Information Theory and Visualization

Tutorial on Information Theory in Visualization

Min Chen
University of Oxford
The Role of a Theoretic Framework

- **Facts**
- **Wisdom**
- **Theory**

Diagram: 
- Theoreric Framework

Nodes and Connections: 
- T
- X

Legend: 
- T
- X
Three Visualization Subsystems

Source \(\xrightarrow{\text{raw data}}\) vis-encoder \(\xrightarrow{\text{image}}\) vis-channel \(\xrightarrow{\text{image}}\) vis-decoder \(\xrightarrow{\text{knowledge}}\) Destination

Source \(\xrightarrow{\text{message}}\) Encoder (Transmitter) \(\xrightarrow{\text{signal}}\) Channel \(\xrightarrow{\text{signal}}\) Decoder (Receiver) \(\xrightarrow{\text{message}}\) Destination

compactness \(\uparrow\) error detection \(\uparrow\) error correction
Claude E. Shannon (1916-2001)

Entropy

- Random variable (alphabet)
  - $X$

- It takes values (letters)
  - $x_1, x_2, ..., x_m$

- Probability mass function
  - $p(x_i)$

- Entropy (uncertainty)

$$H(X) = - \sum_{i} p(x_i) \log_2 p(x_i)$$
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary

Min Chen
Shannon’s Definition of Information

\[ H(Z) = - \sum_{z \in Z} p(z) \log_2 p(z) \]

Communication System

Alphabet \( Z \)

G, A, C, T

\( Z \)

\( \leq 2 \) bits uncertainty
Shannon’s Definition of Information

\[ \mathcal{H}(Z) = - \sum_{z \in Z} p(z) \log_2 p(z) \]

Communication System

Alphabet \( Z \)

G, A, C, T

\( G \)

a variable \( Z \)

\( \leq 2 \) bits information

Shannon Information
An Alphabet and its Letters

- English alphabet: A, B, C, ..., X, Y, Z
- All English prefixes: bio, geo, pre, pro, ..., un
- All English words: a, ..., silicosis, ..., titin, ...
- All sentences in a text corpus: pneumonoultrami
  microscopicsilico
  volcanocnosis
- All published BioVis papers: ...
- ...

189,819 letters: a word or a formula?
Data Processing Inequality

\[ I(X;Y) \geq I(X;Z) \]

- “No clever manipulation of data can improve the inferences that can be made from the data”
  [Cover and Thomas, 2006]
Mutual Information (shared uncertainty)

\[ I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \left( \frac{p(x, y)}{p_X(x)p_Y(y)} \right) \]
An Example Data Analysis and Visualization Process

- $r$ time series
- 720 data point each series
- $2^{32}$ valid value each point
- $r$ decisions
- 3 valid values each (e.g., buy, sell, hold)

- $Z_1$ Raw Data
  - 1 hour long
  - at 5-second resolution
  - $r$ time series
  - $\times 720$ data points
  - $\times 2^{32}$ valid values

- $Z_2$ Aggregated Data
  - at 1-minute resolution
  - $r$ time series
  - $\times 60$ data points
  - $\times 2^{32}$ valid values

- $Z_3$ Time Series Plots
  - $r$ time series
  - $\times 60$ data points
  - $\times 128$ valid values
  - $\mathcal{H}_\text{max} = 420r$

- $Z_4$ Feature Recognition
  - $r$ time series
  - $\times 10$ features
  - $\times 8$ valid values
  - $\mathcal{H}_\text{max} = 30r$

- $Z_5$ Correlation Indices
  - $r(r-1)/2$ data points
  - $\times 2^{30.7}$ valid values
  - $\mathcal{H}_\text{max} \approx 1.16r(r-1)$

- $Z_6$ Graph Visualization
  - $r(r-1)/2$ connections
  - $\times 5$ valid values
  - $\mathcal{H}_\text{max} \approx 1.58r$

- $Z_7$ Decision
  - $r$ decisions
  - $\times 3$ valid values

$\mathcal{H}_\text{max} = 23040r$

$\mathcal{H}_\text{max} = 1920r$

$\mathcal{H}_\text{max} = 30r$

$\mathcal{H}_\text{max} = 420r$

$\mathcal{H}_\text{max} = 30r$

$\mathcal{H}_\text{max} \approx 1.16r(r-1)$

$\mathcal{H}_\text{max} \approx 1.58r$
Data Processing Inequality

Decreasing of Mutual Information

\[ I_1 \geq I_2 \geq \ldots \geq I_{L-1} \geq I_L \]
Data Processing Inequality: Big Data Input?

