tinyML® On Device Learning Forum

Enabling Ultra-low Power Machine Learning at the Edge

"Merging insights from artificial and biological neural networks for neuromorphic edge intelligence"

Charlotte Frenkel – Assistant professor, Delft University of Technology

May 16, 2023

TINY ML

www.tinyML.org

The Dawn of On Device Learning in TinyML

The goal of On Device Learning (ODL) is to make edge devices "smarter" and more efficient by observing changes in the data collected and self-adjusting/reconfiguring the device's operating model. Optionally the "knowledge" gained by the device is shared with other deployed devices.

Danilo Pau, Elias Fallon, Evgeni Gousev, Davis Sawyer, Ira Feldman, Christopher B. Rogers

TINY ODL

On device learning Forum

tinyML On Device Learning Forum 8/31 – 9/1 , 2022 Online

- Accademia on 8/31/2022
  - On-Device Learning Under 256KB Memory, Song HAN, Assistant Professor, MIT EECS
  - Neural Network ODL for Wireless Sensor Nodes ,Hiroki MATSUTANI, Professor, Keio University
  - Scalable, Heterogeneity-Aware and Trustworthy Federated Learning, Yiran CHEN, Professor, Duke University
  - On-Device Learning For Natural Language Processing with BERT, Warren J. GROSS, Professor, McGill University
  - Is on-device learning the next "big thing" in TinyML? Manuel ROVERI, Associate Professor, Politecnico di Milano
  - ODL Professors Panel
- Industry on 9/1/2022
  - TinyML ODL in industrial IoT, Haoyu REN, PhD Student, Technical University of Munich/Siemens
  - NeuroMem® wearable, hardwired sub milliwatt real time machine learning with wholly parallel access to "neuron memories" fully explainable, Guy PAILLET, Co-founder, General Vision
  - Using Coral Dev Board Micro for ODL innovations, Bill LUAN, Senior Program Manager, Google
  - Platform for Next Generation Analog AI Hardware Acceleration, Kaoutar EL MAGHRAOUI, Principal Research Scientist, IBM T.J Watson Research Center
  - Enabling on-device learning at scale, Joseph SORIAGA, Sr. Director of Technology, Qualcomm
  - Training models on tiny edge devices, Valeria TOMASELLI, Senior Engineer, STMicroelectronics

https://www.tinyml.org/event/on-device-learning/

tinyML EMEA Forum - On Device Learning 9/12 , 2022 Cyprus, In person

TINY ML

On device learning Forum

- A framework of algorithms and associated tool for on-device tiny learning, Danilo PAU, Technical Director, IEEE and ST Fellow, STMicroelectronics

- In Sensor and On-device Tiny Learning for Next Generation of Smart Sensors Michele MAGNO, Head of the Project-based learning Center, ETH Zurich, D-ITET

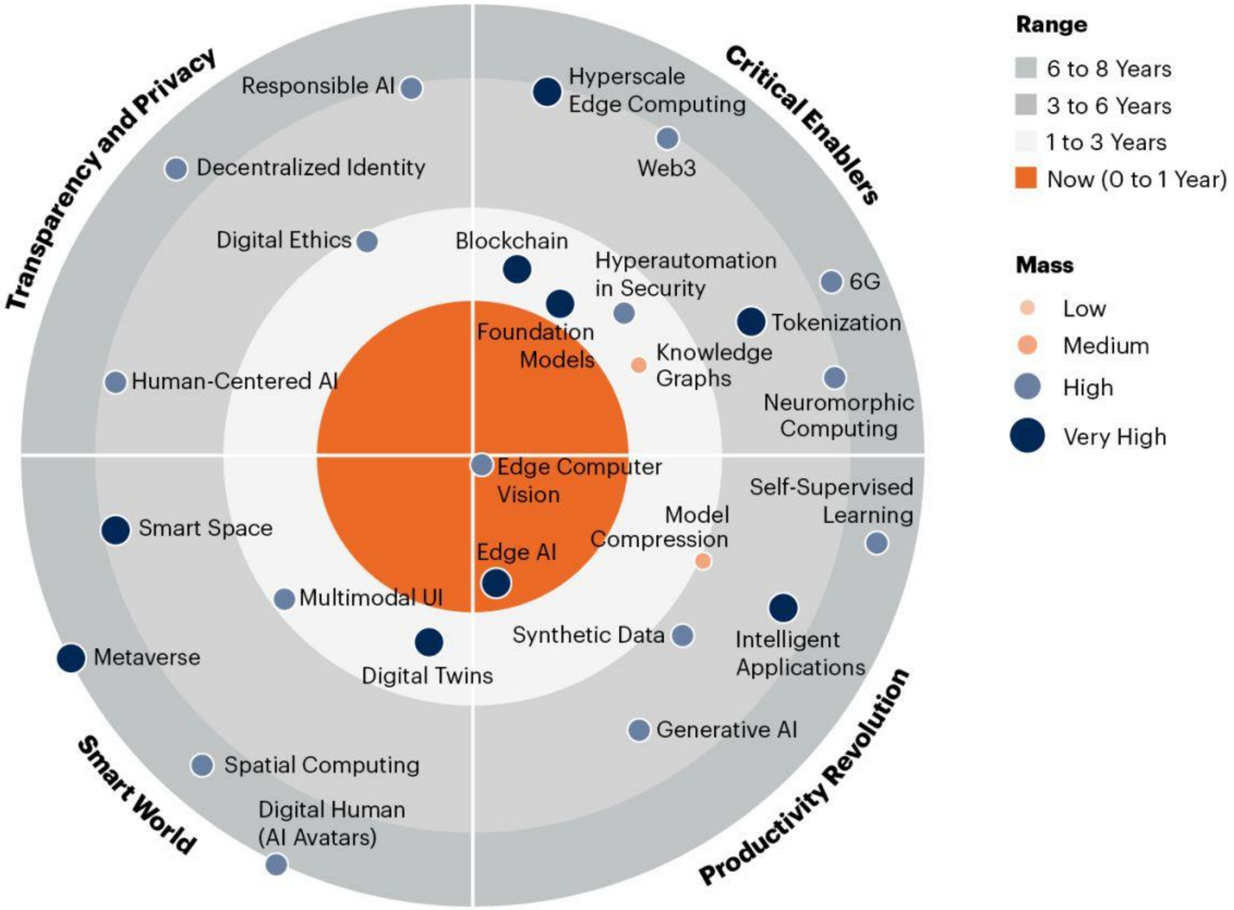
- Continual On-device Learning on Multi- Core RISC-V MicroControllers Manuele RUSCI, Embedded Machine Learning Engineer, Greenwaves

- On-device continuous event-driven deep learning to avoid model drift, Bijan MOHAMMADI, CSO, Bondzai

https://www.tinyml.org/event/on-device-learning/

TINY ML
On device learning Forum

2023 Gartner Emerging Technologies and Trends Impact Radar



gartner.com

Note: Range measures number of years it will take the technology/trend to cross over from early adopter to early majority adoption. Mass indicates how substantial the impact of the technology or trend will be on existing products and markets.

Source: Gartner

Gartner

On Device Learning Forum 2023, May 16 2023

- 8:00 - 8:10 Opening remarks by **Danilo Pau**
- 8:10 - 8:40 **Charlotte Frenkel** "Merging insights from artificial and biological neural networks for neuromorphic edge intelligence"
- 8:40 - 9:40 **Giorgia Dellaferrera** "Forward Learning with Top-Down Feedback: Solving the Credit Assignment Problem without a Backward Pass"
- 9:40 - 10:10 **Guy Paillet** "NeuroMem®, Ultra Low Power hardwired incremental learning and parallel pattern recognition"
- 10:10 - 10:40 **Aida Todri-Sanial** "On-Chip Learning and Implementation Challenges with Oscillatory Neural Networks"
- 10:40 - 11:10 **Eduardo S. Pereira** “Online Learning TinyML for Anomaly Detection Based on Extreme Values Theory”
- 11:10 - 11:15 Closing remarks by Danilo Pau

Pacific Time

life.augmented

TINY ML
On device learning Forum

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

On device learning Forum

Executive Strategic Partners

TINY ML

On device learning Forum

EDGE IMPULSE

The Leading Development Platform for Edge ML

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

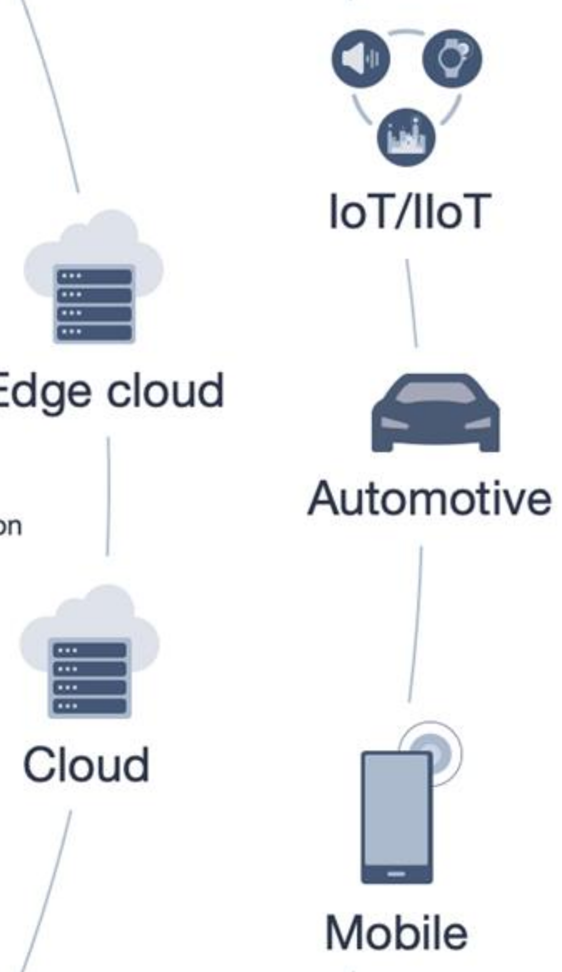
Efficient learning
Robust learning through minimal data, unsupervised learning, on-device learning

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



Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

Platinum Strategic Partners

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.

renesas.com

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML
On device learning Forum

Gold Strategic Partners

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

Where what if becomes what is.

Witness potential made possible at analog.com.

arm AI

Powering tinyML Innovation

Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more:
www.arm.com/techtalks

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Microsoft


The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product
SensiML
Analytics Toolkit Suite
AutoML
Data Collection
Test & Validation
Data Labeling
Code Generation
Model Building
Team Collaboration
Version Control and Model Management
sensiml.com/tinyML

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

synaptics®

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML

On device learning Forum

Silver Strategic Partners

AIZIP

brainchip — The Neuromorphic Computing Company

GREENWAVES TECHNOLOGIES

Grovety Inc.

IBM

imagimob

Nota AI

NXP

POLYN TECHNOLOGY

Qeexo

Schneider Electric

SILICON LABS

Join Growing tinyML Communities:

meetup

14.7k members in
47 Groups in 39 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

