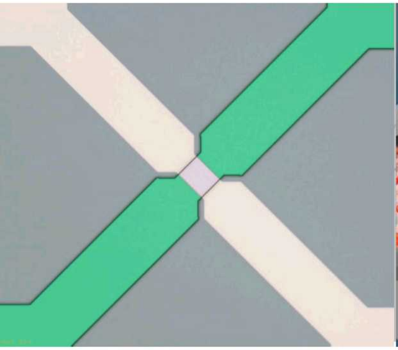
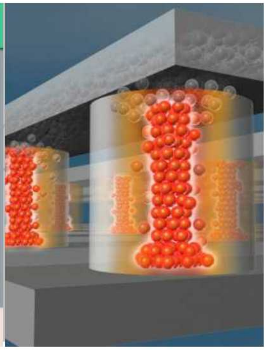
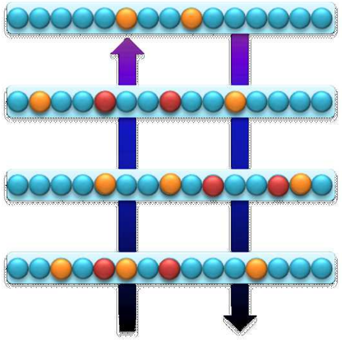
SAND2018-6790C







Bridging Theory to Application in Neuromorphic Computing: Mapping the Liquid State Machine on to the STPU

Michael R. Smith
Sandia National Laboratories

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.SAND NO.

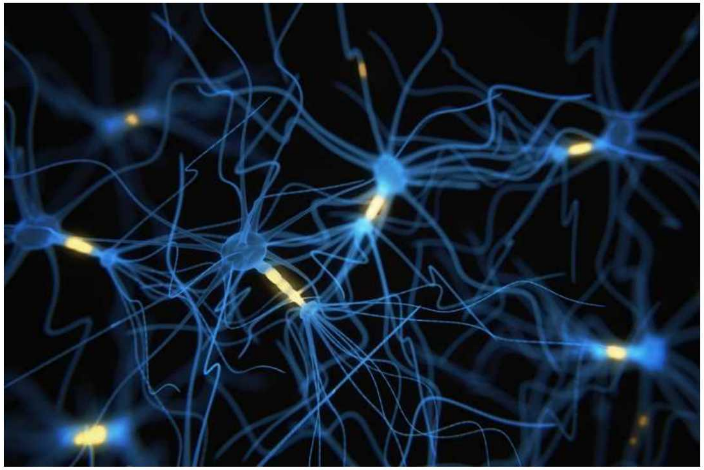
Neuromorphic Computing

Several Architectures:

- IBM True North
- Intel Loihi
- SpiNNaker
- BrainScaleS

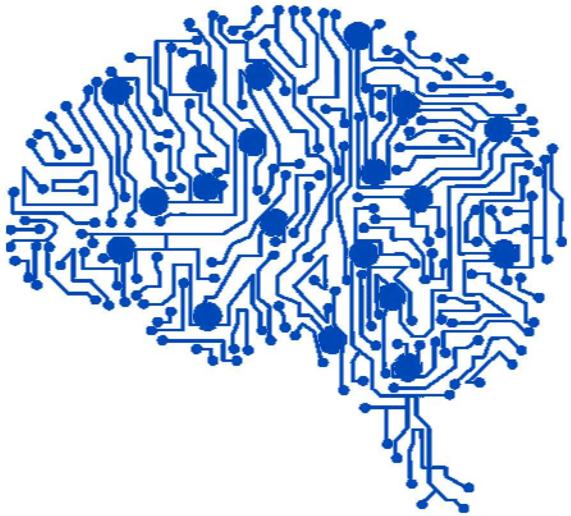
Bottom-Up Approach:

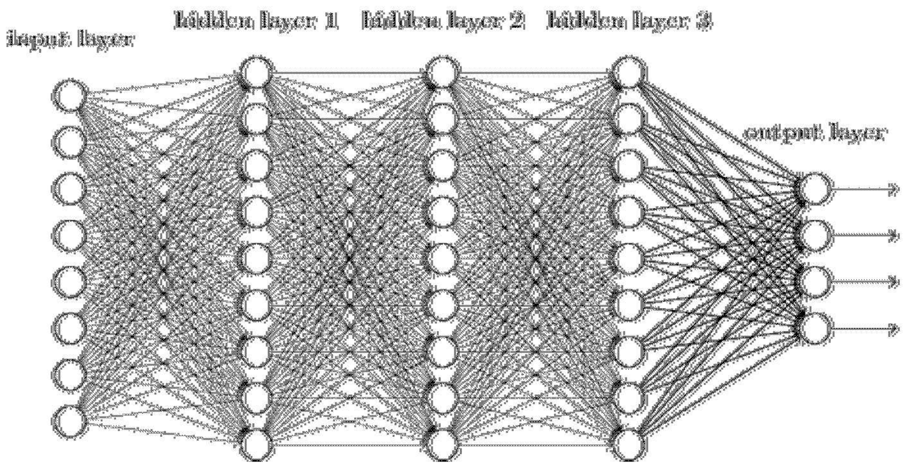
- Sound neuroscience foundation
- Theoretical computational and power benefits



Neural Networks and Deep Learning

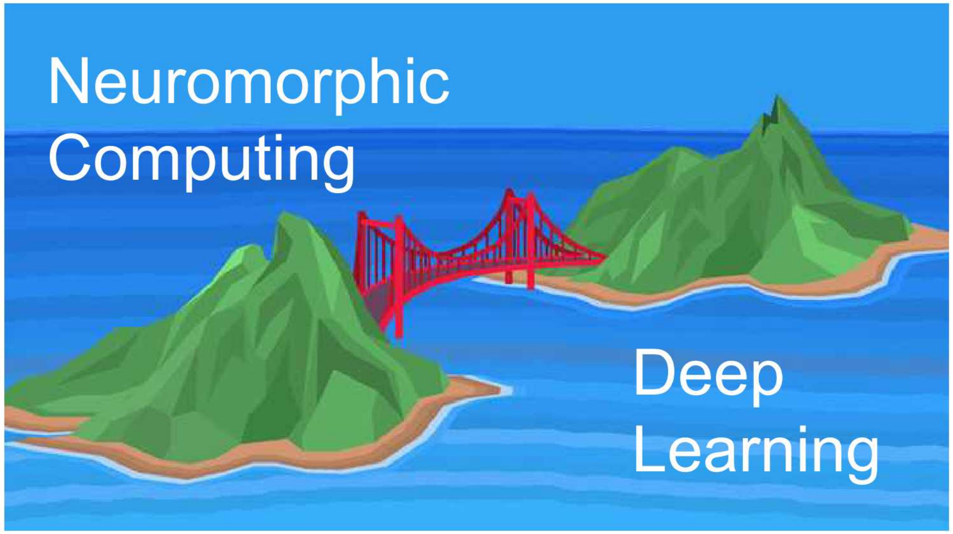
- "Neural inspired" machine learning algorithms
- Neuroscience developed in the 1940's and 50's
  - 1943-McCulloch –Pitts neuron model
  - 1949-Hebbian learning
  - 1957-Perceptron model proposed by Rosenblatt
- Top-down approach
- State-of-the-art in several domains
  - Image Processing
  - Speech





