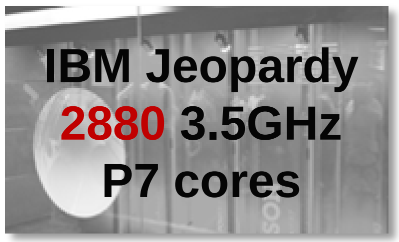
Efficient Neuromorphic Computing with the Feedforward Inhibitory Motif

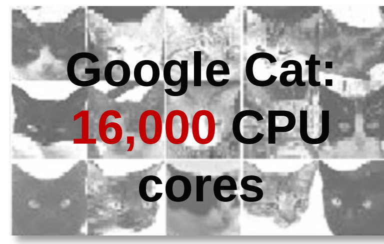
[1]Yu (Kevin) Cao, [2]Maxim Bazhenov,
[1]Jae-sun Seo, [1]Shimeng Yu, [3]Visar Berisha

[1]School of ECEE, ASU; [2]School of Medicine, UCSD;
[3]Dept. of SHS, ASU

Machine Learning Today

- A **top-down** approach: better for CPU/GPU
  - *Pros*: mathematical, accurate, scalable
  - *Cons*: **computation** cost, energy **efficiency**, off-line learning

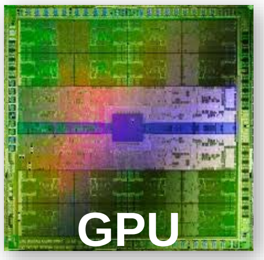




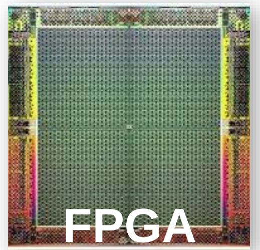
- **Edge computing** needs novel hardware/algorithms
  - **Local** to the sensor, **real-time**, **reliable**, low-power
  - **On-line**, personalized learning with continuous data

Acceleration Needs

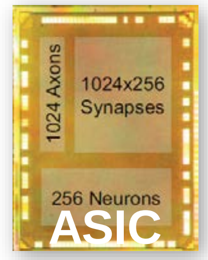
- $10^3 – 10^5$ speedup required to achieve real-time training of HD images at 30 frames/second



