Data Visualization

or Graphical Data Presentation

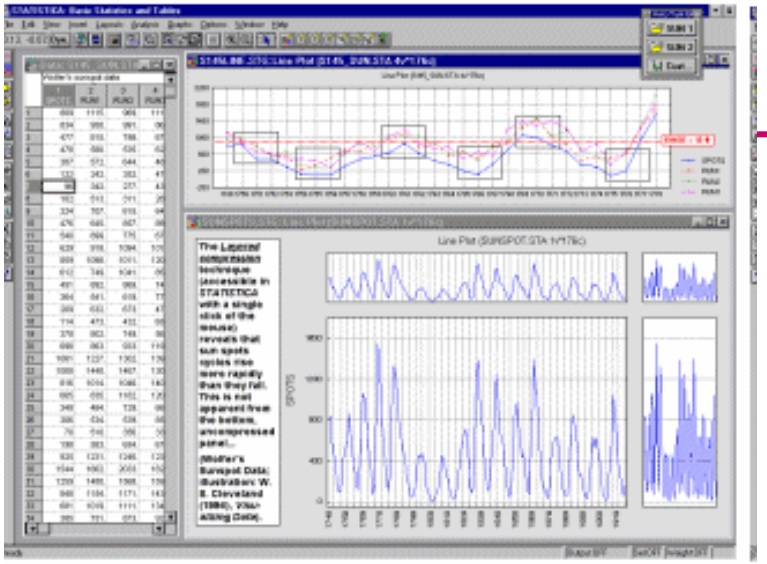
Jerzy Stefanowski
Instytut Informatyki

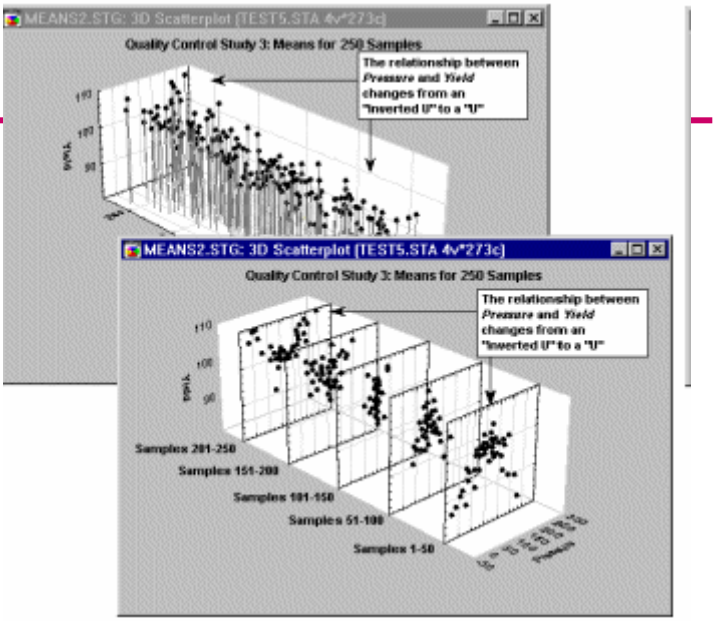
Data mining for SE -- 2013

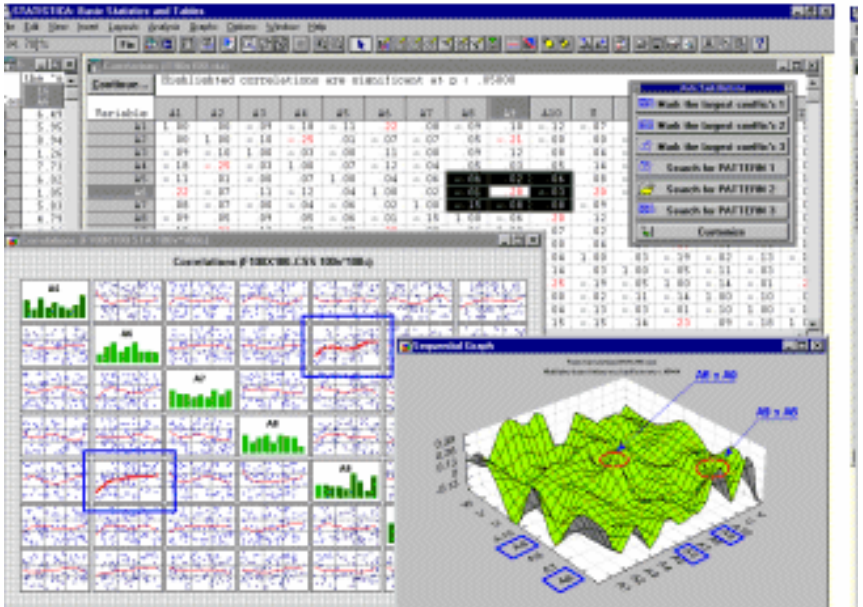
Ack.

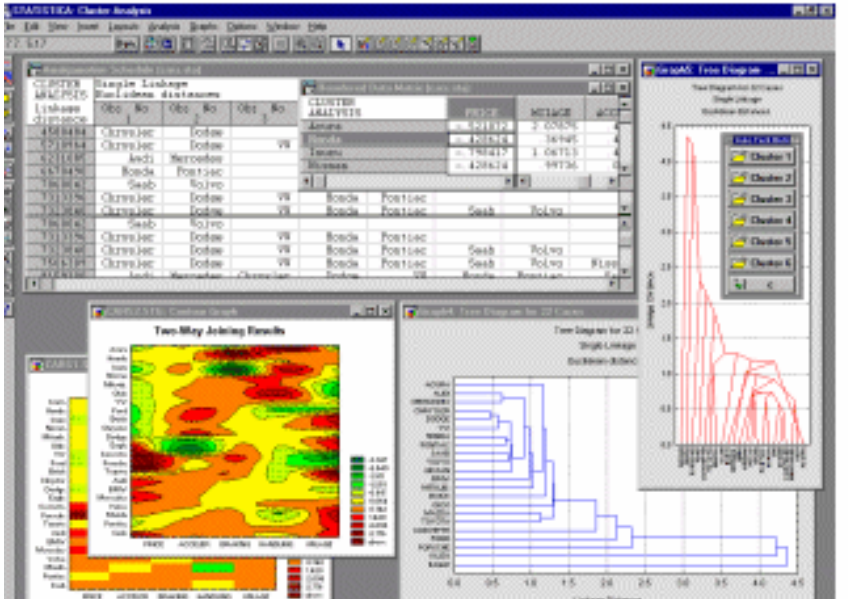
Inspirations are coming from:

- G.Piatetsky Schapiro lectures on KDD
- J.Han on Data Mining
- Ken Brodlie “Envisioning Information”
- Chris North “Information Visualisation”




MEANS2.STG: 3D Scatterplot (TEST5.STA 4v*273c)
Quality Control Study 3: Means for 250 Samples
The relationship between Pressure and Yield changes from an "Inverted U" to a "U"
Yield
Samples 201-250
Samples 151-200
Samples 101-150
Samples 51-100
Samples 1-50
Pressure




Two-Way Joining Results

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

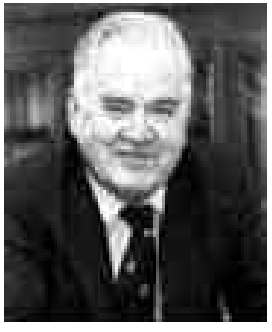
A graph is best when:

- 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

- Graphs and tables serve different purposes. Choose the appropriate data display to fit your purpose.

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



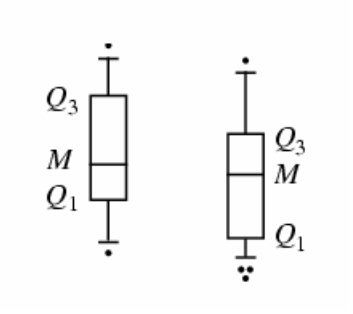
- Two stage process:
  - Exploratory: Search for evidence using all tools available
  - Confirmatory: evaluate strength of evidence using classical data analysis

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:

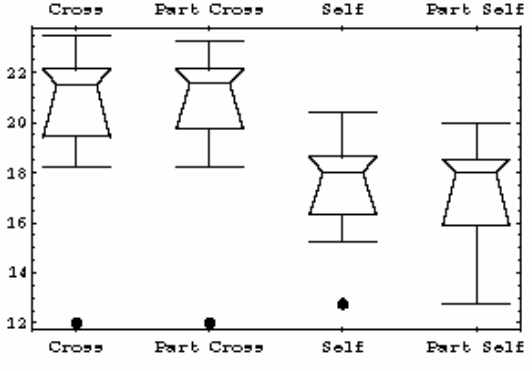
http://mathworld.wolfram.com/Box-and-WhiskerPlot.html

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



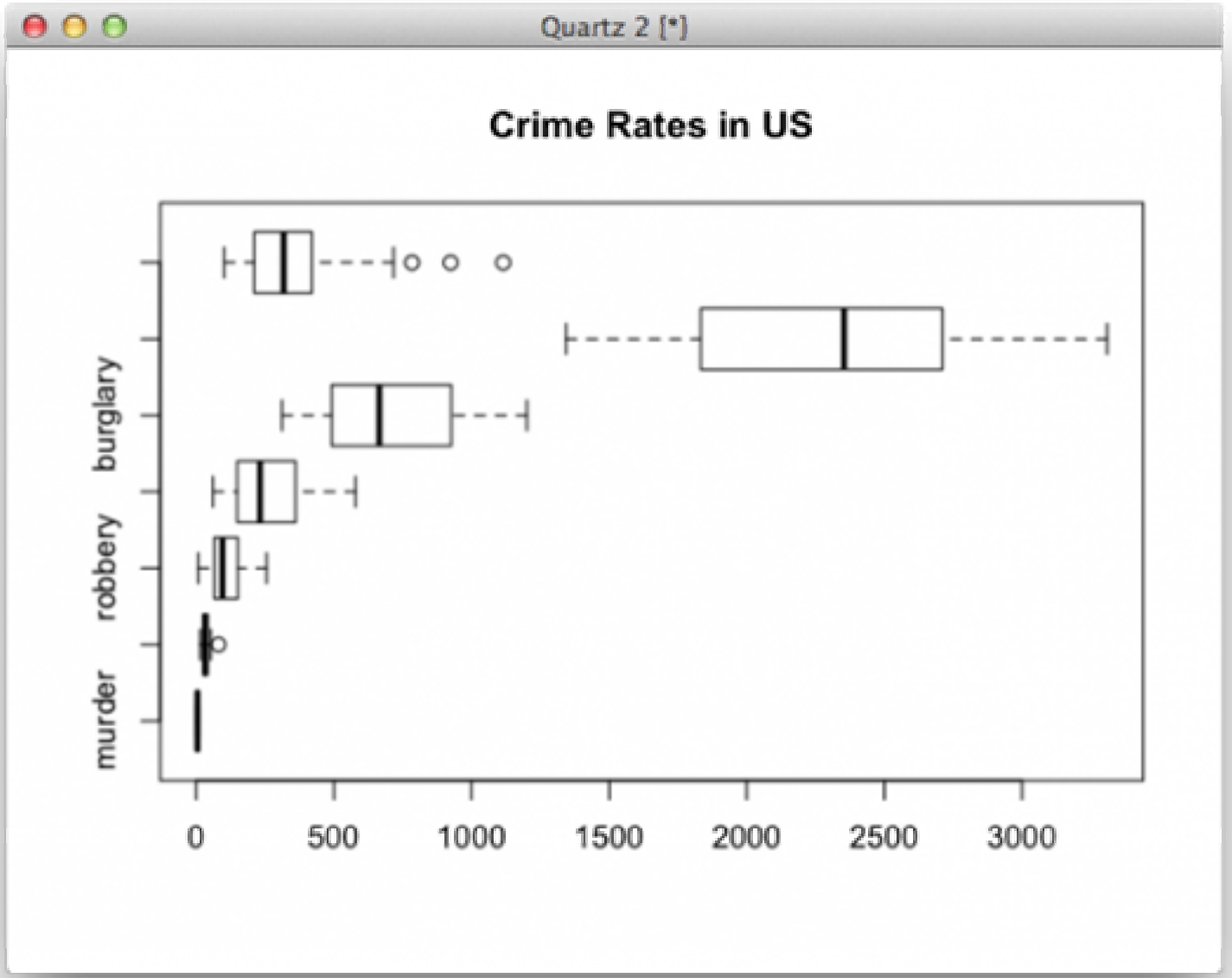
M – median
Q1, Q3 – quarrtiles
Whiskers –
1.5 * interquartile range
Dots - outliers

