Swiss Physical Society Annual Meeting 2024

Neuromorphic Intelligence: spiking neural network and on-line learning circuits for brain-inspired technologies

Giacomo Indiveri

Institute of Neuroinformatics
University of Zurich and ETH Zurich

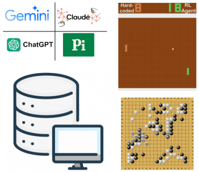
Universität Zürich UZH

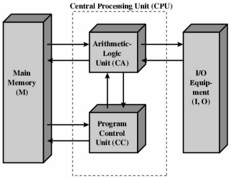
ETH zürich

Outline

Artificial vs natural

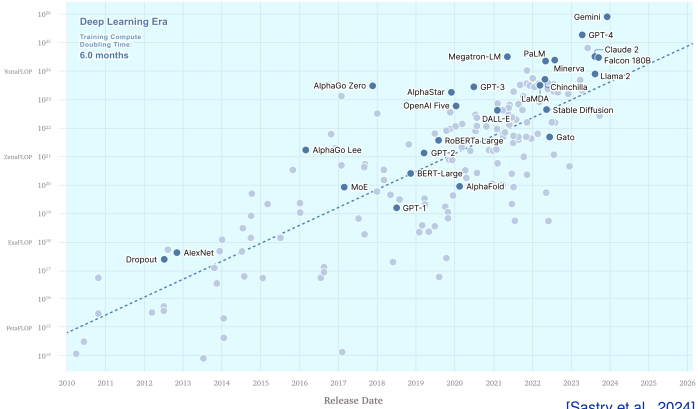
"Intelligent"





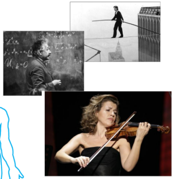
~1MW

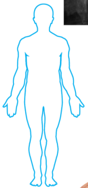
Compute Used for AI Training Runs (Deep Learning Era)



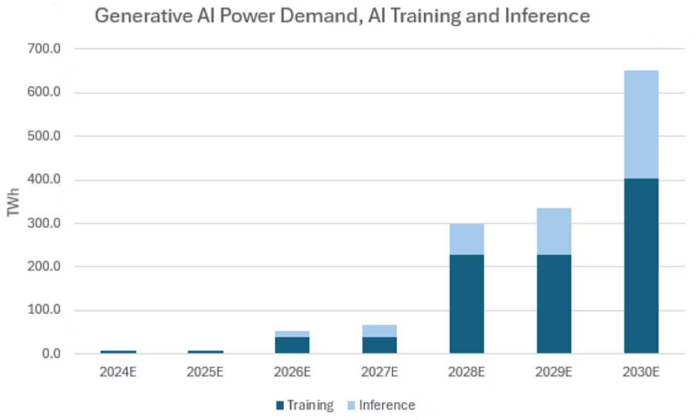
[Sastry et al., 2024]

"Intelligent"

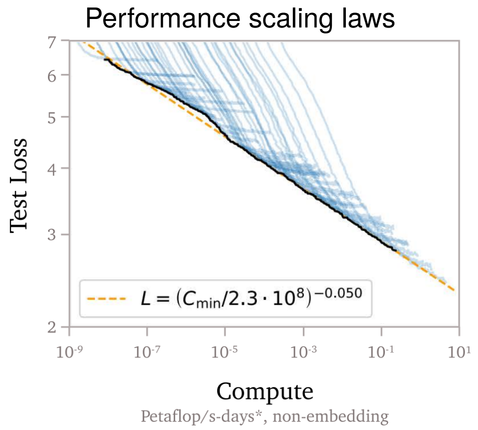




~10W



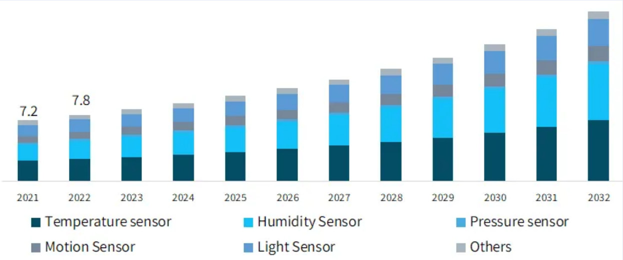
[io-fund.com]



[Sastry et al., 2024]

AI training is expected to drive the power demand to 402 TWh by 2030 (about the same demand of the whole of France or Germany in 2023)

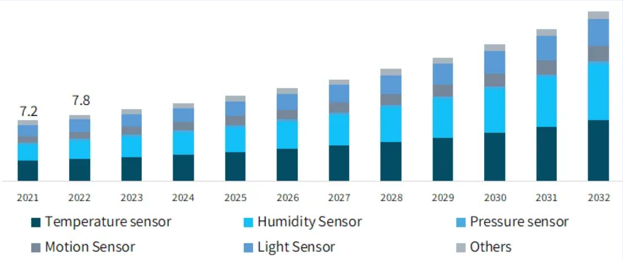
Global passive sensors market (USD billion)



[gminsights.com]

- More than 50 billion Internet of Things (IoT) devices are expected by 2030
- Embedded devices with sensors and/or actuators are the key components of the IoT
- Local “intelligence” is key to reducing communication, bandwidth and energy consumption.

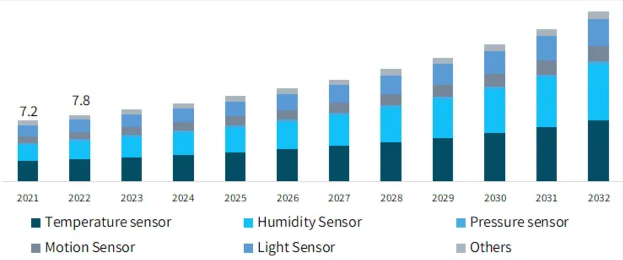
Global passive sensors market (USD billion)



[gminsights.com]

- More than 50 billion Internet of Things (IoT) devices are expected by 2030
- Embedded devices with sensors and/or actuators are the key components of the IoT
- Local "intelligence" is key to reducing communication, bandwidth and energy consumption.

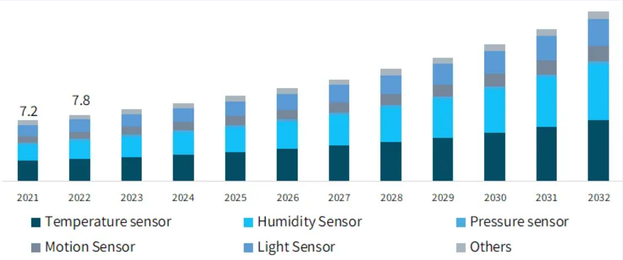
Global passive sensors market (USD billion)



[gminsights.com]

- More than 50 billion Internet of Things (IoT) devices are expected by 2030
- Embedded devices with sensors and/or actuators are the key components of the IoT
- Local "intelligence" is key to reducing communication, bandwidth and energy consumption.

Global passive sensors market (USD billion)



[gminsights.com]

- More than 50 billion Internet of Things (IoT) devices are expected by 2030
- Embedded devices with sensors and/or actuators are the key components of the IoT
- Local “intelligence” is key to reducing communication, bandwidth and energy consumption.

Clearly it is not possible to use conventional large-scale AI methods to endow IoT devices with intelligence.

Reducing energy in embedded systems

Conventional approaches

1. Make application specific (lose general purpose flexibility)
2. Quantize parameters (reduce bit precision)
3. Minimize resource usage (reduce accuracy)

Novel approaches

1. Reduce data movement (implement in-memory computing)
2. Reduce clock switching (use asynchronous circuits)
3. Exploit all the physics of the devices (mix analog and digital)

Reducing energy in embedded systems

Conventional approaches

1. Make application specific (lose general purpose flexibility)
2. Quantize parameters (reduce bit precision)
3. Minimize resource usage (reduce accuracy)

Novel approaches

1. Reduce data movement (implement in-memory computing)
2. Reduce clock switching (use asynchronous circuits)
3. Exploit all the physics of the devices (mix analog and digital)

Conventional approaches

1. Make application specific (lose general purpose flexibility)
2. Quantize parameters (reduce bit precision)
3. Minimize resource usage (reduce accuracy)

Novel approaches

1. Reduce data movement (implement in-memory computing)
2. Reduce clock switching (use asynchronous circuits)
3. Exploit all the physics of the devices (mix analog and digital)

$\Downarrow$

Novel computing paradigms: brain-like computation

- Co-localize memory and computation (local processing, local state variables)
- Maximize fine-grain parallelism (massively parallel arrays of memory and processing)
- Use the "physics of computation" (exploit properties of computing substrate)

[Indiveri Sandamirskaya, IEEE Signal Processing Magazine, 2019; Indiveri Liu, Proceedings of IEEE, 2015]

Animal brains

- Slow, noisy and variable processing elements.
- Local connectivity, small world networks.
- Massively parallel distributed computation.
- Closed-loop interaction with the environment.
- Real-time spatio-temporal signal processing.
- Continual always-on learning.

Existence proof

Bee brain specs

weight:	1 mg
volume:	1 mm^3
# neurons:	960’000
energy/op:	10^{-15} J/spike

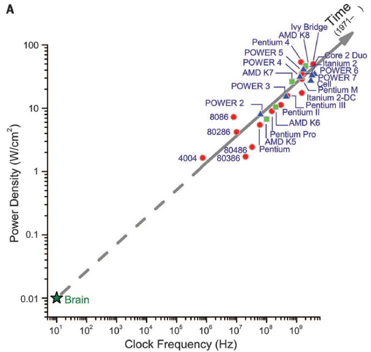
Animal brains

- Slow, noisy and variable processing elements.
- Local connectivity, small world networks.
- Massively parallel distributed computation.
- Closed-loop interaction with the environment.
- Real-time spatio-temporal signal processing.
- Continual always-on learning.

Time represents itself

The brain uses the time evolution of the physical system to implement its computations. Neural circuits compute by exploiting the natural time evolution of their hardware substrate. [Sterling & Laughlin, 2017]

Clock speed

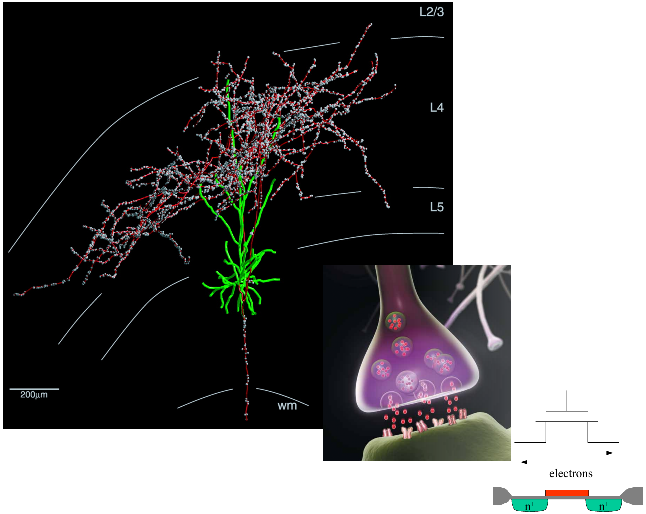


Brains outperform faster computing systems in many sensory processing tasks at lower speeds, with less power.

Neuromorphic Intelligence (according to me)

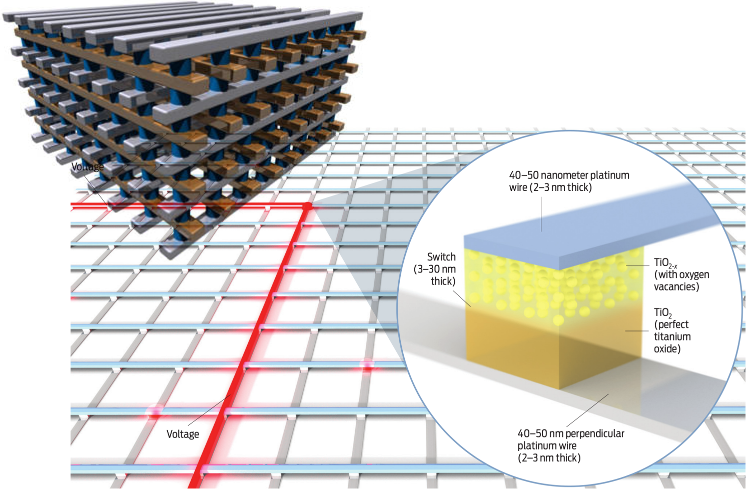
Neuroscience

- Study the principles of computation in animal brains
- Identify them at the neural circuit level
- Emulate the bio-physics of neurons and synapses using analog electronic circuits
- Validate/invalidate hypotheses of neural computation



Electronics

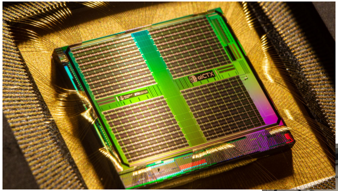
- Include novel devices and emerging memory technologies
- Exploit (all) the physics of these nanoscale devices
- Integrate CMOS and memristive devices together
- Engineer efficient "in-memory computing" architectures

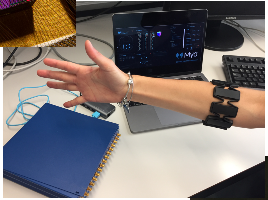


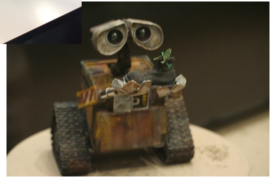
Neuromorphic Intelligence (according to me)

Computing

- Implement “neural processing” systems in custom ASICS
- Integrate processors with sensors and actuators
- Apply them to closed-loop sensory processing tasks
- Develop *cognitive agents* that produce autonomous behavior.







Neuroscience

- Study the principles of computation in animal brains
- Identify them at the neural circuit level
- Emulate the bio-physics of neurons and synapses using analog electronic circuits
- Validate/invalidate hypotheses of neural computation

Device physics

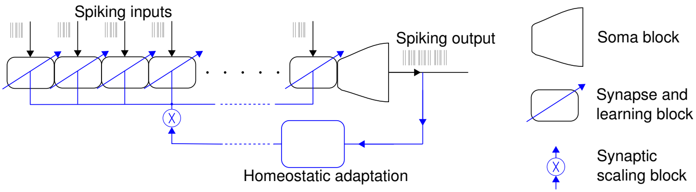
- Include novel devices and emerging memory technologies
- Exploit (all) the physics of these nanoscale devices
- Integrate CMOS and memristive devices together
- Engineer efficient "in-memory computing" architectures

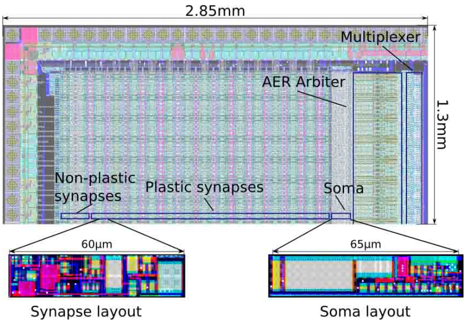
Computing

- Implement "neural processing" systems in custom ASICS
- Integrate processors with sensors and actuators
- Apply them to closed-loop sensory processing tasks
- Develop *cognitive agents* that produce autonomous behavior.

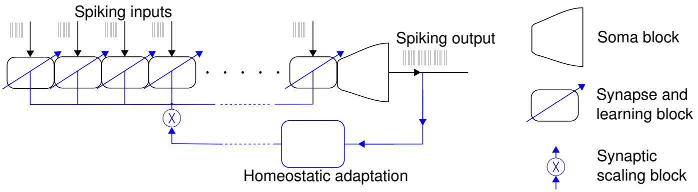
Outline

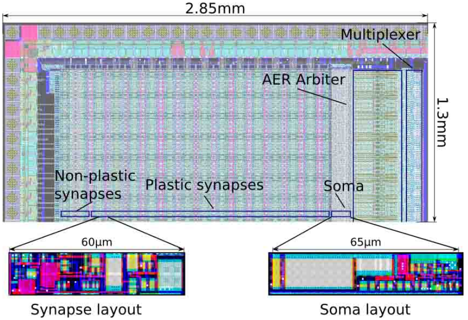
Mixed-signal neuromorphic chips

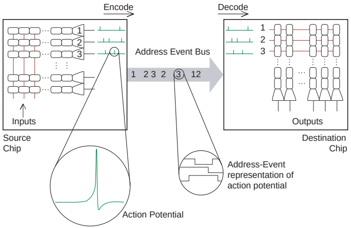




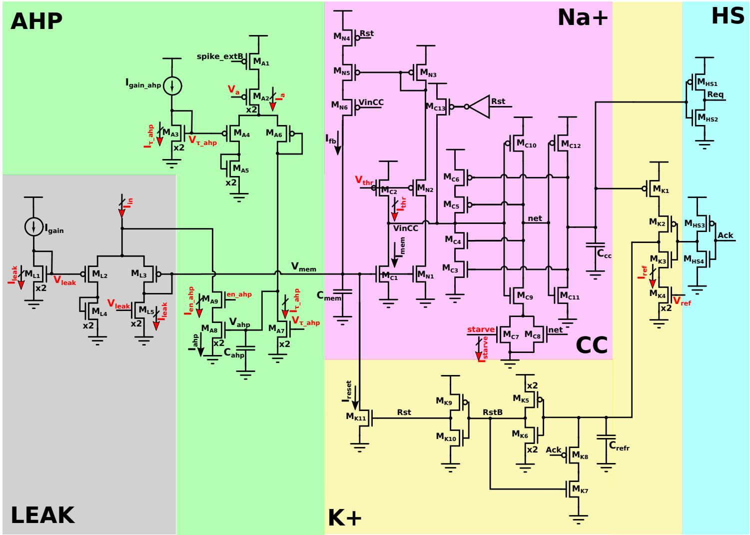
- Spiking neural networks (SNNs)
- Analog subthreshold circuits.
- Slow temporal, non-linear dynamics.
- Massively parallel operation.
- Compatible with memristive devices
- Inhomogeneous, imprecise, and noisy.

