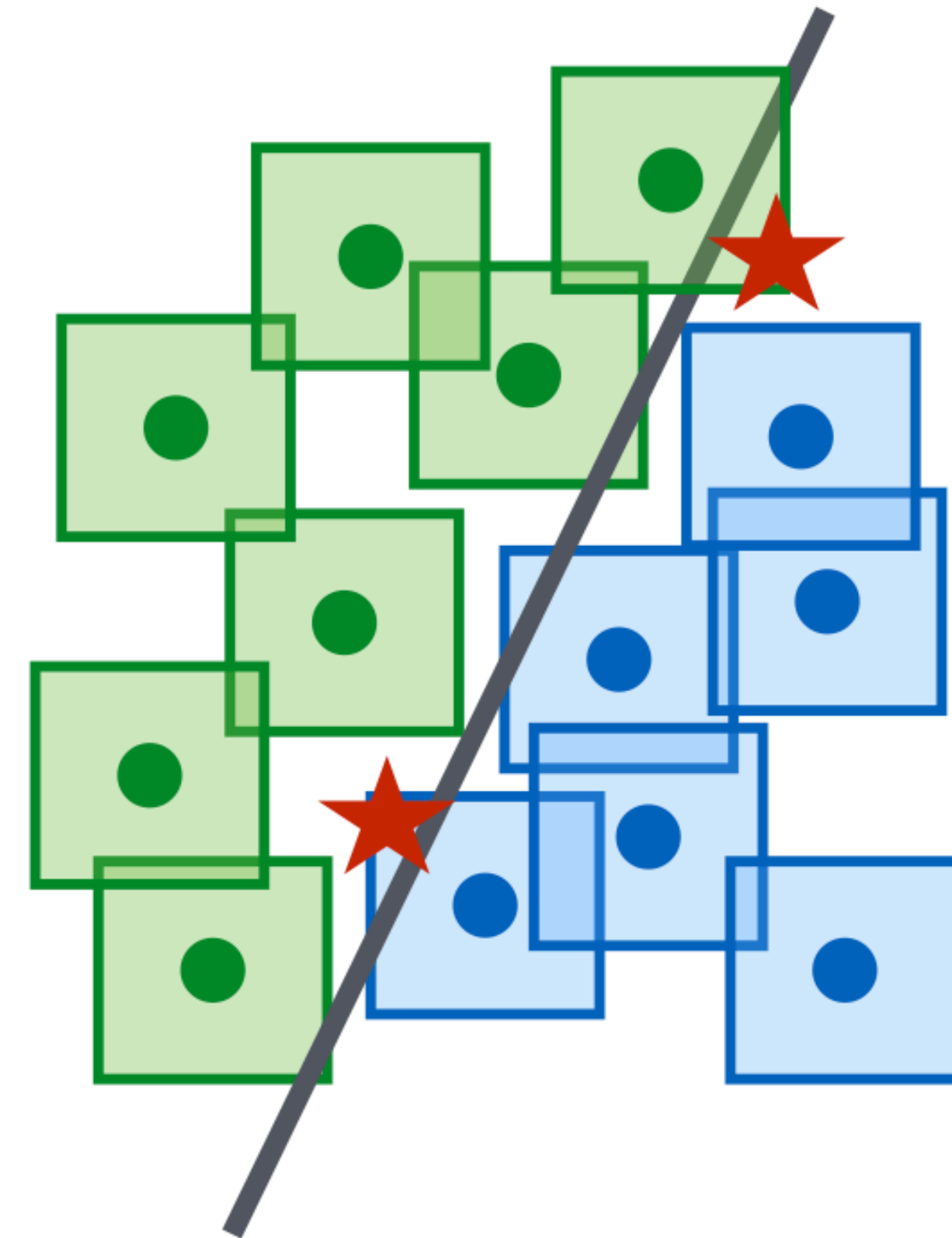
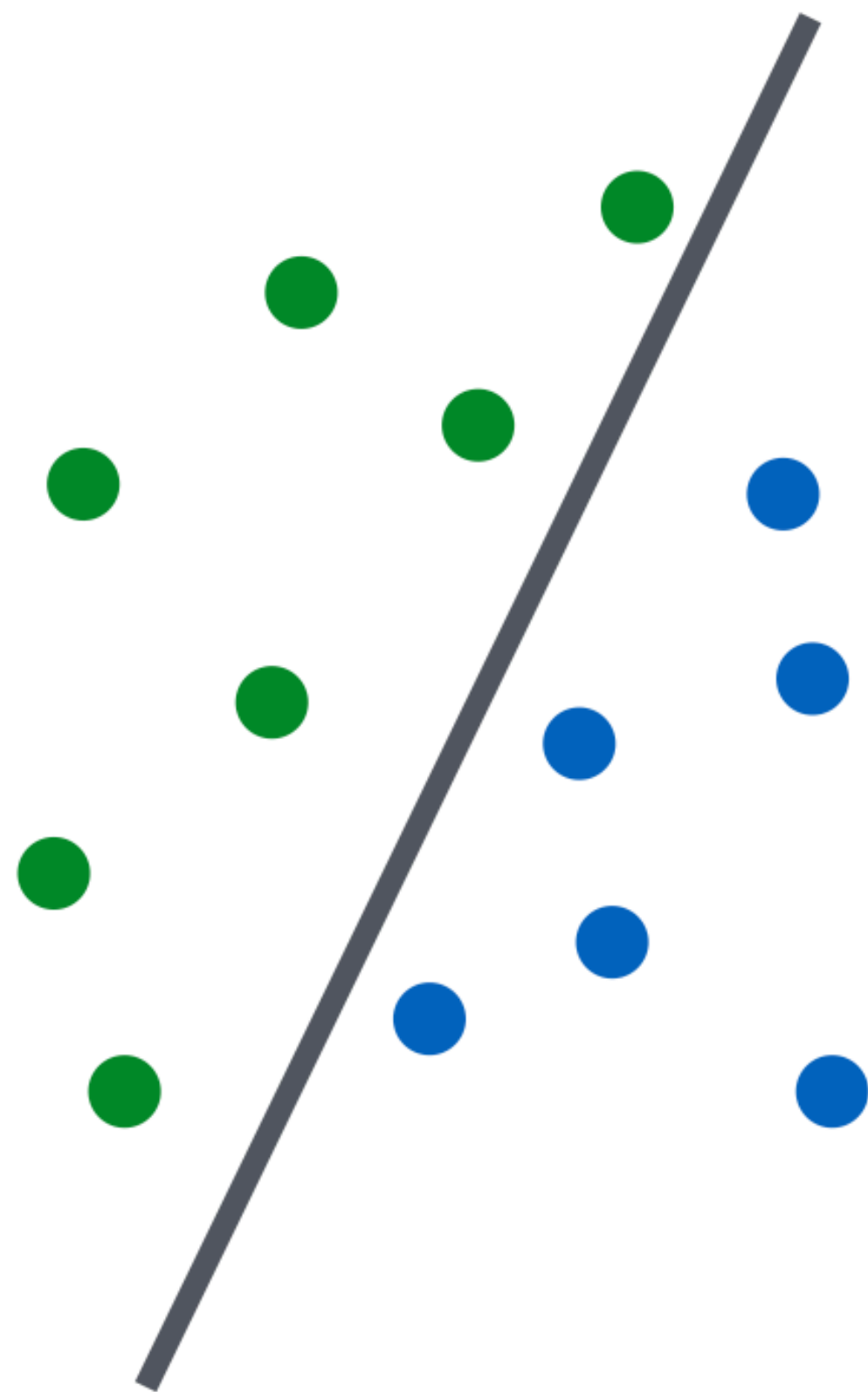
Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Test-time Attack

