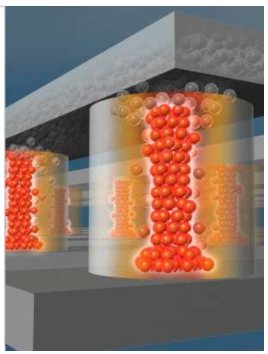
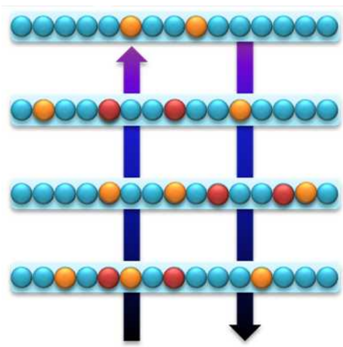
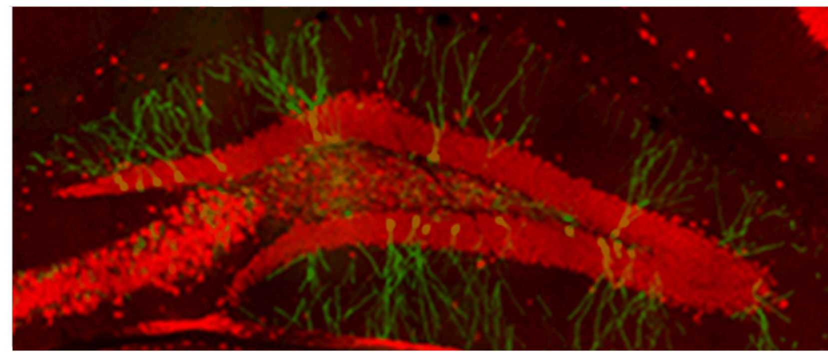
SAND2016-4712PE

Exceptional service in the national interest

Sandia National Laboratories







Sandia Neural Computing Overview
2016 AFOSR Digital Electronics Working Group
Arlington, VA

Kris Carlson, Ph.D.
Sandia National Laboratories

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.

Cognitive robotics and spiking neural networks

Embodied Cognition Theory: You can learn more about cognition if you study the whole brain-body system

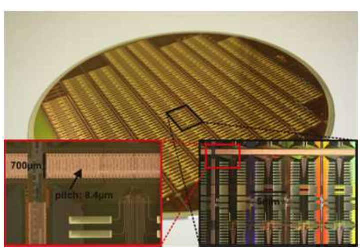
Carl from Jeff Krichmar's Lab at UC Irvine

What are SNNs?

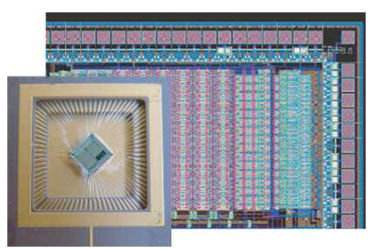
- Neural networks that model neuronal/synaptic temporal dynamics
- Spike only when the membrane voltage exceeds a threshold

Why use SNNs?

- Spike events are rare: average brain activity 1-10 Hz
- Event-driven nature of SNNs fits well with neuromorphic hardware
- Use "**Address Event Representation**" (**AER**) to minimize communication.
- SNNs provide temporal coding but can still use rate coding
- SNNs support biologically plausible learning rules



BrainScaleS Chip from Heidelberg University



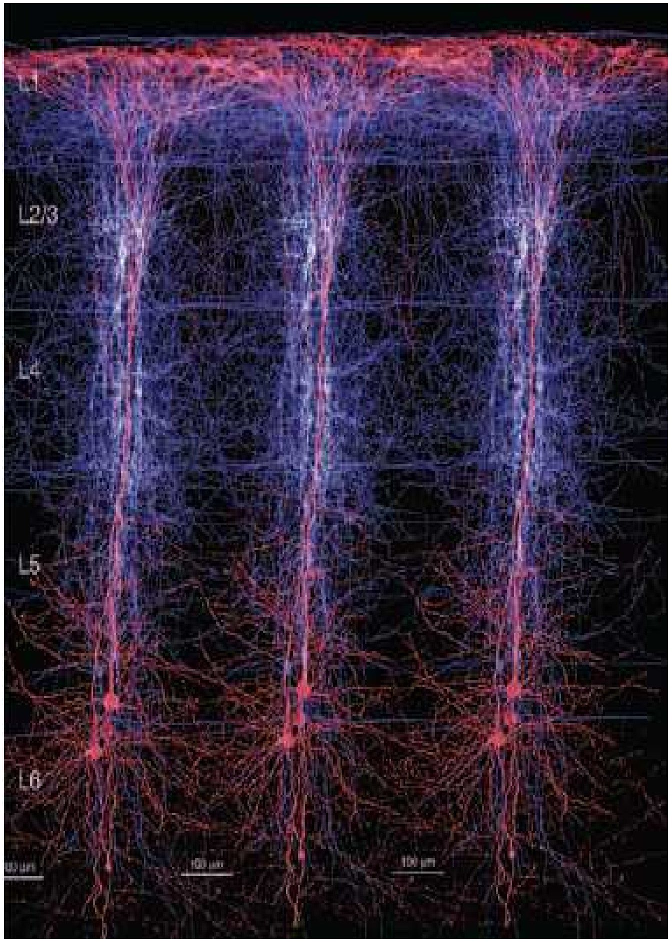
Neuromorphic Chip from INI in Zurich

TrueNorth chip from IBM

DVS camera from iniLabs in Zurich

Properties of neural systems

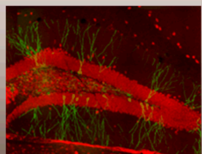
- Massive parallelism (10^{11} neurons)
- Massive connectivity (10^{15} synapses)
- Excellent power-efficiency
  - ~ 20W for 10^{16} flops
- Probabilistic responses and fault-tolerant
- Autonomous, on-line learning
- Low-performance components (~100 Hz)
- Low-speed comm. (~meters/sec)
- Low-precision synaptic connections



Neuromorphic computing at SNL leverages a broad research foundation

Neural Theory

- Modeling and Simulation
- Computational Neuroscience
- Neuro-informatics



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

Neural-enabling hardware

- Memory technology
- Micro-sensors
- Non-von Neumann architectures

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

Neural Computing Capabilities

- Neural Machine Learning Algorithms
- Formal Neural Computing Theory
- UQ / SA of Neural Algorithms and Neural Architectures
- Configurable CMOS Neural Architectures
- Adaptive post-CMOS Neural Architectures

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

Enabling Advanced Scientific and Data-driven Computing

- Neural-inspired communication paradigms
- Adaptive memory management
- Robust machine learning

Deployable National Security Applications

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

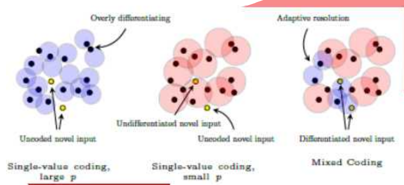
CCR Center for Computing Research

Sandia National Laboratories

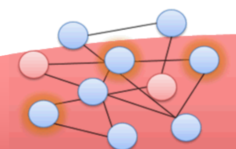
Computational neuroscience research spans several domains

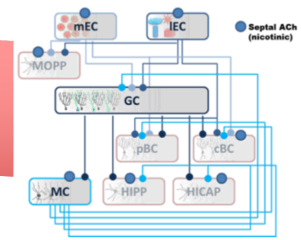
Neuroscience

Theory

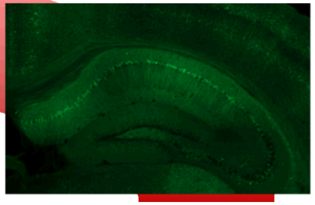


Computational models

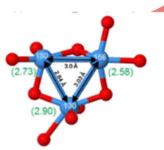




Data analytics



Tools

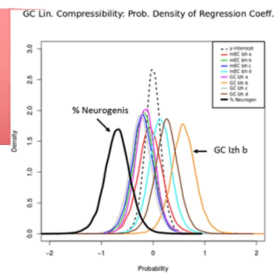


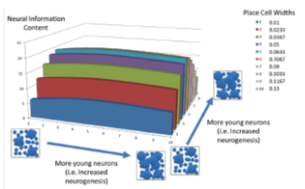
Simulation tools

Xyce

N2A

Novel Metrics





Desirable properties of an adaptive neuromorphic autonomous system

Power-efficient

Highly adaptive

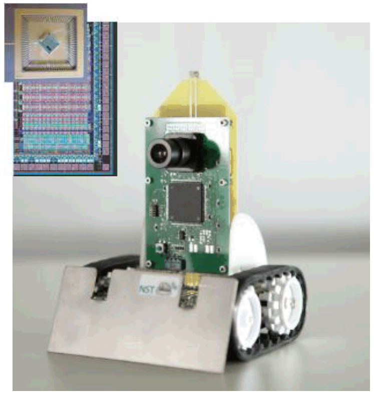
Built on solid CS/Neuro theoretical foundations

Handles streams of information in real-time

Utilizes a prototyping framework for testing & development

Verified & validated using proven engineering principles

Built on existing ML/DNN techniques



Pushbot from iniLabs in Zurich with neuromorphic chip from INI in Zurich

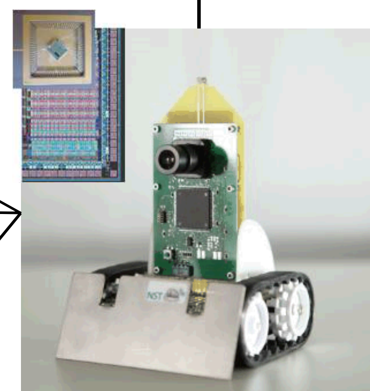
- Extremely low power camera/chip device only uses power when something happens
- Pattern recognition capabilities enable actions to be taken when specific event occurs
- DVS camera can operate at μs resolution
- Plasticity enables online learning of new patterns or habituation of irrelevant input