Bridging Neuromorphic Computing and Deep Learning

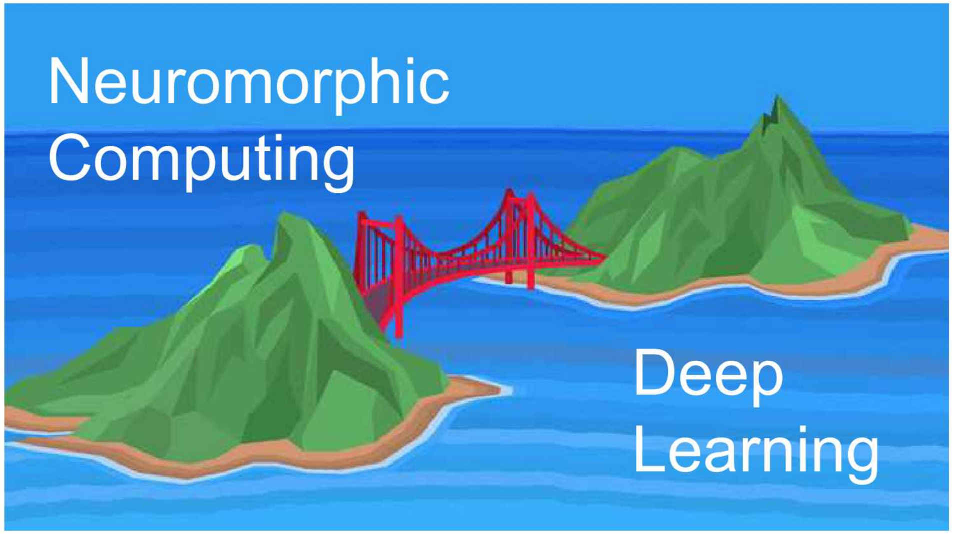
- Steve Furber (SpiNNaker): "There currently is a lot of hype about neuromorphic computing. … there is currently no compelling demonstration of a high-volume application where neuromorphic outperforms the alternative."
- William J. Dally (Stanford/Nvidia): "People who do conventional neural networks get results and win the competitions. The neuromorphic approaches are interesting scientifically, but they are nowhere close on accuracy."

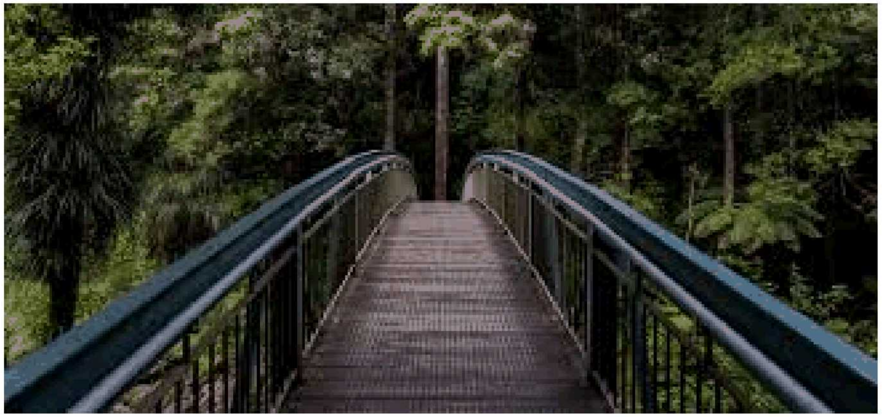


- Brian Van Essen (Lawrence Livermore National Laboratory): been able to get neural networks to run on TrueNorth but that the task of picking the right network and then successfully porting it over remains "a challenge."

Bridging the gap between theory and application (Why is this difficult?)

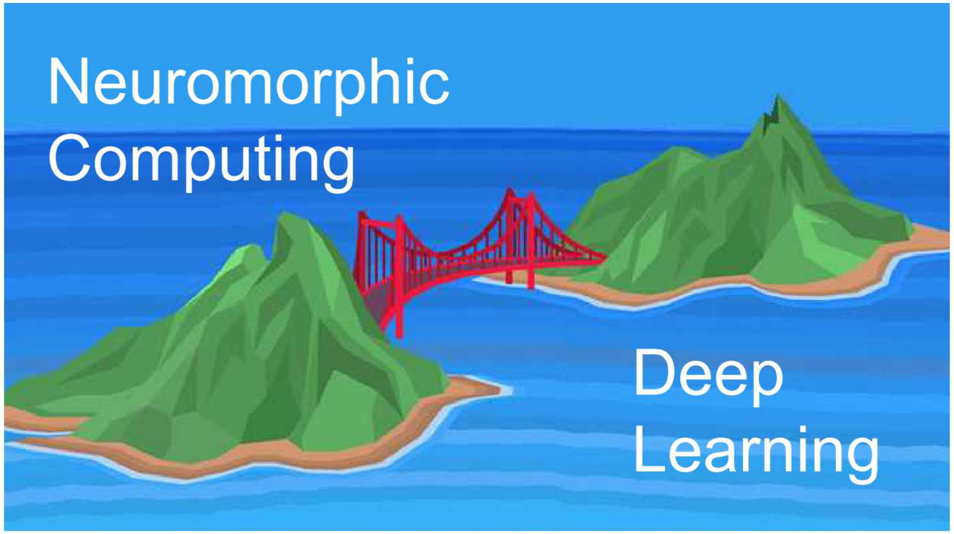
In theory, the difference between theory and practice is small. In practice, the difference between theory and practice is large. *In theory, theory and practice are the same. In practice, they are not.* -- Albert Einstein





