Nano-Neurons for Artificial Intelligence

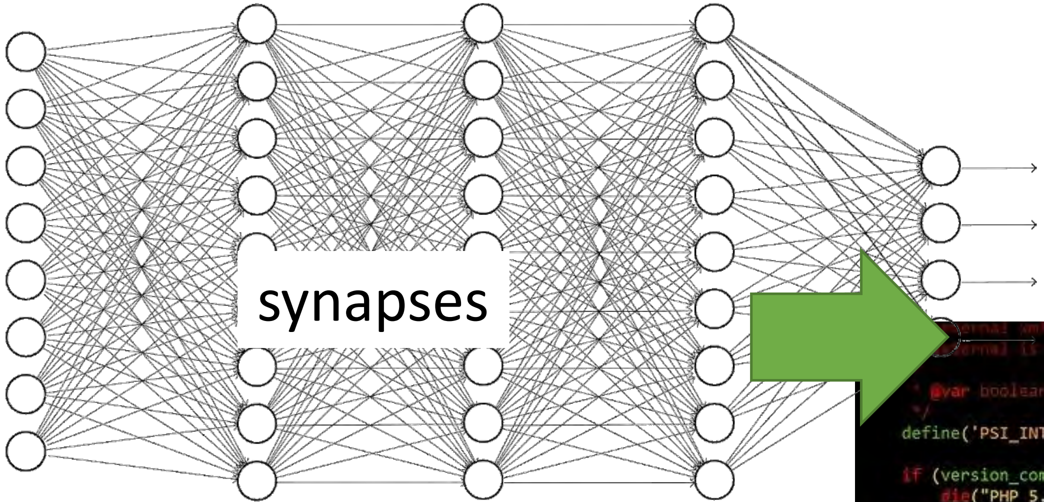
Julie Grollier[1]

Nathan Leroux[1], Danijela Marković[1], Jérémie Laydevant[1], Dedalo Sanz Hernandez[1], Philippe Talatchian[1], Miguel Romera[1], Mathieu Riou[1], Jacob Torrejon[1], Flavio Abreu Araujo[1], Paolo Bortolotti[1], Juan Trastoy[1], Erwann Martin[1], Teodora Petrisor[1], Vincent Cros[1], Guru Khalsa[2], Mark Stiles[2], Sumito Tsunegi[3], Kay Yakushiji[3], Akio Fukushima[3], Hitoshi Kubota[3], Shinji Yuasa[3], Ricardo Ferreira[4], Alex Jenkins[4], Leandro Martins[4], Tifenn Hirtzlin[5], Maxence Ernoult[5], Alice Mizrahi[1], Damien Querlioz[5]

[1]CNRS/Thales, France *[2]NIST, USA* *[3]AIST, Japan* *[4]INL, Portugal* *[5]C2N, France*

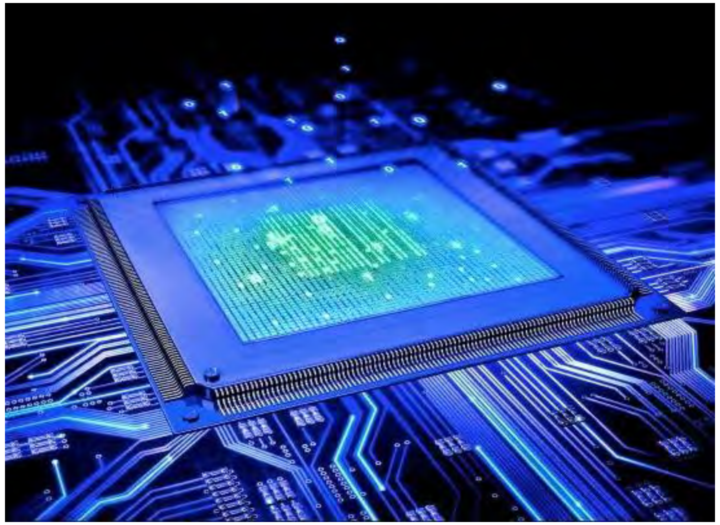
cnrs THALES université PARIS-SACLAY erc bioSPINspired Quantum Materials

Deep Neural networks run on unoptimized hardware

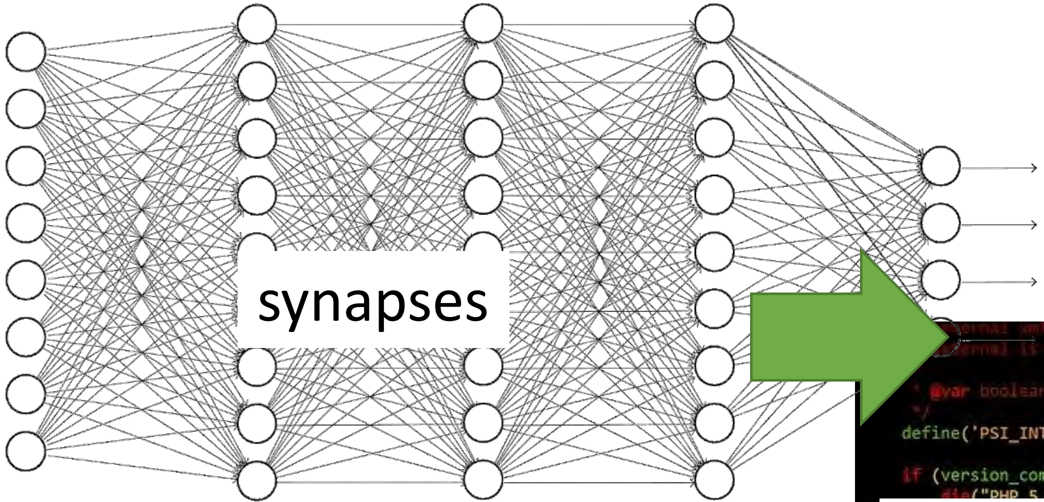


0011011

GPUs, TPUs, FPGAs



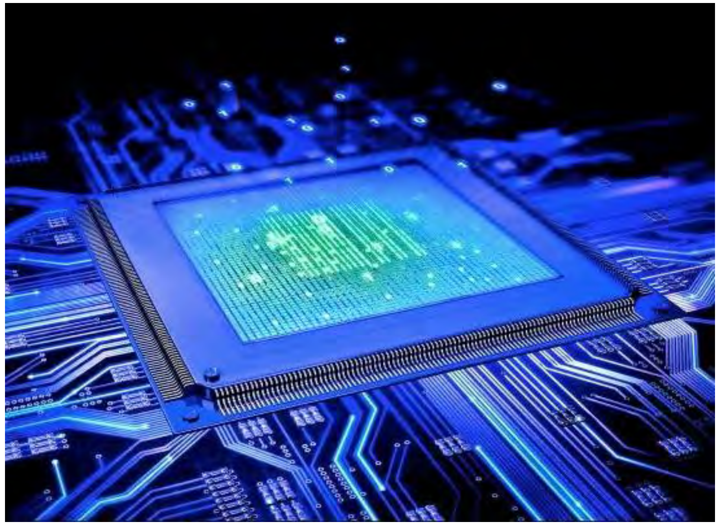
Deep Neural networks run on unoptimized hardware



0011011

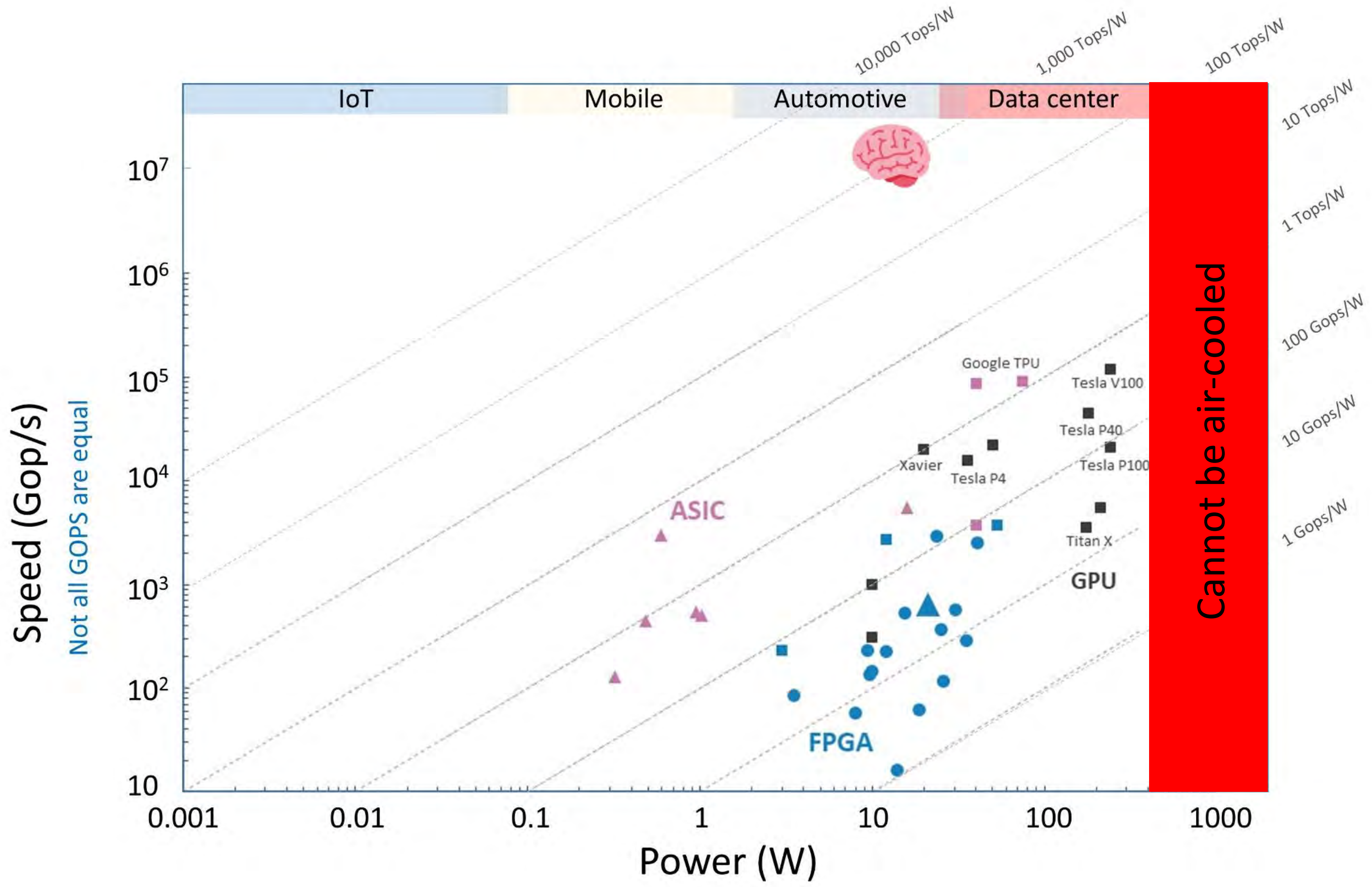
Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

GPUs, TPUs, FPGAs



E. Strubell et al,
https://arxiv.org/abs/1906.02243v1

Current CMOS processors cannot run future AI

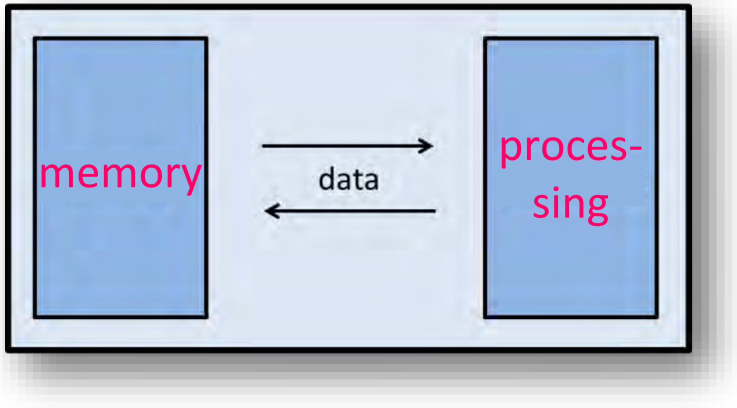


[Based on https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/]

Training neural networks on current computers is extremely power inefficient

Digital computer:

CPUs, GPUs, TPUs, FPGAs



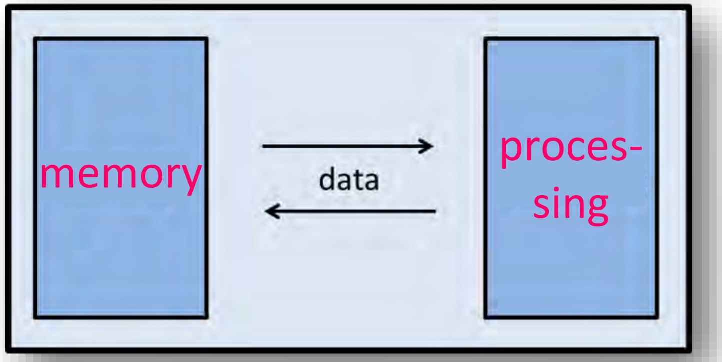
Operation	Energy consumption
Addition of data	1x
Access data (onchip cache)	60x
Access data (offchip RAM)	3500x

Pedram *et al*, *IEEE Xplore* (2017)

Training neural networks on current computers is extremely power inefficient

Digital computer:

CPUs, GPUs, TPUs, FPGAs

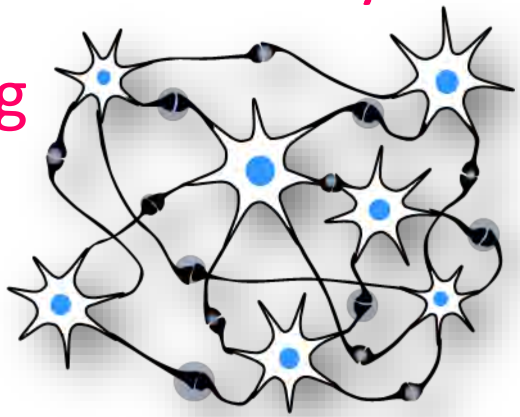


1000 kW.h to train a
Natural Language Processor

Brain : 20 W

synapse
memory

neuron
processing

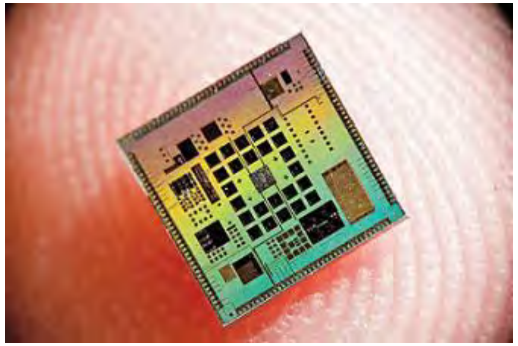


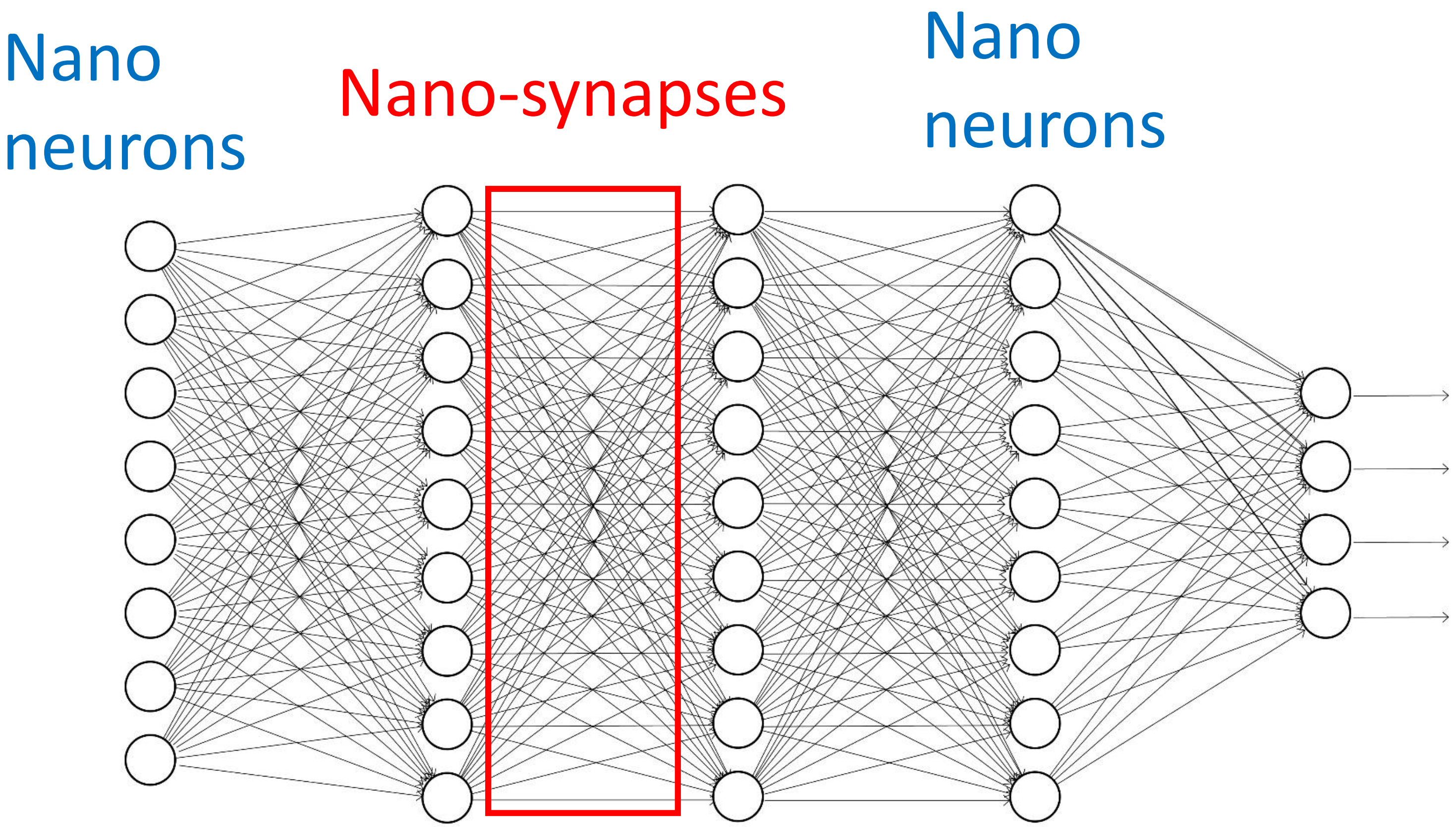
neuron
processing

6 years of brain operation

D. Marković et al, "Physics for neuromorphic computing", Nature Review Physics 2020

Orders of magnitude in energy can be saved by assembling physical synapses and neurons in neuromorphic chips





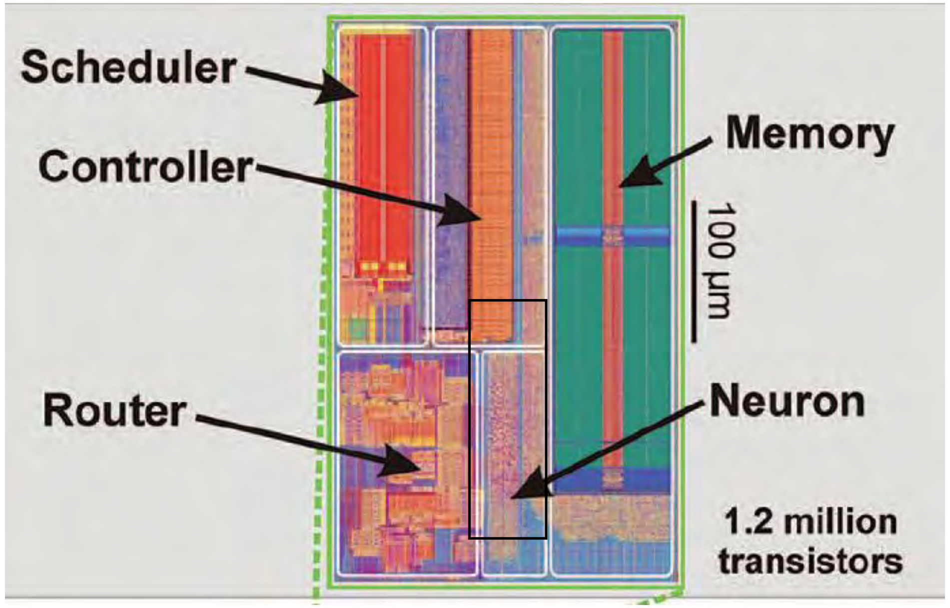
Hundred millions of neurons and synapses in a 1 cm^2 chip

→ Each device smaller than 1 μm^2

CMOS neurons and synapses are complex circuits

- A transistor is nanoscale but it is just a switch
- CMOS does not provide memory (volatile)

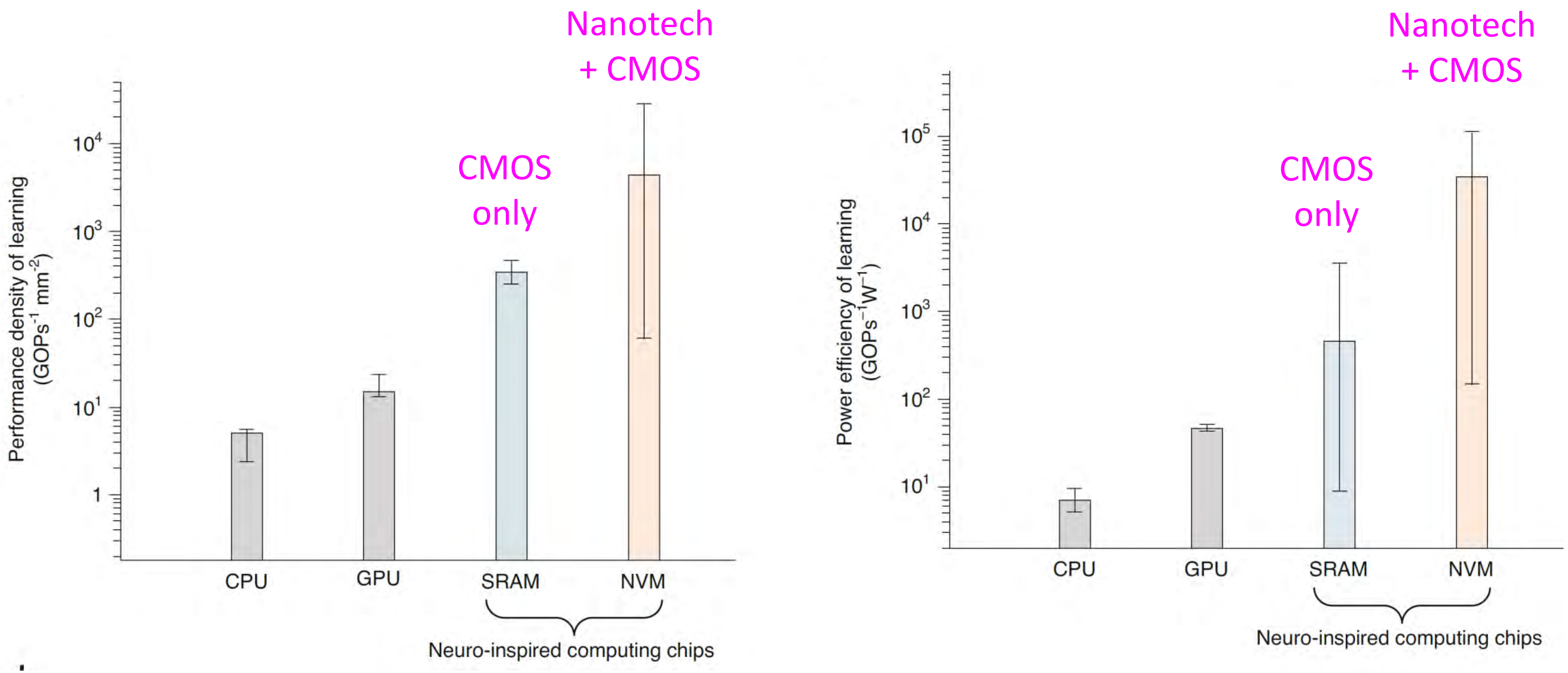
CMOS neuron **10-100 μm**
CMOS synapse **10 μm**



Merolla et al, *Science* **345**, 668 (2014)
Davies et al, *IEEE Micro*. **38**, 82–99 (2018)

Brainscales 20 wafer machine. 4M neurons, 1B synapses

Transistors alone won’t do the job: they should be complemented by emerging nanotechnologies



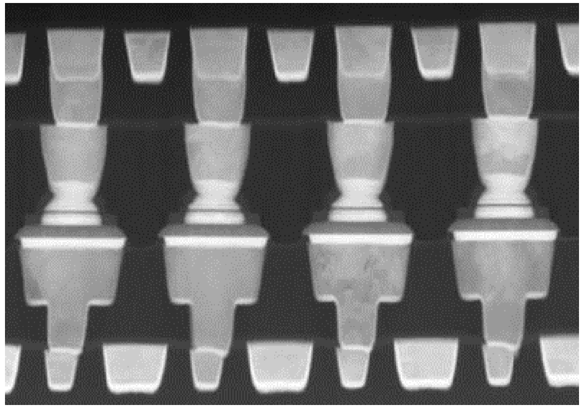
Zhang et al, Nature Electronics 3, 371 (2020)

The power of novel nanotechnologies for AI

Novel nanotechnologies are monothically integrated in major foundry process: they are commercially available and bring memory at the closest to compute

Spintronics

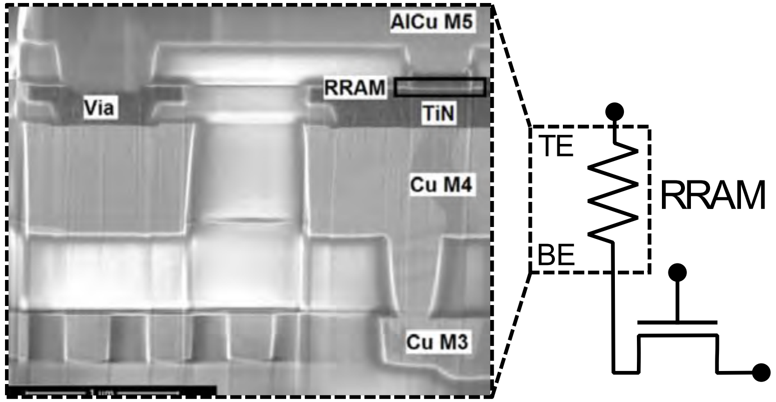
magnetic tunnel junctions



Intel: MRAM integrated into 22nm FinFET CMOS

Resistive-Switching

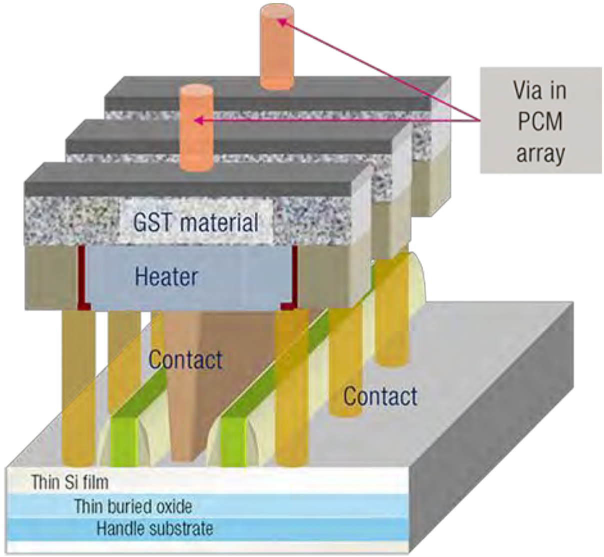
ReRAMs



CEA LETI: 130nm CMOS + HfO_2 RRAM

Bocquet, …, Vianello, Portal, Querlioz, IEEE IEDM, 2018

Phase Change



ST microelectronics

Memristors

They are multifunctional: they can emulate many features of neurons and synapses

Filamentary switching

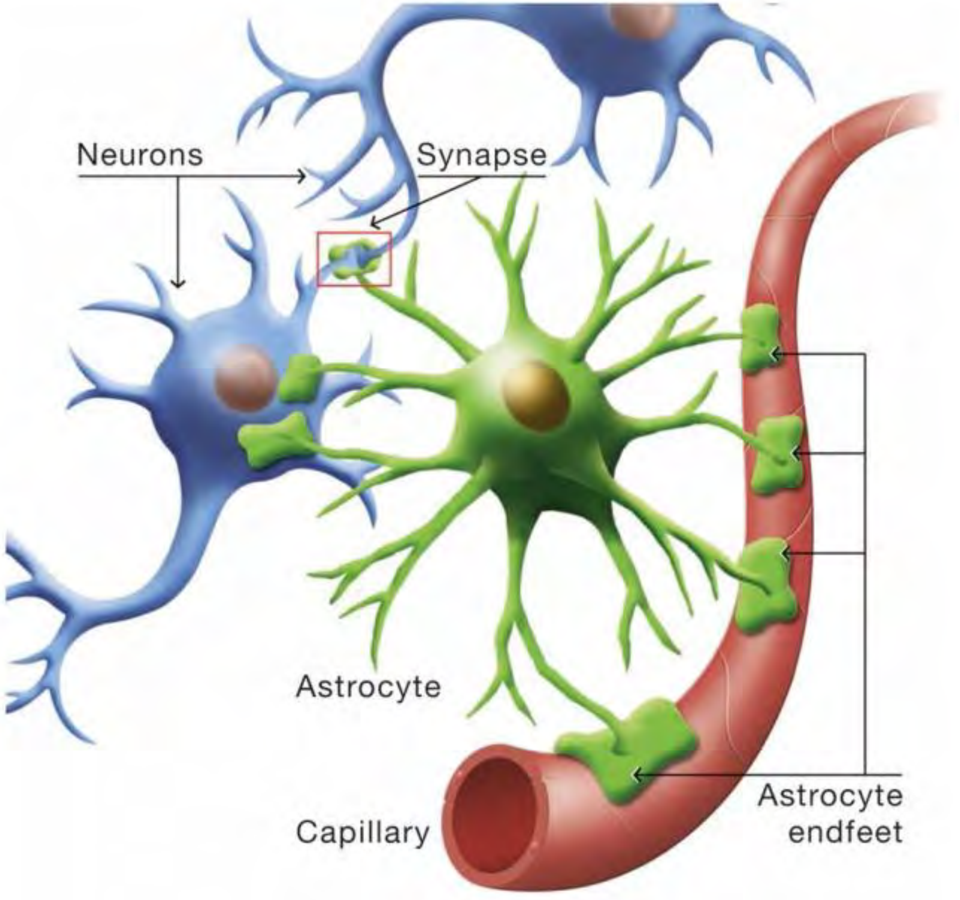
Phase-change

Optics

Ferroelectrics

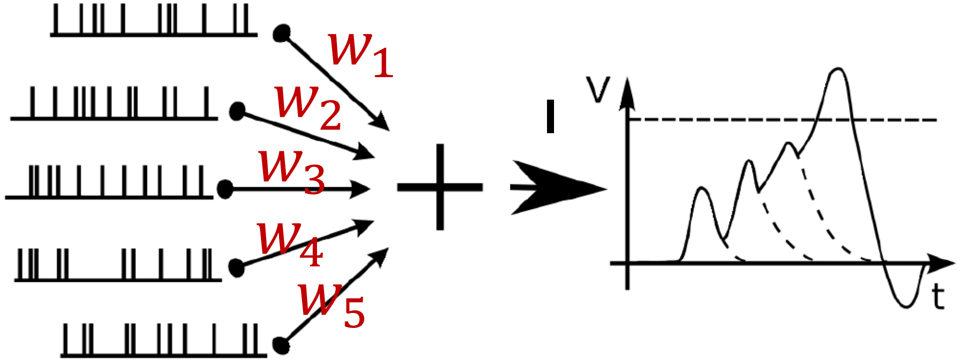
Organics

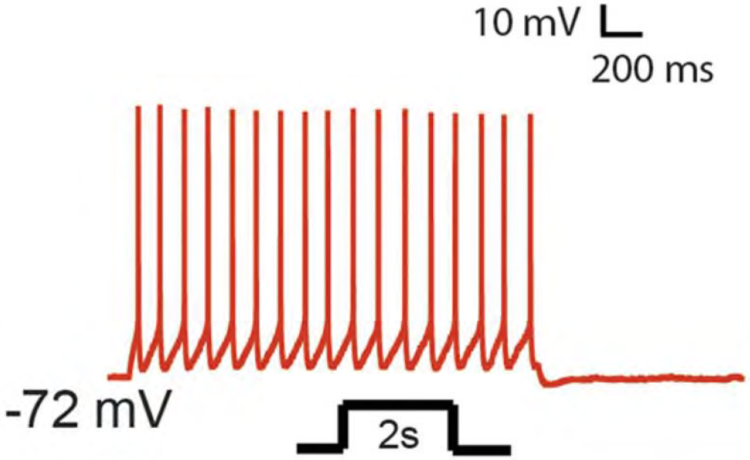
Spintronics



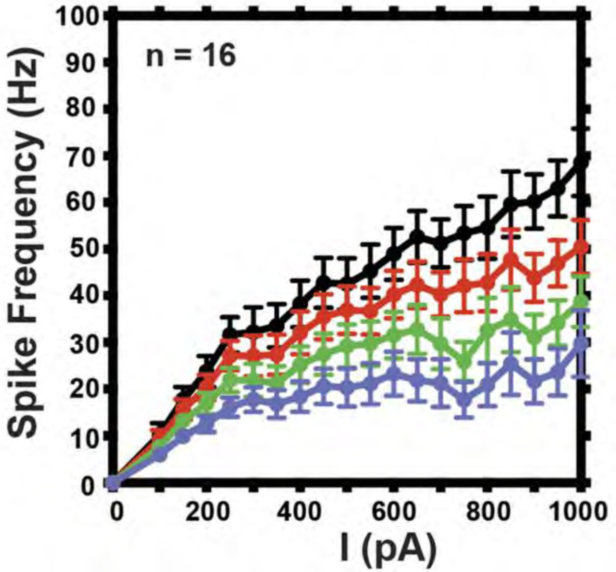
Neurons are non-linear and synapses are valves with memory