Desirable properties of an adaptive neuromorphic autonomous system

Handles streams of information in real-time

Power-efficient

Utilizes a prototyping framework for testing & development



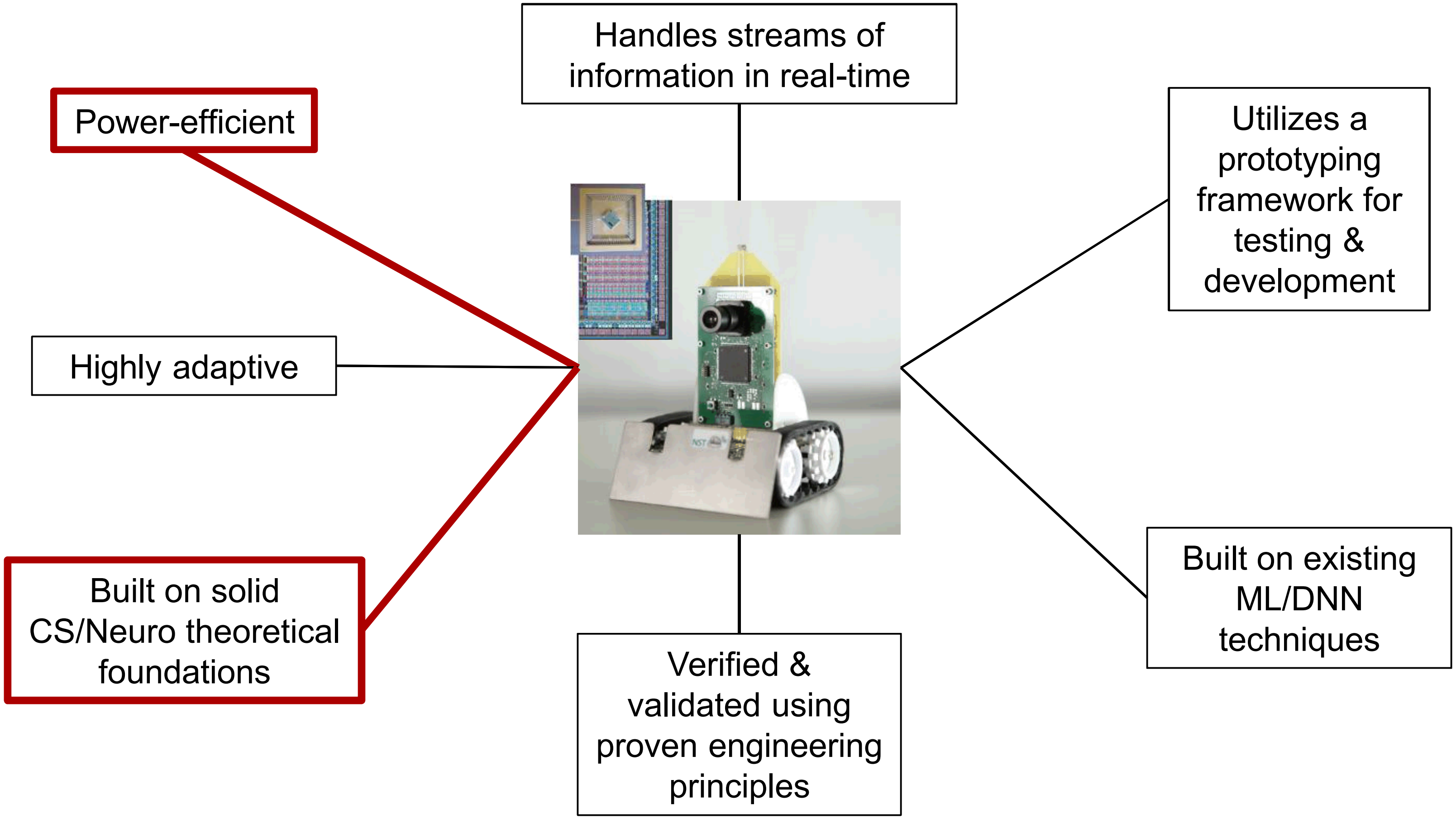
Highly adaptive

Built on solid CS/Neuro theoretical foundations

Verified & validated using proven engineering principles

Built on existing ML/DNN techniques

Desirable properties of an adaptive neuromorphic autonomous system



Theoretical computing advantages for neural computation

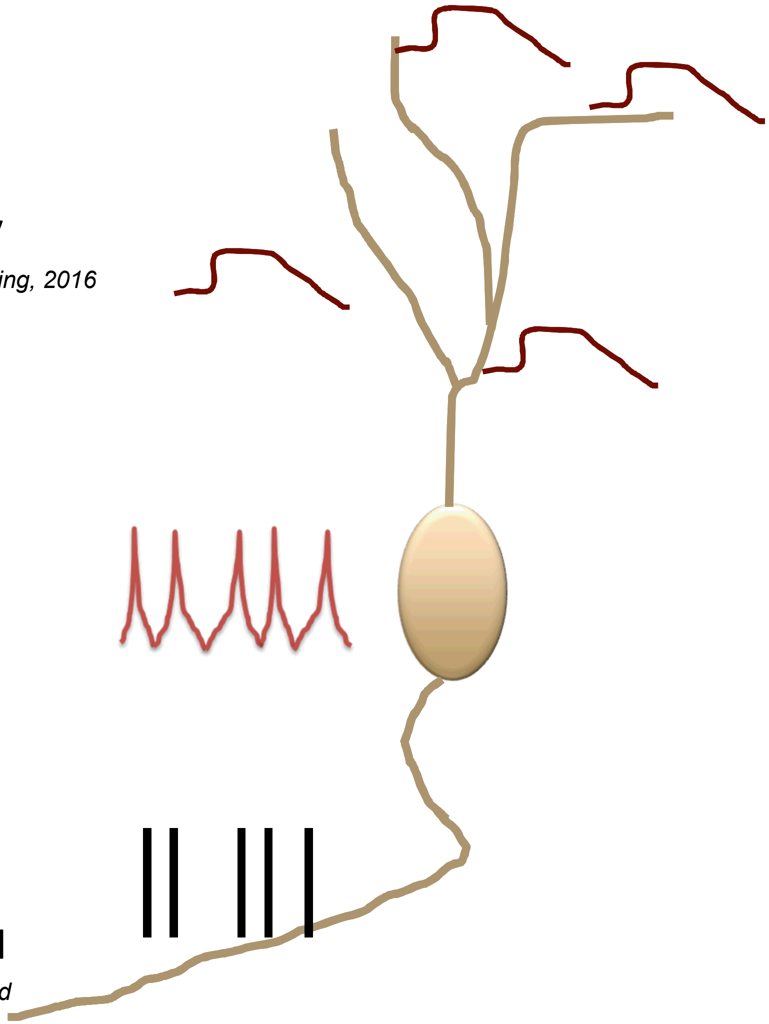
- Synaptic memories are independently accessed and integrated → Local Analog Computation
  - ***Example: O(N³) -> O(N²) advantage*** in energy of sparse coding due to analog processing at memory

Agarwal et al., Frontiers in Neuromorphic Computing, 2016

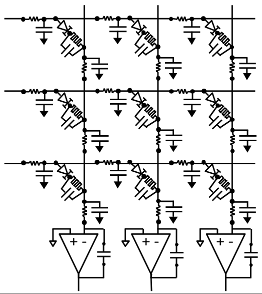
- Integrated inputs are transformed into "spike" trains → Information communicated in temporal domain (binary in voltage, analog in time)
  - ***Example: O(log N) -> O(1) advantage*** in energy due to temporal coding of communication

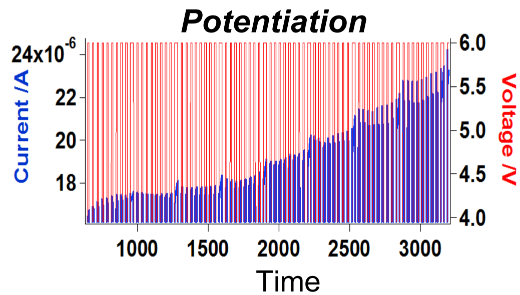
Verzi et al., in preparation

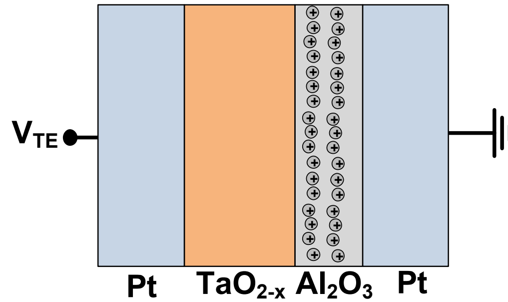
- Synapse memory, neuron dynamics, and even whole neurons change over time during learning → Continuously adaptive algorithms
  - ***Example: O(N⁴)->O(N²) advantage*** in amortized cost of training and running "deep learning" neural algorithms in changing world *Draelos et al., submitted*

