Multivariate Data &
Tables and Graphs

CS 7450 - Information Visualization
Aug. 29, 2016
John Stasko

Learning Objectives

• Explain different types of data models
• Describe different variable types (categories)
• Define metadata
• Know when to use a table versus a graph
• Identify four fundamental types of symbolic displays
• Explain marks and mark properties
• Identify effective techniques for low-dimensional (<=3) data
• Given raw data, be able to analyze, model, and transform into tabular data
Data

- Data is taken from and/or representing some phenomena from the world
- Data models something of interest to us
- Data comes in many different forms
  - Typically, not in the way you want it

- What is available to me (in the raw)?

Example

- Cars
  - make
  - model
  - year
  - miles per gallon
  - cost
  - number of cylinders
  - weights
  - ...

Example

• Web pages

Data Models

• Often characterize data through three components
  – Objects
    Items of interest
    (students, courses, terms, ...)
  – Attributes
    Characteristics or properties of data
    (name, age, GPA, number, date, ...)
  – Relations
    How two or more objects relate
    (student takes course, course during term, ...)

Fall 2016 CS 7450
Data Tables

- We take raw data and transform it into a model/form that is more workable
- Main idea:
  - Individual items are called cases
  - Cases have variables (attributes)

Statistical Model

- Independent and Dependent variables
- Dimensions
  - Discrete, categorical info
- Measures
  - Continuous, quantitative info
## Data Table Format

<table>
<thead>
<tr>
<th></th>
<th>Case₁</th>
<th>Case₂</th>
<th>Case₃</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable₁</td>
<td>Value₁₁</td>
<td>Value₂₁</td>
<td>Value₃₁</td>
<td></td>
</tr>
<tr>
<td>Variable₂</td>
<td>Value₁₂</td>
<td>Value₂₂</td>
<td>Value₃₂</td>
<td></td>
</tr>
<tr>
<td>Variable₃</td>
<td>Value₁₃</td>
<td>Value₂₃</td>
<td>Value₃₃</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Think of as a function

\[ f(\text{case}_1) = \langle \text{Val}_{11}, \text{Val}_{12}, \ldots \rangle \]

---

## Example

<table>
<thead>
<tr>
<th>Mary</th>
<th>Jim</th>
<th>Sally</th>
<th>Mitch</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN</td>
<td>145</td>
<td>294</td>
<td>563</td>
<td>823</td>
</tr>
<tr>
<td>Age</td>
<td>23</td>
<td>17</td>
<td>47</td>
<td>29</td>
</tr>
<tr>
<td>Hair</td>
<td>brown</td>
<td>black</td>
<td>blonde</td>
<td>red</td>
</tr>
<tr>
<td>GPA</td>
<td>2.9</td>
<td>3.7</td>
<td>3.4</td>
<td>2.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

People in class
Or

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Age</th>
<th>Hair</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>145</td>
<td>23</td>
<td>brown</td>
<td>2.9</td>
</tr>
<tr>
<td>Jim</td>
<td>294</td>
<td>17</td>
<td>black</td>
<td>3.7</td>
</tr>
<tr>
<td>Sally</td>
<td>563</td>
<td>47</td>
<td>blonde</td>
<td>3.4</td>
</tr>
<tr>
<td>Mitch</td>
<td>823</td>
<td>29</td>
<td>red</td>
<td>2.1</td>
</tr>
</tbody>
</table>

