Visualization and Data Mining

Outline

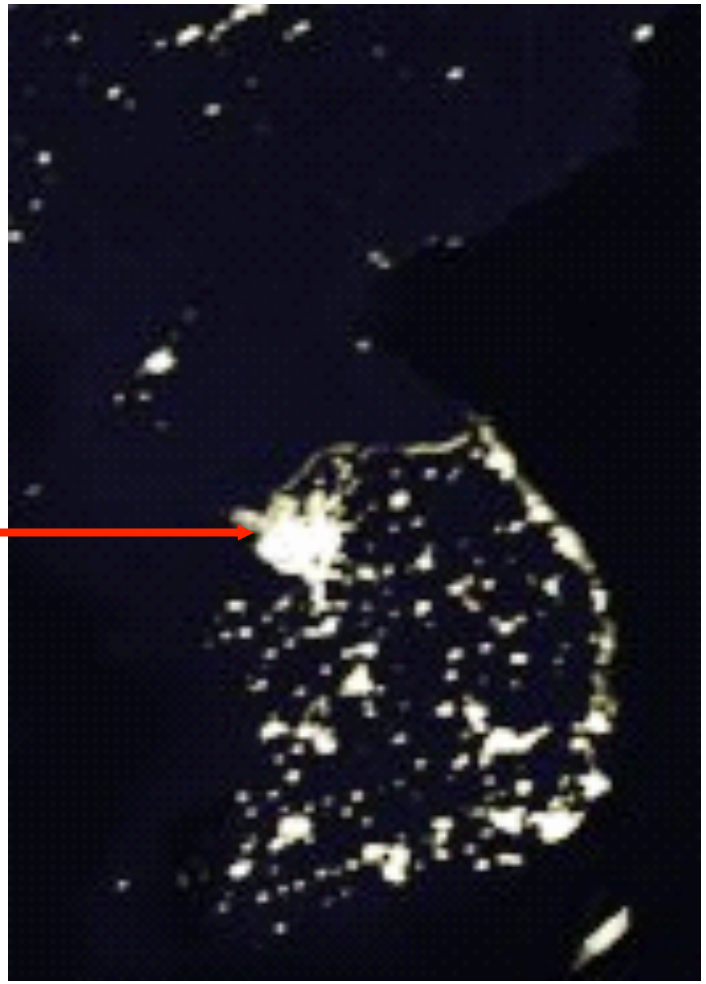
- Representing data in 1,2, and 3-D
- Time series data
- Spatial data
- Network and graph

Asia at night

South and North Korea at night

North Korea
Notice how dark it is

Seoul,
South Korea

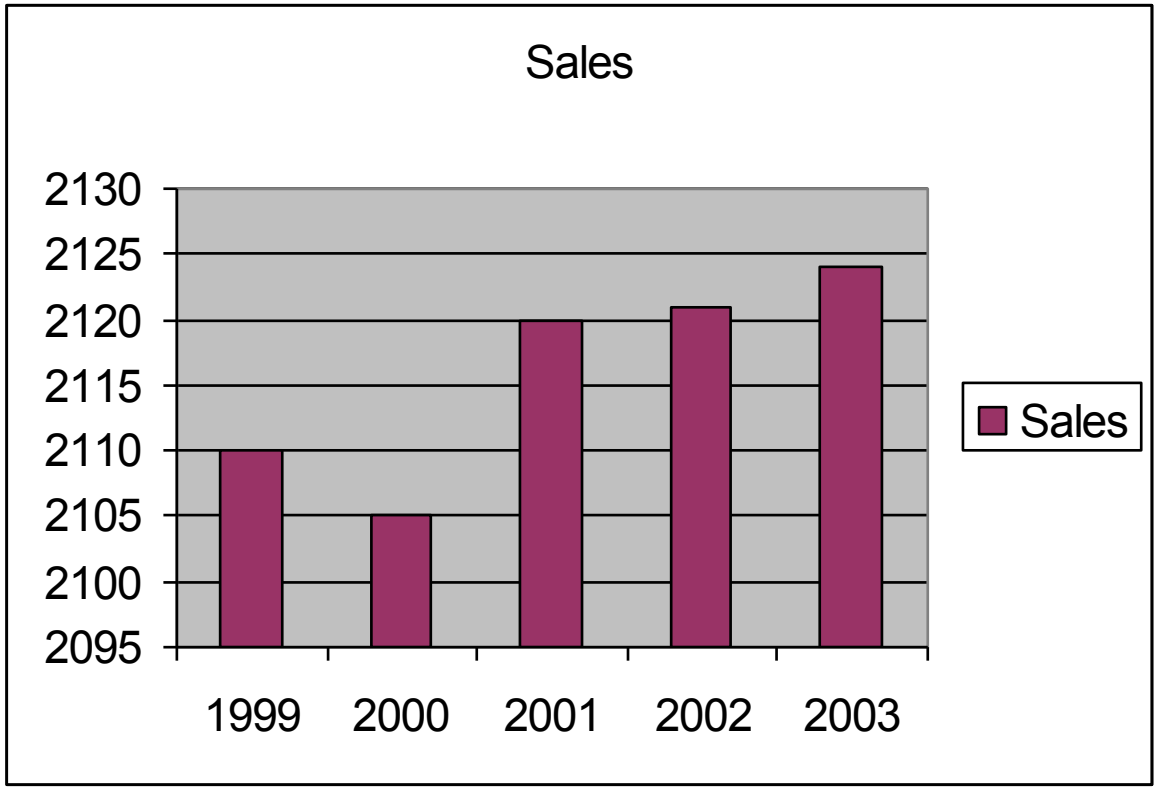


Visualization Role

- Support interactive exploration
- Help in result presentation
- Disadvantage: requires human eyes
- Can be misleading

Bad Visualization: Spreadsheet

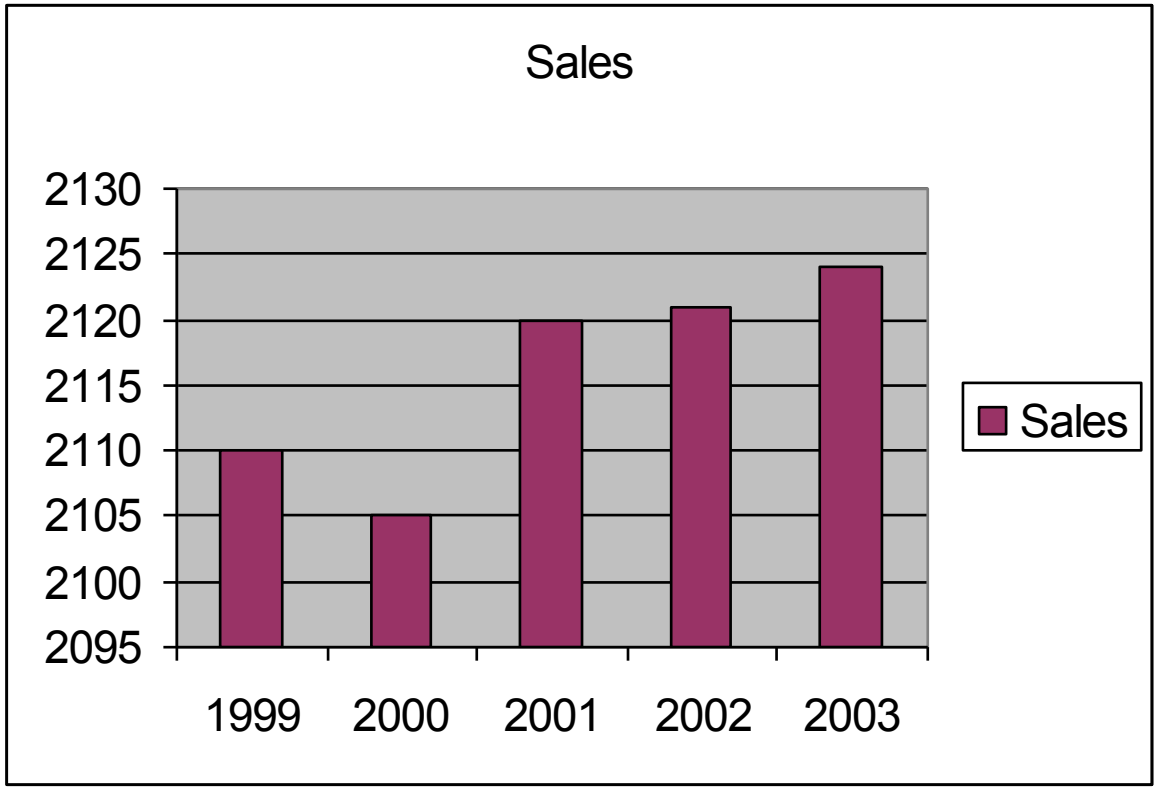
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



