# An Example Application of Entropy : Information Theory in Scientific Visualization (Cited from *Entropy* 2011)

Prof. Ja-Ling Wu

Dept. of CSIE and GINM

National Taiwan University

# 1. Introduction

- The field of **visualization** is concerned with the **creation of images from data** to **enhance the user's ability to reason and understand properties related to the underlying problem**.
- Over the past twenty years, visualization has become a standard means to perform **data analysis** for a variety of data intensive applications.

- Numerical simulations for **fluid flow modeling**, high resolution **biomedical imaging**, and **analysis** of **genome and protein sequences** are some examples that can benefit from effective visual data analysis. For these applications, visualization as a fast maturing discipline offers many standard techniques such as **iso-surfaces**, direct **volume rendering**, and **particle tracing** to analyze scalar and vector data defined in the **spatial domain.**

- For **non-spatial data** which is more common for **business applications**, methods such as parallel coordinates, tree maps, and node-link diagrams are widely used.
- Currently, the visual analysis process is mostly operated by the user through **trial and error** in an ***ad hoc* manner**.

- Important **parameters for visualization** algorithms, such as transfer functions, values of iso-contours, levels of detail, and camera positions and directions, often need to be **frequently updated and refined** before satisfactory visualization results are obtained.

- As the **size of data** continues to grow, however, it becomes increasingly difficult to generate useful visualization using this *ad hoc* approach.
- Even after many visualizations have been produced, it may be still **difficult to determine whether the data have been completely analyzed**, or if **some important features are left undiscovered**.

One major cause of the difficulties in visual analysis of large datasets is the **lack of quantitative metrics to measure the visualization quality relative to the amount of information contained in the data**.

As the **size of data** grows even larger, these problems will become even worse since the **user's ability to move and process the data will be severely limited.**

- Without a **systematic and quantitative way** to guide the user through the visual analysis process, visualization could soon lose its value to be a viable approach for large-scale scientific data analysis.
- In the cited article, the authors interpret **information theory principles in the context of scientific visualization**.

- For data analysis and visualization, one may naturally wonder whether information theory can be applied to **improve our understanding of the data** and furthermore, to assist us to **extract hidden salient data features**.

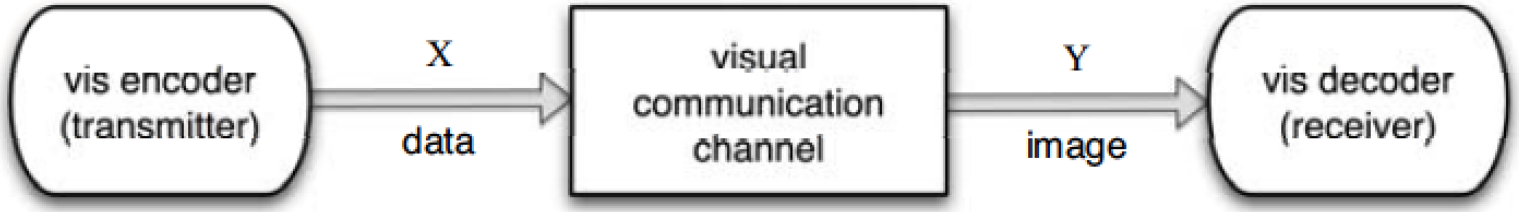
# 2. Visualization and Information Channel

- Figure 1 illustrates the analogy between **data communication** and **data visualization**. In data communication, one attempts to transmit a message *X* through a ***noisy communication channel*** to the destination, the receiver.
- Due to the noisy nature of the channel, **information loss** could be inevitable, resulting in a different version of the message, which we denote as *X'*.

- One obvious goal of data communication is therefore to understand the uncertainty of the symbols embedded in a message so that the that the message can be encoded properly to reduce the possibility of being contaminated in

# Figure 1. The analogy between message transmission and data visualization.


(a) message transmission




(b) data visualization

- Similarly, the **visualization process** can be treated as an **information channel**, *i.e.*, a *visual communication channel* that attempts to communicate the information in the source data to the **destination**, the **viewer**.
- In a typical **visualization pipeline**, the data need to be transformed by a sequence of steps such as **denoising, filtering, visual mapping**, and **projection**.

