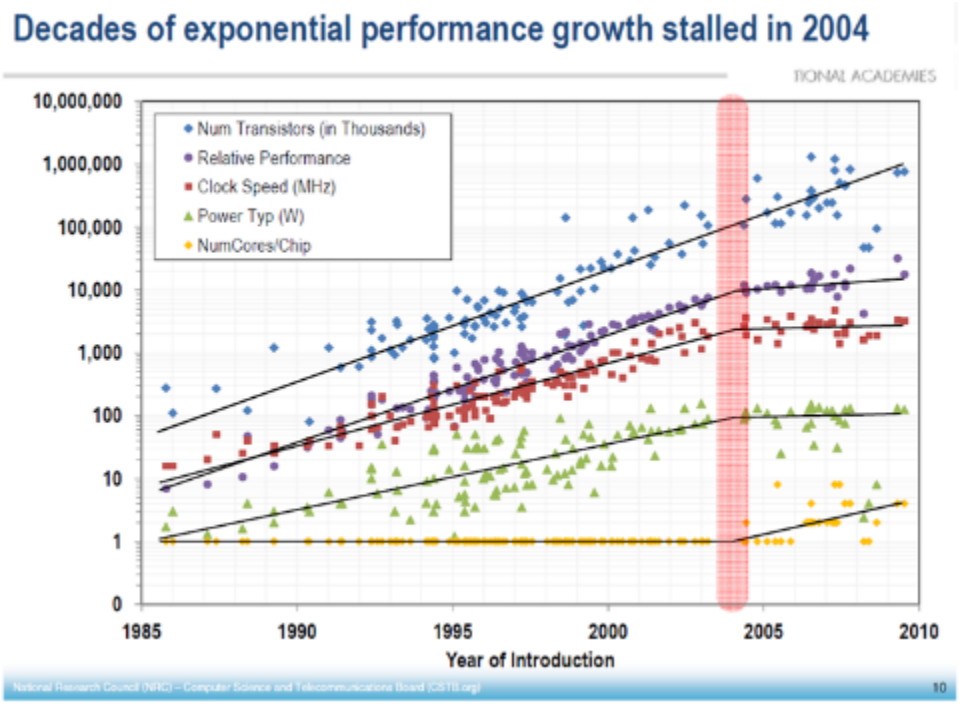
UPSIDE / Cortical Processor Study

Dr. Dan Hammerstrom
Program Manager / MTO

Semiconductors: Where We Are

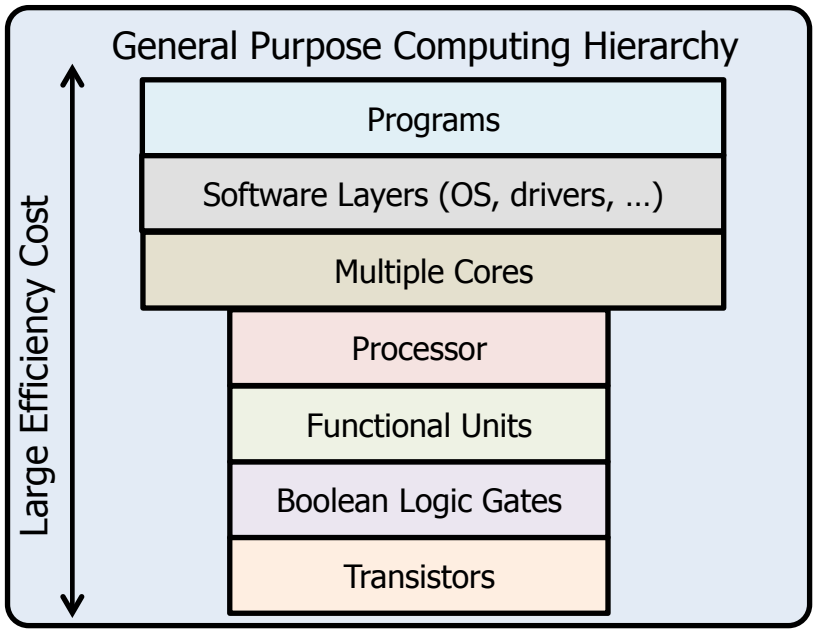


Source: NRC, The Future of Computing Performance, Game Over or Next Level?

- Moore's law continues – we're getting more transistors with each geometry shrink.
- Dennard scaling has stopped – voltage decreases have stalled even as feature sizes shrink. Clock rates would have to decrease in order to hold power constant.
- Hardware offers lots more concurrency. Software in general can't use it all.

For power or energy constrained DoD-embedded systems, greater power efficiency is the only path forward

Solution: Non-Digital, Probabilistic Computing Reduces Hierarchy

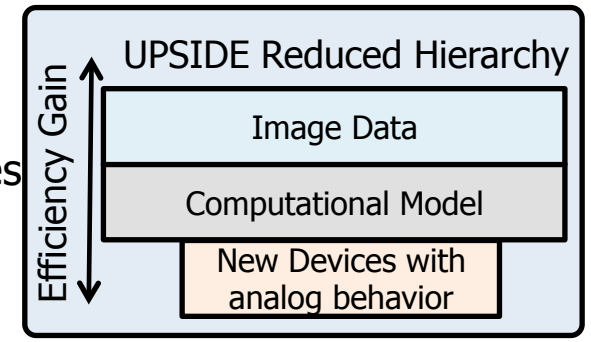


Digital architectures are not well matched to feature extraction from sensor images

- Images are inherently analog
- Digital algorithms are created to search for image structures based on existing digital number crunching architectures
- Digital abstractions limit data analysis
  - 100s to 1,000s of digital operations per image activity
  - Wasted energy in excess operations, data movement and precision

Need new computing approaches matched to image processing

- Use the physics of new emerging devices to extract features
- Data naturally represented in sparse form more suitable to devices and efficient for data transfer



Computing directly with devices eliminates multiple layers of hierarchy/inefficiency

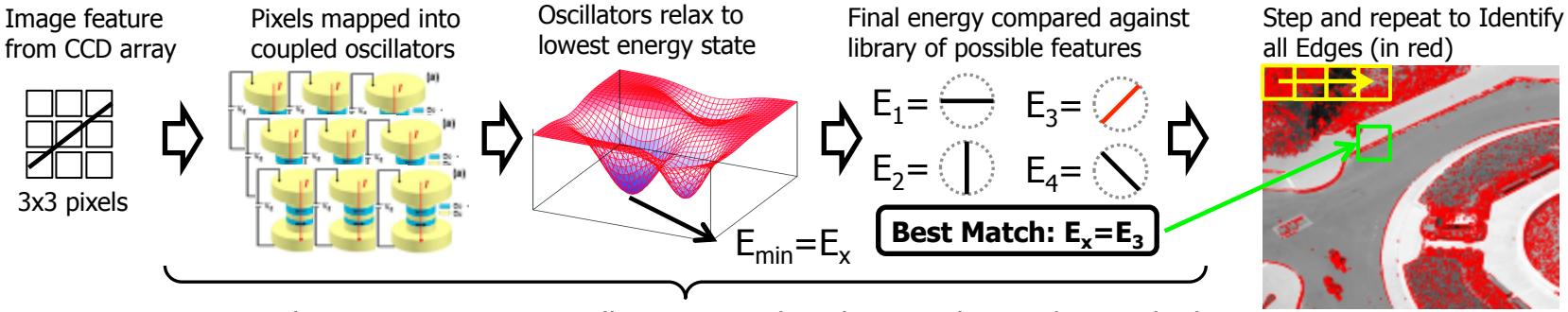
UPSIDE Unconventional Processing of Signals for Data Exploitation

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



UPSIDE eliminates computationally intensive digital CMOS dot product multiplication

