Today

- **Before class**
  - Reading: Ch 1 - What's Vis, and Why Do It?

- **During class**
  - Highlight and discuss Ch 1
  - *We will not cover everything that you are responsible for during class time.*
Topic Objectives

- Define visualization.
- Explain the importance of humans in the visualization process.
- Explain why human vision is particularly well-suited for information transfer.
- Give an example of a visualization idiom.
- Explain why it is best to consider multiple alternatives for vis before selecting a solution.
- Explain at a high-level the "what-why-how" framework for analyzing visualization use.
- Differentiate between R, D3, and Tableau and describe the type of tasks for which each tool might be most appropriate.

What is visualization?

- "The communication of information using graphical representations"
  - Ward, Grinstein, Keim

- "The use of computer-supported interactive visual representations of data to amplify cognition"
  - Card, Mackinlay, Shneiderman, *Readings in Information Visualization: Using Vision to Think*

- "The purpose of visualization is insight, not pictures."
  - Ben Shneiderman
What's vis?

- Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.
- The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations.
- Vis design is full of tradeoffs, and most possibilities in the design space are ineffective for a particular task, so validating the effectiveness of a design is both necessary and difficult.
- Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays.
- Vis usage can be analyzed in terms of why the user needs it, what data is shown, and how the idiom is designed.
Why have a human in the loop?

- Vis allows people to analyze data when they don't know exactly what questions to ask in advance.

- Best path - put a human in the loop
  - exploit the pattern detection properties of human vision

Humans are great at pattern recognition

Create visualizations that lets computers do what computers do well and lets humans do what humans do well.
Uses of vis tools

- Transitional
  - vis works itself out of a job

- Long-term
  - exploratory analysis

- Presentation
  - visual explanations

Why have a computer in the loop?

Munzner, Fig 1.2 (Barsky et al., 2007)
Why use an external representation?

- Vis allows people to offload cognition and memory usage to make space for other operations.

- Diagrams as external representations
  - information can be organized by spatial location
    - search - grouping items needed for problem-solving in one location
    - recognition - grouping relevant info for one item in the same location

Visualization can extend your memory

<table>
<thead>
<tr>
<th>What is 57 x 48?</th>
<th>paper</th>
<th>mental buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>[8 * 7 = 56]</td>
<td></td>
</tr>
<tr>
<td>x 48</td>
<td>[8 * 5 = 40 + 5 = 45]</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>[4 * 7 = 28]</td>
<td></td>
</tr>
<tr>
<td>456</td>
<td>[4 * 5 = 20 + 2 = 22]</td>
<td></td>
</tr>
<tr>
<td>228</td>
<td>[8 + 5 = 13]</td>
<td></td>
</tr>
<tr>
<td>2736</td>
<td>[4 + 2 + 1 = 7]</td>
<td></td>
</tr>
</tbody>
</table>

Example courtesy Tamara Munzner, Univ. of British Columbia
Why depend on vision?

- Visual system provides a high-bandwidth channel to our brains.
- Significant amount of visual information processing occurs in parallel at the pre-conscious level.

Can you find the red dot?

preattentive processing

http://www.csc.ncsu.edu/faculty/healey/PP/index.html
Which state had the highest marriage rate?

U.S. Census Bureau, Statistical Abstract of the United States: 2012
http://www.census.gov/compendia/statatab/2012/tabs/12s0133.pdf,
http://www.census.gov/compendia/statatab/2012/tabs/12s0133.xls
Why show the data in detail?

- Vis tools can allow people to explore data to find patterns or to determine if a statistical model actually fits the data.

- Look out for questionable data
  - "just because it's numbers doesn't mean it's true"
  - is it a typo or something interesting?
    - "make sure you know which one it is"

---

### Anscombe's Quartet

**Anscombe’s Quartet: Raw Data**

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>8.04</td>
<td>9.14</td>
<td>7.46</td>
<td>6.58</td>
</tr>
<tr>
<td>x</td>
<td>8.0</td>
<td>10.0</td>
<td>8.0</td>
<td>10.0</td>
</tr>
<tr>
<td>y</td>
<td>9.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
</tr>
<tr>
<td>x</td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>7.58</td>
<td>8.74</td>
<td>12.74</td>
<td>7.71</td>
</tr>
<tr>
<td>x</td>
<td>9.0</td>
<td>10.0</td>
<td>9.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>8.81</td>
<td>8.77</td>
<td>7.11</td>
<td>8.84</td>
</tr>
<tr>
<td>x</td>
<td>11.0</td>
<td>11.0</td>
<td>11.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>8.33</td>
<td>9.26</td>
<td>7.81</td>
<td>8.47</td>
</tr>
<tr>
<td>x</td>
<td>14.0</td>
<td>14.0</td>
<td>14.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>9.96</td>
<td>8.10</td>
<td>8.84</td>
<td>7.04</td>
</tr>
<tr>
<td>x</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>7.24</td>
<td>6.13</td>
<td>6.08</td>
<td>5.25</td>
</tr>
<tr>
<td>x</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>19.0</td>
</tr>
<tr>
<td>y</td>
<td>4.26</td>
<td>6.10</td>
<td>5.39</td>
<td>12.50</td>
</tr>
<tr>
<td>x</td>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>10.0</td>
<td>9.13</td>
<td>8.15</td>
<td>5.56</td>
</tr>
<tr>
<td>x</td>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>7.26</td>
<td>7.26</td>
<td>6.42</td>
<td>7.91</td>
</tr>
<tr>
<td>x</td>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>8.0</td>
</tr>
<tr>
<td>y</td>
<td>5.0</td>
<td>4.74</td>
<td>5.73</td>
<td>6.89</td>
</tr>
<tr>
<td>mean</td>
<td>9.0</td>
<td>7.5</td>
<td>9.0</td>
<td>7.5</td>
</tr>
<tr>
<td>var.</td>
<td>10.0</td>
<td>3.75</td>
<td>10.0</td>
<td>3.75</td>
</tr>
<tr>
<td>corr.</td>
<td>0.816</td>
<td>0.816</td>
<td>0.816</td>
<td>0.816</td>
</tr>
</tbody>
</table>