People in class

Example

Baseball statistics

<table>
<thead>
<tr>
<th>Name</th>
<th>At Bats</th>
<th>Hits</th>
<th>Home Runs</th>
<th>RbA</th>
<th>Walks</th>
<th>Years in Career</th>
<th>At Career Hits</th>
<th>Career HRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy Allen</td>
<td>293</td>
<td>66</td>
<td>19</td>
<td>6</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Alan Stout</td>
<td>315</td>
<td>81</td>
<td>7</td>
<td>24</td>
<td>39</td>
<td>14</td>
<td>34438</td>
<td>8</td>
</tr>
<tr>
<td>Alex Davis</td>
<td>479</td>
<td>130</td>
<td>10</td>
<td>66</td>
<td>72</td>
<td>7</td>
<td>1524</td>
<td>457</td>
</tr>
<tr>
<td>Andre Dawson</td>
<td>491</td>
<td>141</td>
<td>30</td>
<td>65</td>
<td>70</td>
<td>11</td>
<td>16381</td>
<td>1675</td>
</tr>
<tr>
<td>Andre Galama</td>
<td>521</td>
<td>107</td>
<td>10</td>
<td>39</td>
<td>42</td>
<td>10</td>
<td>2360</td>
<td>101</td>
</tr>
<tr>
<td>Al Mango</td>
<td>594</td>
<td>169</td>
<td>4</td>
<td>74</td>
<td>51</td>
<td>11</td>
<td>44458</td>
<td>11333</td>
</tr>
<tr>
<td>Al Norman</td>
<td>198</td>
<td>37</td>
<td>1</td>
<td>23</td>
<td>6</td>
<td>2</td>
<td>214</td>
<td>42</td>
</tr>
<tr>
<td>Angel Salsa</td>
<td>299</td>
<td>73</td>
<td>0</td>
<td>34</td>
<td>24</td>
<td>7</td>
<td>501</td>
<td>108</td>
</tr>
<tr>
<td>Andre 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>
</tr>
<tr>
<td>Andre Thornton</td>
<td>461</td>
<td>92</td>
<td>17</td>
<td>49</td>
<td>66</td>
<td>13</td>
<td>5296</td>
<td>1333</td>
</tr>
<tr>
<td>Alan Freamer</td>
<td>672</td>
<td>159</td>
<td>21</td>
<td>507</td>
<td>76</td>
<td>68</td>
<td>801</td>
<td>1904</td>
</tr>
<tr>
<td>Alan Terrio</td>
<td>203</td>
<td>53</td>
<td>4</td>
<td>31</td>
<td>30</td>
<td>9</td>
<td>407</td>
<td>407</td>
</tr>
<tr>
<td>Andy Van Sty</td>
<td>419</td>
<td>113</td>
<td>13</td>
<td>48</td>
<td>61</td>
<td>4</td>
<td>1612</td>
<td>352</td>
</tr>
<tr>
<td>Alan Wiggins</td>
<td>239</td>
<td>60</td>
<td>0</td>
<td>30</td>
<td>11</td>
<td>2</td>
<td>1541</td>
<td>510</td>
</tr>
<tr>
<td>Billy Adams</td>
<td>198</td>
<td>43</td>
<td>7</td>
<td>29</td>
<td>27</td>
<td>13</td>
<td>3231</td>
<td>826</td>
</tr>
<tr>
<td>Billy Bari</td>
<td>183</td>
<td>39</td>
<td>3</td>
<td>26</td>
<td>15</td>
<td>3</td>
<td>201</td>
<td>42</td>
</tr>
<tr>
<td>Buddy Bell</td>
<td>568</td>
<td>156</td>
<td>20</td>
<td>68</td>
<td>75</td>
<td>15</td>
<td>18068</td>
<td>2273</td>
</tr>
<tr>
<td>Buddy Bower</td>
<td>190</td>
<td>46</td>
<td>2</td>
<td>24</td>
<td>6</td>
<td>5</td>
<td>479</td>
<td>102</td>
</tr>
<tr>
<td>Bruce Bristle</td>
<td>407</td>
<td>104</td>
<td>6</td>
<td>57</td>
<td>43</td>
<td>12</td>
<td>5231</td>
<td>1478</td>
</tr>
</tbody>
</table>
Variable Types

- Three main types of variables
  - N-Nominal (equal or not equal to other values)
    Example: gender
  - O-Ordinal (obeys < relation, ordered set)
    Example: fr, so, jr, sr
  - Q-Quantitative (can do math on them)
    Example: age

Alternate Characterization

- Two types of data
  - Quantitative
    Relationships between values:
    - Ranking
    - Ratio
    - Correlation
  - Categorical
    How attributes relate to each other:
    - Nominal
    - Ordinal
    - Interval
    - Hierarchical

From S. Few
Metadata

- Descriptive information about the data
  - Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn’t enough
  - Example: if variable1 is “l”, then variable3 can only be 3, 7 or 16

Data Cleaning

- Data may be missing/corrupted
  - Remove?
  - Modify?
- You may want to adjust values
  - Use inverse
  - Map nominal to ordinal/quantitative
  - Normalize values
    - Scale between 0 and 1
Nice Interactive Tool

https://www.trifacta.com/start-wrangling/

How Many Variables?

- Data sets of dimensions 1, 2, 3 are common
- Number of variables per class
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data
Representation

• What are two main ways of presenting multivariate data sets?
  – Directly (textually) → Tables
  – Symbolically (pictures) → Graphs

• When use which?

Strengths?

• Use tables when
  – The document will be used to look up individual values
  – The document will be used to compare individual values
  – Precise values are required
  – The quantitative info to be communicated involves more than one unit of measure

• Use graphs when
  – The message is contained in the shape of the values
  – The document will be used to reveal relationships among values
Effective Table Design

• See *Show Me the Numbers*

• Proper and effective use of layout, typography, shading, etc. can go a long way

• (Tables may be underused)

Example
Basic Symbolic Displays

- Graphs
- Charts
- Maps
- Diagrams

From:
1. Graph

Showing the relationships between variables’ values in a data table

![Graph showing relationships between variables]

Properties

- Graph
  - Visual display that illustrates one or more relationships among entities
  - Shorthand way to present information
  - Allows a trend, pattern or comparison to be easily comprehended
Issues

- Critical to remain task-centric
  - Why do you need a graph?
  - What questions are being answered?
  - What data is needed to answer those questions?
  - Who is the audience?

Graph Components

- Framework
  - Measurement types, scale
- Content
  - Marks, lines, points
- Labels
  - Title, axes, ticks
Quick Aside

• Other symbolic displays
  – Chart
  – Map
  – Diagram

2. Chart

• Structure is important, relates entities to each other
• Primarily uses lines, enclosure, position to link entities

Examples: flowchart, family tree, org chart, ...
3. Map

Representation of spatial relations

Locations identified by labels

4. Diagram

- Schematic picture of object or entity
- Parts are symbolic

Examples: figures, steps in a manual, illustrations,...
Some History

- Which is older, map or graph?

