

Workshop Machine Learning

26 maart 2025

Over mij



Aswin van Woudenberg
NHL Stenden Hogeschool
Toegepaste Wiskunde

Freelance ontwikkelaar/data scientist
reachme@aswinvanwoudenberg.com

Inhoud

Artificial Intelligence versus Machine Learning

Uit welke stappen bestaat een machine learning project?

De gereedschapskist van een data scientist

Aan de slag!

Supervised, unsupervised en reinforcement learning

Gestructureerde en ongestructureerde data

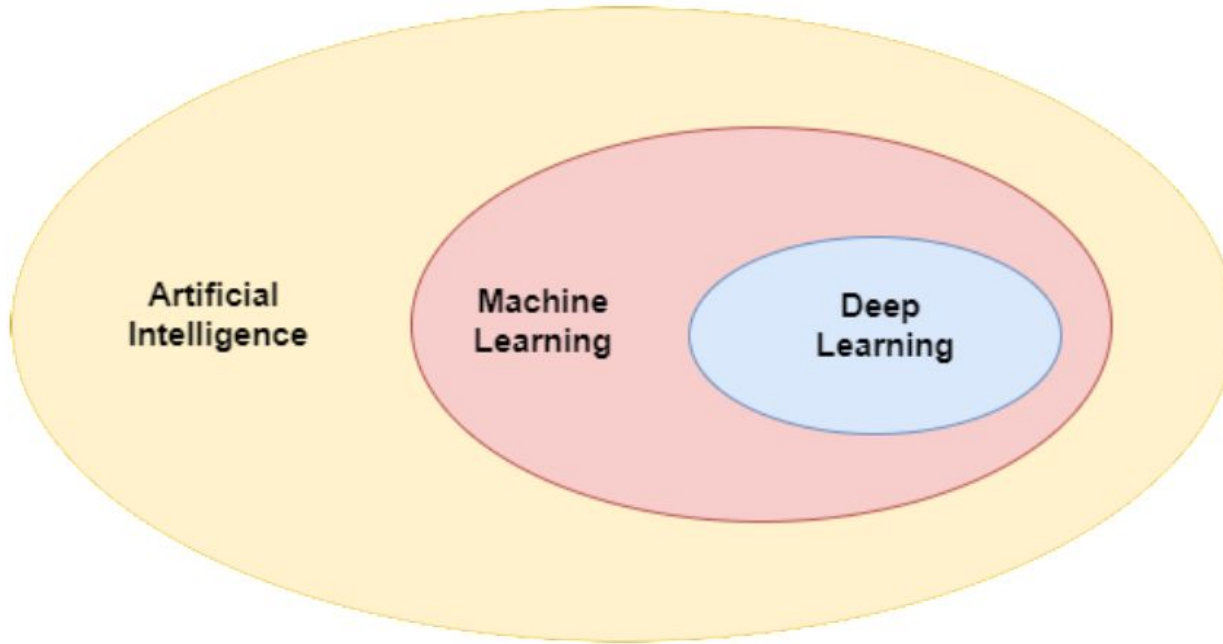
Perceptrons, neurale netwerken en deep learning

Large Language Models (met *structured output* en *function calling*)

Machine learning in robotica

Hoe word ik een pro?

Artificial Intelligence versus Machine Learning



Artificial Intelligence

The science and engineering of making intelligent machines.

- John McCarthy





WHY DID YOU DROP IT?

BT "IT": I ASSUME YOU MEAN THE
SUPERBLOCK.
I CAN'T EXPLAIN A NON-EXISTENT EVENT.



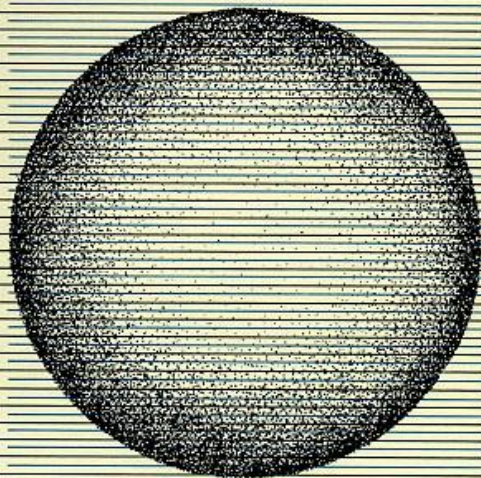
Buchanan
Shortliffe

**RULE-
BASED
EXPERT
SYSTEMS**

RULE-BASED EXPERT SYSTEMS

THE **MYCIN** EXPERIMENTS OF THE STANFORD
HEURISTIC PROGRAMMING PROJECT

Bruce G. Buchanan
Edward H. Shortliffe



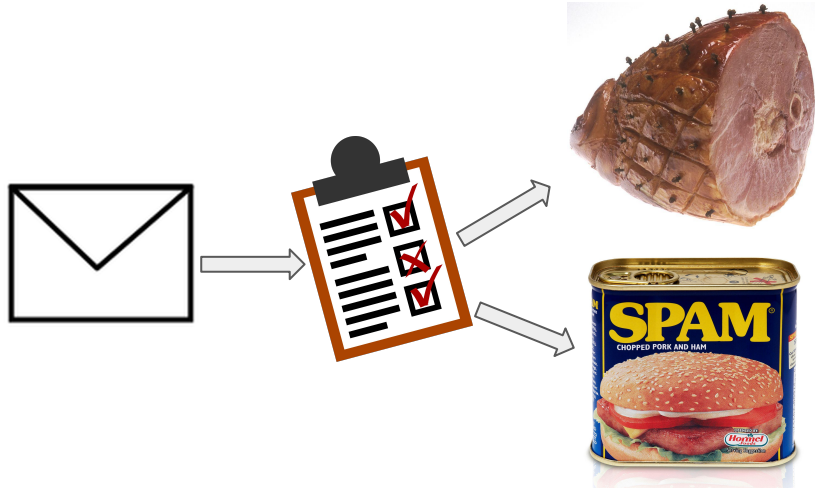
Machine Learning

Field of study that gives computers the ability to learn without being explicitly programmed.

- Arthur Samuel



Voorbeeld van AI: E-mail SPAM filter



Eerste SPAM-filters:

- Handmatig geprogrammeerde set regels
 - Bevatten bepaalde woorden.
 - E-mails van een lijst met verdachte IP-adressen.
 - Etc.
- Voor elke update was een programmeur nodig om de code aan te passen.

Een op regels gebaseerd SPAM-filter bootst een mens na die verdachte e-mails herkent, dus het valt onder AI, maar het is geen Machine Learning.

Voorbeeld van ML: E-mail SPAM-filter

In de moderne versie van een Machine Learning SPAM-filter:

- Een dataset van e-mails, gelabeld door mensen als SPAM of geen SPAM, wordt opgebouwd.
- Deze dataset wordt ingevoerd in een machine learning-algoritme dat een model traint.
- Wanneer gebruikers SPAM rapporteren die door het filter is gekomen, kunnen e-mails worden toegevoegd aan de dataset en kan het model opnieuw getraind worden.

19, 9:08 AM (3 days ago)



