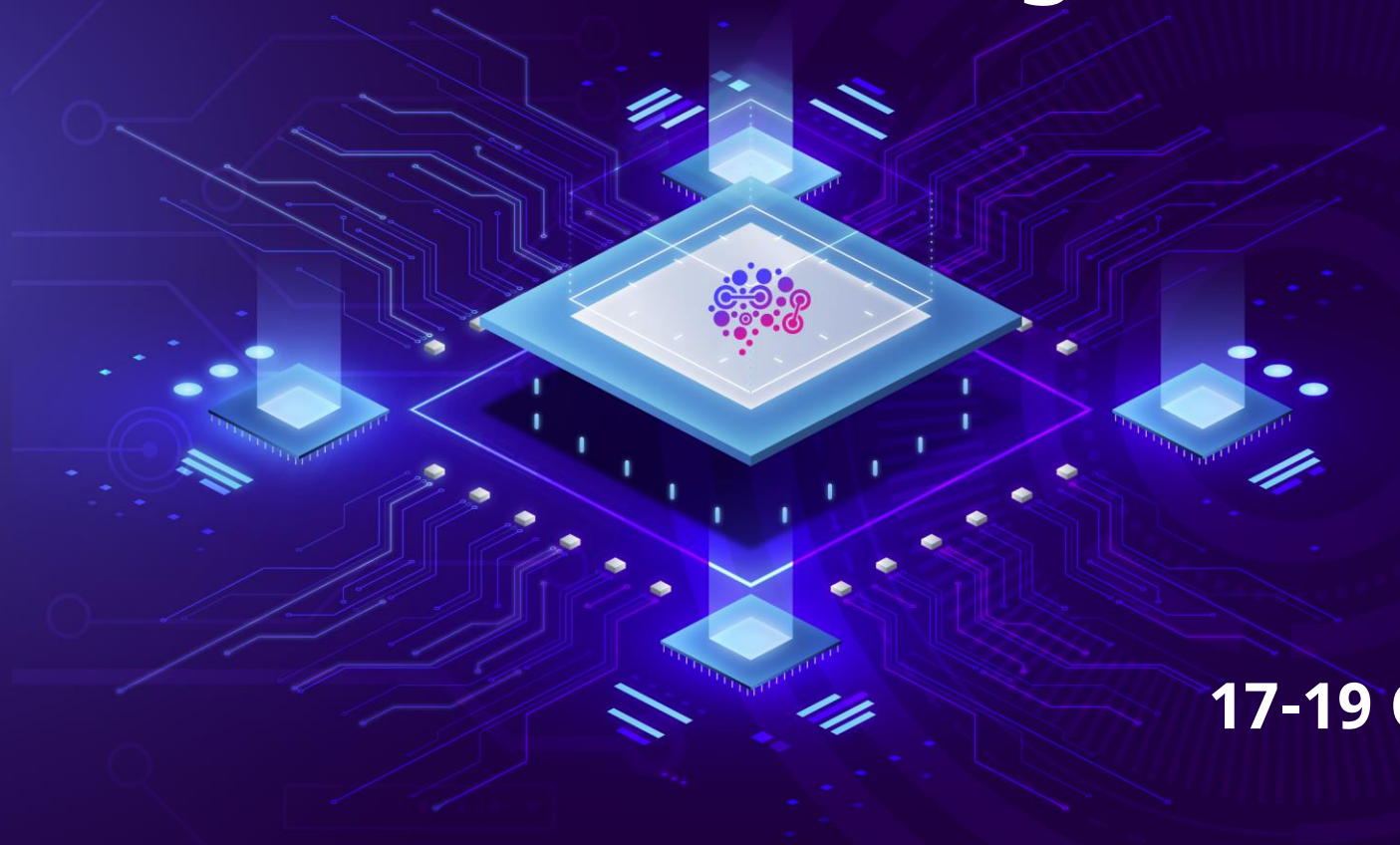
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI

Memristor based Bayesian In-Memory-Computing

Tifenn HIRTZLIN, CEA-LETI, Grenoble, France

17-19 October 2023 Athens, Greece

Copyright © 2023

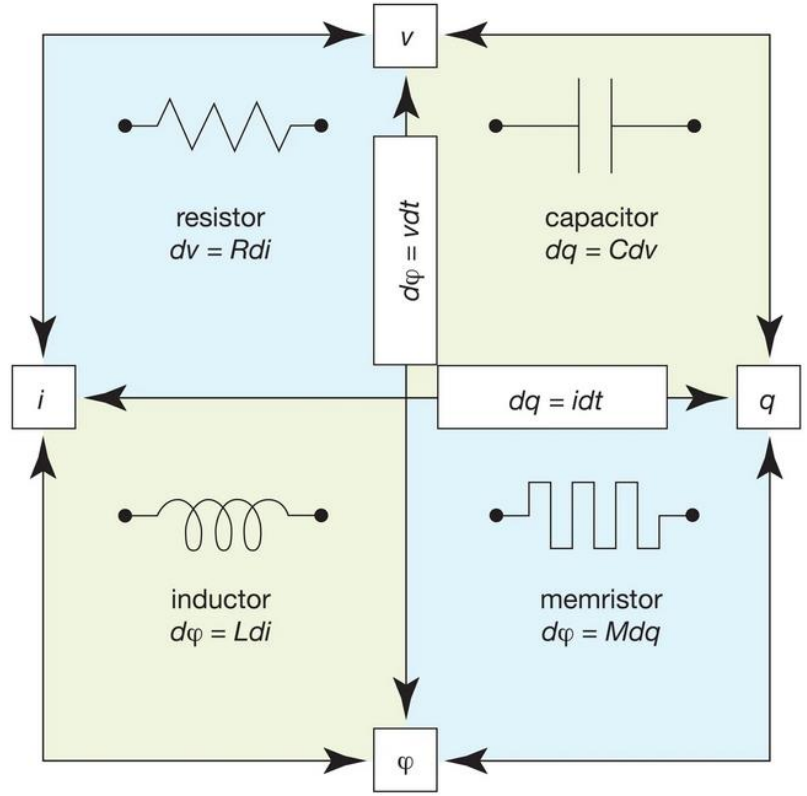
What is a Memristor?

- Introduced in 1971 by **Leon Chua**
- Establish a relationship between charge (q) and magnetic flux (ϕ)
- Propose a new two-terminal circuit element called a **memristor**, (for "memory" and "resistor“)

$$v(t) = M(q(t))i(t)$$

$$M = \frac{\partial \varphi}{\partial q}$$

$$v = \frac{\partial \varphi}{\partial t},$$

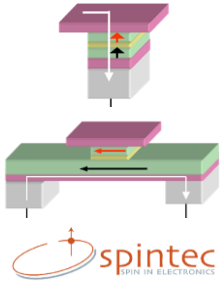
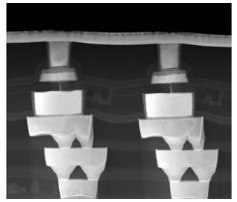
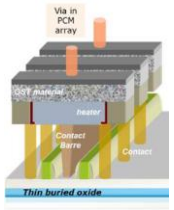


A true memristor is just a matter of math.
For device engineers, whether a true memristor exist or not is not very important

Copyright © 2023

Revolution in the field of electronic memories

Modern Definition uses physics behavior name

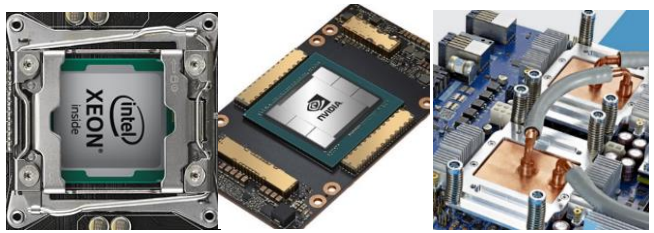
MRAM	OxRAM	PCM	FeRAM
			
	1T1R 28nm	1T1R 28nm	1T1C 28nm

- Many variations (phase change, magnetoresistive, conductive bridge, ferro…)
- Fast, Analog, small footprint, **non-volatile** memory that can be **embedded at the core of CMOS**
- Industrial product (Samsung, TSMC, Intel, IBM, ST Microelectronics...)

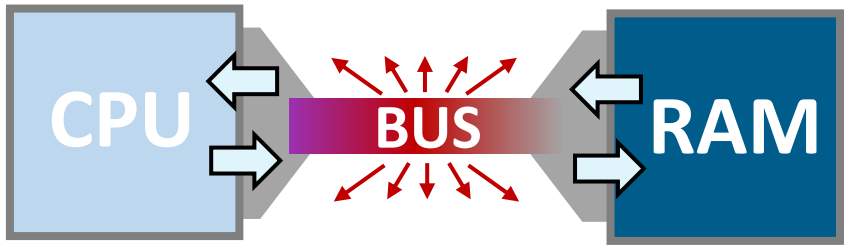
New features available in foundry Design Kits

Copyright © 2023

New features that question current hardware architecture



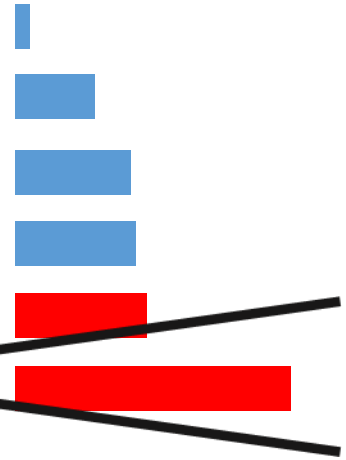
Intel Xeon CPU GPU NVIDIA TPU 3 Google



- **Modern AI algorithms require a lot of data exchange between memory and computation**
- **Memristors offer a lot of opportunities for hardware design engineers:**
  - → instant on/off
  - → integration with logic function
  - → reconfigurabiity
  - → …

