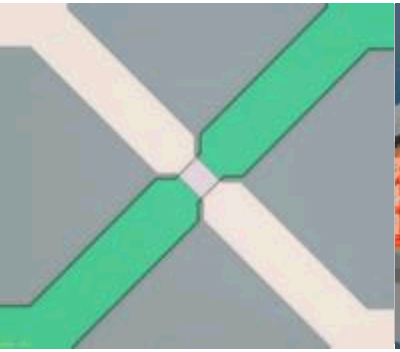
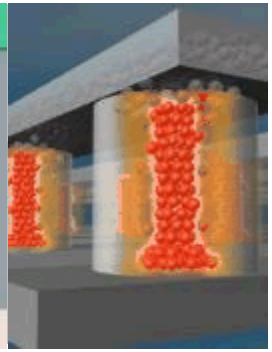
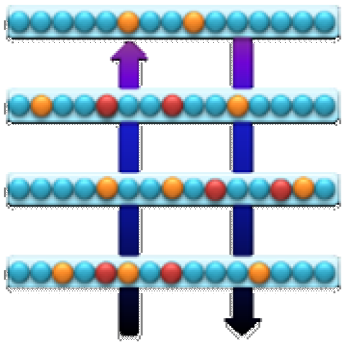
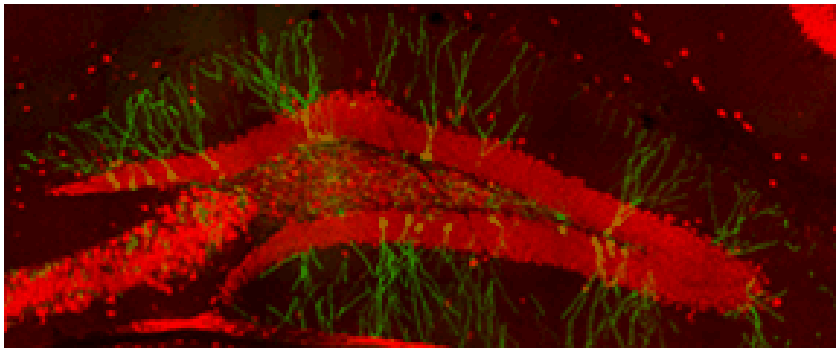
SAND2017-8847PE

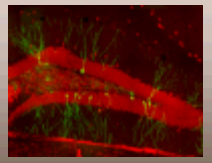
Hardware Acceleration of Adaptive Neural Algorithms

Brad Aimone (jbaimon@sandia.gov)
Deputy PI, HAANA Grand Challenge
Sandia National Laboratories
August 23, 2017

Sandia National Laboratories is a multiprogram laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Neuroscience & neural-inspired computing at SNL impacts the mission broadly

Neuroscience Theory



IARPA MICrONS Government Team for Test & Evaluation of Neural Models and Machine Learning

- Neuro-informatics
- Modeling and Simulation
- Neural Data Analytics
- Computational Neuroscience

Neural Computing Capabilities

HAANA Grand Challenge – Flagship LDRD across computing, materials, and cyber security centers

- Formal Neural Computing Theory
- Neural Inspired Architectures
- Neural Machine Learning Algorithms
- UQ / SA of Neural Algorithms and Neural Architectures

Neural-enabling Hardware

MESA Fabrication Facility provides materials and design research capabilities for next generation neural systems

- Micro-sensors
- Non-von Neumann architectures
- Memory technology
- Neuromorphic Computing Lab

Mission Impacts

Enabling Advanced Simulation and Computing

- Neural-inspired communication paradigms
- Adaptive memory management
- Numerical computing with neurons

ASC

Neuroscience Contributions

- Contribute to the science of understanding the brain
- BRAIN Initiative
- MICrONS

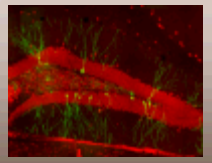
Deployable National Security Applications

- Cyber Defenses
- Embedded Pattern Recognition Systems
- Smart Sensor Technologies

U.S. DEPARTMENT OF ENERGY

Neuroscience & neural-inspired computing at SNL impacts the mission broadly

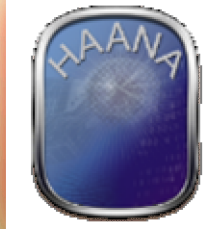
Neuroscience Theory



IARPA MICrONS Government Team for Test & Evaluation of Neural Models and Machine Learning

- Neuro-informatics
- Modeling and Simulation
- Neural Data Analytics
- Computational Neuroscience

Neural Computing Capabilities



HAANA Grand Challenge – Flagship LDRD across computing, materials, and cyber security centers

- Formal Neural Computing Theory
- Neural Inspired Architectures
- Neural Machine Learning Algorithms
- UQ / SA of Neural Algorithms and Neural Architectures

Neural-enabling Hardware

MESA Fabrication Facility provides materials and design research capabilities for next generation neural systems

- Micro-sensors
- Non-von Neumann architectures
- Memory technology
- Neuromorphic Computing Lab

Mission Impacts

Enabling Advanced Simulation and Computing

- Neural-inspired communication paradigms
- Adaptive memory management
- Numerical computing with neurons

