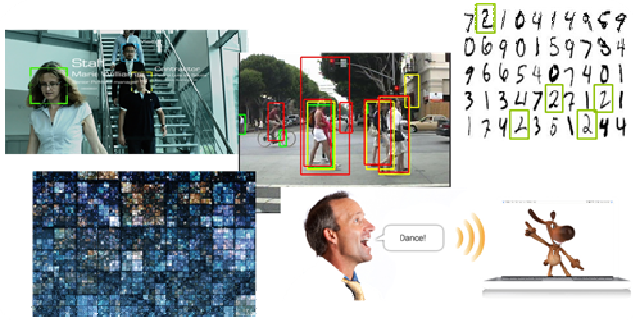
Using OxRAM arrays to mimic bio-inspired short and long term synaptic plasticity

Leti Memory Workshop | Thilo Werner | 27/06/2017

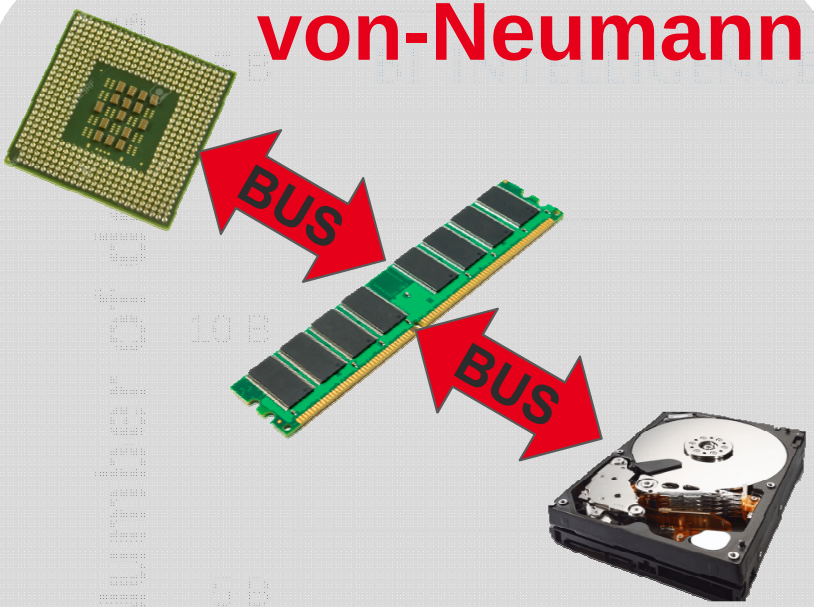
Collaboration by:
T. Werner, E. Vianello, O. Bichler, A. Grossi, E. Nowak, J.-F. Nodin, B. Yvert, B. DeSalvo, L. Perniola

Why Neuromorphic?



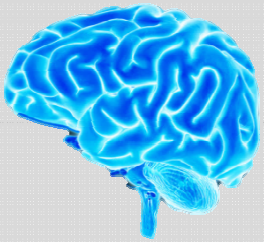
20 B

von-Neumann

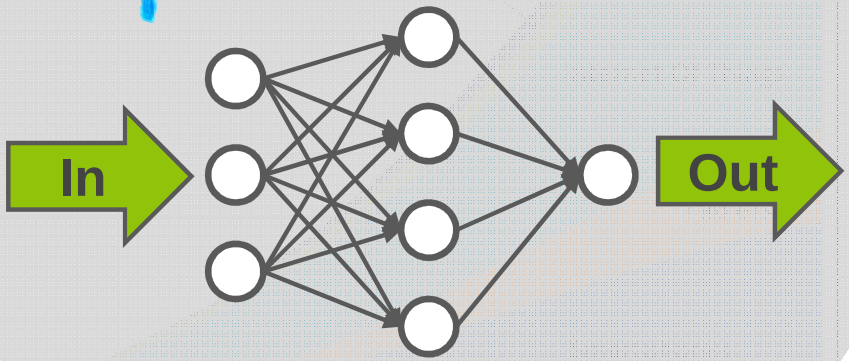


Computation ≠ Memory

non-von-Neumann



Computation
=
Memory



Low power devices

Abundant unstructured data

Real-time response

Moore's law

?

