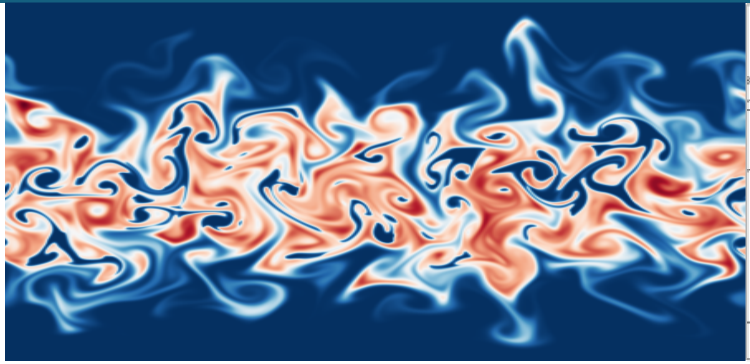
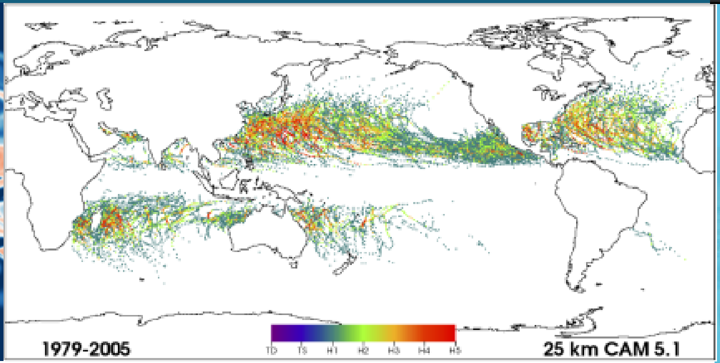
SAND2021-7874C

Sandia National Laboratories

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





PRESENTED BY

Warren L. Davis IV

Collaborators: Hemanth Kolla, Tim Shead, Irina Tezaur, Philip Kegelmeyer, Gabriel Popoola

Platform for Advanced Scientific Computing (PASC) Conference 2021
July 9, 2021

U.S. DEPARTMENT OF ENERGY NNSA

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.

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.

U.S. DOE Base Computer Science Research

- DOE Office of Science - ASCR funded research
- Phase 1: Three-year research, Collaborative research with Stony Brook University
- Phase 2: Recently renewed as a 4-year project

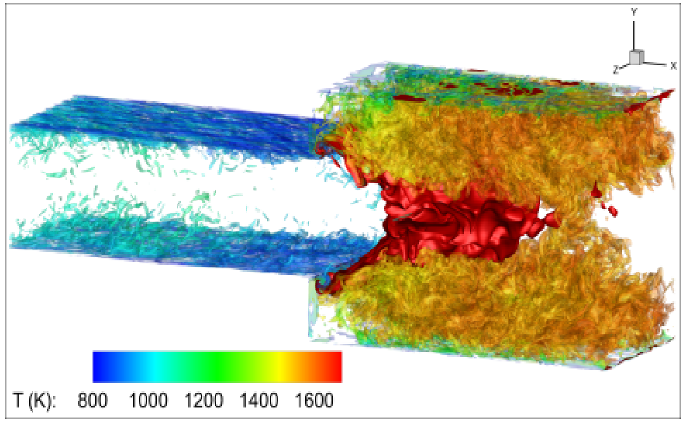
SNL: Warren Davis (PI), Hemanth Kolla, Tim Shead, Irina Tezaur, Philip Kegelmeyer, Gabriel Popoola

Past Members: Kevin Reed (Stony Brook University), Danny Dunlavy (SNL), Julia Ling (Citrine Informatics),

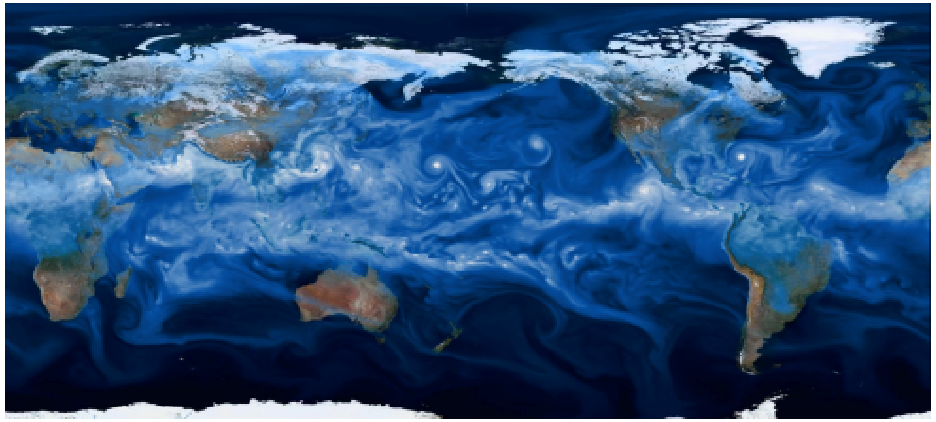
Aditya Konduri (Indian Institute of Science)

Motivation and Context

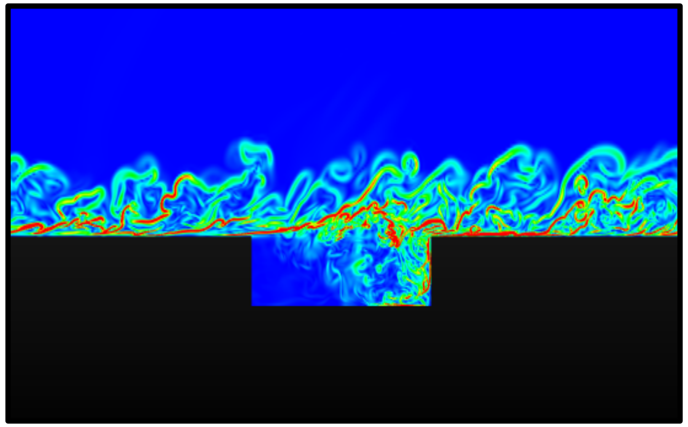
- DOE is interested in many problems that require high-fidelity physics-based HPC simulations



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,

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

Related Research is in the Early Stages

- Domain-specific
  - J. Bennett, A. Bhagatwala, J. Chen, A. Pinar, M. Salloum, and C. Seshadhri. 2016. Trigger Detection for Adaptive Scientific Workflows Using Percentile Sampling. SIAM Journal on Scientific Computing 38, 5 (2016), S240–S263. https: //doi.org/10.1137/15M1027942
  - P. Malakar, V. Vishwanath, C. Knight, T. Munson, and M. E. Papka. 2016. Optimal Execution of Co-analysis for Large-Scale Molecular Dynamics Simulations. In SC ’16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 702–715. https://doi.org/10. 1109/SC.2016.59

