Ideas for a Biologically Inspired Bayesian Computer

Wolfgang Maass
Institut für Grundlagen der Informationsverarbeitung
Technische Universität Graz, Austria

A *paradigm shift* in our understanding of human reasoning

- For over 2000 years researchers have believed that mathematical logic is the best framework for understanding and reproducing human reasoning and intelligence

- But mathematical logic turned out to be of little use for reasoning with uncertain facts (e.g., for everyday reasoning), and related AI approaches have failed

- A new mathematical framework (Bayesian networks, Belief networks, „graphical models“, „probabilistic inference“) was invented around 1990, that provides principled methods for reasoning with unreliable facts and beliefs

- Research in Cognitive Science during the last two decades has shown that this new framework provides a substantially better basis for understanding (and reproducing) human reasoning

- *But:* Probabilistic inference is very computing-intense, and can therefore not yet be used to provide useful reasoning capabilities in regular computers

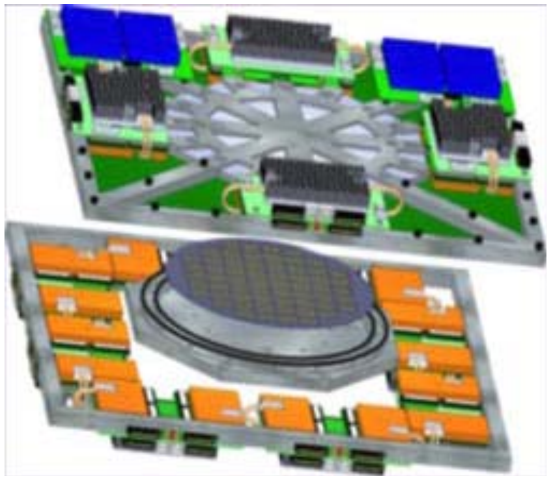
Resulting *challenges* for computer science

- Understand how the brain carries out probabilistic inference (with a 50 Watt power consumption, and stochastic computing elements)

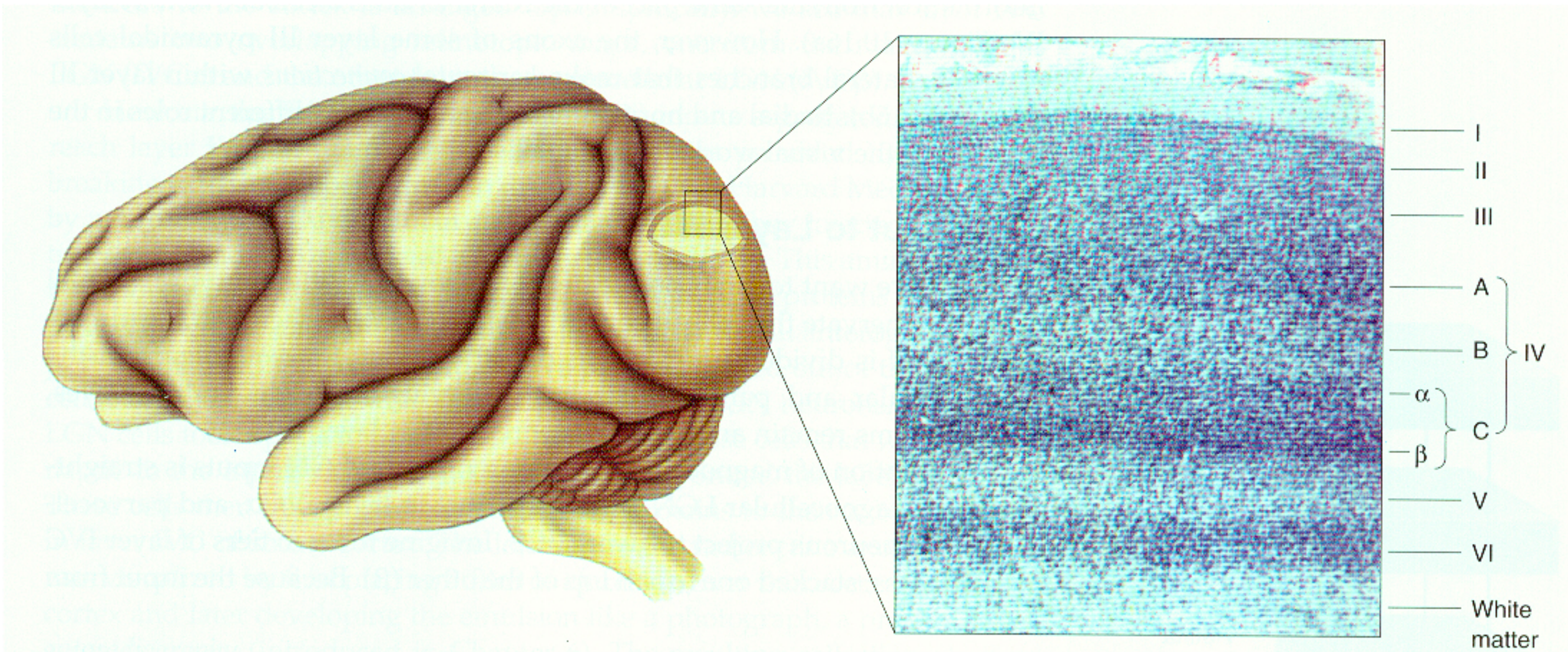
- Learn to make use of inherently stochastic aspects of computing elements on the molecular scale for artificial computing devices

- Build a *Bayesian Computer* based on this insight, that is able to reason with large numbers of uncertain facts and beliefs

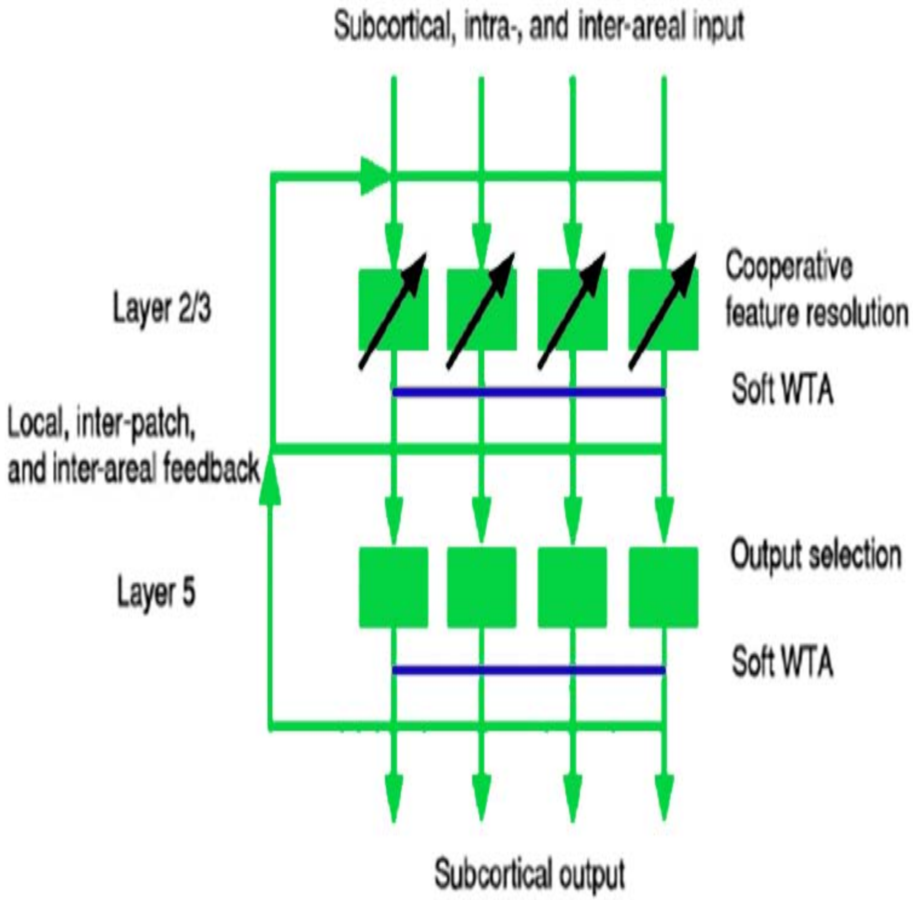
Our first concrete target: Port probabilistic inference capabilities to the spike-based hardware developed in FACETS