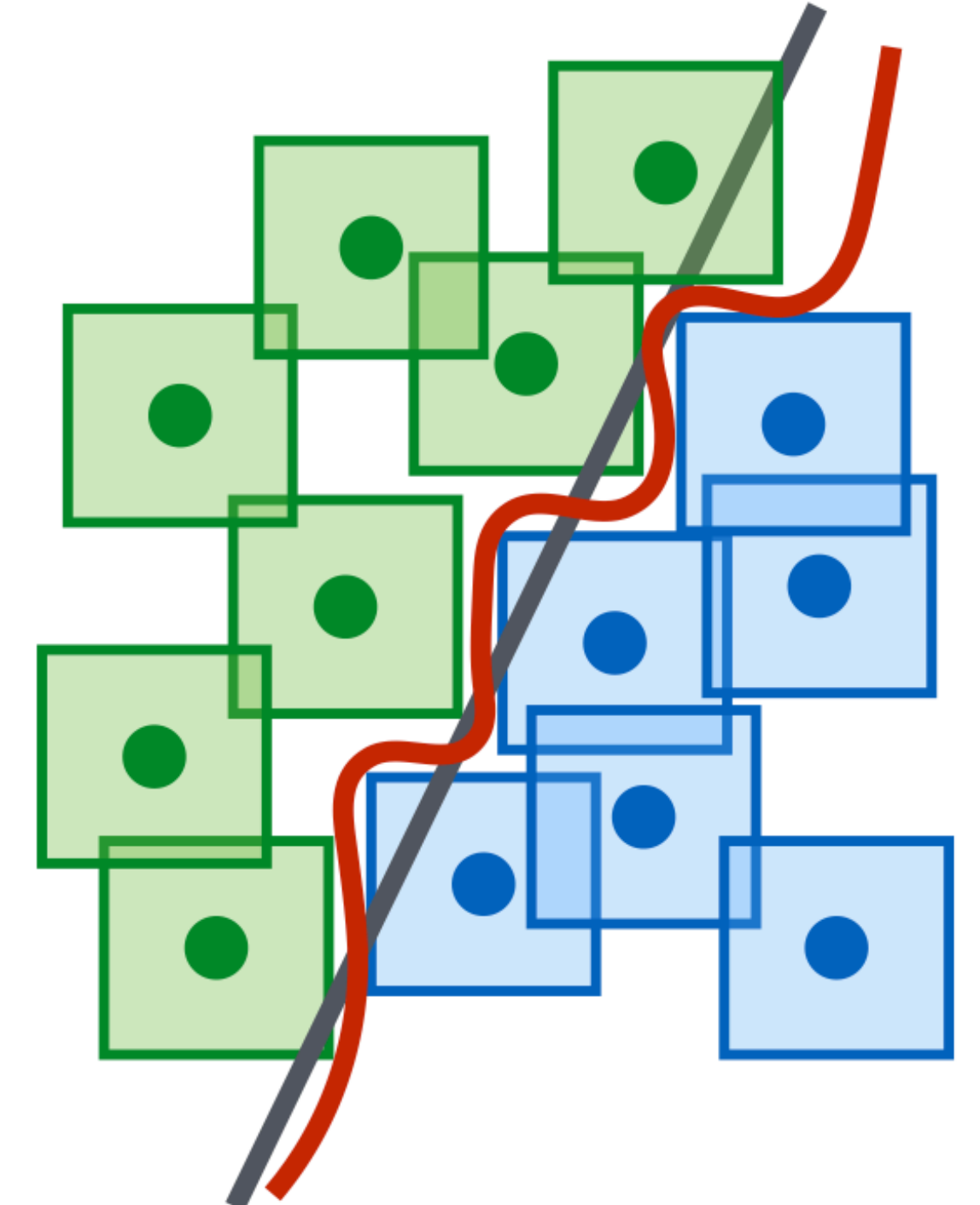
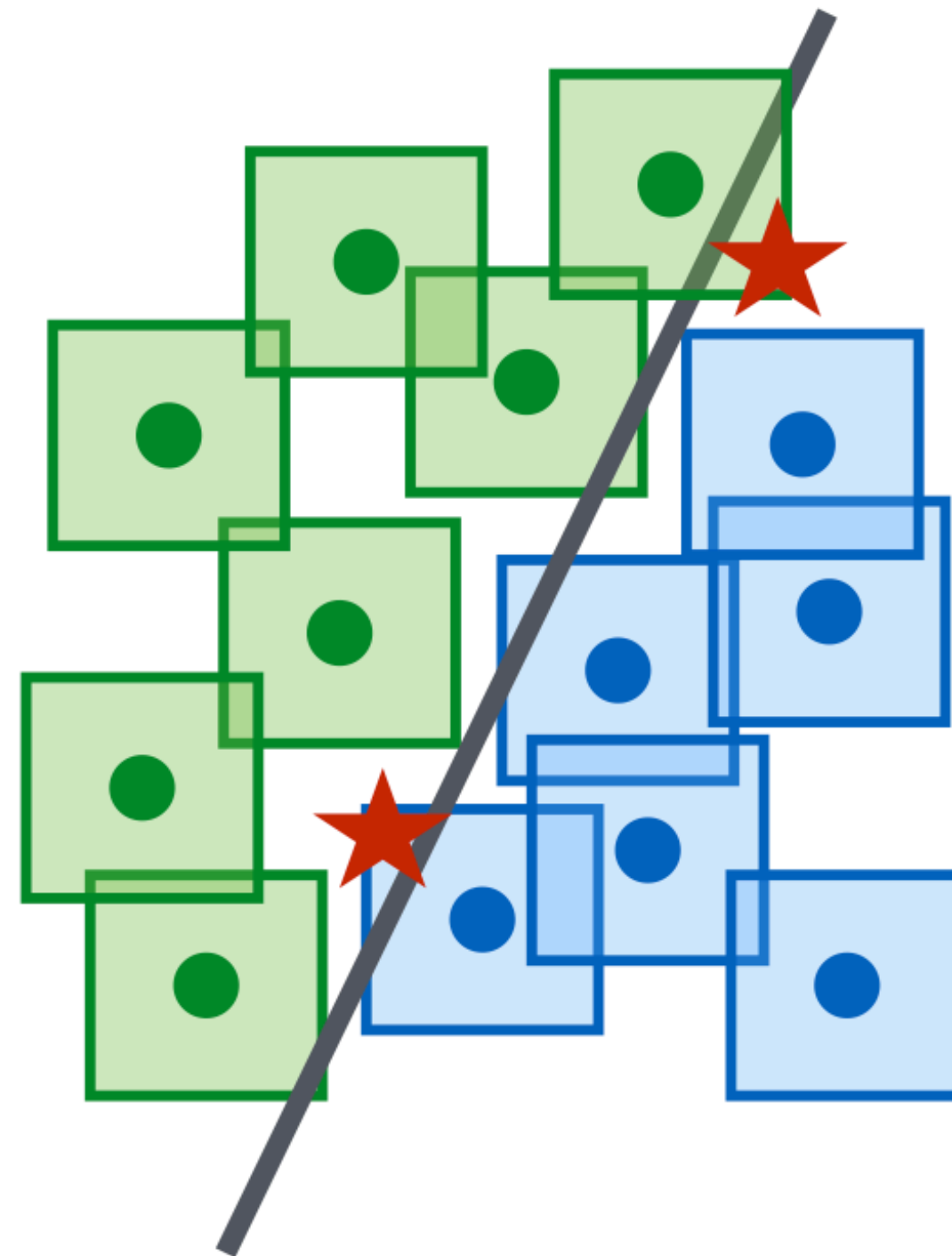