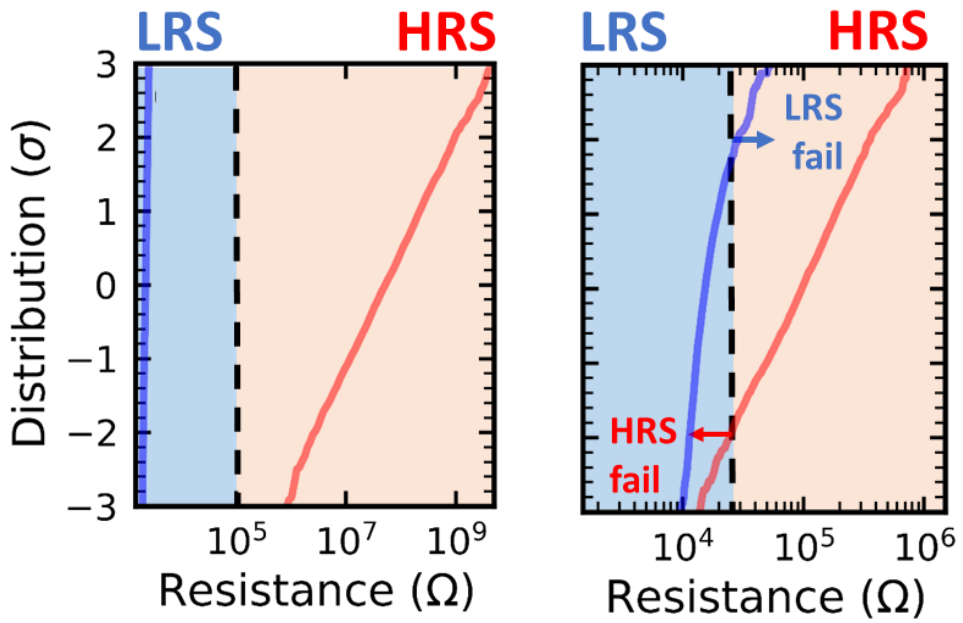
Future of Electronics : reevaluating the role of memory

Operation	Energy (pJ)
32b Addition	0.1
32b FP Addition	0.9
32b Multiplication	3.1
32b FP Multiplication	3.7
32b SRAM Read	5
32b DRAM Read	640



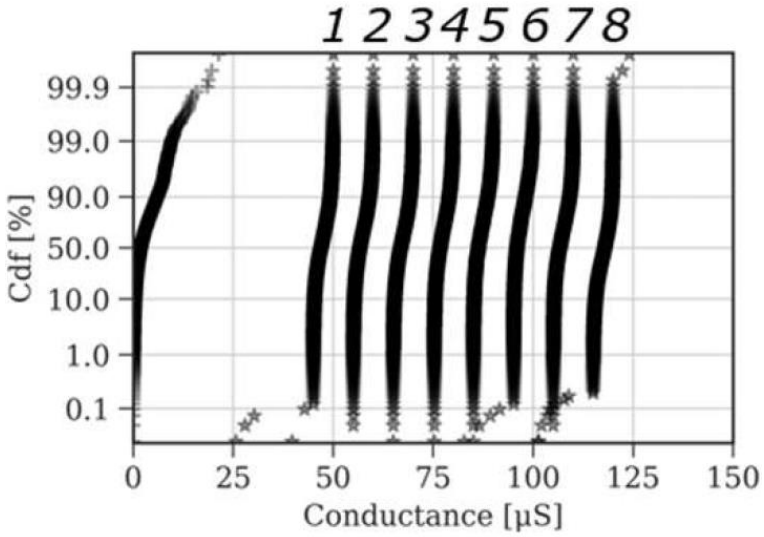
Copyright © 2023

The Challenge with Memristors : Device imperfection



M. Bocquet, T. Hirtzlin* et al. IEDM 2018*

- Scaling to advanced CMOS nodes / using relaxing programming conditions increase the bit error rate



- Possibility to reach multi-level/multiple distributions but requires complex programming procedure

The probabilistic nature of memristors can be exploited to build probabilistic hardware

Copyright © 2023

Probabilistic Hardware for Edge AI

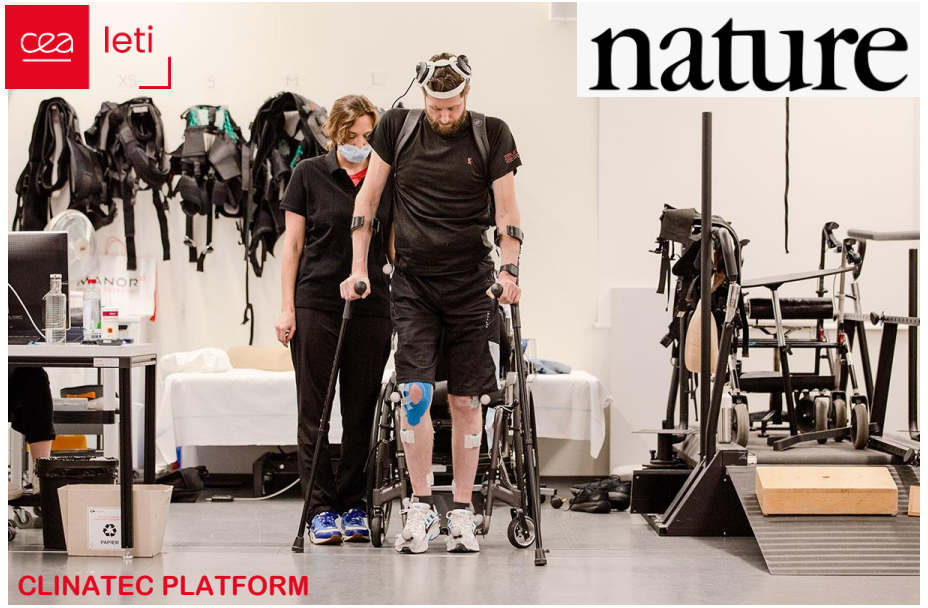
Adapted to new electronic imperative for Safety-Critical Applications

Medical Application

- Neurological symptom detection & control
- Motor decoding Brain Machine Interfaces

Industrial Application

- Monitoring early drifts/faults to avoid accidents
- Taking accurate decision based on uncertain information
- Continuous learning from ongoing data streams
- …



"Walking naturally after spinal cord injury using a brain–spine interface", H. Lorach et al. Nature 2023

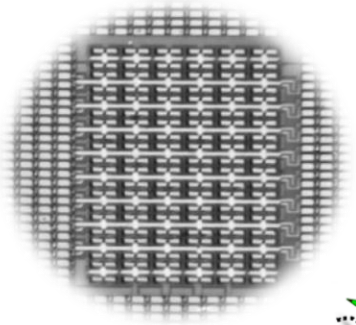
Requirements:

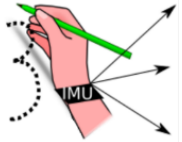
- Fast & Effective action
- Privacy assured
- Secured
- Off-line mobility
- Low energy consumption

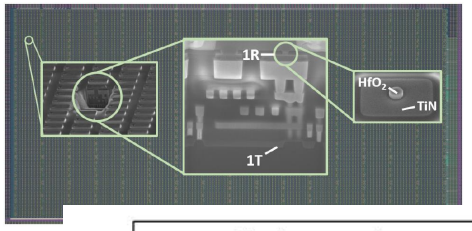
Memristor based Bayesian In-Memory Computing

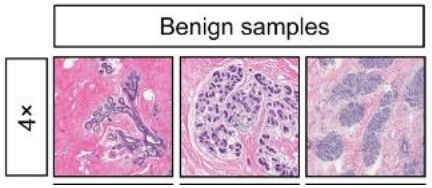
Copyright © 2023

Memristor based Bayesian In Memory Computing









- **The Memristor-Based Bayesian e**

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

- Bayesian Neural Networks with Memristors

  I. Dalgaty et al. (Nature Electronics 2021)

  D. Bonnet, T. Hirtzlin et al, (2023)

Copyright © 2023

Bayesian Reasoning

$$p(Disease \mid Observations) \propto \underbrace{p(Observations \mid Disease)}_{\text{Likelihoods}} \times \underbrace{p(Disease)}_{\text{Prior}}$$