1 B

0

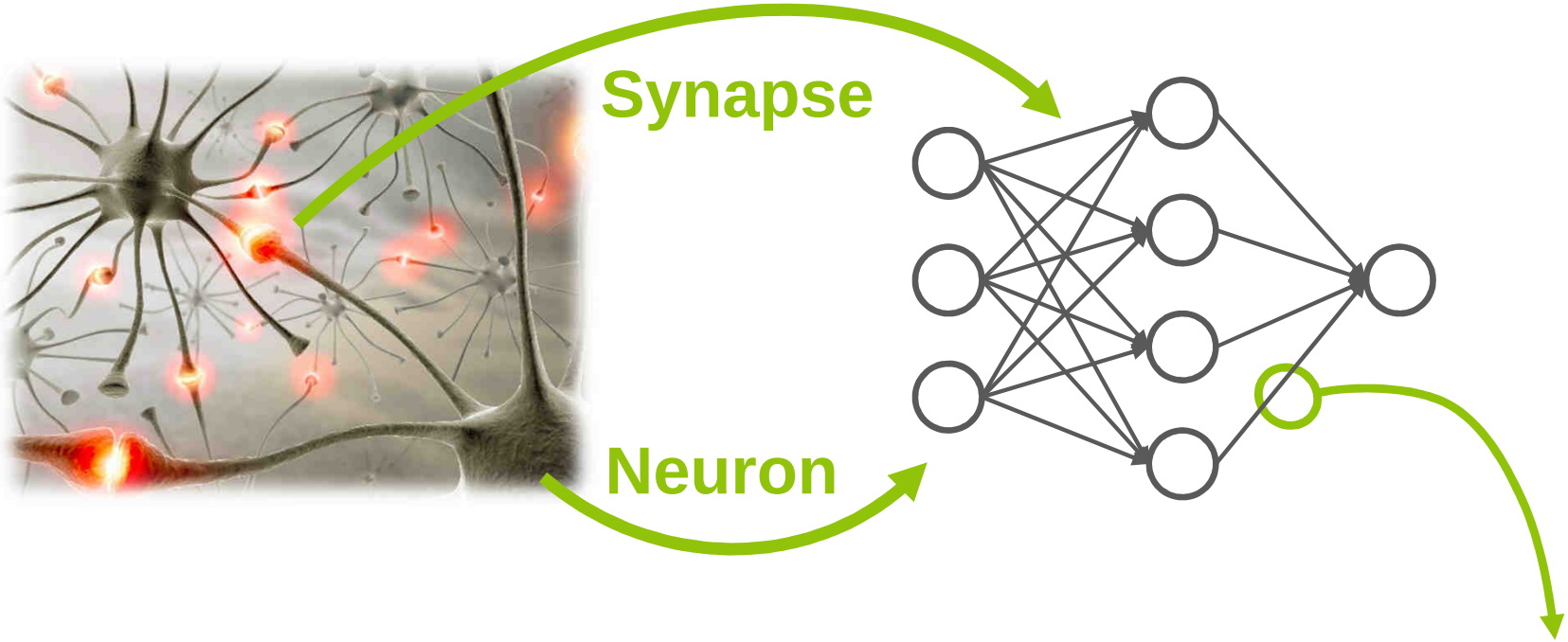
PCs

2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014E 2015E 2016E

Source: BI Intelligence Estimates

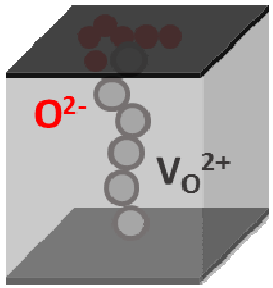
Execution of NVN on VN hardware ➔ $$

Neural Networks

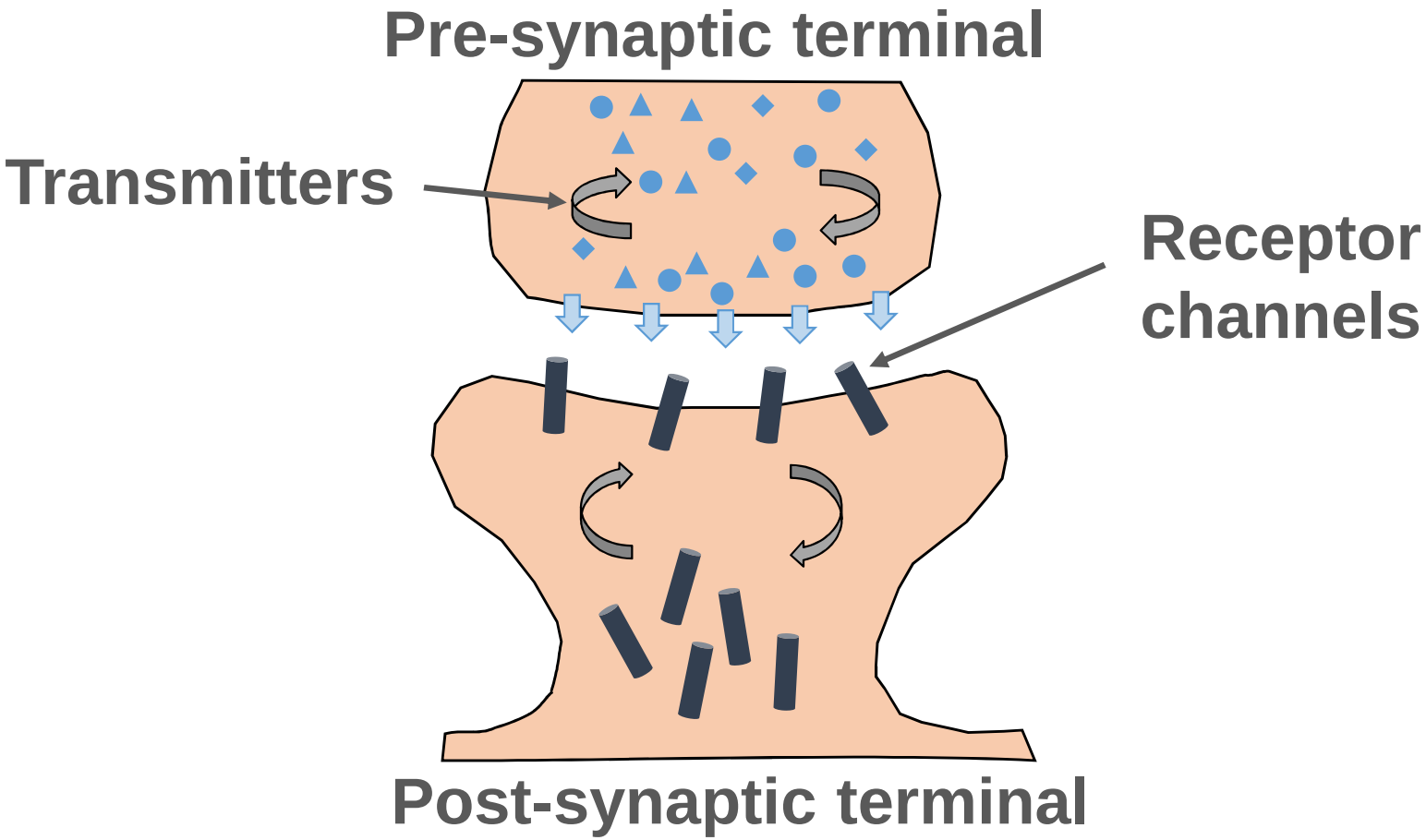


➢ **OxRAM to implement HW synapses:**

- high density
- online learning
- low power consumption



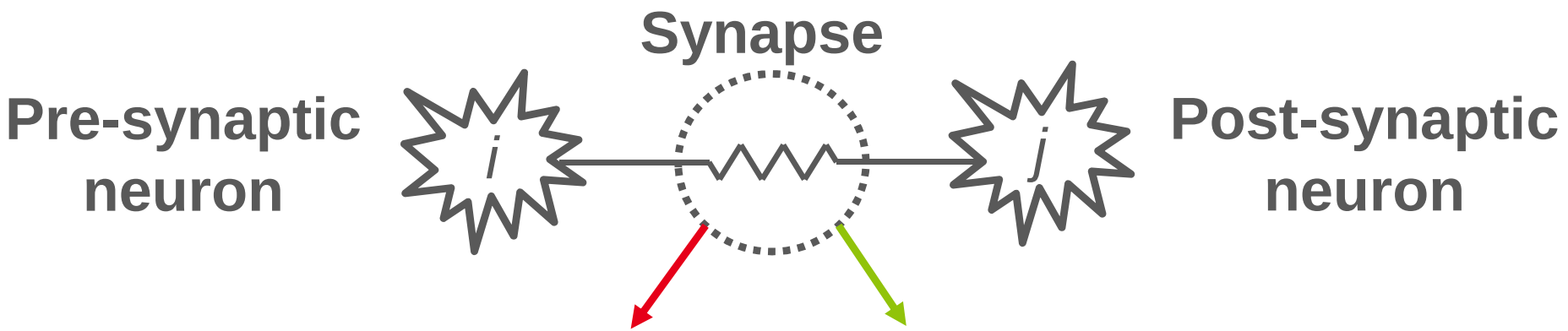
Synaptic plasticity in biology



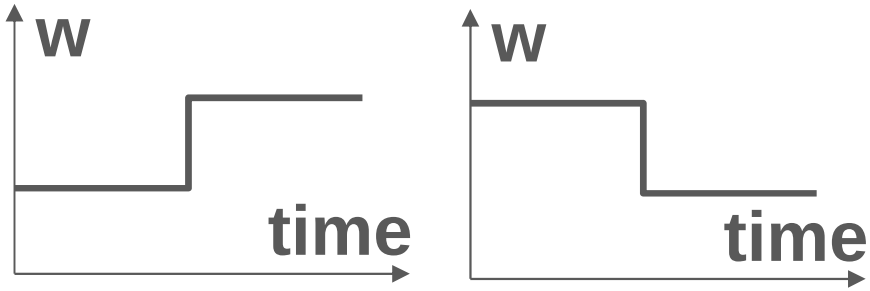
Transmitters ➔ Short Term Plasticity

Channels ➔ Long Term Plasticity

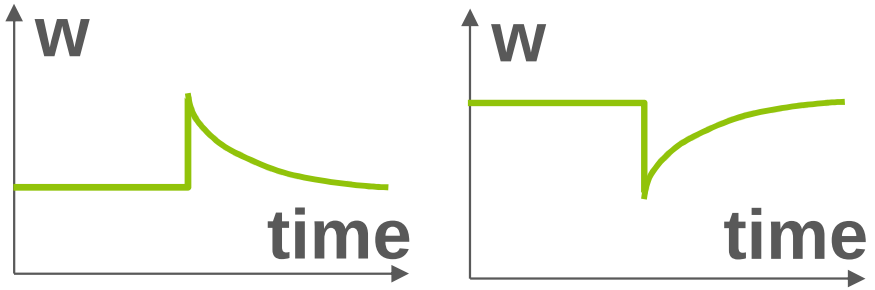
Simplified plasticity models



Long Term Plasticity



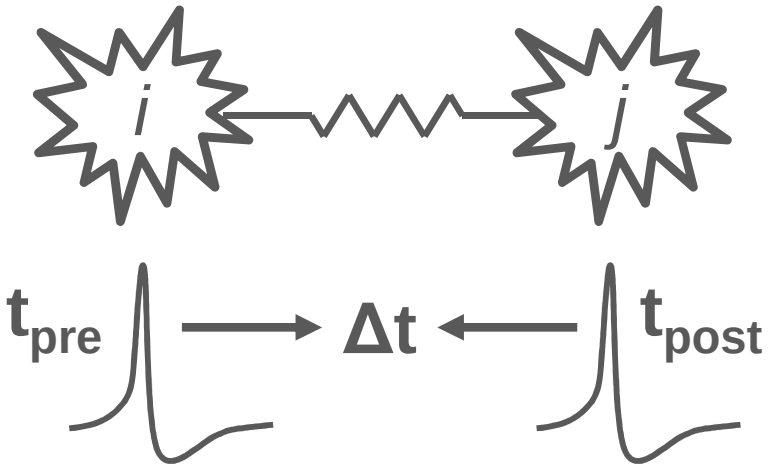
Short Term Plasticity

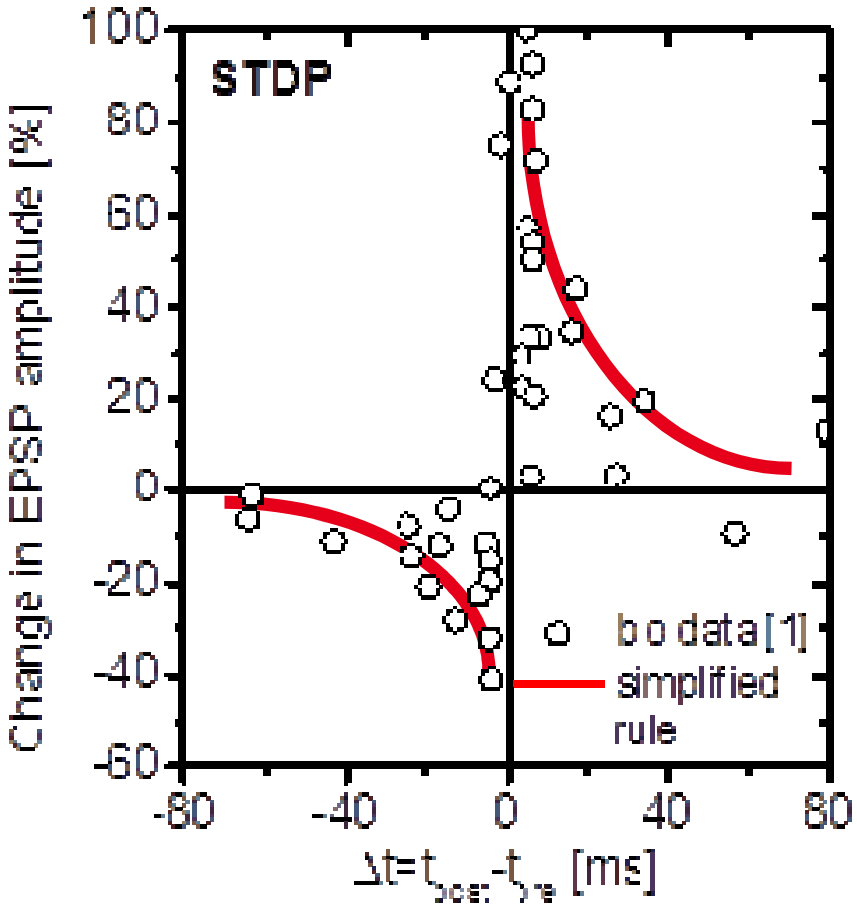


Emerging NVM demonstrated only for LTP

State-of-the-art: RRAM synapses

Long Term Plasticity (LTP)
➔Spike-Timing-Dependent Plasticity





STDP: correlation of pre- and post-synaptic activities

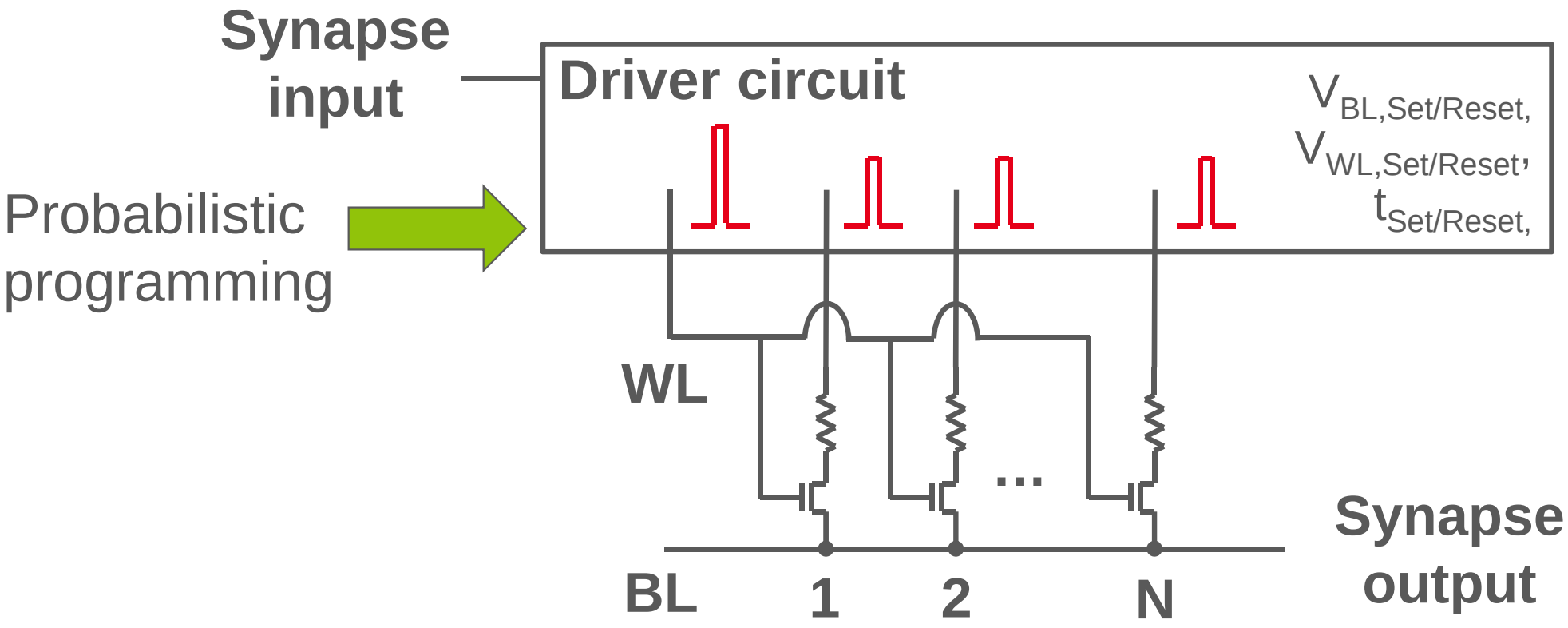
Objective

1. How to implement Short Term Plasticity with non-volatile OxRAM?
2. Is STP (i.e. ‘volatile’) useful for learning?

Outline

- Introduction
- ➢ **OxRAM synapses and STP**
- Co-implementation of STP & LTP
- STP impact in Spiking Neural Networks
- Summary

