**Extreme-Scale Distribution-Based Data Analysis**

Han-Wei Shen, The Ohio State University (Principal Investigator), shen.94@osu.edu

Gagan Agrawal, The Ohio State University (Co-Investigator), agrawal@cse.ohio-state.edu

Huamin Wang, The Ohio State University (Co-Investigator), whmin@cse.ohio-state.edu

Tom Peterka, Argonne National Laboratory (Co-Investigator), tpeterka@mcs.anl.gov

Jonathan Woodring, Los Alamos National Laboratory (Co-Investigator), woodring@lanl.gov

Joanne Wendelberger, Los Alamos National Laboratory (Co-Investigator), joanne@lanl.gov

**Project Description**

Scientists overview and identify regions of interest by transforming data into compact information descriptors that characterize simulation results and allow detailed analysis on demand. Among many existing feature descriptors, **statistical information** derived from data samples is a promising approach to taming the big data avalanche because data distributions computed from a population can compactly describe the presence and characteristics of salient data features with minimal data movement. The ability to computationally summarize and process data using distributions also provides an efficient and representative capture of information that can adjust to size and resource constraints, with the added benefit that uncertainty associated with results can be quantified and communicated. In this project, we will develop a distribution-based data analysis and visualization framework for optimizing in situ processing and analysis of extreme-scale scientific data. Our framework will consist of the following three unique, but tightly integrated, components:

**Computation, Representation, and Indexing of Distributions:** We transform simulation output into a compact representation of the data in the form of distributions (histograms, KDE, or GMM) to capture the main data characteristics. We also develop bitmap indexing techniques to perform fast histogram computation, and novel data structures to compute and store summary distribution information with minimum space overhead. Our goal is to support interactive query of distributions for any arbitrary region from a large data set, both for in situ and offline analysis applications.

**Data Summarization, Reduction, and Triage:** We study how to reduce analysis costs via the proxy knowledge provided by the distributions and the indices. Based on the characterization of data, we perform data summarization, reduction, and triage in spatial, value, and temporal domains. We also study how distributions alone can be used in extreme scale data visualization and analysis.



**Distribution-based Visual Analytics:** We develop distribution-based visual analytics algorithms to efficiently perform analysis and visualization of multivariate data sets. We also study how distributions can be used for feature detection and view prioritization, selection, and control. Finally, we study how the distributions computed in situ can be used for similarity-based feature comparison and search, uncertainty quantification of visualization output, and statistical inferences and correlations for feature identification. Figure 1 shows our distribution-based data analysis pipeline. The first two components will be performed in situ, while the third component can run either in in situ or in post processing.

The project brings together six diverse researchers, from one university (The Ohio State University) and two national laboratories (Argonne National Laboratory and Los Alamos National Laboratory), to tackle the crucial problems in extreme-scale scientific data management and visualization. We will deploy our distribution-based data analysis and visualization algorithms to support the DOE’s science applications. Specifically, we have selected three computational science applications, climate, cosmology, and superconductivity, to motivate the solutions in this proposal.

**Climate:** Climate studies are of high-impact, long-term interest for DOE scientific research. The applications of our research are POP (Parallel Ocean Program) and MPAS-O (Model for Prediction Across Scales-Ocean). In the simulations, there are multiple parameters and quantities of interest, over various time scales, encompassing large ensembles of data. The massive scale of data makes it daunting to explore time-varying or multivariate correlations across multiple simulations. Our research will assist analysts and computation allowing for targeted results through intelligent triage, saving time and resources. With our distribution representation of the data, we will automatically locate correlated data of interest at different spatial and temporal domains, and guide the climate scientists and analysts through a visual analytic process to understand the high-dimensional correlations across time, space, and ensemble runs.

**Cosmology:** Our computational cosmology application is a multi-petaflop high-resolution N-body code, HACC (Hardware/Hybrid Accelerated Cosmology Code). HACC has been co-designed to account for disruptive computational architectures, as it is founded on the principle that a large-scale computational framework must exploit architectural heterogeneity.

The code solves the six-dimensional Vlasov-Poisson gravitational equations using N-body particle methods, as opposed to regular grids or adaptive meshes. We will motivate and evaluate our work in particle sampling and distributions to work with HACC dark matter tracer particles. The resulting distributions will be used to perform targeted analysis such as range distribution query and feature extraction.

**Superconductivity:** Our superconductivity application, OSCon (Optimizing SuperConductor transport properties through large-scale simulation), simulates a high temperature superconductor to optimize current transport through the material. This intensive large-scale calculation is being deployed on leadership class hardware at ANL. Dynamic processes occurring inside the material (i.e., vortices) govern the simulation outcomes, and interpreting simulation data necessitates complex in situ analysis to find, track, correlate, and aggregate statistics of the vortex dynamics, while the data are still in memory. We will use the vortices to motivate our work in distribution-based feature extraction to accelerate in situ data summarization and directed visual analytics.

   