# COMP9208
# Artificial Intelligence and Society

Lecture 5:  5 September 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 | |
| 07/11 | 13: Limits of cognition | |
| 17/11 | 14: STUVAC | # 3: Individual assignment (data analysis report) |

# Brief review of last week

?

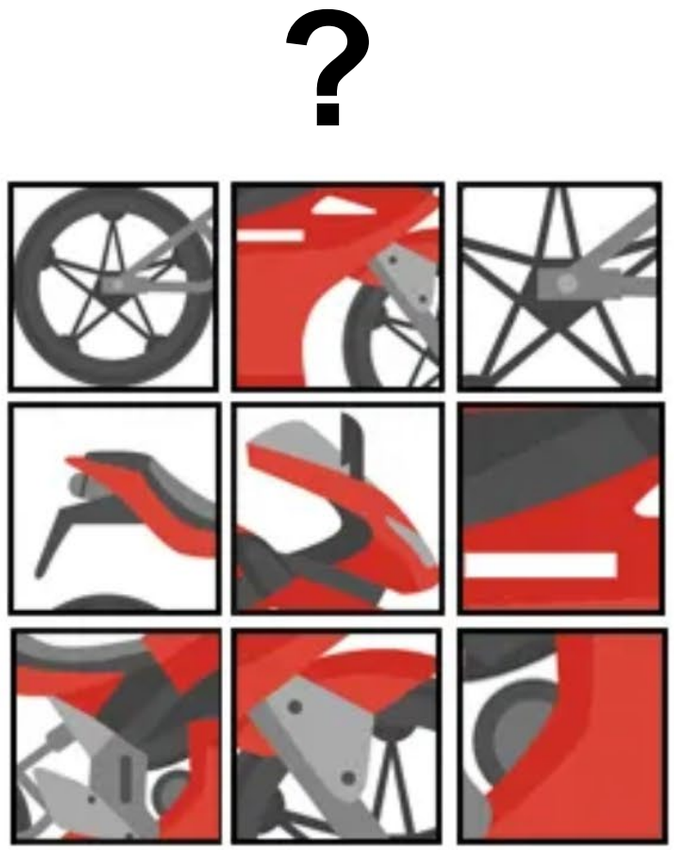
# Machine Learning

?

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

?

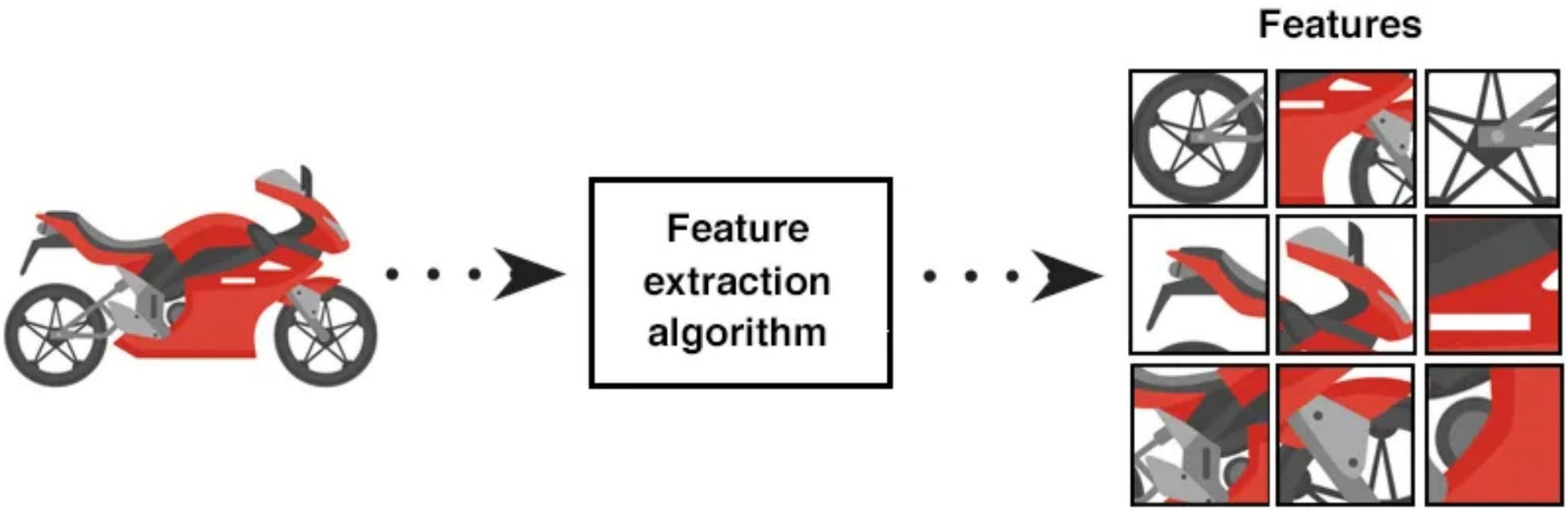
?



# Machine Learning

Feature extraction:

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

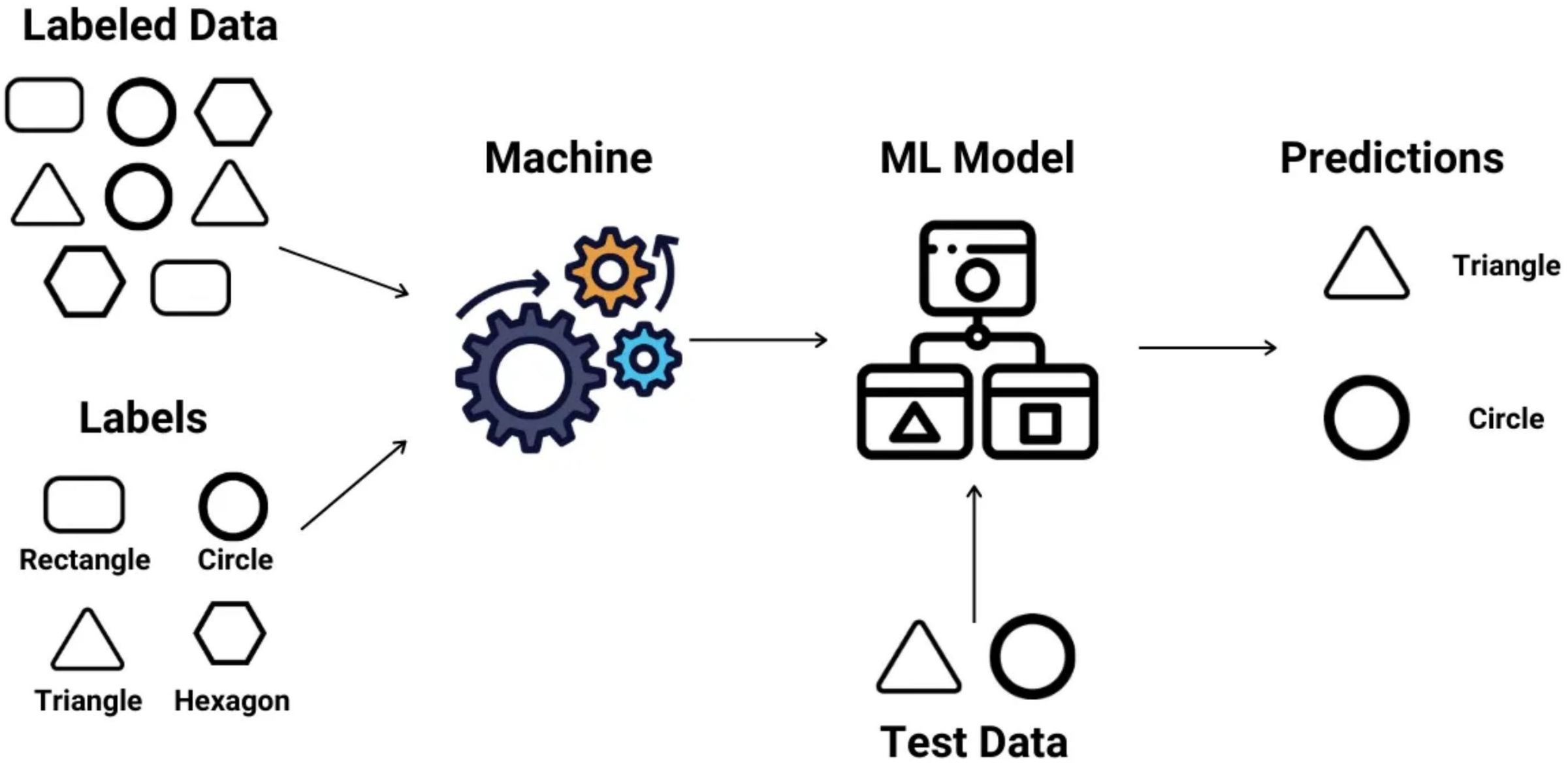



https://medium.com/@evertongomede/feature-engineering-for-computer-vision-f01a76d8058c

# Machine Learning

?

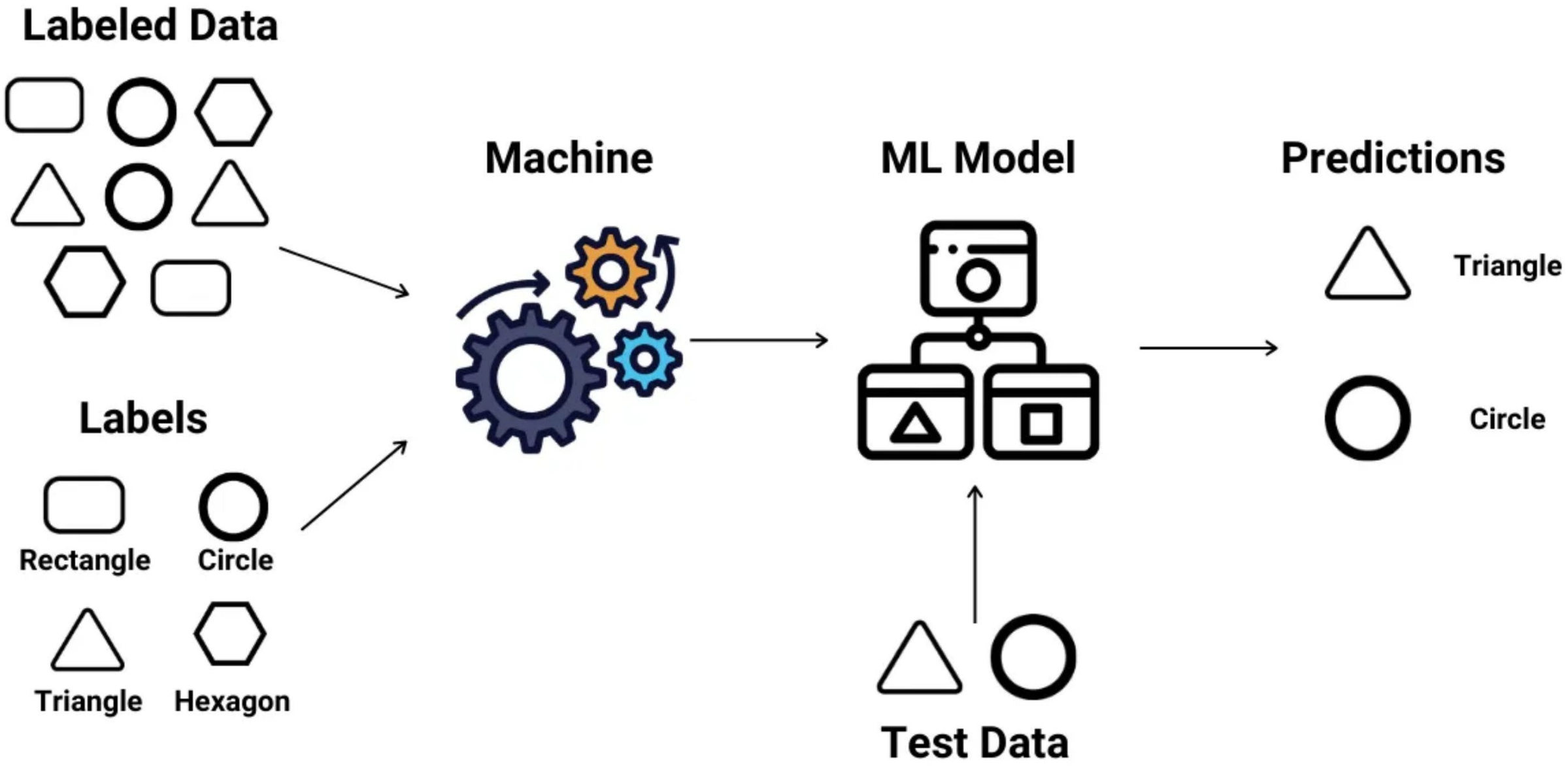
- trained on a labelled dataset: each training example is coupled with an output label
- learns a mapping from inputs to outputs to predict the labels of new, unseen data



# 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

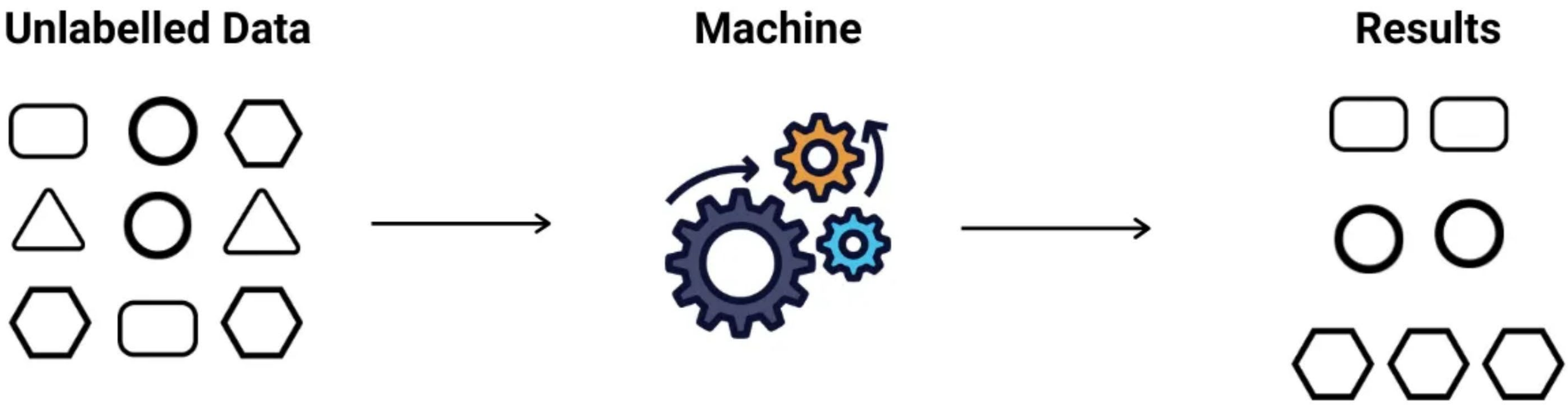



https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644

# Machine Learning

?

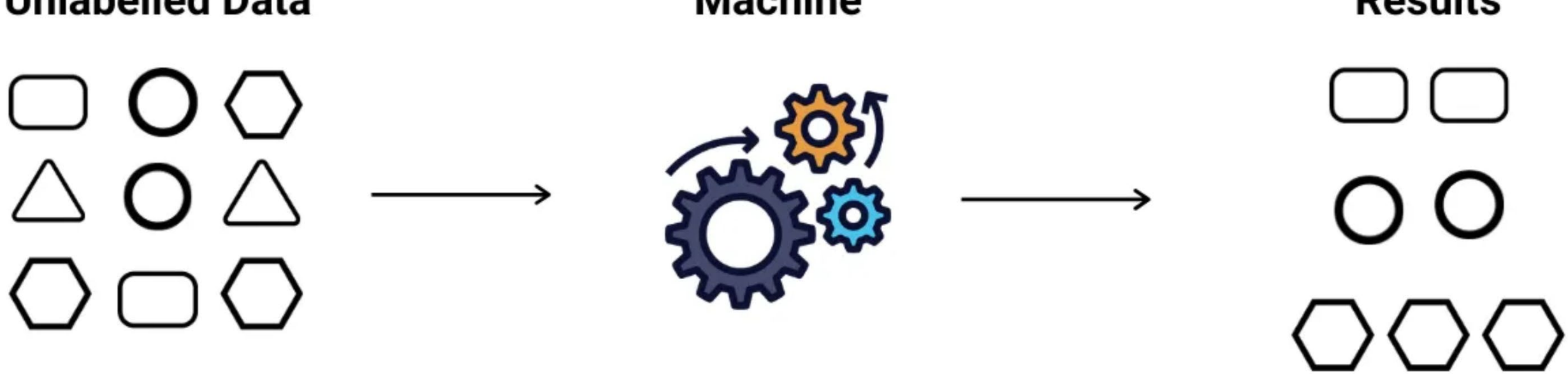
- trained on an unlabelled dataset
- infers the natural (intrinsic) structure present within a set of data points



# Machine Learning

Unsupervised learning:

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

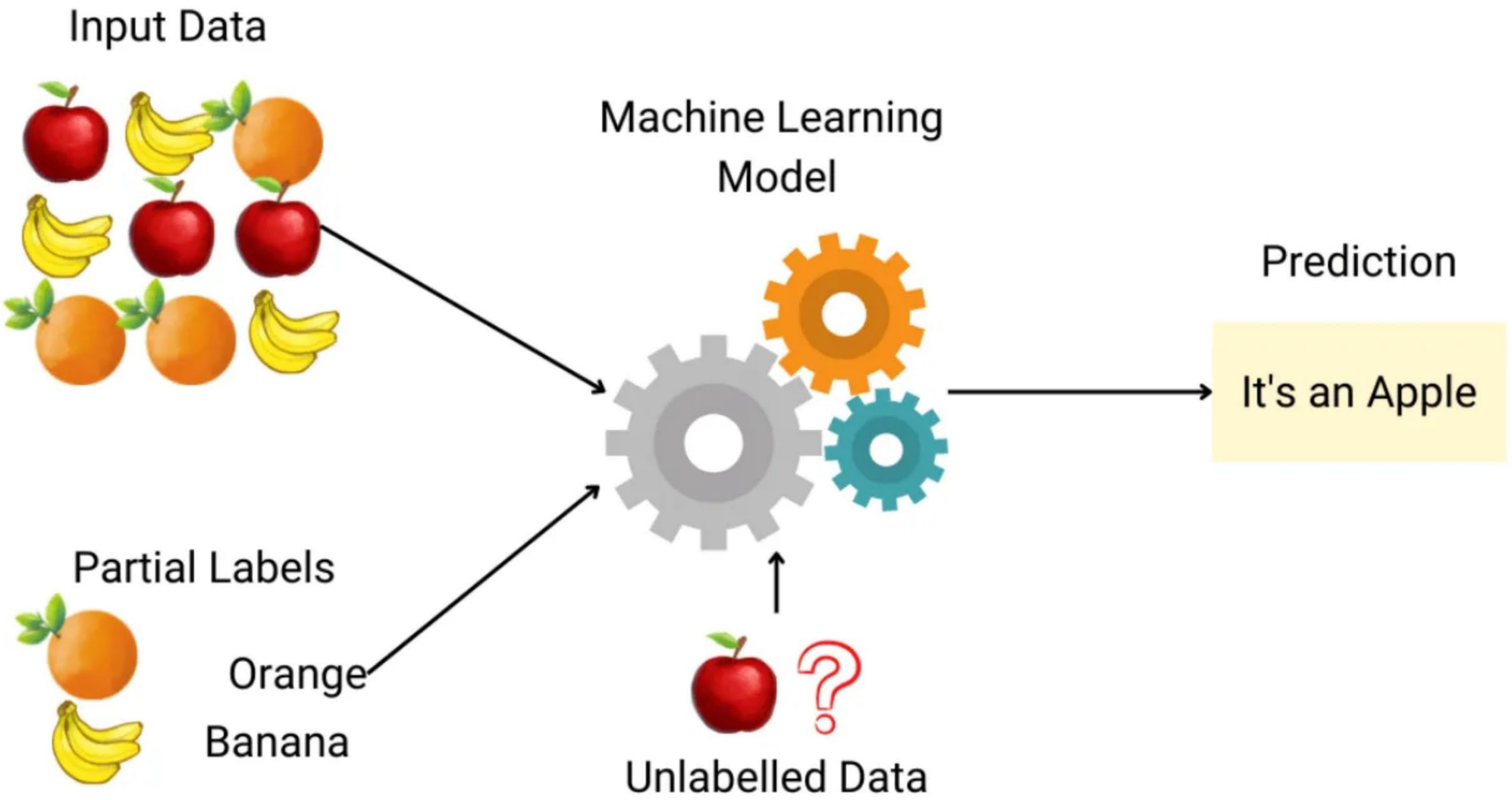



https://medium.com/@metehankozan/supervised-and-unsupervised-learning-an-intuitive-approach-cd8f8f64b644

# Machine Learning

?