Reply



Forward

Filter messages like this

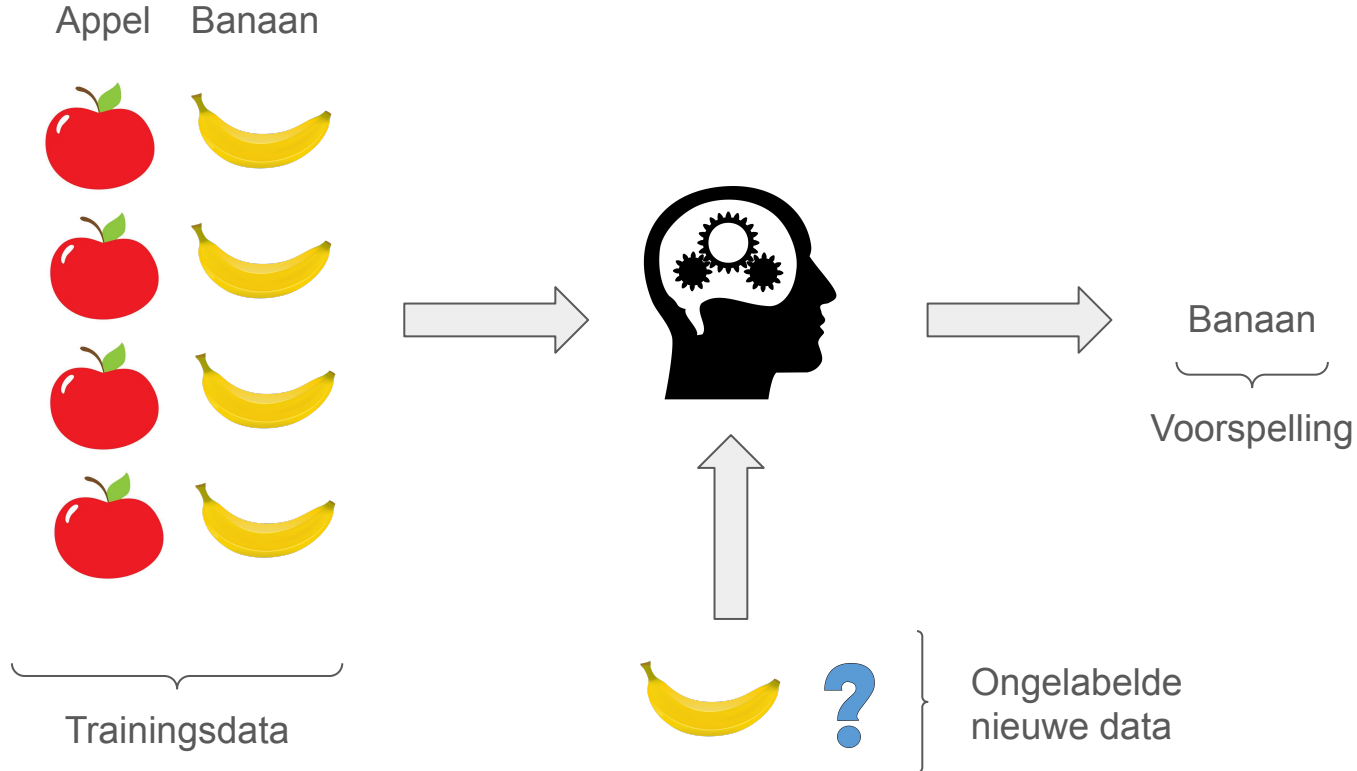
Print

Delete this message

Report spam

Report phishing

Het trainen van een model



Wat doet een machine learning algoritme?

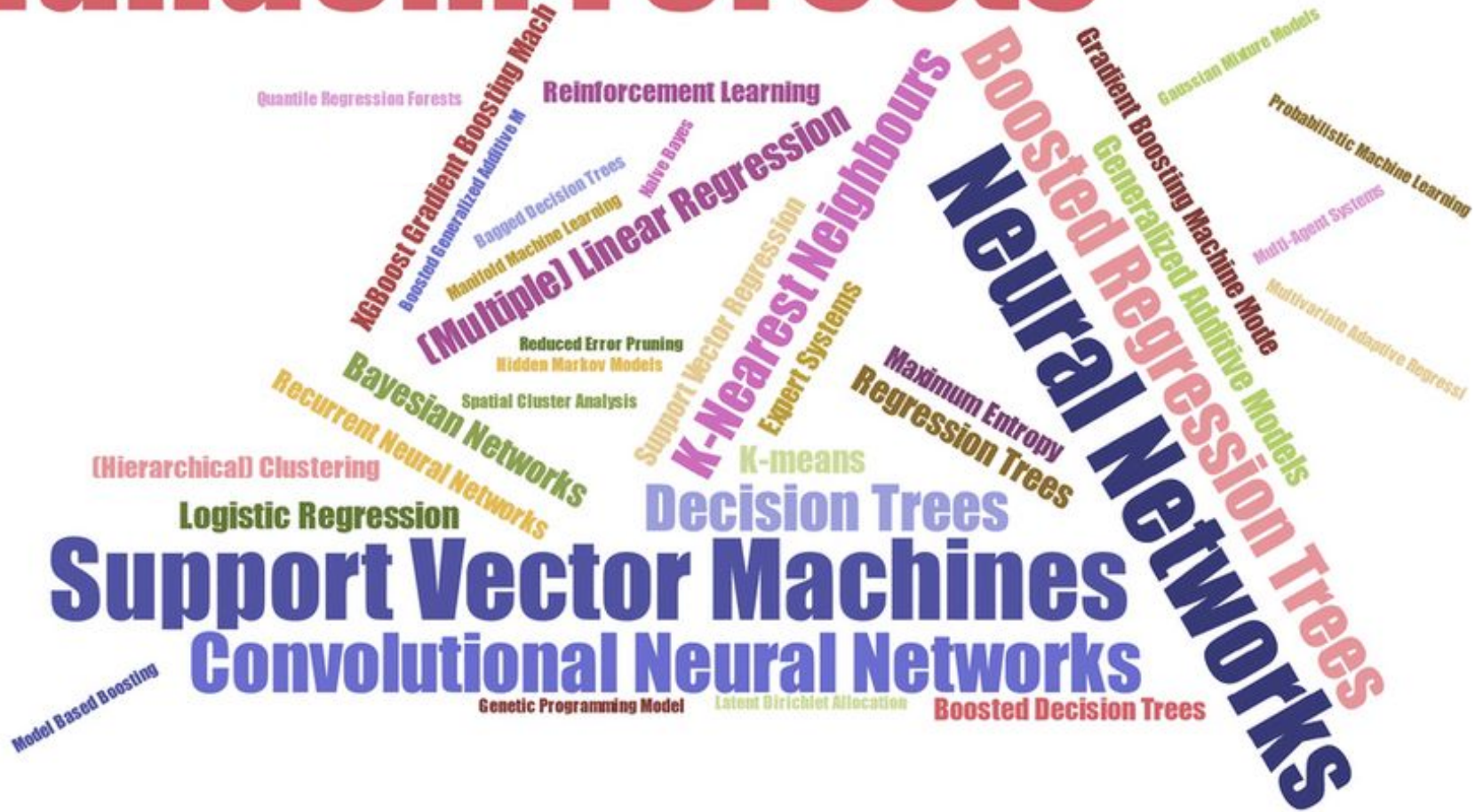
Machine Learning-algoritmen proberen een functie (f) te leren die de relatie beschrijft tussen invoervariabelen (ook wel een **featurevector**, X) en een **doelvariabele** (Y).

De relatie wordt wiskundig weergegeven als:

$$Y = f(X)$$

Eenmaal geleerd (of getraind), wordt deze functie aangeduid als het **machine learning-model**. Het model wordt vervolgens gebruikt om voorspellingen te doen.

Random Forests



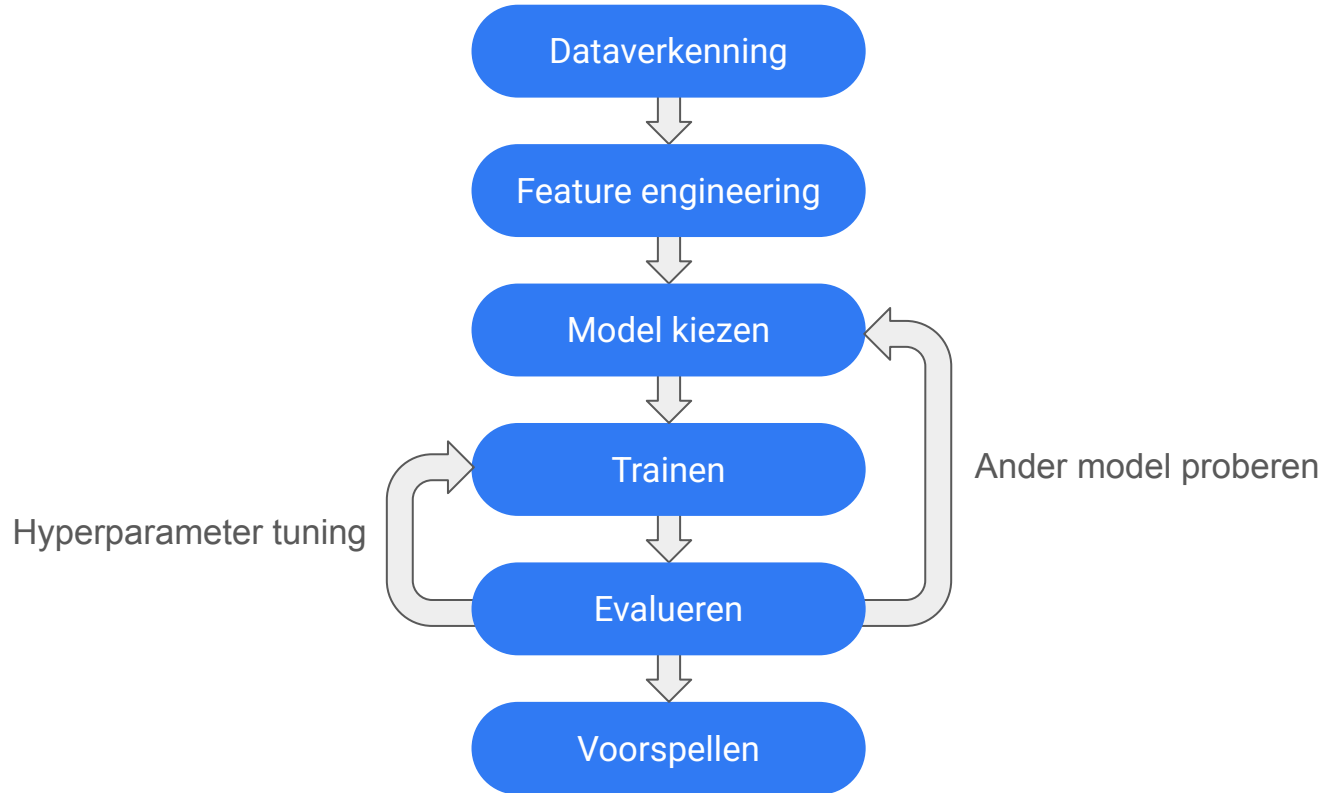
Regressie versus classificatie

De meeste modellen vallen in een van de volgende categorieën:

Regressie wordt gebruikt om een continue numerieke waarde te voorspellen, zoals de prijs van een huis of de temperatuur morgen.

