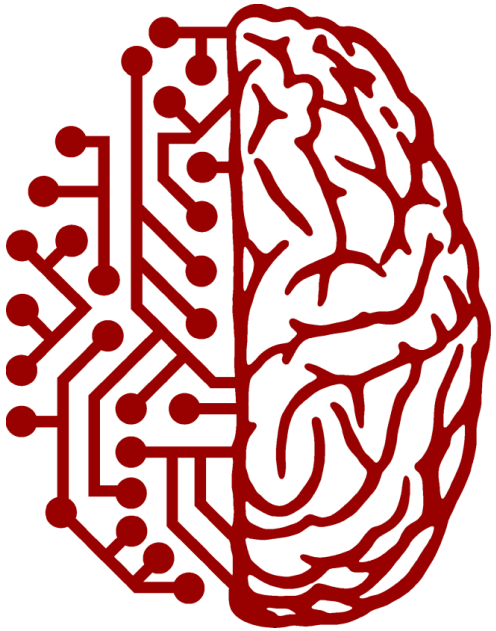
UNIVERSITÄT HEIDELBERG

UNIVERSITÄT BERN

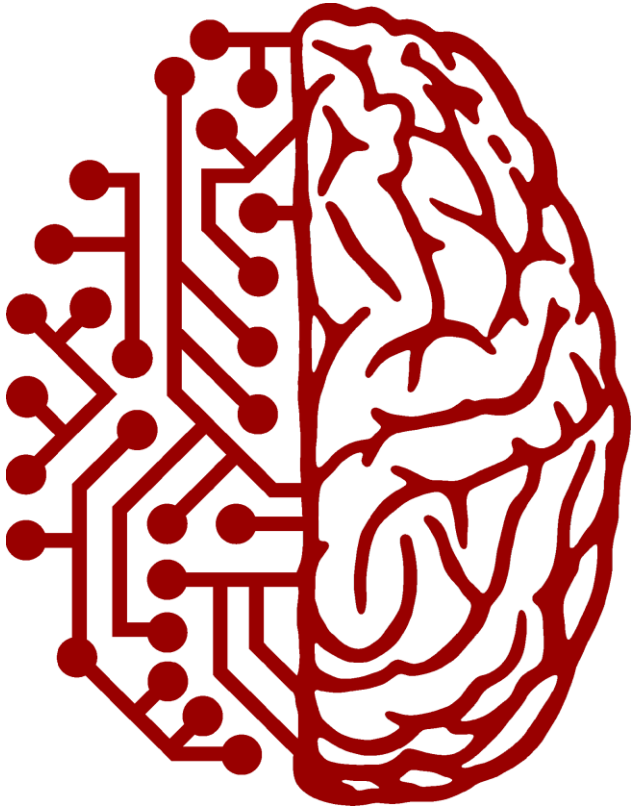


Fast inference with spiking networks

Mihai A. Petrovici

Computantional Neuroscience Group
Department of Physiology
University of Bern

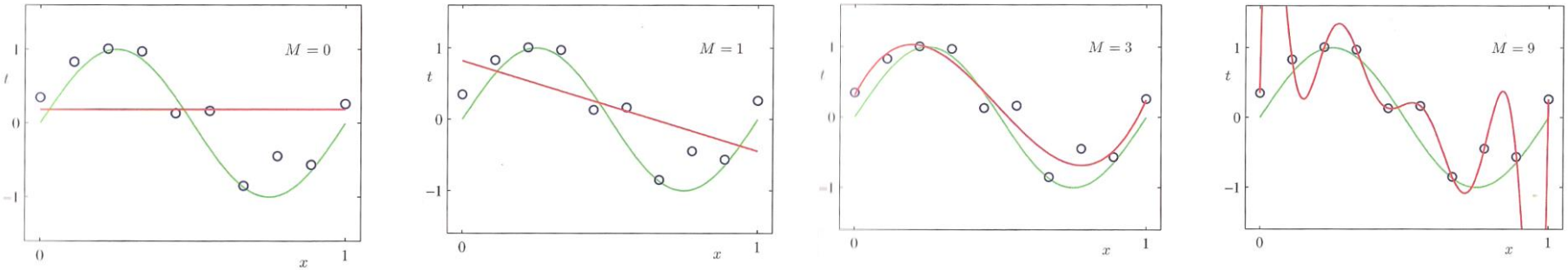
Electronic Vision(s)
Kirchhoff Institute for Physics
University of Heidelberg

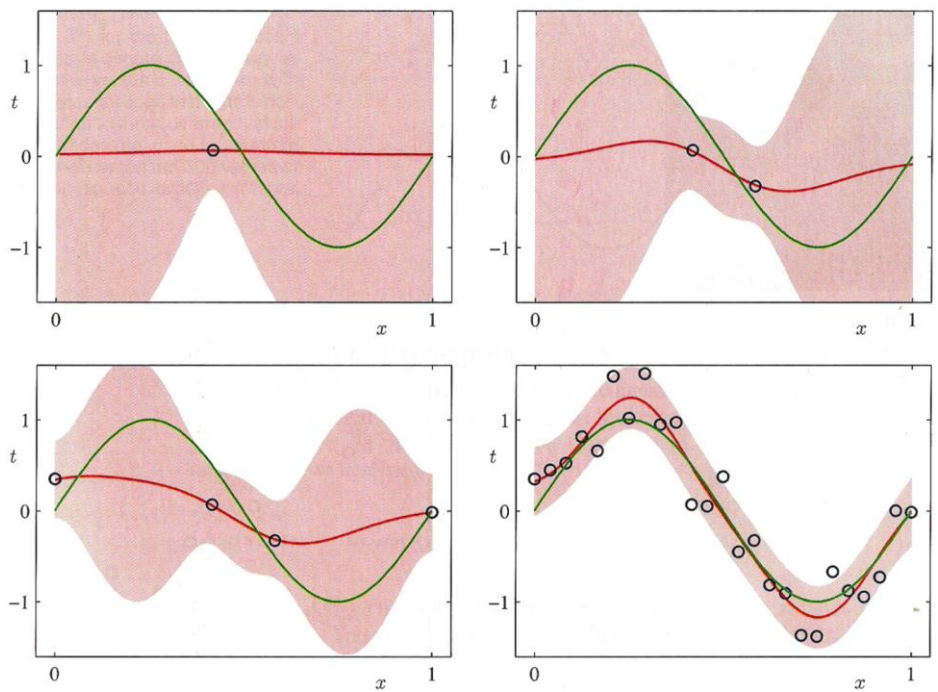


Probabilistic computation with spikes

Probabilistic (Bayesian) computing: motivation

get an estimate of the <u>distribution</u> of the sought function (expected value <u>and</u> uncertainty)

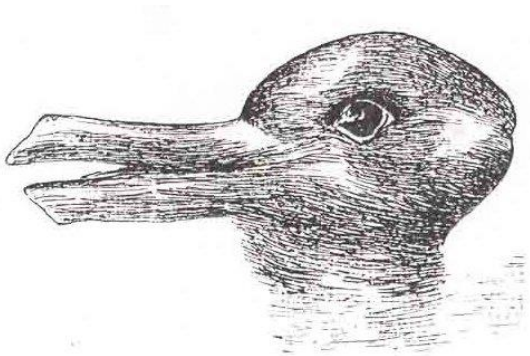


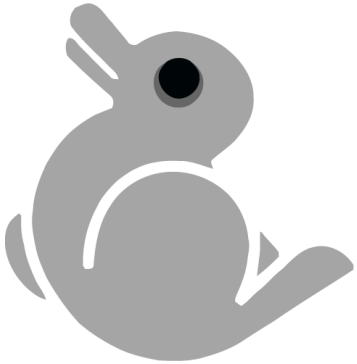


Bishop (2006)

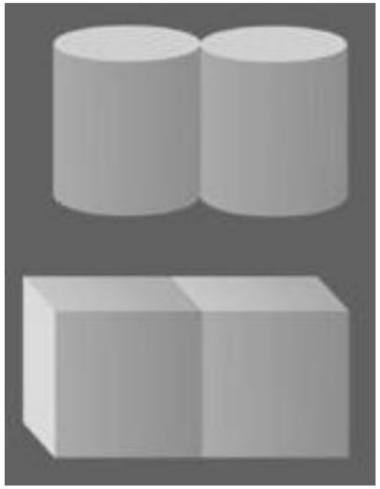
Probabilistic (Bayesian) computing: experimental evidence

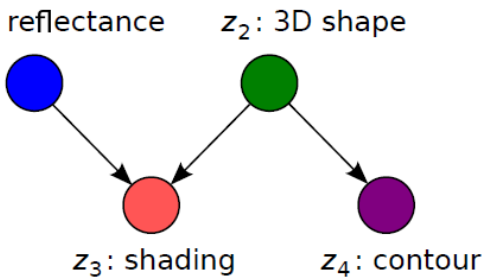
Rabbit/duck ambiguity



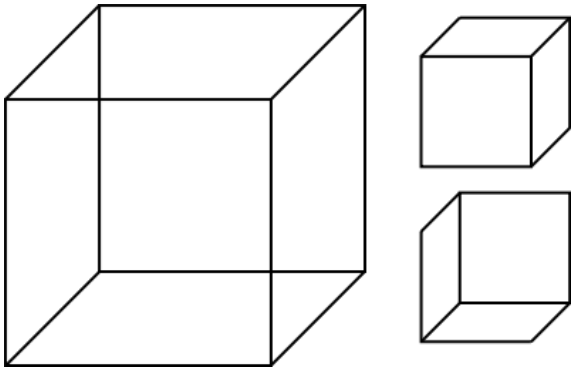


Knill-Kersten illusion



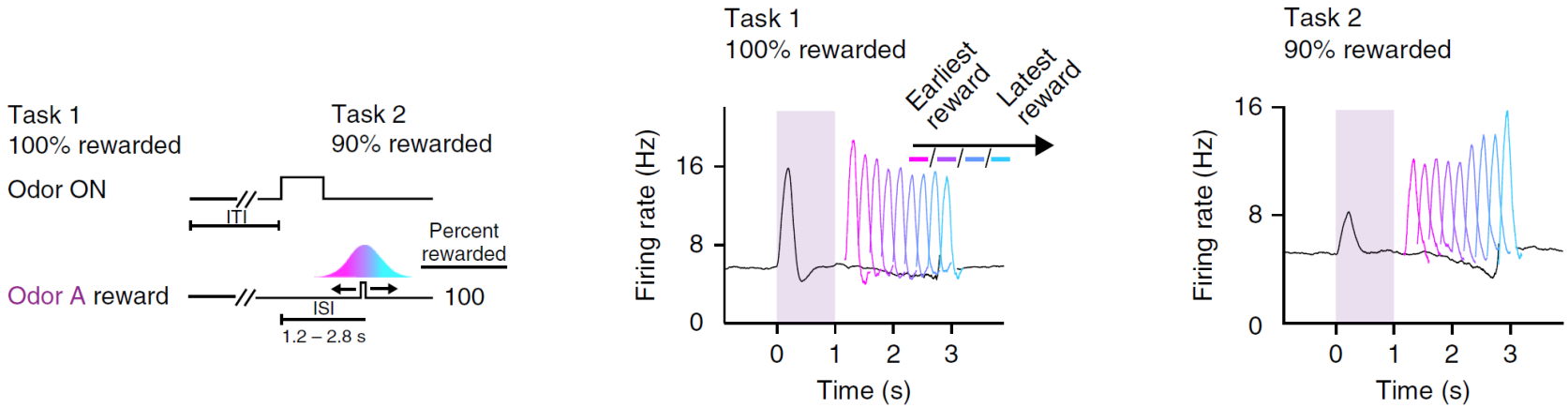


Necker cube

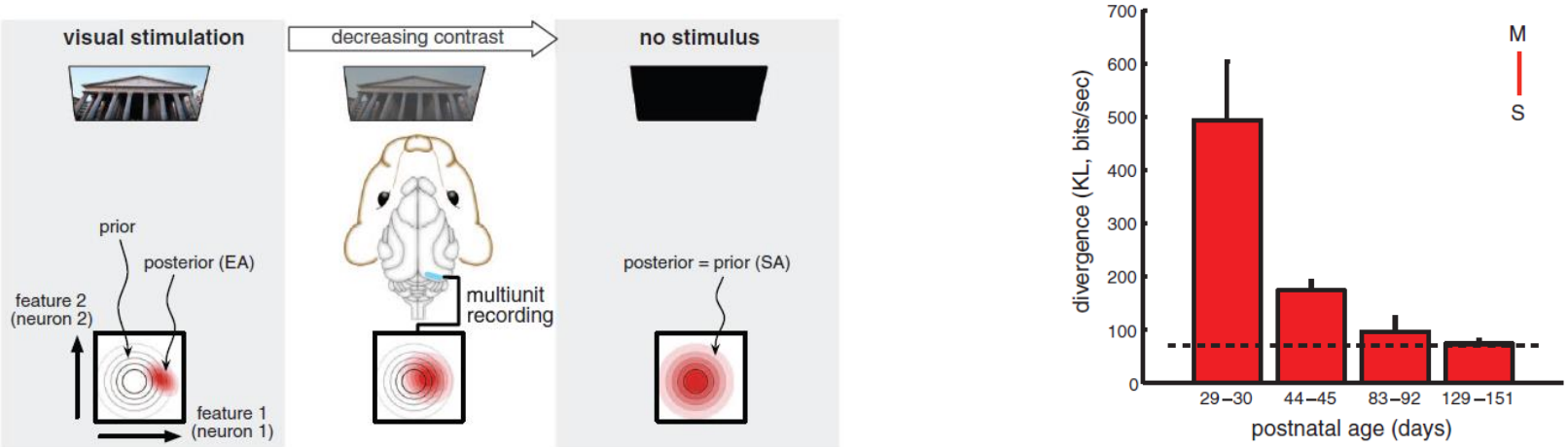


Probabilistic (Bayesian) computing: experimental evidence

Starkweather et al. (2017)



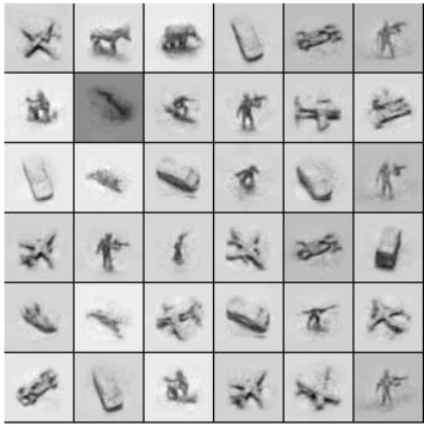
Berkes et al. (2011)



Probabilistic (Bayesian) computing in machine learning

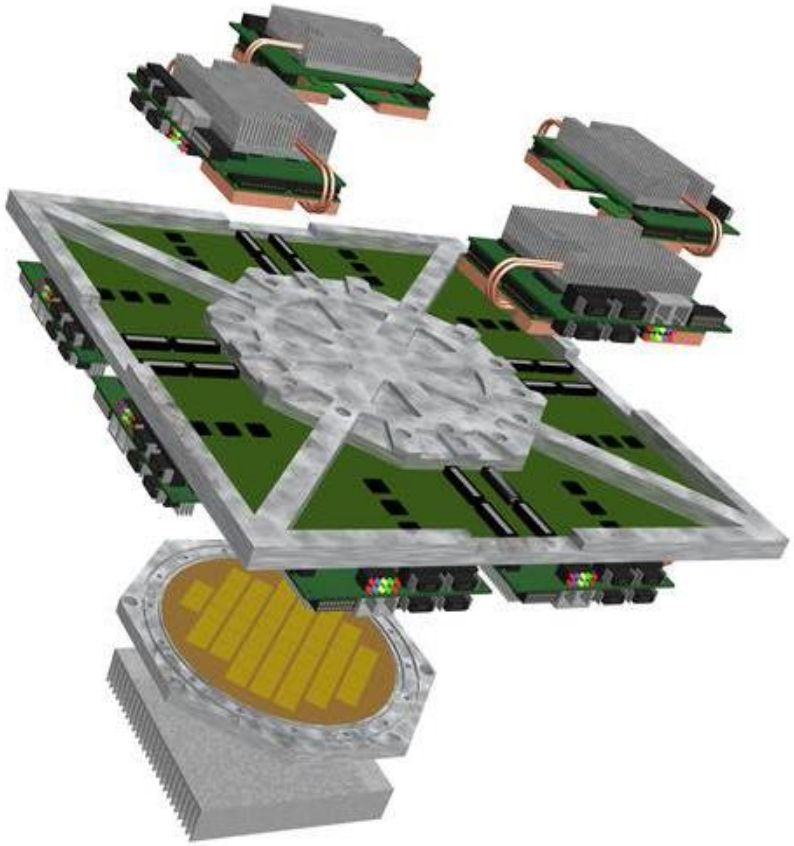
Deep generative models

Salakhutdinov & Hinton (2009)



