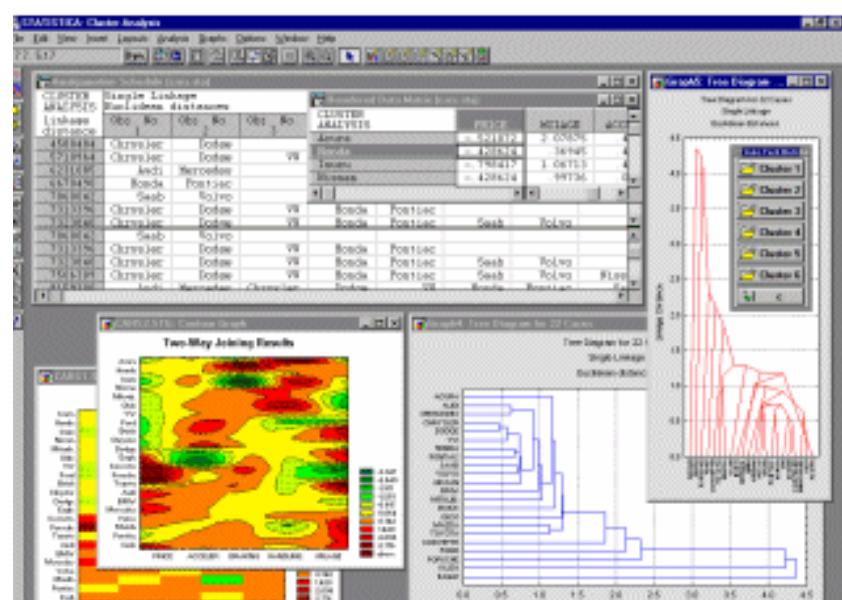
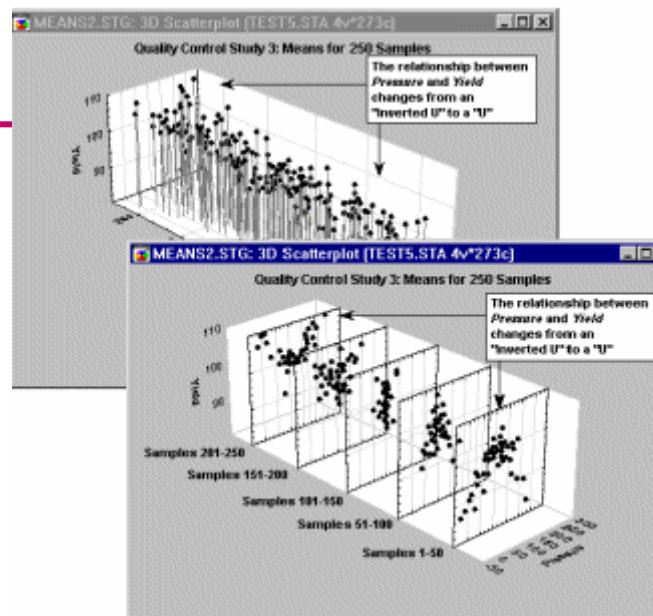
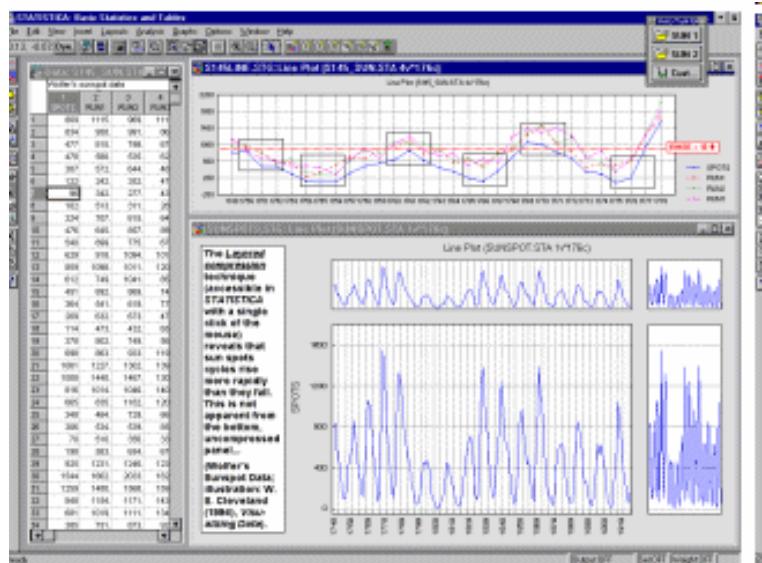

Data Visualization

or Graphical Data Presentation



What is visualization and data mining?

- **Visualize:** “To form a mental vision, image, or picture of (something not visible or present to the sight, or of an abstraction); to make visible to the mind or imagination.”
- **Visualization** is the use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data.
- **Visual Data Mining** is the process of discovering implicit but useful knowledge from large data sets using visualization techniques.

Tables vs graphs

A table is best when:

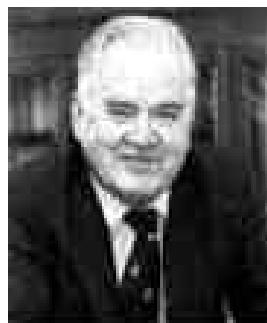
- You need to look up specific values
- Users need precise values
- You need to precisely compare related values
- You have multiple data sets with different units of measure
- Graphs and tables serve different purposes. Choose the appropriate data display to fit your purpose.

A graph is best when:

- The message is contained in the shape of the values
- You want to reveal relationships among multiple values (similarities and differences)
- Show general trends
- You have large data sets

Exploratory Data Analysis

- Pioneer -> John Tukey
 - New approach to data analysis, heavily based on visualization, as an alternative to classical data analysis
 - See its bio
- Two stage process:
 - Exploratory: Search for evidence using all tools available
 - Confirmatory: evaluate strength of evidence using classical data analysis

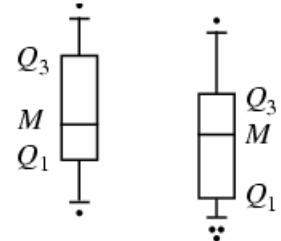


Box Plots

- In some situations we have, not a single data value at a point, but a number of data values, or even a probability distribution
- When might this occur?
- Tukey proposed the idea of a **boxplot** to visualize the distribution of values
- For explanation and some history, see:

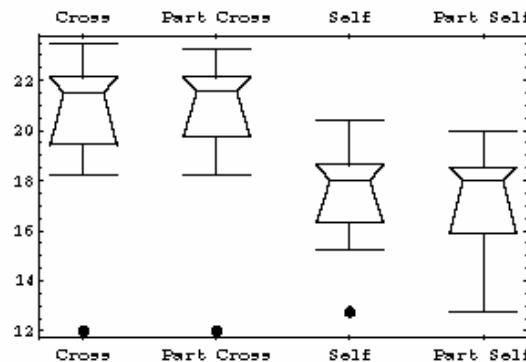
<http://mathworld.wolfram.com/Box-and-WhiskerPlot.html>

http://en.wikipedia.org/wiki/Box_plot

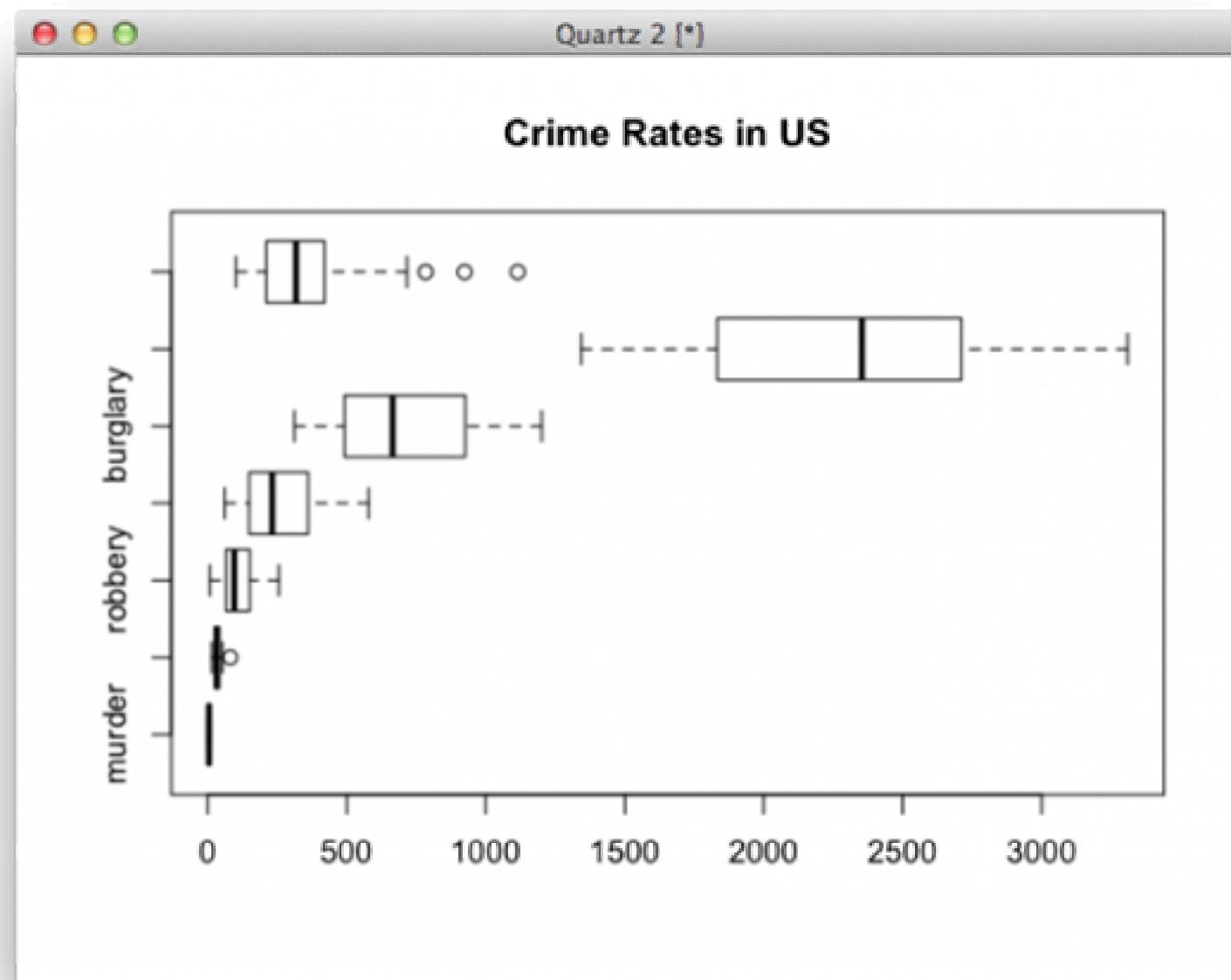


M – median
Q1, Q3 – quartiles
Whiskers –
1.5 * interquartile range
Dots - outliers

Darwin's plant study



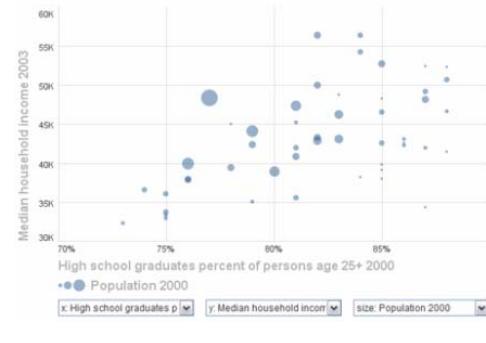
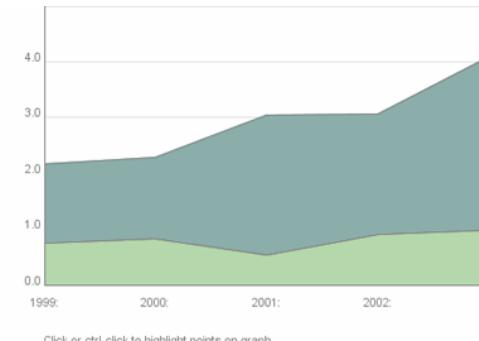
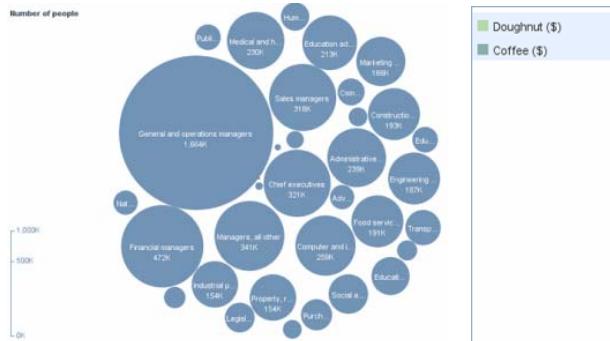
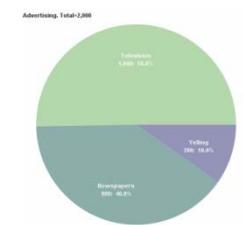
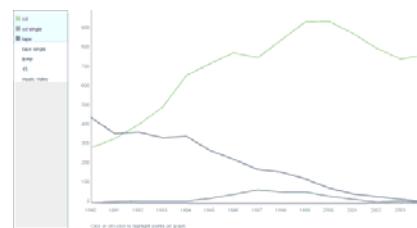
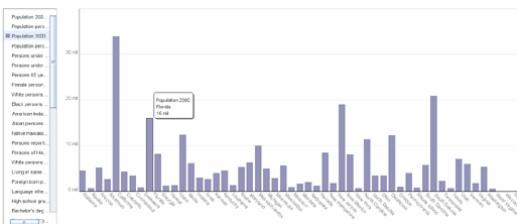
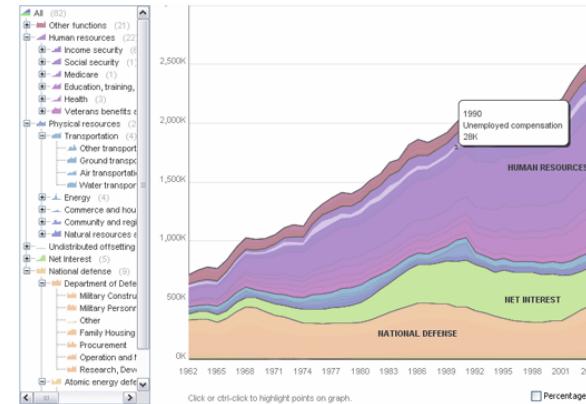
Distribution visualisation – US Crime Story



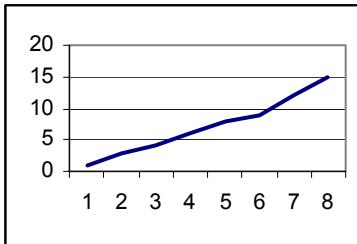
Data Visualization – Common Display Types

Common Display Types

- Bar Charts
- Line Charts
- Pie Charts
- Bubble Charts
- Stacked Charts
- Scatterplots

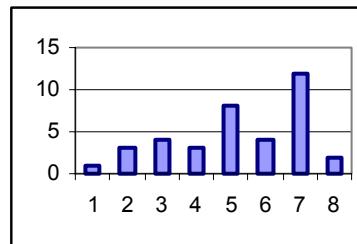


When to use which type?



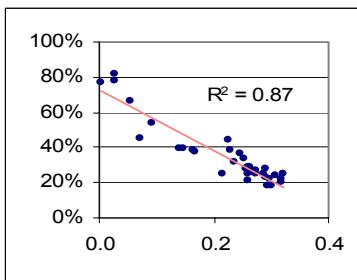
Line Graph

- x-axis requires quantitative variable
- Variables have contiguous values
- Familiar/conventional ordering among ordinals



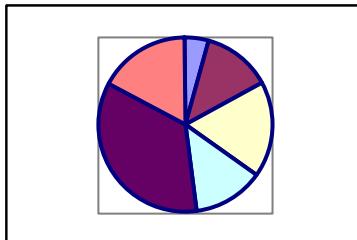
Bar Graph

- Comparison of relative point values



Scatter Plot

- Convey overall impression of relationship between two variables

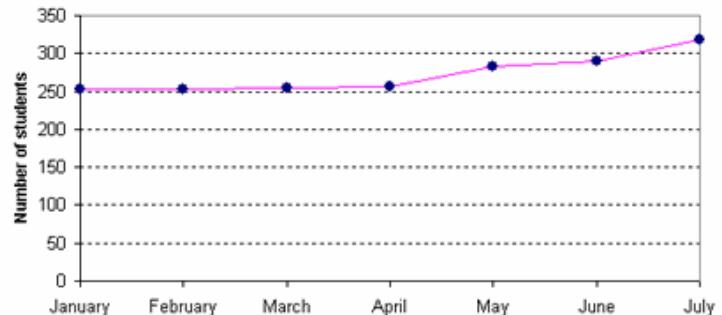


Pie Chart

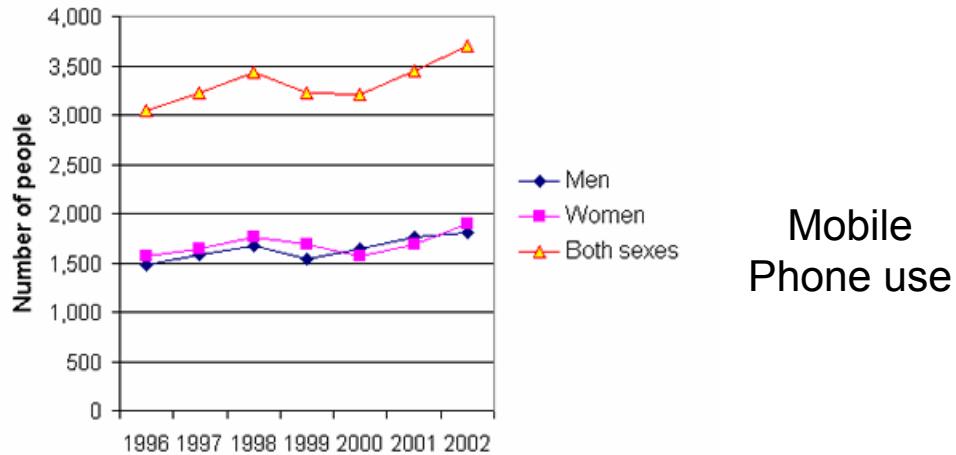
- Emphasizing differences in proportion among a few numbers

Line Graph – Trend visualization

- Fundamental technique of data presentation
- Used to compare two variables
 - X-axis is often the control variable
 - Y-axis is the response variable
- Good at:
 - Showing specific values
 - Trends
 - Trends in groups (using multiple line graphs)



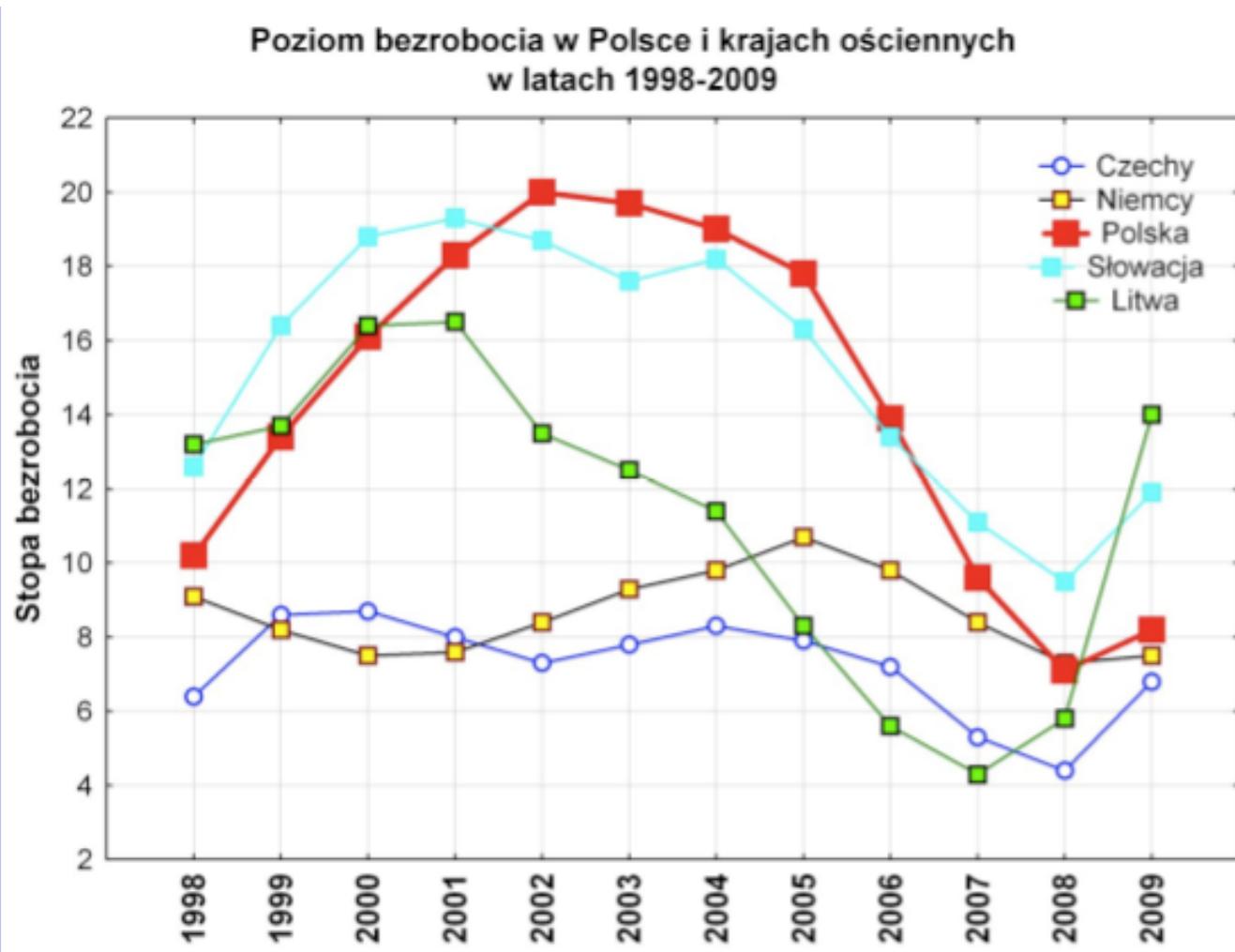
Students participating in sporting activities