Classificatie wordt gebruikt om gegevens in categorieën in te delen, zoals e-mails als "SPAM" of "geen SPAM".

De machine learning workflow



De gereedschapskist van een data scientist

Belangrijke Python-tools voor Machine Learning:

Jupyter Notebook: Interactieve programmeeromgeving.

scikit-learn: Machine learning-algoritmen.

pandas: Gegevensmanipulatie en -analyse.

NumPy: Numerieke berekeningen met arrays.

Matplotlib & Seaborn: Datavisualisatie.

TensorFlow & PyTorch: Deep learning-frameworks.

Aan de slag!

Download de notebooks hier:

<https://rb.gy/jqzz4j>

Pak het zip bestand uit en upload de notebooks hier:

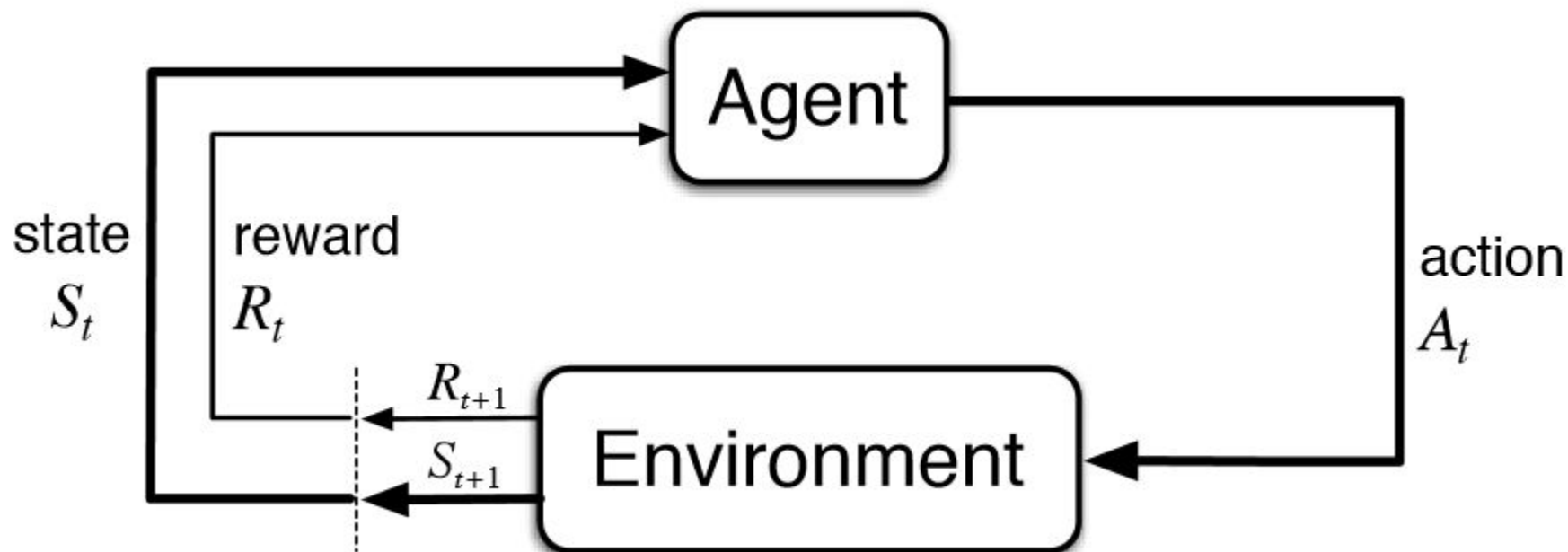
<http://drive.google.com>

Soorten machine learning

In machine learning wordt onderscheid gemaakt tussen de volgende typen:

- Supervised learning
 - Regressie
 - Classificatie
- Unsupervised learning
 - Clustering
 - Dimensionality reduction
- Reinforcement learning
 - Leert door interactie met omgeving

Reinforcement learning



Gestructureerde versus ongestructureerde data

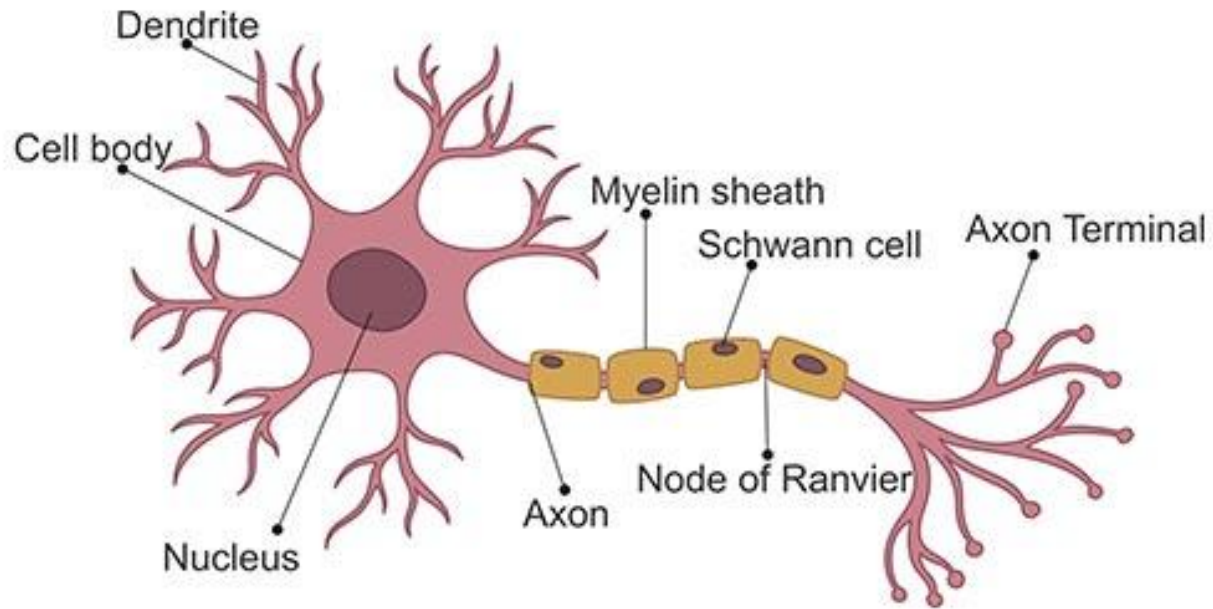
- Gestructureerde data

- Gestructureerde data is georganiseerd in een duidelijk formaat, zoals tabellen met rijen en kolommen.
- Voorbeelden: spreadsheets, relationele databases.
- Geschikt voor traditionele ML-algoritmen.

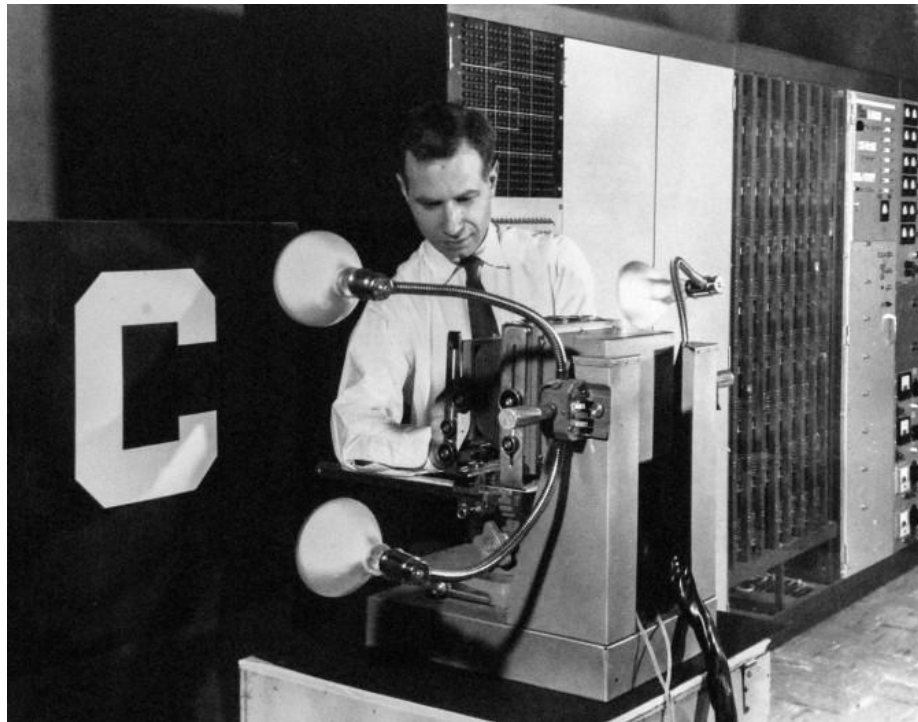
- Ongestructureerde data

- Ongestructureerde data heeft geen vast formaat of organisatie.
- Vaak *high-dimensional*
- Voorbeelden: tekstbestanden, afbeeldingen, audio, video.
- Vereist geavanceerdere ML-technieken.

