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*, *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!
1D: Nominal
1D: Nominal
1D: Nominal
1D: Nominal
1D: Nominal
1D: Nominal
1D: Nominal
1D: Nominal

Raw

Aggregate (Count)
1D: Nominal

Raw

Aggregate (Count)

Origin

Europe
Japan
USA

Count

Europe
Japan
USA
1D: Nominal

Raw

Aggregate (Count)
1D: Nominal

Raw

Aggregate (Count)
1D: Nominal

**Raw**

**Aggregate (Count)**

- Europe
- Japan
- USA
1D: Nominal

**Raw**

- Europe
- Japan
- USA

**Aggregate (Count)**

- Europe
- Japan
- USA

- Europe: Count
- Japan: Count
- USA: Count
Expressive?

Raw

Aggregate (Count)
1D: Quantitative
1D: Quantitative

Raw
1D: Quantitative

Raw

![Bar Chart: Miles per Gallon](Miles_per_Gallon.png)
1D: Quantitative

Raw

![Miles per Gallon Chart]
1D: Quantitative

Raw

Miles_per_Gallon

Miles_per_Gallon

Miles_per_Gallon
1D: Quantitative

Raw
1D: Quantitative

Raw

- Miles_per_Gallon

- Miles_per_Gallon

- Miles_per_Gallon

- Miles_per_Gallon
1D: Quantitative

Raw

Miles_per_Gallon

Miles_per_Gallon

Miles_per_Gallon

Miles_per_Gallon
1D: Quantitative

Raw

Aggregate (Count)
1D: Quantitative

Raw

Aggregate (Count)
1D: Quantitative

Raw

Aggregate (Count)
1D: Quantitative

Raw

Aggregate (Count)
Expressive?

Raw

Aggregate (Count)
Treemap
Treemap

Bubble Chart

Aggregate (Distributions)

middle 50%
(inter-quartile range)

median

Box Plot
2D: Nominal x Nominal
2D: Nominal x Nominal

Raw
2D: Nominal x Nominal

Raw

Europe
Japan
USA

Cylinders
2D: Nominal x Nominal

Raw

<table>
<thead>
<tr>
<th>Origin</th>
<th>Cylinders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>3</td>
</tr>
<tr>
<td>Japan</td>
<td>4, 5</td>
</tr>
<tr>
<td>USA</td>
<td>6, 8</td>
</tr>
</tbody>
</table>
2D: Nominal x Nominal

Raw
2D: Nominal x Nominal

Raw

Aggregate (Count)
2D: Nominal x Nominal

Raw

<table>
<thead>
<tr>
<th>Origin</th>
<th>Cylinders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>3</td>
</tr>
<tr>
<td>Japan</td>
<td>4</td>
</tr>
<tr>
<td>USA</td>
<td>6, 8</td>
</tr>
</tbody>
</table>

Aggregate (Count)

<table>
<thead>
<tr>
<th>Origin</th>
<th>Cylinders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>3, 6</td>
</tr>
<tr>
<td>Japan</td>
<td>4, 8</td>
</tr>
<tr>
<td>USA</td>
<td>5</td>
</tr>
</tbody>
</table>

COUNT

3  108
2D: Nominal x Nominal

**Raw**

- Europe: Cylinders 3, 4, 6, 8
- Japan: Cylinders 4, 5, 6, 8
- USA: Cylinders 3, 4, 5, 6, 8

**Aggregate (Count)**

- Europe: Cylinders 3, 4, 5, 6, 8
- Japan: Cylinders 4, 5, 6, 8
- USA: Cylinders 3, 4, 5, 6, 8

COUNT: 3, 108
2D: Nominal x Nominal

**Raw**

- **Origin**: Europe, Japan, USA
- **Cylinders**: O, 3, 4, 5, 6, 8

**Aggregate (Count)**

- **Origin**: Europe, Japan, USA
- **Cylinders**: O, 3, 4, 5, 6, 8
- **COUNT**: 3, 108, 20, 40, 60, 80, 100
2D: Quantitative x Quantitative
2D: Quantitative x Quantitative