ASC

Neuroscience Contributions

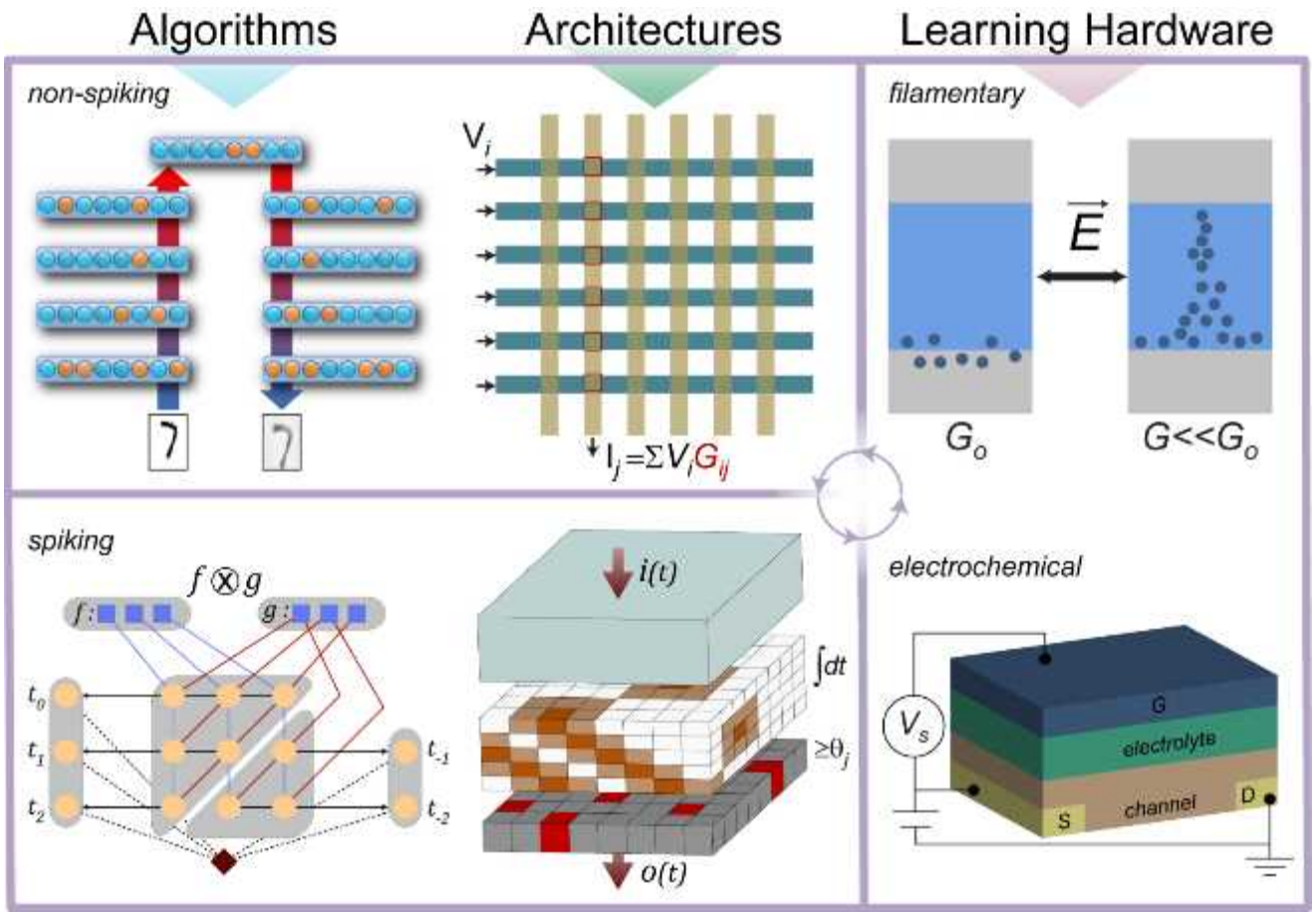
- Contribute to the science of understanding the brain
- BRAIN Initiative
- MICrONS

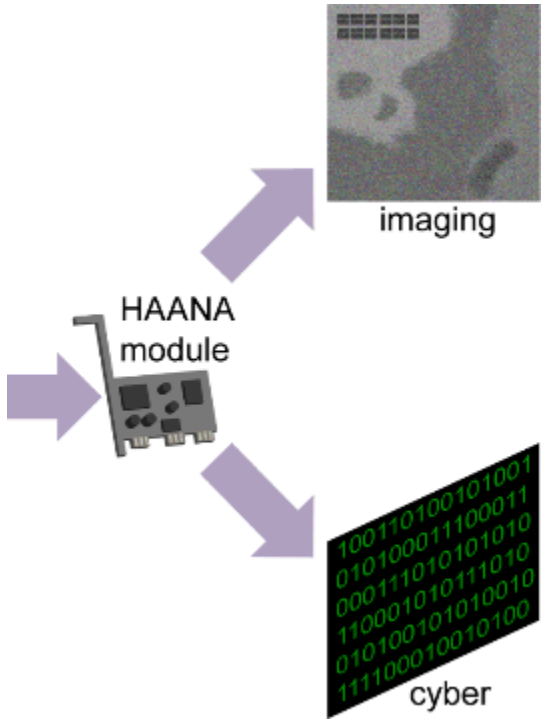
Deployable National Security Applications

- Cyber Defenses
- Embedded Pattern Recognition Systems
- Smart Sensor Technologies

U.S. DEPARTMENT OF ENERGY

Hardware Acceleration of Adaptive Neural Algorithms (HAANA)

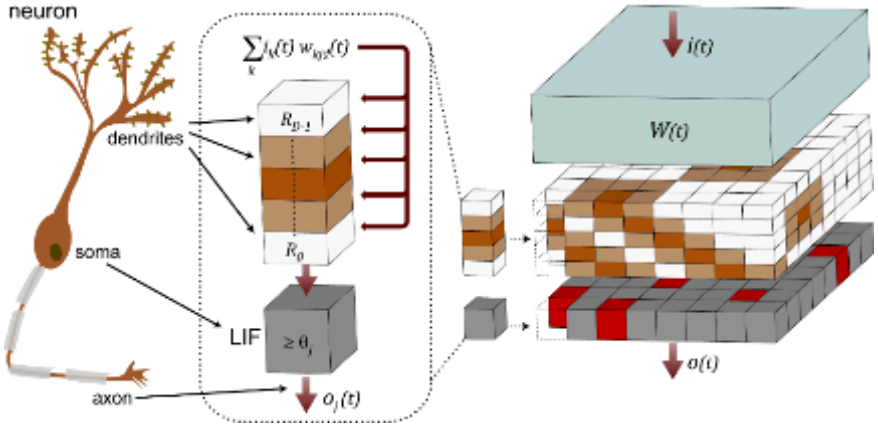




Resistive switching model: Mickel et al, Adv Mater 2014
Adaptive categorization: Vineyard et al., IJCNN 2015
File categorization with DNNs: Cox et al., CAS 2015
Spiking network algorithms: Severa et al., ICRC 2016

Electrochemical transistor: Fuller et al., Adv Mater 2016
Resistive crossbar accelerator: Agarwal et al., IJCNN 2016
Neurogenesis deep learning: Draelos et al, IJCNN 2017
Digital neuromorphic architecture: Smith et al., IJCNN 2017

