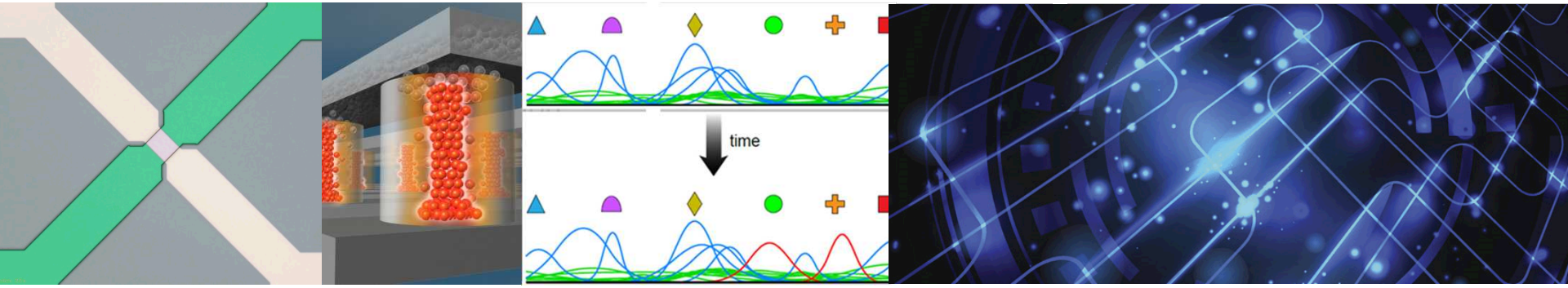
SAND2015-3240PE

Exceptional service in the national interest

Sandia National Laboratories



Hardware Acceleration of Adaptive Neural Algorithms (HAANA)

Conrad D. James, Ph.D. – Project PI
Kevin R. Dixon, Ph.D. – Project PM
Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration | LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT

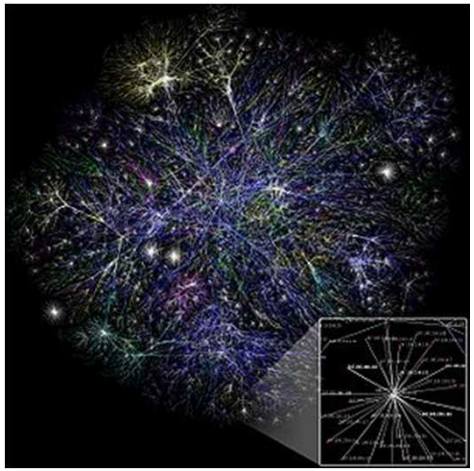
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

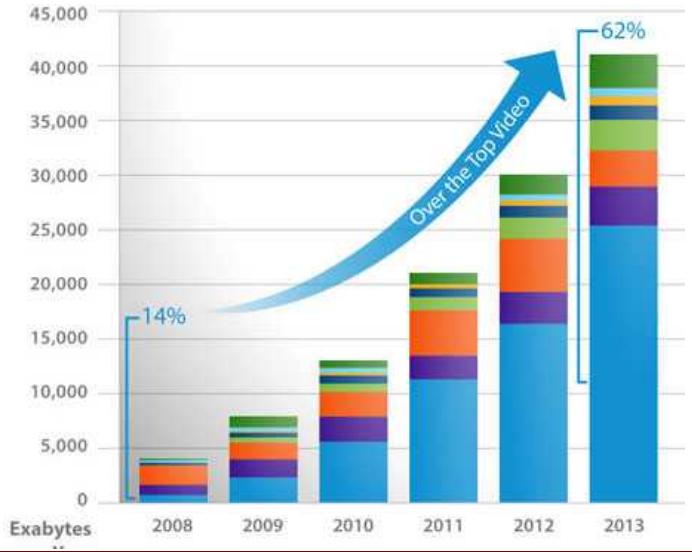
Outline

- **HAANA – hypothesis and objective**
  - Data-driven computing
  - Neural-inspired computing
- Landscape and differentiation
- HAANA project structure
  - Algorithms Core
  - Architecture Core
  - Learning Hardware Core
- Conclusions

The problem

- Detect sophisticated threats to national security
  - Space / airborne: size, weight, power constrained
  - Cyber: time constrained, rapidly evolving
- Current techniques
  - Require human analysts to scale with the data
  - Rely on ever-increasing computation time and power
- Evolutionary improvements in current performance are not sufficient!



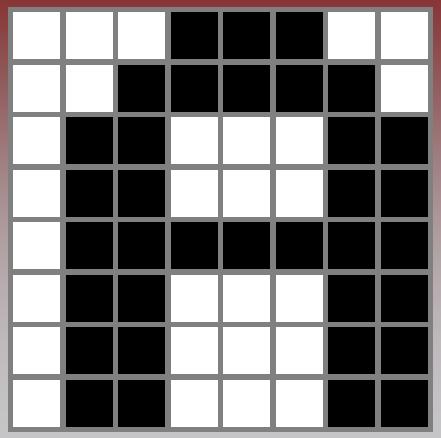


There is a strong need for data-driven computing

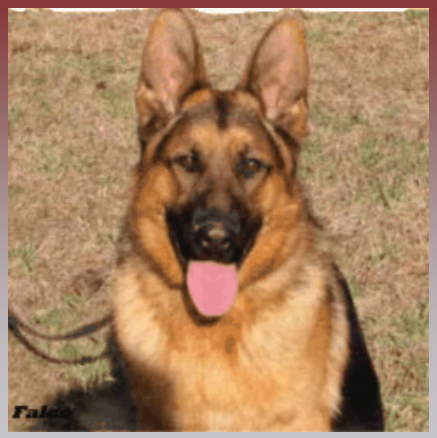
Sandia National Laboratories

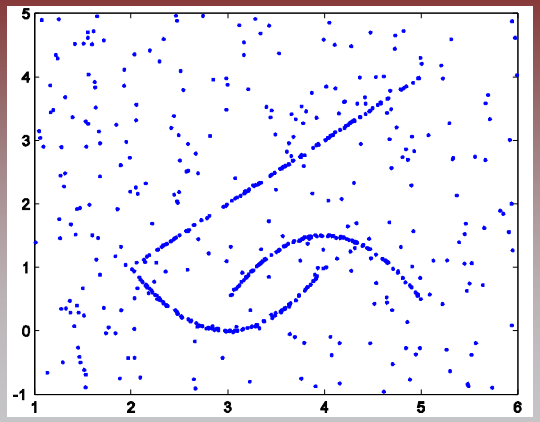
- Data is fed into the system and classifications, regressions, etc. are produced to “understand” data
  - No need to develop equation-based numerical models;
  - ***Train instead of explicitly program***
- Robust to variability and outliers, adaptable to dynamic data

Object recognition problem:



C. Lampert, VRML 2013

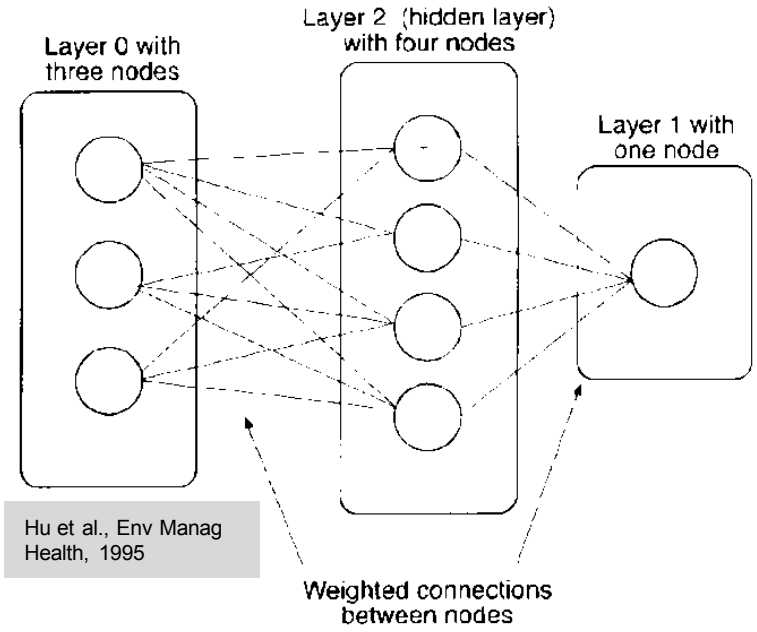




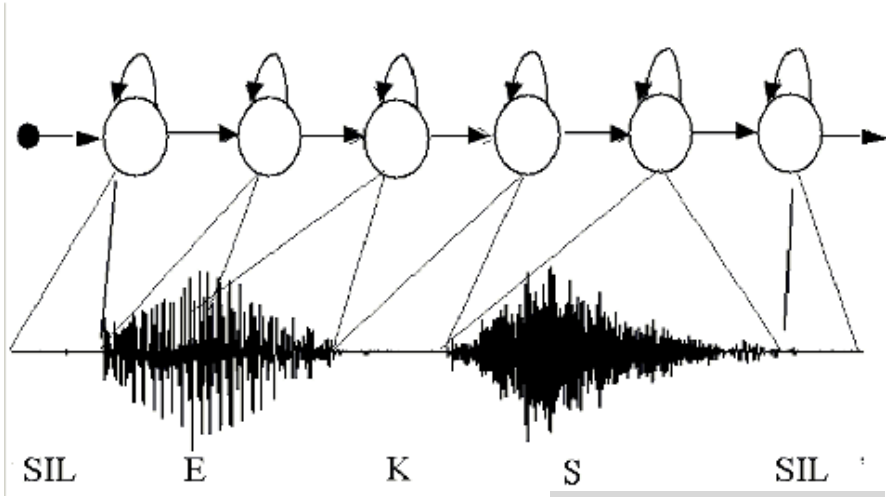
Quatch, SNL 2014

Machine learning approaches are data-driven

- Equation-driven approaches are used when the fundamental principles are well-understood → not always the case
- Issues: require large amounts of labeled data, difficult to instantiate in hardware, slow to train, have difficulty handling "spontaneity"