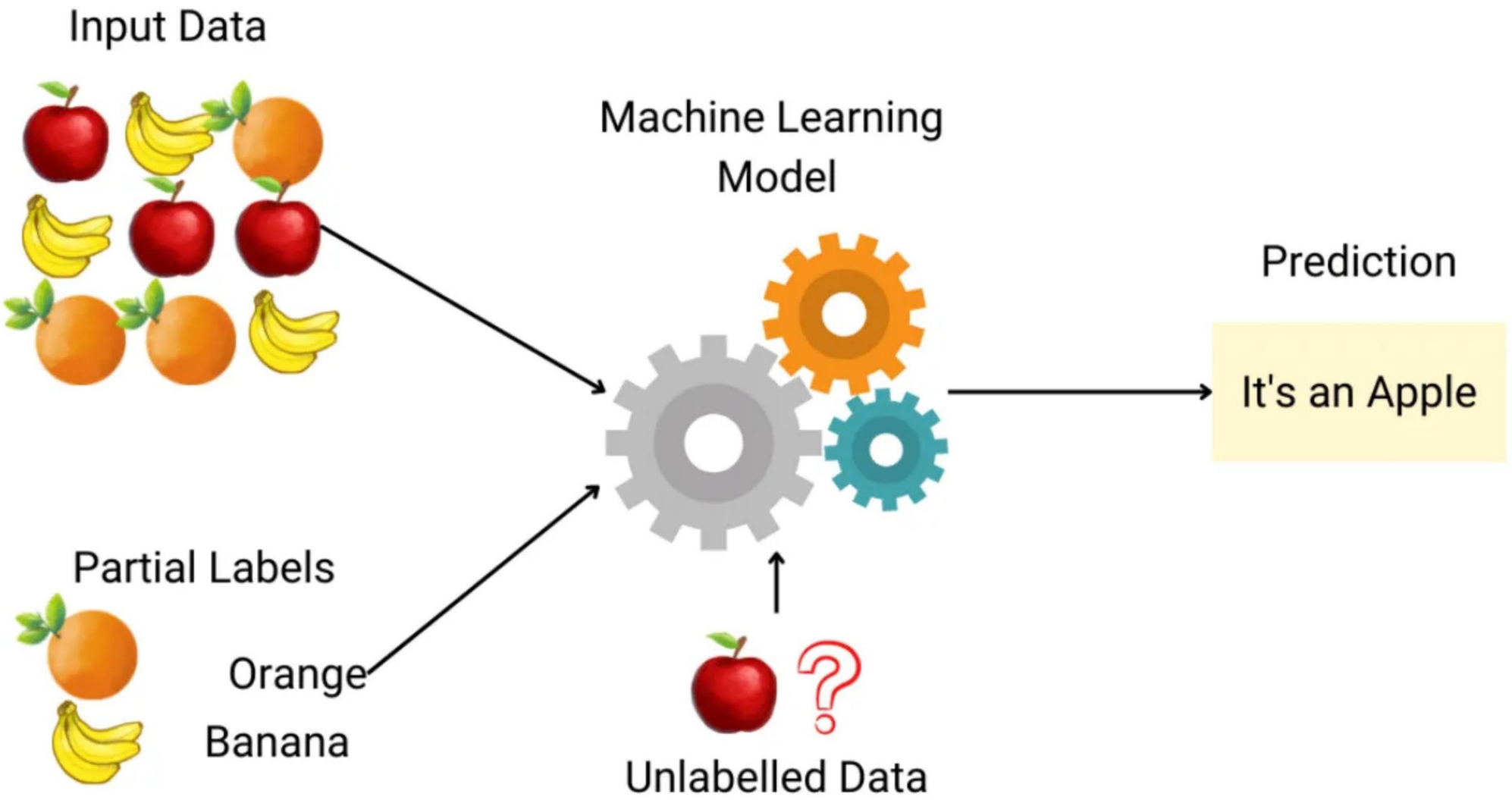
- uses a small amount of labelled data along with a large amount of unlabelled data
- labelled data helps guide the learning process, improving the model's accuracy when labelled data is scarce



# 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

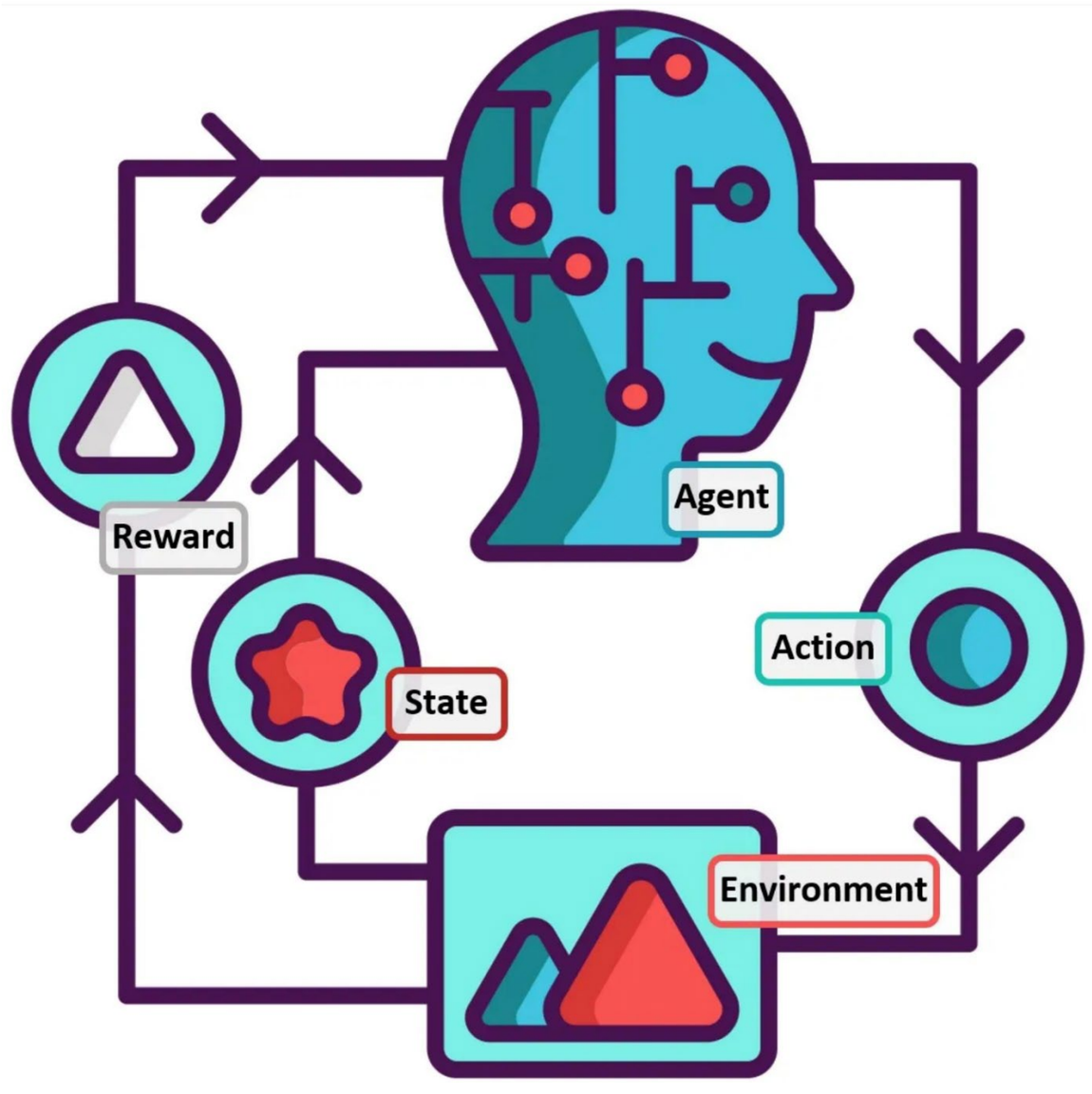



https://medium.com/@gayatri_sharma/a-gentle-introduction-to-semi-supervised-learning-7afa5539beea

# Machine Learning

?

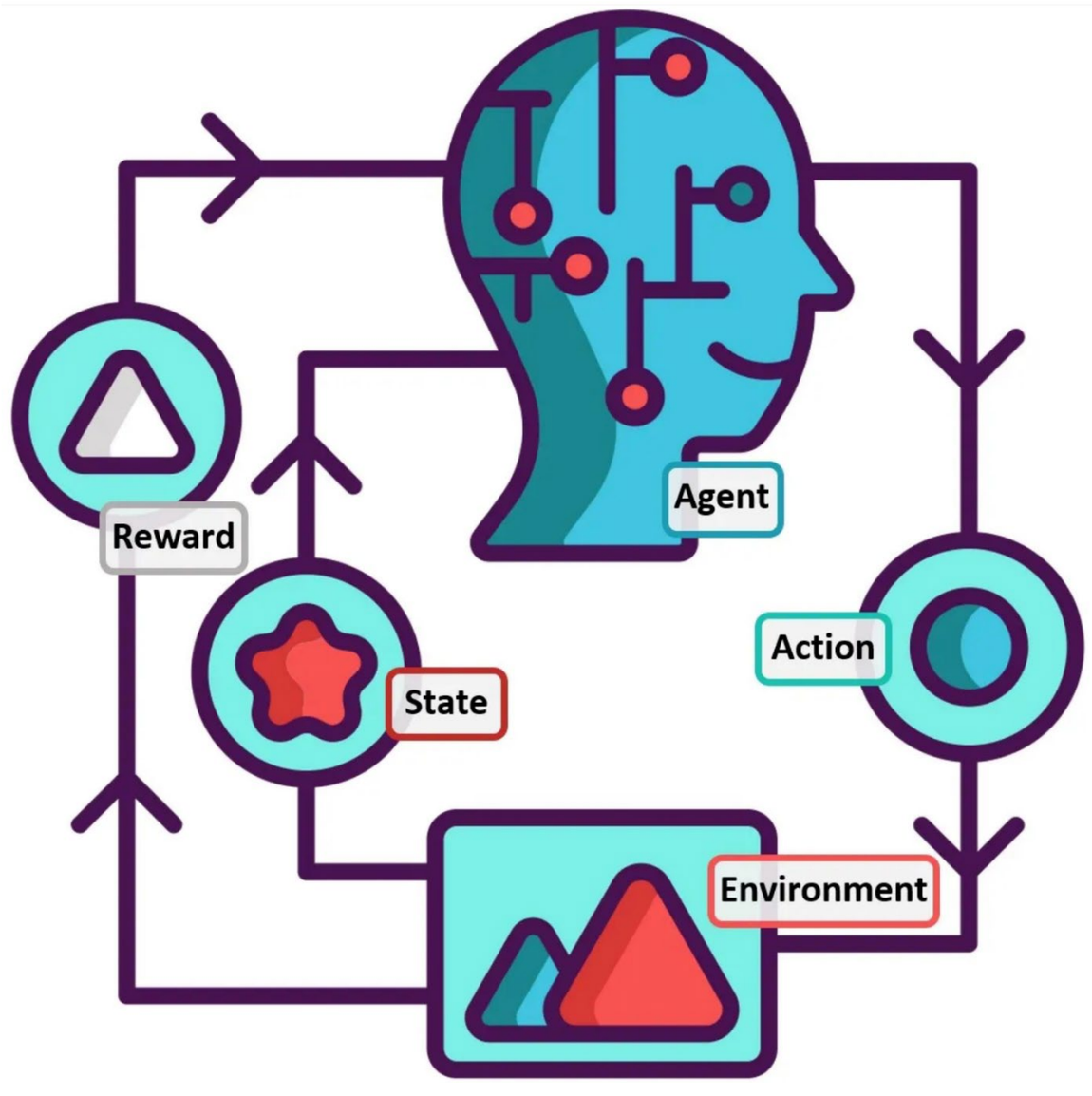
- agent learns to make decisions by doing actions in an environment to maximise cumulative reward
- the agent learns from the effects of its actions rather than from explicit instruction



# 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

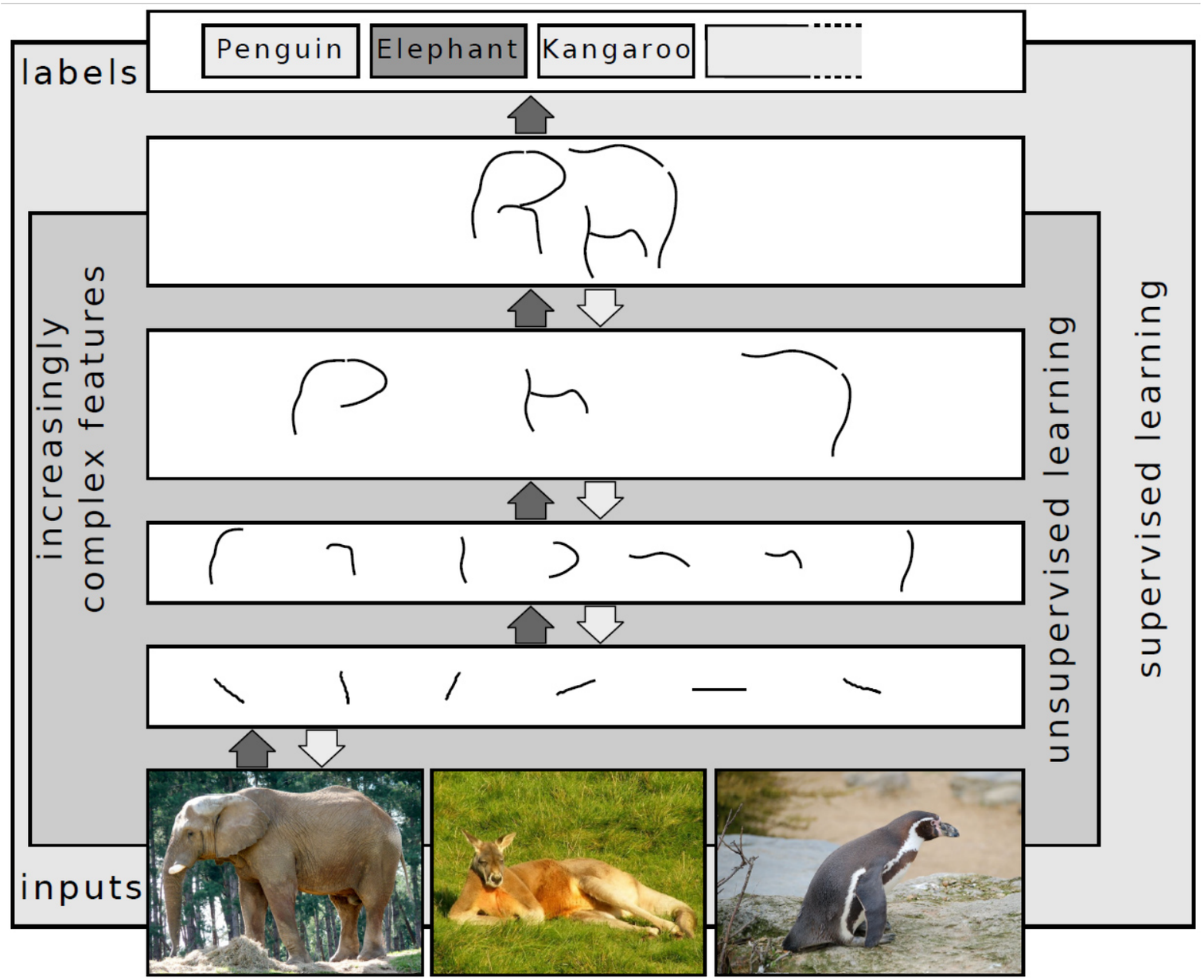



https://towardsdatascience.com/reinforcement-learning-fda8ff535bb6

# Machine Learning

?

