Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Study Materials for Lecture 20

- An Information-Aware Framework for Exploring Multivariate Data Sets, Biswas et al., TVCG
- Multimodal Data Fusion Based on Mutual Information, Bramon et al., TVCG
- In Situ Adaptive Spatio-Temporal Data Summarization, Dutta et al., IEEE BigData

An Introduction to Information Theory Measures

Entropy & Joint Entropy

- Entropy: In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

$$H(X) = -\sum_x p(x) * \log p(x), X \text{ is a discrete random variable}$$

Entropy & Joint Entropy

- Entropy: In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.

$$H(X) = -\sum_x p(x) * \log p(x), X \text{ is a discrete random variable}$$

- Joint Entropy: For a pair of discrete random variables X and Y

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

Conditional Entropy

- Conditional Entropy: Entropy of Y when variable X is observed

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

Conditional Entropy

- Conditional Entropy: Entropy of Y when variable X is observed

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

- Relationship between joint probability, conditional probability and marginal probability:

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y)$$

Properties of Entropy

- Some useful properties:

1. $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
2. $H(X,Y) \leq H(X) + H(Y)$
3. $H(X) \geq H(X|Y) \geq 0$
4. If X and Y are independent, $H(Y|X) = H(Y)$,
   - So, $H(X,Y) = H(X) + H(Y)$

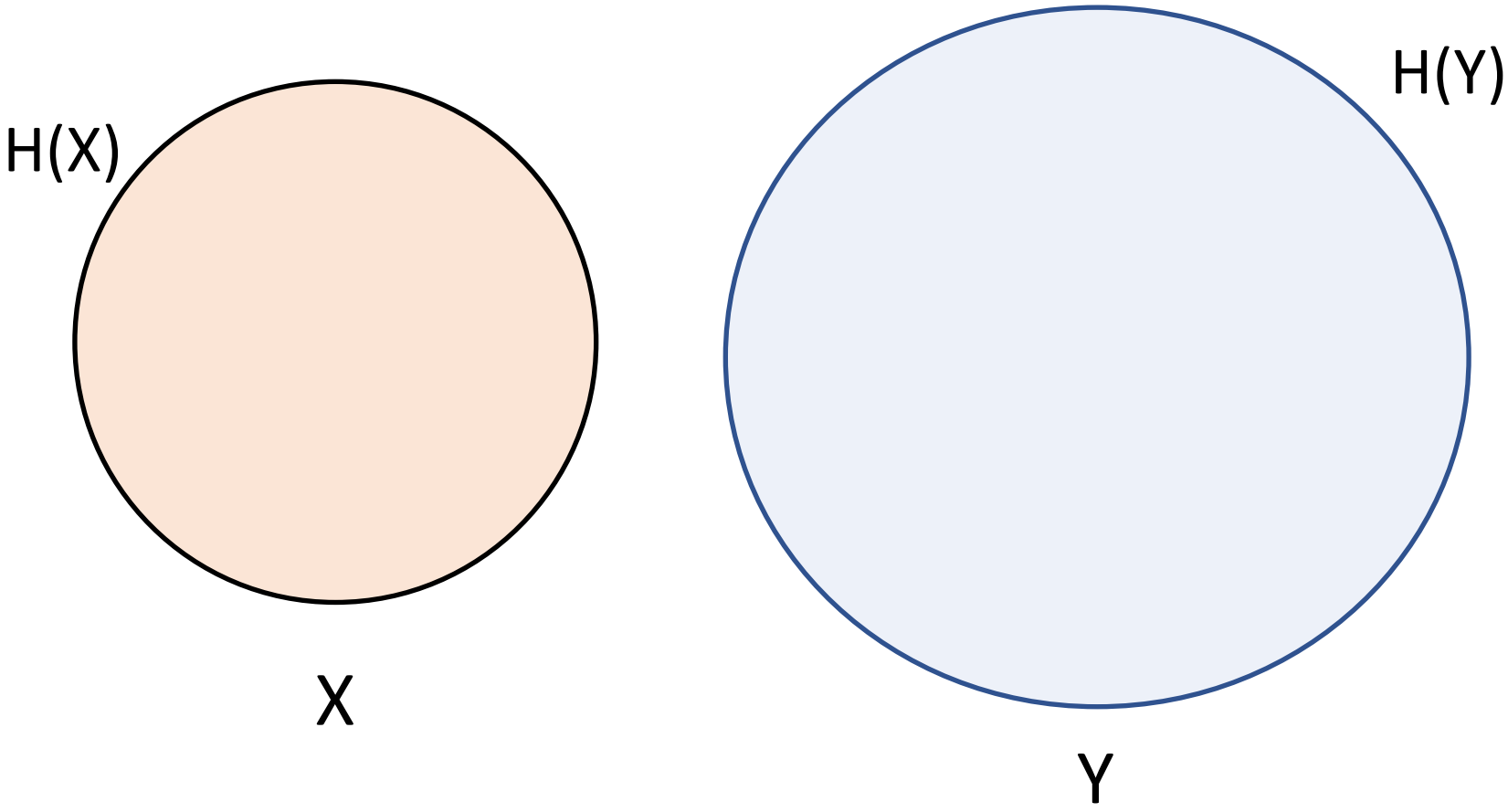
Mutual Information

- Mutual Information: It expresses how much the knowledge of variable Y decreases the uncertainty of X
- It is a measure of mutual dependence between two random variables
- Mutual information is also interpreted as a measure of nonlinear dependence

