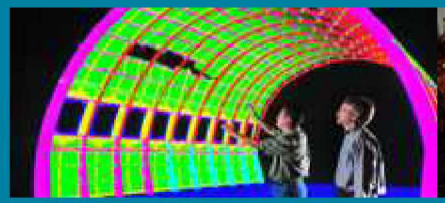
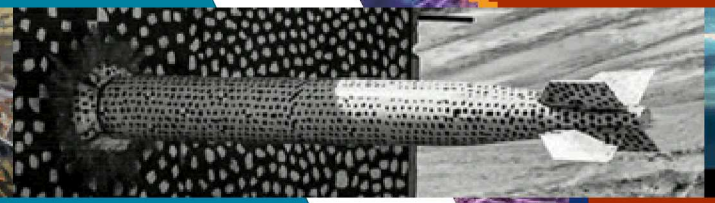
SAND2020-0879C

Exceptional service in the national interest








Selected AI/ML Efforts At Sandia National Laboratories

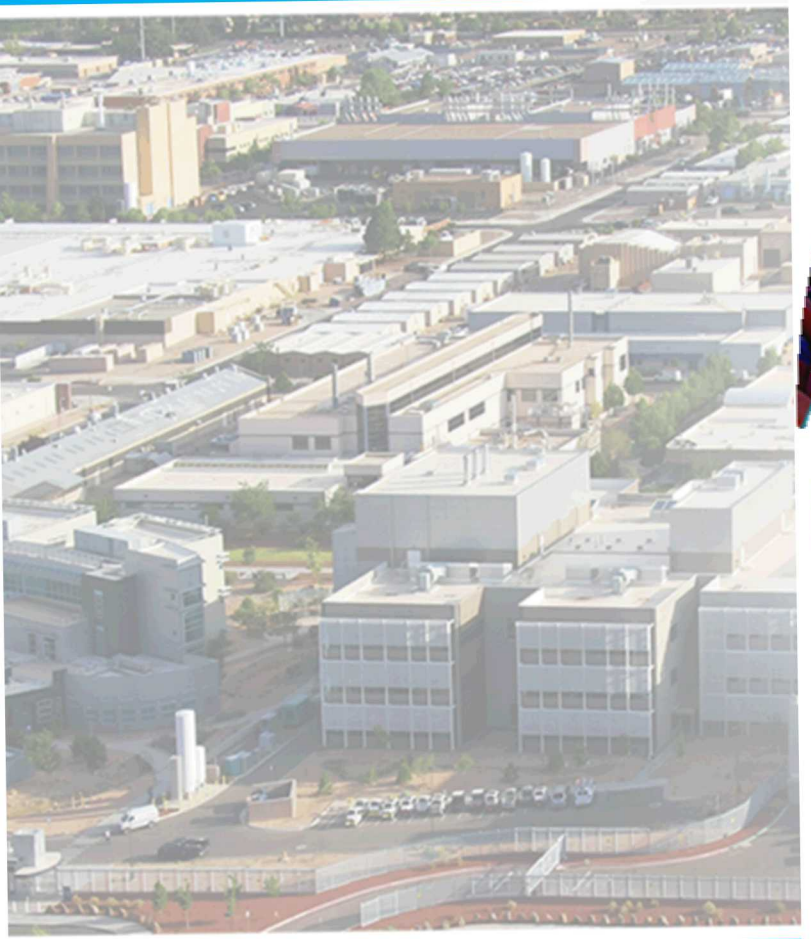
Machine Learning in the Presence of Noise : Early Experiments

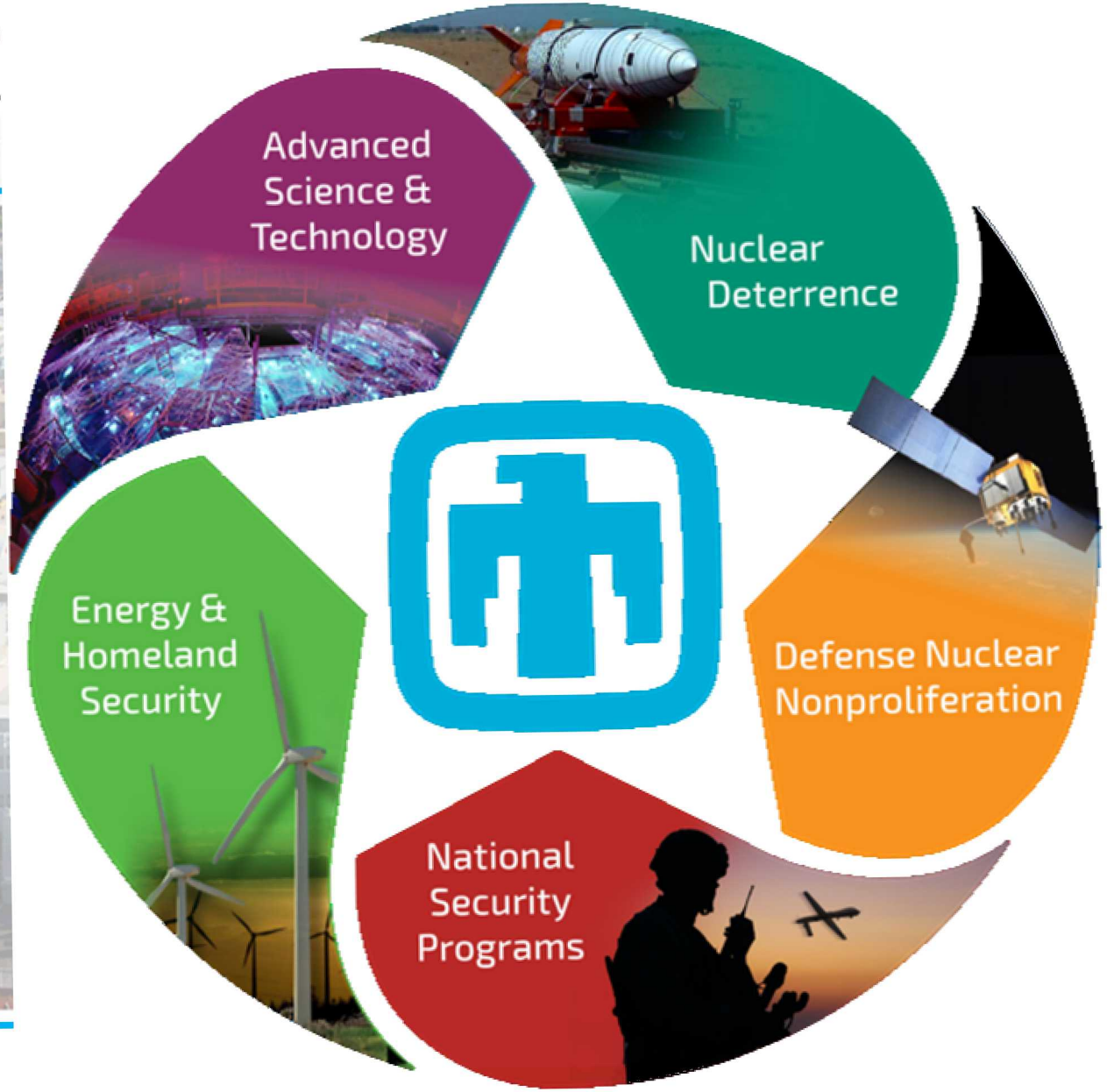
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Presented by: Siva Rajamanickam

srajama@sandia.gov

SANDIA HAS FIVE M/
PROGRAM PORTFOL





Why has ML/AI had so much attention?

Some problems are difficult to solve with a directly-coded algorithm

- Don’t generalize well
- Can be difficult to scale

There have been Machine Learning (ML) successes in a variety of areas

- Recognizing patterns
- Anomaly detection
- Learning predictive models from data
- Creating surrogate models
- Generating synthetic data that models real data
- Assisting human decision making

