COMP9208 Artificial Intelligence and Society

Lecture 4: 29 August 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

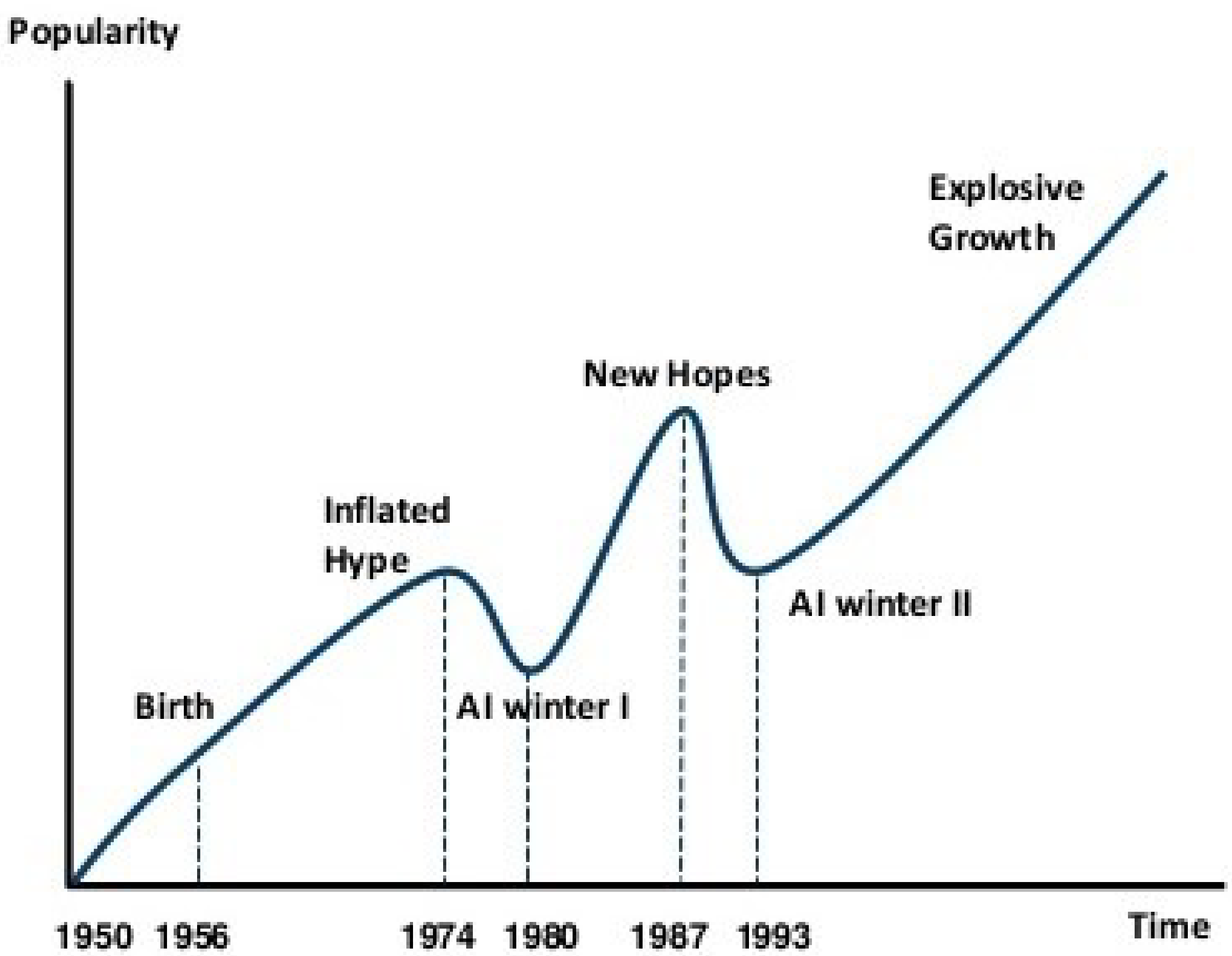
Date	Week: Workshop	Assignment
08/08	1: History of AI	
15/08	2: History of AI + Perception and Actuation	
22/08	3: History of AI + Representation and Reasoning	
29/08	4: Machine Learning	
05/09	5: Natural Interaction	
12/09	6: Distributed AI (swarm intelligence)	
19/09	7: From Good Old-Fashioned AI (GOFAI) 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	
07/11	13: Limits of cognition	
17/11	14: STUVAC	# 3: Individual assignment (data analysis report)

Brief review of last week

?

History of AI: main achievements and setbacks

AI HAS A LONG HISTORY OF BEING "THE NEXT BIG THING"...



Timeline of AI Development

- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM's Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world's champions

https://www.actuaries.digital/2018/09/05/history-of-ai-winters/

Knowledge Representation and Reasoning (KR&R)

Logical Representation

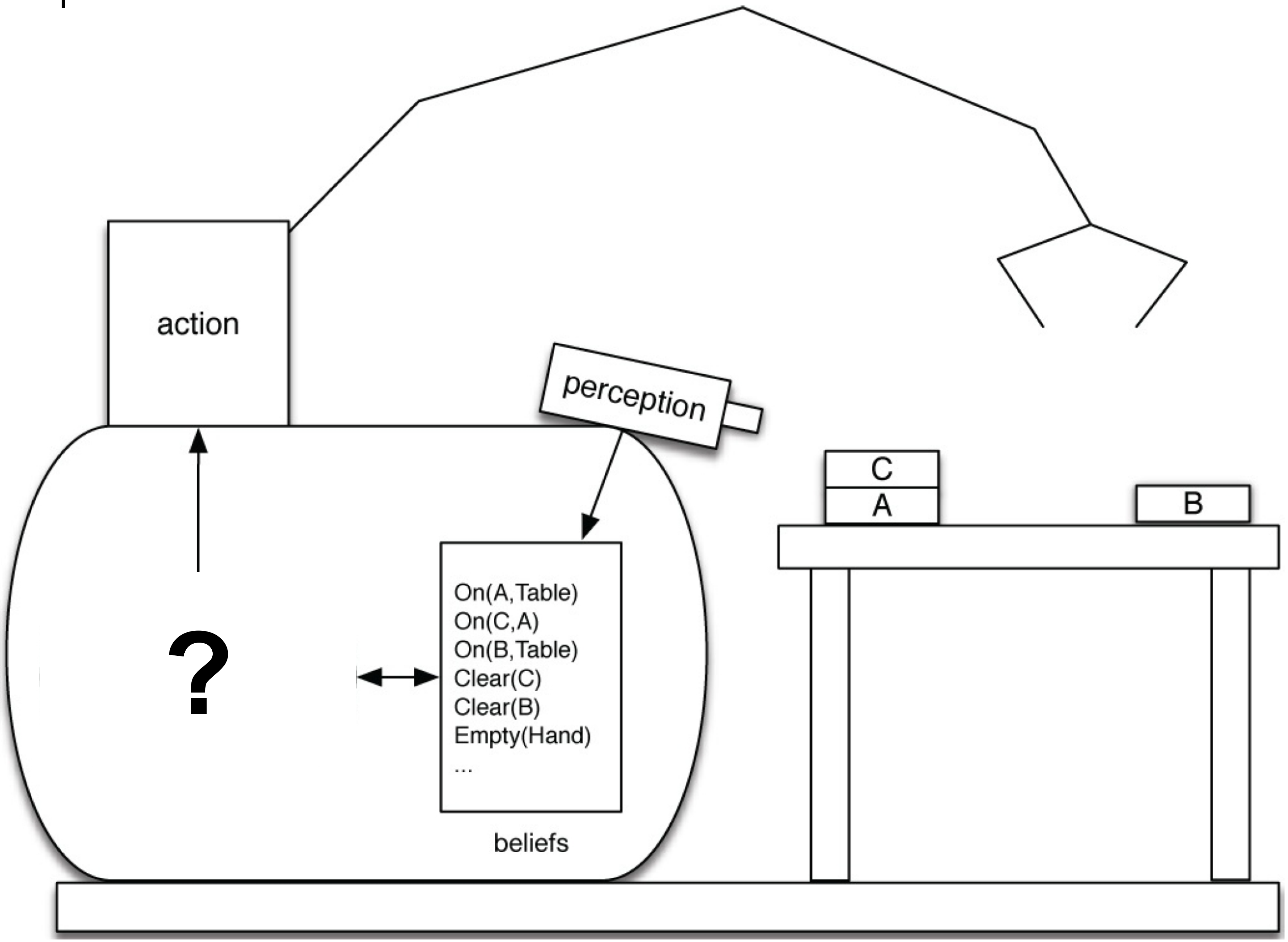


Figure 5: McCarthy's vision for logic-based AI. A logic-based AI system explicitly represents its beliefs about the environment via logical statements, and intelligent decision-making is broken down into a process of logical reasoning.

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

Knowledge Representation and Reasoning (KR&R)

Logical Representation

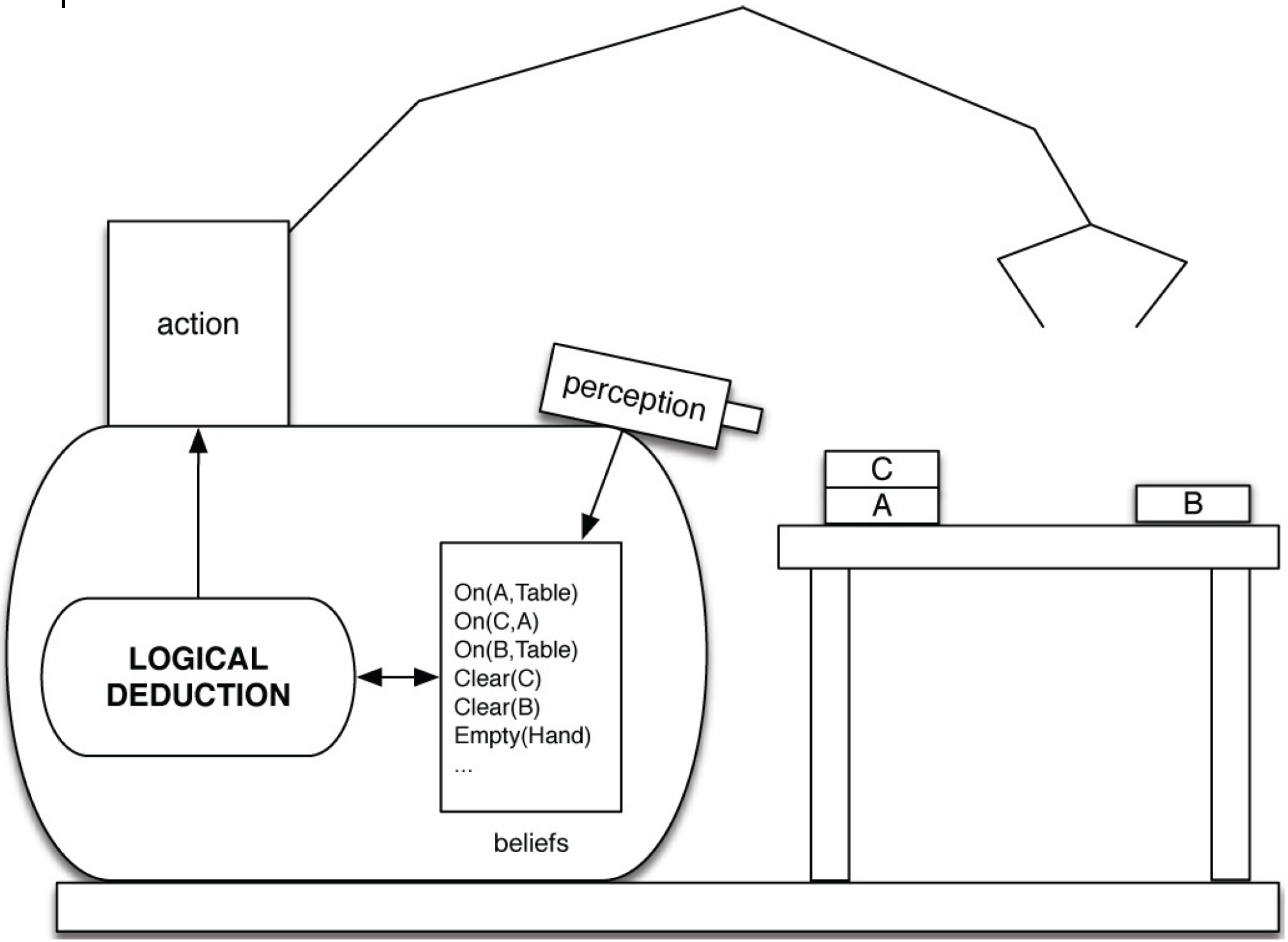


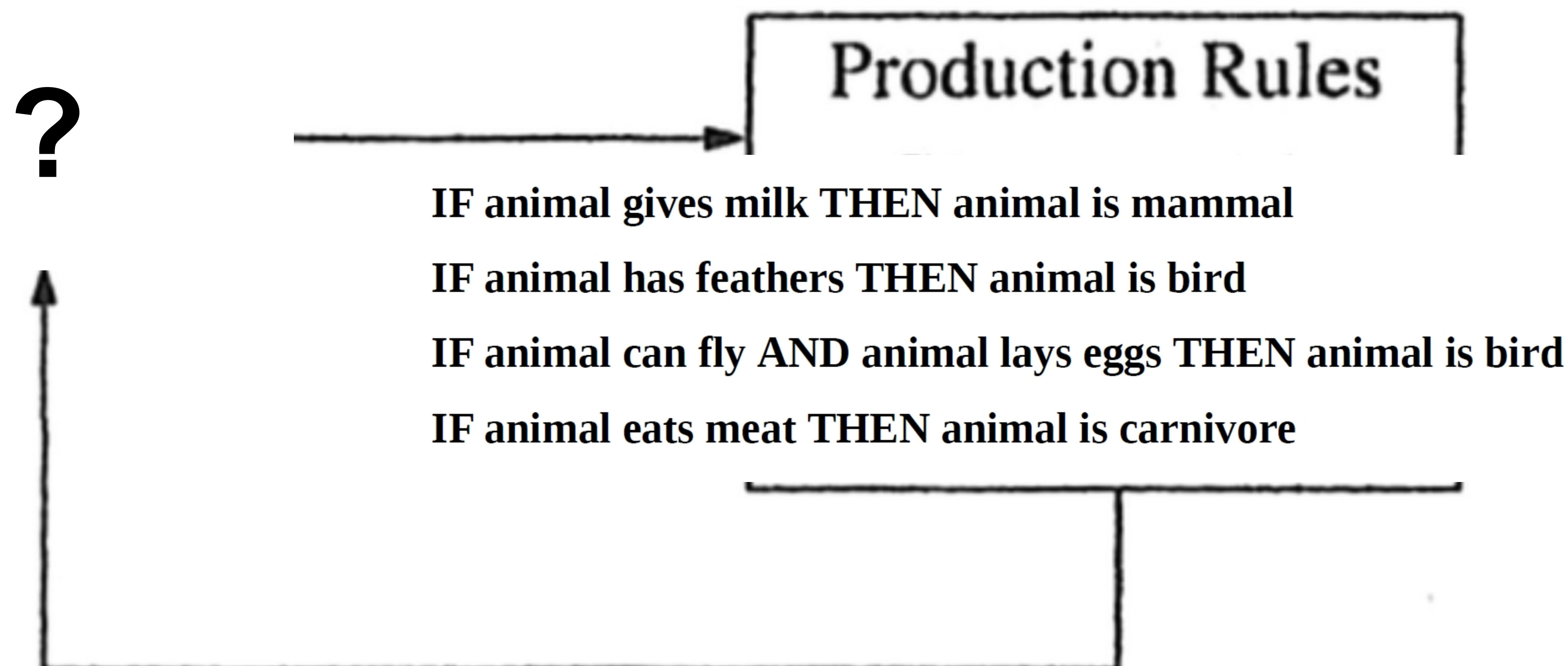
Figure 5: McCarthy's vision for logic-based AI. A logic-based AI system explicitly represents its beliefs about the environment via logical statements, and intelligent decision-making is broken down into a process of logical reasoning.

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

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

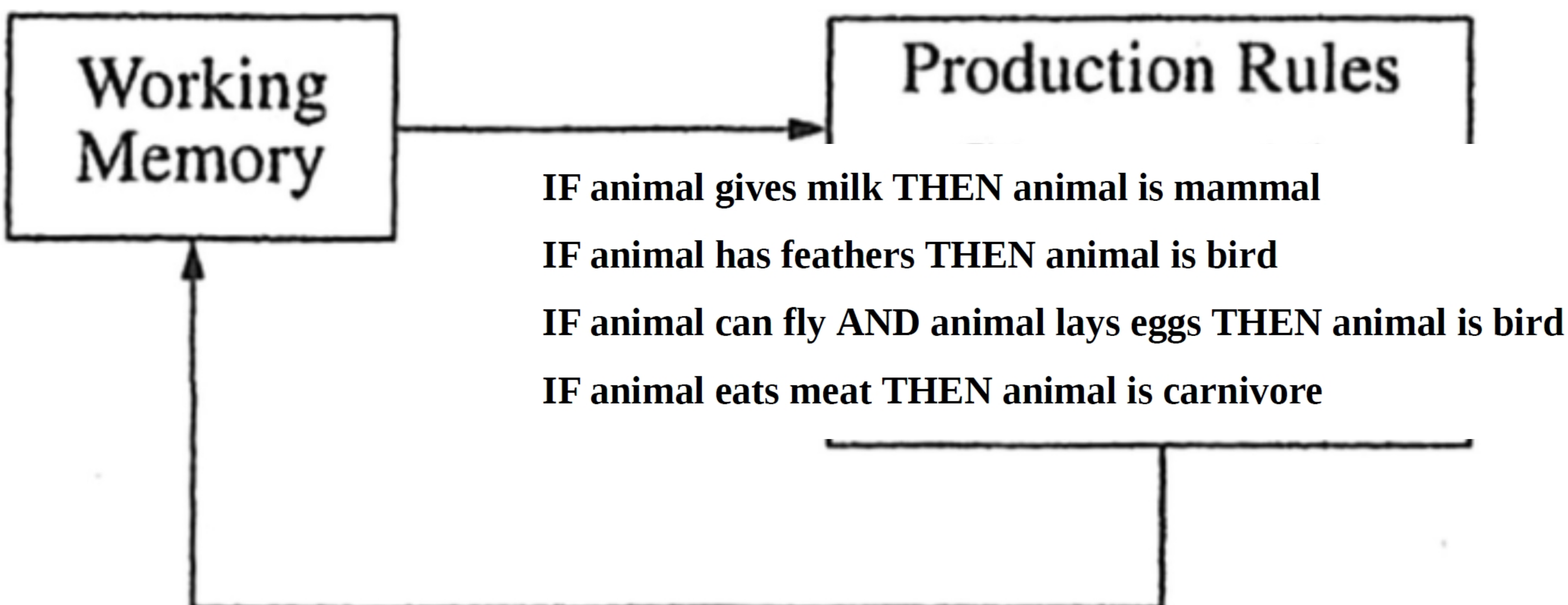


https://link.springer.com/chapter/10.1007/978-3-030-71873-2_1
M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

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



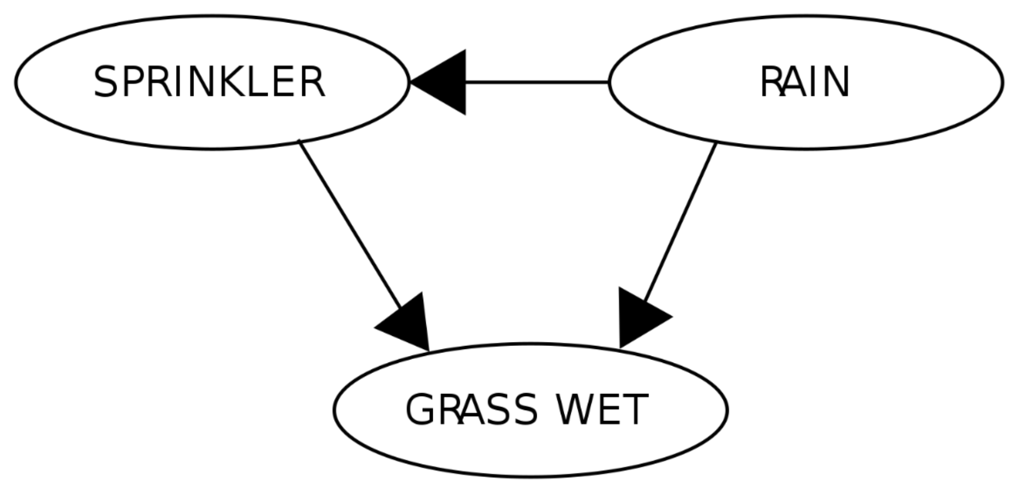
https://link.springer.com/chapter/10.1007/978-3-030-71873-2_1
M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Knowledge Representation and Reasoning (KR&R)

Bayesian networks:

- probabilistic graphical models representing a set of variables and their conditional dependencies
- useful for reasoning under uncertainty
- example: medical diagnosis with symptoms probabilistically linked to disease