Constructed with expert knowledge+Data

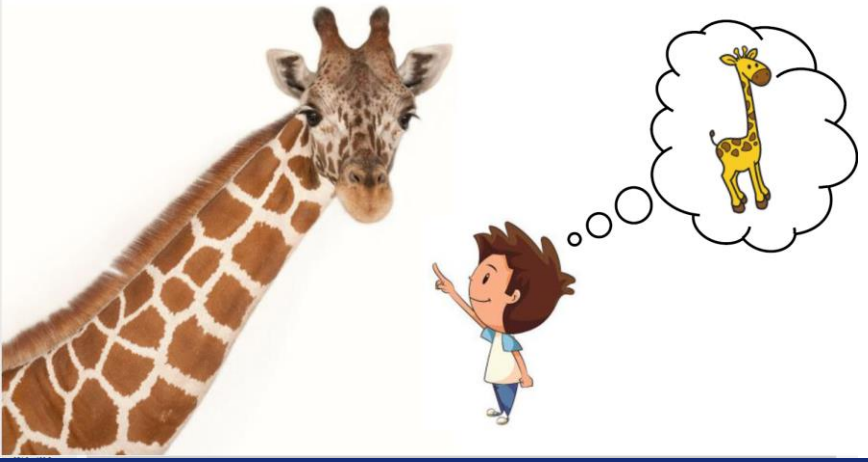
Hard to Compute

Thomas Bayes

Bayesian reasoning

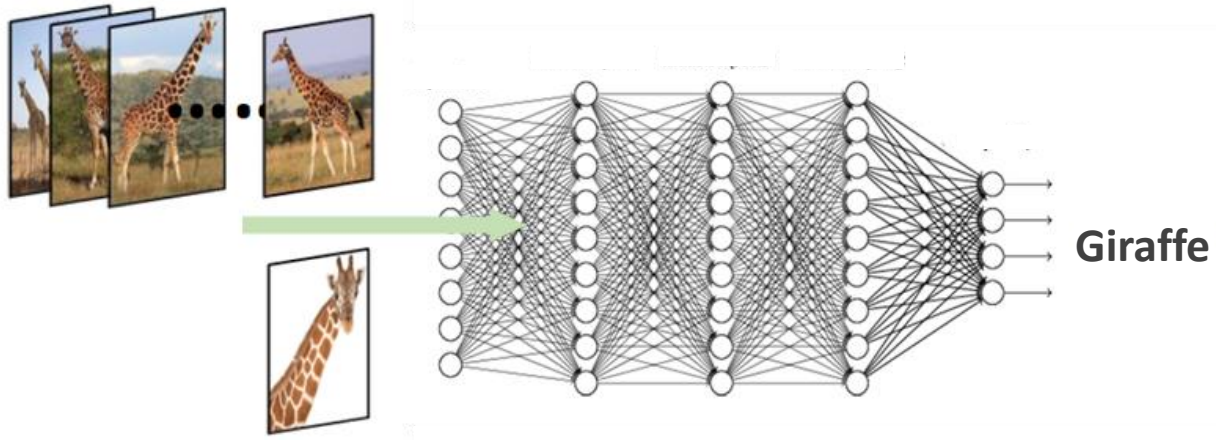
- Similar to human learning methods

(Cognitive neuroscience: Learning in young children)



Deep Learning

- Requires a huge amount of data
- Overconfident output

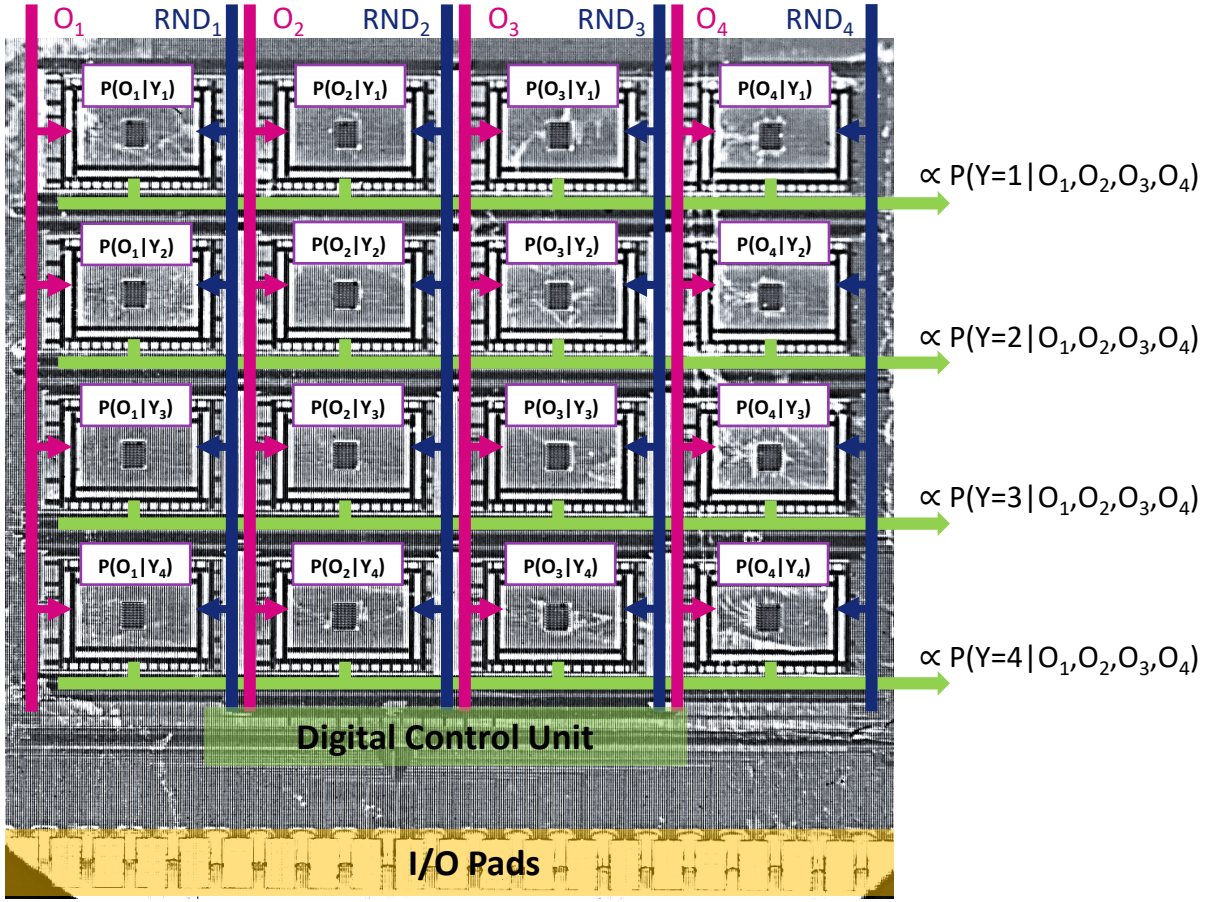


Copyright © 2023

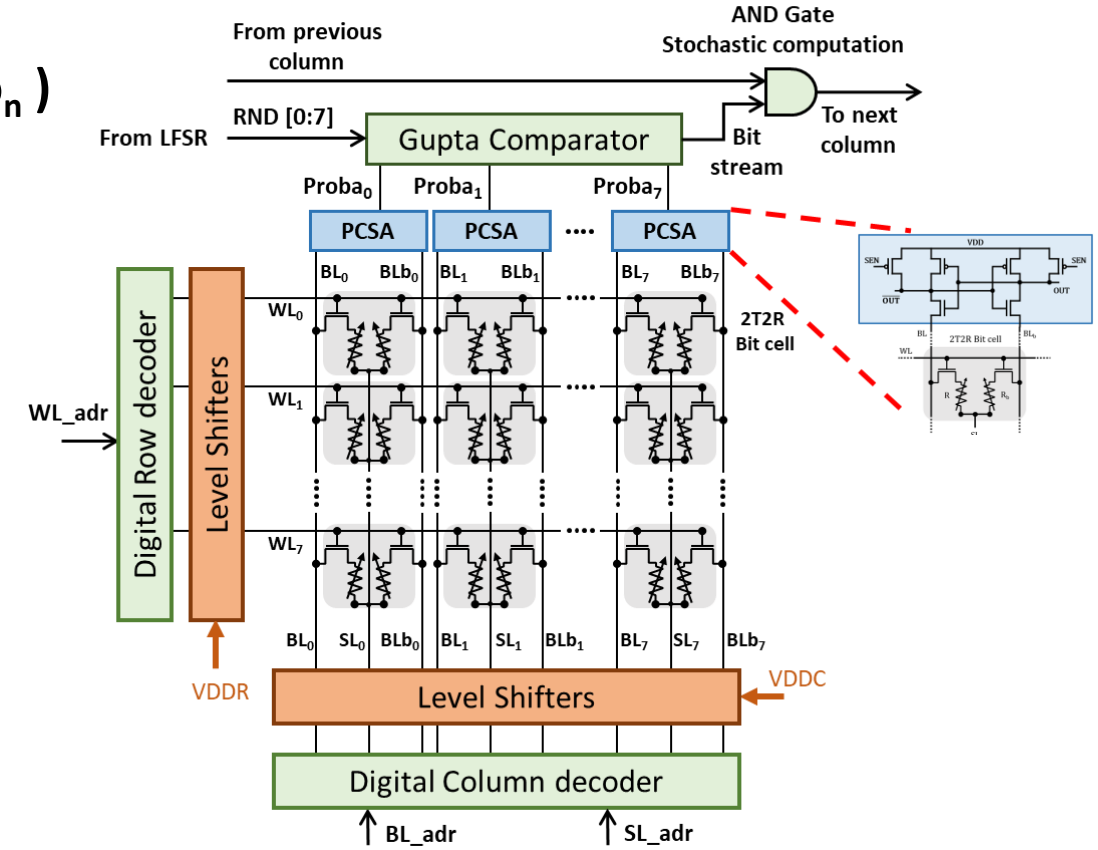
The Memristor-Based Bayesian Machine

Computed equation:

$$P(Y) \times P(O_1|E) \times P(O_2|E) \times P(O_3|E) \times P(O_4|E) \alpha\ P(E|\omega_1, \omega_2 \ldots \omega_n)$$



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



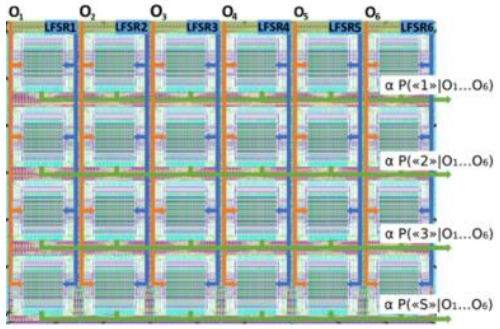
$$P(Y=4|O_1,O_2,O_3,O_4) \propto P(O_1|Y=4) \times \cdots \times P(O_4|Y=4)$$