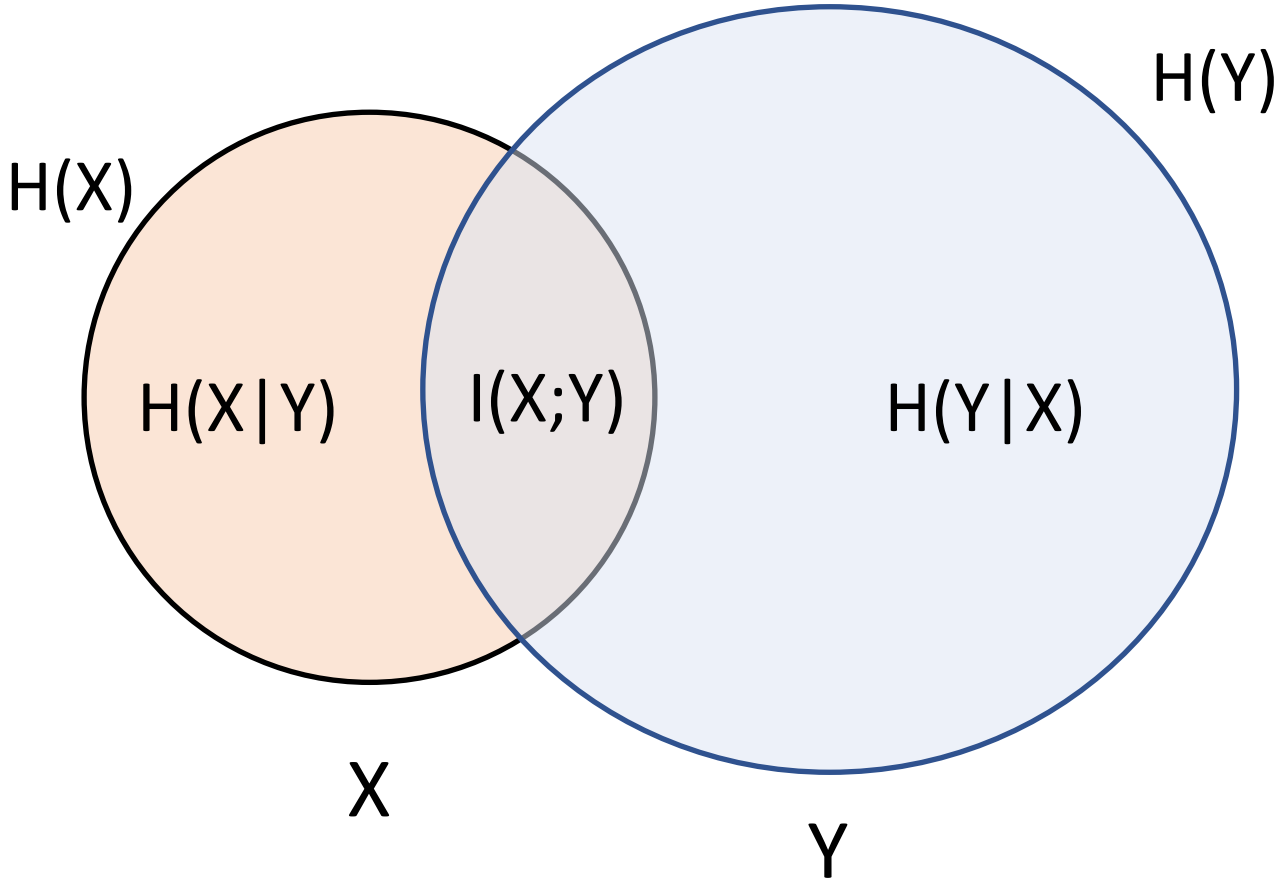
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$
$$= H(X) - H(X|Y) = H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$

$I(X;Y)$ >= 0 and $I(X;Y)$= $I(Y;X)$

Relationship Between Various Measures



Relationship Between Various Measures



Specific Mutual Information Measures

- Mutual information can be decomposed in various ways
- Specific mutual information conveys information for a specific observation when the entire second variable is observed
- ‘Surprise’, ‘Predictability’, and ‘Entanglement’ are three such specific information measures derived from Mutual Information

Specific Mutual Information Measures

- Surprise:

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

$$= \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log \frac{p(y|x)}{p(y)}$$

$$Surprise = I_1(x;Y) = \sum_{y \in Y} p(y|x) log \frac{p(y|x)}{p(y)}$$

- 'Surprise' expresses the surprise about variable Y when x is observed. It means that Surprise quantifies how much you know about Y when x is observed
- It only takes positive values

Specific Mutual Information Measures

- Predictability:

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

$$Predictability = I_2(x;Y) = H(Y) - H(Y|x)$$
$$= -\sum_{y \in Y} p(y) log p(y) + \sum_{y \in Y} p(y|x) log p(y|x)$$

- 'Predictability' expresses the change in uncertainty about Y when x is observed.
- It can take both positive and negative values
  - Positive values mean uncertainty is reduced; negative value means uncertainty is increased

Specific Mutual Information Measures

- Entanglement:

$$I_3(x;Y) = \sum_{y \in Y} p(y|x) I_2(y;X)$$

- The most informative x values are those that are related to the most informative values of y
- $I_3(x;Y)$ can be both positive and negative

Pointwise Mutual Information

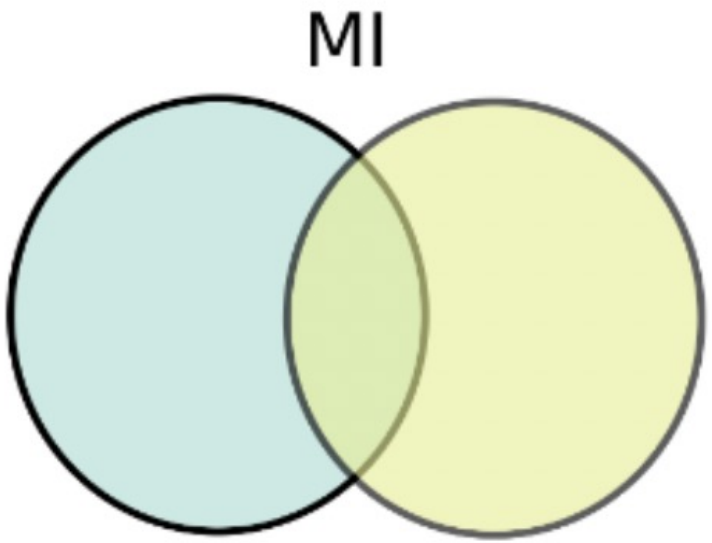
- Given two random variables X and Y, if x is an observation of X and y for Y, then the PMI value for the value pair (x, y) is expressed as

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

$p(x)$ is the probability of a particular occurrence x of X
$p(y)$ is the probability of y of variable Y
$p(x,y)$ is their joint probability