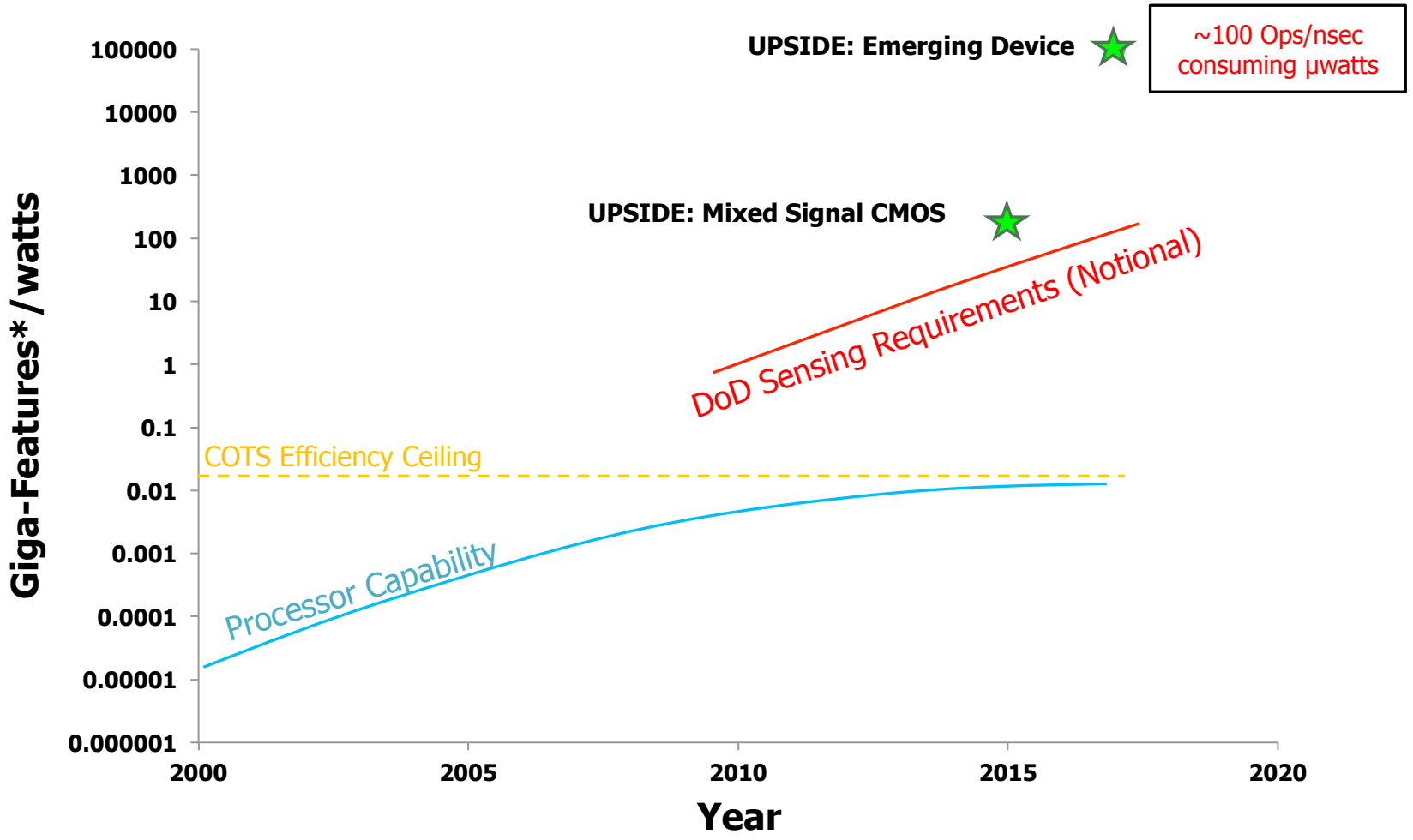
DARPA Neovision2 – Stanford Tower Video

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

BAE Systems – ARGUS IS

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

UPSIDE: Performance Goals



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

UPSIDE Program Tasks

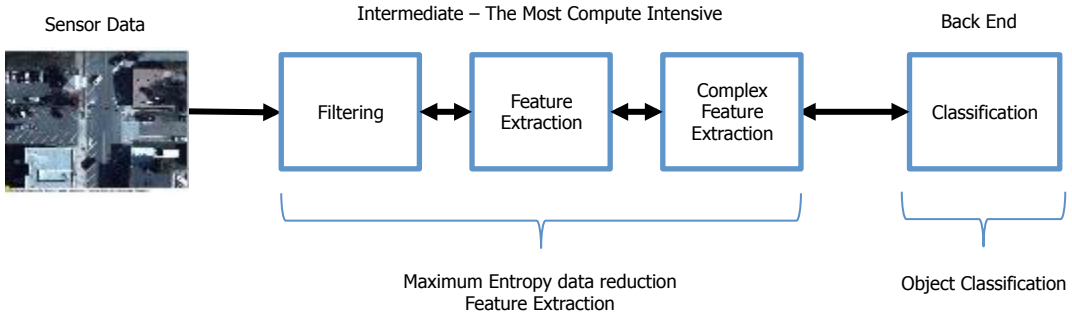
Image Processing Pipeline an application driver

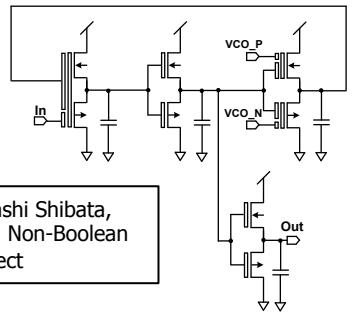
UPSIDE Task 1
IM Development & IPP Implementation

UPSIDE Task 2
Mixed signal CMOS implementation of computational model and system test bed

UPSIDE Task 3
Image processing demonstration combining next-generation devices with new computation model

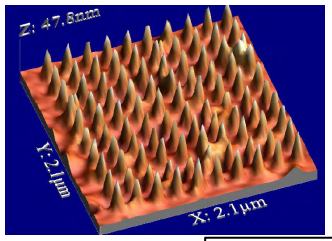
Recreate the traditional image processing pipeline (IPP) hierarchies of Inference Modules (IM)





Tadashi Shibata, Intel Non-Boolean Project

- Design and fabricate a mixed signal representation of the Inference Module in state of the art Mixed Signal CMOS
- Implement a test bed system using Mixed Signal CMOS
- Validate against IPP simulation



- Simulate the mapping of the Inference Module to specialized devices
- Determine systems level performance-price
- Show simple circuit operation

Image of CoFe2O4 epitaxial nanopillars fabricated on MgO, courtesy of R. Comes, J. Lu, and S. Wolf, University of Virginia

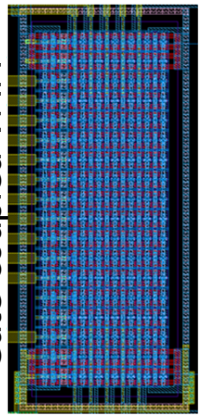
UPSIDE ARGUS-IS Image Processing Pipeline: **40GP/s, 5W**

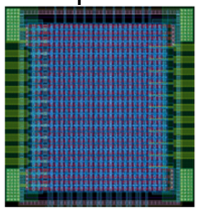
BAE SYSTEMS

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

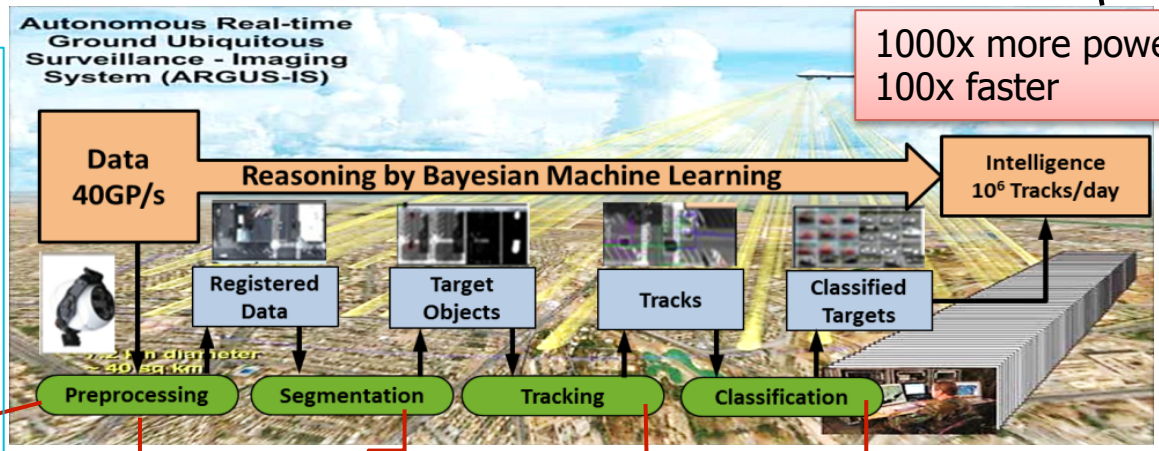
Source coupled VMM

Gate-coupled VMM

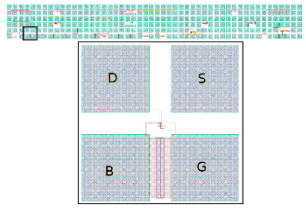


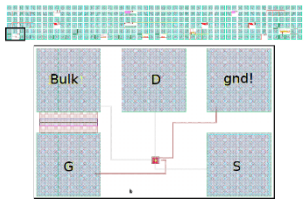


NUC/Debayer

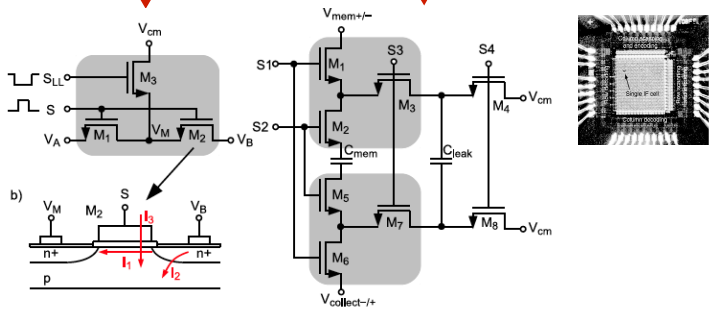


1000x more power efficient
100x faster





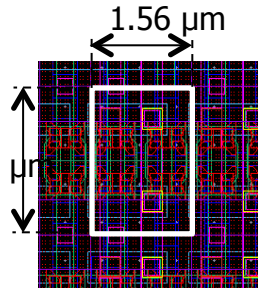
Symmetric FGMOS device to be used in analog NVM circuits



