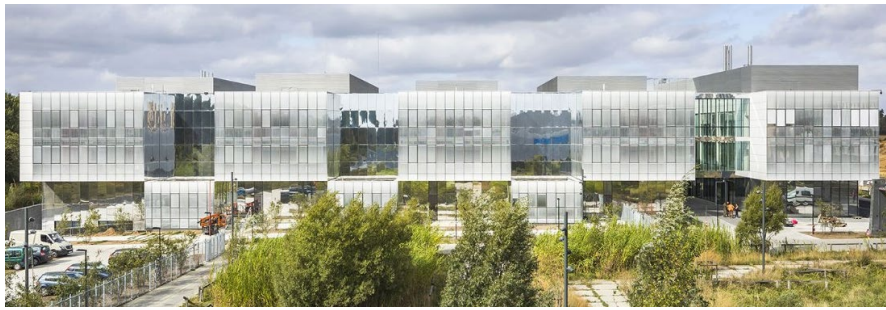
Bayesian In-Memory Computing

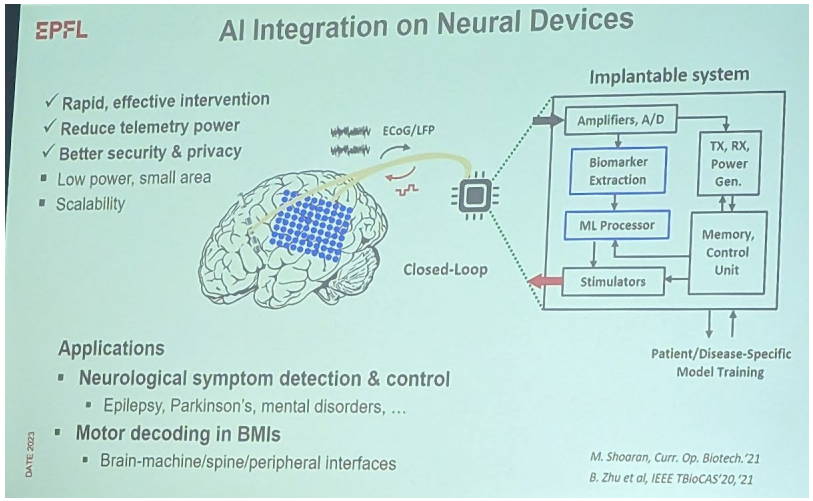
Damien Querlioz
Centre de Nanosciences et de Nanotechnologies
Université Paris-Saclay, CNRS, Palaiseau, France
damien.querlioz@universite-paris-saclay.fr



Joint work with the groups of Marc Bocquet and Jean-Michel Portal (Aix-Marseille Univ.) and Elisa Vianello (CEA-LETI)

Edge AI Has an Incredible Potential for Safety-Critical Applications

Medical: Predicting epileptic seizures, closed-loop Parkinson DBS…

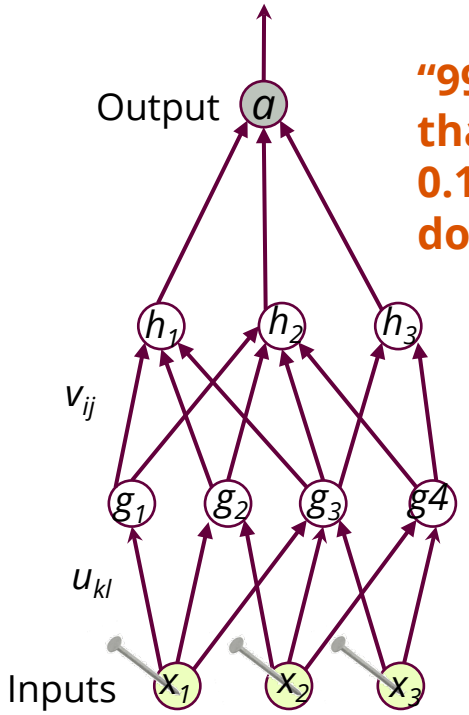


Shoaran Mahsa (EPFL) talk
DATE 2023, Special Day on Human-AI Interaction

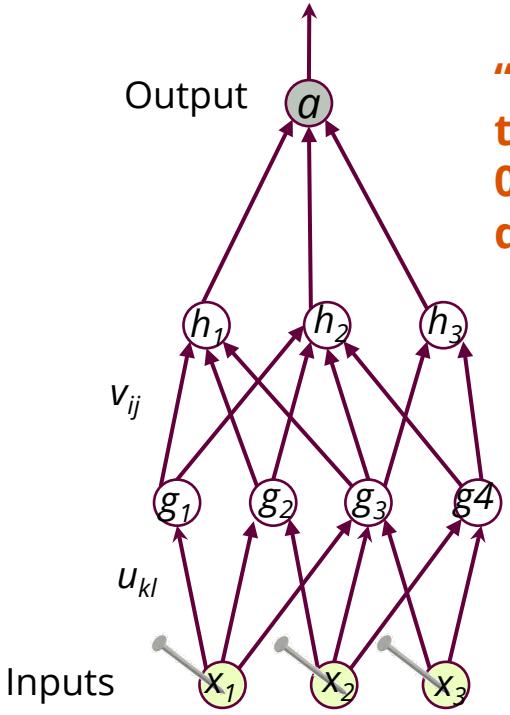
Industrial: Monitoring early drifts/faults to avoid accidents

Modern AI Is Amazing
But It Has a Confidence Problem

- If a network has been trained to distinguish CATS and DOGS



"99.9% proba. that this a cat, 0.1% that it is a dog"

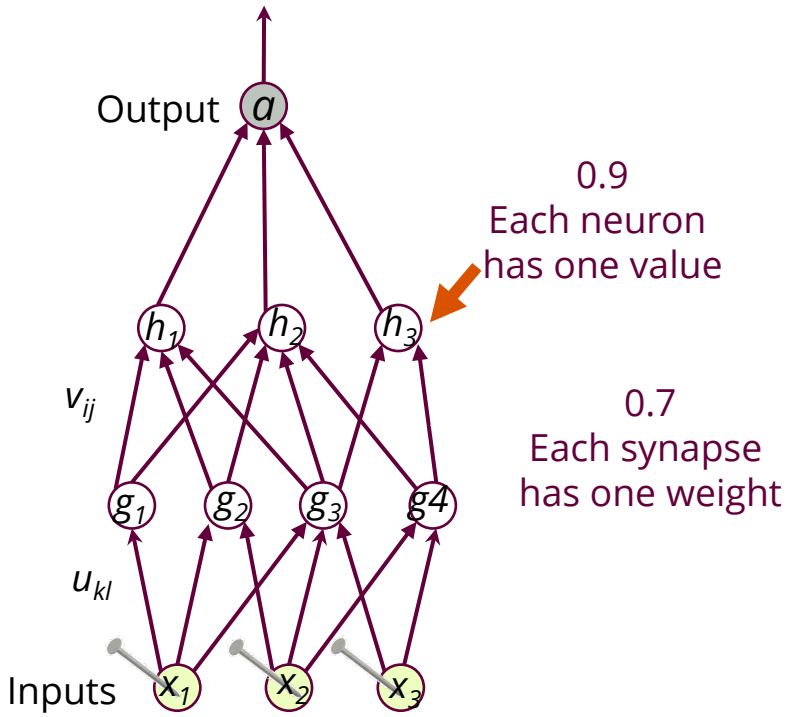


"99.7% proba. that this a cat, 0.3% that it is a dog"

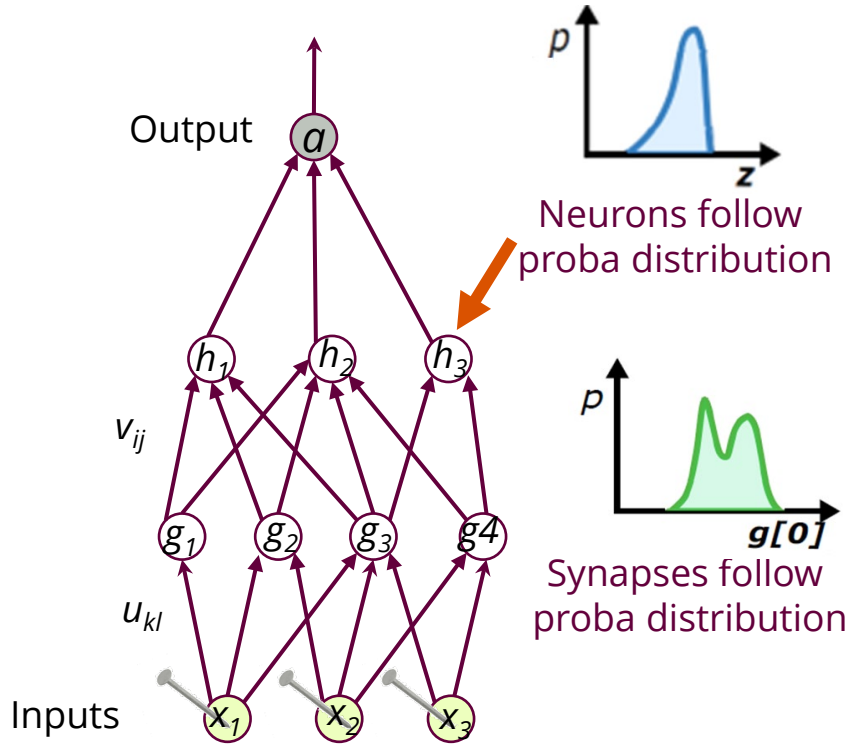
A Lead: Bayesian Models

- In Bayesian models, everything is considered a random variable that follows specific probability distributions

Traditional neural network



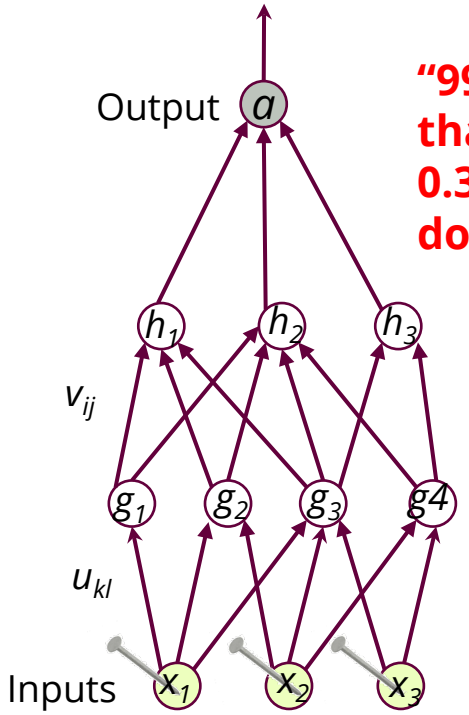
Bayesian neural network



A Lead: Bayesian Models

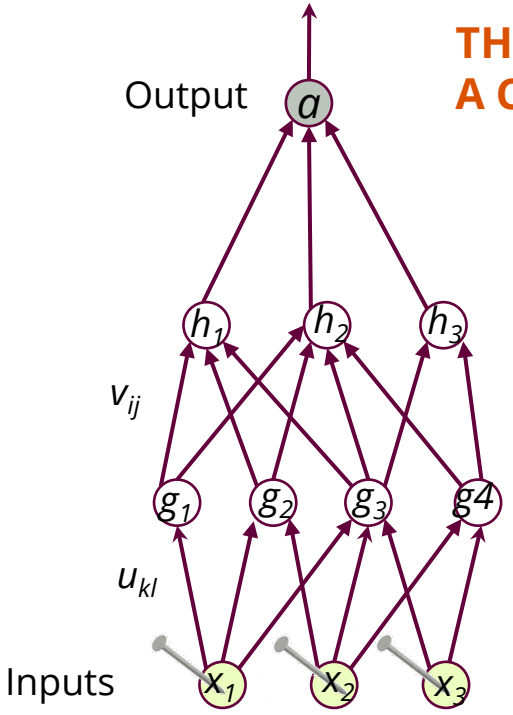
- If a network has been trained to distinguish CATS and DOGS

Traditional neural network



"99.7% proba. that this a cat, 0.3% that it is a dog"

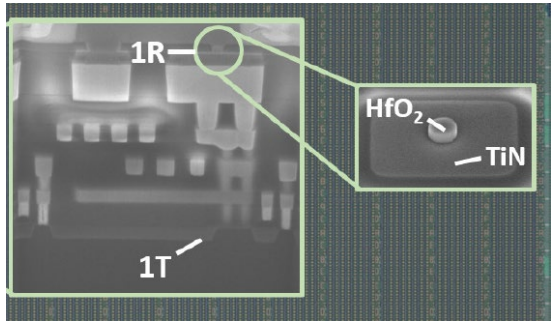
Bayesian neural network

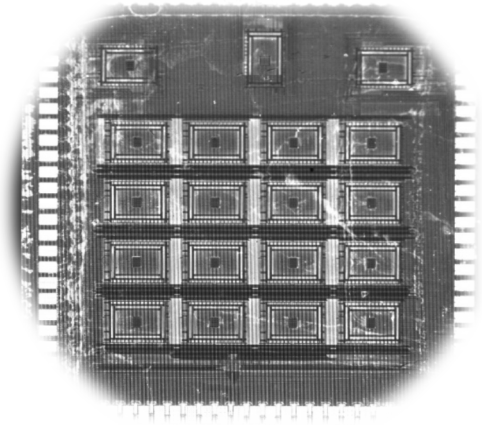


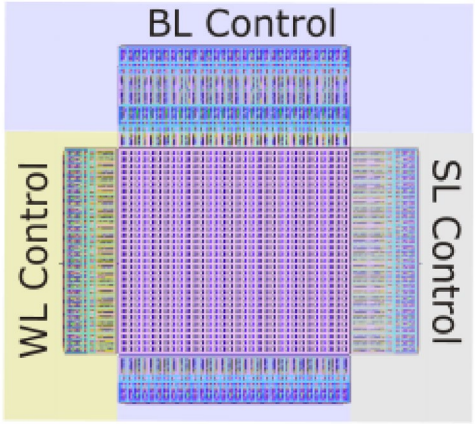
THIS IS NEITHER A CAT NOR A DOG!

