Introduction to Information Visualization

Kai Li
Computer Science Department
Princeton University

About This Talk

- What is information visualization
- Principles of graphical excellence
- Principles of integrity
- Some visualization techniques

- References
  - E.R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, 1983.
  - S.K. Card, J.D. Mackinlay, and B. Shneiderman, *Information Visualization: Using Vision to Think,* Morgan Kaufmann Publishers, 1999.

What is Information Visualization?

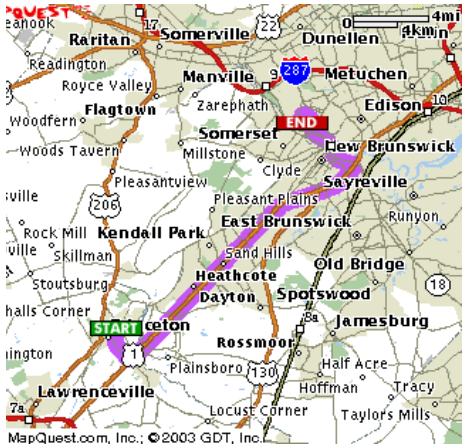
- **Visualization**:
  "The action or fact of visualizing; the power or process of forming a mental picture or vision of something not actually present to the sight; a picture thus formed." (Oxford English Dictionary)

- **Information visualization**:
  "Transformation of the symbolic into the geometric" (McCormick et al., 1987)
- **Information visualization**:
  "... finding the artificial memory that best supports our natural means of perception." (Bertin, 1983)
- **Information visualization:**
  "The use of computer-supported, interactive, visual representations of abstract data to simplify cognition." (Card, Mackinlay, Shneiderman, 1999)

Power of Visualization

From Princeton CS Department to Rutgers' CS Department:

- **Start out going South on OLDEN ST toward PROSPECT AVE.**
- **Turn RIGHT onto PROSPECT AVE.**
- **Turn LEFT onto WASHINGTON RD/ CR-526/ CR-571.**
- **Turn RIGHT.**
- **Turn LEFT onto US-1 N/ BRUNSWICK PIKE. Continue to follow US-1 N.**
- **Merge onto NJ-18 N toward TRENTON/ NEW BRUNSWICK.**
- **NJ-18 N becomes CR-609 N/ METLARS LN.**
- **Turn LEFT onto SUTPHEN RD.**
- **Turn RIGHT onto FRELINGHUYSEN RD.**



Information Visualization

- Problem
  - How to understand massive datasets?
- Solution
  - Convert information into a graphical representation
  - Take better advantage of human perceptual system
- Issues
  - What is a good visualization?
  - How to convert data?

Goals of Information Visualization

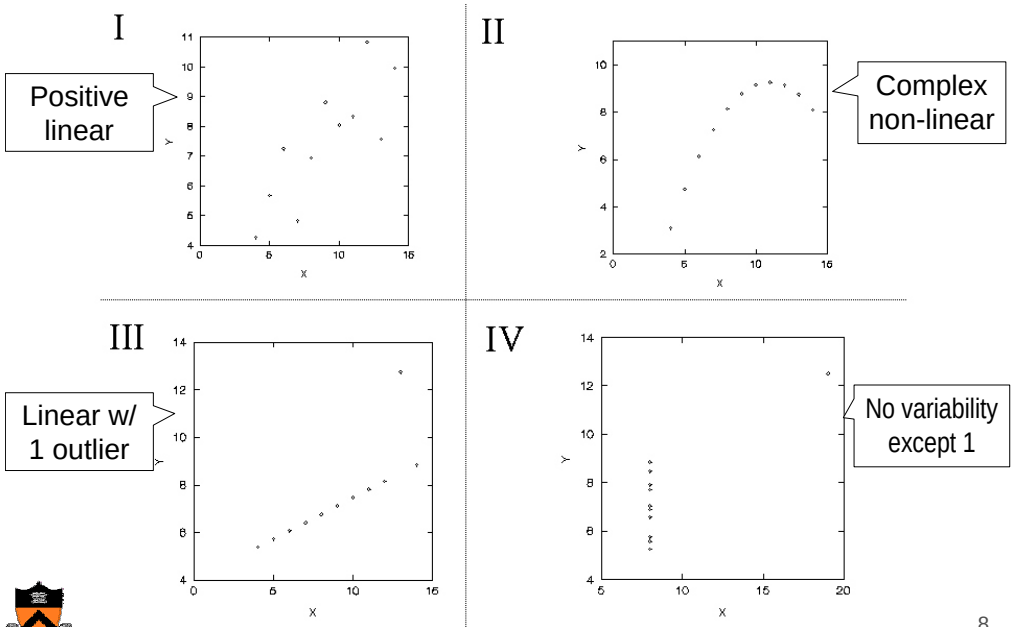
- Make large datasets coherent
- Present huge amounts of information compactly
- Induce the viewer to think about the substance instead of methodology, design, technology, and so on
- Encourage comparisons of different data
- Present information at several levels of detail, from overviews to fine structure
- Tell stories about the data statistically

Statistical Visualization: Anscombe's Quartet

Data Set I		Data Set II		Data Set III		Data Set IV	
X	***Y***	***X***	***Y***	***X***	***Y***	***X***	***Y***
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (Feb 1973), pp17-21