Mobile Phone use

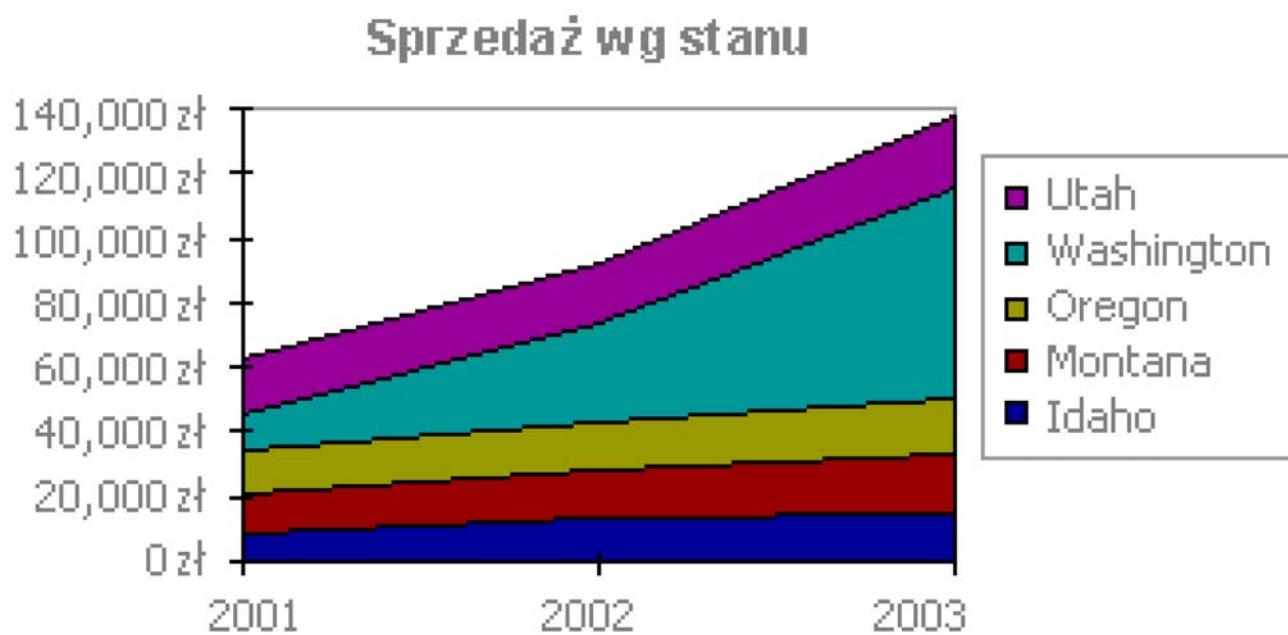
Note: graph labelling is fundamental

Time line graph – show dynamics of measurements

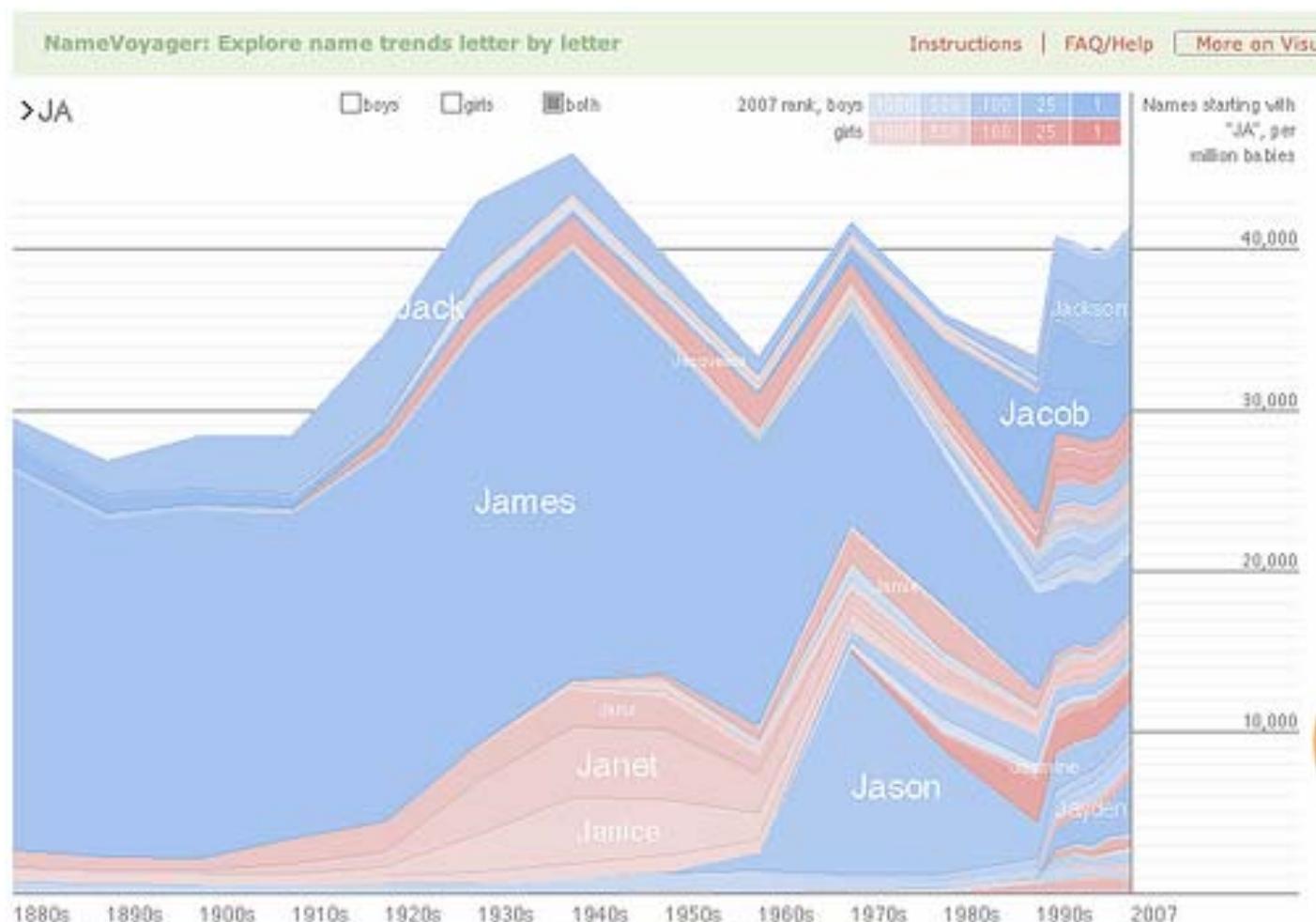


Stratified graphs

- Trends of values with respect to time and different qualitative categories



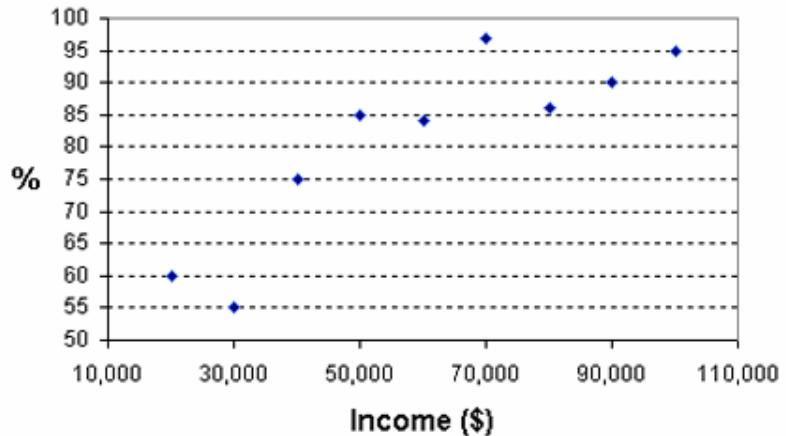
Demo – Baby Names Voyager



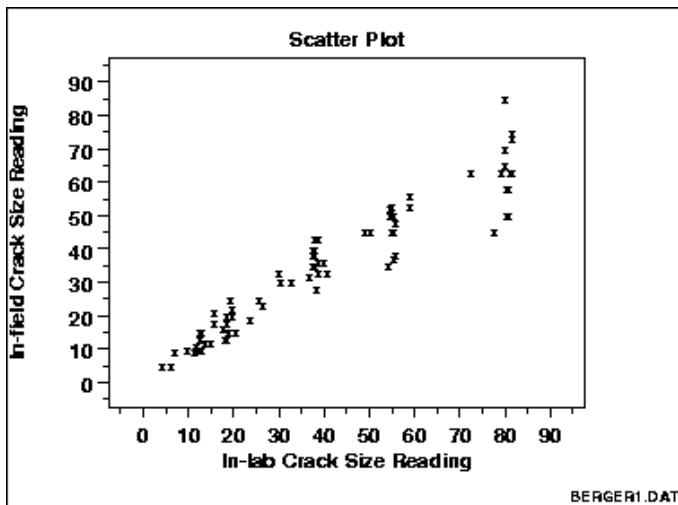
<http://www.babynamewizard.com/voyager>

Scatter Plot – Wykresy rozrzutu XY

- Used to present measurements of two variables
- Effective if a relationship exists between the two variables

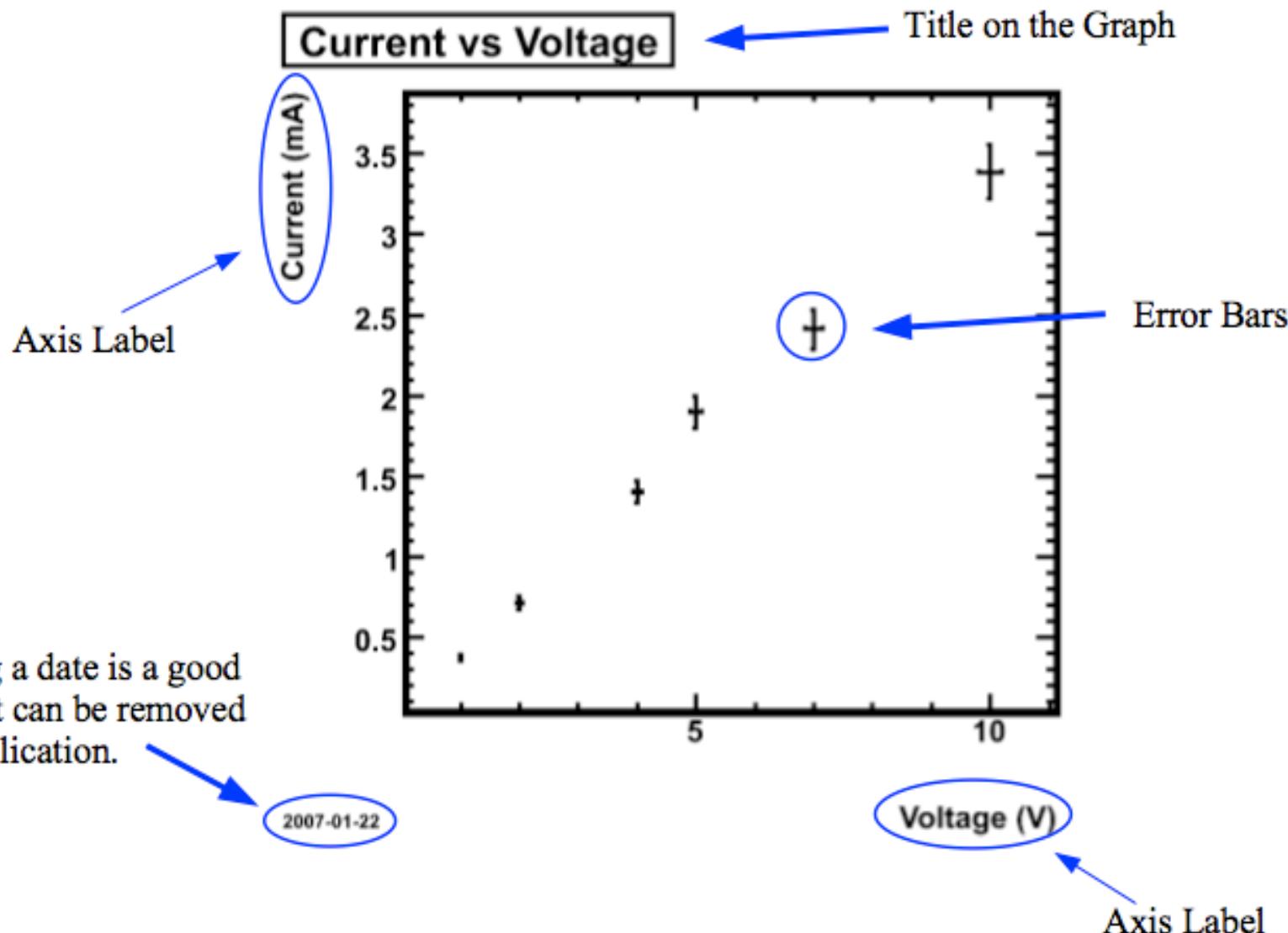


Car ownership by household income



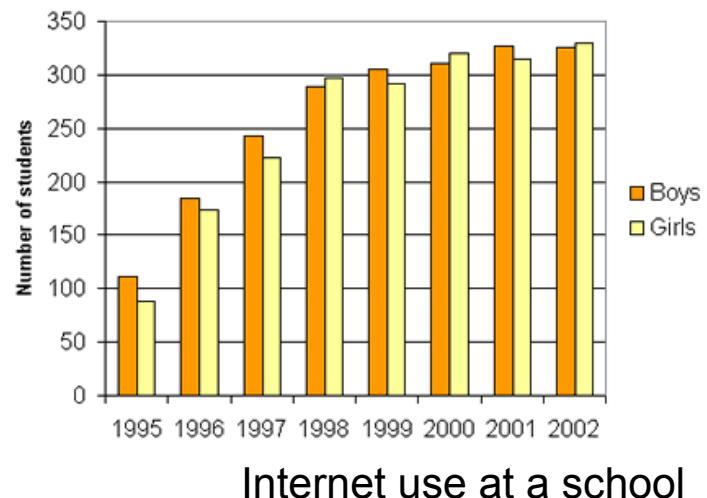
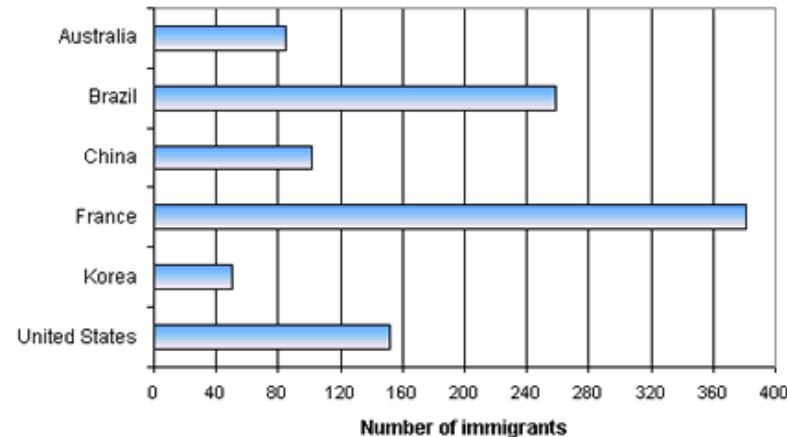
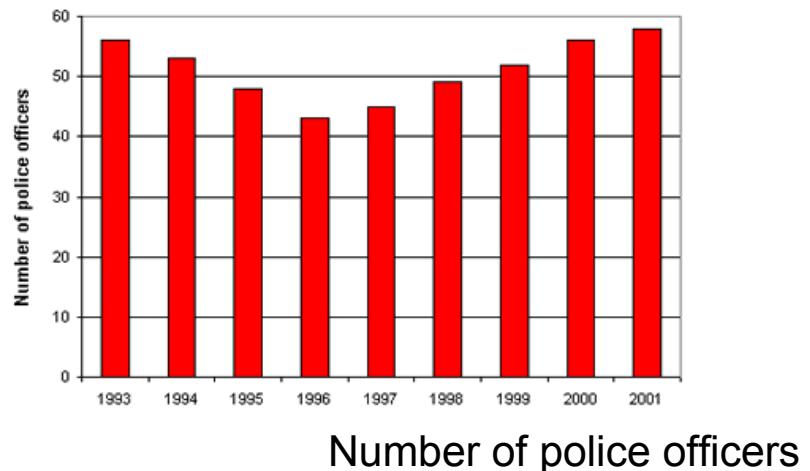
Example taken from
NIST Handbook –
Evidence of strong
positive correlation

Elements of a Good Graph



Simple Representations – Bar Graph

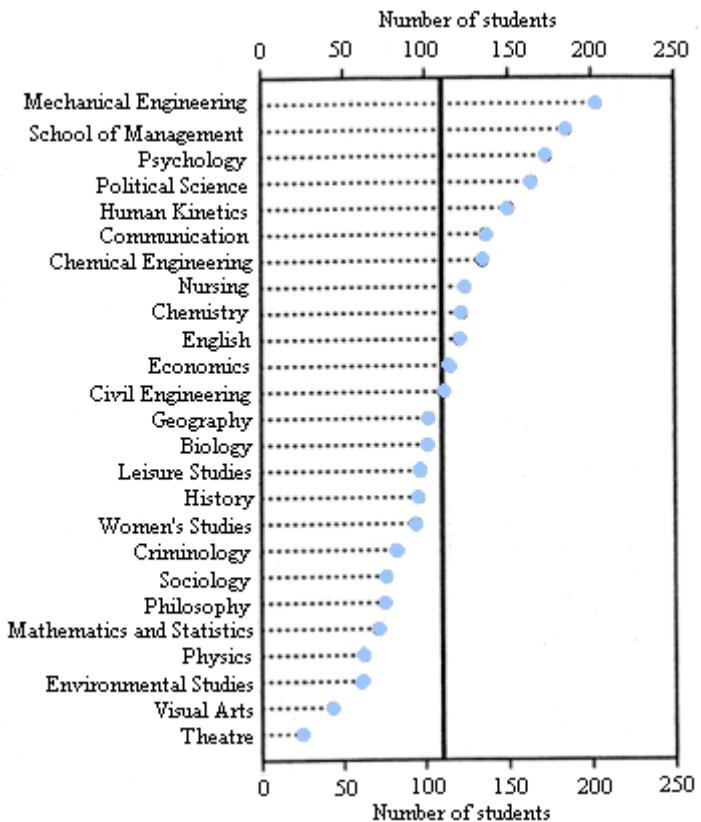
- Bar graph
 - Presents categorical variables
 - Height of bar indicates value
 - Double bar graph allows comparison
 - Note spacing between bars
 - Can be horizontal (when would you use this?)



Note more space for labels

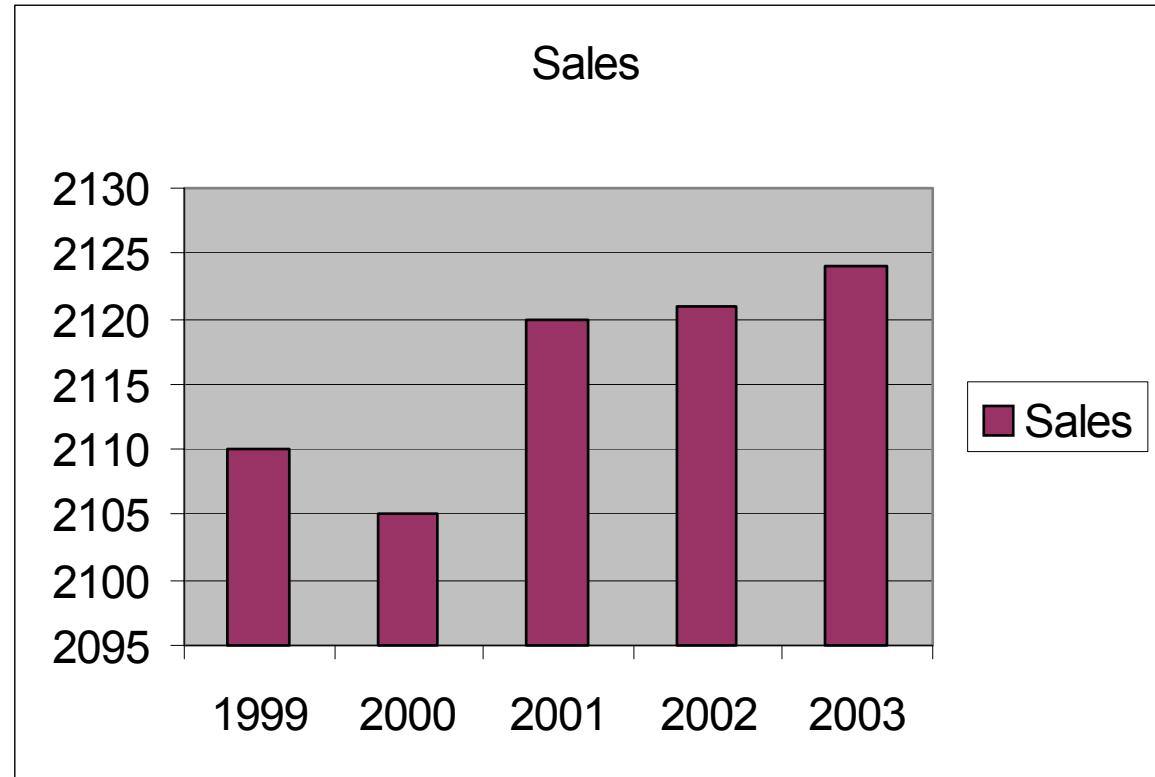
Dot Graph

- Very simple but effective...
- Horizontal to give more space for labelling



Bad Visualization: Spreadsheet

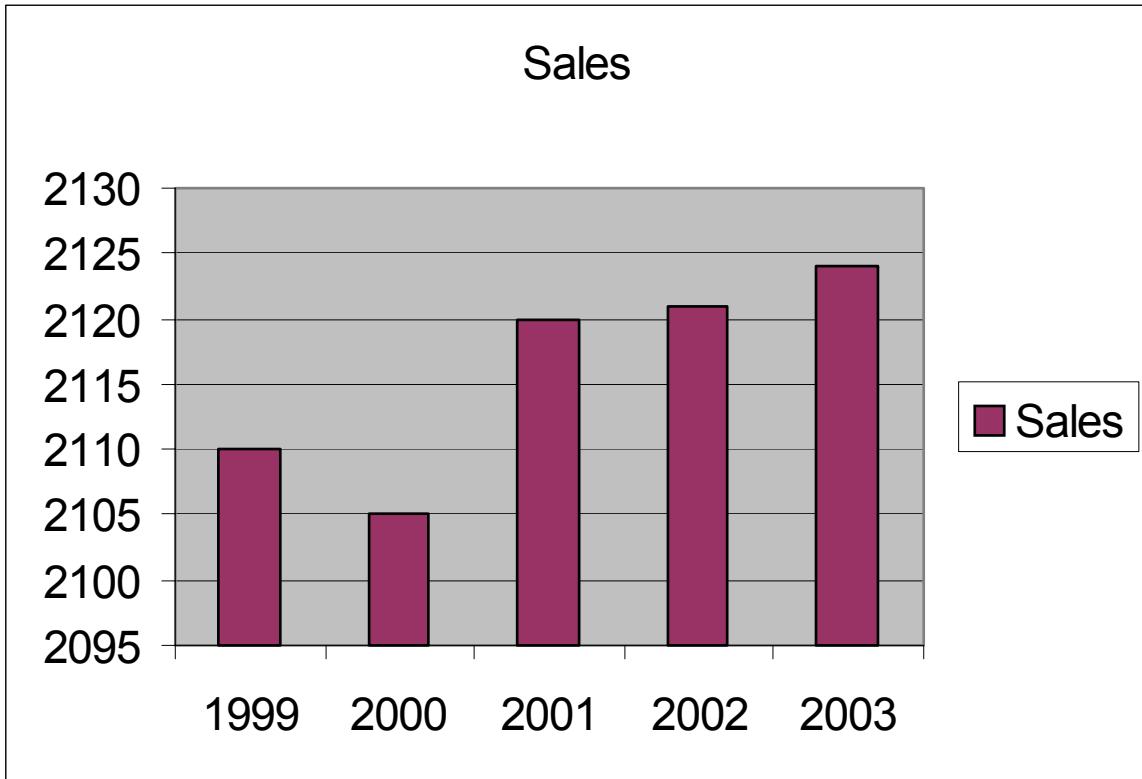
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



What is wrong with this graph?