A very simple replicated memory cut with near memory computing

Copyright © 2023

Uncertainty Evaluation & Energy consumption

Signatures of an uncertain Bayesian machine

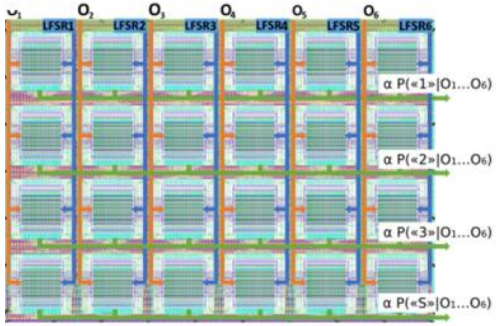


0 ones
0 ones
0 ones
3 ones

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

EPISTEMIC UNCERTAINTY

OR



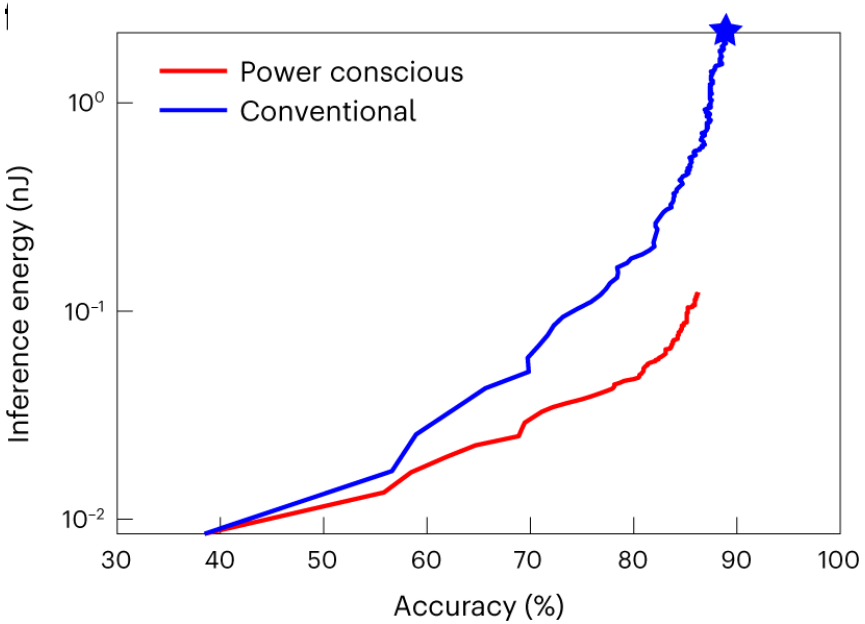
127 ones
100 ones
93 ones
3 ones

Ambivalent situation:
Some outputs produce a balanced number of ones

ALEATORIC UNCERTAINTY

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

Energy consumption

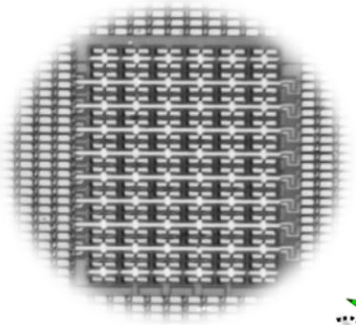


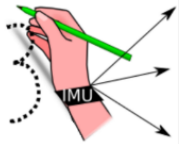
Same task, STM32 microcontroller unit: 2µJ

Two reasons for energy efficiency:
- Near-memory computing
- Stochastic computing

Copyright © 2023

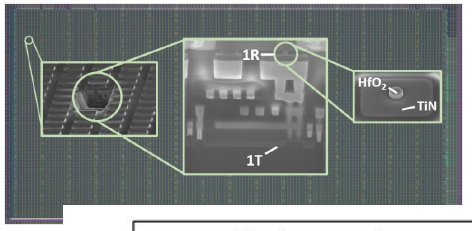
Memristor based Bayesian In Memory Computing

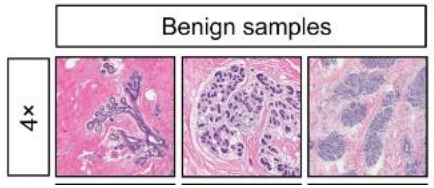




- **The Memristor-Based Bayesian Machine**

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





- Bayesian Neural Networks with Memristors

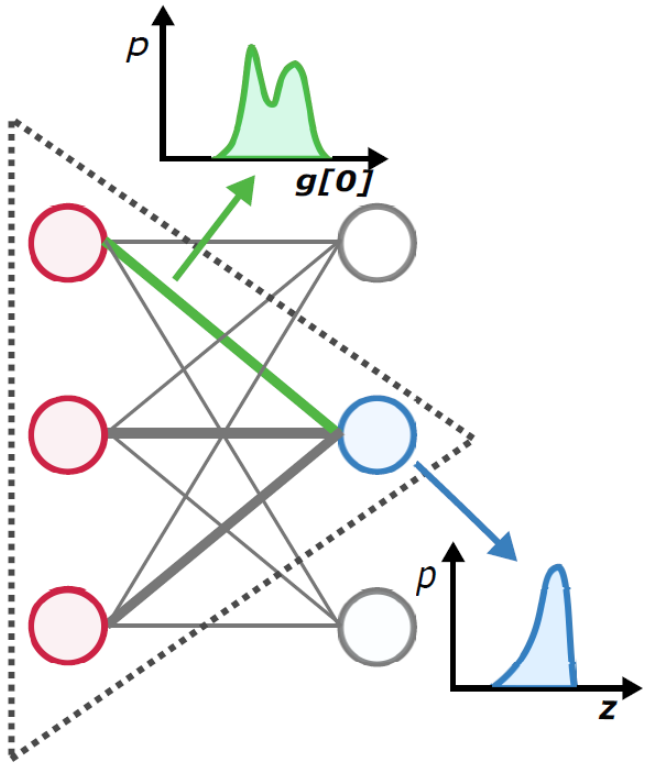
  M. Dalgaty et al. (Nature Electronics 2021)

  D. Bonnet, T. Hirtzlin et al, (2023)

Copyright © 2023

Bayesian Neural Networks

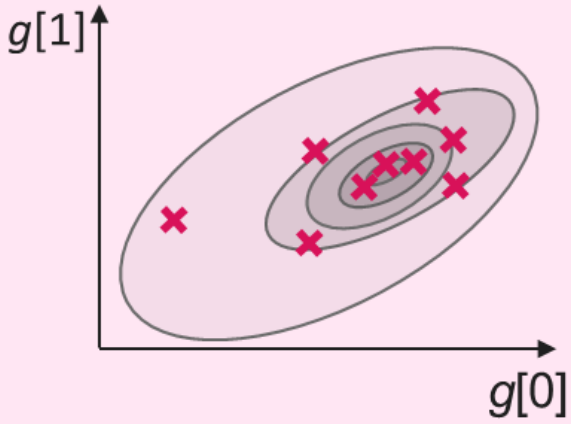
Bayesian Neural Network parameters / activations are probability distributions describing uncertainty



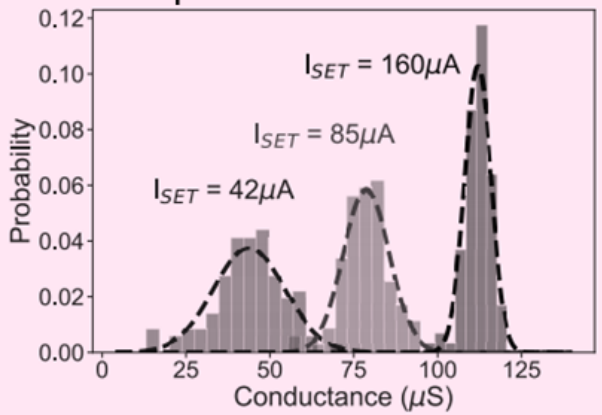
T. Dalgaty et al. (Nature Electronics 2021)

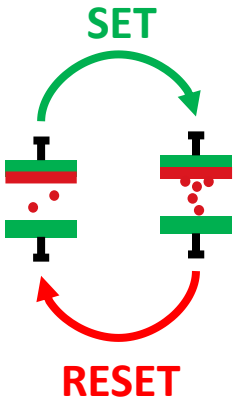
Bayesian approach : Sample over parameters

sample the posterior distribution



Sample the conductance





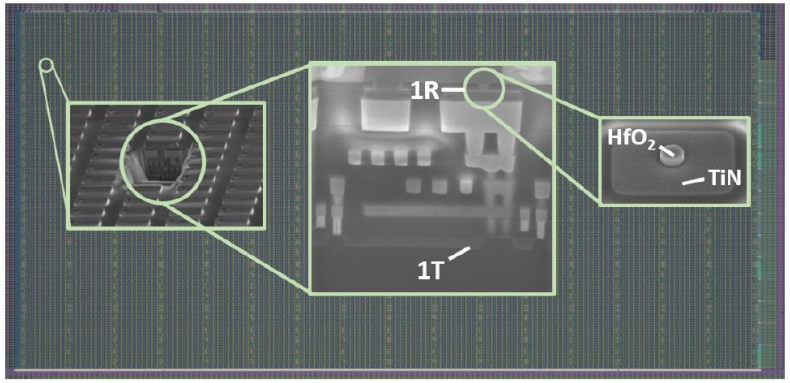
Markov chain Monte Carlo (MCMC) sampling