HAANA neural architectures enable new paradigms for approaching computation

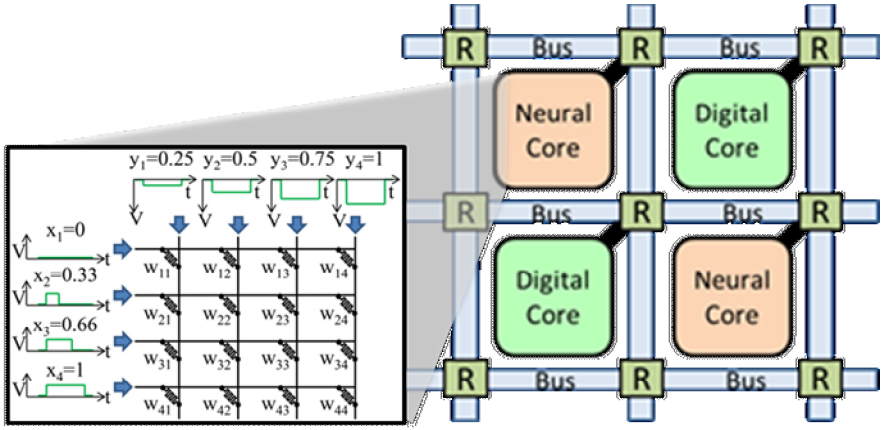


Spiking Temporal Processing Unit

- Digital, FPGA compatible
- Unrestricted connectivity
- Complex temporal representations

Resistive Crossbar Neural Architectur

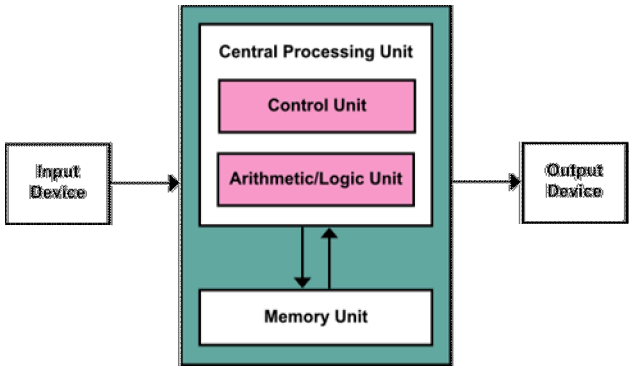
- Analog processing of vector matrix algebra
- Hybrid with conventional digital components



Challenge: What does this neural computing paradigm actually look like?