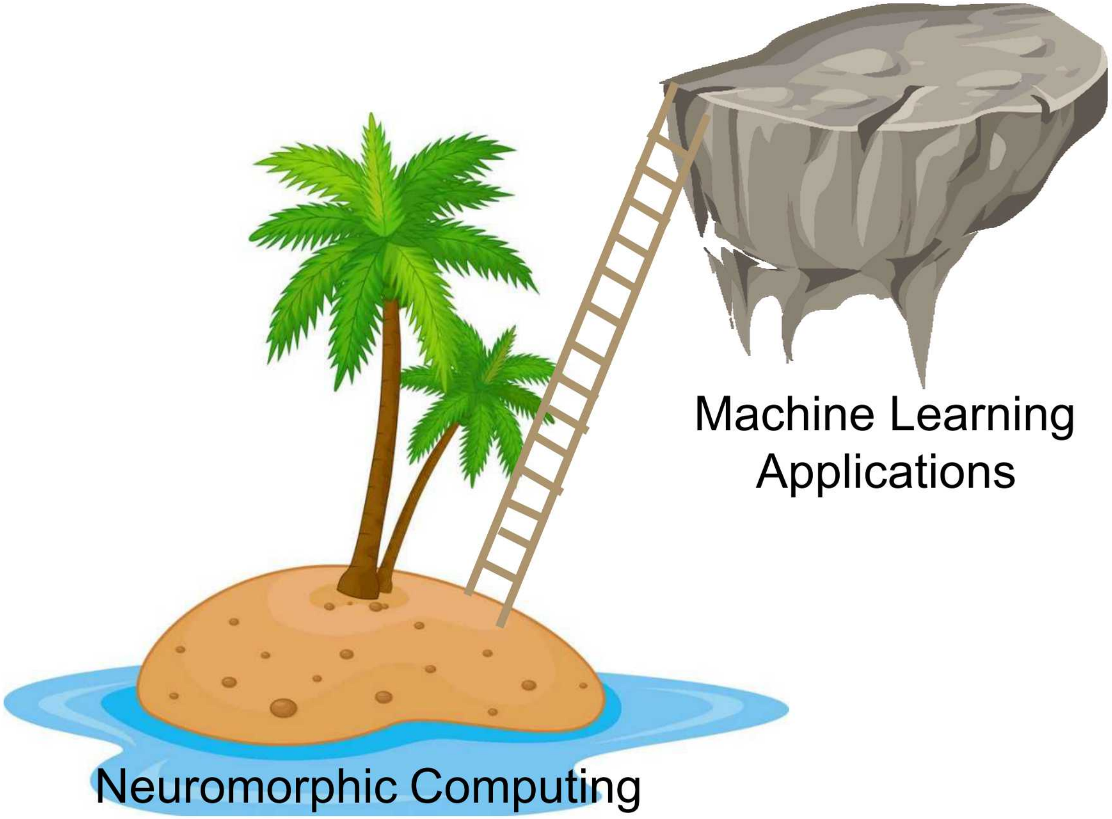
Bridging the gap between theory and application

In theory, the difference between theory and practice is small. In practice, the difference between theory and practice is large. *In theory, theory and practice are the same. In practice, they are not.* -- Albert Einstein



A Foundation with a Destination

- Often forget about the overall goal and focus on the neuroscience or hardware aspects
- Use neuromorphic principles as a foundation to higher-level applications



Our Application: Implement the Liquid State Machine onto the STPU

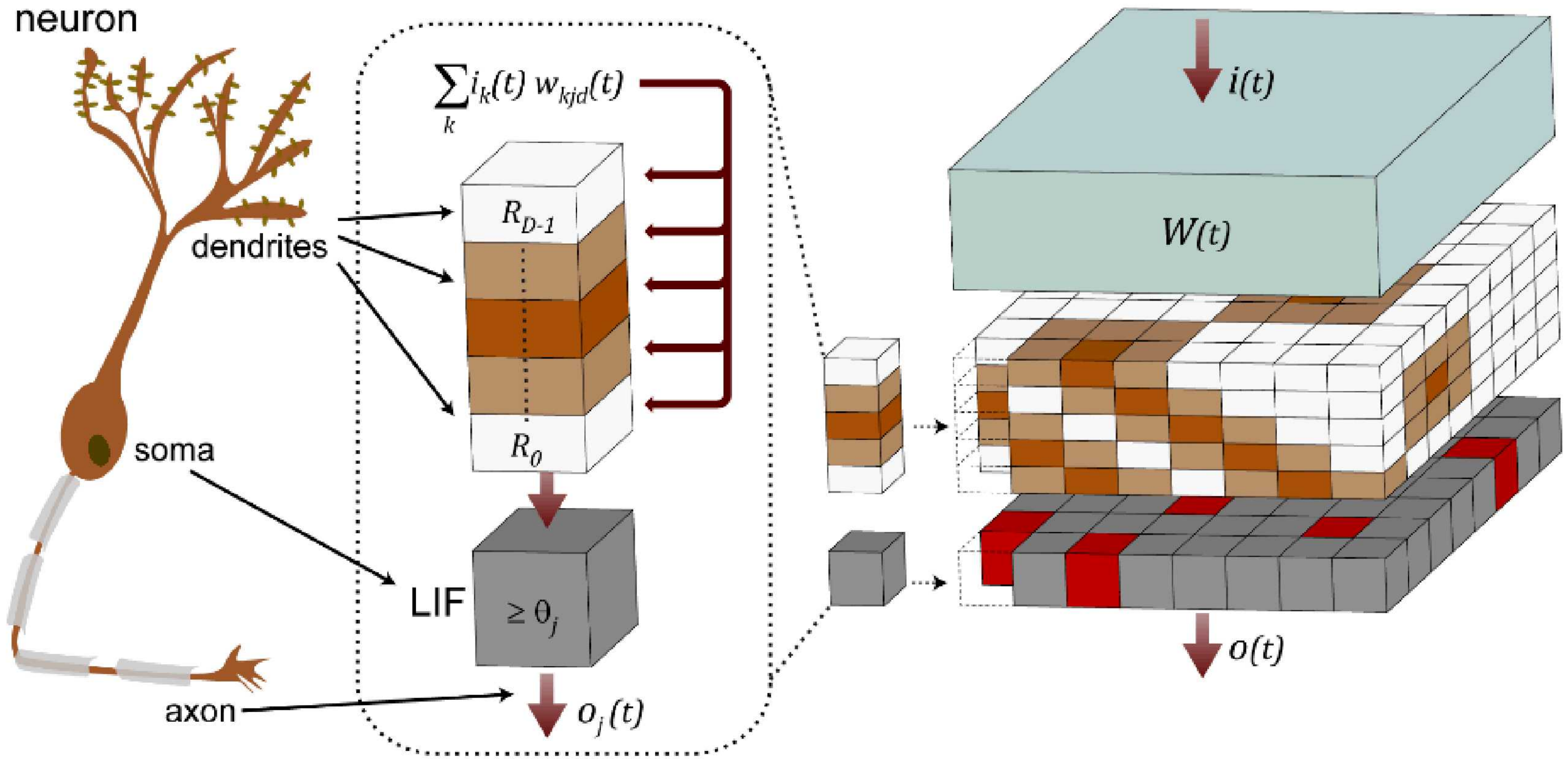
Spiking Temporal Processing Unit (STPU)

- Neuro-inspired
  - Highly parallel
  - LIF neurons
  - Binary spikes
- Fast
- Lower power
- High band in and out

Liquid State Machines (LSM)