These successes have been enabled by

- Large curated (labeled) datasets
- Advancements in computing power

Sandia's Unique Mission Needs

Diversity

Scale

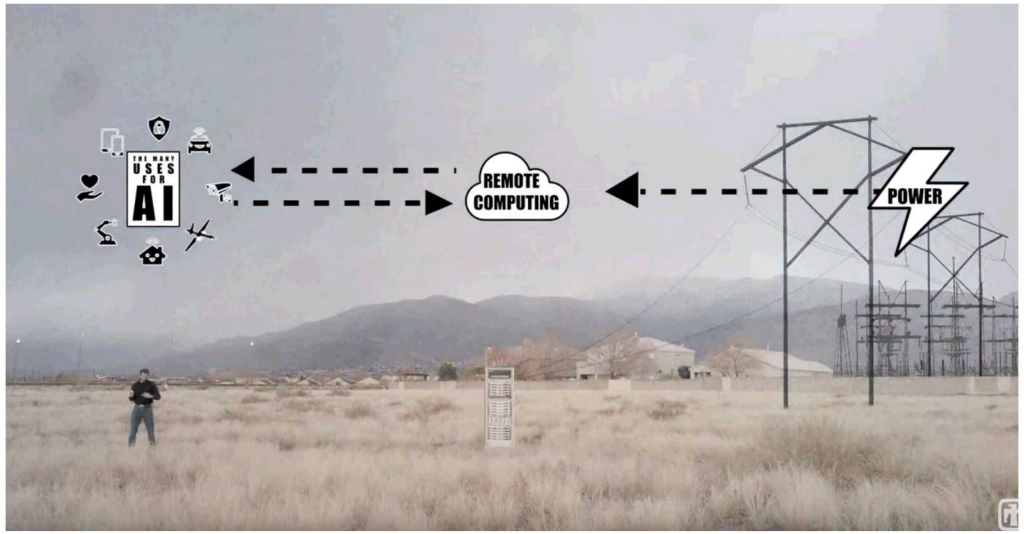
Consequence

Sandia's Unique Mission Needs

Diversity

Scale

Consequence

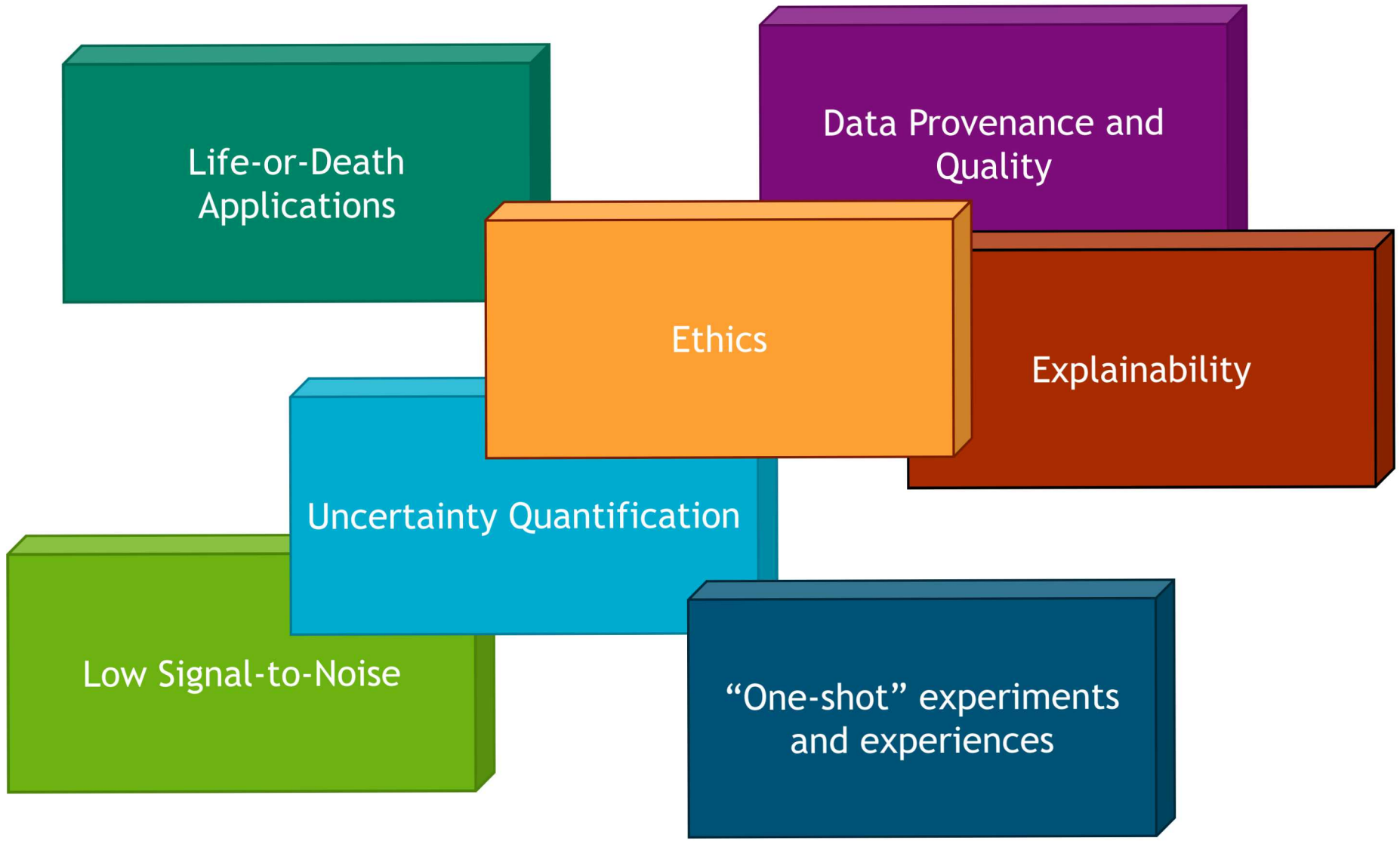


Sandia's Unique Mission Needs

Diversity

Scale

Consequence



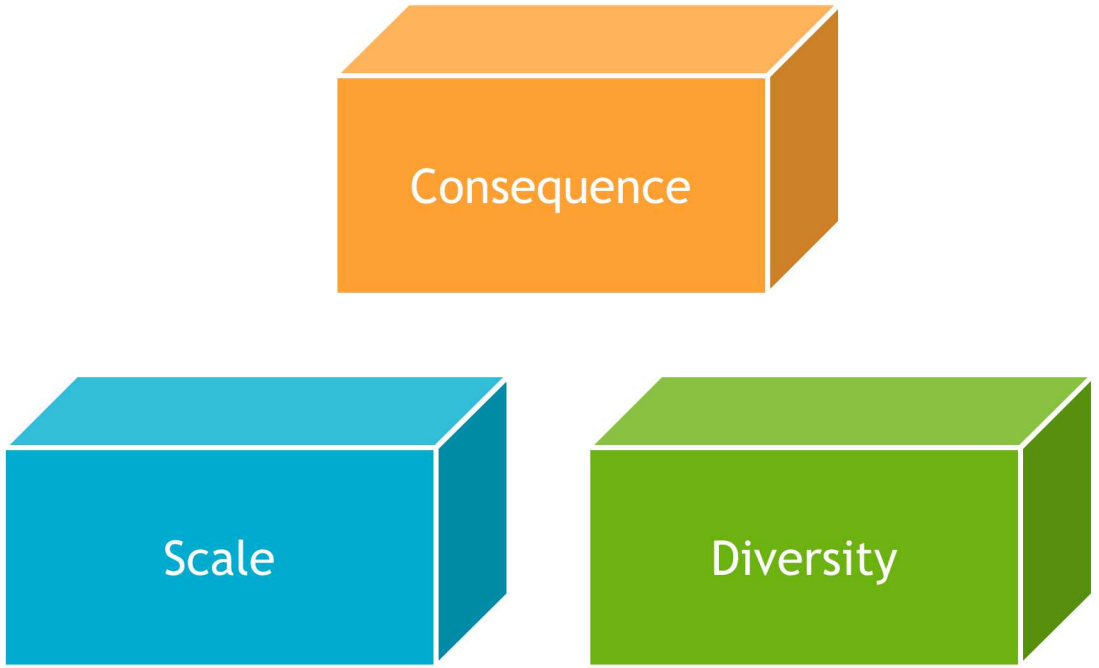
High-confidence decisions

- Typically designing to "Five 9's" of reliability
- Need to assure trust in our solutions
- Need to understand uncertainty of decisions
- Algorithms need to be explainable

classified as rifle
classified as other

Synthesizing Robust Adversarial Examples, Athalye, et.al., 2018

Many Sandia efforts are premised on idea that AI solutions will be instrumental in delivering these requirements



Sandia has a goal of creating a bridge between the broader world of AI and our missions

Extending and developing AI algorithms

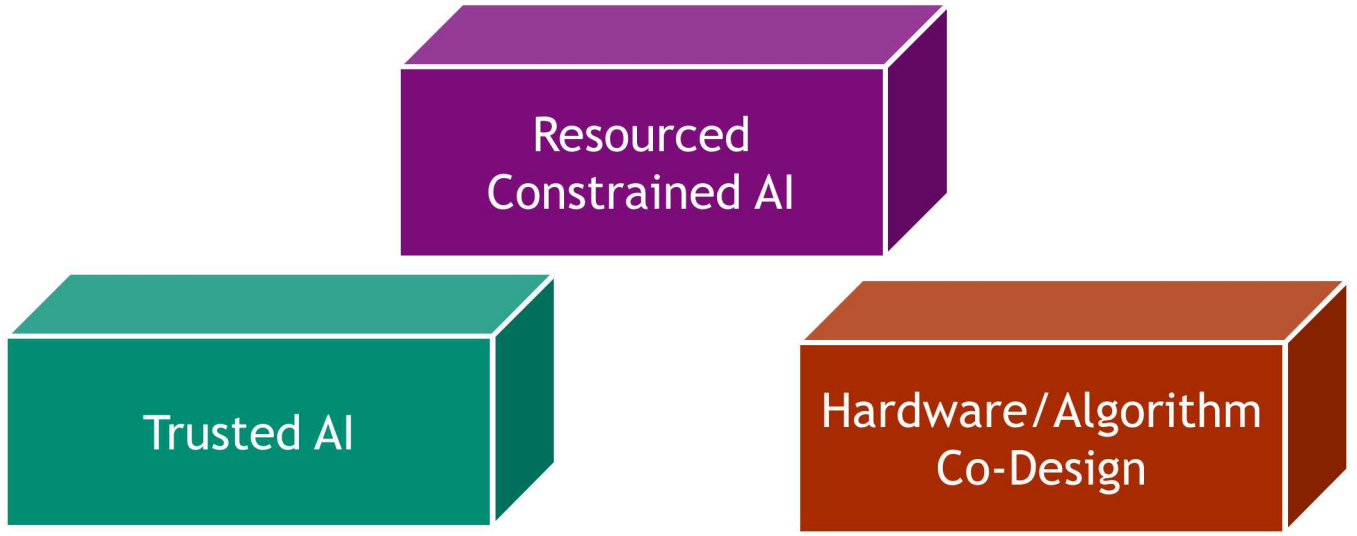
Evaluating novel hardware and accelerators

Explore brain-inspired sensor technology

Identifying opportunities for novel AI impact

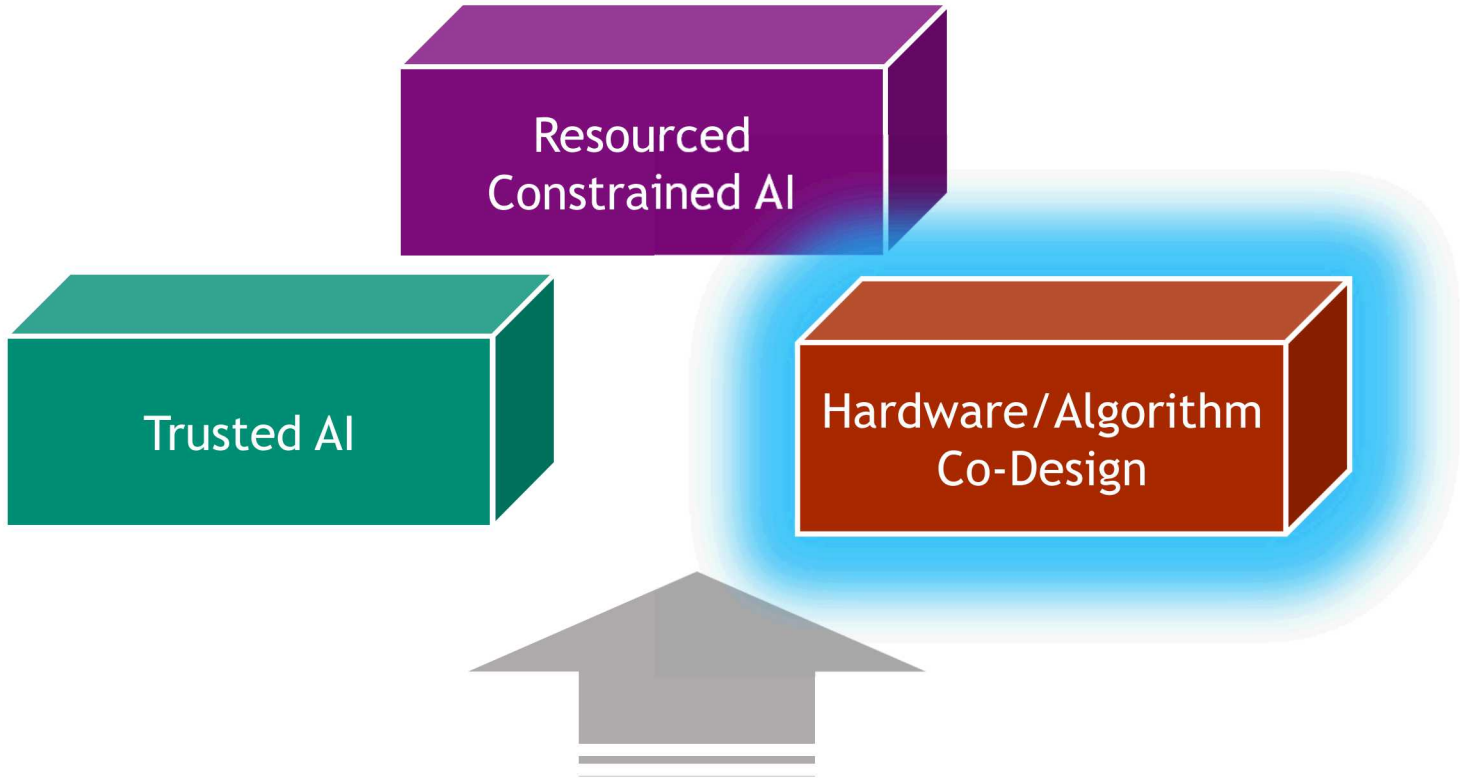
Developing tools and analyses suitable for widespread adoption of emerging AI technologies

Capabilities

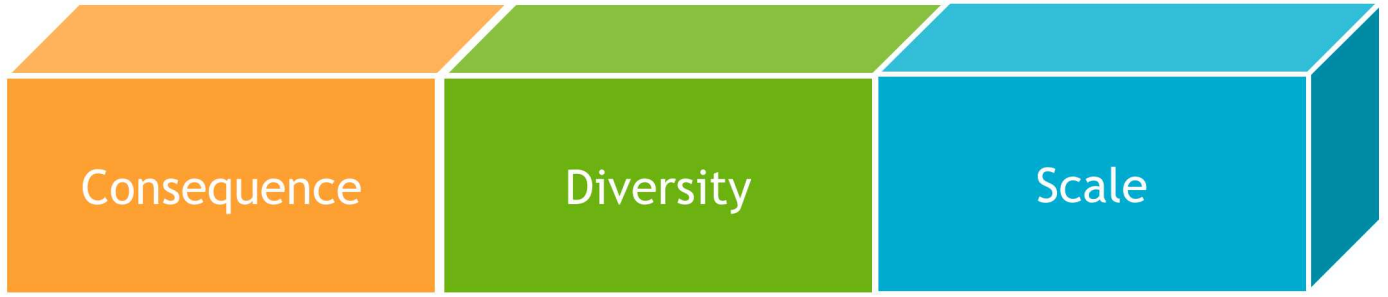


Challenges

Capabilities



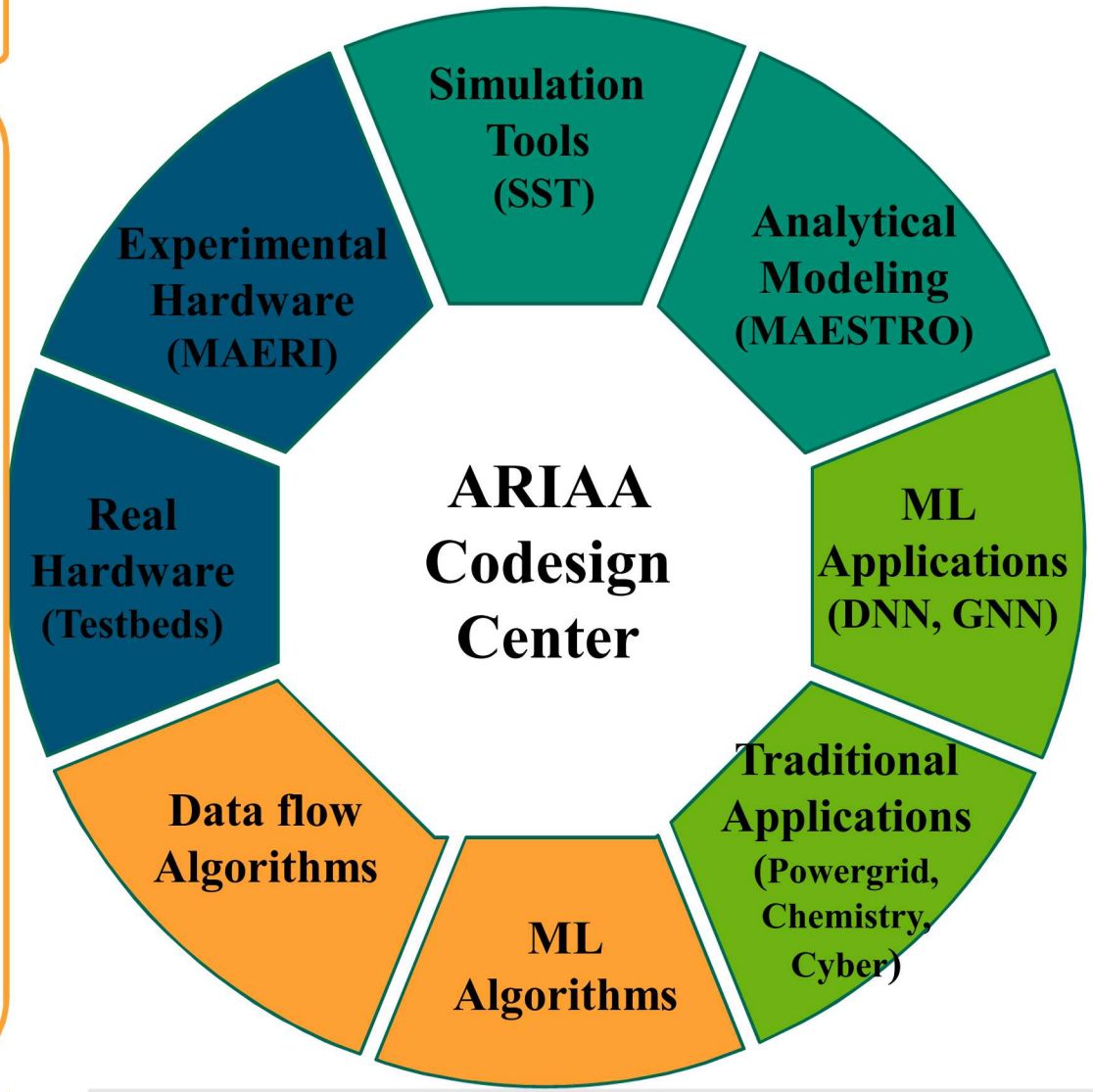
Challenges



Center for co-design of ARtificial Intelligence focused Architectures and Algorithms (ARIAA)