Big Data → Process 1 → Process 2 → ...... → Process L-1 → Process L → Decision

\[ H(Z_1) \]

\[ H(Z_{L+1}) \]

\[ I(Z_1; Z_{L+1}) \]

mutual information
DPI is not Ubiquitous

- Markov chain conditions
  - Closed coupling: \((X, Y), (Y,Z)\)
  - \(X\) and \(Z\) are conditionally independent
- What if one of the conditions is broken?
- In visual analytics, both conditions are usually broken.

\[ p(x, y, z) = p(x) p(y|x) p(z|y) \]

\[
\begin{align*}
I(X; Y) &\geq I(X; Z) \\
I(Y; Z) &
\end{align*}
\]

Soft Knowledge in Data Intelligence

All possible decisions under different conditions

a) totally data-driven
b) totally instinct-driven
c) data-informed
d) due to unknown or uncontrollable factors

$\mathcal{H}(Z_1)$

$x \in \mathcal{X}$ is a piece of soft knowledge

$I(Z_1; X)$

mutual information
Data \rightarrow \text{Process 1} \rightarrow \cdots \rightarrow \text{Process s} \rightarrow \cdots \rightarrow \text{Process L} \rightarrow \text{Decision}

\text{forward mapping } F

\text{Alphabet Compression}

\text{backward mapping } G

A Sequential Workflow and Two Basic Metrics

- The $s^{th}$ Function (Process):

- Alphabetic Compression Ratio (ACR):

- A Reverse “Guessing” Process:

- Potential Distortion Ratio (PDR):
Kullback–Leibler divergence

\[ \mathcal{D}_{KL}(Z' \| Z) = \sum_{(z=z') \in Z} p(z') \log_2 \frac{p(z')}{q(z)} \]
Cost-Benefit Ratio

- Effectual Compression Ratio (ECR):

\[ \Psi_{ECR}(F_s) = \frac{\mathcal{H}(Z_{s+1}) + \mathcal{D}_{KL}(Z'_s||Z_s)}{\mathcal{H}(Z_s)} \]

- Incremental Cost-Benefit Ratio (ICBR):

\[ \Upsilon(F_s) = \frac{\mathcal{B}(F_s)}{\mathcal{C}(F_s)} = \frac{\mathcal{H}(Z_s) - \mathcal{H}(Z_{s+1}) - \mathcal{D}_{KL}(Z'_s||Z_s)}{\mathcal{C}(F_s)} \]

- Cost can be measured in energy, time, money, etc.

Cost-Benefit Optimization

Data → Process 1 → ... → Process s → ... → Process L → Decision

Alphabet Compression — Potential Distortion

Cost

Process s

Alphabet Compression — Potential Distortion

Cost

Process s+1

Composite Process for s and s+1
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary

Min Chen
Three Spaces and Three Measures

- Entropy of Input Data Space: $H(X)$
- Visualization Capacity: $V(G)$
- Display Capacity: $D$

Visual Mapping Ratio (VMR) = \( \frac{V(G)}{H(X)} \)

Information Loss Ratio (ILR) = \( \frac{\max(H(X) - V(G), 0)}{H(X)} \)

Display Space Utilization (DSU) = \( \frac{V(G)}{D} \)
**Example of V(G)**

- **Entropy of Data Alphabet**
  \[ H(Z) = -\sum_{i=0}^{255} \frac{1}{256} \log_2 \frac{1}{256} = 512 \]

- **V(G) = H(Z)**

- **Binary Pixel Plot: D**
  - 4x4 pixels per bit
  - D: \(2^9 \times 2^4 = 2^{13}\) bits

- **Time Series Plot: D**
  - Minimal 256x64 pixels
  - D: \(2^8 \times 2^6 = 2^{14}\) bits

- **The more compact, the better?**
- **Cost?**
- **Reconstructability?**
Information Loss Ratio (ILR)

- Display Space Restriction
  - 64x64 pixels
- Evenly distributed probability mass function
- Linear visual mapping
- ILR is a probabilistic measure about
  - a data space $X$
  - not an instance $x_i$
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary
Information Loss Ratio (ILR)