Conventional and quantum computing benefit from concrete theoretical models…

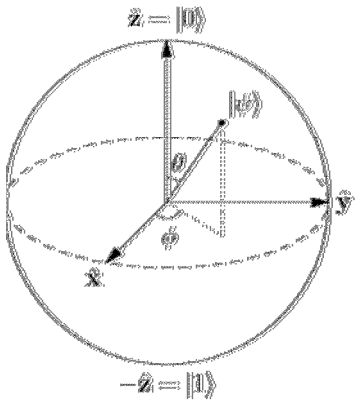
Classic Algorithms



Materials Science & Device Physics

Von Neumann computing

Quantum Algorithms



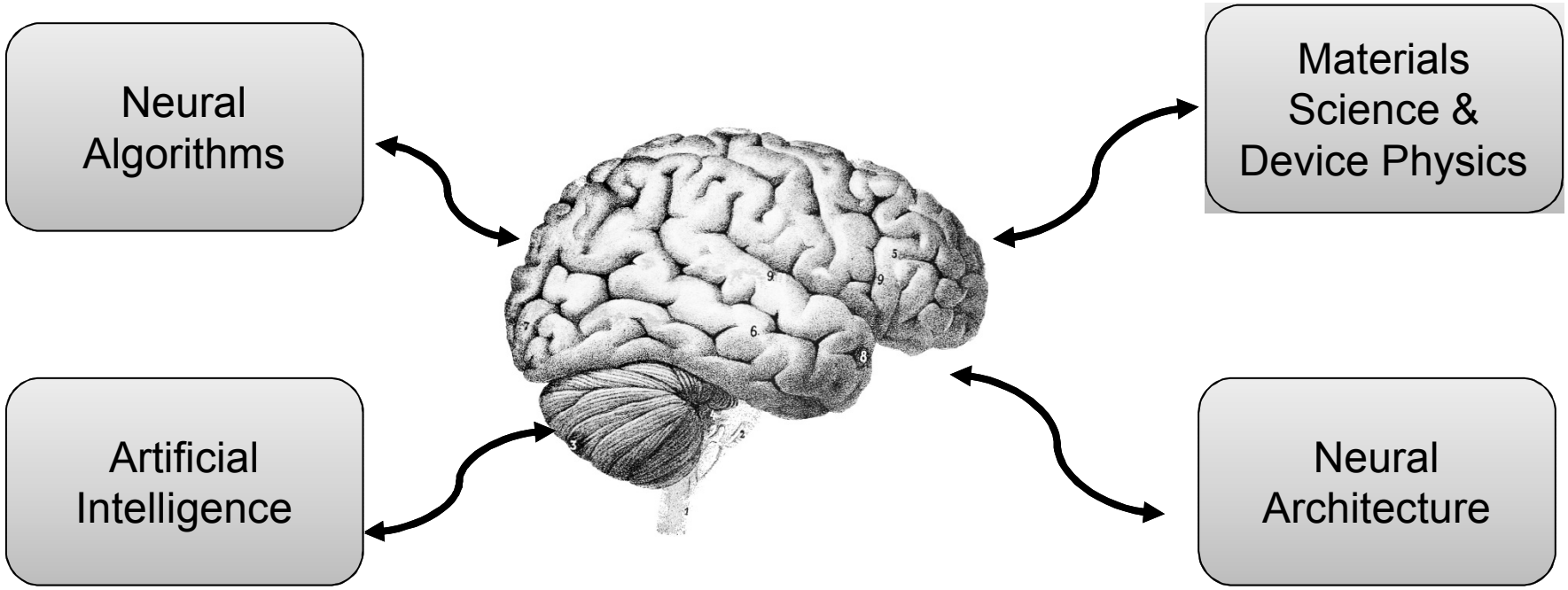
Quantum physics

Computer scientists know what to program

Quantum computing

Engineers and physicists know what to build

…whereas neural computing means something different to everybody



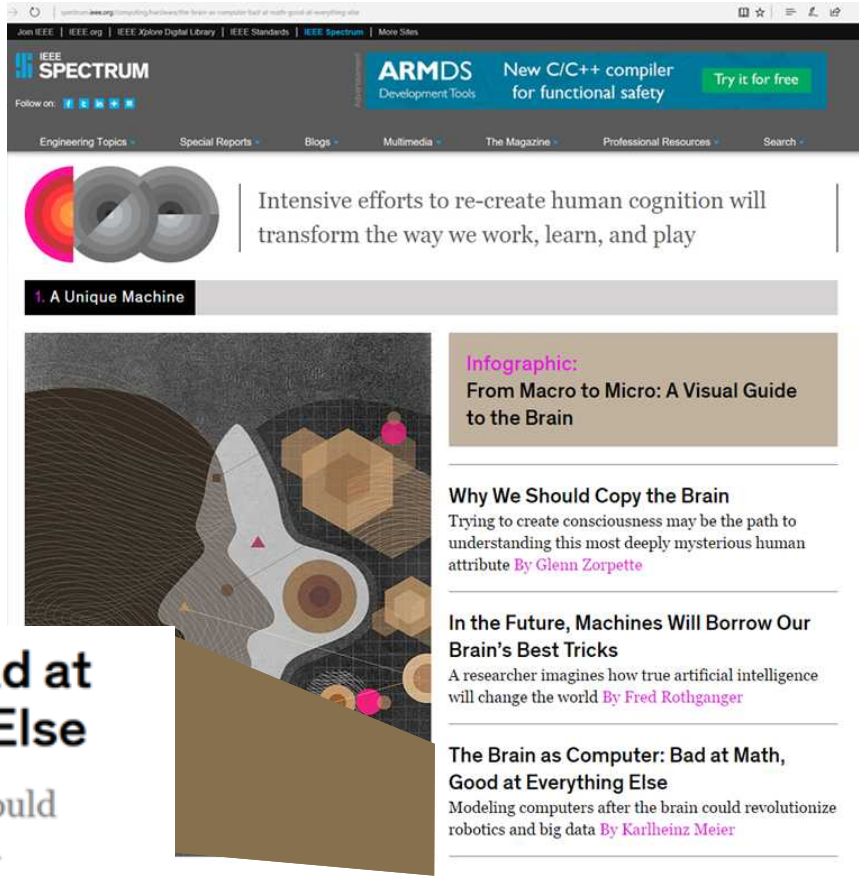
What is the brain as inspiration?

HAANA specifically aimed to make this theoretical framework explicit to bridge algorithms and architectures

Established conventional wisdom: neural-inspired computing is bad at math

Why?

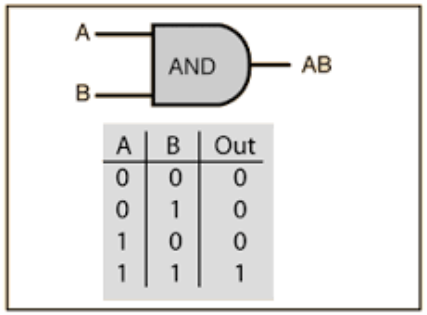
- It is a challenge to separate *brains* (cognitive capability) from *neurons* (low-energy mechanism)
- Belief that neurons are noisy
- Moore's Law – It has always been easier to wait for faster processors than to re-invent numerical computing on specialized parallel architecture



Emerging HAANA Hypothesis: This presumption may not be correct…

Spiking neurons are a more powerful version of classic logic gates

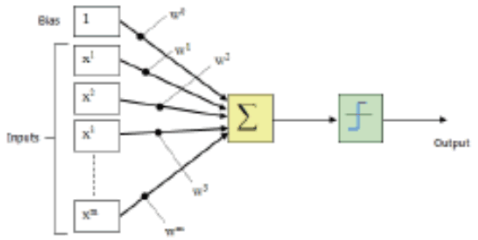
Spiking threshold gates provide high degree of parallelism at very low power



A	B	Out
0	0	0
0	1	0
1	0	0
1	1	1

High fan-in

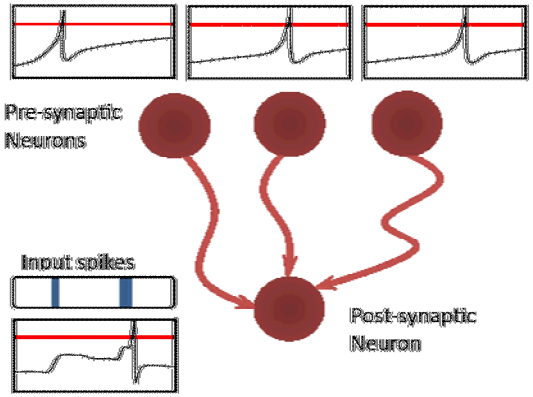
Based on a simple McCullogh-Pitts model:



Outputs a 1 if and only if: $w_0 + \sum_{i>0} w_i x_i \geq 0$.

Compute more powerful logic functions

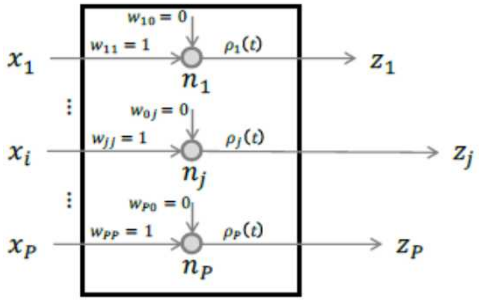
Spiking

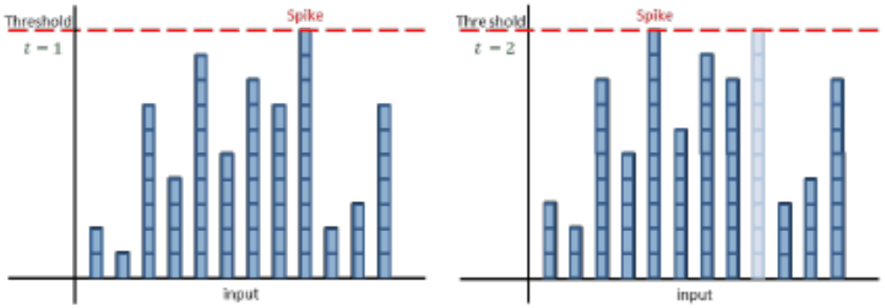


Incorporate time into logic

HAANA has produced a number of spiking numerical algorithms

- Cross-correlation
  - Severa et al., ICRC 2016
- SpikeSort
  - Verzi et al., *submitted*
  - SpikeMin
  - SpikeMax
- SpikeOptimization
  - Verzi et al., IJCNN 2017

