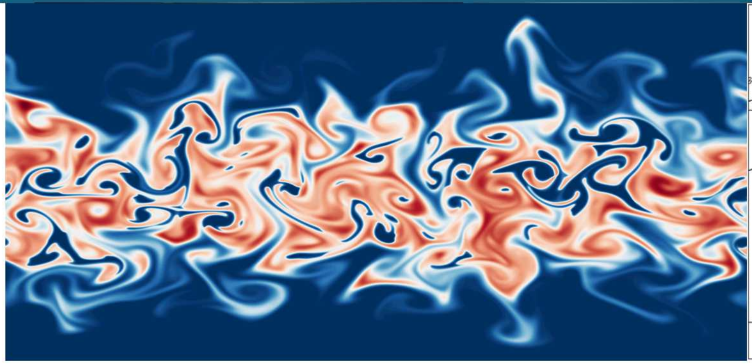
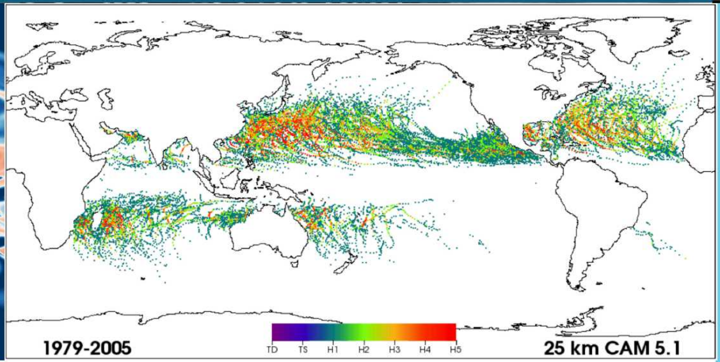
SAND2020-0613PE

In-Situ Machine Learning for Intelligent Data Capture on Exascale Platforms





PRESENTED BY

Warren L. Davis IV

Collaborators: Tim Shead, Hemanth Kolla, Kevin Reed, Philip Kegelmeyer, Gabriel Popoola

Artificial Intelligence for Robust Engineering & Science Workshop (AIRES), January 22-24, 2023

SAND 2020-XXXXX

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

DOE Base CS Research with Academic Collaboration

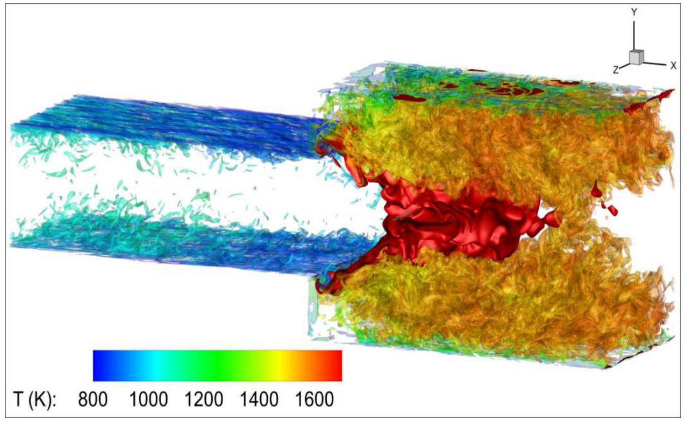
- DOE Office of Science - ASCR funded research (PM: Robinson Pino)
- Collaborative research with Stony Brook University
- Three-year research ($500K) – leftover funding for wrap-up publications and conferences

SNL: Warren Davis (PI), Tim Shead, Hemanth Kolla, Philip Kegelmeyer
Stony Brook: Kevin Reed (PI)
North Carolina A&T: Gabriel Popoola

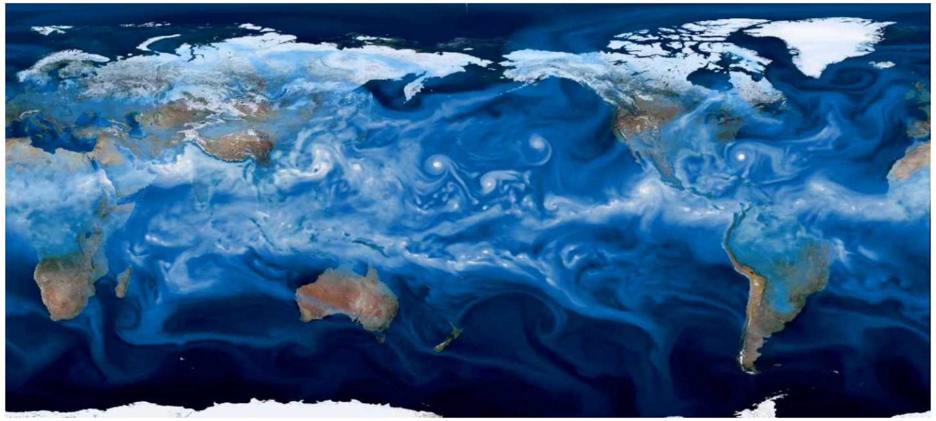
Past Members: Danny Dunlavy (SNL), Julia Ling (Citrine Informatics), Aditya Konduri (Indian Institute of Science)