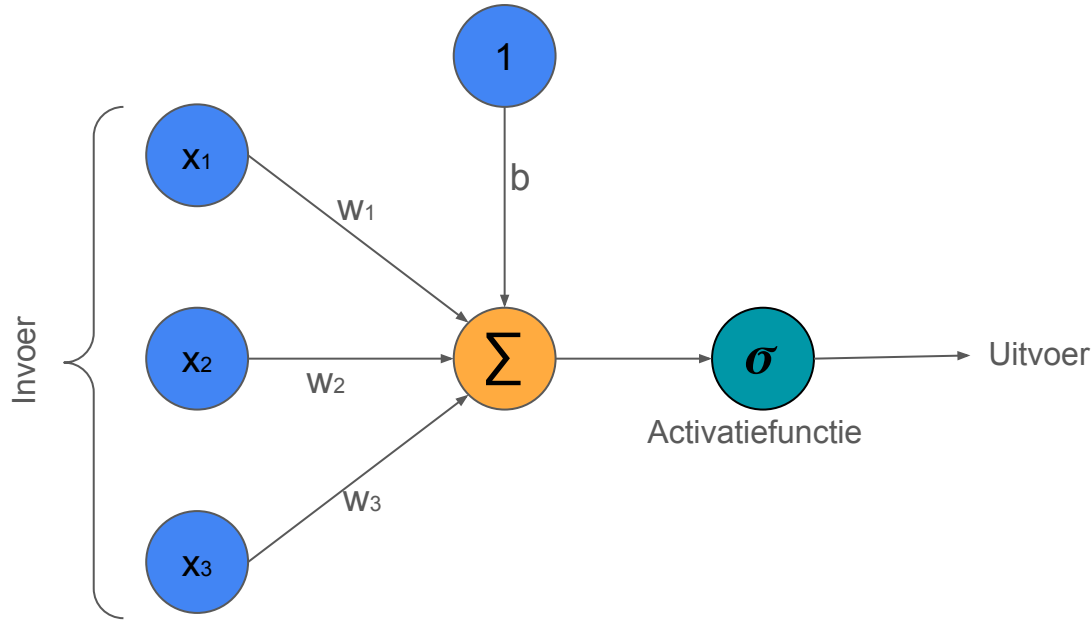
Neuron



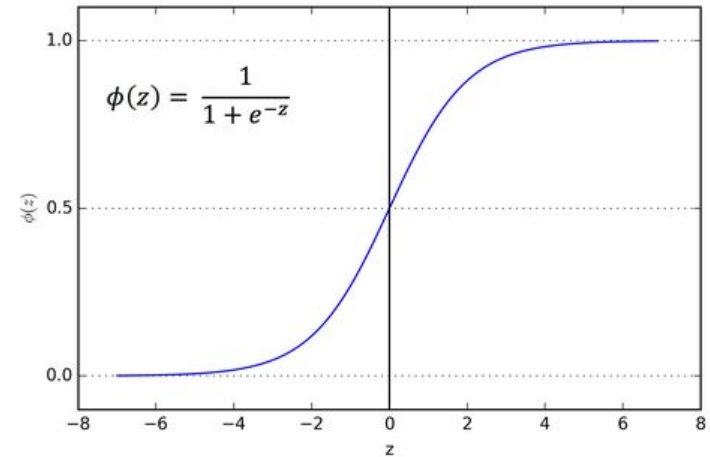
Frank Rosenblatt



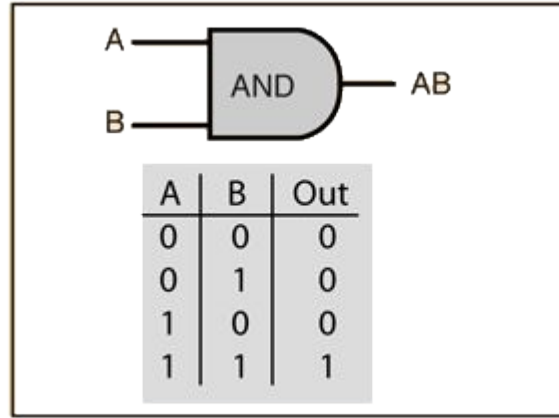
Perceptron



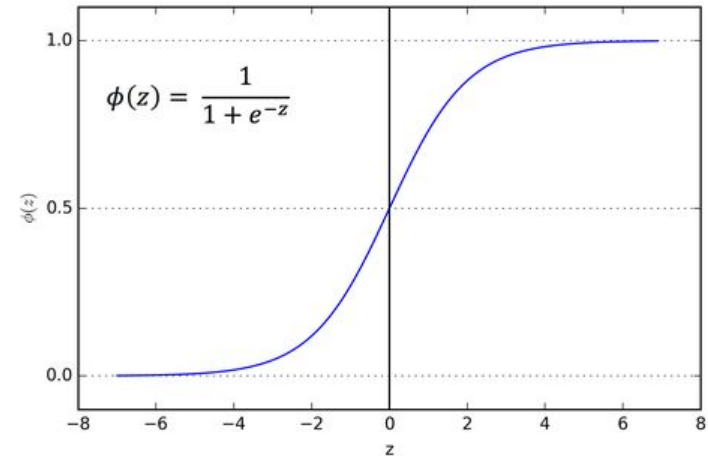
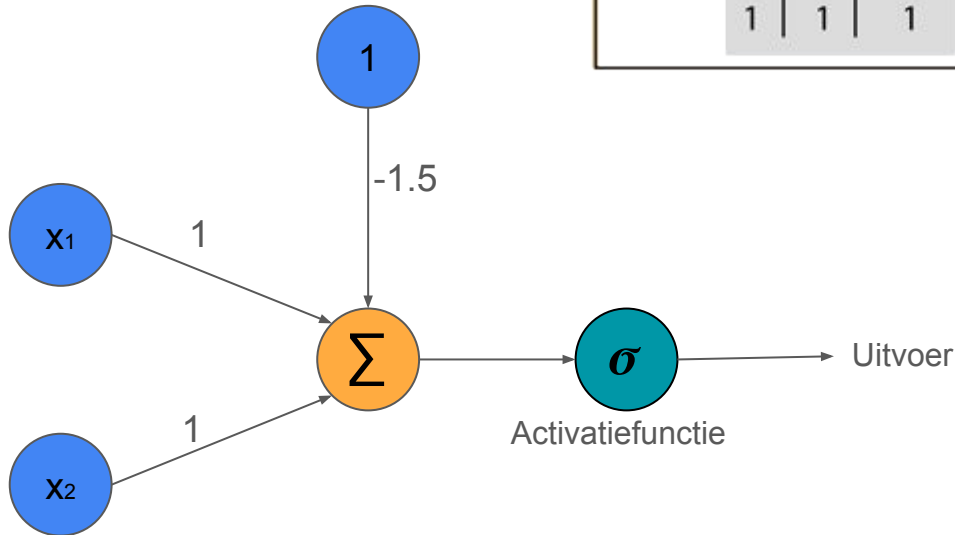
$$z = \sum_i w_i \cdot x_i + b$$