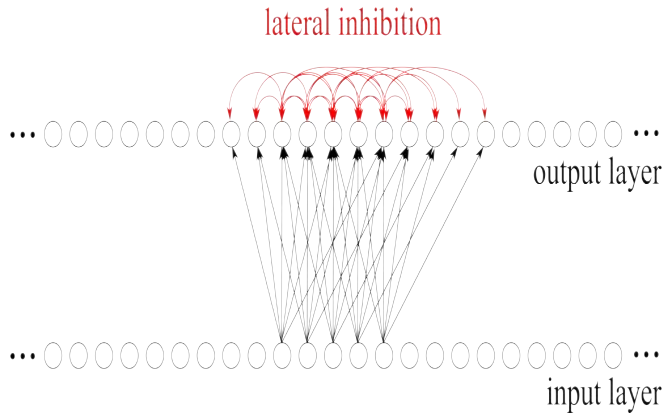
A quick look at recent progress in *understanding how networks of neurons in the brain could implement probabilistic inference*



Neuroanatomy and neurophysiology suggest that cortical microcircuits are composed of stochastic WTA circuits

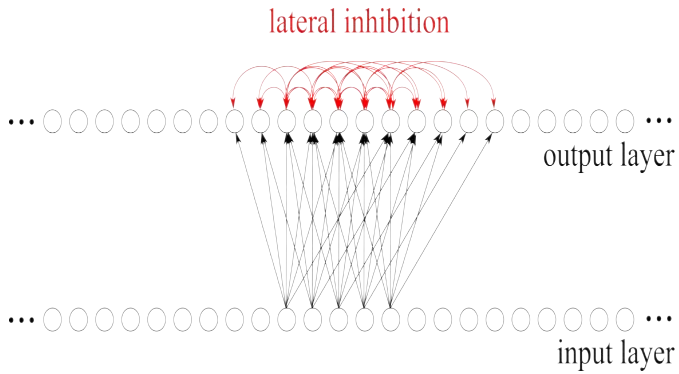


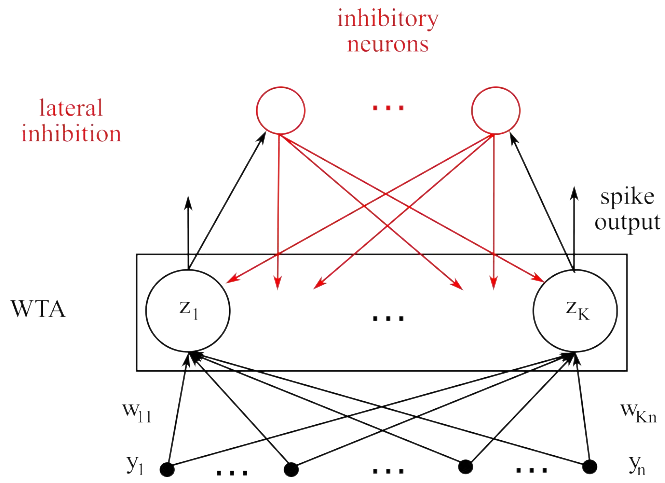
[Douglas and Martin, 2004]:
„canonical microcircuit“ of the cortex

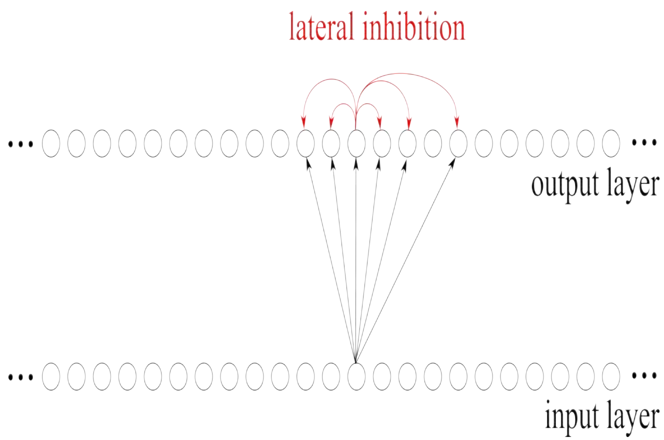


WTA sheet with local lateral inhibition

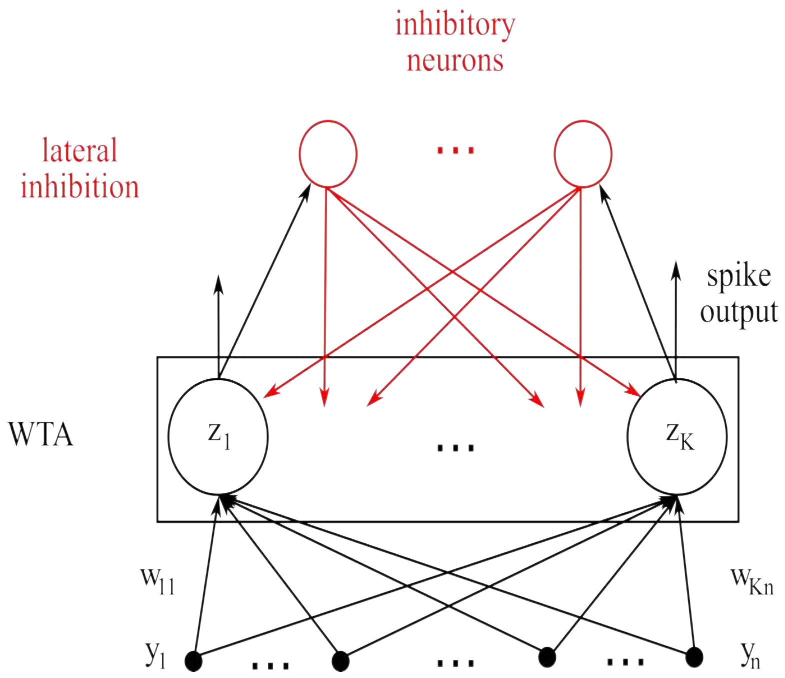
Extraction of a theoretically tractable basic WTA circuit







