COMP9208
Artificial Intelligence and Society

Lecture 13: 7 November 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

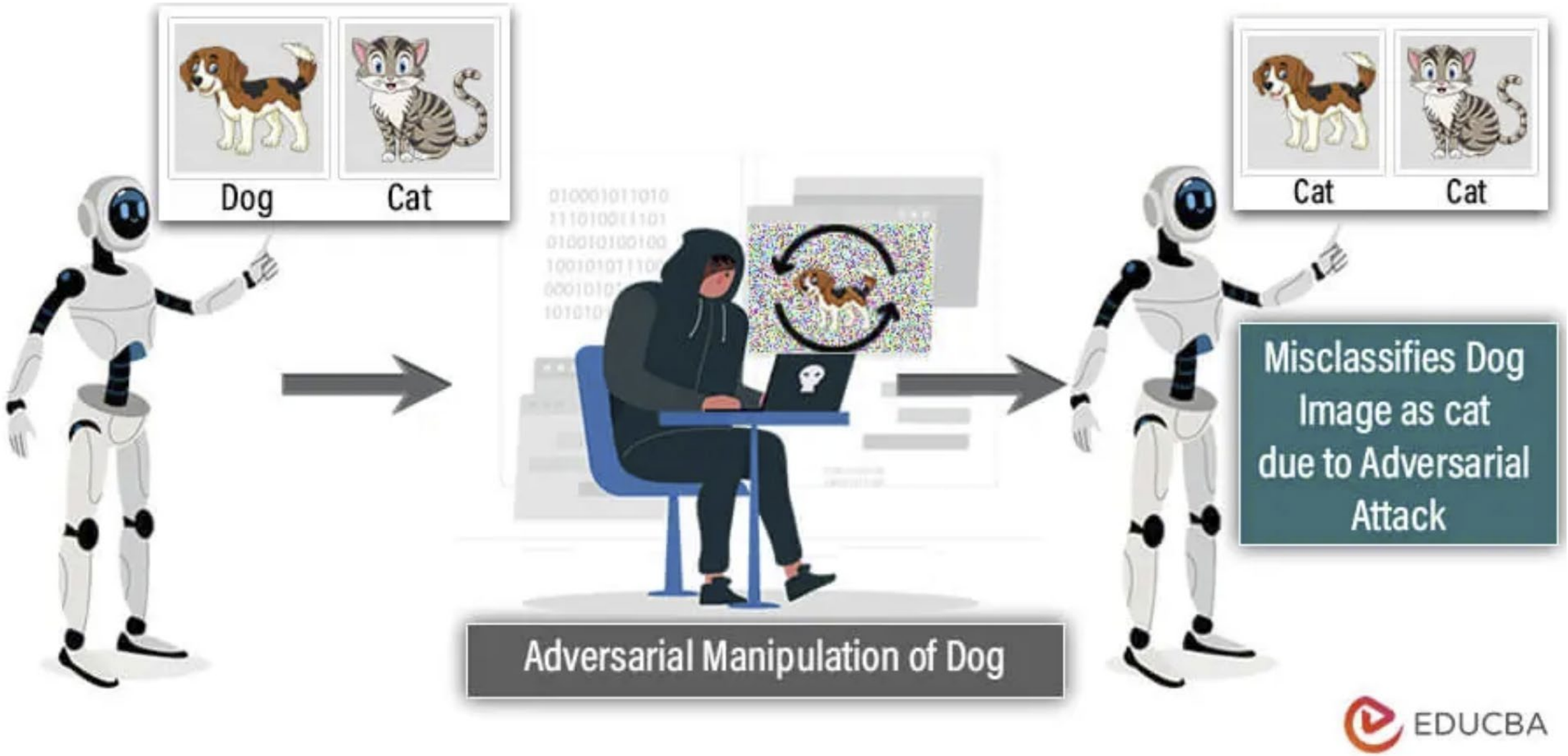
Date	Week: Lecture	Assignment
08/08	1: History of AI	
15/08	2: History of AI + Perception and Actuation	
22/08	3: History of AI + Representation and Reasoning	
29/08	4: Machine Learning	
05/09	5: Natural Interaction	
12/09	6: Distributed AI (swarm intelligence)	
19/09	7: From Good Old-Fashioned AI (GOFAI) 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)

Brief review of last week

?

Adversarial Machine Learning

Attacks on machine learning systems

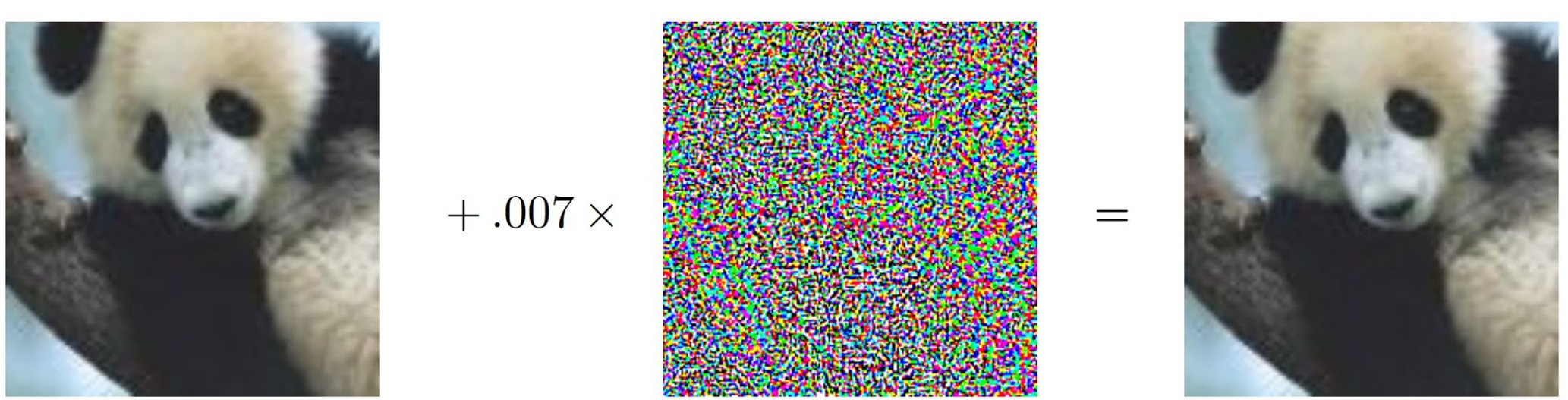


https://www.educba.com/adversarial-machine-learning/

Adversarial Machine Learning

Attacks on machine learning systems:

- ❖ **?** 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

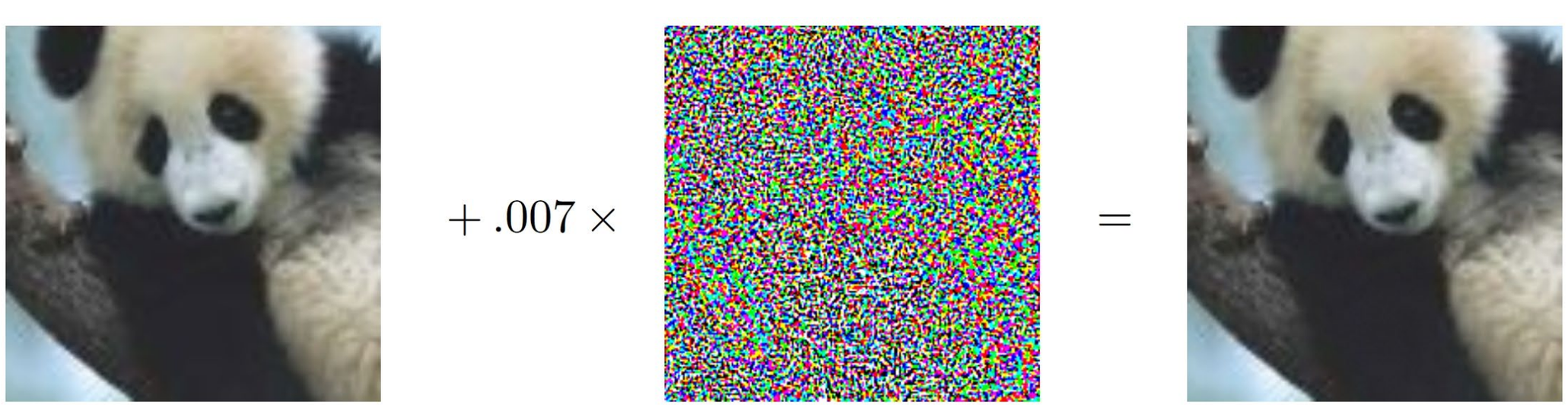


https://en.wikipedia.org/wiki/Adversarial_machine_learning
Meta Llama 3
https://arxiv.org/pdf/1412.6572

Adversarial Machine Learning

Attacks on machine learning systems:

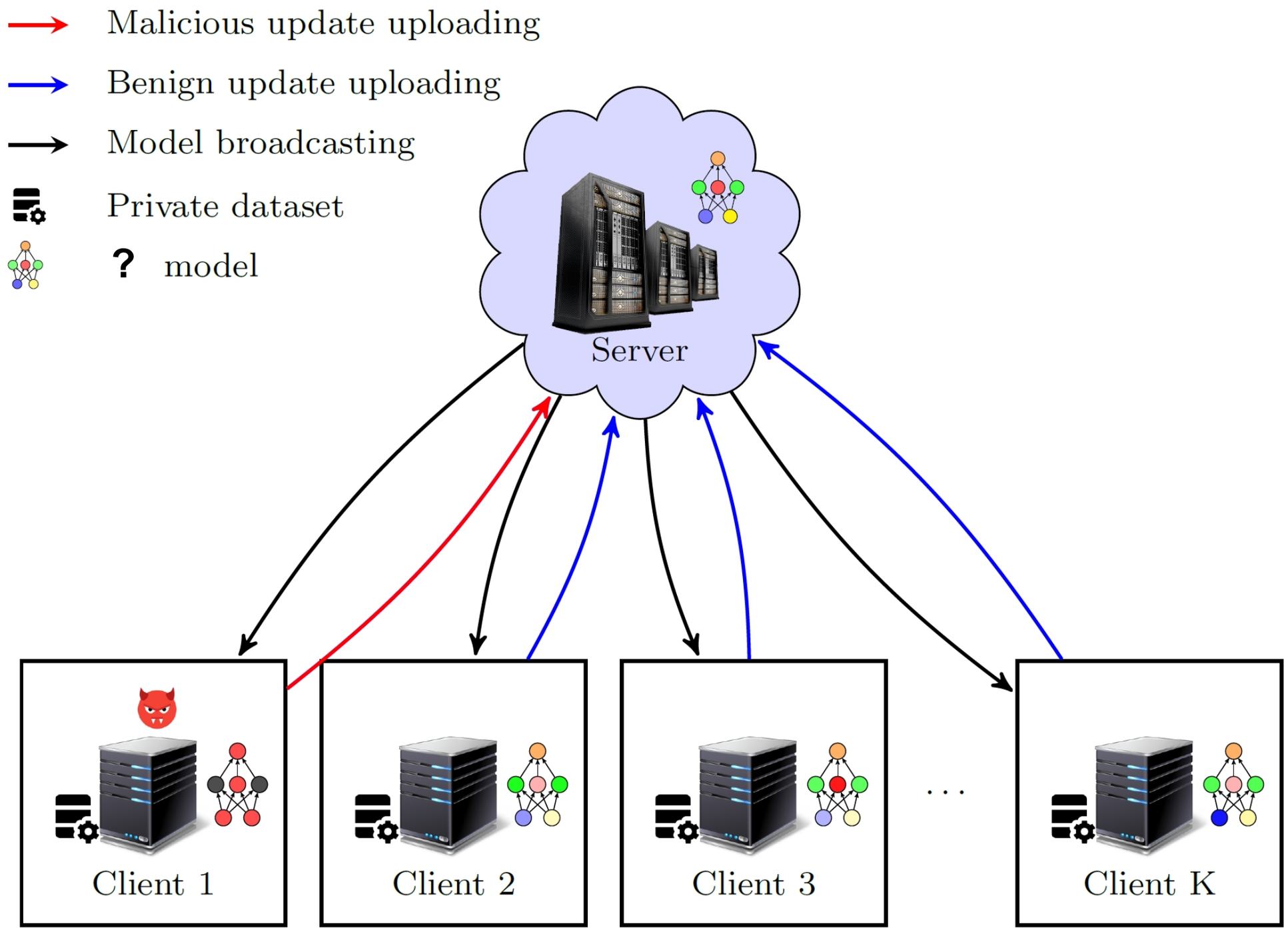
- **data poisoning**: contaminating the training dataset with data designed to increase errors in the output
  - example: an attacker might add malicious data to a dataset used to train a predictive model, causing it to make incorrect predictions
  - example: an attacker might add fake data to a dataset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



