In Situ Generated Probability Distribution Functions for Interactive Post Hoc Visualization and Analysis

Yucong Ye[1], Tyson Neuroth[1], Franz Sauer[1], Kwan-Liu Ma[1],
Giulio Borghesi[2], Aditya Konduri[2], Hemanth Kolla[2], and Jacqueline Chen[2]

1 VIDi VISUALIZATION & INTERFACE DESIGN INNOVATION

2 Sandia National Laboratories

VIS 2016

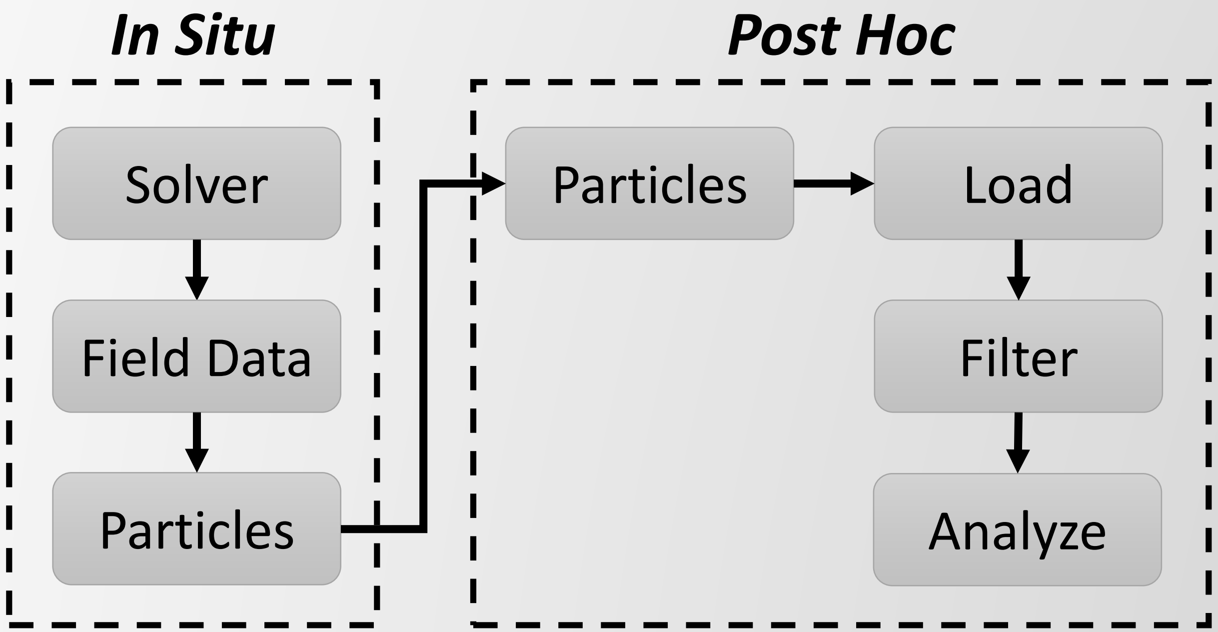
IEEE

Introduction: Problem Statement

- In situ processing to support post hoc analysis → particle selection
- "Trial and error" based exploration
  - Processing large datasets takes time
  - Selection criteria are complex
- How can we…
  - …leverage extra information available in situ?
  - …make particle/feature selection fast and efficient?

Introduction: Combustion Simulations and S3D

- Sandia National Laboratories
- S3D Combustion Simulation
  - Field Data (very high resolution)
  - Particle Data (more manageable)