Multilevel OxRAM synapses



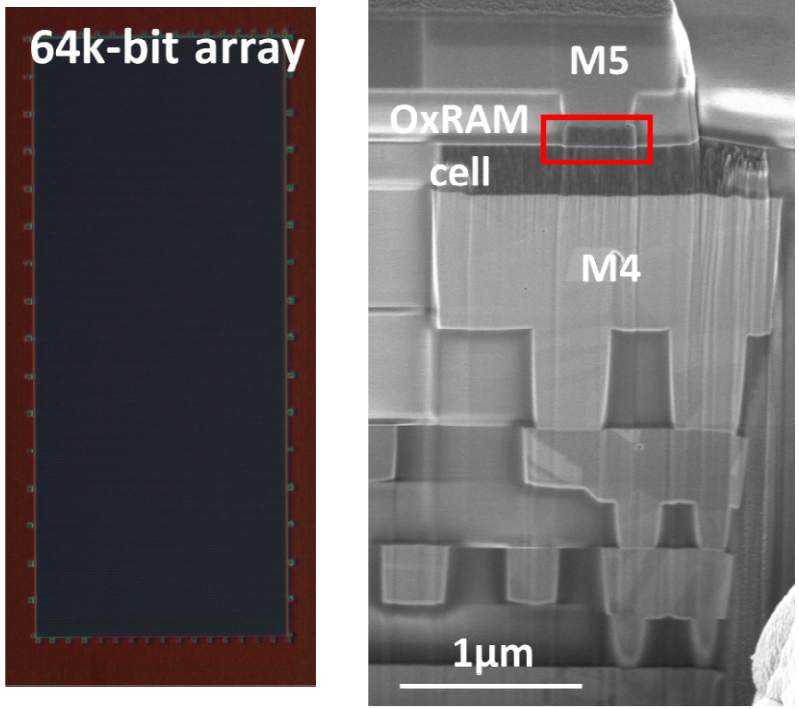
- **Binary OxRAM cells in parallel → (N+1) levels**
- **Switching probability → gradual tuning**

Read more: D. Garbin et al., T-ED, 2015

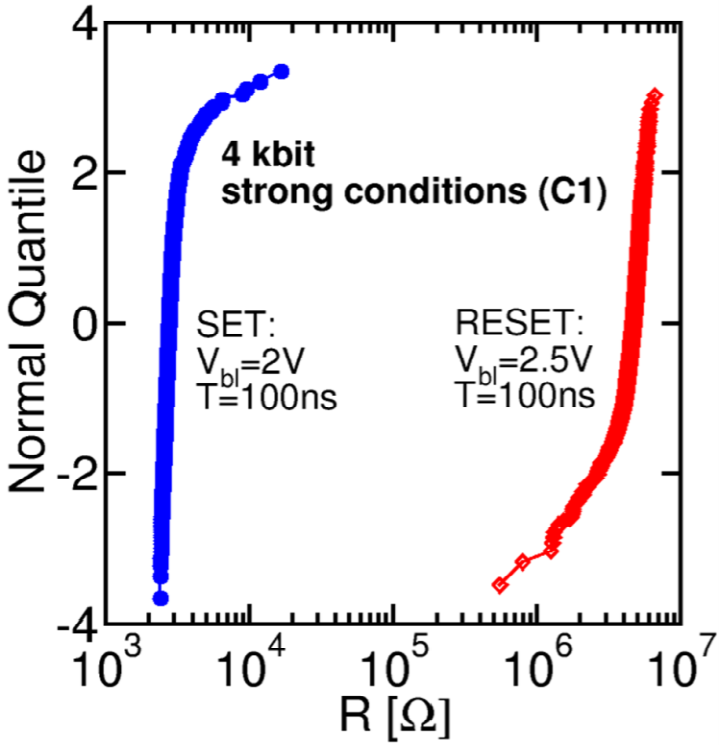
OxRAM test vehicle

- 64kbit HfO_2 OxRAM array integration

SEM

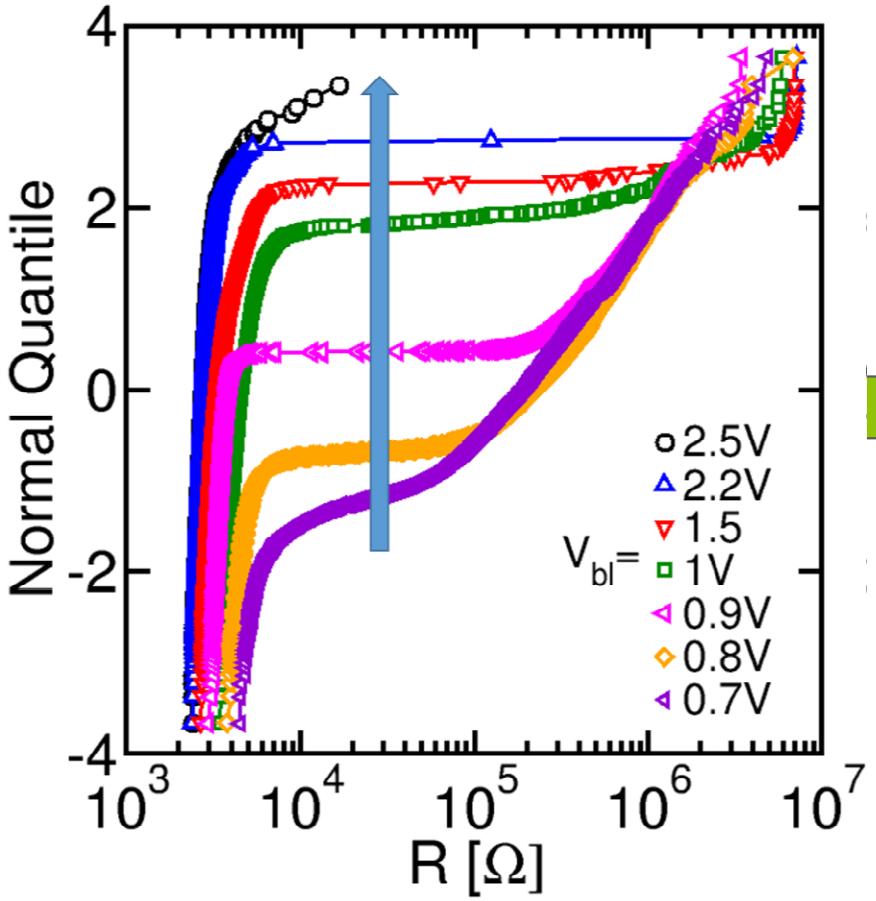


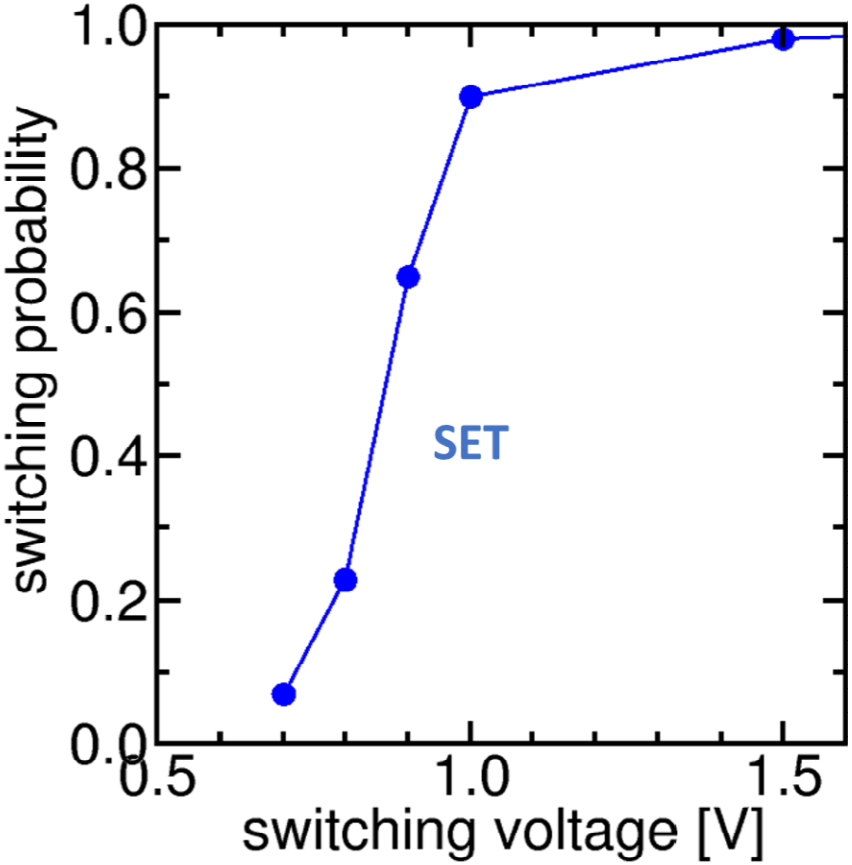
CC~100uA



→ Electrical data from Leti MAD wafers: A. Grossi et al., IEDM 2016

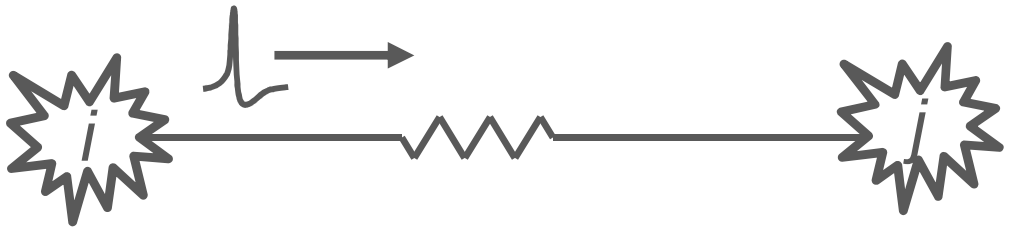
Set: HRS→LRS



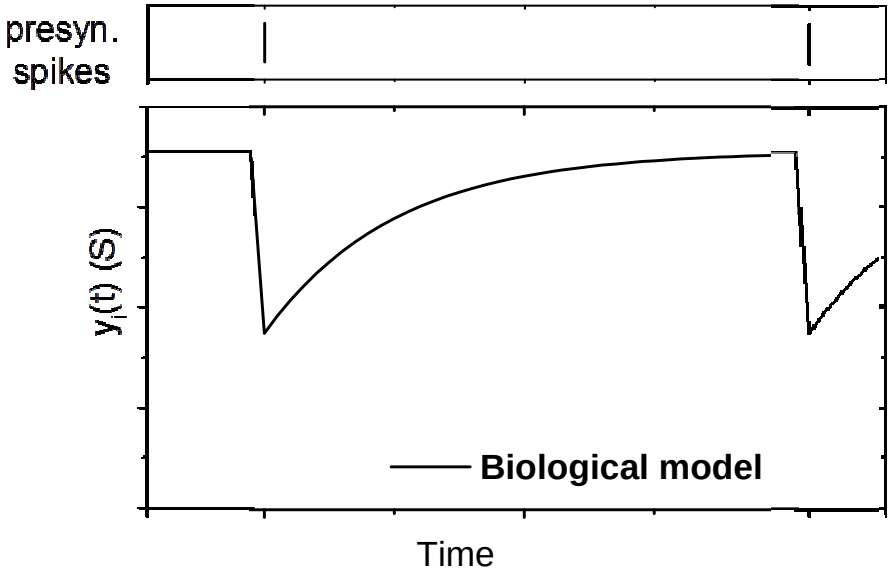


Set probability ~ V_{Set}

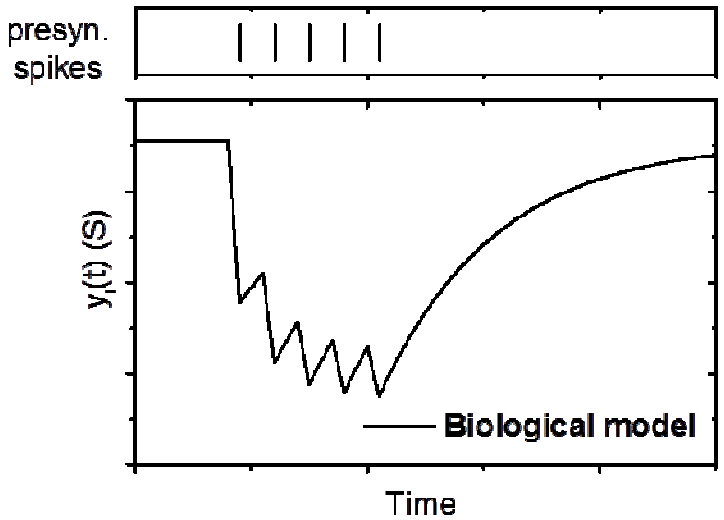
Short Term Plasticity (STP)