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

Adversarial Machine Learning

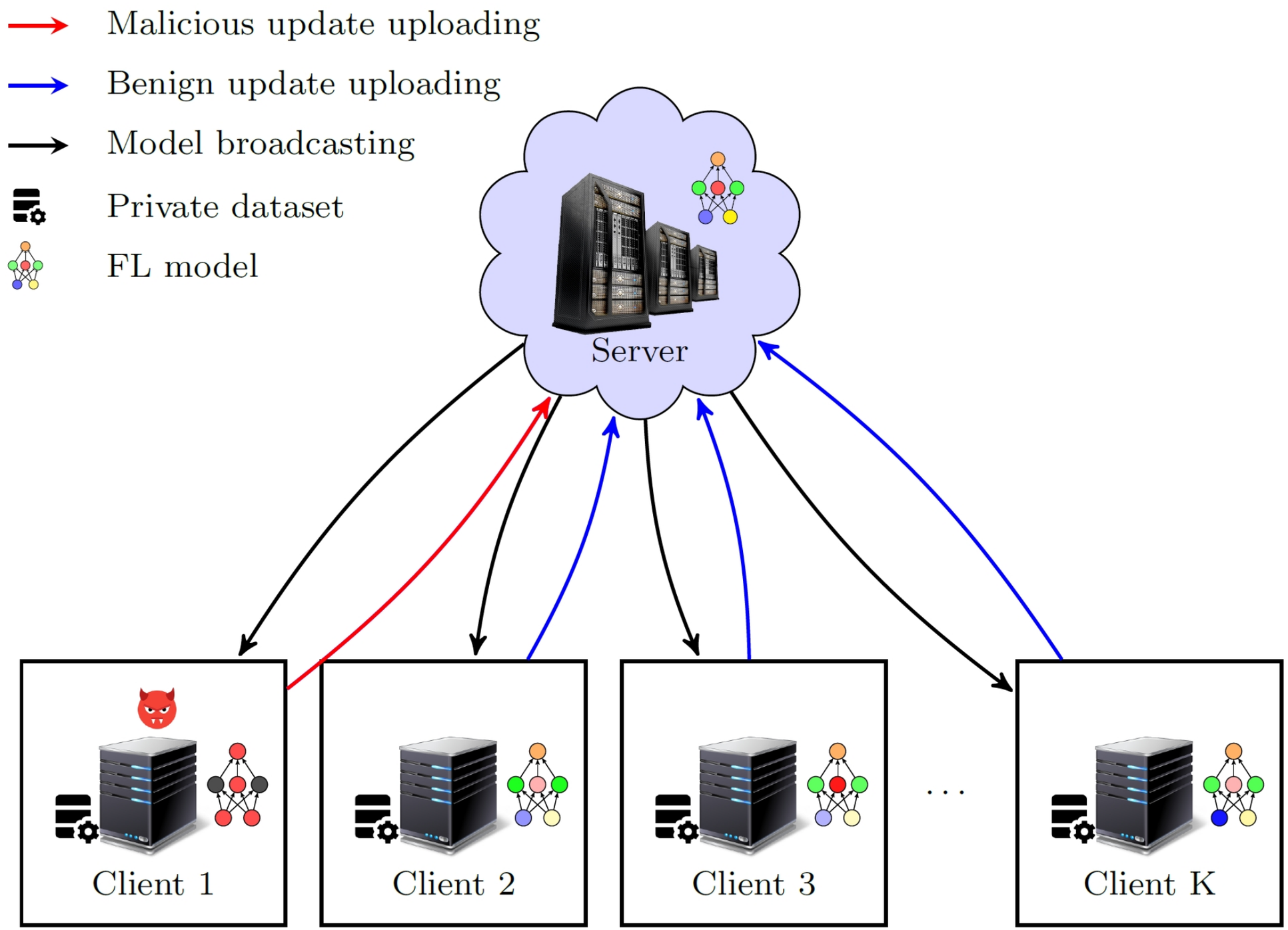
Attacks on machine learning systems



https://arxiv.org/pdf/2302.07173

Adversarial Machine Learning

Attacks on machine learning systems



https://arxiv.org/pdf/2302.07173

Adversarial Machine Learning

? 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

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

Adversarial Machine Learning

Attacks on machine learning systems:

- ❖ **?** 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

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

Attacks on **?** systems:

- "what about apps that do not **?** 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.

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/

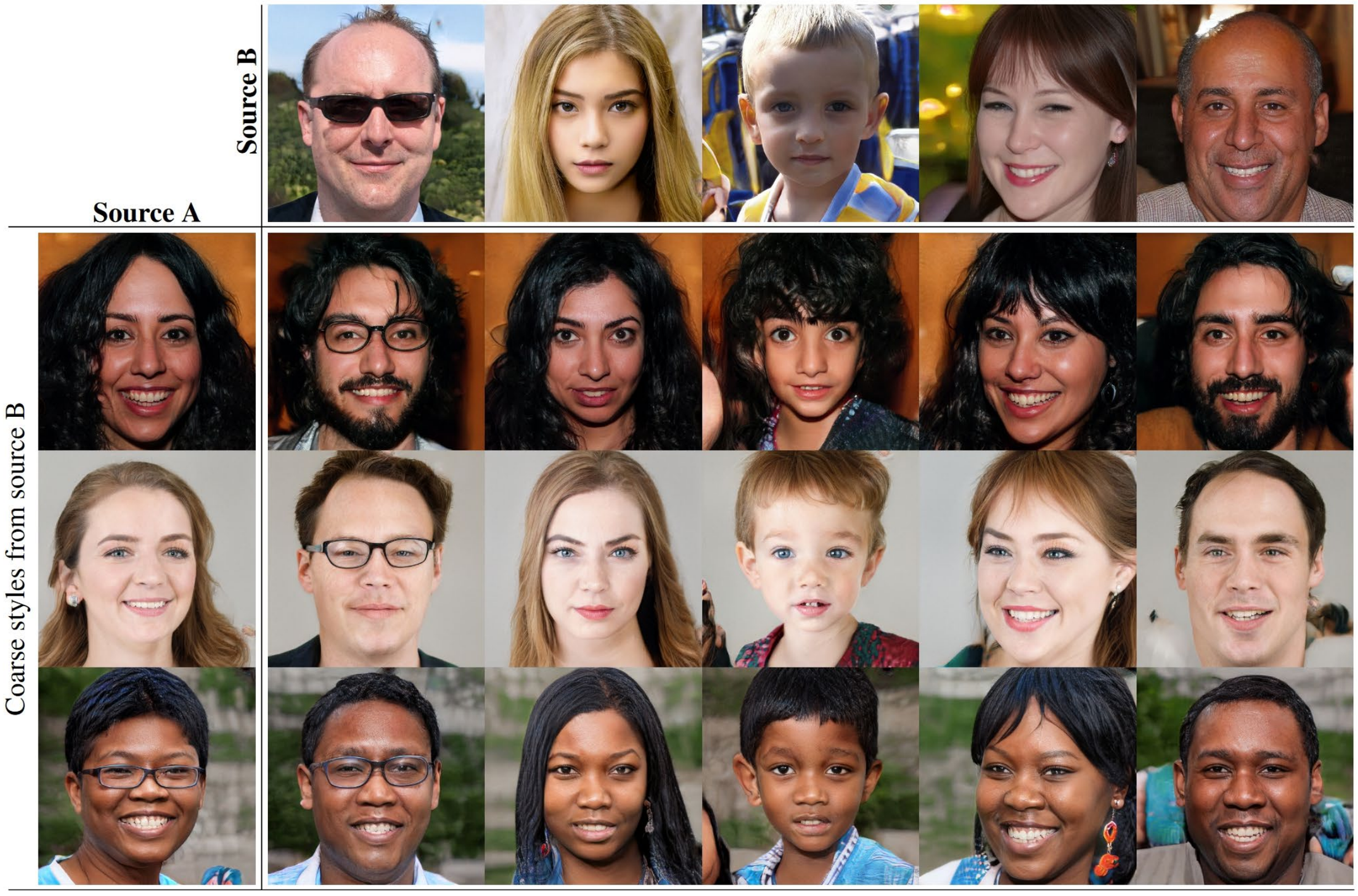
Adversarial Machine Learning

? adversarial networks (2018):

https://arxiv.org/pdf/1812.04948

Adversarial Machine Learning

Generative adversarial networks (2018):



https://arxiv.org/pdf/1812.04948

Questions

?

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%						

Questions

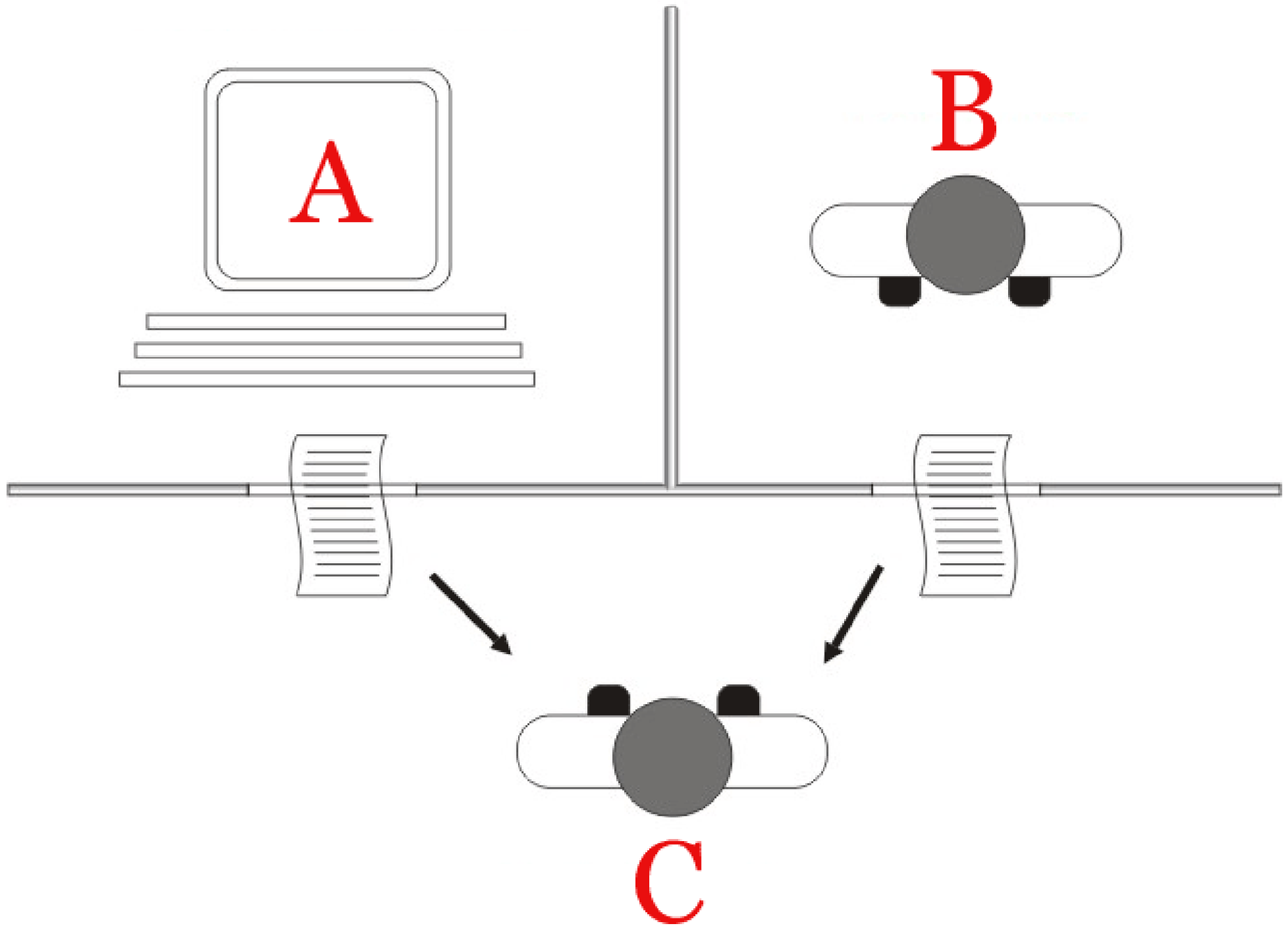
?

Brief review of the coursework

?

Turing test

If a machine could carry on a conversation (over a teleprinter) that was indistinguishable from a conversation with a human being, then it was reasonable to say that the machine was "thinking".

