Large Scale Information Visualization

Jing Yang
Fall 2007
Pre-requisite

- Information Visualization
- Course slides:  www.cs.uncc.edu/~jyang13
- Basic techniques will be covered in the first two weeks
- Please learn other contents by yourself
Motivation - Data Explosion

- Society is more complex: computers, internet, web, ...

- How much data?
  - Between 1 and 2 exabytes of unique info produced per year
    - 10000000000000000000 bytes
    - 250 meg for every man, woman and child
    - Printed documents only .003% of total

Peter Lyman and Hal Varian, 2000
Cal-Berkeley, Info Mgmt & Systems
www.sims.berkeley.edu/how-much-info

Slide courtesy of John Stasko
Data Overload

- Confound: How to make use of the data
  - How do we make sense of the data?
  - How do we harness this data in decision-making processes?
  - How do we avoid being overwhelmed?
Human Vision

- Highest bandwidth sense
- Fast, parallel
- Pattern recognition
- Pre-attentive
- Extends memory and cognitive capacity
- People think visually
The Challenge

- Transform the data into information (understanding, insight) thus making it useful to people
Example

Which state has the highest income?

Questions:
- Is there a relationship between income and education?
- Are there any outliers?

Example courtesy of Chris North
Visualize the Data
Illustrate Our Approach

- Provide tools that present data in a way to help people understand and gain insight from it
- Cliches
  - “Seeing is believing”
  - “A picture is worth a thousand words”
Don’t you believe it?

This graphic is worth at least 700 words.
Visualization

- Visualization - *the use of computer-supported, interactive visual representations of data to amplify cognition*
  
  From [Card, Mackinlay Shneiderman ’98]

- Often thought of as process of making a graphic or an image
- Really is a cognitive process
  - Form a mental image of something
  - Internalize an understanding
- “The purpose of visualization is insight, not pictures”
  - Insight: discovery, decision making, explanation

Slide courtesy of John Stasko
What is Information Visualization

Information Visualization is NOT Scientific Visualization

Scientific Visualization is primarily related to and represents something physical or geometric

- Examples:
  - Air flow over a wing
  - Stresses on a girder
  - Weather over Pennsylvania
What is Information Visualization

- It is about abstract data
InfoVis Is About Numerical Data
InfoVis Is About Ordinal Data

Figure 3: AP data from July - August 1990. A wide current in the river indicates heavy use of a topic, while changes in color distribution correlate to changes in themes.

ThemeRiver: Visualizing Theme Changes Over Time [Havre et al. Infovis 00]
InfoVis Is About Nominal Data

Swiss mountain map, L. Matterhorn, 1983
InfoVis Is About Structured Data
InfoVis Is About Space

Statistics Bureau, Tokyo, 1985
InfoVis Is About Time

New York Times, January 1982
InfoVis Is About Change
InfoVis Is About Motion and Process

Illustration of magic turning a silver coin into a copper coin
InfoVis Is About All Kinds of Information
What is Information Visualization

- It is about analyzing, communicating, and decision making

Of all method for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful  - E. Tufte
What is Information Visualization

- It is about scale and dimensionality

Scale
- Essential problem in reasoning is comparison
- Comparisons must be enforced within scope of eyespan

Dimensionality
- The world is multi-dimensional
- The paper and computer is 2 dimensional
- Escape from the flatland
What is Information Visualization

- It is about interactive exploration
  - Want to show multiple different perspectives on the data
  - A way to increase scalability and dimensionality
What is Information Visualization

- It is about applications
  - Document, images, videos, multimedia
  - Financial/business data
  - Internet information
  - Software
In five years, 100 million people will be using an information-visualization tool on a near-daily basis. And products that have visualization as one of their top three features will earn $1 billion per year.

Ramana Rao, founder and chief technology officer, Inxight Software Inc., Sunnyvale, Calif.

http://www.computerworld.com/databasetopics/data/story/0,10801,80243,00.html

Slide courtesy of John Stasko
Basic Techniques 1: Multi-dimensional Visualization