Neuromorphic hardware

Schemmel et al. (2010)



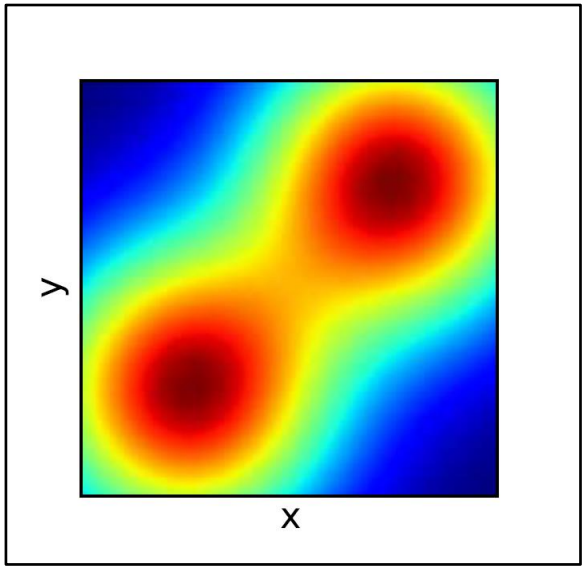
Requirements for probabilistic inference

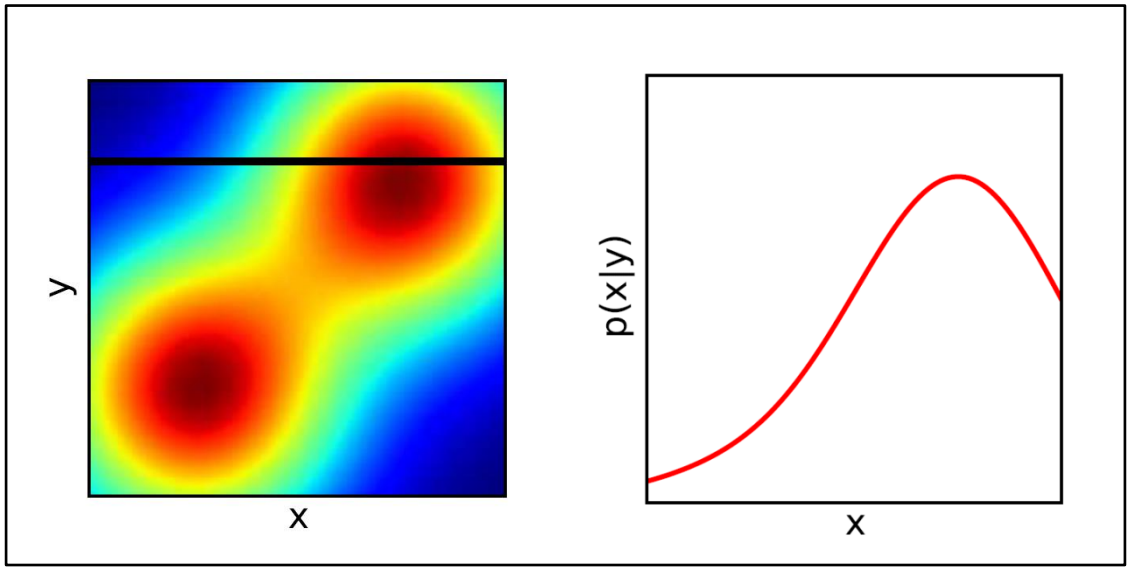
A system that performs probabilistic inference has to

→ represent probability distributions $p(z_1, z_2, \dots)$

→ calculate posterior (conditional) distributions $p(z_1, z_2, \dots | z_k, z_{k+1}, \dots)$

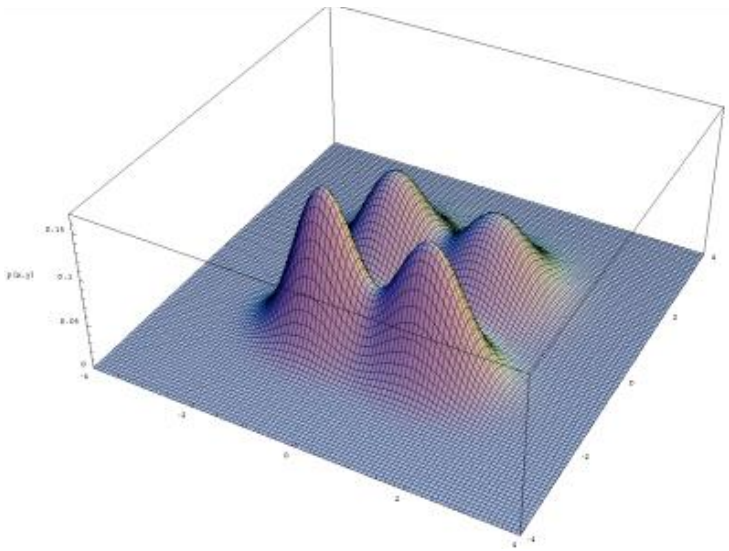
→ evaluate marginal distributions $p(z_1, z_2) = \sum_{z_3, z_4, \dots} p(z_1, z_2, z_3, \dots)$





Representation of probability distributions

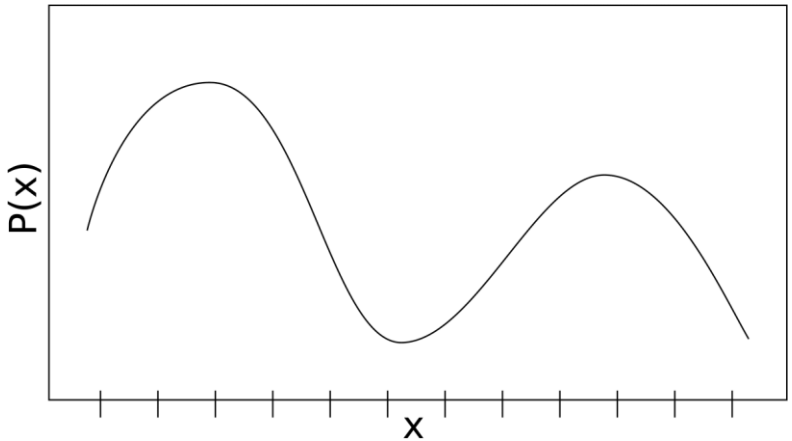
full state space



analytic / parametric

$$p(\boldsymbol{z}) = \frac{1}{Z}\exp\left[\frac{1}{2}\boldsymbol{z}^T\boldsymbol{W}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{b}\right]$$

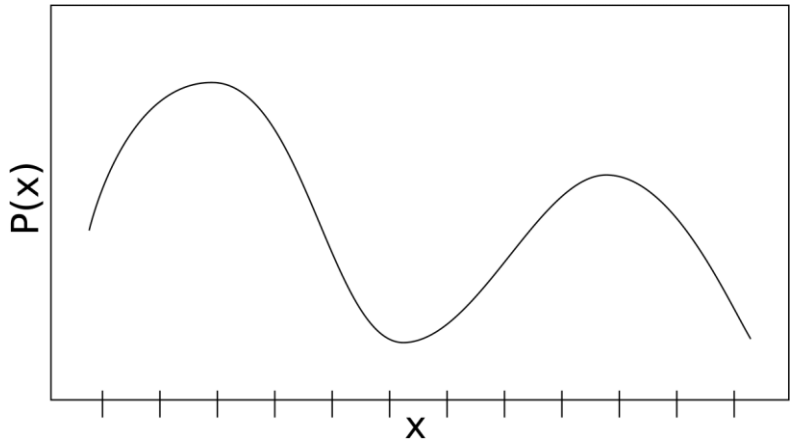
sampling



Sampling vs. parametric representation

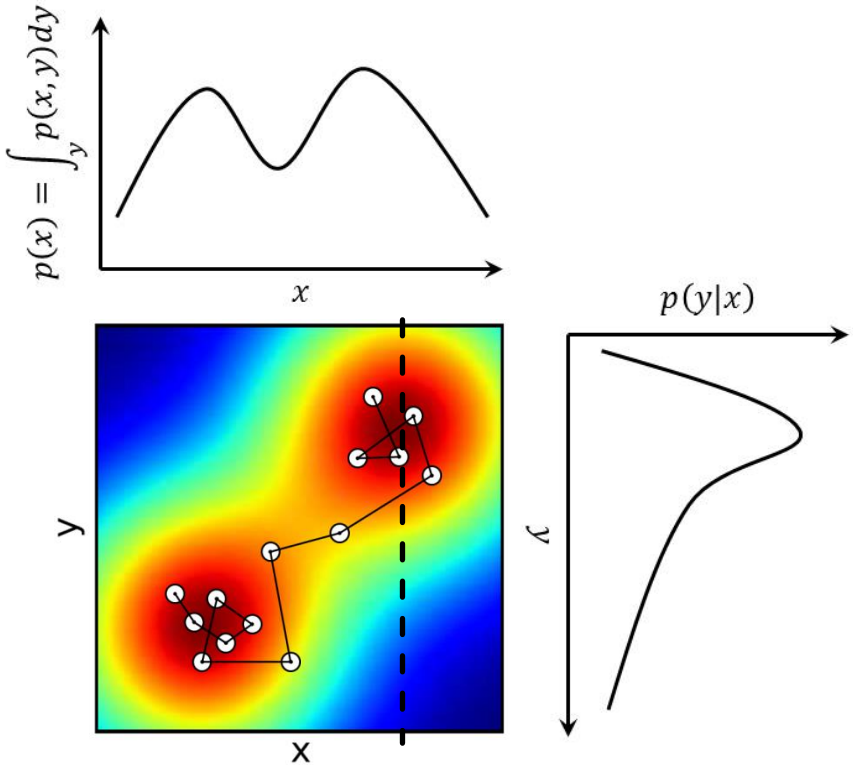
temporal aspects:

- increasingly correct representation
- anytime computing



computational complexity aspects:

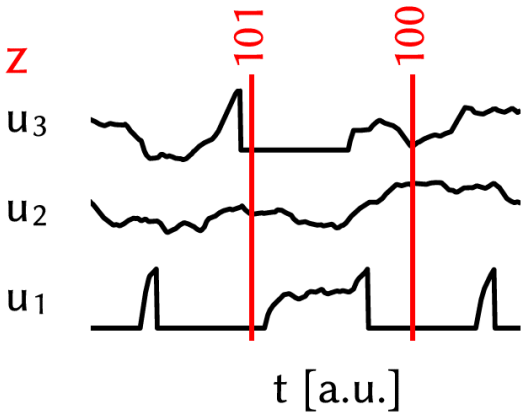
- computation of contitionals is simple
- marginalization is free



Spike-based encoding of an ensemble state

$z_k = 1 \;\Leftrightarrow\;$ neuron has spiked in $[t-\tau, t)$

$\rightarrow$ spike pattern encodes states $\boldsymbol{z}^{(t)}$

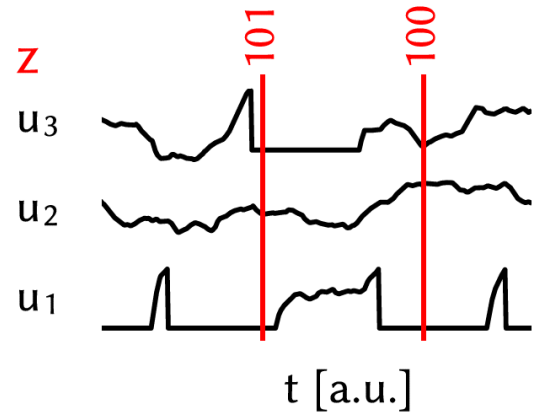


Büsing et al. (2011), Petrovici & Bill et al. (2016)

Emulation of Boltzmann machines

$z_k = 1 \;\Leftrightarrow\;$ neuron has spiked in $[t-\tau, t)$

$\rightarrow$ spike pattern encodes states $\boldsymbol{z}^{(t)}$



Neural computability condition

$$u_k = \log \frac{p(z_k = 1 \,|\boldsymbol{z}_{\setminus k})}{p(z_k = 0 \,|\boldsymbol{z}_{\setminus k})}$$

(which is equivalent to a logistic activation function $p(z_k = 1 \,|\boldsymbol{z}_{\setminus k}) = \frac{1}{1+\exp(-u_k)}$).

mediated by synaptic weights:

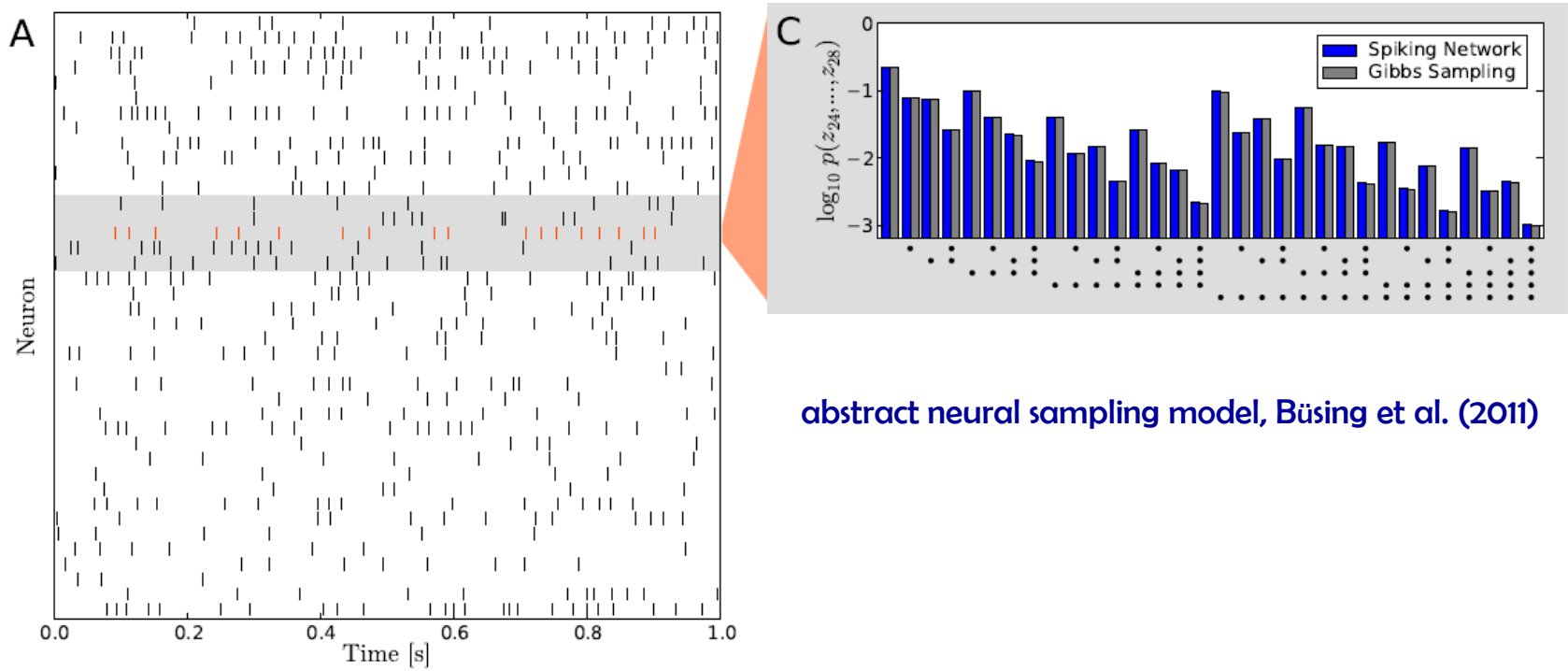
$$u_k = \sum_{i=1}^{K} W_{ki} z_i + b_k$$

Boltzmann distribution over $z_k \in \{0, 1\}$

$$p(\boldsymbol{z}) = \frac{1}{Z} \exp\left[\frac{1}{2}\boldsymbol{z}^T \boldsymbol{W} \boldsymbol{z} + \boldsymbol{z}^T \boldsymbol{b}\right]$$

Büsing et al. (2011), Petrovici & Bill et al. (2016)

