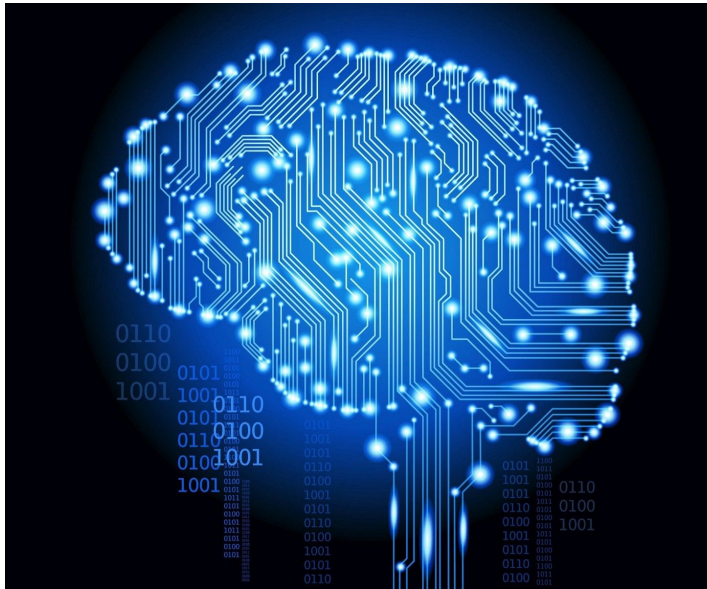


Neuromorphic computing



CS4575

Sustainable Software Engineering

05.03.2025

Nergis Tömen

Introduction



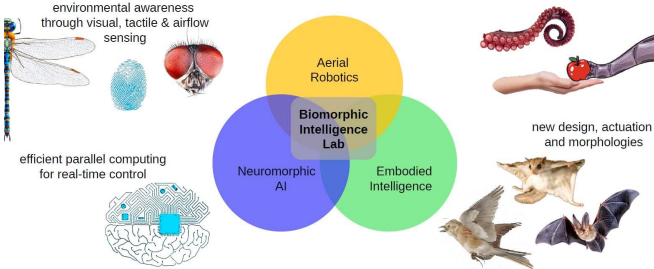
Nergis

Computer Vision Lab

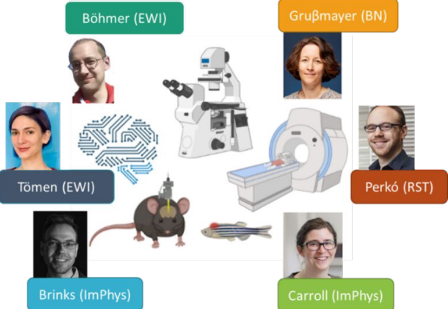


Biomorphic Intelligence Lab

Bio-inspired solutions for aerial robotics



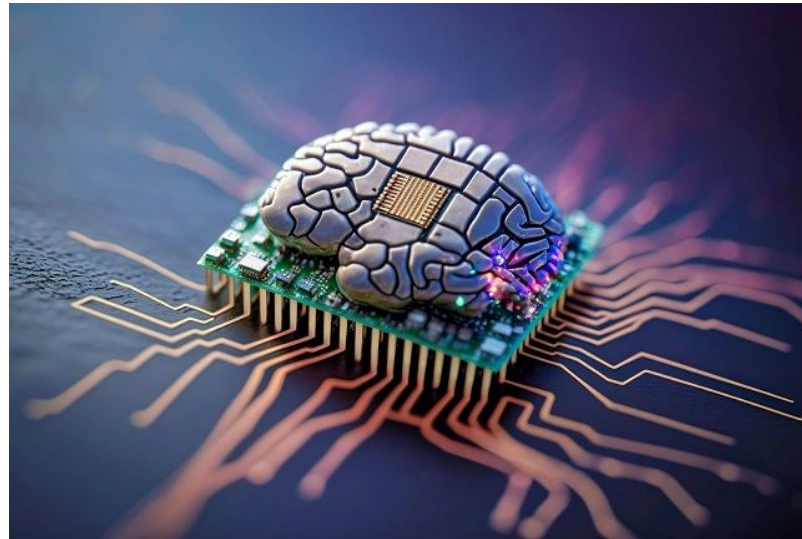
Biomedical Intervention Optimisation Lab



Neuromorphic computing

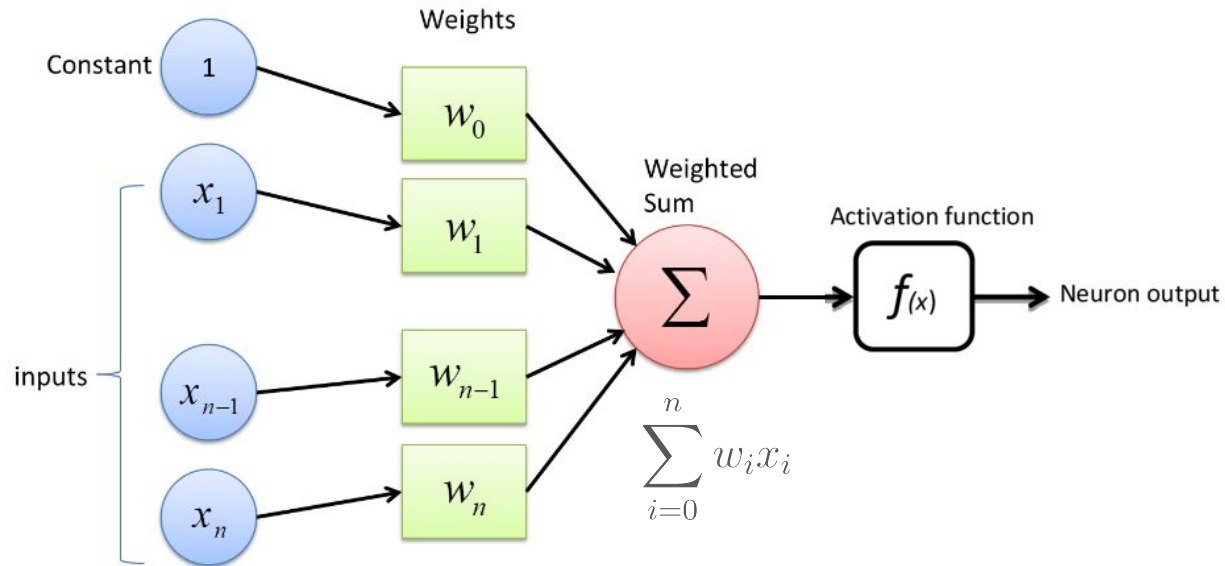
Neuromorphic computing is an approach to computing that is inspired by the **structure and function of the human brain**.

A neuromorphic computer/chip is any **device** that uses **physical artificial neurons** to do computations.

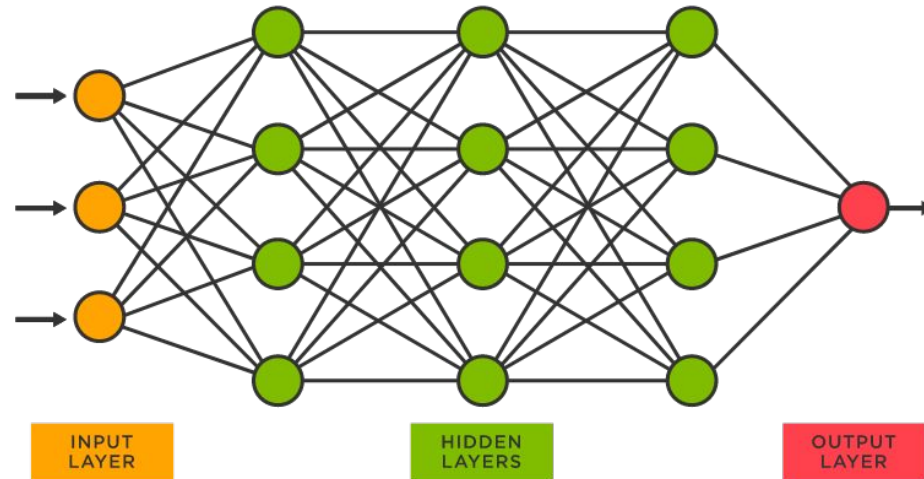


How many of you are familiar
with neural networks?

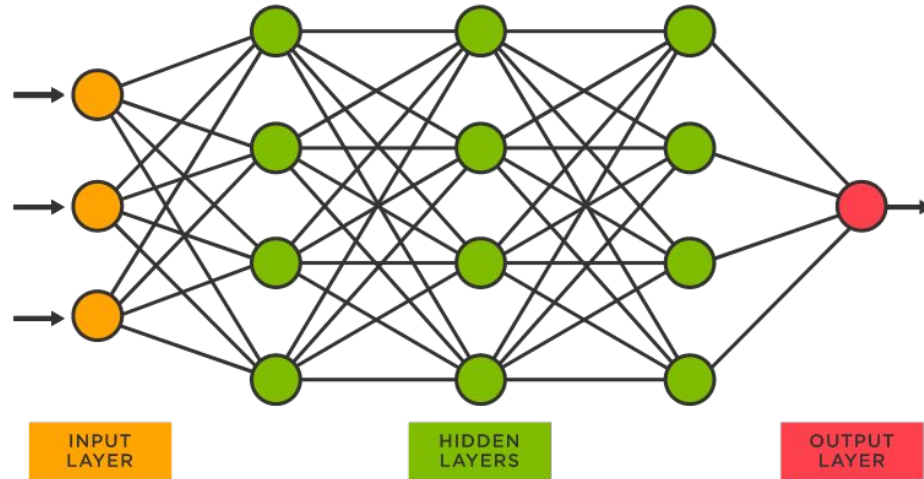
Simple model of an artificial neuron



What is a neural network?



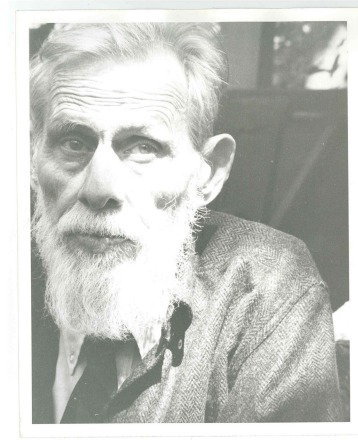
Why is it called a 'neural network'?



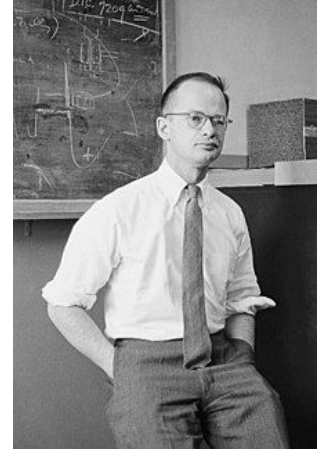
What is a neural network?

Why is it called a 'neural network'?

McCulloch-Pitts neuron [1, 2] (1943)



Warren Sturgis McCulloch
(Neurophysiologist)



Walter Pitts
(Logician)

I. Introduction

Theoretical neurophysiology rests on certain cardinal assumptions. **The nervous system is a net of neurons**, each having a soma and an axon. Their adjunctions, or synapses, are always between the axon of one neuron and the soma of another. At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse.

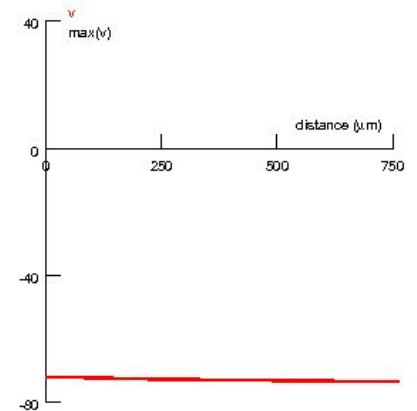
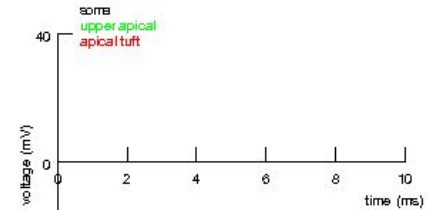
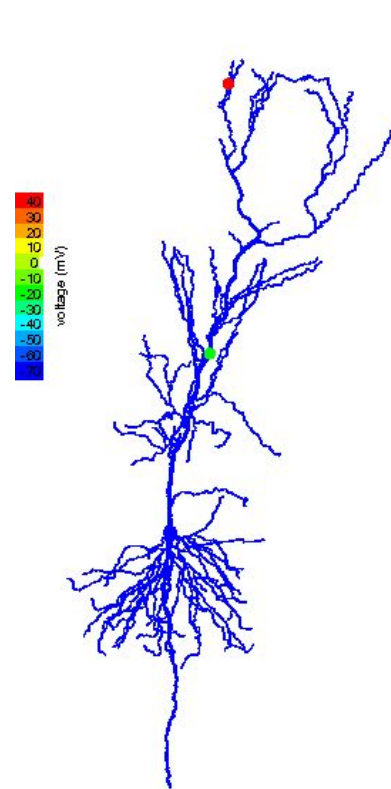
How do biological neurons work?

'Input current' travels down the dendrites (top),

get **integrated** (summed!) in the cell body

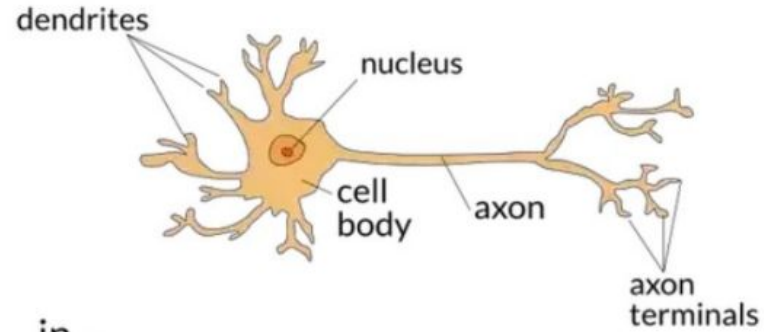
which generates an '**output current**' (bottom)

which is chemically transmitted to the dendrites of other neurons.