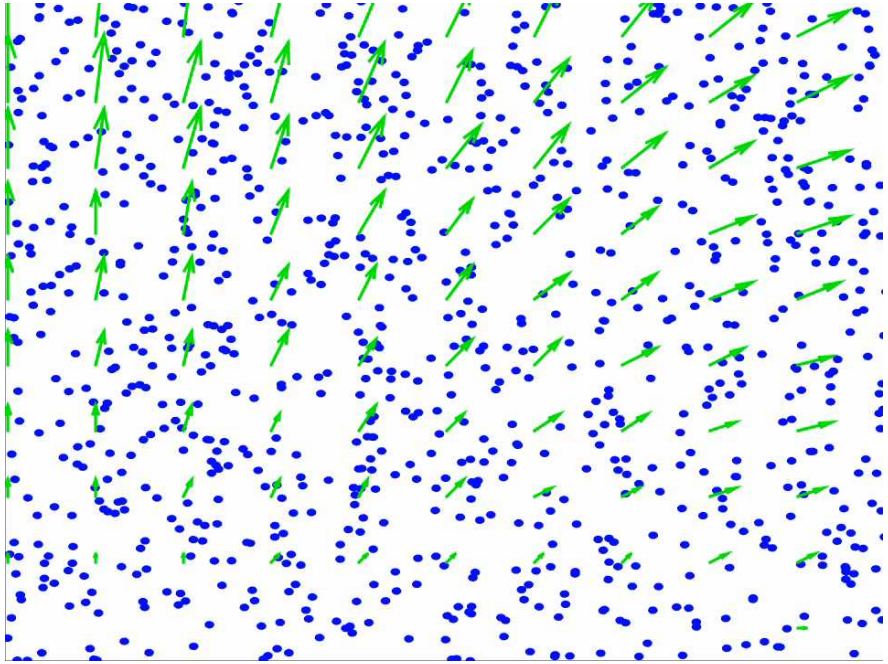




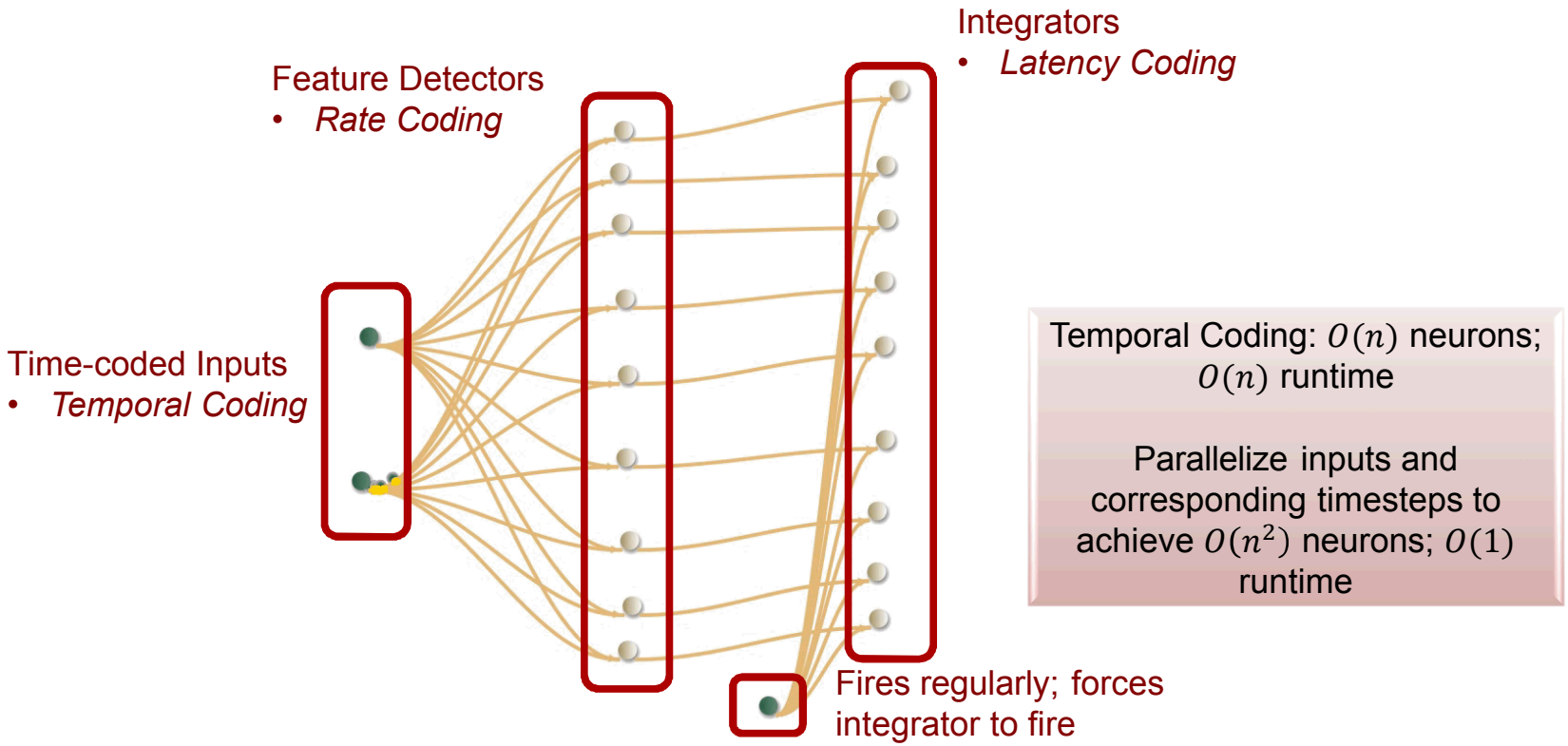
- Sub-cubic (i.e., Strassen) constant depth matrix multiplication
  - Parekh et al., *submitted*

