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

Han-Wei Shen
The Ohio State University

Motivation

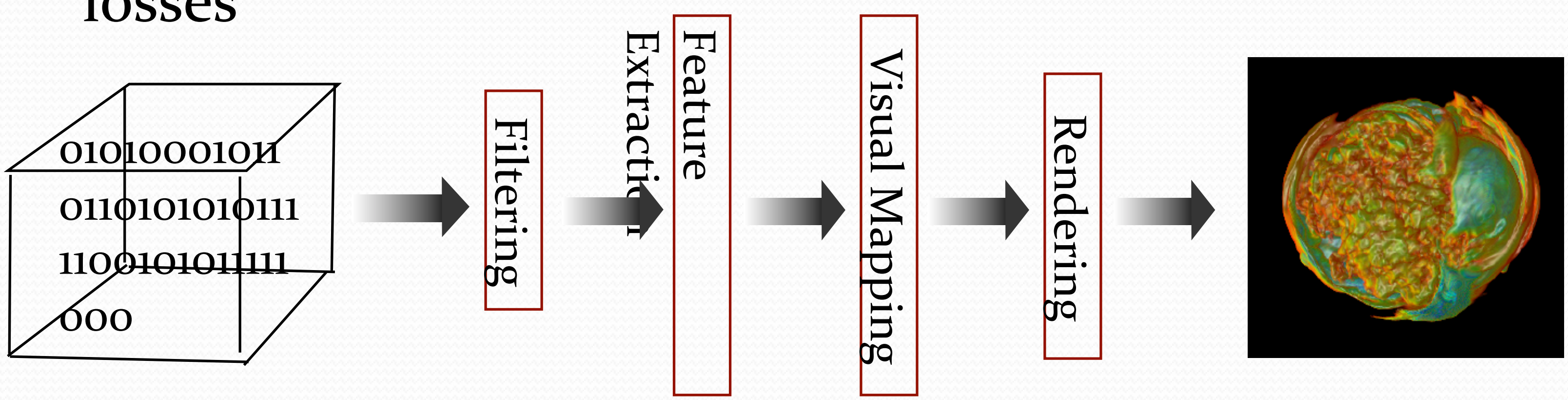
- The speed and capacity of storage cannot keep pace with the advance of computation power
  - I/O becomes a major bottleneck
- Throw away and triage data
  - It is often difficult to decide what data are the most essential for analysis
- In-situ visualization
  - The parameter space for visualization algorithms is often huge

Visual Analytic Sample Questions

- Data reduction and triage
  - Where are the most salient regions?
  - What resolution to use?
- Visual mapping
  - How to choose the best algorithm parameters?
  - How much information in the data is being revealed by the visualization?
- Image Analysis
  - Is this a good view point?
  - Is this a good transfer function?

Approach

- Develop a quantitative model to measure the flow of information across the entire data analysis and visualization pipeline
  - Quantify the information content in the data set
  - Measure the amount of information losses in each stage of the visualization pipeline
  - Choose parameters that can minimize the information losses



Visualization pipeline

Information Theory

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

Shannon's Entropy

- The random variable takes a sequence of symbols $\{a_1, a_2, a_3, \ldots, a_n\}$ with probabilities $\{p_1, p_2, p_3, \ldots, p_n\}$
- The information content of each symbol a_i is defined as

$$\log(1/p_i) = -\log p_i$$

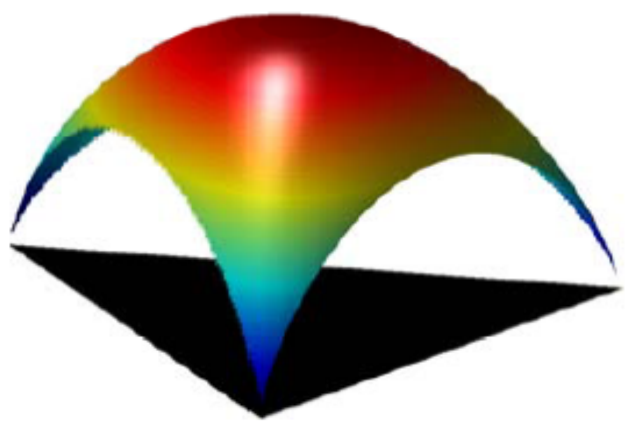
- The average amount of information expressed by the random variable is