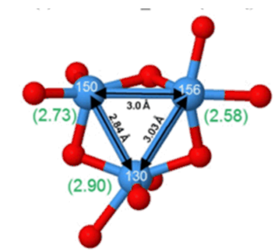


ReRAM crossbar research from device to system







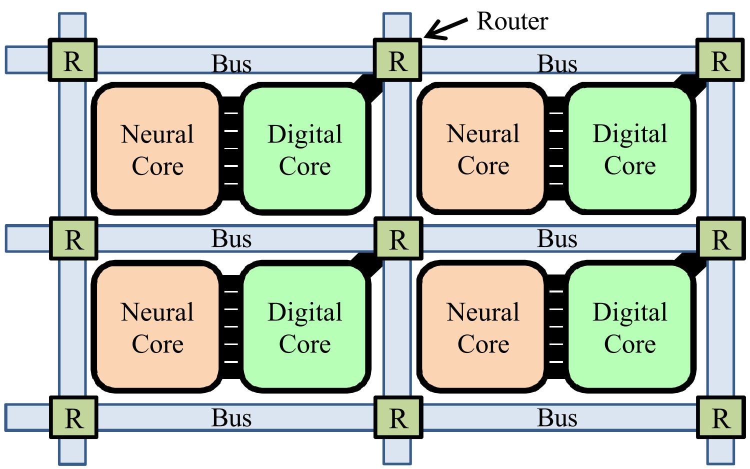


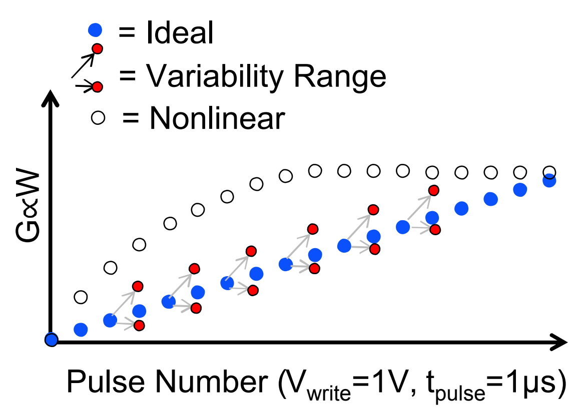
Learning Hardware System Model ⟷ **Device Electrical Characterization and Compact Modeling** ⟷ **Learning Device Development** ⟷ **Physical Mechanisms S&T**

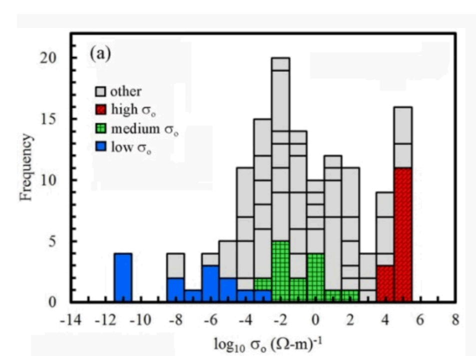
Run *any* neural algorithm on the same hardware

Variability and Nonlinearity

Conductance distribution for different atomic structures







CCR Center for Computing Research

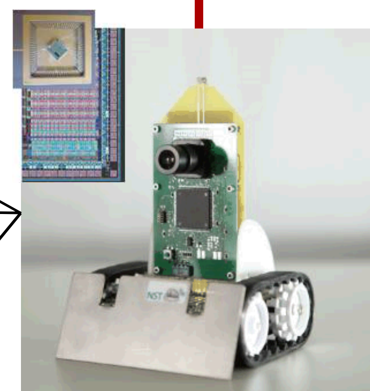
Sandia National Laboratories

Desirable properties of an adaptive neuromorphic autonomous system

Handles streams of information in real-time

Power-efficient

Utilizes a prototyping framework for testing & development



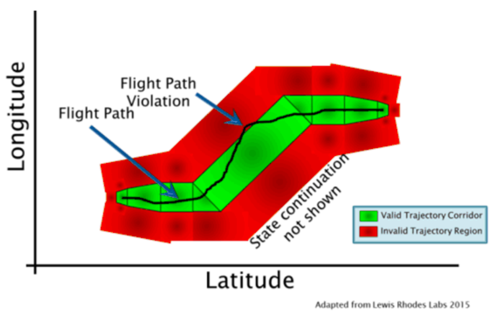
Highly adaptive

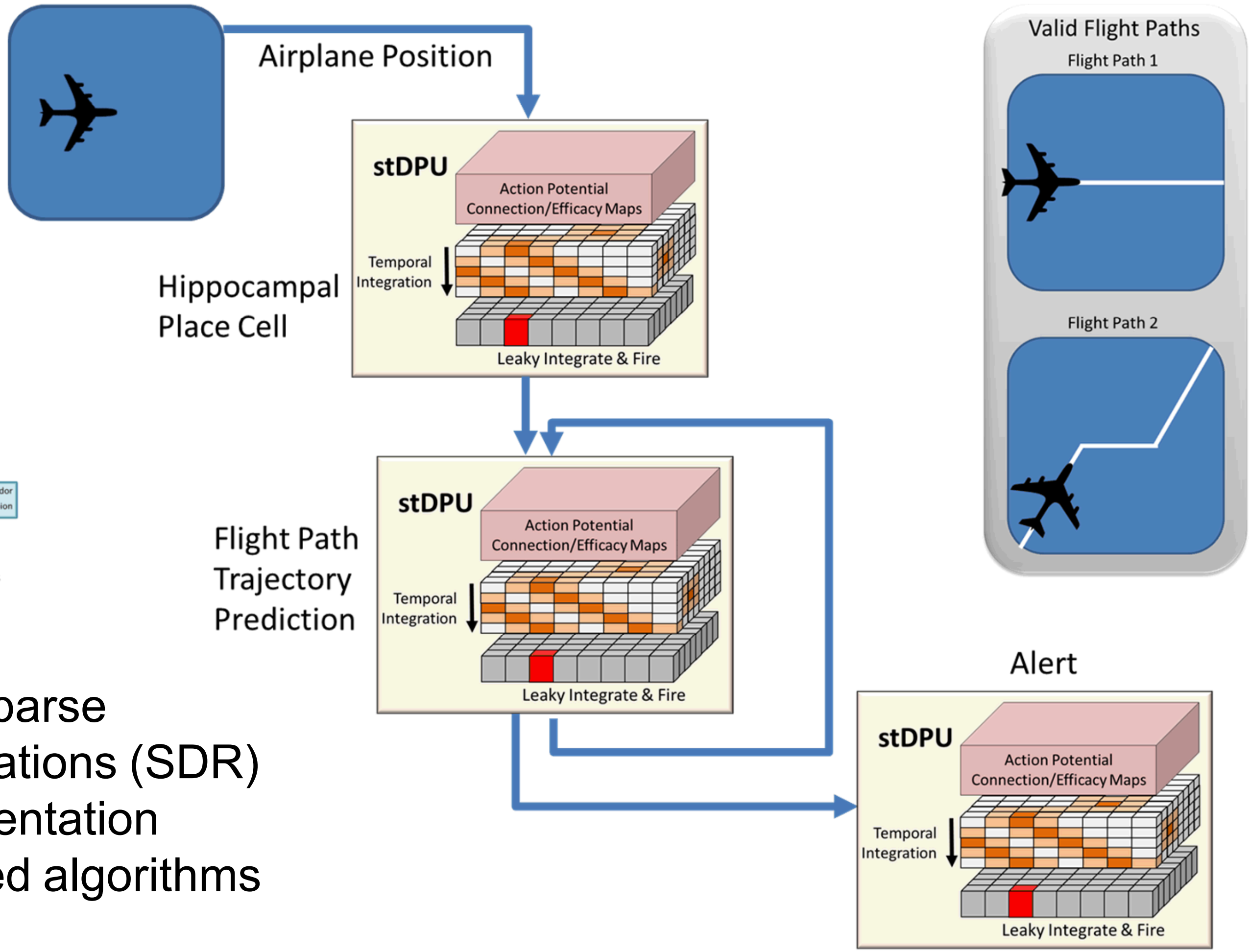
Built on solid CS/Neuro theoretical foundations

Verified & validated using proven engineering principles

Built on existing ML/DNN techniques

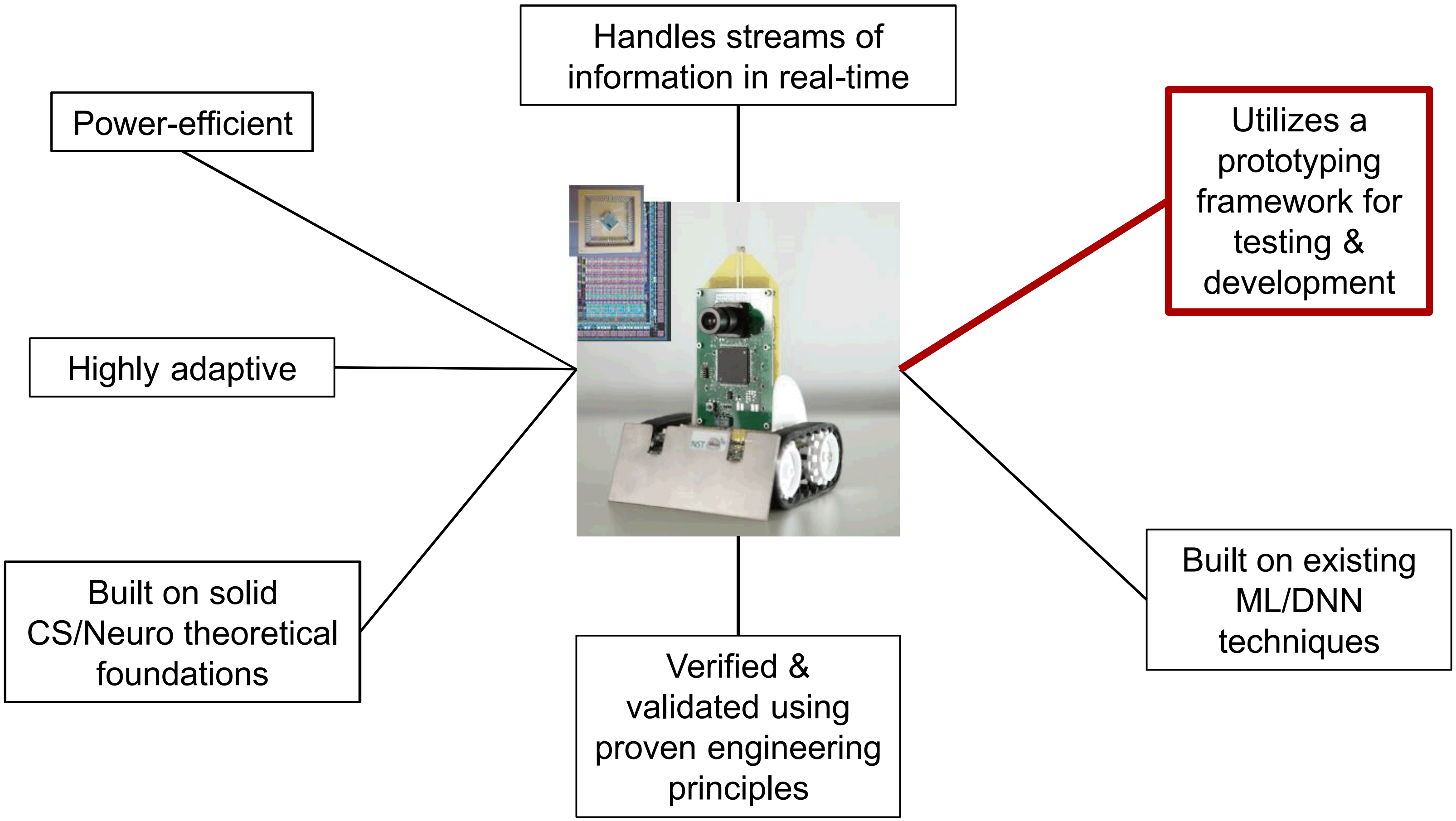
HAANA spiking architecture enables neural-inspired pattern recognition





