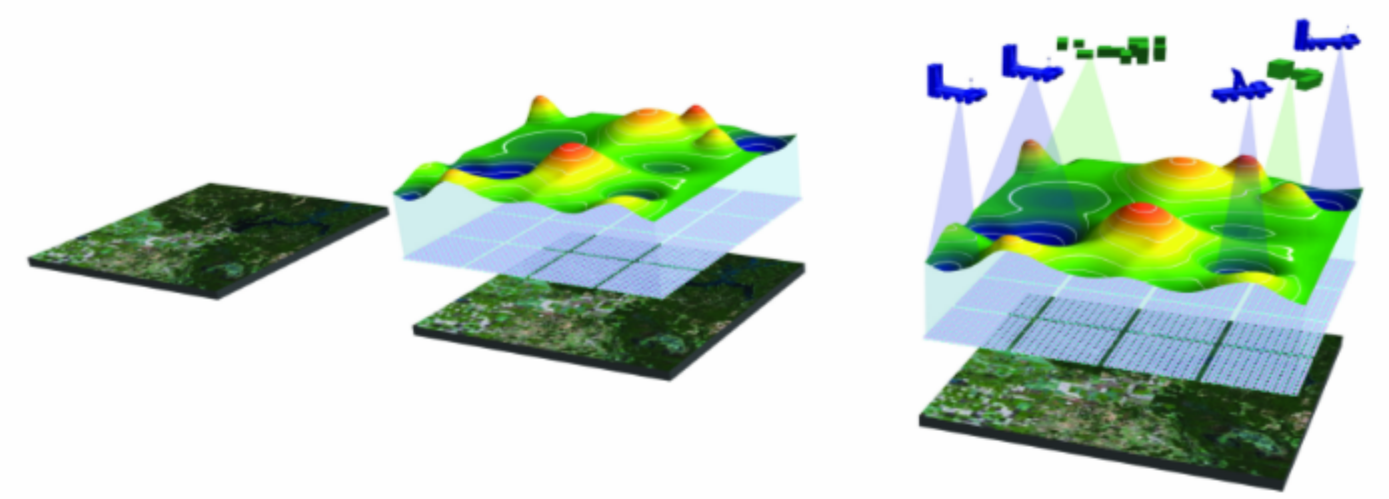
UPSIDE / Cortical Processor Study

Dr. Dan Hammerstrom

Program Manager / MTO

The Unconventional Processing of Signals for Intelligent Data Exploitation (UPSIDE) Program



Large coordinated multi-disciplinary teams with multiple subs

Two teams using neural-inspired solutions

BAE Systems
University of Massachusetts
Johns Hopkins
UCSB
Stony Brook University
SEMATECH

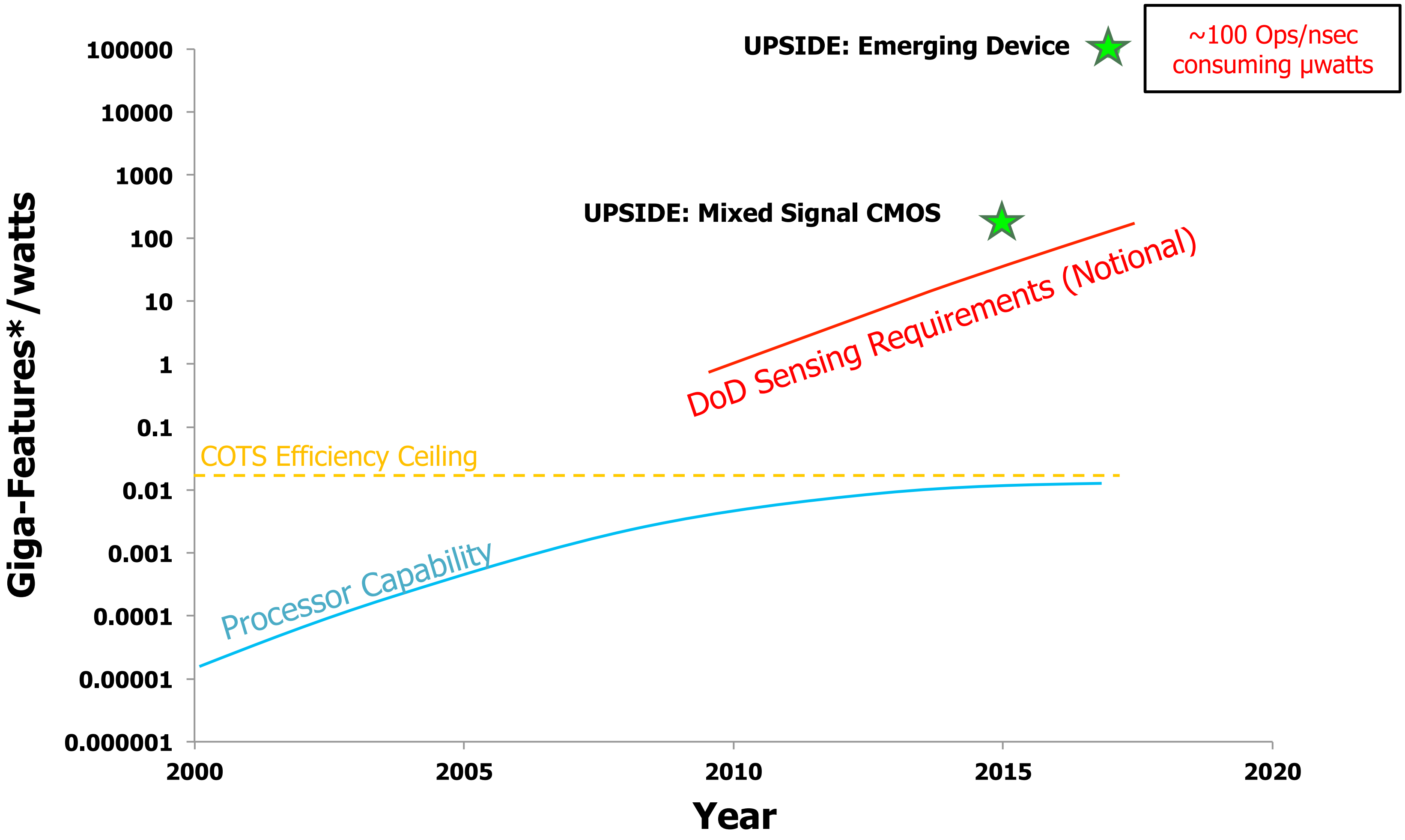
HRL Laboratories, LLC
Purdue University
University of Notre Dame
University of Pittsburgh
Intel Corp.
NIST

University of Michigan
Portland State University
New Mexico Consortium, Los Alamos National Lab

The University of Tennessee
Oak Ridge National Laboratory
Stanford University.

Distribution Statement A: Approved for Public Release, Distribution Unlimited

UPSIDE: Performance Goals



UPSIDE Goals: **3** orders of magnitude in throughput, **4** orders of magnitude in power efficiency, no loss in accuracy

UPSIDE Unconventional Processing of Signals for Data Exploitation

DARPA Insight #1: Exploit the physics of emerging devices and mixed signal CMOS to perform extremely fast, low power computation.

