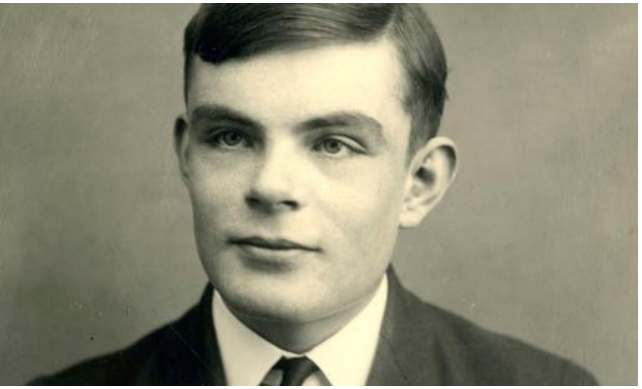
More than Moore with Neuromorphic Computing Architectures

Colloquium GSI Darmstadt
December 19, 2017

Karlheinz Meier
Ruprecht-Karls-Universität Heidelberg

meierk@kip.uni-heidelberg.de
@brainscales

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



„We may compare a man in the process of computing a real number with a machine which is only capable of a finite number of conditions“

On computable numbers, with an application to the Entscheidungsproblem
published 1937 in *Proceedings of the London Mathematical Society*

The Turing Machine

Modelled after a human executing a set of computational instructions:

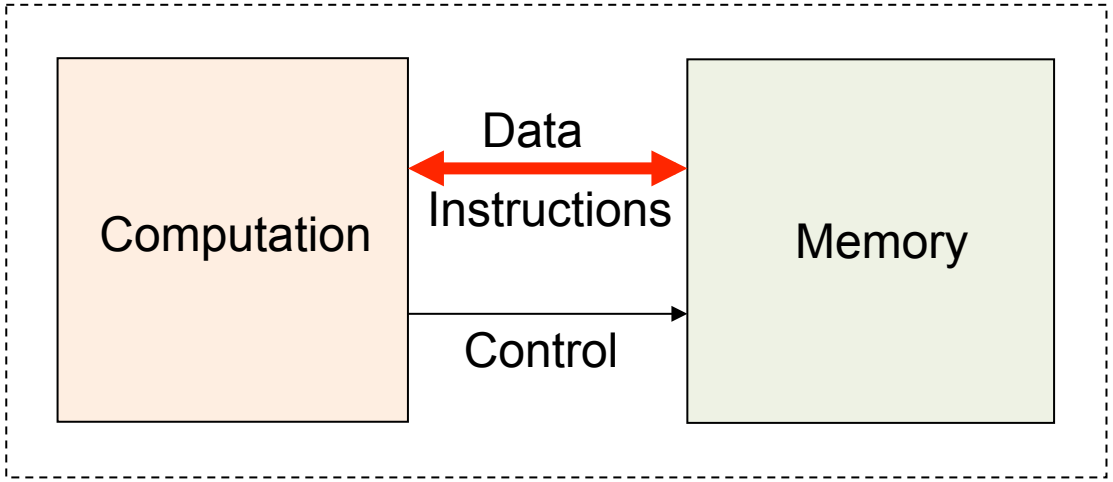
1. Limited size of internal storage
   → finite number of states
2. Infinite size of external storage but only a fraction can be used at a given time t

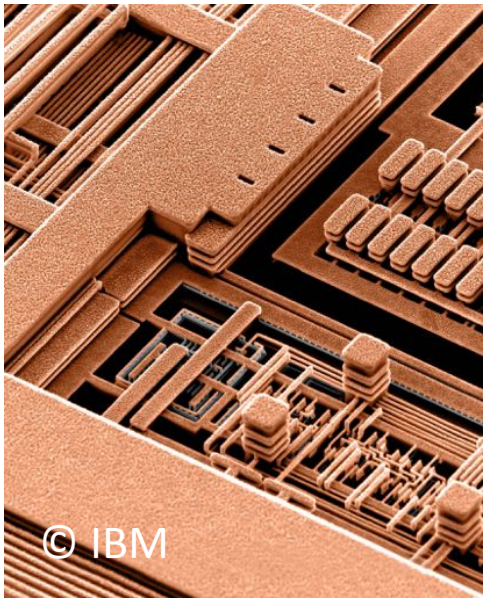
1. + 2. uniquely determine the state of the machine at time t+1

Each Turing machine uniquely represents an algorithm

Realizing the Turing Machine
The von Neumann Architecture

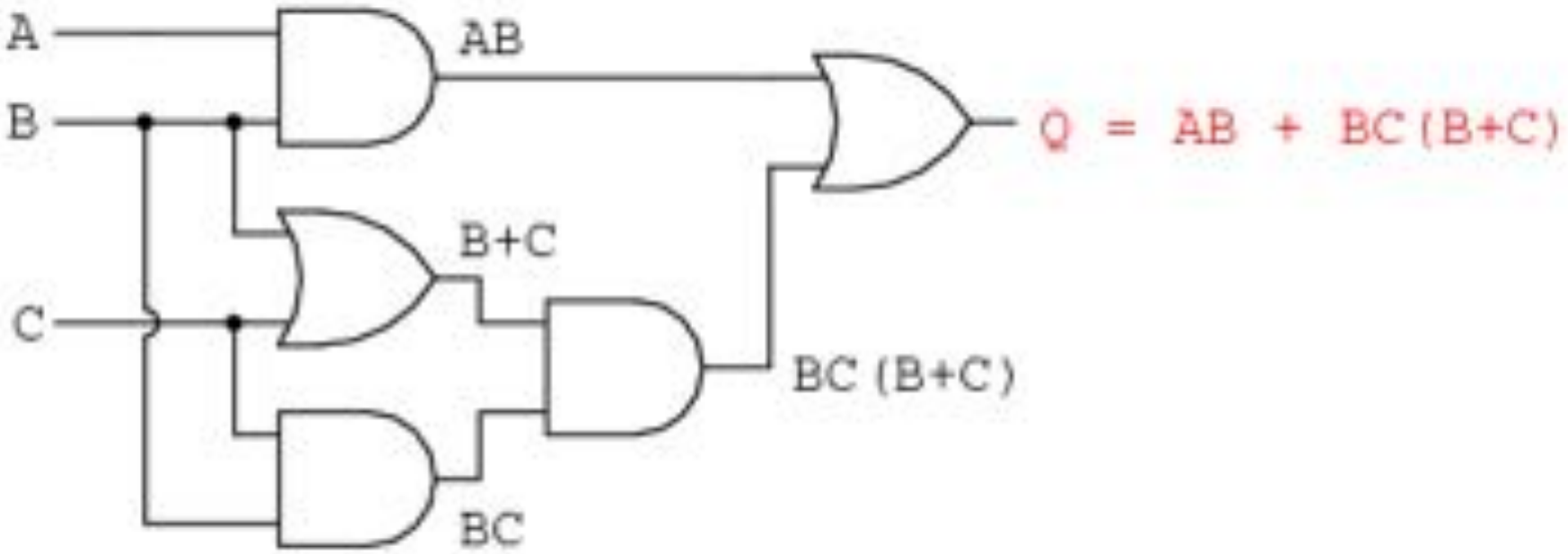
- Data and instructions stored in memory
- Content of memory addressable by location
- Instructions executed sequentially unless order is explicitly modified
- Memory and Computation physically separated

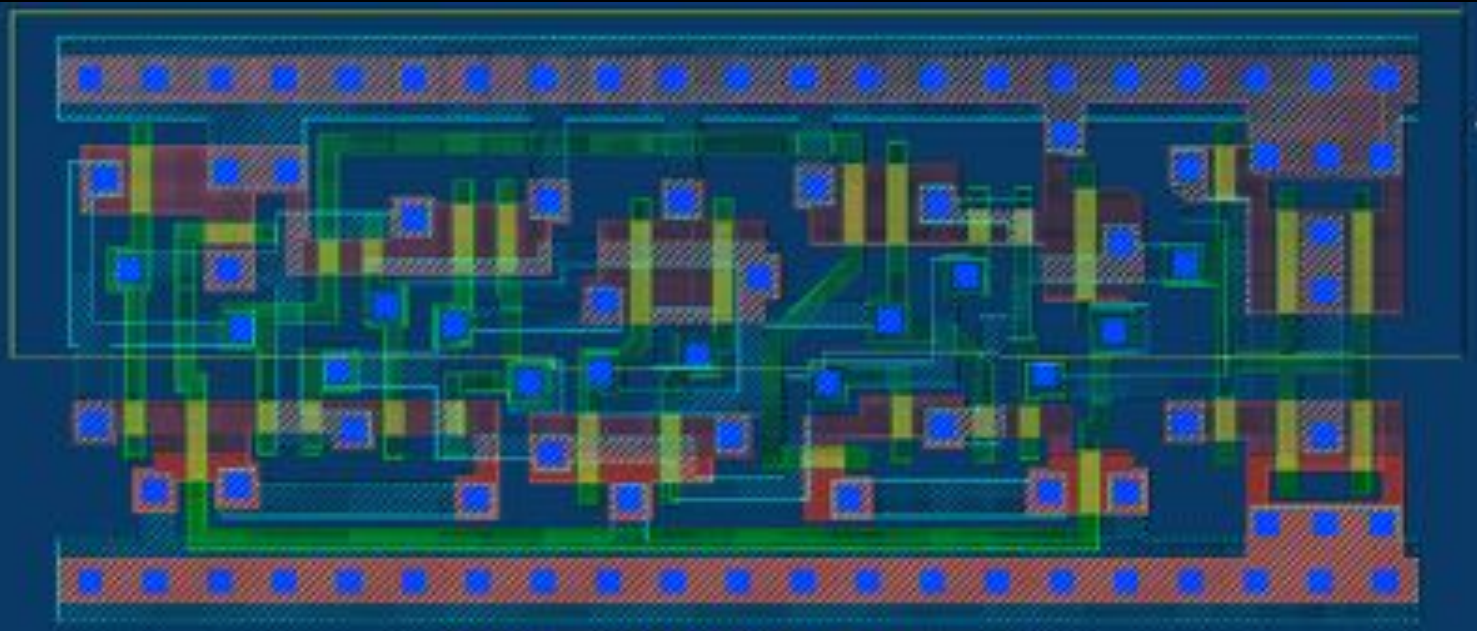


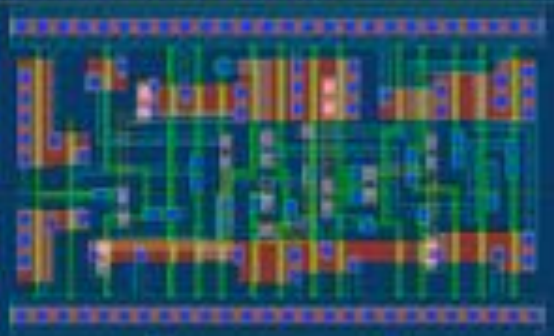


Realizing the von Neumann Architecture

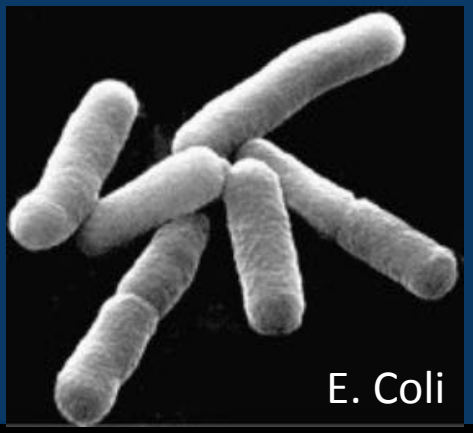
Claude Shannon : From algebra to logic circuits

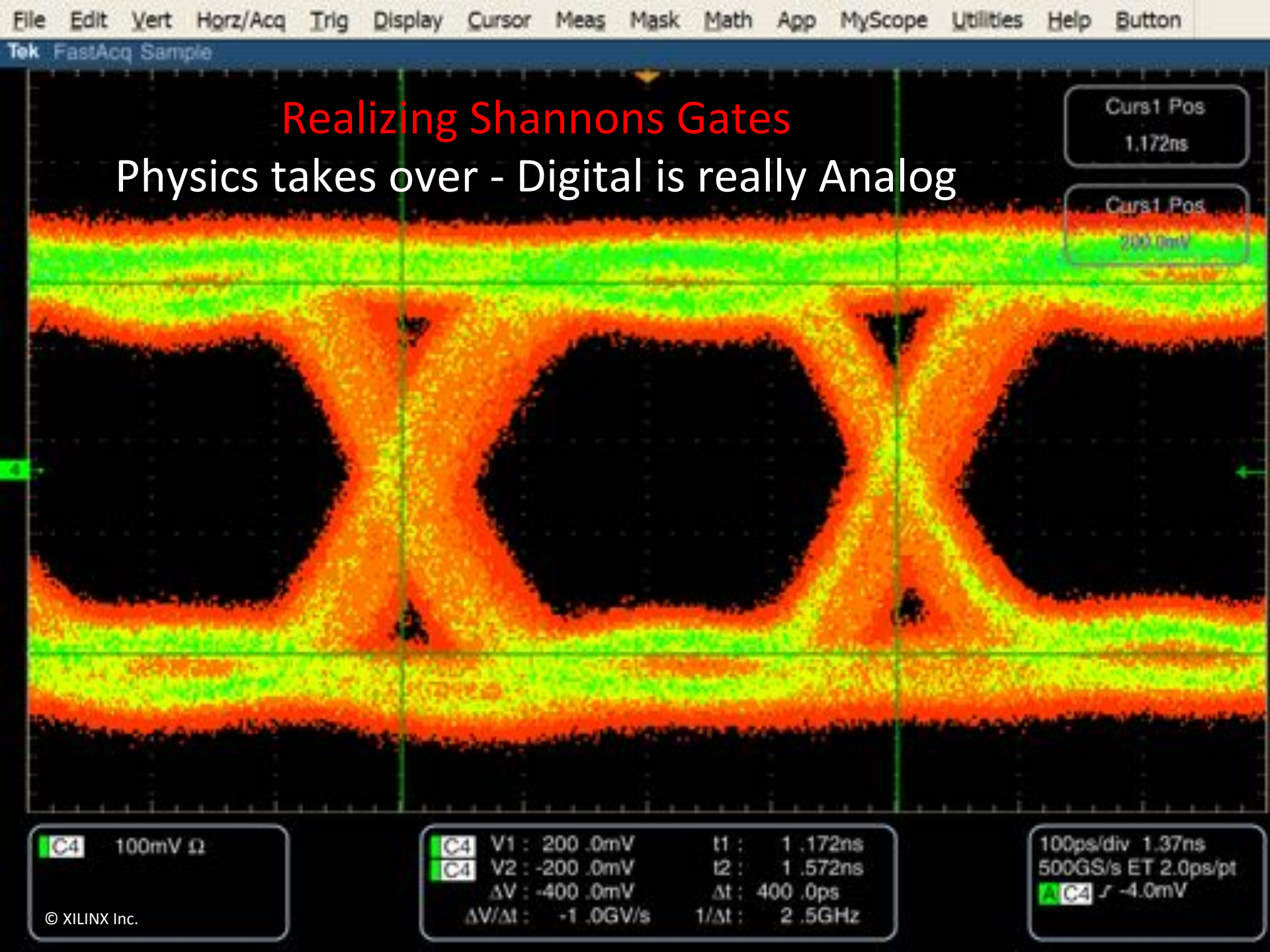






D-Flip-Flop
Drawn to Scale
1000 nm




File Edit Vert Horz/Acq Trig Display Cursor Meas Mask Math App MyScope Utilities Help Button
Tek FastAcq Sample
Realizing Shannons Gates
Physics takes over - Digital is really Analog
Curs1 Pos
1.172ns
Curs1 Pos
C4 100mV Ω
C4 V1 : 200 .0mV t1 : 1 .172ns
C4 V2 : -200 .0mV t2 : 1 .572ns
ΔV : -400 .0mV Δt : 400 .0ps
ΔV/Δt : -1 .0GV/s 1/Δt : 2 .5GHz
100ps/div 1.37ns
500GS/s ET 2.0ps/pt
A C4 -4.0mV
© XILINX Inc.

Limits of Technology Scaling

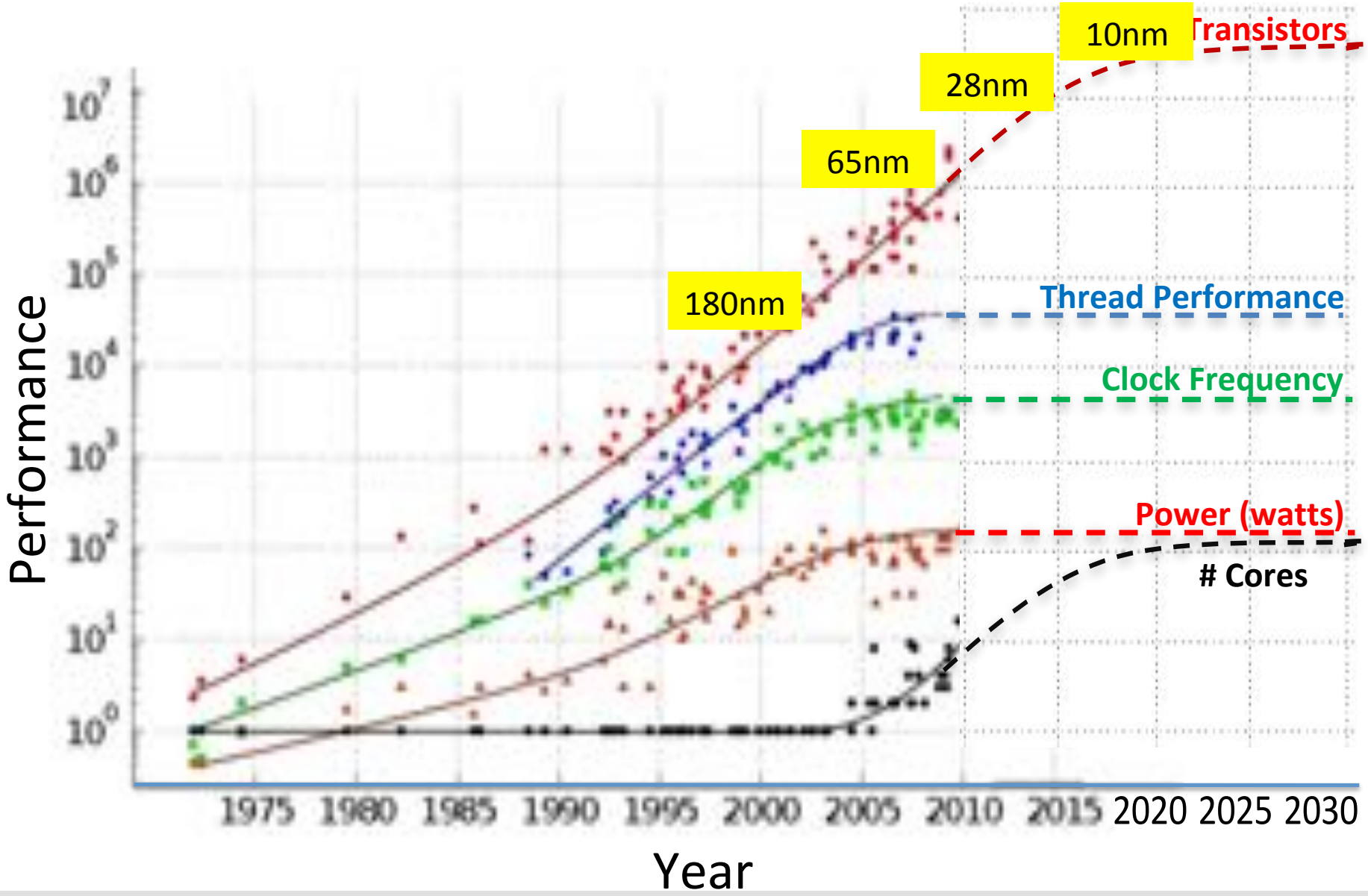
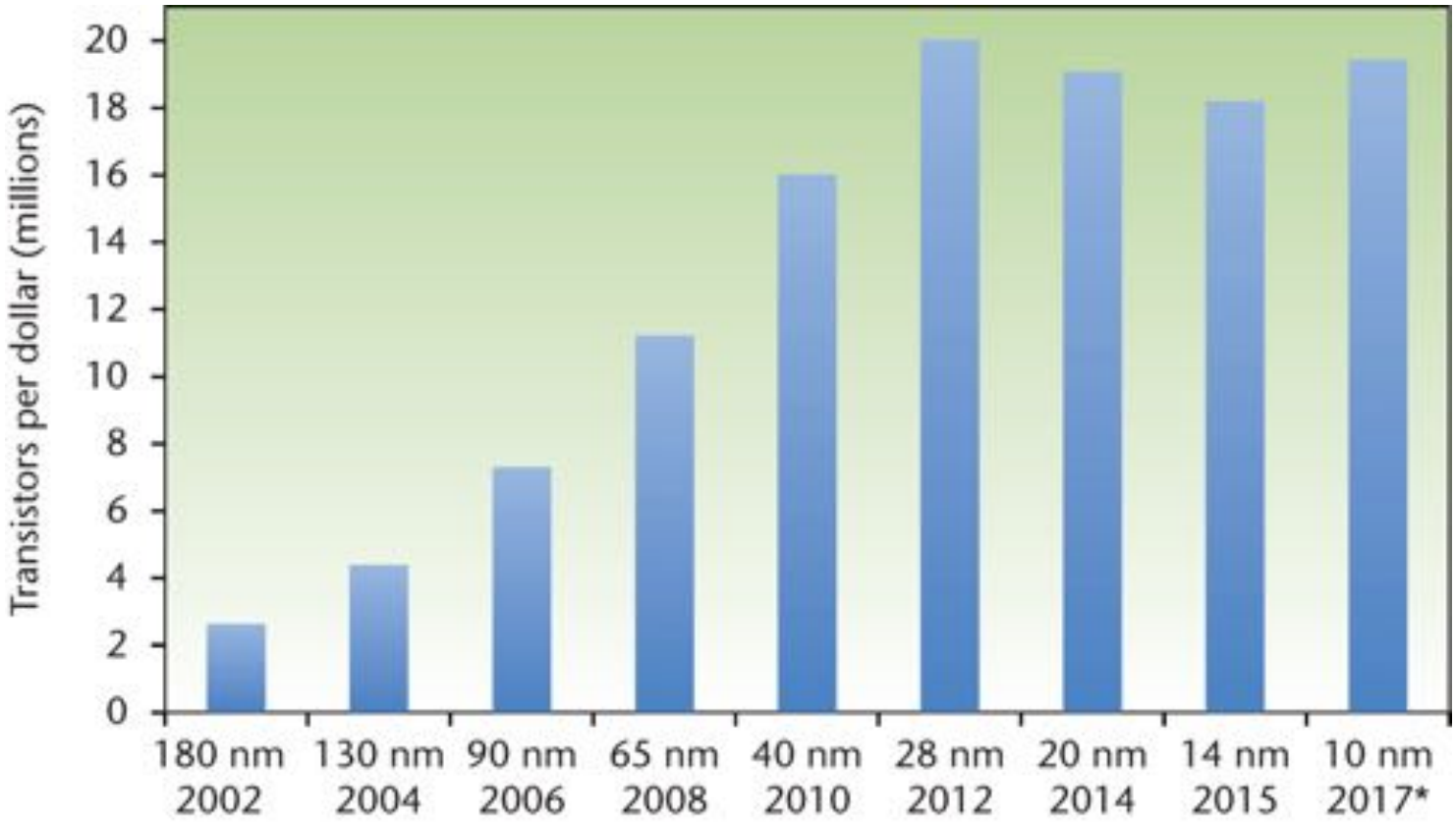


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith



No more progress from smaller transistors

New ARCHITECTURES suddenly interesting !