- States encoded as sparse distributed representations (SDR)
- High capacity representation
- Hippocampus-inspired algorithms

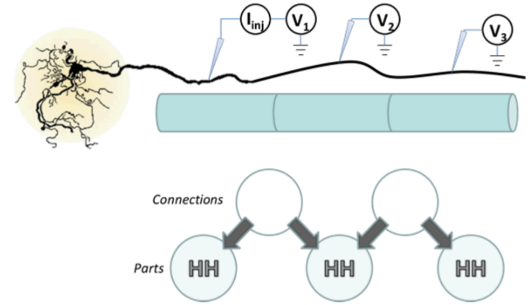
Desirable properties of an adaptive neuromorphic autonomous system



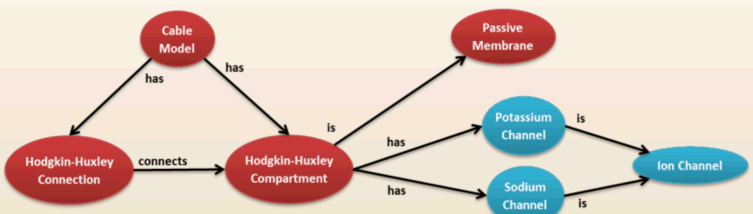
N2A –Compiler of neuroscience information into compute-friendly representations

Neuroscience data is vast and exists at many scales…

… but many neuroscience concepts can be reduced to dynamical physics-based models



N2A compositionally represents neural dynamics and compiles to conventional or neuromorphic systems




```
HH Connection
A     =  $connect ("HH Compartment")
B     =  $connect ("HH Compartment")
A.V' =+ (B.V - A.V) / R
B.V' =+ (A.V - B.V) / R
R     =  10

HH Compartment
parent = $inherit ("Passive Membrane")
K      = $include ("Potassium Channel")
Na     = $include ("Sodium Channel")

Cable Model
HH     = $include ("HH Compartment")
HH.$n = 3
C      = $include ("HH Connection")
C.A    = HH
C.B    = HH
C.$p   = C.A.$index == C.B.$index - 1
```

```
Sodium Channel
parent   =  $inherit ("Ion Channel")
I        =  G * m^3 * h * (E - V)
m'       =  alpha_m * (1 - m) - beta_m * m
h'       =  alpha_h * (1 - h) - beta_h * h
alpha_m =: (25 - V) / (10 * (exp ((25 - V) / 10) - 1))
beta_m  =: 4 * exp (- V / 18)
alpha_h =: 0.07 * exp (- V / 20)
beta_h  =: 1 / (exp ((30 - V) / 10) + 1)
G        =  120
E        =  115

Passive Membrane
V'       = (G * (V_rest - V) + I_inj) / C
G        = 0.3
V_rest = 10.613
C        = 1
```

C++

```
class Model : public _Compartment
{
public:
  class HHmod : public _Compartment
  {
  public:
    ...
    virtual void update (float _24t, float & _24dt)
    {
      float Na_2ealpha_5fh;
      Na_2ealpha_5fh = 0.07f * exp(-V / 20.0f);
      float Na_2ealpha_5fm;
      Na_2ealpha_5fm = (25.0f - V) / (10.0f * (exp((25.0f - V) / 10.0f) - 1.0f));
      float K_2ealpha_5fn;
      K_2ealpha_5fn = (10.0f - V) / (100.0f * (exp((10.0f - V) / 10.0f) - 1.0f));
      float Na_2ebeta_5fh;
      Na_2ebeta_5fh = 1.0f / (exp((30.0f - V) / 10.0f) + 1.0f);
      float Na_2ebeta_5fm;
      Na_2ebeta_5fm = 4.0f * exp(-V / 18.0f);
      float K_2ebeta_5fn;
      K_2ebeta_5fn = 0.125f * exp(-V / 80.0f);
      _nextV_27 += (0.3f * (10.613f - V) + I_5finj) / 1.0f + K_2eI / 1.0f + Na_2eI / 1.0f;
      Na_2eh_27 = Na_2ealpha_5fh * (1.0f - Na_2eh) - Na_2ebeta_5fh * Na_2eh;
      Na_2eI = 120.0f * pow (Na_2em, 3.0f) * Na_2eh * (115.0f - V);
      K_2eI = 36.0f * pow (K_2en, 4.0f) * (-12.0f - V);
      Na_2em_27 = Na_2ealpha_5fm * (1.0f - Na_2em) - Na_2ebeta_5fm * Na_2em;
      K_2en_27 = K_2ealpha_5fn * (1.0f - K_2en) - K_2ebeta_5fn * K_2en;
      if (_24index == 0.0f)
        I_5finj = 10.0f;
    }
    ...
  };
};
```

Rothganger et al., Frontiers in Neuroscience 2014

Center for Computing Research

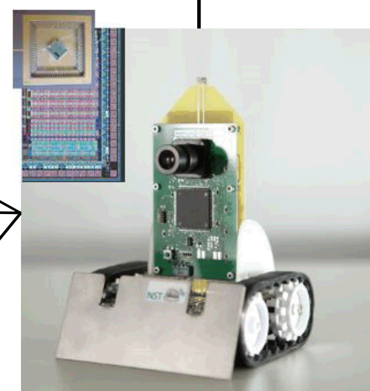
Sandia National Laboratories

Desirable properties of an adaptive neuromorphic autonomous system

Handles streams of information in real-time

Power-efficient

Utilizes a prototyping framework for testing & development



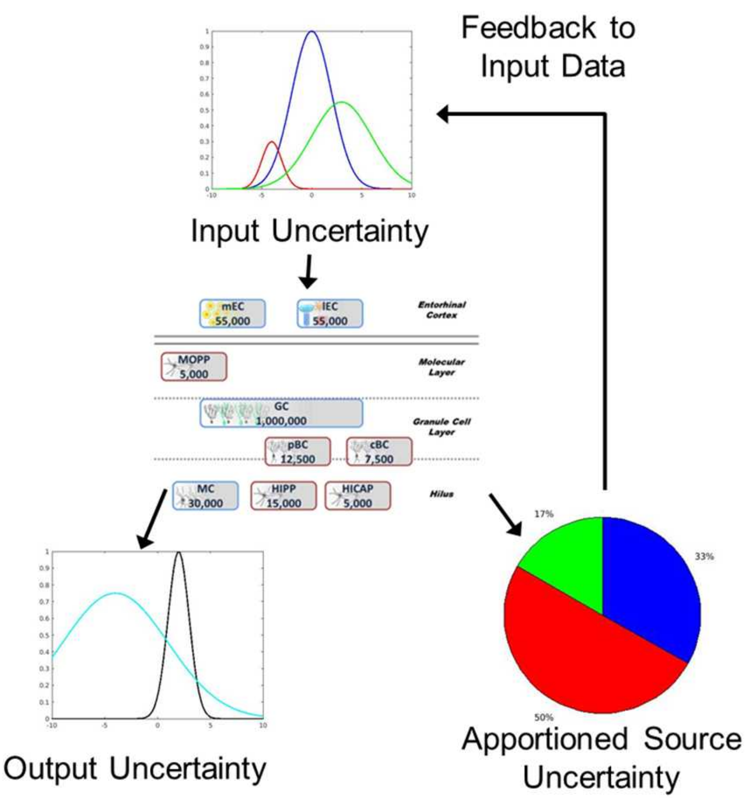
Highly adaptive

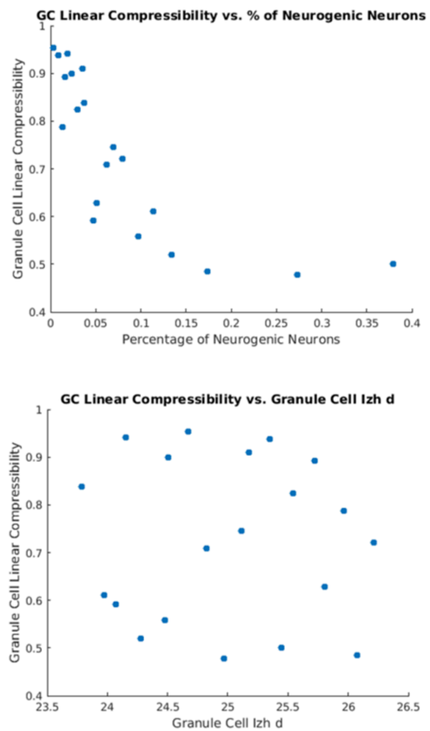
Built on solid CS/Neuro theoretical foundations

