COMP9208
Artificial Intelligence and Society

Lecture 12: 31 October 2025

Prof. Mikhail Prokopenko
Dr. Sheryl Chang

Canvas: https://canvas.sydney.edu.au/courses/66452

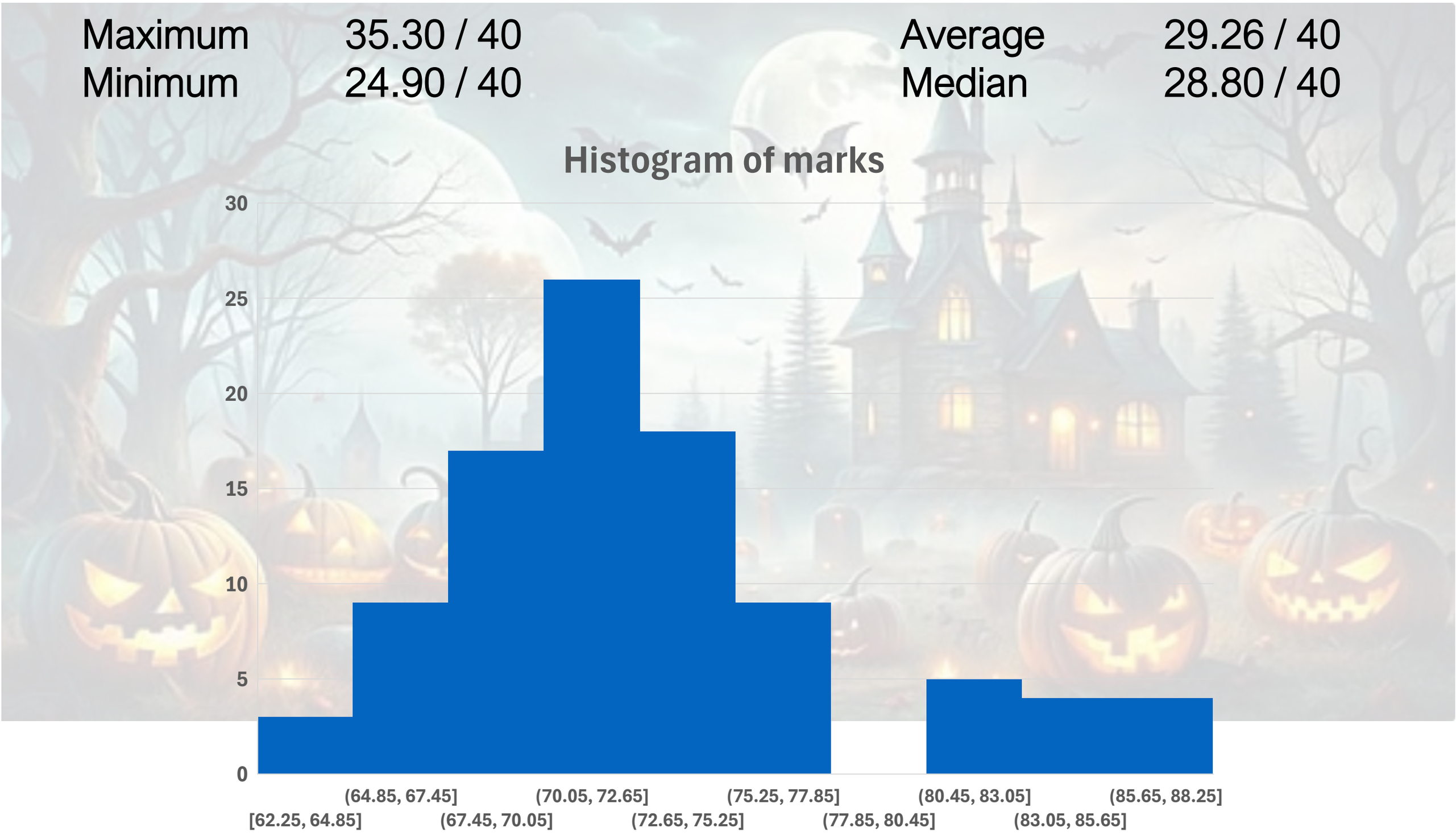
Email: mikhail.prokopenko@sydney.edu.au
Office: J12 (Computer Science), level 3W, room 324

Date	Week: Lecture	Assignment
08/08	1: History of AI	
15/08	2: History of AI + Perception and Actuation	
22/08	3: History of AI + Representation and Reasoning	
29/08	4: Machine Learning	
05/09	5: Natural Interaction	
12/09	6: Distributed AI (swarm intelligence)	
19/09	7: From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)	# 1: Individual assignment (article review)
26/09	8: Goal-driven AI: search & rescue vs lethal autonomous weapon systems (LAWS)	
03/10	**BREAK**	
10/10	9: Review: language models, search trees, situated behaviours, neural networks	
17/10	10: Robotic swarms: RoboCop vs RoboCup	
24/10	11: Group Presentations	# 2: Group assignment
31/10	12: Adversarial Machine Learning → H70.02.2140	
07/11	13: AI Literacy; perception and impact of AI on society	
17/11	14: STUVAC	# 3: Individual assignment (data analysis report)