- Display Space Restriction
  - 64x64 pixels

- Evenly distributed probability mass function

- Linear visual mapping

- ILR is a probabilistic measure about
  - a data space $X$
  - not an instance $x_i$

Non-uniform Distribution

- Linear visual mapping

<table>
<thead>
<tr>
<th>probability</th>
<th>Z</th>
<th>linear</th>
<th>Z'</th>
</tr>
</thead>
<tbody>
<tr>
<td>D: $p = \frac{1}{8}$</td>
<td>256</td>
<td>224</td>
<td>56</td>
</tr>
<tr>
<td>C: $p = \frac{1}{8}$</td>
<td>256</td>
<td>192</td>
<td>48</td>
</tr>
<tr>
<td>B: $p = \frac{1}{4}$</td>
<td>256</td>
<td>160</td>
<td>40</td>
</tr>
<tr>
<td>A: $p = \frac{1}{2}$</td>
<td>256</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>96</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Information loss:

(a) evenly distributed $p$

(b) unevenly distributed $p$

- Information loss: 25.0%
- Information loss: 25.8%
Non-uniform Distribution

- Nonlinear visual mapping

- Information loss:
  - (a) evenly distributed $p$: 25.0%
  - (b) unevenly distributed $p$: 25.8%
  - (c) 4 regional mappings: 22.6%
Non-uniform Distribution

- Logarithmic visual mapping

A: $p = \frac{1}{2}$
B: $p = \frac{1}{4}$
C: $p = \frac{1}{8}$
D: $p = \frac{1}{8}$

$k \to \infty$

$p = \frac{1}{2^k}$
$p = \frac{1}{2^{k-1}}$
$p = \frac{1}{2^3}$
$p = \frac{1}{2^2}$
$p = \frac{1}{2}$

information loss:
(a) evenly distributed $p$: 25.0%
(b) unevenly distributed $p$: 25.8%
(c) 4 regional mappings: 22.6%
(d) logarithmic plot: 0%
Display capacity $D$ is limited.

Data space: noticeable non-uniform distribution: $H(X) << H_{max}(X)$

Visualization capacity $V(G)$: a visual representation exhibiting a similar non-uniform distribution of space requirement can reduce information loss.

Can adjacency matrixes be improved for trees?

http://hci.stanford.edu/jheer/files/zoo/
http://en.wikipedia.org/wiki/Treemapping
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary
Map Overlay and its Generalization in Visualization

http://learnpracticalgis.com/how-to-overlay-maps/
- Frequency-division multiplexing (FDM)
- Time-division multiplexing (TDM)
- Space-division multiplexing (SDM)
- Code-division multiplexing (CDM)

Location $p$ can be associated with $X$ in the source data or determined by a spatial mapping.

$X = \langle x_1, x_2, \ldots, x_k \rangle$ at $p$

Other signals and noise

Spatial domain $D$

Temporal domain $T$

Perceived information may include estimated values and relationships with data conveyed by other signals.

$X$ can be a data record or a set of partially encoded visual attributes.

10 Types of Visual Multiplexing

(a) Type A: Partition a space
(b) Type B: Partition a time period
(c) Type C: Introduce partial occlusion
(d) Type D: Use a ‘hollow’ visual channel
(e) Type E: Introduce translucent occlusion
(f) Type F: Use an integrated visual channel
(g) Type G: Depict a continuous field
(h) Type H: Shift a visual channel
(i) Type I: Use periodic motion
(j) Type J: Assume a priori knowledge
(k) Type J (continued): Acquired knowledge
(l) Type J (continued): Visual language
Type C: Introduce Partial Occlusion
Type G: Depict a Continuous Field
Type H: Shift a Visual Channel
Type J: Assume A Priori Knowledge
How about When Every Pixel is Used?

Data Space

Visualization Space

Display Space

\[ H \]

Data Space Entropy

\[ V(G) \]

Visualization Capacity

(Visualization Space Entropy)

\[ D \]

Display Space Capacity

\[ V(G) \]

\[ D \]

\[ \ll 1 \]
C: Introduce Partial Occlusion

E: Introduce Translucent Occlusion

G: Depict a Continuous Field

J: Assume A Priori Knowledge
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary
1. Disseminative Level \( O(1) \)
   - This is “a”!