Darwin's plant study



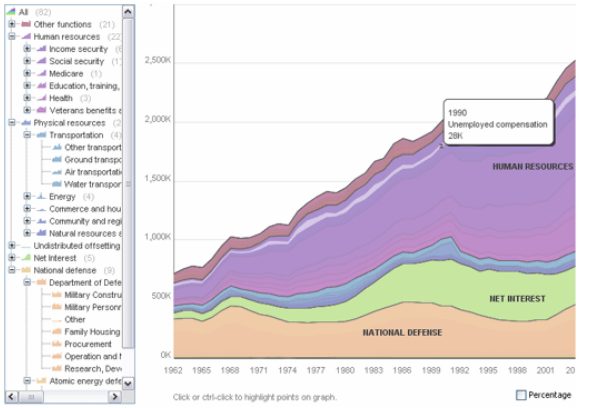
http://www.upscale.utoronto.ca/GeneralInterest/Harrison/Visualisation/Visualisation.html

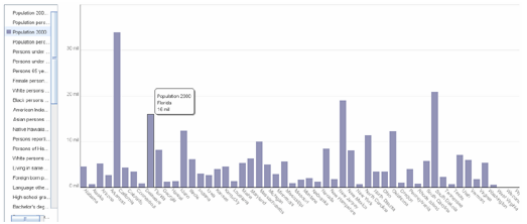
Distribution visualisation – US Crime Story

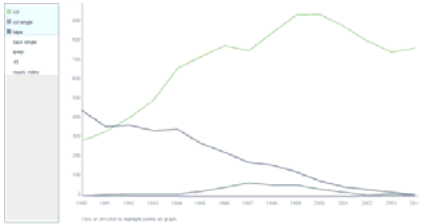


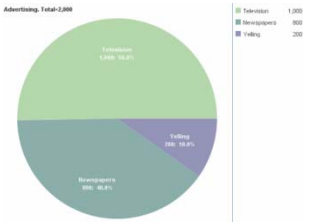
Common Display Types

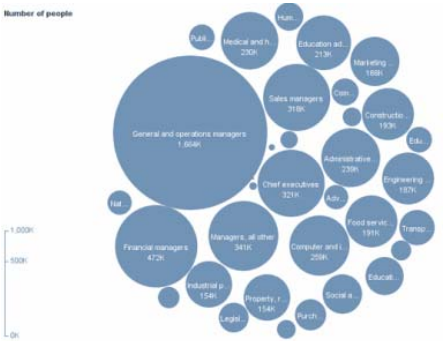
- Bar Charts
- Line Charts
- Pie Charts
- Bubble Charts
- Stacked Charts
- Scatterplots

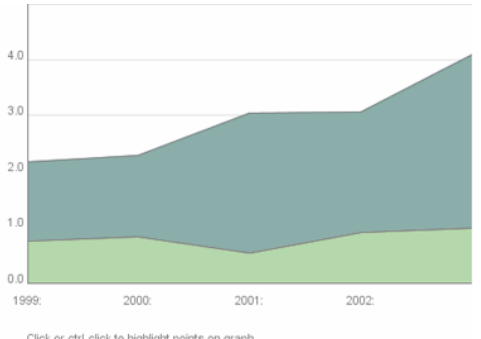


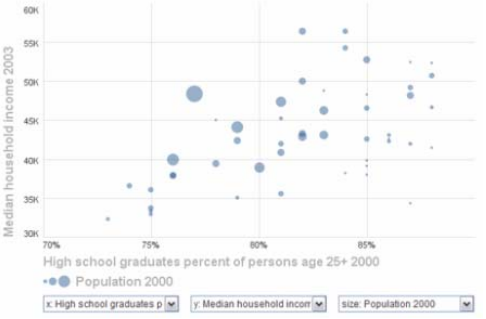




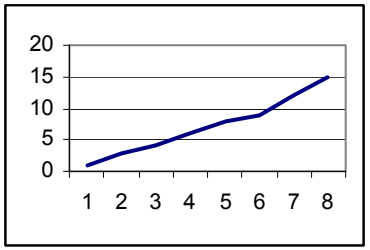






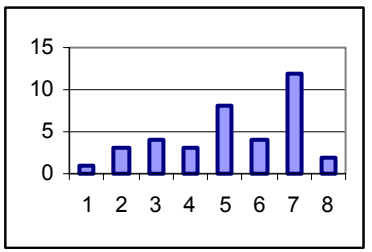


When to use which type?



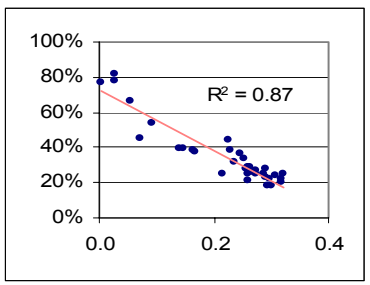
Line Graph

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



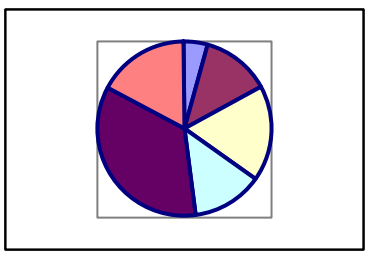
Bar Graph

- Comparison of relative point values



Scatter Plot

- Convey overall impression of relationship between two variables

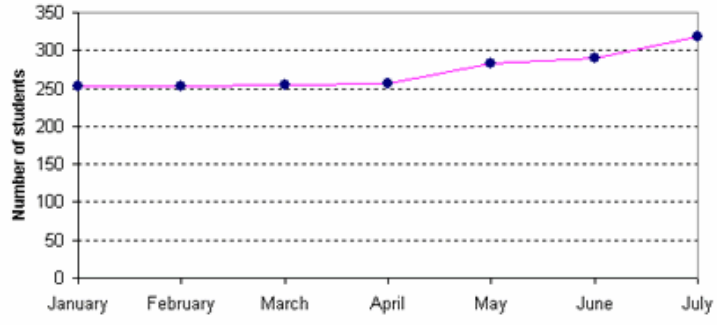


Pie Chart

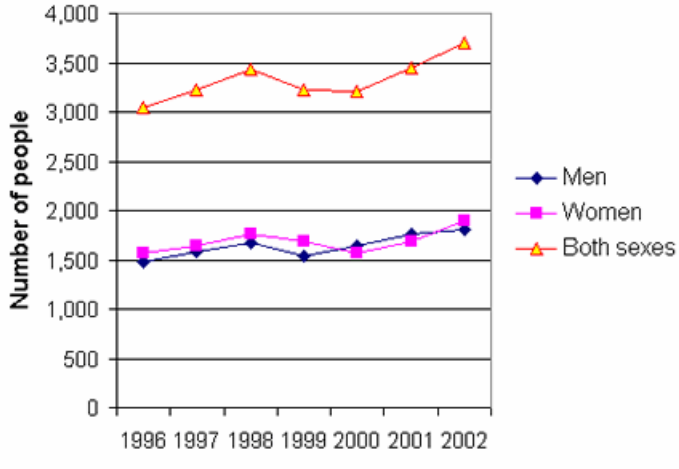
- Emphasizing differences in proportion among a few numbers