- Maps from about 2300 BC

- Graphs from 1600’s
  - Rene Descartes
  - William Playfair, late 1700’s

Details

- What are the constituent pieces of these four symbolic displays?

- What are the building blocks?
Visual Structures

- Composed of
  - Spatial substrate
  - Marks
  - Graphical properties of marks

Space

- Visually dominant
- Often put axes on space to assist
- Use techniques of composition, alignment, folding, recursion, overloading to
  1) increase use of space
  2) do data encodings
Marks

- Things that occur in space
  - Points
  - Lines
  - Areas
  - Volumes

Graphical Properties

- Size, shape, color, orientation...

<table>
<thead>
<tr>
<th>Expressing extent</th>
<th>Spatial properties</th>
<th>Object properties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
<td>Grayscale</td>
</tr>
<tr>
<td></td>
<td>Size</td>
<td></td>
</tr>
<tr>
<td>Differentiating marks</td>
<td>Orientation</td>
<td>Color</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shape</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Texture</td>
</tr>
</tbody>
</table>
Back to Data

- What were the different types of data sets?
- Number of variables per class
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data

Univariate Data

- Representations
What Goes Where?

• In univariate representations, we often think of the data case as being shown along one dimension, and the value in another.

Alternative View

• We may think of graph as representing independent (data case) and dependent (value) variables.

• Guideline:
  – Independent vs. dependent variables
    Put independent on x-axis
    See resultant dependent variables along y-axis
**Bivariate Data**

- **Representations**

  - Scatter plot is common
    - Two variables, want to see relationship
    - Is there a linear, curved or random pattern?

  - Each mark is now a data case

**Trivariate Data**

- **Representations**

  - 3D scatter plot is possible

  - horsepower
  - price
  - mileage
Alternative Representation

Still use 2D but have mark property represent third variable

Alternative Representation

Represent each variable in its own explicit way
Hypervariate Data

- Ahhh, the tough one
- Number of well-known visualization techniques exist for data sets of 1-3 dimensions
  - line graphs, bar graphs, scatter plots
  - We see a 3-D world (4-D with time)
- What about data sets with more than 3 variables?
  - Often the interesting, challenging ones

Multiple Views

Give each variable its own display

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
**Scatterplot Matrix**

Represent each possible pair of variables in their own 2-D scatterplot

Useful for what?
Misses what?

---

**More to Come...**

- Subsequent day will explore other general techniques for handling hypervariate data
Advice

• Take DB & IR courses
  – Learn about query languages, relational data models, datacubes, data warehouses, ...

Design Challenge

Data about dogs

<table>
<thead>
<tr>
<th>Variety</th>
<th>Group</th>
<th>Size</th>
<th>Smartness</th>
<th>Popularity</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>O</td>
<td>N</td>
<td>Q</td>
<td>Q</td>
</tr>
</tbody>
</table>

Design a representation
Dear Data

Dear Data is a year-long, analog data drawing project by Ursula Lopez and Stephanie Powers, two award-winning information designers living on different sides of the Atlantic.

By collecting and sending their personal data according to instructions by the other, they became friends.

http://www.dear-data.com
http://www.dear-data.com/all
Learning Objectives

- Explain different types of data models
- Describe different variable types (categories)
- Define metadata
- Know when to use a table versus a graph
- Identify four fundamental types of symbolic displays
- Explain marks and mark properties
- Identify effective techniques for low-dimensional (<=3) data
- Given raw data, be able to analyze, model, and transform into tabular data

Administratia

- Office hours posted
- John S.
- John T.
- Ayshwarya
Piazza

- Add yourself
- Feel free to post...

---

HW 1 Discussion

- What findings did you make?
- What was difficult?
- What help did you want?
Project

- Overview
  - Examine details on Assignments webpage

- Milestones
  - Teams & topics 2 weeks from Wednesday

- Topic ideas

Project Teams

- 3-4 people (3 preferred)
- Self-forming or designated
- Students wiki page on t-square
  - Add your info by Weds.
- Email me if you want me to pair you
  - Email header: [7450 team]
What are you Listening to?

- Represent music listening histories
- What would you want to show?
- How might you visualize it?

Nice example of a project

LastHistory

- Visualizing a person’s listening history from last.fm
- Want to support
  - Analysis
  - Reminiscing
- Potential to synchronize with photos and calendar entries from that time

Baur et al
TVCG (InfoVis) ’10
**Vis of the Day**

- Everyone will find one interesting new visualization
- Explanation on Assignments page
- Details (tumblr, your day) can be found in t-square
Reading

- *Dear Data* website

Upcoming

- Statistical Charts & Graphs

- **No class**: Labor Day
Sources Used

Few book
CMS book
Referenced articles
Marti Hearst SIMS 247 lectures
Kosslyn '89 article
A. Marcus, *Graphic Design for Electronic Documents and User Interfaces*
W. Cleveland, *The Elements of Graphing Data*