10 – 30 X



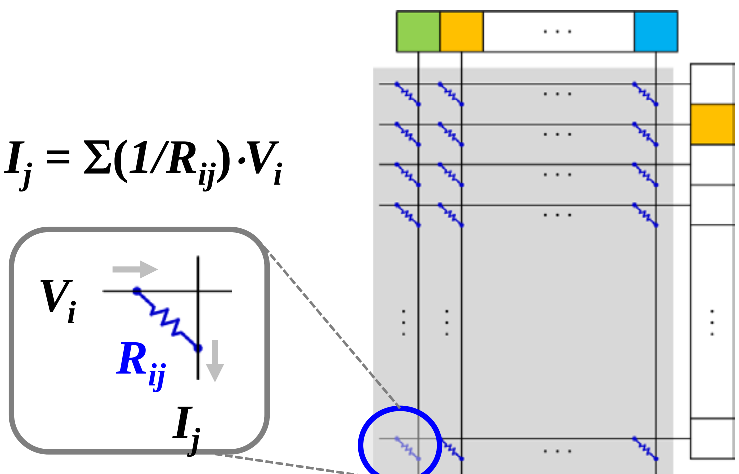
10 – 50 X

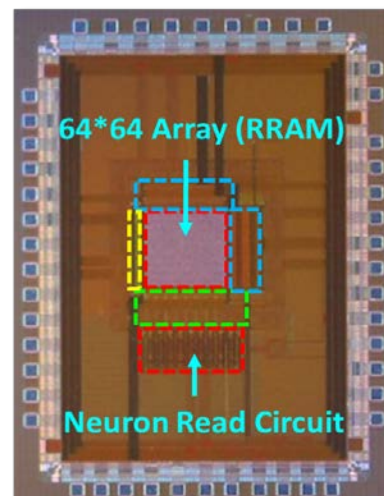


$10^2 – 10^3$ X

$>10^3$ X

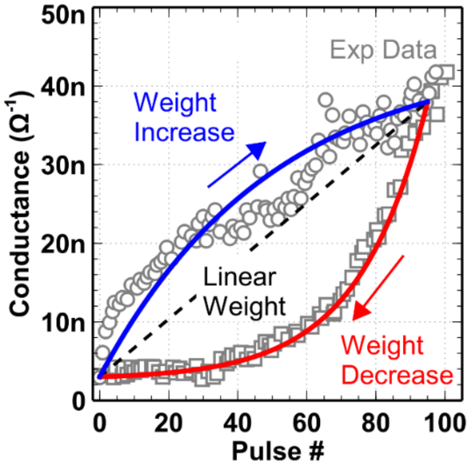
- Resistive Crossbar Architecture

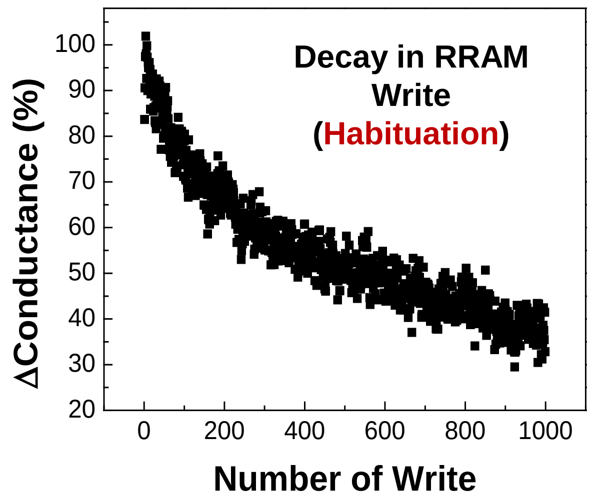


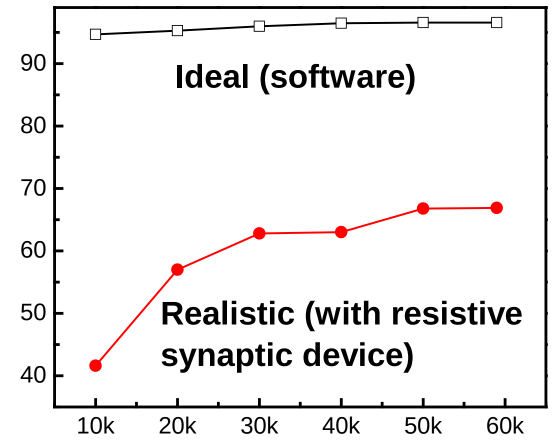


Physical Challenges

- **Nonlinear**, **noisy**, poor endurance (**habituation** in programming)







- These hardware problems (variations, unreliable synapse) and application demands (real time, on-line learning, and mobile) exist in **biological** cortical and sensory systems!

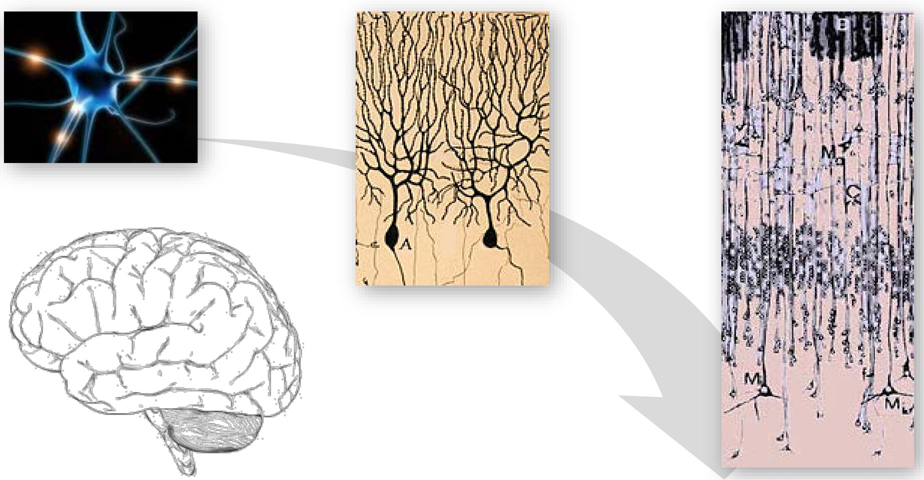
A bio-plausible hardware-algorithm solution:

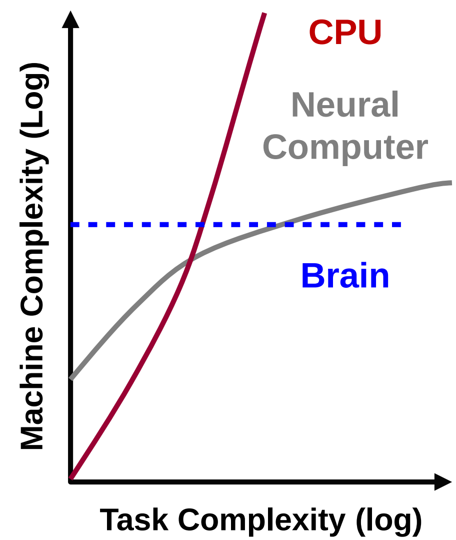
robust, low-power, low-precision, accurate, on-line

Brain-inspired Computing

- A **bottom-up** approach: better integration with sensors
  - *Pros*: **energy efficiency**, simpler computing, real time, reliable
  - *Cons*: complicated dynamics, limited scale and **accuracy**

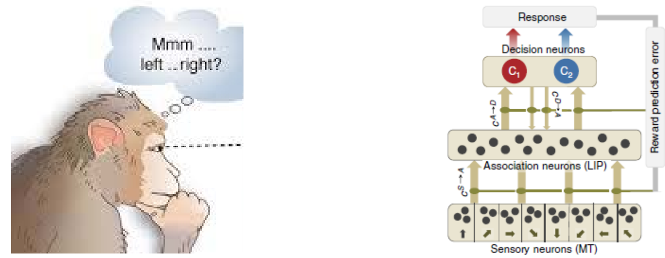
Neuron	Microcircuit	System
4-100μm	***FO = 1K-100K***	***100B, 100Hz, 20W, 30% ER/neuron***
[*22nm*]	[*FO = 4*]	[*1.4B, 3.7GHz, 45W, <10^{-9} BER*]



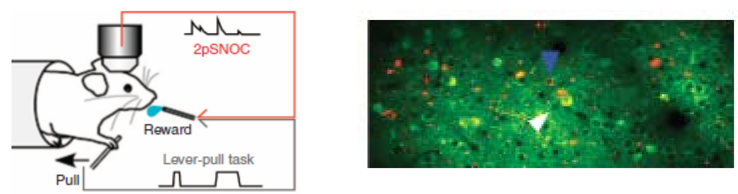


Neurobiological Basis of Learning

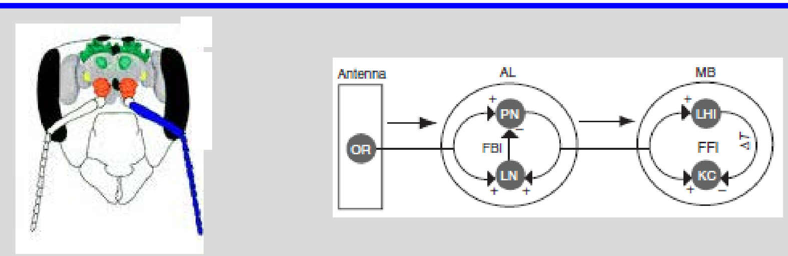
- **Reward** (supervision): global feedback signal
- **Inhibition**: unsupervised sparse feature extraction
- Synapse: non-linear, **habituation** (local), **noisy**
- Neurons: continuous leaky-integrate-fire
- Learning: local, feed forward STDP or SRDP on each plastic synapse



