Data Visualization (CIS 468)

Filtering & Aggregation

Dr. David Koop
Superimposition

Temperature (°F)

New York

San Francisco

Austin

M. Bostock
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)
D3 Multiple Views and Interaction

• https://codepen.io/dakoop/pen/oQxxmx

• Process `mouseover` and `mouseout` events
  - Get selected element
  - Provide **feedback** (e.g. highlighting)

• Find matching items in other view(s)
  - Can use **filter** for this
  - Highlight them
  - Make sure that if they overlap, the highlighted item is on top
Assignment 4

• Link
• Interaction, Network, and Multiple Views
• Due before Thanksgiving, but remember Quiz 2!
Quiz 2

• Next Tuesday

• Similar Format to Quiz 1:
  - Multiple Choice and Free Response

• First half of the class period

• Focus on material from Midterm through this Thursday
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**
- **Aggregate**
- **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
Example: NYC Health Dept. Restaurant Ratings

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.

[Gracie's Cafe] Grade pending 27

[FIND A RESTAURANT] [FIND A LOCATION] [FILTER] [All grades] [All violations] [All cuisines]

[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
  - Sliders, 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

Occupation
- Artist / Art Teacher
- Athlete
- Auctioneer
- Author
- Baggage man
- Baker
- Bank Teller
- Barber / Beautician
- Bartender
- Bill Collector
- Blacksmith
- Blaster
- Boarding House Keeper
- Boatman
- Boilermaker
- Bookbinder
- Bookkeeper
- Bootblack
- Building Manager
- Bus Driver
- Buyer - Farm
- Buyer - Store
- Cabinet Maker
- Car Washer
- Carpenter

Scale
- Total People Count
- % of Work Force
- # of times viewed

Figure 3: The number and variety of j

[Willett et al., 2007]
## Scented Widgets

In designing a framework for encoding scent within widgets, we aim to interface with widgets collaborators are using. Our scented widgets enable synchronous collaboration, such that users can see in real-time the visualizations provided by widgets. This includes anti-aliased text and smooth, steady presentation of data.

### Scented Widget Encoding

Scented widgets rely on link text to indicate paths. However, there are other data-based cues that can be used to provide information scent cues based on the author of a comment or an edit. These cues can support social discussions, while the del.icio.us social bookmarking service, for example, relies on link text to a text query and then sorts documents. ScentTrails offers navigation cues about other data types that can serve as valuable social cues for users.

### Encoding Modes

Visual encodings should be used with discretion. Modes of encoding should include the number of variables to be encoded, the types of encoding to support these variables, and consistency across the interface. More complicated metrics can be computed from the data, and they should be applied to different types of information. For example, position encodings are typically invisible to users, but which could serve as valuable visual navigation cues rather than synchronous visualization to provide “anti-afterglow effects to highlight widgets”.

### Table: Scent Encodings

<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><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Saturation</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Opacity</td>
<td>Varies the saturation of the widget (or of a visualization embedded in it)</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Text</td>
<td>Inserts one or more small text figures into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Icon</td>
<td>Inserts one or more small icons into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Bar Chart</td>
<td>Inserts one or more small bar chart visualizations into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
<tr>
<td>Line Chart</td>
<td>Inserts one or more small line chart visualizations into the widget</td>
<td><img src="image" alt="Option A" /> <img src="image" alt="Option B" /></td>
</tr>
</tbody>
</table>

### Rationale and Implementation

Many of these guidelines are covered by our implementation. The inclusion of too many scented widgets can serve as a hindrance to the user experience. In this case, we discuss the design decisions made in our implementation, with a focus on what the scent cues correspond to. For example, displaying the number of visits to a dataset, as is location, can be embedded into the widgets. Examples include bar and line charts over a slider (e.g., Figure 1, [8]).

### Guidelines for Scented Widgets

- **Consistency:** Ensure consistency across the interface. More complicated metrics can be computed from the data, and they should be applied to different types of information. For example, position encodings are typically invisible to users, but which could serve as valuable visual navigation cues rather than synchronous visualization to provide “anti-afterglow effects to highlight widgets”.