4k members
&
11.6k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

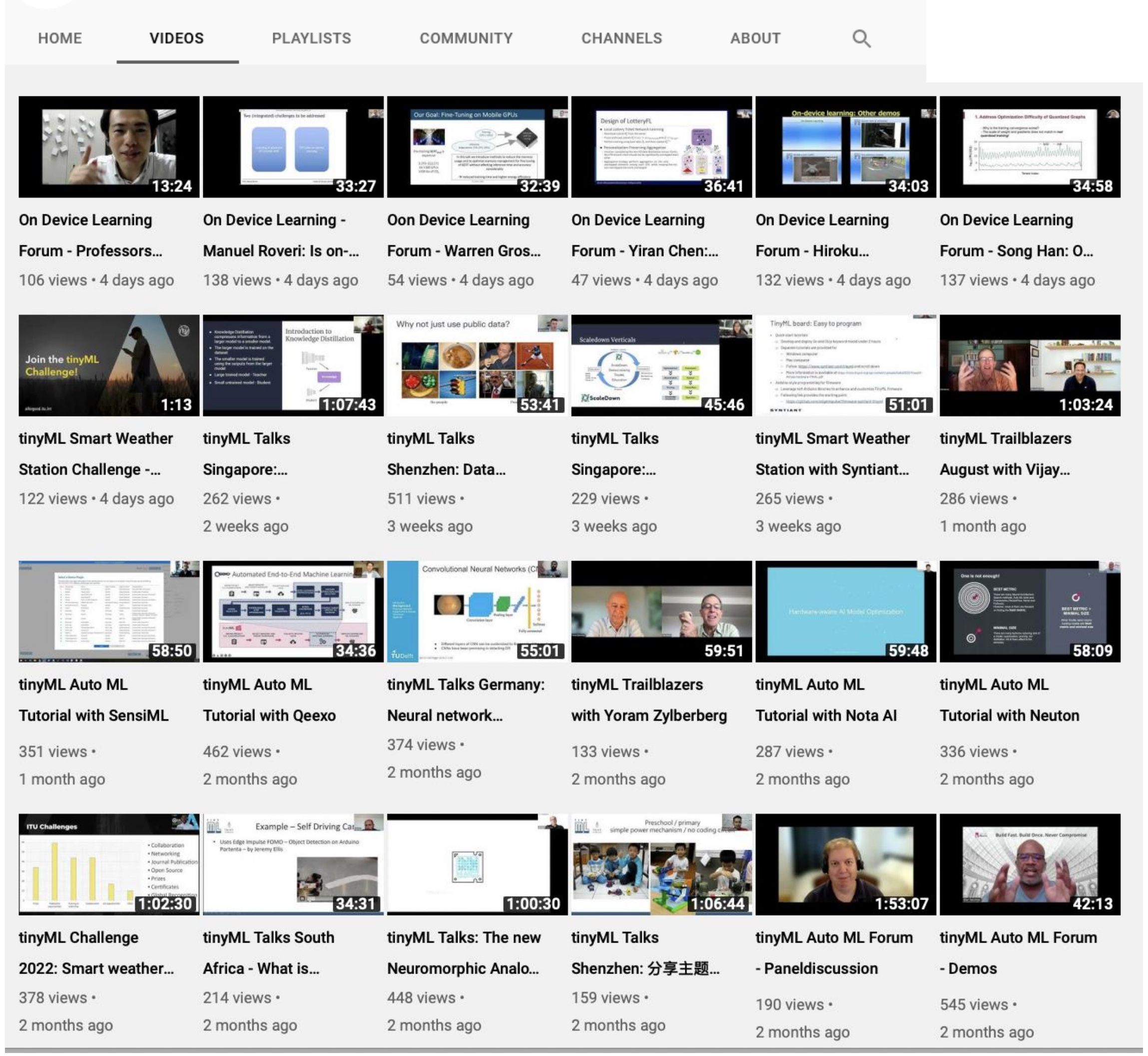
TINY ML

TINY ML

On device learning Forum

YouTube

9.4k subscribers, 559 videos with 327k views



Subscribe to
tinyML YouTube Channel
for updates and notifications
(including this video)

www.youtube.com/tinyML

On device learning Forum

FOUNDATION

EMEA 2023

https://www.tinyml.org/event/emea-2023

More sponsorships are available: sponsorships@tinyML.org

TINY ML
On device learning Forum

Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums   youtube.com/tinyml

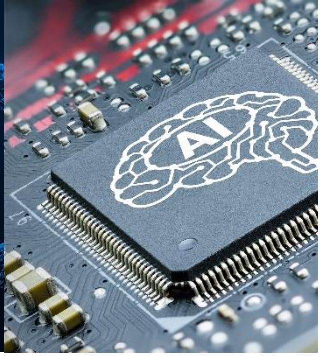
Please use the Q&A window for your questions

TINY ML

On device learning Forum

Charlotte Frenkel

Charlotte Frenkel received the Ph.D. degree from Université catholique de Louvain (UCLouvain), Belgium, in 2020. After a postdoc at the Institute of Neuroinformatics, UZH and ETH Zürich, Switzerland, she joined Delft University of Technology, The Netherlands, as an Assistant Professor in July 2022. Her research focuses on neuromorphic integrated circuit design and learning algorithms for adaptive edge computing. She received a best paper award at the IEEE ISCAS 2020 conference, as well as the FNRS Nokia Bell Labs Scientific Award, the FNRS IBM Innovation Award and the UCLouvain/ICTEAM Best Thesis Award for her Ph.D. thesis. She serves as a TPC member for the tinyML Research Symposium and for the IEEE ESSCIRC, ISLPED, and DATE conferences.



Merging insights from artificial and biological neural networks for neuromorphic intelligence

Charlotte Frenkel

Delft University of Technology, Microelectronics Department
c.frenkel@tudelft.nl

tinyML On-Device Learning Forum 2023
Online, May 16th 2023

Outline

1. From neuroscience to AI and back again… …which perspective?
   …which starting point?

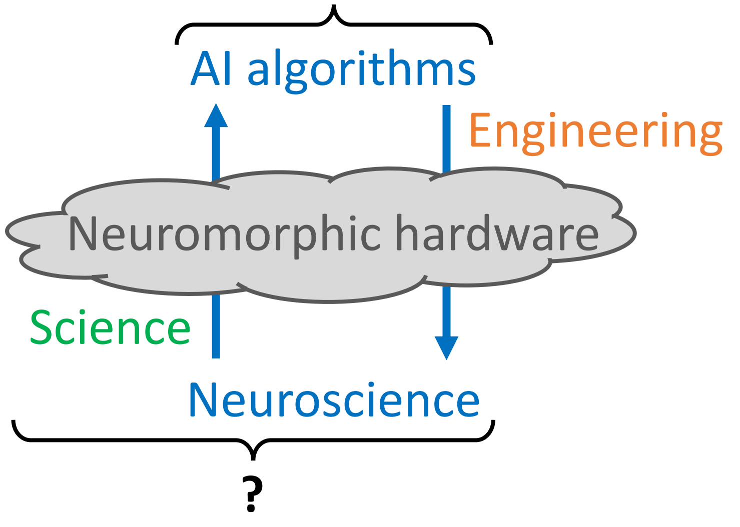
2. Why should we bother with neuroscience?

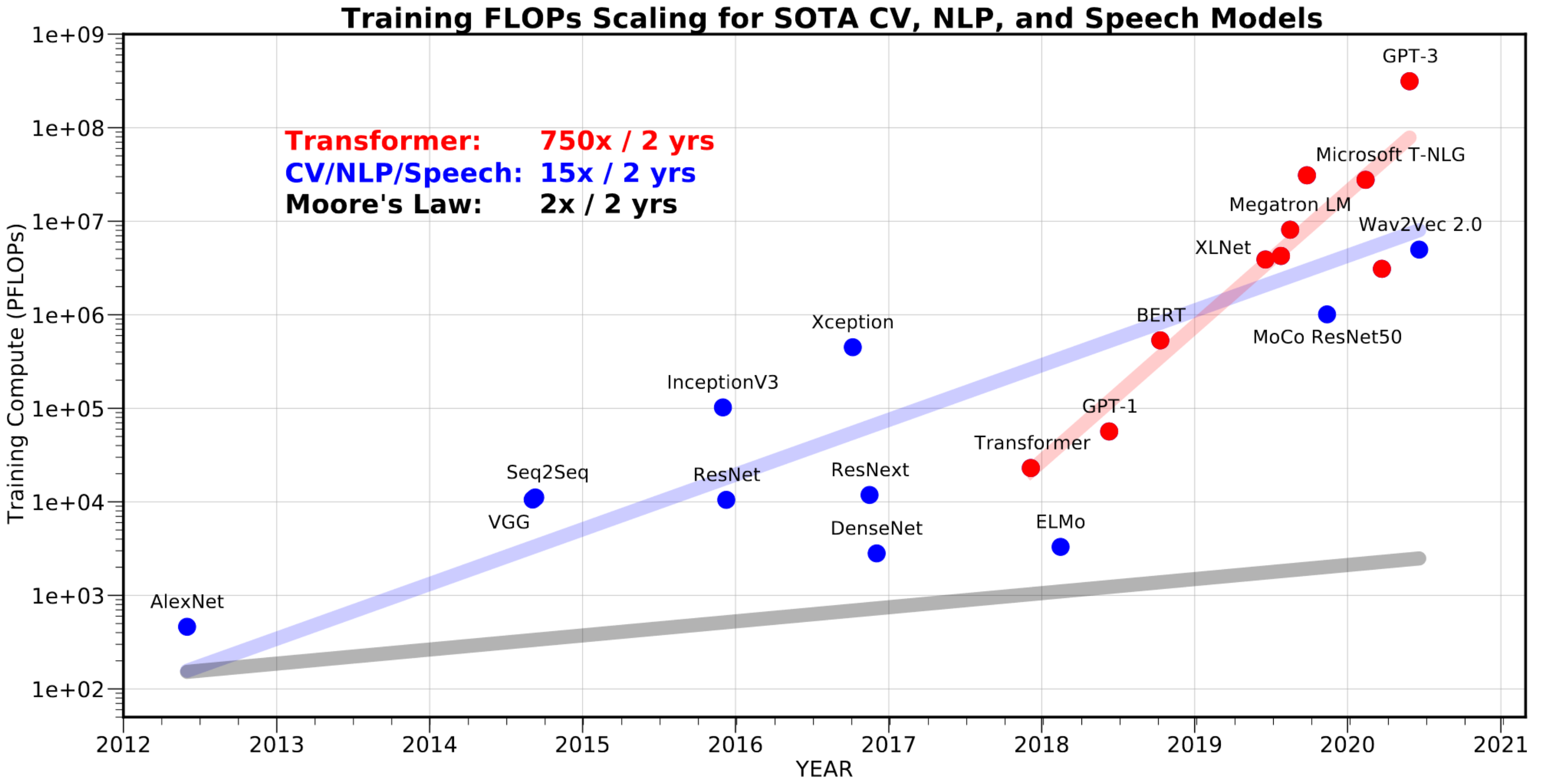
3. How can we morph these questions into interesting solutions for on-device-learning?

From neuroscience to AI and back again

Which starting point? *Which perspective?*

AI without hardware is unsustainable





[A. Gholami, *RiseLab Medium Post,* 2021]

Outline

1. From neuroscience to AI and back again… …which perspective? …which starting point?

2. Why should we bother with neuroscience?

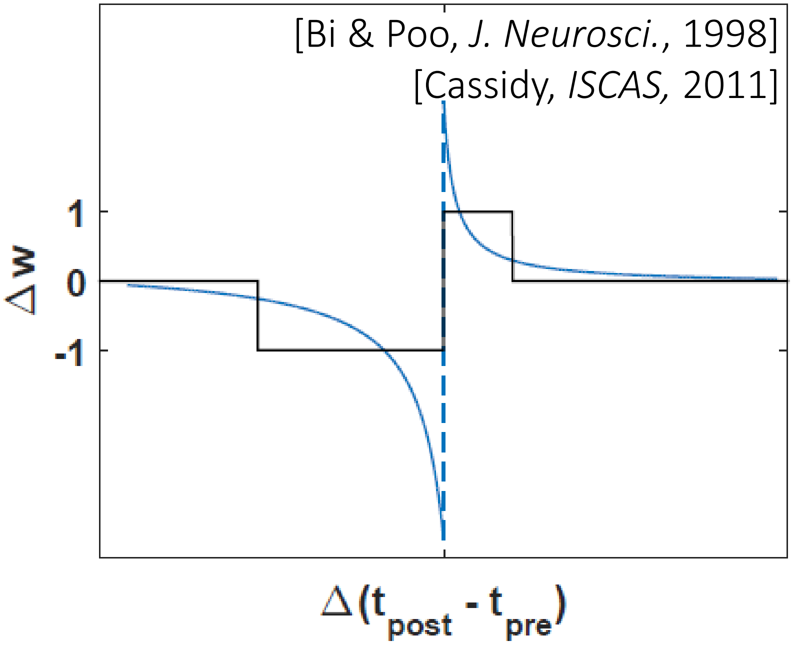
3. How can we morph these questions into interesting solutions for on-device-learning?

Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

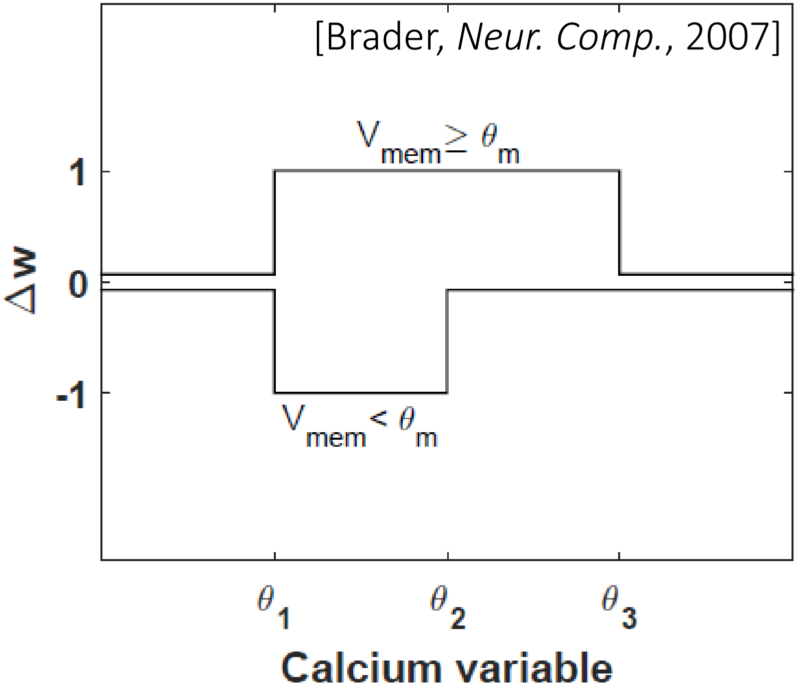
AI algorithms
Neuroscience

Spike-timing-dependent plasticity (STDP)

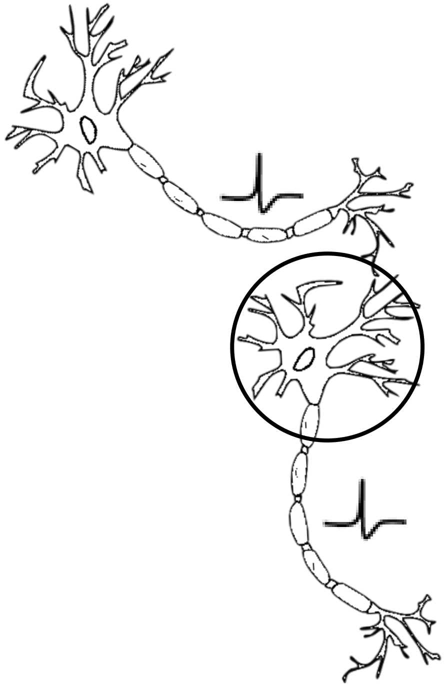


✔ **Local**

Spike-dependent synaptic plasticity (SDSP)



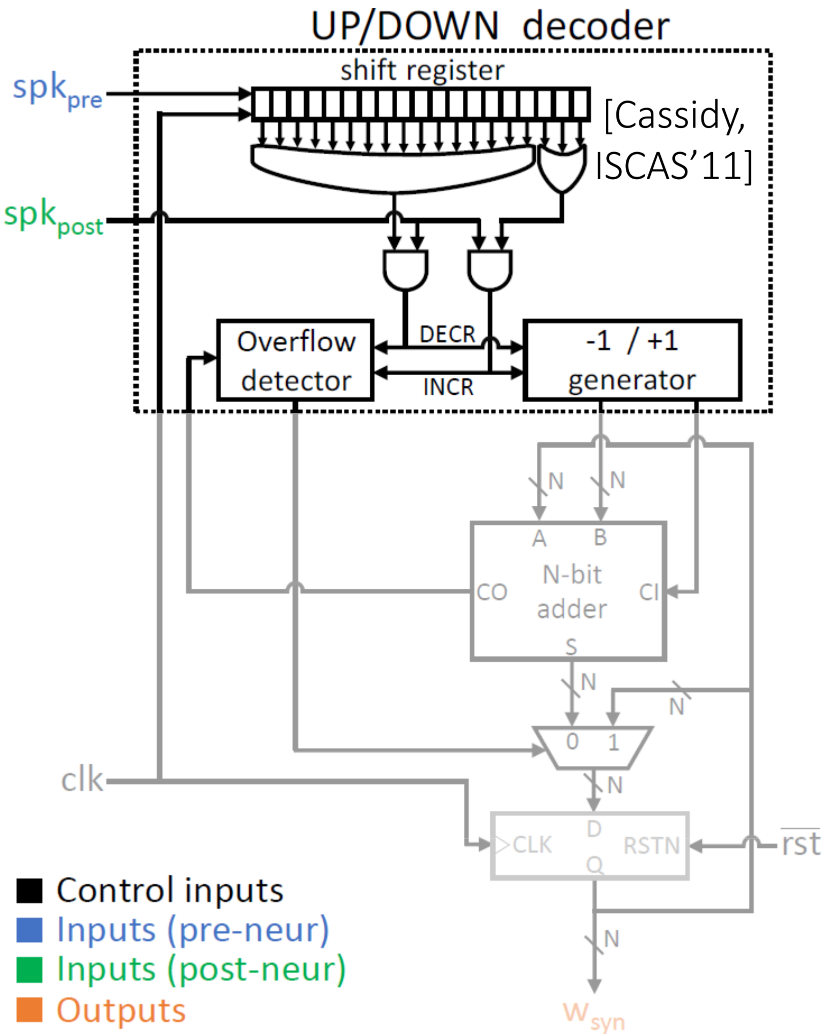
✔ **Local**



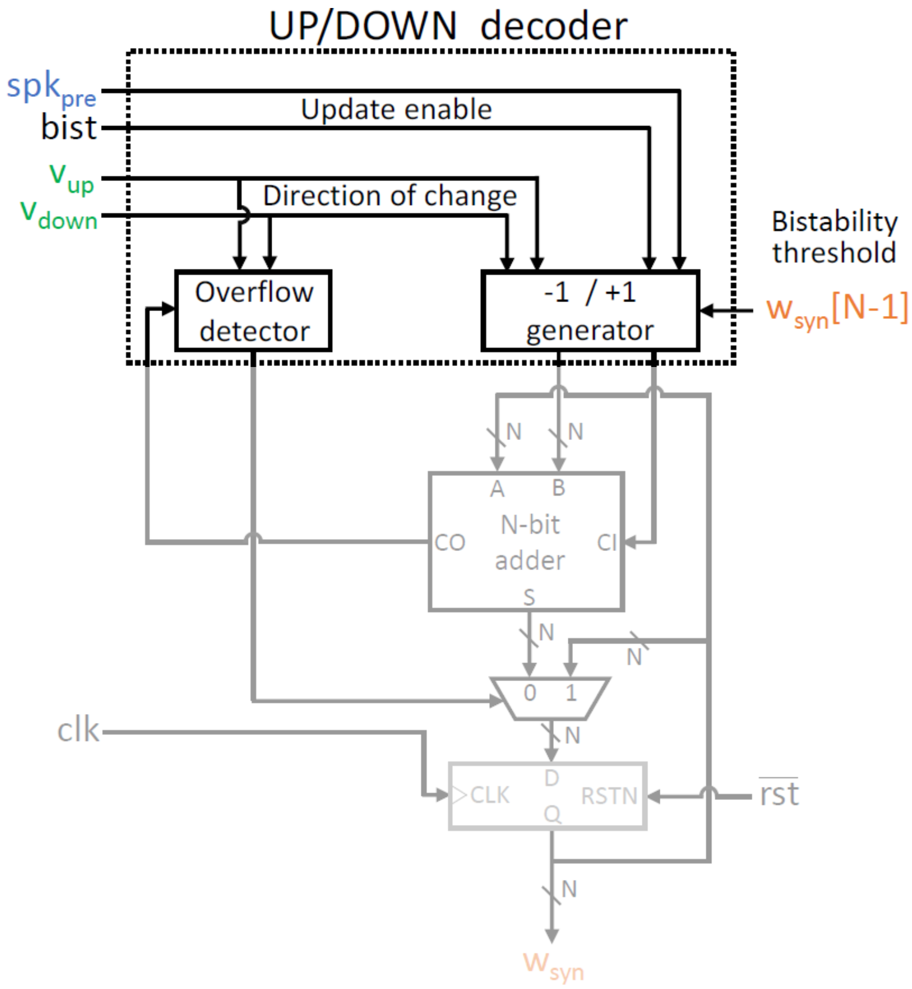
Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

AI algorithms
↑
Neuroscience



STDP



SDSP

[Frenkel, *Trans. BioCAS*, 2019]

AI algorithms

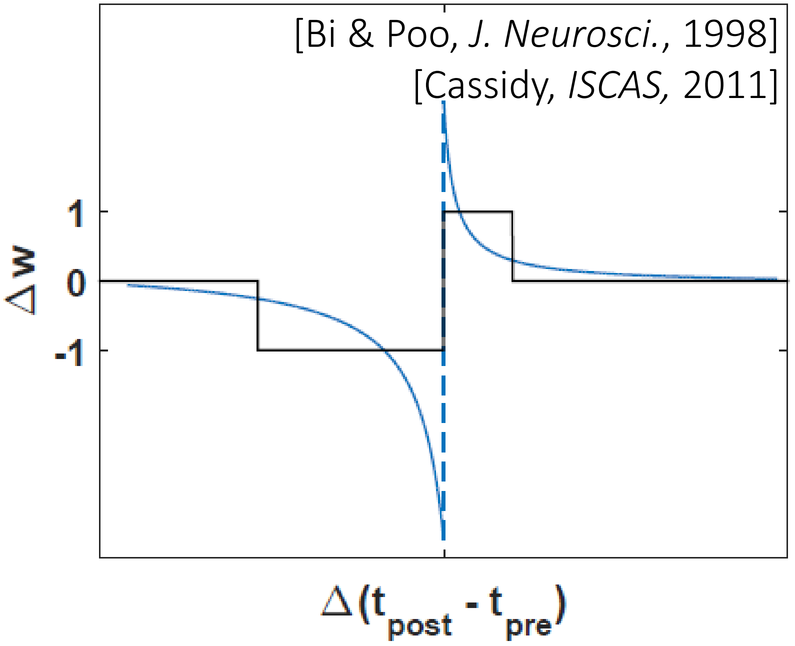
↑

Neuroscience

Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

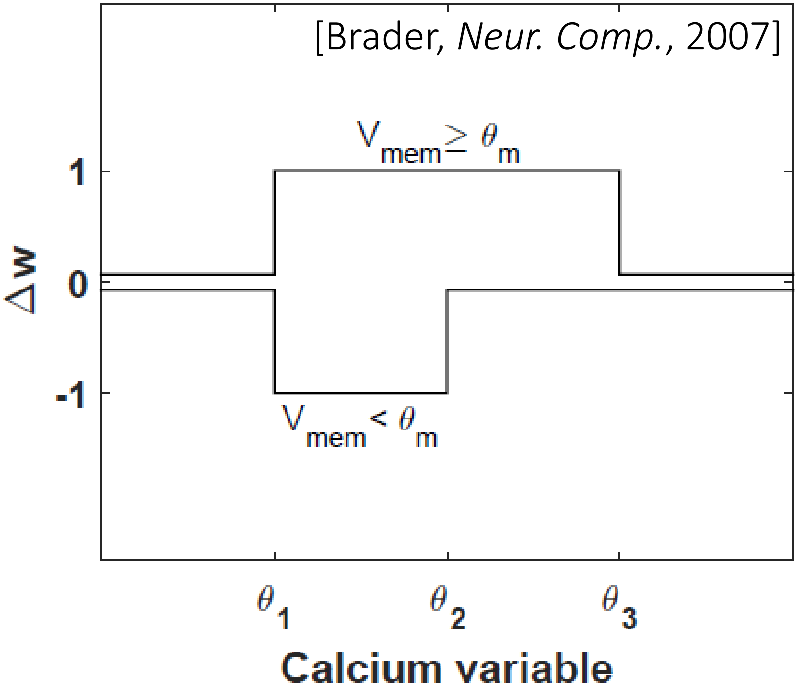
Spike-timing-dependent plasticity (STDP)



✔ **Local** in space

✖ **Non-local** in time

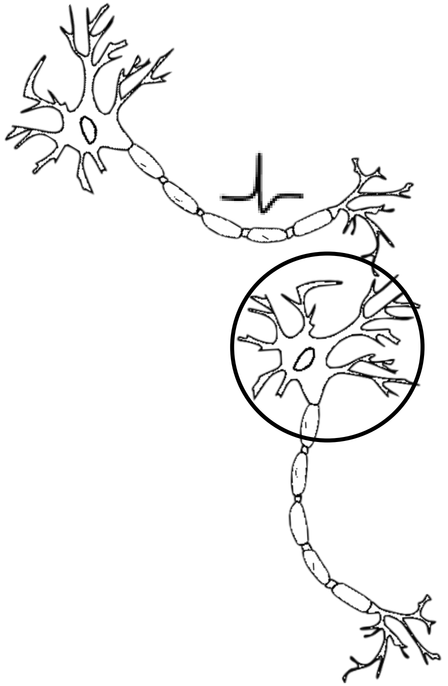
Spike-dependent synaptic plasticity (SDSP)



✔ **Local** in space

✔ **Local** in time

Huge savings in silicon



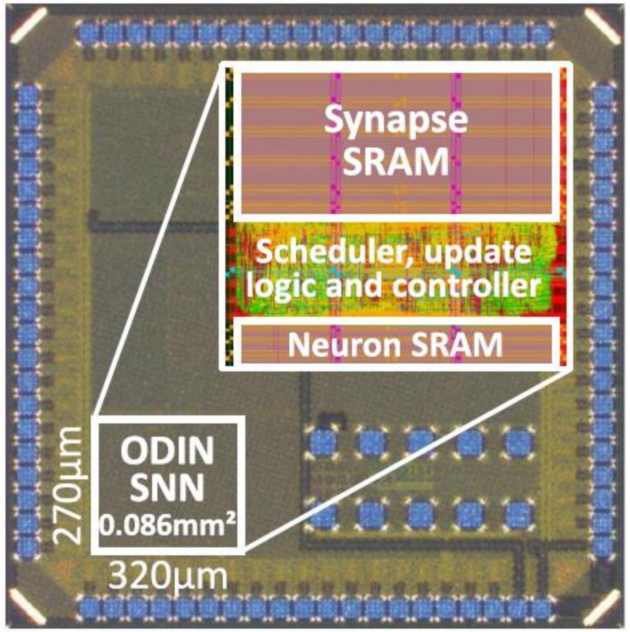
[Clopath and Gerstner, *Front. Syn. Neuro.*, 2010]