ARIAA Objectives

- **Co-design novel AI/ML architectures and algorithms** to enable traditional and ML-based DOE applications
- Understand potential impact of AI-focused architectures on **future leadership class systems**
- Understand how AI/ML accelerators can work with **sparse, irregular, and/or streaming data**
- **Independently evaluate** AI/ML accelerators from different startups and **nudge them for DOE needs**



Ongoing activities

Architecture

- ✓ SST/MAESTRO integration strategy
- ✓ SST accelerator abstractions
- ✓ Initial evaluation of NVIDIA DNN accelerator

Software

- ✓ Programming abstractions for representative kernels
- ✓ Initial integration of MCL with SST and NVIDIA DNN accelerator

Applications

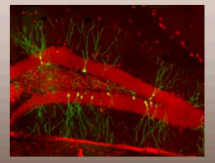
- ✓ Identify first set of ML/Lin Alg. kernels and algorithms
- ✓ Graph coloring, all-to-all hashing (streaming), SchNet

Sandia National Laboratories · Pacific Northwest National Laboratory · Georgia Tech

Codesign of AI/ML accelerators with algorithms and applications will enable the development of this key technology to suite DOE HPC and AI/ML needs

Neural Computing at Sandia Labs Leverages a Large Research Foundation

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

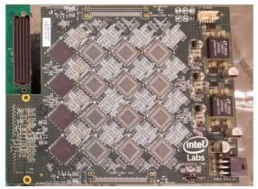
Deployable National Security Applications

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

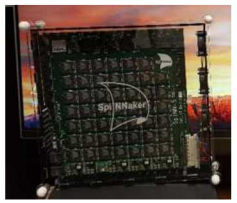
U.S. DEPARTMENT OF ENERGY

- Enables researchers to explore the boundaries of neural computation
- Consists of a variety of neuromorphic hardware & neural algorithms providing a testbed facility for comparative benchmarking and new architecture exploration

