Introduction to Information Visualization
The Value of Information Visualization

At hundreds of words: “A picture is worth a thousand words” example shown in Figure 1. Part (a) shows a spreadsheet with data for the 50 states and the District of Columbia in the U.S. Also shown are the percentage of citizens of each state with a college degree and the per capita income of the states’ citizens.

Given just the spreadsheet, answering a question such as, “Which state has the highest average income?” is not too difficult. A simple scan of the income column likely will produce the correct answer in a few seconds. More complex questions can be quite challenging given just the data, however. For example, are the college degree percentage and income correlated? If they are correlated, are there particular states that are outliers to the correlation? These questions are much more difficult to answer using only the spreadsheet.

Now, let us turn to a graphical visualization of the data. If we simply draw the data in a scatterplot as shown in part (b), the questions now become much easier to answer. Specifically, there does appear to be an overall correlation between the two attributes and states such as Nevada and Utah are outliers on the correlation. The simple act of plotting the spreadsheet data in this more meaningfully communicative form makes these kinds of analytic queries easier to answer correctly and more rapidly.

Note that the spreadsheet itself is a visual representation of the data that facilitates queries as well. Consider how difficult the three questions would be if the data for each state was recorded on a separate piece of paper or webpage. Or worse yet, what if the data values were read to you and you had to answer...
---

- mean of the x values = 9.0
- mean of the y values = 7.5
- equation of the least-squared regression line is: \( y = 3 + 0.5x \)
- sums of squared errors (about the mean) = 110.0
- regression sums of squared errors (variance accounted for by x) = 27.5
- residual sums of squared errors (about the regression line) = 13.75
- correlation coefficient = 0.82
- coefficient of determination = 0.67.

Introduction

• Problem
  • How do we make sense of small data? and big data?

• Solution
  • Take advantage of the human visual system
  • Convert data into graphical form

• Issues
  • How do we convert abstract data into graphical form?
  • Are visualization better than other methods?
Motivation

• Data Explosion: estimated info added to digital universe each year will soon approach 1 ZB (zettabyte)

• 1000000000000000000000 (10^21) bytes

• From: http://www.emc.com/digital_universe viewed December 8, 2008
Motivation

• 6 million FedEx transactions per day [http://www.fedex.com/us/about/today/companies/corporation/facts.html]

• Average of 98 million Visa credit-card transactions per day in 2005 [http://www.corporate.visa.com/md/nr/press278.jsp]

• Average of 5.4 petabytes of data crosses AT&T’s network per day [http://att.sbc.com/gen/investor-relations?pid=5711]

• Average of 610 to 1110 billion e-mails worldwide per year (based on estimates in 2000) [http://www2.sims.berkeley.edu/research/projects/how-much-info/internet.html]
Purpose of Visualization

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

• “The purpose of visualization is insight, not pictures”

• Insight: discovery, decision making, explanation
  • Visuals help us think
  • Provide a frame of reference, a temporary storage area

• External cognition:
  • Role of external world in thinking and reason
Information Visualization

“the use computer-supported, interactive, visual representations of abstract data to amplify cognition”


• What kinds of data?
  • Information that does not have a direct physical correspondence

• How is it different from Scientific Visualization?
  • SciVis relates to and represents something physical or geometric
A Model of Visualization

Source: Jarke J. van Wijk, The Value of Visualization (2005)
Historical Highlights

• 2nd. century: tabular data in Egypt
• 17th century: Descartes’ two-dimensional graphs for math
• 18th century: William Playfair (bar graph, line graphs for quantity change in time, pie-chart)
• 1913: First graphing college course on Iowa State Univ.
• 1977: John Tukey’s book “Exploratory Data Analysis”
• 1984: Macintosh, first affordable computer with graphics as a mode of interaction
• 1999: “Readings in Information Visualization: Using Vision to Think”, representations of abstract information emerged as a distinct area of study from representations of physical phenomena.