AI algorithms ↑ Neuroscience

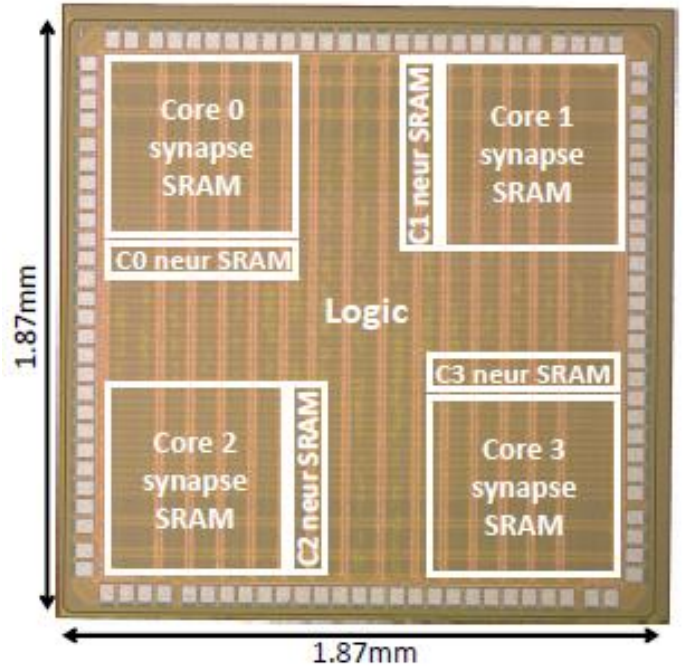
Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

ODIN (single-core)



MorphIC (quad-core)



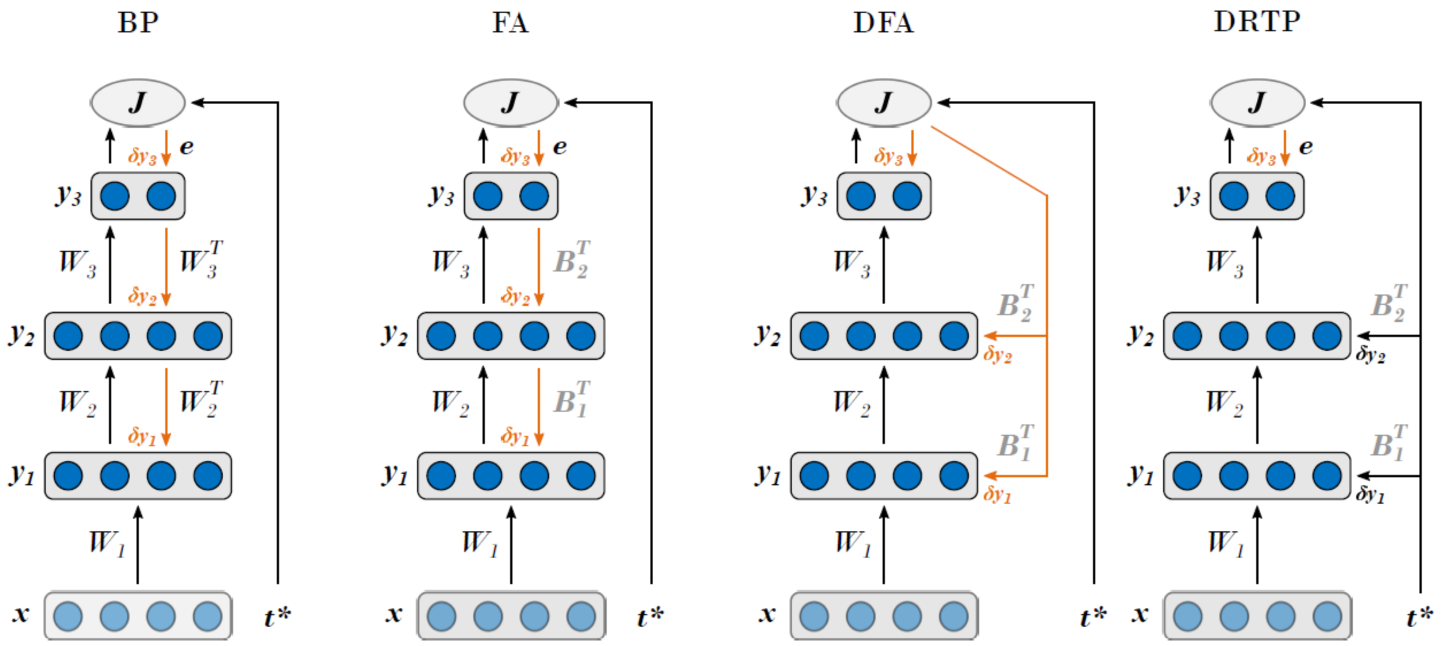
open source hardware

- ✔ Record synaptic density
- ✔ Energy efficiency competitive with analog mixed-signal designs
- ✔ Large feature set (incl. synaptic plasticity)
  ...but quite painful to exploit!

AI algorithms → Neuroscience

Neural network training – Bio-plausibility as the end goal

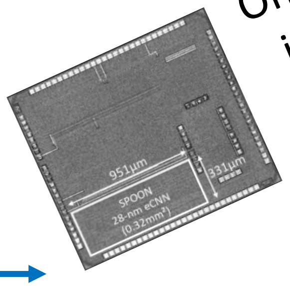
Synergy with hardware: latency, memory access patterns



	BP	FA	DFA	DRTP
δy_k	$\frac{\partial J}{\partial y_k} = W_{k+1}^T \delta z_{k+1}$	$B_k^T \delta z_{k+1}$	$B_k^T e$	$B_k^T t^*$
Weight-transport-free	✗	✓	✓	✓
Update-unlocked	✗	✗	✗	✓

↘ Computational and memory cost ↘

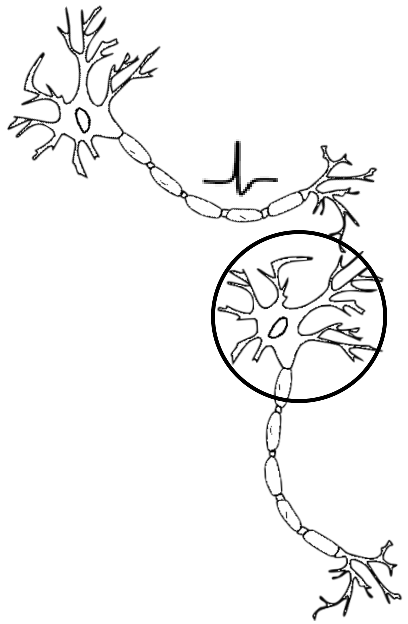
Output-independent target signals are also found in the brain! [Magee & Grienberger, *Ann. Rev. Neuro.*, 2020]



Only ~15% overhead in power and area [Frenkel, *ISCAS'20*] (🏆 Best paper award)

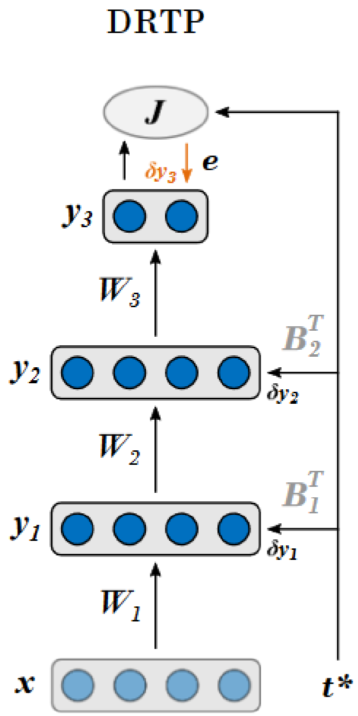
HW efficiency and bio-plausibility are often two sides of the same coin!

Many more examples: quantization, stochastic computing, event-driven computation,…



Designing efficient hardware hints toward bio-plausible mechanisms

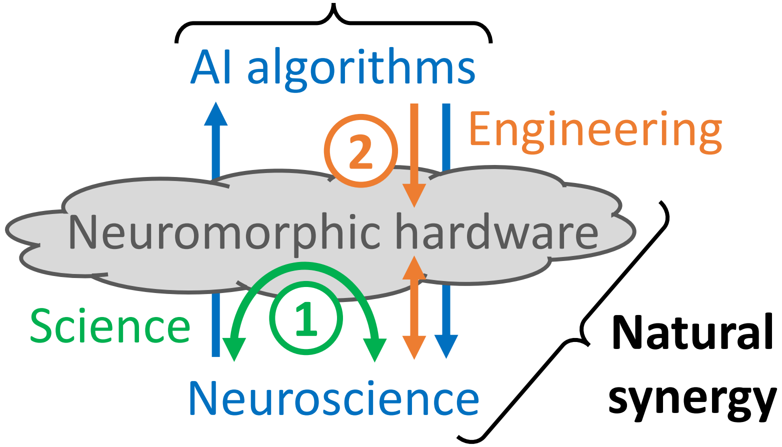
Bringing AI closer to neuroscience leads to hardware efficiency



From neuroscience to AI and back again

Which starting point? Which perspective?

AI without hardware is unsustainable



1 **Bottom-up science-driven approach**

- ✔ Analysis-by-synthesis
- ✘ Difficult to scale efficiently to real-world problems

2 **Top-down engineering-driven approach**

- ✔ Starts from working solutions to real-world problems
- ✘ Which "salt & pepper" from neuroscience?

Neuromorphic intelligence:

2 should be fed by 1

Outline

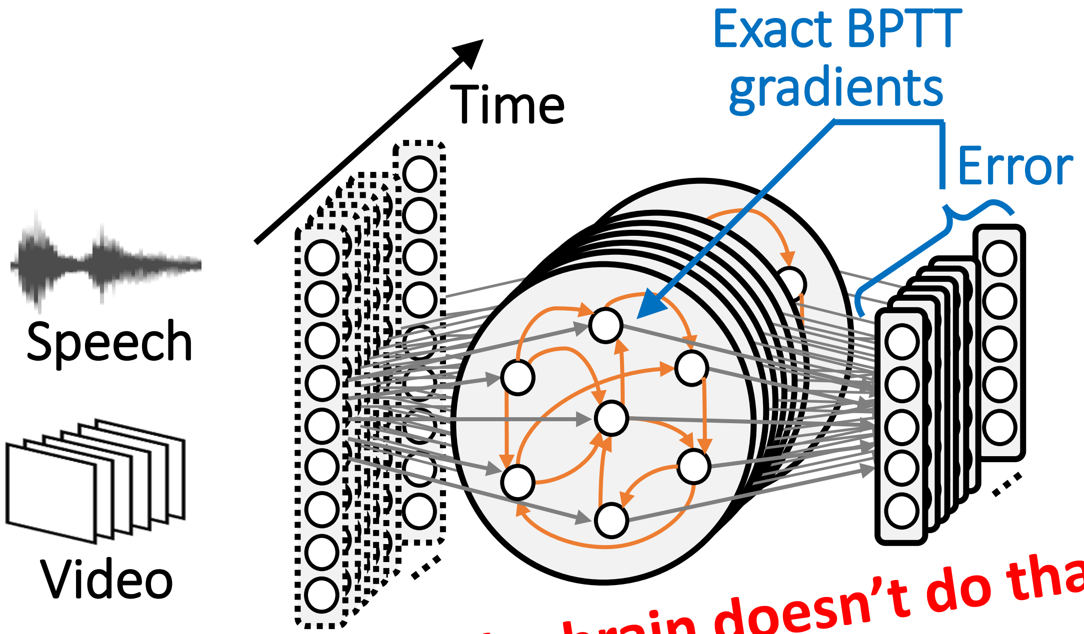
1. From neuroscience to AI and back again… …which perspective? …which starting point?

2. Why should we bother with neuroscience?

3. How can we morph these questions into interesting solutions for on-device-learning?

Why is on-chip learning over second-long timescales difficult?

Let's solve a yet unsolved engineering challenge!



Obviously, the brain doesn't do that!

- Unrolling in time: very deep network
  (current learning ICs for static stimuli: ≤3 layers)
- Intractable memory/latency requirements
- No end-to-end on-chip solution to date
  (you still need costly external memory!)

