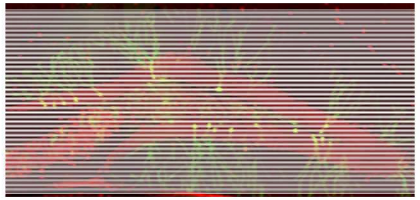
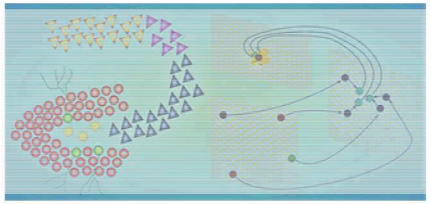
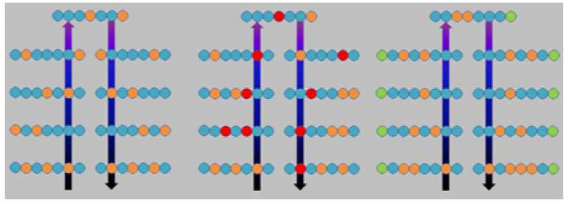
SAND2020-11105C

Preparing for the Next Generation of Brain-Inspired AI







PRESENTED BY

Brad Aimone; jbaimon@sandia.gov

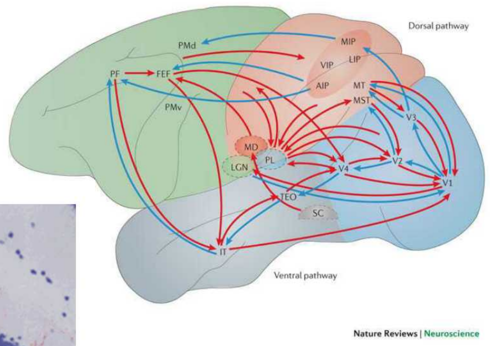
2020 ValleyML

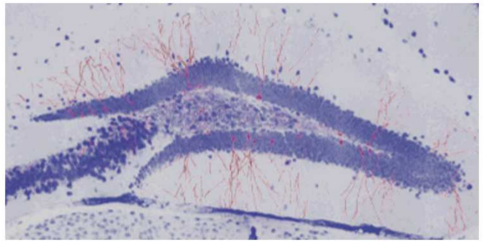
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

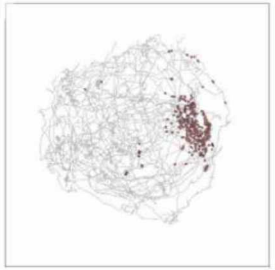
Brain-Inspired Computing Proposition

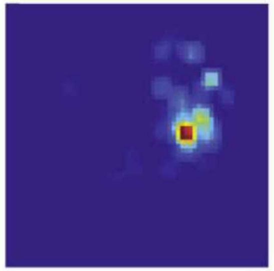
Leveraging knowledge of how the brain processes information can impact a wide range of science and technology applications

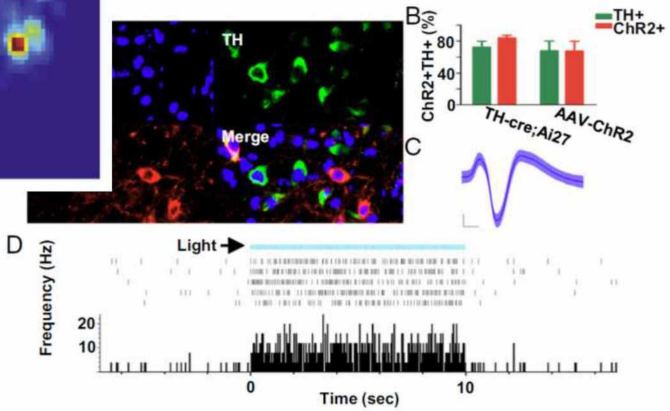
Leveraging knowledge of how the brain processes information can impact a wide range of science and technology applications

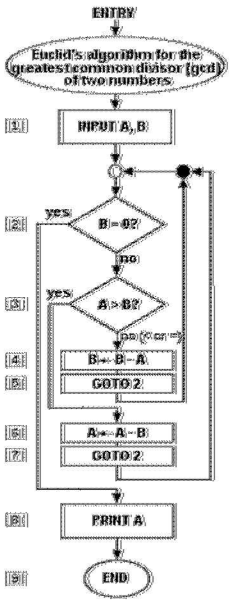


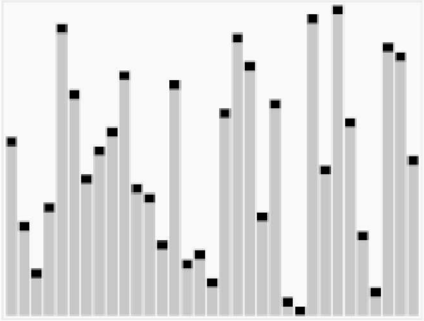


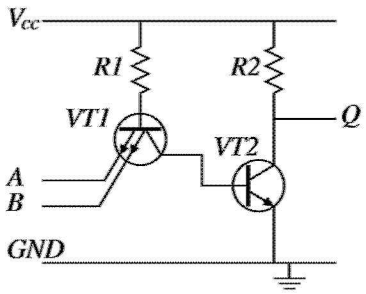


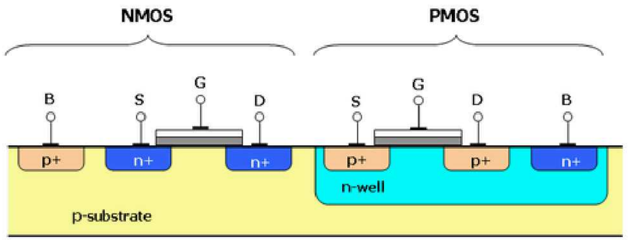




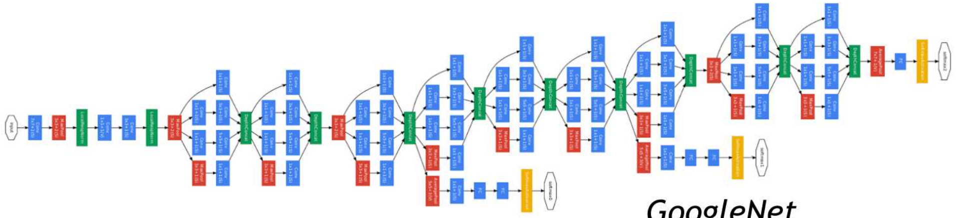






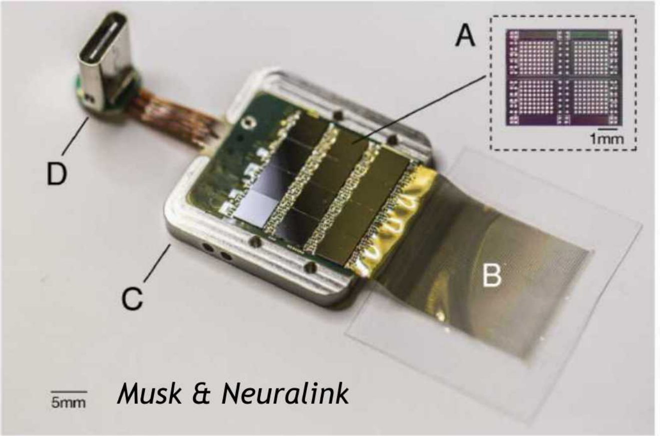


Leveraging knowledge of how the brain processes information can impact a wide range of science and technology applications

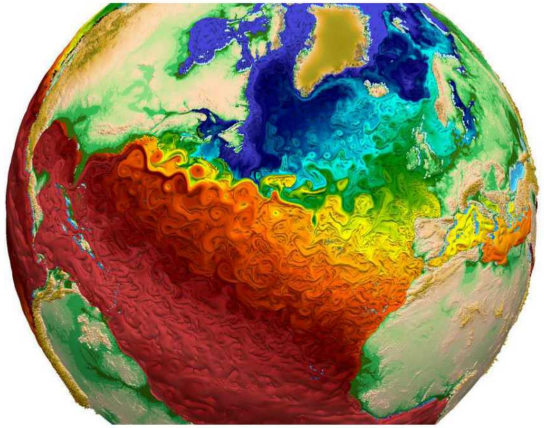


GoogleNet

Tesla Autopilot

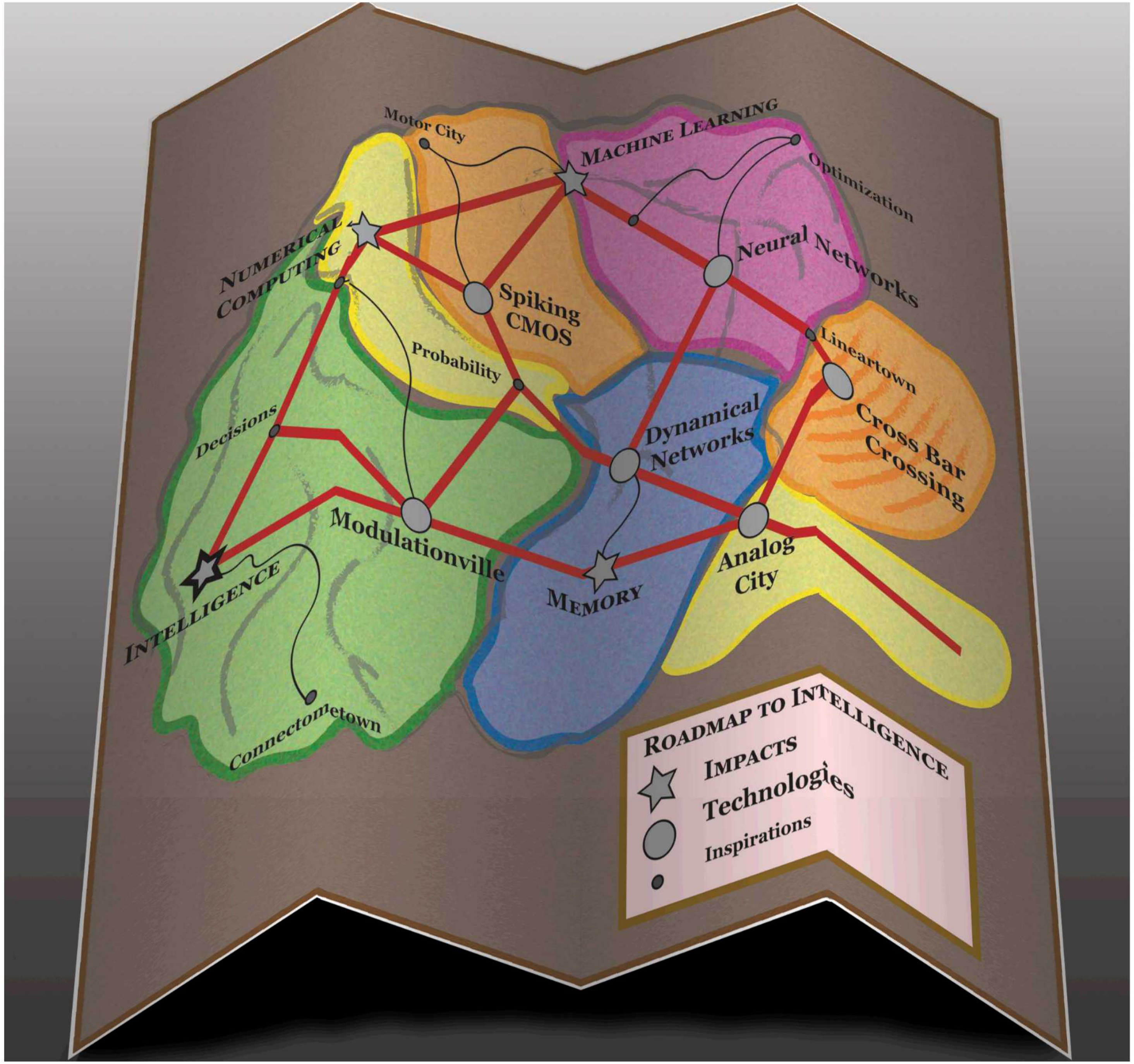


Musk & Neuralink



Large-scale modeling & simulation

Oak Ridge National Laboratory Summit

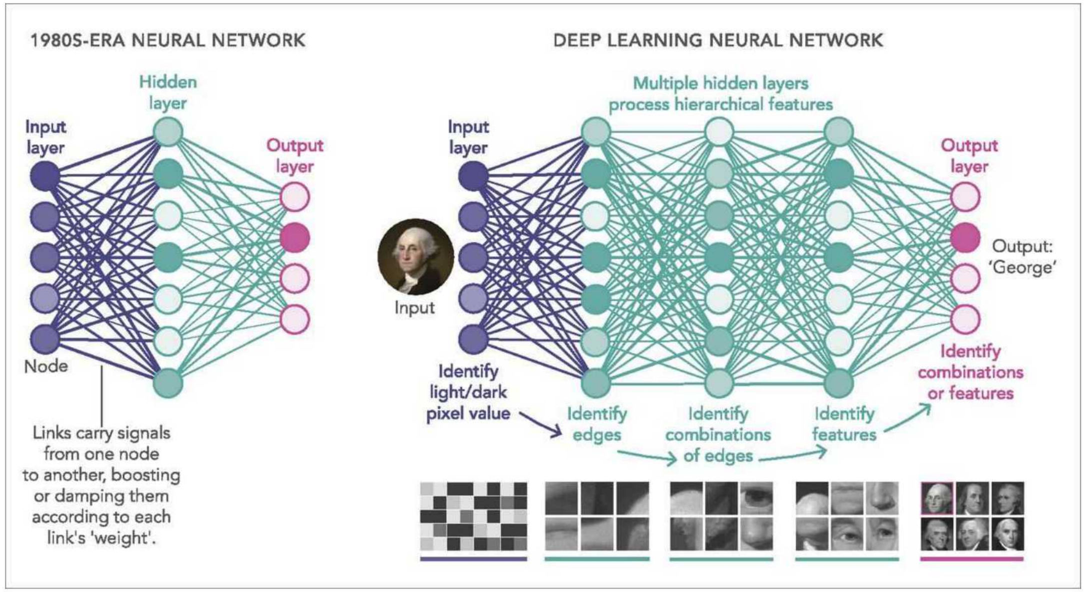

Motor City
Machine Learning
Optimization
Numerical Computing
Neural Networks
Spiking CMOS
Probability
Lineartown
Cross Bar Crossing
Decisions
Dynamical Networks
Modulationville
Analog City
Intelligence
Memory
Connectometown
Roadmap to Intelligence
Impacts
Technologies
Inspirations

The recent rise in AI has many causes

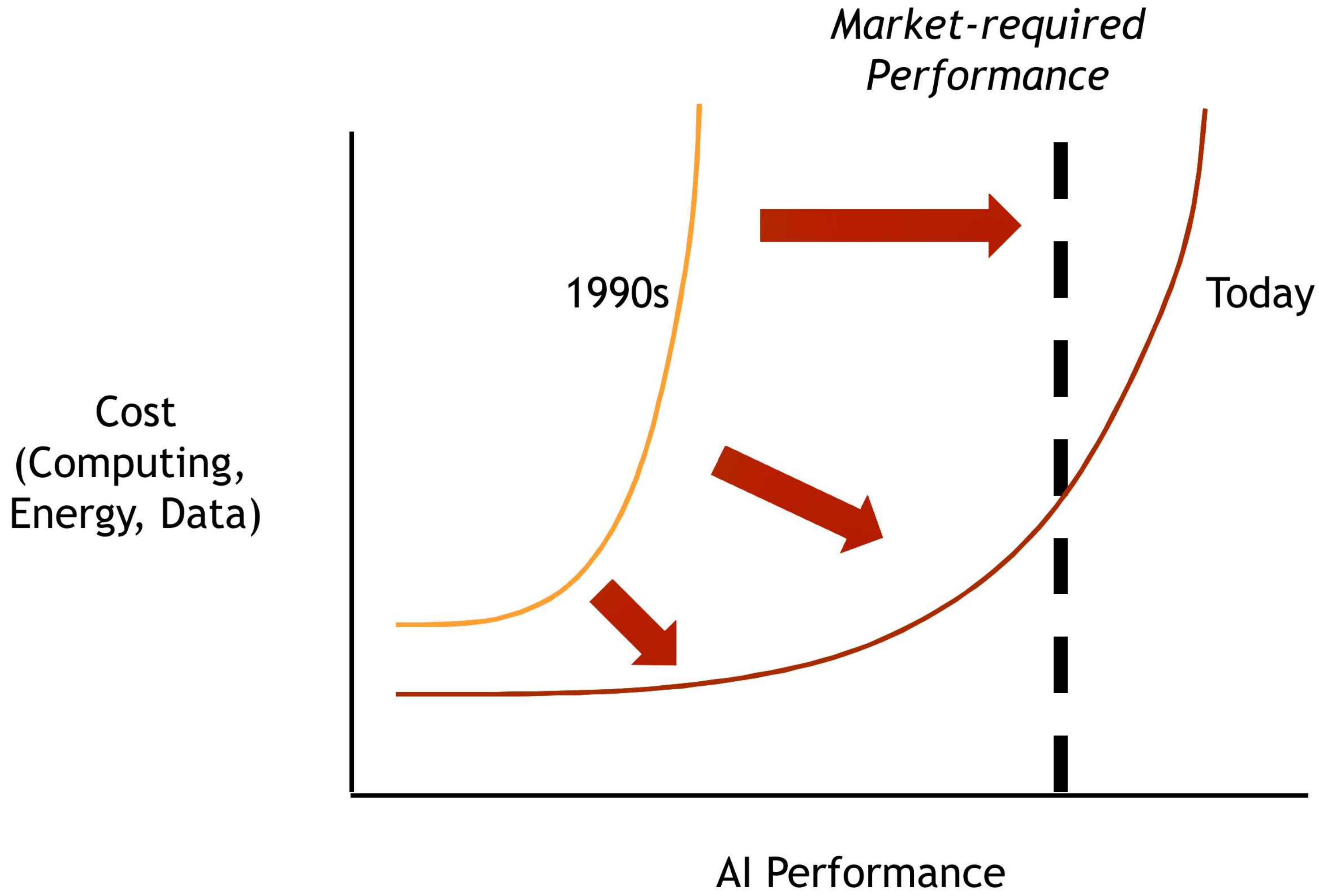
- Moore's Law! – There is always a bigger computer!
  - GPUs…

- The Internet! – Endless supply of unlimited data!
  - Social Media…

- Model-free Learning! – Deep networks can do anything!
  - Pre-training, drop-out, etc…



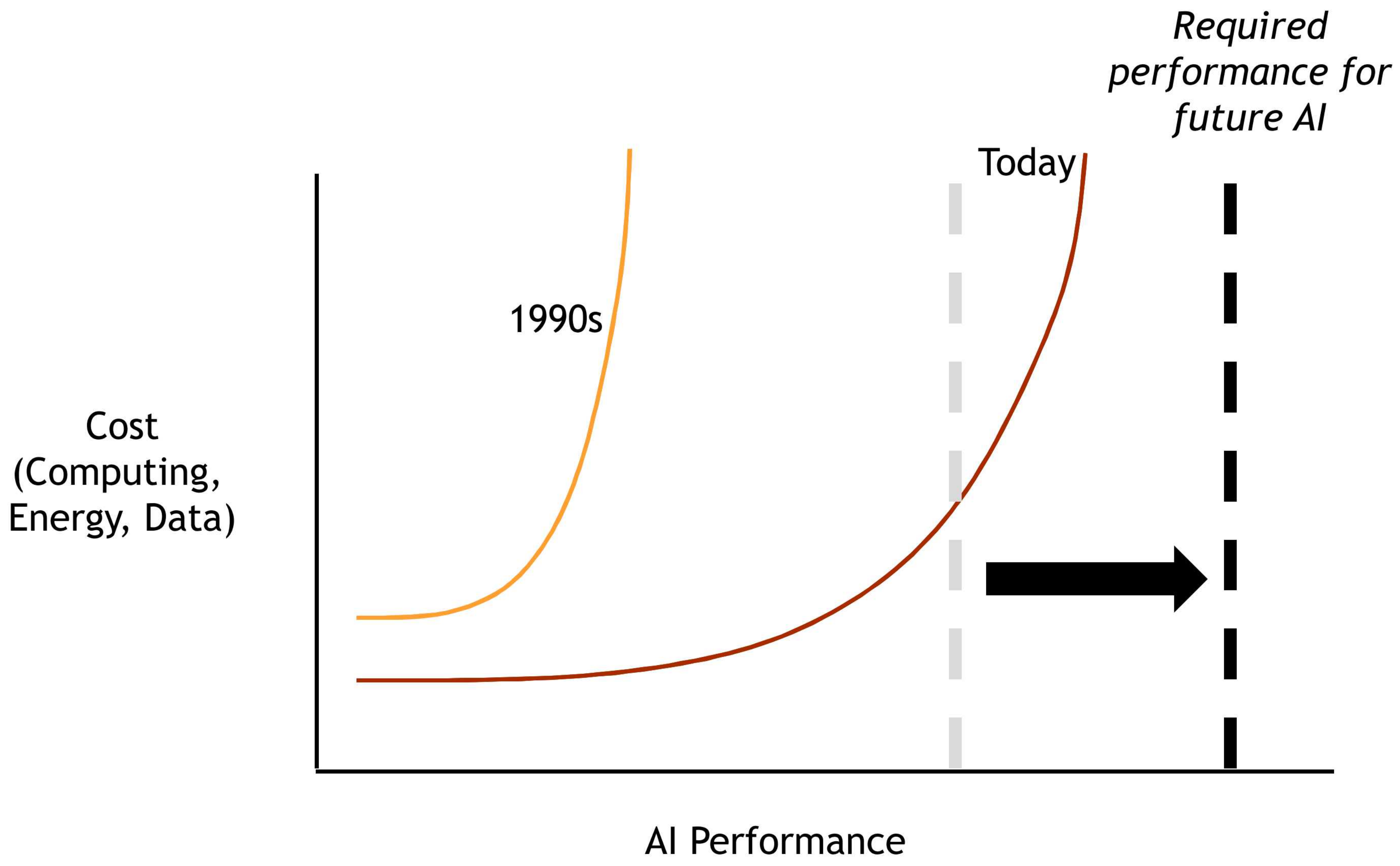
Waldrop PNAS 2019



Efficiency Drivers

- Cheaper computing
- Data, data, data
- Some new theory

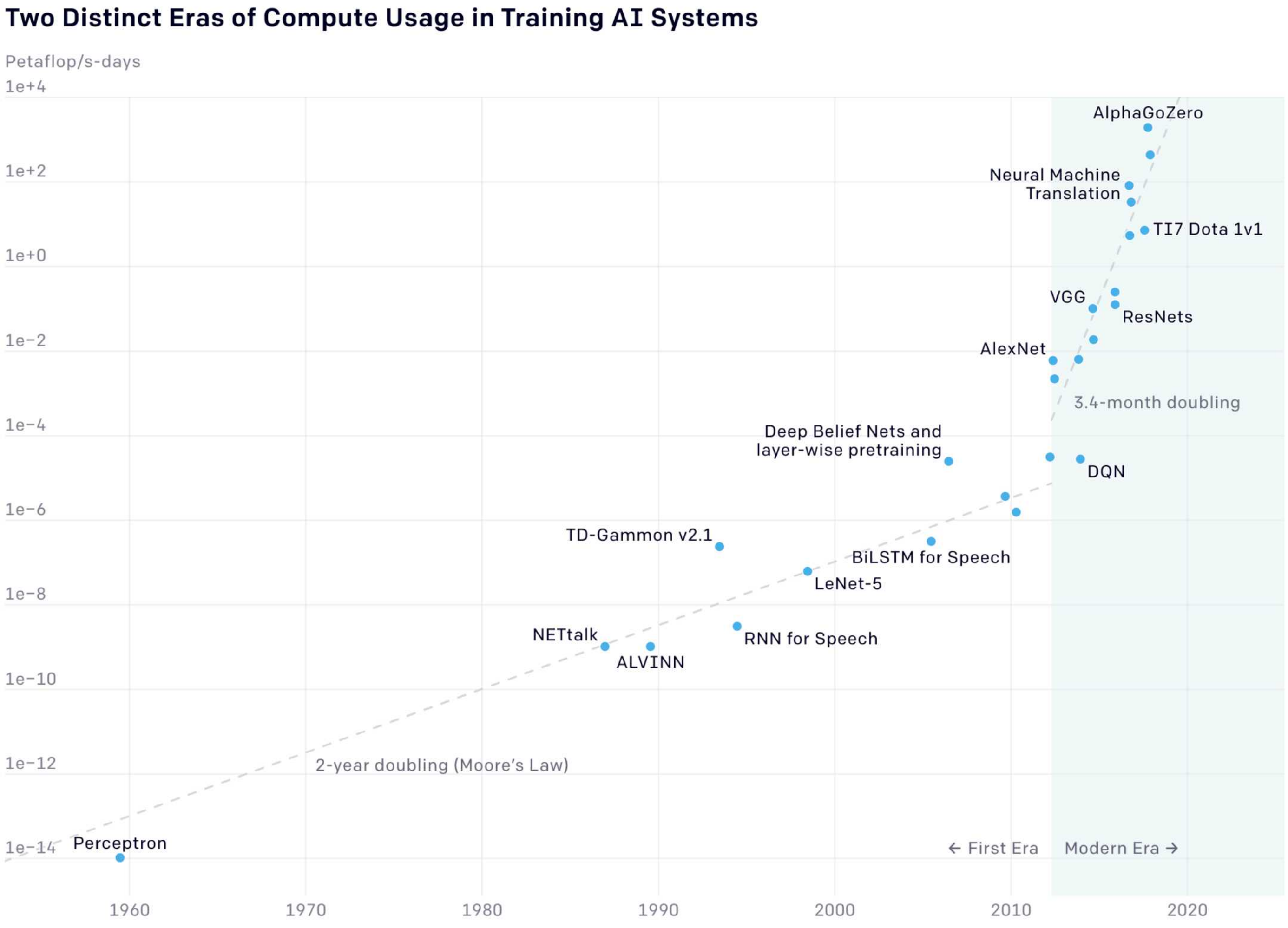
Extending AI to different applications requires further efficiency scaling