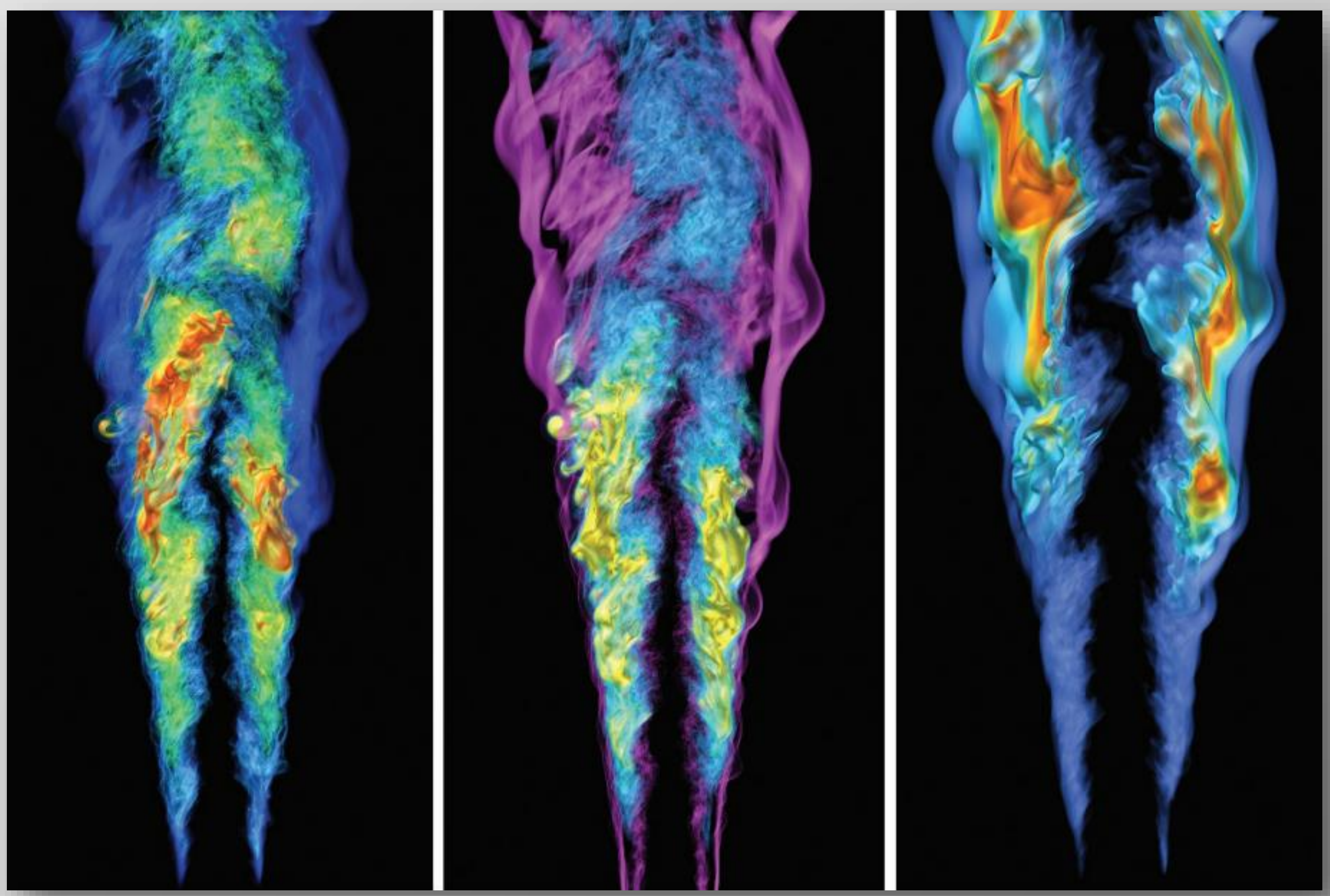
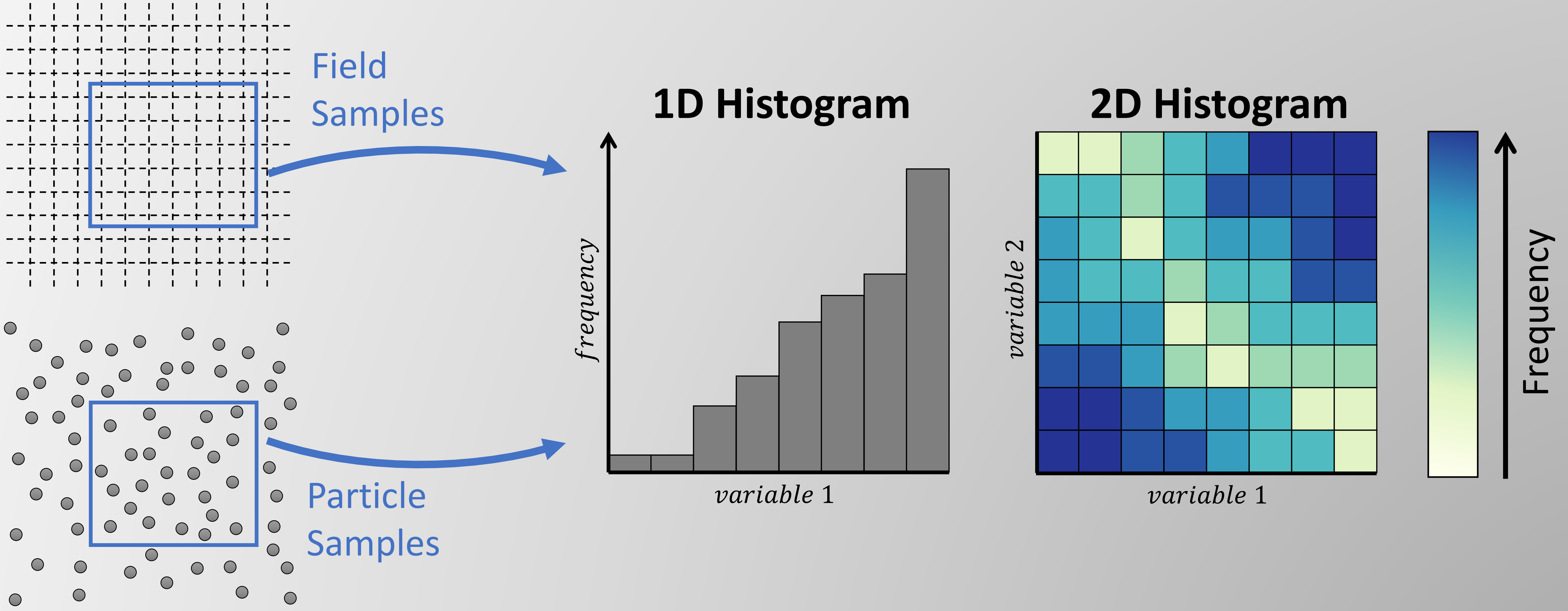


Image courtesy of [Yu et al. 2010]

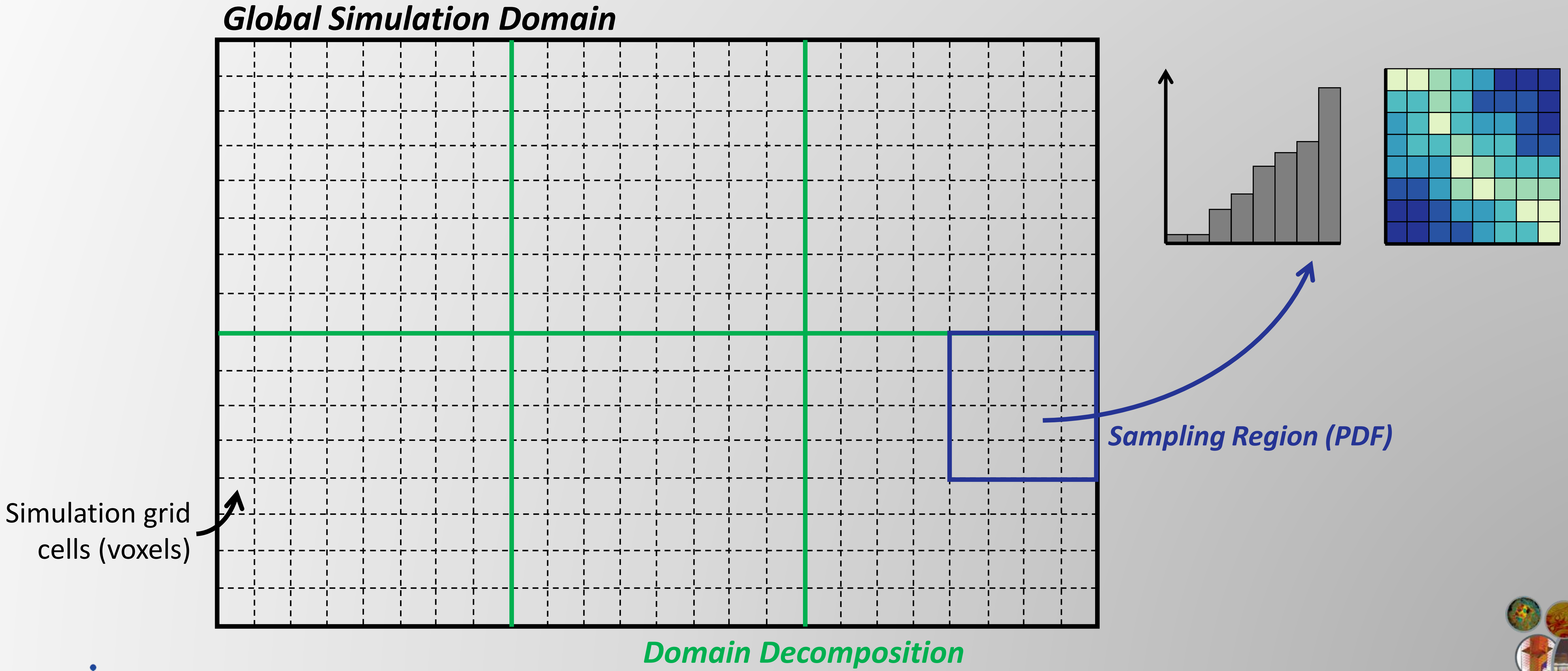
Background: Probability Distribution Functions