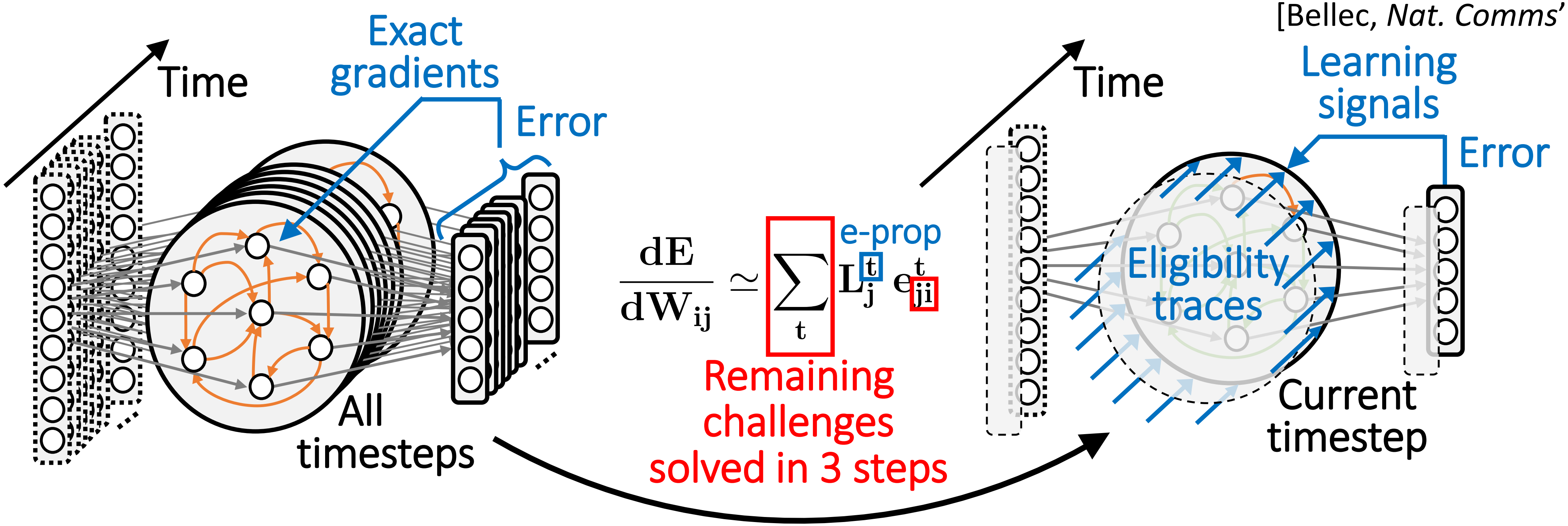
<u>Key challenge</u>: End-to-end on-chip learning over long timescales while keeping a fine-grained temporal resolution

The bio-inspired solution

Backward- vs. forward-mode training

Backprop through time (BPTT, backward) | **Eligibility propagation (e-prop, forward)**

[Bellec, *Nat. Comms*'20]



Biological plausibility ↗

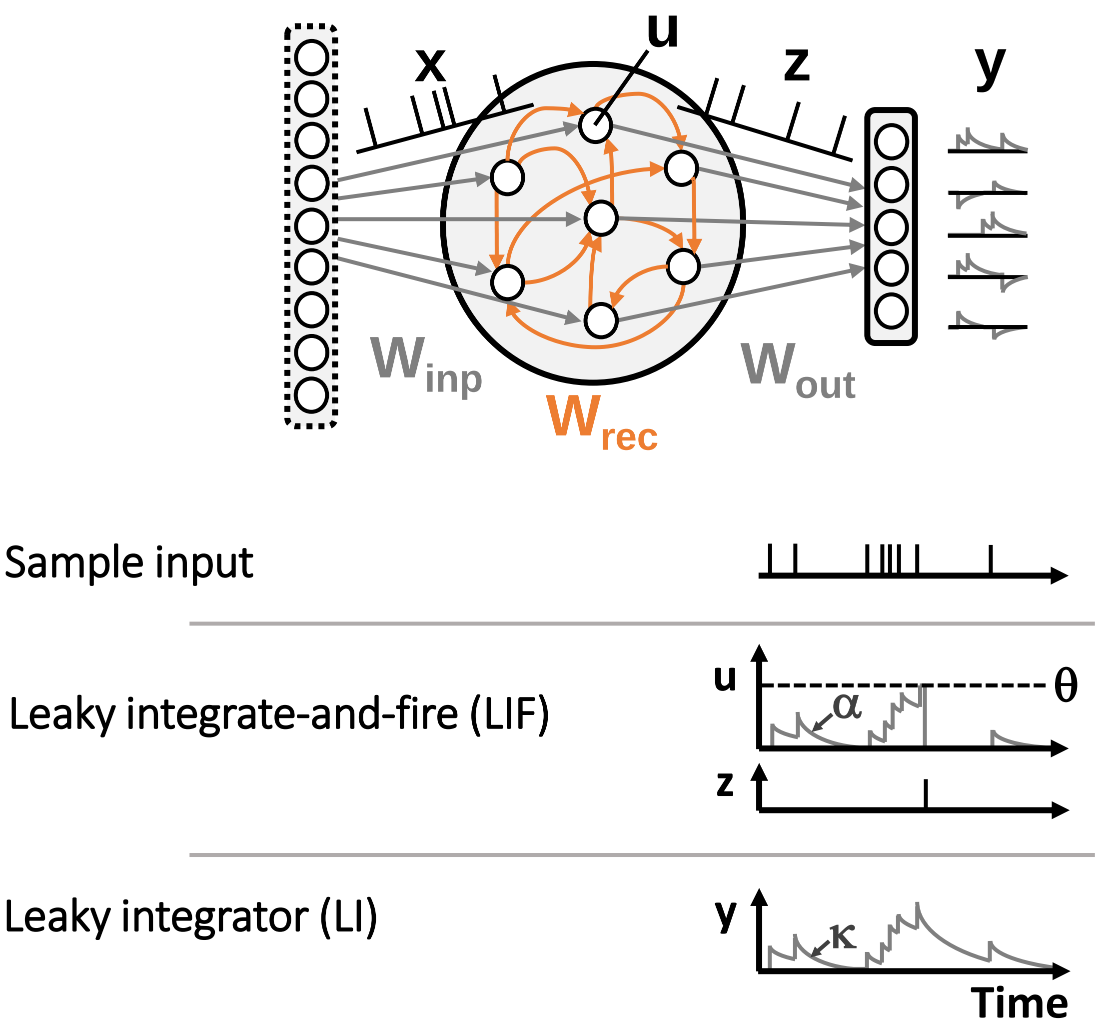
Space and time locality ↗

On-chip memory requirements ↘

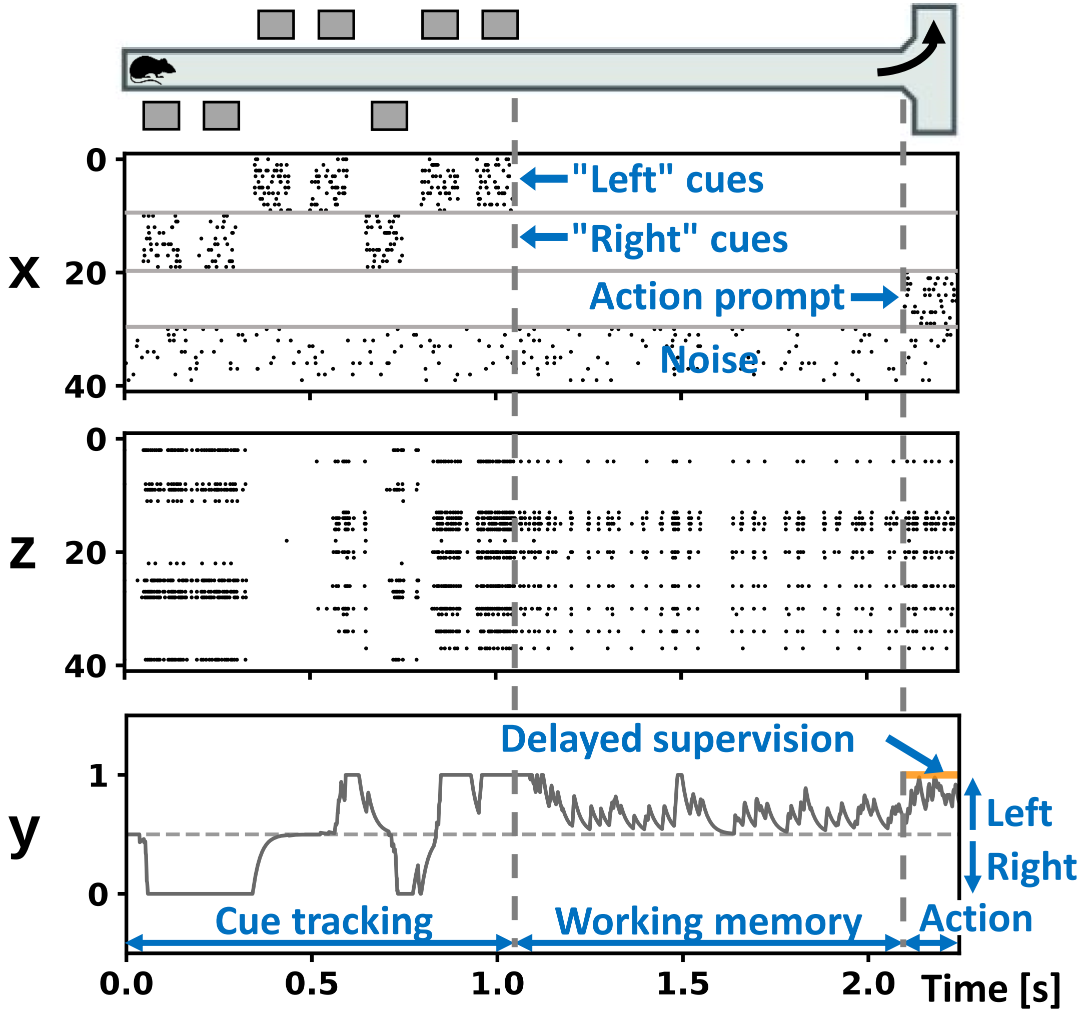
Algorithmic developments toward efficient long-term on-chip training

