CSE 442 - Data Visualization

Visual Encoding Design

Leilani Battle  University of Washington
Lecture Overview

Learning Goals
How do we apply existing encoding principles to univariate, bivariate, and multivariate data?

Topics
Exploring the Visual Design Space
Encoding Effectiveness
Scales & Axes
Multidimensional Data
A Design Space of Visual Encodings
Assign **data fields** (e.g., with $N$, $O$, $Q$ types) to **visual channels** ($x$, $y$, *color*, *shape*, *size*, ...) for a chosen **graphical mark** type (*point*, *bar*, *line*, ...).

Additional concerns include choosing appropriate **encoding parameters** (*log scale*, *sorting*, ...) and **data transformations** (*bin*, *group*, *aggregate*, ...). These options define a large combinatorial space, containing both useful and questionable charts!
Expressive?

Raw

Aggregate (Count)

Origin
Europe
Japan
USA

COUNT

Origin
Europe
Japan
USA

COUNT

Origin
Europe
Japan
USA

COUNT

Origin
Europe
Japan
USA

COUNT

Origin
Europe
Japan
USA

COUNT

Origin
Europe
Japan
USA

COUNT
1D: Quantitative

Raw

Aggregate (Count)
Expressive?

Raw

Aggregate (Count)

Miles_per_Gallon

COUNT

BIN(Miles_per_Gallon)

COUNT

BIN(Miles_per_Gallon)
Effective?

**Raw**

- Miles_per_Gallon
  - Raw data distribution

**Aggregate (Count)**

- Miles_per_Gallon
  - Distribution with count bars
  - Count values: 20, 40, 60, 80

- BIN(Miles_per_Gallon)
  - Count distribution (9, 47, 98)

- Effective status: Green check mark
A histogram subdivides a numerical range into bins, and counts the number of data points within each segment. This chart provides a discrete estimate of the probability density function.
Violin Plot Example

A violin plot visualizes a distribution of quantitative values as a continuous approximation of the probability density function, computed using kernel density estimation (KDE). The densities are additionally annotated with the median value and interquartile range, shown as black lines. Violin plots can be more informative than classical box plots.
2D: Nominal x Nominal

Raw

Aggregate (Count)

Origin

Europe

Japan

USA

Cylinders

Europe

Japan

USA

Cylinders

Cylinders

Cylinders

Cylinders

COUNT

3

108

COUNT

20

40

60

80

100

COUNT

3

4

5

6

8

20

40

60

80

100
2D: Quantitative x Quantitative

Raw

Aggregate (Count)
2D: Nominal x Quantitative

Raw

Aggregate (Mean)
Treemap

Bubble Chart

Beeswarm Plot

Origin:
- Europe
- Japan
- USA
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?
Encoding Effectiveness
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
### Effectiveness Rankings

**Quantitative**
- Position
- Length
- Angle
- Slope
- Area (Size)
- Volume
- Density (Value)
- Color Sat
- Color Hue
- Texture
- Connection
- Area (Size)
- Volume
- 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]
Effectiveness Rankings  [Mackinlay 86]

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
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

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

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
Gene Expression Time-Series [Meyer et al '11]

Color Encoding

Position Encoding
Artery Visualization [Borkin et al ‘11]

Rainbow Palette | Diverging Palette

2D

3D

pollev.com/leibatt
Artery Visualization [Borkin et al ‘11]

Rainbow Palette

2D

Shear Stress (Pa)

62%

39%

3D

Diverging Palette

92%

71%
# 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>★ Color Sat</td>
<td>Containment</td>
<td>Color Sat</td>
</tr>
<tr>
<td>★ Color Hue</td>
<td>Length</td>
<td>Shape</td>
</tr>
<tr>
<td>Texture</td>
<td>Angle</td>
<td>Length</td>
</tr>
<tr>
<td>Connection</td>
<td>Slope</td>
<td>Angle</td>
</tr>
<tr>
<td>Containment</td>
<td>Area (Size)</td>
<td>Slope</td>
</tr>
<tr>
<td>Shape</td>
<td>Volume</td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
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

[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 vs. 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?
Break Time!
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]
Scatterplots for pairwise comparison of each data dimension.

Scatterplot Matrix (SPLOM)
Scatterplot Matrix (SPLOM)

Scatter plots for pairwise comparison of each data dimension.
Parallel Coordinates
Parallel Coordinates [Inselberg]
Parallel Coordinates [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

[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).
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]
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.
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!
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.
t-SNE projection of latent space of language translation model.

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

- ENGLISH
- KOREAN
- JAPANESE

성층권은 고도 약 10km부터 약 50km까지 확장됩니다.

成層圏は、高度 10km から 50km の範囲にあります.
Time Curves  [Bach et al. ’16]

Timeline:

1  2  3  4  5  6  7

Time difference

Circles are data cases with a time stamp.
Similar colors indicate similar data cases.

Folding:

1  2  3  4  5  6  7

Wikipedia “Chocolate” Article

U.S. Precipitation over 1 Year

(a) Folding time

The temporal ordering of data cases is preserved.
Spatial proximity now indicates similarity.
Summary: 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!
Administrivia
A1: Expository Visualization

Pick a **guiding question**, use it to title your vis. Design a **static visualization** for that question. You are free to **use any tools** (inc. pen & paper).

**Deliverables** (submit on Gradescope; see A1 page)
- Image of your visualization (PNG or JPG format)
- Short description + design rationale (≤ 4 paragraphs)

Due by **11:59 pm, Wed Jan 11**.
Tableau Tutorial  (Optional)

Friday Jan 13, 4:30-6pm, Gates G20
Led by Erin and Sonia
Zoom link will be available on Canvas
Session will be recorded
I Like… / I Wish… / What If?

I LIKE…
Praise for design ideas and/or well-executed implementation details. Example: "I like the navigation through time via the slider; the patterns observed as one moves forward are compelling!"

I WISH…
Constructive statements on how the design might be improved or further refined. Example: "I wish moving the slider caused the visualization to update immediately, rather than the current lag."

WHAT IF?
Suggest alternative design directions, or even wacky half-baked ideas. Example: "What if we got rid of the slider and enabled direct manipulation navigation by dragging data points directly?"