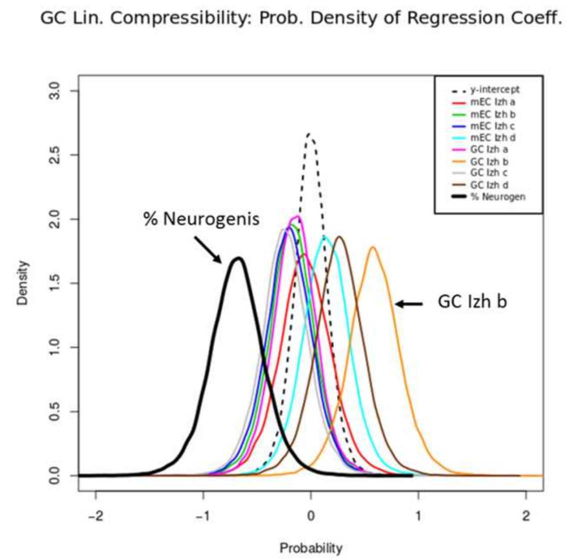
Verified & validated using proven engineering principles

Built on existing ML/DNN techniques

UQ / SA of computational neuroscience models and algorithms



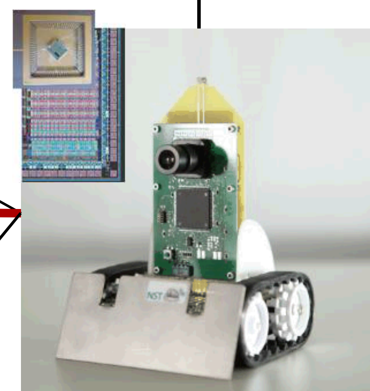




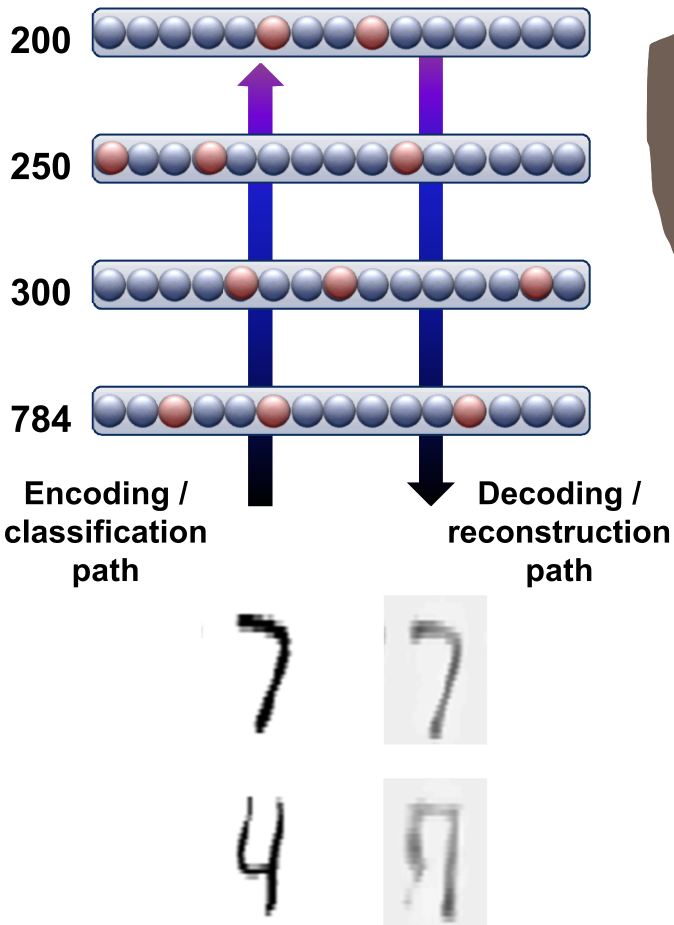
Carlson et al., in preparation

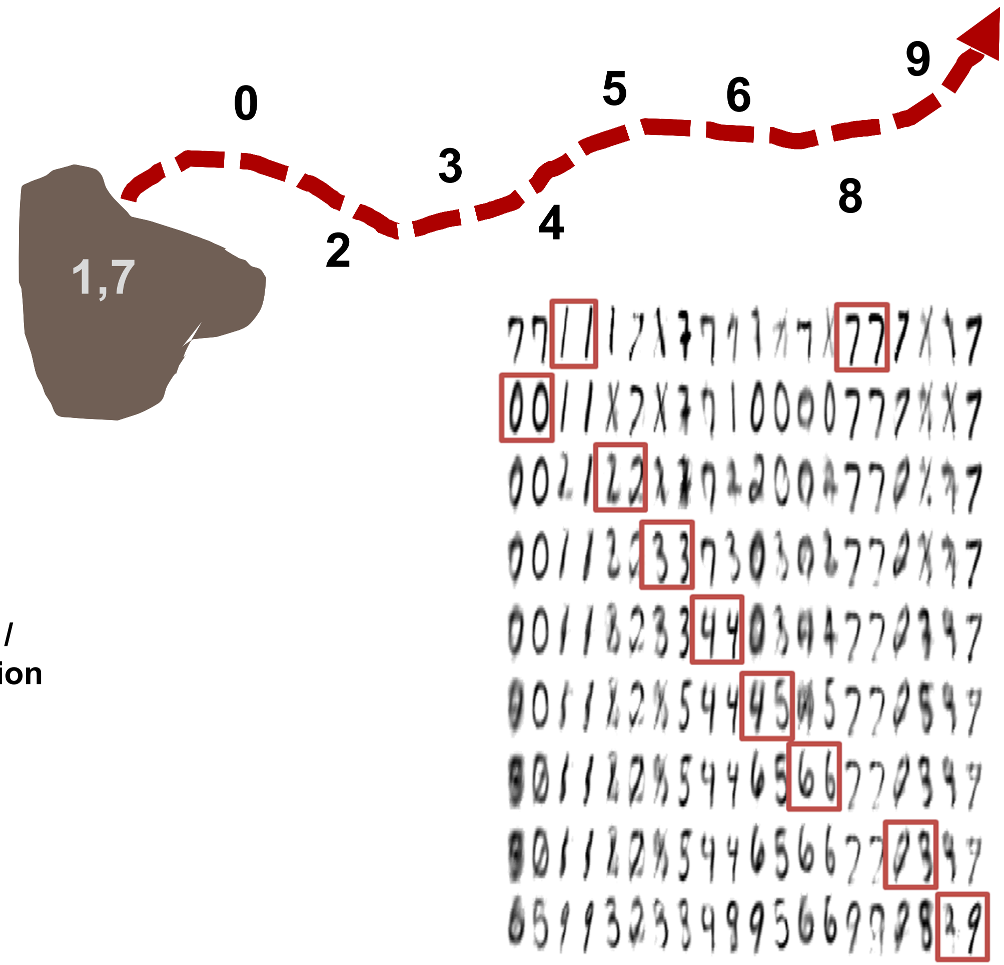
Desirable properties of an adaptive neuromorphic autonomous system

- Handles streams of information in real-time
- Power-efficient
- Utilizes a prototyping framework for testing & development
- Highly adaptive
- Built on existing ML/DNN techniques
- Built on solid CS/Neuro theoretical foundations
- Verified & validated using proven engineering principles



“Neurogenesis deep learning” enables adaptation to changing threats





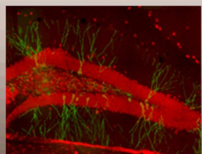
Draelos et al, submitted 2016

Neuromorphic computing at SNL leverages a broad research foundation

CCR Center for Computing Research — Sandia National Laboratories

Neural Theory

- Modeling and Simulation
- Computational Neuroscience
- Neuro-informatics



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

Neural-enabling hardware

- Memory technology
- Micro-sensors
- Non-von Neumann architectures

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

Neural Computing Capabilities

- Neural Machine Learning Algorithms
- Formal Neural Computing Theory
- UQ / SA of Neural Algorithms and Neural Architectures
- Configurable CMOS Neural Architectures
- Adaptive post-CMOS Neural Architectures

HAANA

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

Enabling Advanced Scientific and Data-driven Computing

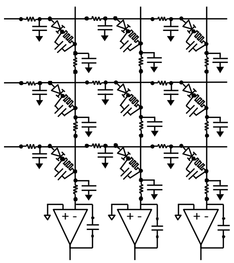
- Neural-inspired communication paradigms
- Adaptive memory management
- Robust machine learning

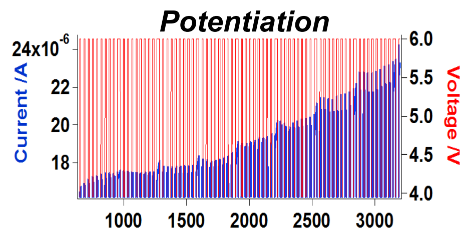
Deployable National Security Applications

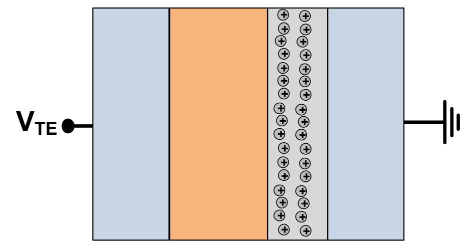
- Cyber Defenses
- Embedded Pattern Recognition Systems
- Smart Sensor Technologies

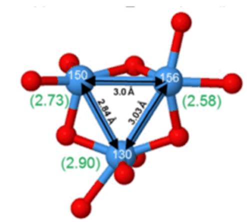
Additional SNL capabilities

We are working on reconfigurable hardware solutions that leverage the theoretical benefits of spiking and analog processing w/ Matt Marinella and Sapan Argawal in department 1768