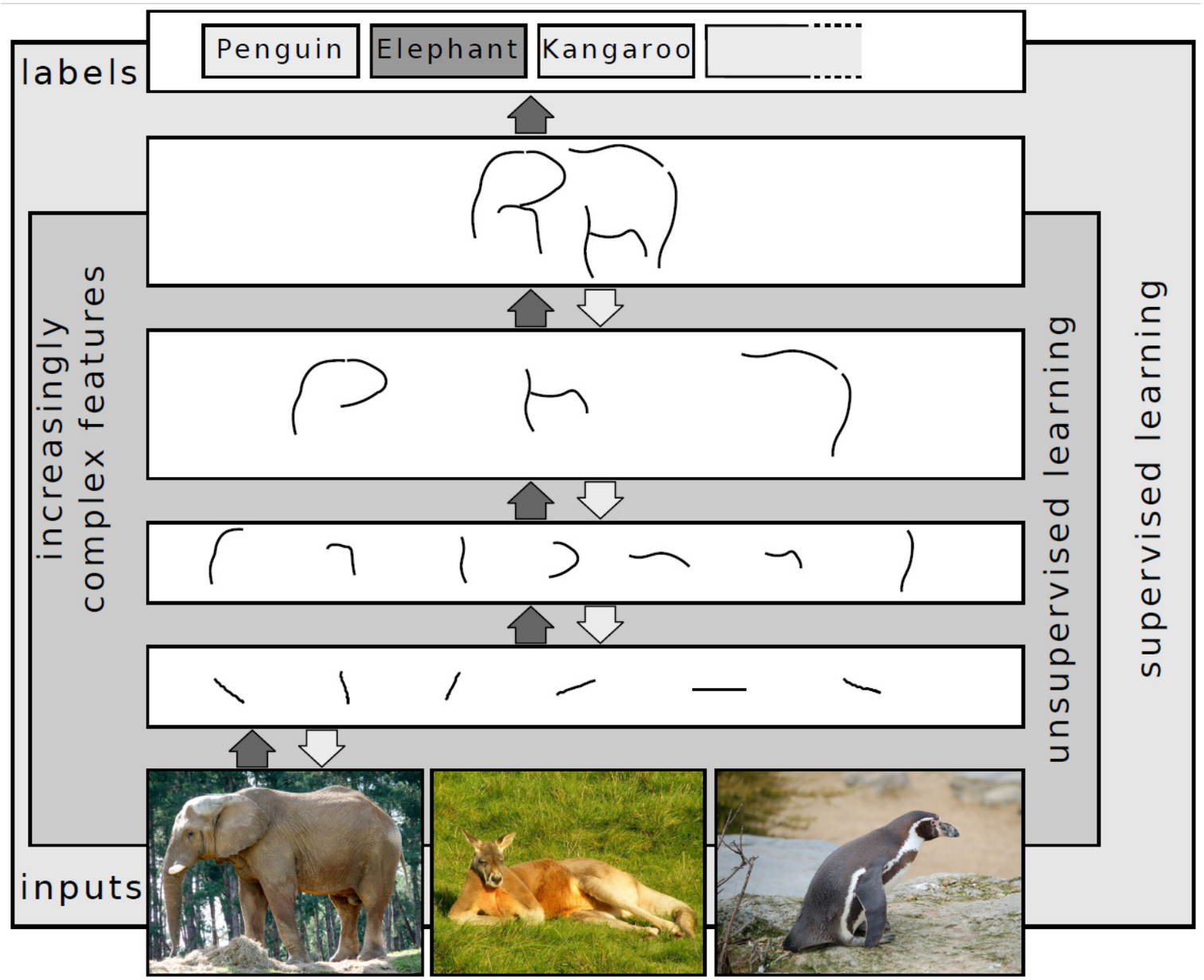
- uses neural networks with many layers (deep neural networks) to learn abstract representations of data
- works for unstructured data like images, audio and text



# Machine Learning

Deep learning:

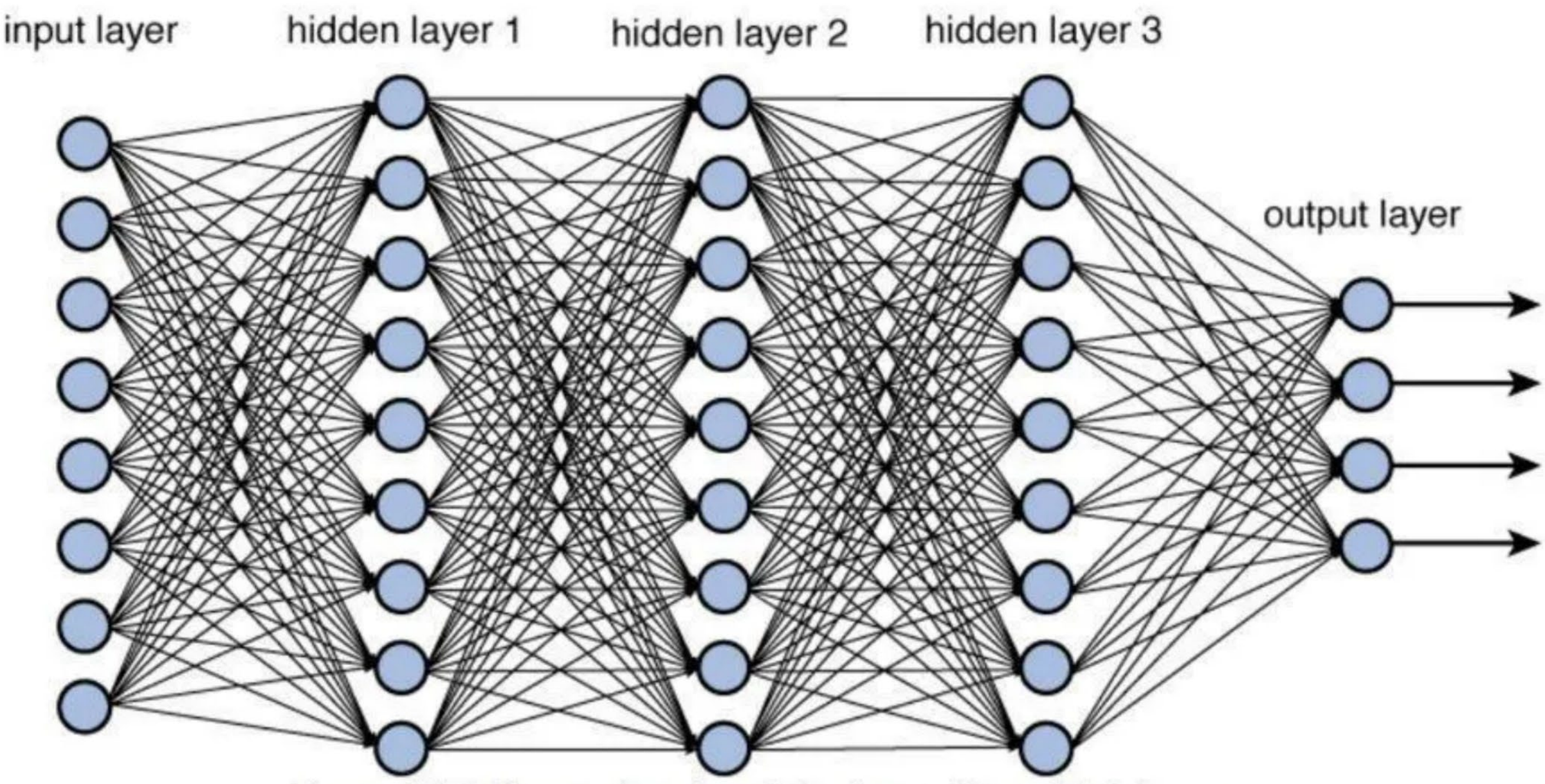
- uses neural networks with many layers (deep neural networks) to learn abstract representations of data
- works for unstructured data like images, audio and text




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

# Machine Learning

Deep learning




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

# 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

# Questions

?

# 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)**

# 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% | | | | | | |

# Articles to review

1. “Computing Machinery and Intelligence”, Alan Turing (1950)
2. “A Chess-Playing Machine”, Claude E. Shannon (1950)
3. “Why Should Machines Learn”, Herbert Simon (1983)
4. “Flocks, Herds, and Schools: A Distributed Behavioral Model”, Craig Reynolds (1987)
5. “Elephants Don't Play Chess”, Rodney A. Brooks (1990)
6. “The Myth of the Last Metaphor”, Joseph Weizenbaum (1995)
7. “RoboCup: The Robot World Cup Initiative”, Hiroaki Kitano et al. (1996)
8. “The RoboCup Synthetic Agent Challenge”, Hiroaki Kitano et al. (1997)

# Articles to review

9. “Perceptron, Encyclopedia of Computer Science”, Laveen Kanal (2003)

10. “Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm”, Nathan Ensmenger (2012)

11. “A Few Useful Things to Know About Machine Learning”, Pedro Domingos (2012)

12. “Overview on DeepMind and Its AlphaGo Zero AI”, Sean Holcomb et al. (2018)

13. “Conversations with ELIZA on Gender and Artificial Intelligence”, Pedro Costa & Luisa Ribas (2018)

14. “The Five Tribes of Machine-Learning: A Brief Overview”, Jens Pohl (2019)

15. “Deep New: The Shifting Narratives of Artificial Intelligence from Deep Blue to AlphaGo”, Paolo Bory (2019)

16. “ChatGPT Is a Blurry JPEG of the Web”, Ted Chiang (2023)

# Articles to review

17. "Steps Toward Artificial Intelligence", Minsky (1961)

18. "Learning representations by back-propagating errors", Rumelhart et al. (1986)

19. "Heuristic Problem Solving The Next Advance in Operations Research", Simon & Newell (1958)

20. "Intelligence Without Representation", Brooks (1991)

21. "Deep learning", LeCun et al. (2015)

22. "ImageNet Classification with Deep Convolutional Neural Networks", Krizhevsky et al. (2012)

23. "Datasheets for Datasets", Gebru et al. (2018)

24. "There is a blind spot in AI research", Crawford & Calo (2016)

# Questions

?

# Generative Artificial Intelligence

1. Do not enter confidential, personal, proprietary or otherwise sensitive information into ChatGPT or other cloud-based generative AI platforms
    - Once information is fed into a cloud-based generative AI tool, it can be incorporated into the datasets that the tool uses to generate content for all users and could therefore be presented to another user at a later date (including those maliciously seeking to collect and aggregate it)

2. Do not rely on the accuracy of outputs
    - Be aware that generative AI tools like ChatGPT often produce incorrect information
    - Generative AI tools use probabilistic models that are trained on enormous datasets to generate plausible new content: they use existing, uncorroborated content from many sources and can produce inaccurate, biased or creatively fictitious content. Repeated use of a generative AI tool can create different content each time it is run
    - You remain responsible for your work: you must independently verify and edit AI-generated content to ensure the integrity, accuracy, and suitability of the output

3. Openly acknowledge your use of generative AI

4. Check for restrictions and conditions

https://intranet.sydney.edu.au/strategy-governance/gen-ai.html

# Typical advice on using generative AI

1. Assistance purely with the language
    - paraphrasing or polishing the author's original content, rather than for suggesting new content is typically acceptable

2. Literature search
    - generative AI may be used for search assistance, e.g. to identify relevant literature
    - authors must read and discuss such references
    - ensure citation accuracy
    - consider possible biases in suggested citations

3. Low-novelty text
    - provide relevant citations for both the source of the text and and the source of the idea, including quotes for verbatim copying

4. New ideas or content
    - acknowledge the use of the generative AI model
    - check for known sources for the ideas and acknowledge them too
    - ensure that new content is "correct, coherent, original and does not have missing citations"

https://2023.aclweb.org/blog/ACL-2023-policy/

# Responsible Generative AI Principles (Elsevier)

1. Consider the real-world impact of AI solutions on people
   - “are people’s health, livelihood (including career prospects) or rights affected in any way?”

2. Take action to prevent the creation or reinforcement of unfair bias
   - output may “lead to less favourable outcomes for individuals or groups based on gender, ethnicity, socio-economic status, and other personal attributes”

3. Explain how AI solutions work
   - “appropriate level of transparency for each application and use-case to ensure that different users can understand and trust the output”

4. Create accountability through human oversight
   - “important that humans have ownership and accountability over the development, use, and outcomes of AI systems”

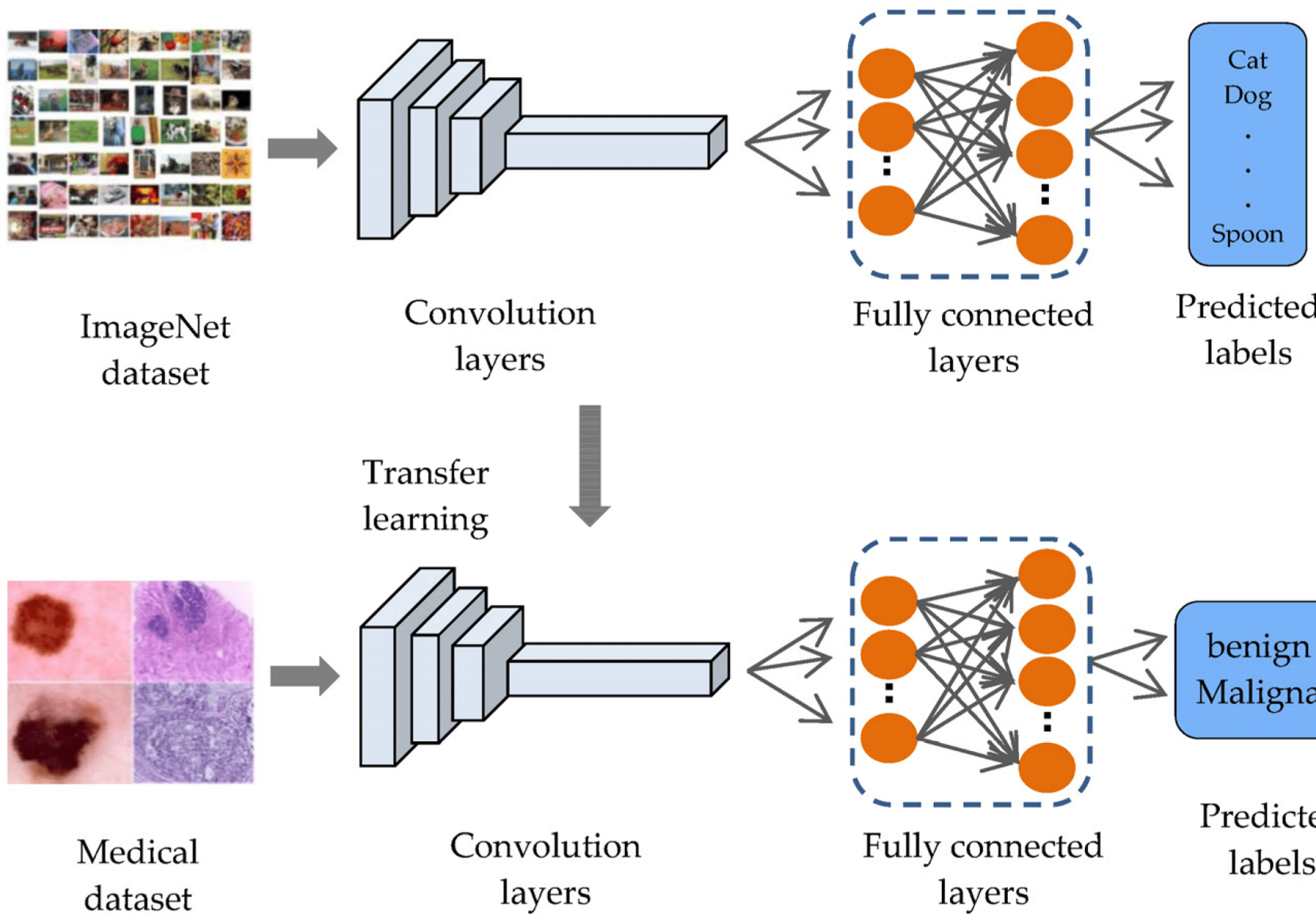
5. Respect privacy and champion robust data governance
   - ensure “robust data management and security policies and procedures”

https://www.elsevier.com/about/policies-and-standards/responsible-ai-principles

# Machine Learning

Transfer learning:

- a model is pre-trained on one task and is adapted to a different but related task
- improves learning efficiency and performance on the new task

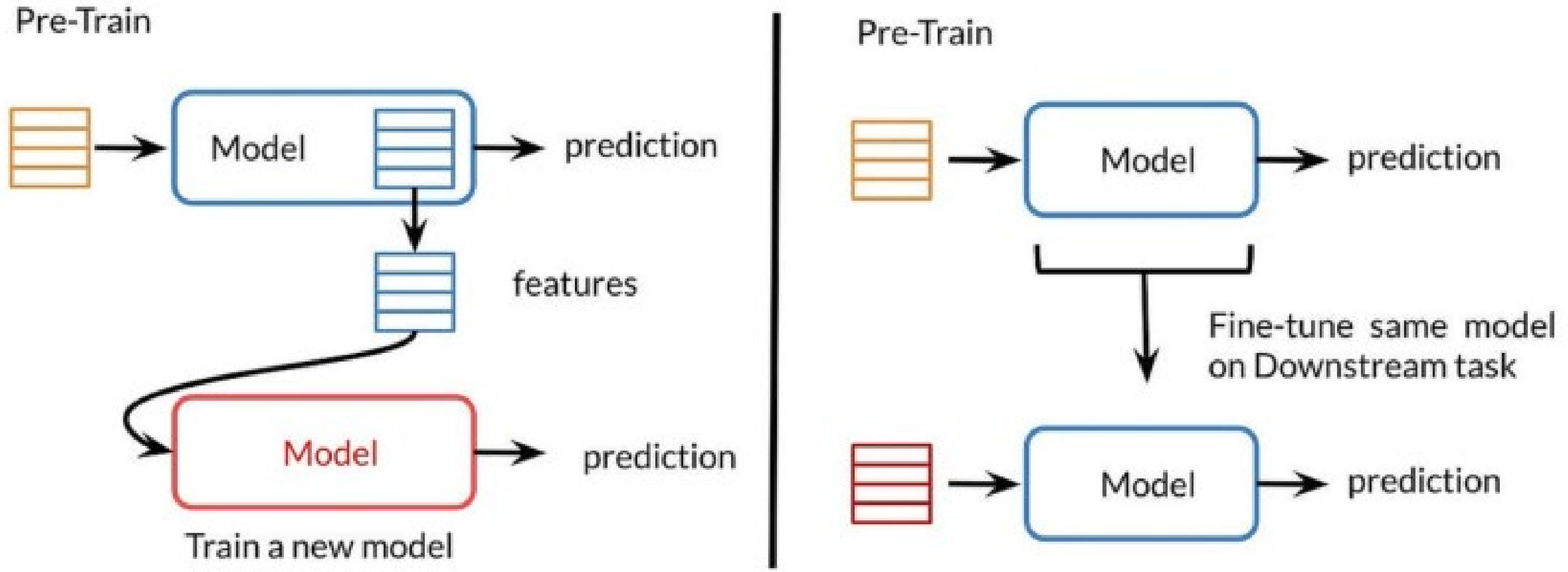



https://www.mdpi.com/1424-8220/23/2/570

# Machine Learning

Transfer learning:

- common algorithms:
  - feature extraction: using the pre-trained model's features as input to a new model
  - fine-tuning: adjusting the weights of a pre-trained model on a new dataset
- applications:
  - image classification with limited data
  - natural language processing