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

Properties of Shannon's Entropy

- Entropy is to measure the average uncertainty of the random variable
- Entropy is a concave function, which has a maximum value when all outcomes are equally possible

$$p_1 = p_2 = p_3$$

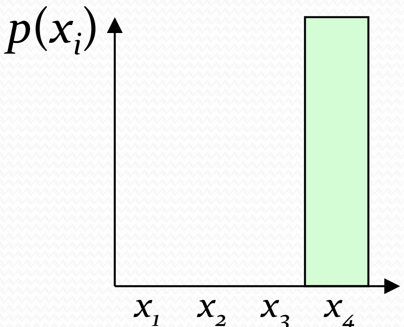


Information and Entropy

Information theory: quantitatively measures the amount of information contained in a data source

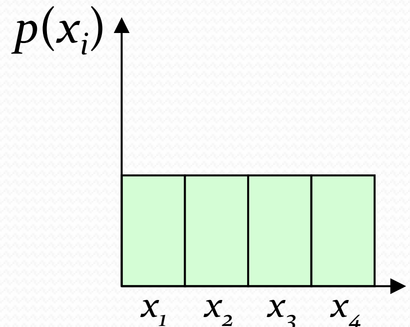
Entropy of **X** : $H(\mathbf{X}) = -\sum p(x_i) \log_2 p(x_i)$

Minimal Entropy



Only one possible outcome

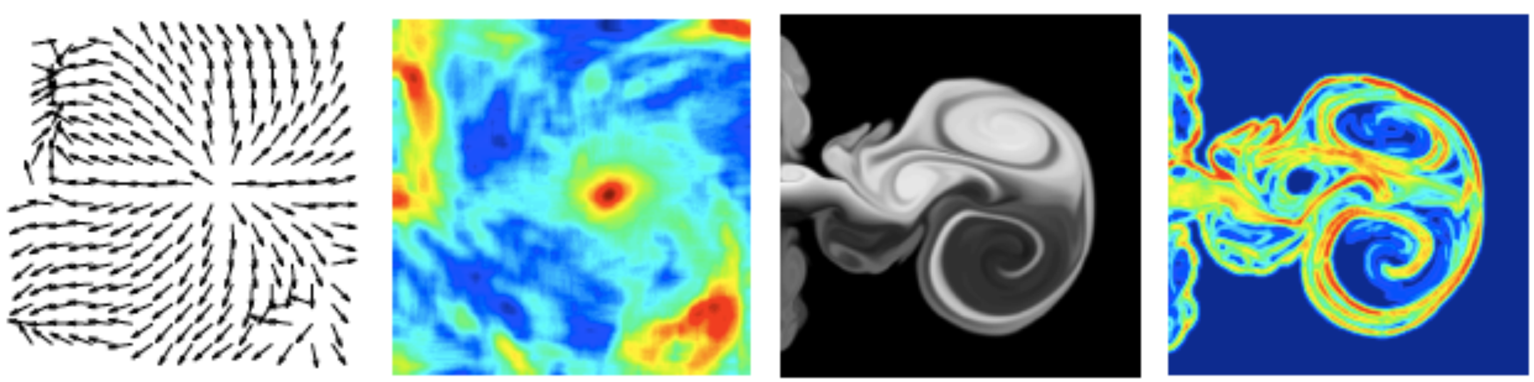
Maximal Entropy



More equally possible outcomes

Entropy for Scientific Data

- A data set can be considered as a random variable
- Each data point can be considered as an outcome of the random variable
- We can estimate the information content for the whole data set or for local regions