A Velocimetry Application

- A motivating application is the determination of the local velocity in a flow field
- The maximal cross-correlation between two sample images provides a velocity estimate
- SNN algorithms are straightforward; exemplify core concepts
  - Highly parallel
  - Different neural representations
  - Modular, precise connectivity
  - Time/Neuron tradeoff



Time Multiplexed Cross Correlation

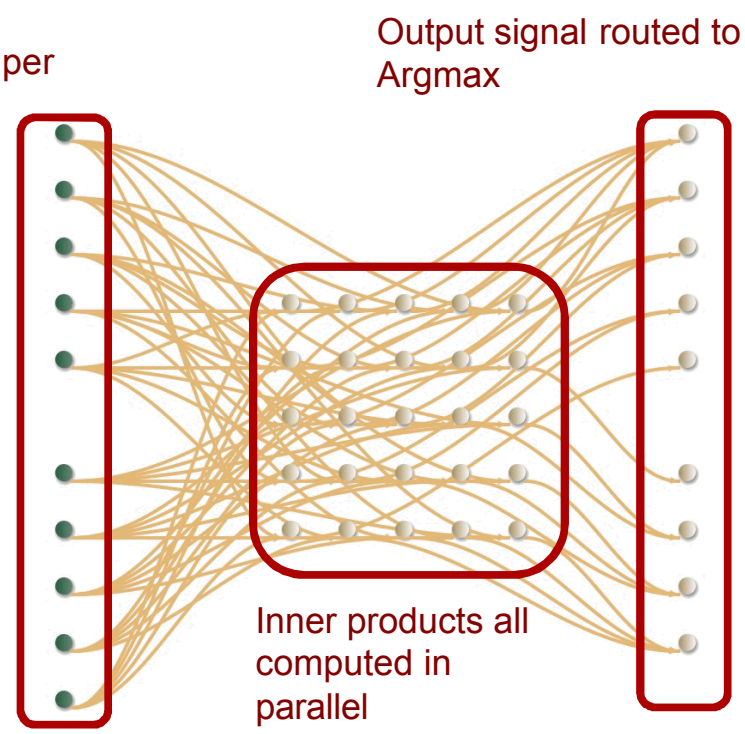


Temporal Coding: $O(n)$ neurons; $O(n)$ runtime

Parallelize inputs and corresponding timesteps to achieve $O(n^2)$ neurons; $O(1)$ runtime

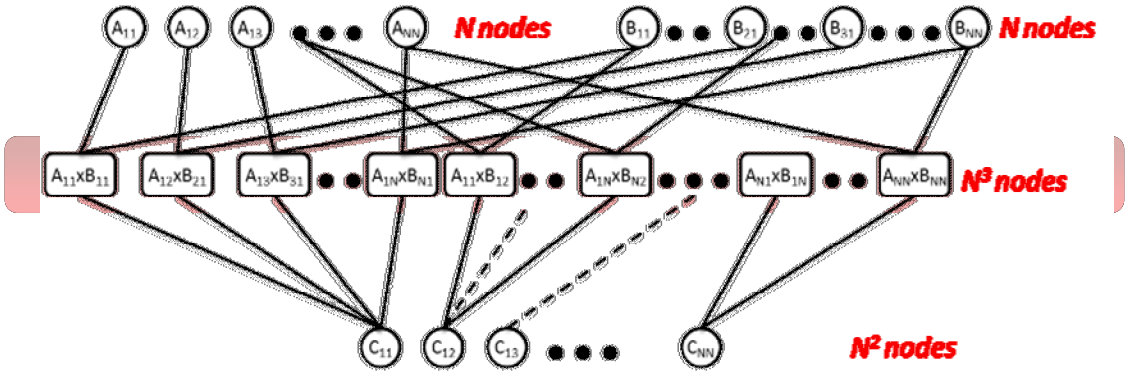
Cross-Correlation Exhibits Time/Neuron Tradeoff

- Exchange Time Cost ↔ Neuron Cost
- Complexity is unchanged
- **Neurons:** $O(n^2) \leftrightarrow O(n)$
- **Time:** $O(1) \leftrightarrow O(n)$

