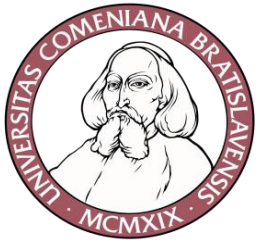


Introduction to cognitive science

Session 5: Connectionist paradigm

Martin Takáč
Centre for cognitive science
DAI FMFI Comenius University in Bratislava

Príprava štúdia matematiky a informatiky na FMFI UK v anglickom jazyku
ITMS: 26140230008

Previous: Symbolic paradigm

- Universal Turing machine
  - Can compute anything that is computable.
- Symbol systems
  - Manipulate symbols by syntactical rules
  - How do symbols acquire meaning?
- Problems
  - Frame problem
  - Symbol grounding problem

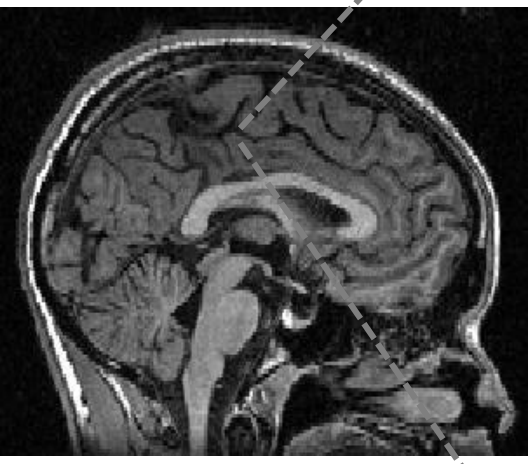
In this session:

- Symbolic versus subsymbolic representation
- Distributed representation
- Gradedness
- Graceful degradation
- Robustness
- Feedback
- Neural architecture & knowledge

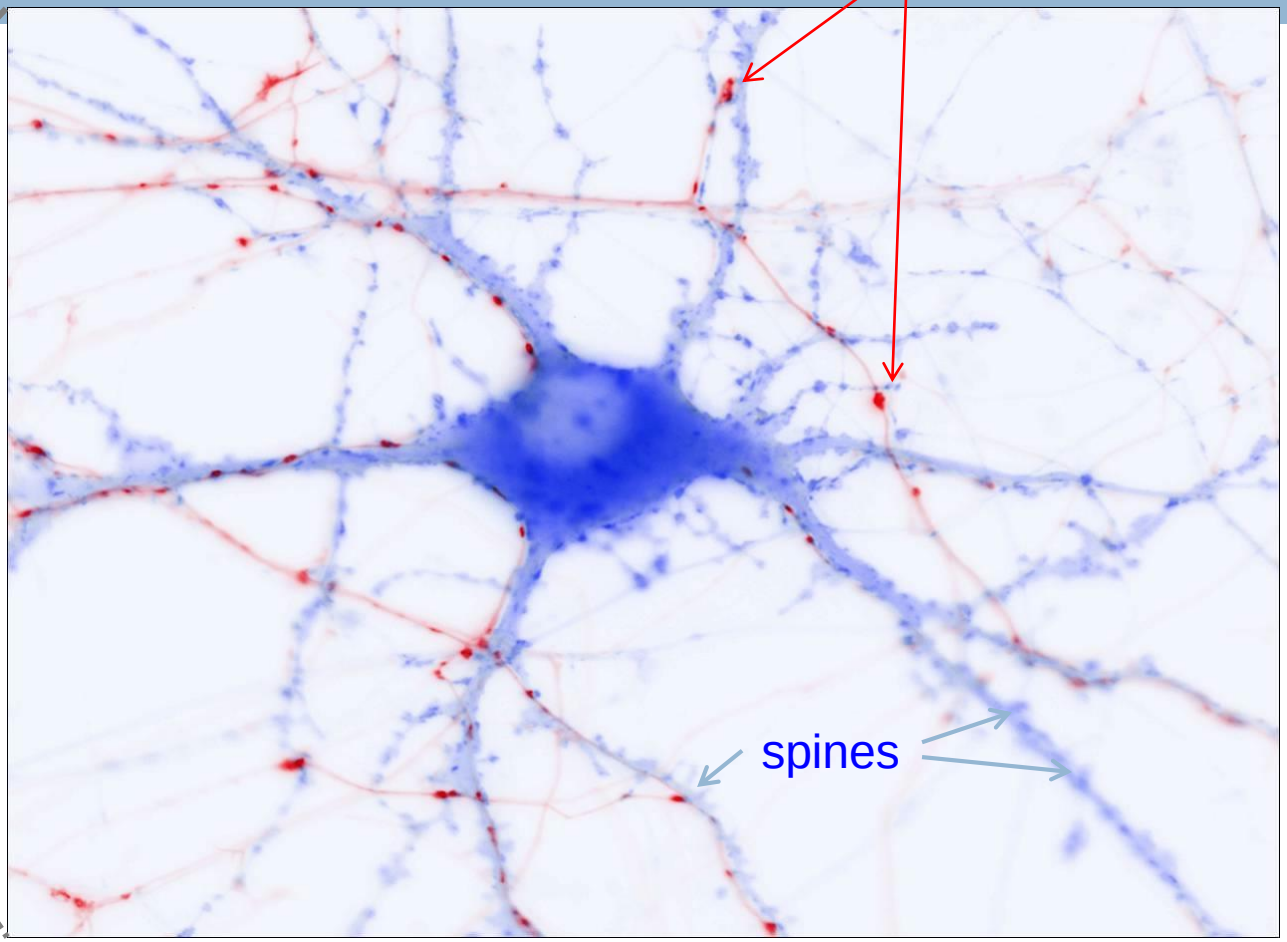
Connectionist (sub-symbolic) paradigm

- Inspired by biological brains
- Network
  - Many simple processors
    - Neurons <-> Units
  - Connectivity
    - Axons/synapses <-> Weighted connections
  - Parallel processing
  - Distributed representation
  - Representation - **not static**

Brain is comprised of networks of neurons connected and communicating via synapses



10^{11} neurons



10^{4} synapses in and out

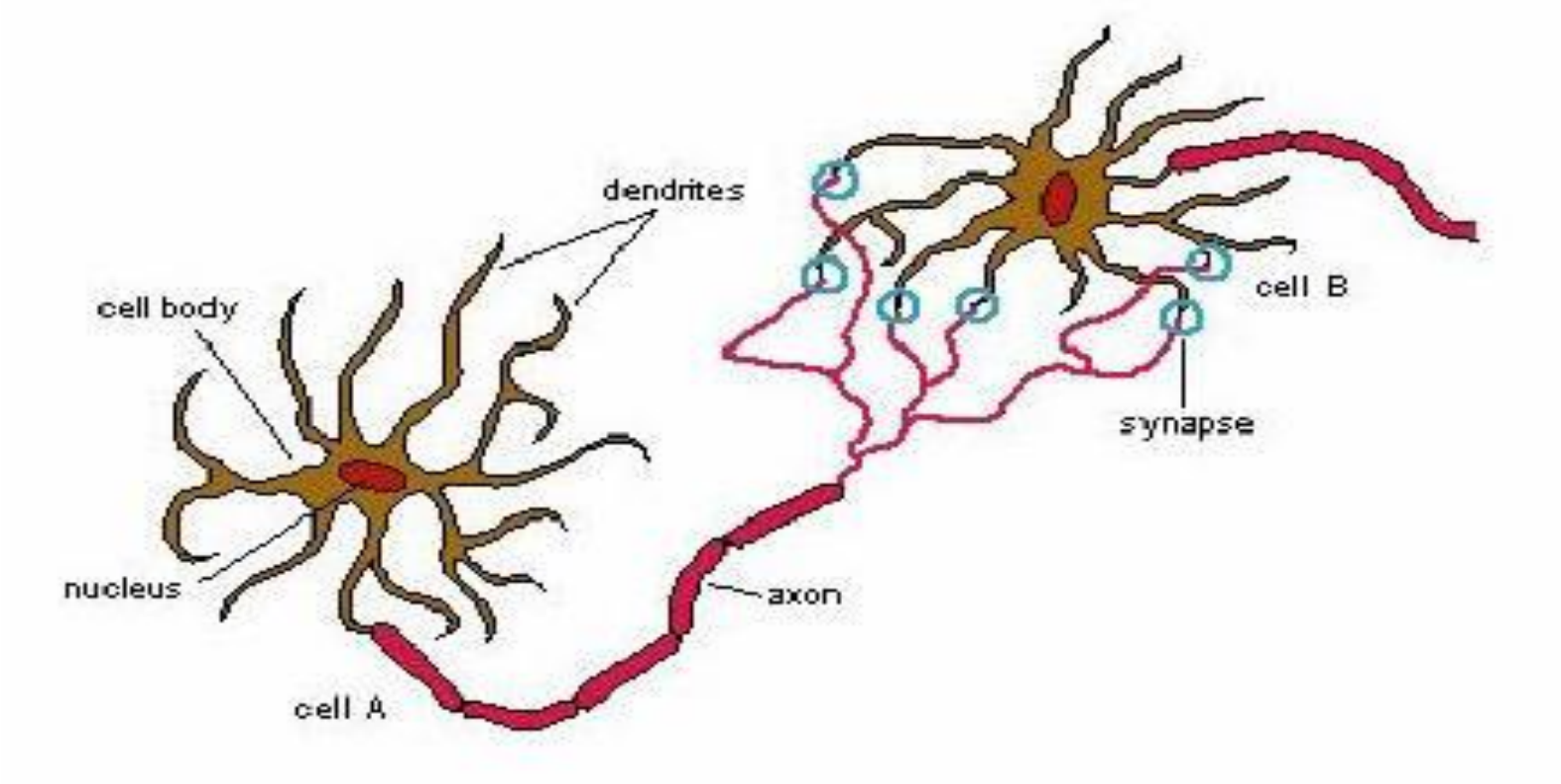
Learning and brain plasticity

6 Traditionally: learning is an acquisition of memories.

- Memory is an organism's ability to store, retain, and subsequently recall information.
- Neural correlate of learning is **brain plasticity**.
- Brain plasticity (neuroplasticity) is a lifelong ability of the brain to reorganize connectivity of neural circuits based on new experience.
- Brain plasticity is based on **synaptic plasticity**.

Hebb's rule of synaptic plasticity (1949):

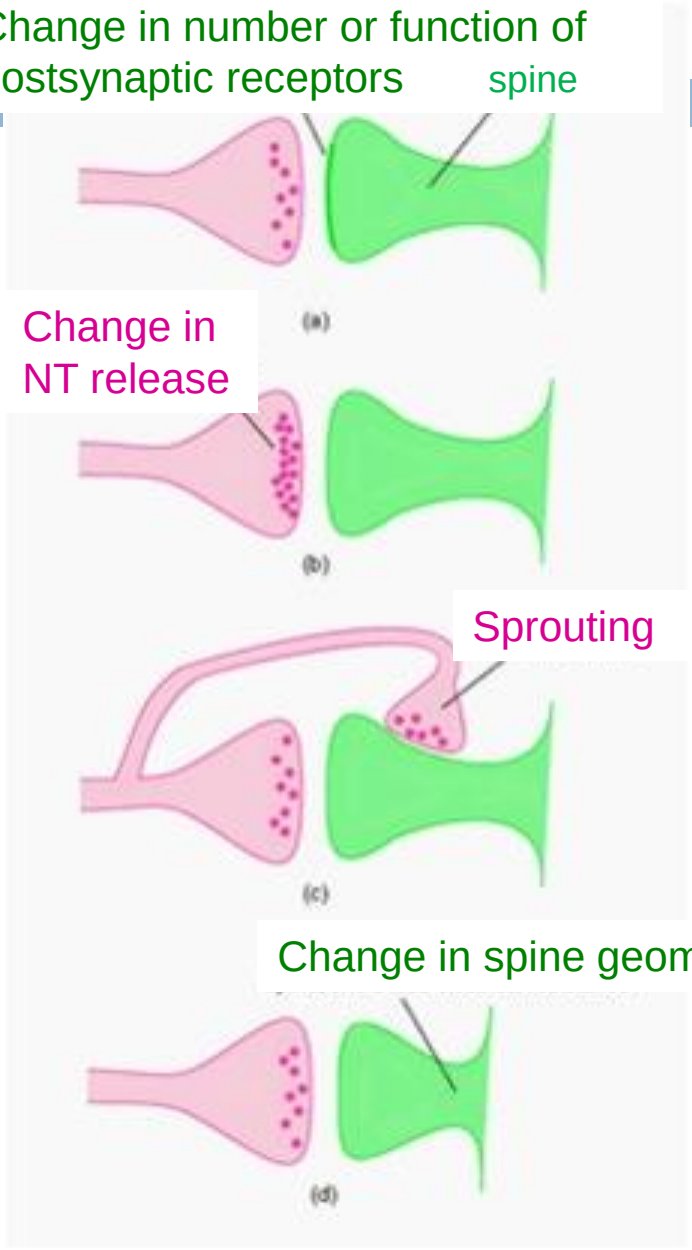
When an axon of cell A is near enough to excite a cell B and ***repeatedly or persistently*** *takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."*