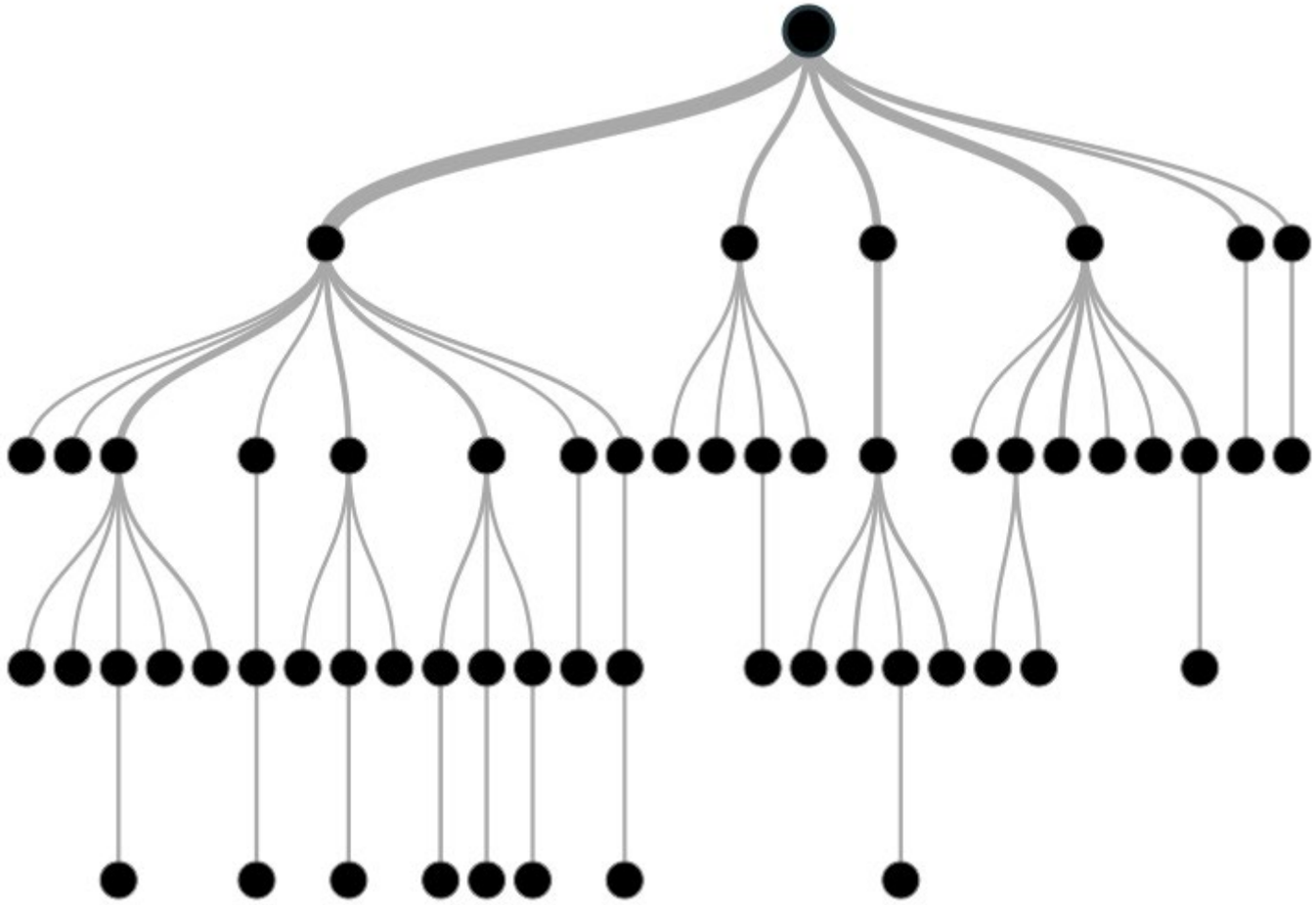


https://en.wikipedia.org/wiki/Turing_test

History of AI: main achievements and setbacks

Early successes (1956 – 1974)

- “Reasoning as search”:
  - the problem-solving process is framed as a search through a space of potential solutions, where each potential solution is represented as a state in the “search-space”,
  - the goal is to find a path to a solution by iteratively exploring and evaluating different states in the search space



https://www.justinmath.com/the-first-wave-of-ai-reasoning-as-search/

History of AI: main achievements and setbacks

Boom (1980 – 1987)

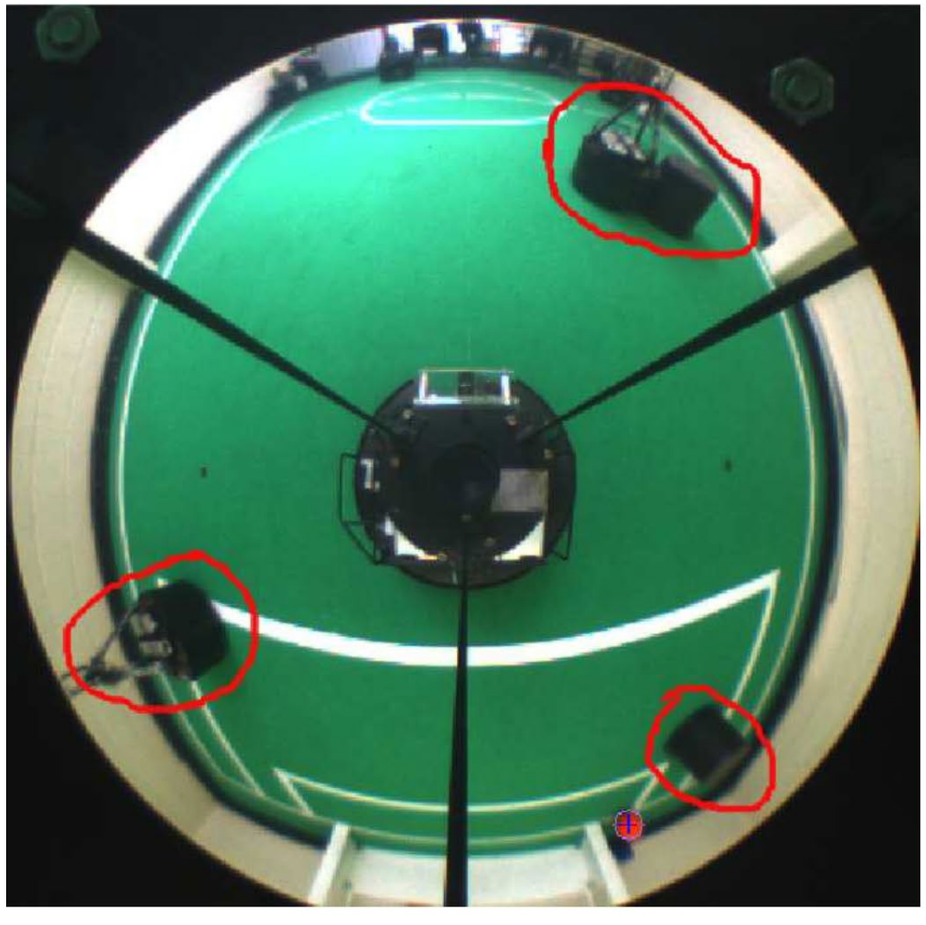
- Expert systems limitations:
  - ❖ domain-specific
  - ❖ knowledge acquisition problem: require knowledge engineers to distil and input the data
  - ❖ no real understanding of concepts and their relationships, no commonsense
  - ❖ simple task can potentially become computationally expensive
  - ❖ hard to scale up, for example, increase the number of rules
  - ❖ hard to modify and maintain consistency of large knowledge bases
  - ❖ do not learn: “could not learn concepts that most children know by the time they are 3 years old” (Marvin Minsky)

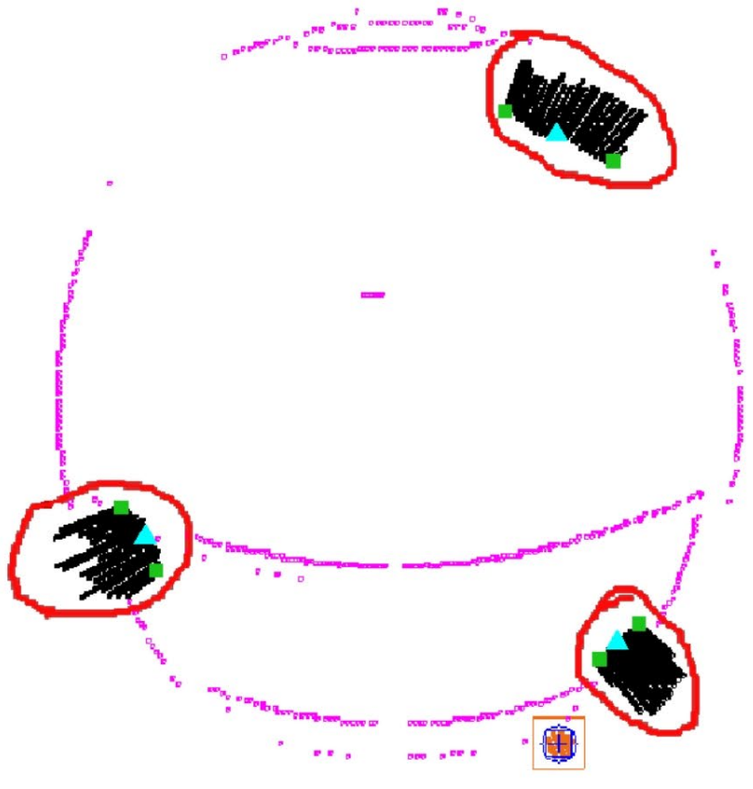
https://en.wikipedia.org/wiki/Expert_system
https://www.javatpoint.com/advantages-and-disadvantages-of-expert-system

Perception and Actuation

Computer vision:

- image processing: filtering, edge detection, and image segmentation
- object recognition: identifying and classifying objects within images
- face recognition: detecting and identifying human faces using features like facial landmarks
- scene understanding: combining object recognition and spatial context to understand the overall scene, for example, semantic segmentation



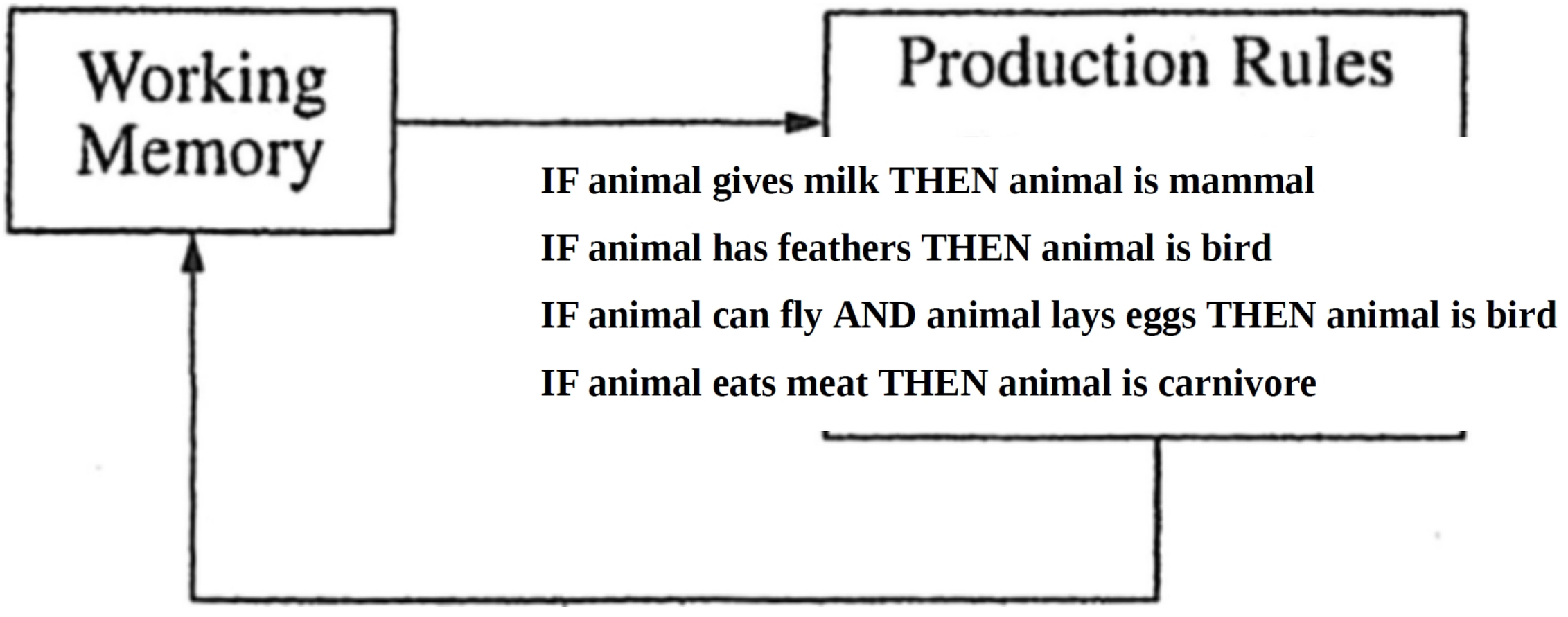


https://link.springer.com/chapter/10.1007/978-3-642-04686-5_29

Knowledge Representation and Reasoning (KR&R)

Production rules:

- knowledge is represented as a set of condition-action pairs (rules)
- when a condition is satisfied, the corresponding action is executed
- working memory contains currently produced results which can trigger new rules
- example: IF it is raining THEN carry an umbrella



