tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

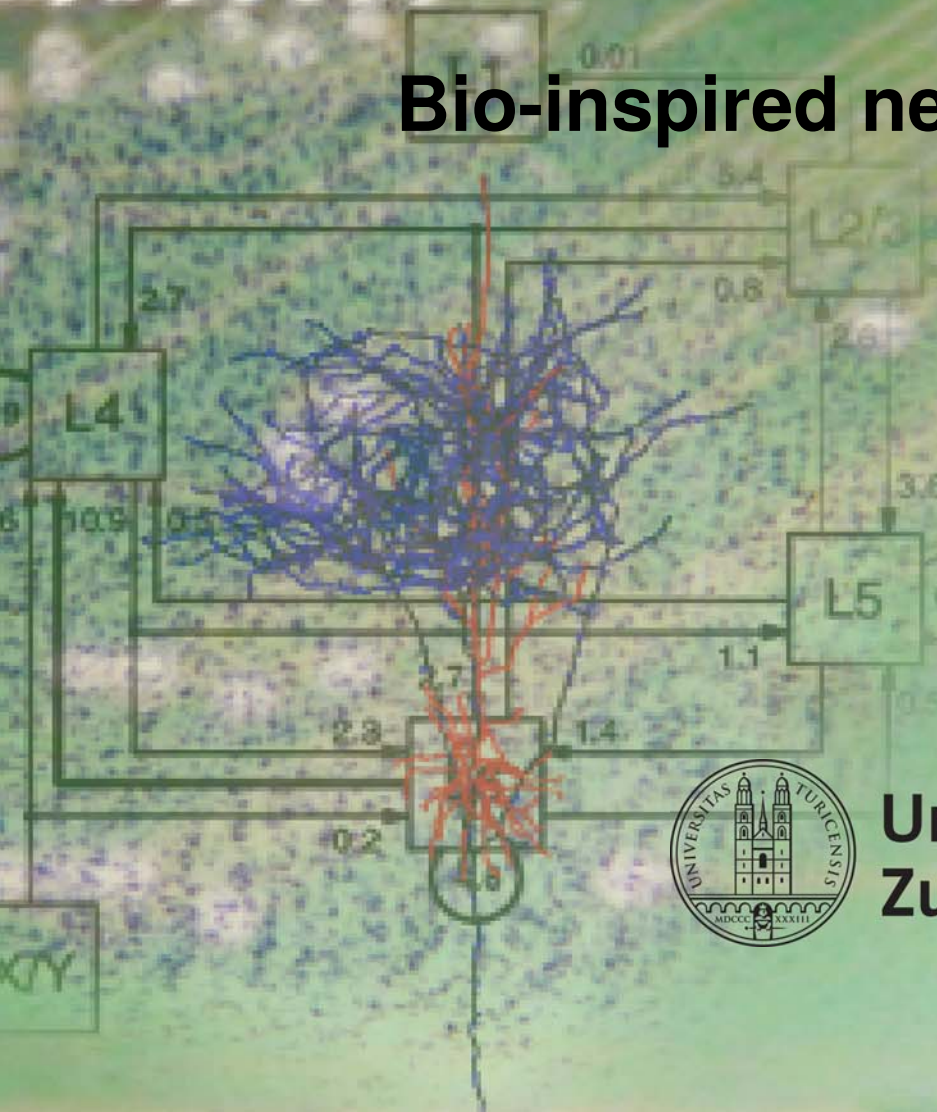
tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

www.tinyML.org

Bio-inspired neuromorphic circuits architectures



Giacomo Indiveri

Institute of Neuroinformatics
University of Zurich and ETH Zurich

June 7, 2021

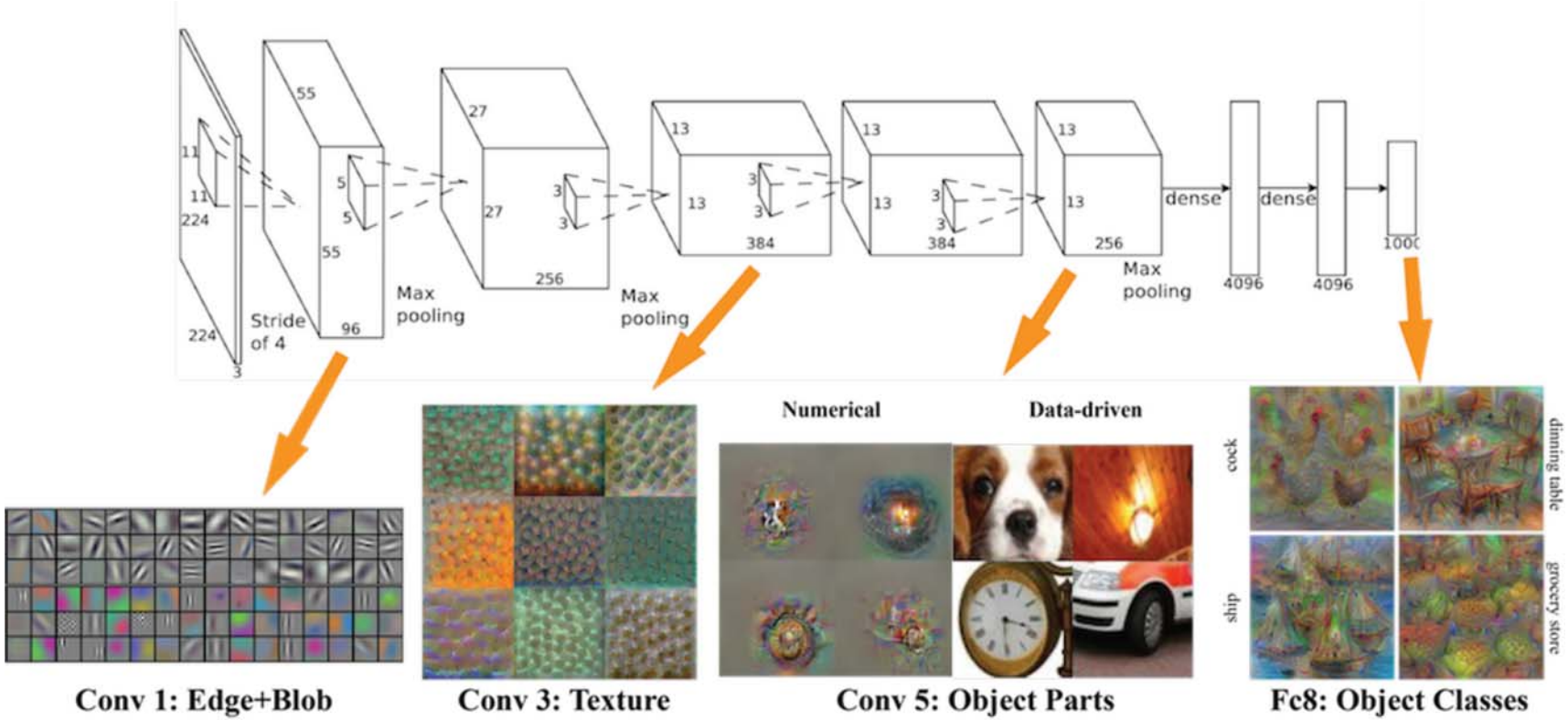
TINY ML

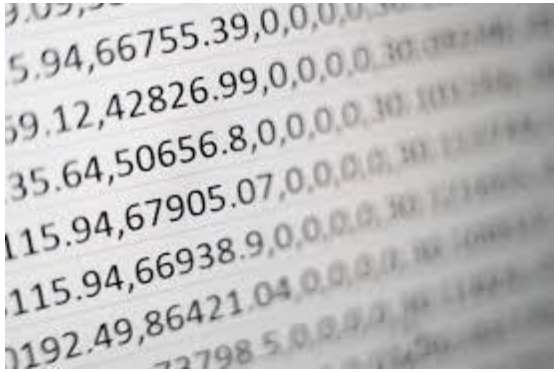
University of Zurich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Artificial Intelligence: The Deep Network revolution

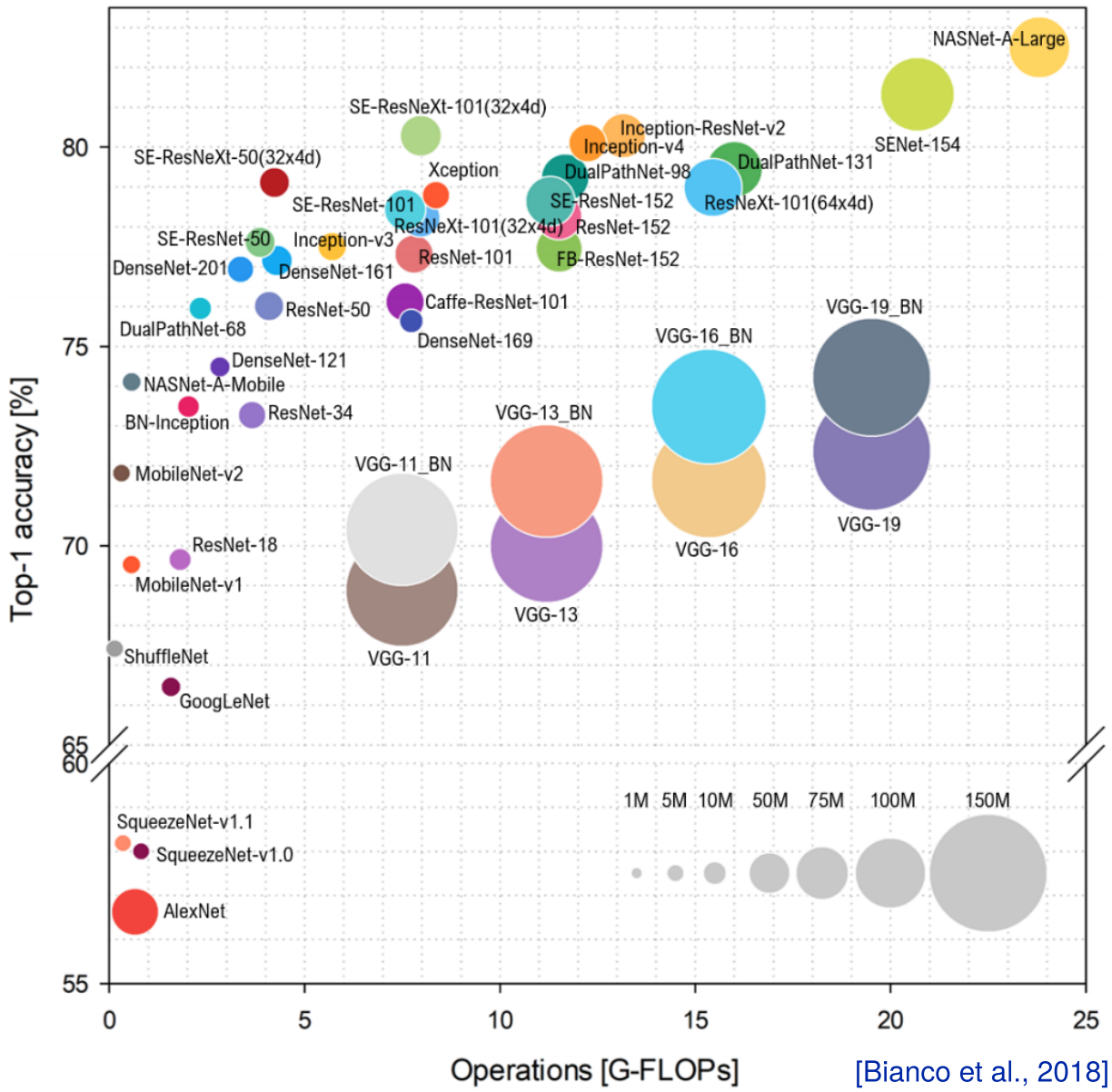
- Although the first successes of ANNs were first demonstrated in the 1980's they only started to outperform classical optimization and engineering approaches from 2009 on.
- In 2011 CNNs trained using backproagation on GPUs achieved for the first time superhuman performance in a visual pattern recognition contest.





[T. Sejnowski, 2018]

Deep Networks galore



[Bianco et al., 2018]

- As CNNs and DNNs outperformed classical approaches many research groups started to extend and optimize them.
- The AI field is now (mostly) dominated by attempts to improve accuracy on standard benchmarks, mostly by scaling up network size and parameter count.
- GPT-3 is a network with 175-billion parameters. It's memory size is exceeding 350GB, and training it requires an estimated $12 million.