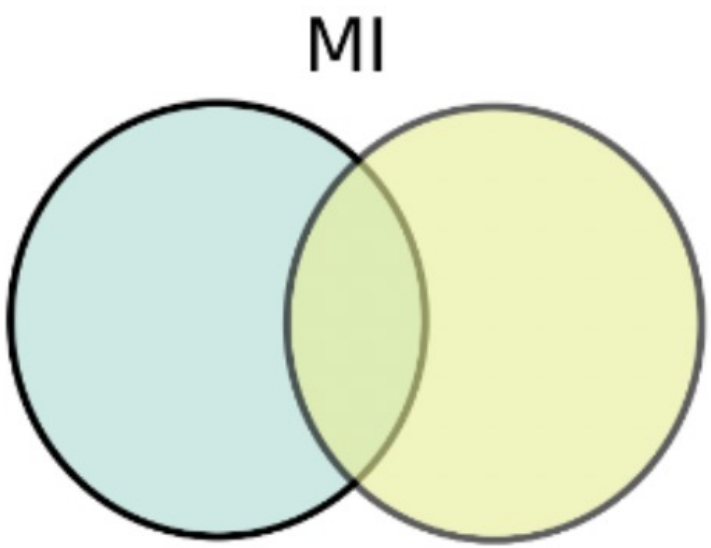
- When $p(x,y) > p(x)p(y)$, PMI$(x,y) > 0$,
- When $p(x,y) < p(x)p(y)$, PMI$(x,y) < 0$,
- When $p(x,y) \approx p(x)p(y)$, PMI$(x,y) \approx 0$

MI vs. SMI vs. PMI

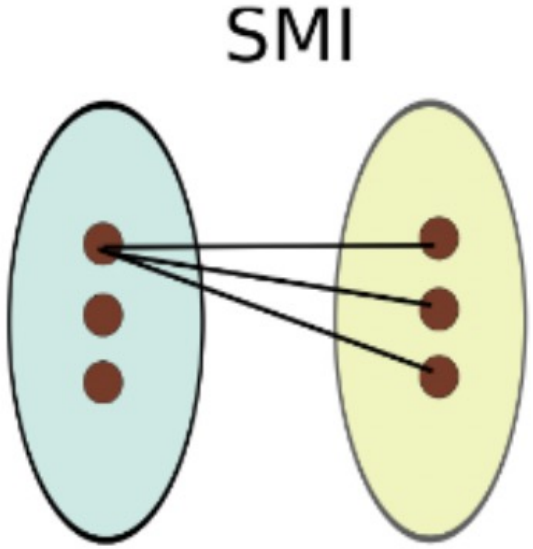


Total shared information
A single number

MI vs. SMI vs. PMI



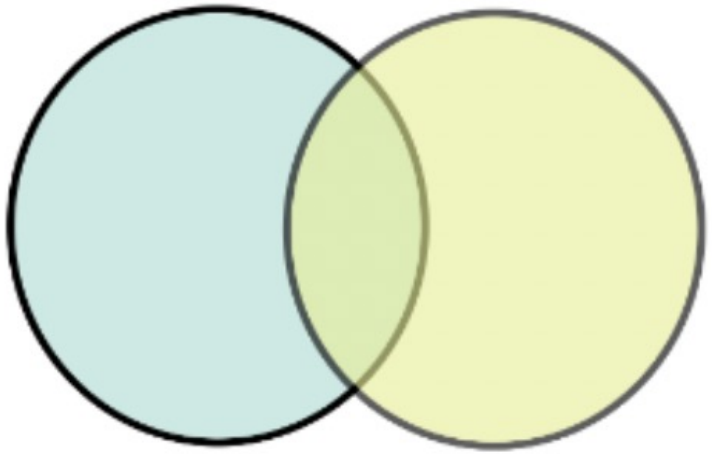
Total shared information
A single number



One to all mapping
One value for each observation

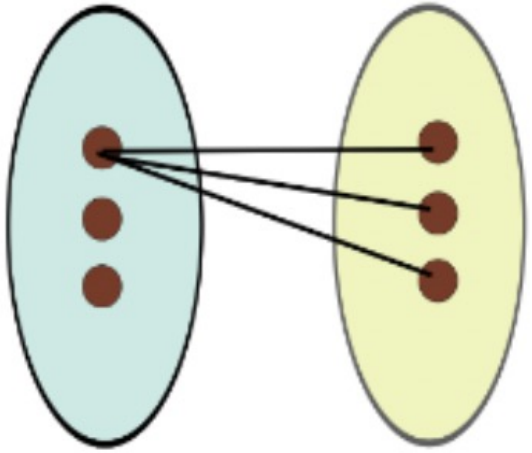
MI vs. SMI vs. PMI

MI



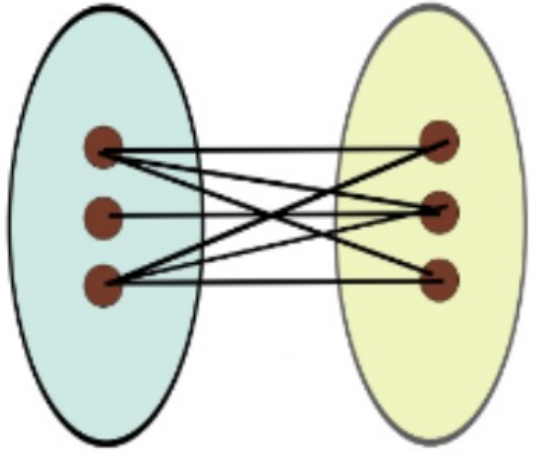
Total shared information
A single number