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	?	0.01

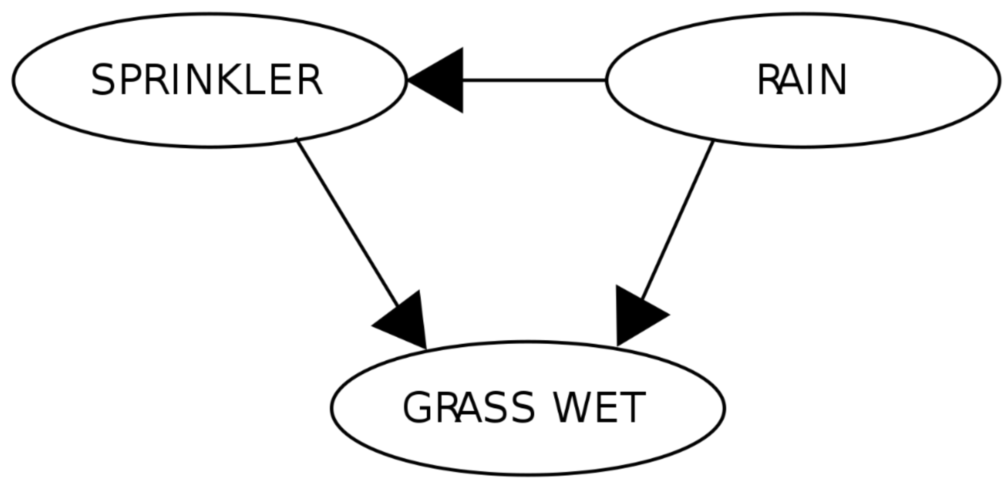
https://en.wikipedia.org/wiki/Bayesian_network

Knowledge Representation and Reasoning (KR&R)

Bayesian networks:

- probabilistic graphical models representing a set of variables and their conditional dependencies
- useful for reasoning under uncertainty
- example: medical diagnosis with symptoms probabilistically linked to disease

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

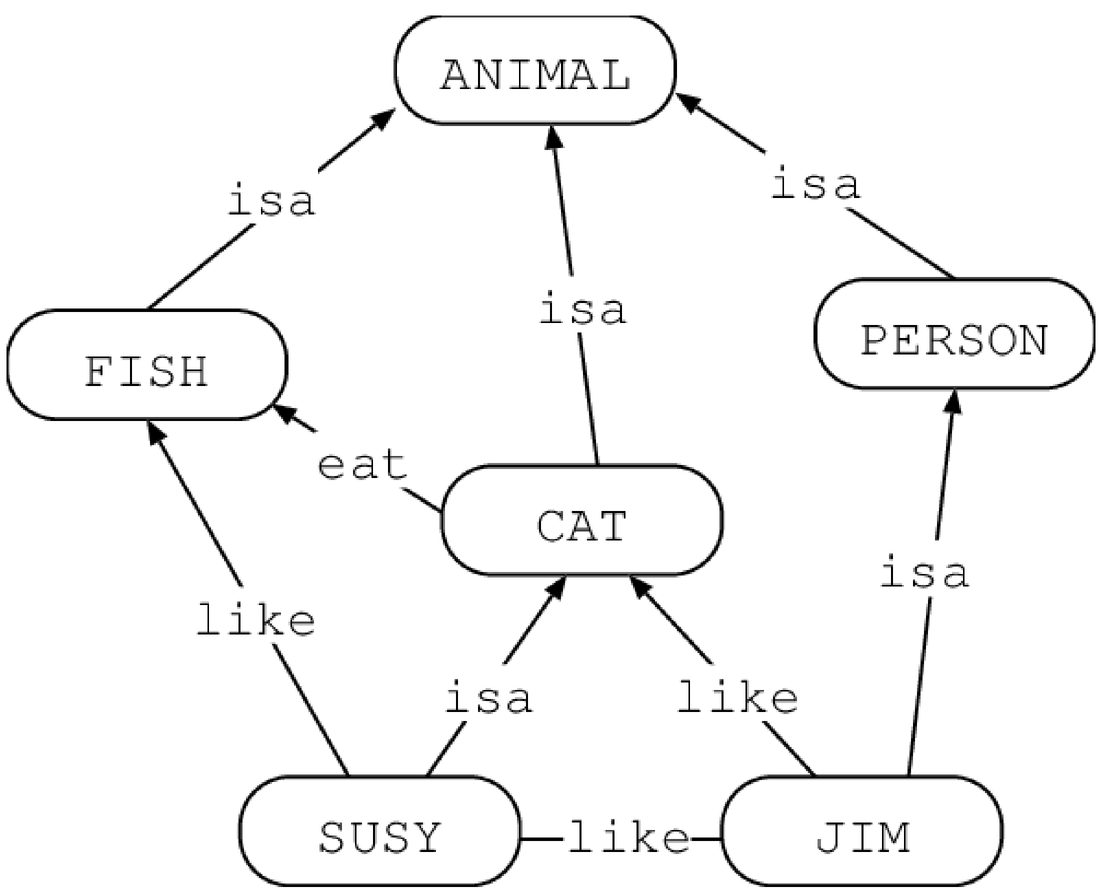
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

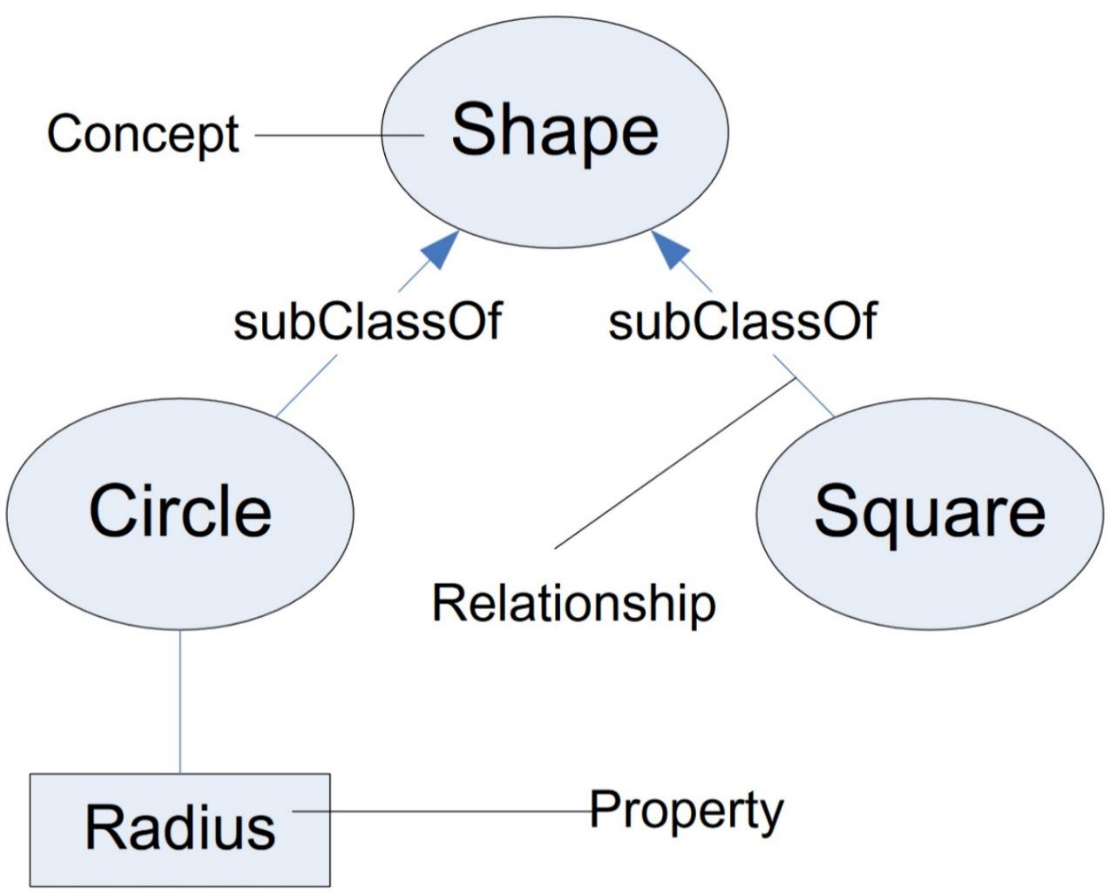
https://en.wikipedia.org/wiki/Bayesian_network

Knowledge Representation and Reasoning (KR&R)

Semantic networks vs ontologies

- semantic networks are more intuitive and more simple, less focused on interoperability
- ontologies are more formal and rigorous, often defined using formal languages, to represent complex and domain-specific knowledge
- ontologies are designed to enable interoperability and integration across different systems and domains





Knowledge Representation and Reasoning (KR&R)

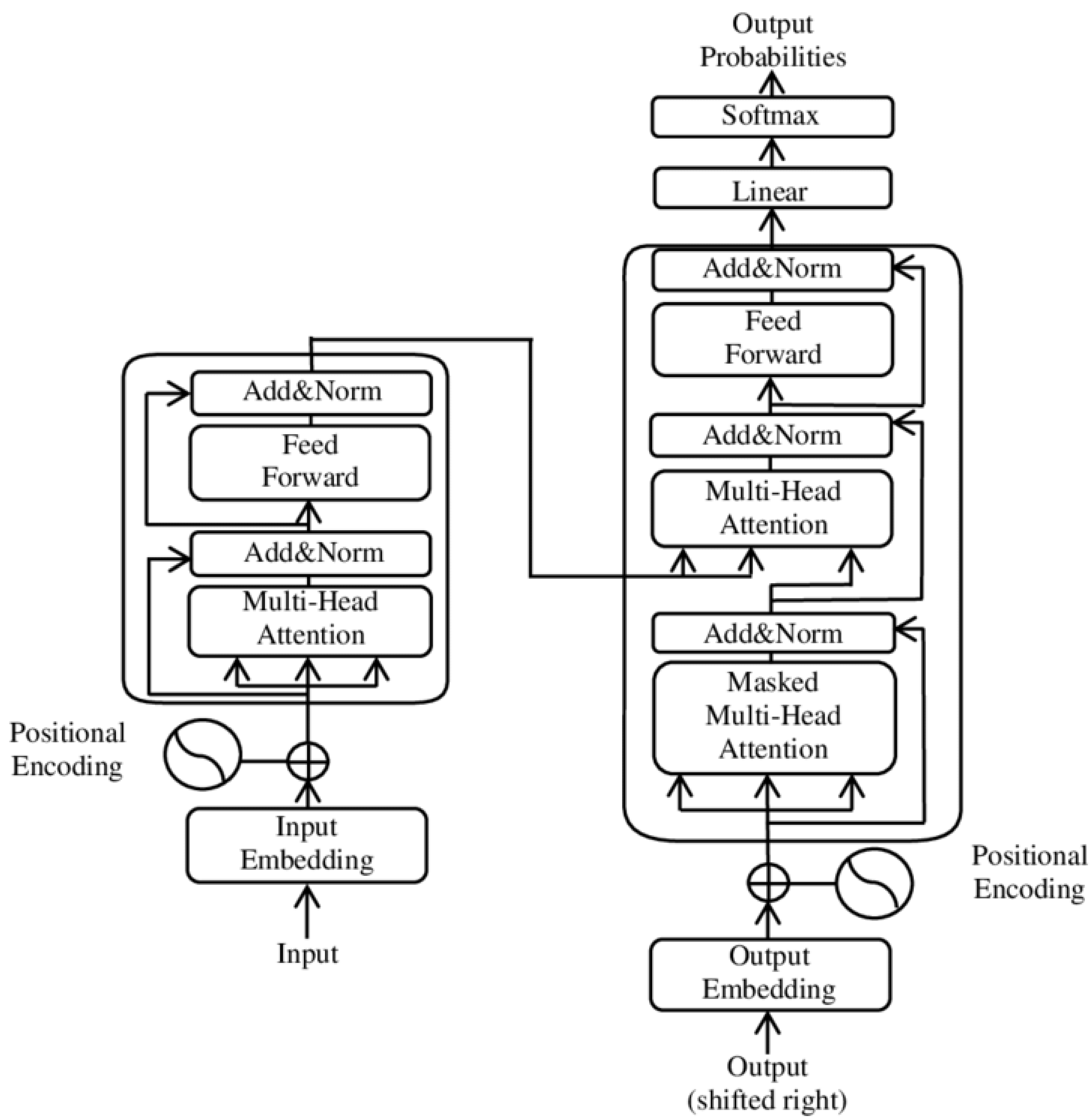
Ethical concerns:

- ❖ misuse of information
  - o AI systems may generate false or misleading information based on reasoning processes, contributing to misinformation campaigns
  - o enhanced surveillance capabilities using knowledge representation to track and profile individuals
- ❖ intellectual property and ownership
  - o disputes over the ownership of knowledge graphs created from proprietary data sources
  - o legal challenges regarding the use of copyrighted material in building knowledge bases
- ❖ ethical reasoning
  - o autonomous systems making decisions that involve ethical trade-offs, such as prioritising one life over another in critical situations
  - o AI systems used in healthcare or social services need to consider ethical implications of their recommendations and actions

ChatGPT 4o

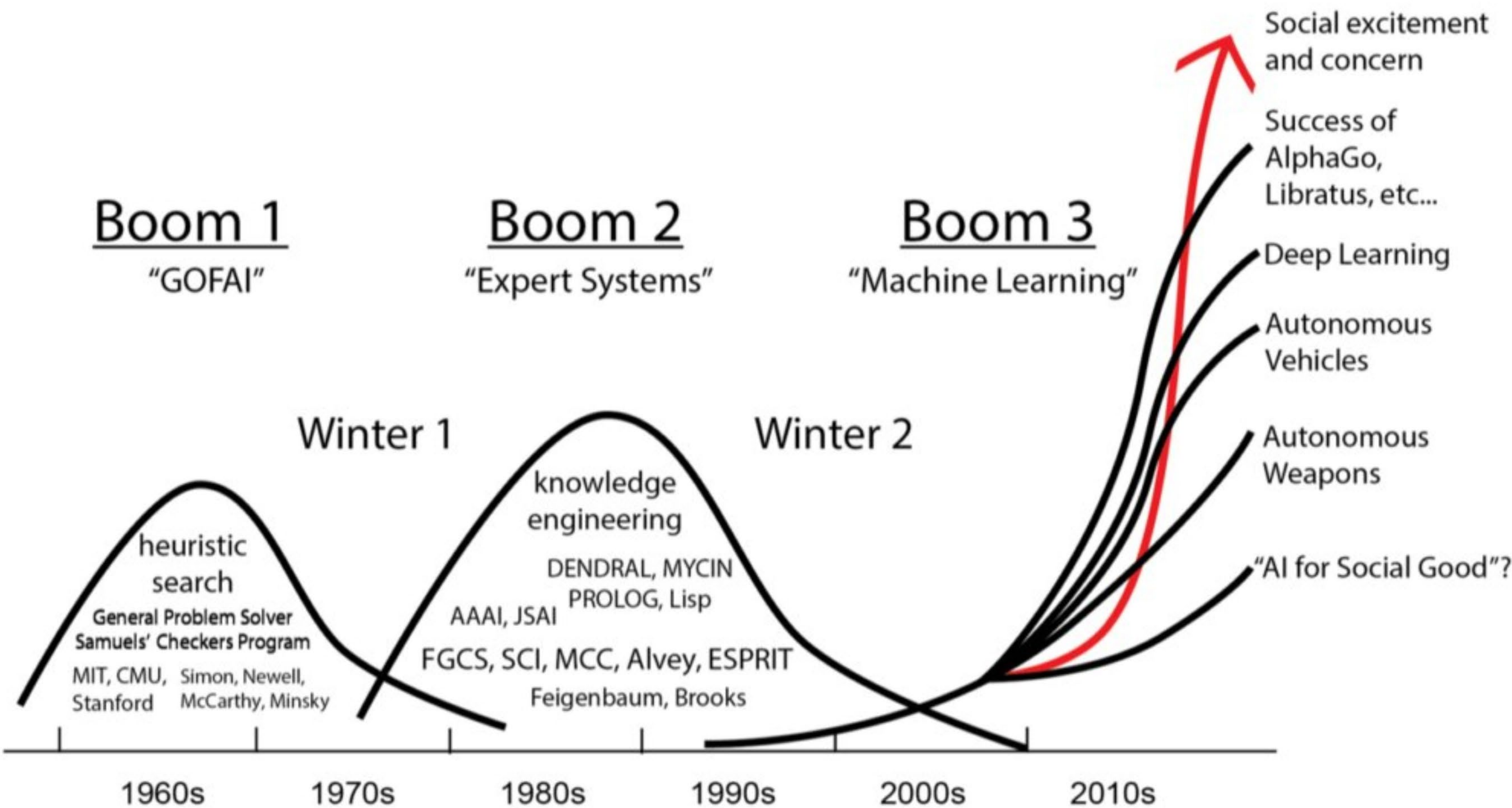
History of AI: main achievements and setbacks

"Attention Is All You Need" (2017, the transformer architecture)



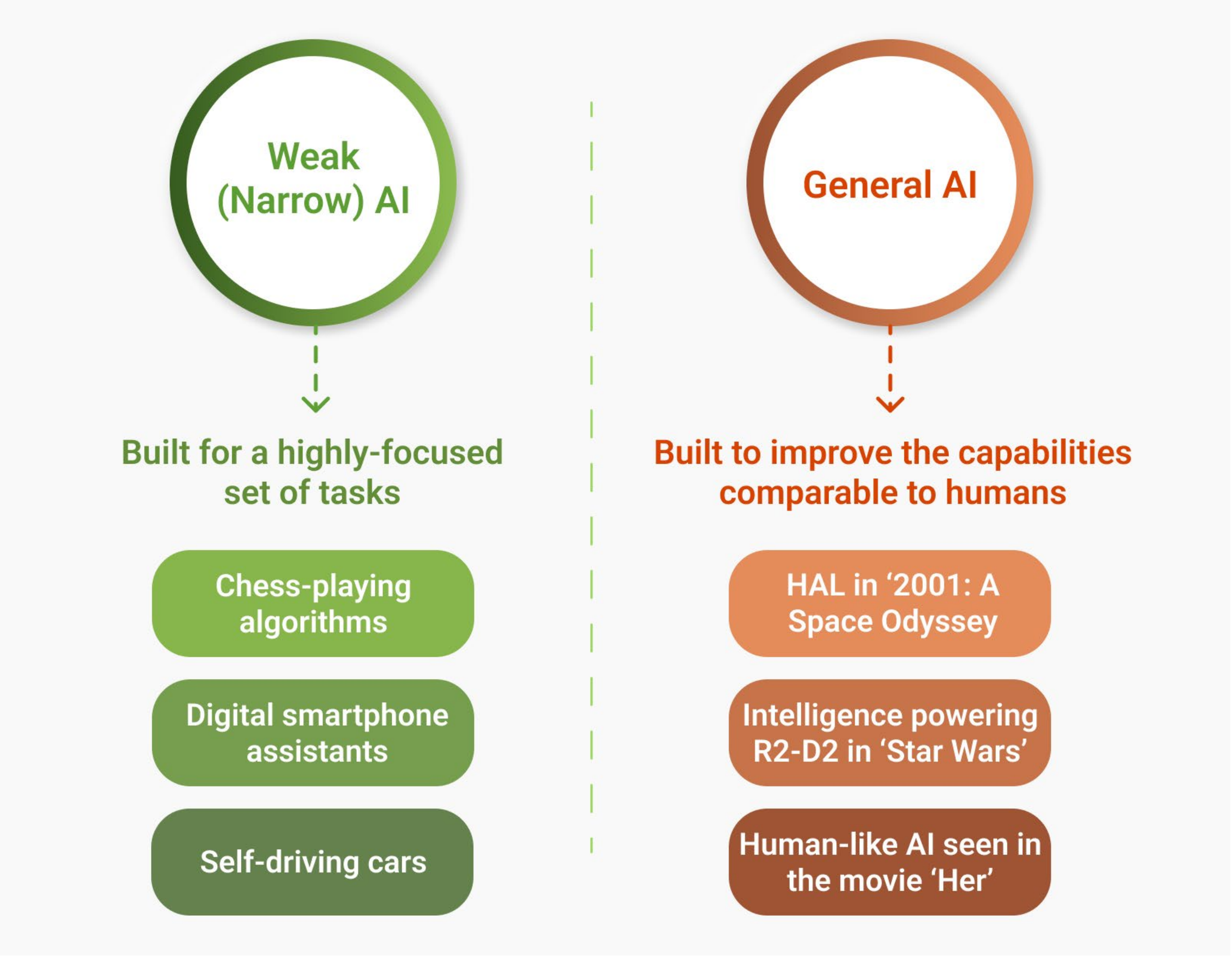
https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

History of AI: main achievements and setbacks



https://jaylatta.net/history-of-ai-from-winter-to-winter/

History of AI: main achievements and setbacks



https://www.plugger.ai/blog/general-ai-vs-narrow-ai-2022-guide

History of AI: main achievements and setbacks

Narrow AI

- AI focused on a specific, singular or limited task
- Examples include image recognition, hyperpersonalization, chatbots, predictive text
- Trained on specific tasks by data scientists
- Correlates questions or assignments to a specific data set to accomplish a task
- No self-awareness, consciousness, ability to think

Artificial general intelligence (AGI)

- Not fully realized, with some developers questioning if it will be possible
- Seeks machines that can handle a range of cognitive tasks with little oversight
- The ability to learn, generalize, apply knowledge and plan for the future
- Must consistently pass the Turing Test
- Single, general intelligence that possesses common sense and creativity and expresses emotions

https://www.techtarget.com/searchenterpriseai/feature/General-AI-vs-narrow-AI-comes-down-to-adaptability

Homework

- check the information sources provided on different slides (at least some)

- "AlphaGo—The movie. Full documentary."
  YouTube: https://www.youtube.com/watch?v=WXuK6gekU1Y

- please try to watch this documentary before the next lecture, and prepare to give a few brief comments about it:
  - what was most striking about the AI technology in the movie?
  - what societal impact did the AI make?
  - what were the main AI limitations, biases and risks?

?

Discussion

?

Assignments (example)

There are 3 assignment activities for this subject:

(1) Individual Report (30%)	– week 7	55/100 $\rightarrow$ 16.5 / 30
(2) Group Presentation (40%)	– week 11	64/100 $\rightarrow$ 25.6 / 40
(3) Individual Report (30%)	– week 14	78/100 $\rightarrow$ 23.4 / 30
Total (100%)		65.5 $\approx$ **66**

❖ Text matching software (Turnitin) will be used for plagiarism detection in (1) and (3)

Assignment #1

Assignment #1
(report: article review)

You will find which article is allocated to you in Canvas, in instructions for Assignment #1

Please submit in Canvas by Turnitin (pdf) a report with your Student ID in the file name, 3000 words maximum, with answers to specific questions (see next slides)

Deadline: **18 September, Thursday, 11:59 pm (week 7)**

Section 1: Background and contextual analysis

1.1 Analyse how the article follows earlier AI traditions (e.g., symbolic AI, connectionism, cybernetics) by identifying two key publications that influenced its theory or method. How does the article reconcile, challenge, or expand on these influences within their historical context?