Problems and limitations of Artificial Intelligence

Energy Intensive

At this pace, by 2025 the ICT industry will consume 20% of the entire world's electricity

[International Renewable Energy Agency, Internet of Things innovation landscape brief]

High cost of data movement

DRAM access is at least 1500x more costly than a MAC operation in CNN accelerators. [Tu et al., 2018]

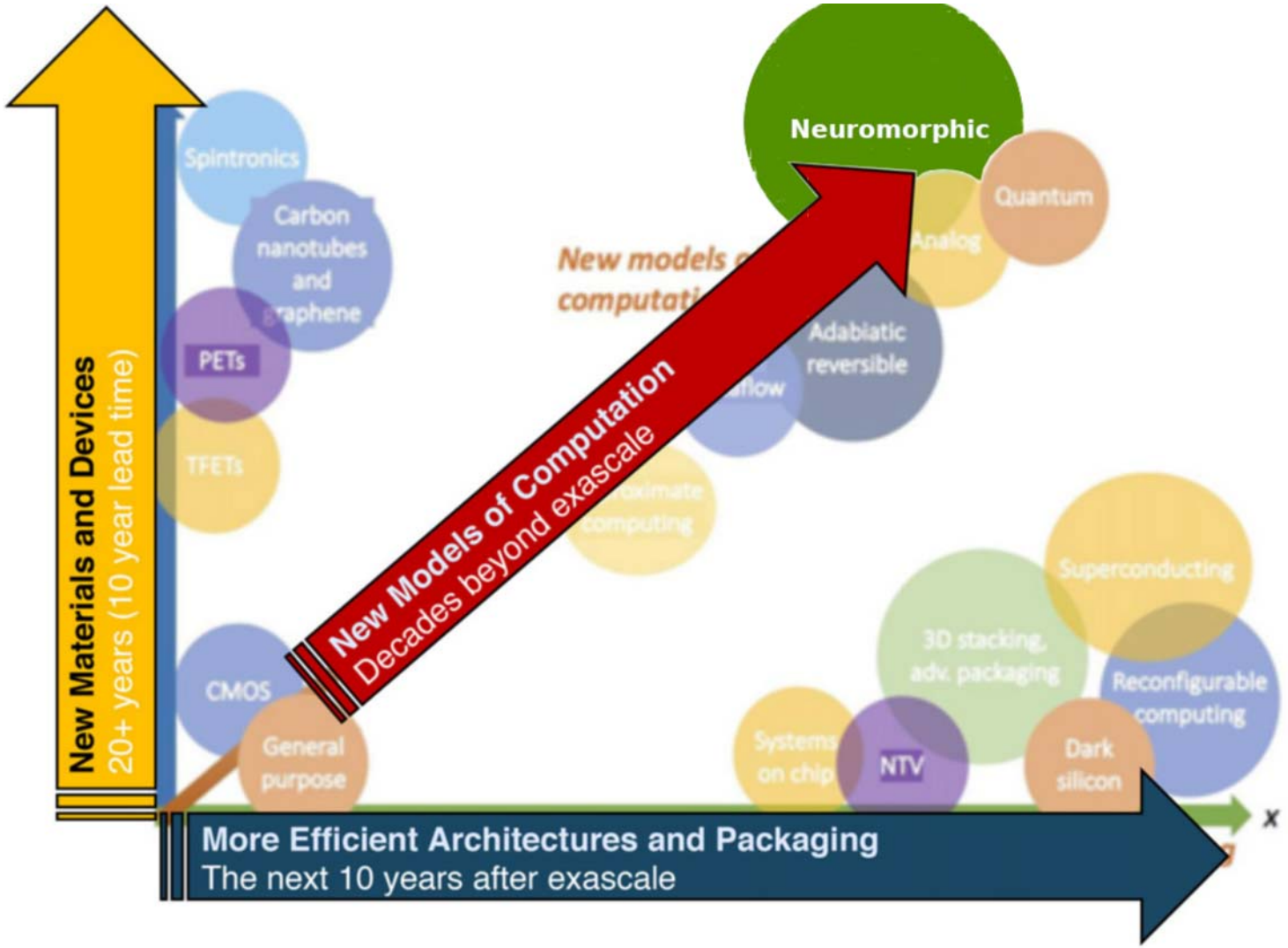
Narrow AI

DNNs programmed to perform a limited set of tasks.They operate within a pre-determined, pre-defined range. [medium.com]

Algorithmic limitations

[I am] deeply suspicious of back-propagation.
I don't think it's how the brain works.
The future depends on some graduate student who is deeply suspicious of everything I have said. [Geoff Hinton]

Paths forward to performance improvements

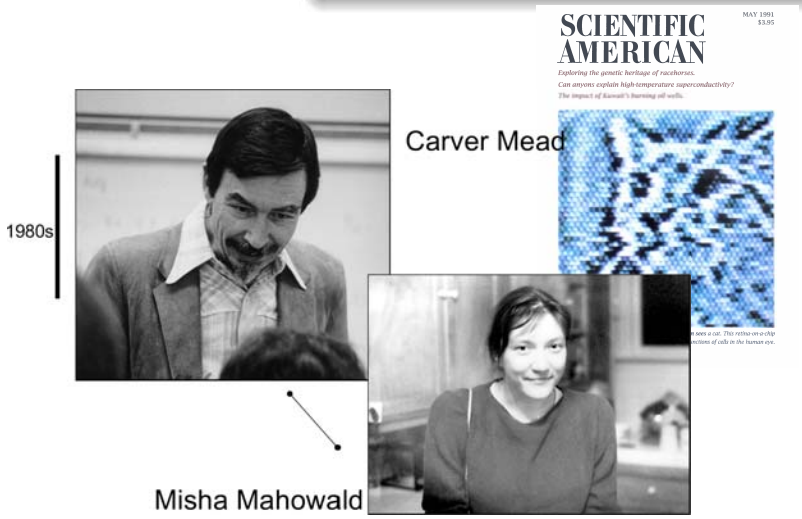


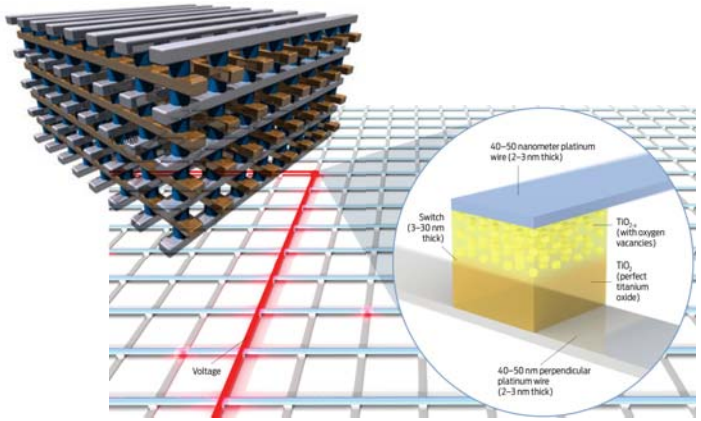
[J. Shalf, The future of computing beyond Moore's Law, 2020]

Neuromorphic Intelligence: bridging multiple communities

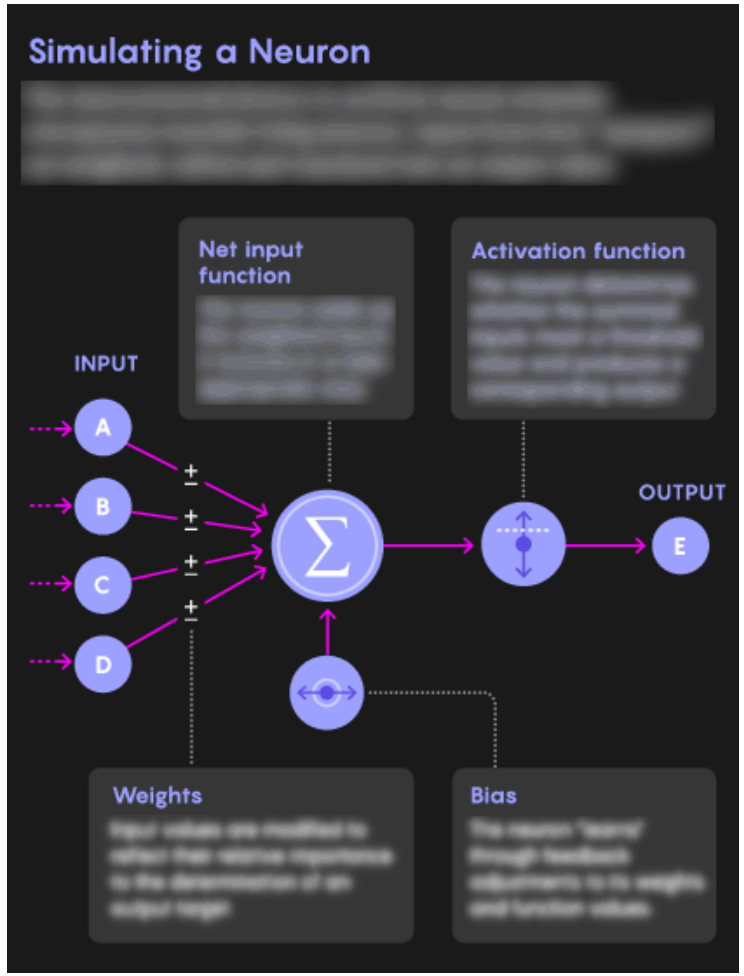
Neuromorphic Intelligence

- Deeply rooted in neuro-biology and neuroscience
- Employs the physics of both silicon and memristive devices to directly emulate neural computation
- Combines analog, asynchronous digital, and logic circuits
- Yelds application-specific devices optimal for edge-computing

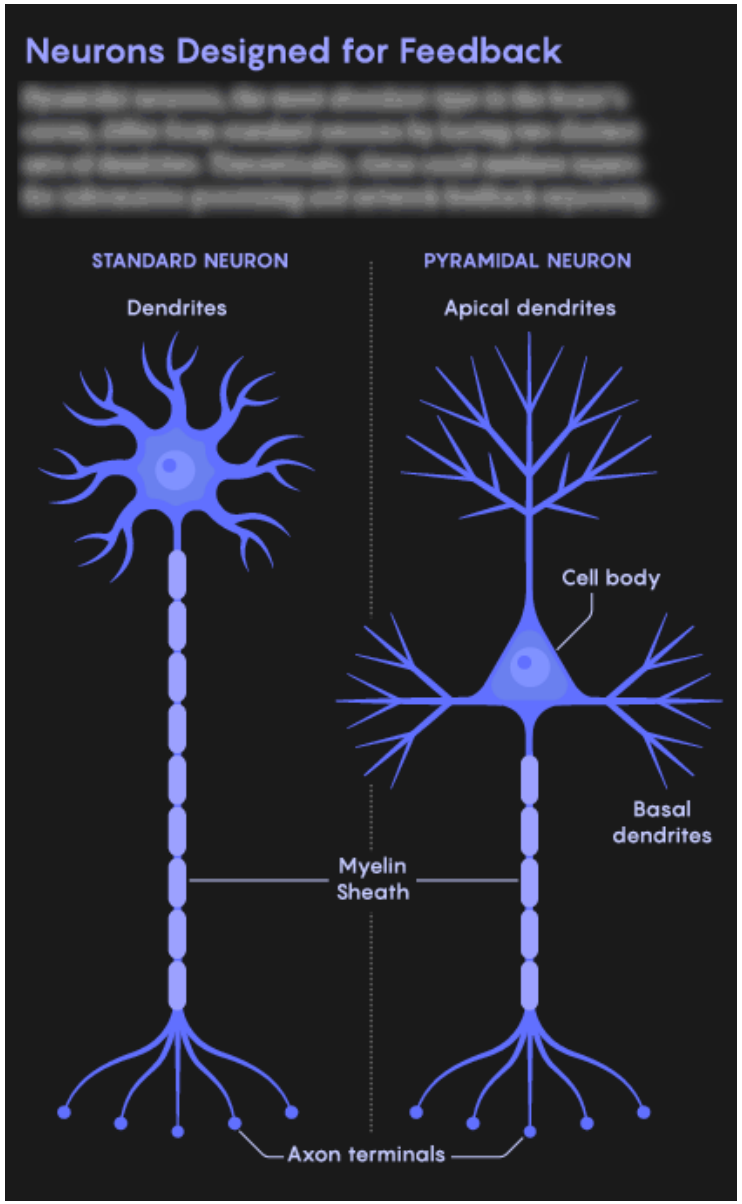




Artificial vs biological neural networks



Source: Quanta Magazine



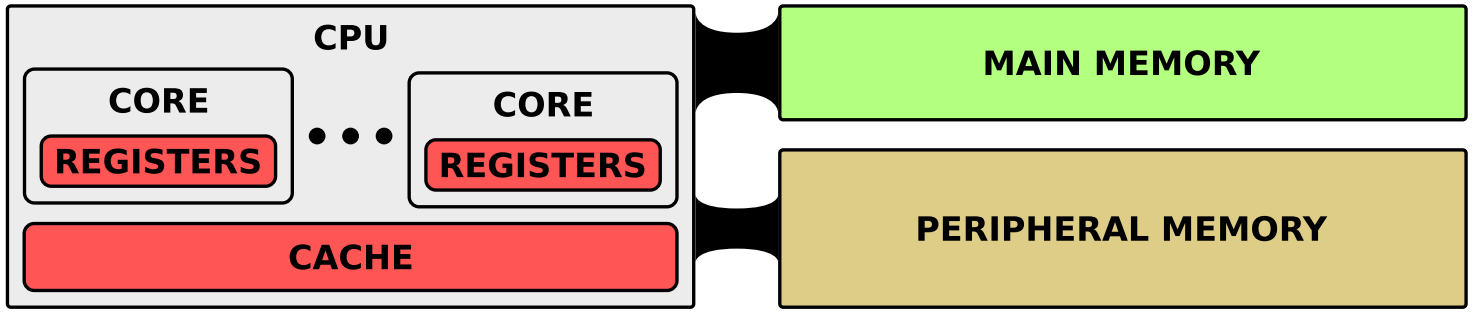
Artificial neural networks

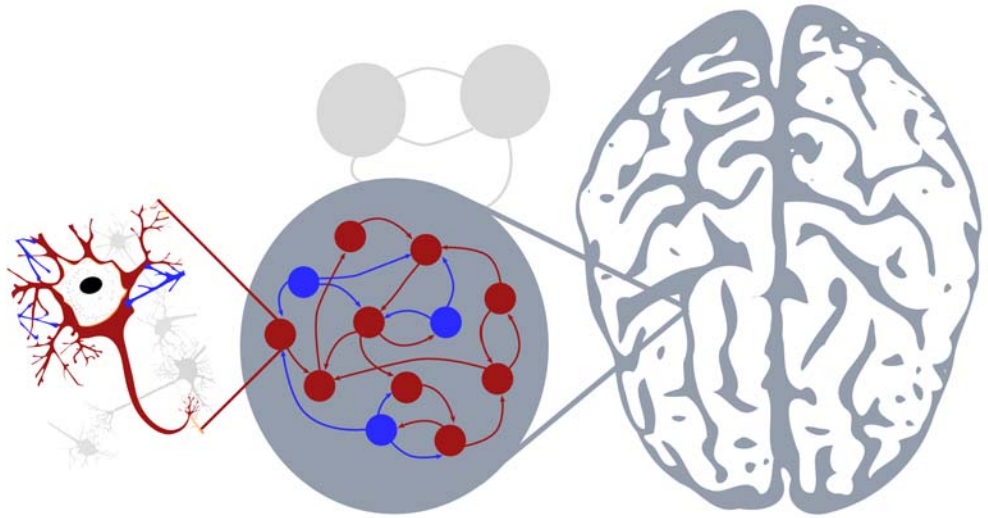
are algorithms that simulate abstract brain-inspired computing architectures using digital, time-multiplexed computing hardware.