### Example: Scented Widget Implementation

In the domain of collaboration, scenting cues should be used with discretion. Modes of encoding should include the number of variables to be encoded, the types of encoding to support these variables, and consistency across the interface. More complicated metrics can be computed from the data, and they should be applied to different types of information. For example, position encodings are typically invisible to users, but which could serve as valuable visual navigation cues rather than synchronous visualization to provide “anti-afterglow effects to highlight widgets”.

[Willett et al., 2007]
Star Plots

Aberfeldy
Malty, Fruity
Nutty, Smoky
Spicy
Honey
Aberlour
Floral
Body
Sweetness
AnCnoc
Ardbeg
Ardmore
Malty, Fruity
Nutty
Spicy
Honey
Smoky
ArranIsleOf
Floral
Body
Sweetness
Auchentoshan
Auchroisk

[K. Schaul]
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></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
<td></td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>8.0</td>
<td>12.50</td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
<td></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

\[ \begin{align*}
Y_1 & = 4 + 0.5X_1 + \epsilon_1 \\
Y_2 & = 8 + 0.5X_2 + \epsilon_2 \\
Y_3 & = 8 + 0.5X_3 + \epsilon_3 \\
Y_4 & = 8 + 0.5X_4 + \epsilon_4
\end{align*} \]

\text{[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]
Binning Scatterplots

- At some point, cannot see density
- Blobs on top of blobs
- 2D Histogram is a histogram in 2D encoded using color instead of height
- Each region is aggregated
Hexagonal Binning

- Hexagonal bins are more circular
- Distance to the edge is not as variable
- More efficient aggregation around the center of the bin
Spatial Aggregation

In cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results.
Spatial Aggregation

In cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results.
Spatial Aggregation

In cartography, changing the boundaries of the regions used to analyze data can yield dramatically different results.
Modifiable Areal Unit Problem

- How you draw boundaries impacts the type of aggregation you get
- Similar to bins in histograms
- Gerrymandering

[Wonkblog, Washington Post, Adapted from S. Nass]
Fig. 8. The left side shows the discrete scatterplot of the "b-lunt-fin" data set, whereas the continuous version is shown on the right. Both types of scatterplots visualize the scalar data value along the horizontal axis and the magnitude of the gradient along the vertical axis. Choosing these data dimensions, material and boundary identification is possible by finding arc-like structures.

Fig. 9. In the upper part, both types of scatterplots are shown for the "tornado" data set. The upper-left images show the discrete scatterplot, to the right is the continuous version. Both scatterplots employ the same color lookup table as the previous example. The lower part shows three volume visualizations of the data set. The lower-left image (a) shows the tornado visualized by a representative isosurface of velocity magnitude. The image in the middle (b) shows highlighted voxels (yellow) that were marked in the continuous scatterplot. This highlighting corresponds to the upper-right selection rectangle in the continuous scatterplot. The other selection rectangle in the lower-middle part of the continuous scatterplot highlights different voxels, as shown in the lower-right volume-visualization image (c). In image (c), highlighted voxels (yellow) and the velocity magnitude are simultaneously visualized by rather transparent volume rendering in order to show selected features at different depths. Therefore, we can see that different voxels than in (b) are highlighted, especially not the ones in the center of the tornado.

CONCLUSION AND FUTURE WORK

We have presented continuous scatterplots as a generalization of conventional scatterplots. One aspect of generalization is the support of any dimension of the domain of the data set and of the scatterplot. The other aspect of generalization is the extension to data defined on continuous domains. The basis for continuous scatterplots is provided in the form of a generic mathematical model. This mathematical model maps an arbitrary density value defined on an n-D input data set to m-D scatterplots. We have shown how continuous scatterplots are related to conventional discrete histograms and to histograms of isosurface statistics. In particular, the 2-D version of continuous scatterplots is, by construction, identical to conventional discrete scatterplots in the limit process of infinitely dense sample points. Therefore, continuous scatterplots lead to the same basic visual mapping as traditional histograms, scatterplots, or other frequency plots, utilizing their proven visualization power. We have provided typical examples of multi-attribute visualization—such as 2-D transfer function specification and flow visualization—to demonstrate the applicability of our approach. The difference to discrete scatterplots is especially visible for low-resolution data sets and for data sets defined on grids with largely different cell sizes.