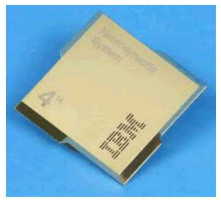
Intel Loihi



SpiNNaker 48 Node Board



IBM TrueNorth*



IBM TrueNorth NS16e*

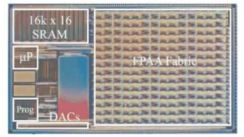
Intel Neural Compute Stick

Google Coral

Google EdgeTPU

Inilabs DAVIS 240C DVS

Georgia Tech FPAA



Intel Loihi

SNL STPU on FPGA

Xilinx PYNQ FPGA

Nengo FPGA

Nvidia Jetson TX1

Nvidia Jetson Nano

GPU Workstations

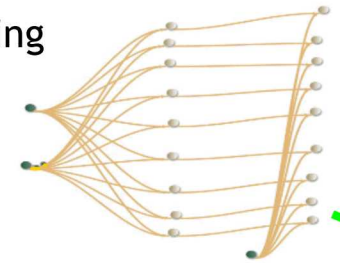
Cognimem CM1K

KnuPath Hermosa

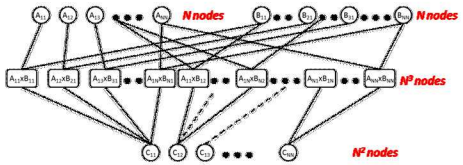
*Remote access

Impacting Broad Areas of Computation

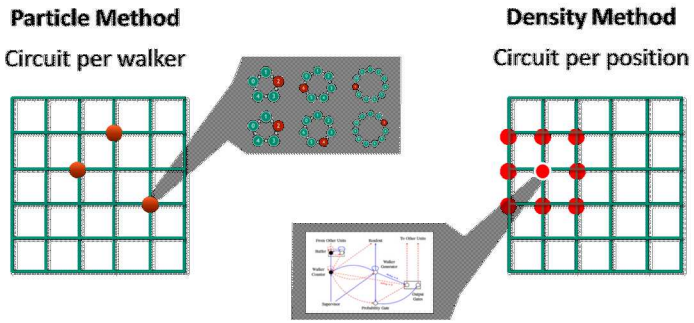
Pattern Matching



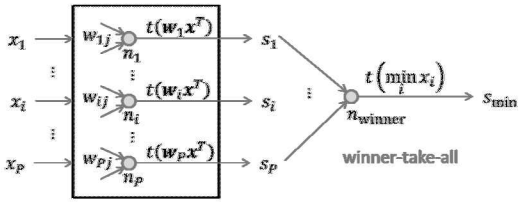
Linear Algebra

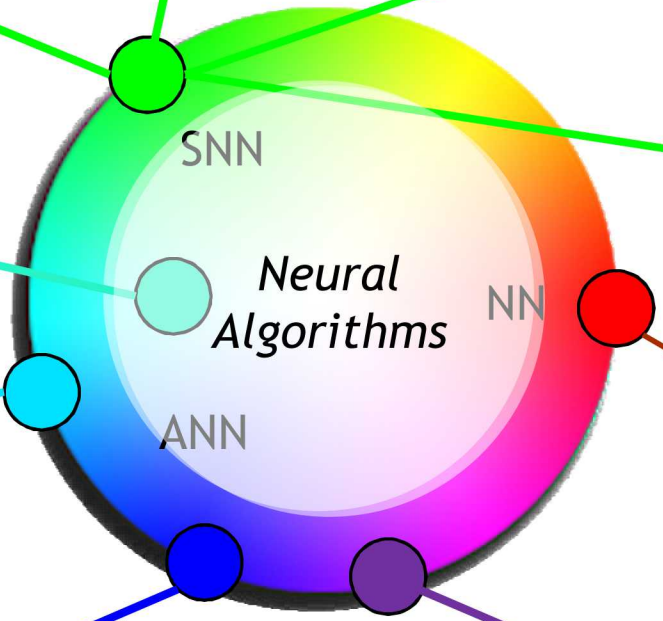


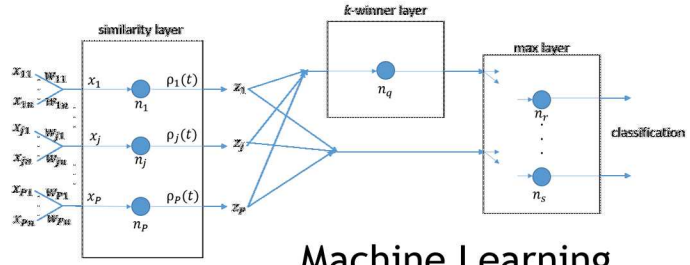
Scientific Computing



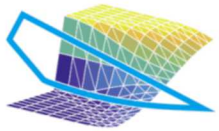
Optimizations



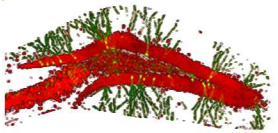




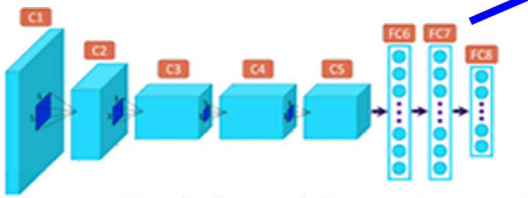
Machine Learning



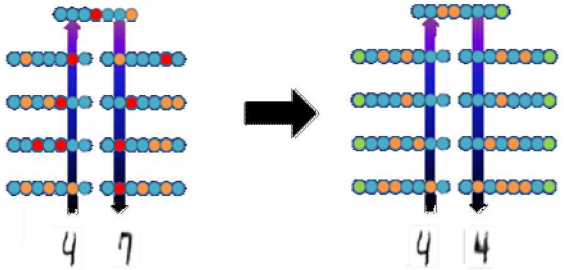
WHETSTONE



Intelligent Storage

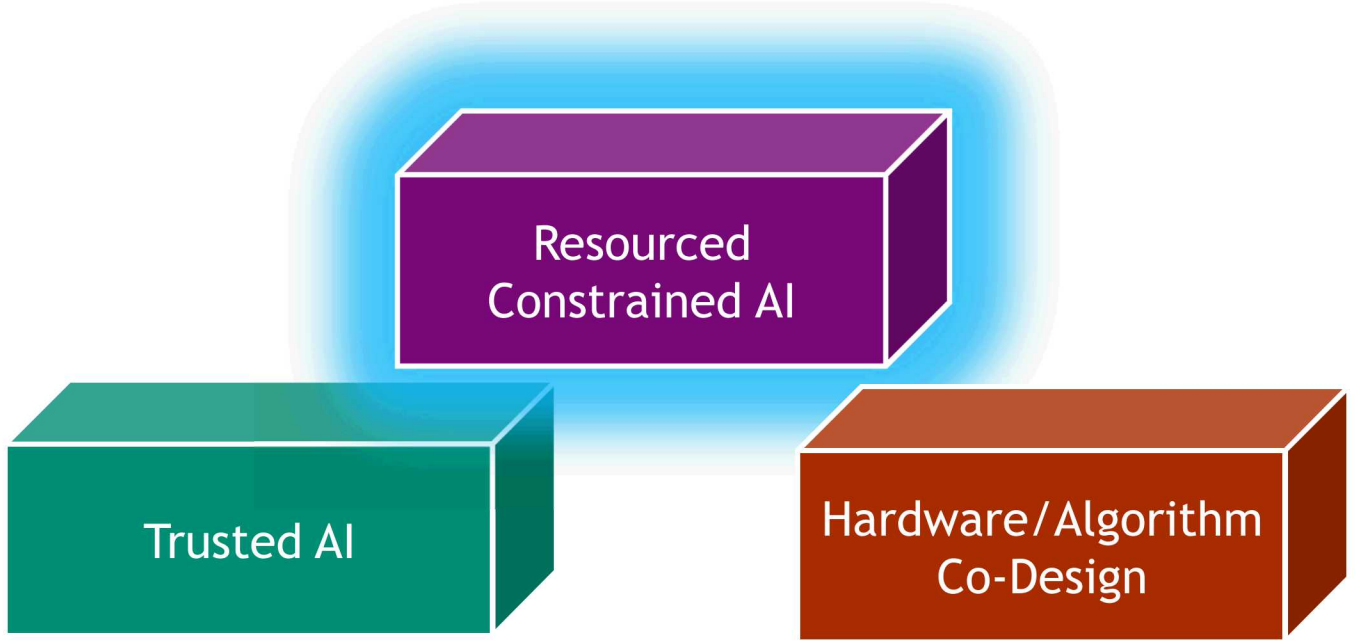


Context Modulated Deep Learning



Adaptive Deep Learning

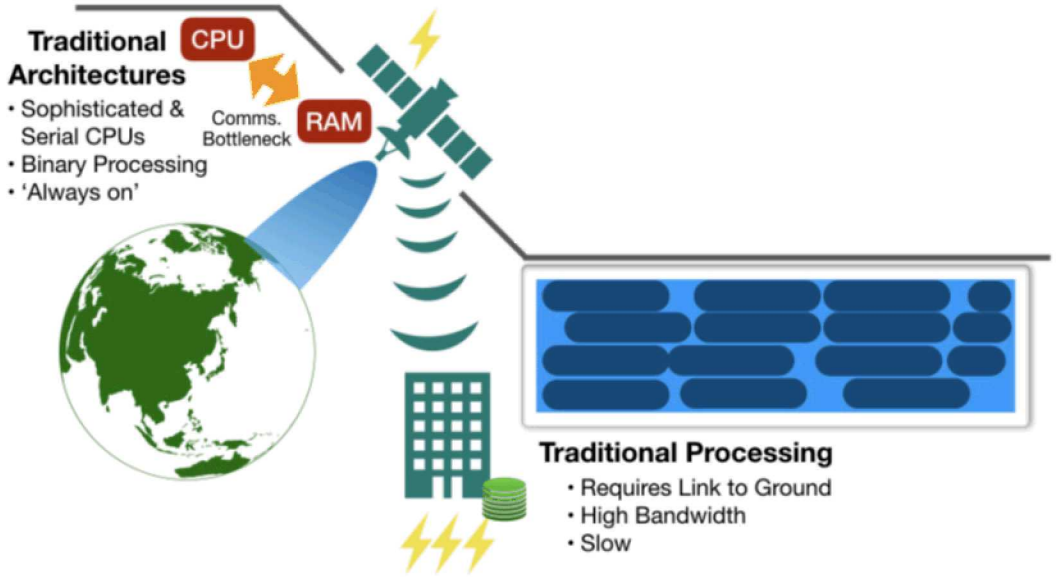
Capabilities

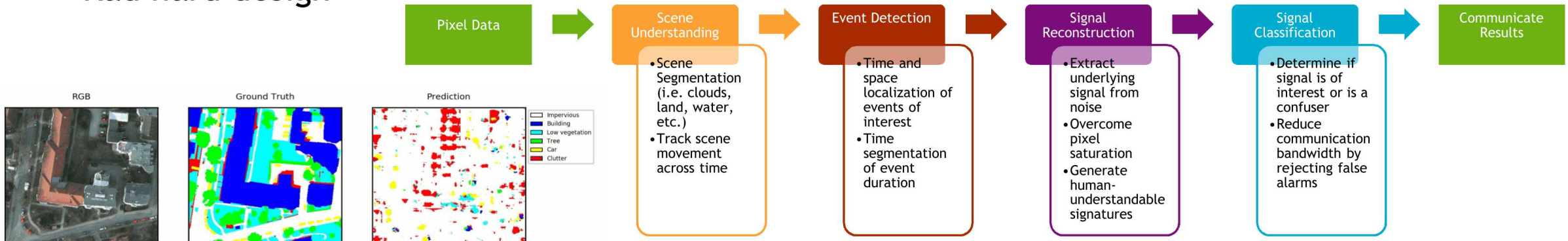


Challenges

Challenges in classic remote sensing

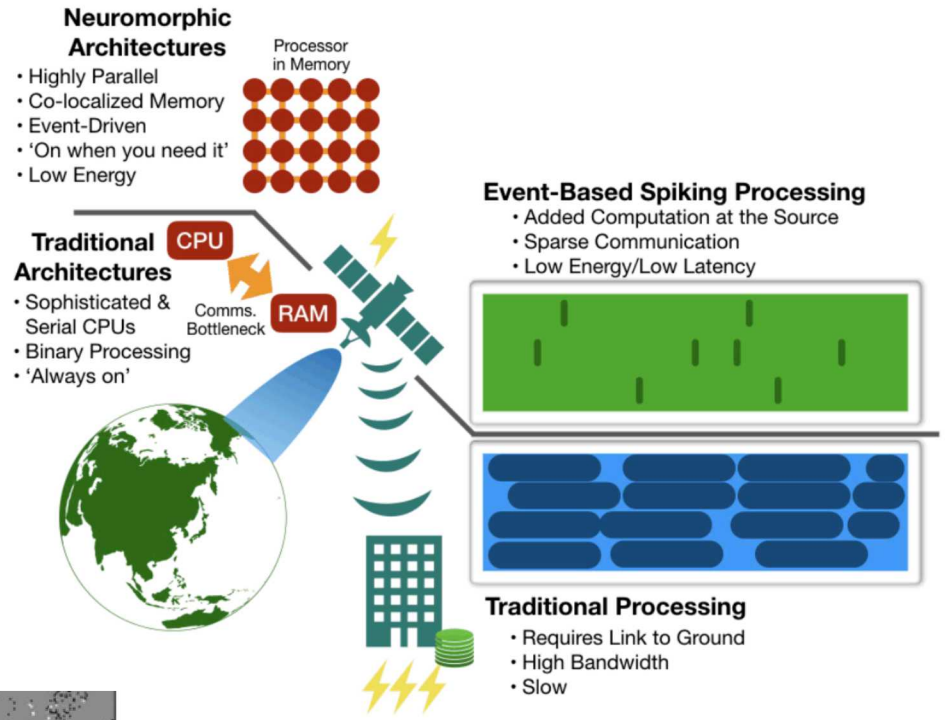
- Growth of sensor technologies outpacing communication bandwidth
- High Consequence Decisions
- Limited algorithm capabilities
- Limited onboard processing capability
- Rad hard design

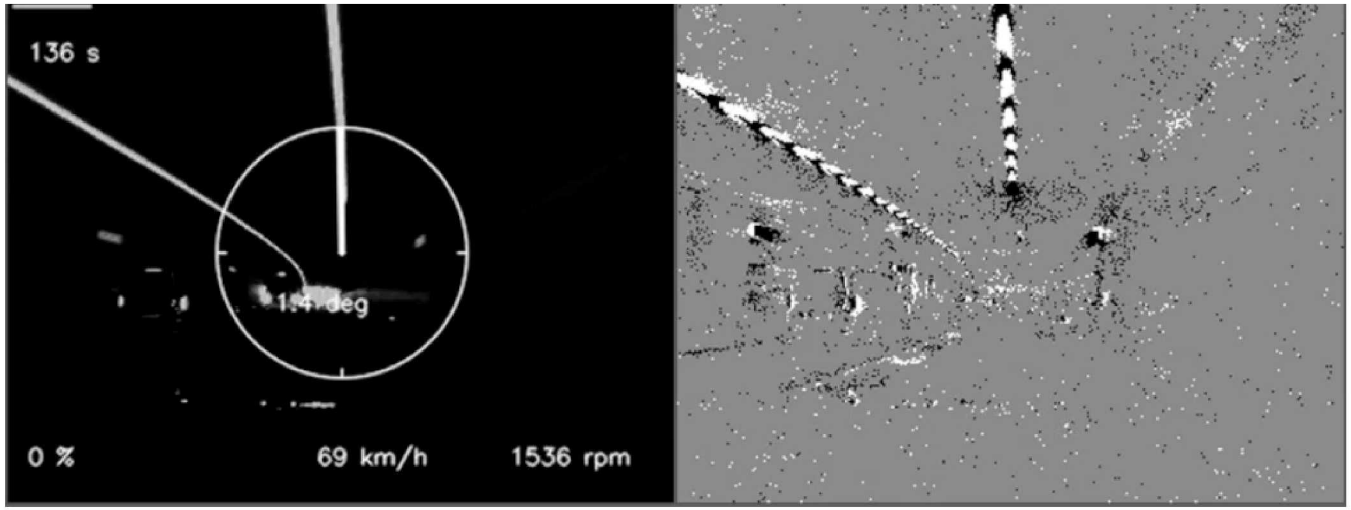




Compute at the sensor

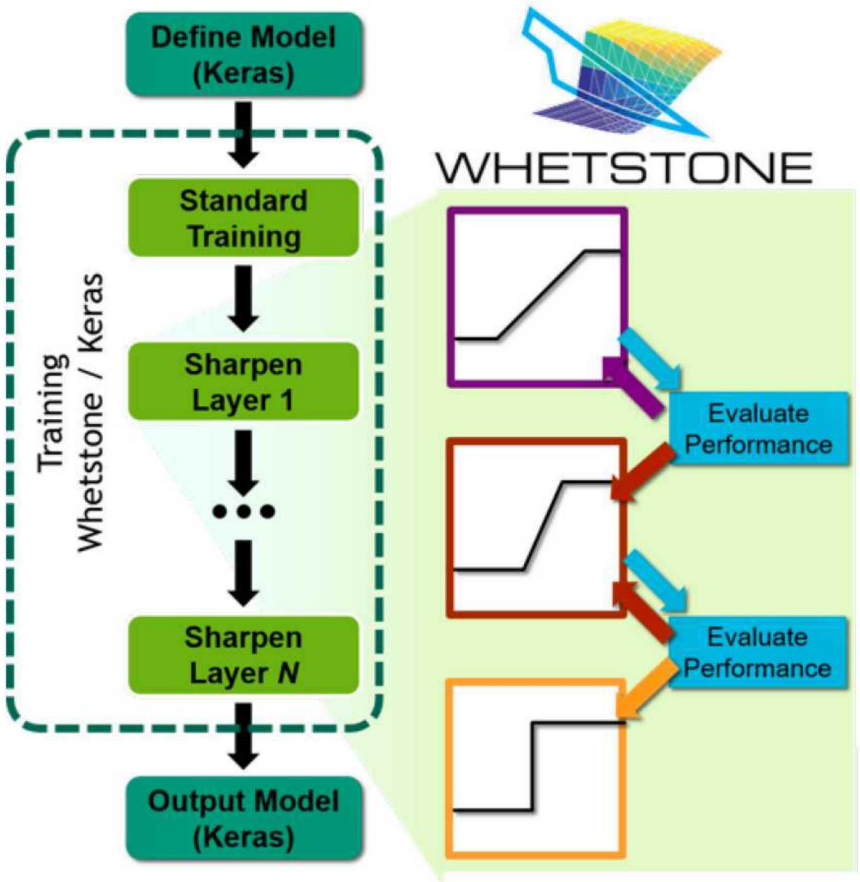
- Improve bandwidth utilization (send only what you need)
- Distributed computation avoiding single-point of failure
- May reduce preprocessing required (e.g. whitening)



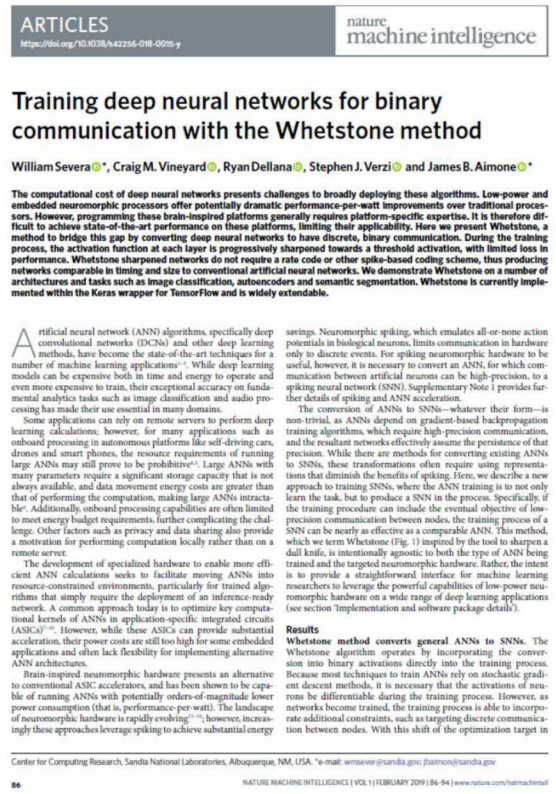


DAVIS 240C Event-Driven Camera

Whetstone for Low-Power Spiking Deep Learning



- Automatically converts deep learning networks from continuous valued neurons to binary activations, making them compatible with neuromorphic hardware
- Open sourced
- Beginning to port onto neuromorphic platforms
  - SpiNNaker Results look great



ARTICLES https://doi.org/10.1038/s42256-018-0015-y

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*

The computational cost of deep neural networks presents challenges to broadly deploying these algorithms. Low-power and embedded neuromorphic processors offer potentially dramatic performance-per-watt improvements over traditional processors. However, programming these brain-inspired platforms generally requires platform-specific expertise. It is therefore difficult to achieve state-of-the-art performance on these platforms, limiting their applicability. Here we present Whetstone, a method to bridge this gap by converting deep neural networks to have discrete, binary communication. During the training process, the activation function at each layer is progressively sharpened towards a threshold activation, with limited loss in performance. Whetstone sharpened networks do not require a rate code or other spike-based coding scheme, thus producing networks comparable in timing and size to conventional artificial neural networks. We demonstrate Whetstone on a number of architectures and tasks such as image classification, autoencoders and semantic segmentation. Whetstone is currently implemented within the Keras wrapper for TensorFlow and is widely extendable.

Artificial neural network (ANN) algorithms, specifically deep convolutional networks (DCNs) and other deep learning methods, have become the state-of-the-art techniques for a number of machine learning applications[1–3]. While deep learning models can be expensive both in time and energy to operate and even more expensive to train, their exceptional accuracy on fundamental analytics tasks such as image classification and audio processing has made their use essential in many domains.

Some applications can rely on remote servers to perform deep learning calculations; however, for many applications such as onboard processing in autonomous platforms like self-driving cars, drones and smart phones, the resource requirements of running large ANNs may still prove to be prohibitive[4,5]. Large ANNs with many parameters require a significant storage capacity that is not always available, and data movement energy costs are greater than that of performing the computation, making large ANNs intractable[6]. Additionally, onboard processing capabilities are often limited to meet energy budget requirements, further complicating the challenge. Other factors such as privacy and data sharing also provide a motivation for performing computation locally rather than on a remote server.

The development of specialized hardware to enable more efficient ANN calculations seeks to facilitate moving ANNs into resource-constrained environments, particularly for trained algorithms that simply require the deployment of an inference-ready network. A common approach today is to optimize key computational kernels of ANNs in application-specific integrated circuits (ASICs)[7–10]. However, while these ASICs can provide substantial acceleration, their power costs are still too high for some embedded applications and often lack flexibility for implementing alternative ANN architectures.

Brain-inspired neuromorphic hardware presents an alternative to conventional ASIC accelerators, and has been shown to be capable of running ANNs with potentially orders-of-magnitude lower power consumption (that is, performance-per-watt). The landscape of neuromorphic hardware is rapidly evolving[11–14]; however, increasingly these approaches leverage spiking to achieve substantial energy savings. Neuromorphic spiking, which emulates all-or-none action potentials in biological neurons, limits communication in hardware only to discrete events. For spiking neuromorphic hardware to be useful, however, it is necessary to convert an ANN, for which communication between artificial neurons can be high-precision, to a spiking neural network (SNN). Supplementary Note 1 provides further details of spiking and ANN acceleration.

The conversion of ANNs to SNNs—whatever their form—is non-trivial, as ANNs depend on gradient-based backpropagation training algorithms, which require high-precision communication, and the resultant networks effectively assume the persistence of that precision. While there are methods for converting existing ANNs to SNNs, these transformations often require using representations that diminish the benefits of spiking. Here, we describe a new approach to training SNNs, where the ANN training is to not only learn the task, but to produce a SNN in the process. Specifically, if the training procedure can include the eventual objective of low-precision communication between nodes, the training process of a SNN can be nearly as effective as a comparable ANN. This method, which we term Whetstone (Fig. 1) inspired by the tool to sharpen a dull knife, is intentionally agnostic to both the type of ANN being trained and the targeted neuromorphic hardware. Rather, the intent is to provide a straightforward interface for machine learning researchers to leverage the powerful capabilities of low-power neuromorphic hardware on a wide range of deep learning applications (see section 'Implementation and software package details').