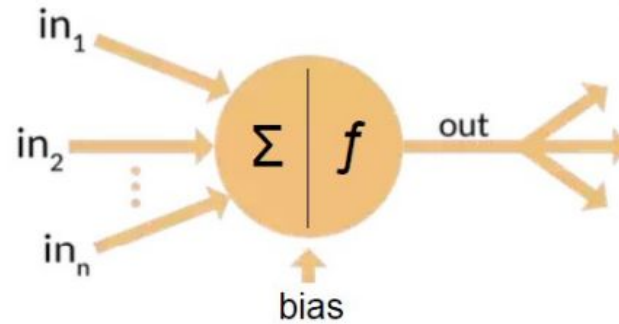


Simplified picture

Real neuron



Artificial neuron



What is a neural network?

McCulloch-Pitts neuron [1, 2]

1940s: How do biological neurons compute basic logic functions? (e.g. logic gates)

Note: Ref. [3] gives a nice brief history on the ideas which lead to the McCulloch-Pitts neuron.

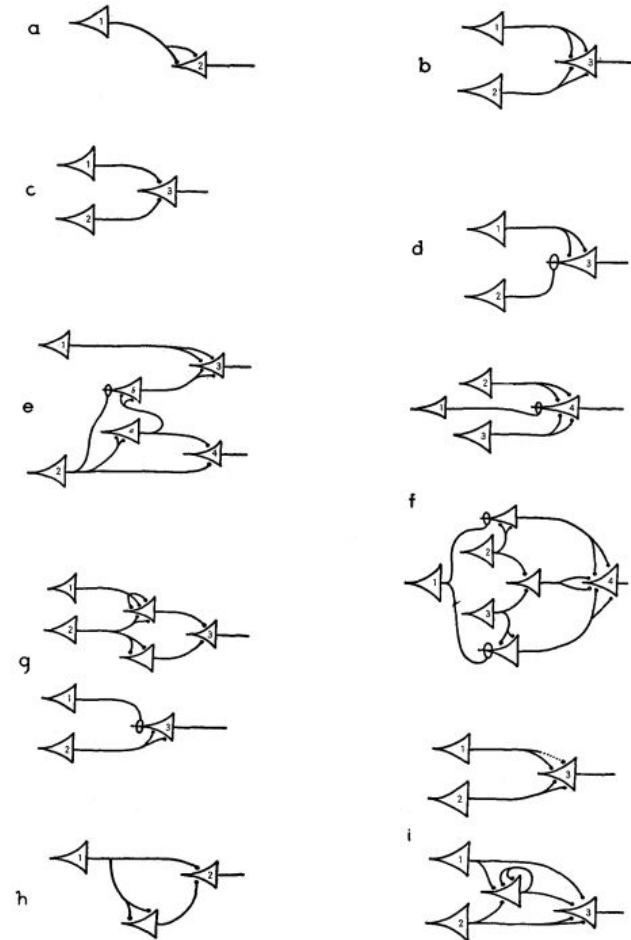


FIGURE 1

What is a neural network?

1950s: How are neurons organized to perform sensory perception?

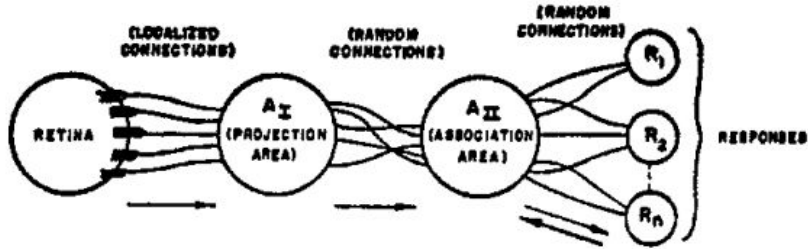


FIG. 1. Organization of a perceptron.

ticular response; i.e., the information is contained in *connections* or *associations* rather than topographic representations. (The term *response*, for

What is a neural network?

1950s: How are neurons organized to perform sensory perception?

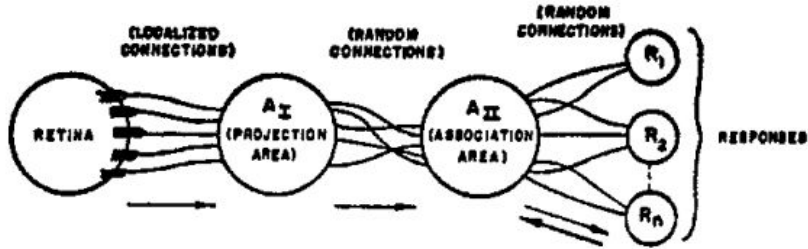
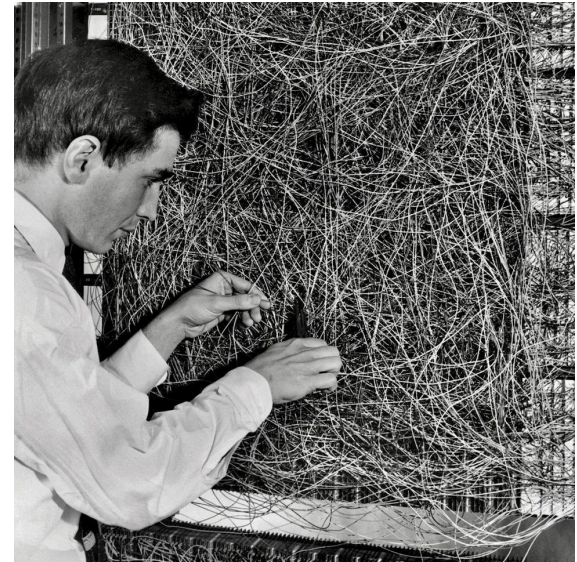


FIG. 1. Organization of a perceptron.

ticular response; i.e., the information is contained in connections or associations rather than topographic representations. (The term response, for

Frank Rosenblatt (Psychologist)
with a Mark I Perceptron computer in 1960



The first "neural network": Perceptron (1958). [4]

What is a neural network?

1950s: How are neurons organized to perform sensory perception?

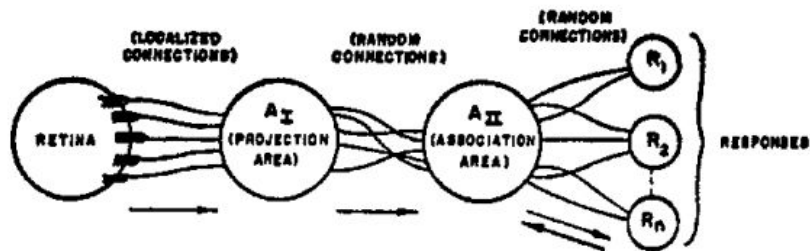
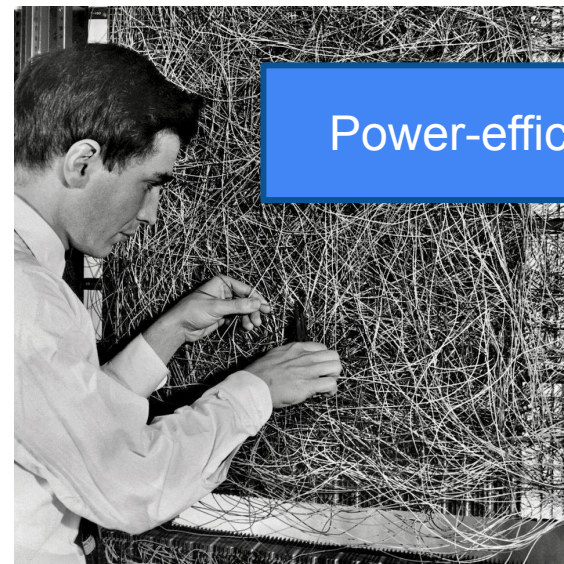


FIG. 1. Organization of a perceptron.

ticular response; i.e., the information is contained in connections or associations rather than topographic representations. (The term response, for

Frank Rosenblatt (Psychologist)

with a Mark I Perceptron computer in 1960



The first "neural network": Perceptron (1958). [4]

Neuromorphic computing

Why neuromorphic computing?

Neuromorphic computing is an approach to computing that is inspired by the **structure and function of the human brain**.

A neuromorphic computer/chip is any **device** that uses **physical artificial neurons** to do computations.

Neuromorphic computing

Why neuromorphic computing?

'**Biological inspiration**' for artificial neural networks (ANNs) is not a new idea.

Emulation (as opposed to simulation) of neural networks in hardware is not a new idea.

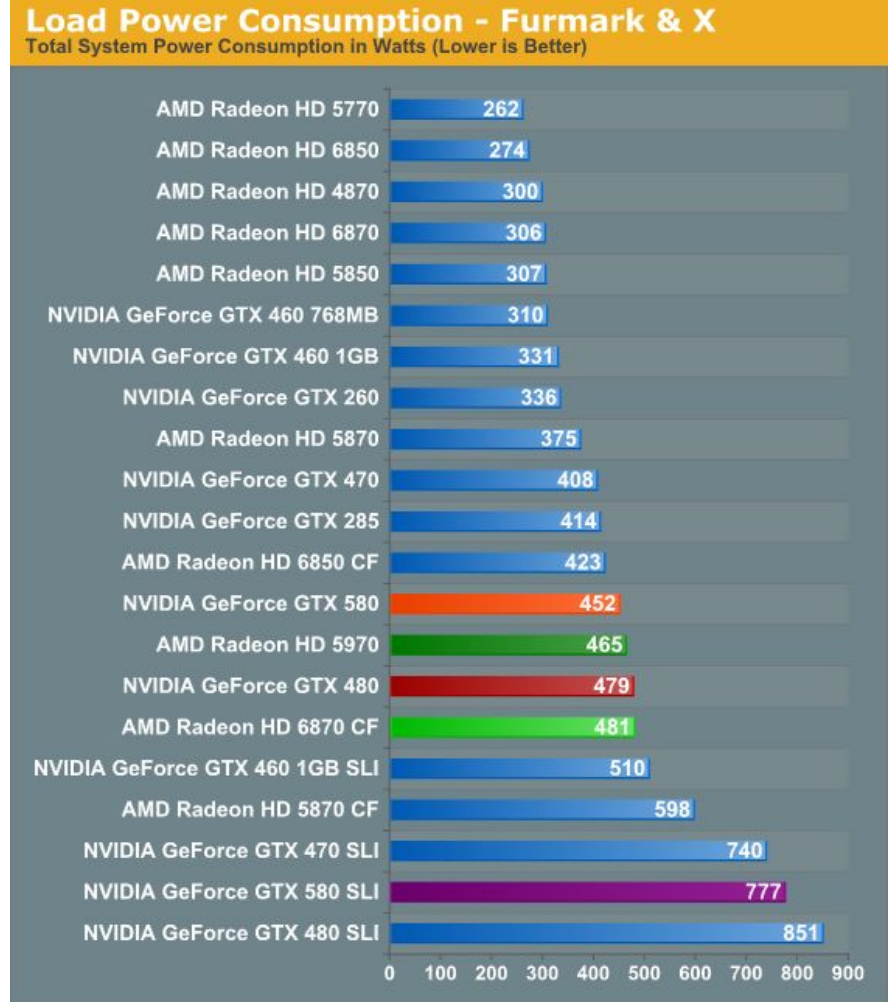
Question: What can we **gain** from **increasing** biological realism in **existing** neural networks?

Power-efficiency at scale

Modern, deep neural networks*
are trained using **GPUs**.

* It is estimated that ChatGPT was trained on 10,000-20,000 GPUs and that it will require **30,000 GPUs** to keep running stably in the future.

* It is estimated that ChatGPT has **10-20 billion parameters**.



Power-efficiency at scale

Modern, deep neural networks*
are trained using **GPUs**.

* It is estimated that ChatGPT was trained on 10,000-20,000 GPUs and that it will require **30,000 GPUs** to keep running stably in the future.

* It is estimated that ChatGPT has **10-20 billion parameters**.

Load Power Consumption - Furmark & X

Total System Power Consumption in Watts (Lower is Better)



Single model with 20 billion parameters:

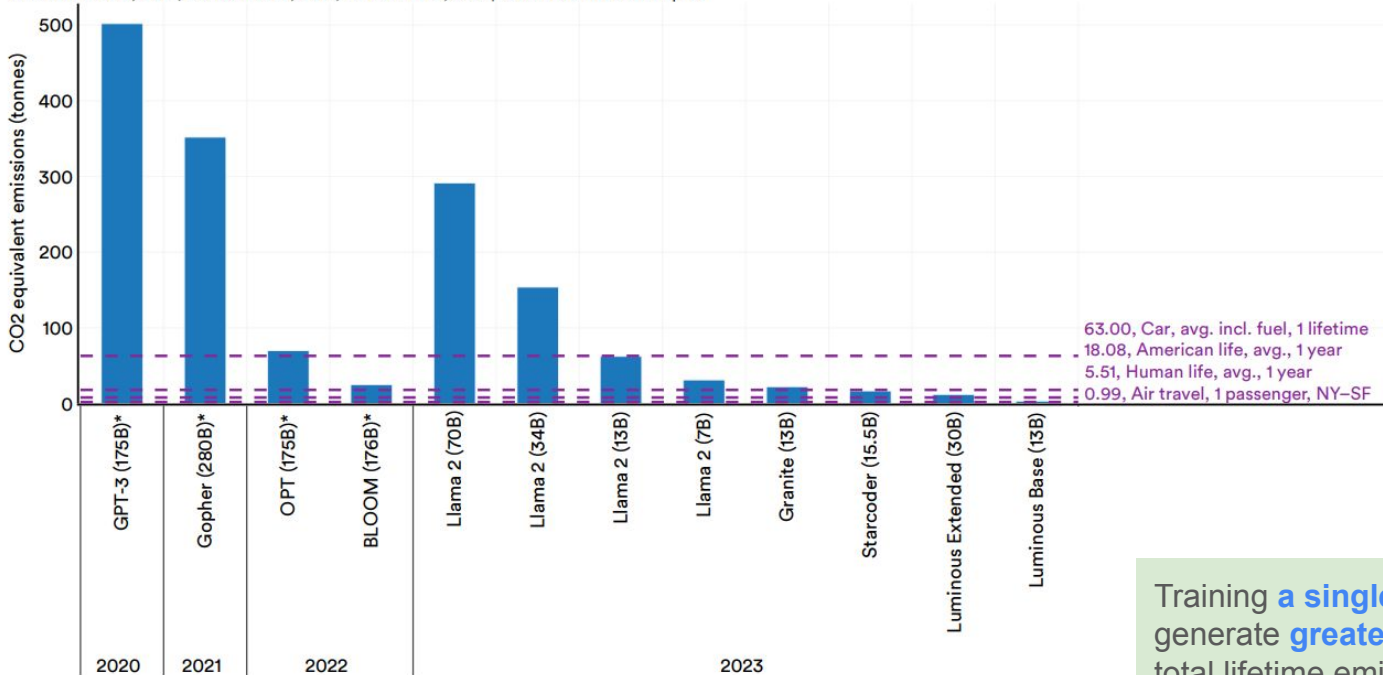
200 Watts x 30,000 GPUs = **6M Watts**



Power-efficiency at scale

CO2 equivalent emissions (tonnes) by select machine learning models and real-life examples, 2020–23

Source: AI Index, 2024; Luccioni et al., 2022; Strubell et al., 2019 | Chart: 2024 AI Index report



Training a **single** large language model can generate **greater CO2** emissions than the total lifetime emissions of 8 cars (in **2020**)!