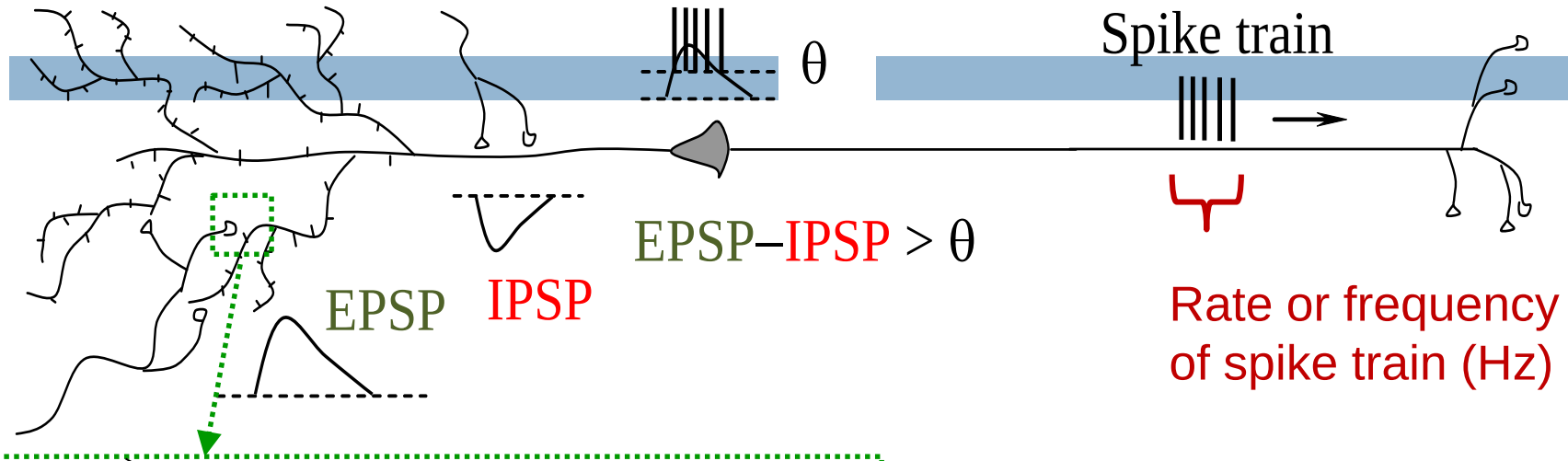
Synaptic plasticity: mechanisms

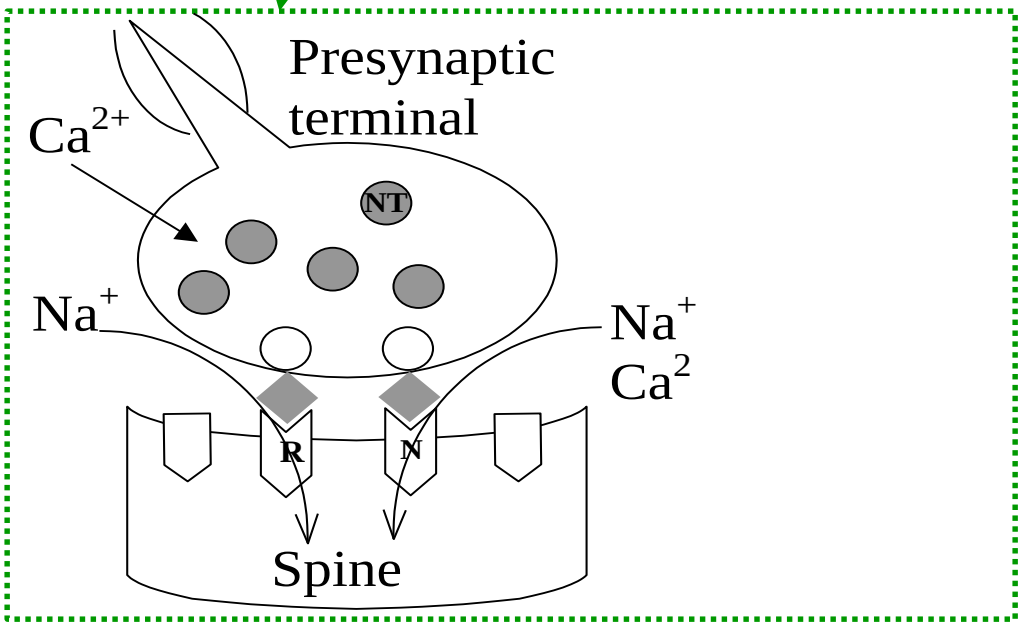
- Brain plasticity = refinement of the connectivity of neural networks based on **synaptic plasticity**.

- Synaptic plasticity is the ability of the synapse to change its strength (pre- and postsynaptic mechanisms).



Firing threshold and spikes

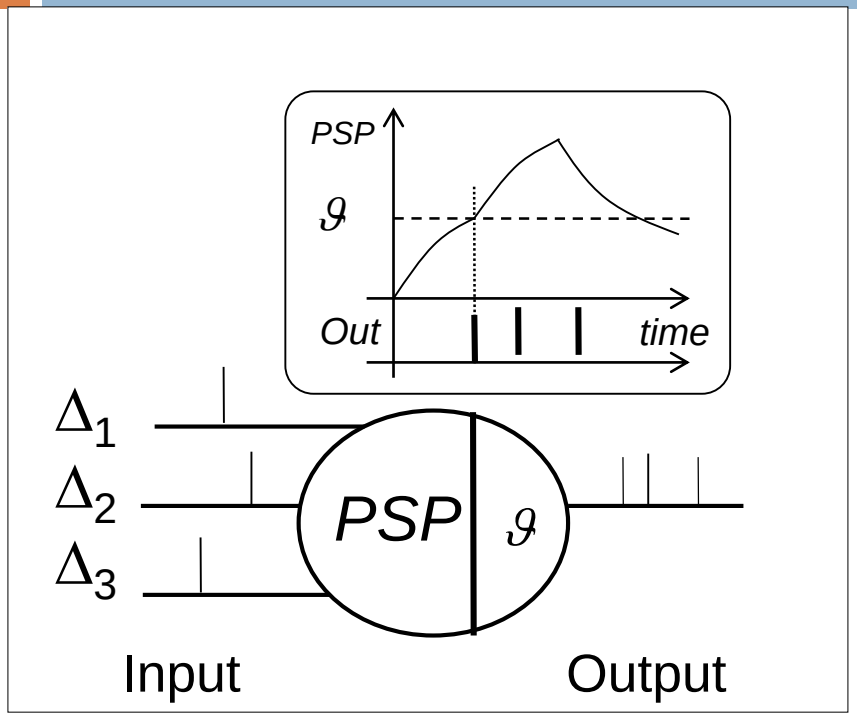


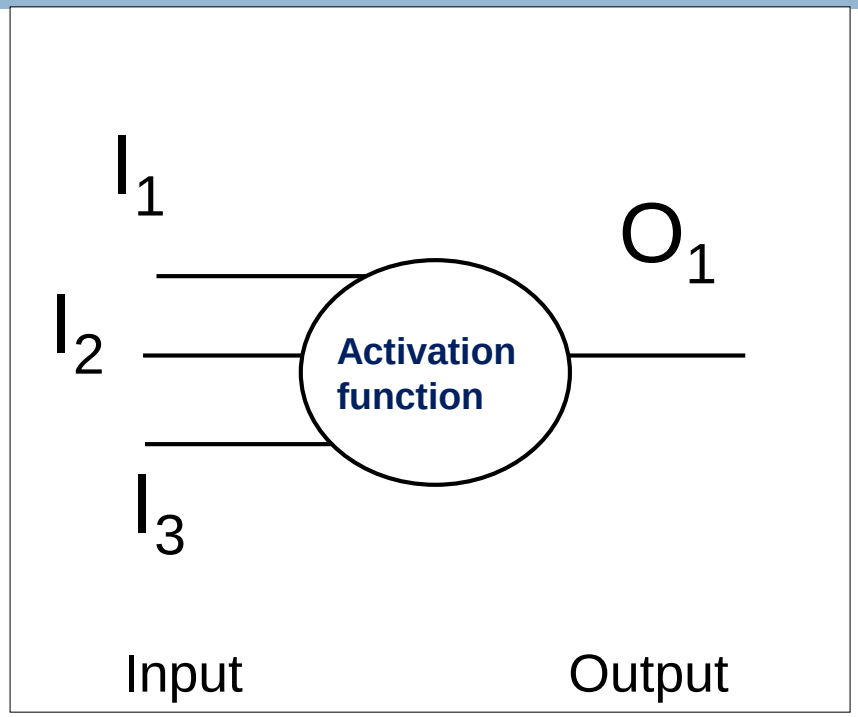


Excitation: EPSP

Inhibition: IPSP

Spiking versus rate neuron models

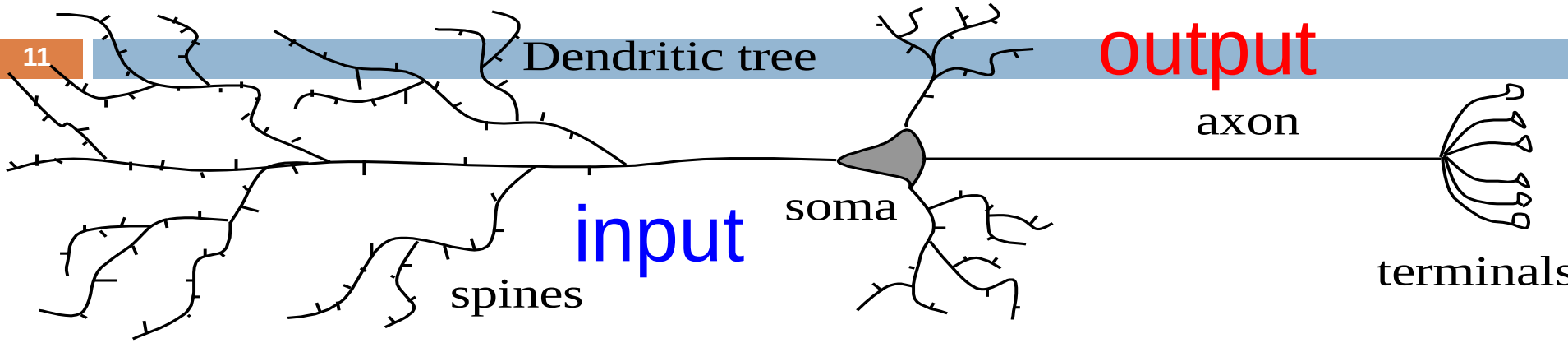


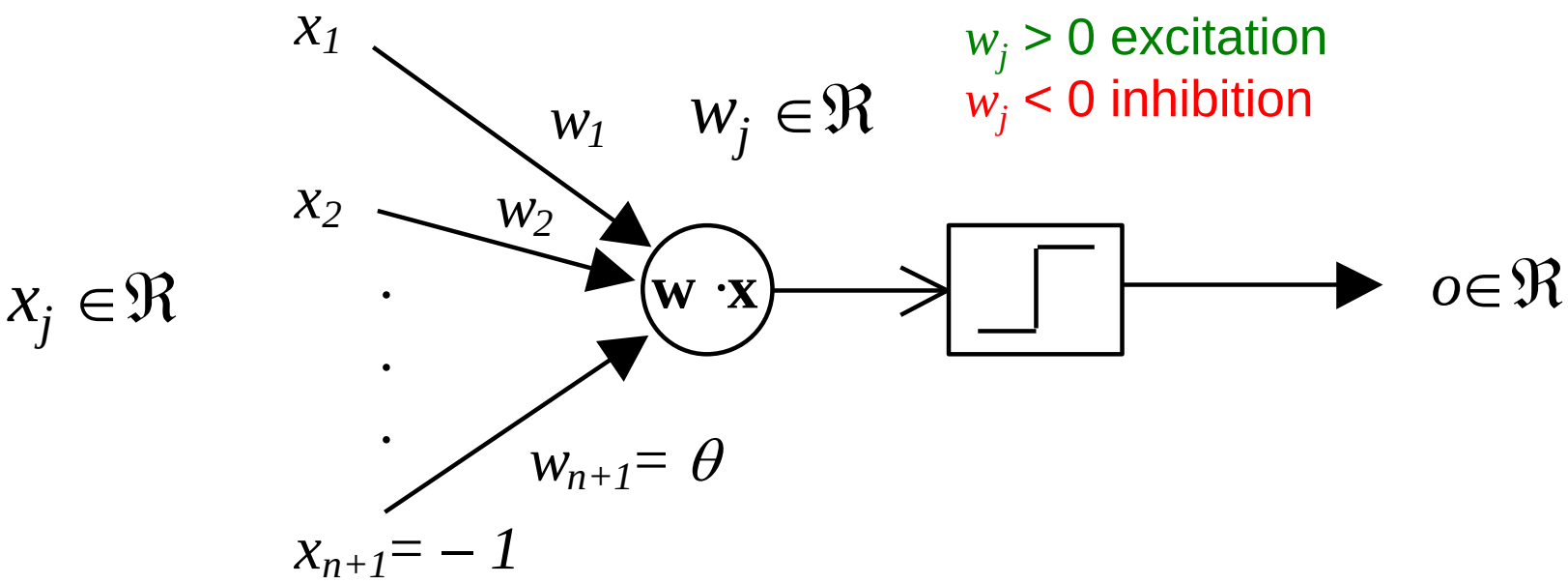


Spiking model: output depends on timing of input spikes

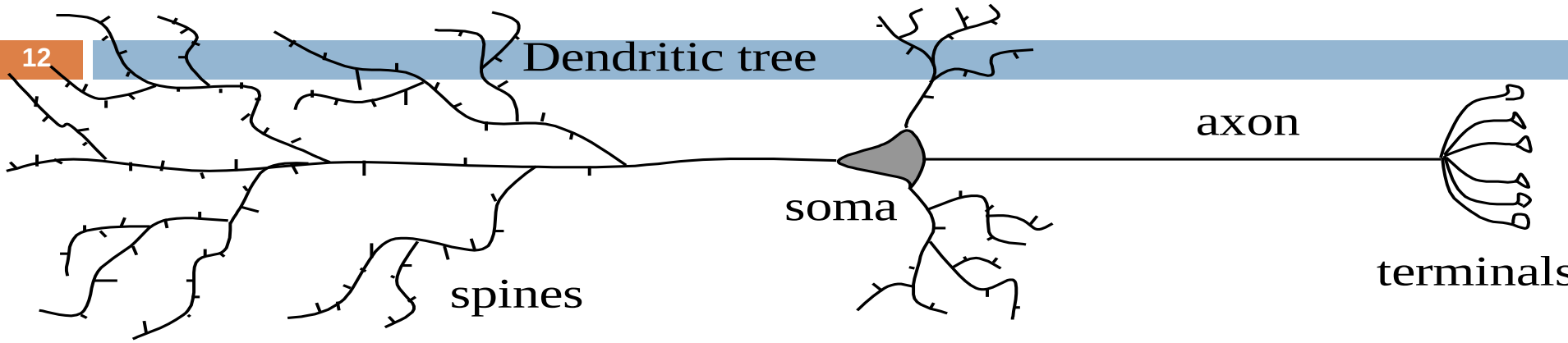
Rate model: output depends on the sum of input rates