Non-von Neumann

Non-Turing

The Brain – *Extreme Matter*

10^{15} connections (synapses)

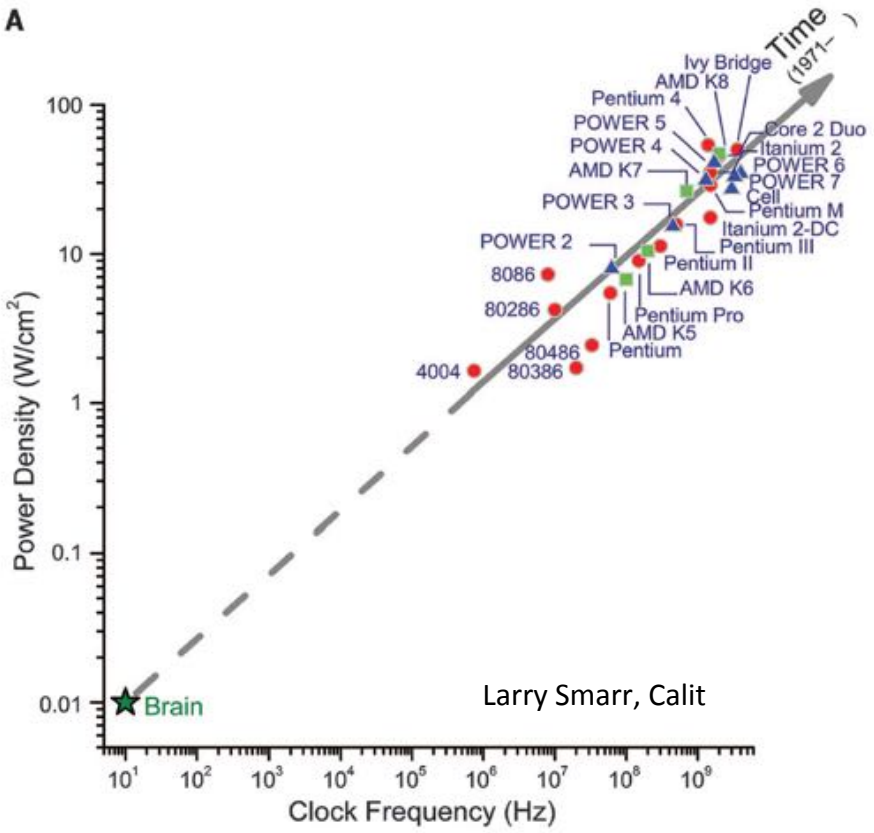
10^{11} nodes (neurons)

Open system driven by external world

Dynamic long-range and short-range interactions

Timescales from milliseconds to years

Stochastic on the microscopic level

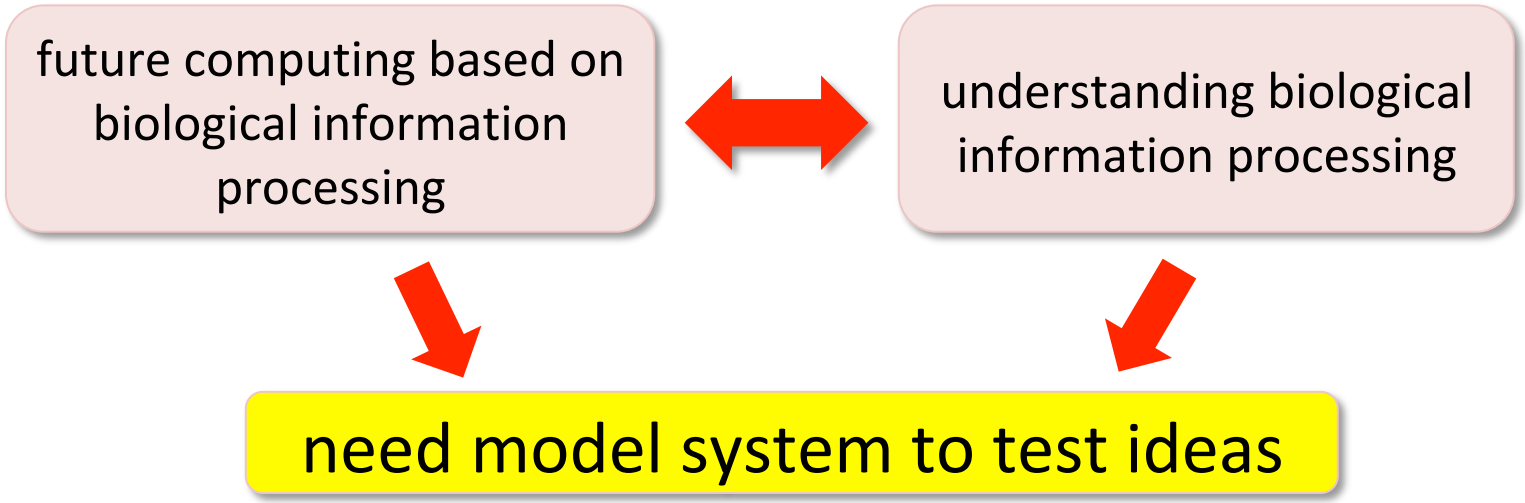


Assets of brain computing

- Energy efficiency
- Compactness
- Fault tolerance
- Speed
- Configuration and learning replace programming
- Scalability

Conventional computing is moving away from the brain

Why brain inspired computing ?



Two **fundamentally different** modeling approaches:

- **NUMERICAL MODEL (Turing)**

  represents model parameters as binary numbers

- **PHYSICAL MODEL (not Turing)**

  represents model parameters as physical quantities

  → **voltage, current, charge** (like the biological brain)

can be combined to form a hybrid system

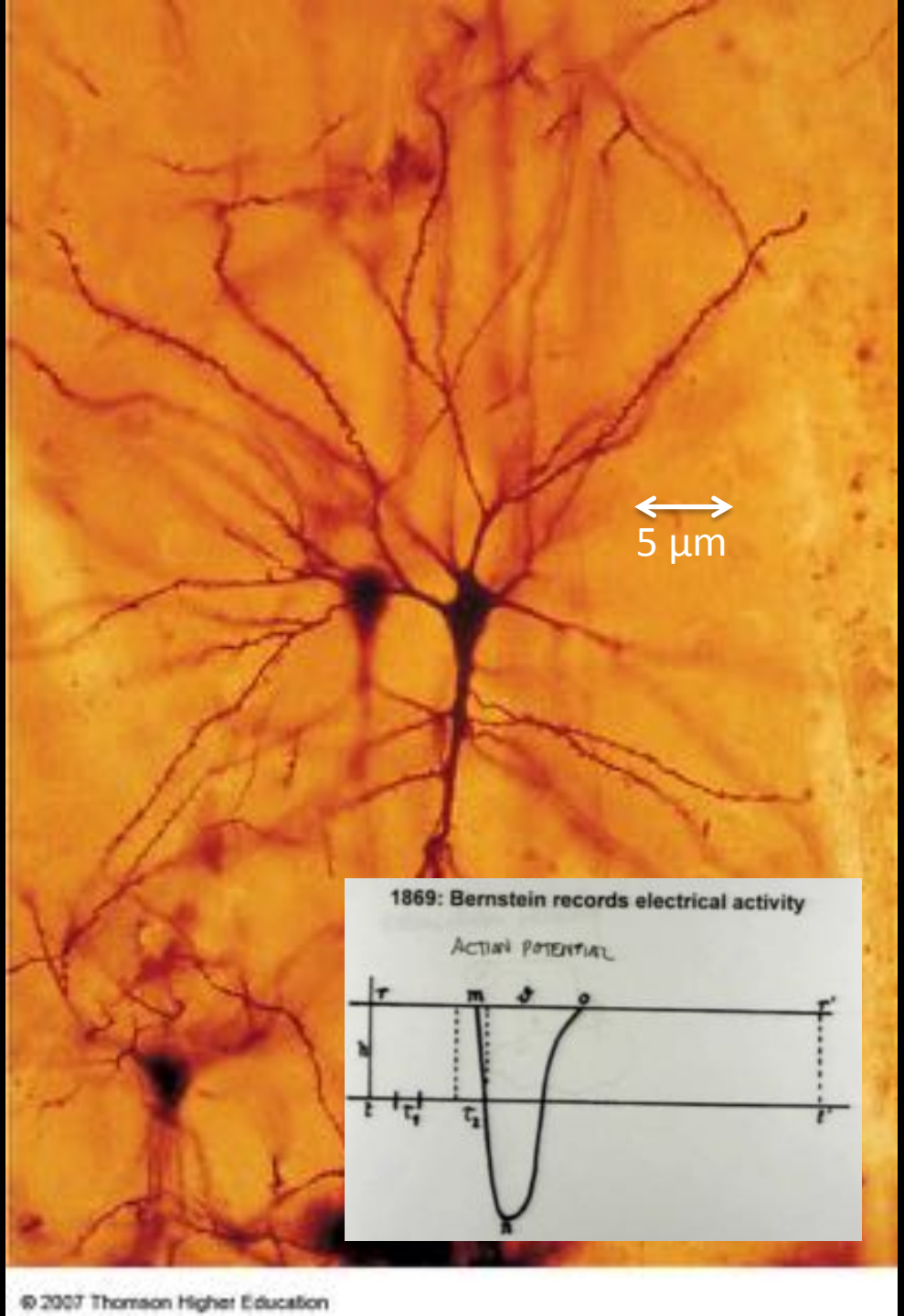
Herrmann v. Helmholtz (1821-1894)
Julius Bernstein (1839-1917)
Santiago Ramón y Cajal (1852-1934)

Individual cells in the brain are spatially separated constituents

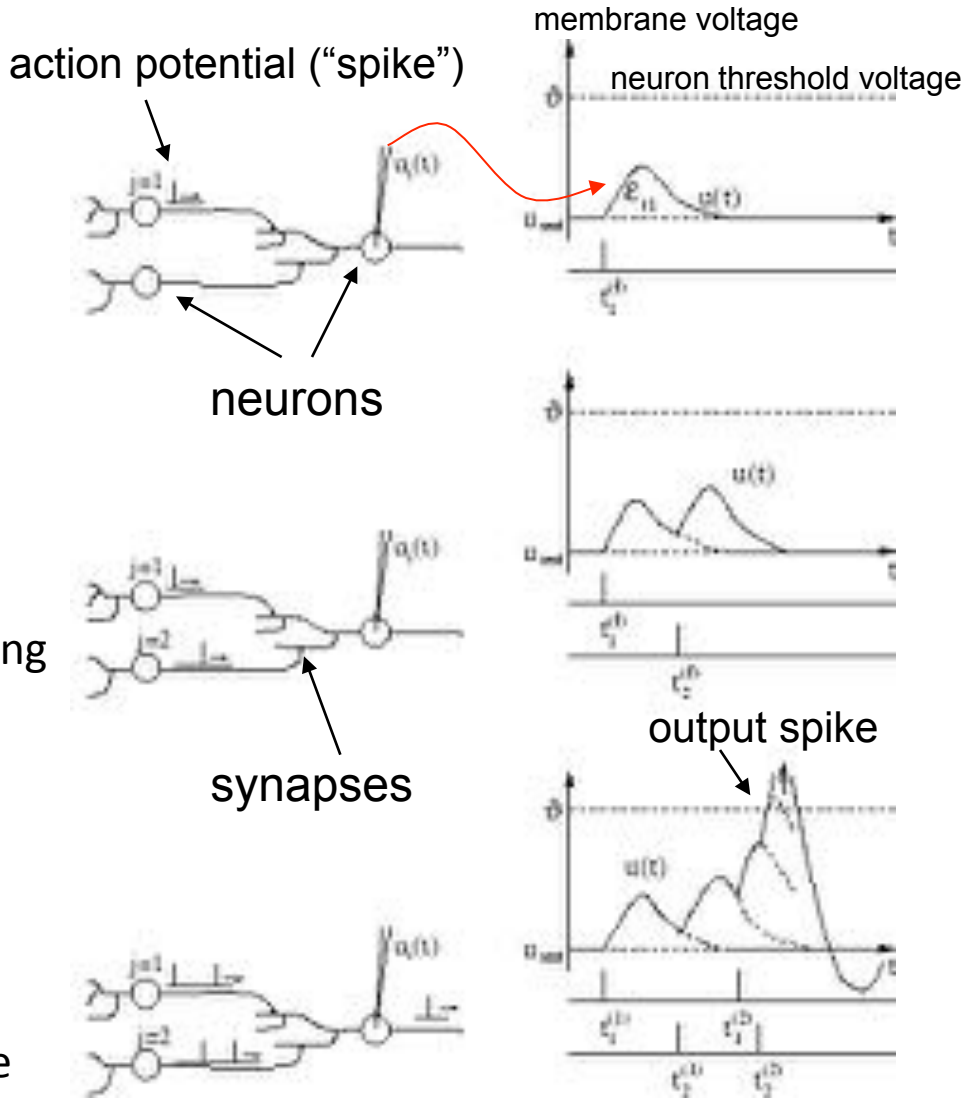
"interaction over a distance"

and

"spatial and temporal integration"

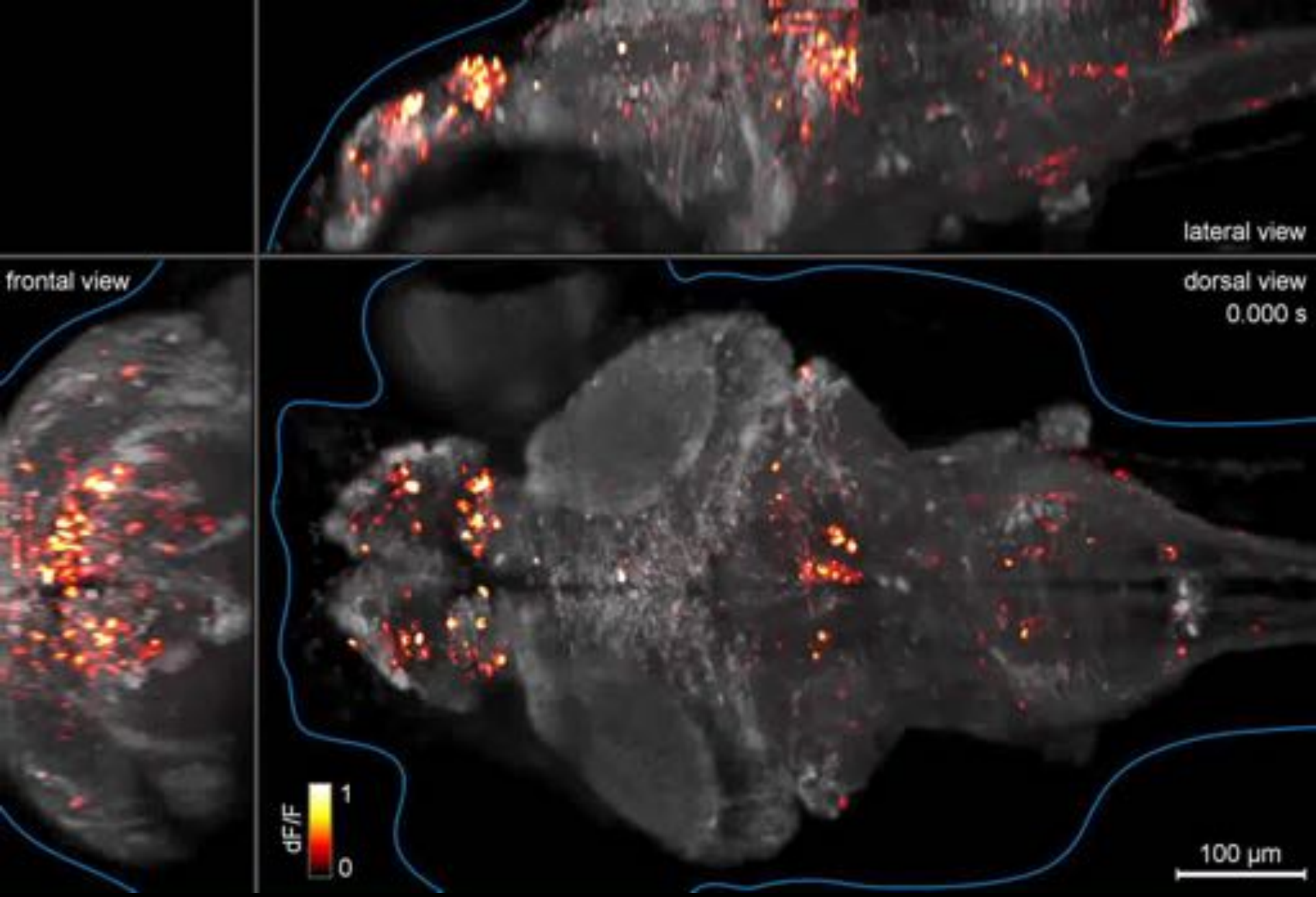


Basics of neural communication



- Threshold for non-linear response leading to all-or-none law (spikes)
- neurons integrate over space and time-characteristic time constants
- temporal correlation is important
- mixed-signal system:

  action potential ↔ membrane voltage

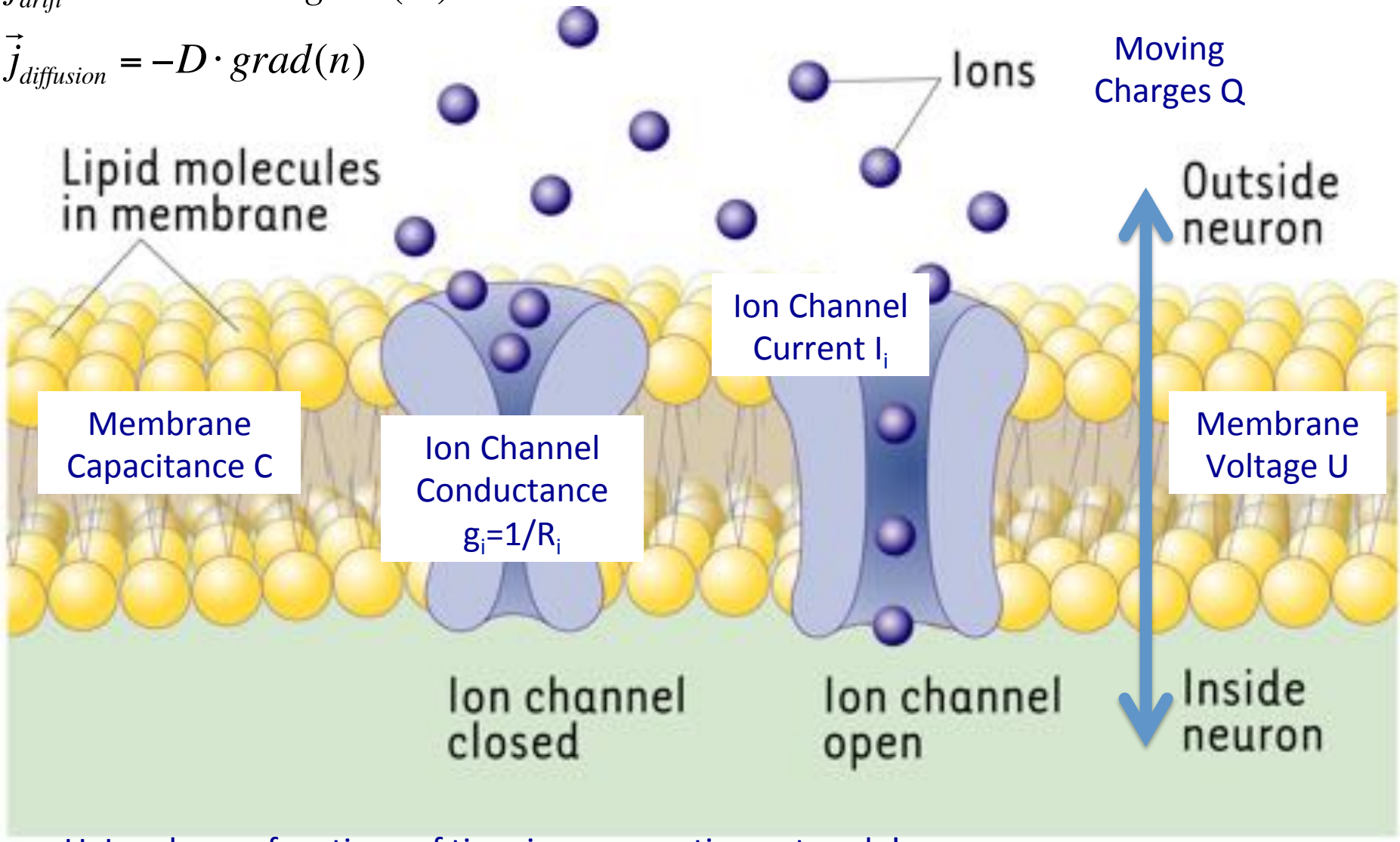


Ahrens et al, *Nature Methods* **10**, 413–420 (2013)