What is wrong with this graph?

Bad Visualization: Spreadsheet with misleading Y –axis

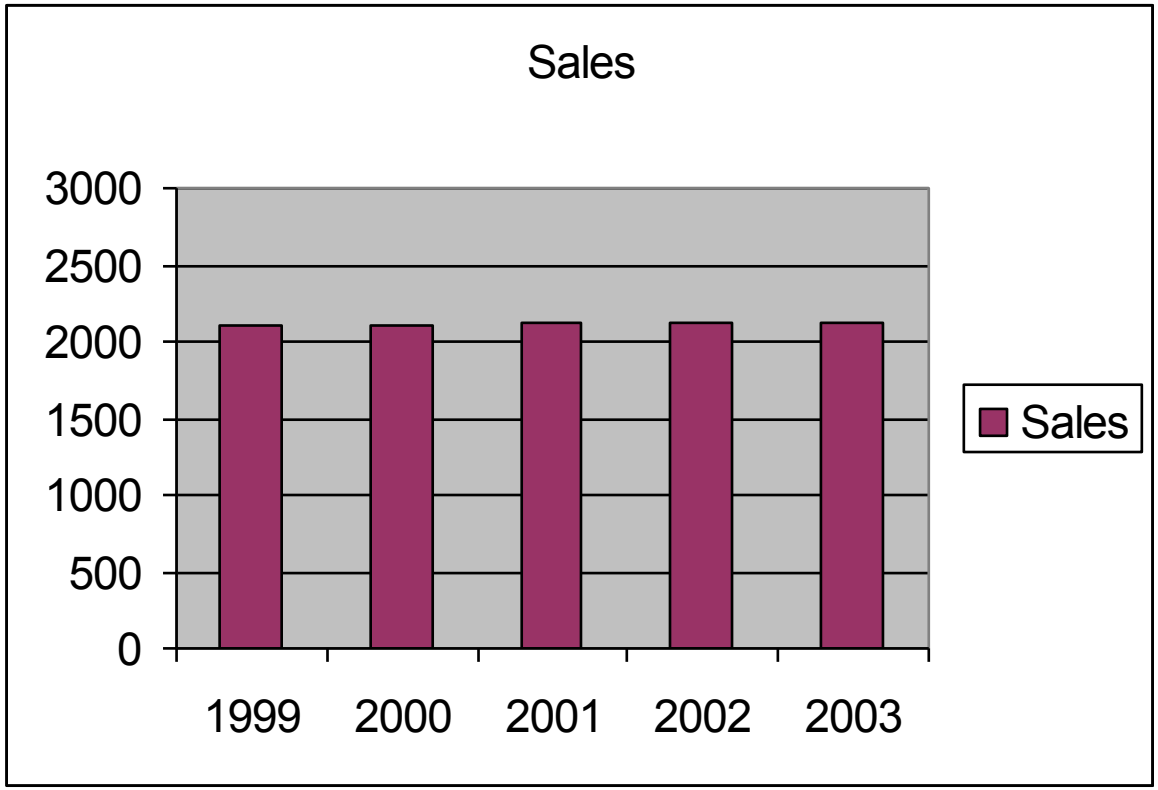
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



Y-Axis scale gives **WRONG** impression of big change

Better Visualization

Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



Axis from 0 to 2000 scale gives
correct impression of small change

Principles of Graphical Excellence

- Give the viewer
  - the greatest number of ideas
  - in the shortest time
  - with the least ink in the smallest space.

- Tell the truth about the data!

Visualization Methods

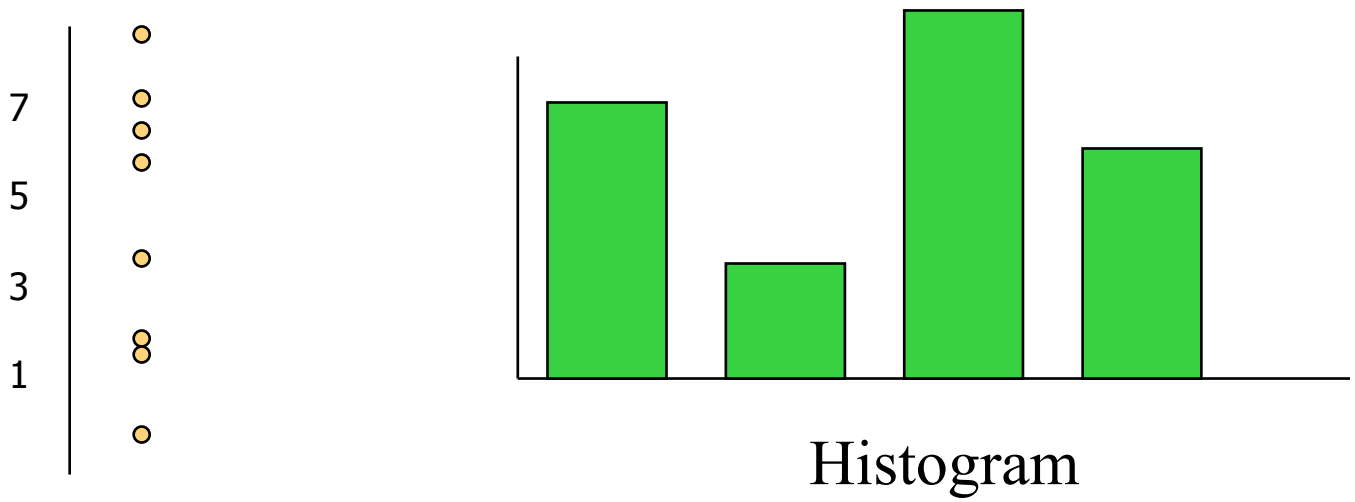
- Visualizing in 1-D, 2-D and 3-D
- Different methods are available for visualization of data based on type of data, where data can be
  - Univariate
  - Bivariate
  - Multivariate