Low spiking frequency



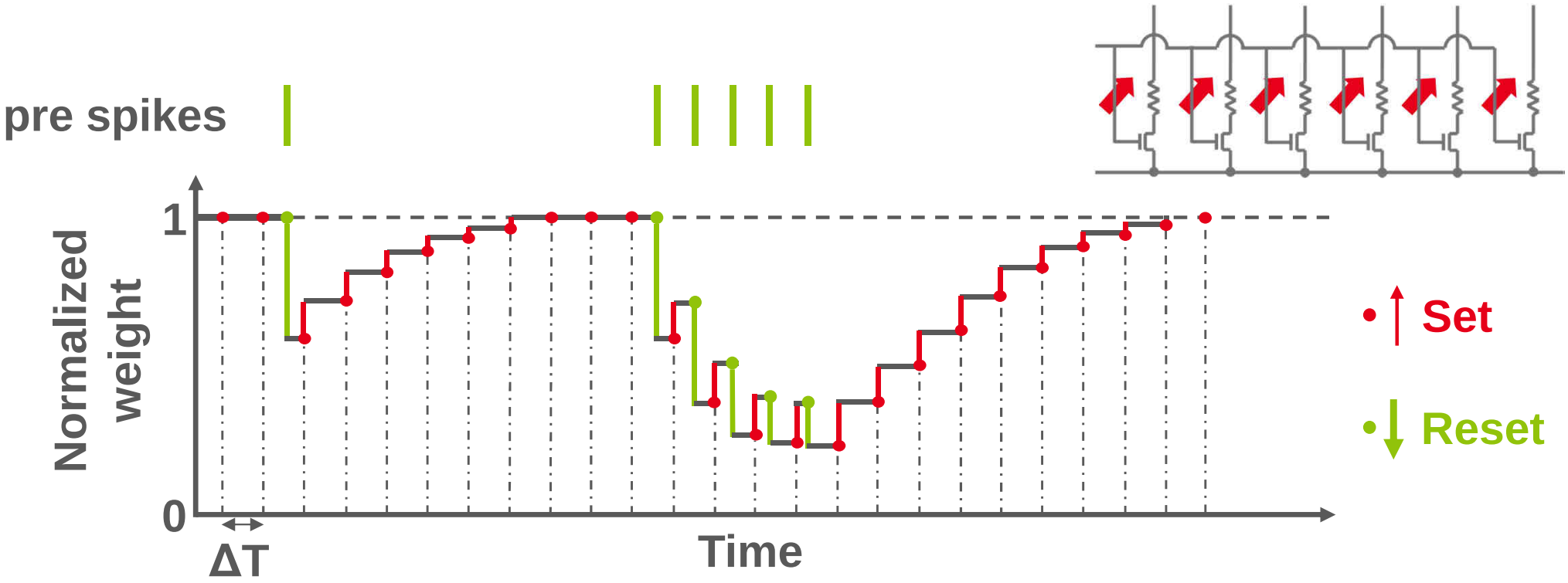
High spiking frequency



STP depends only on presynaptic activity and is volatile

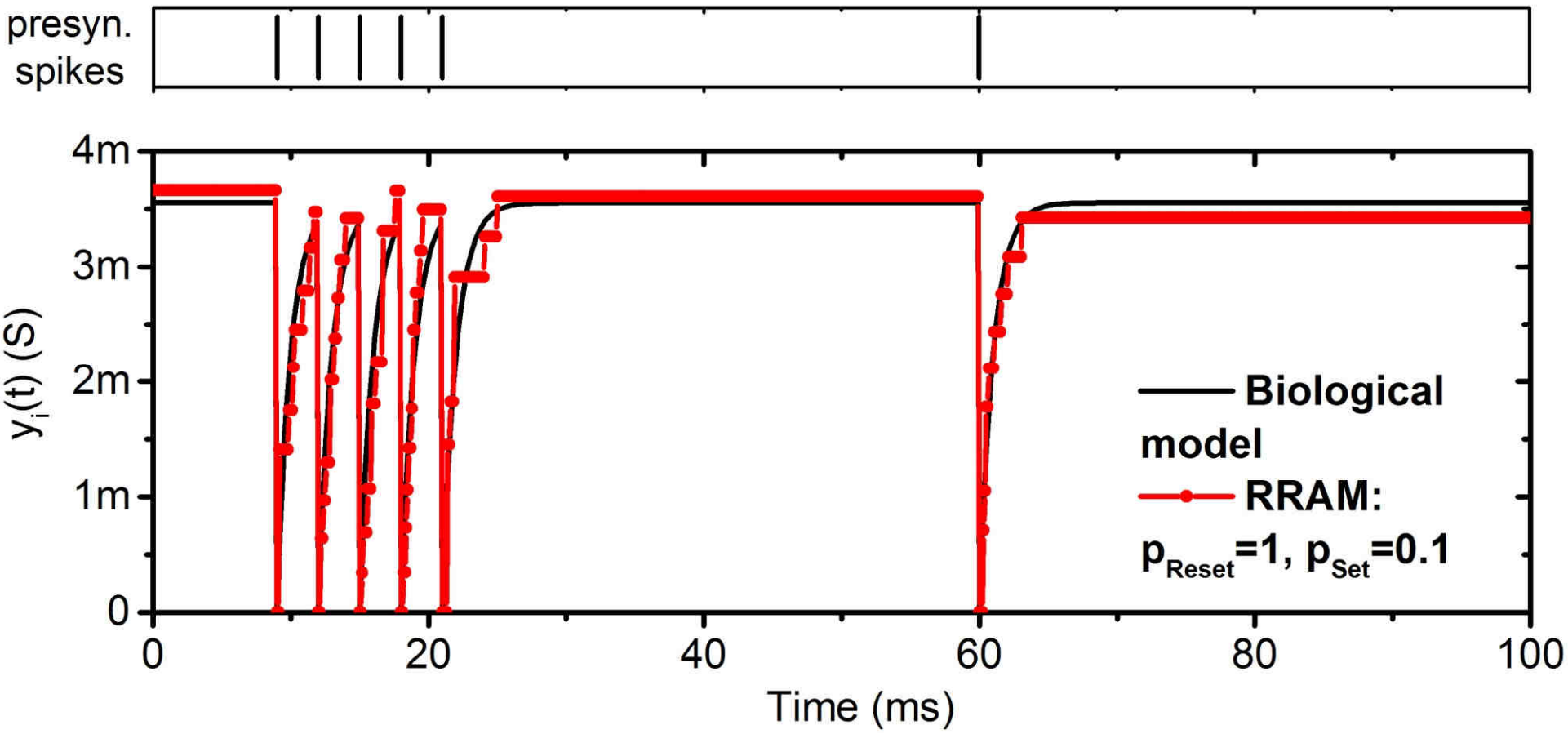
→ Impact for high spiking frequencies

STP programming scheme



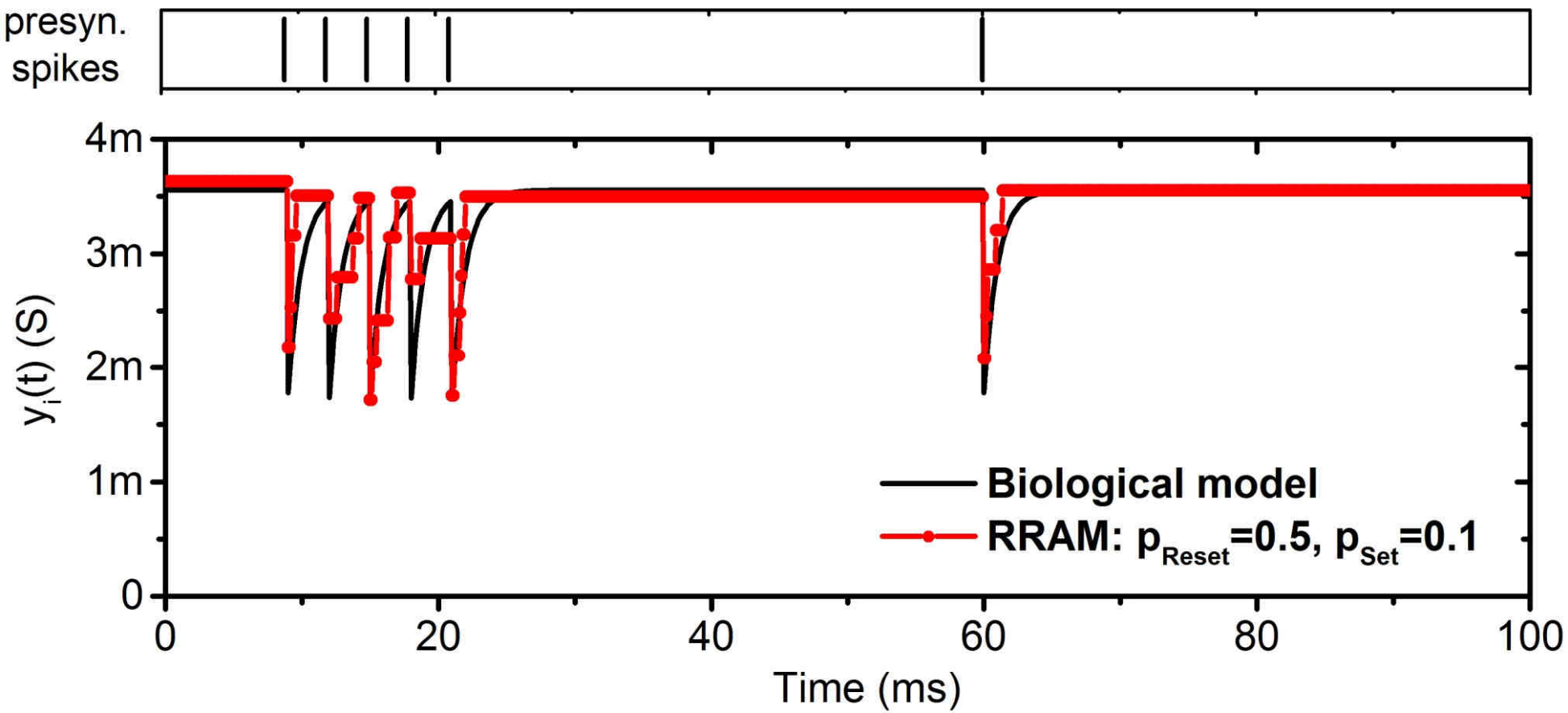
- Set pulse every ΔT
- Reset pulse every pre-synaptic spike