65 nm UCE Tape –out #1: October 2014, includes computing with capacitor arrays

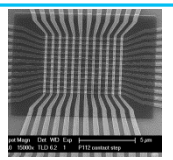
"Training and operation of an integrated neuromorphic network based on metal-oxide memristors", M. Prezioso et al., Nature Letter, 7 May 2015, Vol. 521

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



Fabricated and tested on May 14

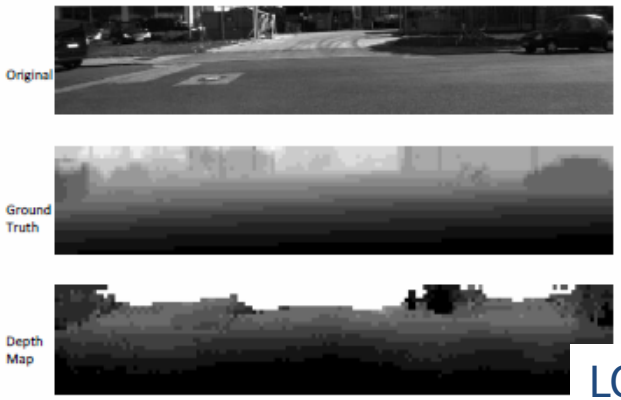
- First demo of this kind
- Physical memristor x-bar implementation
- 3x3 binary input images
- 4 classes (X,I,C,V)

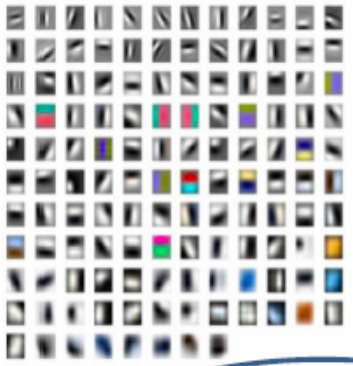


Micrograph of 12x12 Mermistor Array

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

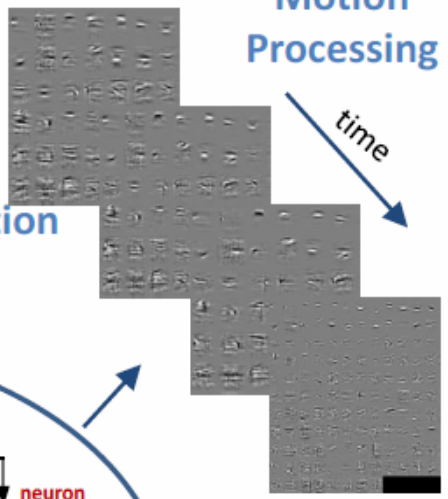
Depth from Stereo



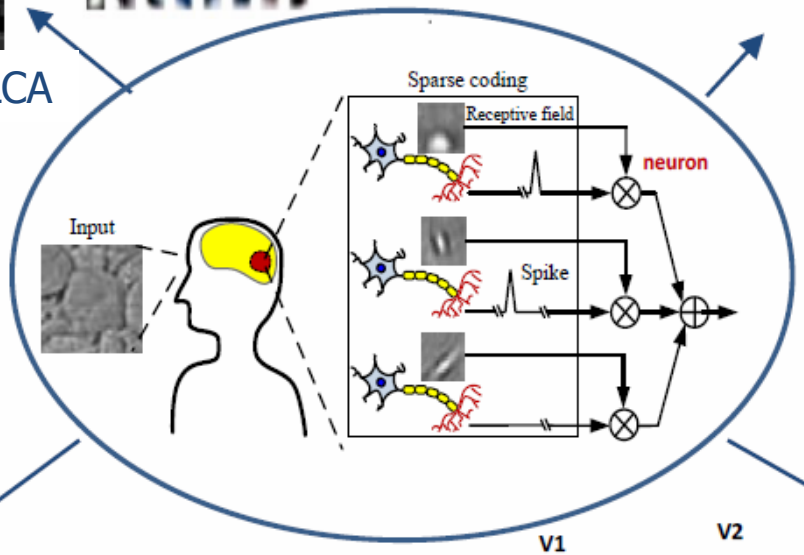


Self Organization

Motion Processing

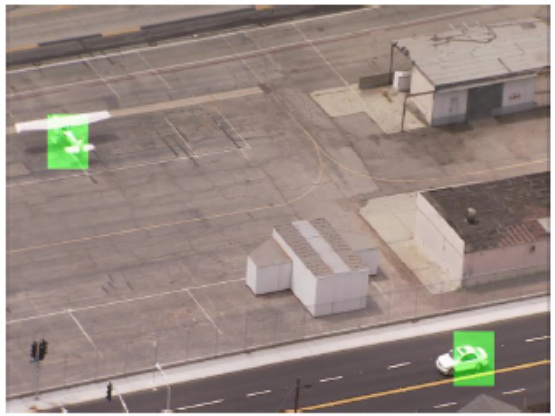


LCA



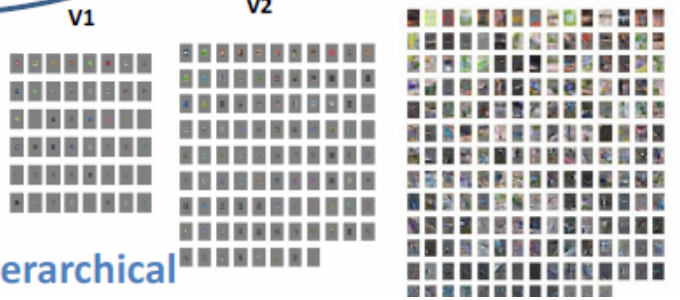
Analytics

Sensing



Object Detection

Neural Inference Module



Hierarchical

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

Task 1 Result: Accuracy of UPSIDE Method

UPSIDE Objective: Lower power with no loss in accuracy.
Task 1 Result: UPSIDE probabilistic computation approach yields equivalent image-processing accuracy to high-precision conventional digital method.

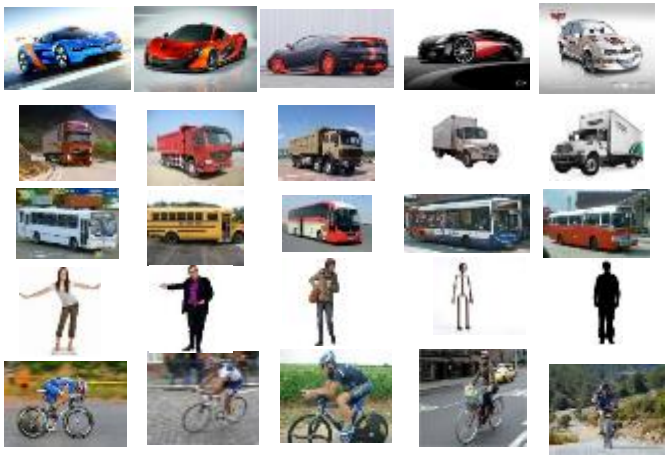
Conventional IPP

UPSIDE IPP

	True Pos UPSIDE	False Pos UPSIDE	False Neg UPSIDE
	Conventional Score (#)		
Car	1050 (1039)	63 (74)	2 (2)
Truck	1 (0)	12 (11)	1 (3)
Bus	0 (0)	43 (43)	1 (1)
Person	1342 (1210)	508 (623)	29 (48)
Cyclist	541 (643)	339 (278)	9 (18)

UPSIDE Accuracy Score = 0.677 (vs 0.675)

Google Images
50 unique object centered images/class



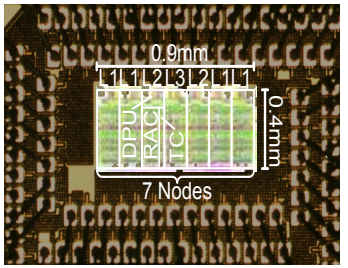
	True Pos UPSIDE*	False Pos UPSIDE*
	Conventional Score (#)	
Car	36 (31)	2 (7)
Truck	26 (27)	6 (5)
Bus	33 (34)	4 (3)
Person	38 (39)	1 (0)
Cyclist	36 (35)	5 (6)