Bayesian Models Are All About Probabilities

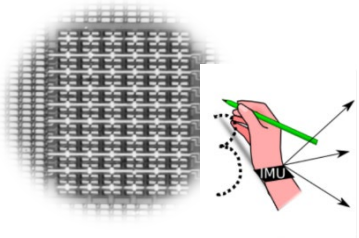
- Probabilities are hard to do on computers
- **Nanoelectronics offers many opportunities**



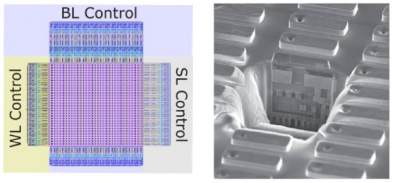




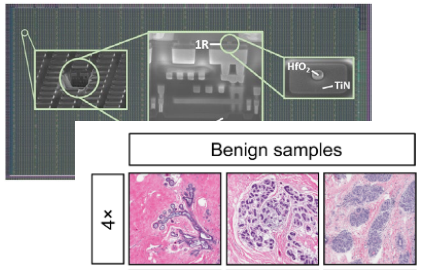
Bayesian In-Memory Computing



- The Memristor-Based Bayesian Machine

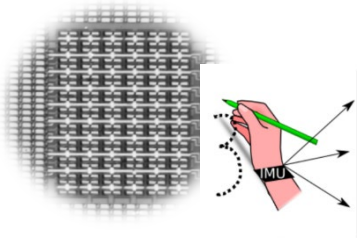


- Bayesian Neural Networks with Memristors

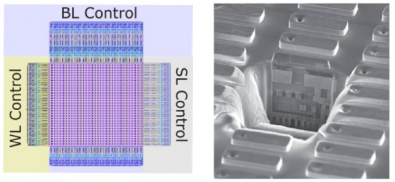


- Bayesian Learning with Memristors

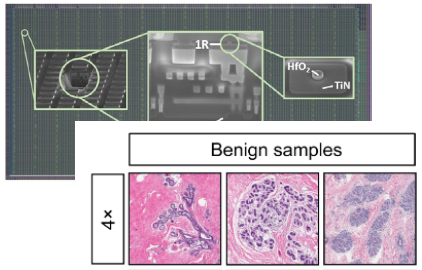
Bayesian In-Memory Computing



- The Memristor-Based Bayesian Machine

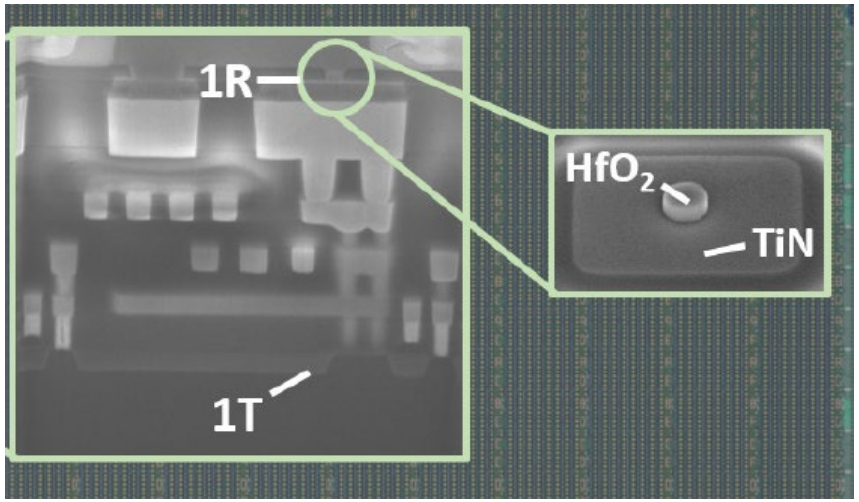


- Bayesian Neural Networks with Memristors

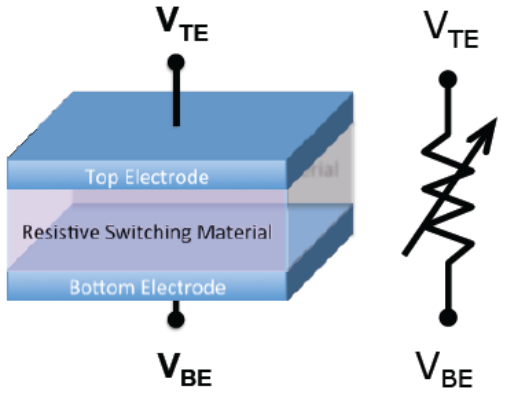


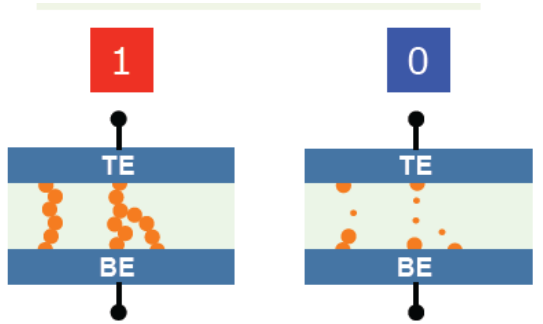
- Bayesian Learning with Memristors

Memristor/RRAM:
A Nanomemory Embeddable at the Core of CMOS



cea leti





Low resistance High resistance

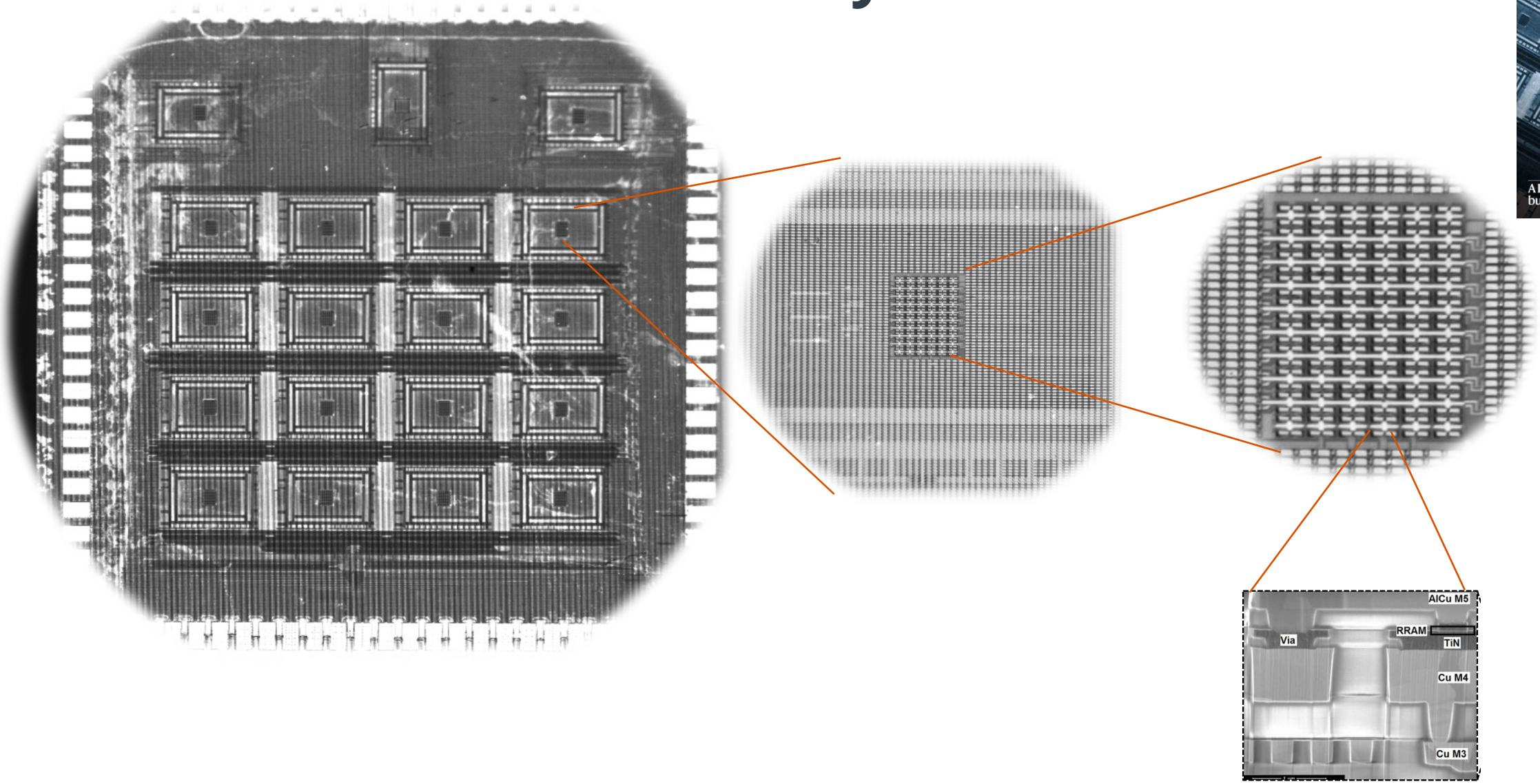
TiN/HfO$_x$/Ti/TiN stack

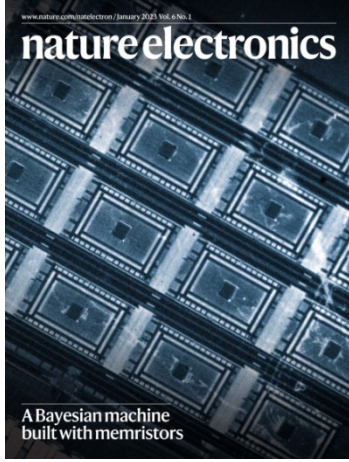
High voltage: move atoms to switch memristor between low/high resistance

Low voltage: allows reading the resistance

cnrs université PARIS-SACLAY

The Memristor-Based Bayesian Machine





Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Bayesian Reasoning: *Better at Small Data*

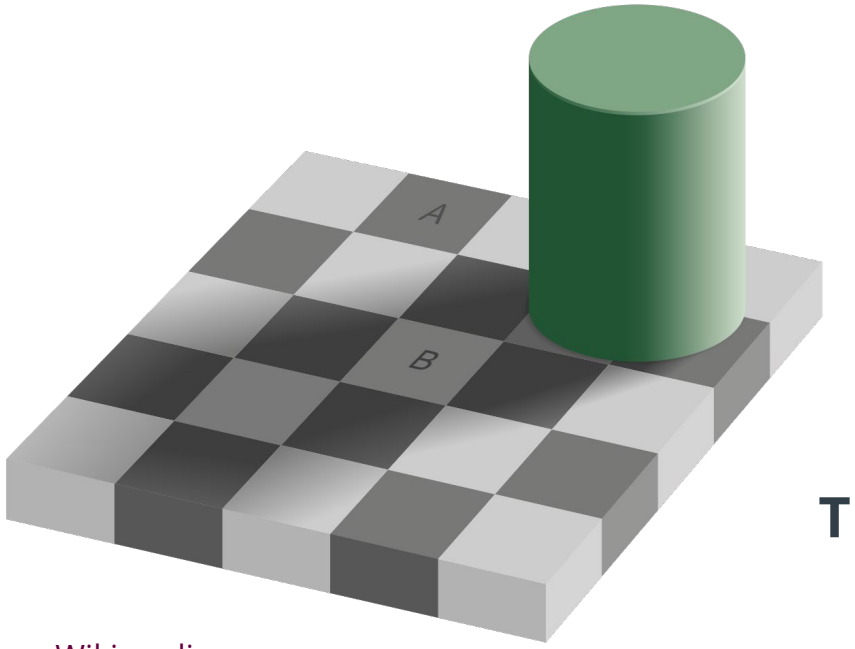
Thomas Bayes

Likelihoods *Prior*

$$p(Disease \mid Observations) \propto p(Observations \mid Disease) \times p(Disease)$$

Constructed with expert knowledge+Data

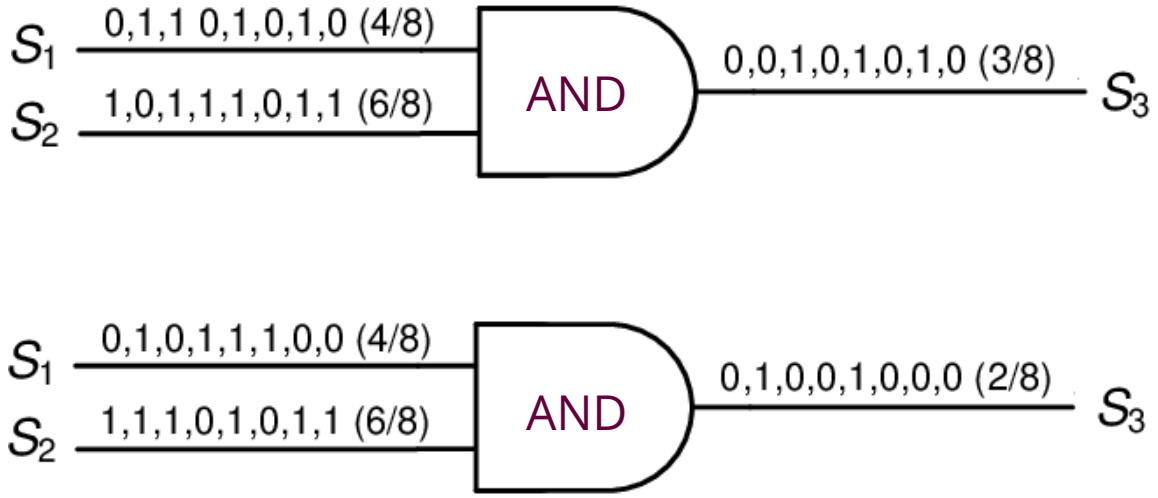
Hard to Compute



Wikimedia

The Brain Might Be Using Bayesian Reasoning

Stochastic Computing

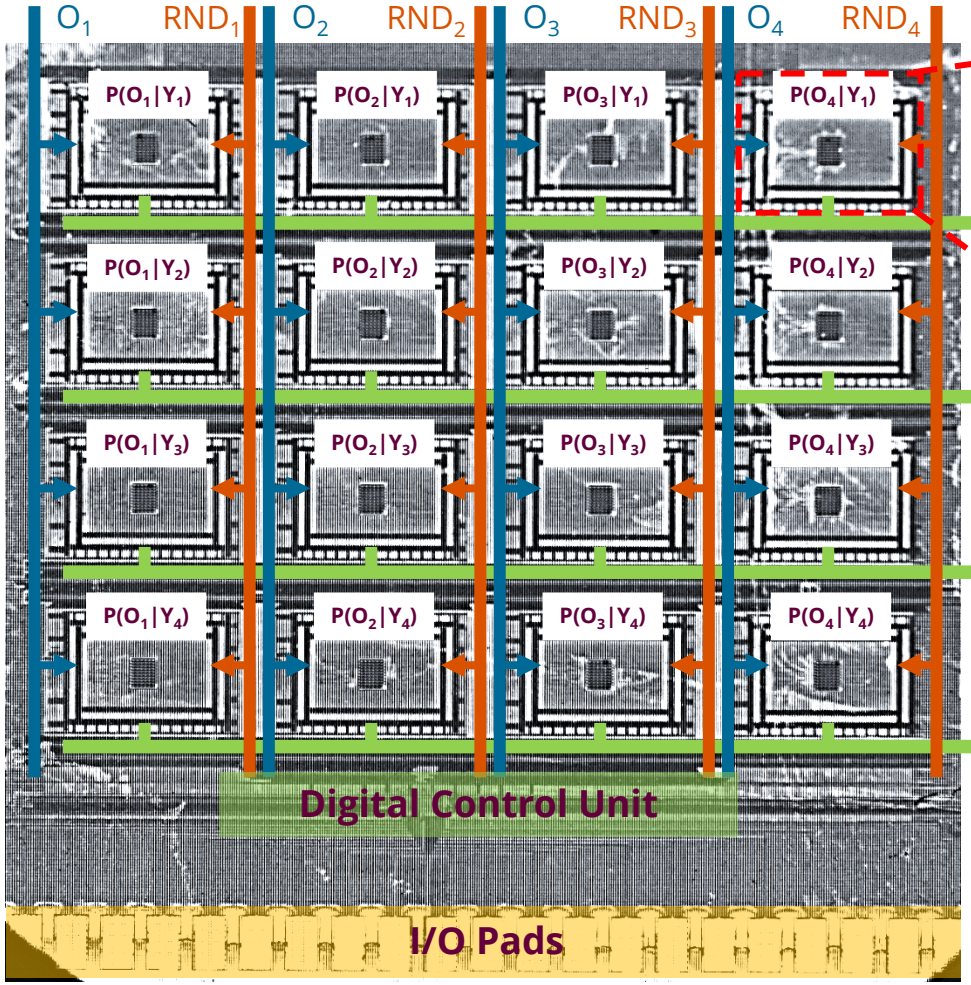


$p(S_3)=p(S_1)p(S_2)$

A AND gate implements the multiplication of two probabilities!