Model for the spike output of such WTA circuit



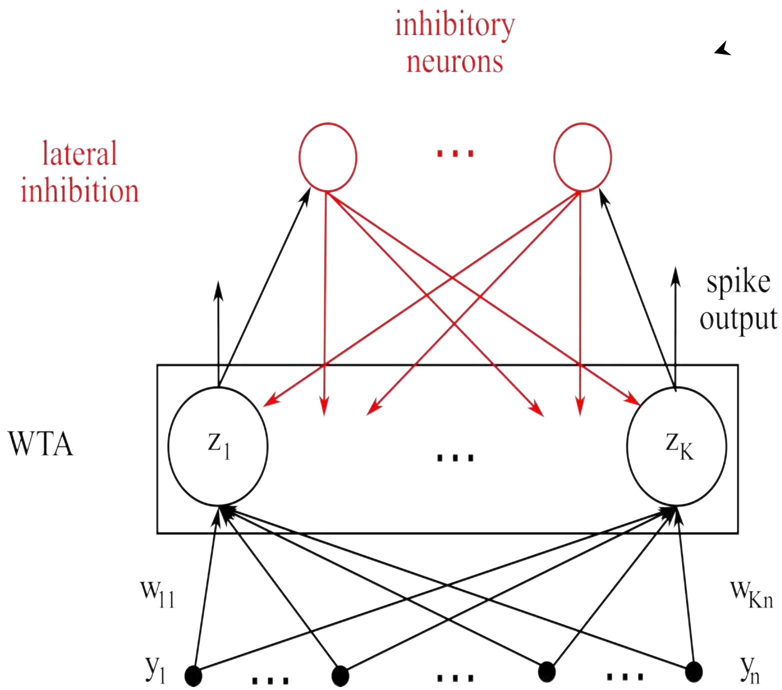
$$p(z_k \text{ fires at time } t|\mathbf{y}) = \frac{e^{u_k(t)}}{\sum_{l=1}^{K} e^{u_l(t)}}$$

Membrane potential of this neuron:

$$u_k(t) = \sum_{i=1}^{n} w_{ki}\tilde{y}_i(t) + w_{k0}$$

This exponential firing rule fits experimental data quite well, as shown by Gerstner et al.

Model for the spike output of such WTA core



$$p(z_k \text{ fires at time } t|\mathbf{y}) = \frac{e^{u_k(t)}}{\sum_{l=1}^{K} e^{u_l(t)}}$$

Membrane potential of this neuron:

$$u_k(t) = \sum_{i=1}^{n} w_{ki}\tilde{y}_i(t) + w_{k0}$$

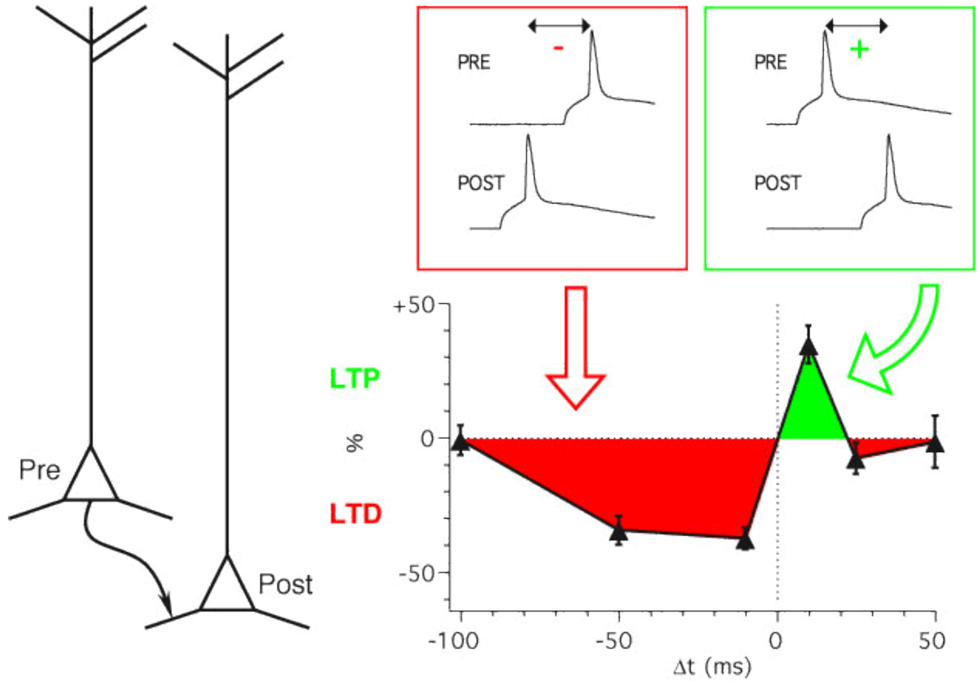
This exponential firing rule fits experimental data quite well, as shown by Gerstner et al.

Bayesian computation emerges in this simple WTA circuit through STDP

[Nessler, Pfeiffer, Maass, NIPS 2009]

STDP (= Spike-Timing-Dependent plasticity) is currently the best understood experimental method for inducing synaptic plasticity

(see work by Grant, Markram, Shulz, Fregnac, Gerstner in FACETS.)



difference in spike times:
POST -PRE

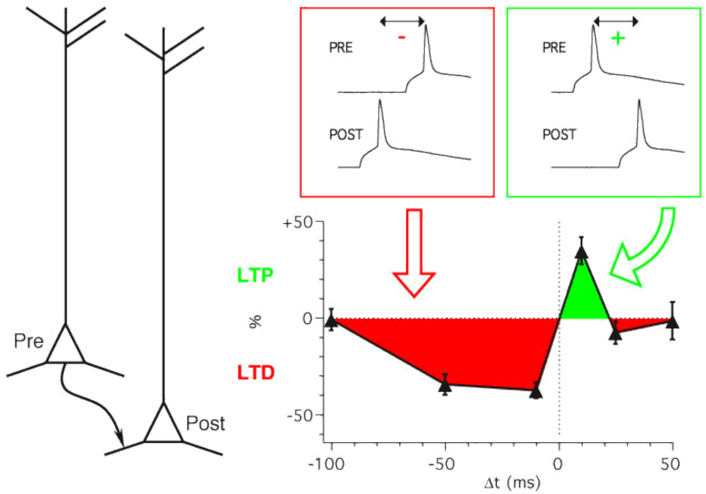
Our theoretical analysis favors the following scaling of weight changes in dependence of the current weight :

$$\Delta w_{ki} = \begin{cases} c \cdot e^{-w_{ki}} - 1, & \text{if the presynaptic neuron } y_i \text{ fires shortly before the postsynaptic neuron } z_k \\ -1, & \text{else,} \end{cases}$$

Mathematical basis for a link between STDP (i.e., the *world of synaptic plasticity*) and probabilistic inference (i.e., the *world of probability theory*):

This STDP rule

$$\Delta w_{ki} = \begin{cases} c \cdot e^{-w_{ki}} - 1, & \text{if the presynaptic neuron } y_i \text{ fires shortly before the postsynaptic neuron } z_k \\ -1, & \text{else,} \end{cases}$$



causes each synaptic weight to converge (in fact, optimally fast) to

log p (PRE has fired just before time t | *POST fires at time t)*

[This relationship to probability theory only becomes visible when one rescales the weights into the negative domain]

Consequence of this link between STDP and probabilistic inference:

The neurons in the WTA circuit *self-organize for creating an internal model* **for the distribution of high-dimensional spike inputs y to the WTA circuit**

