Data Visualization (CIS/DSC 468)

Partitioning, Filtering, & Aggregation

Dr. David Koop
Multiple Views

• Why have just one visualization?

• Sometimes data is best examined in more than one view
  - Clutter/visual overload
  - Different attributes (cannot show all attributes in one view)
  - Different scales (task requires overview or detail)
  - Different encodings (no single encoding is optimal for all tasks)

• Eyes Beat Memory (Ch. 6)
  - Aiding working memory:
    side-by-side/layers > animated > jump cuts
  - Showing all visual elements at once → don't need to remember
Strategies for Multiple Views

• Juxtaposition
• Superimposition
• Nesting
## Multiple Views

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Data</th>
<th>All</th>
<th>Subset</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Redundant</td>
<td></td>
<td>Overview/Detail</td>
<td>Small Multiples</td>
</tr>
<tr>
<td>Different</td>
<td>Multiform</td>
<td></td>
<td>Multiform, Overview/Detail</td>
<td>No Linkage</td>
</tr>
</tbody>
</table>
Multiform

[Improvise, Weaver, 2004]
Small Multiples

• Same encoding, but different data in each view (e.g. SPLOM)

[http://bl.ocks.org/mbostock/4063663]
Assignment 3

- Soccer data
  - Draw two choropleth maps
  - Draw a teammate graph using force-directed layout
  - Country lookup: Create an associative array, d3.map or Map
Exam 2

- Similar Format to Exam 1: Multiple Choice + Free Response
- Cumulative but emphasis on the topics covered since Exam 1
  - Trees
  - Geospatial Data
  - Color & Colormaps
  - Interaction
  - Multiple Views
- Reading: Chs. 8.1-8.3, 10, 6, 11, 12, 13
- Be prepared for a similar D3 question to the last exam
- More information to be made available on the course web page
Juxtaposition

[ComVis, K. Matkovic et al., 2008]
Juxtaposition Guidelines

• Benefits:
  - The component visualizations are independent and can be composed without interference
  - Easy to implement

• Drawbacks:
  - Implicit visual linking is not always easy to see, particularly when multiple objects are selected
  - Space is divided between the views, yielding less space for each view

• Applications: Use for heterogeneous datasets consisting of many different types of data, or for where different independent visualizations need to be combined.

[W. Javed and N. Elmqvist, 2012]
Integration: Juxtaposition + Links

[VisLink, Collins and Carpendale, 2007]
The "best statistical graphic ever"

(later known as a Sankey Diagram)

[Napoleon's March to Moscow, C. J. Minard, 1869]
Superimposition

Figure 6: Mapgets [38] (Superimposed Views). Presentation stack, with superimposed layers for rivers, borders, and labels, in Mapgets.

Figure 7: GeoSpace [22] (Superimposed Views). A crime data layer superimposed on a geographical map of the Cambridge, MA area.

Superimposed views overlay two or more visual spaces on top of each other (Figures 6 and 7). The resulting visualization becomes the visual combination of the component visualizations, often using transparency to enable seeing all views. Superimposed views are generally used to highlight spatial relations in the component visualizations. In other words, the spatial linking present in these views is one-to-one, i.e., all the overlay visualizations share the same underlying visual space. Line graph visualizations with several data series, where more than one graph is superimposed in a single chart (e.g., [19]), is a very commonly used example of this design pattern. The spatial linking in the superimposed views allows for easy comparison across different datasets because the user does not have to split their attention between different parts of the visual space. Furthermore, the fact that visualizations are stacked means that they can each use the full available space in the view. However, because the composition simply adds the component visualizations together, the visual clutter may become significant, and it is also likely to cause conflicts arising from one visualization occluding another.

5.1 Mapgets
Mapgets [38] is a geographic visualization system that allows users to interactively perform map editing and querying of geographical datasets. The maps generated using Mapgets are built on an underlying presentation stack that superimposes multiple dataset layers on top of each other. The users can dynamically select the dataset to use for each layer and the total number of layers to compose. Different layers in the presentation stack allow users to independently interact with each of the associated visualization and control the layer attributes. The technique also allows the users to reorder layers in the presentation stack to achieve the desirable map result. Figure 6 shows an example of a European map generated in Mapgets. The presentation stack associated with this map consists of three layers: the bottom layer visualizes rivers, the center layer is used to depict the country borders, and the topmost layer is used to display the country labels.

