Data Analysis and Information Visualization

Texas Advanced Computing Center
Napoleon Vs. Russia, 1812-1813

Napoleon's March to Moscow  The War of 1812

This chart of Charles Joseph Minard (1785-1850), the French engineer, shows the terrible fate of Napoleon's army in Russia. Described by E. J. May as seeming to defy the laws of the human by its brutal eloquence, this combination of a map and time-series, drawn in silks, portrays the devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (1,25,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was then sacked and destroyed. Scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, for more enlightening than just a single number bouncing along over time. So variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and
Florence Nightingale Cox Comb

**Diagram of the Causes of Mortality**

*In the Army in the East*

1. April 1854 to March 1855
2. April 1855 to March 1856

*The areas of the blue, red, & black wedges are each measured from the centre as the common series.*

*The blue wedges measured from the centre of the circle represent area for the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.*

*The black line across the red triangle in No. 1, 1854 marks the boundary of the deaths from all other causes during the month.*

In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black.

*The entire areas may be compared by following the blue, the red & the black lines enclosing them.*
Data Analysis and Information Visualization

• Data Sources
  • CSV
  • Excel
  • Databases – SQL, no-SQL, Map-Reduce...

• Data Analysis – Extract Information from Data
  • Statistics: PCA, regression...
  • Machine Learning: clustering, classification...
  • Data Mining

• Visualization – Information Representation
  • histograms, dendograms, tree maps..
Analysis vs. Visualization

• Anscombe's Quartet

<table>
<thead>
<tr>
<th>Data Set I</th>
<th>Data Set II</th>
<th>Data Set III</th>
<th>Data Set 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>

Why Visualize?

- Simple statistical analysis

<table>
<thead>
<tr>
<th></th>
<th>Data Set I</th>
<th>Data Set II</th>
<th>Data Set III</th>
<th>Data Set IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>X</strong></td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>8.0</td>
</tr>
<tr>
<td><strong>Y</strong></td>
<td>8.04</td>
<td>9.14</td>
<td>7.46</td>
<td>6.58</td>
</tr>
<tr>
<td><strong>X</strong></td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
</tr>
<tr>
<td><strong>Y</strong></td>
<td>6.95</td>
<td>8.14</td>
<td>6.77</td>
<td>5.76</td>
</tr>
<tr>
<td><strong>X</strong></td>
<td>13.0</td>
<td>13.0</td>
<td>13.0</td>
<td>8.0</td>
</tr>
<tr>
<td><strong>Y</strong></td>
<td>7.58</td>
<td>8.74</td>
<td>12.74</td>
<td>7.71</td>
</tr>
<tr>
<td><strong>X</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Y</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td>9.0</td>
<td>7.5</td>
<td>9.0</td>
<td>7.5</td>
</tr>
<tr>
<td><strong>variance</strong></td>
<td>10.0</td>
<td>3.75</td>
<td>10.0</td>
<td>3.75</td>
</tr>
<tr>
<td><strong>correlation</strong></td>
<td>0.816</td>
<td>0.816</td>
<td>0.816</td>
<td>0.816</td>
</tr>
<tr>
<td><strong>regression</strong></td>
<td>Y=3+0.5x</td>
<td>Y=3+0.5x</td>
<td>Y=3+0.5x</td>
<td>Y=3+0.5x</td>
</tr>
<tr>
<td></td>
<td>4.82</td>
<td>7.26</td>
<td>6.42</td>
<td>8.91</td>
</tr>
<tr>
<td></td>
<td>5.68</td>
<td>4.74</td>
<td>5.73</td>
<td>6.89</td>
</tr>
</tbody>
</table>
Why Visualize?