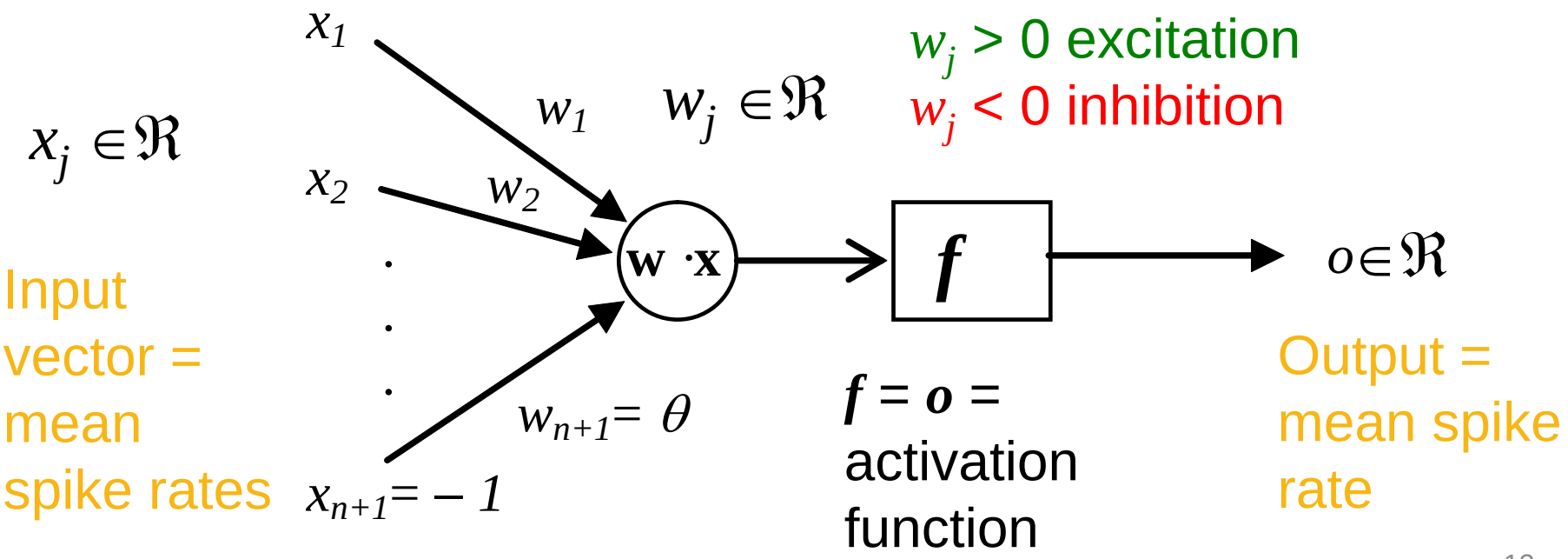
Perceptron – neuron (rate) model



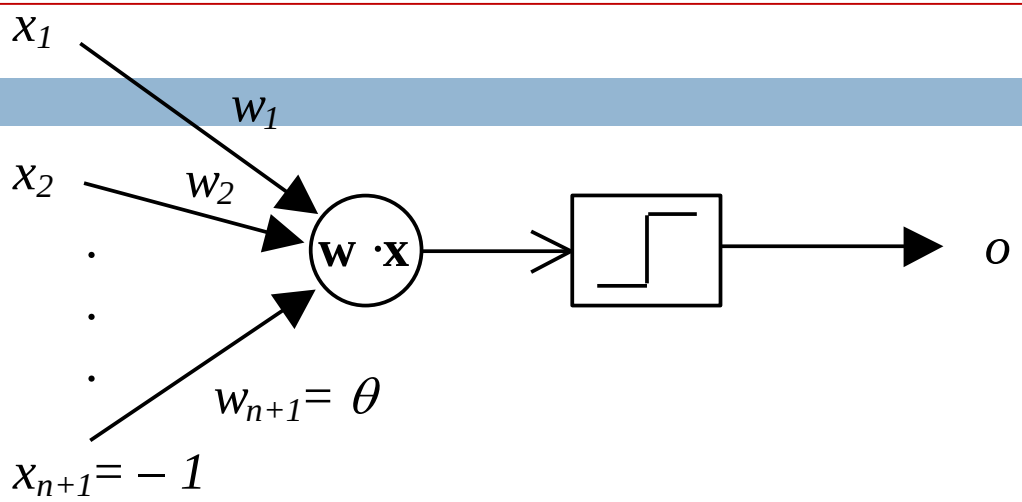


Neuron rate model





Rosenblatt's binary perceptron (hyperplane)



$$o = f(net) = f(\mathbf{w} \cdot \mathbf{x}) = f\left(\sum_{j=1}^{n+1} w_j x_j\right) = f\left(\sum_{j=1}^{n} w_j x_j - \theta\right)$$

$$f(net) = sign(net) = \begin{cases} +1 & if\ net \geq 0 \Leftrightarrow \sum_{j=1}^{n} w_j x_j \geq \theta \\ -1 & if\ net < 0 \Leftrightarrow \sum_{j=1}^{n} w_j x_j < \theta \end{cases}$$

General learning rule

- The weight vector changes as a function of the product of the input vector x and the learning signal $s(t)$

$$w_j(t+1) = w_j(t) + \Delta w_j(t) = w_j(t) + \alpha \; s(t)\, x_j(t)$$

- $0 < \alpha \leq 1$ is the learning speed

- Based on the type of the learning signal we distinguish:
  - Supervised learning
  - Reinforcement learning
  - Unsupervised learning

Types of learning

- **Supervised learning:** weights are adjusted according to the ***desired*** output (perceptron, MLP, RBF, RNN)

$$s = s(\mathbf{w}, \mathbf{x}, d)$$

- **Reinforcement learning:** weights are adjusted according to the ***reward*** (MLP, RBF) – extension of the supervised learning

$$s = s(\mathbf{w}, \mathbf{x}, r)$$

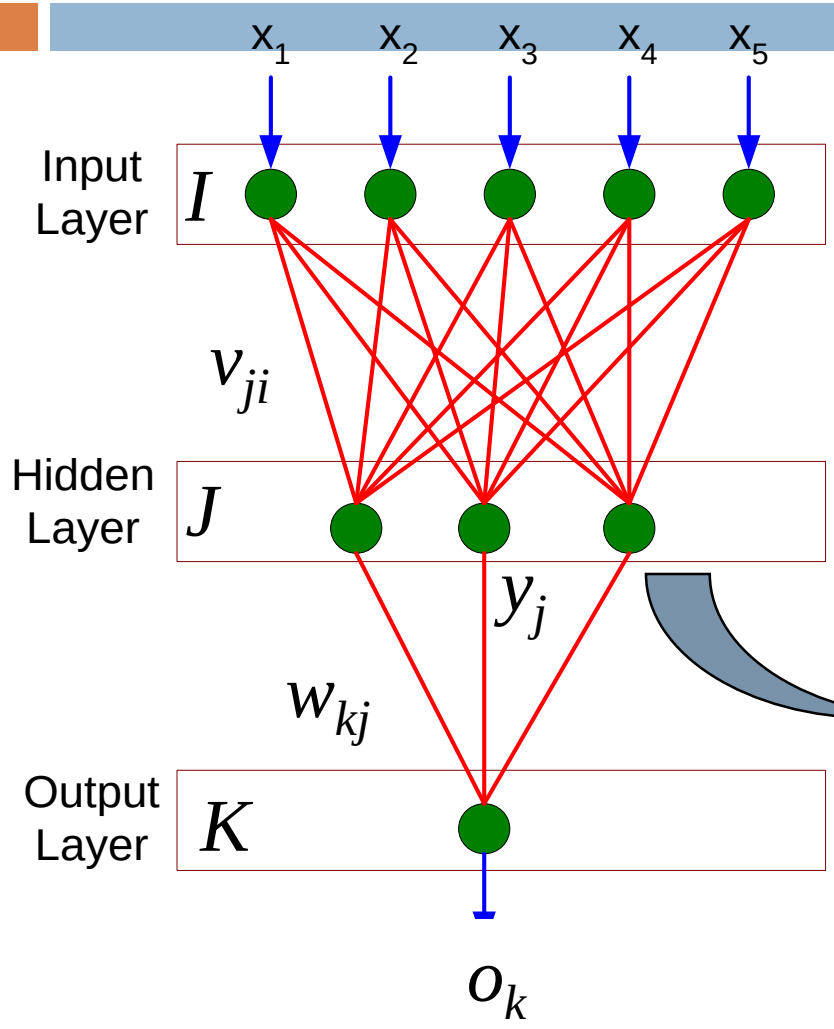
- **Unsupervised learning:** weights are adjusted according to the ***statistics*** of the input:

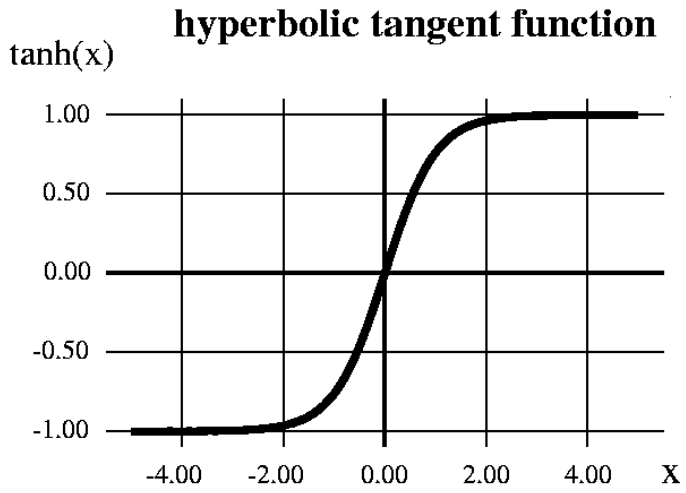
$$s = s(\mathbf{w}, \mathbf{x})$$

History of connectionism

- 1950' – perceptron (Rosenblatt, 1958)
- 1960' Minsky&Papert – Perceptrons (1969)
  - Limtations
    - Linearly separable classification problems
- 1970' - SOM
- 1980' – multiple layers
  - Hidden layer
  - Back propagation algorithm
- 1990' – recurrent networks
- 2000 – deep networks

MLP (Multilayer Perceptron)

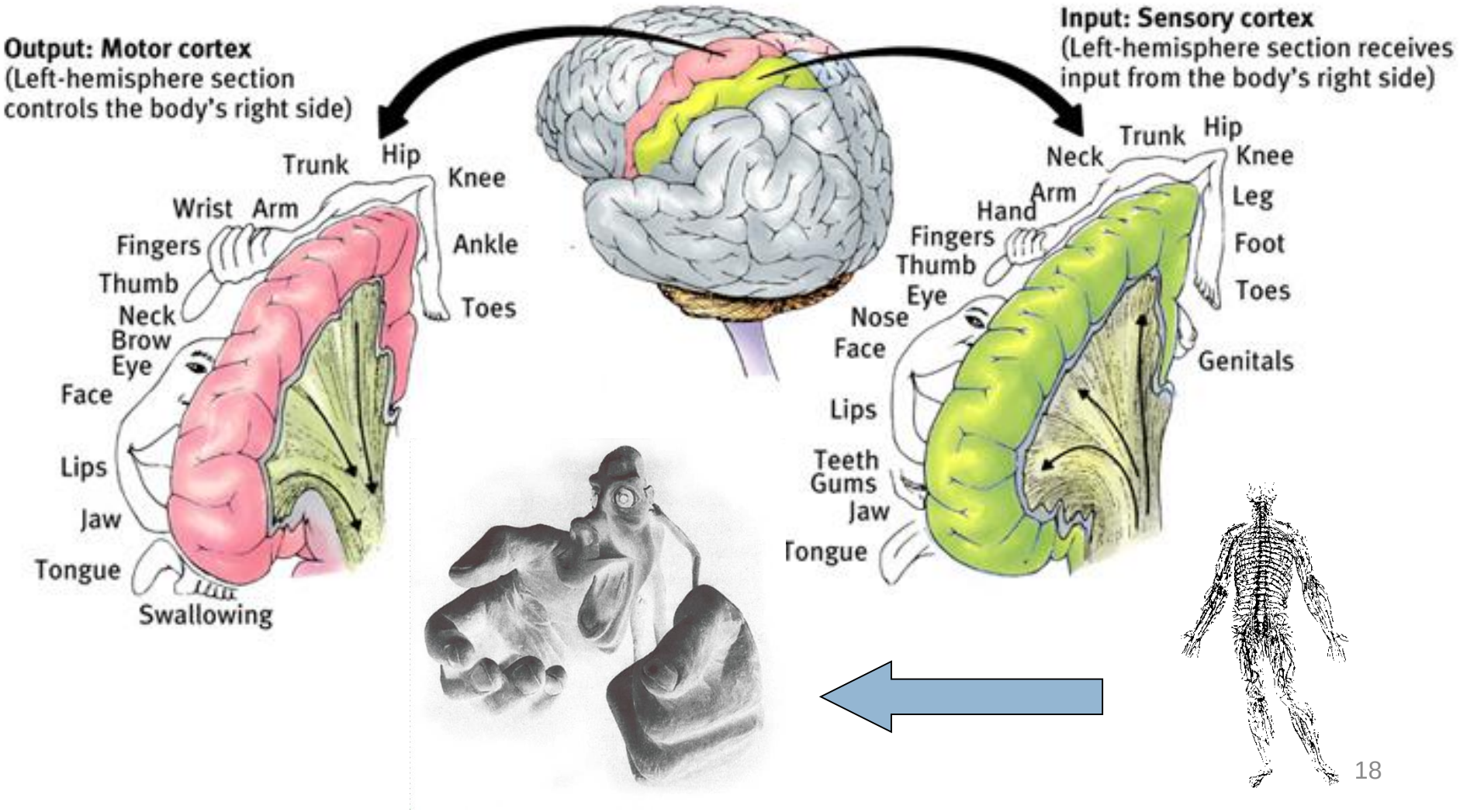




$f(net)$ is a nonlinear differentiable function (hyperbolic tan, sigmoid, Gaussian, etc.)