Front End Filtering (Edge Detection)

Approach is being implemented in MS CMOS for near term gains

Final Result: Filtered Image

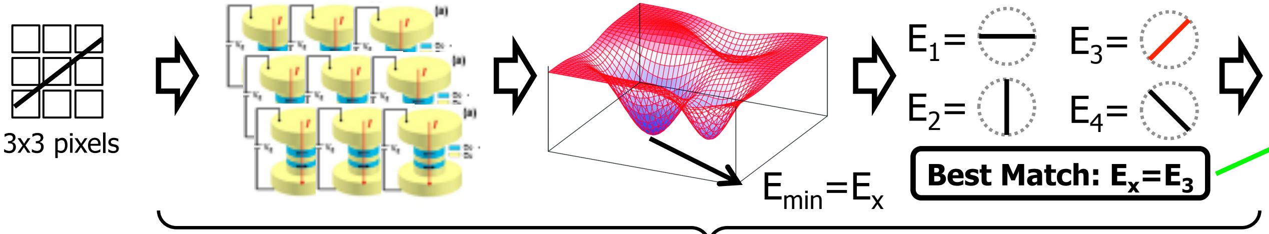
Image feature from CCD array

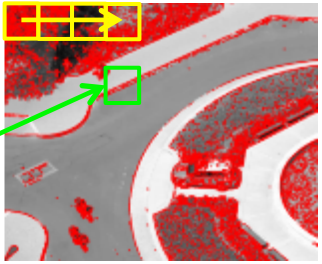
Pixels mapped into coupled oscillators

Oscillators relax to lowest energy state

Final energy compared against library of possible features

Step and repeat to Identify all Edges (in red)



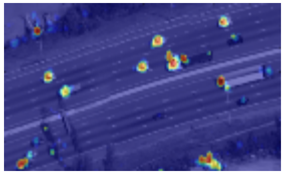


UPSIDE eliminates computationally intensive digital CMOS dot product multiplication

DARPA Neovision2 – Stanford Tower Video

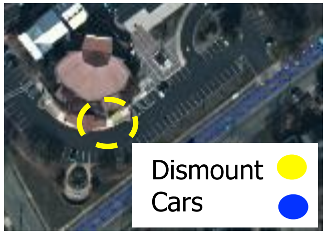
DARPA Insight #2: Computational method can be applied universally to almost every computing function in the front end of the Image Processing Pipeline

Object Detection



Object Saliency/Tracking

Object Classification



BAE Systems – ARGUS IS

Reduce ISR computational power budget from kW to W, while increasing speed >100x

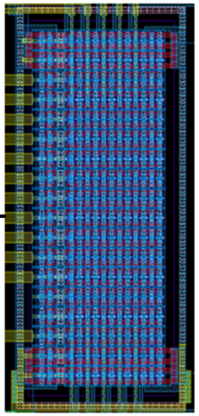
UPSIDE ARGUS-IS Image Processing Pipeline: **40GP/s, 5W**

1000x more power efficient
100x faster

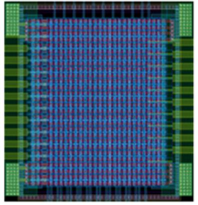
BAE SYSTEMS

NVM Tape-out #2:June 10th- successfully tested

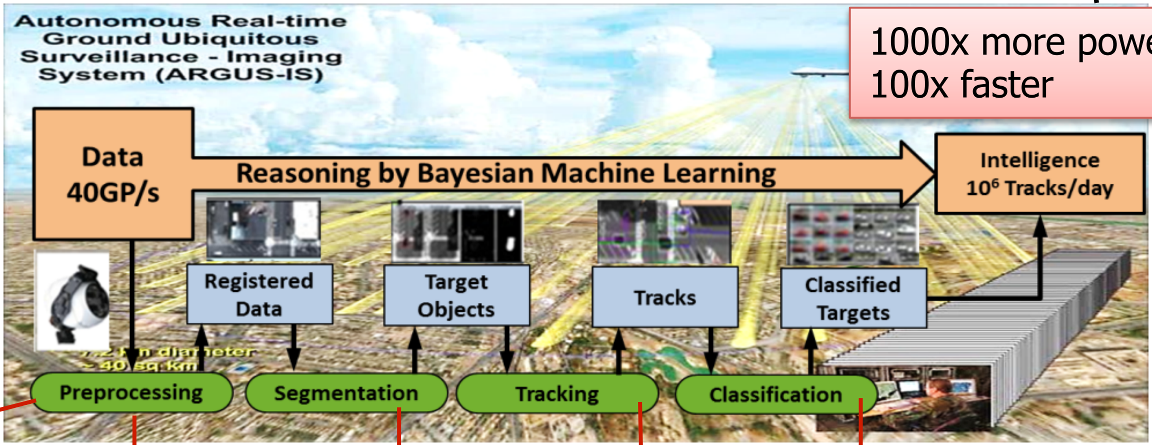
Source coupled VMM



Gate-coupled VMM



NUC/Debayer



Autonomous Real-time Ground Ubiquitous Surveillance - Imaging System (ARGUS-IS)

Data 40GP/s

Reasoning by Bayesian Machine Learning

Intelligence 10⁶ Tracks/day

Registered Data

Target Objects

Tracks

Classified Targets

Preprocessing

Segmentation

Tracking

Classification

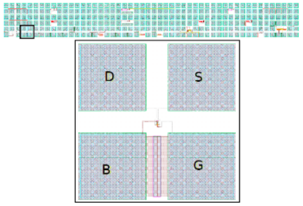


Figure 5. Test structure for jhu_nhmos4_4p5_1.

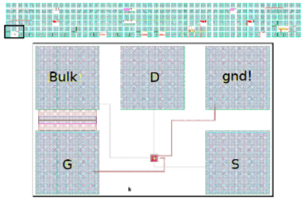
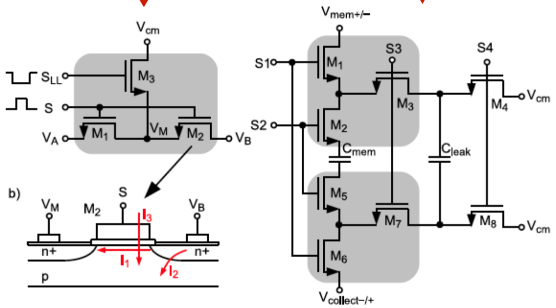
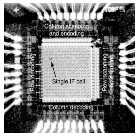


Figure 6. Test structure for jhu_phmos4_0p8_1.