Univariate data

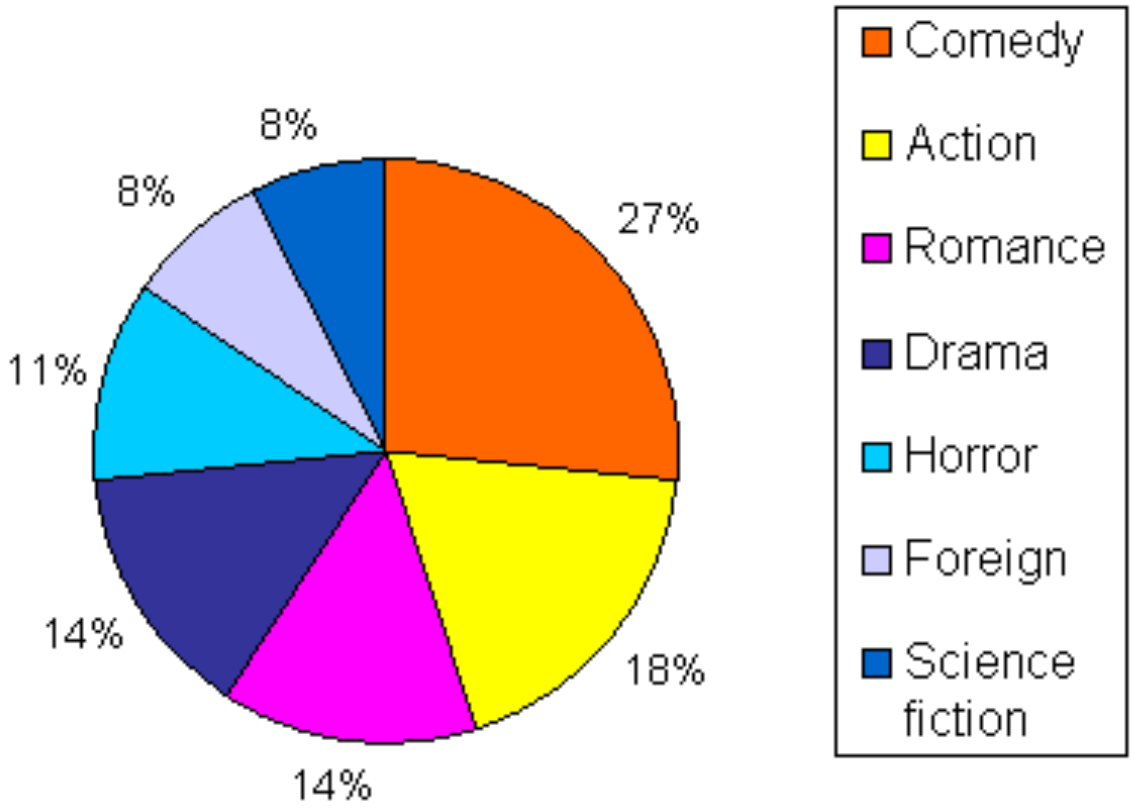
- Measurement of single quantitative variable
- Characterize distribution
- Represented using following methods
  - Histogram
  - Pie Chart

1-D (Univariate) Data

- Representations



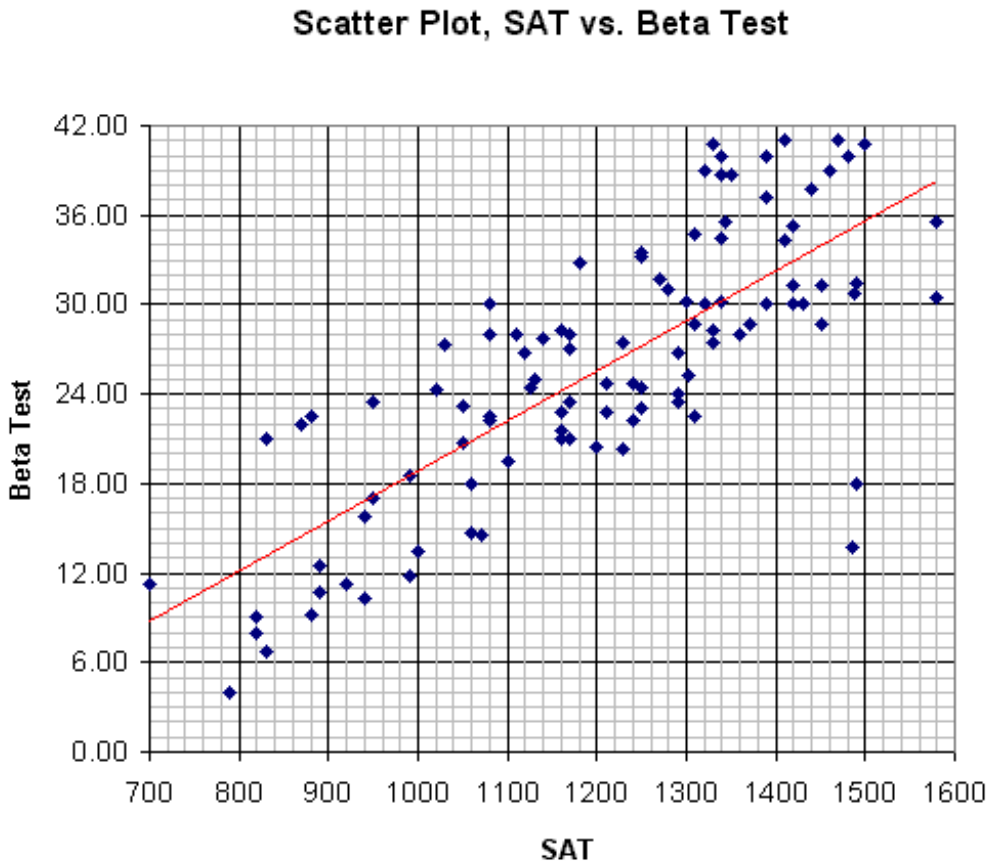
Pie Chart


8%
8%
27%
11%
14%
18%
14%
Comedy
Action
Romance
Drama
Horror
Foreign
Science fiction