Bad Visualization: Spreadsheet with misleading Y –axis

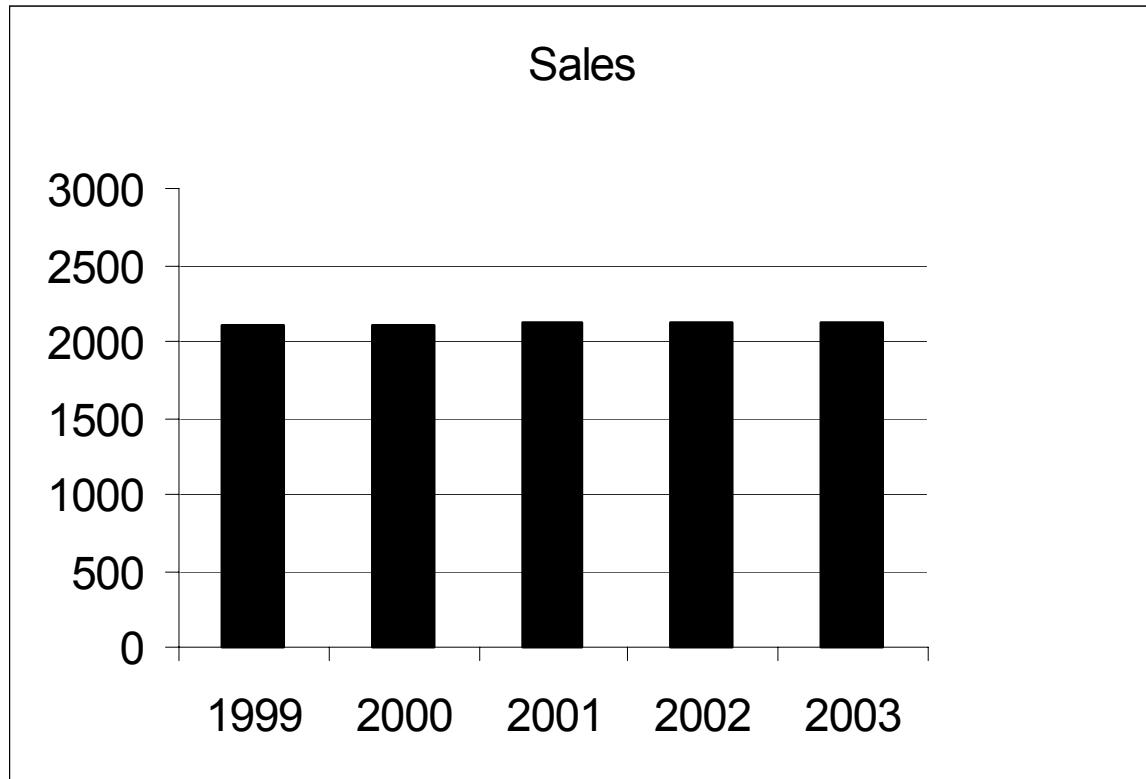
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



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

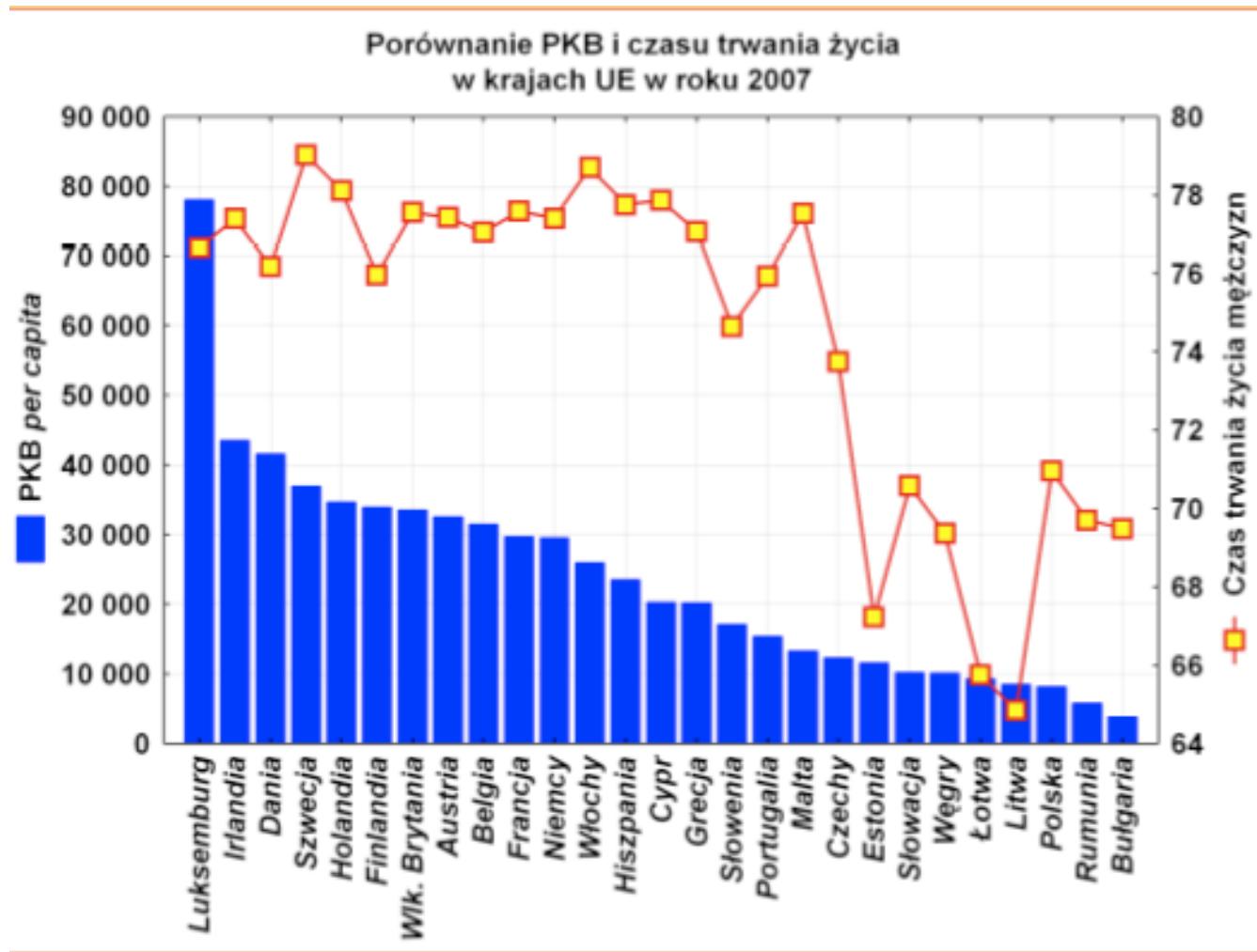
Better Visualization

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



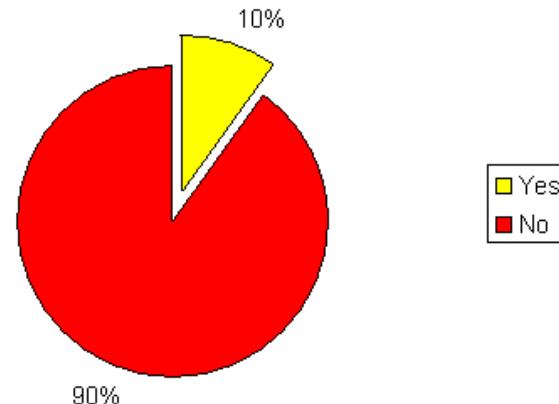
Axis from 0 to 2000 scale gives
correct impression of small change + small formatting tricks

Integrating various graphs

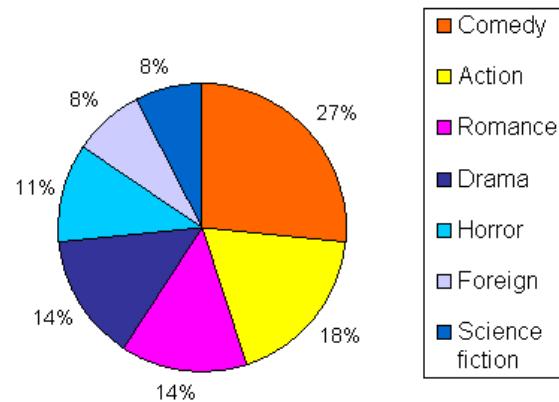


Pie Chart

- Pie chart summarises a set of categorical/nominal data
- But use with care...
- ... too many segments are harder to compare than in a bar chart



Should we have a long lecture?



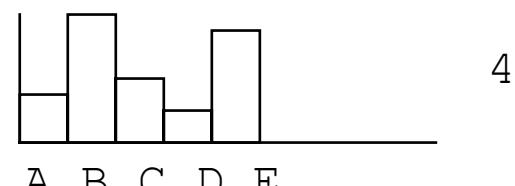
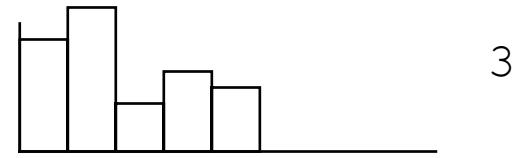
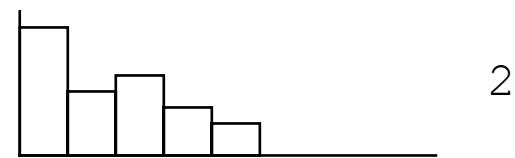
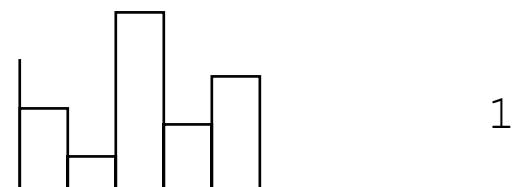
Favourite movie genres

Visualizing in 4+ Dimensions

- Extensions of Scatterplots
- Parallel Coordinates
- Radar Figures
- Other tools
- ...

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



Problem: does not show correlations

Tableau bar comparisons

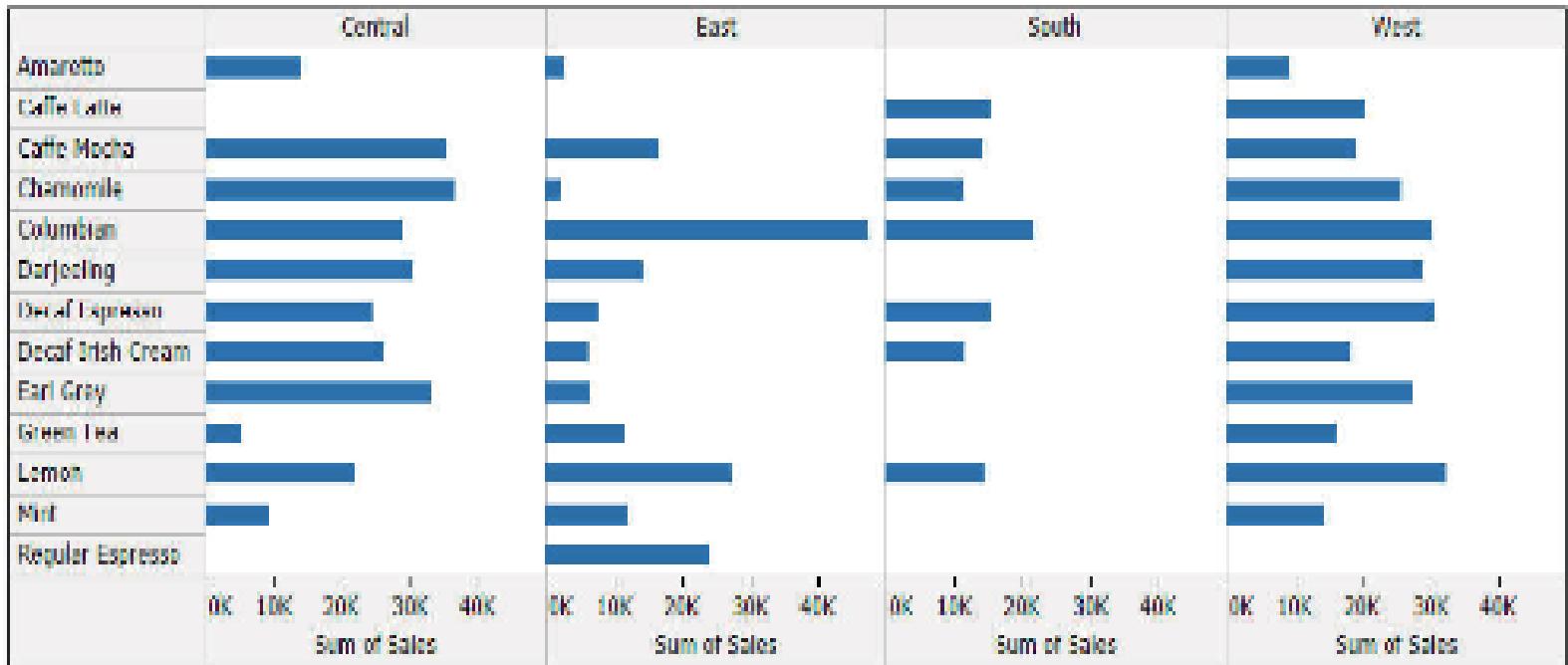
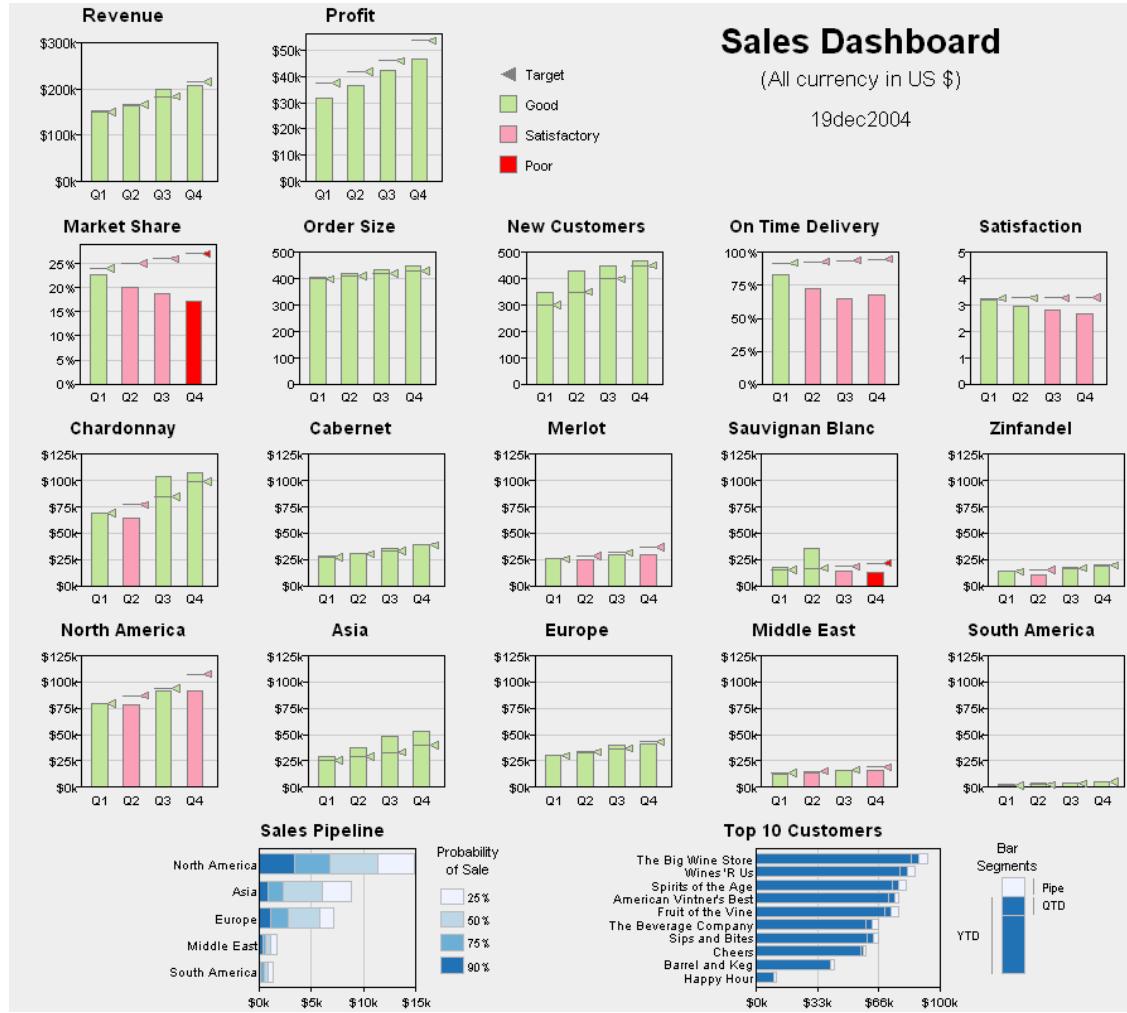


Figure 1: A typical Tableau display of product sale values by region, arranged alphabetically.

Buisness Analytics Tools – Manager Dashboards



Scatterplot Matrix

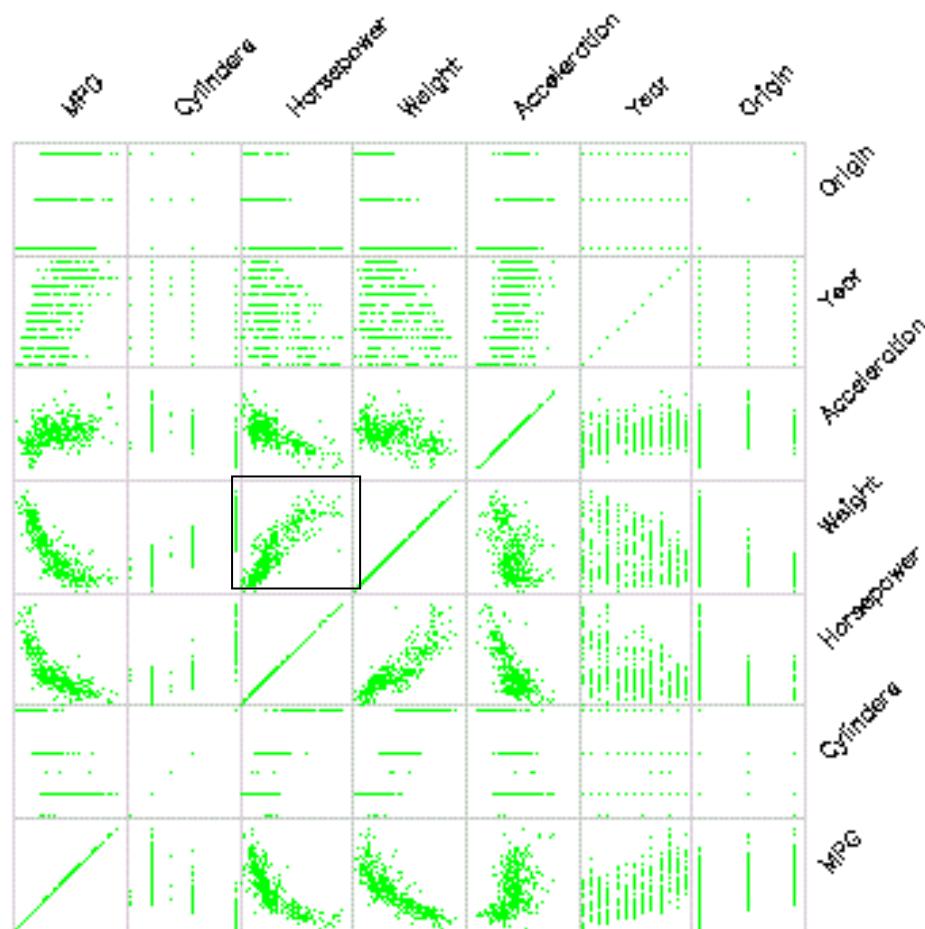
Represent each possible pair of variables in their own 2-D scatterplot (car data)

Q: Useful for what?

A: linear correlations
(e.g. horsepower & weight)

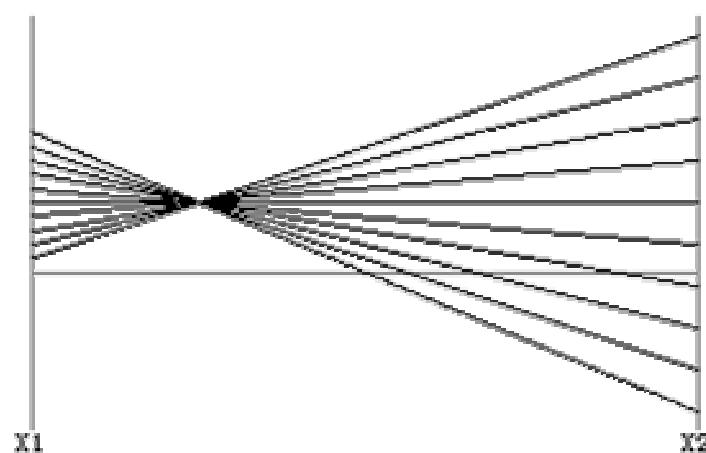
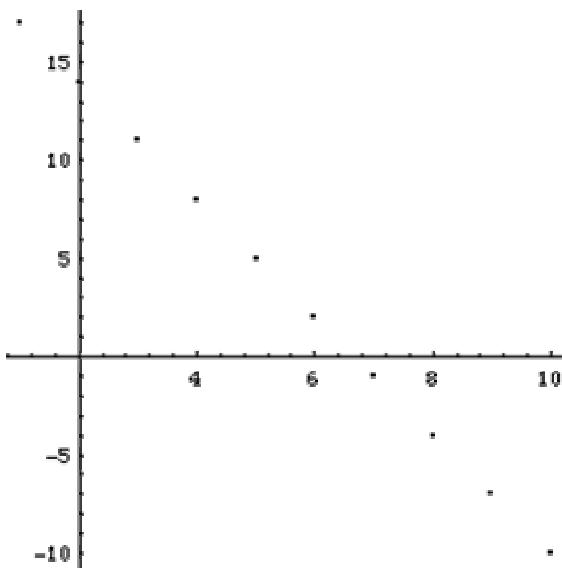
Q: Misses what?