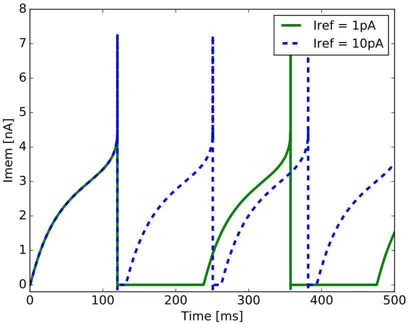






- Spiking neural networks (SNNs)
- Analog subthreshold circuits.
- Slow temporal, non-linear dynamics.
- Massively parallel operation.
- Compatible with memristive devices
- Inhomogeneous, imprecise, and noisy.
- Fast asynchronous digital circuits for routing spikes.
- Reprogrammable network topology



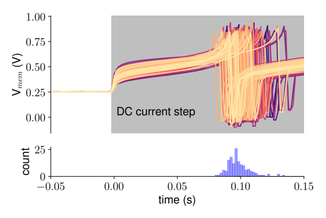


Work	[12]	[19]	[36]	**This work**
Techn.	180 nm CMOS	28 nm CMOS	28 nm FDSOI	22 nm FDSOI
Type	Mixed	Mixed	Mixed	Mixed
V_{dd}	1.8 V	0.7-1 V	1 V	0.8 V
Freq	30 Hz	-	30 Hz	30 Hz
Results	Experimental	Experimental	Simulation	Simulation
En./spike	883 pJ	2.3 nJ-30 nJ	50 pJ	14 pJ

[Brette and Gerstner, 2005, Rubino et al., IEEE TCAS, 2020]

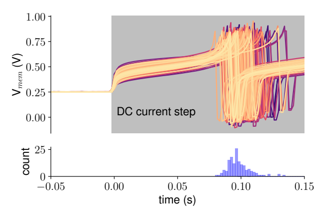
Adaptive Exponential I&F neuron model (beyond HH)

Device mismatch effects



Time-to-first spike measured across 256 different neurons

Device mismatch effects

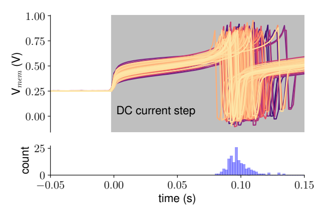


Time-to-first spike measured across 256 different neurons

How to cope with mismatch?

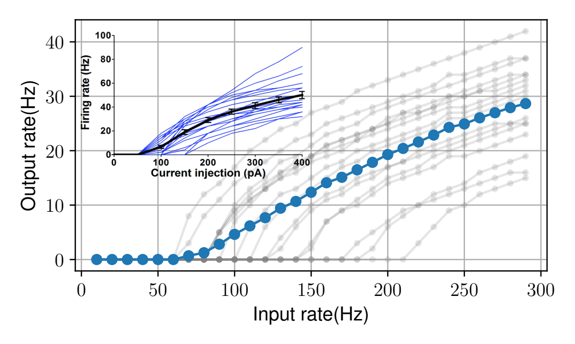
- Use populations of neurons and average over space and time
- Employ negative feedback, adaptation, and learning mechanisms

Device mismatch effects



Time-to-first spike measured across 256 different neurons

Population codes and averaging

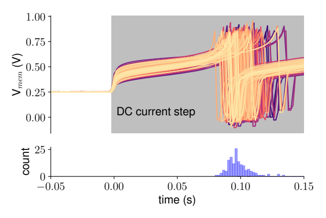


How to cope with mismatch?

- Use populations of neurons and average over space and time
- Employ negative feedback, adaptation, and learning mechanisms

Robust computation with inhomogeneous devices

Device mismatch effects

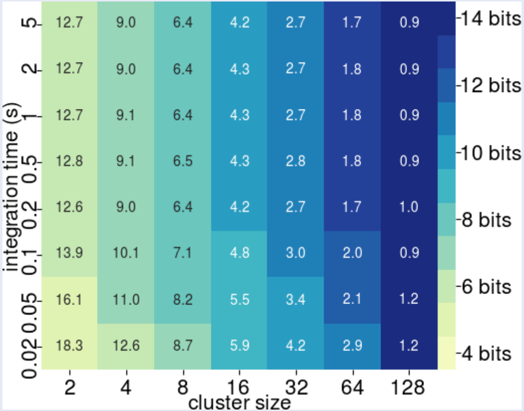


Time-to-first spike measured across 256 different neurons

How to cope with mismatch?

- Use populations of neurons and average over space and time
- Employ negative feedback, adaptation, and learning mechanisms

Choosing bit resolution

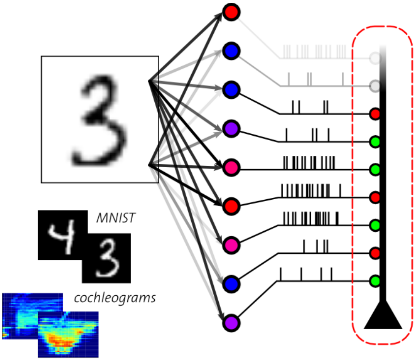


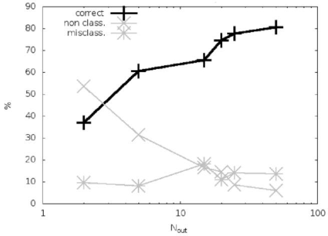
Coefficient of variation and Equivalent Number of Bits (ENOB)

[Zendrikov et al., 2023]

Outline

Ensemble learning techniques exploit variability of inhomogeneous synapses.





MNIST	deep/CNN (Hinton et al. 2012)	98.4%
	random + bistable synapses	∼ 85%
	random + bistable synapses + (mod. protocol)	∼ 96%
TIMIT	deep/CNN (Hinton et al. 2012)	77%
	VLSI cochlea + bistable synapses	∼ 60%

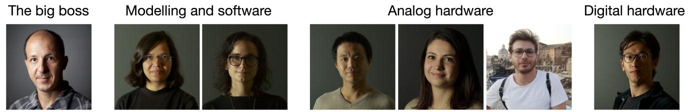
On-line bagging techniques *require* variability

AdaBoost theorem: $1-\text{error}(H_{final}) \geq 1 - e^{-2\gamma^2 N}$ [Freund and Schapire, 1997]

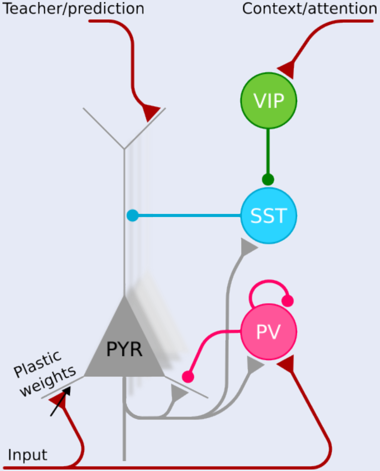
Features

- Compatible with "biological" and "electronic" computing substrate
- Based on latest dendritic multi-compartment models
- Exploits properties of multiple inhibitory cell types
- Makes use of population coding and dynamics
- Implements a "stop learning" mechanism to automatically switch between training and inference

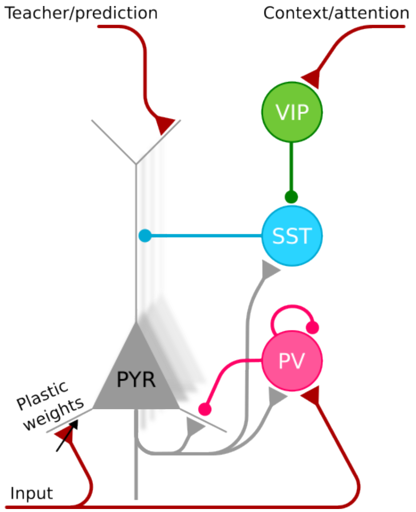
Design team (FDSOI 22 nm, 2024)



Cortical motif

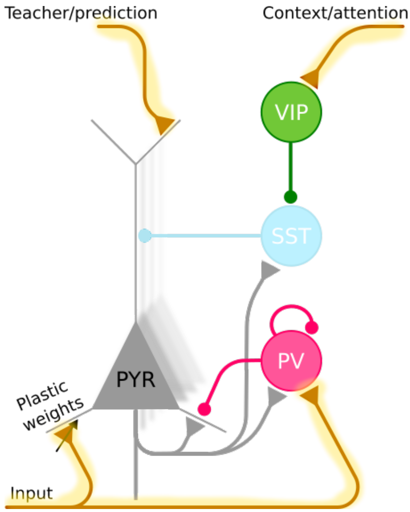


Circuit operating principle



Spiking inputs arrive from both top-down pathways (context, attention, prediction signals) and bottom-up pathways (sensory signals).

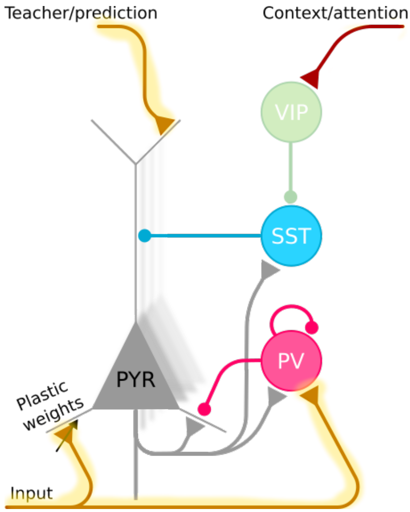
- An E-I balance maintains the population in a proper operating range at all times.
- During training, teacher signals reach the soma and change plastic weights
- During inference, SST inhibitory cells block top-down inputs, the neurons respond only to bottom-up inputs with lower firing rates, and synaptic weights "stop learning".



Spiking inputs arrive from both top-down pathways (context, attention, prediction signals) and bottom-up pathways (sensory signals).

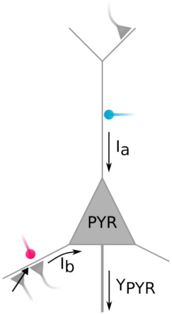
- An E-I balance maintains the population in a proper operating range at all times.
- During training, teacher signals reach the soma and change plastic weights
- During inference, SST inhibitory cells block top-down inputs, the neurons respond only to bottom-up inputs with lower firing rates, and synaptic weights "stop learning".

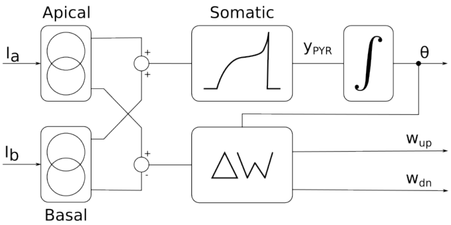
Circuit operating principle



Spiking inputs arrive from both top-down pathways (context, attention, prediction signals) and bottom-up pathways (sensory signals).

- An E-I balance maintains the population in a proper operating range at all times.
- During training, teacher signals reach the soma and change plastic weights
- During inference, SST inhibitory cells block top-down inputs, the neurons respond only to bottom-up inputs with lower firing rates, and synaptic weights “stop learning”.





$$Y_{PYR} = \sigma\left(I_a + I_b\right)$$

$$\theta = \int Y_{PYR} dt$$

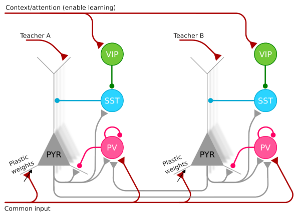
$$\Delta w_{up} = \eta(I_a - I_b)\sigma_{LTP} \quad \text{if} \quad I_a \geq I_b$$
$$\Delta w_{dn} = \eta(I_a - I_b)\sigma_{LTD} \quad \text{if} \quad I_a < I_b$$

$$
\begin{aligned}
I_a &= \left[I_T - I_{SST}\right]^+ \\
I_b &= \left(I_{in} + I_{PYR} - I_{PV}\right) \\
I_{in} &= \int \sum_i w_i \alpha\left(t - \delta(t_i)\right) dt
\end{aligned}
$$

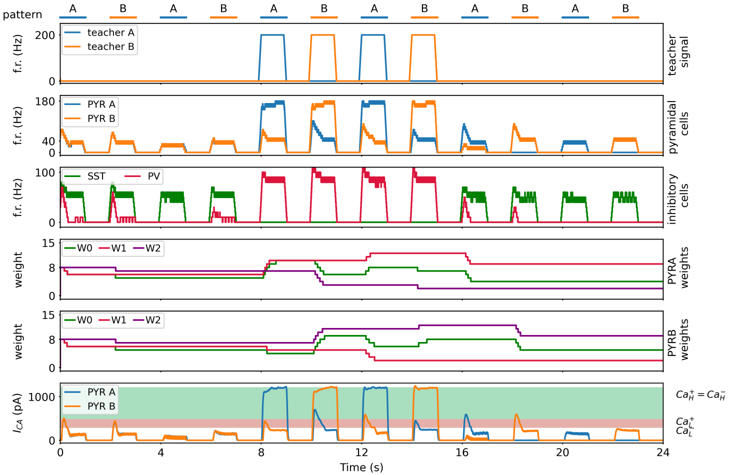
$$\sigma_{LTP} = \begin{cases} 1 & \text{if} \quad \theta_{LTP_-} < \theta < \theta_{LTP_+} \\ 0 & \text{otherwise} \end{cases}$$

$$\sigma_{LTD} = \begin{cases} 1 & \text{if} \quad \theta_{LTD_-} < \theta < \theta_{LTD_+} \\ 0 & \text{otherwise} \end{cases}$$

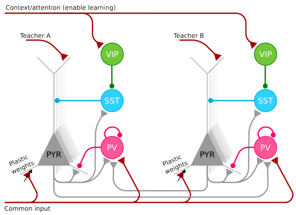
Binary classification: changing only relevant weights

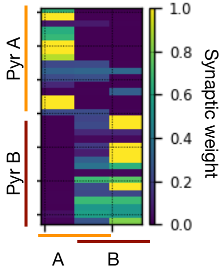


SPICE circuit simulations of two neurons

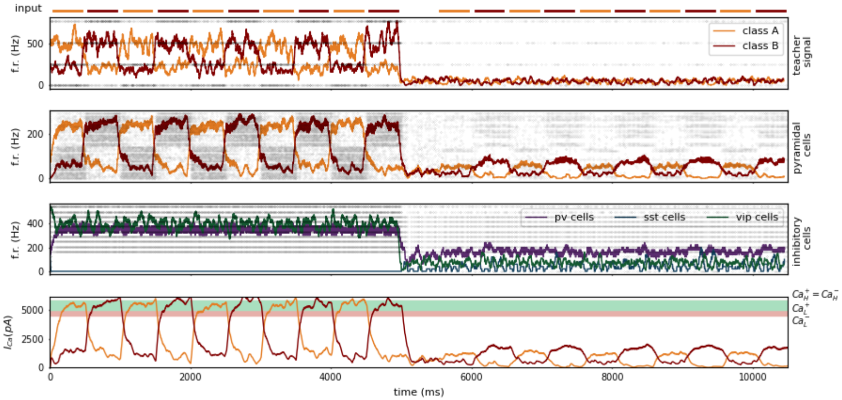


Binary classification: changing only relevant weights





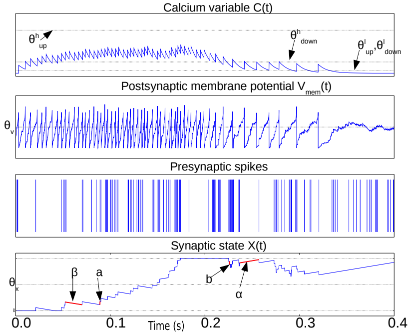
NEST SW simulations at full population level (with mismatch)



Robust spike-based learning mechanisms

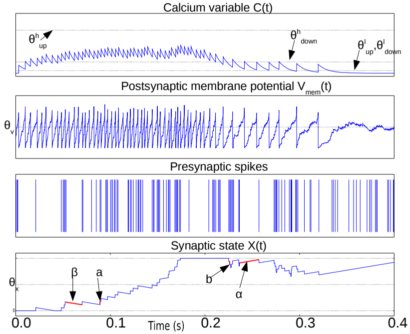
There are many hardware-friendly spike-driven learning algorithms that (go beyond STDP).

W. Senn, S. Fusi, N. Brunel, S. Sheik, E. Neftci, R. Zecchina, M. Memmesheimer, etc.



There are many hardware-friendly spike-driven learning algorithms that (go beyond STDP).

W. Senn, S. Fusi, N. Brunel, S. Sheik, E. Neftci, R. Zecchina, M. Memmesheimer, etc.



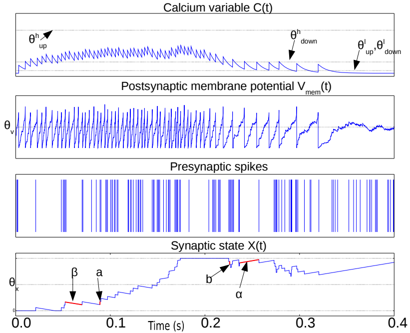
All rule have the following *requirements*:

- Redundancy (population codes)
- Bi-stable or multi-stable weights
- Variability and heterogeneity
- Analog, continuous-time state variables

Robust spike-based learning mechanisms

There are many hardware-friendly spike-driven learning algorithms that (go beyond STDP).