Symmetric FGMOS device to be used in analog NVM circuits

Johns Hopkins Neuromorphic Circuits

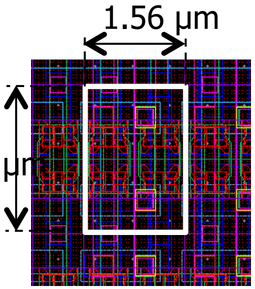




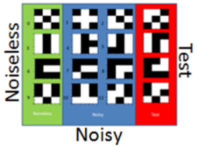
65 nm UCE includes computing with capacitor arrays

"Training and operation of an integrated neuromorphic network based on metal-oxide memristors", M. Prezioso et al., Nature Letter, 7 May 2015, Vol. 521

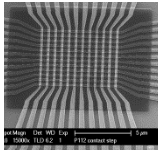
Modified commercial NVM memory technology: >**500x** density advantage over state-of-the-art



Fabricated and tested on May 14



- First demo of this kind
- Physical memristor x-bar implementation
- 3x3 binary input images
- 4 classes (X,I,C,V)

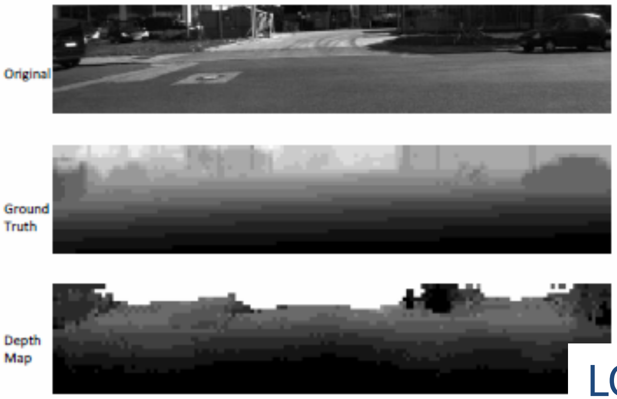


Micrograph of 12x12 Mermistor Array

Distribution A. Approved for public release: distribution unlimited.

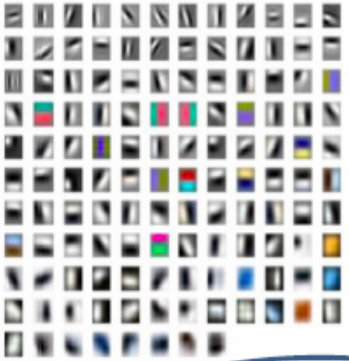
Sparse Adaptive Local Learning: A universal platform for efficient sensing and analytics (U Michigan)

Depth from Stereo



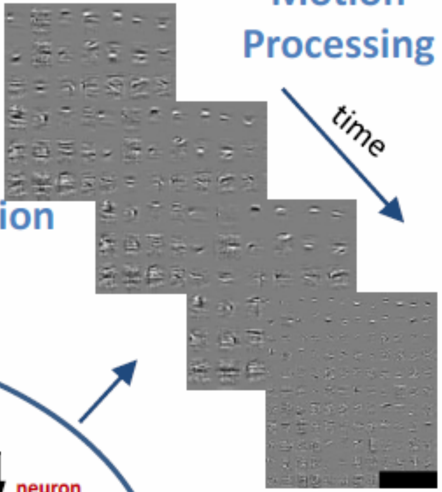
LCA

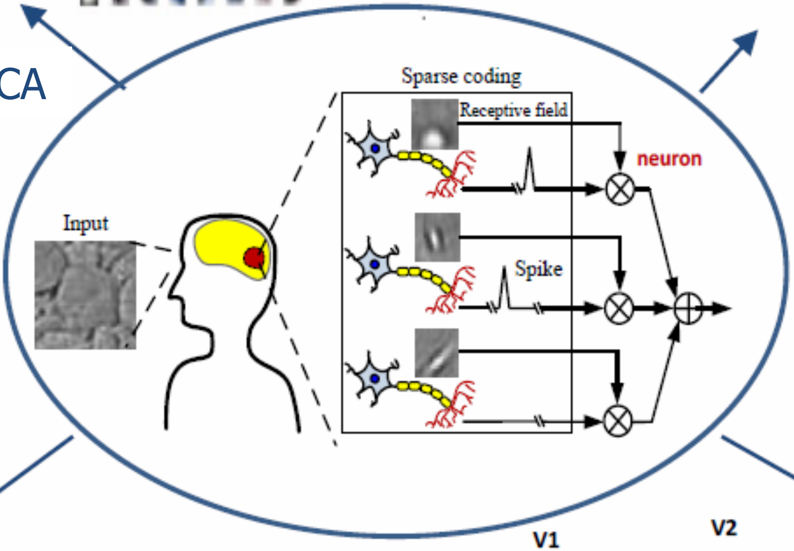
Analytics



Self Organization

Motion Processing



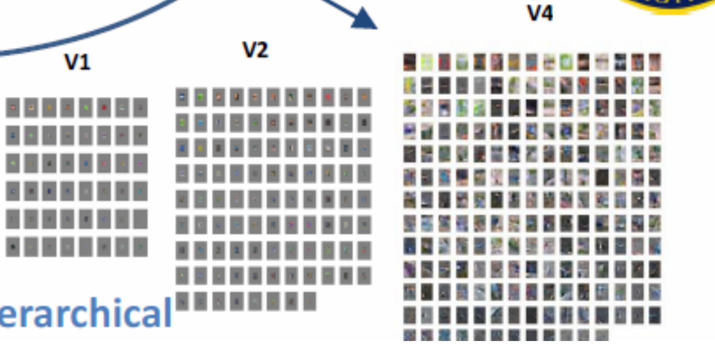


Sensing

Gar Kenyon

Object Detection

Neural Inference Module



Hierarchical

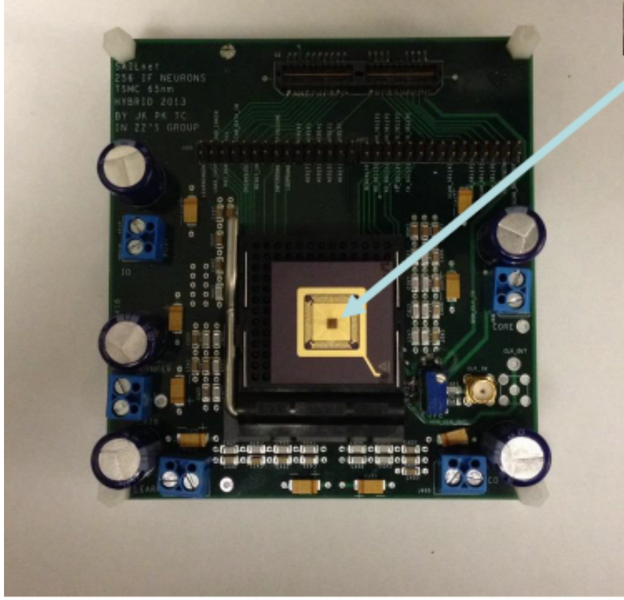
Work performed under DARPA Cooperative Agreement Award HR0011-13-2-0015

Mapping The Neural Inference Module Onto A Low-power Sparse Coding Processor

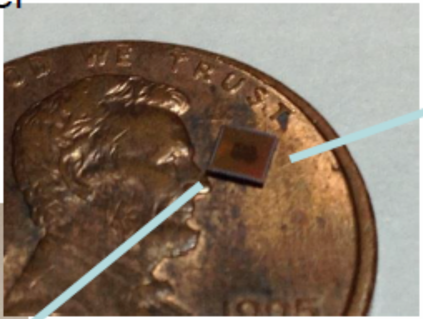
3.1 mm² sparse coding processor in 65nm CMOS
Inference throughput: 1.24 Gpixels/s
Learning throughput: 188 Mpixels/s
Energy: as low as 47.6 pJ/pixel