Monkey, Parietal cortex, *Nature Communications*, 2015



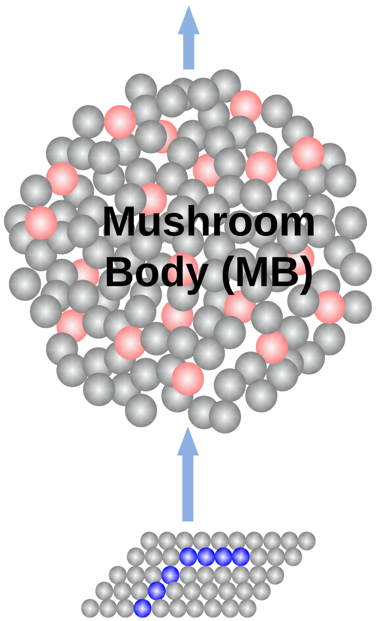
Mouse, Motor cortex, *Nature Communications*, 2014



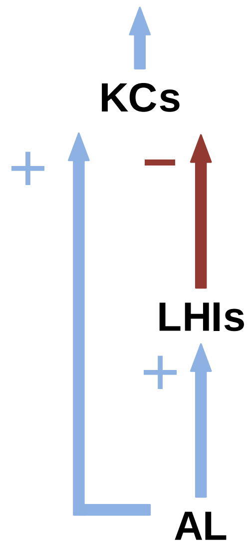
Insect, olfactory system, *Nature Neuroscience*, 2007

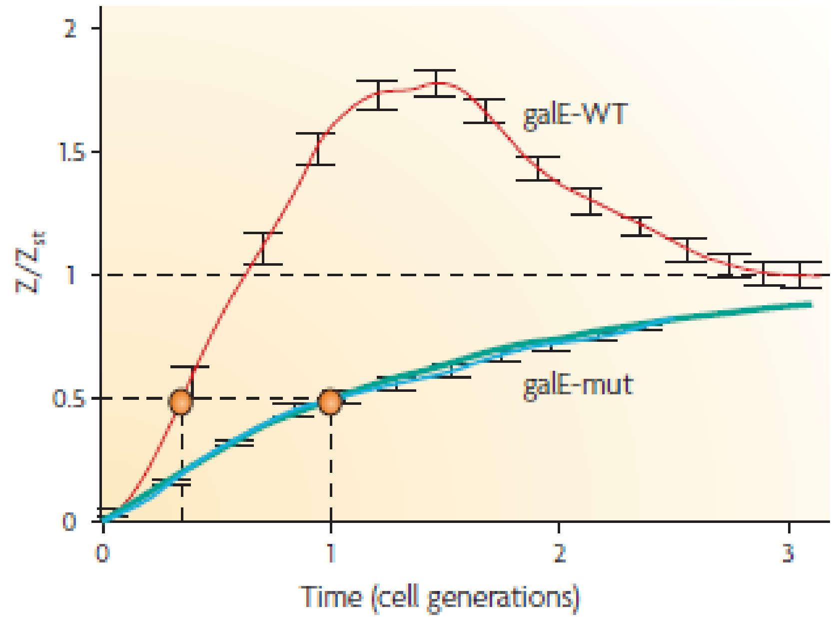
RHINO: A Biomimetic Solution

- **R**eward, **H**abituation, **I**nhibition, **NO**ise
- **Motif**: a recurring network element; general in biological process



Antennal Lobe (AL)