- **Brain**



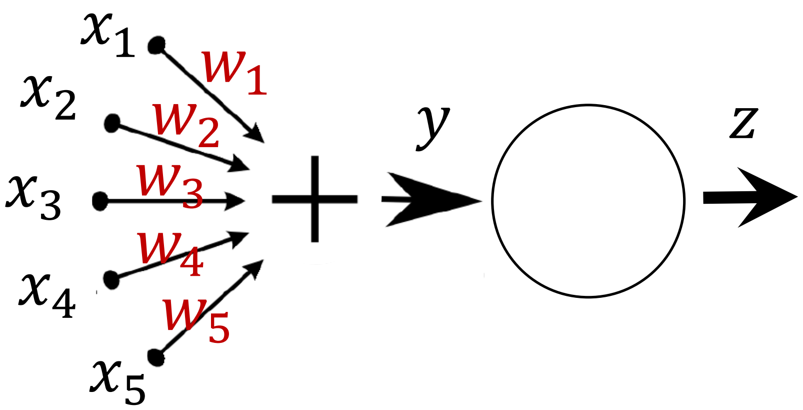


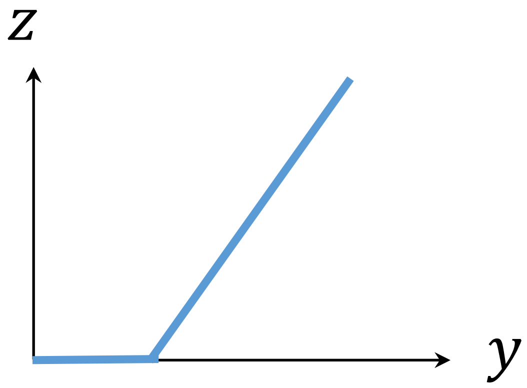
D. Guan et al, J Neurophysiol. 113, 2014 (2015)



- **Most neural networks today**

Real values

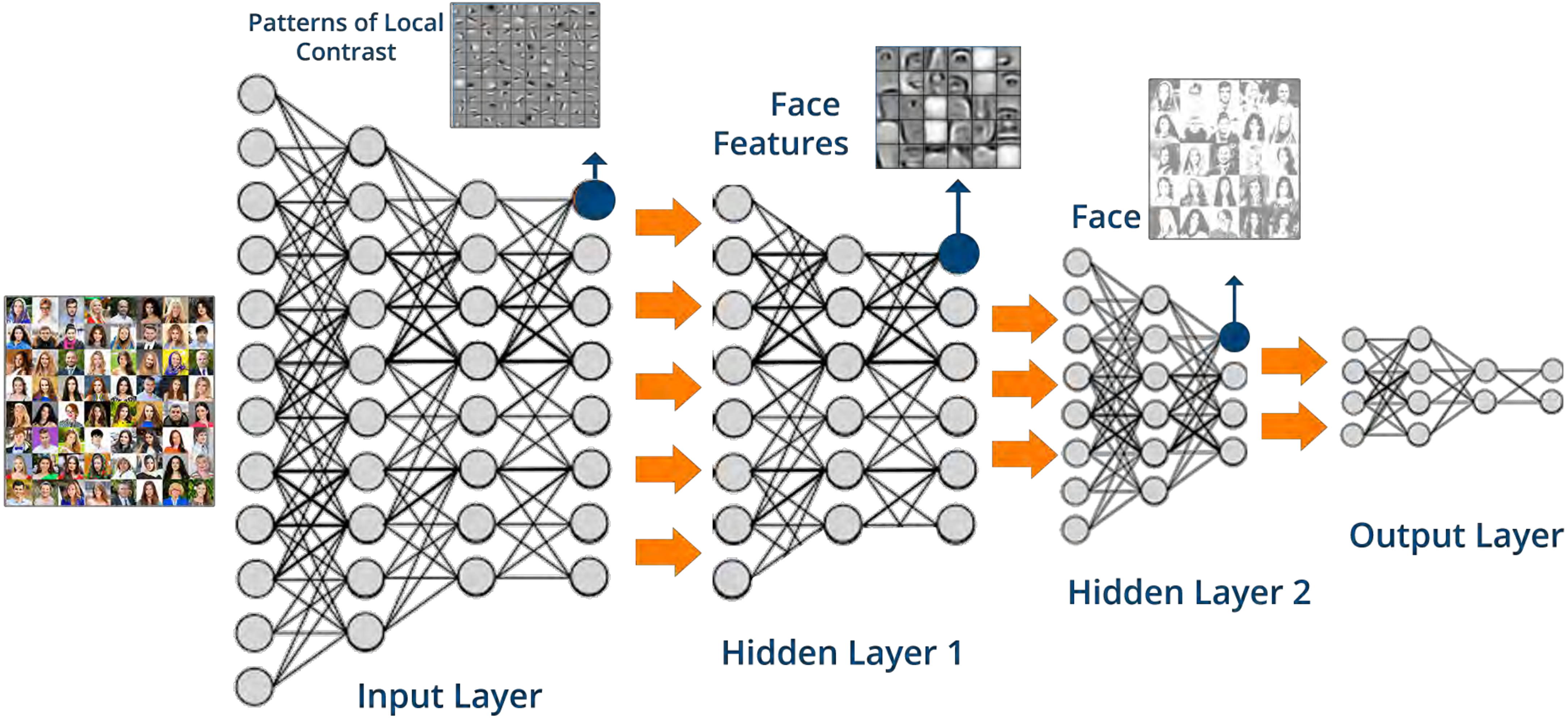




$$y = \sum w_i x_i$$

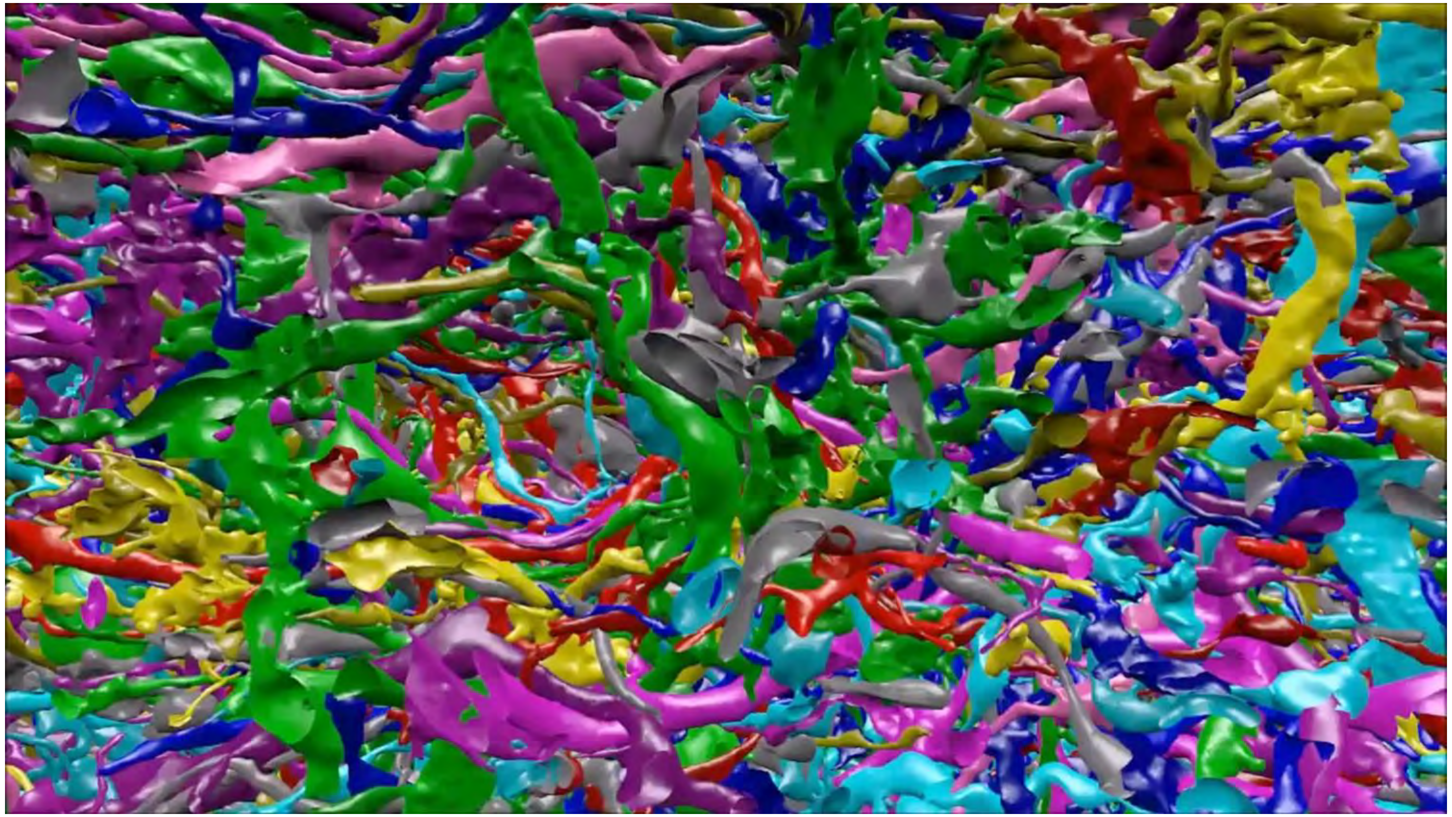
is called a Multiply and Accumulate (**MAC**) operation

State-of-the-art neural networks are deep: they extract features layer by layer



Synapses and neurons should be densely interconnected

Cortex: 10^4 synapses / neurones = 10^4 wires/neurons



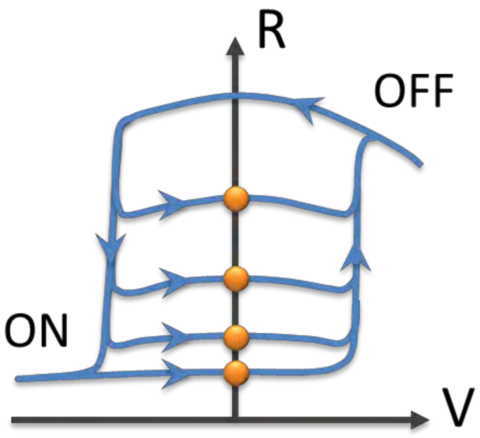
Moritz Helmstaedter lab, retina flight 2013

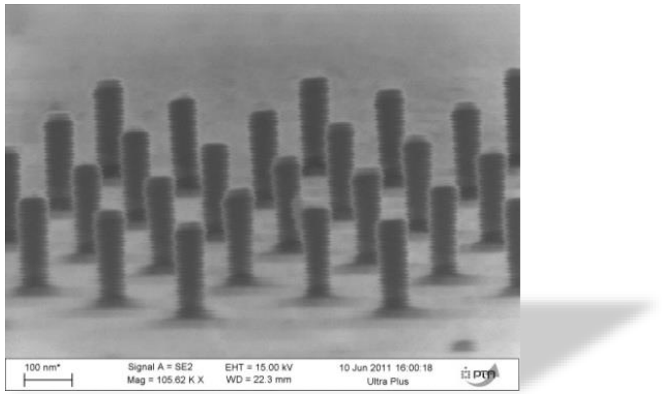
- Memristive neural nets

- Spintronics neural nets

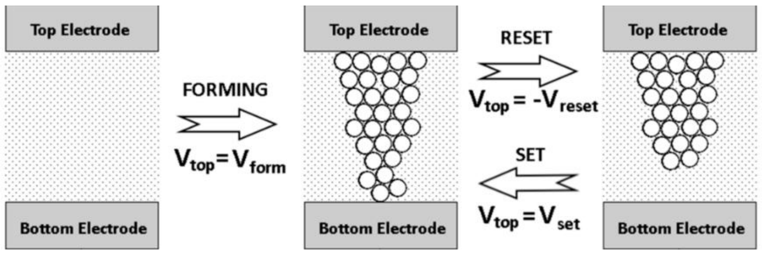
Non-volatile memristors emulate synapses

Chua, IEEE Trans.
Circuit Theory (1971)



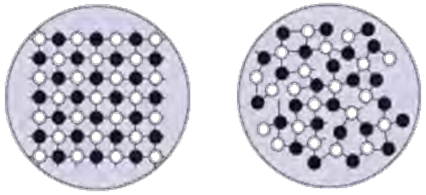


Filamentary switching



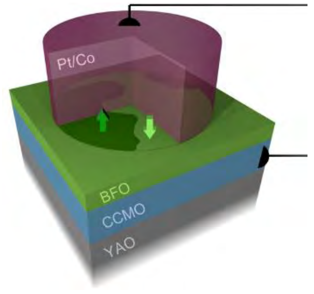
Yang et al.,
Nature Nano. (2013)

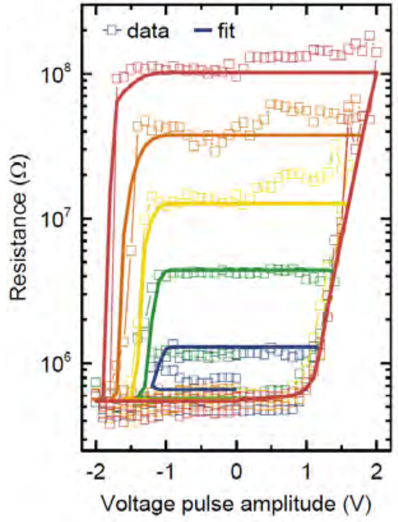
Phase change



Kuzum et al,
Nanotechnology (2013)