Motivation and Context

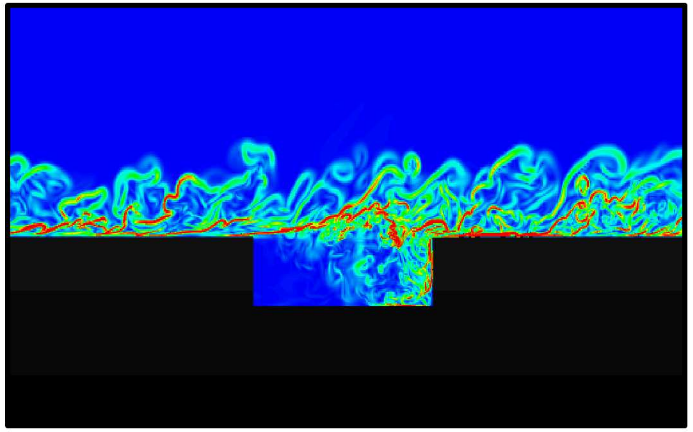
- DOE is interested in many problems that require high-fidelity physics-based HPC simulations



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)

Current state-of-the-art for HPC simulation analysis

- Take "snapshots" in space and time (1/1000th or 1/10000th)
- Post-process snapshot data with standard algorithms

Problems with the current methods:

- Interesting events may happen between or outside of these snapshots
- Important information leading up to the captured event could be lost
- Rerunning simulations to capture lost information is expensive

- This problem will only get worse as the amount of data and fidelity of the simulations increases

Is there a way to detect the anomalies *in-situ*,
thus facilitating more precisely targeted event capture?

Changing the Paradigm with *In-Situ* Event Detection

- **Develop techniques to detect interesting spatial and temporal events *in-situ* for HPC physics simulations**

- **Scalable : Can't significantly hinder the runtime of the application**
- **Unsupervised : To enable discovery, should not require labeling of interesting events**
- **Generalizable : Not focused on one specific event or domain**
- **Online : Don't require having access to all the data from every time step (post-processing)**

This is foundational research, with a focus on algorithms that can motivate changes to simulation code and facilitate more intelligent, focused data capture

Anomaly Detection Framework

Signatures

A condensed, information-rich, representation of the simulation data on a node

- E.g., descriptive statistics, embeddings

Measures

A representation of how close a signature is to other signatures in the simulation

- E.g., distances, densities, estimators

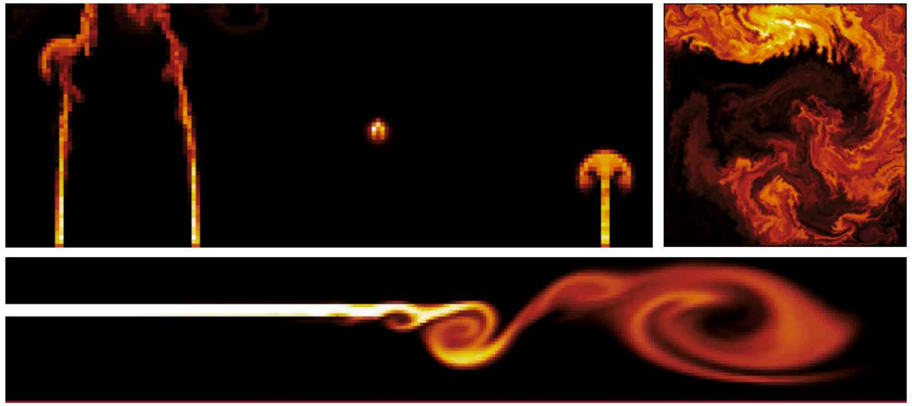
Decisions

An arbitration of the measures to determine which nodes contain "interesting" data, given the signatures and measures

- E.g., threshold, momentum

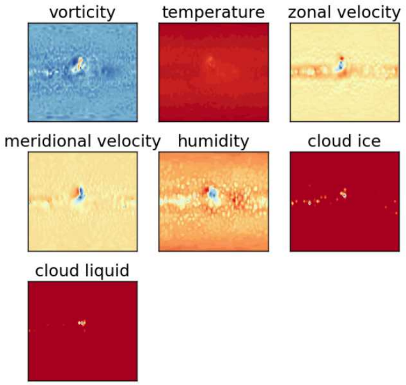
Vehicles for Exploration and Experimentation

