An Information-Theoretic Framework for Enabling Extreme-Scale Science Discovery

Han-Wei Shen (Technical Lead), The Ohio State University
Rob Ross, Tom Peterka, Argonne National Laboratory
Yi-Jen Chiang, Polytechnic Institute of New York University

Overview

As scientists eagerly anticipate the benefits of extreme-scale computing, roadblocks to science discovery at scale threaten to impede their progress. The disparity between computing and storing information, and the gap between stored information and the understanding derived from it, are two of the main barriers to success. **This project addresses two difficulties faced by computational scientists. The first is deciding what data are the most essential for analysis, given that only a small fraction can be retained. The second is transforming these data into visual representations that rapidly convey the most insight to the viewer.** We will quantify the amount of information in data using information-theoretic approaches. Computing the information entropy of data allows decisions to be made as to how data should be stored and subsequently analyzed. Data saliency will further be used to inform and steer visualization algorithms automatically, including temporal analyses, and it can enable new types of analyses to be performed. **We will construct a data analysis and visualization framework based on information theory that allows us to evaluate the information content of simulation output, and we will test our approaches in applications that represent the next generation of extreme-scale science.** We will work together with scientists to evaluate the results of our information-theoretic algorithms. With these tools, scientists will be able to preserve important and discard irrelevant data, enabling them to see results sooner. Informed visualization algorithms will generate more meaningful displays. This will result in more knowledge, faster, and will impact decisions critical to the mission of the Department of Energy.

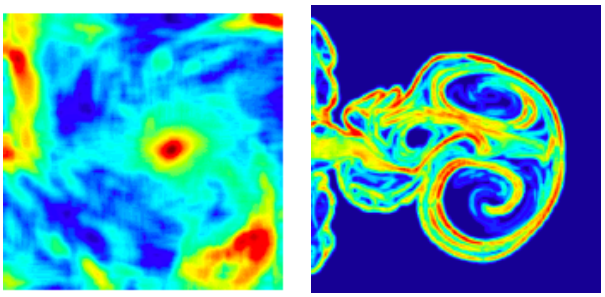

Figure 1: A scalar field (left), and its corresponding entropy measure shown on the right.

Impact

This research has the potential to impact any computational science domain that is data-intensive and faces storing, analyzing, and visualizing vast quantities of data. Initially, we will work with two science applications, Nek5000 (a Navier-Stokes solver) and FLASH (an adaptive mesh code for astrophysics and cosmology), to help motivate the solutions in this project and to evaluate their efficacy. We selected these as exemplars of current petascale applications that have the potential to scale further in coming years. They share several characteristics of the extreme scale execution, namely managing bounded resources, exploiting locality, and executing at a high degree of parallelism. Assisting scientists to store data and to generate visualizations according to their information content will enable the preservation of the most salient features without incurring prohibitive computation and storage costs.

THE OHIO STATE UNIVERSITY
Argonne NATIONAL LABORATORY
NYU·poly
U.S. DEPARTMENT OF ENERGY

Current Activities

To apply our theoretical findings to the benefit of knowledge discovery from exascale data, we have started building up a parallel yet lightweight library called Information Theoretic Library **(**ITL) that is to be run on DOE's leadership computing facility. **Our objective is to make this library generic in nature so that it can be seamlessly integrated into large-scale simulations to provide in-situ data analysis and visualization**. To keep our library flexible and easy-to-incorporate, we have carefully adopted some design principles. First, it is developed as a template-based library containing generic classes that can be customized for different types of data and metrics. We plan to include support for a wide range of data and grid types, namely, structured, unstructured and adaptive grids for both scalar and vector data. At present, the library is capable of parallel computation of a rich set of metrics including Shannon's entropy, joint entropy etc. We believe that our generic API can be easily extended to support even more complex data and more advanced metrics. In addition to offering versatility, we also plan to keep the memory footprint for the metric computation as low as possible, since the cost of information theoretic analysis should be barely perceivable to the entire visualization pipeline. Another important goal that we plan to achieve is scalability in terms of computation time. Our initial results (as presented in Figure 2) show promise in that context.

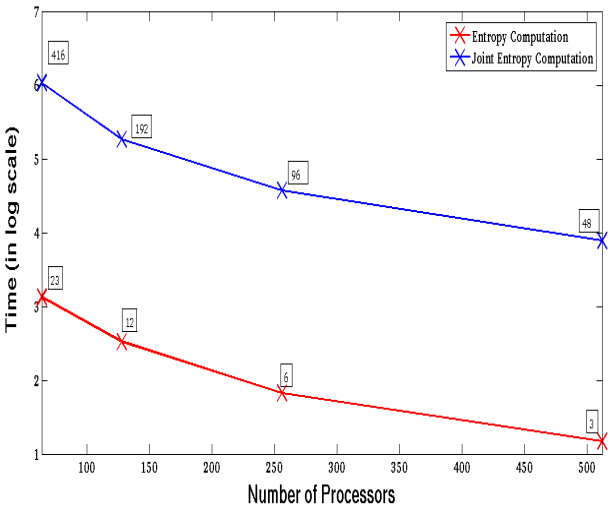

Figure 2: Performance of ITL run on NERSC's Franklin (Cray XT4). Our initial results showed that satisfactory scalability can be achieved.

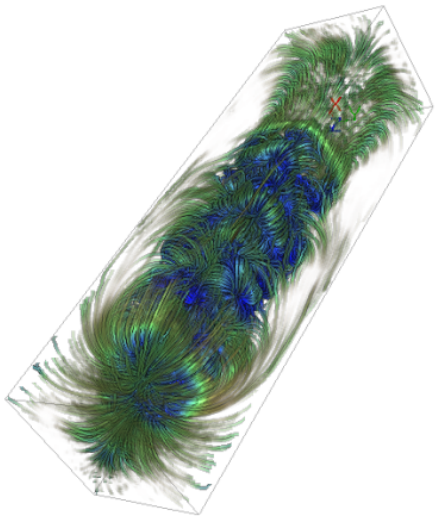

Figure 3: An image generated using our information-theoretic streamline placement algorithm.

In addition to software building, we are also developing novel visual analysis techniques to leverage the result of entropy-based information analysis. One such example is to use the concept of conditional entropy to assist streamlines placement so that the amount of information returned by the visualization is maximized. Our information-theoretic framework allows us to quantitatively evaluate the effectiveness of visualization, which can in turn assist in optimizing the parameters of various visualization algorithms. In addition to this example, we have also developed techniques based on information theory for automatic view selection for volume and streamline rendering.

For more information, please contact Han-Wei Shen <hwshen@cse.ohio-state.edu>.

THE OHIO STATE UNIVERSITY

Argonne NATIONAL LABORATORY

NYU·poly

U.S. DEPARTMENT OF ENERGY