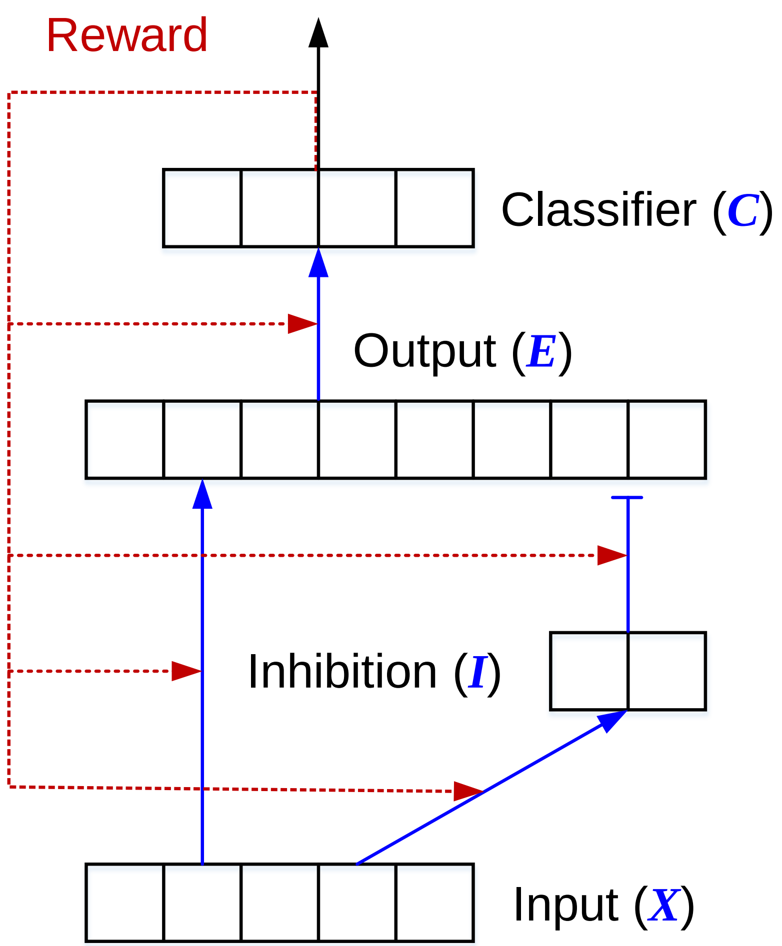
Kenyon Cells (KCs) 15,000

Lateral Horn Interneurons (LHIs), 100

[Nature Review, 2007]

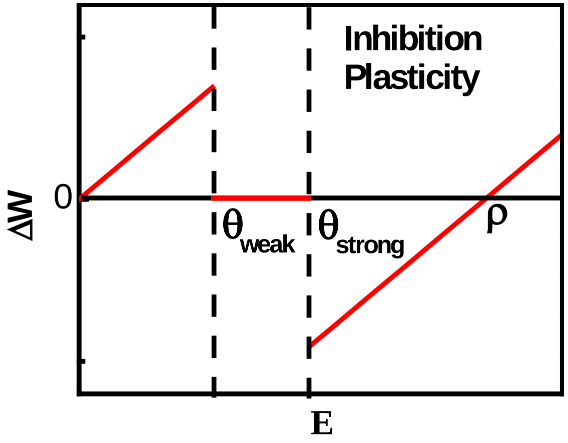
Network Structure

- Rewarding for associative (supervised) learning
- Inhibition to speed up the formation of sparsity
- Habituation (decay in learning rate) to achieve the convergence
- **Non-gradient** based: no backward propagation
- **Local adaptation**: no crosstalk among synapses