Raw
2D: Quantitative x Quantitative

Raw

![Graph showing the relationship between Horsepower and Miles_per_Gallon. The points are scattered throughout the graph, indicating a correlation between the two variables.](image-url)
2D: Quantitative x Quantitative

Raw
2D: Quantitative x Quantitative

Raw
2D: Quantitative x Quantitative

Raw

Aggregate (Count)
2D: Quantitative x Quantitative

Raw

Aggregate (Count)
2D: Nominal x Quantitative
2D: Nominal x Quantitative

Raw
2D: Nominal x Quantitative

Raw
2D: Nominal x Quantitative

Raw

<table>
<thead>
<tr>
<th>Origin</th>
<th>Miles_per_Gallon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td></td>
</tr>
</tbody>
</table>
2D: Nominal x Quantitative

Raw
2D: Nominal x Quantitative

Raw
2D: Nominal x Quantitative

Raw

Aggregate (Mean)
2D: Nominal x Quantitative

**Raw**

- **Europe**
- **Japan**
- **USA**

**Aggregate (Mean)**

- **Europe**
- **Japan**
- **USA**
2D: Nominal x Quantitative

Raw

Aggregate (Mean)
Treemap

Raw (with Layout Algorithm)
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

[McKinlay 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
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
Effectiveness Rankings

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
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]
Area Encoding
<table>
<thead>
<tr>
<th></th>
<th>QUANTITATIVE</th>
<th>ORDINAL</th>
<th>NOMINAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position</td>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Density (Value)</td>
<td>Color Hue</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Color Sat</td>
<td>Texture</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Color Hue</td>
<td>Connection</td>
<td>Connection</td>
</tr>
<tr>
<td><strong>Area (Size)</strong></td>
<td></td>
<td>Containment</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td></td>
<td>Length</td>
<td>Length</td>
</tr>
<tr>
<td><strong>Density (Value)</strong></td>
<td></td>
<td>Angle</td>
<td>Angle</td>
</tr>
<tr>
<td>Color Sat</td>
<td></td>
<td>Slope</td>
<td>Slope</td>
</tr>
<tr>
<td>Color Hue</td>
<td></td>
<td>Area (Size)</td>
<td>Area</td>
</tr>
<tr>
<td>Texture</td>
<td></td>
<td>Volume</td>
<td>Volume</td>
</tr>
<tr>
<td>Connection</td>
<td></td>
<td>Shape</td>
<td>Shape</td>
</tr>
<tr>
<td>Containment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shape</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Gene Expression Time-Series [Meyer et al '11]
Gene Expression Time-Series [Meyer et al ʼ11]

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

Color Encoding

Position Encoding
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
Volume
Length
Angle
Slope
Area (Size)
Volume
Shape

**NOMINAL**

Position
Color Hue
Texture
Connection
Containment
Density (Value)
Color Sat
Shape
Length
Angle
Slope
Area
Volume
Artery Visualization

[Borkin et al '11]

Rainbow Palette

Diverging Palette

2D

Shear Stress (Pa)

3D

Shear Stress (Pa)
Artery Visualization [Borkin et al ’11]

Rainbow Palette

2D

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><strong>Density (Value)</strong></td>
<td><strong>Containment</strong></td>
<td><strong>Color Sat</strong></td>
</tr>
<tr>
<td><strong>Color Sat</strong></td>
<td>Length</td>
<td><strong>Shape</strong></td>
</tr>
<tr>
<td><strong>Color Hue</strong></td>
<td>Angle</td>
<td><strong>Length</strong></td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td><strong>Angle</strong></td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td><strong>Slope</strong></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>
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?
Include Zero in Axis Scale?

Violates Expressiveness Principle!
Include Zero in Axis Scale?

