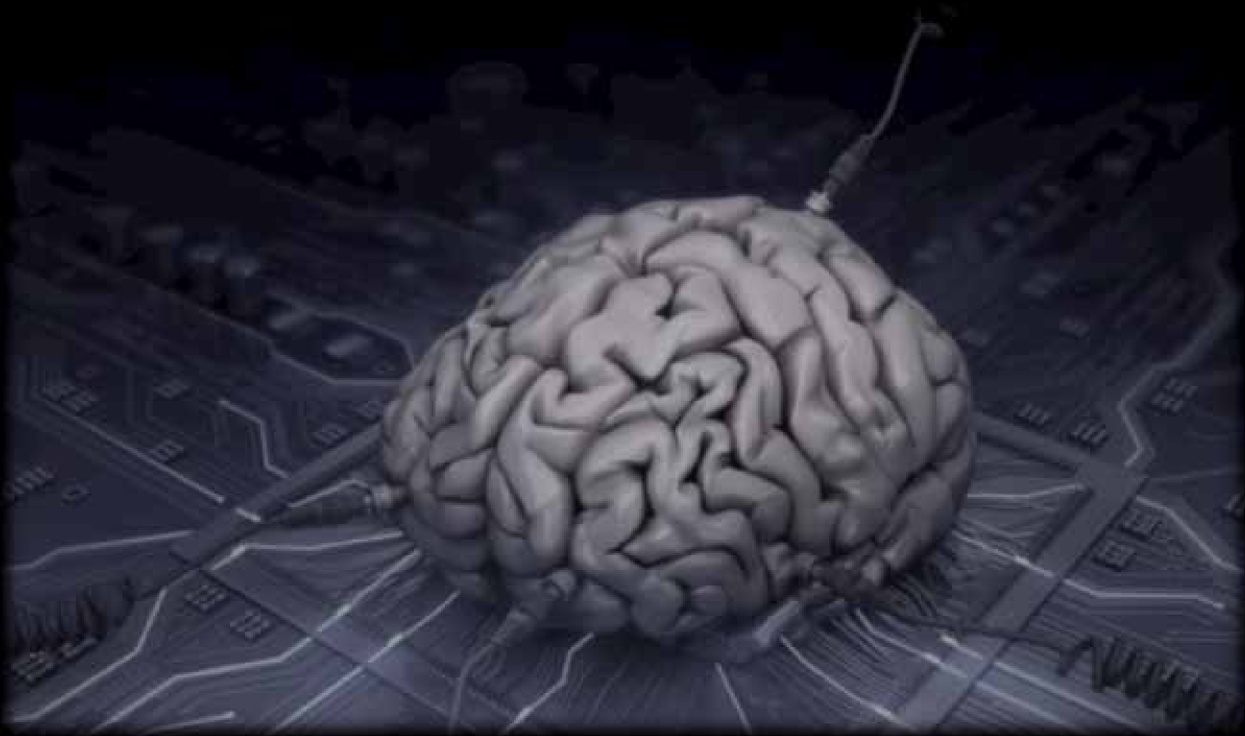
The Return of Neuro-Inspired Computing

Why Now?



Credit: iStockphoto/Andrey Volodin

Jan M. Rabaey
University of California @ Berkeley

Lund, Sept 11 2014

How Viktor is keeping me on my toes!

Talks at Lund over the past decade

SOS

- 2014:The Return of Neuro-Inspired Computing - Why Now?
- 2013: Innovation is in the Mind (Mind of Innovation Conference)
- 2012: The Wireless Revolution Continues – From Mobiles to Swarms (Hon. Doctorate)
- 2011: The Swarm at the Edge of the Cloud – A New Face of Wireless
- 2009: Exploring the Boundaries of Ultra-Low Power Design - Microscopic Wireless

CCCD

- 2007: Design without Borders (A Tribute to Richard Newton)
- 2005: Traveling the Wild Frontiers of Ultra Low-Voltage Design
- 2004: Design in the Late-Silicon Age
- 2001: Picoradio – LP WSN

A Pertinent 21st Century Question …

"How to perform high-fidelity efficient computing on platforms that feature huge numbers of lousy components (aka nano-devices)?"

One Plausible Answer:
Abandon Determinism

Neuro-inspired scalable computational paradigms based on statistical inference, massive redundancy, and low resolution

Not a novel idea. Many have tried and failed …

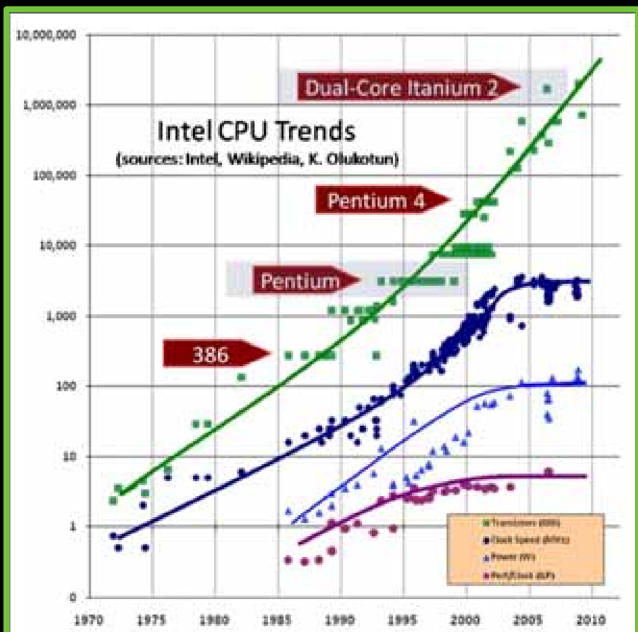
Why Now?

Recurring Waves of Neuro-inspired Computing

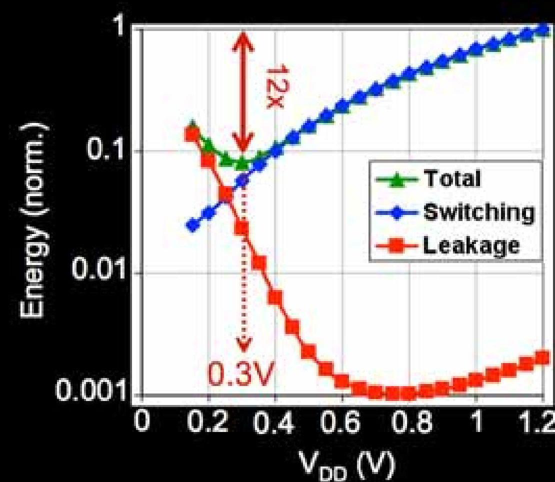
- Wave 1: forties – sixties
  - McCullough-Pitts, Hebbian learning
  - Ended with Marvin Minsky's paper (1969)
- Wave 2: eighties – nineties
  - Re-emergence of ANNs – Hopfield networks
  - Spurred by Carver Mead (neuromorphic)
- Wave 3: zeros – tens
  - Better understanding of neural functions
  - The success of deep learning (Google Brain, Watson)
  - Emergence of nano-devices