- Each of the **transformation steps in the visualization pipeline** can be thought of as an **encoding process** where the **goal** is to **preserve** the **maximum amount of information** from the **input** and generate the output for the next stage of the pipeline.

- When **information loss is inevitable**, such as in the case of projecting 3D data to 2D images, special care is needed so that **appropriate parameters are chosen to preserve as much information as possible**. Only in doing so, are we able to **faithfully reveal the information embedded** in the data through visualization.

# 3. **Entropy**

- As a measure of the **average uncertainty** in *X*, the **entropy** is always **nonnegative**, and indicates the **number of bits on average** required **to describe the random variable**.

- **The higher the entropy, the more information the variable contains**.

- An important property of the entropy is that ***H*(*X*) is a concave function** and reaches its maximum of log $|X|$ if and only if ***p*(*x*) is equal** for all *x*, *i.e.*, when the probability distribution is uniform.
- The notion of ***"equal probability, maximum entropy"*** is at the heart of probability function design in many of the visualization examples.

- The key of applying the concept of entropy to visualization problems lies in **how to properly specify the random variable *X*** and **define the probability function *p*(*x*)**.
- In most cases, these **probability functions** can be **defined heuristically** to **meet** the **need of individual applications**.

- To apply the **Shannon entropy**, we can model a scientific dataset as a **discrete random variable** where each data point in the domain carries a value as the outcome.
- The **probability mass function *p*(*x*)** of the random variable *X* can be **estimated using histogram**. That is, we can use the **normalized frequency of each histogram bin** as the probability ***p*(*x*)**.

- In a simple example, given a **3D volume dataset**, we can model the entire dataset as a discrete random variable ***X*** where each voxel carries a **scalar value**.
- The entropy $H(X)$ indicates how much information the dataset contains.

- If the distribution in the histogram is uniform across all bins, then it is difficult to predict the value of a voxel. Thus the **entropy** of the dataset is **high**.
- On the contrary, if the histogram distribution is **highly skewed into a few bins**, then it is easy to guess the value of a voxel. Thus the **entropy** of the dataset is **low**.

- In Figure 2, we show an example 2D hurricane dataset and its derived entropy fields. For Figure 2(b) and (c), a **constant-size 2D local window centered at each pixel** is used to compute the **entropy** in the pixel's neighborhood.
- We **discretize the velocity magnitude** or **direction** into a certain number of bins and compute a **1D histogram** for each local window accordingly.

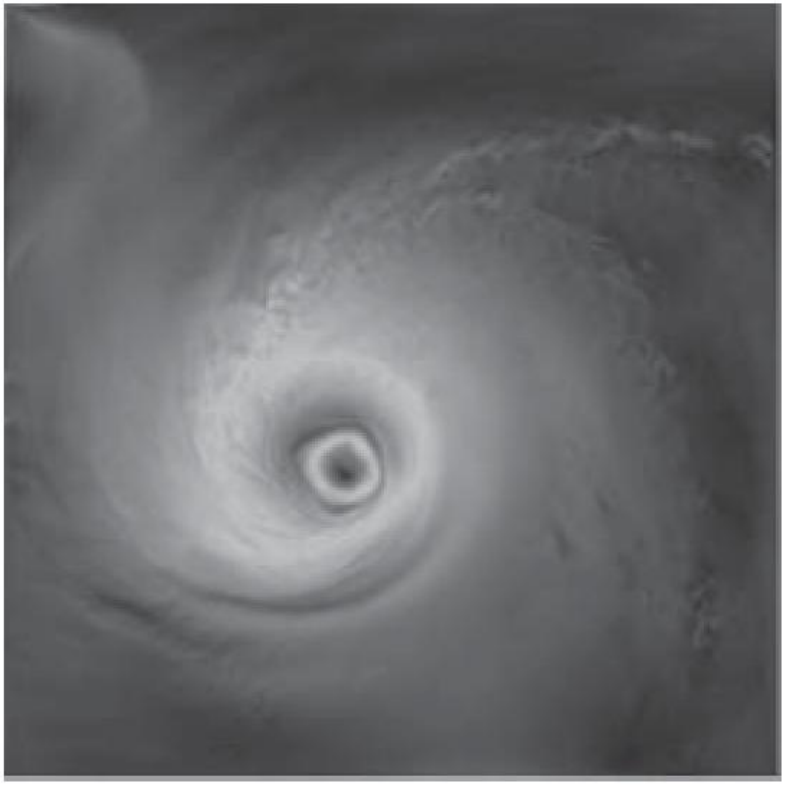
- As we can see in (b) and (c), around the **center** of the hurricane, the **entropy is high** in both evaluations.
- Unlike the velocity magnitude, **the velocity direction** also varies greatly around local regions on the right side of the hurricane's center (as we can see that those regions have high entropies as well).