1.2 Consider the socio-technical conditions at the time — such as dominant paradigms, funding and social attitudes — and how these shaped the article's problem framing and assumptions. How might these differ if written today?

Section 2: Novelty and societal implications

2.1 Identify one major technical concept or approach introduced or popularised by the article. How has this concept propagated across disciplinary boundaries and application domains?

2.2 Provide at least one concrete example of its use (or misuse) in a societal context, critically evaluating the implications.

Section 3: Comparative research impact

3.1 How was the article received in academic and non-academic communities over time? Analyse its research impact: consider citations, follow-up studies, and critiques, and identify which factors have been amplified, ignored, or distorted, and why.

3.2 Then, compare it with one contemporary or near-contemporary publication that offers a contrasting view to the article's core claims (e.g., opposing methodology or ethics), and describe the key difference(s).

Section 4: Future impact

4.1 Does the article make explicit or implicit claims about the future of AI or intelligence? Analyse the assumptions used to make these projections. With hindsight, how accurate were they?

4.2 What can this tell us about the limitations and power of futurism in AI research?

Section 5: Ethics of AI

5.1 If ethical considerations are present or absent in the article, how should this be interpreted given the historical moment of its publication? What are the main limitations, biases and risks?

5.2 How would the inclusion of today's ethical frameworks (e.g., fairness, accountability, transparency, privacy) alter the article's claims, limitations or conclusions?

Marking individual report (article review)

	Weight	0-49%	50-64%	65-74%	75-84%	85-100%	Mark
Background and contextual analysis	20%	Not placed in the context of AI research and applications	Partially placed in the context of AI research and applications	Placed in the context of AI research and applications, with a brief or unstructured description	Placed in the context of AI research and applications, with a brief but structured description	Placed in the context of AI research and applications, with a detailed and structured description	
Novelty and societal implications	20%	Significantly mis-identified novelties and implications	Some novelties and implications identified with a brief or unstructured description	Majority of novelties and implications identified with a brief or unstructured description	Majority of novelties and implications identified with a brief but structured description	Majority of novelties and implications identified with a detailed and structured description	
Comparative research impact	20%	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	
Future impact	20%	Impact described inadequately	Impact described incompletely	Impact described but without sufficient detail, structure or argumentation	Impact described in sufficient detail, but without strong argumentation	Impact described adequately, with sufficient detail, structure and argumentation	
Ethics of AI: limitations: biases and risks	20%	Significantly mis-identified limitations	Some limitations identified with a brief or unstructured description	Majority of limitations identified but without sufficient detail, structure or argumentation	Majority of limitations identified in sufficient detail but without strong argumentation	Majority of limitations fully identified, with sufficient detail, structure and argumentation	
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

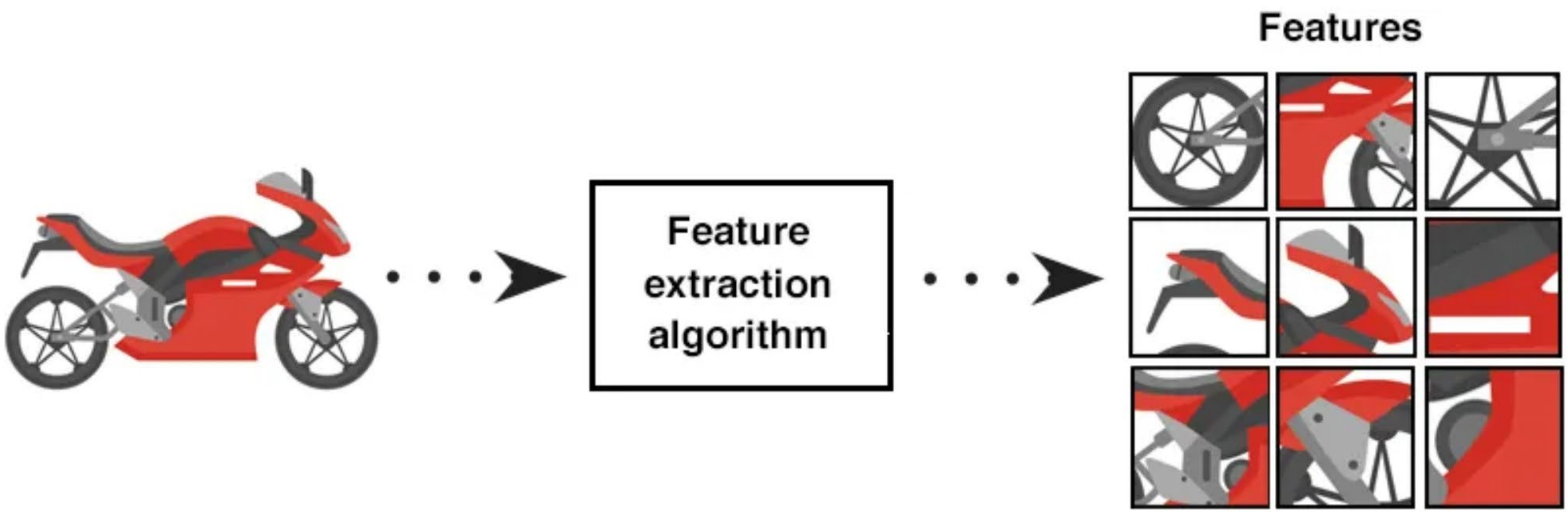
Questions

?

Machine Learning

Feature extraction:

- identifies patterns (visual, audio, ...) within data (image, audio, ...) that can be used to distinguish one object from another

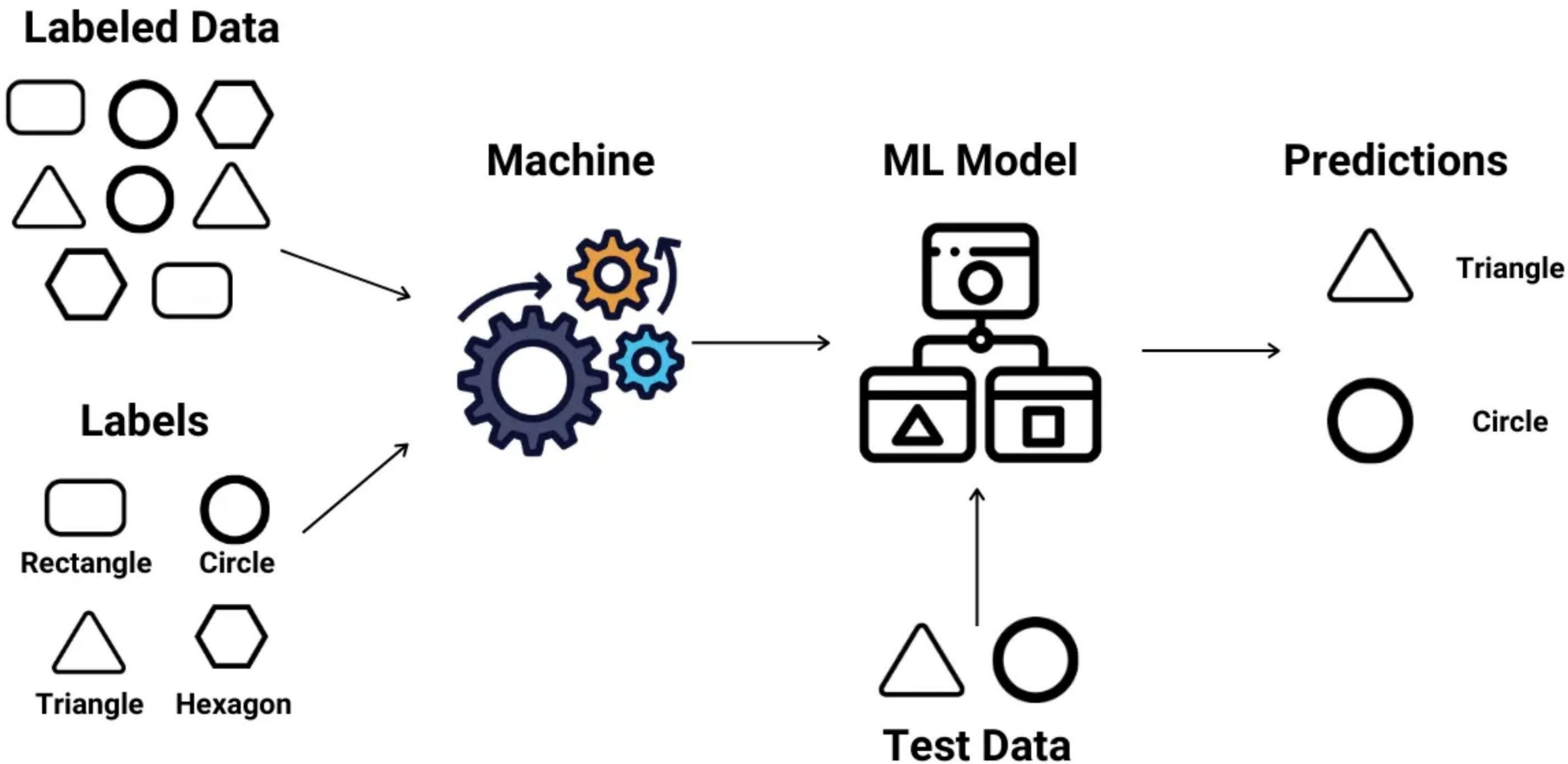


https://medium.com/@evertongomede/feature-engineering-for-computer-vision-f01a76d8058c

Machine Learning

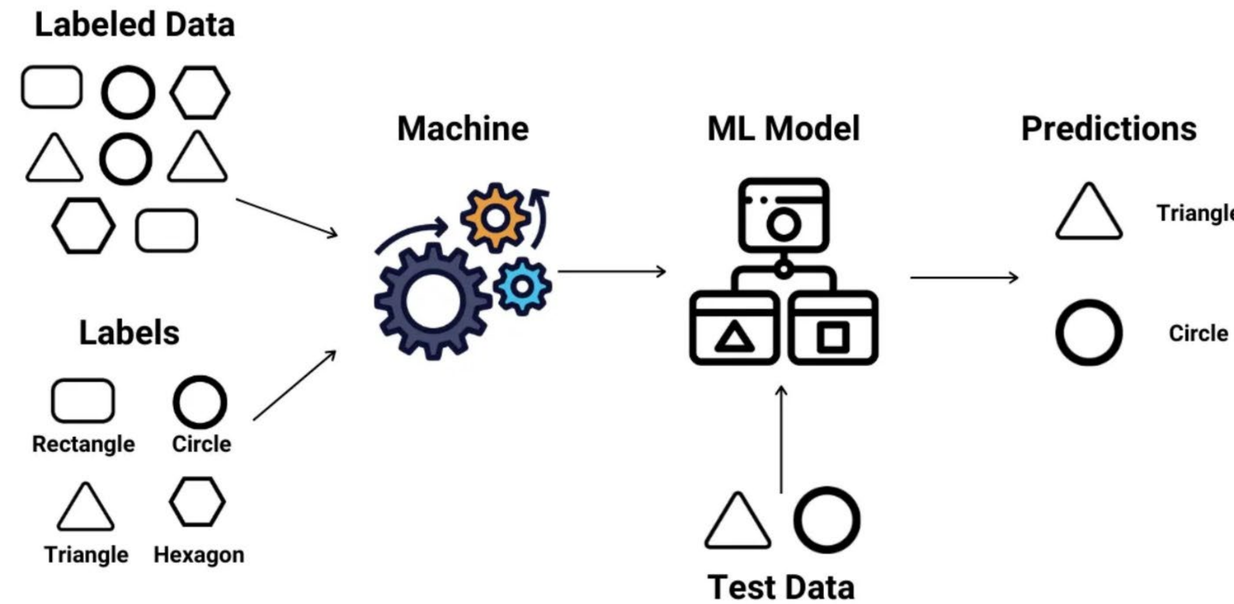
Supervised learning:

- trained on a labelled dataset: each training example is coupled with an output label
- learns a mapping from inputs to outputs to predict the labels of new, unseen data



https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644

Machine Learning



Supervised learning:

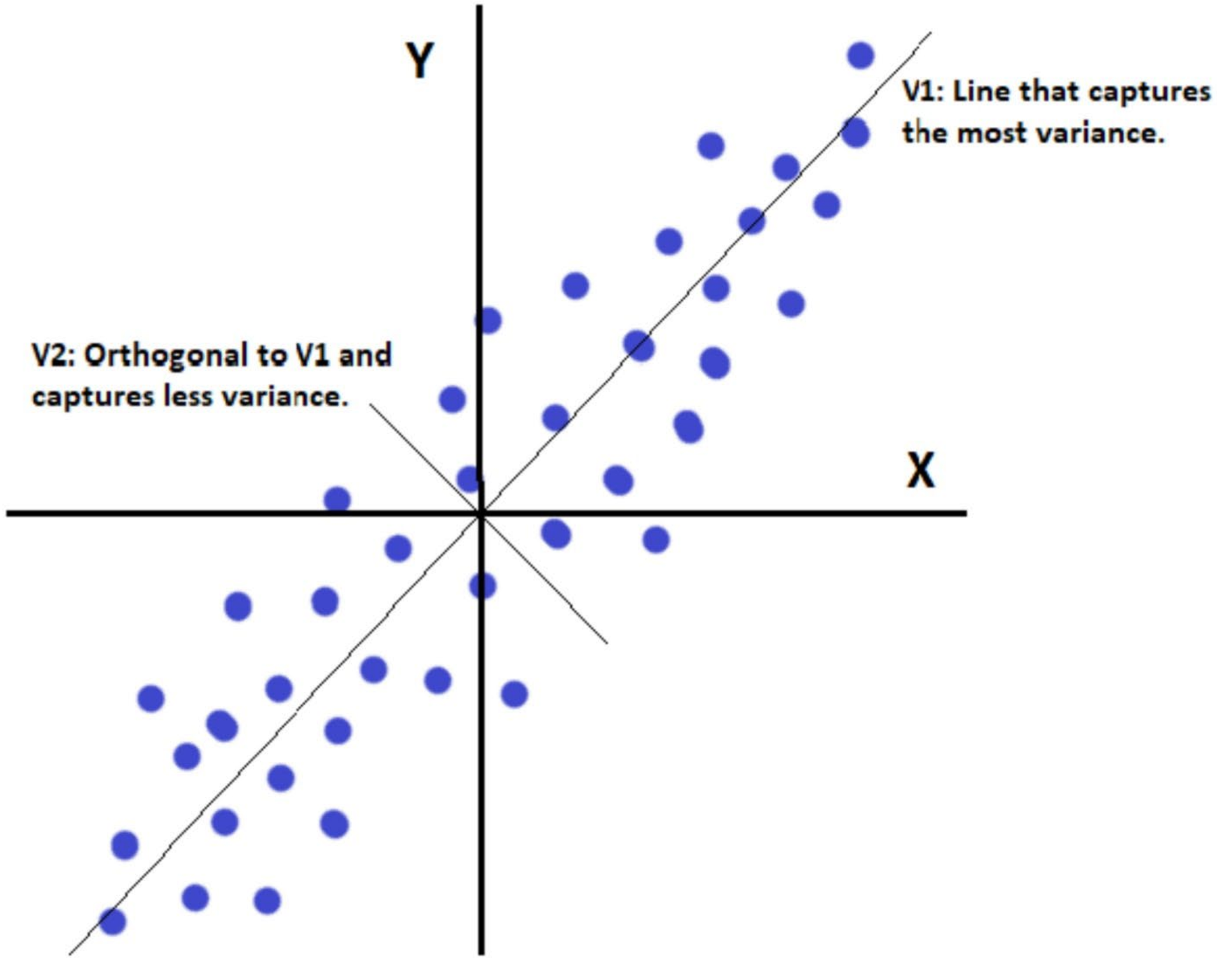
- common algorithms:
  - linear regression: predicts a continuous output using a linear relationship between input variables
  - logistic regression: predicts probabilities of classes, used for binary classification problems
  - decision trees: models decisions and their possible consequences as a tree structure
  - Support Vector Machines (SVM): finds the hyperplane that best separates the classes in the feature space
  - neural networks: layers of interconnected nodes that learn complex patterns in data
  - k-Nearest Neighbours (k-NN): classifies data points based on their proximity to other labelled data points
  - ensemble methods: combine multiple models to improve performance (e.g., Random Forest, Gradient Boosting)

https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644

Machine Learning

Supervised learning: feature extraction

- manual feature engineering: features are manually crafted based on domain knowledge and intuition about the data
  - example: in a credit scoring model, features might include income, credit history, and loan amount derived from raw financial data
- automated feature extraction: special feature selection algorithms can automatically extract and select the most relevant features from the data
  - example: Principal Component Analysis (PCA) can reduce the dimensionality of a dataset while preserving as much variability as possible

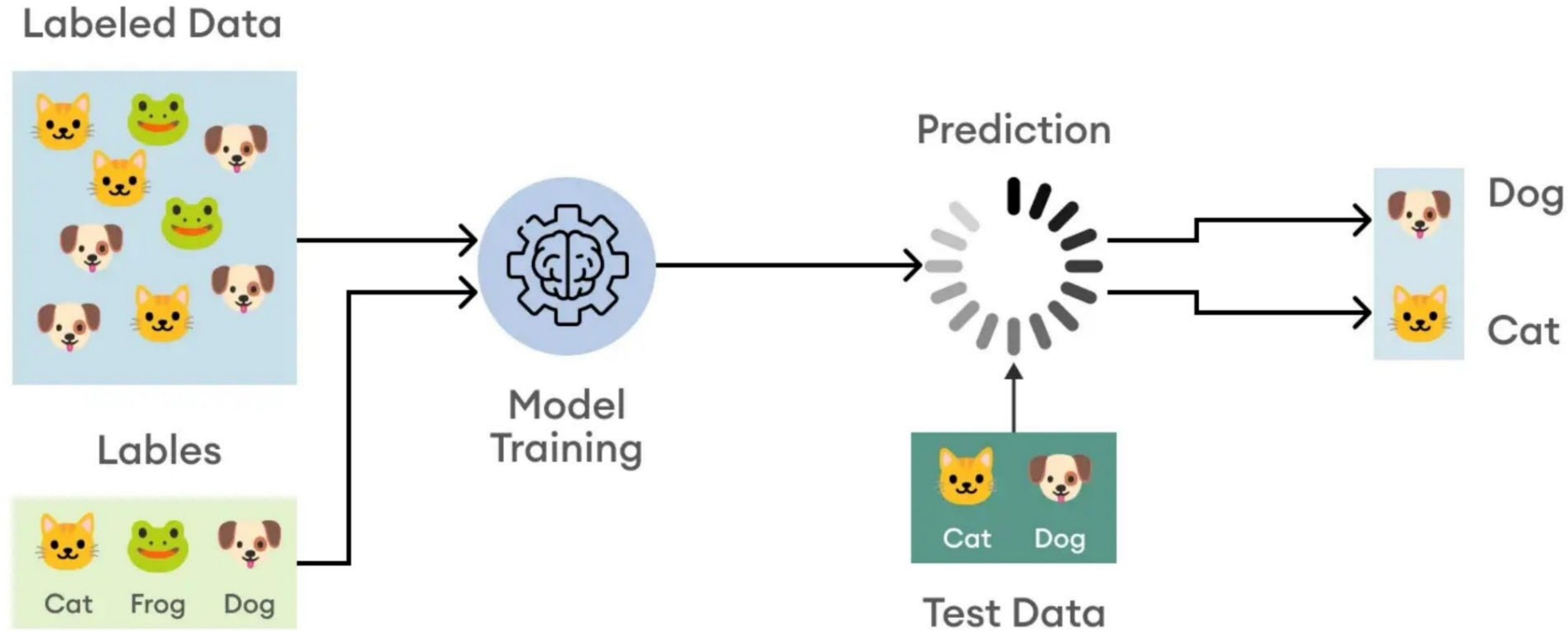


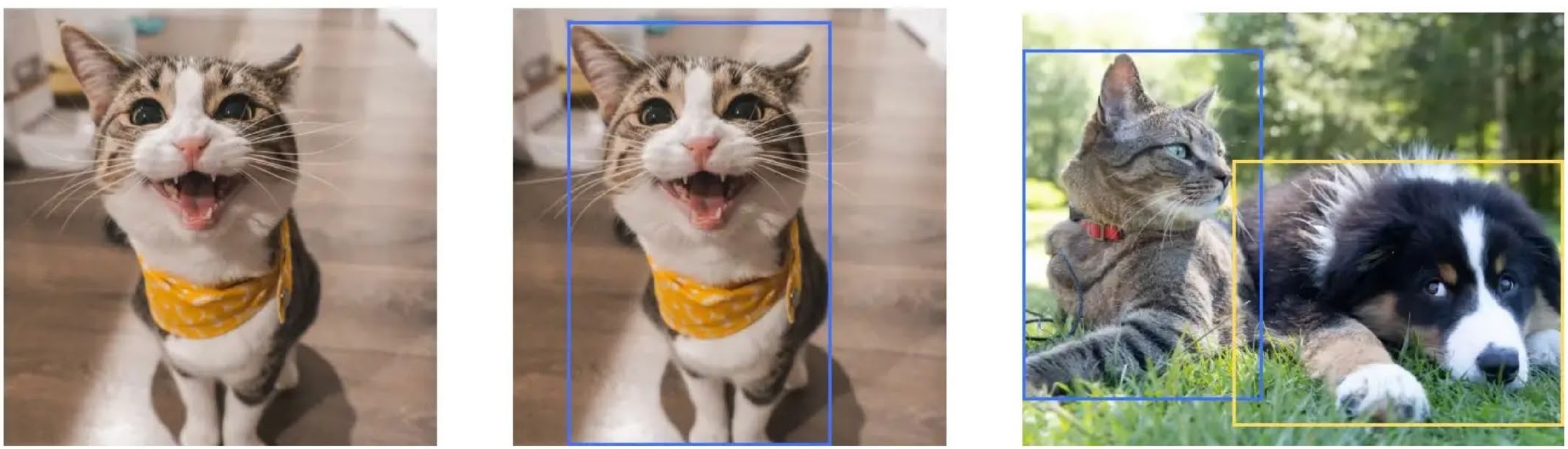
https://www.nature.com/articles/s41598-023-35846-6
https://statisticsbyjim.com/basics/principal-component-analysis/

Machine Learning

Supervised learning:

- applications:
  - image classification
  - speech recognition
  - spam detection
  - medical diagnosis





Classification Cat | **Classification, Localization** Cat | **Object Detection** Cat, Dog

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

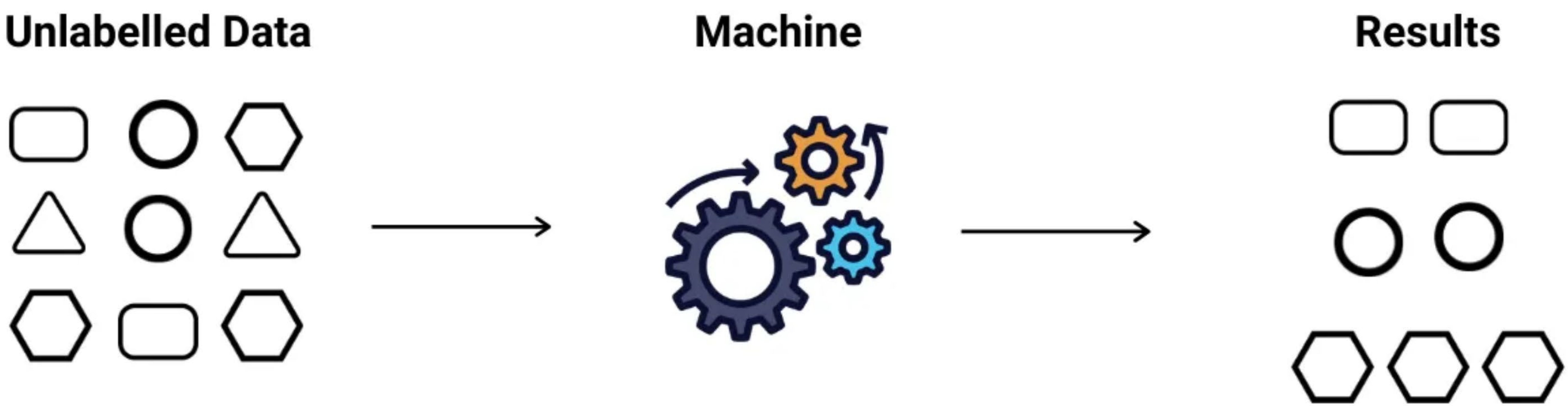
Questions

?