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)



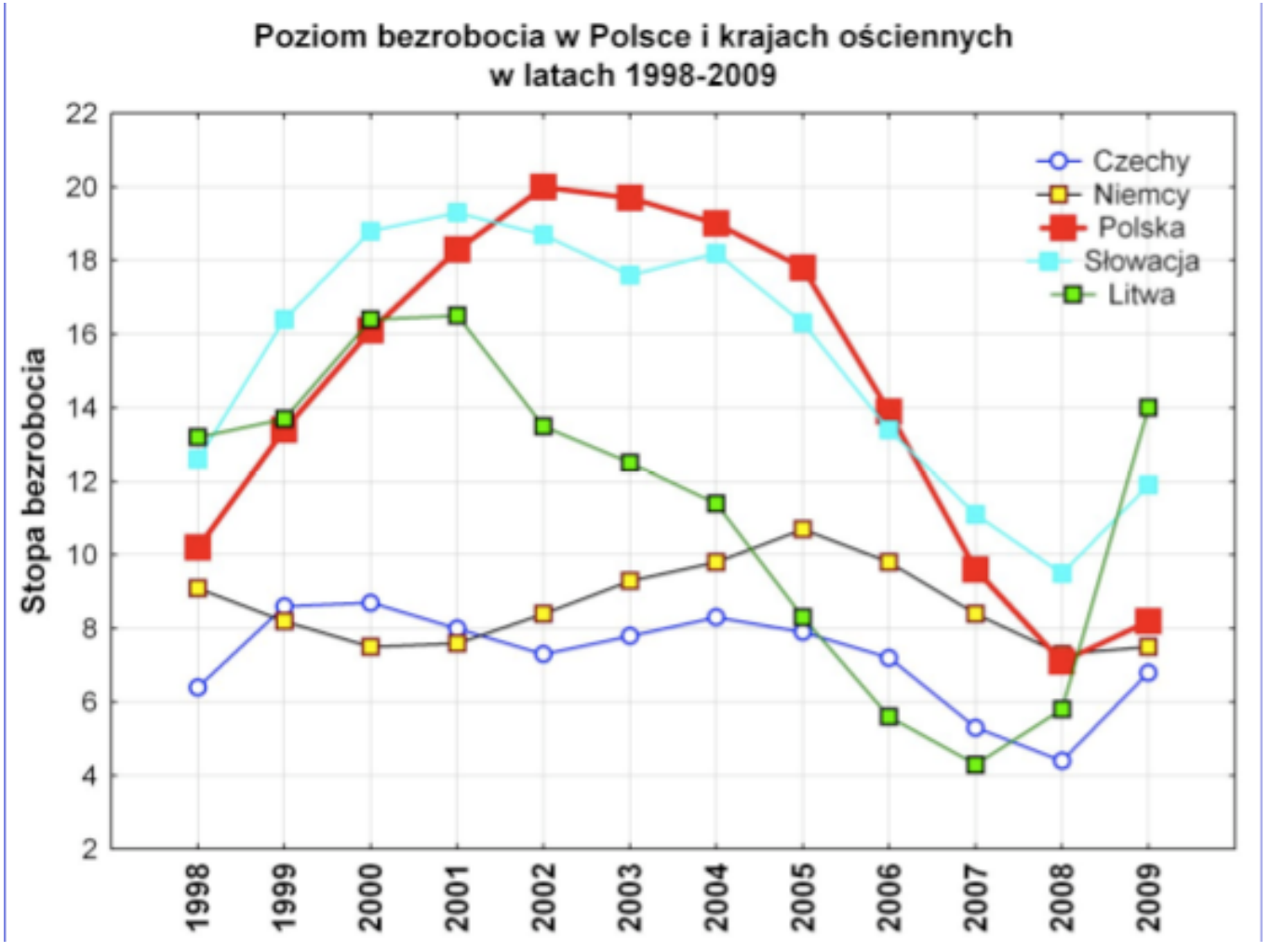
Students participating in sporting activities