Munzner, Figure 1.3
The four data sets are not the same

"Graphics reveal data"
- Edward Tufte, *The Visual Display of Quantitative Information*

Datasaurus Dozen
https://www.autodeskresearch.com/publications/samestats
Why use interactivity?

- The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations.

- Interaction allows for
  - handling complexity
  - displaying multiple aspects of a dataset

Why is the idiom design space huge?

- The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations.

- Vis idioms - approaches to creating and manipulating visual representations

- Simple examples: scatterplots, bar charts, line charts
Visual Vocabulary

Deviation

Diverging bar
A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar
Perfect for presenting survey results which involve sentiment (e.g. disagree/neutral/agree).

Spine chart
Splits a single value into 2 contrasting components (e.g. Male/Female).

Surplus/deficit filled line
The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Scatterplot
The standard way to show the relationship between two continuous variables, each of which has its own axis.

Line + Column
A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot
Usually used to show how the relationship between 2 variables has changed over time.

Buble
Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap
A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.
Ranking Distribution

Ordered bar
Standard bar charts display the ranks of values much more easily where sorted into order.

Ordered column
See above.

Ordered proportional symbol
Use when there are big variations between values where seeing the differences between data is not so important.

Dot strip plot
Data placed in order on a strip are a space-saving method of laying out ranks across multiple categories.

Lollipop chart
Lollipop charts draw more attention to the data value than standard bar charts and can also show rank and value effectively.

Histogram
The standard way to show a statistical distribution — keep the gaps between columns small to highlight the shape of the data.

Bar plot
Similar to a box plot but more effective with complex distributions (data that cannot be summed over to simple averages).

Population pyramid
A standard way for showing the age and sex breakdown of a population. Helps us see change effectively, back to back histograms:

Dot strip plot
Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Dot plot
A simple way of showing the change in the frequency (y-axis) of data across multiple categories.

Bombe plot
Like dot strip plots, good for displaying all values in a table/row, work best when highlighting individual values.

Condense curve
A good way of showing how unequal a distribution is as it is always considerably resistant to changes in a mean.

Change Over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader.

Line
The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column
Columns work well for showing change over time — but usually best with only one series of data at a time.

Line + column
A good way of showing the relationship over time between an amount (columns) and a rate (line).

Stock price
Usually focused on day-to-day activity, these charts show opening/closing and low/high points of each day.

Slope
Good for showing changing data as long as the data can be simplified into 2 or 3 points without messing a key part of story.

Area chart
Use with care — these are good of showing changes to Y1/Y2 but swamping change in components can be very difficult.

Fan chart (projections)
Use to show the uncertainty in future projections — usually this grows the further forward to projection.

Connected scatterplot
A great way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap
A great way of showing temporal patterns (daily, weekly, monthly) — at the expense of showing precision in quantity.

Pine tree timeline
Great when data and duration are key elements of the story in the data.

Circle timeline
Good for showing discrete values of varying size across multiple categories (e.g., earthquakes by continent).

Seismogram
Another alternative to the circle timeline for showing series where there are big variations in the data.
Part-to-Whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Stacked column
A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Proportional stacked bar
A good way of showing the size and proportion of data at the same time - as long as the data are not too complicated.

Pie
A common way of showing part-to-whole data - but be aware that it’s difficult to accurately compare the size of the segments.

Donut
Similar to a pie chart - but the centre can be a good way of making space to include more information about the data (eg. total).

Tree map
Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Column
The standard way to compare the size of things. Must always start at 0 on the axis.

Bar
See above. Good when the data are not time series and labels have long category names.

Paired column
As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar
See above.

Pie chart
Lollipop chart
Lollipop charts draw more attention to the data value than standard bar/column - does not HAVE to start at zero (but preferable).

Radar chart
A space-efficient way of showing value of multiple variables - but make sure they are organized in a way that makes sense to reader.