Results

Whetstone method converts general ANNs to SNNs. The Whetstone algorithm operates by incorporating the conversion into binary activations directly into the training process. Because most techniques to train ANNs rely on stochastic gradient descent methods, it is necessary that the activations of neurons be differentiable during the training process. However, as networks become trained, the training process is able to incorporate additional constraints, such as targeting discrete communication between nodes. With this shift of the optimization target in

Center for Computing Research, Sandia National Laboratories, Albuquerque, NM, USA. *e-mail: wmsever@sandia.gov; jbaimon@sandia.gov

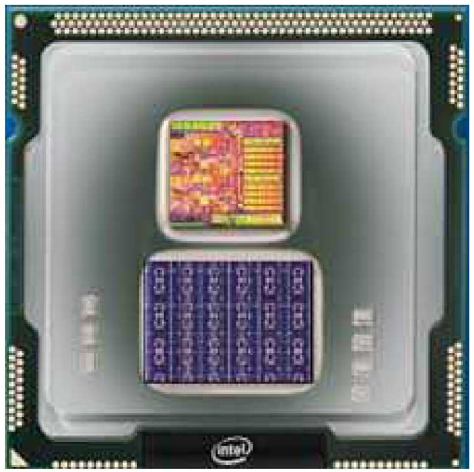
POC: W. Severa

Whetstone and Vehicle Perception

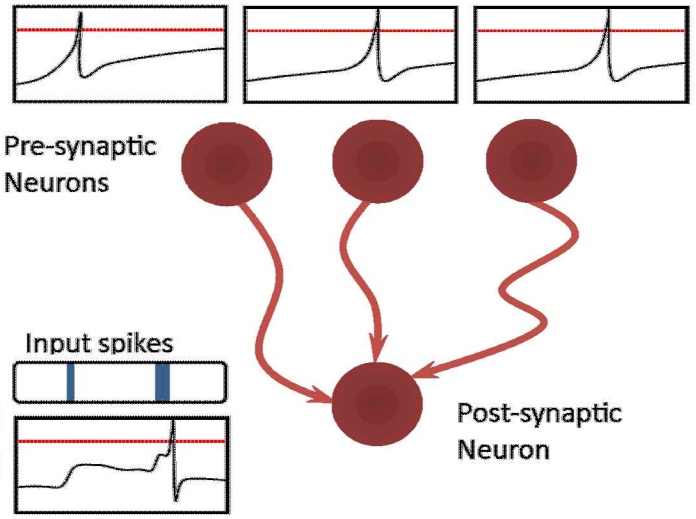
- AI power draw is a key limiting factor especially for electric powered vehicles: 3kW now; HPC-level for fully self-driving
- Prototype vehicles use a trunk full of GPUs
- Forecasting current tech ~1TeraOp/Watt
- Neuromorphic Hardware:
  - Enables event-driven computation
  - Opportunity for extremely low power consumption