W. Senn, S. Fusi, N. Brunel, S. Sheik, E. Neftci, R. Zecchina, M. Memmesheimer, etc.



All rule have the following *requirements*:

- Redundancy (population codes)
- Bi-stable or multi-stable weights
- Variability and heterogeneity
- Analog, continuous-time state variables

Many mixed-signal hardware implementations have been demonstrated:

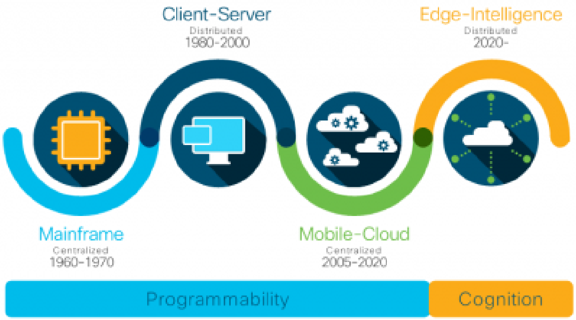
- Supervised learning, mean rates
- Unsupervised learning, precise spike-timing
- Hopfield/attractor networks
- Reservoir computing, liquid state and perceptron
- Ensemble learning (random forest, bagging)

[Khacef et al., 2023]

Outline

Edge intelligence

Mixed-signal neuromorphic systems are optimally suited for extreme-edge computing applications, which require resource constrained electronic systems. They are ideal for always-on in-sensor and in-memory computing applications that need to perform closed-loop interactions the environment, in real-time.



Example: wearables and health monitoring

- Neuromorphic CPG for adaptive pace-makers [Abu-Hassan et al., 2019]
- ECG anomaly detection [Bauer et al., 2019,Corradi et al., 2019]
- EMG signal classification [Donati et al., 2019,Ma et al, 2020]
- High-Frequency Oscillation (HFO) detection [Sharifhazileh et al., 2021,Burelo et al., 2022]
- Neuromorphic Heart Rate Monitors [Carpegna et al., 2024]

Outline

Conclusions

nature

EDITORIAL · 06 FEBRUARY 2018

Big data needs a hardware revolution

Artificial intelligence is driving the next wave of innovations in the semiconductor industry.

Software companies make headlines but research on computer hardware could bring bigger rewards. Credit: Morris MacMatzen/Getty

- Conventional AI increasing power requirements are unsustainable.
- New emerging memory technologies will benefit from massively parallel processing architectures.
- Neuroscience and machine learning are uncovering powerful and robust neural processing methods.
- Hardware implementations of spiking neural networks and sparse event-based sensory-processing systems are starting to show their advantages.
- This is the perfect time to follow the “neuromorphic intelligence” approach for starting a hardware revolution.

Conclusions

nature

EDITORIAL · 06 FEBRUARY 2018

Big data needs a hardware revolution

Artificial intelligence is driving the next wave of innovations in the semiconductor industry.

Software companies make headlines but research on computer hardware could bring bigger rewards. Credit: Morris MacMatzen/Getty

- Conventional AI increasing power requirements are unsustainable.
- New emerging memory technologies will benefit from massively parallel processing architectures.
- Neuroscience and machine learning are uncovering powerful and robust neural processing methods.
- Hardware implementations of spiking neural networks and sparse event-based sensory-processing systems are starting to show their advantages.
- This is the perfect time to follow the “neuromorphic intelligence” approach for starting a hardware revolution.

Conclusions

nature

EDITORIAL · 06 FEBRUARY 2018

Big data needs a hardware revolution

Artificial intelligence is driving the next wave of innovations in the semiconductor industry.

Software companies make headlines but research on computer hardware could bring bigger rewards. Credit: Morris MacMatzen/Getty

- Conventional AI increasing power requirements are unsustainable.
- New emerging memory technologies will benefit from massively parallel processing architectures.
- Neuroscience and machine learning are uncovering powerful and robust neural processing methods.
- Hardware implementations of spiking neural networks and sparse event-based sensory-processing systems are starting to show their advantages.
- This is the perfect time to follow the “neuromorphic intelligence” approach for starting a hardware revolution.

Conclusions

nature

EDITORIAL · 06 FEBRUARY 2018

Big data needs a hardware revolution

Artificial intelligence is driving the next wave of innovations in the semiconductor industry.

Software companies make headlines but research on computer hardware could bring bigger rewards. Credit: Morris MacMatzen/Getty

- Conventional AI increasing power requirements are unsustainable.
- New emerging memory technologies will benefit from massively parallel processing architectures.
- Neuroscience and machine learning are uncovering powerful and robust neural processing methods.
- Hardware implementations of spiking neural networks and sparse event-based sensory-processing systems are starting to show their advantages.
- This is the perfect time to follow the “neuromorphic intelligence” approach for starting a hardware revolution.

Conclusions

nature

EDITORIAL · 06 FEBRUARY 2018

Big data needs a hardware revolution

Artificial intelligence is driving the next wave of innovations in the semiconductor industry.

Software companies make headlines but research on computer hardware could bring bigger rewards. Credit: Morris MacMatzen/Getty

- Conventional AI increasing power requirements are unsustainable.
- New emerging memory technologies will benefit from massively parallel processing architectures.
- Neuroscience and machine learning are uncovering powerful and robust neural processing methods.
- Hardware implementations of spiking neural networks and sparse event-based sensory-processing systems are starting to show their advantages.
- This is the perfect time to follow the “neuromorphic intelligence” approach for starting a hardware revolution.

Acknowledgments

institute of neuroinformatics

- Elisa Donati
- Chiara De Luca
- Sapta Ghosh
- Chenxi Wen
- Dmitrii Zendrikov
- Farah Baracat
- Junren Chen
- Yigit Demirag
- Maryada
- Shyam Narayanan
- Arianna Rubino
- Zhe Su

Swiss National Science Foundation

Thank you for your attention

Backup slides

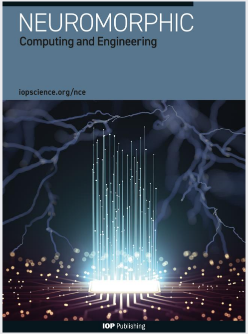
Journal on Neuromorphic Computing and Engineering

Universität Zürich[UZH] ETH zürich

NEUROMORPHIC
Computing and Engineering

OPEN ACCESS

No APCs in 2024



A multidisciplinary, open access journal devoted to the application and development of neuromorphic computing, devices, and systems in advancing new scientific discovery and realising emerging new technologies.

Editor-in-Chief
Giacomo Indiveri University of Zurich, Switzerland

Indexed in Scopus and Web of Science

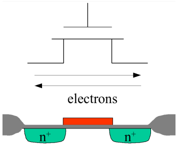
IMPACT FACTOR COMING IN JUNE **2024**

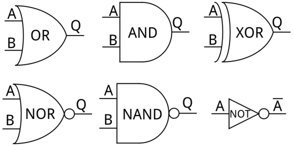
IOP Publishing

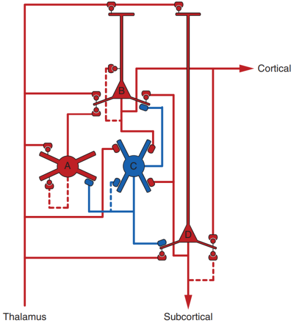
Iopscience.org/nce
nce@ioppublishing.org

@IOPneuromorphic

In addition to using populations of neurons and use learning and plasticity to improve robustness of neural processing, it is useful to identify and adopt basic building blocks that implement key principles of computation.



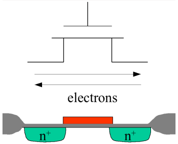


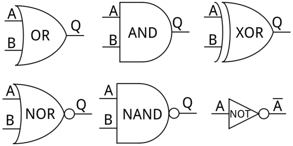


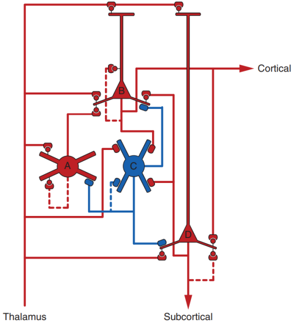
Neural computational primitives

In addition to using populations of neurons and use learning and plasticity to improve robustness of neural processing, it is useful to identify and adopt basic building blocks that implement key principles of computation.

- Attractor networks
- E-I balanced networks
- Winner-Take-All networks
- Relational networks
- Coupled oscillators
- Neural State Machines

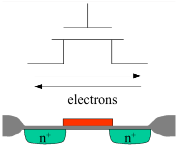


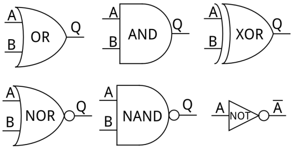




Neural computational primitives

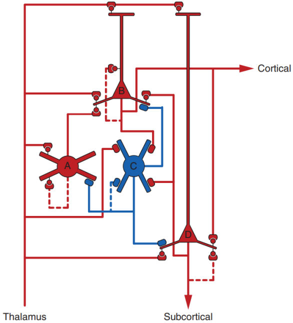
In addition to using populations of neurons and use learning and plasticity to improve robustness of neural processing, it is useful to identify and adopt basic building blocks that implement key principles of computation.





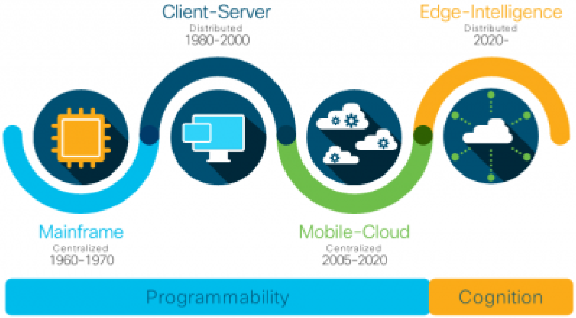
- Attractor networks
- E-I balanced networks
- Winner-Take-All networks
- Relational networks
- Coupled oscillators
- Neural State Machines

  see also poster by Maryada et al. (Calcium-based dendritic plasticity)



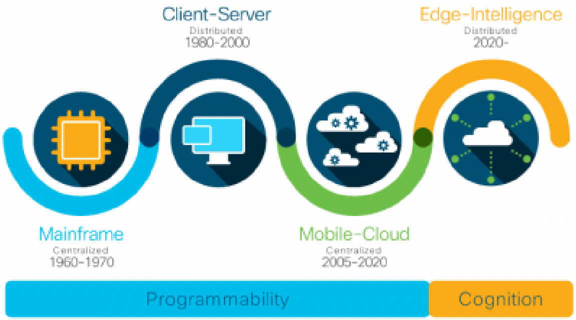
Edge intelligence

Mixed-signal neuromorphic systems are optimally suited for extreme-edge computing applications, which require resource constrained electronic systems. They are ideal for always-on in-sensor and in-memory computing applications that need to closed-loop interactions the environment, in real-time.



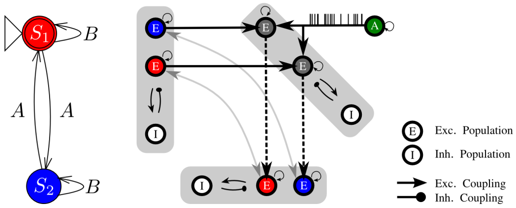
Edge intelligence

Mixed-signal neuromorphic systems are optimally suited for extreme-edge computing applications, which require resource constrained electronic systems. They are ideal for always-on in-sensor and in-memory computing applications that need to closed-loop interactions the environment, in real-time.

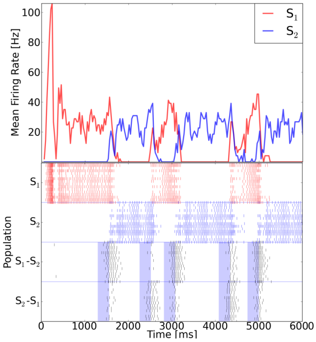


Example: wearables and health monitoring

- Neuromorphic CPG for adaptive pace-makers [Abu-Hassan et al., 2019]
- ECG anomaly detection [Bauer et al., 2019,Corradi et al., 2019]
- EMG signal classification [Donati et al., 2019,Ma et al, 2020]
- High-Frequency Oscillation (HFO) detection [Sharifhazileh et al., 2021,Burelo et al., 2022]
- Neuromorphic Heart Rate Monitors [Carpegna et al., 2024]

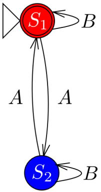


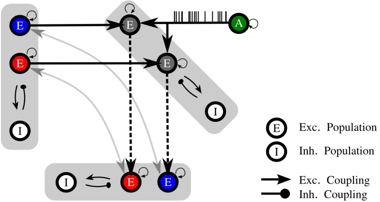
S_1
B
A
A
S_2
B
E
E
A
E
E
I
I
E
E
I
E Exc. Population
I Inh. Population
Exc. Coupling
Inh. Coupling

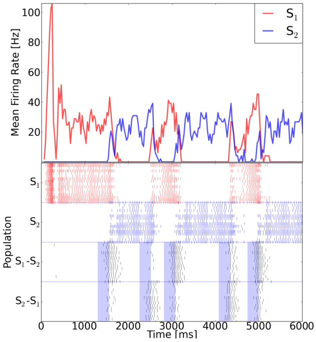


Mean Firing Rate [Hz]
100
80
60
40
20
S_1
S_2
Population
S_1-S_2
S_2-S_1
0
1000
2000
3000
4000
5000
6000
Time [ms]

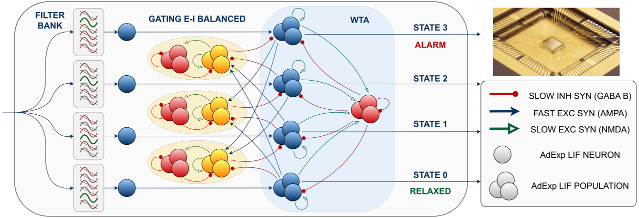
Neural state machines and slow monotonic changes



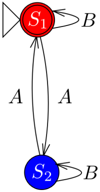


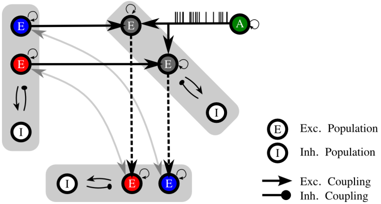


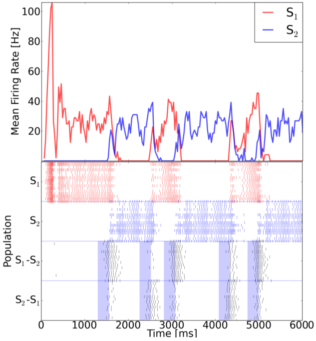
Detecting "agitation states" by monitoring monotonic increases in heart rates, over long-time periods.

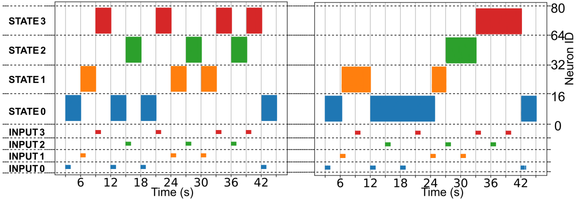


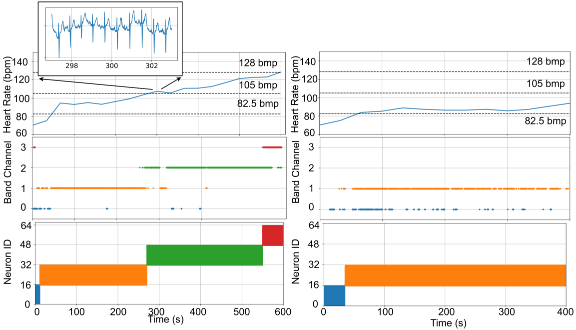
Neural state machines and slow monotonic changes











Neuromorphic vs conventional processors

Pros

- Low-power (< 1 mW)
- Low latency
- ...

Cons

- High area
- High variability, noisy
- Low(er) accuracy

Neuromorphic vs conventional processors

Universität Zürich[UZH] ETHzürich

Pros

- Low-power (< 1 mW)
- Low latency
- ...

Cons

- High area
- High variability, noisy
- Low(er) accuracy

What are they good for?

- Closed-loop sensory-motor processing
- Multi-modal sensory fusion
- Always-on on-line learning

What are they bad at?

- High precision number crunching
- High accuracy pattern recognition
- Batch processing of large data sets

Neuromorphic vs conventional processors

Pros

- Low-power (< 1 mW)
- Low latency
- ...

Cons

- High area
- High variability, noisy
- Low(er) accuracy

What are they good for?

- Closed-loop sensory-motor processing
- Multi-modal sensory fusion
- Always-on on-line learning

What are they bad at?

- High precision number crunching
- High accuracy pattern recognition
- Batch processing of large data sets

Open challenges

- How to obtain robust and reliable computation using a noisy and heterogeneous computing substrate.
- How to program networks of spiking neurons (hint: compose *computational primitives* and use *learning*).

Background

We are building physical, real-time, signal processing systems for real-world sensory data.

Background

We are building physical, real-time, signal processing systems for real-world sensory data.

Requirements

1. Robust communication of analog signals across long distances through noisy channels.
2. Local processing, multi-core architectures and distributed computing.
3. Low power and low-latency.

Why spikes?

Background

We are building physical, real-time, signal processing systems for real-world sensory data.