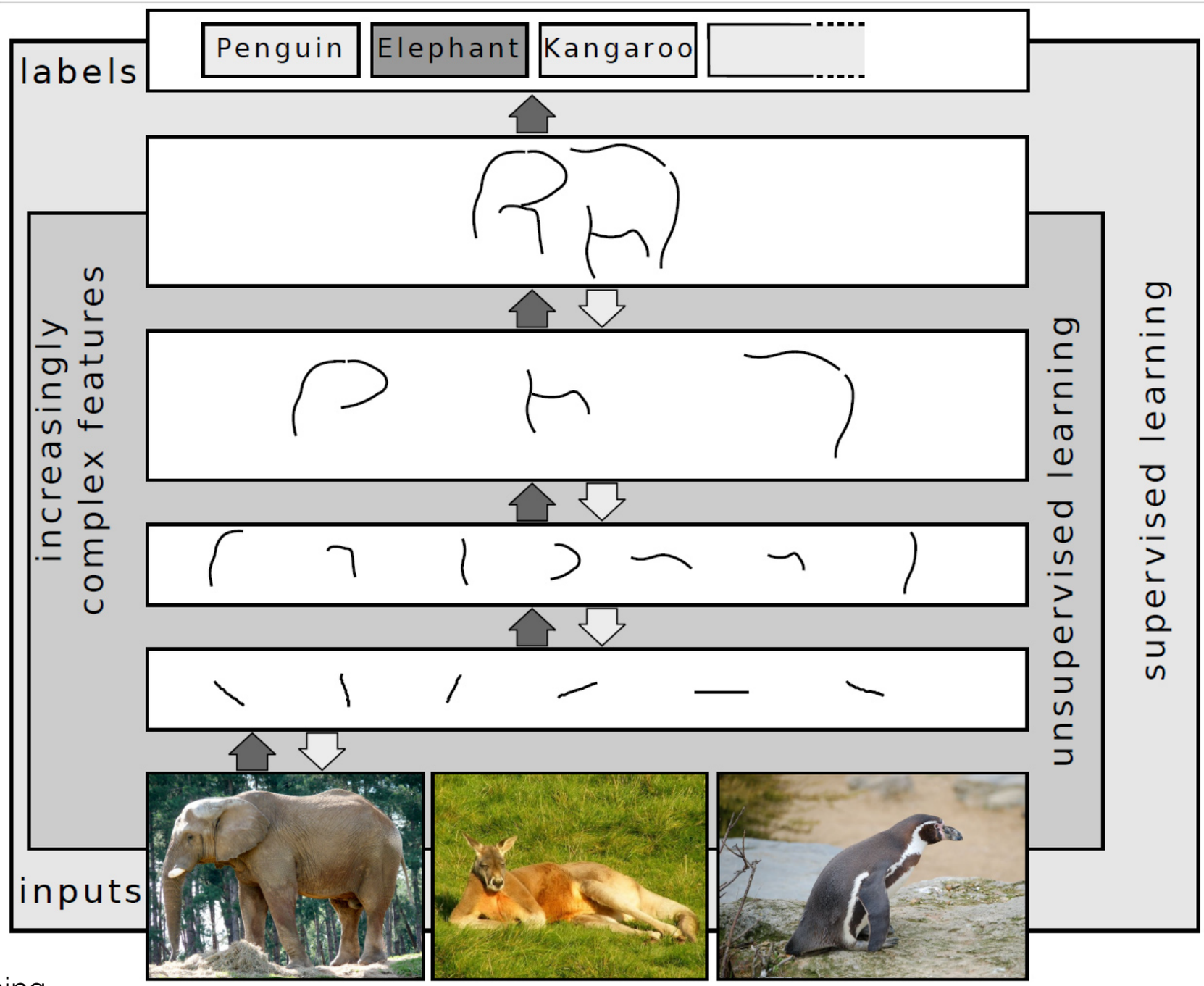
https://link.springer.com/chapter/10.1007/978-3-030-71873-2_1

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

History of AI: main achievements and setbacks

Deep Learning: neural networks + representation learning

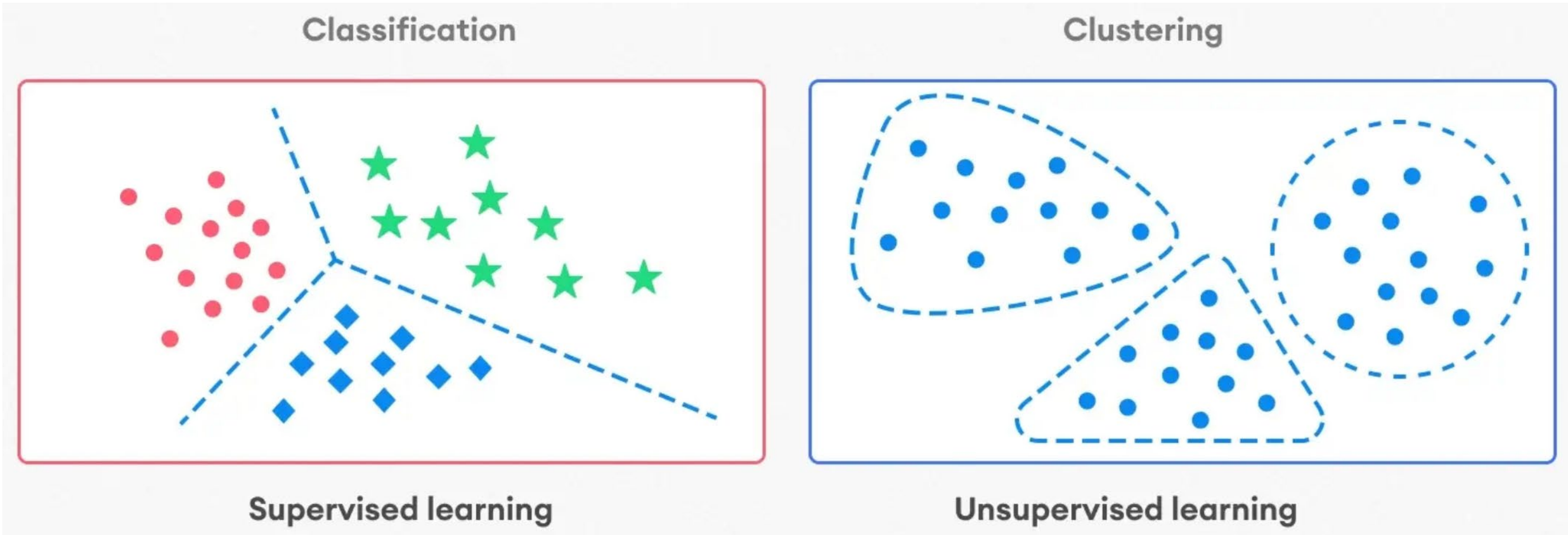
- a hierarchy of layers transforming input data into composite abstractions



https://en.wikipedia.org/wiki/Deep_learning

Machine Learning

Supervised vs unsupervised learning

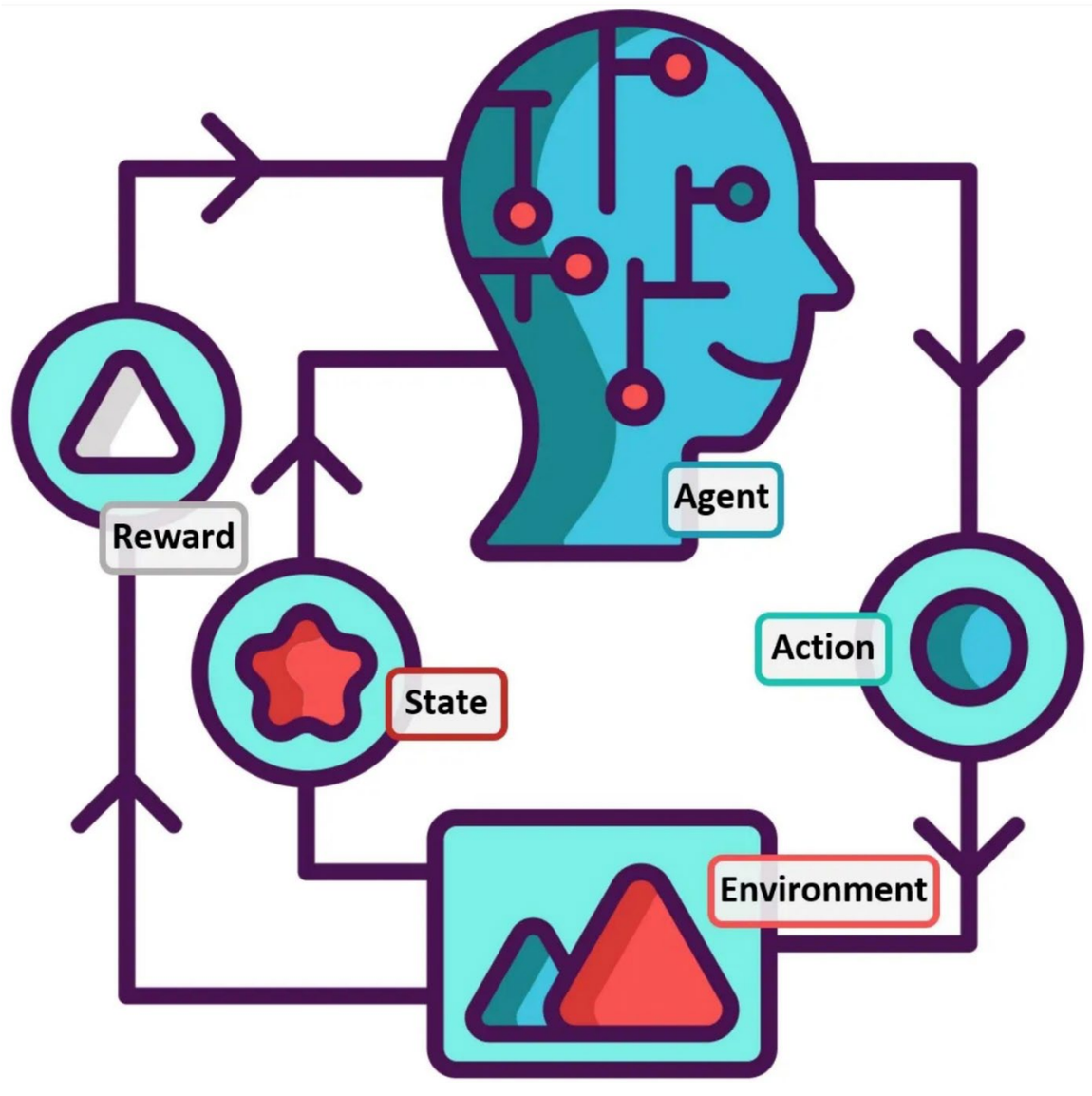


https://www.superannotate.com/blog/image-classification-basics

Machine Learning

Reinforcement learning:

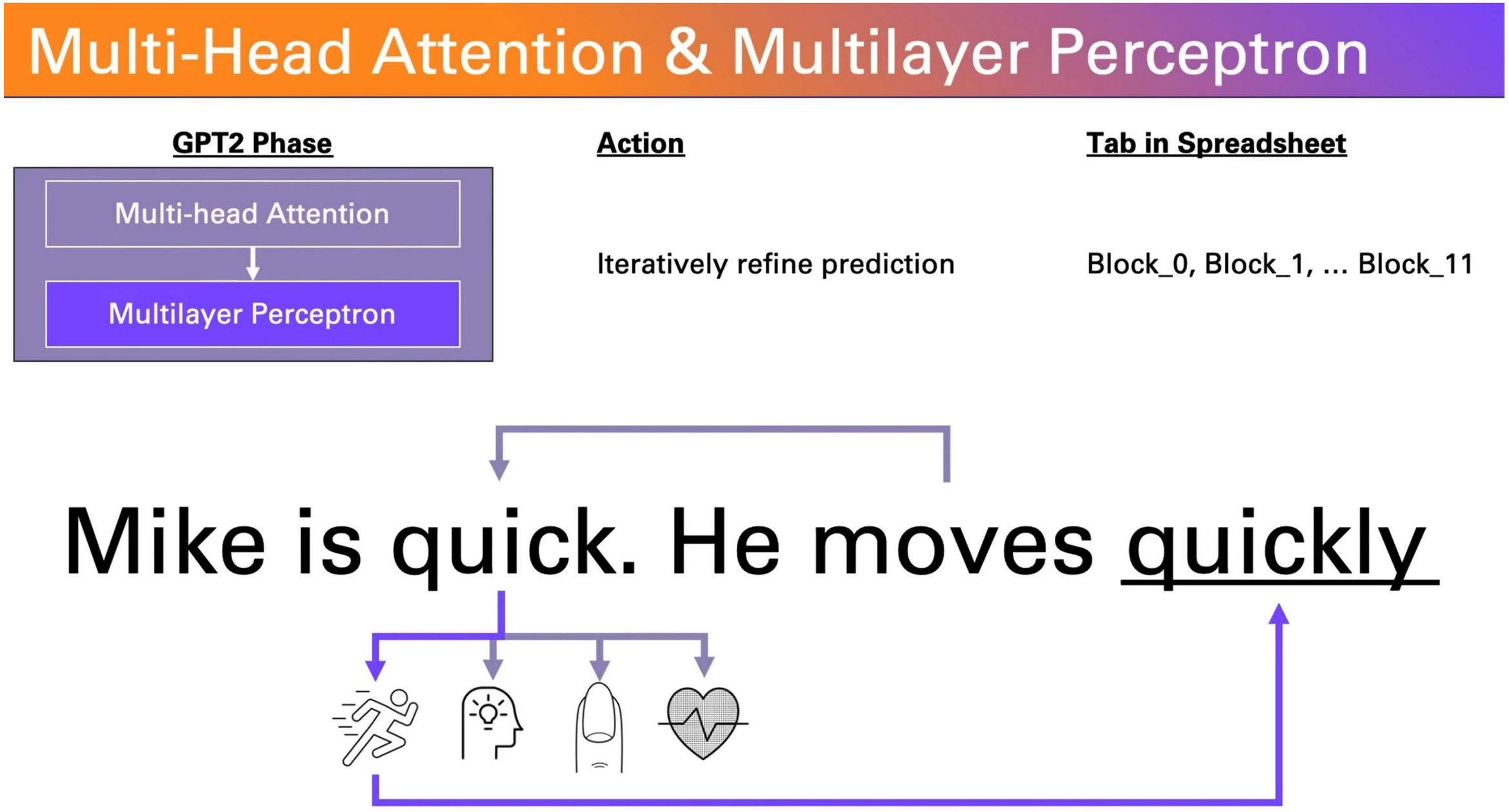
- agent learns to make decisions by doing actions in an environment to maximise cumulative reward
- the agent learns from the effects of its actions rather than from explicit instruction