UPSIDE Accuracy Score = 0.904 (vs 0.888)

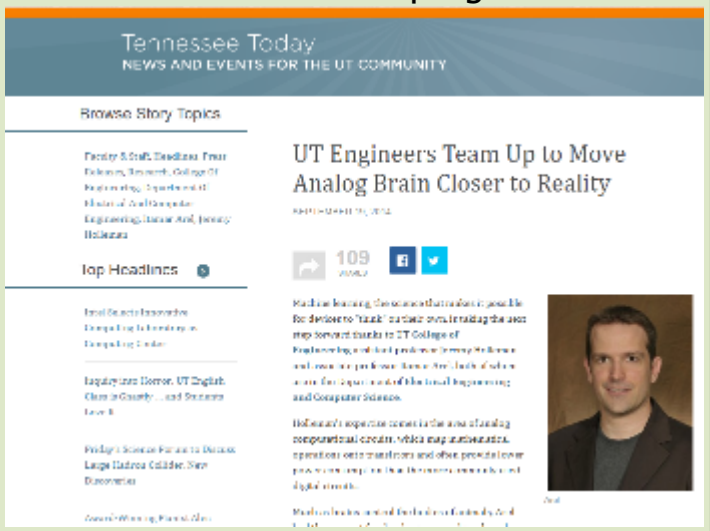
Task 2 Result: Deep Learning Analog Chip

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

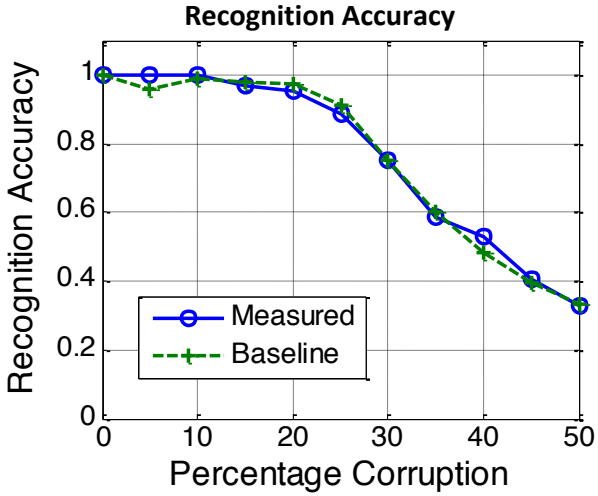


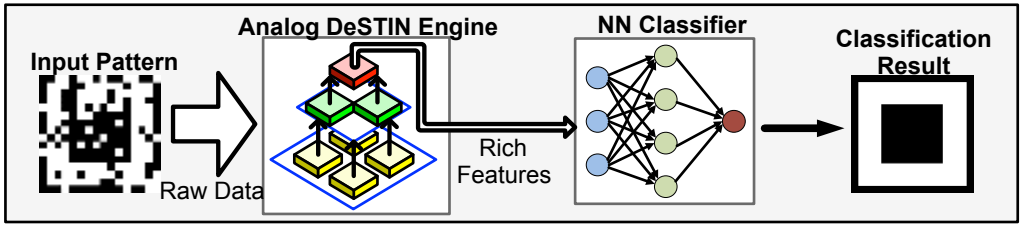
J. Lu, S. Young, I. Arel, J. Holleman, "A 1TOPS/W Analog Deep Machine-Learning Engine with Floating-Gate Storage in 0.13um CMOS," IEEE Journal of Solid-State Circuits, Vol. 50, Issue 1, pp. 270-281, Jan. 2015.

Performance & Efficiency

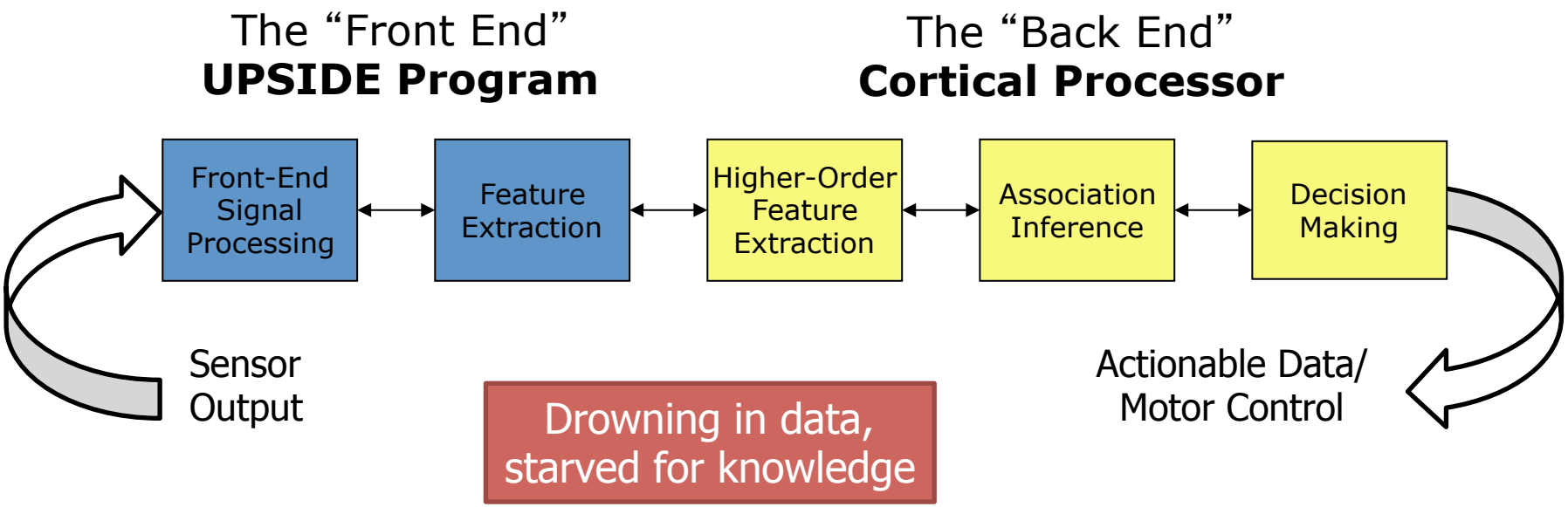
Accuracy comparable to s/w, with **282x lower training energy** than synthesized custom digital equivalent.

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





Sensor Processing Pipeline: A Data Analysis Crisis



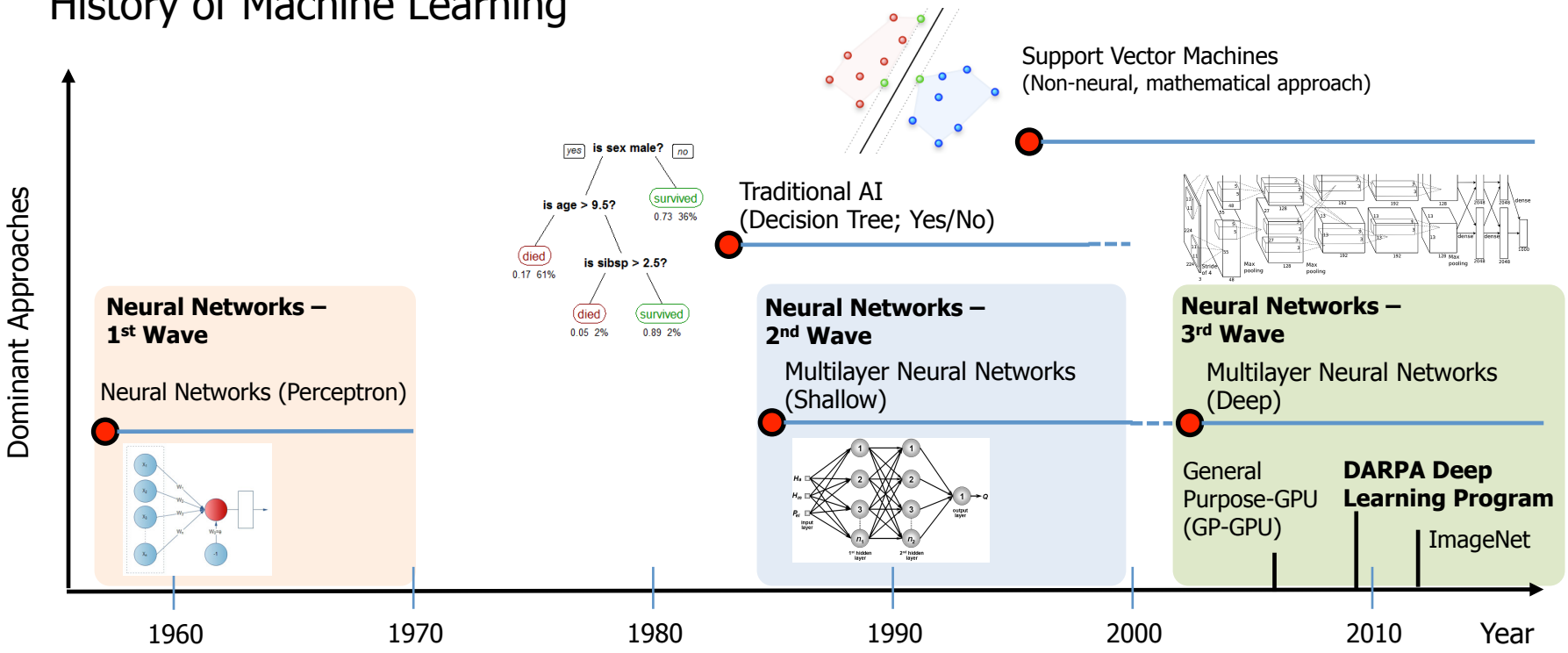
- Sensor data bandwidth exceeding processing capabilities, particularly for embedded systems
- Data become more knowledge / context intensive, containing both spatial and temporal information, as they move through the pipeline
- Current computational approaches do not adequately represent complex spatial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