Future reality is not so rosy

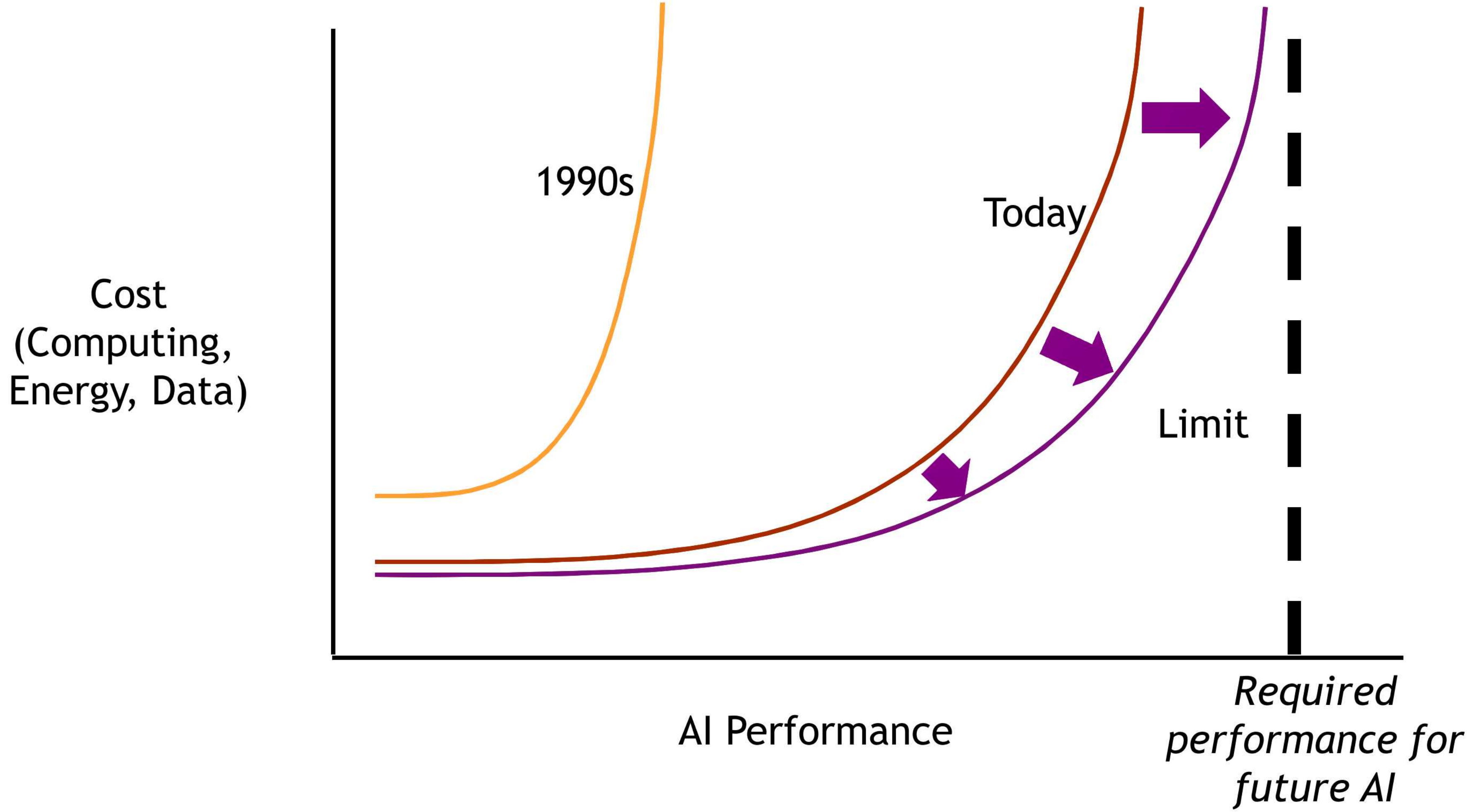
- Moore's Law! – There is always a bigger computer!
  - ***Dennard scaling is over, Moore's Law is slowing***
- The Internet! – Endless supply of unlimited data!
  - ***Data is not equally available, and not all data is AI-friendly***
- Model-free Learning! – Deep networks can do anything!
  - ***Theory and trust in algorithms remains poor, little physics in current algorithms***

Unending push towards bigger and bigger and bigger networks…



Open AI – *"AI and Compute"* May 2018

Slowing of Moore's Law limits computing scalability



High-performing AI algorithms often depend on a lot of data…

IMAGENET

SEARCH

14,197,122 images, 21841 synsets indexed

Home Explore
About Download

Not logged in. Login | Signup

About ImageNet

- Overview
- Research Team
- **Summary and Statistics**
- Citations and Publications
- Interesting Articles
- Join ImageNet Mailing List
- API Documentation
- Sponsors

Summary and Statistics (updated on April 30, 2010)

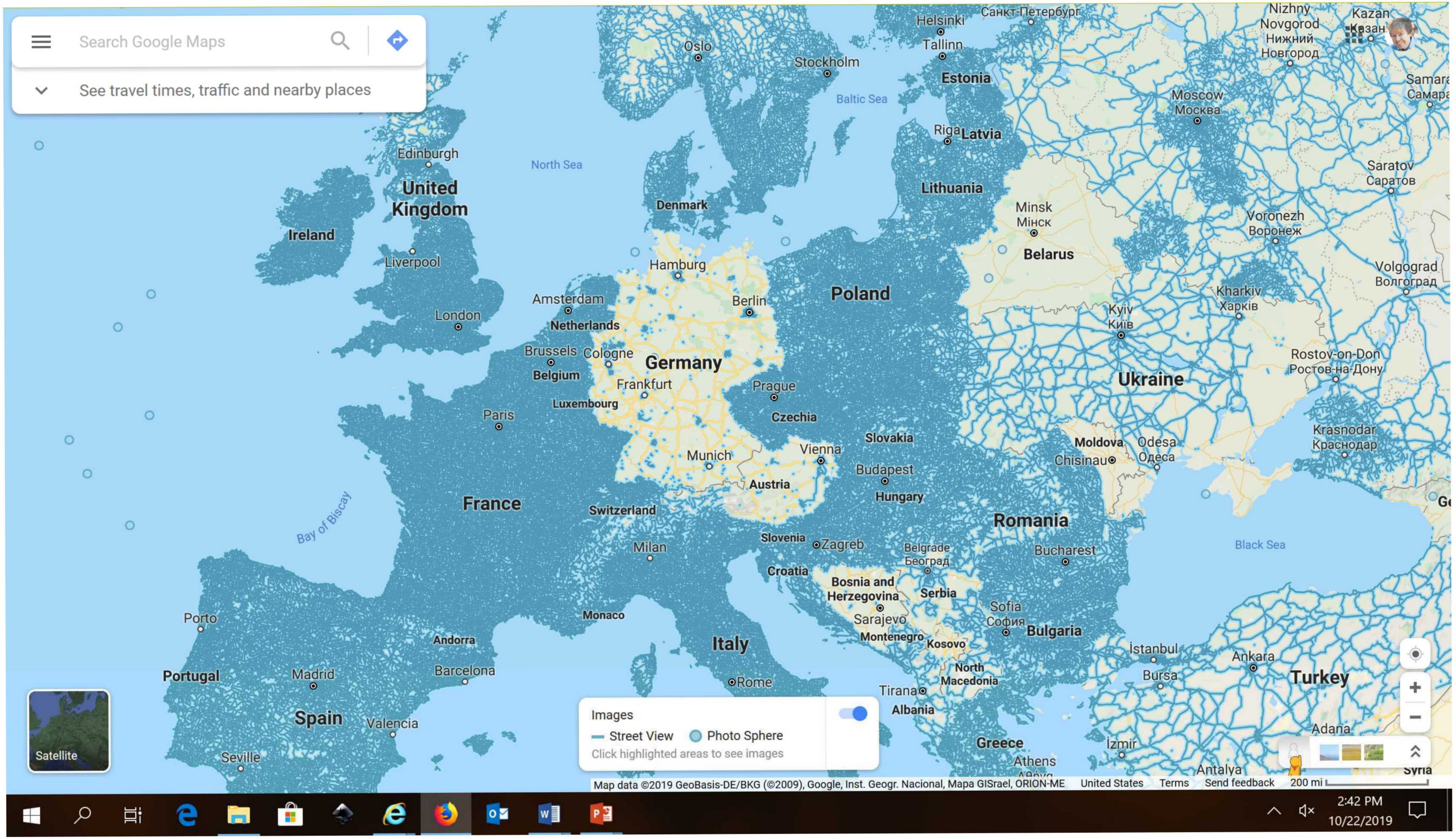
Overall

- Total number of non-empty synsets: 21841
- Total number of images: 14,197,122
- Number of images with bounding box annotations: 1,034,908
- Number of synsets with SIFT features: 1000
- Number of images with SIFT features: 1.2 million

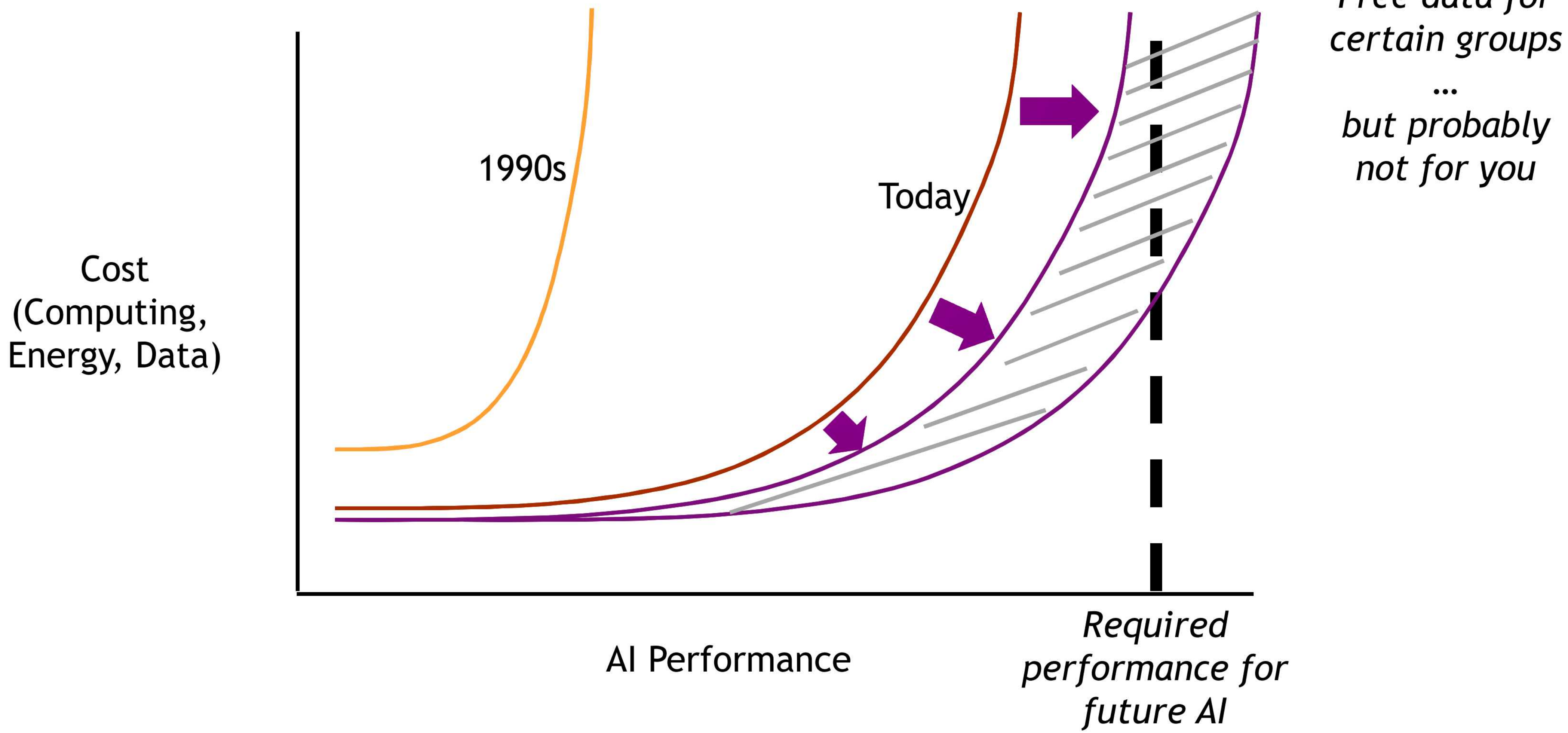
Statistics of high level categories

High level category	# synset (subcategories)	Avg # images per synset	Total # images
amphibian	94	591	56K
animal	3822	732	2799K
appliance	51	1164	59K
bird	856	949	812K
covering	946	819	774K
device	2385	675	1610K
fabric	262	690	181K
fish	566	494	280K
flower	462	735	339K

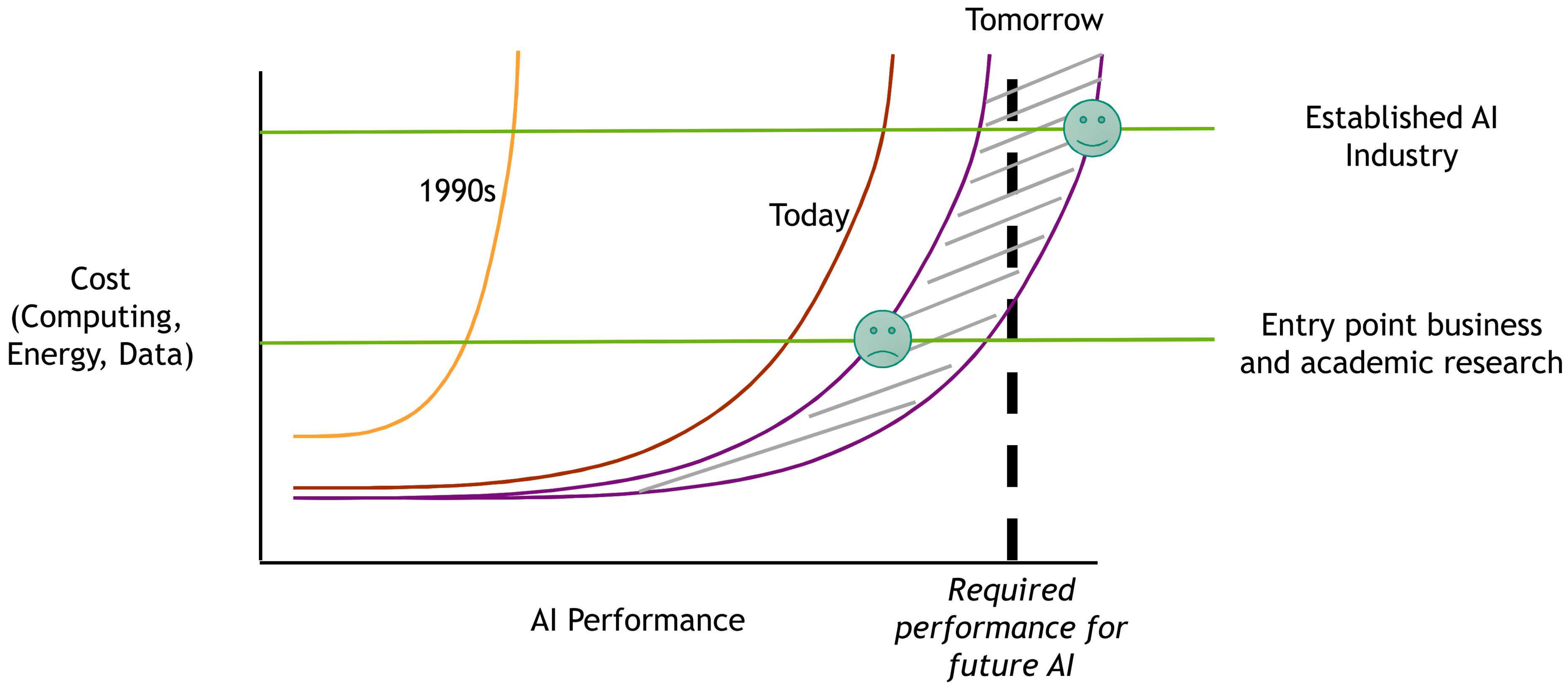
Good data is not uniformly available in all domains



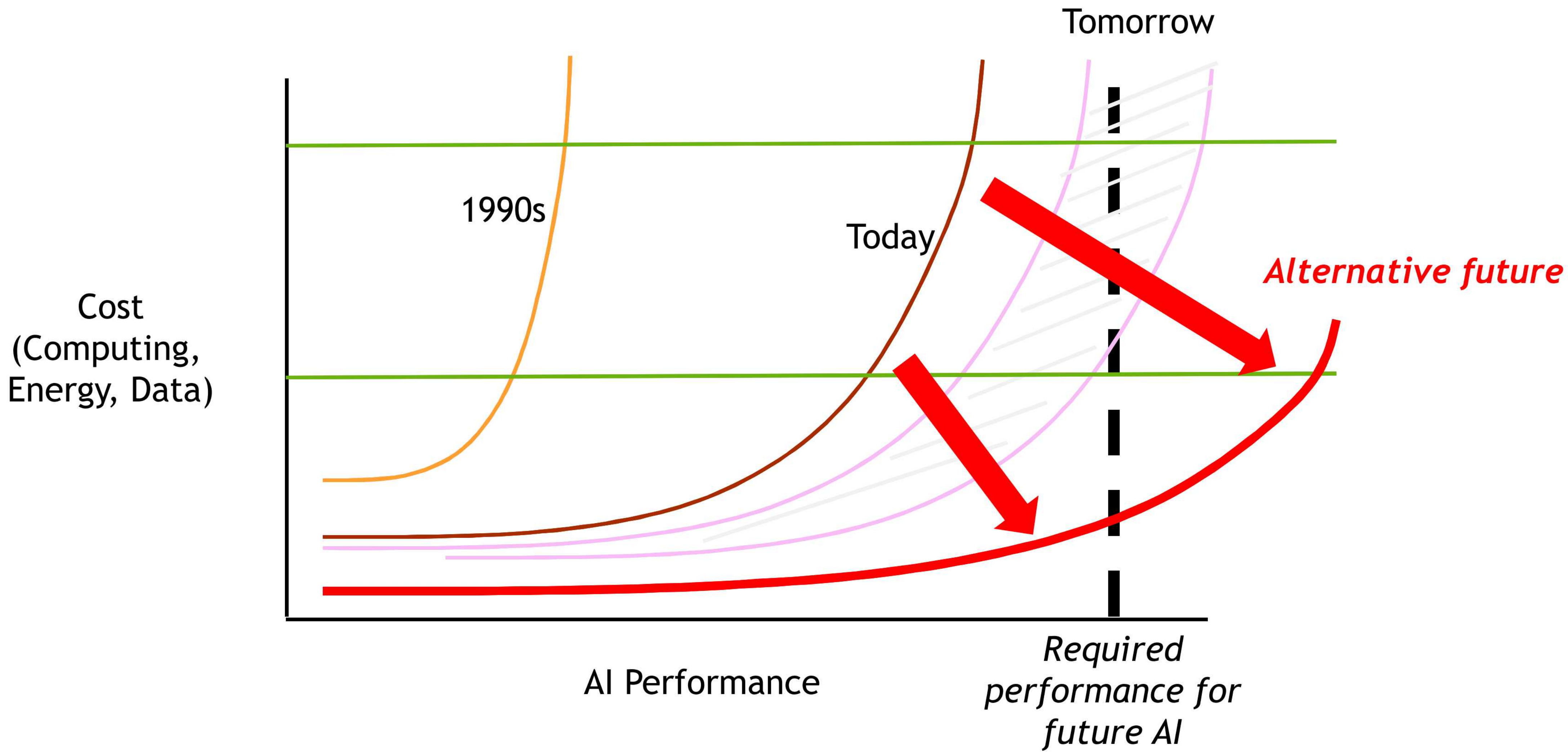
Future of data (privacy, cost, etc.) ensures unequal availability…