- The convergence of each synaptic weight to *log p (PRE has fired just before time t | POST fires at time t)* causes each neuron z_k in the WTA circuit to build in its synaptic weights an internal model $p(\mathbf{y}|z_k \text{ fires}, \mathbf{w})$

  for the typical spike input **y** that makes this neuron fire

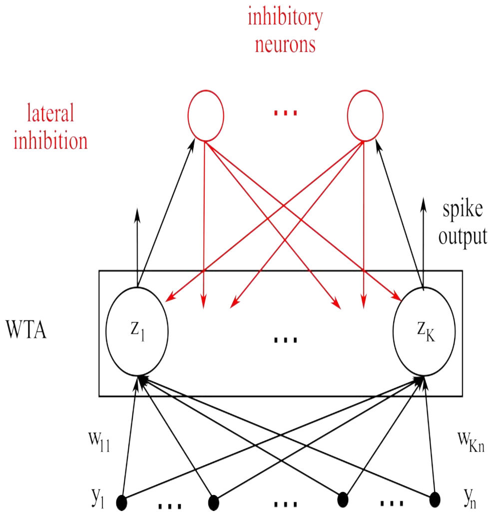
- The competition of the z_k neurons causes the *ensemble of these neurons to create an internal model* $p(\mathbf{y}|\mathbf{w}) = \frac{1}{Z}\sum_{k=1}^{K} e^{u_k(\mathbf{y})}$

  for the distribution p*(**y**) *of spike inputs* **y** *to the circuit*

- This *self-organization process is guaranteed to converge* because it approximates a well known method for fitting an internal model to a given distribution p*(**y**) of inputs: *EM*

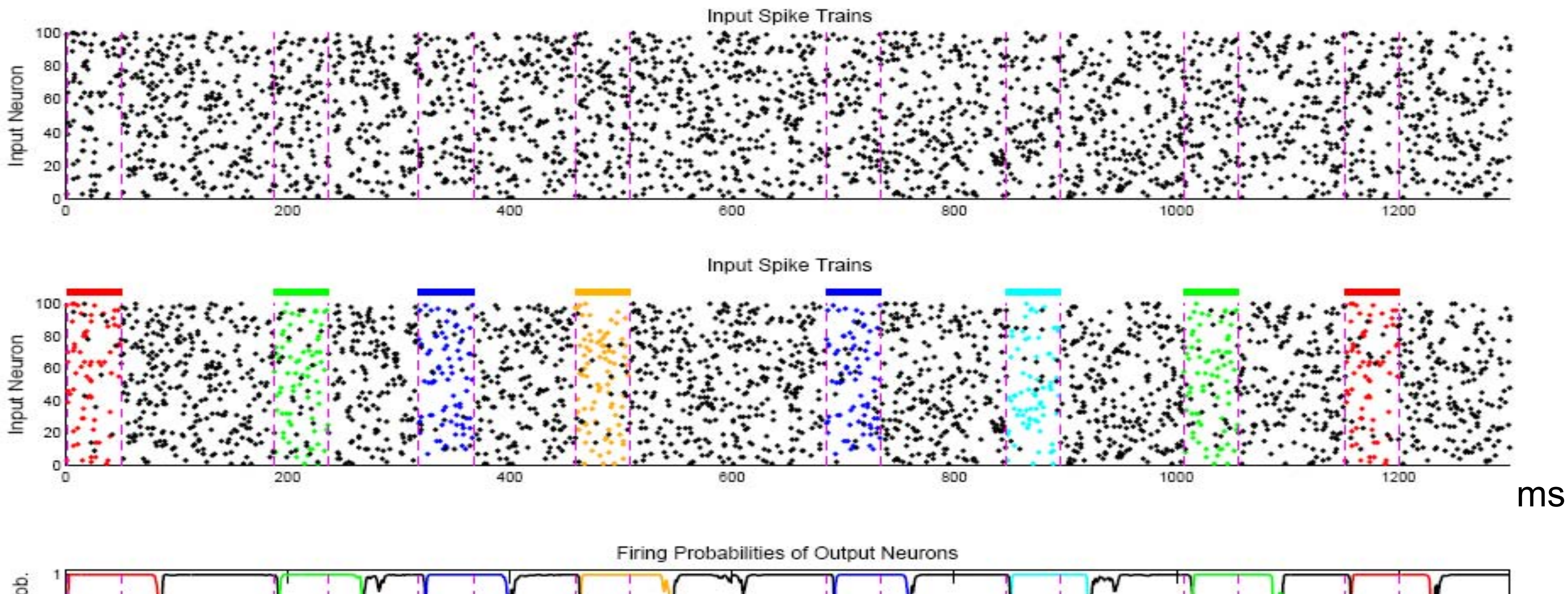
What is EM, and why is it useful ?

EM („Expectation Maximization") solves a *„chicken-and-egg" problem* that always appears when one wants to extract hidden causes („latent variables") from a stream of data without any „supervisor" or teacher:



Each neuron z_k *should learn to fire when a particular one of several potential hidden causes has created the current spike input y.*

But if they do not know what the hidden causes are, they do not know when they should fire.



EM offers a surprising solution to this problem

- One starts with a „random guess“ (as replacement for a teacher) that lets each z-neuron specialize on some arbitrary spike input **y** *(M-step)*

- In the next step this initial random guess is replaced by the stochastic circuit response *(E-step)*. Now only the winner of the competition for responding to the next spike input **y** can adjust its weights via STDP to **y** *(M-step)*

- *Iterate* (practically this simply means that one applies STDP to the weights of the z-neurons in an online manner; no separate E- or M- steps are necessary)

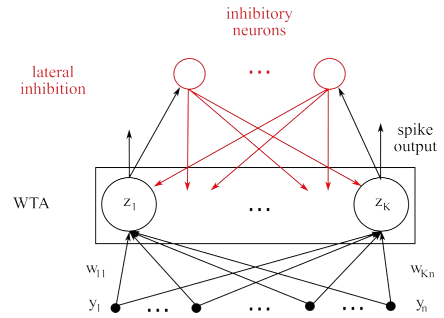
- The *theory of EM* [Dempster et al., 1977] *guarantees*, that these iterations do not lead to a random walk in weight-space, but rather yield *convergence* to a (local) minimum of the KL-distance between the resulting internal model $p(\mathbf{y}|\mathbf{w}) = \frac{1}{Z}\sum_{k=1}^{K} e^{u_k(\mathbf{y})}$ and the external