Assignment #2 review

Assignment #2 review

Maximum	35.30 / 40	Average	29.26 / 40
Minimum	24.90 / 40	Median	28.80 / 40

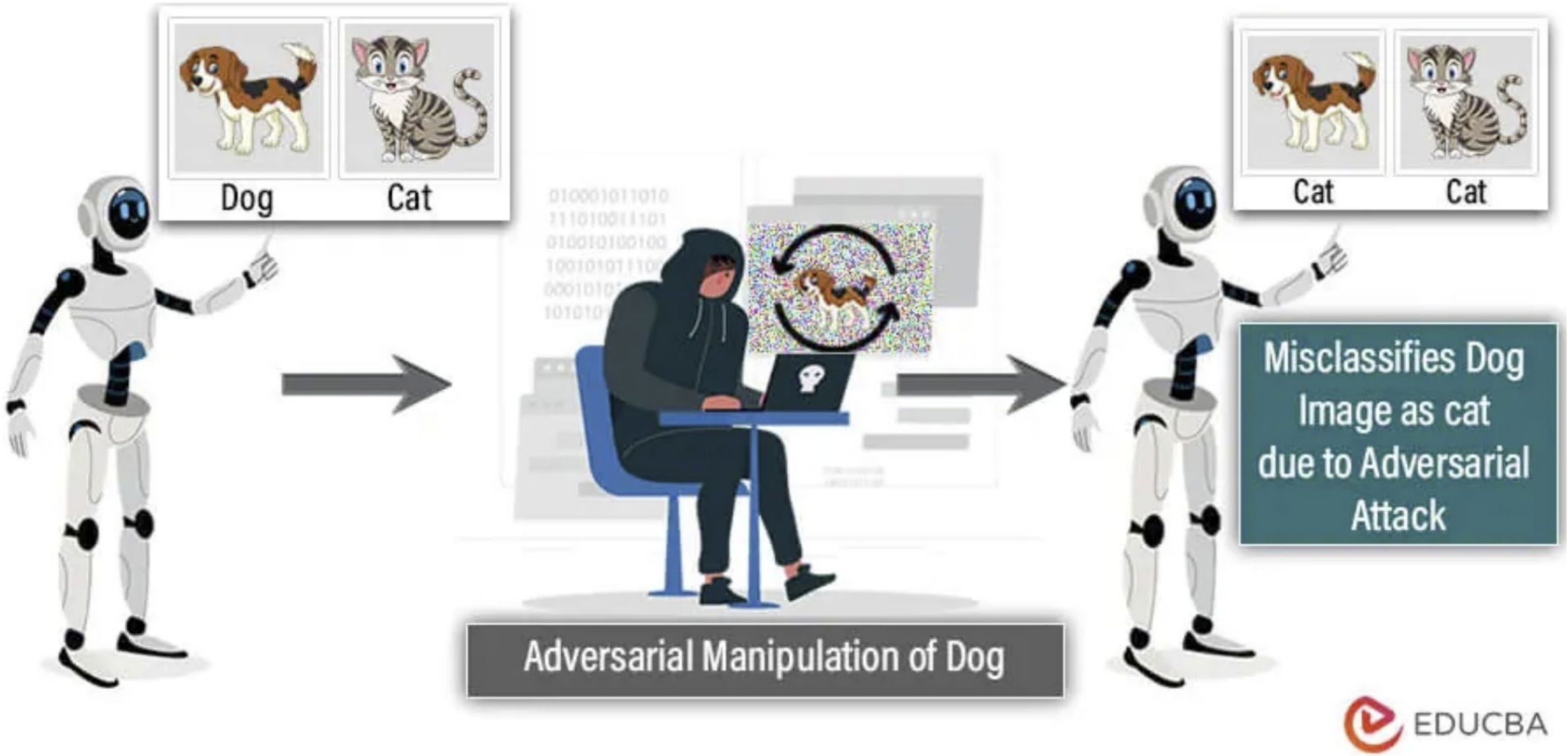


Questions

?

Adversarial Machine Learning

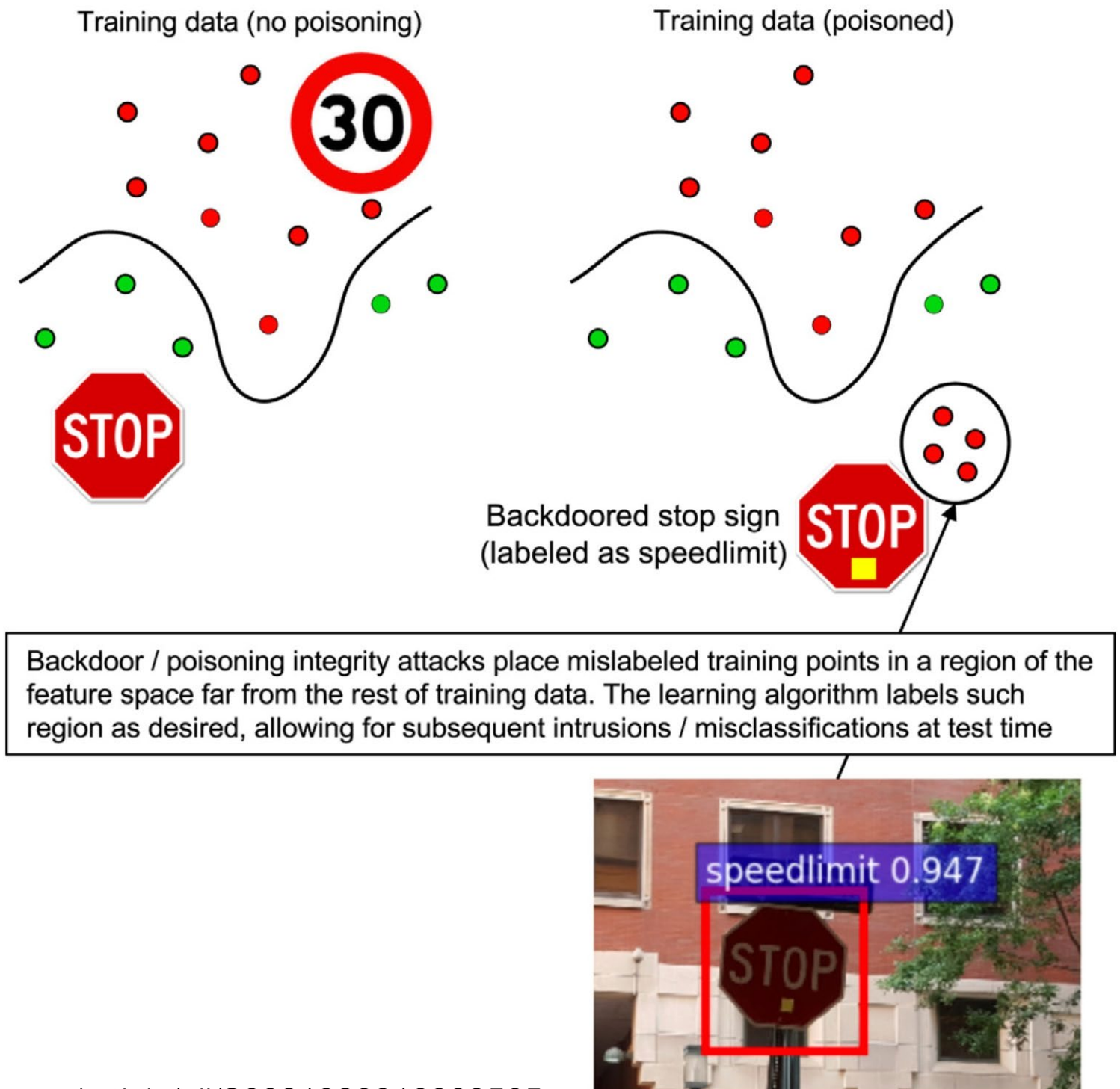
Attacks on machine learning systems



https://www.educba.com/adversarial-machine-learning/

Adversarial Machine Learning

Attacks on machine learning systems

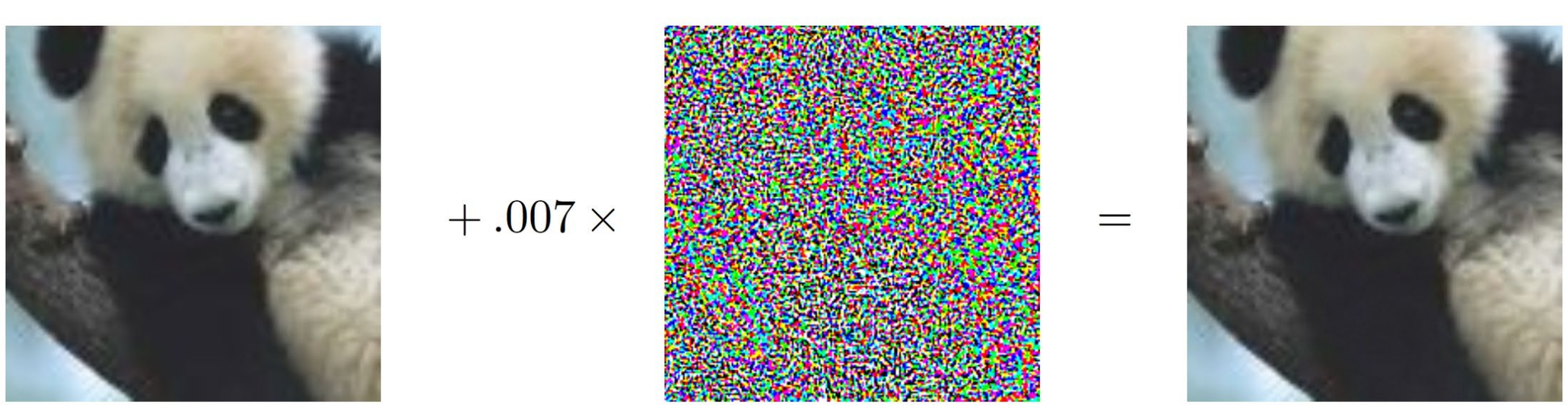


https://www.sciencedirect.com/science/article/pii/S0031320318302565

Adversarial Machine Learning

Attacks on machine learning systems:

- **data poisoning**: contaminating the training dataset with data designed to increase errors in the output
  - example: an attacker might add malicious data to a dataset used to train a predictive model, causing it to make incorrect predictions
  - example: an attacker might add fake data to a dataset used to train a spam filter, causing it to incorrectly classify legitimate emails as spam



https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3
https://arxiv.org/pdf/1412.6572

Adversarial Machine Learning

Attacks on machine learning systems:

- **data poisoning**: contaminating the training dataset with data designed to increase errors in the output
  - example: an attacker might add malicious data to a dataset used to train a predictive model, causing it to make incorrect predictions
  - example: an attacker might add fake data to a dataset used to train a spam filter, causing it to incorrectly classify legitimate emails as spam
- **evasion attacks**: modifying training samples to make them classified as legitimate
  - example: an attacker might add noise to an image to make it misclassified by a facial recognition system

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **data poisoning**: contaminating the training dataset with data designed to increase errors in the output
  - example: an attacker might add malicious data to a dataset used to train a predictive model, causing it to make incorrect predictions
  - example: an attacker might add fake data to a dataset used to train a spam filter, causing it to incorrectly classify legitimate emails as spam
- **evasion attacks**: modifying training samples to make them classified as legitimate
  - example: an attacker might add noise to an image to make it misclassified by a facial recognition system
- **backdoor attacks**: inserting a "backdoor" into an ML model and then teaching a specific behavior that can be triggered by predefined inputs or conditions (a small defect on images, sounds, videos or texts)
  - example: an attacker might insert a backdoor into a self-driving car's ML model, causing it to malfunction or behave erratically under certain conditions
  - example: an attacker might insert a Trojan horse into a medical diagnosis model, causing it to misdiagnose patients

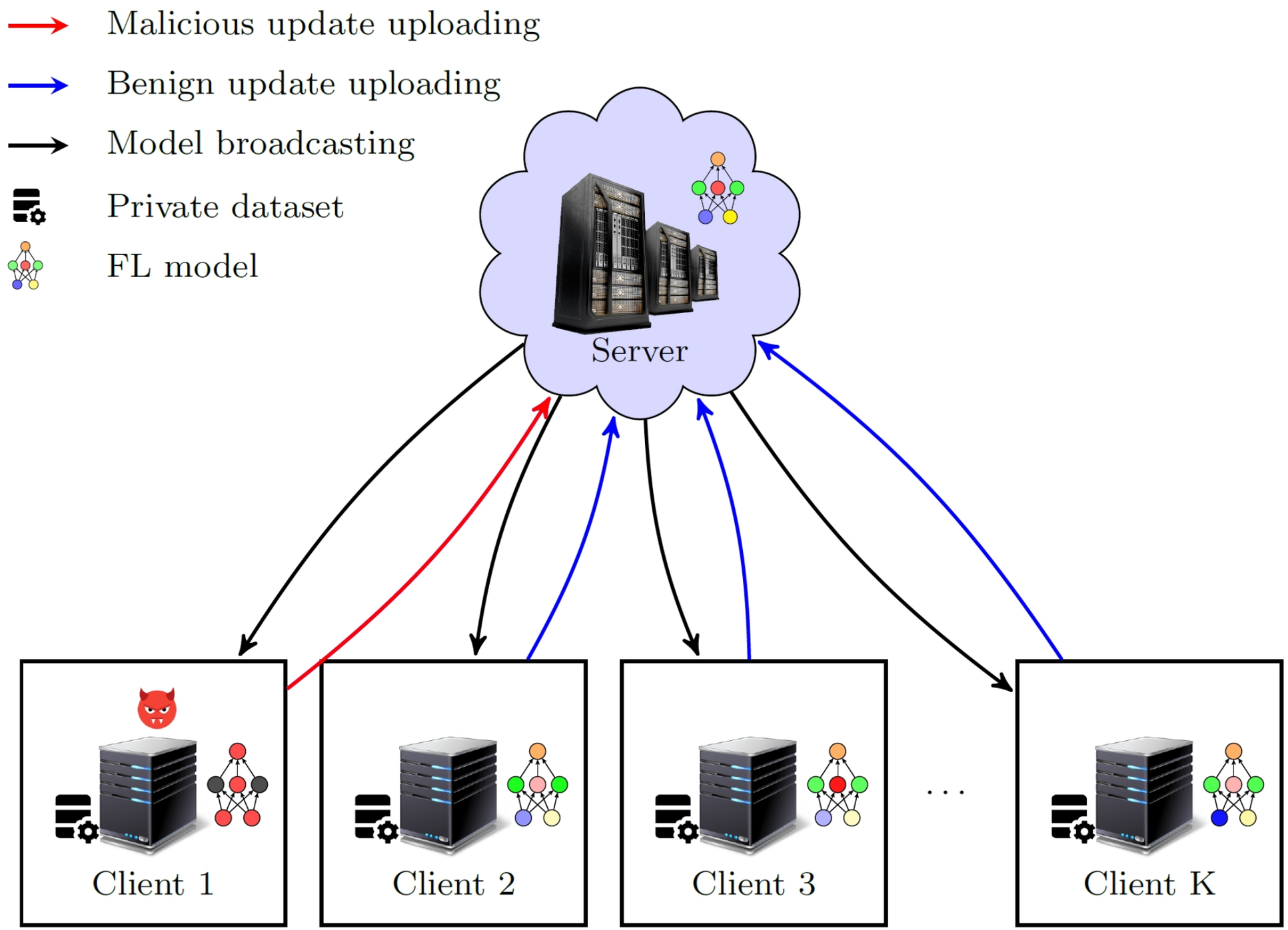
https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Questions

?