Class 1, Part B
Multi-dimensional (Multivariate) Dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>At Bats</th>
<th>Hits</th>
<th>Home Run</th>
<th>RBIs</th>
<th>Walks</th>
<th>Years In M</th>
<th>Career At</th>
<th>Career Hit</th>
<th>Career RBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy Allison</td>
<td>293</td>
<td>66</td>
<td>1</td>
<td>30</td>
<td>29</td>
<td>1</td>
<td>293</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>Alan Ashby</td>
<td>315</td>
<td>81</td>
<td>7</td>
<td>24</td>
<td>38</td>
<td>14</td>
<td>3449</td>
<td>835</td>
<td></td>
</tr>
<tr>
<td>Alvin Davis</td>
<td>479</td>
<td>130</td>
<td>18</td>
<td>66</td>
<td>72</td>
<td>76</td>
<td>1624</td>
<td>457</td>
<td></td>
</tr>
<tr>
<td>Andre Dawson</td>
<td>496</td>
<td>141</td>
<td>20</td>
<td>65</td>
<td>76</td>
<td>3</td>
<td>5620</td>
<td>1575</td>
<td></td>
</tr>
<tr>
<td>Andre Galarraga</td>
<td>321</td>
<td>87</td>
<td>10</td>
<td>39</td>
<td>42</td>
<td>30</td>
<td>396</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>Alfredo Griffin</td>
<td>594</td>
<td>169</td>
<td>4</td>
<td>74</td>
<td>51</td>
<td>35</td>
<td>4040</td>
<td>1133</td>
<td></td>
</tr>
<tr>
<td>Al Newman</td>
<td>185</td>
<td>37</td>
<td>1</td>
<td>23</td>
<td>8</td>
<td>21</td>
<td>214</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Argenis Salaza</td>
<td>298</td>
<td>73</td>
<td>0</td>
<td>24</td>
<td>24</td>
<td>7</td>
<td>509</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Andres Thomas</td>
<td>323</td>
<td>81</td>
<td>6</td>
<td>26</td>
<td>32</td>
<td>8</td>
<td>341</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Andrea Thornton</td>
<td>401</td>
<td>92</td>
<td>17</td>
<td>49</td>
<td>66</td>
<td>65</td>
<td>5206</td>
<td>1332</td>
<td></td>
</tr>
<tr>
<td>Alan Trammell</td>
<td>574</td>
<td>159</td>
<td>21</td>
<td>107</td>
<td>75</td>
<td>59</td>
<td>4631</td>
<td>1300</td>
<td></td>
</tr>
<tr>
<td>Alex Treviso</td>
<td>202</td>
<td>53</td>
<td>4</td>
<td>31</td>
<td>26</td>
<td>27</td>
<td>1876</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>Andy Van Slyke</td>
<td>418</td>
<td>113</td>
<td>13</td>
<td>61</td>
<td>47</td>
<td>4</td>
<td>1512</td>
<td>392</td>
<td></td>
</tr>
<tr>
<td>Alan Wiggins</td>
<td>239</td>
<td>60</td>
<td>0</td>
<td>30</td>
<td>11</td>
<td>22</td>
<td>1911</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>Bill Almon</td>
<td>196</td>
<td>43</td>
<td>7</td>
<td>29</td>
<td>27</td>
<td>30</td>
<td>3321</td>
<td>825</td>
<td></td>
</tr>
<tr>
<td>Billy Beane</td>
<td>183</td>
<td>39</td>
<td>3</td>
<td>20</td>
<td>15</td>
<td>11</td>
<td>201</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Buddy Bell</td>
<td>568</td>
<td>158</td>
<td>20</td>
<td>89</td>
<td>76</td>
<td>73</td>
<td>8068</td>
<td>2273</td>
<td></td>
</tr>
<tr>
<td>Buddy Biancal</td>
<td>190</td>
<td>46</td>
<td>2</td>
<td>24</td>
<td>8</td>
<td>15</td>
<td>479</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>Bruce Bochette</td>
<td>407</td>
<td>104</td>
<td>6</td>
<td>57</td>
<td>43</td>
<td>65</td>
<td>5233</td>
<td>1478</td>
<td></td>
</tr>
</tbody>
</table>

Microsoft Excel - baseball

Slide courtesy of John Stasko
Data Item (Object, Record, Case)