Requirements

1. Robust communication of analog signals across long distances through noisy channels.
2. Local processing, multi-core architectures and distributed computing.
3. Low power and low-latency.

Optimal solution for communication and computation

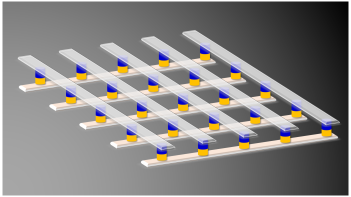
- The optimal method that minimizes bandwidth and power consumption for achieving this goal, under these constraints, is pulse-frequency modulation. [A. Mortara et al., 1995, K. Boahen, 1998]
- "*Counter to intuition, computing with spikes can be extremely efficient on neuromorphic hardware even when the problem being solved is mathematically formulated in terms of activity rates.*" [M. Davies, Intel, 2019]

Advantages

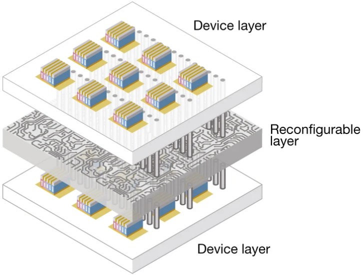
- Compute through the dynamics of the circuits
- No need to “count time”: avoid use of clock trees
- Avoid large DAC/ADC overhead
- Exploit the full potential of memristors
  - Exploit intrinsic non-linearities [Brivio et al., 2021]
  - Exploit intrinsic stochasticity [Gaba et al., 2013, Payvand et al., 2018]

Disadvantages (?)

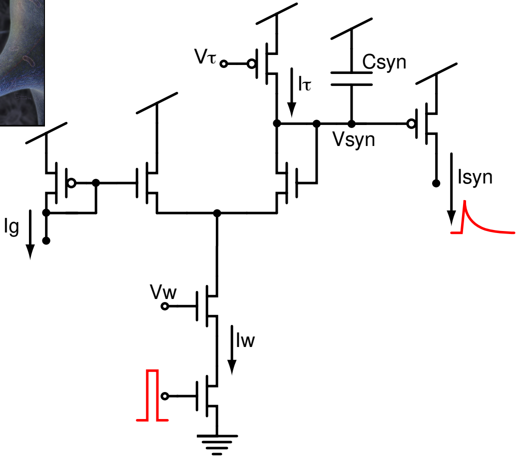
- Noisy $\Longrightarrow$ average across multiple neurons (exploit population coding and heterogeneity)
- Large area requirements $\Longrightarrow$ employ memristive devices and 3D VLSI (exploit low power)

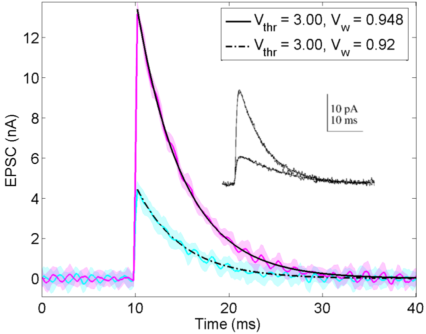


PCM cross-bar array [Source: IBM]



“Dendrocentric learning for synthetic intelligence” [Boahen, 2022]



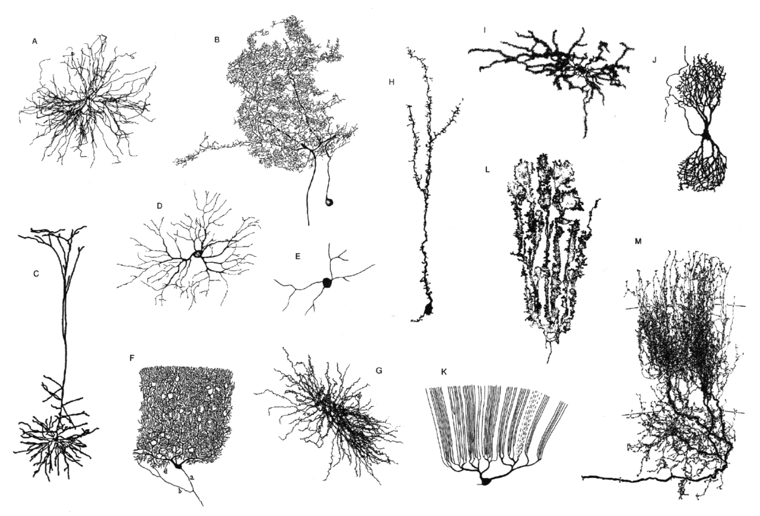


$$\tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_g I_w}{I_\tau}$$

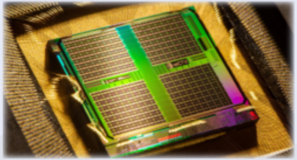
$$\tau = \frac{C U_T}{\kappa I_\tau}$$

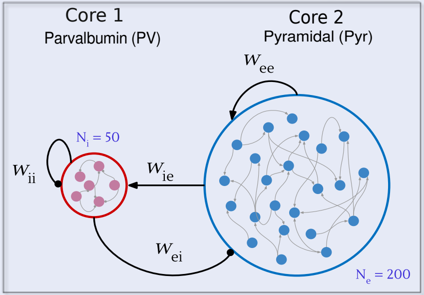
[Bartolozzi, Indiveri, NECO 2007]

Also real neurons are diverse and inhomogeneous



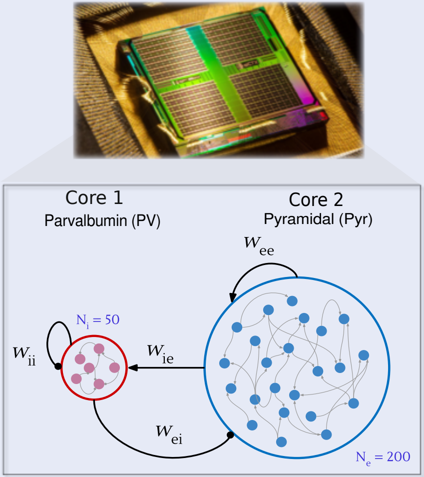
Balanced HW E-I networks



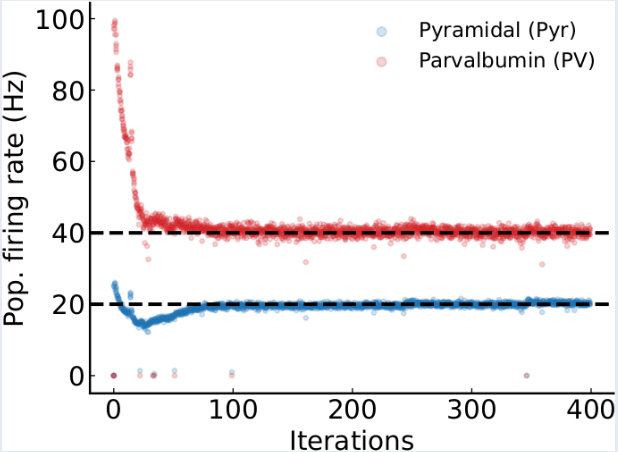


Maryada

Balanced HW E-I networks



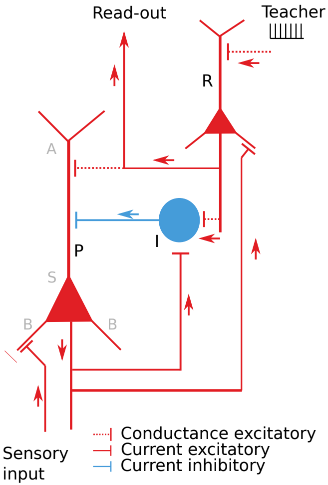
Cross-homeostatic plasticity induced stability

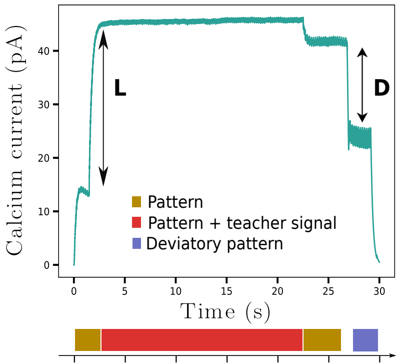


[Maryada et al., 2023]

Local learning rules

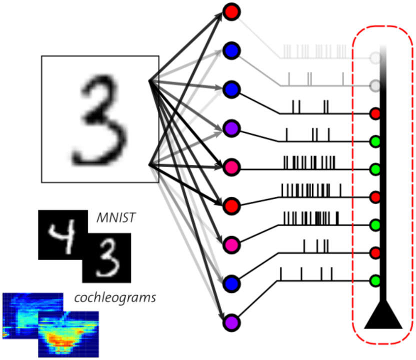
- Fusi et al. 2000
- Brader et al. 2007
- Urbanczik, Senn 2014
- Baldassi et al. 2016
- Neftci et al. 2017
- Sacramento et al. 2018
- Bellec et al. 2019
- Zenke, Vogels 2021
- Siddique et al. 2023
- ...

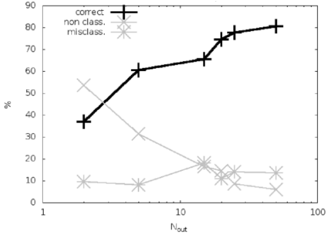




[Cartiglia et al., AICAS 2019]

Ensemble and stochastic learning can exploit variability of inhomogeneous synapses.





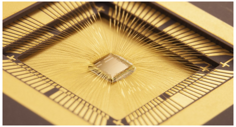
MNIST	deep/CNN (Hinton et al. 2012)	98.4%
	random + bistable synapses	∼ 85%
	random + bistable synapses + (mod. protocol)	∼ 96%
TIMIT	deep/CNN (Hinton et al. 2012)	77%
	VLSI cochlea + bistable synapses	∼ 60%

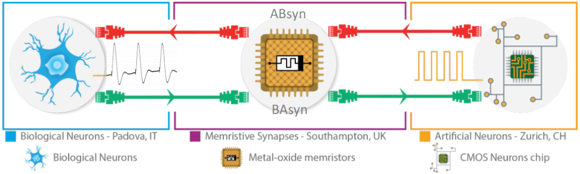
On-line bagging techniques

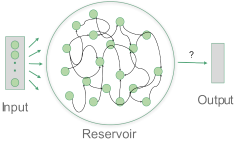
AdaBoost theorem: $1 - \text{error}(H_{final}) \geq 1 - e^{-2\gamma^2 N}$

[Y. Freund And R. E. Schapire, 1995]

Outline



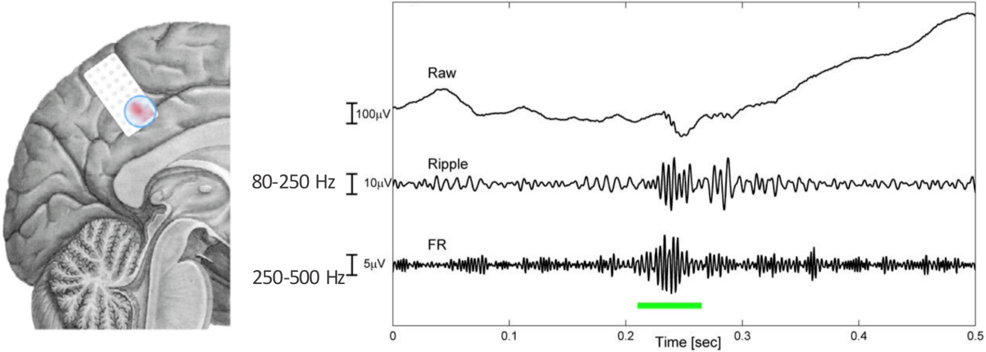




- Zebra-finch "Bird's Own Song" classification [Corradi et al., 2015]
- Closed-loop bidirectional brain machine interfaces with in rats and cell-cultures [Boi et al., 2016] [Serb et al. 2020]
- Adaptive pace-maker with neuromorphic CPG network [Abu-Hassan et al., 2019]
- On-line ECG anomaly detection [Bauer et al., 2019]
- On-line classification of EMG signals [Donati et al., 2019]
- Closed-loop obstacle avoidance on roving robot [Milde et al. 2017]
- Closed-loop robot head position control with a neuromorphic processor [Zhao et al., 2020]
- Neuromorphic pattern generation circuits for bioelectronic medicine [Donati et al., 2021]
- Instantaneous stereo depth estimation of real-world stimuli with a stereo-vision setup [Risi et al., 2021]
- On-line detection of vibration anomalies using balanced spiking neural networks [Dennler et al., 2021]
- High-Frequency Oscillation (HFO) detection [Sharifhazileh, Burelo et al., 2021]

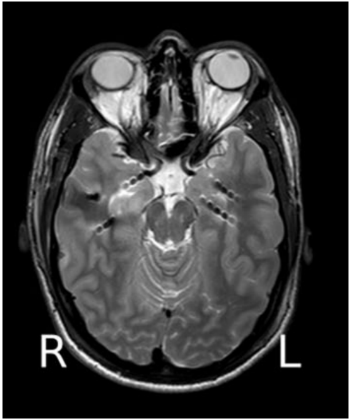
What is an HFO?

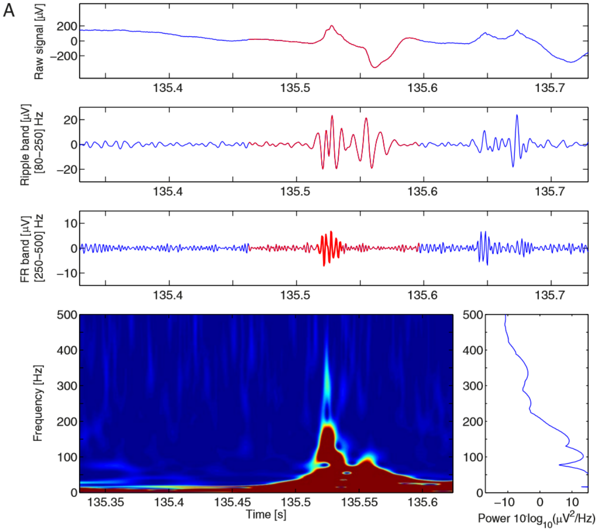
Spontaneous EEG events in the frequency range between 80 and 500 Hz consisting of at least four oscillations that clearly stand out from the baseline.



[Fedele et al. 2017]

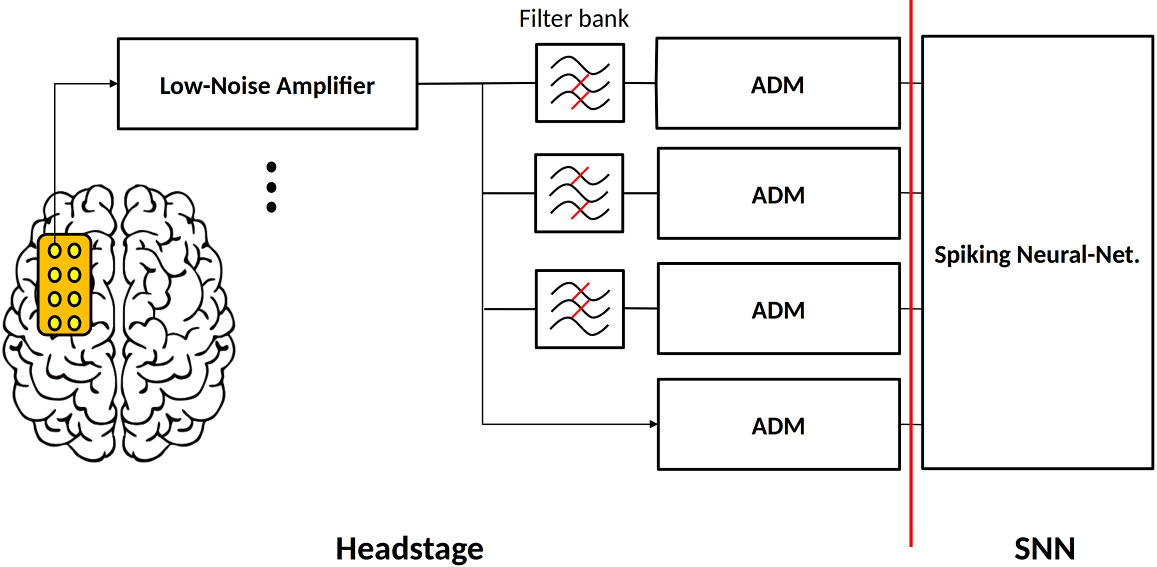
HFO are biomarkers for epileptogenic brain tissue.