A: multivariate effects



Parallel Coordinates

- Encode variables along a horizontal row
- Vertical line specifies values



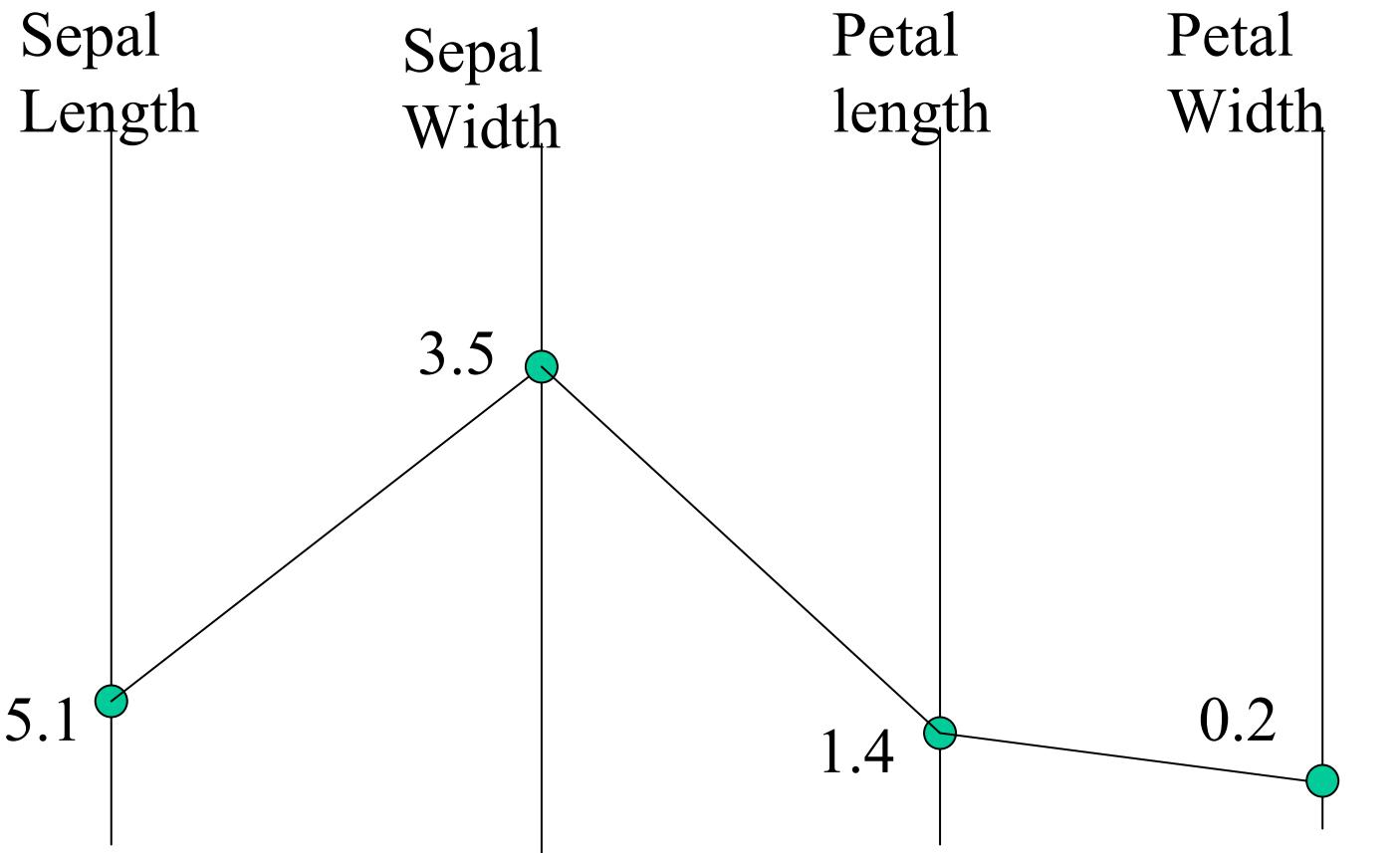
Same dataset in parallel coordinates

Dataset in a Cartesian coordinates

Invented by
Alfred Inselberg
while at IBM,
1985

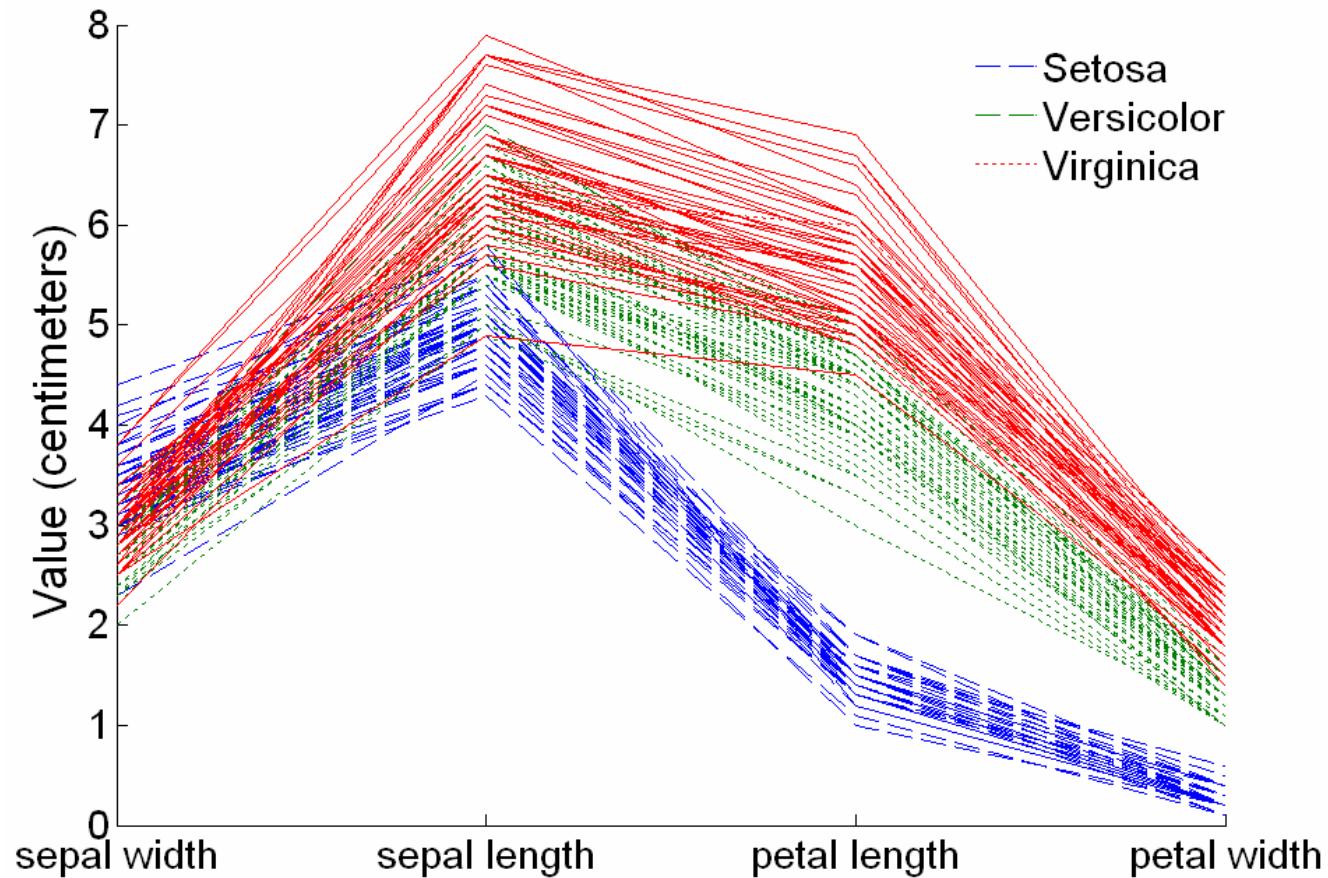


Parallel Coordinates: 4 D



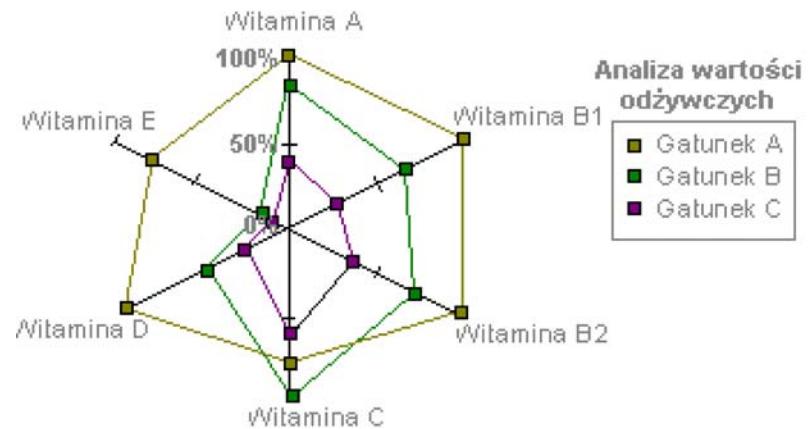
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates Plots for Iris Data

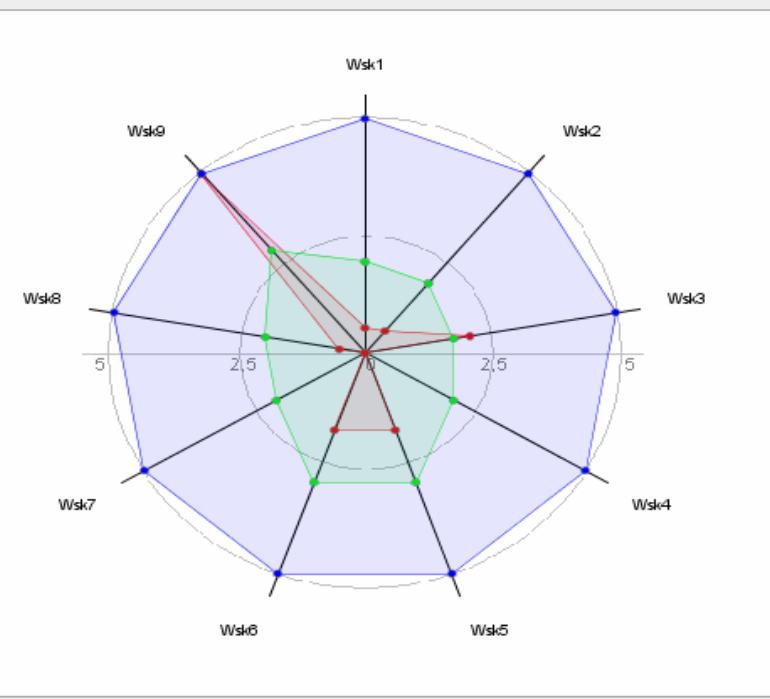


Radar Figures

- Aggregate multidimensional observations
- Each observation gets a separate colour or graph symbols
- Variables corresponds to angles



Wybrana dziedzina



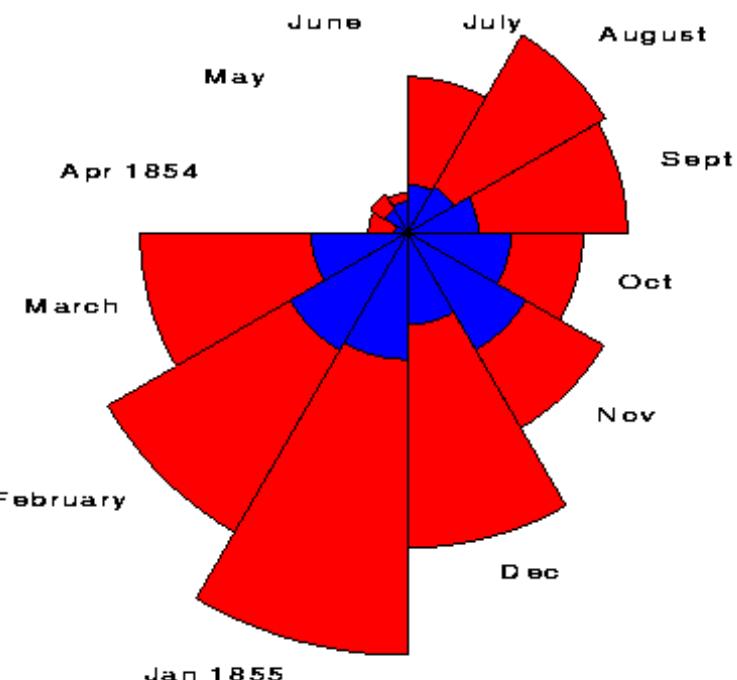
Wykres radarowy – oceny wskaźników w ramach dziedziny I poziom oceny

F. Nightingale (1856) – abstract representation

Causes of Mortality in the Army in the East
April, 1854 to March 1855

■ Non-Battle

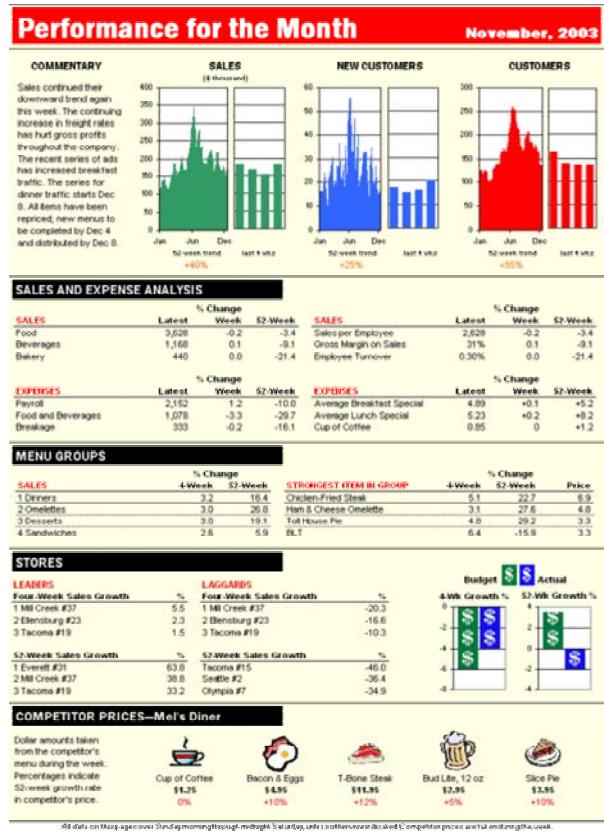
■ Battle



From: F. Nightingale, "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army", 1858