Other Entropy Measures

- Joint Entropy

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

- Relative Entropy

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$

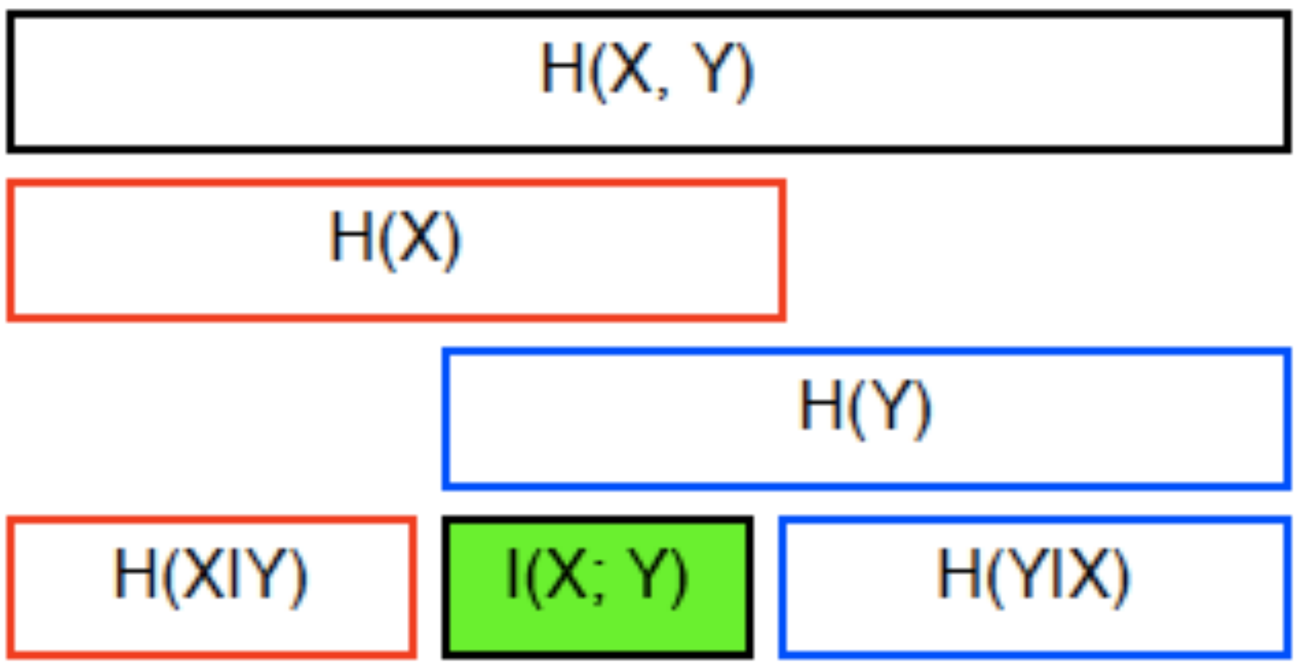
- Conditional Entropy

$$H(X|Y) = \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log p(x|y)$$

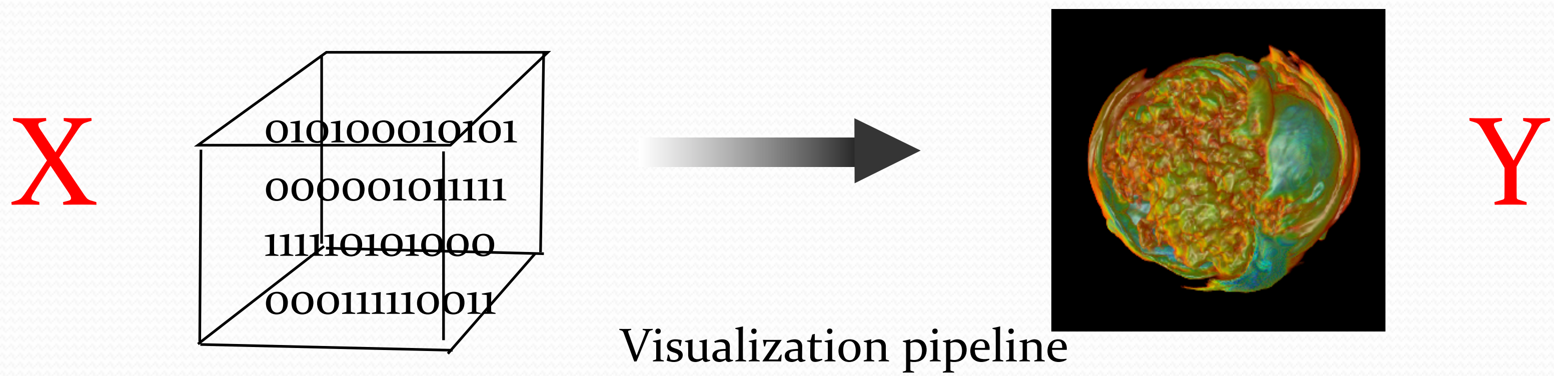
- Mutual Information

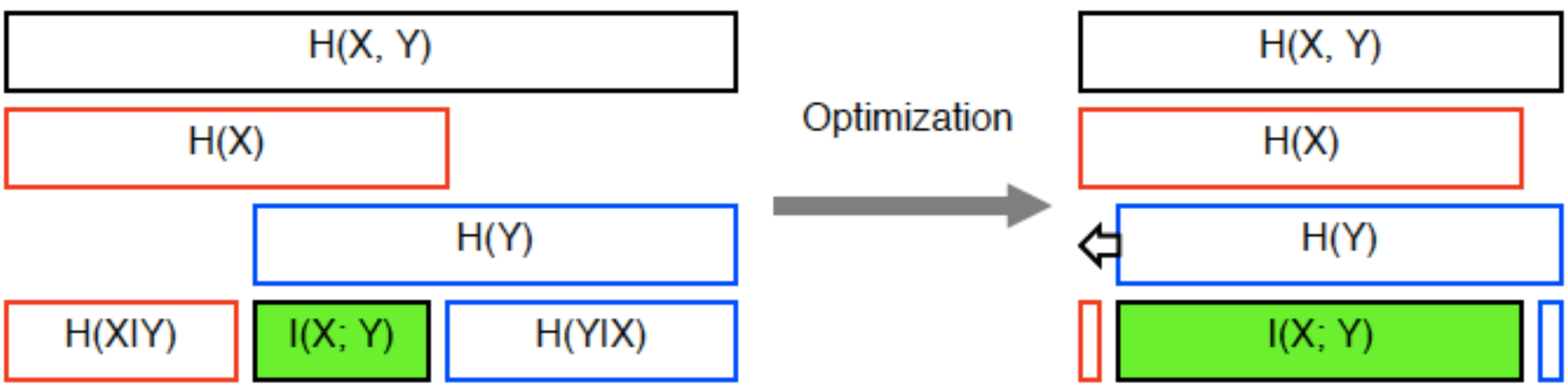
$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Relations of Entropy Measures