Network definitions and evaluation task

Network model



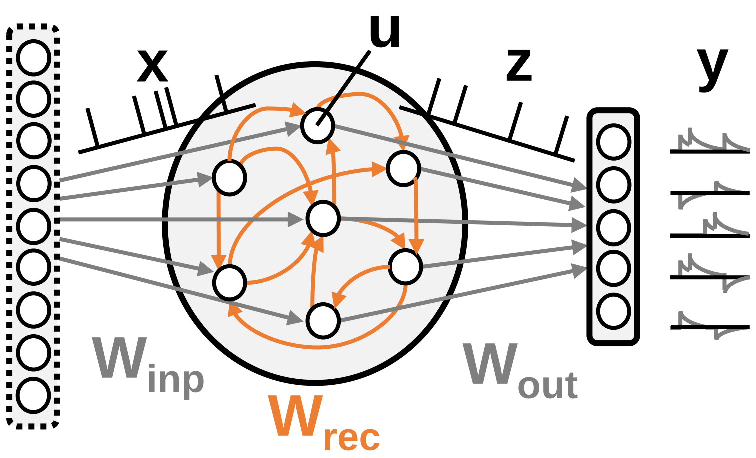
Delayed-supervision navigation task



Algorithmic developments – Step 1

Neuron model selection

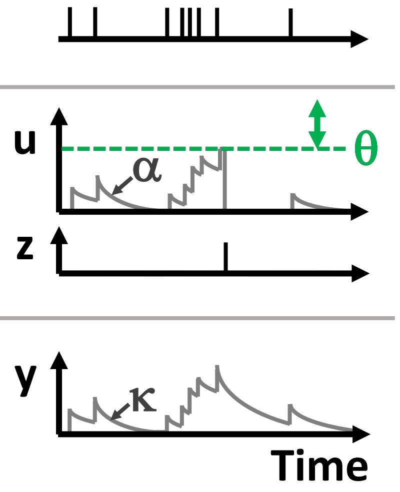
Network model



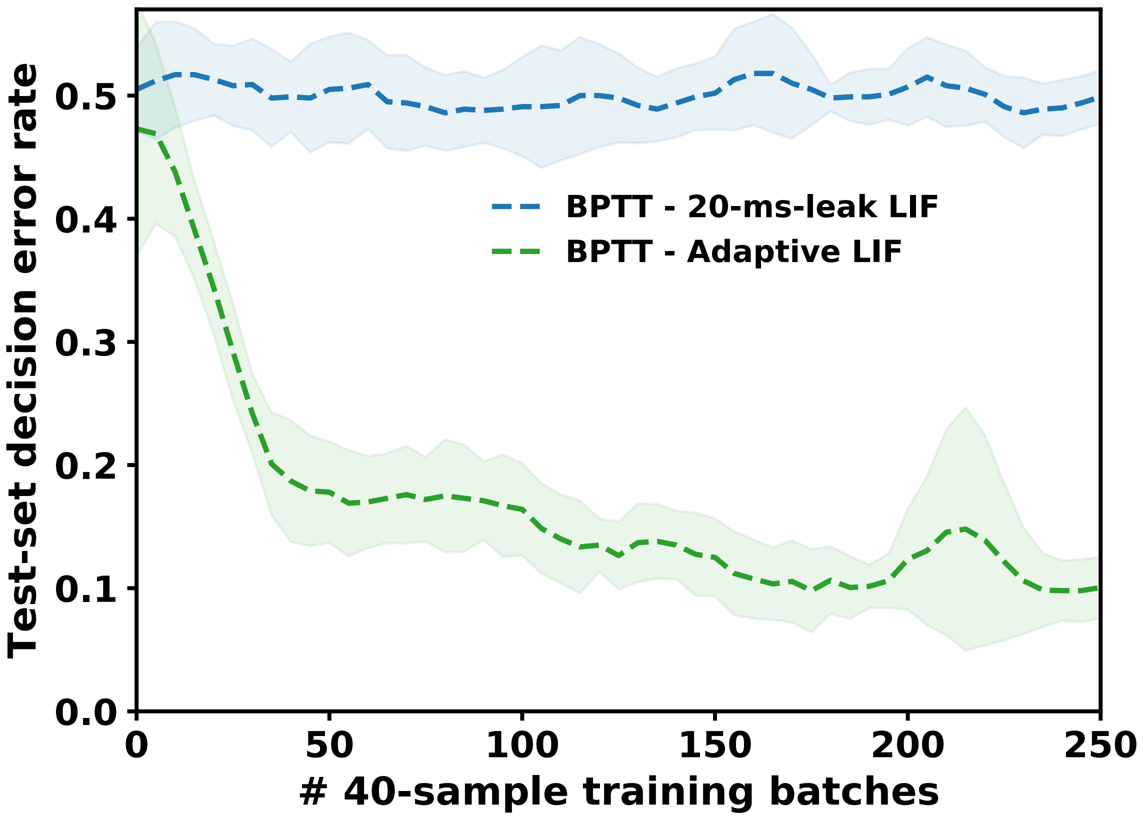
Sample input

Leaky integrate-and-fire (LIF)
or adaptive LIF (ALIF)

Leaky integrator (LI)



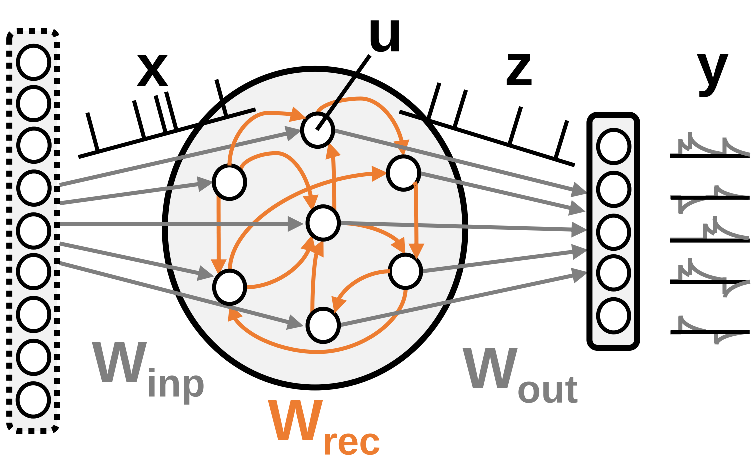
Task performance



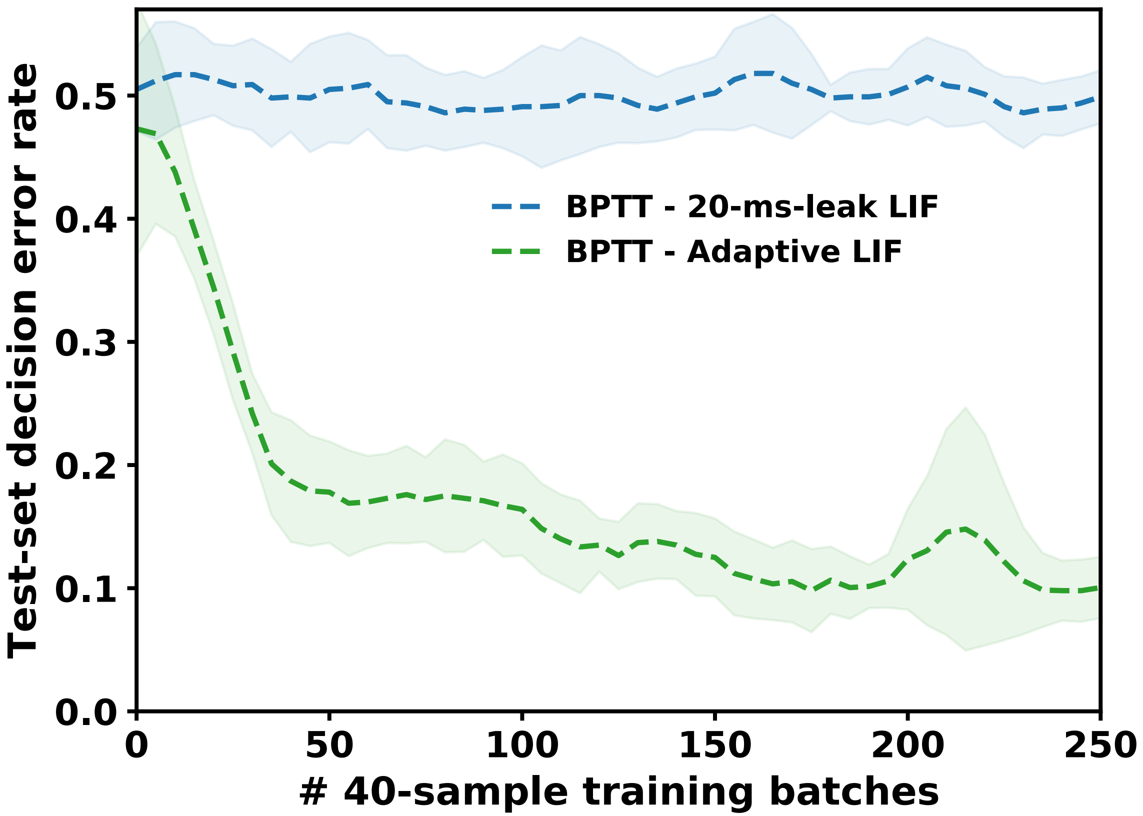
Algorithmic developments – Step 1

Neuron model selection

Network model



Task performance



Leaky integrate-and-fire (LIF):

- ❌ Only a short time constant (~20-ms leak)

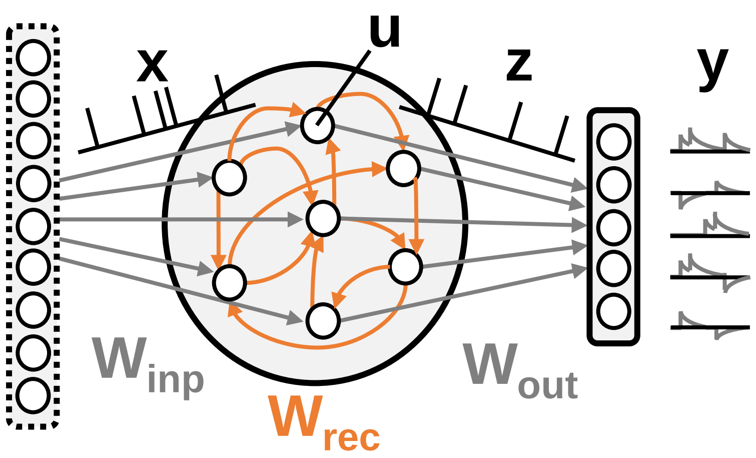
Adaptive LIF (ALIF):

- ✔ Embeds threshold adaptation over 100s of ms

Algorithmic developments – Step 1

Neuron model selection

Network model



Leaky integrate-and-fire (LIF):

- ❌ Only a short time constant (~20-ms leak)
- ✅ Eligibility traces: simple activity LPF

Adaptive LIF (ALIF):

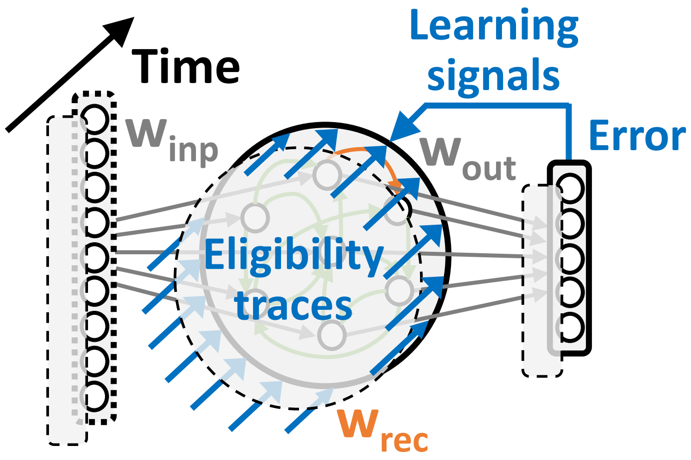
- ✅ Embeds threshold adaptation over 100s of ms
- ❌ ET: Complex per-synapse multi-scale filtering

LIF with configurable leak:

- ✅ Flexible time constant (ms to sec)
- ✅ Eligibility traces: simple activity LPF
- ≈ Less biologically plausible

Algorithmic developments – Step 2

Space and time locality



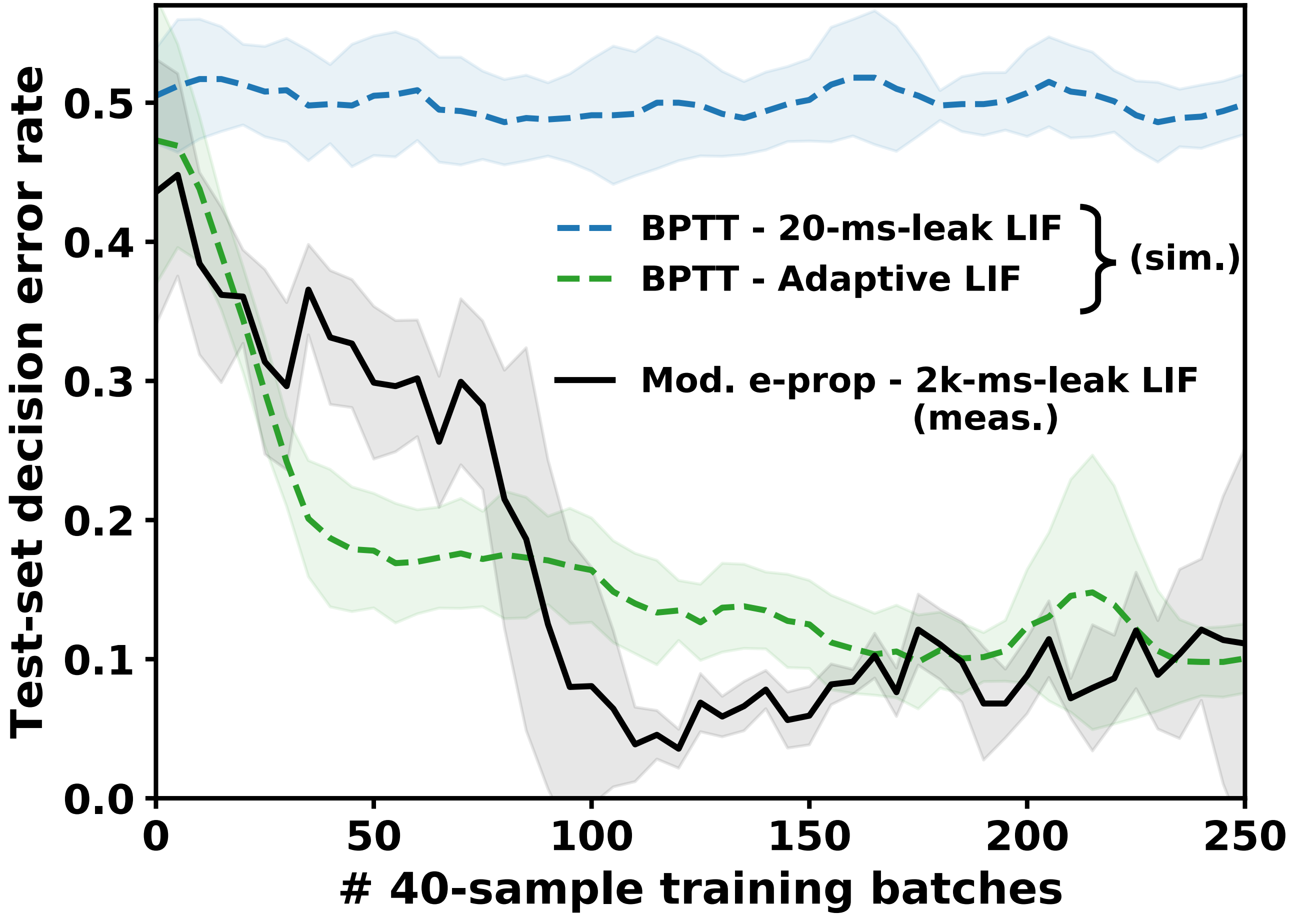
$$\frac{\mathbf{dE}}{\mathbf{dW_{ij}}} \simeq \cancel{\sum_{\mathbf{t}}} \mathbf{L_j^t} \; \mathbf{e_{ji}^t}$$

Local decoupling of space and time

Step 3 – Stochastic weight updates allow reducing weight resolution to 8 bits.