But First Some Background

- Intelligent computing has been highly dependent on the Moore's law bounty
- And although there have been algorithm improvements, the 5 orders of magnitude increases in processor speed and memory capacity over the last 40 years account for most of the the ability of computers to operate in the real world
  - We have not, as yet, been able to implement truly intelligent computing
  - We still have significant problems:
    - Drowning in sensor data
    - Complex weapon systems that have exceeded our ability to write error-free software
    - Cybersecurity as a major system liability
  - And, we can no longer rely on Moore's Law to save us
- Neural-inspired computing is still the best (perhaps only) answer, but its near-term future is far from obvious
- Emerging are a few successful applications built primarily from neural components and which is mainly driven by Deep Learning and GPUs

A Brief History

History of Machine Learning



2008/9 – DARPA identifies and invests in the potential of deeply-layered neural networks or Deep Learning machines for rapidly analyzing sensory input and identifying salient or anomalous features and events

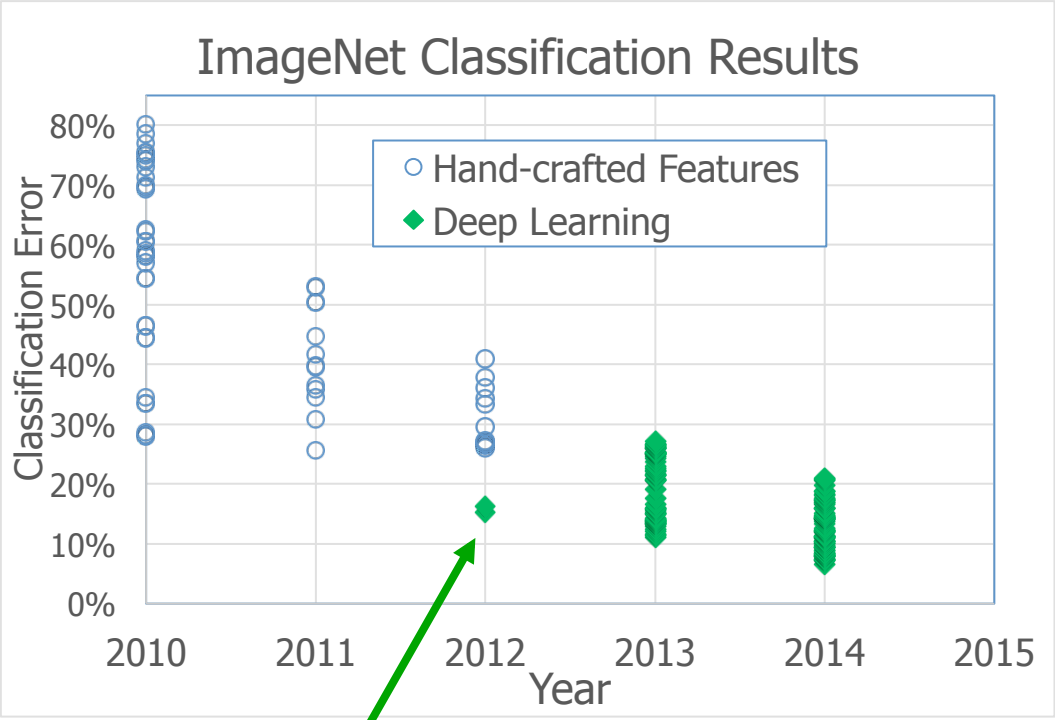
2012 – ImageNet win spurs new commercial interest in Machine Learning Algorithms for Image Processing

Breakthrough: Machine Learning Dominates ImageNet

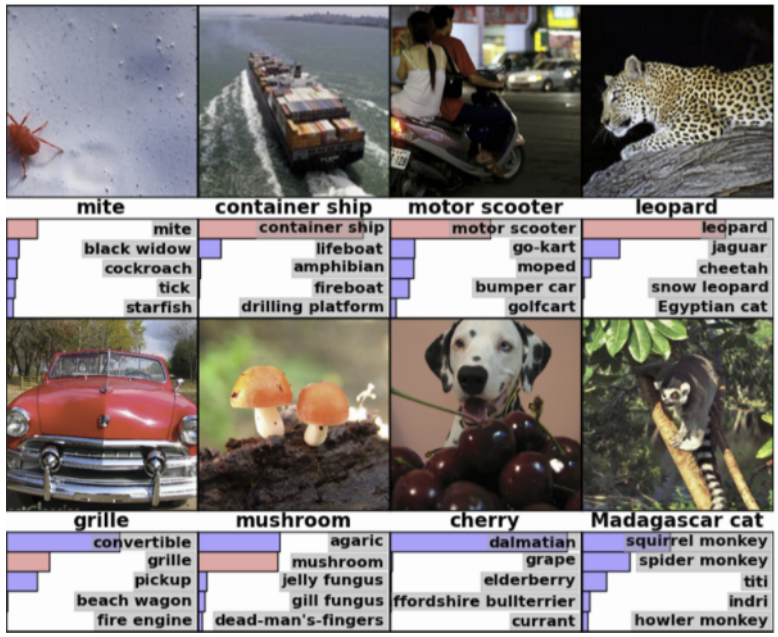
ImageNet: The most challenging annual competition for object class identification algorithms.

- Goal: Identify object classes in the images
- Training: 1.2M labeled images depicting 1,000 object categories
- Test: Algorithms identify the 5 most probable object classes
- Less well known test: find object's position (results are poor)

www.image-net.org - Large Scale Visual Recognition *Challenge*



Deep Learning Object Classifications

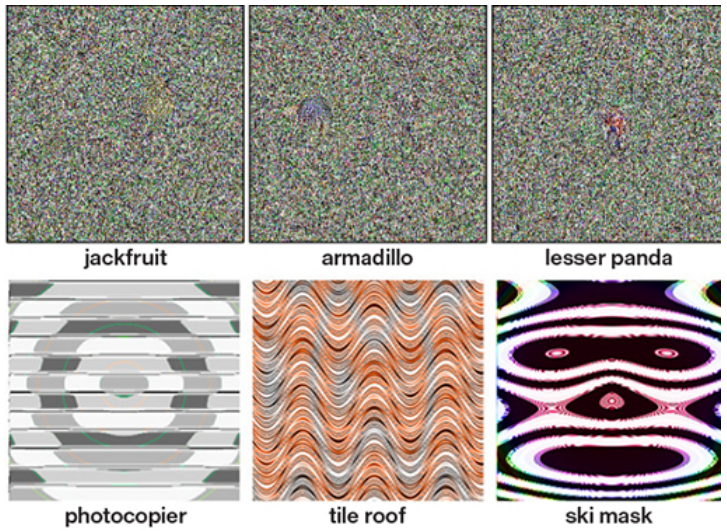


Machine Learning algorithm (Deep Learning) enters in 2012 as new approach with hand beats other machine learning with highest accuracy.

Motivation for Bio-Inspired Algorithms

The key question: is where do we go next with machine learning technology?