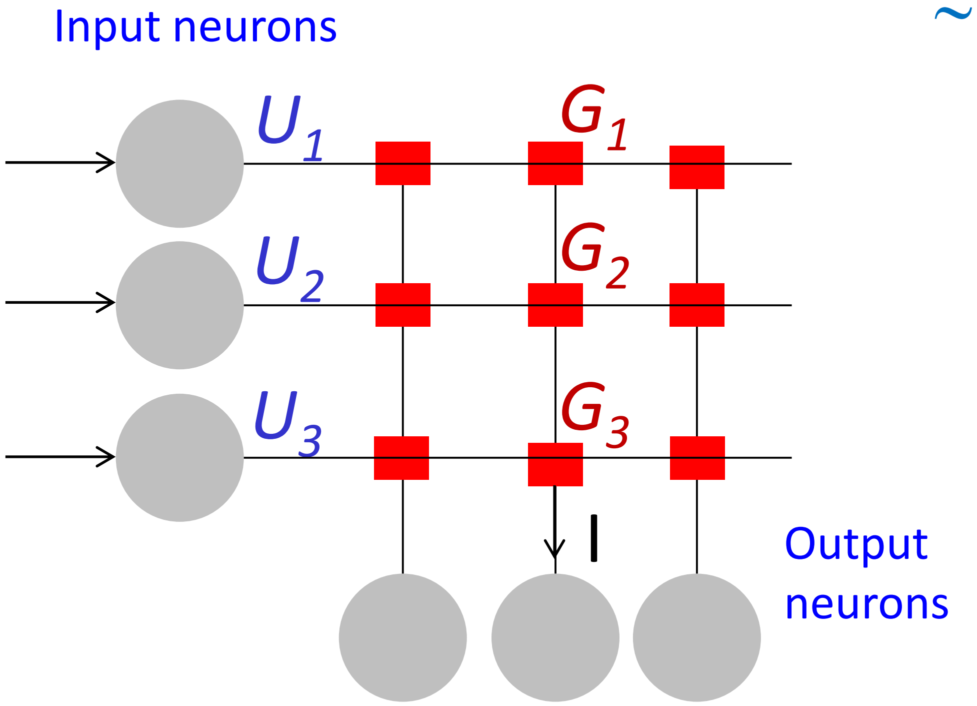
Ferroelectric





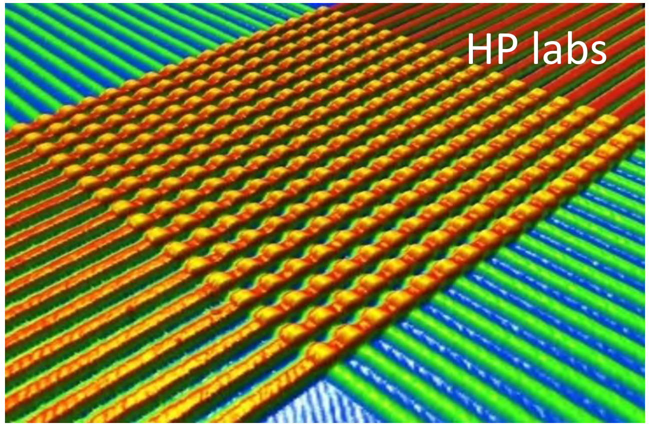
Chanthbouala et al,
Nature Mat. (2012)

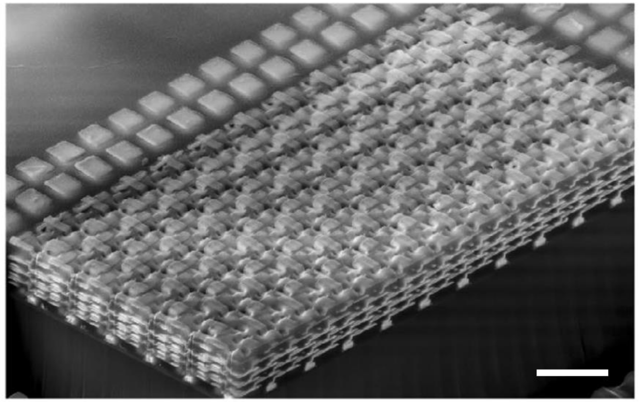
Going deep: crossbar arrays of memristors physically implement the multiply and accumulate operation



Current $I = \Sigma\, G_i\, U_i$

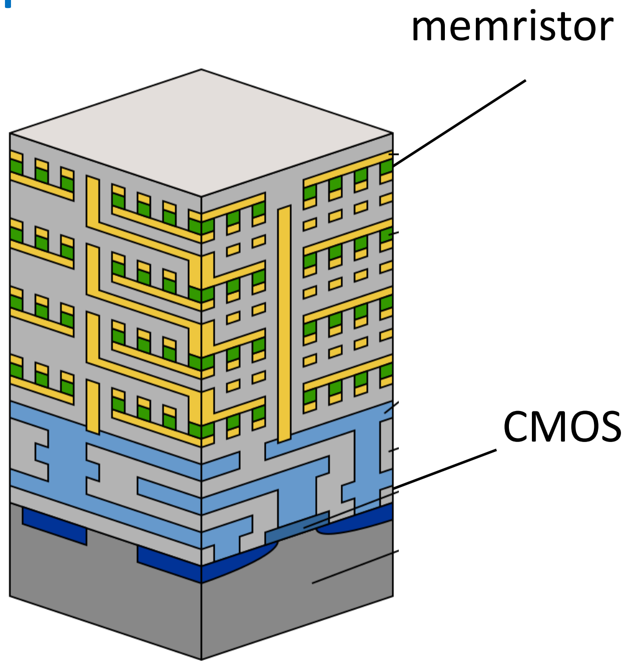
~ 100 synapses per neuron





Lin et al, Nature Electronics 3, 225 (2020)

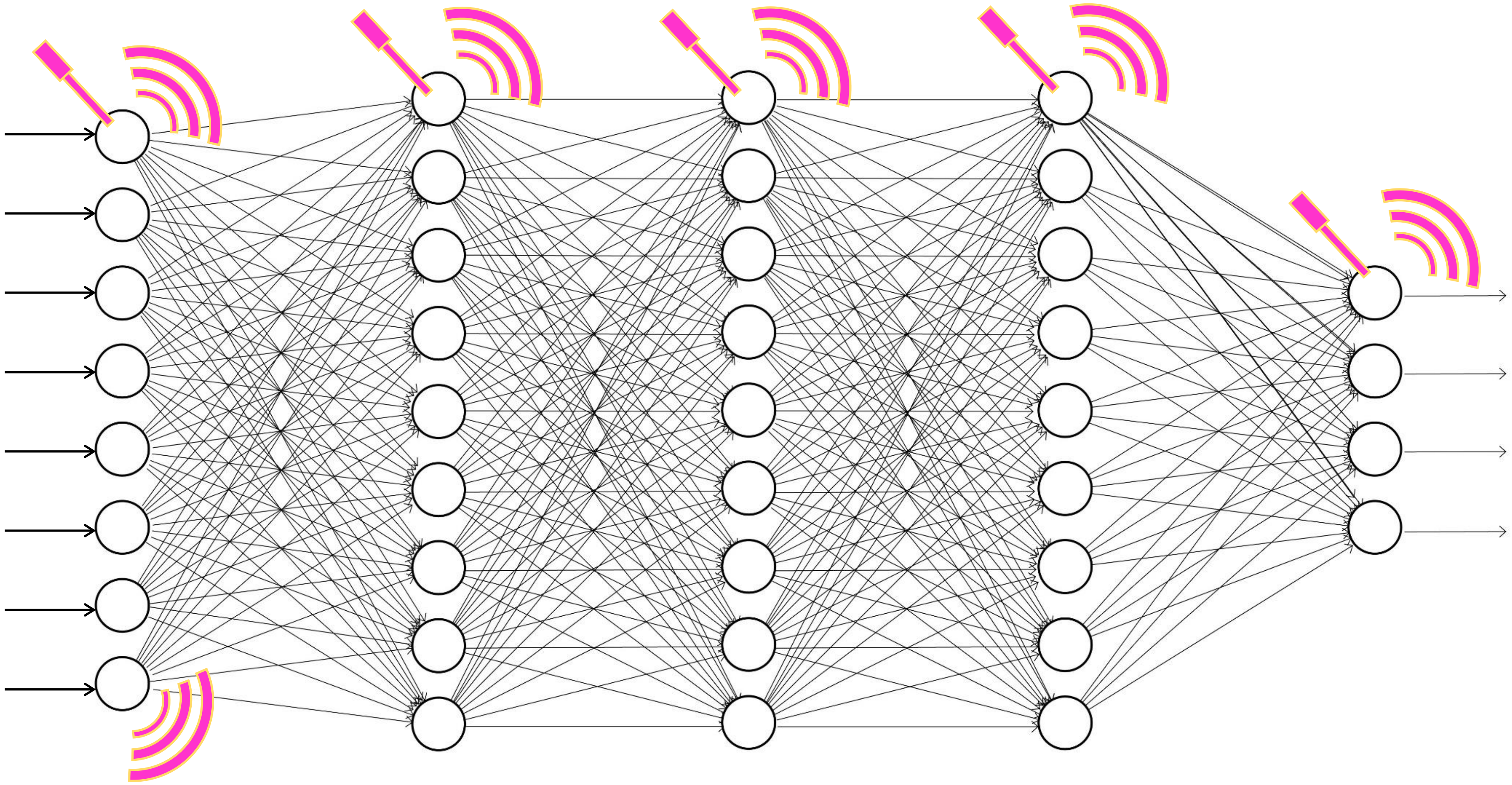
~10,000 synapses per neuron ?



Strukov and Williams, *PNAS* 106, 20155 (2009)

- Memristive neural nets
- Spintronics neural nets

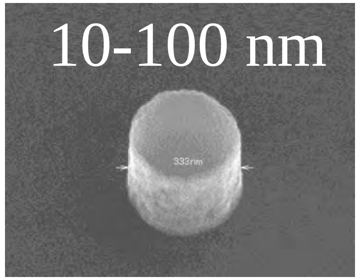
Deep learning through RF communications?



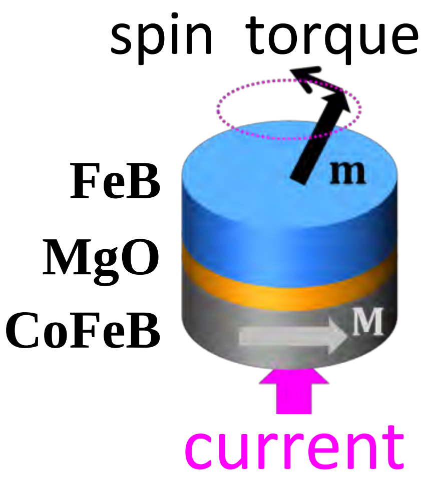
Magnetic tunnel junctions can be used as radio-frequency neurons

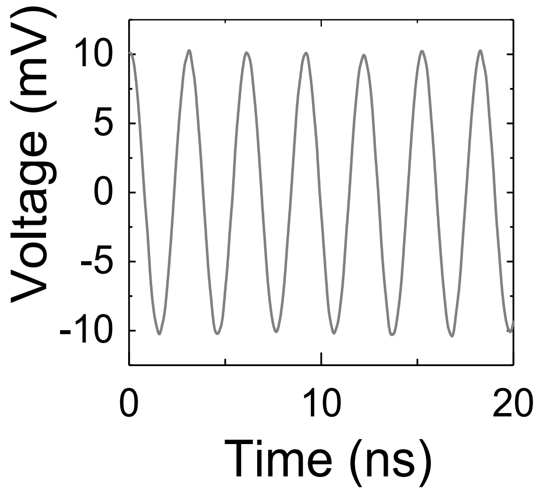
Nanoscale, fast (GHz), non-linear and easily measurable

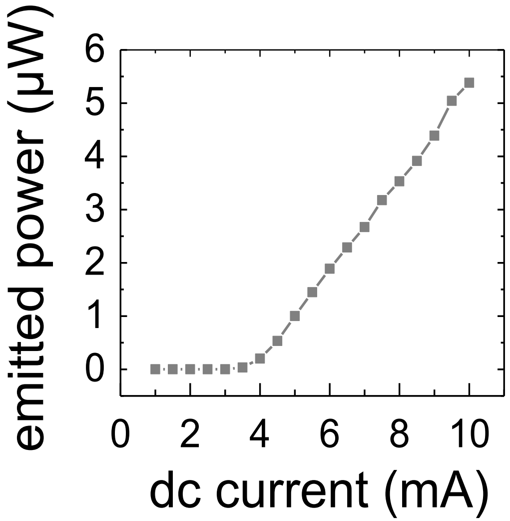
magnetic tunnel junction



compatible with CMOS







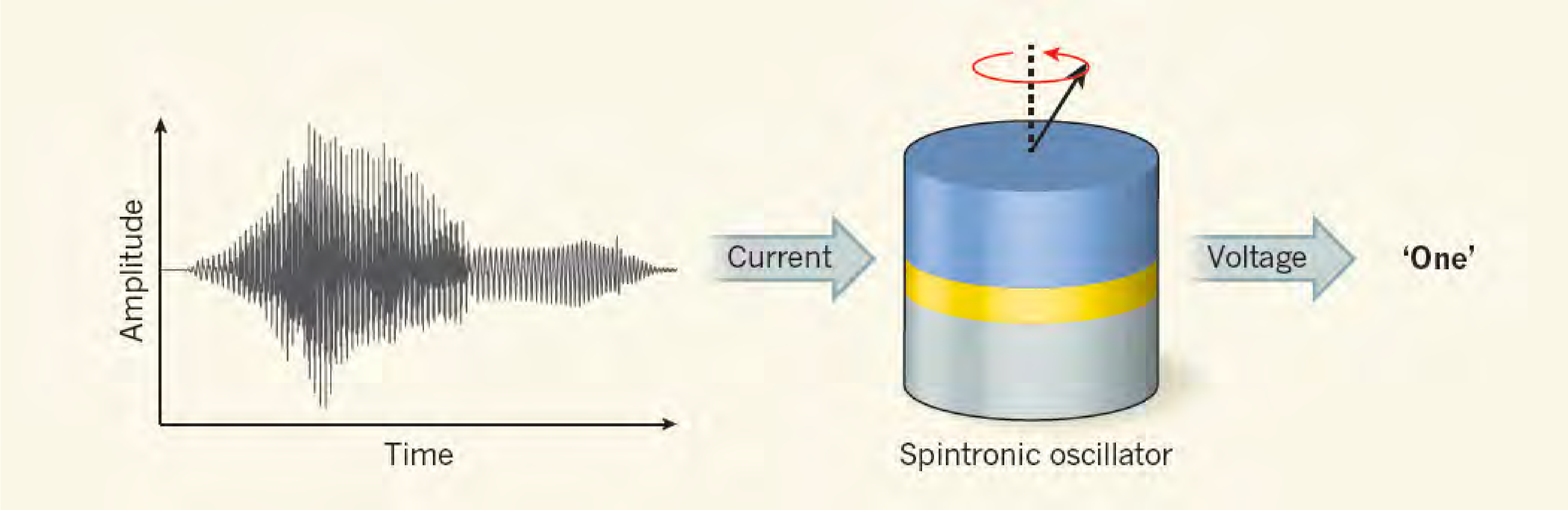
Same structure as magnetic memories

J. Grollier et al, "Neuromorphic Spintronics", Nature Electronics (2020)

Step 1: Single junction

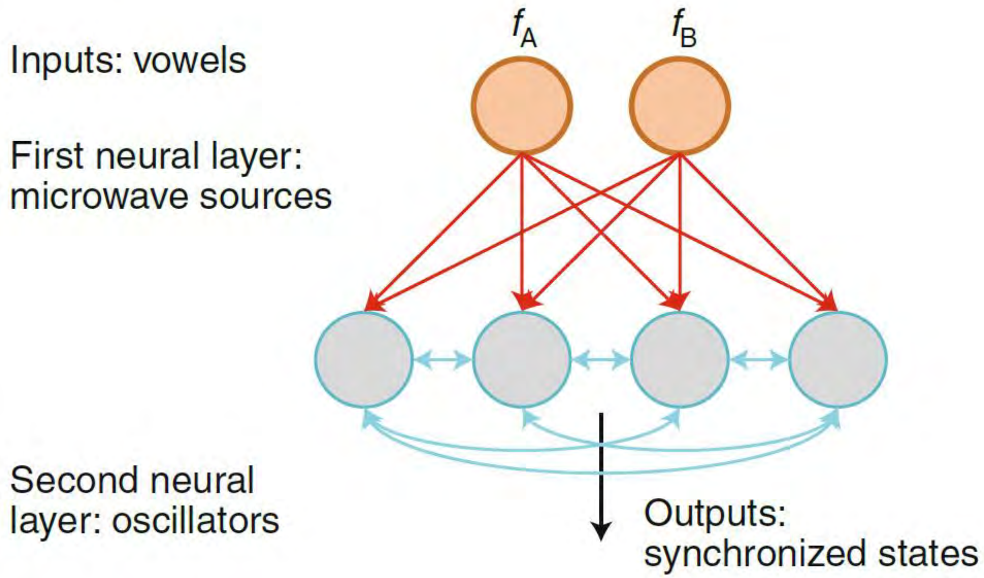
Due to its rich dynamics the nano-oscillator recognizes spoken digits with a success rate > 99.6%