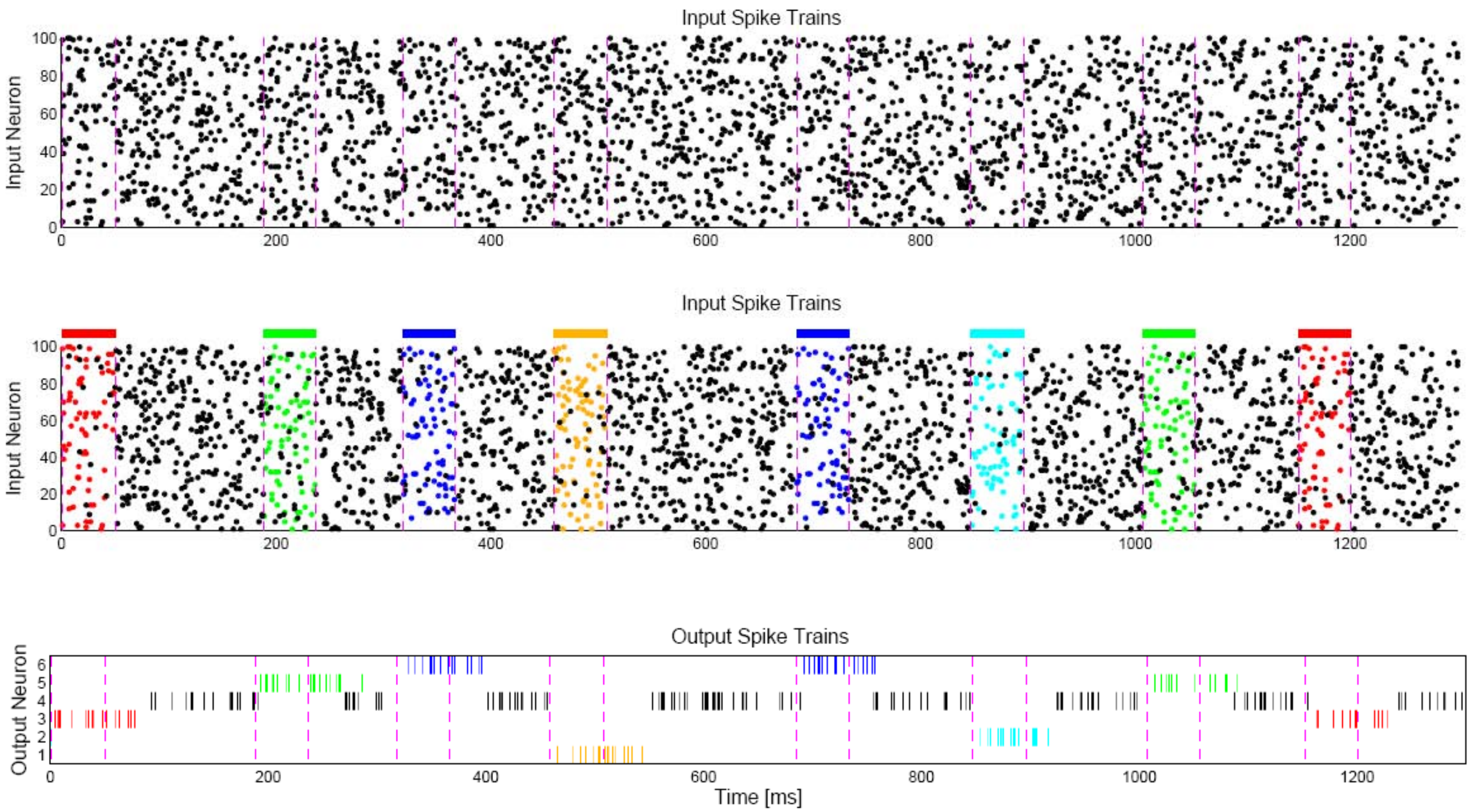
  distribution p*(**y**) of inputs **y**

  (more precisely, the circuit executes an approximation to online stochastic EM)

Result after running STDP in the WTA circuit for a 20 s stream of such spike inputs

(shown for test inputs that had never occurred previously)

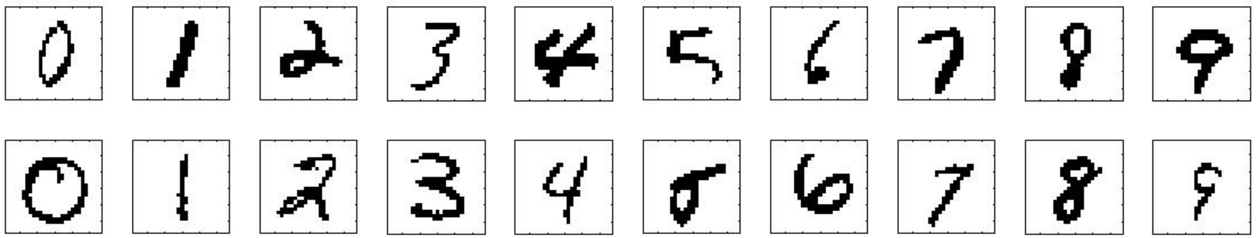


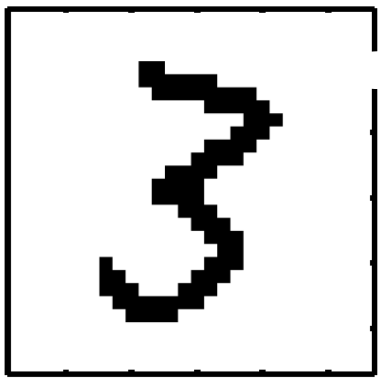


Application of STDP in the same circuit (but with different numbers of z-neurons) to two completely different problems

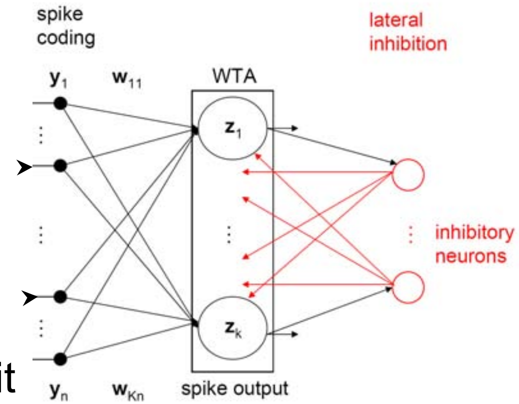
1. Handwritten digits MNIST dataset (WITHOUT supervision):

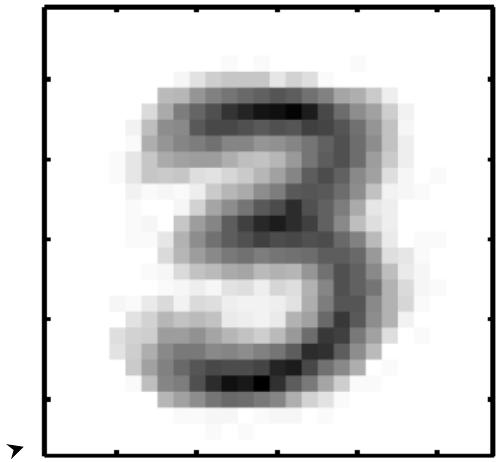
These are 20 random samples from the 70 000 samples in the MNIST dataset.