- There are still opportunities for improvement in the general paradigm:
  - Attention, feed-forward, feed-back
  - Real-time learning, smaller training sets, adding new information continuously
  - More efficient use of temporal information, objects are generally recognized by spatial-temporal characteristics
  - Improved scalability
  - Better capture of high order structure in data



Misclassified Test Data Examples by Non-Bio-Inspired Algorithms

"Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", Anh Nguyen, Jason Yosinski, and Jeff Clune, CVPR 2015

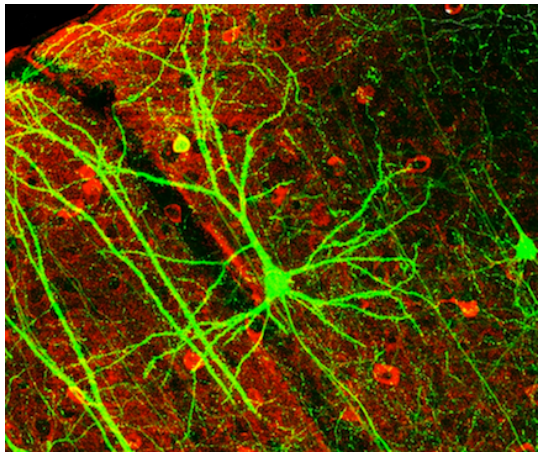
(Image Source)

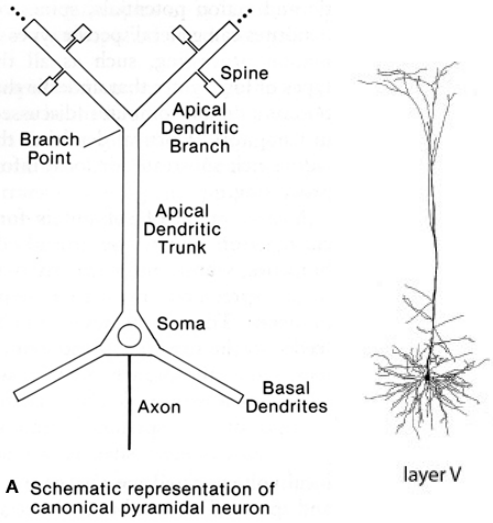
- Solution: leverage biology
  - We may not know how the brain works, but we know enough to start abstracting biological principles and modifying ML algorithms accordingly

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: 2.6×10^{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)



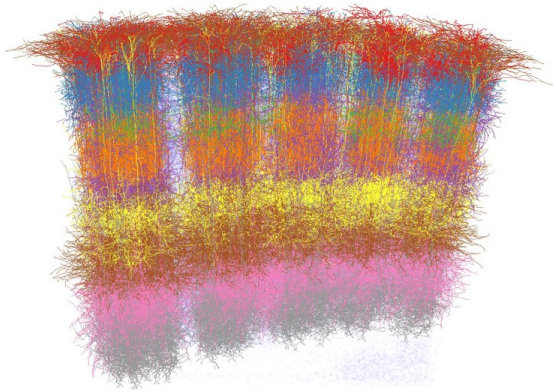


A Schematic representation of canonical pyramidal neuron

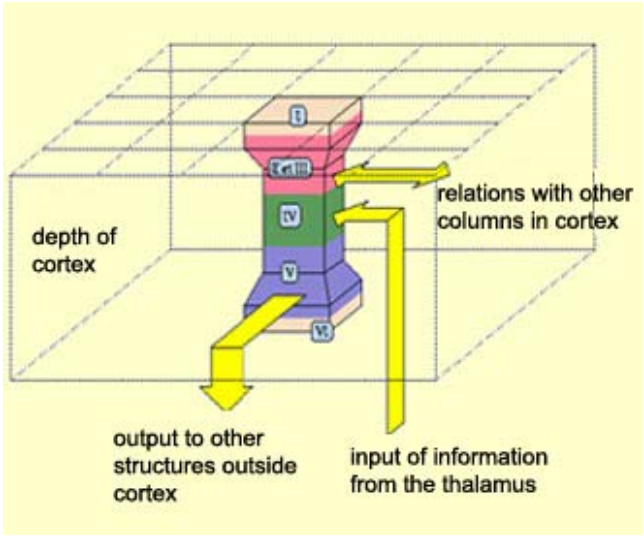
"Pyramidal neurons: dendritic structure and synaptic integration", Nelson Spruston, Nature Reviews Neuroscience 9, 206-221 (March 2008)

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



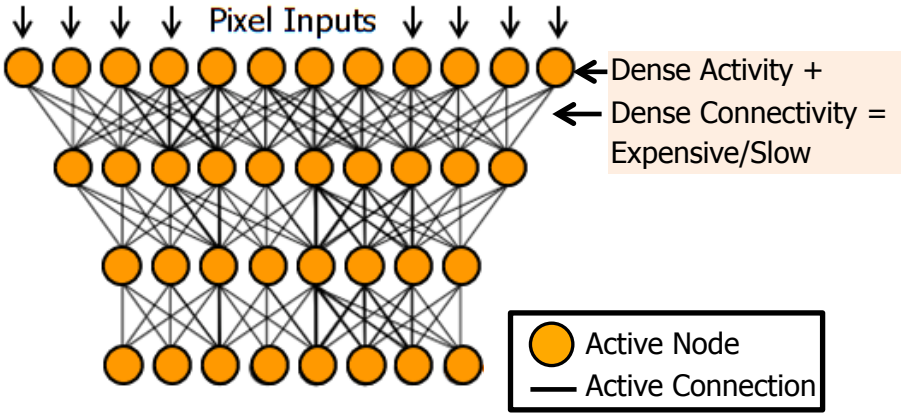
Cell-type-specific 3D reconstruction of five neighboring barrel columns in rat vibrissal cortex (credit: Marcel Oberlaender et al., Cerebral Cortex October 2012;22:2375±2391)



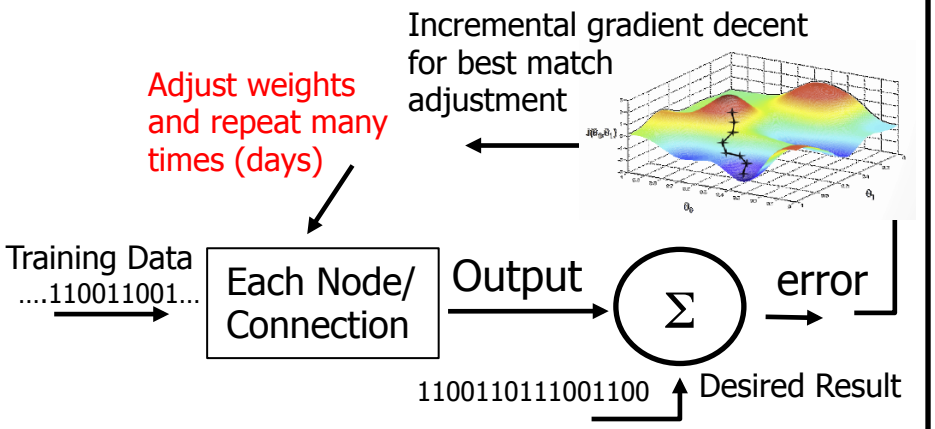
The Cortical Column: http://www.metz.supelec.fr/metz/recherche/ersidp/Projects/Cortical/Root.html

Faster Learning, But Is It Good Enough?

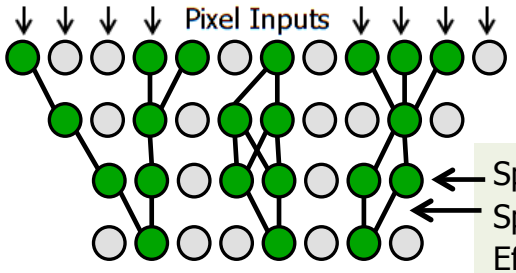
Deep Learning



Deep Learning – Learning Cycle

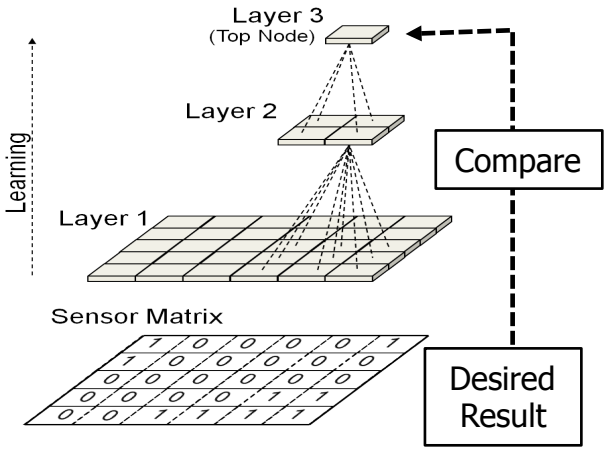


Hierarchical Temporal Memory (HTM)



- Numenta
- Neurithmic
- ART

HTM adaption – Single-pass Learning



HTM performs on-line dynamic adaption and learning. New objects can be added continuously.

Good fit for managing data complexity in DoD applications.

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 incrementally 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

DARPA Cortical Processor Concept

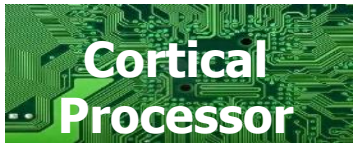
Heilmeier Questions

What are you trying to do?

- Create an algorithm and computer architecture that mimics selected characteristics of human intelligence, such as learning and pattern recognition, to address the challenges of data recognition, control, and complexity in continuously evolving environments for DoD systems.