Zhengya Zhang
Michael Flynn

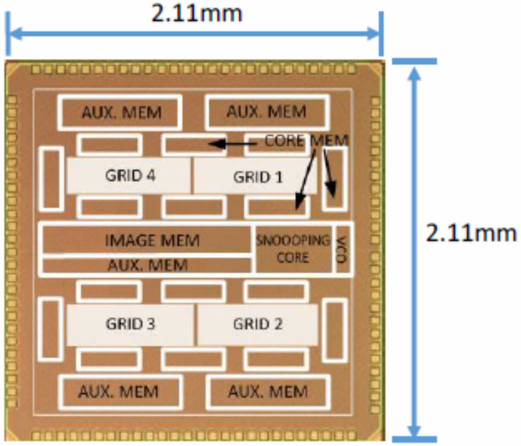
Evaluation board



Sparse coding chip



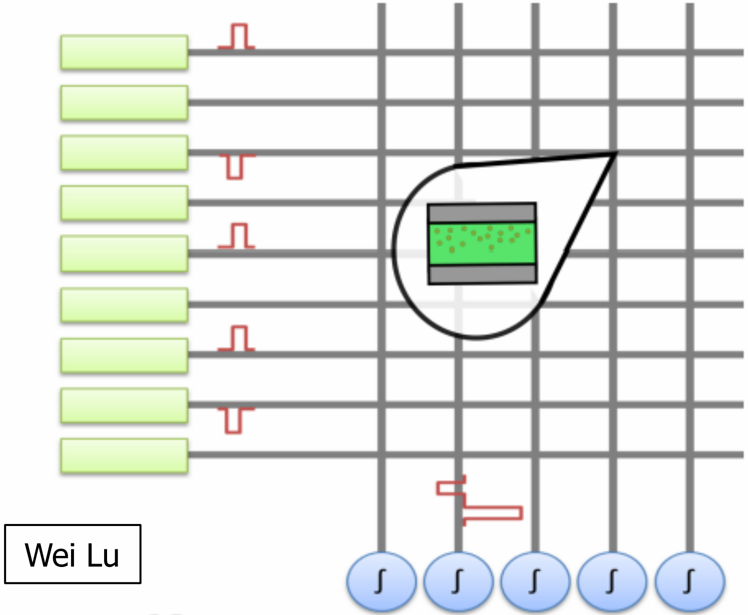
Microphotograph



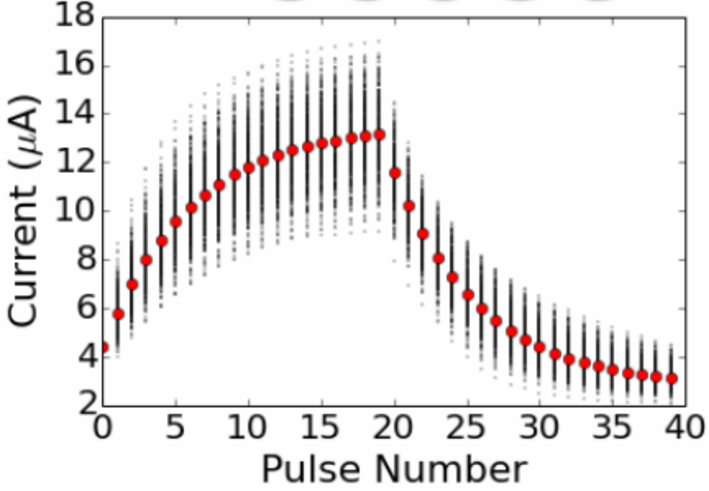
- Low spike rate translates to low power consumption
- Efficient sharing of neuron communication enables scalable architecture
- Excellent quantized performance for efficient memory usage
- Soft processing and error resilient for low power approximate computing

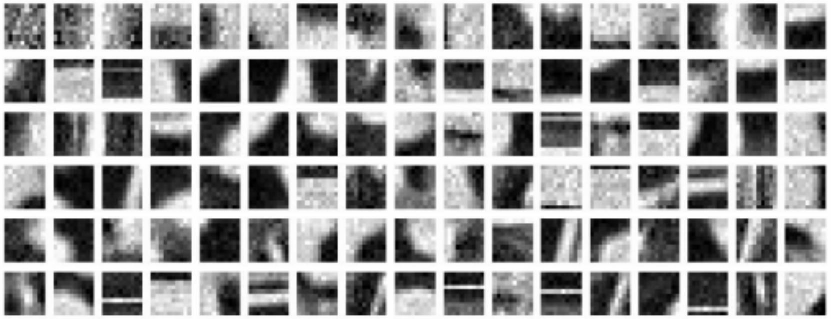
Work performed under DARPA Cooperative Agreement Award HR0011-13-2-0015

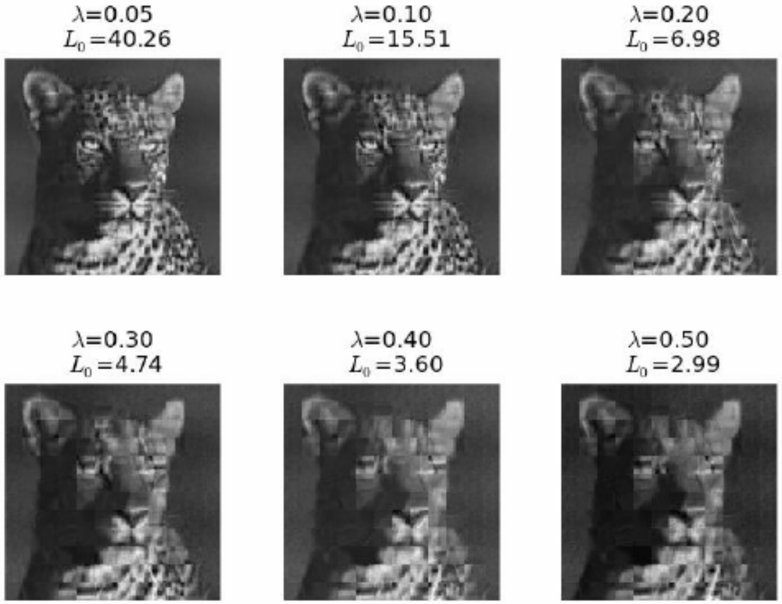
Mapping IM onto Memristor-Based Hardware



Wei Lu



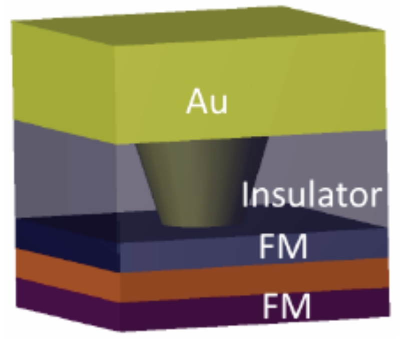




Results: Emerging Device Characteristics

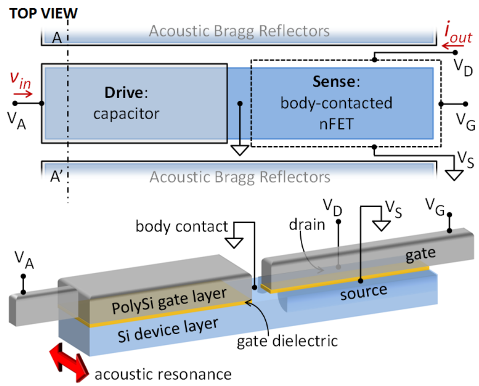
Emerging device specifications indicate that both power and frequency data are near the projected values for UPSIDE program goal.