5.2 GeoSpace
GeoSpace [22] allows users to interactively explore complex visual spaces using superimposed views. It permits progressively overlaying different datasets, based on the user queries, in a single view. Beyond allowing users to explore datasets through dynamic queries, GeoSpace also supports pan and zoom operations for navigation. Figure 7 shows GeoSpace system being used for exploring crime around the Cambridge, MA area. The figure shows a 2D view of the visualization, where red dots that are spatially coupled to the underlying layer show the reported crime cases in the region.

Figure 8: SPPC [45] (Overloaded Views). This tool overloads points into the region bounded by two axes in the parallel coordinate plot.

Figure 9: Links on treemaps [14] (Overloaded Views). The tool identifies a tree structure in a graph and visualizes it using a treemap.
Superimposition Guidelines

• Benefits:
  - Allows direct comparison in the same visual space.

• Drawbacks:
  - May cause occlusion and high visual clutter.
  - The client visualization must share the same spatial mapping as the host visualization.

• Applications: In settings where comparison is common, or where the component visualization views need to be as large as possible (potentially the entire available space).

[W. Javed and N. Elmqvist, 2012]
Nesting: Vis inside vis (Inception)

[NodeTrix, N. Henry et al., 2007]
Nesting Guidelines

• Benefits:
  - Very compact representation
  - Easy correlation

• Drawbacks:
  - Limited space for the client visualizations
  - Clutter is high
  - Visual design dependencies are high

• Applications: Situations that call for augmenting a particular visual representation with additional mapping

[W. Javed and N. Elmqvist, 2012]
Partitioned Views

• Split dataset into groups and visualize each group
• Extremes: one item per group, one group for all items
• Can be a hierarchy
  - Order: which splits are more "related"?
  - Which attributes are used to split? usually categorical
Example: Grouped Bar Chart

[Example chart showing population by age group for different states.]

[M. Bostock, http://bl.ocks.org/mbostock/3887051]
Example: Small Multiples Bar Chart

Group 4

Group 3

Group 2

Group 1

Q108  Q208  Q308  Q408  Q109  Q209  Q309  Q409

[M. Bostock, http://bl.ocks.org/mbostock/4679202]
Matrix Alignment & Recursive Subdivision

• Matrix Alignment:
  - regions are placed in a matrix alignment
  - splits go to rows and columns
  - main-effects ordering: use summary statistic to determine order of categorical attribute

• Recursive subdivision:
  - Designed for exploration
  - Involves hierarchy
  - User drives the ways data is broken down in recursive manner
I page. In Figure 2 there are 6 panels, 1 column, 6 rows, and 1 page. Later, we will show a Trellis display with more than one page. We refer to the rectangular array as the trellis because it is reminiscent of a garden trelliswork.

Each panel of a trellis display shows a subset of the values of panel variables; these values are formed by conditioning on the values of conditioning variables.

In Figure 1 the panel variables are variety and yield, and the conditioning variables are site and year. On each panel, values of yield and variety are displayed for one combination of year.

[Becker et al., 1996]
Example: HiVE System

[Slingsby et al., 2009]
Example: HiVE System

[Slingsby et al., 2009]
Superimposed Layers

• Put different layers in the same spatial region, overlay information
• Usually each layer spans the entire view
• Must be **identifiable**: visually distinguishable
• Cartography has to deal with this a lot
• May be static or dynamic (user controls which layers are shown)
Example: Superimposed Line Charts

![Superimposed Line Charts](http://bl.ocks.org/mbostock/3884955)
Fig. 13. A software system and its associated call graph (caller = green, callee = red). (a) and (b) show the system with bundling strength $\beta = 0.85$ using a balloon layout (node labels disabled) and a radial layout, respectively. Bundling reduces visual clutter, making it easier to perceive the actual connections than when compared to the non-bundled versions (figures 2a and 11a). Bundled visualizations also show relations between sparsely connected systems more clearly (encircled regions); these are almost completely obscured in the non-bundled versions. The encircled regions highlight identical parts of the system for (a), (b), and figure 15.

Fig. 14. Using the bundling strength $\beta$ to provide a trade-off between low-level and high-level views of the adjacency relations. The value of $\beta$ increases from left-to-right; low values mainly provide low-level, node-to-node connectivity information, whereas high values provide high-level information as well by implicit visualization of adjacency edges between parent nodes that are the result of explicit adjacency edges between their respective child nodes.

The majority of the participants regarded the technique as useful for quickly gaining insight in the adjacency relations present in hierarchically organized systems. In general, the visualizations were also regarded as being aesthetically pleasing. SIG and FEI Company Eindhoven are currently supporting further development by providing us with additional data sets and feedback regarding the resulting visualizations.