Emulation of Boltzmann machines



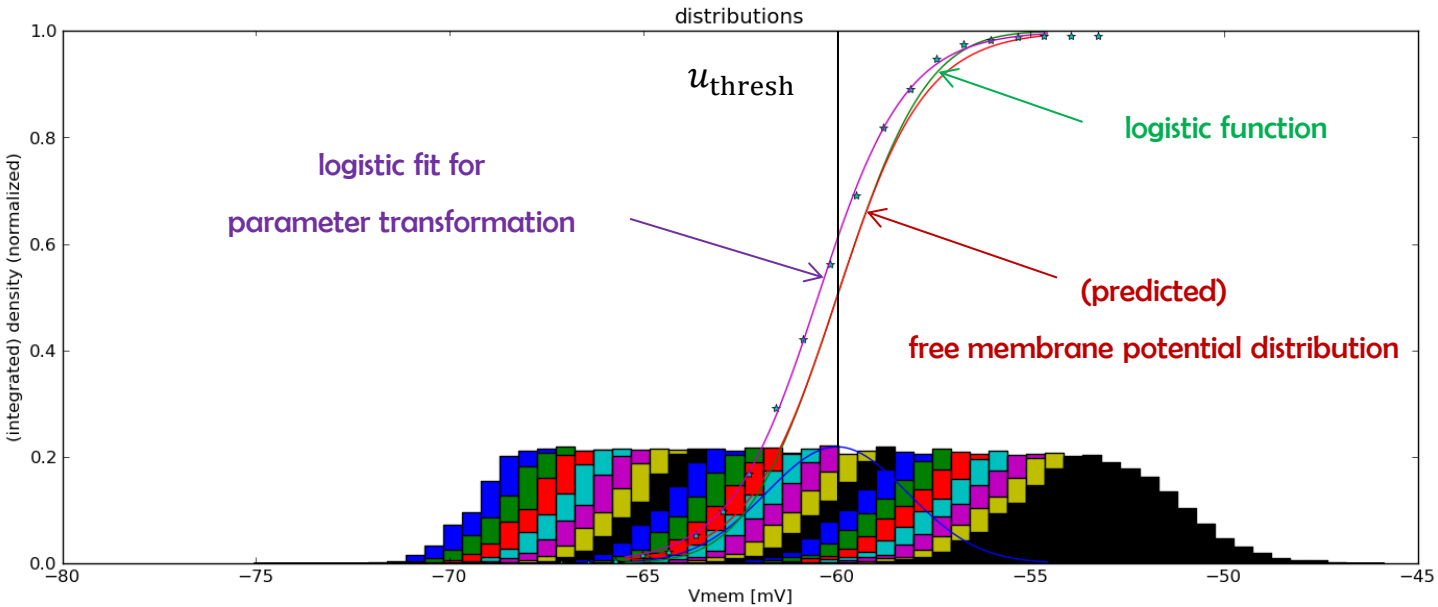
abstract neural sampling model, Büsing et al. (2011)

Neural computability condition

$$u_k = \log \frac{p(z_k = 1 \,|\boldsymbol{z}_{\setminus k})}{p(z_k = 0 \,|\boldsymbol{z}_{\setminus k})}$$

(which is equivalent to a logistic activation function $p(z_k = 1 \,|\boldsymbol{z}_{\setminus k}) = \frac{1}{1+\exp(-u_k)}$).

Idea: Stochasticity by Poisson background



$$\Rightarrow \quad p_{\text{spike}} \approx \text{erf}[\alpha \cdot (u_{\text{eff}} - \langle u_0 \rangle)]$$

$$\approx \sigma[\alpha \cdot (u_{\text{eff}} - \langle u_0 \rangle)]$$

unfortunately, neurons are a bit more complicated...

The diffusion approximation

noise source: Poisson spike trains
high background firing rates
relatively low synaptic weights

$\Rightarrow$ membrane as Ornstein-Uhlenbeck process

$$du(t) = \Theta \cdot [\mu - u(t)]dt + \sigma\, dW(t)$$

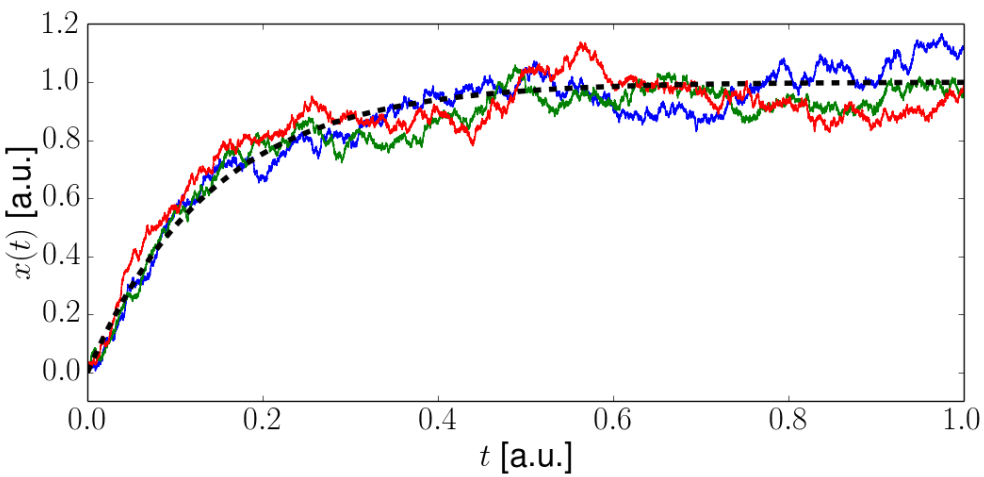
Ricciardi & Sacerdote (1979)

for COBA LIF:

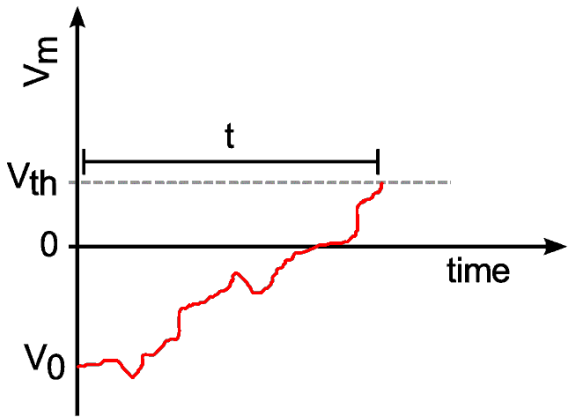
$$\Theta = \frac{1}{\tau_{\text{syn}}}$$

$$\mu = \frac{I^{\text{ext}} + g_{\text{l}} E_{\text{l}} + \sum_{i=1}^{n} \nu_i w_i E_i^{\text{rev}} \tau_{\text{syn}}}{\langle g^{\text{tot}} \rangle}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} \nu_i [w_i (E_i^{\text{rev}} - \mu)]^2 \tau_{\text{syn}}}{2 \langle g^{\text{tot}} \rangle^2}$$



First-passage-time calculations



$$\text{FPT}(a,0) \;:=\; \langle \inf\{t \geq 0 : V_t = a | V_0 = 0\}\rangle$$
$$= \sqrt{\frac{\pi}{\rho\sigma^2}} \int_0^{|a|} \left[1 + \operatorname{erf}\left(\sqrt{\frac{\rho}{\sigma^2}}\, x\right)\right] \exp\left(\frac{\rho x^2}{\sigma^2}\right) dx$$

Thomas (1975)

however, remember that $\tau_{\text{ref}} \approx \tau_{\text{syn}}$!

$\Rightarrow$ membrane does not forget !

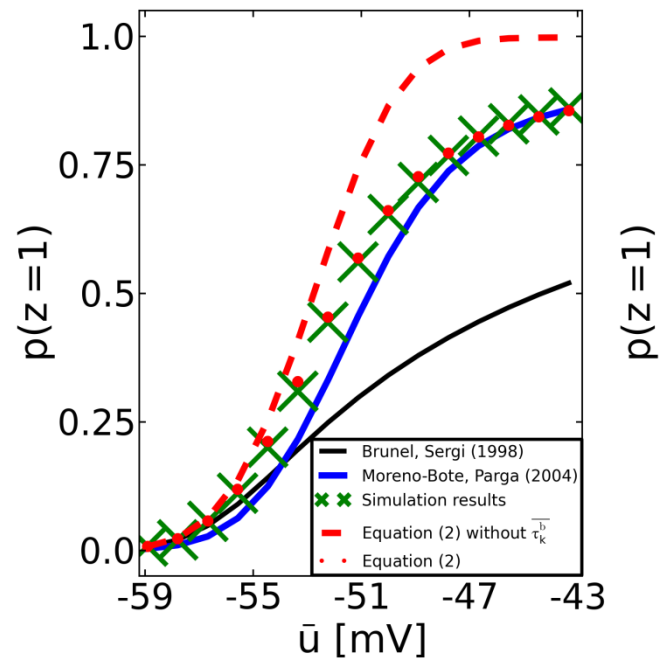
Brunel & Sergi (1998)

assumption: $\tau_{\text{syn}} \ll \tau_{\text{m}}$

this allows expansion in $\sqrt{\frac{\tau_{\text{syn}}}{\tau_{\text{m}}}}$:

$$\langle T \rangle = \tau\sqrt{\pi} \int_{\rho-\mu/\sigma}^{\vartheta_{\text{eff}}-\mu/\sigma} [1 + \operatorname{erf} x] \exp(x^2)\, dx$$

$$\nu_k = \frac{1}{\tau_{\text{ref}} + \text{FPT}(V_{th}, V_0)}$$

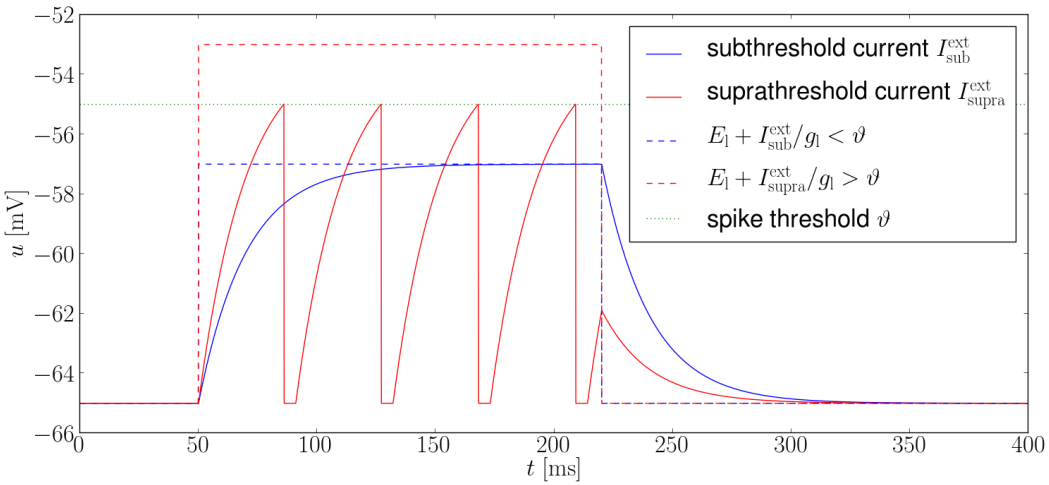


The adiabatic approximation

Moreno-Bote & Parga (2004)

assumption: $\tau_{\mathrm{syn}} \gg \tau_{\mathrm{m}}$

then, the synaptic input appears quasistatic to the membrane

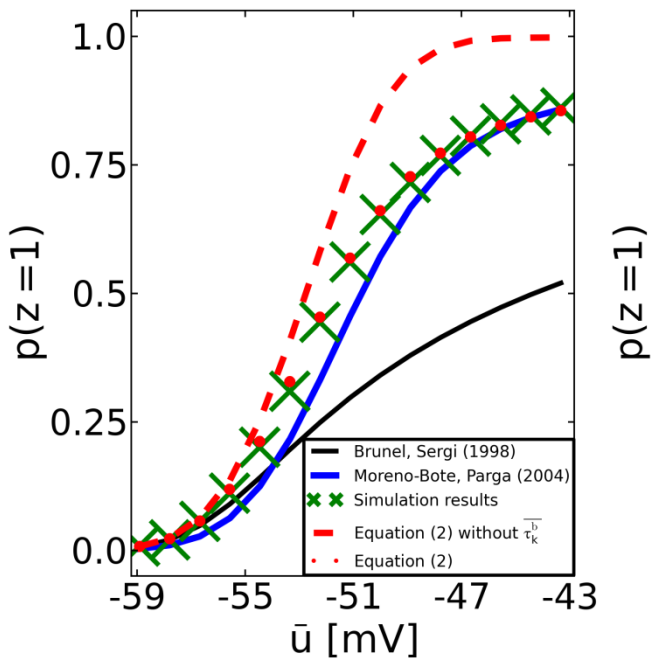


$$\tilde{\nu} = \left(\tau_{\mathrm{m}} \frac{\varrho - \tilde{u}}{\vartheta - \tilde{u}}\right)^{-1}$$

$$\nu = \int_{\vartheta}^{\infty} \tilde{\nu}(\tilde{u})\, p(\tilde{u})\, d\tilde{u}$$

however, remember that $\tau_{\mathrm{ref}} \approx \tau_{\mathrm{syn}}$!

$\Rightarrow$ adiabatic approximation does not hold !



The membrane autocorrelation propagation

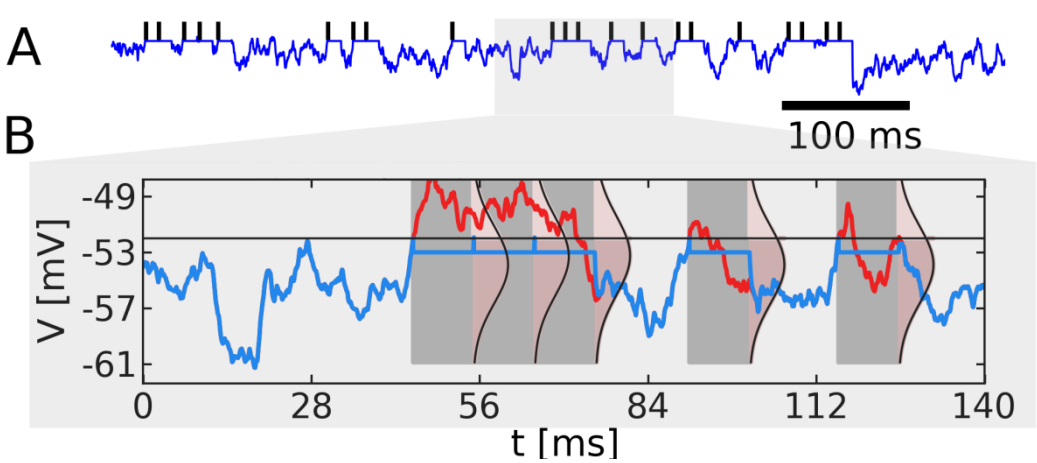
$$p(z_k = 1) = \frac{t_{k,\,\text{refractory}}}{t_{\text{total}}} = \frac{\sum_n P_n \, n \, \tau_{\text{ref}}}{\sum_n P_n \cdot \left(n \, \tau_{\text{ref}} + \sum_{k=1}^{n-1} \overline{\tau_k^b} + T_n \right)}$$

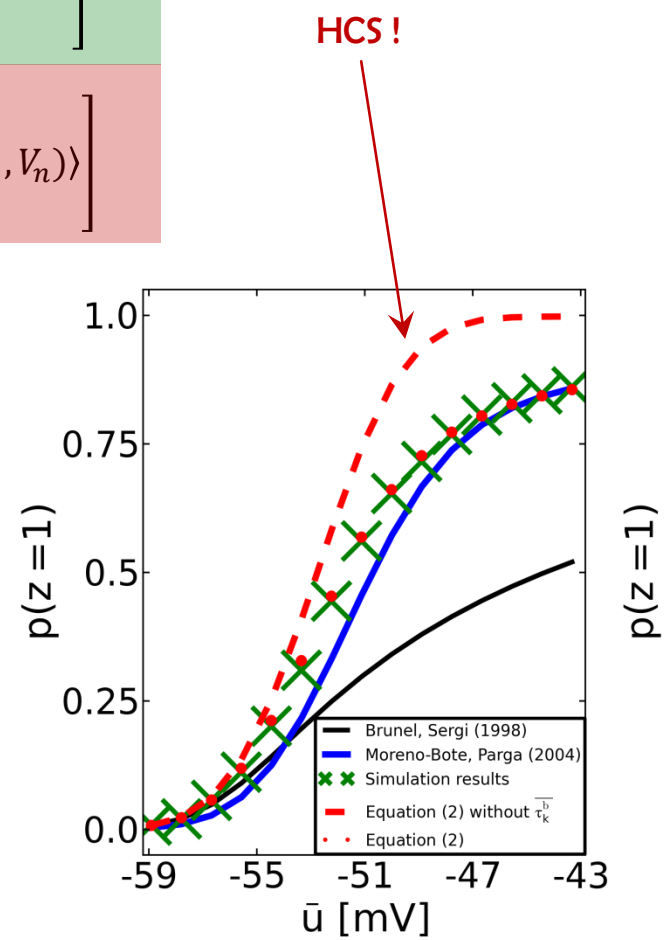
Petrovici & Bill et al. (2016)

$$P_n = \left(1 - \sum_{i=1}^{n-1} P_i \right) \cdot \int_{V_{\text{thr}}}^{\infty} dV_{n-1} \; p(V_{n-1} | V_{n-1} > V_{\text{thr}}) \left[\int_{-\infty}^{V_{\text{thr}}} dV_n \; p(V_n | V_{n-1}) \right]$$