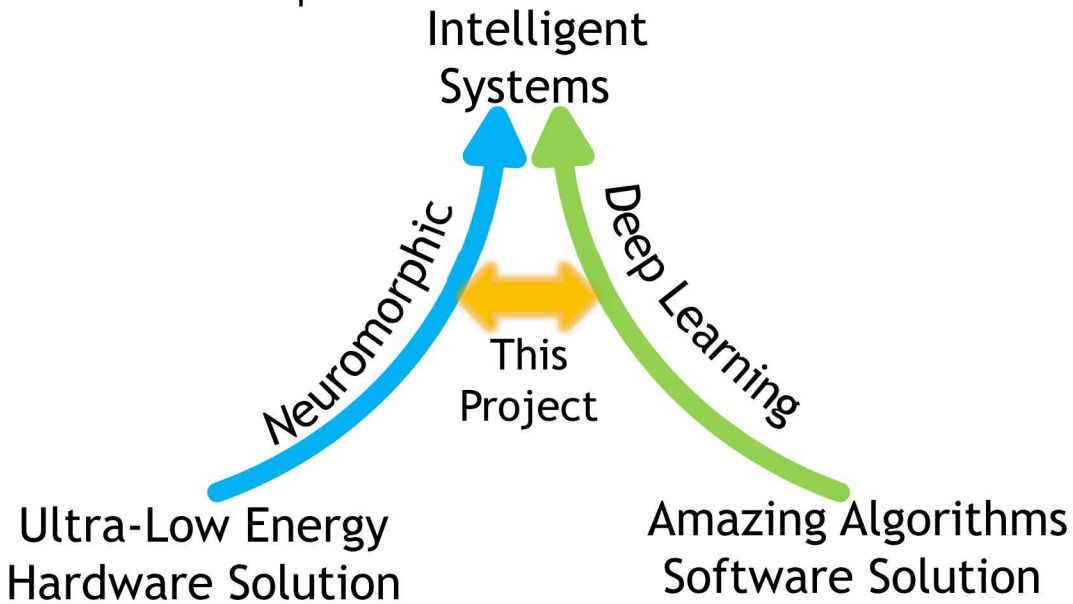
SpiNNaker, Univ. of Manchester

Example Image
From Berkeley DeepDrive



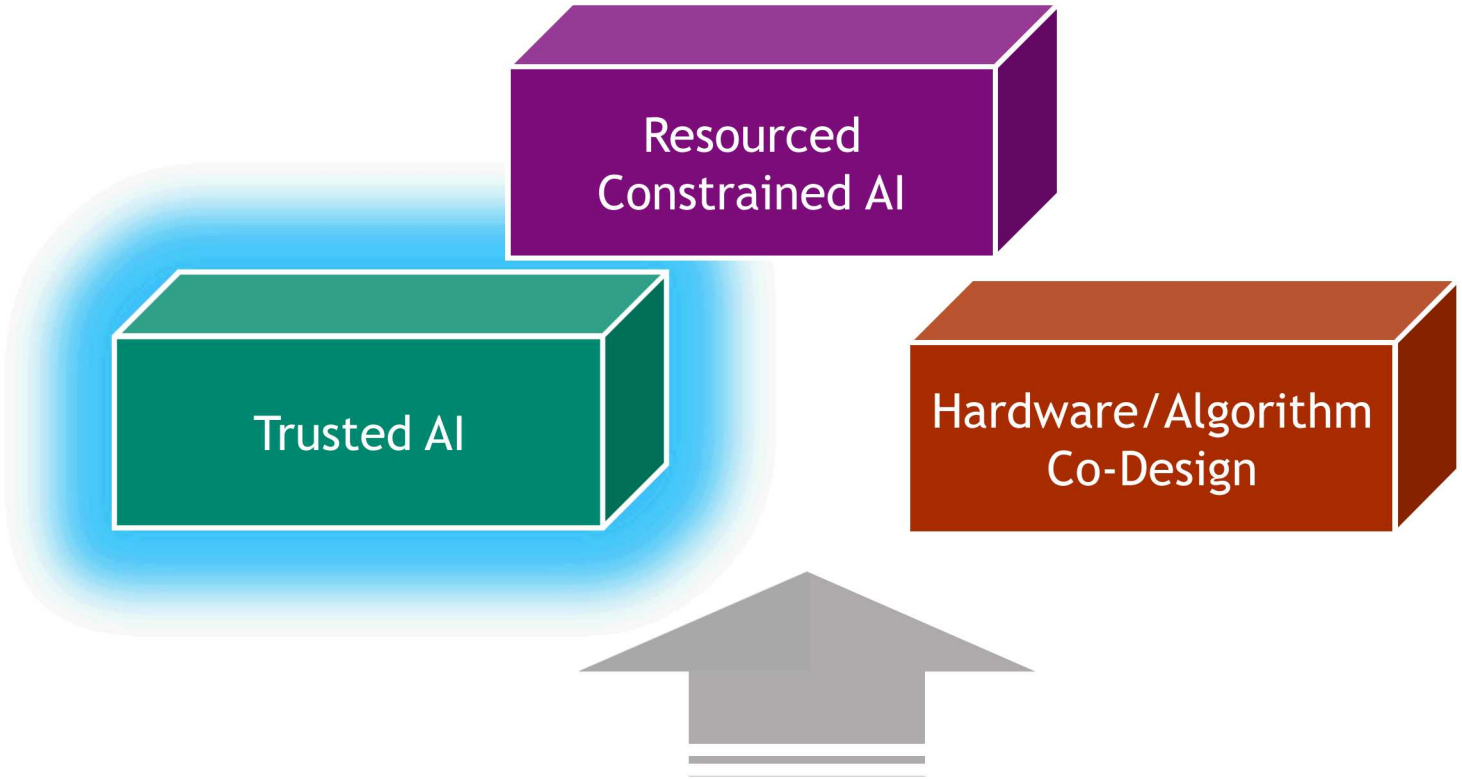
Intel Loihi
Photo: intel.com



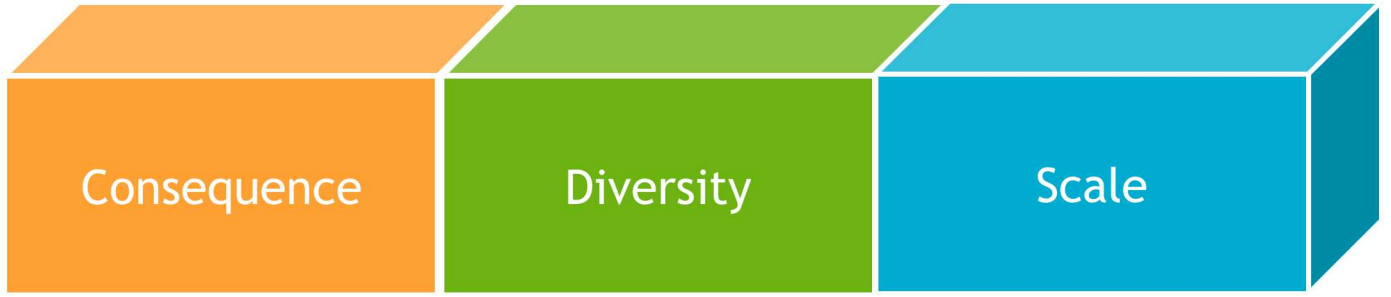


POC: W. Severa

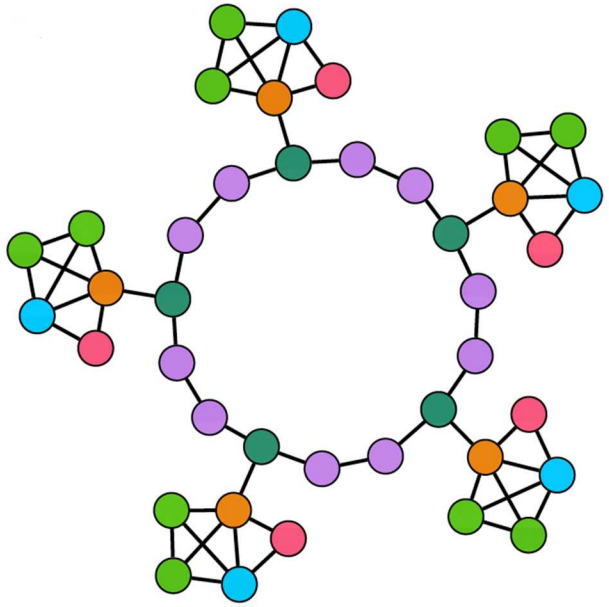
Capabilities



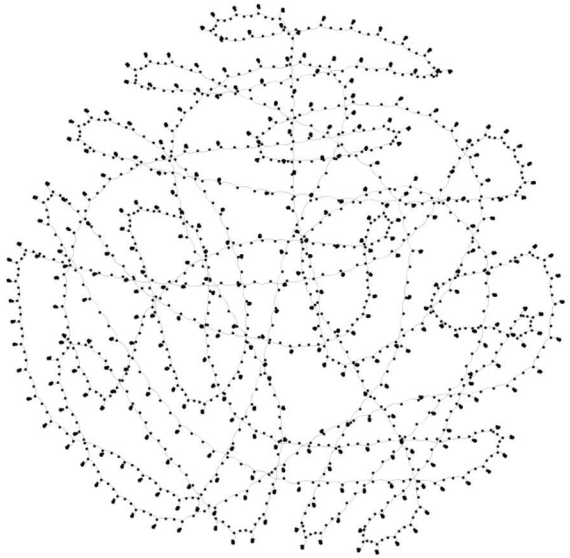
Challenges



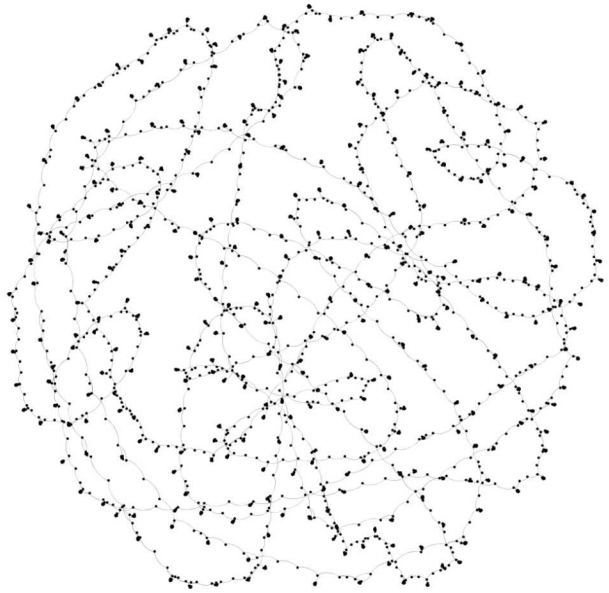
Graph Neural Networks



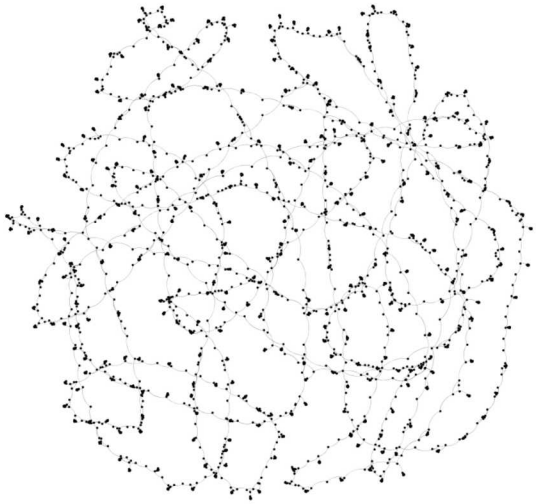
Ring of Houses



Larger graph



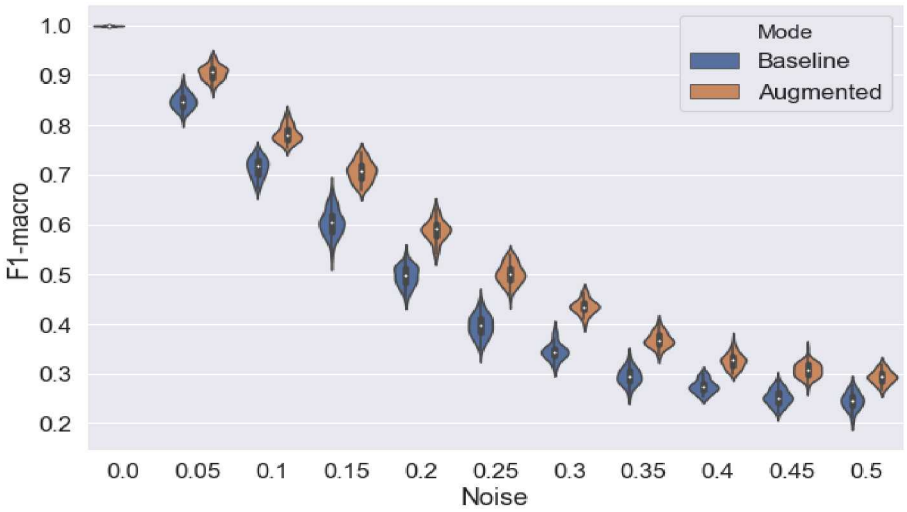
Distance-2 noise



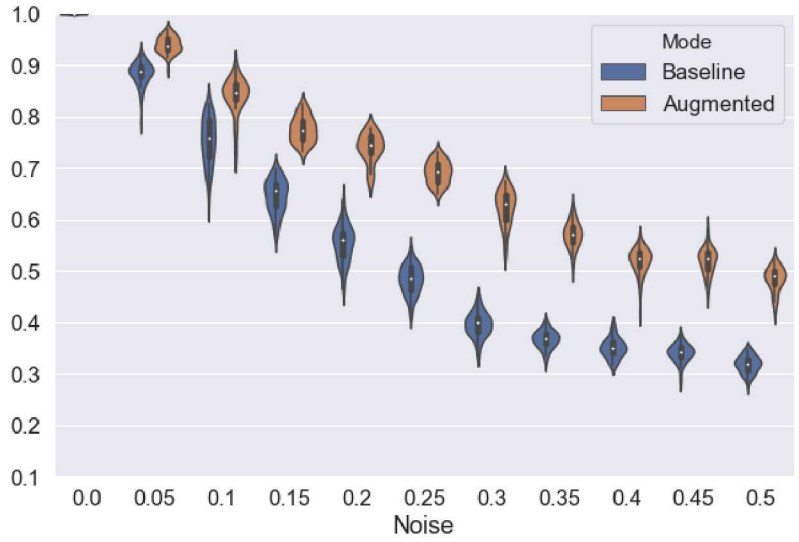
Distance-3 noise

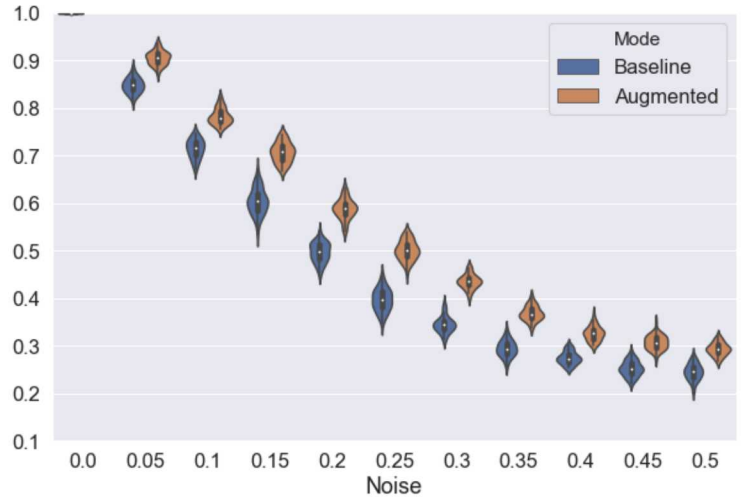
- Graph Neural Networks (GNNs) are a powerful abstraction for learning embeddings on graph structured data
- GNNs have been used in several domains including drug discovery, material science, molecular toxicity prediction
- Evaluate a powerful GNN (Xu et al. 2018) in the presence of noise

Graph Neural Networks – Noisy Data



Test F1 score of GIN model with varying levels of structural noise added to input graph, across 3 different modes of noise constraint.





Augmented vs. non-augmented training (baseline) for node classification on Gp. Y-axis is F1 score, x-axis is random edge addition ratio.

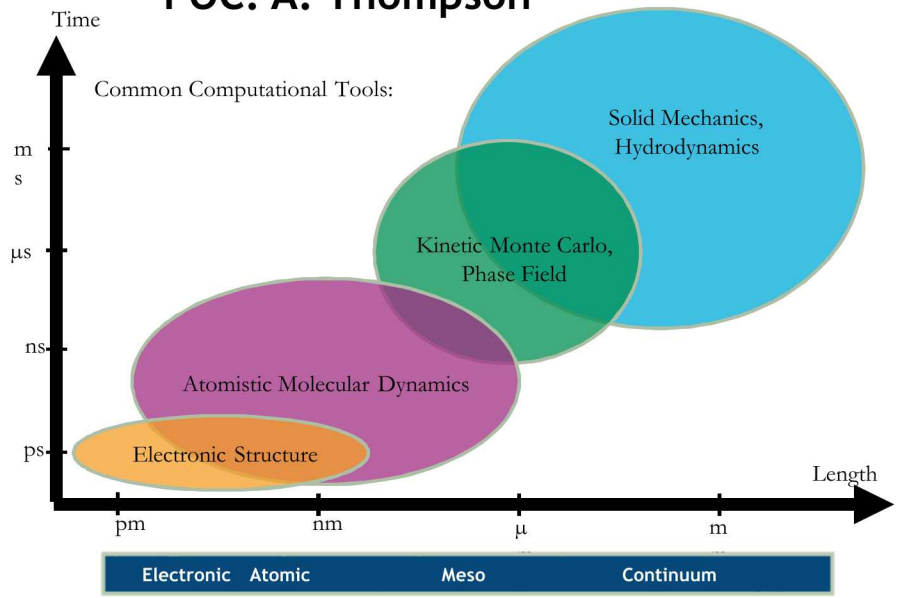
- GNNs can predict the six classes with perfect accuracy with no noise
- The class prediction accuracy drops quite fast even at the presence of small amount of noisy edges 0.1-0.15
- The prediction accuracy can be improved by training on augmented noisy graphs

"How Robust are Graph Neural Networks to Structural Noise", J. Fox, S. Rajamanickam, DLGMA workshop, AAAI 2019.

SNAP (Spectral Neighbor Analysis Potential): Machine Learned Interatomic Potential

- Continuum models require underlying models of the materials behavior
- QM, MD, KMC, SM: Each method makes different approximations and captures different physics
- Quantum methods can provide very complete description for 100s of atoms
- Molecular Dynamics acts as the "missing link"
  - Bridges between quantum and continuum models
  - Moreover, extends quantum accuracy to continuum length scales; retaining atomistic information

POC: A. Thompson

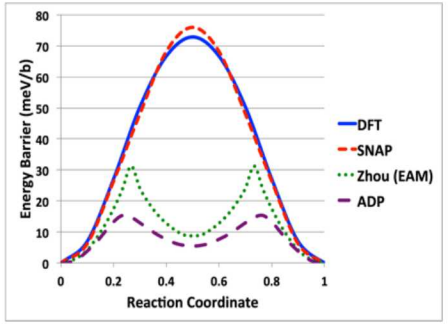


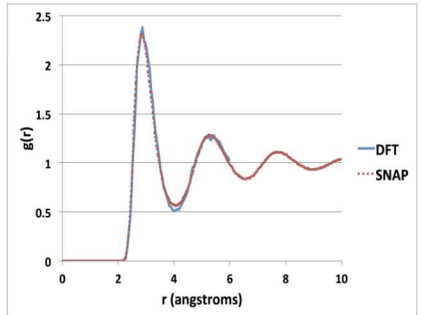
QM: $< 10^{-3}$ μm, $< 10^{-5}$μs, *MD:* $< 10^{1}$ μm, $< 10^{1}$μs, *Meso:* ~1μm, ~1μs, *Continuum:* > 1μm, > 1μs

Current Areas of SNAP Development

- Plasticity in Tantalum
- Plasma-Facing Materials (SciDAC-4): Tungsten/Beryllium, W/Be/H (in progress)
- Radiation Damage in III-V Semiconductors: New Multi-element SNAP formulation for Indium Phosphide
- SNAP Accuracy: Quadratic SNAP, SNAP + Neural Networks, Better descriptors
- SNAP Computational Speed in LAMMPS: SNAP with KOKKOS (CPU, GPU,...) [ECP CoPA project], Exploring new GPU algorithms [NERSC/NESAP]

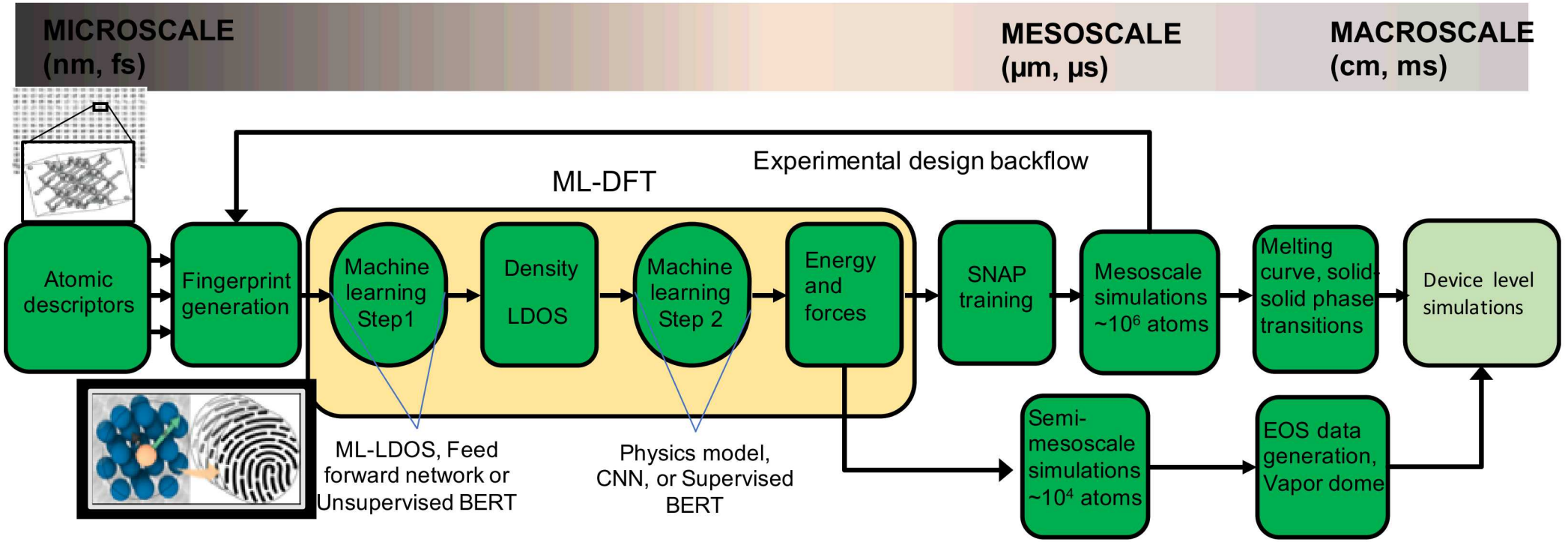
SNAP potential agrees well with DFT calculations





A. P. Thompson , L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker, *J. Comp. Phys.*, **285** 316 (2015).

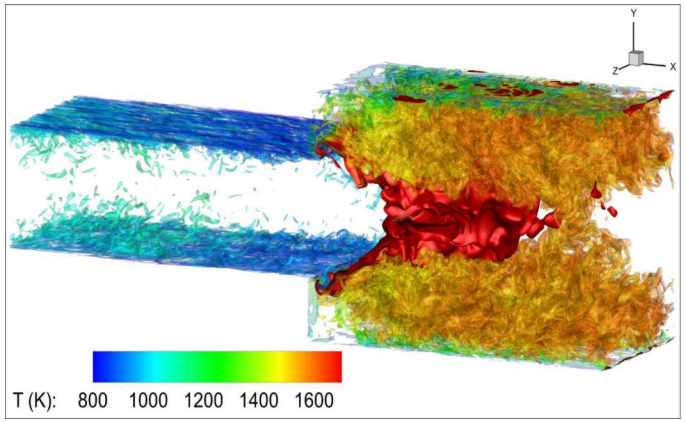
ML-DFT: Accelerating Multiscale Materials Modeling with Machine Learning



