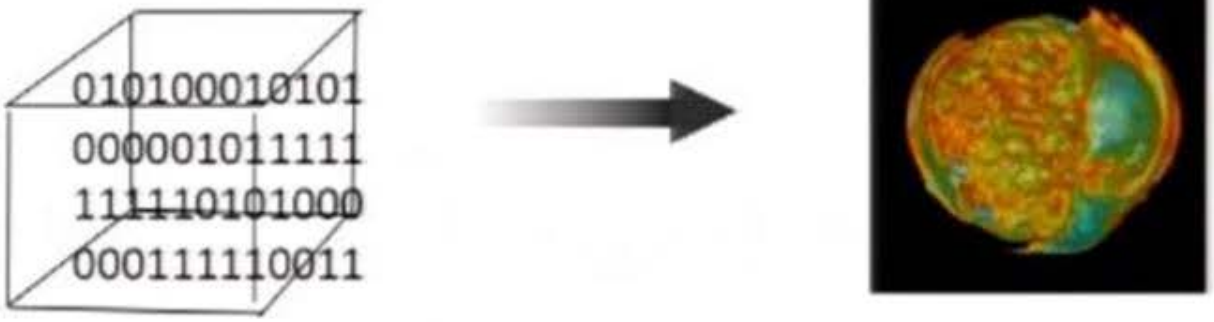
Large Scale Scientific Data Analysis and Visualization