Spin-Torque Oscillator (STO)



Parameter	Simulated	Measured	Projected
Power consumption per STO	45 μW	100 μW (10mv, 10 ma)	10 μW (10mV, 1mA)
Power consumption, 16 STOs			360 μW
Nanocontact size		100 nm	40 nm
Time to phaselock, 2 STOs	3-5 ns	3-5 ns	2.5 ns
Frequency		1-40 GHz	10-40GHz (CMOS dependent)
Footprint , unit cell 2 STOs + resistive coupling			1 x 5 μm^2
Footprint, 16 RBO cluster			2 x 20 μm^2

Resonant Body Oscillator (RBO)



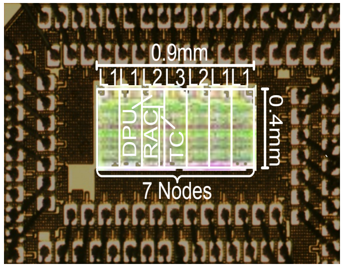
Parameter	Simulated	Measured	Projected
Power consumption per RBO	22.5 μW		20 μW
Power consumption, 16 RBOs			360 μW
Energy dissipated in coupling resistance, 2 RBOs, anti-phase to in-phase transient	99 aJ		
Energy dissipated in coupling resistances, 16 RBOs, anti-phase to in-phase transient			1.6fJ
Time to phaselock, 2 RBOs anti-phase to in-phase transient	33 ns		3.3ns
Frequency	1-10 GHz	11.1GHz	10GHz
Footprint , unit cell 2 RBOs + resistive coupling	40 x 40 μm^2		
Footprint, 16 RBO cluster			80x160 μm^2

Distribution Statement A: Approved for Public Release, Distribution Unlimited

Deep Learning Analog Chip

Deep Learning Chip Architecture Implemented with Custom Analog Elements

- Floating-gate analog memory for non-Boolean, probabilistic pattern matching performing on-chip, real-time training
- *Approach enables highly efficient computation for object recognition, classification and tracking*



University Tennessee press release and news articles about chip and DARPA UPSIDE program

Tennessee Today
NEWS AND EVENTS FOR THE UT COMMUNITY

Browse Story Topics

Faculty & Staff Headlines From [illegible]

Top Headlines

[illegible]

UT Engineers Team Up to Move Analog Brain Closer to Reality

109

Machine learning, the science that makes it possible for devices to "think" on their own, is taking the next step forward thanks to UT College of Engineering [illegible] and Computer Science.

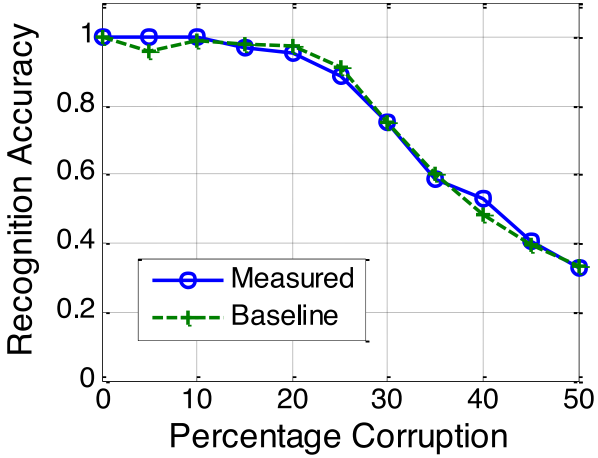
[illegible]

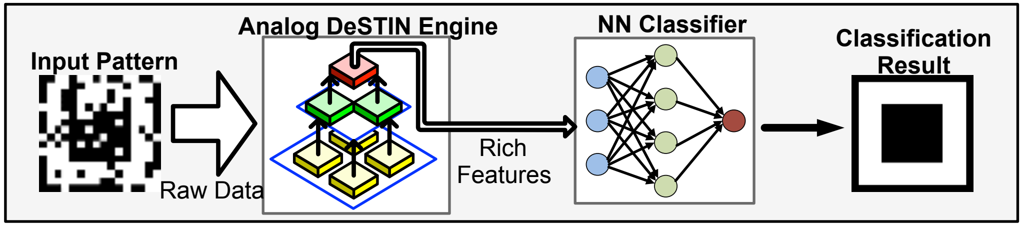
Performance & Efficiency

Accuracy comparable to s/w, with **282x lower training energy** than synthesized custom digital equivalent.

UPSIDE Chip Performance	
Training Efficiency	
Digital Design	**UPSIDE Chip**
1.7 GOPS/W	**480 GOPS/W**
282x Improvement	

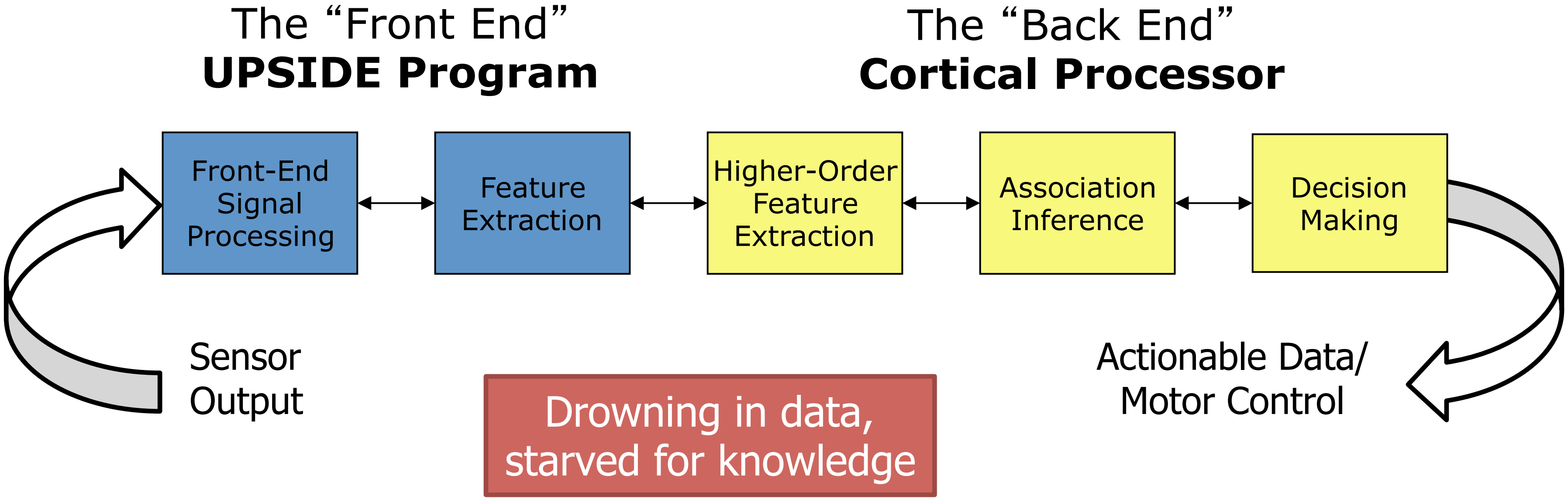
Recognition Accuracy





J. Lu, S. Young, I. Arel, J. Holleman, "A 1TOPS/W Analog Deep Machine-Learning Engine with Floating-Gate Storage in 0.13um CMOS," IEEE Journal of Solid-State Circuits, Vol. 50, Issue 1, pp. 270-281, Jan. 2015.