Buisness Analytics Tools – Typical Reports

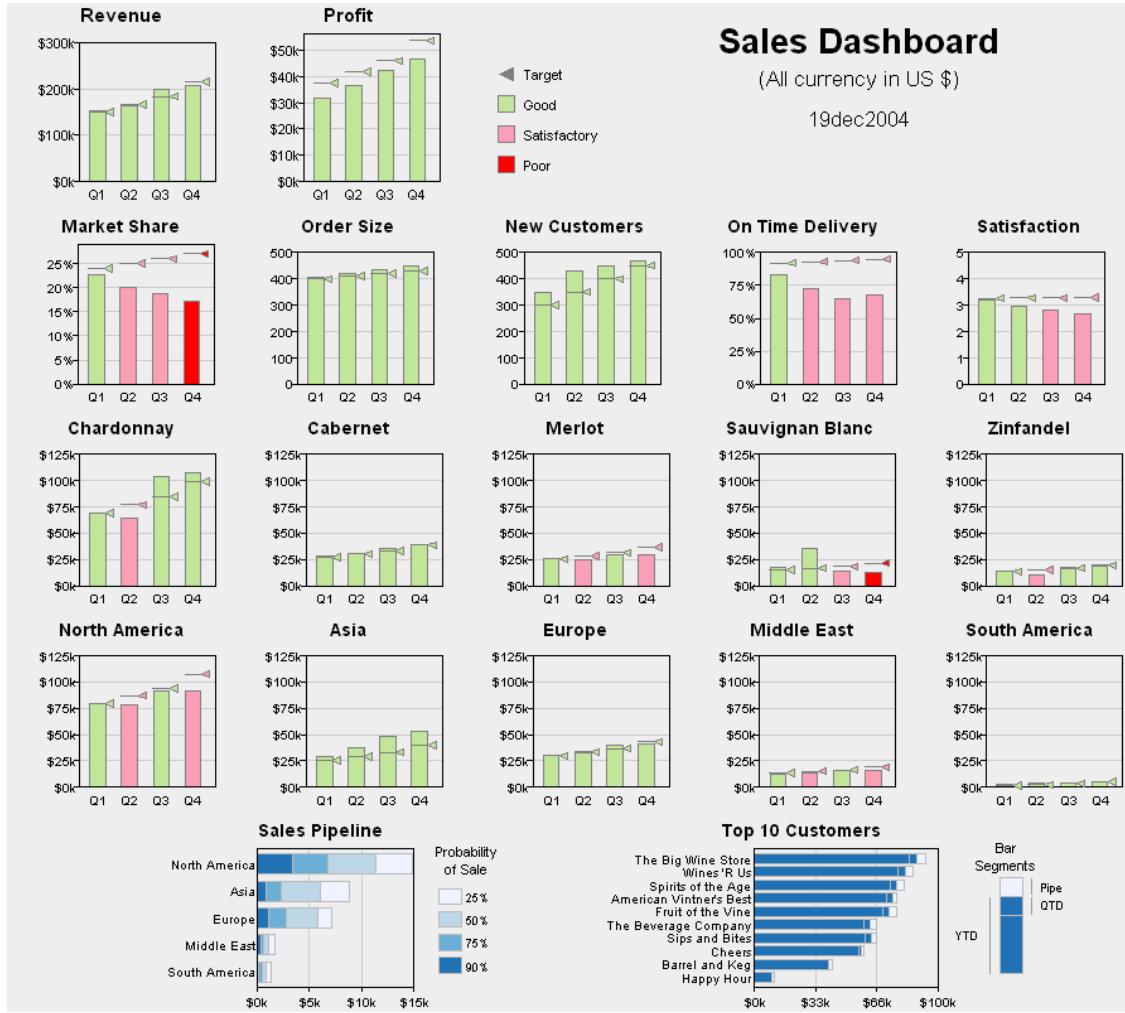
Raport more traditional



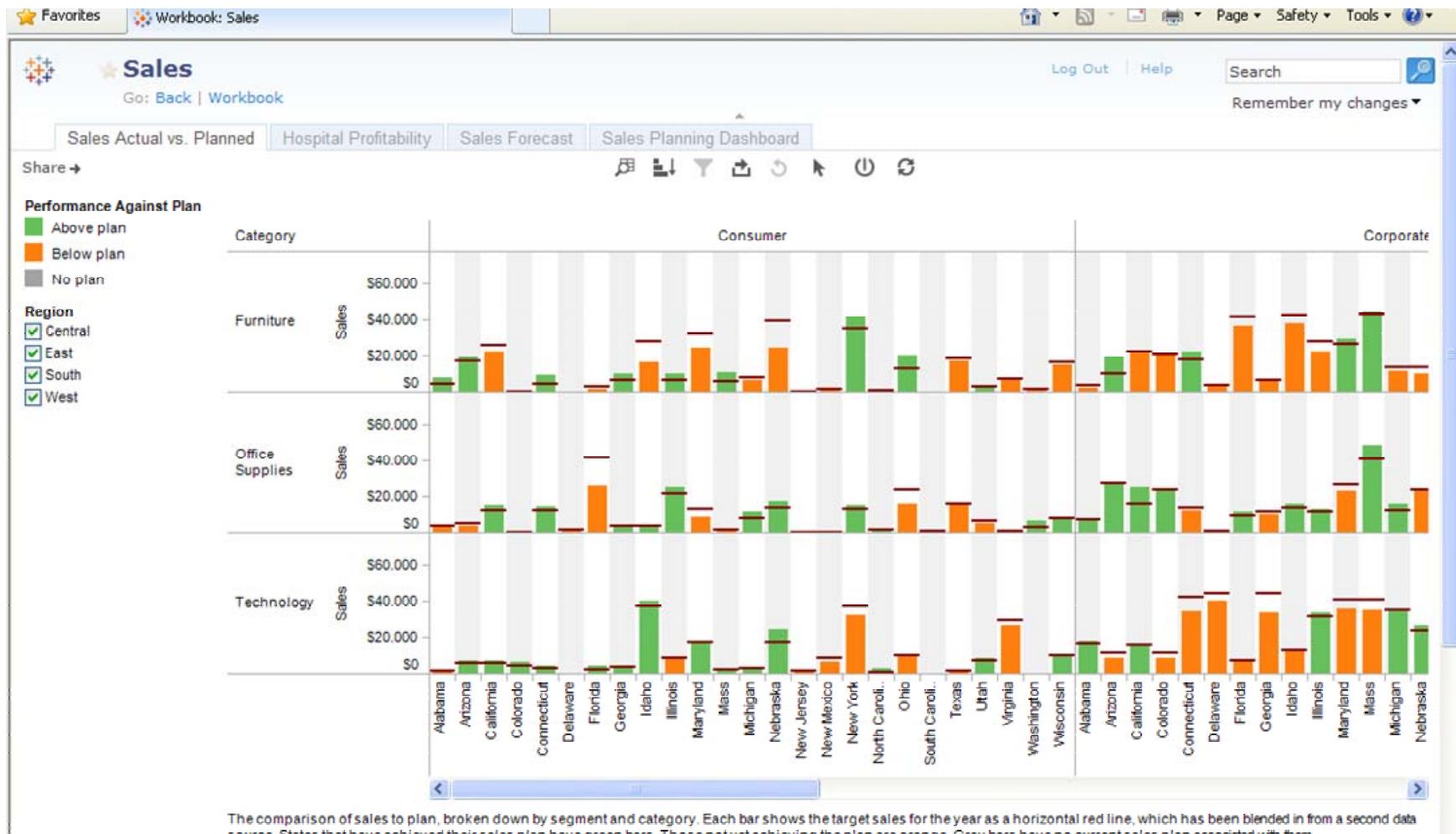
Other forms



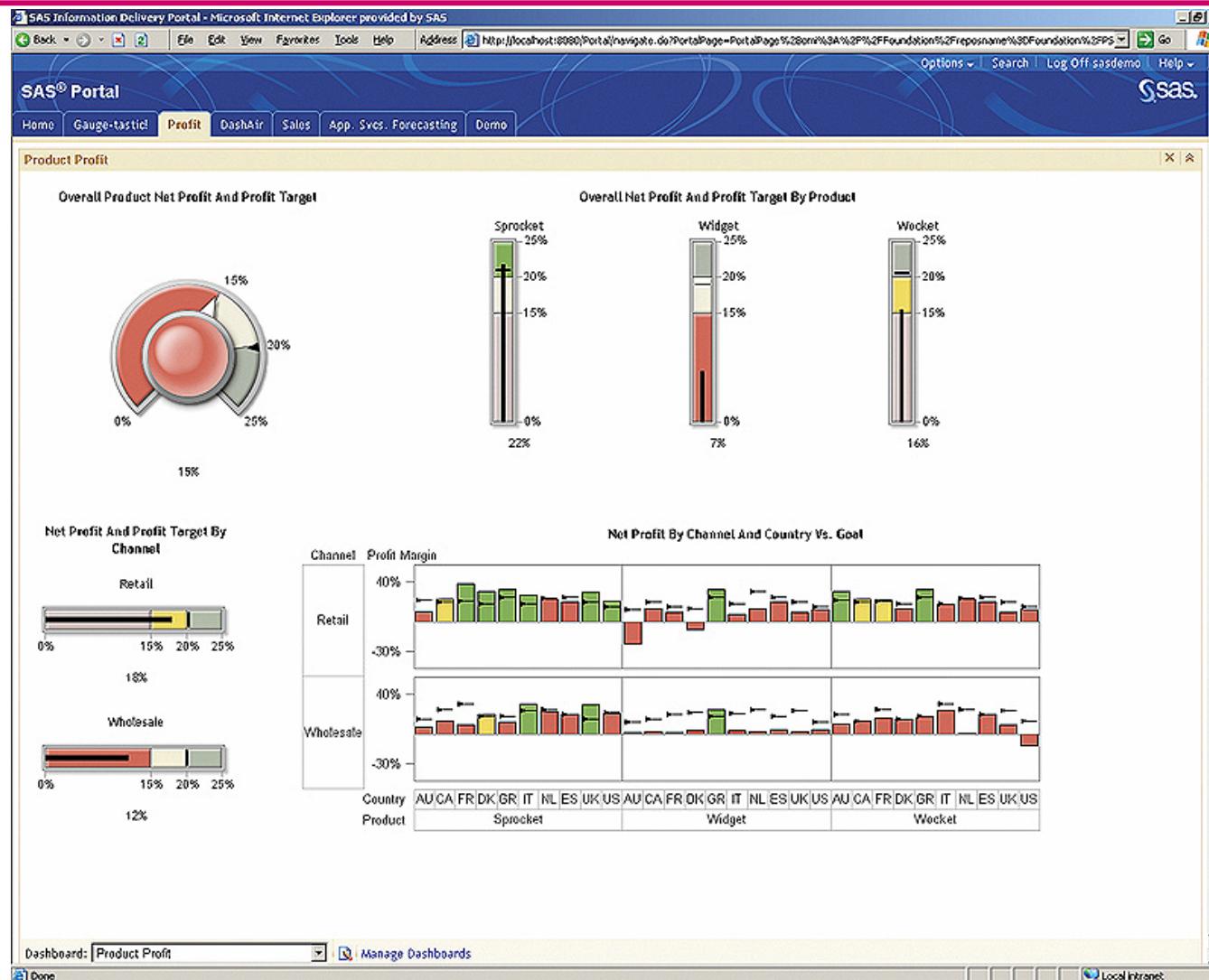
Buisness Analytics Tools – Manager Dashboards



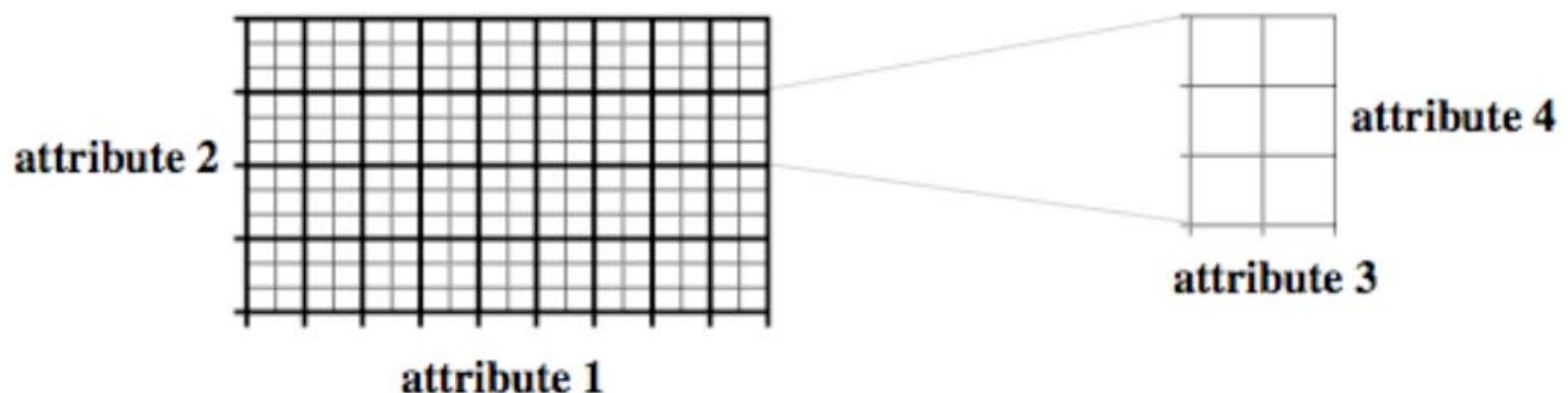
Bars in business dashboards – Tableau Software



Data analytics – kokpity menadżerskie



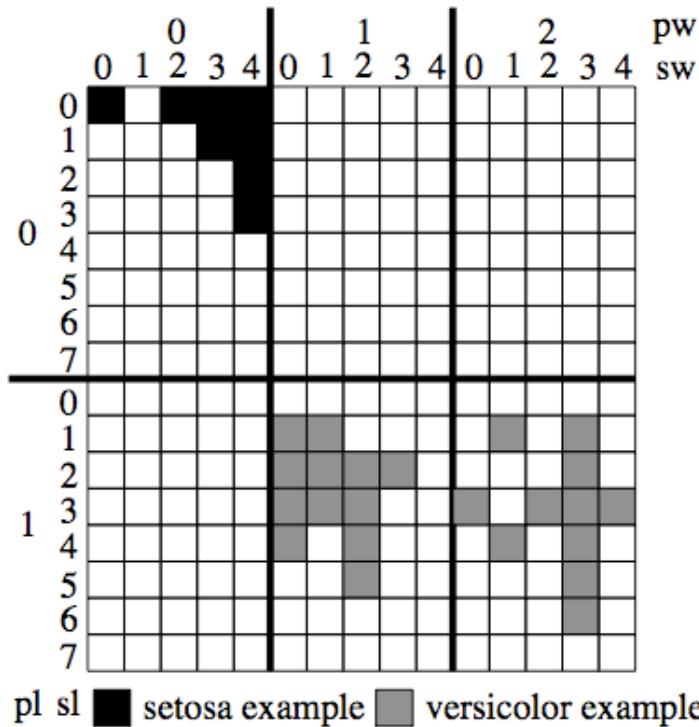
Multidimensional Stacking



Multidimensional presentation of nominal attributes

- VL1 diagrams (Michalski 70) for machine learning

Visualization of Training Examples: Discrete Version of the Iris Data Set



		Size	Size	Size
		S	M	L
Green	T	■		
	C		■	
	R			■
Blue	T	■		
	C		■	
	R			■
Red	T	■		
	C		■	
	R			■
Color	Shape			
Target				
concept,				
$t = 1 \dots 40.$				
Target				
concept,				
$t = 41 \dots 80.$				
Target				
concept, $t = 81 \dots 120.$				

STAGGER and concept drift

Hierarchiczne wizualizacje - Treemaps

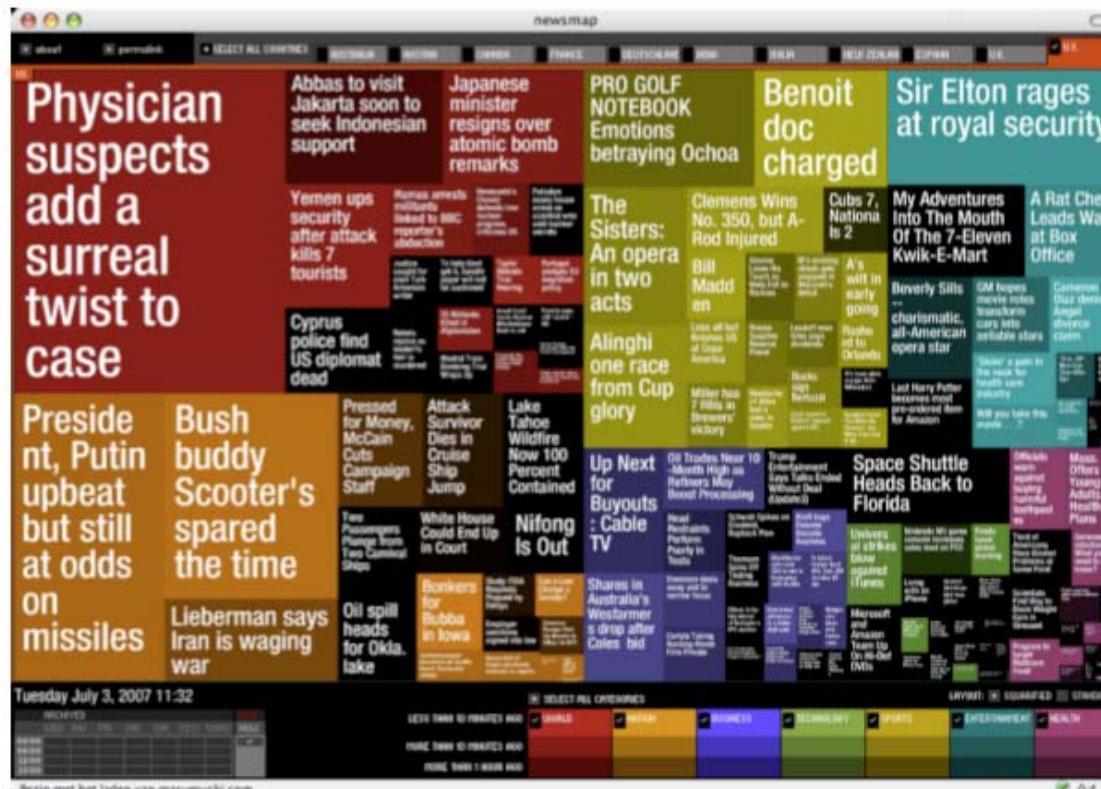


Figure 4: Treemap used for displaying news <http://newsmap.jp/>

- Treemaps display hierarchical data using rectangles. Each branch of the tree is assigned a rectangle. Then each sub-branch gets assigned to a rectangle and this continues recursively until a leaf node is found.
 - Depending on choice the rectangle representing the leaf node is colored, sized or both according to chosen attributes.

The Top 1 Percent: What Jobs Do They Have?

Explore the occupations and industries of the nation's wealthiest households.

Rectangles are sized according to the number of people in the top 1 percent. Color shows the percentage of people within that occupation and industry in the top 1 percent.

1% 10% 20%

With 376,076 members, the largest single group in the 1 percent are those who listed their occupation as a manager.

[ZOOM TO](#)

Lawyers who work on Wall Street are twice as likely as those in general practice to make the top 1 percent.

[ZOOM TO](#)

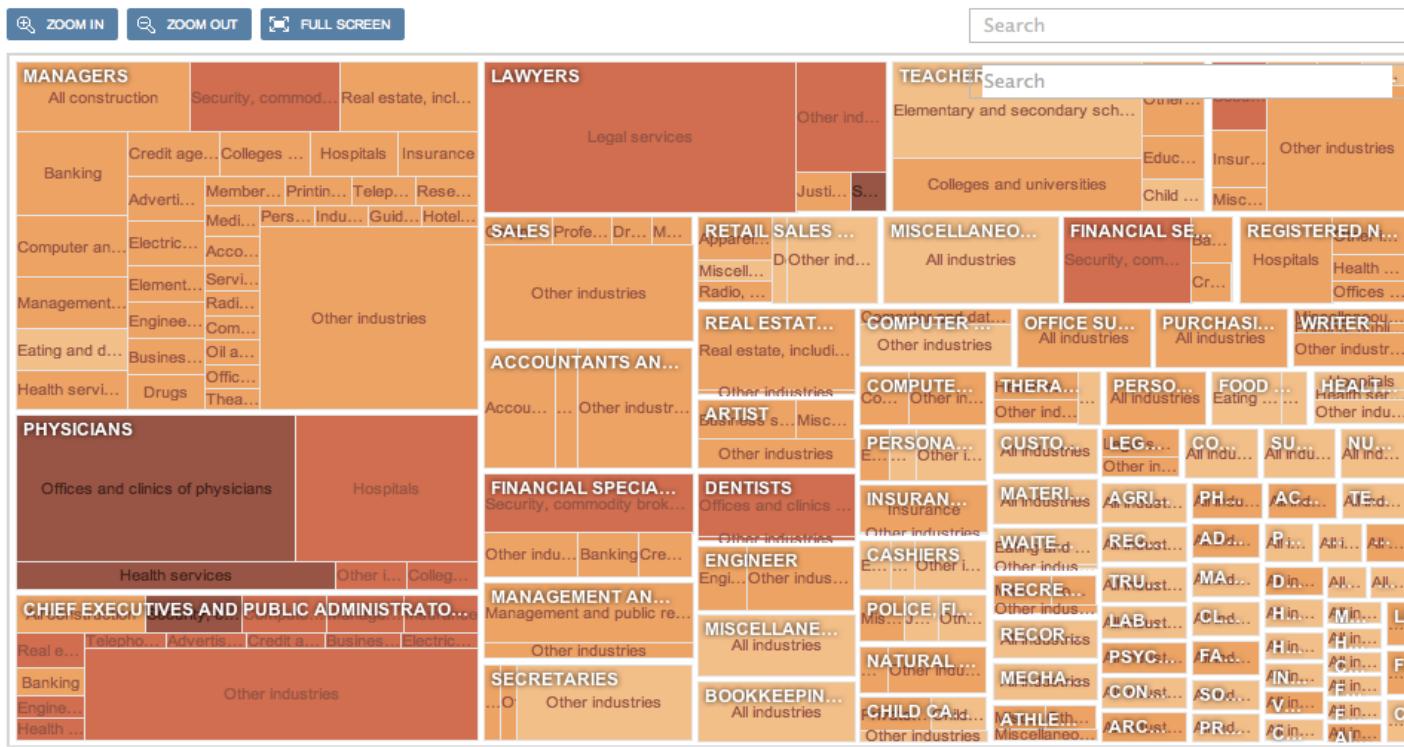
Physicians who work primarily in doctor's offices are somewhat more likely to make the cutoff, though all doctors are well-represented in the group.

[ZOOM TO](#)

School teachers don't earn enough to make the top 1 percent on their own, but many live in 1-percent households, primarily through marriage.

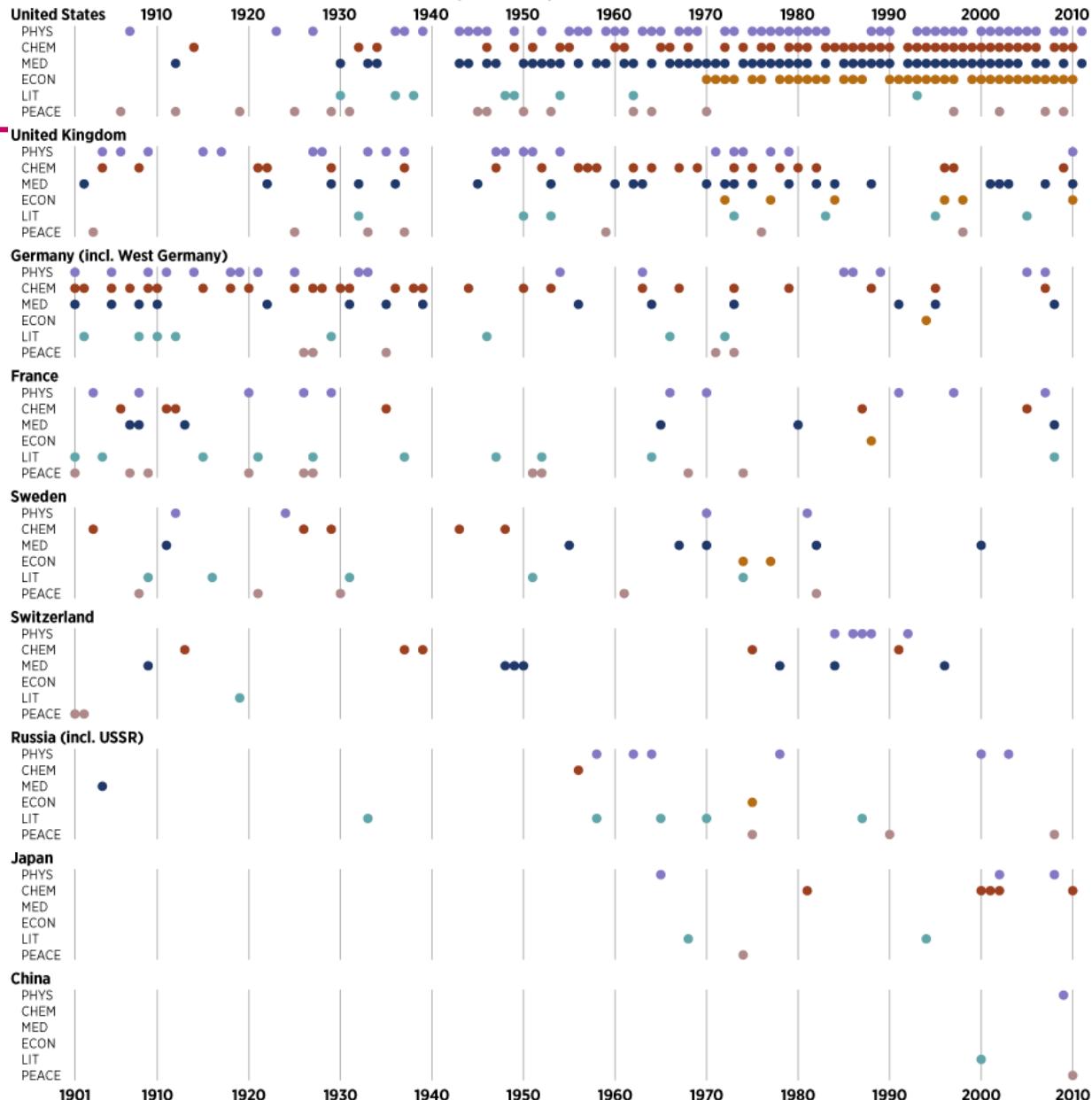
[ZOOM TO](#)

Note: The chart counts the number of individual workers living in households with an overall income in the top 1 percent nationwide.