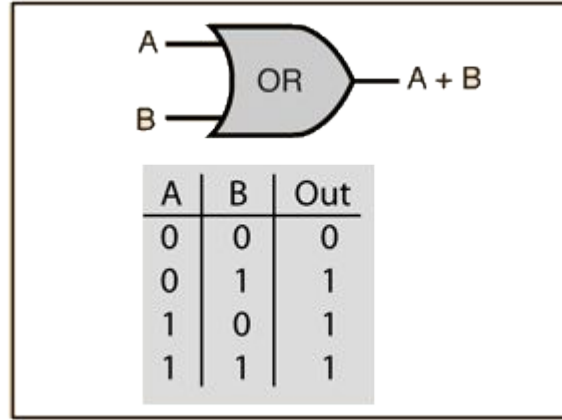
Perceptron: AND



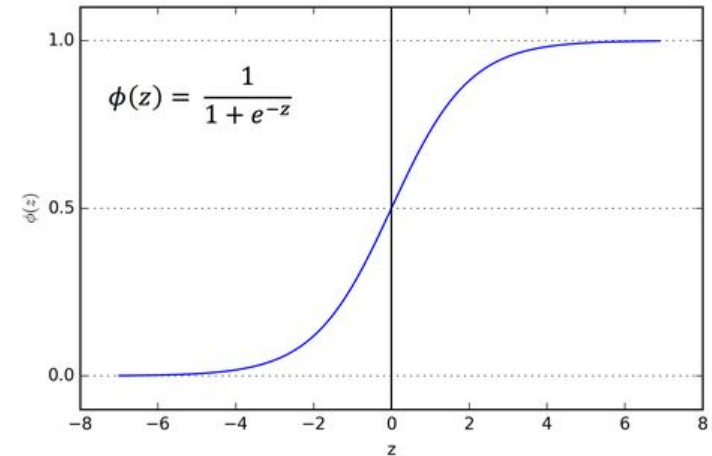
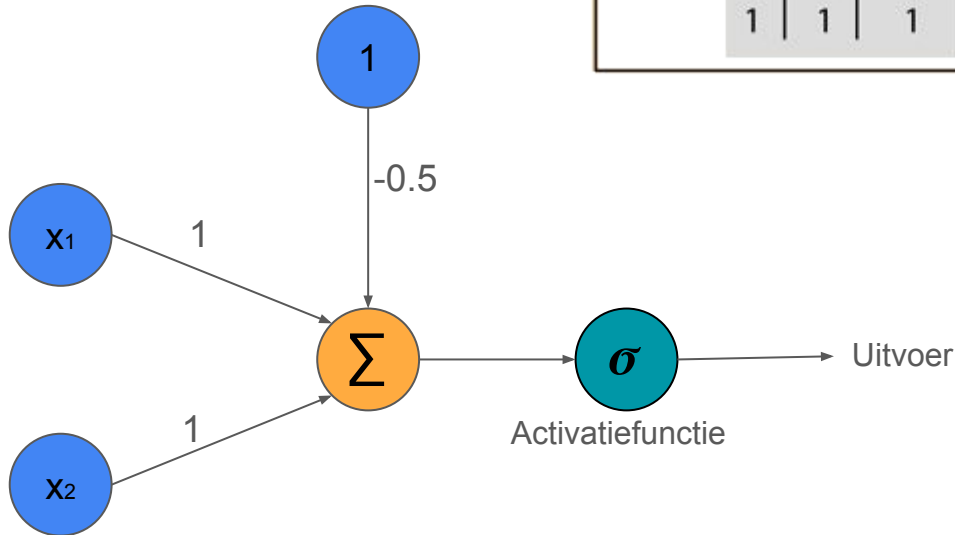
$$z = \sum_i w_i \cdot x_i + b$$



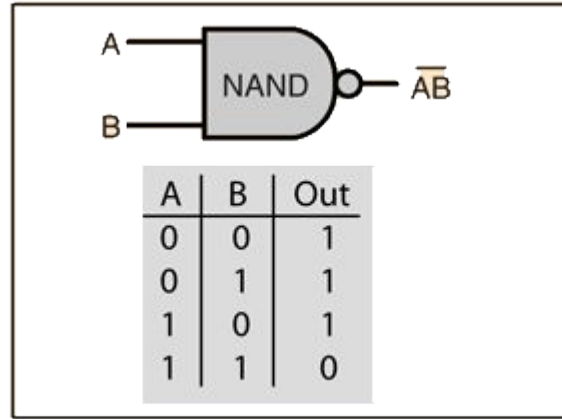
Perceptron: OR



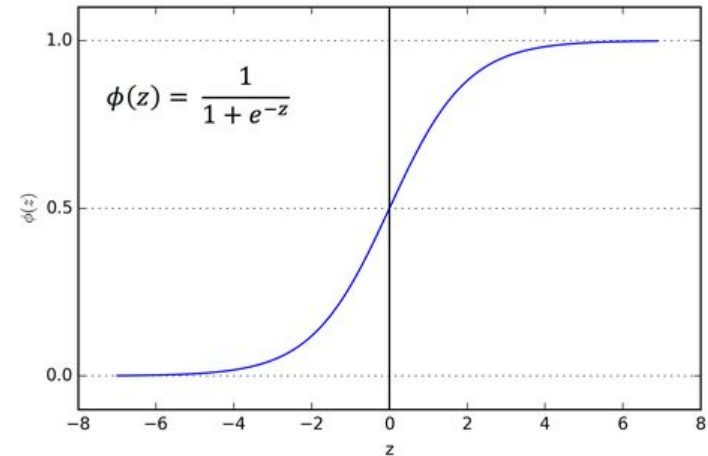
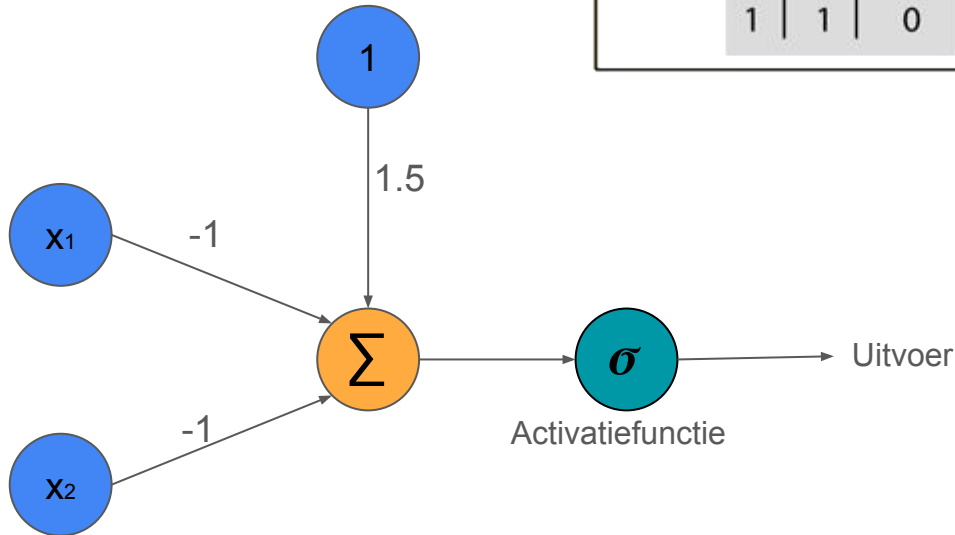
$$z = \sum_i w_i \cdot x_i + b$$