2. Observational Level \( O(n) \)
   - “a”, “b”, “c”, ... what, when, where?

3. Analytical Level \( O(n^k) \)
   - Are “a”, “b”, “c” related? Why?

4. Model-developmental Level \( O(kn), O(n!) \)
   - How does “a” lead to “b”?
A Composite Workflow

1. **Historical Data**: Creating a new weather forecasting model.
2. **Known Errors**: Optimizing a weather forecasting model.
3. **Initial Conditions**: Running an ensemble of simulation models.
4. **Simulation Results**: Producing a 5-day weather forecast.
5. **Forecasting Results**: Presenting a 5-day TV weather forecast.
6. **Forecast Presentation**: Making a decision about what to wear.
Observational Visualization

- Real-time or offline annotation results in a huge spreadsheet of events
Observational Visualization

Input Video

Events  Timeline  Occurrences
Information Theory and Visualization

1. Data Intelligence — a big picture
2. Visualization — a small picture
3. Measurement, Explanation, and Prediction
4. Example: Visual Multiplexing
5. Example: Error Detection and Correction
6. Example: Process Optimization
7. Summary
<table>
<thead>
<tr>
<th>Information Theory Taxonomy</th>
<th>Relevance in Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fundamental Concepts</strong></td>
<td>A possible mathematical framework that underpins the subject of visualization.</td>
</tr>
<tr>
<td><strong>Major Quantities and Properties</strong></td>
<td>Quantitative measurements about the data and visualization space, and the relationship between input and output of a process or subsystem at different stages of a visualization pipeline.</td>
</tr>
<tr>
<td>Entropy</td>
<td>Measuring information content (see Section 5.1); salience in visualization.</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>Uncertainty reduction in visualization (see Section 5.3); information-assisted visualization.</td>
</tr>
<tr>
<td><strong>Major Theorems</strong></td>
<td>Many theorems can be used to explain visualization phenomena and events.</td>
</tr>
<tr>
<td>Information balance (conservation law)</td>
<td>Given two visualizations, A and B, the amount of information about A contained in B is the same as that about B in A; overview + detail visualization; multi-view visualization.</td>
</tr>
<tr>
<td>Data processing inequality</td>
<td>After visual mapping, the visualization cannot contain more information than the original data (see Section 6.2); Information cannot be recovered after being degraded by some processes or subsystems in a visualization pipeline.</td>
</tr>
<tr>
<td><strong>Channel Types</strong></td>
<td>Providing a theoretical basis for classifying visualization subsystems (see Section 6).</td>
</tr>
<tr>
<td>Noiseless channel</td>
<td>Not common in practical visualization pipelines (see Section 3).</td>
</tr>
<tr>
<td>Noisy channel</td>
<td>Most visualization processes and subsystems can be affected by noise (see Section 3).</td>
</tr>
<tr>
<td><strong>Channel Capacity</strong></td>
<td>It can be adapted to define the maximum amount of information that can be visualized or displayed (see Sections 5.1 and 6).</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Efficiency of a visual mapping; Error detection and correction (see Section 8).</td>
</tr>
<tr>
<td><strong>Source Coding (for Noiseless Channels)</strong></td>
<td>Inspiration for developing new data abstraction and visual encoding techniques.</td>
</tr>
<tr>
<td>Coding Schemes</td>
<td>Applicable, for example, to the following visualization algorithms:</td>
</tr>
<tr>
<td>Entropy coding (e.g., Huffman, arithmetic coding)</td>
<td>Logarithmic plots (see Section 7.1); importance-based visualization; information-assisted visualization; magic lens; illustrative deformation.</td>
</tr>
<tr>
<td>Dictionary-based coding</td>
<td>Legend design; icon design; visualization literacy; knowledge-assisted visualization.</td>
</tr>
<tr>
<td>Run-length encoding</td>
<td>Spatial clustering; cluttering reduction.</td>
</tr>
<tr>
<td>Differential encoding</td>
<td>Video visualization; time-varying data visualization.</td>
</tr>
<tr>
<td>Subband coding</td>
<td>Multi-resolution modeling; transfer function design in volume rendering.</td>
</tr>
<tr>
<td>Transform coding</td>
<td>Perceptually-based visual encoding, frequency-domain volume rendering.</td>
</tr>
<tr>
<td>Quantization</td>
<td>Color mapping; multi-resolution modeling.</td>
</tr>
<tr>
<td>Multiplexing</td>
<td>Comparative visualization; volume rendering; multi-field visualization.</td>
</tr>
</tbody>
</table>