A data reduction tool that maintains distributions (i.e. a histogram)



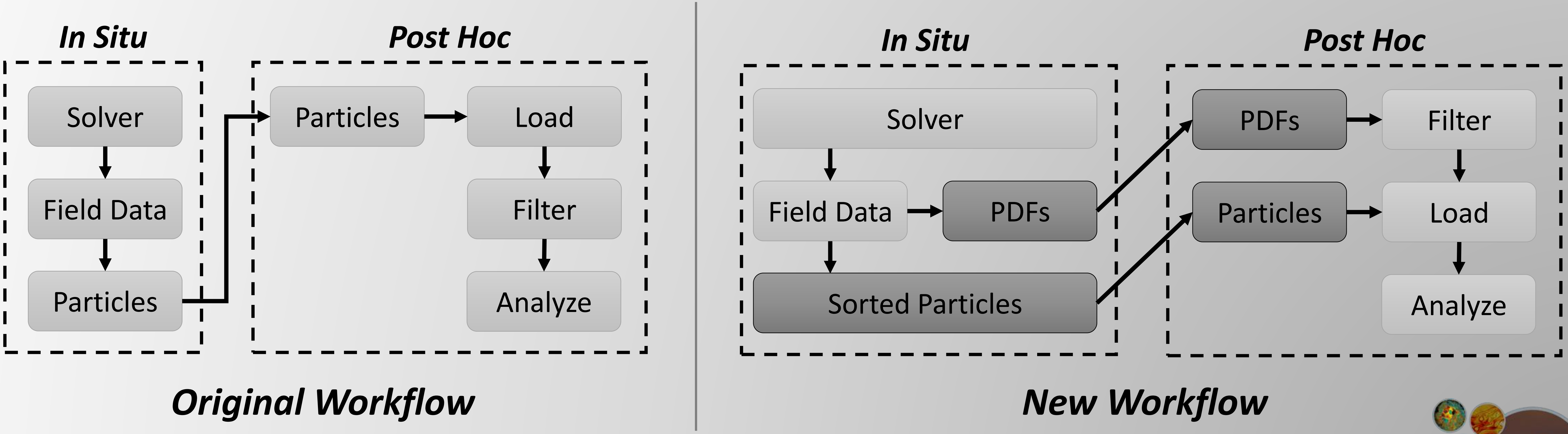
Background: Domain Subdivisions and Terminology

Insert a new level into the simulation hierarchy



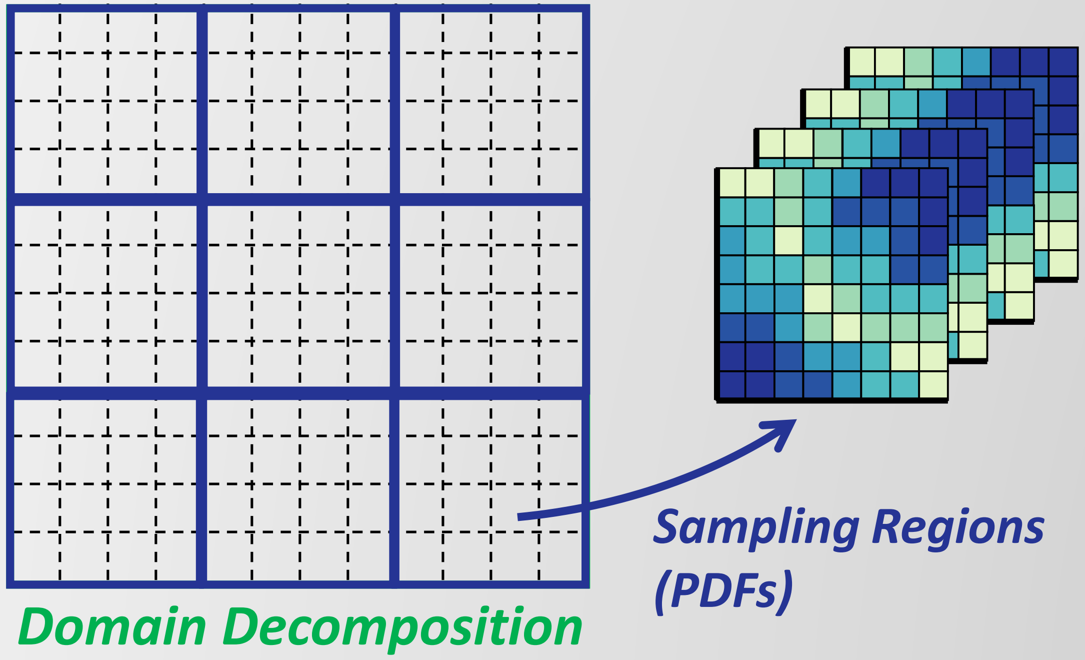
Methods: Overview and Workflow

- Modification to scientists’ normal workflow
  - Construct PDFs from field data (in situ)
  - Sort particles according to PDF sampling regions (in situ)
  - Perform filtering on PDFs to select particles quickly (post hoc)



Methods: PDF Generation (in situ)

- Routines are called within each domain decomposition
- PDFs may be 1D, 2D, or 3D
- Representation used (to minimize storage):
  - Dense matrix representation (frequency of all bins)
  - Sparse matrix representation (frequency of non-zero bins + location)