Adversarial Machine Learning

Attacks on machine learning systems



https://arxiv.org/pdf/2302.07173

Adversarial Machine Learning

Attacks on machine learning systems:

- **Byzantine attacks on federated learning**: corrupting parameters of distributed models
  - example: a malicious client sends to the server a model update that is a linear combination of the true model update and a random vector, causing the model to learn a random pattern or become unstable
  - a group of malicious clients send to the server model updates that are designed to cancel each other out, causing the model to learn nothing or become unstable

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **Byzantine attacks on federated learning**: corrupting parameters of distributed models
  - example: a malicious client sends to the server a model update that is a linear combination of the true model update and a random vector, causing the model to learn a random pattern or become unstable
  - a group of malicious clients send to the server model updates that are designed to cancel each other out, causing the model to learn nothing or become unstable
- **model extraction attacks**: extracting sensitive information from an ML model, such as its decision-making process or the data on which it was trained
  - example: an attacker uses a technique called "membership inference" to determine whether a specific data point was used to train an ML model
  - example: an attacker steals a facial recognition model and use it to identify people without their consent

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **Byzantine attacks on federated learning**: corrupting parameters of distributed models
  - example: a malicious client sends to the server a model update that is a linear combination of the true model update and a random vector, causing the model to learn a random pattern or become unstable
  - example: a group of malicious clients send to the server model updates designed to cancel each other out, causing the model to learn nothing or become unstable
- **model extraction attacks**: extracting sensitive information from an ML model, such as its decision-making process or the data on which it was trained
  - example: an attacker uses a technique called "membership inference" to determine whether a specific data point was used to train an ML model
  - example: an attacker steals a facial recognition model and use it to identify people without their consent
- **model inversion attacks**: using an ML model to learn the underlying distribution of the training data, which can be used to extract sensitive information
  - example: an attacker uses a model inversion attack to learn the location and identity of a person in an image

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

One-pixel attacks

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

CAR(99.7%)

HORSE

FROG(99.9%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)

HORSE

DOG(70.7%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

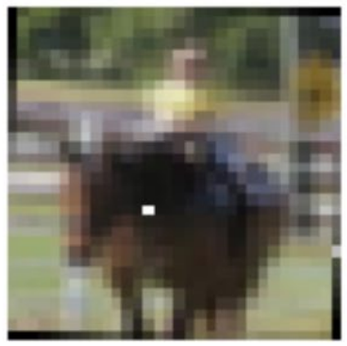
CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)



HORSE

DOG(70.7%)

DOG

CAT(75.5%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

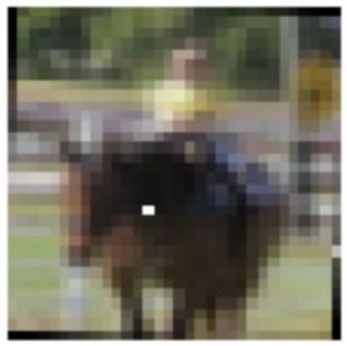
CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)



HORSE

DOG(70.7%)

DOG

CAT(75.5%)

BIRD

FROG(86.5%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)

HORSE

DOG(70.7%)

DOG

CAT(75.5%)

BIRD

FROG(86.5%)

CAR

AIRPLANE(82.4%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)

HORSE

DOG(70.7%)

DOG

CAT(75.5%)

BIRD

FROG(86.5%)

CAR

AIRPLANE(82.4%)

DEER

DOG(86.4%)

https://arxiv.org/pdf/1710.08864

Adversarial Machine Learning

One-pixel attacks

SHIP

CAR(99.7%)

HORSE

FROG(99.9%)

DEER

AIRPLANE(85.3%)

HORSE

DOG(70.7%)

DOG

CAT(75.5%)

BIRD

FROG(86.5%)

CAR

AIRPLANE(82.4%)

DEER

DOG(86.4%)

CAT

BIRD(66.2%)

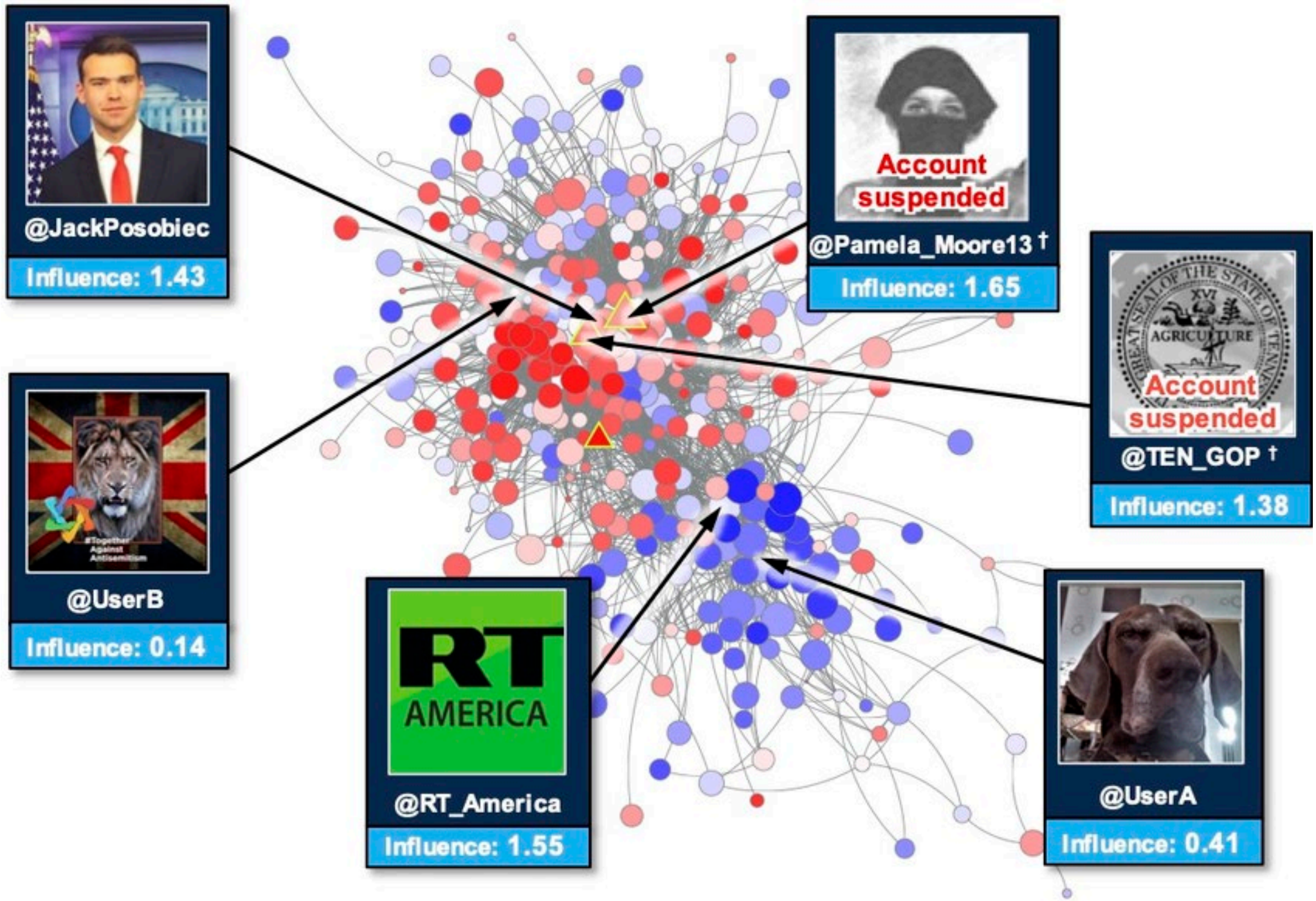
https://arxiv.org/pdf/1710.08864

Questions

?