- We can also **trace streamlines** from the 2D **flow field** and evaluate the entropy associated with each control point along the streamlines.
- For Figure 2(d), a constant-size 1D local window centered at each control point along each streamline is used to evaluate the **entropy** at the **control point**.

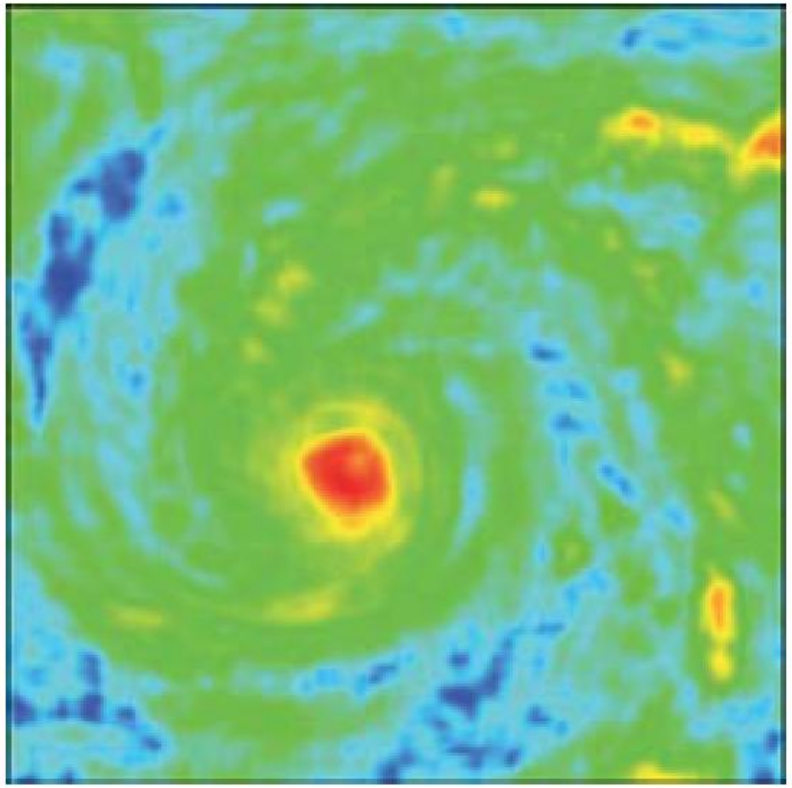
- We create a **2D histogram** in this case for each local window with one dimension for **velocity magnitude** and the other dimension for **velocity direction**.
- We can see that **the streamlines close to the hurricane's center have high entropies**, mainly due to the changes of velocity direction (as evident by the circular flow pattern).

- Intuitively, the **entropy images highlight which regions** in the data are **important** or **interesting** in terms of exhibiting **more variation or change in their local neighborhood** compared with other regions.