MantaFlow



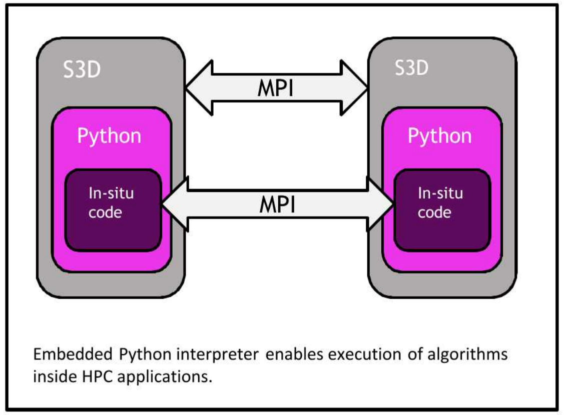
Fluid dynamics

CESM/CAM5



Climate

S3D



Combustion

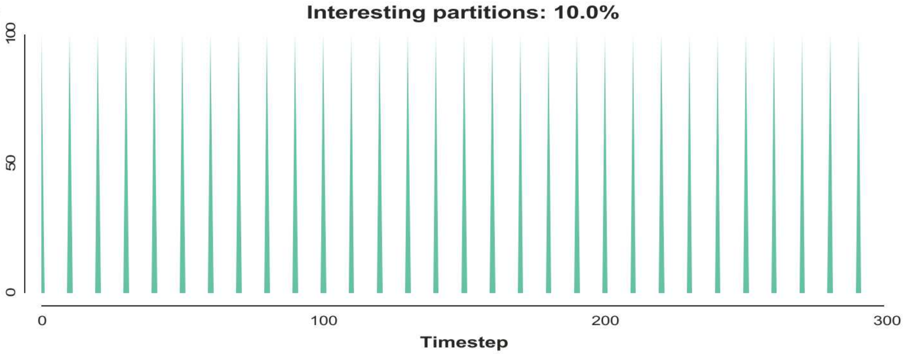
Current research tools require ability to run Python from within the simulation code

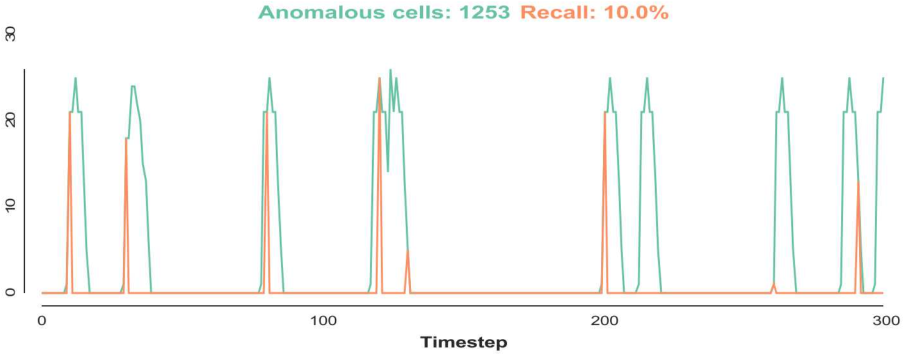
In-Situ Detection is More Accurate and Efficient

Turbulent Flow anomaly detection in Mantaflow

Snapshot (Conventional)