SMI



One to all mapping
One value for each observation

PMI



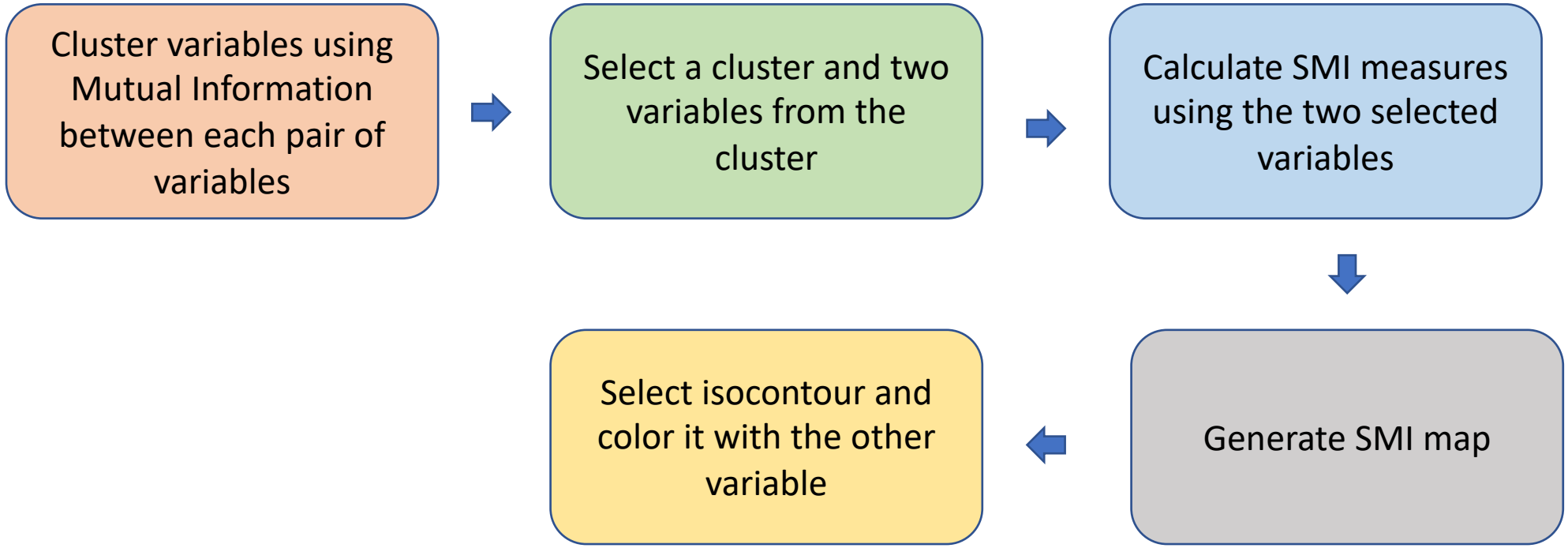
One to one mapping
Defined for each value pair

Applications

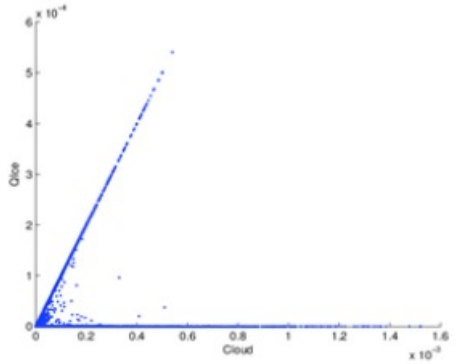
- Multivariate analysis framework
- Spatial data fusion
- Temporal data fusion

Multivariate Data Analysis Framework

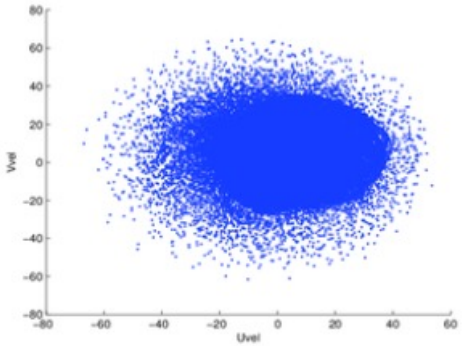
Multivariate Data Analysis Framework



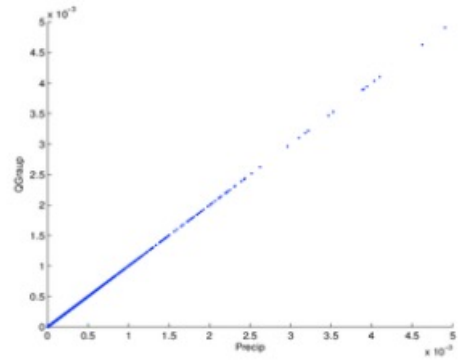
Different Degrees of Correlation among Variables



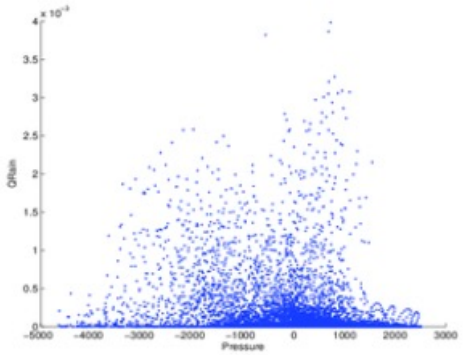
(a) Cloud vs QIce.



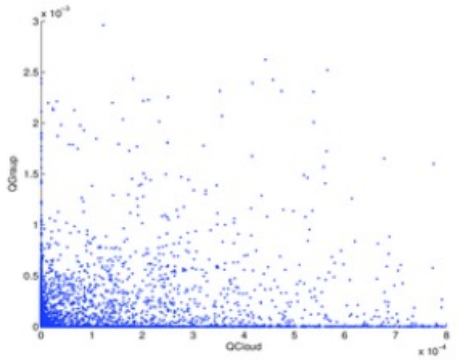
(b) V-vel vs U-vel.



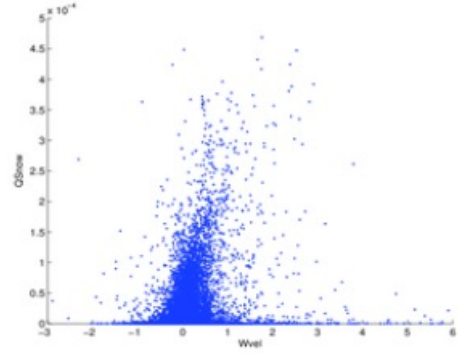
(c) Precip vs QGraup.



(d) Pressure vs QRain.