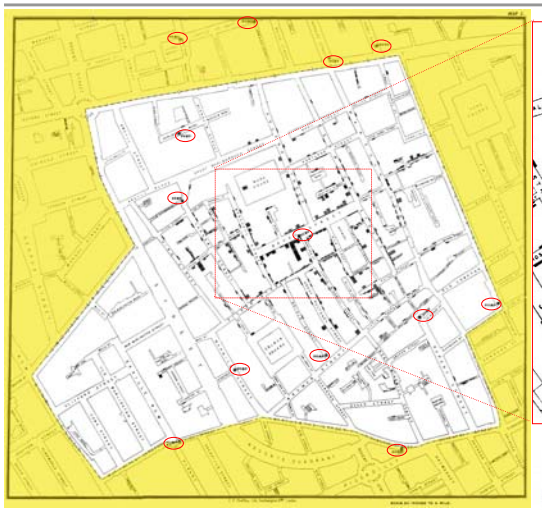
Anscombe's Scatter Plots

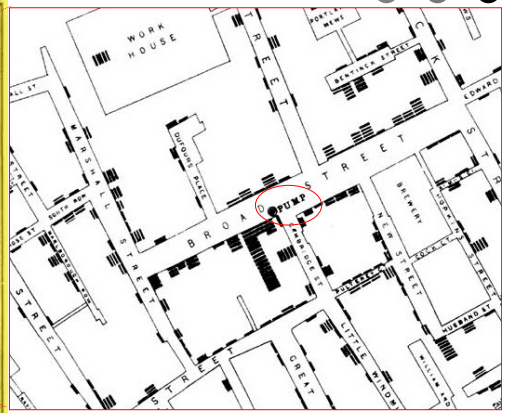


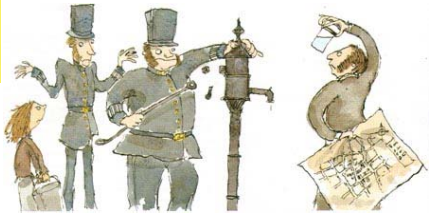
Cholera Outbreak in London in 1854

- The first death caused by cholera was found in London in 1831.
- The year 1853 saw outbreaks in Newcastle and Gateshead as well as in London, where a total of 10,675 people died of the disease.
- On August 31 of 1854, the outbreak of cholera hit London Soho area: 127 people died in the next three days and 500 within 10 days.
- What is causing a cholera epidemic in London in 1854?
- Dr. John Snow suspected cholera was transmitted by water, but could not prove it, then he used a map …

John Snow's Map of Cholera Deaths

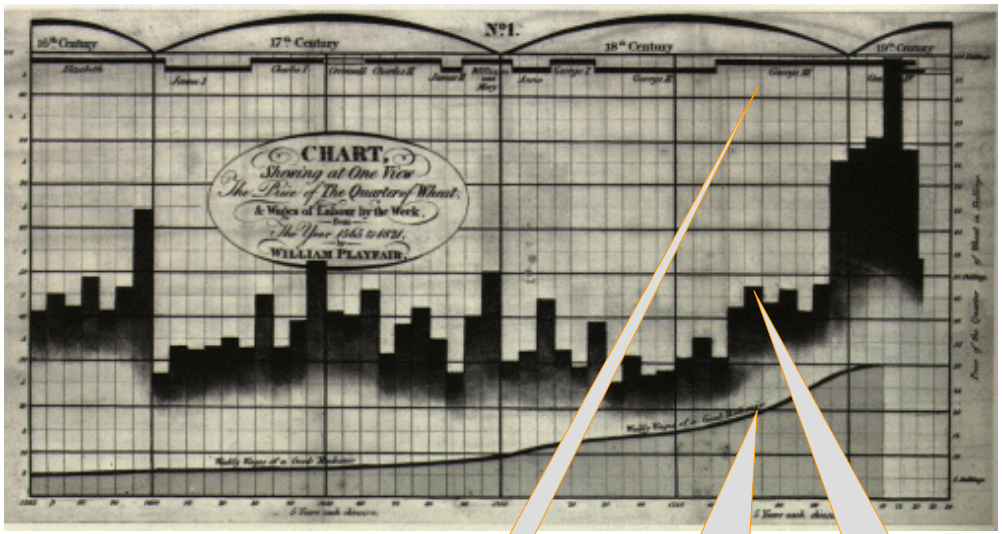






Dr. John Snow plotted the location of deaths from cholera in central London for Sept 1854. Deaths are marked by black dots. Water pumps are marked with red cycles.

Time Series: Wheat Prices, Wages and Kings and Queens (William Playfair, 1786)