Biological neural networks

use the time evolution of the physical elements in the system, and their dynamics, to implement computation.
The physical hardware substrate IS the algorithm.

Brain-inspired computing: a radical paradigm shift



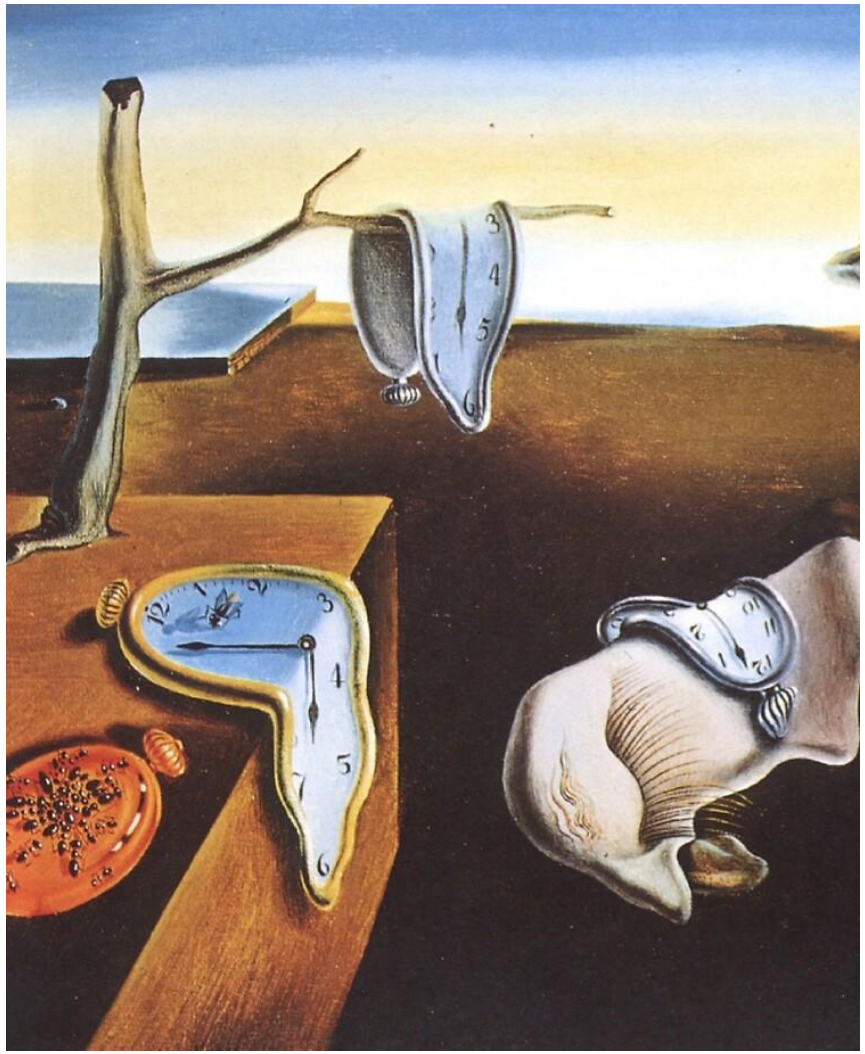


Exploit physical space

- Use the physics of the electronic devices to emulate neural dynamics
  - Exploit *all* the properties of transistors and memristors
- Use parallel arrays of processing elements
  - Maximize fine grain parallelism (no time-multiplexing)
  - Co-localize memory and computation

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

Brain-inspired computing: a radical paradigm shift

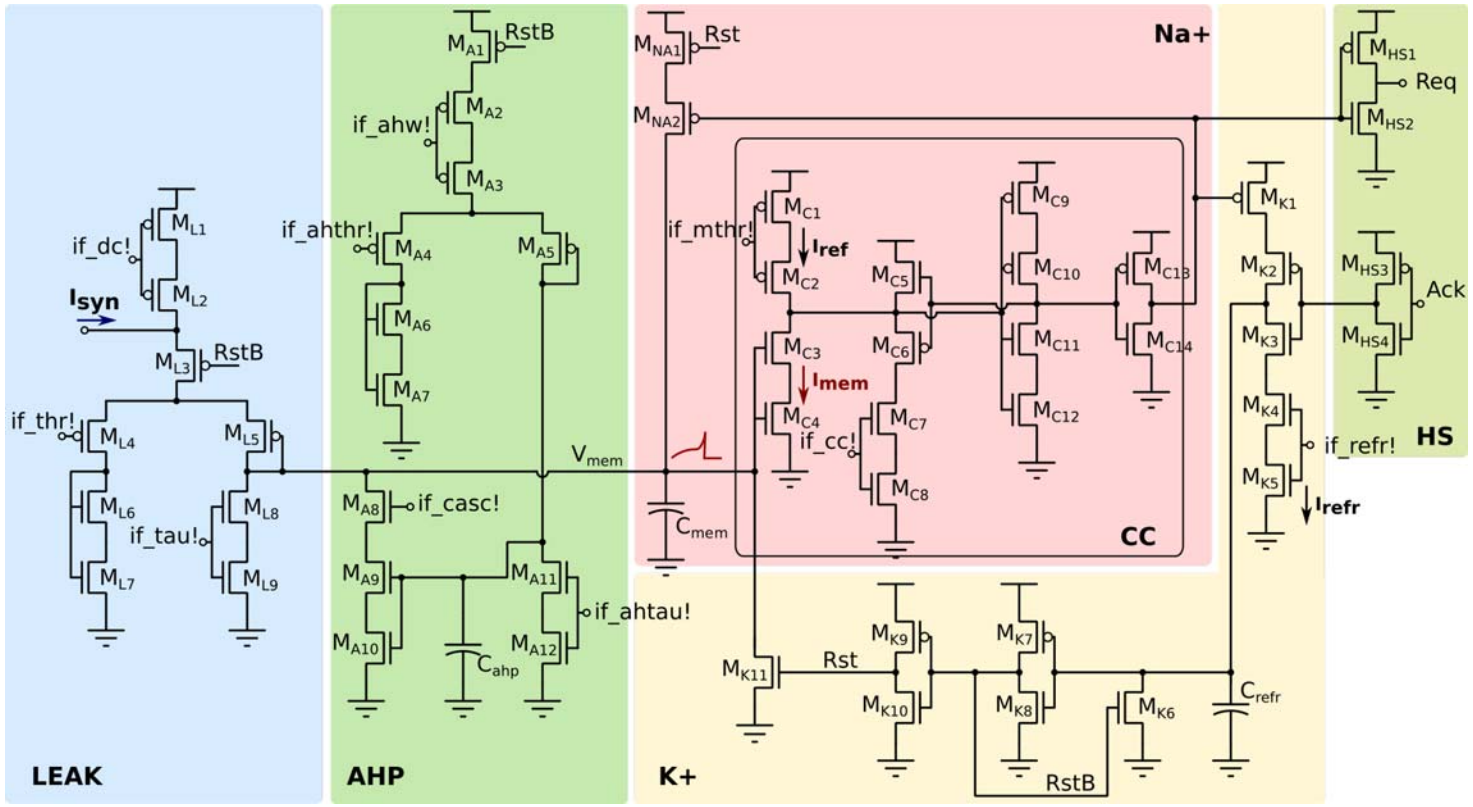


Let time represent itself

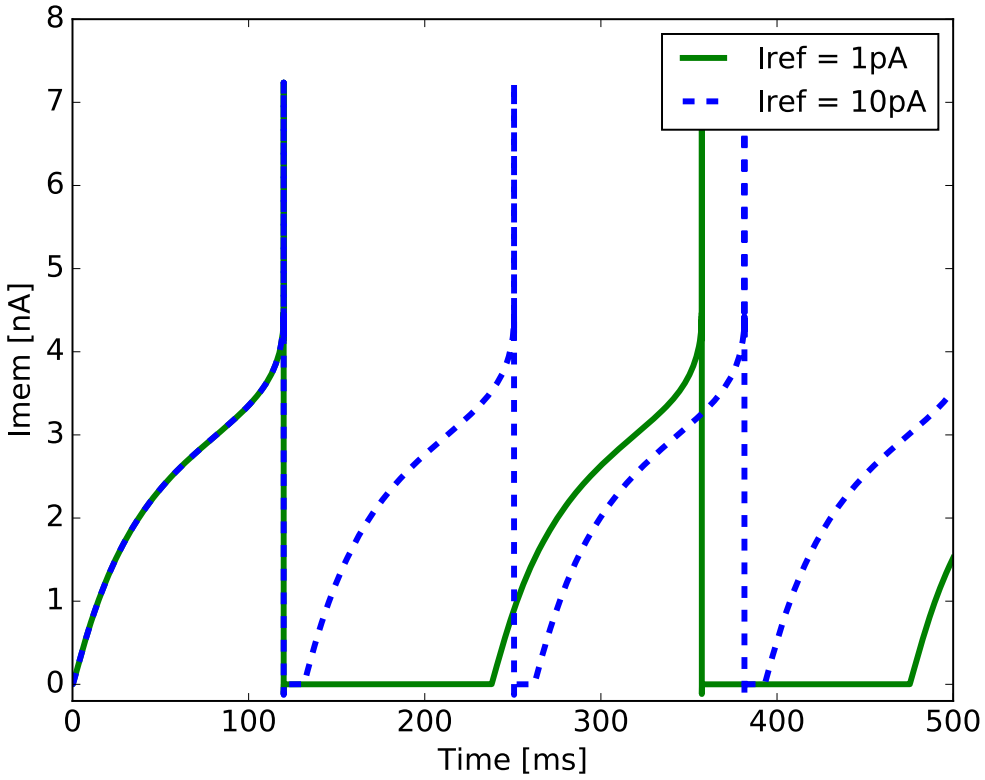
- For interacting with the environment in real-time.
- To match the circuit time constants to the input signal dynamics.
- For inherently synchronizing with the real-world natural events.
- To process sensory signals efficiently.

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

Time and space in silicon neurons

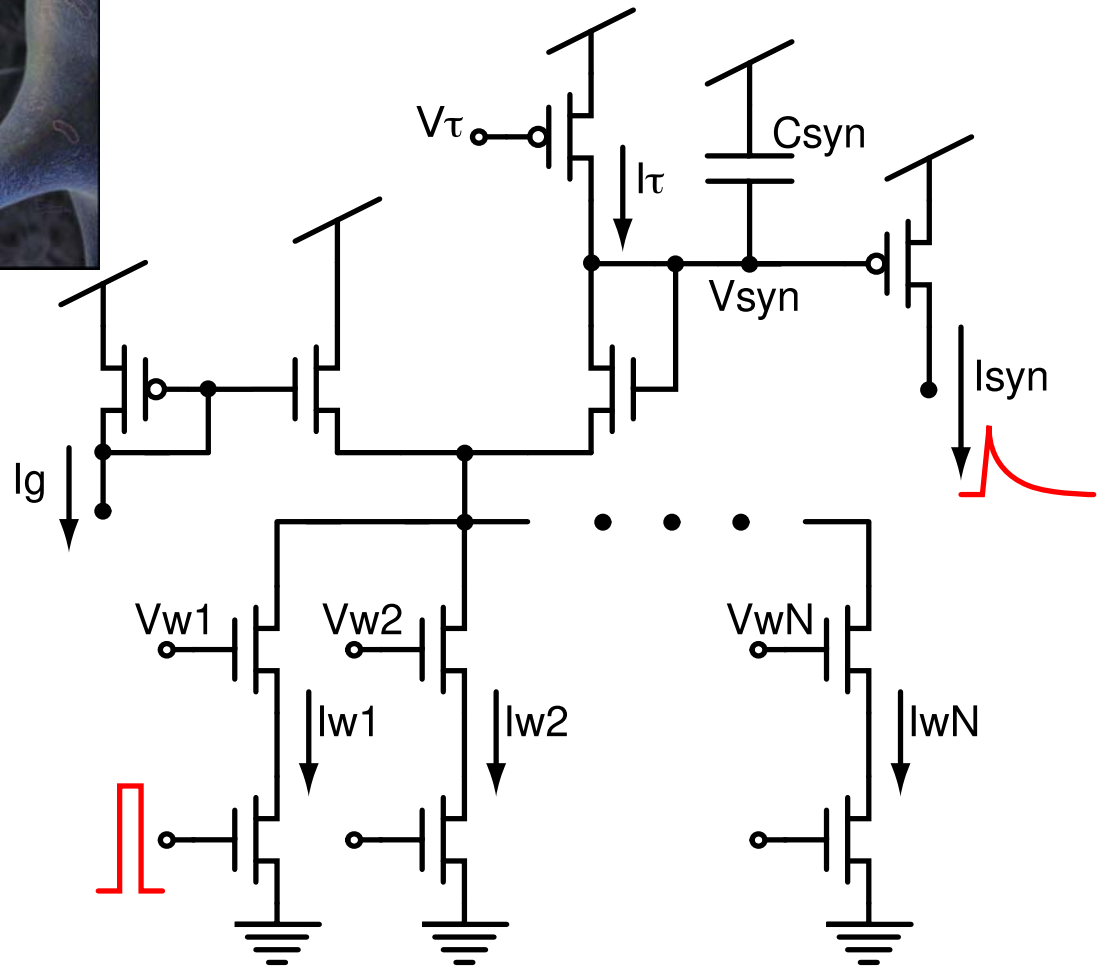


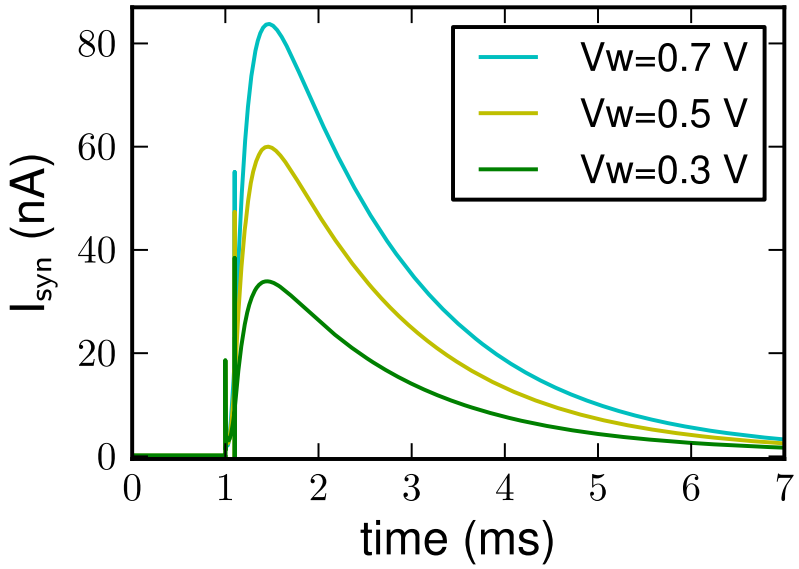
[Rubino et al., IEEE TCAS, 2020]