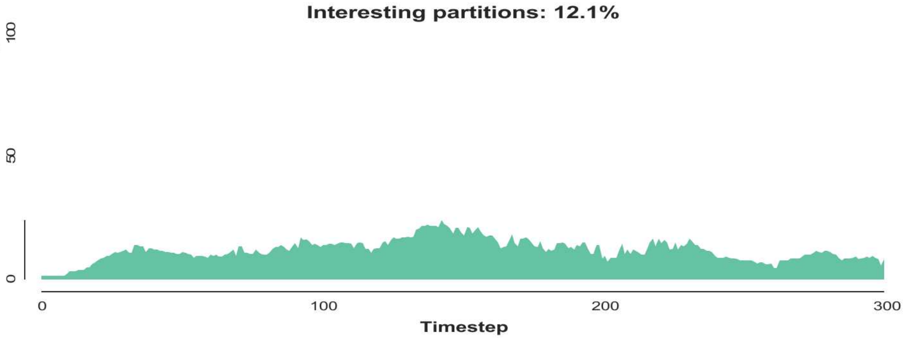
Percentage of data captured

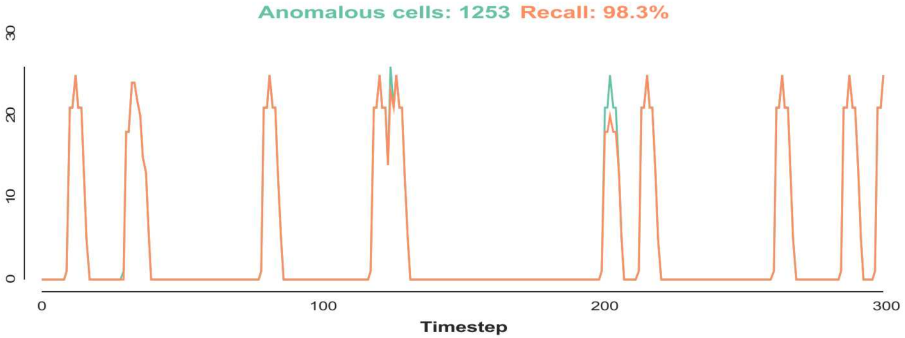




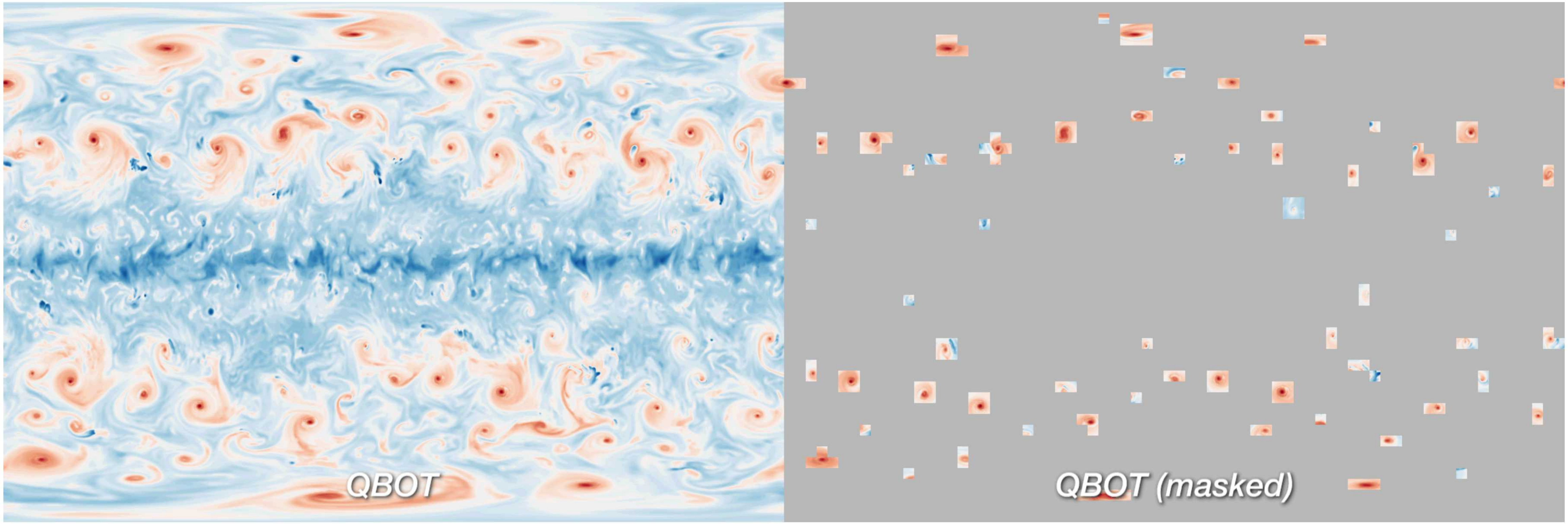
In-Situ Detection

Percentage of data captured





Anomaly Detection in Climate Modeling



Climate Modeling

Towards Exascale

- Formally integrate *in-situ* capability to S3D
- Integration into CAM5
- Categorization of signatures, measures, and decisions
  - What signature works best for this type of data?
  - What measures best capture this particular type of change?
- Exploration of new domains / Increased scaling