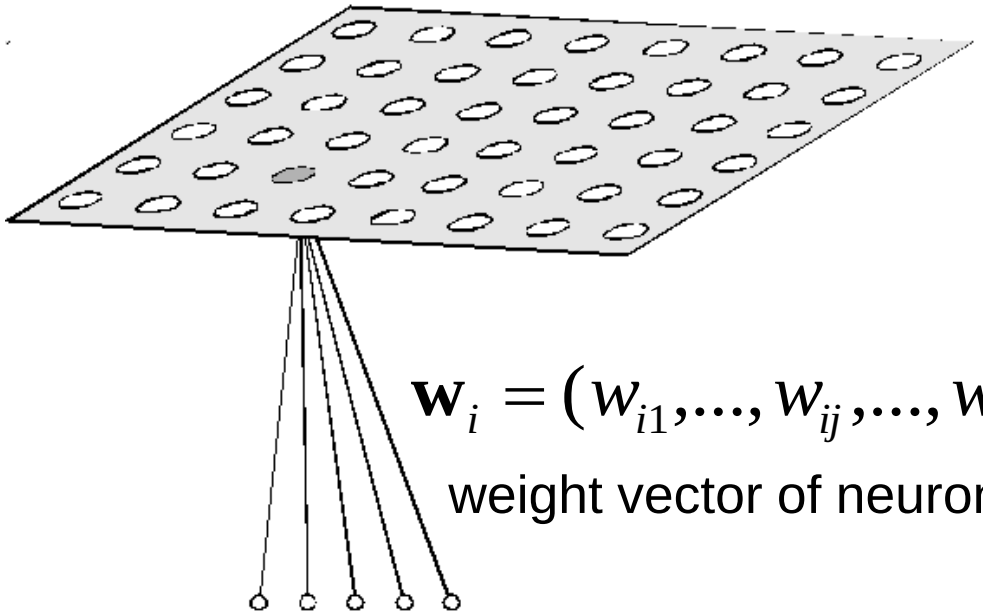
Somatosensory and motor systems in animals

- topological mapping from body to cortex



Self-Organizing Map (SOM): architecture

n linear neurons in the output layer: $o_i = \sum_{j=1}^{m} w_{ij} x_j = \mathbf{w}_i \mathbf{x}$



$\mathbf{w}_i = (w_{i1}, \ldots, w_{ij}, \ldots, w_{im})$

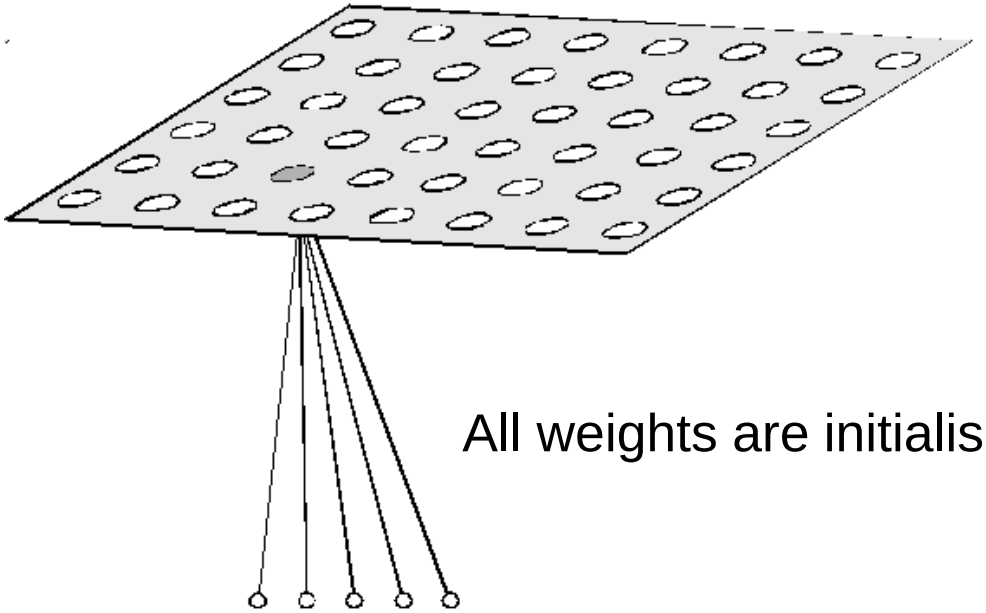
weight vector of neuron i

$\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_m)$

Input pattern = vector of real values. They feed each neuron.

SOM training: competition phase

- For each input vector in the training set, we find a winner neuron
  - According to maximal dot product: $i^* = \text{argmax}\left(\mathbf{w}_i\,\mathbf{x}\right)$
  - Or according to minimal Euclidean distance: $i^* = \underset{i}{\text{argmin}}\, d_{\text{E}}(\mathbf{w}_i, \mathbf{x})$



All weights are initialised as small random numbers

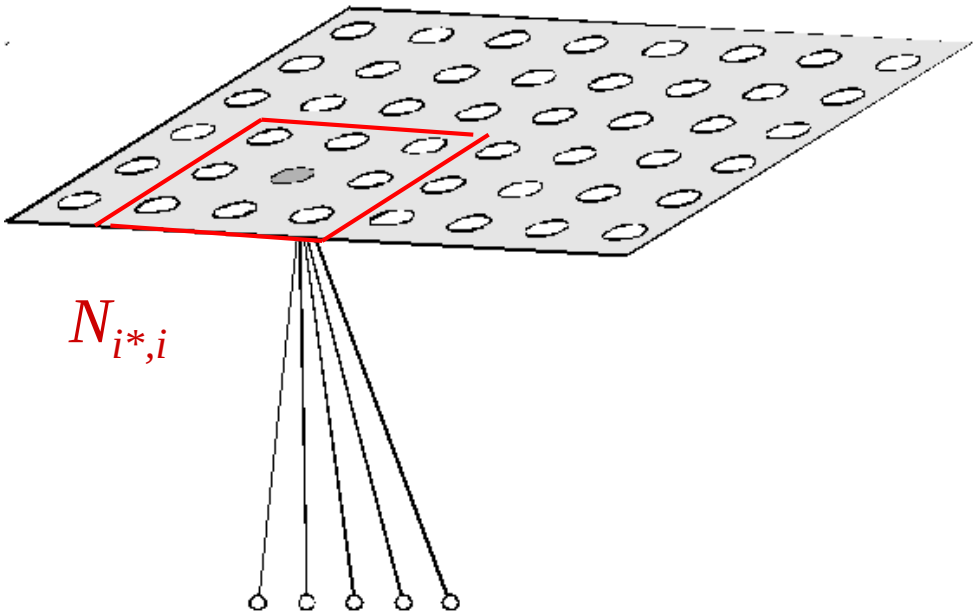
Training set consists of input vectors only, which are presented in random order:

$$A_{train} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^p, ..., \mathbf{x}^P\}$$

SOM training: weight update & cooperation

The weights of the winner and its neighbours in $N_{i^*,i}$ are updated:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t) \cdot N_{i^*,i}(t) \cdot [\mathbf{x}(t) - \mathbf{w}_i(t)]$$



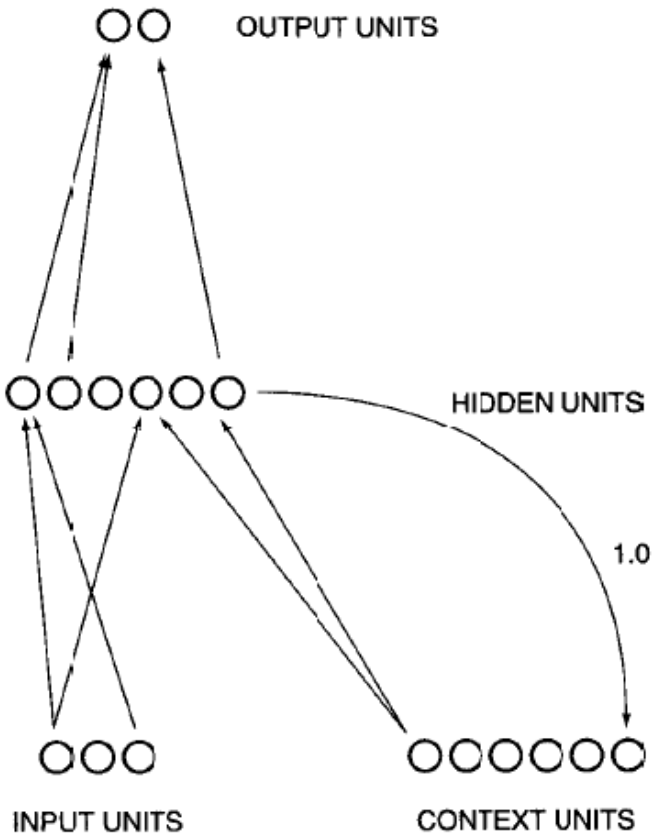
The weight vectors of $i \in N_{i^*,i}$ move closer to the current input

Training set consists of input vectors only, which are presented in random order:

$$A_{train} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^p, ..., \mathbf{x}^P\}$$

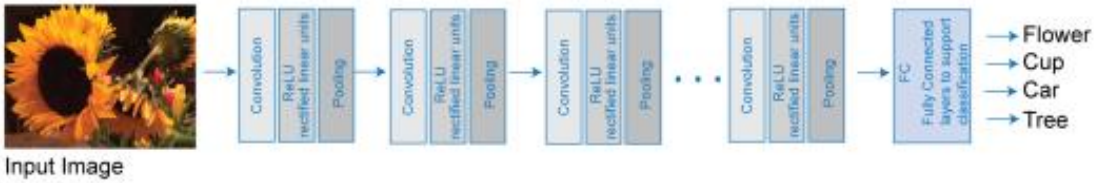
Recurrent neural networks

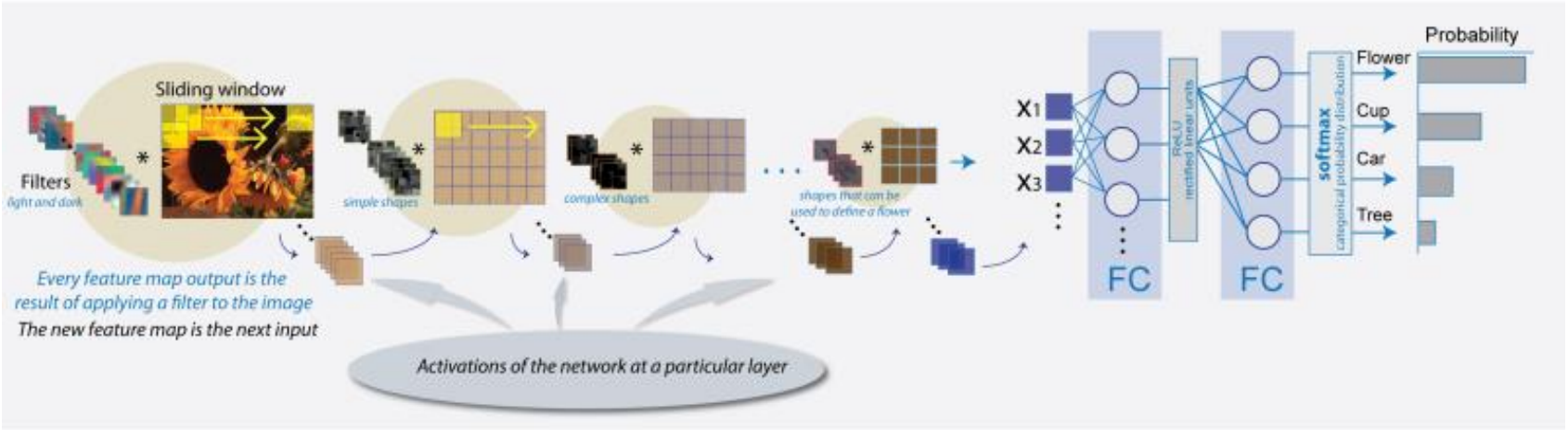
- Able to process dynamic temporal patterns
- Additional **feedback loop**
  - recycles some aspects of the networks activity at time *t1* together with input in *t2*.



(Clark, 2001)

Deep networks





Symbolic vs. connectionist

Symbolic	Connectionist
Sequential	Parallel
Logic & Deduction	Patterns & Induction
Algorithm must be known	Learning from examples
Noise intolerant	Noise tolerant
Semantically interpretable	Semantic interpretation often not feasible
Not robust	Robust (graceful degradation)

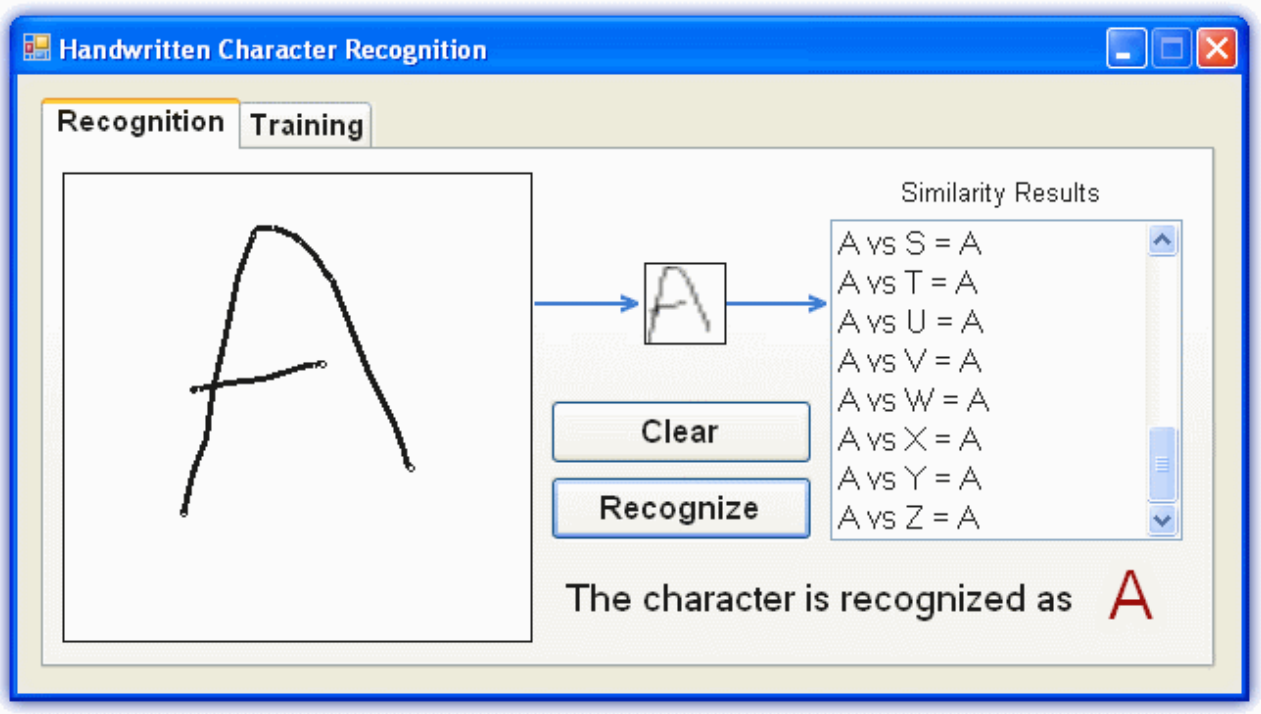
Distributed representation

- As opposed to a body of declarative statements
- Knowledge in the set of connection weights and structure of the network