A Need for Novel Computation Models

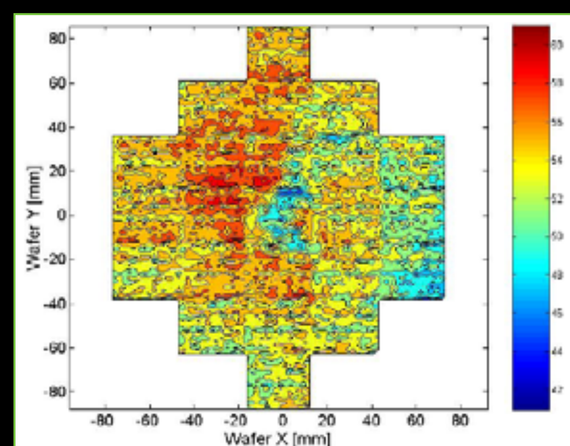
The waning days of Moore's Law



Speed, energy and efficiency
(and economics) plateauing



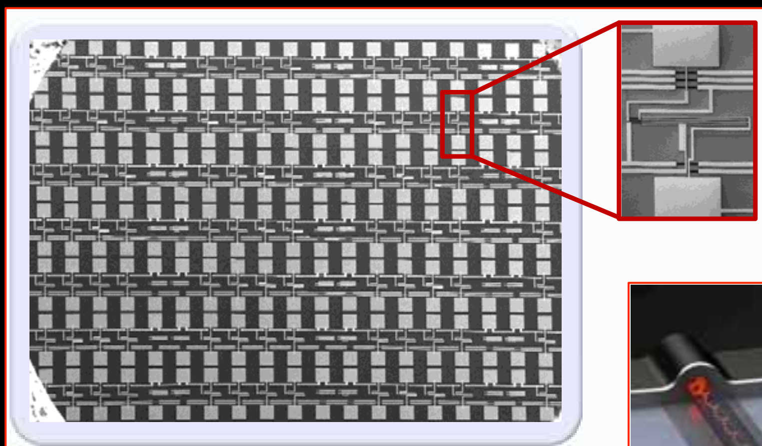
Energy Minimum Set by Leakage



Variance and uncertainty dictate operational margins

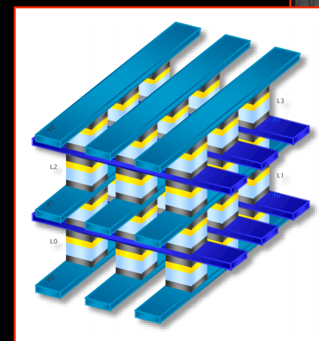
A Need for Novel Computation Models

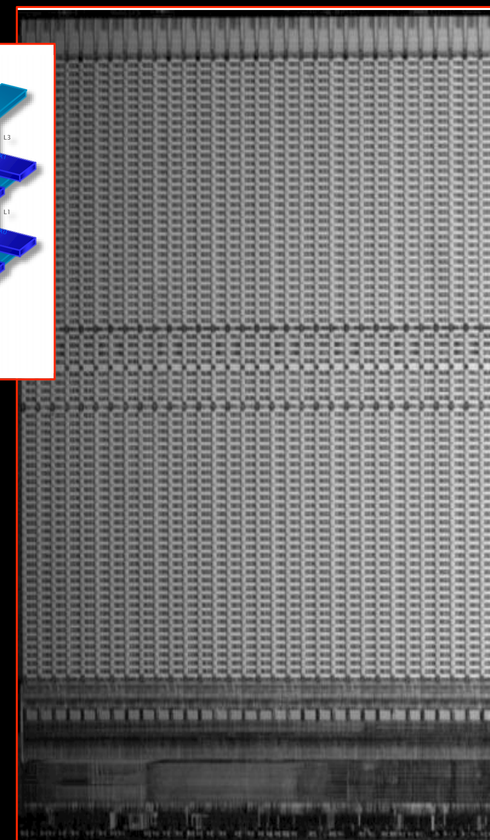
Emergence of Nano-devices



CNT microprocessor
[Courtesy: Mitra, Wong, ISSCC13]

Others: TFETs, Graphene, Spin, DNA, organic, True 3D …





16/32 Gbit RRAM
[ISSCC 2014]

Main challenges: reliability, variability, performance/energy, …

A Need for Novel Computation Models

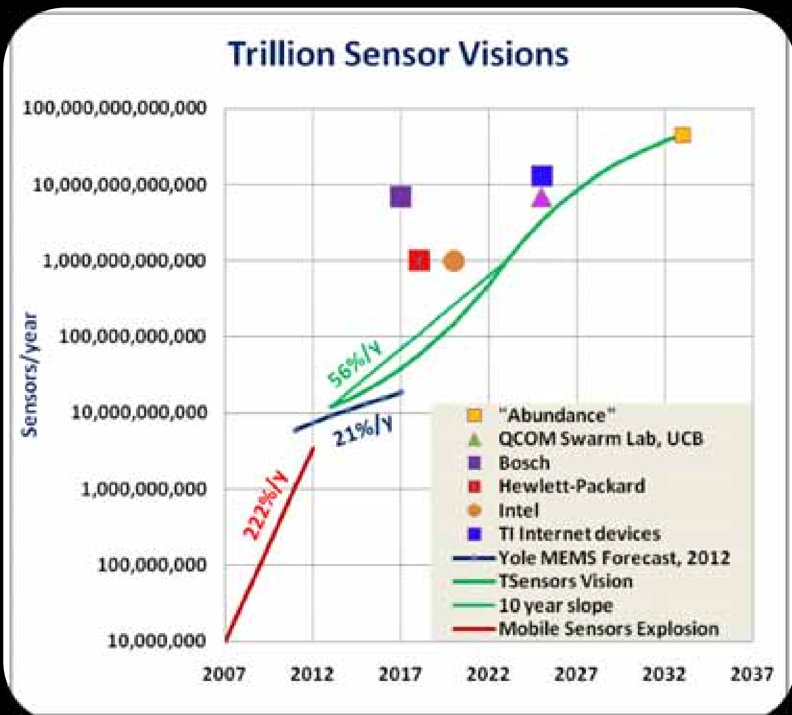
Data-Abundant Computation

From Big Data to Big Knowledge

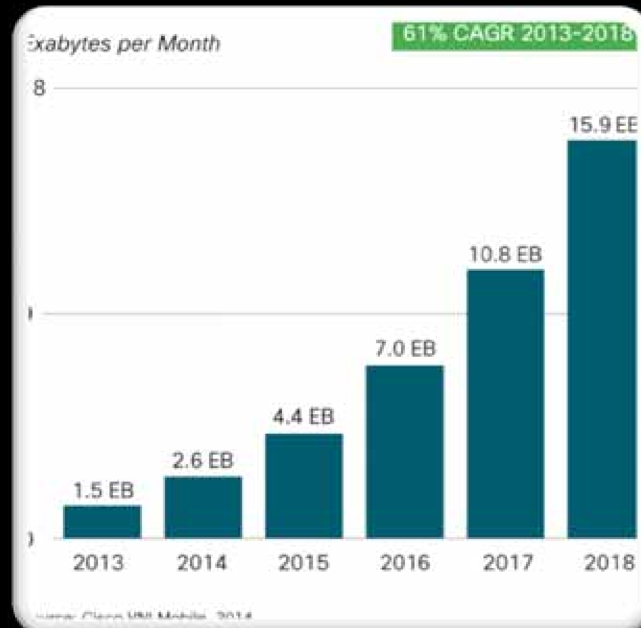
- Interactive analysis of "abundant data" using machine learning kernels
- Requiring 1000's of servers consuming MWs of power today
- Need memory-centric architectures

A Need for Novel Computation Models

The data deluge



Tsensor Summit 2013 [J. Bryszek]