Algorithmic developments

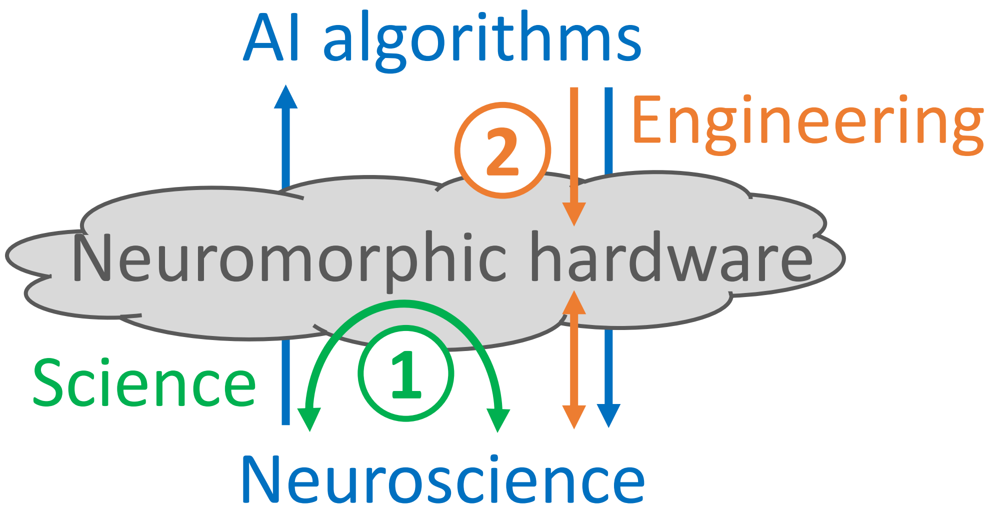
Final performance



- Competitive with a BPTT-trained network of ALIF neurons
- Memory overhead reduced to 0.8% of the inference-only network

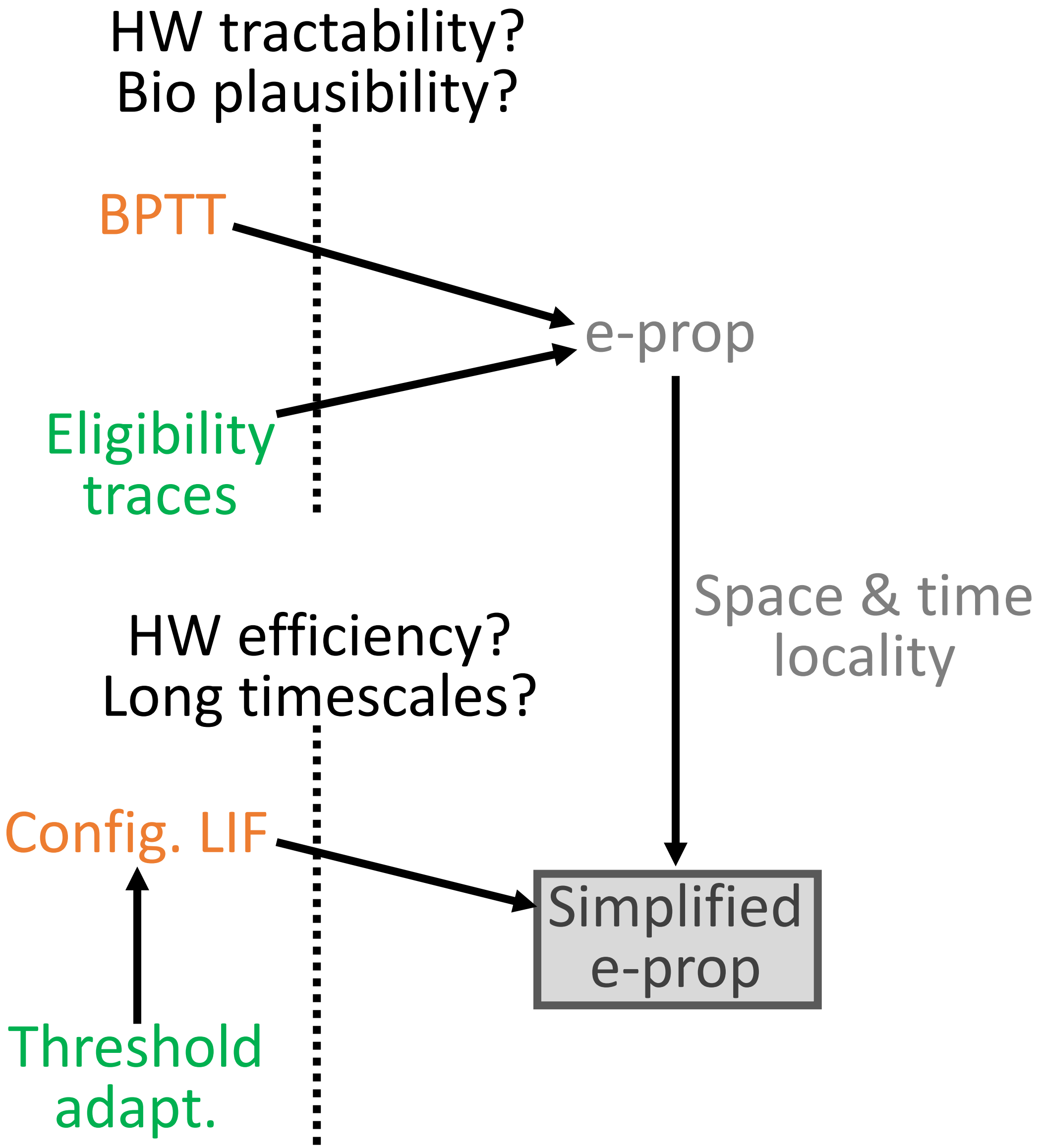
From neuroscience to AI and back again

Which starting point? Which perspective?



Neuromorphic intelligence:

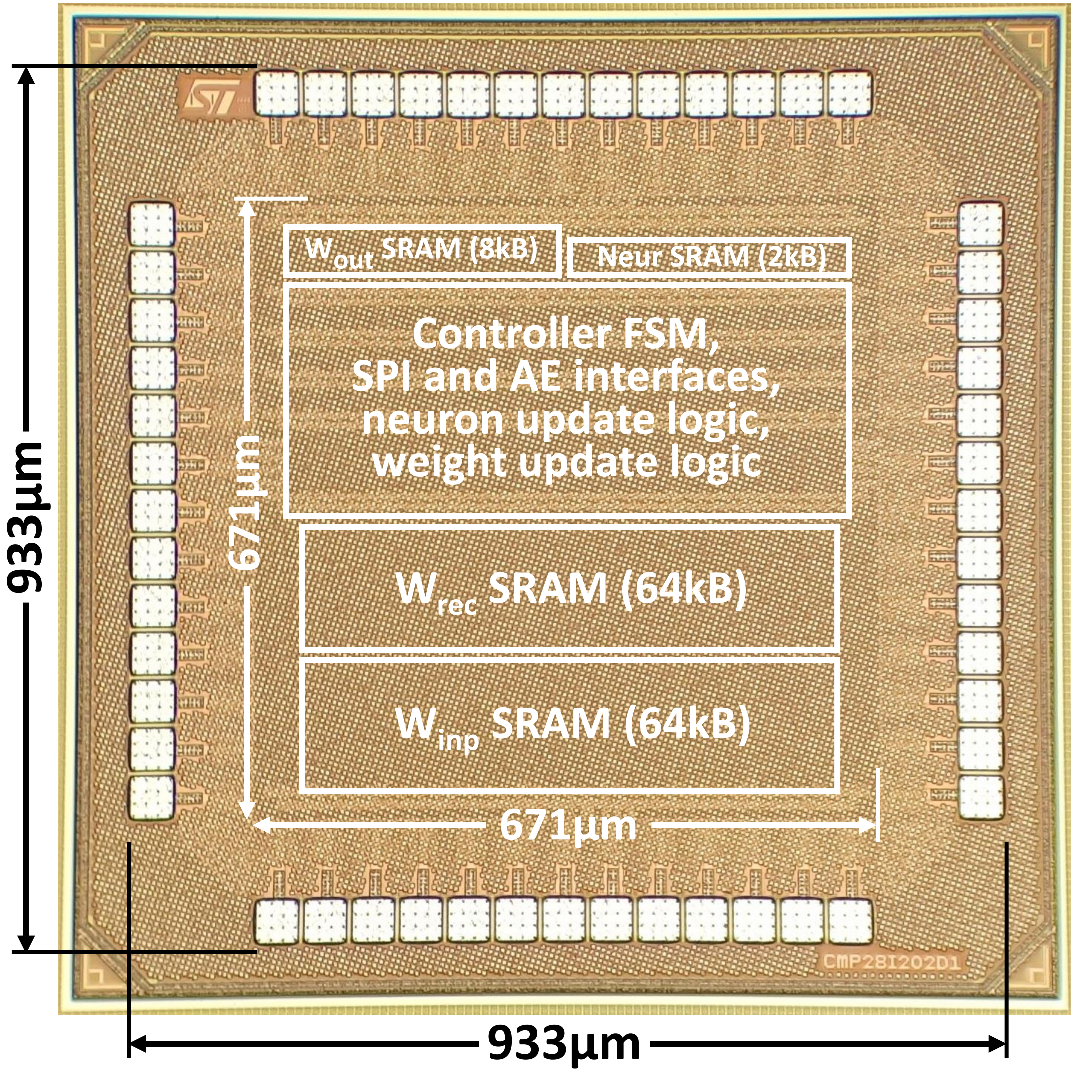
2 shoud be fed by 1



ReckOn – Neuroscience and AI meet in efficient hardware

Chip microphotograph and summary

open source hardware

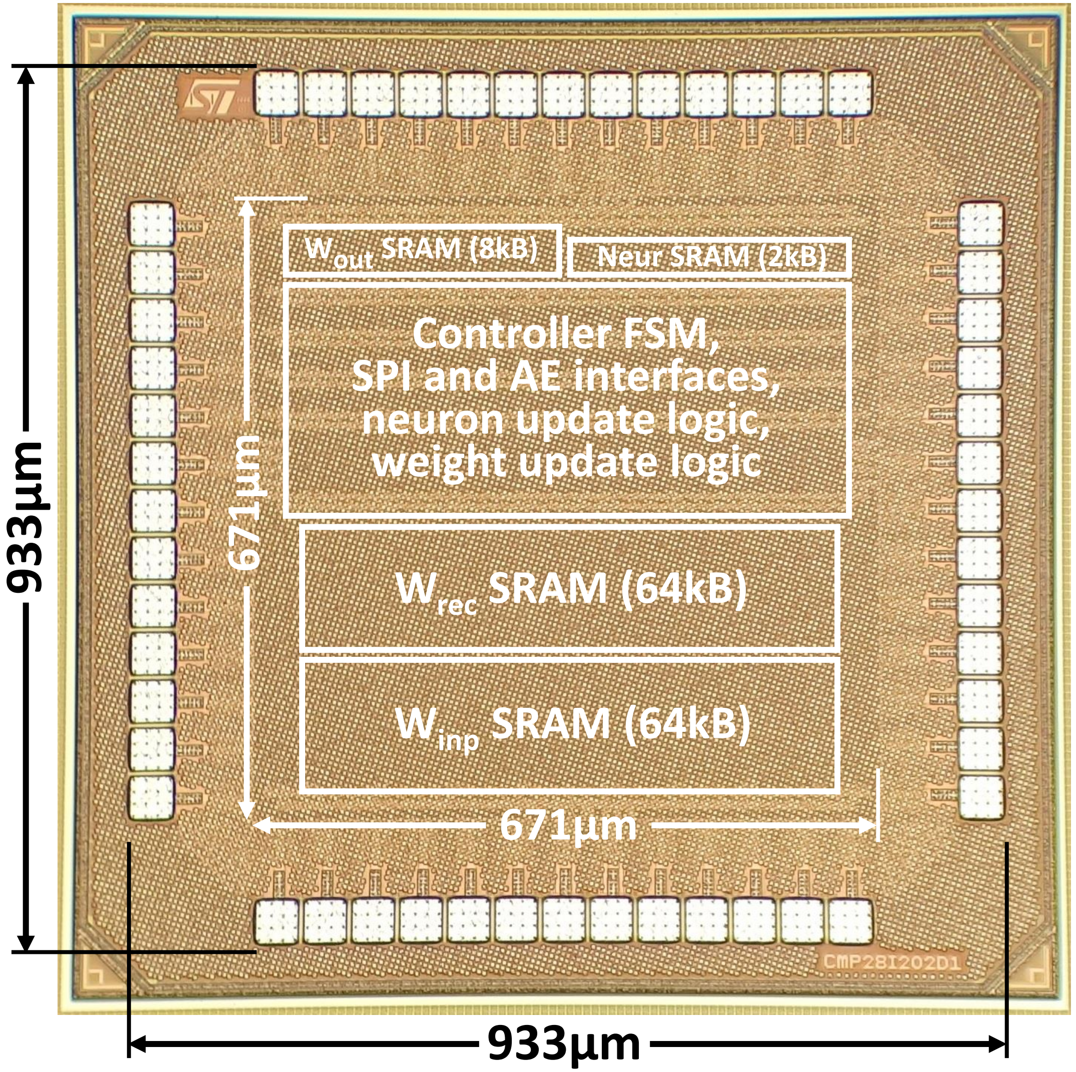


Technology	28nm FDSOI CMOS
Core size	0.67 x 0.67 mm² 0.45mm²
Die size	0.93 x 0.93 mm²
SRAM	138kB + 0kB ext. DRAM!
Network	Spiking RNN
Training timespan	Max. 32k steps