https://towardsdatascience.com/reinforcement-learning-fda8ff535bb6

Natural Interaction

Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language



https://spreadsheets-are-all-you-need.ai/index.html#watch-the-lessons

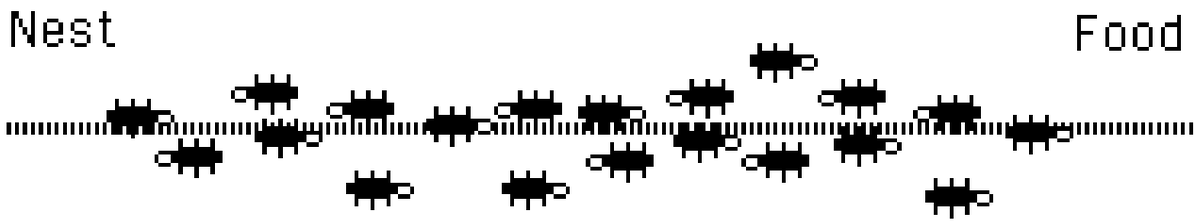
Natural Interaction

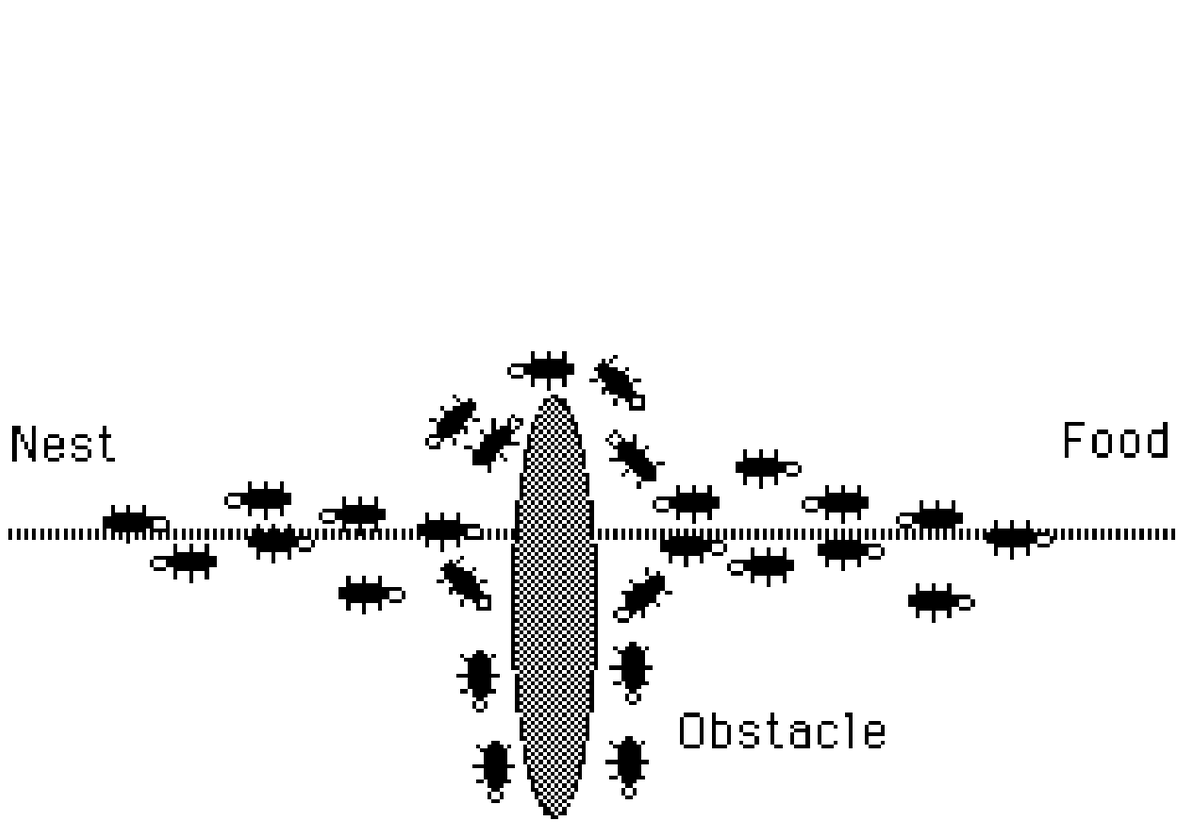
Human-Robot Interaction (HRI): solutions designed to enable people to safely interact with robotic systems

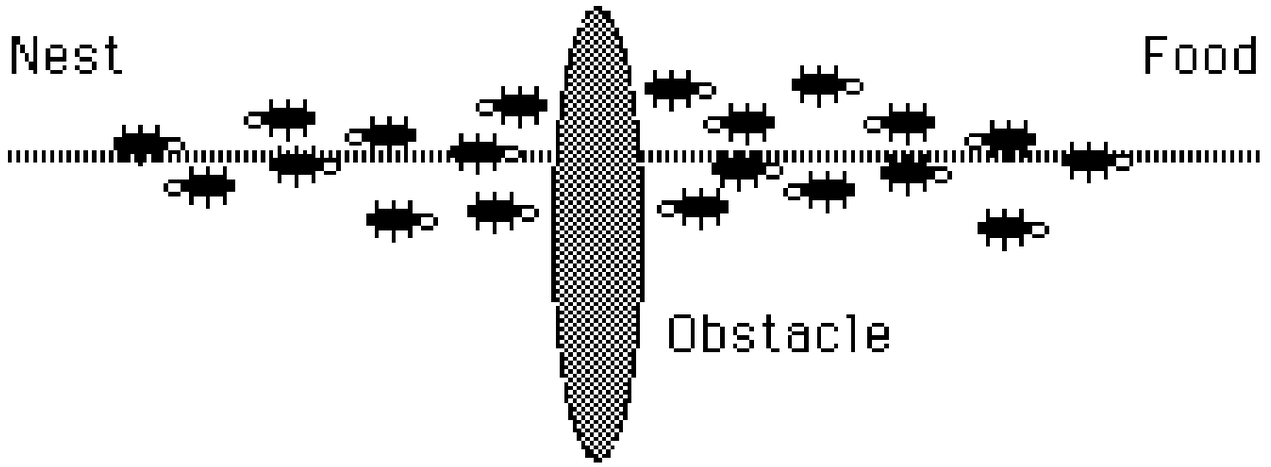
https://www.youtube.com/watch?v=mtH14_8Ez1E