Some Electrical Quantities of a real Neuron Membrane

$$\vec{j}_{drift} = \sigma \cdot \vec{E} = -\sigma \cdot grad(\Phi)$$

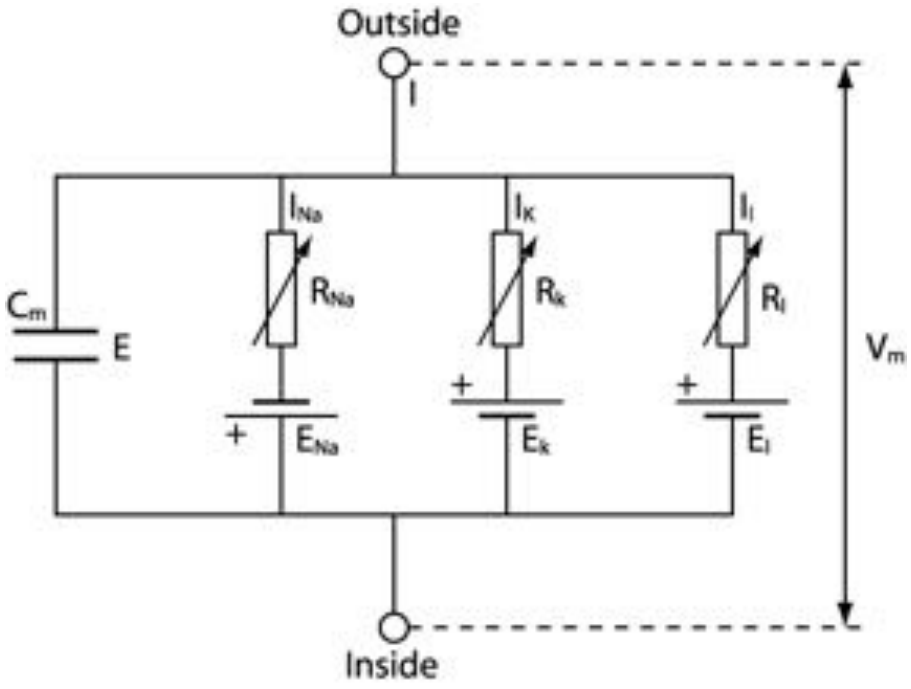
$$\vec{j}_{diffusion} = -D \cdot grad(n)$$

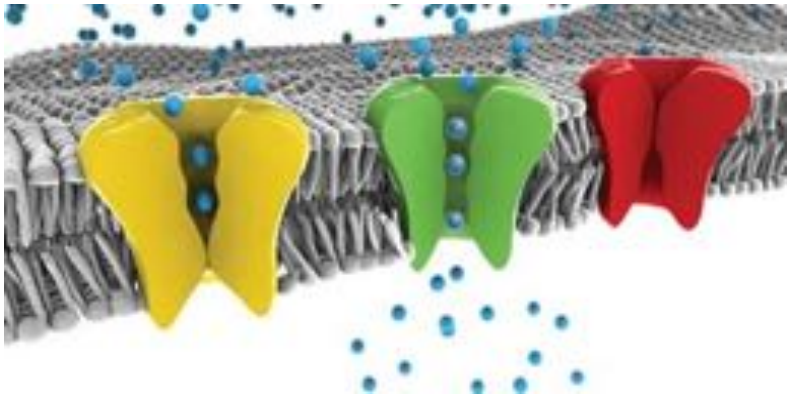


U, I and g are functions of time in an operating network !

Current theories and modelling are treating these quantities only (few exceptions)

Hodgkin-Huxley 1952
Describing the non-linearity



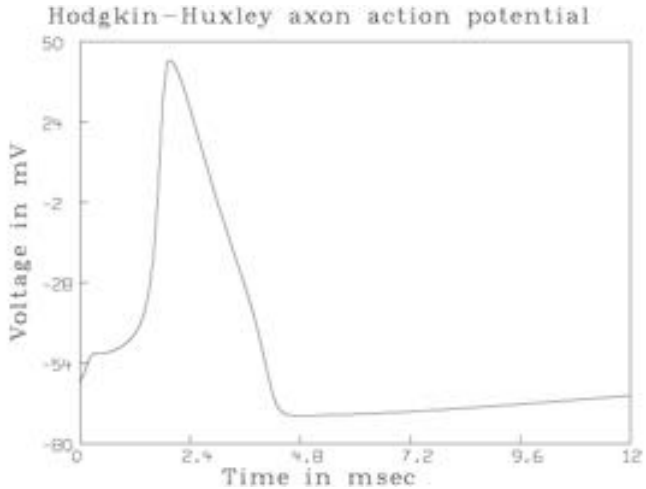


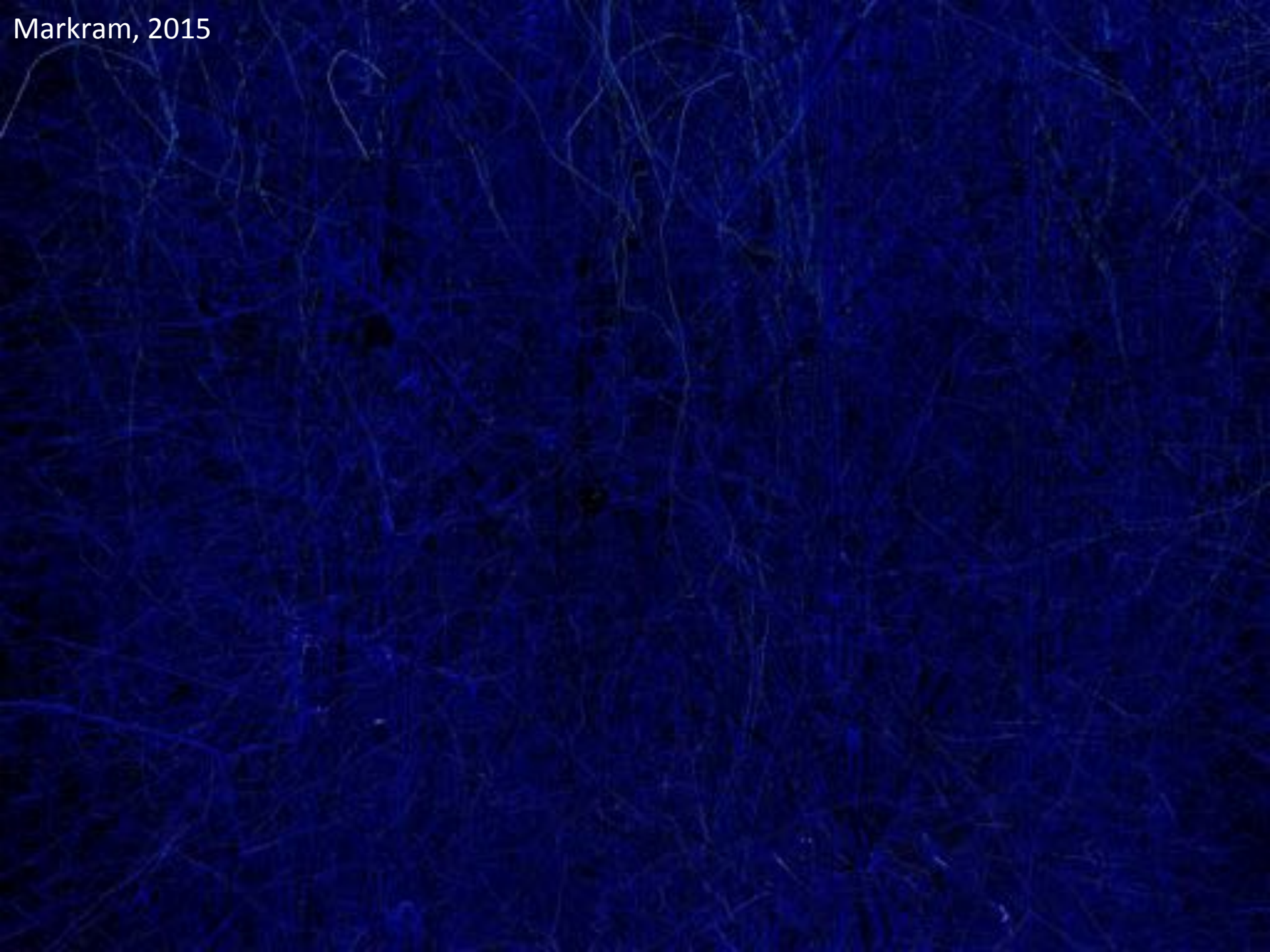
$$I = C_m \frac{\mathrm{d}V_m}{\mathrm{d}t} + \bar{g}_{\mathrm{K}} n^4 (V_m - V_K) + \bar{g}_{\mathrm{Na}} m^3 h (V_m - V_{Na}) + \bar{g}_l (V_m - V_l),$$

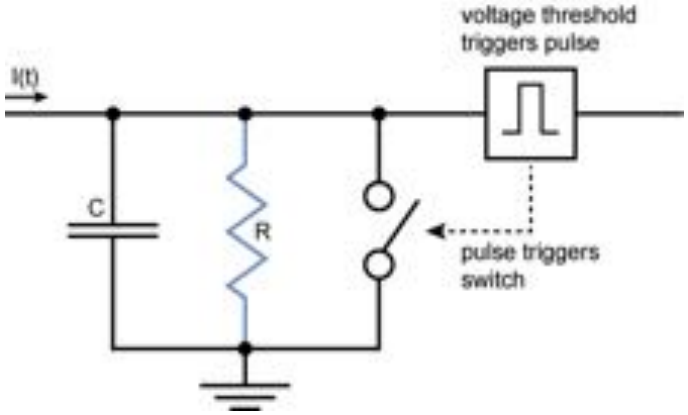
$$\frac{dn}{dt} = \alpha_n(V_m)(1-n) - \beta_n(V_m)n$$

$$\frac{dm}{dt} = \alpha_m(V_m)(1-m) - \beta_m(V_m)m$$

$$\frac{dh}{dt} = \alpha_h(V_m)(1-h) - \beta_h(V_m)h$$



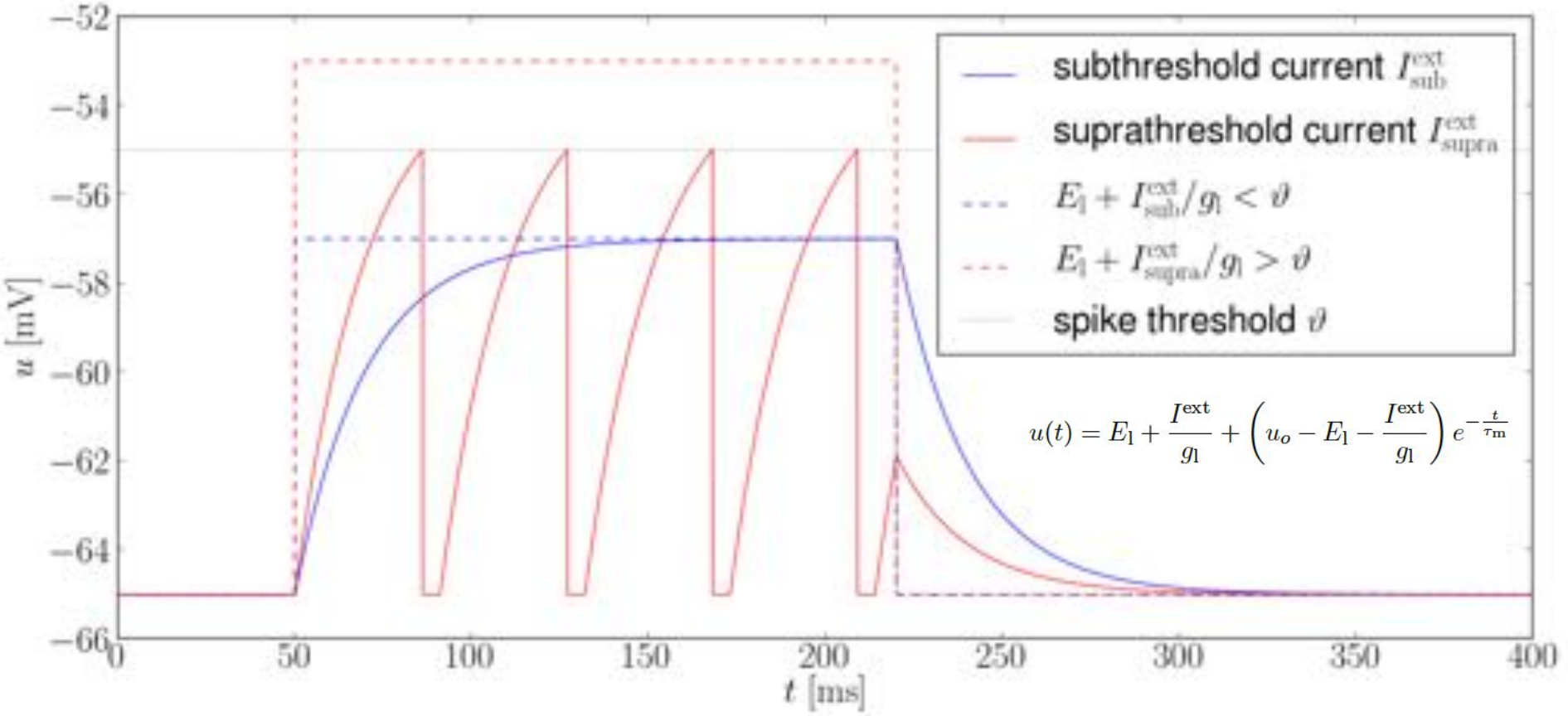

Markram, 2015



Leaky-integrate-and-fire (LIF)

$$C_{\mathrm{m}} \frac{du}{dt} = g_{\mathrm{l}}(E_{\mathrm{l}} - u) + I^{\mathrm{syn}} + I^{\mathrm{ext}}$$

$$u(t_{\mathrm{spike}} < t \leq t_{\mathrm{spike}} + \tau_{\mathrm{ref}}) = \varrho$$

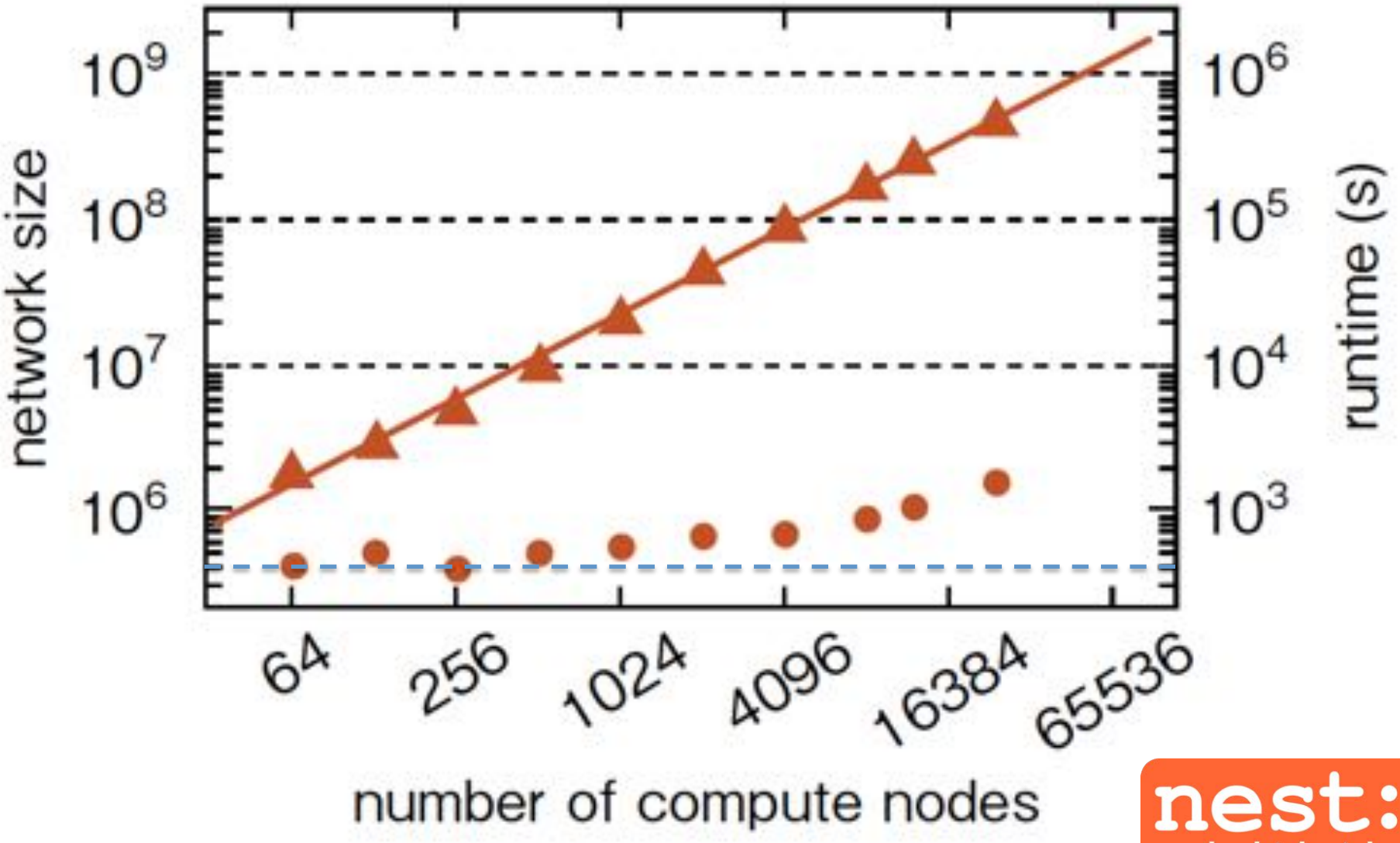


PhD, Mihai Petrovici, 2015

K-Computer, RIKEN Lab, 12.6 MW

Processor-to-Neural Cell Ratio 1 : 20.000

Simulation speed 1.520 : 1 compared to biological real-time

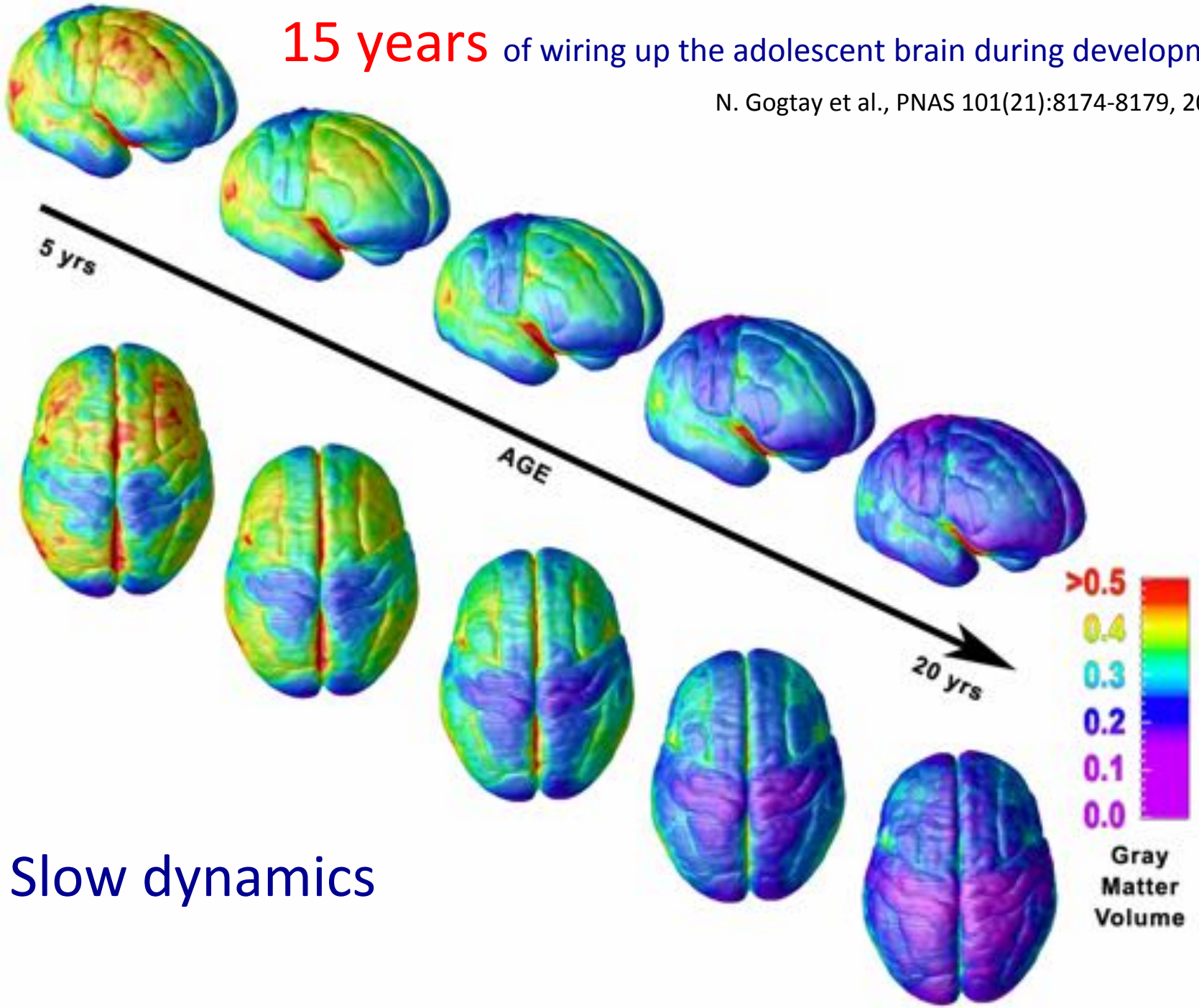


Diesmann, Proceedings of the 4th Biosupercomputing Symposium, Tokyo, 2012

nest:: initiative

15 years of wiring up the adolescent brain during development

N. Gogtay et al., PNAS 101(21):8174-8179, 2004



Slow dynamics

TimeScales	Nature + Real-time	Simulation
Causality Detection	10^{-4} s	0.1 s
Synaptic Plasticity	1 s	1000 s
Learning	Day	1000 Days
Development	Year	1000 Years
12 Orders of Magnitude		
Evolution	> Millenia	> 1000 Millenia
> 15 Orders of Magnitude		

Energy*Scales*

Computational Primitive :
Energy used for a synaptic transmission

10 - 14 orders of magnitude
difference for „*the same thing*“

From : HBP project report

How much does a Neural Computation cost ?