Mobile Phone use

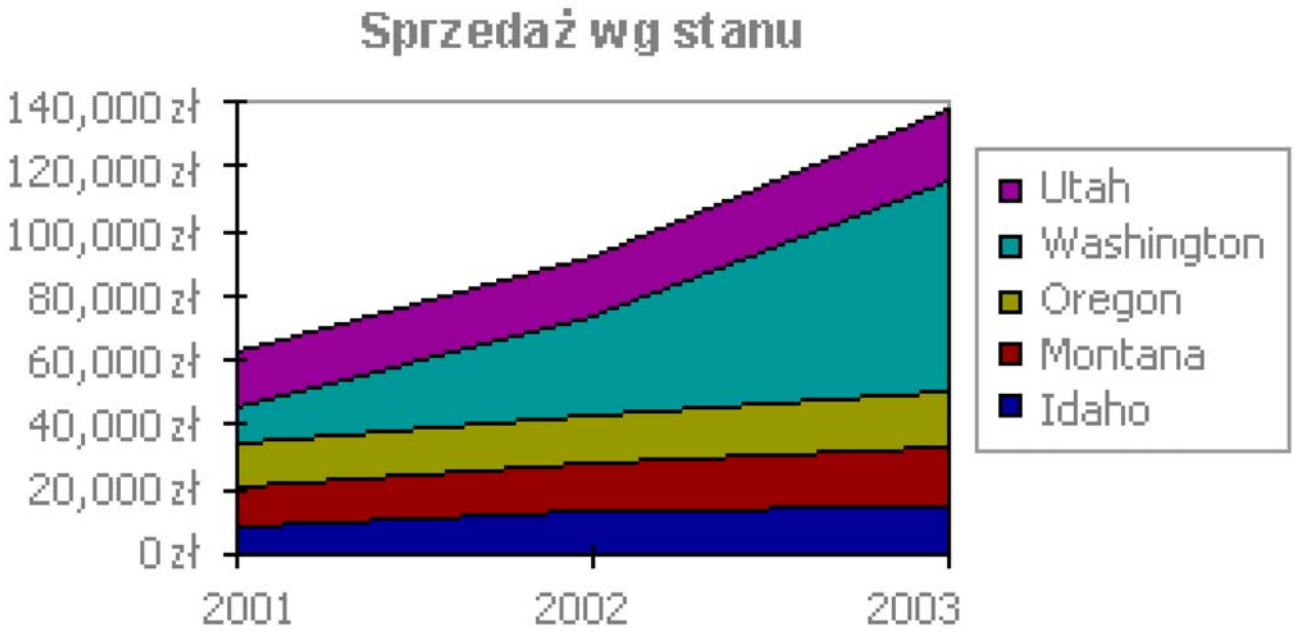
Note: graph labelling is fundamental

Time line graph – show dynamics of measurements

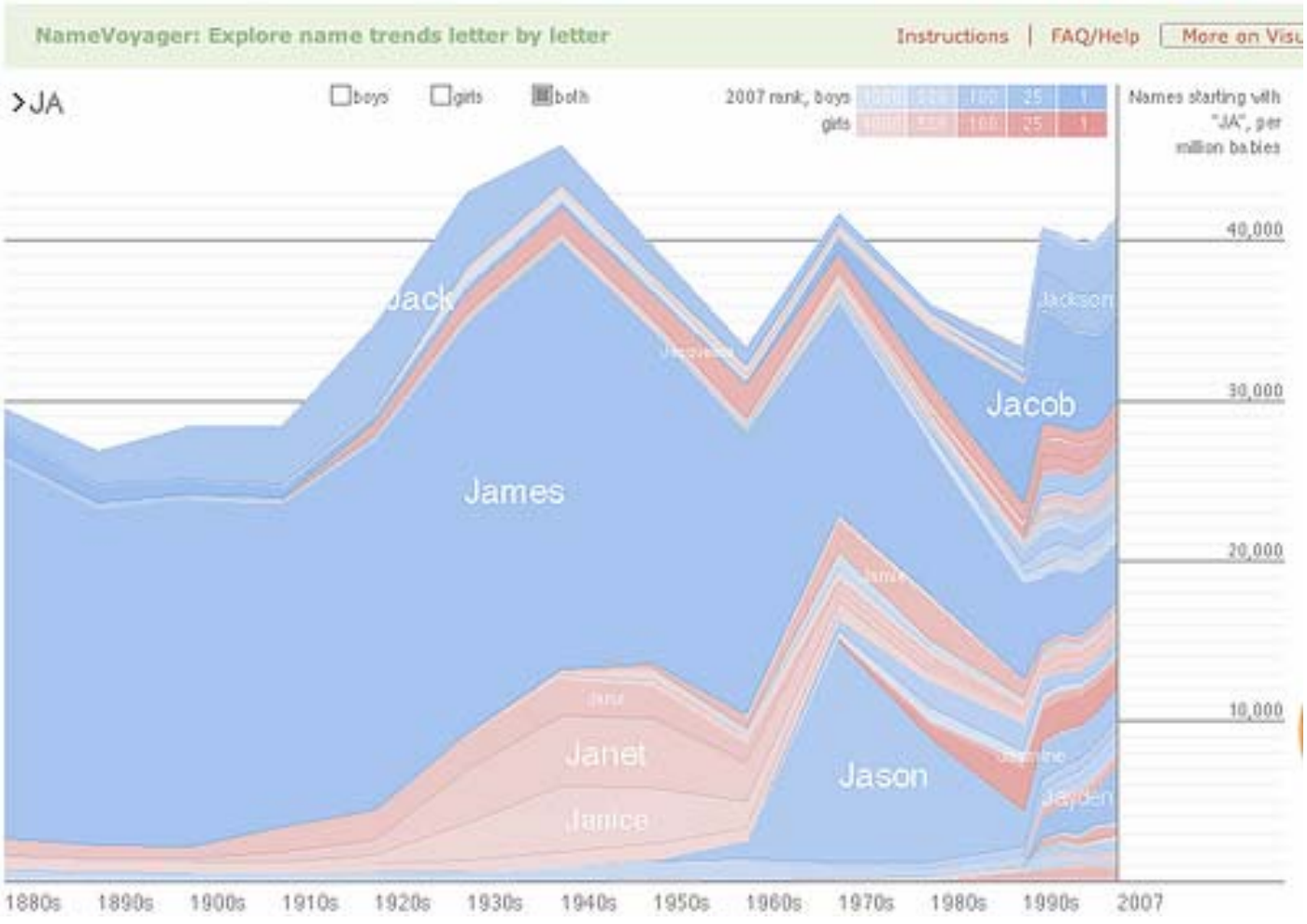


Stratified graphs

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



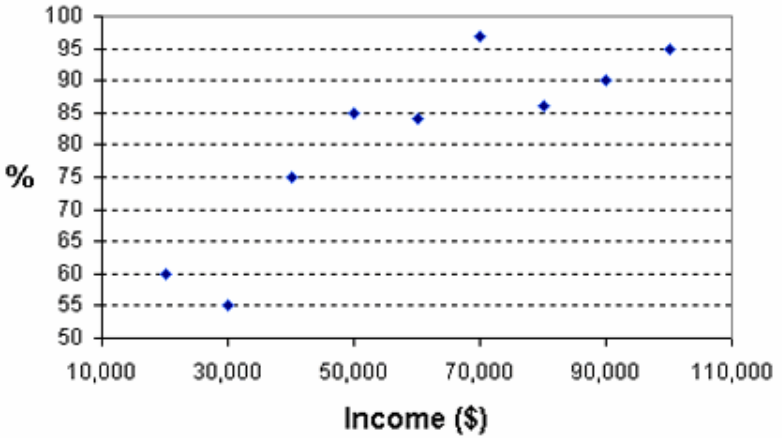
Demo – Baby Names Voyager