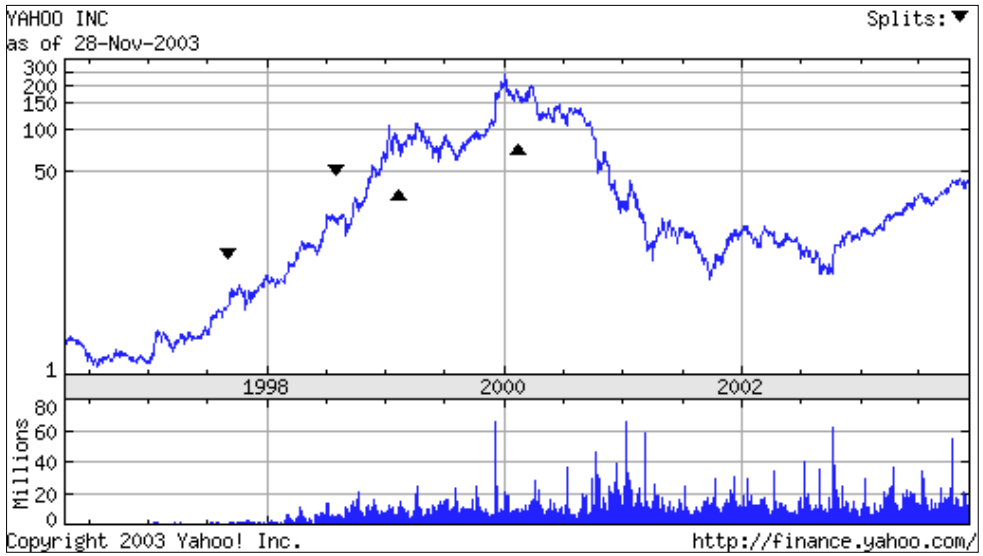


King or Queen

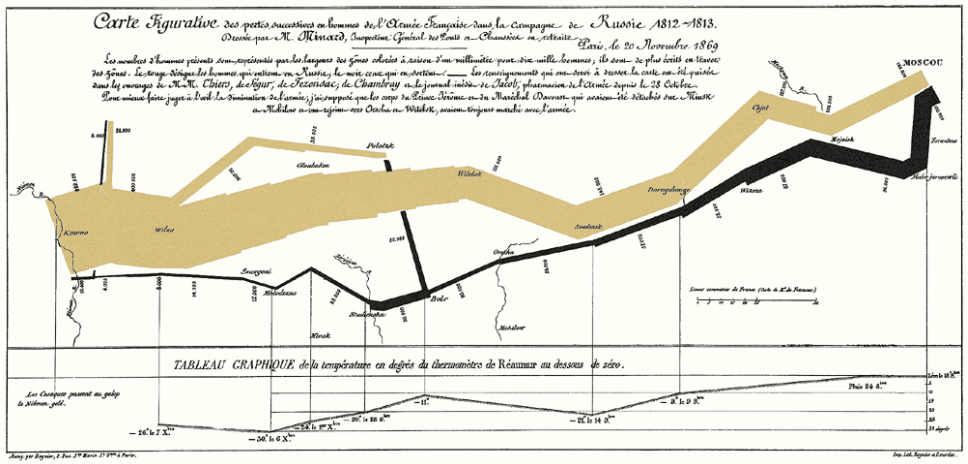
Weekly wages of good mechanics

Price of quarter of wheat

Today's Time Series

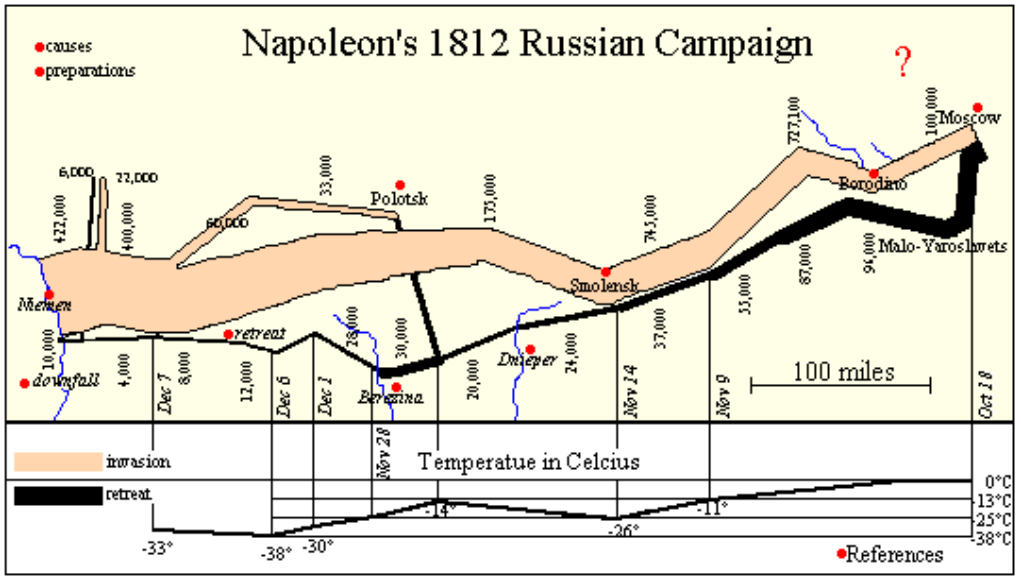


Space & Time: Napoleon's Army in Russia (Charles Joseph Minard, 1861)



"It may well be the best statistical graphic ever drawn."
Edward R. Tufte, 1983

A More Readable Version



Principles of Graphical Excellence

- Graphical excellence is the well-designed presentation of interesting data – a matter of substance, of statistics, and of design
- Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency
- Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
- Graphical excellence is nearly always multivariate
- Graphical excellence requires telling the truth about the data

E.R. Tufte 1983

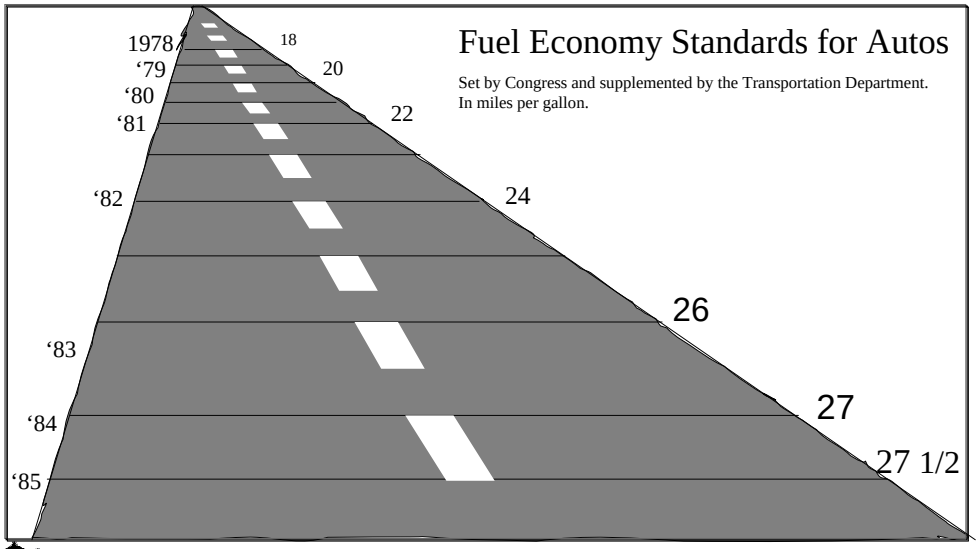
Integrity Principle I

- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented
- Measure of violation

$$\text{Lie Factor (LF)} = \frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$

- Use logarithm of the Lie Factor to compare
  - Overstating *log* LF > 0
  - Understating *log* LF < 0
  - Most distortions involve overstating; LF = 2-5 are common

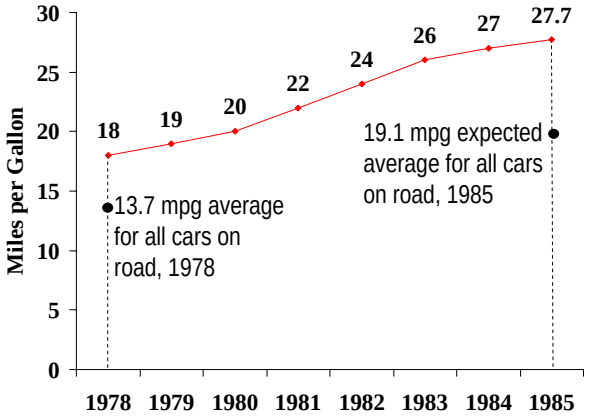
Example Violated the Principle



Adapted from The New York Times, August 9, 1978, p. D-2

Integrity Principle II

- Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity
- Write out explanations of the data on the graphic itself
- Label important events in the data



Integrity Principle III

- Show data variation, not design variation

Integrity Principle IV

- In time-series displays of money, deflated and standardized units of monetary measurement are always better than nominal units