Fast Sensor Data Stream

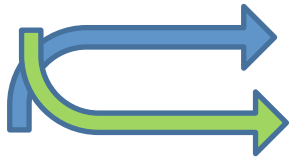
UPSIDE



Continuous Real Time Learning

Predictions

Anomalies



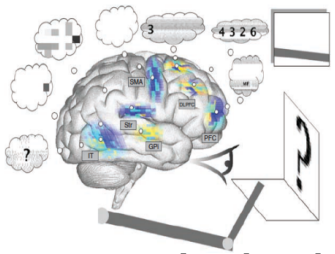
Automated Action

How is it done today?

- The most common data recognition approach is to utilize application specific, hand-crafted and hand-tuned algorithms
- Learning approaches have shown promise in solving a wider range of problems in less constrained environments, but require high-precision and long compute times limiting their ability to learn large data sets rapidly or adapt in real time in the field

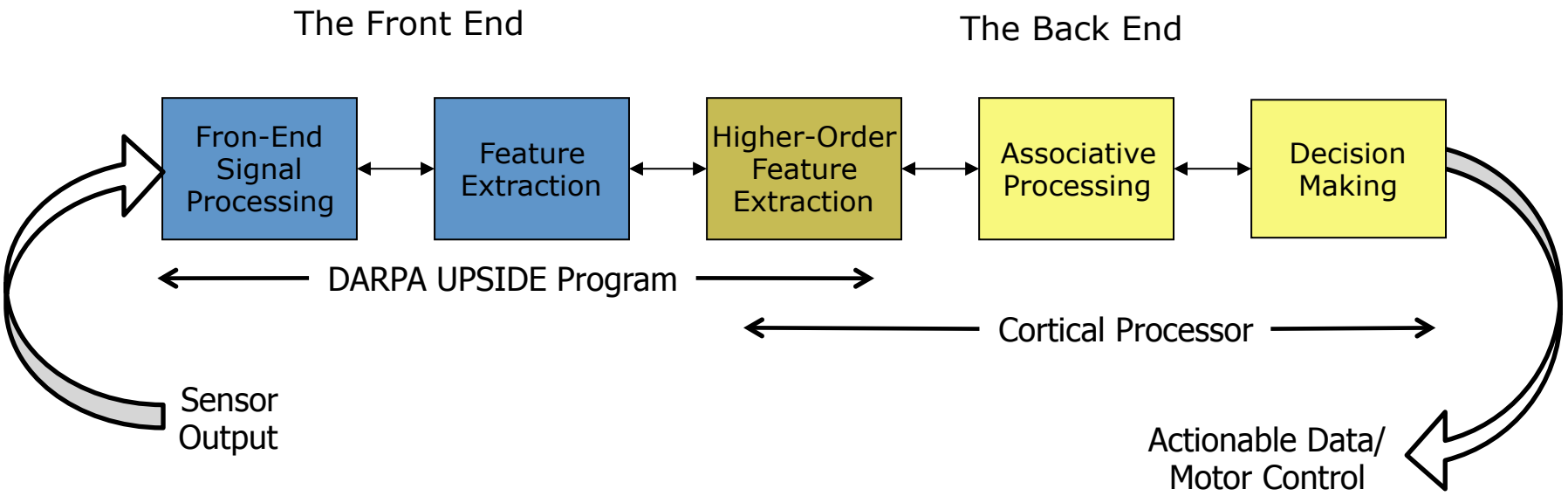
What is the new approach?

- Brain-inspired neural algorithms that use a low-precision, hierarchical, temporal memory structure that rapidly evolves with changing data and eliminates the need for long training times
- Optimized silicon architectures for these algorithms will result in high performance and low power for real-time, embedded system operation or large-scale applications



Computational Models and Hardware Based on Neuroscience

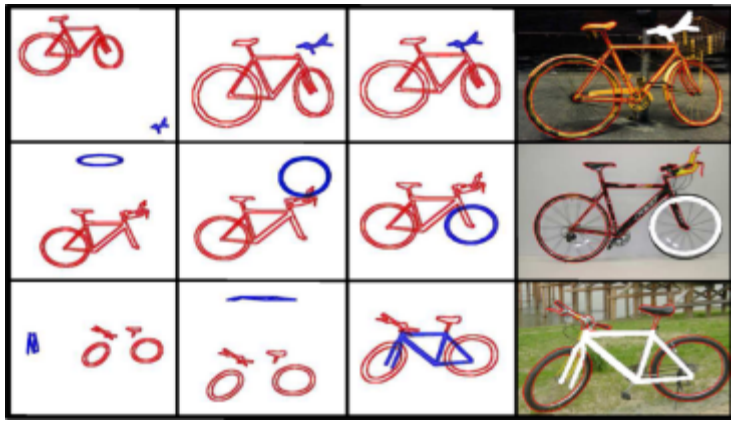
The Back End – Opportunity!



- We know a lot about the front end of most sensor data processing problems. Although there are hardware constraints, there are also solutions – the current crop of ML algorithms do reasonably well there
- It is the back end, capturing complex relationships in the data where we struggle for broader, more capable solutions
  - We have a limited set of algorithms in this toolbox, most require extensive manual development

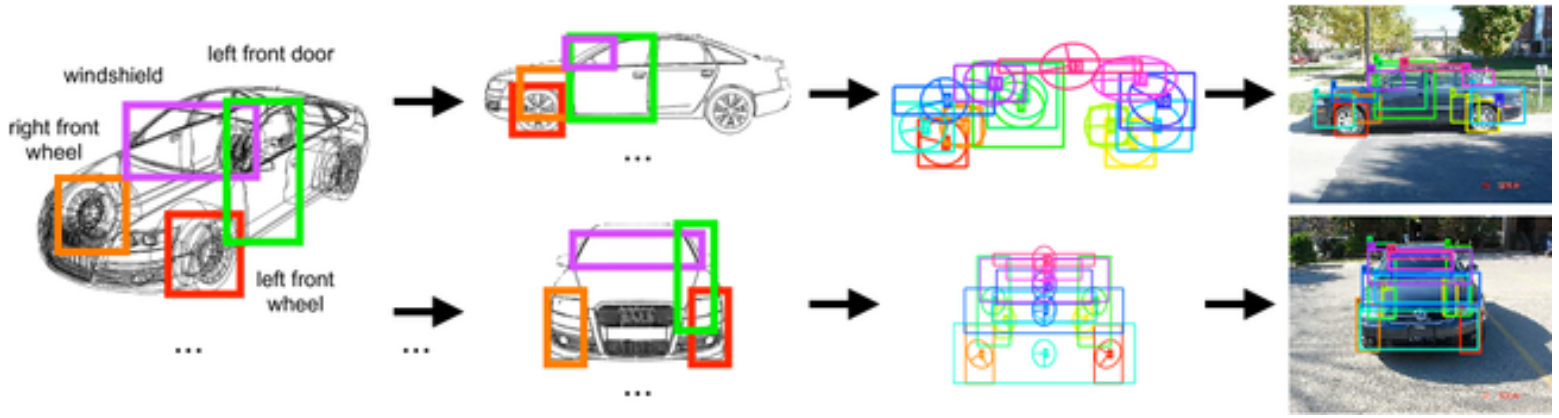
A Back End Opportunity

- One of the most important potential capabilities of hierarchical sparse models is the ability to capture the high-order structure that is a fundamental to real world data
- Creating and acting on complex structured data remains a significant problem and it is a basic requirement for any approach to cognition
- Traditional solutions are still limited in this regard, for most systems the needed data structures are hand crafted



Moving from the back of the brain to the front – Gill Pratt

Object Detection: Need algorithms that have the potential to capture complex, subtle structure in the data and to recognize parts and their relationships

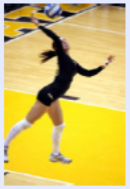


Structure in Data

A:


Tennis forehand


Croquet shot


Volleyball smash

. . .

O:

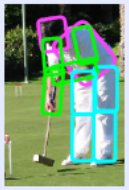

Tennis racket

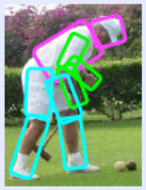

Croquet mallet

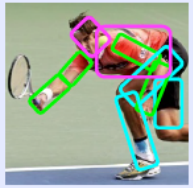
Volleyball

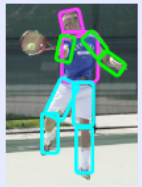
. . .