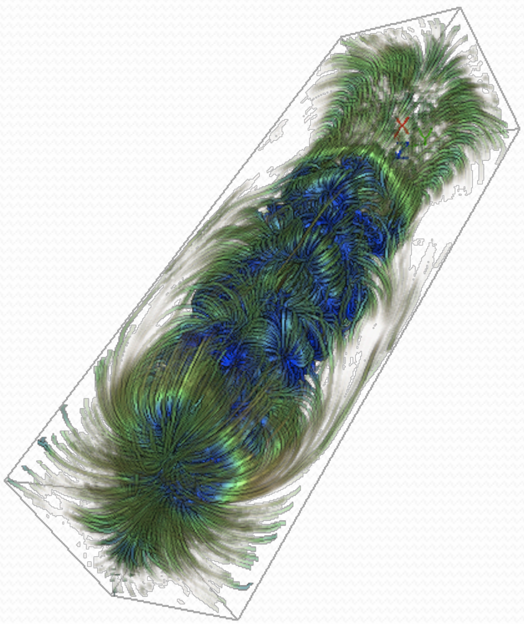
Evaluate Visualization

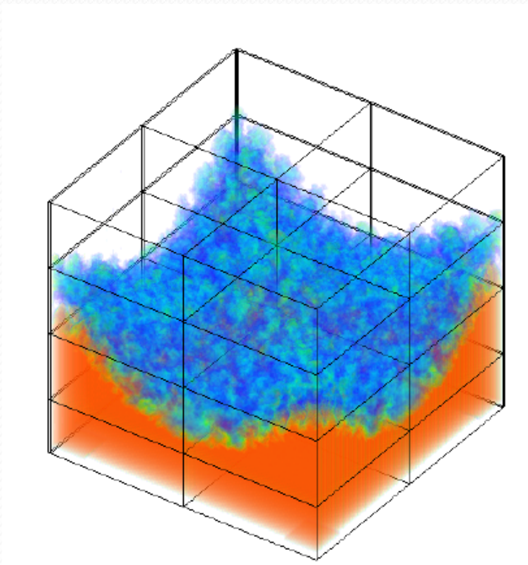


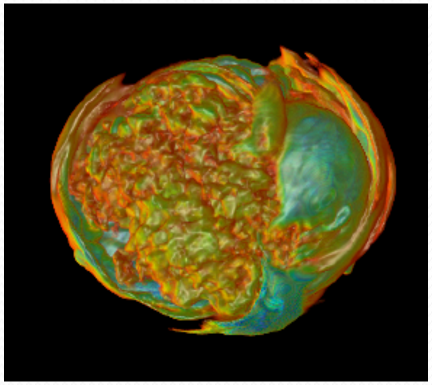


Applications in Visualization

- Streamline placement
- LOD selection
- Viewpoint selection for static and time-varying volume data



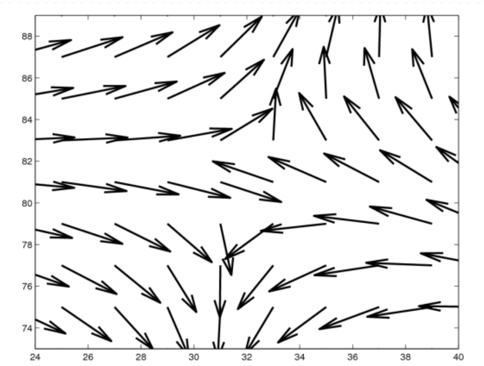




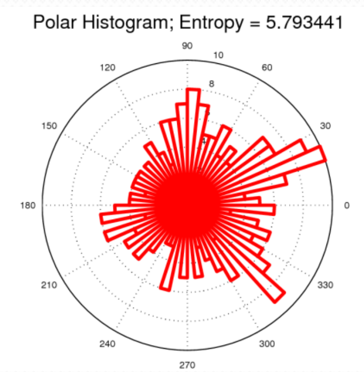
Information in Vector Fields

- Concept
  - Treat the vector field as a data source that generates vector orientation as outcome
  - The more diverse the vector orientations, the more information is contained in the vector field