Hu et al., Env Manag Health, 1995

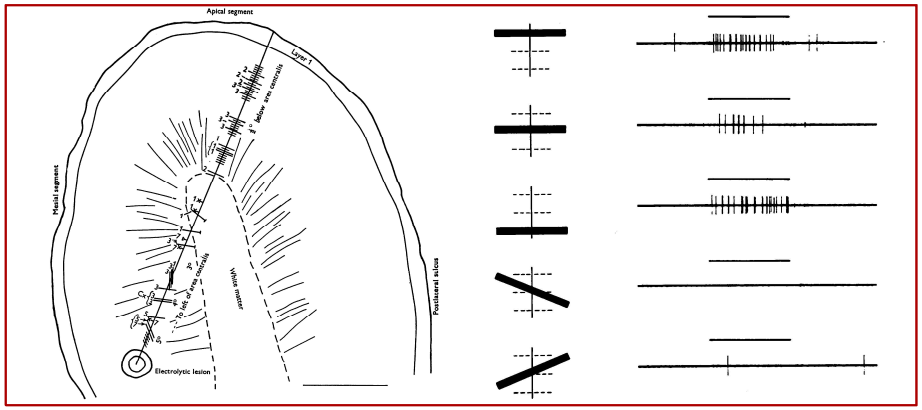


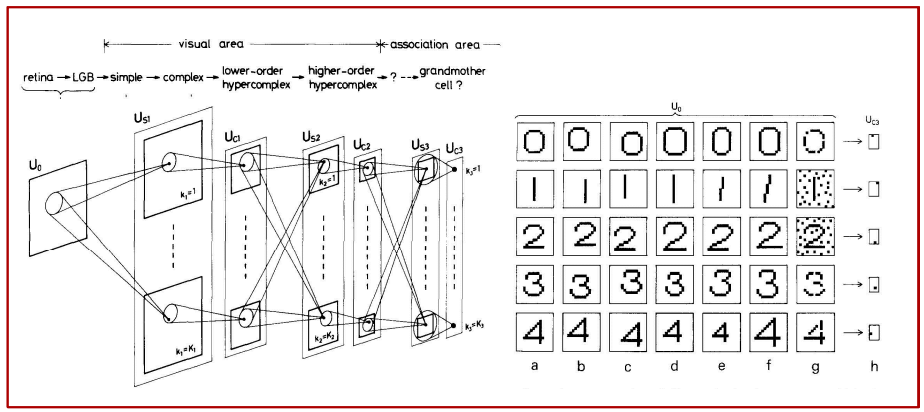
U. East Anglia Speech Group, www.uea.ac.uk/computing

A resurgence in neural machine learning research has addressed many problems…

Incorporation of neural-inspired concepts has shown promise…

Receptive fields and convolutional nets: Hubel & Wiesel 1962, Fukushima 1980





Deep learning networks: Hinton lab, NIPS 2009, ICML 2012

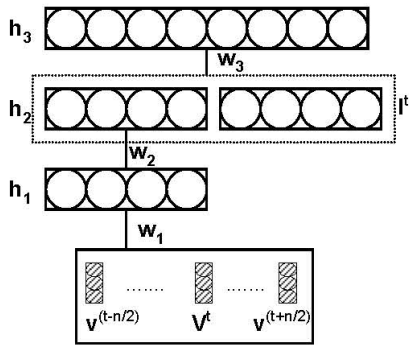
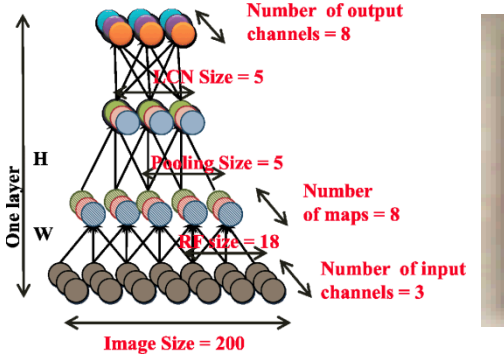
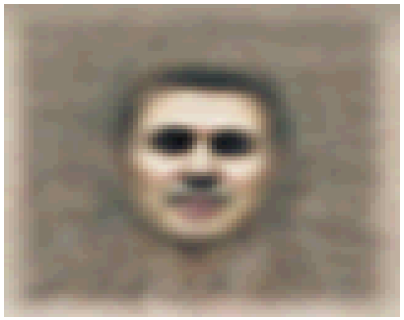


Table 4: *Reported results on TIMIT core test set*

Method	PER
Stochastic Segmental Models [27]	36%
Conditional Random Field [28]	34.8%
Large-Margin GMM [29]	33%
CD-HMM [4]	27.3%
Augmented conditional Random Fields [4]	26.6%
Recurrent Neural Nets [30]	26.1%
Bayesian Triphone HMM [31]	25.6%
Monophone HTMs [32]	24.8%
Heterogeneous Classifiers [33]	24.40%
Deep Belief Networks(DBNs) (this work)	**23.0%**

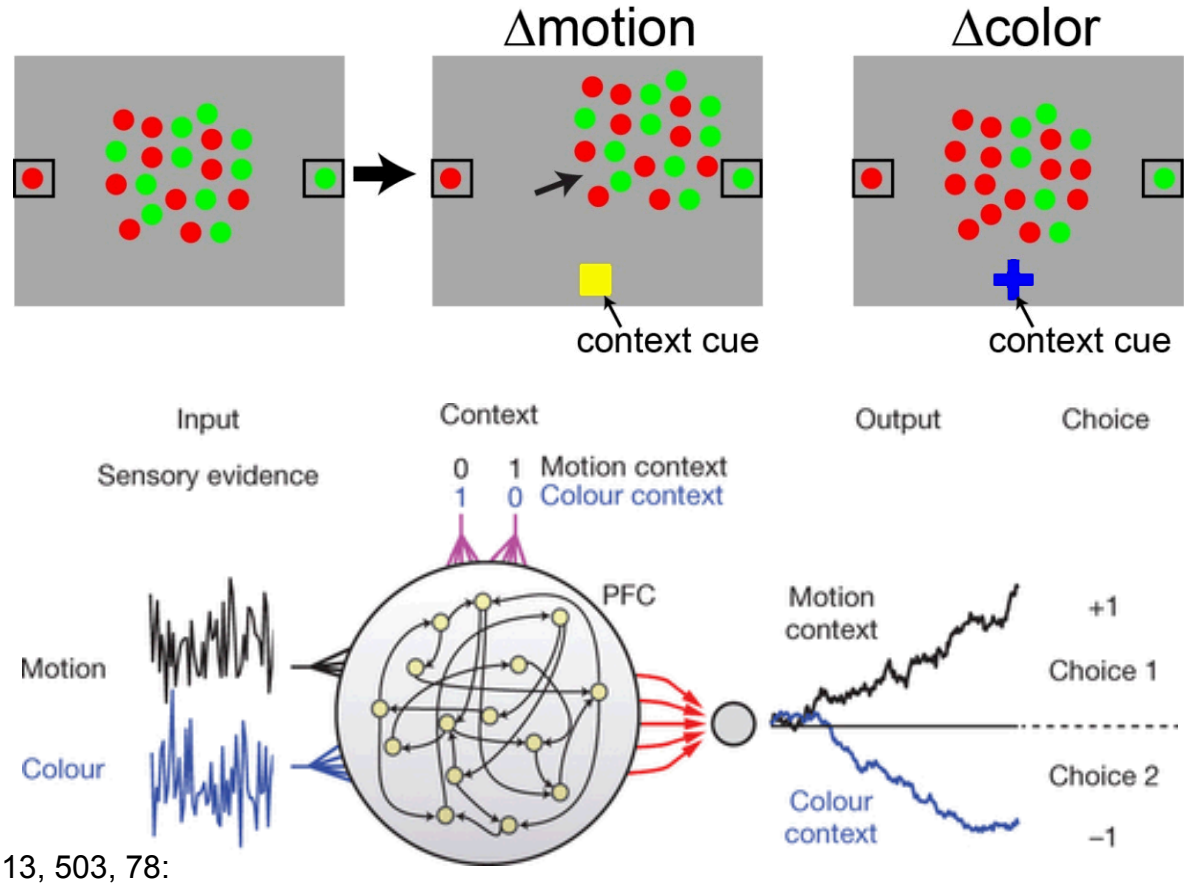




Moore's law has enabled many advances in neural machine learning approaches…

Neural-inspired **algorithms** are well-suited for detecting intelligent threats

- Multi-sensor data fusion is directly incorporated into neural computation
- Historical and contextual information are inherently integrated

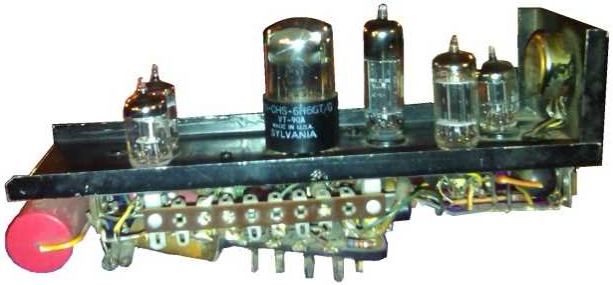


Mante et al., Nature 2013, 503, 78:

Long history of hardware innovation targeting neural-inspired computing

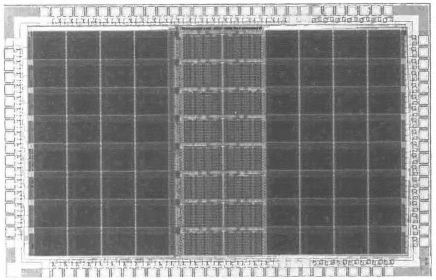
Sandia National Laboratories

SNARC; Minsky 1951:

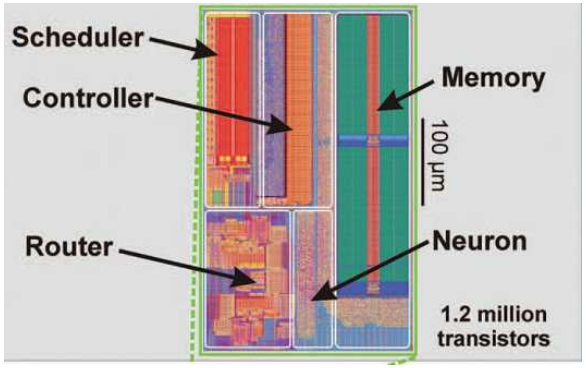


www.cyberneticzoo.com; Gregory Loan

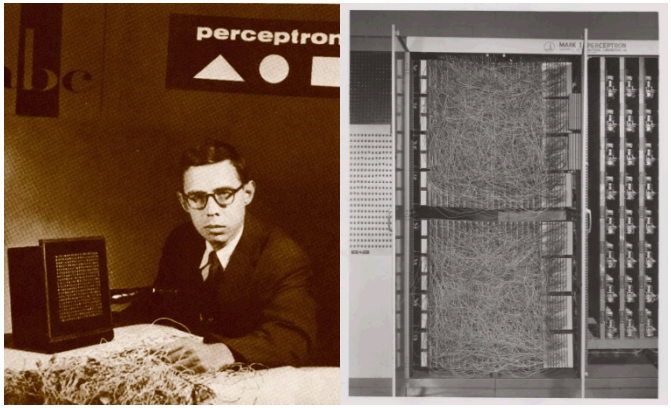
Neural chip (Graf 1990);



DARPA , IBM TrueNorth (2014):

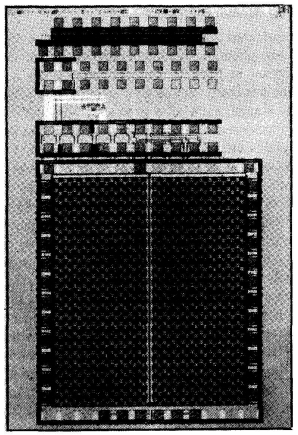


Mark I Perceptron (Rosenblatt 1960):