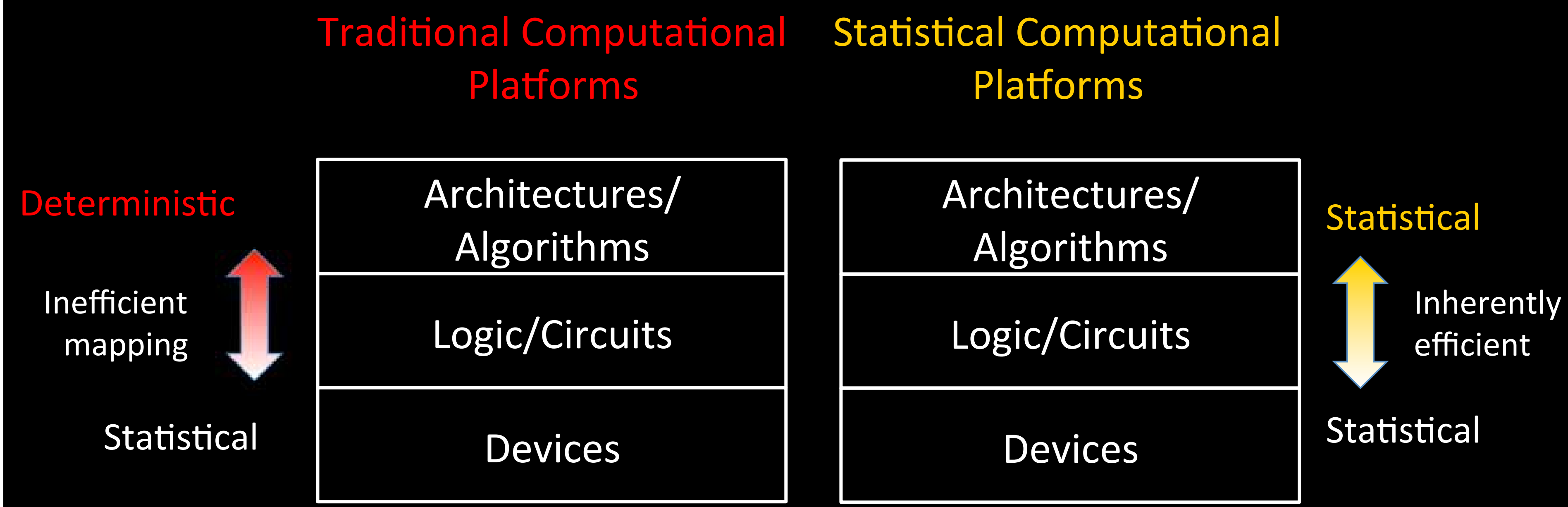
15.9 Exabyte/month of mobile data by 2018 [Cisco14]

Transmit sensory information as knowledge rather than raw data

Requires energy-efficient processing at the source

Neuro-Inspired Statistical Computing as an Attractive Alternative

Alternative Computational Paradigms



Functional non-determinism present in most applications related to human-cyber interfaces

(feature extraction, classification, synthesis, recognition, decision making, learning)

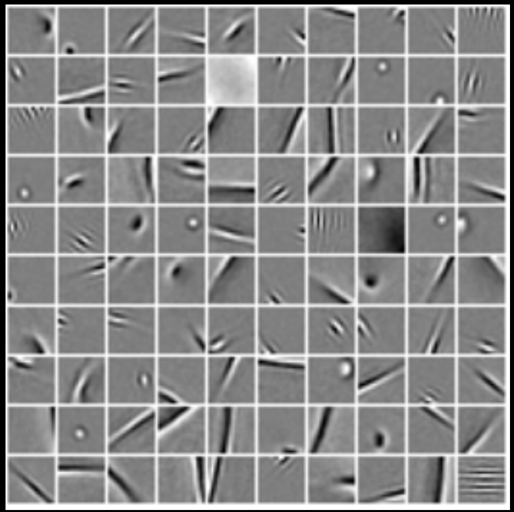
Features of (Bio) Neural Computation

- 2-3 orders more efficient than today's silicon equivalent (>10^{16} FLOPS with ~20 W)
- Robustness in presence of component failure and variations
  - Neural response is highly variable ($\sigma/\mu \approx 1$) [Faisal]
- Amazing performance with mediocre components
  - Auditory system: can tell difference of time arrival within 10 μs with cells having time constant of 1ms [Sarpeshkar]
  - Olfactory system: can discriminate 10^4-10^5 odors with slight difference of chemical structure with olfactory receptors having broad reception range [Buck]
- Seamless interaction with the physical world

In other words, welcome to Nanotechnology

Opportunity of Neuro-Inspired Computing

- Exploit properties of neural systems
  - Massively parallel, high density, major redundancy (hyper-dimensional)
  - Low resolution (SNR) processing
  - Efficiency through sparsity
  - Robustness through exploitation of randomness and variability
  - Adapting to variations through learning



Overcomplete representation

- To efficiently realize some hard cognitive problems
  - E.g. Artificial Olfaction, Vision, Classification, Detection, Decision making
- While mitigating the properties of deeply scaled nanometer CMOS or post-CMOS devices (CNT, Graphene, MEMS, RRAM, Spin, PC, …)
  - Large numbers of devices, possibly in multiple layers (3D)
  - Intertwined memory and computation
  - Huge variability and fault-density

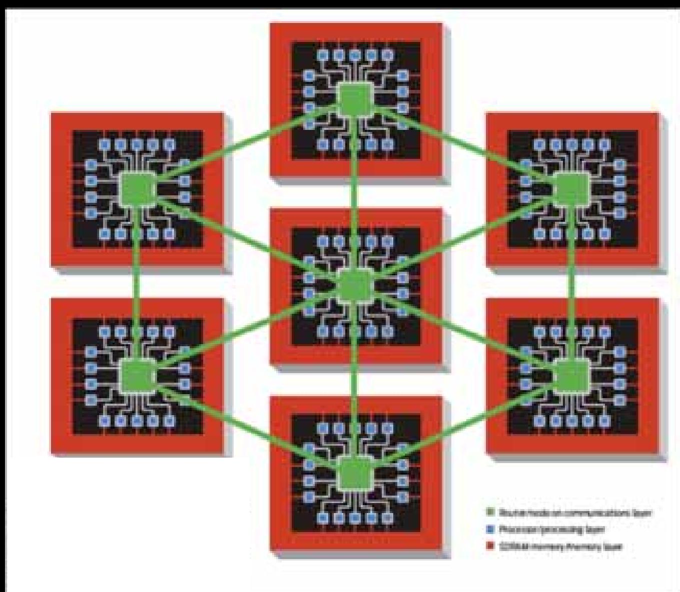
Some Resonance in Industry

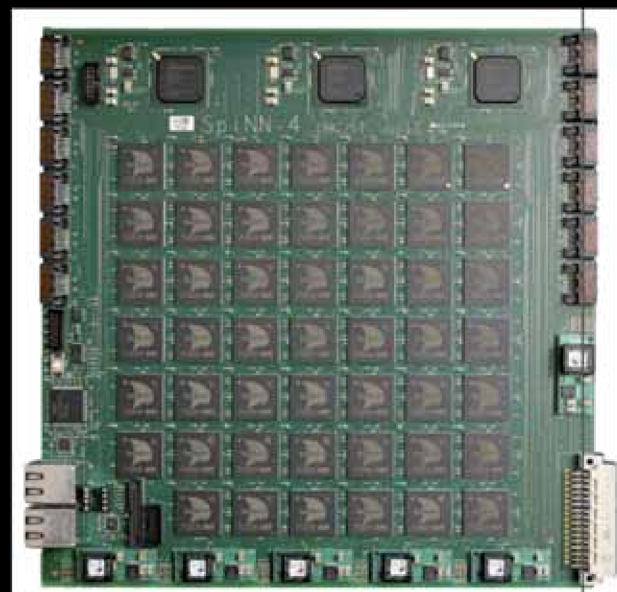
- Qualcomm Neural Processor
- Google BRAIN
- Intel "Approximate Computing"
- IBM Watson
- Micron Probabilistic Graph Processor
- Various start-ups (e.g. Nervana Systems)

Neuro-inspired: What its is not!