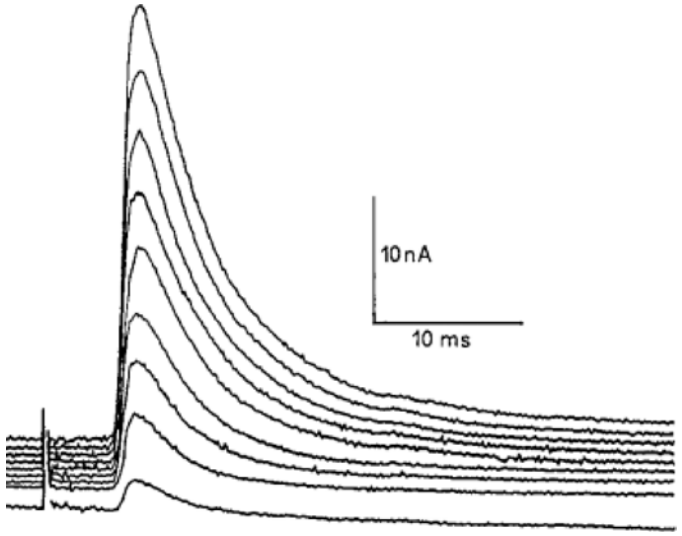


Work	[12]	[19]	[36]	**This work**
Techn.	180 nm CMOS	28 nm CMOS	28 nm FDSOI	22 nm FDSOI
Type	Mixed	Mixed	Mixed	Mixed
V_{dd}	1.8 V	0.7-1 V	1 V	0.8 V
Freq	30 Hz	-	30 Hz	30 Hz
Results	Experimental	Experimental	Simulation	Simulation
En./spike	883 pJ	2.3 nJ-30 nJ	50 pJ	14 pJ

Synaptic dynamics



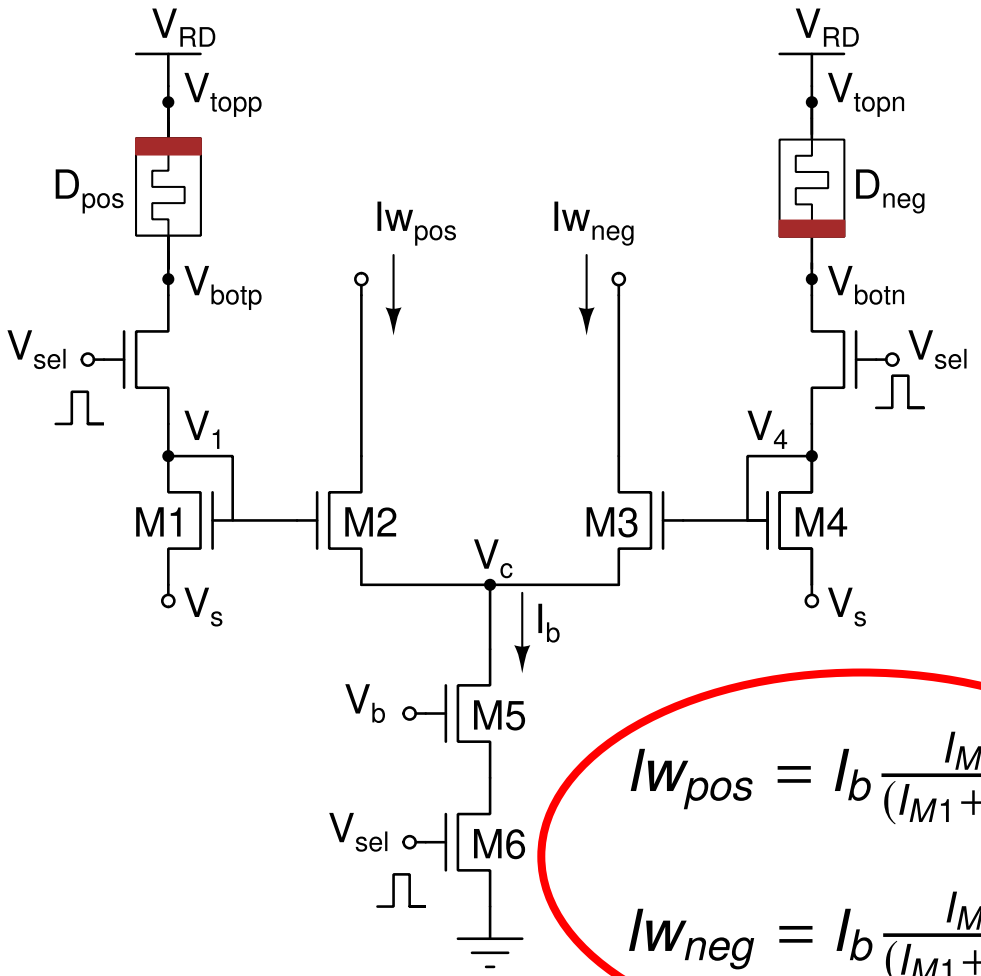




[Bartolozzi, Indiveri, NECO 2007; Sumislawska et al., ISCAS 2016]

Synaptic plasticity using memristive devices

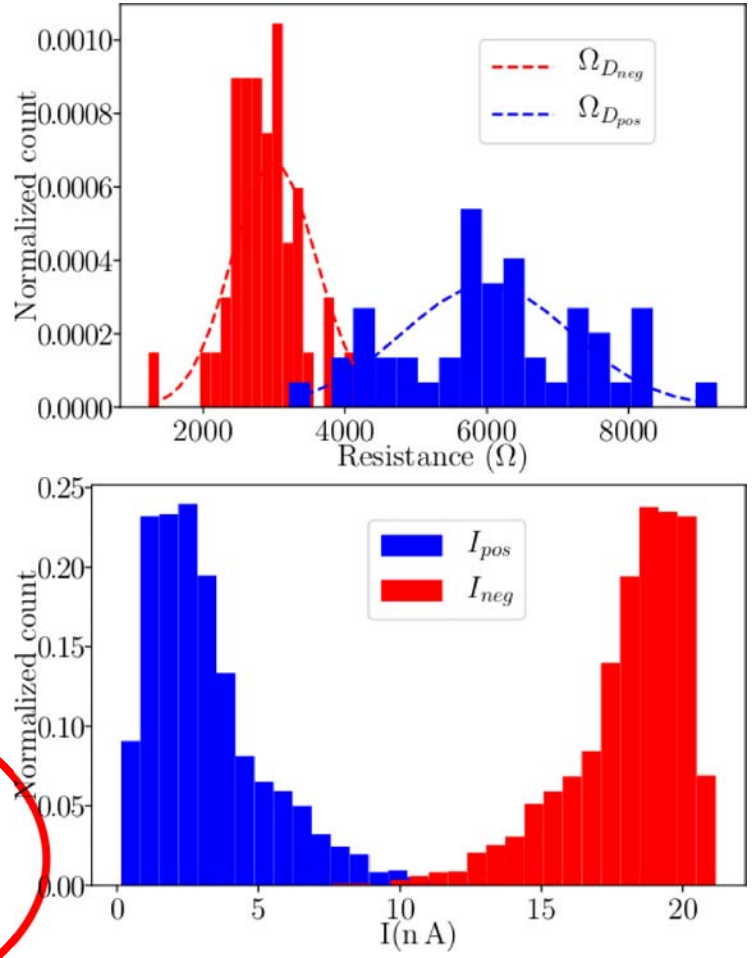
A (Gilbert) normalizer memristive synapse circuit



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

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

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



$CV = 0.429$

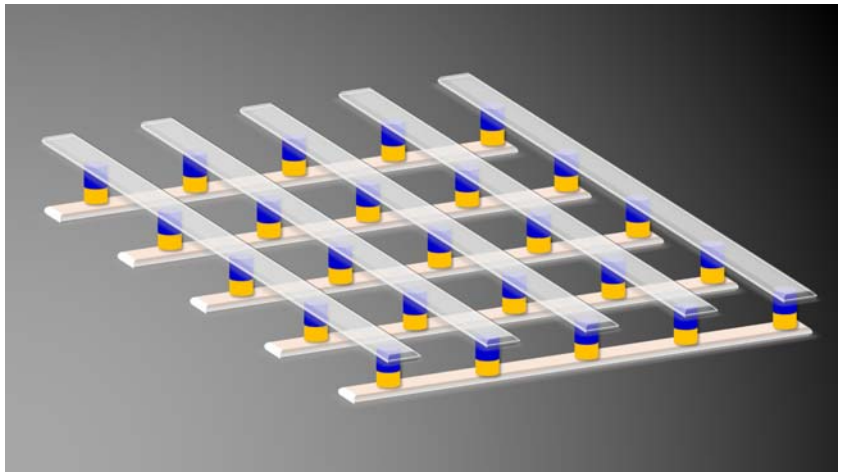
$CV = 0.284$

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

Analog subthreshold circuits

Advantages

- Avoid use of digital clock circuitry
- Avoid large DAC/ADC overhead
- Minimize power consumption
- Exploit the full potential of emerging memory technologies
  - Control multi-level properties with analog pulse heights
  - Exploit intrinsic non-linearities [Brivio et al., 2021]
  - Exploit intrinsic stochasticity [Gaba et al., 2013, Payvand et al., 2018]
  - Exploit non-volatility properties [Berdan et al., 2016, Demirag et al., 2021]



[Source: IBM]

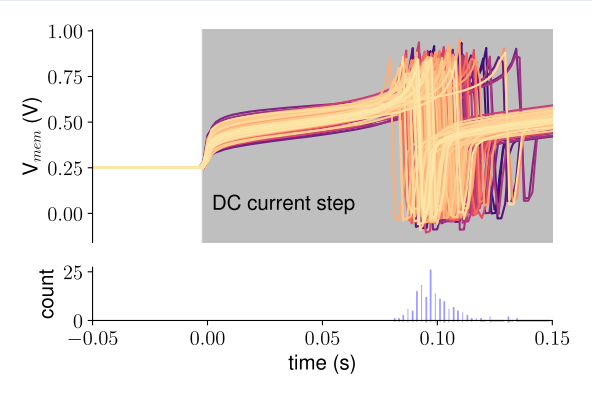
PCM-trace - [Y. Demirag et al., ISCAS 2021]

Exploit the drift of PCM devices to implement long-lasting *eligibility traces*. These enable the construction of powerful learning mechanisms for solving complex tasks by bridging the synaptic (~ms) and behavioral time-scales (~minutes).