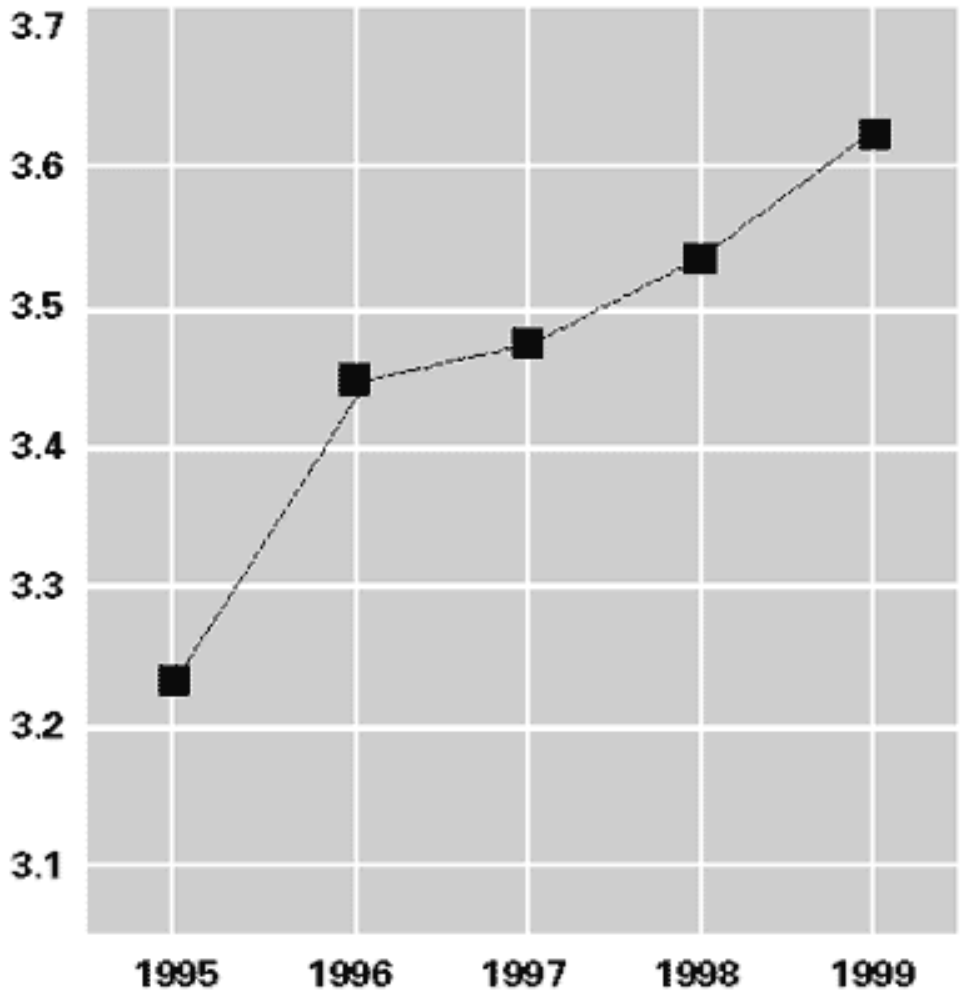
Bivariate Data

- Constitutes of paired samples of two quantitative variables
- Variables are related
- Represented using following methods
  - Scatter plots
  - Line graphs