Neuromorphic computing

- reconstructing the brain bottom-up
- Mostly intended to be a simulation and modeling tool





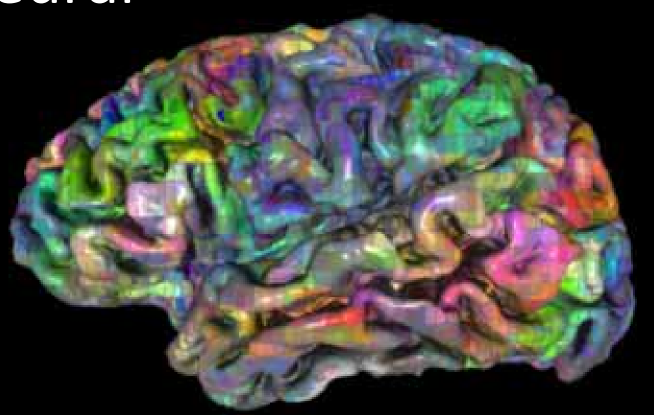
Example: SpiNNaker (Manchester)
1 million ARM9 processors, 100 kW,
1 billion neurons

Others: Blue Brain (EPFL), IBM Almaden, Neurogrid (Stanford)

Note: The human brain houses 100 billion neurons and 1 quadrillion synapses!

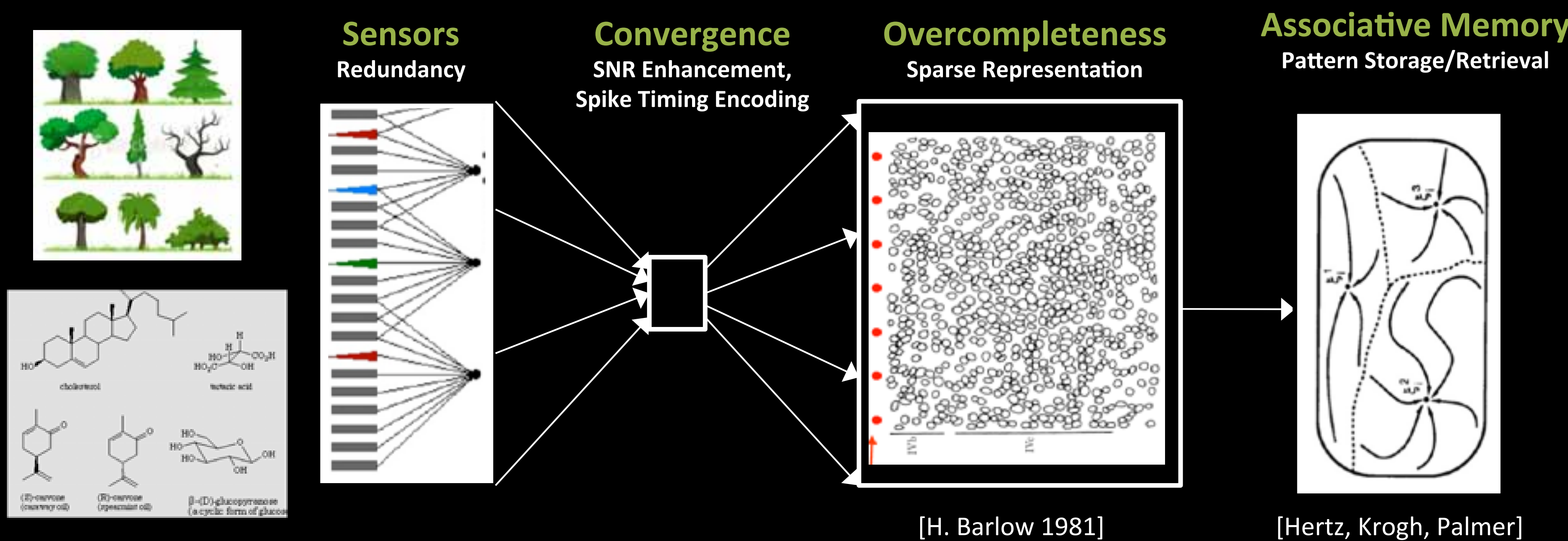
How to Gain Insights?

- Study what the brain does, and how well it does it (psychophysics/behavior)
- Study the brain's anatomical structure and neural response properties (neuroanatomy/physiology)
  - Improved imaging/BMI techniques to provide insights
- Formulate theories and test against neural data and performance (computational modeling)
  - Collaboration between computational neuroscience and engineering



[Courtesy: B. Olshausen, UCB]

The Sensory Pathway

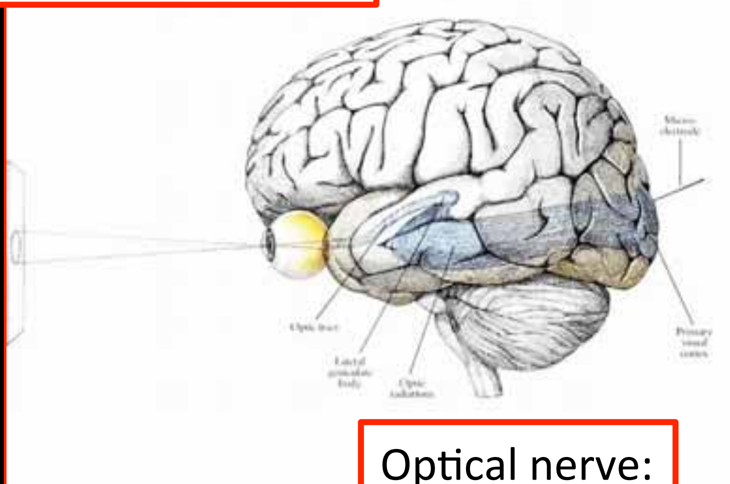


	Sensors	Convergence	Overcompleteness	Associative Memory
Visual	Retina	Ganglion Cells/LGN	Primary Visual Cortex (V1)	Higher-level Cortex
Olfactory	Olfactory Epithelium (OE)	Olfactory Bulb (OB)	Primary Olfactory Cortex	Higher-level Cortex

It's all about data representations!