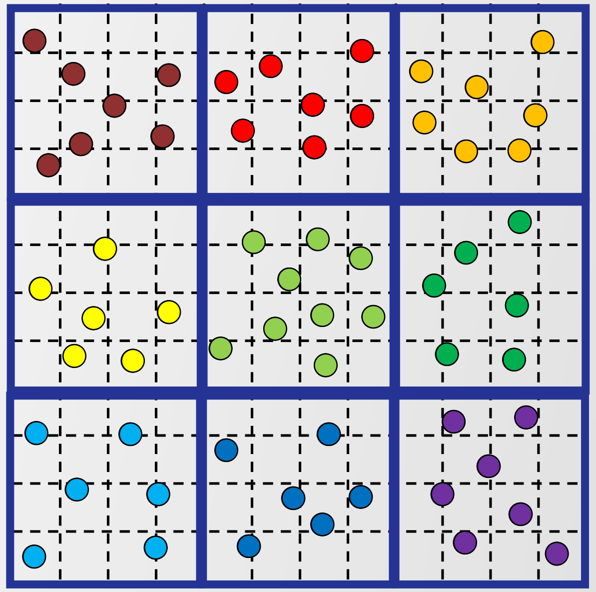


Dense: $\{f_1, f_2, f_3, \ldots, f_n\}$

Sparse: $\{f_1, l_1, f_2, l_2, \ldots, f_m, l_m\}$

Methods: Particle Sorting (in situ)

- Routines are called within each domain decomposition
- Particle in the same sampling regions are placed in contiguous chunks
- A separate set of indexes point to the start of each chunk



Sampling Regions (PDFs)

$$\{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, \dots\}$$

$$\Downarrow$$

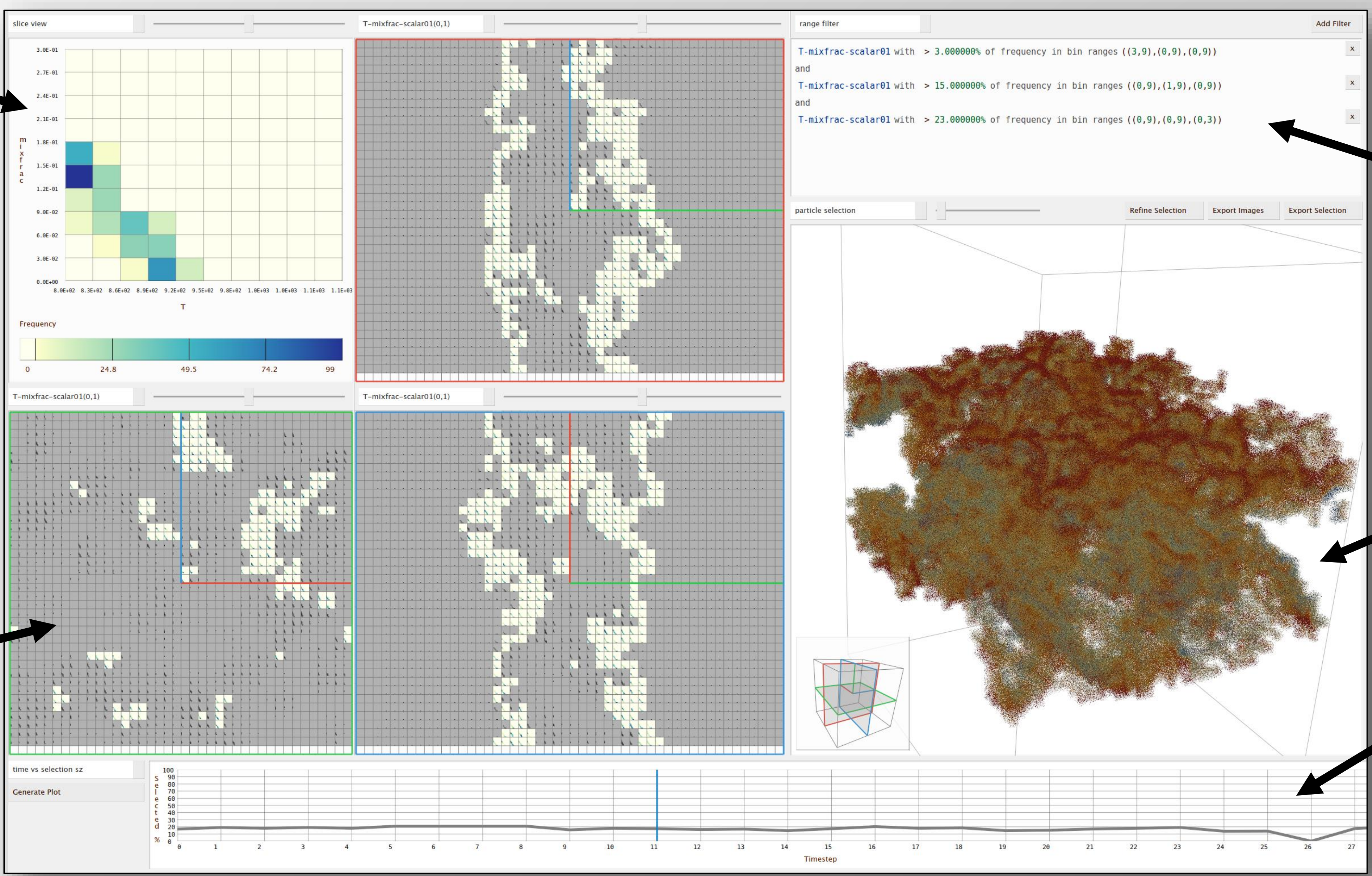
$$\{p_1, p_8, p_{10}, p_{12}, p_6, p_{13}, p_7, p_2, p_4, p_9, p_{11}, p_3, p_5, \dots\}$$

$$\{l_1, l_2, l_3, \dots\}$$

Methods: Analysis and Visualization Tool (post hoc)