Scatter plots



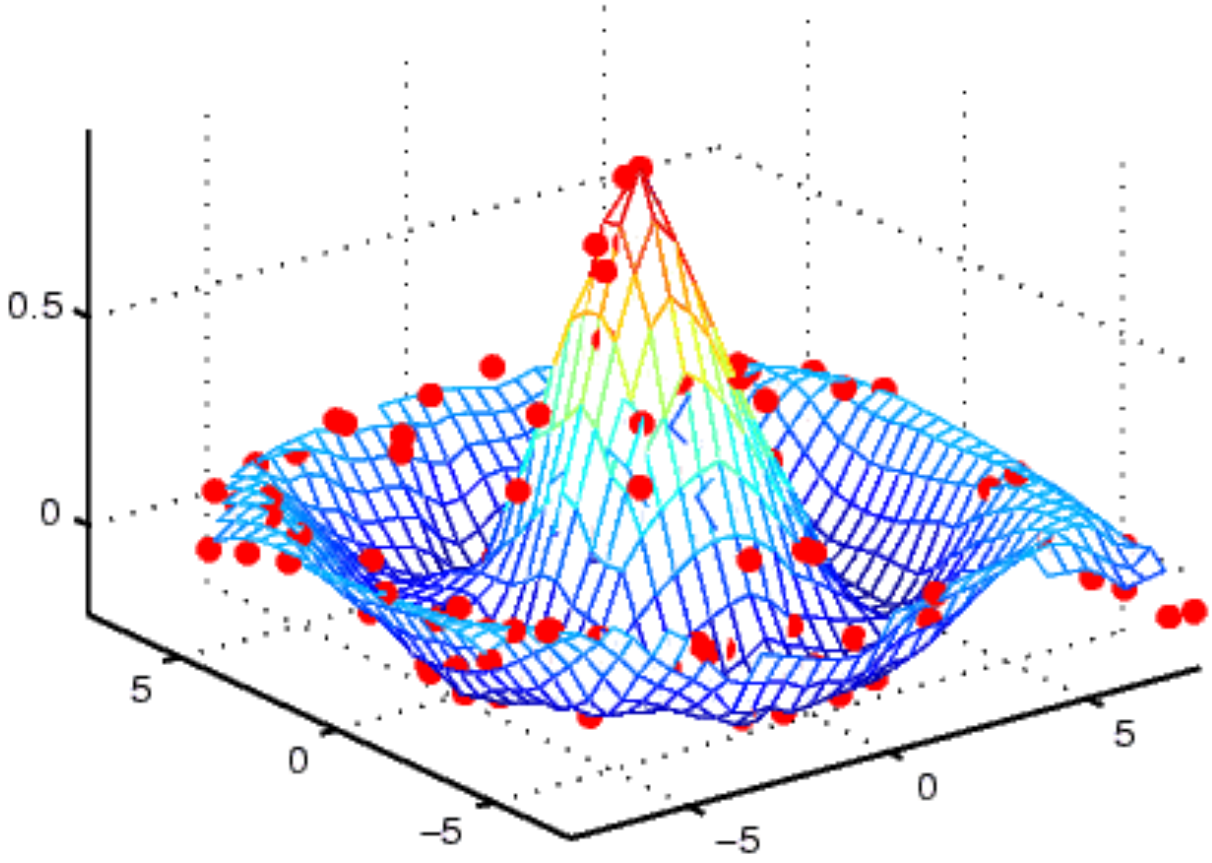
Line graphs



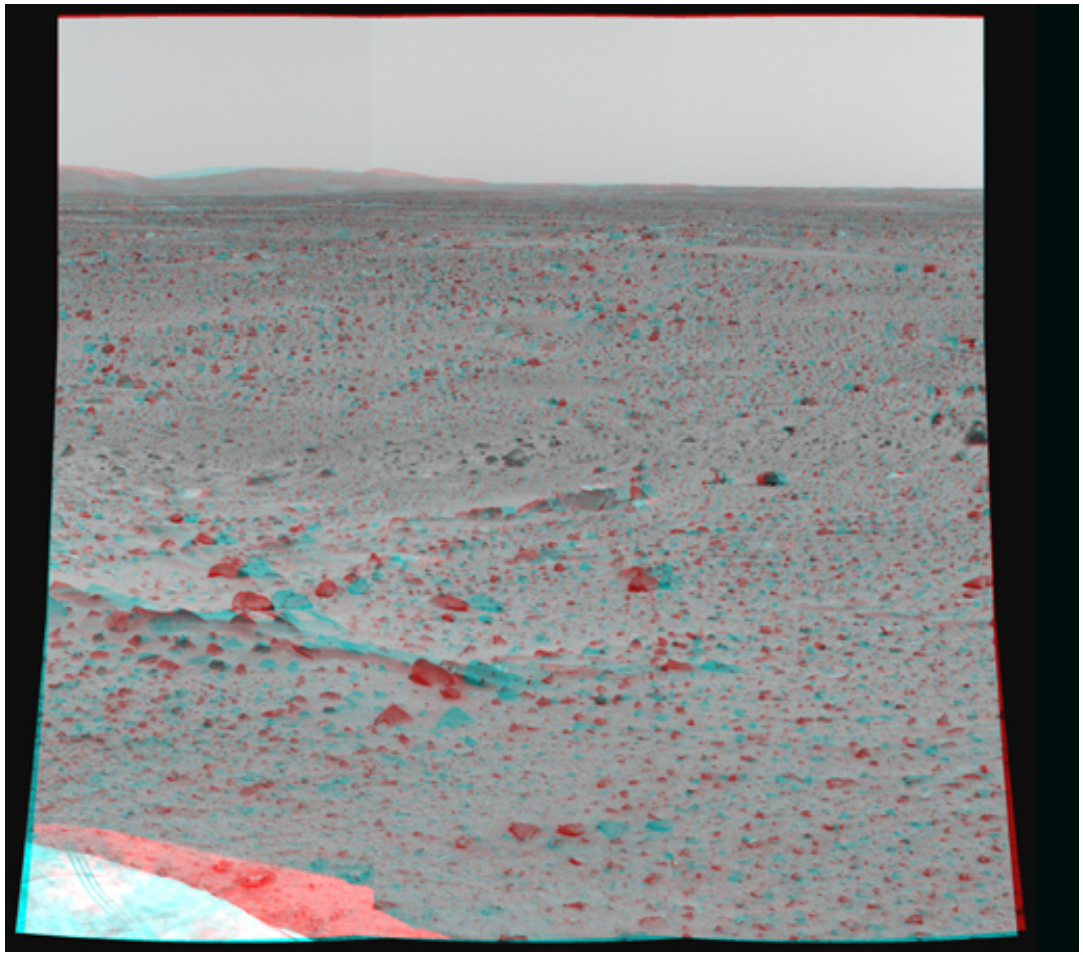
Multivariate Data

- Multi dimensional representation of multivariate data
- Represented using following methods
  - 3-D projection

3-D Data (projection)



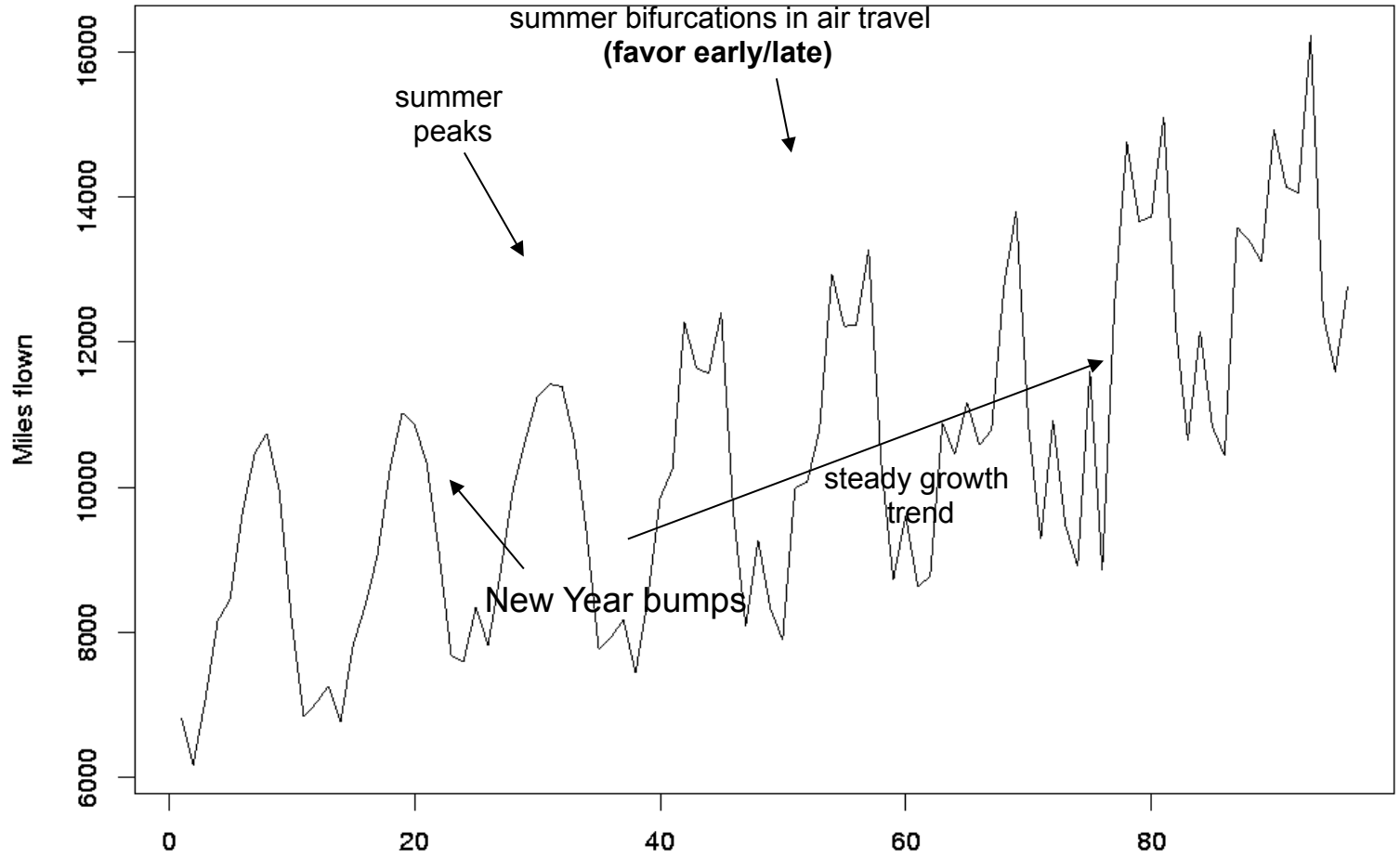
3-D image
(requires 3-D blue and red glasses)