FROM TOP TO BOTTOM (Human Brain)

20 W total Power equally shared

100 Billion neurons firing at 1 Hz 10^{-10} Joule per action potential

10^{15} Synapses transmitting at 1 Hz 10^{-14} Joule per synaptic transmission

FROM BOTTOM TO TOP

Approx. 10^{9} ATP molecules to be hydrolyzed for action potential

Approx. 10^{5} ATP molecules to be hydrolyzed for synaptic transmission

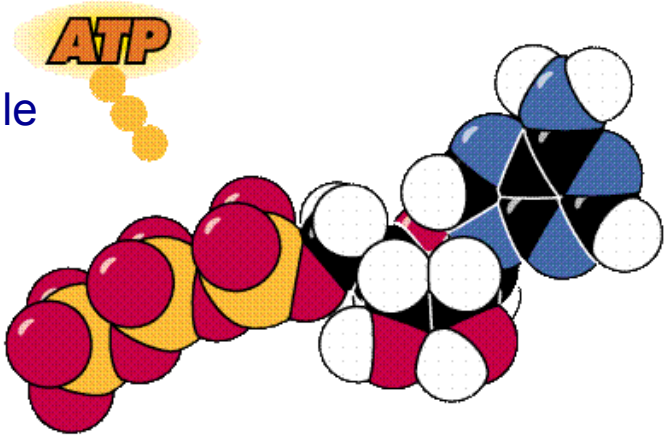
D. Attwell and S. B. Laughlin

Obtain 10^{-19} Joule (approx. 1 eV) per ATP molecule

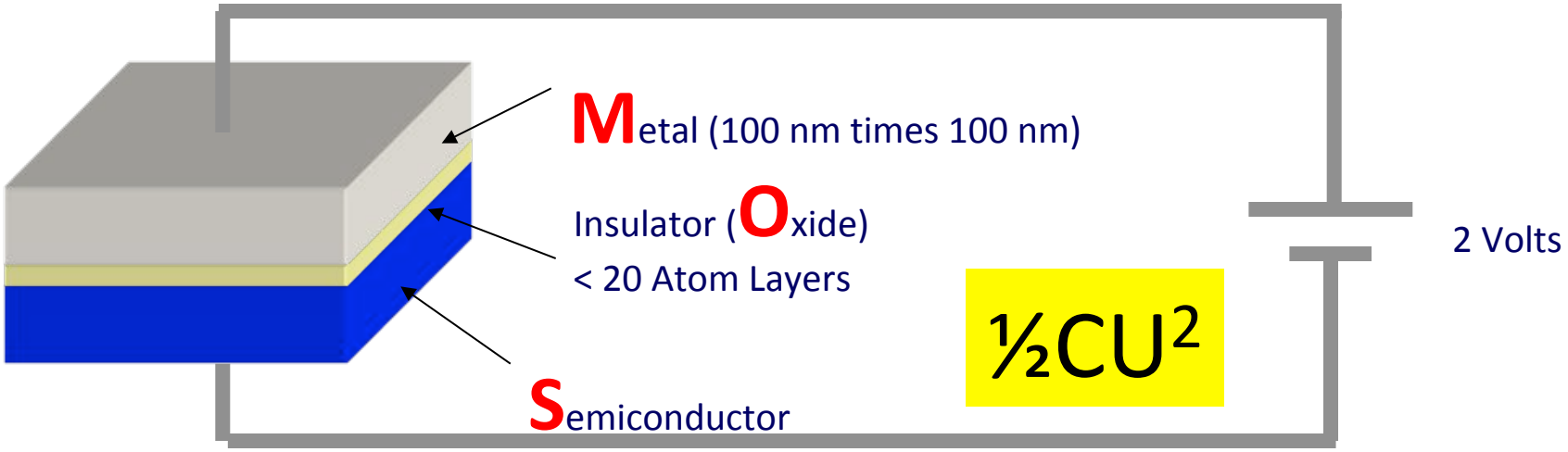
Bray, Dennis. Cell Movements. New York: Garland, 1992

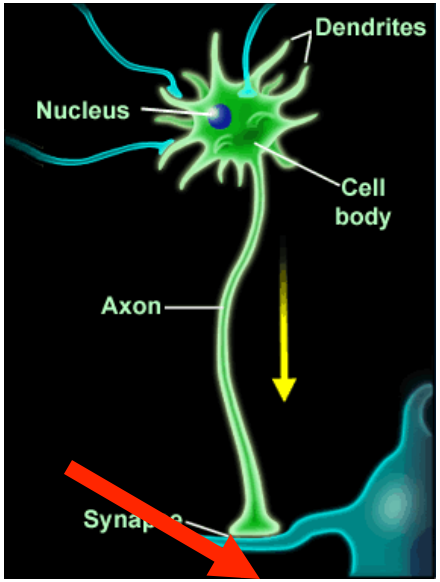
10^{-10} Joule (100.000 fJ = 0.1 nJ) per action potential

10^{-14} Joule (10 fJ) per synaptic transmission



Electronics vs. Biology on the device level - Not a big difference !





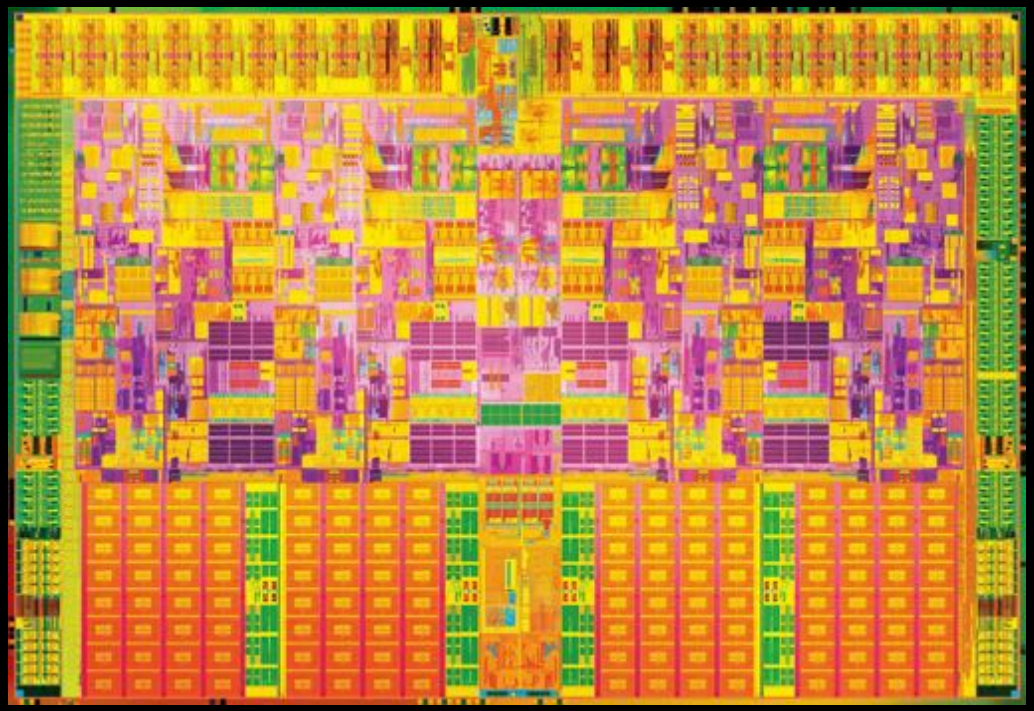
„Switching“ of a MOS transistor : approximately 0.5 fJ

Synaptic Transmission :

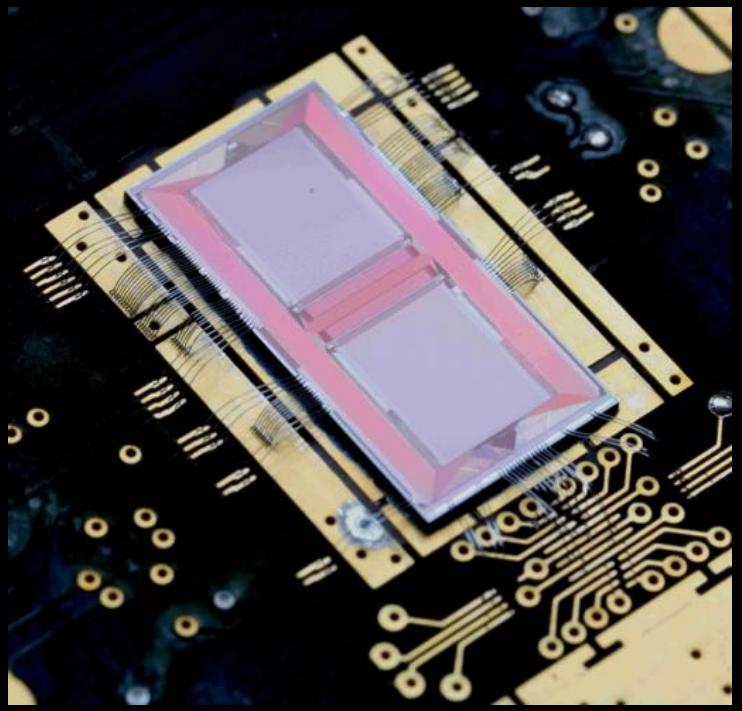
approximately 10 fJ

20 CMOS Transistors

It‘s not the devices, it‘s the ARCHITECTURE and the COMPUTATIONAL MODEL



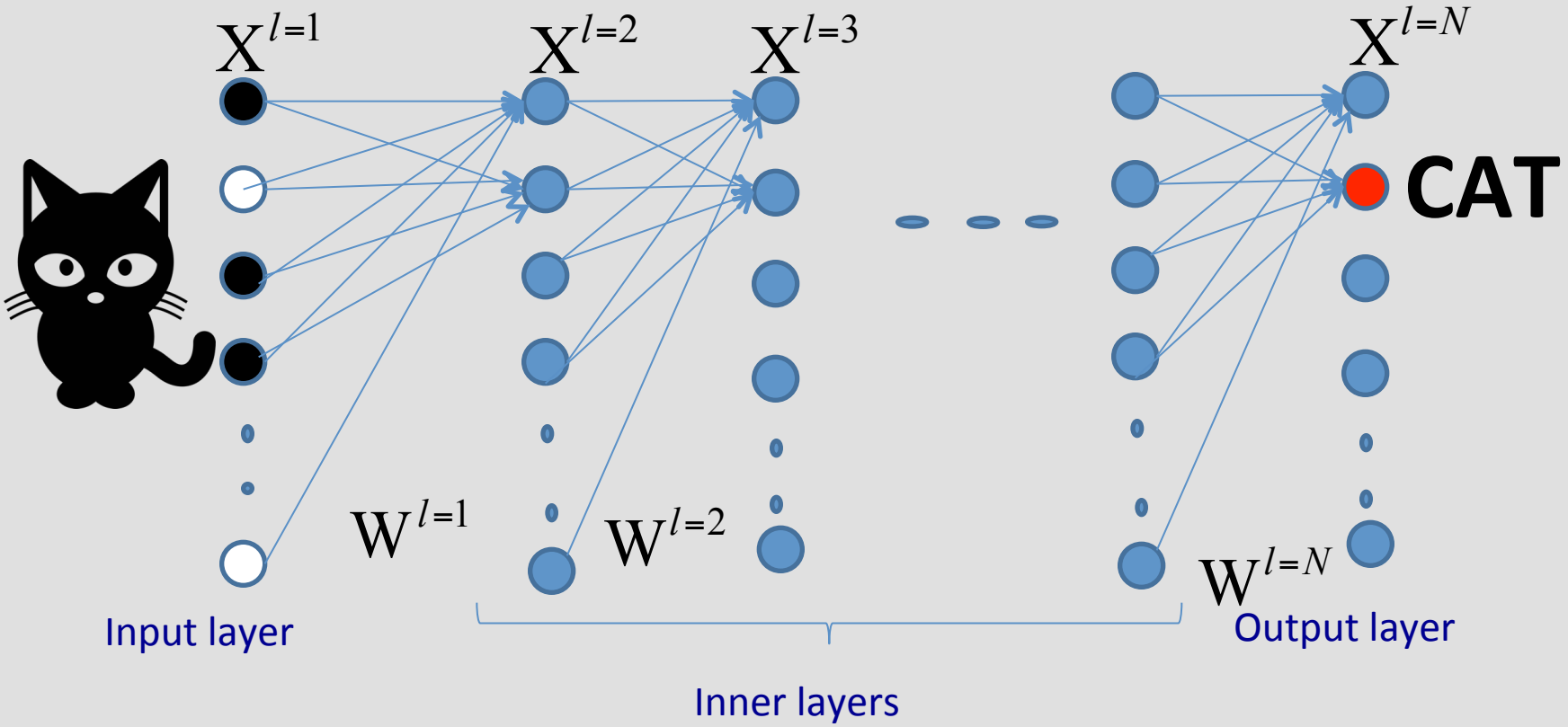
The Brain in a Computer ?



Computers like Brains ?

Artificial Neuronal Networks

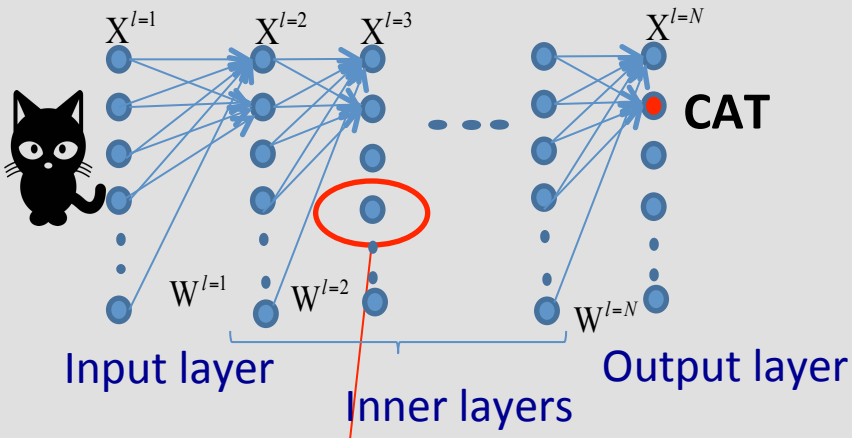
ignore time evolution ….

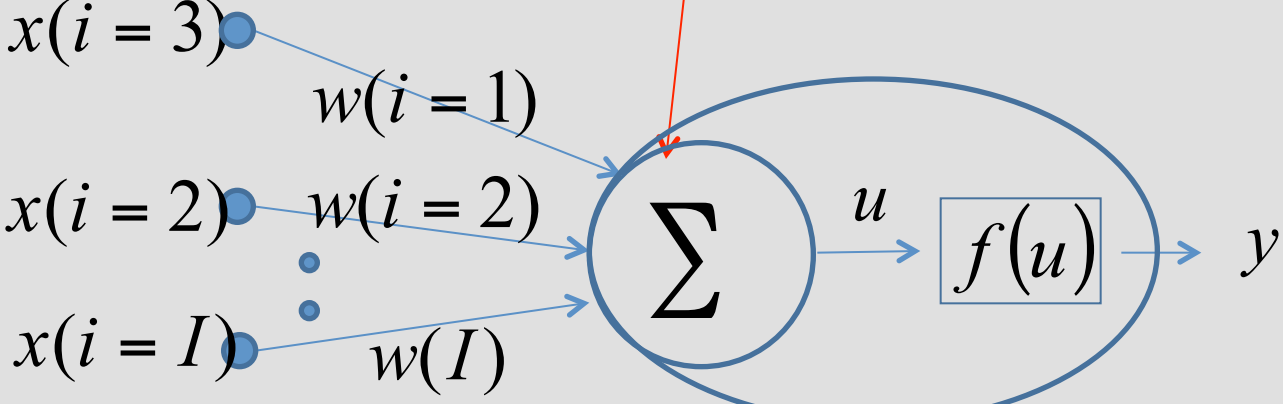


Here : local, no recurrency *feed-forward*

Pairs of neurons connected by weights

Neuron performs integration (summing)

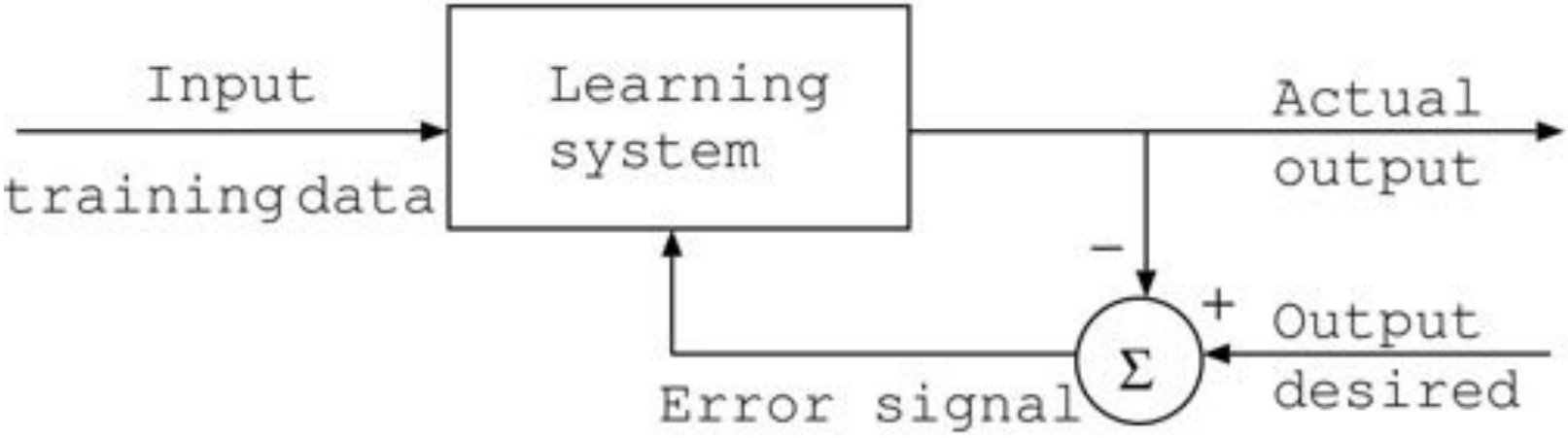




Learning Example : Supervised

Labelled input data

Actual output



Deviation

Desired output

Jumpstart

Strategic Network

Supervised Learning

Predict human moves

database of existing matches

160.000 matches, 30 Million positions

Policy Network

Reinforcement Learning

Network self-matches

128.000.000 matches

Value Network

Combination of first 2 steps

30 Million self-matches

One year learning time, 0.5 MW

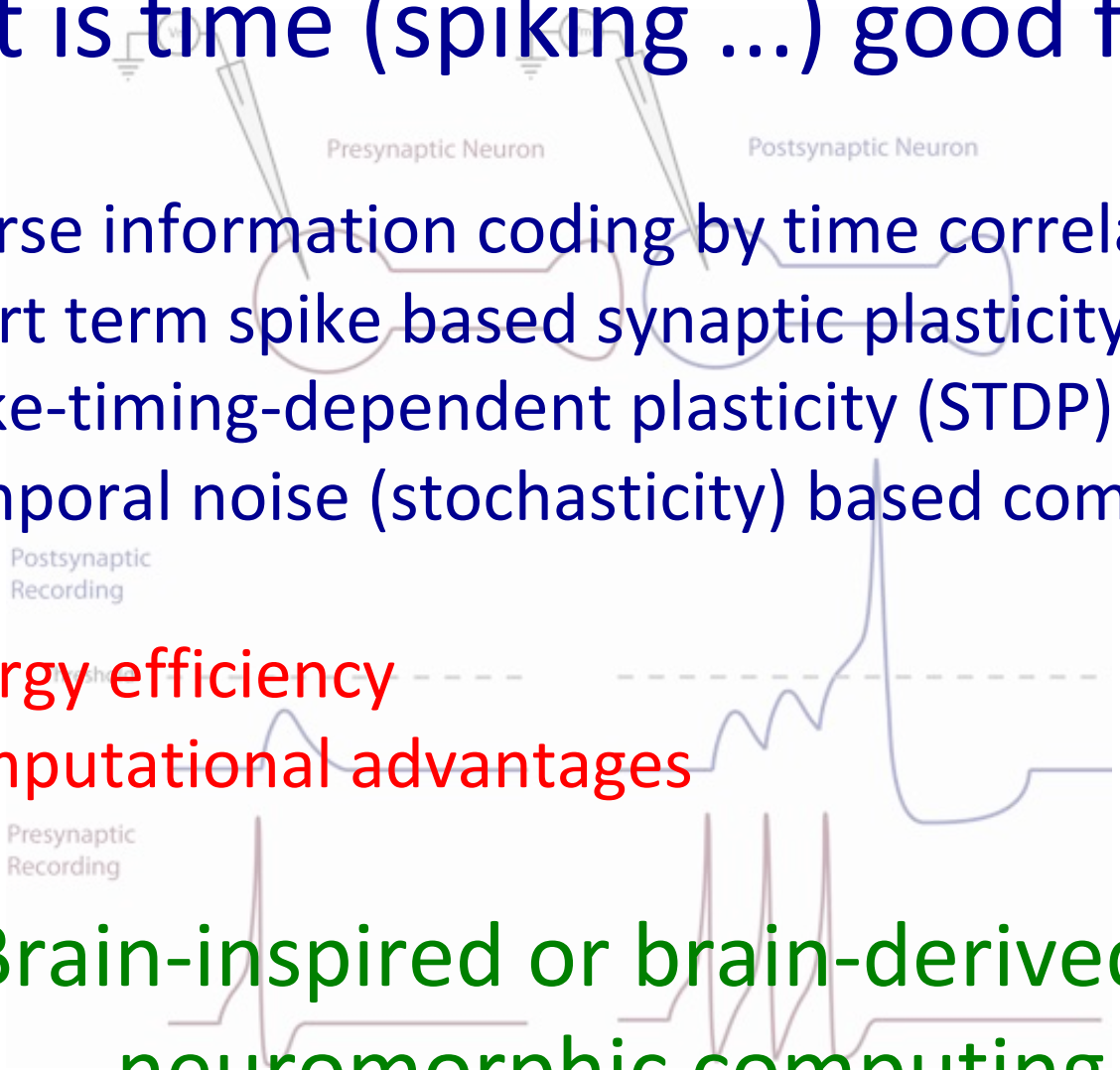
Energy : 183 MWh

Excessive training samples

Learning is slow and expensive

Application is fast

What is time (spiking ...) good for ?