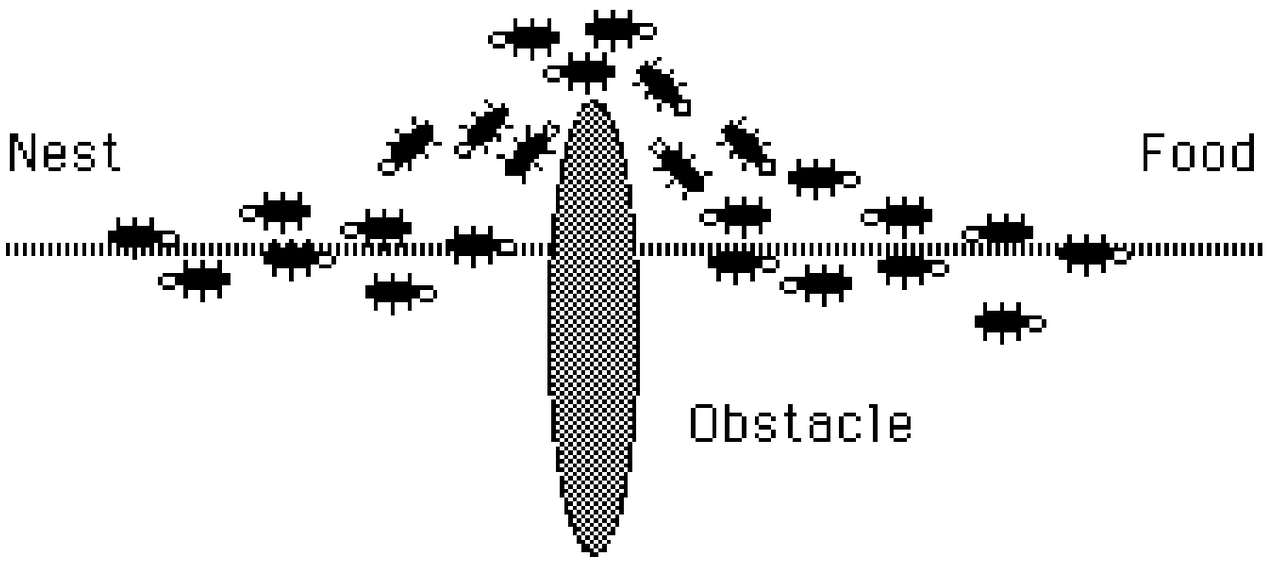
Distributed AI

Swarm intelligence: Ant Colony Optimisation



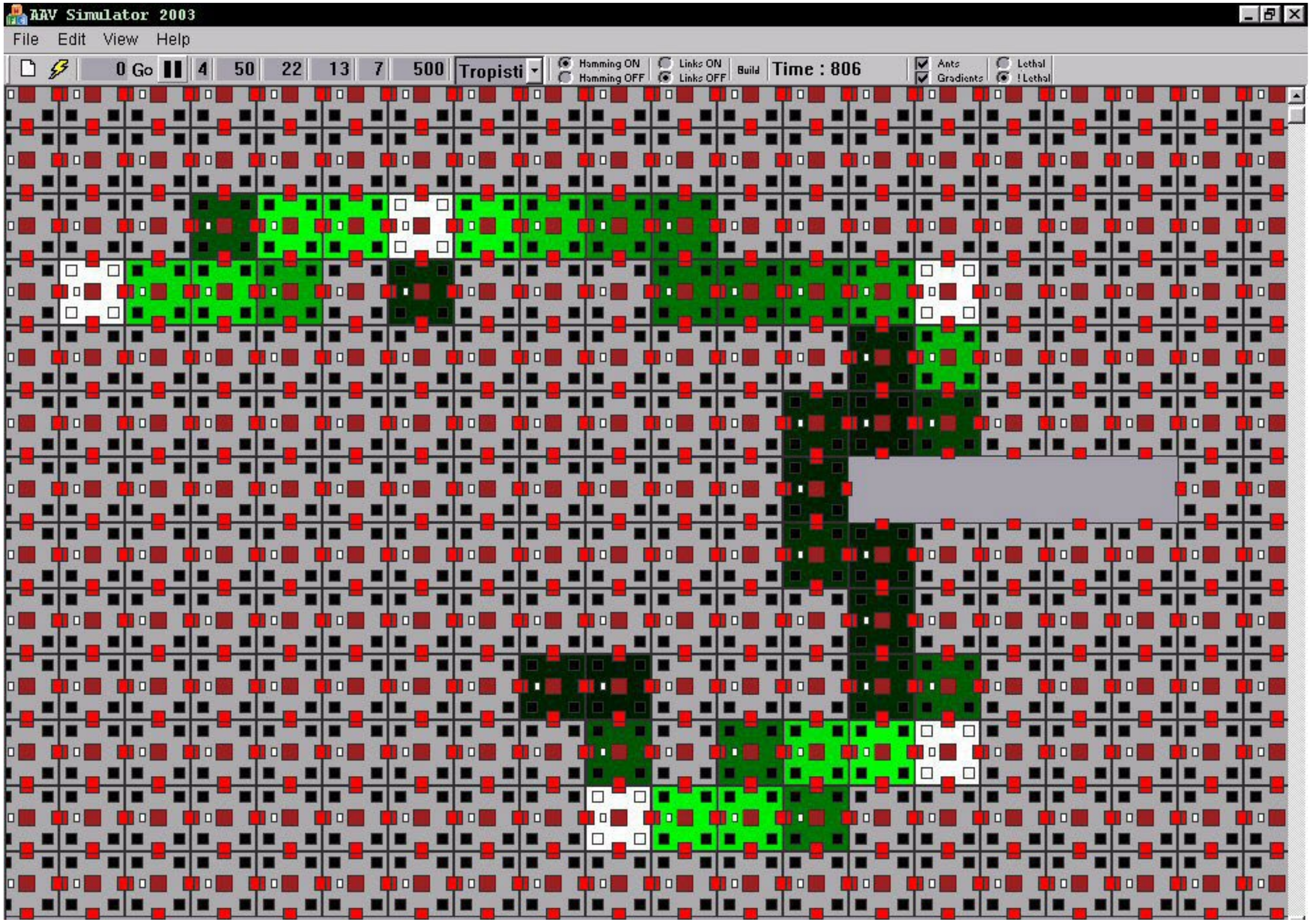




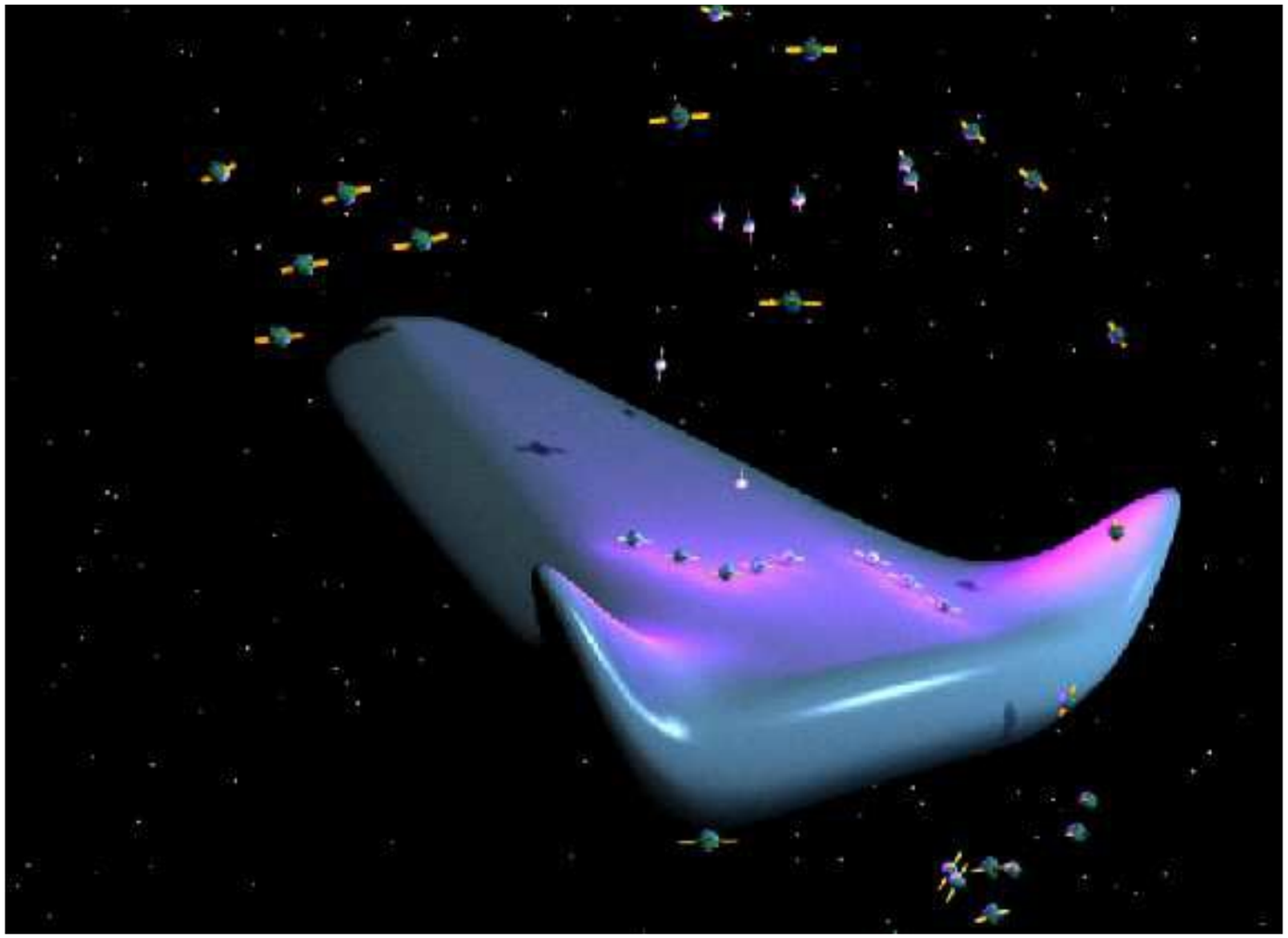


https://slideplayer.com/slide/12455505/

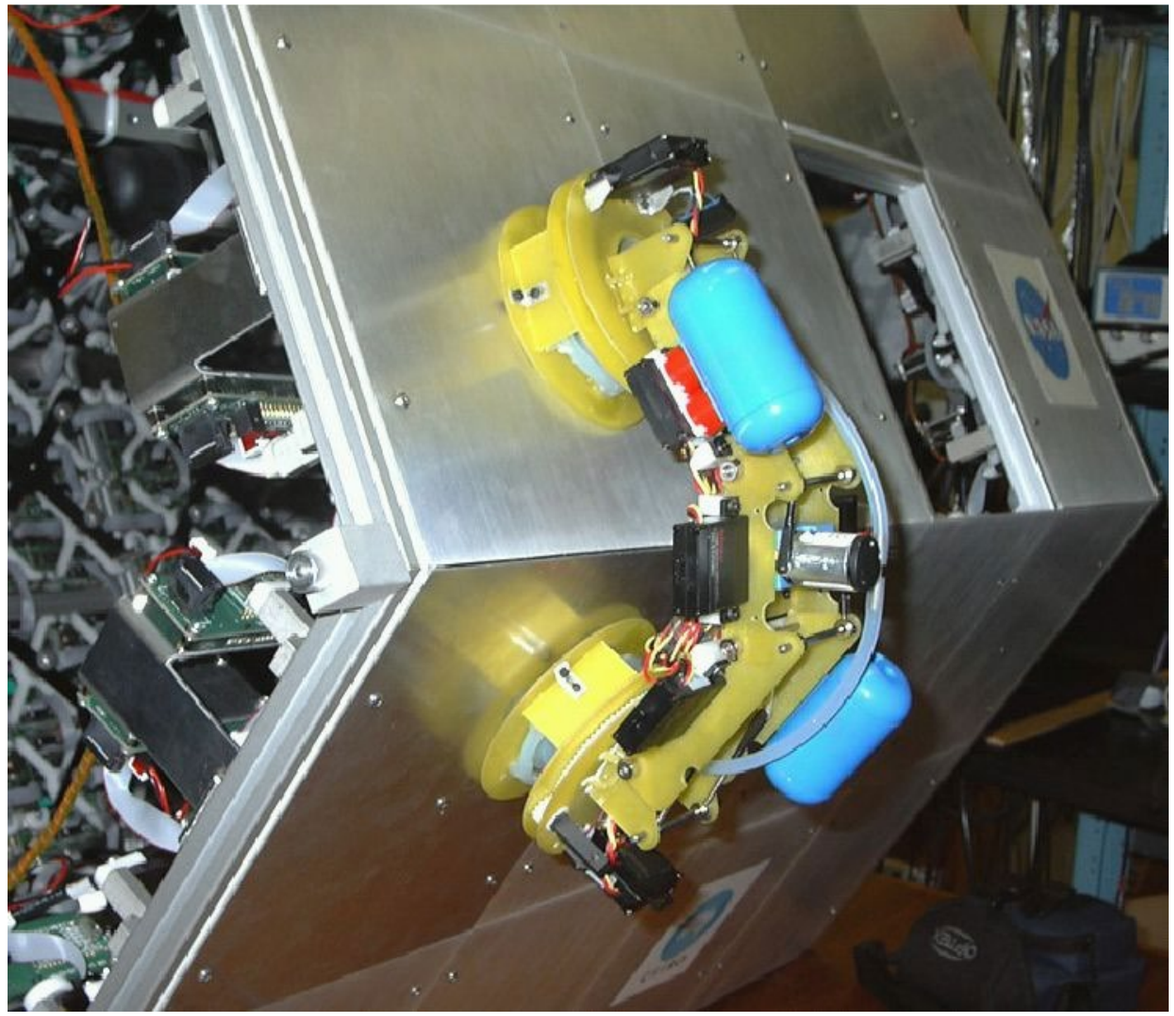
Self-organising impact networks: connecting impact points



Design vs Self-organisation

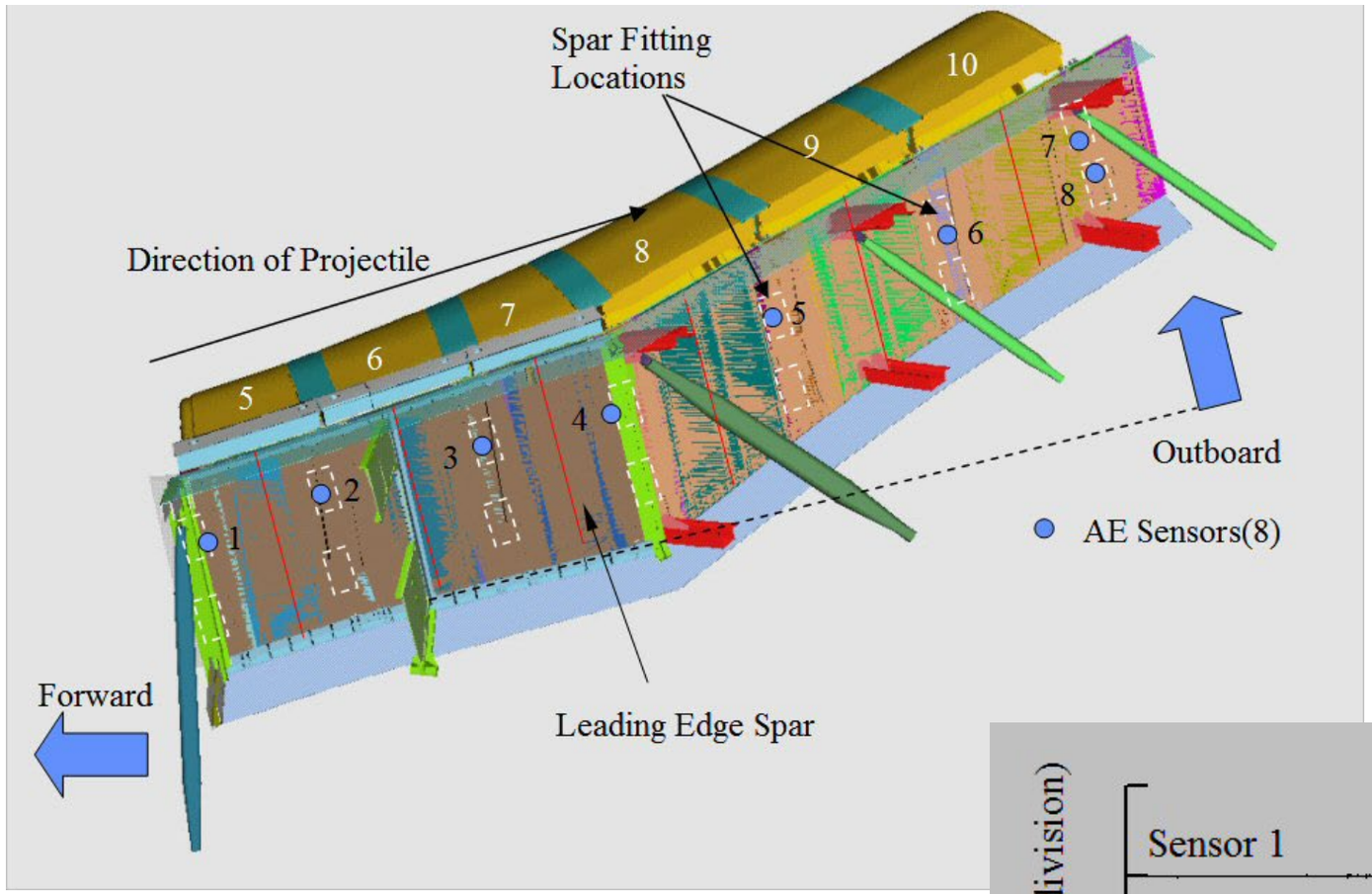


Swarm robots in a loop with embedded SHMM networks



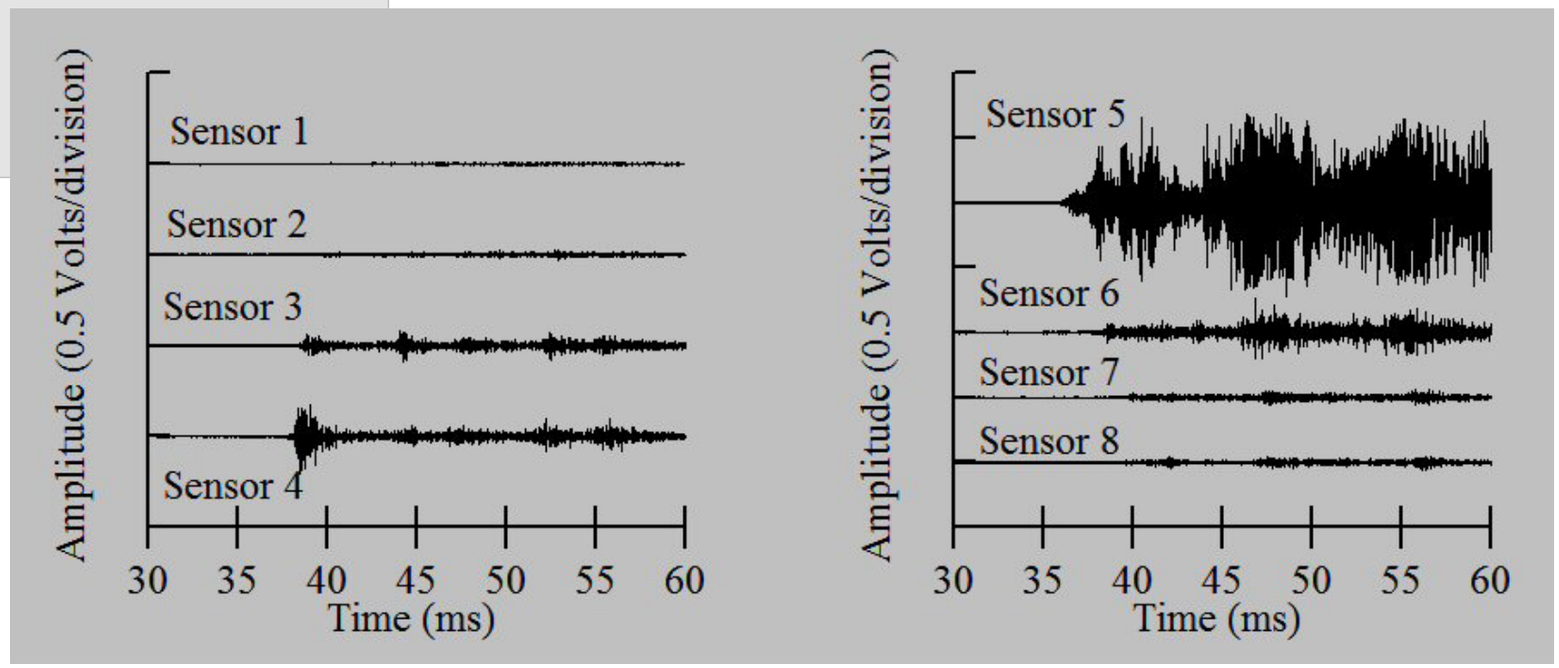
AAV project (2005)

Columbia shuttle accident investigation



Schematic of Shuttle wing leading edge test article showing RCC panels, location of impact and position of AE sensors

AE signals from foam impact on Shuttle RCC wing leading edge



W. H. Prosser, S. G. Allison, S. E. Woodard, R. A. Wincheski, E. G. Cooper, D. C. Price, M. Hedley, M. Prokopenko, D. A. Scott, A. Tessler, J. L. Spangler (2004). Structural Health Management for Future Aerospace Vehicles, *The 2nd Australasian Workshop on Structural Health Monitoring (2AWSHM)*, Melbourne, Australia, December 2004.