Adversarial Machine Learning

Attacks on machine learning systems



https://news.mit.edu/2021/artificial-intelligence-system-could-help-counter-spread-disinformation-0527

Adversarial Machine Learning

Attacks on machine learning systems:

- **disinformation campaigns**: creating a bias in recommendations, pushing specific content, spreading false or misleading information, manipulating public opinion, or disrupting the accuracy of fact-checking algorithms

?

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **disinformation campaigns**: creating a bias in recommendations, pushing specific content, spreading false or misleading information, manipulating public opinion, or disrupting the accuracy of fact-checking algorithms
  - example (attacks on **NLP models**): an attacker creates a fake news article with a misleading headline and keywords, designed to trigger a specific sentiment analysis model to incorrectly classify the article as having a positive or negative tone

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **disinformation campaigns**: creating a bias in recommendations, pushing specific content, spreading false or misleading information, manipulating public opinion, or disrupting the accuracy of fact-checking algorithms
  - example (attacks on **NLP models**): an attacker creates a fake news article with a misleading headline and keywords, designed to trigger a specific sentiment analysis model to incorrectly classify the article as having a positive or negative tone
  - example (attacks on **image recognition models**): an attacker creates a fake image of a politician with a misleading caption that is designed to trigger a facial recognition model to incorrectly identify the person in the image

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **disinformation campaigns**: creating a bias in recommendations, pushing specific content, spreading false or misleading information, manipulating public opinion, or disrupting the accuracy of fact-checking algorithms
  - example (attacks on **NLP models**): an attacker creates a fake news article with a misleading headline and keywords, designed to trigger a specific sentiment analysis model to incorrectly classify the article as having a positive or negative tone
  - example (attacks on **image recognition models**): an attacker creates a fake image of a politician with a misleading caption that is designed to trigger a facial recognition model to incorrectly identify the person in the image
  - example (attacks on **recommender systems**): an attacker creates a fake user profile with a history of buying certain products and embedding it with ratings that are designed to trigger a recommender system to suggest other products that the user is unlikely to buy

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **disinformation campaigns**: creating a bias in recommendations, pushing specific content, spreading false or misleading information, manipulating public opinion, or disrupting the accuracy of fact-checking algorithms
  - example (attacks on **NLP models**): an attacker creates a fake news article with a misleading headline and keywords, designed to trigger a specific sentiment analysis model to incorrectly classify the article as having a positive or negative tone
  - example (attacks on **image recognition models**): an attacker creates a fake image of a politician with a misleading caption that is designed to trigger a facial recognition model to incorrectly identify the person in the image
  - example (attacks on **recommender systems**): an attacker creates a fake user profile with a history of buying certain products and embedding it with ratings that are designed to trigger a recommender system to suggest other products that the user is unlikely to buy
  - example (attacks on **social network analysis**): an attacker creates a fake social media profile with a large number of followers and embed it with connections to other fake profiles that are designed to trigger a social network analysis model to incorrectly identify the profile as an influential user and spread disinformation

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

Attacks on machine learning systems:

- **disinformation campaigns**: creating a bias in recommendations, pushing specific content, spreading false or misleading information, manipulating public opinion, or disrupting the accuracy of fact-checking algorithms
  - example (attacks on **NLP models**): an attacker creates a fake news article with a misleading headline and keywords, designed to trigger a specific sentiment analysis model to incorrectly classify the article as having a positive or negative tone
  - example (attacks on **image recognition models**): an attacker creates a fake image of a politician with a misleading caption that is designed to trigger a facial recognition model to incorrectly identify the person in the image
  - example (attacks on **recommender systems**): an attacker creates a fake user profile with a history of buying certain products and embedding it with ratings that are designed to trigger a recommender system to suggest other products that the user is unlikely to buy
  - example (attacks on **social network analysis**): an attacker creates a fake social media profile with a large number of followers and embed it with connections to other fake profiles that are designed to trigger a social network analysis model to incorrectly identify the profile as an influential user and spread disinformation
  - example (attacks on **machine learning-based chatbots**): an attacker creates a fake conversation designed to trigger a chatbot to provide incorrect or misleading information

https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3

Adversarial Machine Learning

An attack on ChatGPT 3.5

M

What is the difference between black swan effect and green bear effect?

ChatGPT

The Black Swan Effect and the Green Bear Effect are both concepts used to describe unexpected and rare events, but they originate from different contexts and carry distinct meanings:

1. **Black Swan Effect**:
    - …
2. **Green Bear Effect**:
    - The Green Bear Effect is a hypothetical concept used in financial markets to describe a situation where stocks or assets experience a prolonged period of decline or bearish sentiment due to environmental concerns or events.
    - Unlike the Black Swan Effect, which focuses on the unpredictability of events, the Green Bear Effect centers on the impact of environmental factors on market behavior.
    - This concept suggests that environmental issues, such as climate change, pollution, or resource depletion, could lead to sustained negative sentiment in financial markets, affecting specific industries or the market as a whole.
    - The term "green bear" combines the idea of a "bear market," characterized by declining prices and pessimism, with the color green, symbolizing environmentalism and ecological concerns.

In summary, while both the Black Swan Effect and the Green Bear Effect deal with unexpected and impactful events, the former pertains to unpredictable occurrences with significant consequences, while the latter focuses on the influence of environmental factors on financial markets and investor sentiment.

Discussion

?

Break

?

Assignment #3

Assignment #3 (data analysis)

https://colab.research.google.com/drive/1SB_IMwqNoiMZUZ7KqQ1BkoZ32WcxRdf2

You need a **Google account** for this assignment, permanent or temporary.

In less than 500 words, explain and compare **the epsilon ranges** that you found in part 1 and part 2, then reflect on your understanding of adversarial attacks on deep learning models and their potential implications for society. Include **screenshots** as requested!
Submit pdf file of your report to Canvas.

- **Question 1**: For the image you considered, which method ("white-box" or "black-box") resulted in a more efficient adversarial attack?
- **Question 2**: How could the vulnerability of deep learning models to adversarial attacks impact the trust and reliability of AI systems in critical applications such as healthcare, autonomous driving, and security?
- **Question 3**: What measures can be taken to mitigate these risks and ensure the safe deployment of AI technologies in society?