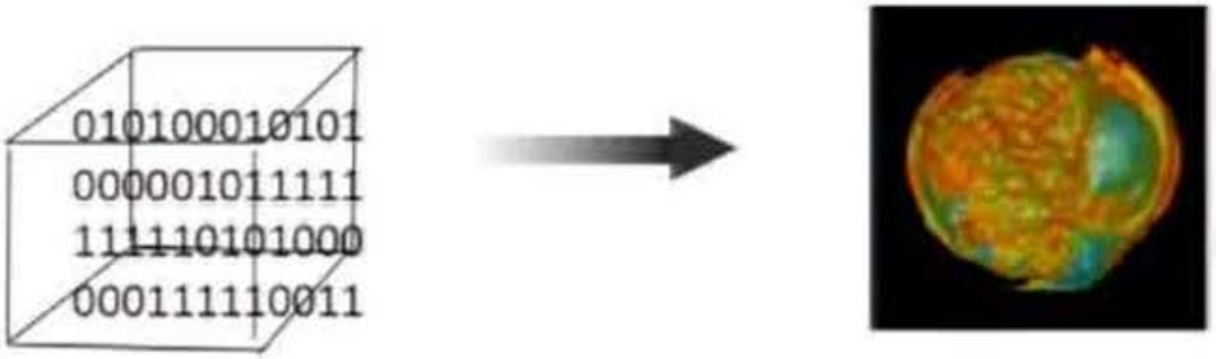
Han-Wei Shen

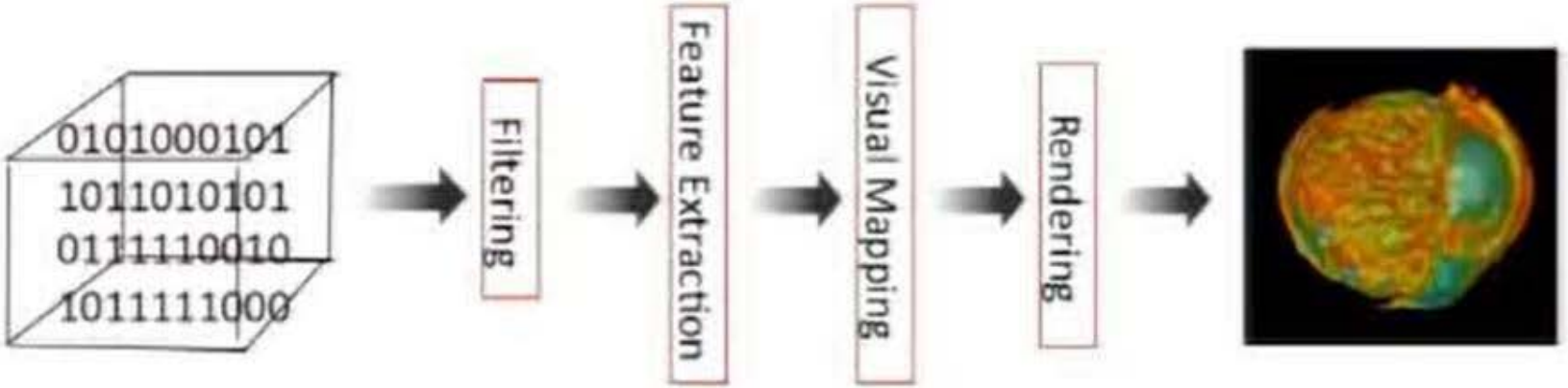
The Ohio State University

Driven Analysis and Visualization



Driven Analysis and Visualization





Visualization pipeline

Visual Analytic Questions

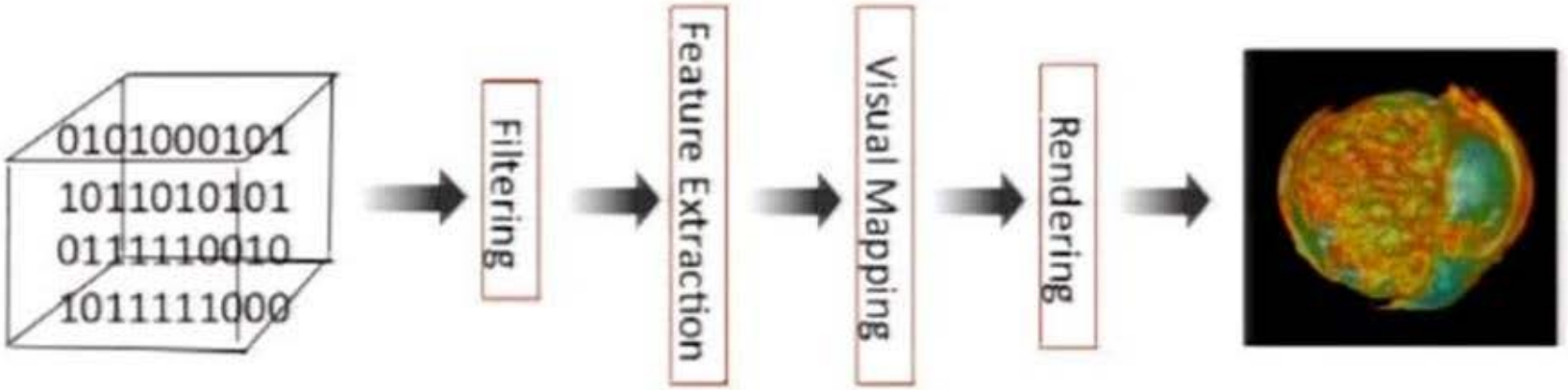
THE OHIO STATE UNIVERSITY

Visual Analytic Questions

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

Information Flow

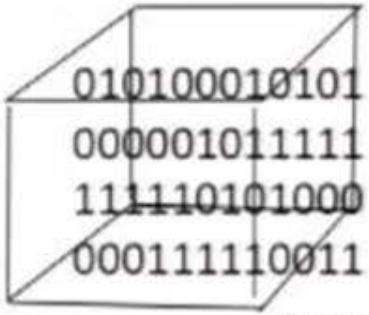
- Measure the flow of information across the entire data analysis and visualization pipeline
  - Quantify the information content in the data set
  - Measure the amount of information losses in each stage of the visualization pipeline
  - Choose parameters that can minimize the information losses