- Non *In-Situ*
  - Bo Zhou and Yi-Jen Chiang. 2018. Key Time Steps Selection for Large-Scale Time-Varying Volume Datasets Using an Information- Theoretic Storyboard. Computer Graphics Forum (2018). https: //doi.org/10.1111/cgf.13399

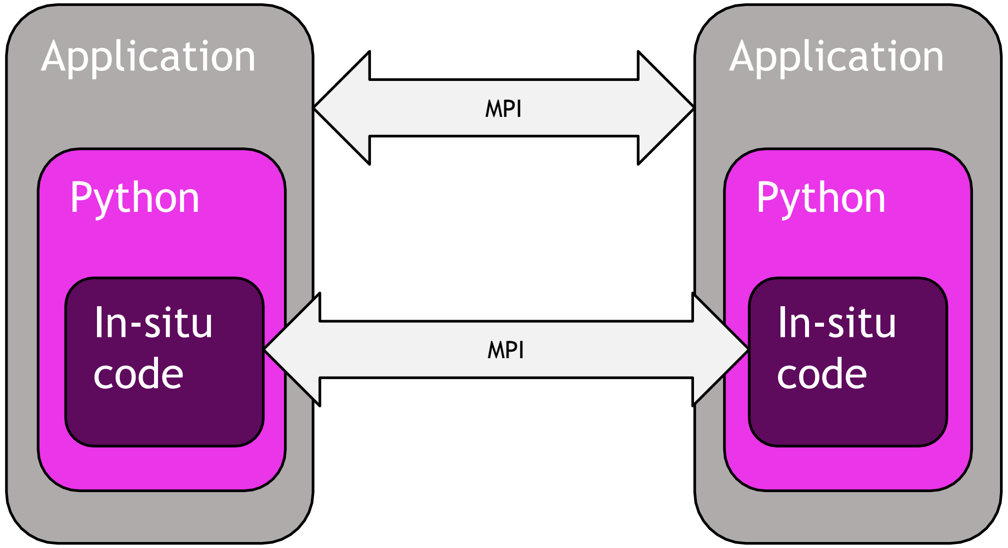
- Domain Agnostic/*In-Situ*
  - K. Myers, E. Lawrence, M. Fugate, C. McKay Bowen, L. Ticknor, J. Woodring, J. Wendelberger, and J. Ahrens. 2014. Partitioning a Large Simulation as It Runs. ArXiv e-prints (Sept. 2014). arXiv:stat.ME/1409.0909
  - Larsen, M., Woods, A.L., Marsaglia, N., Biswas, A., Dutta, S., Harrison, C., & Childs, H. (2018). A flexible system for in situ triggers. *ISAV@SC*.

Vehicles for Exploration and Experimentation

- Sandia 3D Direct Numerical Solver (S3D)
  - Used for reacting flows (e.g., combustion)

- Python Interpreter
- *In-Situ* code has access to state variables.

- Enabled immediate use of OTS algorithms and facilitated the development of new algorithms



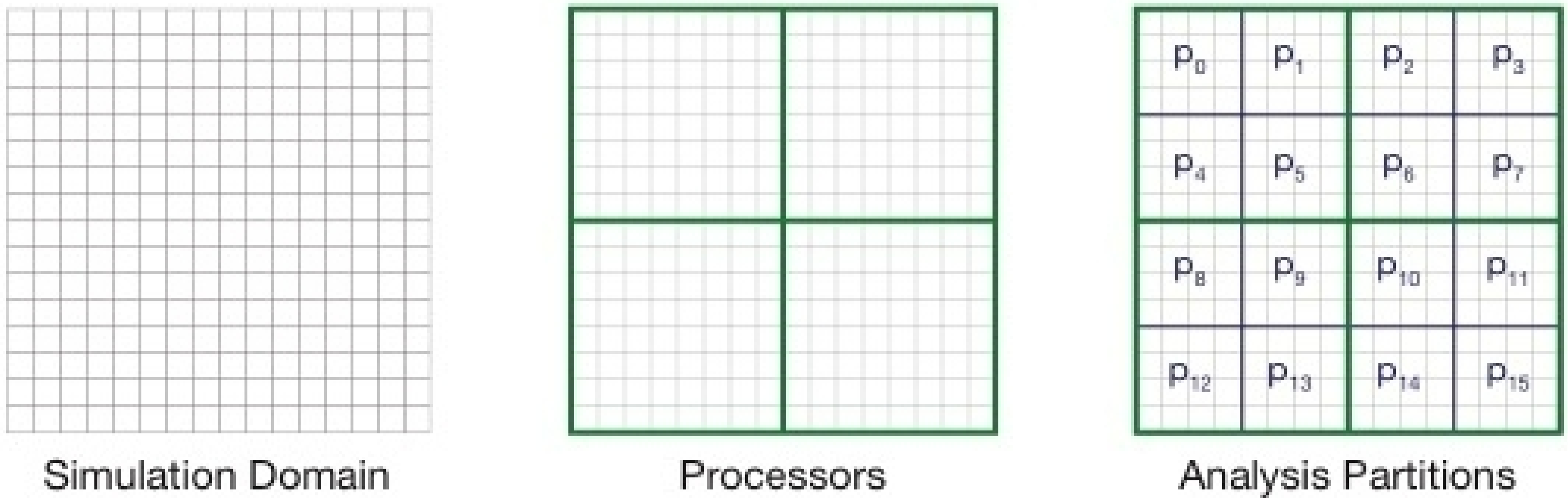
Embedded Python interpreter enables execution of algorithms inside HPC applications.

- Tested algorithms on combustion *in-situ*, climate offline (e.g., LOF, DBSCAN, i-Forests)

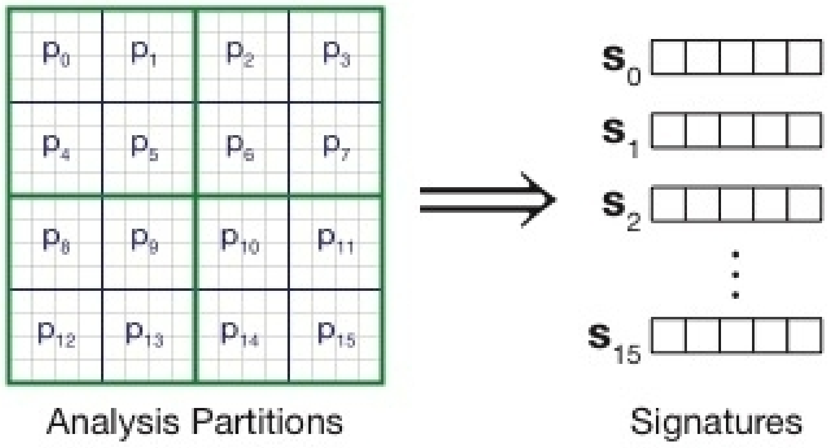
Timothy M. Shead et al. "Embedding Python for In-Situ Analysis." SAND2018-9009. August 2018.