[Gerstner et al., 2018; Bellec et al., 2020]

Analog subthreshold circuits

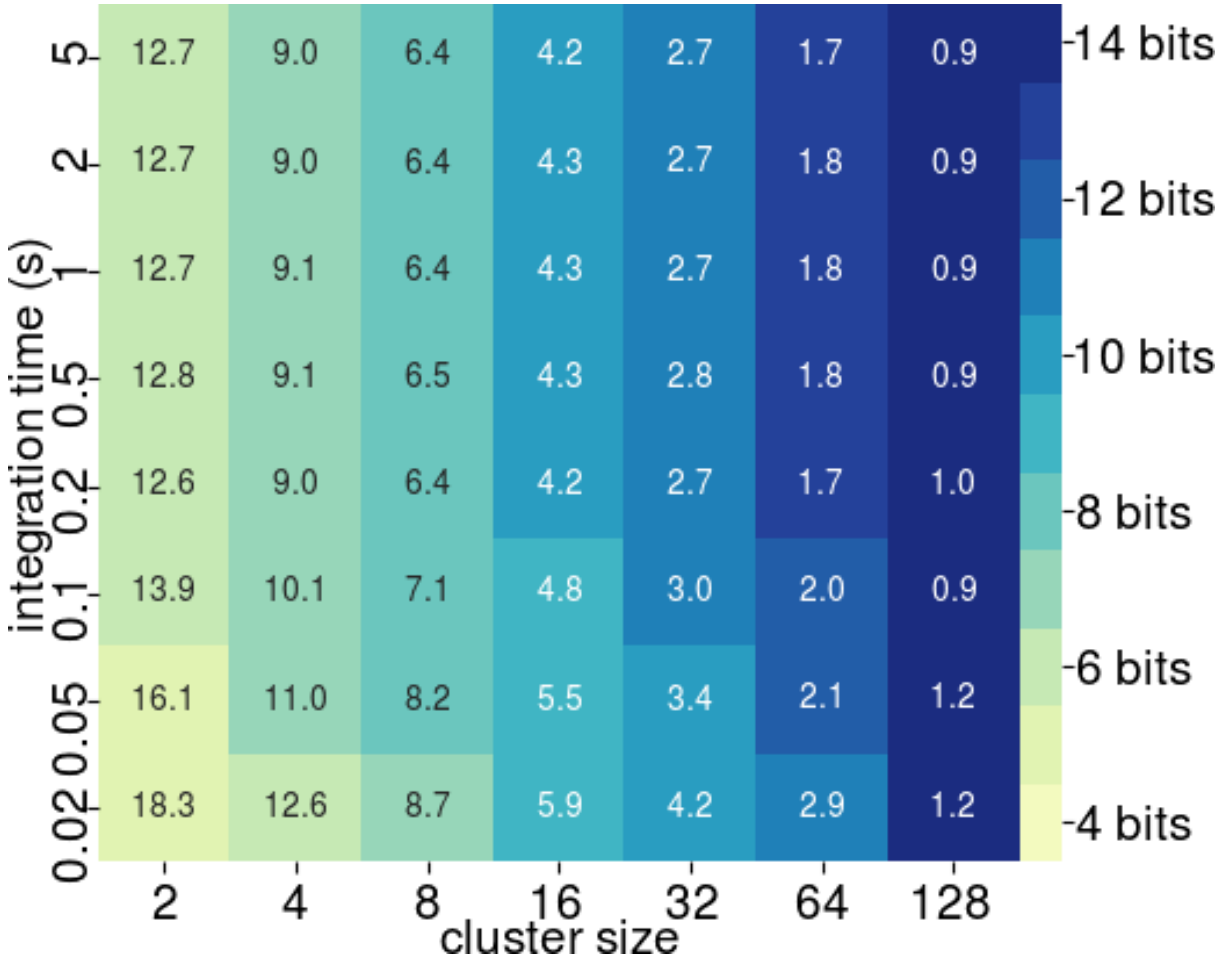
Disadvantages (?)



Membrane currents measured across 256 different neurons, in response to the same inputs

How to cope with mismatch and noise?

- Integrate over space (**populations** of neurons)
- Integrate over time (calculate mean rates)



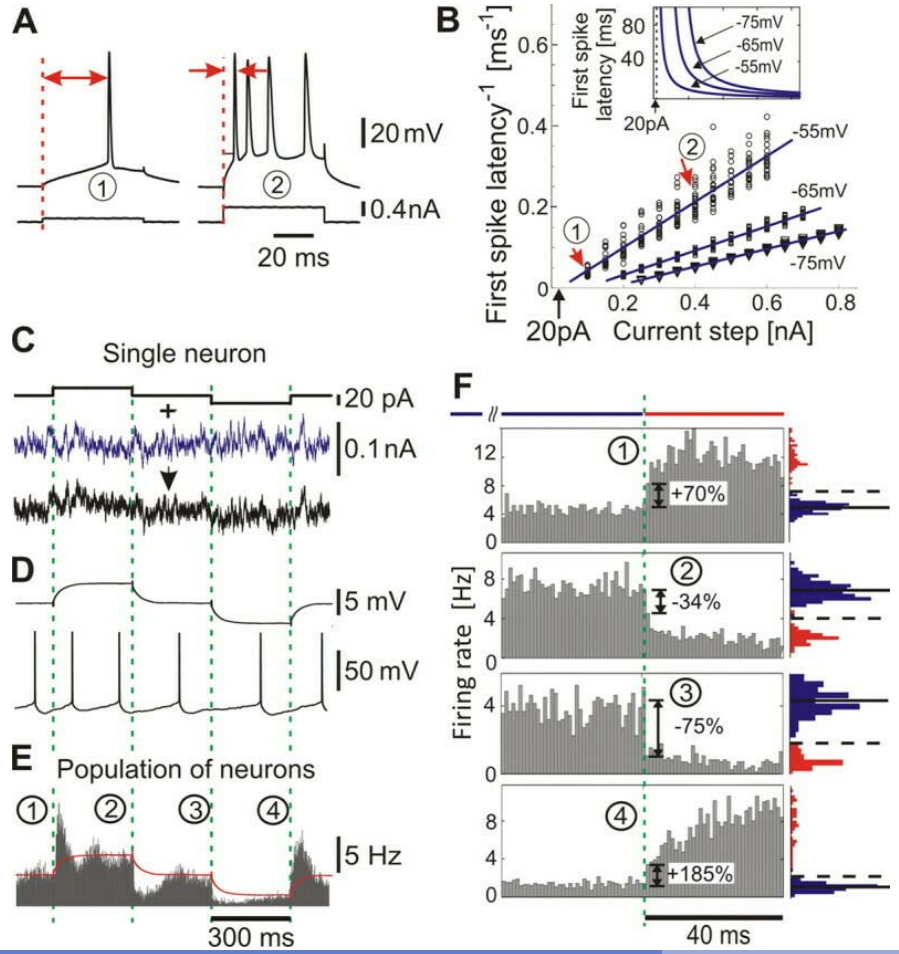
Coefficient of variation and Equivalent Number of Bits (ENOB)

[Zendrikov et al., (in preparation)]

False myths about analog neural responses

- Neural responses are slow.
  WRONG

Population firing rates of neurons can reliably encode weak signal changes ≈50 times faster than individual neurons.

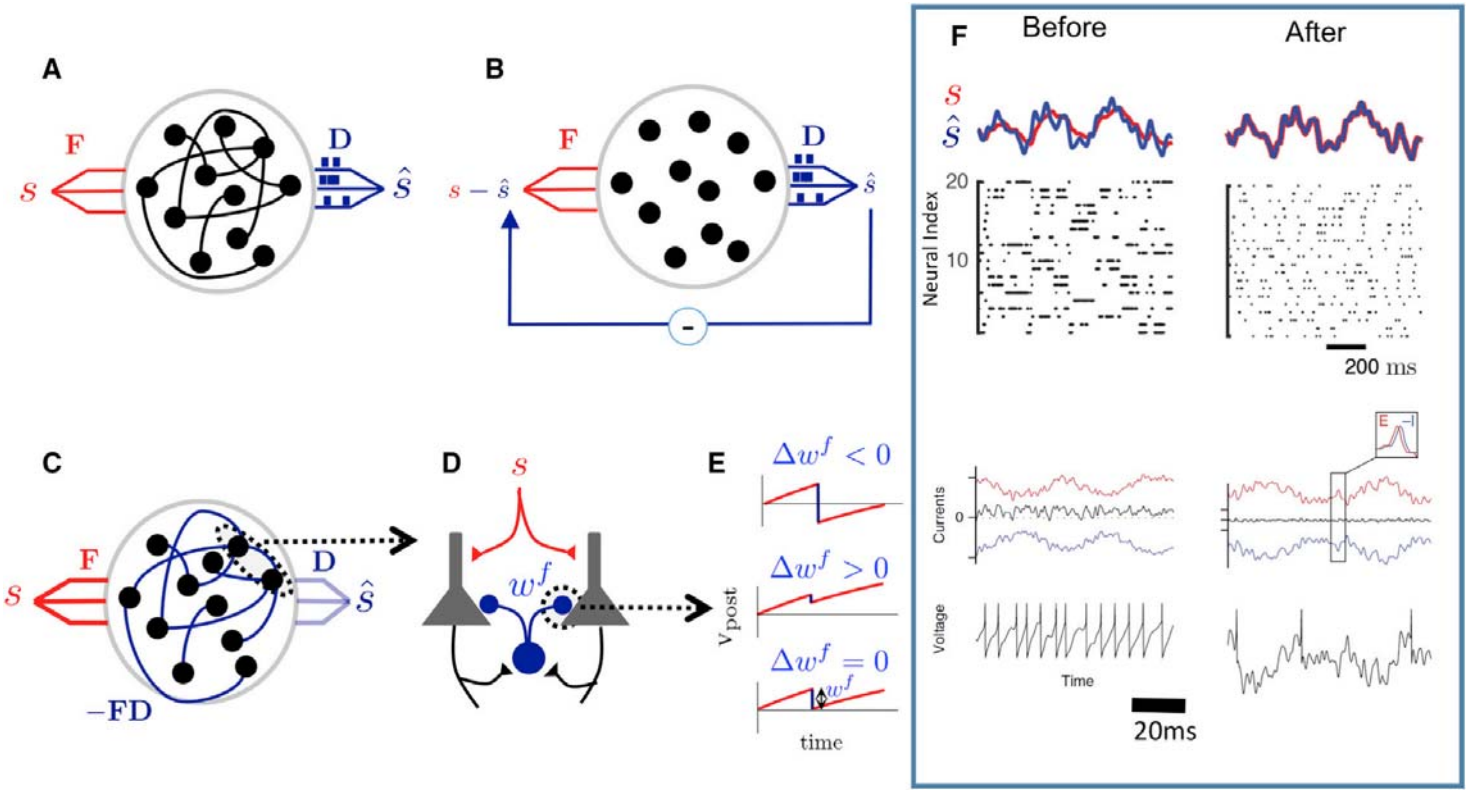


[T. Tchumatchenko et al., 2011]

False myths about analog neural responses

- Neural responses are slow.
- Neurons need to fire at high firing rates to achieve high precision.
  WRONG

E-I Balanced **populations** can encode signals with high precision and low rates



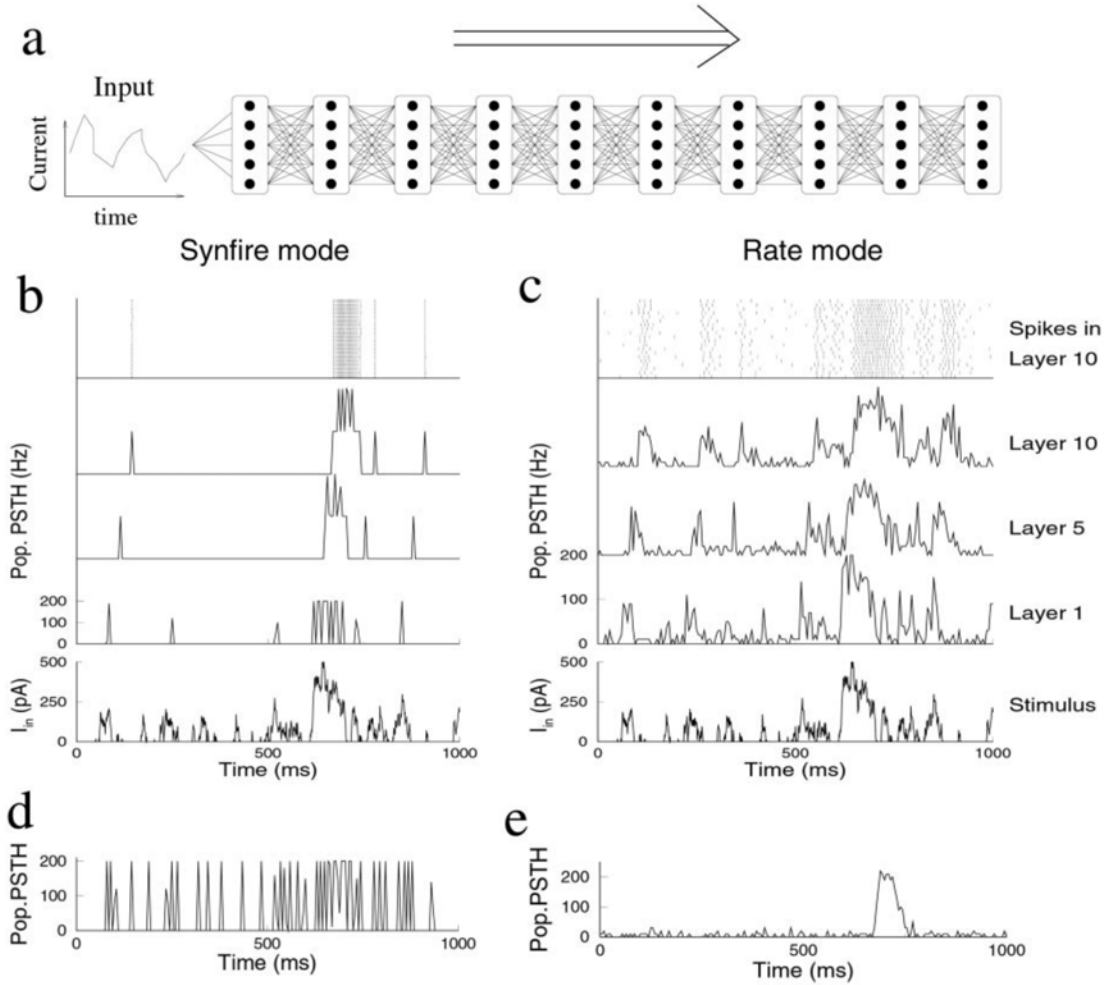
[Denève at al., 2017]

[Implementing efficient balanced networks with mixed-signal spike-based learning circuits, Büchel et al., ISCAS 2021]

False myths about analog neural responses

- Neural responses are slow.
- Neurons need to fire at high firing rates to achieve high precision.
- Neurons need to be accurate to propagate precise information across **population** layers.
  WRONG

Neurons need to be noisy to propagate neural activity reliably



[M.C.W. van Rossum, et al., 2002]

False myths about analog neural responses

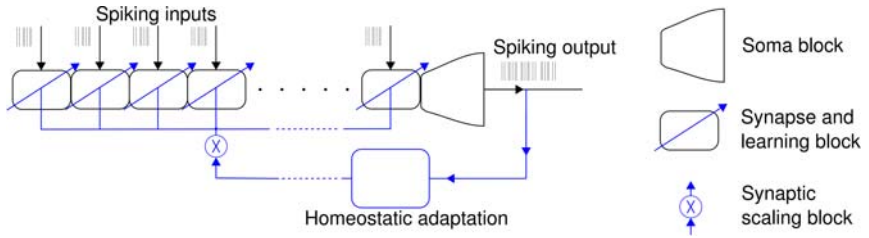
- Neural responses are slow. $\Longrightarrow$ **Populations** of noisy neurons have very fast response times.
- Neurons need to fire at high firing rates to achieve high precision. $\Longrightarrow$ Sparse neural **population** activity (in space *and* time) can represent signals with high accuracy.
- Neurons need to be accurate to propagate precise information across layers. $\Longrightarrow$ To propagate signals using low firing rates, it is necessary to use inhomogeneous **populations** of neurons.