Machine Learning

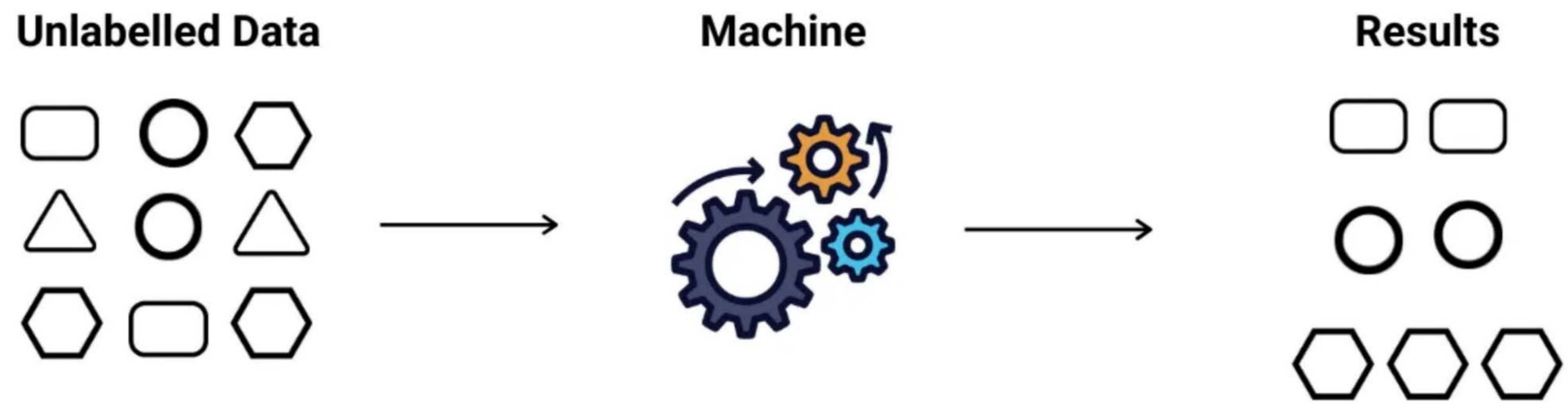
Unsupervised learning:

- trained on an unlabelled dataset
- infers the natural (intrinsic) structure present within a set of data points



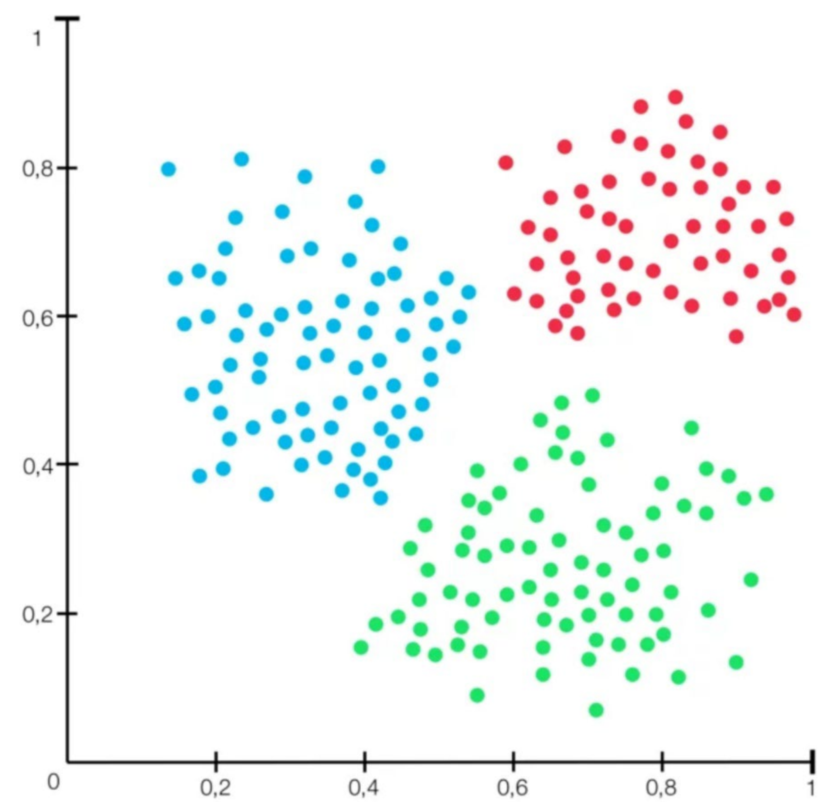
https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644

Machine Learning



Unsupervised learning:

- common algorithms:
  - clustering: groups data points into clusters based on similarity (example: “grouping your customers with similar characteristics”)
  - dimensionality reduction: reduces the number of variables under consideration
  - anomaly detection: Identifies unusual data points that do not fit the general pattern
  - association rule learning: discovers interesting relationships between variables in large databases (example: “customers who bought this also bought these”)
  - self-organising maps (Kohonen maps)

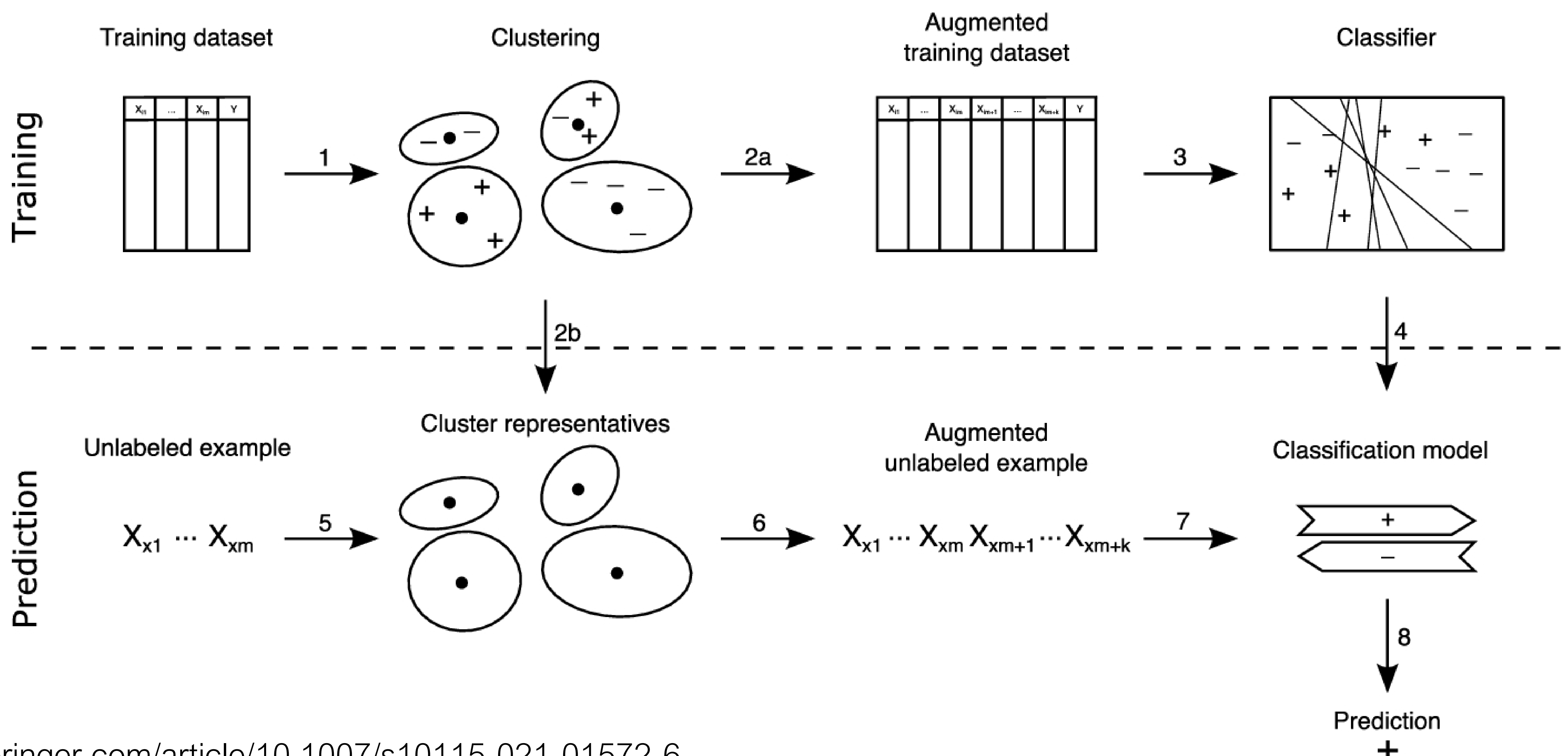


https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644
https://rocketloop.de/en/blog/clustering-machine-learning-comprehensive-guide/

Machine Learning

Unsupervised learning: feature extraction

- clustering algorithms benefit from feature extraction to group data points more effectively
  - example: extracting features such as colour histograms from images for clustering
- dimensionality reduction techniques: methods like PCA can be used to extract meaningful features that capture the essence of the data
  - example: high-dimensional data is transformed into a lower-dimensional space for easier clustering

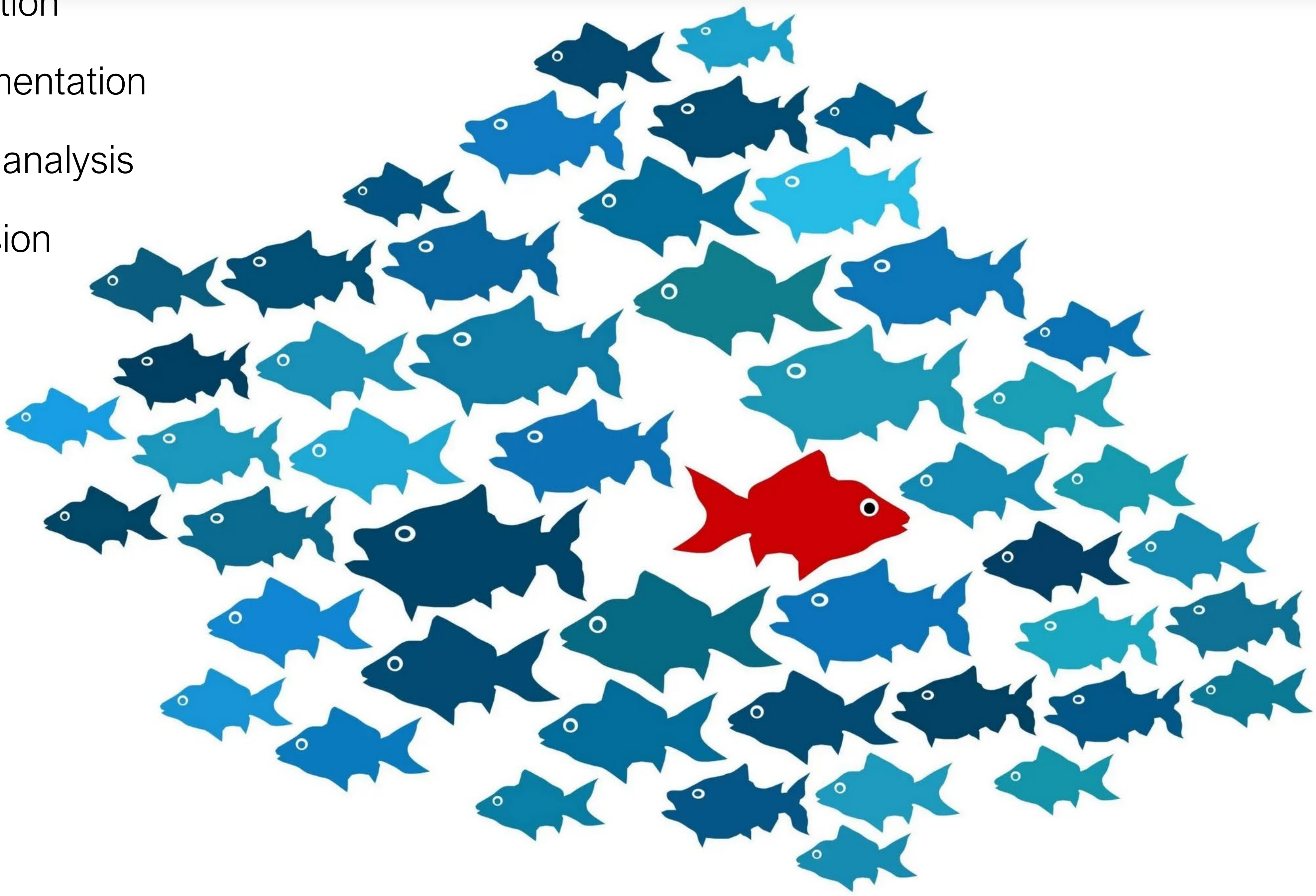


https://link.springer.com/article/10.1007/s10115-021-01572-6

Machine Learning

Unsupervised learning:

- applications:
  - anomaly detection
  - customer segmentation
  - market basket analysis
  - data compression



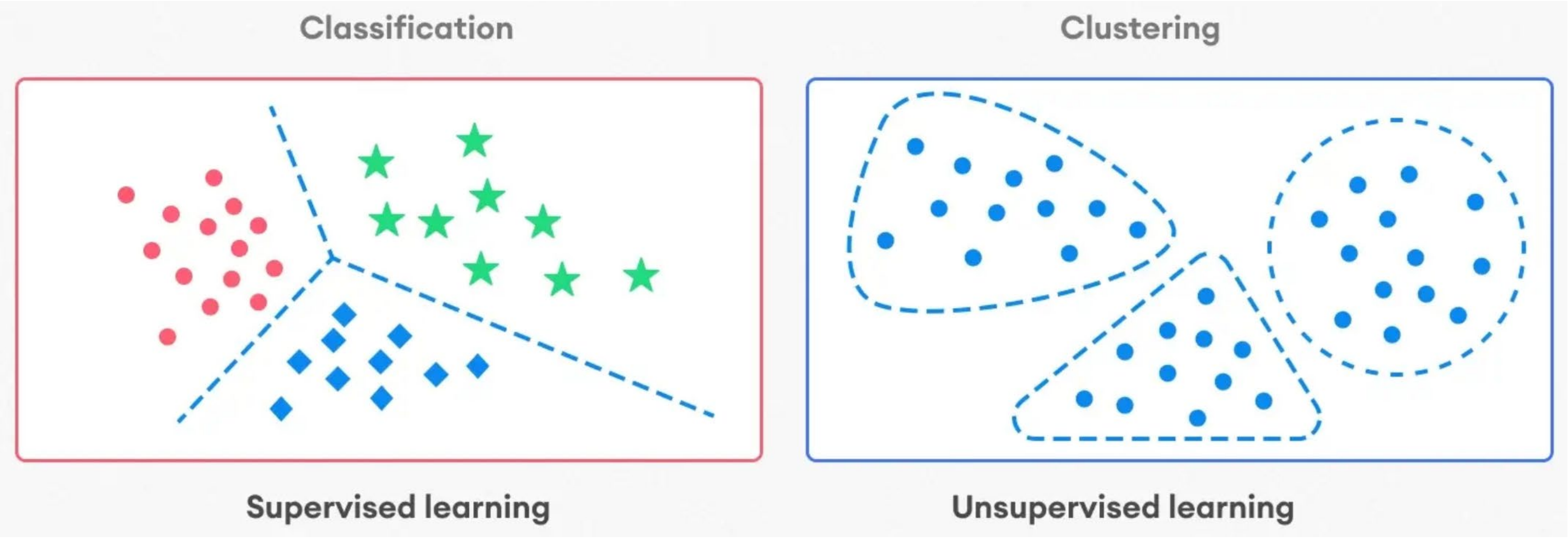
https://www.tatvic.com/blog/detecting-real-time-anomalies-using-r-google-analytics-360-data/

Questions

?