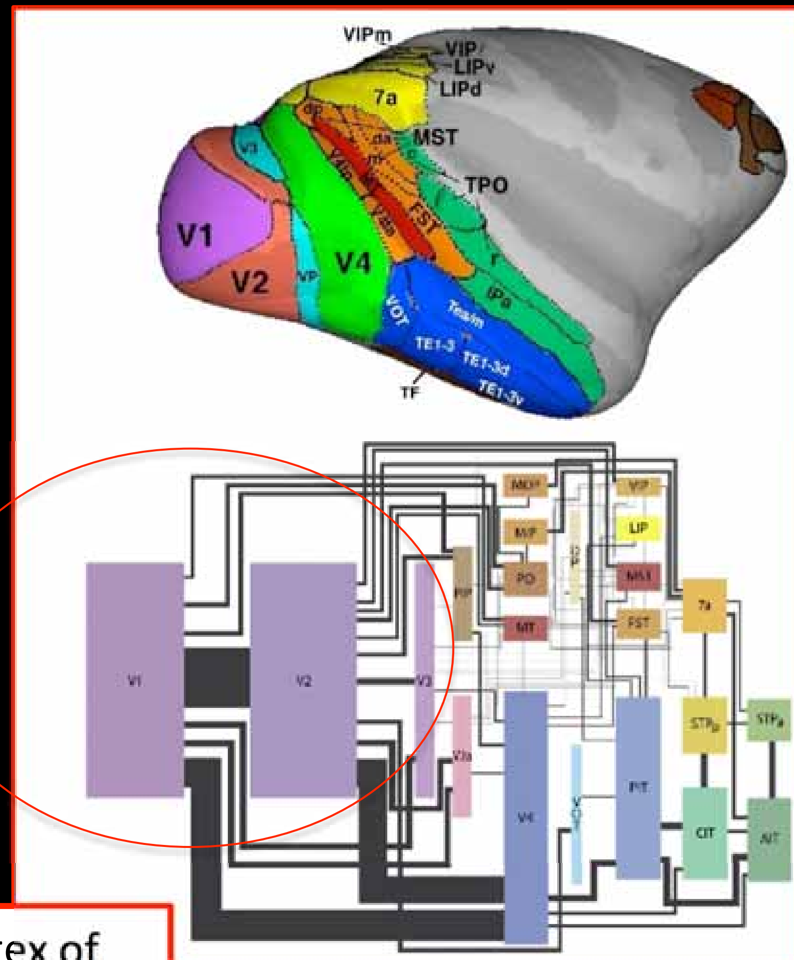
Sparse Representations and Coding

Retina
130 Million photoreceptors



Optical nerve:
1 Million fibers
10-100 Mb/sec

Date compression in retina



Massive expansion in V1 and V2

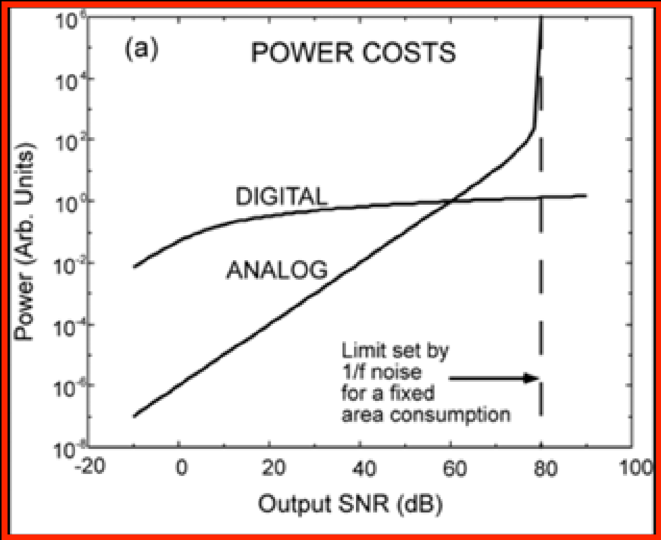
Visual cortex of
Macaque monkey

[Courtesy: B. Olshausen, UCB]

Various forms of data representations

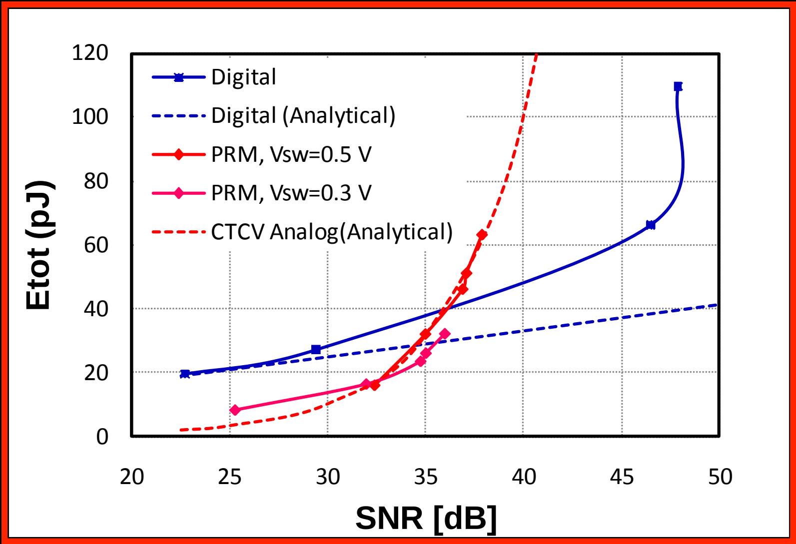
Low Precision Representations

Computation



[R. Sarpeshkar, Ultra-Low Power Bioelectronics,2010]

Communication

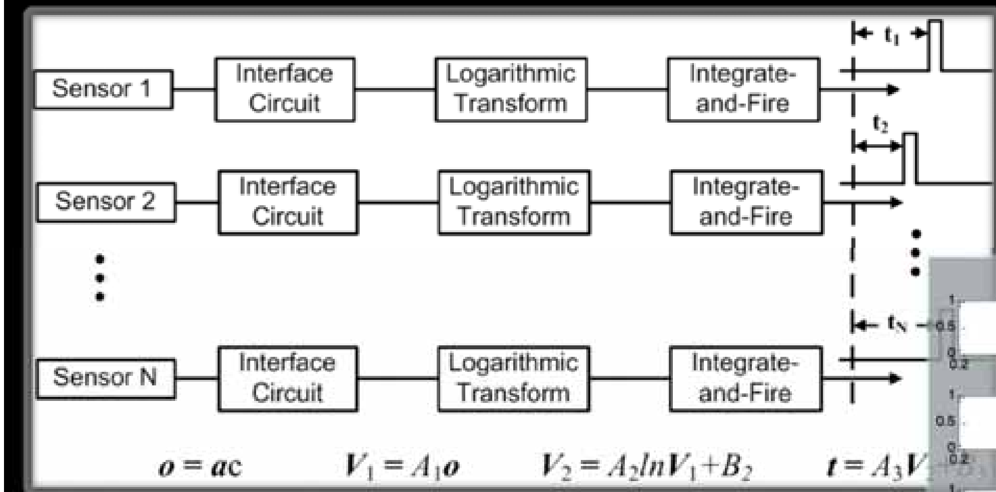


[PC. Huang, SIPS 2011]

Digital is supreme when high precision is needed, while analog (voltage, time) is most efficient at low SNR

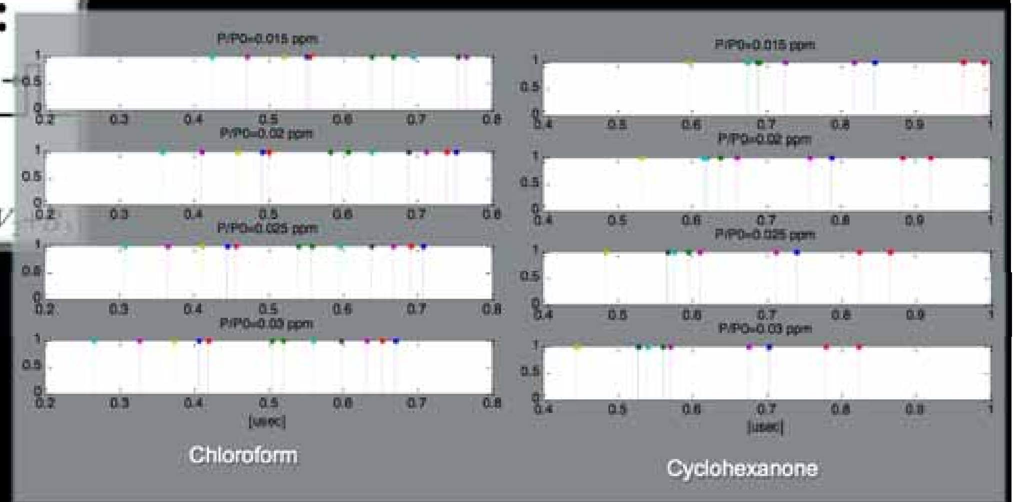
Use of slow (digital) feedback moves analog curves further to the right

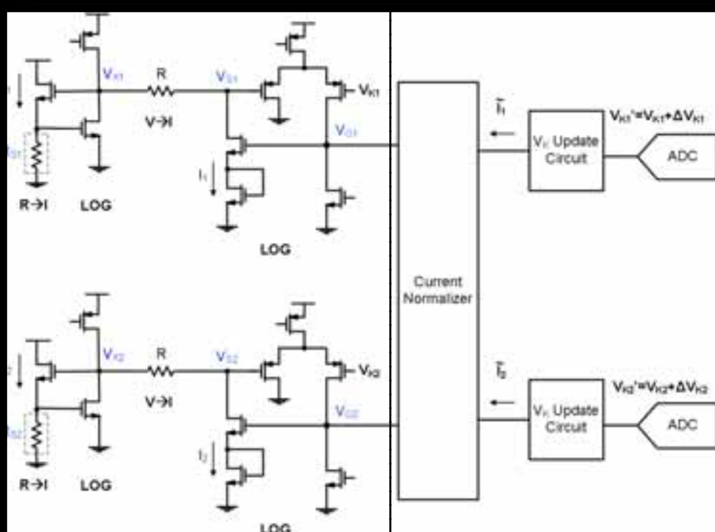
Example: Concentration-Invariant Encoding