50ms per digit

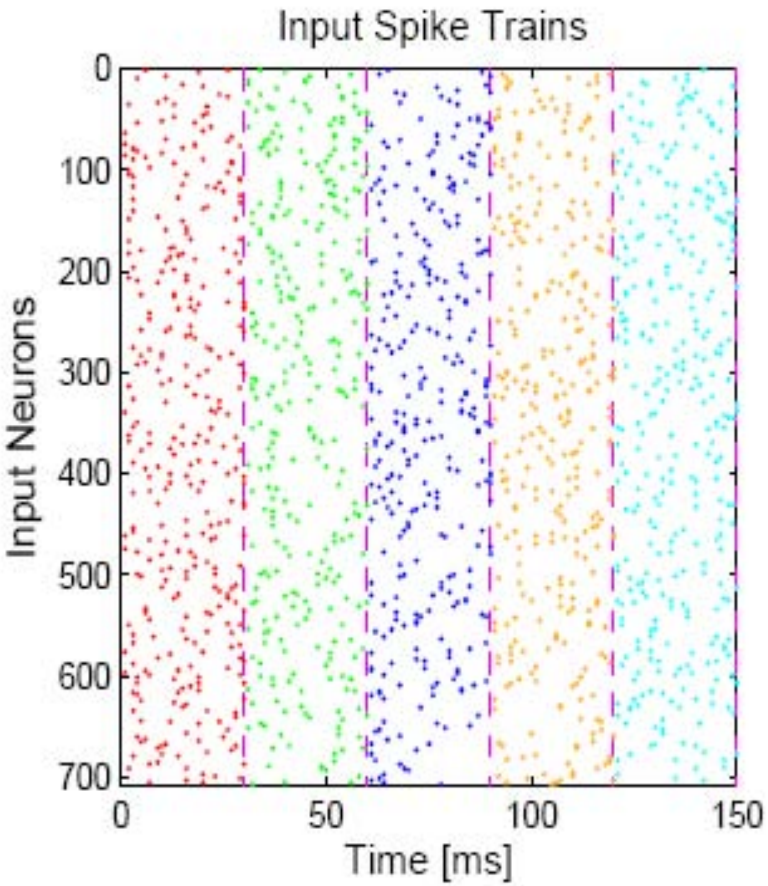




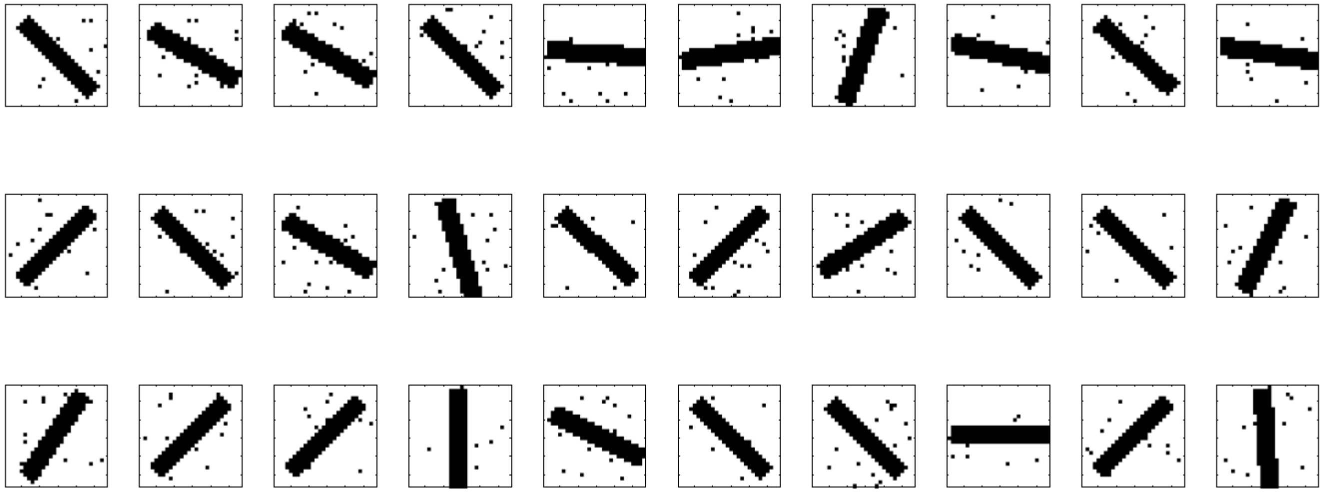
weights of one neuron

Resulting implicit generative models of 100 z-neurons $p(\mathbf{y}|z_k \text{ fires}, \mathbf{w})$

after exposing the circuit to 300 s of spike inputs, where a different sample of a handwritten digit is encoded in each window of 30 ms

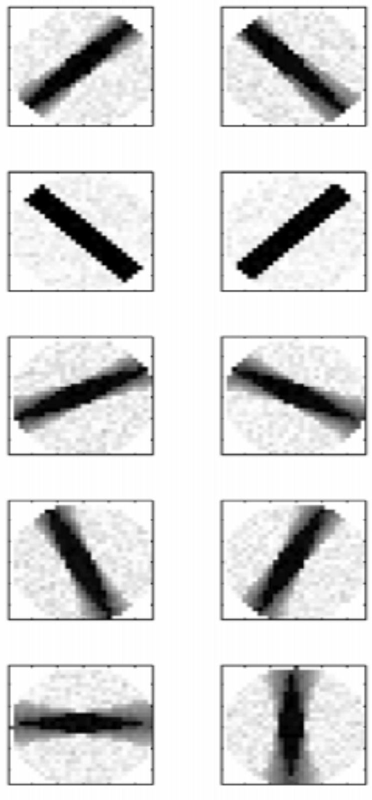


Emergence of orientation selective neurons



30 samples from an „infinite“ set of randomly generated directions
each encoded by spike trains **y** for 50 ms

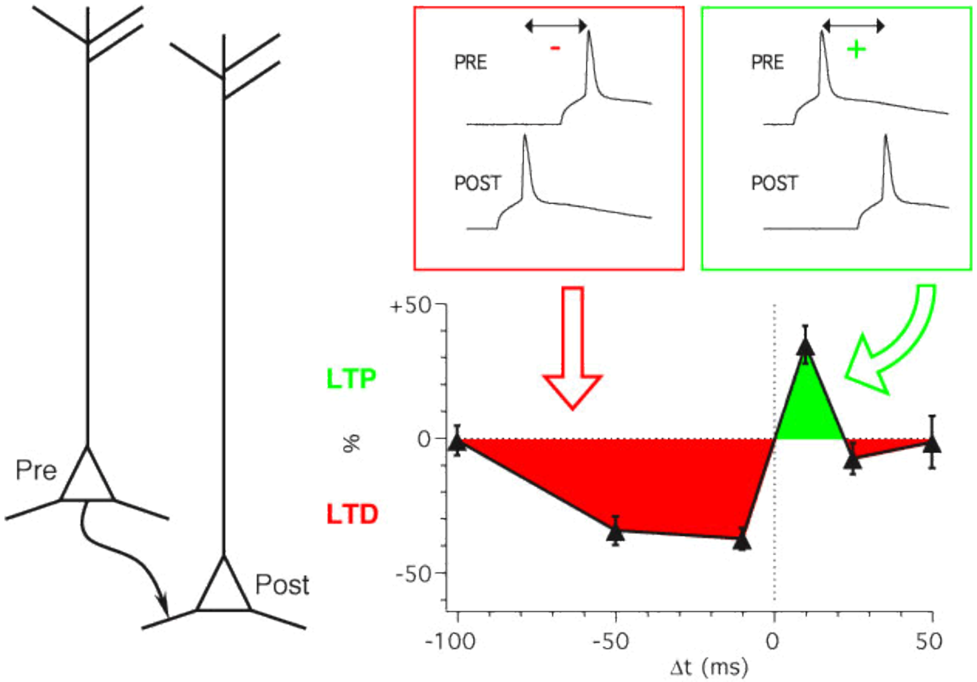
Resulting internal models of the 10 z-neurons after applying STDP for 200 s



Note that these 10 z-neurons have self-organized to cover the continuous space of arbitrary orientations
(note that there are no „clusters“ of directions in the input stream)

Our theoretical analysis relied on a particular dependence of weight updates on the current weight in STDP:

Is this consistent with biological data ?