- Mimics cortical column functionality
- Temporal input
- Differences between the patterns are amplified by the liquid
- Liquid of LIF neurons

Spiking Temporal Processing Unit

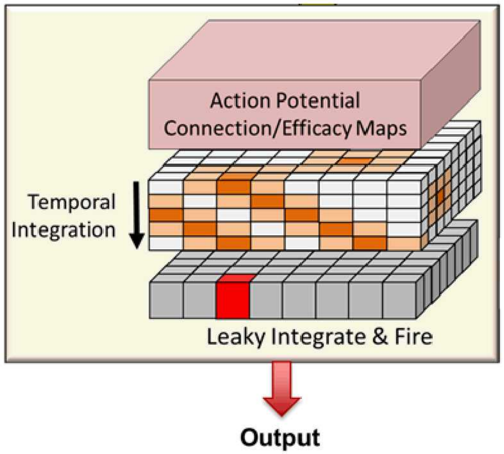


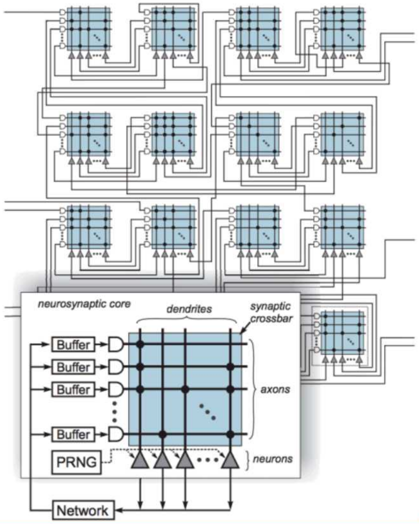
$$v_j^n = v_j^{n-1} - \frac{v_j^{n-1}}{\tau_j} + \sum_k \sum_d w_{kjd} \cdot s(t - t_{kd} - \Delta_{kd})$$

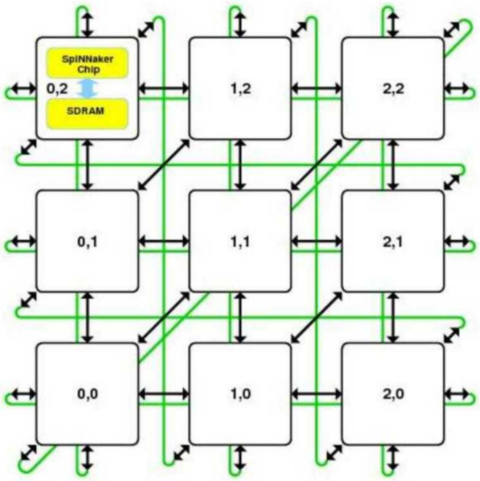
Neuromorphic Comparison

Platform	STPU	TrueNorth	SpiNNaker
Interconnect:	3D mesh multicast	2D mesh unicast	2D mesh mutlicast
Neuron Model:	Basic LIF	Programmable LIF	Programmable
Synapse Model:	Programmable	Binary	Programmable

- The STPU 3D mesh is enabled due to the temporal buffer
- The synapse model in the STPU is implemented via the temporal buffer