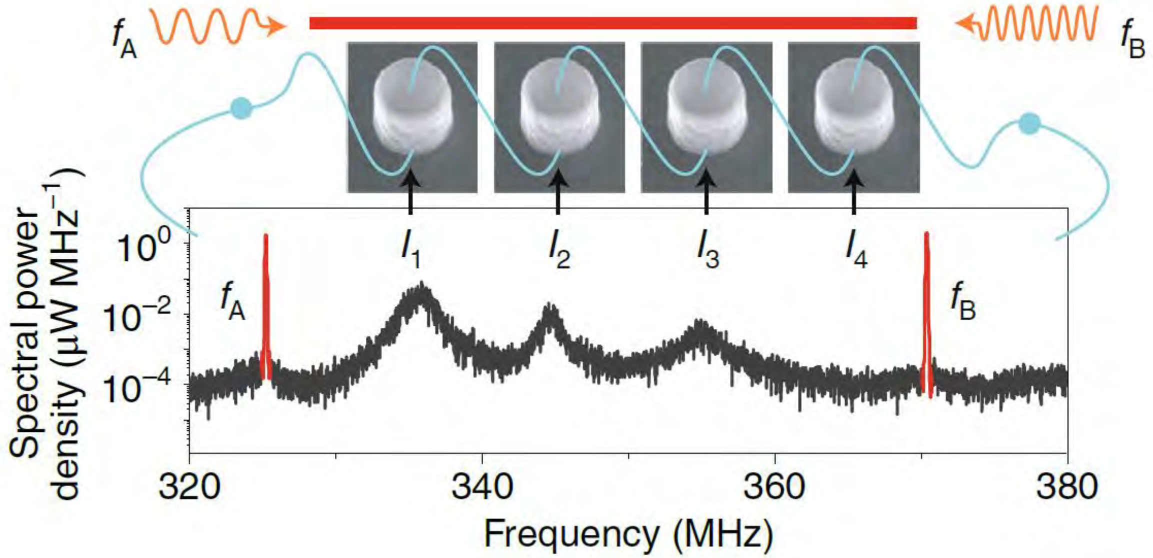
TI-46 database, 5 female speakers, cochlear pre-processing



J. Torrejon, M. Riou, F. Abreu Araujo et al, Nature 547, 428 (2017)

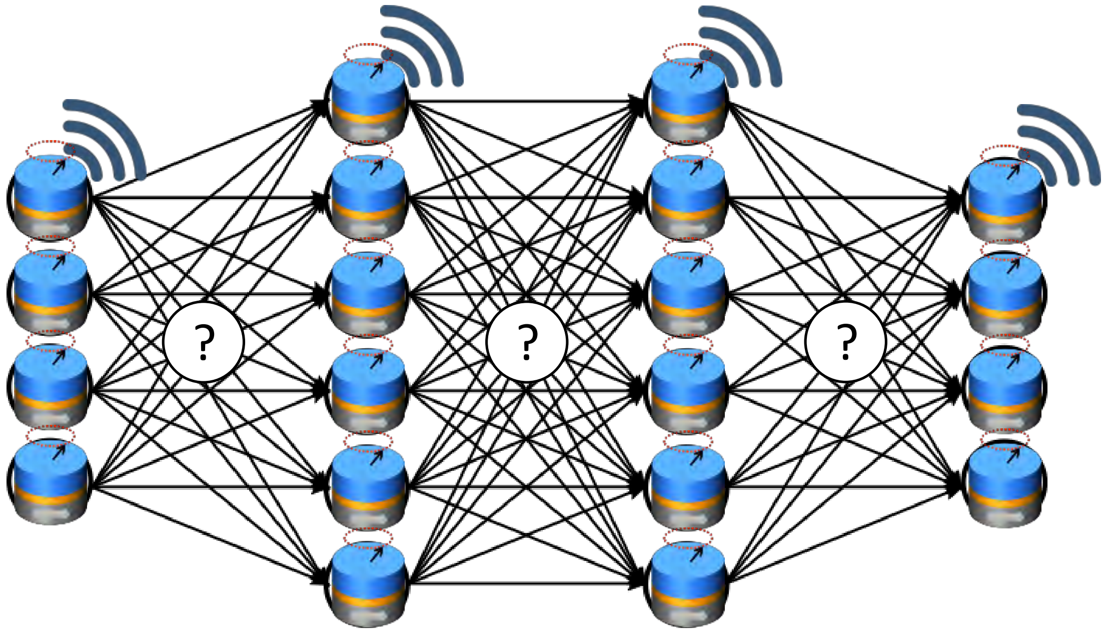
Step 2: RF communication between the two layers of a magnetic neural network



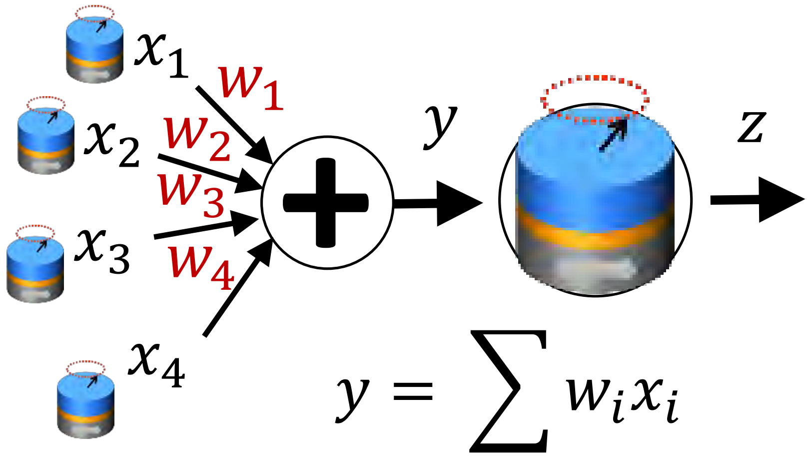


M. Romera, P. Talatchian et al, Nature 563, 230 (2018)

Step 3: connect layers of radio-frequency neurons with tunable synapses

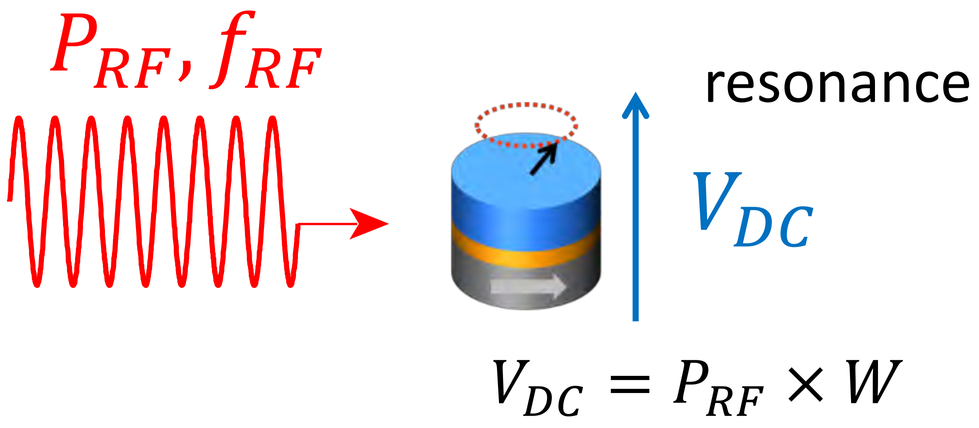


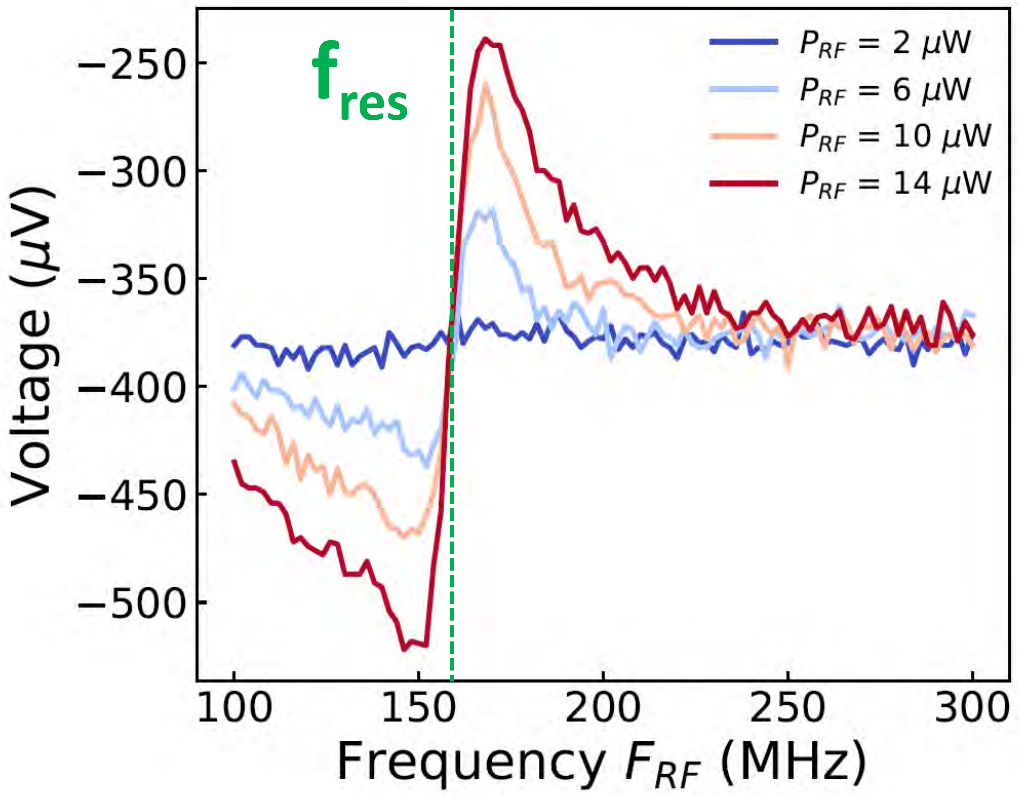
- **Multiply-And-Accumulate (MAC)**



N. Leroux et al, Radio-Frequency Multiply-And-Accumulate Operations with Spintronic Synapses, arxiv:2011.07885

A magnetic tunnel junction can perform the multiplication operation on an RF signal





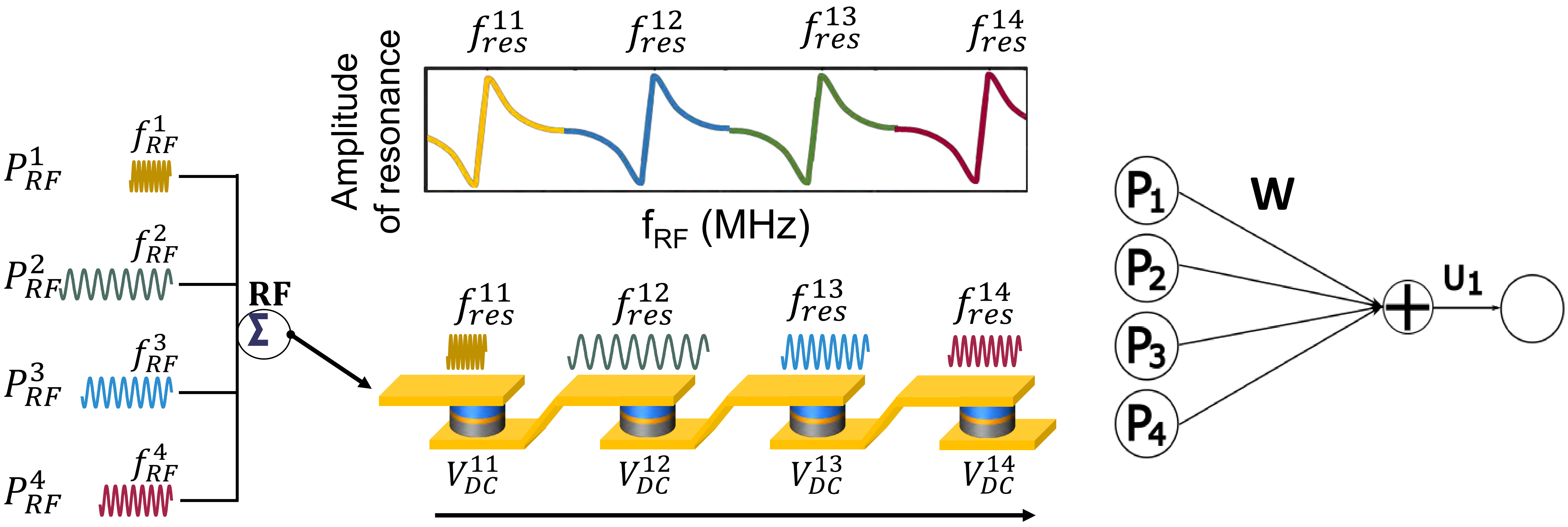
Output = **Input** * **Weight**

Input: RF Power received by MTJ

Output: DC Voltage accross the MTJ

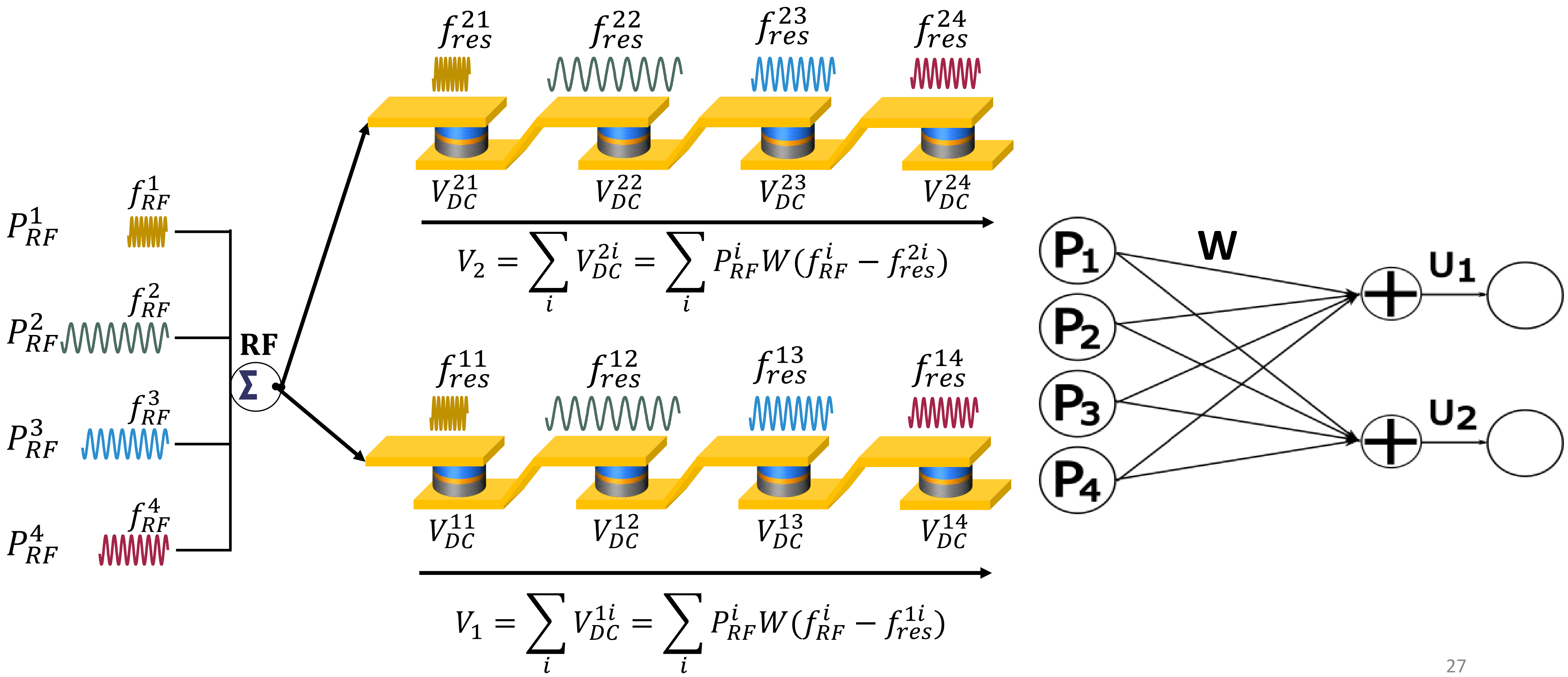
Weight is a function of frequency mismatch **$W(f_{RF} – f_{res})$**