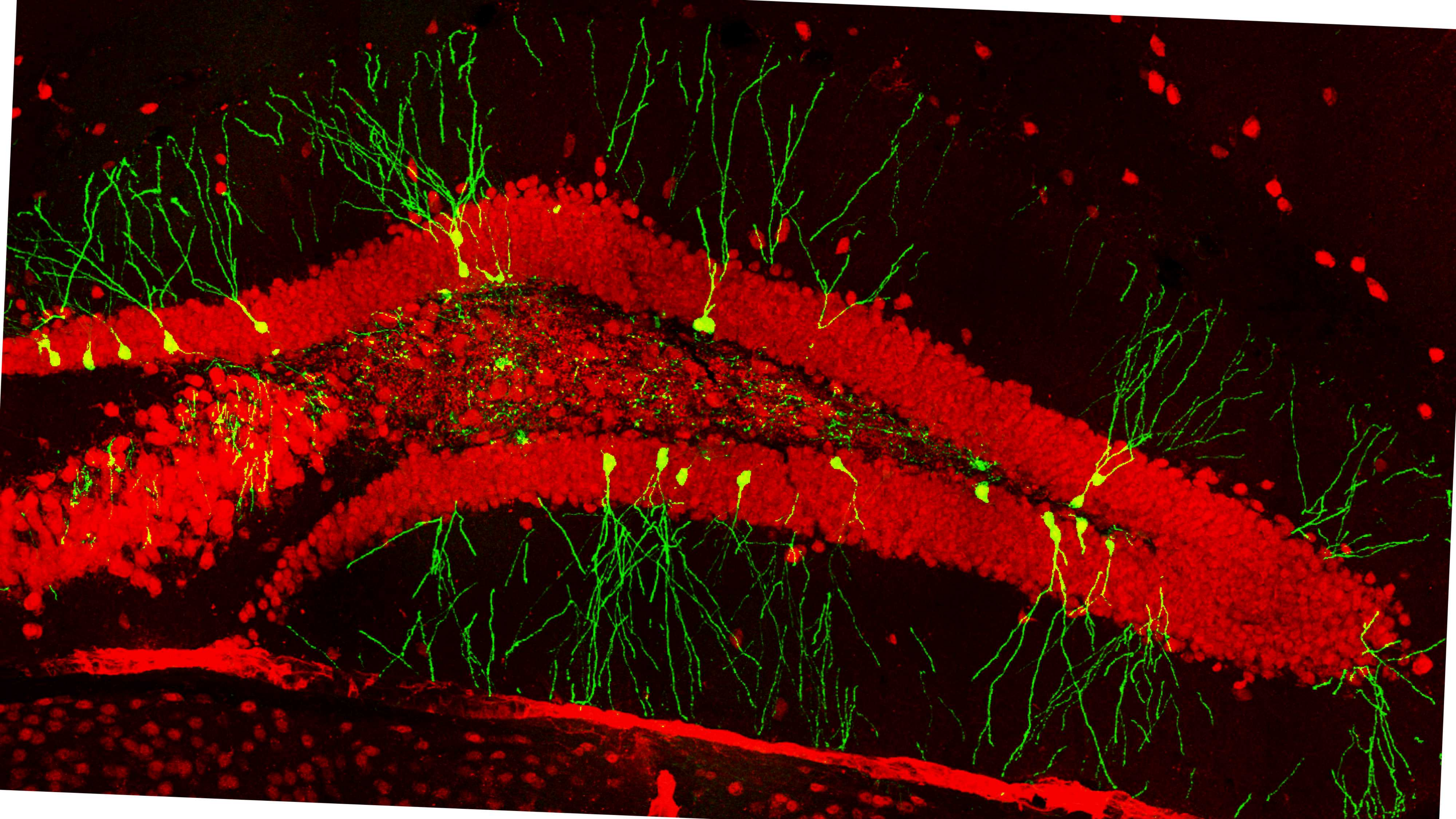


Data should be seen as a potential barrier to entry for AI

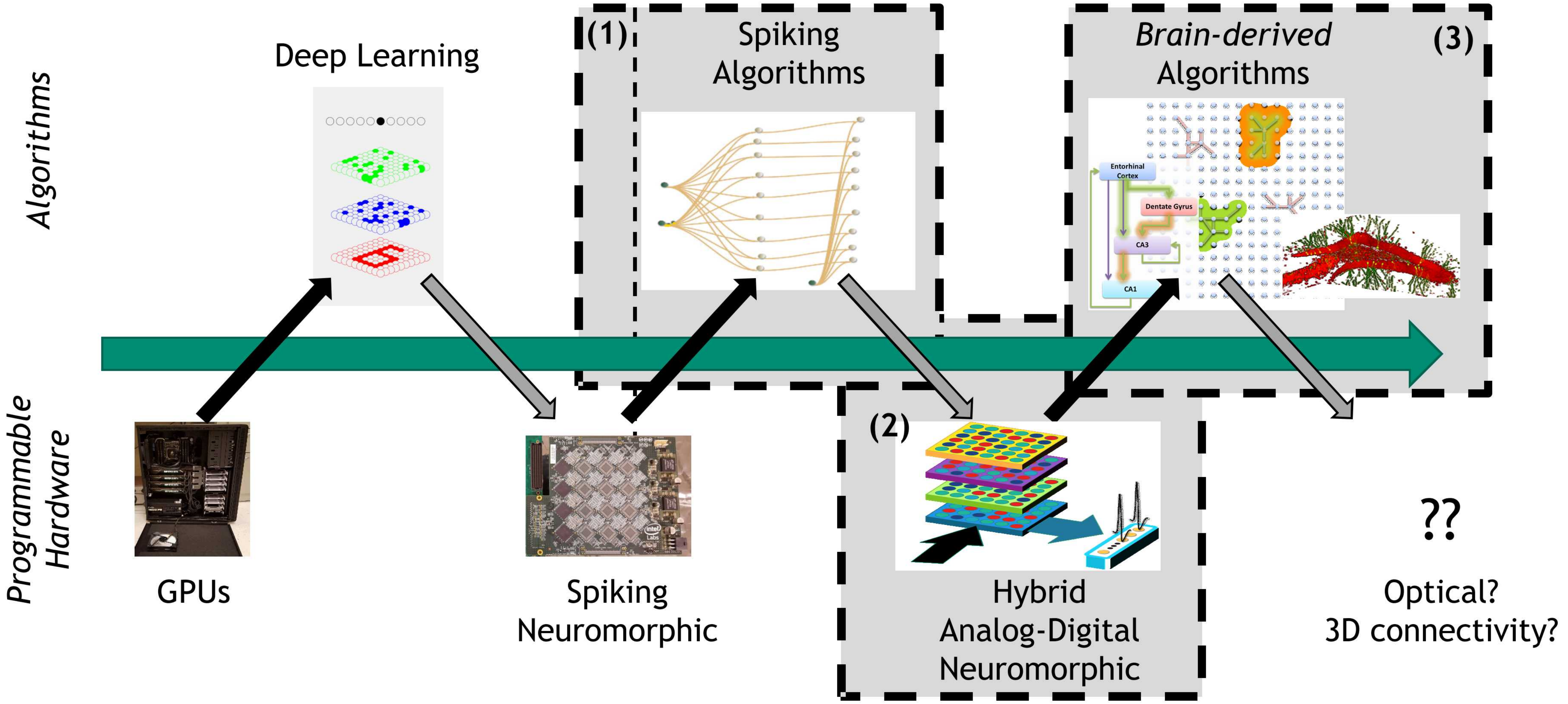


Can we envision an alternative AI future that is more scalable?



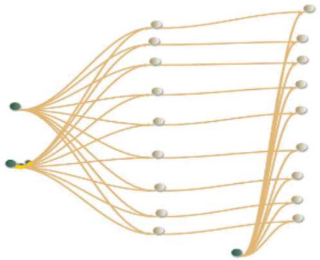


Neuromorphic computing is embarking on a co-design future

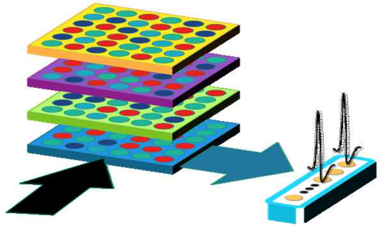


A roadmap for neuromorphic computing

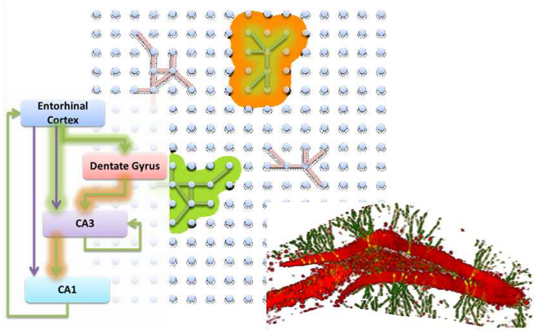
- *Today*: High-density spiking CMOS chips
  - Is spiking deep learning realistic?
  - Can these chips do anything beyond deep learning?



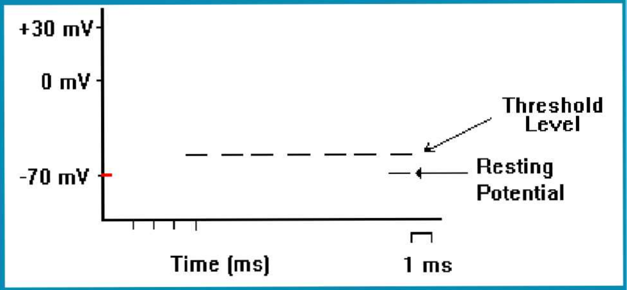
- *Tomorrow*: Hybrid analog-spiking processors as part of heterogeneous architecture
  - Is energy-savings enough to justify a loss in precision?
  - Can I create an efficient neural memory algorithm?



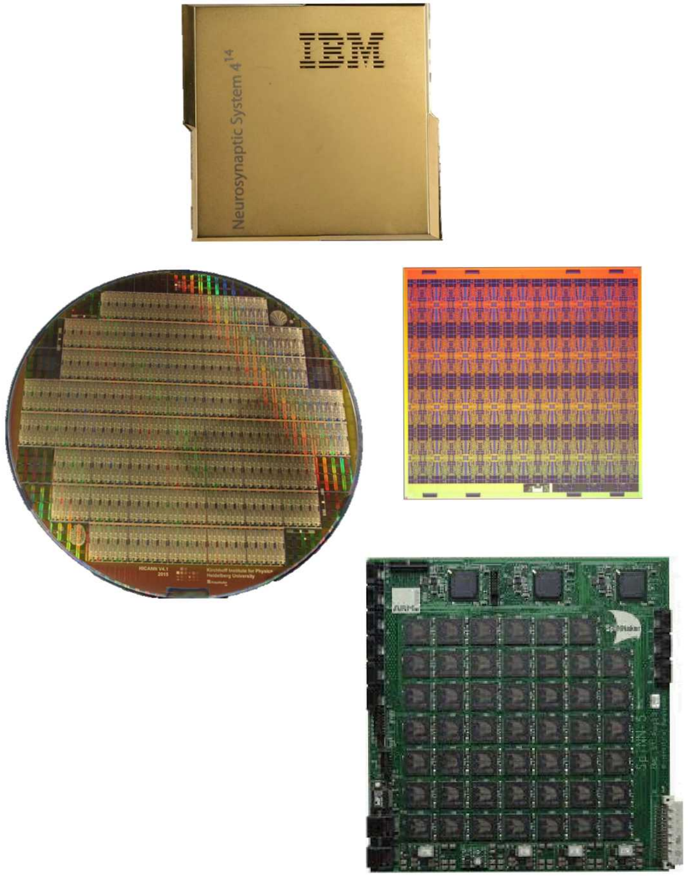
- *Future*: Brain-derived algorithms and hardware
  - What is the path to a data-efficient brain-inspired AI method?
  - Is current hardware path sufficient? Or do we need something radically different?



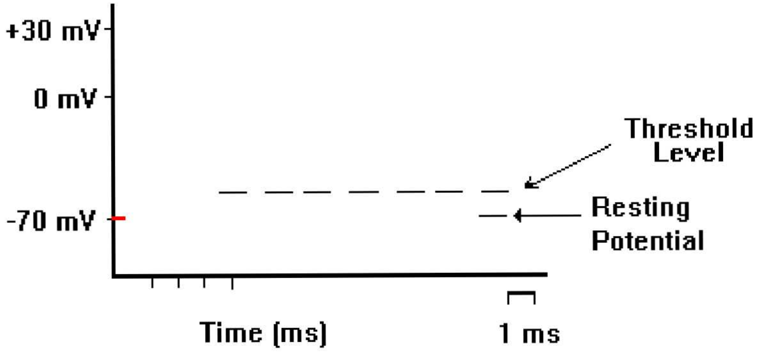
Part 1:
Can spiking actually be useful for computing?



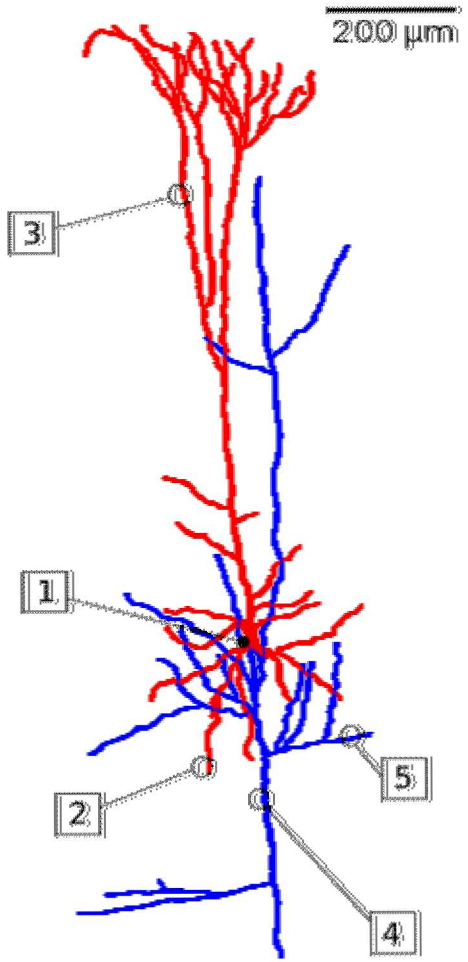
The hardware industry is pushing towards *spiking* chips



Clockwise from top: IBM TrueNorth, Intel Loihi, SpiNNaker, BrainScales



https://faculty.washington.edu/chudler/ap.html



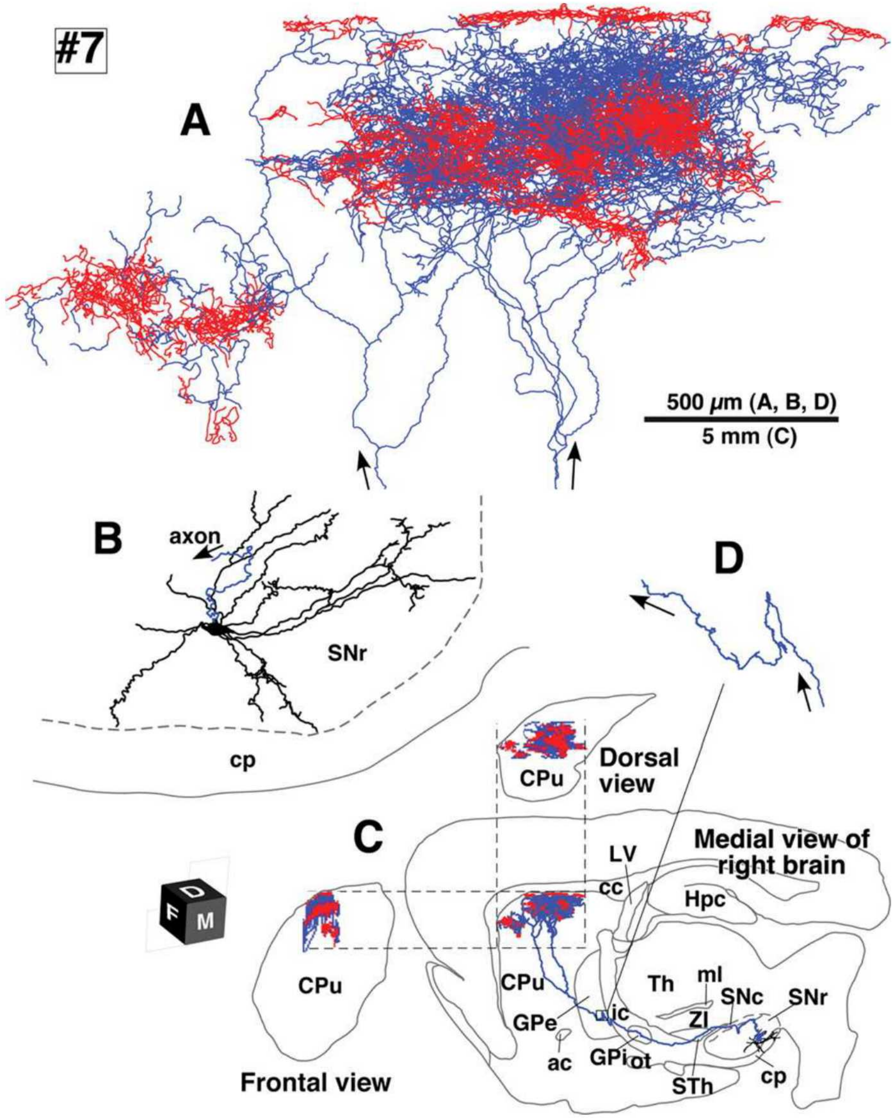
Pyramidal Cell -- Wikipedia

Why spiking?

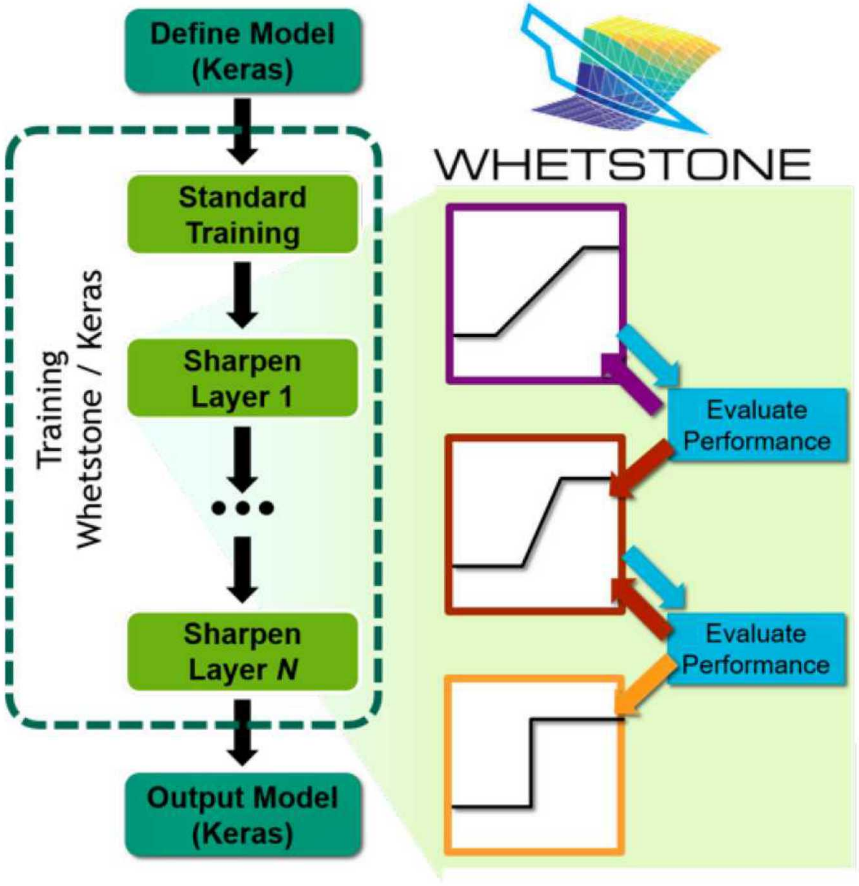
- Event-driven
  - Only expend energy when neuron crosses threshold

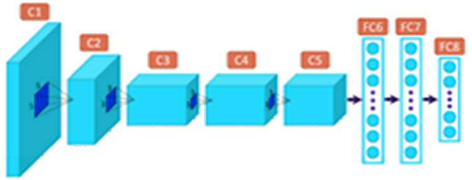
- Reliable and efficient over long distances
  - Neurons often project across brain or whole body…

- Robust to noise
  - Away from threshold, biophysical noise should not accidently cause spikes



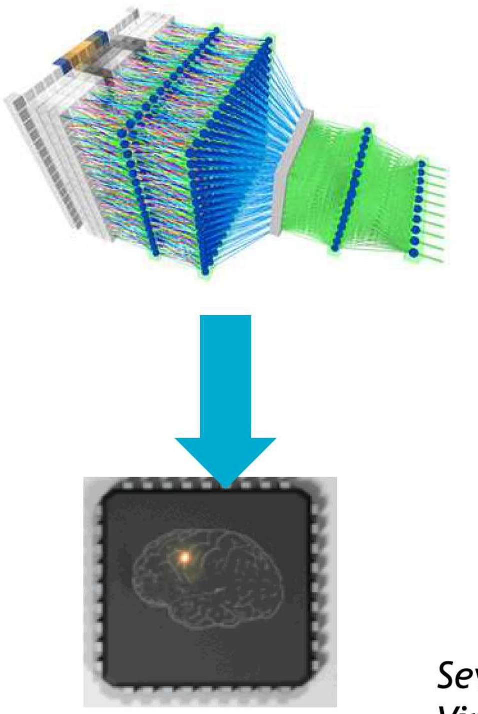
What can you do with spiking neurons?

