CSE 442 - Data Visualization

Visual Encoding Design

Matthew Conlen  University of Washington
Re-Design Exercise
Re-Design Exercise

Task: Analyze and Re-design visualization
Identify data variables (N/O/Q) and encodings
Critique the design: what works, what doesn’t
Sketch a re-design to improve communication
Be ready to share your thoughts with the class

Break into groups with those sitting near you (~4 people per group)
<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>
Teacher Salaries: Is It Really That Bad?

National and State averages for K-12 Public-School Teachers

UNITED STATES

<table>
<thead>
<tr>
<th>State</th>
<th>Average Salary</th>
<th>Average Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>NJ</td>
<td>$56,850</td>
<td>$57,300</td>
</tr>
<tr>
<td>NY</td>
<td>$56,200</td>
<td>$53,500</td>
</tr>
<tr>
<td>MI</td>
<td>$46,700</td>
<td>$46,400</td>
</tr>
<tr>
<td>IL</td>
<td>$46,300</td>
<td>$40,900</td>
</tr>
<tr>
<td>MA</td>
<td>$30,600</td>
<td>$34,400</td>
</tr>
<tr>
<td>RI</td>
<td>$59,800</td>
<td>$52,700</td>
</tr>
<tr>
<td>PA</td>
<td>$65,500</td>
<td>$66,500</td>
</tr>
</tbody>
</table>

HOURSLY

- Hours per week on-site: 30.5
- Public-School Teacher: $34.06
- Average Worker: $25.06
- Police: $22.64
- Fire: $17.91

CANADA

<table>
<thead>
<tr>
<th>State</th>
<th>Average Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>ON</td>
<td>$54,900</td>
</tr>
<tr>
<td>QC</td>
<td>$54,300</td>
</tr>
<tr>
<td>AB</td>
<td>$52,100</td>
</tr>
<tr>
<td>MB</td>
<td>$51,700</td>
</tr>
<tr>
<td>SK</td>
<td>$51,300</td>
</tr>
<tr>
<td>NB</td>
<td>$49,200</td>
</tr>
<tr>
<td>NS</td>
<td>$48,700</td>
</tr>
<tr>
<td>NL</td>
<td>$47,700</td>
</tr>
</tbody>
</table>

HOURSLY: $30.18

- Hours per week on-site: 55.6


AVERAGE Workers' salaries used for comparison are those of white-collar, non-sales employees.

Source: Good Magazine
1982: Ronald Reagan issues an order that greatly strengthens agencies' ability to classify documents. The order is commonly paraphrased inside and outside the executive branch as "When in doubt, classify."

1985: Classification reaches heights that alarm its official overseers, who step in to impose clear rules about what can and cannot be classified.

1989–1992: With the end of the Cold War, classification activity falls off, but declassification remains relatively rare.

1994: The National Reconnaissance Office and the CIA recognize that technology privately available even outside the U.S. has made images from the first-generation "Keyhole" satellites obsolete. Millions of pages of aerial images and image interpretation fall into the public record in an effort led by Vice President Al Gore, who advertises their value in climate change research.

1995: An order by Bill Clinton sets standards for timely declassification and produces a wave of historical-document declassification, totaling more than a billion pages in the next decade. Clinton officials emphasize transparency: "When in doubt, let it out" becomes a mantra for classification authorities.

2001–present: Declassification stagnates, and classification jumps. Attorney General John Ashcroft issues a memo urging government agencies to resist declassification requests whenever legally possible.

2004: The Office of the Vice President bars classification monitors from conducting their regular on-site inspection. Classification costs reach $7 billion per year.

Source: The Atlantic 300 no. 2 (September 2007)
Number of Classified U.S. Documents
Washington Dulles Airport Map
Source: United Airlines *Hemispheres*
Silver, Mark. "High School Give-and-Take."
IT WAS A VERY GOOD YEAR?