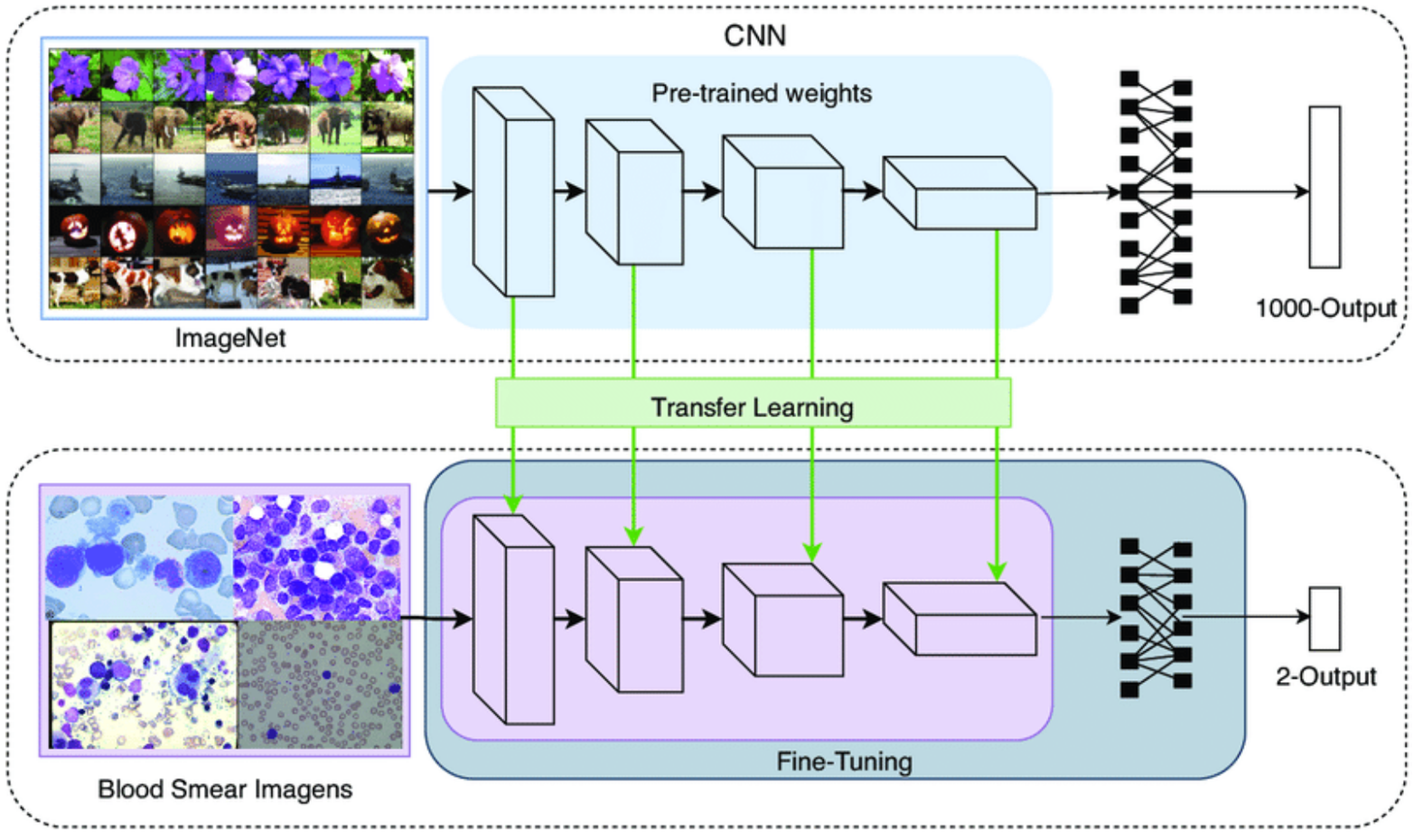



https://vitalflux.com/transfer-learning-vs-fine-tuning-differences/

# Machine Learning

## Transfer learning: feature extraction

- feature transfer: models pre-trained on large datasets extract features from smaller datasets
  - example: using a pre-trained model to extract image features (from ImageNet) for a new classification task with limited data
- fine-tuning: pre-trained model's features are fine-tuned on the new dataset
  - example: fine-tuning a model pre-trained on a specific text classification task to extract better text features



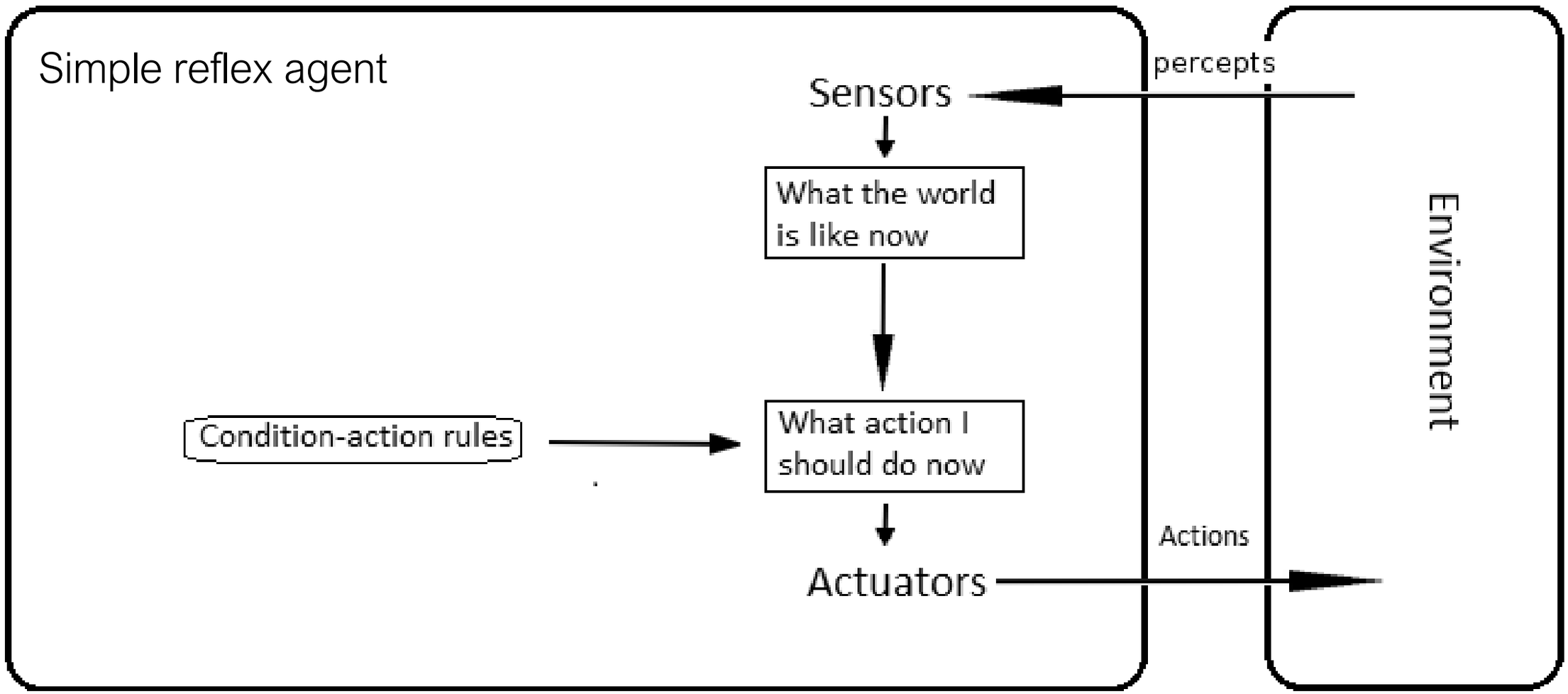

https://www.researchgate.net/figure/The-transfer-learning-and-fine-tuning-techniques-used-in-the-development-of-the-proposed_fig3_351085550

# Questions

?

# Intelligent agents

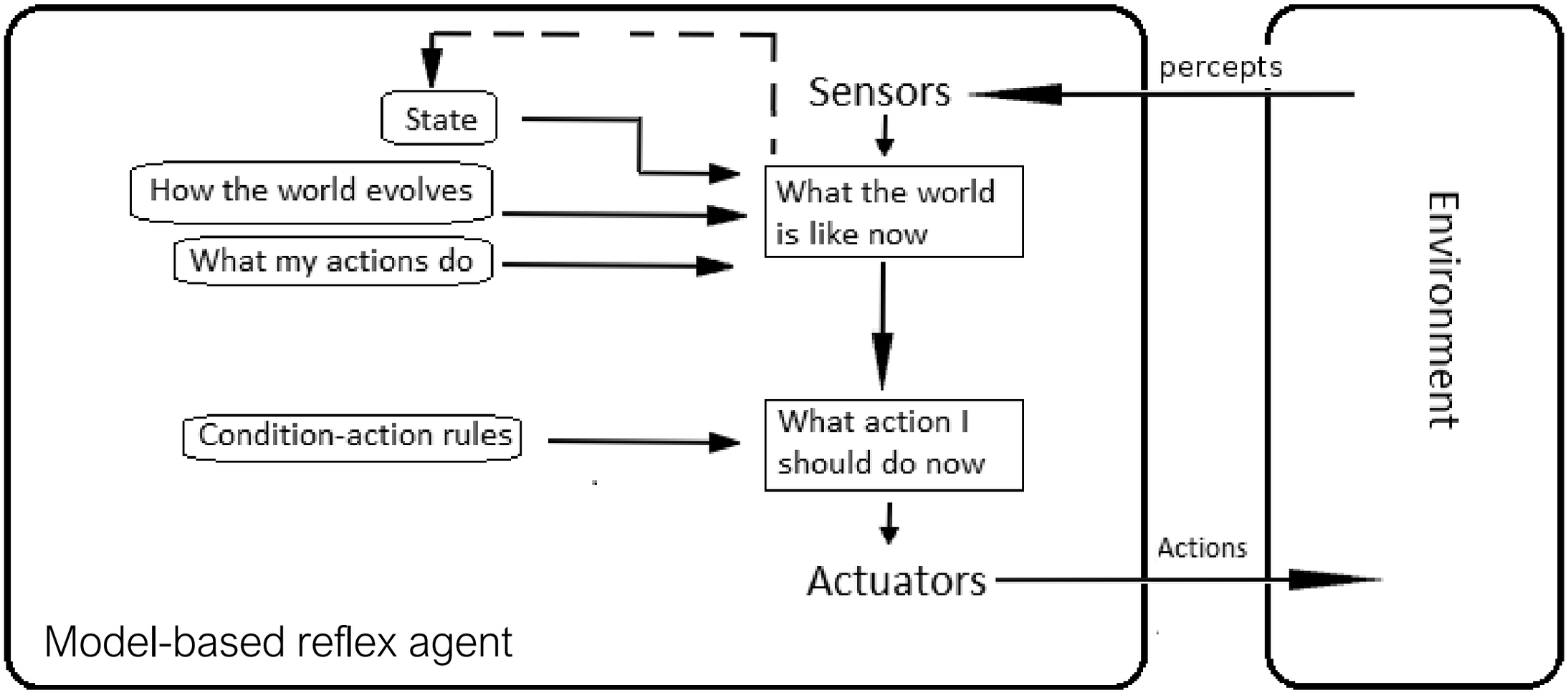
- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge




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

# Intelligent agents

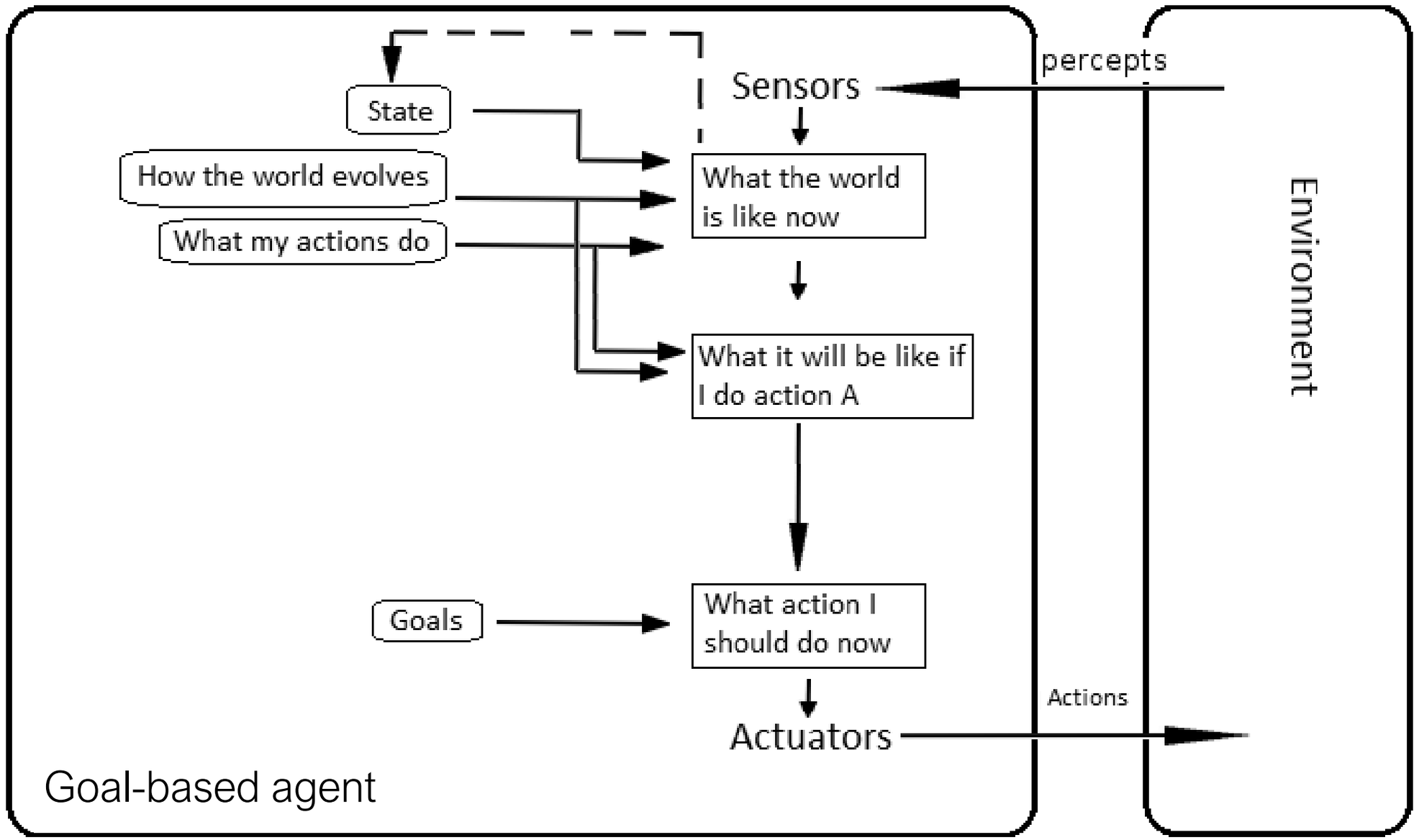
- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge




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

# Intelligent agents

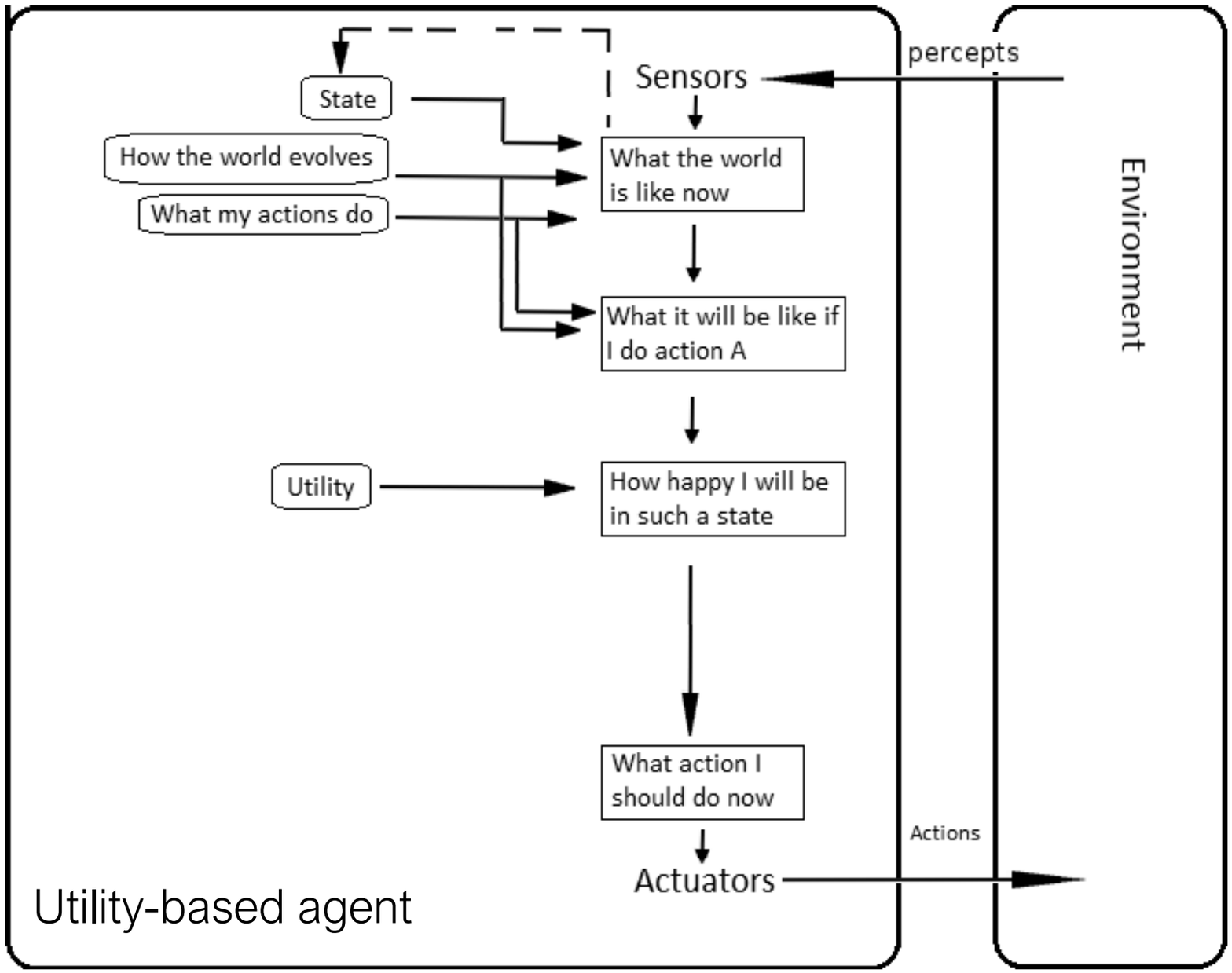
- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge




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

# Intelligent agents

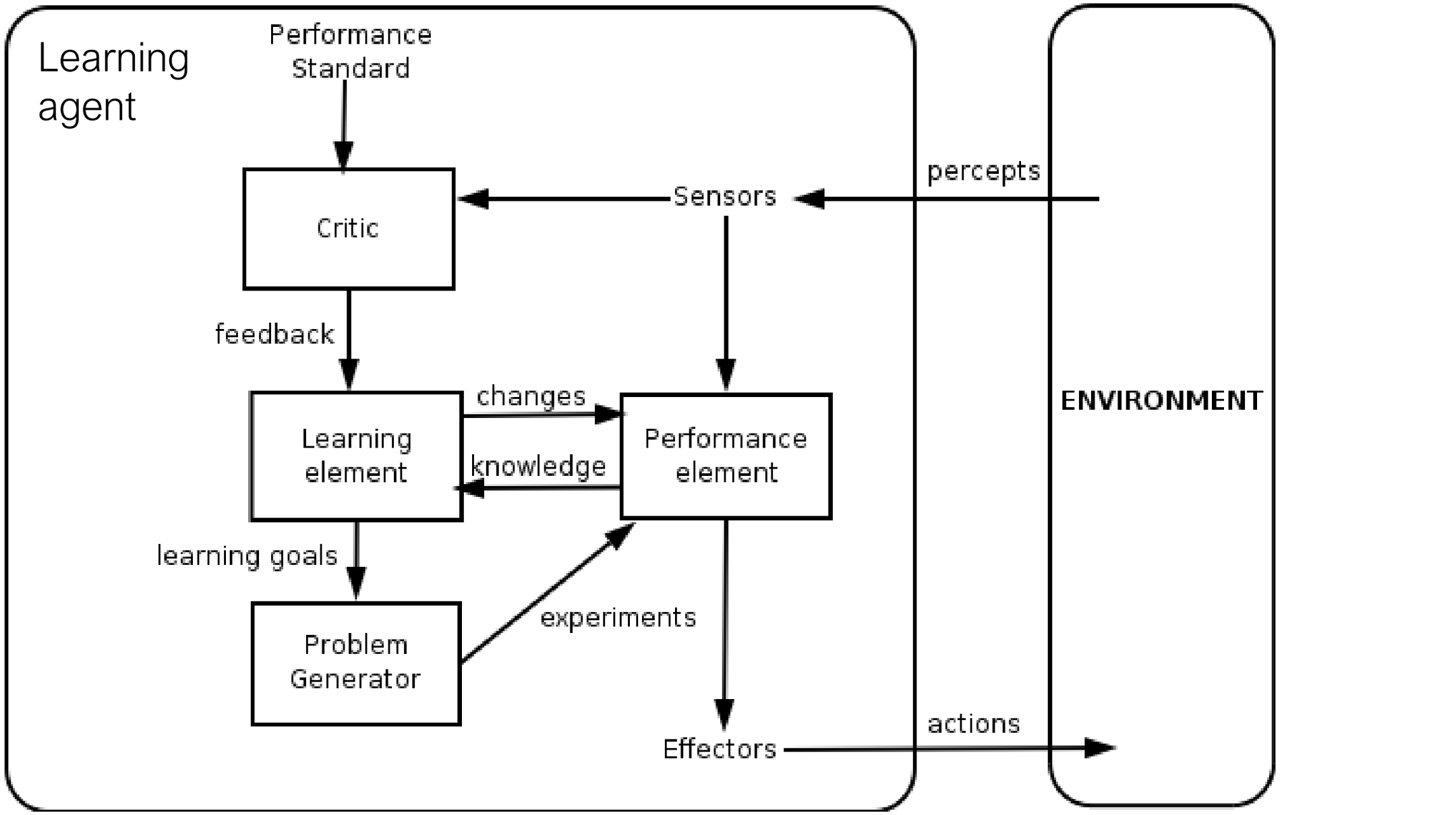
- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge




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

# Intelligent agents

- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge

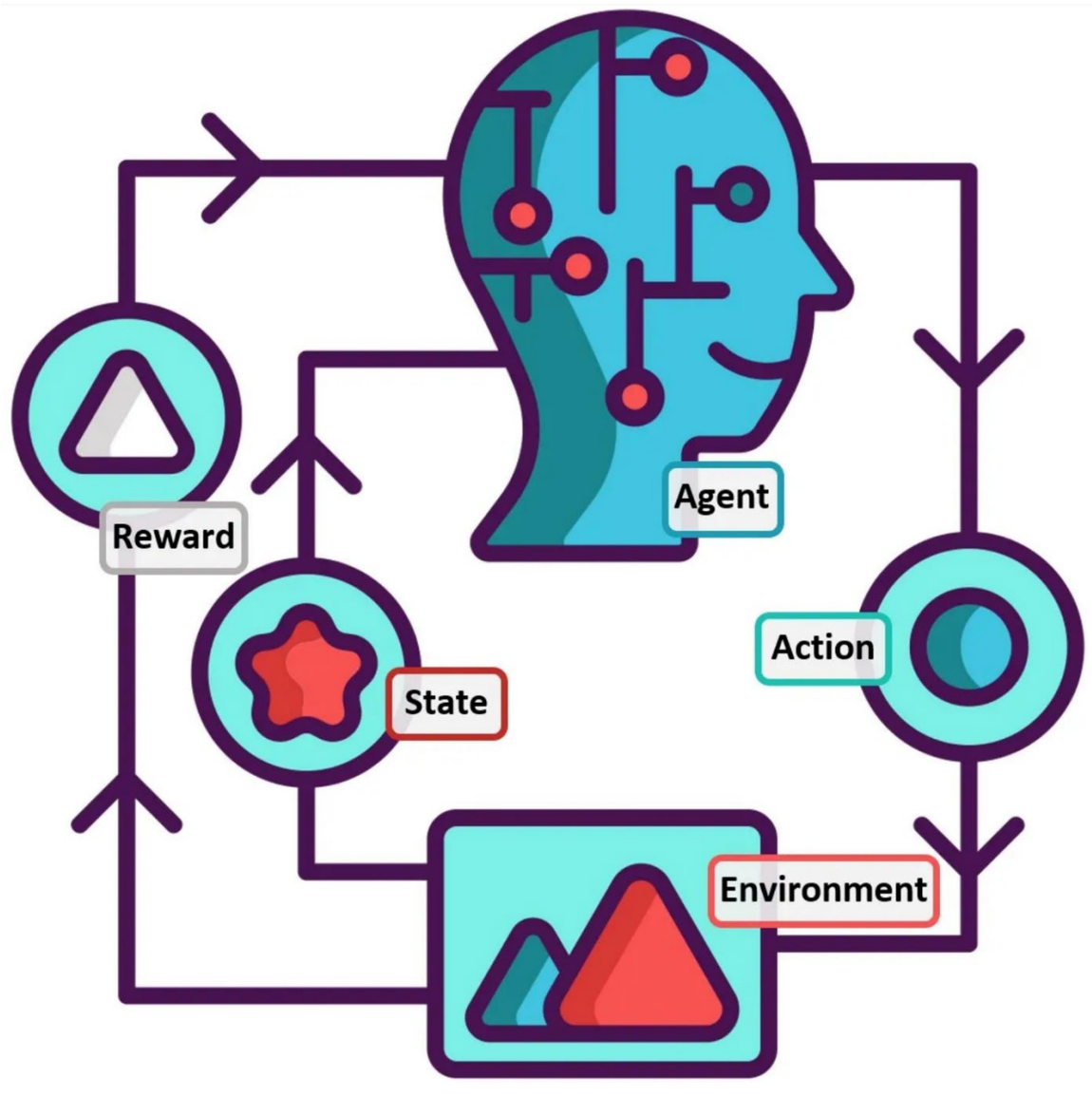



https://en.wikipedia.org/wiki/Intelligent_agent

# 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

# Questions

?

# Break

?

# Natural Interaction

- Natural Language Processing (NLP)
- speech and voice recognition
- computer vision
- multimodal interaction
- Human-Robot Interaction (HRI):
  - natural language commands: allowing users to interact with robots using spoken language
  - gesture-based control: using hand or body movements to control robot actions
  - social robotics: designing robots that can understand and respond to social cues and behaviors
- affective computing:
  - emotion recognition: analysing facial expressions, voice tones and physiological signals to detect emotions
  - sentiment analysis: determining the emotional tone of text or speech
  - adaptive responses: tailoring interactions based on detected emotions
- context-aware computing (smart homes, adaptive interfaces, …)

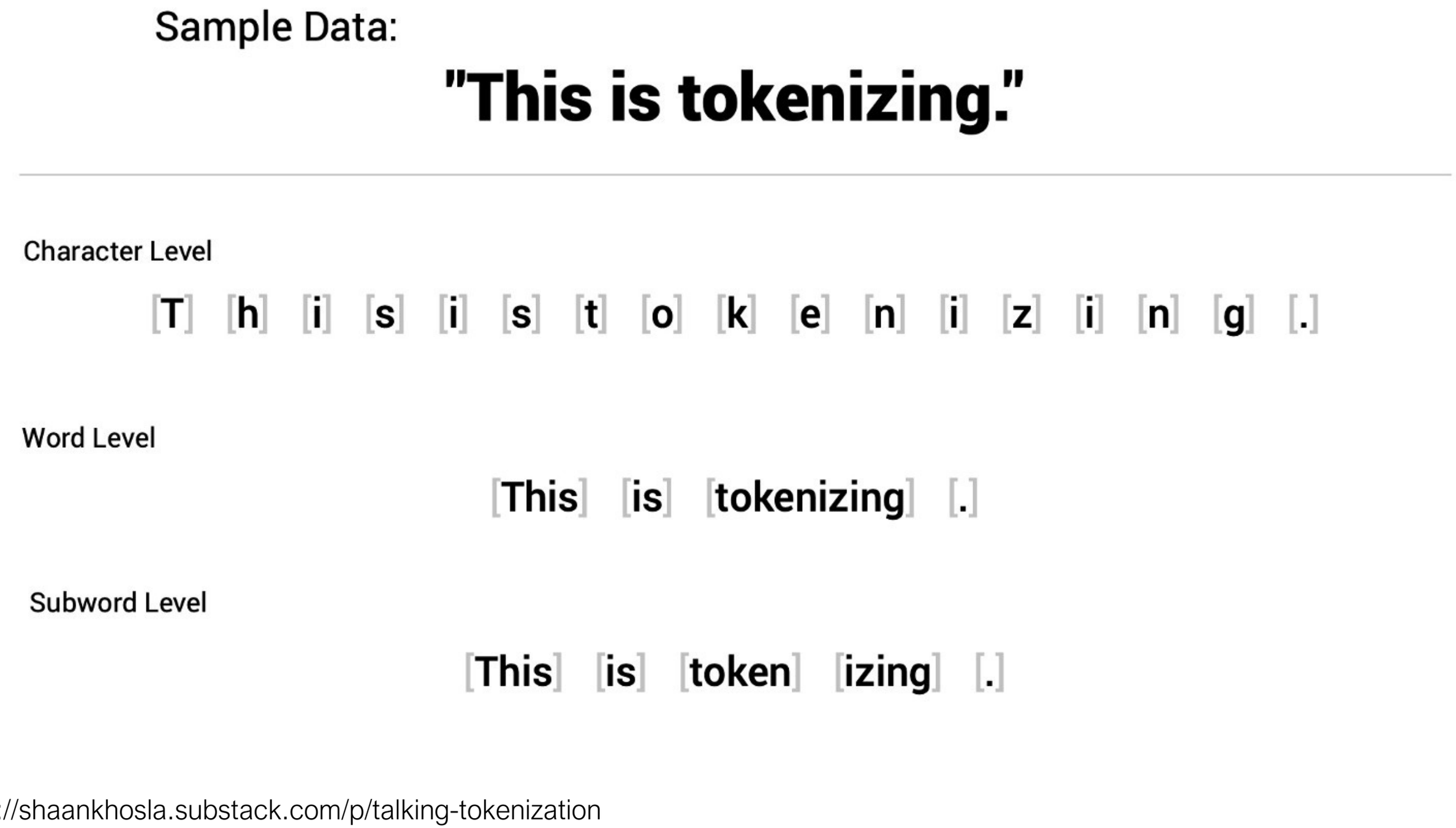
# Natural Interaction

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

- key techniques:
  - tokenization: breaking text into individual words or phrases
  - syntax and semantic analysis: understanding the grammatical structure and meaning of sentences
  - Named Entity Recognition (NER): identifying and classifying entities (names, dates, …) within text
  - sentiment analysis: determining the sentiment or emotional tone of text
  - machine translation: translating text from one language to another
  - speech recognition: converting spoken language into text
  - Text-to-Speech (TTS): generating spoken language from text
  - dialogue systems and chatbots: engaging in conversations with users

# Natural Interaction

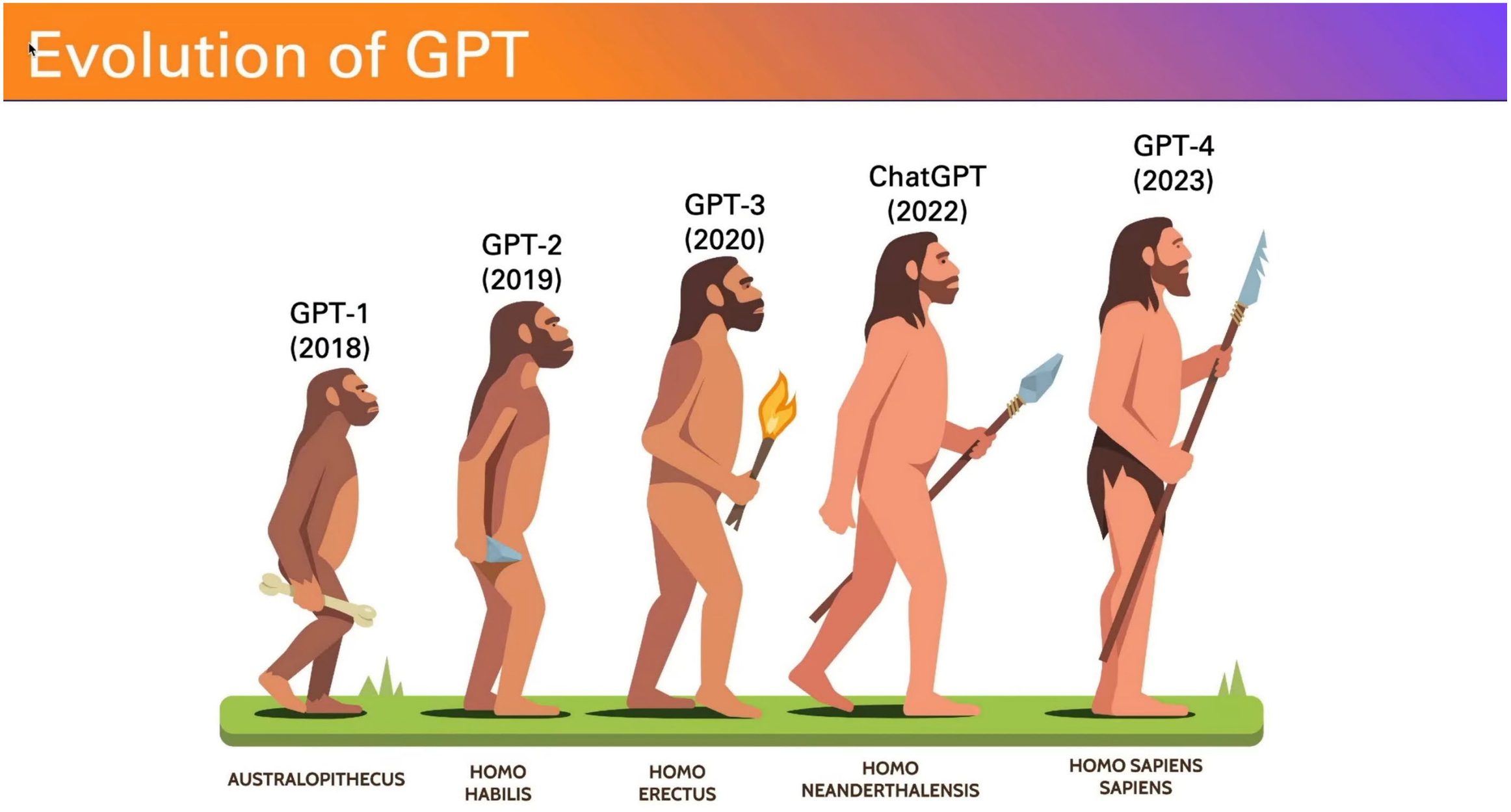
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




https://shaankhosla.substack.com/p/talking-tokenization

# Natural Interaction

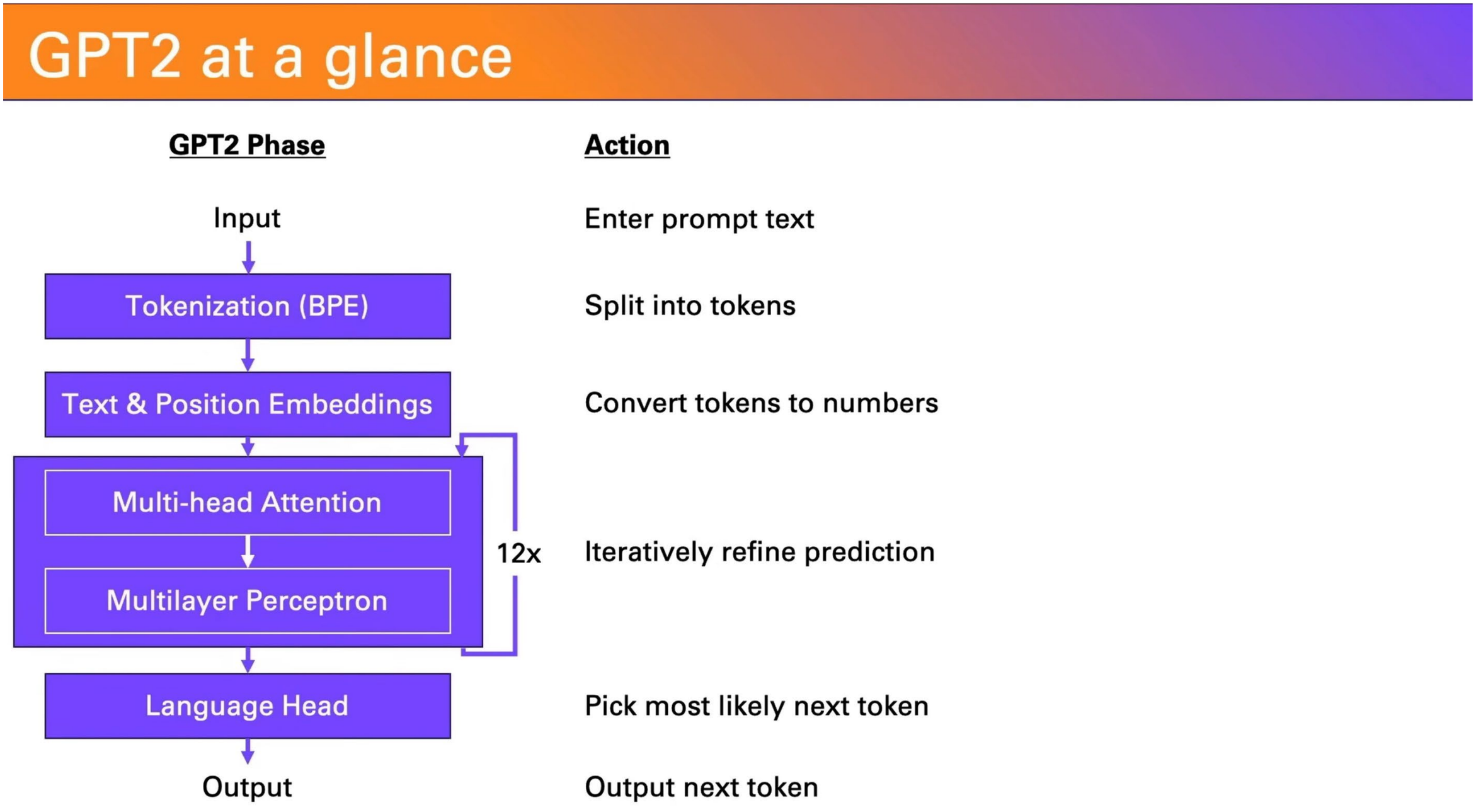
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# Natural Interaction

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

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

## Tokenization

| GPT2 Phase | Action | Tab in Spreadsheet |
|---|---|---|
| Tokenization (BPE) | Split into tokens | Prompt_to_Tokens, Constants |

| Tokens | Mike | is | quick | . | He | moves |
|---|---|---|---|---|---|---|
| Token IDs | 16073 | 318 | 2068 | 13 | 679 | 6100 |