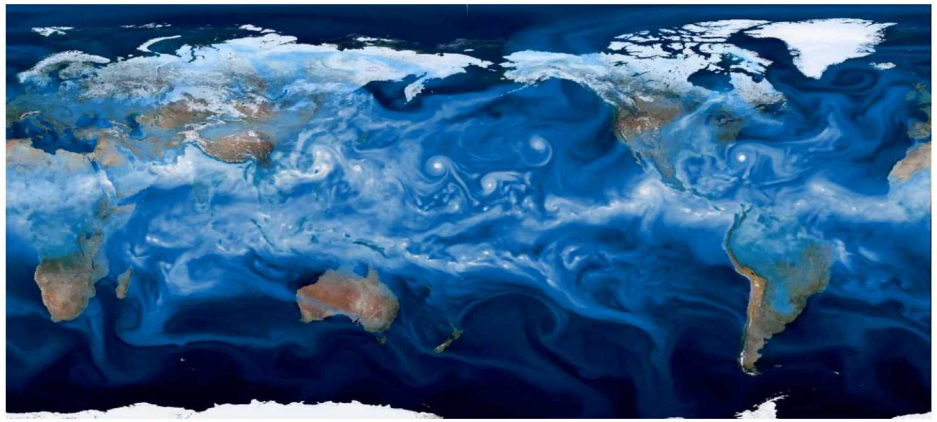
- **Physics-informed Machine Learning (ML)** model based Multiscale Materials Modeling (MMM) toolchain.
- Accelerate first-principles data generation, and increase fidelity and robustness of predictive atomistic material simulations.
- **Applications**: rad-hard semiconductors, advanced manufacturing, shock compression, and energetic materials.
- ML to accelerate *interpolation* of microscale data (10^2 atoms) and enable *extrapolation* to mesoscale (10^4 atoms)
- *High fidelity training data generation*, *optimal experimental design* to select ML training data, *three ML approaches*, and extensions to macroscopic scale.

"Trilinos for (ML-based) Material simulations"

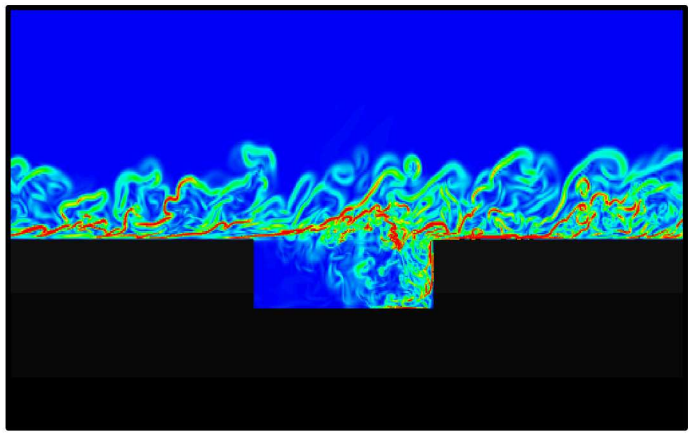
Finding Needles in Noisy Data



Combustion



Climate Modeling



Fluid Dynamics

Want to find "interesting" events, anomalies, state changes, etc.

Examples may include cyclones, onset of combustion, or other things that the scientists may not prescribe a priori and may be difficult to perform via rule-based detection

Desired solution would be to take all the data and run the appropriate detection algorithms (e.g., LOF, isolation forests, clustering)

These simulations produce massive amounts of data (problems for storage capacity, bandwidth)

Signatures: A condensed, information-rich, representation of the simulation data on a node (E.g., descriptive statistics, embeddings)

Measure: A representation of how close a signature is to other signatures in the simulation (E.g., distances, densities, estimators)

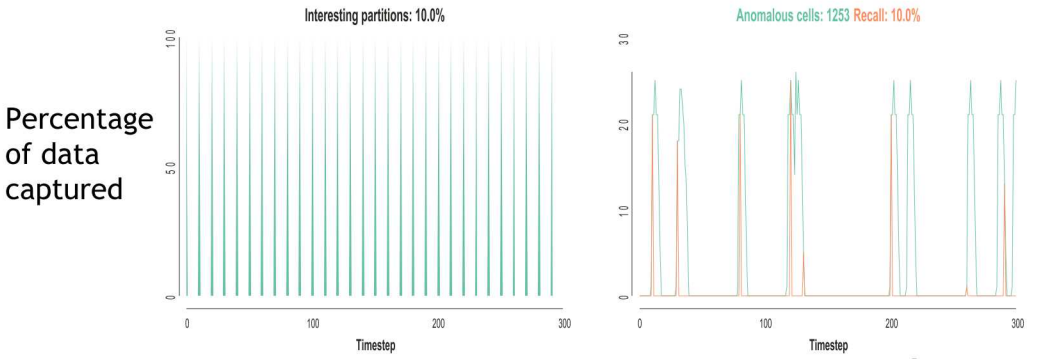
Decisions: An arbitration of the measures to determine which nodes contain "interesting" data, given the signatures and measures (E.g., threshold, momentum)

Slide Courtesy: W. Davis

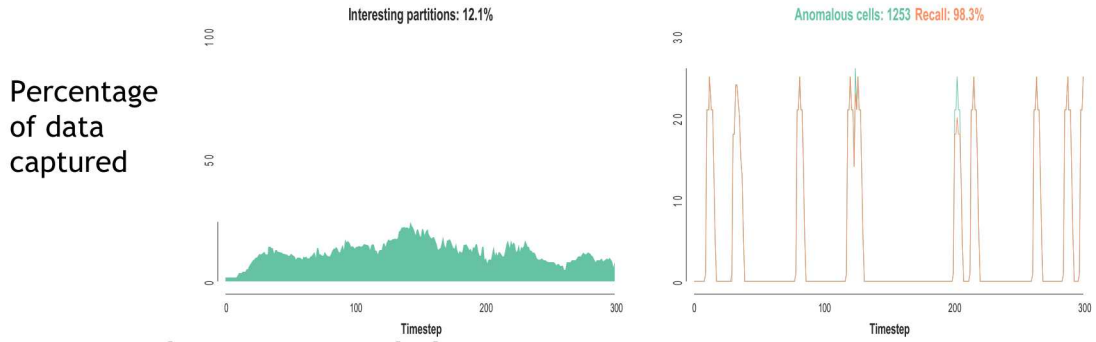
In-Situ Detection is More Accurate and Efficient

Turbulent Flow anomaly detection in Mantaflow

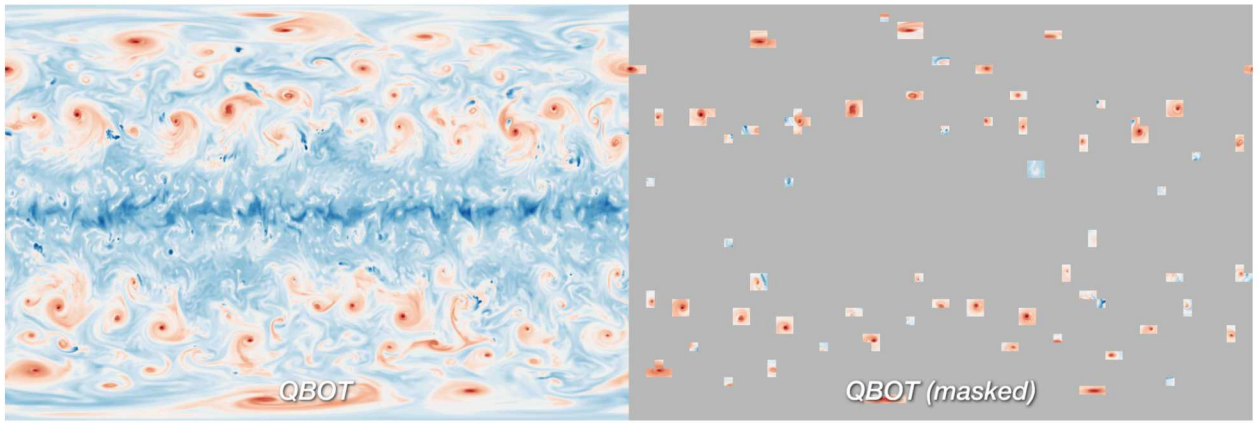
Snapshot (Conventional)



In-Situ Detection



Anomaly Detection in Climate Modeling



See Flash Talk by Warren Davis tomorrow

POC: W. Davis (SNL)

Higher Order Moment Tensors for Combustion Analysis

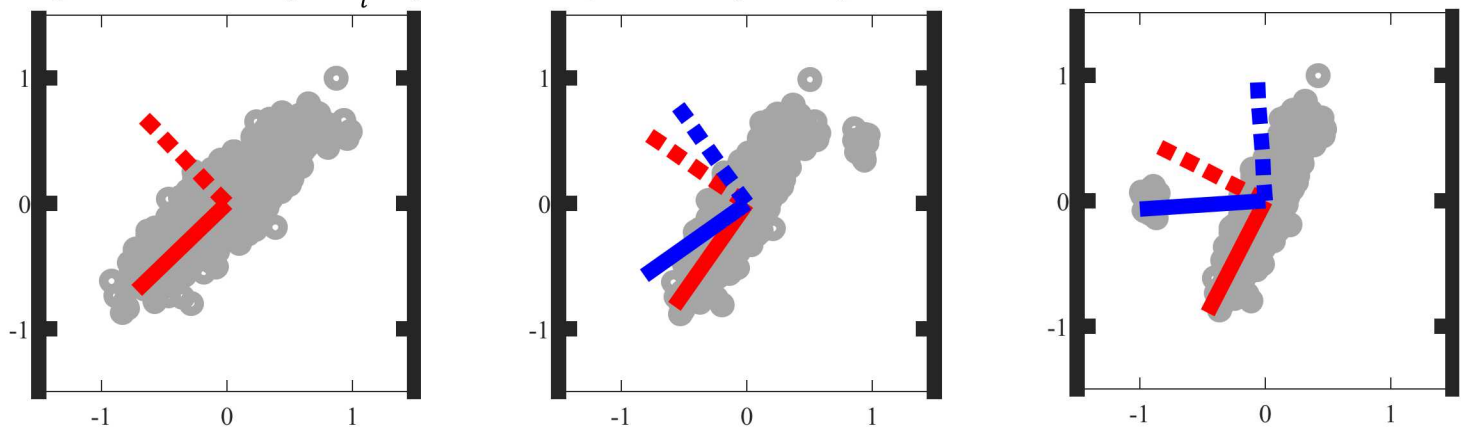
POC: H. Kolla

Hypothesis and proposed solution

- Information **of anomalous events present in higher order statistical moments**, e.g. kurtosis.
- Identify **principal vectors of kurtosis** (analogous to PCs in PCA) in the variable (a.k.a feature) space.
- ***Anomalies manifest as principal kurtosis vectors (PKVs) that are "distinct".***

Simple Moment-Tensor Decomposition

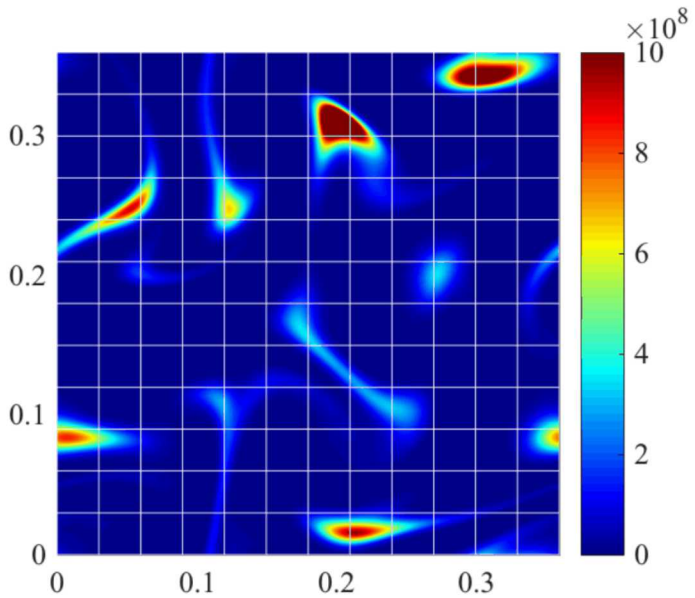
- Motivated by connections to Independent Component Analysis (ICA).
- Operate on fourth cumulant tensor (Lathauwer & Moore 2001, Comon & Jutten 2010, Anandkumar *et al.* 2014)
  - $\mathcal{M}_4 := \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] - \mathbb{E}[x_{i1}x_{i2}]\,\mathbb{E}[x_{i3}x_{i4}] - \mathbb{E}[x_{i1}x_{i3}]\,\mathbb{E}[x_{i2}x_{i4}] - \mathbb{E}[x_{i1}x_{i4}]\,\mathbb{E}[x_{i2}x_{i3}]$
- A simple way to decompose $\mathcal{M}_4$: matricize and SVD (Anandkumar *et al.* 2014):
  - $\text{mat}(\mathcal{M}_4) = \mathbf{M} = \sum_i \kappa_{s_i} a_i \otimes vec(ai \otimes a_i \otimes a_i)$



Higher Order Moment Tensors for Combustion Analysis

POC: H. Kolla

Putting It Together: HCCI Data Set



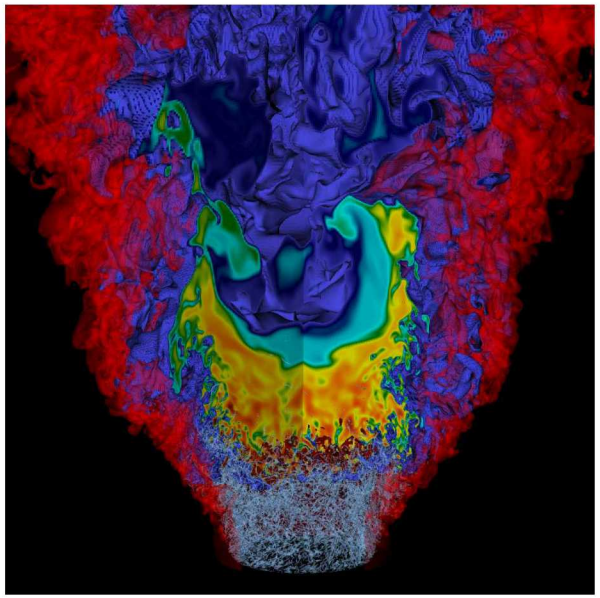
- Extract Principal Kurtosis Vectors (PKVs) on each MPI rank.
- Transform PKVs to a "moment (kurtosis) metric per feature (variable)".
- Moment metrics quantify contribution of a feature to overall kurtosis.
- Normalized (between 0-1), and also sum to 1 (like a discrete distribution).
- Compare moment metrics across MPI ranks (Hellinger distance).