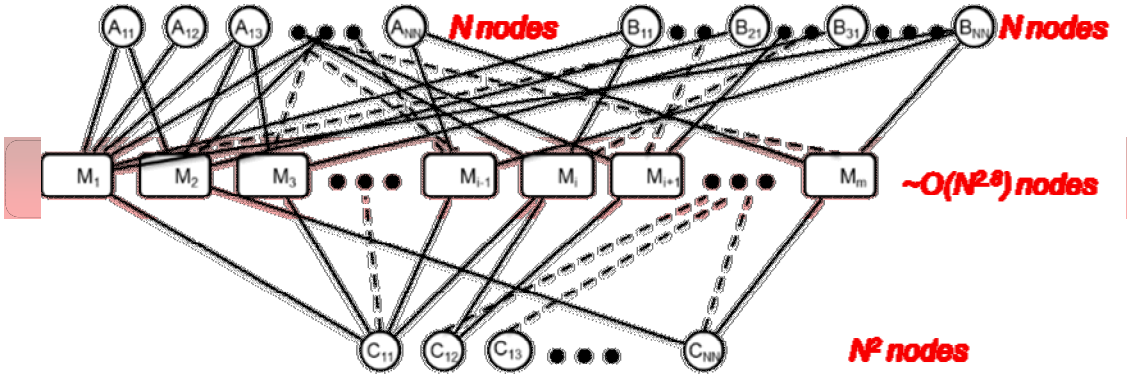


“Neural” network for matrix multiplication

Standard:
8Ms, 4As →
$O(N^3)$



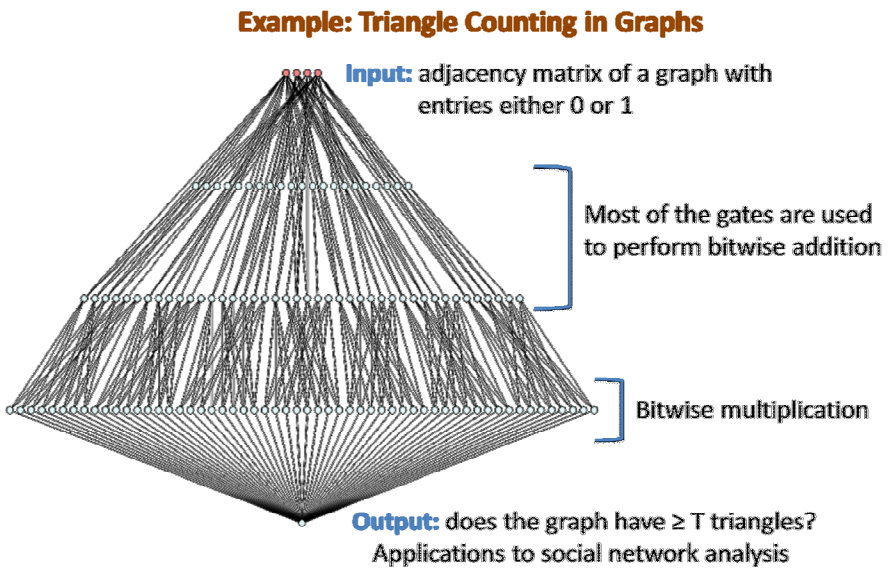
Strassen:
7Ms, 18A/Ss→
$O(N^{2+e})$

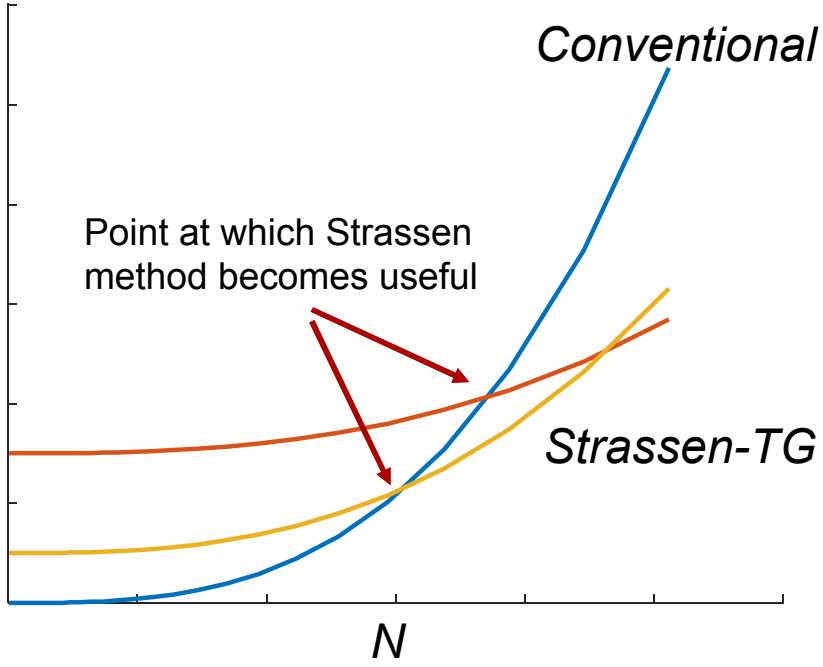


Strassen formulation of matrix multiply enables less than O(N^3) neurons – resulting in less power consumption

Strassen multiplication in neural hardware may show powerful advantages

	Depth	# Gates	Value of ε
Standard	3	$O(N^3)$	–
"Direct" Strassen	d	$O(N^{\omega+\varepsilon})$	1/d
Refined Strassen	d	$O(N^{\omega+\varepsilon})$	$O(1/c^d)$
Non-constant Depth	O(log log N)	$O(N^{\omega})$	–



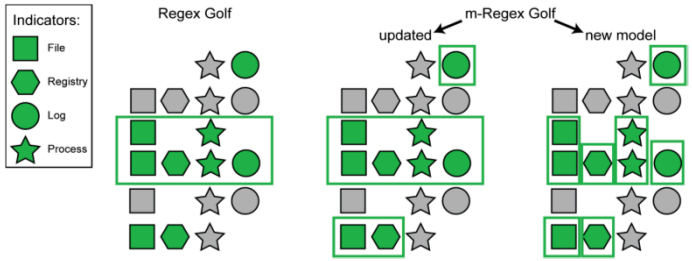


What comes after HAANA?

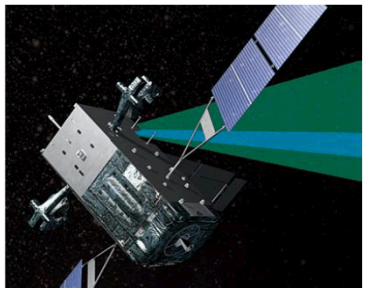
- DOE
  - ASC Beyond Moore computing portfolio
  - ASC Machine learning for scientific computing applications
  - “Big Idea” on Beyond Moore’s Law computing
- DARPA
  - Lifelong Learning in Machines program
- IARPA
  - MICrONS program

HAANA has identified several application areas to impact Sandia's mission

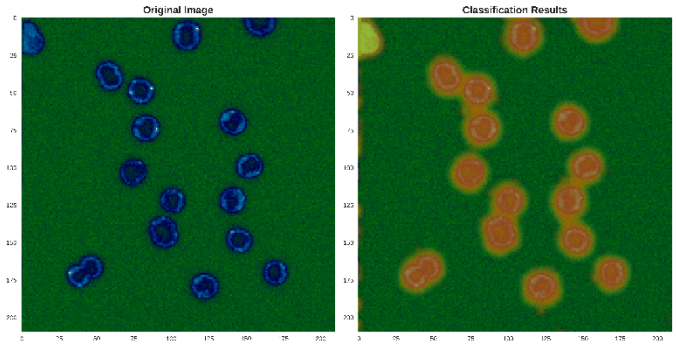
- Cyber
  - "Tracking the Known" streaming network analysis
  - Digital file forensics
  - Host system calls analytics



- Remote Sensing
  - Contour-based image classification
  - Deep learning on neuromorphic architectures