- Finite-element simulations

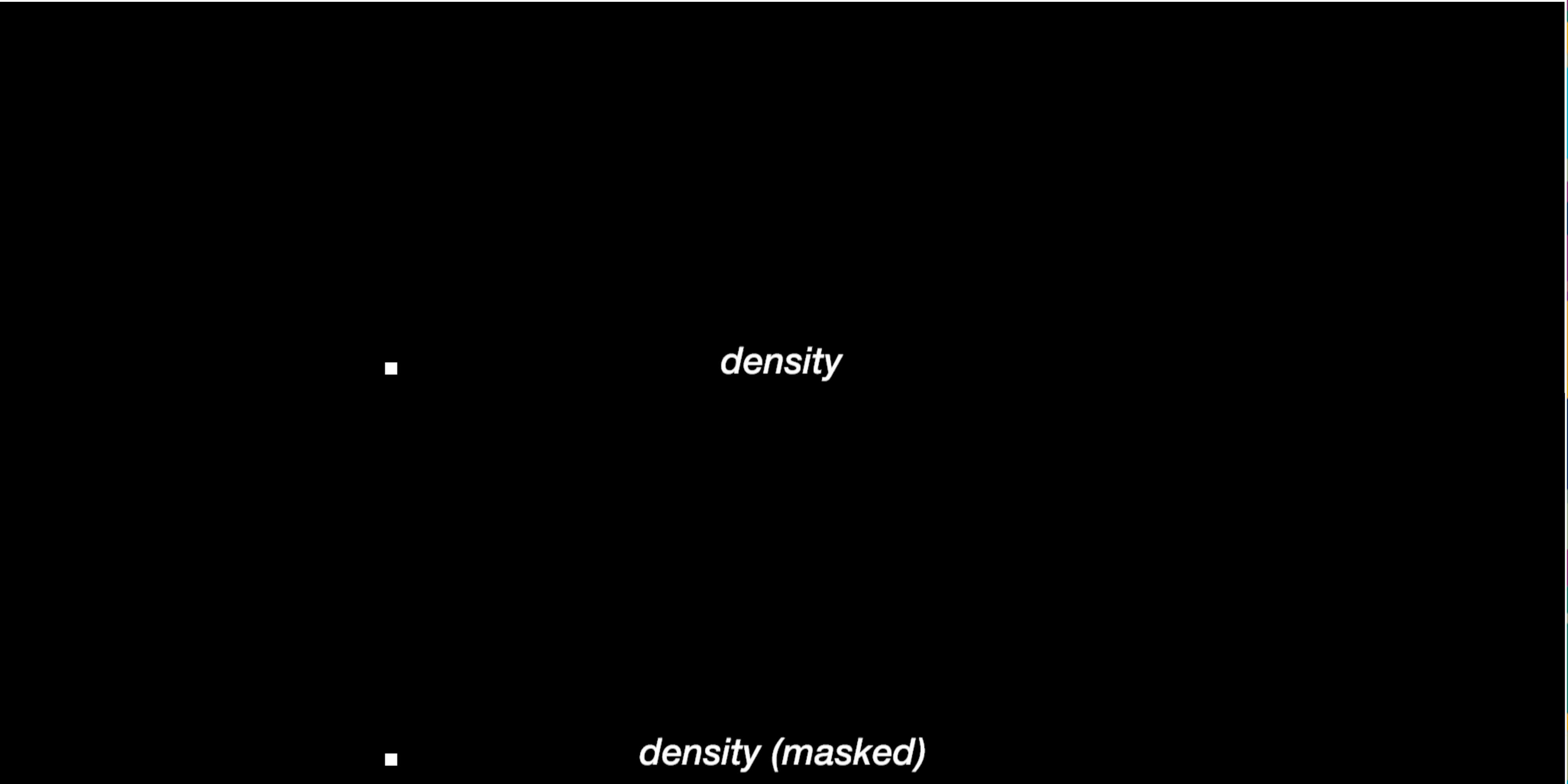
More Information

- Publications/Presentations
  - Ling, Julia, W. Philip Kegelmeyer, Aditya Konduri, Hemanth Kolla, Kevin A. Reed, Timothy M. Shead and Warren L. Davis IV. "Using feature importance metrics to detect events of interest in scientific computing applications." *2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)* (2017): 55-63.
  - Kolla, Hemanth, Aditya Konduri, Prashant Rai, Tamara G. Kolda, Warren Leon Davis. "Tensor Decomposition to Perform Change of Basis in Multi-Variate HPC Data to Preserve Higher Order Statistical Moments," *Presentation,* SIAM Parallel Processing 2018, March 2018.
  - Konduri, Aditya, Hemanth Kolla, Julia Ling, W. Philip Kegelmeyer, Timothy Shead, Daniel Dunlavy, Warren Leon Davis. Event Detection in Multi-Variate Scientific Simulations Using Feature Anomaly Metrics," *Presentation,* SIAM Parallel Processing 2018, March 2018.
  - Aditya K, Kolla H, Kegelmeyer WP, Shead TM, Ling J, Davis IV, Warren L. "Anomaly detection in scientific data using joint statistical moments", Journal of Computational Physics, Vol 387.
  - Timothy M. Shead, Konduri Aditya, Hemanth Kolla, Daniel M. Dunlavy, W. Philip Kegelmeyer, Warren L. Davis IV. "Embedding Python for In-Situ Analysis." SAND2018-9009. August 2018.
  - Davis IV, Warren Leon; Shead, Timothy M.; Kolla, Hemanth; Popoola, Gabriel; Kegelmeyer, Philip; Konduri, Aditya. "The Potential of Integrated Machine Learning Algorithms for Tropical Cyclone Detection in Advanced Climate Modeling." American Geophysical Union Fall Meeting, December 2019.

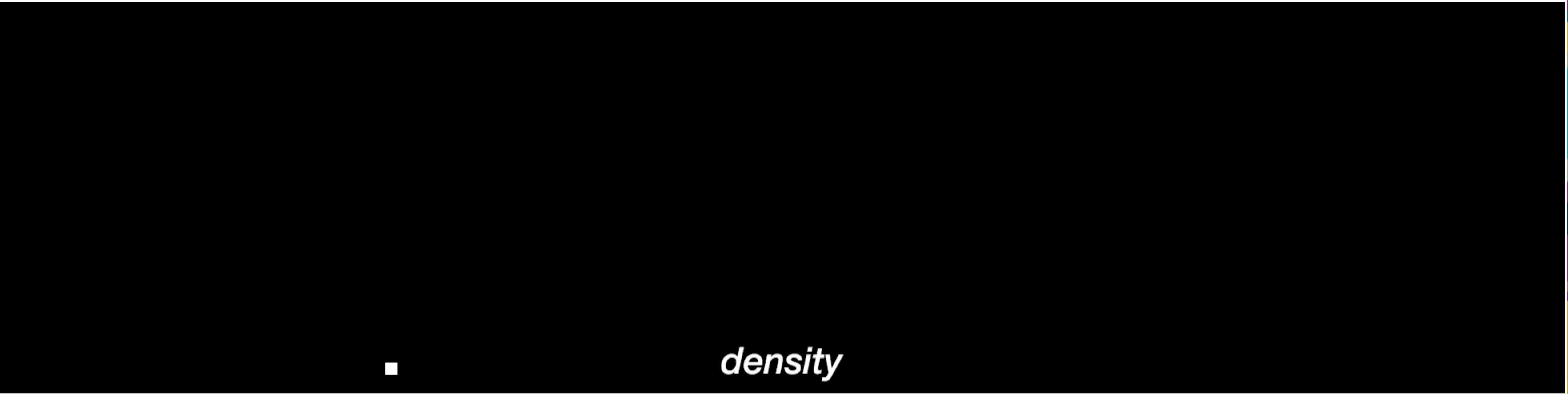
- For more information, contact: Warren L. Davis IV (wldavis@sandia.gov)