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>At Bat</th>
<th>Hits</th>
<th>Home Run</th>
<th>Runs</th>
<th>Rbi</th>
<th>Walks</th>
<th>Years In M</th>
<th>Career At</th>
<th>Career Hits</th>
<th>Career Bat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>STRING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Andy Allison</td>
<td>293</td>
<td>66</td>
<td>1</td>
<td>30</td>
<td>29</td>
<td>14</td>
<td>1</td>
<td>293</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Alan Ashby</td>
<td>315</td>
<td>81</td>
<td>7</td>
<td>24</td>
<td>38</td>
<td>39</td>
<td>13</td>
<td>3449</td>
<td>835</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Alvin Davis</td>
<td>479</td>
<td>130</td>
<td>18</td>
<td>66</td>
<td>72</td>
<td>76</td>
<td>3</td>
<td>1624</td>
<td>457</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Andre Dawson</td>
<td>496</td>
<td>141</td>
<td>20</td>
<td>65</td>
<td>78</td>
<td>37</td>
<td>11</td>
<td>5628</td>
<td>1575</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Andres Galarraga</td>
<td>321</td>
<td>87</td>
<td>10</td>
<td>39</td>
<td>42</td>
<td>30</td>
<td>2</td>
<td>396</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Alfredo Griffin</td>
<td>694</td>
<td>169</td>
<td>4</td>
<td>74</td>
<td>51</td>
<td>35</td>
<td>11</td>
<td>4408</td>
<td>1133</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Al Newman</td>
<td>185</td>
<td>37</td>
<td>1</td>
<td>23</td>
<td>8</td>
<td>21</td>
<td>2</td>
<td>214</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Argenis Salazar</td>
<td>298</td>
<td>73</td>
<td>0</td>
<td>24</td>
<td>24</td>
<td>7</td>
<td>3</td>
<td>509</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Andres Thomas</td>
<td>323</td>
<td>81</td>
<td>6</td>
<td>26</td>
<td>32</td>
<td>8</td>
<td>2</td>
<td>341</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Andrea Thornton</td>
<td>401</td>
<td>92</td>
<td>17</td>
<td>49</td>
<td>66</td>
<td>65</td>
<td>13</td>
<td>5206</td>
<td>1332</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Alan Trammell</td>
<td>574</td>
<td>159</td>
<td>21</td>
<td>107</td>
<td>75</td>
<td>59</td>
<td>10</td>
<td>4631</td>
<td>1300</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Alex Trevino</td>
<td>202</td>
<td>53</td>
<td>4</td>
<td>31</td>
<td>26</td>
<td>27</td>
<td>9</td>
<td>1876</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Andy Van Slyke</td>
<td>418</td>
<td>113</td>
<td>13</td>
<td>61</td>
<td>47</td>
<td>4</td>
<td>1512</td>
<td>392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Alan Wiggins</td>
<td>239</td>
<td>60</td>
<td>0</td>
<td>30</td>
<td>11</td>
<td>22</td>
<td>6</td>
<td>1941</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Bill Almon</td>
<td>196</td>
<td>43</td>
<td>7</td>
<td>29</td>
<td>27</td>
<td>30</td>
<td>13</td>
<td>3231</td>
<td>825</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Billy Beane</td>
<td>183</td>
<td>39</td>
<td>3</td>
<td>20</td>
<td>15</td>
<td>11</td>
<td>3</td>
<td>201</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Buddy Bell</td>
<td>568</td>
<td>158</td>
<td>20</td>
<td>89</td>
<td>75</td>
<td>73</td>
<td>15</td>
<td>8068</td>
<td>2273</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Buddy Biancalda</td>
<td>190</td>
<td>46</td>
<td>2</td>
<td>24</td>
<td>8</td>
<td>15</td>
<td>5</td>
<td>479</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Bruce Bochte</td>
<td>407</td>
<td>104</td>
<td>6</td>
<td>57</td>
<td>43</td>
<td>65</td>
<td>12</td>
<td>5233</td>
<td>1478</td>
<td></td>
</tr>
</tbody>
</table>

Microsoft Excel - baseball
Dimension (Variable, Attribute)