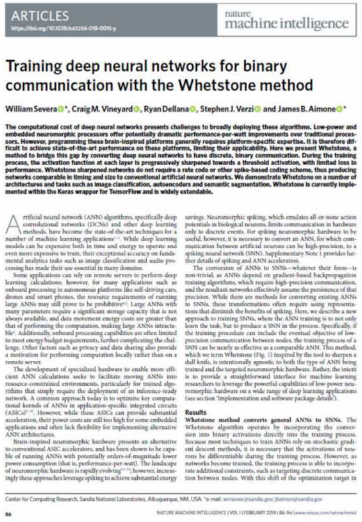




Spiking deep neural networks

- Whetstone allows us to use spiking communication with *no time penalty* and minimal accuracy reduction



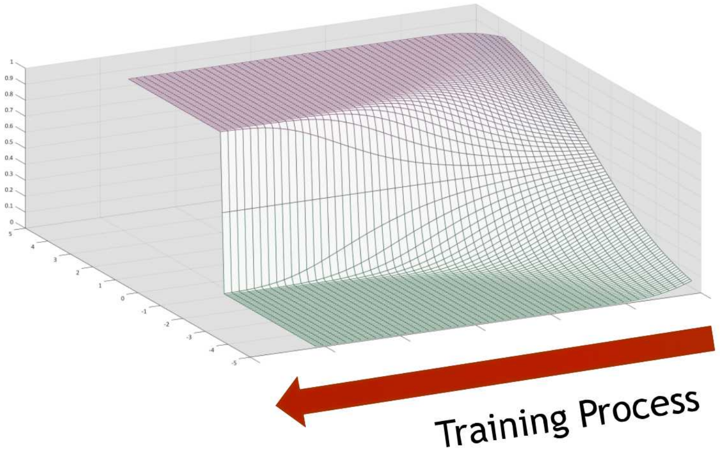

ARTICLES
nature machine intelligence

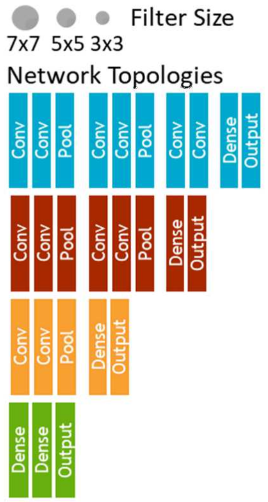
Training deep neural networks for binary communication with the Whetstone method

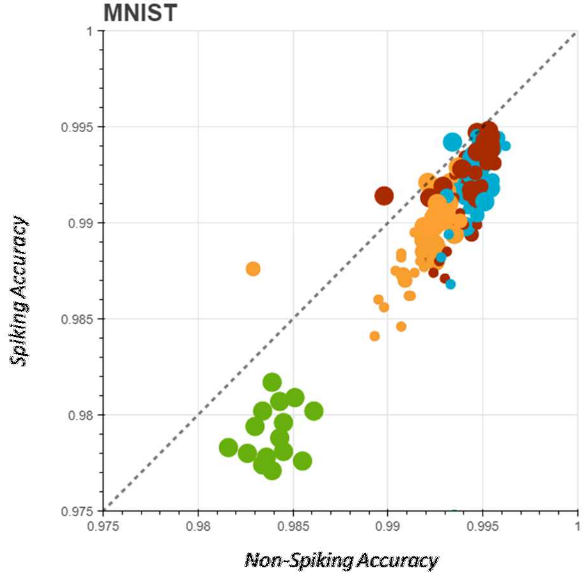
William Severa, Craig M. Vineyard, Ryan Dellana, Stephen J. Verzi and James B. Aimone

Severa et al., Nature Machine Intelligence, Feb 2019
Vineyard et al., NICE Proceedings, 2019

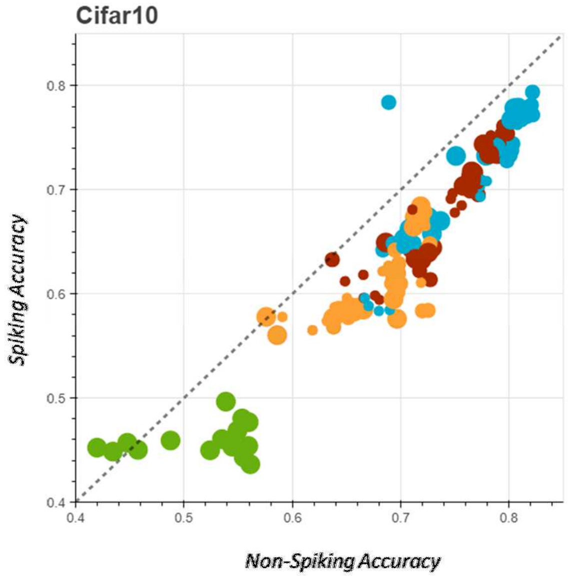
Whetstone has only minimal penalty for binary activations

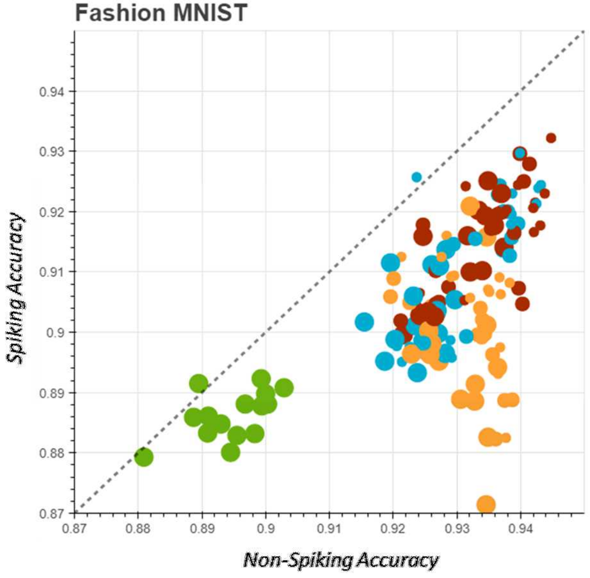




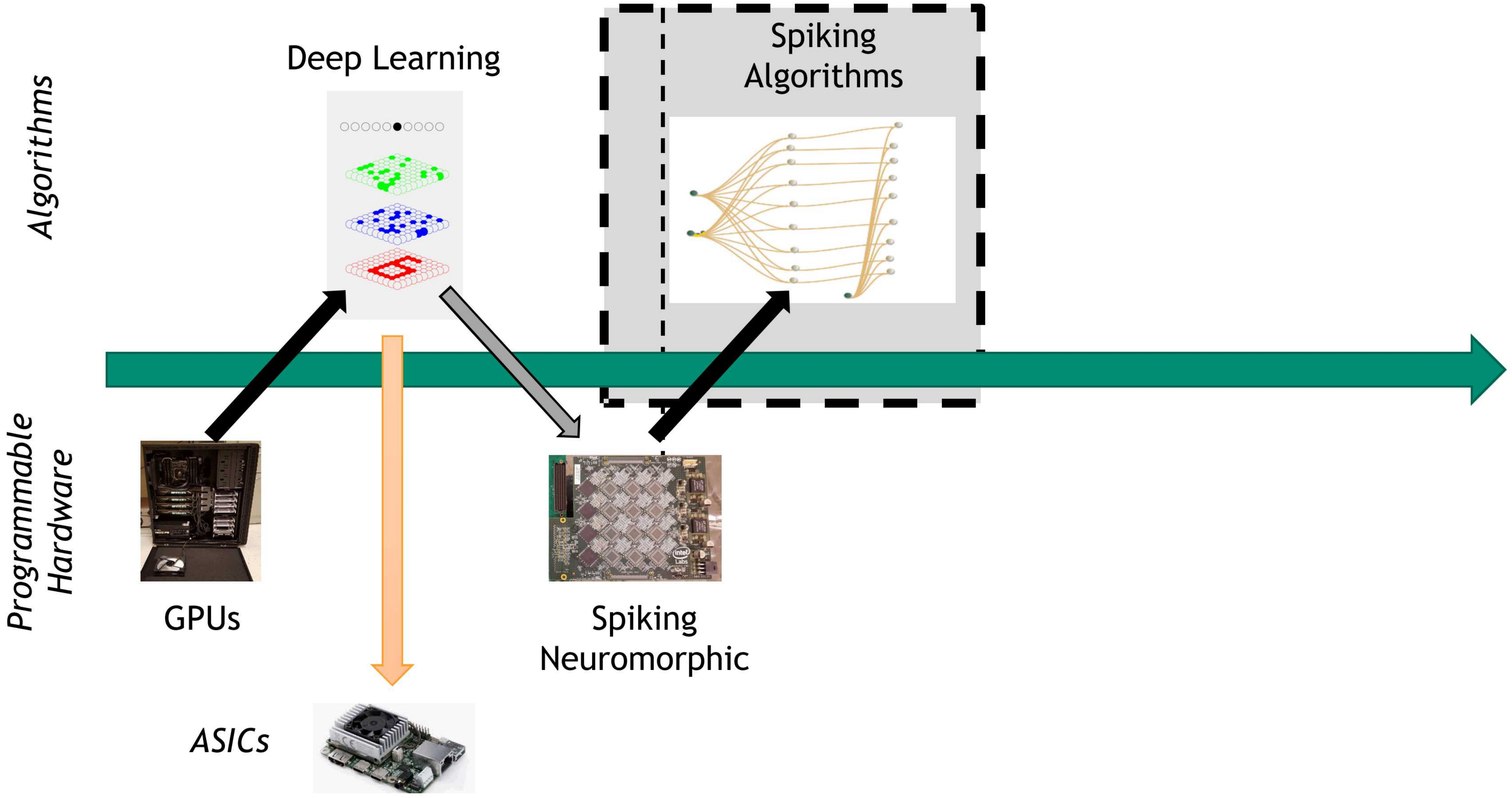


Method	MNIST	CIFAR-10
Whetstone (VGG-like)	**0.9953**	**0.8467**
Whetstone (10-net ensemble)	**0.9953**	**0.8801**
Eliasmith, et al.	0.9912	0.8354
EEDN	0.9942	0.8932
Rueckauer, et al.	0.9944	0.9085
BinaryNet	0.9904	0.8985

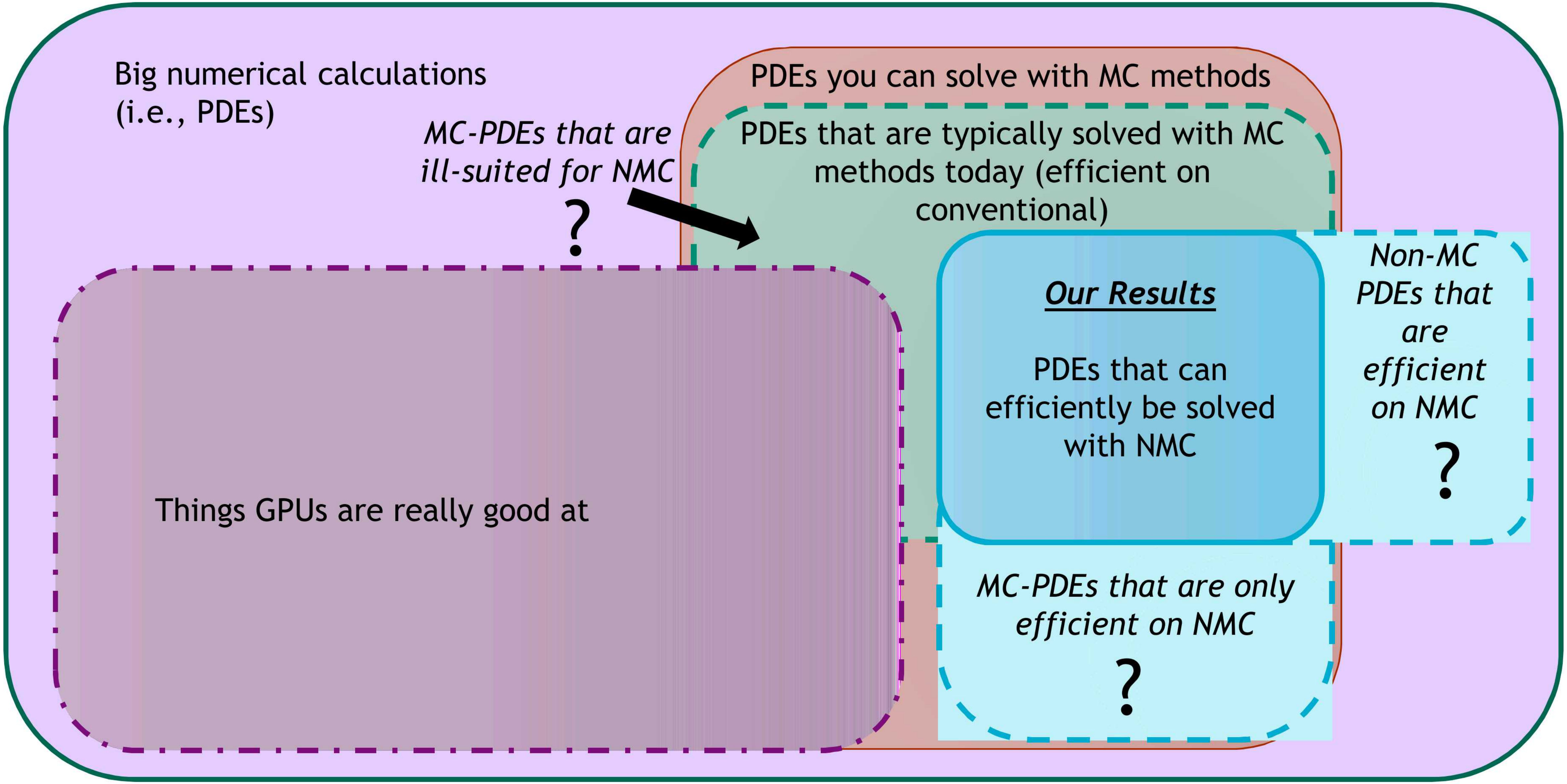




Spiking neuromorphic hardware needs more than just deep learning

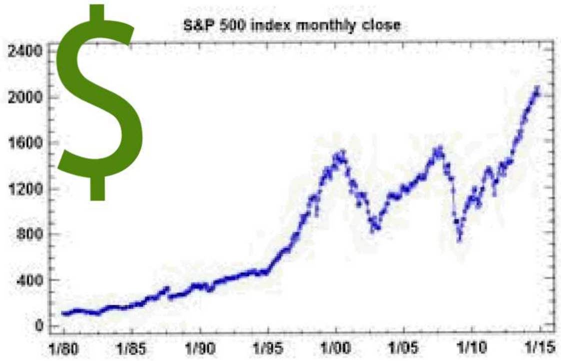


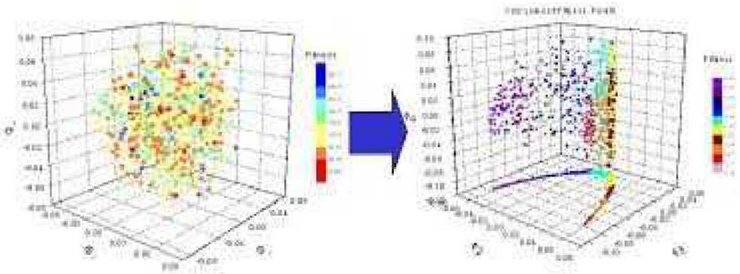
Our hypothesis: There exists a class of scientific computing algorithms for which neuromorphic computing is *efficient*

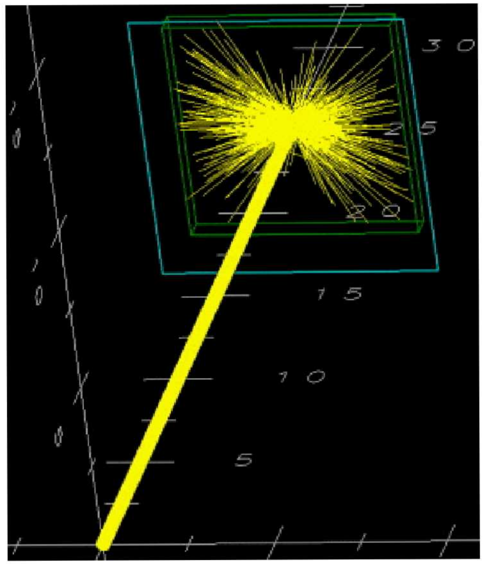


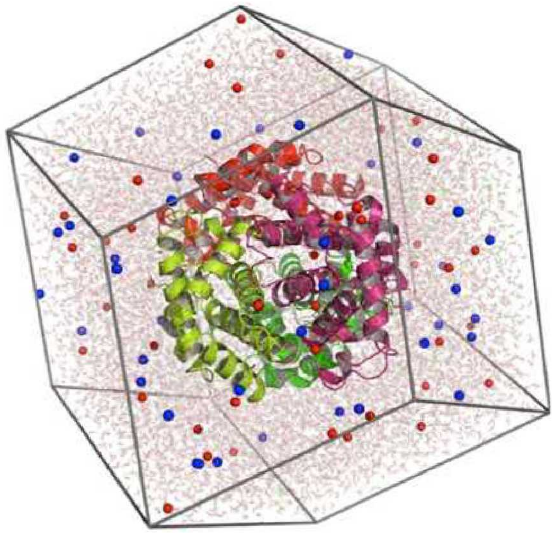
Spiking circuits can efficiently solve stochastic differential equations

Diffusion: $$\frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial x^2}$$

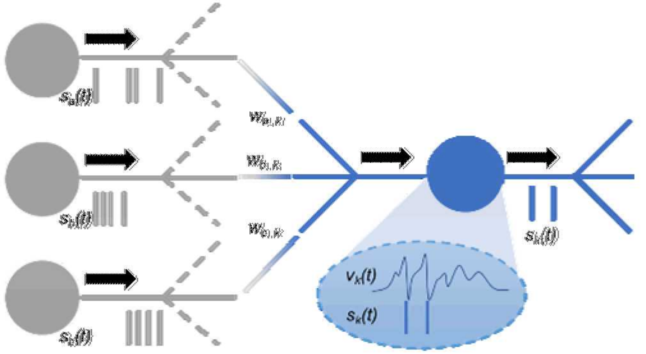




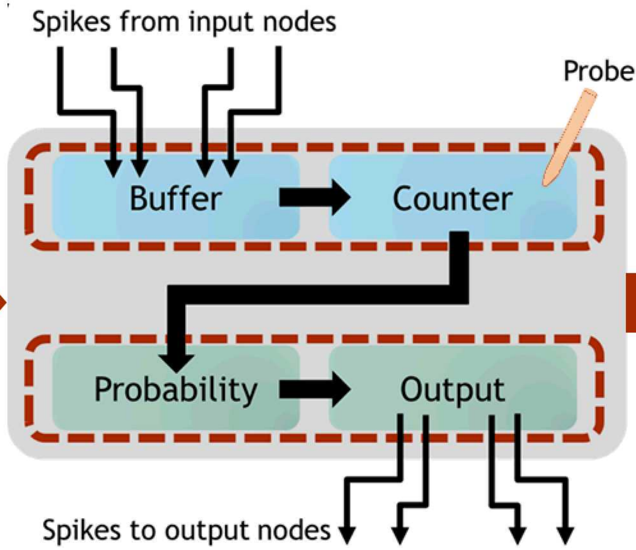


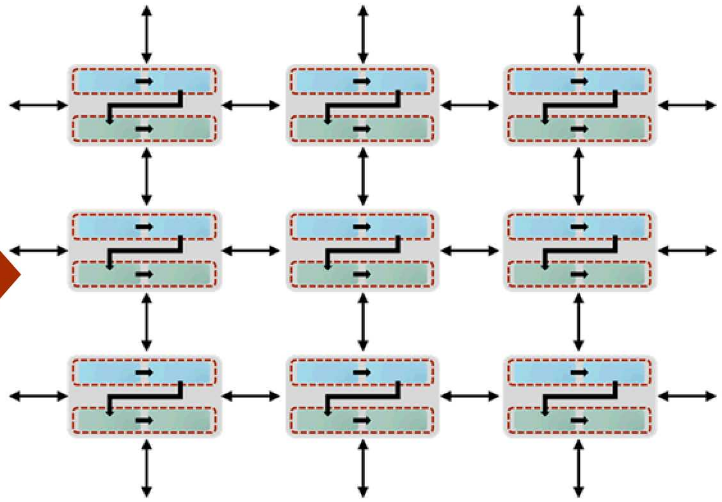


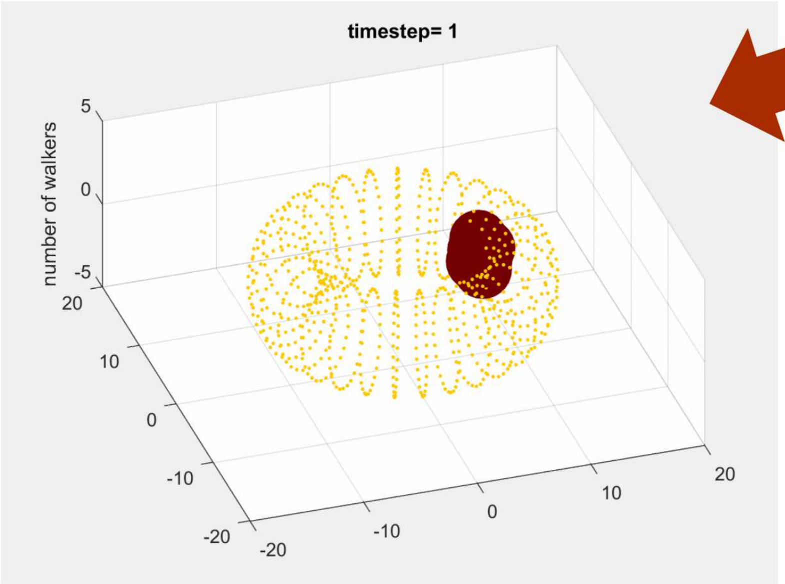
Neuromorphic algorithm can simulate random walks



Leaky Integrate and Fire Neuron





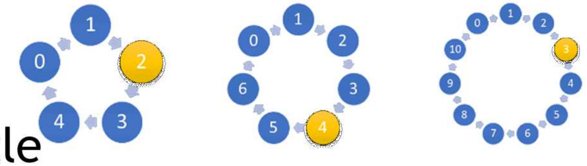


Smith et al., in preparation 2020

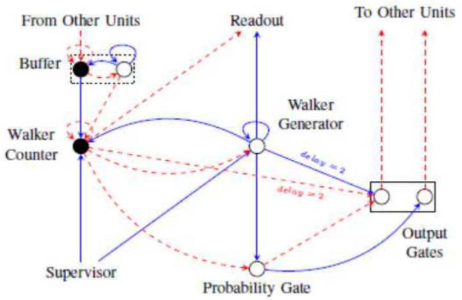
Spiking circuits can efficiently solve stochastic differential equations

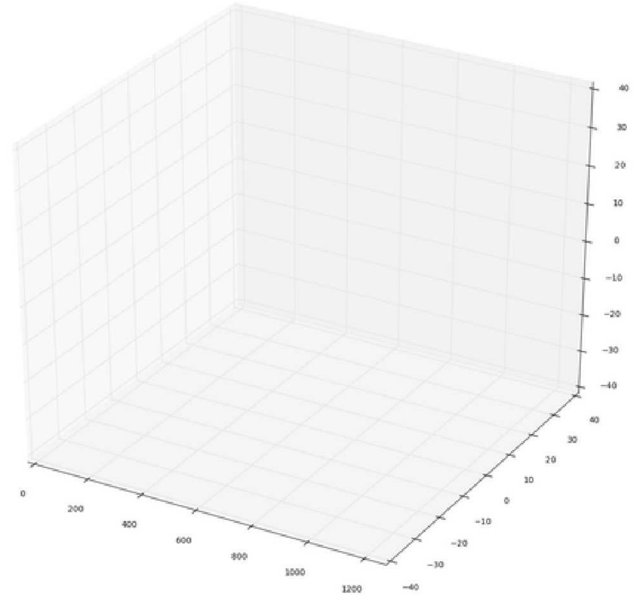
Diffusion: $$\frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial x^2}$$