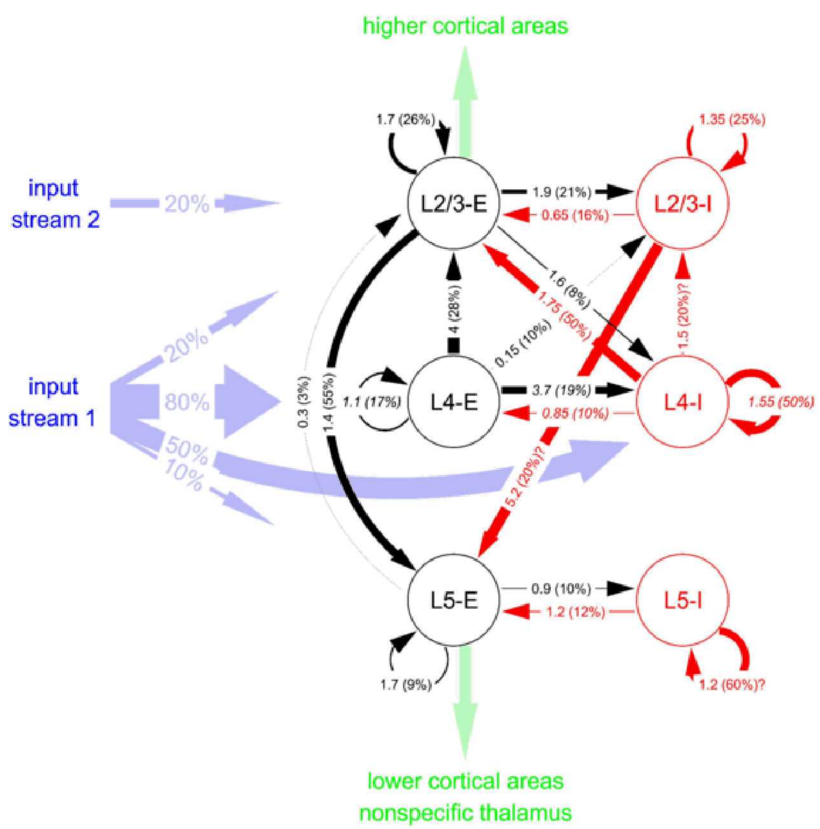


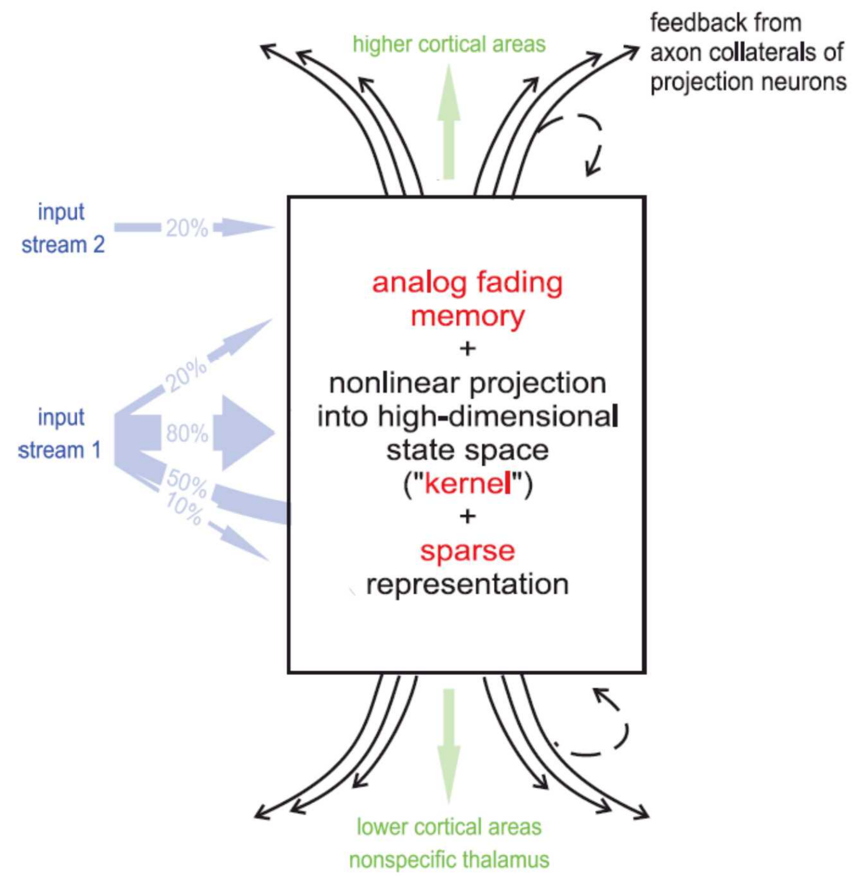


Original STPU Limitations

- No native on chip recurrence
  - Other use case was being devised
  - Assumed that all meaningful input could be provided as an input stream
  - Only 32 or 64 time steps are allowed
- All output is binary
  - Helps with power efficiency
- Limited to 8 bits
  - All numerical values are represented with 8 bits of precision

LSM motivation: Cortical Microcircuit

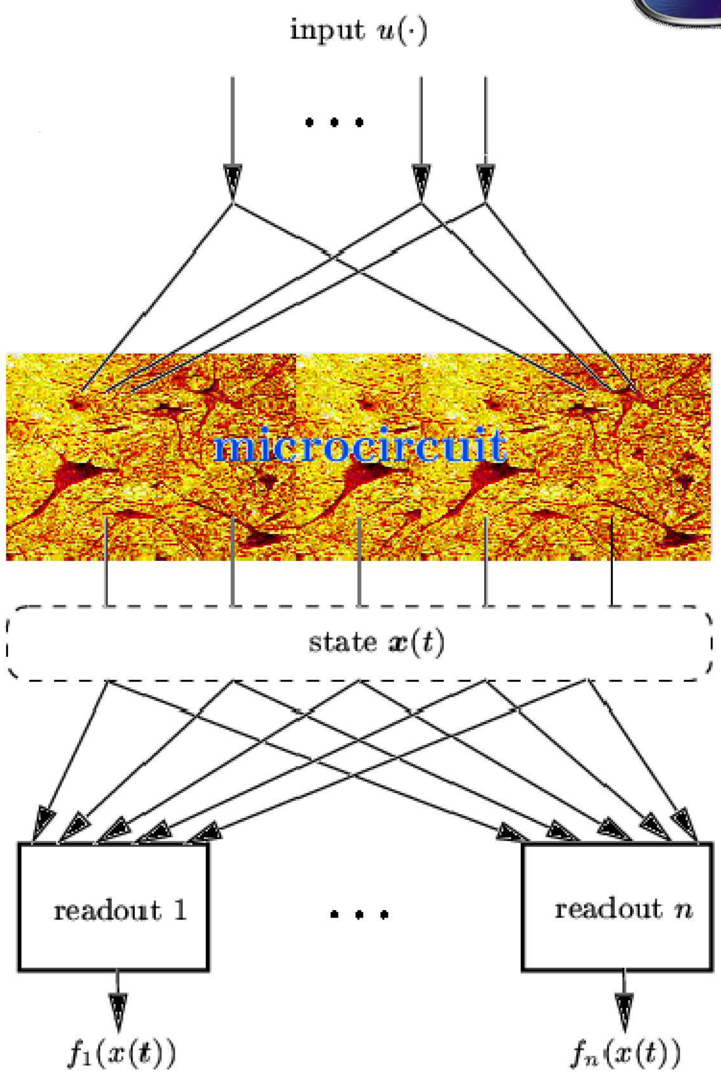




Maass, W., Liquid state machines: motivation, theory, and applications. In: Computability in context: computation and logic in the real world (Cooper B, Sorbi A, eds), pp 275–296 (2010)

Liquid State Machine

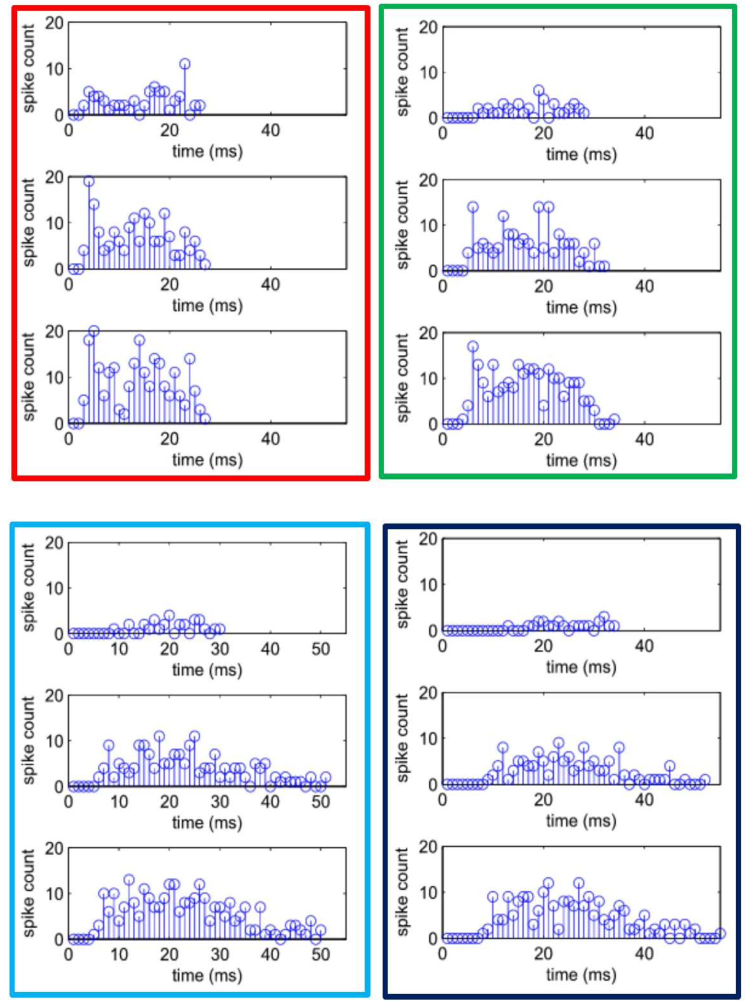
- Input (spike trains)
  - Maps input streams to output streams
- Liquid (or microcircuit)
  - A recurrent neural network of spiking neurons (leaky integrate and fire)
  - Randomly connected 80/20
  - Acts a preprocessor (temporal)
- State
  - Measure the state of the liquid at any given time t
- Readout neurons
  - Plastic synapses
  - By assumption, has no temporal integration capability of its own