STP implemented in OxRAM



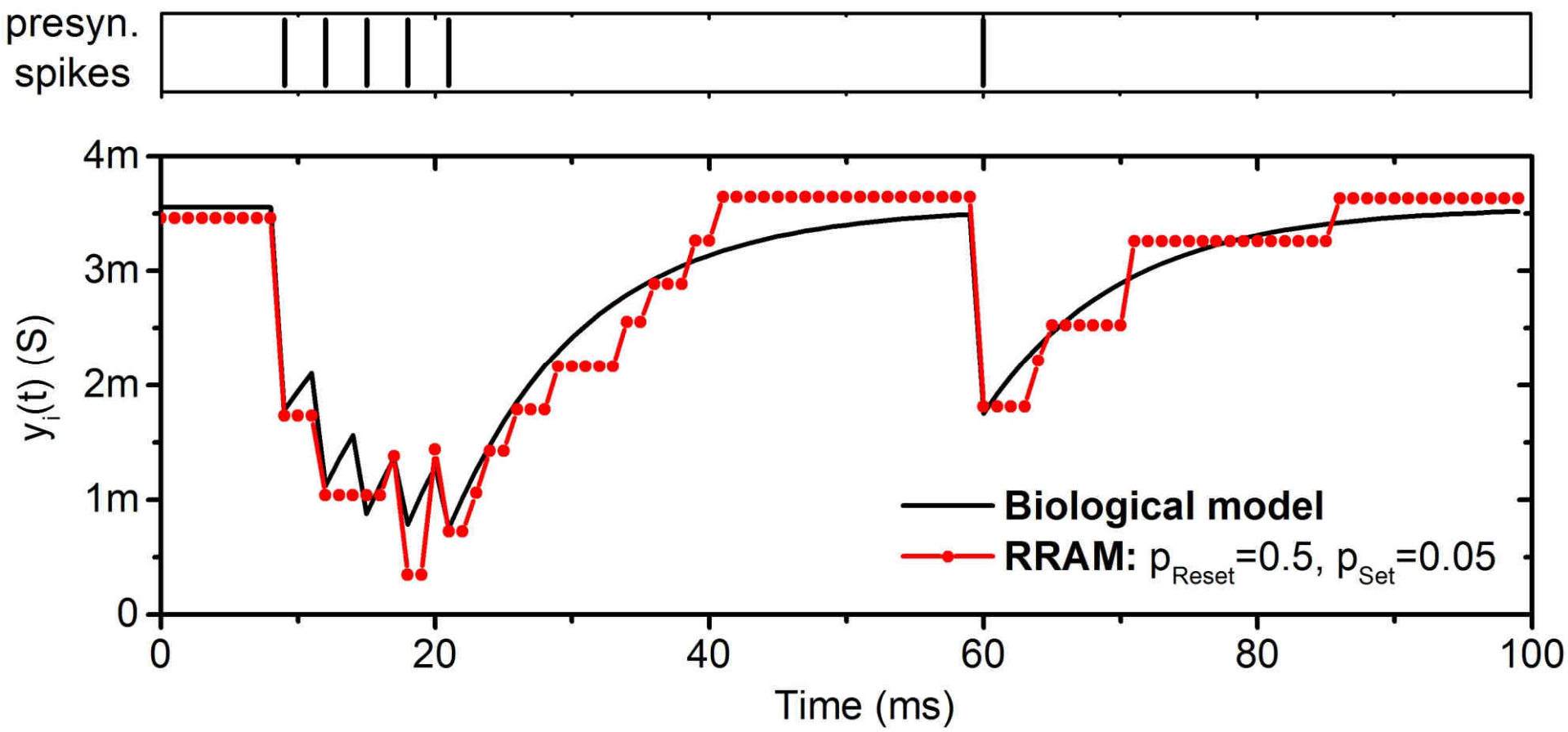
Good fit btw data and 10 OxRAM device/syn

Impact of p_{Reset}



p_{Reset} controls strength of STP depression

Impact of p_{Set}

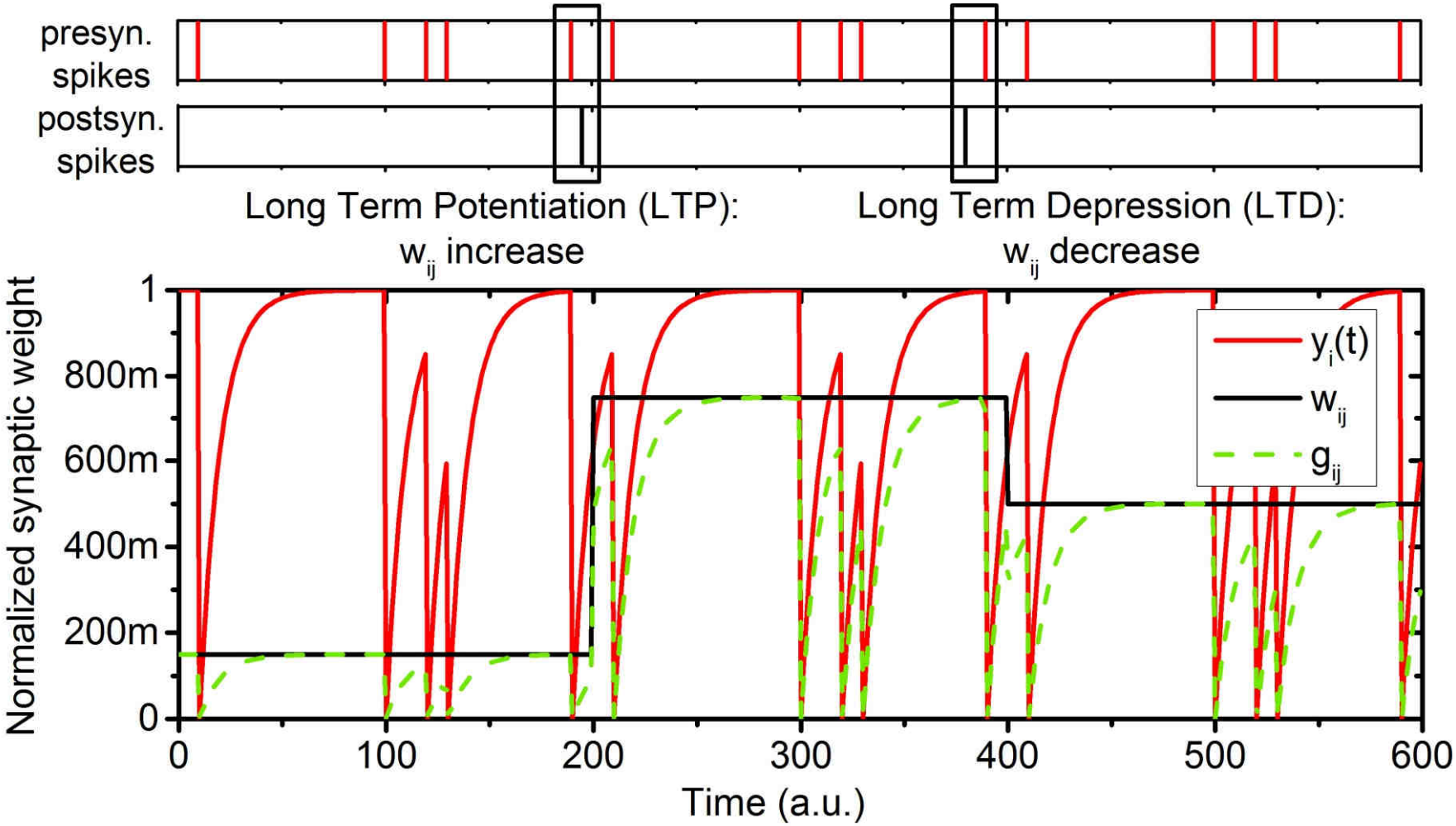


p_{Set} controls relaxation time

Outline

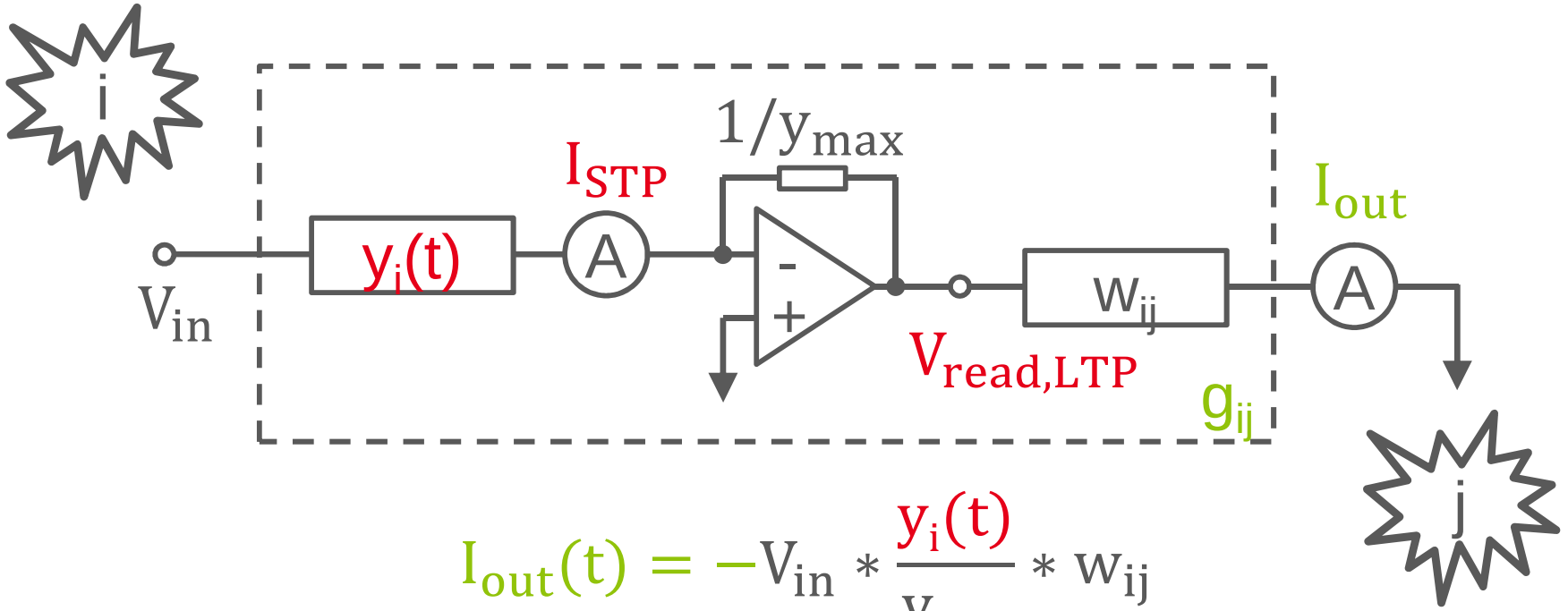
- Introduction
- OxRAM synapses and STP
- ➢ **Co-implementation of STP & LTP**
- Spiking Neural Networks
- Summary

