Multivariate Data &
Tables and Graphs

CS 4460/7450 - Information Visualization
Jan. 13, 2009
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
      - (students 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>Dimensions</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

$$f(\text{case₁}) = <\text{Val₁₁}, \text{Val₂₁}, \ldots>$$

---

## Example

<table>
<thead>
<tr>
<th></th>
<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>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>23</td>
<td>17</td>
<td>47</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Hair</td>
<td>brown</td>
<td>black</td>
<td>blonde</td>
<td>red</td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>2.9</td>
<td>3.7</td>
<td>3.4</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

People in class
### People in class

<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>

### Example

#### Baseball statistics

<table>
<thead>
<tr>
<th>Name</th>
<th>Bats</th>
<th>Pos</th>
<th>G</th>
<th>AB</th>
<th>H</th>
<th>2B</th>
<th>3B</th>
<th>HR</th>
<th>SB</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy Adams</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Ace Addy</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Ann Davis</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Andy Carson</td>
<td>R</td>
<td>1B</td>
<td>176</td>
<td>434</td>
<td>52</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>.203</td>
</tr>
<tr>
<td>Abby Johnson</td>
<td>R</td>
<td>3B</td>
<td>231</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Al Everest</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Aliexpress</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Amy Adams</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Andy Adams</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Ace Addy</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Ann Davis</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Andy Carson</td>
<td>R</td>
<td>1B</td>
<td>176</td>
<td>434</td>
<td>52</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>.203</td>
</tr>
<tr>
<td>Abby Johnson</td>
<td>R</td>
<td>3B</td>
<td>231</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Al Everest</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Aliexpress</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Amy Adams</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Andy Adams</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Ace Addy</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Ann Davis</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Andy Carson</td>
<td>R</td>
<td>1B</td>
<td>176</td>
<td>434</td>
<td>52</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>.203</td>
</tr>
<tr>
<td>Abby Johnson</td>
<td>R</td>
<td>3B</td>
<td>231</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Al Everest</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Aliexpress</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Amy Adams</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Andy Adams</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Ace Addy</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Ann Davis</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Andy Carson</td>
<td>R</td>
<td>1B</td>
<td>176</td>
<td>434</td>
<td>52</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>.203</td>
</tr>
<tr>
<td>Abby Johnson</td>
<td>R</td>
<td>3B</td>
<td>231</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Al Everest</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Aliexpress</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Amy Adams</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Andy Adams</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Ace Addy</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Ann Davis</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Andy Carson</td>
<td>R</td>
<td>1B</td>
<td>176</td>
<td>434</td>
<td>52</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>.203</td>
</tr>
<tr>
<td>Abby Johnson</td>
<td>R</td>
<td>3B</td>
<td>231</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Al Everest</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Aliexpress</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</td>
</tr>
<tr>
<td>Amy Adams</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Andy Adams</td>
<td>R</td>
<td>LF</td>
<td>228</td>
<td>539</td>
<td>238</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>.441</td>
</tr>
<tr>
<td>Ace Addy</td>
<td>R</td>
<td>CF</td>
<td>234</td>
<td>520</td>
<td>120</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>.231</td>
</tr>
<tr>
<td>Ann Davis</td>
<td>R</td>
<td>RF</td>
<td>220</td>
<td>469</td>
<td>81</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.175</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 variable 1 is “l”, then variable 3 can only be 3, 7 or 16

**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
Many Examples

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></th>
<th>Spatial properties</th>
<th>Object properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expressing extent</td>
<td>Position</td>
<td>Grayscale</td>
</tr>
<tr>
<td></td>
<td>Size</td>
<td></td>
</tr>
<tr>
<td>Differentiating</td>
<td>Orientation</td>
<td>Color</td>
</tr>
<tr>
<td>marks</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]

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

- 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

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

A  B  C  D  E
1  4  1  8  3  5
2  6  3  4  2  1
3  5  7  2  4  3
4  2  6  3  1  5

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

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

*S Few*

“Effectively Communicating Numbers”

http://www.perceptualedge.com/articles/Whitepapers/Communicating_Numbers.pdf

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

Loaded with examples of how to redesign ineffective tables and graphs

---

**Administrativa**

- HW 1 due Thursday
- Office hours posted
- Lecture slides availability
### Upcoming

- **Case Studies**
  - Reading:
    - Weaver paper

- **Good day for stimulating project ideas**

### 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*