- Sparse information coding by time correlations
- Short term spike based synaptic plasticity (STP)
- Spike-timing-dependent plasticity (STDP)
- Temporal noise (stochasticity) based computing

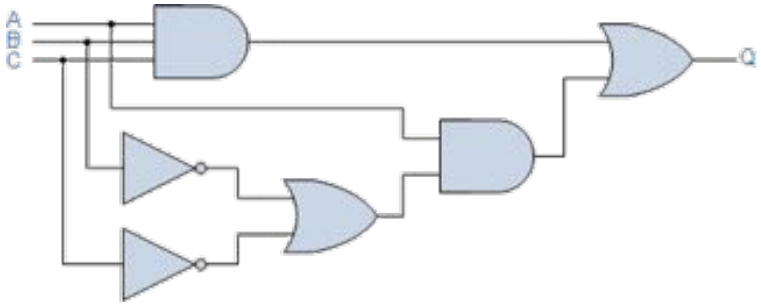
- Energy efficiency
- Computational advantages

Brain-inspired or brain-derived or neuromorphic computing

https://upload.wikimedia.org/wikipedia/commons/0/0a/Temporal_summation.JPG

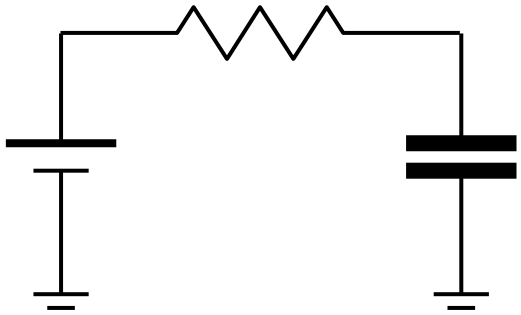
Digital

- Discrete values of physical variables
- Computation by Boolean algebra
- One wire one bit of information
- Signal restored after gate



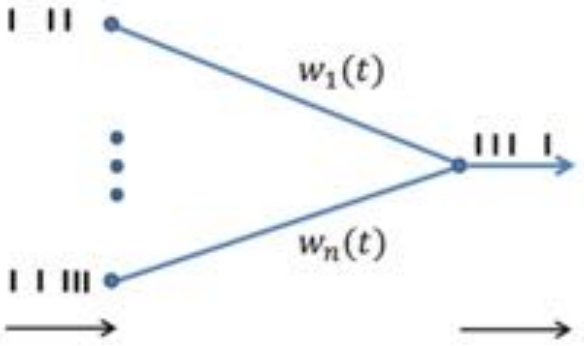
Analog

- Continuous values of physical variables
- Computation by component physics
- One wire many bits of information
- Signal not restored after stage



Nature / mixed-signal

- Local analogue computation
- Binary communication by spikes
- Signal restoration



Large-scale Neuromorphic Computing – compare

➢ Commodity microprocessors	SpiNNaker, HBP	Soft-binary-code
➢ Custom fully digital	TrueNorth, IBM	Hard-binary-code
➢ Custom Mixed-Signal	BrainScaleS, HBP	Physical model

Anything in common ?

+ Massively parallel (close to perfect weak scaling)
+ Asynchronous communication
+ Configurability
- Limited flexibility and complexity in neural models

COMPLEMENTARITY OF APPROACHES ESSENTIAL !

SpiNNaker
Biologically Inspired Massively Parallel Architectures

IBM®

BrainScaleS
ScaleS

HBP Neuromorphic Computing Concepts

MANY-CORE NUMERICAL MODEL SYSTEM

0.5 – 1 Million ARM processors – address-based, small packet, asynchronous communication – real-time simulation

Location : Manchester (UK)

PHYSICAL MODEL SYSTEM

Local analog computing with 4 Million neurons and 1 Billion synapses – binary, asynchronous communication – x 10 000 accelerated emulation

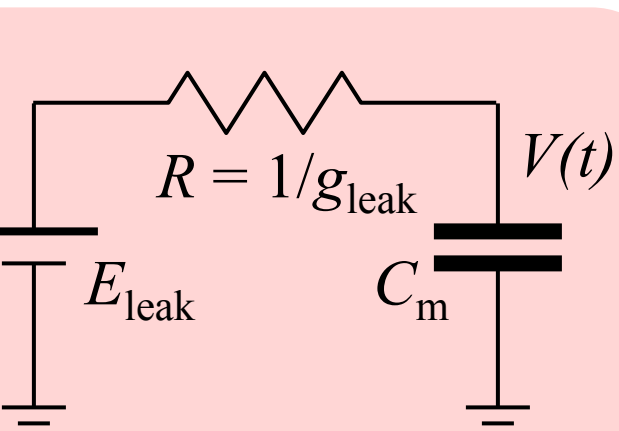
Location : Heidelberg (Germany)

Physical Model System

Continuous Time Integrating Neural Cell Membrane

(+ non-linearity)

$$C_m \frac{dV}{dt} = -g_{leak}\left(V - E_{leak}\right)$$



	g_{leak} [S]	C_m [F]
Biology(*)	10^{-8}	10^{-10}
VLSI	10^{-6}	10^{-13}

(*) Brette/Gerstner, J. Neurophysiology, 2005

$$c_m \frac{dV}{dt} = -g_{leak}\left(V - E_l\right) + \sum_k p_k g_k \left(V - E_x\right) + \sum_l p_l g_l \left(V - E_i\right)$$

$p_{k,l}(t)$ exponential onset and decay (post-synaptic potential shape)

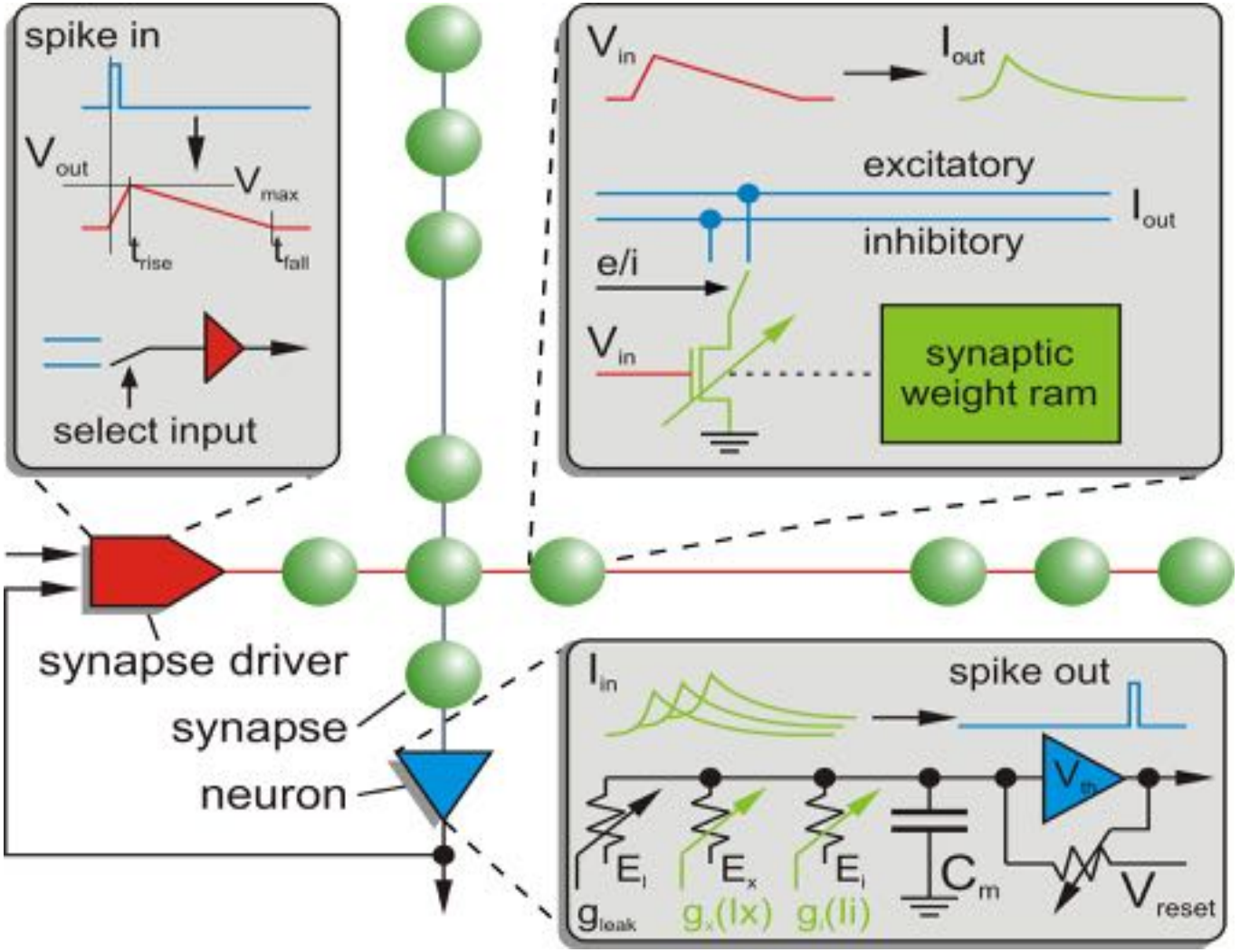
$g_{k,l}$ 0 to g_{max} (“weights”)

effective membrane time-constant c_m / g_{total} is time-dependent

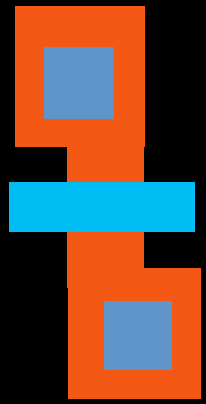
„Time“ is imposed by internal physics, not by external control

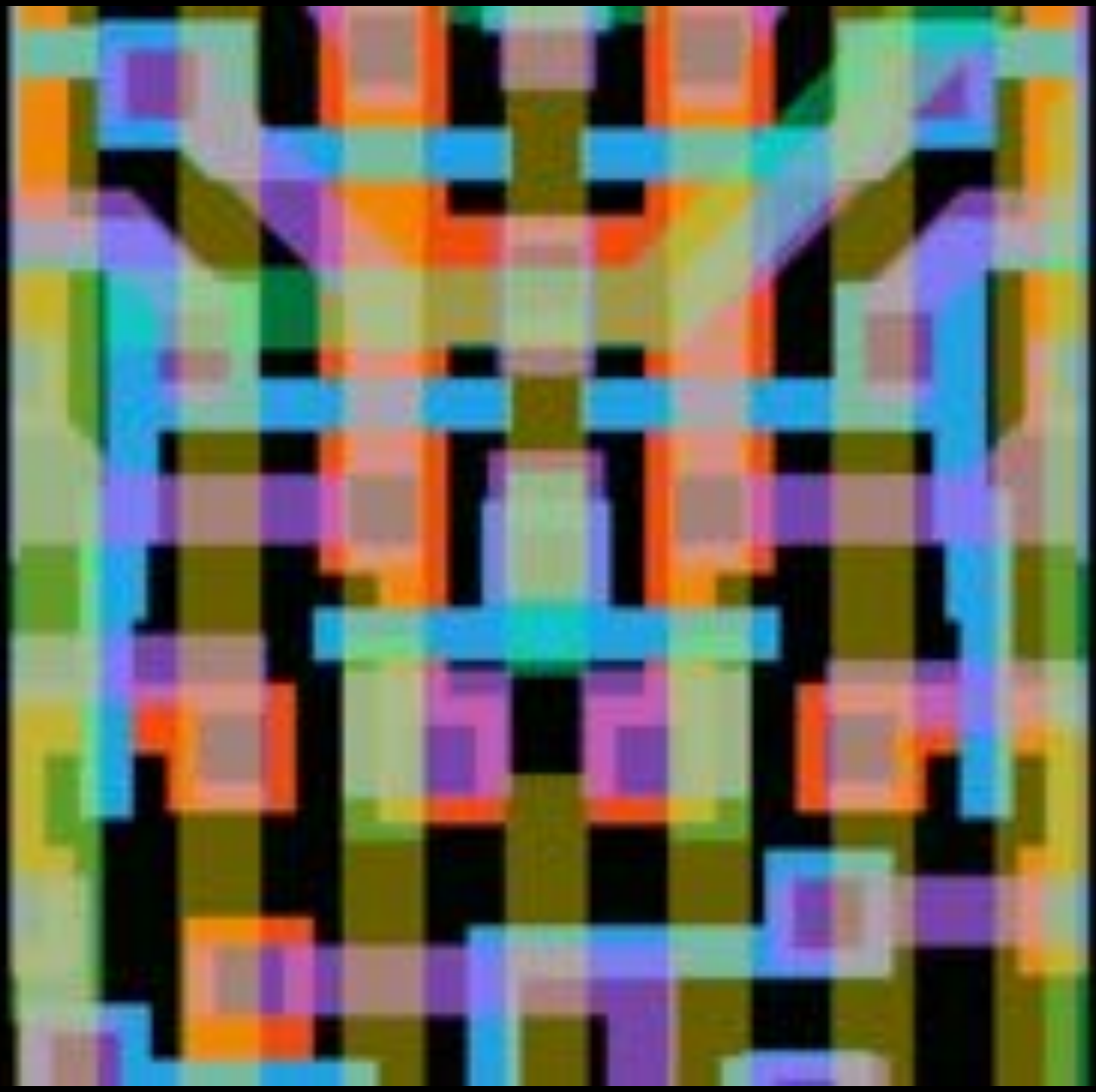
Implementation example with synaptic inputs and neuron non-linearity

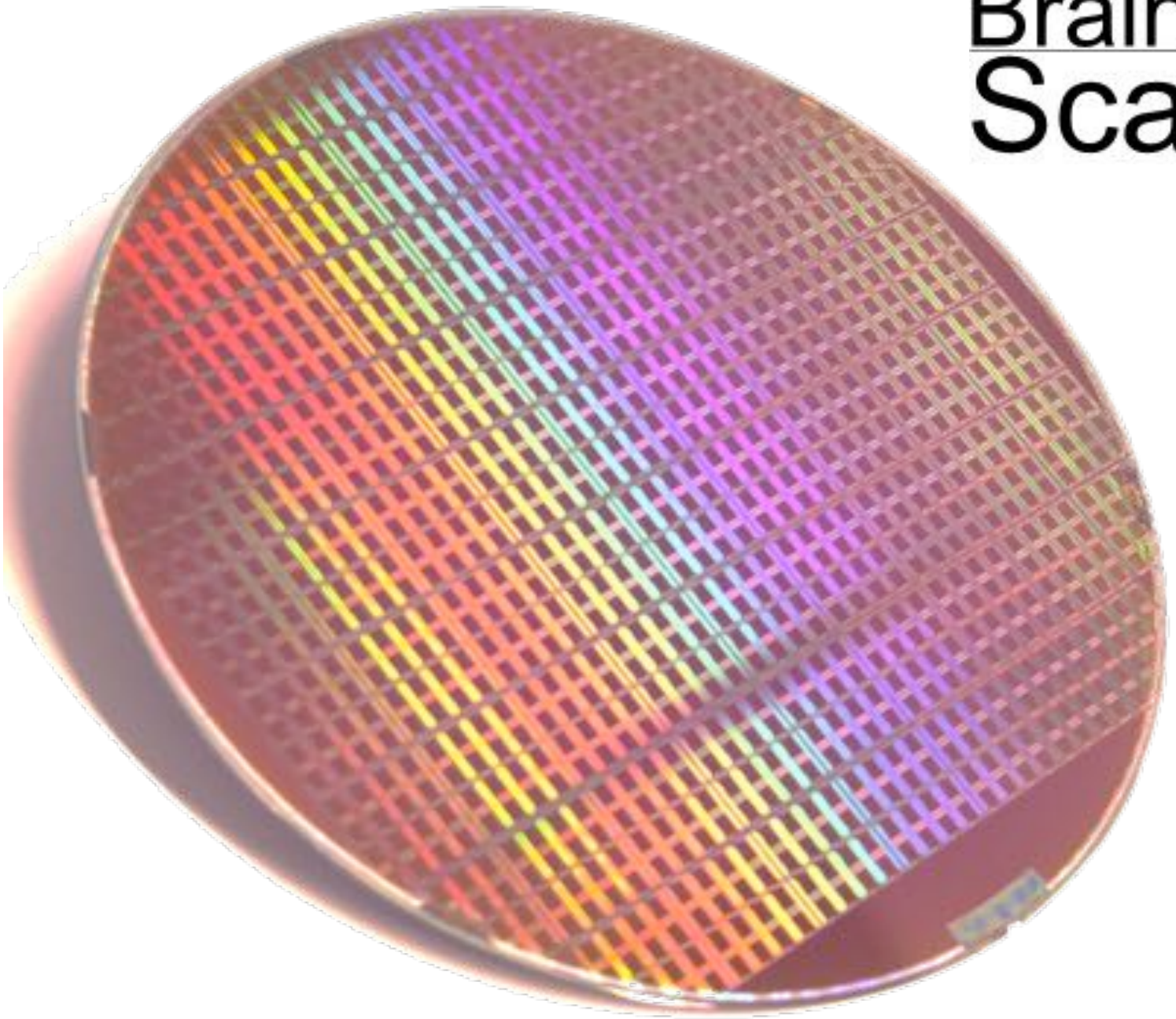
mixed-signal : analog cores, binary communication



A New VLSI Model of Neural Microcircuits Including Spike Time Dependent Plasticity, Johannes Schemmel, Karlheinz Meier, Eilif Muller, Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04), IEEE Press, pp. 1711-1716, 2004







BrainScaleS

ScaleS

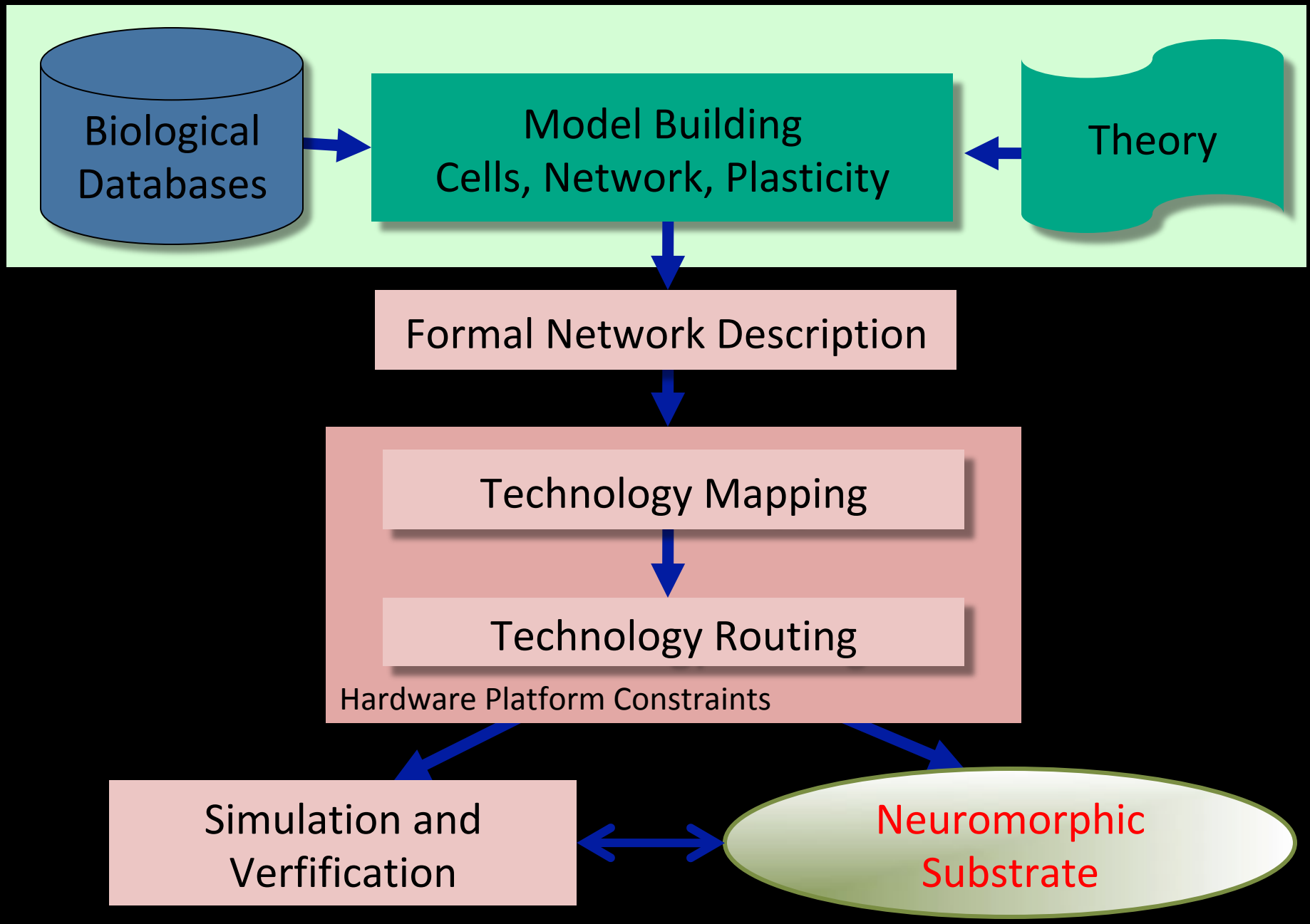
BrainScaleS

ScaleS

Physical Model, local analogue computing, binary continuous time communication

Wafer-Scale Integration of 200.000 neurons and 50.000.000 synapses on a single 20 cm wafer