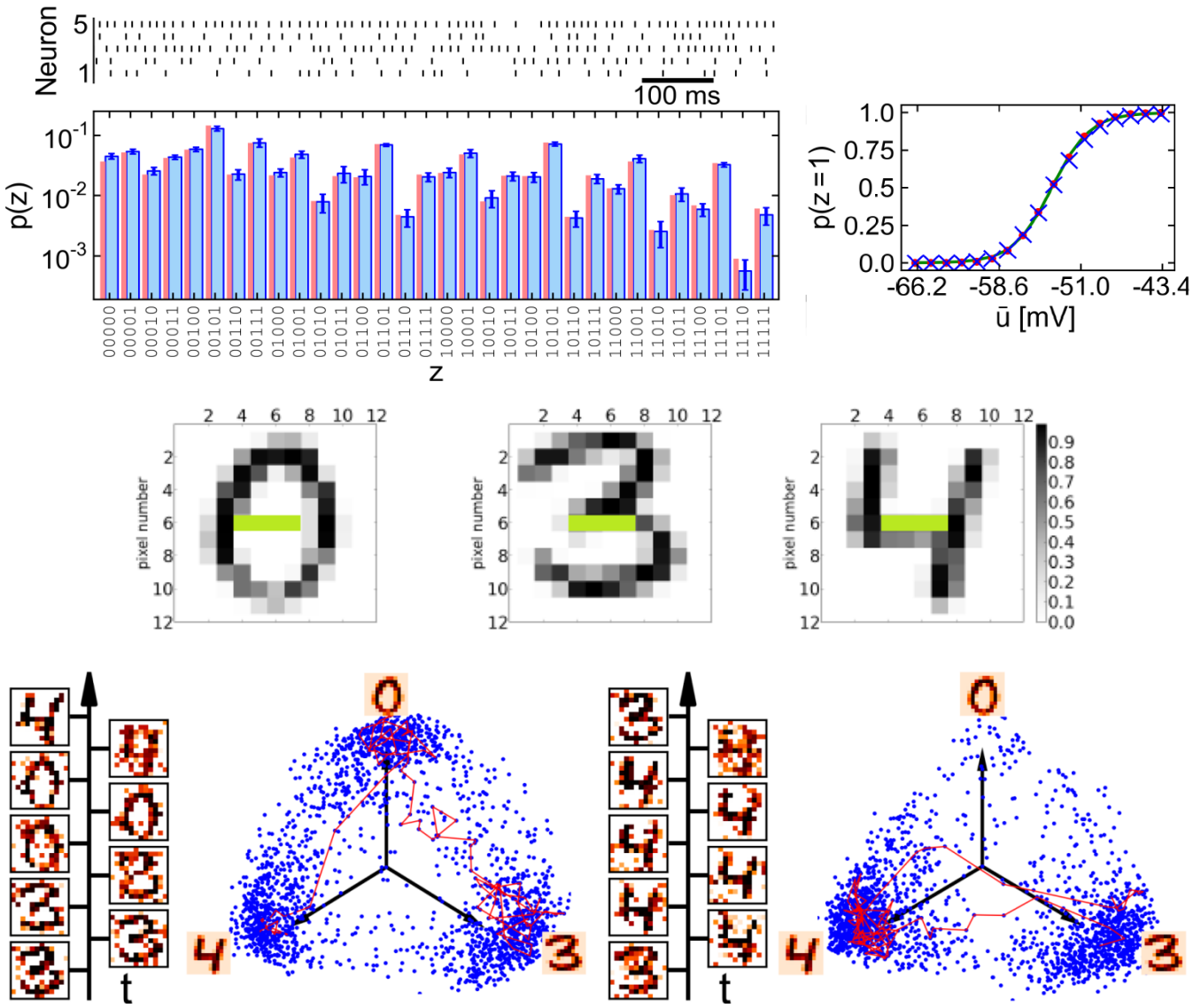
$$T_n = \int_{V_{\text{thr}}}^{\infty} dV_{n-1} \; p(V_{n-1} | V_{n-1} > V_{\text{thr}}) \left[\int_{-\infty}^{V_{\text{thr}}} dV_n \; p(V_n | V_{n-1}) \langle FPT(V_{\text{thr}}, V_n) \rangle \right]$$

$$\overline{\tau_k^b} = \int_{\vartheta}^{\infty} du_k \, \tau_{\text{eff}} \ln \left(\frac{\varrho - u_k}{\vartheta - u_k} \right) p(u_k | u_k > \vartheta, u_{k-1})$$



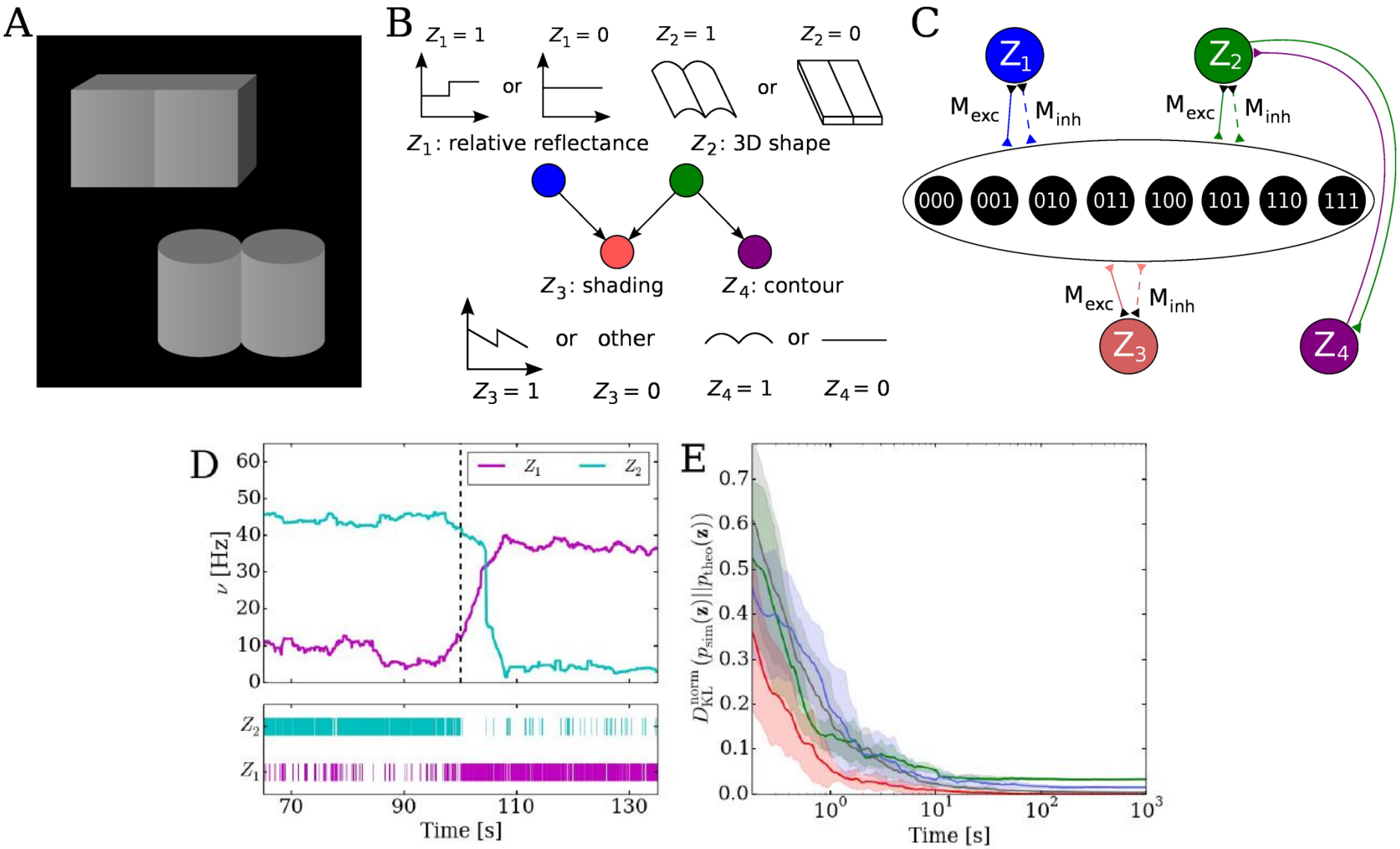


(Fully visible) LIF-based Boltzmann machines



Petrovici & Bill et al. (2016)

Beyond Boltzmann: Spiking Bayesian networks



Probst & Petrovici et al. (2015)

Deep spiking discriminative architectures

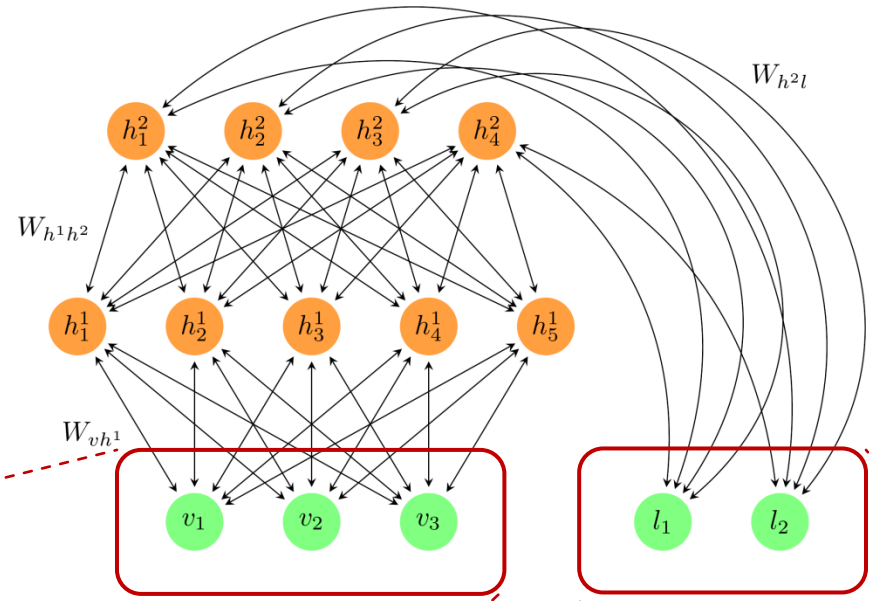
Training of RBMs/DBMs on MNIST:

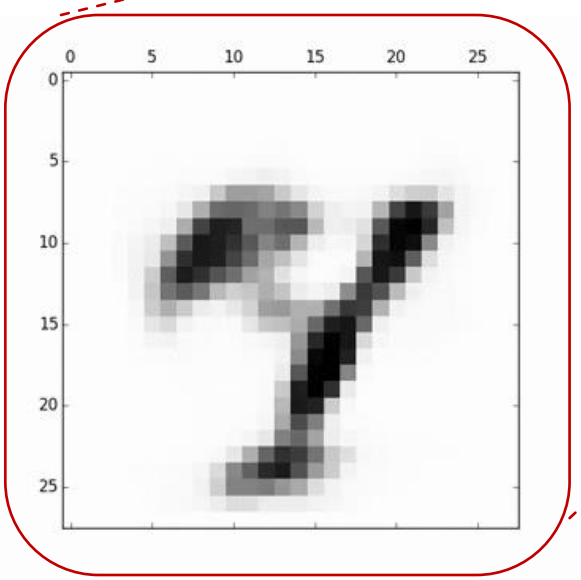
- maximum likelihood learning

$$\Delta w_{ij} \propto \langle z_i z_j \rangle_{\text{data}} - \langle z_i z_j \rangle_{\text{model}}$$

$$\Delta b_i \propto \langle z_i \rangle_{\text{data}} - \langle z_i \rangle_{\text{model}}$$

- coupled adaptive tempering

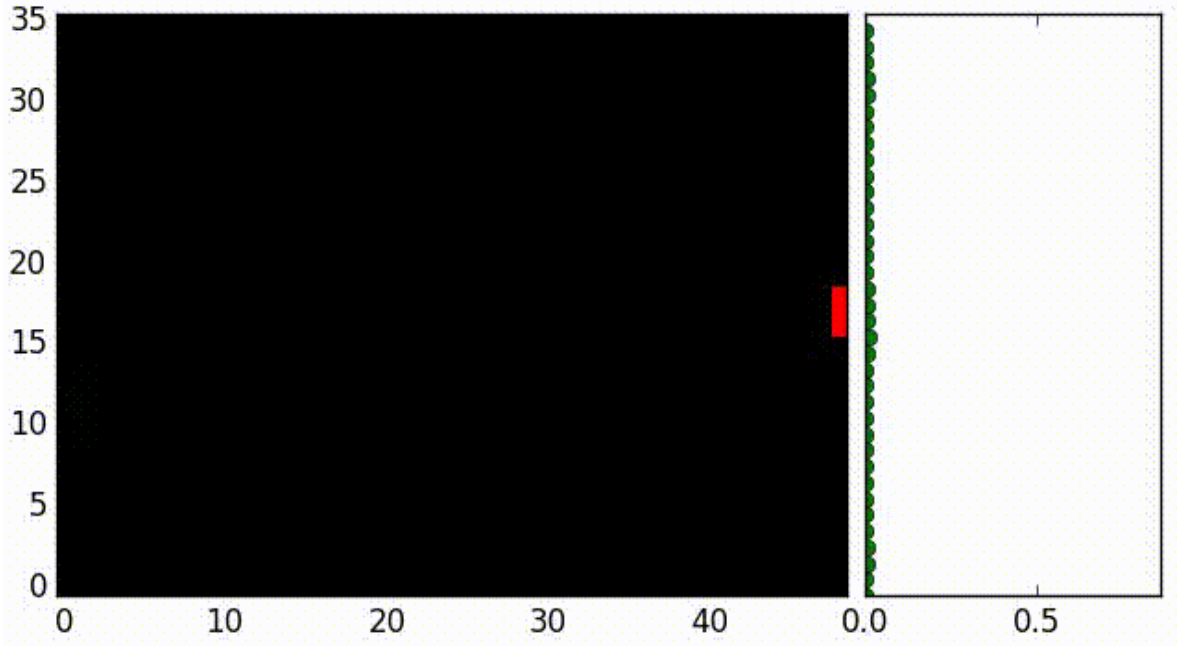


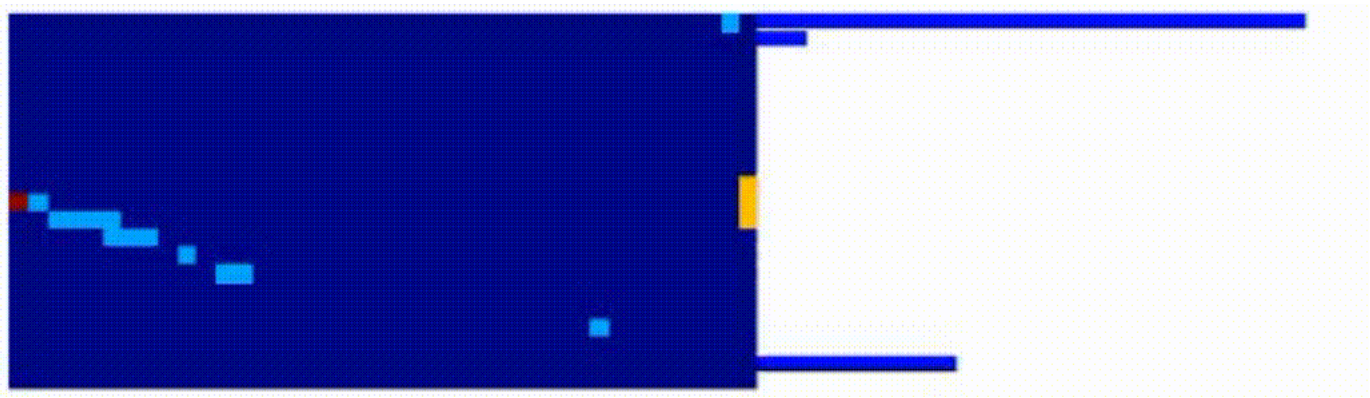


96.9 % correct classification with less than 2000 neurons

Leng & Petrovici et al. (2016)