- Measurement
  - Estimate the distribution of the vector orientation
  - Compute the entropy of this distribution as the measurement

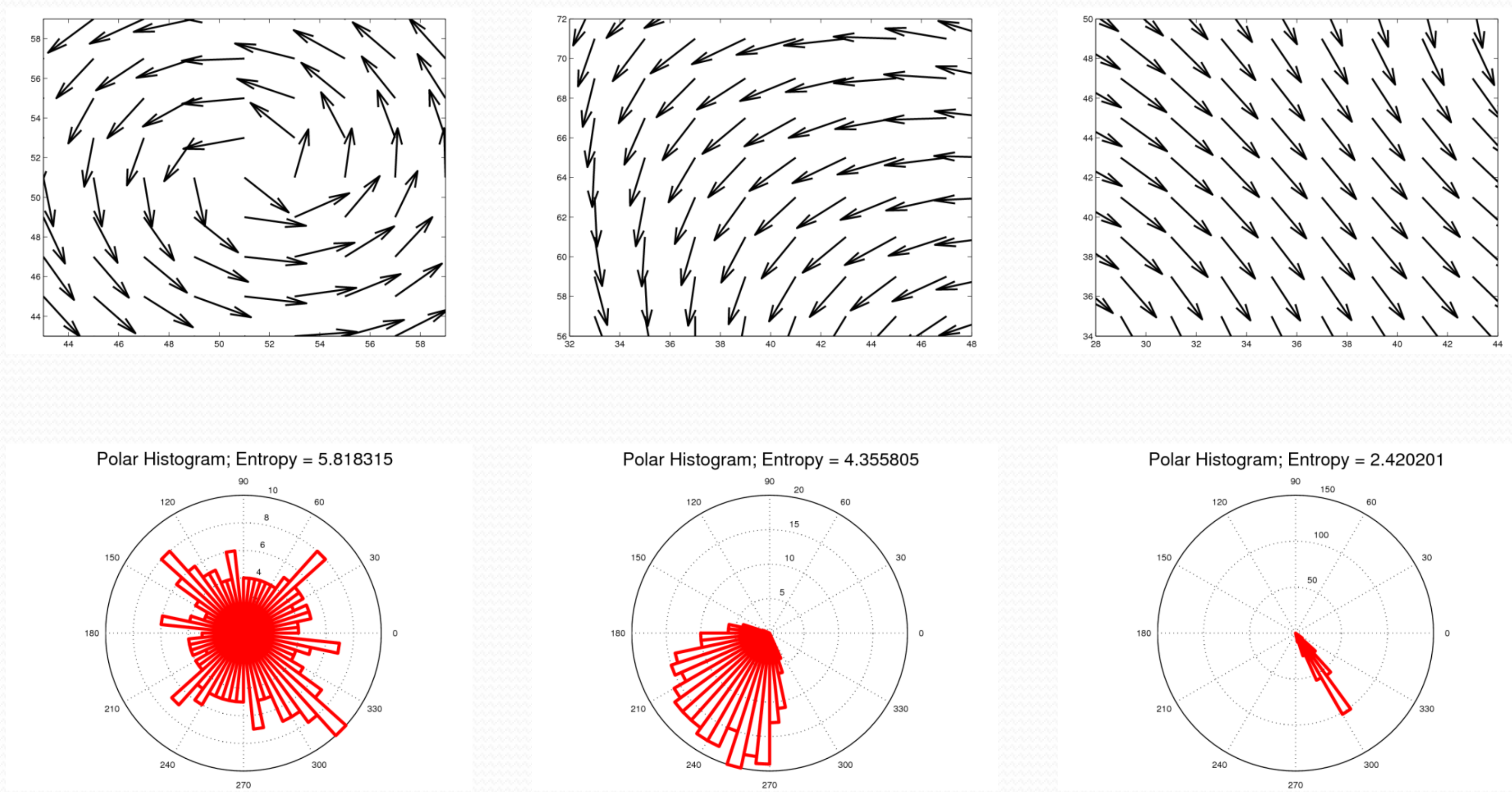


Vector field



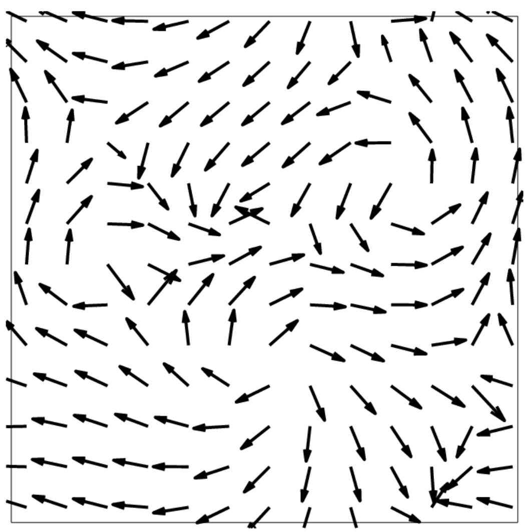
Polar Histogram

Information in Vector Fields