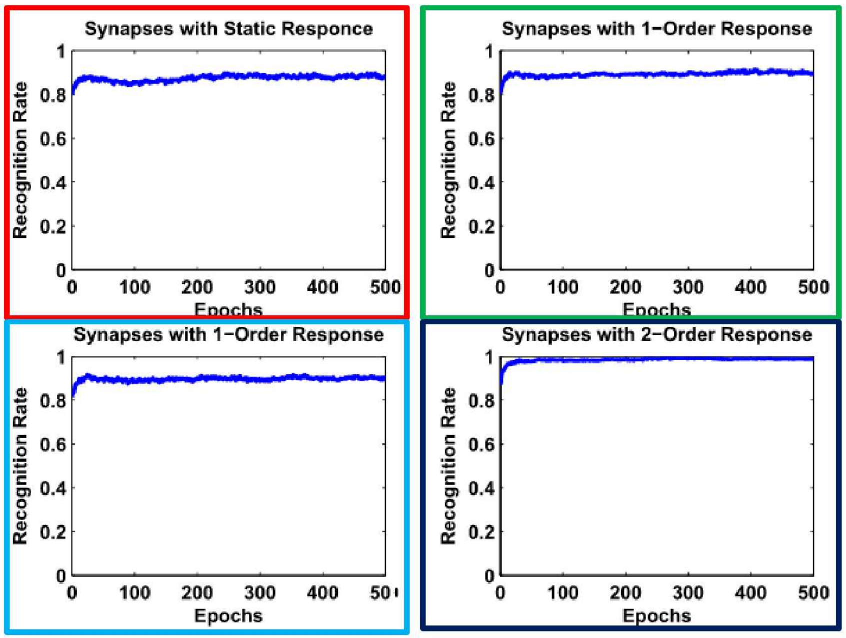


Natschläger, T., "The Liquid State Machine Framework." *Neural Micro Circuits,* http://www.lsm.tugraz.at/learning/framework.html. Accessed 26 September 2016

Living on the Edge of Chaos

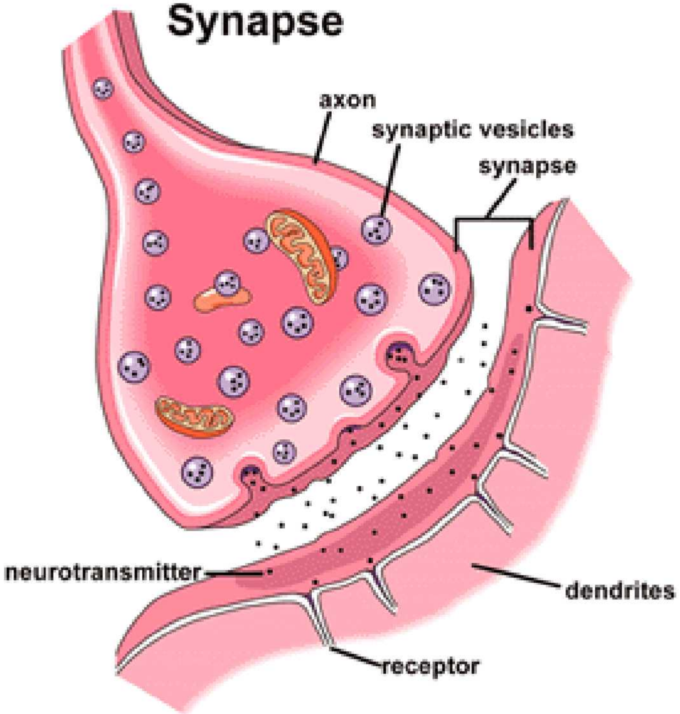
- Fading memory
  - Feedback loops and synaptic properties
  - Do not want to evolve to a steady state





Zhang, Y., Li, P., Jin, Y. & Choe, Y. (2015). A Digital Liquid State Machine With Biologically Inspired Learning and Its Application to Speech Recognition.. *IEEE Trans. Neural Netw. Learning Syst.*, 26, 2635-2649.

Spiking Temporal Processing Unit



- Signals along axon are digital (binary spikes)
- Synapses are analog and have exponential behavior
- Static

$$\delta(t - t_{ij} - d_j)$$

- First-order response

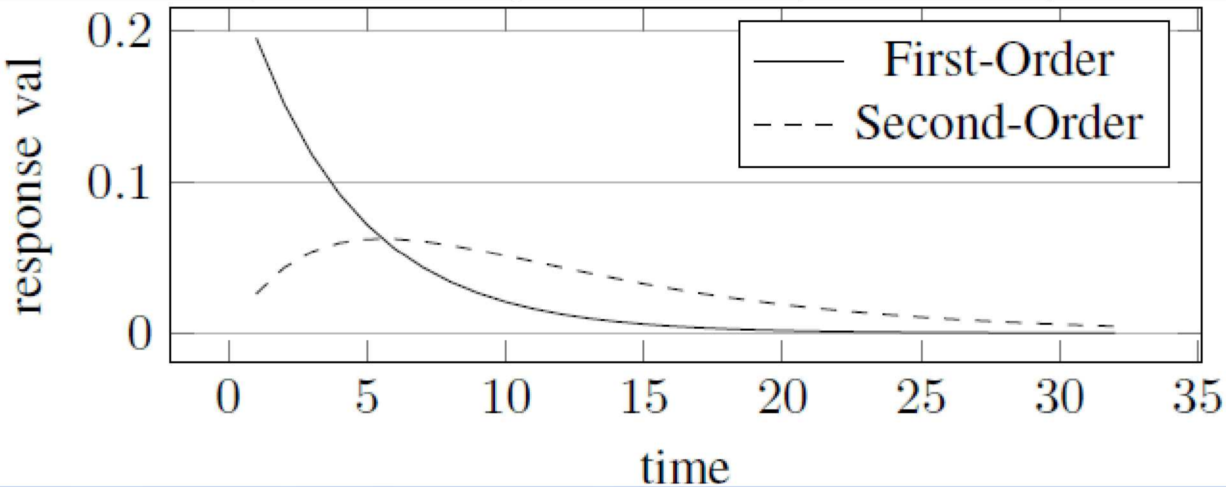
$$\frac{1}{\tau_m} e^{-\frac{t - t_{ij} - d_j}{\tau^s}} \cdot H(t - t_{ij} - d_j)$$

- Second-order response

$$\frac{1}{\tau_1^s - \tau_2^s} (e^{-\frac{t - t_{ij} - d_j}{\tau_1^s}} - e^{-\frac{t - t_{ij} - d_j}{\tau_2^s}}) \cdot H(t - t_{ij} - d_j)$$