We perform the MAC operation through frequency multiplexing



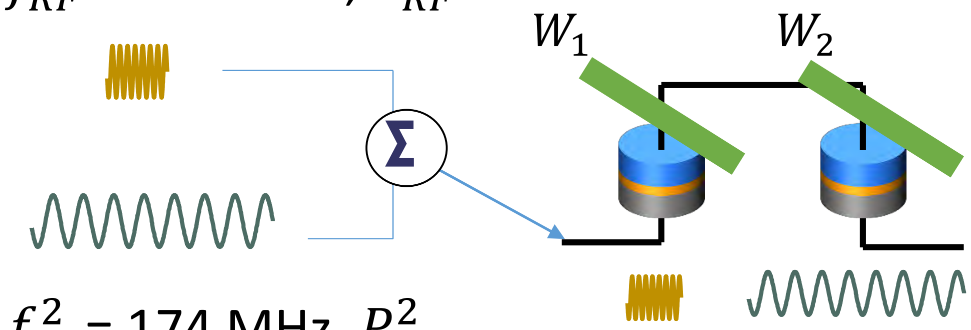
$$V_1 = \sum_i V_{DC}^{1i} = \sum_i P_{RF}^i W(f_{RF}^i - f_{res}^{1i})$$

Frequency multiplexing make high density connectivity possible



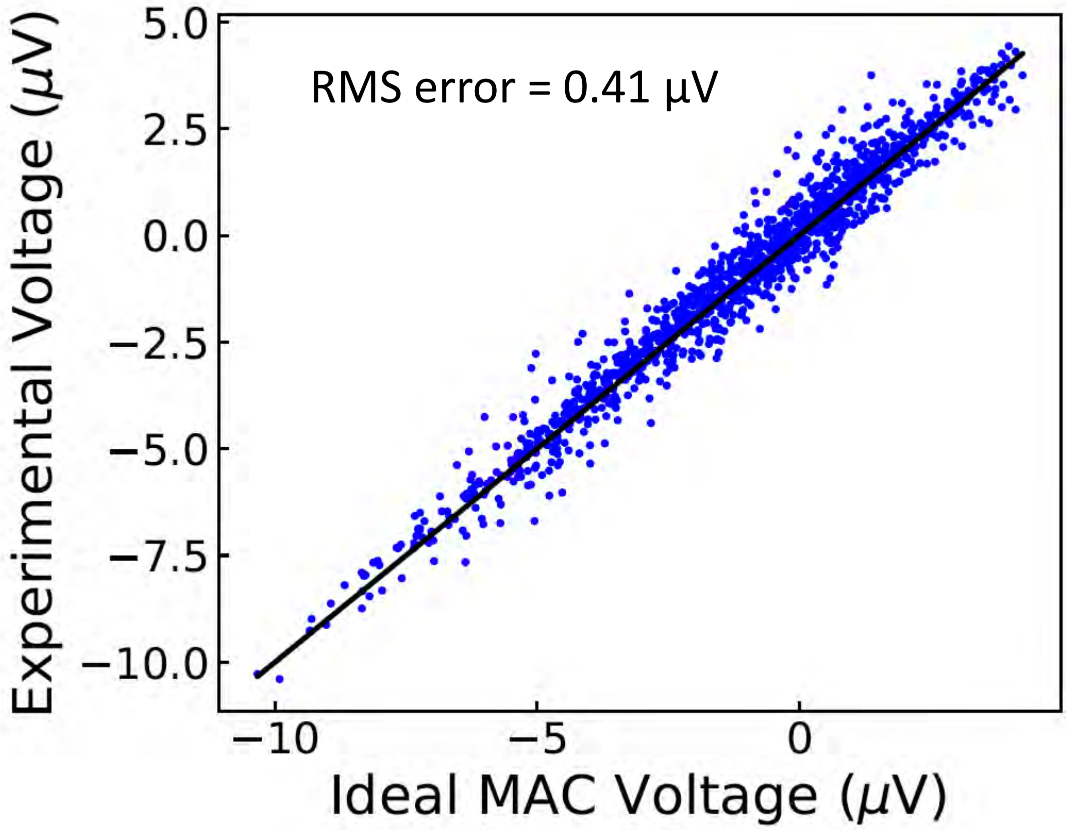
Two magnetic tunnel junctions perform the MAC on RF signals

f_{RF}^{1} = 540 MHz, P_{RF}^{1}

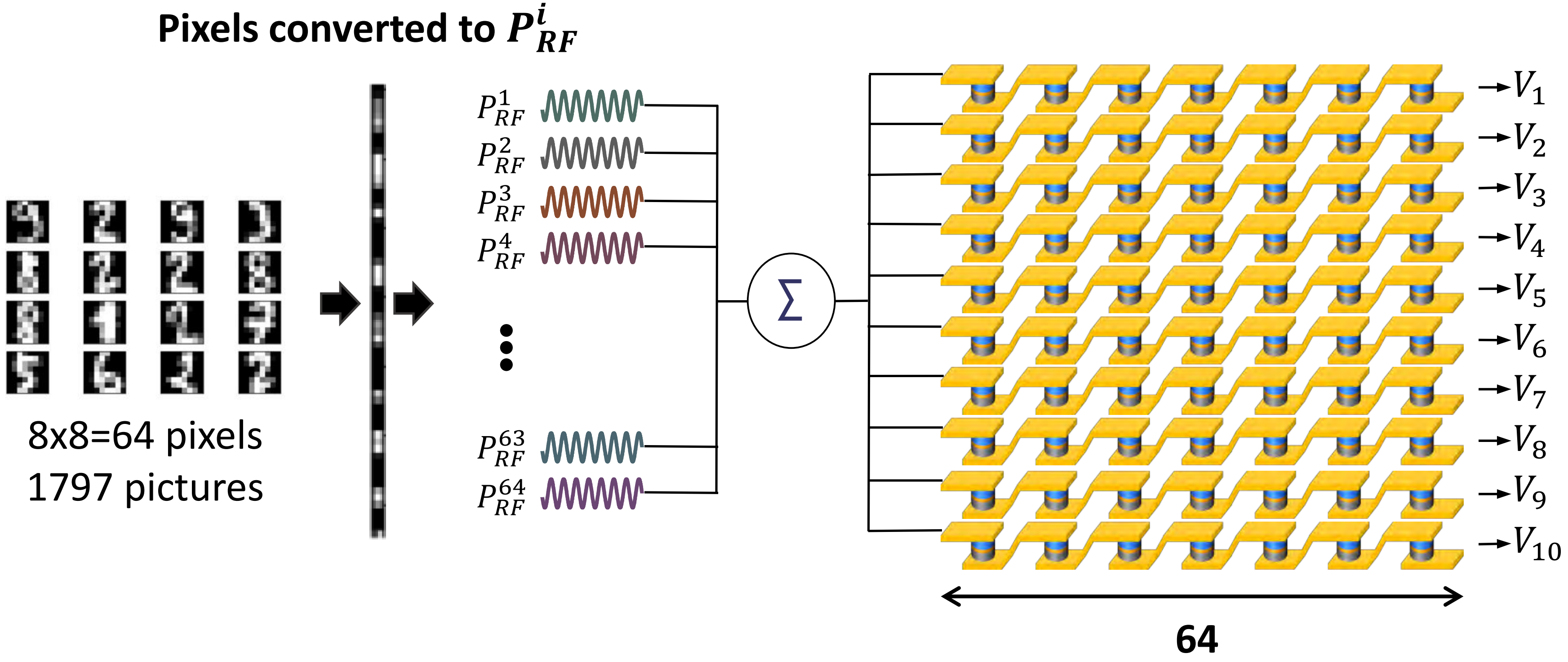


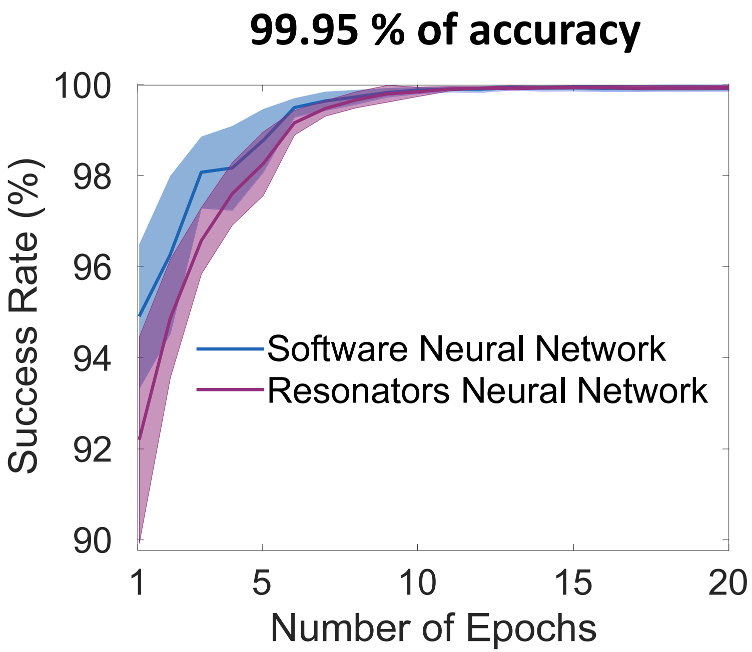
f_{RF}^{2}= 174 MHz, P_{RF}^{2}

$$V_{th} = P_{RF}^{1} \times W_1\left(f_{RF}^{1} - f_{res}^{1}\right) + P_{RF}^{2} \times W_2\left(f_{RF}^{2} - f_{res}^{2}\right)$$

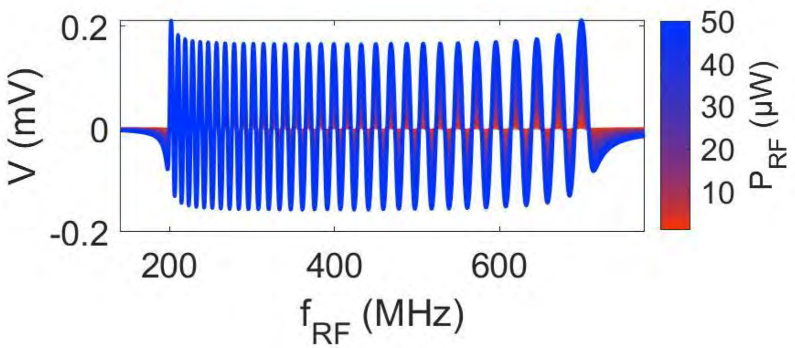


A simulated single synaptic layer perceptron recognizes digits database



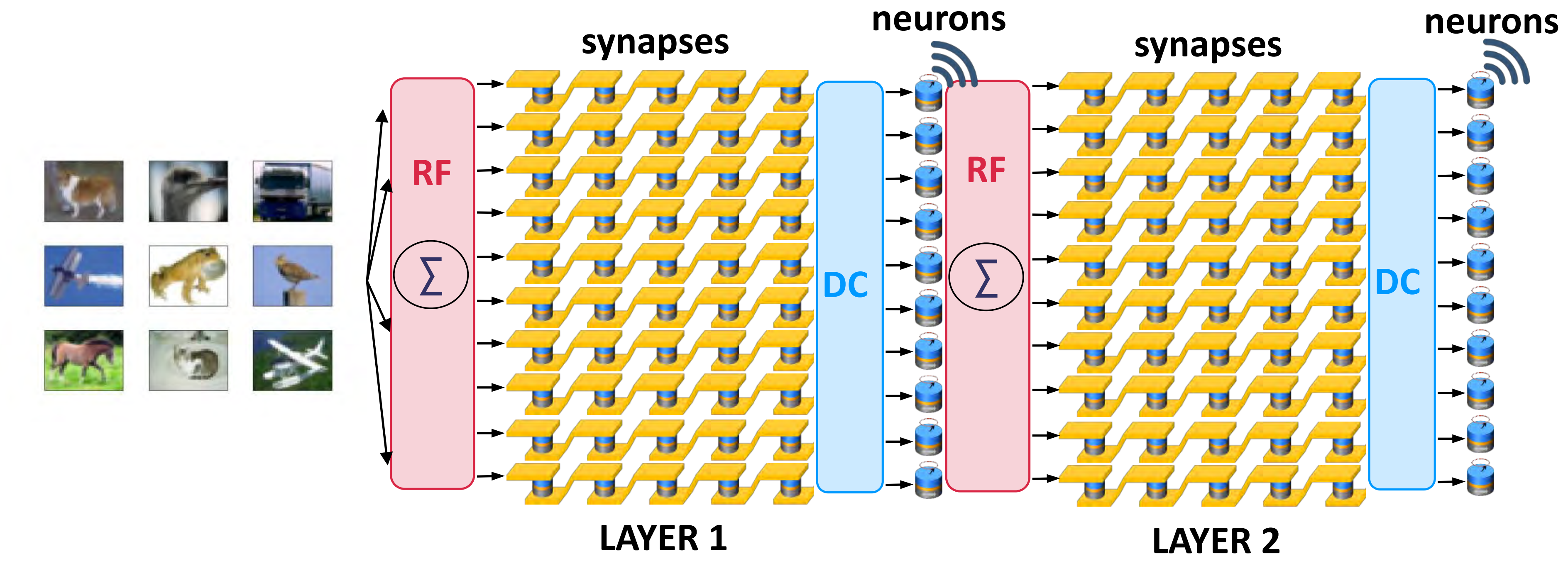


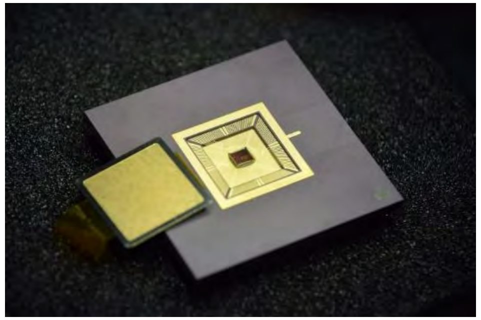
- Simulations realized with *PyTorch*
- Analytical model for the spin-diodes



N. Leroux et al, arxiv:2011.07885

Our goal is to design and build a deep neural network made of spintronic nano-synapses and nano-neurons with RF interconnexions