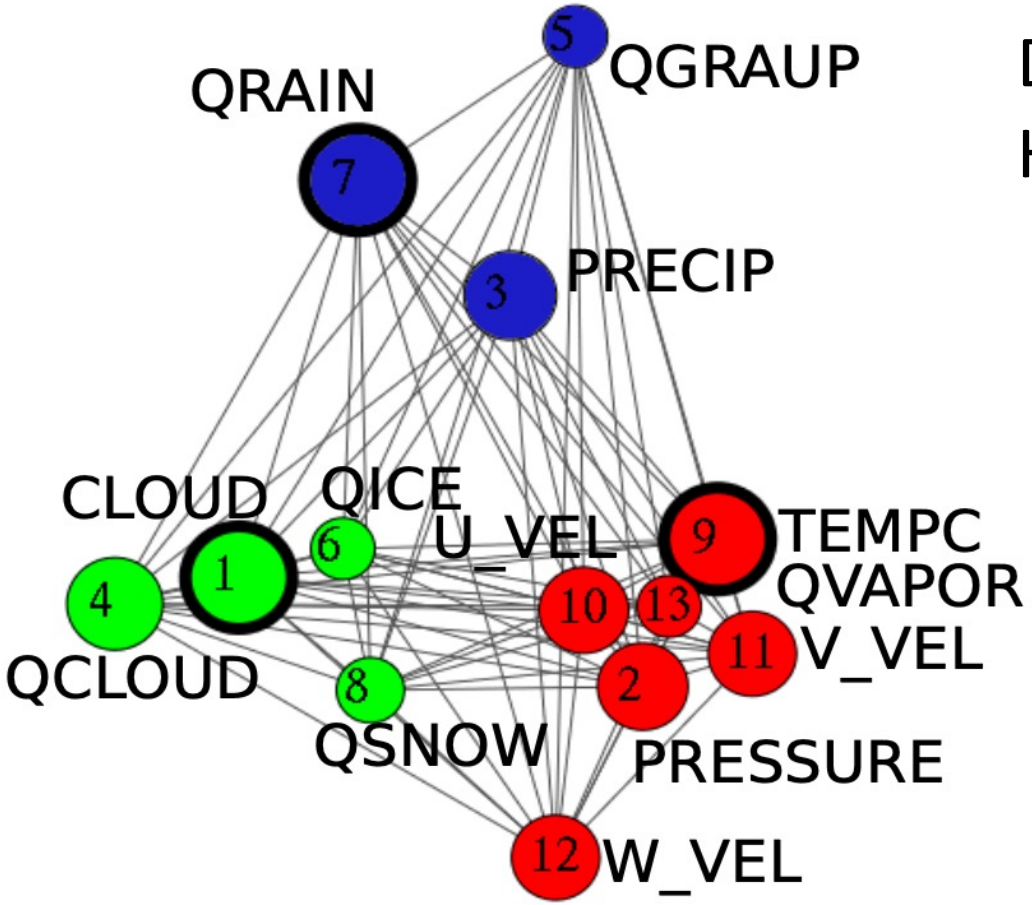


(e) QCloud vs QGraup.



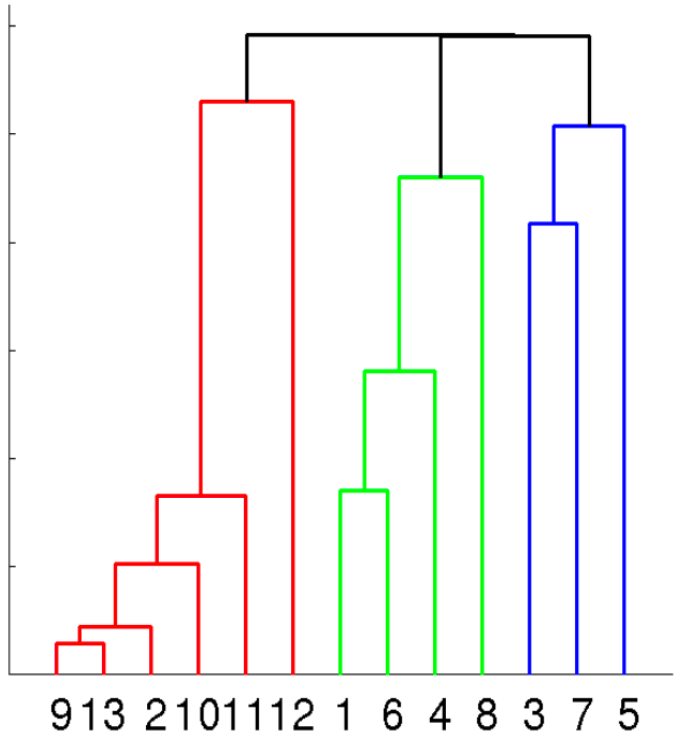
(f) W-velocity vs QSnow.