Power-efficiency at scale

Environmental impact of select models

Source: AI Index, 2024; Luccioni et al., 2022 | Table: 2024 AI Index

Model and number of parameters	Year	Power consumption (MWh)
Gopher (280B)	2021	1,066
BLOOM (176B)	2022	433
GPT-3 (175B)	2020	1,287
OPT (175B)	2022	324
Llama 2 (70B)	2023	400
Llama 2 (34B)	2023	350
Llama 2 (13B)	2023	400
Llama 2 (7B)	2023	400
Granite (13B)	2023	153
StarCoder (15.5B)	2023	89.67
Luminous Base (13B)	2023	33
Luminous Extended (30B)	2023	93

Training **a single** model can consume more than **1000 MWh** of power!

Power-efficiency at scale

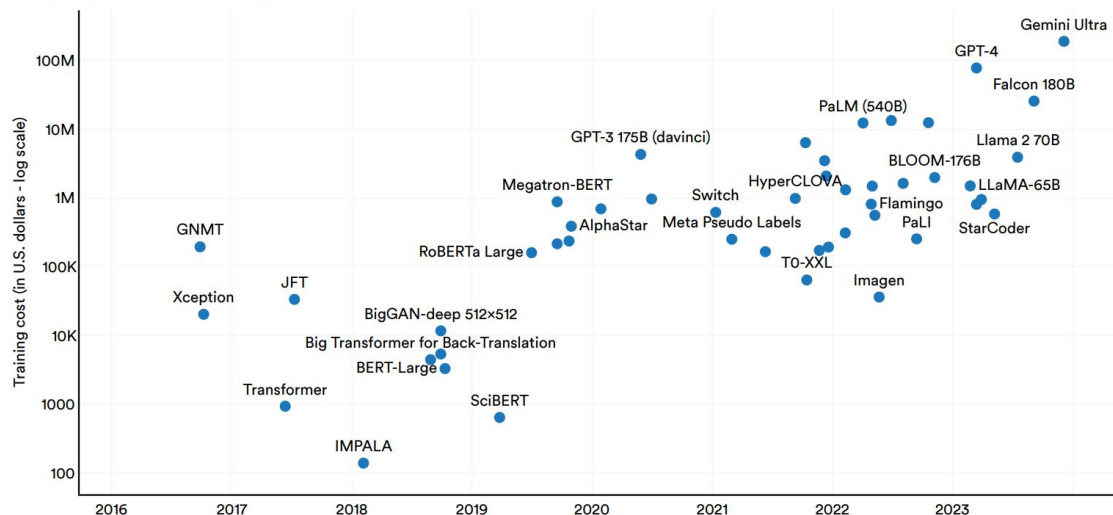
Environmental impact of select models

Source: AI Index, 2024; Luccioni et al., 2022 | Table: 2024 AI Index

Model and number of parameters	Year	Power consumption (MWh)
Gopher (280B)	2021	1,066
BLOOM (176B)	2022	433
GPT-3 (175B)	2020	1,287
OPT (175B)	2022	324
Llama 2 (70B)	2023	400
Llama 2 (34B)	2023	350
Llama 2 (13B)	2023	400
Llama 2 (7B)	2023	400
Granite (13B)	2023	153
StarCoder (15.5B)	2023	89.67
Luminous Base (13B)	2023	33
Luminous Extended (30B)	2023	93

Estimated training cost of select AI models, 2016–23

Source: Epoch, 2023 | Chart: 2024 AI Index report



Training **a single** model can consume more than **1000 MWh** of power!

... with energy costs reaching **200M USD!**

Power-efficiency at scale

Your brain runs on:



Power-efficiency at scale

Your brain runs on:

High estimate ~3000 kcal a day

≈145 Watts

* Human brain has **~600 trillion synapses**
(≈parameters).



Power-efficiency at scale

Your brain runs on:

High estimate ~300 W
≈145 W

* Human brain has ~600 trillion
(≈parameters).

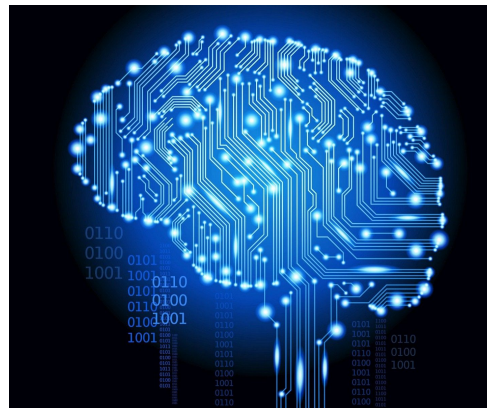
Oversimplification

There are also multiple other advantages...



Human vs. computer computation

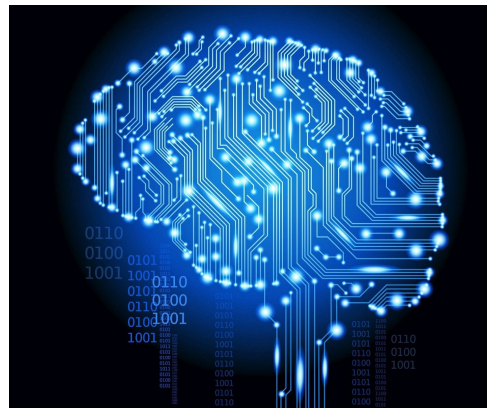
- **Fast** real-time decision making, e.g. sports, e-sports
- **Adaptive**, e.g. context-aware and employs selective attention
- **Energy efficient**: Close to 100 billion neurons in the brain
- **Robust**, for example to changes in illumination or obstructions in object tracking



Human vs. computer computation

Brains are **energy efficient**: Why?

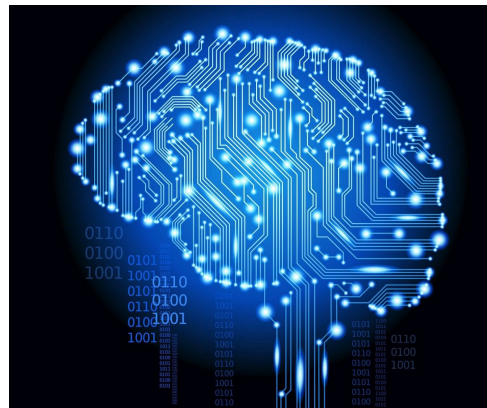
1. High **temporal resolution** (more computation with less neurons)
2. **Sparse** encoding



Human vs. computer computation

Brains are **energy efficient**: Why?

1. High **temporal resolution** (more computation with less neurons)
2. **Sparse** encoding

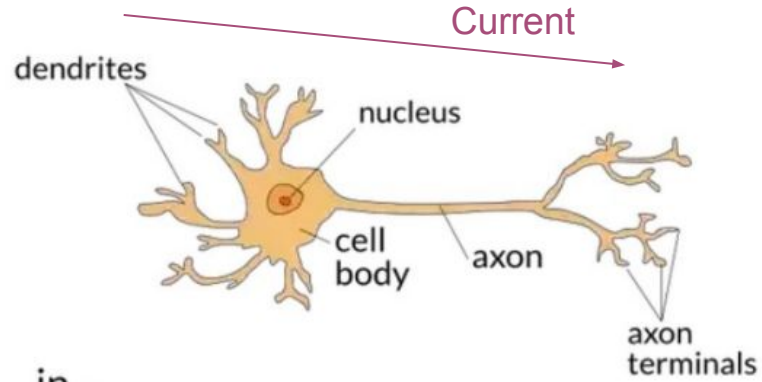


Questions?

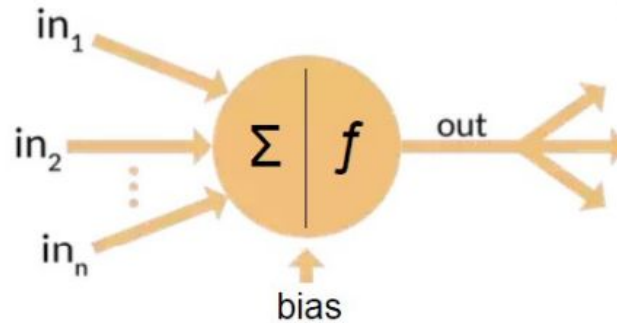
How do biological neurons communicate?

Analogy to artificial neural networks

Real neuron

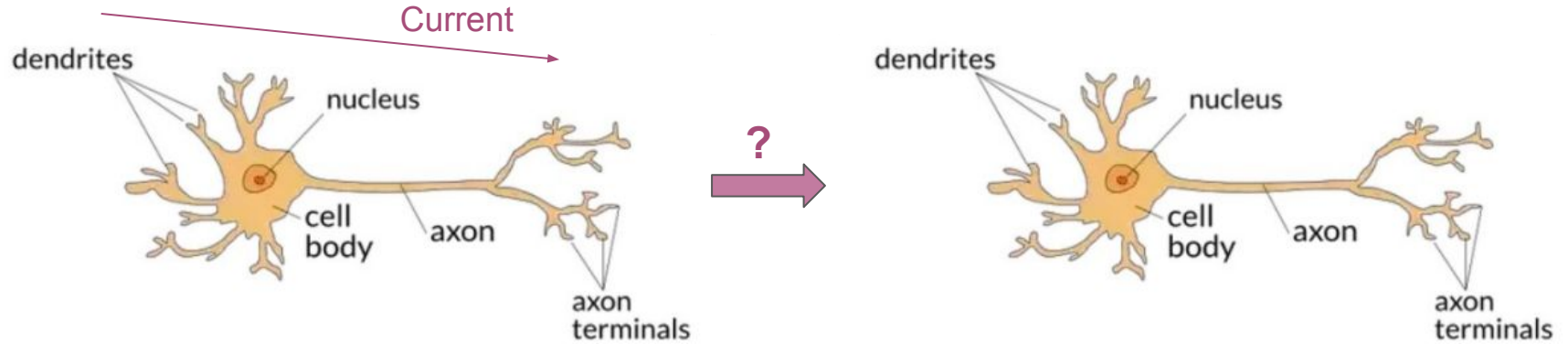


Artificial neuron

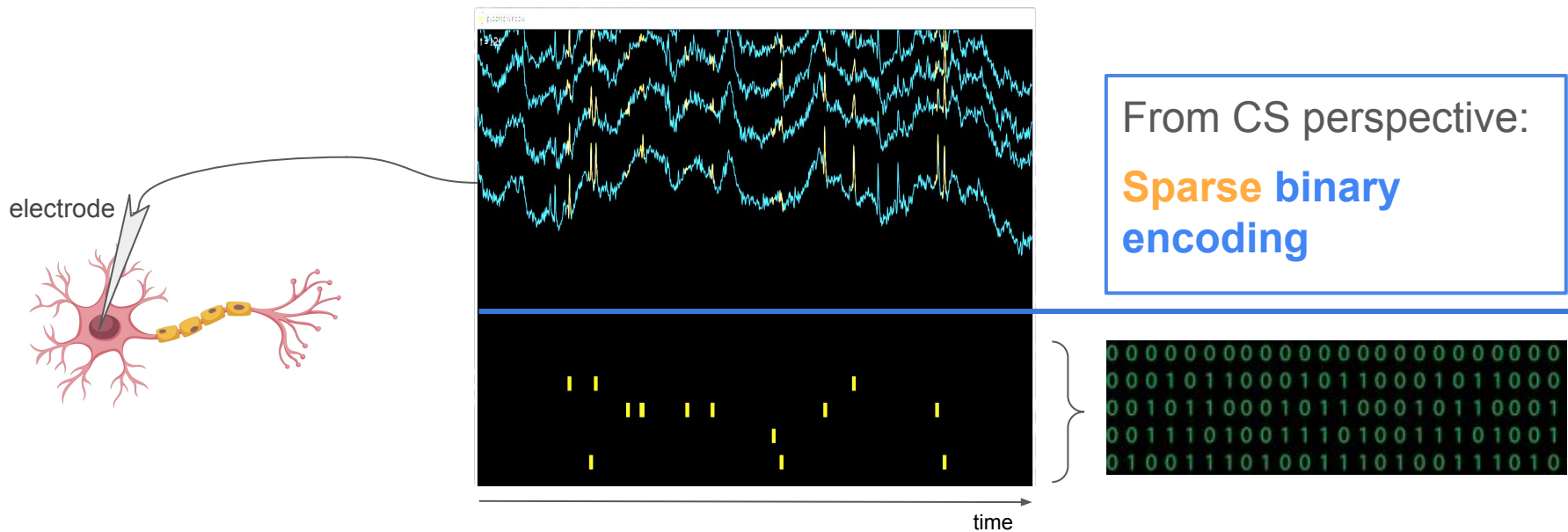


How do biological neurons communicate?