Arvin Calspan Advanced Technology Center; Hecht-Nielsen, R. *Neurocomputing* (Reading, Mass.: Addison-Wesley, 1990); Cornell Library;

Electronic cochlea (Lyons and Mead 1988);

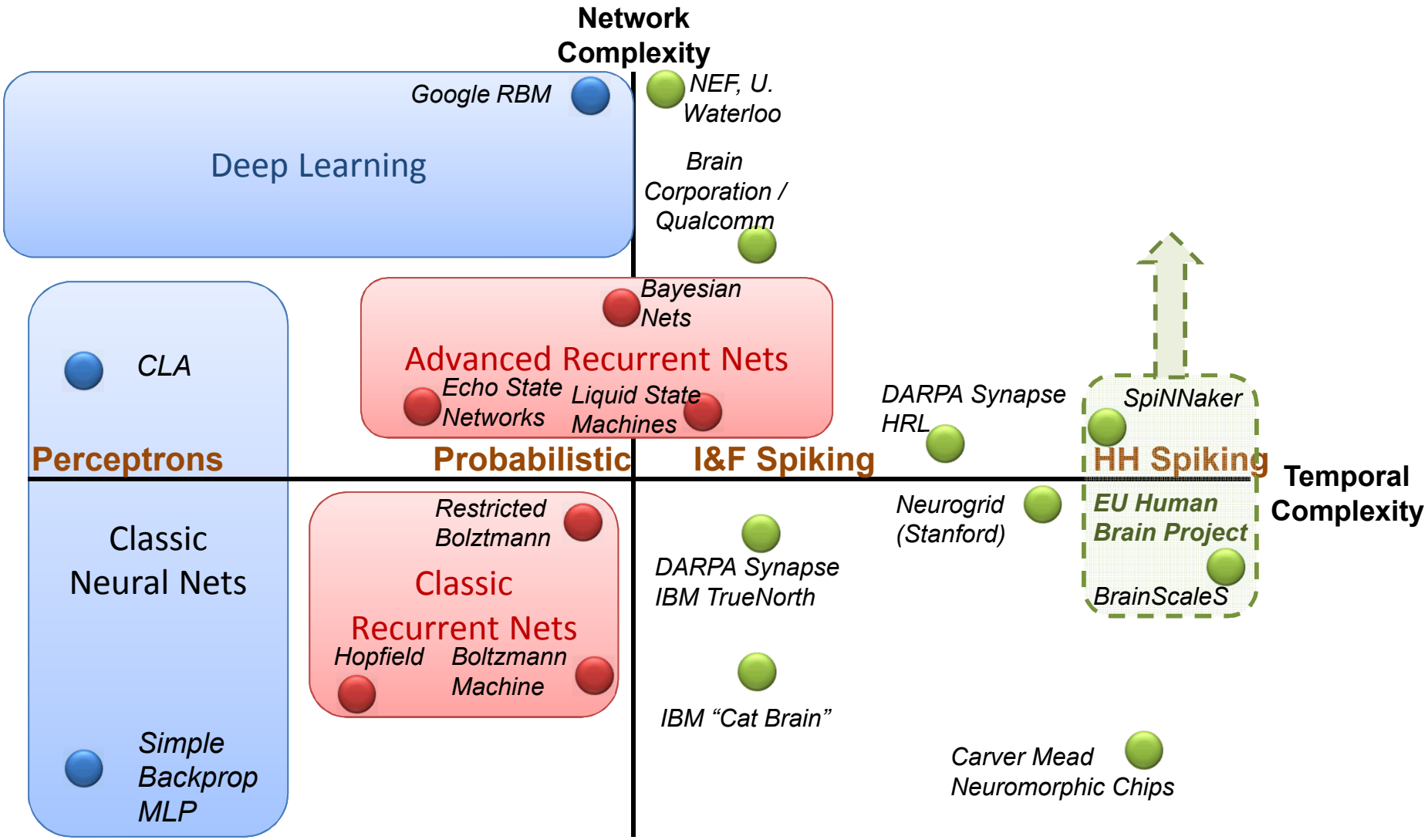


EU HBP, SpiNNaker (2014):

Outline

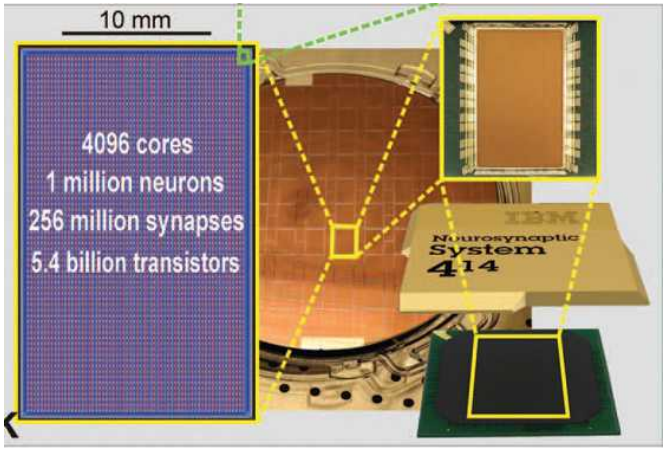
- HAANA – hypothesis and objective
  - Data-driven computing
  - Neural-inspired computing
- **Landscape and differentiation**
- HAANA project structure
  - Algorithms Core
  - Architecture Core
  - Learning Hardware Core
- Conclusions

Competitive landscape of neuro-inspired computing efforts



Limitations of the current state-of-the-art

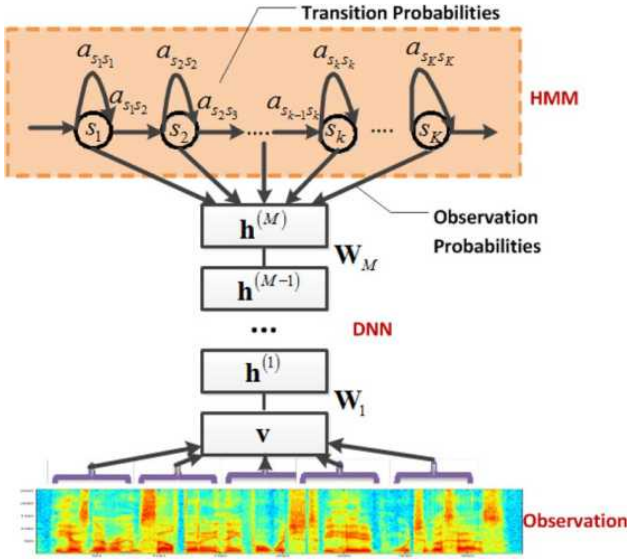
- Current hardware tends to be…
  - Highly specialized for specific problems
  - General purpose and relatively inefficient



Merolla (IBM), Science 2014

- Current algorithms tend to rely on…
  - Large amounts of labeled training data
  - Static decision boundaries

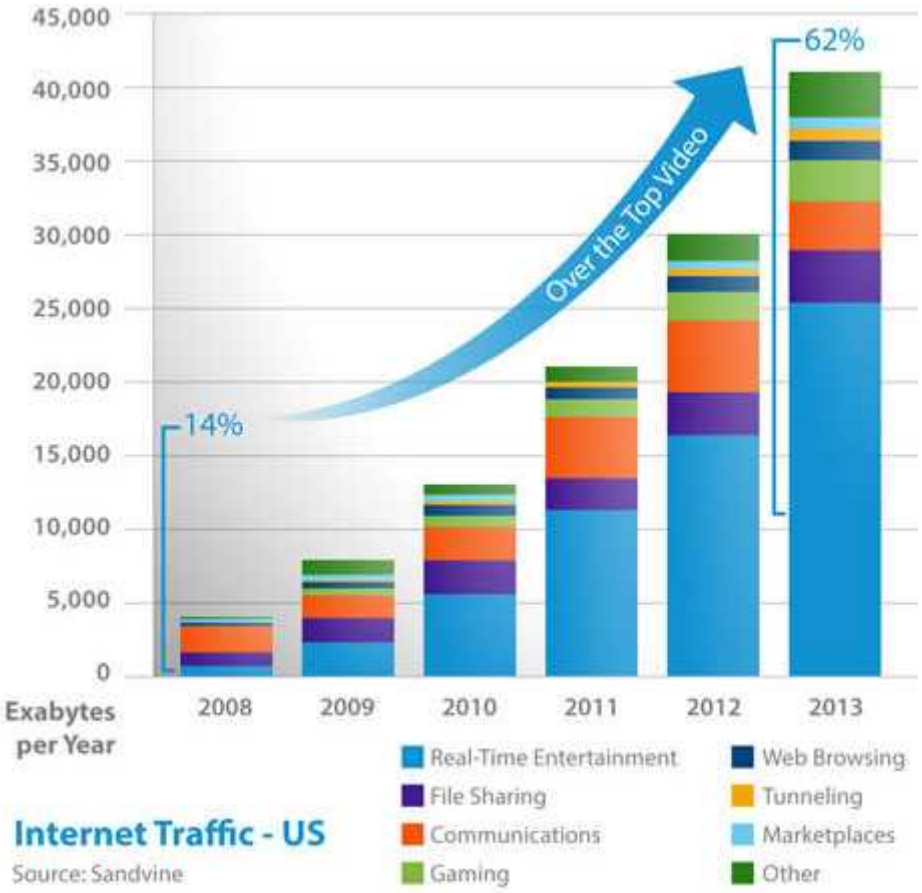
- <u>Rapidly evolving, evasive threats</u> are not the focus of academic or commercial research



Dahl, IEEE Trans Aud Speech Lang Proc 2012

A different approach is needed

- Moore's law is ending and data volume is increasing
- Algorithm development is needed
  - e.g. statistical learning theory → support vector machines
- Novel hardware development is needed



HAANA's Hypothesis and Objective

- **Hypothesis:** There exists a class of machine-learning methods, used in conjunction with neural-inspired algorithms and learning hardware, that will reduce computation time and power consumption by orders of magnitude in national-security applications.