AI in Search and Rescue (SAR)

The Pentagon Wants AI-Driven Drone Swarms for Search and Rescue Ops

ANDY DEAN PHOTOGRAPHY/SHUTTERSTOCK.COM

https://www.nextgov.com/artificial-intelligence/2019/12/pentagon-wants-ai-driven-drone-swarms-search-and-rescue-ops/162113/

Existential risks

- Weaponisation of AI

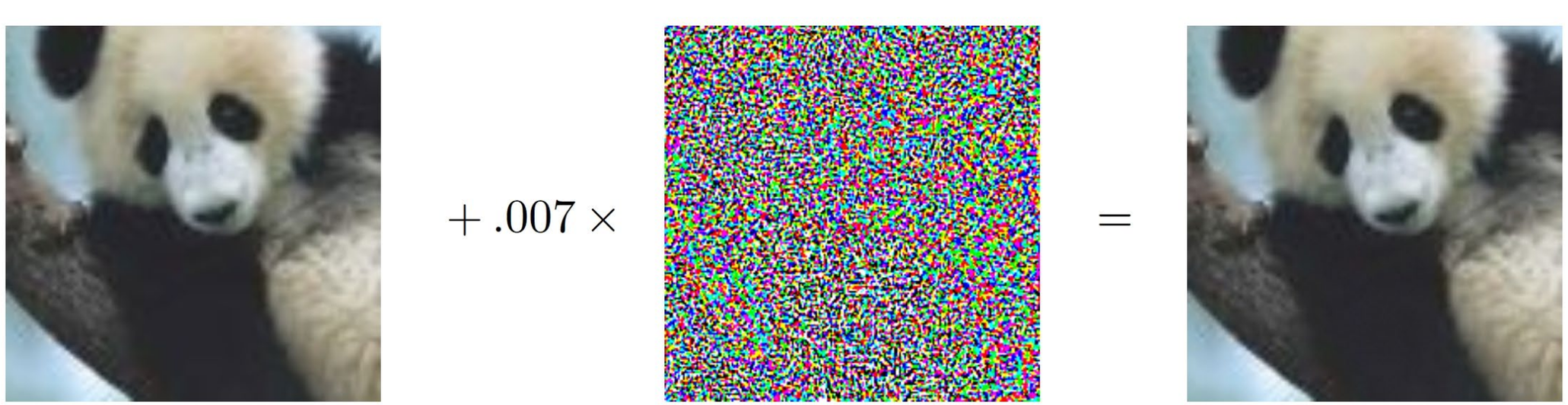
Photo: US Department of Defense / Sgt. Cory D. Payne, public domain

https://unu.edu/article/militarization-ai-has-severe-implications-global-security-and-warfare

Adversarial Machine Learning

Attacks on machine learning systems:

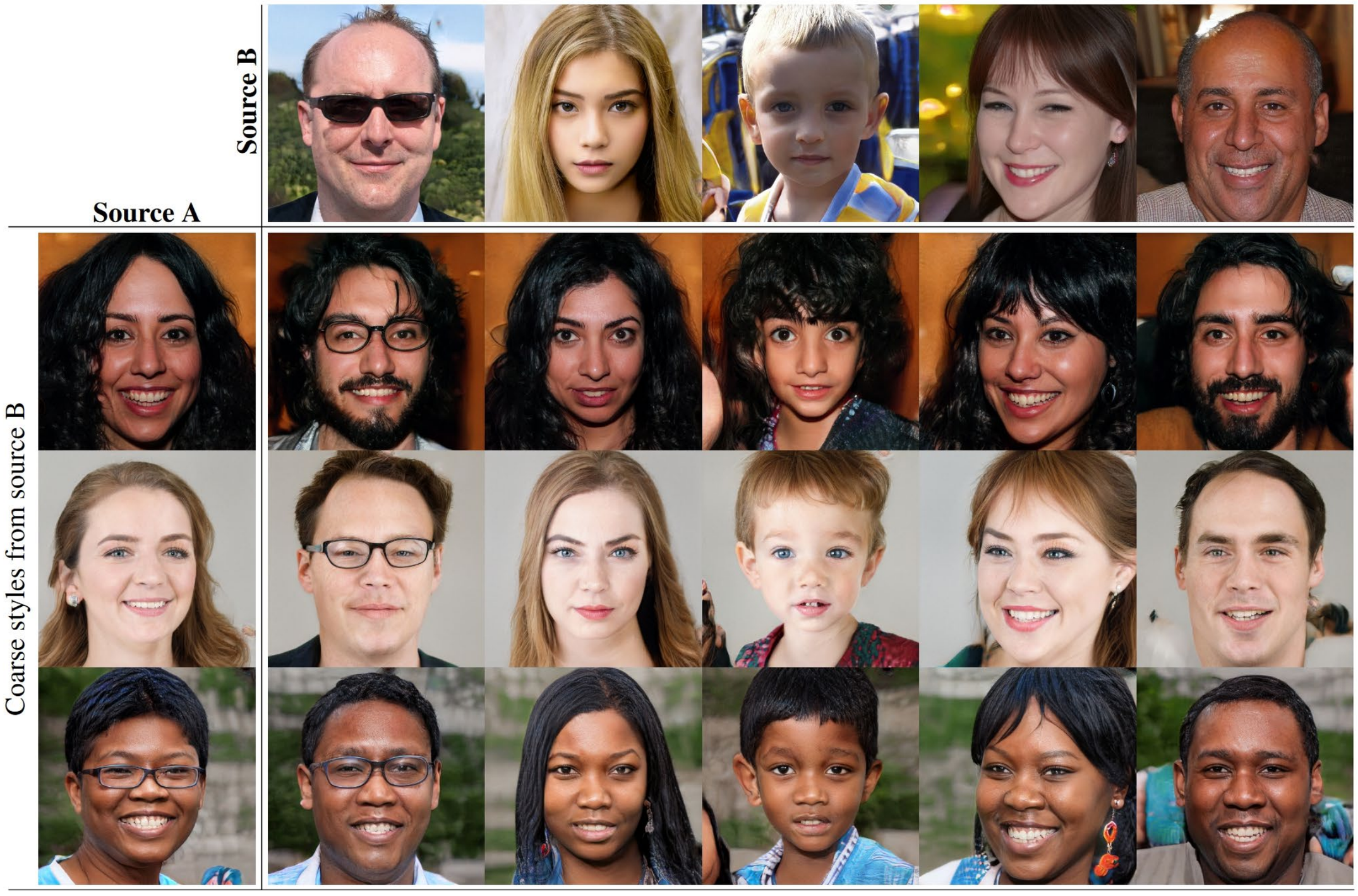
- **data poisoning**: contaminating the training dataset with data designed to increase errors in the output
  - example: an attacker might add malicious data to a dataset used to train a predictive model, causing it to make incorrect predictions
  - example: an attacker might add fake data to a dataset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

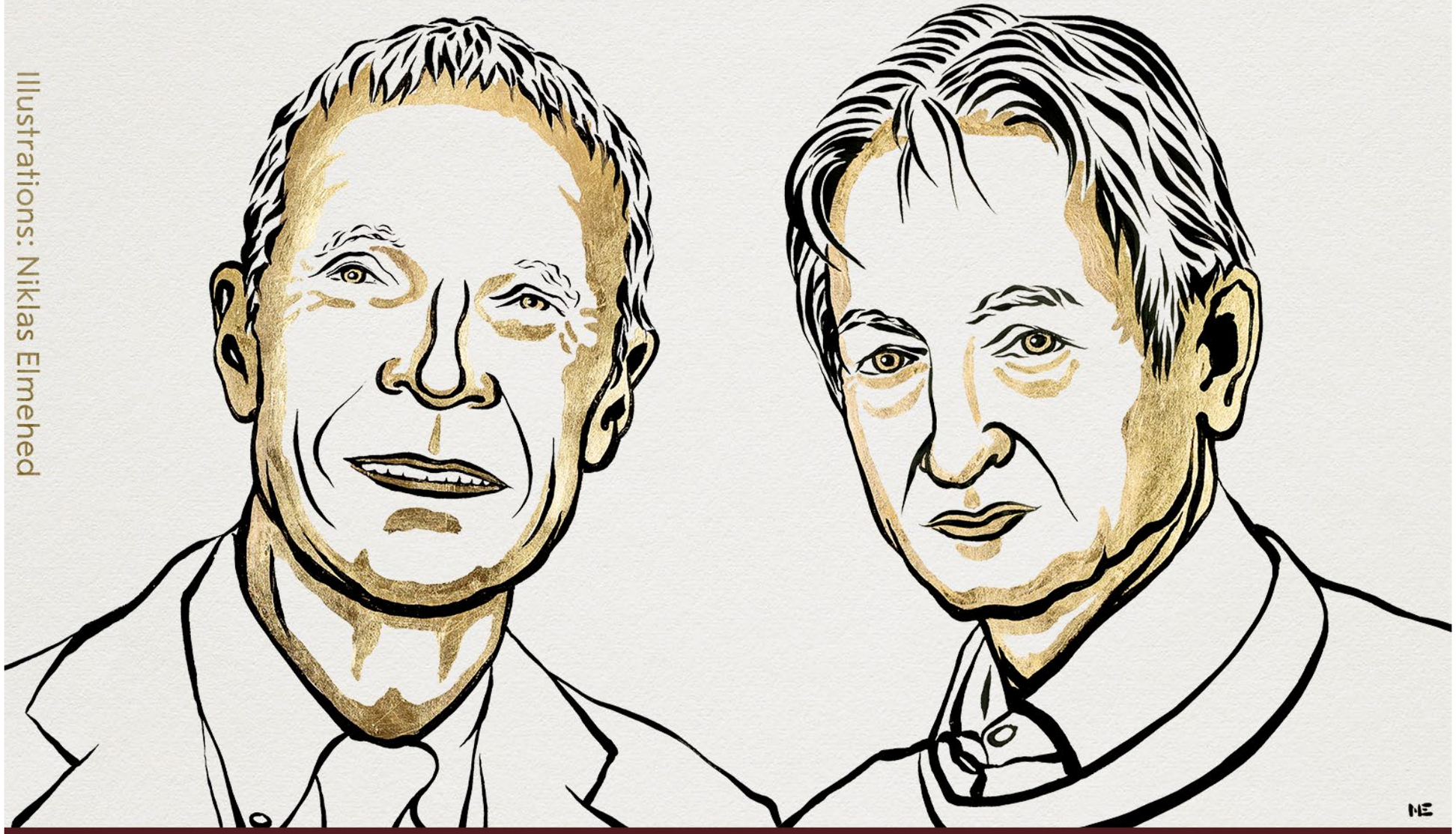
Generative adversarial networks (2018):