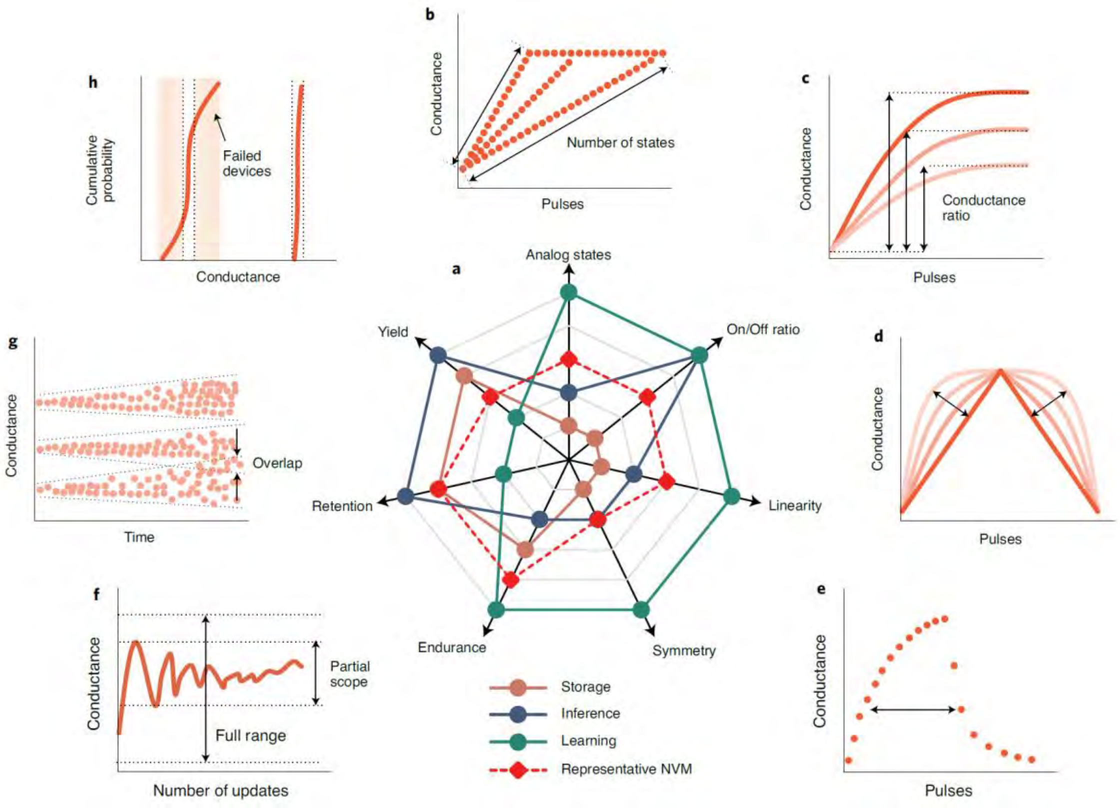




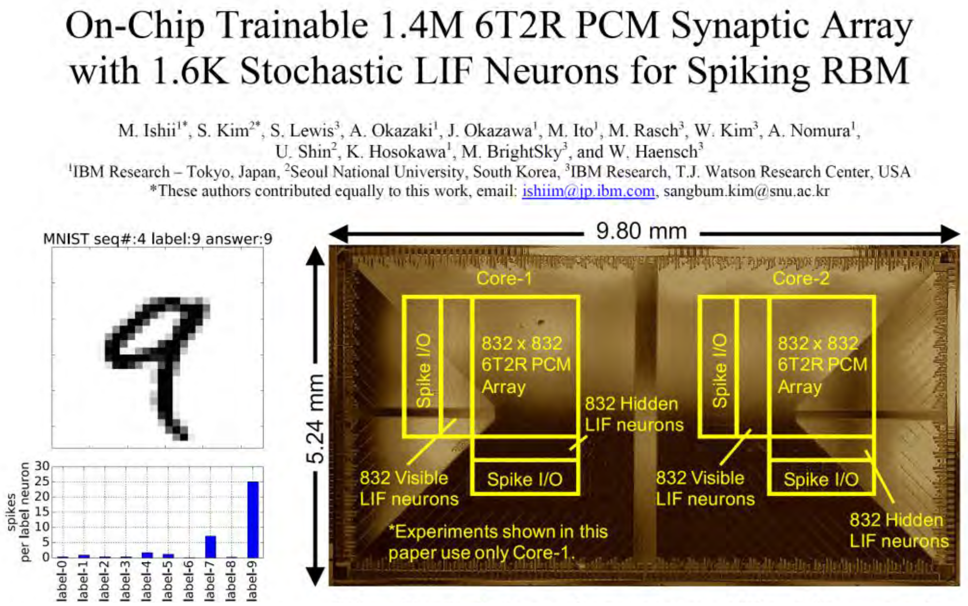
The downside of novel nanotechnologies for AI

Nanodevices are by essence noisy, imperfect and highly variable from device to device

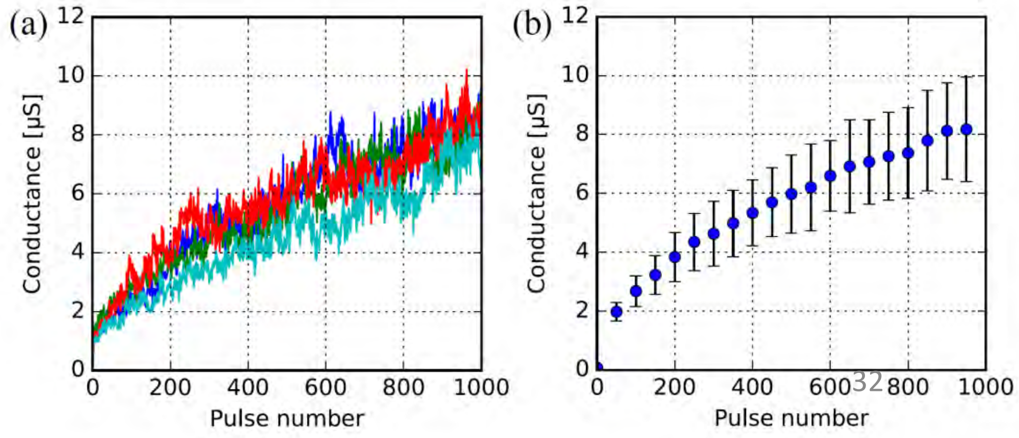
Panorama of memristor synapse faults



Zhang et al, Nature Electronics 3, 371 (2020)

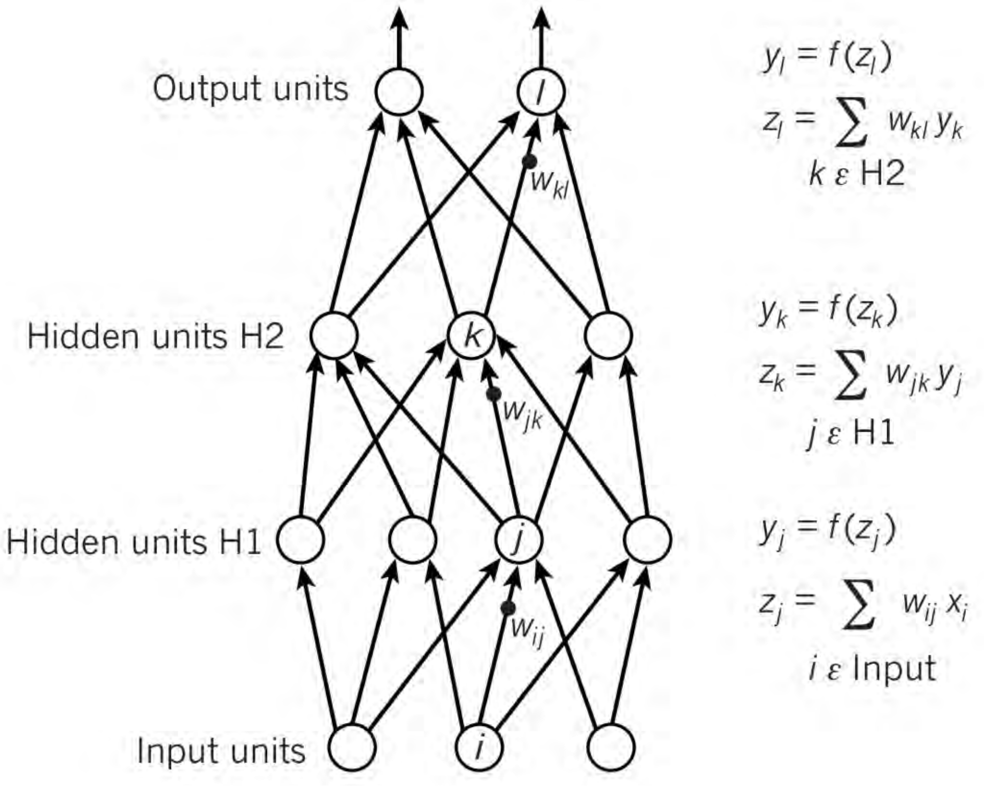


First fully integrated memristor/CMOS chip: only 92% on MNIST due to device variability



They are hardly compatible with the flagship training algorithm of deep neural networks: backpropagation of errors

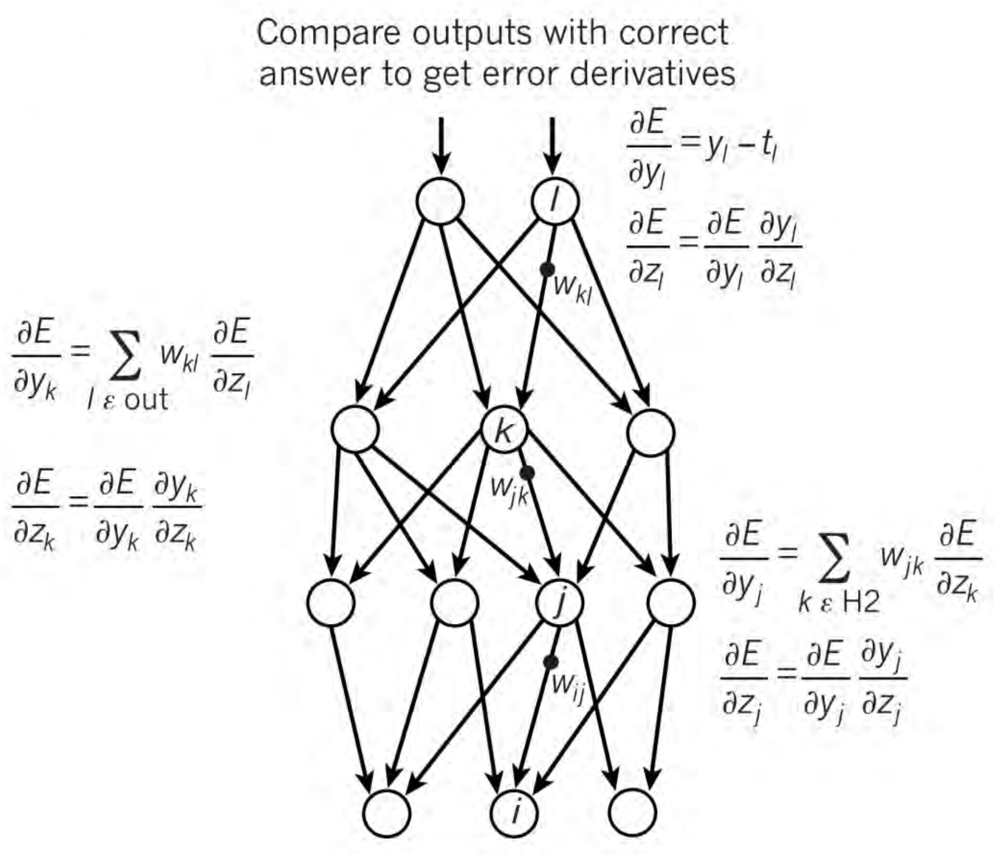
Forward pass: inference



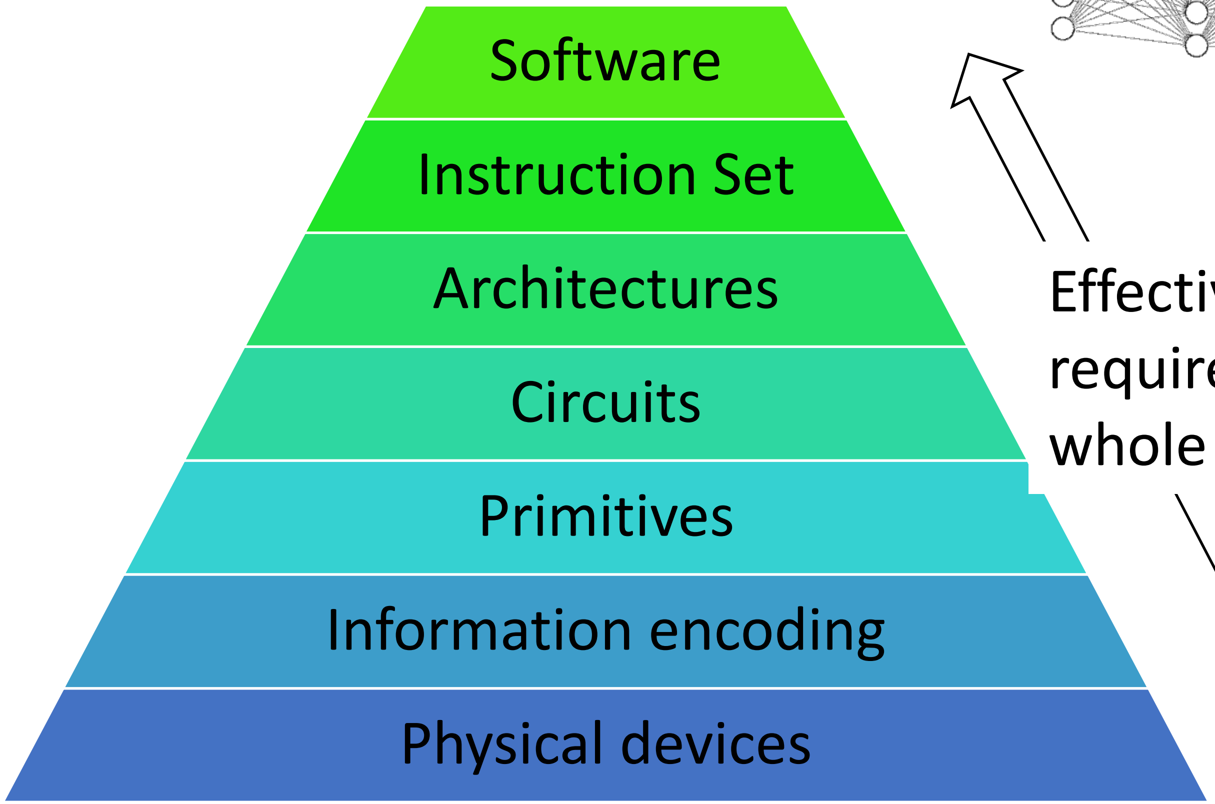
$$\Delta w = -\alpha \frac{\partial E}{\partial w}$$

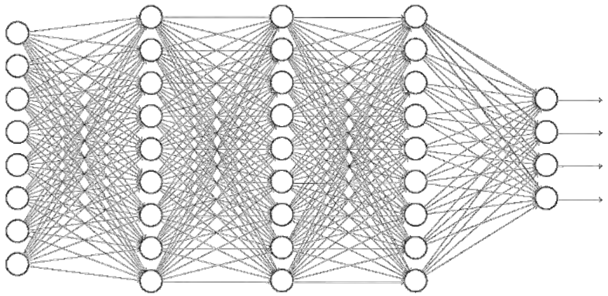
$$\frac{\Delta w}{w} < 10^{-5}$$

Backward pass

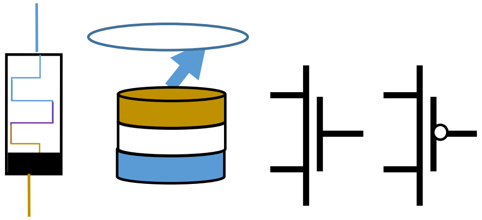


Yann Lecun, Yoshua Bengio and Geoffrey Hinton, *Nature* 521, 436 (2015)





Effective use of new devices requires working across the whole computational stack



Three main approaches

1- implement backpropagation AI inspired

2- make backpropagation more hardware-compatible (top-down)

3 - find new ways to perform hardware-compatible learning (bottom-up)

Neuroscience & AI inspired

Three main approaches

1- implement backpropagation AI inspired

2- make backpropagation more hardware-compatible (top-down)

3 - find new ways to perform hardware-compatible learning (bottom-up)