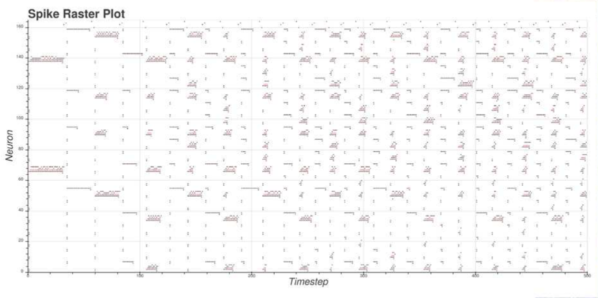
Modular circuit of spiking neurons per random walk particle



RW counting circuit of spiking neurons per simulation mesh vertex



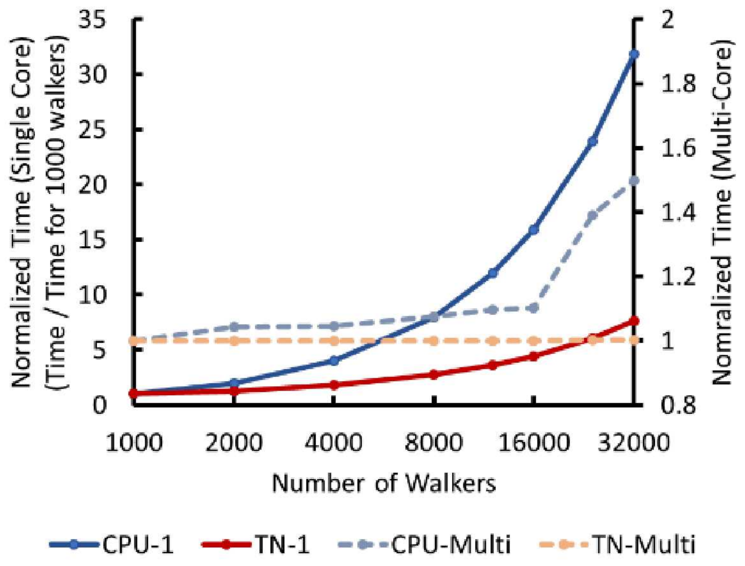




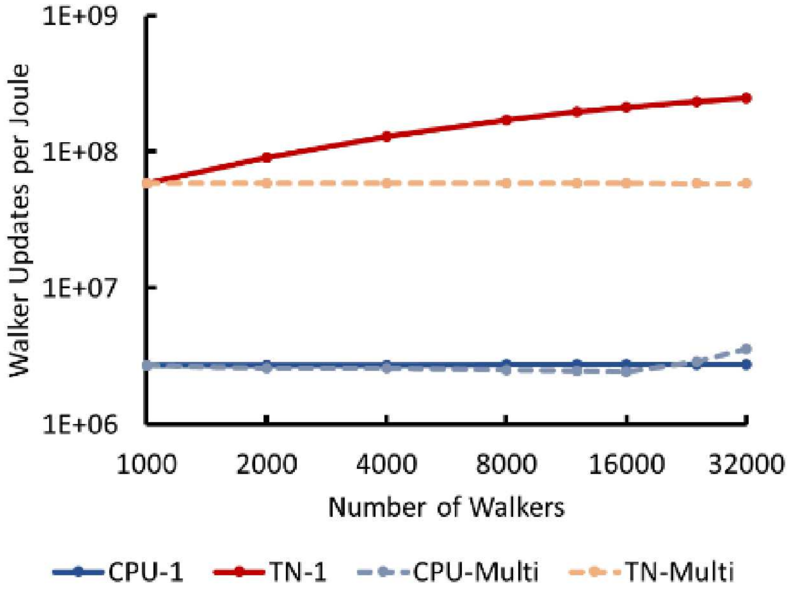
Severa et al., IJCNN 2018

We can identify a *neuromorphic advantage* for simulating random walks

We define a *neuromorphic advantage* as an algorithm that shows a demonstrable **advantage** in terms of one resource (e.g., energy) while exhibiting comparable **scaling** in other resources (e.g., time).

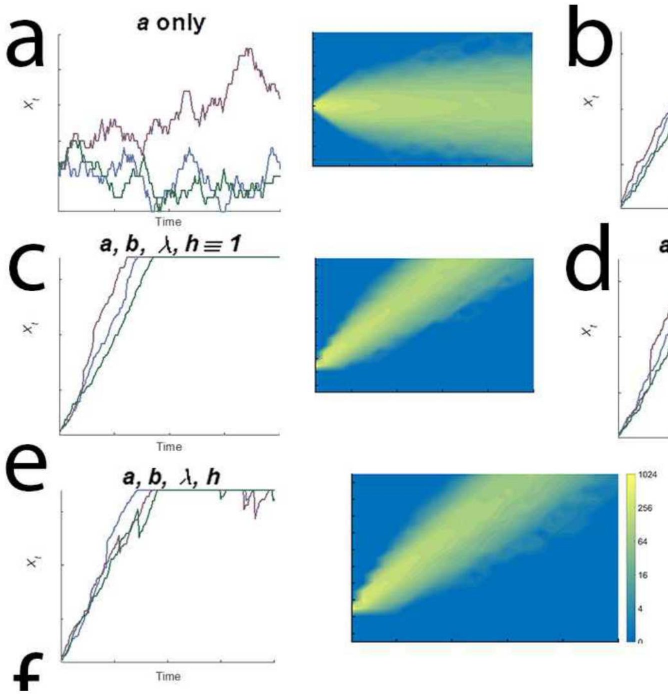


- We show a *neuromorphic advantage* for implementing simple random walks on neuromorphic hardware compared to CPU implementation
  - Same task, architecture specific algorithms
  - TrueNorth and Loihi are slower, but NMC algorithm time scales better
  - **Overall energy consumption (speed / power) is markedly better (20x-100x) on NMC**



Smith et al., in review 2020

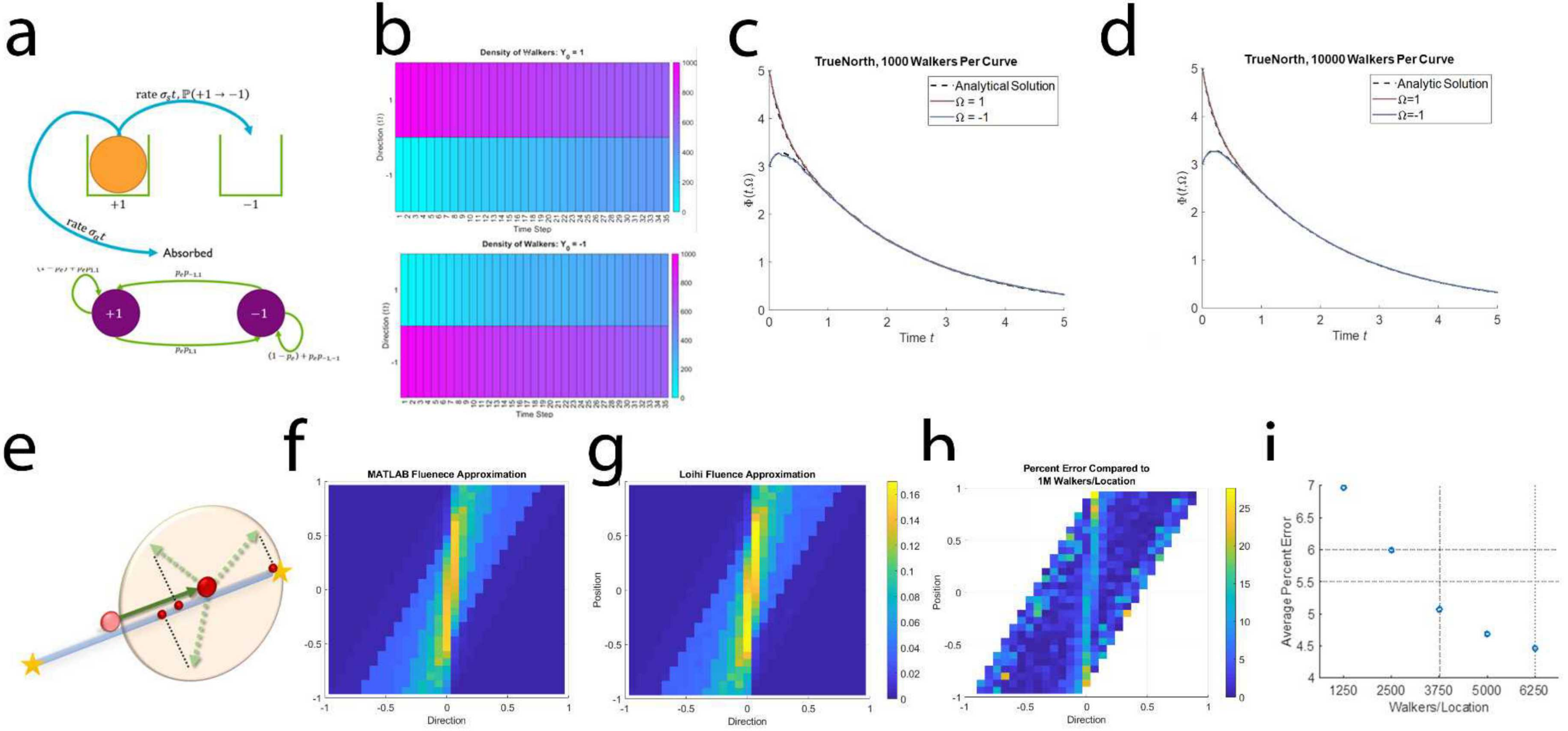
What PDEs can these stochastic processes be useful for?



Non-Zero Terms	Example Application
Time-dependent problems	
a, b, c, f	European Option Pricing
λ, c, h	Simplified Particle Flux Density (See **Fig. 3a-d**)
λ, b, c, f, h	Boltzmann Flux Density
a, c	Heat Equation with Dissipation (See **Fig. 4c**)
Steady-state problems	
a, f	Electrostatic Scalar Potential, Heat Transport, or Simple Beam Bending [23]
λ, b, f, h	Simplified Particle Fluence (See **Fig 3e-i**)

Smith et al., in review 2020

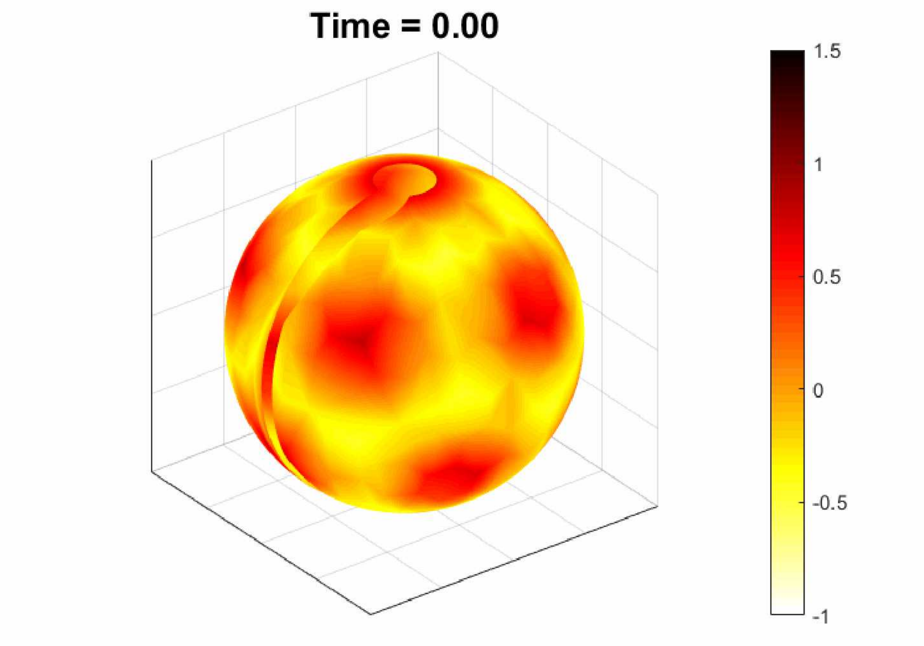
Simulating Particle Transport on TrueNorth and Loihi



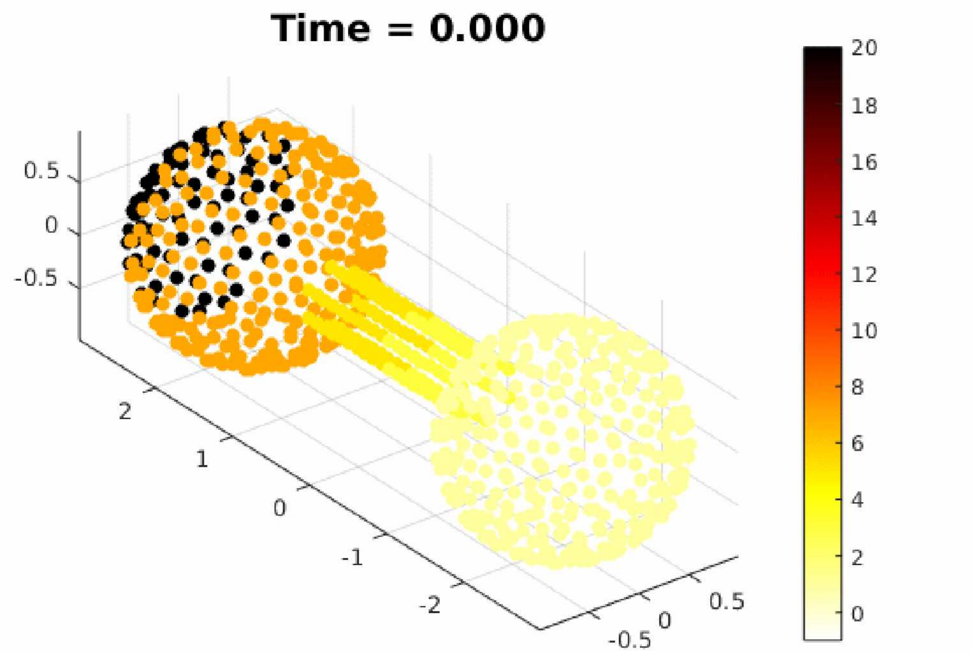
Smith et al., in review 2020

Algorithm can implement non-Euclidean geometries

- Stochastic process can be over any mesh, in theory there are no restrictions on geometry (beyond number of mesh-points and hardware size)
- Implemented random walks over surface of sphere and across barbell shape
- Can extend to any graph / network



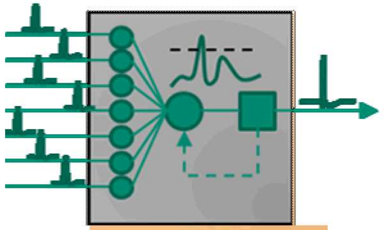
On Loihi



Neural Simulation

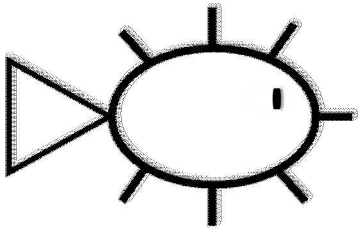
Smith et al., in review 2020

What can you do with spiking neurons?

