CSE 442 - Data Visualization

Visual Encoding Design

Jeffrey Heer  University of Washington
A Design Space of Visual Encodings
Mapping Data to Visual Variables

Assign **data fields** (e.g., with N, O, Q types) to **visual channels** \((x, y, \text{color}, \text{shape}, \text{size}, \ldots)\) for a chosen **graphical mark** type \((\text{point, bar, line,} \ldots)\).

Additional concerns include choosing appropriate **encoding parameters** \((\text{log scale, sorting,} \ldots)\) and **data transformations** \((\text{bin, group, aggregate,} \ldots)\).

These options define a large combinatorial space, containing both useful and questionable charts!
1D: Nominal

Raw

Aggregate (Count)
Expressive?

Raw

Aggregate (Count)
1D: Quantitative

**Raw**

**Aggregate (Count)**
Expressive?

Raw

Aggregate (Count)
Effective?

Raw

Aggregate (Count)

![Bar chart](image)
2D: Nominal x Nominal

Raw

Aggregate (Count)
2D: Quantitative x Quantitative

Raw

Aggregate (Count)
2D: Nominal x Quantitative

Raw

Aggregate (Mean)
3D and Higher

**Two variables** \([x, y]\)
Can map to 2D points.
Scatterplots, maps, …

**Third variable** \([z]\)
Often use one of size, color, opacity, shape, etc. Or, one can further partition space.

*What about 3D rendering?*

[Bertin]
Other Visual Encoding Channels?

Wind Map

April 1, 2015
11:35 pm EST
(time of forecast download)

Top speed: 30.5 mph
Average: 10.2 mph
Encoding Effectiveness
## Effectiveness Rankings

<table>
<thead>
<tr>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Sat</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color Hue</td>
<td>Connection</td>
</tr>
<tr>
<td>Area (Size)</td>
<td>Texture</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Connection</td>
<td>Density (Value)</td>
</tr>
<tr>
<td>Density (Value)</td>
<td>Containment</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Sat</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
# Effectiveness Rankings

[McKinlay 86]

<table>
<thead>
<tr>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position</strong></td>
<td><strong>Position</strong></td>
<td><strong>Position</strong></td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Sat</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color Hue</td>
<td>Connection</td>
</tr>
<tr>
<td>Area (Size)</td>
<td>Texture</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Connection</td>
<td>Density (Value)</td>
</tr>
<tr>
<td>Density (Value)</td>
<td>Containment</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Sat</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
Effectiveness Rankings

**QUANTITATIVE**
- Position
- Length
- Angle
- Slope
- Area (Size)
- Volume
- Density (Value)
- Color Sat
- **Color Hue**
- Texture
- Connection
- Containment
- Shape

**ORDINAL**
- Position
- Density (Value)
- Color Sat
- **Color Hue**
- Texture
- Connection
- Containment
- Length
- Angle
- Slope
- Area (Size)
- Volume
- Shape

**NOMINAL**
- Position
- **Color Hue**
- Texture
- Connection
- Containment
- Density (Value)
- Color Sat
- Shape
- Length
- Angle
- Slope
- Area
- Volume

[Mackinlay 86]
Color Encoding (Choropleth Map)
# Effectiveness Rankings

<table>
<thead>
<tr>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Sat</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color Hue</td>
<td>Connection</td>
</tr>
<tr>
<td><strong>Area (Size)</strong></td>
<td>Texture</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Connection</td>
<td>Density (Value)</td>
</tr>
<tr>
<td><strong>Density (Value)</strong></td>
<td>Containment</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Sat</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
Color Encoding (Choropleth Map)
Gene Expression Time-Series [Meyer et al ‘11]

