

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

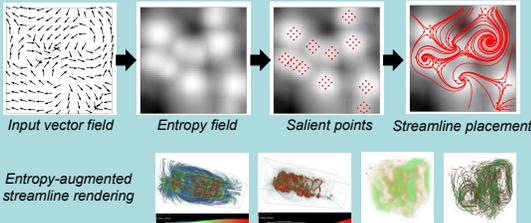
Abon Chaudhuri, Teng-Yok Lee, Han-Wei Shen
The Ohio State University

Tom Peterka
Argonne National Laboratory

Cong Wang, Tiantian Xu, Bo Zhou, Yi-Jen Chiang
Polytechnic Institute of NYU

Applications in Vector Field Analysis

- Compute the entropy based on the vector orientations
- High entropy near the critical points
- Seed streamlines at these points and advect through vector field

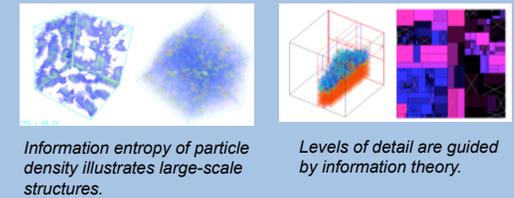


Project Overview

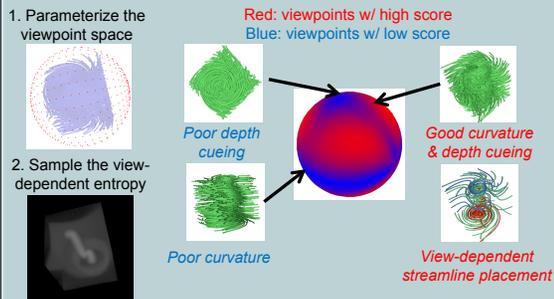
- **Research goals**
 - ❑ Create quantitative data analysis models to analyze information flow
 - ❑ Minimize information losses and maximize quality of analysis
 - ❑ Transform data into effective representations that convey the most insight
- **Specific aims**
 - ❑ Data triage with precise quality indicators for prioritized data retrieval
 - ❑ Streamlined parameter selection for visualization and analysis algorithms
 - ❑ In situ data analysis and reduction
- **Information-theoretic approach**
 - ❑ Quantify information content using information entropy measures
 - ❑ Steer analysis of data based on information saliency
 - ❑ Focus attention on small percentage of informative data

Applications in Cosmology

Simulations: FLASH, HACC
Application: Dark matter and energy
Goal: Feature identification quality
Impact: Dark matter accounts for 80% of all matter content and may hold the key to understanding the distribution of galaxies in the universe.



Applications in Viewpoint Selection

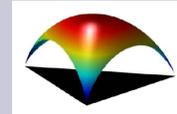


What is Information Theory?

- Study of fundamental limits to reliably transmitting messages through a noisy channel
- Model messages as random variables whose value is taken from a sequence of symbols
- Information content can be measured by Shannon's Entropy

Shannon's Entropy

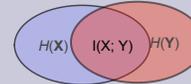
If a random variable takes a sequence of symbols $\{a_1, \dots, a_n\}$ with probabilities $\{p_1, \dots, p_n\}$, the average amount of information expressed by the random variable is Shannon's entropy.



$$H(x) = - \sum_{i=1}^n p_i \log p_i$$

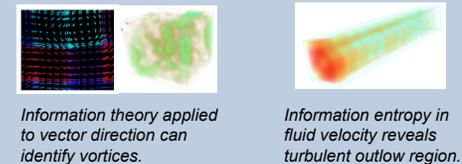
Mutual Information

Given two data sources X and Y, their mutual information $I(X; Y)$ indicates amount of information overlap.



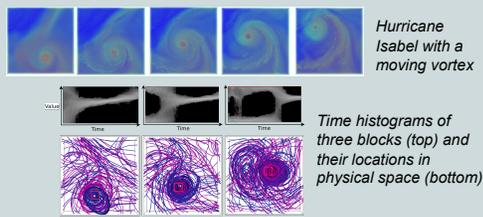
Applications in Fluid Dynamics

Simulation: Nek5000
Application: Thermal hydraulics
Goal: Vortex detection, velocity distribution
Impact: Turbulence affects cooling efficiency and safety in nuclear reactor cores.



Applications in Spatio-Temporal Analysis

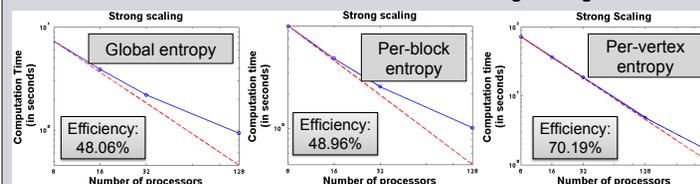
- Compute time histograms from large-scale time-varying data
- Capture dynamic behavior of data over space and time



ITL: Information Theory Library

- A C/C++ library for entropy and distribution computation for large-scale datasets
- ❑ Different information-theoretic measurements
 - ❑ Distributed parallel computation via DIY
 - ❑ Support for various data types and grids

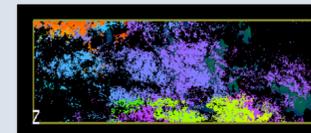
Strong Scaling Performance



Platform: Cray XT4 supercomputer with 38,128 compute cores, 78 TB of memory, and 436 TB of disk
Dataset: NEK5000 CFD thermal hydraulics data, resolution 512³

Applications in Climate

Simulation: Madden-Julian Oscillation (MJO)
Application: Tropical storm systems
Goal: Identify and track MJO system
Impact: Large-scale atmospheric systems like the MJO impact global climate.



Salient temporal trends computed using information theory affect the Madden-Julian Oscillation.