Deep pong





Roth, Zenk (2017)

Deep spiking generative architectures

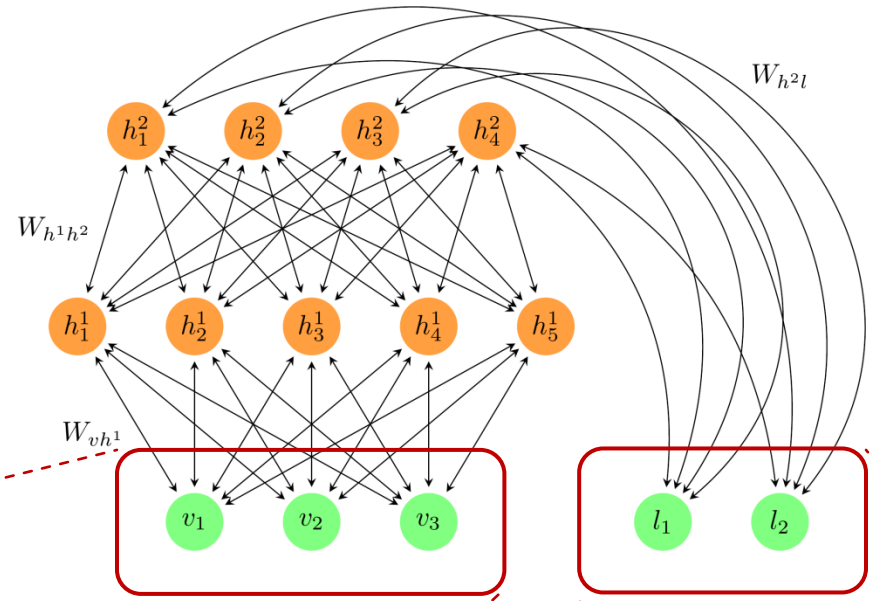
Training of RBMs/DBMs on MNIST:

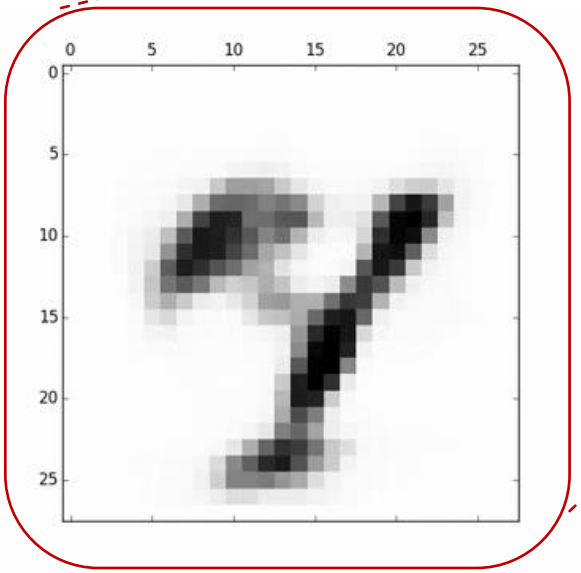
- maximum likelihood learning

$$\Delta w_{ij} \propto \langle z_i z_j \rangle_{\text{data}} - \langle z_i z_j \rangle_{\text{model}}$$

$$\Delta b_i \propto \langle z_i \rangle_{\text{data}} - \langle z_i \rangle_{\text{model}}$$

- coupled adaptive tempering



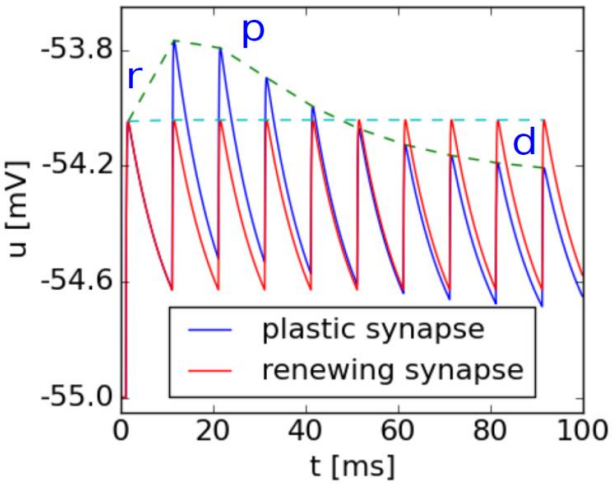


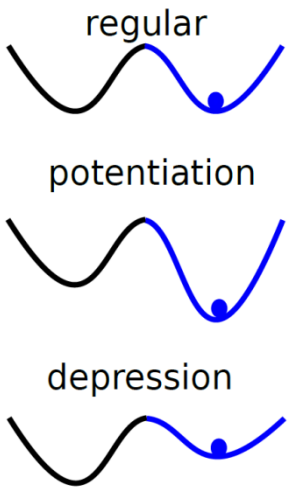
96.9 % correct classification with less than 2000 neurons

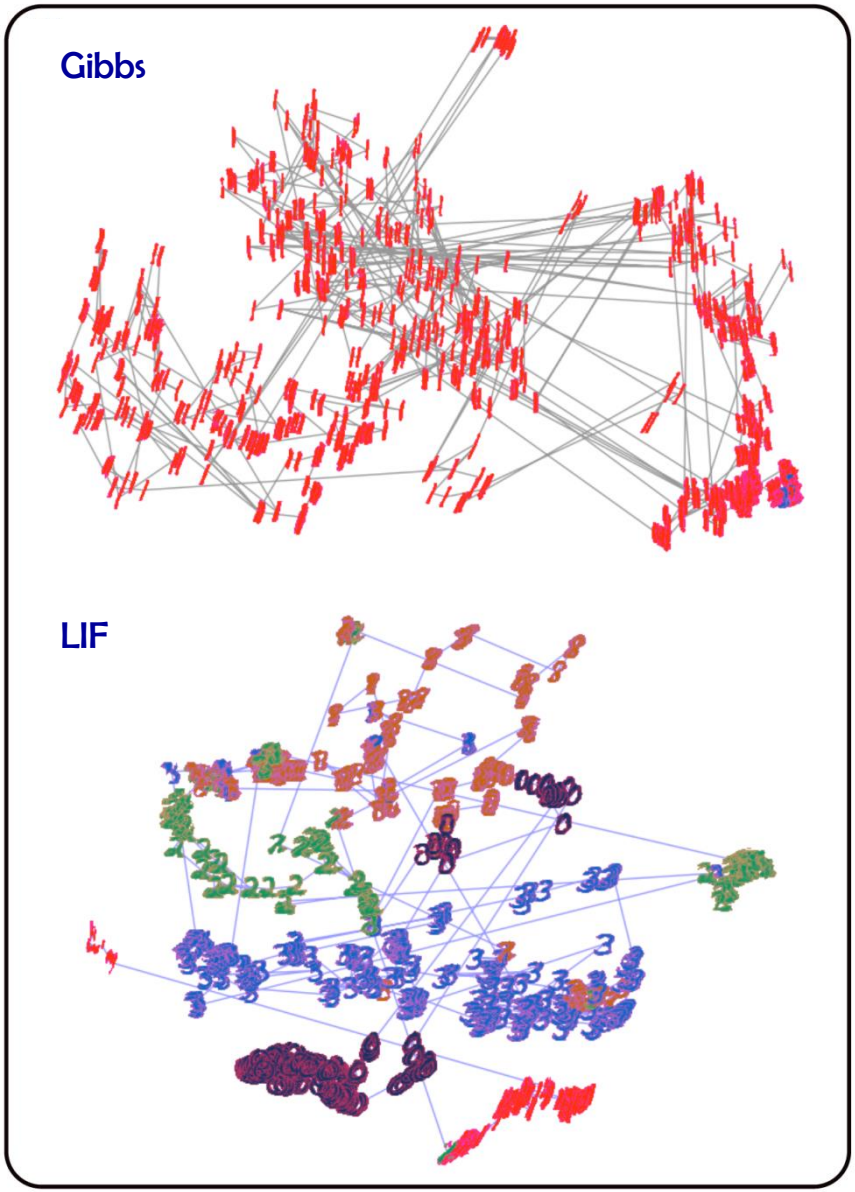
Leng & Petrovici et al. (2016)

Short-term plasticity enables superior mixing

STP model: Tsodyks & Markram (1997)



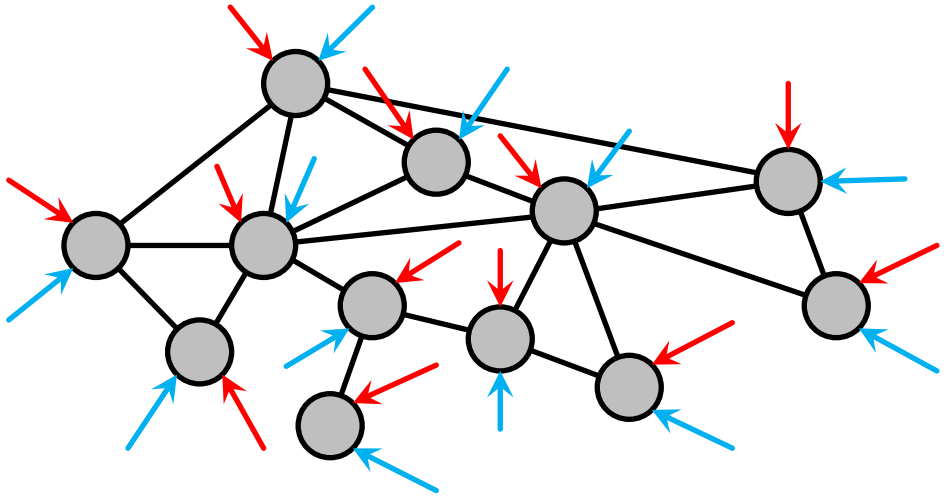




Leng & Petrovici et al. (2016)

... so where does the noise come from?

1st approximation: independent Poisson sources



unrealistic in both biological & artificial systems

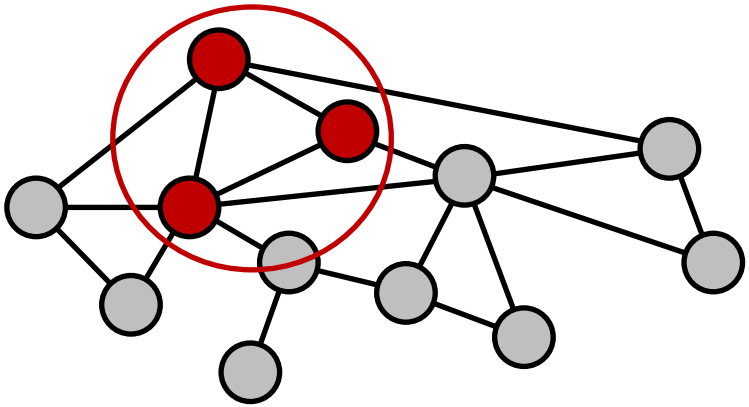
better: common pool of presynaptic partners

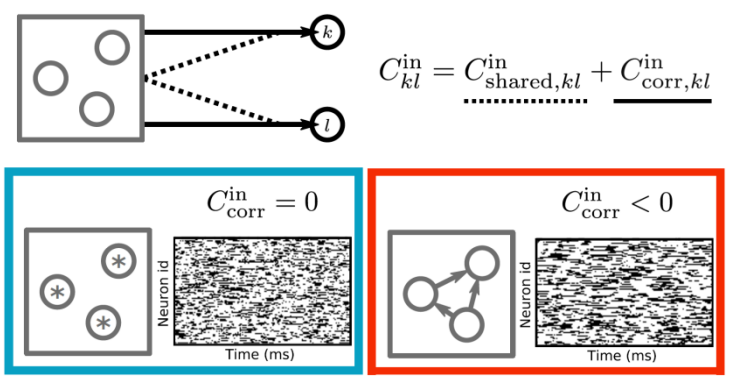
$\Rightarrow$ correlated inputs

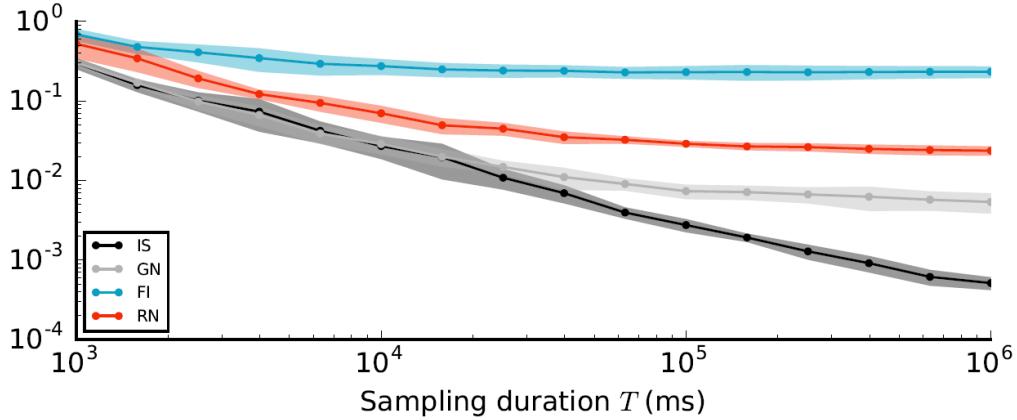
$\Rightarrow$ deviation from target distribution