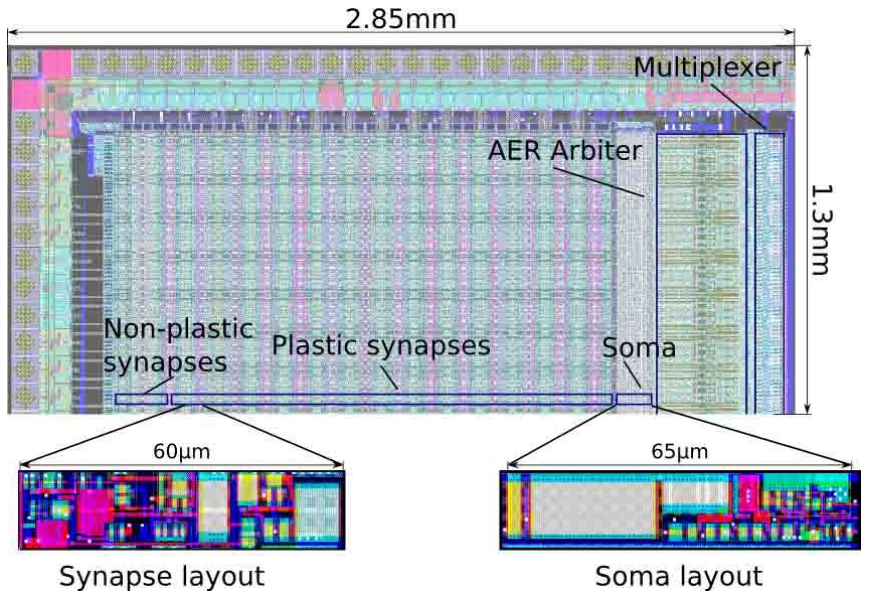
Brain inspired spiking neural network hardware

The perfect recipe

- Mix populations of mixed-signal silicon neurons and synapses.
- Add capacitors and volatile memristors for state dynamics and memory traces.

- Sprinkle distributed memory elements for parameter storage (SRAM, TCAM, non-volatile memristors)
- Include asynchronous digital circuits for event-based communication
- Serve with always-on, on-chip, self-supervised learning methods



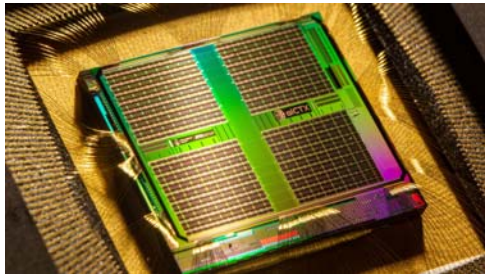


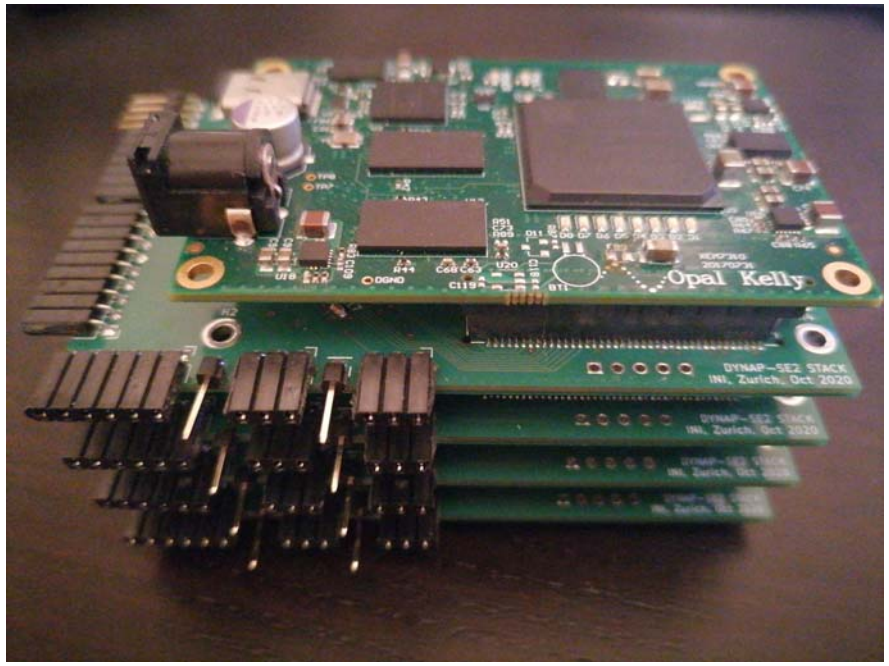
[Chicca & Indiveri, Applied Physics Letters 2020]

The Dynamic Neuromorphic Asynchronous Processor

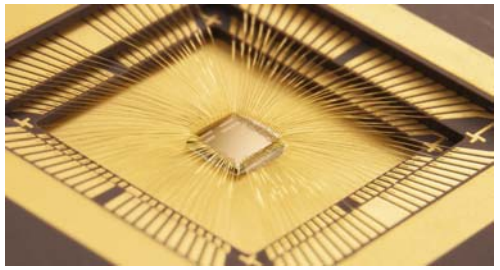
The DYNAP-SE2

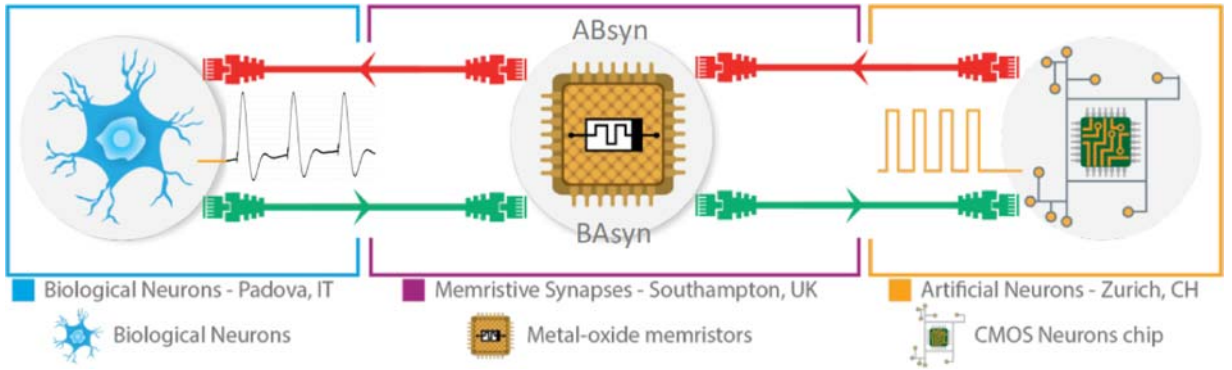
- Standard CMOS 180 nm process
- Four cores of 256 AdExp I&F neurons/core
- 64 synapses/neuron, 4 bit synaptic weight
- Four dendritic compartments
- Short-Term Plasticity (~ms)
- Homeostatic plasticity (~hours)
- Synaptic Delays
- Multi-cast event-based routing
- Tag based TCAM addressing
- On-chip bio-amplifiers and filters
- On-chip asynchronous delta modulation ADC

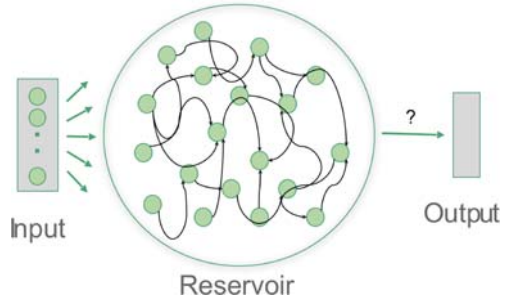




Applications: extreme edge computing

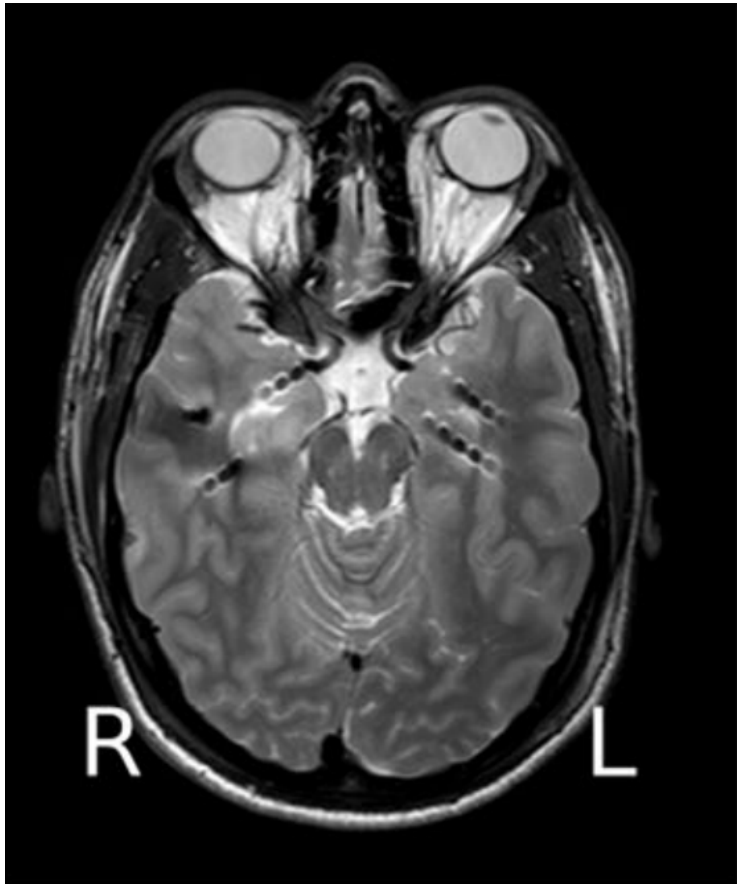




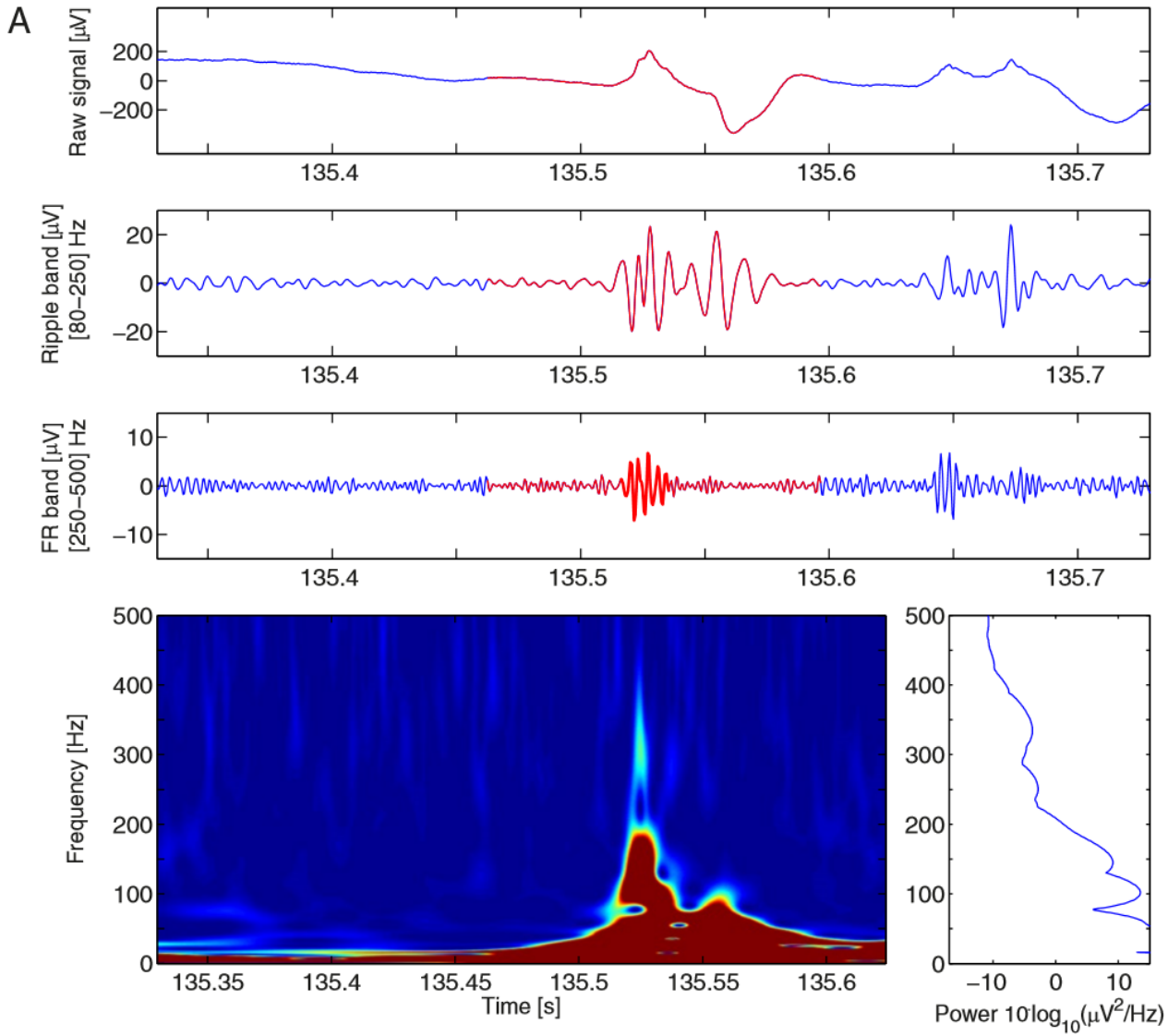


- Zebra-finch "Bird's Own Song" classification [Corradi et al., 2015]
- Closed-loop bidirectional brain machine interfaces [Boi et al., 2016]
- Closed-loop coupled biological-silicon neuron network [Serb et al. 2020]
- Adaptive pace-maker with neuromorphic CPG network [Abu-Hassan et al., 2019]
- On-line ECG anomaly detection [Bauer et al., 2019]
- On-line classification of EMG signals [Donati et al., 2019]
- Closed-Loop Spiking Control on a Neuromorphic Processor Implemented on the iCub [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 Neurom. Stereo-Vision Setup [Risi et al., 2021]
- On-line High-Frequency Oscillation (HFO) detection [Burelo, et al., 2021]
- Online Detection of Vibration Anomalies Using Balanced Spiking Neural Networks [Dennler et al., 2021]