- **Objective:** Leverage Sandia's unique capabilities to build and deploy the first versatile neural-inspired computing system that addresses Sandia's core Mission Challenges

The need for HAANA

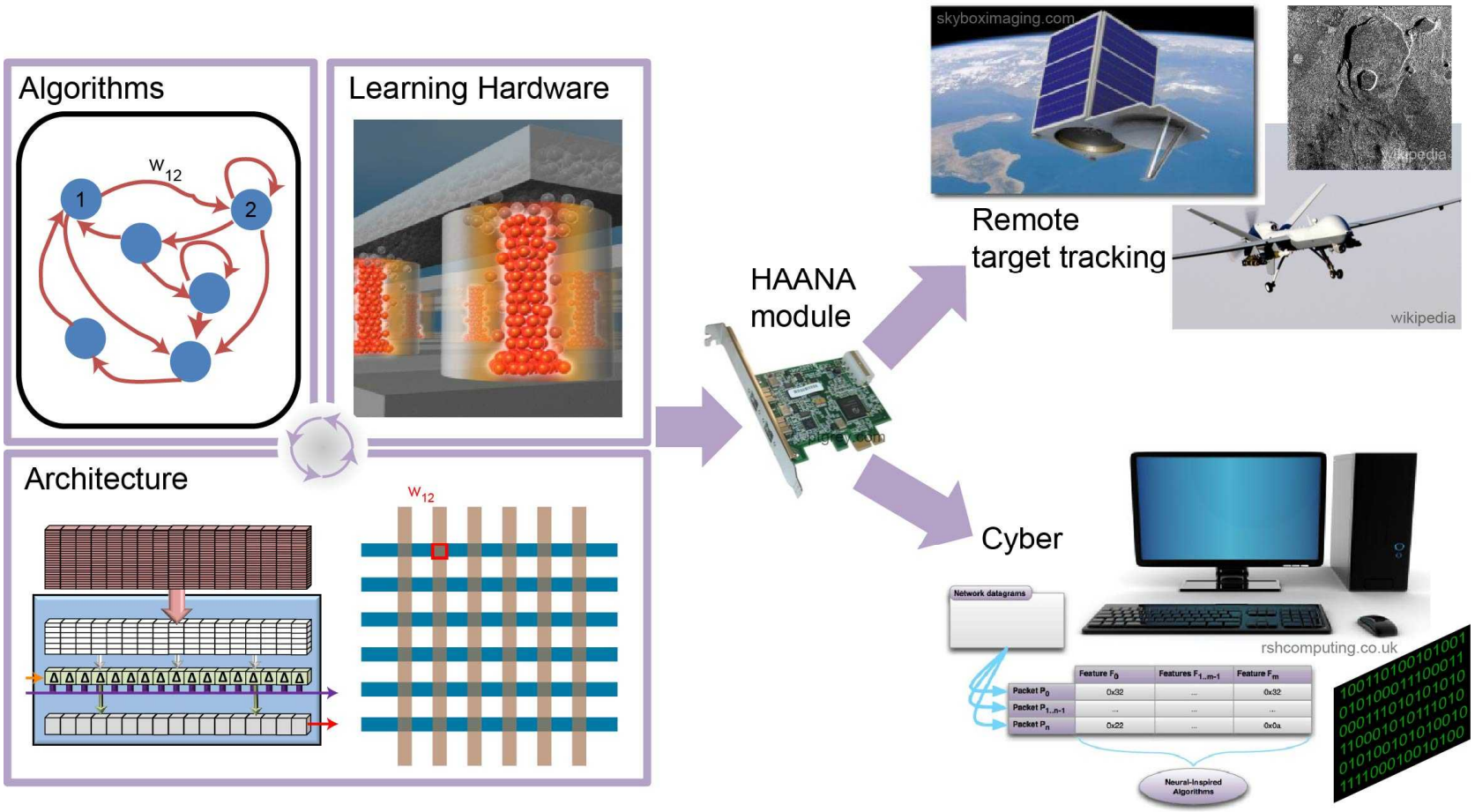
Sandia National Laboratories

- We need to transition machine learning from the quasi-static time domain into highly dynamical and multi-modal domains
- We need algorithms that leverage recent neuroscience developments
- We need hardware that can handle the temporal complexity of neural algorithms

Outline

- HAANA – hypothesis and objective
  - Data-driven computing
  - Neural-inspired computing
- Landscape and differentiation
- **HAANA project structure**
  - Algorithms Core
  - Architecture Core
  - Learning Hardware Core
- Conclusions

Real-time, low-power, small footprint, embedded threat-detection system



Neural architecture emulations: ≥ 3 orders of magnitude improvements in speed, $, power consumption

HAANA project structure

Sandia National Laboratories

Algorithms

Theory

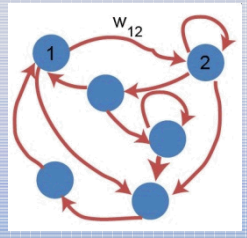
$$R(\alpha) = \int \frac{1}{2} |y - f(\vec{x}, \alpha)| dP(\vec{x}, y)$$

$$w^T x + b = \{-1, 0, 1\}$$

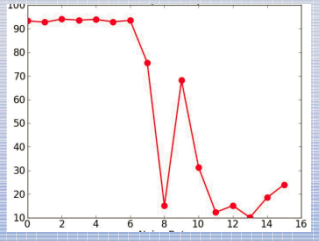
$$w_i^n := w_i^{n-1} - \eta \frac{\partial E_n(w)}{\partial w_i}$$

$$w_i^n := w_i^{n-1} - \eta \left(y(x_n, w) - t_n \right) \frac{\partial y(x_n, w)}{\partial w_i}$$

Design

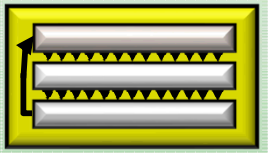


Model

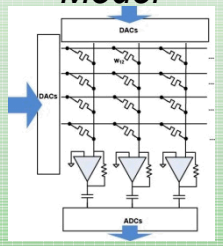


Architecture

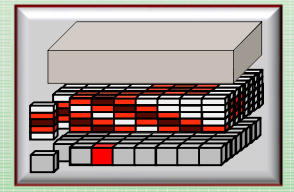
Build



Model

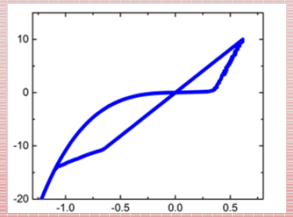


Design

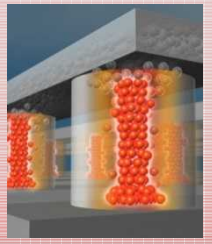


Learning Hardware

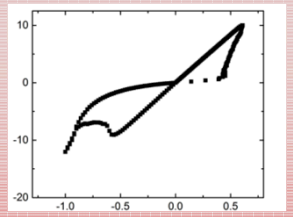
Model



Build



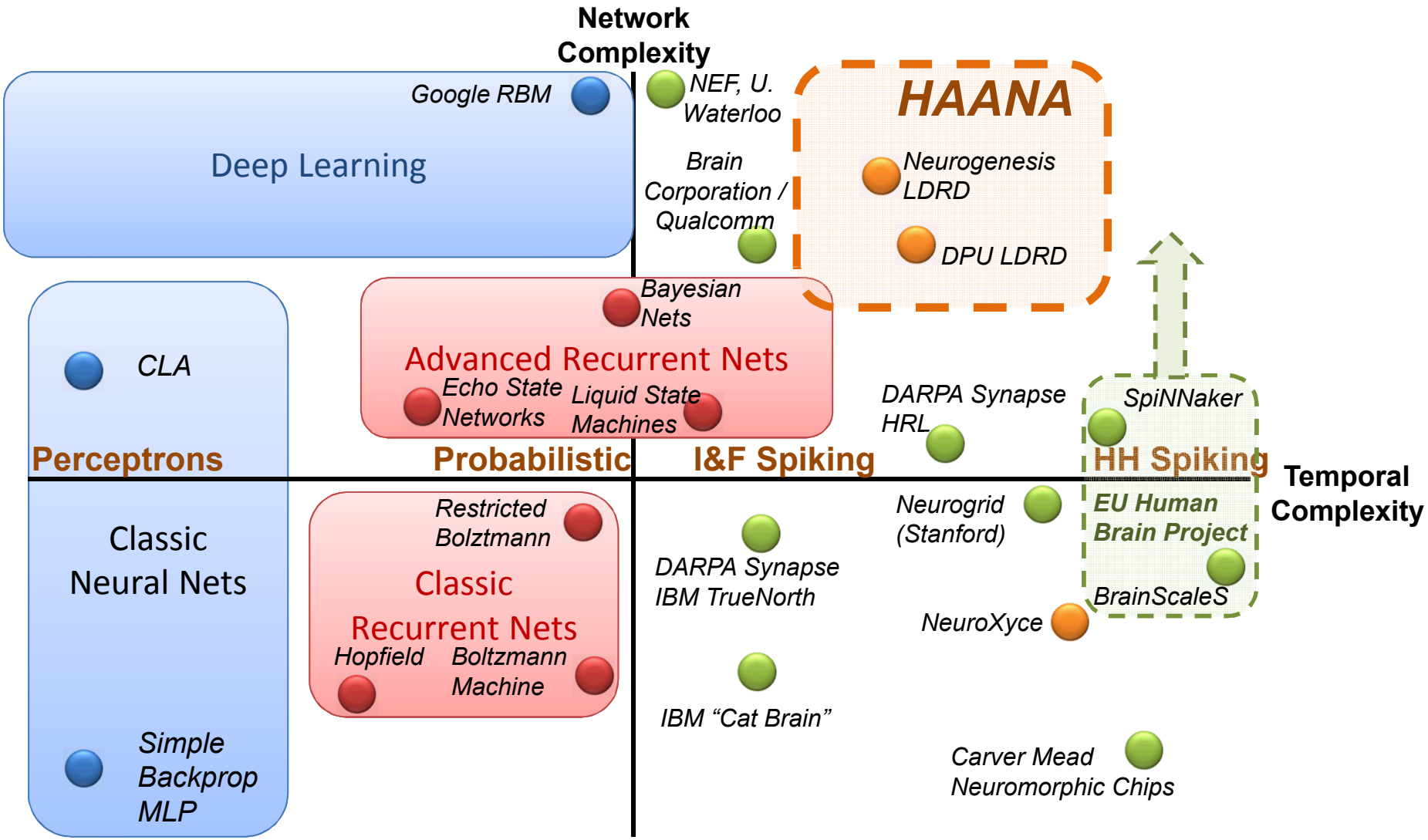
Measure



FY15

Updated competitive landscape of neuro-inspired computing efforts

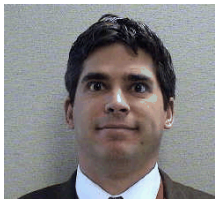
Sandia National Laboratories



Core Team

Sandia National Laboratories

Conrad James
1714
PI

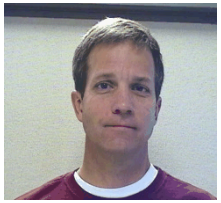


Kevin Dixon
5621
PM

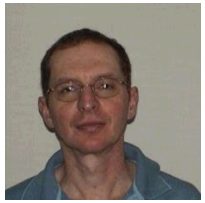
John Naegle
9336
Architecture Lead

Brad Aimone
1462
Algorithms Lead

Matt Marinella
1748
Hardware Lead



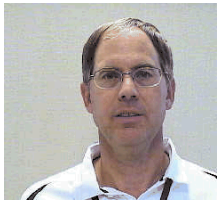
Steve Plimpton
1444
Modeling Lead



Alec Talin
8656
Device Design

Sandra Faust
5563
Dynamical Systems

JD Doak
5635
Cybersecurity



Tim Draelos
5563
Deep Learning

Nadine Miner
6114
Neural Networks

Rich Schiek
1355
Device Modeling

LPS
The Laboratory for Physical Sciences

Lewis Rhodes Labs

UNIVERSITY of DAYTON

AIR FORCE RESEARCH LABORATORY

hp

Full Team

Sandia National Laboratories

Algorithms: Brad Aimone, Ojas Parekh, Nadine Miner, Sandra Faust, Steve Verzi, Fred Rothganger, Frances Chance, Tu-Thach Quach, Chris Lamb
Architecture: John Naegle, Alex Hsia, Eric Debenedictis, Craig Vineyard, John Donaldson
Hardware: Matt Marinella, Tom Beechem, John Ihlefeld, Alec Talin, Paul Kotula, Jim Stevens, Stephen Howell, David Hughart, Patrick Mickel, Andy Armstrong, David Henry, Gaadi Haase, Steve Wolfley
Modeling: Steve Plimpton, Richard Schiek, Christy Warrender, Robert Bondi, Fred Rothganger
Application areas: Tim Draelos, Justin Doak, Jonathan Cox