Some resemblance with neurons (one wire = one real number)

The Memristor-Based Bayesian Machine

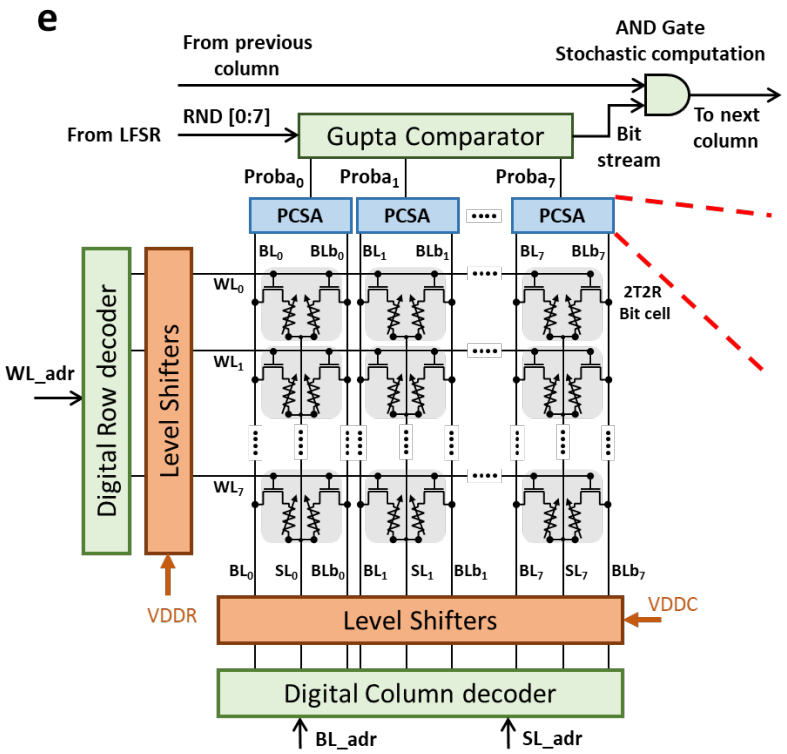


$\propto P(Y=1|O_1,O_2,O_3,O_4)$

$\propto P(Y=2|O_1,O_2,O_3,O_4)$

$\propto P(Y=3|O_1,O_2,O_3,O_4)$

$\propto P(Y=4|O_1,O_2,O_3,O_4)$

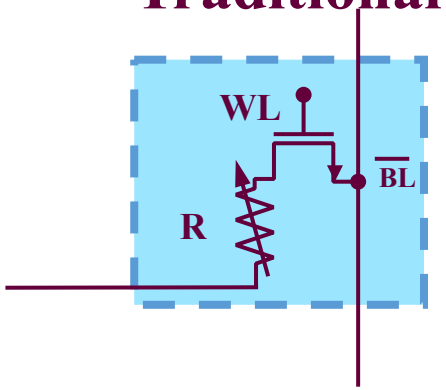


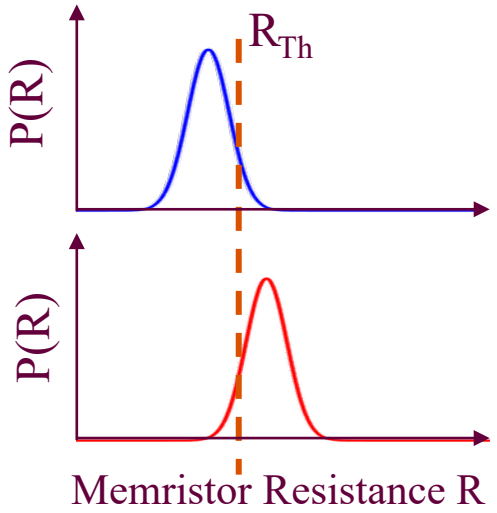
$P(Y=4|O_1,O_2,O_3,O_4) \propto P(O_1|Y=4) \times \cdots \times P(O_4|Y=4)$

Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Reducing the Error Rate without ECC

Traditional approach

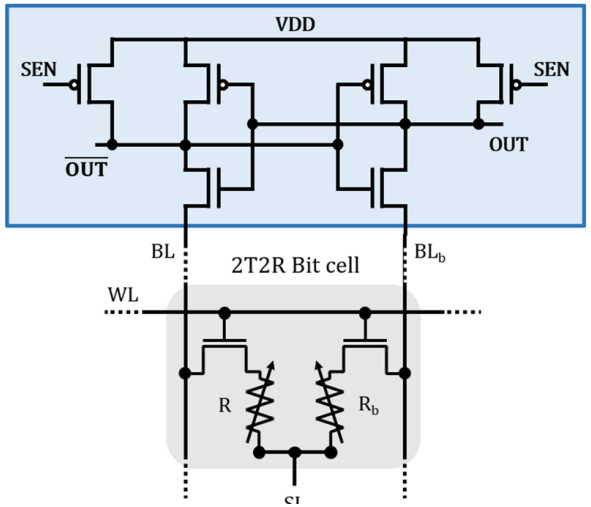




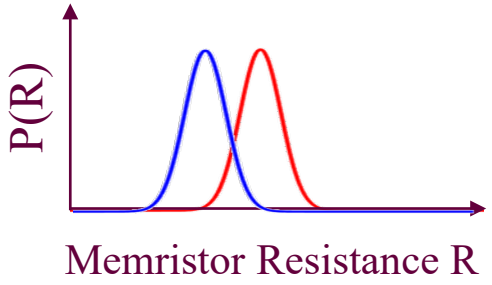
State 0 **LRS**
Error if $R(LRS)>R_{Th}$

State 1 **HRS**
Error if $R(HRS)<R_{Th}$

Our approach (Bayesian Machine)



R	R_b	
LRS	HRS	**State 1**
HRS	LRS	**State 0**



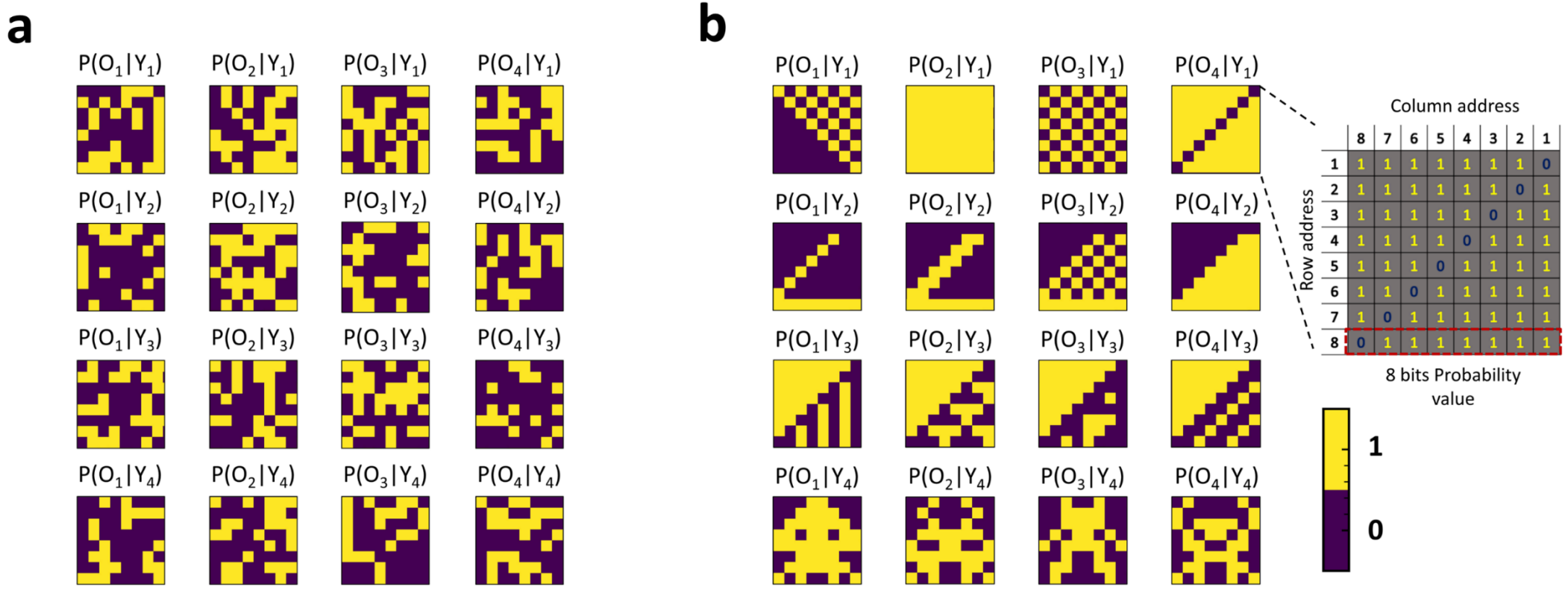
An error happens if
R(LRS)>R(HRS)

Bocquet et al, IEEE IEDM, p. 20.6.1, 2018; Hirtzlin et al., Front. Neurosci. 13, p. 1383, 2020