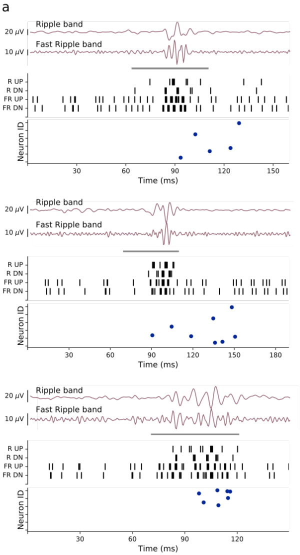


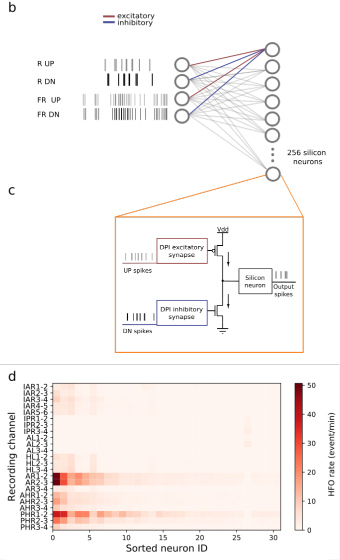
[Fedele et al., 2017]

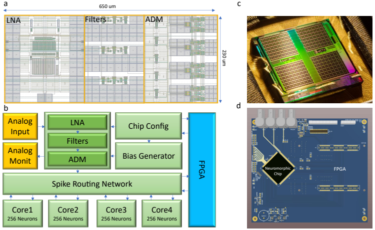
A neuromorphic HFO sensory-processing device



On-line detection of HFO



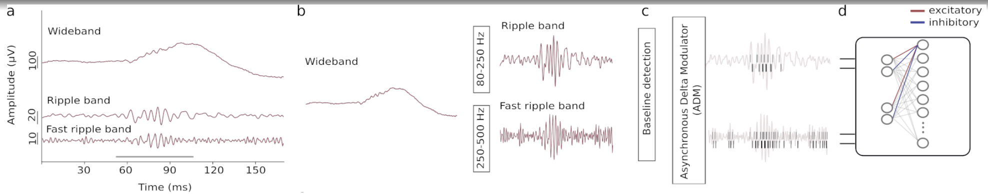


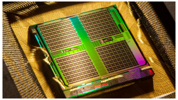


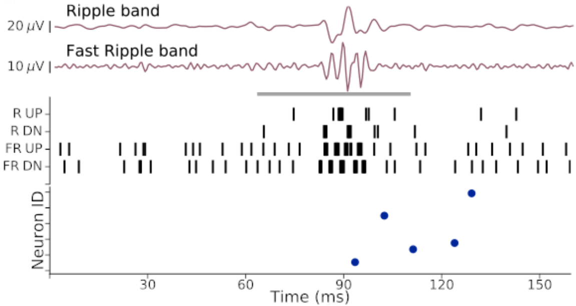
Accuracy: 78 % (vs. 67%)

Power consumption: 614 μW

[Sharifhazileh, Burelo, et al., 2021]







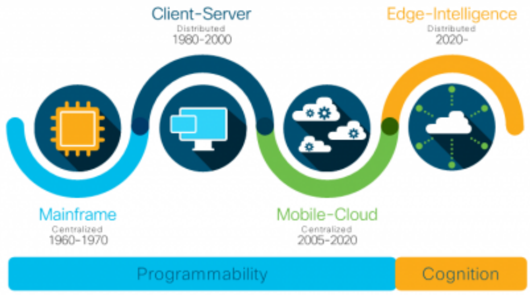
[Sharifhazileh, Burelo et al., Nat. Comms 2021]

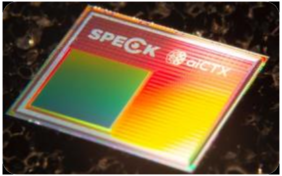
- Stand-alone sensor+processor
- 256 neurons, 512 synapses
- “Backpropless” two layer network
- One-bit weights
- Inhomogeneous parameters
- Matched time constants
- Power consumption: **614** μ**W**
- Accuracy: **78%** (vs. 67% from s-o-a)

Outline

Edge intelligence

We are now entering the era of *neuromorphic intelligence* in which dedicated cognitive "chiplets" will be used to provide intelligence to a multitude of extreme edge-computing devices





- Health monitoring
- Wearable sensors
- Environmental sensing
- Industrial monitoring
- Intelligent machine vision
- Consumer applications

http://capocaccia.cc/

- Interdisciplinary, international, diverse
- Morning lectures, afternoon hands-on work-groups
- Active and lively discussions (no powerpoint)
- Concrete results, establishment of long-term collaborations

Capo Caccia, Sardinia, Italy. April 28 - May 11, 2024

Take home message for neuromorphic intelligence

Academic/basic research

- Study real brains, start from small neural circuits/systems
- Take into account all properties of neurons and synapses
- Focus on fundamental problems (ignore incremental benchmarks)
- “There’s plenty of room at the bottom” (large scale is not all)

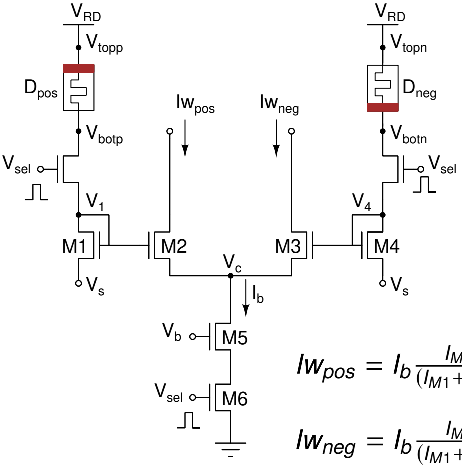
Applied/industrial research

- Choose a specific problem to solve that is not being solved yet
- Consider it’s requirements in it’s entirety, from end to end
- Be open to using the best of all possible approaches (analog and digital)
- Build the full ecosystem for your solution (devices, software, users)

Early access:

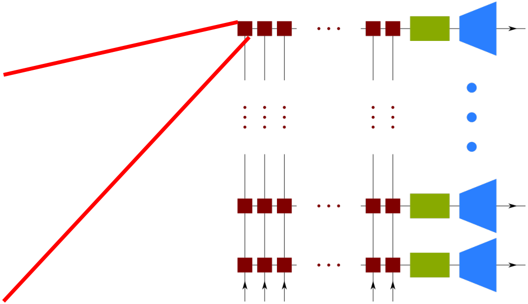
Bottom-Up and Top-Down Approaches for the Design of Neuromorphic Processing Systems: Tradeoffs and Synergies Between Natural and Artificial Intelligence, Frenkel and Indiveri, Proceedings IEEE, 2023.

A (Gilbert) normalizer memristive synapse circuit



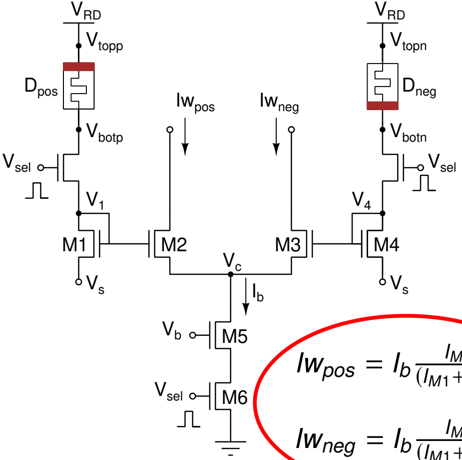
$$Iw_{pos} = I_b \frac{I_{M1}}{(I_{M1}+I_{M4})}$$

$$Iw_{neg} = I_b \frac{I_{M4}}{(I_{M1}+I_{M4})}$$



[M. Nair et al., Nano Futures, 2017; Payvand et al., Faraday Discuss., 2019]

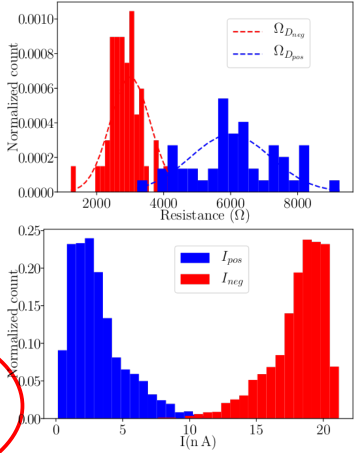
A (Gilbert) normalizer memristive synapse circuit



$$Iw_{pos} = I_b \frac{I_{M1}}{(I_{M1}+I_{M4})}$$

$$Iw_{neg} = I_b \frac{I_{M4}}{(I_{M1}+I_{M4})}$$

Divisive non-linearity "squashes" distributions and reduces mismatch effects



$CV = 0.429$

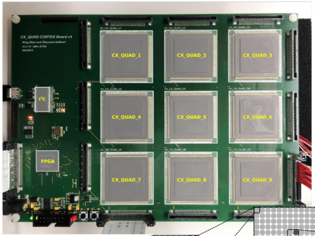
$CV = 0.284$

[M. Nair et al., Nano Futures, 2017; Payvand et al., Faraday Discuss., 2019]

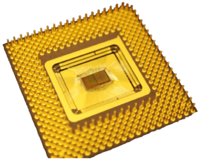
Neuromorphic processors for sensory processing

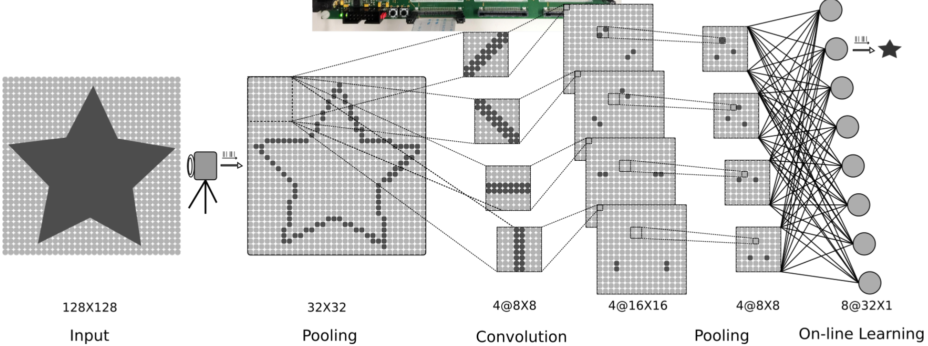
DVS

3x3 cxQuad PCB



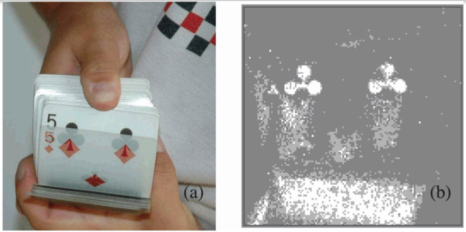
ROLLS

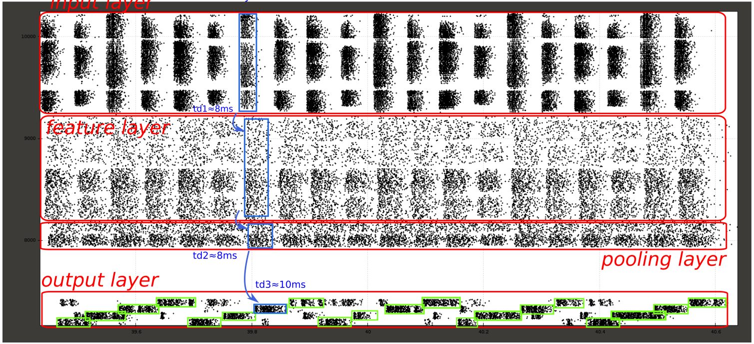


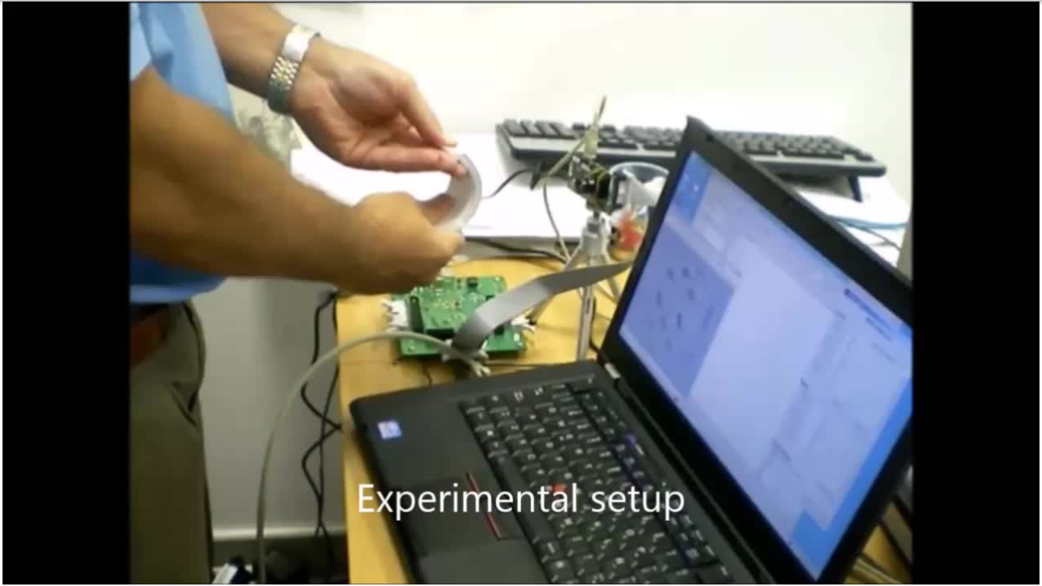


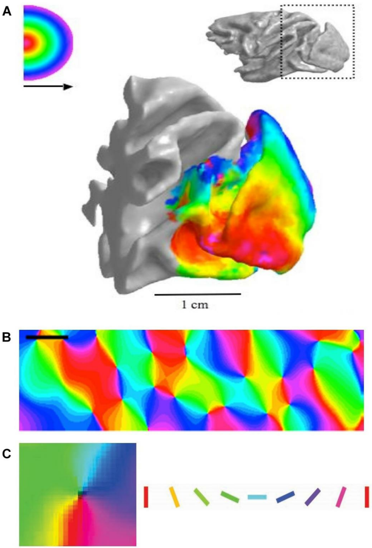
[Indiveri et al., IEDM 2015]

Real-time low-latency convolutional neural networks

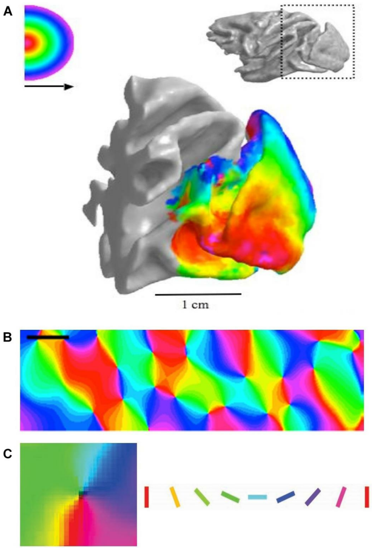




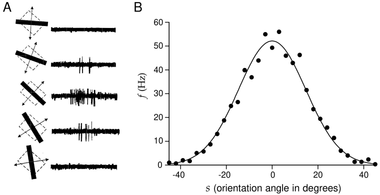

Experimental setup



⟵ Retinotopic and orientation maps representing the preference of neurons in the visual cortex for the location and orientation of a stimulus on the visual field.

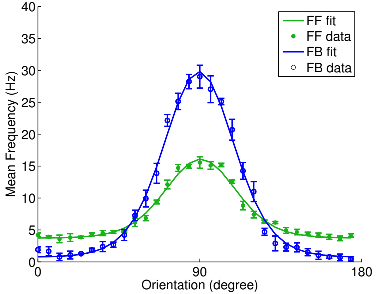


⟵ Retinotopic and orientation maps representing the preference of neurons in the visual cortex for the location and orientation of a stimulus on the visual field.



[Dayan & Abbott, 2005]

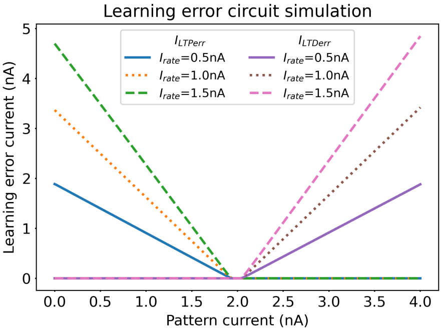
Orientation tuning: ⟶ Non-human primate response to moving bars (top); Neuromorphic processor response to flashing bars (bottom)

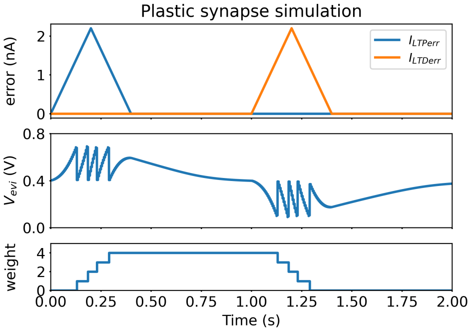


[Chicca et al., 2007]

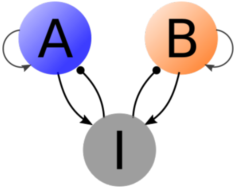
Feature tuning via populations of neurons

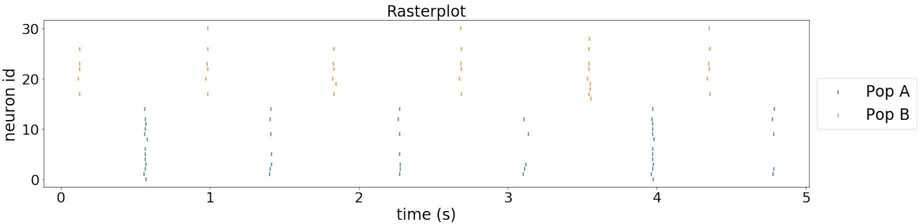
On-chip learning implementation

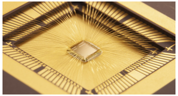


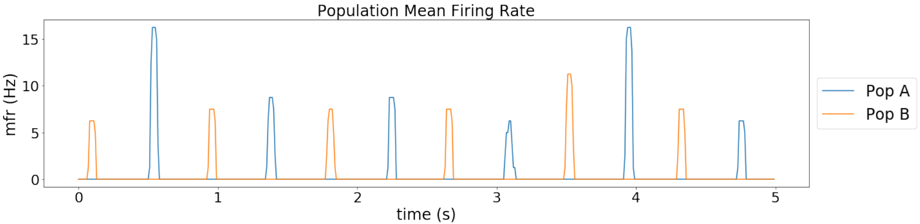


Neural oscillators and Central Pattern Generators



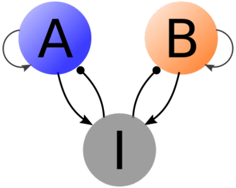


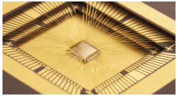




[R. Krause et al., 2021]