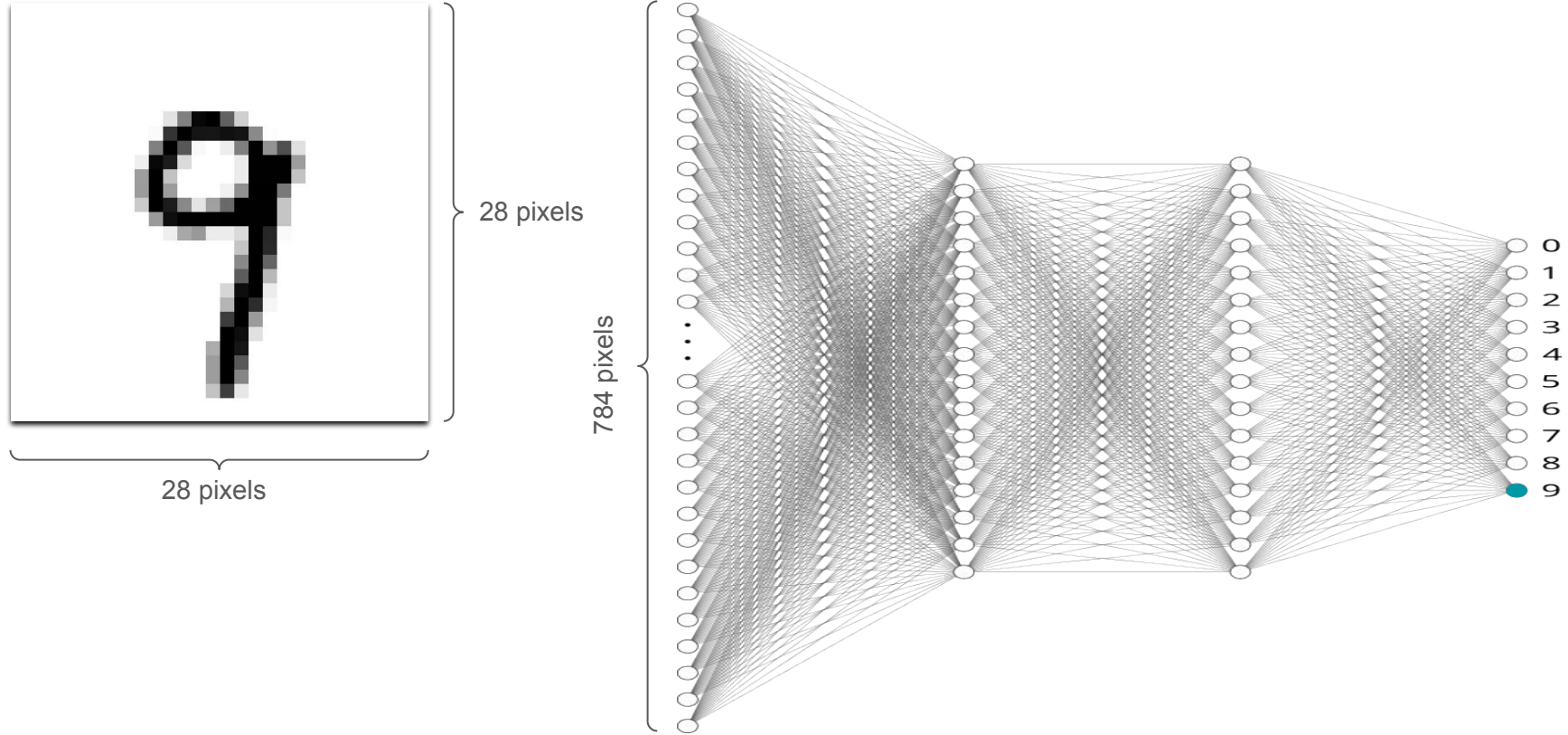
Perceptron: NAND



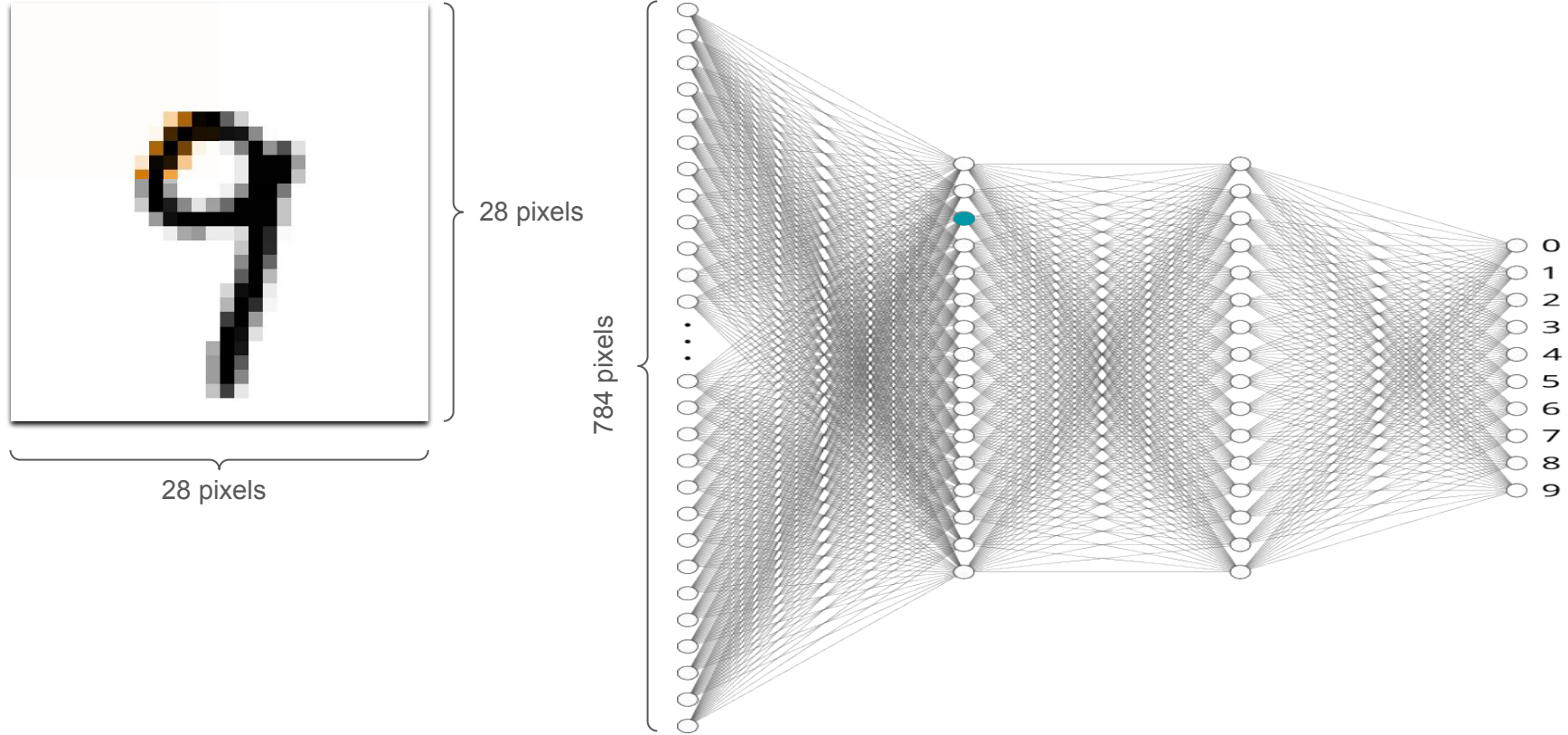
$$z = \sum_i w_i \cdot x_i + b$$



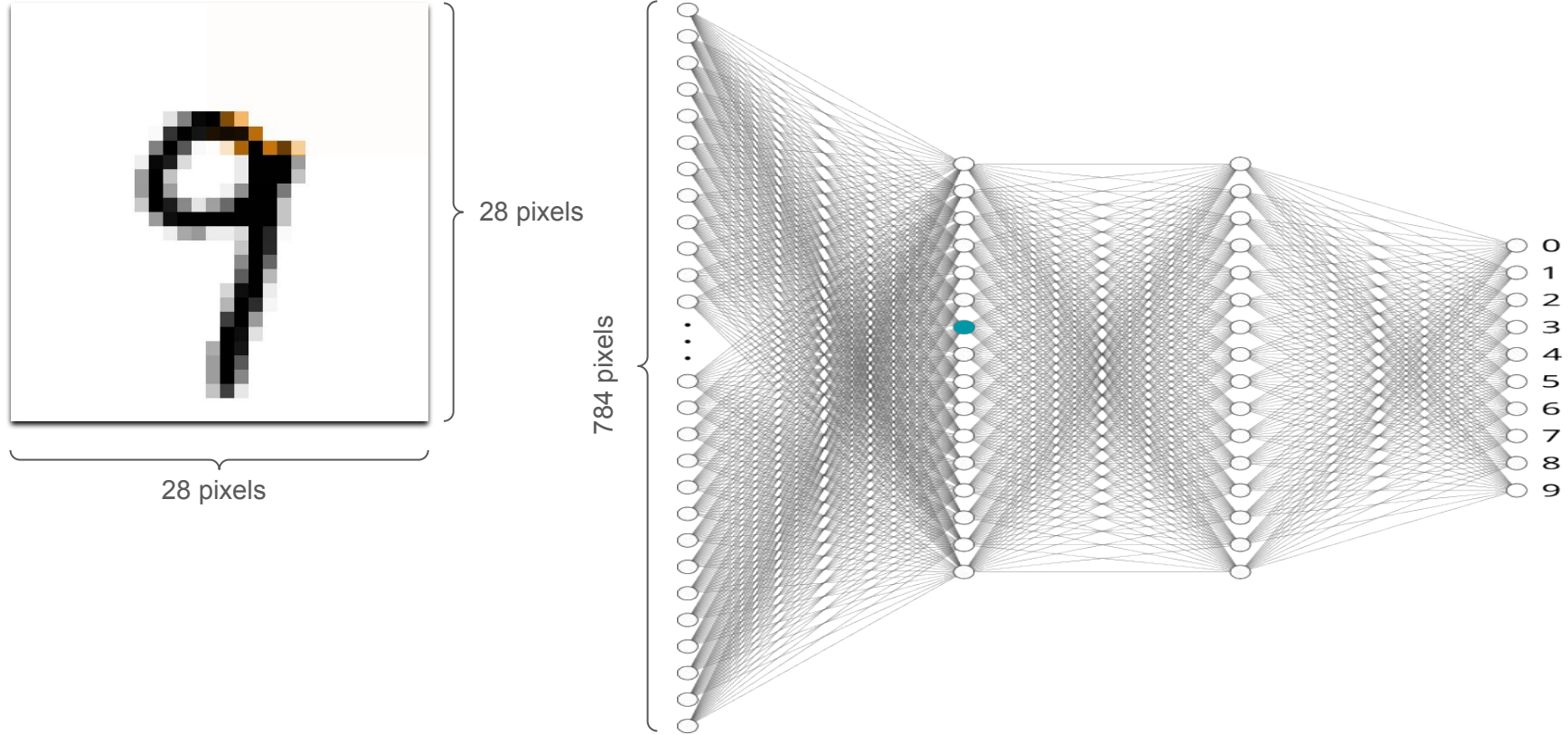
Multi-layer perceptron (MLP)