Color Encoding
## Effectiveness Rankings

<table>
<thead>
<tr>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position</strong></td>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Sat</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color Hue</td>
<td>Connection</td>
</tr>
<tr>
<td>Area (Size)</td>
<td>Texture</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Connection</td>
<td>Density (Value)</td>
</tr>
<tr>
<td><strong>Density (Value)</strong></td>
<td>Containment</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Sat</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td><strong>Color Hue</strong></td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
Gene Expression Time-Series [Meyer et al. ’11]

Color Encoding

Position Encoding
Artery Visualization [Borkin et al ’11]

Rainbow Palette

2D

Shear Stress (Pa)

39%

Diverging Palette

3D

Shear Stress (Pa)

71%

62%

92%
Effectiveness Rankings

**QUANTITATIVE**
- Position
- Length
- Angle
- Slope
- Area (Size)
- Volume
- Density (Value)
- Color Sat
- Color Hue
- Texture

**ORDINAL**
- Position
- Density (Value)
- Color Sat
- Color Hue
- Texture
- Connection

**NOMINAL**
- Position
- Color Hue
- Texture
- Connection
- Containment
- Density (Value)
- Color Sat
- Shape
- Length
- Angle
- Slope
- Area
- Volume
Scales & Axes
Include Zero in Axis Scale?

Government payrolls in 1937 [How To Lie With Statistics. Huff]
Include Zero in Axis Scale?

Yearly CO$_2$ concentrations  [Cleveland 85]
Include Zero in Axis Scale?

Violates Expressiveness Principle!

Compare Proportions (Q-Ratio)

Compare Relative Position (Q-Interval)
Axis Tick Mark Selection

What are some properties of “good” tick marks?
Axis Tick Mark Selection

Simplicity - numbers are multiples of 10, 5, 2
Coverage - ticks near the ends of the data
Density - not too many, nor too few
Legibility - whitespace, horizontal text, size
How to Scale the Axis?
One Option: Clip Outliers
Clearly Mark Scale Breaks

Violates Expressiveness Principle!

Poor scale break [Cleveland 85]  Well-marked scale break [Cleveland 85]
Scale Break vs. Log Scale

Scale Break

Log Scale

[Cleveland 85]
Scale Break vs. Log Scale

Both increase visual resolution

Scale break: difficult to compare (cognitive – not perceptual – work)
Log scale: direct comparison of all data
Logarithms turn multiplication into addition.

\[ \log(x \cdot y) = \log(x) + \log(y) \]

Equal steps on a log scale correspond to equal changes to a multiplicative scale factor.
Linear Scale vs. Log Scale

Linear Scale

Log Scale
Linear Scale
Absolute change

Log Scale
Small fluctuations
Percent change
\[ d(10,30) > d(30,60) \]
When To Apply a Log Scale?

Address data skew (e.g., long tails, outliers)
Enables comparison within and across multiple orders of magnitude.

Focus on multiplicative factors (not additive)
Recall that the logarithm transforms $\times$ to $+!$
Percentage change, not linear difference.

Constraint: positive, non-zero values
Constraint: audience familiarity?
Administrivia
A2: Deceptive Visualization

Design two static visualizations for a dataset:
1. An earnest visualization that faithfully conveys the data
2. A deceptive visualization that tries to mislead viewers

Your two visualizations may address different questions.

Try to design a deceptive visualization that appears to be earnest: *can you trick your classmates and course staff?*

You are free to choose your own dataset, but we have also provided some preselected datasets for you.

Submit two images and a brief write-up on Canvas.

Due by **Wed 10/20 11:59pm**.
Design Exercise
Visual Encoding Exercise

5 17

How many visualizations can you think of for conveying these two numbers? Feel free to invent tasks or contexts. Sketch as many as you can!

Don’t stress over quality, go for quantity.