Integrity Principle V

- The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data

Dollar value shrinks in one dimension, but the dollar sizes shrinks in 2 dimensions

Adapted from The Washington Post, October 25, 1978, p.1


1958 - Eisenhower: $1.00


1963 - Kennedy: 94¢


1968 - Johnson: 83¢


1973 - Nixon: 64¢


1978 - Carter: 44¢

Purchasing Power of the Diminishing Dollar

Source: Labor Department

LF = 1

- Now the area size of the dollar shrinks at the same rate as the dollar value


1958 - Eisenhower: $1.00


1963 - Kennedy: 94¢


1968 - Johnson: 83¢


1973 - Nixon: 64¢

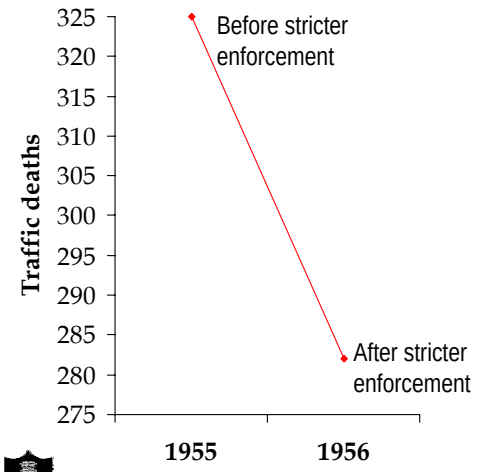

1978 - Carter: 44¢

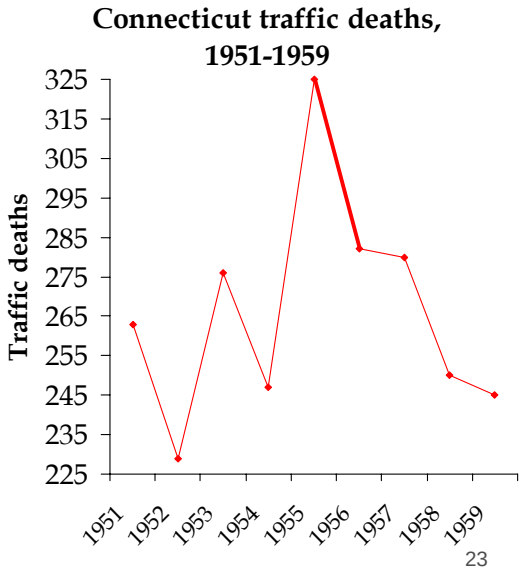
Purchasing Power of the Diminishing Dollar

Source: Labor Department

Integrity Principle VI

- Graphics must not quote data out of context

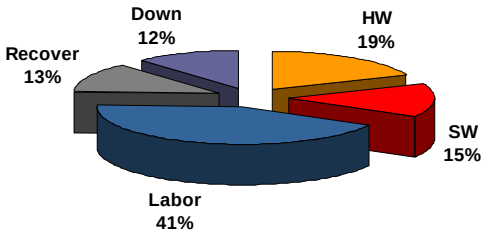


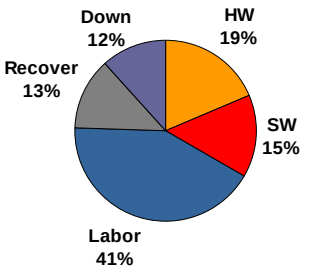


Data Ink Principle (Tufte, 1983)

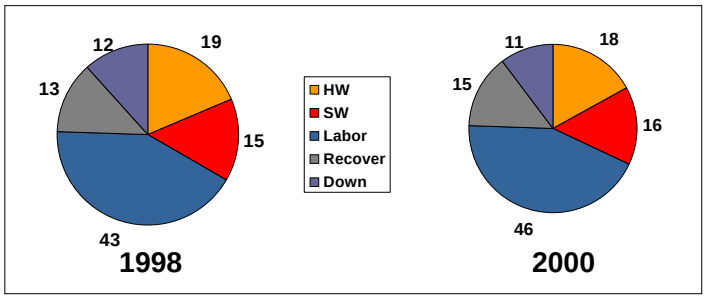
$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

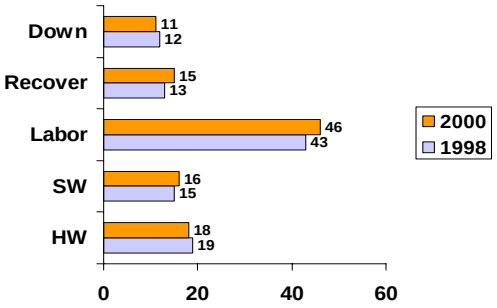
- Maximize the data-ink ratio
- Erase non-data-ink
- Erase redundant data-ink





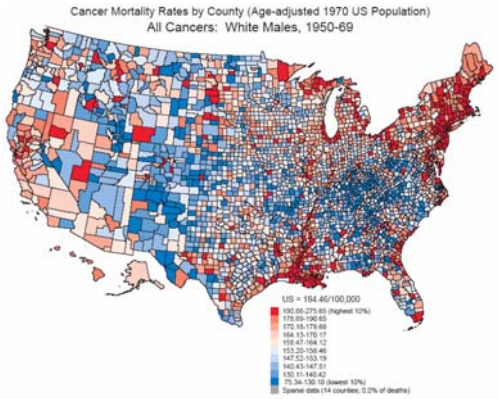
Example: Erase Redundant Data-Ink

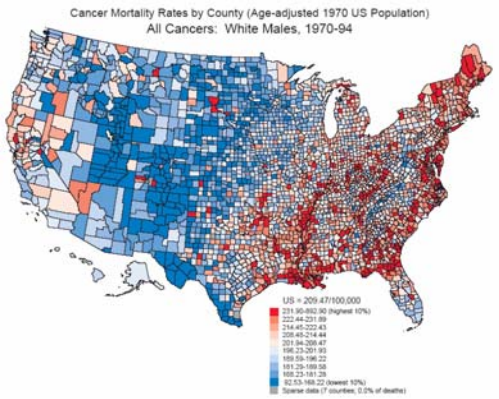




Data Maps: Cancer Mortality by County

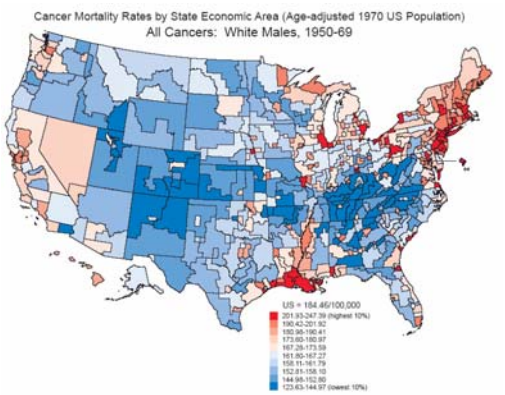
- What do we learn from the maps?
- What's wrong with the data maps?