Connecting STP with LTP



$$g_{ij} \sim y_i(t) \times w_{ij}$$

Circuit for co-implementation



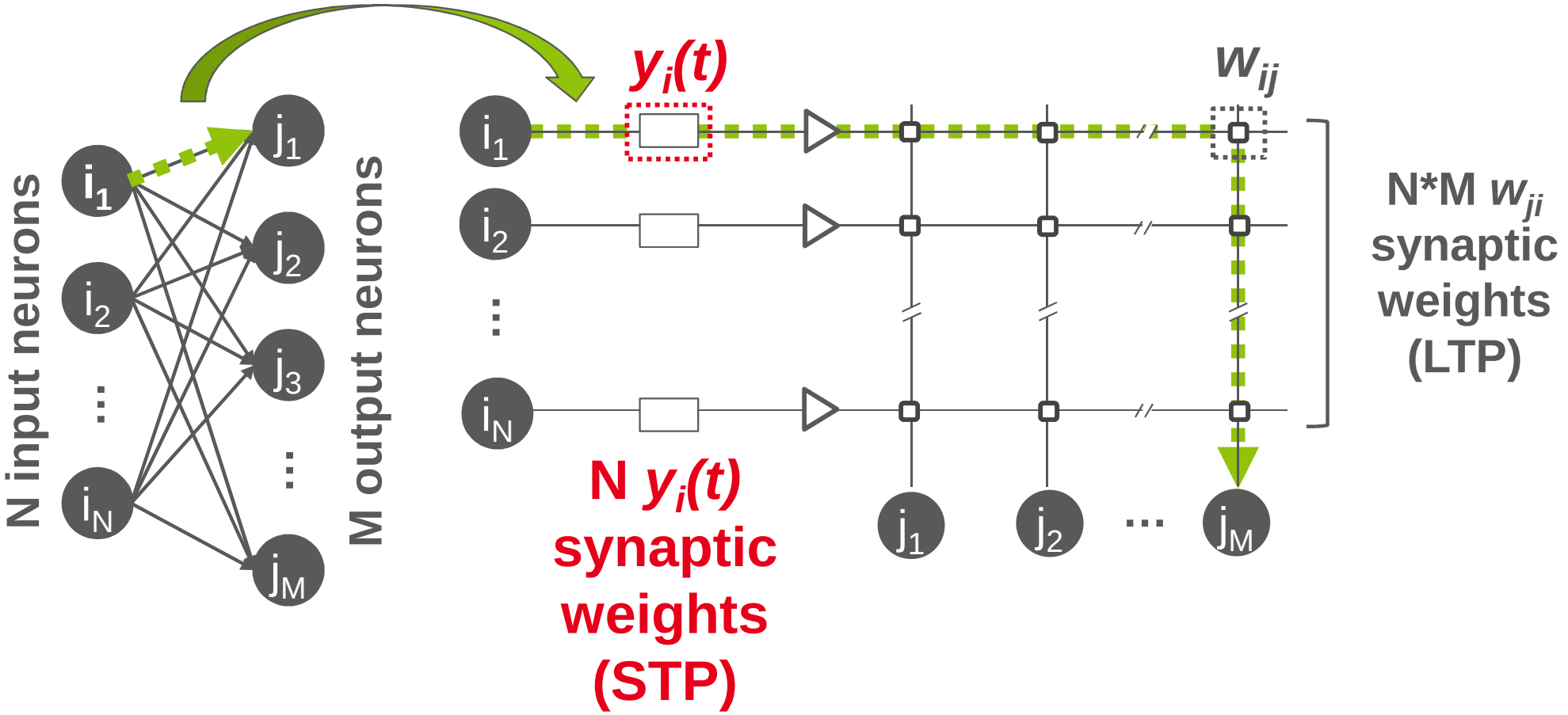
$$I_{out}(t) = -V_{in} * \frac{y_i(t)}{y_{max}} * w_{ij}$$

If $y_i(t) = y_{max} \rightarrow V_{read,LTP}(t) = V_{in}$ (no STP impact)

If $y_i(t) = y_{min} \rightarrow V_{read,LTP}(t) = V_{in} * \frac{y_{min}}{y_{max}}$ (highest STP impact)

Read current of $y_i(t)$ modulates I_{out}

Array implementation



LTP: NxM weights

STP: N weights

Outline

- ✓ Introduction to Synaptic Plasticity
- ✓ OxRAM synapses
- ✓ Co-implementation of STP & LTP
- ➢ **STP impact on Spiking Neural Networks**
- Summary

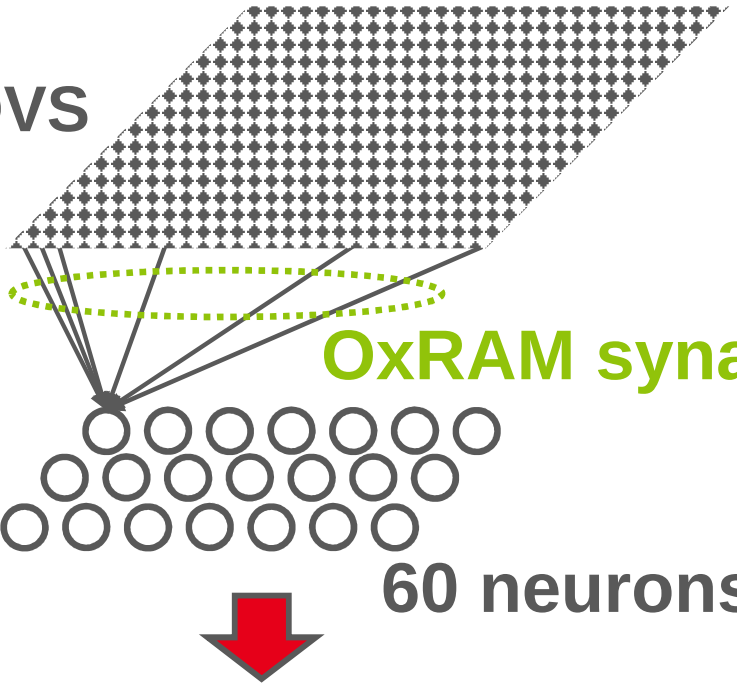
Visual pattern extraction

O. Bichler, T-ED, 2012

Cars on freeway

128x128 px

Retina inspired DVS

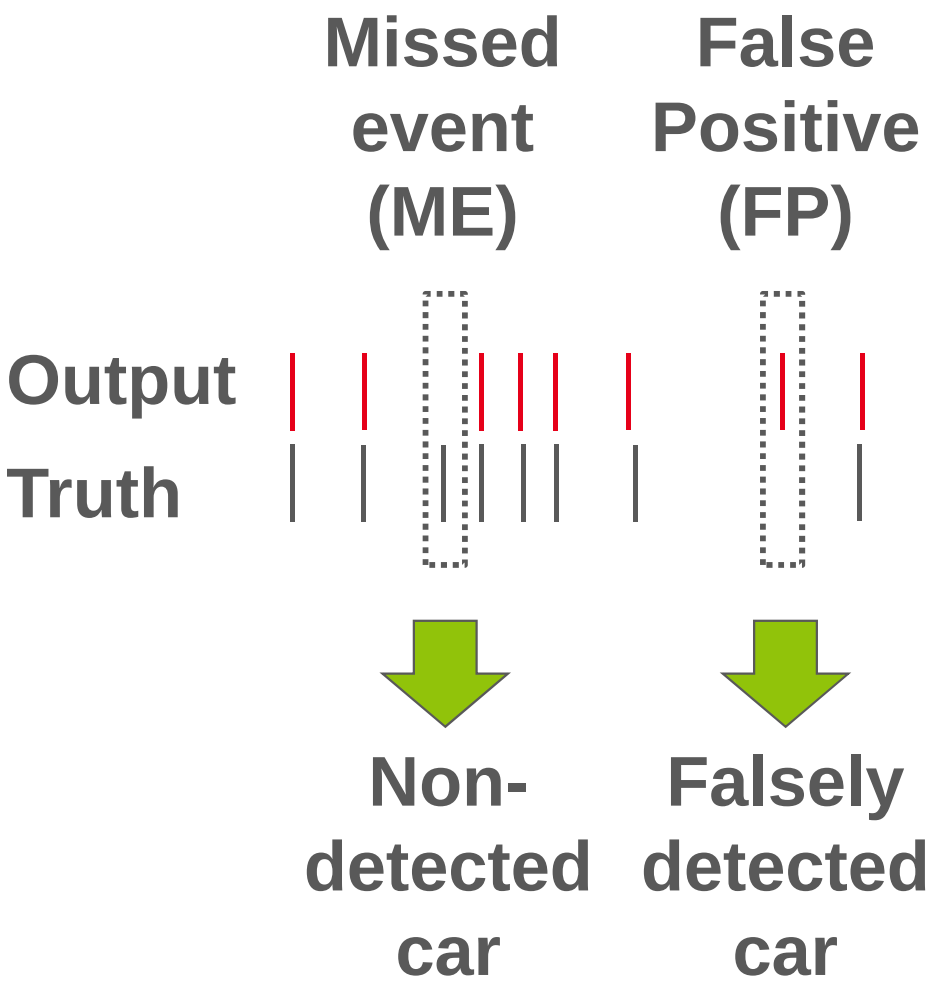


Online learning

- unsupervised
- Competitive → winner-takes-all

Neurons become selective to lanes

Quantification



Detection Rate (DR)

$$DR = \frac{truth - ME}{truth}$$

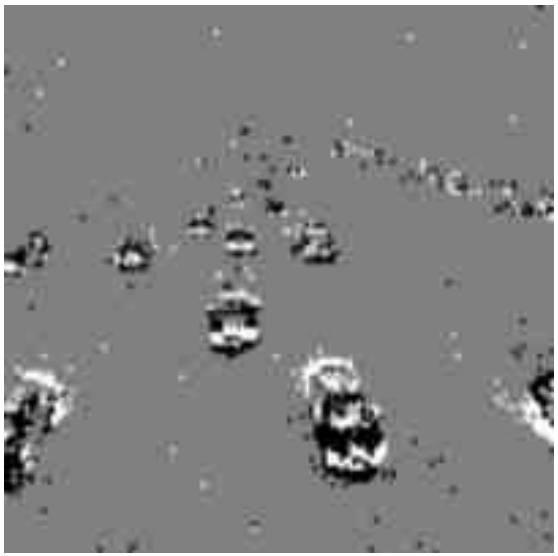
False Positive Rate (FPR)

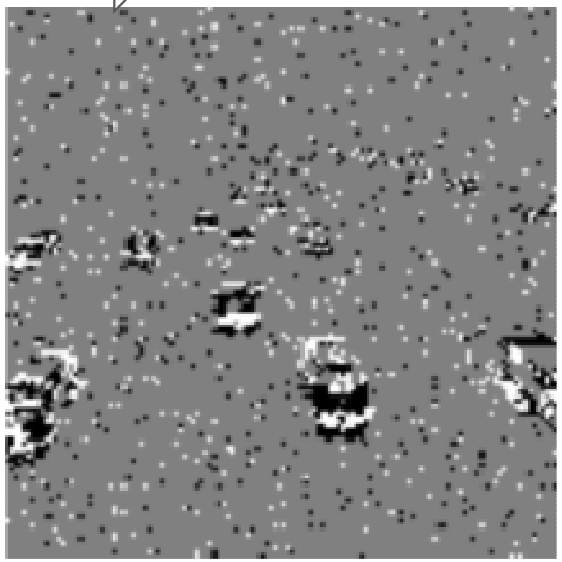
$$FPR = \frac{FP}{truth}$$

Noise

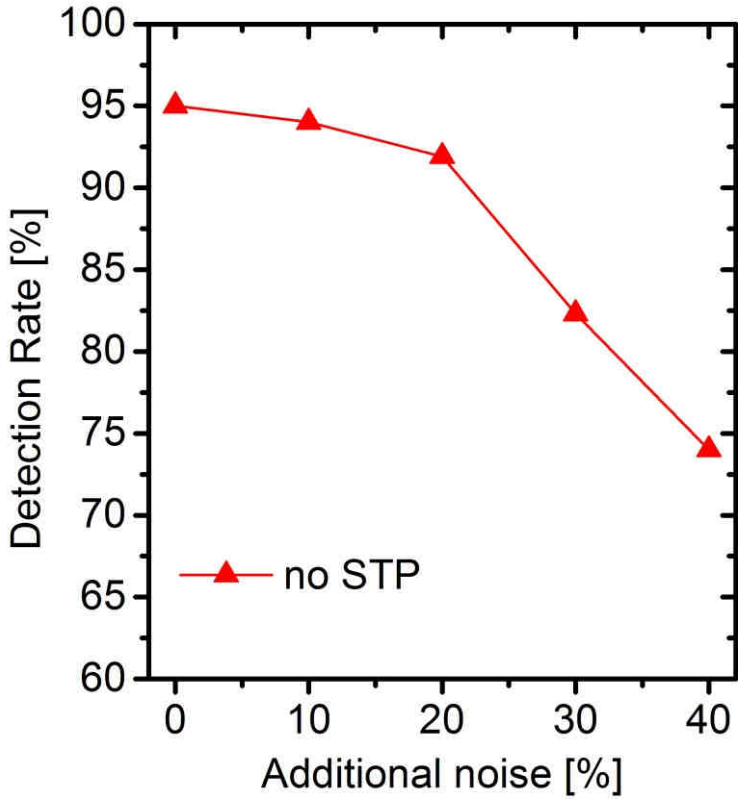
- **Additional noise introduced by random spiking activity of retina**

