Extreme-Scale Distribution-Based Data Analysis

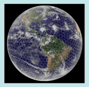
Han-Wei Shen（Lead PI), Gagan Agrawa, Huamin Wang
The Ohio State University

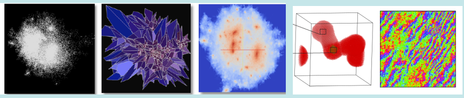
Tom Peterka
Argonne National Laboratory

Jonathan Woodring, Joanne Wendelberger
Los Alamos National Laboratory

Science Applications

- ❑ Climate: POP and MPAS-O (LANL)
- ❑ Superconductivity: SOScon (ANL)
- ❑ Cosmology: HACC (ANL)





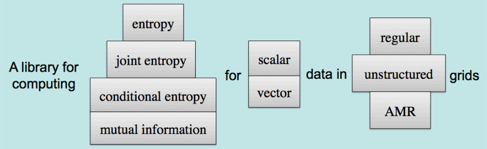
Why Distributions?

- ❑ A compact representation of data
  - Many statistics of the data can be derived
  - Information flow across the data analytics pipeline can be analyzed
  - Regions of high information content can be identified
  - Parameters of various visualization algorithms can be optimized
  - Allow detailed data analysis and inferences
- ❑ Support many needs of in situ data analysis
  - Data reduction
  - Data summarization
  - Data triage
  - Feature extraction and indexing

Software Library (ITL)

A C/C++ library for entropy and distribution computation for large scale datasets

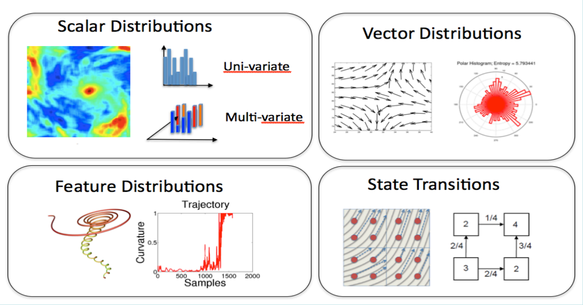
- ❑ Different information-theoretic measurement
- ❑ Distributed computation via MPI
- ❑ Support of various data types



A library for computing entropy / joint entropy / conditional entropy / mutual information for scalar / vector data in regular / unstructured / AMR grids

Research Goals

- Develop distribution-based data analysis and visualization techniques to support DOE's exascale applications
  - ❑ Efficient computation, representation, and query of data distributions
  - ❑ In situ data reduction, summarization, and triage
  - ❑ Distribution-based data analytics and visualization



Scalar Distributions

Vector Distributions

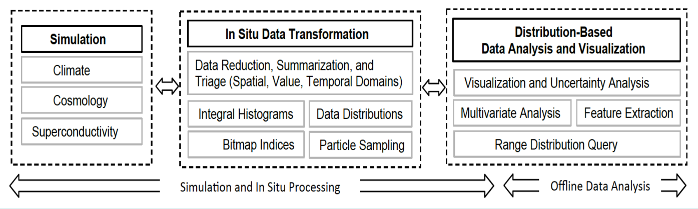
Feature Distributions

State Transitions

Research Tasks

- ❑ Computation and Representation of Distributions
  - Computing distributions from bitmap indices
  - Supporting efficient range distribution query
  - Statistics-preserving block decomposition
- ❑ Data Summarization, Reduction, and Triage
  - Spatial domain data summarization, reduction, and triage
  - Value domain summarization, reduction, and triage
  - Temporal domain summarization, reduction, and triage
- ❑ Distribution-based Visual Analytics
  - Distribution-based multivariate data analysis
  - Feature-driven view seleciton and control
  - Query-driven visual analysis with distributions

In Situ Data Analysis/Visualization Pipeline



Simulation: Climate; Cosmology; Superconductivity

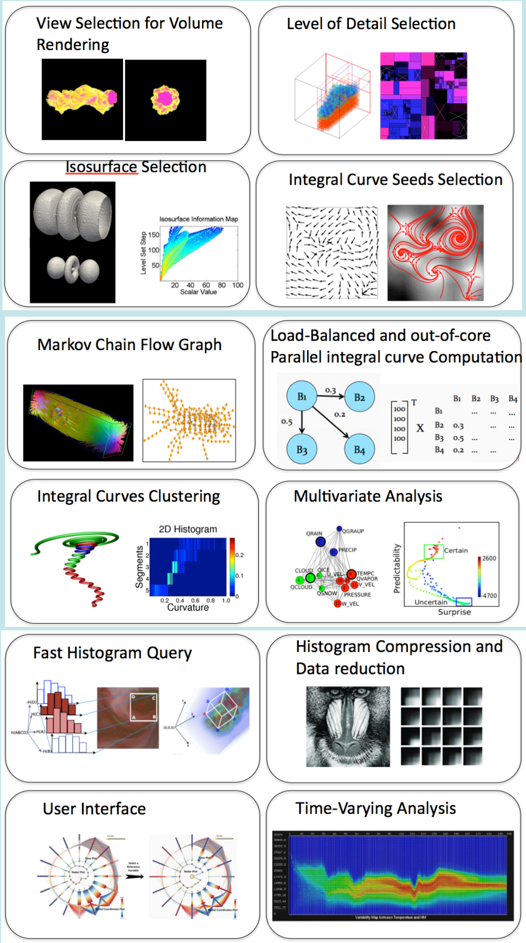
In Situ Data Transformation: Data Reduction, Summarization, and Triage (Spatial, Value, Temporal Domains); Integral Histograms; Data Distributions; Bitmap Indices; Particle Sampling

Distribution-Based Data Analysis and Visualization: Visualization and Uncertainty Analysis; Multivariate Analysis; Feature Extraction; Range Distribution Query

Simulation and In Situ Processing

Offline Data Analysis

Distribution-Based Visual Analytics



View Selection for Volume Rendering

Level of Detail Selection

Isosurface Selection

Integral Curve Seeds Selection

Markov Chain Flow Graph

Load-Balanced and out-of-core Parallel integral curve Computation

Integral Curves Clustering

Multivariate Analysis

Fast Histogram Query

Histogram Compression and Data reduction

User Interface

Time-Varying Analysis

Acknowledgement

This work was supported by DOE Office of Science, Advanced Scientific Computing Research, under award number DE-DC0012495, program manager Lucy Nowell.