<u>Strategic partnerships:</u>
Algorithms: Mark McLean – Laboratory for Physical Sciences, U. MD
Architecture: David Follett – Lewis Rhodes Labs
Hardware: Tarek Taha – U. of Dayton, R. Stanley Williams (HP)

SNL Facilities – Microelectronics and High Performance Computing

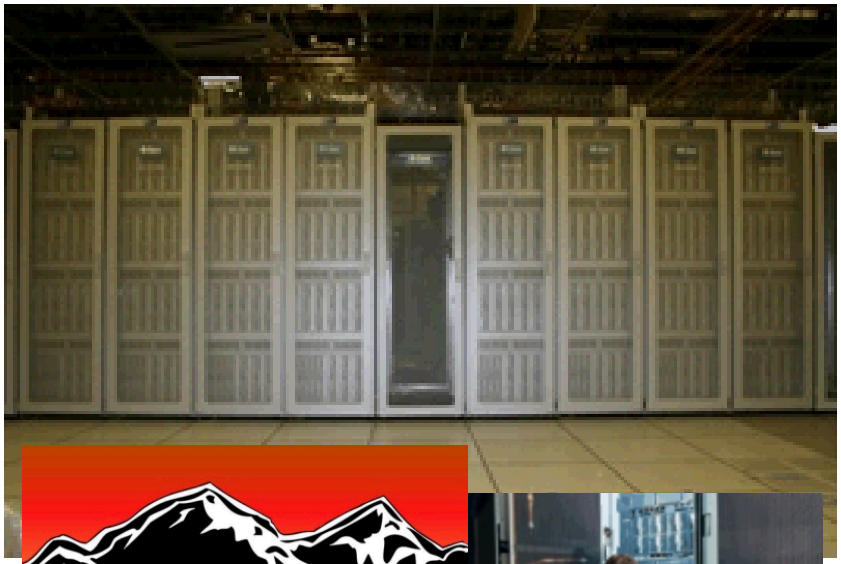
Sandia National Laboratories

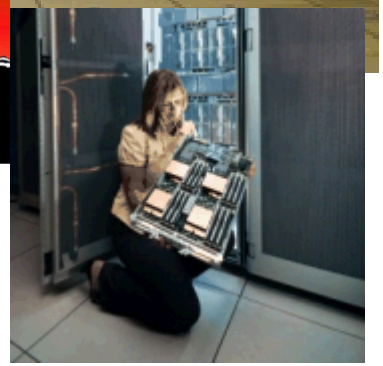
Microsystems and Engineering Science Applications (MESA)

- Trusted ASIC design foundry (350,180,130, 90nm)

Red Sky Supercomputer

- 264 Tflop Linux cluster with Infiniband, 2.93 GHz Intel Nehalum processors, 2846 nodes→22584 cores





Outline

- HAANA – hypothesis and objective
  - Data-driven computing
  - Neural-inspired computing
- Landscape and differentiation
- HAANA project structure
  - **Algorithms Core**
  - Architecture Core
  - Learning Hardware Core
- Conclusions

HAANA project structure

Algorithms

Theory

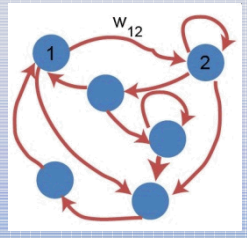
$$R(\alpha) = \int \frac{1}{2} |y - f(\vec{x}, \alpha)| dP(\vec{x}, y)$$

$$w^T x + b = \{-1, 0, 1\}$$

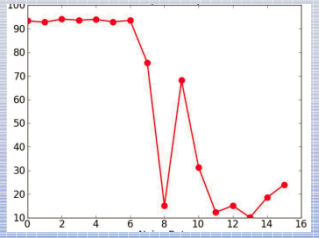
$$w_i^n := w_i^{n-1} - \eta \frac{\partial E_n(w)}{\partial w_i}$$

$$w_i^n := w_i^{n-1} - \eta \left(y(x_n, w) - t_n \right) \frac{\partial y(x_n, w)}{\partial w_i}$$

Design



Model



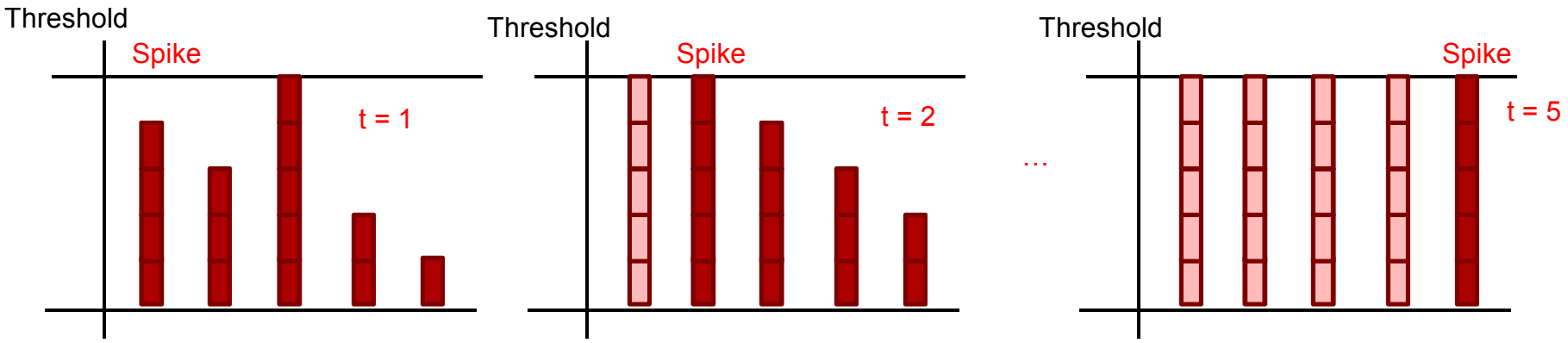
FY15

Architecture — Build — Model — Design

Learning Hardware — Model — Build — Measure

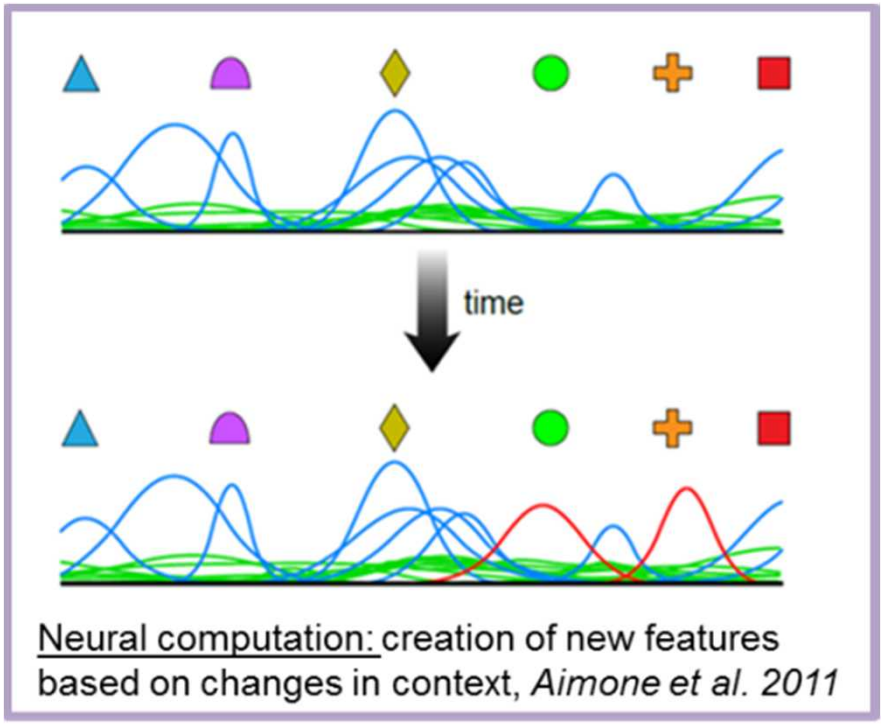
Algorithms Core Objective - Theory

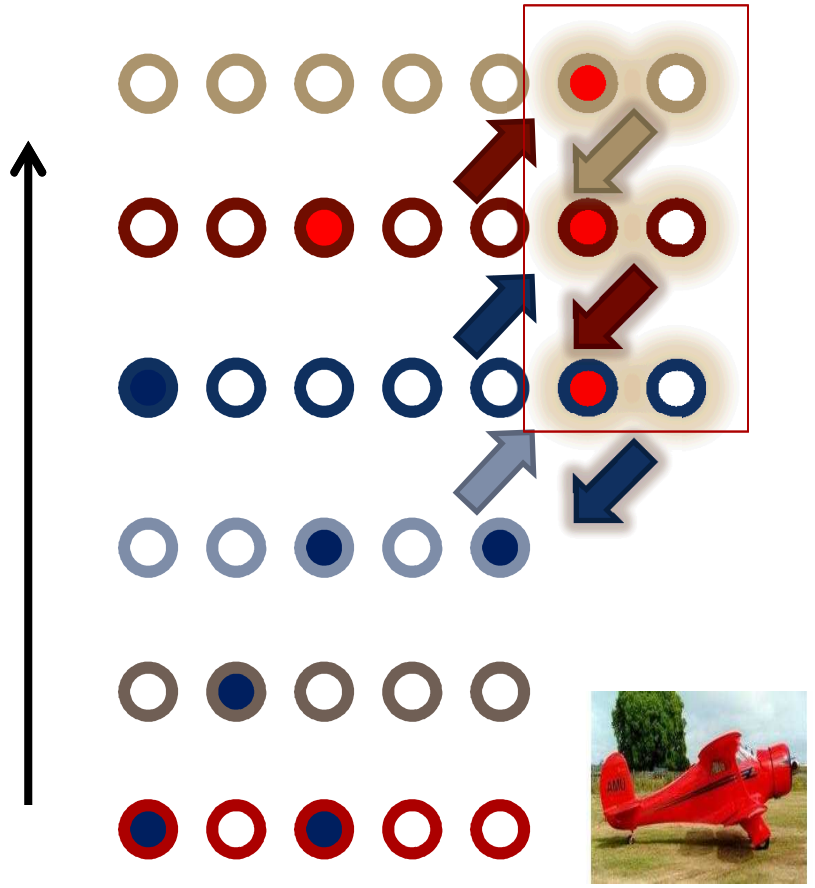
1. Devise machine learning algorithms with functional neural concepts
   - SpikeSort: trade space for computational complexity (time domain);