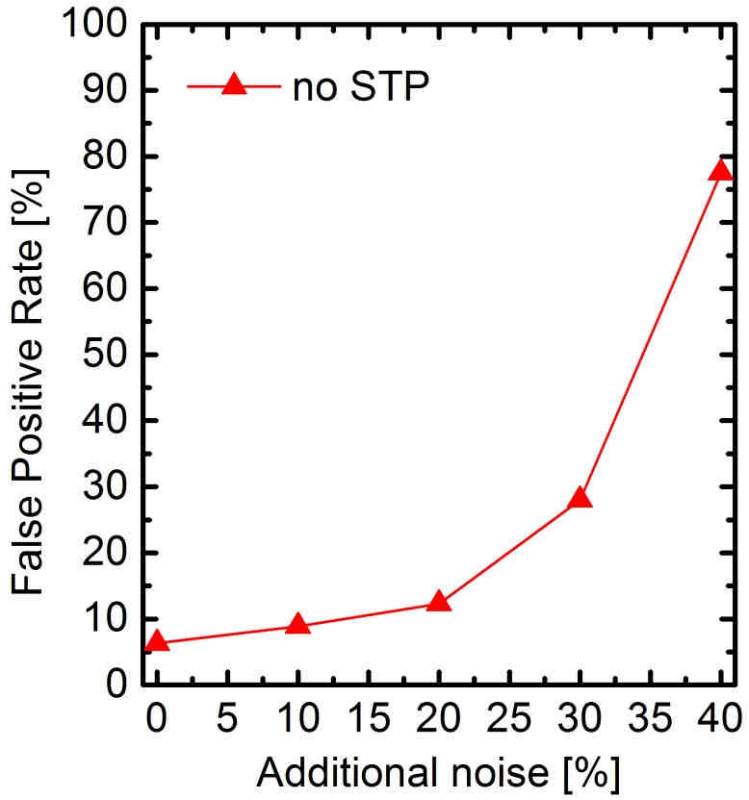
+20 %Noise





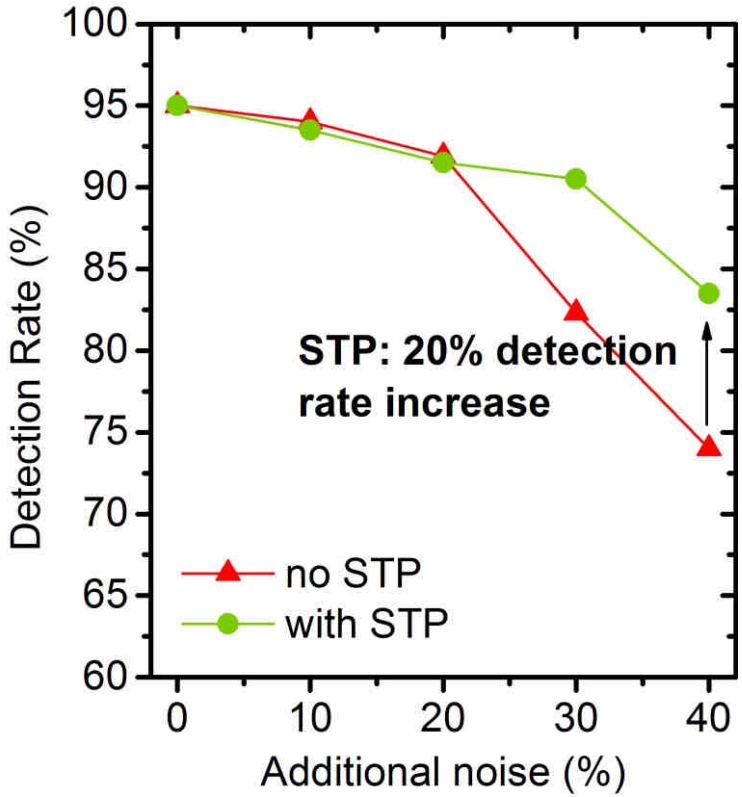
Noise impact

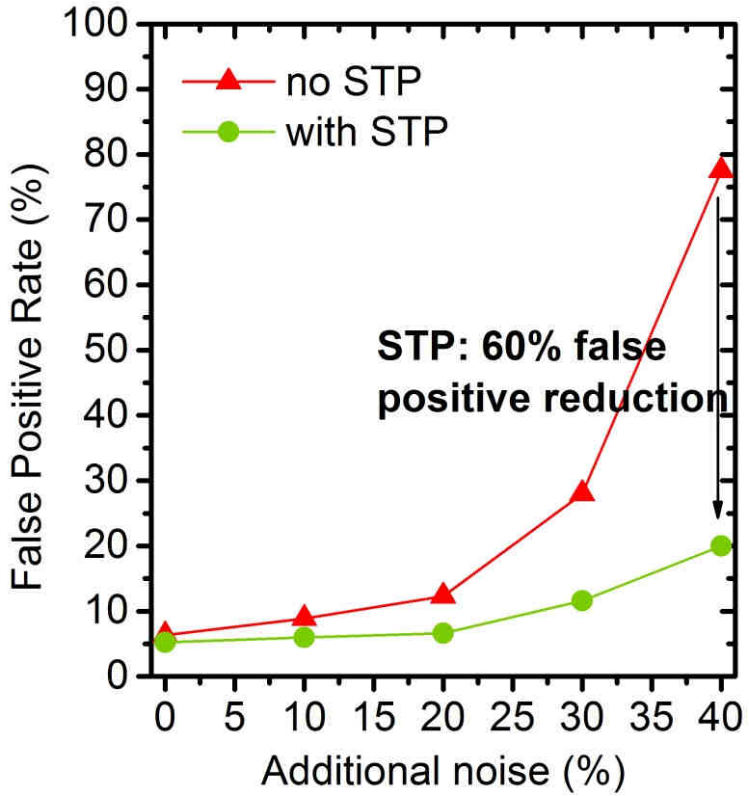




Noisy input signal tends to decrease DR and increase FP

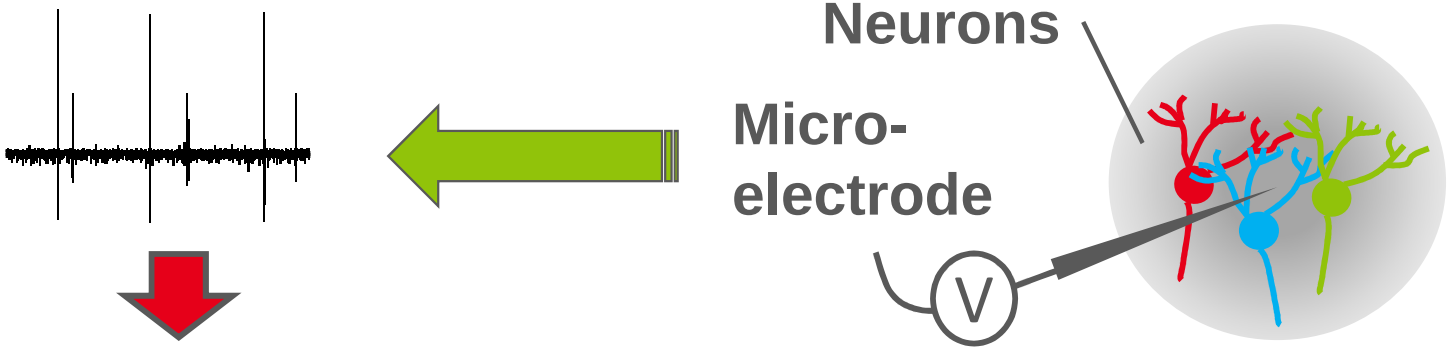
STP effect





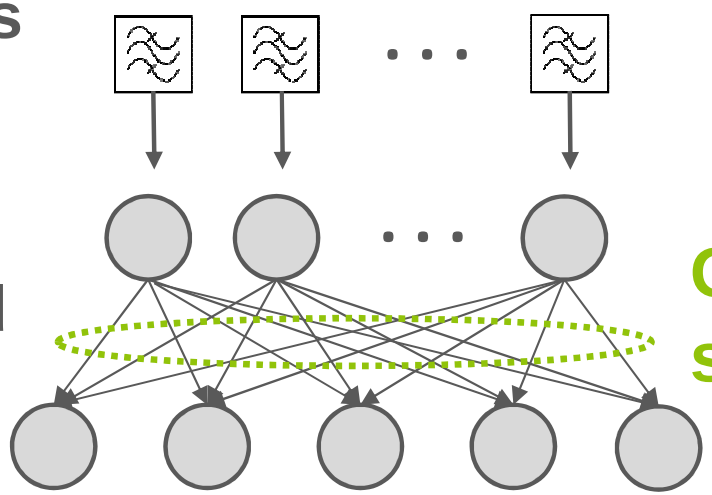
Thanks to STP: DR and FPR strongly improve in case of highly noisy data

Decoding of neural activity



Band-pass filters

Fully Connected Neural Network



OxRAM synapses

Online learning

- unsupervised
- Competitive → winner-takes-all

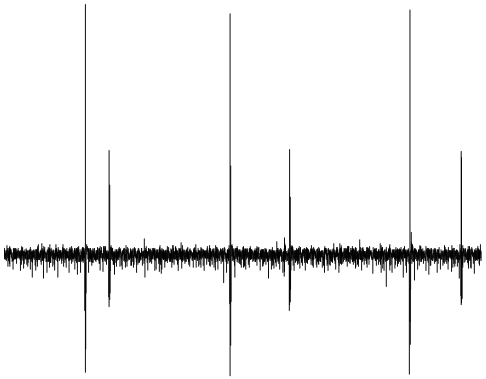
Spike detection

Noisy input data

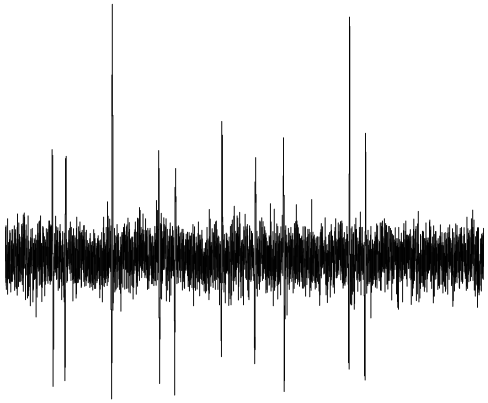
→ **Signal-Noise-Ratio of artificial data varied**