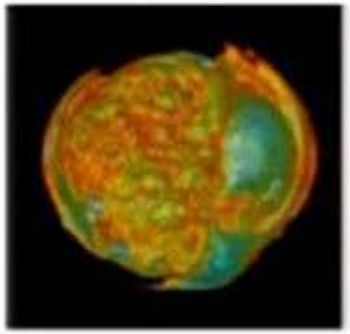


Visualization pipeline

Information-Driven Analysis and Visualization

X



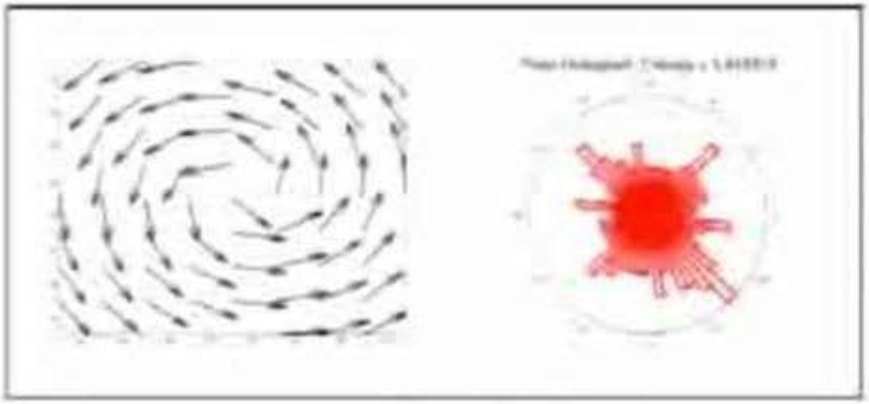


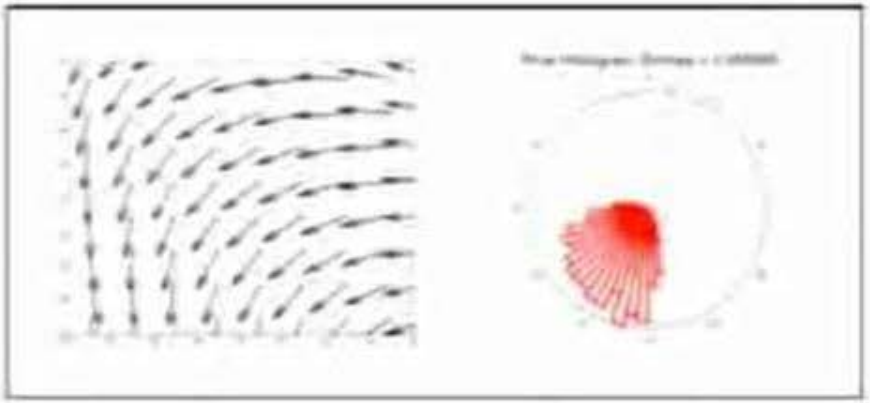
Y

Data Analysis and Visualization

Information Complexity

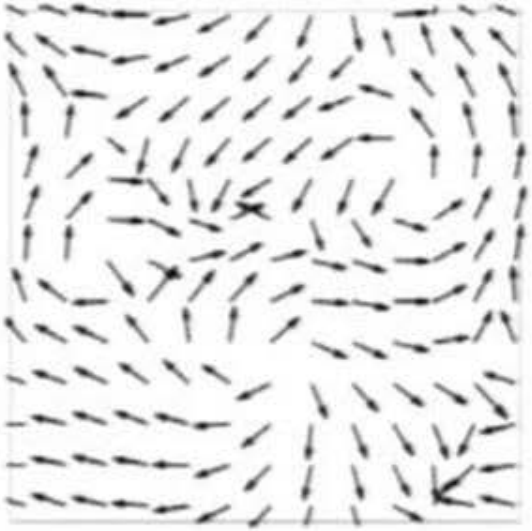
- Treat the vector data as a random variable
- The complexity of a data block can be represented by the distribution of the vectors
- Measure the amount of information contained in the local regions based on entropy measures