Backup

Fluid Flow (Snapshots)

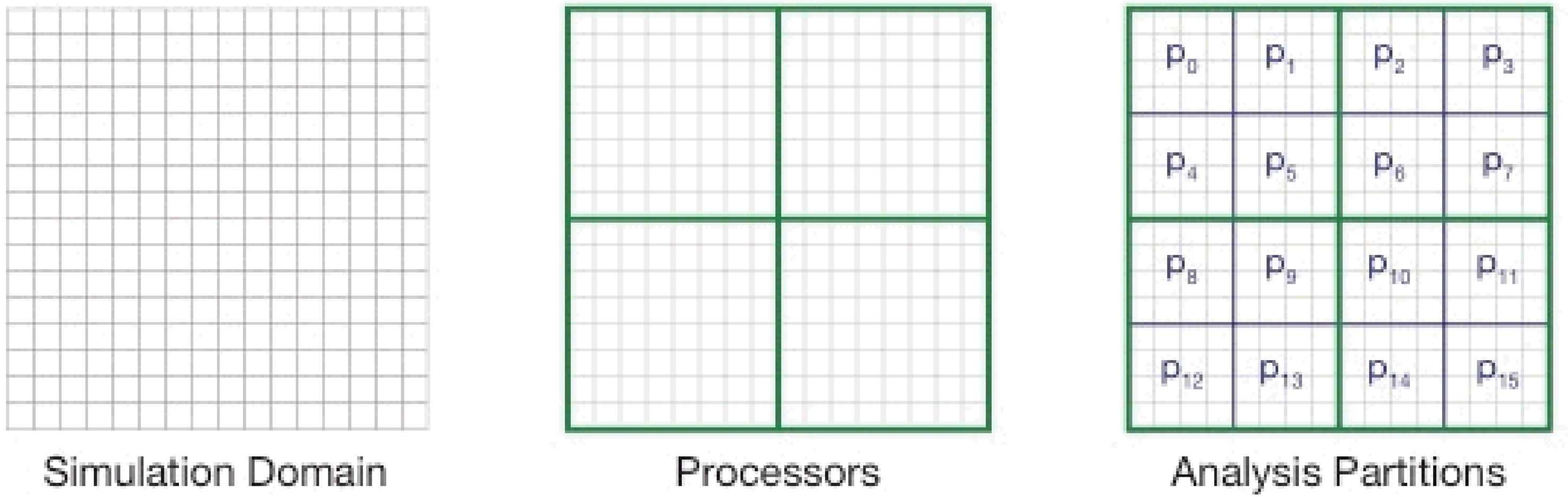


Fluid Flow (In-Situ)

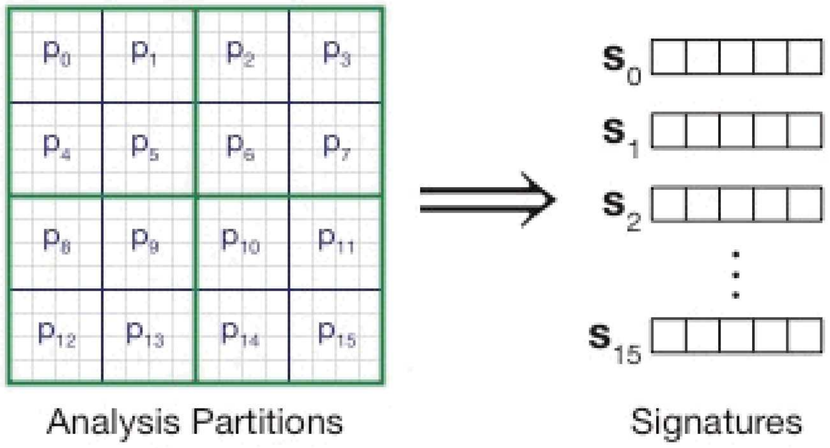


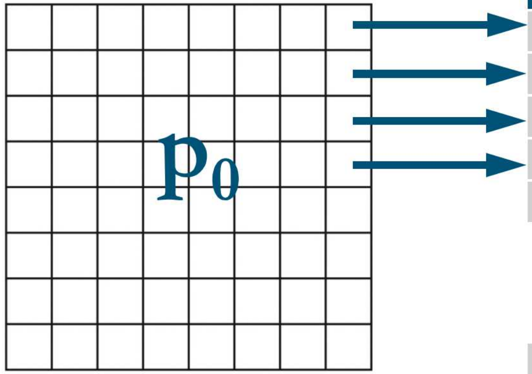
A New Anomaly Detection Framework: Signatures, Measures, and Decisions