Taken by Mars Rover Spirit, Jan 2004

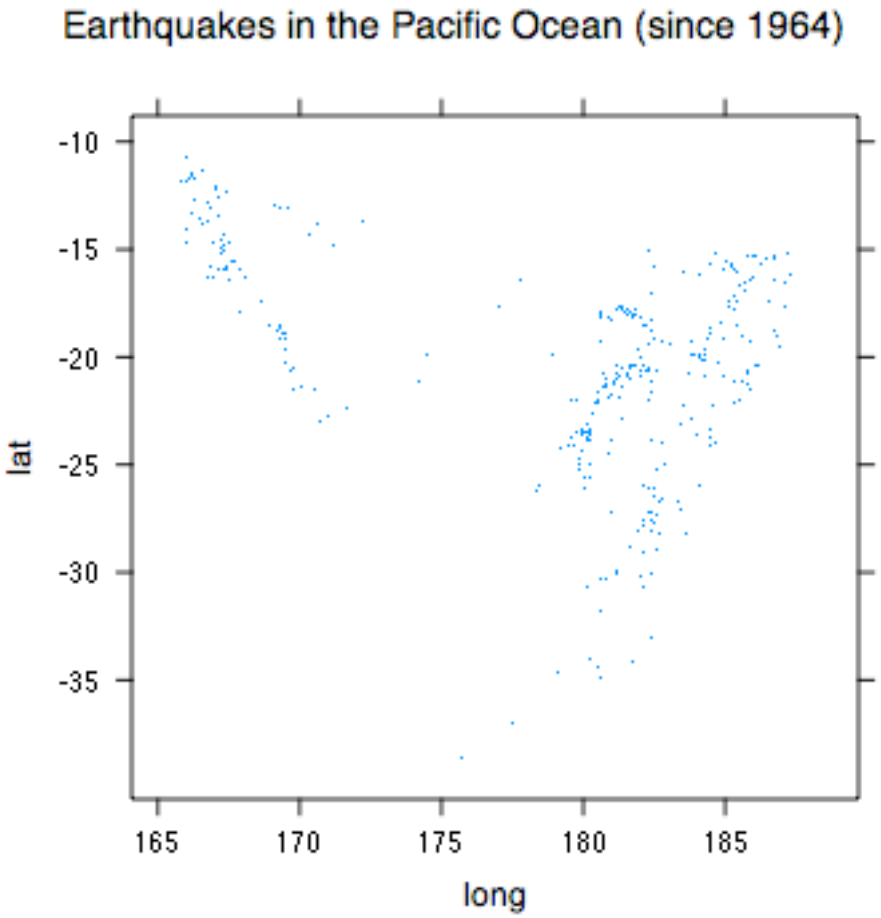
Time Series

If your data has a temporal component, be sure to exploit it

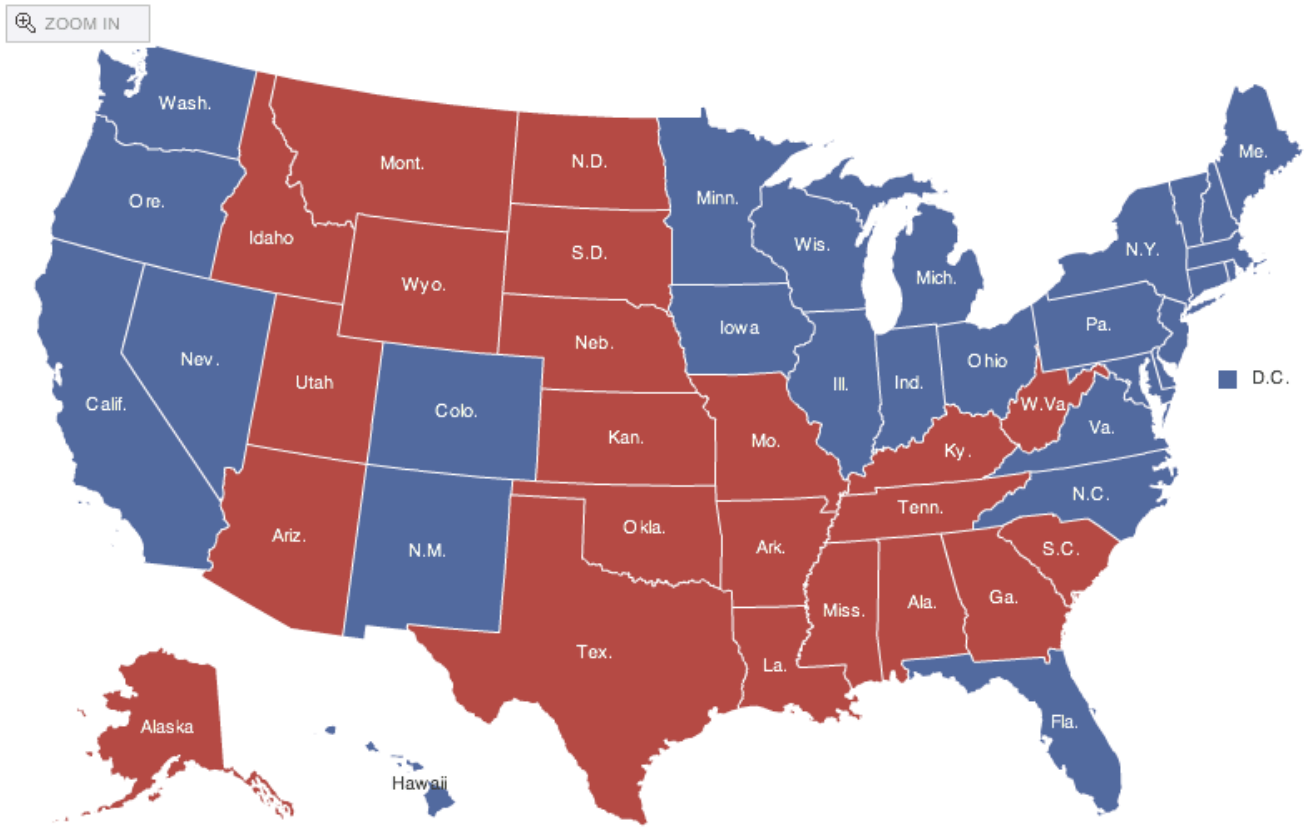


Spatial Data

- If your data has a geographic component, be sure to exploit it
- Data from cities/states/ zip cods – easy to get lat/long
- Can plot as scatterplot



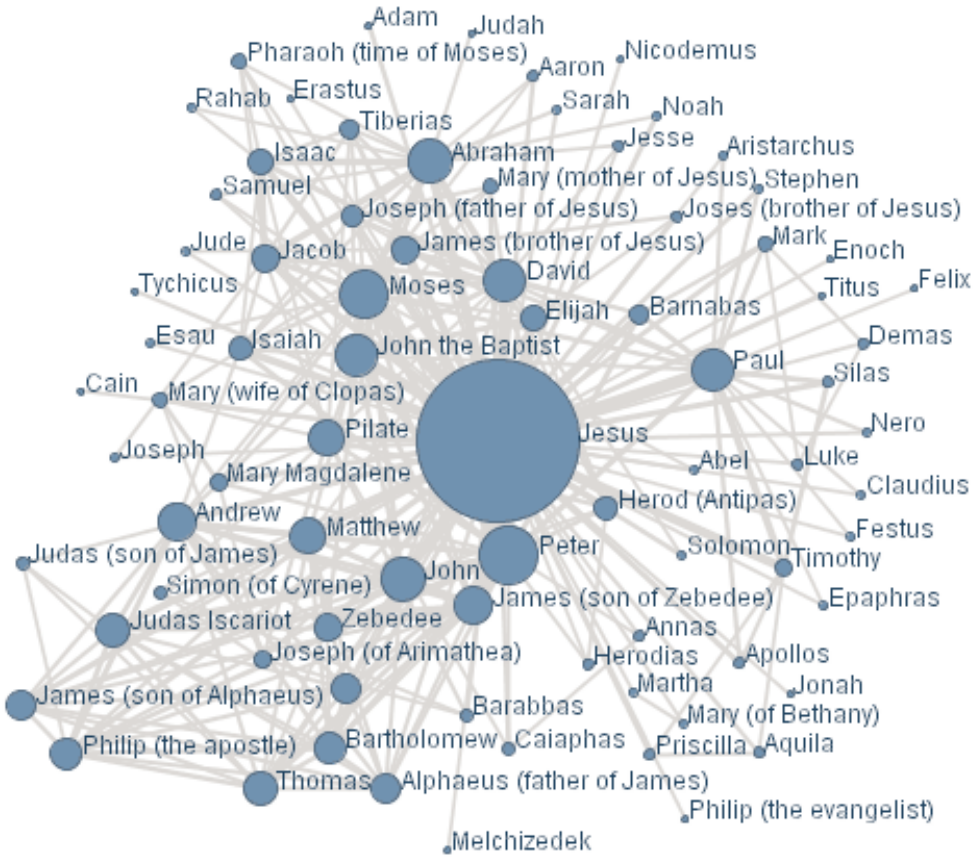
Spatial data: choropleth Maps



- Maps using color shadings to represent numerical values are called chloropleth maps
- http://elections.nytimes.com/2008/results/president/map.html