How does the electrical activity propagate?



How do biological neurons work?



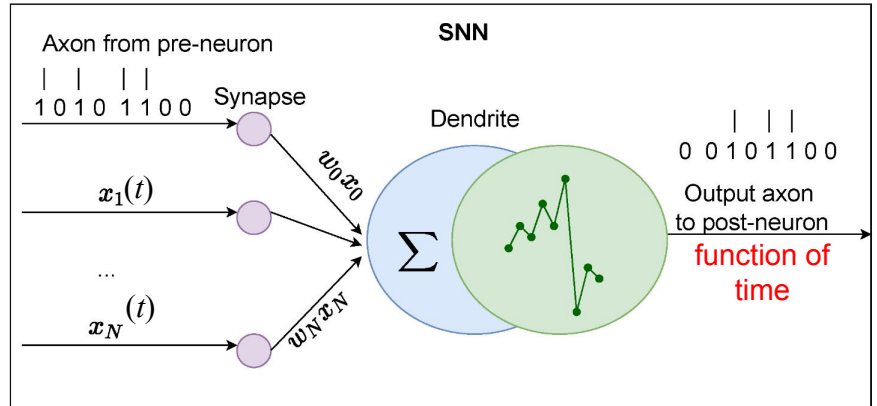
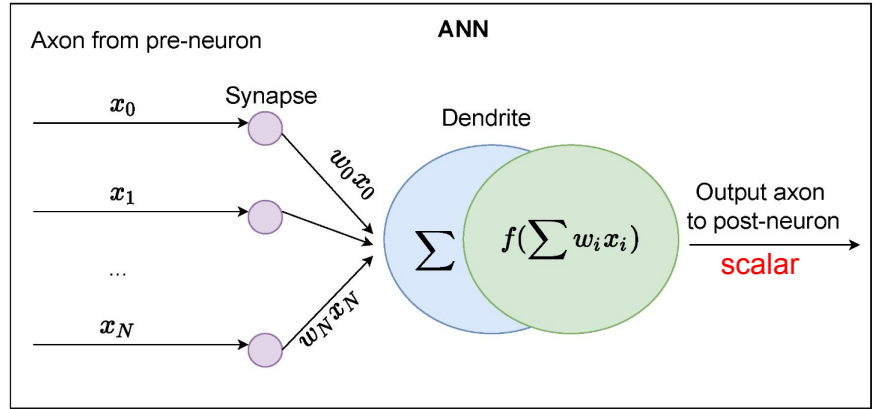
Quick **electrical pulses** trigger **chemical signals** for the next neuron → **Spikes**

Biologically realistic **spiking** neuron models

Biologically realistic neuron models have a new dimension: **Time!**

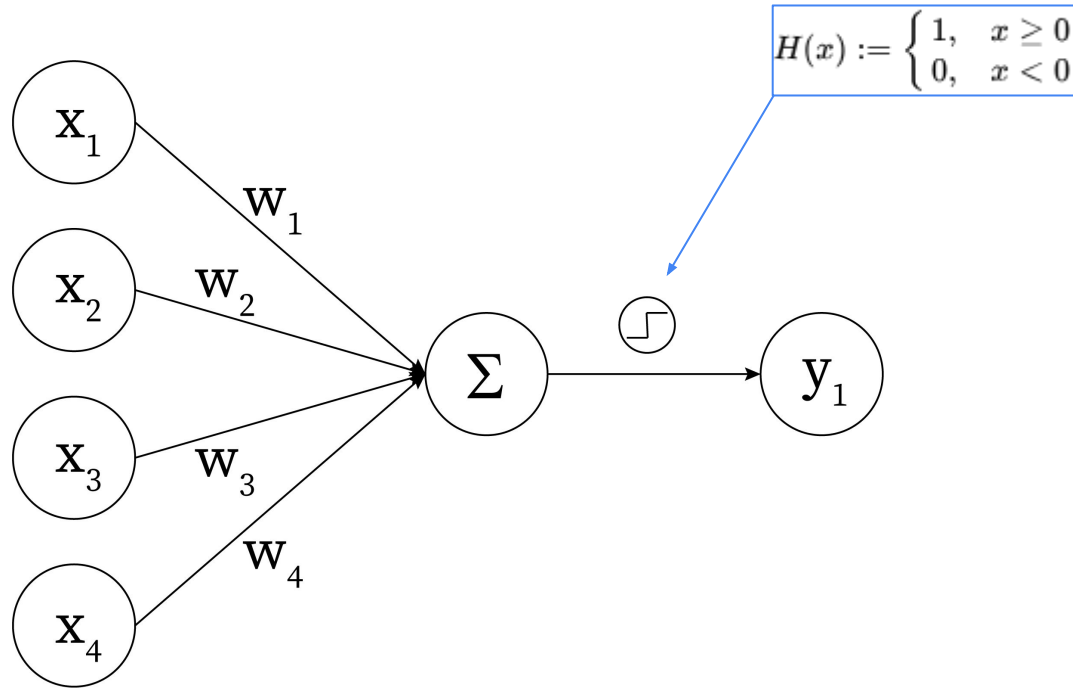
Spiking neural networks (SNNs):

The input $\mathbf{x}(t)$ to each neuron is summed (integrated) over time.



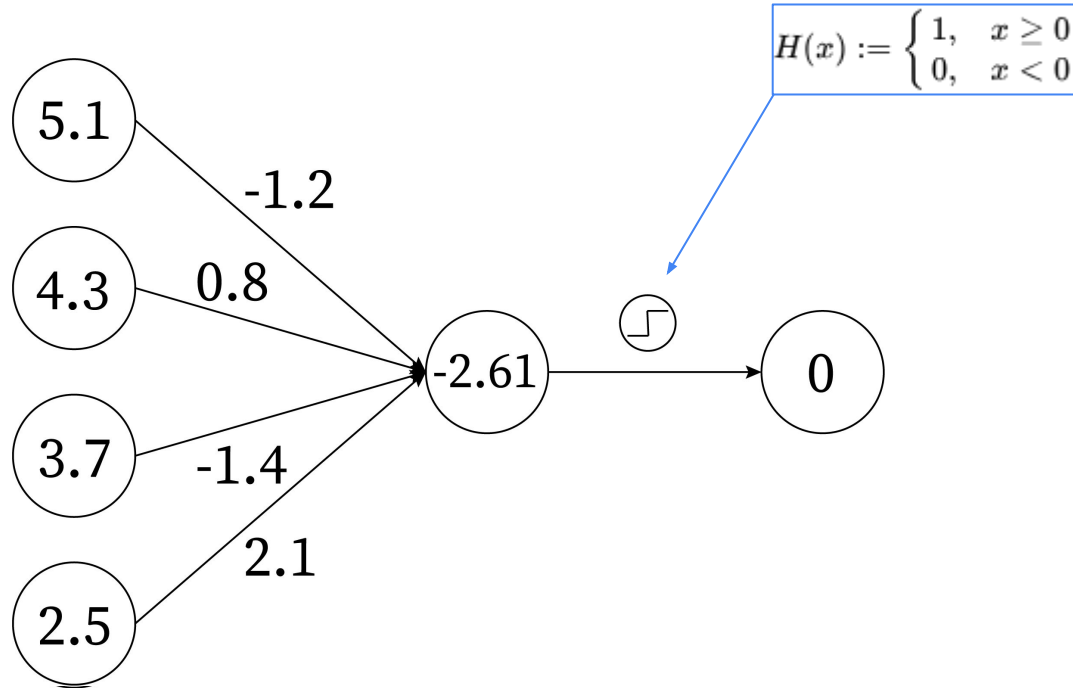
How to model spiking neurons?

ANN: Perceptron, threshold activation function:



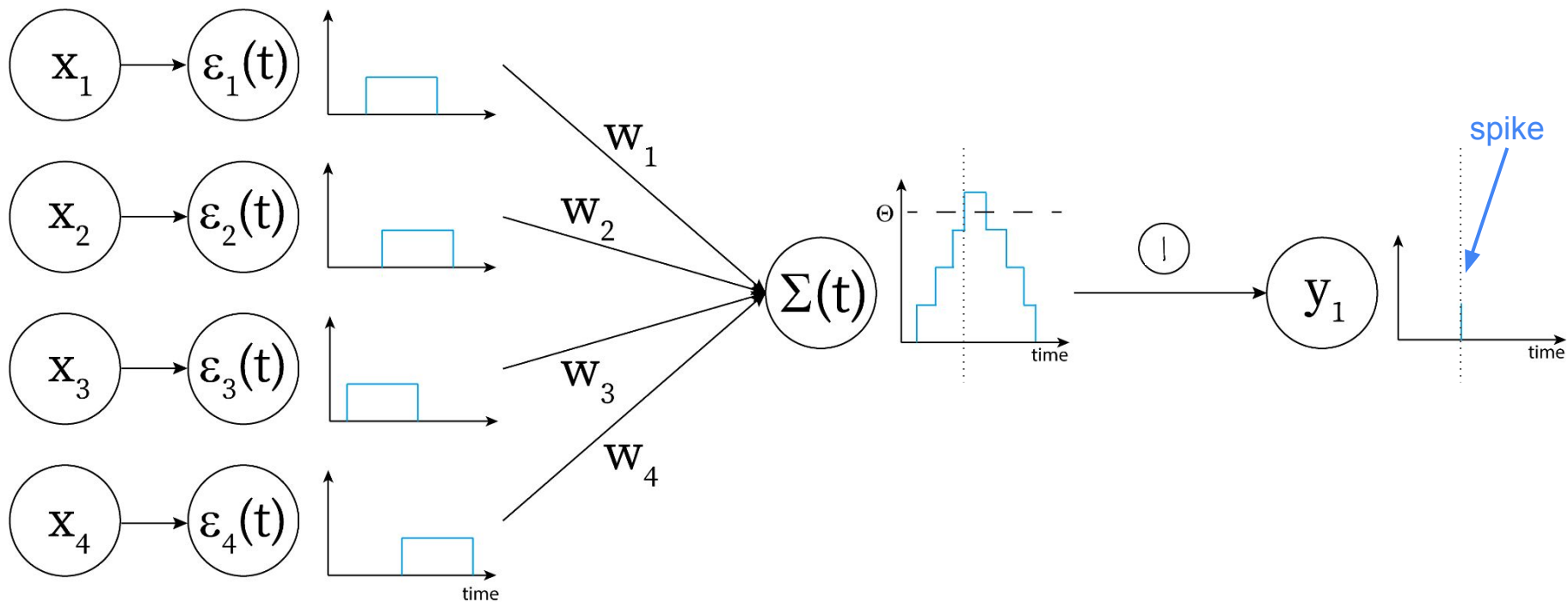
How to model spiking neurons?

ANN: Perceptron, threshold activation function:



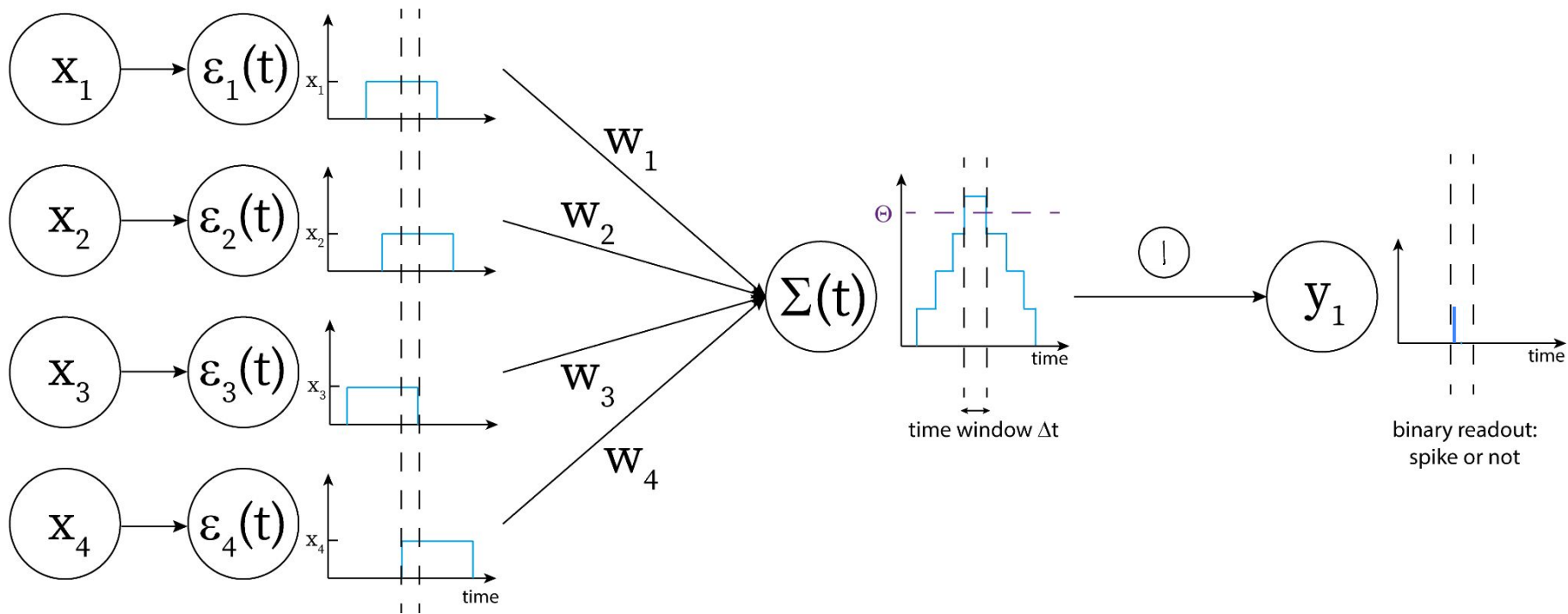
How to model spiking neurons?

Spiking neural network (SNN): The 'input current' $\epsilon(t)$ is integrated over time.