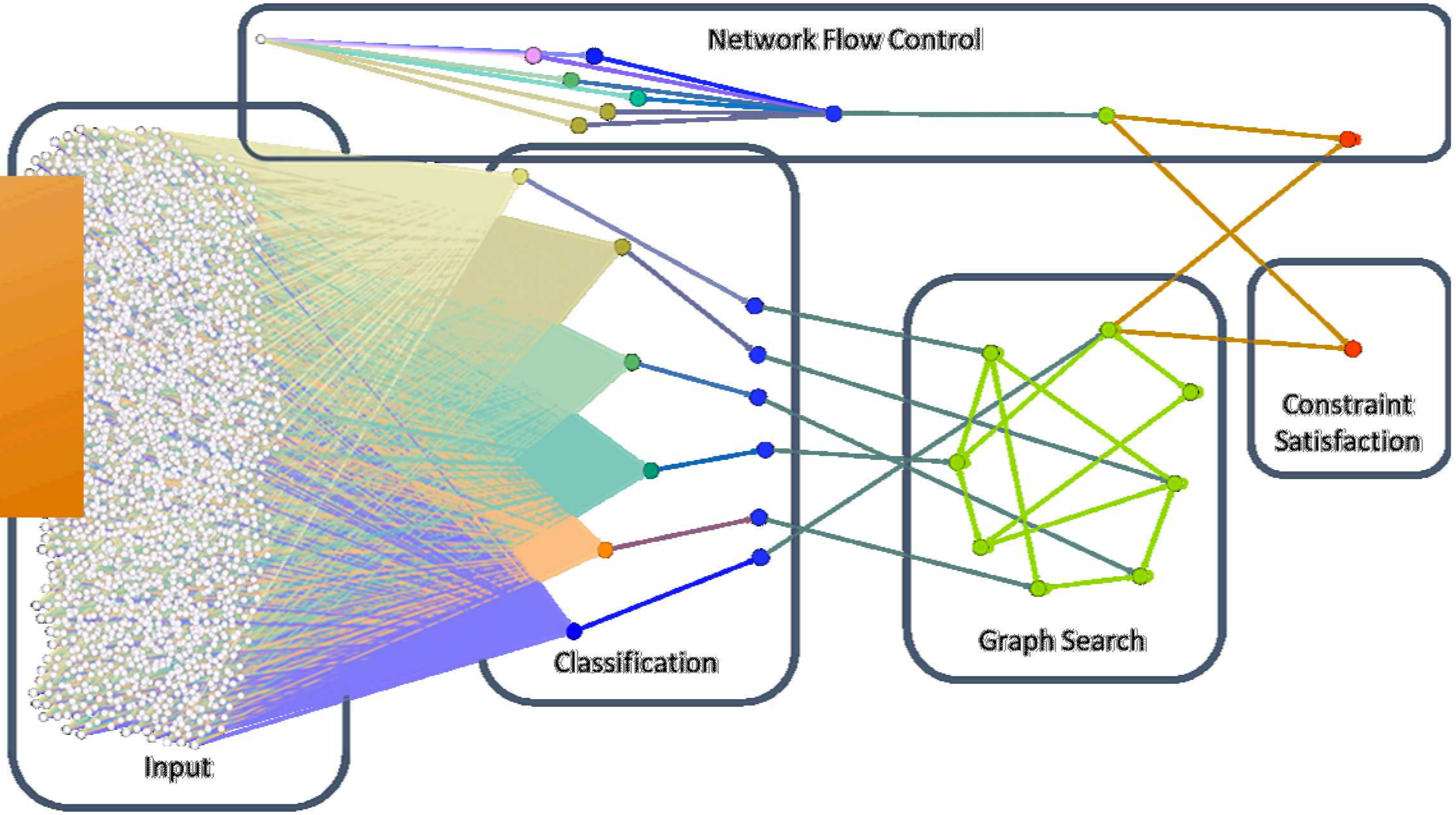


Treat neurons as powerful logic gates

Algorithms are circuits…

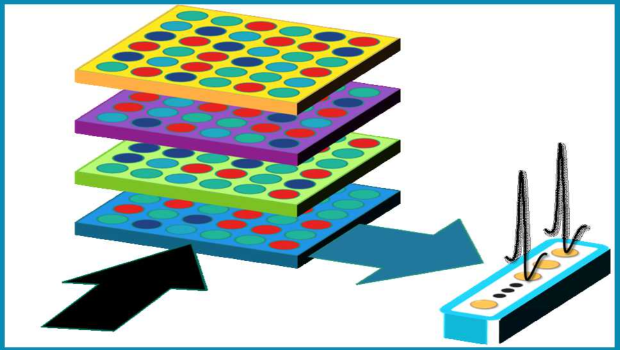


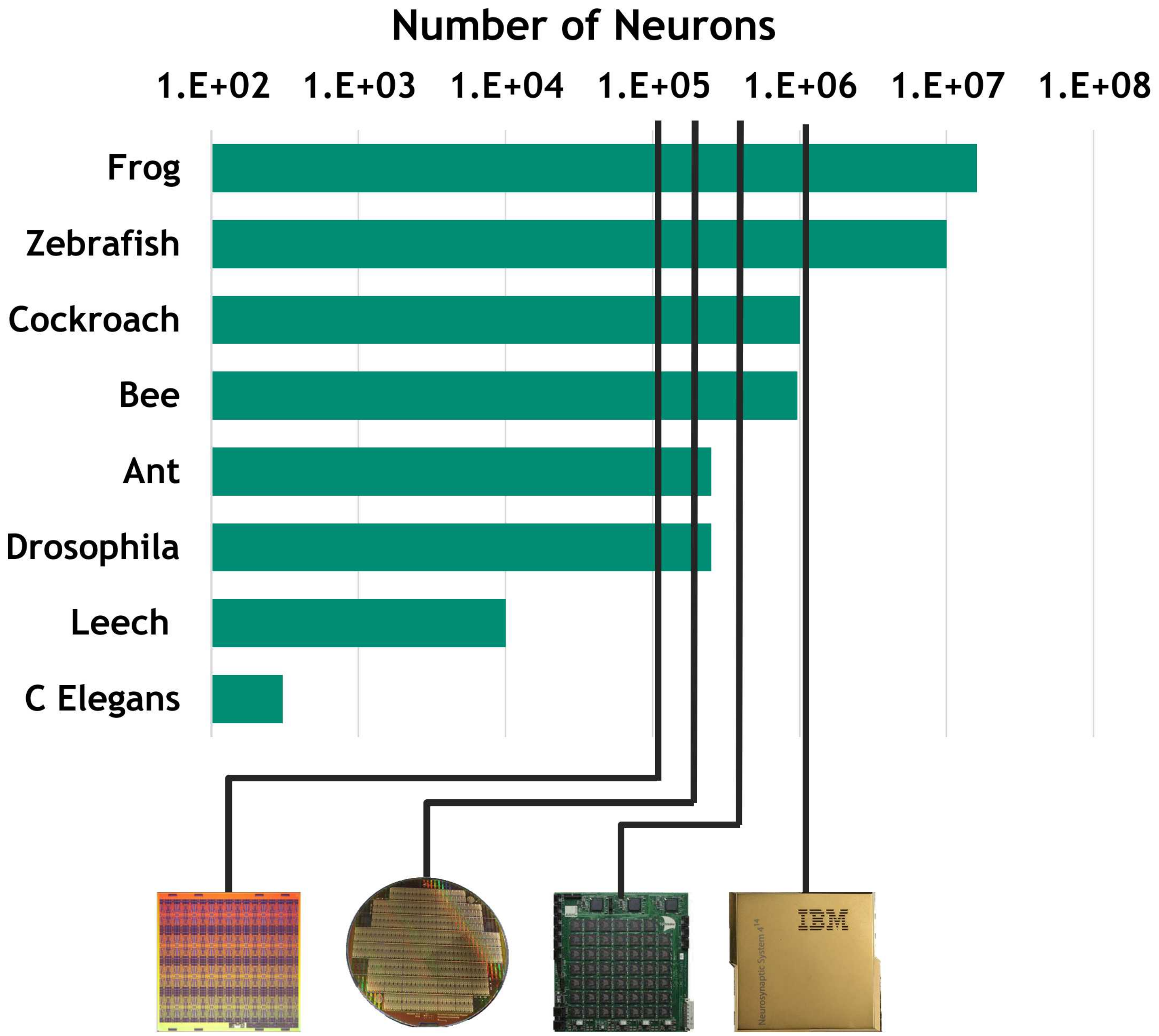
Fugu



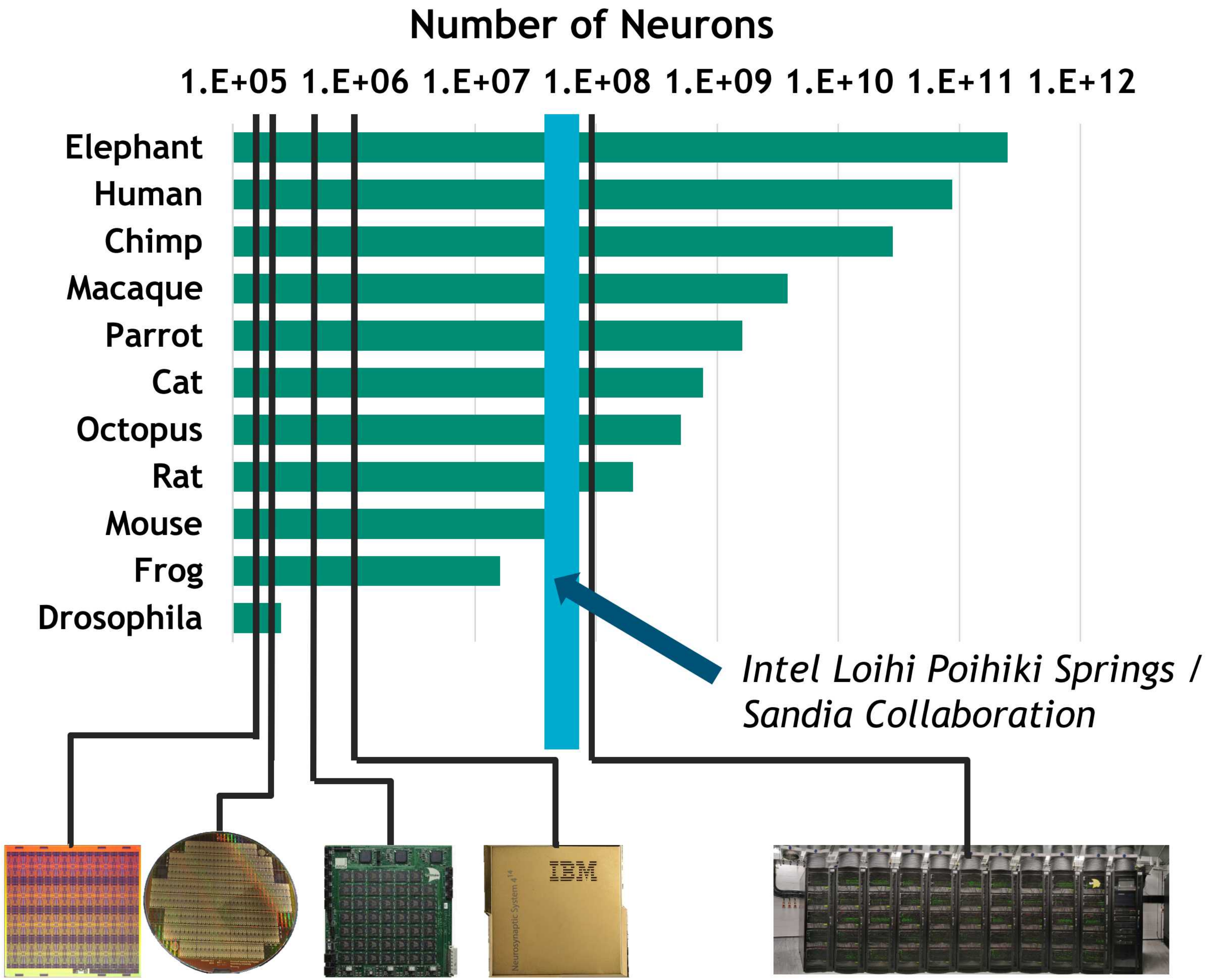
Aimone et al, ICONS 2019

Part 2:
Scaling Neuromorphic Architectures to the Next Level





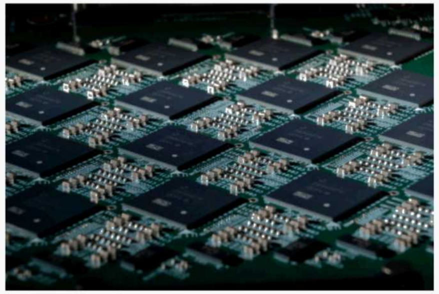
Number of Neurons
1.E+02
1.E+03
1.E+04
1.E+05
1.E+06
1.E+07
1.E+08
Frog
Zebrafish
Cockroach
Bee
Ant
Drosophila
Leech
C Elegans
IBM


Number of Neurons
1.E+05 1.E+06 1.E+07 1.E+08 1.E+09 1.E+10 1.E+11 1.E+12
Elephant
Human
Chimp
Macaque
Parrot
Cat
Octopus
Rat
Mouse
Frog
Drosophila
Intel Loihi Poihiki Springs / Sandia Collaboration
IBM

Overview About Intel Strategic Priorities Segments News & Events Financial Info Stock I

Overview | Press Releases | IR Calendar | Annual Stockholders' Meeting | Investor Meeting

Intel and Sandia National Labs Collaborate on Neuromorphic Computing



A close-up shot of an Intel Nahuku board, each of which contains 8 to 32 Intel Loihi neuromorphic chips. Intel's latest neuromorphic system, Pohoiki Beach, is made up of multiple Nahuku boards and contains 64 Loihi chips. Pohoiki Beach was introduced in July 2019. (Credit: Tim Herman/Intel Corporation)

OCTOBER 02, 2020 9:00AM EDT

Download as PDF

SANTA CLARA, Calif.--(BUSINESS WIRE)-- **What's New:** Today, Intel Federal LLC announced a three-year agreement with Sandia National Laboratories (Sandia) to explore the value of

Sandia National Laboratories

About News Research Partnerships Careers

News Releases Publications Media Contacts Media B-roll Events Video Image Gallery Subscribe

Sandia Labs News Releases

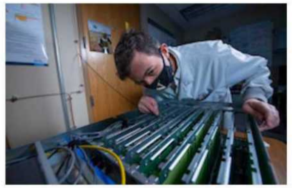
October 2, 2020

50 million artificial neurons to facilitate machine-learning research at Sandia

Total number in final system could reach 1 billion or more

ALBUQUERQUE, N.M. — Fifty million artificial neurons — a number roughly equivalent to the brain of a small mammal — were delivered from Portland, Oregon-based Intel Corp. to Sandia National Laboratories last month, said Sandia project leader Craig Vineyard.

The neurons will be assembled to advance a relatively new kind of computing, called neuromorphic, based on the principles of the human brain. Its artificial components pass information in a manner similar to the action of living neurons, electrically pulsing only when a synapse in a complex circuit has absorbed enough charge to produce an electrical spike.



Sandia National Laboratories researcher J. Darby Smith does an initial examination of computer boards containing artificial neurons designed by Intel Corp. (Photo by Regina Valenzuela) Click on the thumbnail for a high-resolution image

"With a neuromorphic computer of this scale," Vineyard said, "we have a new tool to understand how brain-based computers are able to do impressive feats that we cannot currently do with ordinary computers."

Improved algorithms and computer circuitry can create wider applications for neuromorphic computers, said Vineyard.

Sandia manager of cognitive and emerging computing John

THENEXTPLATFORM

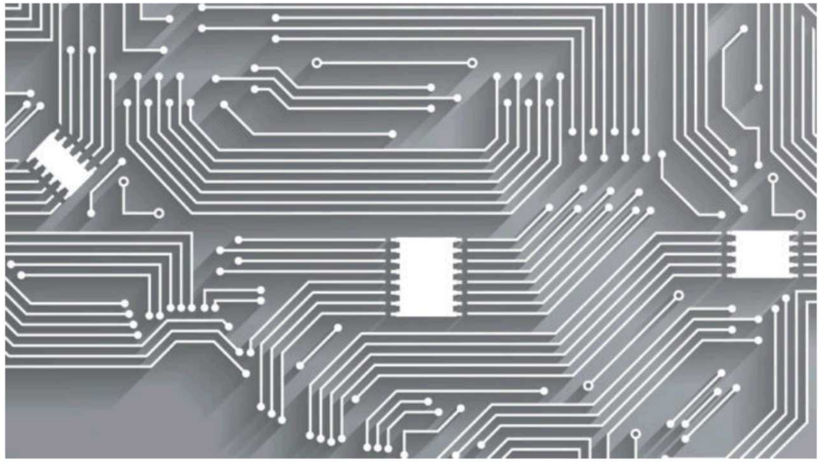
HOME COMPUTE STORE CONNECT CONTROL CODE AI HPC ENTERP

LATEST › With "Crossroads" Supercomputer, HPE Notches Another DOE Win › HPC

HOME › COMPUTE › On the Fringes of Useful Neuromorphic Scalability

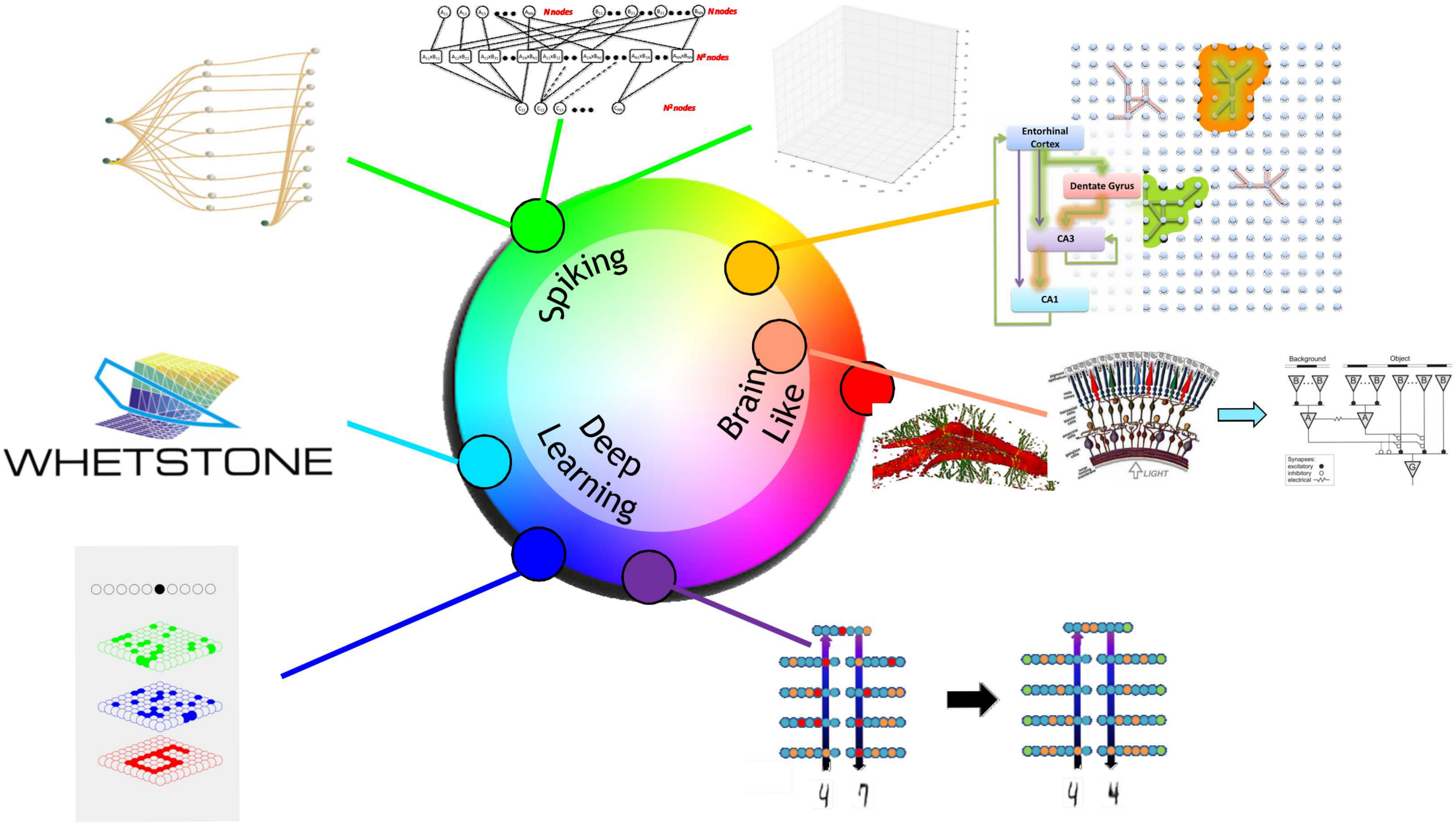
ON THE FRINGES OF USEFUL NEUROMORPHIC SCALABILITY

October 5, 2020 Nicole Hemsoth

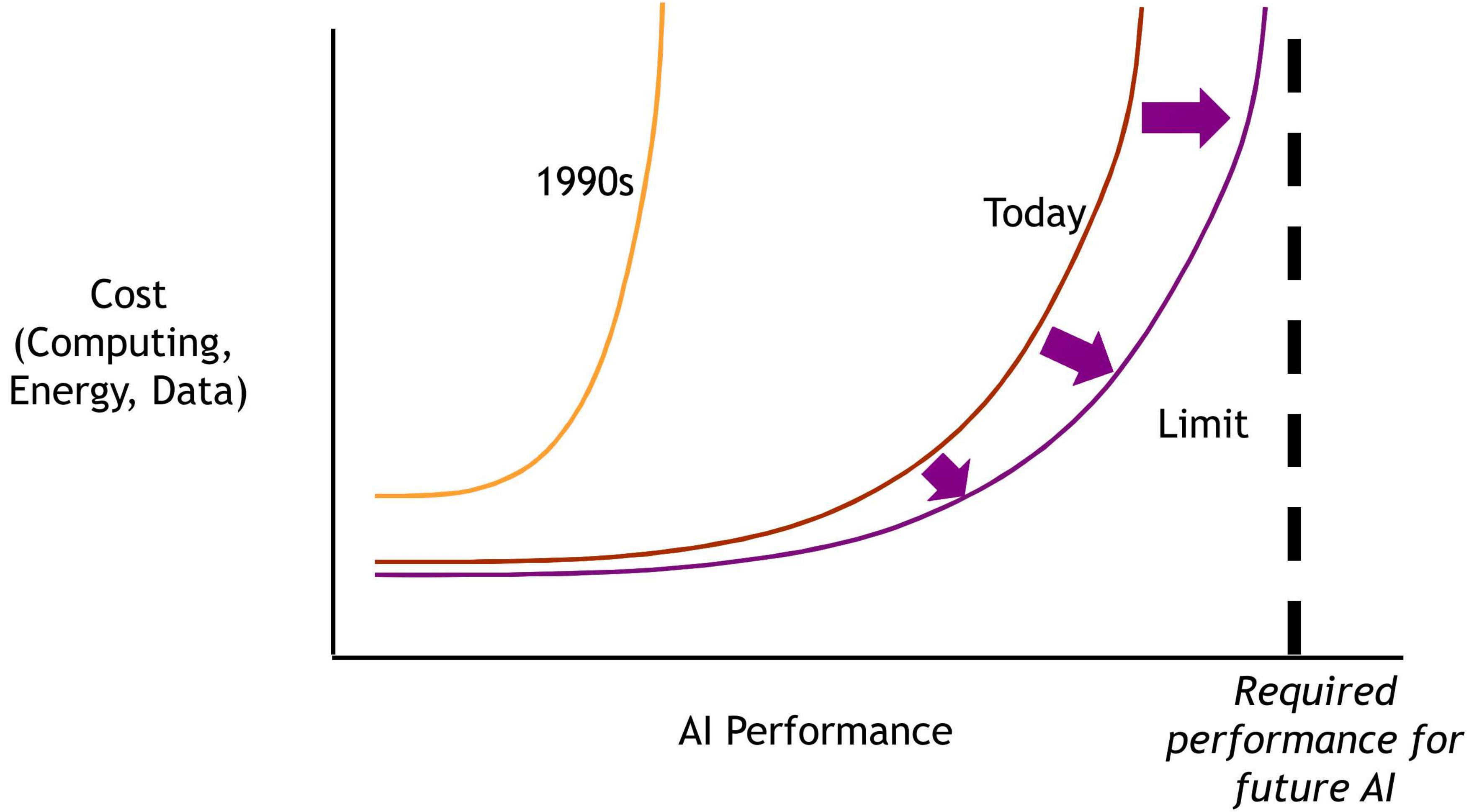


When it comes to novel computing architectures, whether in quantum, deep learning, or

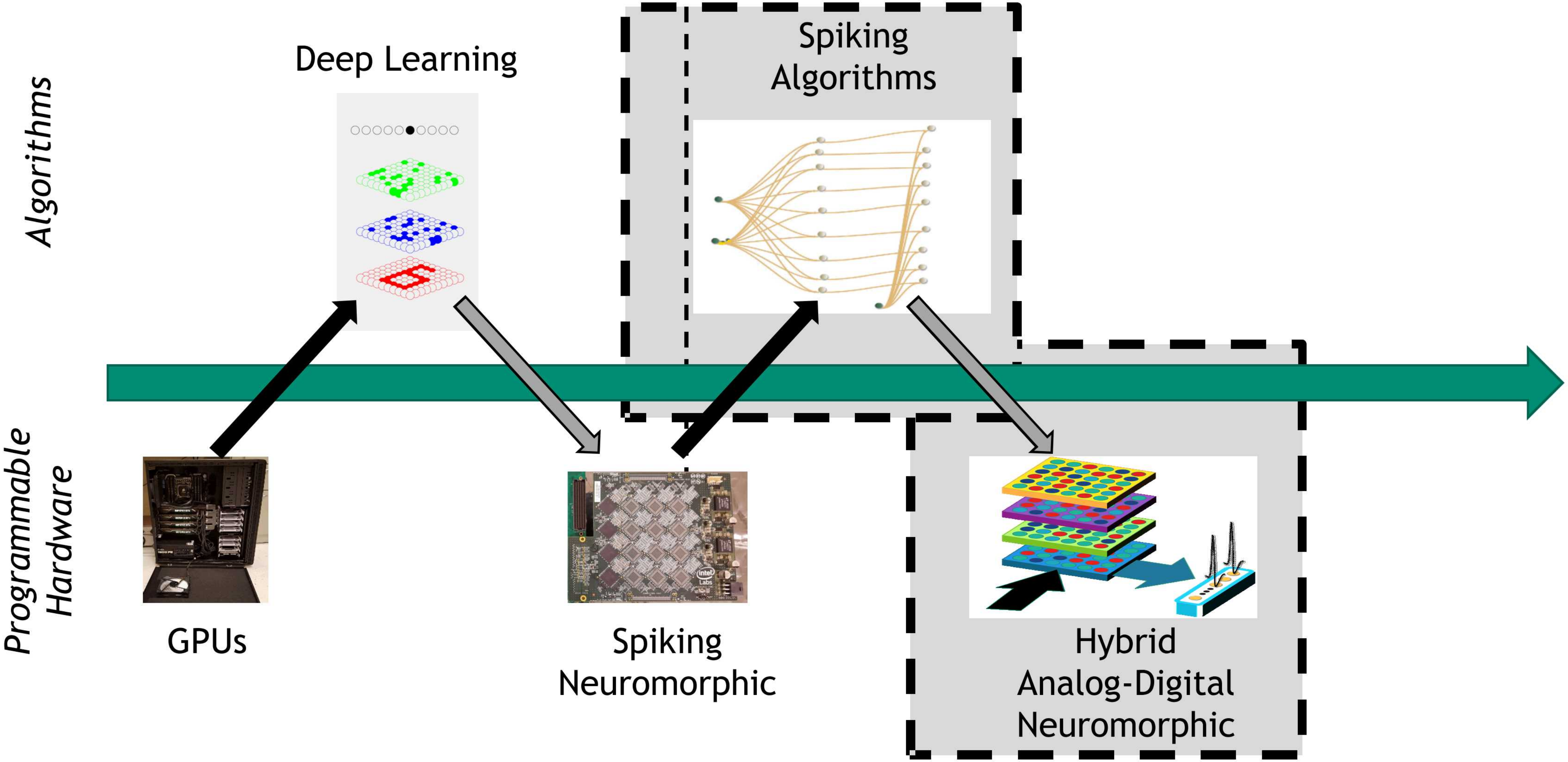
More neurons = better



We need a new post-Moore's Law path to cheaper computing



Scaling to real-world applications will require future hardware solutions



Neuromorphic Processors

Analog

- Focus on Kirchhoff Law – enabled computation
  - Neurons sum current across weighted synapses
  - Neural nodes sum current over weighted memristors
- Substantial energy and time savings
  - Non-trivial costs of precision
  - Practical issues limit size and integration with digital logic
- Ideal scenario
  - Train weights in situ
  - Compatible with linear algorithms

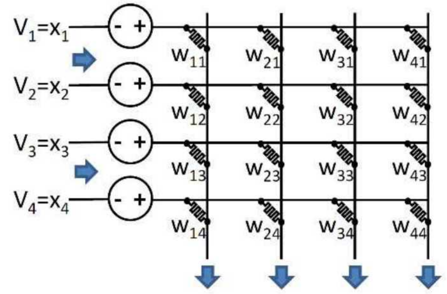
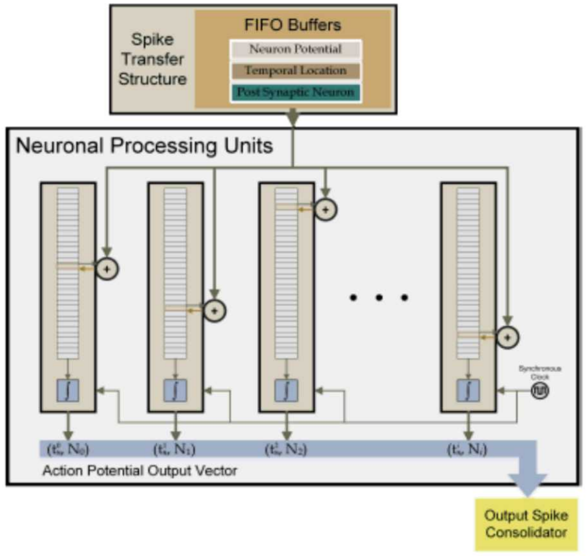


Fig 1: Analog RRAMs can be used to reduce the energy of a vector matrix multiply. The conductance of each RRAM represents a weight. Analog input values are represented by the input voltages or input pulse lengths. This allows all the read operations, multiplication operations and sum operations to occur in a single step. A conventional architecture must perform these operations sequentially for each weight resulting in a higher energy and delay.