- For anomaly detection in scientific data, statistical models based on higher order moments may be promising.
- Use of "principal vectors of Kurtosis" as indicators of anomalous events.
- Metrics quantify change in the principal kurtosis vectors and identify anomalous subdomains.
- Construction of PKVs as a symmetric tensor decomposition problem.

K. Aditya, H. Kolla, W. P. Kegelmeyer, T. M. Shead, J. Ling, Warren L Davis IV, 2019, "Anomaly detection in scientific data using joint statistical moments", Journal of Computational Physics, vol. 387, pp:522.

Data Compression based on Implicit Voronoi Tessellation for Combustion

POC: M. Ebeida



Challenge: Reduce the size of the massive data generated during a numerical simulation on the fly without missing important features

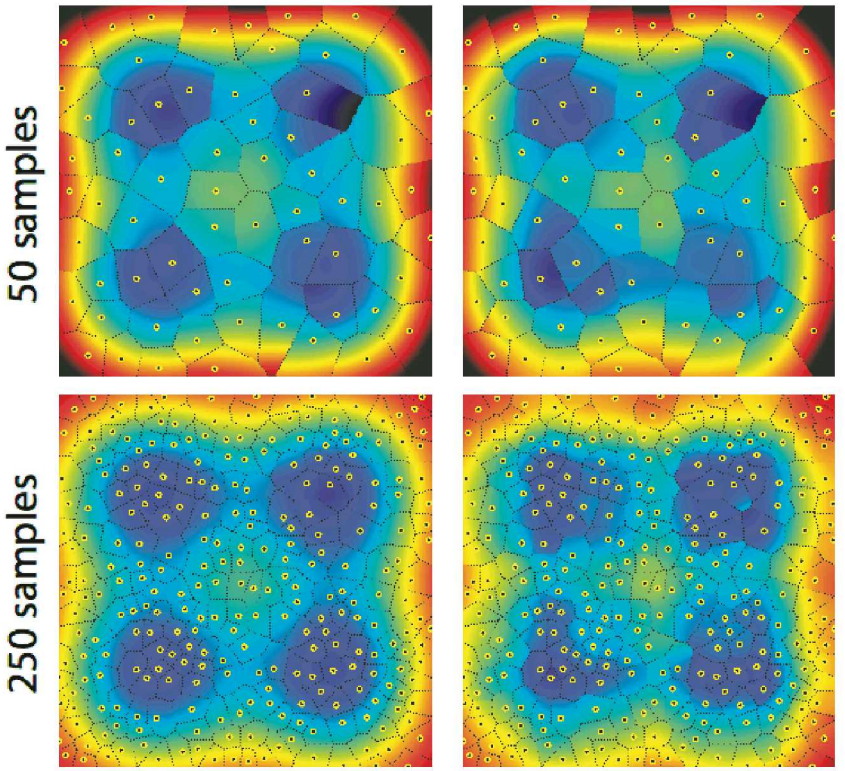
Current Practice: **Store a snapshot of the 3d field of some quantity of interest every few time steps based**, Apply **Tucker MPI** to compress on all the stored snapshots to reduce the size of the stored data

Concerns: The frequency of the storage snapshots might be insufficient to capture all the important features of the simulation. Moreover, Tucker MPI is not accurate and hence **require that the QOI of interest** is known before running the simulation. For new QOI, we need to re-run the expensive simulation

Data Compression based on Implicit Voronoi Tessellation for Combustion

POC: M. Ebeida

Our Approach: We **adaptively sample the four-dimensional spatio-temporal space as time evolves**, and we utilize the **implicit Voronoi Cells** around each of these samples to **construct a local surrogate** that approximate the underlying field with a user-input desired global accuracy.



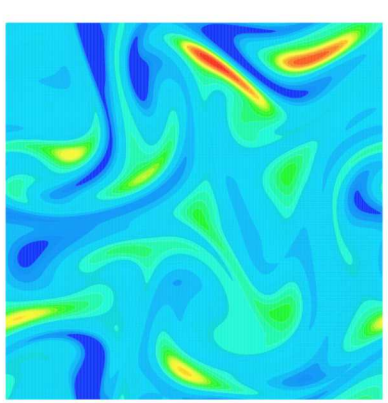
Merits:

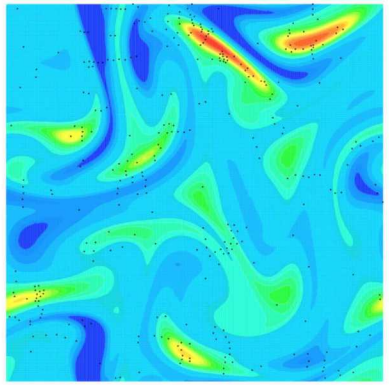
1. The **adaptive sampling automatically decides where and how to approximate the generated data throughout the 4 dimensional space**. This should eliminate the current heuristic intermittent storage of generated data.
2. Since out approach guarantees accuracy, we apply it to the primitive data directly not the QOI. This enable using the compressed data in future analysis.
3. In addition, our local surrogates can provide analytical derivatives of the approximated surface.

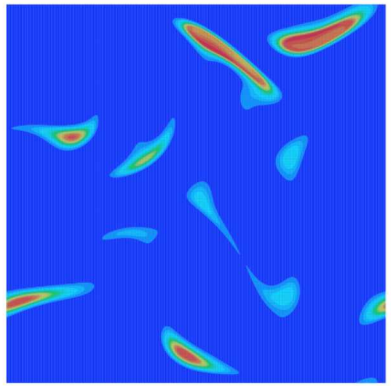
EXASCALE COMPUTING PROJECT

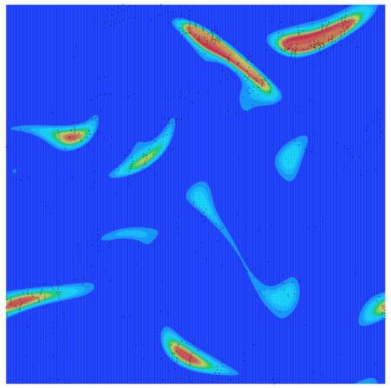
Initial Results of a prototype implementation:

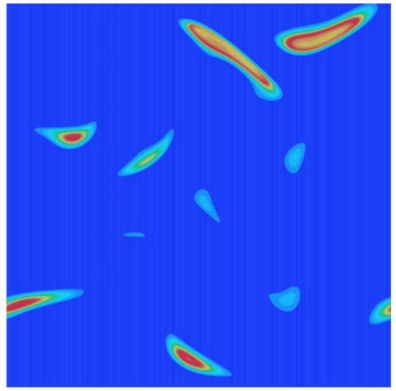
POC: M. Ebeida

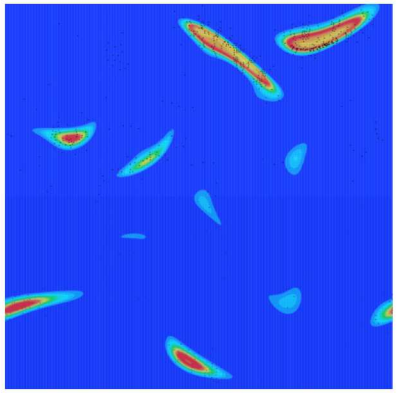

Temperature field


Temp. Compressed

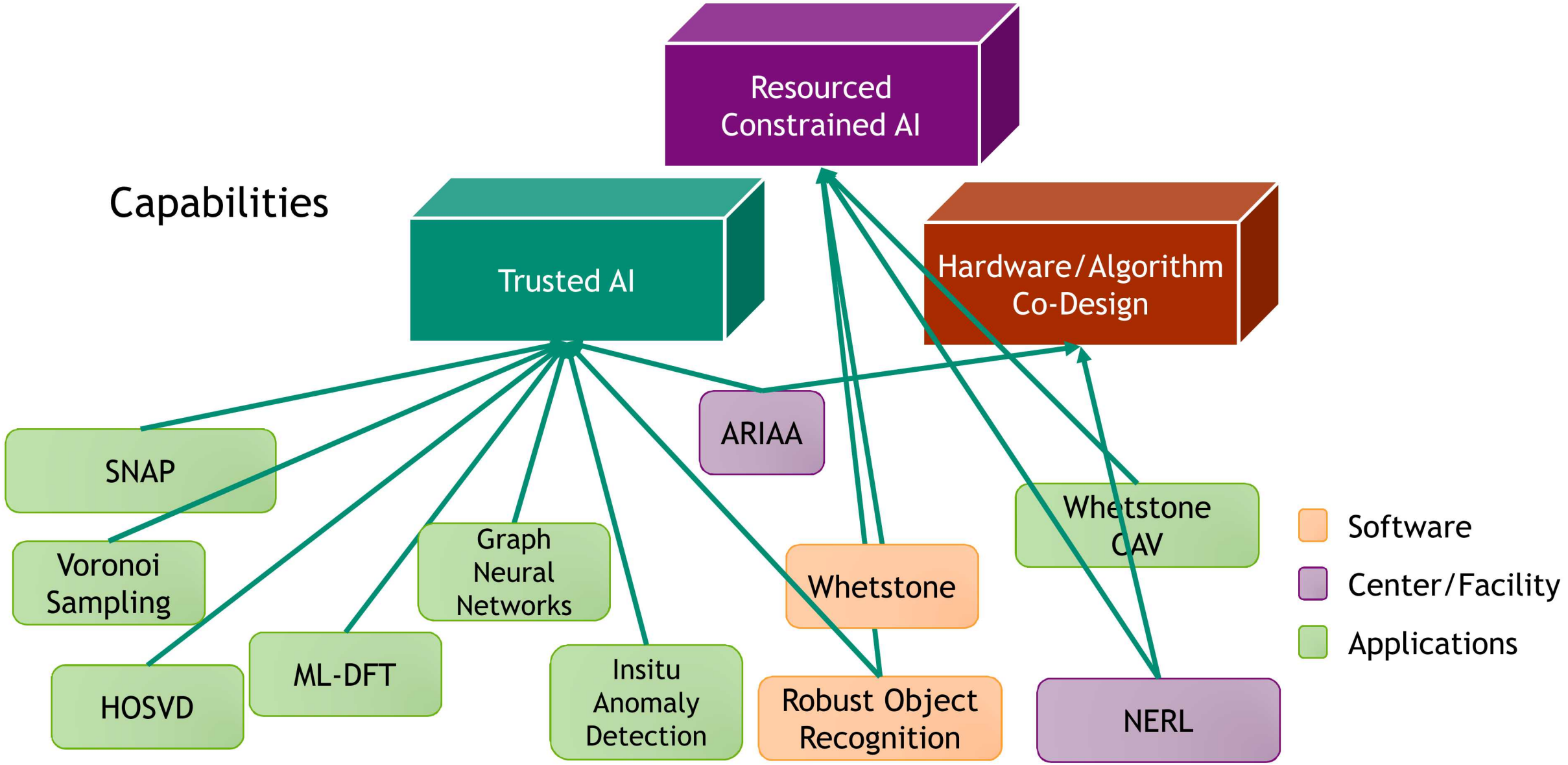

CH3 field


CH3 compressed


OH field


OH compressed

- A prototype implementation of our proposed approach was applied to **three primitive fields** (Temperature, CH3 and OH) generated by two-dimensional combustion simulation
- We were able to **achieve 3 orders of magnitude in compression ratio**. The code was executed in 210 seconds.
- The combustion team reported that they found these results promising and that they couldn't get comparable accuracy with Tucker MPI.
- We are currently working on implementing a **Voronoi Compression Tool based on kokkos.**


Capabilities
Resourced Constrained AI
Trusted AI
Hardware/Algorithm Co-Design
SNAP
Voronoi Sampling
HOSVD
ML-DFT
Graph Neural Networks
Insitu Anomaly Detection
ARIAA
Whetstone
Robust Object Recognition
Whetstone CAV
NERL
Software
Center/Facility
Applications

Thanks! Questions?

Exceptional service in the national interest