Clustering of Variables

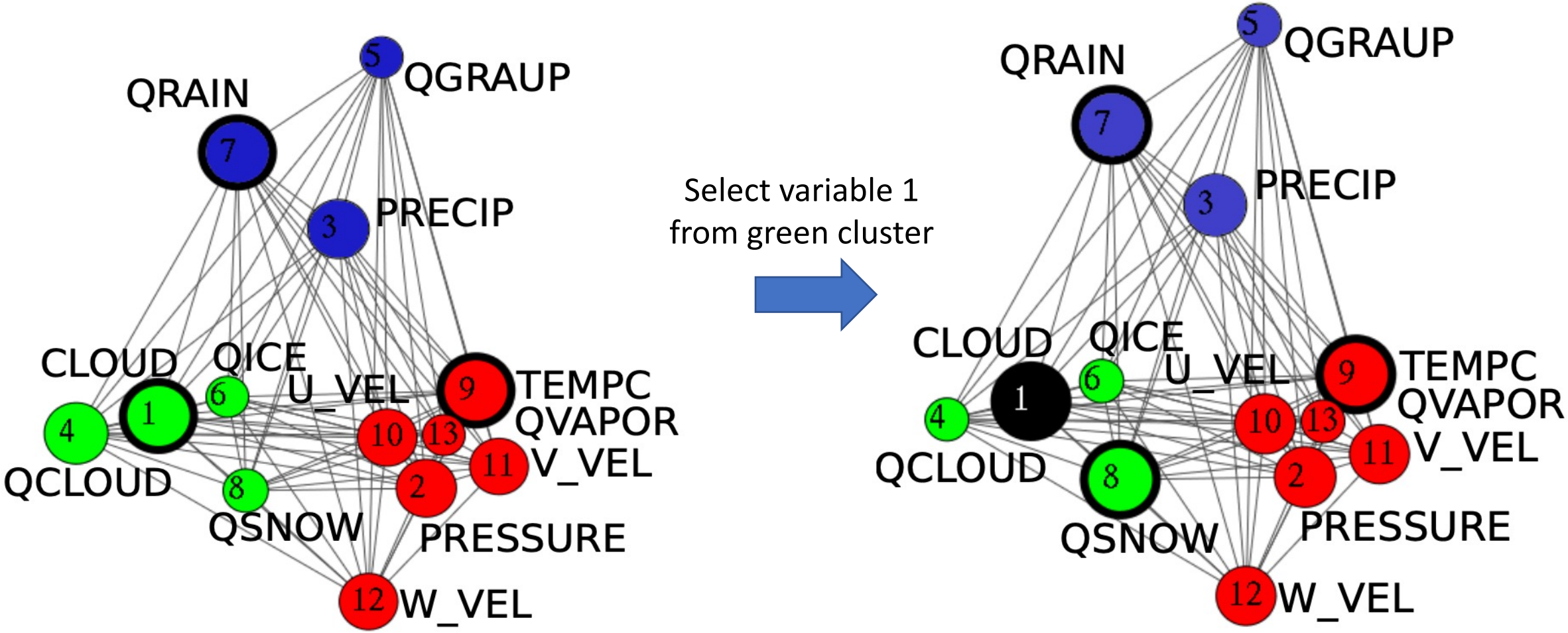


Distance between variables = Inverse of MI

Hierarchical clustering to select cluster number



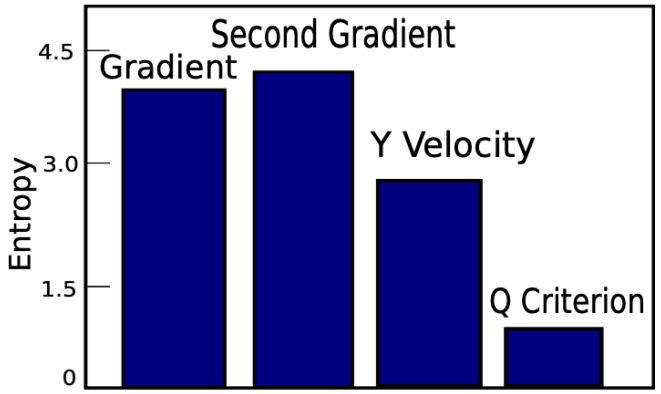
Clustering of Variables



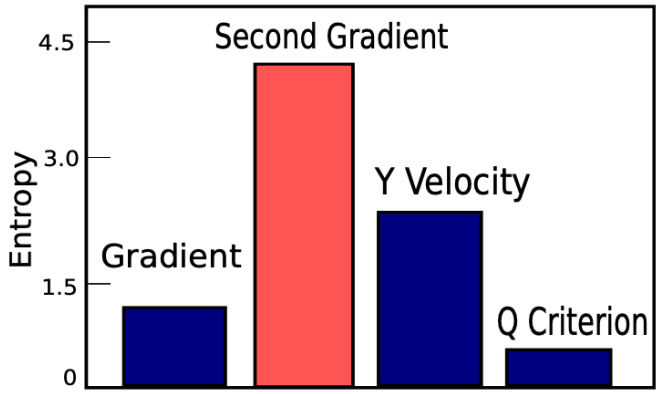
Variable Selection from a Cluster

- Use conditional entropy for variable selection and estimation of remaining information in the system

$$H(X_1,\ldots,X_n|X_{k1},\ldots,X_{km}) = H(X_1,\ldots,X_n) - H(X_{k1},\ldots,X_{km})$$



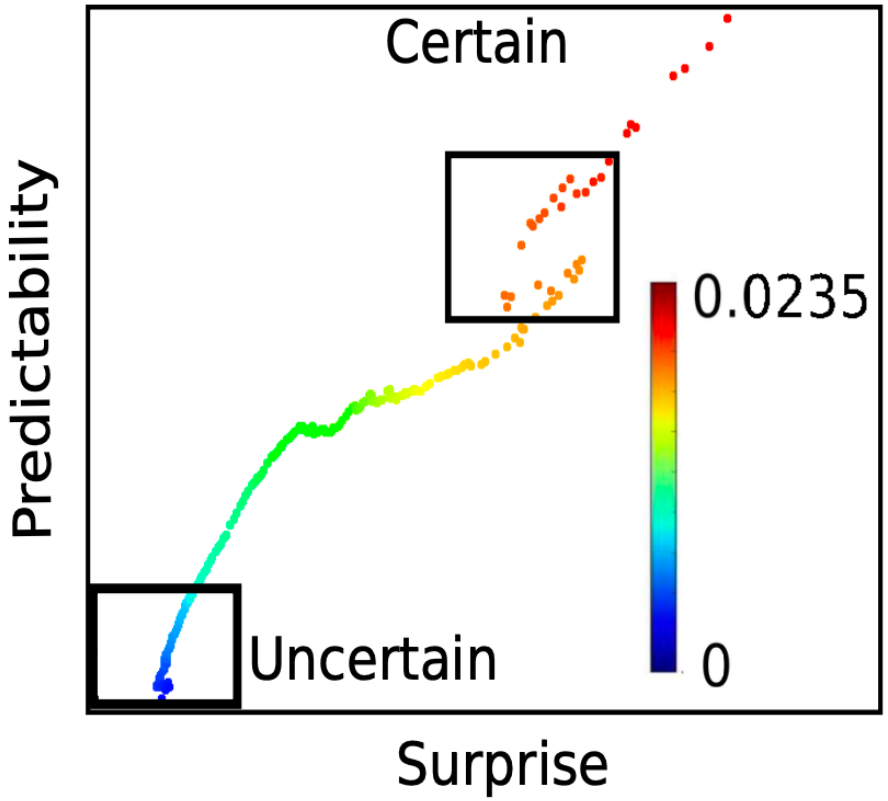
(a) Uncertainty remaining in the variables before selection.



(b) Uncertainty remaining in the variables after selection of Second Gradient.