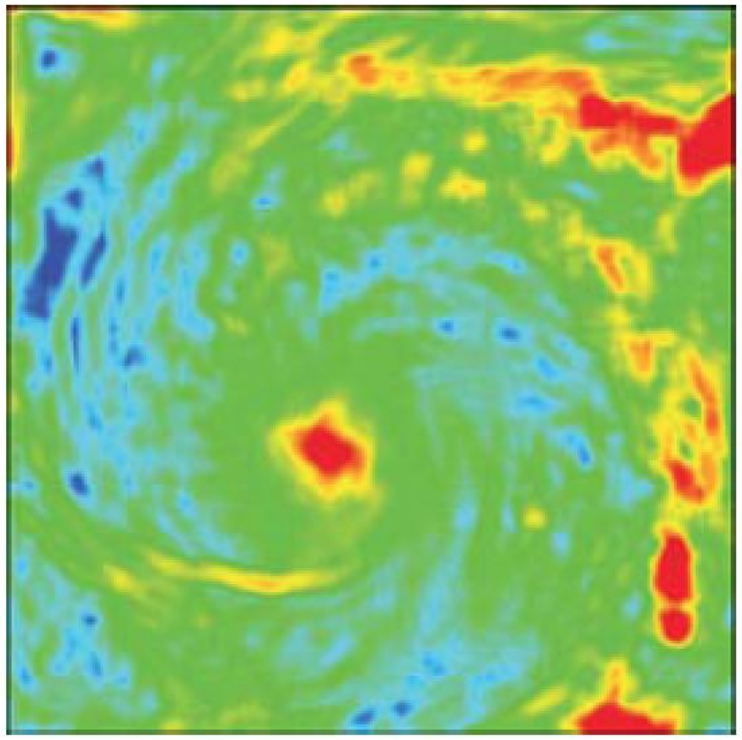
- Figure 2. (a) a 2D hurricane field of velocity magnitude. (b) the entropy field derived from velocity magnitude. (c) the entropy field derived from velocity direction. (d) uniformly placed streamlines with color coded entropy derived from velocity direction and magnitude. The **entropy value increases from blue to green to red** in (b), (c), and (d).



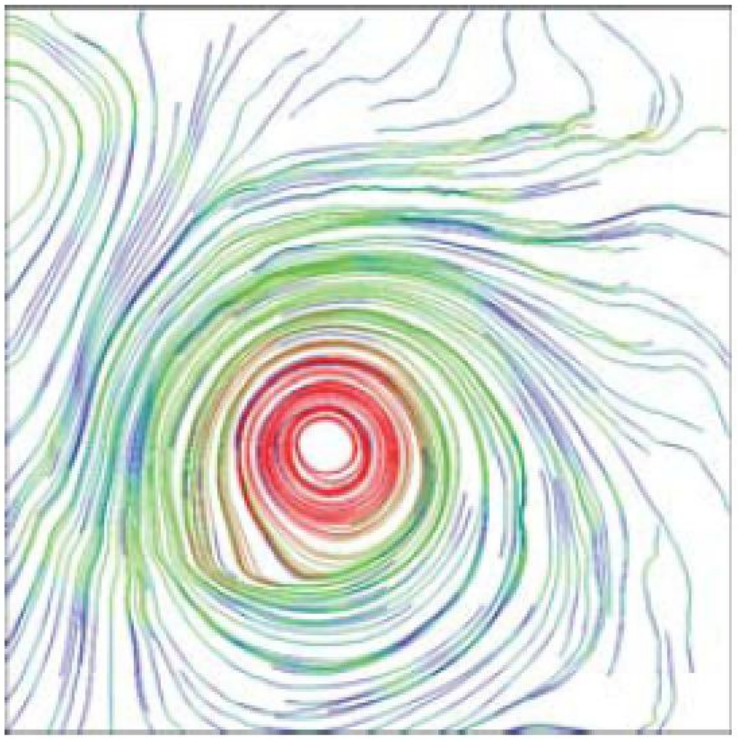
(a)



(b)



(c)



(d)

- It has been demonstrated that in practice, the evaluation of data entropy can be more flexible.
- For instance, depending on the need, we can **partition** a large volume dataset into **individual blocks** and evaluate the **entropy on a per-block basis**.

- We can also consider more than just a single scalar field when building a histogram. This means that the **histogram can be multidimensional**, including **not only the raw data**, but **also other derived quantities** such as **local features** (e.g., gradient magnitude or direction) and/or **domain-specific derivatives**.