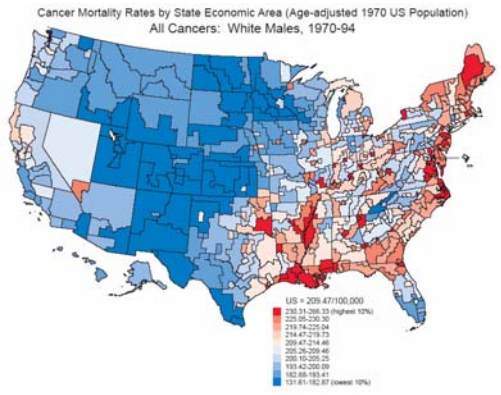




Data Maps: Cancer Mortality by SEA

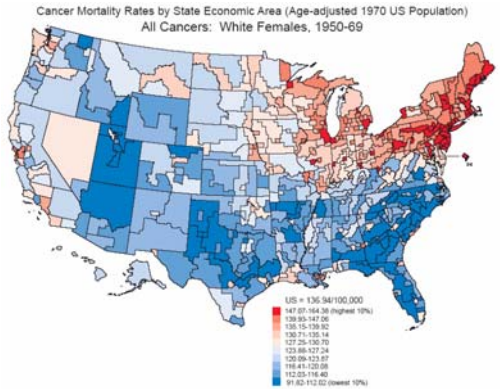
- What do you think about these maps?

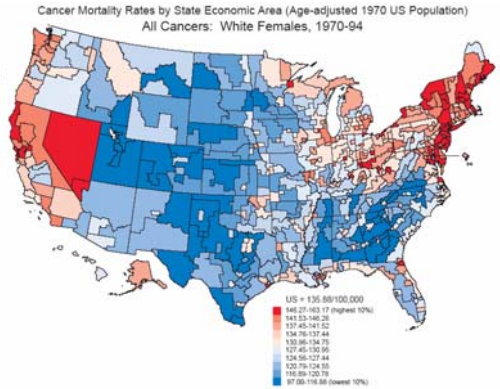




Data Maps: Cancer Mortality by SEA

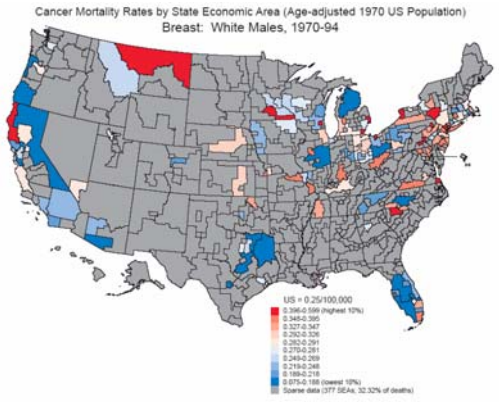
- What do we learn from these?

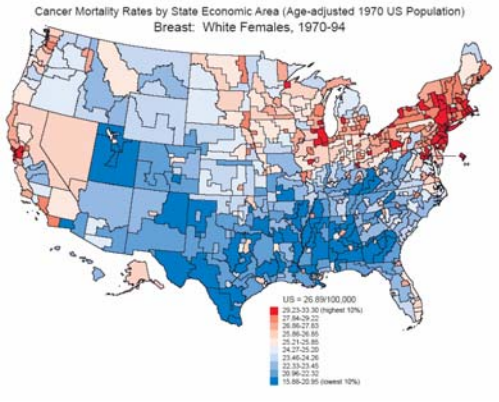




Data Maps: Breast Cancer by SEA

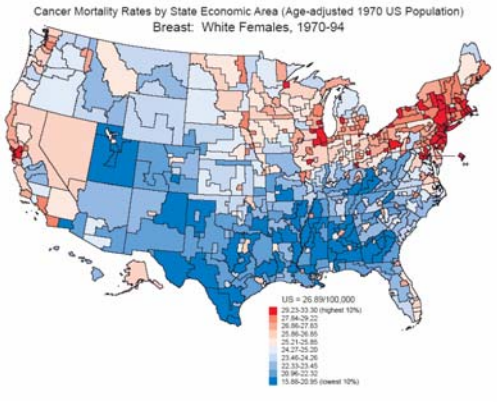
- Big difference between male and female?