More specifically, most of the participants particularly valued the fact that relations between items at low levels of the hierarchy were automatically lifted to implicit relations between items at higher levels by means of bundles. This quickly gave them an impression of the high-level connectivity information while still being able to inspect the low-level relations that were responsible for the bundles by interactively manipulating the bundling strength. This is illustrated in figure 14, which shows visualizations using different values for the bundling strength $\beta$. Low values result in visualizations that mainly provide low-level, node-to-node connectivity information. High values result in visualizations that provide high-level information as well by implicit visualization of adjacency edges between parent nodes that are the result of explicit adjacency edges between their respective child nodes.

Another aspect that was commented on was how the bundles gave an impression of the hierarchical organization of the data as well, thereby strengthening the visualization of the hierarchy. More specifically, a thick bundle shows the presence of two elements at a fairly
Reducing Complexity

- Too many items or attributes lead to visual clutter
- Interaction and Multiple Views can help, but often lose the ability to start understanding an entire dataset at first glance
- **Reduction** techniques show less data to reduce complexity
- Can reduce items or attributes (both are elements)
- **Filtering**: eliminate elements from the current view
  - "out of sight, out of mind"
- **Aggregation**: replace elements with a new element that represents the replaced elements
  - summarization is often challenging to design
- Another method is **focus+context**: show details in the context of an overview
Overview

Reducing Items and Attributes

- **Filter**
  - Items
  - Attributes

- **Aggregate**
  - Items
  - Attributes

Reduce

- **Filter**
  - Filter

- **Aggregate**
  - Aggregate

- **Embed**
  - Embed

[Munzner (ill. Maguire), 2014]
Filtering

• Just don't show certain elements
• Item filtering: most common, eliminate marks for filtered items
• Attribute filtering:
  - attributes often mapped to different channels
  - if mapped to same channel, allows many attributes (e.g. parallel coordinates, star plots), can filter
• How to specify which elements?
  - Pre-defined rules
  - User selection
Filter vs. Query

• Queries start with an empty set of items and **add** items
• Filters start with all items and **remove** items
Restaurant locations are derived from the New York City Department of Health and Mental Hygiene database. Due to the limitations of the Health Department’s database, some restaurants could not be placed.

By JEREMY WHITE

© 2013 The New York Times Company

New York Health Department Restaurant Ratings Map

The New York City Department of Health and Mental Hygiene performs unannounced sanitary inspections of every restaurant at least once per year. Violation points result in a letter grade, which can be explored in the map below, along with violation descriptions. The information on this map will be updated every two weeks. For menus and reviews by New York Times critics, visit our restaurants guide.

Related Article »

Gracie’s Cafe
Grade pending
Violation points 27
Click for details

Example: NYC Health Dept. Restaurant Ratings

[J. White, New York Times]
Dynamic Queries

• Interaction need not be with the visualization itself
• Users interact with **widgets** that control which items are shown
  - Slides, Combo boxes, Text Fields
• Often tied to attribute values
• Examples:
  - All restaurants with an "A" Grade
  - All pizza places
  - All pizza places with an "A" Grade
Scented Widgets

For each task, subjects with one of the three following task hypotheses were instructed to make at least seven observations that provided evidence either for or against the current task hypothesis. The task hypotheses were intended to be of similar depth and diversity.

Hypothesis 1: The number and variety of jobs stereotypically male jobs remain almost entirely male.

Hypothesis 2: Technology is costing jobs by making occupations obsolete.

Hypothesis 3: Women have joined the work force, but jobs have not.

We gave them an introductory tutorial to the system, and then asked subjects to complete three tasks. Participants were asked to note their observations by commenting up to two of the observations had to be unique findings on views not yet visited, leaving new comments on the corresponding views.

Social data across conditions. Subjects in the previous sense.us study used a small amount of manual seeding to balance the metrics from a study of the sense.us system [12] and supplemented it with a dynamic query widgets to show visitation patterns and likely had an effect on these correlations. We believe that the semantics of the tasks also affect how users would visit the same views that were visited in the seed data than users in conditions where they are not scented.

Next, we analyzed the data to check if scented widgets help us make unique discoveries. Our hypotheses were that scented widgets are used in the dynamic query widgets to show visitation patterns and likely had an effect on these correlations. We believe that the semantics of the tasks also affect performance would improve over subsequent trials, regardless of condition. We removed the starting overview from consideration, representing the number of visits to each view in each scenting condition. We then compared these visitation vectors to the visitation vector for the seed data, in which we used no scenting conditions on perceived utility and user experience. Because both visits (493) = 0.217 for no scent, and comments (493) = 0.200 for no scent, in all cases). These results suggest that no scent exhibited a power law distribution, and so we scaled them logarithmically for display in the scented widgets.