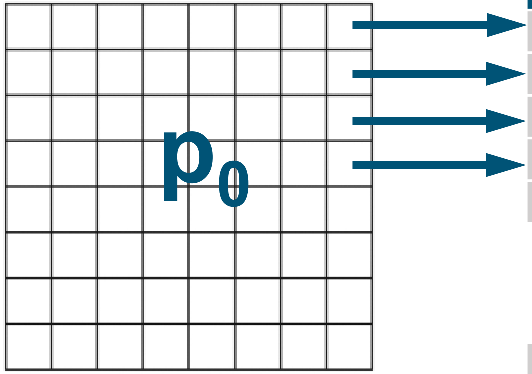
A New Anomaly Detection Framework: Signatures, Measures, and Decisions

Communication is a constraint for In-Situ HPC Anomaly Detection



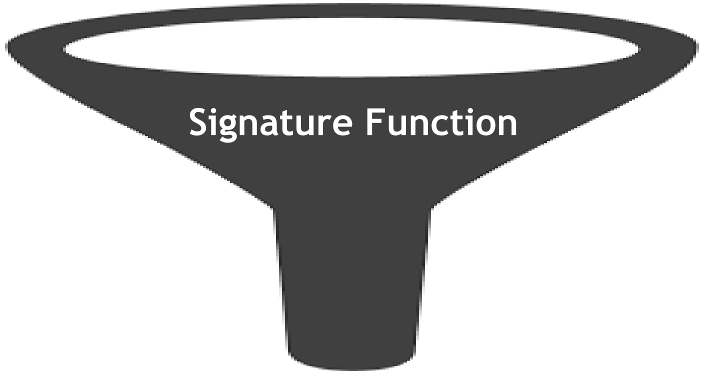
Signatures Represent the Data on a Partition





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

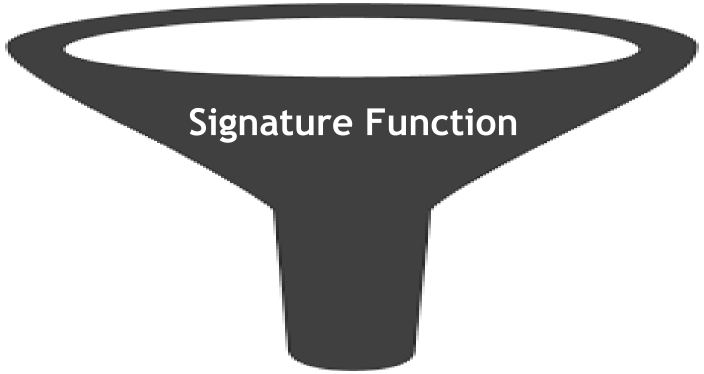
Examples

- 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.

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

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

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.

New, Effective *In-Situ* Anomaly Detection Algorithms

Using Feature Importance Metrics to Detect Events of Interest in Scientific Computing Applications

Julia Ling, Citrine Informatics
W. Philip Kegelmeyer, Sandia National Labs
Kondur Aditya, Sandia National Labs
Hemanth Kolla, Sandia National Labs
Kevin A. Reed, Stony Brook University
Timothy M. Shead, Sandia National Labs
Warren L. Davis IV, Sandia National Labs

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.

Anomaly detection in scientific data using joint statistical moments

Konduri Aditya, Hemanth Kolla, W. Philip Kegelmeyer, Timothy M. Shead, Julia Ling, Warren L. Davis IV

Abstract

We propose an anomaly detection method for multi-variate scientific data based on analysis of high-order joint moments. Using kurtosis as a reliable measure of outliers, we suggest that principal kurtosis vectors, by analogy to principal component analysis (PCA) vectors, signify the principal directions along which outliers appear. The inception of an anomaly, then, manifests as a change in the principal values and vectors of kurtosis. Obtaining the principal kurtosis vectors requires decomposing a fourth order joint cumulant tensor for which we use a simple, computationally less expensive approach that involves performing a singular value decomposition (SVD) over the matricized tensor. We demonstrate the efficacy of this approach on synthetic data, and develop an algorithm to identify the occurrence of a spatial and/or temporal anomalous event in scientific phenomena. The algorithm decomposes the data into several spatial sub-domains and time steps to identify regions with such events. Feature moment metrics, based on the alignments of the principal kurtosis vectors, are computed at each sub-domain and time step for all features to quantify their relative importance towards the overall kurtosis in the data. Accordingly, spatial and temporal anomaly metrics for each sub-domain are proposed using the Hellinger distance of the feature moment metric distribution from a suitable nominal distribution. We apply the algorithm to two turbulent auto-ignition combustion cases and demonstrate that the anomaly metrics reliably capture the occurrence of auto-ignition in relevant spatial sub-domains at the right time steps.

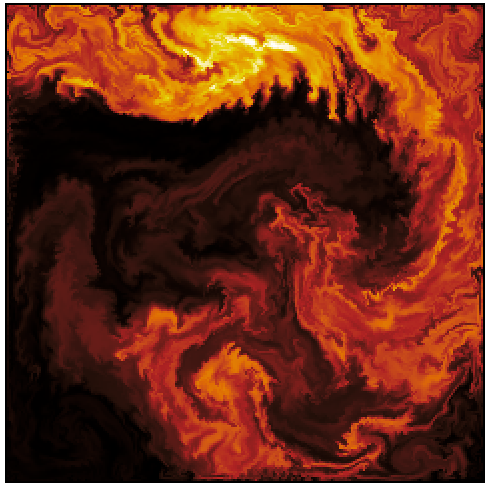
Keywords:
Anomaly detection, Scientific computing, Co-Kurtosis, Tensor decomposition, Hellinger distance, Auto-ignition

1. Introduction

Aditya et al. "Anomaly detection in scientific data using joint statistical moments", Journal of Computational Physics, Vol 387, June 15 2019, pp. 522-538.