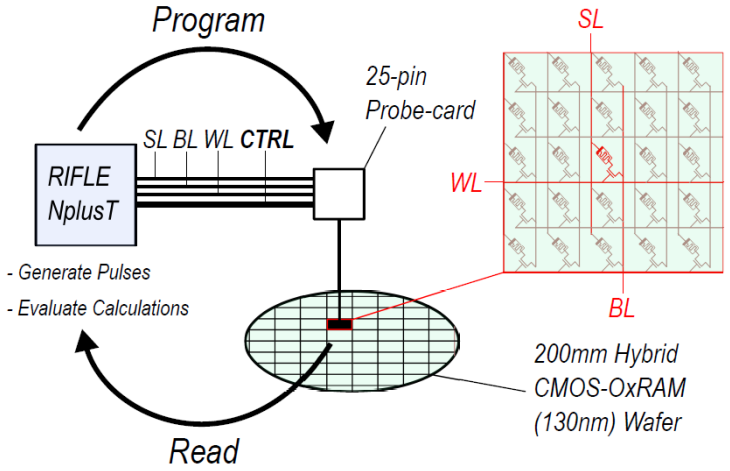
ACCEPT/REJECT sampling from posterior distribution

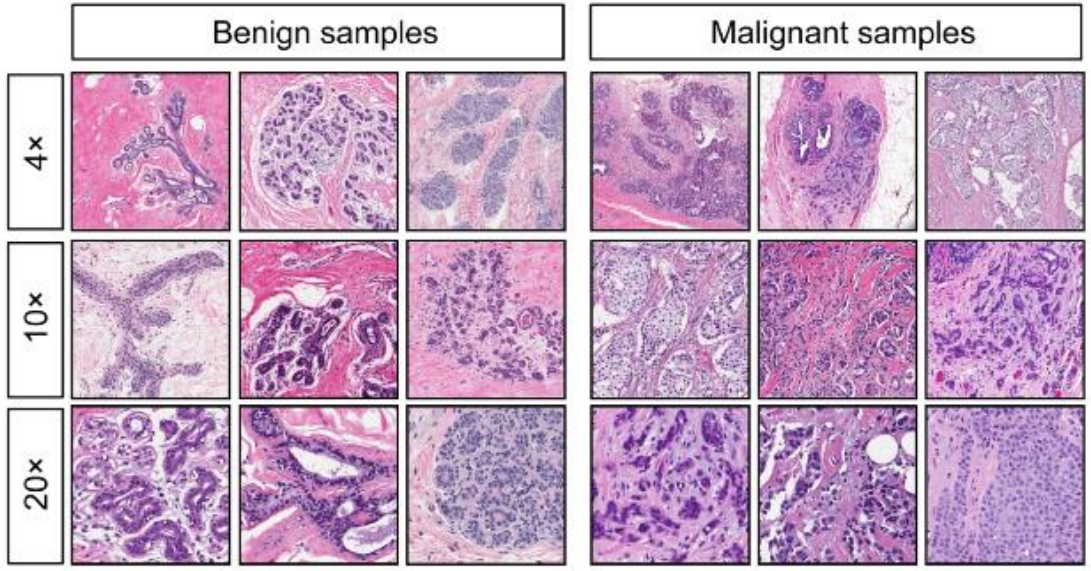
Copyright © 2023

Supervised learning with RRAM-based MCMC

Computer-in-the-loop with a 16kb device array







Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, *43*(4), 570-577.

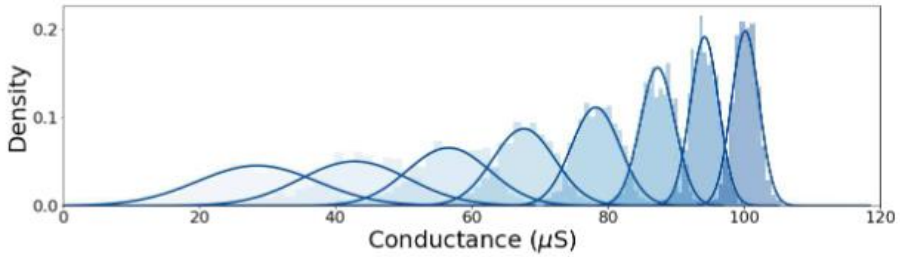
The experimental system was able to detect maligant tissue with 98% accuracy

MCMC difficult to scale to large/deep neural networks

Copyright © 2023

Bayes by Backprop and technological loss for variational inference

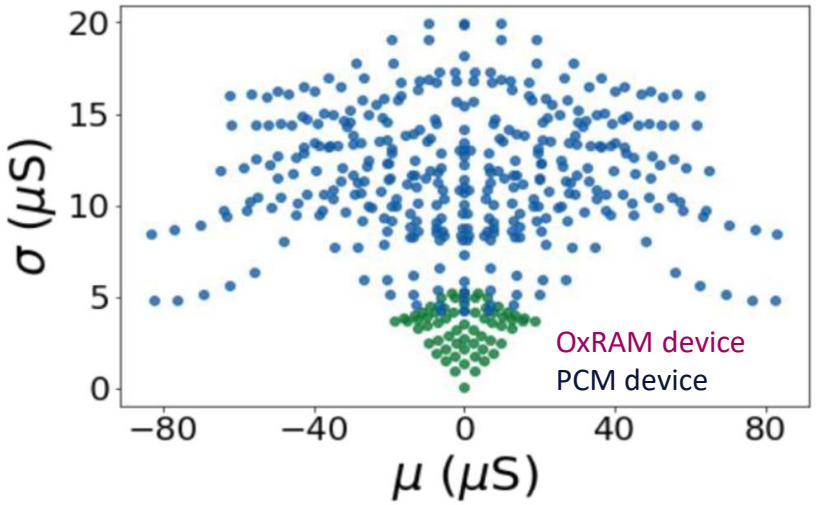
Distribution using one memristor

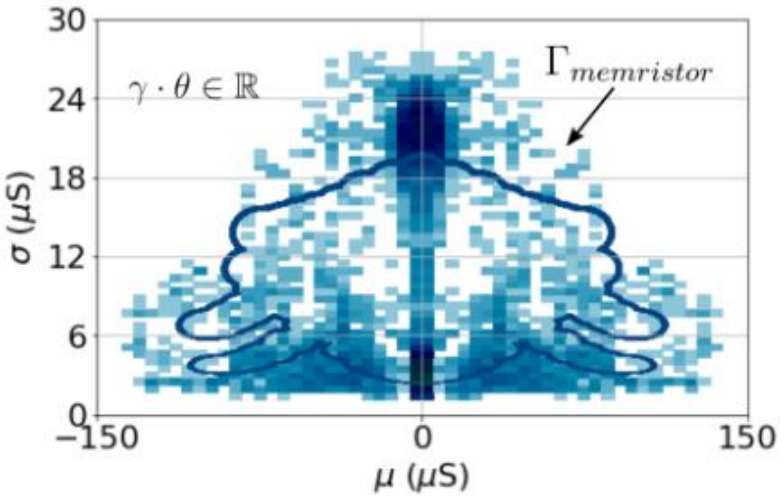


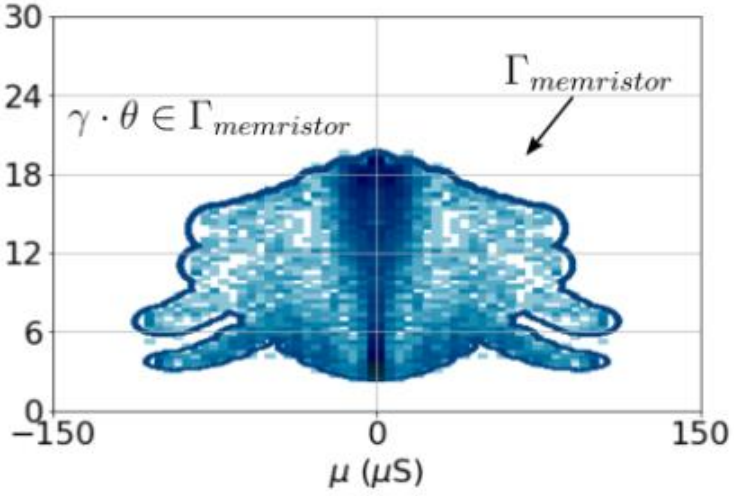
- Scaling up Bayesian Neural network with backpropagation
- Learning with backpropagation on (μ, σ) parameters
- **Hardware & software co-development :**
  → Technological loss