H:







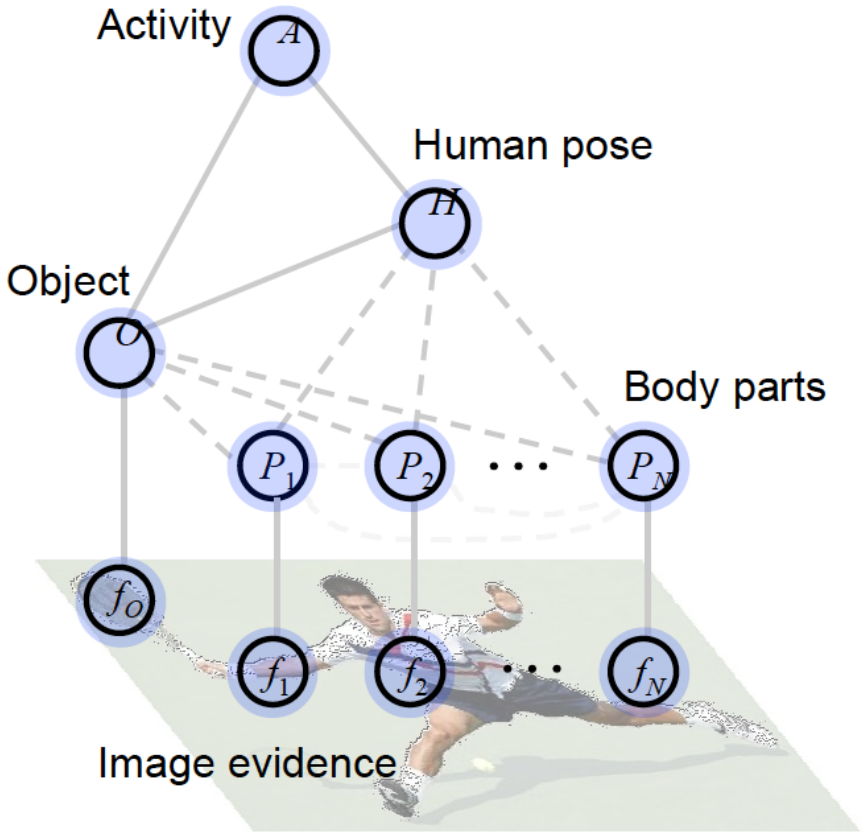


Intra-class variations

- 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]



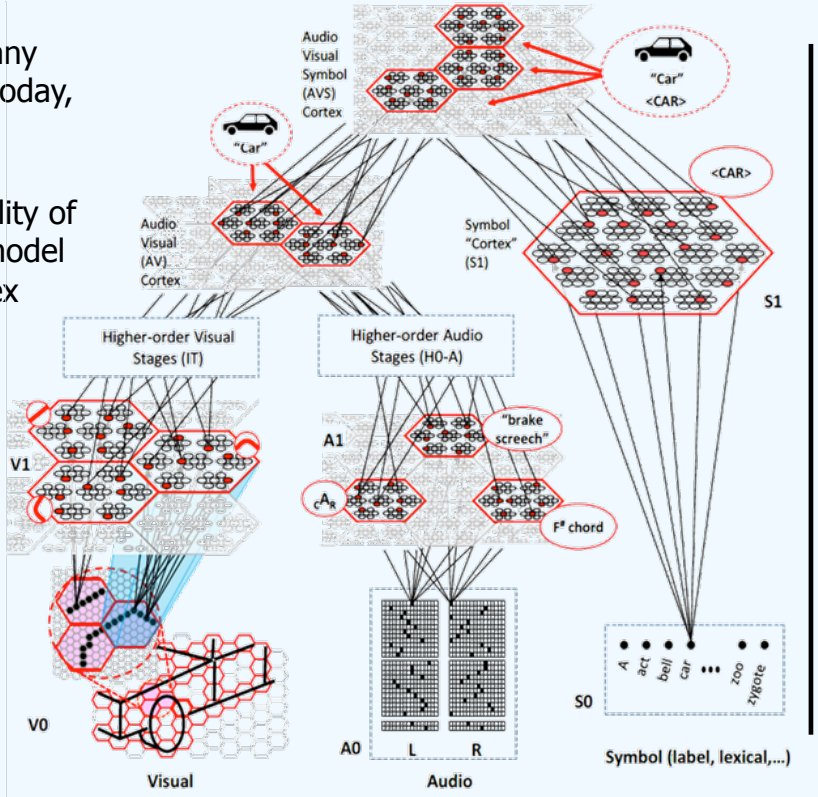
Yao and Fei-Fei, CVPR 2010

Sensor Fusion – Leverage Structure in Data

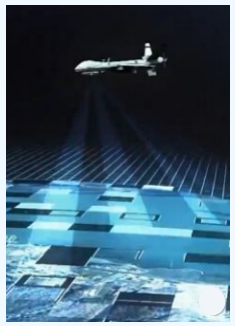
Cortical-like algorithms have the potential to solve the most challenging DoD sensor problems (*Sensor streams can be gracefully added dynamically in the field)*

Sensor Data Fusion

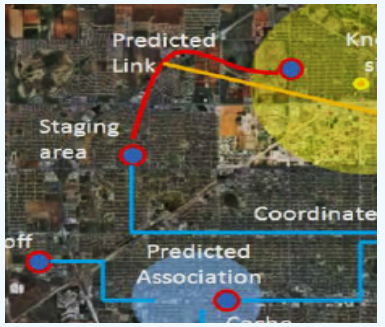
- Not possible with any neural algorithms today, nor with traditional techniques
- Creates the capability of using learning to model and control complex systems
- Helps manage signal and system complexity by automating higher order relationships



Sensor Applications



Surveillance imaging



Scenario awareness Complex tructure

Tracking convoy of vehicles

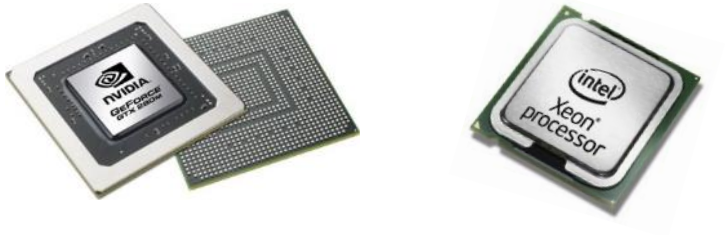
Mapping Bio-Inspired Algorithms to Hardware

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**

Example Hardware Options

- **Good:** COTS (FPGA and GPU)

- **Better:** Available Hardware (non-COTS)
  - University of Manchester SpiNNaker, ARM nodes, custom spike routing hardware, 130nm 16 ARM core, ASIC, as part of HBP they will move to a newer process
  - TrueNorth – 1M neurons, 256M connections, 28nm, not optimized to our algorithms, connectivity structure, and does not implement learning, but an enhanced learning version may be possible
  - BrainScales portable processor
  - Sparse processors
  - Micron AP (Automata Processor) – a regular expression processor for localized pattern match

- **Best:** New Architectures
  - Paul Franzon, NCSU, Cortical Processor, architecture requirements

Cortical Processor Hardware Possibilities

Deep Learning

- Does not train in real time
- Requires extensive off-line training
- Adding new objects essentially requires re-training entire network

Cortical Processor

- Operates in real time
- Performs on-line dynamic adaption and learning
- New objects can be added continuously

Off-line Learning

Fixed Training data

On-line Adaptation

Learning during operation

Best hardware mapping is GPU
- Double precision floating point operations
- All nodes active

= 2.3 Images/sec, 212 Joules/image

Ref (1)

Maps well to custom HW
- Fixed point precision
- Sparse node activation

= 1000 Images/sec, 0.0004 Joules/image

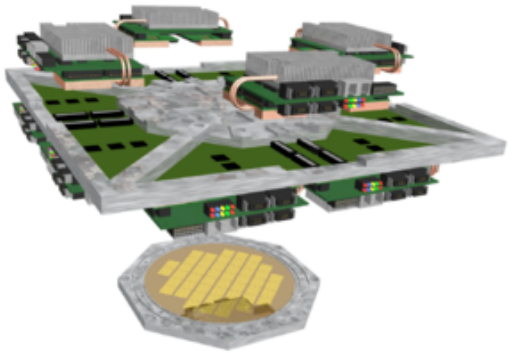
Estimate based on Ref (2)

Managing Data Complexity
500x faster
500,000x less power

Ref 1: Krizhevsky, Alex, Sutskever, Hinton. "ImageNet Classification with Deep Convolutional Neural Networks."
http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf

Ref 2: Performance/price estimate for cortex-scale hardware: A design space exploration, Zaveri, M., et al. JNN 24 (2011)

Related Efforts



- The European Human Brain Project: BrainScales (Heidelberg University)
  - Wafer-scale neurocomputers

- White House OSTP: A Nanotechnology-Inspired Grand Challenge for Future Computing
  - Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.

- IARPA MICrONS (Machine Intelligence from Cortical Networks)
  - Revolutionize machine learning by reverse-engineering the algorithms of the brain

- Chinese Brain Inspired Computing Research (CBICR) program
  - Stan Williams (HP) trip report

- IBM: Machine Intelligence: Cognitive systems which learn continuously and without supervision, predict patterns and sequences (i.e. the future) and act on it