The study employed a 3 (Task) x 3 (Scent) between subjects design. Task and scent pairings and presentation order were counter balanced using a Latin Square. All tests were carried out in a laboratory environment using standard desktop PCs connected to a web server hosting the visualization and usage data. After completing the tasks, subjects filled out a survey that asked them to rate their experience of the system. Task and scent pairings and presentation order were counter balanced using a Latin Square. All tests were carried out in a laboratory environment using standard desktop PCs connected to a web server hosting the visualization and usage data. After completing the tasks, subjects filled out a survey that asked them to rate their experience of the system.

Figure 3: The number and variety of jobs stereotypically male jobs remain almost entirely male.

[Willett et al., 2007]
Scented Widgets

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hue</td>
<td>Varies the hue of the widget (or of a visualization embedded in it)</td>
<td>Option A</td>
</tr>
<tr>
<td>Saturation</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td>Option A</td>
</tr>
<tr>
<td>Opacity</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td>Option A</td>
</tr>
<tr>
<td>Text</td>
<td>Inserts one or more small text figures into the widget</td>
<td>(2) Option A</td>
</tr>
<tr>
<td>Icon</td>
<td>Inserts one or more small icons into the widget</td>
<td>Option A</td>
</tr>
<tr>
<td>Bar Chart</td>
<td>Inserts one or more small bar chart visualizations into the widget</td>
<td>Option A</td>
</tr>
<tr>
<td>Line Chart</td>
<td>Inserts one or more small line chart visualizations into the widget</td>
<td>Option A</td>
</tr>
</tbody>
</table>

Figure 2. Examples of several scent encodings.

Baudisch et al's Chi's ScentTrails system [16] explores the use of web data that provides a toolkit to accommodate scenting. Examples include bar charts, line charts, and histograms, which are useful for visualizing data such as the number of visits or recency. These visualizations can be embedded into standard interface widgets such as sliders, buttons, or combo boxes to provide visual cues for navigating semantic views.

Temporal data regarding changes in any of the application state it leads to has been manipulated. The framework provides a toolkit for displaying aggregated activity patterns. Blogs and discussion forums are usually associated with other data and metadata that can be useful aids for navigation. For example, position encodings are useful views. Similarly, indicating the kinds of visual encodings used to convey this data can be useful for displaying highlighting, selection, and editing activity throughout the document.

Social activity metrics are another potential data source, providing a sense of the density of data across the information. For example, displaying the number of visits to a hyperlink text to indicate paths outperforms both searching and browsing. However, temporal data regarding changes in any of the application state it leads to has been manipulated. The framework provides a toolkit for displaying aggregated activity patterns. Blogs and discussion forums are usually associated with other data and metadata that can be useful aids for navigation. For example, position encodings are useful views. Similarly, indicating the kinds of visual encodings used to convey this data can be useful for displaying highlighting, selection, and editing activity throughout the document.

[Willett et al., 2007]
Star Plots

Aberfeldy
- Malty
- Fruity

Aberlour
- Floral

AnCnoc
- Smoky

Ardmore
- Body
- Sweetness

ArranIsleOf
- Nutty

Auchentoshan
- Spicy
- Honey
- Smoky

Attribute Filtering on Star Plots

[Yang et al., 2003]
Attribute Filtering

• How to choose which attributes should be filtered?
  - User selection?
  - Statistics: similarity measures, attributes with low variance are not as interesting when comparing items

• Can be combined with item filtering
Aggregation

• Usually involves **derived** attributes

• Examples: mean, median, mode, min, max, count, sum

• Remember expressiveness principle: still want to avoid implying trends or similarities based on aggregation

<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>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>
## Aggregation

- Usually involves **derived** attributes
- Examples: mean, median, mode, min, max, count, sum
- Remember expressiveness principle: still want to avoid implying trends or similarities based on aggregation

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
</tr>
</tbody>
</table>

Mean of $x$: 9
Variance of $x$: 11
Mean of $y$: 7.50
Variance of $y$: 4.122
Correlation: 0.816
Anscombe's Quartet

[F. J. Anscombe]
Aggregation: Histograms

- Very similar to bar charts
- Often shown without space between (continuity)
- Choice of number of bins
  - Important!
  - Viewers may infer different trends based on the layout

[Munzner (ill. Maguire), 2014]