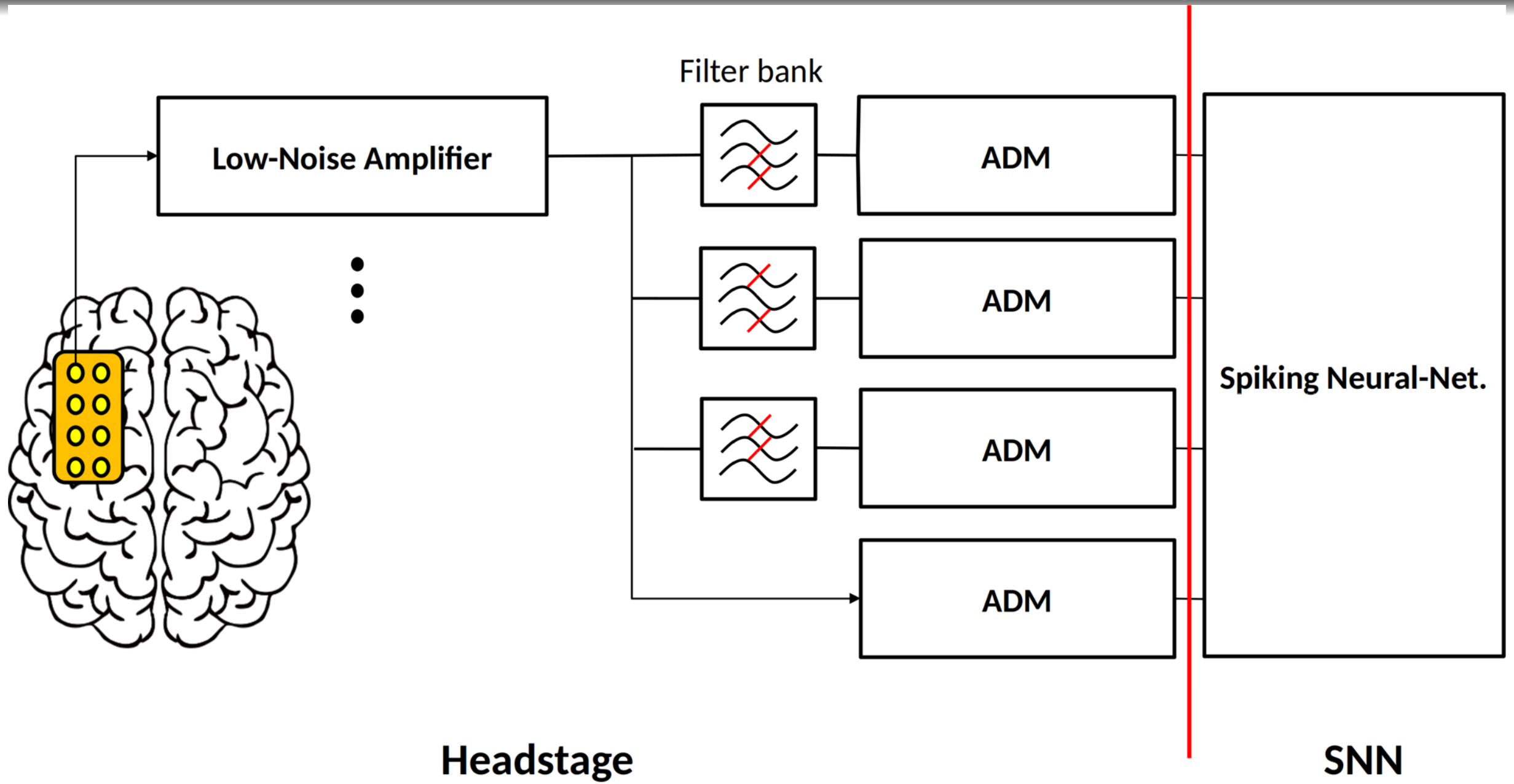
High-Frequency Oscillations (HFO) in iEEG



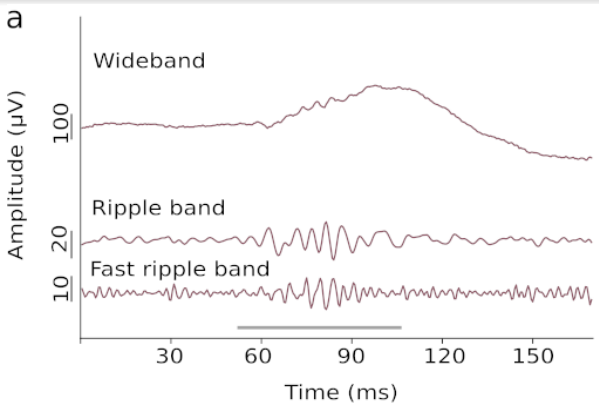
[Fedele et al., 2017]

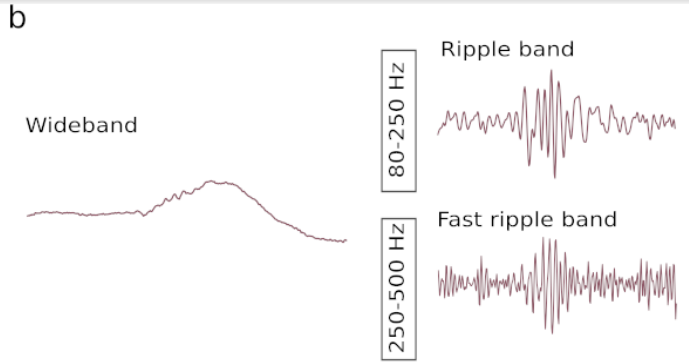


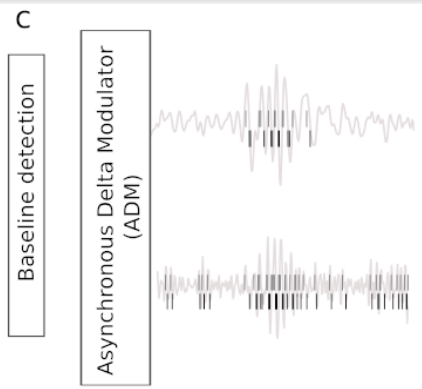
A neuromorphic HFO detection device

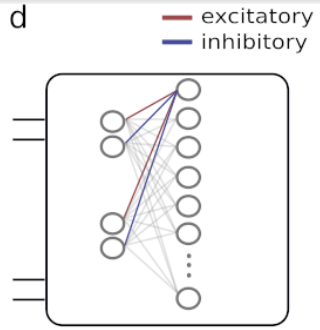


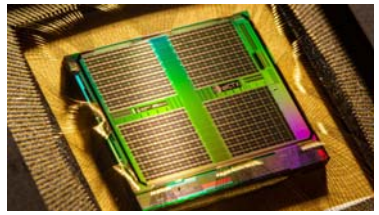
On-line iEEG HFO detection: results

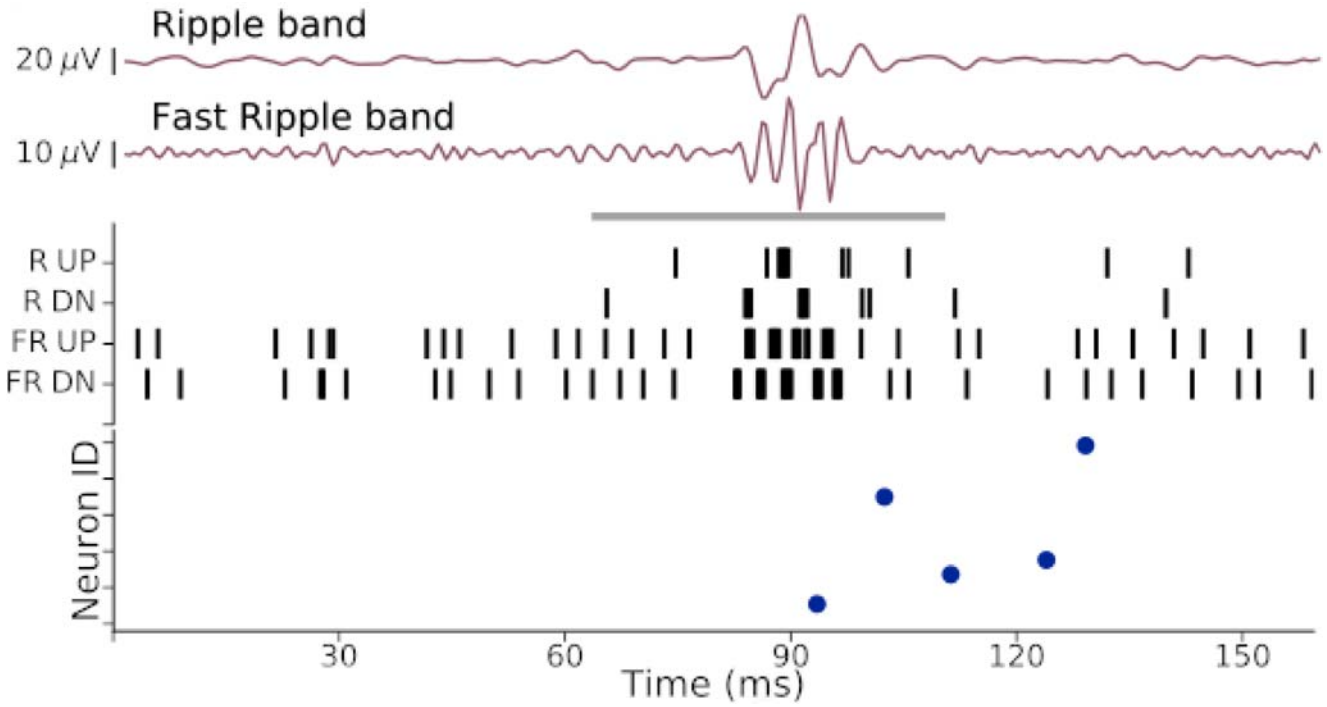












- Two layer network
- 256 neurons, 512 synapses
- One-bit weights
- Matched time constants
- Inhomogeneous parameters
- Power consumption: **614 μW**
- Accuracy: **78%** (vs. 67%)

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

Neuromorphic vs conventional computing

Pros

- Low latency
- Ultra low-power (**<1 mW**)

Cons

- Limited resolution (<8 bits)
- High variability, noisy

What are they good for?

- Small-scale network emulation
- Real-time sensory processing
- Sensory-fusion and on-line classification
- Low-latency decision making

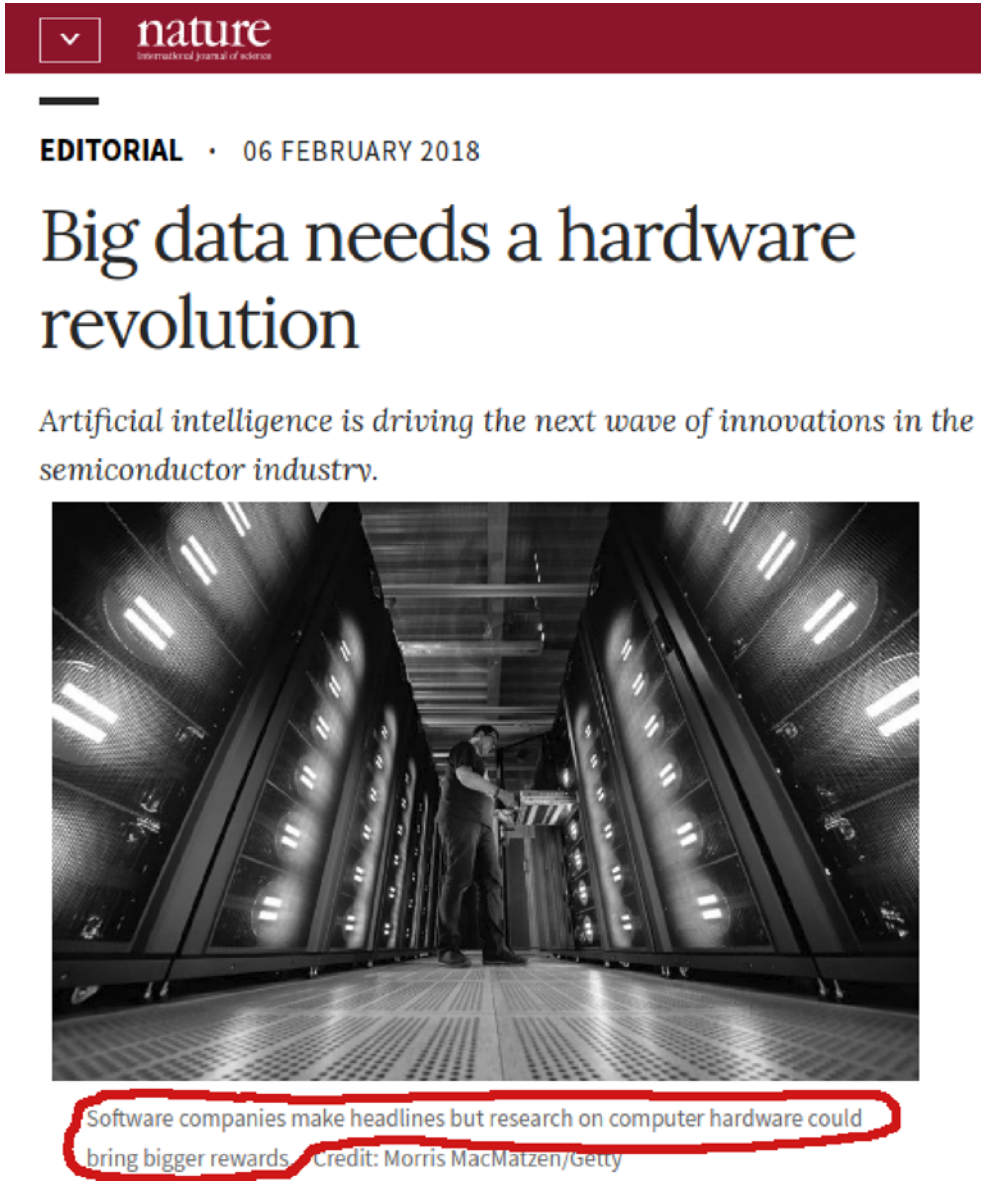
What are they bad at?