difference in spike times:
POST -PRE

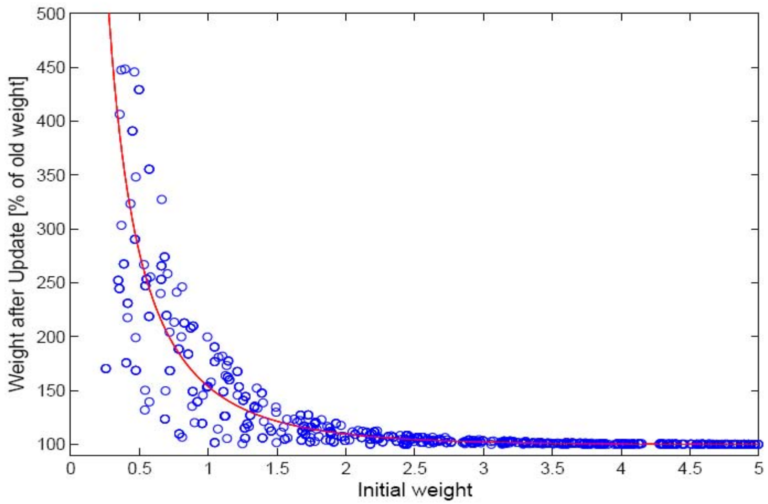
Our theoretical analysis favors the following scaling of weight changes in dependence of the current weight :

$$\Delta w_{ki} = \begin{cases} c \cdot e^{-w_{ki}} - 1, & \text{if the presynaptic neuron } y_i \text{ fires shortly before the postsynaptic neuron } z_k \\ -1, & \text{else,} \end{cases}$$

Our theoretical analysis favors the following two rules

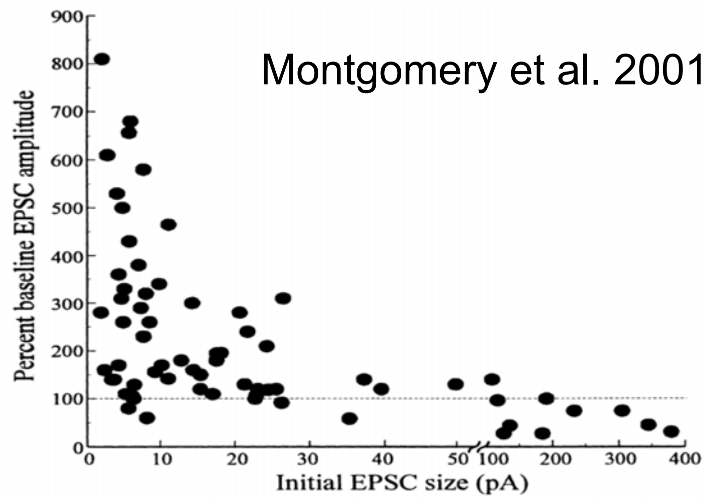
1. „Weight increases become exponentially smaller in dependence of the current weight size“

Theoretical prediction



Noise does not harm the effictiveness of the STDP rule in our model

Experimental data



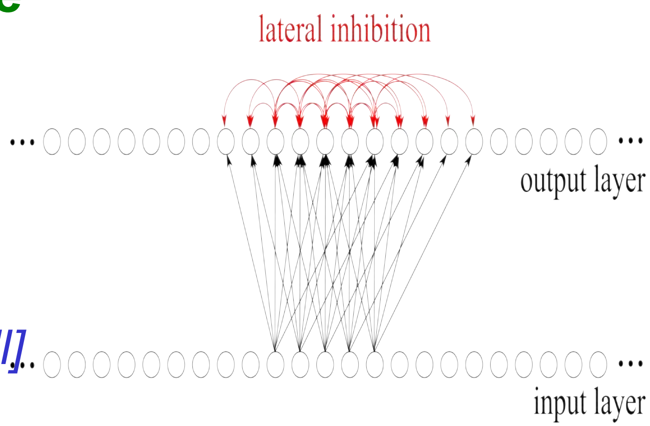
See similar data by [Liao et al., 1992], [Bi and Poo, 1998], [Sjöström et al., 2001]

2. „Weight decreases are independent of the current weight size“

Experiments by Dan Shulz (FACETS-partner CNRS) with STDP in-vivo show that weight decreases of STDP are not correlated with the current weight size

This is just the beginning of understanding the anatomy and dynamics of networks of neurons from the perspective of probabilistic inference

Work in progress:



- More expressive internal models by STDP in biologically more realistic sheets of neurons with local lateral inhibition (allowing several „winners") *[Lars Büsing, J. Bill]*

- A new perspective of neuronal dynamics from the perspective of sampling in Bayesian networks (MCMC) *[Lars Büsing]*

- Learning Gaussians and other expontential family distributions as internal models through STDP *[Stefan Habenschuss]*

- Emergence of HMM-capabilities when lateral excitatory connections are added *[David Kappel, Bernhard Nessler]*

- Emulation of probabilistic inference in arbitrary Bayesian networks via Gibbs sampling *[Dejan Pecevski]*

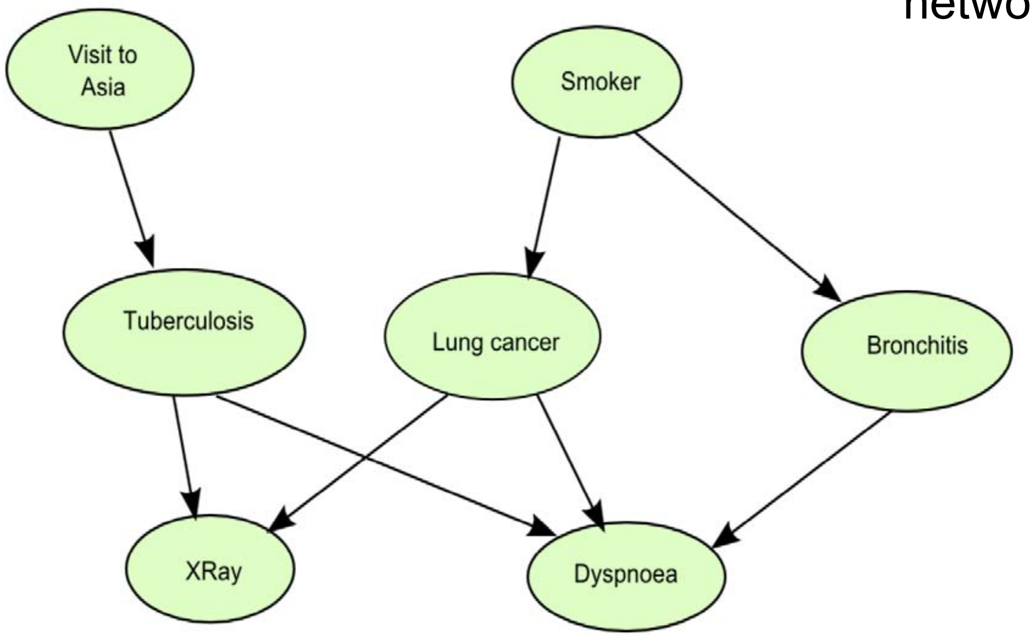
Emulation of probabilistic inference in arbitrary Bayesian networks via Gibbs sampling (and learning via STDP)

Bayesian network: nodes = random variables, arrows = direct causes.