<table>
<thead>
<tr>
<th>Data Set I</th>
<th>Data Set II</th>
<th>Data Set III</th>
<th>Data Set IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X}$</td>
<td>$\bar{Y}$</td>
<td>$\bar{X}$</td>
<td>$\bar{Y}$</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>7.11</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>7.81</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.84</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>

---

<table>
<thead>
<tr>
<th>Data Set I</th>
<th>Data Set II</th>
<th>Data Set III</th>
<th>Data Set IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X}$</td>
<td>$\bar{Y}$</td>
<td>$\bar{X}$</td>
<td>$\bar{Y}$</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>7.11</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>7.81</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.84</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>

---

In the simple case

Line Graph
- x-axis requires quantitative variable
- Variables have contiguous values
- Familiar/conventional ordering among ordinals

Scatter Plot
- Convey overall impression of relationship between two variables

Bar Graph
- Comparison of relative point values

Pie Chart
- Emphasizing differences in proportion among a few numbers
- Histogram vs. Pie
From Data to Graph

• Information Type:
  • Easy case: 1D, 2D, 3D spatial
  • What about more dimensions?

• Structured data
  • Tree
  • Network
  • Graph

• Text and document collections
Example: A movie database

Attributes

Types:
- Quantitative
- Ordinal
- Nominal/Categorical

Items
(aka: cases, tuples, data points, ...)

Note: No spatial info!

<table>
<thead>
<tr>
<th>Year</th>
<th>Length</th>
<th>Title</th>
<th>Subject</th>
<th>Actor</th>
<th>Actress</th>
<th>Director</th>
<th>Popularity</th>
<th>Awards</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>135</td>
<td>HUNT FOR RED OCTOBER</td>
<td>Drama</td>
<td>Connery, Sean</td>
<td>McTiernan, J.</td>
<td>8</td>
<td>No</td>
<td>Nicholas Cage gif</td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>105</td>
<td>ALIENS</td>
<td>Sci-Fi</td>
<td>Sigourney, Weaver</td>
<td>Ripley, Ellen</td>
<td>10</td>
<td>Yes</td>
<td>Sigourney, Weaver gif</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>130</td>
<td>JURASSIC PARK</td>
<td>Sci-Fi</td>
<td>Grant, Sam</td>
<td>Sattler, A.</td>
<td>8</td>
<td>No</td>
<td>Grant, Sam gif</td>
<td></td>
</tr>
</tbody>
</table>

Microsoft Excel - film.txt
Visual Mapping

1. Map: data items → visual marks

Visual marks:

- Points
- Lines
- Areas
- Volumes
- Glyphs
Visual Mapping

1. Map: data items → visual marks
2. Map: data attributes → visual properties of marks

Visual properties of marks:
- Position, x, y, z
- Size, length, area, volume
- Orientation, angle, slope
- Color, gray scale, texture
- Shape
- Animation, blink, motion
Example: A Movie database

- Year → X
- Length → Y
- Popularity → size
- Subject → color
- Award? → shape
Accuracy of Visual Attributes

- Position
- Length
- Angle, Slope
- Size
- Color
- Shape

Increased accuracy for quantitative data
Map n-D space onto 2-D screen

- Visual representations:
  - Continuous
    - Heatmap, heightfield, volume
  - Multiple views
    - E.g. plot matrices, brushing histograms, ...
  - Complex glyphs
    - E.g. star glyphs, faces ...
  - More axes
    - E.g. Parallel coords, star coords, ...
Continuous approximations

- Reduce a high-dimensional data set to 2D or 3D
- Principal component analysis (PCA):
  - determine 2-3 significant vectors
  - Represent data as linear combinations of those vectors
- Topological Landscapes (Weber et al. 07, Harvey et al. 10)

- Are PCA axes relevant?
Continuous Descriptors

- Transform spatial data into another domain
  - Histogram
  - Fourier transform, other spectra
- Fourier spectra of 7000 carbon molecules with 6 atoms or less
Multiple Views

• Basic idea:
  • Showing multiple views of same data set at the same time.
  • Each individual visualizations might be of same or different types.
• Brushing and linking
  • With interactive visualizations, All views might be linked so that action, such as selection, on one view might be reflected in all other views.
• Example: Scatter plot matrix
  • Create a 2d views for all attributes pairs
## Example Data