Effects of Synaptic Response Function

Synaptic Response	Train Sep	Train Rate	Test Sep	Test Rate	SVM
Dirac Delta	0.129	0.931	0.139	0.931	0.650
First-Order	0.251	0.845	0.277	0.845	0.797
Second-Order	**0.263**	**0.261**	**0.290**	**0.255**	**0.868**
First-Order 30	**0.352**	0.689	**0.389**	0.688	0.811
First-Order 40	0.293	0.314	0.337	0.314	0.817
First-Order 50	0.129	0.138	0.134	0.138	0.725



Red indicates the best values for default parameters

Blue indicates values that improved over second-order

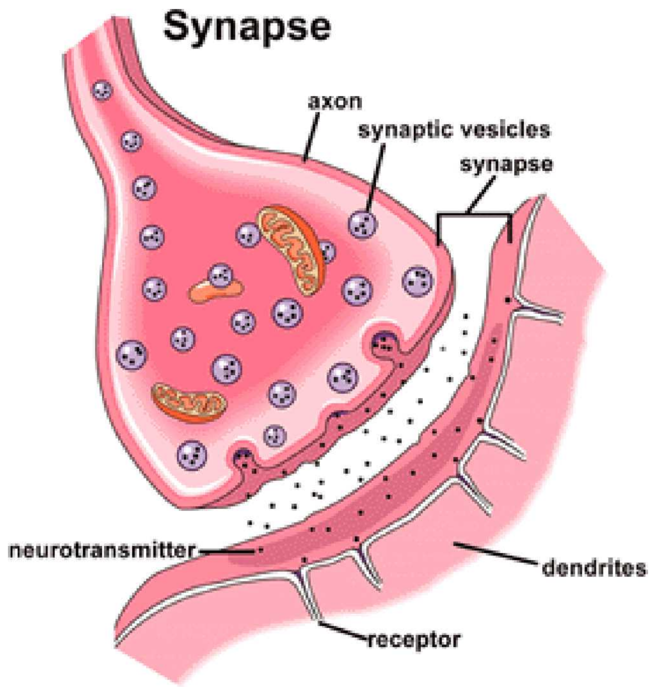
Requirements for the LSM

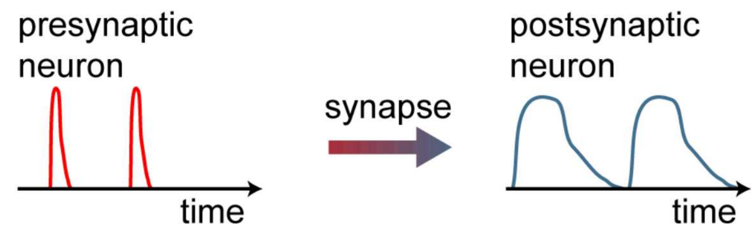
- Recurrence
- Analog to digital conversion
  - Synaptic Response Function (exponential equations)
- Dynamic/Plastic Weights
  - Learn the readout function
- Unknown number of time steps in the input.

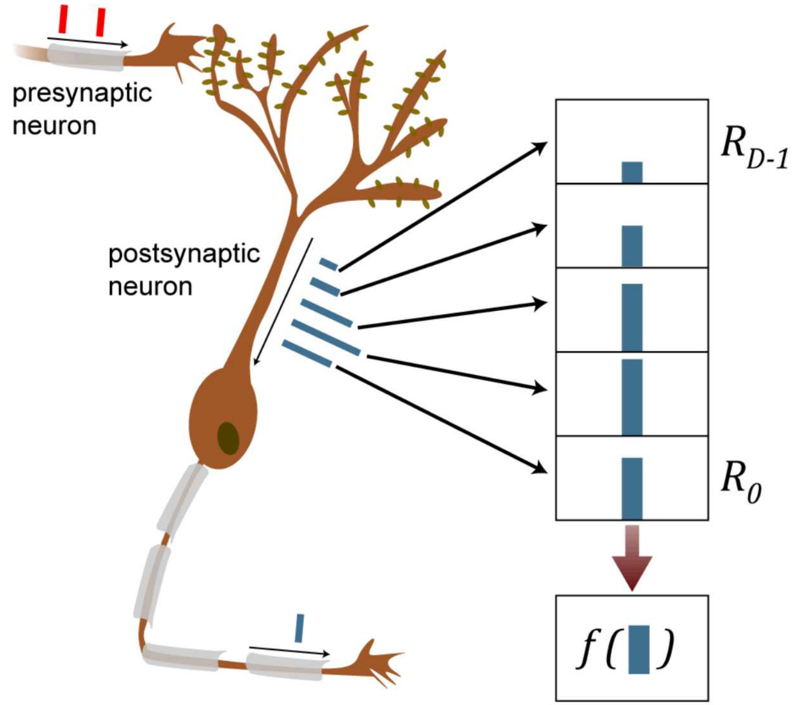
Co-design

- Used the requirements of the LSM to inform the design of the STPU V2
  - On chip recurrence
  - Arbitrary input lengths
  - Mechanism to update all weights
- Used features of STPU to keep everything digital
- Reduced all numerical values to 8 bit representation

Synaptic Response Functions







First-order response

$$\frac{1}{\tau_m} e^{-\frac{t-t_{ij}-d_i}{\tau^S}} \cdot H(t - t_{kd} - \Delta_{kd})$$

Second-order response

$$\frac{1}{\tau_1^S - \tau_2^S}(e^{-\frac{t-t_{kd}-\Delta_{kd}}{\tau_1^S}} - e^{-\frac{t-t_{kd}-\Delta_{kd}}{\tau_2^S}}) \cdot H(t - t_{kd} - \Delta_{kd})$$