100 Years of Nobel Laureates

By Country and Prize



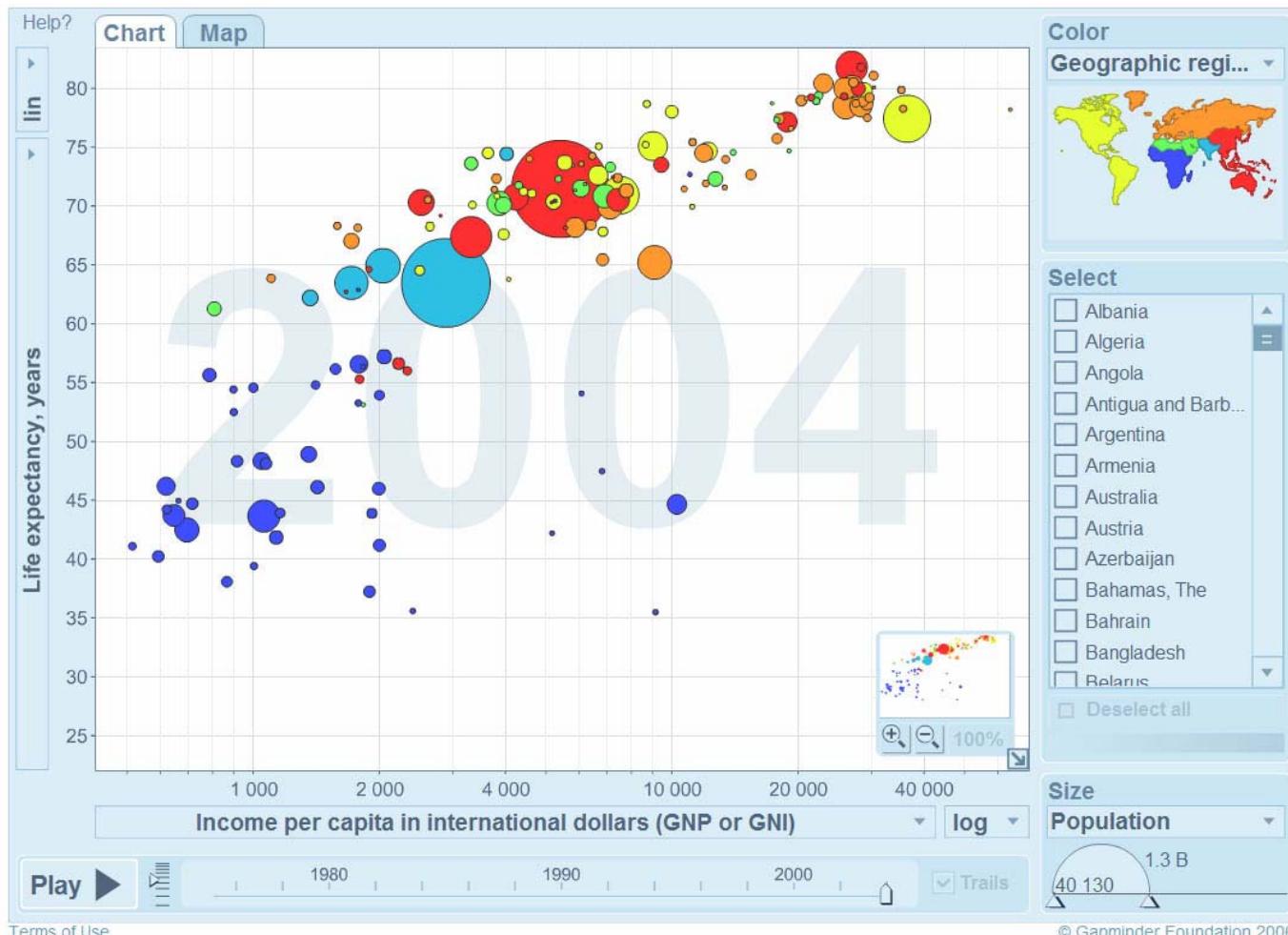
SOURCE: NOBELPRIZE.ORG

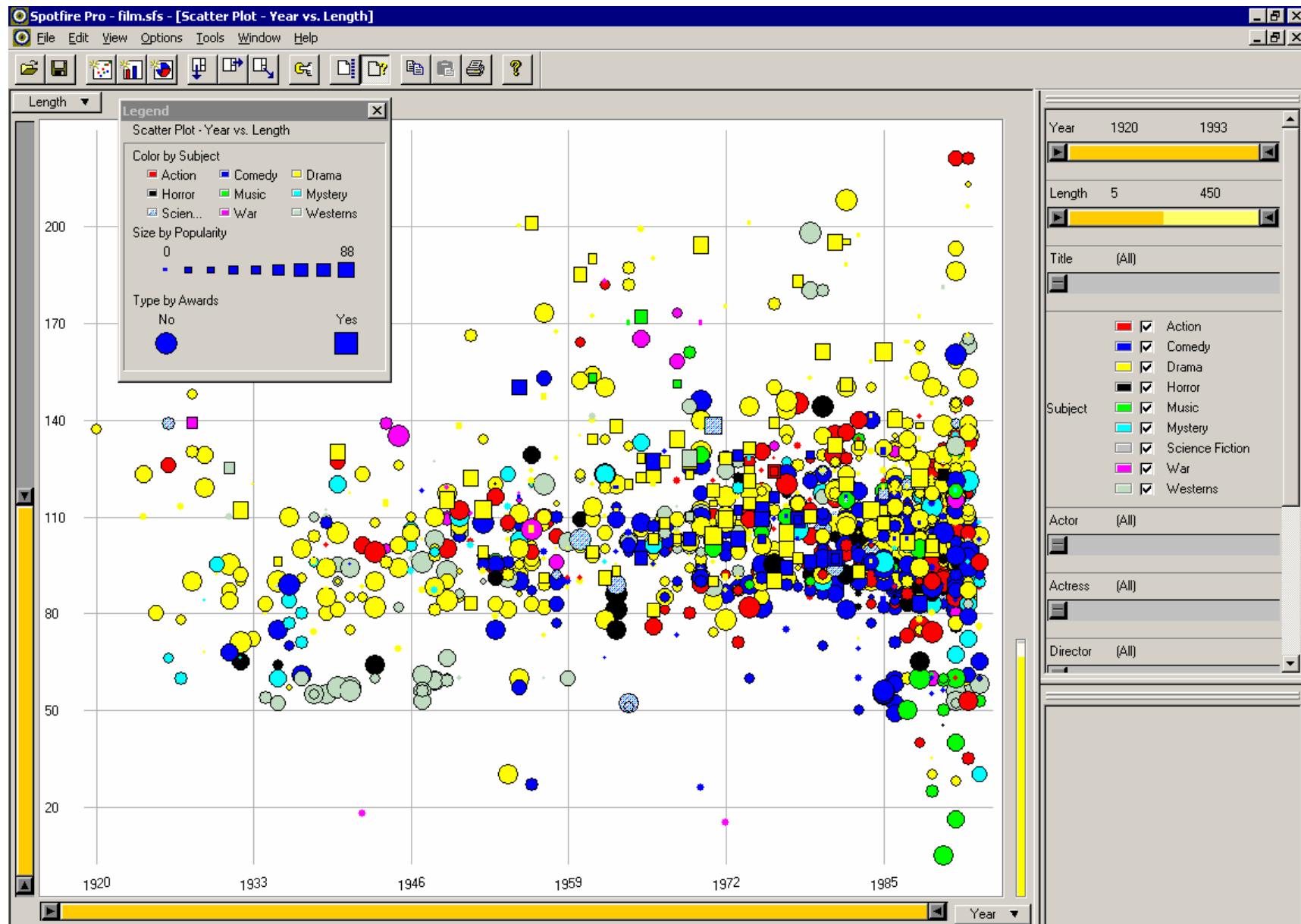
Laureates are shown in the country that hosted their research at the time of award

Last updated on October 4, 2011

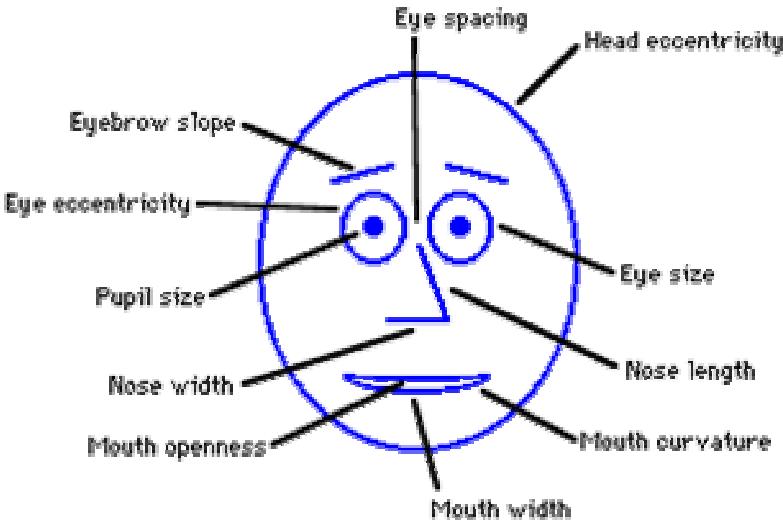
Forbes blogs.forbes.com/jonbruner

Gapminder – Motion Charts





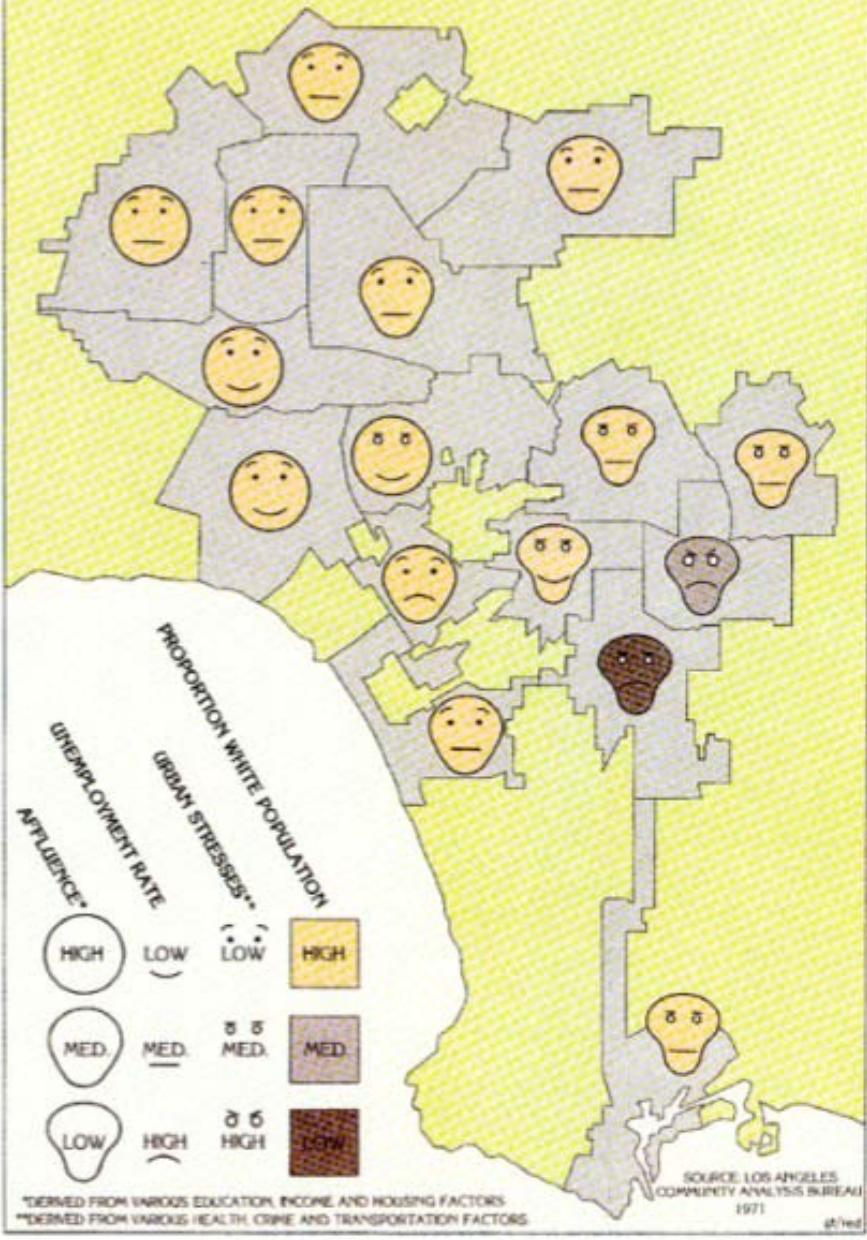
Encode different variables' values in characteristics of human face

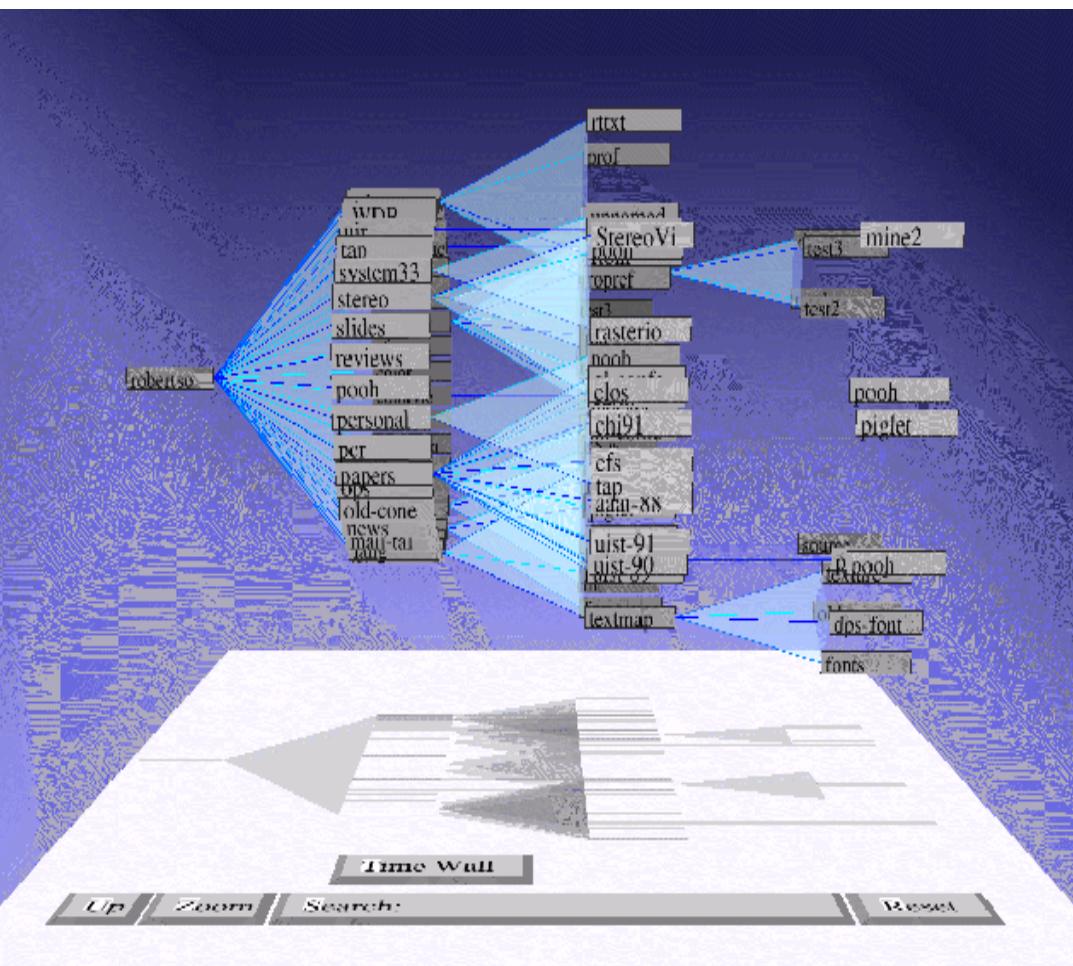


Cute applets: <http://www.cs.uchicago.edu/~wiseman/chernoff/>

<http://hesketh.com/schampeo/projects/Faces/chernoff.html>

Life in Los Angeles

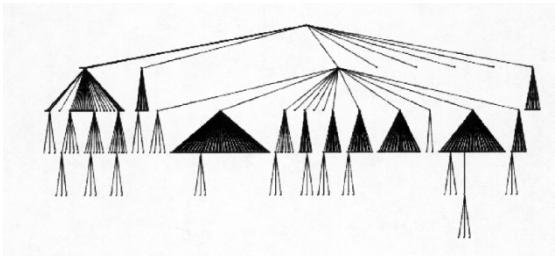




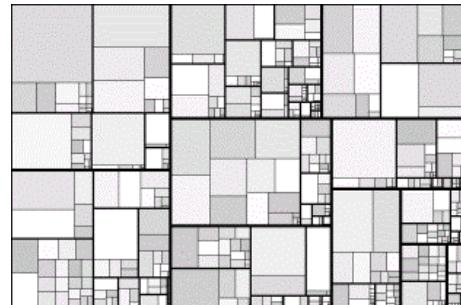
Cone Trees [RMC91]

- animated 3D visualizations of hierarchical data
- file system structure visualized as a cone tree

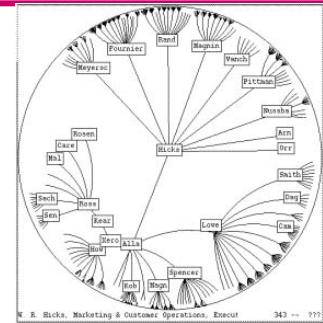
Abstract → Hierarchical Information – Preview



Traditional



Treemap



Hyperbolic Tree

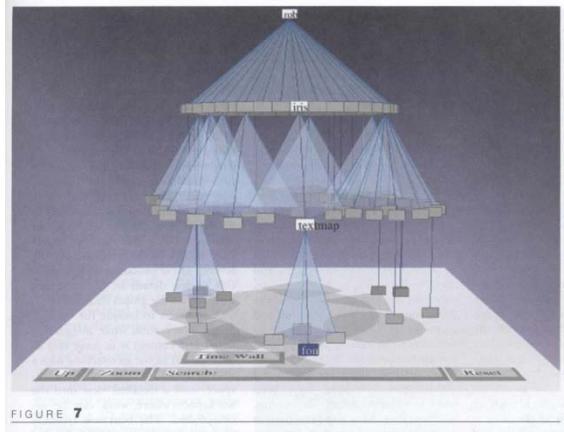
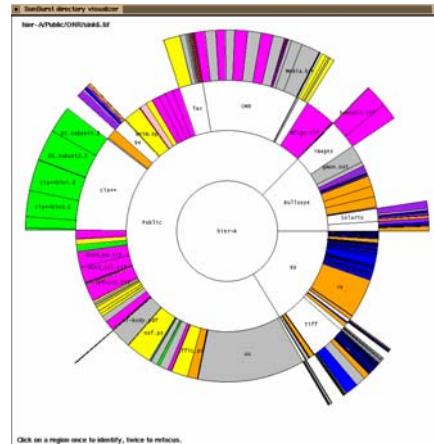
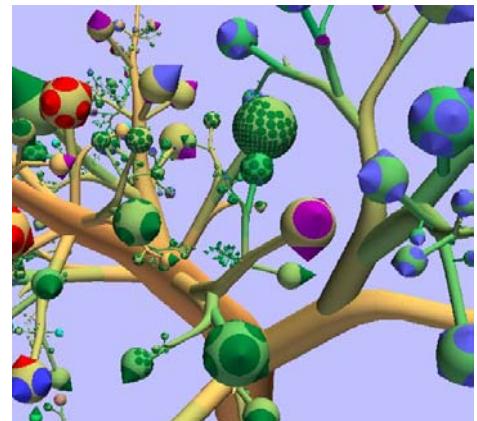


FIGURE 7
Cone Tree visualization of a directory hierarchy

ConeTree

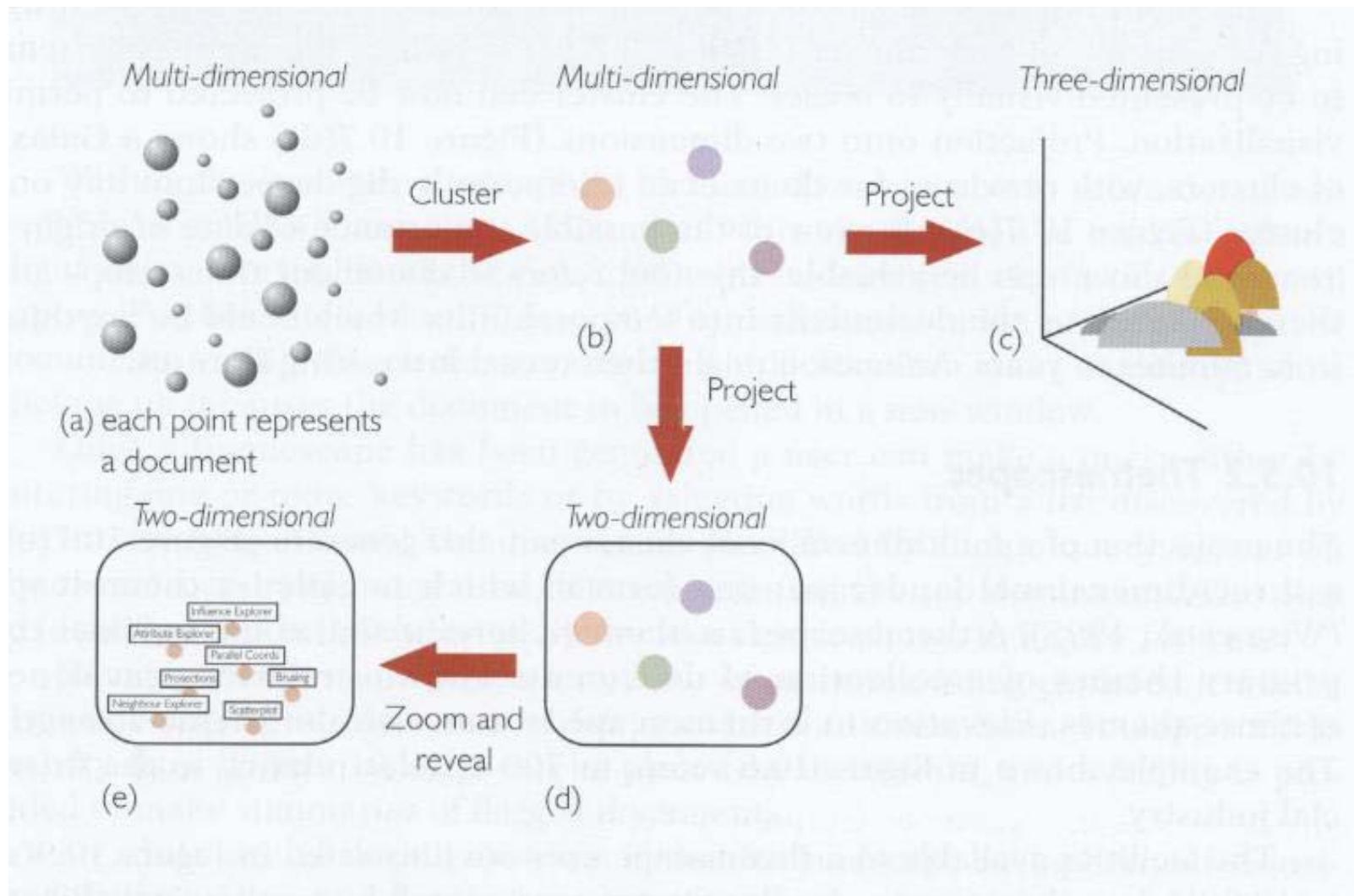


SunTree

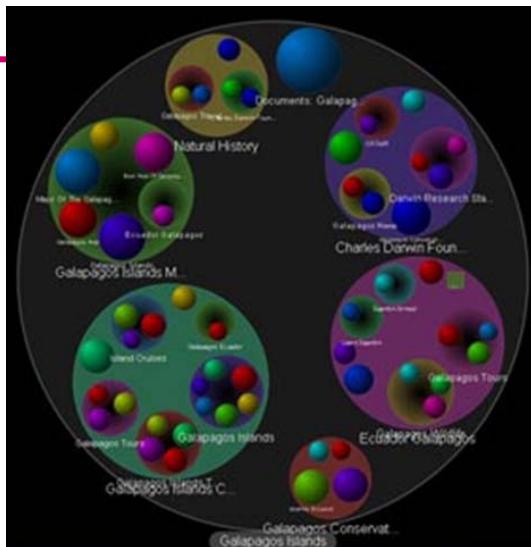


Botanical

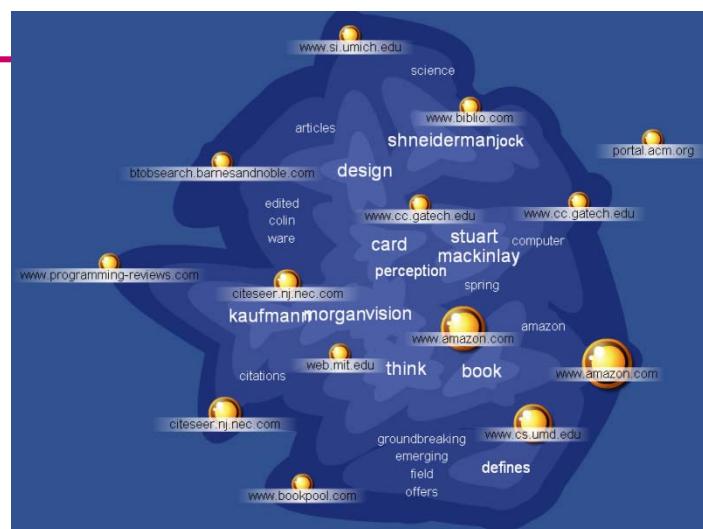
Visualization of Search Results & Inter-Document Similarities