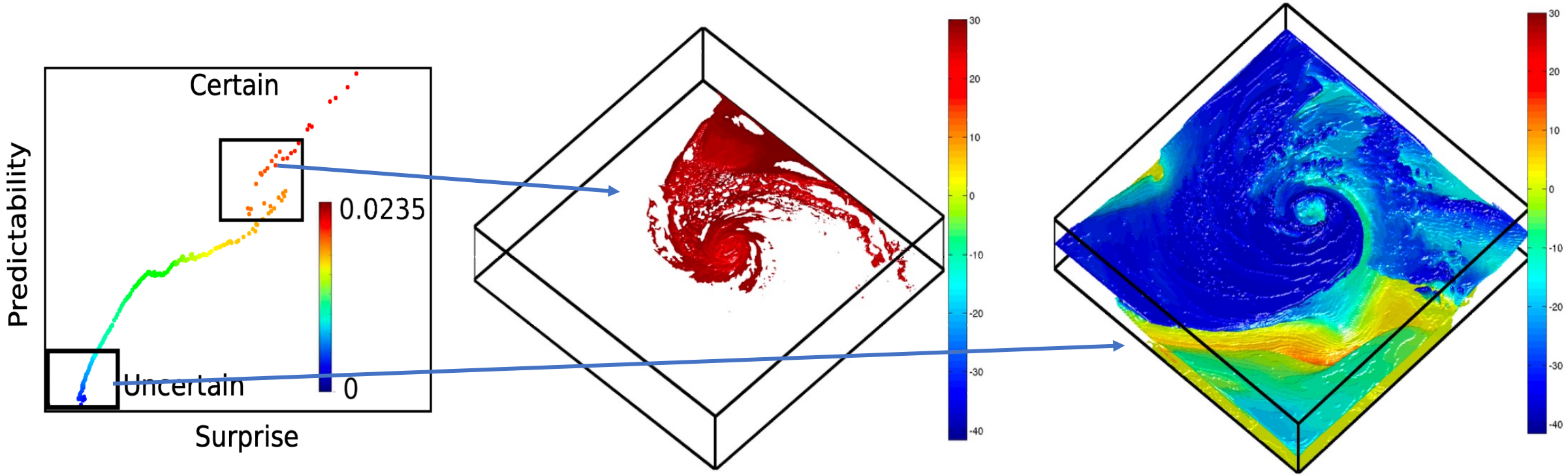
Compute Surprise and Predictability

- Compute I1 and I2 between the two selected variables
- Plot the I1-I2 map where the points are colored by the selected variable
- Now we can select points that has low surprise and predictability or high surprise and predictability
- Visualization is done using Isosurface
- Color of Isosurface indicates amount of information gained



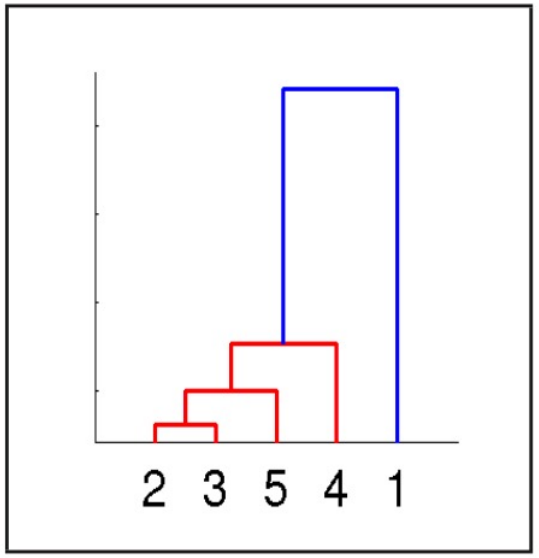
Explore Isosurface Uncertainty for Isabel Data