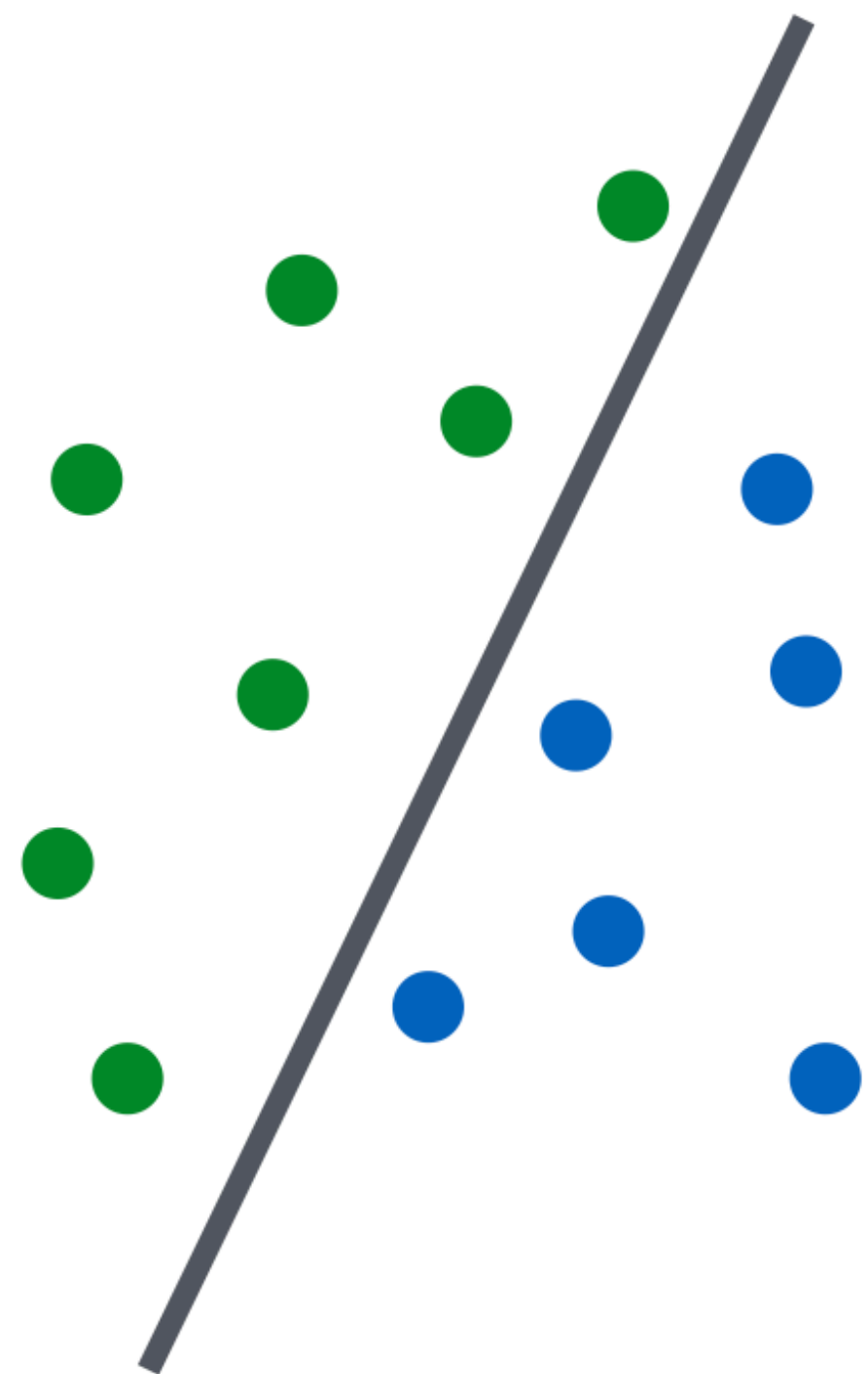
$$\max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$