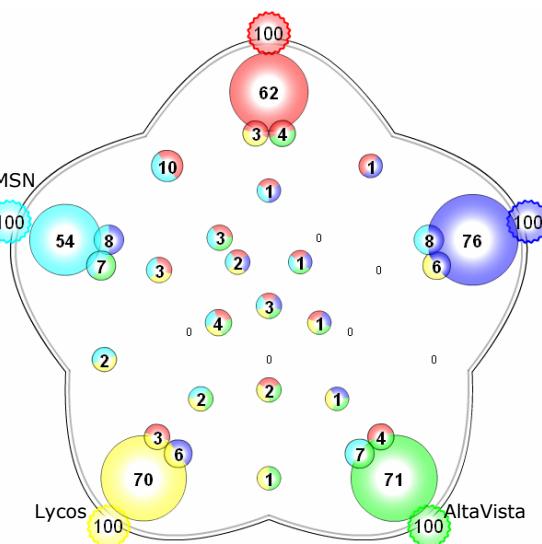
Abstract → Text – **MetaSearch** Previews



Grokker



Kartoo

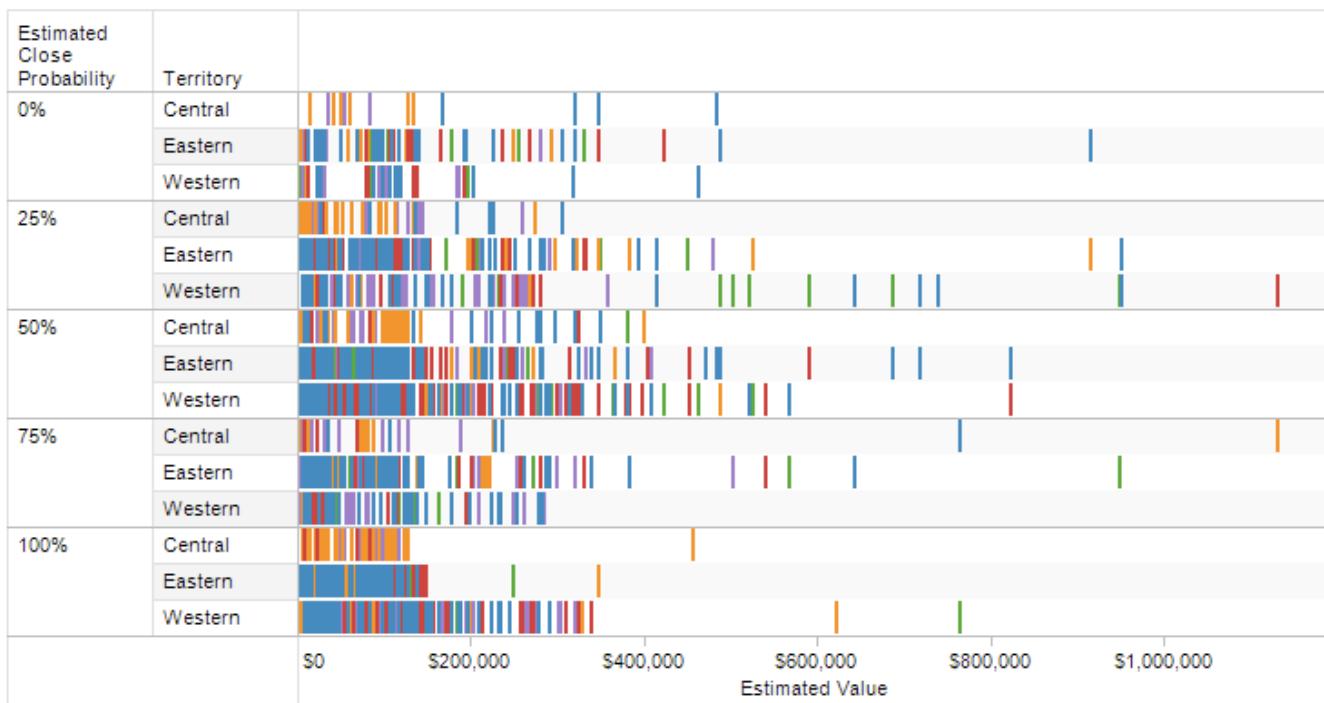


MetaCrystal → searchCrystal

Analysis of CRM data

[Opportunity Map](#)
[Actual vs. Forecast](#)
[Distribution of Deal Size](#)
[Current Pipeline](#)
[Current Pipeline Report](#)

Opportunity Map



INDUSTRY NAME

- Automotive
- Financial Services
- Healthcare
- Manufacturing
- Retail

INDUSTRY NAME

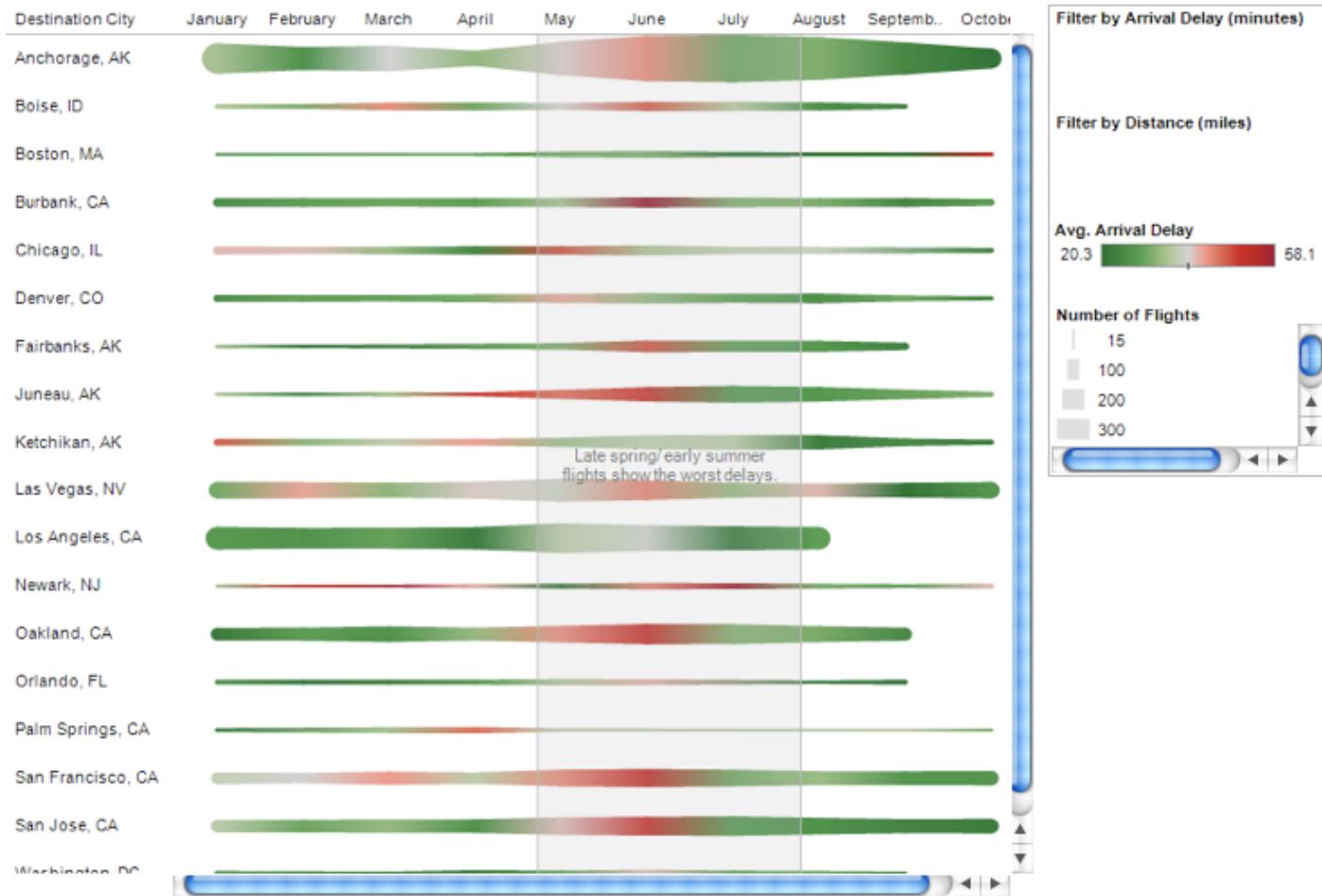
- (All)
- Automotive
- Financial Services
- Healthcare
- Manufacturing
- Retail

Visualization shows a mark for each deal in the pipeline; mark is placed based on deal size, region and probability stage (estimated percent chance a deal will closed this quarter). Color is by Industry. Click the "+" above territory to drill down to individual salespeople.

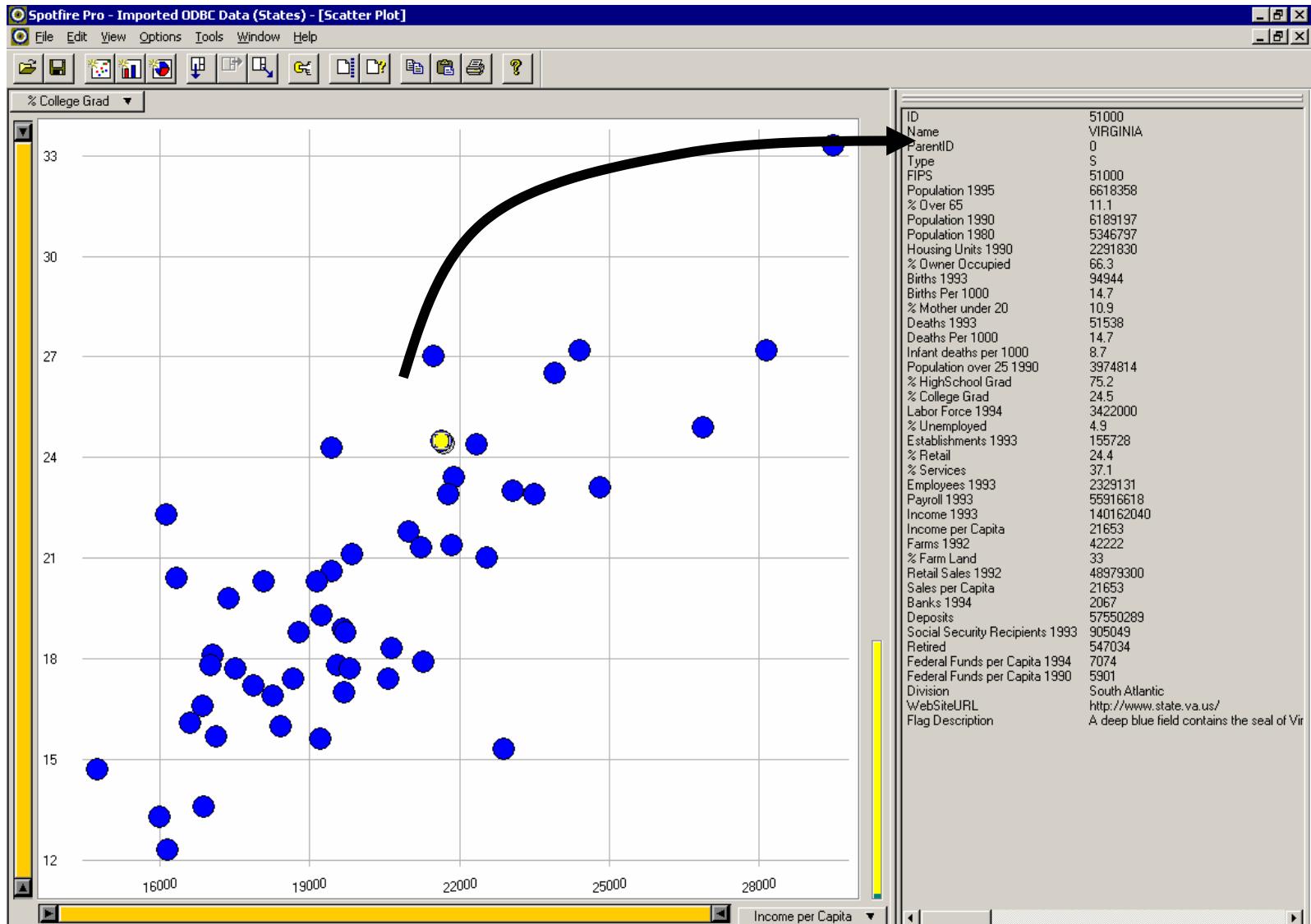
Visualization of different conditions

Severity of Flight Delays

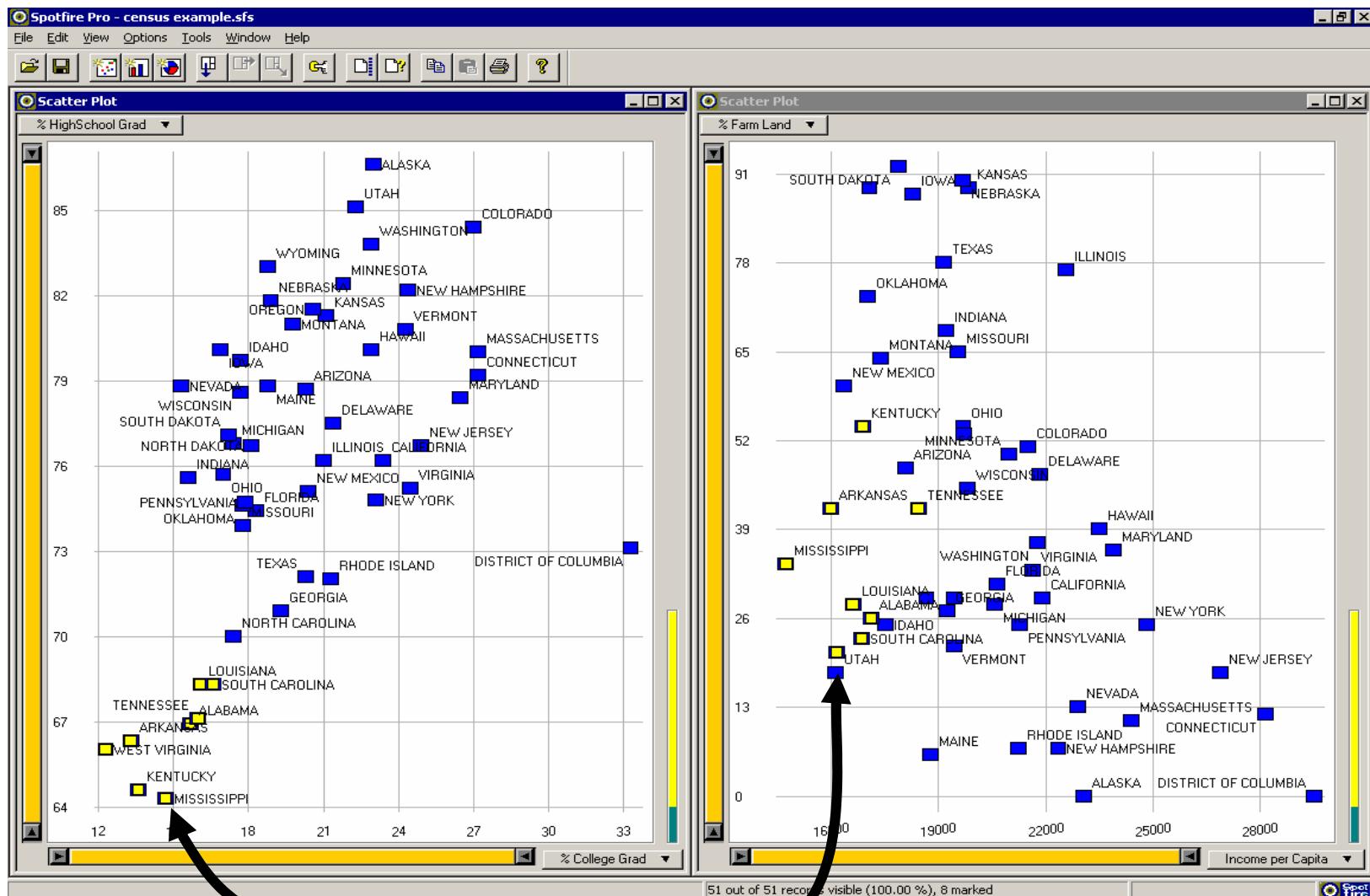
Flight Delays from Seattle



Overview and Detail



Brushing and Linking



Census Data

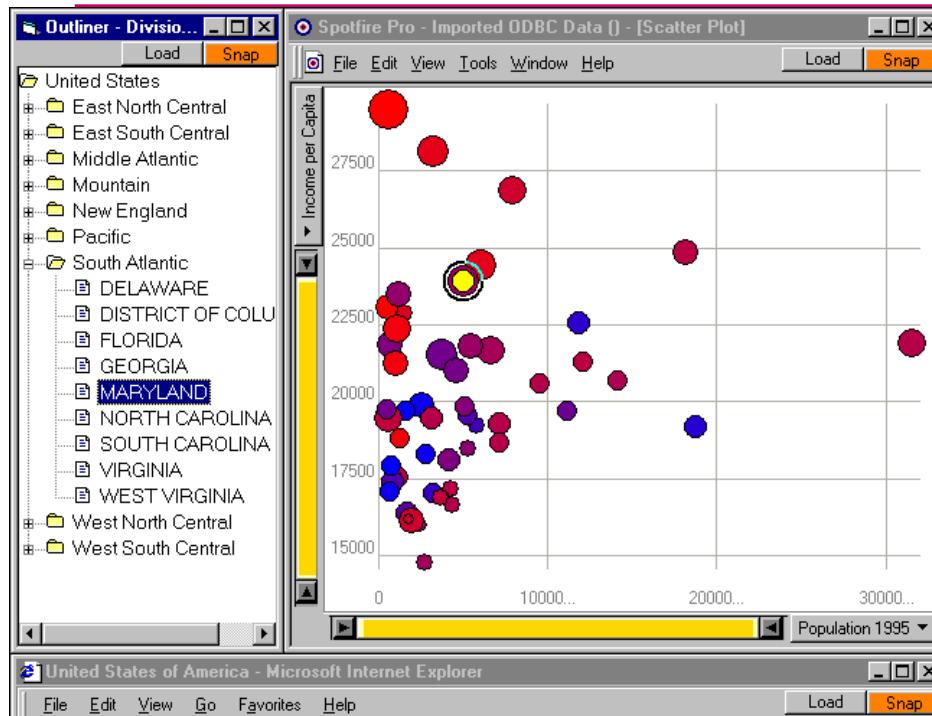
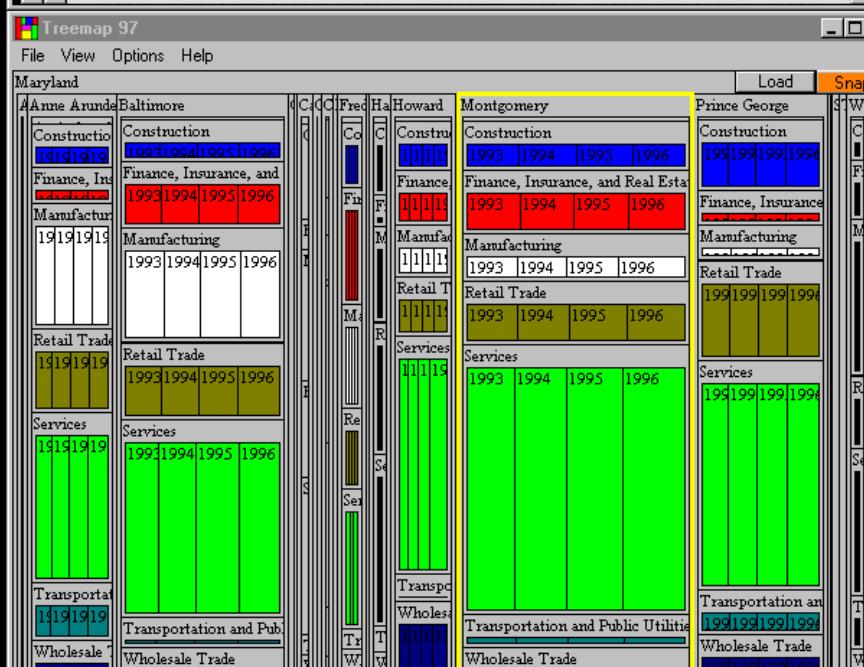
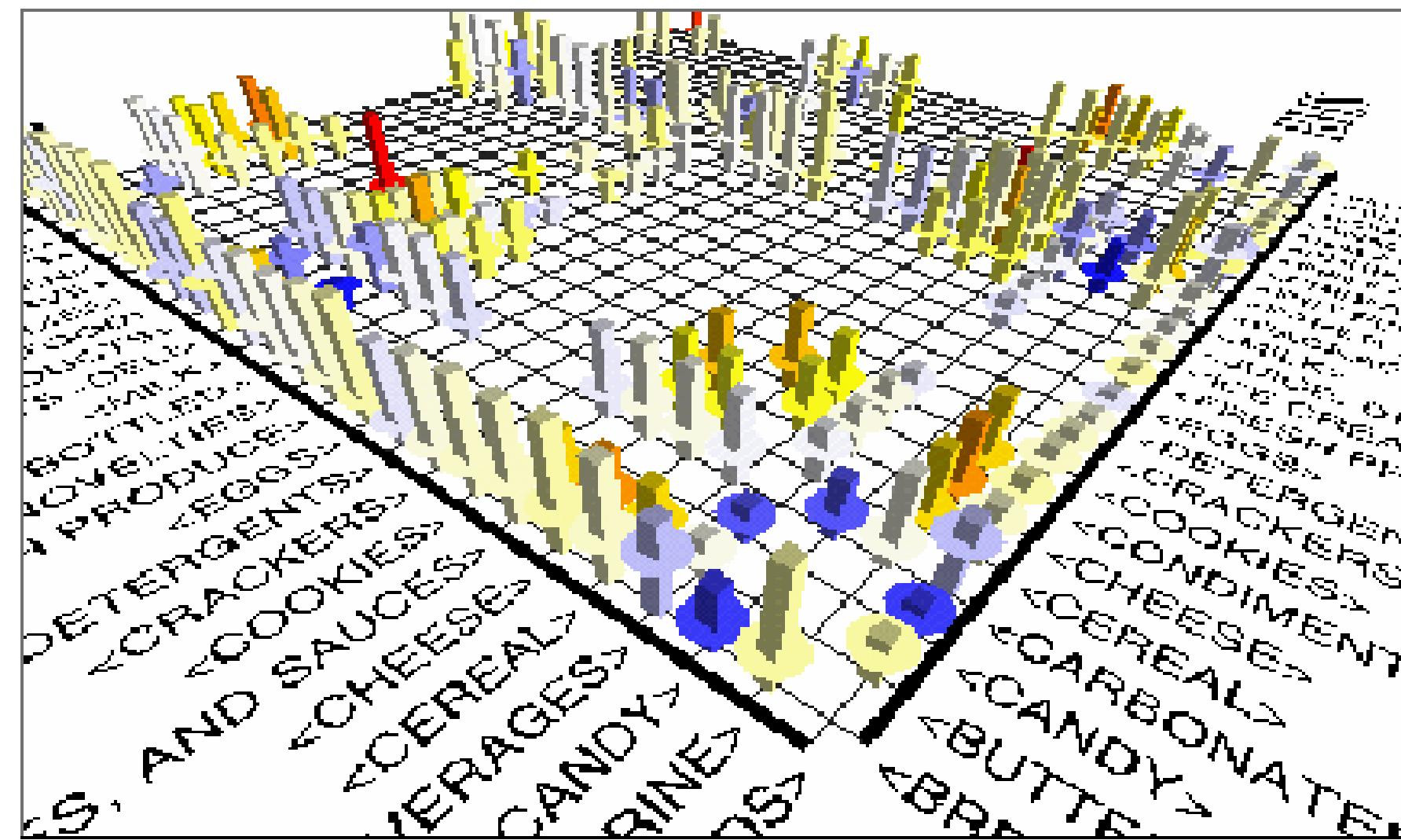