Entropy Field and Seeding

Measure the entropy around each point's neighborhood

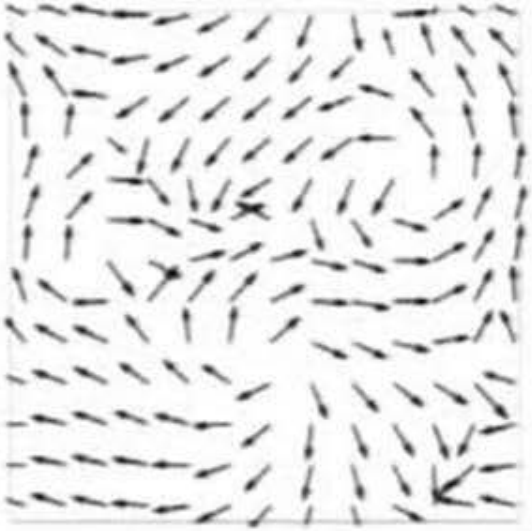


Vector Field

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

Entropy Field and Seeding

Measure the entropy around each point's neighborhood

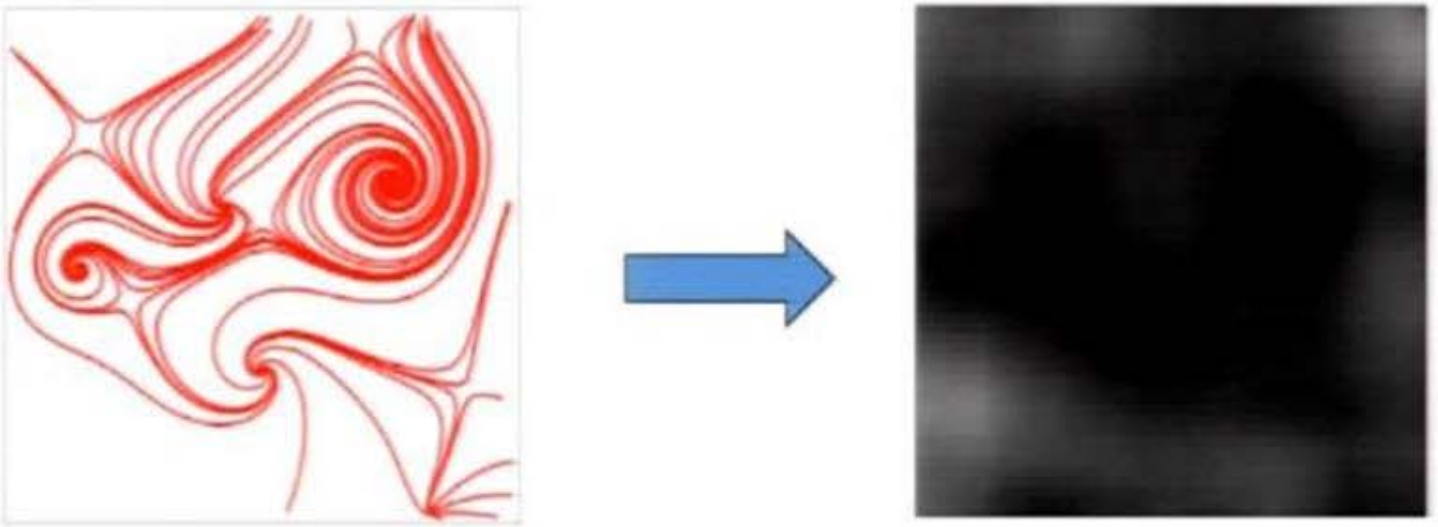


Vector Field

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

Conditional Entropy Field and Seeding

Measure the under-represented information in each region



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

Information Convergence

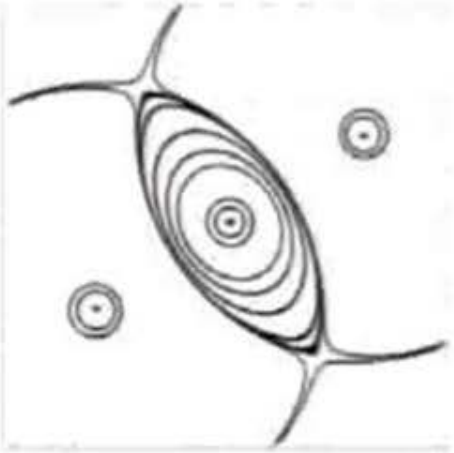
1st iteration: Entropy-based seeding