Algorithms Core Objective - Design

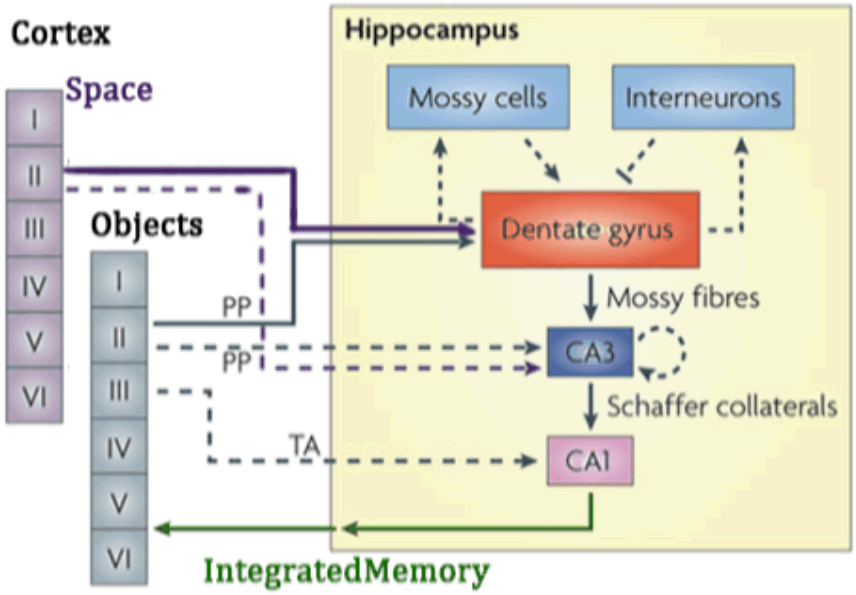
2. Incorporate neuroscience concepts (e.g. structural plasticity) to provide machine learning methods with novel capabilities (e.g. evolving feature sets)

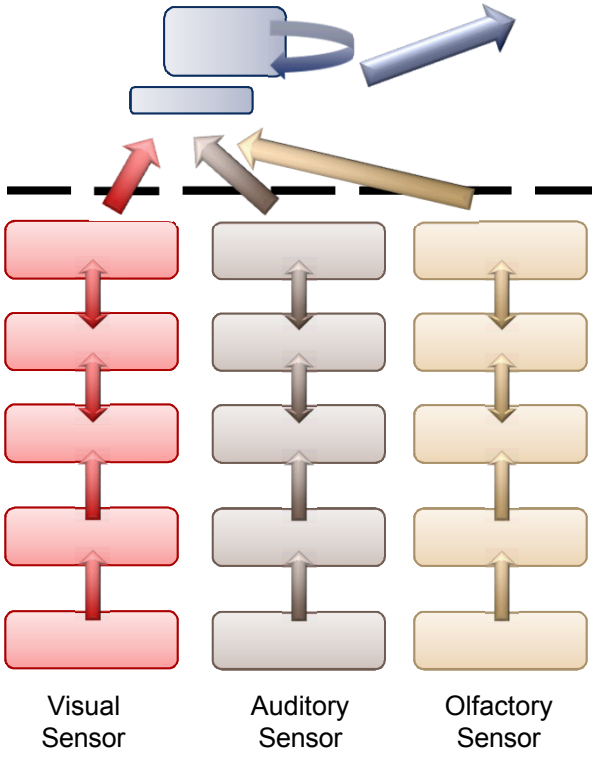




Algorithms Core Objective - Model

3. Extract novel algorithms from neural circuit models
   - multimodal integration and historical context referencing based on hippocampal circuit





Deng, Aimone, and Gage, 2010

Outline

- HAANA – hypothesis and objective
  - Data-driven computing
  - Neural-inspired computing
- Landscape and differentiation
- HAANA project structure
  - Algorithms Core
  - **Architecture Core**
  - Learning Hardware Core
- Conclusions

HAANA project structure

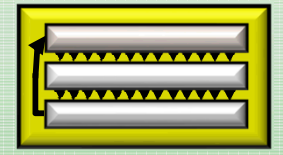
Sandia National Laboratories

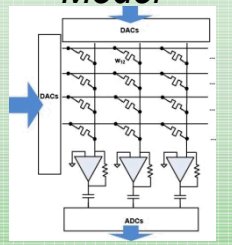
Theory | Design | Model

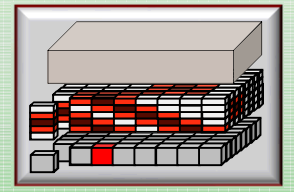
Algorithms

Architecture

Build | Model | Design



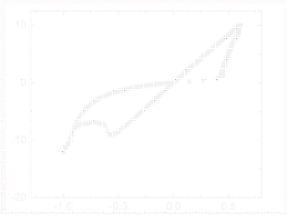




FY15

Model | Build | Measure

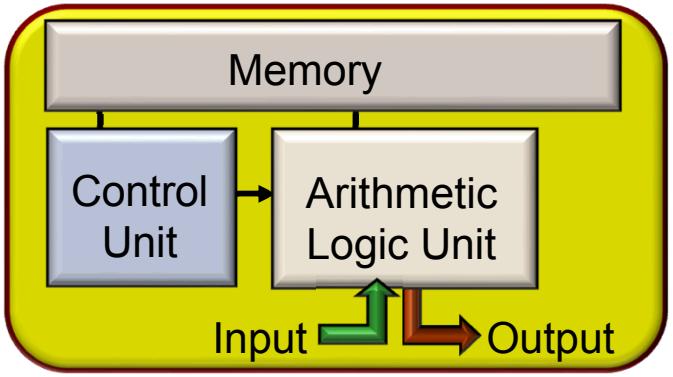
Learning Hardware



Architecture Core Objective - Build

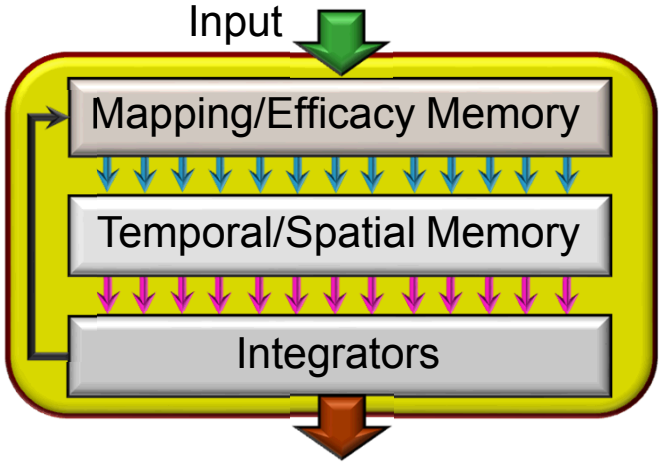
1. Demonstrate the extreme scaling potential of neural-inspired architectures
   - leverage previously developed emulator for an FPGA implementation

Conventional CPU:



- Complex processor core
- Simple memory

Temporal DPU (tDPU):

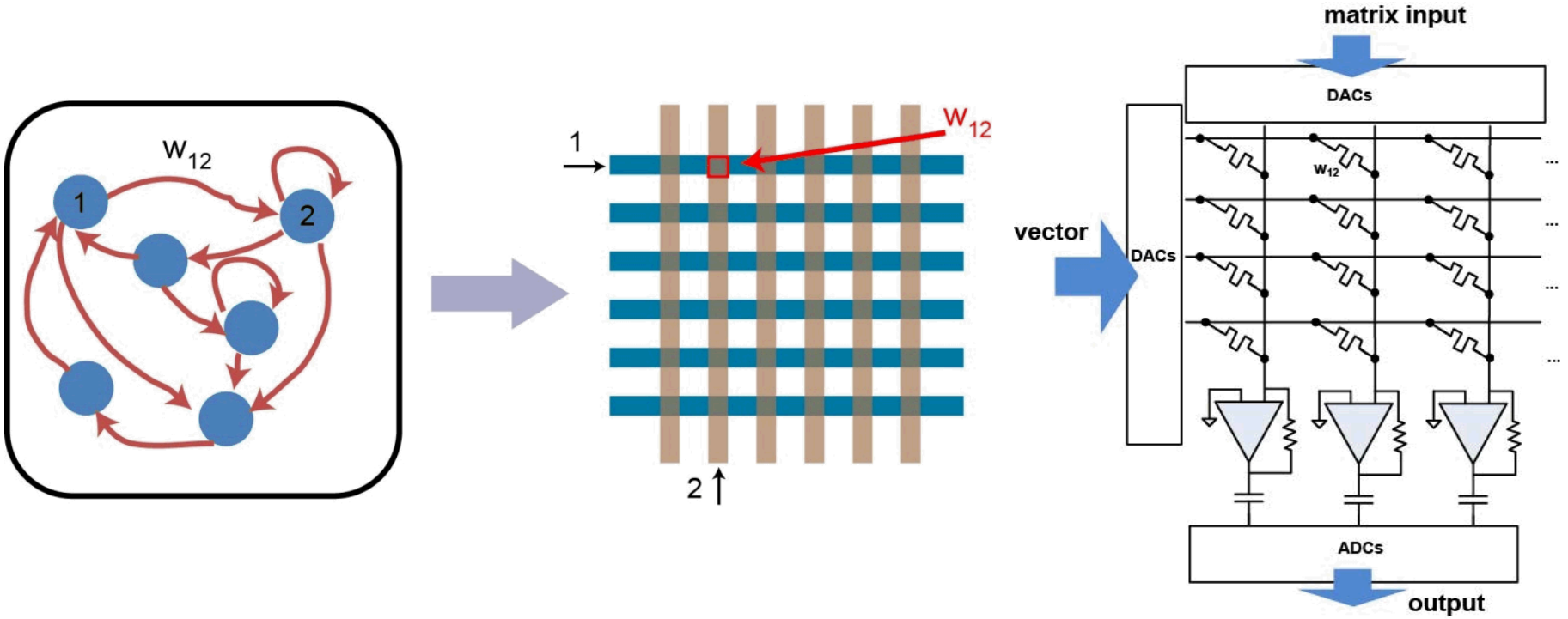


- Simple processor core
- Complex memory

Lewis Rhodes Labs

Architecture Core Objective - Model

2. Co-design novel neural-inspired algorithms and architectures that enable solutions for previously intractable problems
   - matrix multiply and accumulate operations for weight updates, correlation calculations, etc.