Short term and long term plasticity, 10.000 faster than real-time


Biological Databases
Model Building
Cells, Network, Plasticity
Theory
Formal Network Description
Technology Mapping
Technology Routing
Hardware Platform Constraints
Simulation and Verfification
Neuromorphic Substrate

Configuration Space 40 MB for a full Wafer

Scope	Name	Type	Description
Neuron circuits (A)			membrane voltage
Synapse line drivers (B)			amp
Synapses (B)			
STP related (C)			on
STDP related (D)	n/a	i_t	Bias current controlling delay for presynaptic correlation pulse (for calibration purposes)
	$A_{+/-}$	s_t	Two voltages dimensioning charge accumulation per (anti-)causal correlation measurement
	n/a	s_t	Two threshold voltages for detection of relevant (anti-)causal correlation
	τ_{STDP}	g	Voltage controlling STDP time constants

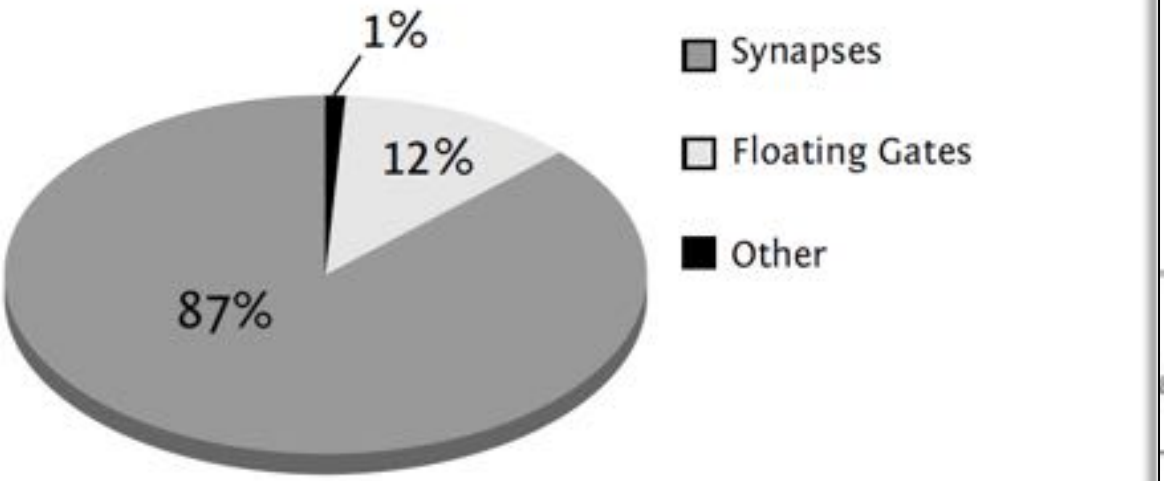
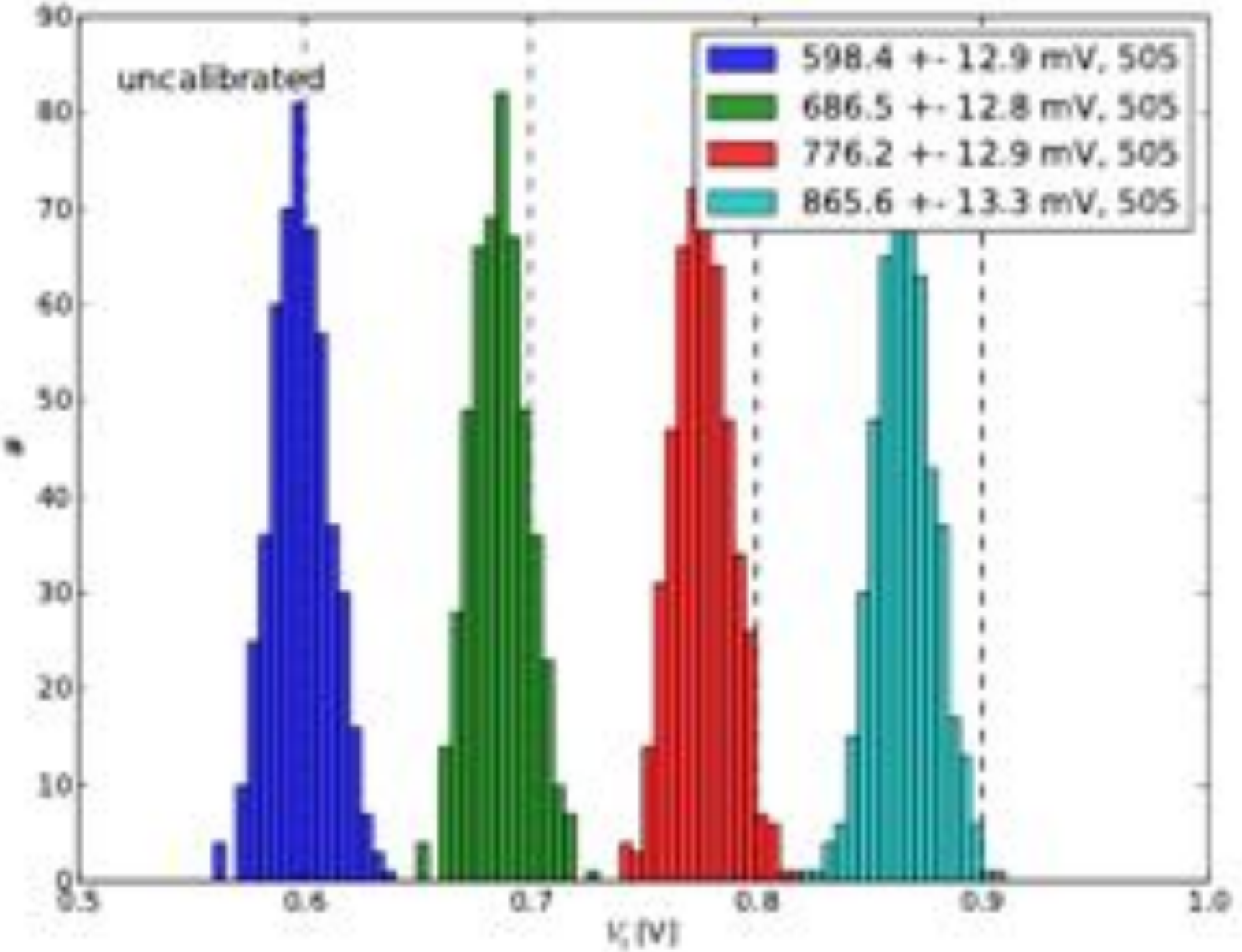


Fig. 4: Sector diagram of the parameter space to configure one HICANN chip. For a full wafer, the configuration data volume is 44 MB large.

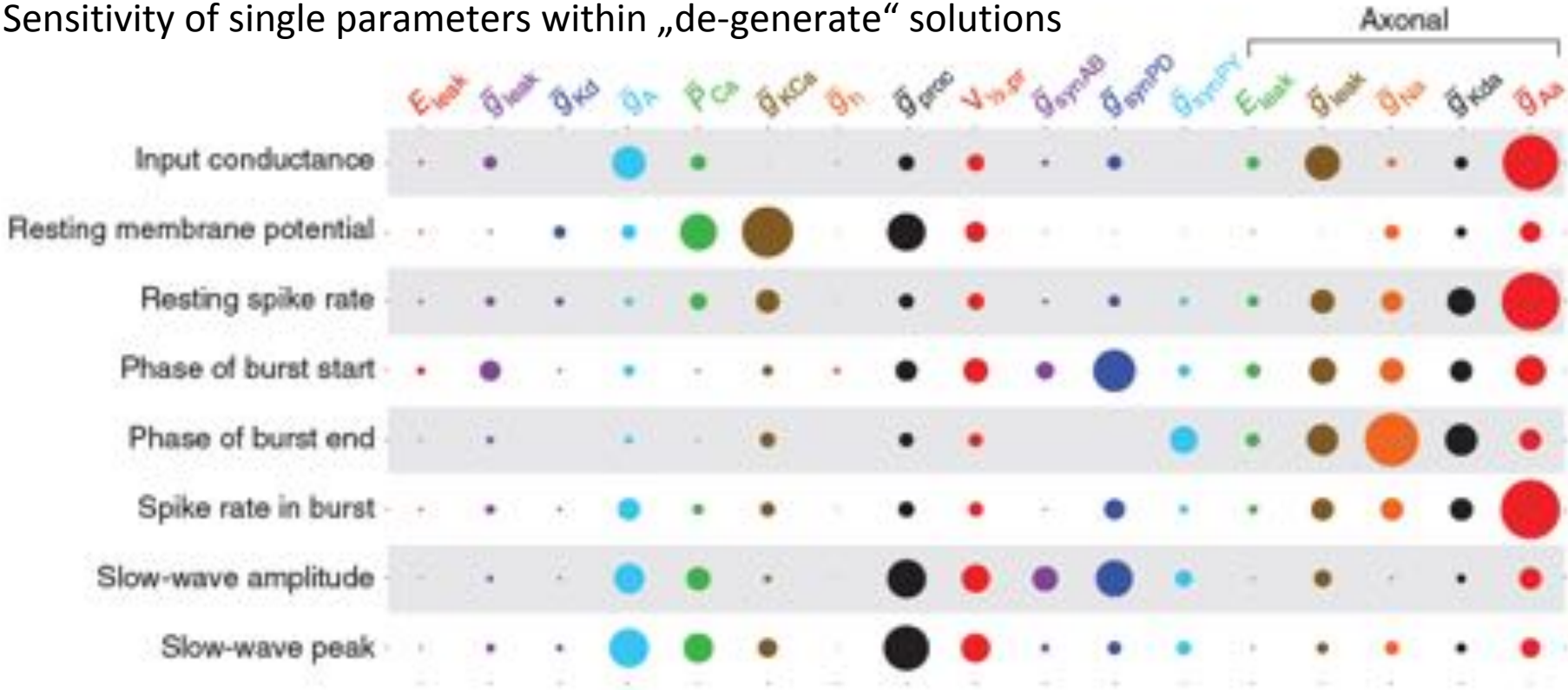
Challenge and Opportunity : Variability



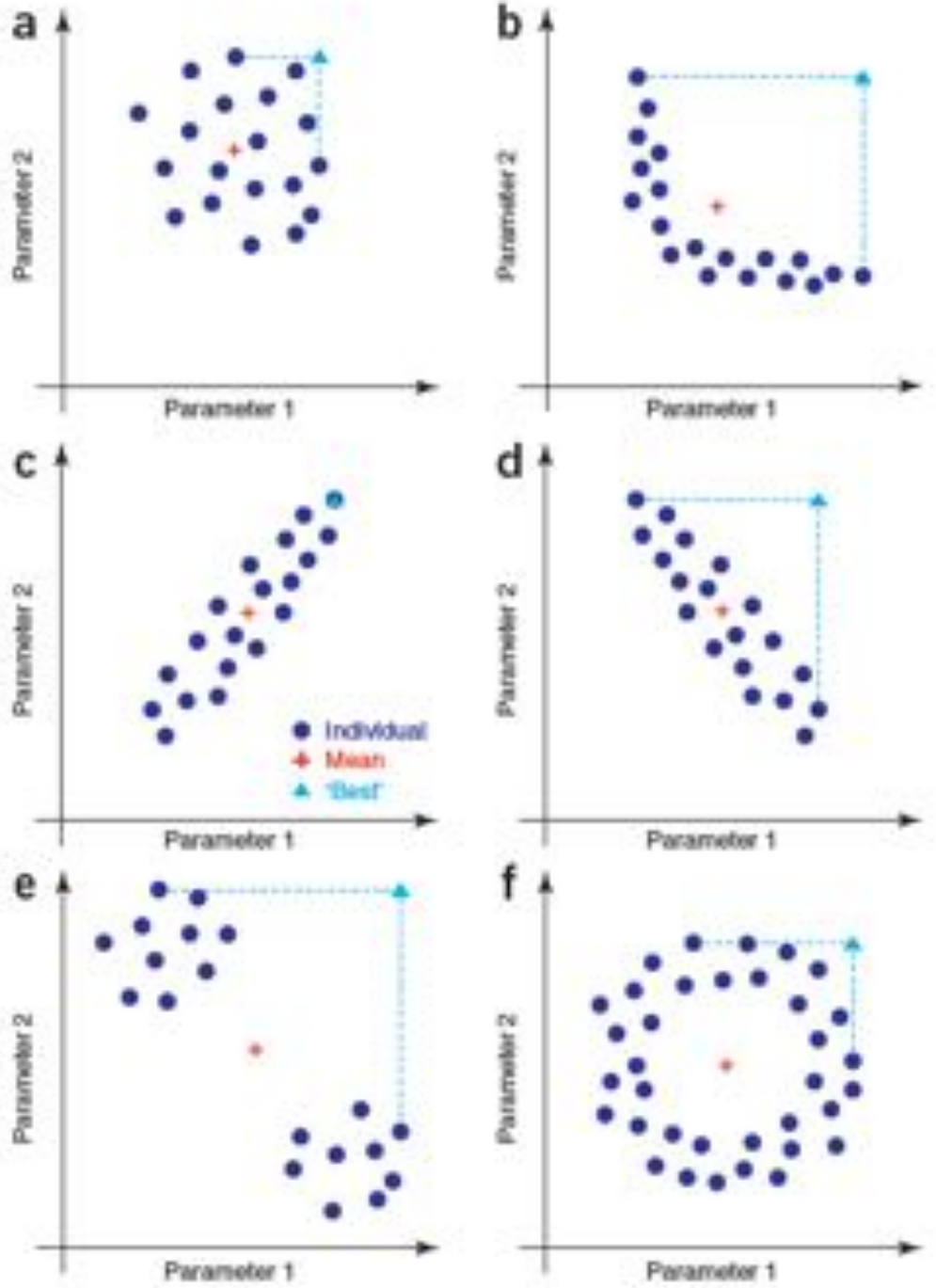
Pyloric rhythm of the crustacean stomatogastric ganglion

20.000.000 model networks created with 17 random cell parameters, fixed connectivity (Neuron)
400.000 networks found with „identical (de-generate)“ timing behaviour in measured biological range

Sensitivity of single parameters within „de-generate“ solutions



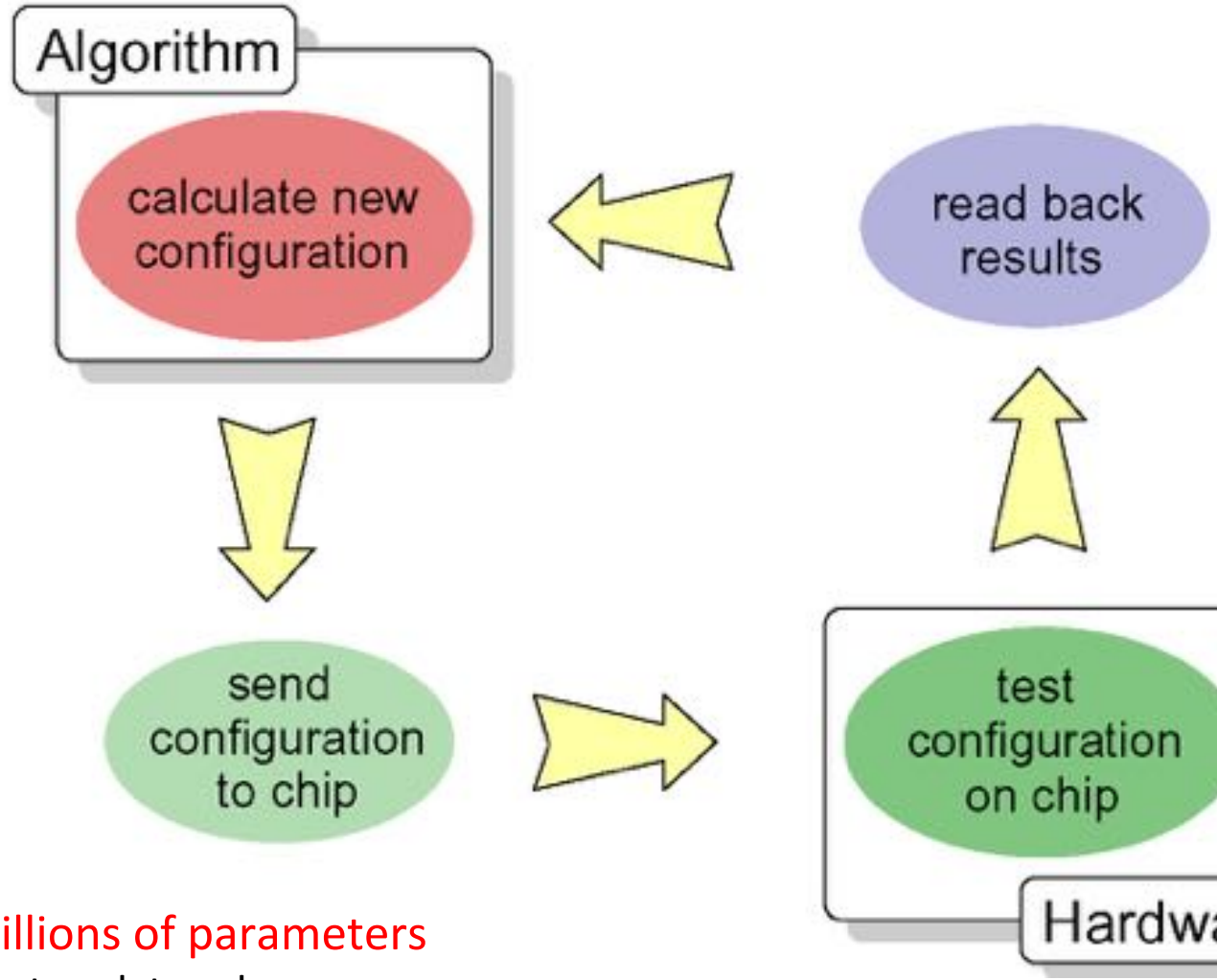
Marder, Taylor Nature Neuroscience 14, Nr 2, 2011



Variability has to be at the right place ...

Marder, Taylor Nature Neuroscience 14, Nr 2, 2011

Hardware-In-the-Loop



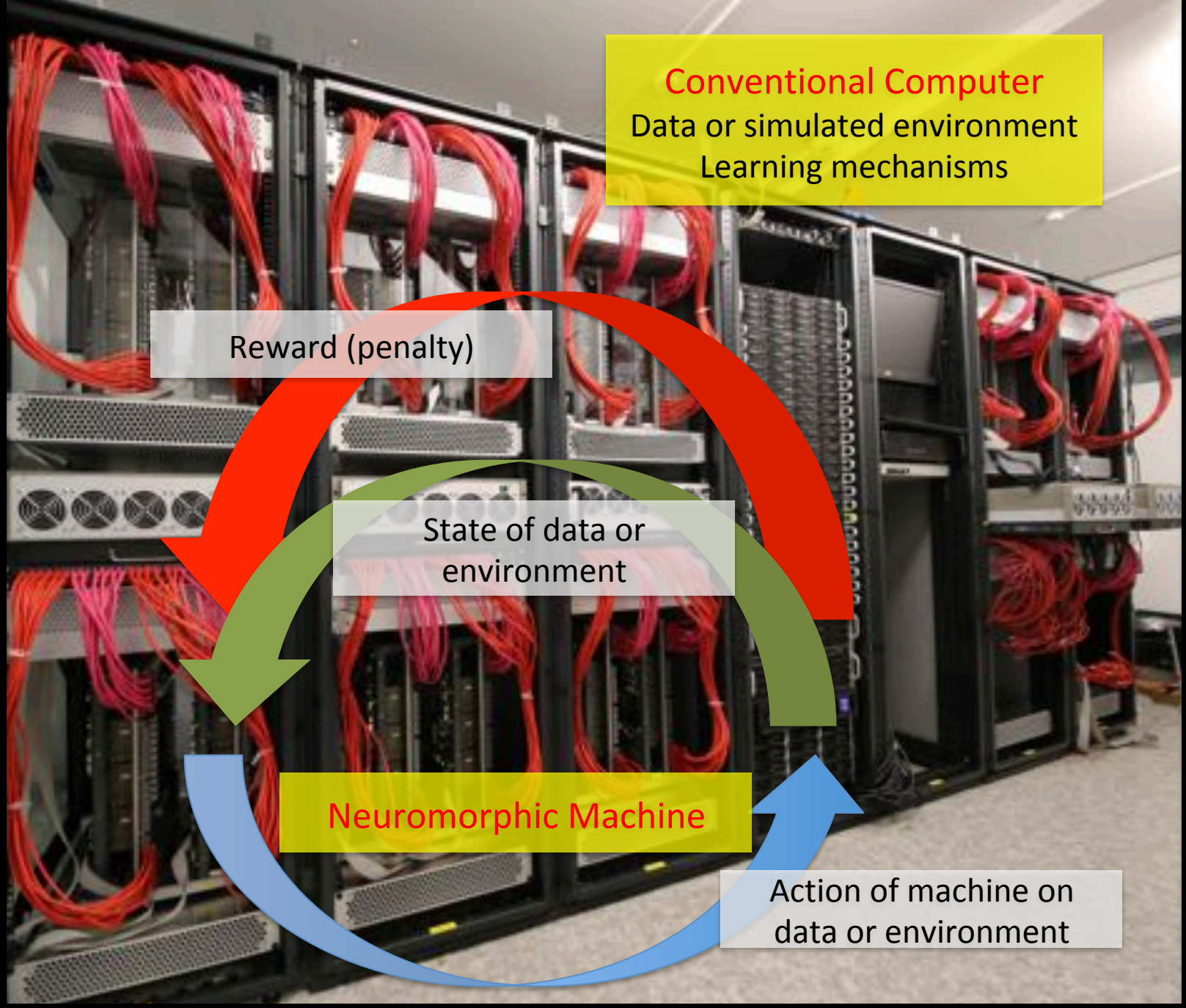
What for ?