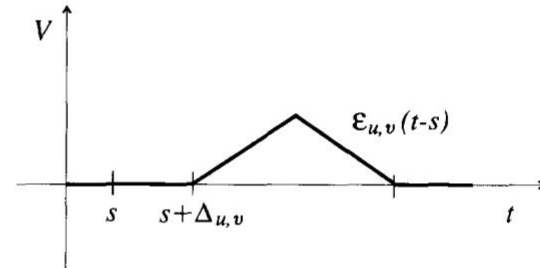
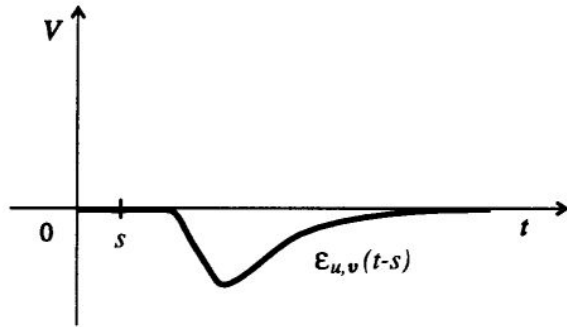
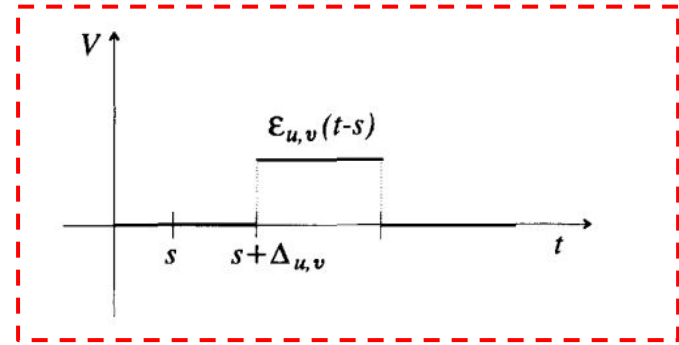
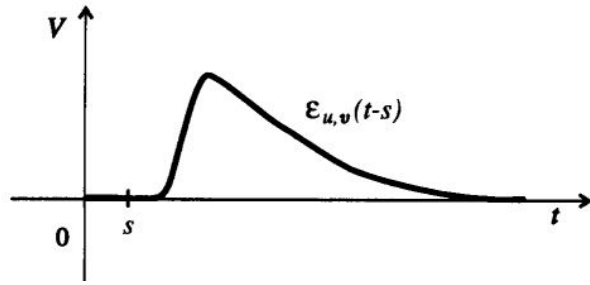
How to model spiking neurons?

Equivalence to perceptron: Computation at least as complex as a perceptron.



Non-leaky integrate-and-fire (IF) neuron

The temporal profile of the input current $\epsilon(t)$ can be chosen differently, for different computations.



Questions?

Computation with spiking neurons

Coincidence detection

we consider the concrete boolean function $CD_n: \{0, 1\}^{2n} \rightarrow \{0,1\}$, which is defined by

$$CD_n(x_1, \dots, x_n, y_1, \dots, y_n) = \begin{cases} 1, & \text{if } x_i = y_i = 1 \\ & \text{for some } i \in \{1, \dots, n\} \\ 0, & \text{otherwise.} \end{cases}$$

This function appears to be relevant in a biological context, since it formalizes some form of pattern-matching, respectively, *coincidence-detection*.

Computation with spiking neurons

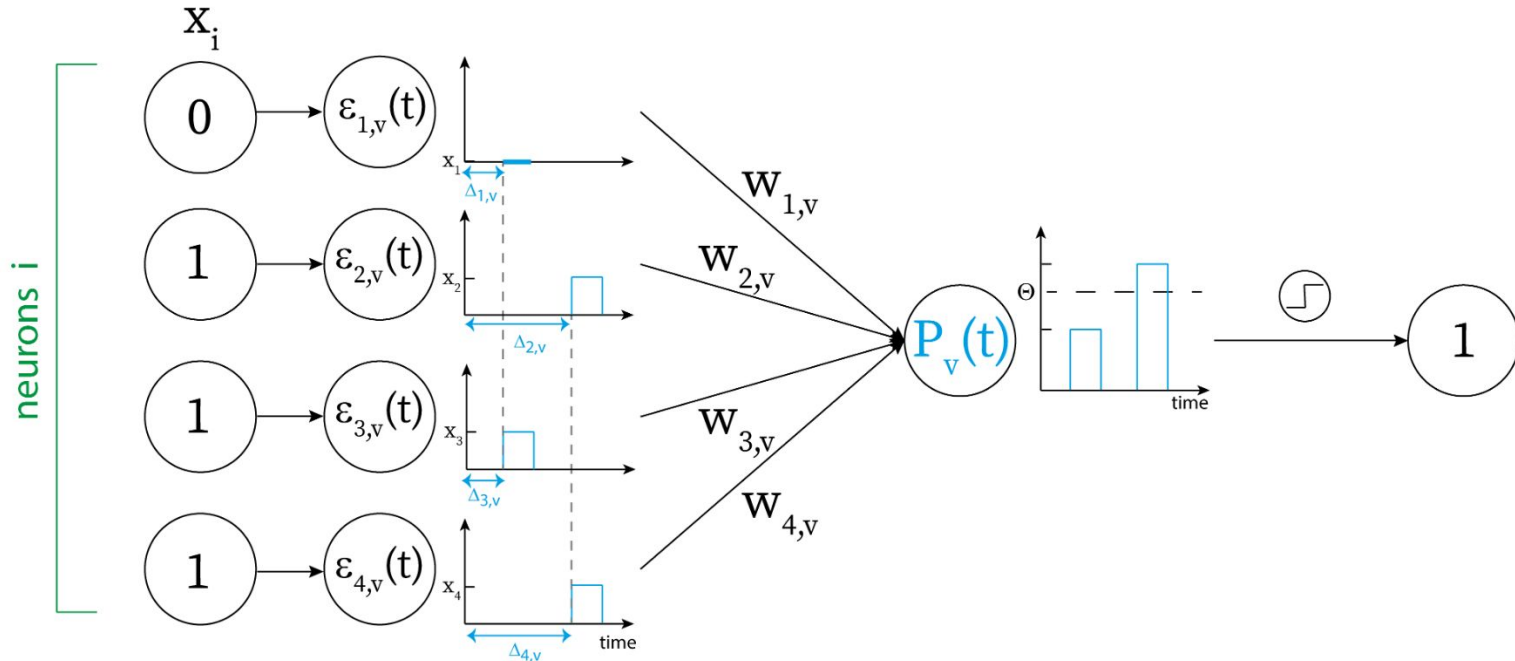
Coincidence detection

$$CD_n(x_1, \dots, x_n, y_1, \dots, y_n) = \begin{cases} 1, & \text{if } x_i = y_i = 1 \\ & \text{for some } i \in \{1, \dots, n\} \\ 0, & \text{otherwise.} \end{cases}$$

$$x, y \in \{0, 1\}^n \longrightarrow \text{Example: } n=2 \longrightarrow \text{input} = \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \longrightarrow \text{output?}$$

Computation with spiking neurons

Coincidence detection



Computation with spiking neurons

Coincidence detection

$$CD_n(x_1, \dots, x_n, y_1, \dots, y_n) = \begin{cases} 1, & \text{if } x_i = y_i = 1 \\ & \text{for some } i \in \{1, \dots, n\} \\ 0, & \text{otherwise.} \end{cases}$$

Can be trivially computed with a **single** spiking neuron! Requires at least $n/\log(n+1)$ hidden units for a perceptron (proof in [7]).

Computation with spiking neurons

Coincidence detection

$$CD_n(x_1, \dots, x_n, y_1, \dots, y_n) = \begin{cases} 1, & \text{if } x_i = y_i = 1 \\ & \text{for some } i \in \{1, \dots, n\} \\ 0, & \text{otherwise.} \end{cases}$$

Can be trivially computed with a **single** spiking neuron! Requires at least $n/\log(n+1)$ hidden units for a perceptron (proof in [7]).

Brains are **energy efficient**:

1. High **temporal resolution** (more computation with less neurons)

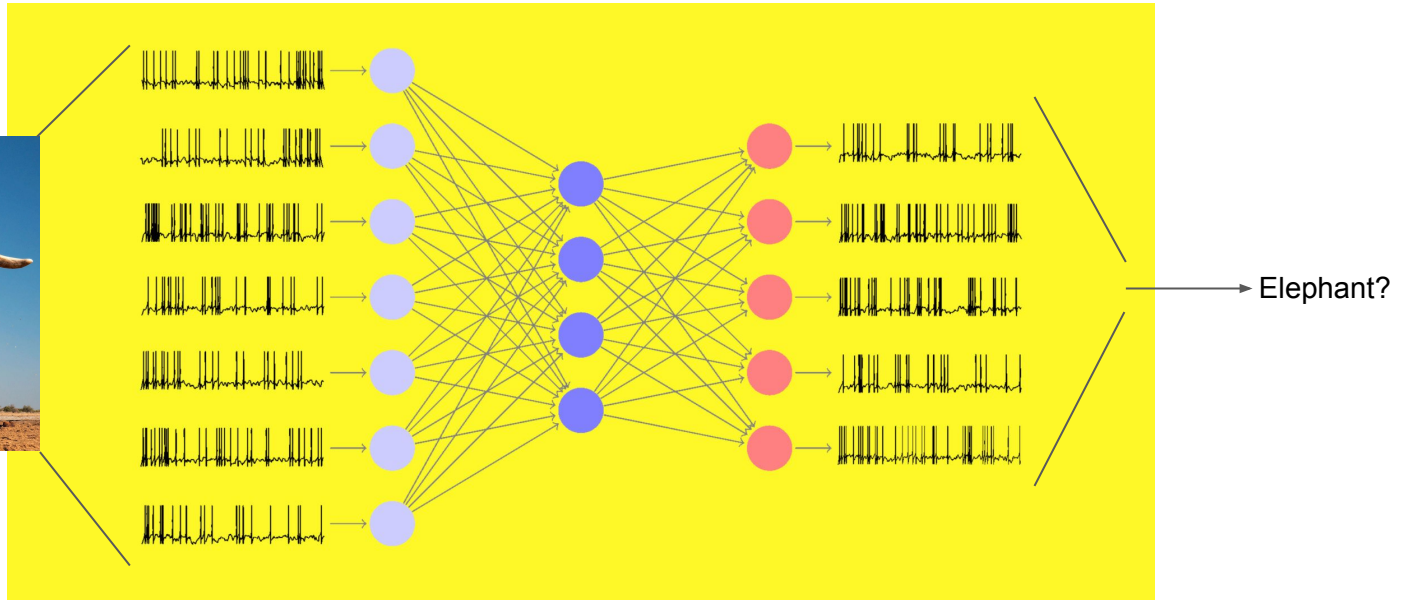
Encoding strategies

We considered **single neurons** with Boolean output ('spike'=1 or 'no spike'=0).

How should we encode information about 'features' in a **large network with many spikes?**



<https://www.nationalgeographic.com/animals/mammals/facts/african-elephant>



Firing rates

Classical view of the brain:

- Each neuron is selective for one **specific feature** in the input.

- **Higher firing rate** (spikes per unit time) for 'selected' feature.

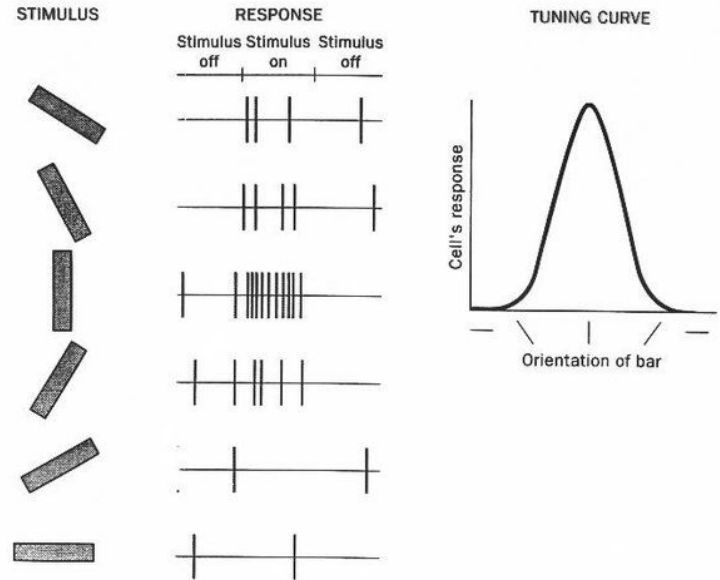
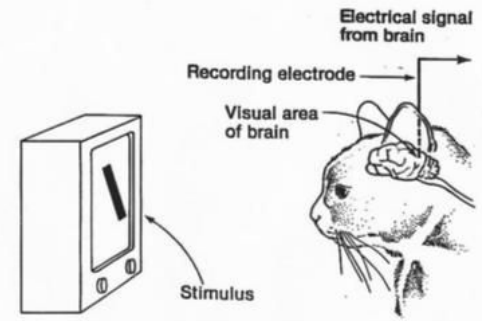


FIGURE 4.8 Response of a single cortical cell to bars presented at various orientations.

Firing rates

Classical view of the brain:

- Each neuron is selective for one **specific feature** in the input.
- **Higher firing rate** (spikes per unit time) for 'selected' feature.
- Link to **modern ANNs**: The scalar output of an artificial neuron is interpreted as the firing rate.