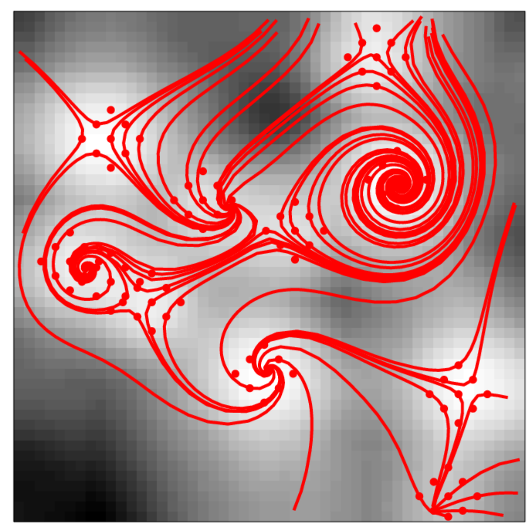


Entropy Field and Seeding

Measure the entropy around each point's neighborhood



Vector Field

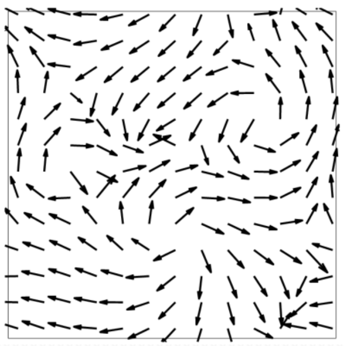


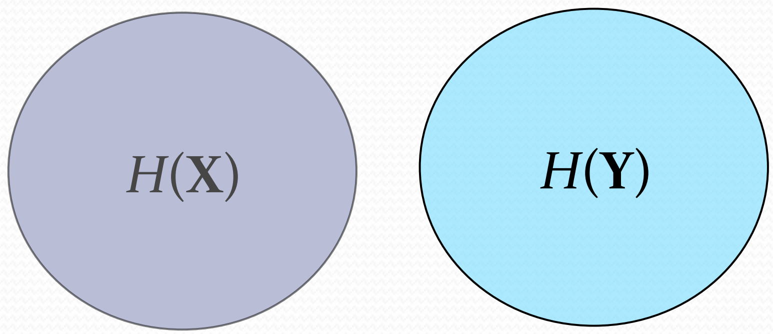
Entropy field: higher value means more information in the corresponding region