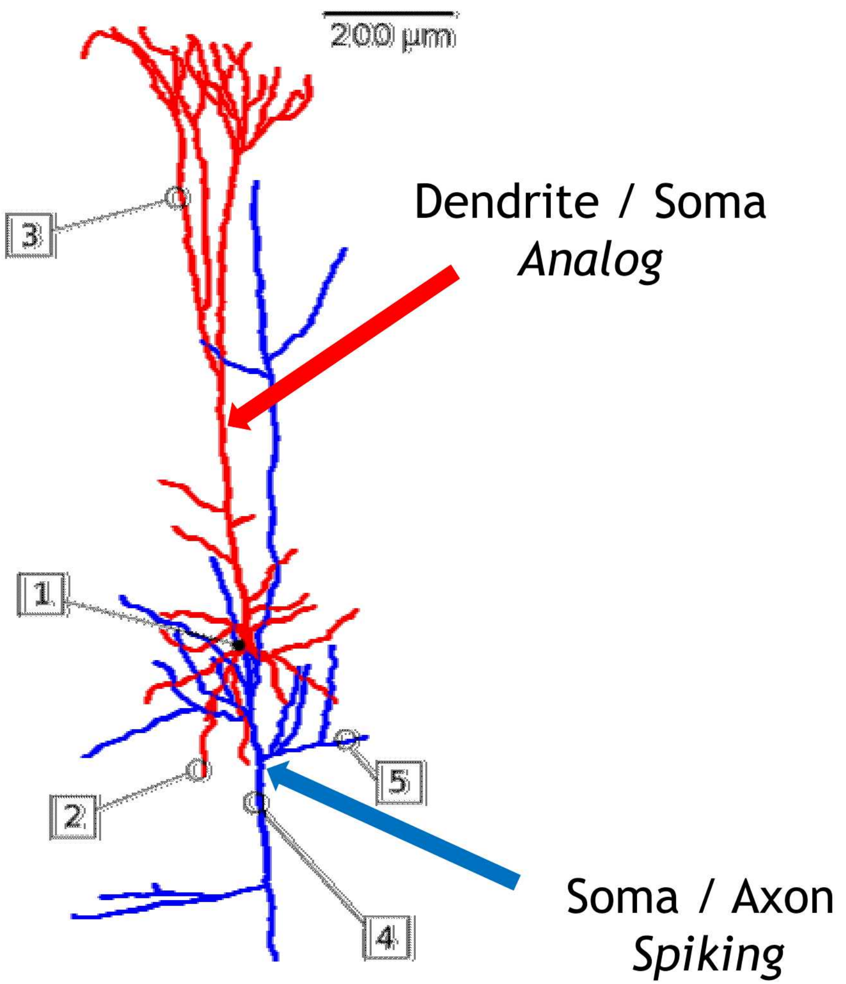
Agarwal et al., E3S 2015

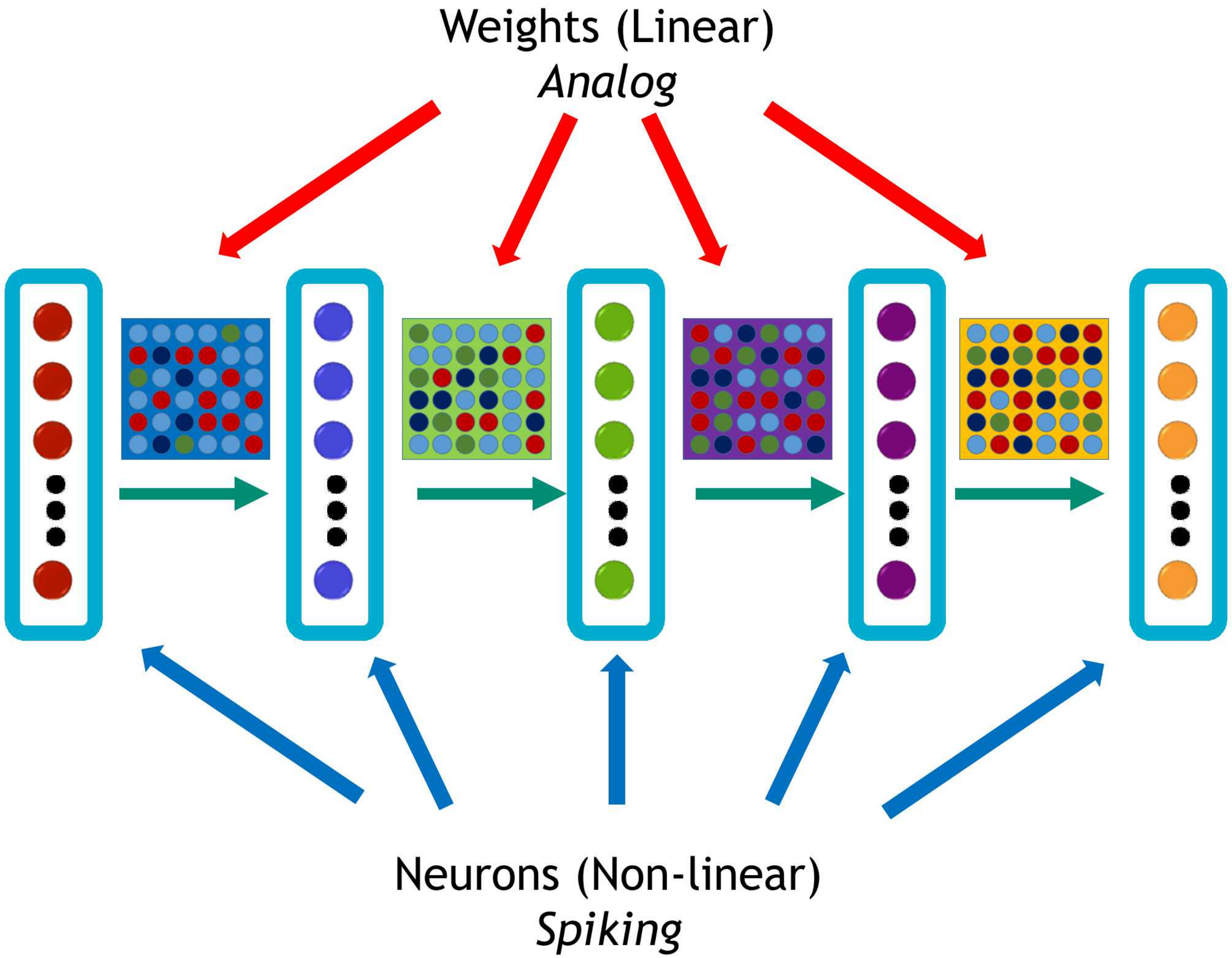
Digital

- Rely on event-driven "spiking" for communication
  - Communication only needed for '1's', not otherwise
  - Equivalent to large threshold gate networks + time dimension
- Substantial energy savings
  - Information in time dimension; limiting time savings
- Compatible and scalable using conventional technology
- Ideal scenario
  - Algorithms can be reframed in discrete spiking form
  - Learning algorithms are reformulated for spiking approaches

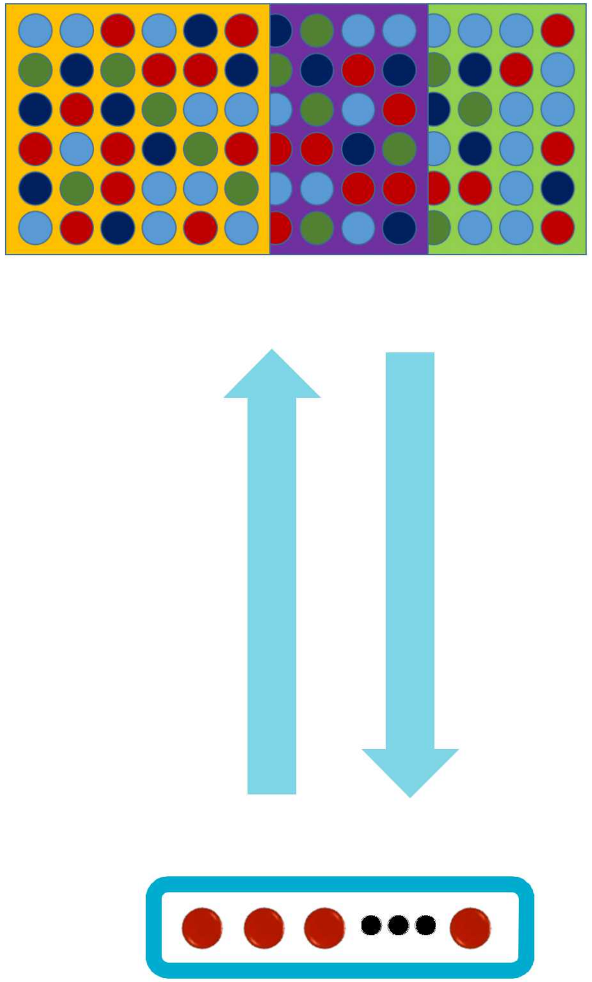


Brains, and neural networks, do both…

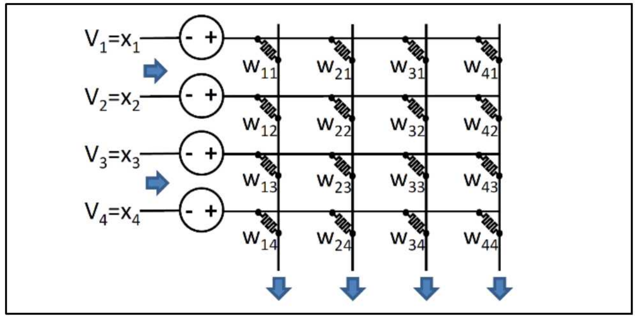




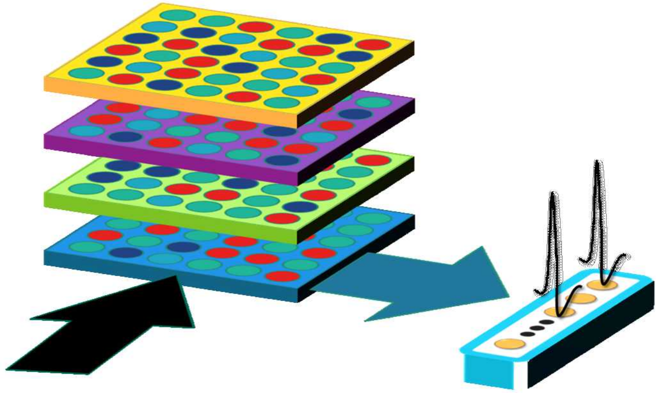
Future of neuromorphic is likely a hybrid spiking / analog system



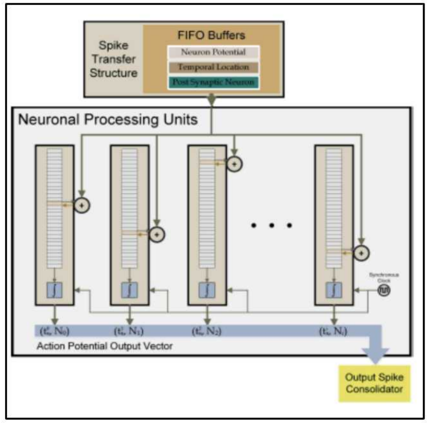
Analog Synapses



3d Hybrid System for Communication

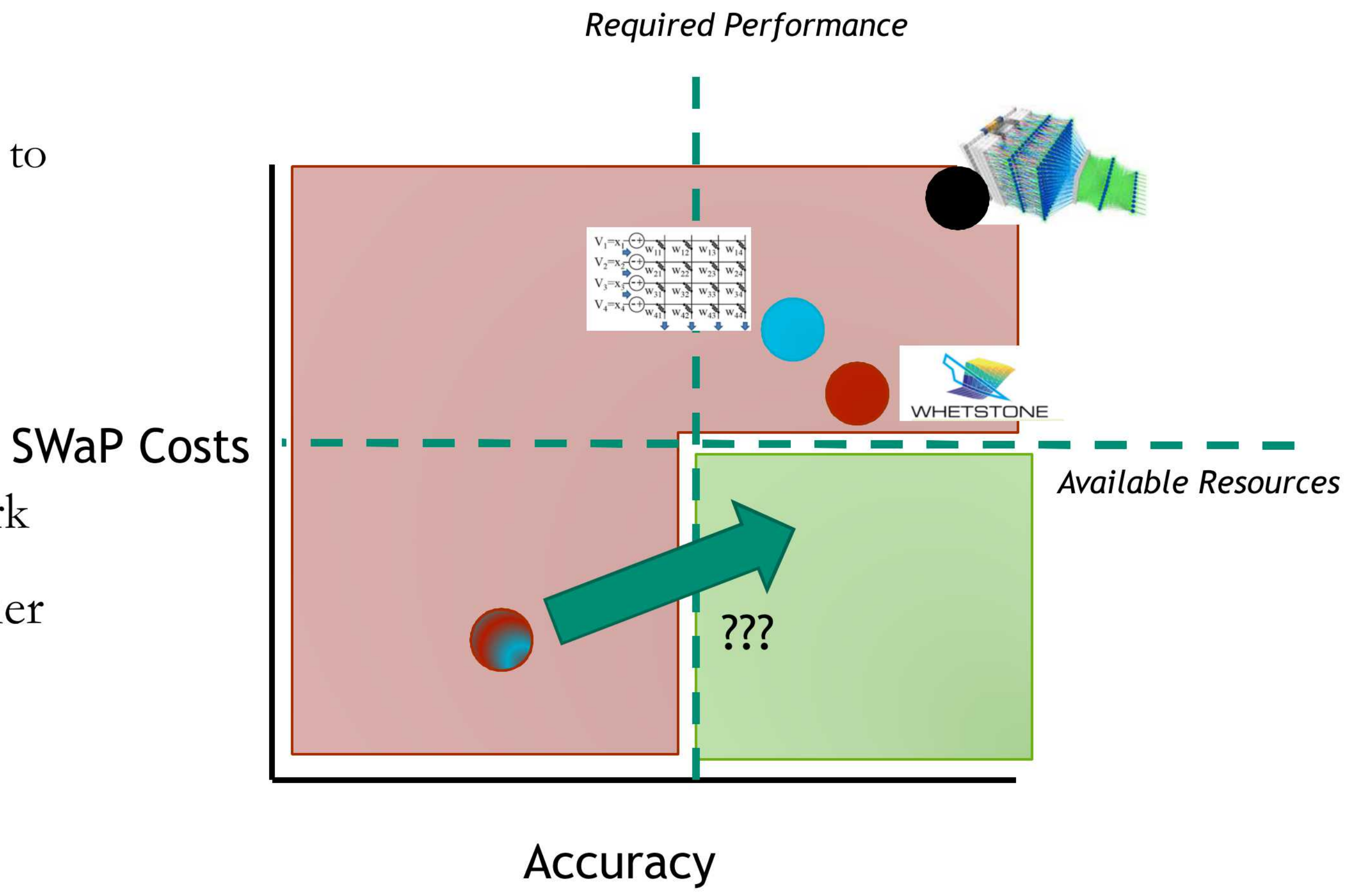


Digital Neurons

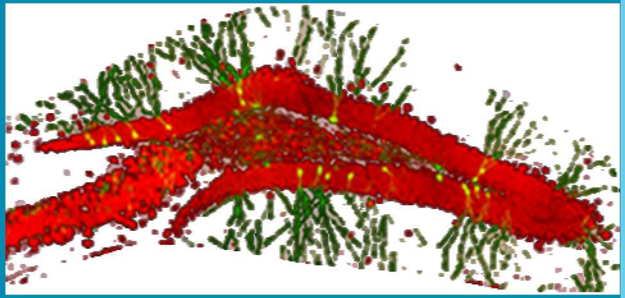


Implications of analog noise + spiking accuracy

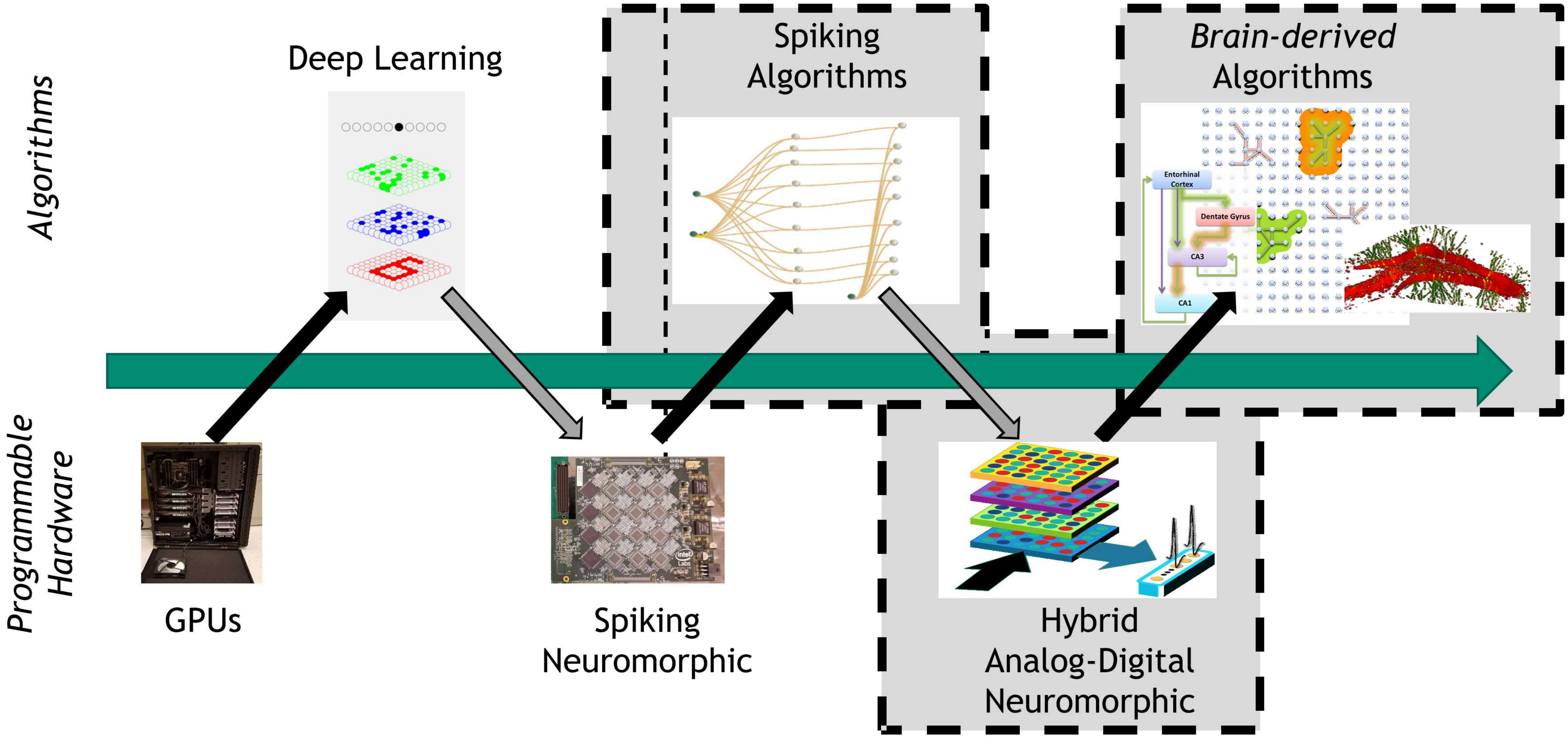
- Some accuracy / efficiency trade-off is likely okay
  - Most applications have requirements
  - Most neural network solutions can be tweaked to change resources / performance
- Spiking or Analog alone probably is not sufficient
- However, we're seeing that without some modification, analog + digital likely won't work
- We need to somehow mitigate errors in either training or architecture implementation



Part 3:
Looking to the brain for intelligence beyond deep learning



Can we really get the brain into algorithms?



DOI:10.1145/3231589

Advances in neurotechnologies are reigniting opportunities to bring neural computation insights into broader computing applications.

BY JAMES B. AIMONE

Neural Algorithms and Computing Beyond Moore's Law

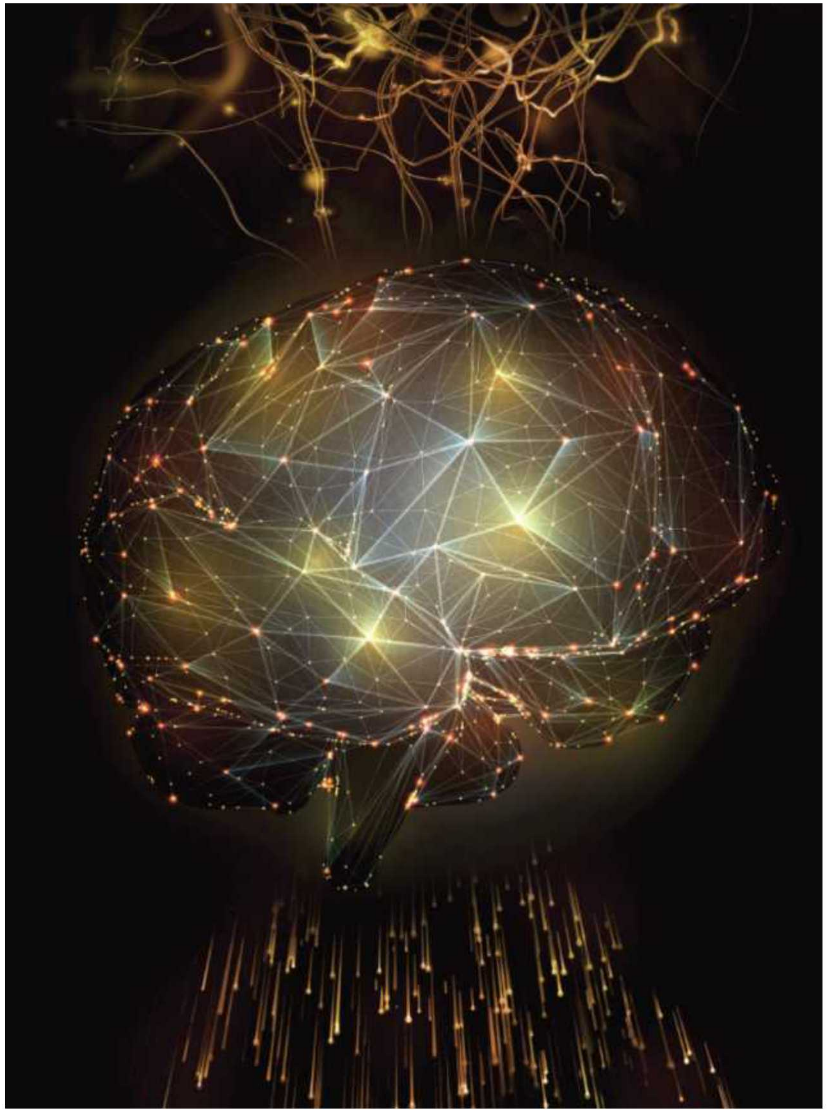
THE IMPENDING DEMISE of Moore's Law has begun to broadly impact the computing research community.[38] Moore's Law has driven the computing industry for many decades, with nearly every aspect of society benefiting from the advance of improved computing processors, sensors, and controllers. Behind these products has been a considerable research industry, with billions of dollars invested in fields ranging from computer science to electrical engineering. Fundamentally, however, the exponential growth in computing described by Moore's Law was driven by advances in materials science.[30,37] From the start, the power of the computer has been limited by the density of transistors. Progressive advances in how to manipulate silicon through advancing lithography methods and new design tools have kept advancing computing in spite of perceived limitations of the dominant fabrication processes of the time.[37]

There is strong evidence that this time is indeed different, and Moore's Law is soon to be over for good.[1,38] Already, Dennard scaling, Moore's Law's lesser known but equally important parallel, appears to have ended.[11] Dennard's scaling refers to the property that the reduction of transistor size came with an equivalent reduction of required power.[8] This has real consequences—even though Moore's Law has continued over the last decade, with feature sizes going from ~65nm to ~10nm; the ability to speed up processors for a constant power cost has stopped. Today's common CPUs are limited to about 4GHz due to heat generation, which is roughly the same as they were 10 years ago. While Moore's Law enables more CPU cores on a chip (and has enabled high power systems such as GPUs to continue advancing), there is increasing appreciation that feature sizes cannot fall much further, with perhaps two or three further generations remaining prior to ending.