Sensor Processing Pipeline: A Data Analysis Crisis



Drowning in data, starved for knowledge

- Sensor data bandwidth exceeding processing capabilities, particularly for embedded systems
- Data become more knowledge / context intensive, containing both spatial and temporal information, as they move through the pipeline
- Current computational approaches do not adequately represent complex spatial and temporal data, limiting the ability to effectively perform complex recognition for important DoD tasks like anomaly detection and scenario prediction

Explosion of Data and Software Complexity

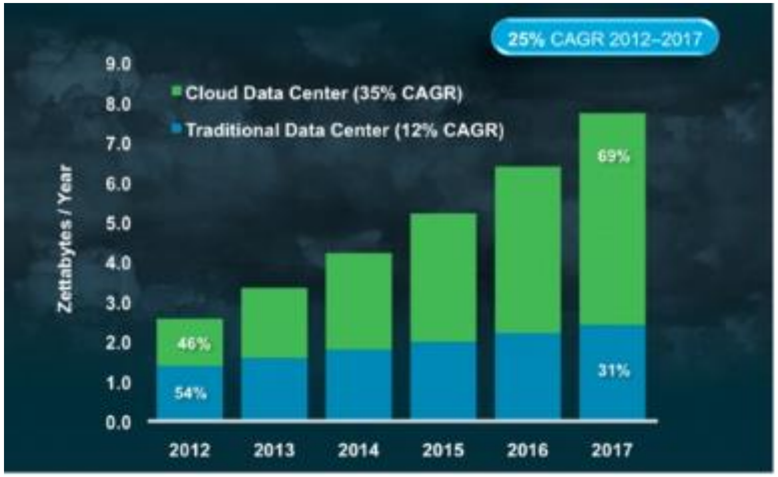
Can learning be leveraged as an efficient system construction alternative, for system control as well as data processing?

- Expose the agent to reality rather than trying to approximate it through programmed equations
- Learn complex and subtle relationships in the data and perform inference over those structures Rich models allow more robust anomaly detection
- Continue learning and adaptation *in situ*

Global Hawk

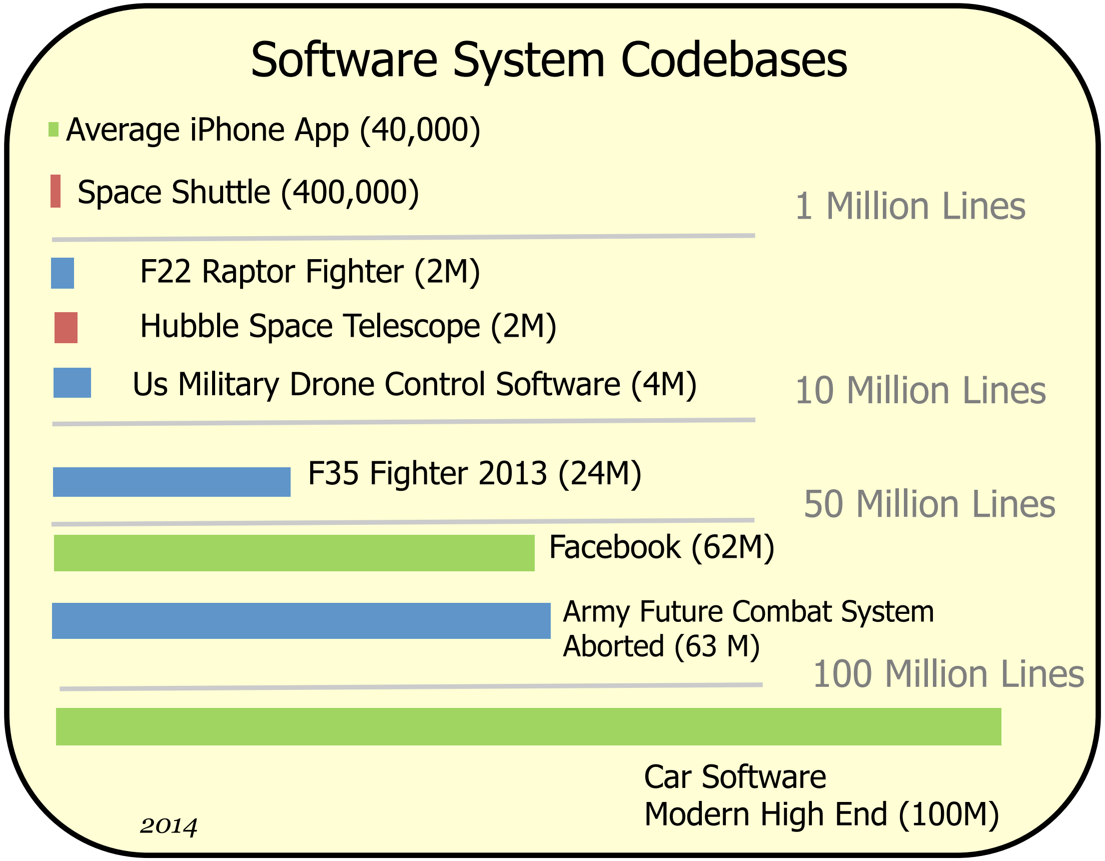
DoD Sensing: A single Global Hawk requires 500 Mbps → 5x the total SatCom bandwidth that the entire U.S. military used during the Gulf War

Big Data: Global Data Center Traffic Projection



2012-2017 © Cisco Global Cloud Index

Software System Codebases



Cortical Processor Study

- Study consists of 12 performers and runs from Q2 2015 to Q2 2016

- MTO Cortical Processor Study investigates systems that:
  - Eliminate the need for large training sets as a prerequisite to training
  - Train in real time in an unsupervised or weakly supervised environment
  - Recognize temporal as well as spatial patterns for recognition of action and anomalies
  - Learn and perform inference over complex structure in data, scenarios

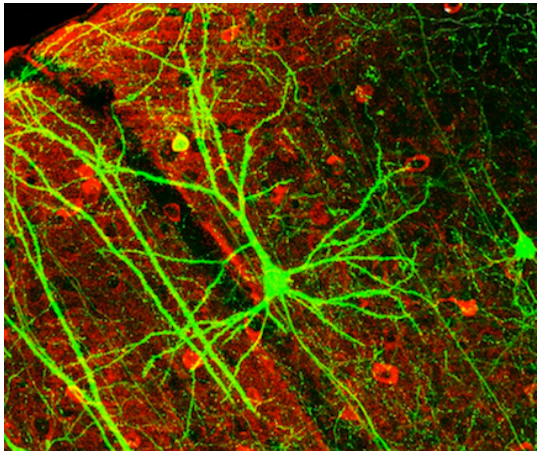
- How: Leverage elements of computational neuroscience
  - Spatial/temporal pattern recognition
  - One shot learning – network re-use
  - Efficient performance – sparsity and lower precision reduces HW requirements