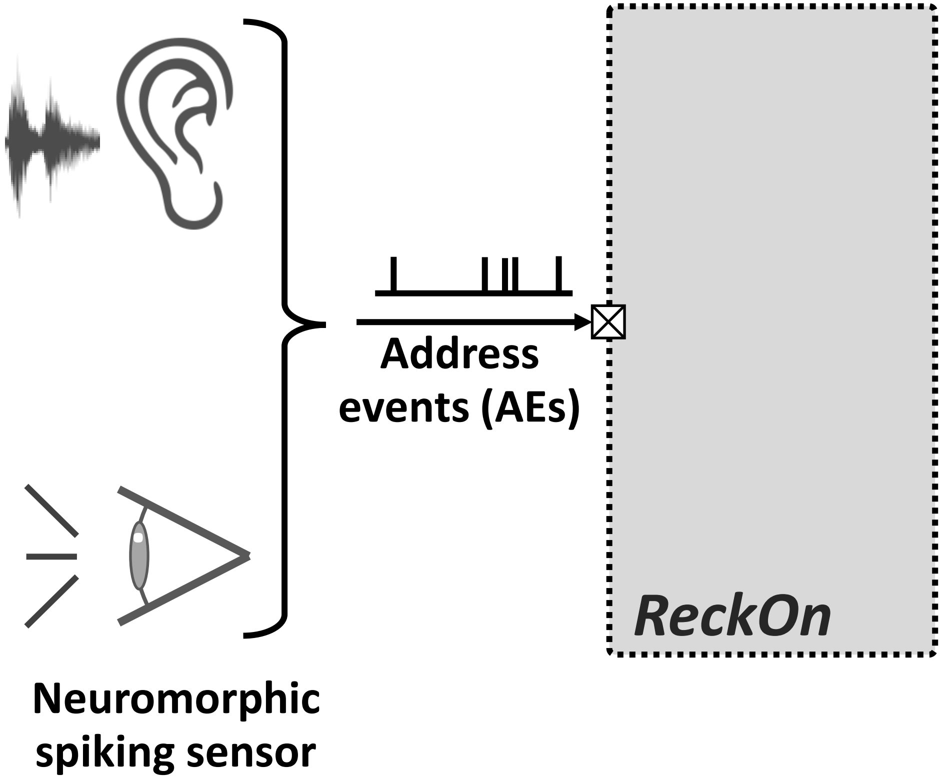
[Frenkel, *ISSCC*, 2022]

ReckOn – Neuroscience and AI meet in efficient hardware

Chip microphotograph and summary



- Event-driven / sparsity-aware computation
- Sensor-agnostic raw-data processing
- Task-agnostic processing and learning



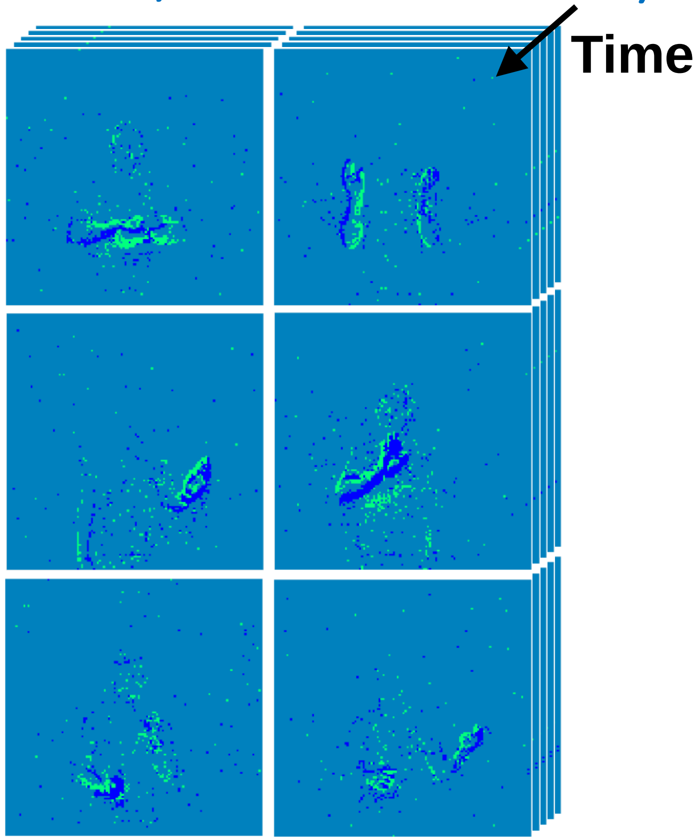
[Frenkel, *ISSCC*, 2022]

RecKon – Measurements and benchmarking

Three benchmarks that demonstrate task-agnostic learning

Vision

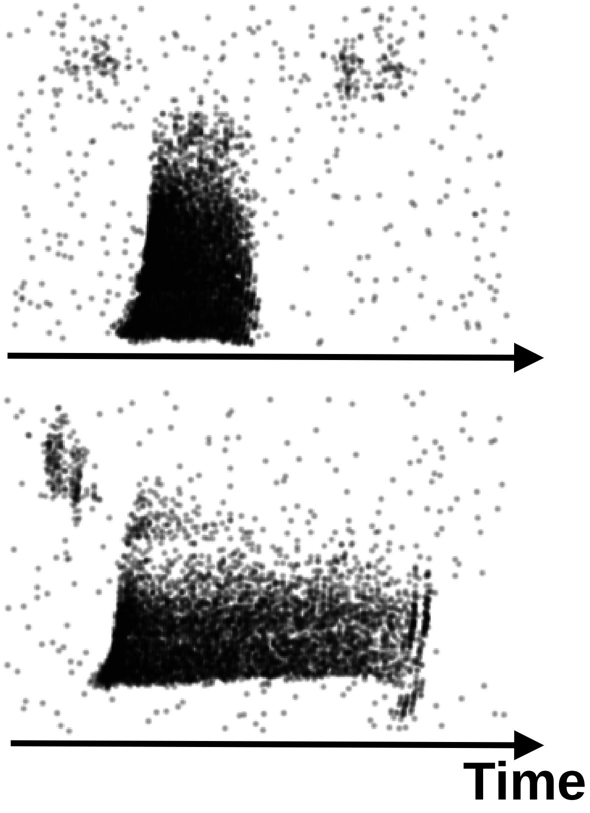
IBM DVS Gestures dataset
(10 classes, shrinked to 16x16)



Accuracy: 87.3% (28µW @0.5V)

Audition

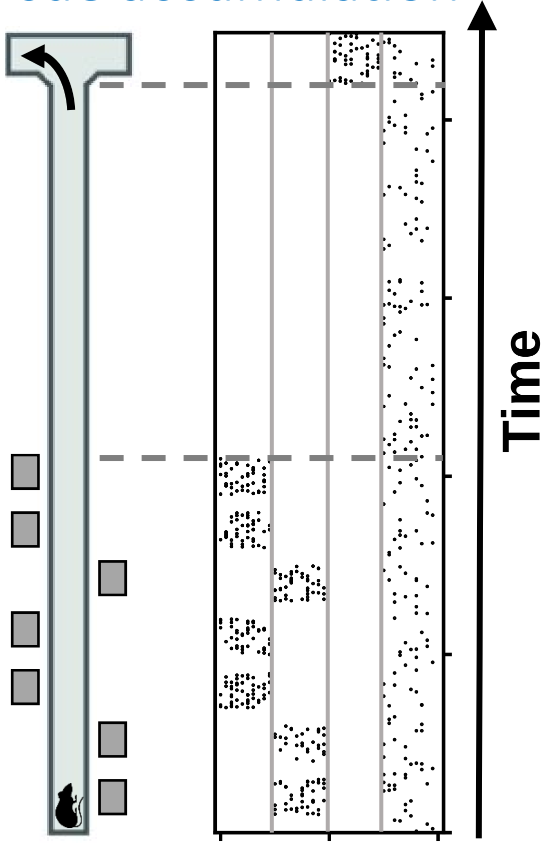
Spiking Heidelberg Digits (EN) dataset
(1-word KWS, 1:1 target vs. filler, 1:3 sub)



Accuracy: 90.7% (46µW @0.5V)

Navigation

Delayed-supervision
cue accumulation



Accuracy: 96.4% (14µW @0.5V)

[Frenkel, *ISSCC*, 2022]

ReckOn – What you should remember

Key elements toward neuromorphic edge intelligence

ReckOn merges AI, neuroscience and hardware to

achieve end-to-end on-chip learning over second-long timescales while keeping a milli-second temporal resolution, a yet unsolved challenge,

provide a low-cost solution: 0.45-mm² core area, <50µW for training at 0.5V, and 0.8-% memory overhead vs. inference-only network,

demonstrate task-agnostic learning with a spike-based encoding toward user customization and chip repurposing at the edge.

Neuromorphic intelligence outlines an exciting future for tiny on-device learning!

The *Cognitive Sensor Nodes and Systems* (CogSys) Team

We bridge the bottom-up (bio-inspired) and top-down (engineering-driven) design approaches toward neuromorphic intelligence.

Things we like (non-exhaustive list!):

- designing neuromorphic ICs and tinyML accelerators (mostly digital, going mixed-signal)
- bio-plausible training algorithms and synaptic plasticity mechanisms
- system-level optimization for autonomous sensorimotor agents, from sensing to decision

Positions will open soon!

Questions?

@C_Frenkel
cfrenkel
ChFrenkel
Charlotte-Frenkel
c.frenkel@tudelft.nl
chfrenkel.github.io

Main references:

- ODIN: [C. Frenkel et al., “A 0.086-mm² 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28nm CMOS,” *IEEE Trans. BioCAS*, 2019]
- MorphIC: [C. Frenkel et al. “MorphIC: A 65-nm 738k-synapse/mm² quad-core binary-weight digital neuromorphic processor with stochastic spike-driven online learning,” *IEEE Trans. BioCAS*, 2019]
- DRTP: [C. Frenkel, M. Lefebvre et al., “Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks,” *Frontiers in Neuroscience*, 2021]
- SPOON: [C. Frenkel et al., “A 28-nm convolutional neuromorphic processor enabling online learning with spike-based retinas,” *IEEE ISCAS*, 2020]
- **Review:** [C. Frenkel, D. Bol and G. Indiveri, “Bottom-up and top-down approaches for the design of neuromorphic processing systems: Tradeoffs and synergies between natural and artificial intelligence,” *Proceedings of the IEEE* (to appear), 2023]
- **ReckOn:** [C. Frenkel and G. Indiveri, “ReckOn: A 28-nm Sub-mm² Task-Agnostic Spiking Recurrent Neural Network Processor Enabling On-Chip Learning over Second-Long Timescales,” *IEEE International Solid-State Circuits Conference (ISSCC)*, 2022]

Open-sourced! github.com/ChFrenkel/ODIN

Open-sourced! github.com/ChFrenkel/DirectRandomTargetProjection

Already available in arxiv.org/abs/2106.01288

Open-sourced! github.com/ChFrenkel/ReckOn

TINY ML
On device learning Forum

Copyright Notice

www.tinyml.org

On device learning Forum

Copyright Notice