- Calibration
- Learning
- Environment
- Data

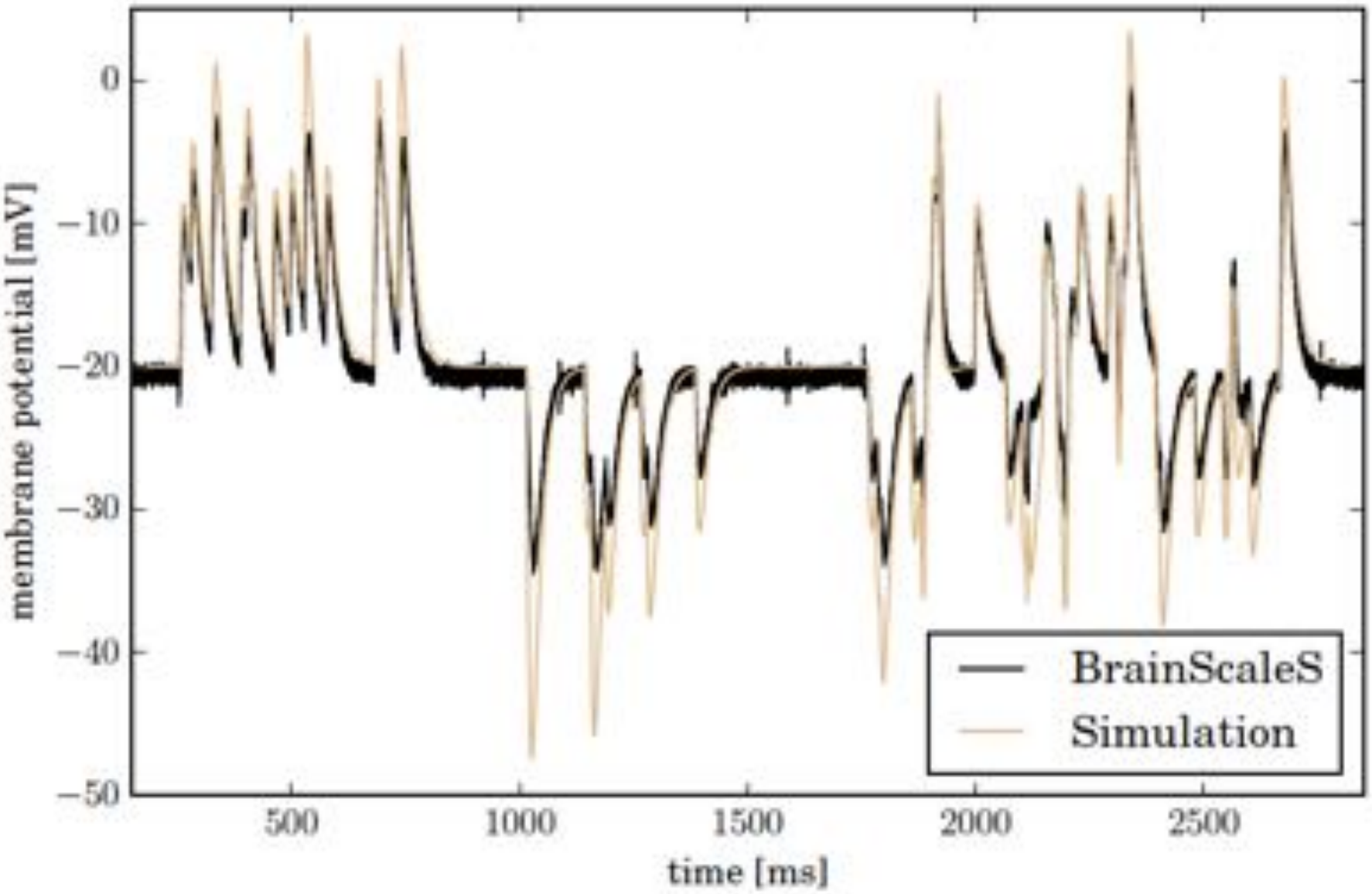
Separated ?

Millions of parameters

- network topology
- neuron sizes and parameters
- synaptic strengths

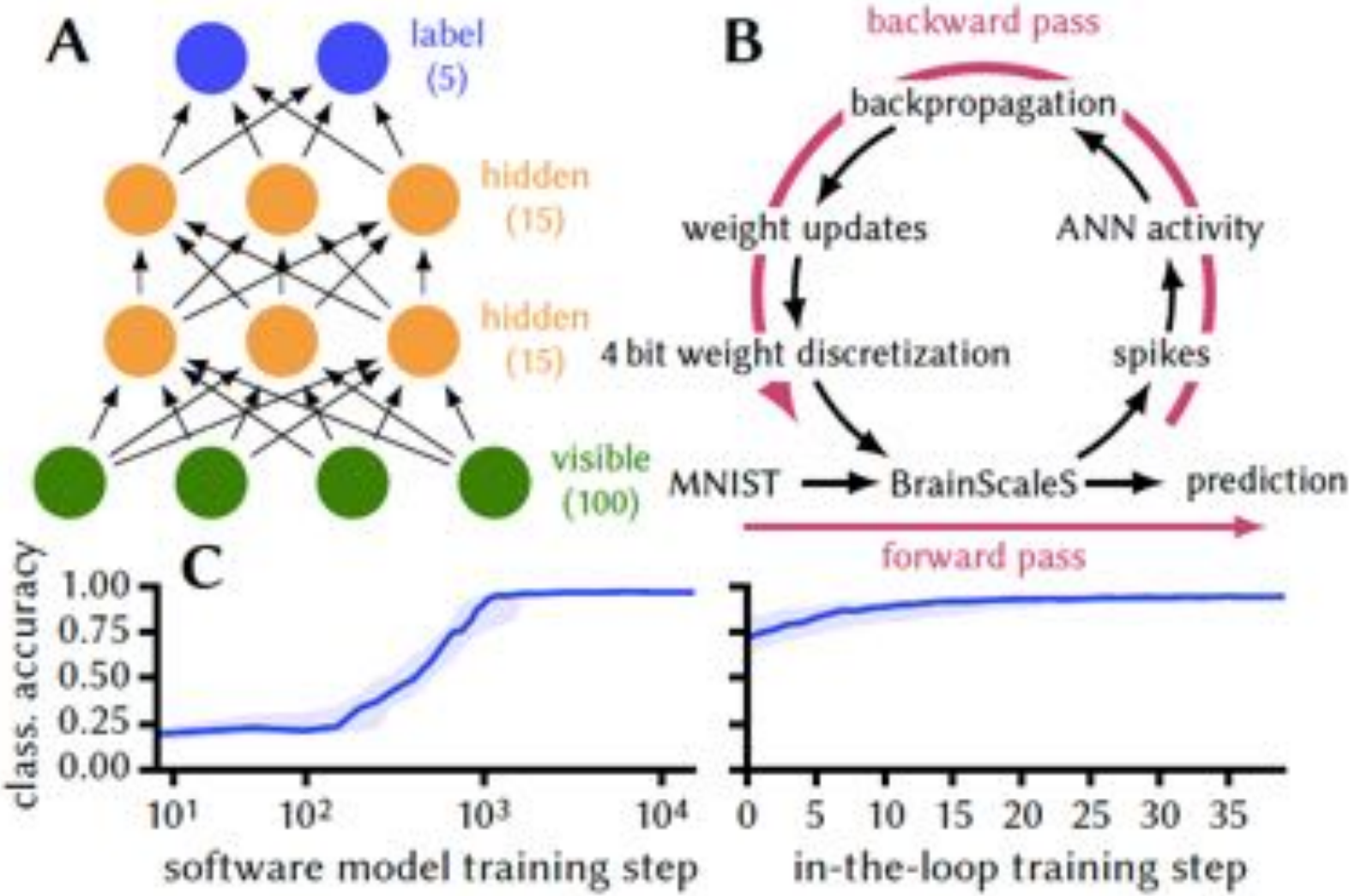

Conventional Computer
Data or simulated environment
Learning mechanisms
Reward (penalty)
State of data or environment
Neuromorphic Machine
Action of machine on data or environment

Physical model emulation



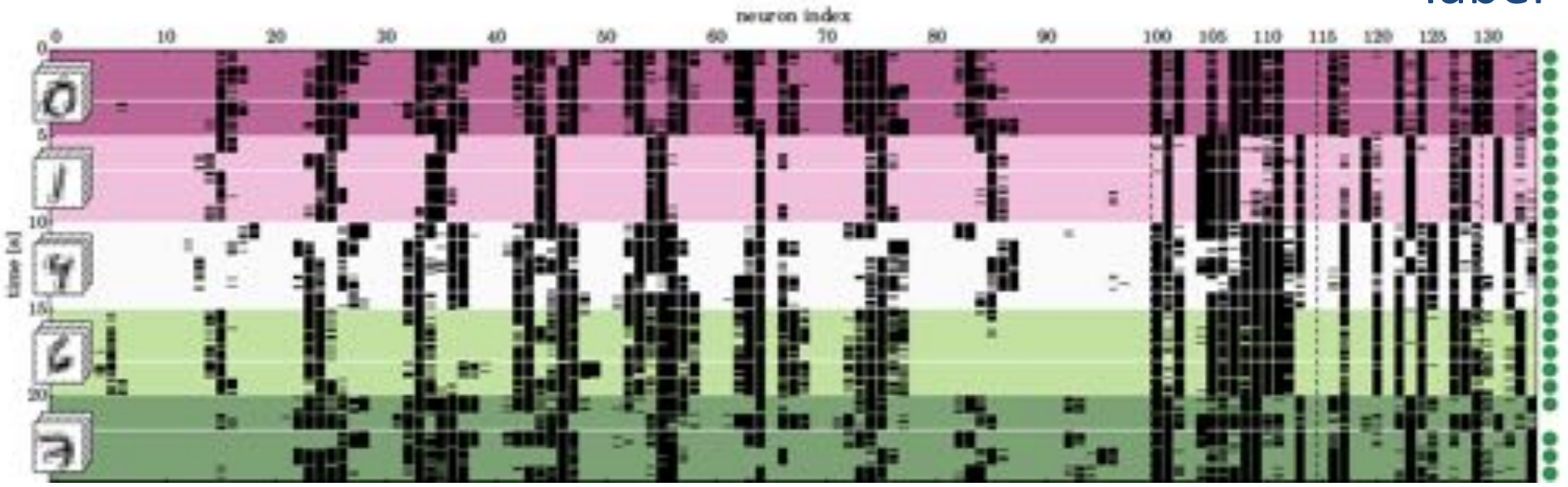
Sebastian Schmitt et al., accepted IJCNN 2017

Feed-forward, rate-based. 4-layer spiking network
MNIST classification on a physical model machine
performance before and after hardware in-the-loop learning



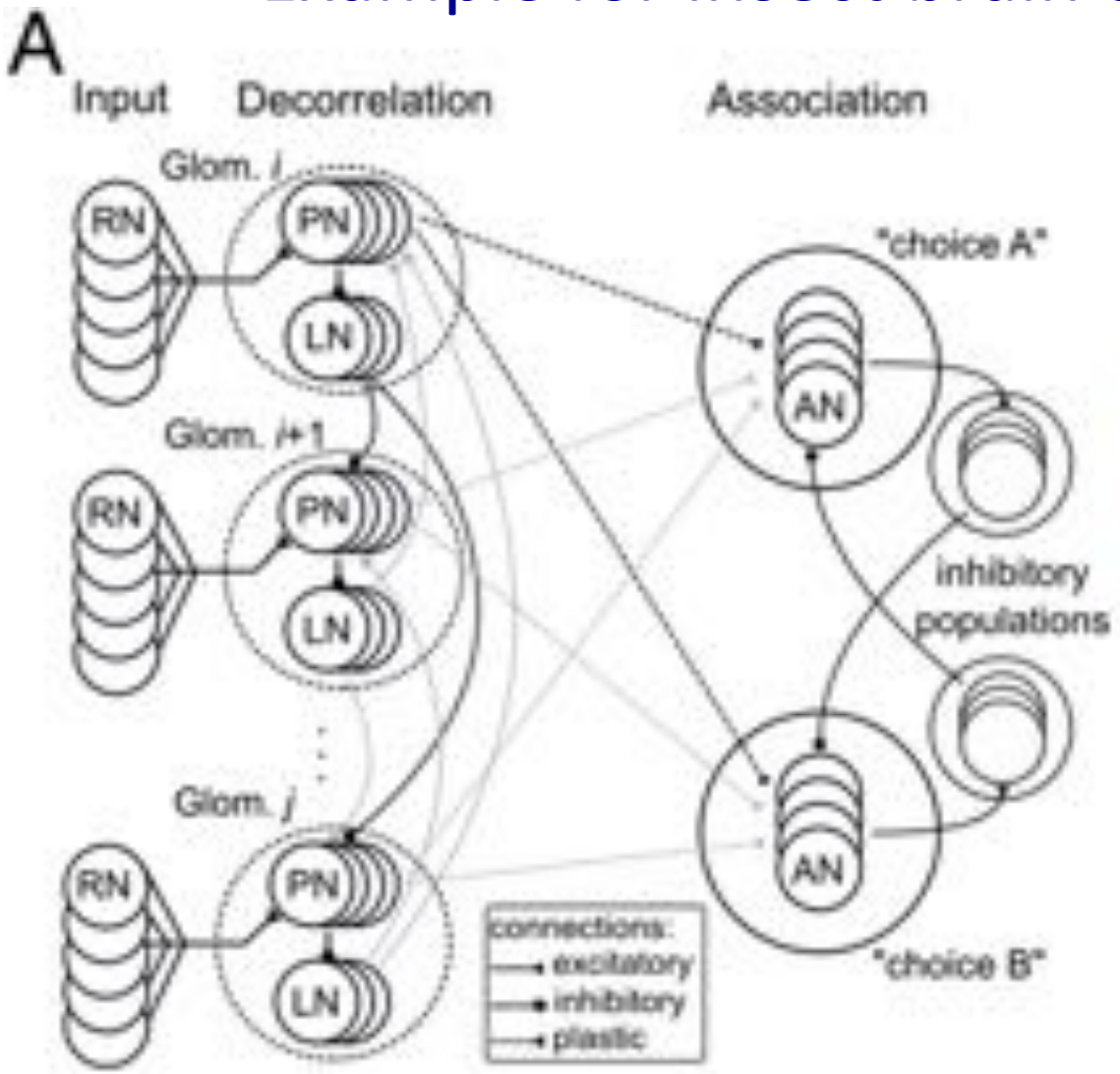
https://arxiv.org/abs/1703.01909

MNIST classification on a physical model machine
Neuronal firing activity after hardware in-the-loop learning



https://arxiv.org/abs/1703.01909

Example for insect brain derived circuit



3 Layer Spiking Neuron Network derived from Insect Olfactory System

L I : Receptor Neurons

L II : Decorrelation through lateral inhibition (Glomeruli)

L III : Association (Soft WTA through strong inhibitory populatuions)

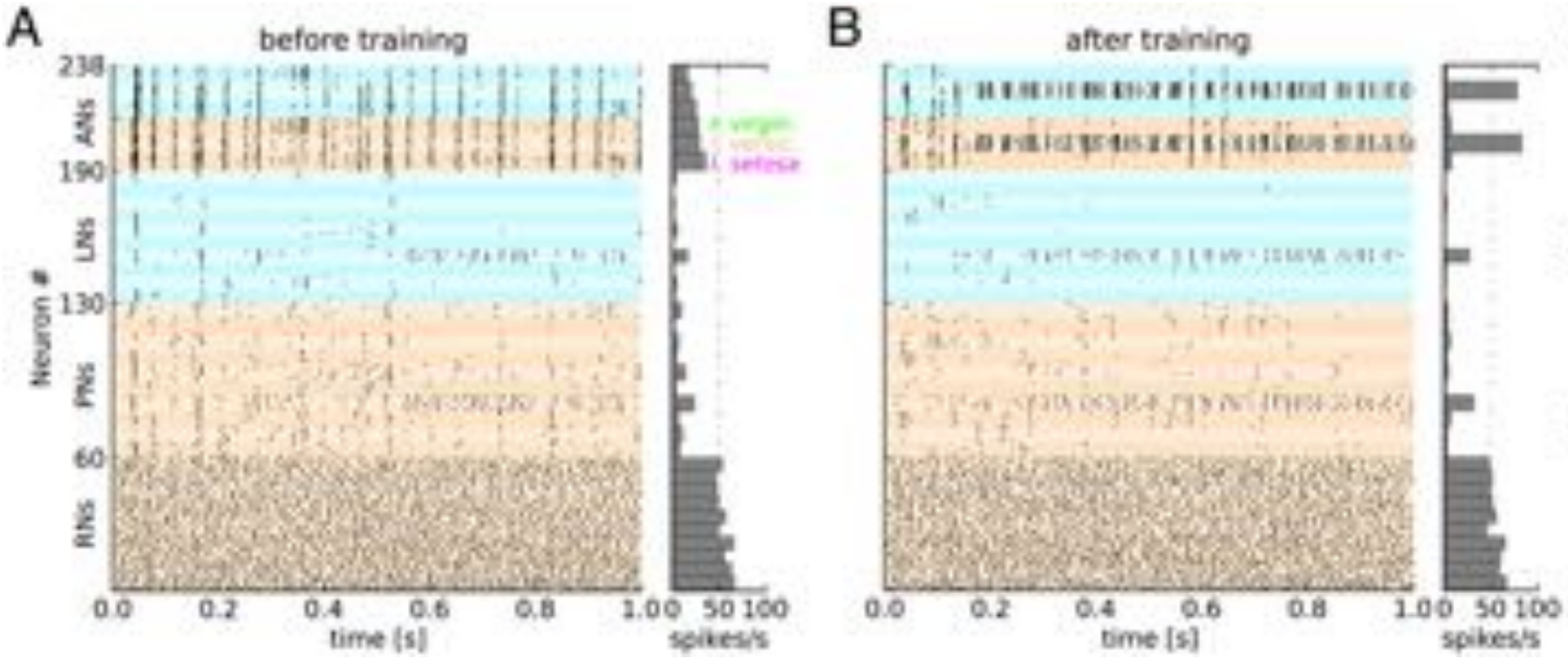
Supervised Learning

Synaptic Projections from Layer 2 to Layer 3

Schmuker, M. et al., "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

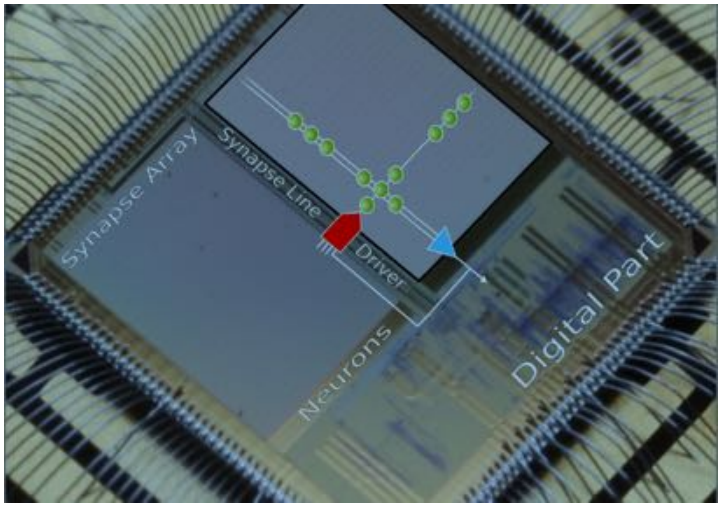
Neuronal firing activity before and after learning

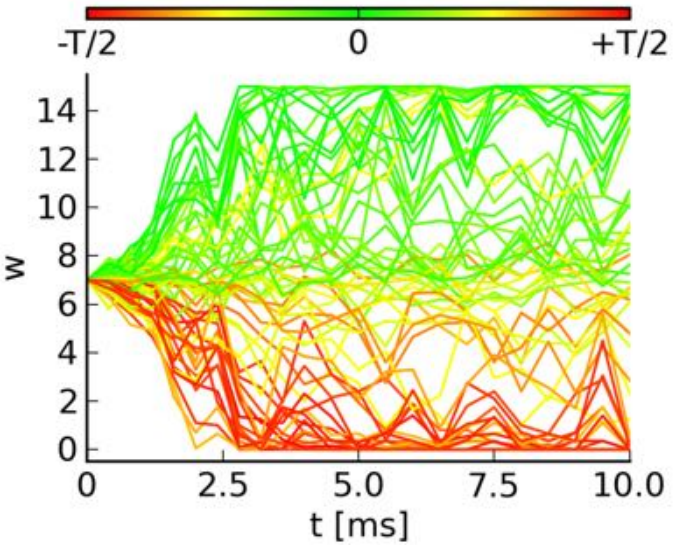
Application in generic multivariate data classification

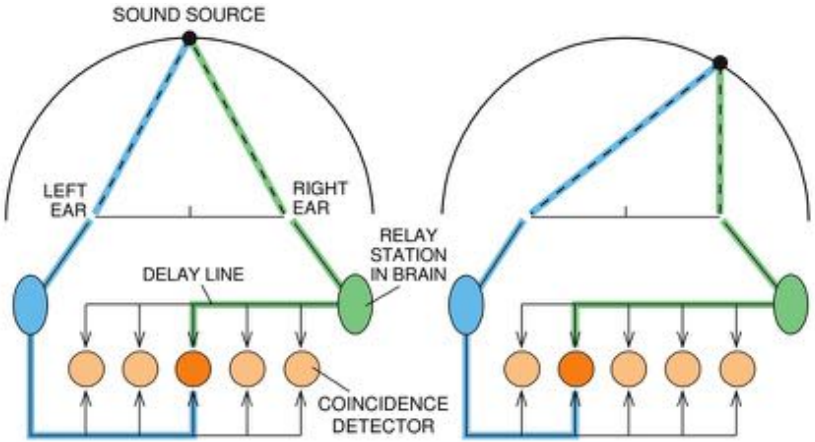


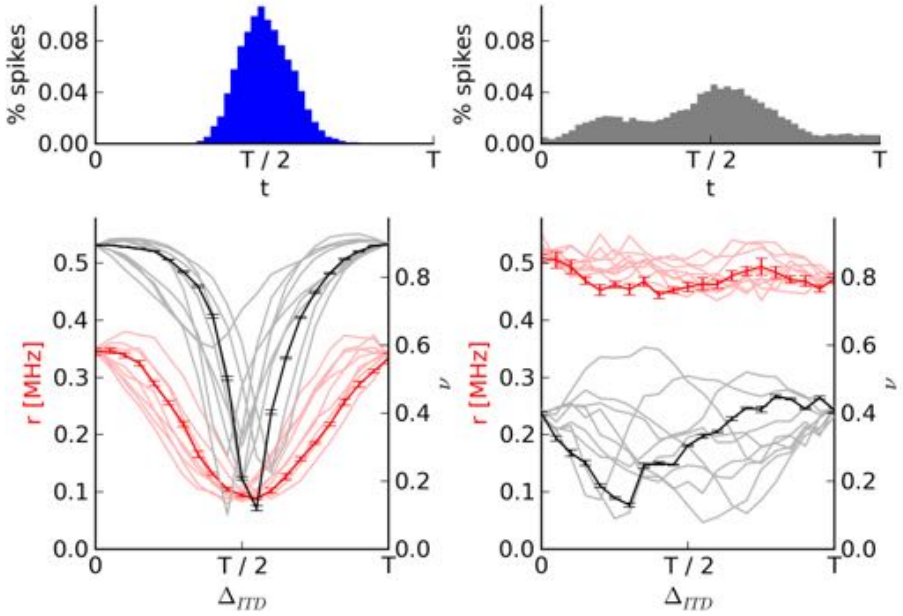
Schmuker, M.et al., "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