The Memristor-Based Bayesian Machine

Before programming the memristors

After programming the memristors

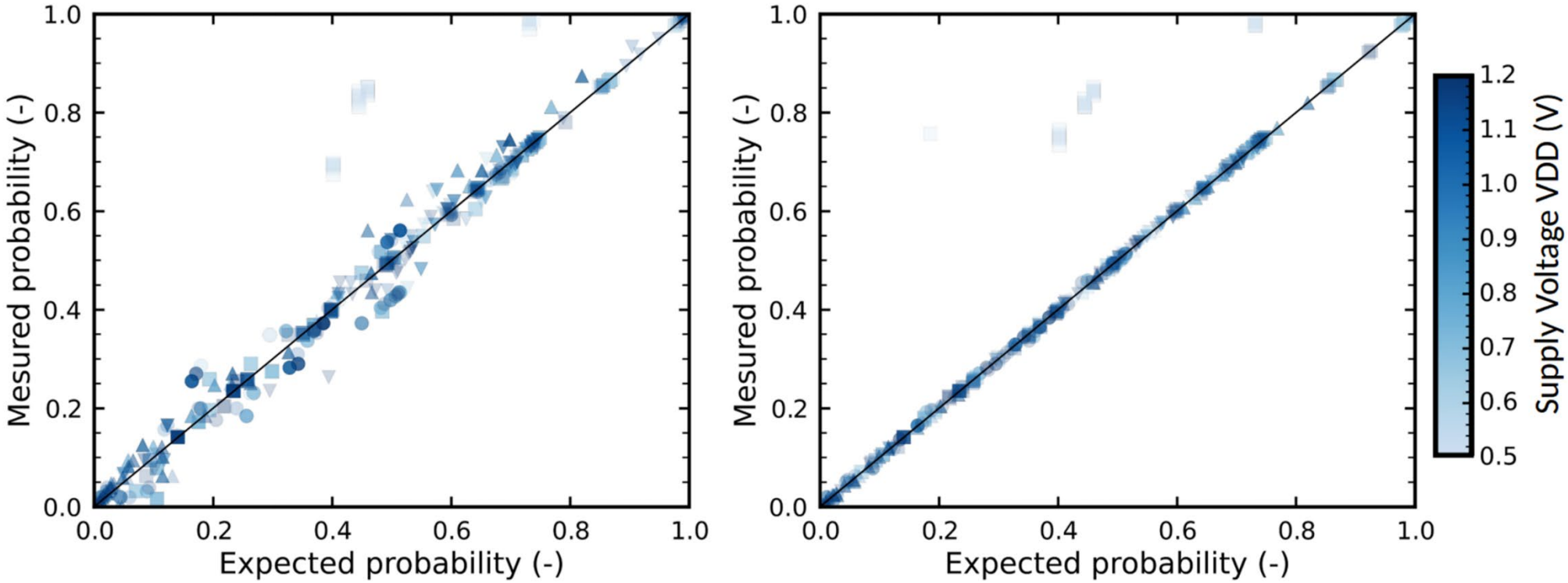


Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

The Memristor-Based Bayesian Machine

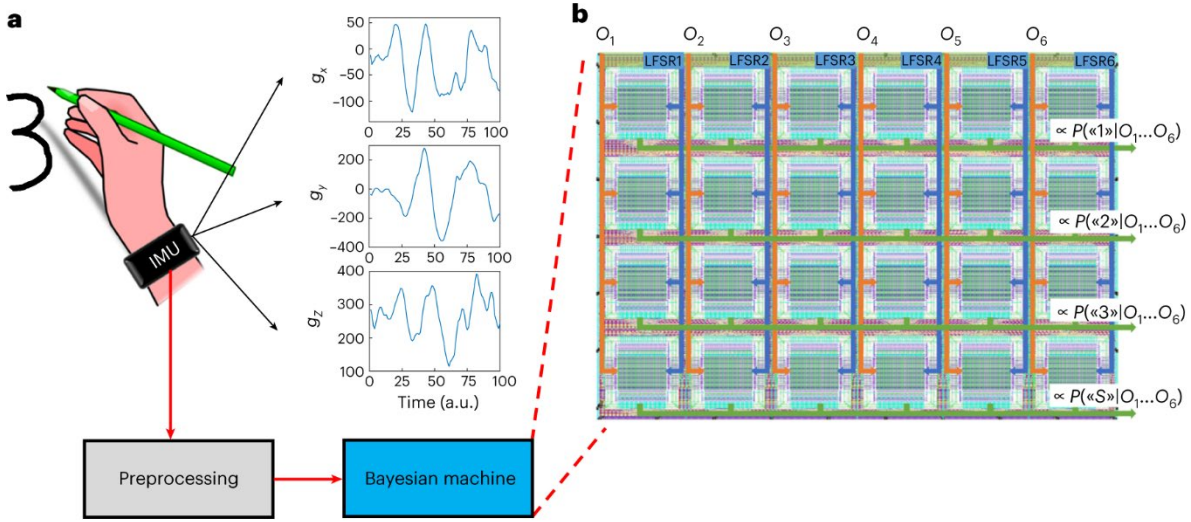
Randomly chosen pseudo-RNG seeds

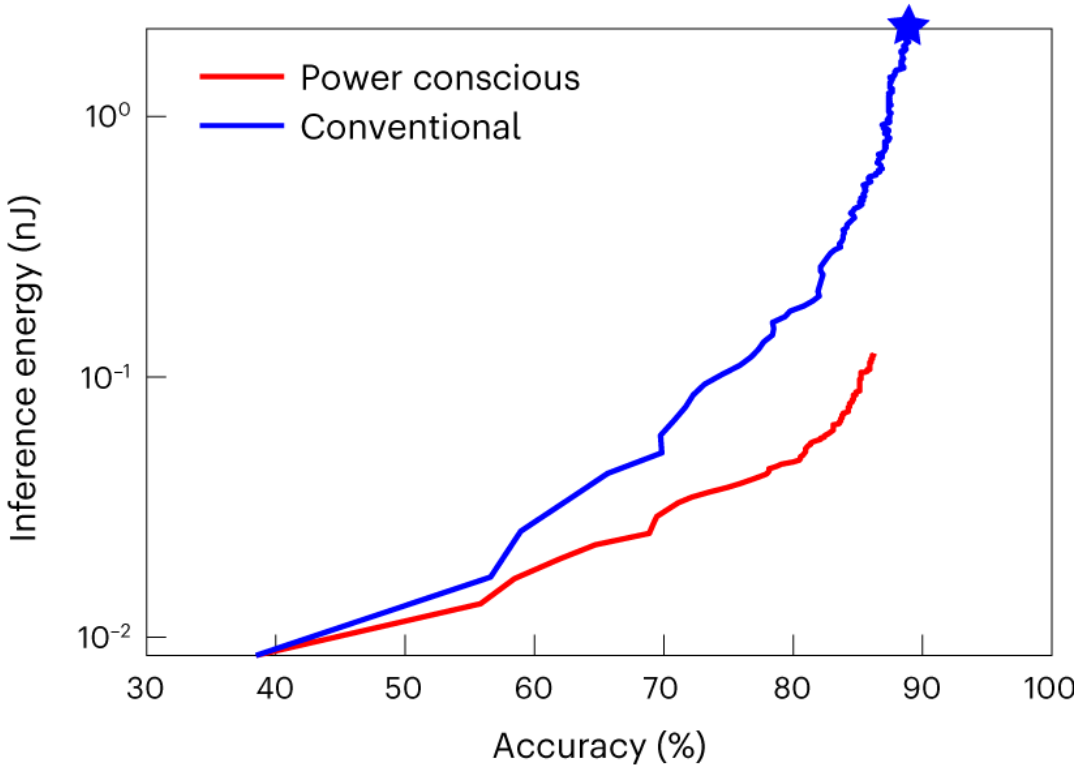
Best choice of pseudo-RNG seeds



Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Energy Consumption Is Very Low





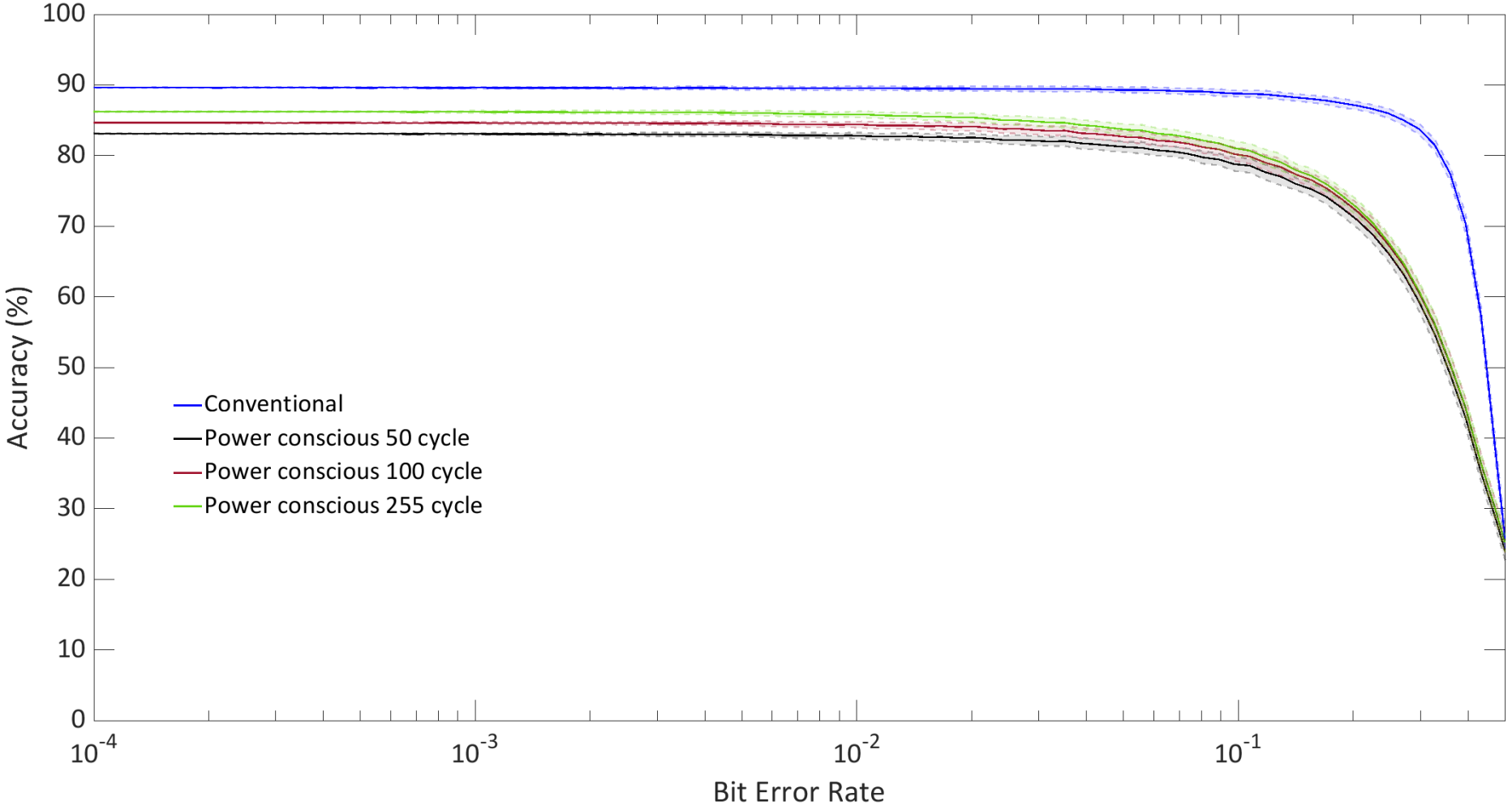
Two reasons for energy efficiency:

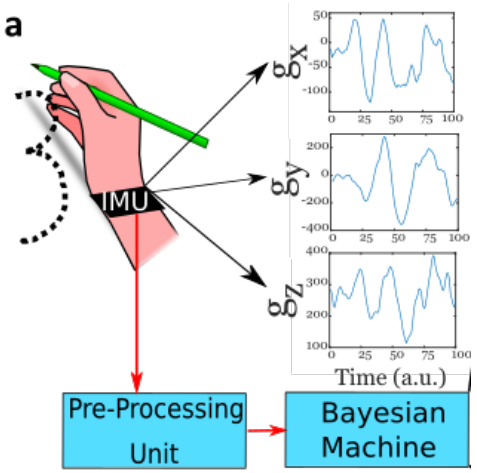
- **Near-memory computing**
- **Stochastic computing**

Same task, STM32 microcontroller unit: 2µJ

Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Benefits of Stochastic Computation

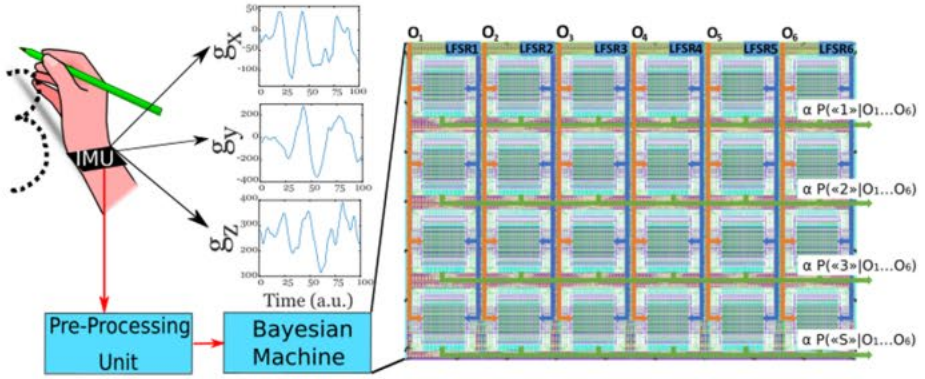




Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Uncertainty Evaluation

Signatures of an uncertain Bayesian machine

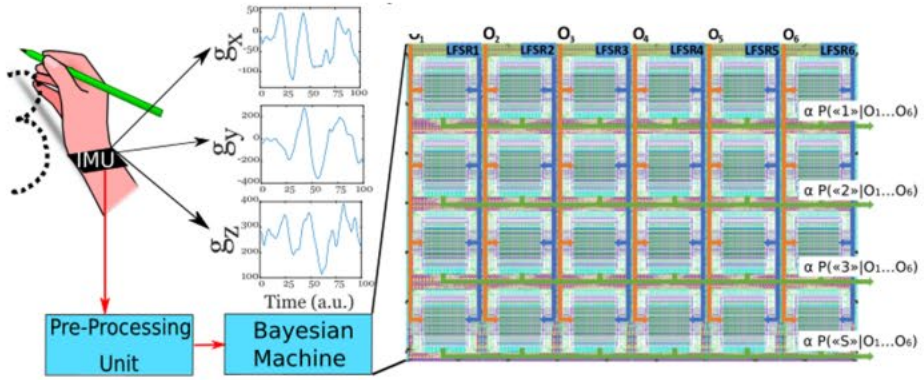


0 ones
0 ones
0 ones
3 ones

Contradictory observations:
All outputs produce a very small number of ones

EPISTEMIC UNCERTAINTY

OR



127 ones
100 ones
93 ones
3 ones

Ambivalent situation:
Some outputs produce a balanced number of ones

ALEATORIC UNCERTAINTY

Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

Uncertainty Evaluation

Case when a gesture is presented to the Bayesian machine **corresponding to a different subject**

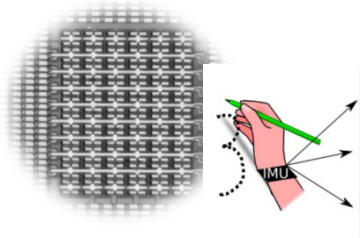
	T=0	T=5%	T=10%
Bayesian machine was uncertain (desired behavior)	93%	97%	98%
Bayesian machine provided a certain output	7%	3%	2%

Case when a gesture is presented to the Bayesian machine **corresponding to the appropriate subject**

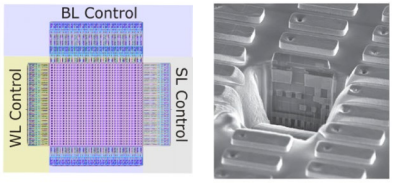
	T=0	T=5%	T=10%
Bayesian machine was certain about the correct gesture (desired behavior)	88%	80%	71%
Bayesian machine was certain about an incorrect gesture	7%	6%	5%
Bayesian machine was uncertain	5%	14%	24%

Harabi, Hirtzlin, Turck et al, Nature Electronics 6, 53 (2023)

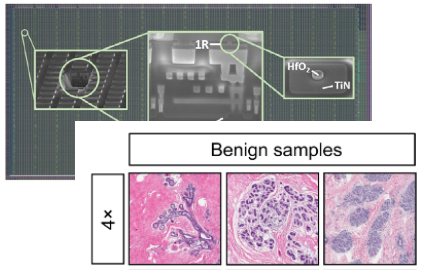
Bayesian In-Memory Computing



- The Memristor-Based Bayesian Machine



- Bayesian Neural Networks with Memristors

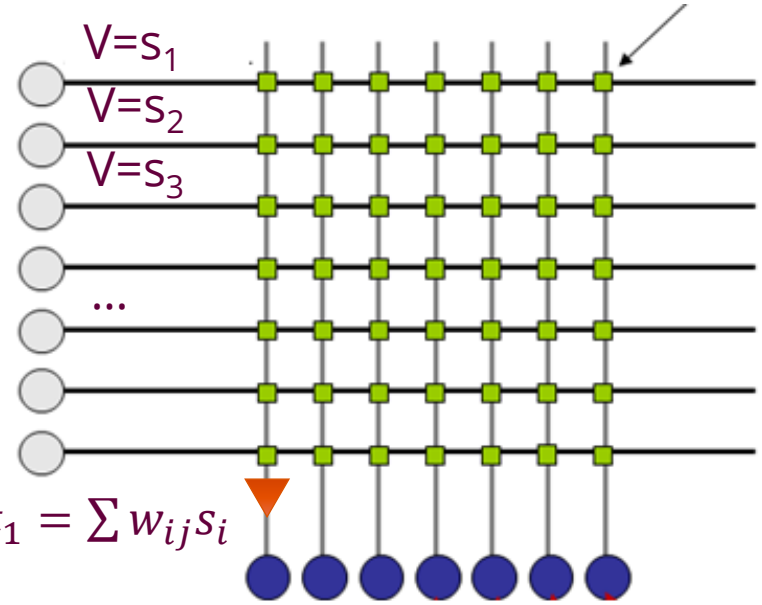


- Bayesian Learning with Memristors

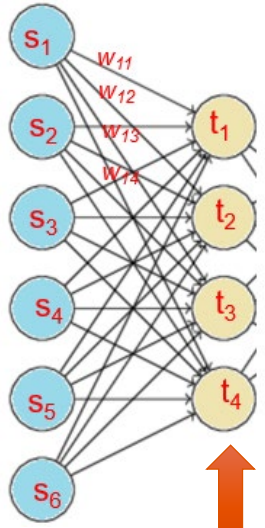
How to Make a Neural Network with Memristors

- A matrix of analog memristors naturally implements a layer of neural network with Kirchhoff laws!