- Large scale network simulation
- High accuracy pattern recognition
- High precision number crunching
- Batch processing of data sets

Ideal technologies for extreme edge-computing applications

Bio-inspired neuromorphic processors complement conventional ANN accelerators.

Conclusions



nature

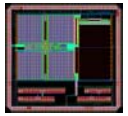
EDITORIAL · 06 FEBRUARY 2018

Big data needs a hardware revolution

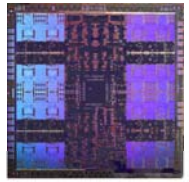
Artificial intelligence is driving the next wave of innovations in the semiconductor industry.

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 can benefit from massively parallel processing architectures.
- Neuroscience and machine learning are uncovering powerful and robust neural processing methods.
- This is the perfect time to follow the neuromorphic engineering approach for starting a hardware revolution.



DYNAP (160 μW)



NVIDIA Fermi GPU (145 W)

Swallow (45 g)

Boeing 737 (40000 kg)

A team effort

institute of neuroinformatics

University of Zurich[UZH] | ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

- Hannah Bos
- Elisa Donati
- Charlotte Frenkel
- Melika Payvand
- Ning Qiao (collaborator)
- Sergio Solinas (collaborator)
- Farah Baracat
- Karla Burelo
- Matteo Cartiglia
- Junren Chen
- Yigit Demirag
- Renate Krause
- Vanessa Leite
- Maryada
- Shyam Narayanan
- Nicoletta Risi
- Arianna Rubino
- Mohammad Ali Sharif
- Zhe Su
- Chenxi Wu
- Jingyue Zhao
- Dmitrii Zendrikov

Funding Sources

SWISS NATIONAL SCIENCE FOUNDATION

SynSense

institute of **neuroinformatics**

TINY ML EMEA
Virtual
JUNE 7-10,2021

Premier Sponsor

Intelligent Agent
Neuton

Automated TinyML

Zero-code SaaS solution

Create tiny models, ready for embedding, in just a few clicks!

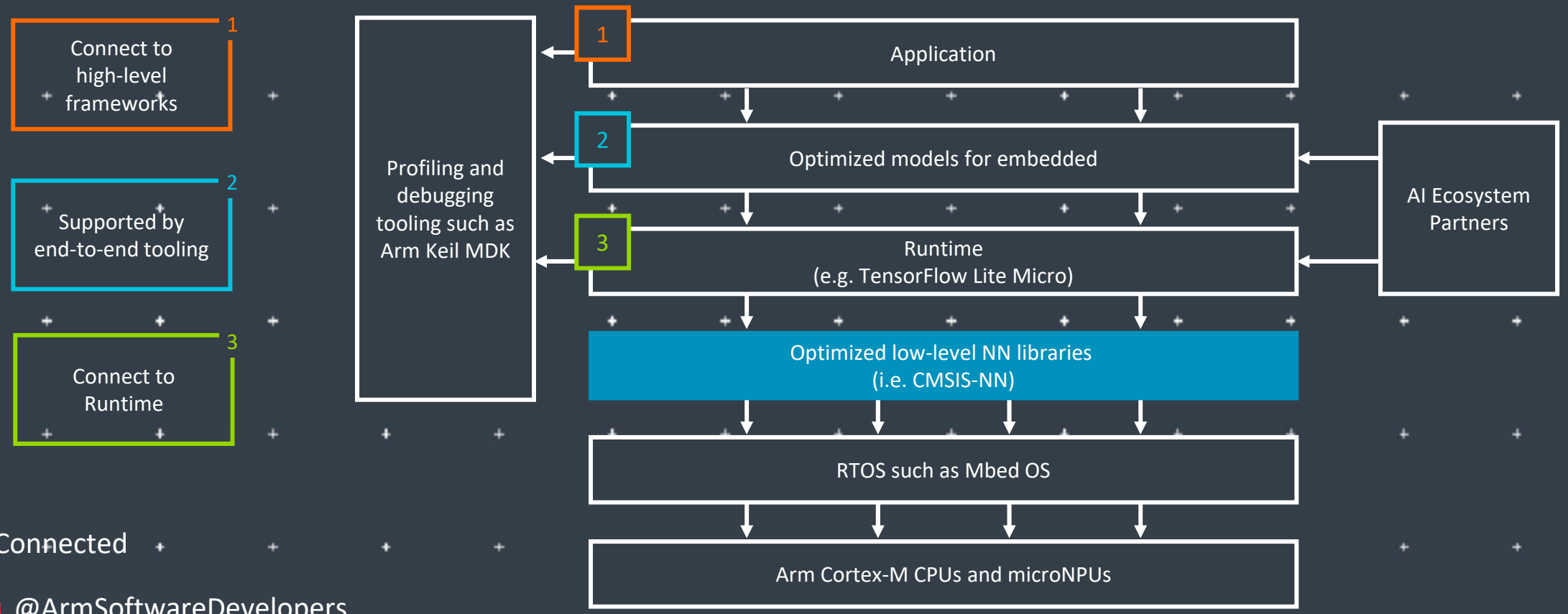
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.

TINY ML EMEA
Virtual
JUNE 7-10,2021

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

© 2020 Arm Limited (or its affiliates)

arm

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

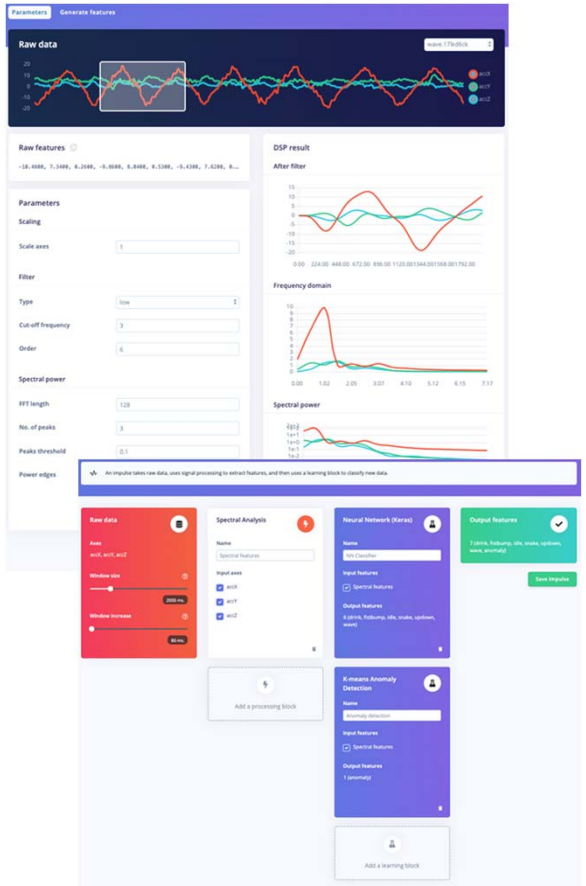
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



www.edgeimpulse.com

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

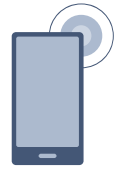
Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive



Mobile

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

SYNTIANT

Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp

TINY ML EMEA
Virtual
JUNE 7-10,2021

Platinum Sponsors

Part of your life. Part of tomorrow.

www.infineon.com

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai info@reality.ai @SensorAI Reality AI

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions

Prebuilt sound recognition models for indoor and outdoor use cases

Solution for industrial anomaly detection

Pre-built automotive solution that lets cars "see with sound"

Reality AI Tools® software

Build prototypes, then turn them into real products

Explain ML models and relate the function to the physics

Optimize the hardware, including sensor selection and placement

TINY ML EMEA
Virtual
JUNE 7-10,2021

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

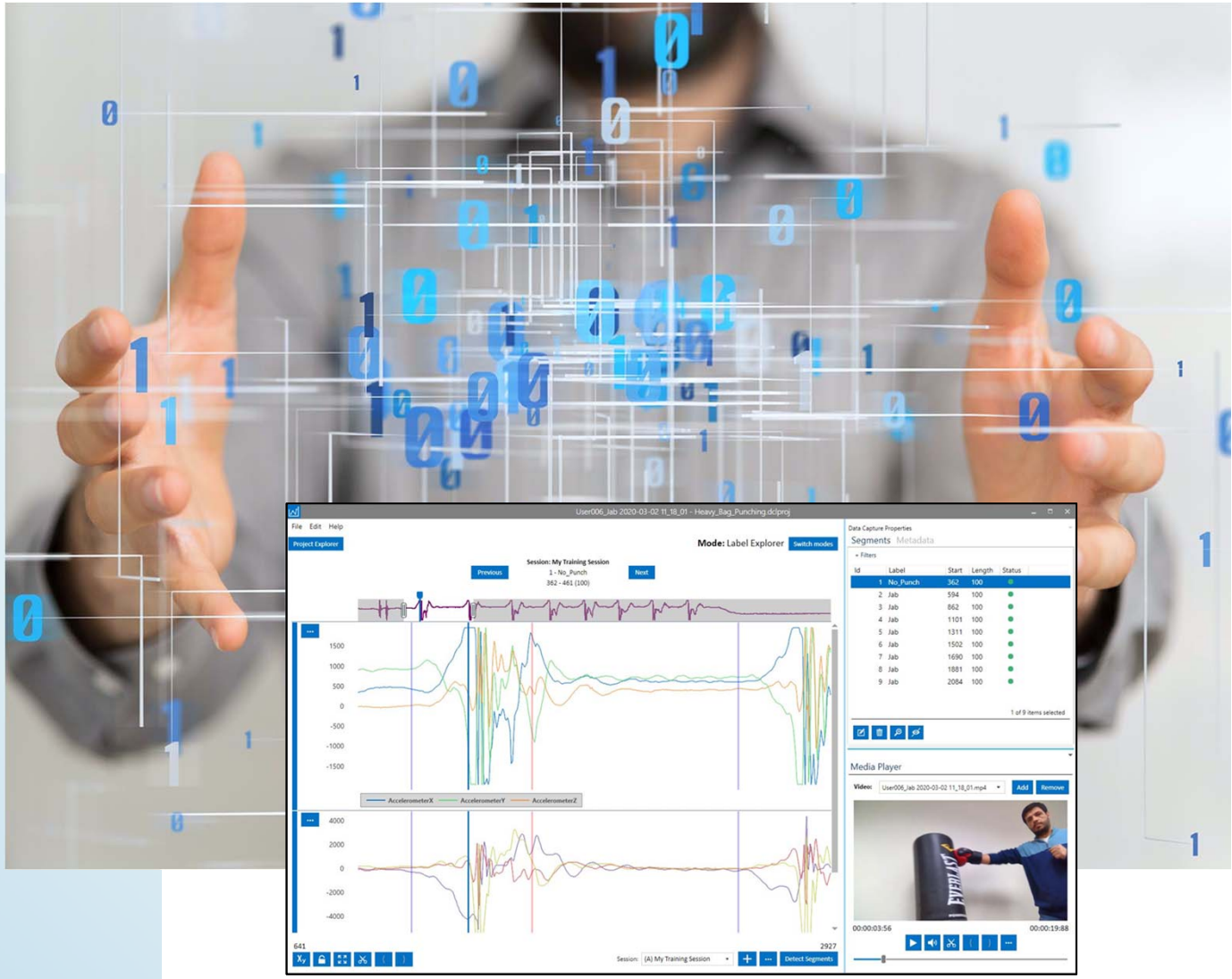
SensiML™

Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

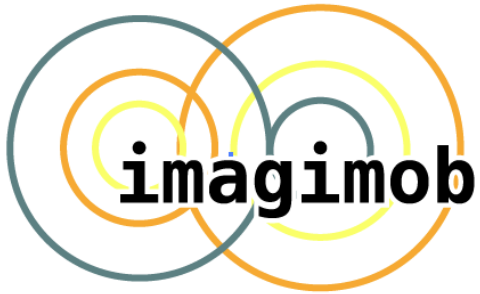
We enable the creation of production-grade smart sensor devices.



sensiml.com

TINY ML EMEA Virtual JUNE 7-10,2021

Silver Sponsors



Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® EMEA Technical Forum 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at tinyML EMEA. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org