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

# Natural Interaction

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

## Text & Position Embeddings

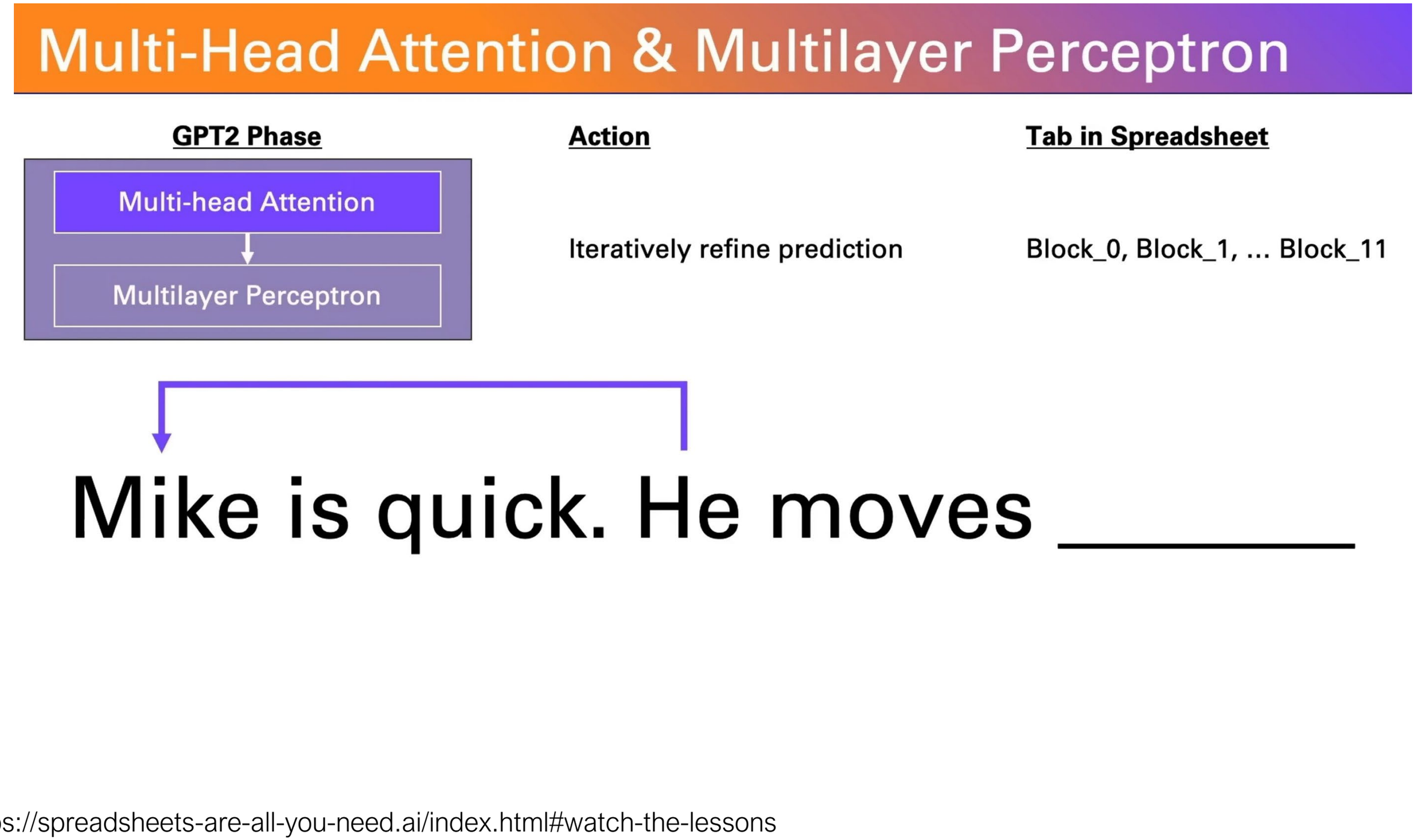
| GPT2 Phase | Action | Tab in Spreadsheet |
|---|---|---|
| Text & Position Embeddings | Convert tokens to numbers | Token_to_Text_Embeddings |

| | | | | | |
|---|---|---|---|---|---|
| Mike → | -0.0723849 | -0.237343 | 0.22738675 | -0.1565971 | -0.0809315 |
| is → | 0.01423593 | -0.0437407 | -0.0392998 | -0.1253222 | -0.1242475 |
| quick → | -0.0041666 | -0.1149568 | 0.25142141 | -0.1039098 | 0.03202019 |
| . → | 0.04635777 | -0.085105 | 0.13385425 | 0.04700242 | 0.02346371 |
| He → | -0.0551851 | -0.0318704 | 0.18177843 | 0.04345733 | -0.207683 |
| moves → | -0.0551851 | -0.0318704 | 0.18177843 | 0.04345733 | -0.207683 |

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

# Natural Interaction

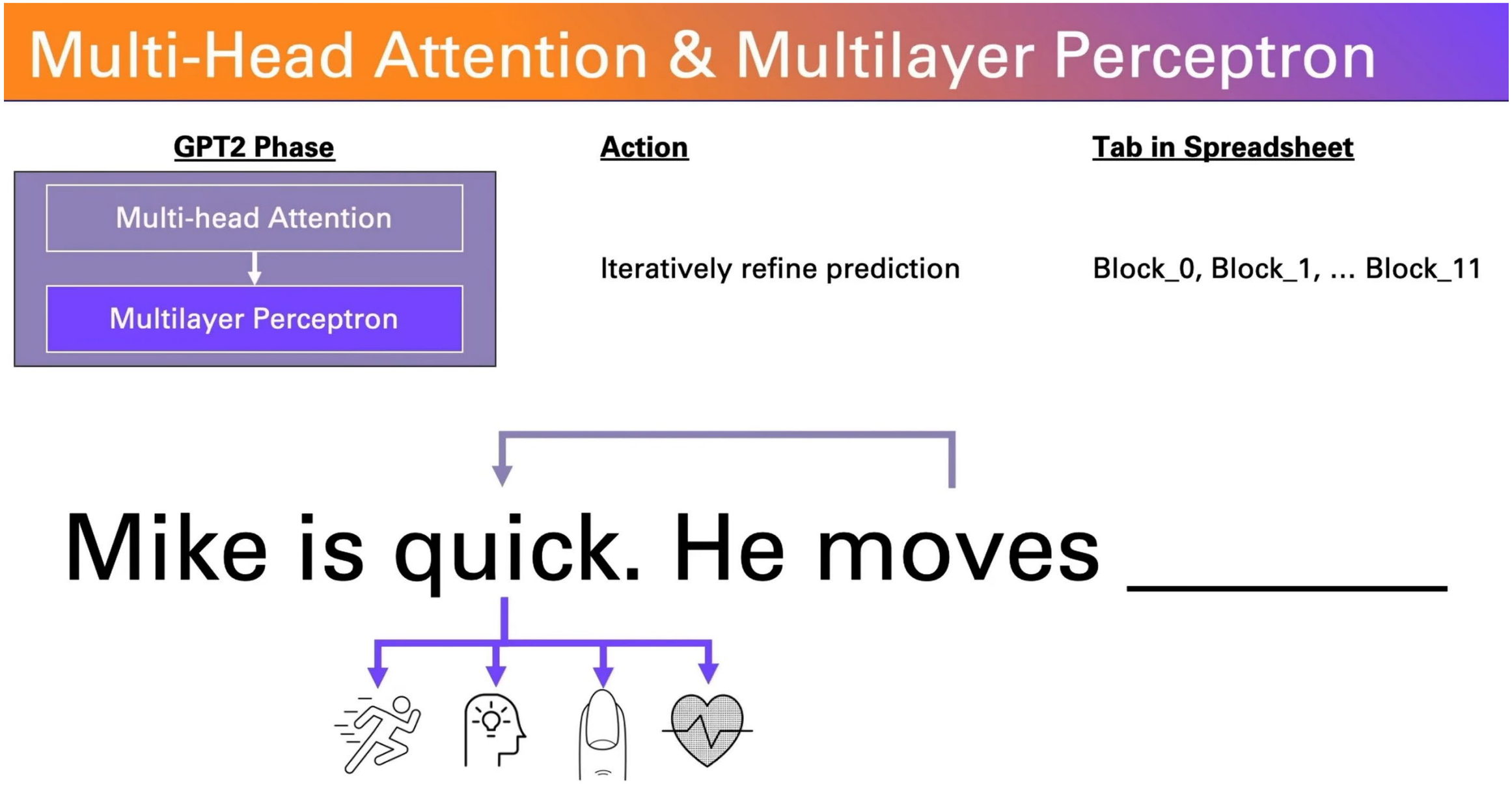
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# Natural Interaction

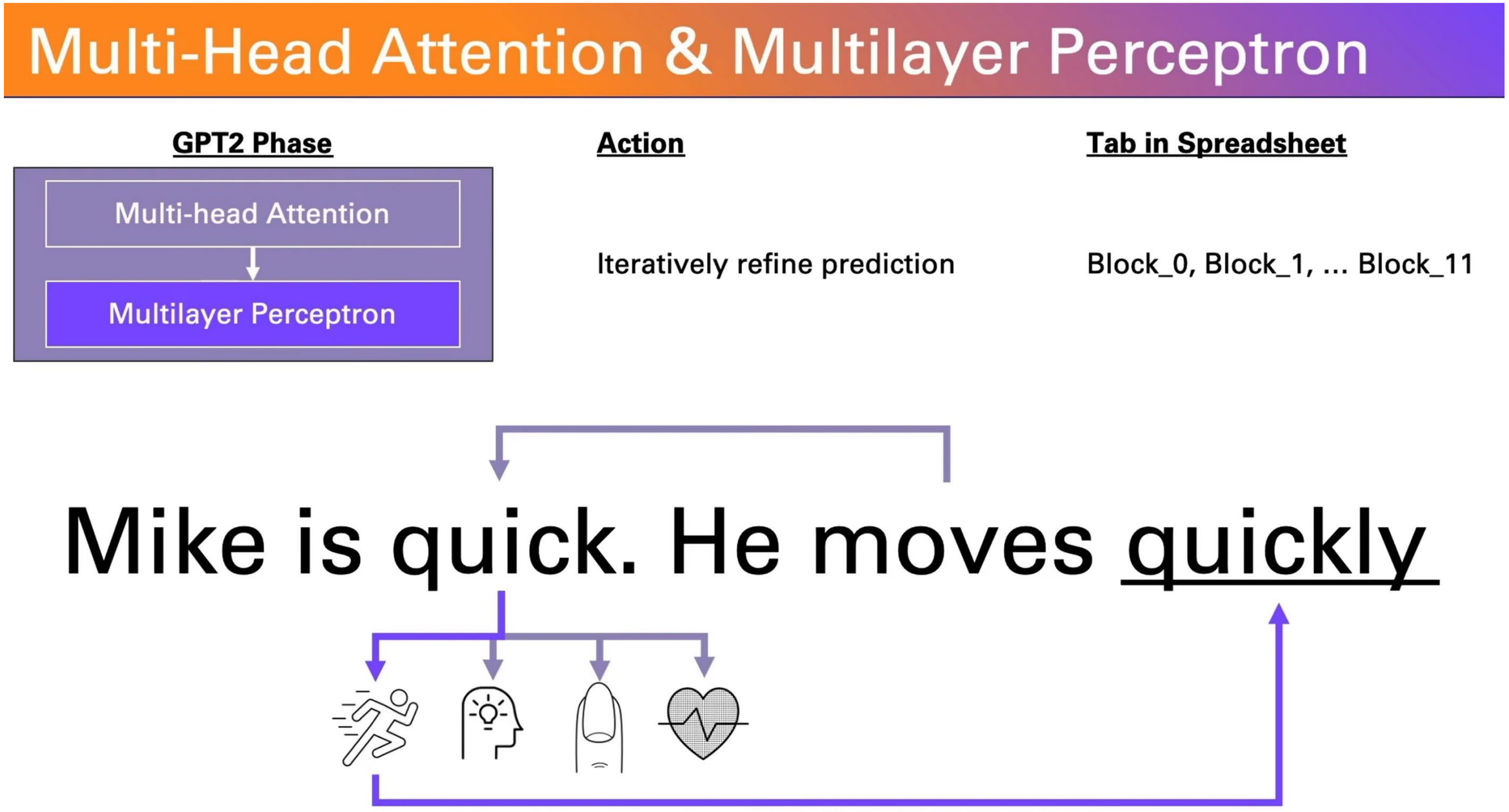
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# Natural Interaction

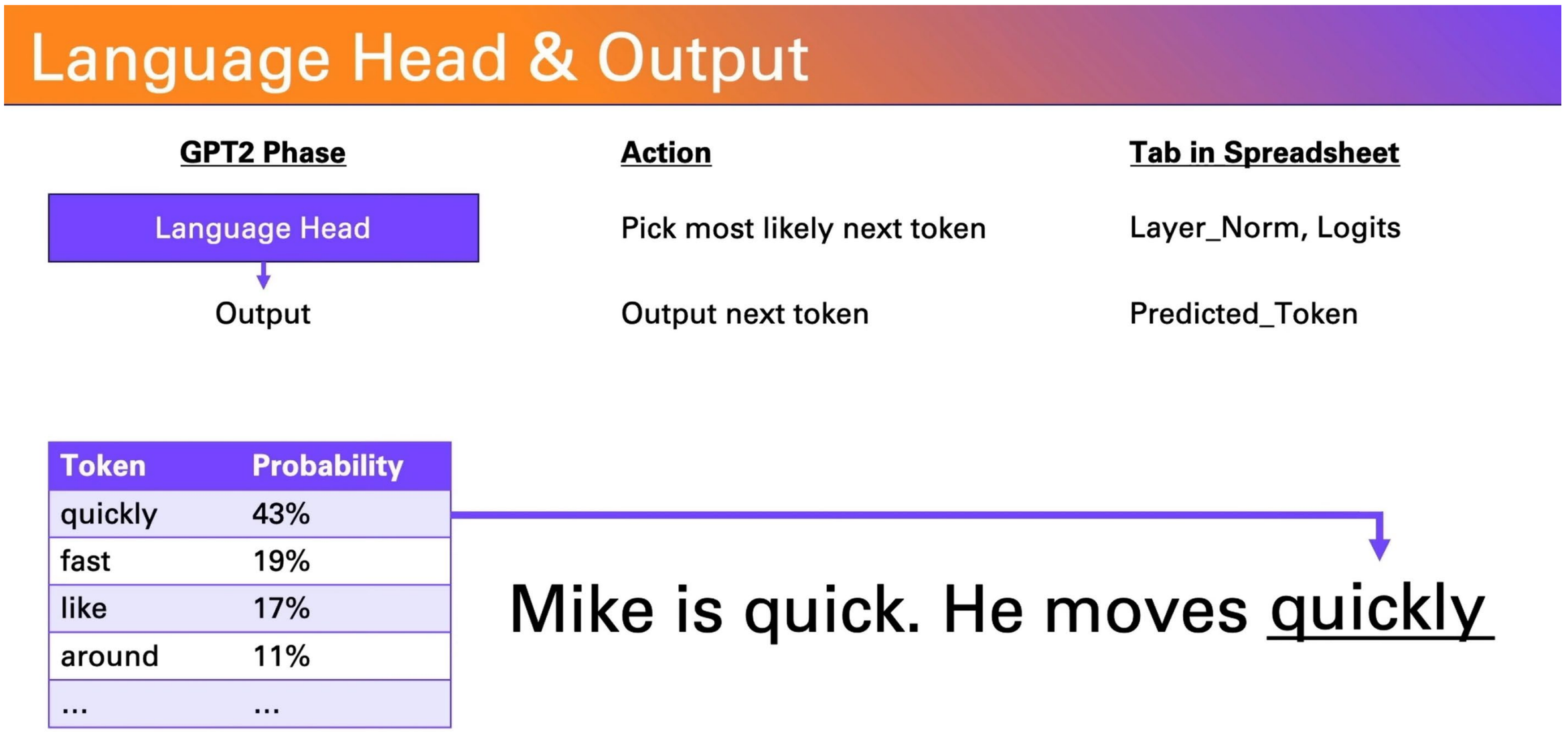
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# Natural Interaction

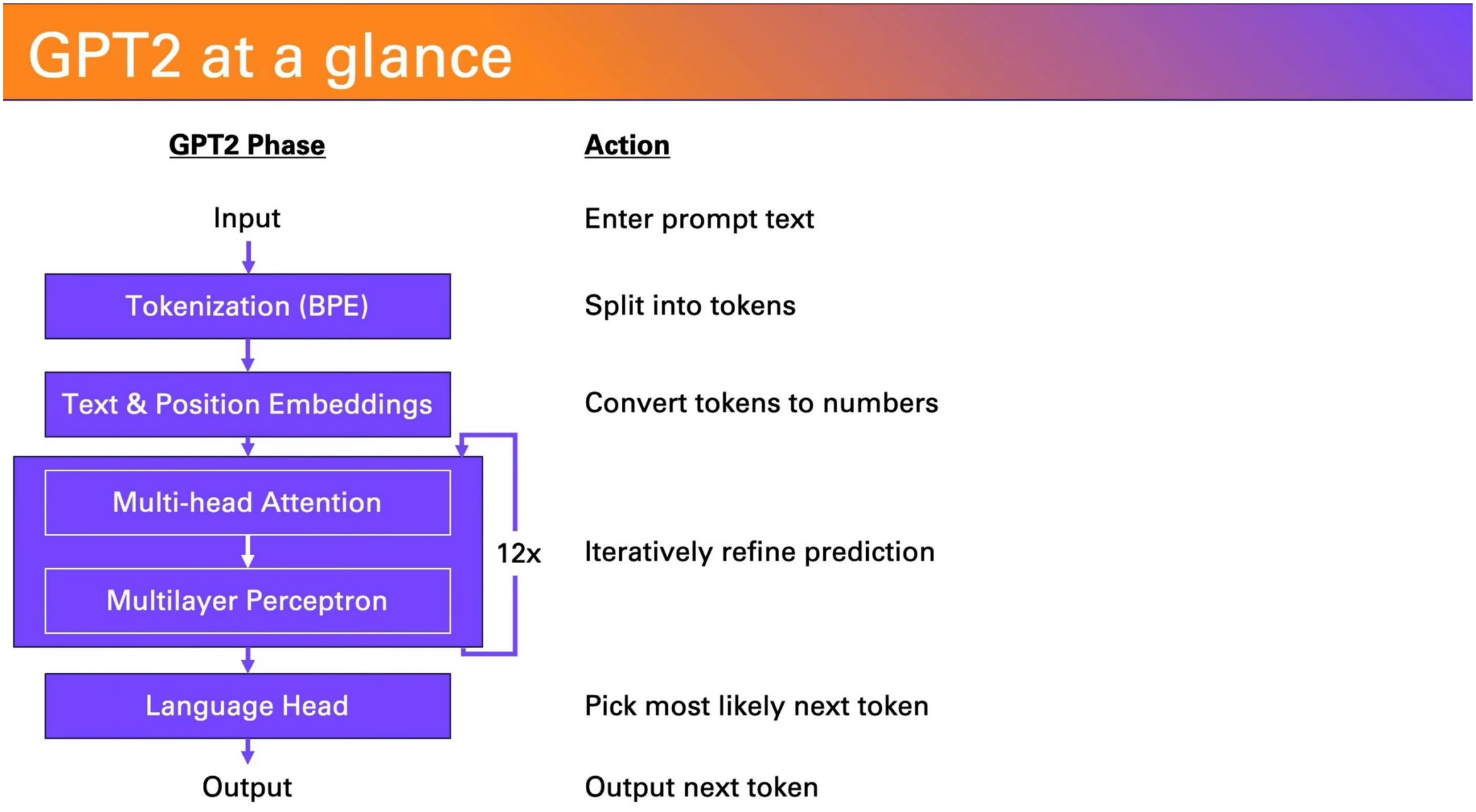
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# Natural Interaction