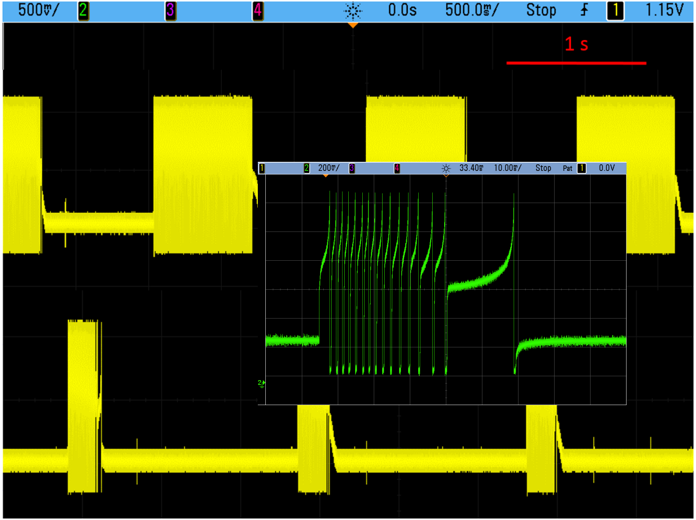
[E. Donati et al., 2021]



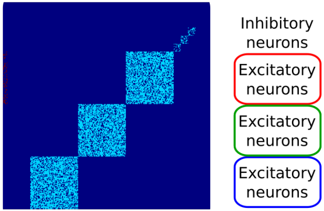


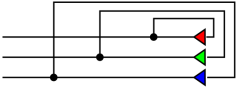
[R. Krause et al., 2021]

[E. Donati et al., 2021]

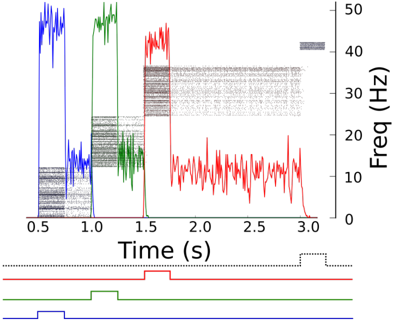


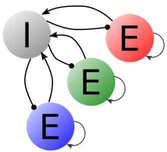
Synaptic matrix





Output spikes





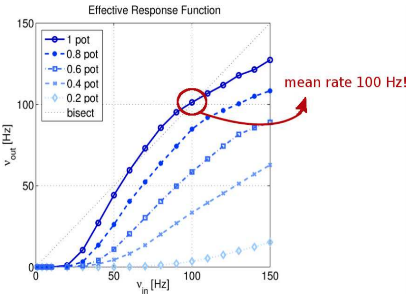
Methods and tools:

- Mean Field Theory
- Effective Response Function
- Self consistency condition

$$\tau \frac{d}{dt}\mu = -\mu + W\nu_{in} - \beta$$

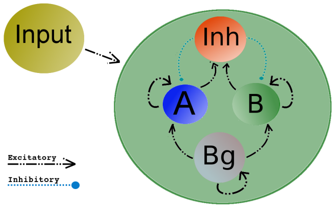
$$\nu_{out} = \Phi(\mu(\nu_{in}), \sigma)$$

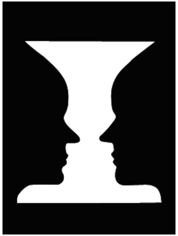
$$\nu_{out} = \nu_{in} = \nu$$

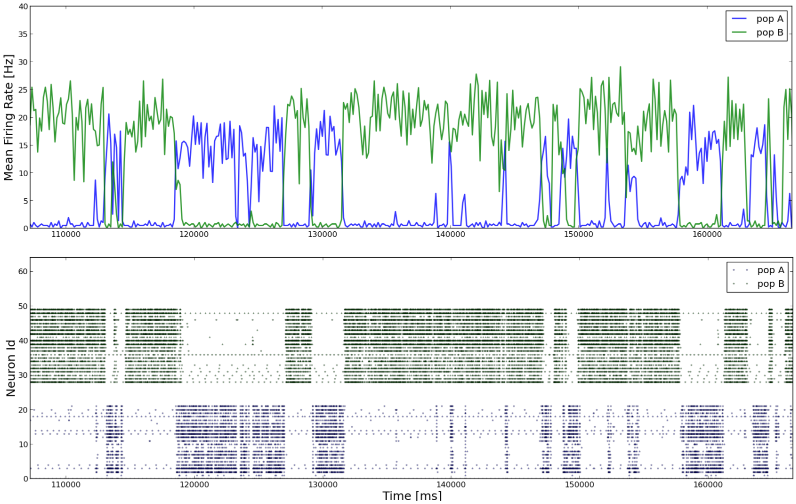


[M. Giulioni et al., 2012, M. Giulioni et al., 2015]

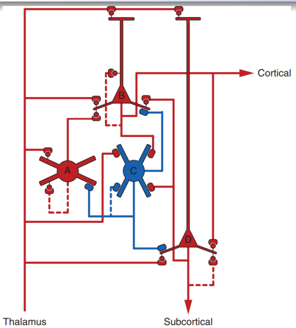
Excitation-Inhibition balanced networks

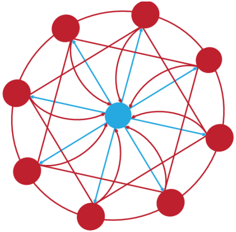


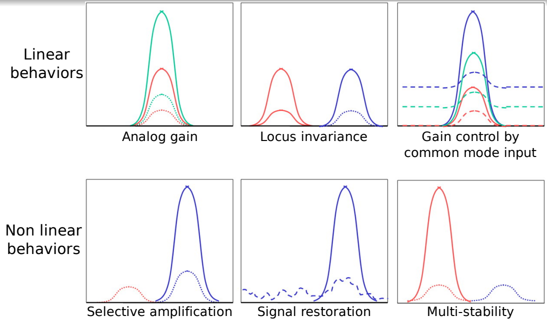


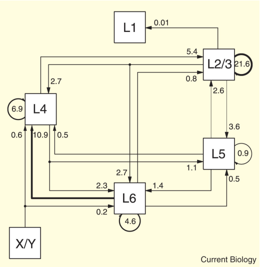


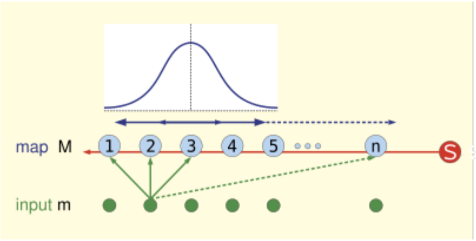
The canonical micro-circuit as a soft-WTA

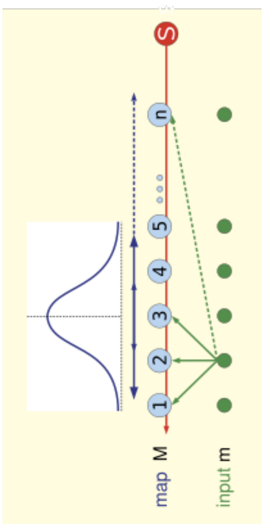


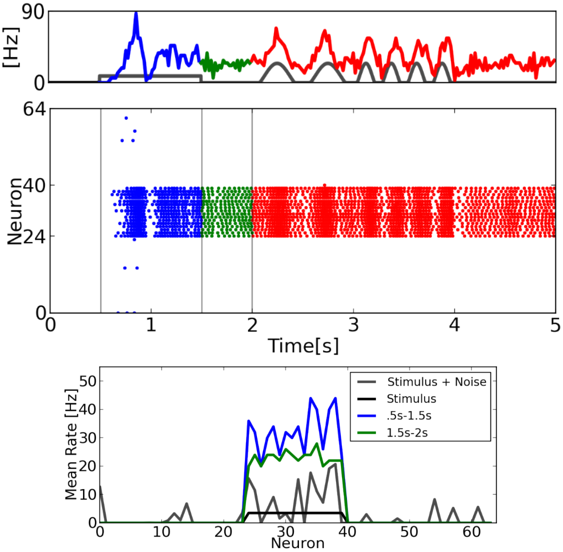






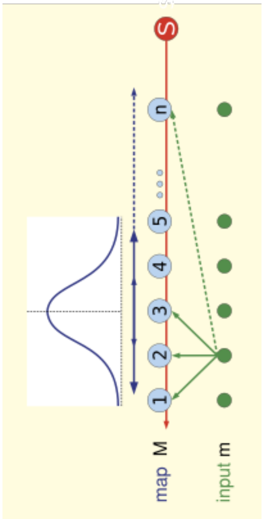


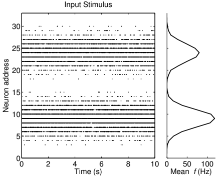




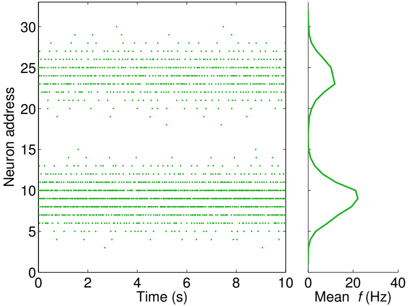
[Neftci et al., 2013]

Soft Winner-Take-All networks

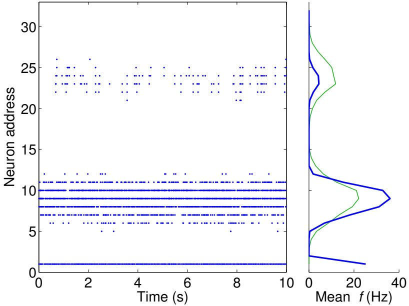


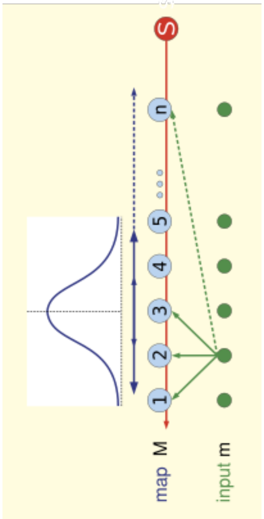


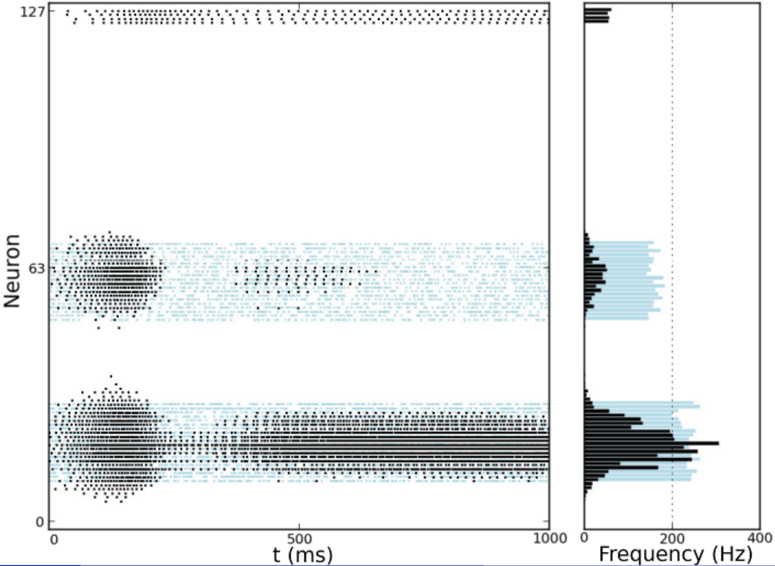
Feedforward Network



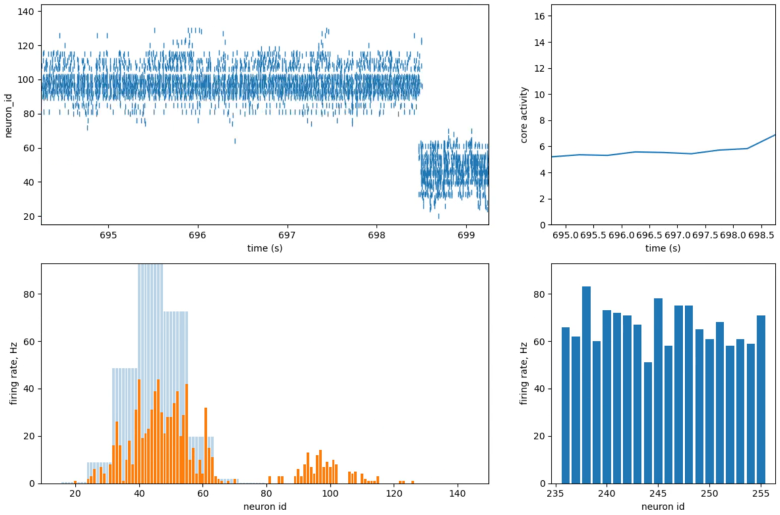
Feedback Network




S
n
5
4
3
2
1
map M
input m

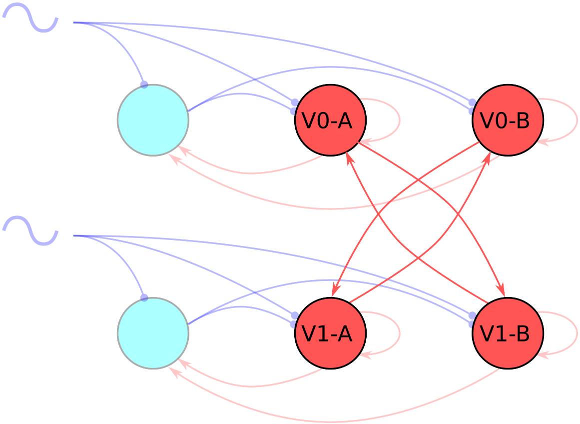

127
63
0
Neuron
0
500
1000
t (ms)
0
200
400
Frequency (Hz)

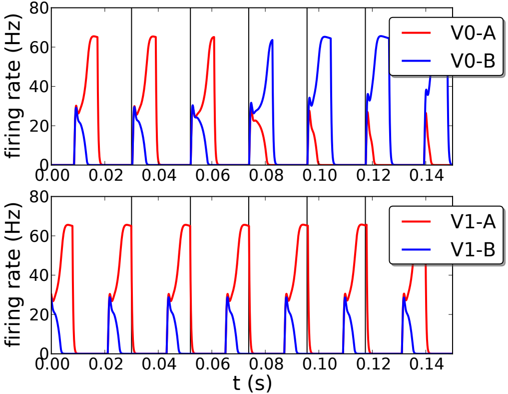
Soft Winner-Take-All drifting



[Zendrikov et al., 2023]

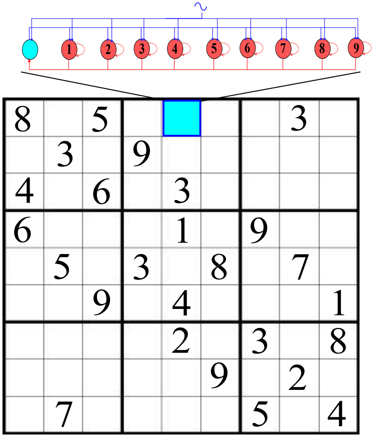
Binary variables V0,V1: V0≠V1

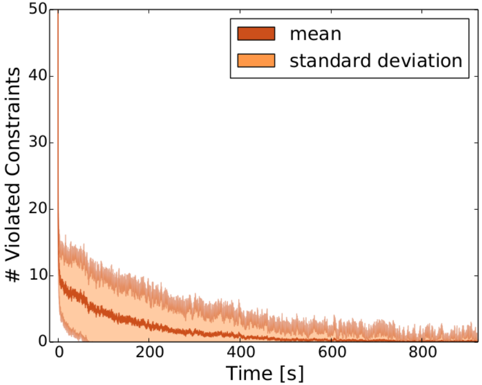




In absence of external input (evidence), the network settles to the lowest energy state (all constraints satisfied).

[Mostafa et al., 2015]



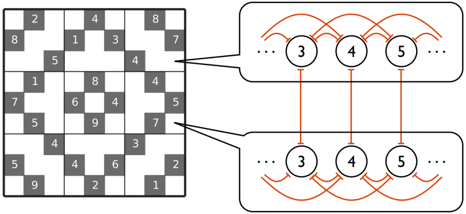


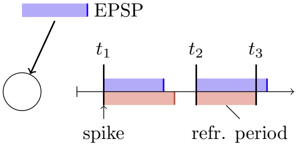
Can be applied to all Boolean satisfiability problems, such as graph coloring problem, SAT, etc.

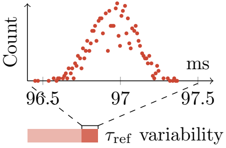
[Mostafa et al., 2015]

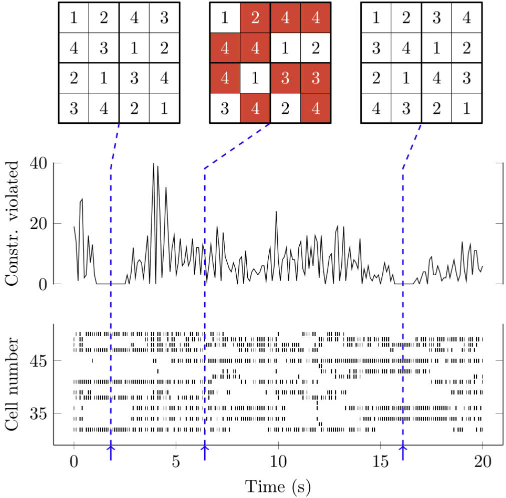
WTA networks and neural sampling

Exploiting the device mismatch in the neuron's refractory period.