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

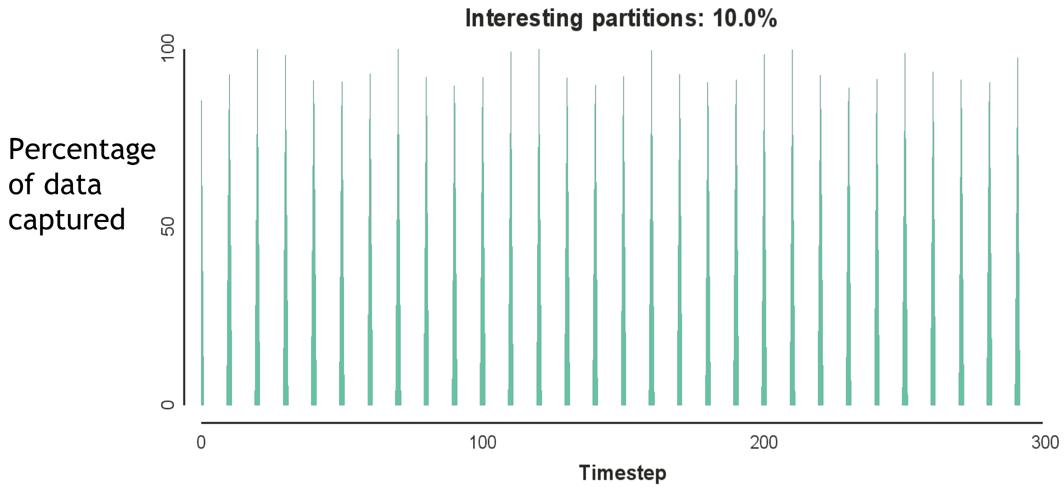
Experiments and the Complexity of Measuring Performance

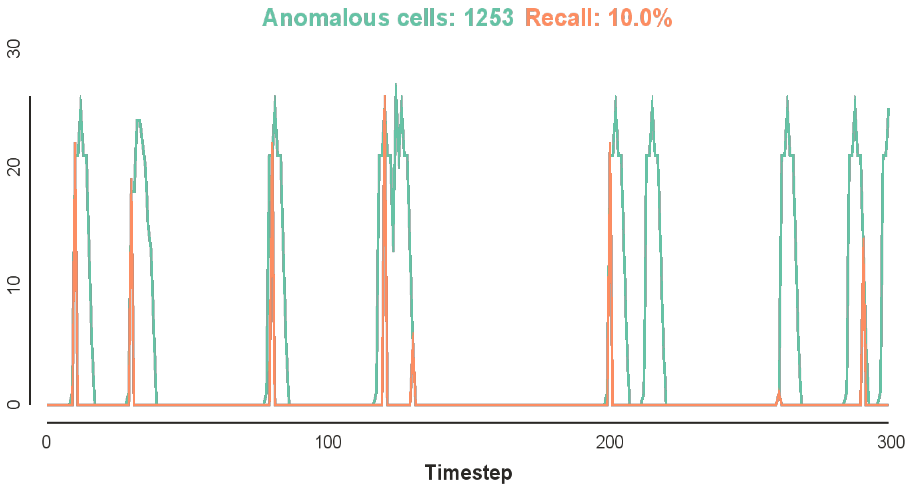
- **Buoyant fluid injections simulated in Mantaflow**
- **Various algorithms capture different aspects of the simulation**
  - **Hard to get a crisp definition of accuracy vs. data efficiency**
  - **We devised a way of adding anomalies independent of the flow simulation**
    - **Modifications to mesh 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.

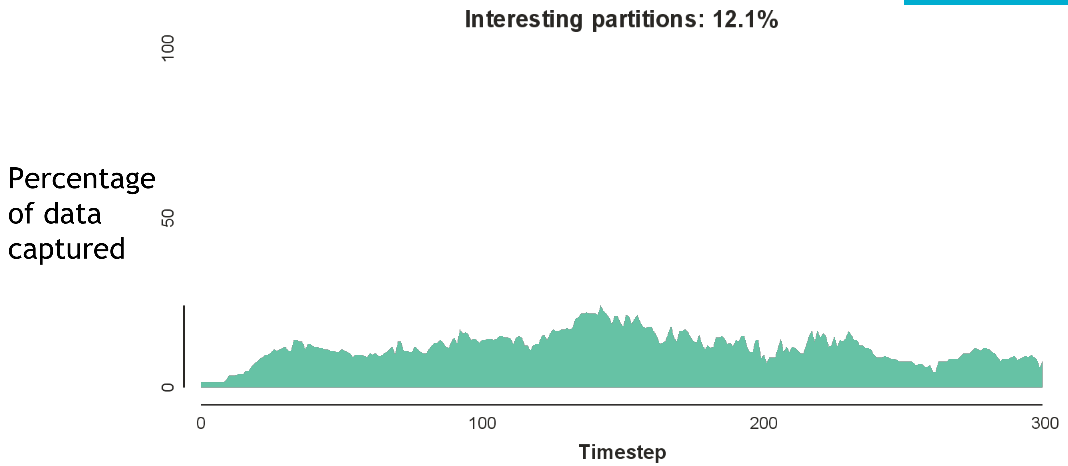
In-Situ Detection is More Accurate and Efficient

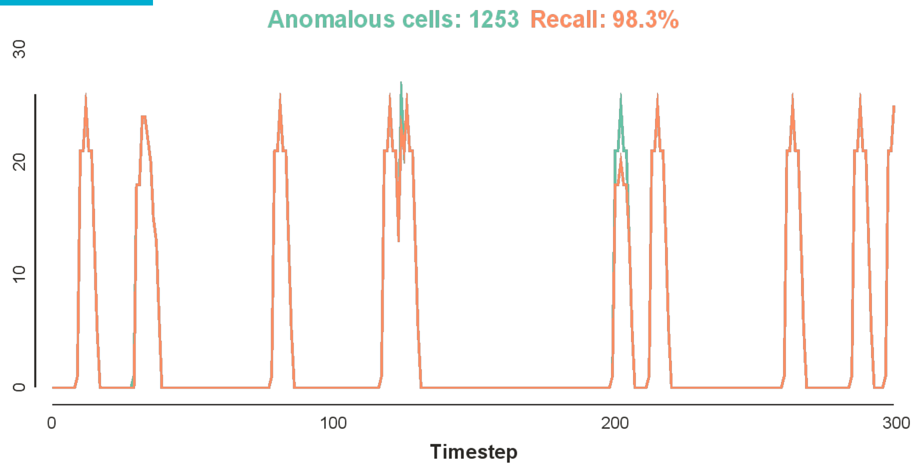
Snapshot (Conventional)





In-Situ Detection

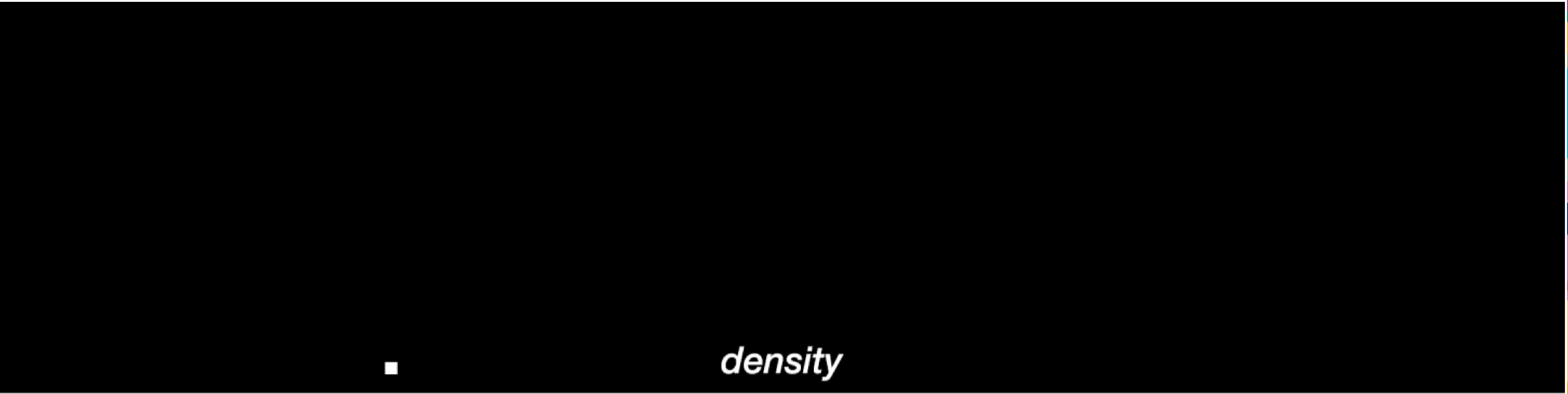




Fluid Flow (Snapshots)