2nd iteration: Cond.-entropy-based seeding

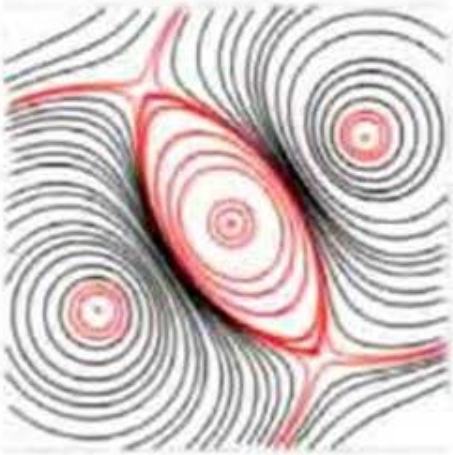
Conditional entropy

Information Convergence

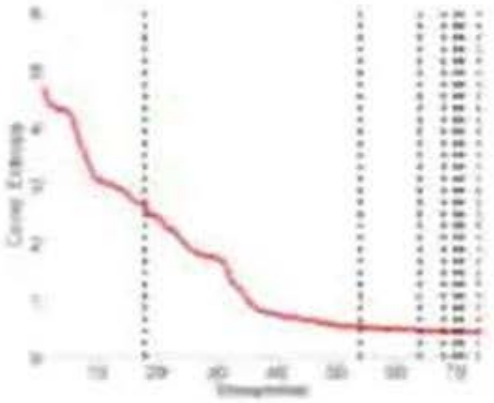
1st iteration: Entropy-based seeding



2nd iteration: Cond.-entropy-based seeding



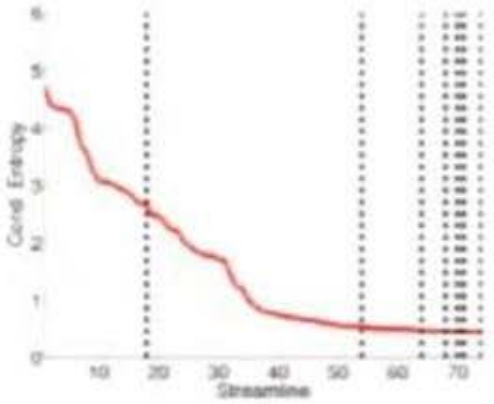
Conditional entropy



Information Convergence

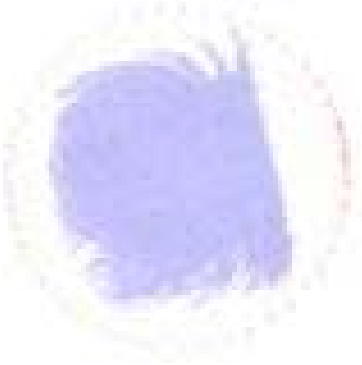
When conditional entropy converges

Conditional entropy



Application in View Selection

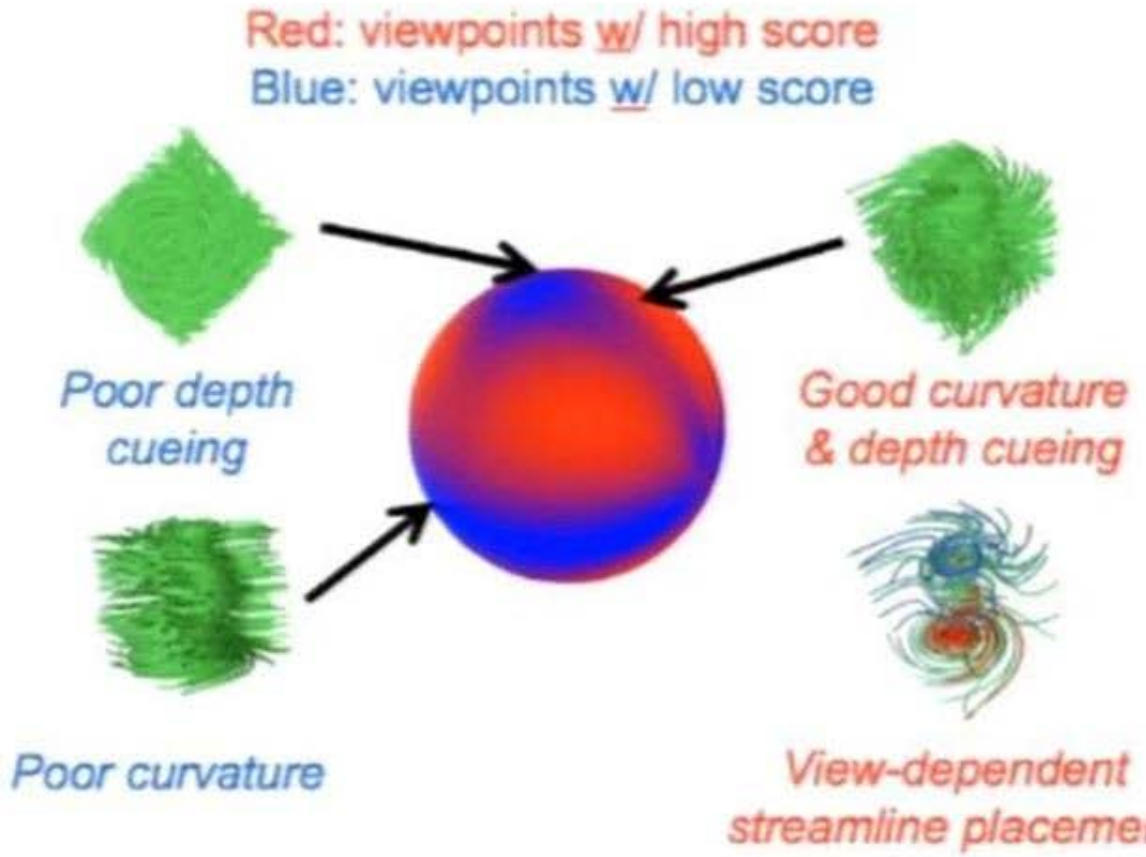
1. Parameterize the viewpoint space



Application in View Selection

1. Parameterize the viewpoint space