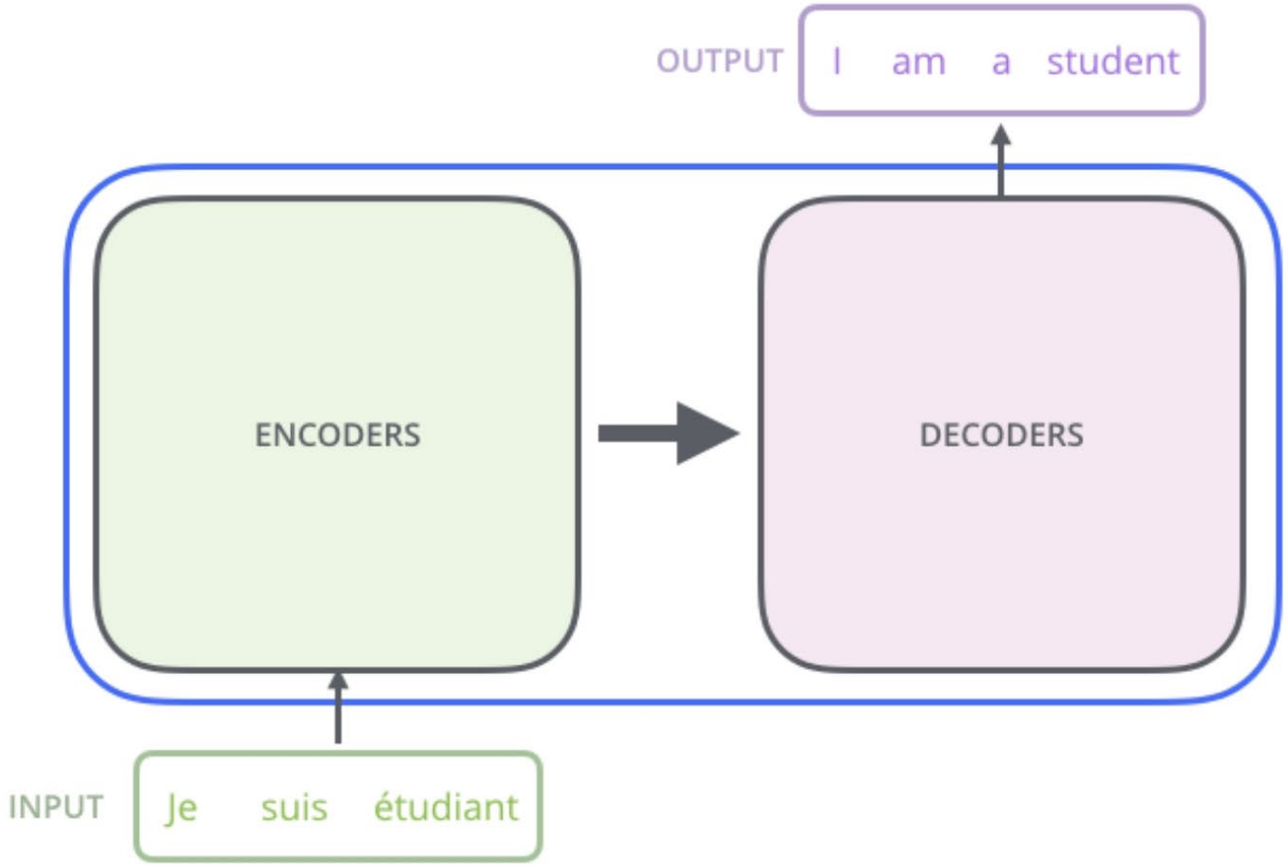
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# Natural Interaction

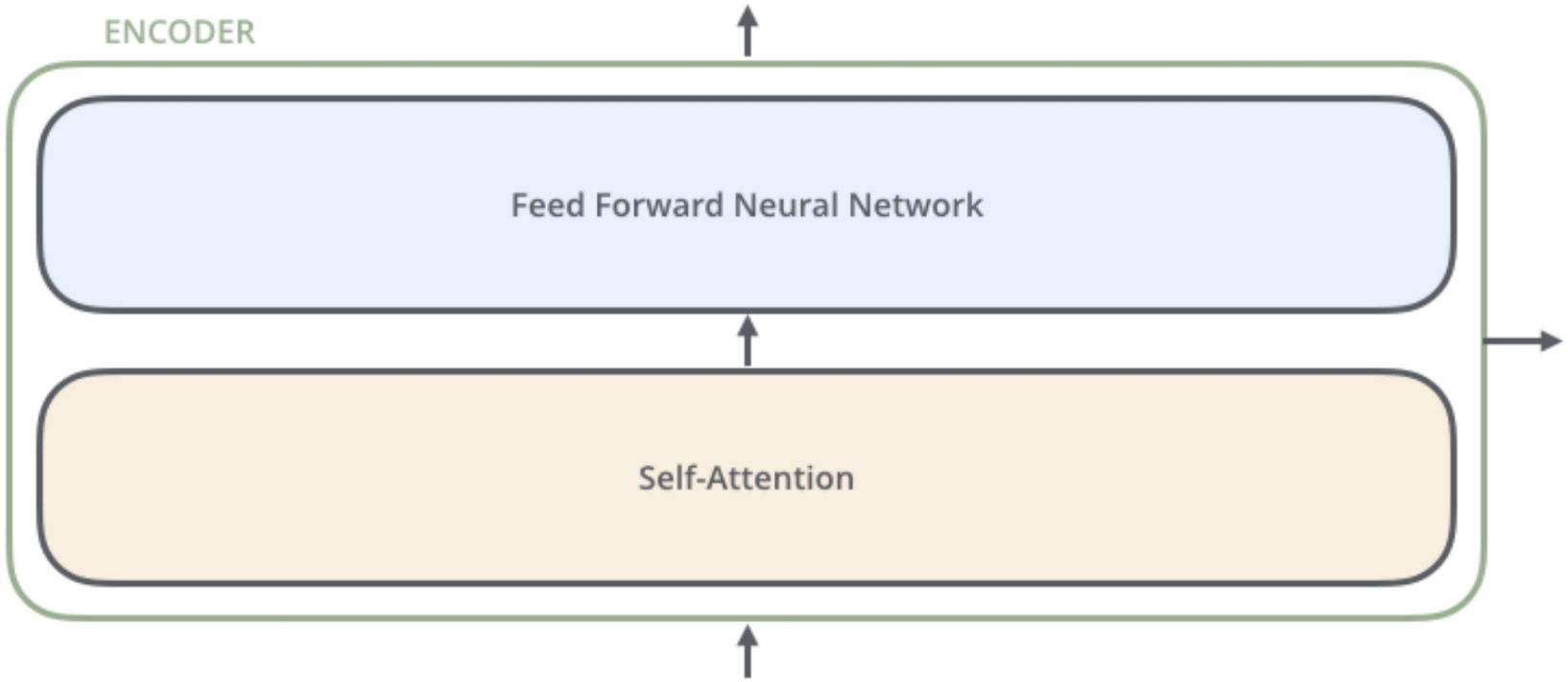

Popping open that Optimus Prime goodness, we see an encoding component, a decoding component, and connections between them.




https://jalammar.github.io/illustrated-transformer/

# Natural Interaction

The encoders are all identical in structure (yet they do not share weights). Each one is broken down into two sub-layers:




The encoder's inputs first flow through a self-attention layer – a layer that helps the encoder look at other words in the input sentence as it encodes a specific word. We'll look closer at self-attention later in the post.

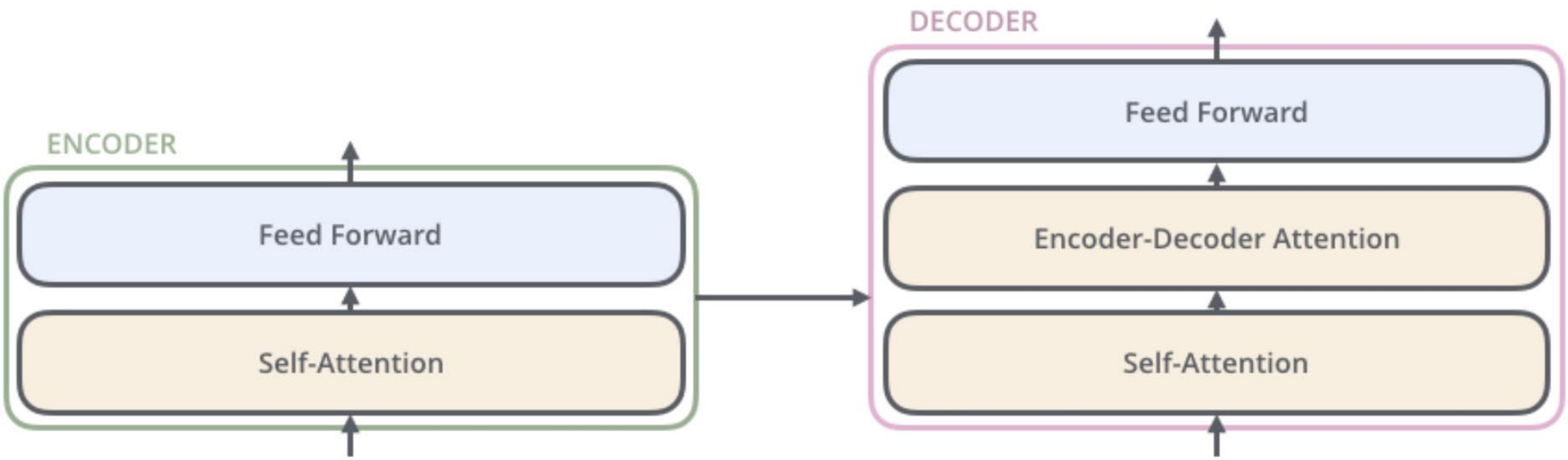
The outputs of the self-attention layer are fed to a feed-forward neural network. The exact same feed-forward network is independently applied to each position.

The decoder has both those layers, but between them is an attention layer that helps the decoder focus on relevant parts of the input sentence (similar what attention does in seq2seq models).

https://jalammar.github.io/illustrated-transformer/

# Natural Interaction

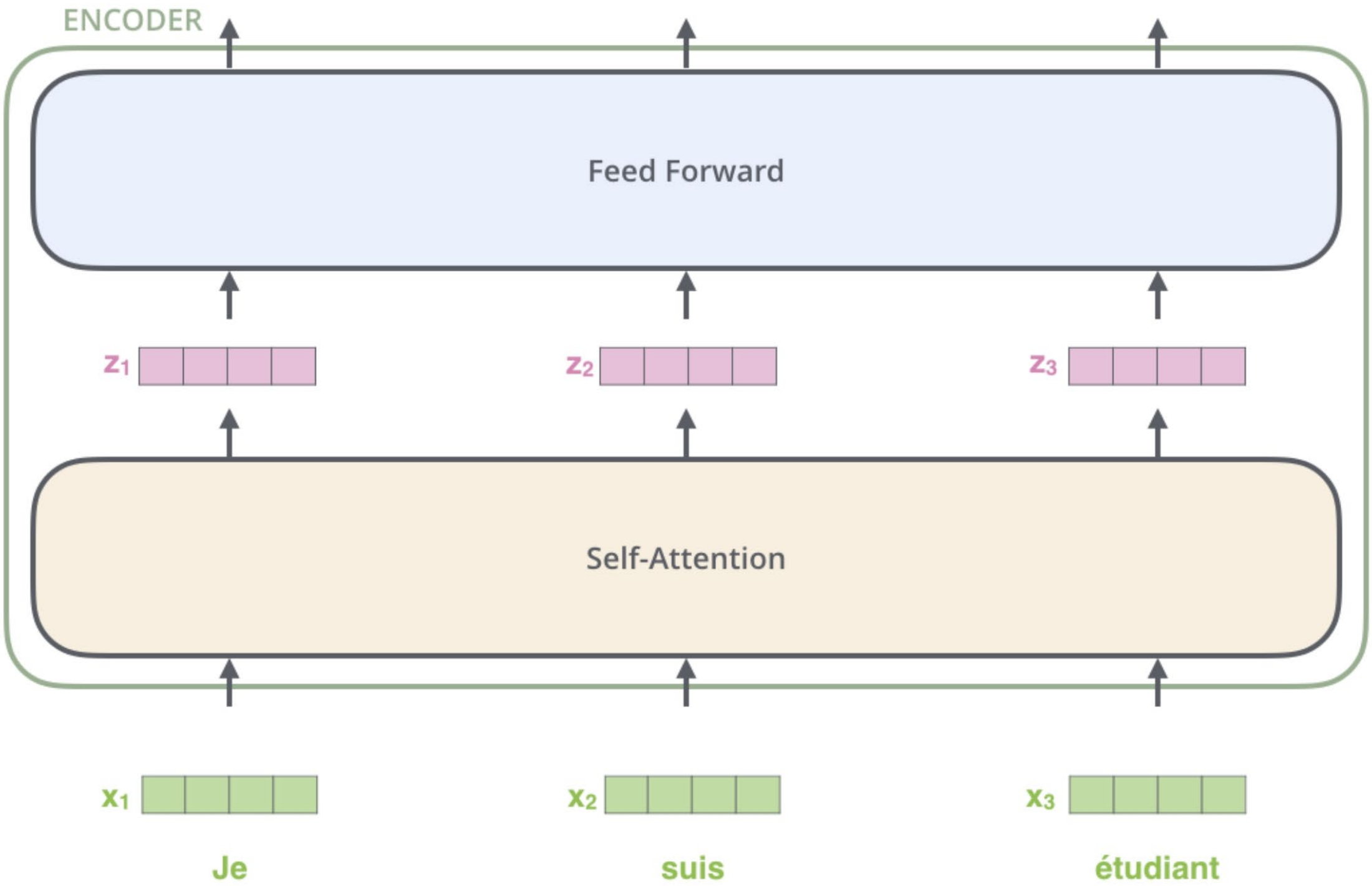
The decoder has both those layers, but between them is an attention layer that helps the decoder focus on relevant parts of the input sentence (similar what attention does in seq2seq models).




https://jalammar.github.io/illustrated-transformer/

# Natural Interaction

After embedding the words in our input sequence, each of them flows through each of the two layers of the encoder.




https://jalammar.github.io/illustrated-transformer/

# Natural Interaction

Say the following sentence is an input sentence we want to translate:

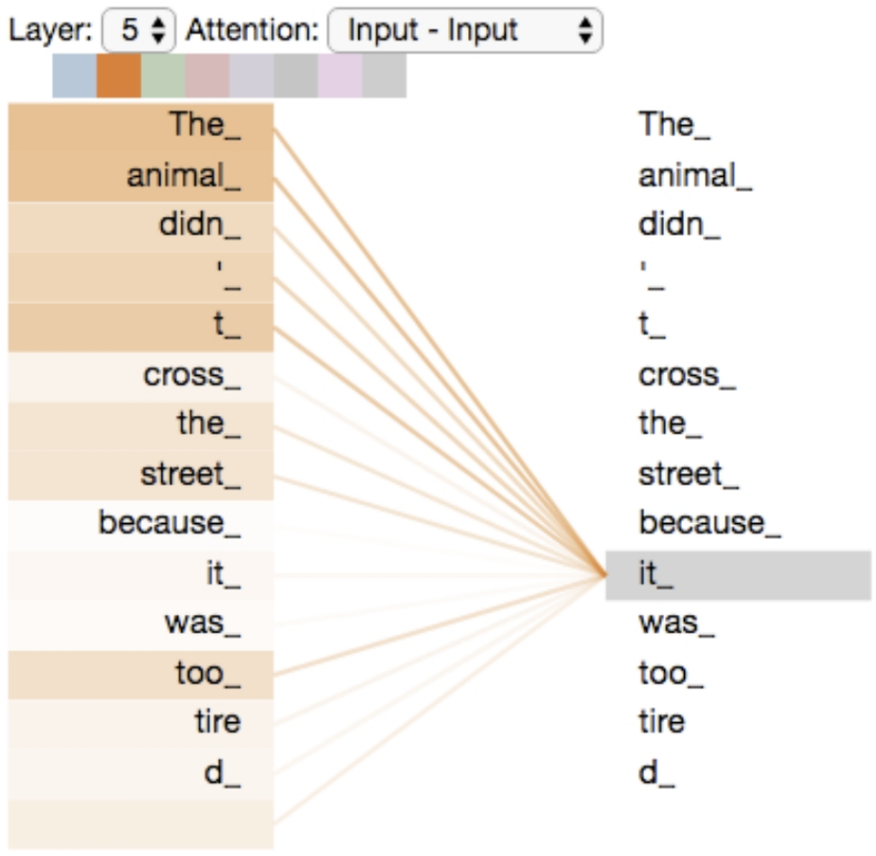
"`The animal didn't cross the street because it was too tired`"

What does “it” in this sentence refer to? Is it referring to the street or to the animal? It’s a simple question to a human, but not as simple to an algorithm.

When the model is processing the word “it”, self-attention allows it to associate “it” with “animal”.