$$Loss = KL[q(\Omega|\theta)||p(\Omega|D)]$$

Distribution using two memristors & complex programming





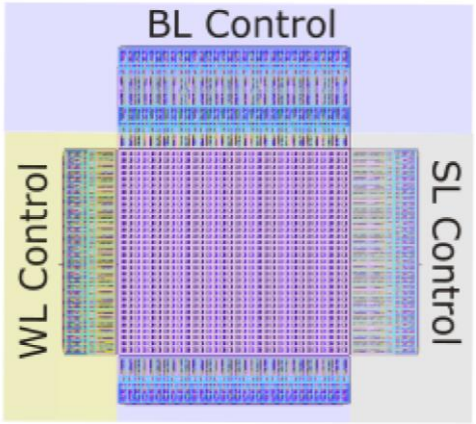


Bonnet, Hirtzlin et al, preprint DOI: 10.21203/rs.3.rs-2458251/v1

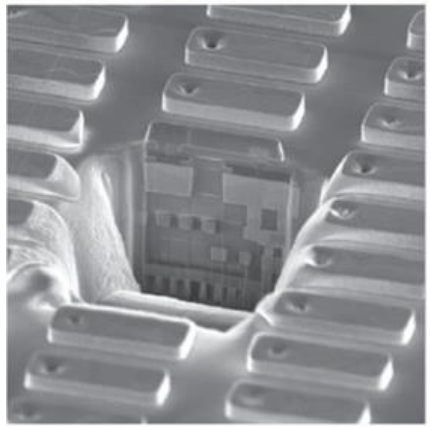
Copyright © 2023

Fully Experimental Arrythmia Recogniton with Uncertainty Evaluation

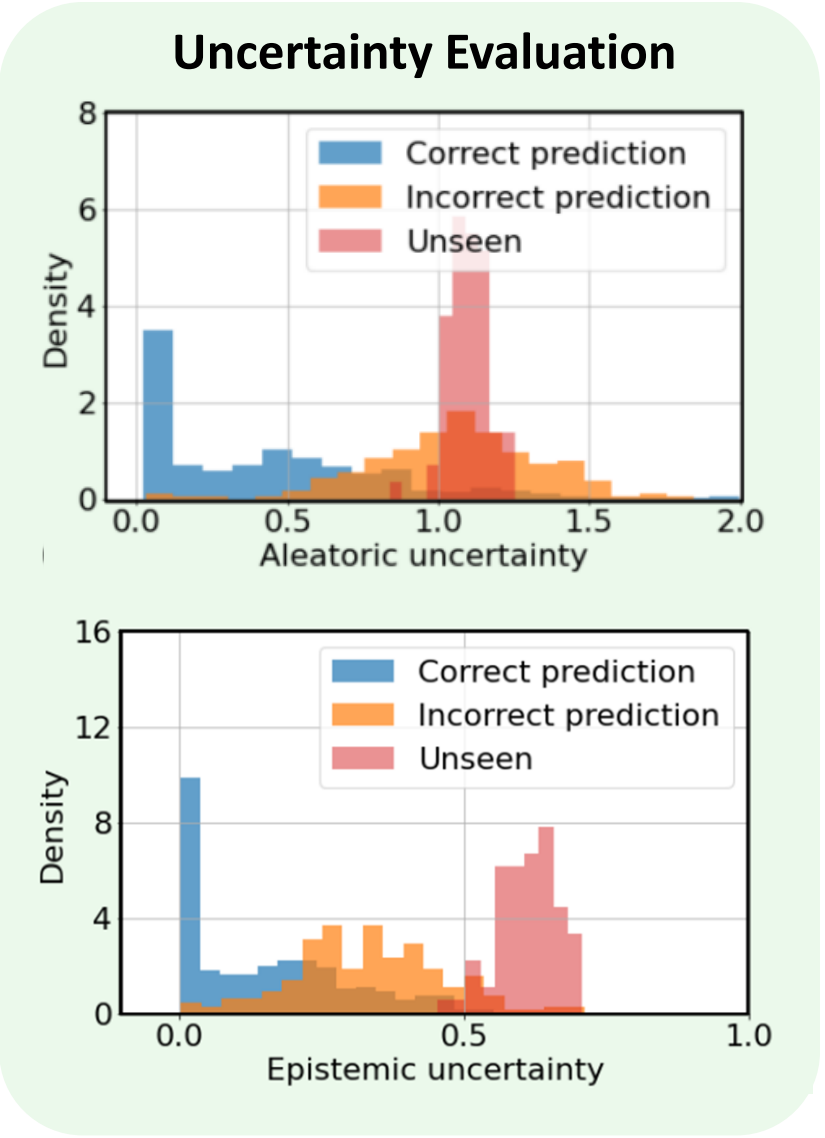
- 79% raw accuracy (software: 80%)
- Unknown types of arrythmia are easily recognized



In-Memory-Computing macro



Memristor device (OxRAM) integration



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

Copyright © 2023

Conclusion

- Memristors offers new possibilities for AI based hardware

  → *non volatility feature*
  → *in memory computing*
  → *intrinsic probabilistic nature can be exploited for learning*
  → *…*

- Bayesian framework is ideal to model uncertainty

  → *Especially adapted for "small data"/ high uncertainty situations (where wrong answers have dramatic impact)*

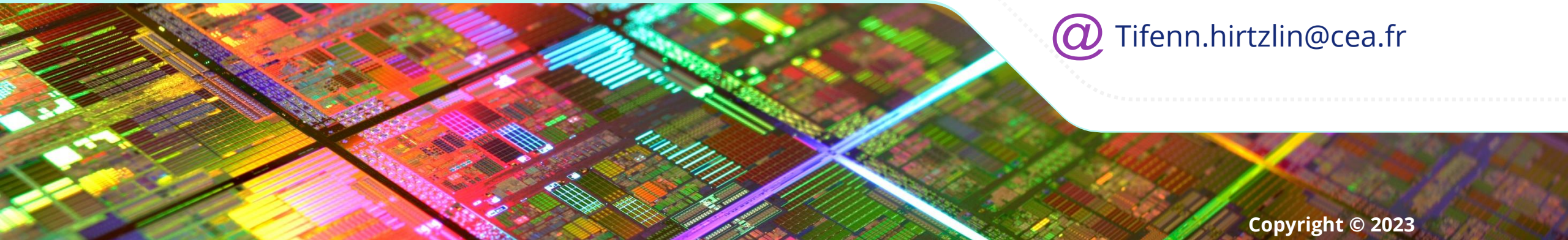
- Expertise on device / circuit / architecture / software & algorithm offers large possibilities for developing new Edge AI hardware

Copyright © 2023

EDGE AI

Thank You

For your attention



@ Tifenn.hirtzlin@cea.fr

Copyright © 2023

Acknowledgments

- Kamel-Eddine Harabi
- Clément Turck
- Jacques-Olivier Klein
- Damien Querlioz

- Elisa Vianello
- Djohan Bonnet
- Thomas Dalgaty
- 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

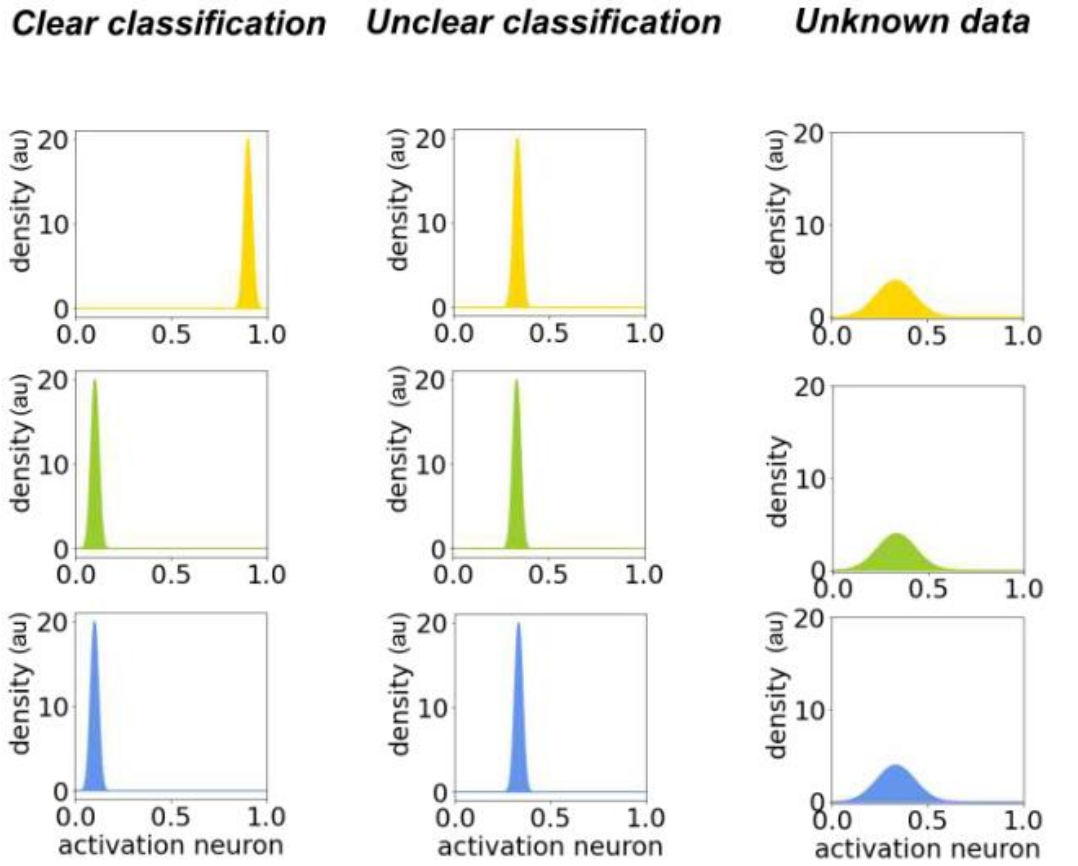
Copyright © 2023

Uncertainty calculation

$$U_a = -\frac{1}{M}\sum_{s=1}^{M}\sum_{c=1}^{N} p(y=c|X,W_s) log\ p(y=c|X,W_s).$$

$$U_p = -\sum_{c=1}^{N}\left(\frac{1}{M}\sum_{s=1}^{M} p(y=c|X,W_s)\right) log\left(\frac{1}{M}\sum_{s=1}^{M} p(y=c|X,W_s)\right)$$