On-chip : Spike-Timing-Dependent -Plasticity

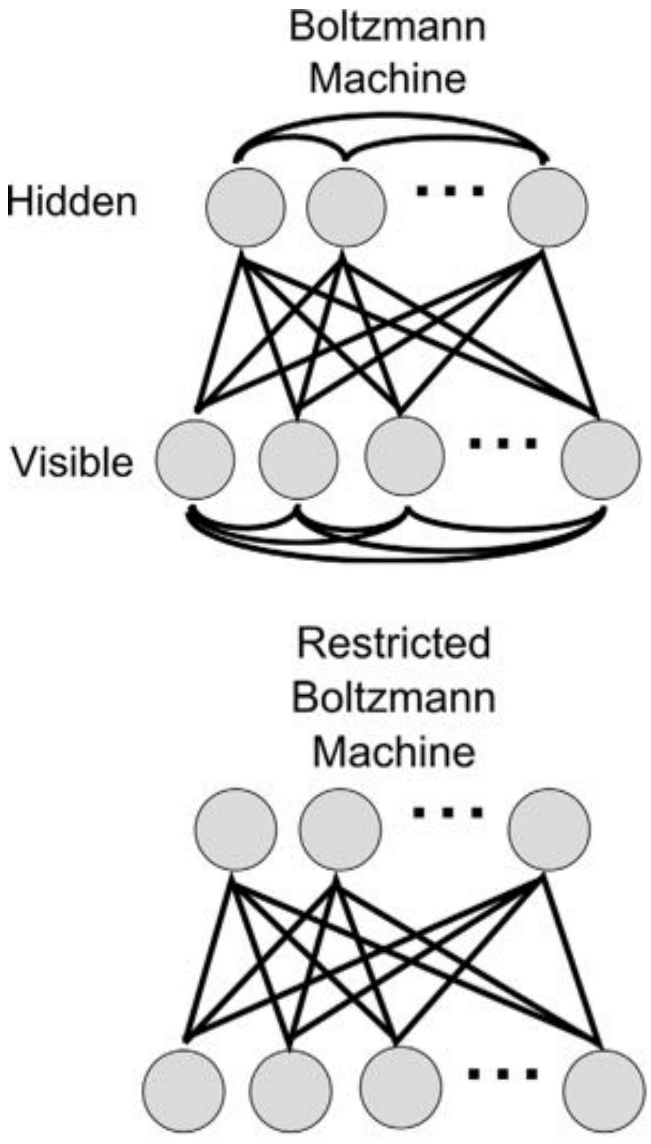








T. Pfeil, A.-C. Scherzer, J. Schemmel and K. Meier,
Neuromorphic Learning towards Nano Second Precision,
Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN).
Dallas, TX, USA: IEEE Press, 2013, pp. 869-873.



Boltzmann Machines

Networks of symmetrically connected stochastic nodes k

State of nodes described by vector of binary random variables z_k (0,1)

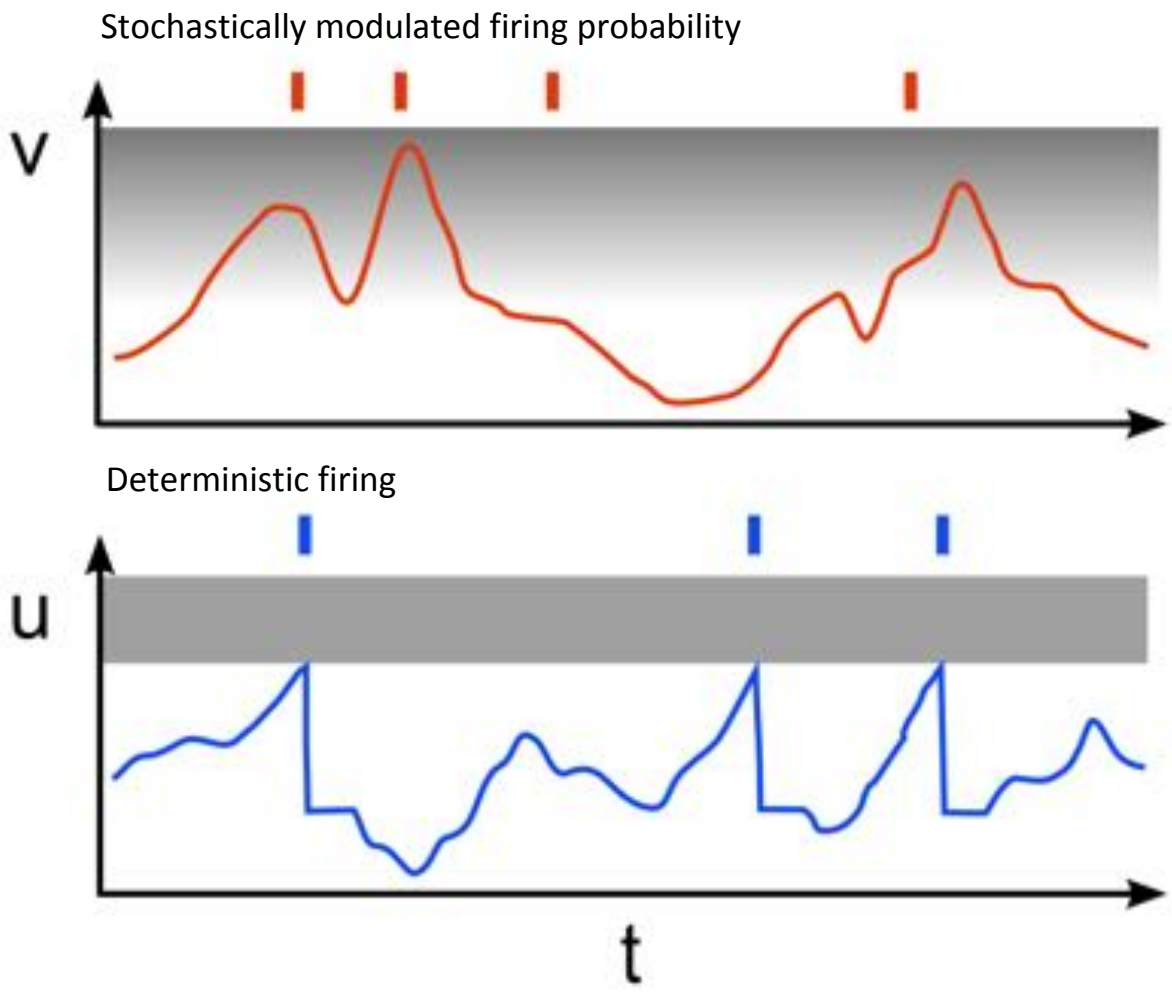
Probability for state-vector converges to a target Boltzmann-distribution

$$p(\vec{z}) = \frac{1}{Z} \exp\left[-E\left(\vec{z}\right)\right]$$

Energy function

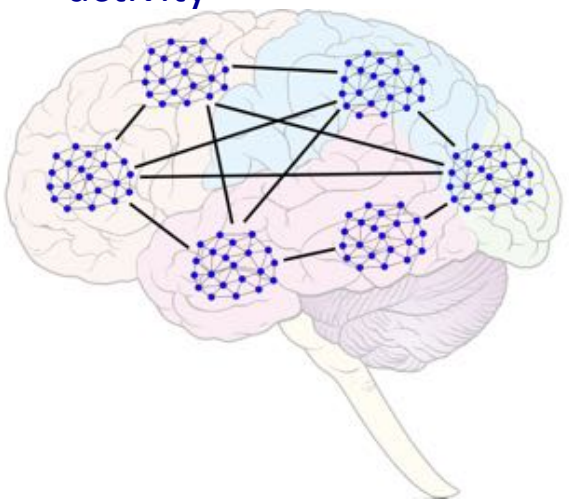
$$E(\vec{z}) = -\frac{1}{2}\sum_{i \neq j} w_{ij} z_i z_j - \sum_i b_i z_i$$

WHAT FOR ? Learn internal stochastic model of input space – Generate or discriminate



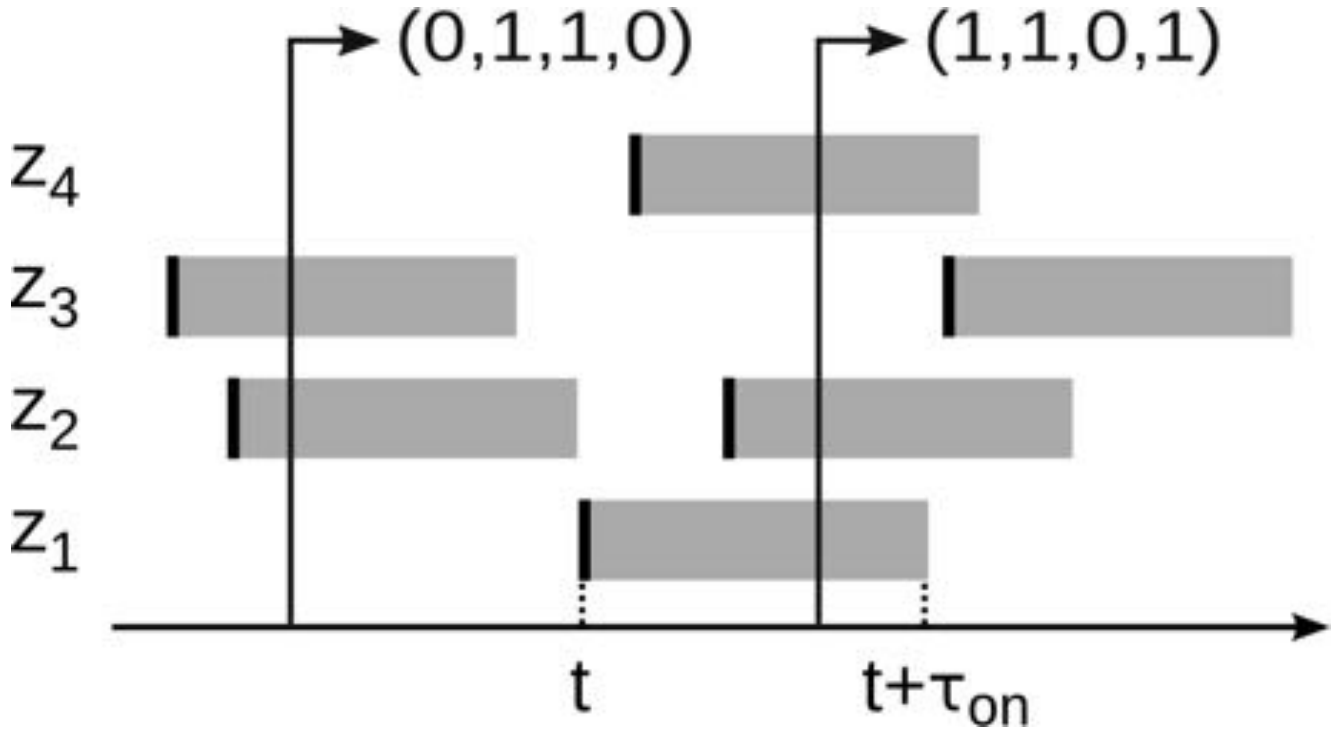
Stochasticity from

- External noise source
- Internal noise source
- Ongoing network activity



M. A. Petrovici, J. Bill, I. Bytschok, J. Schemmel, and K. M.:
Stochastic inference with spiking neurons in the high-conductance state. Physical Review E 94, 2016

Spiking LIF Neurons as a 2-state (binary) system



Neural computability condition

A network of spiking neurons draws samples from the joint distribution p if the membrane potentials u_k of the neurons follows

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

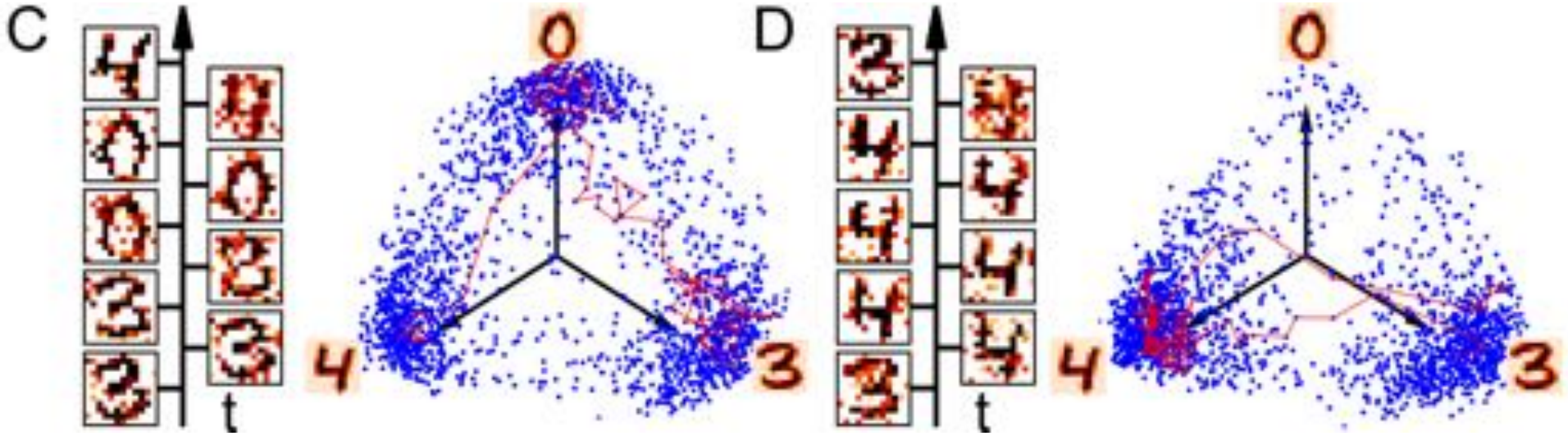
Corresponds to a logistic activation function of the neuron

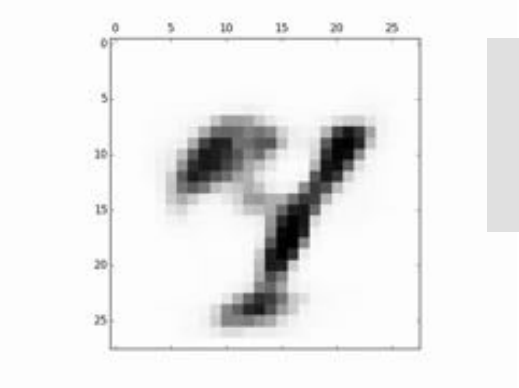
Petrovici, Mihai A., et al. "Stochastic inference with deterministic spiking neurons." *arXiv preprint arXiv:1311.3211*

Buesing, Lars, et al. "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons." *PLoS Comput. Biol* 7.11 (2011): e1002211.

Learning specific input distributions by adjusting LOCAL interactions

- Clamp visible units to value of particular pattern – reach thermal equilibrium
- Incremement interaction between any 2 nodes that are both on
- Run network freely and sample from stored probability distribution
- Infer from clamped input

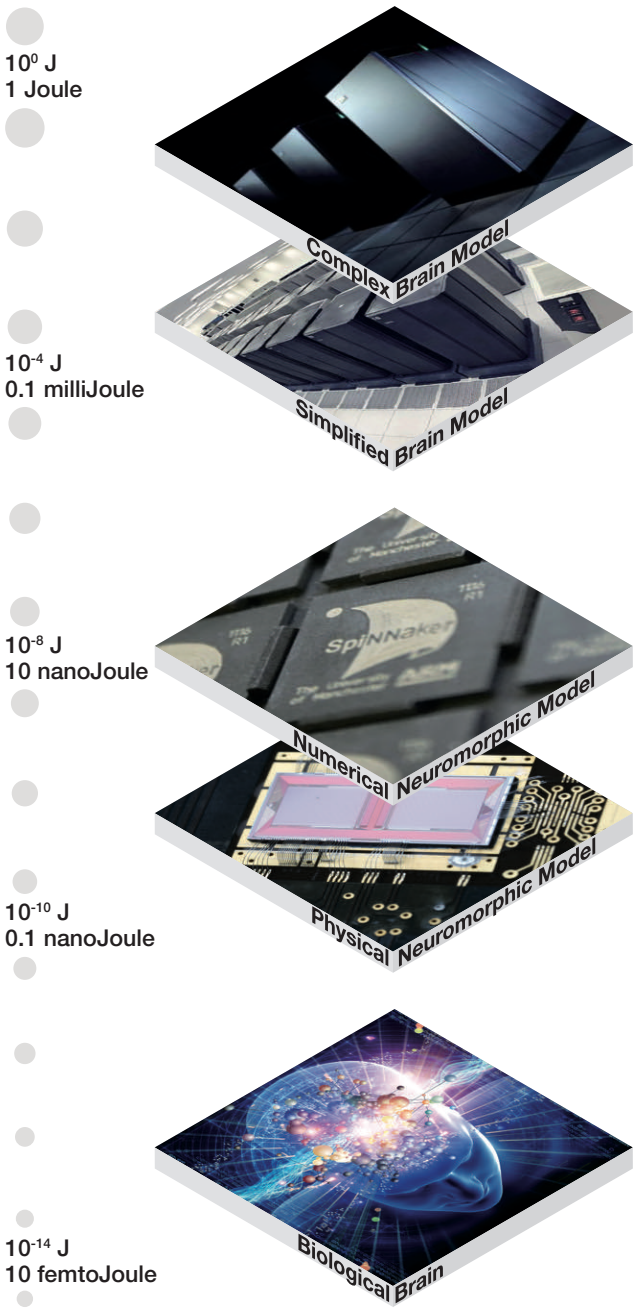




Free running
„Dreaming“
Generative

Inferring
Input incompatible with 0
Discriminative

PhD, Mihai Petrovici, BA Luziwei Leng 2015



E F F I C I E N C Y

Energy*Scales*

Energy used for a synaptic transmission

Filling the Gap

- Typically 10.000.000 times more energy efficient than state-of-the art HPC (comparable model)
- 10.000 less efficient than biology

From : HBP project report

TimeScales	Nature + Real-time	Simulation	Accelerated Model
Causality Detection	10^{-4} s	0.1 s	10^{-8} s
Synaptic Plasticity	1 s	1000 s	10^{-4} s
Learning	Day	1000 Days	10 s
Development	Year	1000 Years	3000 s
12 Orders of Magnitude			
Evolution	> Millenia	> 1000 Millenia	> Months
> 15 Orders of Magnitude			

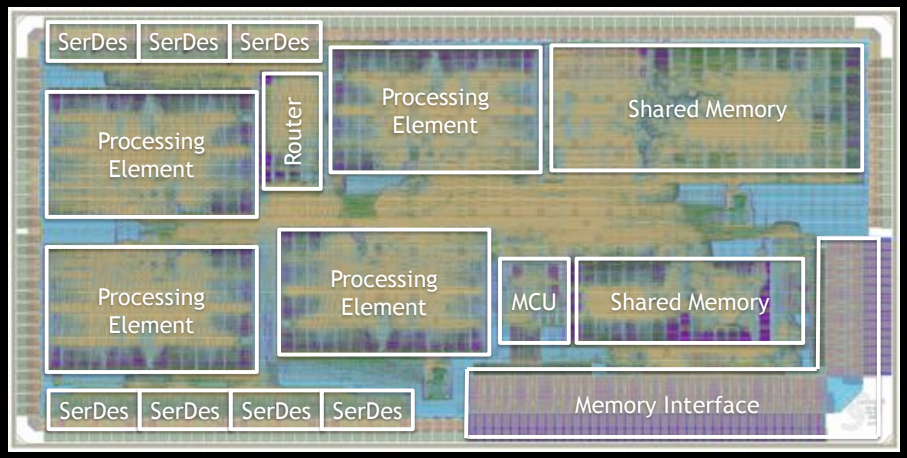
Next generation of NM computing in the HBP

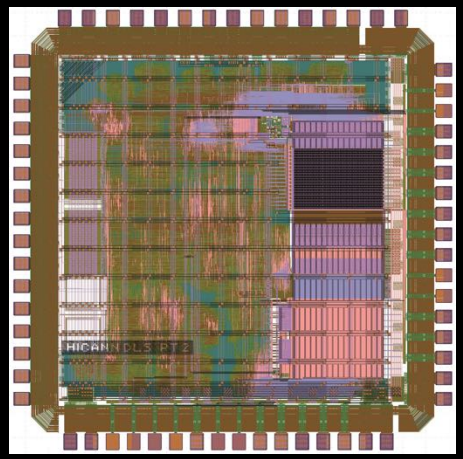
SpiNNaker-2

4-core Quad Processing Element
25 GIPS/W on a single die
Floating point precision
True random numbers

BrainScales-2

Flexible local learning
On-the-fly network reconfiguration
Structured neurons
Dendritic computation





Today : Working prototypes *2020 : Operational systems*

Goal : learning cognitive machines

Final Thoughts

- After 10 years of development the BrainScaleS large scale physical hardware system is being commissioned and delivers first results
- Fully non-Turing, physical model computing can solve established machine learning tasks
- 2nd generation physical model systems start to offer very advanced accelerated local learning capabilities and exploitation of dendritic computation

Goal : Build a continuously learning cognitive machine