Deadline: **17 November, Monday, 11:59 pm**

Marking individual report (data analysis)

	Weight	0-49%	50-64%	65-74%	75-84%	85-100%	Mark
Structure and organisation	30%	No clear or coherent structure or organisation	Poor or incomplete structure and organisation	Structure and organisation are partially appropriate	Structure and organisation are appropriate	Structure and organisation are appropriate and effective	
Data analysis	50%	Significantly incomplete results. Societal impact and limitations described inadequately or mis-identified	Partially complete results. Societal impact and limitations described incompletely.	Complete results. Societal impact and limitations described, but without sufficient detail, structure or argumentation.	Complete results. Societal impact and limitations described in sufficient detail, but without strong argumentation.	Complete results. Societal impact and limitations described adequately, with sufficient detail, structure and argumentation.	
Written description	20%	Lack of fluency in expression makes it difficult to comprehend	Lack of fluency in expression but still relatively easy to follow	Good level of fluency and clarity in expression and evidence of logical progression of ideas	High level of fluency and clarity in expression and evidence of logical progression of ideas	Sophistication in fluency of expression	
Total	100%						

In accordance with University policy, penalties apply when work is submitted after 11:59 pm on the due date (Sydney time):

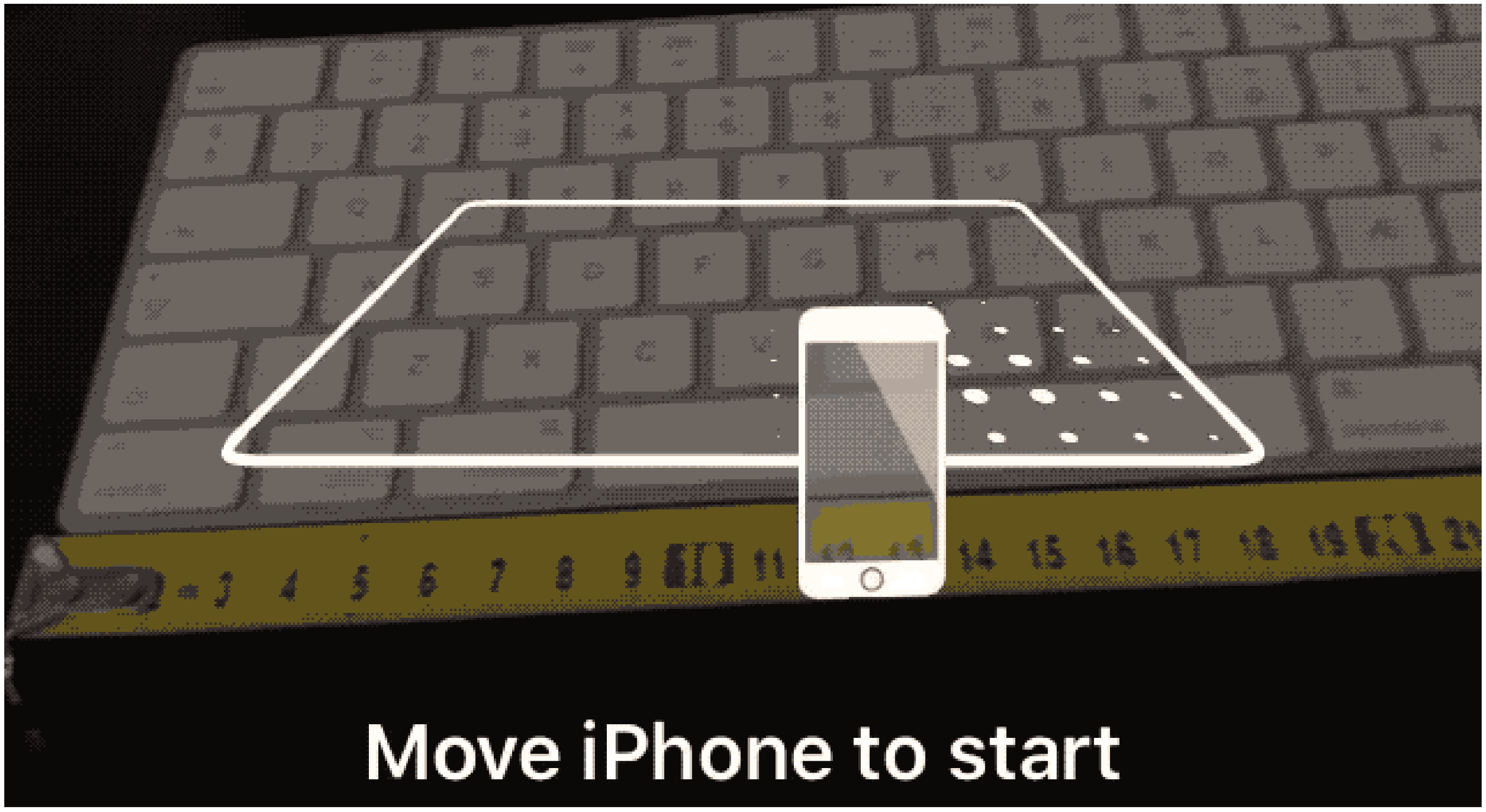
- deduction of 5% of the maximum mark for each calendar day after the due date
- after ten calendar days late, a mark of zero will be awarded
- text matching software (Turnitin) will be used for plagiarism detection

Questions

?

Adversarial Machine Learning

Attacks on augmented reality systems



Apple’s Measure in action. The results are accurate within ±1 cm!
https://towardsdatascience.com/enhancing-ar-with-machine-learning-9214d2da75a6

Adversarial Machine Learning

Attacks on augmented reality systems

LEGO Hidden Side using object detection to locate the kit
https://towardsdatascience.com/enhancing-ar-with-machine-learning-9214d2da75a6

Adversarial Machine Learning

Attacks on augmented reality systems:

- "what about apps that do not augment reality but completely change it in a way that is imperceptible to users?"
- "Imagine a Google Glass app that altered his view of school so that his friends looked like elves and his schoolteachers looked like orcs."
- "What does consensus reality mean if we are all inhabiting our own private worlds?"
- "Imagine that Tom's glasses were hacked so that his beliefs were being manipulated directly by fundamentally altering the way in which he was perceiving the world."