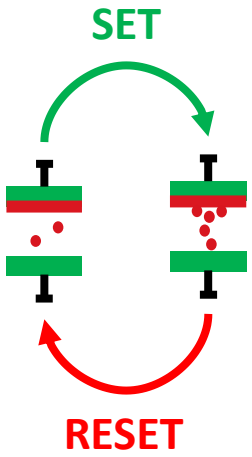
$$U_p = U_a + U_e.$$



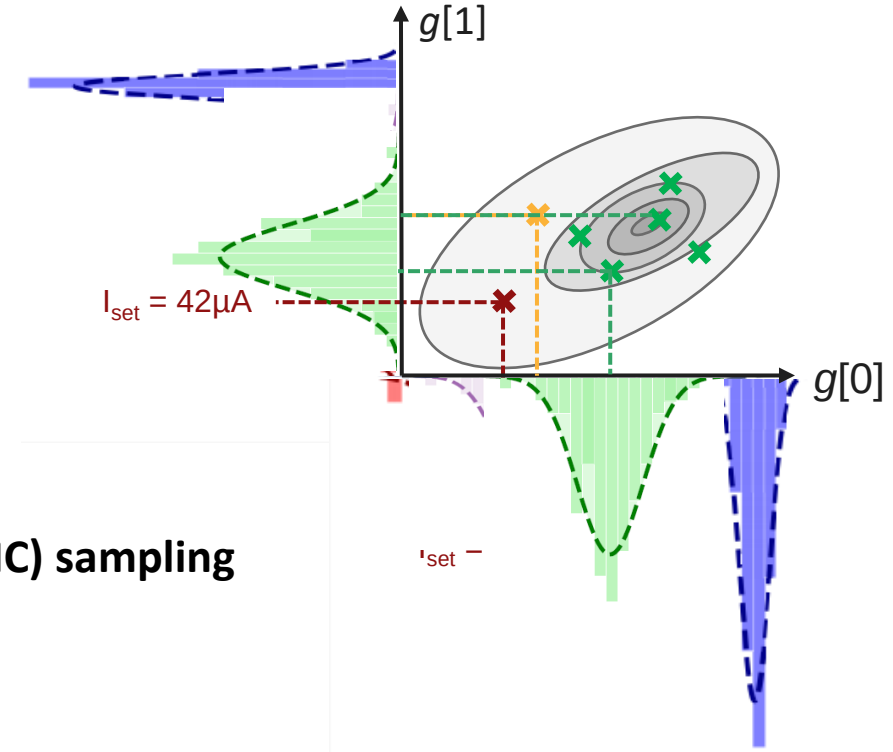
Copyright © 2023

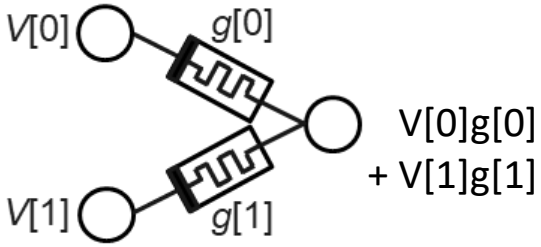
MCMC Resistive memory compatible learning

Instead of fighting against intrinsic device properties, embrace them with a more suitable algorithm



Markov chain Monte Carlo (MCMC) sampling





(1) SET initialization

(2) RESET/SET with corresponding compliance current

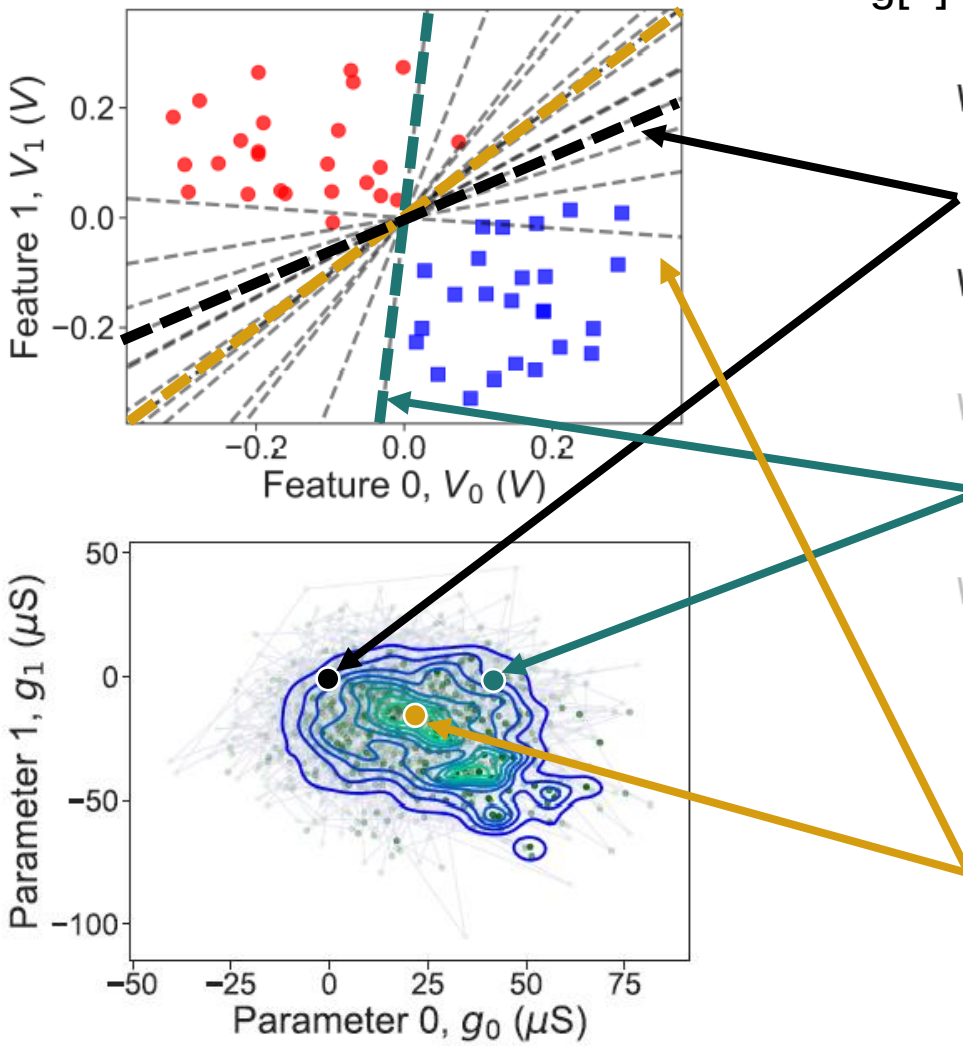
(3) RESET/SET with corresponding compliance current