- *density*
- *density (masked)*

Fluid Flow (In-Situ)



Climate

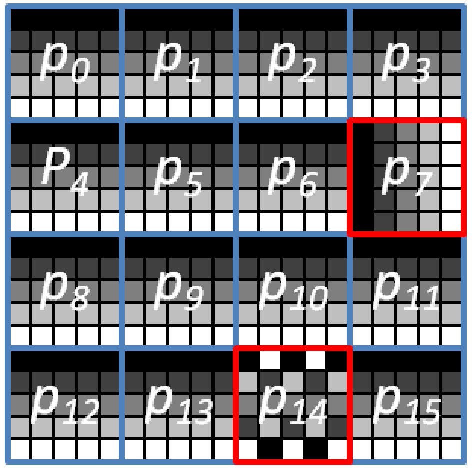
Summary

- Conventional approaches to anomaly detection in HPC simulations are 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

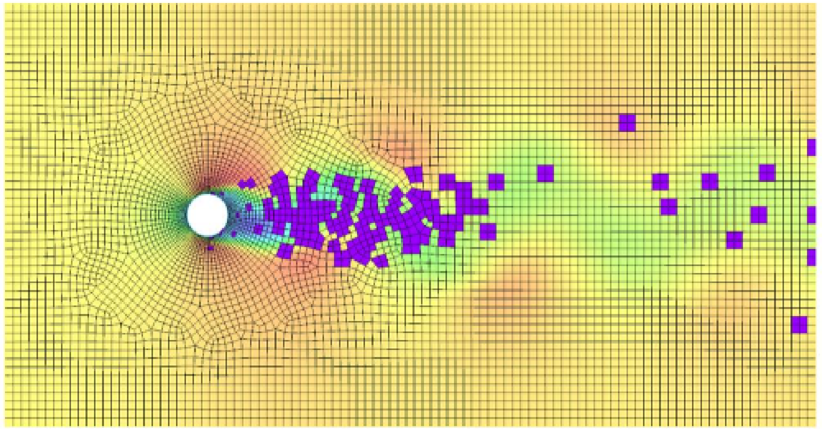
Towards Exascale

- Signature families
  - Gradient
  - Frequency
  - Simulation-driven error indicators
  - Multi-fidelity/scale signatures
- Partition/time-window effects
- Utilization of simulation data
  - Unsupervised to supervised (e.g., anomaly classification)
  - Creating supervised signatures without event labels
  - Creating supervised signatures with labels
  - Using generative models to *predict* anomalies
- Reduced-order modeling (ROM)

Our research will enhance the utility of machine learning in HPC scientific computation, creating new capabilities and increasing our overall efficacy.



Detection of statistically homogenous anomalous patterns



Mesh sampling for ROM refinement

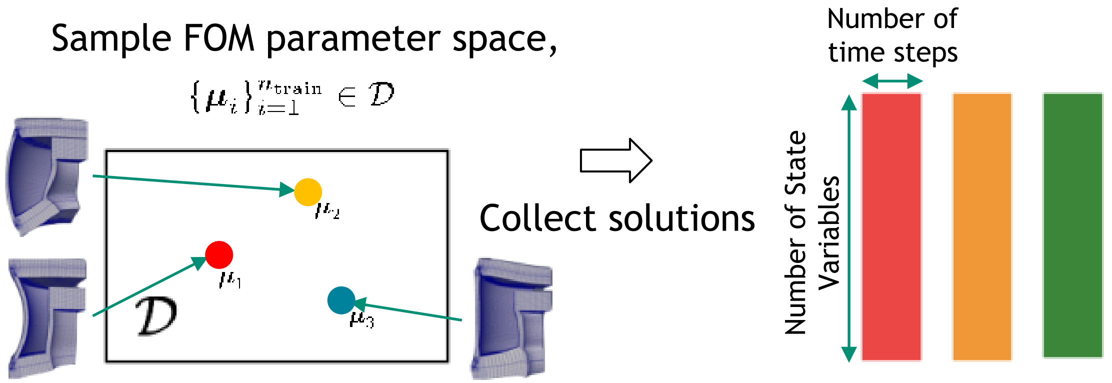
Projection-Based Reduced Order Models (ROMs)

ROMs = physics-based surrogates tied directly to a "full-order model" (FOM) that can enable full-field predictions in real time.

1. Data Acquisition

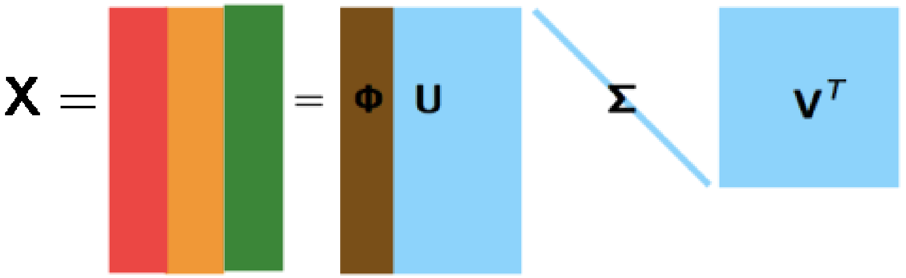
Sample FOM parameter space,

$\{\boldsymbol{\mu}_i\}_{i=1}^{n_{\text{train}}} \in \mathcal{D}$



2. Manifold Learning

Unsupervised Learning of Reduced Basis
(e.g. via Principal Component Analysis or Nonlinear Methods):



3. Projection and Reduction

Choose ODE Temporal Discretization

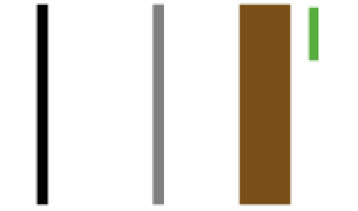
$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}; t, \boldsymbol{\mu})$$

$$\Downarrow$$

$$\mathbf{r}^n(\mathbf{x}^n; \boldsymbol{\mu}) = \mathbf{0}, \quad n = 1, \ldots, T$$