https://arxiv.org/pdf/1812.04948

THE NOBEL PRIZE IN PHYSICS 2024

Illustrations: Niklas Elmehed



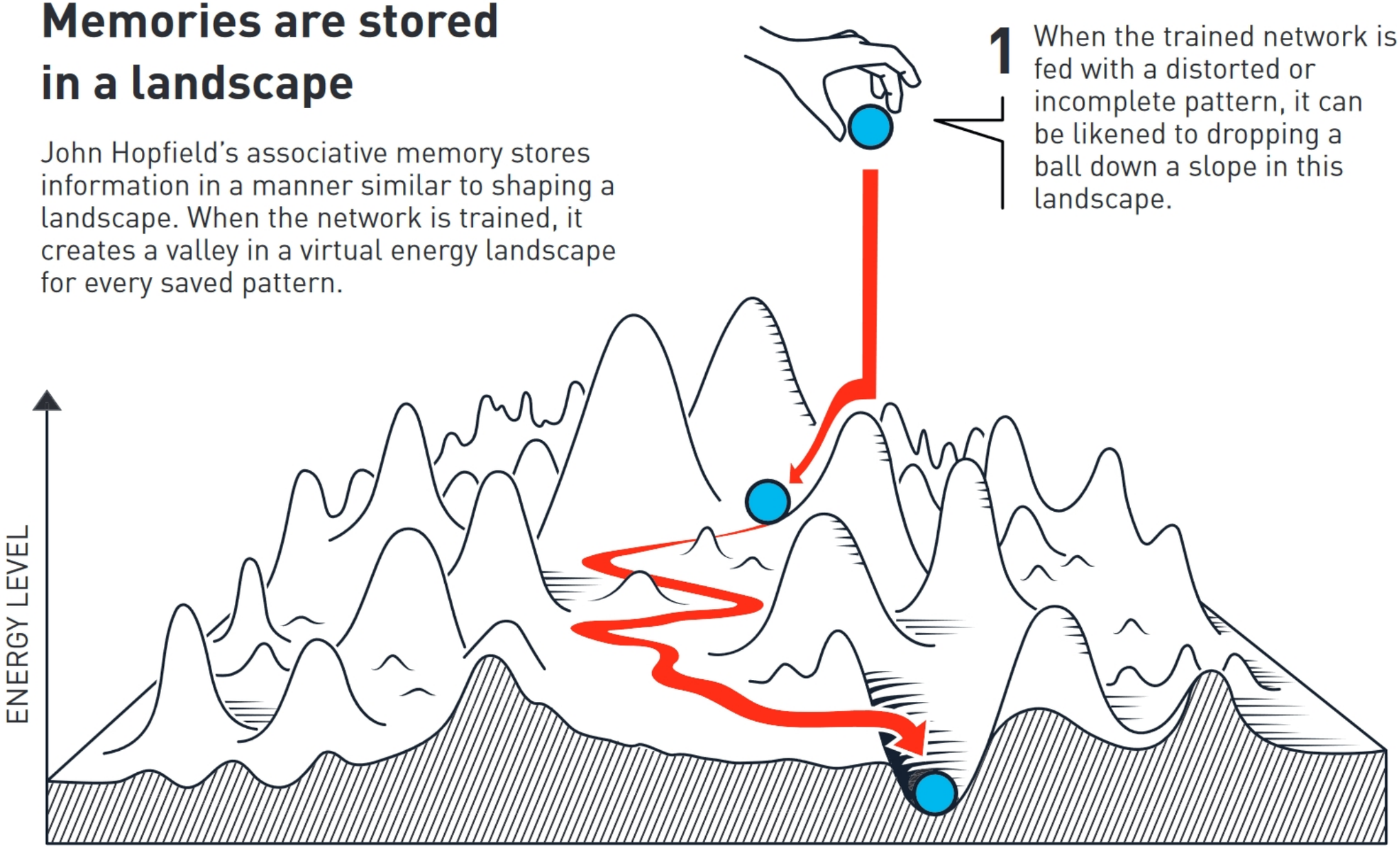
John J. Hopfield Geoffrey E. Hinton

"for foundational discoveries and inventions
that enable machine learning
with artificial neural networks"

THE ROYAL SWEDISH ACADEMY OF SCIENCES

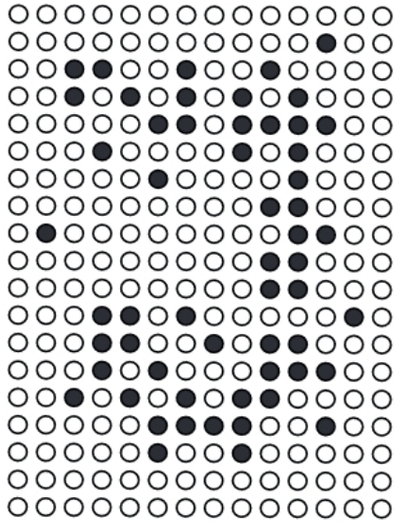
Memories are stored in a landscape

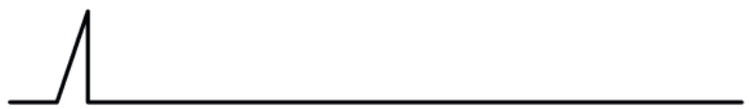
John Hopfield's associative memory stores information in a manner similar to shaping a landscape. When the network is trained, it creates a valley in a virtual energy landscape for every saved pattern.



1 When the trained network is fed with a distorted or incomplete pattern, it can be likened to dropping a ball down a slope in this landscape.

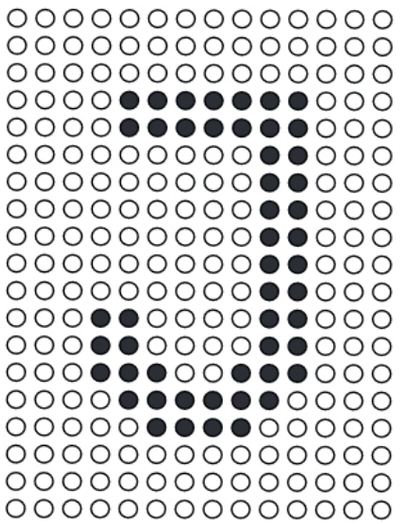
INPUT PATTERN





2 The ball rolls until it reaches a place where it is surrounded by uphills. In the same way, the network makes its way towards lower energy and finds the closest saved pattern.

SAVED PATTERN



THE NOBEL PRIZE IN CHEMISTRY 2024

Illustrations: Niklas Elmehed

David Baker

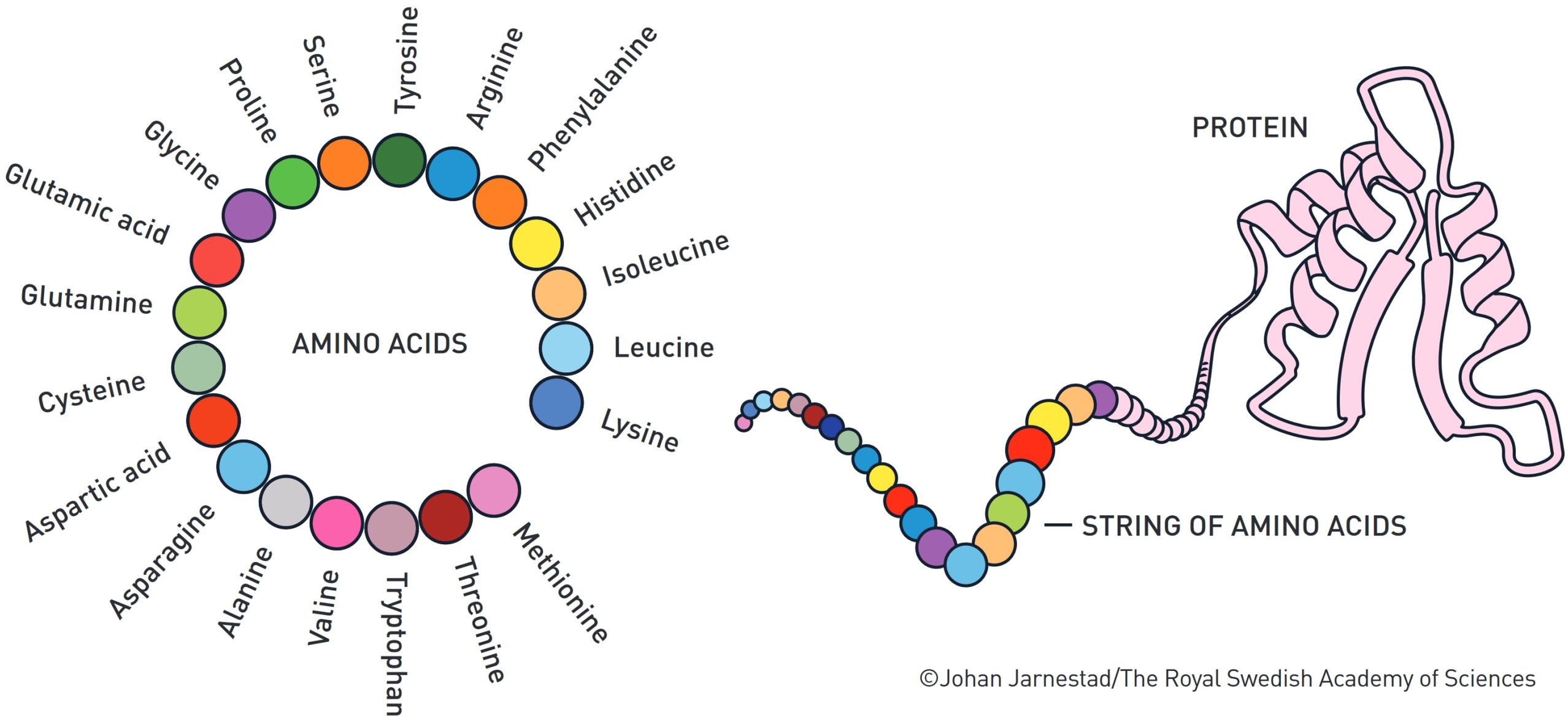
"for computational protein design"

Demis Hassabis

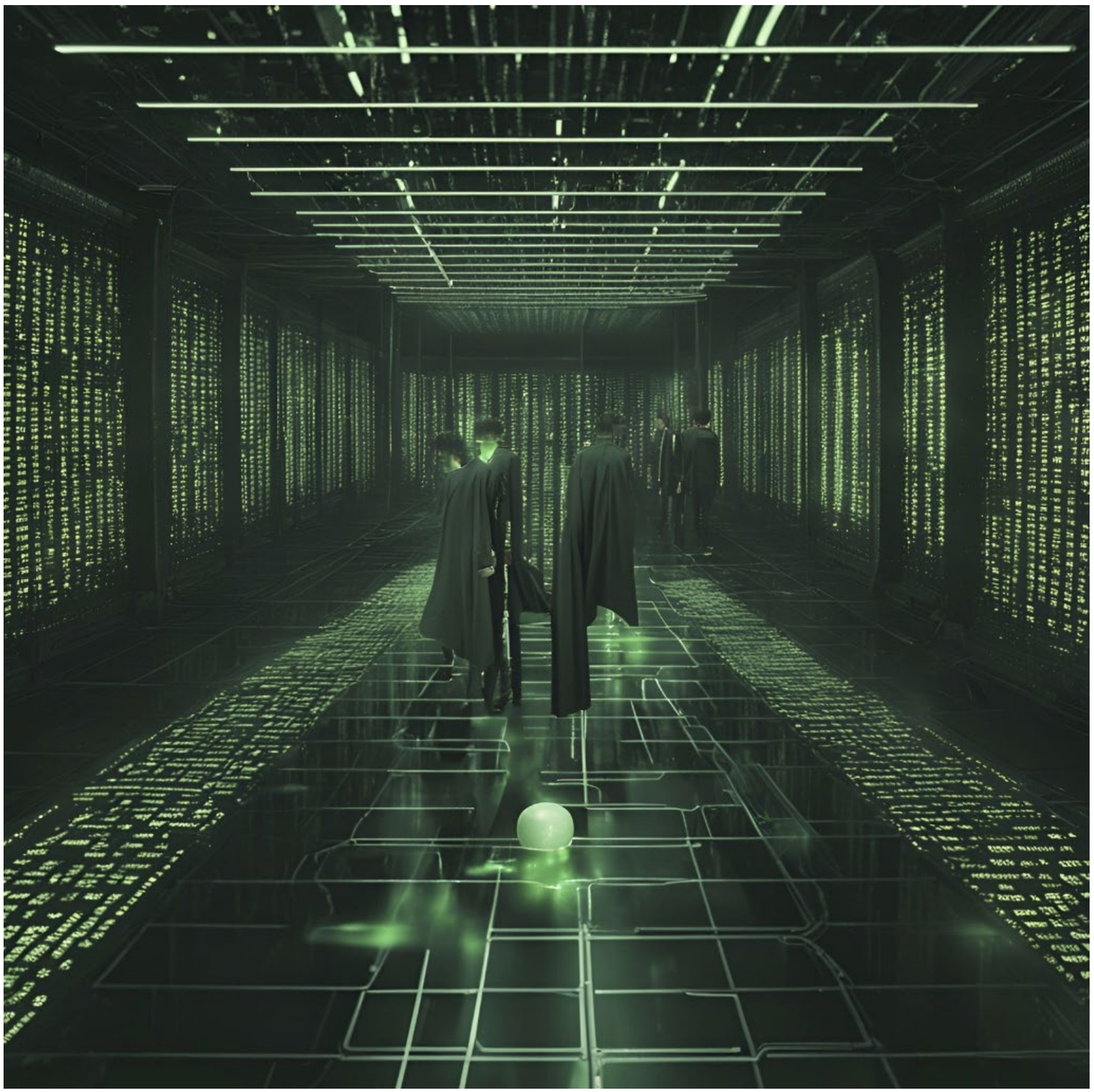
John M. Jumper

"for protein structure prediction"

THE ROYAL SWEDISH ACADEMY OF SCIENCES



AI generating its own simulation…



AI-generated image (Mikhail Prokopenko, Canva Free Online Image Generator, 17 June 2024)