Compare Proportions (Q-Ratio)

Violates Expressiveness Principle!
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

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

Linear Scale

Log Scale
**Linear Scale vs. Log Scale**

**Linear Scale**
- Absolute change

**Log Scale**
- Small fluctuations
- Percent change
  \[ d(10,20) = d(30,60) \]
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 absolute value.

Constraint: positive, non-zero values
Constraint: audience familiarity?
Regression Lines
By Eye ...
Linear regression ...

[The Elements of Graphing Data. Cleveland 94]
Linear regression w/out outlier ...
Linear regression w/out outlier ...
Transforming Data

How well does the curve fit the data?

[Cleveland 85]
Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

[Cleveland 85]
Multiple Plotting Options

Plot model in data space

Plot data in model space

[Cleveland 85]
Administrivia
A2: Exploratory Data Analysis

Use visualization software to form & answer questions

First steps:
Step 1: Pick domain & data
Step 2: Pose questions
Step 3: Profile the data
Iterate as needed

Create visualizations
Interact with data
Refine your questions

Author a report
Screenshots of most insightful views (10+)
Include titles and captions for each view

Due by 11:59pm Tuesday, Oct 16
Multidimensional Data
Visual Encoding Variables

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

~8 dimensions?
Example: Coffee Sales

Sales figures for a fictional coffee chain

<table>
<thead>
<tr>
<th>Sales</th>
<th>Q-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profit</td>
<td>Q-Ratio</td>
</tr>
<tr>
<td>Marketing</td>
<td>Q-Ratio</td>
</tr>
<tr>
<td>Product Type</td>
<td>N {Coffee, Espresso, Herbal Tea, Tea}</td>
</tr>
<tr>
<td>Market</td>
<td>N {Central, East, South, West}</td>
</tr>
</tbody>
</table>
Encode “Sales” (Q) and “Profit” (Q) using Position
Encode “Product Type” (N) using Hue
Encode "Market" (N) using Shape
Encode “Marketing” (Q) using Size
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.
Multiple Coordinated Views
Multiple Coordinated Views

select high salaries
Multiple Coordinated Views

- **Years**
  - How long in majors

- **Log(1+Salary)**
  - Select high salaries

- **Assists - PutO**

- **CHits/Years - C**

- **Position**
  - 3B, C, 2B, SS, CF, 1B, RF, OF, LF, DH, UT
Multiple Coordinated Views

- **how long in majors**
- **select high salaries**
- **avg assists vs avg putouts (fielding ability)**
Multiple Coordinated Views

- how long in majors
- select high salaries
- avg assists vs avg putouts (fielding ability)
- avg career HRs vs avg career hits (batting ability)
Multiple Coordinated Views

- how long in majors
- select high salaries
- avg assists vs avg putouts (fielding ability)
- avg career HRs vs avg career hits (batting ability)
- distribution of positions played
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
Dimensionality Reduction
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.
PCA of Genomes  [Demiralp et al. '13]
Many Reduction Techniques!

**General Strategies:**
Matrix Factorization
Nearest Neighbor (Topological) Methods

**Popular Techniques:**
Principal Components Analysis (PCA)
t-Dist. Stochastic Neighbor Embedding (t-SNE)
Uniform Manifold Approx. & Projection (UMAP)
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.
Visualizing t-SNE [Wattenberg et al. '16]
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

Time curve:

1  4  5  6  7

Similarity

The temporal ordering of data cases is preserved. Spatial proximity now indicates similarity.

(a) Folding time
Time Curves [Bach et al. ‘16]

Timeline:

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

Folding:

The temporal ordering of data cases is preserved. Spatial proximity now indicates similarity.

(a) Folding time

Wikipedia “Chocolate” Article
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

Wikipedia “Chocolate” Article

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

U.S. Precipitation over 1 Year
Visual Encoding Design

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

Avoid **over-encoding**

**Reduce** the problem space

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!