Noise level

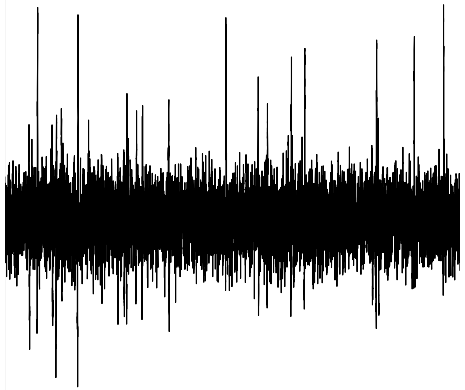
SNR=80



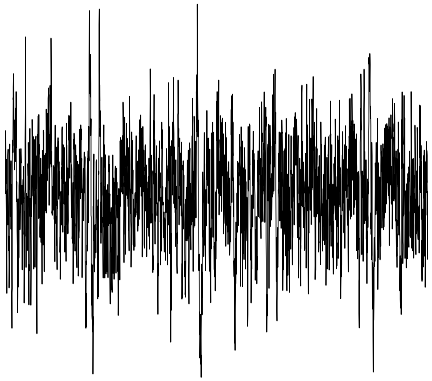
SNR=27



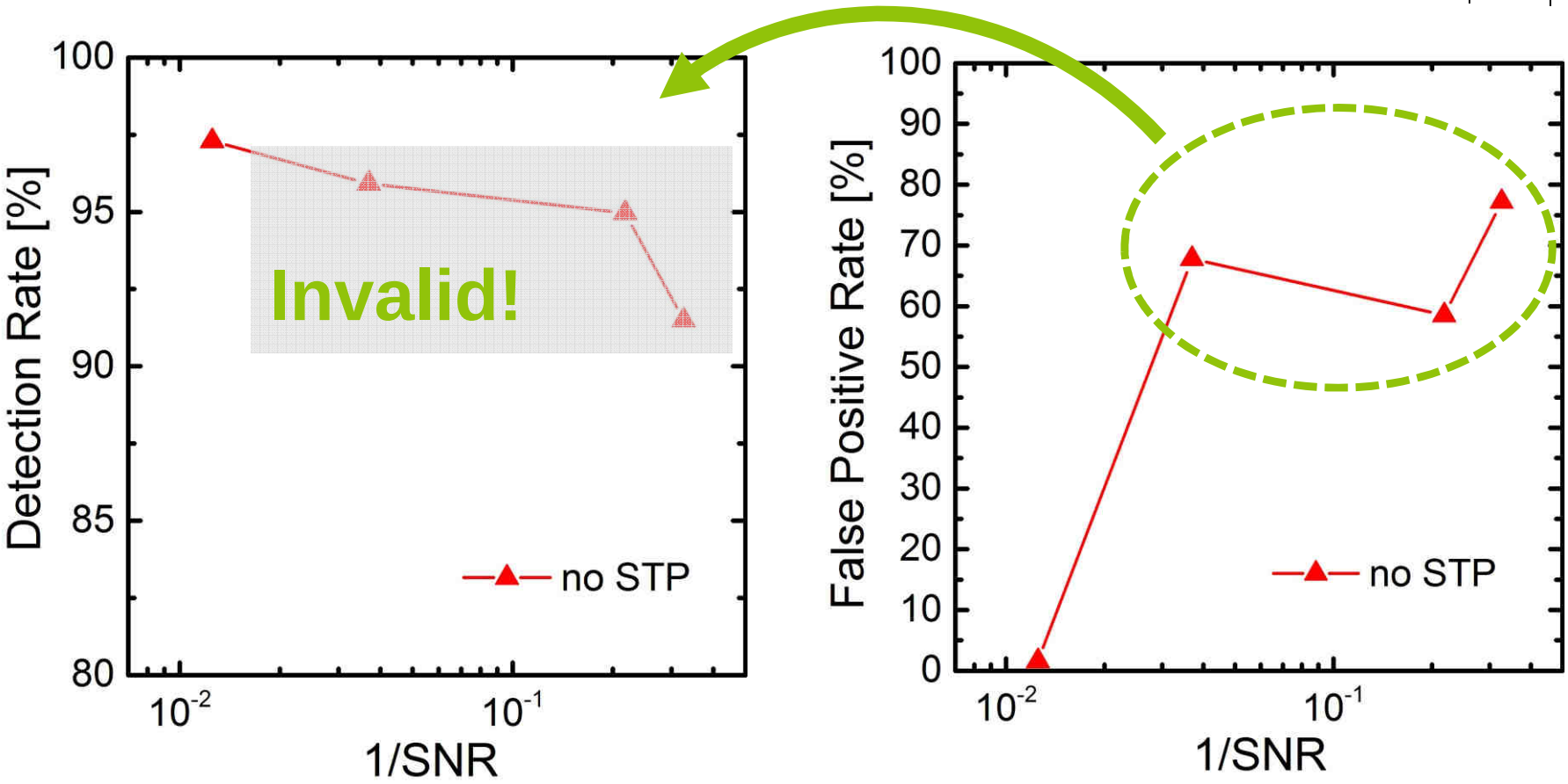
SNR=4.6



SNR=3

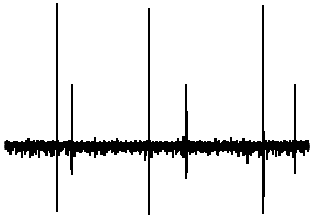


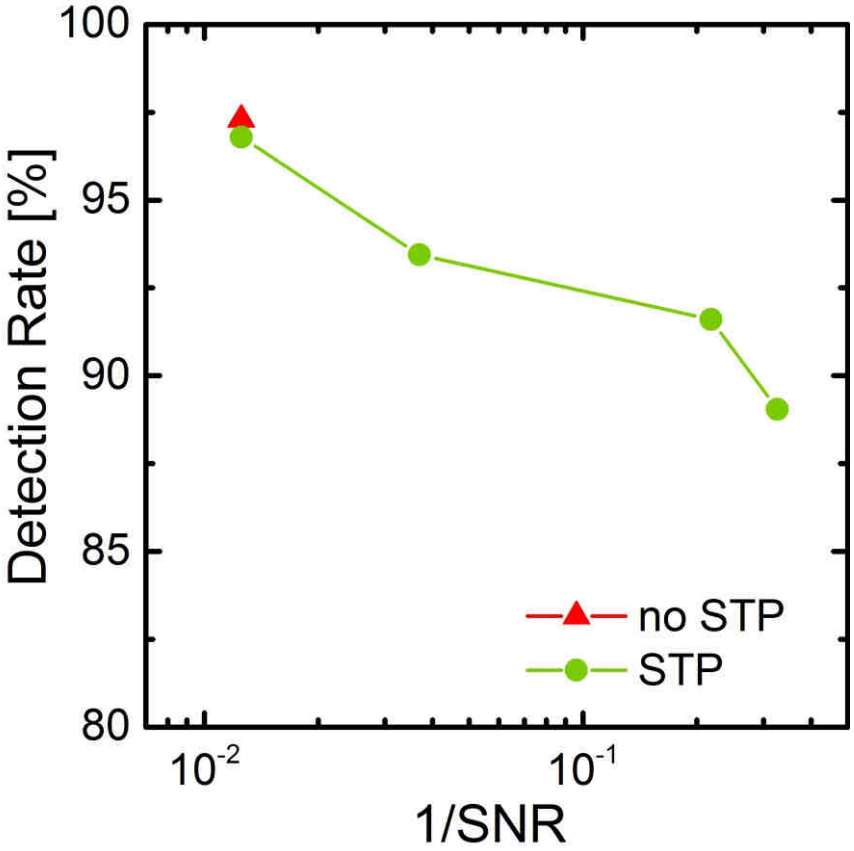
Noise impact

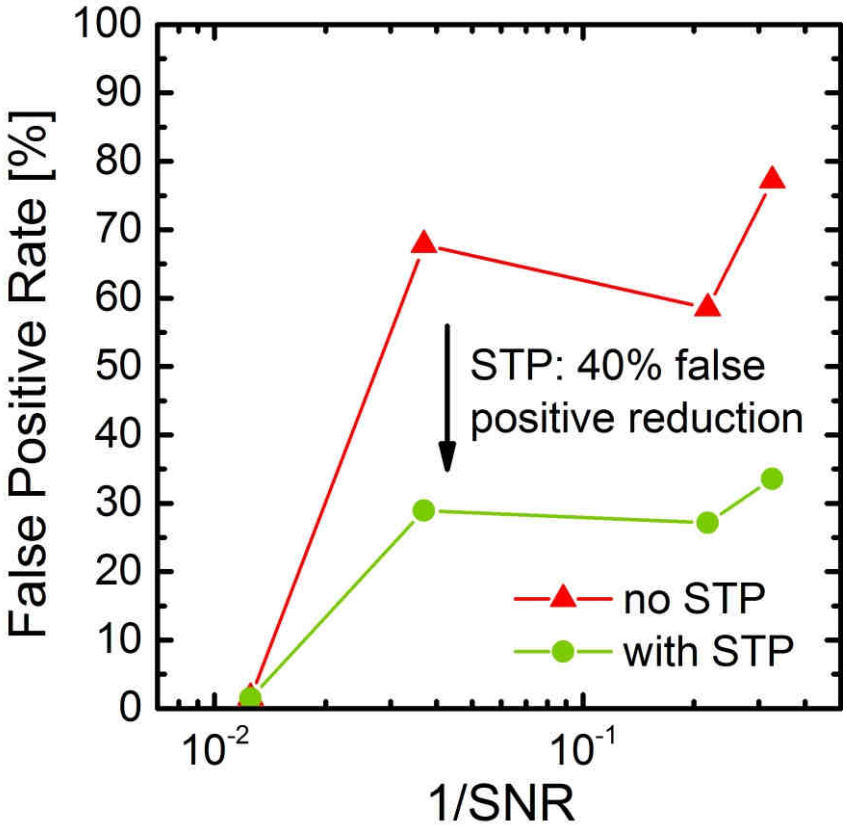


Very high FPR due to noise!

STP effect







STP enables spike detection in noisy signals

Outline

- ✓ Introduction to Synaptic Plasticity
- ✓ OxRAM synapses
- ✓ Co-implementation of STP & LTP
- ✓ Spiking Neural Networks
- ➢ **Summary**

Summary

- ✓ **Bio-inspired system with non-volatile OxRAM synapses to reproduce Long and Short Term Plasticity**

- ✓ **Short Term Plasticity achieved with:**
  - ✓ 1 additional synapse per input neuron
  - ✓ 10 binary OxRAM cells per synapse

- ✓ **Adding STP to LTP allows to suppress noise and improves learning → strongly reduced False Positive Rate (50% @SNR~3)**

Thank you for your attention.

thilo_werner@gmx.de

Leti, technology research institute
Commissariat à l'énergie atomique et aux énergies alternatives
Minatec Campus | 17 rue des Martyrs | 38054 Grenoble Cedex | France
www.leti.fr

References

- M. V. Tsodyks et al., Proc. Natl. Acad. Sci. vol.94, 1997.
- G. Q. Bi et al., J. Neuroscience, vol.18, 1998.
- A. Grossi et al., Proc. IEDM, 2016.
- D. Garbin et al., T-ED, 2015