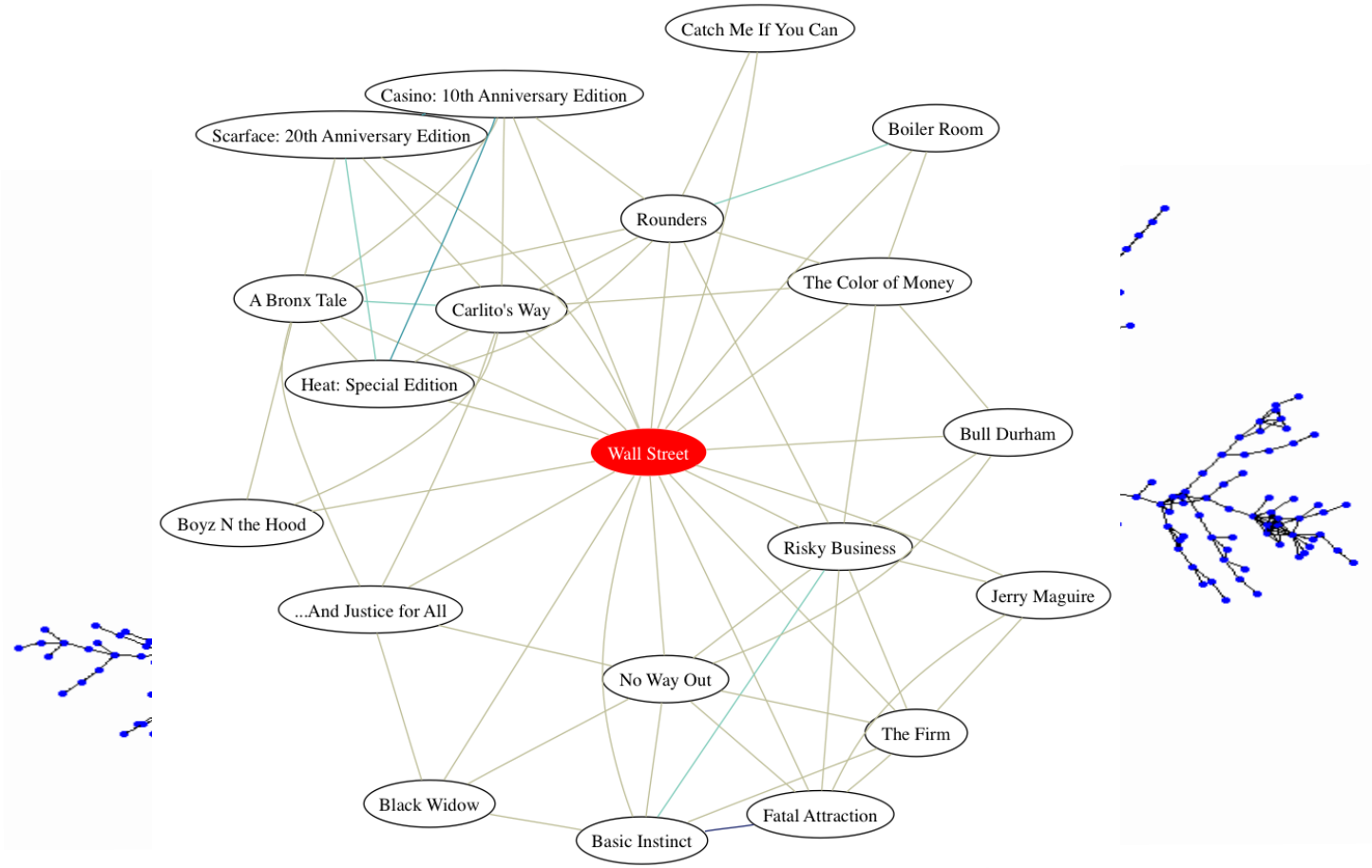
Networks and Graphs

- Visualizing networks is helpful, even if is not obvious that a network exists



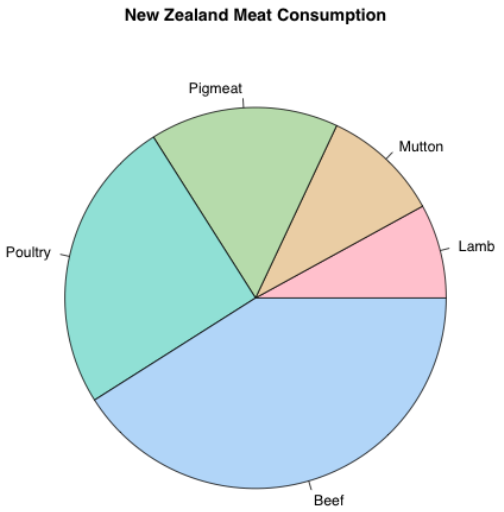
Network Visualization

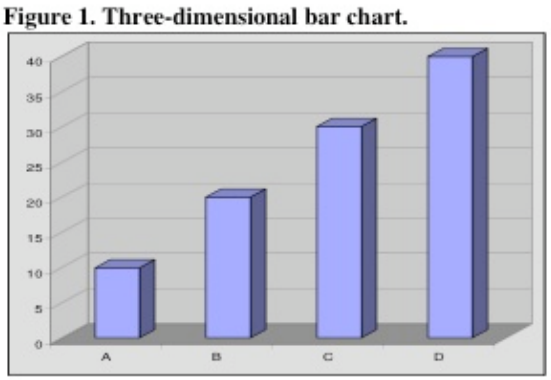
- Graphviz (open source software) is a nice layout tool for big and small graphs

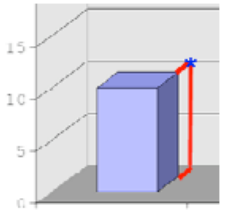


What's missing?

- pie charts
  - very popular
  - good for showing simple relations of proportions
  - Human perception not good at comparing arcs
  - barplots, histograms usually better (but less pretty)
- 3D
  - nice to be able to show three dimensions
  - hard to do well
  - often done poorly
  - 3d best shown through "spinning" in 2D
    - uses various types of projecting into 2D
    - http://www.stat.tamu.edu/~west/bradley/