- Furthermore, **each bin** in such a multidimensional histogram can carry a **weight** indicating its **relative importance** in the entropy calculation.
- This is the place where **domain knowledge about the data** or **visualization-specific quantities** can be **leveraged**.

- For example, in **volume visualization**, the user needs to specify **a transfer function** so that **scalar data values** can be mapped to **optical quantities** such as **colors** and **opacities**.
- The **opacity value** can be used to set the **weight** for its corresponding histogram bin. A bin with a higher opacity value is likely to have more contribution to the resulting image per voxel, and therefore, should be assigned with a higher weight.

# *3.1 Joint Entropy and Relative Entropy*

- To measure the **distance between two distributions**, we can use the ***Kullback-Leibler divergence***, or ***relative entropy***. Given two random variables *P* and *Q*, the Kullback-Leibler divergence between them is defined as

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},$$

where $p(x)$ and $q(x)$ are the probability mass functions of *P* and *Q*, respectively.

- *P* represents the true distribution of data or observations and *Q* represents a model or approximation of *P*.
- $D_{KL}(P||Q)$ is used to describe **the deficiency of using one distribution *q* to represent the true distribution *p***, which is useful for comparing two related distributions, e.g., two different resolutions of the same dataset.
- The Kullback-Leibler divergence is always **nonnegative** and equals **zero** if and only if $P = Q$.

- There are some issues with the Kullback-Leibler divergence measure that make it less than ideal.
- First, it is not a true metric, *i.e.*, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.
- Second, if $q(x) = 0$ and $p(x) = 0$ for any $x$, then $D_{KL}(P||Q)$ is undefined.
- Third, the Kullback-Leibler divergence does not offer any nice upper bounds.

- To overcome these problems, we may consider the **symmetric *Jensen-Shannon divergence* measure**

$$D_{JS}(P||Q) = D_{JS}(Q||P) = \frac{1}{2}\Big(D_{KL}(P||M) + D_{KL}(Q||M)\Big),$$

$$\text{where } M = \frac{P+Q}{2}.$$

- The Jensen-Shannon divergence can be expressed in terms of entropy, *i.e.*,

$$D_{JS}(P||Q) = H\Big(\frac{1}{2}P + \frac{1}{2}Q\Big) - \frac{1}{2}\Big(H(P) + H(Q)\Big).$$

In general, the Jensen-Shannon divergence has the following form

$$D_{JS}(\lambda_1, \lambda_2, \ldots, \lambda_n; P_1, P_2, \ldots, P_n) = H\Big(\sum_{i=1}^{n} \lambda_i P_i\Big) - \sum_{i=1}^{n} \lambda_i H(P_i),$$

where $\lambda_i \in [0.0, 1.0]$ and $\sum_{i=1}^{n} \lambda_i = 1.0$.

- Wang and Ma [1] utilized the Kullback-Leibler divergence to quantify the **difference between wavelet coefficient distributions** of the **original and distorted data**.


- [1]. Wang, C.; Ma, K.L. A statistical approach to volume data quality assessment. *IEEE Trans. Vis.Comput. Graph.* 2008, *14*, 590–602.


- Bordoloi and Shen [2] utilized the **Jensen-Shannon divergence** to evaluate the **similarity of two viewpoints**, The similarity values were used to generate a view space partitioning and select representative views.

- [2]. Bordoloi, U.D.; Shen, H.W. View selection for volume rendering. In Proceedings of IEEE Visualization Conference, Minneapolis, MN, USA, October 2005; pp. 487–494.

# *3.2 **Mutual Information** and Conditional Entropy*

- We can **measure how much information of a random variable *X* is conveyed by another random variable *Y*** using the concept of *mutual information*.
- **Mutual information** can be treated as a special case of **relative entropy**: it is the relative entropy between the **joint distribution *p*(*x, y*)** and the **product distribution *p*(*x*)*p*(*y*)**.