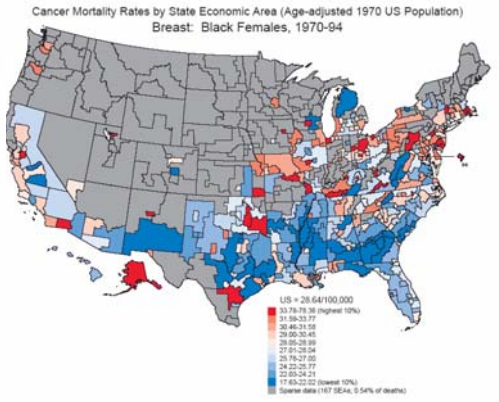




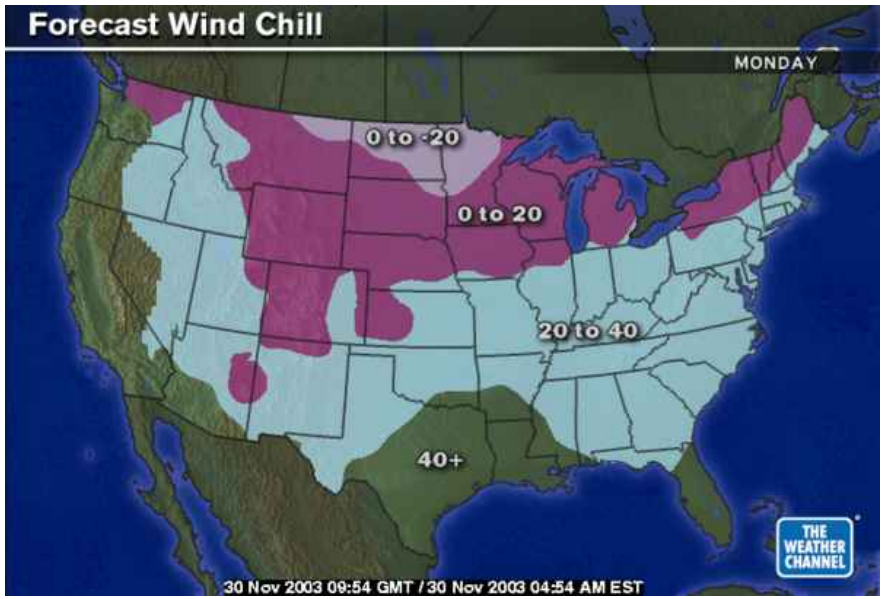
Data Maps: Breast Cancer by SEA Black vs. White Female

- What can

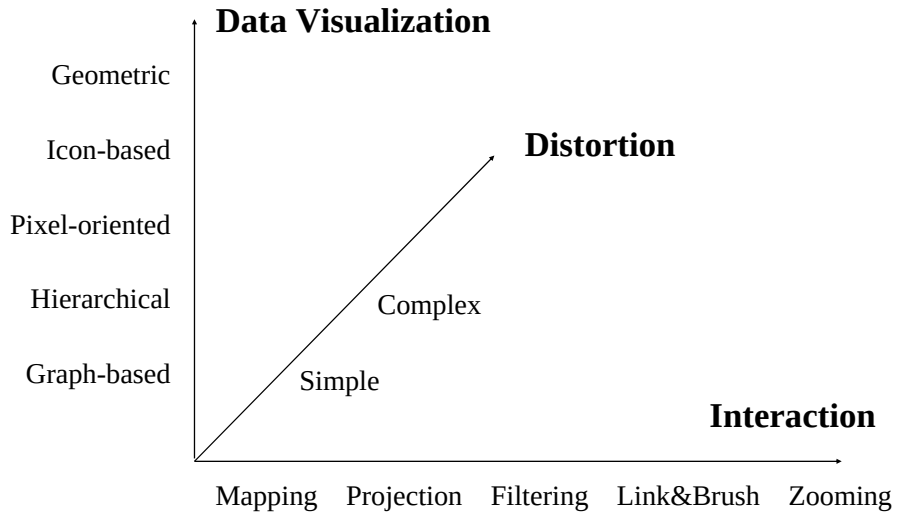




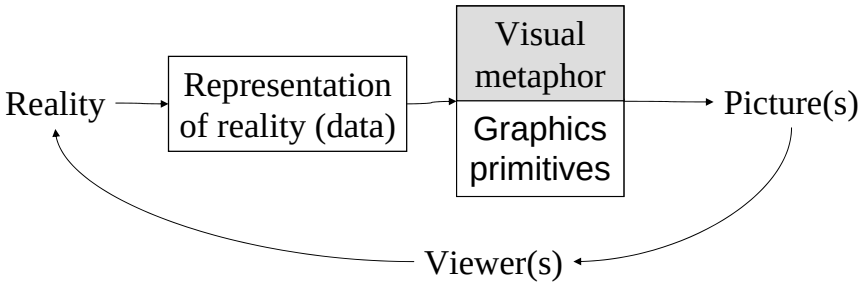
What About This Familiar Data Map



Dimensions to Explore Data (Keim 97)



The Key Is Visual Metaphor



- Metaphor examples
  - Weather map
  - Scatter plots

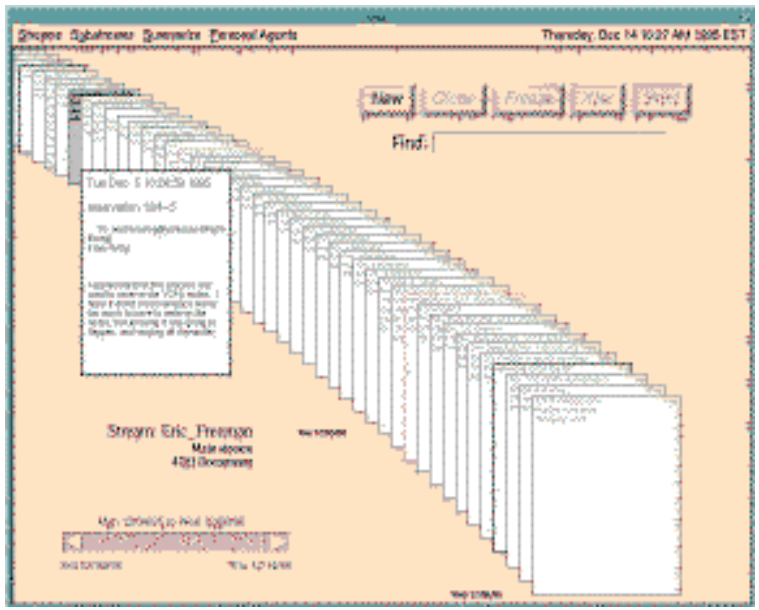
Basic Techniques

- Scatter plot for point data
  - 1-D: The data are points on a single axis
  - 2-D: The pairs of values are points in a plane
  - 3-D: Use color, brightness, and animation
- Scalar Data
  - Line graph: Draw a line through a set of data points, interpolating using straight line, spline curve, etc. (Lambert proposed drawing through middle of two points in 1765)
  - Multiple line graph: Display several plots on the same graph, using line patterns (continuous, dot, dash), thickness, and color
  - Bar chart (pie charts similar): Depict values by lengths of bars
  - Histogram: Data elements are placed into bins according to value ranges, and draw bar charts on the numbers of elements of bins