Pulse pattern independent of concentration

Analyte information represented in time





Redundant Arrays of low-precision analog processing units

50 nW /channel at 0.5 V

Training and adaptation essential

[Courtesy: PC Huang, UCB]

Example: Analyzing Sensor Signals with Reduced Precision

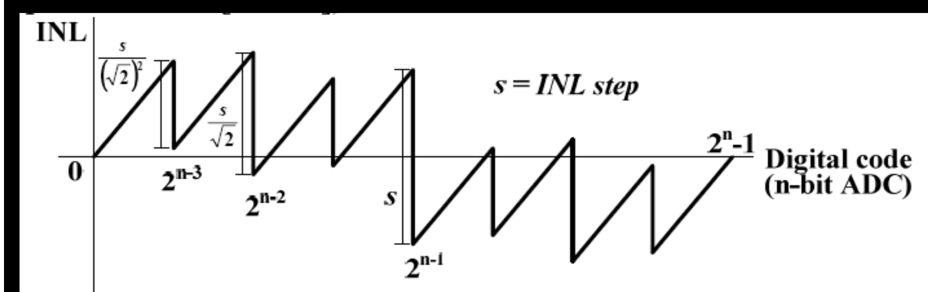
Classification System:

Sensor data → Feature Extraction → Data-driven Classification → Classifier output

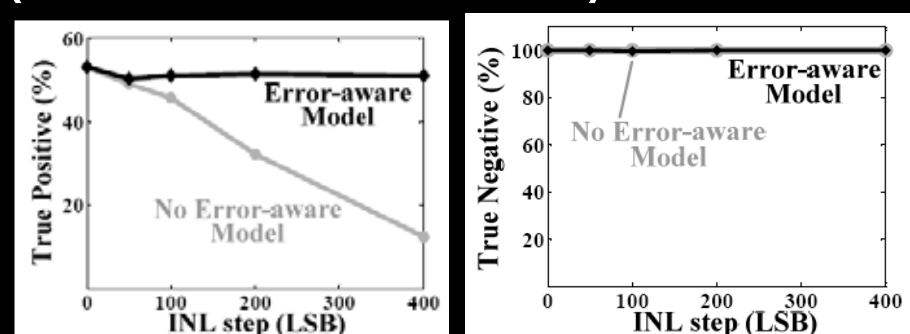
Permit imprecise signal-acquisition, conversion, feature-extraction (low-energy analog)…

Learn feature statistics due to imprecisions via data-driven classifier training

E.g. ADC Integral Non-Linearity (INL)



E.g. SVM training to INL: error-aware model
(EEG-based seizure detector)



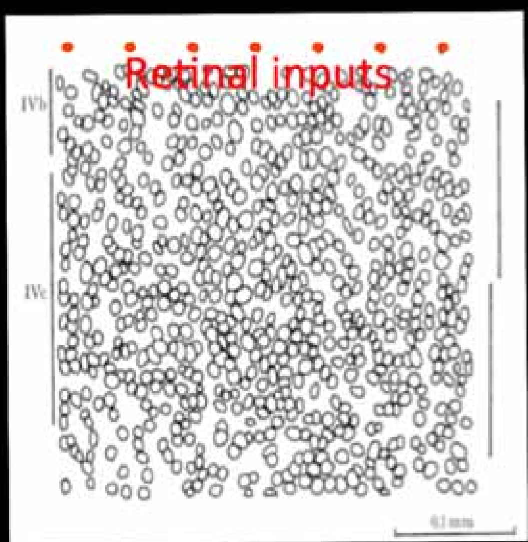
[Courtesy: N. Verma, Princeton]

Hyper-dimensional Representations

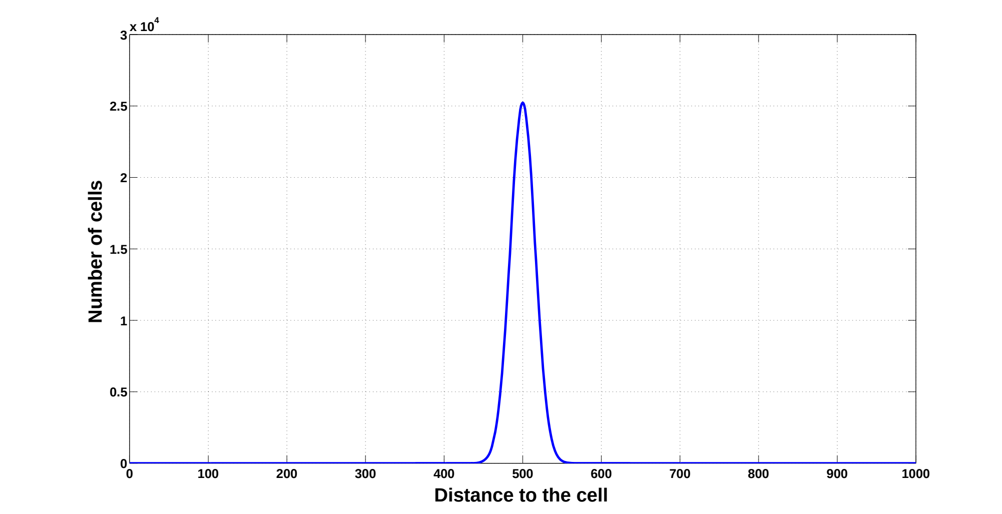
Representation is *hyper-dimensional* when number of dimensions is "much" (> 1000?) larger than needed to cover space.

- Extremely robust against most failure mechanisms and noise
- Purely statistical, thrives on randomness
- Supports full algebra

V1 (Layer 1 of visual cortex) is HIGHLY overcomplete

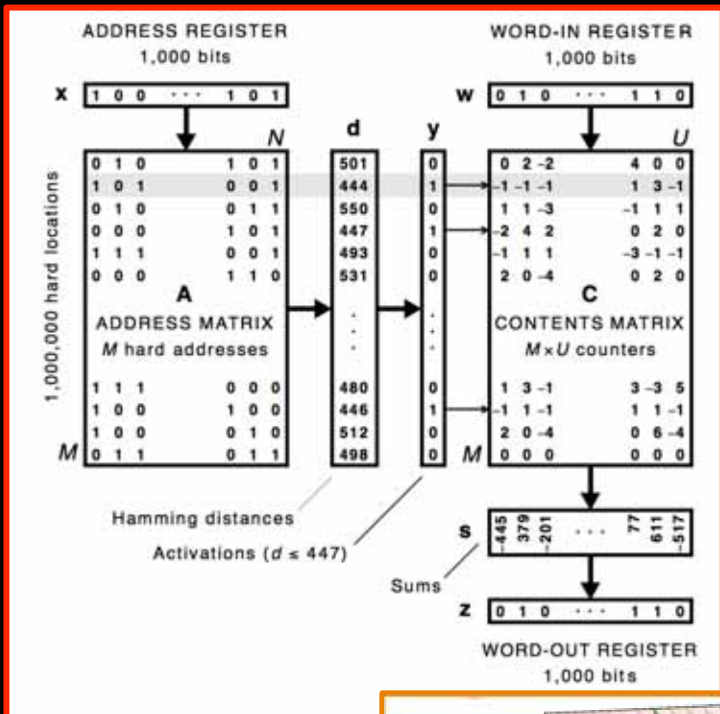


[Barlow 1981]