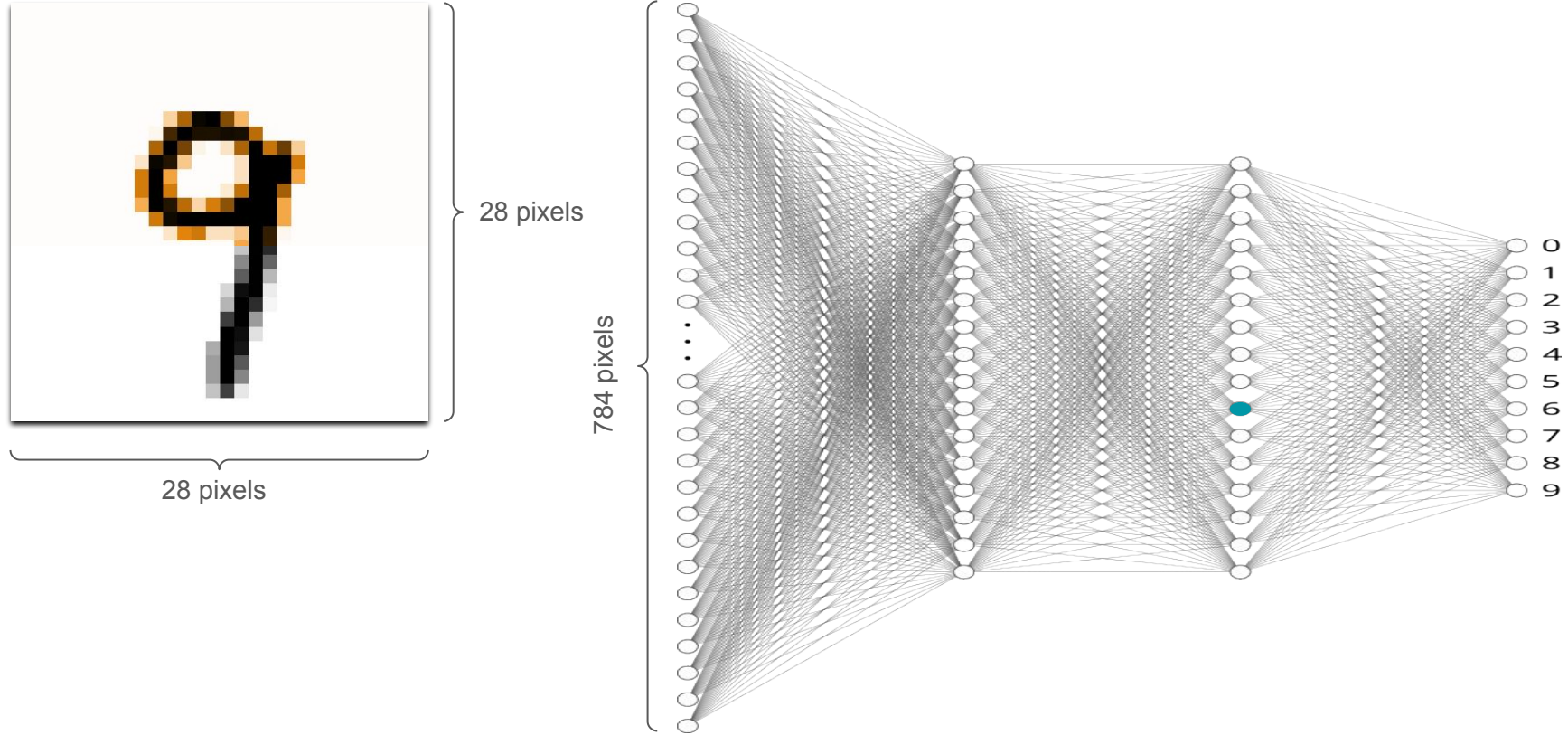
Multi-layer perceptron (MLP)



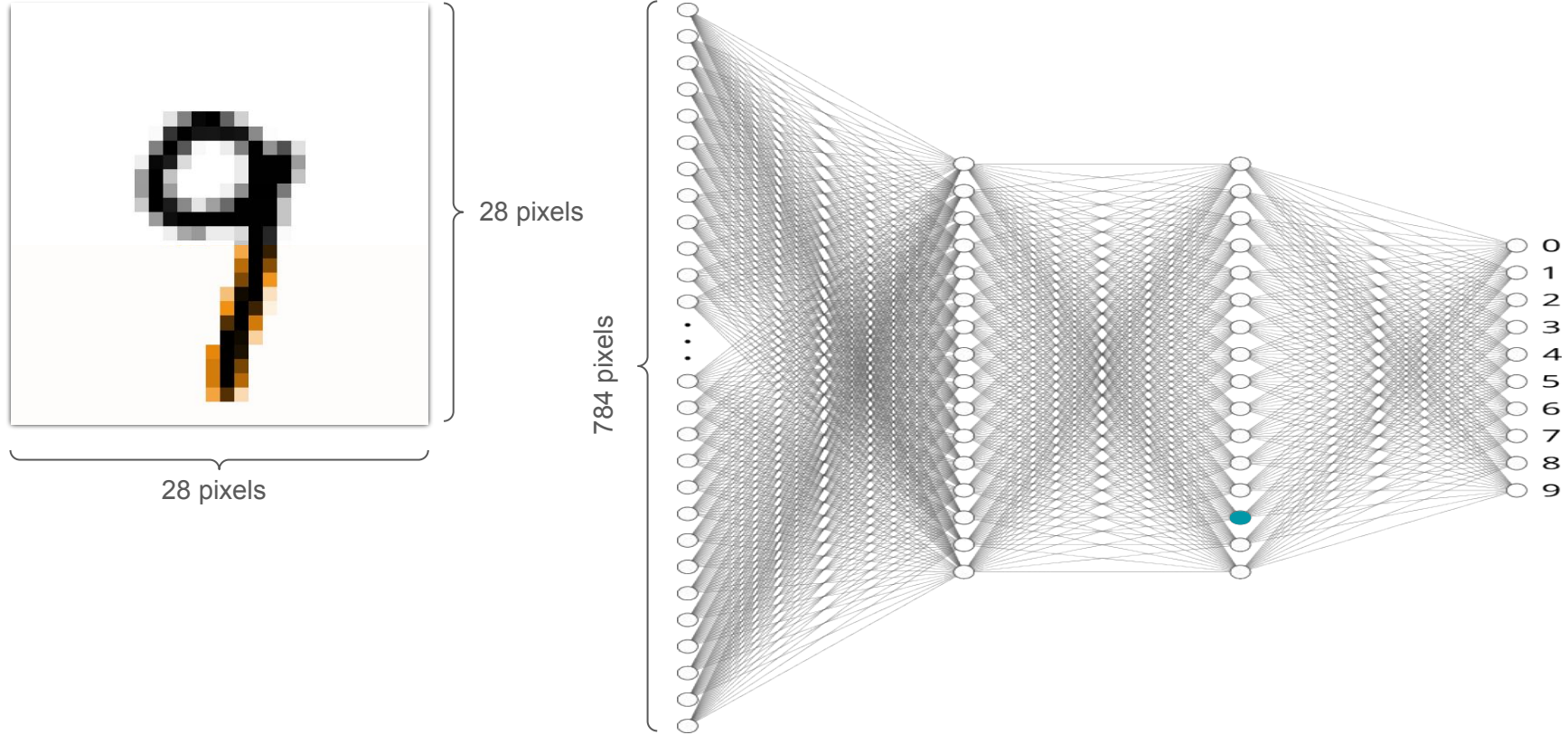
Multi-layer perceptron (MLP)



Multi-layer perceptron (MLP)



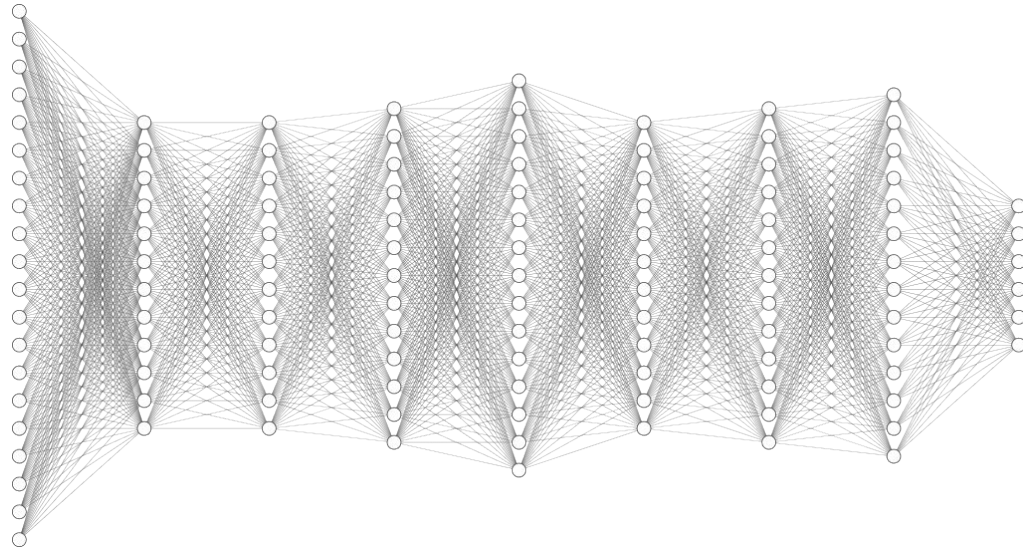
Multi-layer perceptron (MLP)