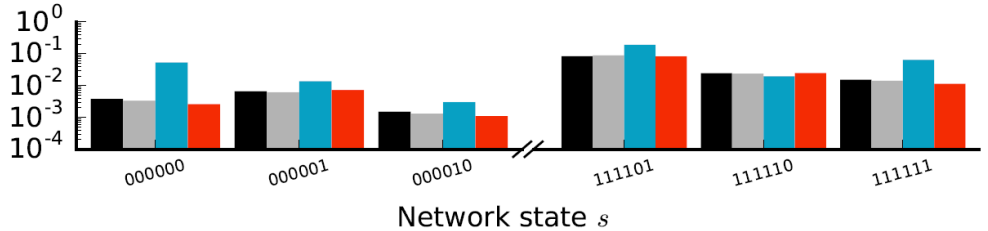
Embedded stochastic inference machines

more realistic: sea of noise



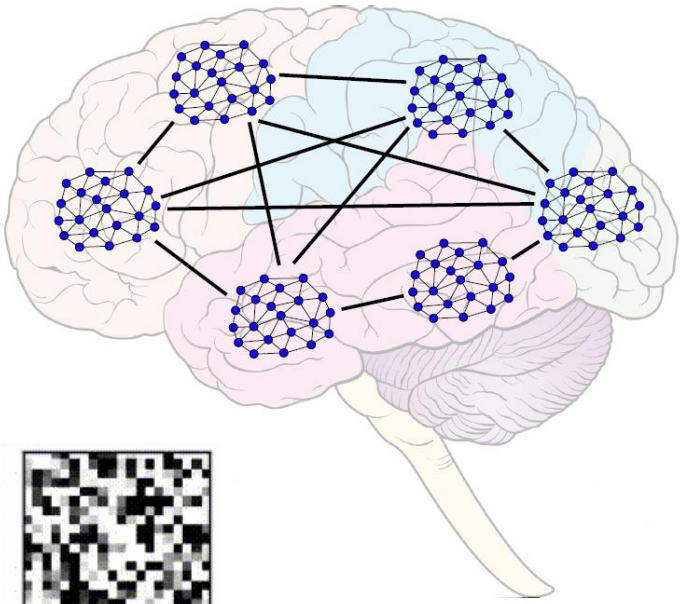


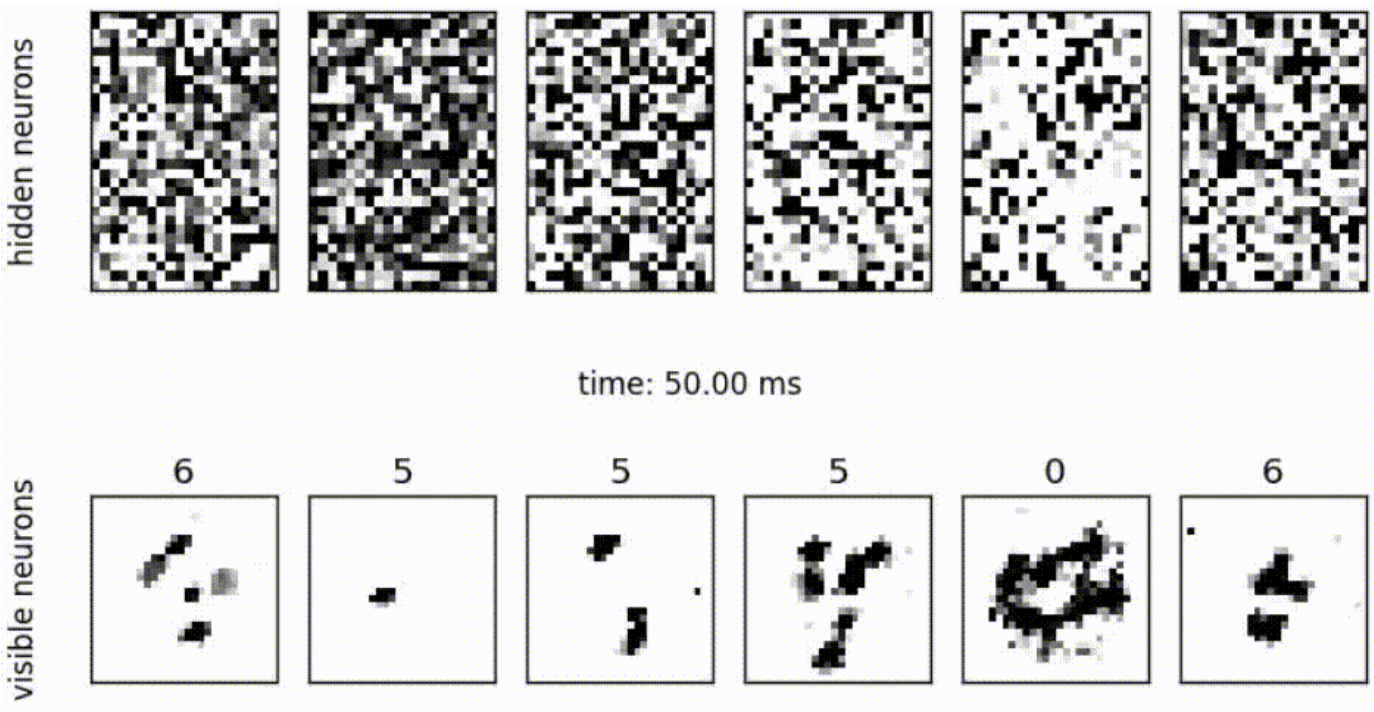




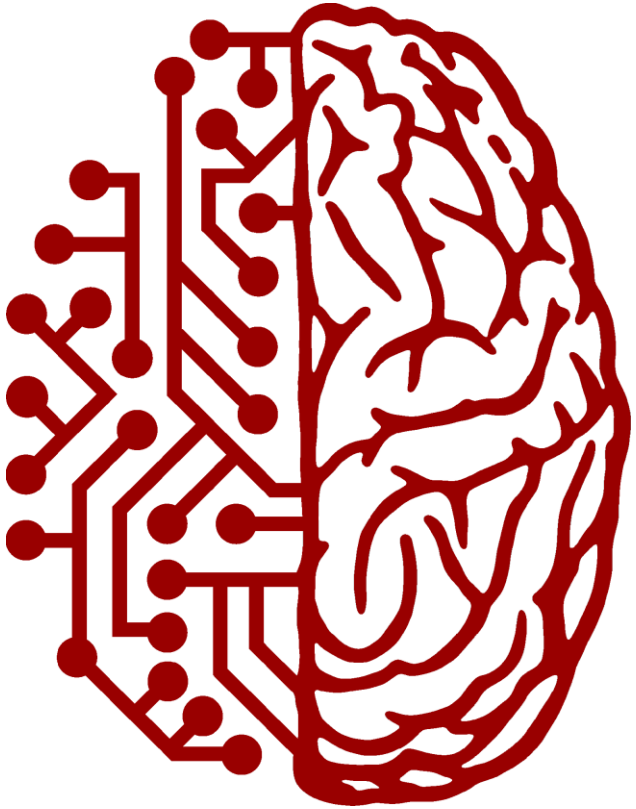
Jordan et al. (2015)

Noiseless stochastic computation



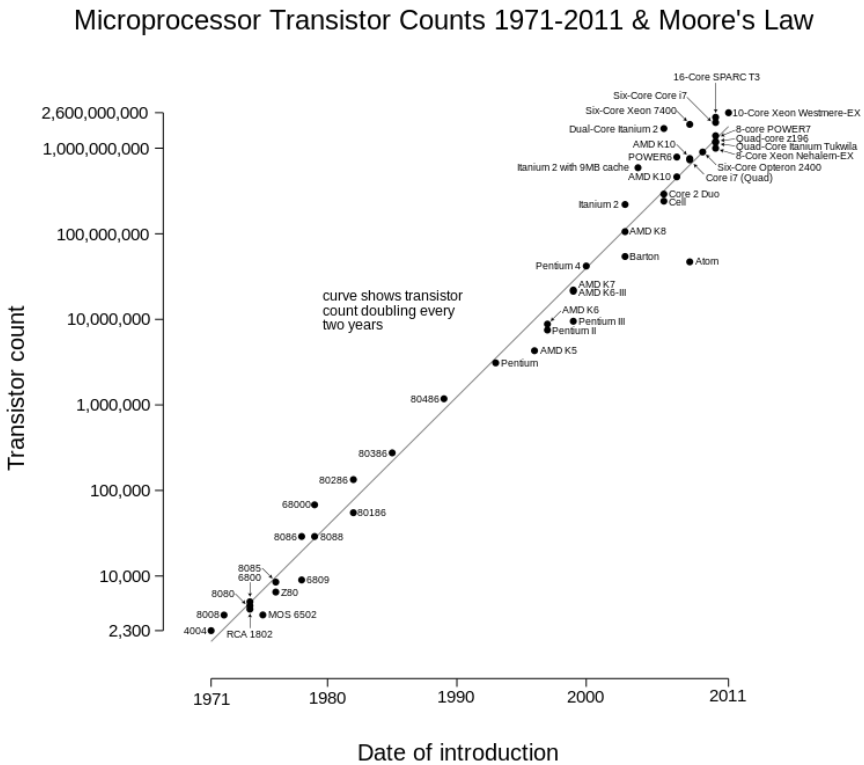


ongoing work with Dominik Dold and Ilja Bytschok



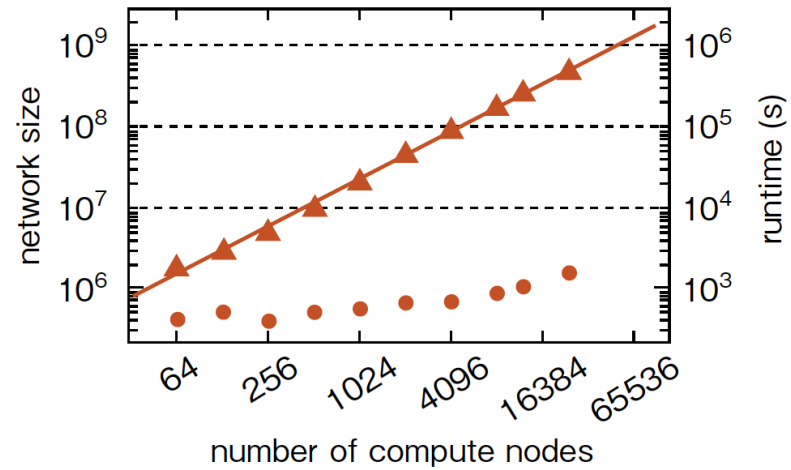
Physical emulation of spiking networks

Simulation: size & time



simulation speed 1520:1

compared to biological real-time



Diesmann (2012)

	nature	simulation
synaptic plasticity	seconds	hours
learning	days	years
development	years	millennia
evolution	> millennia	> millions of years

Simulation & emulation: energy scaling

Energy Scales

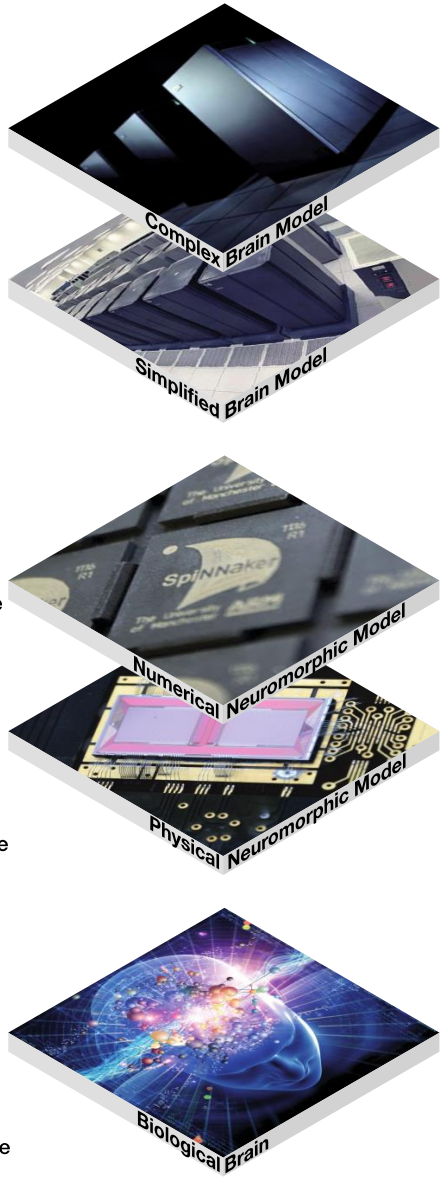
10^{0} J
1 Joule

10^{-4} J
0.1 milliJoule

10^{-8} J
10 nanoJoule

10^{-10} J
0.1 nanoJoule

10^{-14} J
10 femtoJoule

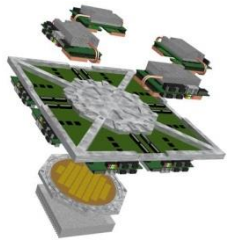


E F F I C I E N C Y

Blue Brain Project

nest:: initiative

SpiNNaker



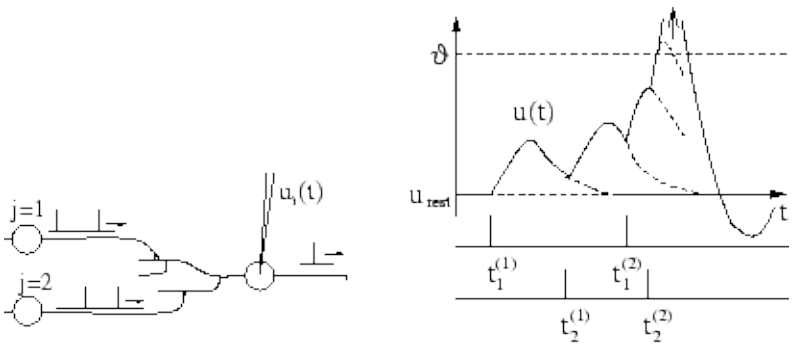
BrainScaleS

1.000.000 times more energy-efficient

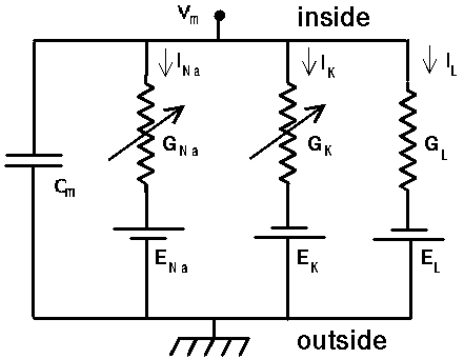
10.000 times less energy-efficient

Analog neuromorphic hardware

Electrophysiology



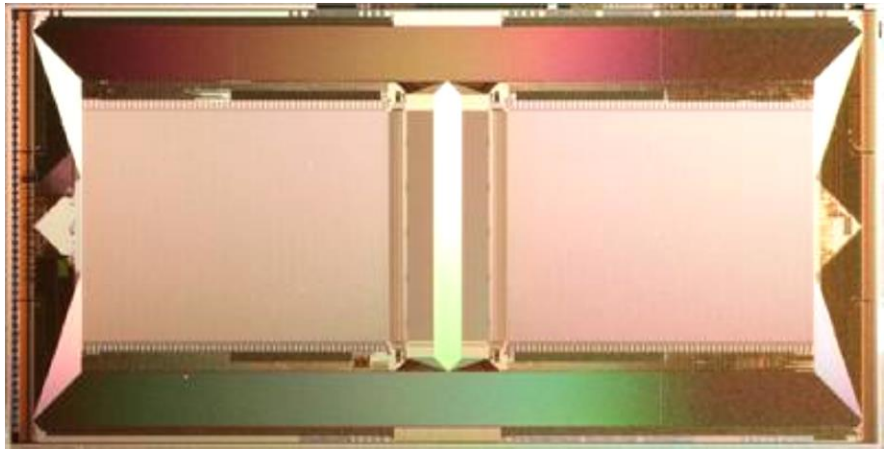
Hodgkin-Huxley-Model



$$C_m \dot{u} = g_L(u - E_L) + g_{\text{syn}}\,(u - E_{\text{syn}})$$

$$\cancel{g_{\text{K}}\, n^4 (u - E_{\text{K}}) + g_{\text{Na}}\, m^3 h (u - E_{\text{Na}})}$$

HICANN chip



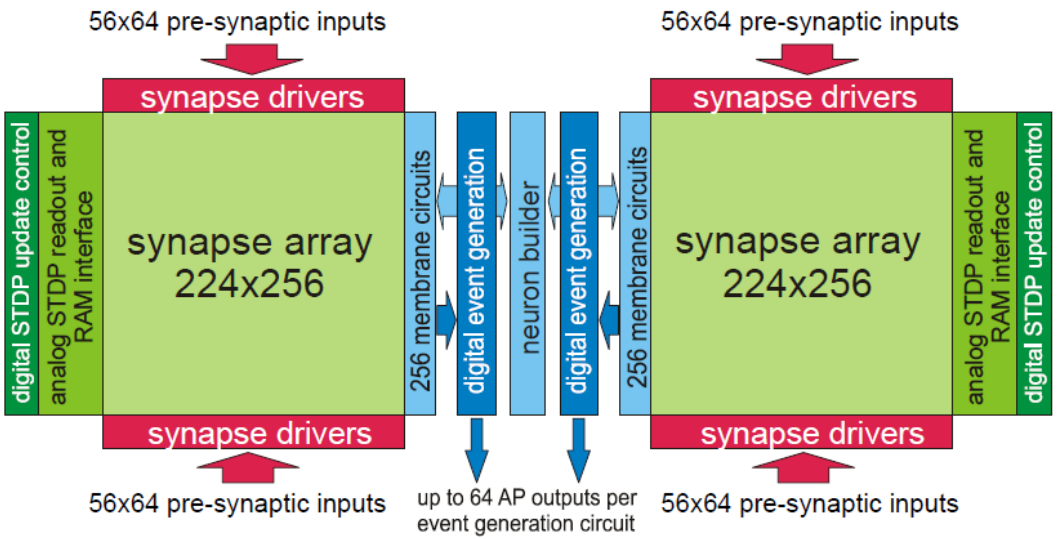
Schemmel et al. (2010)

Adaptive Exponential I&F Model

$$C_m \dot{u} = g_L(u - E_L) + g_{\text{syn}}\left(u - E_{\text{syn}}\right) + g_L \Delta_T \exp\left(\frac{V - V_T}{\Delta_T}\right) - w$$

$$\tau_w \dot{w} = a(V - E_L) - w$$

Analog neuromorphic hardware



HICANN chip

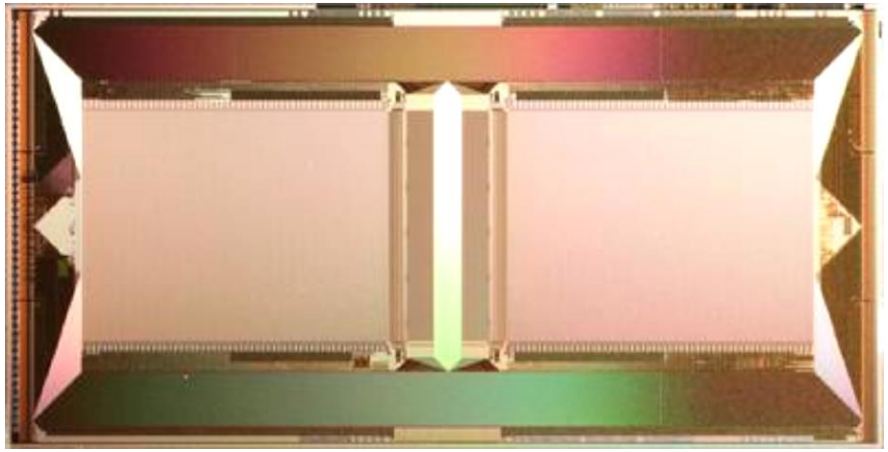
- mixed-signal VLSI:

  membrane → analog

  spikes → digital

- inherent speedup:

  $10^3 - 10^5$



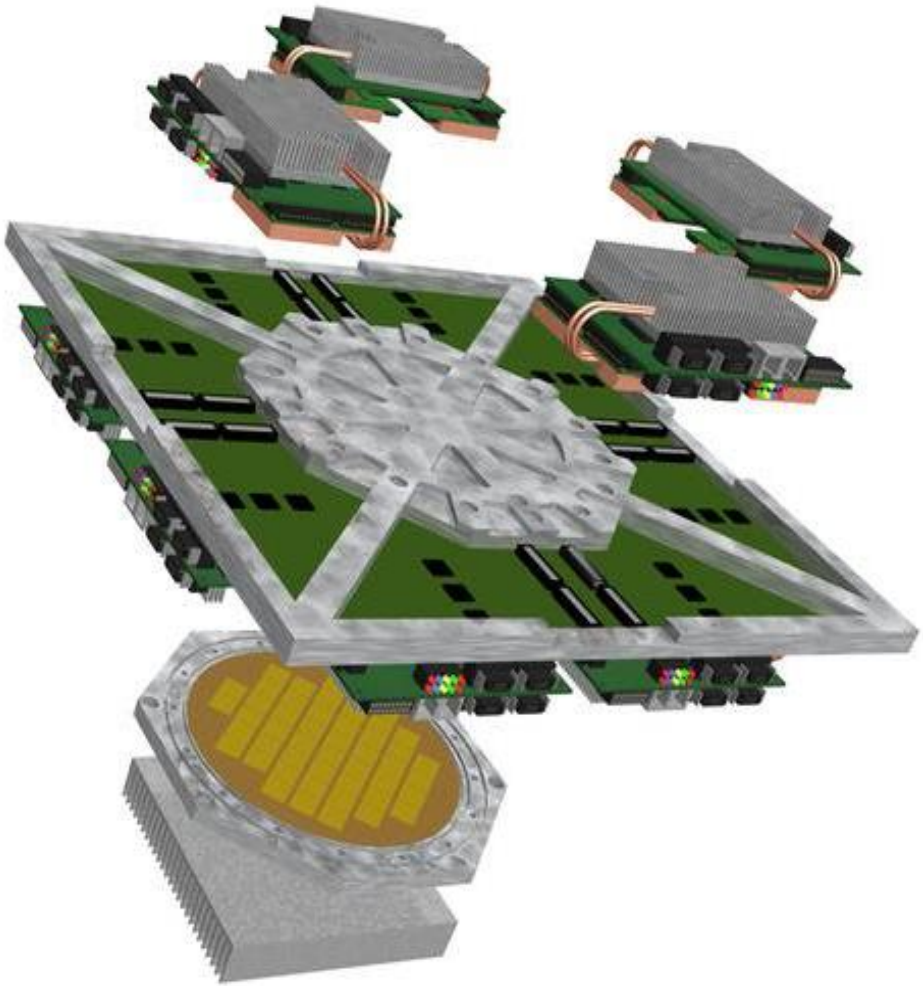
Schemmel et al. (2010)

Adaptive Exponential I&F Model

$$C_m \dot{u} = g_L(u - E_L) + g_{\text{syn}}(u - E_{\text{syn}}) + g_L \Delta_T \exp\left(\frac{V - V_T}{\Delta_T}\right) - w$$

$$\tau_w \dot{w} = a(V - E_L) - w$$

Waferscale integration → BrainScaleS system

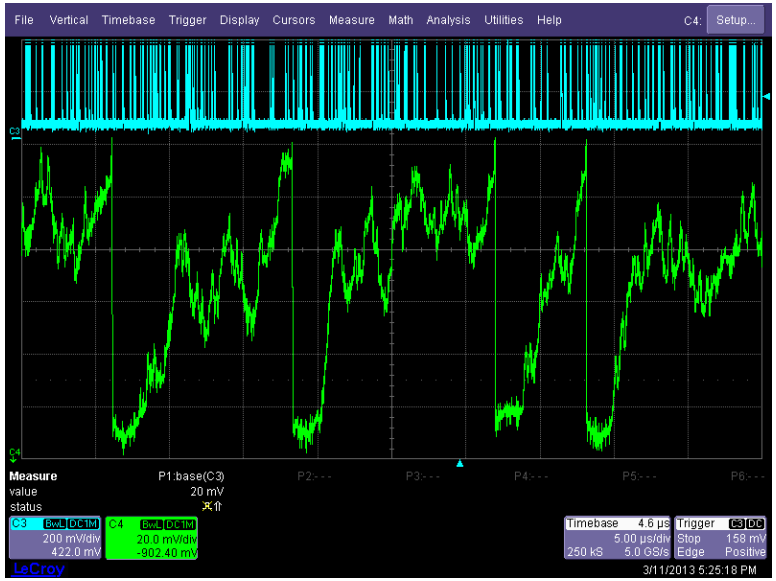


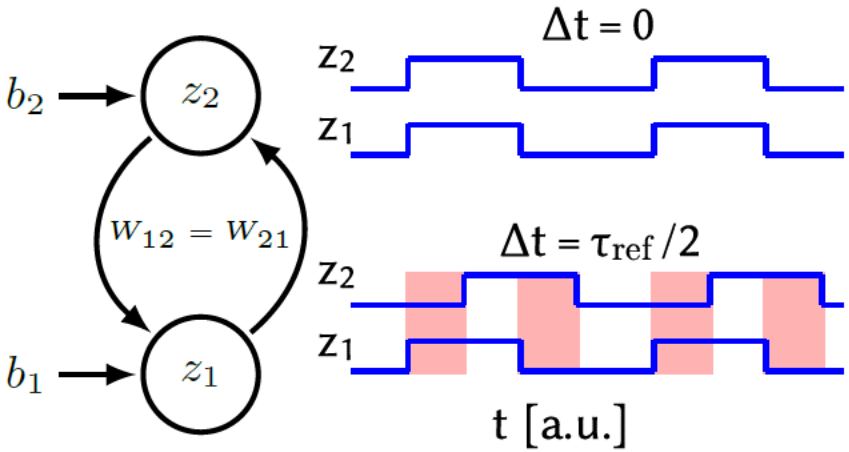
Schemmel et al. (2010)

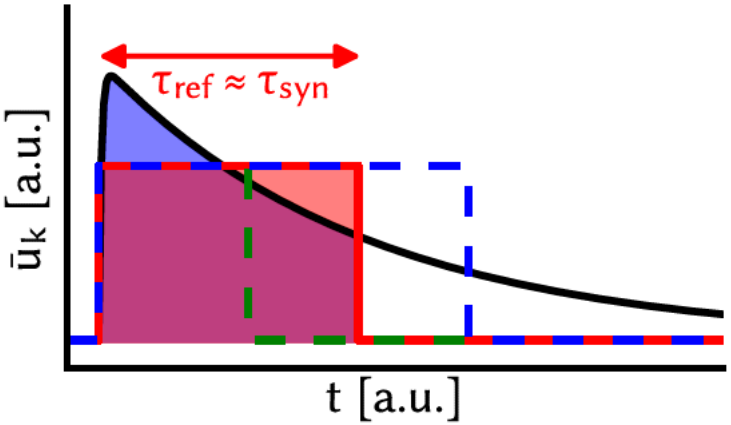
The Hybrid Modeling Facility in Heidelberg

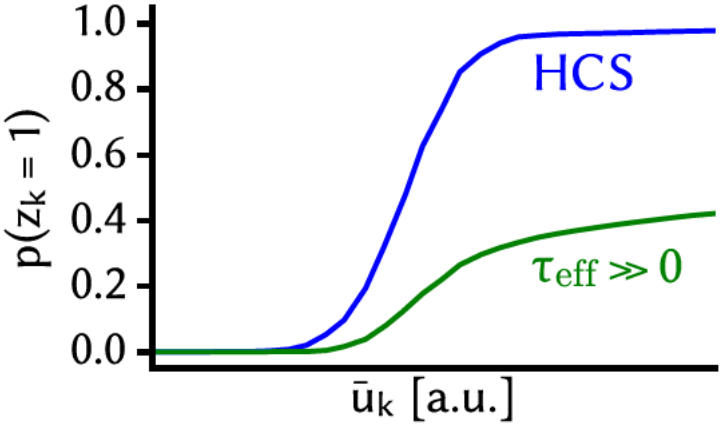
4 million AdEx neurons, 1 billion conductance-based synapses, under construction

Hardware is not software...

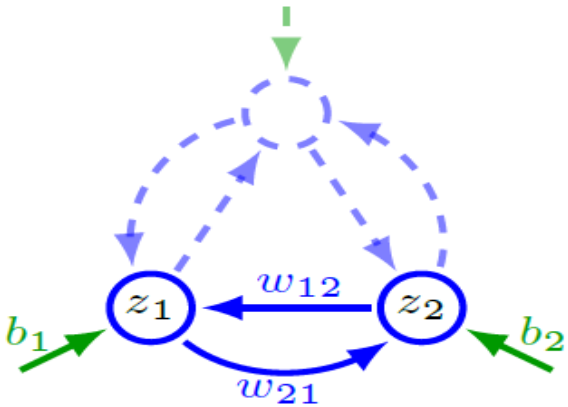


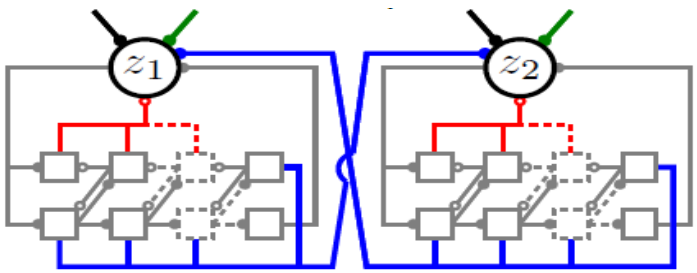


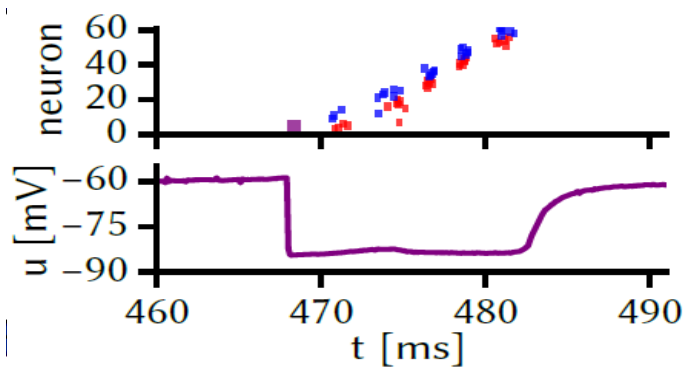


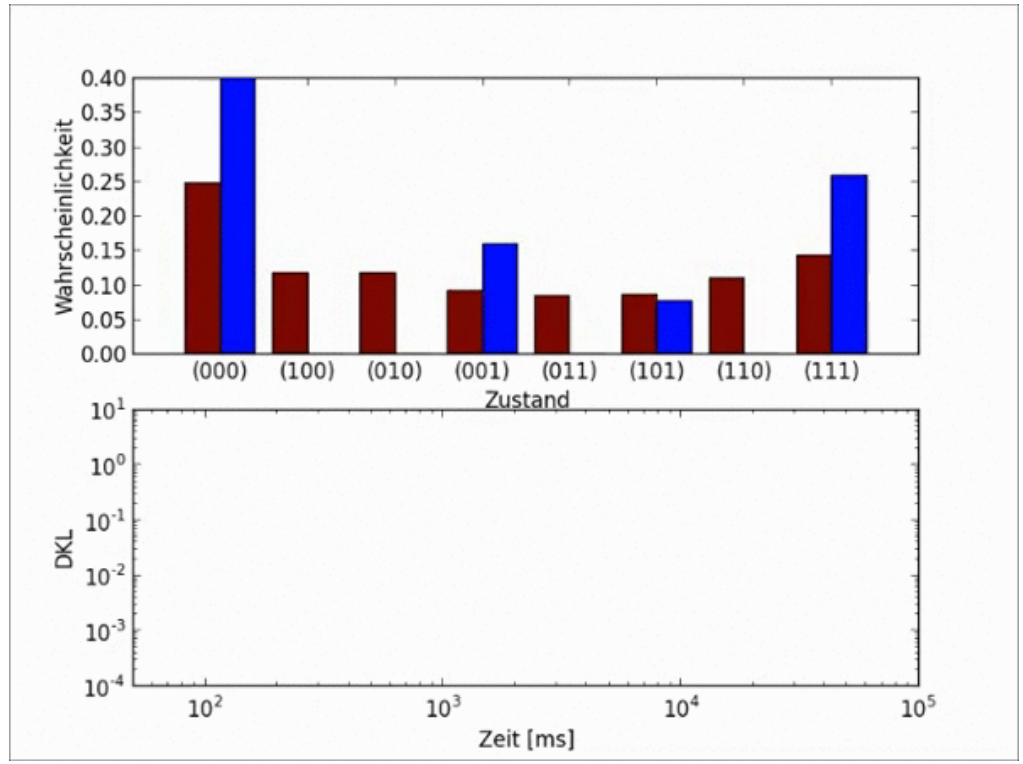


LIF sampling on accelerated hardware









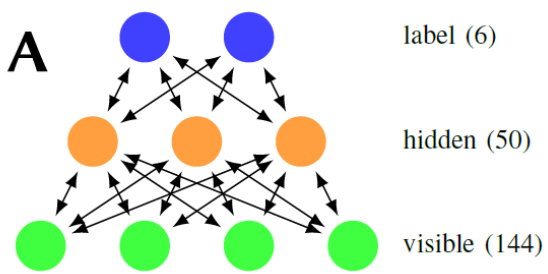
bio time: 100 s

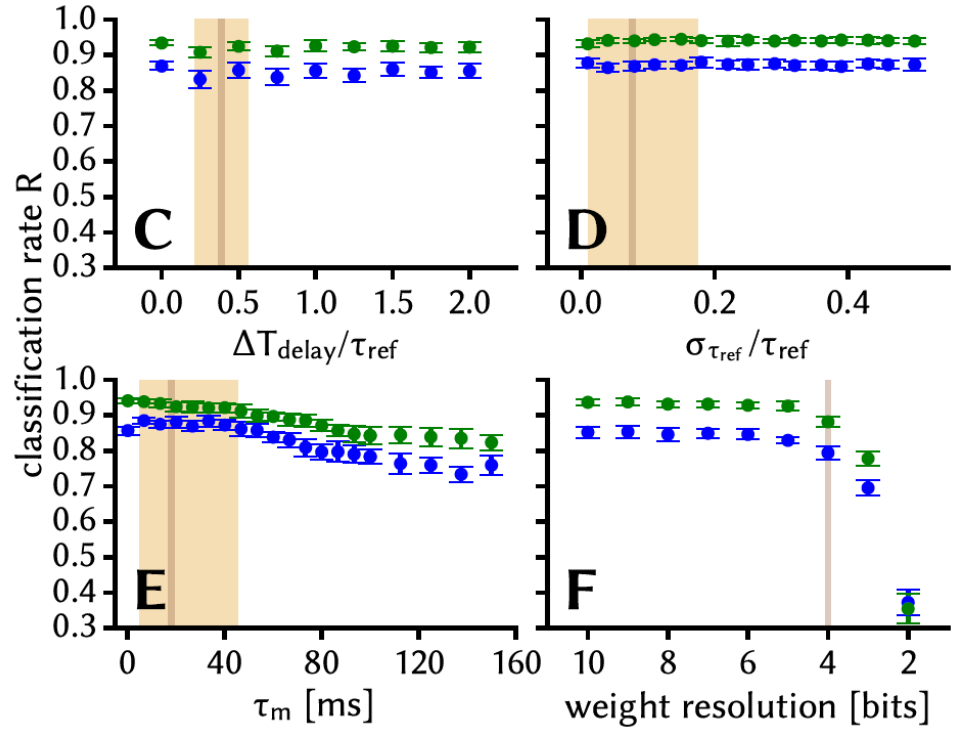
software simulation: 1 s

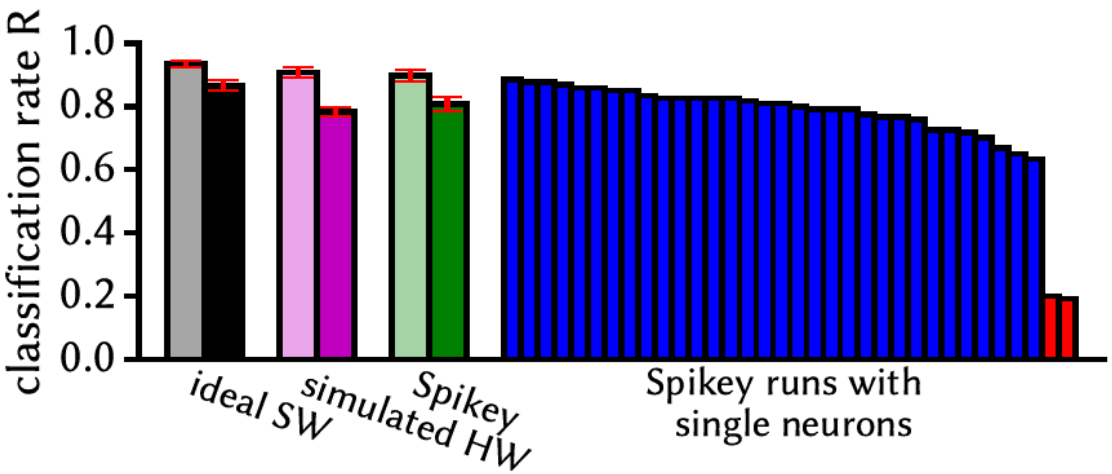
hardware emulation: 10 ms

Petrovici & Stöckel et al. (2015), Petrovici et al. (2017)