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>At Bats</th>
<th>Hits</th>
<th>Home Run</th>
<th>Runs</th>
<th>Rbi</th>
<th>Walks</th>
<th>Years</th>
<th>M Career At</th>
<th>M Career Hits</th>
<th>Career Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Andy Allanson</td>
<td>293</td>
<td>66</td>
<td>1</td>
<td>30</td>
<td>29</td>
<td>14</td>
<td>1</td>
<td>293</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Alan Ashby</td>
<td>315</td>
<td>81</td>
<td>7</td>
<td>24</td>
<td>38</td>
<td>39</td>
<td>14</td>
<td>3449</td>
<td>835</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Alvin Davis</td>
<td>479</td>
<td>130</td>
<td>18</td>
<td>66</td>
<td>72</td>
<td>76</td>
<td>3</td>
<td>1524</td>
<td>457</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Andre Dawson</td>
<td>496</td>
<td>141</td>
<td>20</td>
<td>65</td>
<td>70</td>
<td>37</td>
<td>11</td>
<td>5628</td>
<td>1575</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Andres Galarraga</td>
<td>321</td>
<td>87</td>
<td>10</td>
<td>39</td>
<td>42</td>
<td>30</td>
<td>2</td>
<td>396</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Alfredo Griffin</td>
<td>594</td>
<td>169</td>
<td>4</td>
<td>74</td>
<td>51</td>
<td>15</td>
<td>11</td>
<td>4408</td>
<td>1133</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Al Newman</td>
<td>185</td>
<td>37</td>
<td>1</td>
<td>23</td>
<td>8</td>
<td>21</td>
<td>2</td>
<td>214</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Argenis Salaza</td>
<td>298</td>
<td>73</td>
<td>0</td>
<td>24</td>
<td>24</td>
<td>7</td>
<td>3</td>
<td>509</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Andres Thomas</td>
<td>323</td>
<td>81</td>
<td>6</td>
<td>26</td>
<td>32</td>
<td>8</td>
<td>2</td>
<td>341</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Andre Thornton</td>
<td>401</td>
<td>92</td>
<td>17</td>
<td>49</td>
<td>66</td>
<td>65</td>
<td>13</td>
<td>5206</td>
<td>1332</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Alan Trammell</td>
<td>574</td>
<td>159</td>
<td>21</td>
<td>107</td>
<td>75</td>
<td>59</td>
<td>10</td>
<td>4831</td>
<td>1300</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Alex Trevino</td>
<td>202</td>
<td>53</td>
<td>4</td>
<td>31</td>
<td>26</td>
<td>27</td>
<td>9</td>
<td>1876</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Andy Van Slyke</td>
<td>418</td>
<td>113</td>
<td>13</td>
<td>48</td>
<td>61</td>
<td>47</td>
<td>4</td>
<td>1512</td>
<td>392</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Alan Wiggins</td>
<td>239</td>
<td>60</td>
<td>0</td>
<td>30</td>
<td>11</td>
<td>22</td>
<td>6</td>
<td>1941</td>
<td>610</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Bill Almon</td>
<td>196</td>
<td>43</td>
<td>7</td>
<td>29</td>
<td>27</td>
<td>30</td>
<td>13</td>
<td>3231</td>
<td>825</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Billy Beane</td>
<td>183</td>
<td>39</td>
<td>3</td>
<td>15</td>
<td>11</td>
<td>15</td>
<td>3</td>
<td>201</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Buddy Bell</td>
<td>568</td>
<td>158</td>
<td>20</td>
<td>89</td>
<td>75</td>
<td>73</td>
<td>15</td>
<td>8068</td>
<td>2273</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Buddy Biancalle</td>
<td>190</td>
<td>46</td>
<td>2</td>
<td>24</td>
<td>8</td>
<td>15</td>
<td>5</td>
<td>479</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Bruce Bochte</td>
<td>407</td>
<td>104</td>
<td>6</td>
<td>57</td>
<td>43</td>
<td>65</td>
<td>12</td>
<td>5233</td>
<td>1478</td>
<td></td>
</tr>
</tbody>
</table>

Microsoft Excel - baseball
Multidimensional Data
Example: Iris Data

- Scientists measured the sepal length, sepal width, petal length, petal width of many kinds of iris...
Multidimensional Data
Example: Iris Data

<table>
<thead>
<tr>
<th>sepal length</th>
<th>sepal width</th>
<th>petal length</th>
<th>petal width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>4.9</td>
<td>3</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5.9</td>
<td>3</td>
<td>5.1</td>
<td>1.8</td>
</tr>
</tbody>
</table>
Classification

- Table-based techniques
  - HeatMap, tableLens

- Geometric techniques
  - Scatterplot matrices, parallel coordinates, landscapes, Dust&Magnet ...

- Icon-based techniques
  - glyphs, shape-coding, *color icons, ...

- Hierarchical techniques
  - Dimensional stacking, worlds-within-worlds, ...