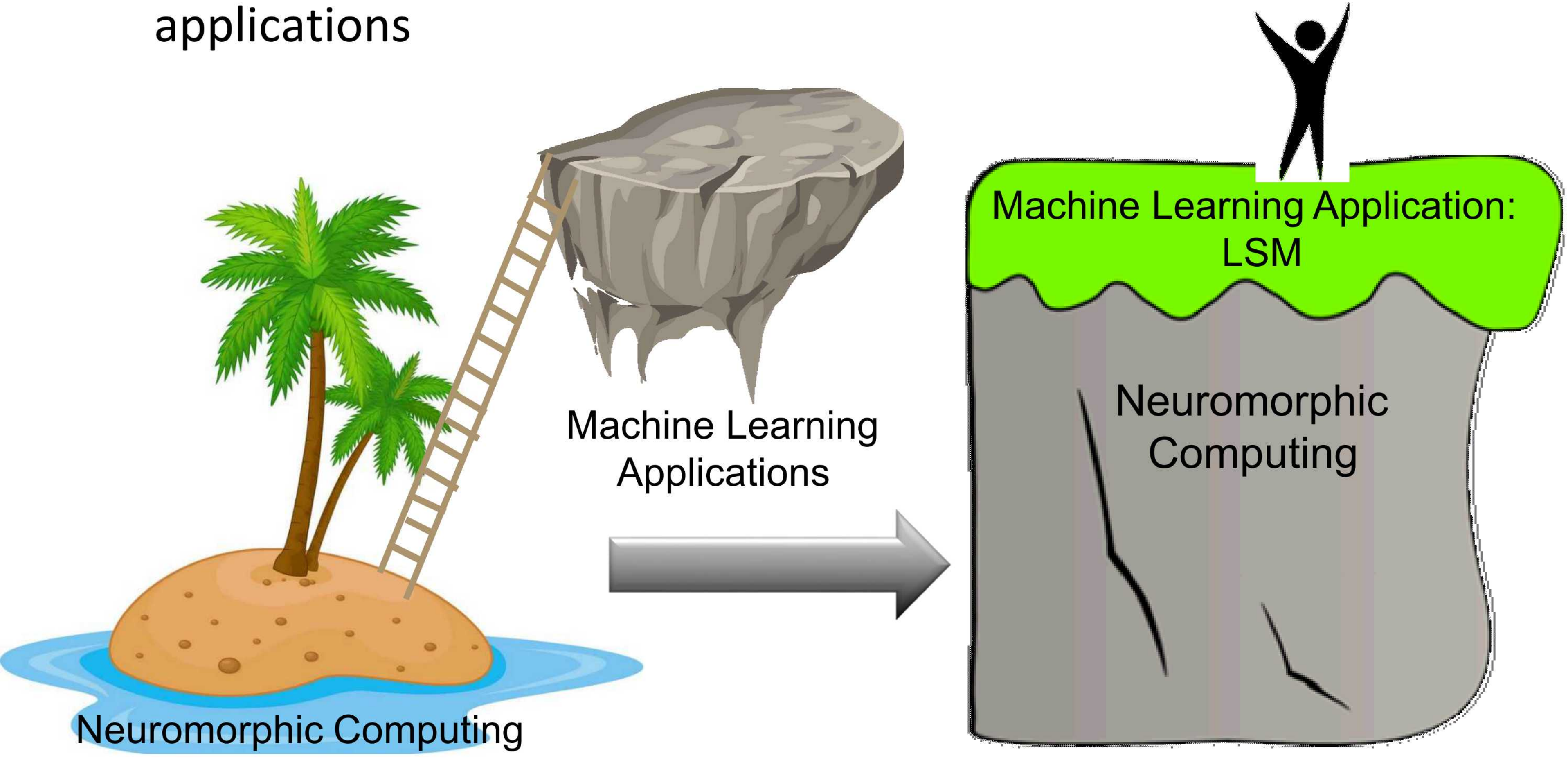
Bit Representation

- Did not have a significant effect on the classification accuracy

num bits	Train Sep	Train Sparse	Test Sep	Test Sparse
64	0.211	0.290	0.192	0.290
32	0.211	0.290	0.192	0.290
20	0.211	0.290	0.192	0.290
20	0.211	0.290	0.192	0.290
16	0.210	0.290	0.193	0.290
10	0.208	0.287	0.202	0.289
8	0.198	0.276	0.213	0.279
4	0.177	0.220	0.196	0.215
2	0.153	0.109	0.113	0.111

Neuromorphic Foundation

- Go from two separate efforts to being more unified
- Neuromorphic principles provide a foundation to higher-level applications



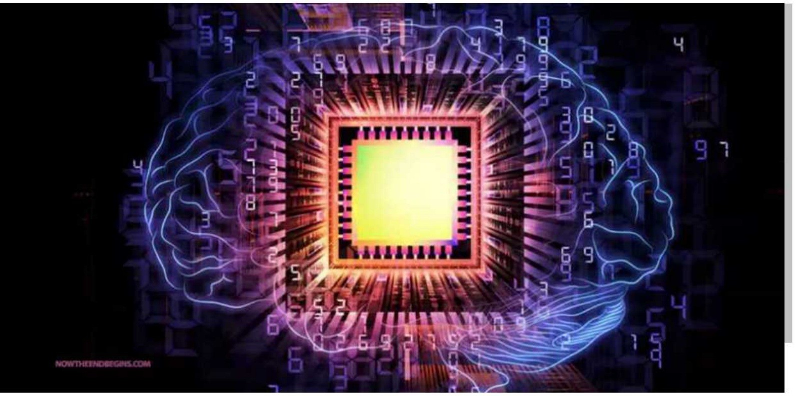
Lessons Learned

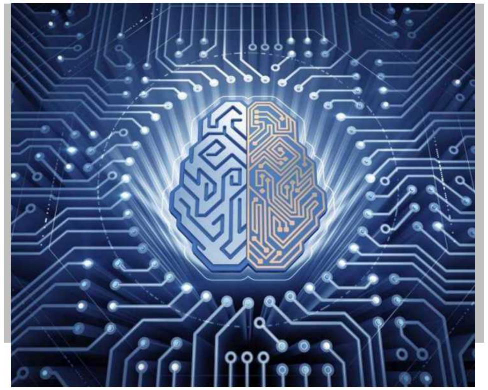
- Implemented LSM on the STPU
  - Co-design is essential—Need a unified approach including neuromorphic principles (bottom-up) and a real application or algorithm (top-down)
  - A diverse team is key—neuroscientists, computer scientists, hardware design, hardware architecture, machine learning, mathematicians, practitioners, etc.
  - Compromise is needed
    - LSM needed recurrence—was added in V2
    - STPU was limited pure digital computations—devised a way to do analog in digital processing
- Moving Forward
  - LSM is a simple algorithm—started looking at deep networks
  - How to using spiking neurons in place of sigmoid or ReLU neurons

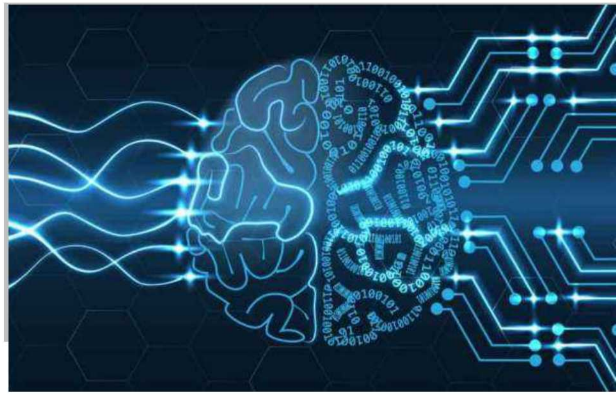
Bright Future for Neuromorphic

Need to find a good application/algorithm and design toward it.

- Gain more visibility for the achievement
- Integrate neuroscience and neuromorphic principles into the algorithm
- Make the hardware more amenable to being used
- Highlight short-comings in the hardware and/or algorithm







Thank You

Questions?

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.SAND NO. 2011-XXXXP