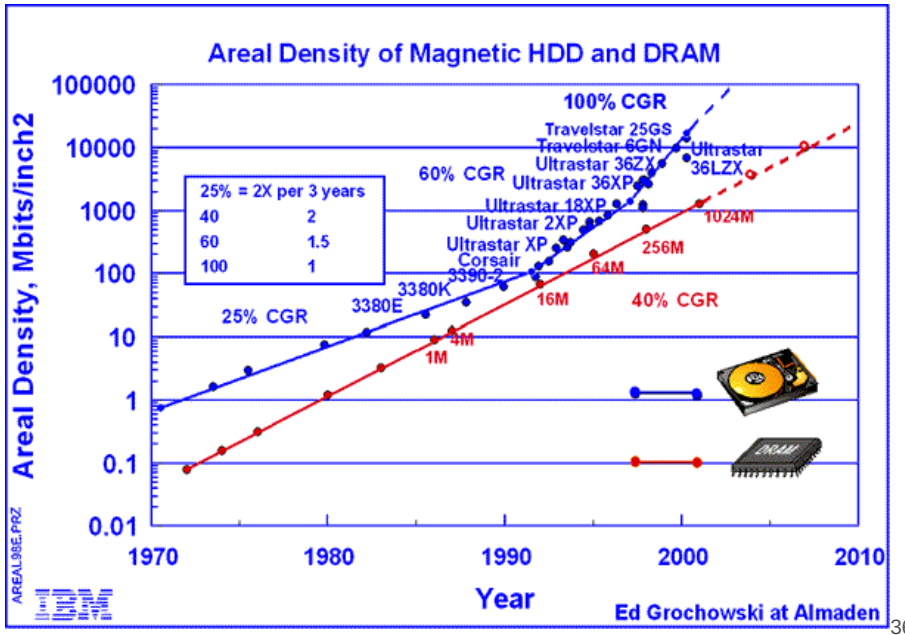
Lifestreams

(Fertig, S., Freeman, E. and Gelernter, D. 1996)

- View by time



2-D: Disk Areal Density vs. Time

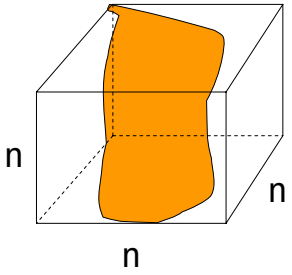


Techniques for Multi-Dimensions

- Place data using 2-D placements
  - Scatter-plot matrices, hyperslice, prosection
  - Parallel coordinates
  - Icon shapes
  - Stick figures
  - Pixel-oriented with tour, spiral, axes, circle segments
  - Colors
- Place data using 3-D projection, landscape
  - Isosurface
  - Volume rendering
  - Vector visualization

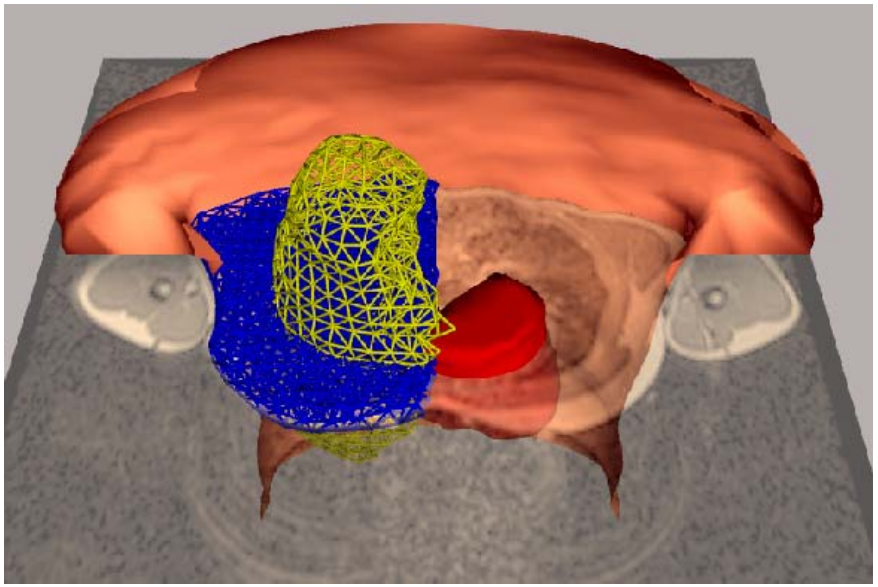
Isosurface Cells

- For a given isovalue, only a smaller portion of cells are isosurface cell.
- For a volume with n x n x n cells, the average number of the isosurface cells is n x n (ratio of surface v.s. volume)
- The classical approach is called “Marching cubes” which marches through all cells and figure out the isosurface



Isosurface Extraction of Visible Woman

3-D Volume Rendering



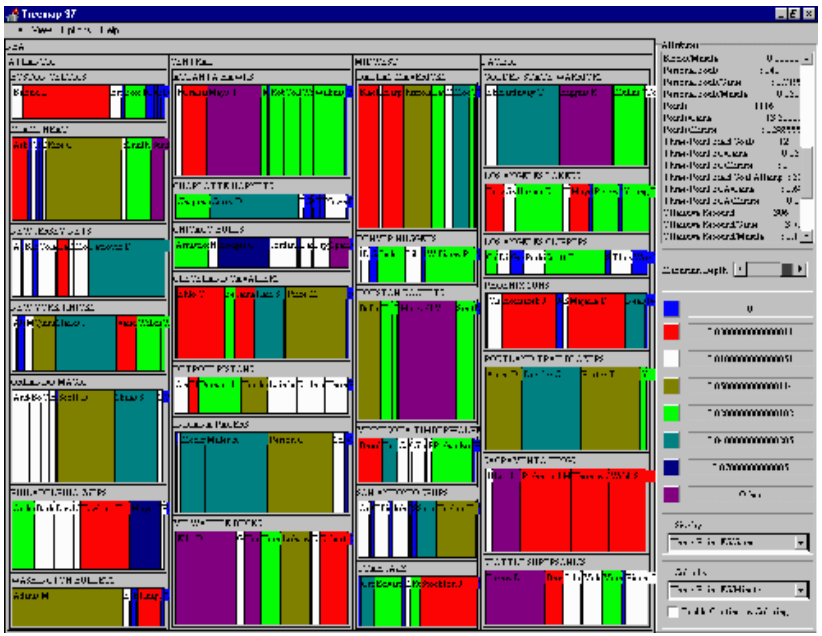
Graphs

- Types
  - Undirected graphs
  - Directed graphs
- CS Examples
  - Networks and wiring diagrams
  - Finite state machines
  - Dependencies
  - Call graphs
  - Pointers