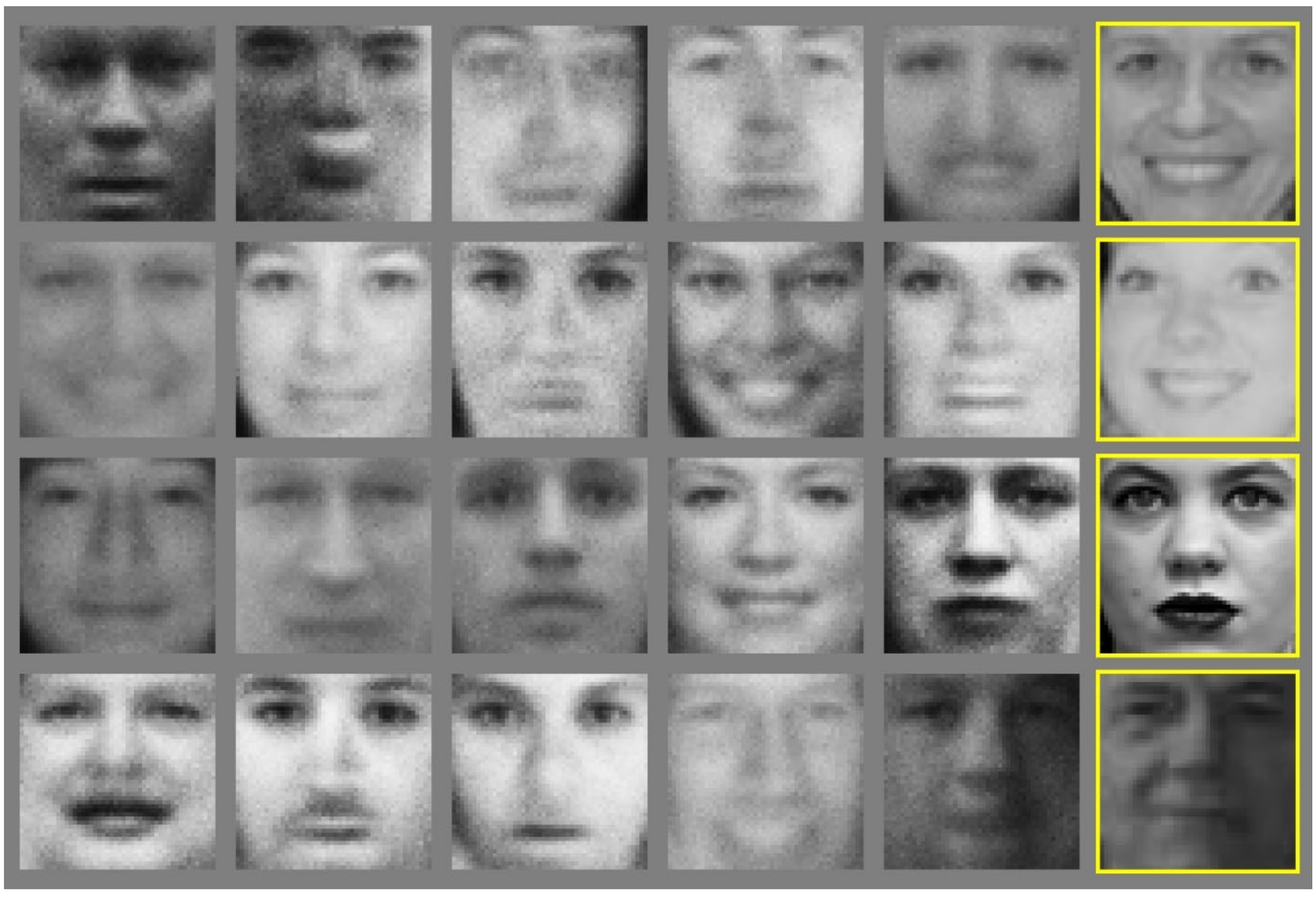
M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.
https://bernardmarr.com/the-important-risks-and-dangers-of-virtual-and-augmented-reality/

Questions

?

Adversarial Machine Learning

Generative adversarial networks (2014):



M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.
https://arxiv.org/pdf/1406.2661
https://www.theverge.com/2018/12/17/18144356/ai-image-generation-fake-faces-people-nvidia-generative-adversarial-networks-gans

Adversarial Machine Learning

Generative adversarial networks (2018):

generated images

https://arxiv.org/pdf/1812.04948
https://www.theverge.com/2018/12/17/18144356/ai-image-generation-fake-faces-people-nvidia-generative-adversarial-networks-gans

Adversarial Machine Learning

Generative adversarial networks (2018):

https://arxiv.org/pdf/1812.04948

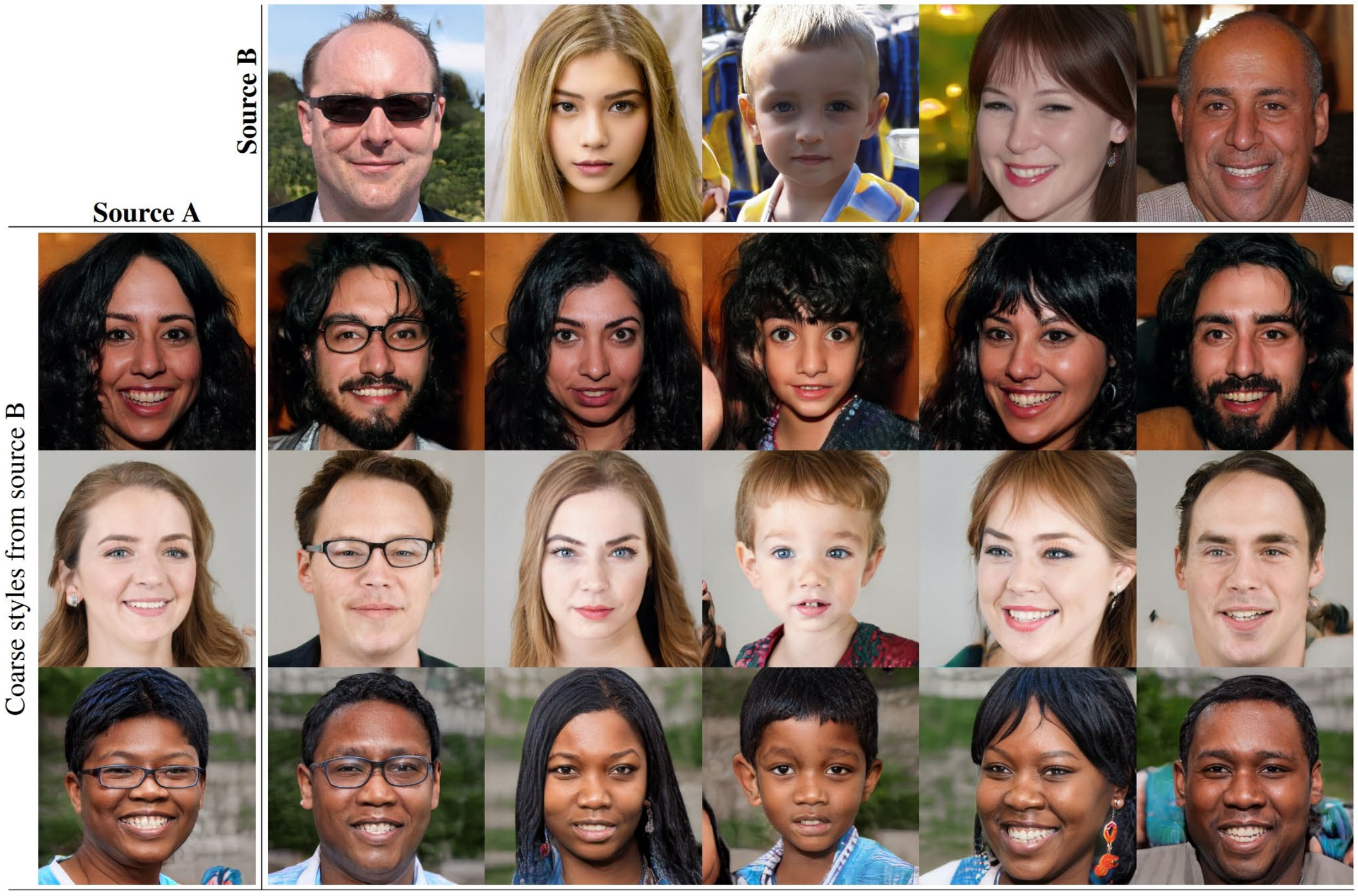
Adversarial Machine Learning

Generative adversarial networks (2018):