<table>
<thead>
<tr>
<th>country</th>
<th>model name</th>
<th>mpg</th>
<th>weight</th>
<th>ratio</th>
<th>hp</th>
<th>disp.</th>
<th>cyl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Buick Estate Wagon</td>
<td>16.9</td>
<td>4.360</td>
<td>2.73</td>
<td>155</td>
<td>350</td>
<td>8</td>
</tr>
<tr>
<td>USA</td>
<td>Ford Country Squire Wagon</td>
<td>15.5</td>
<td>4.054</td>
<td>2.26</td>
<td>142</td>
<td>351</td>
<td>8</td>
</tr>
<tr>
<td>USA</td>
<td>Chevy Malibu Wagon</td>
<td>19.2</td>
<td>3.605</td>
<td>2.56</td>
<td>125</td>
<td>267</td>
<td>8</td>
</tr>
<tr>
<td>USA</td>
<td>Chrysler LeBaron Wagon</td>
<td>18.5</td>
<td>3.940</td>
<td>2.45</td>
<td>150</td>
<td>360</td>
<td>8</td>
</tr>
<tr>
<td>USA</td>
<td>Chevette</td>
<td>30.0</td>
<td>2.155</td>
<td>3.70</td>
<td>68</td>
<td>98</td>
<td>4</td>
</tr>
<tr>
<td>Japan</td>
<td>Toyota Corona</td>
<td>27.5</td>
<td>2.560</td>
<td>3.05</td>
<td>95</td>
<td>134</td>
<td>4</td>
</tr>
<tr>
<td>Japan</td>
<td>Datsun 510</td>
<td>27.2</td>
<td>2.300</td>
<td>3.54</td>
<td>97</td>
<td>119</td>
<td>4</td>
</tr>
<tr>
<td>USA</td>
<td>Dodge Omni</td>
<td>30.9</td>
<td>2.230</td>
<td>3.37</td>
<td>75</td>
<td>105</td>
<td>4</td>
</tr>
<tr>
<td>Germany</td>
<td>Audi 5000</td>
<td>20.3</td>
<td>2.830</td>
<td>3.90</td>
<td>103</td>
<td>131</td>
<td>5</td>
</tr>
<tr>
<td>Sweden</td>
<td>Volvo 240 GL</td>
<td>17.0</td>
<td>3.140</td>
<td>3.50</td>
<td>125</td>
<td>163</td>
<td>6</td>
</tr>
<tr>
<td>Sweden</td>
<td>Saab 99 GLE</td>
<td>21.6</td>
<td>2.795</td>
<td>3.77</td>
<td>115</td>
<td>121</td>
<td>4</td>
</tr>
<tr>
<td>France</td>
<td>Peugeot 694 SL</td>
<td>16.2</td>
<td>3.410</td>
<td>3.58</td>
<td>133</td>
<td>163</td>
<td>6</td>
</tr>
<tr>
<td>USA</td>
<td>Buick Century Special</td>
<td>20.6</td>
<td>3.380</td>
<td>2.73</td>
<td>105</td>
<td>231</td>
<td>6</td>
</tr>
<tr>
<td>USA</td>
<td>Mercury Zephyr</td>
<td>20.8</td>
<td>3.070</td>
<td>3.08</td>
<td>85</td>
<td>200</td>
<td>6</td>
</tr>
<tr>
<td>USA</td>
<td>Dodge Aspen</td>
<td>18.6</td>
<td>3.620</td>
<td>2.71</td>
<td>110</td>
<td>225</td>
<td>6</td>
</tr>
</tbody>
</table>
Scatter plot Matrix Example
Brushing
Brushing Across Different Projection Types
Glyphs

• Glyph
  – composite graphical objects where different geometric and visual attributes are used to encode multidimensional data structures in combination.