Neuroscience & AI inspired

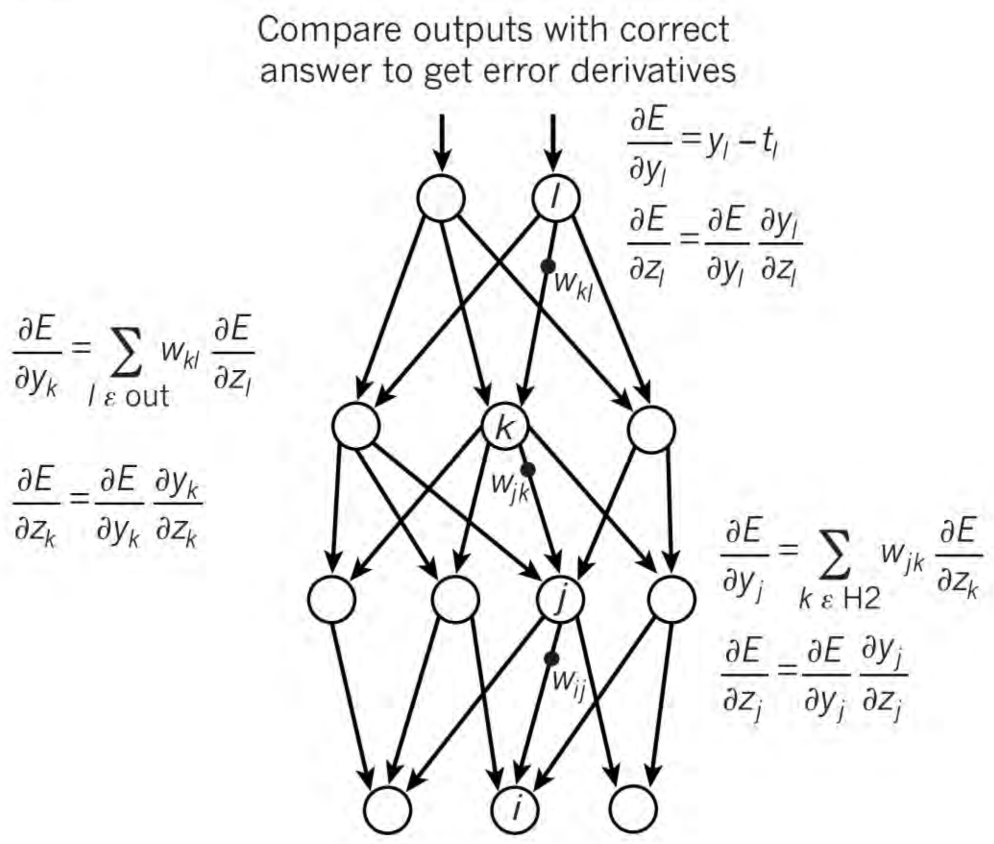
Geoffrey Hinton
AI pioneer
Turing Prize

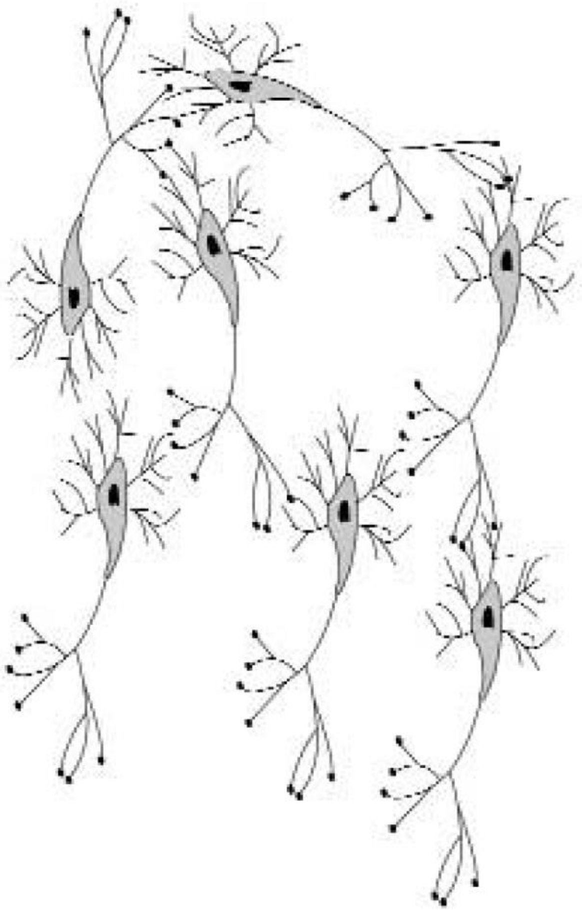
Stanford Seminar - Can the brain do back-propagation?

Can the brain do a form of backpropagation?

Backpropagation requires cumbersome external circuits and additional memories to store activations and gradients

Backward pass





There are no external circuits, no additional memories in the brain: how are gradients computed, stored and applied to synapses ?

Lillicrap et al, "Backpropagation and the brain", Nature Reviews Neuroscience (2020)

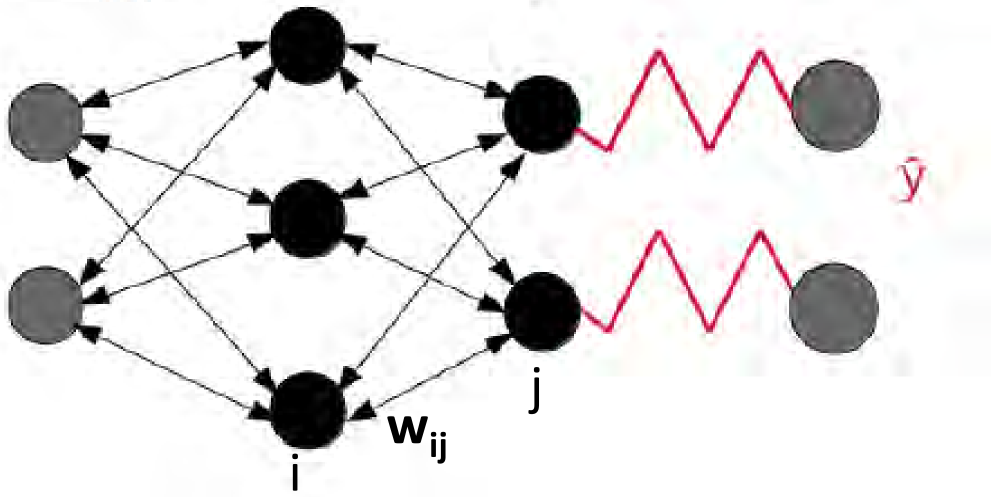
Learning through physics: networks that minimize their error at the same time as they minimize their energy

$$\frac{ds}{dt} = -\frac{\partial F}{\partial s}$$

$$F = E(s) + \beta C(y, \hat{y})$$

Cost function

B. Scellier & Y. Bengio, Front. Comput. Neuroscience 04 May 2017



s → neuron state

ρ → neuron rate = neuron output

Learning rule: $$\frac{dw_{ij}}{dt} = \dot{\rho}(s_i)\rho(s_j) + \dot{\rho}(s_j)\rho(s_i)$$

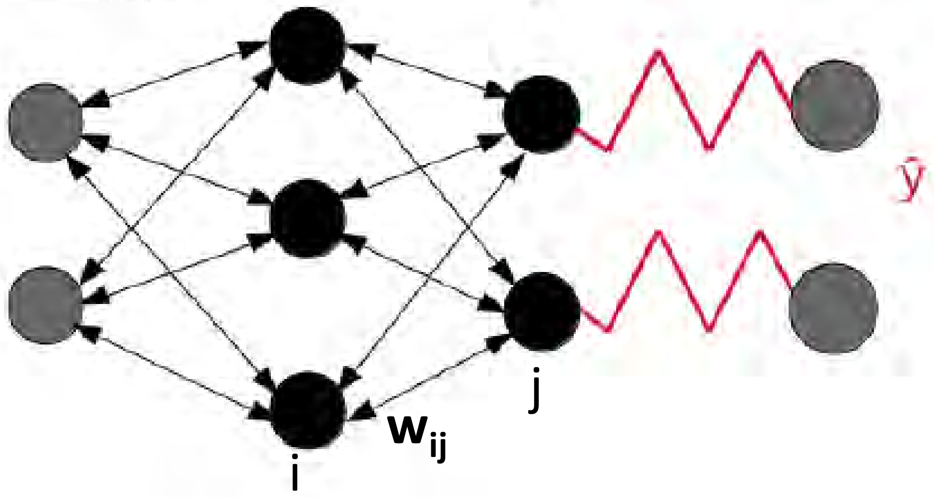
Learning through physics: networks that minimize their error at the same time as they minimize their energy

$$\frac{ds}{dt} = -\frac{\partial F}{\partial s}$$

$$F = E(s) + \beta C(y, \hat{y})$$

Cost function

B. Scellier & Y. Bengio, Front. Comput. Neuroscience 04 May 2017



s → neuron state

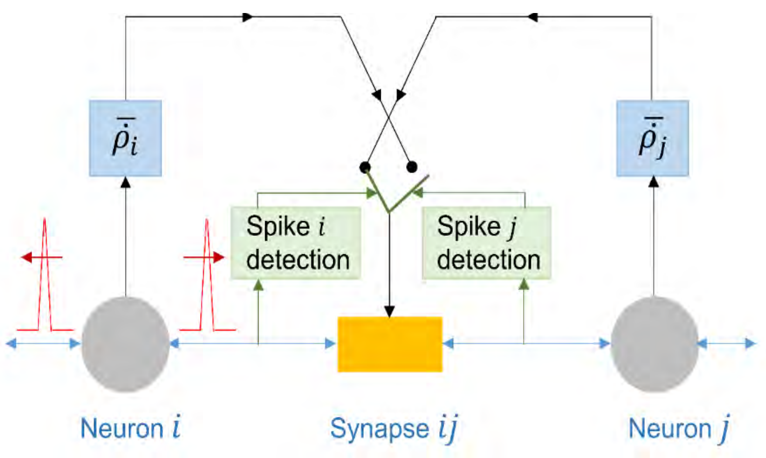
ρ → neuron rate = neuron output

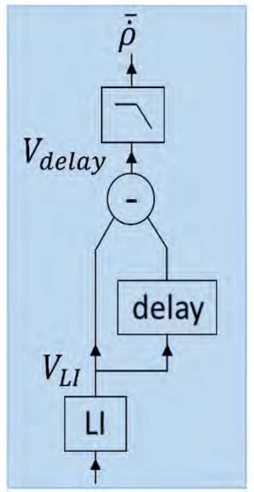
Learning rule:

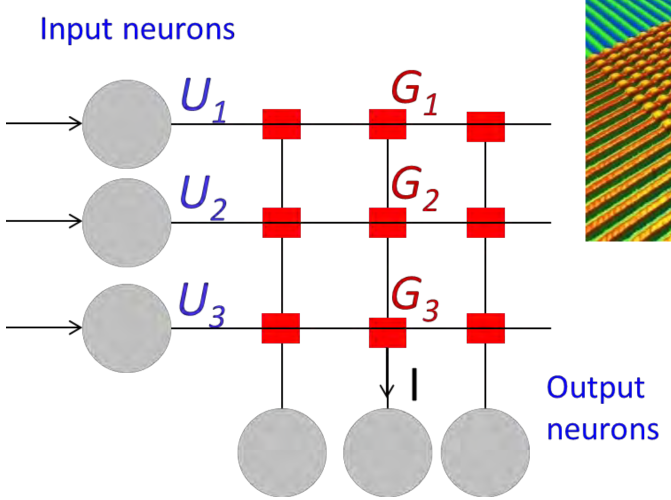
$$\frac{dw_{ij}}{dt} = \dot{\rho}(s_i)\rho(s_j) + \dot{\rho}(s_j)\rho(s_i)$$

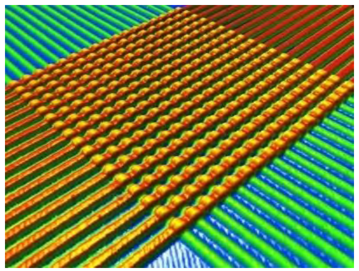
The EP learning rule is equivalent to Backpropagation through time M Ernoult, J Grollier, D Querlioz, Y Bengio, B Scellier, NeurIPS 2019

EqSpike is a spiking version of Equilibrium Propagation compatible with neuromorphic implementations









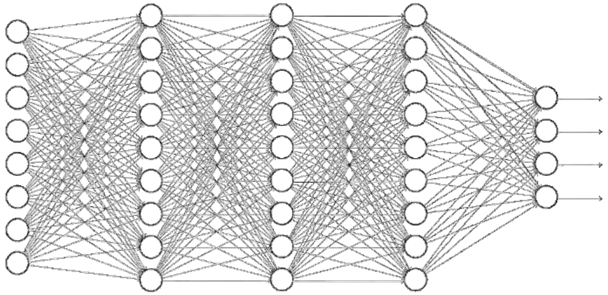
Bidirectional SNN (784 -300-10), 97.6% on MNIST (SOA for online-trained SNNs)

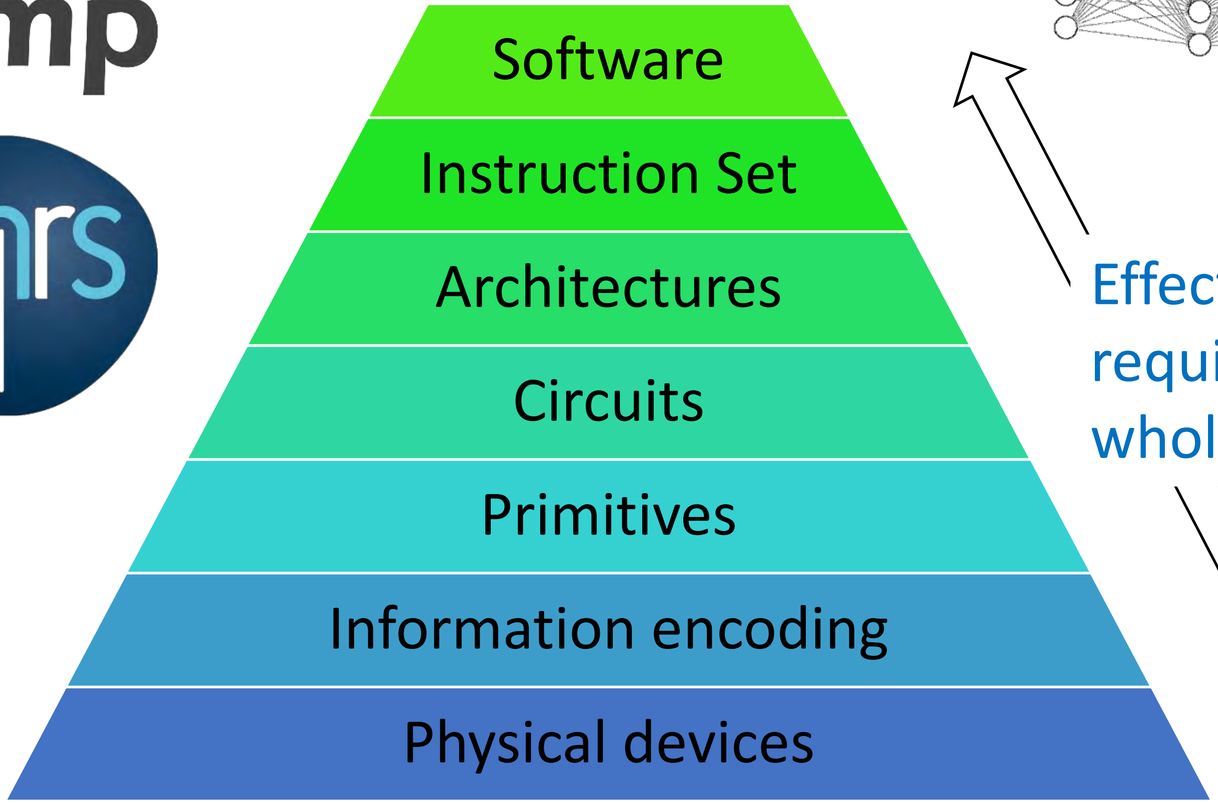
Towards intrinsic learning

E. Martin et al, EqSpike: Spike-driven Equilibrium Propagation for Neuromorphic Implementations, arXiv:2010.07859

Conclusion

Future high performance, low power AI requires emerging nanotechnologies and physics





Effective use of new devices requires working across the whole computational stack