Entropy-based seeding: Places streamlines on the region with high entropy

The Information Comparison between Data/Visualization

Vector Field **X**

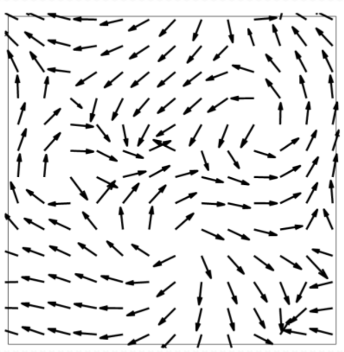


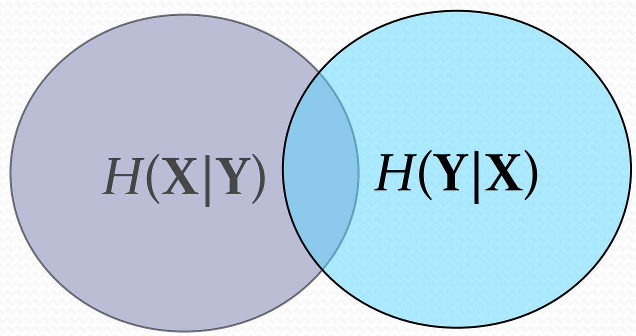


Streamlines **Y**

The Information Comparison between Data/Visualization

Vector Field **X**





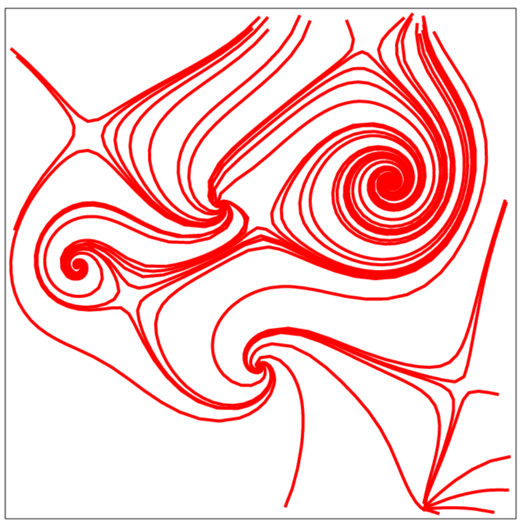
Streamlines **Y**

Conditional entropy $H(\mathbf{X}|\mathbf{Y})$:
The information in **X** not represented by **Y**

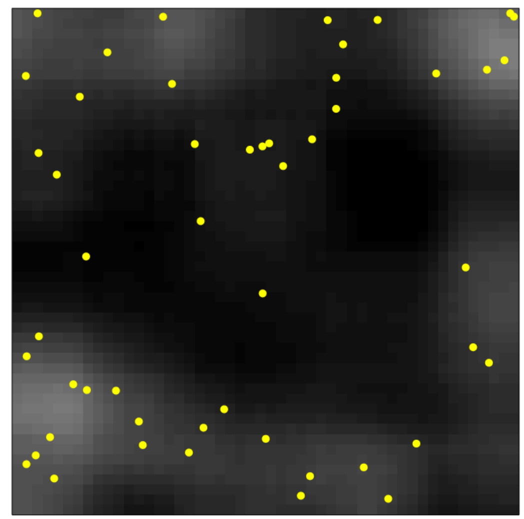
An effective visualization should represent most information in the data, i.e. $H(\mathbf{X}|\mathbf{Y})$ should be small

Conditional Entropy Field and Seeding

Measure the under-represented information in each region



Streamlines

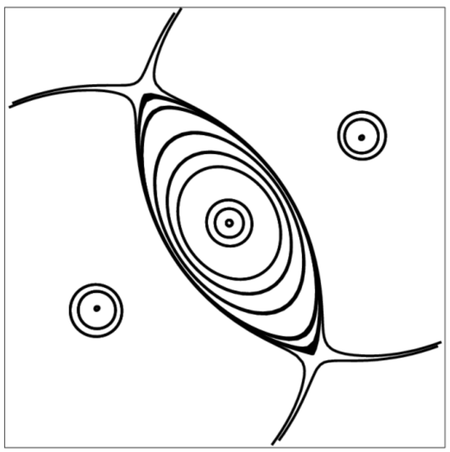


Conditional entropy field

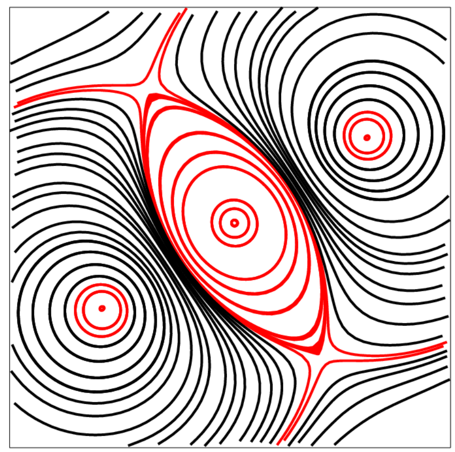
Conditional-entropy-based seeding: Place more seeds on regions with higher under-represented information

Result: 2D Vector Fields

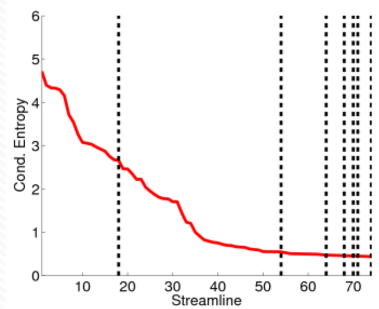
1st iteration: Entropy-based seeding



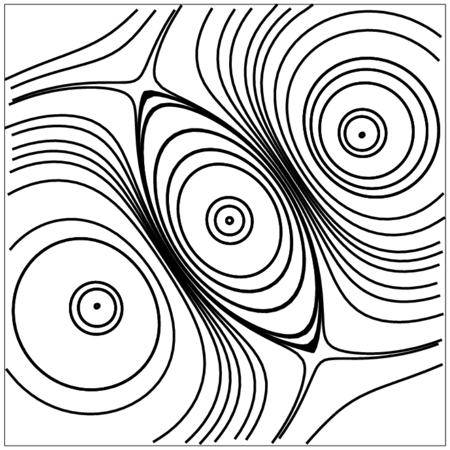
2nd iteration: Cond.-entropy-based seeding