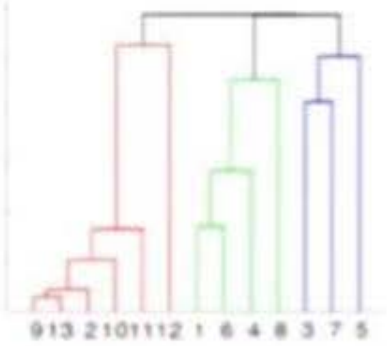
2. Sample view-dependent entropy



Multivariate Analysis

Step-by-Step guidance for multivariate exploration

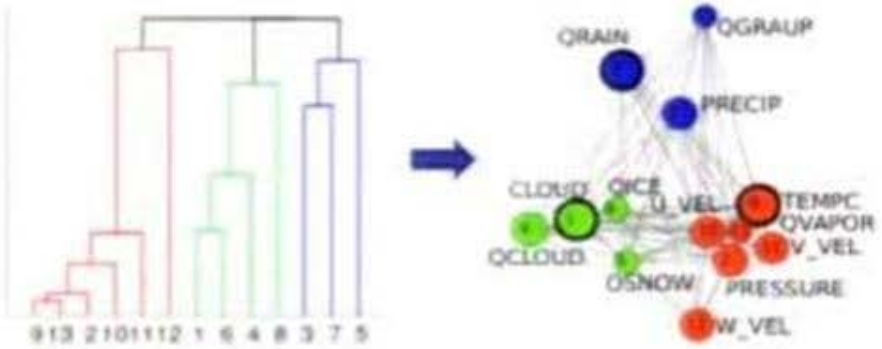
- Calculate all pair mutual information
- Generate a hierarchical cluster tree



Multivariate Analysis

Step-by-Step guidance for multivariate exploration

- Calculate all pair mutual information
- Generate a hierarchical cluster tree
- Group variables and calculate their relative imp.
- Compute the relationship between variables

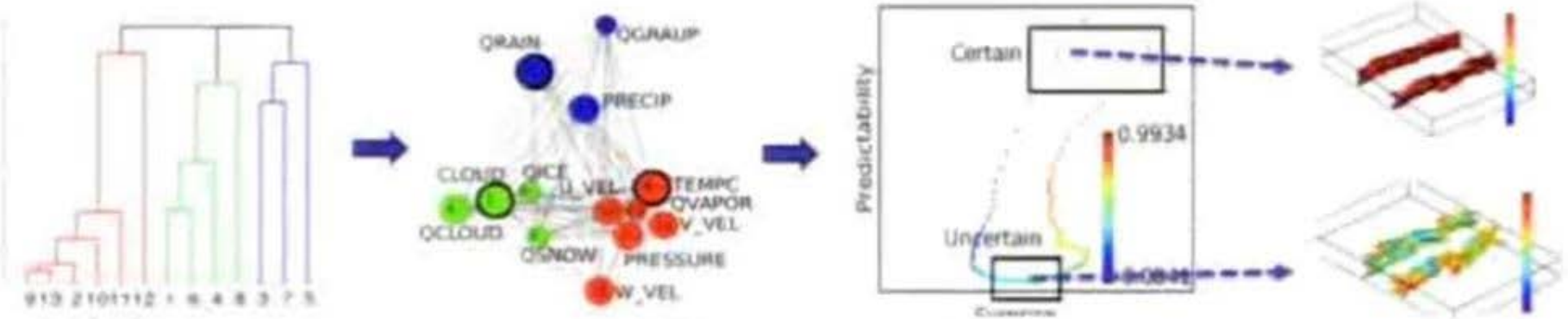


- Surprise: $$I_1(x;Y) = \sum_{y \in Y} \left(p(y|x) \log \frac{p(y|x)}{p(y)} \right)$$
- Predictability: $$I_2(x;Y) = H(Y) - H(Y|x) = -\sum_{y \in Y} (p(y) \log p(y)) + \sum_{y \in Y} (p(y|x) \log p(y|x))$$

Multivariate Analysis

Step-by-Step guidance for multivariate exploration

- Calculate all pair mutual information
- Generate a hierarchical cluster tree
- Group variables and calculate their relative imp.
- Compute the relationship between variables



- **Surprise:** $$I_1(x;Y) = \sum_{y \in Y} \left(p(y|x) \log \frac{p(y|x)}{p(y)} \right)$$
- **Predictability:** $$I_2(x;Y) = H(Y) - H(Y|x) = -\sum_{y \in Y} (p(y) \log p(y)) + \sum_{y \in Y} (p(y|x) \log p(y|x))$$