Robert Parker’s ratings for vintages of Napa Valley cabernet sauvignon

<table>
<thead>
<tr>
<th>Year</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>90T</td>
</tr>
<tr>
<td>2004</td>
<td>91R</td>
</tr>
<tr>
<td>2003</td>
<td>92I</td>
</tr>
<tr>
<td>2002</td>
<td>95E</td>
</tr>
<tr>
<td>2001</td>
<td>96T</td>
</tr>
<tr>
<td>2000</td>
<td>78C</td>
</tr>
<tr>
<td>1999</td>
<td>88T</td>
</tr>
<tr>
<td>1998</td>
<td>85R</td>
</tr>
<tr>
<td>1997</td>
<td>94I</td>
</tr>
<tr>
<td>1996</td>
<td>90T</td>
</tr>
<tr>
<td>1995</td>
<td>94T</td>
</tr>
<tr>
<td>1994</td>
<td>95E</td>
</tr>
<tr>
<td>1993</td>
<td>90E</td>
</tr>
</tbody>
</table>

2001 points. It was a relatively modest year in terms of yield from the vineyards, and that worked to the vintner’s advantage. The results: some of Napa’s most concentrated, structured, long-lived wines. Built for aging, they are rich, densely colored, fruity, and intense. They are beginning to soften, but the best are only beginning to peak, with 2001 and 2003 standing out.

Source: Business Week, June 18, 2007
Preparing for a Pandemic

Source: *Scientific American*, 293(5). November, 2005, p. 50
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

Raw

Aggregate (Count)
Expressive?

Raw

Aggregate (Count)
1D: Quantitative

Raw

Aggregate (Count)
Expressive?

Raw

Aggregate (Count)
Raw (with Layout Algorithm)

Treemap

Bubble Chart

Aggregate (Distributions)

middle 50%
(inter-quartile range)

low

median

high

Box Plot

Violin Plot
2D: Nominal x Nominal

Raw

Aggregate (Count)
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?

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

[Mackinlay 86]

<table>
<thead>
<tr>
<th><strong>QUANTITATIVE</strong></th>
<th><strong>ORDINAL</strong></th>
<th><strong>NOMINAL</strong></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></td>
</tr>
<tr>
<td>Color Sat</td>
<td>Length</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Angle</td>
<td>Shape</td>
</tr>
<tr>
<td>Texture</td>
<td>Slope</td>
<td>Length</td>
</tr>
<tr>
<td>Connection</td>
<td>Area (Size)</td>
<td>Angle</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Slope</td>
</tr>
<tr>
<td>Shape</td>
<td></td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td></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
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  
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
# 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>Angle</td>
</tr>
<tr>
<td>Containment</td>
<td>Volume</td>
<td>Slope</td>
</tr>
<tr>
<td>Shape</td>
<td>Shape</td>
<td>Area</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Volume</td>
</tr>
</tbody>
</table>
Artery Visualization [Borkin et al '11]

Rainbow Palette

2D: 62%

3D: 39%

Diverging Palette

2D: 92%

3D: 71%
Effectiveness Rankings

**QUANTITATIVE**
- Position
- Length
- Angle
- Slope
- Area (Size)
- Volume

**Density (Value)**
- Color Sat
- Color Hue
- Texture

**Connection**
- Containment
- Length
- Angle
- Slope
- Area (Size)
- Volume

**SHAPE**
- Shape

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

**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

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
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
Linear regression ...
Linear regression w/out outlier ...

[The Elements of Graphing Data. Cleveland 94]
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
Images of annotated visualizations
(8+ images; min 4 views of the data)
Include titles and labels for each view

Due by 11:59pm Tuesday, Jan 28
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

Sales Q-Ratio
Profit Q-Ratio
Marketing Q-Ratio
Product Type N \{Coffee, Espresso, Herbal Tea, Tea\}
Market N \{Central, East, South, West\}
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

- select high salaries
- how long in majors
- avg assists vs avg putouts (fielding ability)
- avg career HRs vs avg career hits (batting ability)
- distribution of positions played

[Detailed diagrams and visualizations as shown in the image]
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!
“Parallel” dimensions in polar coordinate space
Best if same units apply to each axis
Dimensionality Reduction
Dimensionality Reduction

http://www.ggobi.org/
Principal Components Analysis

1. Mean-center the data.

2. Find 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)
The Beginner's Guide to Dimensionality Reduction

Explore the methods that data scientists use to visualize high-dimensional data.

By: Matthew Conlen and Fred Hohman
July 16, 2018

Dimensionality reduction is a powerful technique used by data scientists to look for hidden structure in data. The method is useful in a number of domains, for example document categorization, protein disorder prediction, and machine learning model debugging.²

The results of a dimensionality reduction algorithm can be visualized to reveal patterns and clusters of similar or dissimilar data. Even though the data is displayed in only two or three dimensions, structures present in higher dimensions are maintained, at least roughly.⁷

The technique is available in many applications, for...
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:

(a) Folding time

U.S. Precipitation over 1 Year

Wikipedia “Chocolate” Article

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