Memristor conductance G = synaptic weight w



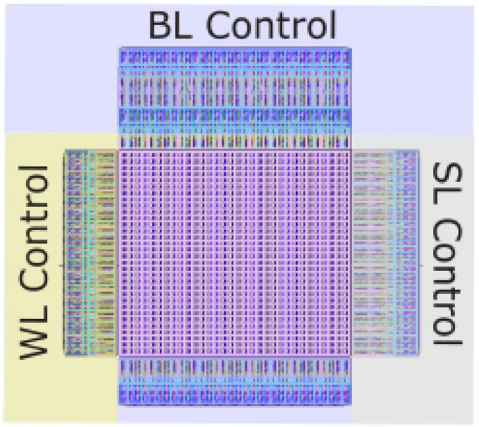
$I_1 = t_1 = \sum w_{ij} s_i$

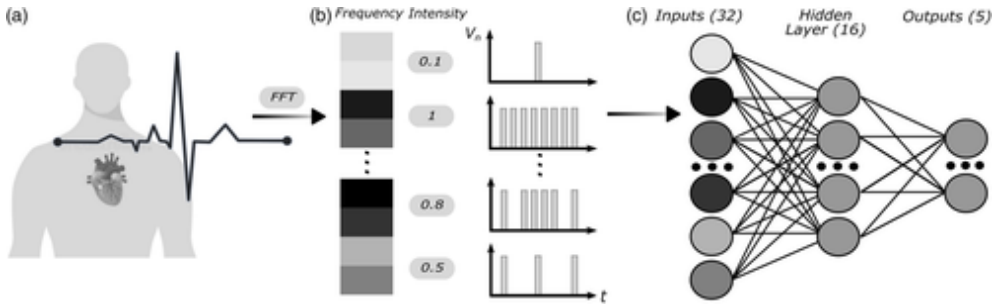


Neuron value

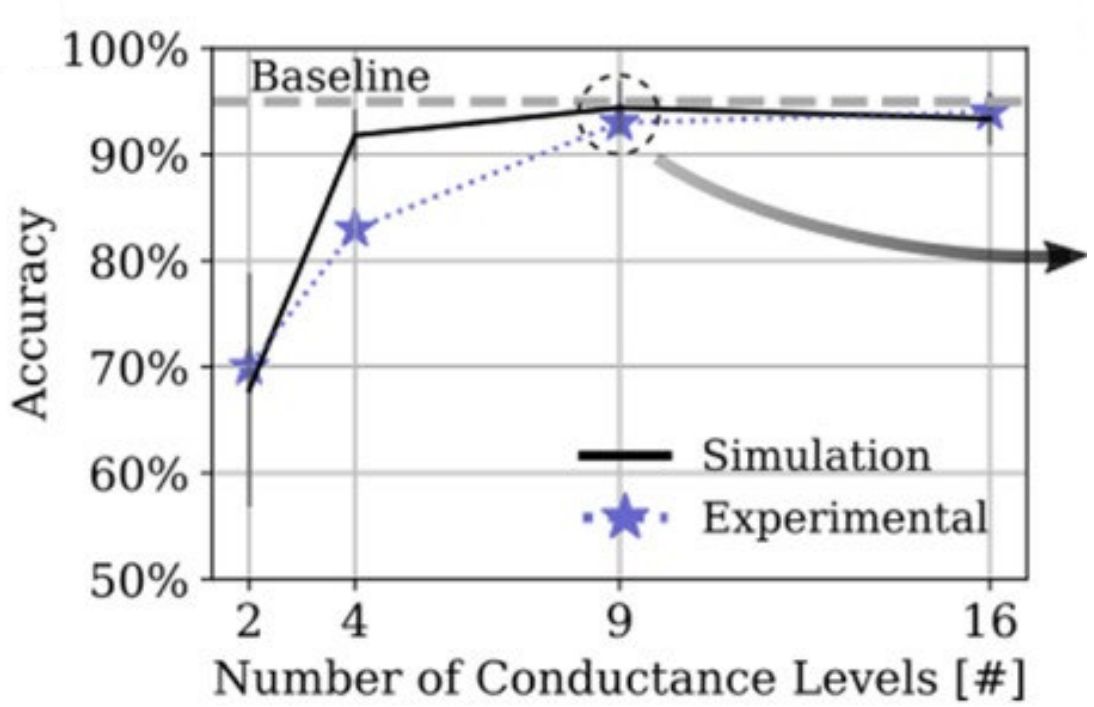
$t_j = f\left(\sum w_{ij} s_i\right)$

Experimental Realization with HfO_x Memristors



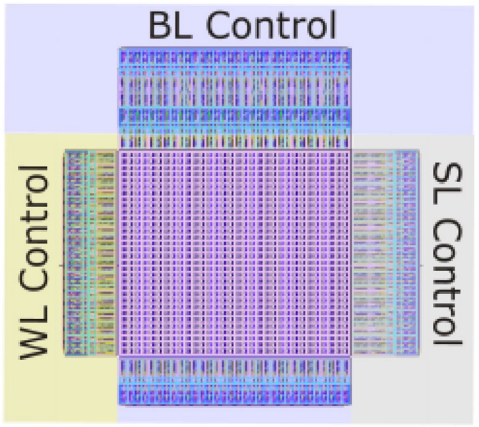


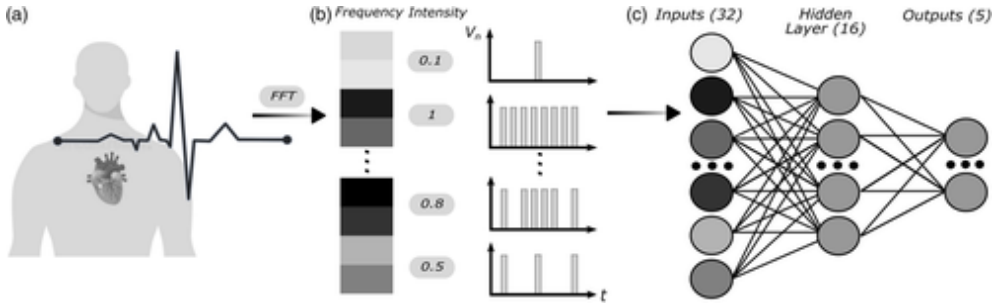
Accuracy on arrhythmia identification



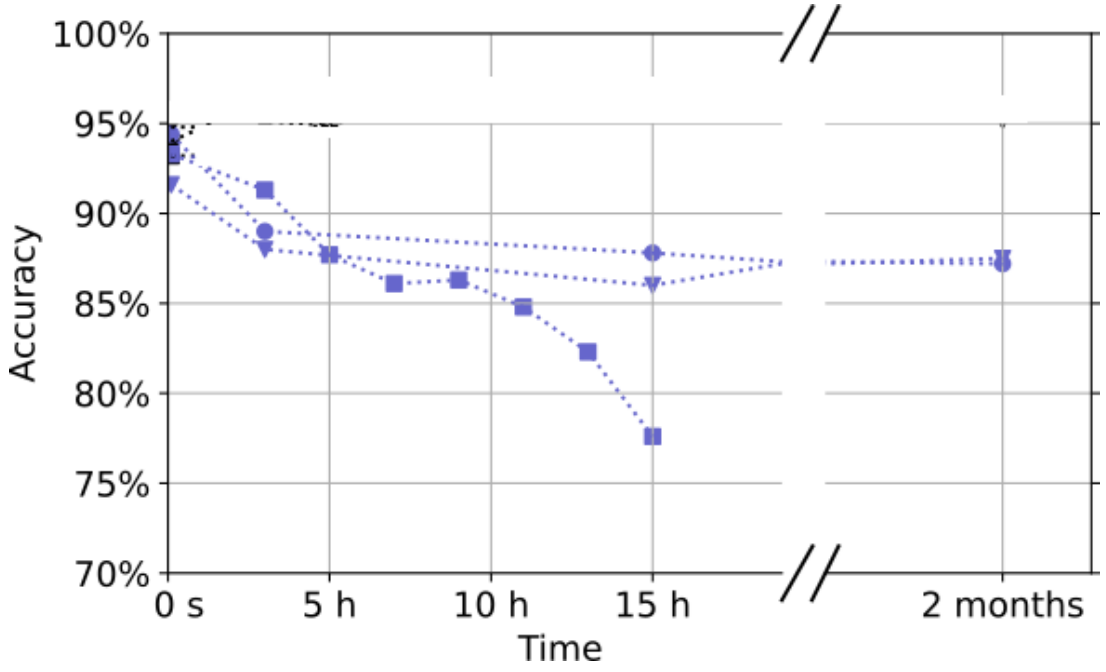
Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

Experimental Realization with HfO_x Memristors





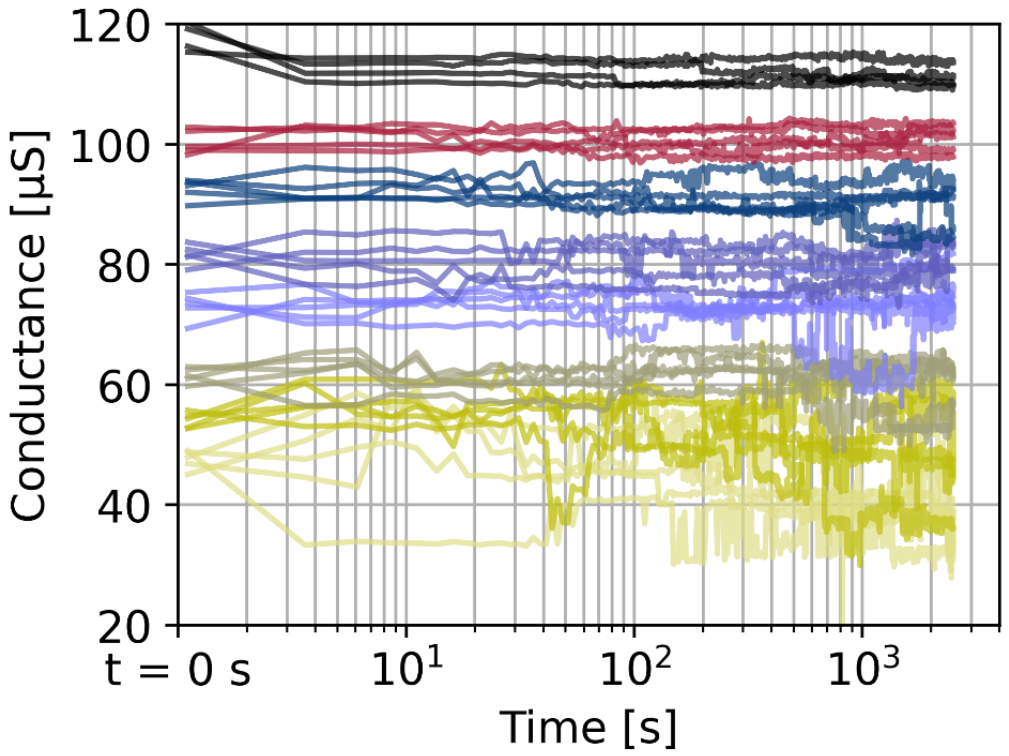
Accuracy on arrhythmia identification



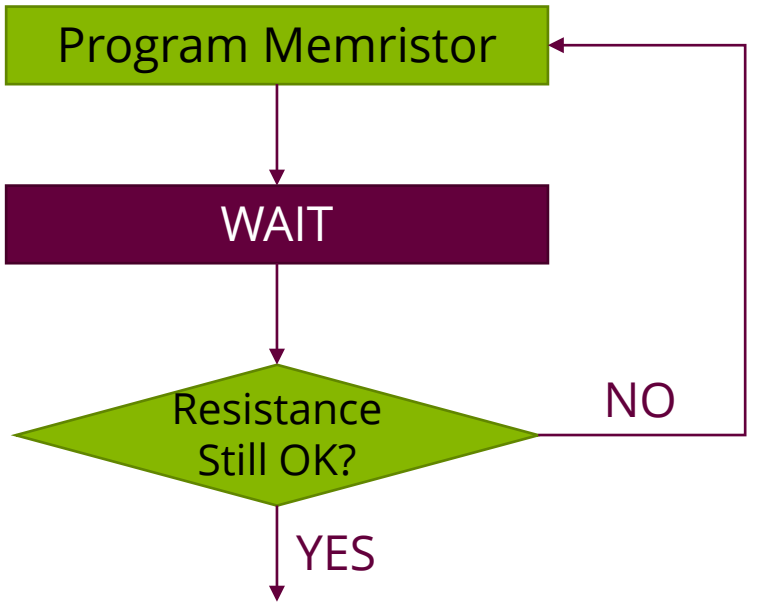
Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