Detail View

Filter Editor

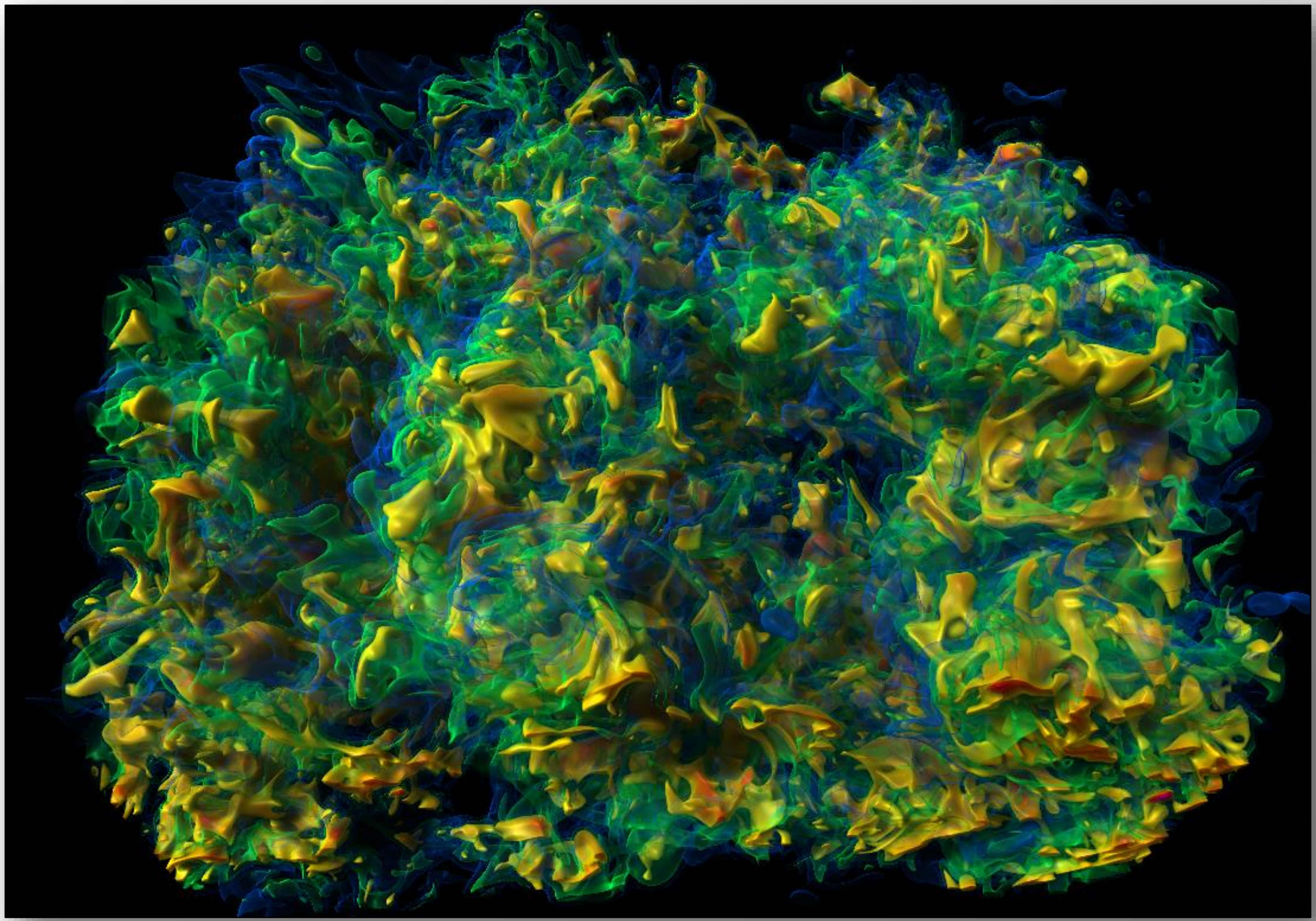


Particle View

Slice Views

Timeline View

Results: Test Dataset



- AirDodecane Dataset
  - n-dodecane & diluted air
  - 1400 x 1500 x 1000 cells
  - ~40 million particles
  - ~100 raw variables
- Large scale run on Titan
  - 80,000 computing processors