- Bioinformatics
  - Deep learning on hyperspectral data

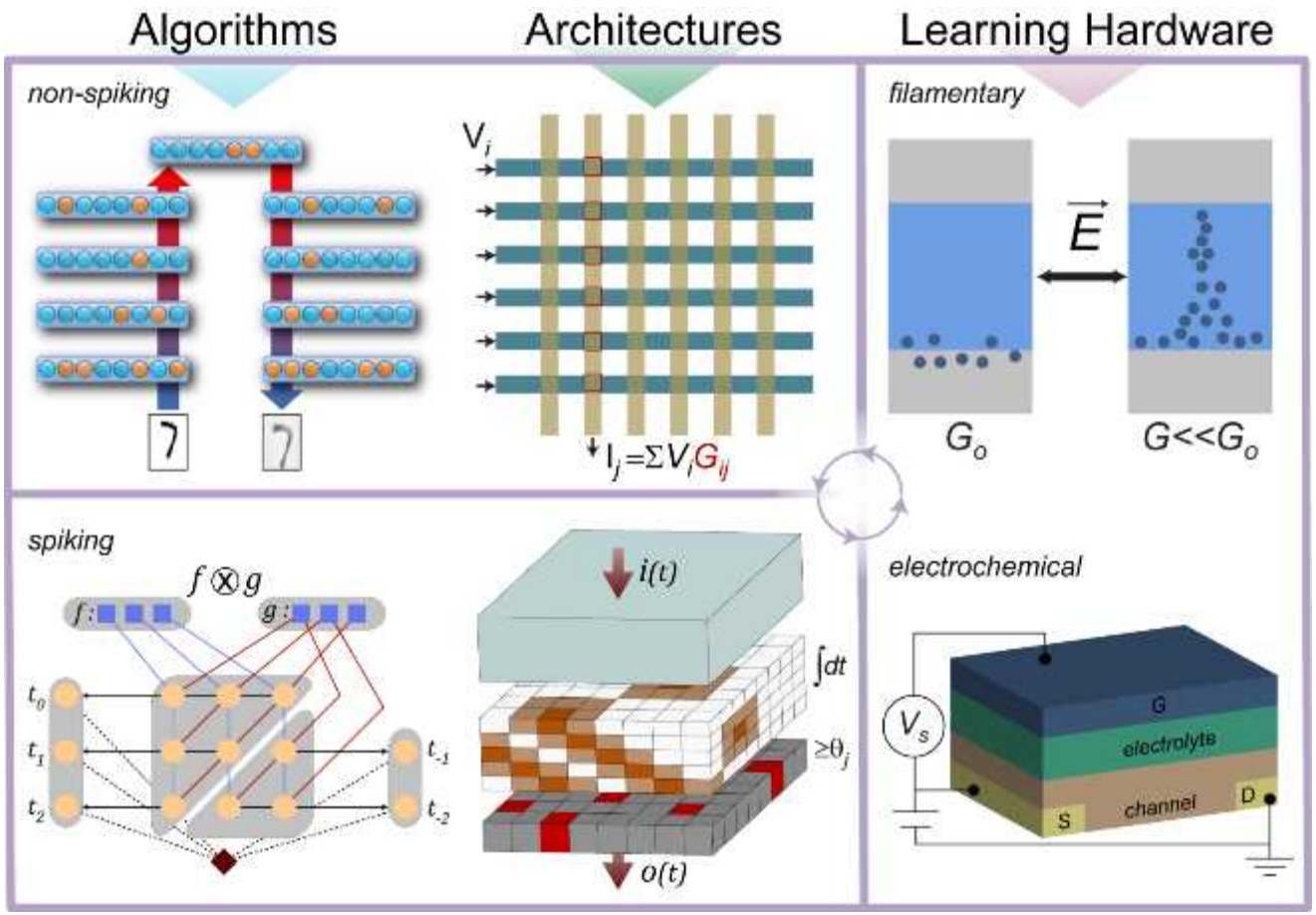


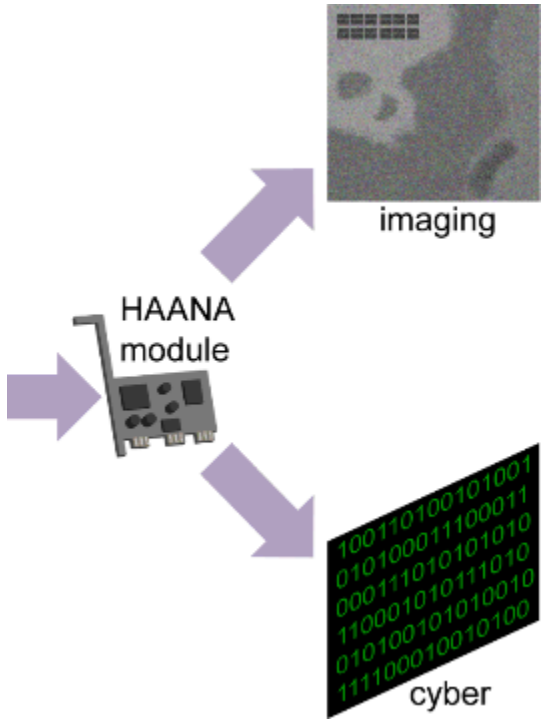
- Scientific computing
  - Numerical computation with spiking neurons
  - Threshold gate linear algebra

Quick highlights from HAANA

- Publications
  - ~30 conference and journal papers published (~10 submitted / in final stages)
  - ~50 talks, conference, and workshop presentations
  - ~8 patents filed
- Growing suite of spiking and adaptive neural algorithms
- Two general neural architectures in development
  - Resistive crossbar array, with emphasis on evaluation
  - Spiking temporal processing unit (STPU), currently in FPGA prototype
- Novel hardware devices for higher precision analog computing
- Application Impact
  - Demonstrated impact of neural acceleration of streaming cyber analytics
  - Identified other areas of cyber security potentially well-suited for neural algorithms (file forensics, system call analysis, volatile memory assessment)
  - Potential role of neural approaches in scientific computing

Hardware Acceleration of Adaptive Neural Algorithms (HAANA)





Resistive switching model: Mickel et al, Adv Mater 2014
Adaptive categorization: Vineyard et al., IJCNN 2015
File categorization with DNNs: Cox et al., CAS 2015
Spiking network algorithms: Severa et al., ICRC 2016

Electrochemical transistor: Fuller et al., Adv Mater 2016
Resistive crossbar accelerator: Agarwal et al., IJCNN 2016
Neurogenesis deep learning: Draelos et al, IJCNN 2017
Digital neuromorphic architecture: Smith et al., IJCNN 2017

Multidisciplinary algorithms team

Computational Neuroscience



Neural Computing Theory



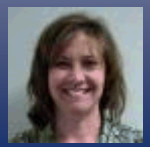


Neural Algorithm Design




Neural Architecture Interfaces



Neural Cyber Application




Neural Image Processing Apps