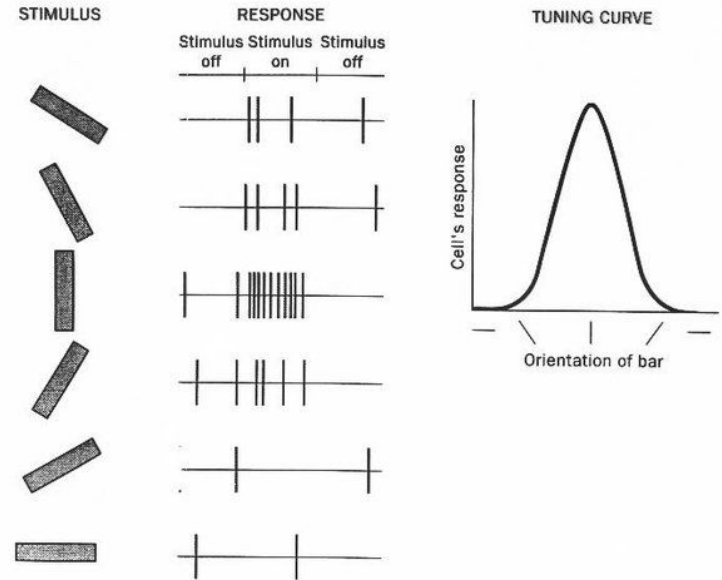
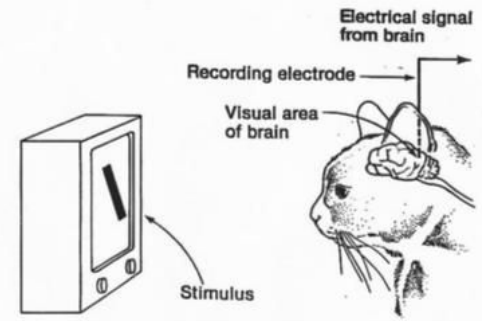


FIGURE 4.8 Response of a single cortical cell to bars presented at various orientations.

Firing rates

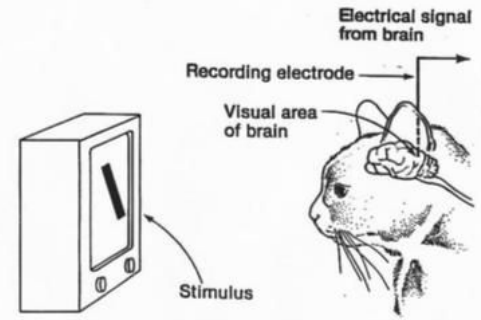
Classical view of the brain:

- Each neuron is
feature in the input

- **Higher firing rate**
'selected' feature

- Link to **modern**
an artificial neural
rate.

But rate coding is **inefficient** and **slow**...
(i.e. each neuron needs to fire many spikes to get good precision)
both *in vivo* and *in silico*.



TUNING CURVE

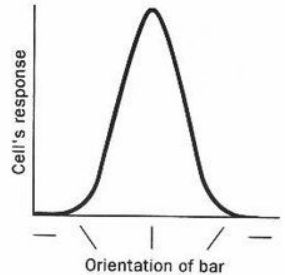
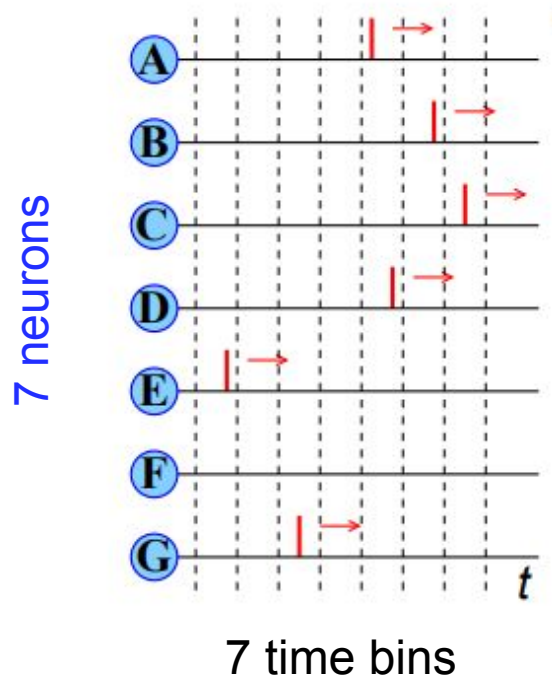
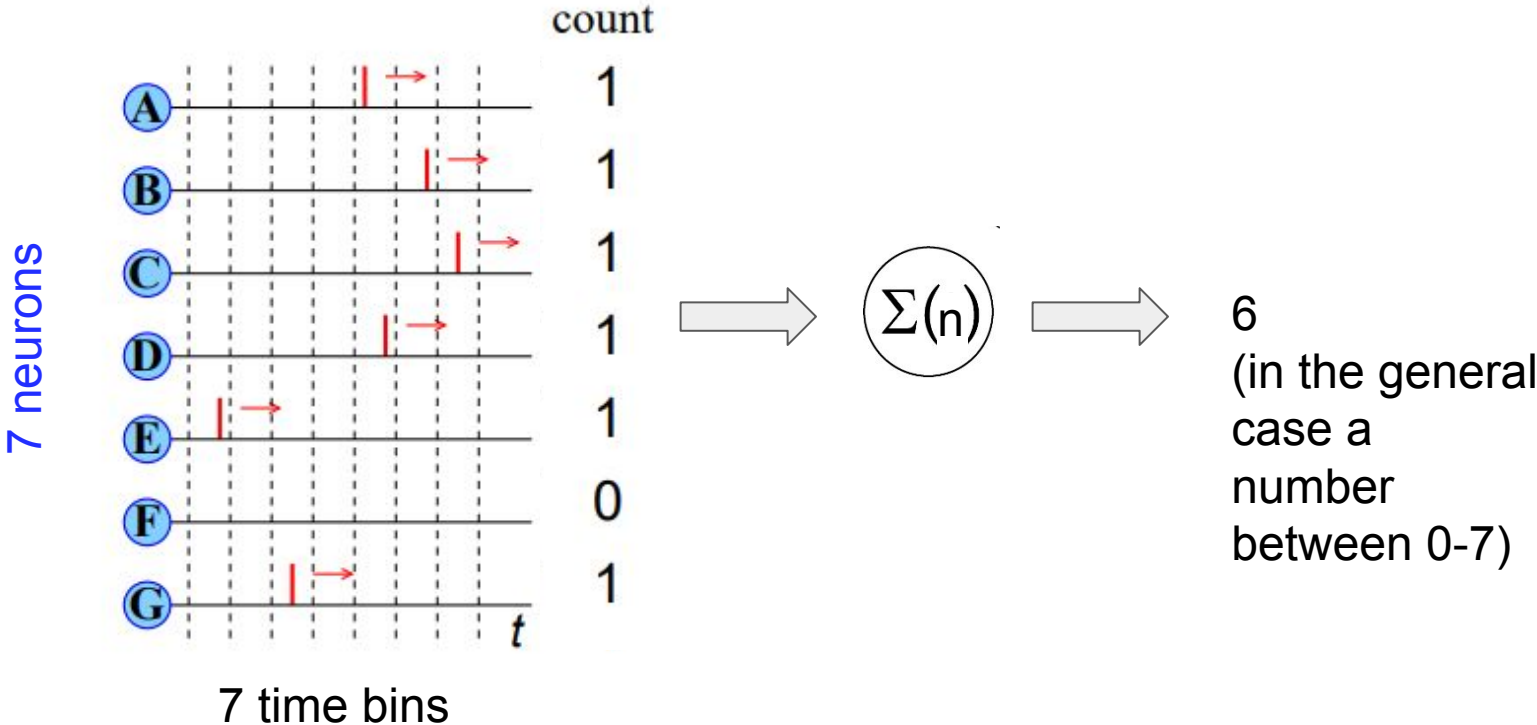


FIGURE 4.8 Response of a single cortical cell to bars presented at various orientations.

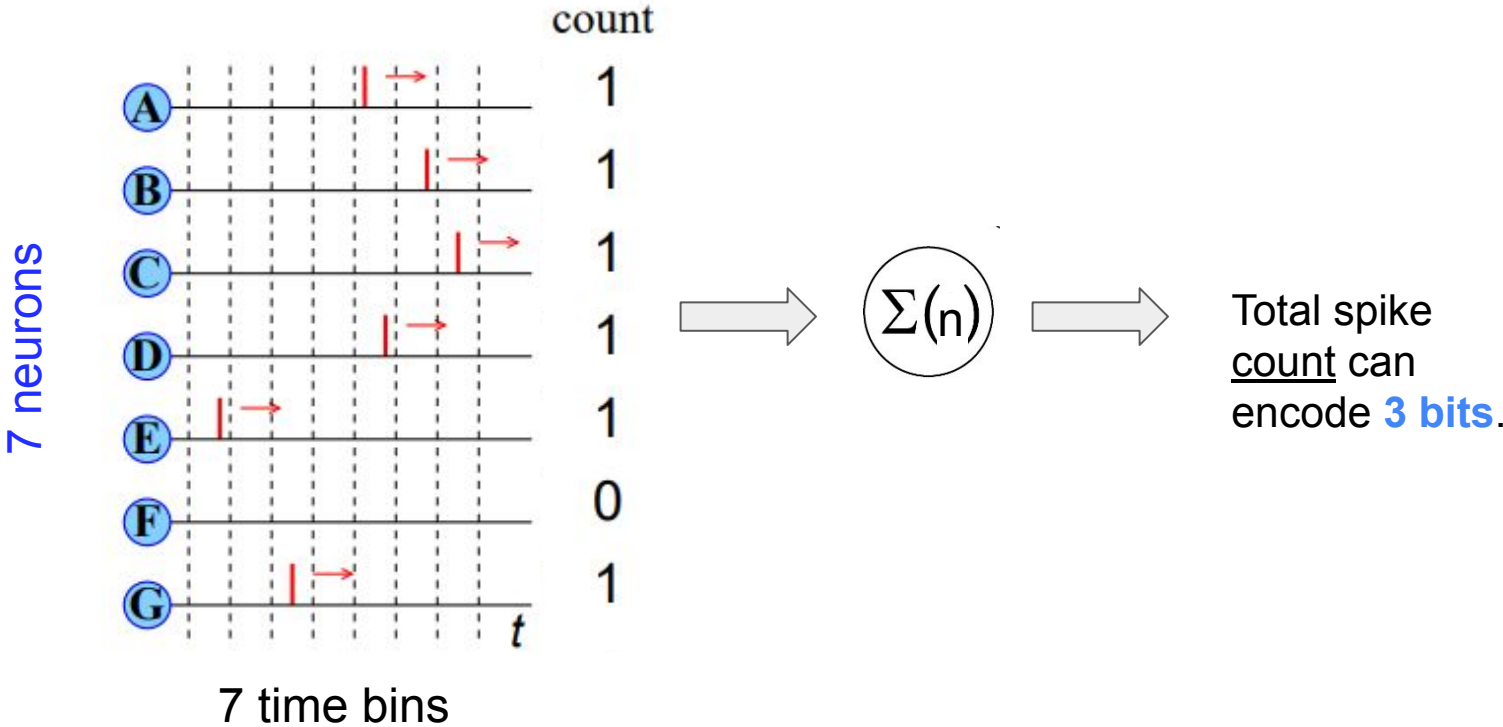
Different encoding strategies with spiking neurons [8]



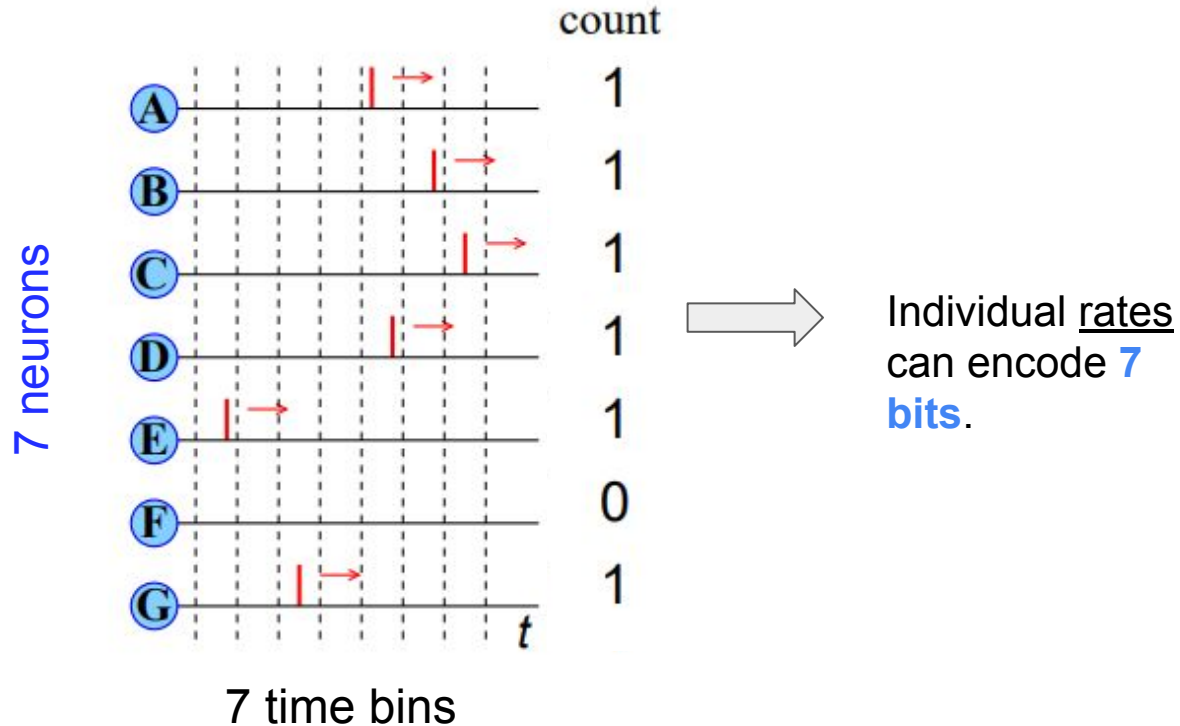
Different encoding strategies with spiking neurons