Machine Learning

Supervised vs unsupervised learning

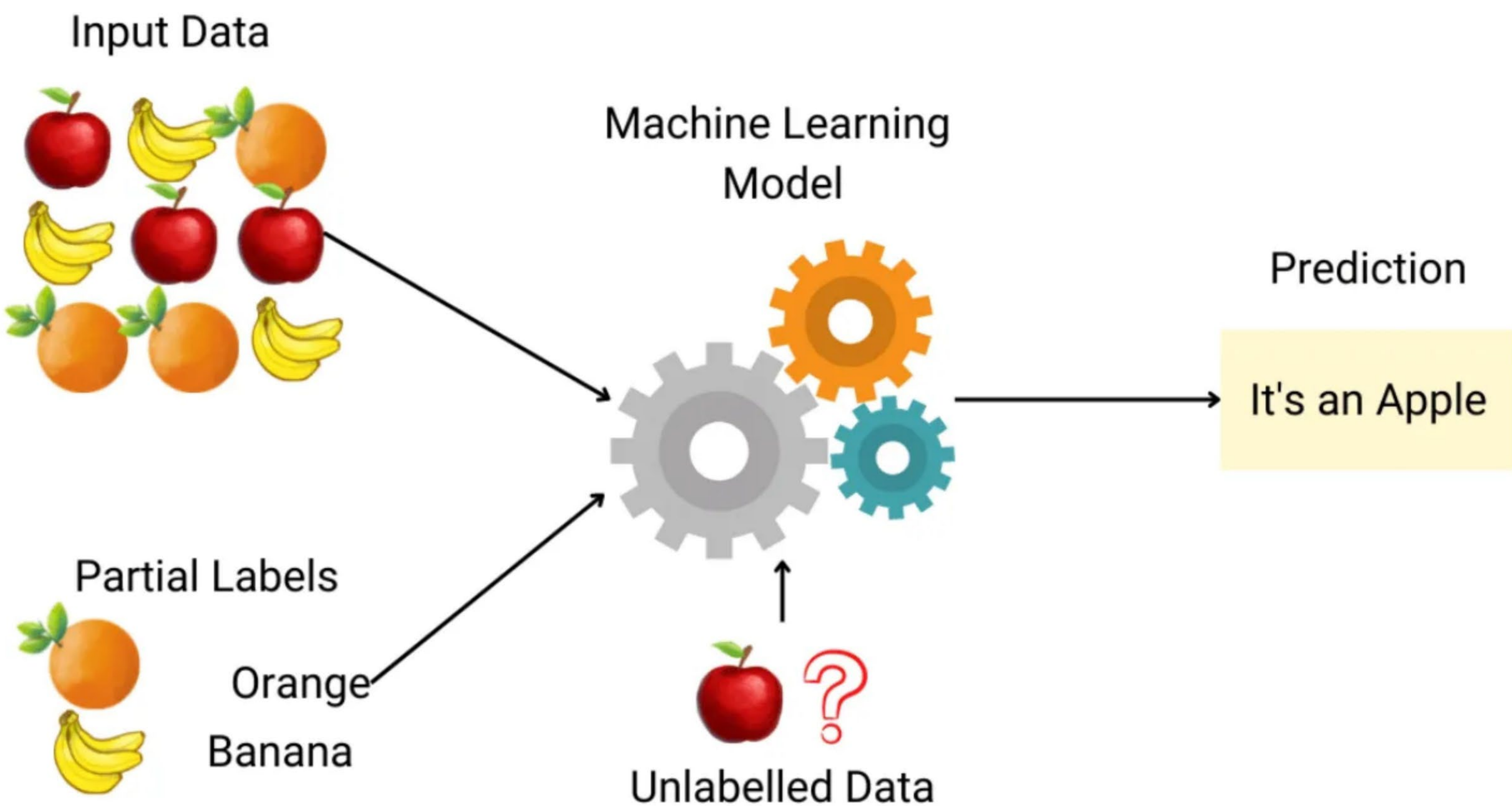


https://www.superannotate.com/blog/image-classification-basics

Machine Learning

Semi-supervised learning:

- uses a small amount of labelled data along with a large amount of unlabelled data
- labelled data helps guide the learning process, improving the model's accuracy when labelled data is scarce

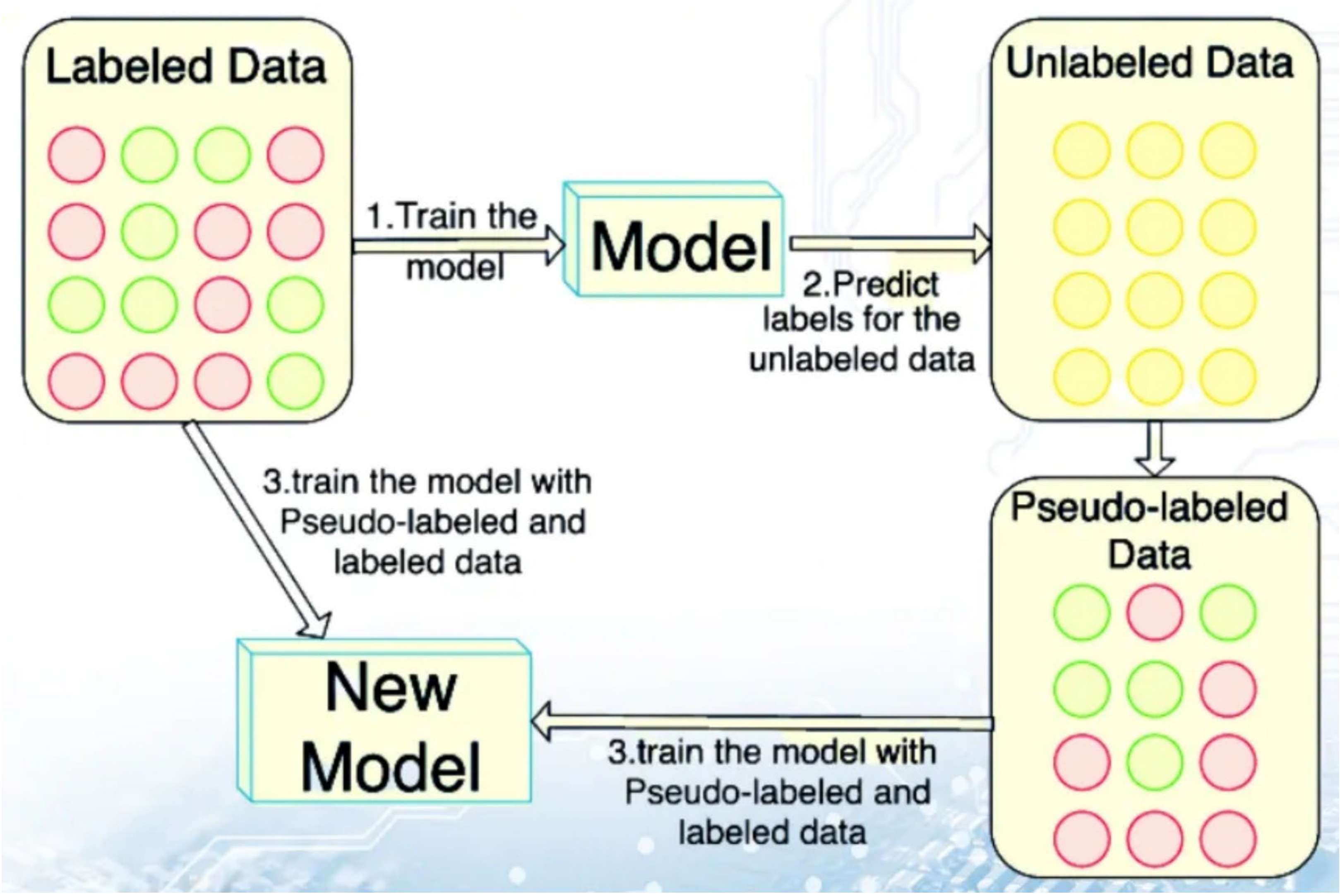


https://medium.com/@gayatri_sharma/a-gentle-introduction-to-semi-supervised-learning-7afa5539beea

Machine Learning

Semi-supervised learning:

- common algorithms:
  - self-training: a model trained on labelled data is used to predict labels for unlabelled data, which are then added to the training set
  - co-training: two models are trained on different views of the same data and help each other by labelling unlabelled data

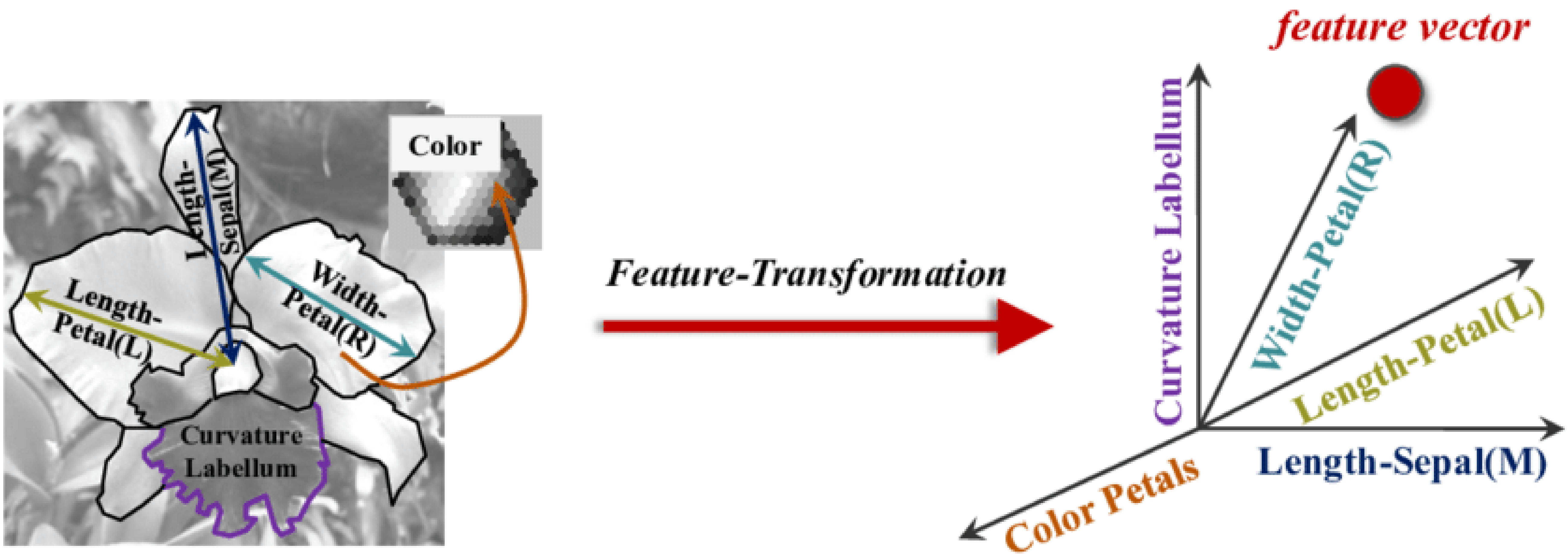


https://medium.com/@datasciencewizards/what-is-semi-supervised-learning-a-guide-for-beginners-a7452a597b8c

Machine Learning

Semi-supervised learning: feature extraction

- feature transformation: self-training or co-training often involve transforming features from unlabelled data to improve model learning
  - example: use labelled data to learn a good feature representation, and then apply it to label the unlabelled data, iteratively improving the feature set

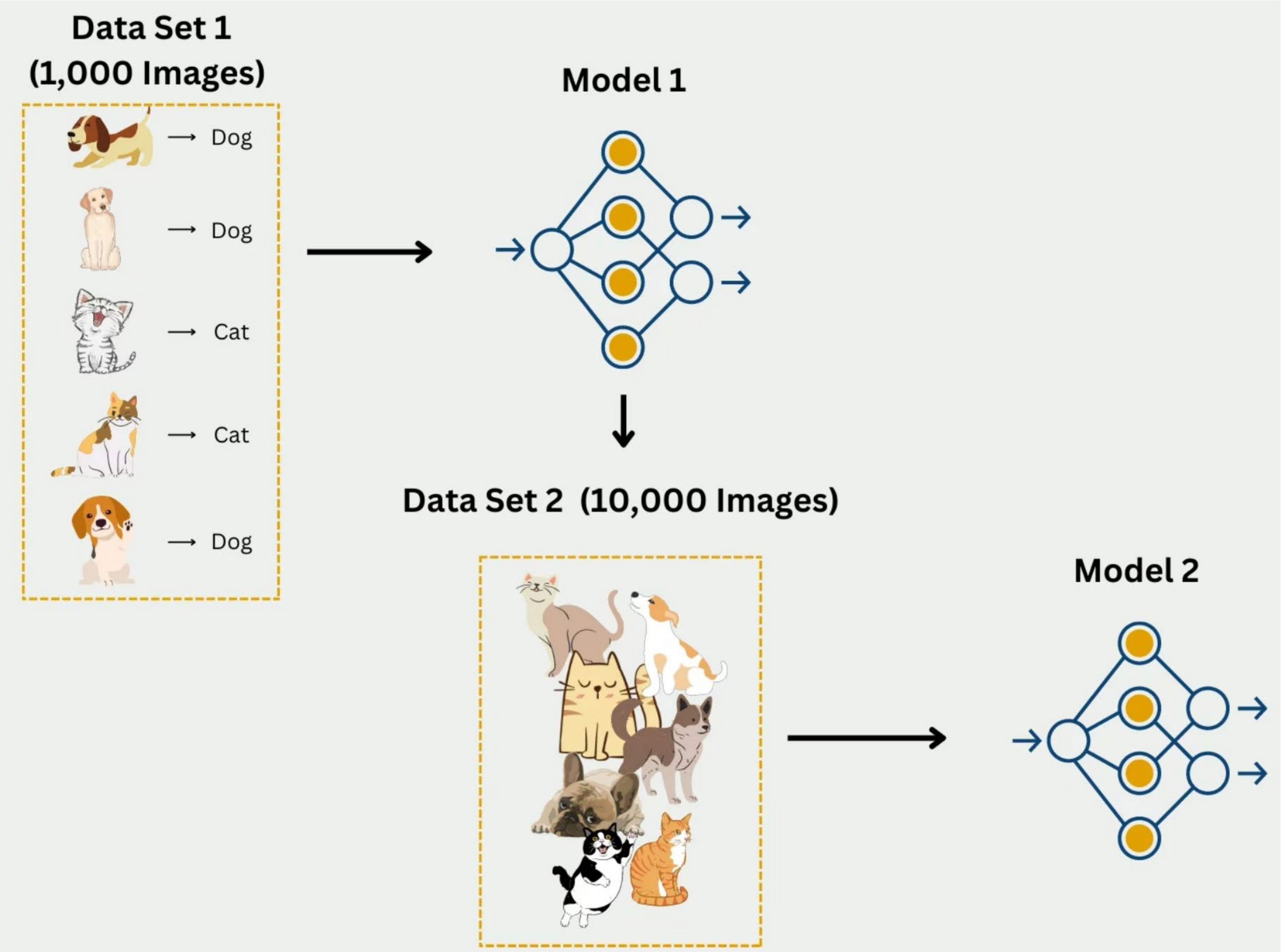


https://www.researchgate.net/figure/Example-for-feature-transformation-Five-features-are-measured-and-transformed-into-a_fig2_265664947

Machine Learning

Semi-supervised learning:

- applications:
  - text classification
  - image recognition
  - bioinformatics



https://databasecamp.de/en/ml/semi-supervised-learning-en

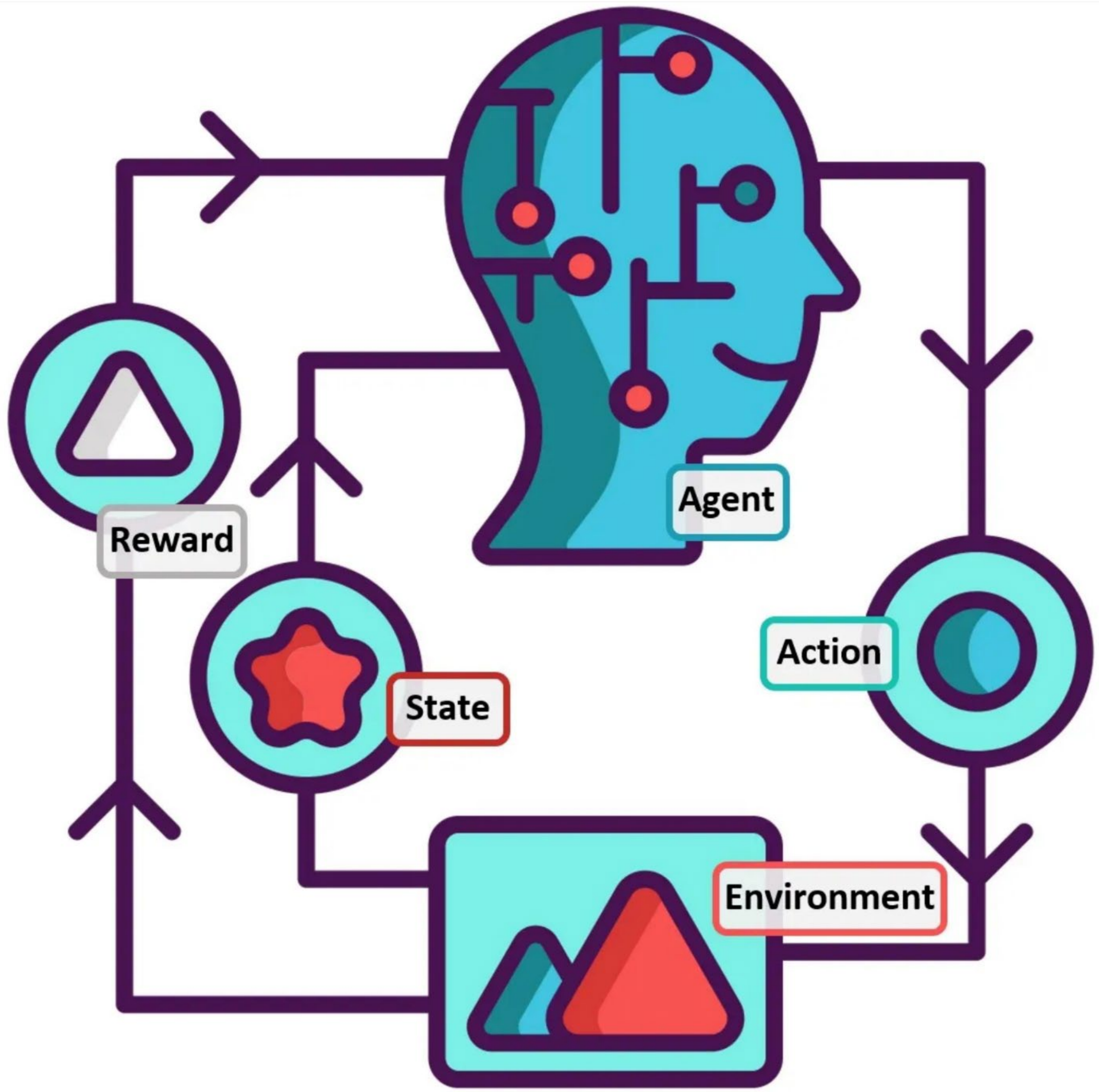
Break

?

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

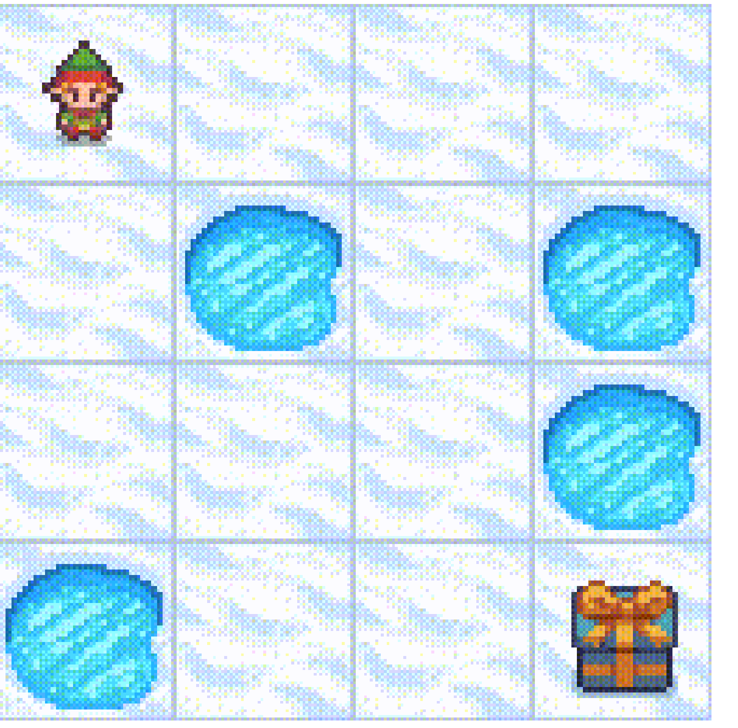


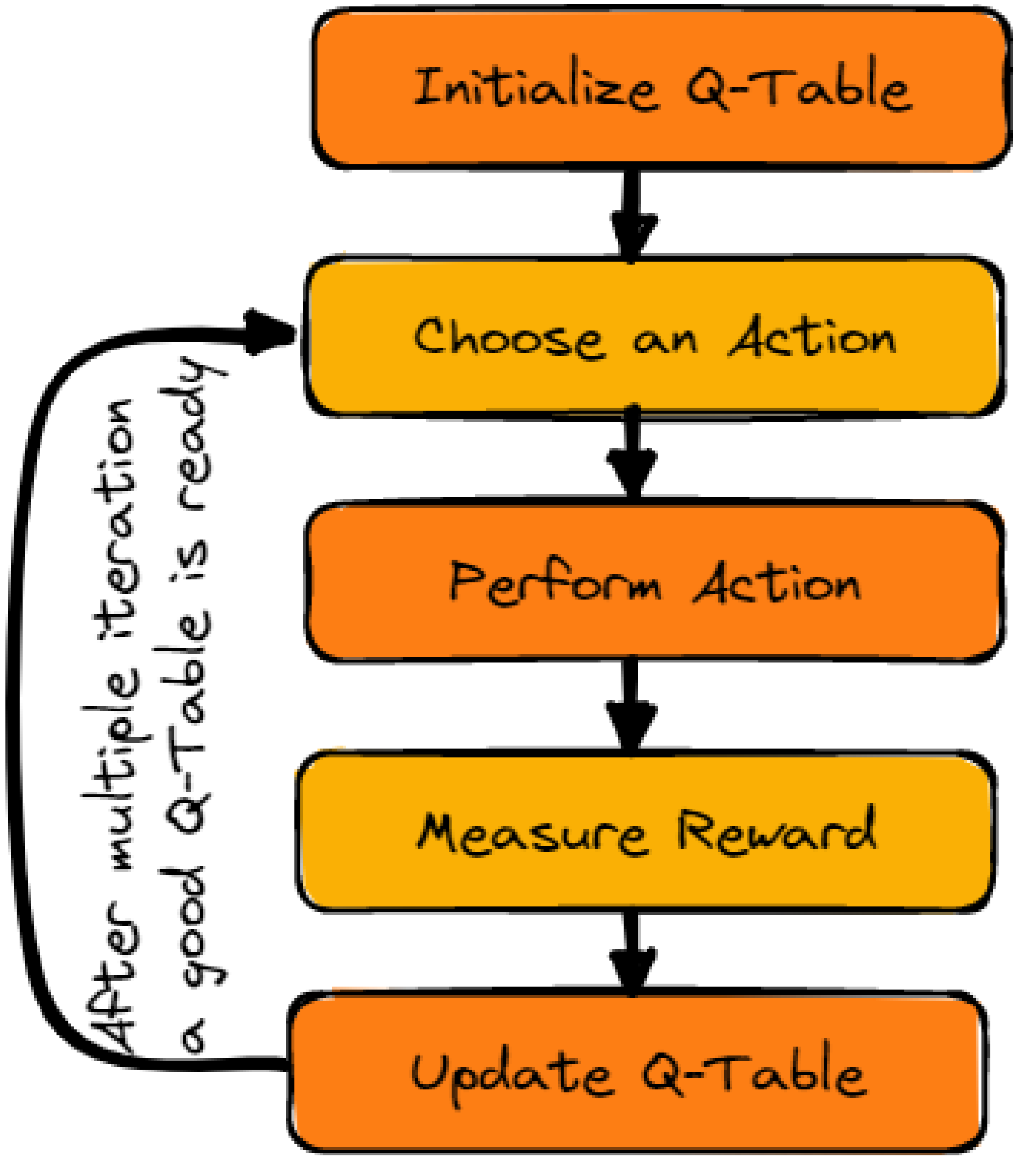
https://towardsdatascience.com/reinforcement-learning-fda8ff535bb6

Machine Learning

Reinforcement learning:

- common algorithms:
  - Q-Learning: a value-based method where the agent learns the value of actions in states to derive an optimal policy
  - deep Q-Networks (DQN): combines Q-learning with deep neural networks to handle high-dimensional state spaces
  - policy gradient methods: directly learn the policy that maps states to actions by optimising the expected reward
  - Actor-Critic methods: combine value-based and policy-based methods



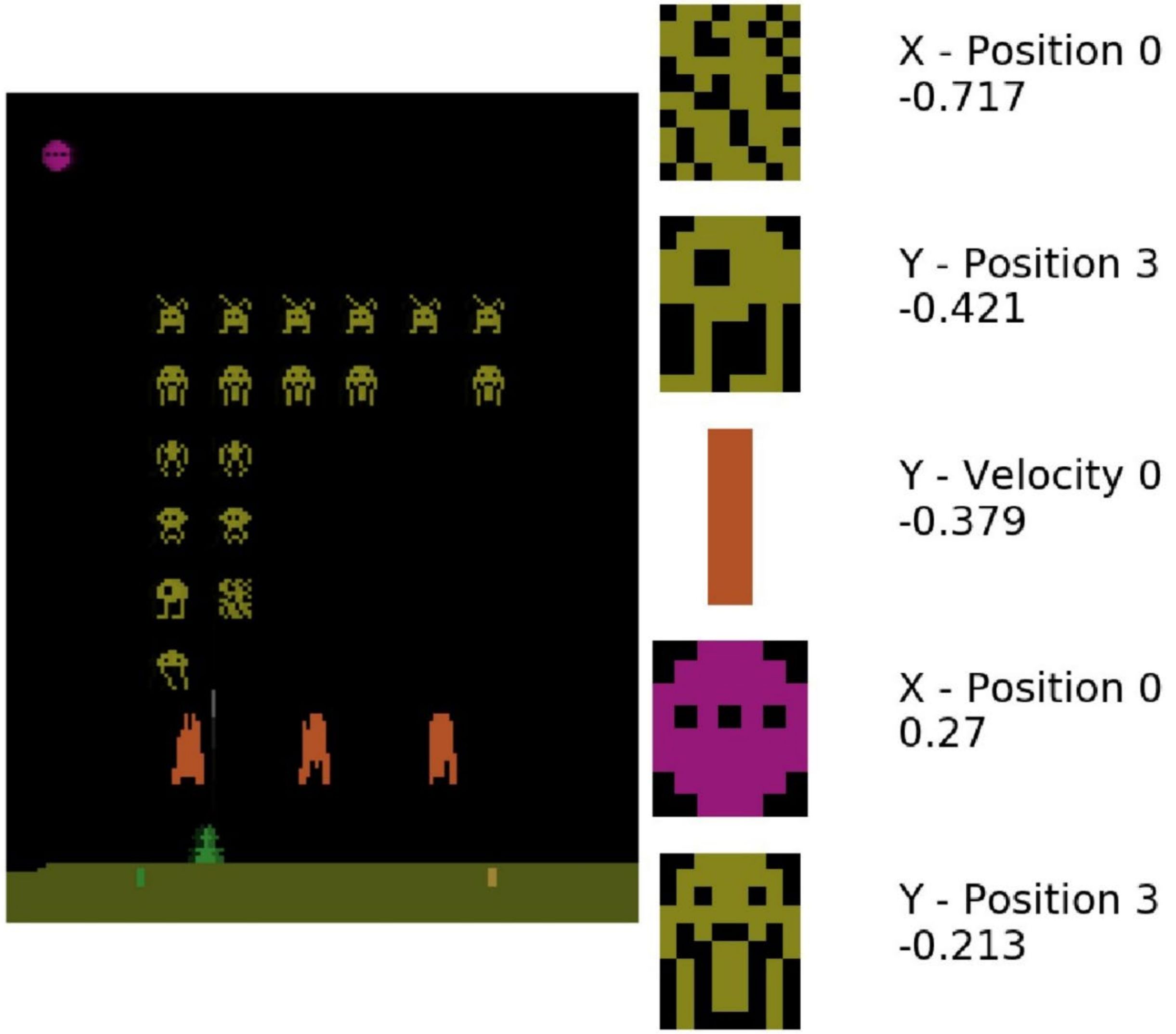


https://www.datacamp.com/tutorial/introduction-q-learning-beginner-tutorial

Machine Learning

Reinforcement learning: feature extraction

- state representation: extract feature to create a good representation of the state space
  - example: in a game-playing agent, features can include the positions of all pieces on the board and possible moves
- deep reinforcement learning: techniques like CNNs are used to automatically extract features from high-dimensional inputs such as images
  - example: a deep Q-network uses CNNs to extract features from raw pixel inputs in Atari games

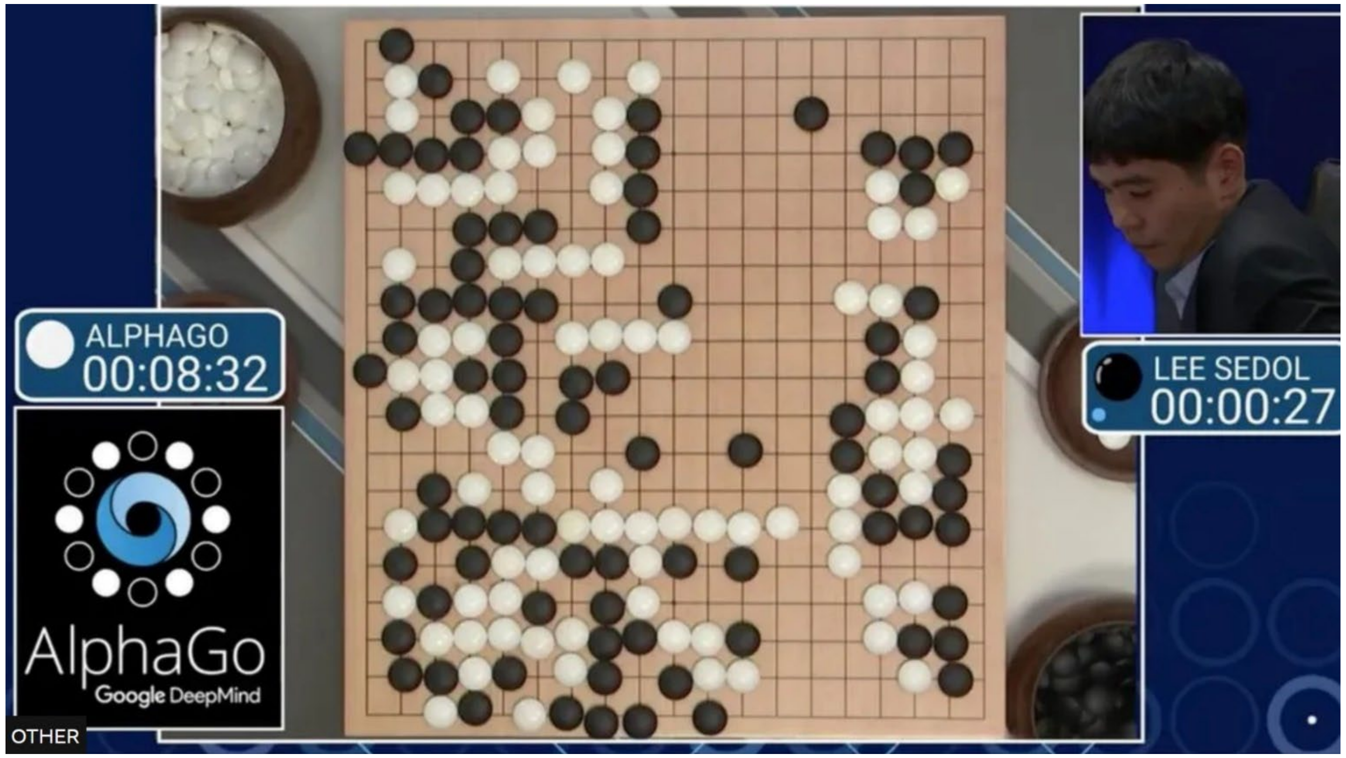


https://arxiv.org/pdf/2110.03184

Machine Learning

Reinforcement learning:

- applications:
  - game playing (example: AlphaGo)
  - robotic control
  - recommendation systems
  - autonomous driving



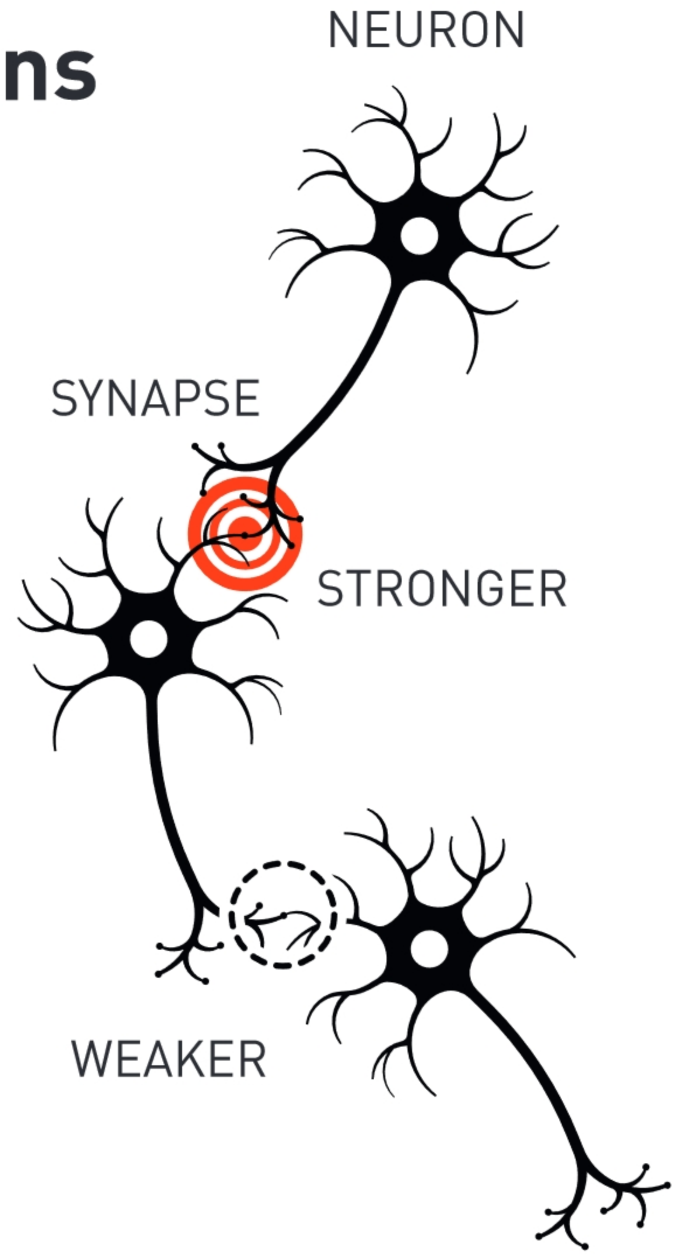
https://www.bbc.com/news/technology-35785875

Questions

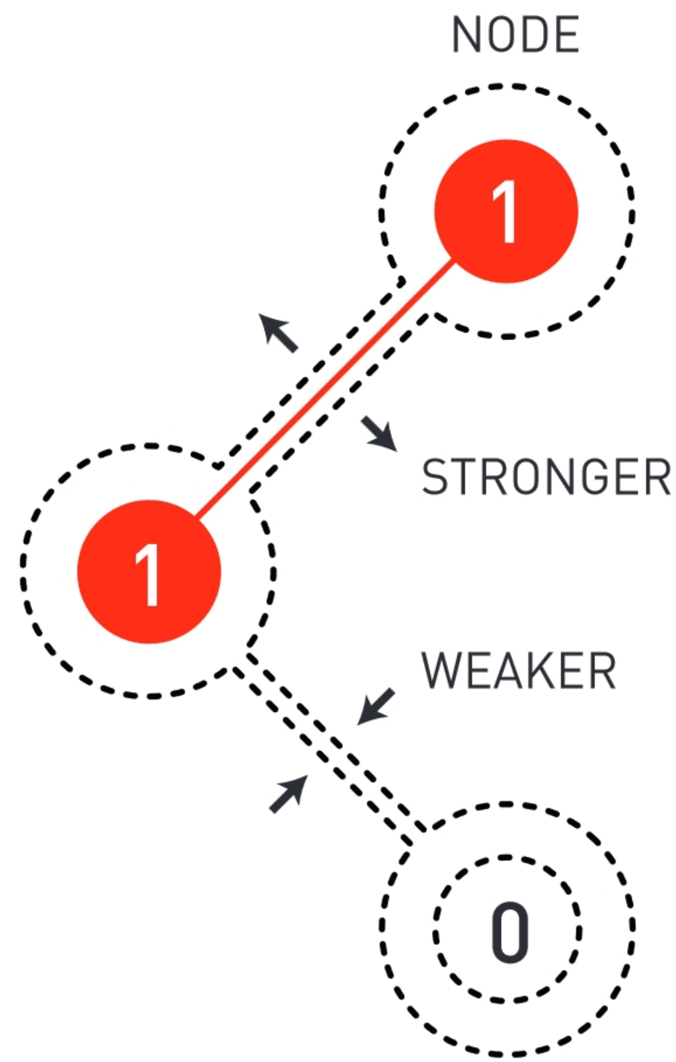
?

Natural and artificial neurons

The brain's neural network is built from living cells, neurons, with advanced internal machinery. They can send signals to each other through the synapses. When we learn things, the connections between some neurons gets stronger, while others get weaker.



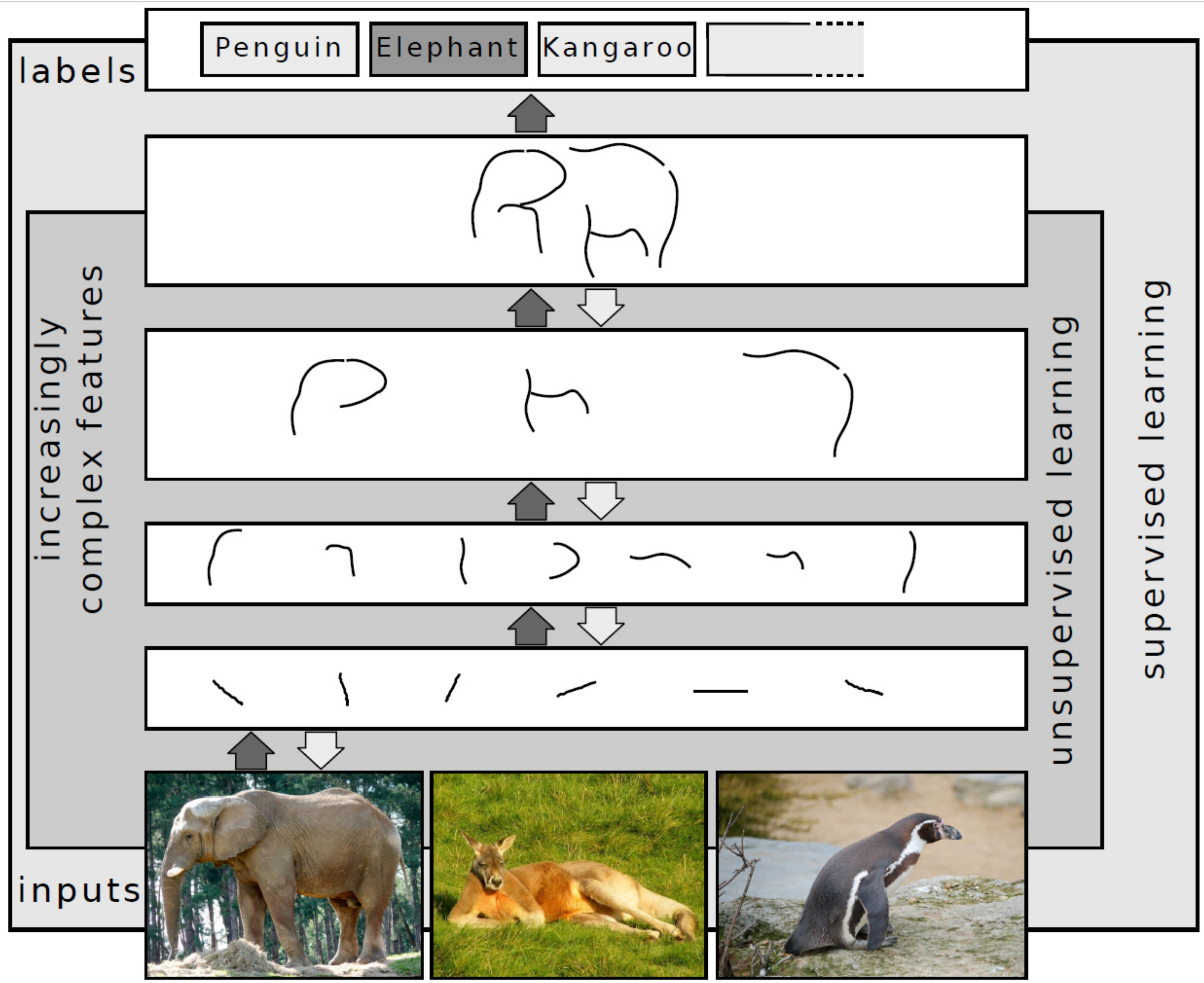
Artificial neural networks are built from nodes that are coded with a value. The nodes are connected to each other and, when the network is trained, the connections between nodes that are active at the same time get stronger, otherwise they get weaker.



Machine Learning

Deep learning:

- uses neural networks with many layers (deep neural networks) to learn abstract representations of data
- works for unstructured data like images, audio and text

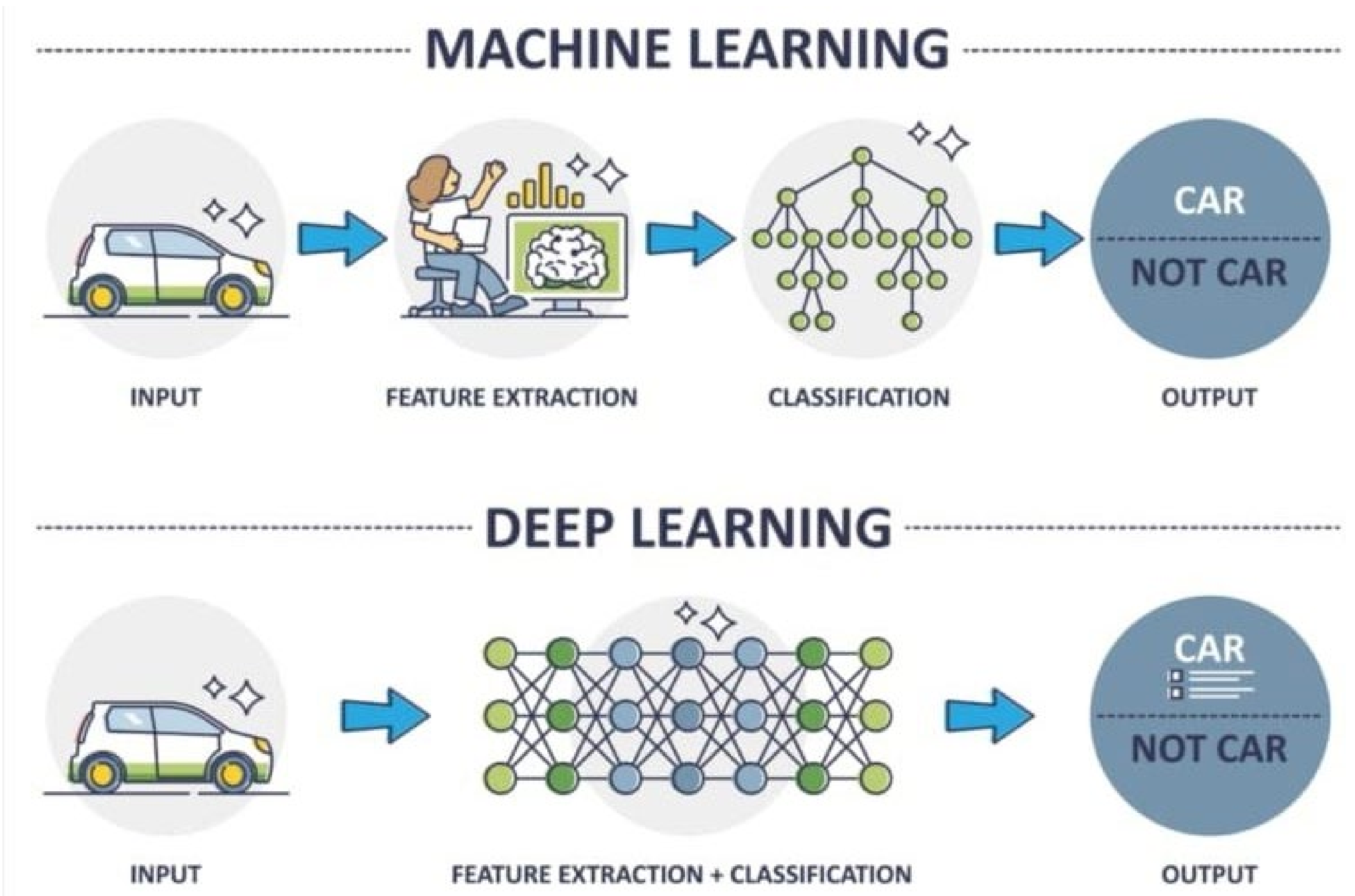


https://en.wikipedia.org/wiki/Deep_learning

Machine Learning

Deep learning: feature extraction

- automatic feature learning: learn hierarchical features directly from raw data through multiple layers of processing
  - example: in a CNN, the initial layers may learn to detect edges and textures, while deeper layers learn to recognise objects and more complex patterns



https://www.linkedin.com/posts/harinivas-sn_datascience-machinelearning-deeplearning-activity-7121358027458297859-BP6I

Machine Learning

High-level overview of deep learning:

1. Forward pass (make a guess): the input flows through the network layer by layer, producing an output (a prediction)

2. Error calculation (how wrong was the guess): The prediction is compared with the correct answer, and the difference (error) is measured