Parallel coordinates
An alternative to radar charts - again, the arrangement of the variables is important. Usually benefits from highlighting values.
Why focus on tasks?

- Vis usage can be **analyzed in terms of why** the user needs it, **what** data is shown, **and how** the idiom is designed.

- The intended task is **just as important** as the data to be visualized.

- Four categories of tasks
  - presentation
  - discovery
  - enjoyment of information
  - producing more information for later use
Why focus on effectiveness?

- Vis design is full of **tradeoffs**, and **most possibilities in the design space are ineffective** for a particular task, so validating the **effectiveness** of a design is both necessary and difficult.

- Effectiveness is an important measure for understanding if the user task was supported.
  - "The purpose of visualization is insight, not pictures." - Ben Shneiderman

- But, no picture can tell the truth, the whole truth, and nothing but the truth.

Why are most designs ineffective?

- Vis design is full of **tradeoffs**, and **most possibilities in the design space are ineffective** for a particular task, so validating the **effectiveness** of a design is both necessary and difficult.

- Design may not match with human perception

- Design may not match with intended task
Why is validation difficult?

- Vis design is full of tradeoffs, and most possibilities in the design space are ineffective for a particular task, so validating the effectiveness of a design is both necessary and difficult.

- How do you know if your visualization "works"?
  - How do you measure insight?

- How do you argue that one design is better than another?
  - What does "better" mean? faster? more fun? more effective?
  - What does "effectively" mean?
Why are there resource limitations?

- Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays.
  - computational capacity
  - human perceptual and cognitive capacity
  - display capacity

Why analyze vis?

- Vis usage can be analyzed in terms of why the user needs it, what data is shown, and how the idiom is designed.
- Analyzing existing systems is a good stepping stone to designing new ones.
- High-level framework for analyzing vis use
  - *what* data the user sees
  - *why* the user intends to use a vis tool
  - *how* the visual encoding and interaction idioms are constructed in terms of design choices
Tools of the Trade

Arnet.js
A library of force-directed layout algorithms plus abstractions for graph, organization and network handling.

CartoDB
A web service for mapping, analyzing and building applications with data.

Chrome.js
Interactive color space editor that allows to preview a set of linear, exponential, HSV, CMYK or grayscale colors.

Circo
A software package for visualizing data in a circular layout.

Cell.js
A library for arranging networks using community-based optimization techniques.

ColorBrowser
A web tool for selecting colors for maps.

Custom.js
A library for creating interactive time series and network graphs based on D3.js.

Cytoscape
An open-source web and efficient technology to create and manipulate interactive graphs based on graphs.

D3.js
A simple, data-driven visualization framework based on D3.js and D3-force.js.

D3.js
A data representation framework providing a uniform abstraction to domain data.

Data.js
An interactive web application for data discovery and data exploration.

Degas
A library for creating interactive graphs and visualizations.

Emissary.js
A library for creating interactive timelines.

Graph
A set of software tools for creating rich interactive data visualizations in Autodesk.

GeoCommons
Academic community and set of tools to access, visualize and analyze data with competing map visualizations.

Geosprit
A visualization and exploration platform for networks with directed and hierarchical graphs.

Google Chart Tools
A collection of simple, fast, and customizable tools to create interactive charts and data visualizations.

http://selection.datavisualization.ch/
Excel

http://chandoo.org/wp/2008/09/03/6-charts-to-never-use/

http://www.juiceanalytics.com/writing/recreating-ny-times-cancer-graph/

http://www.r-project.org
Shiny

Interactive visualizations
Shiny is designed for fully interactive visualization, using JavaScript libraries like d3, Leaflet, and Google Charts.

Start simple
If you’re new to Shiny, these simple but complete applications are designed for you to study.

Tableau

http://www.tableausoftware.com
Many, many others

"What I Learned Recreating One Chart Using 24 Tools" - assigned reading

Tools for Analysis vs. Presentation


Flexibility of Tools

## Apps vs. Libraries and Static vs. Interactive

<table>
<thead>
<tr>
<th>STATIC APPS</th>
<th>WEB - INTERACTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILLUSTRATOR, NODEBOX, EXCEL, POLESTAR, RAW</td>
<td>HIGHCHARTS CLOUD, QUADRIGRAM, EASYCHRT, DATARWRAPPER, TABLEAU, PLOTLY, GOOGLE SHEETS</td>
</tr>
<tr>
<td>GGPLLOT2, MATPLOTLIB, R, SEABORN, BOKEH, PROCESSING</td>
<td>D3, D4, C3, NVD3, GGVIS, HIGHCHARTS, SHINY, VEGA, VEGA-LITE</td>
</tr>
</tbody>
</table>


---

### Topic Objectives

- Define visualization.
- Explain the importance of humans in the visualization process.
- Explain why human vision is particularly well-suited for information transfer.
- Give an example of a visualization idiom.
- Explain why it is best to consider multiple alternatives for vis before selecting a solution.
- Explain at a high-level the "what-why-how" framework for analyzing visualization use.
- Differentiate between R, D3, and Tableau and describe the type of tasks for which each tool might be most appropriate.