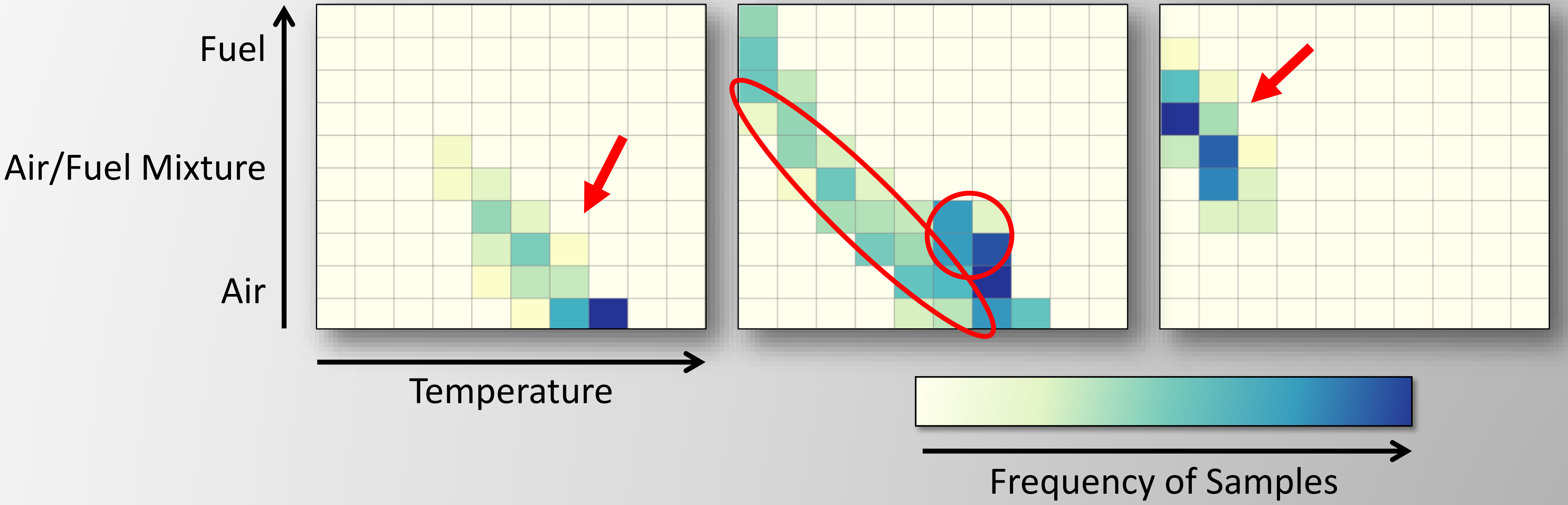
Results: Test Dataset

Distributions between variables describe system behavior

Outer Edges:
Only high temperature air is present

Select Locations:
Conditions are right for combustion

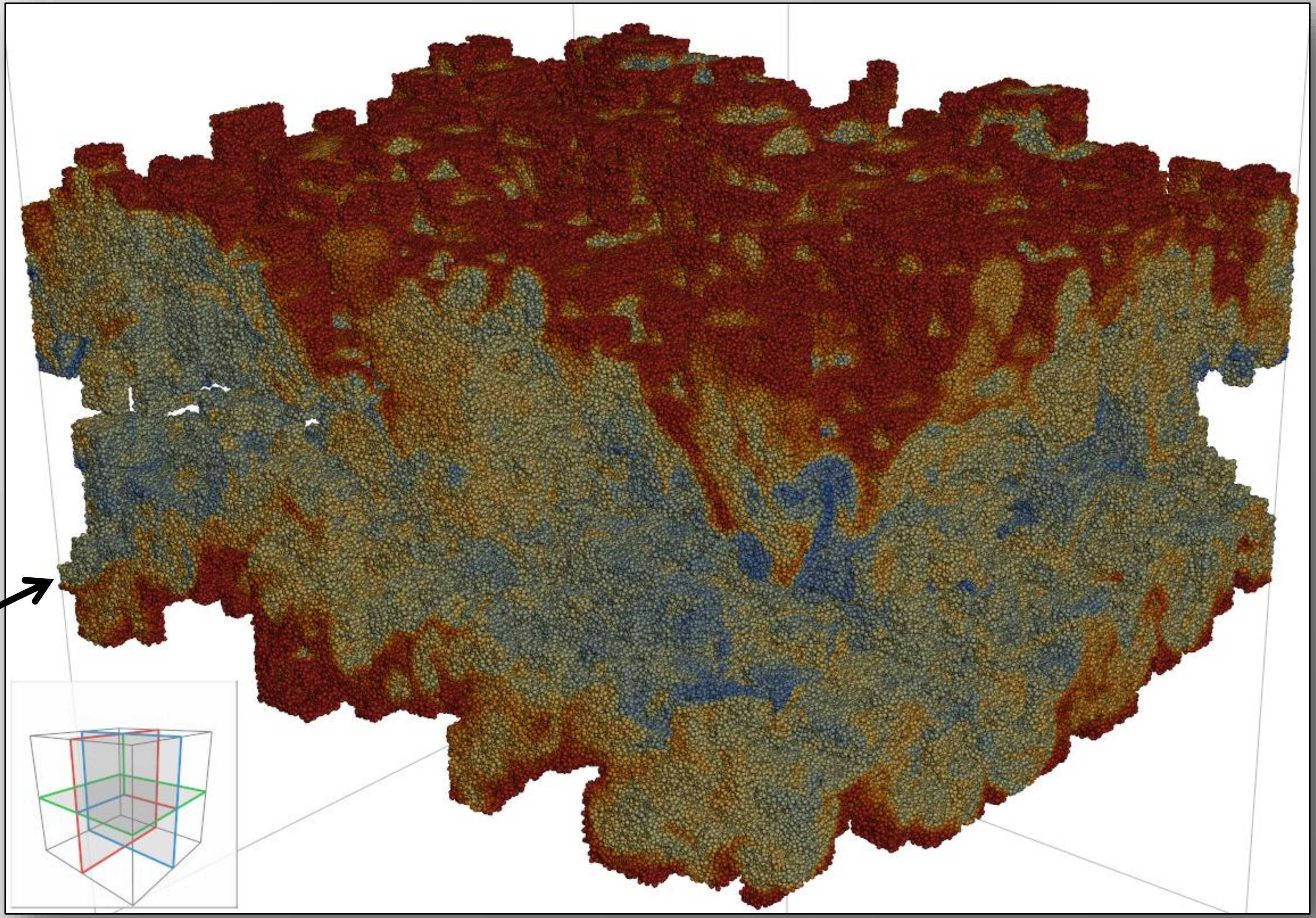
Domain Center:
Only low temperature fuel is present



Results: Test Dataset

Quickly select particles based on histogram distributions

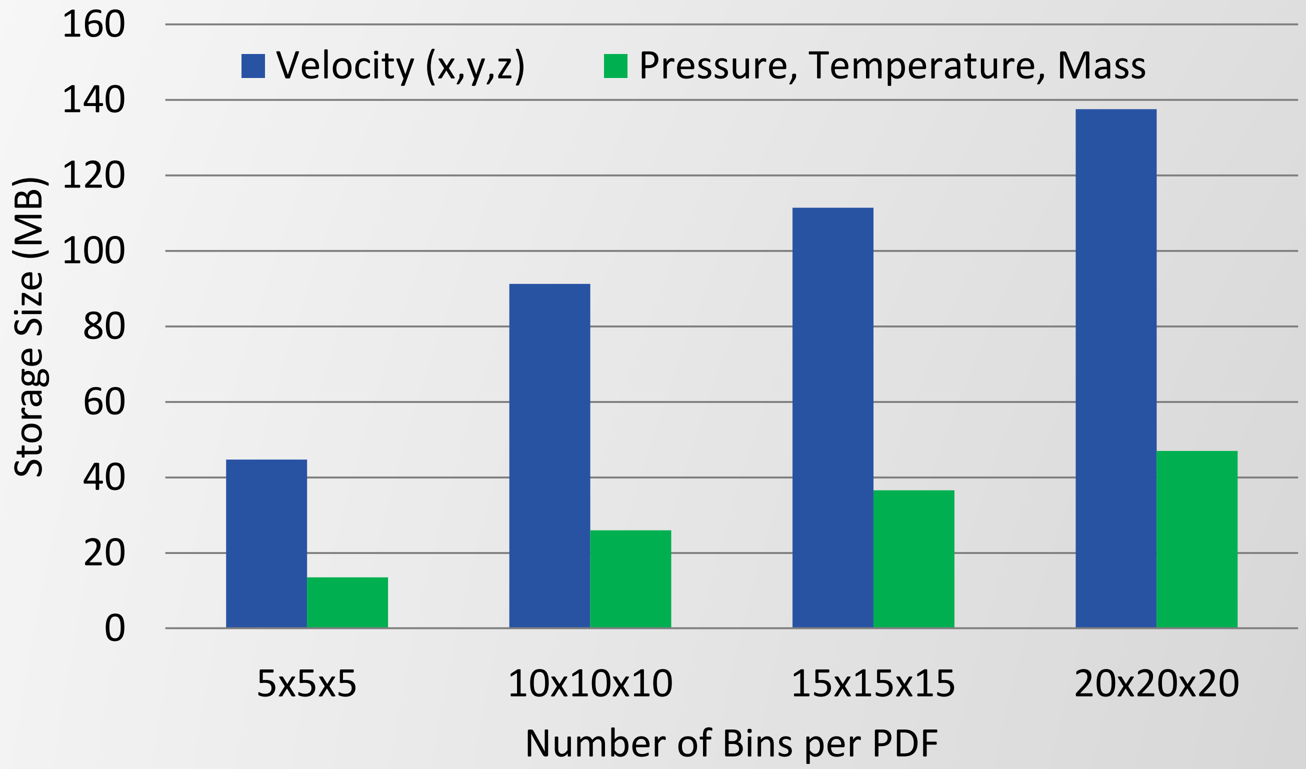
Example of a selection based on distributions between *mixture fraction*, *temperature*, and *scalar dissipation*