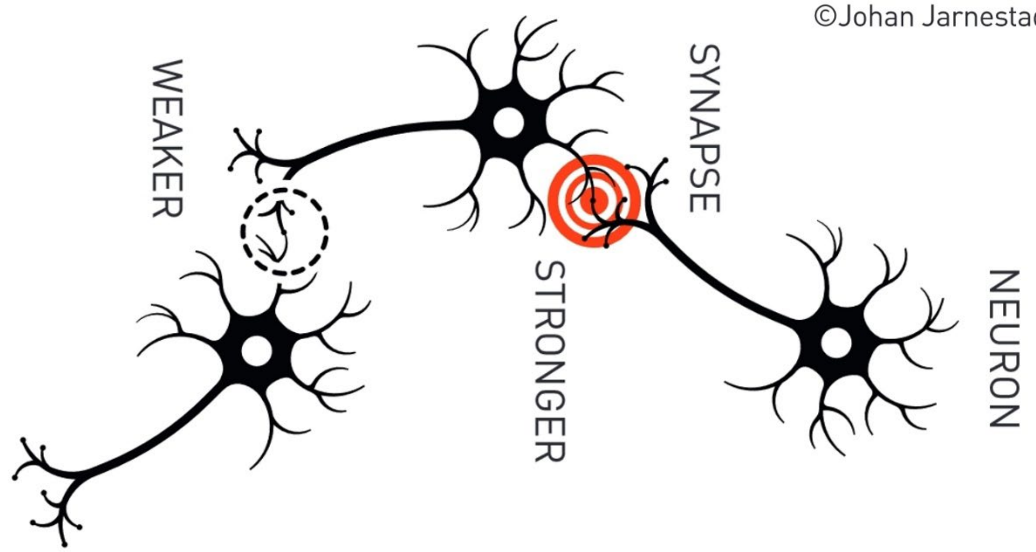
3. Backward pass – "backpropagation" (figure out why it was wrong): the error is sent backwards through the network, layer by layer, so that each neuron learns how much it contributed to the error

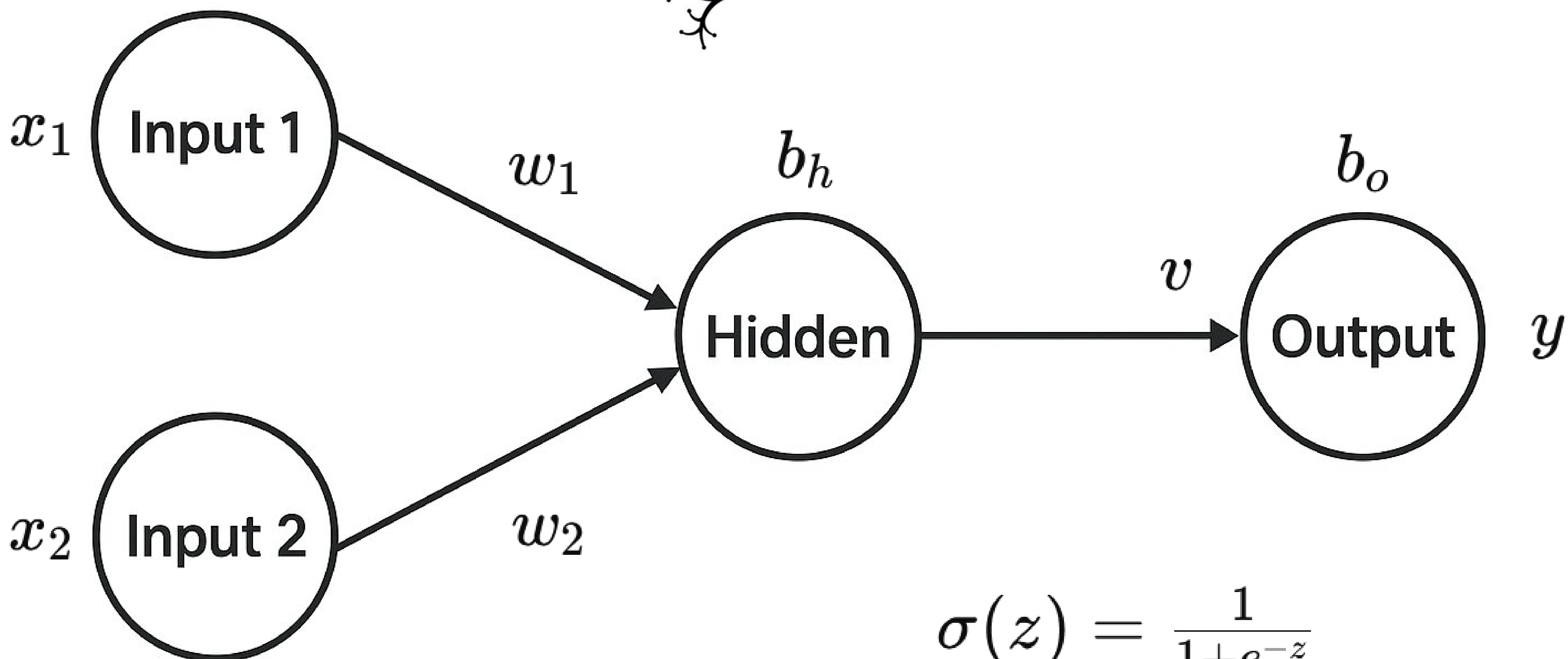
4. Weight updates (learn a little): using this information, the network slightly adjusts its internal "connection strengths" (weights) to reduce the error next time

5. Repeat (keep learning): this process is repeated many times on many examples, gradually improving the network's accuracy

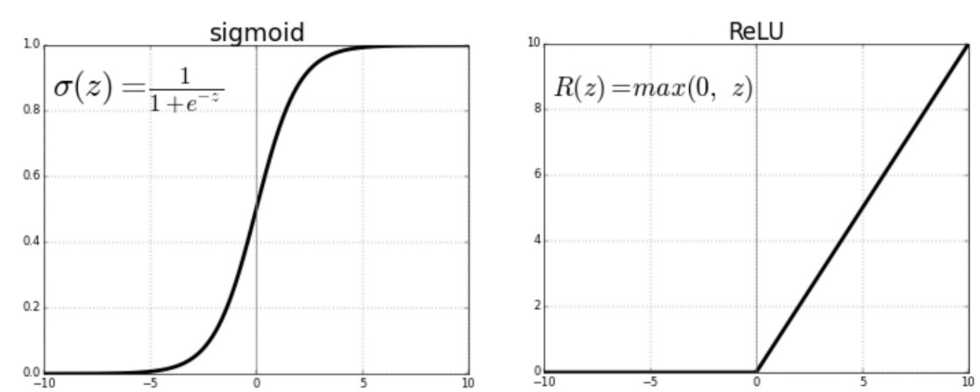
Machine Learning

Example





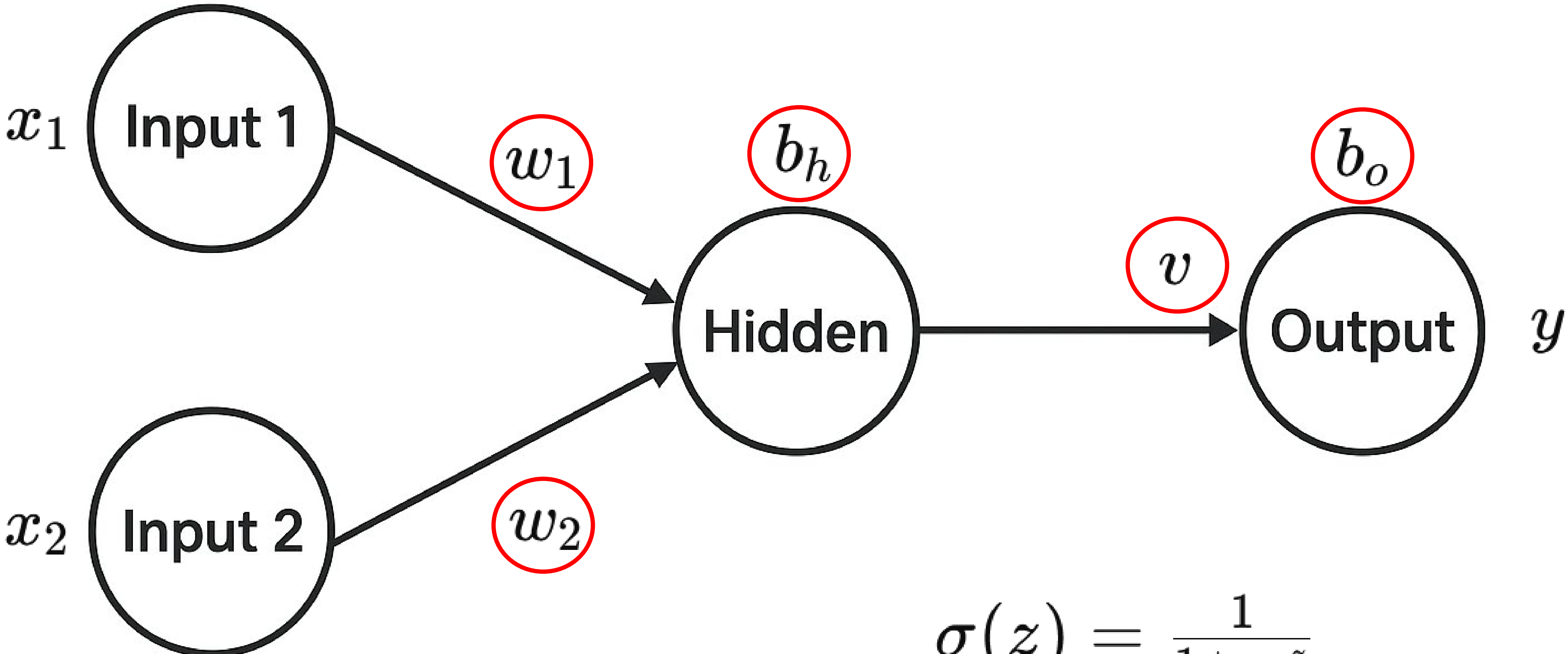
$$\sigma(z) = \frac{1}{1+e^{-z}}$$



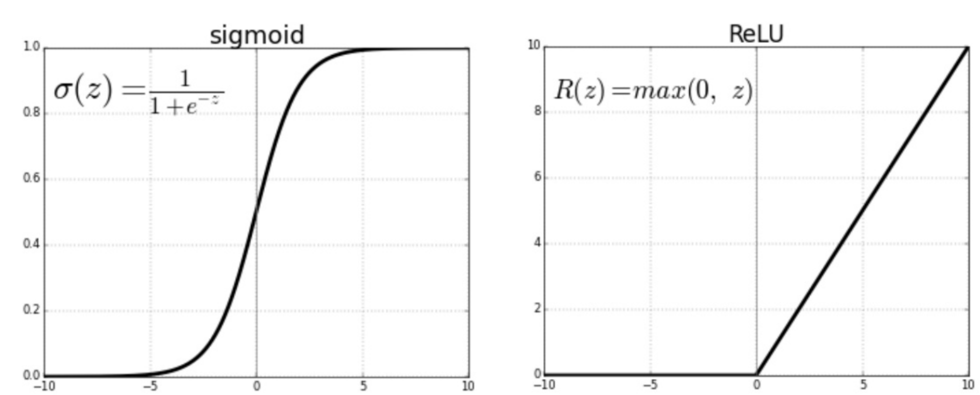
ChatGPT 5
https://towardsdatascience.com/activation-functions-in-neural-networks-how-to-choose-the-right-one-cb20414c04e5/

Machine Learning

Example

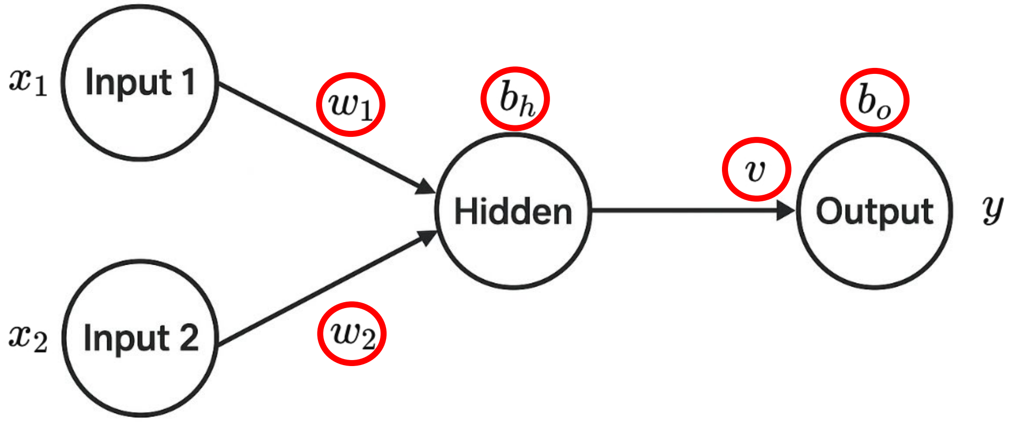


$$\sigma(z) = \frac{1}{1+e^{-z}}$$



ChatGPT 5
https://towardsdatascience.com/activation-functions-in-neural-networks-how-to-choose-the-right-one-cb20414c04e5/

Machine Learning



Example:

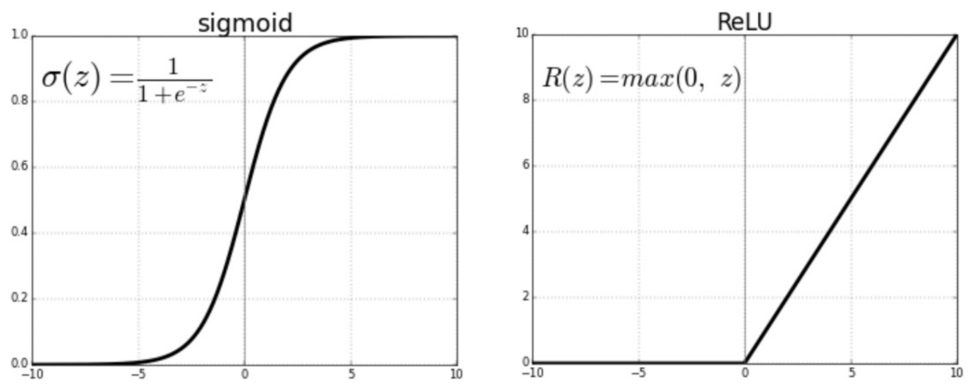
Setup

- Architecture: **2 inputs → 1 hidden neuron (sigmoid) → 1 output neuron (sigmoid)**
- Inputs: $x_1 = 1.0,\ x_2 = 0.5$
- Target: $y = 1.0$
- Loss: $L = \frac{1}{2}(y - \hat{y})^2$
- Learning rate: $\eta = 0.1$

Initial parameters

$$w_1 = 0.4,\ w_2 = -0.2,\ b_h = 0.1, \quad v = -0.3,\ b_o = 0.2$$

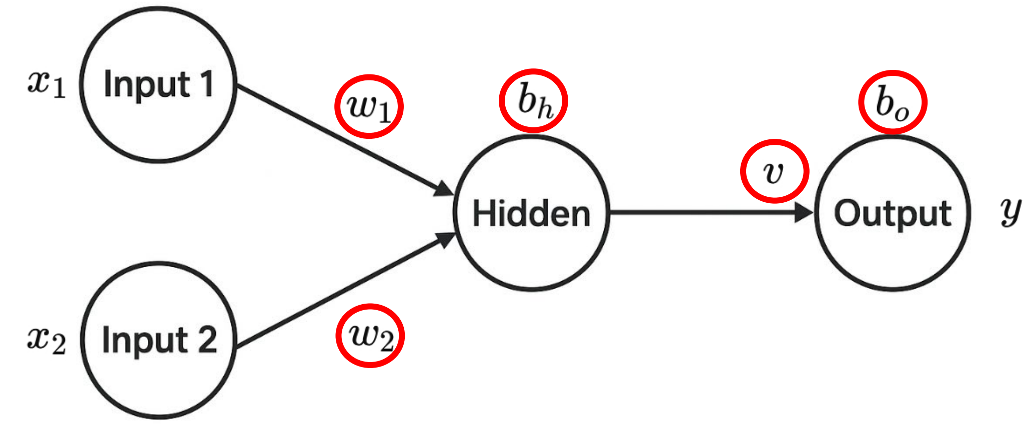
Sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$, $\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$.



ChatGPT 5

https://towardsdatascience.com/activation-functions-in-neural-networks-how-to-choose-the-right-one-cb20414c04e5/

Machine Learning



Example:

1) Forward pass (make a guess)

Hidden pre-activation:

$$z_h = w_1 x_1 + w_2 x_2 + b_h = 0.4(1) + (-0.2)(0.5) + 0.1 = 0.4$$

Hidden activation:

$$a_h = \sigma(z_h) = \sigma(0.4) = 0.5986876601$$

Output pre-activation:

$$z_o = v a_h + b_o = (-0.3)(0.5986876601) + 0.2 = 0.0203937020$$

Prediction:

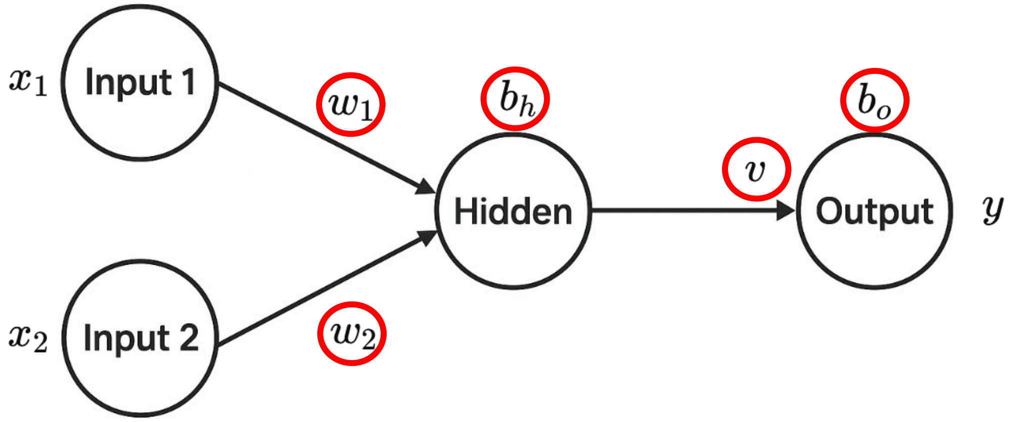
$$\hat{y} = \sigma(z_o) = 0.5050982488$$

2) Loss: (how wrong was the guess)

$$L = \frac{1}{2}(1 - \hat{y})^2 = 0.1224638717$$

ChatGPT 5

Machine Learning



Example:

3) **Backward pass (gradients)** (why the guess was wrong: how much the loss would change if the weight was adjusted slightly)

For the output:

$$\frac{\partial L}{\partial \hat{y}} = \hat{y} - y = -0.4949017512, \quad \frac{\partial \hat{y}}{\partial z_o} = \hat{y}(1-\hat{y}) = 0.249974$$

$$\frac{\partial L}{\partial z_o} = (\hat{y} - y)\hat{y}(1-\hat{y}) = -0.1237125742$$

$$\frac{\partial L}{\partial v} = \frac{\partial L}{\partial z_o} a_h = \underline{-0.0740651916}, \quad \frac{\partial L}{\partial b_o} = \frac{\partial L}{\partial z_o} = \underline{-0.1237125742}$$

Back to the hidden unit:

$$\frac{\partial L}{\partial a_h} = \frac{\partial L}{\partial z_o} v = 0.0371137723$$

$$\frac{\partial a_h}{\partial z_h} = a_h(1-a_h) = 0.2402607455$$

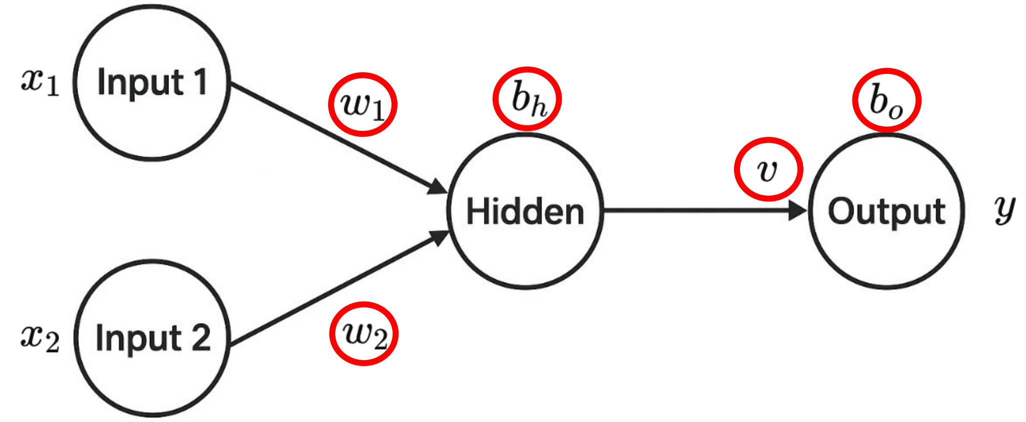
$$\frac{\partial L}{\partial z_h} = \frac{\partial L}{\partial a_h} a_h(1-a_h) = 0.0089169826$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_h} x_1 = \underline{0.0089169826}, \quad \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_h} x_2 = \underline{0.0044584913}, \quad \frac{\partial L}{\partial b_h} = \frac{\partial L}{\partial z_h} = \underline{0.0089169826}$$

ChatGPT 5

Machine Learning

Example:



4) Parameter update (gradient descent) (learn a little)

$$\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta}$$

$$w_1 \leftarrow 0.4 - 0.1(0.0089169826) = \mathbf{0.3991083017}$$

$$w_2 \leftarrow -0.2 - 0.1(0.0044584913) = \mathbf{-0.2004458491}$$

$$b_h \leftarrow 0.1 - 0.1(0.0089169826) = \mathbf{0.0991083017}$$

$$v \leftarrow -0.3 - 0.1(-0.0740651916) = \mathbf{-0.2925934808}$$

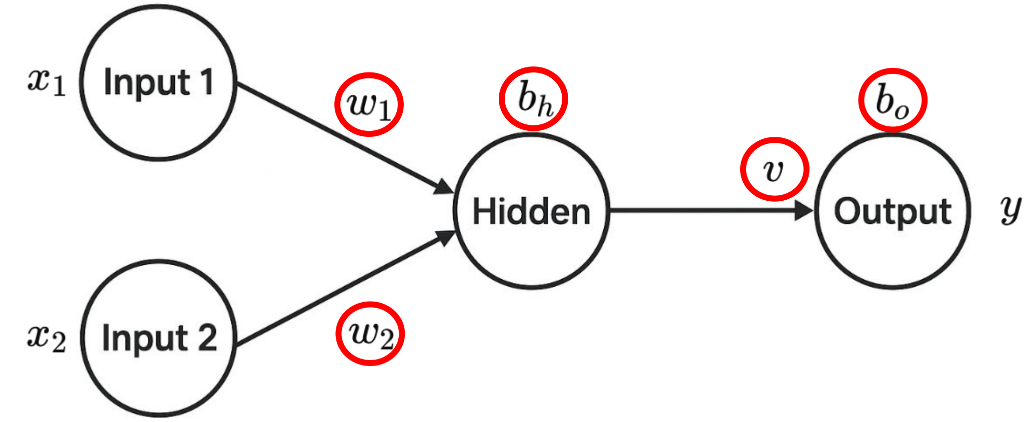
$$b_o \leftarrow 0.2 - 0.1(-0.1237125742) = \mathbf{0.2123712574}$$

the weights are adjusted to reduce the loss in the opposite direction of the gradient (steepest descent)

ChatGPT 5

Machine Learning

Example:



5) Forward pass again (after one update) (keep learning)

$$\hat{y}_{\text{new}} = 0.5093339707, \quad L_{\text{new}} = \frac{1}{2}(1 - \hat{y}_{\text{new}})^2 = \mathbf{0.1203765762}$$

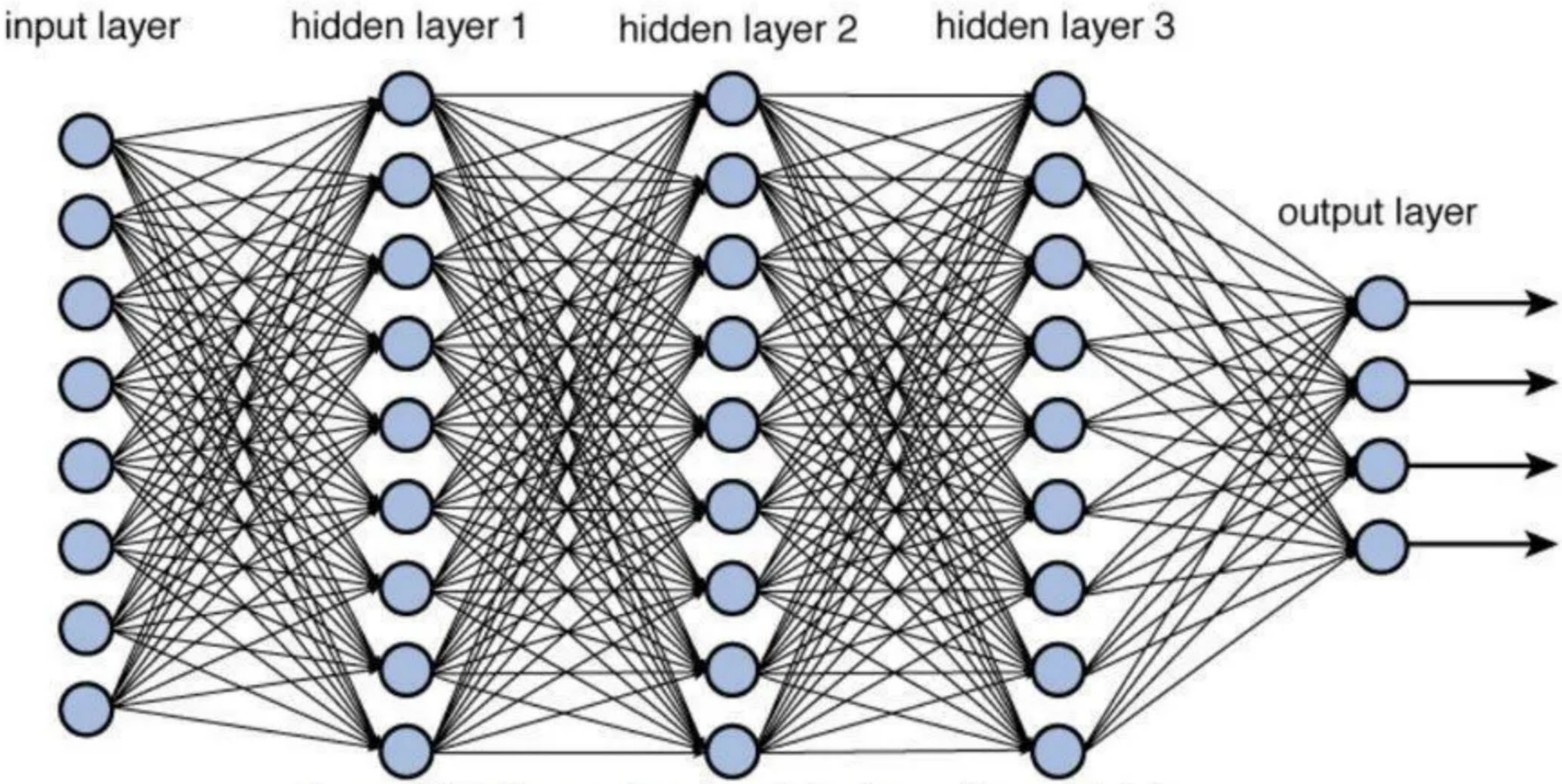
Loss decreased: $0.1224638717 \to 0.1203765762.$

the network moves toward a set of weights that minimises the loss and the model becomes more accurate

ChatGPT 5

Machine Learning

Deep learning:

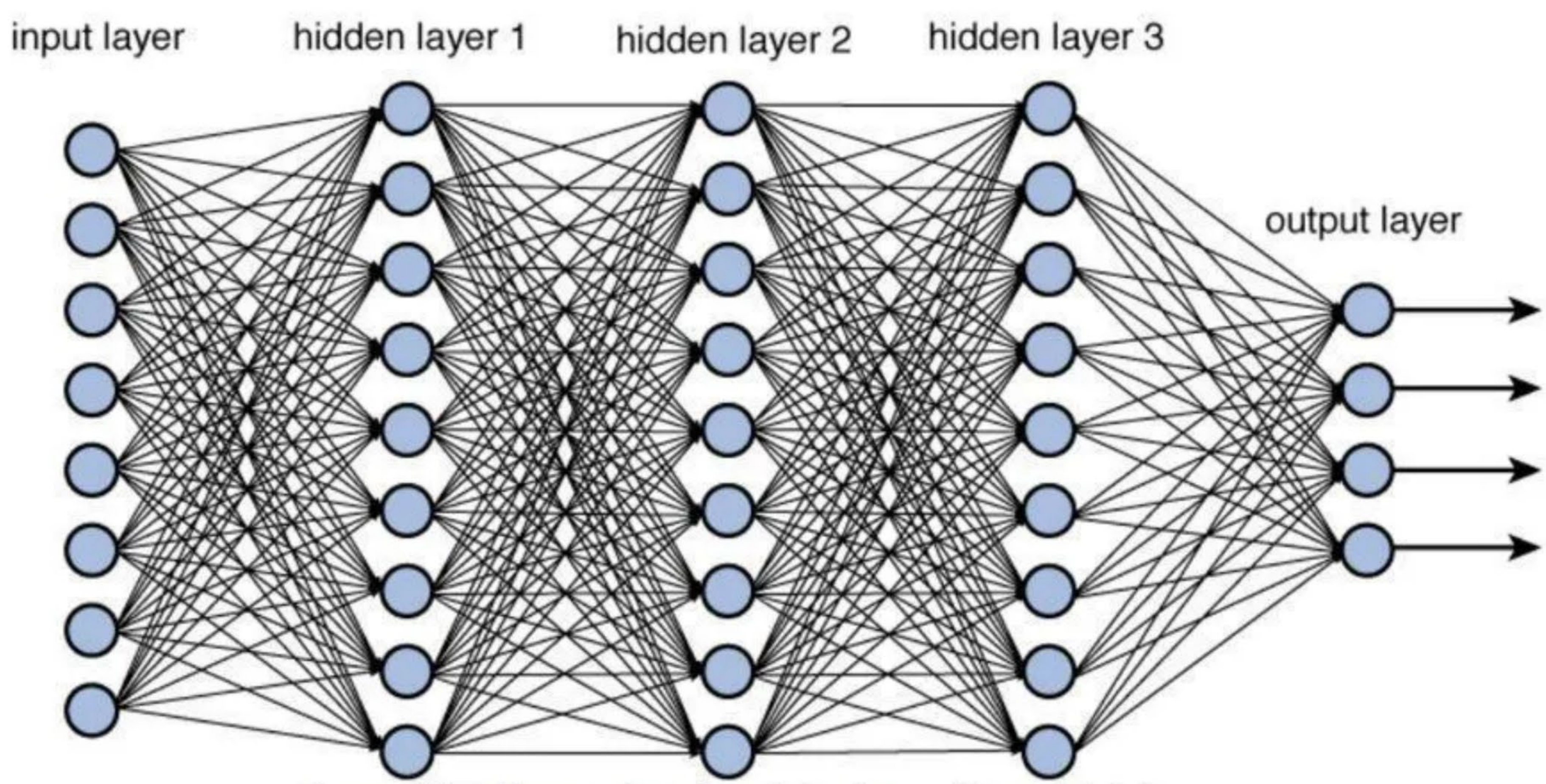


https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964

Machine Learning

Deep learning:

- common architectures:
  - Convolutional Neural Networks (CNNs): specialised for processing grid-like data (images)
  - Recurrent Neural Networks (RNNs): suitable for sequential data (time series or text)
  - Generative Adversarial Networks (GANs): contains two networks (generator and discriminator) that compete to improve data generation quality
  - Transformers: used primarily in natural language processing (NLP) for tasks like translation, summarisation, and question answering



https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964

Machine Learning

Generative Adversarial Networks (GANs) is like a game between two players:

- Generator (the actor)
  - tries to create fake data (like images, music or text) that look real
  - starts with random noise and transforms it into something that could pass as real
- Discriminator (the critic)
  - looks at data and decides: is this real (from the dataset) or fake (from the generator)?
  - its job is to catch the generator's fakes

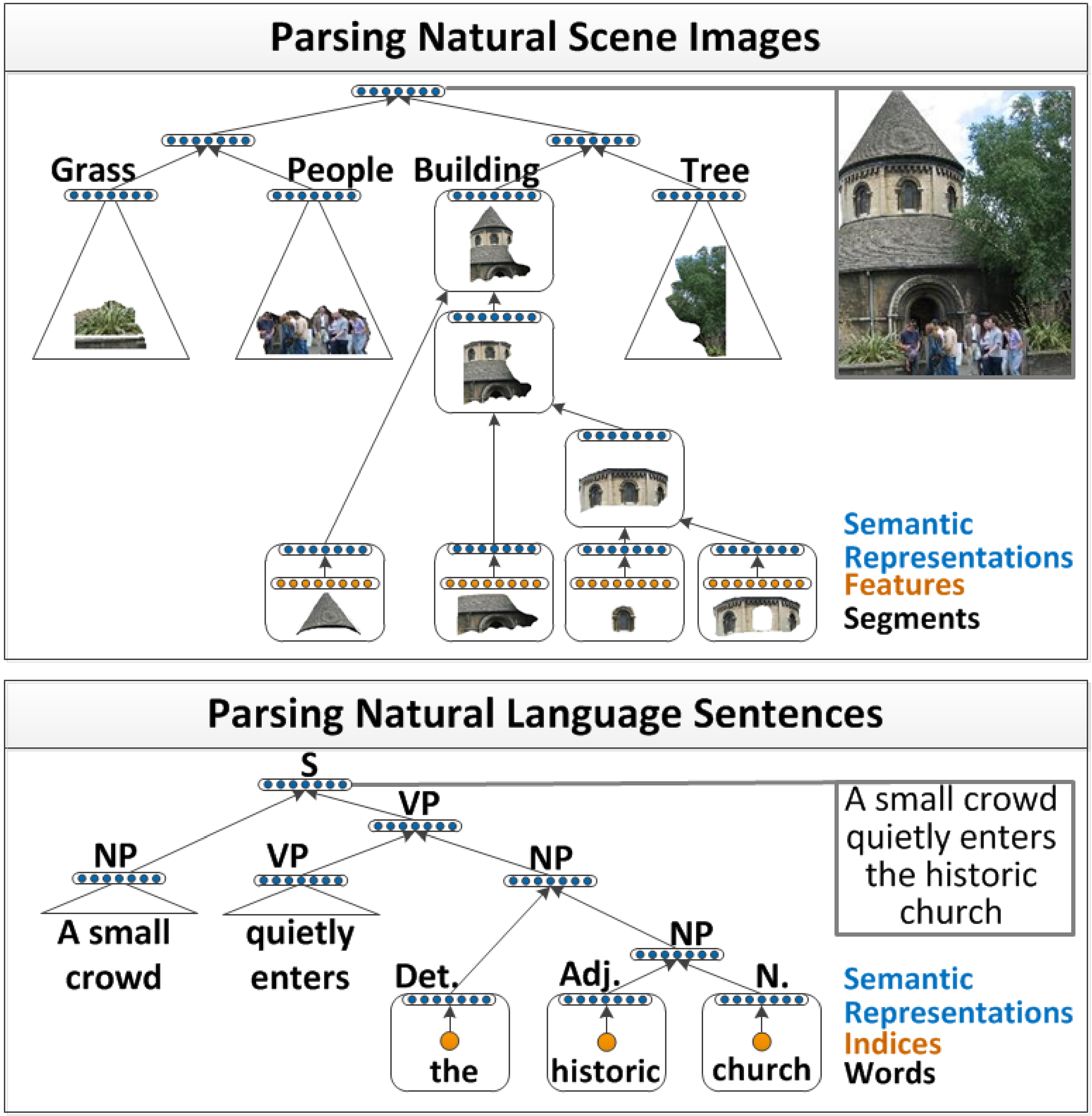
- The generator keeps improving because it wants to fool the discriminator
- The discriminator keeps improving because it wants to correctly spot fakes
- Over many rounds, they push each other to get better
- Eventually, the generator becomes so good that the discriminator can no longer tell the difference

ChatGPT 5

Machine Learning

Deep learning:

- applications:
  - Image and speech recognition
  - natural language processing
  - generative tasks



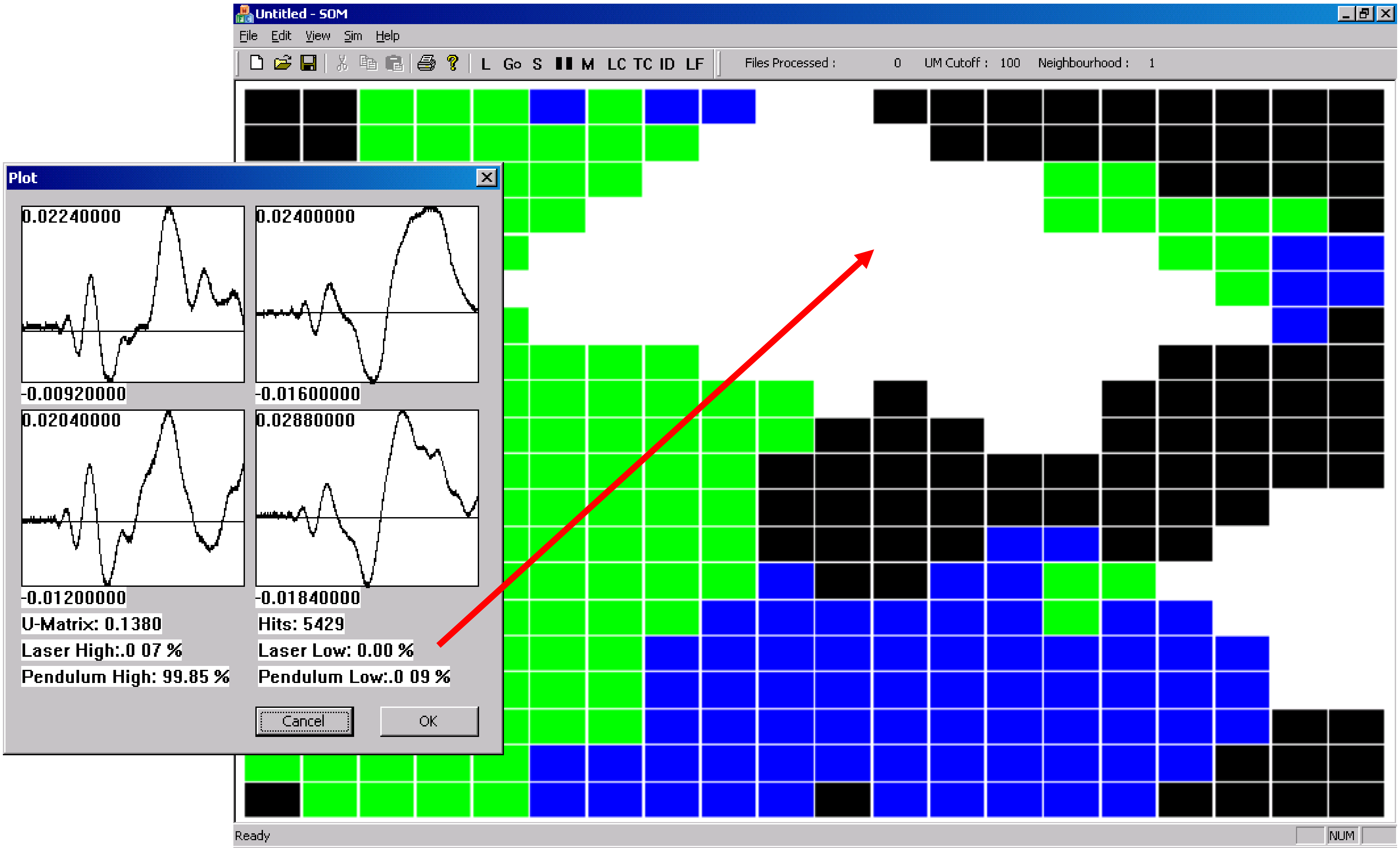
https://nlp.stanford.edu/projects/DeepLearningInNaturalLanguageProcessing.shtml

Questions

?

Self-organising (Kohonen) maps:

Signals in nearby clusters have more similar values than signals in distant clusters



Break

?

Machine Learning

Ethical concerns:

- bias and fairness
    - hiring algorithms may favour candidates from certain backgrounds due to biased training data
- security
    - adversarial examples may cause image recognition systems to misclassify objects
    - poisoning attacks where the training data is tampered with to degrade model performance
- accountability and transparency
    - autonomous vehicles may make split-second decisions without clear explanations, complicating accountability in the event of an accident
    - credit scoring models may deny loans without providing understandable reasons to applicant
- privacy
    - health data being used without proper consent to train predictive models
    - online behaviour data collected and used for targeted advertising without users' knowledge

ChatGPT 4o

Machine Learning

Ethical concerns:

- impact on employment
  - AI systems may replace jobs in manufacturing, retail, and customer service
  - automated decision-making systems may reduce the need for human involvement in various processes
- ethical use of data
  - using patient data for research without adequate consent
  - leveraging personal financial data for predictive modelling without users' explicit permission
- dual-use concerns
  - AI models used for cybersecurity can also be used for developing advanced cyber attacks
  - facial recognition technology can be used for security purposes but also for mass surveillance
- societal and cultural impact
  - social media algorithms may shape public opinion and spread misinformation
  - recommendation systems may reinforce echo chambers and filter bubbles

ChatGPT 4o

Machine Learning

Mitigation of ethical concerns:

- bias and fairness
  - using diverse and representative datasets
  - regularly auditing ML models for bias and implementing fairness-aware algorithms
- security
  - robust training and testing procedures to detect and defend against adversarial attacks
  - conducting thorough security assessments of ML models before deployment
  - regularly updating and patching ML systems to address security vulnerabilities
- accountability and transparency
  - developing explainable AI (XAI) techniques to make model decisions more interpretable
  - establishing clear accountability frameworks that define responsibility for ML system decisions
  - ensuring transparency in how data is collected, processed (algorithms) and used
- privacy
  - implementing data anonymisation and encryption techniques
  - ensuring compliance with data protection regulations
  - obtaining informed consent from users for data collection and use

ChatGPT 4o

Machine Learning

Mitigation of ethical concerns:

- impact on employment
  - investing in retraining and reskilling programs for workers displaced by automation
  - developing policies to support workers in transitioning to new roles and industries
  - ensuring that the economic benefits of AI and ML are distributed fairly
- ethical use of data
  - establishing clear guidelines and ethical standards for data use
  - ensuring that data usage is transparent and that individuals have control over their personal data
  - conducting ethical reviews of data usage in ML research and applications
- dual-use concerns
  - developing and adhering to ethical guidelines that govern dual-use technologies
  - engaging in public and policy discussions about the appropriate use of ML technologies
  - implementing strict controls and oversight for high-risk applications
- societal and cultural impact
  - designing ML systems that promote diversity of opinions and information
  - encouraging interdisciplinary research to understand and mitigate the societal impacts of ML

ChatGPT 4o

Discussion

?