• Examples:
  – Superquadrics for DTI
  – Chernoff Face*
    • mapping $k$-dimensions to facial features

Superquadric glyphs for DT-MRI

- Determine structure of brain tissue, examining movement along N different axes
- G. Kindlmann, University of Utah / University of Chicago
Glyphs: Chernoff Faces

- http://www.stat.harvard.edu/People/Faculty/Herman_Chernoff/
Chernoff Face Example

- Map to 10 dimension binary vector
  
  \[
  [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
  \]
  
  \[
  [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
  \]

- Evaluation Of Judges

![Chernoff Faces Example](image.png)
• What’s a problem with using star glyphs?
Using additional axes

• Easy example:
  – 2D scatter plot $\rightarrow$ 3D scatter plot

• Space $>$ 3D?
Parallel Coordinates

• Instead of orthogonal axes, let’s go parallel

• \((0,1,-1,2) = (0,0,0,0)\)

Inselberg, “Multidimensional detective” (parallel coordinates)
Parallel Coordinates

• Important factors:
  – the scaling of the axes.
  – the order of the axes
  – the rotation of the axes
Parallel Coordinates

The image shows two parallel coordinate plots. The plots compare different parameters such as weight, drive ratio, horsepower, displacement, and cylinder count. The plots are used to visualize and compare multiple variables simultaneously, making it easier to identify patterns and relationships among the variables.
Parallel Coordinates

- Better visualizations

http://davis.wpi.edu/xmdv/
Parallel Coordinates

- 3D parallel coordinates
Visualizing Structured Data

• Some data contains relationships between entities
  – Tree Structures
    • Phylogenetic Trees
    • Presidential voting by state, county and precinct
  – Generalized relations
    • Who knows whom in a college dorm
    • Who follows whom on Twitter
Tree-Structured Data

- Phylogenetetic Trees
Can Get Busy
Tree-Structure With Values

• Suppose we know the breakdown of votes for each precinct in the country....
Treemaps
Generalized Relation Data

Bob knows Bill
Bill knows Ted
Ted knows Ann
Bill knows Ann
Bob knows Ken
...

TACC
THE UNIVERSITY OF TEXAS AT AUSTIN
TEXAS ADVANCED COMPUTING CENTER
Gets Busy Fast...
Relation Data With Weights

Bob likes Bill a little
Bill likes Ted a lot
Ted dislikes Ann
Bill really likes
Bob despises Ken
...

Eigengene adjacency heatmap
Information visualization ...

• General Aims
  – Use human perceptual capabilities
  – To gain insights into large and abstract data sets that are difficult to extract using standard query languages

• Exploratory Visualization
  – Look for structure, patterns, trends, anomalies, relationships
  – Provide a qualitative overview of large, complex data sets
  – Assist in identifying region(s) of interest and appropriate parameters for more focused quantitative analysis
Analysis and Visualization Tools

• Data Management and Applications
  – Database systems
  – Statistical packages: R

• Visualization Tools
  – Geographical Information Systems (GIS) - ESRI
  – Tableau, Spotfire...

• Toolkits
  – Web-native Javascript: D3, Protovis, OpenLayers
  – Python interfaces to DBs, R, Matplotlib...
Techniques

• Lots and Lots
  – And as many permutations as there are graduate students

• Few really general applications
  – Excel

• Lots of Toolkits
  – In particular, in Javascript for web applications
Good visualization

• Use of computer-supported, interactive, visual representations of abstract data to amplify cognition
  – Visual representation can enhance recognition
    • Recognition of patterns
    • Abstraction and aggregation
    • Perceptual interference
  – Facilitate data exploration
    • Interactive medium
    • High data density
    • Greater access speed
  – Increased analytic resources
    • Parallel perceptual processing
    • Offload work from cognitive to perceptual system
Fun Websites

• Atlas of Science (Katy Borner, IU)
  - http://scimaps.org/atlas/maps
• Many Eyes: a project to encourage sharing and conversation around visualizations (need java)
• New York Times Infographics
• Gap Minder http://www.gapminder.org
• How to visualize data with Chernoff face using R
Reference materials

• References

• Software
  – Matplotlib/Python
  – Google charts/Javascript
  – InfoVis ToolKit
  – Prefuse
  – Titan Libraries/VTK InfoVis Libraries