Conditional entropy



When conditional entropy converges



ITL Software

- Information-Theoretic Library (ITL)
- Entropy analysis for exascale data sets
- Integrated into large-scale simulations to provide in situ data reduction and analysis
- Also used in post-processing for quality quantification and parameter tuning

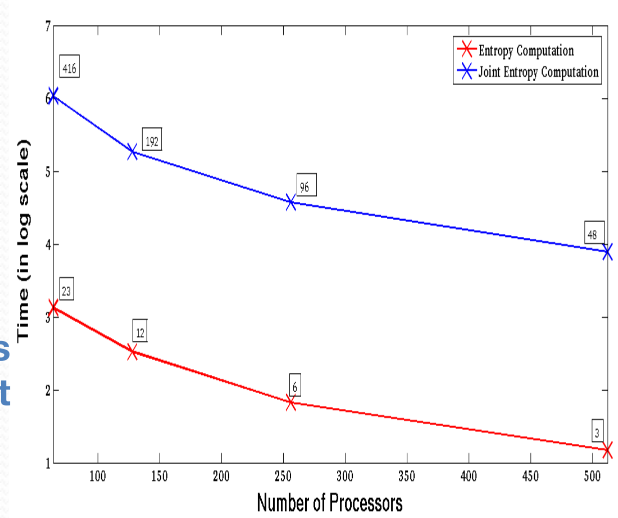
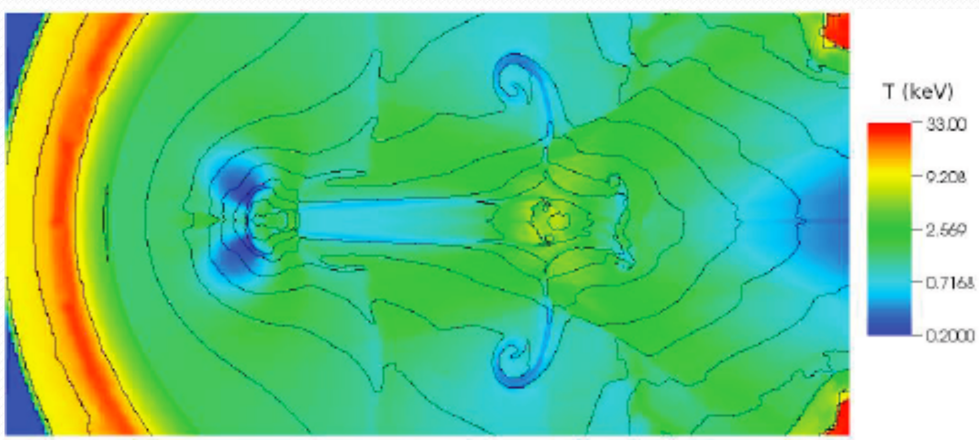
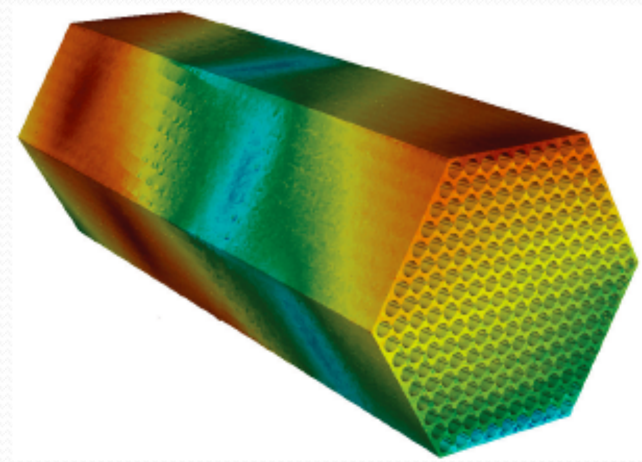


Figure : Performance of ITL run on NERSC's Franklin (Cray XT4). Our initial results showed that satisfactory scalability can be achieved.

Science Applications

- Nek5000: A Navier-Stokes solve for fluid flow, convective heat, and magnetohydrodynamics simulations
- Flash: Adaptive mesh code for astrophysics and cosmology





Collaborators

- Tom Peterka, Rob Ross at Argonne National Laboratory
- Yi-Jen Chiang at Polytechnic Institute of NYU
- Science collaborators: Paul Fischer, Aleksandr Obabko, Paul Ricker, Boyana Norris,

THE OHIO STATE UNIVERSITY

Argonne NATIONAL LABORATORY

NYU-poly POLYTECHNIC INSTITUTE OF NYU

U.S. DEPARTMENT OF ENERGY