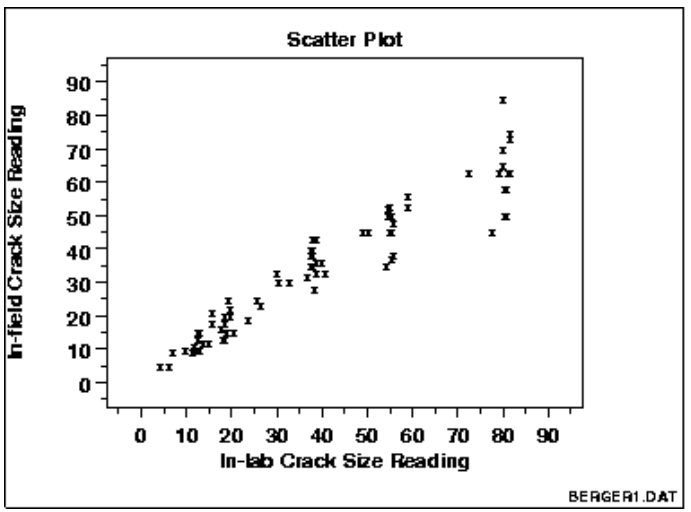
http://www.babynamewizard.com/voyager

Scatter Plot – Wykresy rozrzutu XY

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

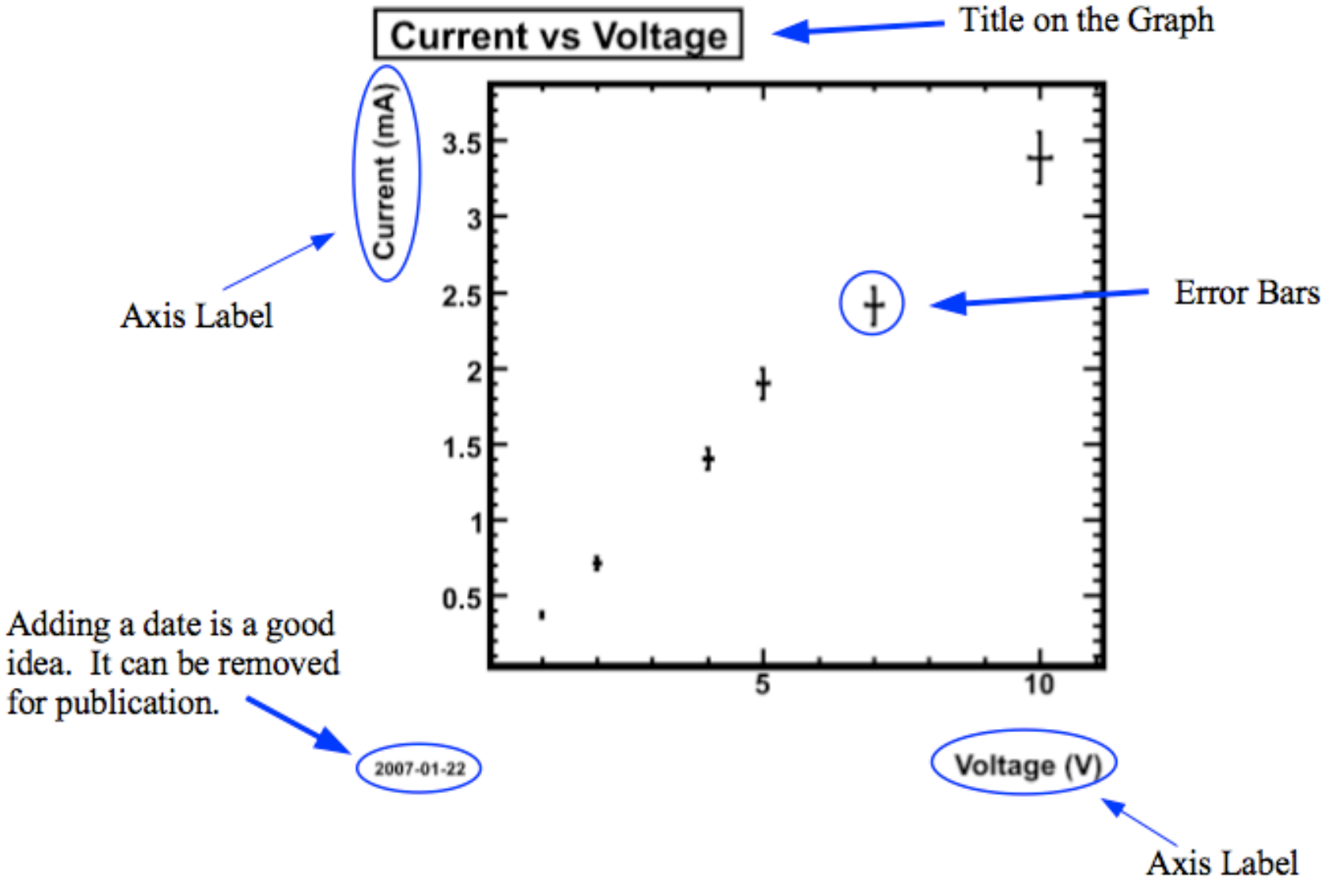


Car ownership by household income



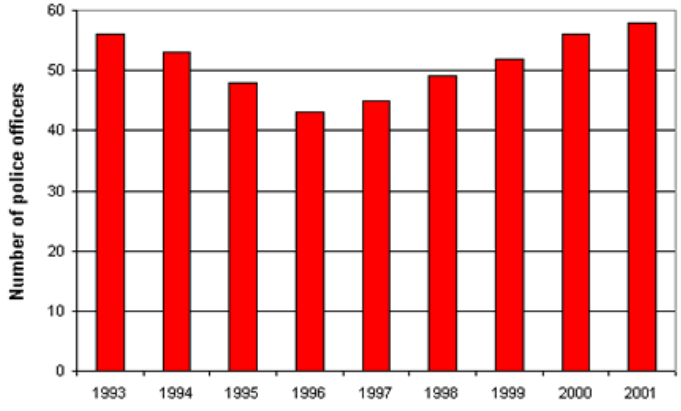
Example taken from NIST Handbook – Evidence of strong positive correlation