Two selected variables are: QVAPOR and Temperature



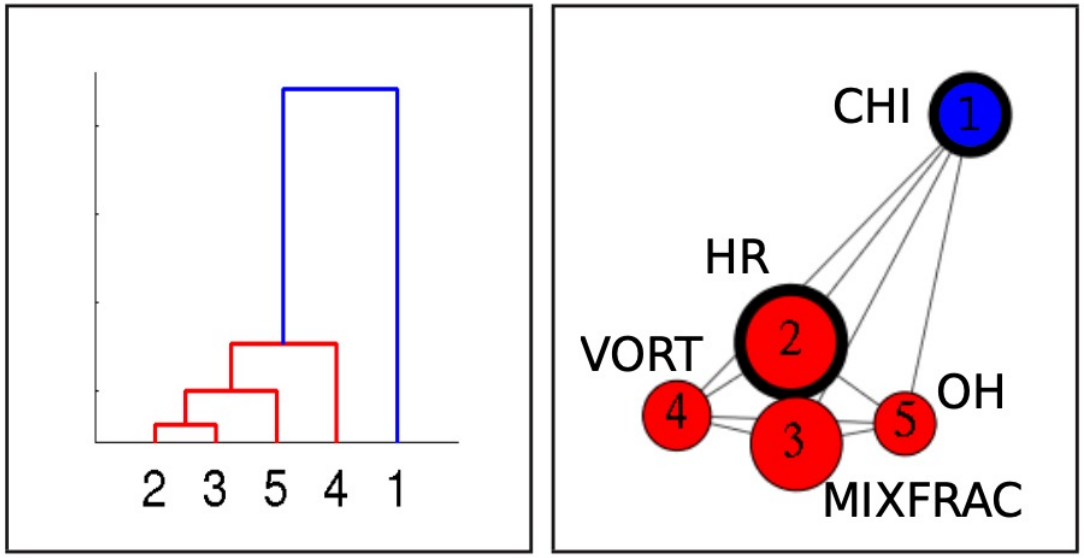
Example from Combustion Data

Two selected variables are: HR and MIXFRAC



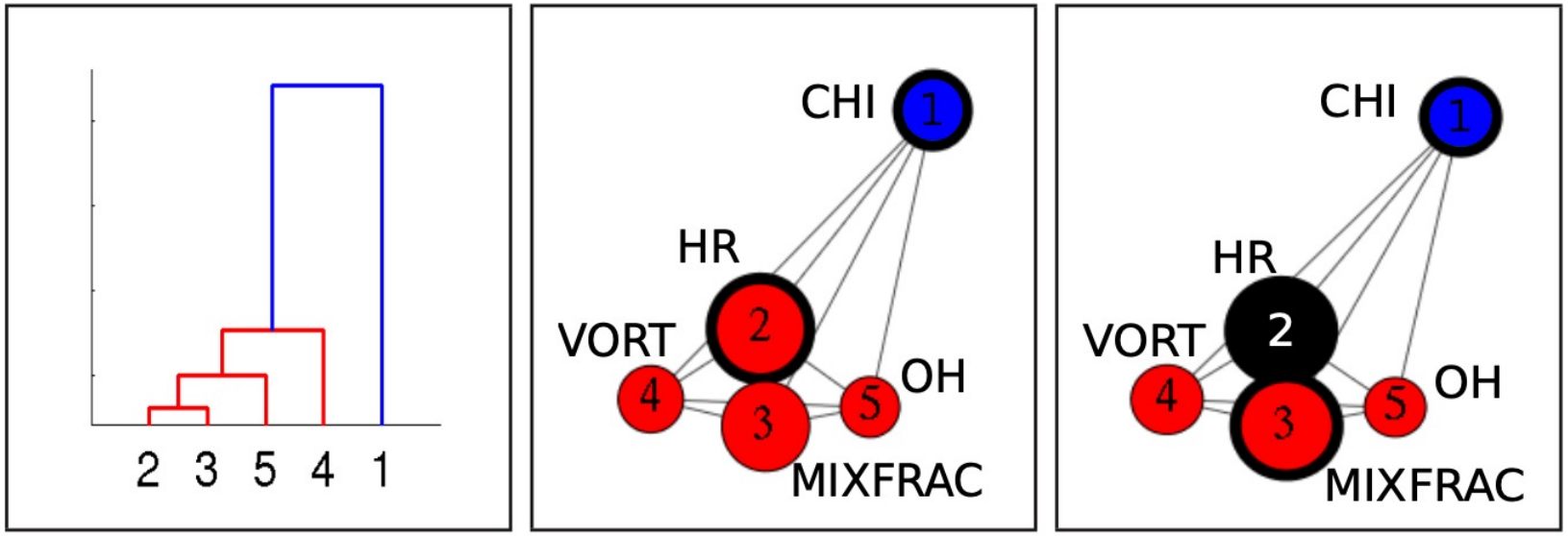
Example from Combustion Data

Two selected variables are: HR and MIXFRAC



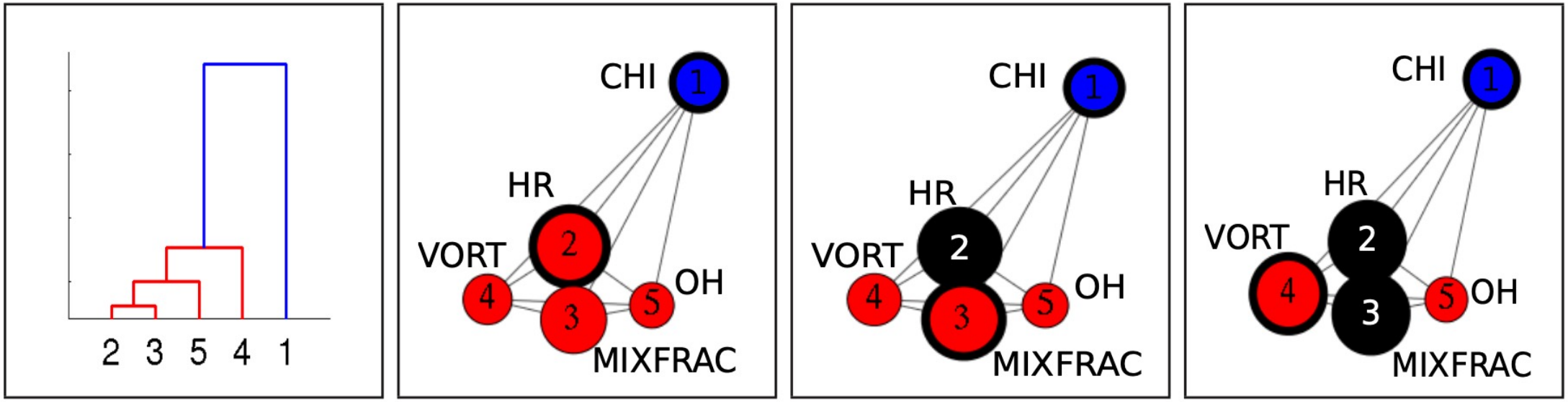
Example from Combustion Data

Two selected variables are: HR and MIXFRAC



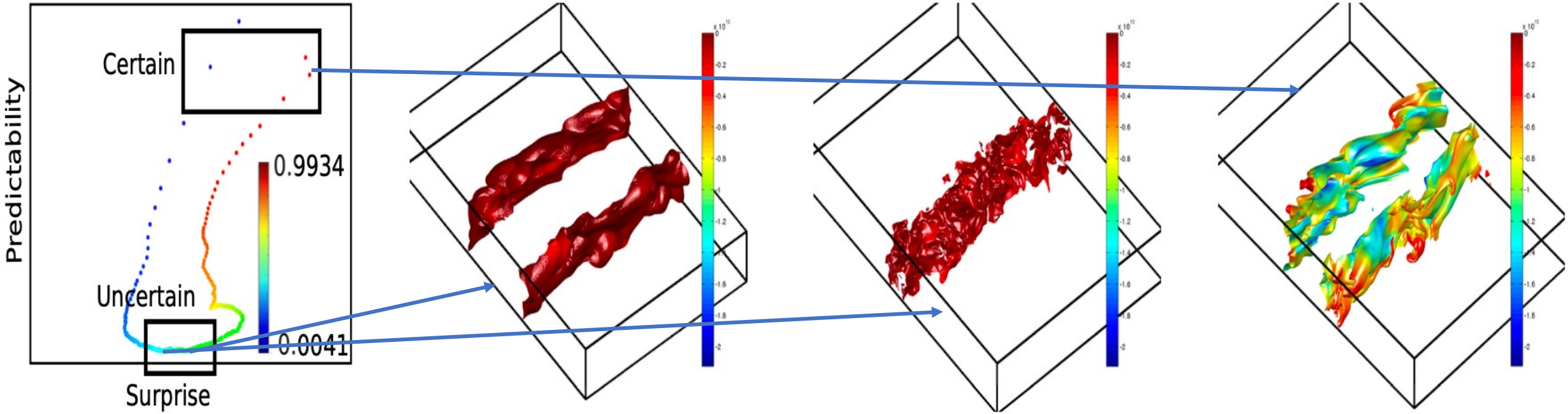
Example from Combustion Data

Two selected variables are: HR and MIXFRAC



Example from Combustion Data

Two selected variables are: HR and MIXFRAC



I1-I2 Map

Certain Isosurfaces

Uncertain Isosurface