Communication is a constraint for In-Situ HPC Anomaly Detection



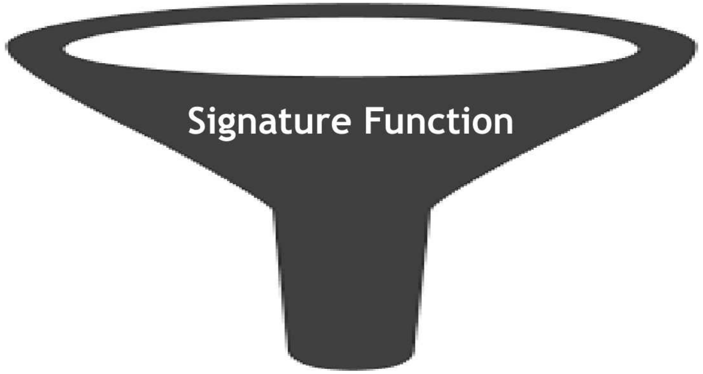
Signatures Represent the Data on a Partition





Individual mesh attributes for P_0

Density	Pressure	Vx	Vy
10	4	2	4
20	10	8	8
30	8	8	12
40	6	2	16
.	.	.	.
40	6	2	16



P_0 signature

25	7	5	10

m Number of mesh points

a Attributes per mesh point

t *m***a*, the total number of values on a partition

Signatures can be shorter or longer than *a*, as long as they are shorter than *t*

Signatures Can Take Many Forms

Examples

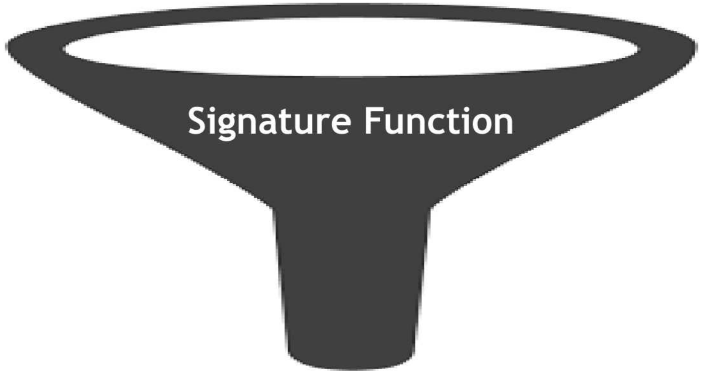
- Mean
  - Individual attribute mean values over the mesh points on a partition
- FIEDA Feature Importance Metric (FIM) scores *
  - First, kernel-density estimation to produce a probability distribution over the state variables
  - Next, random forests to predict the pdf given the state variables
  - Lastly, extract feature importance values from the random forest and use as a signature

*Ling et al. "Using feature importance metrics to detect events of interest in scientific computing applications." *2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)* (2017): 55-63.

Signatures are significantly smaller than all the data on a partition, and can be communicated with little cost, comparatively.

Individual mesh attributes for P_0

Density	Pressure	Vx	Vy
10	4	2	4
20	10	8	8
30	8	8	12
40	6	2	16
		.	
		.	
		.	
40	6	2	16



P_0 signature

25	7	5	10

Measures Indicate the Distance of a Signature From Neighbors

Measures take as input a list of T $P \times S$ matrices where T is the number of elapsed timesteps and each $P \times S$ matrix contains the signatures for the partitions at a given timestep.

Measures can be specific to a type of signature, or general measures, including typical anomaly detection algorithms

Examples

- Mean-Squared Distance
- DBSCAN
- FIEDA M1*

*Ling et al. "Using feature importance metrics to detect events of interest in scientific computing applications." *2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)* (2017): 55-63.

Decisions Allow for Customization

Measures are scalar values that do not, by themselves, answer whether something is anomalous.