Multiple solutions have been presented for technological extension of Moore's Law,[3,21,38,29] but there are two main challenges that must be addressed. For the first time, it is not immediately evident that future materials

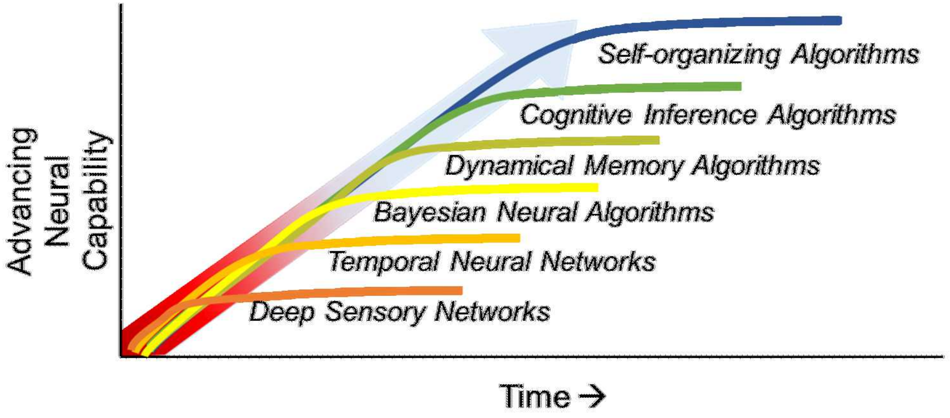
» key insights

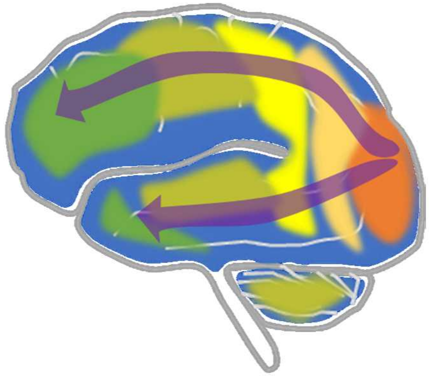
- **While Moore's Law is slowing down, neuroscience is experiencing a revolution, with technology enabling scientists to have more insights into the brain's behavior than ever before and thus positioning the neuroscience field to provide a long-term source of inspiration for novel computing solutions.**
- **Extending the reach of brain-inspiration into computing will not only make current AI methods better, but looking beyond the brain's sensory systems can also expand the reach of AI into new applications.**
- **Realizing the full potential of brain-inspired computing requires increased collaborations and sharing of knowledge between the neuroscience, computer science, and neuromorphic hardware communities.**



How can neuroscience influence AI beyond Deep Learning?

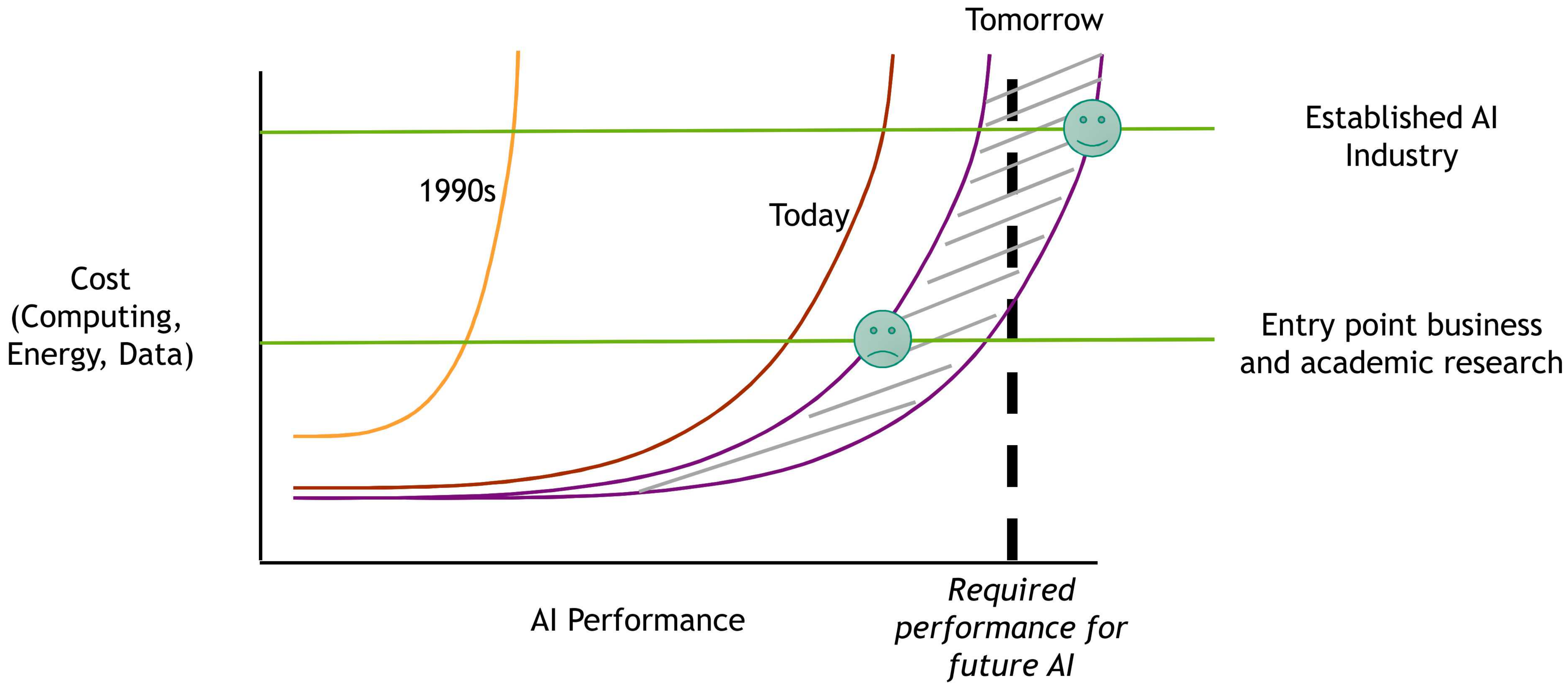
Algorithm Class	Current Algorithms	Inspiration	Application
Deep Vision Processing	Deep Convolutional Networks (VGG, AlexNet, GoogleNet, etc.), HMax, Neocognitron	Hierarchy of sensory nuclei and early sensory cortices	Static feature extraction (e.g., images) & pattern classification
Temporal Neural Networks	Deep Recurrent Networks (e.g., long short-term memory), Hopfield Networks	Local recurrence of most biological neural circuits, especially higher sensory cortices	Dynamic feature extraction (e.g., videos, audio) & classification
Bayesian Neural Algorithms	Predictive Coding, Hierarchical Temporal Memory, Recursive Cortical Networks	Substantial reciprocal feedback between "higher" and "lower" sensory cortices	Inference across spatial and temporal scales
Dynamical Memory and Control Algorithms	Liquid State Machines, Echo State Networks, Neural Engineering Framework	Continual dynamics of hippocampus, cerebellum, and prefrontal and motor cortices	Online learning content-addressable memory & adaptive motor control
Cognitive Inference Algorithms	Reinforcement learning (e.g., Deep Q-learning) Neural Turing Machines	Integration of multiple modalities and memory into prefrontal cortex, which provides top-down influence on sensory processing	Context and experience dependent information processing and decision making
Self-organizing Algorithms	Neurogenesis Deep Learning	Initial development and continuous refinement of neural circuits to specific input and outputs	Automated neural algorithm development for unknown input and output transformations



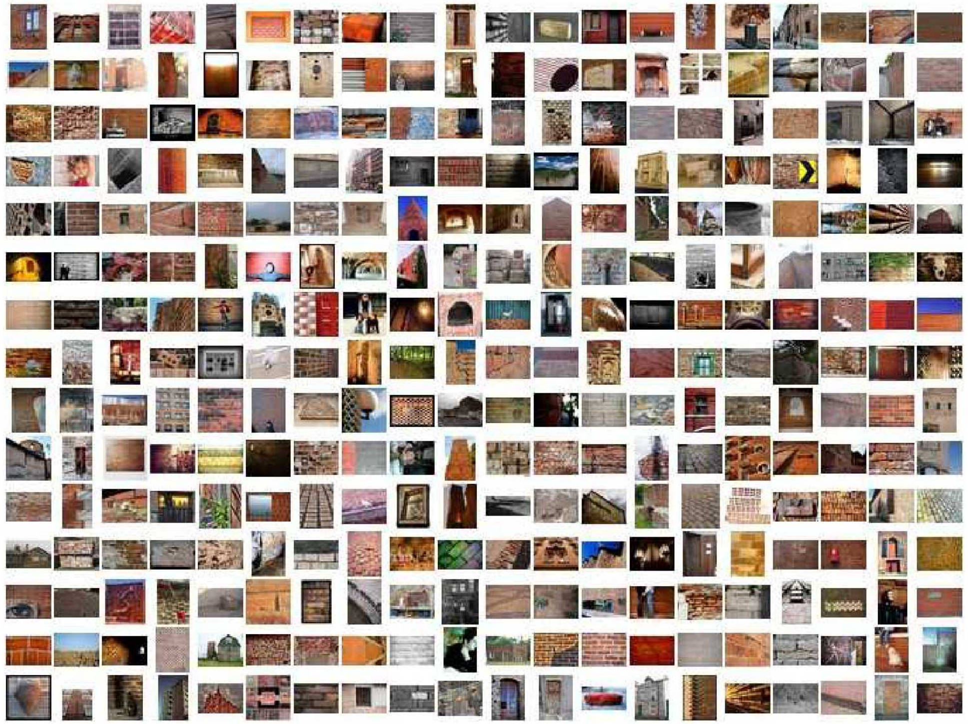


Aimone JB, Communications of ACM, April 2019

Data is a potential unseen barrier to entry for AI

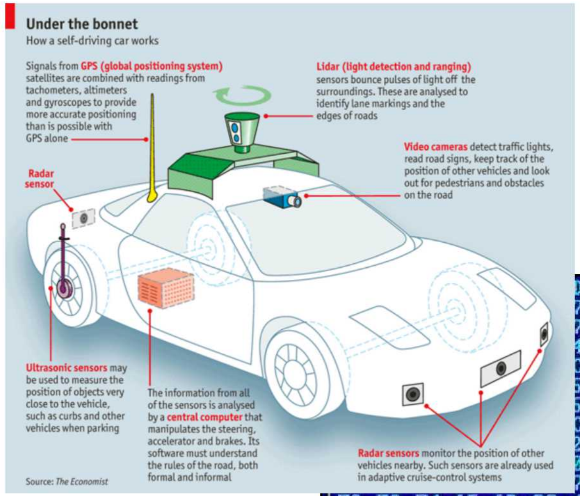


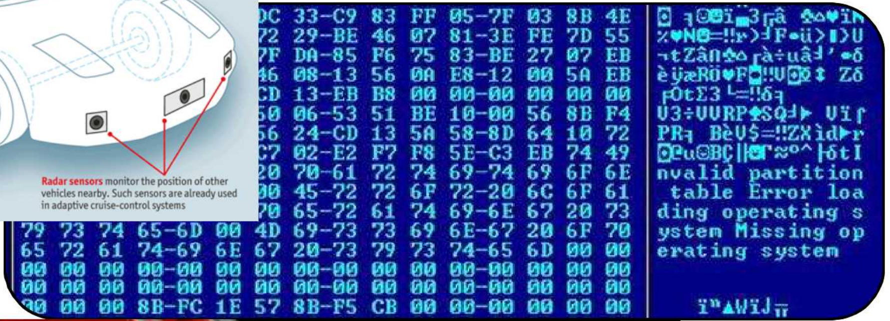
Some types of applications are well-suited for deep ANNs

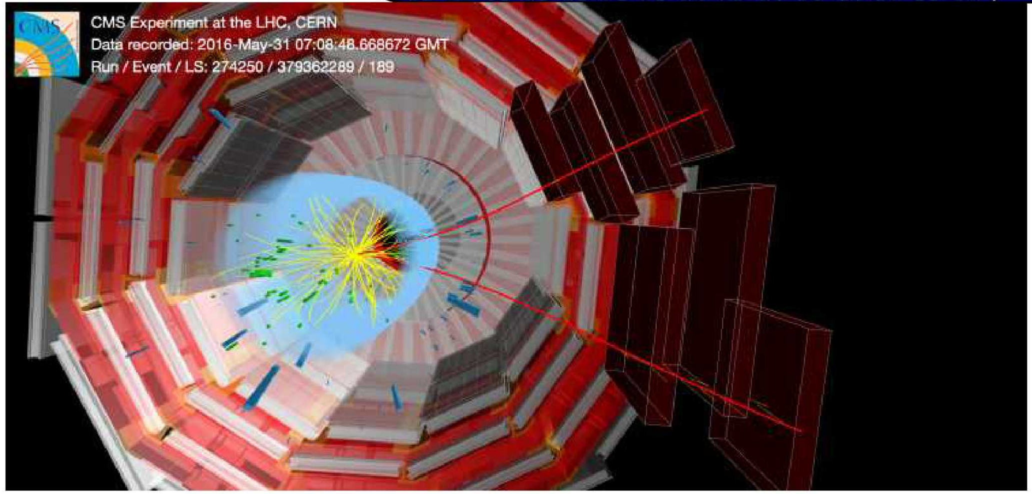


- Deep neural networks benefit from *high volume* of relatively *low-dimensional* data
  - *N>>d*
  - Good (necessary?) for training very large, relatively model-free networks

… other applications are not

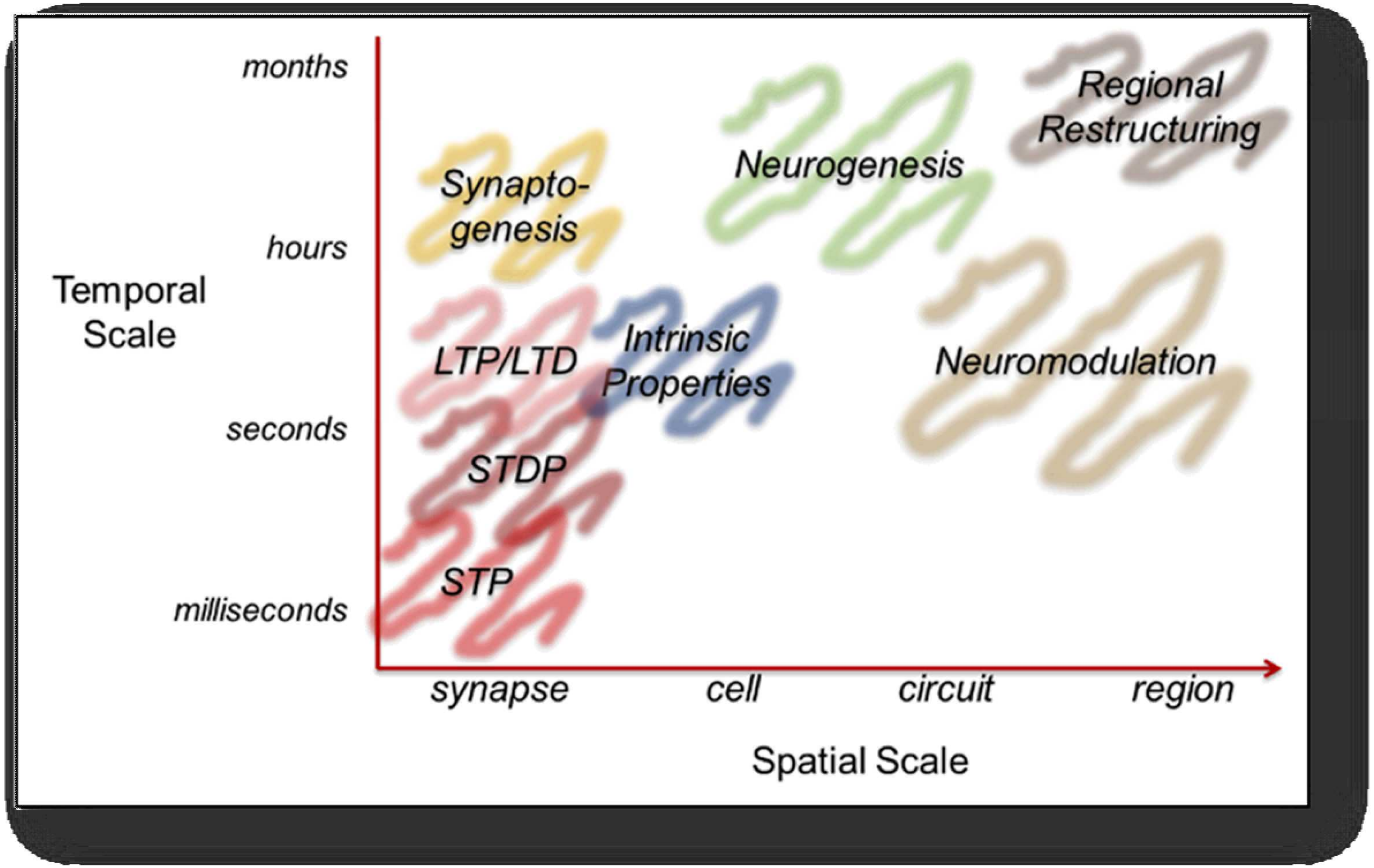




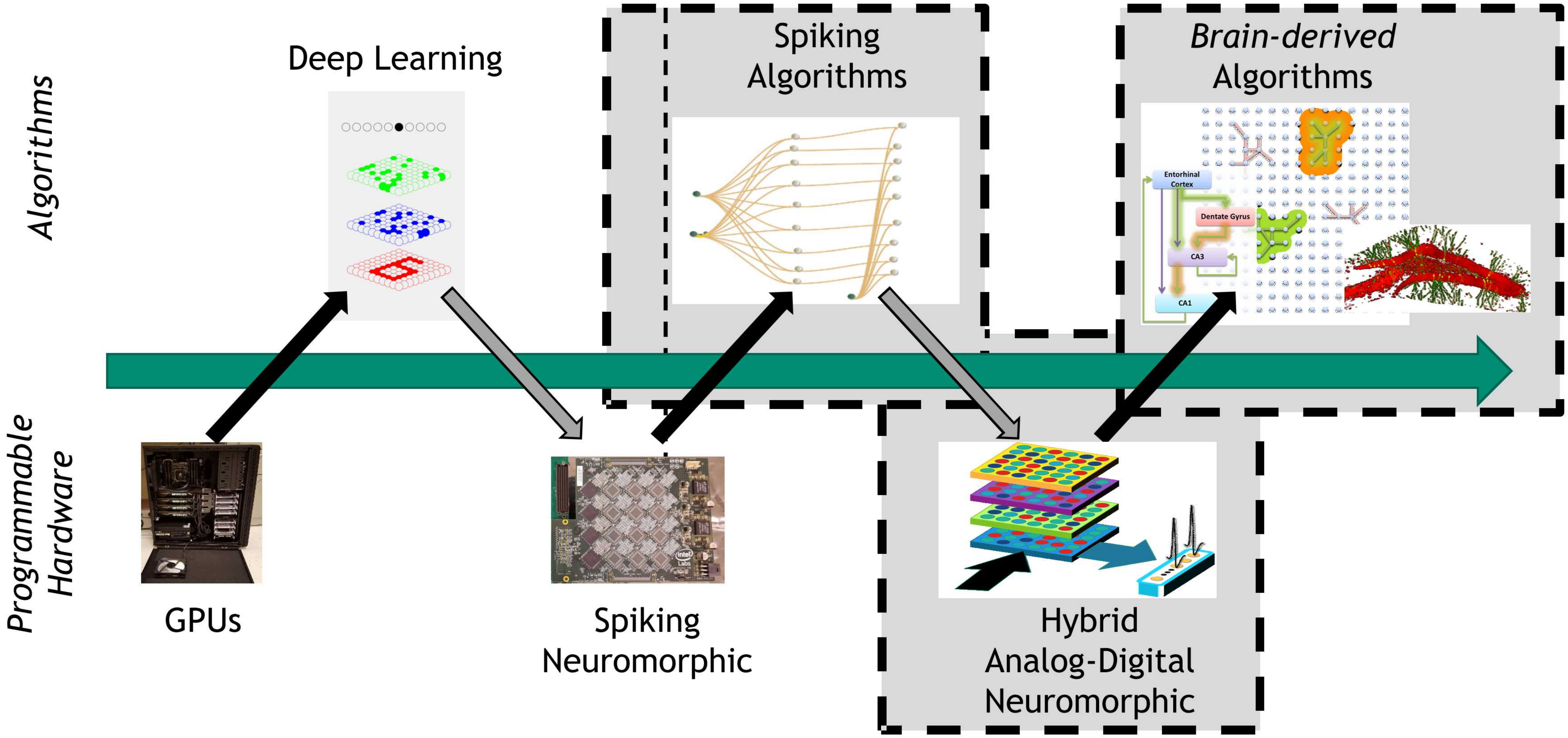


- Deep neural networks benefit from *high volume* of relatively *low-dimensional* data
  - *N>>d*
  - Good (necessary?) for training very large, relatively model-free networks
- Many applications will have *low-volume* or a *skewed-distribution* of relatively *high-dimensional* data
  - Few labels, expensive experiments, changing world, needle-in-haystack, etc.
  - *N≈d,* or *n≈d,* where *n* are relevant observations
  - Not a good fit for large unstructured parameterizable ANNs

The brain exhibits plasticity at many scales



Can we really get the brain into algorithms?



Thanks!

Primary Funding sources

- Sandia Laboratory Directed Research and Development Program
- DOE NNSA Advanced Simulation and Computing

Sandia NERL team

- Brad Aimone, Suma Cardwell, Frances Chance, Srideep Musuvathy, Fred Rothganger, William Severa, Craig Vineyard, Darby Smith, Corinne Teeter, Felix Wang, Ryan Dellana, Mark Plagge

References

- Aimone JB, Neural Algorithms and Computing Beyond Moore's Law; *Communications of ACM,* April 2019
- Aimone JB, A Roadmap for Reaching the Potential of Brain-Derived Computing; *Advanced Intelligent Systems*, in press 2020