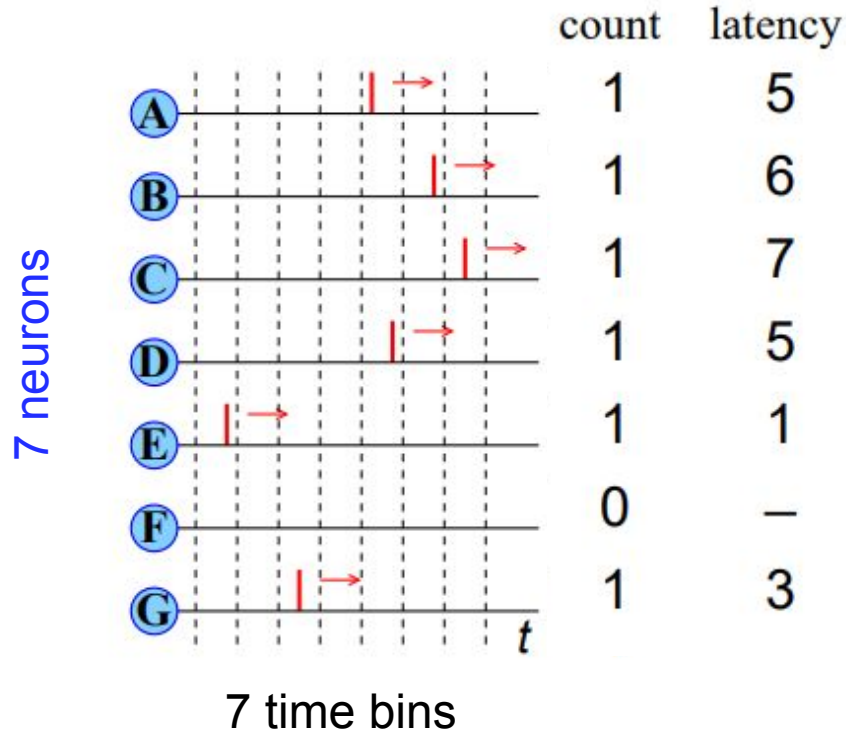
Different encoding strategies with spiking neurons



Different encoding strategies with spiking neurons

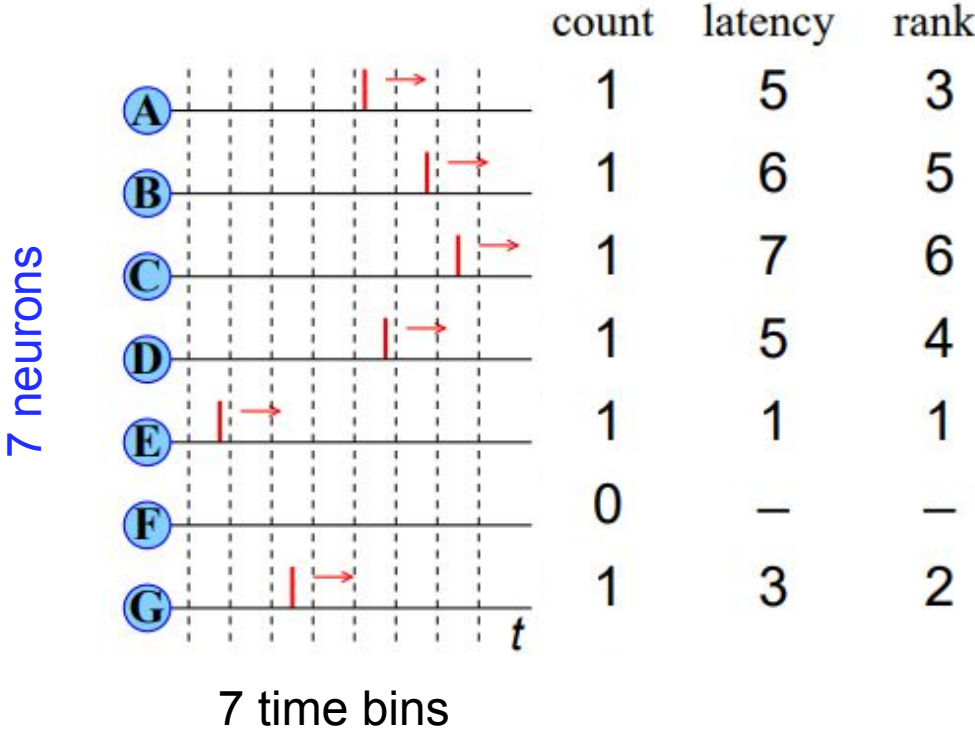


Different encoding strategies with spiking neurons



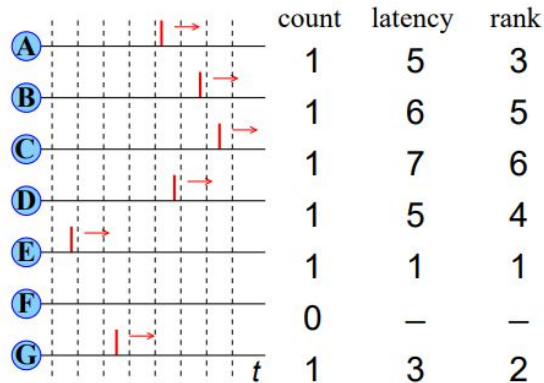
Latency can
encode $\sim 3 \times 7$ or
 ~ 19 bits.

Different encoding strategies with spiking neurons



Rank order can encode ~12 bits.

Different encoding strategies with spiking neurons

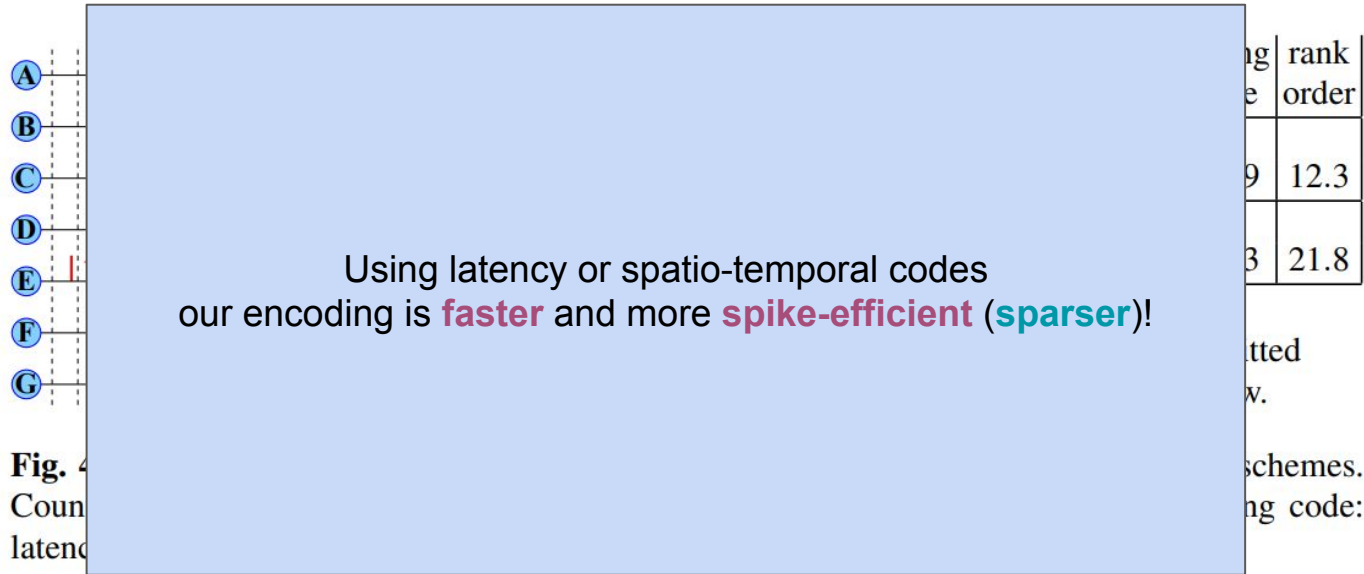


Numeric examples:	count code	binary code	timing code	rank order
left (opposite) figure $n = 7, T = 7ms$	3	7	≈ 19	12.3
Thorpe et al. [164] $n = 10, T = 10ms$	3.6	10	≈ 33	21.8

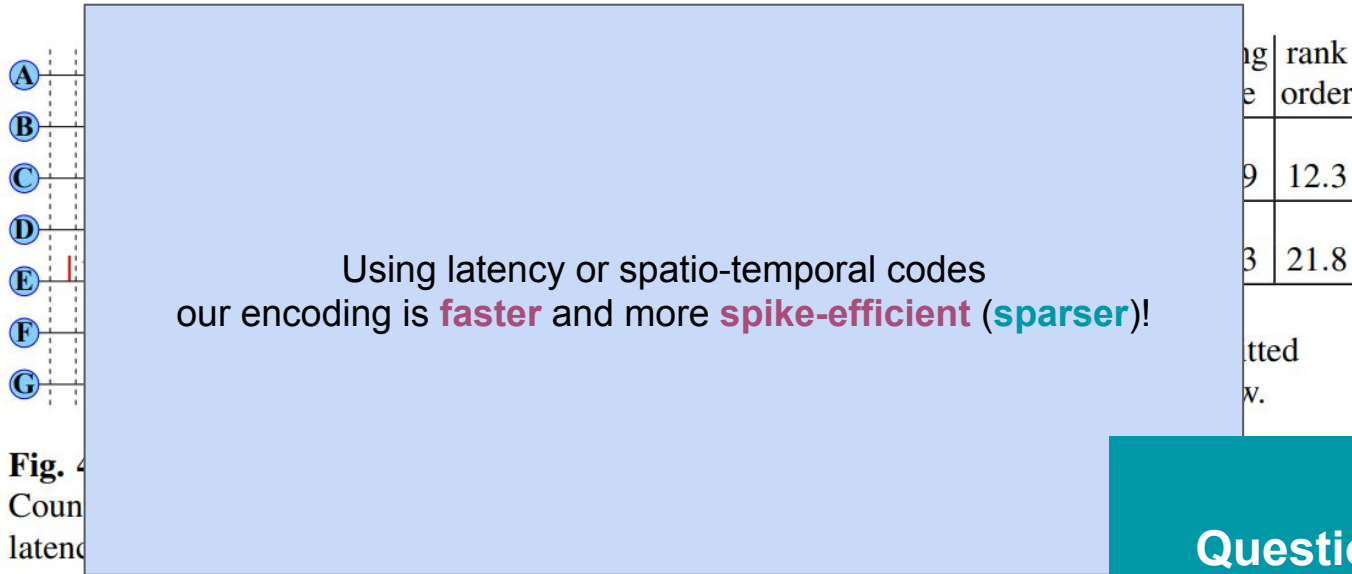
Number of bits that can be transmitted
by n neurons in a T time window.

Fig. 4 Comparing the representational power of spiking neurons, for different coding schemes. Count code: 6/7 spike per 7ms, i.e. $\approx 122 \text{ spikes.s}^{-1}$ - Binary code: 1111101 - Timing code: latency, here with a 1ms precision - Rank order code: $E \geq G \geq A \geq D \geq B \geq C \geq F$.

Different encoding strategies with spiking neurons



Different encoding strategies with spiking neurons



Questions?

Neuromorphic computing

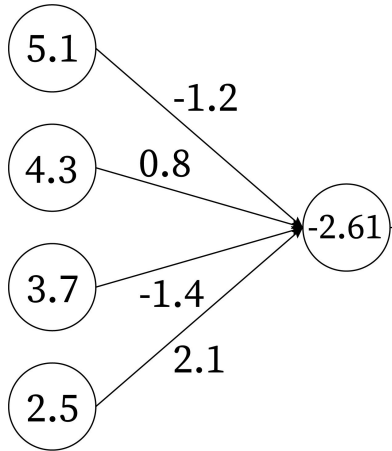
Brains are **energy efficient**:

2. **Sparse** encoding

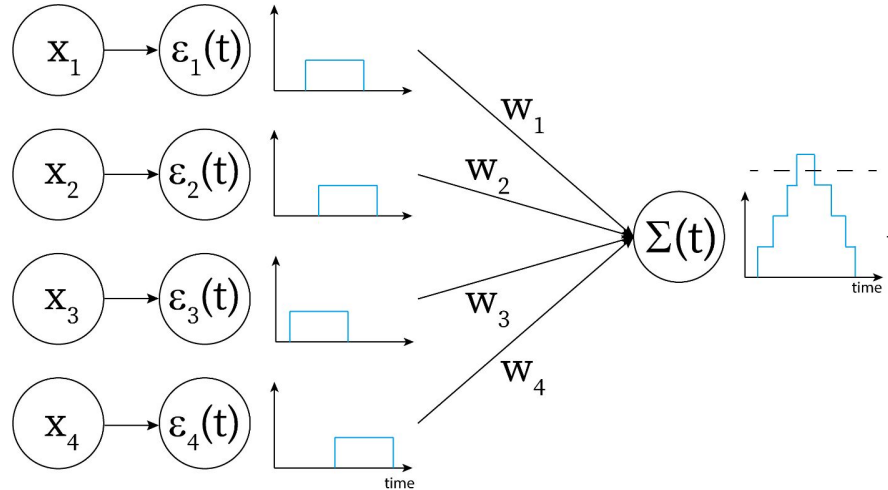
What is the advantage for applications?

- **Less spikes = less energy** consumption in specialized neuromorphic hardware
(e.g. Intel Loihi [12])

Multiply-accumulate (**MAC**) operations:



Normal neuron: Multiplies input with weights, then adds.

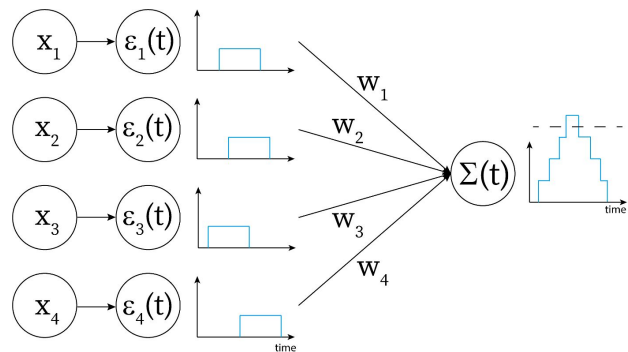
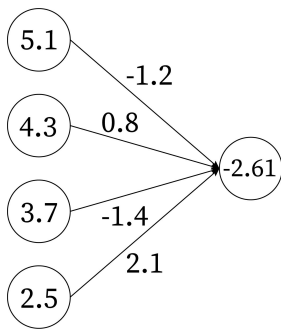


Spiking neuron: Consider binary input (e.g. input currents are piecewise constant and assume values $\{0,1\}$). There is **no multiplication**, only **addition**.