Reduce the Number of Unknowns

$$\mathbf{x}(t) \approx \tilde{\mathbf{x}}(t) = \boldsymbol{\Phi}\,\hat{\mathbf{x}}(t)$$



Minimize the Residual (Galerkin or Petrov-Galerkin Projection)

$$\underset{\hat{\mathbf{v}}}{\text{minimize}} \|\mathbf{r}^n(\boldsymbol{\Phi}\,\hat{\mathbf{v}}; \boldsymbol{\mu})\|_2$$

Projection-Based Reduced Order Models (ROMs)

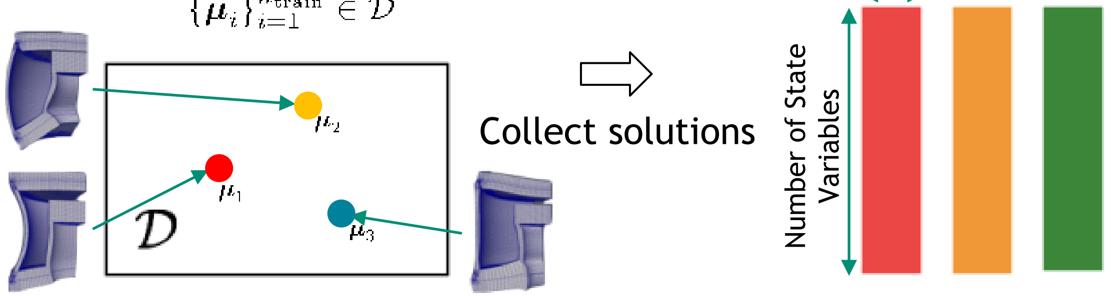
Challenge: *for non-linear problems, residual minimization requires operations scaling like FOM.*

1. Data Acquisition

Sample FOM parameter space,

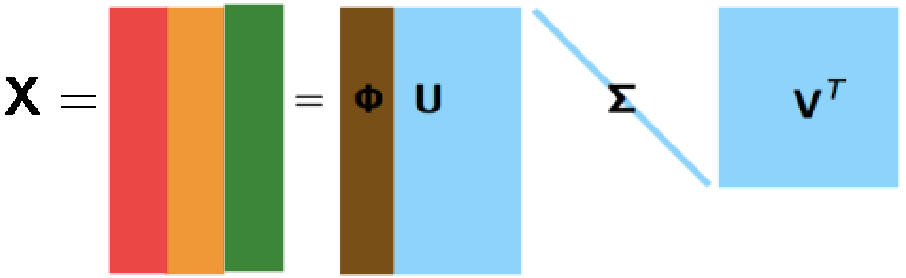
$\{\boldsymbol{\mu}_i\}_{i=1}^{n_{\text{train}}} \in \mathcal{D}$

Collect solutions



2. Manifold Learning

Unsupervised Learning of Reduced Basis
(e.g. via Principal Component Analysis or Nonlinear Methods):



3. Projection and Reduction

Choose ODE Temporal Discretization

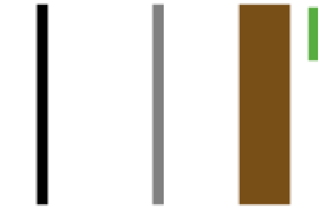
$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}; t, \boldsymbol{\mu})$$

$$\Downarrow$$

$$\mathbf{r}^n(\mathbf{x}^n; \boldsymbol{\mu}) = \mathbf{0}, \quad n = 1, \ldots, T$$

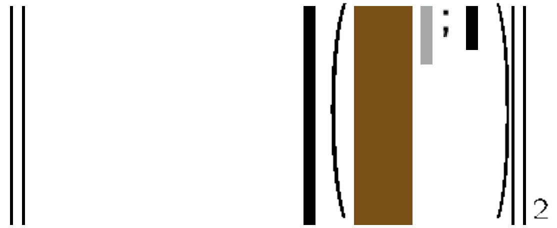
Reduce the Number of Unknowns

$$\mathbf{x}(t) \approx \tilde{\mathbf{x}}(t) = \boldsymbol{\Phi}\,\hat{\mathbf{x}}(t)$$



Minimize the Residual (Galerkin or Petrov-Galerkin Projection)

$$\underset{\hat{\mathbf{v}}}{\text{minimize}} \left\| \mathbf{r}^n(\boldsymbol{\Phi}\,\hat{\mathbf{v}}; \boldsymbol{\mu}) \right\|_2$$



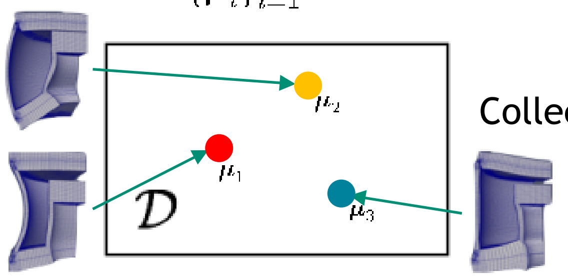
Projection-Based Reduced Order Models (ROMs)

Solution is hyper-reduction: *compute the residual on a small subset of the mesh, represented by matrix* ***A***

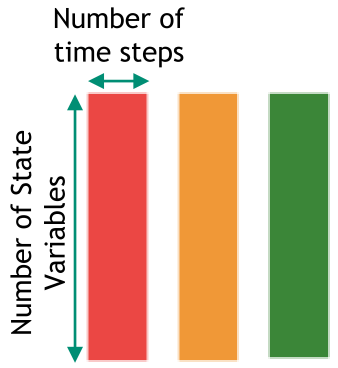
1. Data Acquisition

Sample FOM parameter space,

$\{\boldsymbol{\mu}_i\}_{i=1}^{n_{\text{train}}} \in \mathcal{D}$

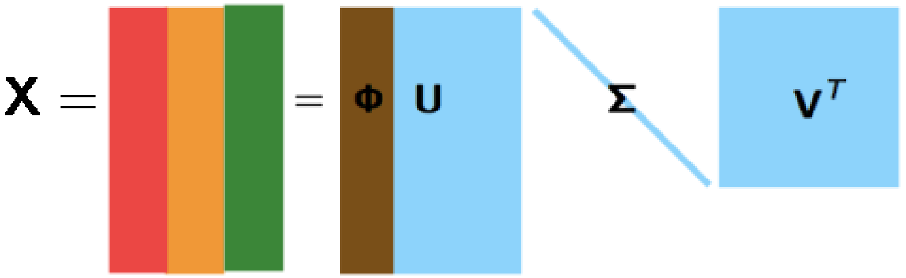


Collect solutions



2. Manifold Learning

Unsupervised Learning of Reduced Basis
(e.g. via Principal Component Analysis or Nonlinear Methods):