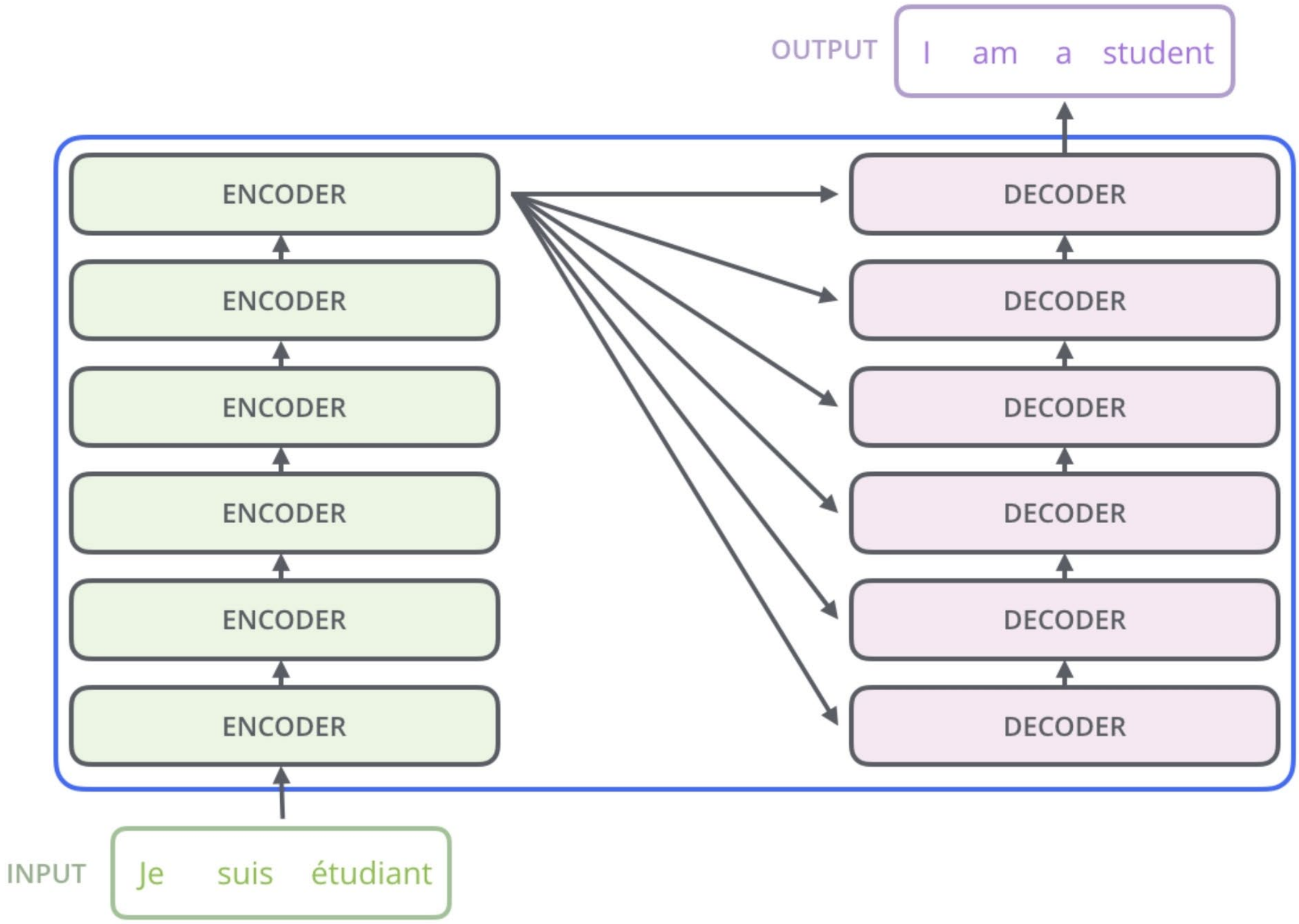
As the model processes each word (each position in the input sequence), self attention allows it to look at other positions in the input sequence for clues that can help lead to a better encoding for this word.

If you’re familiar with RNNs, think of how maintaining a hidden state allows an RNN to incorporate its representation of previous words/vectors it has processed with the current one it’s processing. Self-attention is the method the Transformer uses to bake the “understanding” of other relevant words into the one we’re currently processing.




https://jalammar.github.io/illustrated-transformer/

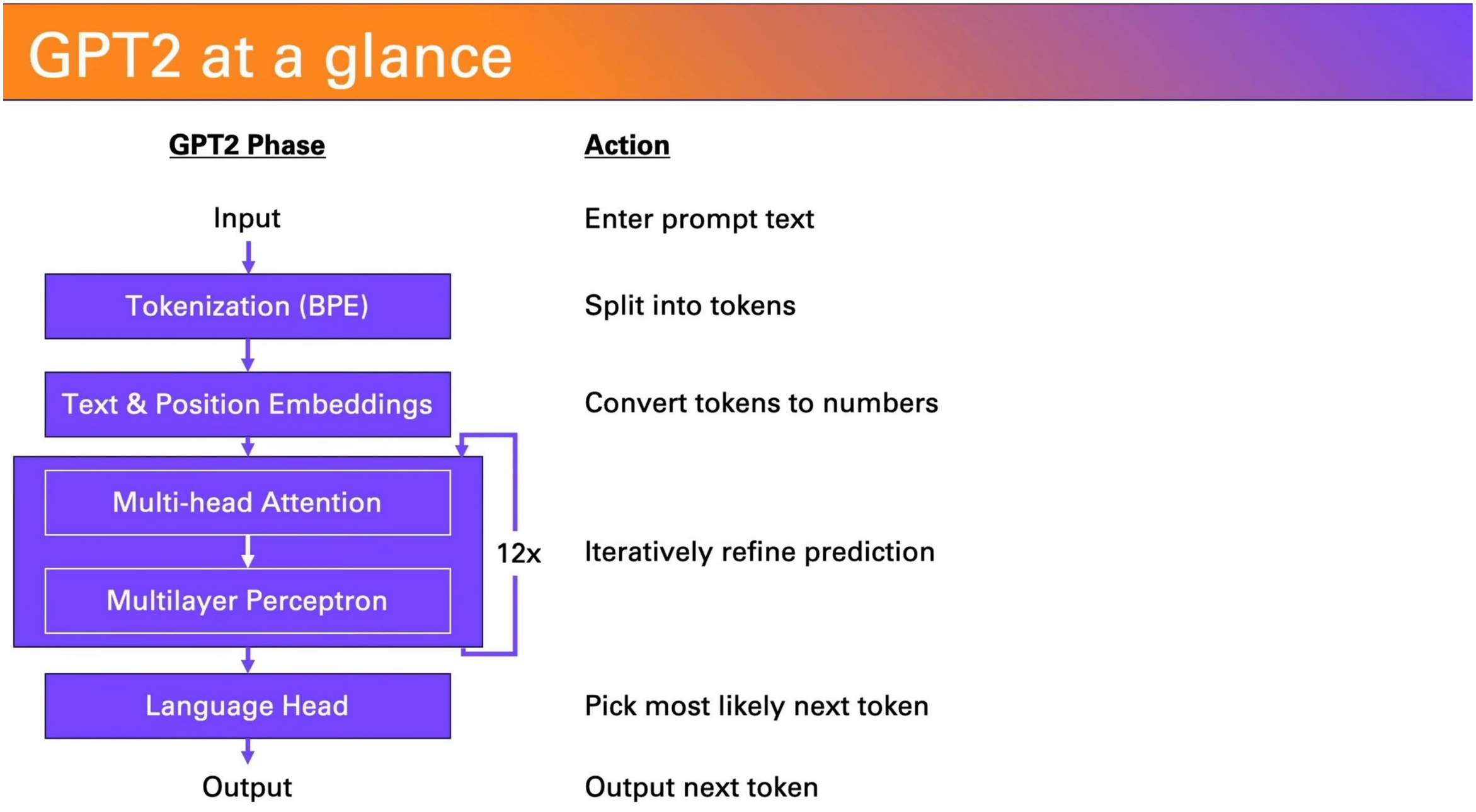
# Natural Interaction




https://jalammar.github.io/illustrated-transformer/

# Natural Interaction

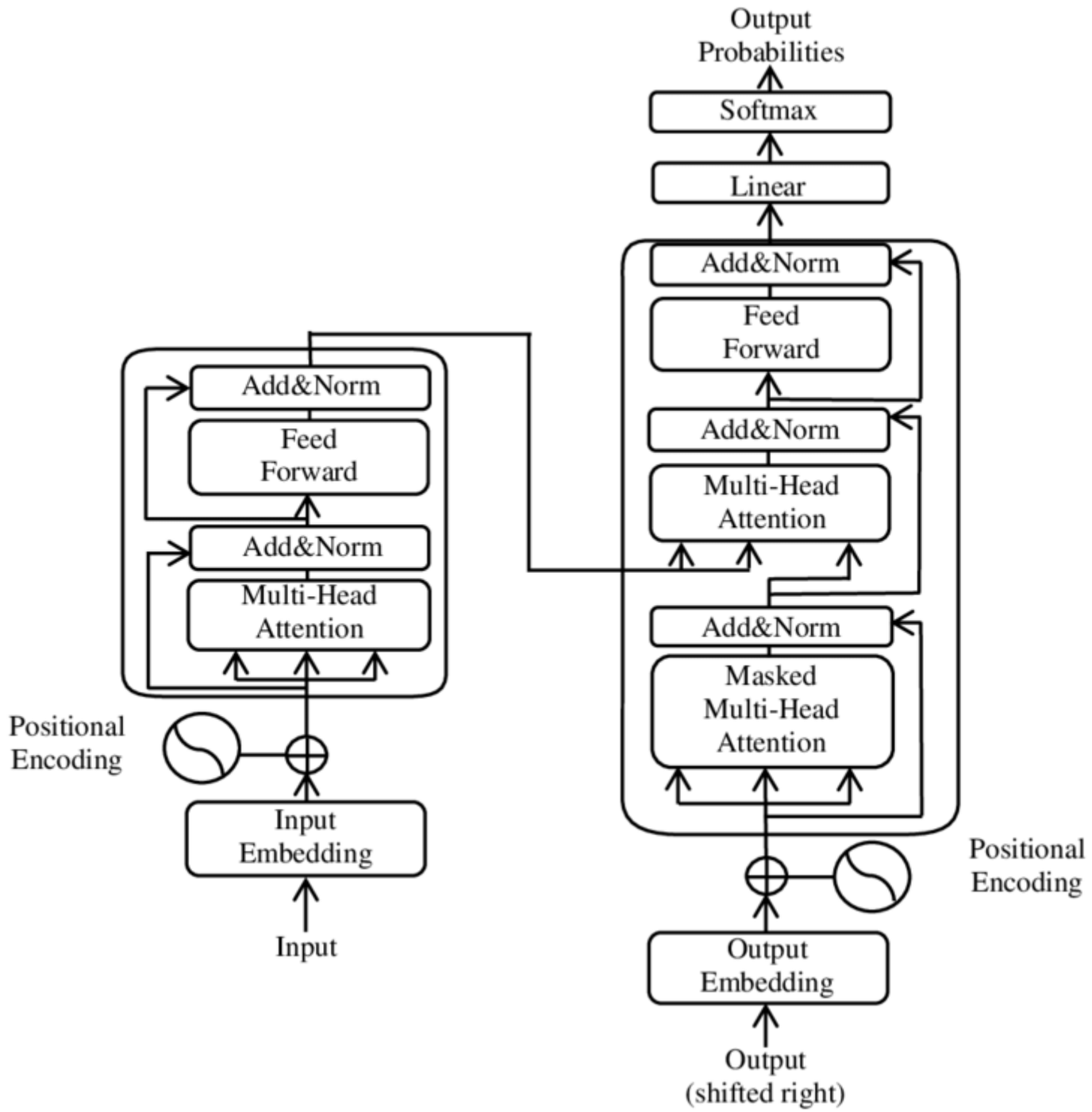
Natural Language Processing (NLP): the ability of AI systems to understand, interpret, and generate human language




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

# 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)

# Natural Interaction

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

**Extracting meaning from text can be hard.**

**Emphasis changes everything.**

**I** never said she stole my money.
*Someone else may have said it, but it wasn't me.*

I **never** said she stole my money.
*I didn't make the claim at any point in time.*

I never **said** she stole my money.
*I may have implied or thought it, but didn't say it.*

I never said **she** stole my money.
*Someone else may have stolen it, but it wasn't her.*

I never said she **stole** my money.
*She may have borrowed or been given it.*

I never said she stole **my** money.
*She stole someone else's money.*

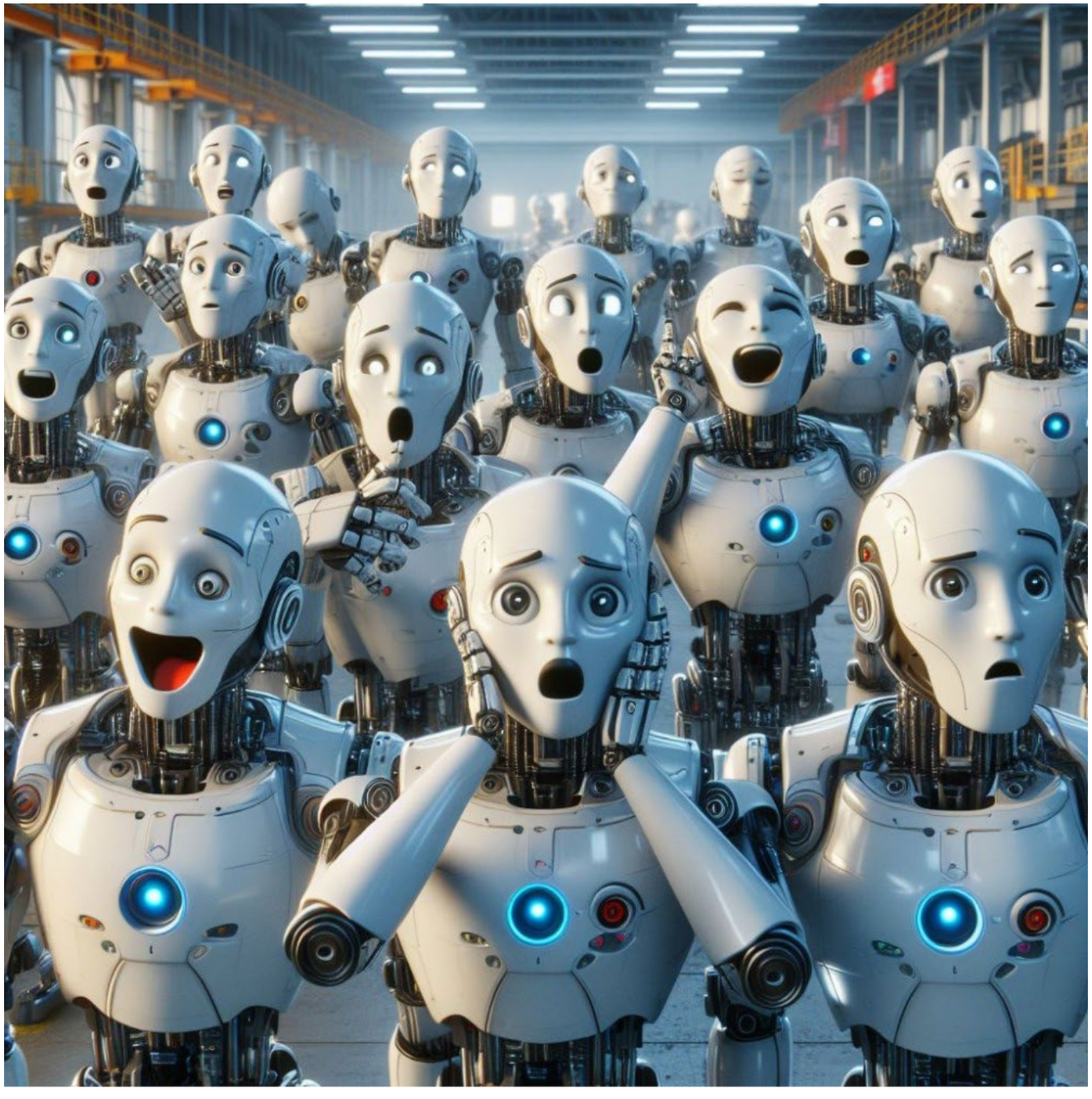
I never said she stole my **money.**
*She may have stolen something else.*

Brian Sacash

Source: Brian Sacash (Internet)

# Natural Interaction

Affective computing: technologies that can recognise, interpret, process and simulate human emotions



https://www.linkedin.com/pulse/ai-uprising-when-robots-ask-seat-table-aditi-kapoor-n8r0f

# Natural Interaction

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

https://www.youtube.com/watch?v=wGWVKkYEHBE
60 Minutes reporter Tom Steinfort interviews Ameca, an AI robot

# 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

# Discussion

?

# Break

?

# Natural Interaction

Ethical concerns:

- bias and discrimination
  - voice recognition systems may perform poorly for certain accents or dialects
  - facial recognition systems may have higher error rates for people with darker skin tones
- surveillance and monitoring
  - facial recognition technology used in public spaces without consent
  - employers using AI to monitor employee activities and communications
- transparency and trust
  - users not knowing whether they are interacting with a human or an AI
  - lack of transparency about how personal data is used and stored
- privacy
  - virtual assistants that record and analyse conversations, potentially capturing sensitive information
  - cameras and microphones in smart devices continuously monitoring user activities

ChatGPT 4o

# Natural Interaction

Ethical concerns:

- misuse of technology
  - deepfake technology used to create realistic but fake videos or audio recordings
  - social engineering attacks using AI to mimic trusted voices or personas
- autonomy and manipulation
  - virtual assistants or chatbots designed to persuade users to make purchases or share personal information
  - AI systems used in political campaigns to influence voter behavior
- psychological impact
  - users developing emotional attachments to AI companions or virtual assistants
  - reduced face-to-face interactions due to reliance on AI communication tools
- accessibility and inclusion
  - speech recognition systems may struggle with non-standard speech patterns or disabilities
  - interfaces may not be accessible to visually or hearing-impaired users

ChatGPT 4o

# Natural Interaction

Mitigation of ethical concerns:

- bias and discrimination
  - using diverse and representative datasets for training AI models
  - regularly auditing AI systems for bias and implementing fairness-aware algorithms
  - engaging diverse teams in the development and testing of AI technologies
- surveillance and monitoring
  - establishing clear regulations and policies to govern the use of surveillance technologies
  - balancing surveillance needs with respect for individual privacy rights communications
- transparency and trust
  - developing explainable AI (XAI) techniques to provide clear and understandable explanations of AI decisions and actions
  - ensuring transparency in data collection, processing (monitoring), and usage practices
  - building trust through user education and communication about AI capabilities and limitations
- privacy
  - implementing strong data protection measures, such as encryption and anonymisation
  - providing users with control over their data and options to delete stored information

ChatGPT 4o

# Natural Interaction

Mitigation of ethical concerns:

- misuse of technology
  - developing and enforcing ethical guidelines and regulations for natural interaction technologies
  - implementing technical safeguards to detect and prevent misuse (deepfake detection algorithms)
  - educating the public and stakeholders about the risks and responsible use of AI technologies
- autonomy and manipulation
  - establishing ethical guidelines to prevent manipulative practices in AI design
  - ensuring users are aware when they are interacting with AI and its potential influence
  - promoting user autonomy by providing choices and respecting user preferences
- psychological impact
  - designing AI systems to complement, rather than replace, human interactions
  - encouraging balanced use of AI technologies and promoting healthy digital habits
  - conducting research to understand the long-term psychological impacts of AI interactions
- accessibility and inclusion
  - designing AI systems with accessibility in mind, following inclusive design principles
  - regularly testing and updating systems to address accessibility issues

ChatGPT 4o

# Questions

?

# Homework

- check the information sources provided on different slides (at least some)
- “A low-code way to learn AI. Learn how AI works from a real LLM implemented entirely in Excel”:
  https://spreadsheets-are-all-you-need.ai/index.html#watch-the-lessons
- please try to watch the first few video lessons before the next lecture, and prepare to give a few brief comments about the video(s):
  - what was most striking about the AI technology in the video(s)?
  - what societal impact did the AI make?
  - what were the main AI limitations, biases and risks?

**?**