Time: ~5 minutes
Visual Encoding Exercise

5 17

Share your designs with fellow students. Introduce yourselves! Then compare your designs. How many ideas are the same? How many are different? Capture your favorite images and post them on the Ed thread “In-Class Design Activity”. 
Multidimensional Data
Visual Encoding Variables

Position (X)
Position (Y)
Area
Value
Texture
Color
Orientation
Shape

~8 dimensions?
A *trellis plot* subdivides space to enable comparison across multiple plots. Typically, nominal or ordinal variables are used as dimensions for subdivision.
Small Multiples

[MacEachren ‘95, Figure 2.11, p. 38]
Small Multiples

[MacEachren '95, Figure 2.11, p. 38]
Scatterplot Matrix (SPLOM)

Scatter plots for pairwise comparison of each data dimension.
Parallel Coordinates [Inselberg]
Parallel Coordinates \cite{Inselberg}

Visualize up to \(~two dozen dimensions at once
1. Draw parallel axes for each variable
2. For each tuple, connect points on each axis
Between adjacent axes: line crossings imply neg. correlation, shared slopes imply pos. correlation.
Full plot can be cluttered. Interactive selection can be used to assess multivariate relationships.
Highly sensitive to axis scale and ordering.
Expertise required to use effectively!
Radar Plot / Star Graph

Parallel dimensions in polar coordinate space
Best if same units apply to each axis

Antibiotics MIC Concentrations

Gram Staining Negative
Salmonella typhi
Salmonella schottmuelleri
Pseudomonas aeruginosa
Proteus vulgaris

Gram Staining Positive
Bacillus anthracis
Streptococcus viridans
Streptococcus pyogenes
Streptococcus pneumoniae
Staphylococcus aureus
Staphylococcus albus

[Loren Yu, CS448B 2009]
Dimensionality Reduction
Dimensionality Reduction (DR)

Project nD data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).
Reduction Techniques

LINEAR - PRESERVE GLOBAL STRUCTURE
Principal Components Analysis (PCA)
Linear transformation of basis vectors, ordered by amount of data variance they explain.

NON-LINEAR - PRESERVE LOCAL TOPOLOGY
t-Dist. Stochastic Neighbor Embedding (t-SNE)
Probabilistically model distance, optimize positions.

Uniform Manifold Approx. & Projection (UMAP)
Identify local manifolds, then stitch them together.
Principal Components Analysis

1. Mean-center the data.
2. Find \( \perp \) basis vectors that maximize the data variance.
3. Plot the data using the top vectors.
Principal Components Analysis

Linear transform: scale and rotate original space.

Lines (vectors) project to lines.

Preserves global distances.
PCA of Genomes [Demiralp et al. '13]
Non-Linear Techniques

Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:
- **t-SNE** - probabilistic interpretation of distance
- **UMAP** - tries to balance local/global trade-off
Visualizing t-SNE [Wattenberg et al. ‘16]

Results can be highly sensitive to the algorithm parameters! Are you seeing real structures, or algorithmic hallucinations?
How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.
MT Embedding [Johnson et al. 2018]

The stratosphere extends from about 10km to about 50km in altitude.

The t-SNE projection of latent space of language translation model.
Time Curves [Bach et al. ’16]

Timeline:
1 2 3 4 5 6 7
Circles are data cases with a time stamp. Similar colors indicate similar data cases.

Folding:
1
2
3
4
5
6
7

Time curve:
1
2
3
4
5
6
7
The temporal ordering of data cases is preserved. Spatial proximity now indicates similarity.

(a) Folding time

Wikipedia “Chocolate” Article

U.S. Precipitation over 1 Year
Visual Encoding Design

Use **expressive** and **effective** encodings

**Reduce** the problem space

Avoid **over-encoding**

Use **space** and **small multiples** intelligently

Use **interaction** to generate *relevant* views

Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is critical!
About the design process...

Visualization draws upon both science and art! Principles like expressiveness & effectiveness are not hard-and-fast rules, but can assist us to guide the process and articulate alternatives. They can lead us to think more deeply about our design rationale and prompt us to reflect. It helps to know “the rules” in order to wisely bend (or break) them at the right times!