Pre-trained modellen en finetuning



ImageNet-1k



<https://www.tensorflow.org/hub>

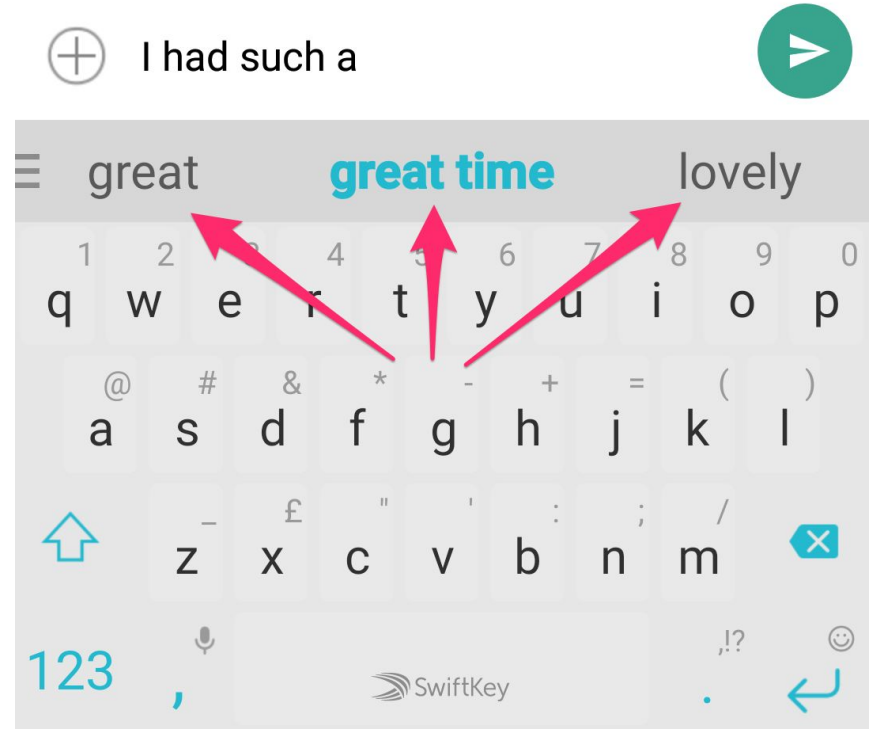
<https://pytorch.org/hub/>

<https://huggingface.co/models>

Large Language Models (LLMs)

Een taalmodel is een type model dat getraind is om een kansverdeling over woorden (tokens) te bepalen.

Simpel gezegd probeert een model het meest passende volgende woord (token) te voorspellen om een lege plek in een zin of zinsnede in te vullen, gebaseerd op de context van de gegeven tekst.



Tokens

Enter text:

The dog eats the apples.



464 3290 25365 262 22514 13

<https://alonsosilva-tokenizer.hf.space/>

Next token prediction

Enter text:

One, two,



3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

<https://alonsosilva-nexttokenprediction.hf.space/>

Instruction fine-tuning

Modellen waarmee je kunt chatten zijn getraind op chat data. Deze data is opgesteld volgens een speciaal template:

```
<s>[INST] <<SYS>>
```

```
{{ system_prompt }}
```

```
<</SYS>>
```

```
{{ user_message_1 }} [/INST] {{ model_answer_1 }} </s>
```

```
<s>[INST] {{ user_message_2 }} [/INST] {{ model_answer_2 }} </s>
```

Verschillende LLMs gebruiken verschillende *tags*, zie bijvoorbeeld ook [deze](#).

Nog steeds next token prediction

Als je een model vervolgens aanroept met de volgende structuur:

```
<s>[INST] <<SYS>>
```

```
{{ system_prompt }}
```

```
<</SYS>>
```

```
{{ user_message }} [/INST]
```

Dan vult het model dit aan met het antwoord op de vraag, token voor token, totdat het model het token `</s>` voorspelt.

Nog steers next token prediction

<s>[INST] <<SYS>>

You are a helpful assistant.

<</SYS>>

Why is the sky blue? [/INST]

Nog steers next token prediction

<s>[INST] <<SYS>>

You are a helpful assistant.

<</SYS>>

Why is the sky blue? [/INST]The sky appears blue due to a phenomenon called Rayleigh scattering.</s>

LLMs lokaal draaien

Ollama

Models

Blog

Discord

GitHub

Download



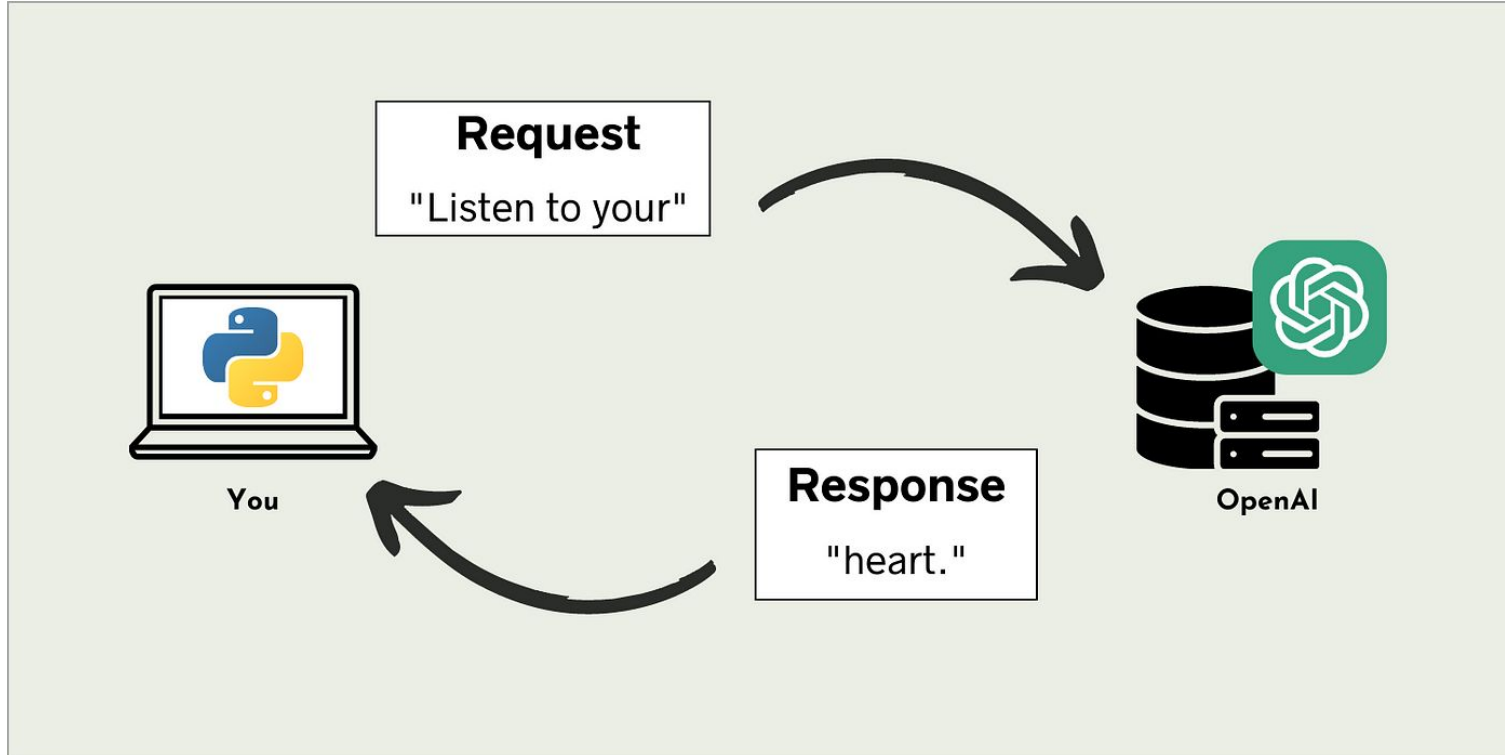
Get up and running with large
language models, locally.

Run Llama 2 and other models on macOS.
Customize and create your own.

Download

<https://ollama.com/>

LLMs aanroepen via een API



Machine learning in robotics

LeRobot



<https://huggingface.co/lerobot>



Hoe word ik een pro?

Waar kan ik meer leren over machine learning?

<https://www.kaggle.com/>

<https://www.youtube.com/@statquest>

<https://www.deeplearning.ai/courses/machine-learning-specialization/>

<https://www.deeplearning.ai/courses/deep-learning-specialization/>