(4) RESET/SET with corresponding compliance current

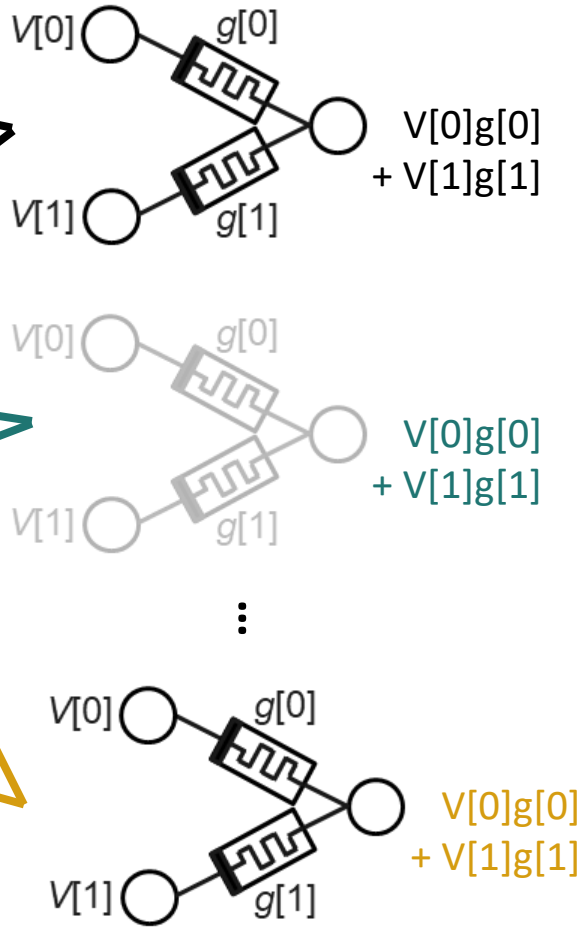
Copyright © 2023

RRAM-based MCMC sampling

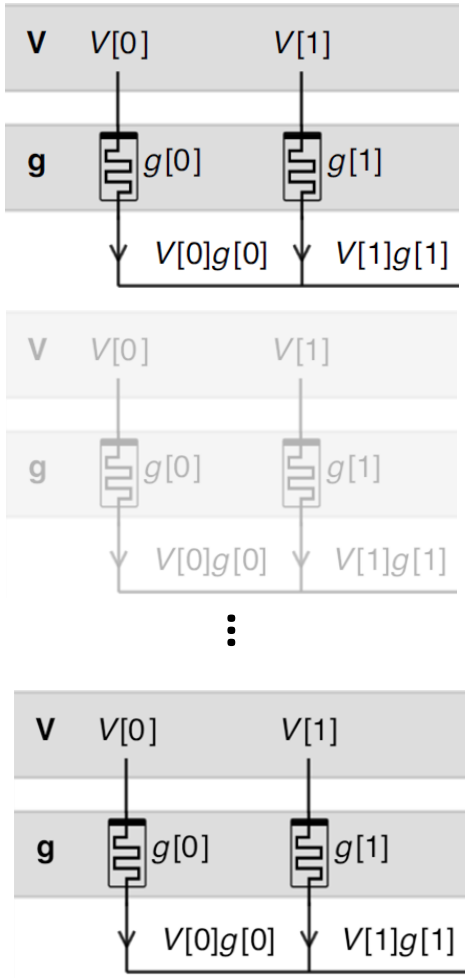
- Simple toy problem



- Simple Neural Network with g[0] & g[1] parameters multiple times

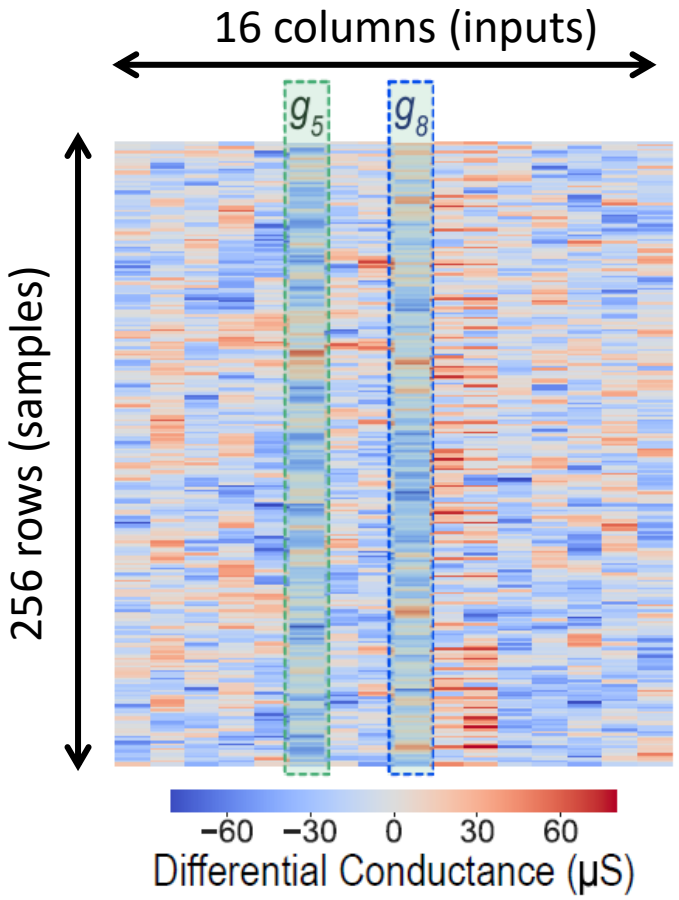


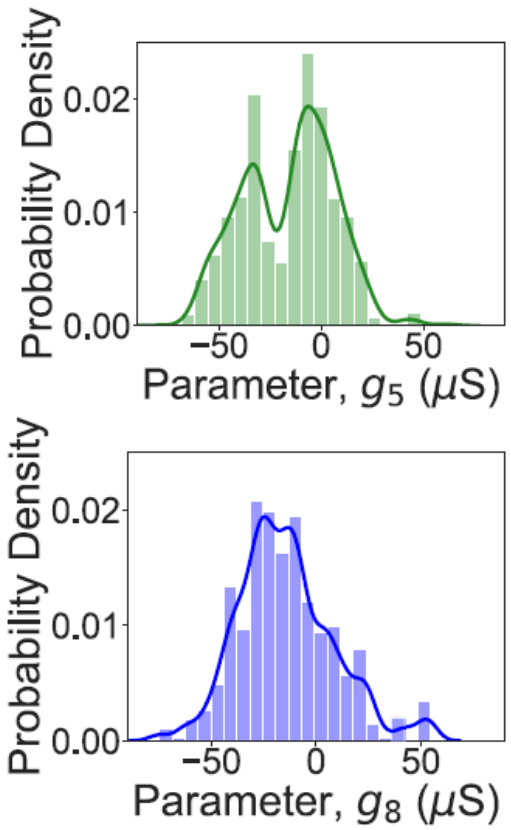
- Simple implementation using Kirchhoff's law

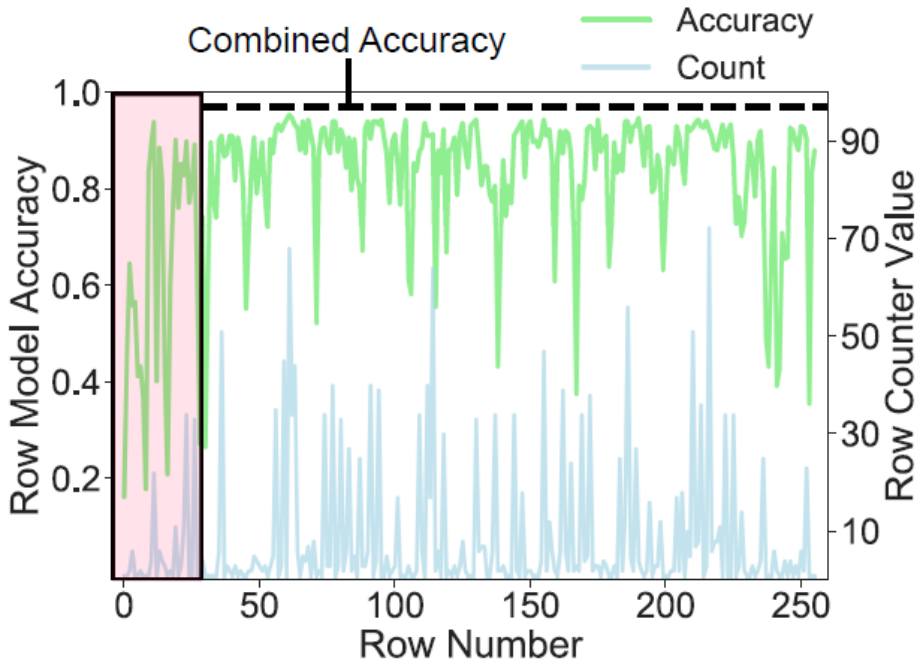


Copyright © 2023

Experimental System



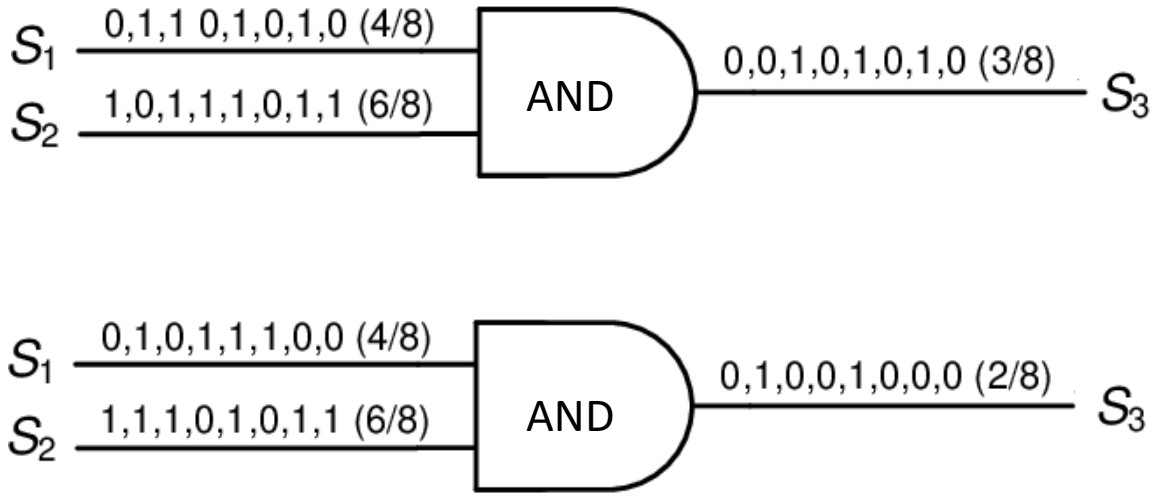




The experimental system was able to detect maligant tissue with 98% accuracy

Copyright © 2023

Stochastic Computing

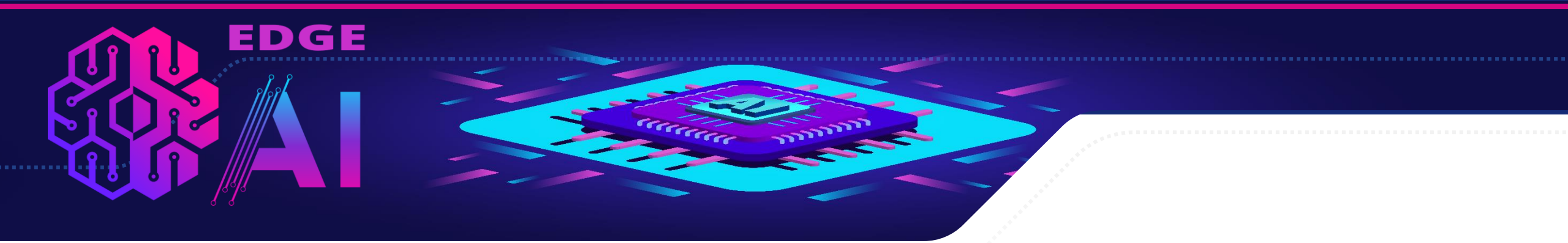


$p(S_3)=p(S_1)p(S_2)$

A AND gate implements the multiplication of two probabilities!

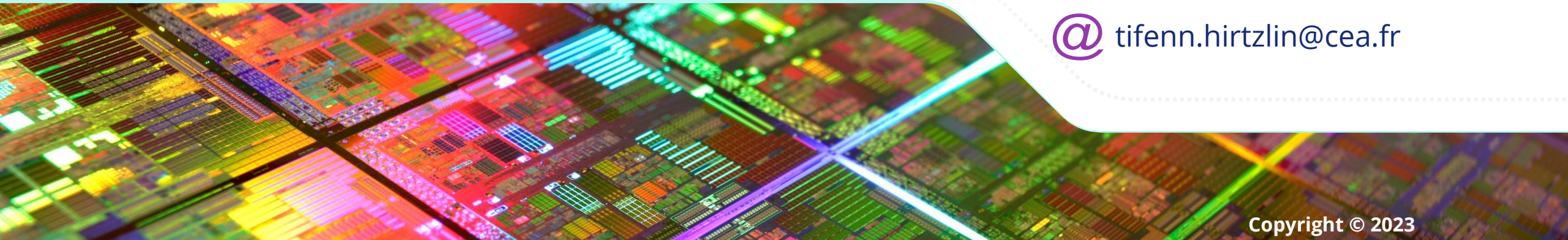
Some resemblance with neurons (one wire = one real number)

Copyright © 2023



Thank You

For your attention



@ tifenn.hirtzlin@cea.fr

Copyright © 2023