Elements of a Good Graph

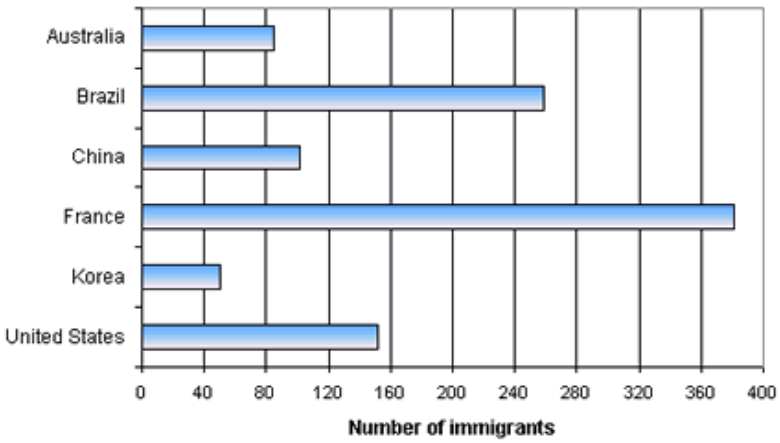


Simple Representations – Bar Graph

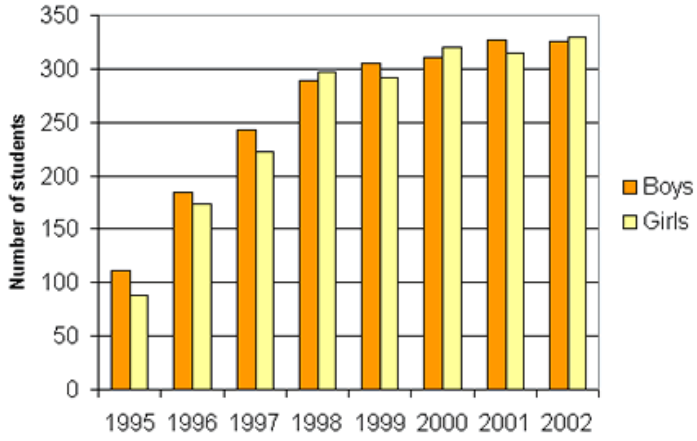
- Bar graph
  - Presents categorical variables
  - Height of bar indicates value
  - Double bar graph allows comparison
  - Note spacing between bars
  - Can be horizontal (when would you use this?)