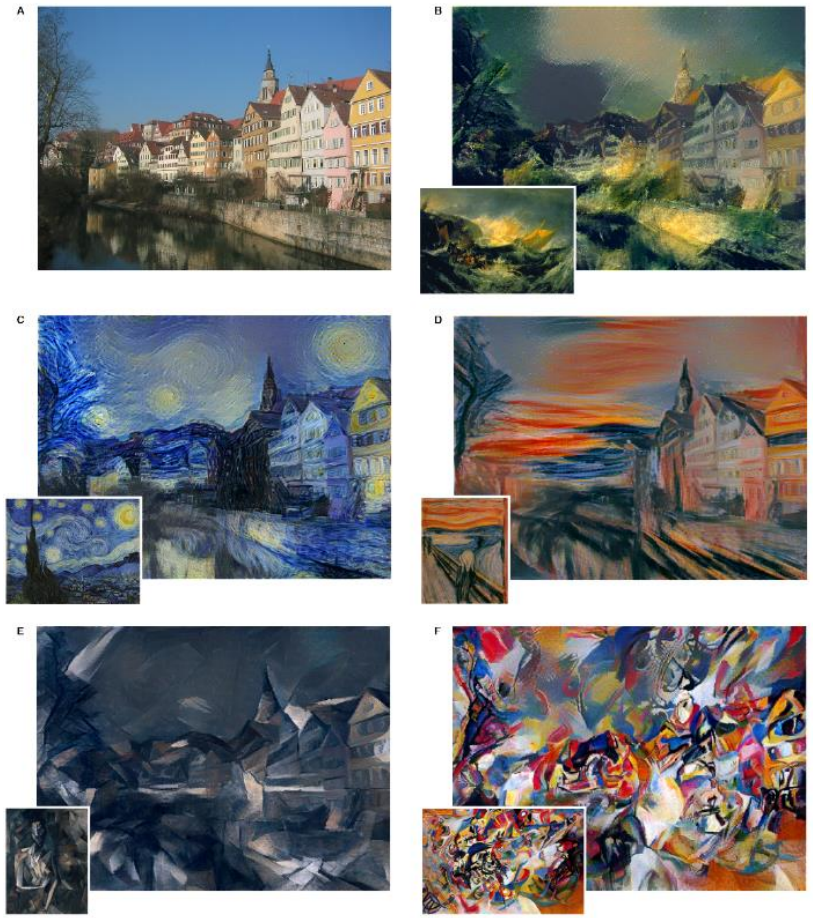
- Expressed by the **simultaneous activity** of a number of units
- Semantically related items are represented by syntactically related (**partially overlapping**) patterns of activation.

Generalization

- If new input resembles an old one in some aspects, it will yield a response in partial overlap



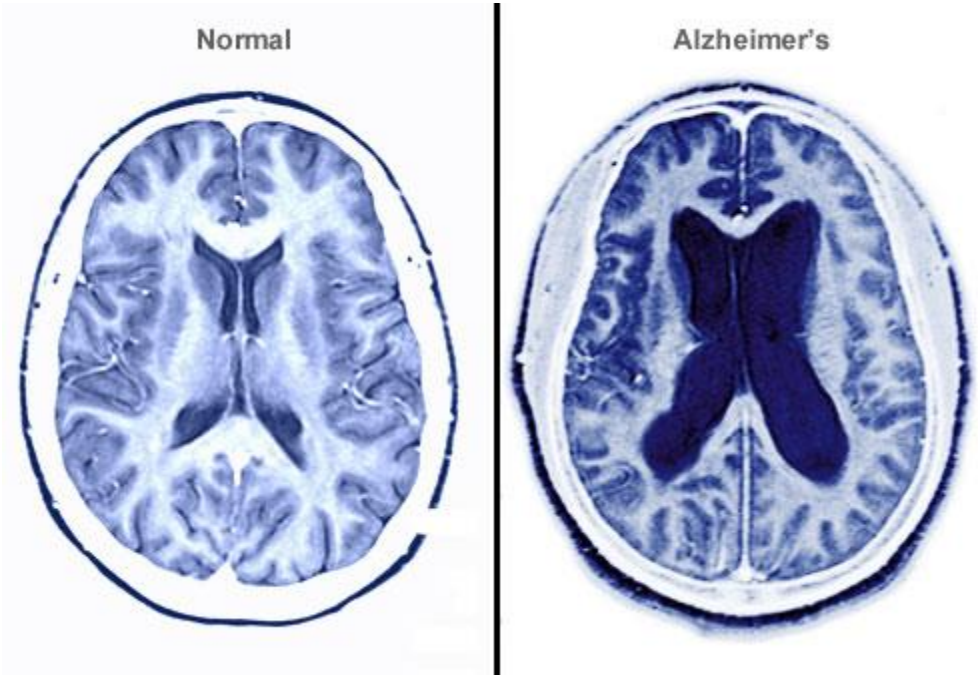
Generalization in deep networks



Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576.*

Graceful degradation

- In case of damage to the network, it can still produce sensible responses



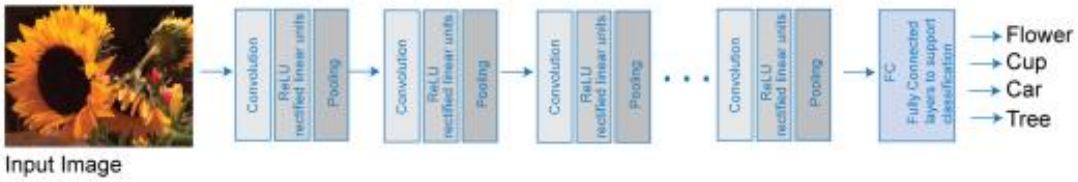
Lack of understanding

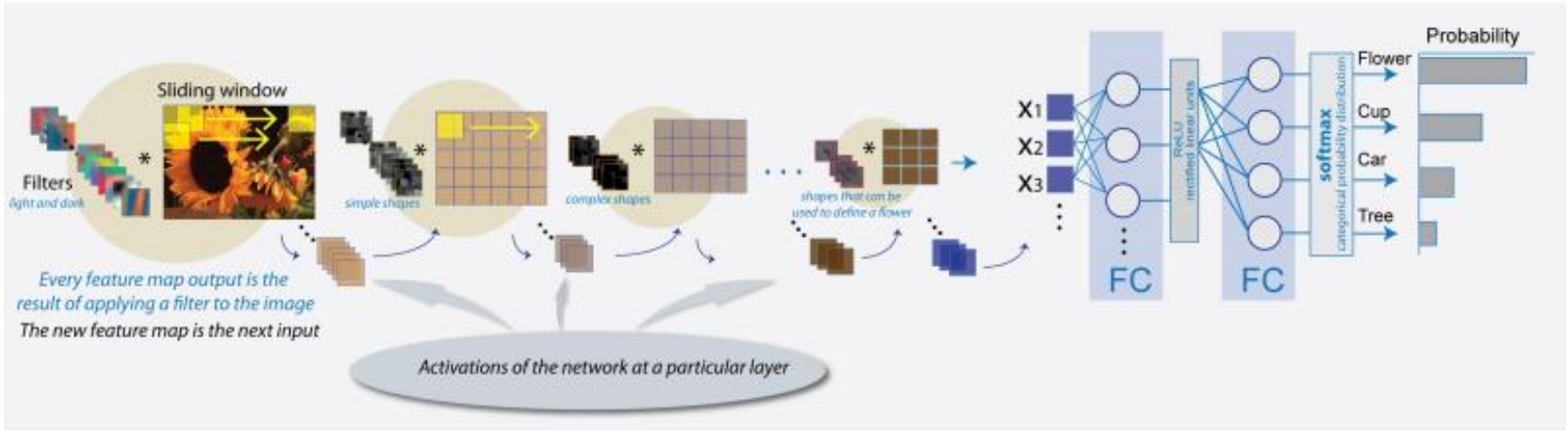
- How to understand the knowledge and strategies that the network is actually using to drive its behavior?
  - Posttraining analysis
    - Statistical analysis
      - Cluster analysis
      - Very different networks with the same training data yield similar statistical properties
    - Systematic interference – damage to units/connections
      - "lesions"

Problems of connectionism (before deep neural networks)

- Simplification of task
  - Usually not dealing with real-world problems
  - Discrete, well-defined problems
  - Not general
- Scaling
  - Models use usually small numbers of units
  - Solutions that work well for small networks with narrow focus fail to deal with large input spaces and multiple tasks
- Level of detail
  - Blue brain??

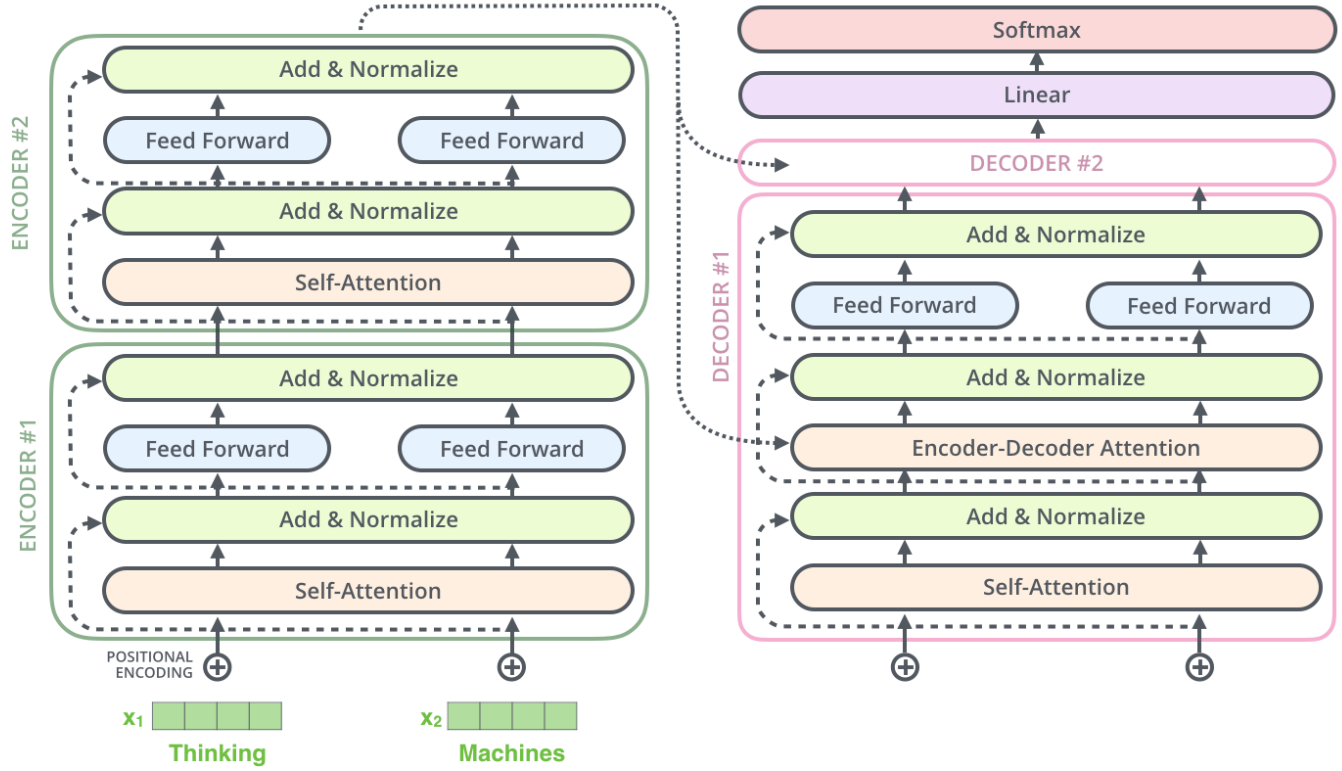
Deep neural networks (DNN)





Deep neural networks (DNN)

- Transformers (Vaswani et al., 2017)



Dreaming and fake videos

- Deep dream
  - (demo1, demo2)

- Deep Fake

Big language models

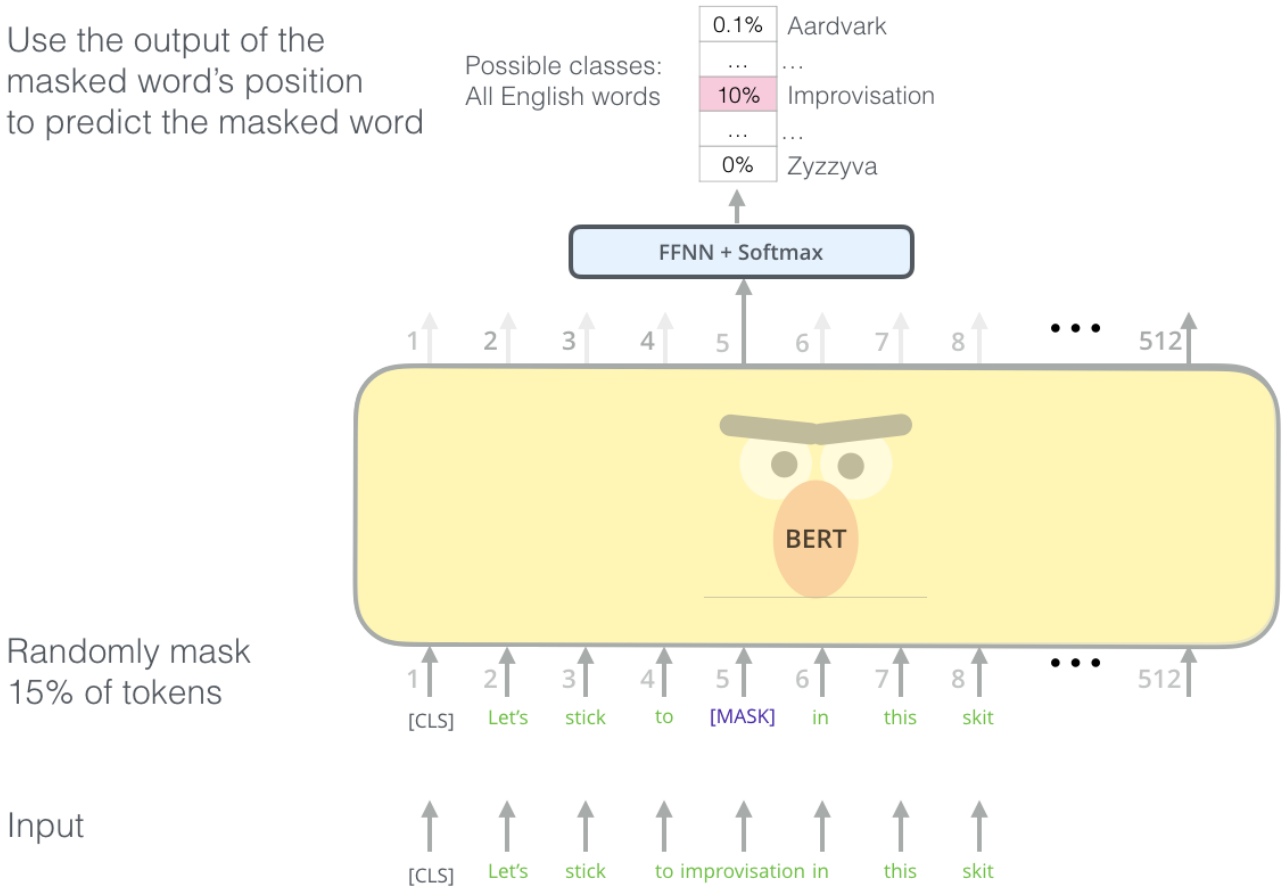
- GPT-3 (Open AI, 2020)
  - 175 billion parameters
  - ~500 billion data tokens:

GPT-3 Training Data

Dataset	# Tokens	Weight in Training Mix
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

  - Applications: Question answering, generating summaries…
  - How it works

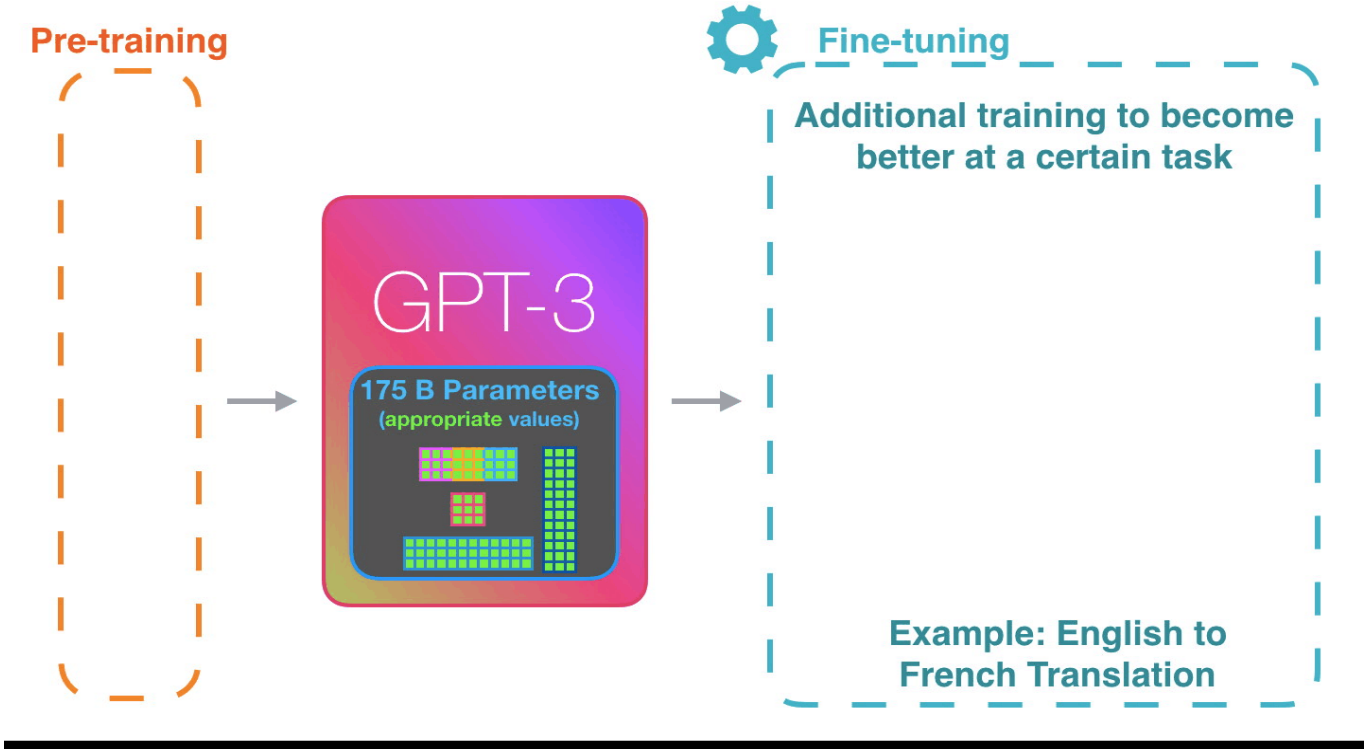
Pretraning – Masked Language Modelling (MLM task)



- Source: https://jalammar.github.io/illustrated-bert/

Fine-tuning, "zero shot" capabilities

- Tasks: classification, topic labeling, sentiment analysis, machine translation, natural language inference



- Source: https://jalammar.github.io/illustrated-bert/

Distributional semantics

- *"You shall know a word by the company it keeps"* (Firth, 1957)

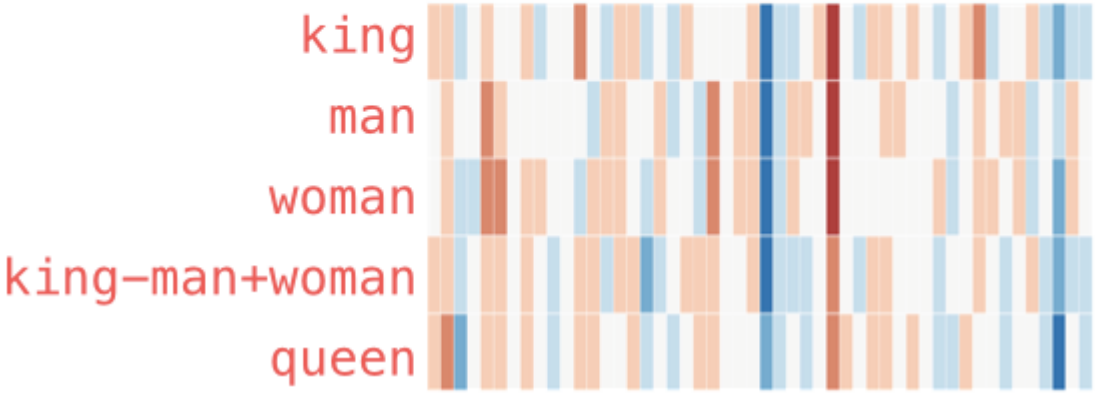
- A meaning of a word is represented as a vector in multidimensional space

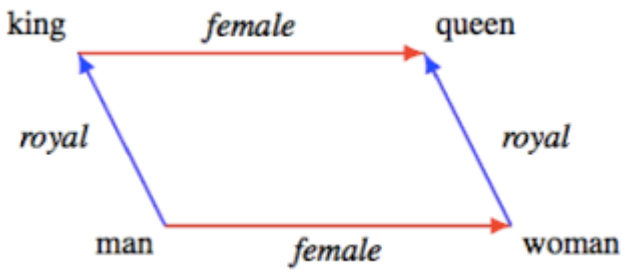
-0.34	-0.84	0.20	-0.26	-0.12	0.23	1.04	-0.16	0.31	0.06	0.30	0.33	-1.17	-0.30	0.03	0.09	0.35	-0.28	-(

  - Approach 1 (e.g. GloVe, Pennington et al., 2014): based on co-occurrences in a large corpora
  - Approach 2 (e.g. Word2Vec, Mikolov et al., 2013) – embedding (a hidden layer activity) of a neural network trained on a large corpus for next word prediction task
  - E.g. BERT model – with transformers trained on masked language task

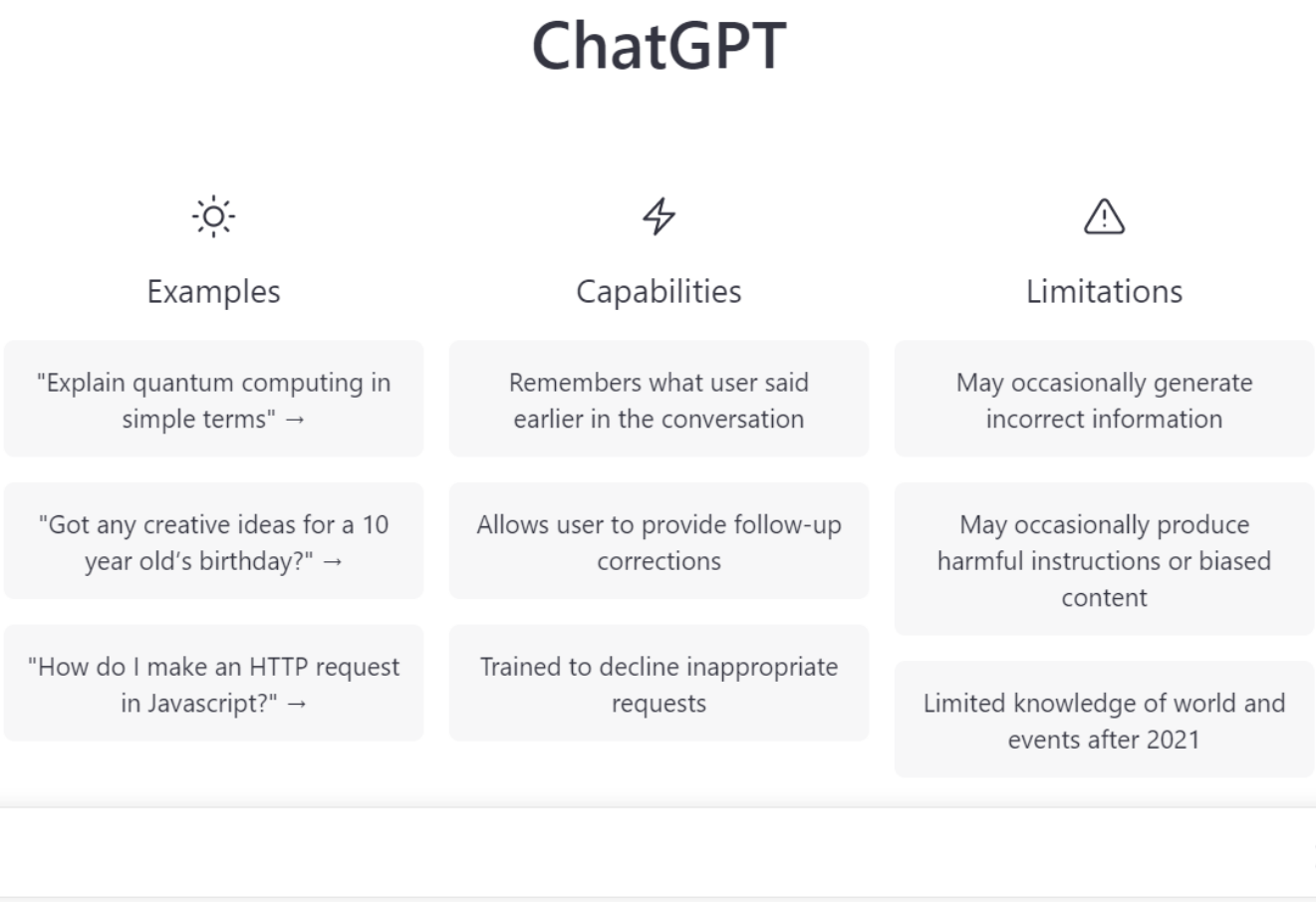
Analogy

king − man + woman ~= queen





ChatGPT


ChatGPT
Examples
"Explain quantum computing in simple terms" →
"Got any creative ideas for a 10 year old's birthday?" →
"How do I make an HTTP request in Javascript?" →
Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests
Limitations
May occasionally generate incorrect information
May occasionally produce harmful instructions or biased content
Limited knowledge of world and events after 2021

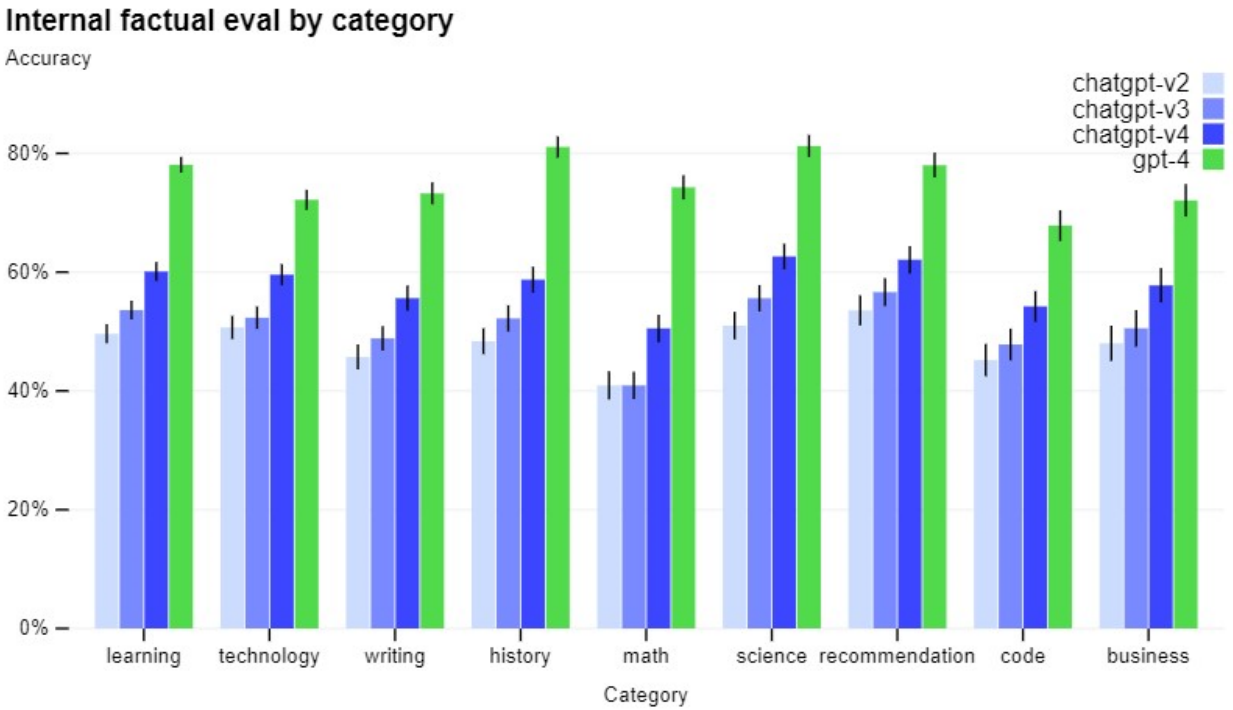
GPT-4 (March 14, 2023)