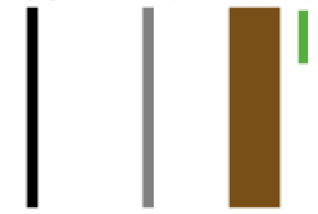
3. Projection and Reduction

Choose ODE Temporal Discretization

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}; t, \boldsymbol{\mu})$$
$$\Downarrow$$
$$\mathbf{r}^n(\mathbf{x}^n; \boldsymbol{\mu}) = \mathbf{0}, \quad n = 1, \ldots, T$$

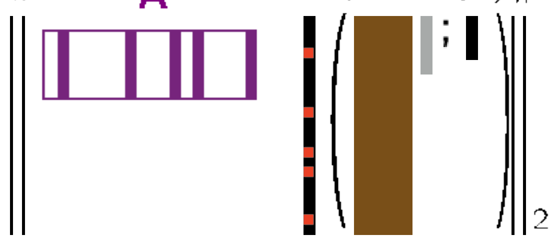
Reduce the Number of Unknowns

$$\mathbf{x}(t) \approx \tilde{\mathbf{x}}(t) = \boldsymbol{\Phi}\,\hat{\mathbf{x}}(t)$$



Minimize the Residual (Galerkin or Petrov-Galerkin Projection)

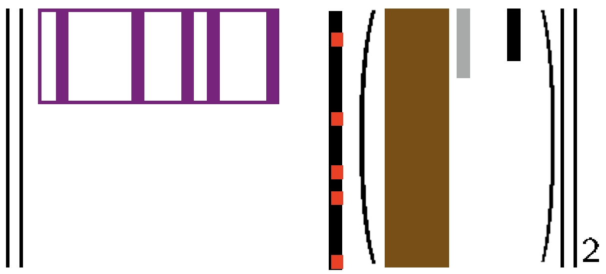
$$\underset{\hat{\mathbf{v}}}{\text{minimize}} \left\| \mathbf{A}\, \mathbf{r}^n(\boldsymbol{\Phi}\,\hat{\mathbf{v}}; \boldsymbol{\mu}) \right\|_2$$



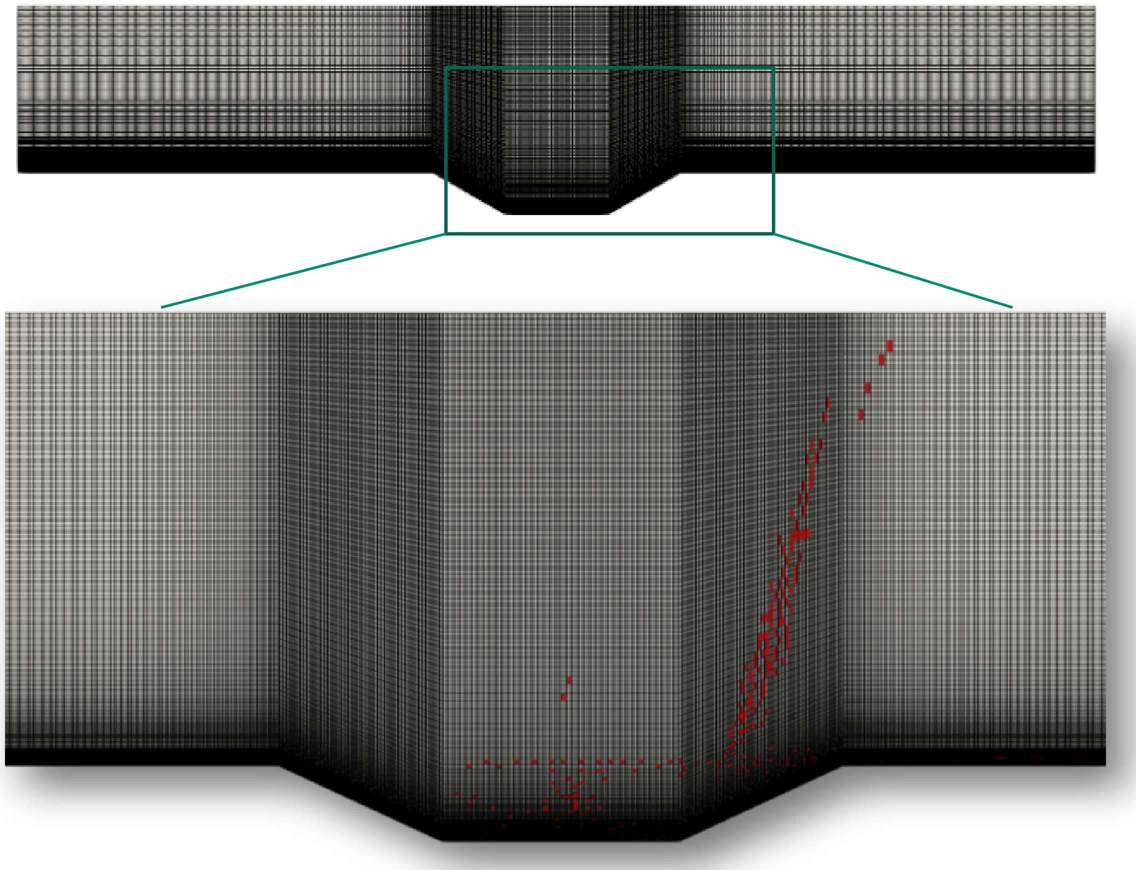
Hyper-reduction using a “Sample Mesh”

Minimize the Residual

$$\underset{\hat{v}}{\text{minimize}} \left\| \mathbf{A}\, r^n(\boldsymbol{\Phi}\,\hat{v};\mu) \right\|_2$$