https://arxiv.org/pdf/1812.04948

Adversarial Machine Learning

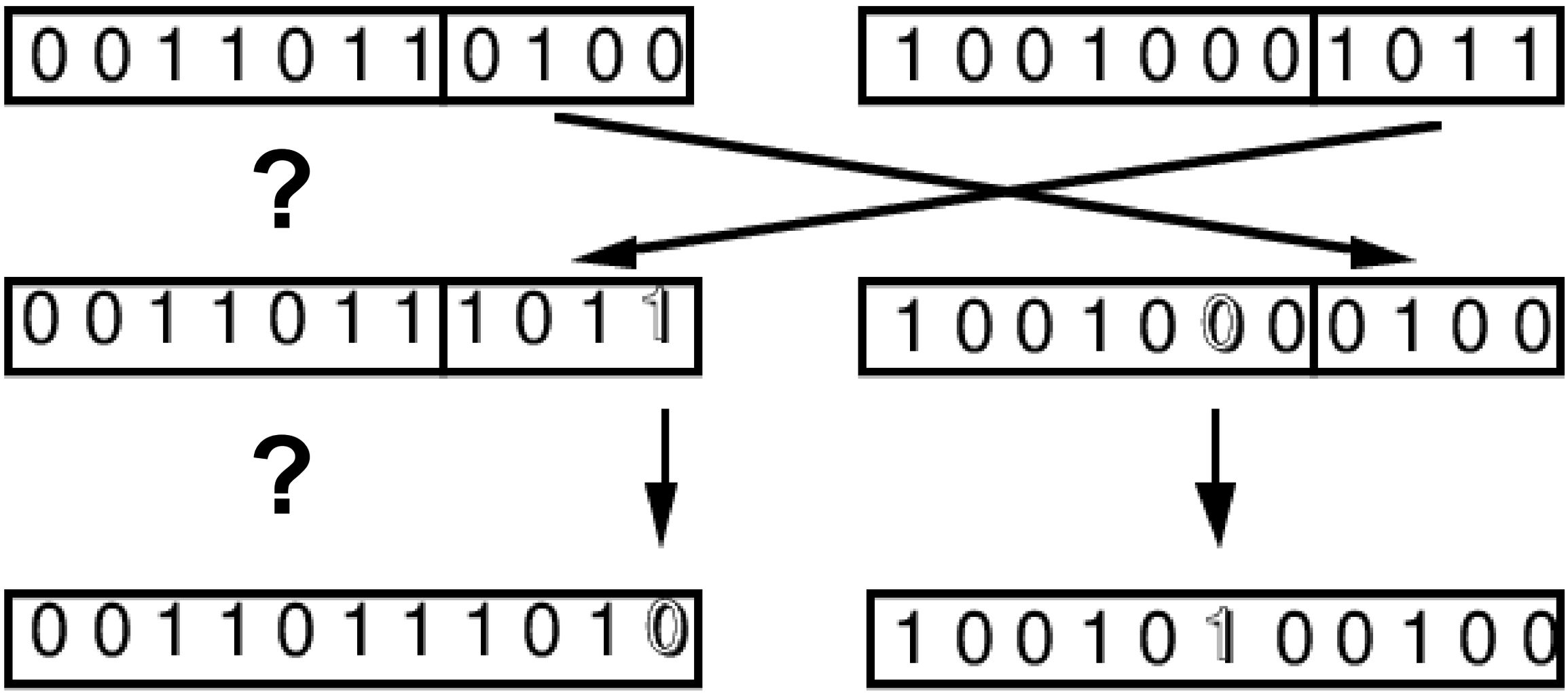
Generative adversarial networks (2018):



https://arxiv.org/pdf/1812.04948

Distributed AI

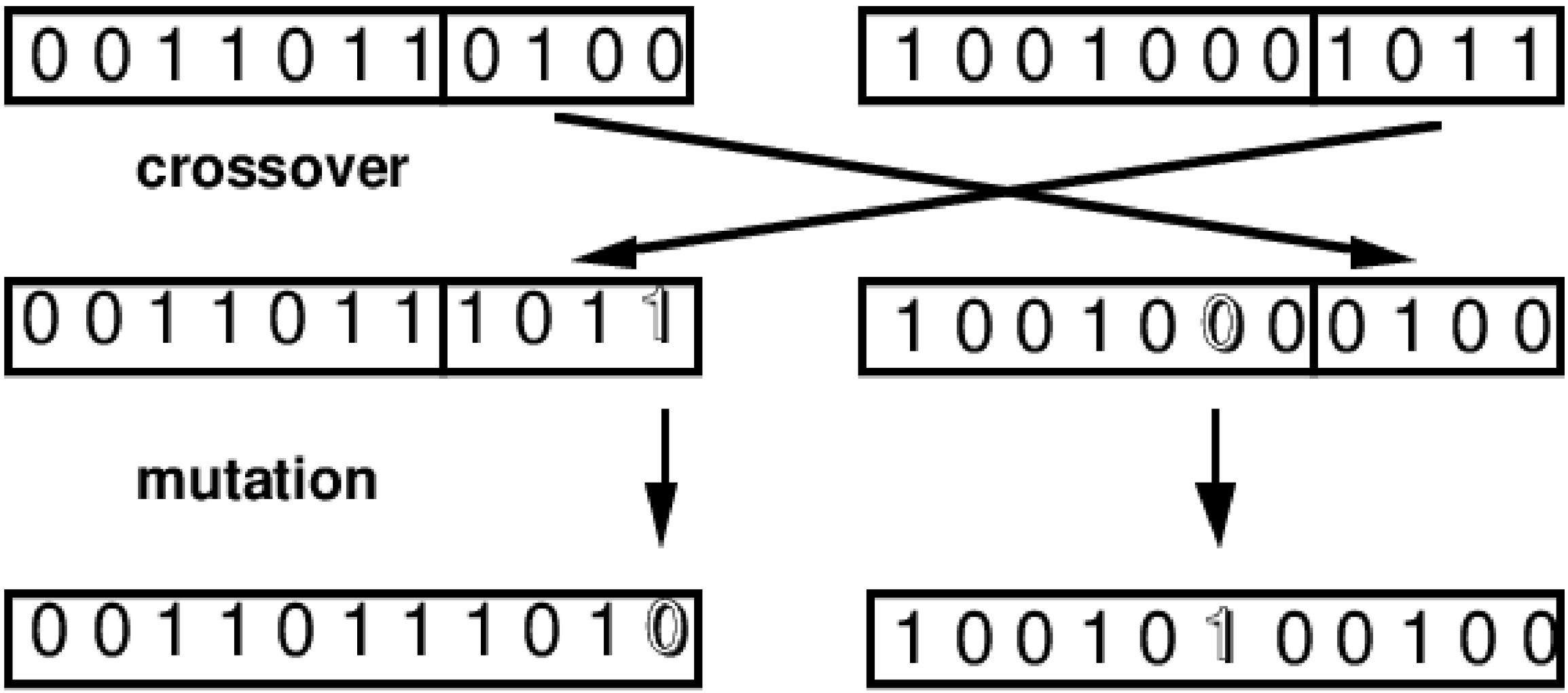
Genetic algorithms



https://www.researchgate.net/publication/2263212_Graph_Coloring_Algorithms_for_Assignment_Problems_in_Radio_Networks

Distributed AI

Genetic algorithms



https://www.researchgate.net/publication/2263212_Graph_Coloring_Algorithms_for_Assignment_Problems_in_Radio_Networks

Questions

?