- Can accept text and images as input, produces text as output
- Its context is 20000 words
- Training: “The data is a web-scale corpus of data including correct and incorrect solutions to math problems, weak and strong reasoning, self-contradictory and consistent statements, and representing a great variety of ideologies and ideas.”
- Steerability: system messages

Hallucinations

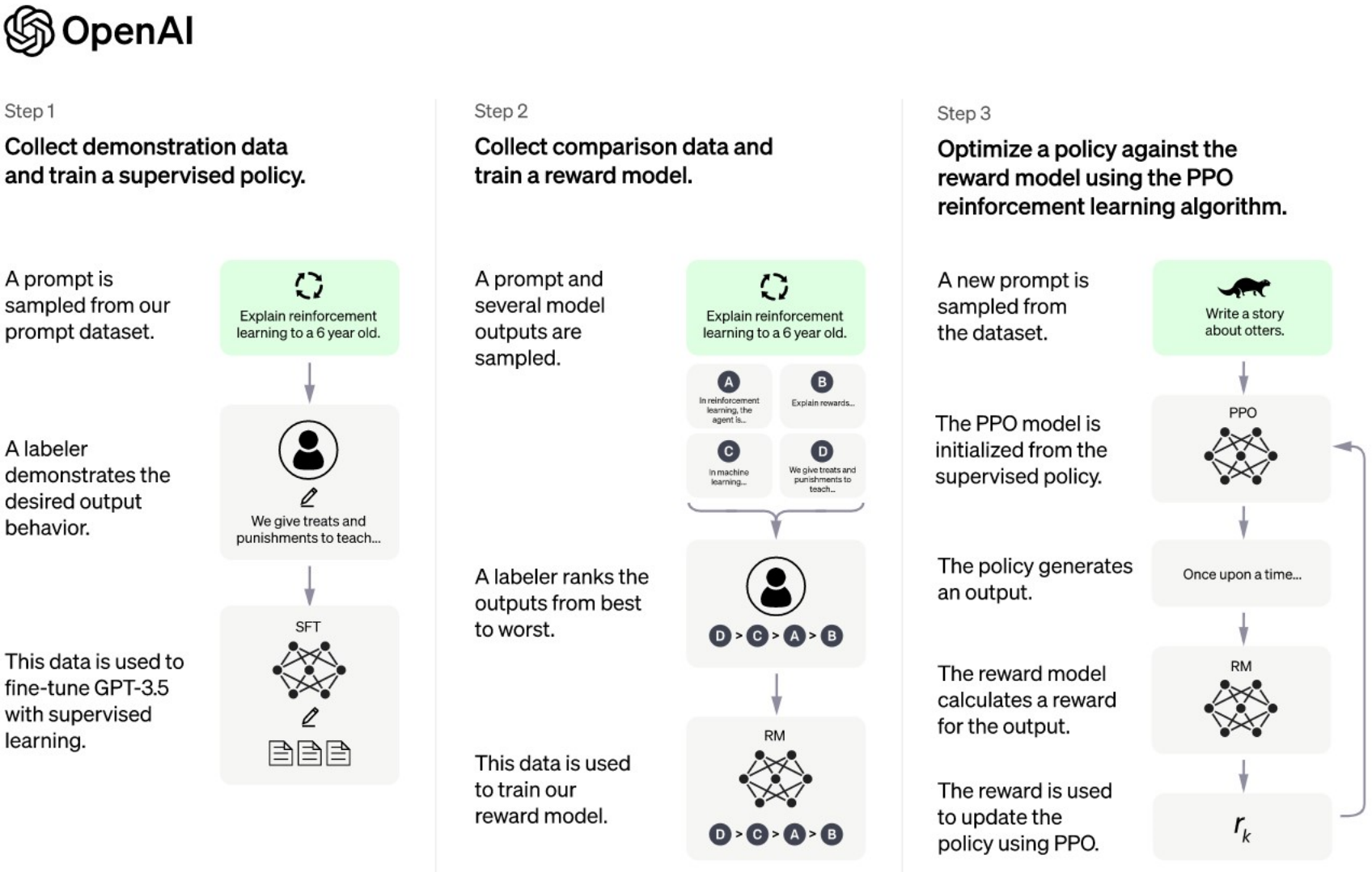
- **Bratislava is a capital of Slovakia. It has a very pleasant medieval inner city with narrow, winding streets, a hill-top castle next to the river Danube, and many historic churches and buildings to visit. The old town is centered on two squares, Hlavne namestie (main square) and Hviezdoslavovo namestie (Hviezdoslav square, named after a famous Slovak poet). Of a rather different architectural character are some of the communist-era buildings found in the modern parts of the city; a prime example is Petrzalka housing estate, the biggest Communist-era concrete block housing complex in Central Europe, which stretches on endlessly just across the river.**
- The city's population is about 430,000 in the city and in more than 1.3 million in the metropolitan area. Bratislava is the capital and the largest city in Slovakia. Bratislava is in southwestern Slovakia on both banks of the Danube River. The city's name is said to derive from the Slovak word "Bratislav", meaning "brotherly", i.e. "peace and happiness"; the full name is often abbreviated as Bratislava.
- The areas that today are Bratislava were already settled in the Stone Age. Bratislava was the site of a Celtic settlement, a Roman fortress, a Slavic fortified town and a royal city of the Kingdom of Hungary. The origins of the first settlement date back to the Bronze Age, between the 5th and 2nd century BC. The first known permanent settlement of Bratislava was the Slavic fortified town of "Aquincum", which existed from the 5th century until the 9th century. In Slavic languages, the town was known as "Priwacz", which means "besieged location

Hallucinations

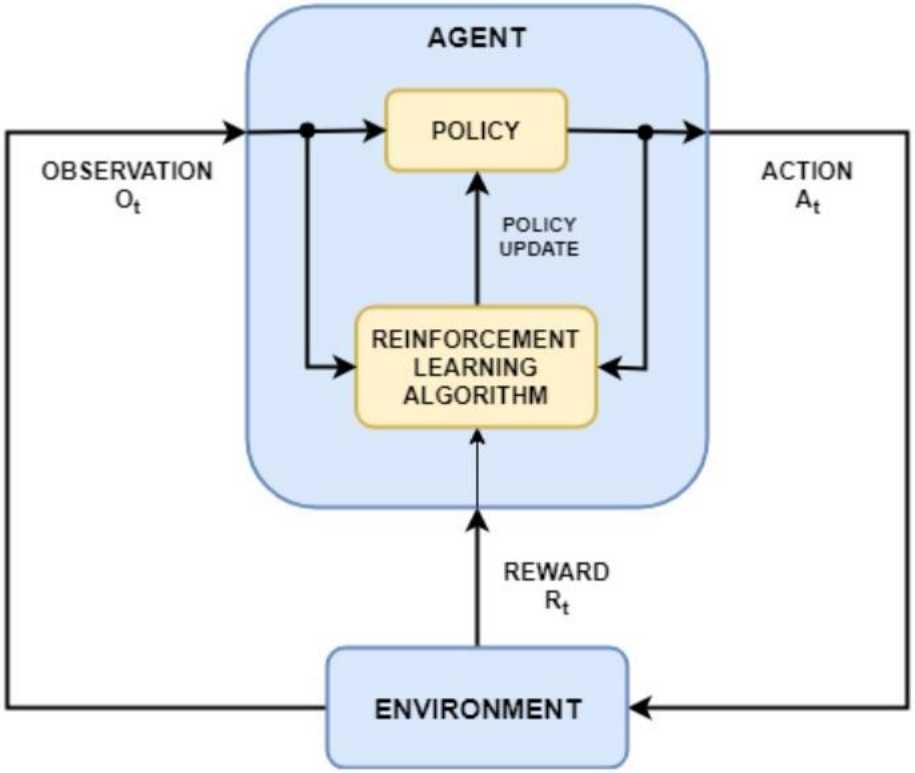


On nine categories of internal adversarially-designed factual evals, we compare GPT-4 (green) to the first three ChatGPT versions. There are significant gains across all topics. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval.

What is ChatGPT and how does it work?