Spatial resolution limited by PDF/sampling region size

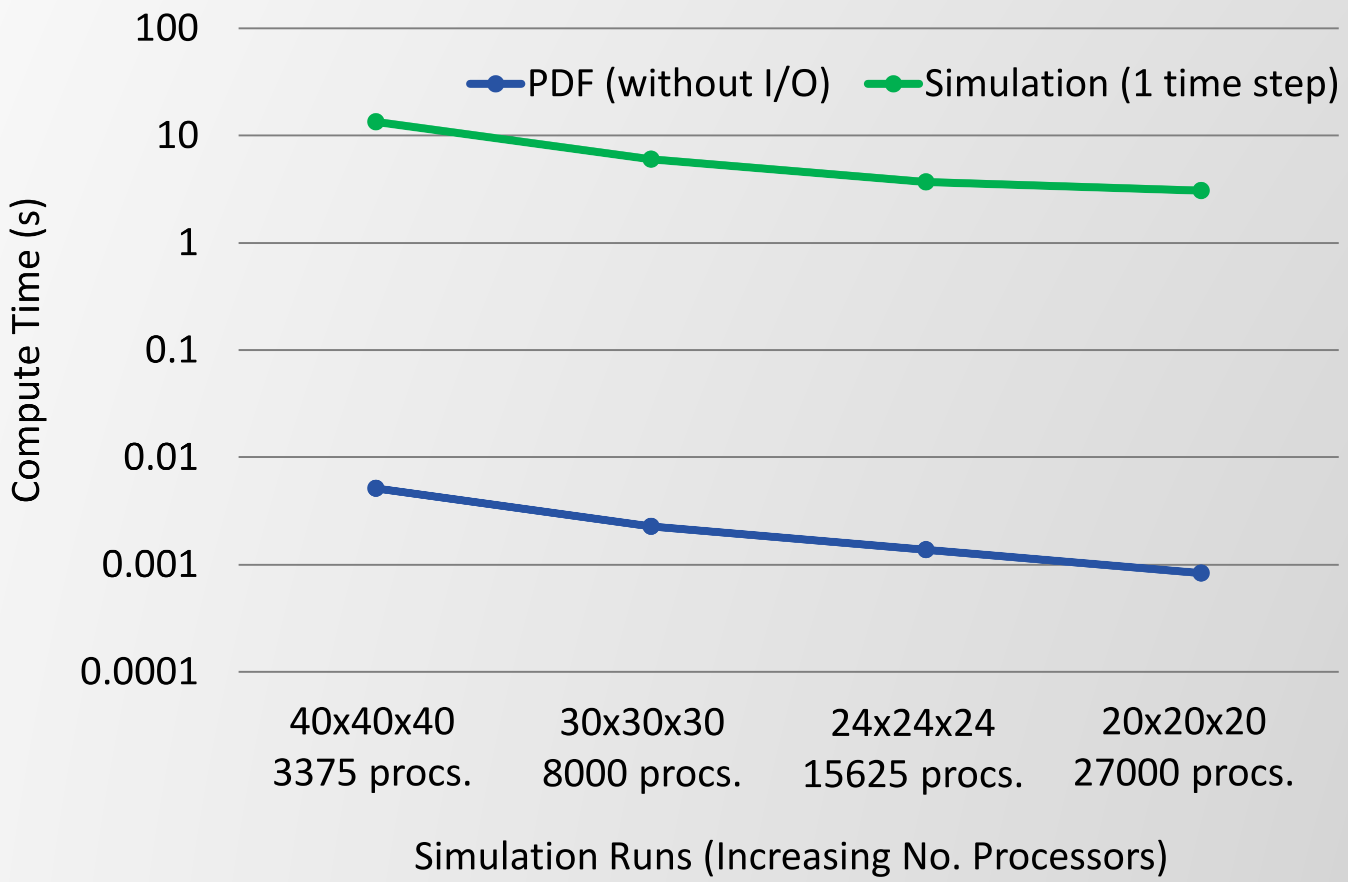
Results: PDF Storage Overhead

Performance testing: rerun simulation with varying sizes and parameters



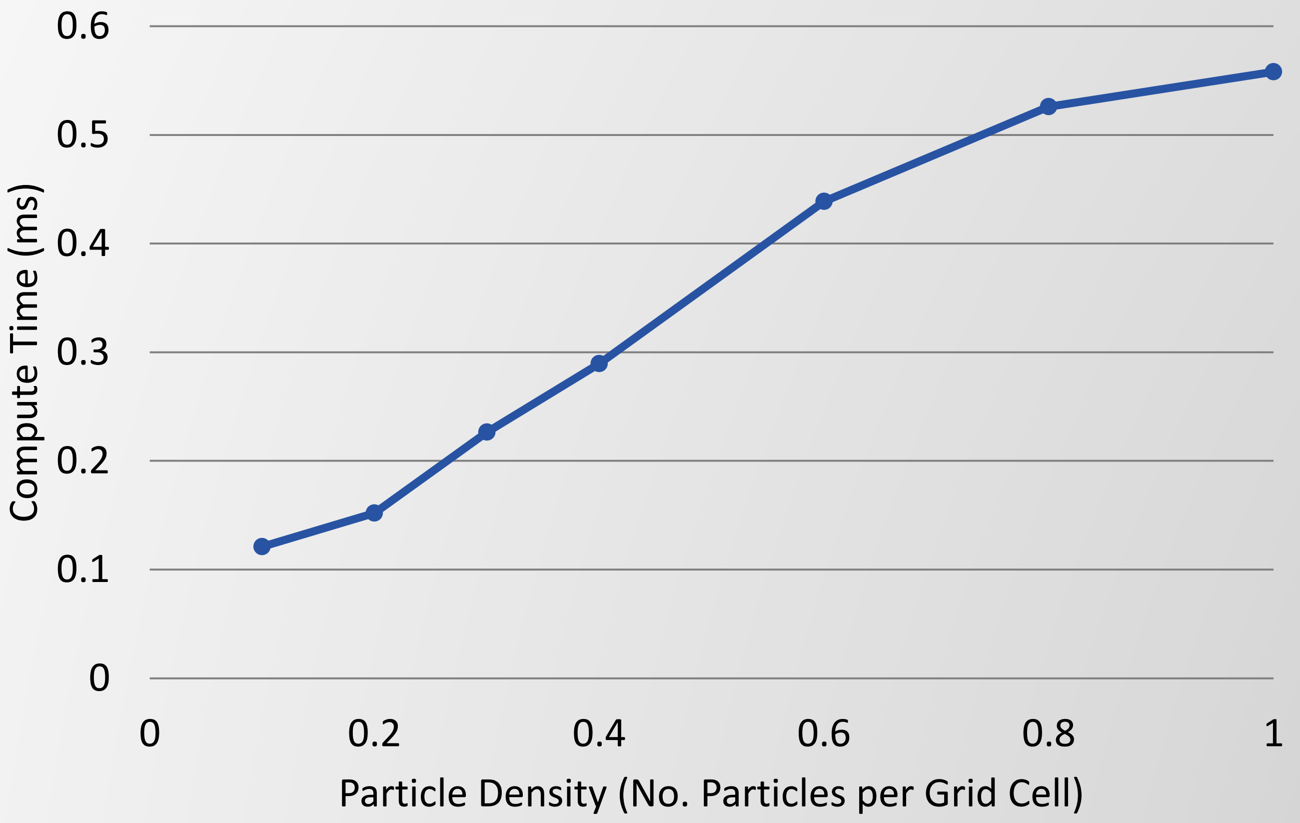
- Depends on data distributions
- More bins → more storage
- In general (per timestep):
  - ~100 MB for PDFs
  - Several GB for particles
  - PDFs use ~5% of particle storage