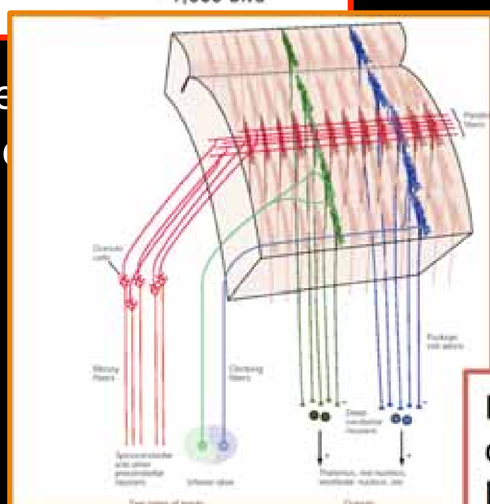


Distance histogram for 1 million points in N-dimensional space (N=1000)

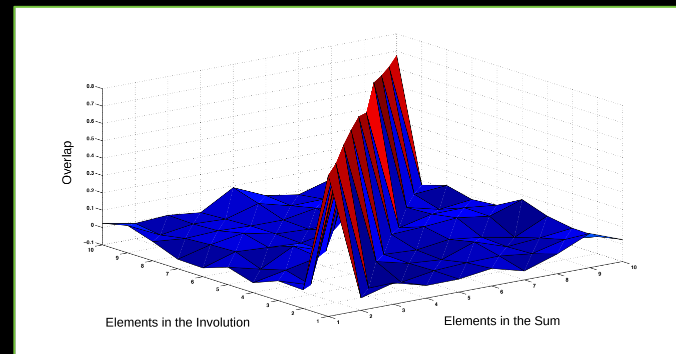
HD Classifier: Sparse Distributed Memory*

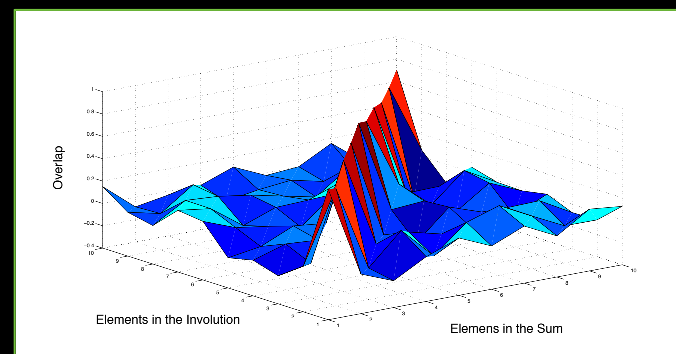


Imprecise set of fe
yields closest mat
library (w)



Basic cerebellum circuit (half of brain neurons)



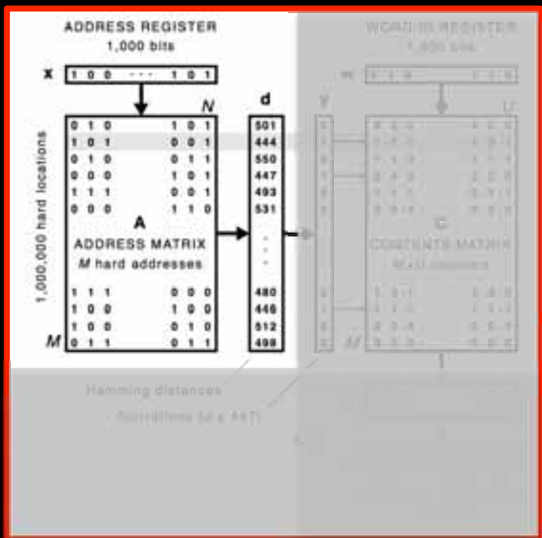


Dimension

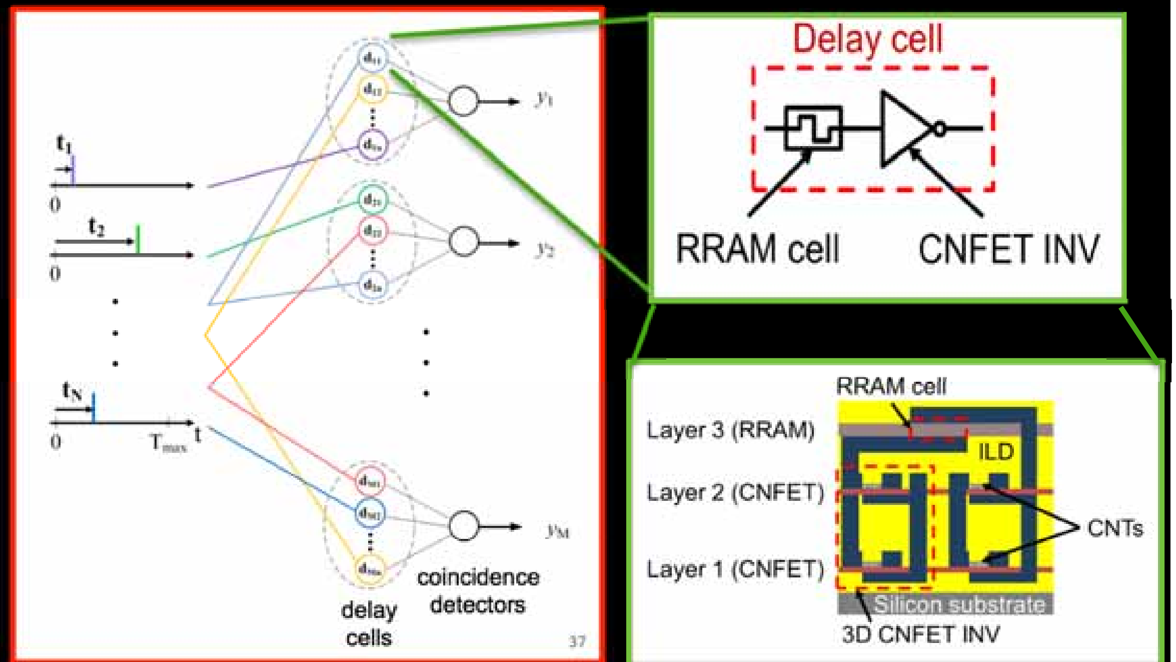
Overlap of stored elements (diagonal) vs. random vectors (non-diagonal)

* A class of associative memory

What is Cool about This?



Random indexing:
Orthogonal transformation of data into hyper-dimensional space

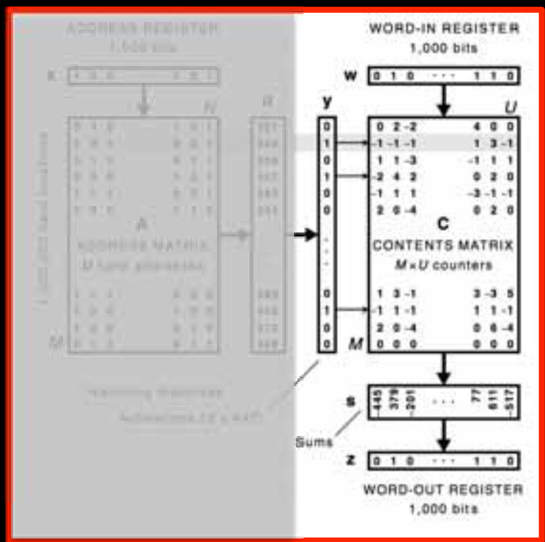


CNT-RRAM combination spreads distributions
3D integration enables scalability
Extremely low energy operation