Writing Memristors Reliably

Problem



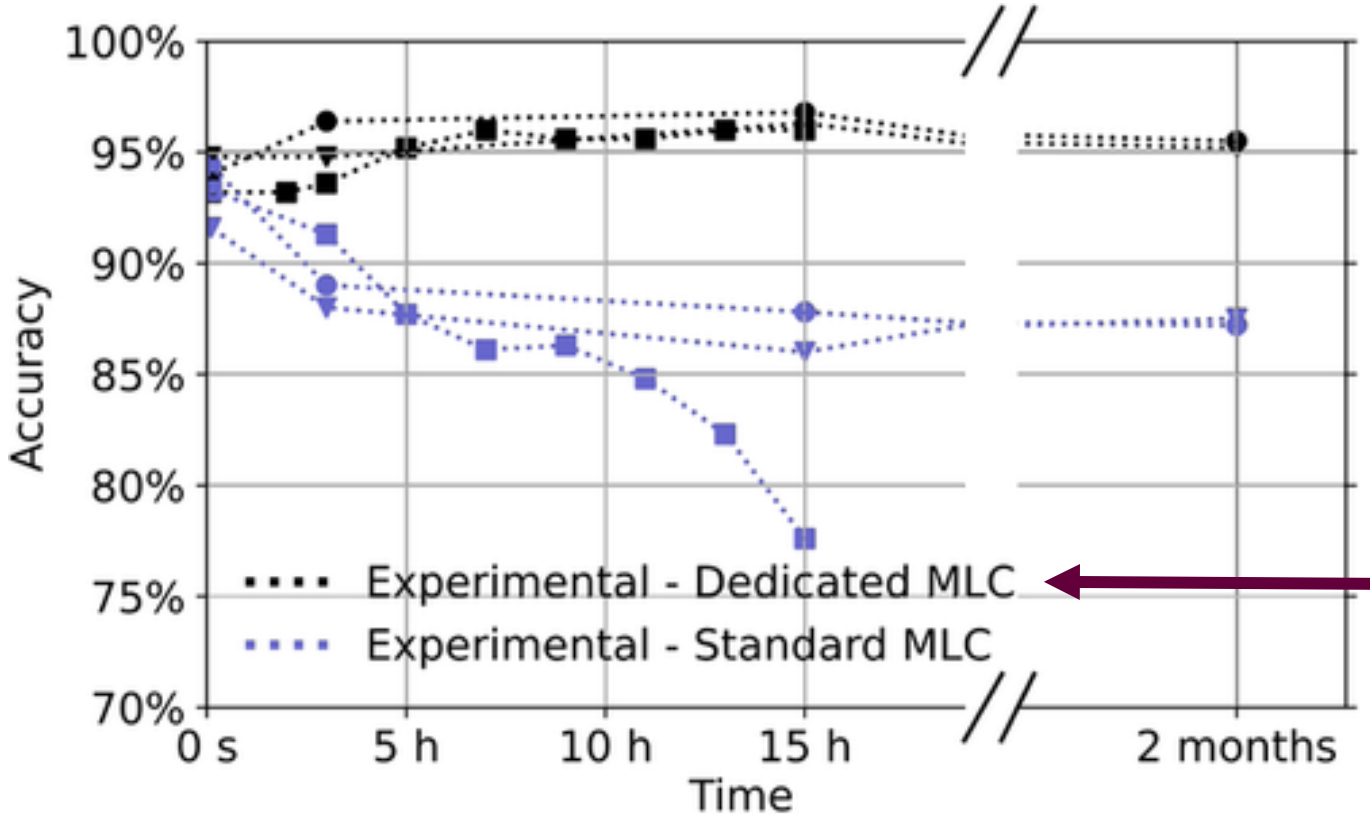
Solutions



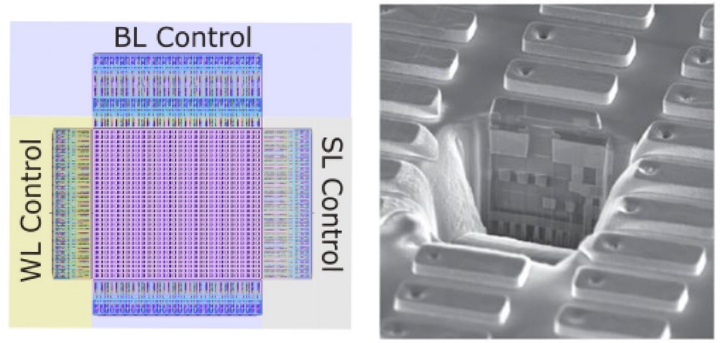
We're good for months!

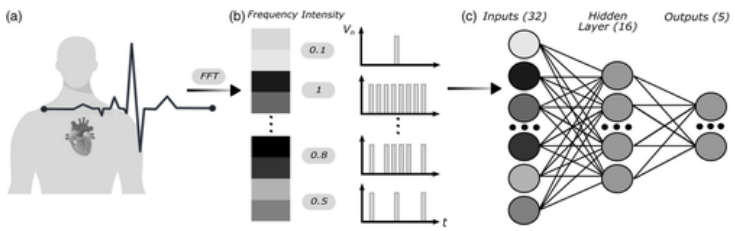
Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

Writing Memristors Reliably



With the Wait

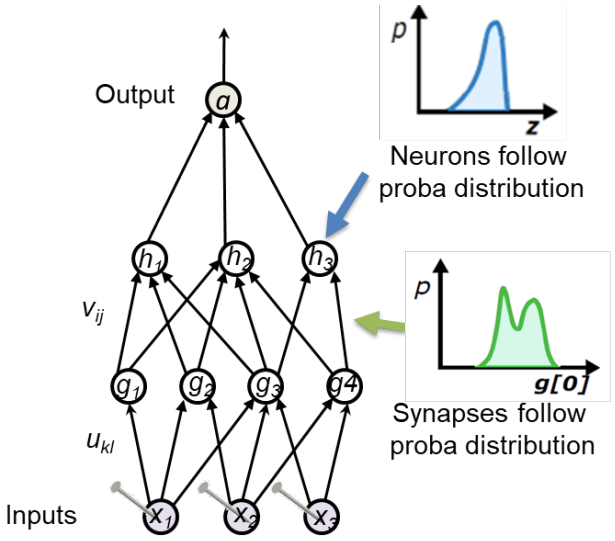




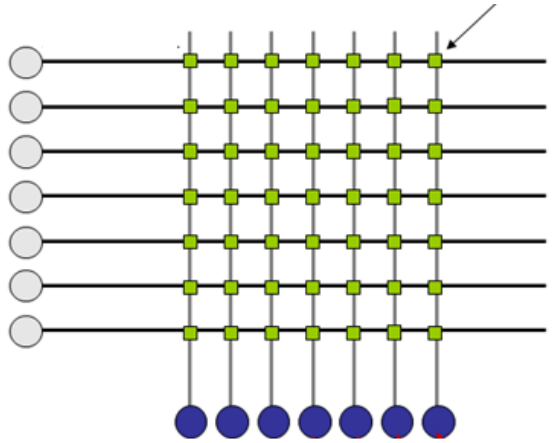
Esmanhotto et al, 2200145, Advanced Intelligent Systems, 2022

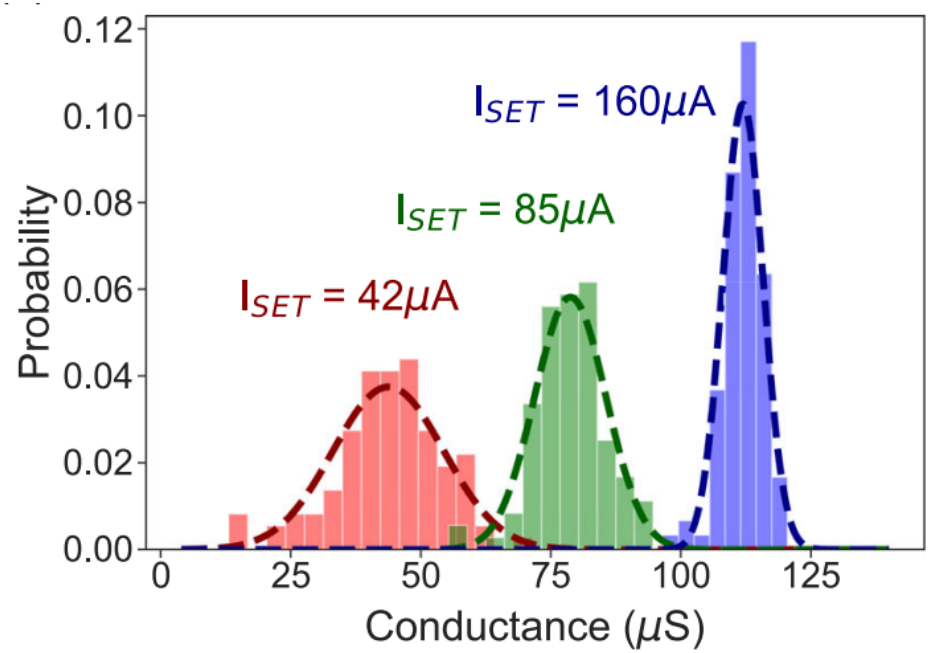
Do Not Fight Memristor Imperfections: They Naturally Produce a Bayesian Neural Network!

In Bayesian models, everything is a random variable that follows specific probability distributions



Memristors actually act as a random variable that follow specific probability distributions!



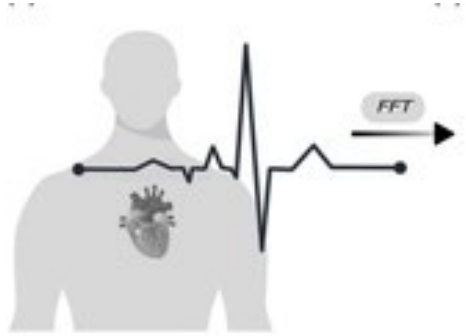


Our concept: Bayesian models can be a "better" way to exploit memristors

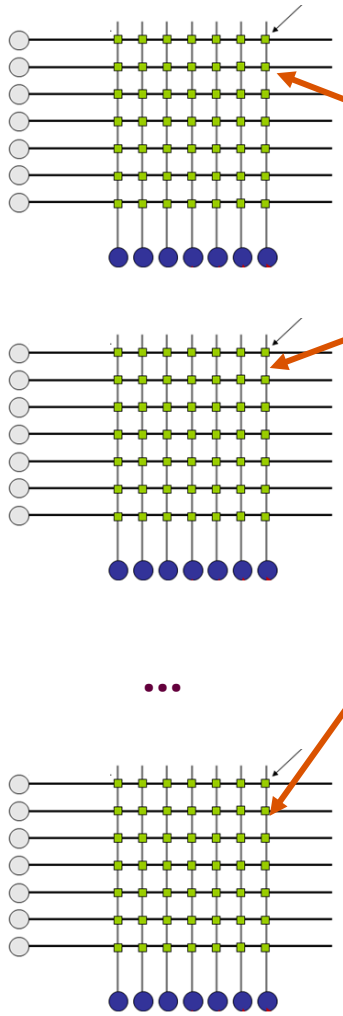
Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

Now Let Us Make a Bayesian Version!

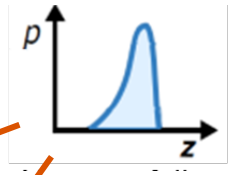
We program 50 memristor arrays, and we apply the same input to them



Input: heartbeat to classify



...



Programmed the same way to represent a probability distribution

We get 50 outputs: Their dispersion tells about the *certainty* of the neural network

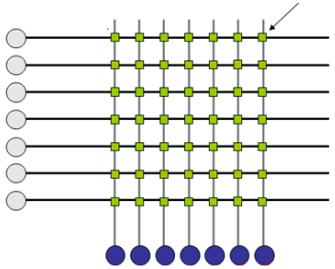
Neural networks needs to have been trained in a special way (variational inference with technological loss)

Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

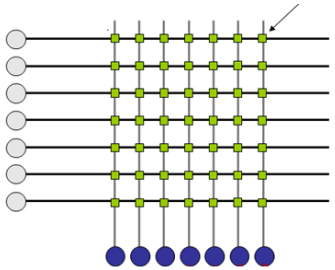
Now Let Us Make a Bayesian Version!

We program 50 memristor arrays, and we apply the same input to them

Input: heartbeat to classify

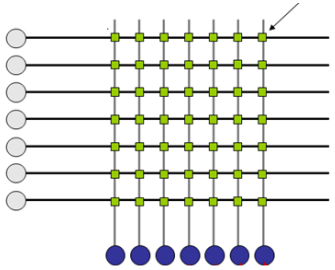


Hesitates between arrythmia type 1 and 2



Hesitates between arrythmia type 1 and 2

...



Hesitates between arrythmia type 1 and 2

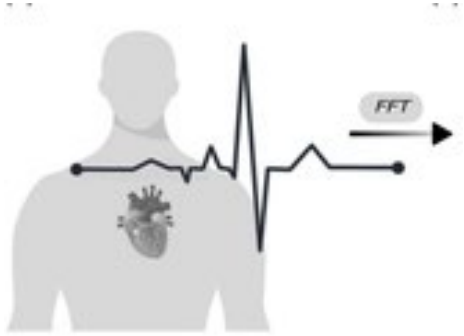
Ambivalence between two types or arrythmias:

ALEATORIC UNCERTAINTY

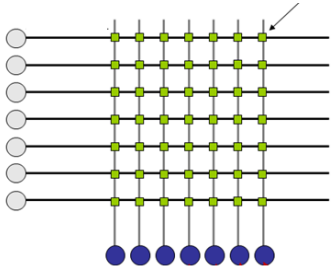
Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

Now Let Us Make a Bayesian Version!

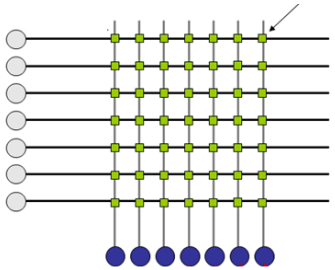
We program 50 memristor arrays, and we apply the same input to them



Input: heartbeat to classify

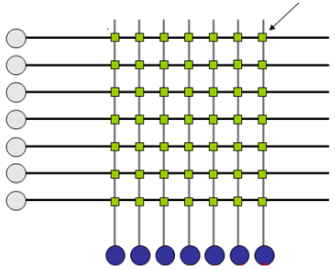


Hesitates between arrythmia type 1 and 2



Hesitates between arrythmia type 3 and 4

...