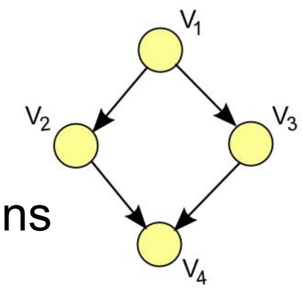
It specifies a joint distribution over these random variables if one adds conditional probability tables for each node (conditioned on parents); these are stored in the weights of our neural circuit model.

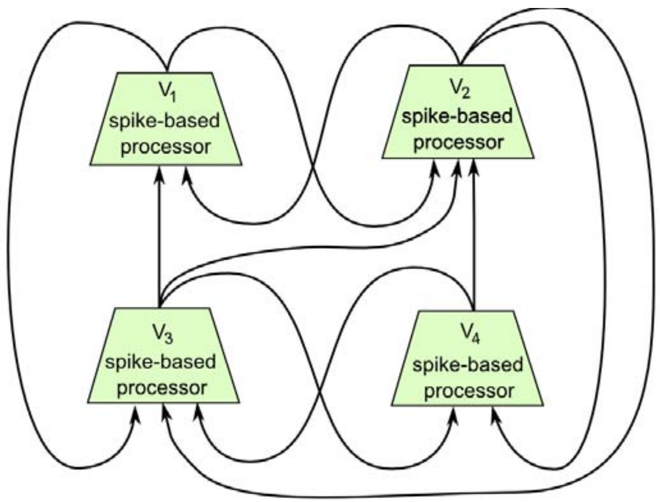
Difficulty: *Information also flows in opposite direction of arrows*, therefore inference is *NP-hard.*

ASIA-network



Transformation to a network of spiking neurons

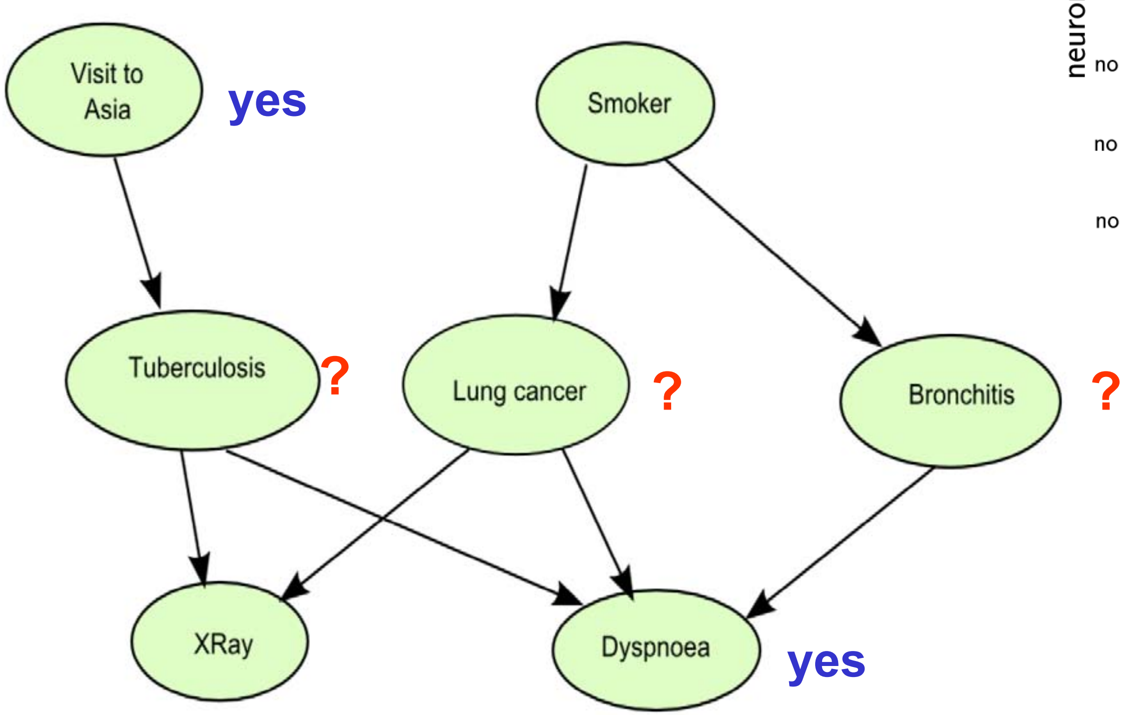




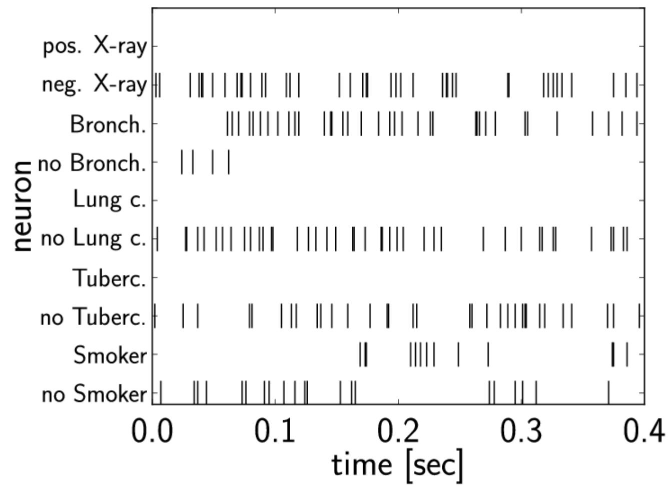
Learning from examples via STDP takes 10s

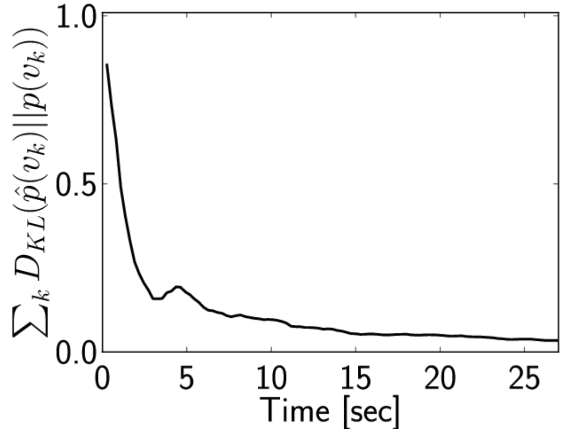
We have implemented inference in this Bayesian network through stochastic firing of spiking neurons that emulates *Gibbs sampling* (after learning through STDP)

Example for a probabilistic inference task:



Neuronal ouputs during sampling





Convergence to target probabilities

Summary

- I have sketched the beginning of a new theory of neural computation.
- Theory predicts that generic moduls for probabilistic computation emerge through STDP in stochastic WTA circuits
- The convergence of the self-organization of neurons is guaranteed here by EM
- Gibbs sampling through stochastic spiking of neurons provides an attractive method for probabilistic inference in neural emulations of Bayesian networks
- These models suggest that spontaneous firing of neurons and trial-to-trial variability of cortical neurons are essential aspects of their computational function
- These models pave the way for an efficient implementation of probabilistic computation in spike-based hardware
- These models pave the way for using stochastic computing elements (e.g. nanoscale switches) as suitable moduls for probabilistic computation

Some of the new talents that have emerged through partial FACETS(-ITN) support in Graz

Bernhard Nessler

Michael Pfeiffer

Lars Büsing

Stefan Habenschuss

Dejan Pecevski

David Kappel

Johannes Bill