- Pixel-oriented techniques
  - Recursive pattern, circle segments, spiral, axes techniques,...
Table-Based Techniques

- Basic idea: Visualization that improves the existing spreadsheet table format
  - HeatMap
  - Table Lens [RC 94]
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
This figure is used by Dr. M. Ward’s permission
Eureka (Table Lens)

# Eureka (Table Lens)

<table>
<thead>
<tr>
<th>Market</th>
<th>Company</th>
<th>Date</th>
<th>Volume</th>
<th>High</th>
<th>Low</th>
<th>Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>Xerox</td>
<td>7/23/1992</td>
<td>79600</td>
<td>71.9</td>
<td>71.3</td>
<td>71.9</td>
</tr>
<tr>
<td>New York</td>
<td>Xerox</td>
<td>7/24/1992</td>
<td>109000</td>
<td>73</td>
<td>71.3</td>
<td>72.5</td>
</tr>
<tr>
<td>New York</td>
<td>Xerox</td>
<td>7/27/1992</td>
<td>164000</td>
<td>73.8</td>
<td>72.5</td>
<td>73.5</td>
</tr>
<tr>
<td>New York</td>
<td>Xerox</td>
<td>7/28/1992</td>
<td>319000</td>
<td>75.1</td>
<td>73.3</td>
<td>75</td>
</tr>
<tr>
<td>New York</td>
<td>Xerox</td>
<td>7/29/1992</td>
<td>402000</td>
<td>74.6</td>
<td>73</td>
<td>73.6</td>
</tr>
</tbody>
</table>
Geometric Techniques

- Basic idea: Visualization of geometric transformations and projections of data
  - Scatterplot-matrices [And 72, Cle 93]
  - Parallel coordinates [Ins 85, ID 90]
  - Parallel Glyphs [Fanea:05]
  - Parallel Sets [Bendix:05]
  - Star coordinates [Kan 2000]
  - Landscapes [Wis 95]
  - Dust & Magnet [Yi 2005]
  - Projection Pursuit Techniques [Hub 85]
  - Prosection Views [FB 94, STDS 95]
  - Hyperslice [WL 93]
Recall 1-Dimensional Visualization

(1.6)
Parallel Coordinates

<table>
<thead>
<tr>
<th>Sepal Length</th>
<th>Sepal Width</th>
<th>Petal Length</th>
<th>Petal Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

The diagram shows a scatter plot with axes labeled as follows:
- Sepal Length
- Sepal Width
- Petal Length
- Petal Width

Key measurements circled on the plot include:
- Sepal Length: 5.1
- Sepal Width: 3.5
- Petal Length: 1.4
- Petal Width: 0.2
Cluster and Outlier

- **Cluster**
  - A group of data items that are similar in all dimensions.

- **Outlier**
  - A data item that is similar to FEW or No other data items.

(a) Parallel Coordinates
(b) Parallel Sets
Scatterplot Matrix

- sepal_length vs sepal_width
- sepal_length vs petal_length
- sepal_length vs petal_width
- sepal_width vs petal_length
- sepal_width vs petal_width
- petal_length vs petal_width
Scatterplot Matrix

<table>
<thead>
<tr>
<th></th>
<th>MPG</th>
<th>Cylinders</th>
<th>Horsepower</th>
<th>Weight</th>
<th>Acceleration</th>
<th>Year</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orig</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Projection Pursuit Techniques

Idea: For High-Dimensional Dataset

1. locate projections to low-dimensional space that reveal most details about dataset structure
2. extract and analyze projection structures from projections

Two general approaches: manual and automatic.
Automatic Projection Pursuit
Friedman and Tukey

Friedman and Tukey’s method

1. characterize a given projection by a numerical index that indicates the amount of structure that is present.

2. heuristically search to locate the ``interesting'' projections using the index.

3. remove detected structures from the data.

4. Iterate until there is no remaining structure detectable within the data.
Hot topics of a news collection

Landscapes

How was the figure generated?

Documents (data items)
   ↓
Keywords (dimensions)
   ↓
N-d vector for each documents
   ↓
Projection from N-d space to 2-d space
   ↓
Landscape view

Dust & Magnet


- Video
Hierarchical Techniques

Basic ideas: Visualization of the data using a hierarchical partitioning into subspaces

- Dimensional Stacking [LWW90]
- Worlds-within-Worlds [FB 90a/b]
- Treemap [Shn 92, Joh 93]
- Cone Trees [RMC 91]
- InfoCube [RG93]
Dimensional Stacking

Imagine each data item (4 attributes) as a small block. We place all blocks on a table.
Dimensional Stacking

Add grids on the table. Place the blocks in the grids according to their values of attribute1.
Dimensional Stacking

Add grids in grids. Place the blocks in the grids according to their values of attribute2.