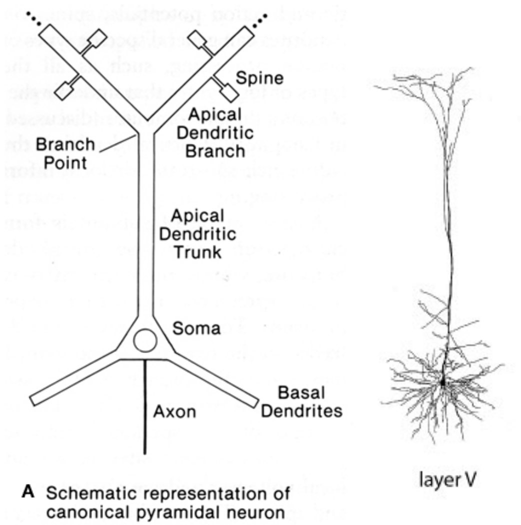
- What the program will do:
  - Take image processing to the next level - systems that learn objects and actions from processing video streams, with minimal labeled training data
  - Model free adaptive control
  - Performance = real-time learning
  - Power and size constraints driving efficient use of hardware, specialized and/or custom

Computational Neuroscience 101

- Lateral inhibition leads to sparse activation and connectivity – creating Sparse Distributed Representations (SDR)
  - Results in a limited distribution sparse activation which, in hardware, can be leveraged for significant efficiency
  - Combinatorics in our favor, e.g. 1000 neurons, 10 active at a time: 2.6x10^{23} possible representations
  - Only a small number of cells are required to recognize a pattern
- Rapid learning – typically one shot - imprint sub-vector on patch of dendritic tree
  - Hebb rule: neurons that fire together, wire together
  - One variation is called One and a half shot learning, where there is some adjustment of imprinted weights
  - Synapses are only possible where axons and dendrites have some physical proximity, providing a wide range of random segments – again combinatorics works in our favor
- Learning is fundamentally unsupervised

  Supervised, weakly supervised, and reinforcement learning also possible
- Weights and activations are typically low precision
  - The expense is in representing and emulating connectivity, not in the arithmetic
- Temporal information is fundamental to neuron construction – delays are ubiquitous in dendritic trees
  - Dendritic trees are active, pulse signals are amplified as they proceed to the soma
- Sequence memory (predicting forward in time) is ubiquitous
  - HTM/CLA Numenta (Hawkins & Ahmad)

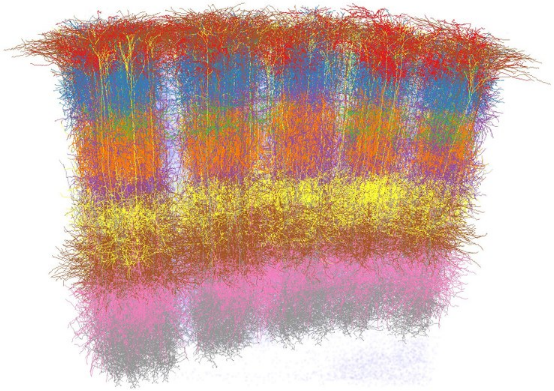




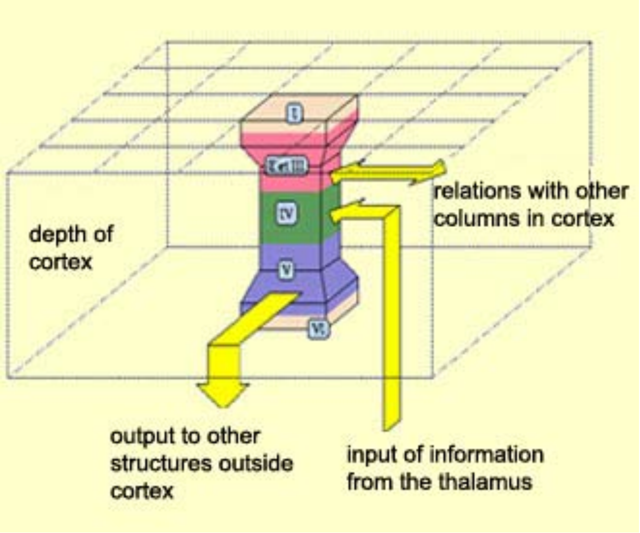
"Pyramidal neurons: dendritic structure and synaptic integration", Nelson Spruston, Nature Reviews Neuroscience 9, 206-221 (March 2008)

Computational Neuroscience 102

- Many models are spiking – which is very favorable for hardware implementations (IBM TrueNorth)
- Feedback as well as feed-forward pathways
  - Hypothesis reinforcement
  - Saliency (directing attention)
  - Spatial and temporal dilation ascending the hierarchy
  - Hierarchical SDRs may allow the efficient capture of and inference over sparse graphs – the ability to capture complex, high level structure
  - IBM's Hierarchical Context Networks (Wilcke)
- Close approximation to Bayesian inference
- Cortical columns: tight intra and local inter column connectivity, sparse longer range connectivity, creates a natural modular structure with more efficient connectivity utilization
- Systems built from more specialized cortical areas are now starting to appear (Eliasmith) – Spaun
  - http://www.extremetech.com/extreme/141926-spaun-the-most-realistic-artificial-human-brain-yet
- Homeostasis
  - Goal is average activity; inactive neurons and synapses continuously reduce threshold to insure uniform activity
  - Keeps all neurons and synapses in the game and actively learning



Cell-type-specific 3D reconstruction of five neighboring barrel columns in rat vibrissal cortex (credit: Marcel Oberlaender et al., Cerebral Cortex October 2012;22:2375±2391)


I
II et III
IV
V
VI
depth of cortex
relations with other columns in cortex
output to other structures outside cortex
input of information from the thalamus

The Cortical Column: http://www.metz.supelec.fr/metz/recherche/ersidp/Projects/Cortical/Root.html

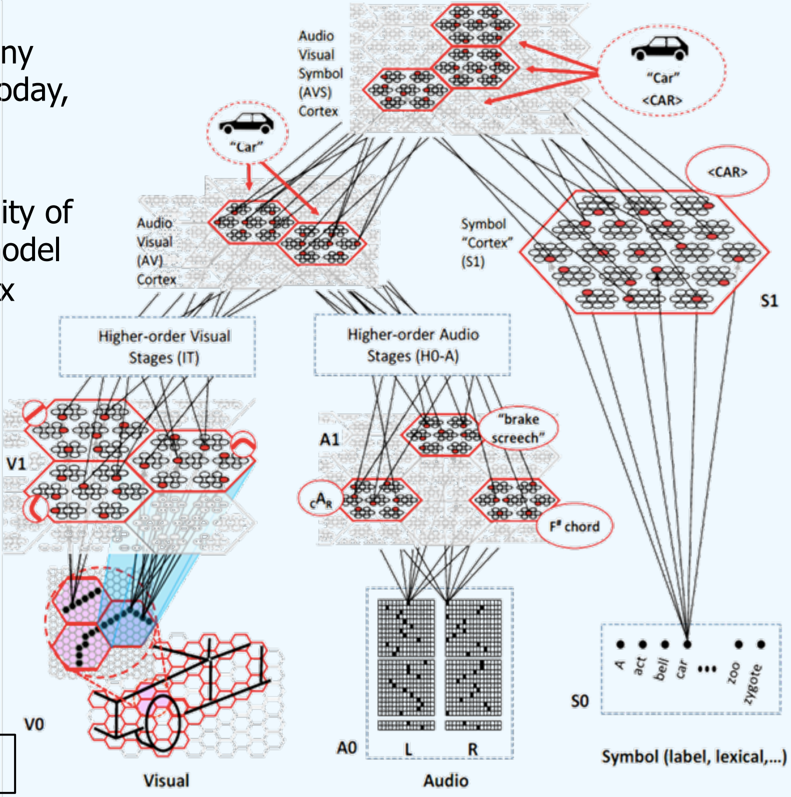
Distribution A. Approved for public release: distribution unlimited.