Different applications can decide an appropriate anomalousness point

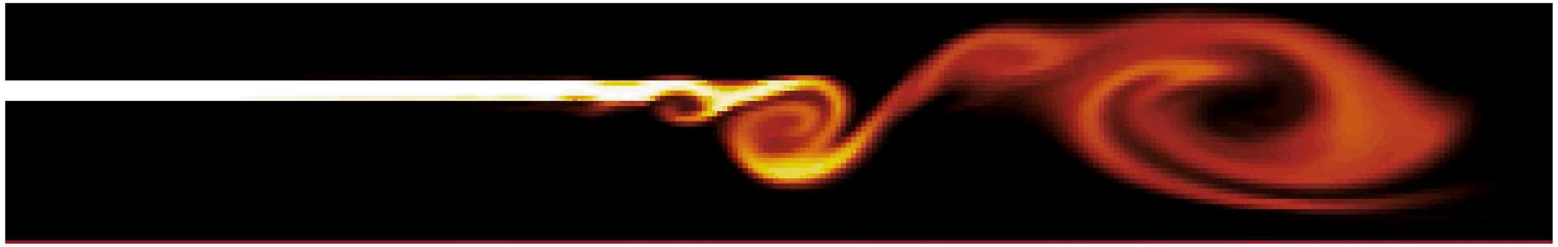
Examples

- Threshold
- Percentile-Change
- Memory / Feathering

Decision functions are meant to be adjustable to fit application needs and are the final arbiter of what is “interesting” in a simulation.

Rapid Development and Testing

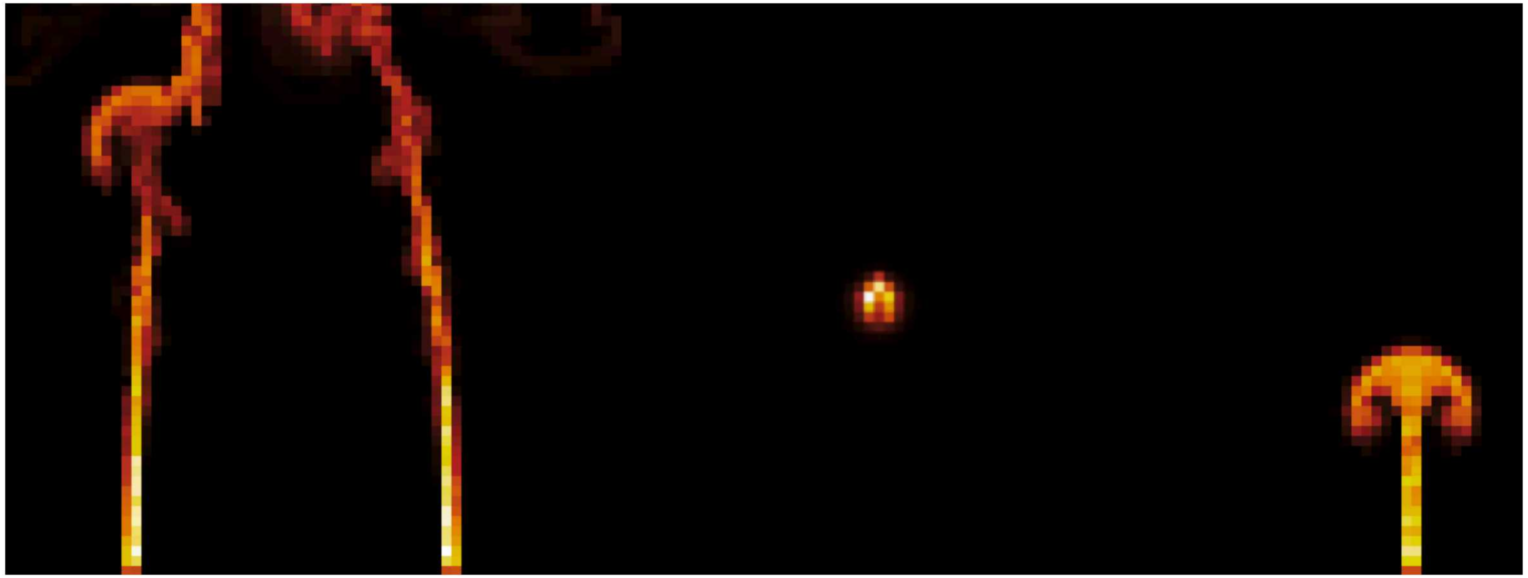
- S3D is useful but too unwieldy for rapid experimentation
- Mantaflow (ETH Zurich, Technical University of Munich)
  - Mini-app that can be run on the desktop
  - Modified to simulate HPC environment (partitioning, inter-partition communication model)



Approximately 30 new viable algorithms, some of which perform better than our previous published algorithms

Experiments and the Complexity of Measuring Performance

- **Buoyant fluid injections simulated in Mantaflow**
- **Various algorithms capture different aspects of the simulation**
  - **Hard to get a crisp definition of accuracy vs. data efficiency**
  - **We devised a way of adding anomalies independent of the flow simulation**
    - **Modifications to mesh attributes that wouldn't be congruent with the simulation**
  - **Determining *recall* in relation to data export is now possible**



We can measure the accuracy of our methods along with the data savings and compare to "snapshotting" and other approaches.

Summary

- Conventional approaches to anomaly detection in HPC simulations is insufficient, and this problem will grow
- Experiments have shown that in-situ anomaly detection is possible, both implementation-wise and algorithmically
- In-situ detection is more accurate and efficient
- Developed new algorithms and a framework for rapid development and testing