Table - Counties of a State (24000) Load Snap

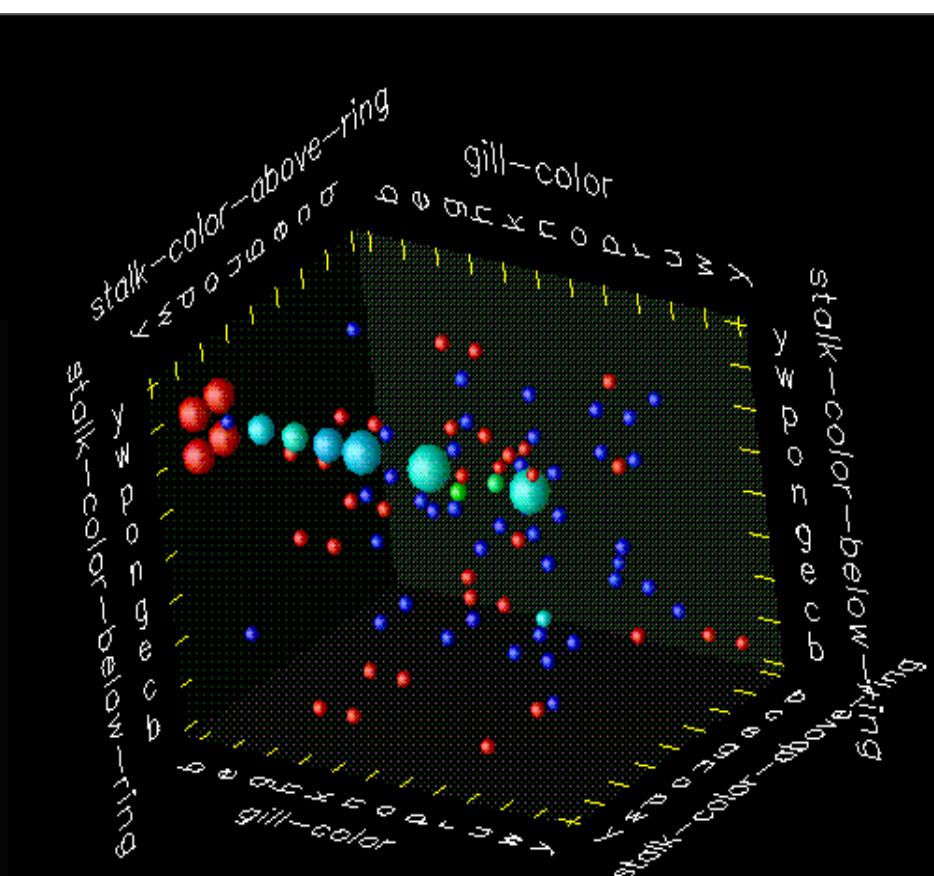
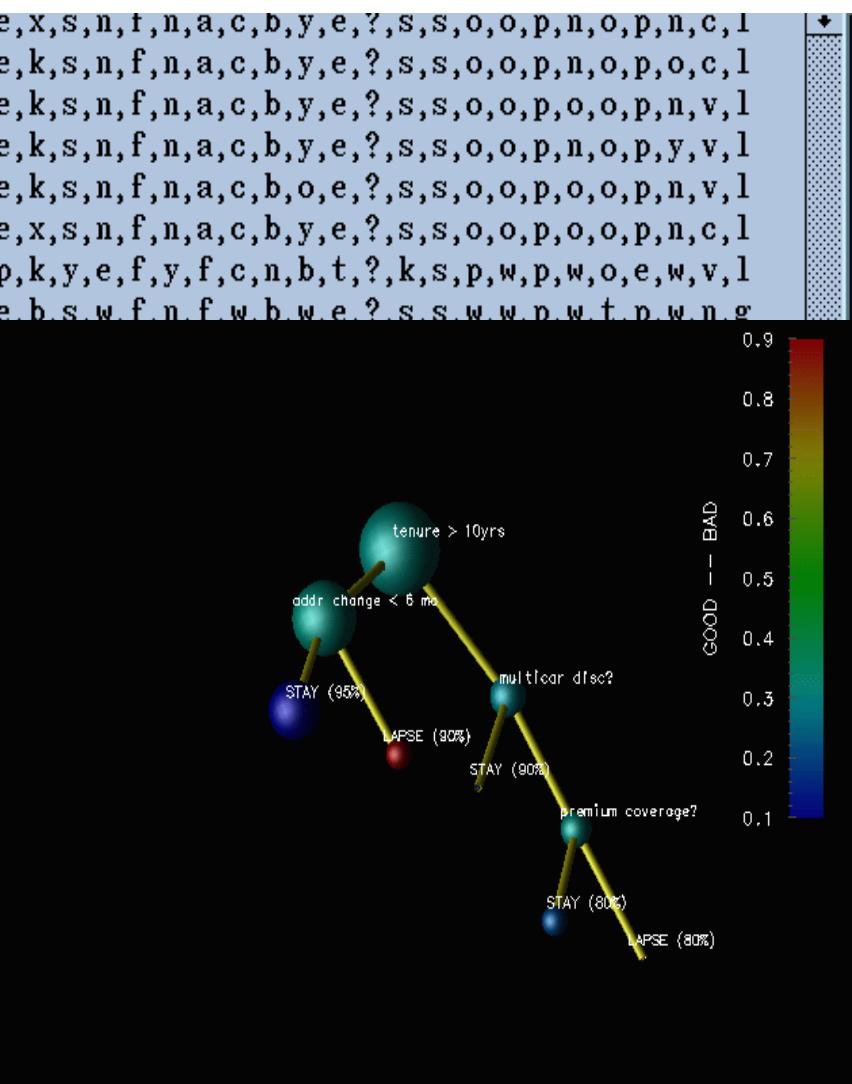
Name	Population 1995	Population 1990	Population 1980	Housing Units 1990
Baltimore, MD	715360	692134	655615	268280
Calvert, MD	64598	51372	34638	16986
Caroline, MD	29072	27035	23143	9983
Carroll, MD	140203	123372	96356	42248
Cecil, MD	78174	71347	60430	24725
Charles, MD	111633	101154	72751	32950
Dorchester, MD	30170	30236	30623	12117
Frederick, MD	175399	150208	114792	52570
Garrett, MD	29461	28138	26490	10110
Harford, MD	205367	182132	145930	63193
Howard, MD	219125	187328	118572	68337
Kent, MD	18736	17842	16695	6702
Montgomery, MD	809569	757027	579053	282228
Prince George's, MD	767413	728553	665071	258011
Queen Anne's, MD	36992	33953	25508	12489
Somerset, MD	24431	23440	19188	7977



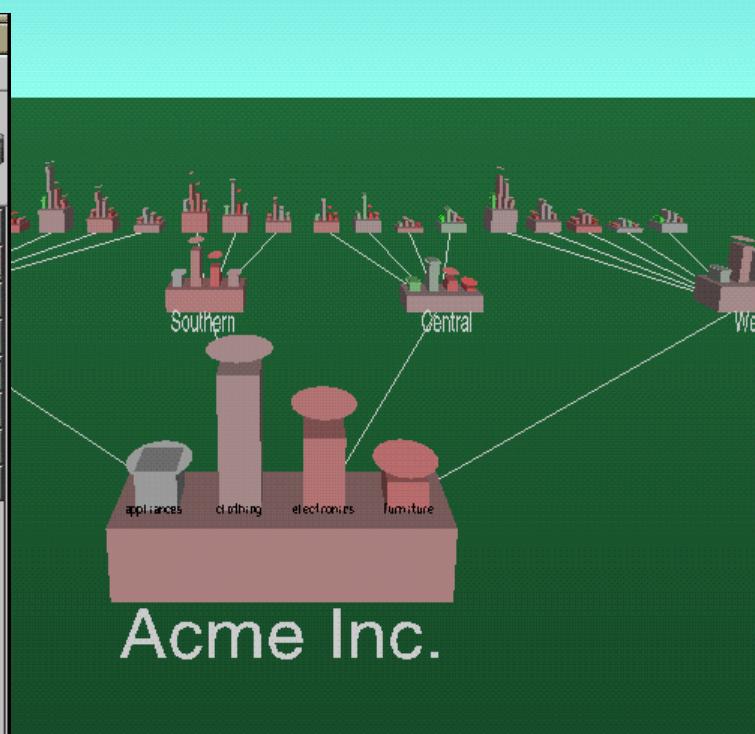
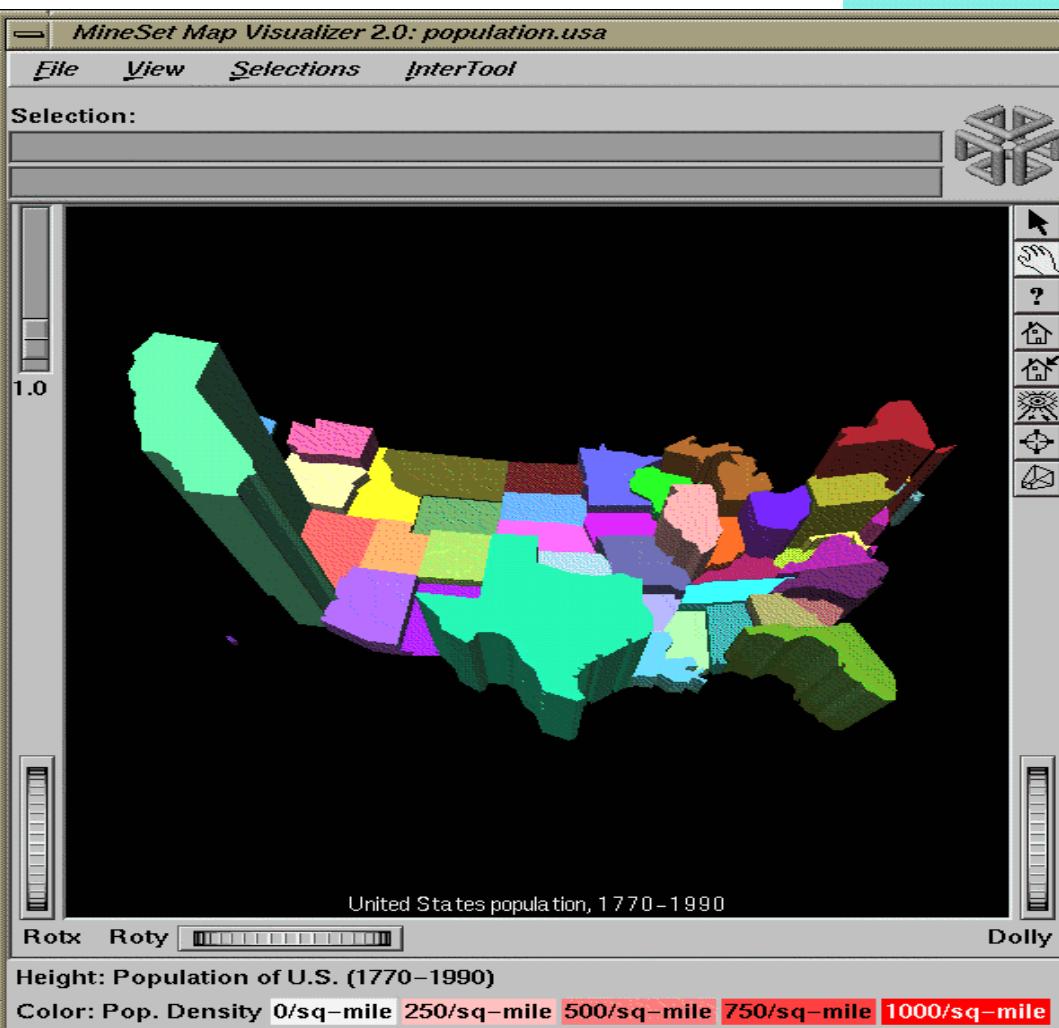
Visualization of Association Rules in SGI/MineSet 3.0



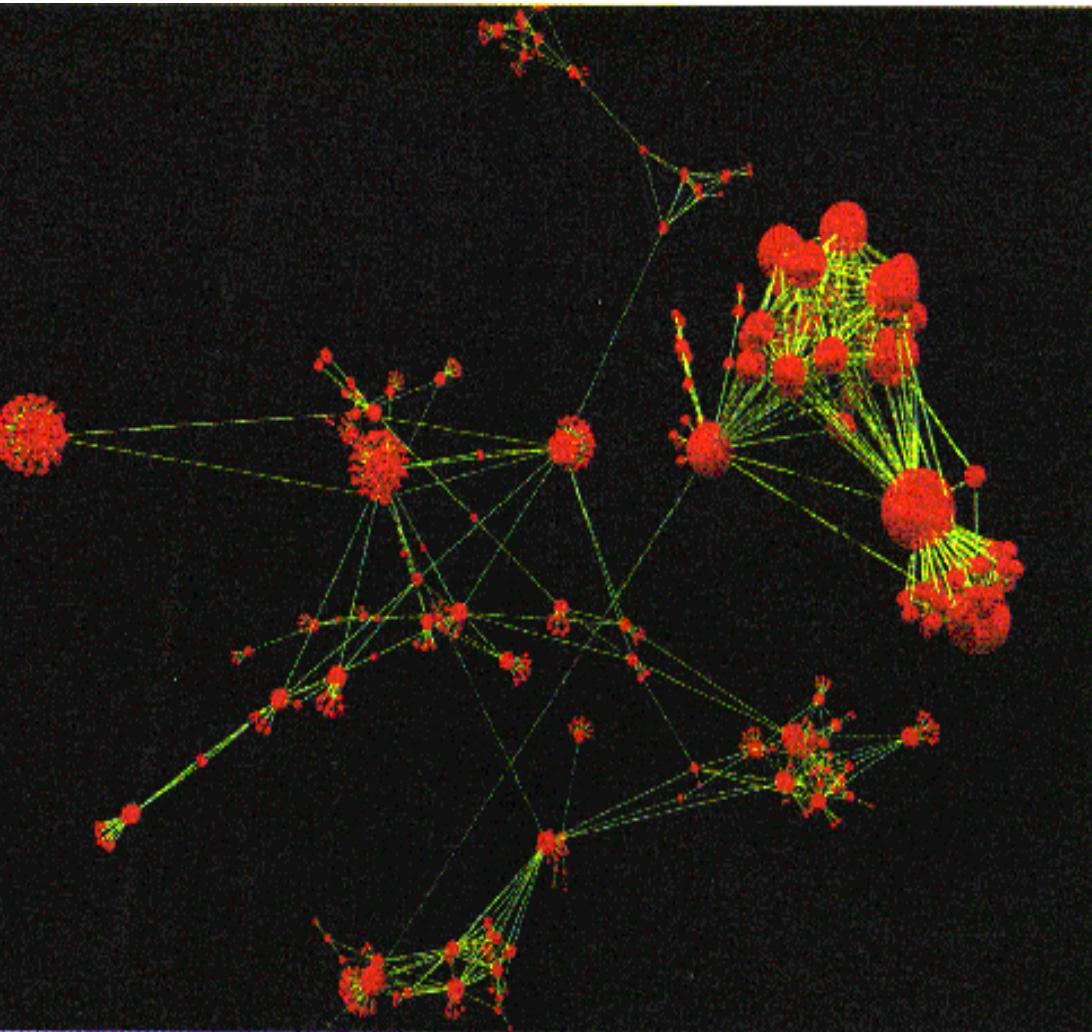
IBM Miner – visualization of mining results



SGI – other tools



Graph-based Techniques



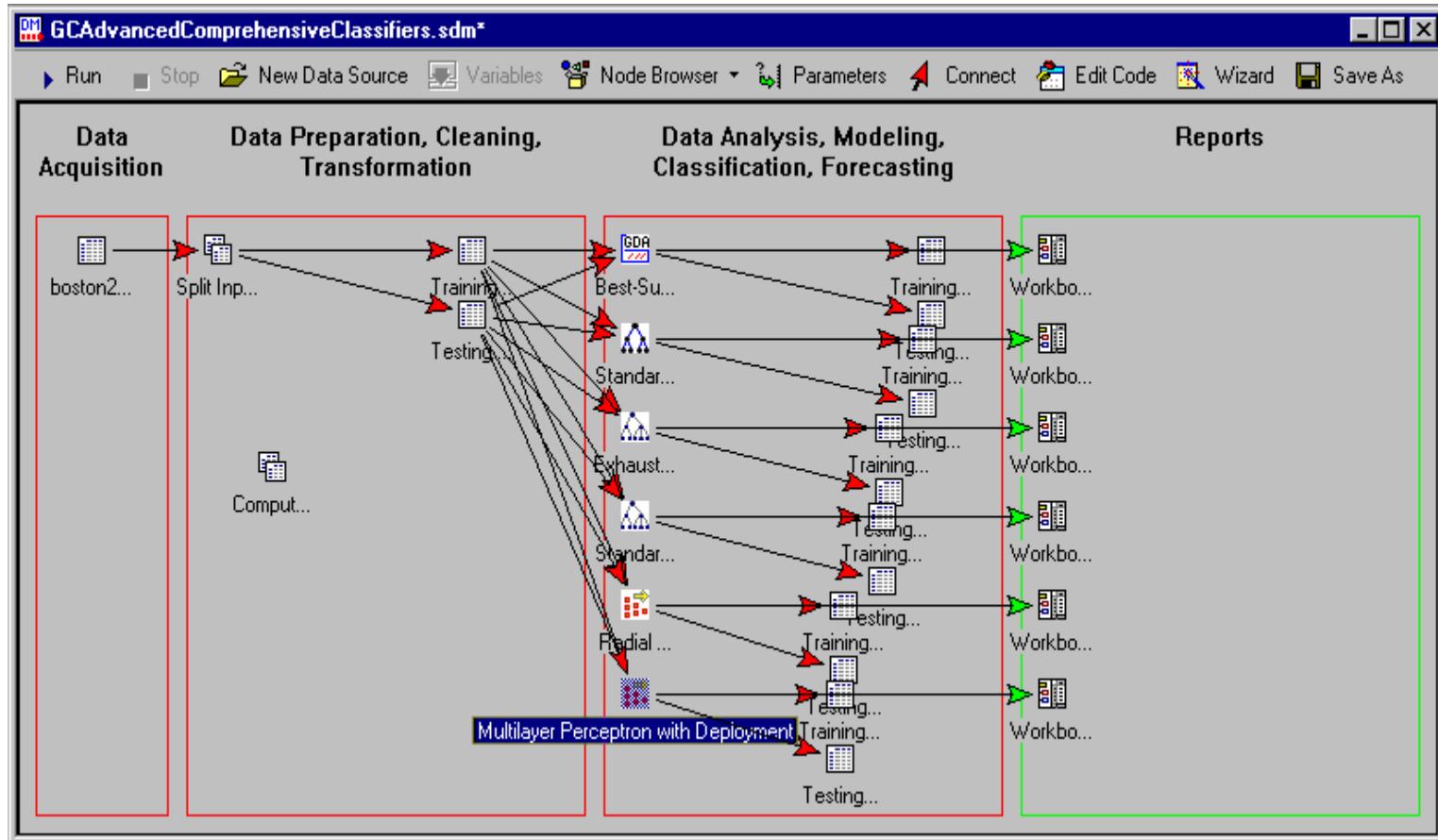
Narcissus

- Visualization of a large number of web pages
- visualization of complex highly interconnected data

Visualization of knowledge discovery process

- A graphical tool for arranging components / steps of KDD
- Just a graph flow of actions
- Graphical objects – plug and place
- Parametrization
- Often → you may produce a kind of script representing a graphical flow of KD process

Statsoft – Data mining graphical panel





[HOME](#) [SEARCH](#) [SITEMAP](#) [LEGAL](#) [CONTACT US](#) [DEUTSCH](#)

[PRODUCTS](#) [DOWNLOADS](#) [SERVICES](#) [COMMUNITY](#) [ABOUT US](#)

TESTIMONIALS

"I have encountered various learning environments, but none so broad, powerful, and easy-to-use as RapidMiner / YALE. Many of us who are not skilled in programming are thankful."

Roberto E. Ferrer, Venezuela

DOWNLOADS

[RapidMiner / YALE](#)

[RapidMiner / YALE Plugins](#)

[RapidMiner / YALE Documentation](#)

[RapidMiner / YALE Interactive Tour](#)

TRAINING SEMINARS

[Data Mining for Marketing and Customer Service](#)

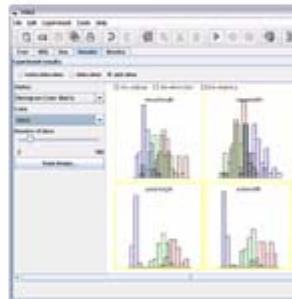
[Data Mining Techniques: Theory and Practice](#)

[Extending RapidMiner and Integration as a Data](#)

[HOME](#) : [PRODUCTS](#) : [RAPIDMINER \(YALE\)](#) : [SCREENSHOTS](#)

RAPIDMINER / YALE SCREENSHOTS

This web page provides a selection of screenshots for RapidMiner (formerly YALE). These pictures might help you to get a first impression of the abilities of RapidMiner. This page contains a large number of images. Please be patient until all pictures were loaded.



Tukey's recommendations

Summary

- We should always take John Tukey's "*There is no excuse for failing to plot and look*" to heart
- "*A picture is worth a thousand words*" is still (mostly) true, but as statisticians we should read it more like "*A full graphical analysis involves drawing a thousand pictures*"
- Following only a few guidelines, we can make sure that we create sensible (non-standard) plots that transport the right message
- Exploration graphics and diagnostic graphics should more and more become one as they serve the same goal – data analysis

Tufte's 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!



E.R. Tufte, “The Visual Display of Quantitative Information”, 2nd edition)



Look for other references
And play with different software tools
Excel is not the only and best software

Thank you for you coming to my lecture and asking questions!



Last Slide
It's not over...