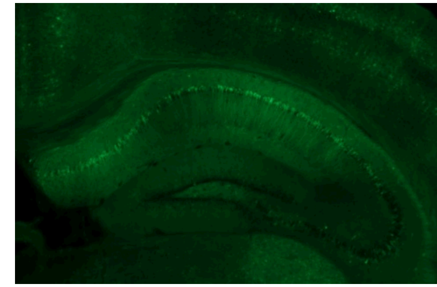


Learning Hardware System Model ⟷ Device Electrical Characterization and Compact Modeling ⟷ Learning Device Development ⟷ Physical Mechanisms S&T

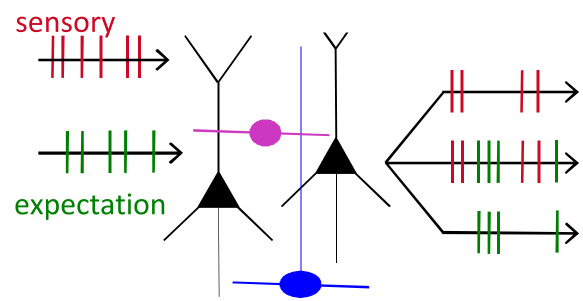
We also have an Intelligent System Controls Department (6533) that has capabilities in:

- Advanced mobility
- Cybernetics
- Advanced Controls
- Small smart machines
- Augmented reality

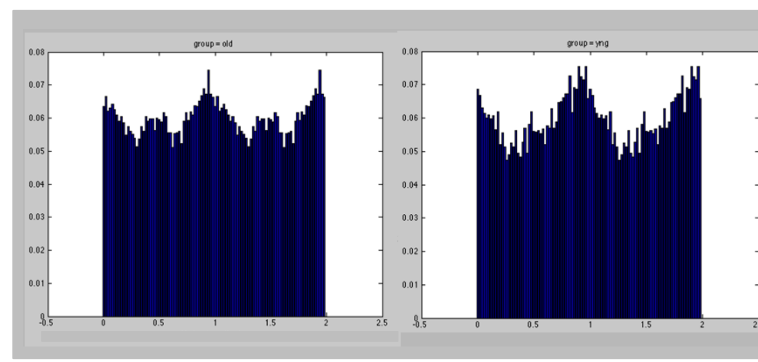
Data analytics of experimental neural data refines neural frameworks



<u>Multimodal Information Multiplexing in the Hippocampus</u>
How is multi-modal information represented and processed in the brain?



Prediction: Sensory and memory information are multiplexed in the spiking outputs of CA1 pyramidal neuron.

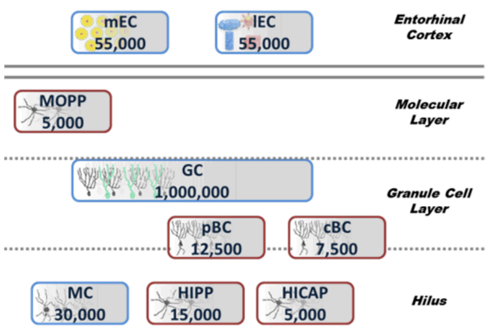


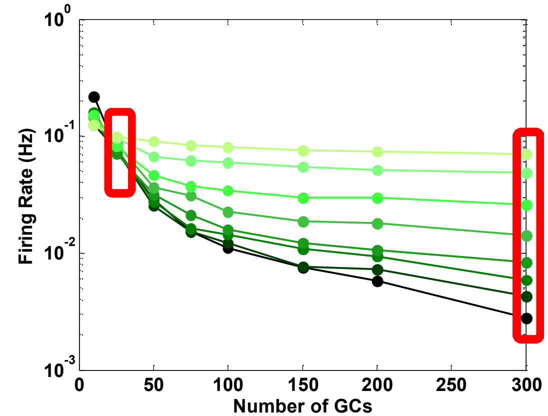
Validation: comparison of model spike patterns with hippocampal CA1 place cells (in vivo recordings from awake and behaving animals)

Courtesy Frances Chance; experimental neuroscience collaborators: Carol Barnes (U. Arizona), Sara Burke (U. Florida), Andrew Maurer (U. Florida)

Large scale neural simulations are required to observe of realistic neural functions

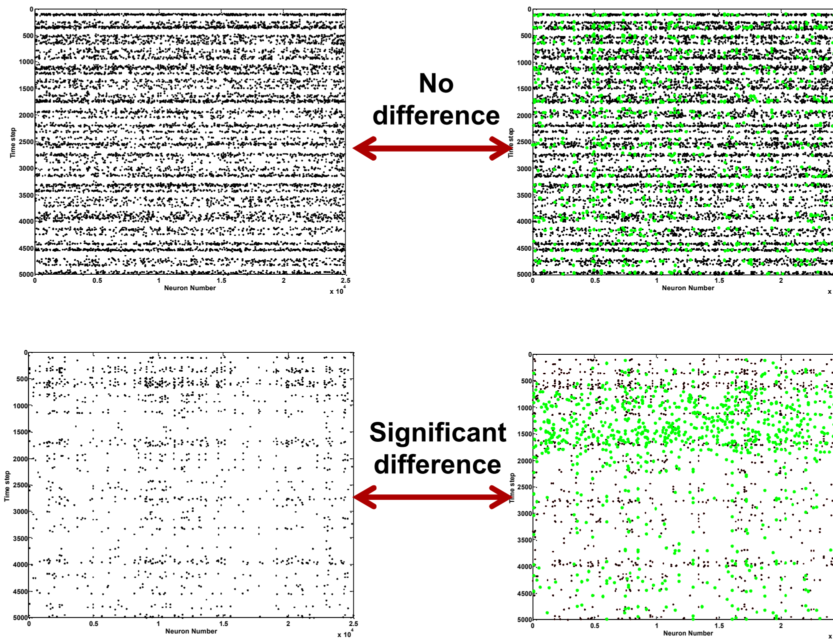
- Neural systems are highly non-linear and can involve complex feedback
- Scaling down neural simulations can have unintended implications
  - Example: Sharp increase in activity of reduced models shown here obfuscates experimental difference
- Models scale at number of interactions (roughly $O(N^2)$) and require substantial node-node communication





