Multivariate Data & Tables and Graphs

CS 7450 - Information Visualization
Aug. 24, 2015
John Stasko

Agenda

- Data and its characteristics
- Tables and graphs
- Design principles
Data

- Data is taken from and/or representing some phenomena from the world
- Data models something of interest to us

Data Sets

- Data comes in many different forms
- Typically, not in the way you want them

- 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, ...)

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)
  - Relational: Relations between cases (not our main focus today)
**Data Table Format**

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

Think of as a function

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

**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>
</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

<table>
<thead>
<tr>
<th>Name</th>
<th>A. Hits</th>
<th>B. Home Runs</th>
<th>C. RBIs</th>
<th>D. WRC+</th>
<th>E. Years</th>
<th>F. At Bat</th>
<th>G. At Home</th>
<th>H. Career HRs</th>
<th>I. Career HRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy</td>
<td>291</td>
<td>36</td>
<td>29</td>
<td>14</td>
<td>1</td>
<td>293</td>
<td>66</td>
<td>36</td>
<td>293</td>
</tr>
<tr>
<td>Jim</td>
<td>315</td>
<td>37</td>
<td>39</td>
<td>14</td>
<td>1</td>
<td>315</td>
<td>61</td>
<td>39</td>
<td>315</td>
</tr>
<tr>
<td>Sally</td>
<td>479</td>
<td>136</td>
<td>66</td>
<td>72</td>
<td>3</td>
<td>479</td>
<td>136</td>
<td>66</td>
<td>479</td>
</tr>
<tr>
<td>Mitch</td>
<td>496</td>
<td>141</td>
<td>30</td>
<td>65</td>
<td>11</td>
<td>496</td>
<td>141</td>
<td>30</td>
<td>496</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Baseball statistics
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
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

Example
Basic Symbolic Displays

• Graphs
• Charts
• Maps
• Diagrams


1. Graph

Showing the relationships between variables’ values in a data table
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>Position</td>
<td>Grayscale</td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Differentiating marks</th>
<th>Orientation</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<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

Tukey box plot

Bill

0 20

Middle 50%

Mean

What Goes Where?

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

Line graph

Bar graph

Y-axis is quantitative variable

See changes over consecutive values

Y-axis is quantitative variable

Compare relative point values
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

  Each mark is now a data case

  Two variables, want to see relationship

  Is there a linear, curved or random pattern?
Trivariate Data

- Representations

3D scatter plot is possible

```
price

horsepower

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?
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>N</td>
<td>N</td>
<td>O</td>
<td>N</td>
<td>Q</td>
<td>Q</td>
</tr>
</tbody>
</table>

Design a representation

More to Come...

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

Back to Graphs

- Design guidance
  - Few provides many helpful principles to design effective graphs
Few’s Selection & Design Process

- Determine your message and identify your data
- Determine if a table, or graph, or both is needed to communicate your message
- Determine the best means to encode the values
- Determine where to display each variable
- Determine the best design for the remaining objects
  - Determine the range of the quantitative scale
  - If a legend is required, determine where to place it
  - Determine the best location for the quantitative scale
  - Determine if grid lines are required
  - Determine what descriptive text is needed
- Determine if particular data should be featured and how

Some examples...

Points, Lines, Bars, Boxes

- Points
  - Useful in scatterplots for 2-values
  - Can replace bars when scale doesn’t start at 0
- Lines
  - Connect values in a series
  - Show changes, trends, patterns
  - Not for a set of nominal or ordinal values
- Bars
  - Emphasizes individual values
  - Good for comparing individual values
- Boxes
  - Shows a distribution of values
Vertical vs. Horizontal Bars

- Horizontal can be good if long labels or many items

Multiple Bars

- Can be used to encode another variable
Multiple Graphs

- Can distribute a variable across graphs too

Sometimes called a trellis display

Examples
Before

You want to present quantitative sales performance data for the 4 regions of your company for the four quarters of the year

After?
Before

Fall 2015  CS 7450  63

After?

Fall 2015  CS 7450  64
Before

After?
**Book Recommendation**

*Show Me the Numbers*
Designing Tables and Graphs to Enlighten

Loaded with examples of how to redesign ineffective tables and graphs

**Advice**

- Take DB & IR courses
  - Learn about query languages, relational data models, data cubes, data warehouses, ...

Fall 2015
CS 7450
67
Administratia

- Our second TA added, Iulian Radu

- Office hours being posted

HW 1 Discussion

- What findings did you make?
- What was difficult?
- What help did you want?
HW 2

- Table and graph design
- Given two (Excel) data sets, design a table and graph for the data, respectively
- Due next Monday

Project

- Start thinking about topics
- Form teams
Upcoming

• S. Few’s Design Guidance
  – Reading:
    *Now You See It*, chapters 5-12

• Multivariate Visual Representations 1
  – Reading:
    Inselberg ’97

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*