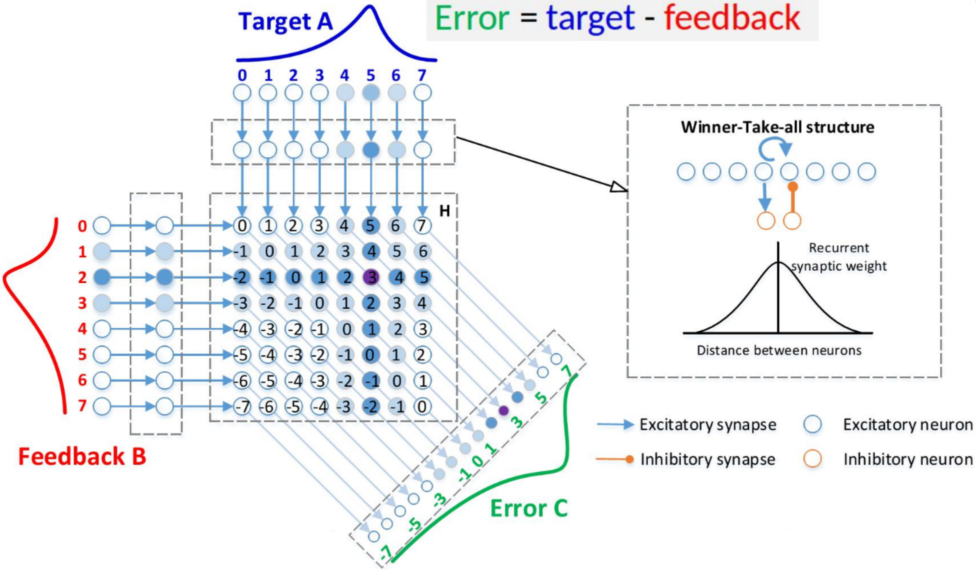






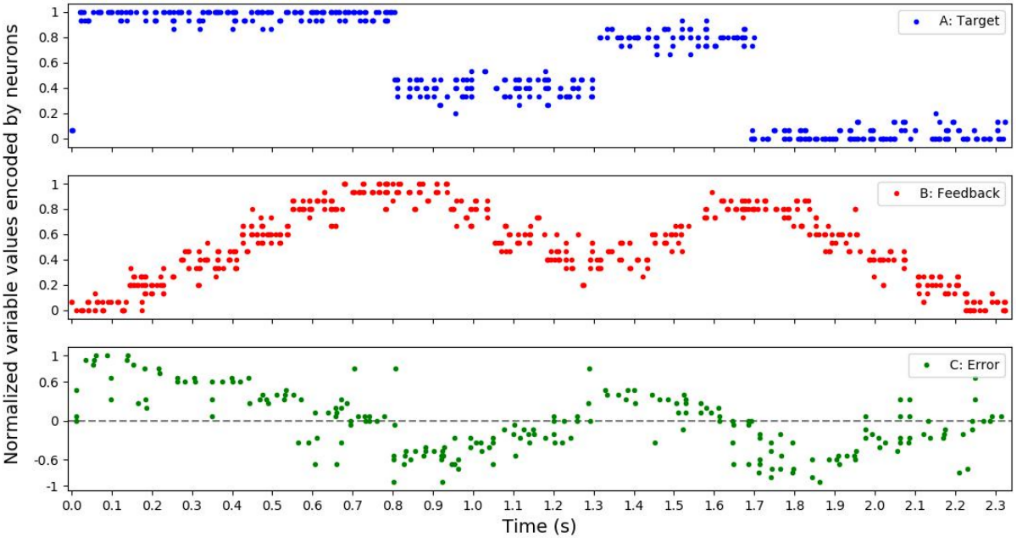


[Bias et al., 2016]



[J. Zhao et al., 2020]

Coupling multiple WTAs to process variables



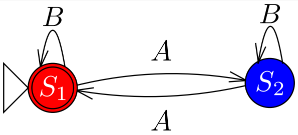
16 neurons

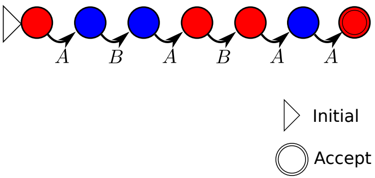
16 neurons

31 neurons

Finite State Machines vs Neural State Machines

A finite-state machine (FSM) is a mathematical model of computation used to design both computer programs and sequential logic circuits. It is conceived as an abstract machine that can be in one of a finite number of states. [Wikipedia]





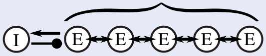
- Recognizes regular expression $B^*[AB^*A]^*$

[Minsky, 1967]

Single WTA

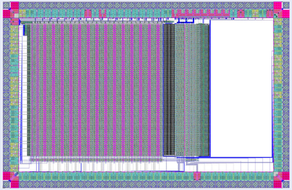
Inhibitory neurons

Excitatory neurons

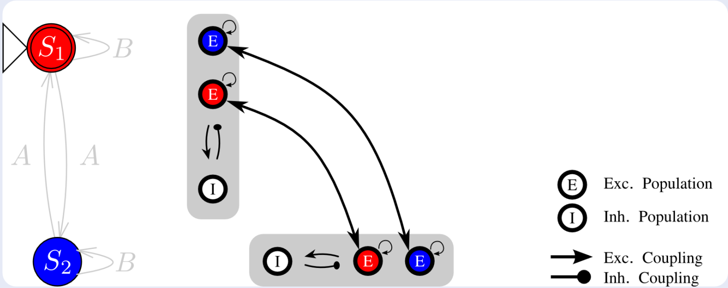


Global Inhibition

Nearest-N Excitation



Coupled WTAs

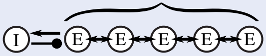


[R. Rutishauser & R.J. Douglas, 2009, R. Rutishauser et al., 2011, E. Neftci et al., 2013]

Single WTA

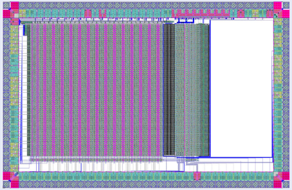
Inhibitory neurons

Excitatory neurons

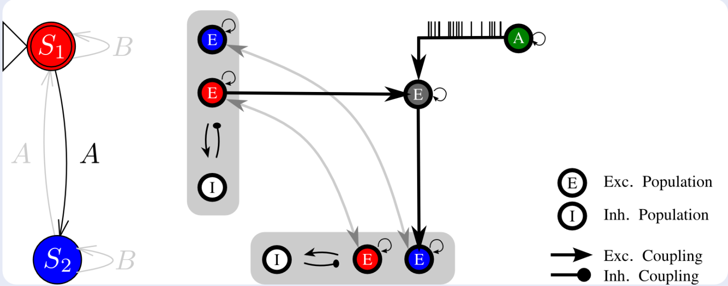


Global Inhibition

Nearest-N Excitation



Coupled WTAs



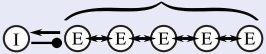
[R. Rutishauser & R.J. Douglas, 2009, R. Rutishauser et al., 2011, E. Neftci et al., 2013]

Neural State Machines: multi-chip implementation

Single WTA

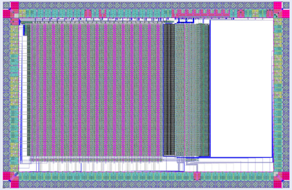
Inhibitory neurons

Excitatory neurons

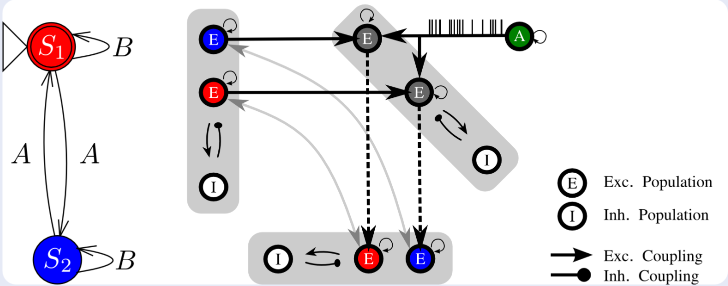


Global Inhibition

Nearest-N Excitation

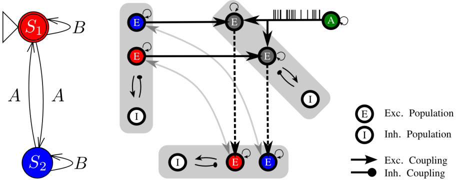


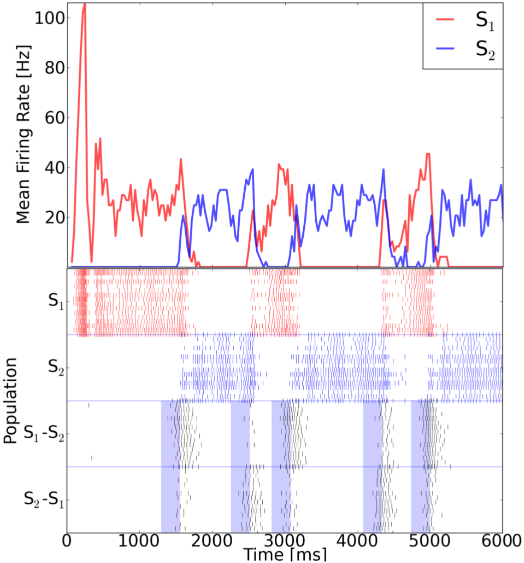
Coupled WTAs



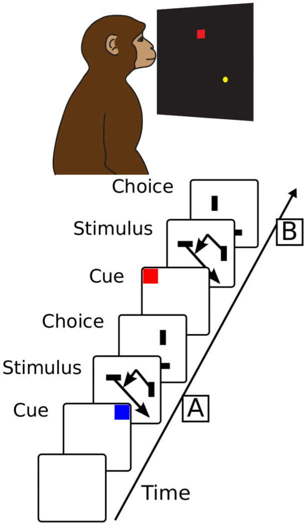
[R. Rutishauser & R.J. Douglas, 2009, R. Rutishauser et al., 2011, E. Neftci et al., 2013]

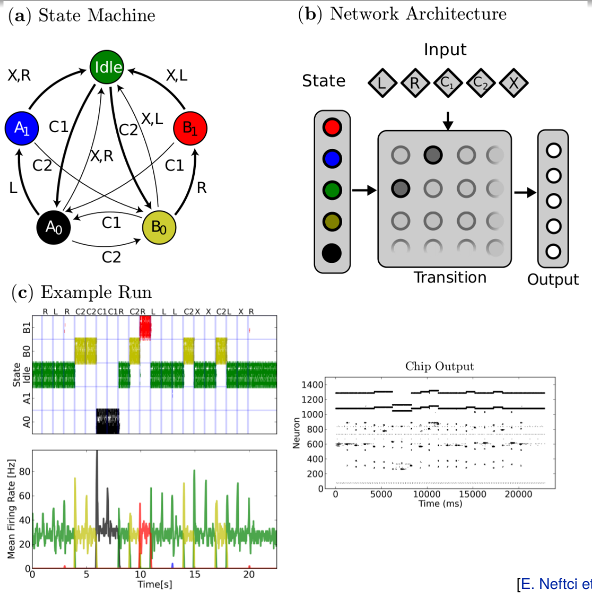
Neural State Machines: experimental results



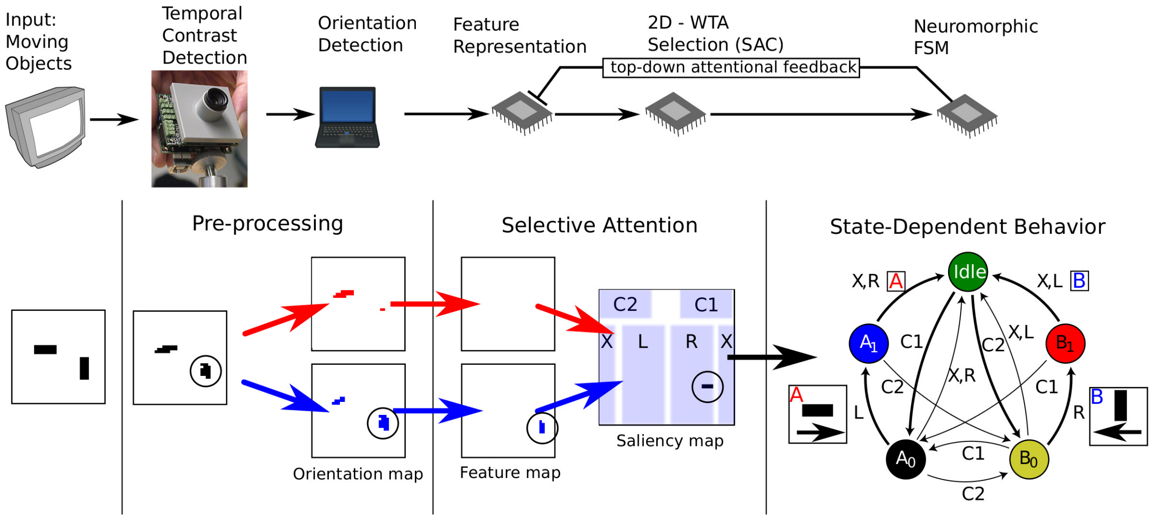


Synthesizing neuromorphic cognitive system



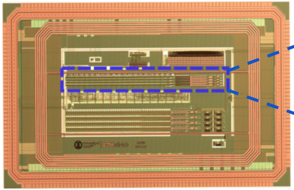


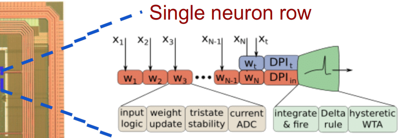
[E. Neftci et al., 2013]

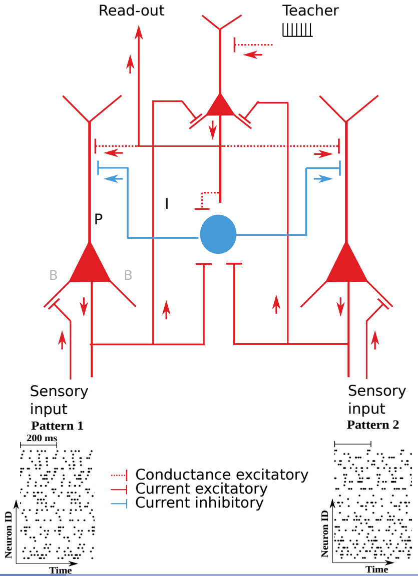

Input:
Moving
Objects
Temporal
Contrast
Detection
Orientation
Detection
Feature
Representation
2D - WTA
Selection (SAC)
Neuromorphic
FSM
top-down attentional feedback
Pre-processing
Selective Attention
State-Dependent Behavior
C2
C1
X
L
R
X
Saliency map
Orientation map
Feature map
Idle
X,R
A
X,L
B
A1
B1
A0
B0
C1
C2
X,R
X,L
C2
C1
L
R
C1
C2

Robust on-chip learning

- Bi-stable synapses with STDP circuits
  [Indiveri et al, 2006]
- Spike-driven synaptic plasticity with stop-learning
  [Mitra et al, 2009,Qiao et al., 2015]
- Error-propagation with local learning
  [Cartiglia et al., 2020]
- Dendritic Hebbian synaptic plasticity with stop-learning
  [Rubino et al, 2023]

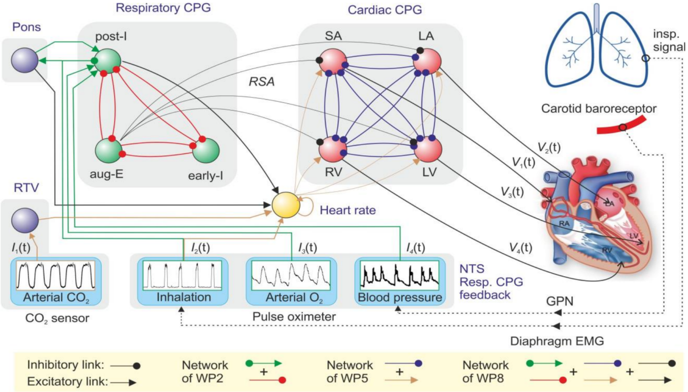






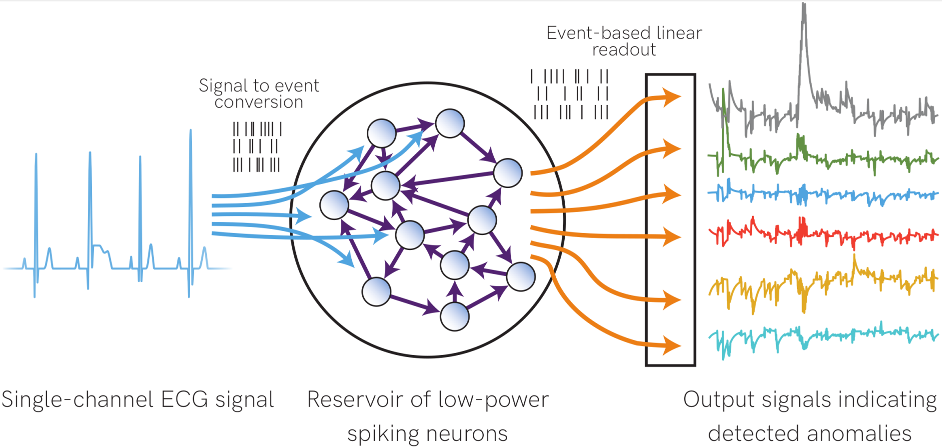
CResPace: Adaptive pace-maker

Build an adaptive pacemaker that responds to physiological feedback in real time to recover heart rate adaptation functionality.