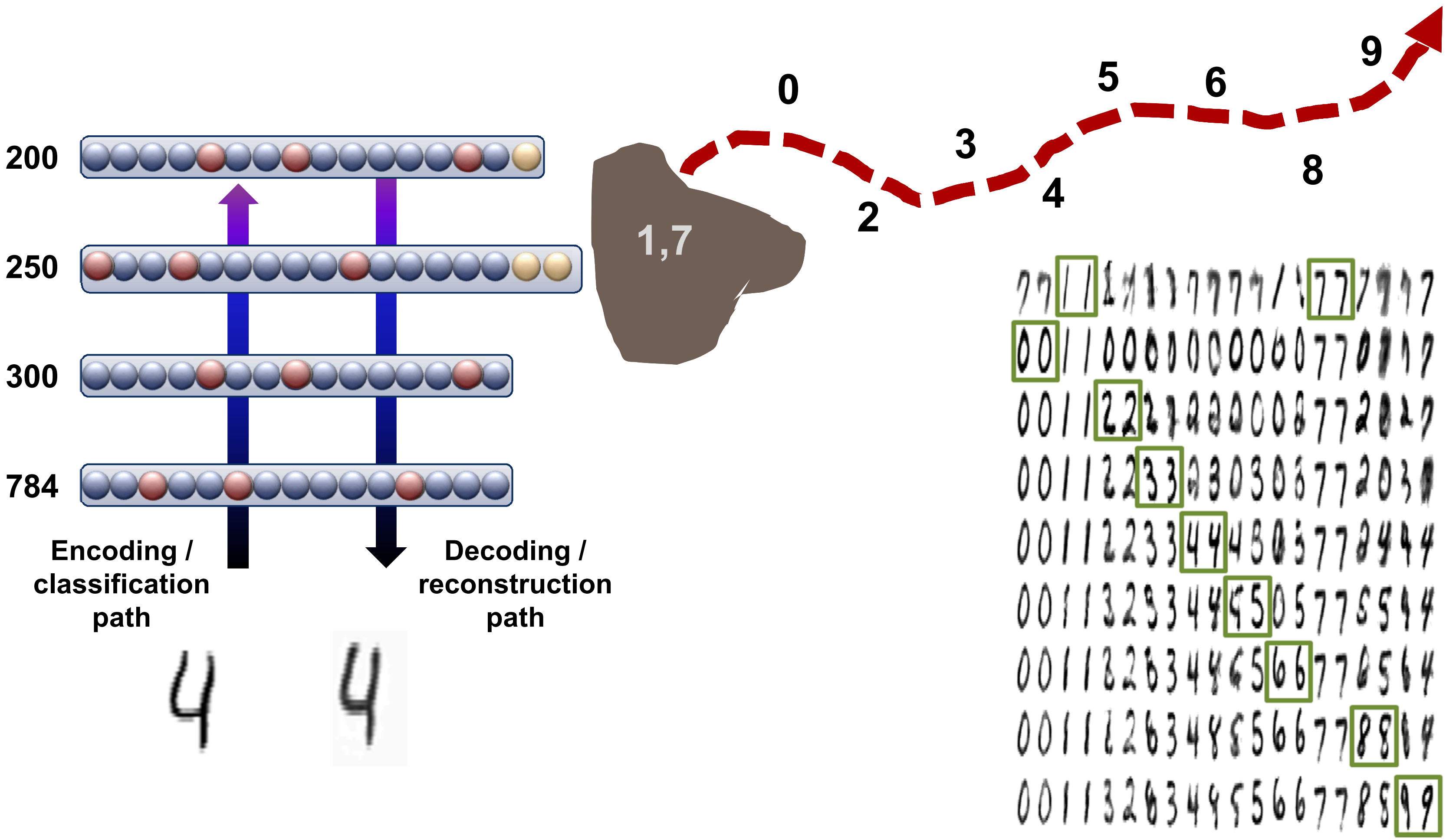
1/10th scale

Full mouse-brain scale



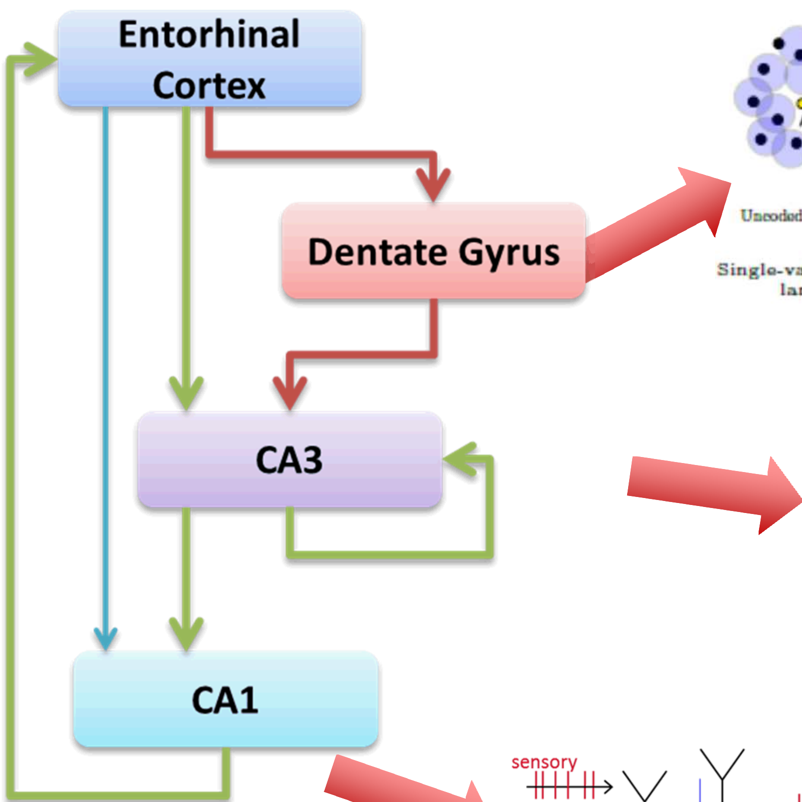
Carlson, Aimone et al., in preparation

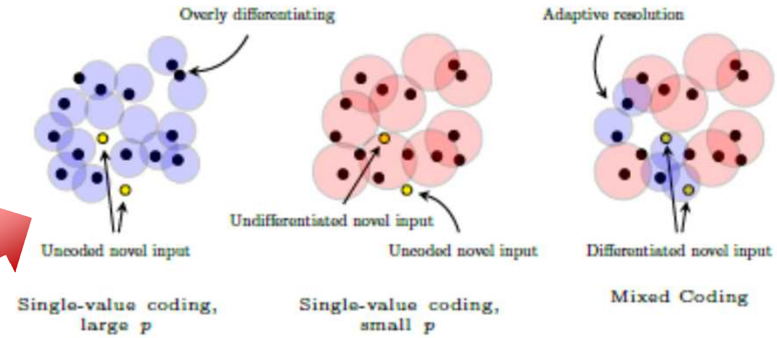
"Neurogenesis deep learning" enables adaptation to changing threats



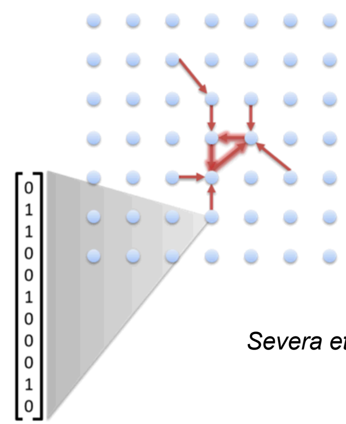
Draelos et al, submitted 2016

Theoretical efforts are seeking to formalize neural computation

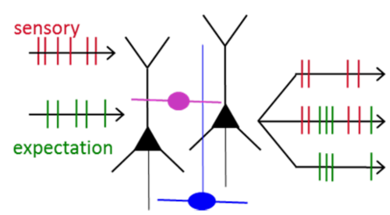




Severa et al., in preparation

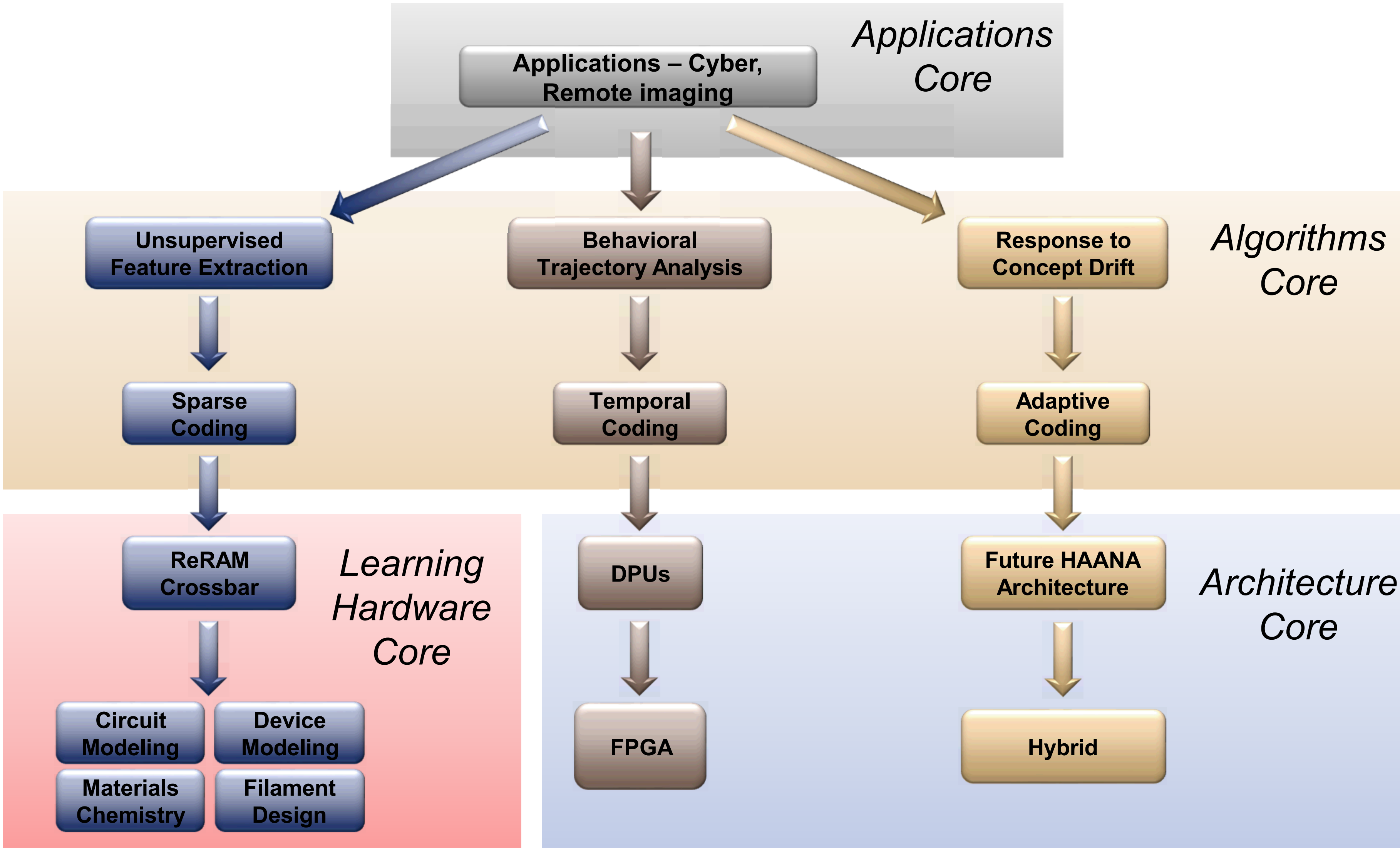


Severa et al., in preparation



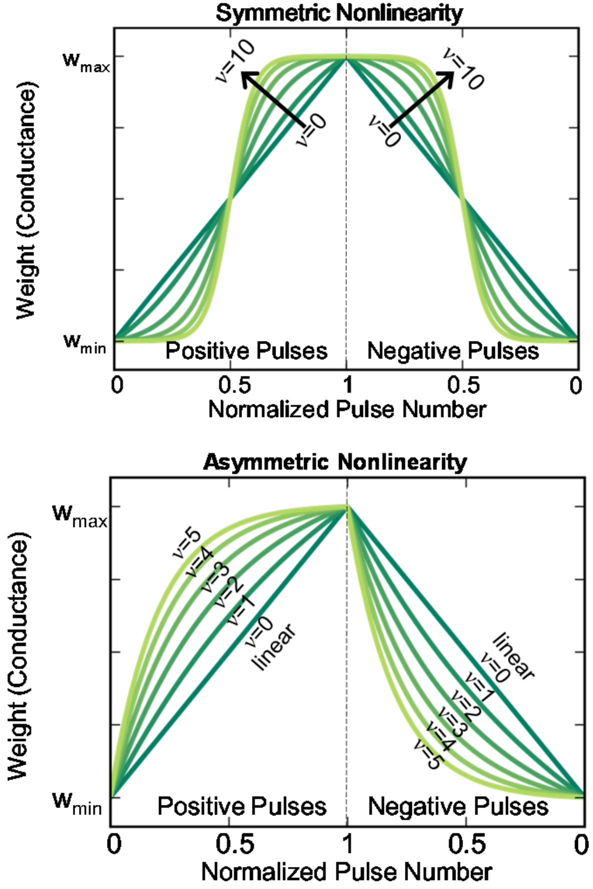
Chance et al., in preparation

HAANA Grand Challenge integrates research from across Sandia



Modeling Algorithms on Neural ReRAM Architectures Defines Device Requirements

	Small Images	Large Images	File Types
Read Noise σ (% Range)	3%	5%	9%
Write Noise σ (% Range)	0.3%	0.4%	0.4%
Asymmetric Nonlinearity (v)	0.1	0.1	0.1
Symmetric Nonlinearity (v)	>20	5	5
Maximum Current	160 nA	13 nA	40 nA



