Learning Goals

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

COUNT

<table>
<thead>
<tr>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
</tbody>
</table>

!?
Raw (with Layout Algorithm)

Treemap

Aggregate (Distributions)

interquartile range (middle 50%)

low median high

Box Plot

Violin Plot
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>5</td>
</tr>
</tbody>
</table>

Aggregate (Count)

<table>
<thead>
<tr>
<th>Origin</th>
<th>Cylinders</th>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Japan</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>USA</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>100</td>
</tr>
</tbody>
</table>
2D: Quantitative x Quantitative

Raw

Aggregate (Count)
2D: Nominal x Quantitative

Raw

Aggregate (Mean)

Origin

Europe
Japan
USA

Miles_per_Gallon

Origin

Europe
Japan
USA

MEAN(Miles_per_Gallon)

Origin

Europe
Japan
USA

MEAN(Miles_per_Gallon)
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?
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
<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>
<tr>
<td>QUANTITATIVE</td>
<td>ORDINAL</td>
<td>NOMINAL</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------</td>
<td>----------------------------------</td>
</tr>
<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>Length</td>
<td>Color Sat</td>
</tr>
<tr>
<td>Color Sat</td>
<td>Angle</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Slope</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Area (Size)</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Volume</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Shape</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td></td>
<td>Volume</td>
</tr>
</tbody>
</table>
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
- Shape
# 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

39%

3D

Shear Stress (Pa)

3
2
1
0

Diverging Palette

2D

62%

3D

92%

71%
Effectiveness Rankings

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

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 (Size)
Volume
Shape
Volume
Scales & Axes
Include Zero in Axis Scale?

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

Yearly CO₂ 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 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?
Aspect Ratio
(width : height)
Banking to 45° [Cleveland]

To facilitate perception of trends, maximize the discriminability of line segment orientations

Two line segments are maximally discriminable when their average absolute angle is 45°
Method: optimize the aspect ratio such that the average absolute angle of all segments is 45°
Alternative: Minimize Arc Length while holding area constant [Talbot et al. 2011]
A Good Compromise

Arc-length banking produces aspect ratios in-between those produced by other methods.

[Talbot et al. 2011]
Trends may occur at different scales!
Apply banking to the original data or to fitted trend lines.
[Heer & Agrawala '06]
Administrivia
Migrating to Gradescope

Students will now submit assignments (A1, A2, etc.) through Gradescope instead of Canvas.

If you submitted A1 through Canvas, we will migrate your submission to Gradescope for you.

Please let us know asap if you run into any issues with Gradescope!
Tableau Tutorial (Optional)

Friday April 8, 1-2pm
Led by Nussara and Chandler
Zoom link available on Canvas
Session will be recorded
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 Gradescope.

Due by **Wed 1/26 11:59pm.**
On Thursday 4/21 you will be assigned two peer A2 submissions to review. For each:

- Try to determine which is earnest and which is deceptive
- Share a rationale for how you made this determination
- Share feedback using the “I Like / I Wish / What If” rubric

Assigned reviews will be posted on the A2 Peer Review page on Canvas, along with a link to a Google Form. You should submit two forms: one for each A2 peer review.

Due by **Fri 4/29 11:59pm**.
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?"
Break Time!
Multidimensional Data
Visual Encoding Variables

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

~8 dimensions?
Example: Coffee Sales

Sales figures for a fictional coffee chain

Sales
Profit
Marketing
Product Type
Tea
Market

Q-Ratio
Q-Ratio
Q-Ratio
N \{Coffee, Espresso, Herbal Tea, Tea\}
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]
Scatterplot Matrix (SPLOM)

Scatter plots for pairwise comparison of each data dimension.
Multiple Coordinated Views

- **how long in majors**
  - Bar chart showing distribution of years in the majors.
- **select high salaries**
  - Logarithmic plot showing salary distribution.
- **avg assists vs avg putouts (fielding ability)**
  - Scatter plot showing assists vs putouts.
- **avg career HRs vs avg career hits (batting ability)**
  - Scatter plot showing career hits vs career HRs.
- **distribution of positions played**
  - Bar chart showing distribution of positions played.
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 graph]

[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).
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
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.
1. Model probability $P$ of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
2. Define a similar probability \( Q \) in the low-dimensional (2D or 3D) embedding space, using a Student’s \( t \) distribution (hence the “\( t \)-“ in “\( t \)-SNE”!). The \( t \)-distribution is heavy-tailed, allowing distant points to be even further apart.
1. Model probability $P$ of one point “choosing” another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

2. Define a similar probability $Q$ in the low-dimensional (2D or 3D) embedding space, using a Student’s $t$ distribution (hence the “$t$-” in “t-SNE”!). The $t$-distribution is heavy-tailed, allowing distant points to be even further apart.

3. Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the $P$ and $Q$ distributions: $KL(P \parallel Q)$
t-SNE projection of photos taken in Paris, France
t-SNE projection of latent space of language translation model.

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

KOREAN
성층권은 고도 약 10km부터 약 50km까지 확장됩니다.

JAPANESE
成層圏は、高度 10km から 50km の範囲にあります.
Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

“Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance.” - McInnes et al. 2018
Figure 1: Variation of UMAP hyperparameters $n$ and min-dist result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of $n$ spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.
Reader Behavior [Conlen et al. 2019]

UMAP projection of reader activity for an interactive article.

“Tentacles” map to activity archetypes, “blob” body maps to sessions that blend behaviors.
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!