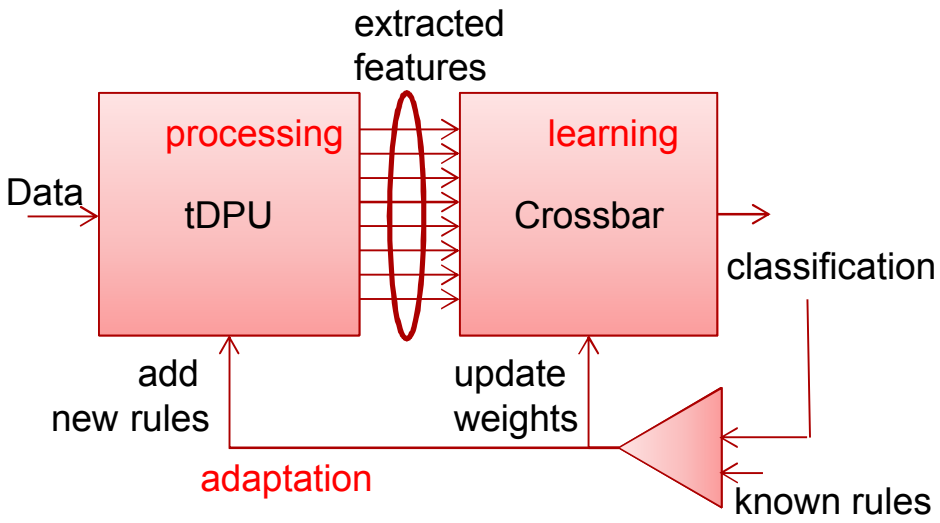
Architecture Core Objective - Design

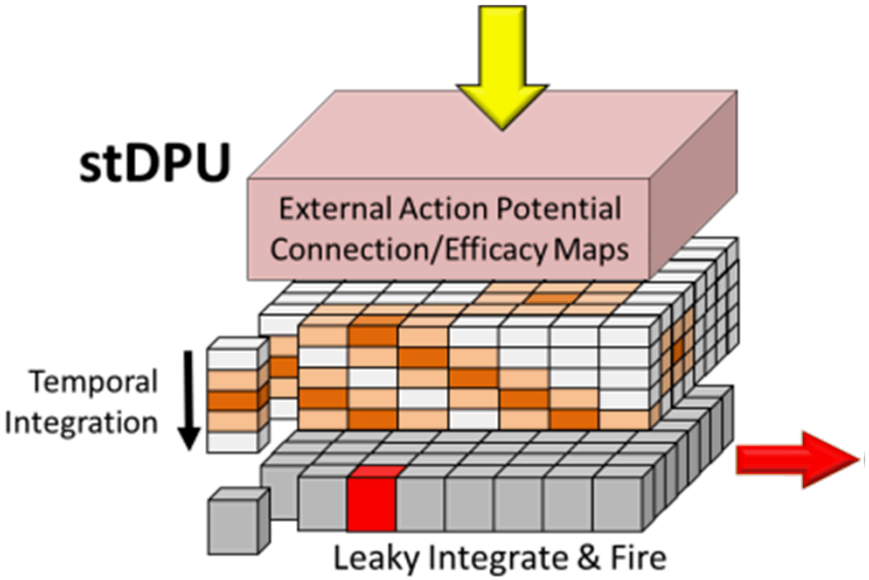
Sandia National Laboratories

3. Develop and demonstrate integrated processing and learning architectures;

Lewis Rhodes Labs

- feature extraction & learning; threat features for weight training; develop stDPU for multi-dimensional data processing





Outline

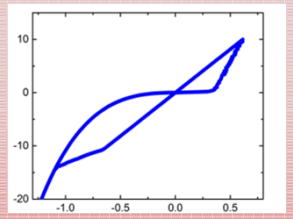
- HAANA – hypothesis and objective
  - Data-driven computing
  - Neural-inspired computing
- Landscape and differentiation
- HAANA project structure
  - Algorithms Core
  - Architecture Core
  - **Learning Hardware Core**
- Conclusions

HAANA project structure

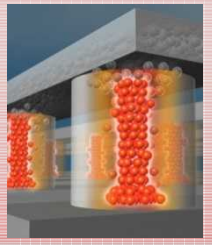
Sandia National Laboratories

Learning Hardware

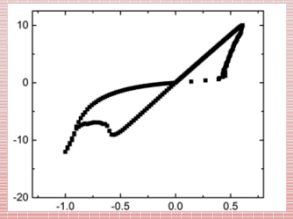
Model



Build



Measure



FY15

Resistive switching device architecture for algorithm performance improvement

Sandia National Laboratories

- Simulations demonstrate significant power savings with a resistive switching-based HW accelerator
- **16x reduction in power, 6x improvement in performance/area over SRAM ASIC**

	Example 1: 25,600 neurons 100,000 iterations/s				
Configuration	# of chips	Chip area (mm^2)	% active	Power (W)	**Power eff. over Xeon**
Memristor Analog (config 4)	1	5.9	38.6%	0.07	**234,859**
Memristor Digital (config 5)	1	18.2	89.6%	0.62	**16,968**
SRAM (config 6)	1	29.1	89.6%	1.13	**8,215**
NVIDIA M2070	12	529.0	99.2%	2700.00	**6**
Intel Xeon X5650	179	240.0	99.9%	17005.00	**1**

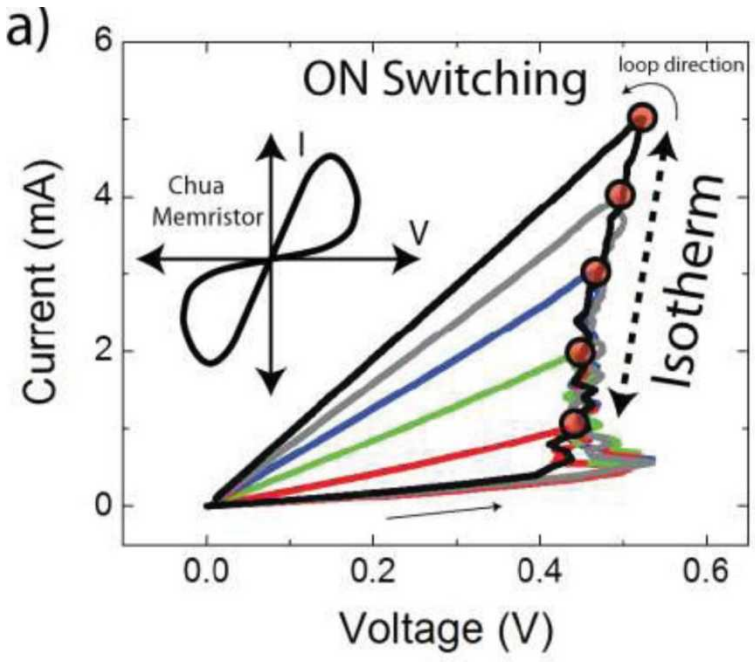
T. Taha, R. Hasan, C. Yakopic, M. McLean, in Proc. IEEE Intl. Joint Conf. on Neural Networks, 2013.

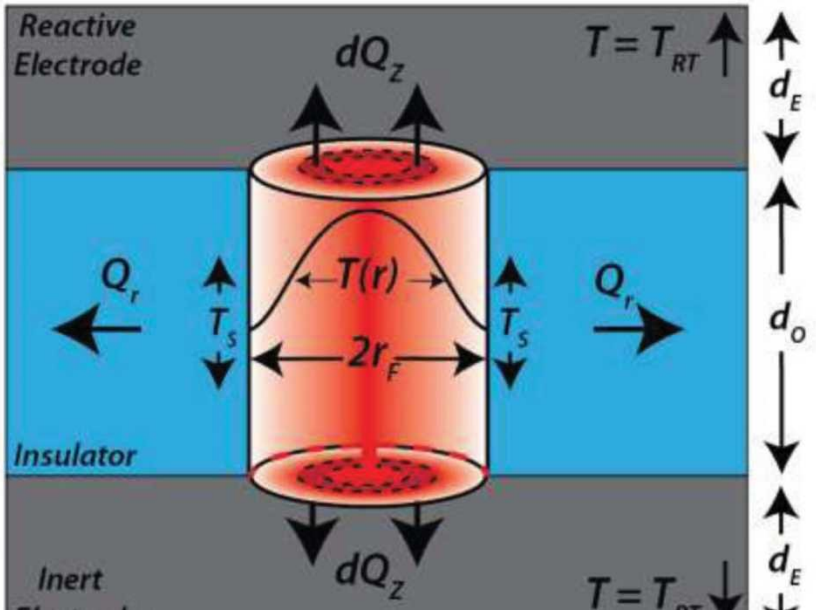
LPS The Laboratory for Physical Sciences

UNIVERSITY of DAYTON

Learning Hardware Core Objective - Model

1. Capture the impact of thermal and e-field driven effects on device switching, resiliency, and precision





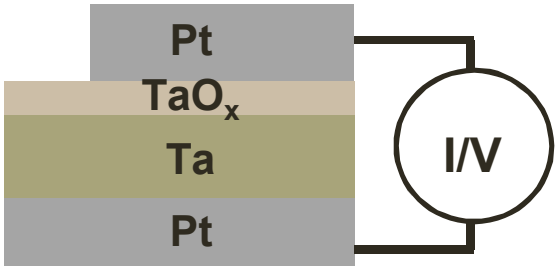
$$P = \frac{\Delta T / R}{\frac{d_E \sigma}{2 k_E d_O} - \frac{r_F^2}{8 L_{WF} T_{crit} d_O^2}}$$

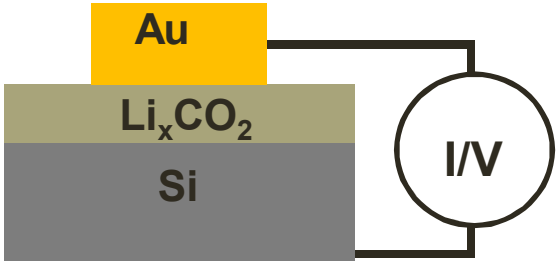
$$T(r) = T_{RT} + \frac{\sigma V^2 d_E}{2 k_E d_O}\left[1 + \frac{k_E}{k_F}\frac{r_F^2 - 2r^2}{4\, d_E d_O}\right]$$