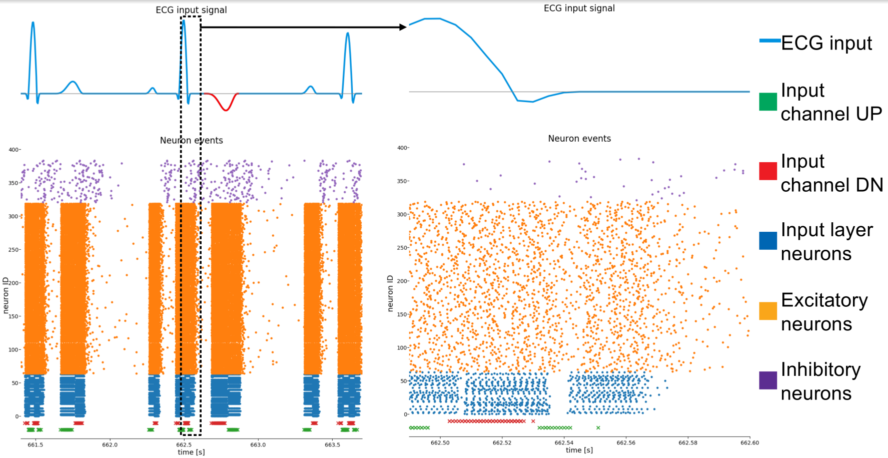
(Elisa Donati, Renate Krause)

ECG anomaly detection using reservoir computing



[H. Jaeger, 2003] [W. Maass et al., 2002] [F. Bauer and D. Muir, SynSense]

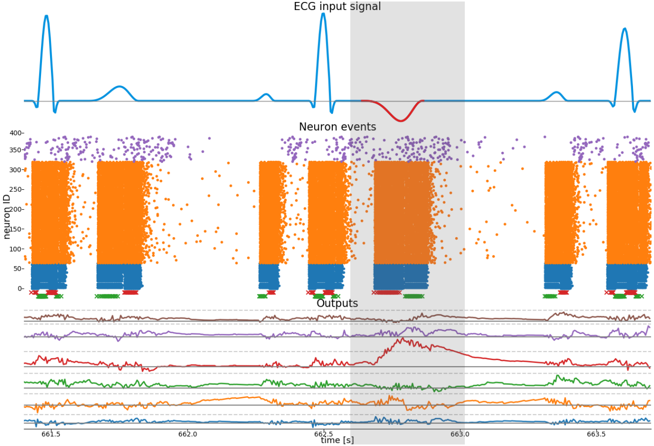
ECG anomaly detection using reservoir computing



[H. Jaeger, 2003] [W. Maass et al., 2002] [F. Bauer and D. Muir, SynSense]

ECG anomaly detection using reservoir computing

- Generic, single-led ECG
- Six different anomaly types
- One read-out unit per anomaly

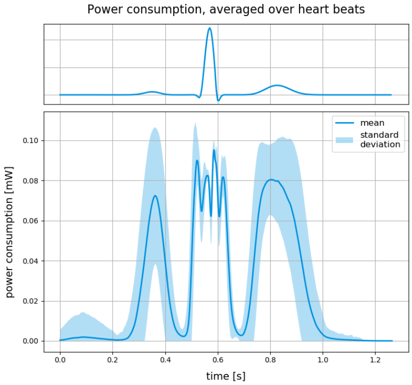


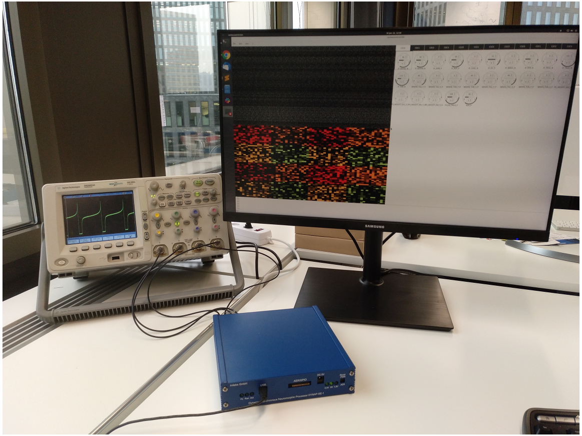
True positives rate (specificity): 91.3%

True negative rate (sensitivity): 97.6%

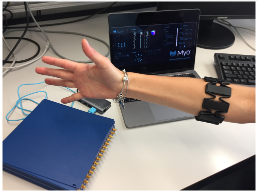
[F. Bauer et al., 2019]

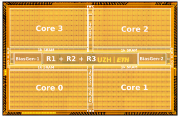
Mean neural event rate: $14.8 \cdot 10^3$/s
Mean synaptic event rate: $787.6 \cdot 10^3$/s
Energy per neural event: 100 pJ
Energy per synaptic event: 40 pJ
Mean power consumption: $< 500\,\mu$W

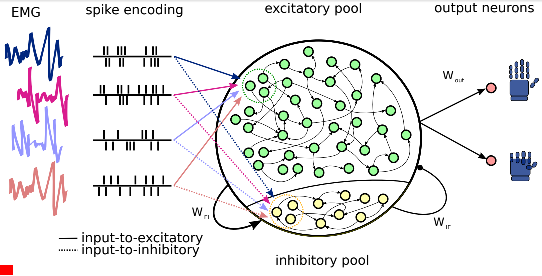


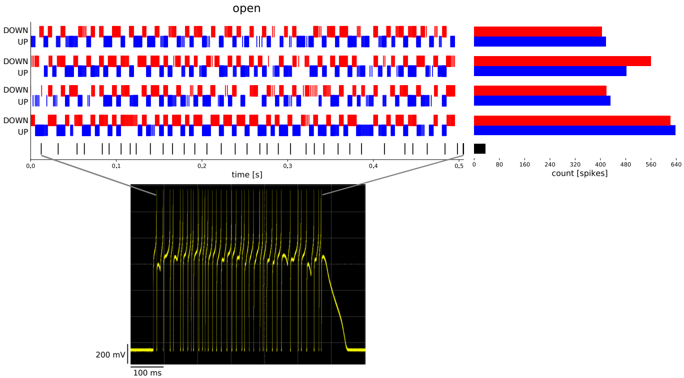


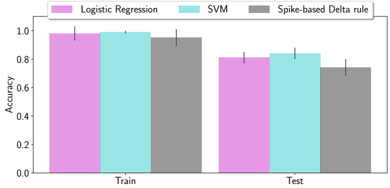
On-line classification of EMG signals



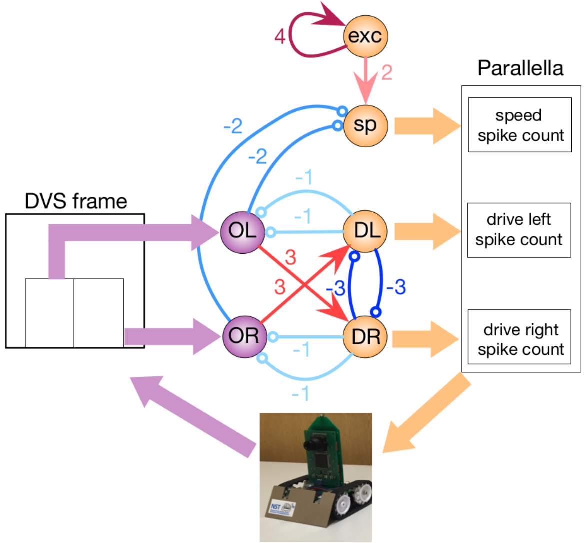


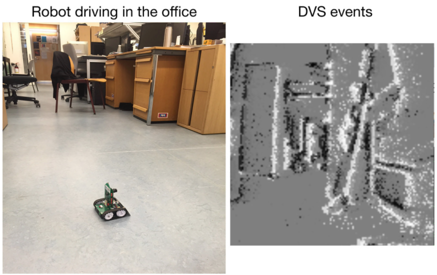






[Donati et al., 2019]





[R. Kreiser et al, Frontiers in Neuromorphic Eng., 2018]

[M. Milde et al., Frontiers in Neurorobotics, 2017]

[H. Blum et al., RSS, 2017]

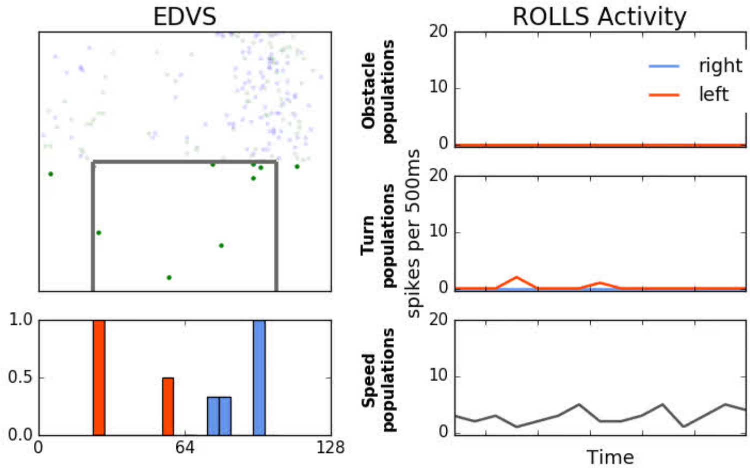
[R. Kreiser et al., ISCAS, 2017]

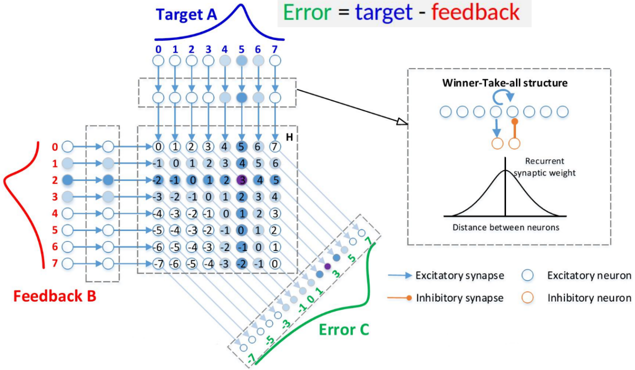
[M. Milde et al., ISCAS 2017]

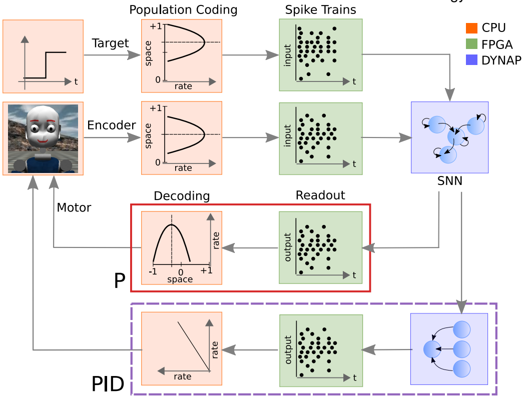
[R. Kreiser et al., IROS, 2018]

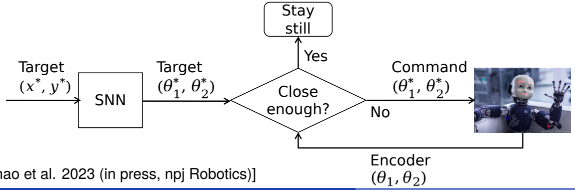
[S. Glatz et al., arXiv:1810.10801, 2018]


Camera


EDVS
1.0
0.5
0.0
0
64
128
ROLLS Activity
Obstacle populations
Turn populations
Speed populations
spikes per 500ms
right
left
20
10
0
Time

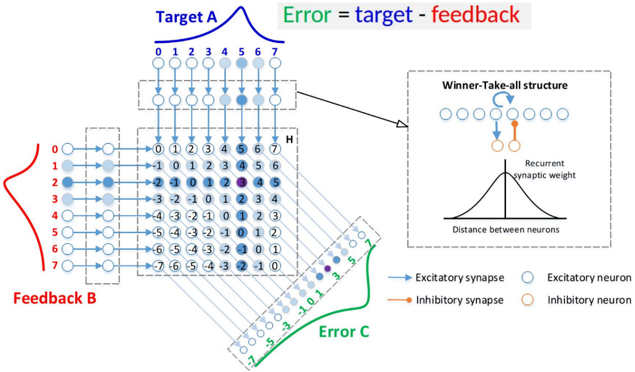


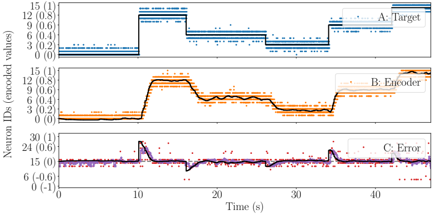


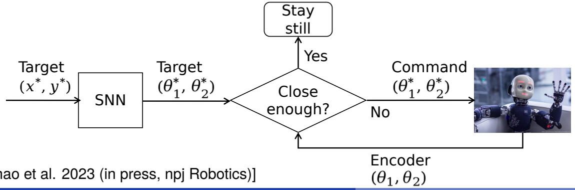


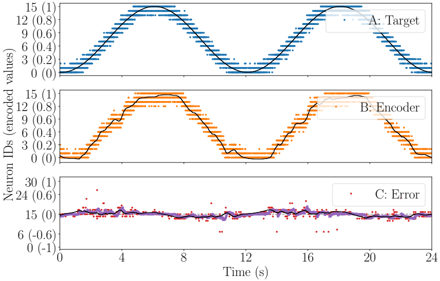
[Zhao et al. 2023 (in press, npj Robotics)]

Application to robotic control





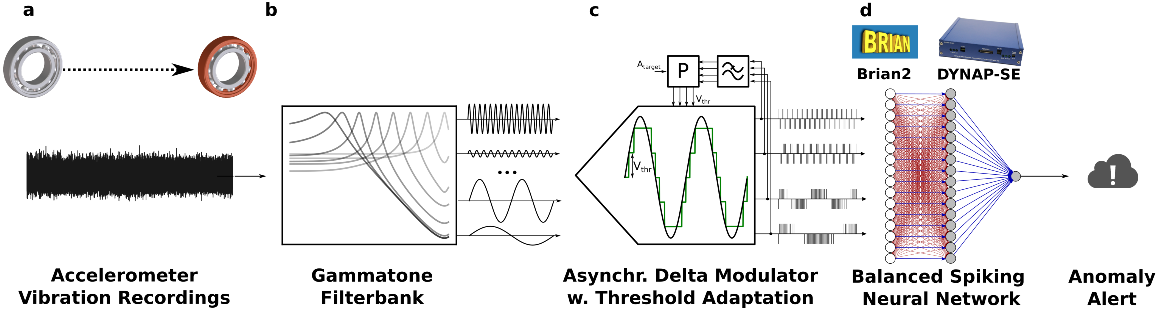


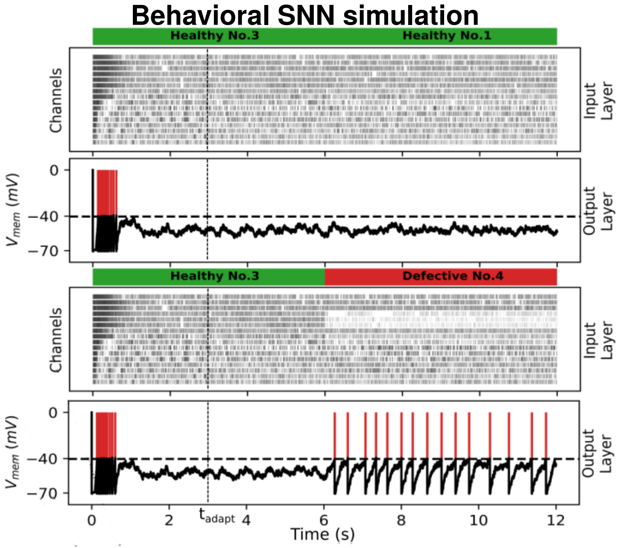


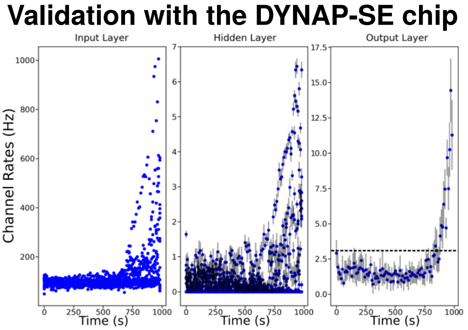
[Zhao et al. 2023 (in press, npj Robotics)]

Industrial Predictive Maintenance (PM)

- Predictive Maintenance involves the health monitoring of a degrading system.
- Vibration patterns yield valuable information about the health state of a running machine.
- PM is typically applied to large industrial tasks, but could be useful for small appliances and robots as well.







DETECTION TIMES (DATAPOINT) FOR RUN-TO-FAILUREDATASET

	b1	b2	b3	b4
LSSVM	**533**	**823**	893	700
AEC	547	-	-	-
This work	543	890	**873**	**683**

[Dennler et al., 2021]