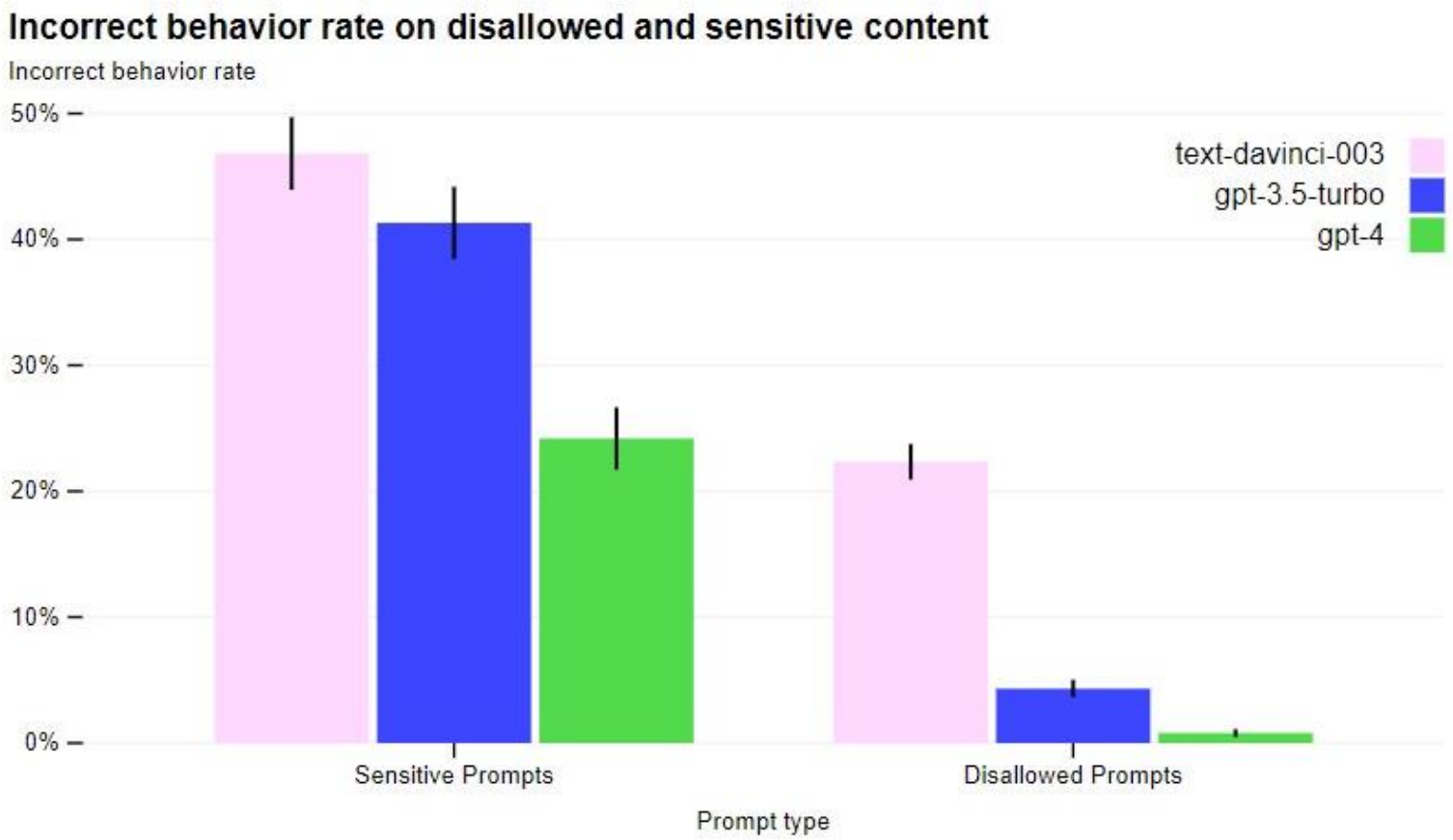


Reinforcement learning



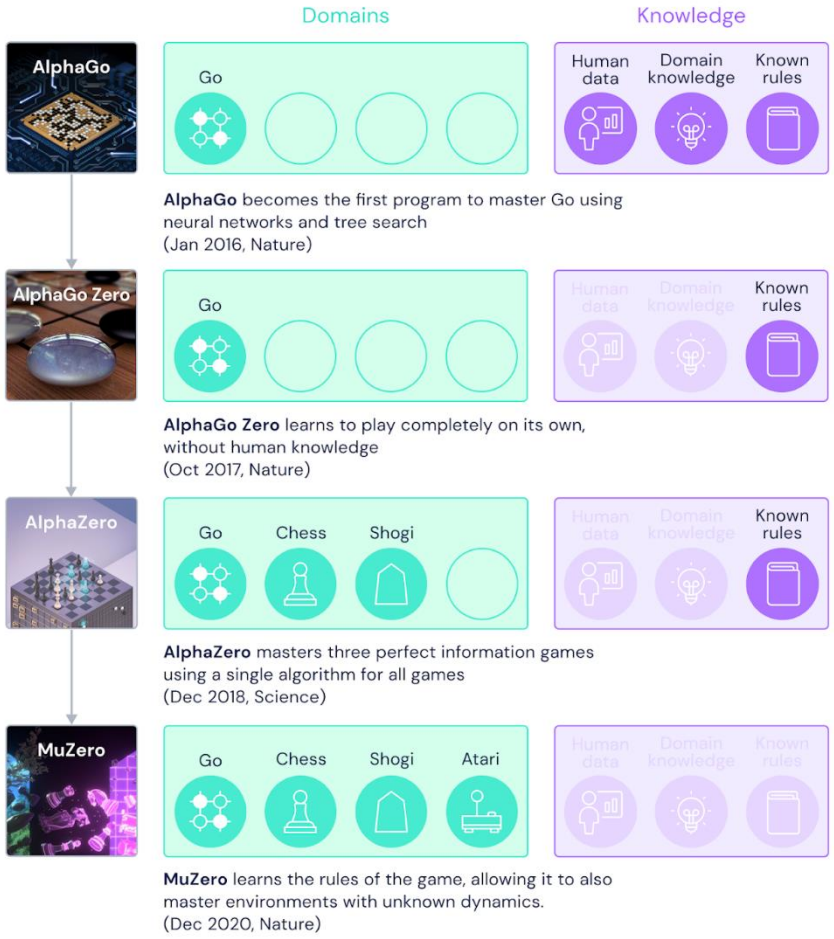
- Source: MathWorks

Safety filters



End-to-end trained general-purpose systems

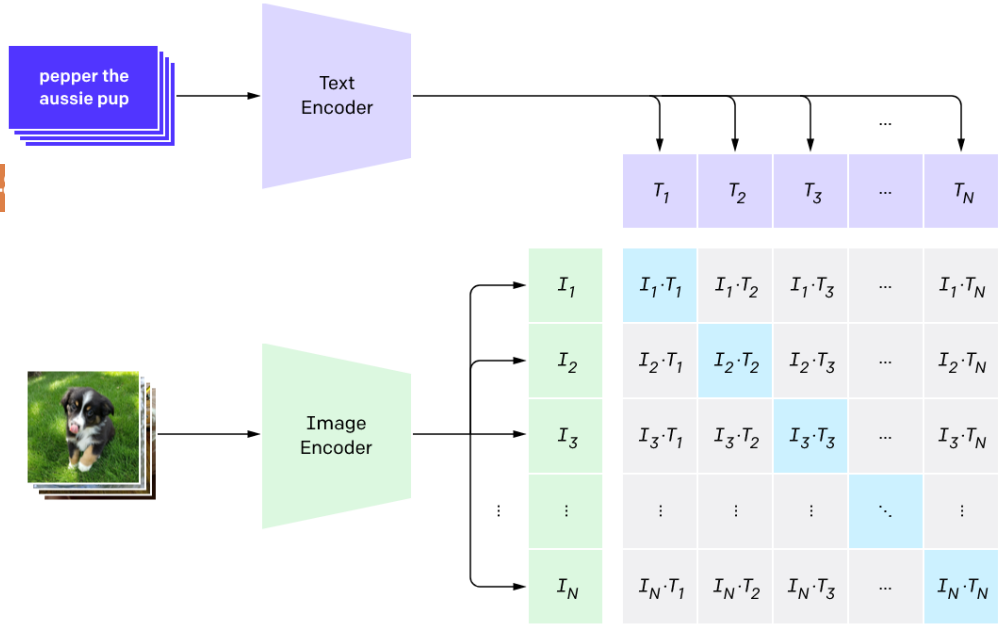
- DeepMind's MuZero (Dec 2020)
- RL algorithm
- Using deep network embeddings as representations
- Only represents relevant features for decision making: value, policy, reward
- Combined with a look-ahead search
- Can improve its planning from the learned model without collecting new data



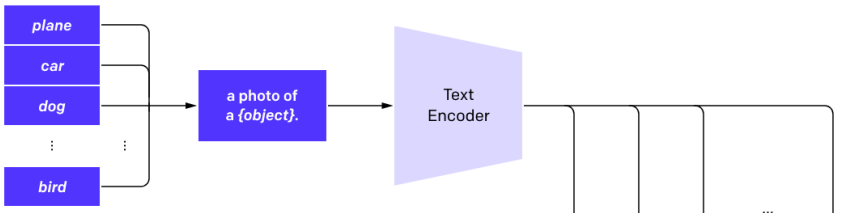
Multimodal models (language and image)

- CLIP (Contrastive Language–Image Pre-training)
  - task-agnostic training from image data paired with text available on Internet
  - training task for CLIP: given an image, predict which out of a set of 32,768 randomly sampled text snippets, was actually paired with it in our dataset.

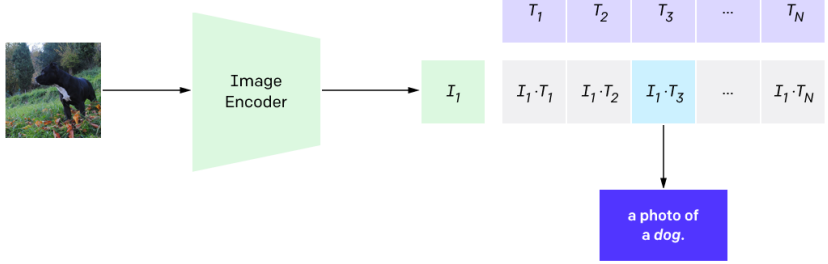
1. Contrastive pre-training



2. Create dataset classifier from label text



3. Use for zero-shot prediction



Multimodal models (language and image)

- DALL-E - a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text–image pairs.
- a simple decoder-only transformer that receives both the text and the image as a single stream of 1280 tokens—256 for the text and 1024 for the image—and models all of them autoregressively.

DALL-E examples

TEXT PROMPT

a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

DALL-E examples

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES

DALL-E examples

TEXT PROMPT

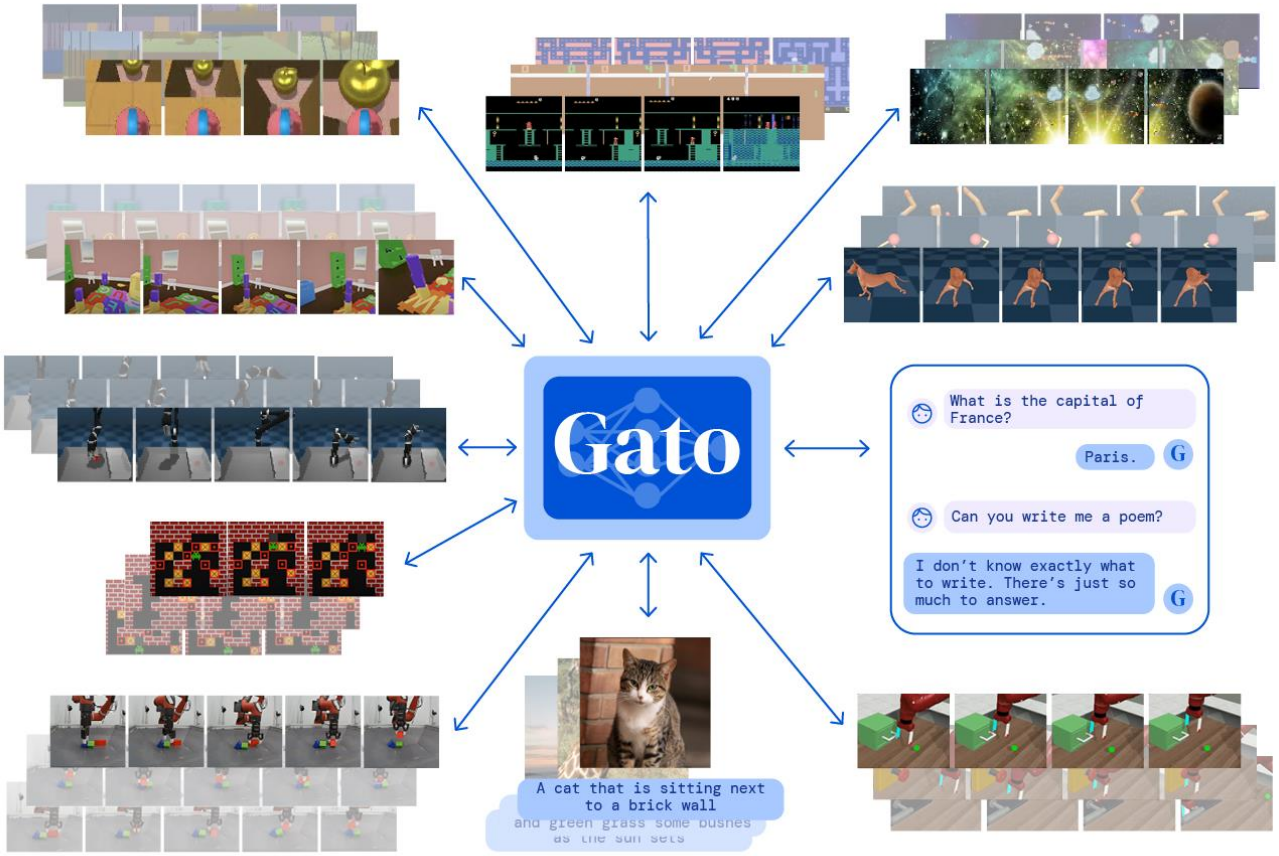
an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

GATO by DeepMind

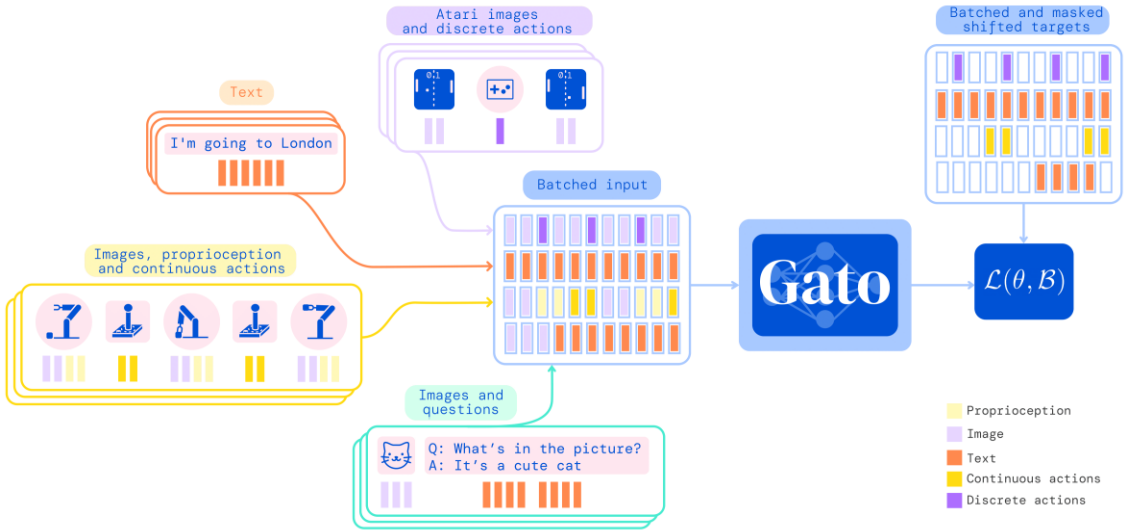
„a multi-modal, multi-task, multi-embodiment generalist policy. The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens. “

GATO (cont.)


Gato
What is the capital of France?
Paris.
Can you write me a poem?
I don't know exactly what to write. There's just so much to answer.
A cat that is sitting next to a brick wall

GATO (cont. 2)

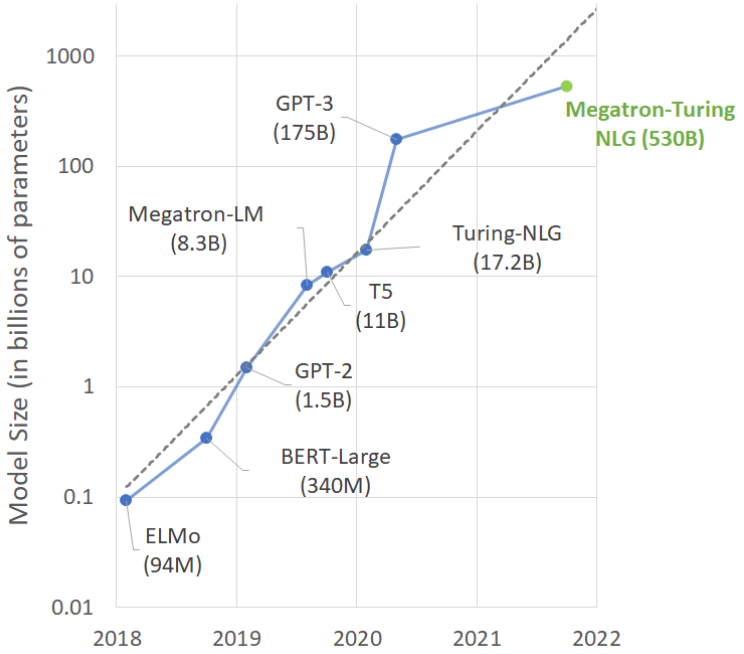
- During the training phase of Gato, data from different tasks and modalities are serialised into a flat sequence of tokens, batched, and processed by a transformer neural network similar to a large language model. The loss is masked so that Gato only predicts action and text targets.



- When deploying Gato, a prompt, such as a demonstration, is tokenised, forming the initial sequence. Next, the environment yields the first observation, which is also tokenised and appended to the sequence. Gato samples the action vector autoregressively, one token at a time.

Vast amount of parameters and training data

Year	Model	Number of parameters	Data
2019	BERT	3.4E+08	16GB
2019	DistilBERT	6.60E+07	16GB
2019	ALBERT	2.23E+08	16GB
2019	XLNet	3.40E+08	126GB
2020	ERNIE-Gen	3.40E+08	16GB
2019	RoBERTa	3.55E+08	161GB
2019	MegatronLM	8.30E+09	174GB
2020	T5-11B	1.10E+10	745GB
2020	T-NLG	1.70E+10	174GB
2020	GPT-3	1.75E+11	570GB
2020	GShard	6.00E+11	–
2021	Switch-C	1.57E+12	745GB
2021	*Wu Dao 2.0*	*1.75E+12*	*4.9TB*



Model descriptions

WuDao (*Enlightenment*) 2.0

- Developed by Beijing Academy of Artificial Intelligence (BAAI)
- 10x larger than GPT-3
- 1.75 trillion params, 4.9 TB of multimodal data (1.2TB Chinese text data in Wu Dao Corpora, 2.5TB Chinese graphic data, 1.2TB English text data in the Pile dataset)
- can solve multimodal tasks: "perform natural language processing, text generation, image recognition, and image generation tasks, captioning images and creating nearly photorealistic artwork, given natural language descriptions."
- can learn different tasks over time, not forgetting what it has learned previously

Is bigger better?

What should be clear by now:

- Symbolic versus subsymbolic representation
- Distributed representation
- Gradedness
- Graceful degradation
- Robustness
- Feedback
- Neural architecture & knowledge

Questions?