According to values of attribute 2
Dimensional Stacking

- Add grids in grids. Place the blocks in the grids according to their values of attribute 3.
Dimensional Stacking

- Add grids in grids. Place the blocks in the grids according to their values of attribute 4.
Dimensional Stacking

- Fix one block!

Expand the block
Dimensional Stacking

Fix another block
Dimensional Stacking

Dimensional stacking!
Dimensional Stacking

visualization of oil mining data with longitude and latitude mapped to the outer x-, y- axes and ore grade and depth mapped to the inner x-, y- axes

M. Ward, Worcester Polytechnic Institute
Icon-Based Techniques

- Basic idea: visualization of the data values as features of icons
  - Chernoff-Faces [Che 73, Tuf 83]
  - Stick Figures [Pic 70, PG88]
  - Shape Coding [Bed 90]
  - Color Icons [Lev 91, KK94]
  - TileBars [Hea 95]
Profile Glyphs

Each bar encodes a variable’s value

3 data items in the iris dataset

1 data item with 4 attributes
Chernoff faces (1973, Herman Chernoff)
Star Glyphs

- Space out variables at equal angles around a circle
- Each arm encodes a variable’s value

1 data item with 4 attributes

4 data items in the iris dataset
Glyphs
Stick Figures

- The mapping
  - Two attributes - display axes
  - Others - angle and/or length of limbs
- Texture patterns in the visualization show certain data characteristics

Stick Figures

5-dim. image data from the great lake region

http://ivpr.cs.uml.edu/gallery/
G. Grinstein, University of Massachusetts at Lowell
Stick Figures

- Stick figure icon family
- Try all different mappings
  - 12X5! = 1440 pictures
- Movie
Stick Figures

The same dataset, different mapping
http://ivpr.cs.uml.edu/gallery/
G. Grinstein, University of Massachusetts at Lowell
Stick Figures

http://ivpr.cs.uml.edu/gallery/
G. Grinstein, University of Massachusetts at Lowell
More

http://ivpr.cs.uml.edu/gallery/
G. Grinstein, University of Massachusetts at Lowell
Shape Coding

- Data item - small array of fields
- Each field - one attribute value

Arrangement of attribute fields (e.g., 12-dimensional data):

1. Missing values → White
2. Low values → Grey
3. Mid Values → White
4. High Values → Black

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Shape Coding

Color Icons [Lev 91, KK94]

- Data item - color icon (arrays of color fields)
- Each field - one attribute value
- Arrangement is query-dependent (e.g., spiral)

6 dimensional dataset

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00.
Color Icons

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Pixel-Oriented Techniques

Basic idea
- Each value - one colored pixel
  (value ranges -> fixed colormap)
- Values for each attribute are presented in separate subwindows
- Values of the same data item are at the same positions of all subwindows

Keim’s tutorial notes in Infovis 00
Major Challenge

- Patterns <- the texture of the subwindows
- How to order the pixels to get informative textures?
Query-Independent Techniques

- Recursive pattern arrangements

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Query-Independent Techniques

- Recursive pattern arrangements

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Query-Dependent Techniques

- Basic idea:
  - data items \((a_1, a_2, ..., a_m)\) & query \((q_1, q_2, ..., q_m)\)
  - distances \((d_1, d_2, ... d_m)\)
  - extend distances by overall distance \((d_{m+1})\)
  - determine data items with lowest overall distances
  - map distances to color (for each attribute)
  - visualize each distance value \(d_i\) by one colored pixel

The slide is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Query-Dependent Techniques

- Spiral technique

Arrangement

The m+1 dimension (overall distance)

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Query-Dependent Techniques

- Spiral technique

The figure is taken from Dr. D. Keim’s tutorial notes in Infovis 00
Major References

- Colin Ware. Information Visualization, 2004
- Daniel Keim. Tutorial note in InfoVis 2000
- John Stasko. Course slides, Fall 2005
- Edward Tufte:
  - The Visual Display of Quantitative Information
  - Envisioning Information
  - Visual Explanation
Homework

1. Read referred papers
2. Download XMDV from http://davis.wpi.edu/~xmdv/downloadxmdv.html
   Binary Release, Microsoft Windows, Xmdv 7.0 and play with it.
3. Prepare your next class by reading papers

The paper list and course slides will be published on my webpage
www.cs.uncc.edu/~jyang13