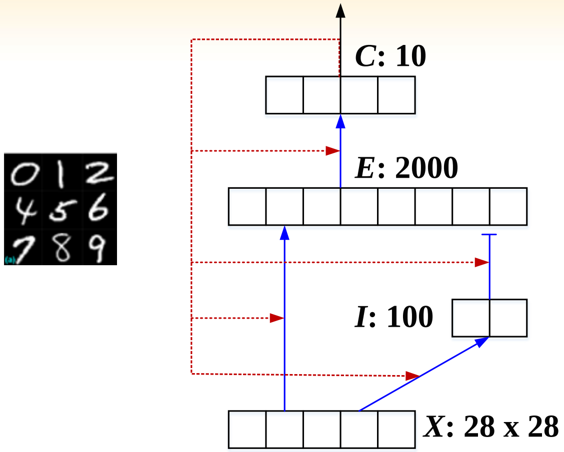
Learning Rules

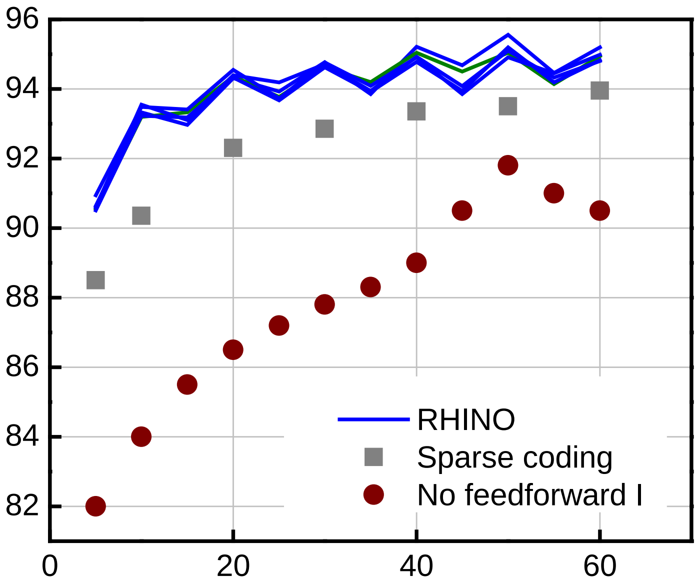
- Reward: A global feedback to all W's
  - C: classification score
  - Correct: $|C\text{-}C_{th}|$; Punish: $-|C\text{-}C_{th}|$; If $|C\text{-}C_{th}| < \theta_C$, no reward feedback
- Classification: Punish only
  - $\Delta W \propto -(C\text{-}C_{th})E/\eta$ for wrong classification
- Excitation: Hebbian learning rule with **habituation**
  - $\Delta W \propto \text{Reward}\cdot E\cdot(X-\theta_{XE})/\eta$
  - η: learning rate decays with training, i.e., habituation per synapse
- Inhibition: **positive feedback** on E
  - If $E < \theta_{weak}$, $\Delta W \propto \text{Reward}\cdot E\cdot I/\eta$
  - If $E > \theta_{strong}$, $\Delta W \propto \text{Reward}\cdot(E-\rho)\cdot I/\eta$
- Neuron: spiking leaky-integrate-fire

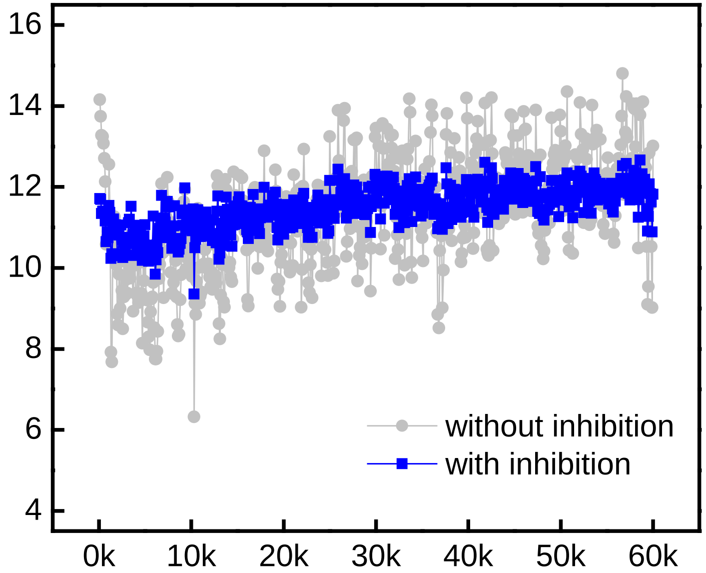


Demonstration: MNIST

- MNIST for handwriting recognition
  - Data represented by 0 – 50 spikes
  - Full image 28 x 28
  - No pooling or normalization
  - 50% connectivity of W_{X2E} and W_{X2I}