- **Mutual information measures the amount of information that *X* and *Y* share.** It is **the reduction in the uncertainty of one random variable due to the knowledge of the other**.
- For example, if *X* and *Y* are independent, *i.e.*, $p(x, y) = p(x)p(y)$, then knowing *X* does not give any information about *Y* and vice versa. Therefore, $I(X; Y) = 0$.

- At the other extreme, if *X* and *Y* are identical, then all information conveyed by *X* is shared with *Y* : knowing *X* determines the value of *Y* and vice versa.
- As a result, ***I*(*X*; *Y* ) is the same as the uncertainty contained in *X* (or *Y* ) alone, namely the entropy of *X* (or *Y* )**.

- $I(X; Y)$ is bounded above by the smaller of log $|X|$ and log $|Y|$. Janicke *et al.* [3] used the **normalized mutual information**, *i.e.*, $\frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$, to compute the **distance between two power spectra** transformed from **climate data**.
- Bruckner and Moller [4] used another version of normalized mutual information, *i.e.*, $2I(X;Y) / H(X)+H(Y)$, to evaluate **the similarity between two iso-surfaces**.

- [3]. Janicke, H.; Bottinger, M.; Mikolajewicz, U.; Scheuermann, G. Visual exploration of climate variability changes using wavelet Analysis. *IEEE Trans. Vis. Comput. Graph.*2009,*15*, 1375–1382.
- [4]. Bruckner, S.; Moller, T. Isosurface similarity maps. *Comput. Graph. Forum* 2010, *29*, 773–782.


# *3.3 Relationships among Information Theory Concepts*

- Mutual information, entropy, joint entropy, and conditional entropy have the following **relationships**

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

- In practice, we can treat $X$ and $Y$ as two relevant random variables, such as two scalar volumes drawn from different time steps of the same dataset.
- Mutual information $I(X; Y)$ indicates the amount of information $X$ and $Y$ share in common,
- conditional entropy $H(X|Y)$ tells how much information about $X$ is still unknown after observing $Y$, and
- joint entropy $H(X, Y)$ indicates the total information the two volumes have.

- Another important property for entropy is the ***chain rule***, which states that **the entropy of a collection of random variables** is the sum of the conditional entropies. Let $X1, X2, \ldots, Xn$ be drawn according to $p(X1), p(X2), \ldots, p(Xn)$ respectively, then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{N} H(X_i | X_{i-1}, \ldots, X_1).$$

- Assuming a **Markov sequence model** for the random variables, *i.e.*, any variable $X_i$ is dependent on variable $X_{i-1}$, but independent of other variables, we have

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}, \ldots, X_1) \\ &= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}). \end{aligned}$$

- The chain rule in conjunction with the Markov sequence model described above was utilized by Bordoloi and Shen [2] to define a **viewpoint goodness measure** for time-varying volume data and by Wang *et al.* [5] to select representative timesteps from time-varying data.


- [5]. Wang, C.; Yu, H.; Ma, K.L. Importance-driven time-varying data visualization. *IEEE Trans. Vis.Comput. Graph.* 2008, *14*, 1547–1554.


- Figure 3 summarizes the relationships among the various measures in information theory between two random variables *X* and *Y* . It also highlights the goal of data visualization on the right.
- Assuming the **input dataset** is denoted as a random variable ***X***, we can **model the visualization** as another random variable ***Y*** , the **output from the *visual communication channel***.

- To produce **insightful visualization**, the amount of mutual information ***I*(*X*; *Y* )** needs to be **as high as possible** (or equivalently, the conditional entropy ***H*(*X*|*Y* )** should be **as low as possible**).
- When *H*(*X*|*Y* ) reaches zero, the visualization fully conveys the information contained in the dataset.

- By **optimization**, we mean **adjusting visualization parameters**, such as the view or transfer function, so that the mutual information $I(X; Y)$ between the input data $X$ and the output visualization $Y$ can be **maximized**.

- Figure 3. Left: Relationships among different entropy measures between two random variables *X* and *Y* . Right: The goal of data visualization is to maximize the mutual information *I*(*X*; *Y* ) between the input data *X* and the output visualization *Y* .