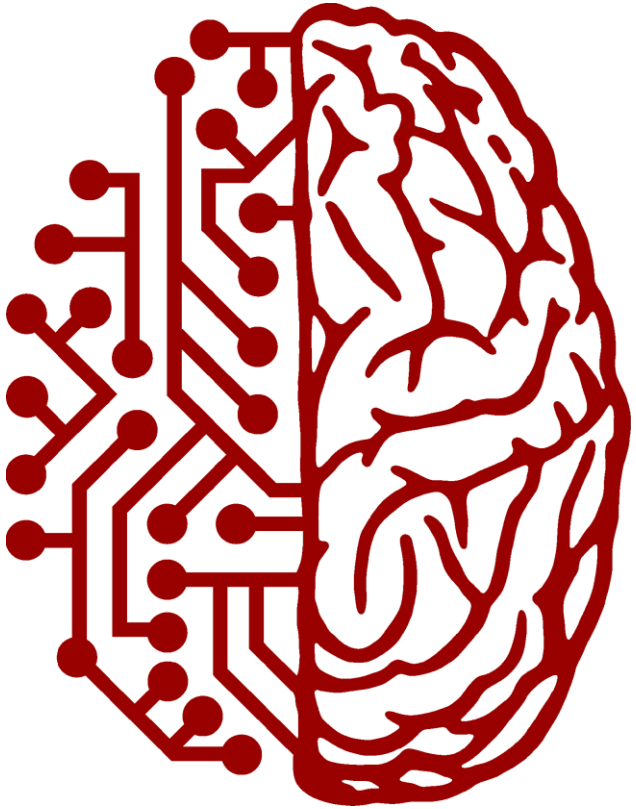
Robustness from structure





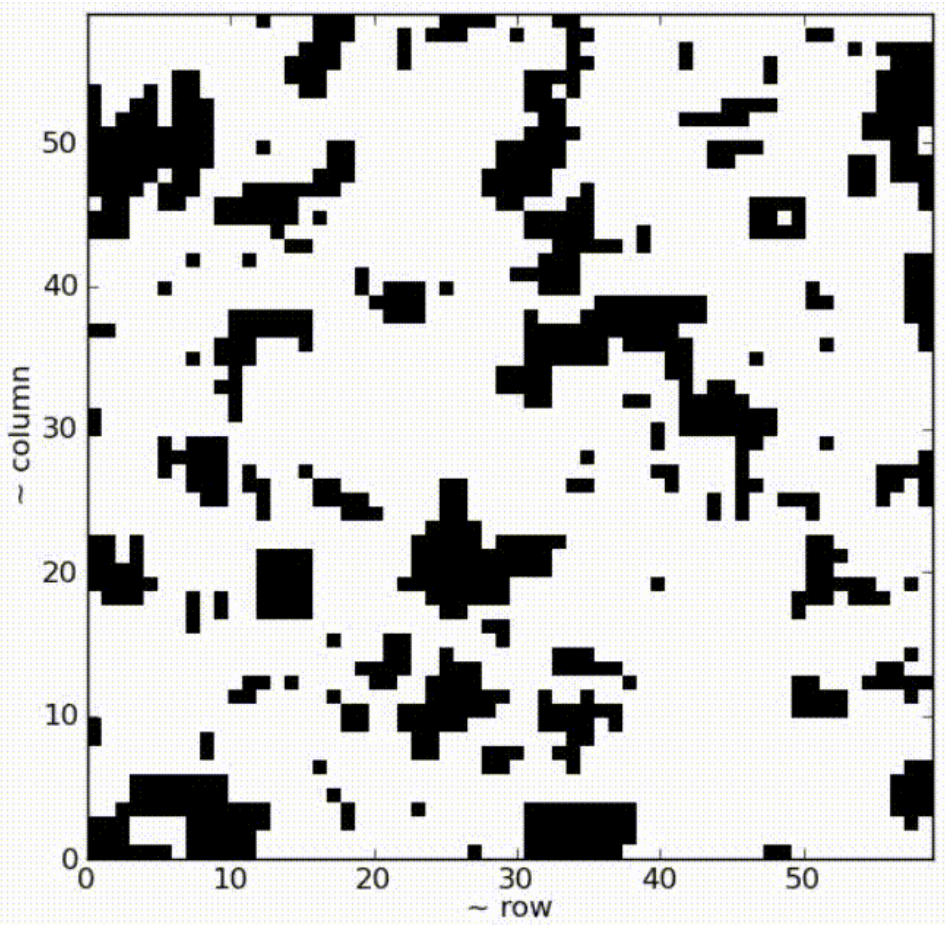


Petrovici & Schröder et al. (2017)



Outlook / Work in progress

Ensemble dynamics



$$p(\sigma) = \frac{1}{Z} \exp\left[\frac{1}{2}\sigma^T J\,\sigma + \sigma^T b\right]$$

spiking networks

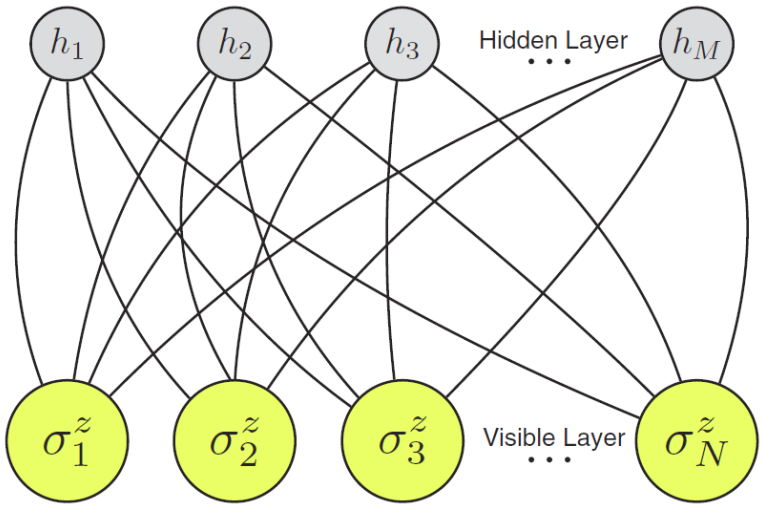
modeling

magnetic systems

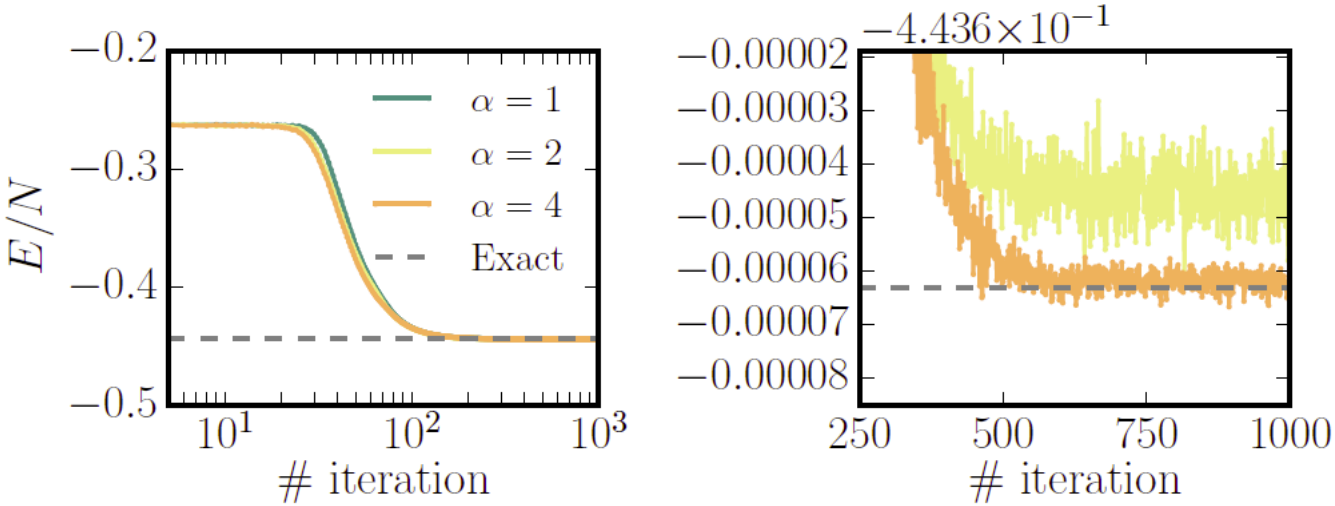
$$p(z) = \frac{1}{Z} \exp\left[\frac{1}{2} z^T W z + z^T b\right]$$

ongoing work with Andreas Baumbach

Quantum many-body problems



$$\Delta w \propto \frac{\partial}{\partial w} \frac{\langle \Psi_M | H | \Psi_M \rangle}{\langle \Psi_M | \Psi_M \rangle}$$



Carleo & Troyer (2017)

Learning rules

$$\dot{w}_k^a = \eta \left(\phi(U) - \phi(\alpha V^a)\right) \mathrm{PSP}_k$$

- **maximum likelihood learning**

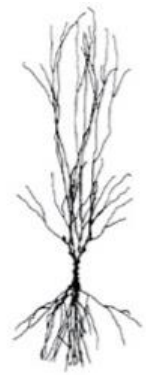
$$\Delta w_{ij} \propto \langle z_i z_j \rangle_{\text{data}} - \langle z_i z_j \rangle_{\text{model}}$$

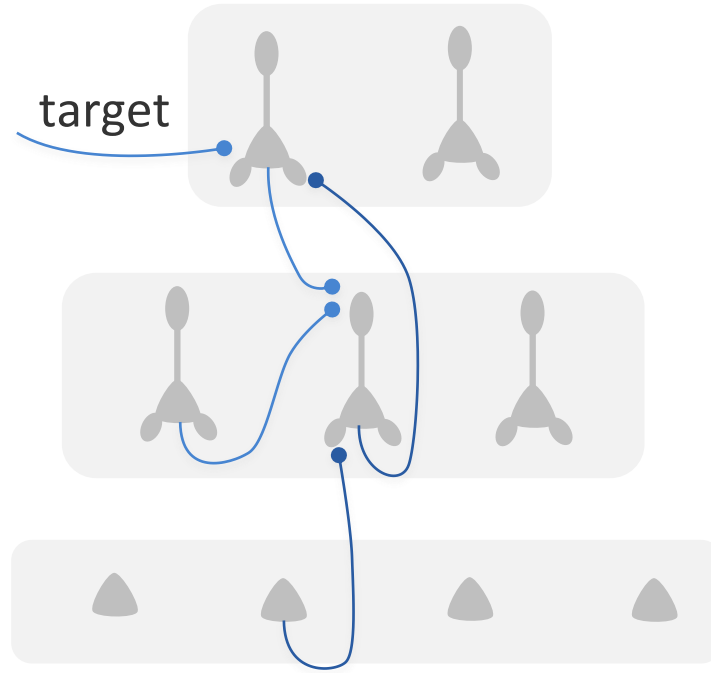
$$\Delta b_i \propto \langle z_i \rangle_{\text{data}} - \langle z_i \rangle_{\text{model}}$$

- **backprop**

$$\Delta w_{ij} \propto \frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial \text{net}_j} \frac{\partial \text{net}_j}{\partial w_{ij}}$$

vs.





ongoing work with Joao Sacramento and Walter Senn

References

- Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128, 1-58.
- Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. Nature Neuroscience, 20(4), 581-589.
- Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*(6013), 83-87.
- Salakhutdinov, R., & Hinton, G. (2009, April). Deep boltzmann machines. In *Artificial Intelligence and Statistics* (pp. 448-455).
- Schemmel, J., Briiderle, D., Griibl, A., Hock, M., Meier, K., & Millner, S. (2010, May). A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on* (pp. 1947-1950). IEEE.
- Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput Biol*, *7*(11), e1002211.
- Petrovici, M. A., Bytschok, I., Bill, J., Schemmel, J., & Meier, K. (2015). The high-conductance state enables neural sampling in networks of LIF neurons. *BMC Neuroscience*, *16*(1), O2.
- Petrovici, M. A., Bill, J., Bytschok, I., Schemmel, J., & Meier, K. (2016). Stochastic inference with spiking neurons in the high-conductance state. *Physical Review E*, *94*(4), 042312.
- Ricciardi, L. M., & Sacerdote, L. (1979). The Ornstein-Uhlenbeck process as a model for neuronal activity. *Biological cybernetics*, *35*(1), 1-9.
- Brunel, N., & Sergi, S. (1998). Firing frequency of leaky intergrate-and-fire neurons with synaptic current dynamics. *Journal of theoretical Biology*, *195*(1), 87-95.
- Moreno-Bote, R., & Parga, N. (2006). Auto-and crosscorrelograms for the spike response of leaky integrate-and-fire neurons with slow synapses. *Physical review letters*, *96*(2), 028101.
- Probst, D., Petrovici, M. A., Bytschok, I., Bill, J., Pecevski, D., Schemmel, J., & Meier, K. (2015). Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons. *Frontiers in computational neuroscience*, *9*.
- Leng, L., Petrovici, M. A., Martel, R., Bytschok, I., Breitwieser, O., Bill, J., ... & Meier, K. (2016). Spiking neural networks as superior generative and discriminative models. *Cosyne Abstracts, Salt Lake City USA*.
- Jordan, J., Tetzlaff, T., Petrovici, M., Breitwieser, O., Bytschok, I., Bill, J., ... & Diesmann, M. (2015). Deterministic neural networks as sources of uncorrelated noise for probabilistic computations. *BMC Neuroscience*, *16*(Suppl 1), P62.
- Diesmann, M. (2013). The road to brain-scale simulations on K. *Biosupercomput. Newslett*, *8*(8).
- Petrovici, M. A.*, Stöckel*, D., Bytschok, I., Bill, J., Pfeil, T., Schemmel, J. & Meier, K. (2015). Fast sampling with neuromorphic hardware. Advances in Neural Information Processing Systems (NIPS).
- Petrovici, M. A., Schroeder, A., Breitwieser, O., Grübl, A., Schemmel, J., & Meier, K. (2017). Robustness from structure: Inference with hierarchical spiking networks on analog neuromorphic hardware. *arXiv preprint arXiv:1703.04145*. (To appear in Proceedings of the ISCAS 2017.)
- Petrovici, M. A., Schmitt, S., Klähn, J., Stöckel, D., Schroeder, A., Bellec, G., ... & Güttler, M. (2017). Pattern representation and recognition with accelerated analog neuromorphic systems. *arXiv preprint arXiv:1703.06043*. (To appear in Proceedings of the IJCNN 2017.)
- Carleo, G., & Troyer, M. (2017). Solving the quantum many-body problem with artificial neural networks. *Science*, *355*(6325), 602-606.