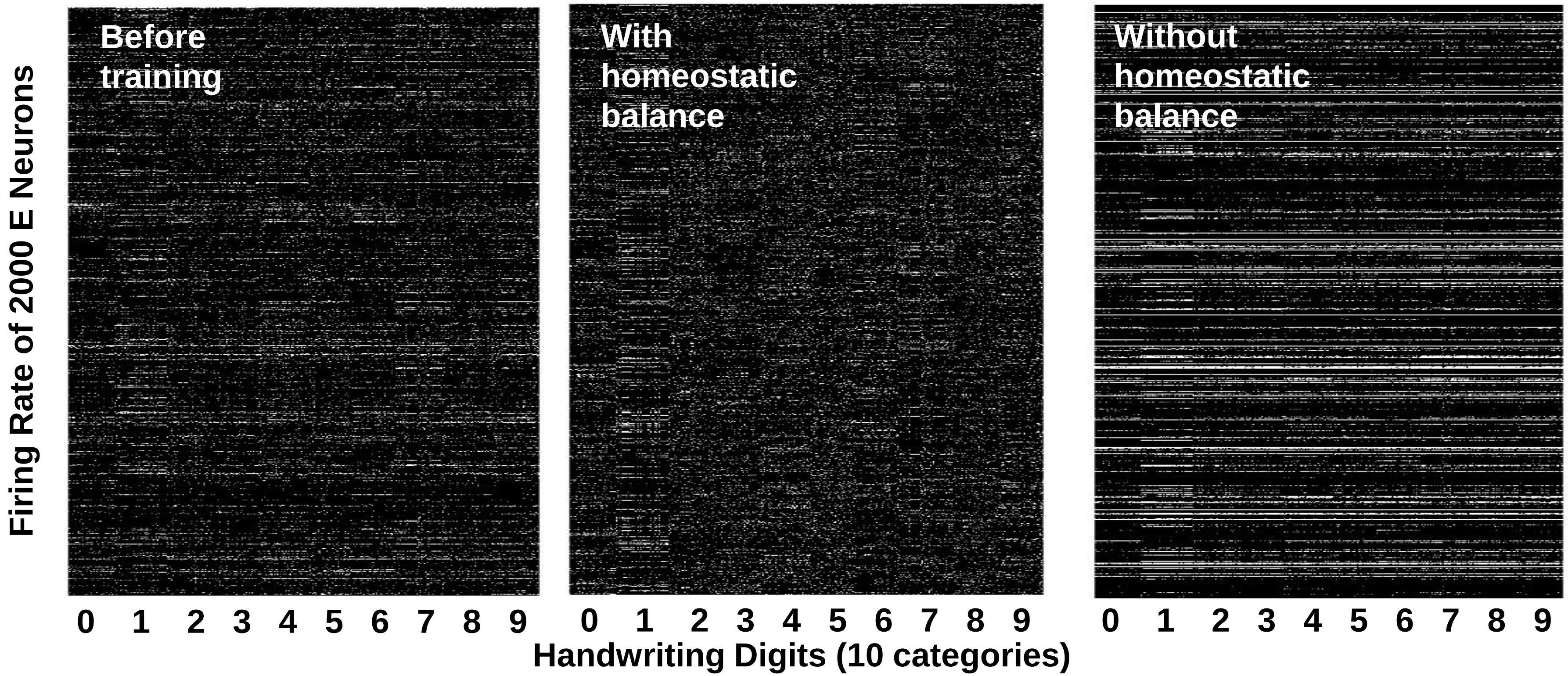






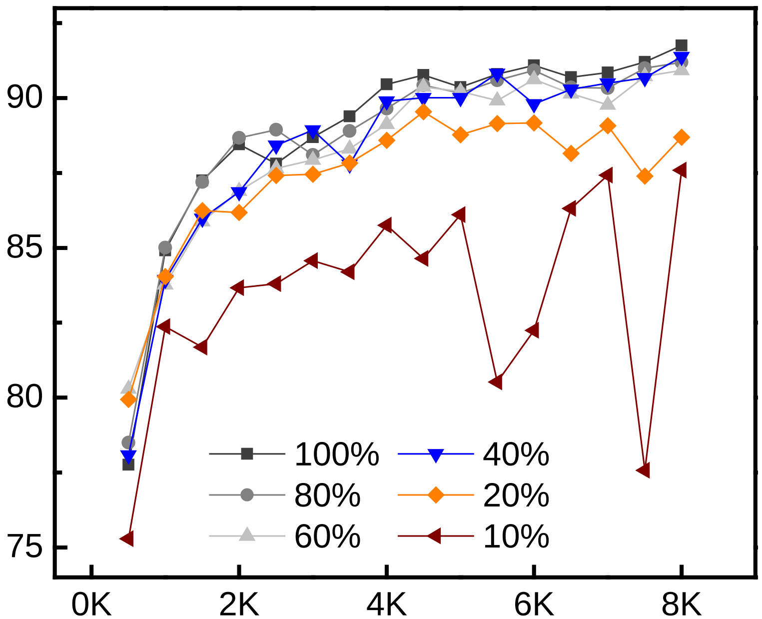
Neuron Firing Rate

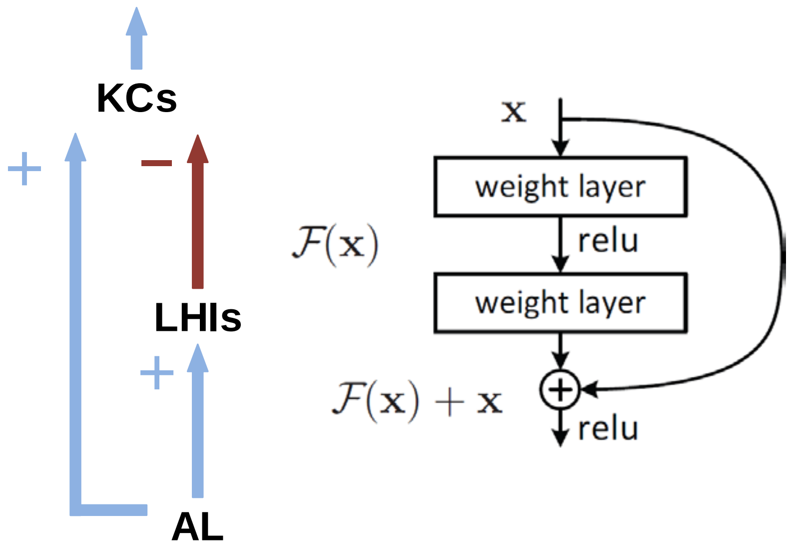
- **Sparsity**: an appropriate range (5-15%) is critical
- Homeostatic **balance**, which controls overfiring of the output neurons, is essential for learning



Factors for Learning Accuracy

- Initial **randomness**: without noise, learning cannot start
- With 100 Is, the network size of E is reduced by **3X** at the same accuracy of **95%**; **~50X** speedup over gradient-based approaches





[Microsoft, 2015]

Results Comparison

Reference	Input	Data format and precision	Learning rules	Number of neurons	Number of parameters	Number of images	Accuracy
Mushroom body	28x28	Spike	Rewarded STDP	50000	5E5	60000	87%
Two layer SNN	28x28	Spike	STDP	300	2.4E5	60000x3	93.5%
Unsupervised SNN	28x28	Spike	STDP	6400	4.6E7	200000	95.0%
This work	28x28	Spike rate in a 50 window	Rewarded SRDP	2100	8.4E5	60000	95.0%
This work	28x28	Spike rate in a 50 window	Rewarded SRDP	6000	2.4E6	60000	96.2%
Spiking RBM	28x28	Spike rate	Contrastive divergence	500	3.9E5	20000	92.6%
Sparse Coding	10x10 patch	3-bit number	Gradient	300	3E4	60000x10	94.0%
Two layer NN	28x28	Floating number	Gradient descent	1000	7.8E5	60000	95.5%
Spiking CNN	28x28	Spike timing	Regenerative learning	5.6E4	1.2E5	60000	99.08%

Summary

- RHINO: A bio-plausible spiking NN
  - Feedforward **inhibitory motif**
  - Reward + **Local** adaptation
- What matters to **efficiency**: spiking, precision, motif,…?
- Algorithms
  - Multi-layer, hierarchical
  - **Low-precision learning**
  - On-line learning
- Hardware
  - Implementation with resistive synaptic array
  - Reliable learning