“A key challenge in information visualization is designing a cognitively useful spatial mapping of a dataset that is not inherently spatial and accompanying the mapping by interaction techniques that allow people to intuitively explore the dataset. Information visualization draws on the intellectual history of several traditions, including computer graphics, human-computer interaction, cognitive psychology, semiotics, graphic design, statistical graphics, cartography, and art.”

http://conferences.computer.org/infovis/
Visualization Success Stories
Power of Visualization

- 1. Start out going Southwest on ELLSWORTH AVE towards BROADWAY by turning right.
- 2. Turn RIGHT onto BROADWAY.
- 3. Turn RIGHT onto QUINCY ST.
- 4. Turn LEFT onto CAMBRIDGE ST.
- 5. Turn SLIGHT RIGHT onto MASSACHUSETTS AVE.
- 6. Turn RIGHT onto RUSSELL ST.
Power of Visualization

London Tube Map by Beck
(Topological vs Geographical)
Power of Visualization

Napolean’s March by Minard (6 variables represented)
Data Types

• Quantitative (e.g. age: 33, 45, 18)
• Ordered (e.g. age group: young, adult, eld)
• Categorical (e.g. continent: South America, North America, Europe)
• Relational Data (e.g. social graph, hierarchies)

How should we visually encode these?
Visual Attributes or Channels

- Examples: spatial position, color, size, shape, orientation.

Image Source: GapMinder.
Figure 27.5. Our ability to perceive information encoded by a visual channel depends on the type of data used, from most accurate at the top to least at the bottom. Redrawn and adapted from (Mackinlay, 1986).
Pre-Attentive Attributes

**FIGURE 5:** Preattentive attributes of visual perception most applicable to data presentation.
Interaction Principles

- Overview first, zoom and filter, details on demand. (Shneiderman 1996).
- Account for Interactivity Costs
- Use Animations with Care
Data Reduction

- Overviews and Aggregation
- Filtering and Navigation
- Focus + Context
- Dimensionality Reduction

Plate L. Dimensionality reduction with the Glimmer multidimensional scaling approach shows clusters in a document dataset (Ingram et al., 2009), © 2009 IEEE. (See also Figure 27.19.)

Multiple Views, Brushing and Linking

Plate XLVIII. The Improvise toolkit was used to create this multiple-view visualization. Image courtesy Chris Weaver. (See also Figure 27.16.)
Some Techniques
Multivariate Data: Scatterplot Matrix
Multivariate Data: Chernoff Faces

• Chernoff Faces

Life in Los Angeles

The Distribution of Voting, Housing, Employment and Industrial Compositions in the 1983 General Election.

Facial features indicate the social & economic characteristics of the constituencies; colour shows the proportions of the vote for the parties.
Multivariate Data: Star Plots

- Star Plots

Connecticut  New Hampshire  Pennsylvania
Maine  New Jersey  Rhode Island
Massachusetts  New York  Vermont
Multivariate Data: Parallel Coordinates

Image Source: http://mbostock.github.com/protovis/ex/cars.html
Hierarchical Data: Standard Trees

- Standard Trees

Graphviz
Hierarchical Data: Standard Trees

- Standard Trees
Hierarchical Data: Phylogenetic Trees
Hierarchical Data: Radial Trees

- Radial Trees
Hierarchical Data: Hyperbolic Trees

- Hyperbolic Trees
Hierarchical Data: Treemaps
Hierarchical Data: Treemaps

- Tree Maps: Layout matters

Slice and Dice  
Cluster  
Squarify  
Pivot by Middle  
Pivot by Size  
Strip
Graph Data: NodeTrix

Fig. 1: NodeTrix Representation of the largest component of the InfoVis Co-authorship Network
Fig. 7: NodeTrix visualization of the information visualization field. This is the largest connected component extracted from the dataset used in the Infovis’04 Contest available at http://www.cs.umd.edu/hcil/iv04contest/. We manually removed a couple of remaining duplicated authors. Colors on axes of matrices represent the number of citations of each author. Color intensity within the matrices represents the strength of each collaboration.
Map Data

- Cartograms

http://www-personal.umich.edu/~mejn/election/2008/
Word Trees

Word Trees

Time Series Visualization

Image Source: Jeffrey Heer and Maneesh Agrawala. Multi-Scale Banking to 45º. (2006)
Calendar Visualization

**Figure 1.** Power demand by ECN, displayed as a function of hours and days

The End