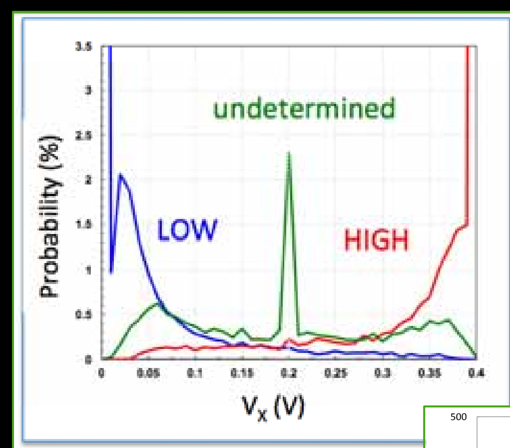
Die	CNT density (CNT/µm)	Delay (µs)		Standard deviation (µs)	Std/(Mean-Min)
		Mean	Min		
1	1	0.73	0.21	0.18	86%
2	0.33	2.23	1.36	0.94	69%
3	0.11	6.79	4.82	2.41	50%

[In collaboration with P. Wong and S. Mitra, Stanford]

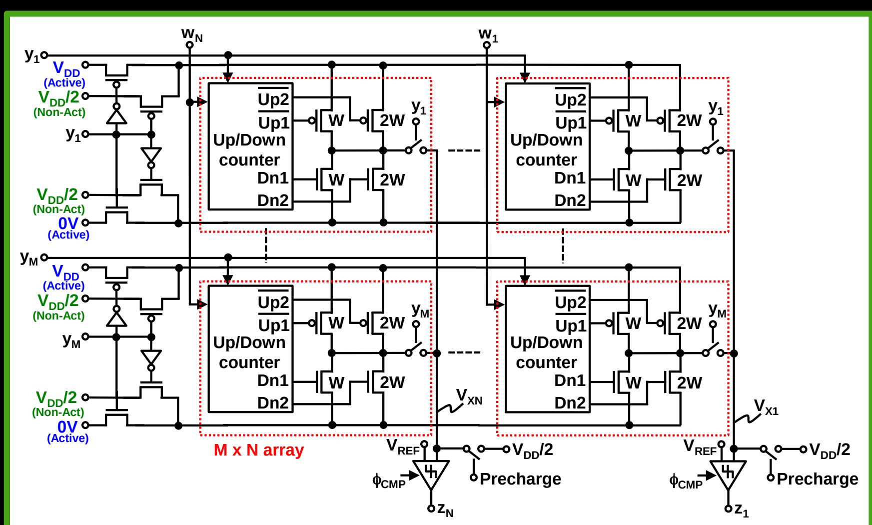
What is Cool about This?

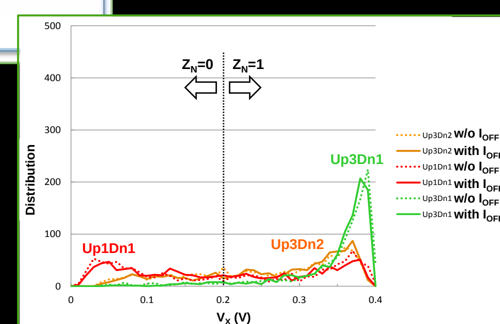


- Close Memory/Compute integration
    - Does NOT scale with current MOS memory technologies
    - Good match to RRAM/STTRAM
    - Only writes during training
- Low resolution distributed analog processing
    - Dimension versus variability and leakage



Probability of wrong decision = 2.3%!
(VDD = 0.4V, 30 active rows of 1000)





Impact of leakage ignorable!
(VDD = 0.4V, 30 active rows of 1000)

[Courtesy: M. Takamiya]

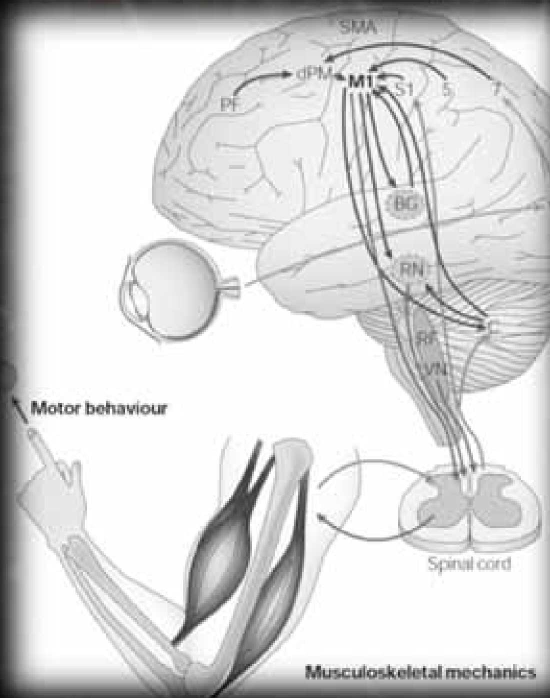
An exciting time …

"Brains work with patterns of neural activity that are not readily associated with numbers. The brain's reliance on high-dimensional distributed representations invites us to study high-dimensional computing, all the more so now that nanotechnology is poised to give us circuits that can scale up to brain-size. To benefit from the technology, we need a theory of computing that matches the technology …"

P. Kanerva, Berkeley, May 2014.

Higher-Order Bits

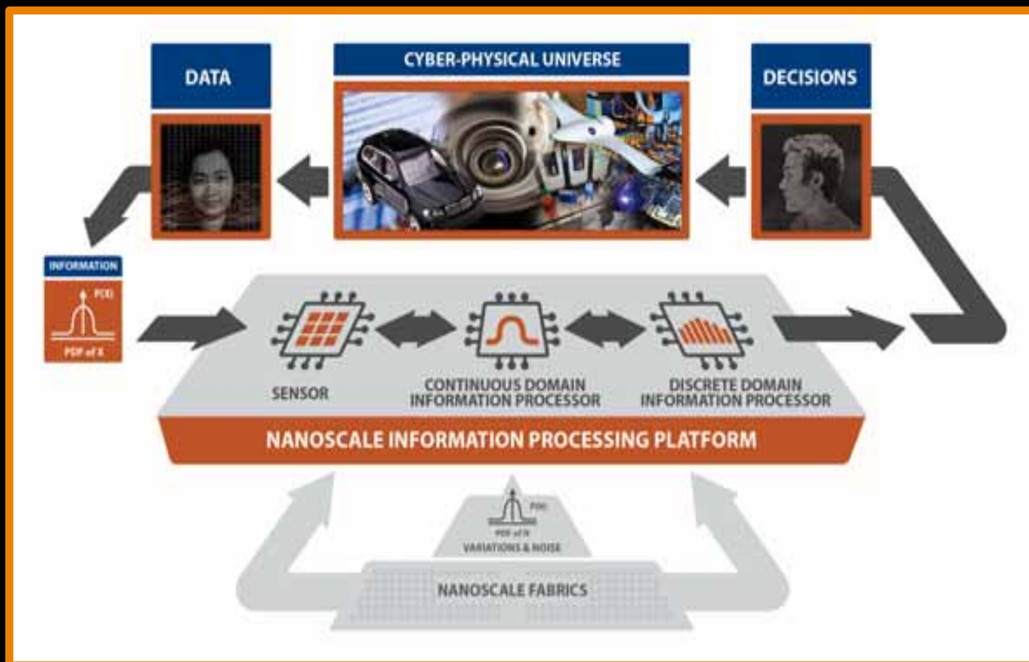
- Neuro-inspired and inference-based computational paradigms may be the perfect match to the next generation of nano devices
  - and as such the novel model of computation
- Prime Target: Addressing the data abundance in both the cloud and the swarm!
- The Search for Generalizable Solutions and Platforms is on
- Requires collaborations between neuroscientists and and architect, circuit and device engineers



Acknowledgements

The many contributions of, Bruno Olshausen, Pentti Kanerva, Ping-chen Huang, Ashkan Borna, Philip Wong, Subhasish Mitra, Jesse Engels, Naveen Verma, Naresh Shanbhag to this presentation are gratefully acknowledged.

The support of the FCRP GSRC and StarNet SONIC centers, as well as the member companies of BWRC is greatly appreciated.



Systems on Nanoscale Information Fabrics