The main advantage of continuous scatterplots is that they are directly designed for input data defined on continuous domains. Therefore, this paper adds one missing piece to the general approach of applying continuous scatterplots to such data.

[Bachthaler & Weiskopf, 2008]
Continuous Scatterplot

Fig. 8. The left side shows the discrete scatterplot of the "b-lunt-fin" data set, whereas the continuous version is shown on the right. Both types of scatterplots visualize the scalar data value along the horizontal axis and the magnitude of the gradient along the vertical axis. Choosing these data dimensions, material and boundary identification is possible by finding arc-like structures.

Fig. 9. In the upper part, both types of scatterplots are shown for the "tornado" data set. The upper-left images show the discrete scatterplot, to the right is the continuous version. Both scatterplots employ the same color lookup table as the previous example. The lower part shows three volume visualizations of the data set. The lower-left image (a) shows the tornado visualized by a representative isosurface of velocity magnitude. The image in the middle (b) shows highlighted voxels (yellow) that were marked in the continuous scatterplot. This highlighting corresponds to the upper-right selection rectangle in the continuous scatterplot. The other selection rectangle in the lower-middle part of the continuous scatterplot highlights different voxels, as shown in the lower-right volume-visualization image (c). In image (c), highlighted voxels (yellow) and the velocity magnitude are simultaneously visualized by rather transparent volume rendering in order to show selected features at different depths. Therefore, we can see that different voxels than in (b) are highlighted, especially not the ones in the center of the tornado.

CONCLUSION AND FUTURE WORK
We have presented continuous scatterplots as a generalization of conventional scatterplots. One aspect of generalization is the support of any dimension of the domain of the data set and of the scatterplot. The other aspect of generalization is the extension to data defined on continuous domains. The basis for continuous scatterplots is provided in the form of a generic mathematical model. This mathematical model maps an arbitrary density value defined on an n-D input data set to an m-D scatterplots. We have shown how continuous scatterplots are related to conventional discrete histograms and to histograms of isosurface statistics. In particular, the 2-D version of continuous scatterplots is, by construction, identical to conventional discrete scatterplots in the limit process of infinitely dense sample points. Therefore, continuous scatterplots lead to the same basic visual mapping as traditional histograms, scatterplots, or other frequency plots, utilizing their proven visualization power. We have provided typical examples of multi-attribute visualization—such as 2-D transfer function specification and flow visualization—to demonstrate the applicability of our approach. The difference to discrete scatterplots is especially visible for low-resolution data sets and for data sets defined on grids with largely different cell sizes.

The main advantage of continuous scatterplots is that they are directly designed for input data defined on continuous domains. Therefore, this paper adds one missing piece to the general approach of our approach. [Bachthaler & Weiskopf, 2008]
Boxplots

- Show **distribution**
- Single value (e.g. mean, max, min, quartiles) doesn't convey everything
- Created by John Tukey who grew up in New Bedford!
- Show **spread** and **skew** of data
- Best for **unimodal** data
- Variations like vase plot for multimodal data
- Aggregation here involves many different marks
Boxplot Example

(a) Overall Activity  (b) Structural Activity  (c) Parameter Activity  (d) Layout Activity

Percentage (%)  

[L. Lins et al., 2008]  

D. Koop, CIS 468, Fall 2018
Four Distributions, Same Boxplot…

[C. Choonpradub and D. McNeil, 2005]
Hierarchical Parallel Coordinates

Figure 4: This image sequence shows a Fatal Accident data set of 230,000 data elements at different levels of detail. The first image shows a cut across the root node. The last image shows the cut chaining all the leaf nodes. The rest of the images show intermediate cuts at varying levels of detail. (See Color Plates).

Figure 6: Left image shows Iris data set without proximity-based coloring. Right image shows Iris data set with proximity-based coloring revealing three distinct clusters depicted by concentrations of blue, green and pink lines. (See Color Plates).

[Fua et al., 1999]