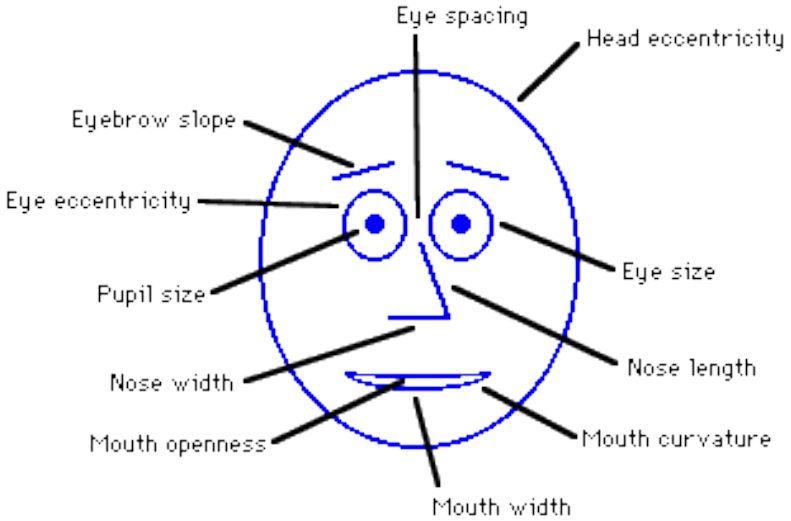






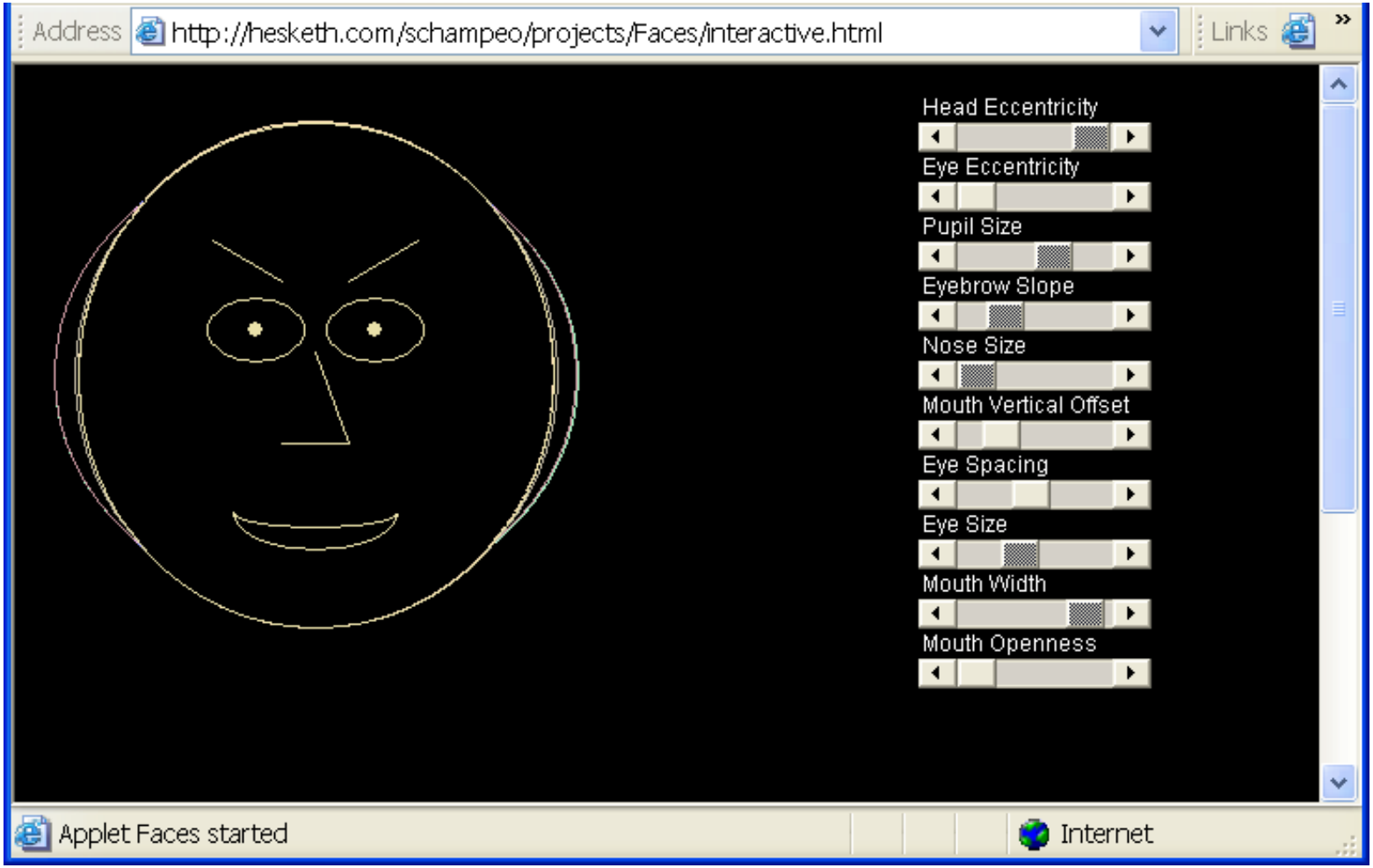
Chernoff Faces

Encode different variables' values in characteristics of human face



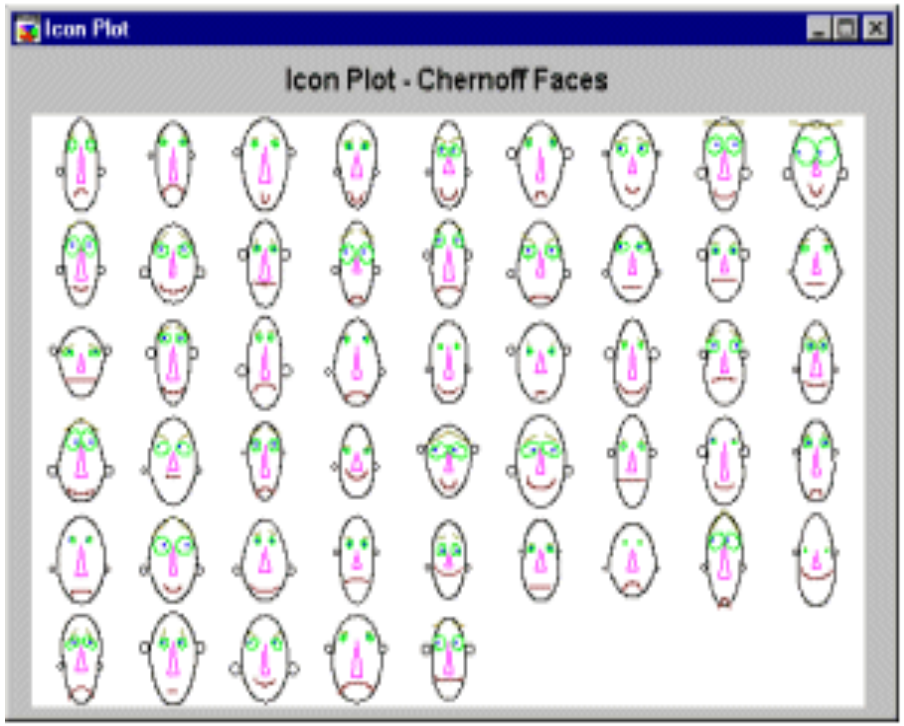
Cute applets: `http://www.cs.uchicago.edu/~wiseman/chernoff/`
`http://hesketh.com/schampeo/projects/Faces/chernoff.html`

Interactive Face



Chernoff's Faces

- described by ten facial characteristic parameters: head eccentricity, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, eye spacing, eye size, mouth length and degree of mouth opening
- Much derided in statistical circles



Visualization software

Free and Open-source

- Ggobi
- Xmdv

- Many more - see www.KDnuggets.com/software/visualization.html

Visualization Summary

- Many methods
- Visualization is possible in more than 3-D
- Aim for graphical excellence
- Method should be chosen depending on the data and your need