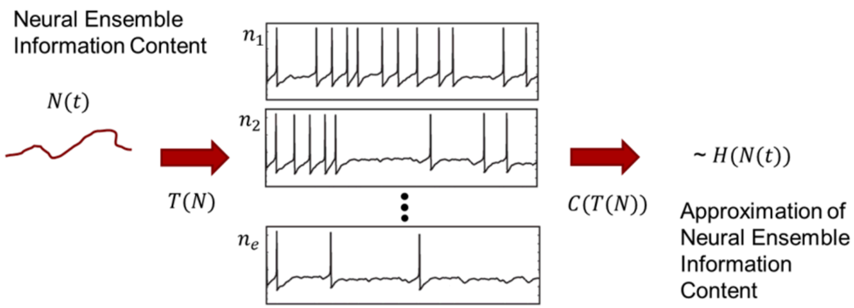
- <u>**Solution: Devices, Circuits, and Algorithms**</u>

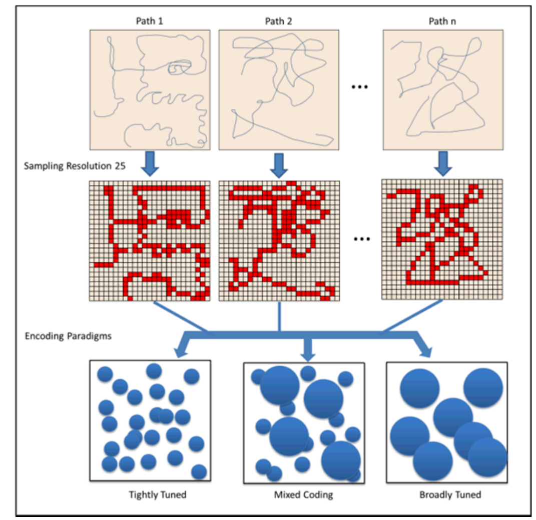
1. **Algorithms: Simulated annealing, HTM, LCO algorithm, HAANA Algorithm**
2. **Circuit: Multi-ReRAM circuit; parasitic compensation**
3. **Devices: Nonfilamentary, seeded/controlled filament**

CCR Center for Computing Research

Sandia National Laboratories

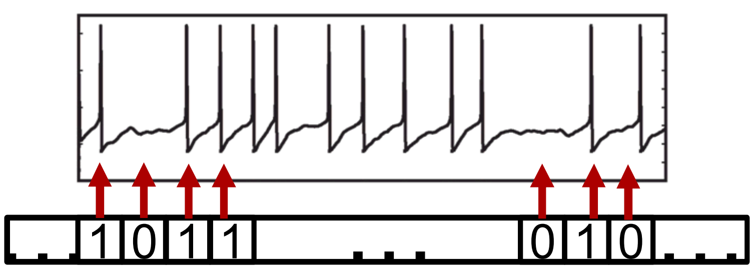
Neural information content metrics make quantifying neural computing concrete

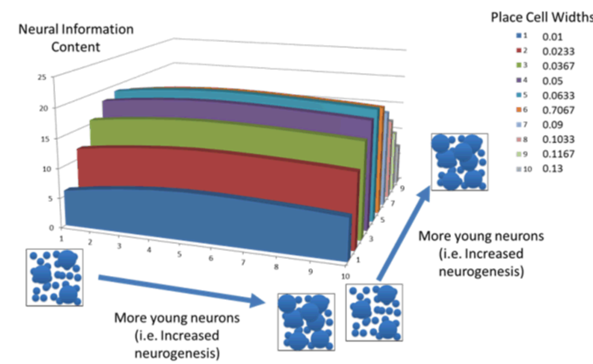




- Use complexity as a measure of compressibility in order to estimate entropy to quantitatively assess the information content of a signal

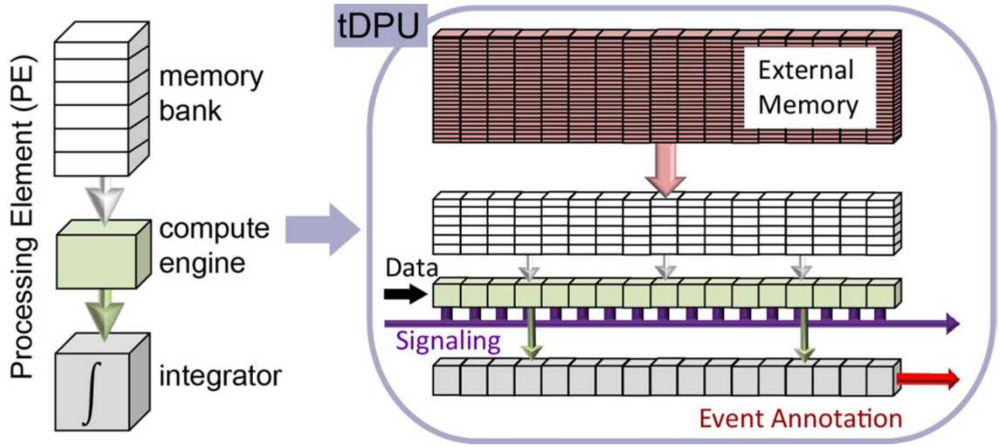
$$c_\alpha(x^n) = \frac{C_\alpha(x^n)}{n} * log_\alpha n$$

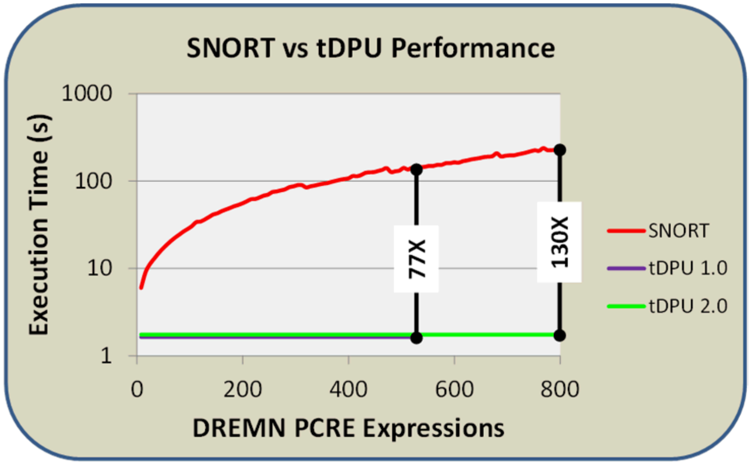




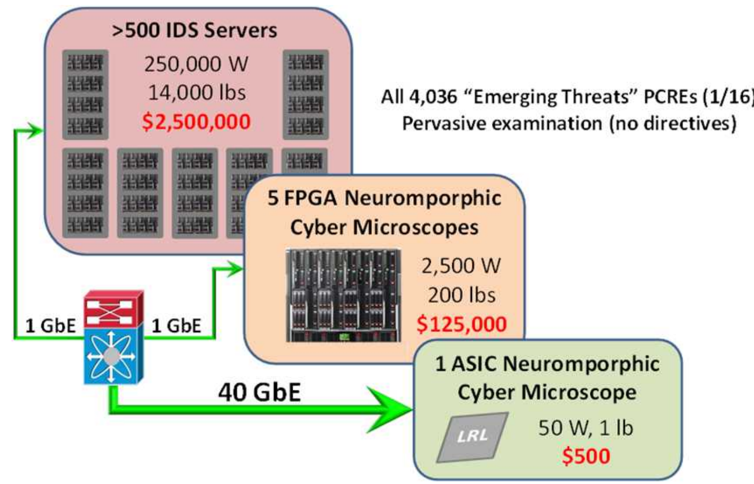
Vineyard, Craig M., et al. "Adult Neurogenesis: Implications on Human And Computational Decision Making." *Foundations of Augmented Cognition*. Springer Berlin Heidelberg, 2013. 531-540.
Vineyard, Craig M., et al. "Quantifying Neural Information Content: A Case Study of the Impact of Hippocampal Adult Neurogenesis" (Accepted to IJCNN 2016)

HAANA streaming architecture provides neural-inspired cyber analytics





- Architecture inspired by brain trauma emulation
- Exploits many features from observed neural processing
- Demonstrated 100X speedup for cyber complex pattern recognition (PCRE rule search) application



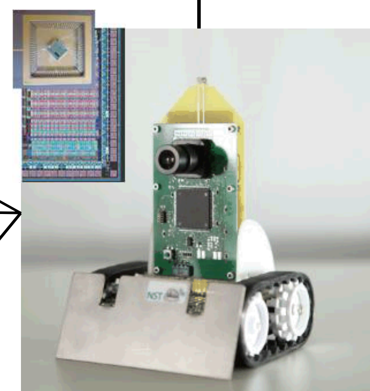
Desirable Properties of an Adaptive Neuromorphic Autonomous System

Handles streams of information in real-time

Power-efficient

Utilizes a prototyping framework for testing & development

Highly adaptive



Built on solid CS/Neuro theoretical foundations

Verified & validated using proven engineering principles

Built on existing ML/DNN techniques

Pushbot from iniLabs in Zurich with neuromorphic chip from INI in Zurich