Multiply-accumulate (**MAC**) operations:

Assume one **multiplier** and one **adder** circuit uses **M** and **A** energy respectively with **A < M**

(e.g., for a 45nm CMOS process, standard energy usage is $A = 0.9$ pJ and $M = 3.7$ pJ).



Normal neuron: $n_{in} \times n_{out}$ multiplications,
 $(n_{in} - 1) \times n_{out}$ additions

Spiking neuron: 0 multiplications, $(n_{active} - 1) \times n_{out}$
additions, with $n_{active} \leq n_{in}$

Energy consumption:

$$E_{normal} = M n_{in} n_{out} + A (n_{in} - 1) n_{out} = 17.5 \text{ pJ}$$

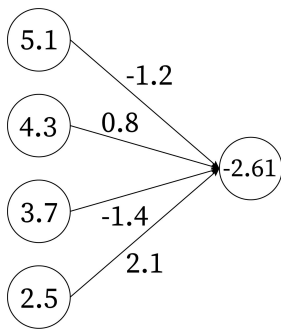
Energy consumption:

$$E_{spiking} = A (n_{active} - 1) n_{out} = 2.7 \text{ pJ}$$

Multiply-accumulate (**MAC**) operations:

Assume one **multiplier** and one **adder** circuit uses **M** and **A** energy respectively with **A < M**

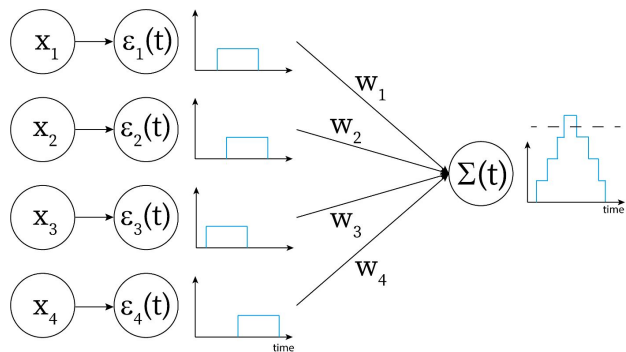
(e.g., for a 45nm CMOS process, standard energy usage is $A = 0.9$ pJ and $M = 3.7$ pJ).



Normal neuron: $n_{in} \times n_{out}$ multiplications,
 $(n_{in} - 1) \times n_{out}$ additions

Energy consumption:

$$E_{\text{normal}} = M n_{in} n_{out} + A (n_{in} - 1) n_{out} = 17.5 \text{ pJ}$$

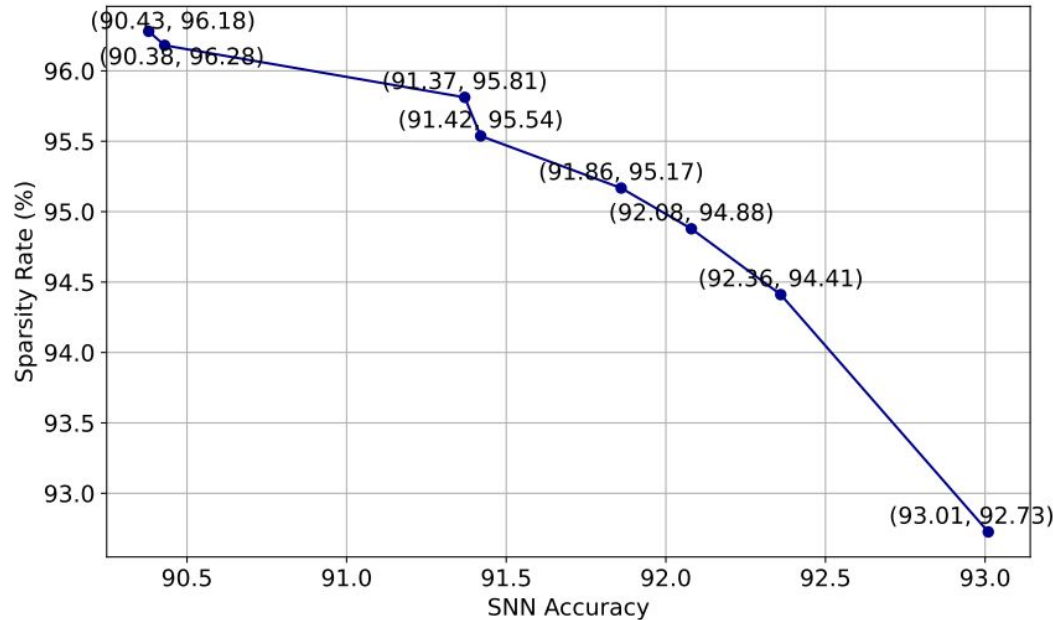


Spiking neuron: 0 multiplications, $(n_{\text{active}} - 1) \times n_{out}$ additions, with $n_{\text{active}} \leq n_{in}$

Energy consumption:

$$E_{\text{spiking}} = A (n_{\text{active}} - 1) n_{out} = 2.7 \text{ pJ}$$

SNN challenge: how to compute with the least amount of spikes!



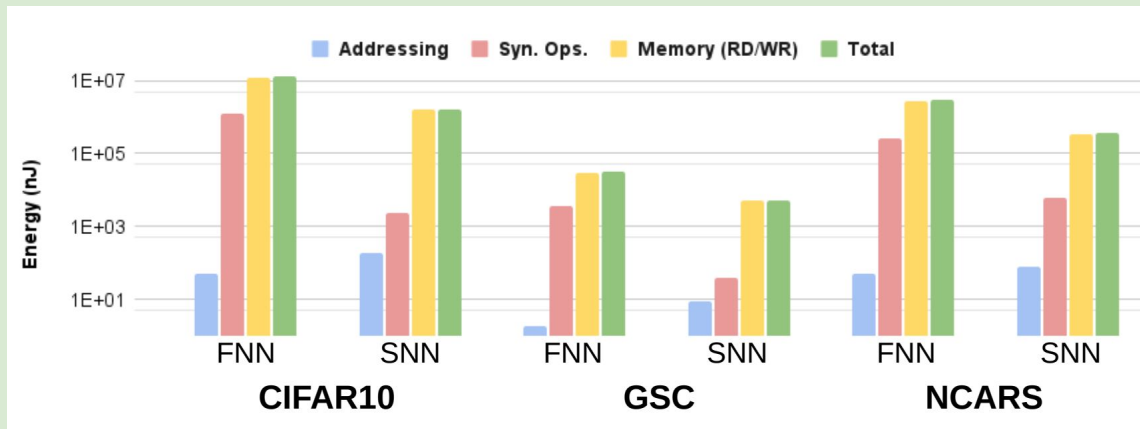
Often, we observe a **sparsity** (energy)-**task accuracy** trade-off

(Left: results for image classification)

Figure 8: SNN accuracy-sparsity trade-off using VGG16 on CIFAR-10 dataset; (x,y) indicates the SNN with accuracy of x% and sparsity rate of y%

SNN challenge: how to compute with the least amount of spikes!

Computing energy consumption



Estimated energy consumption for 3 different datasets (CIFAR10, GSC, NCARS; image, sound and video classification respectively). FNN's are conventional feed-forward neural networks.

In this example: SNNs are **6 to 8 times more energy** efficient than FNNs.

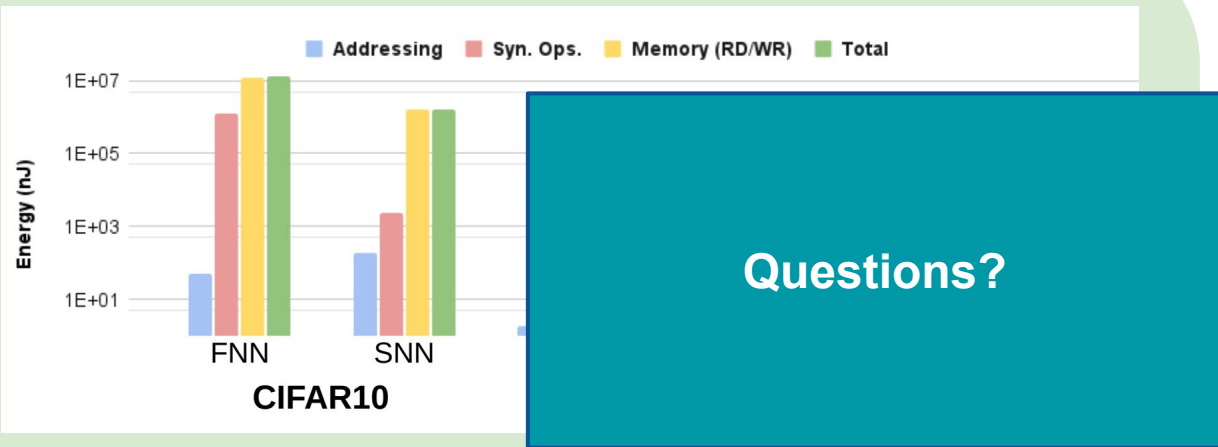
In practice, **energy consumption** computations are complex.

Need to take into account

- memory access,
- addressing,
- auxiliary operations,

in addition to MACs.

Computing energy consumption



Estimated energy consumption for 3 different datasets (CIFAR10, GSC, NCARS; image, sound and video classification respectively). FNN's are conventional feed-forward neural networks.

In this example: SNNs are **6 to 8 times more energy** efficient than FNNs.

In practice, **energy consumption** computations are complex.

Need to take into account

- memory access,
- addressing,
- auxiliary operations,

in addition to MACs.

Reading materials

Main reading:

- Section 1 and Section 3.1 of "Computing with spiking neuron networks." by Paugam-Moisy H, Bohte SM, in Handbook of natural computing (2012).

https://homepages.cwi.nl/~sbohte/publication/paugam_moisy_bohte_SNNChapter.pdf

- Maass W. Networks of spiking neurons: the third generation of neural network models. Neural networks. 1997. 10(9):1659-71. <https://igi-web.tugraz.at/people/maass/psfiles/85a.pdf>

- Neuromorphic computing:

- Based on biology: Zenke F, Bohtë SM, Clopath C, Comşa IM, Göltz J, Maass W, Masquelier T, Naud R, Neftci EO, Petrovici MA, Scherr F. Visualizing a joint future of neuroscience and neuromorphic engineering. Neuron. 2021. 109(4):571-5. <https://www.sciencedirect.com/science/article/pii/S089662732100009X>

- How to train modern spiking networks: Neftci EO, Mostafa H, Zenke F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. IEEE Signal Processing Magazine. 2019. 36(6):51-63. <https://ieeexplore.ieee.org/abstract/document/8891809>

- Rate-based SNNs: Roy K, Jaiswal A, Panda P. Towards spike-based machine intelligence with neuromorphic computing. Nature. 2019. 575(7784):607-17. <https://www.nature.com/articles/s41586-019-1677-2>

Extra reading:

- Converging history of deep networks and biological systems: Sejnowski TJ. The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences. 2020. 117(48):30033-8. <https://www.pnas.org/doi/full/10.1073/pnas.1907373117>

- Also an important part of neuromorphic systems and vision → Event Cameras: Gallego G, Delbrück T, Orchard G, Bartolozzi C, Taba B, Censi A, Leutenegger S, Davison AJ, Conradt J, Daniilidis K, Scaramuzza D. Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence. 2020. 44(1):154-80.

<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9138762>

Basics of 'conventional' neural networks:

- Sections 4.1 to 4.4 from the book "Pattern Recognition" by Theodoridis and Koutroumbas.

- Subsection 4.1.7 from the book "Pattern Recognition and Machine Learning" by Bishop.

References

- 1) McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics. 1943. 5(4):115-33. <https://link.springer.com/content/pdf/10.1007/BF02478259.pdf>
- 2) Pitts W, McCulloch WS. How we know universals the perception of auditory and visual forms. The Bulletin of Mathematical Biophysics. 1947. 9(3):127-47. <https://link.springer.com/content/pdf/10.1007/BF02478291.pdf>
- 3) Abraham TH. (Physio) logical circuits: The intellectual origins of the McCulloch–Pitts neural networks. Journal of the History of the Behavioral Sciences. 2002. 38(1):3-25. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jhbs.1094>
- 4) Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958. 65(6):386. (not publicly available) doi:10.1037/h0042519
- 5) Tappert CC. Who is the father of deep learning? International Conference on Computational Science and Computational Intelligence (CSCI) 2019. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9070967>
- 6) <https://github.com/idsc-frazzoli/retina>
- 7) Maass W. Networks of spiking neurons: the third generation of neural network models. Neural networks. 1997. 10(9):1659-71. <https://igi-web.tugraz.at/people/maass/psfiles/85a.pdf>
- 8) Paugam-Moisy H, Bohte SM, "Computing with spiking neuron networks." in Handbook of Natural Computing (2012). https://homepages.cwi.nl/~sbohte/publication/paugam_moisy_bohte_SNNChapter.pdf
- 9) Kheradpisheh SR, Ganjtabesh M, Thorpe SJ, Masquelier T. STDP-based spiking deep convolutional neural networks for object recognition. Neural Networks. 2018 Mar 1;99:56-67. <https://www.sciencedirect.com.tudelft.idm.oclc.org/science/article/pii/S0893608017302903>
- 10) Gütig R, Sompolinsky H. The tempotron: a neuron that learns spike timing–based decisions. Nature neuroscience. 2006 Mar;9(3):420-8. http://mcn2016public.pbworks.com/w/file/attach/137818197/Gutig_R_The%20tempotron_Nature%20Neuroscience.pdf
- 11) Neftci EO, Mostafa H, Zenke F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to SNNs. IEEE Signal Processing Magazine. 2019. 36(6):51-63. <https://ieeexplore.ieee.org/abstract/document/8891809>
- 12) <https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html>
- 13) <https://neurondynamics.epfl.ch/online/Ch1.S3.html>
- 14) Kron G. Numerical solution of ordinary and partial differential equations by means of equivalent circuits. Journal of Applied Physics. 1945. 16(3):172-86. <https://aip.scitation.org/doi/abs/10.1063/1.1707568>