Sensor Fusion – Leverage Structure in Data

Cortical-like algorithms have the potential to solve the most challenging DoD sensor problems (*Sensor streams can be gracefully added dynamically in the field)*

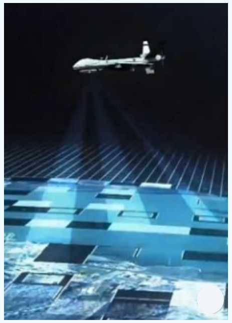
Sensor Data Fusion

- Not possible with any neural algorithms today, nor with traditional techniques
- Creates the capability of using learning to model and control complex systems
- Helps manage signal and system complexity by automating higher order relationships

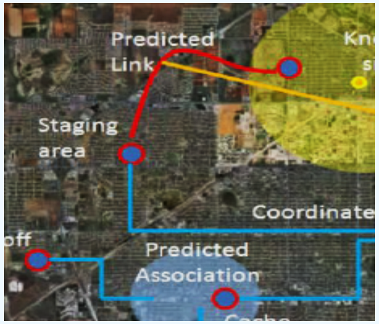


Rinkus - Neurithmic

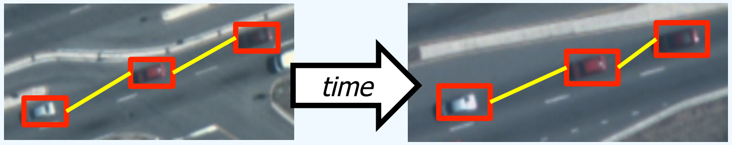
Sensor Applications



Surveillance imaging



Scenario awareness
Complex structure



Tracking convoy of vehicles

Structure in Data

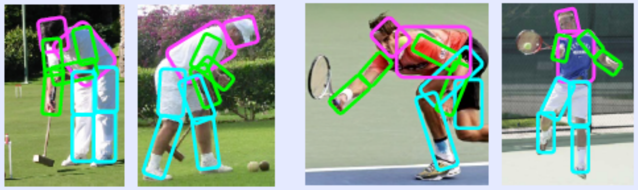
A:

Tennis forehand · Croquet shot · Volleyball smash

O:

Tennis racket · Croquet mallet · Volleyball

H:

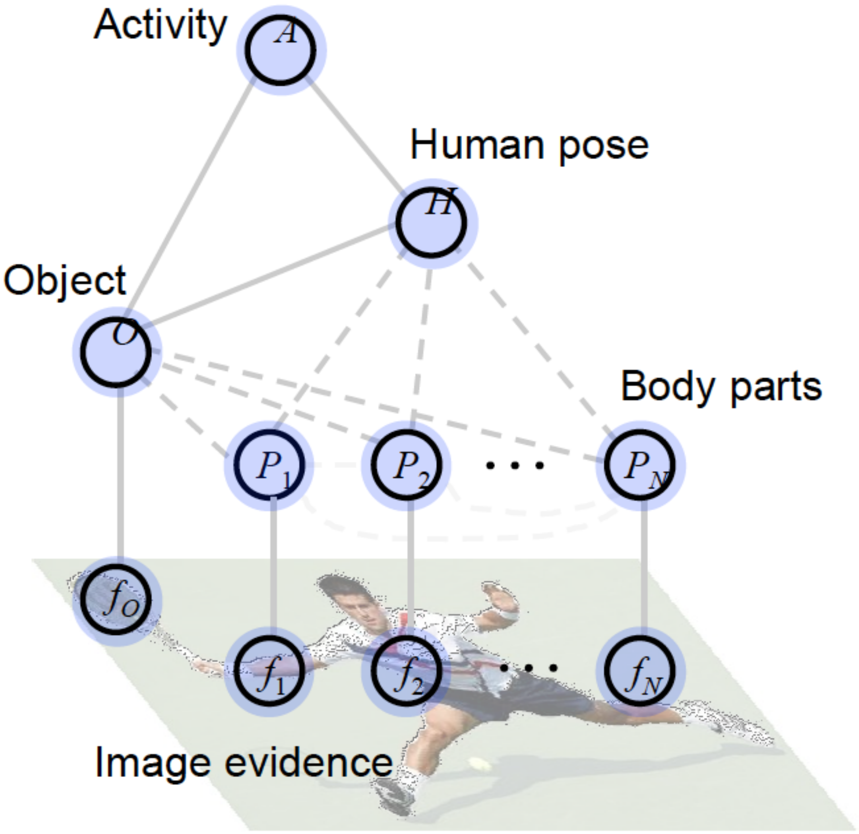


Intra-class variations

- More than one H for each A;
- **Unobserved** during training.

P: l_P: location; θ_P: orientation; s_P: scale.

f: Shape context. [Belongie et al, 2002]



Yao and Fei-Fei, CVPR 2010

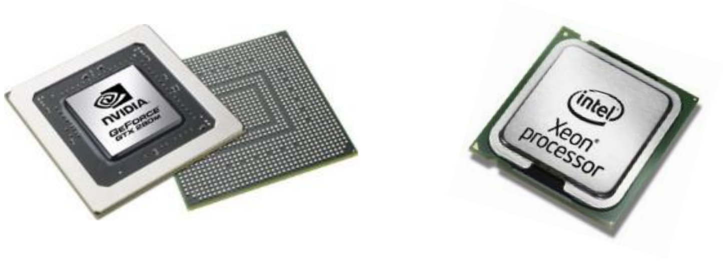
Mapping Bio-Inspired Algorithms to Hardware

Bio-inspired machine learning algorithms require matched hardware

1. High connectivity
2. Local memory and parameter storage
3. Simple, low-precision computation
4. Configurable / Adaptable
5. Sparse activity

Conventional processors are a poor match to cortical algorithms:

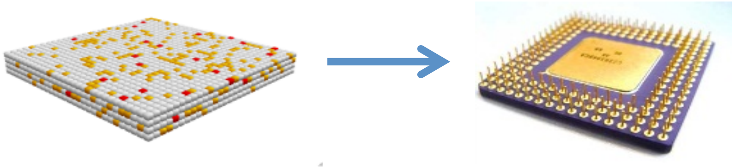
- Constrained: processor/memory partition, limited parallelism
- Excessive: high precision, tiered caches, complex instruction sets, pipelines, etc.



Conventional Solutions

Custom architectures can leverage bio-inspired approach:

- High-risk exotic devices unnecessary
- Utilize conventional CMOS fabrication optimized for neuro architecture/computational model
- Can benefit from latest advances in CMOS



Bio-inspired Algorithms

Specialized cortical processor