Results: PDF Generation Timing



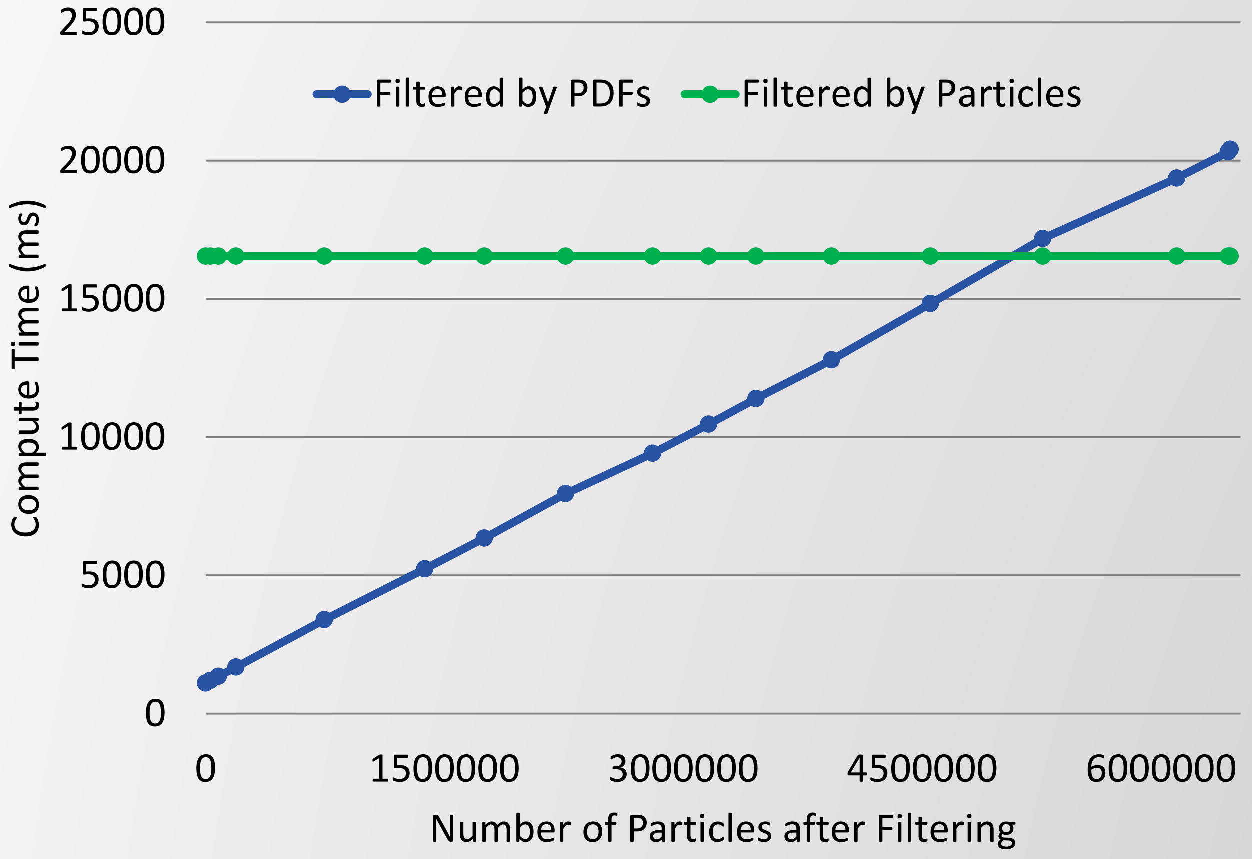
- Compare a simulation timestep with time to compute PDFs
- Horizontal axis: same problem size with increasingly finer subdivisions
- Simulation: 1 – 10 seconds
- PDFs: 0.001 – 0.01 seconds

Results: Particle Sorting Timing



- Storage overhead is negligible
- Sorting time depends on number of particles
- Sort time is a fraction of a ms

Results: Post Hoc Particle Selection Timing



- Filtering by particles: time remains constant
- Filtering by PDFs:
  - Time to process PDFs (constant)
  - Time to load the particles (varies)
- Need to load almost the full dataset before the PDF scheme becomes slower

Discussion: Limitations and Future Work

- Spatial resolution of PDFs/sampling regions limits selection
  - Secondary filtering step done on particle data directly
  - Data sizes will already be smaller from the PDF filtering
- Detailed particle analysis is done using other tools
  - Add temporal analysis of particle selections
  - Provide instant feedback when selection parameters change
- How can we use PDFs for importance driven time step selection?

Discussion: Summary

- Hybrid *in situ* and *post hoc* approach to particle selection
- Combustion research as driving application
- Users can extract representative particle subsets
  - Quickly and interactively to support “trial and error” based exploration
  - Very little overhead to the simulation or storage requirements
- Currently working towards improving the system with Sandia Natl. Labs
- Later plans to generalize the tools for other applications

Thank You
Questions?

Special thanks to:
- Our collaborators at Sandia National Laboratory
- U.S. Department of Energy
- U.S. National Science Foundation
- Lawrence Berkeley National Laboratory
- Oak Ridge National Laboratory

LDAV 2016
October 23rd
Baltimore, Maryland, USA

VIS 2016

IEEE