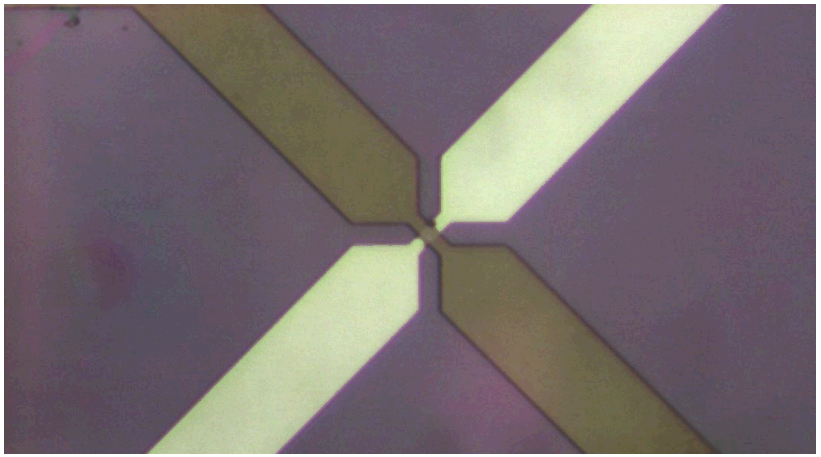
Mickel et al, Adv Mater 26 4486 2014

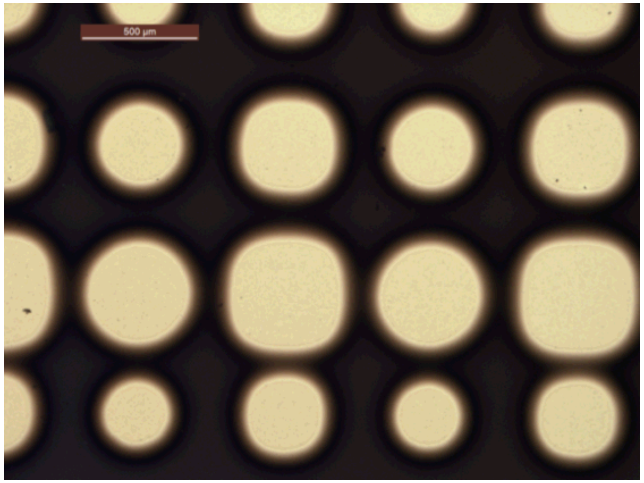
Learning Hardware Core Objective - Build

2. Leverage SNL's resources to design, fabricate, and characterize resistive switching devices to assemble learning hardware







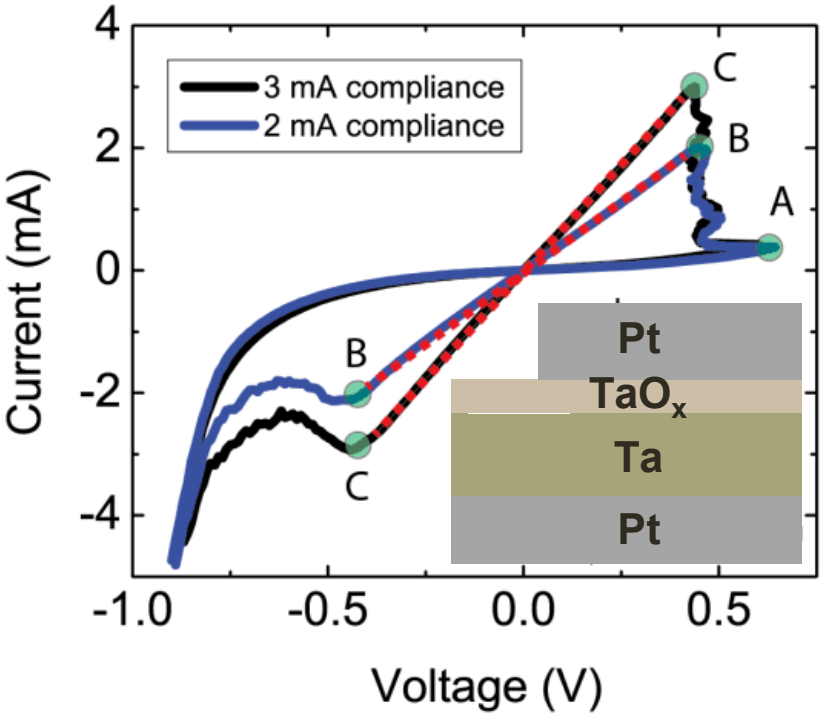


Mickel et al.

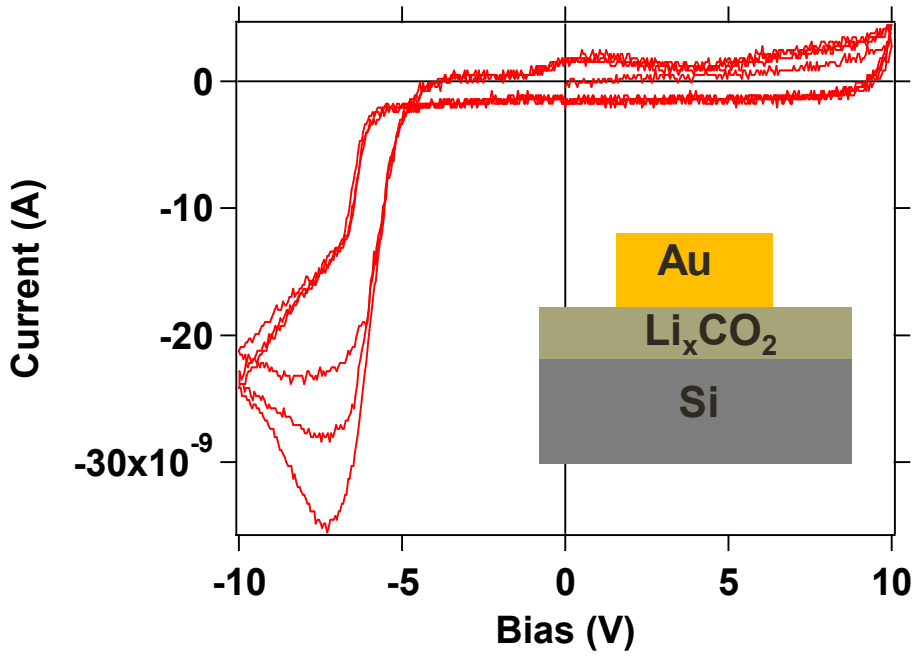
A. Talin et al.

Learning Hardware Core Objective - Measure

3. Leverage SNL's resources to design, fabricate, and characterize resistive switching devices to assemble learning hardware



Mickel et al., Adv Mat 2014, in press

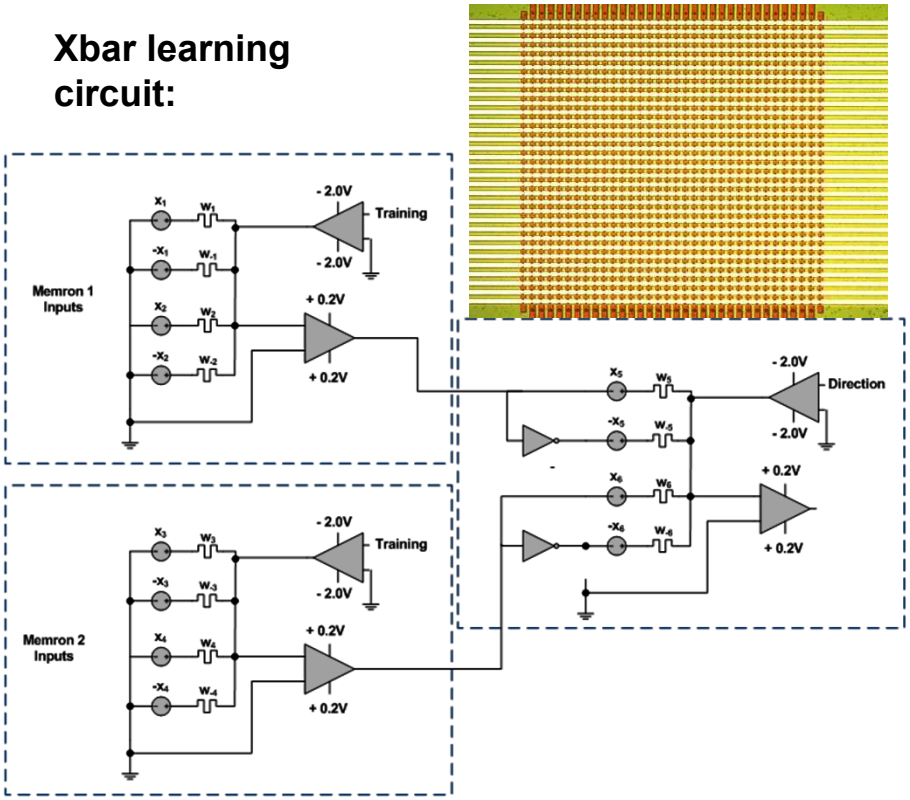


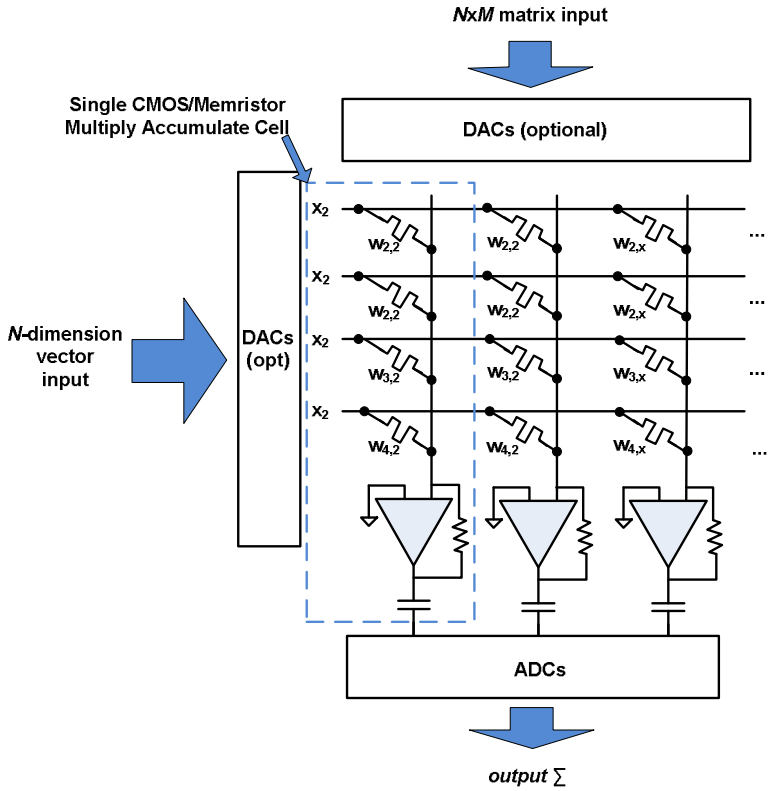
A. Talin et al.

Integration - learning algorithm accelerator

- Learning is computationally intensive
- Crossbar architectures are well suited to accelerate multiply-and-accumulate operations and store weights - demonstrate with candidate algorithms

Xbar learning circuit:





Outline

- HAANA – hypothesis and objective
  - Data-driven computing
  - Neural-inspired computing
- Landscape and differentiation
- HAANA project structure
  - Algorithms Core
  - Architecture Core
  - Learning Hardware Core
- **Conclusions**

Conclusions

- Focus on novel algorithm development for data processing, feature extraction, data fusion, and dynamic context integration
- Leverage existing architectures and integrate capabilities to build a threat detection platform capable of online learning
- Model, design, and build hardware that maps onto the algorithm and architecture structures that have been developed, with a focus on accelerating data processing and learning computations, and on improving SWaP and time constraints