Hesitates between arrythmia type 1 and 5

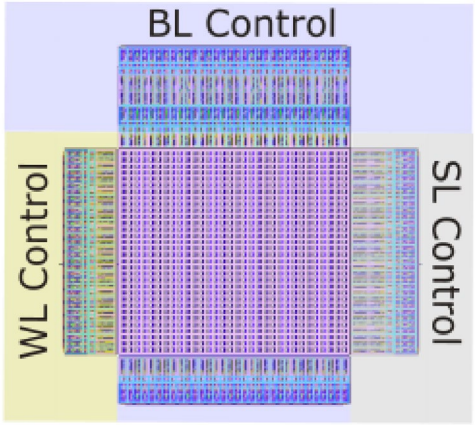
Unknown data:

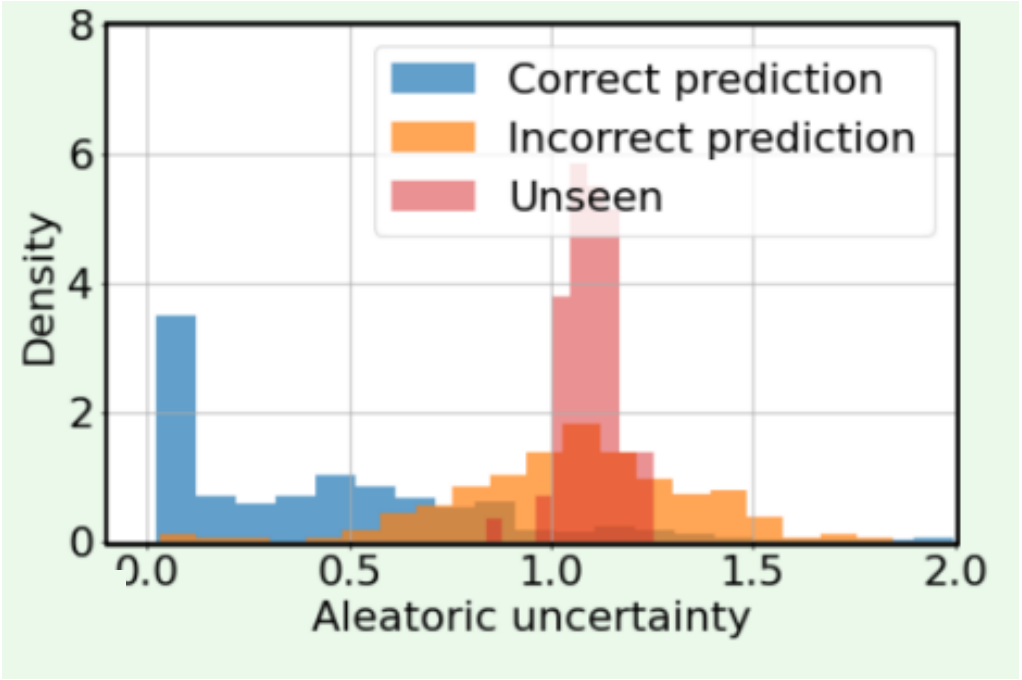
EPISTEMIC UNCERTAINTY

Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

Fully Experimental Arrythmia Recogniton with Uncertainty Evaluation

- **79% raw accuracy (software: 80%)**
- **Unknown types of arrythmia are easily recognized**



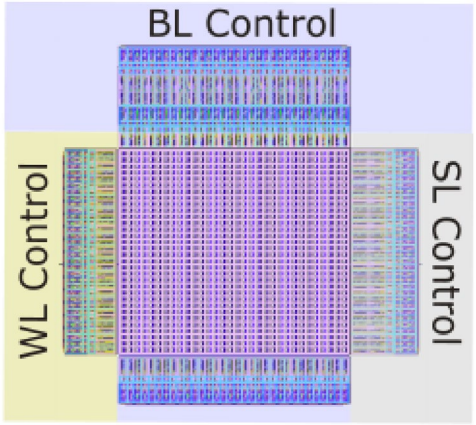


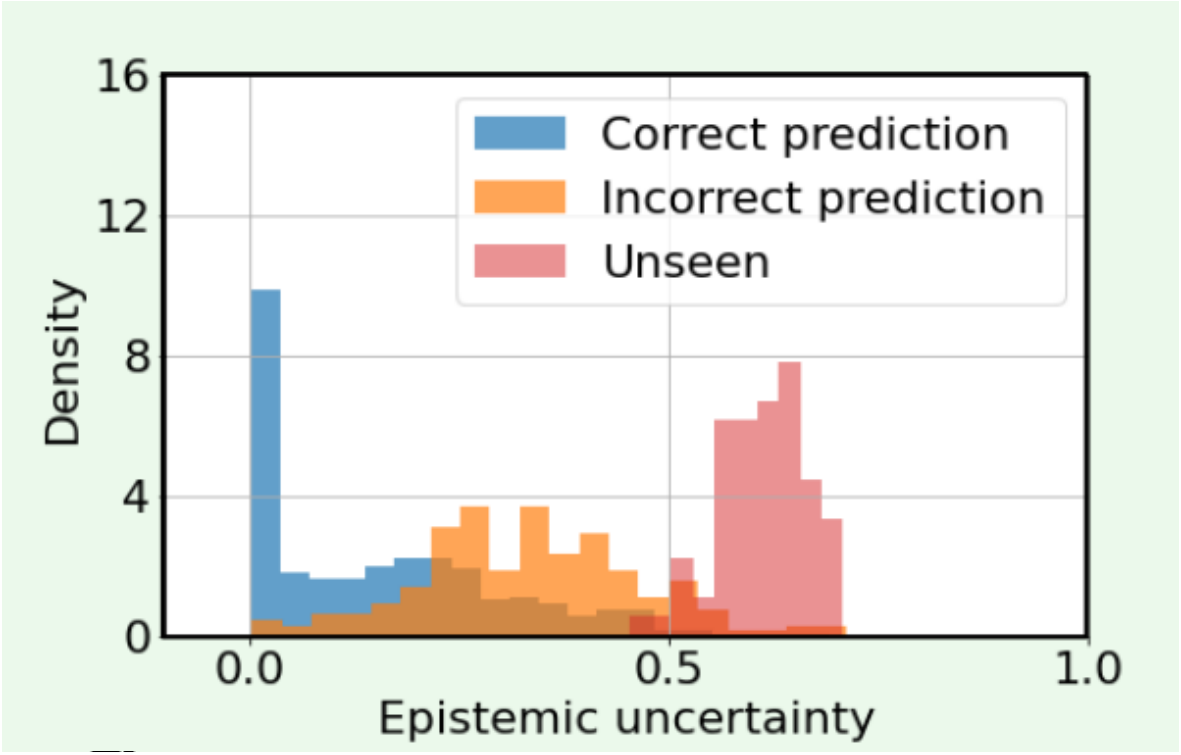
Using 50 memristors neural networks
« Wait and Verify » Programming not needed

Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

Fully Experimental Arrythmia Recogniton with Uncertainty Evaluation

- **79% raw accuracy (software: 80%)**
- **Unknown types of arrythmia are easily recognized**



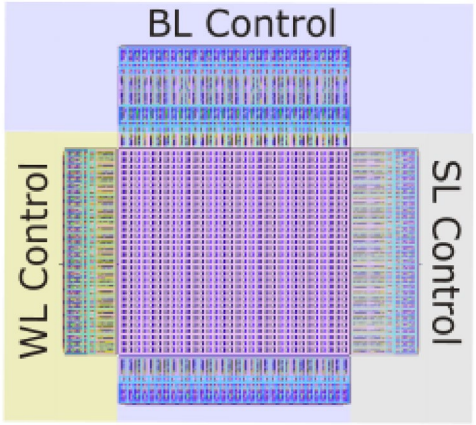


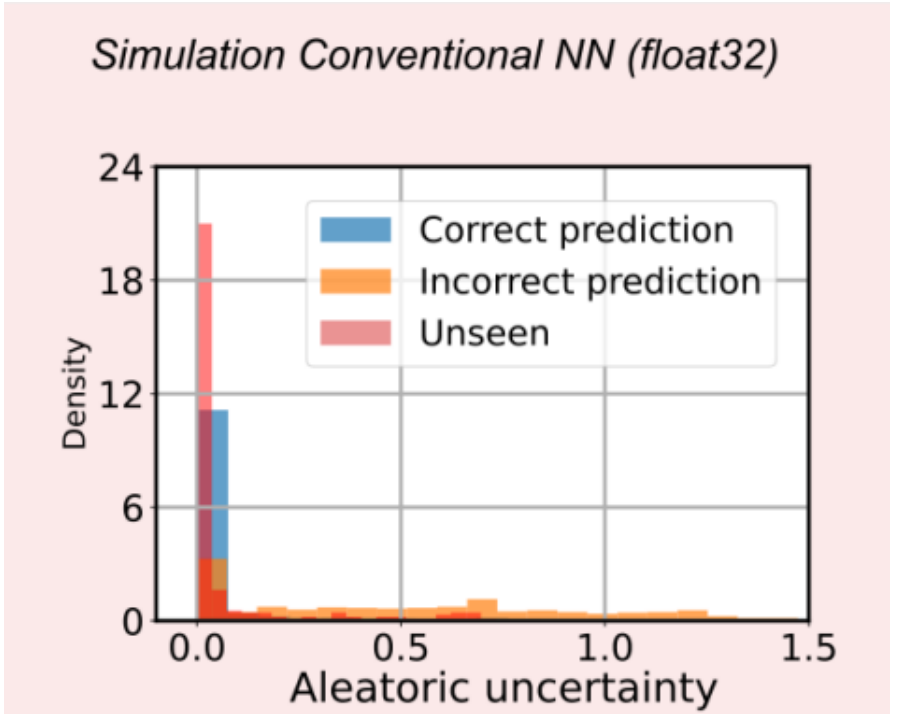
Using 50 memristors neural networks
« Wait and Verify » Programming not needed

Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

Fully Experimental Arrythmia Recogniton with Uncertainty Evaluation

- **79% raw accuracy (software: 80%)**
- **Unknown types of arrythmia are easily recognized**

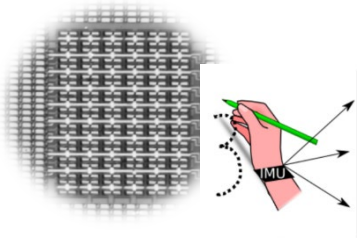




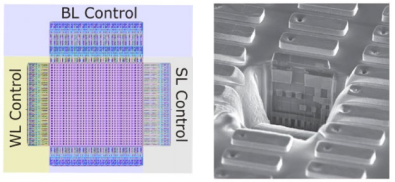
Using 50 memristors neural networks
« Wait and Verify » Programming not needed

Bonnet et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

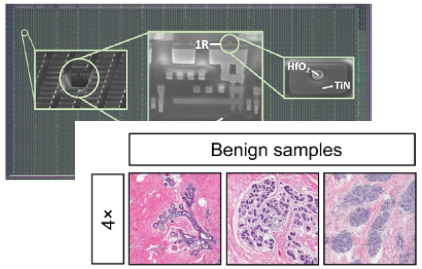
Bayesian In-Memory Computing



- The Memristor-Based Bayesian Machine

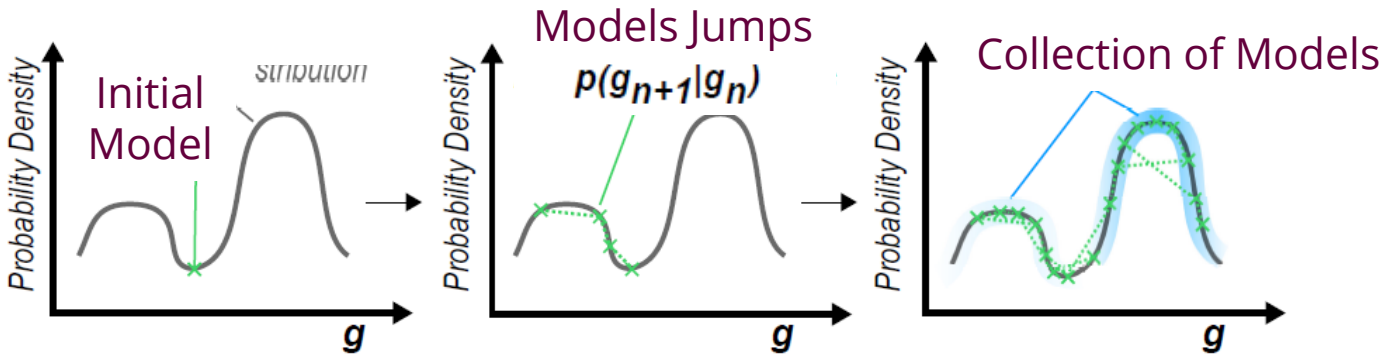


- Bayesian Neural Networks with Memristors

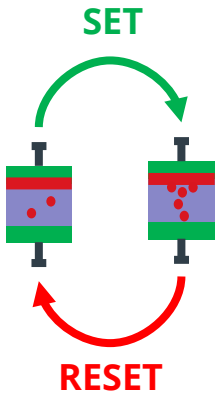


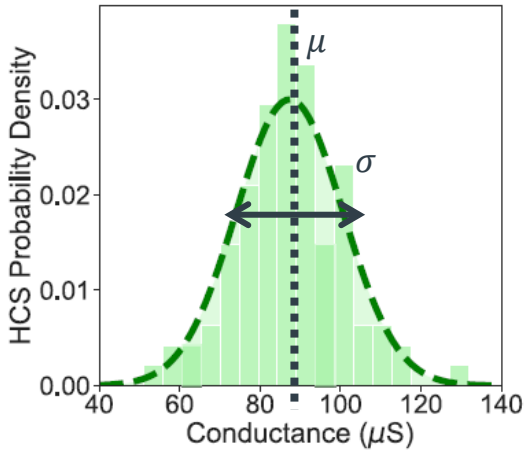
- Bayesian Learning with Memristors

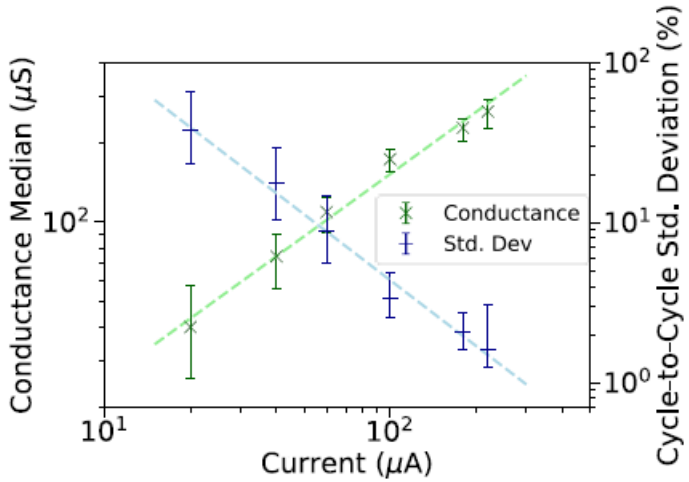
Bayesian Models Can Learn Using Metropolis-Hastings Markov Chain Monte Carlo (MCMC)



The jumps $p(g_{n+1}|g_n)$ can be performed easily using the statistical behavior of memristors!