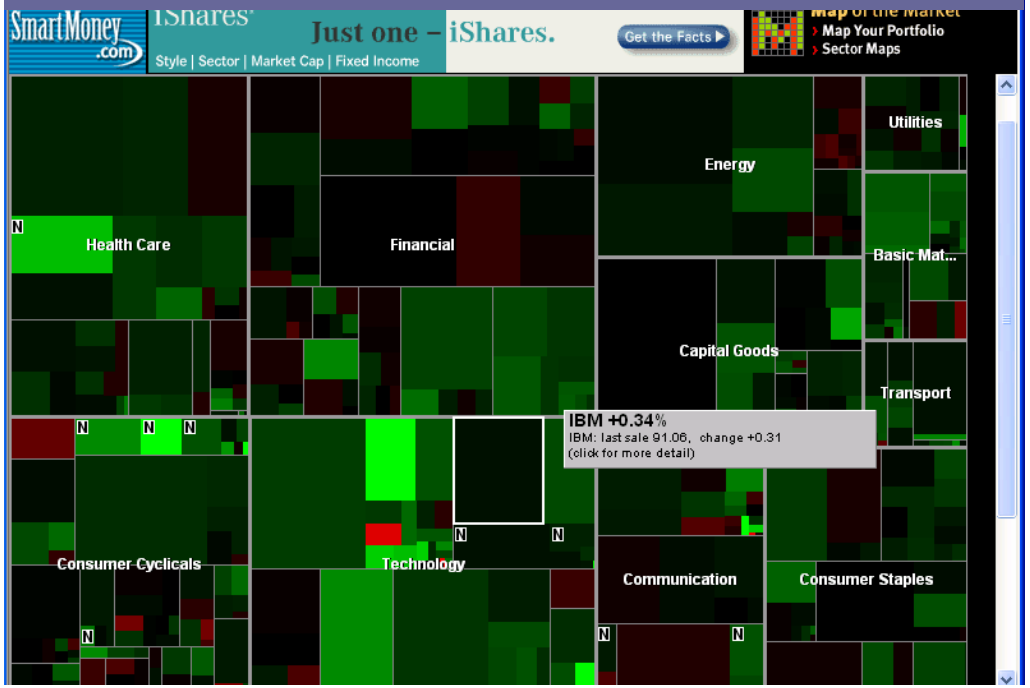
Hierarchies

- Techniques
  - Tree
  - Tree-map: subdivide spaces for multiple dimensions
  - Cone tree
  - Info-cube
- Examples
  - Organization
  - Directory
  - Abstraction

Early Treemap Applied to File System

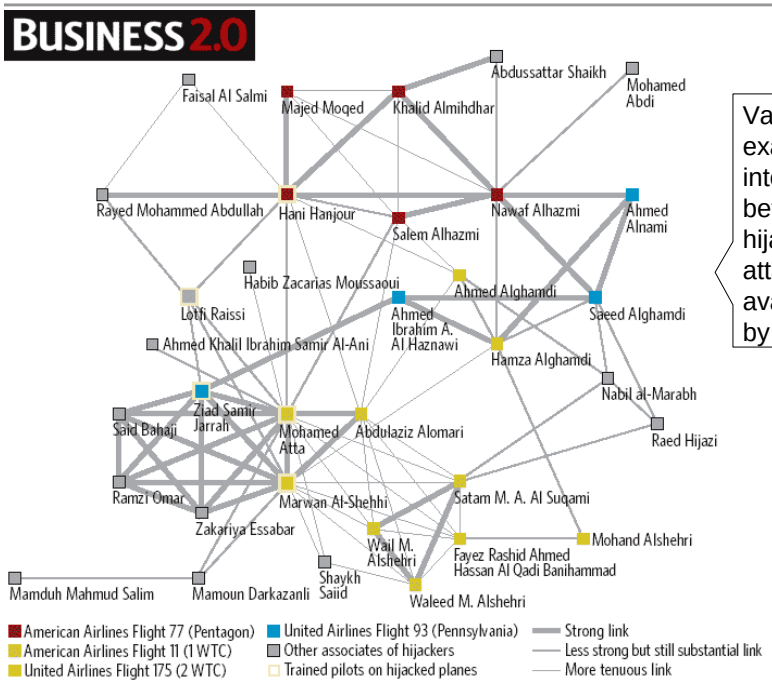


A TreeMap Application



Six Degrees of Mohamed Atta

(T.A. Stewart, December 2001 issue)



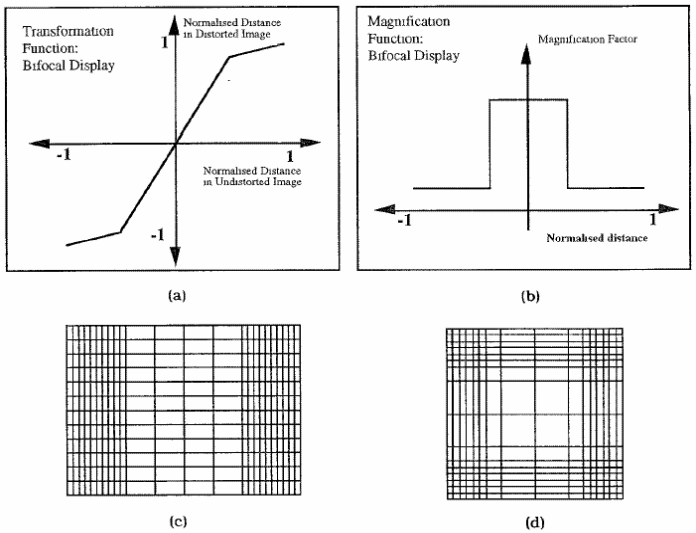
Valdis Krebs's examination of the interrelationships between the 19 hijackers in the 9/11 attack and the available connections by the authorities

Distortion

- Problem
  - Small displays vs. large information space
  - Tunnel vision
  - Seeing the forest through the trees
- Views
  - Local detail or focus
  - Global context
- Taxonomy
  - Multiple windows (overview and detail)
  - Fisheye lens or selective distortion
  - Pan and Zoom

Bifocal Display

- Distortion at 1 or 2 dimensions with linear transformation
- Combination of detailed view and two distorted side views.



Perspective (Bifocal) Wall