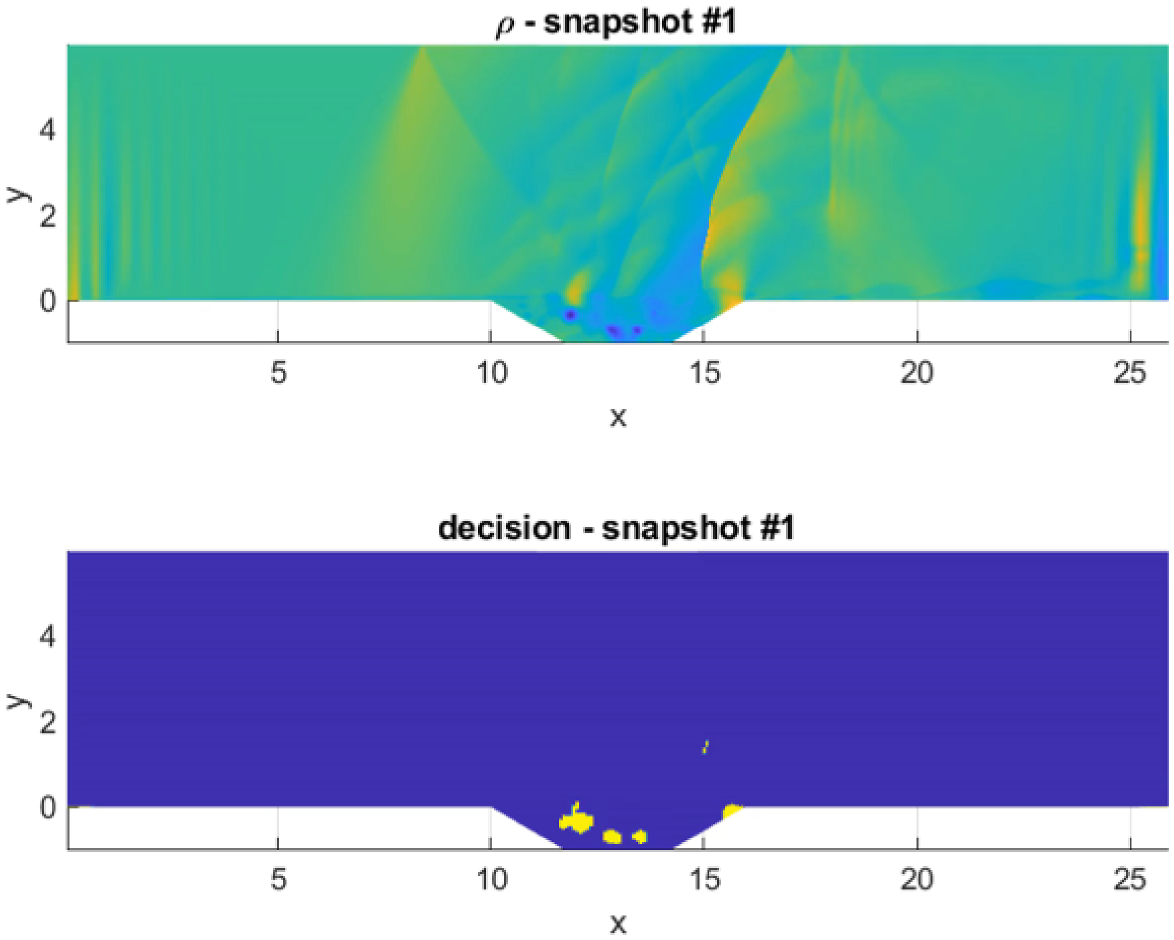


- A ***single “sample mesh”*** is typically computed using a simple greedy algorithm that minimizes reconstruction error of the non-linear function being approximated and ***that same sample mesh is used for hyper-reduction at all time-steps***.

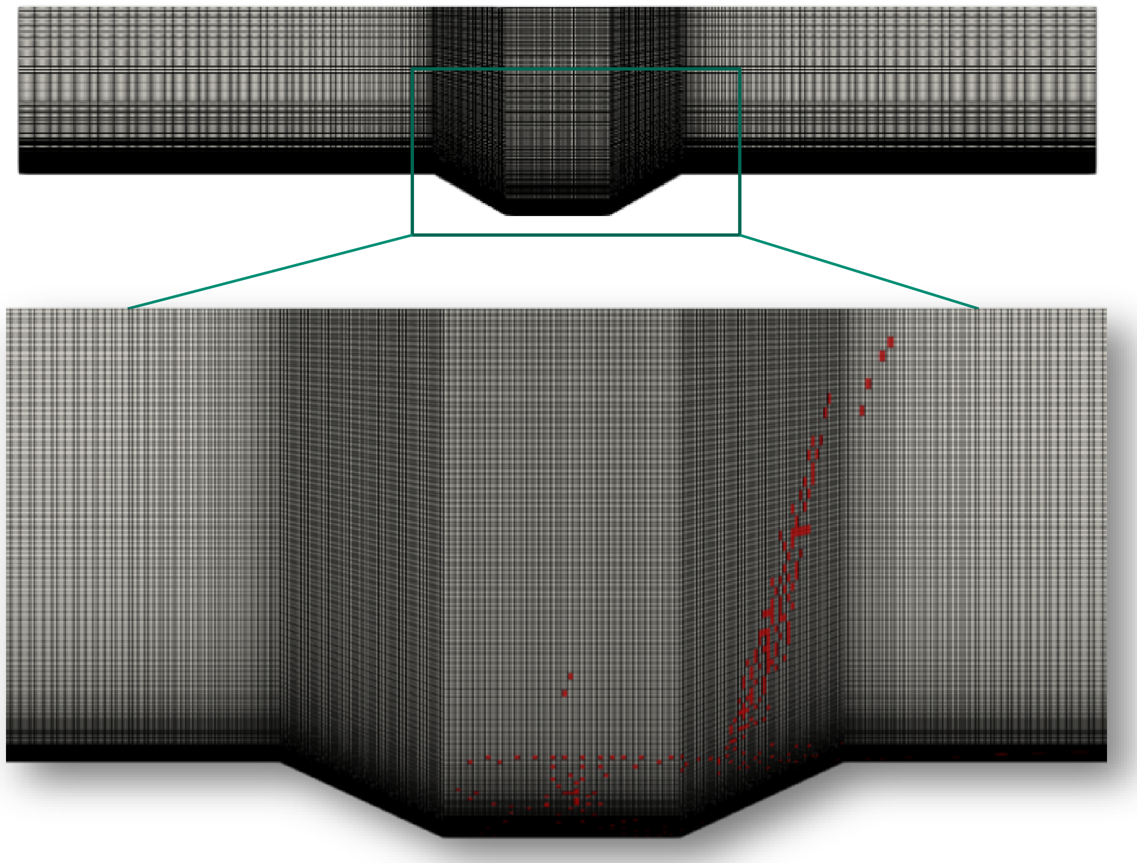


Static sample mesh obtained using q-sampling (Parish et al. 2020)

ISML algorithms have potential for revolutionizing hyper-reduction by calculating an *evolving* sample mesh!



Dynamic sample mesh containing ~4% of the total mesh points, obtained using ISML algorithms



Static sample mesh obtained using q-sampling (Parish et al. 2020)

Publications and 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.
- 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.
- 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.
- Davis, Warren Leon, Timothy Shead, Hemanth Kolla, Kevin Reed, Gabriel Popoola, W. Philip Kegelmeyer, Aditya Konduri, "The Potential of Integrated Machine Learning Algorithms for Tropical Cyclone Detection in Advanced Climate Modeling," *Poster*, American Geophysical Union Fall Conference 2019 in San Francisco, CA, USA, December 2019.
- Davis, Warren Leon, Timothy Shead, Hemanth Kolla, Kevin Reed, W. Philip Kegelmeyer, Gabriel Popoola. "In-Situ Machine Learning for Intelligent Data Capture on Exascale Platforms." *Presentation*, Artificial Intelligence for Robust Engineering & Science Workshop (AIRES), January 22-24, 2020.

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