Distribution-based Visual Analytics

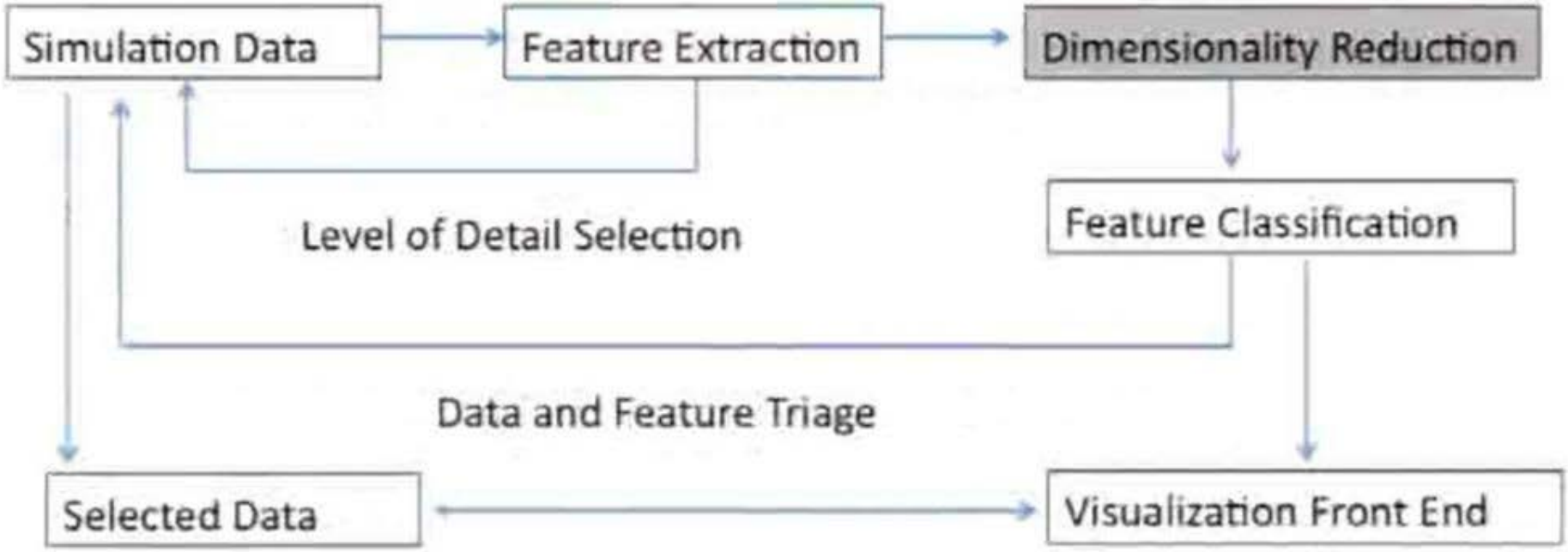
General Steps:

Distribution-based Visual Analytics

General Steps:

- Define the random variable (one or more)
- Define the states of the random variable
- Calculate the probability of each state
- Calculate the entropy measures for each random variable
  - For multivariate problem, calculate the joint entropy between the variables and study their relationship
- Calculate the information content of each variable and the shared information among the variables
- Maximize the information content displayed in the final visualization

Scientific Data Analytics Pipeline



Conclusions

- Use distributions as a compact representation of data
  - Many statistics about the data can be derived
  - Information flow across the visualization pipeline can be analyzed
  - Regions of high information content can be identified
  - Parameters for various visualization algorithms can be optimized
  - It allows detailed analysis and inferences even in the absence of the raw data

Conclusions

- Use distributions as a compact representation of data
  - Many statistics about the data can be derived
  - Information flow across the visualization pipeline can be analyzed
  - Regions of high information content can be identified
  - Parameters for various visualization algorithms can be optimized
  - It allows detailed analysis and inferences even in the absence of the raw data
- Supports the needs of in situ data analysis
  - Data reduction
  - Data summarization
  - Data triage
  - Feature extraction and indexing

Thank You!