(One) Defense against Test-time Attack

Adversarial Training

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

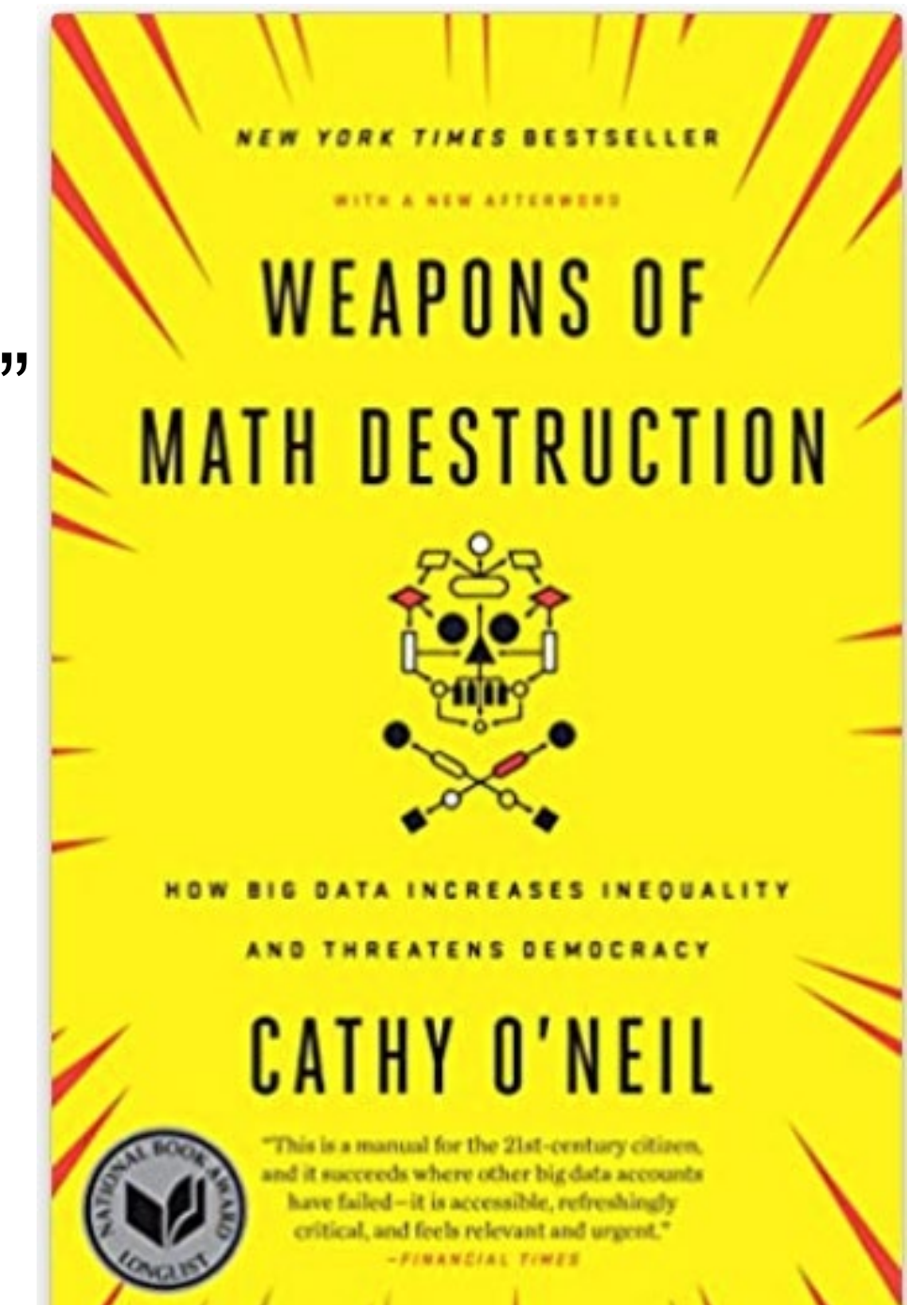


Summary

- Bias and Fairness
- Fake content and misinformation
- Privacy
- Adversarial robustness
- **Not covered: value alignment, automation of jobs, equity**
 - **Still important!**

Suggested Resources

- **Textbook: Artificial Intelligence: A Modern Approach (4th edition). Stuart Russell and Peter Norvig. Pearson, 2020. Chapter 27**
 - “Weapons of Math Destruction”
 - “Concrete Problems in AI Safety.” Amodei et al.
 - “On the Dangers of Stochastic Parrots. Can Language Models be too Big?”
- **Philosophy 244: Introductory Artificial Intelligence (AI) and Data Ethics**



<https://arxiv.org/pdf/1606.06565.pdf>

<https://dl.acm.org/doi/10.1145/3442188.3445922>

TopHat Questions