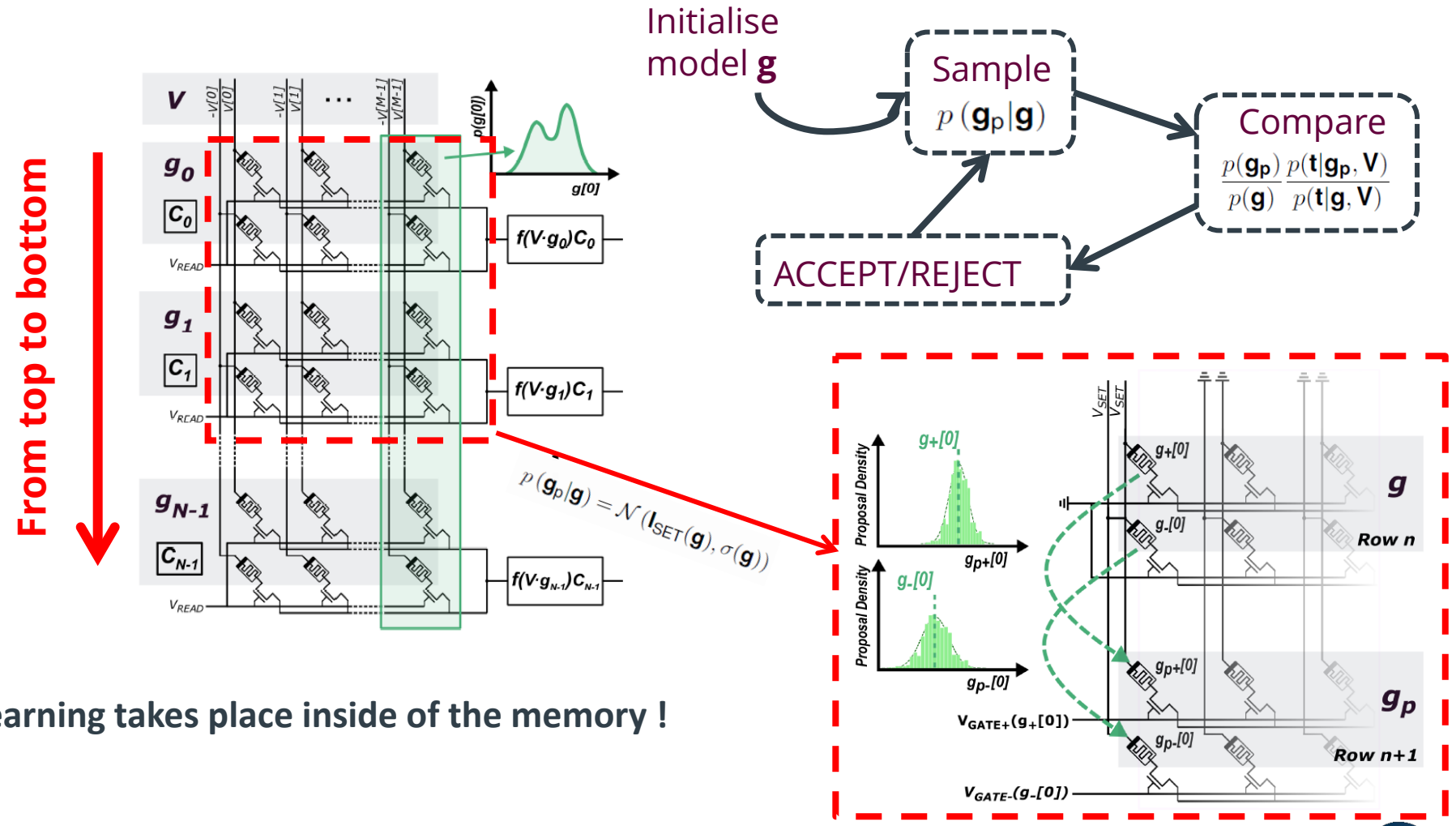






Memristors are ideal for MCMC-based learning!

Memristor-Based MCMC in Practice



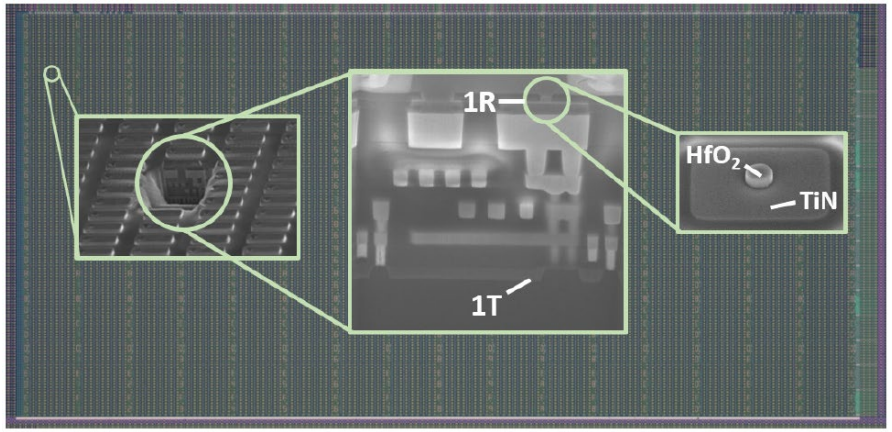
Learning takes place inside of the memory !

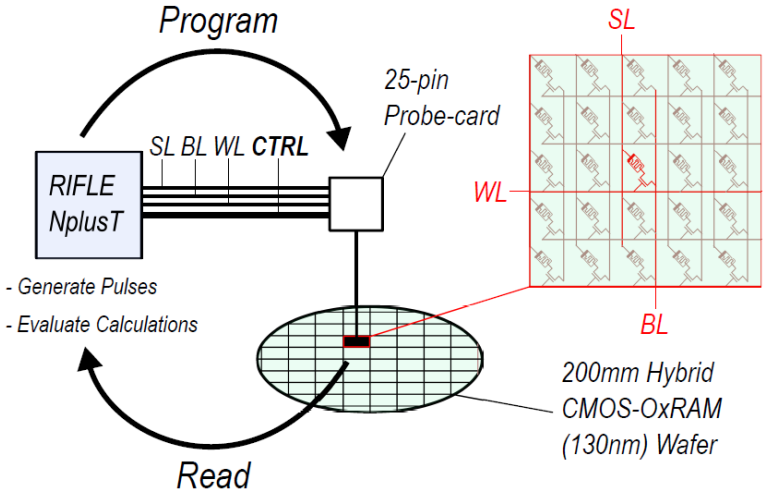
Dalgaty et al, Nature Electronics 4, p. 151, 2021

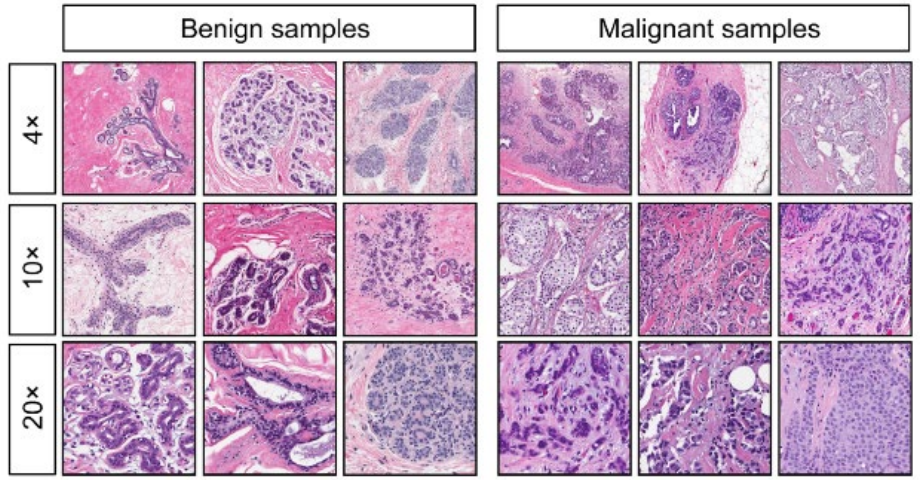
cnrs université PARIS-SACLAY

Memristor-Based MCMC in Practice

Computer-in-the-loop experiment with an array of 16,384 memristors



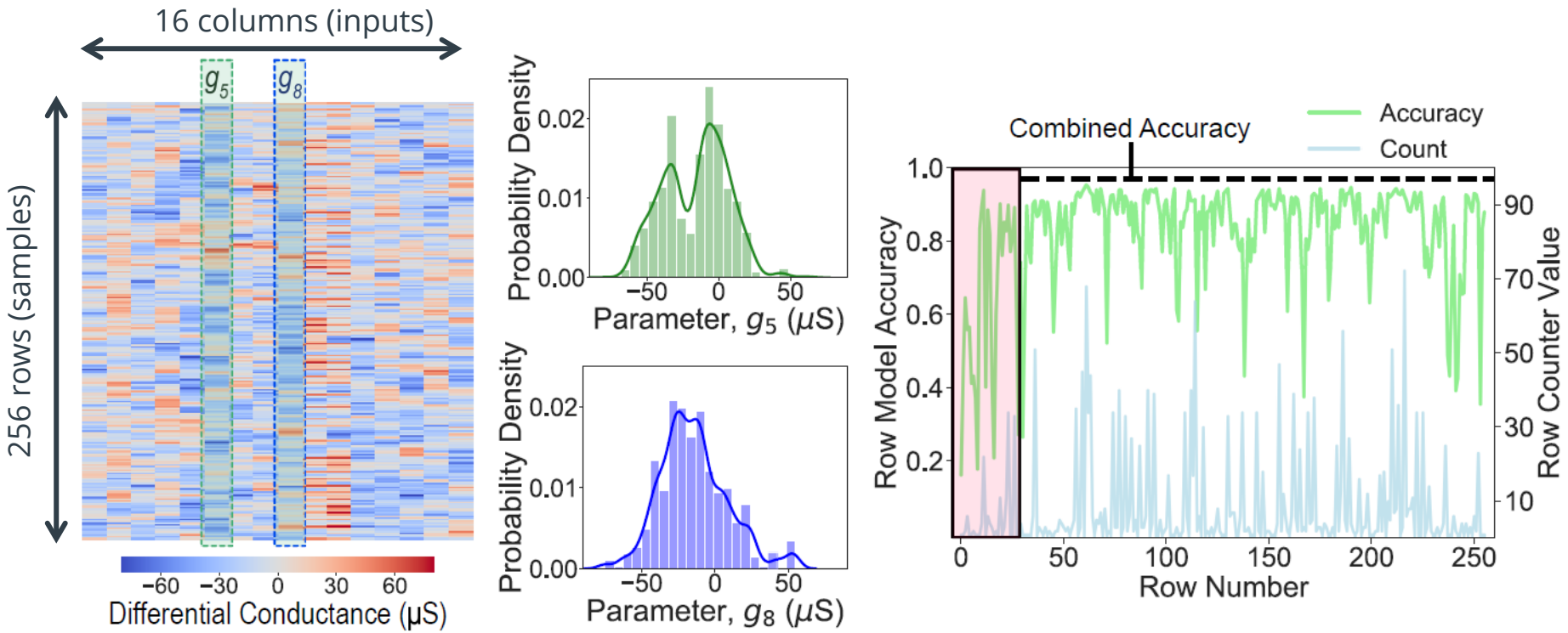




Mangasarian, O. L., et al (1995). *Operations Research*, *43*(4), 570-577.

Dalgaty et al, Nature Electronics 4, p. 151, 2021

Supervised Learning with Memristor-Based MCMC

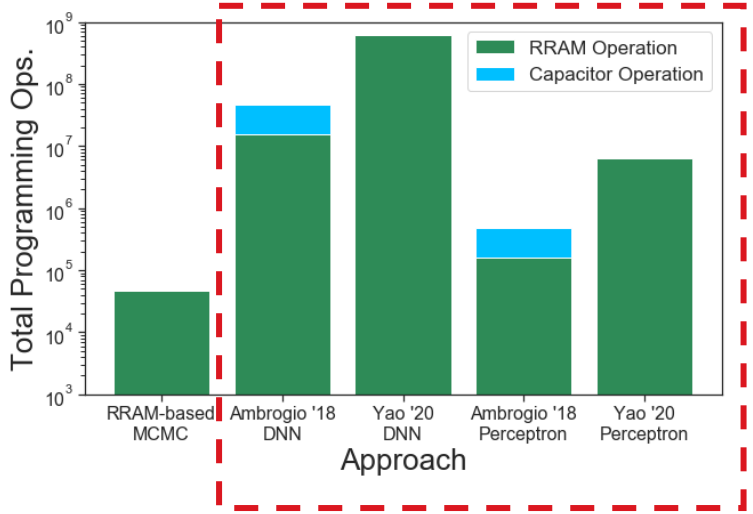


The experimental system was able to detect maligant tissue with 98% accuracy

Dalgaty et al, Nature Electronics 4, p. 151, 2021

cnrs université PARIS-SACLAY

MCMC Learning Is Highly Energy-Efficient for Small Data



Memristor-based backpropagation

Preliminary mixed-signal design results (full model training)

	Step 1 (Model evaluation)	Step 2 (Model acceptance/rejection)	Step 3 (RRAM programming)	Total
Number of repetitions	$500 \times 10 \times 512$	10×512	10×512	
Total energy (130nm)	5.8μJ	$120n$J	1.1μJ	**6.9μJ**
Total energy (28nm)	2.5μJ	$34n$J	1.1μJ	**3.6μJ**

Intel Xeon processor (7nm) implementation of MCMC sampling required **600mJ**

Dalgaty et al, Nature Electronics 4, p. 151, 2021

cnrs université PARIS-SACLAY

Conclusion

- Nanoelectronics enables a wide range of Bayesian concepts
- The probabilistic nature of memristors can be exploited for probabilistic machine learning, i.e., Bayesian models
- This approach can be used for both learning and inference
- Particularly appropriate for “small data”/ high uncertainty situations where wrong answers have dramatic impact, e.g., medical tasks

Acknowledgments

- Kamel-Eddine Harabi
- Clément Turck
- Tifenn Hirtzlin
- Atreya Majumdar
- Marie Drouhin
- Jacques-Olivier Klein

- Elisa Vianello
- Djohan Bonnet
- Thomas Dalgaty
- Tifenn Hirtzlin
- Eduardo Esmanhotto
- Niccolo Castellani
- François Andrieu

- Jacques Droulez
- Pierre Bessière

- Raphaël Laurent

- Jean-Michel Portal
- Marc Bocquet
- Fadi Jebali
- Jean-Pierre Walder
- Eloi Muhr
- Mathieu-Coumba Faye

NEURONIC NANOINFER

Thank you for your attention!

@DamienQuerlioz
damien.querlioz@universite-paris-saclay.fr
https://sites.google.com/site/damienquerlioz/

cnrs université PARIS-SACLAY