Number of police officers



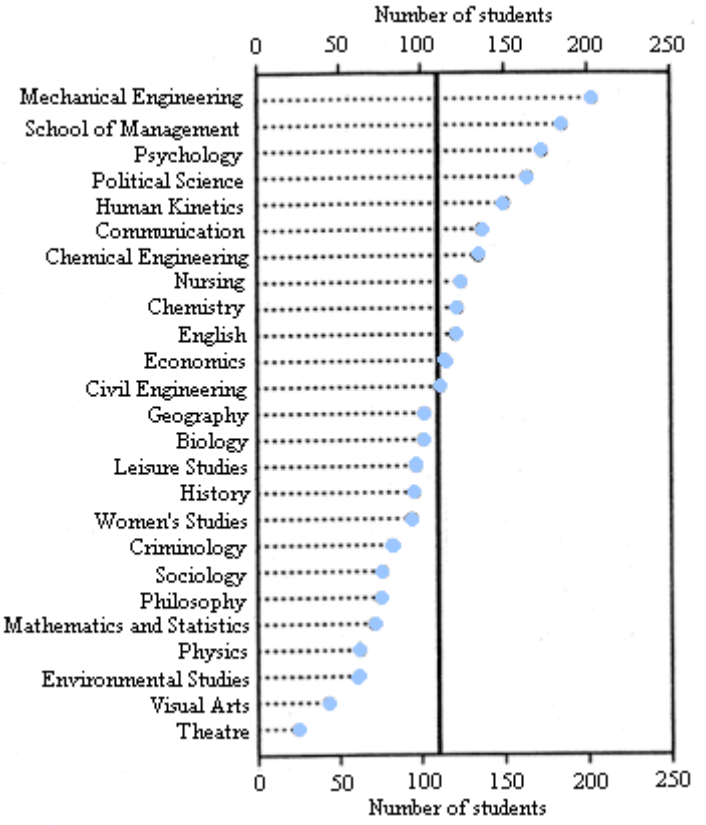
Note more space for labels



Internet use at a school

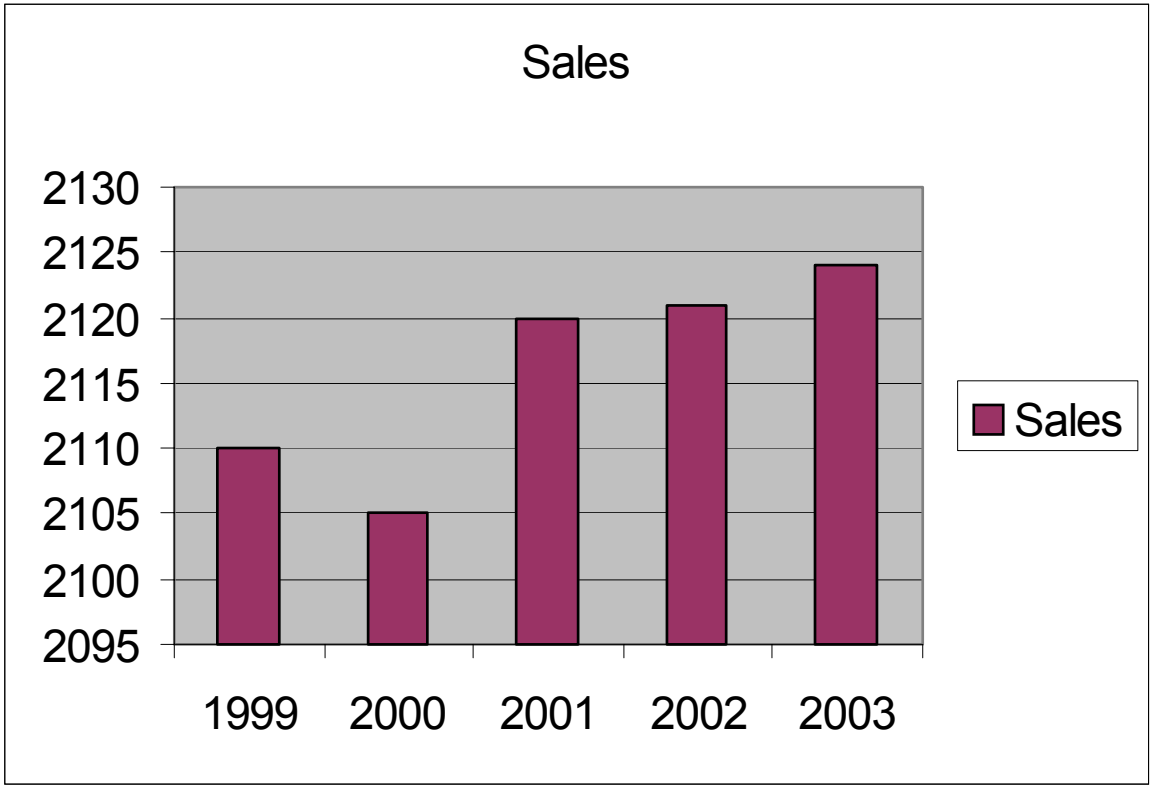
Dot Graph

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



Bad Visualization: Spreadsheet

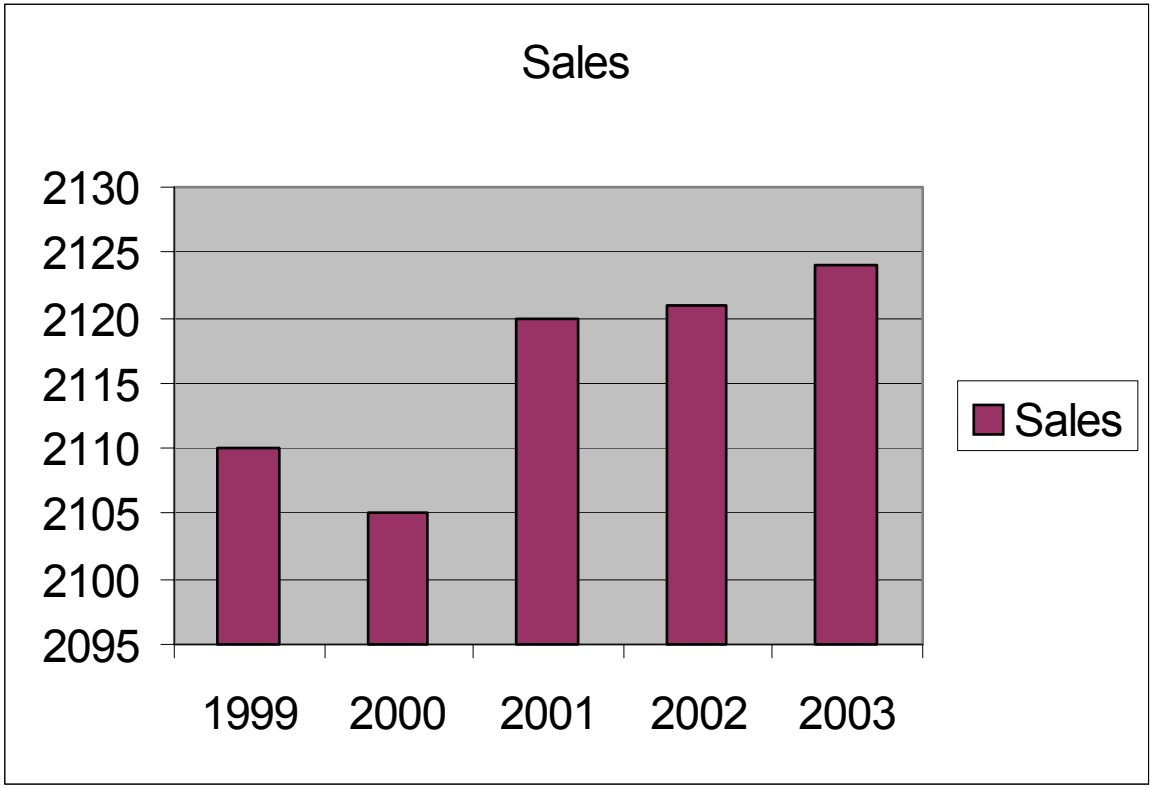
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



What is wrong with this graph?

Bad Visualization: Spreadsheet with misleading Y –axis

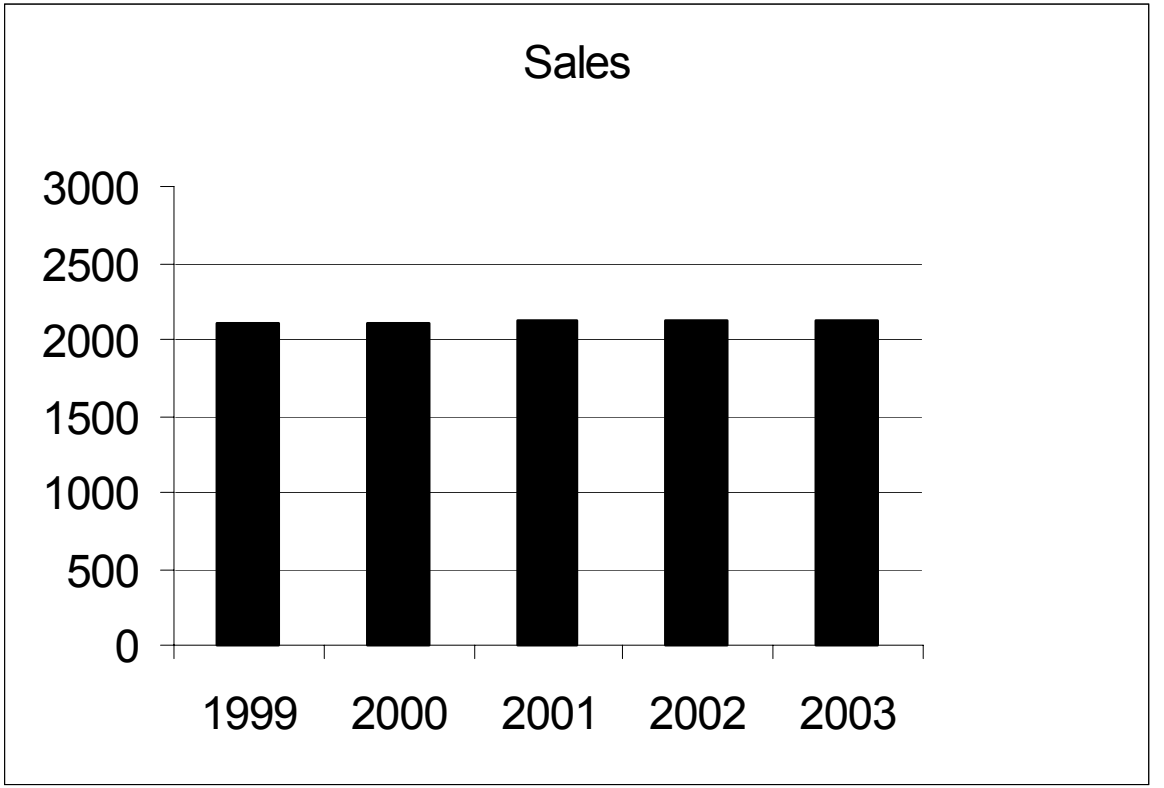
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



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

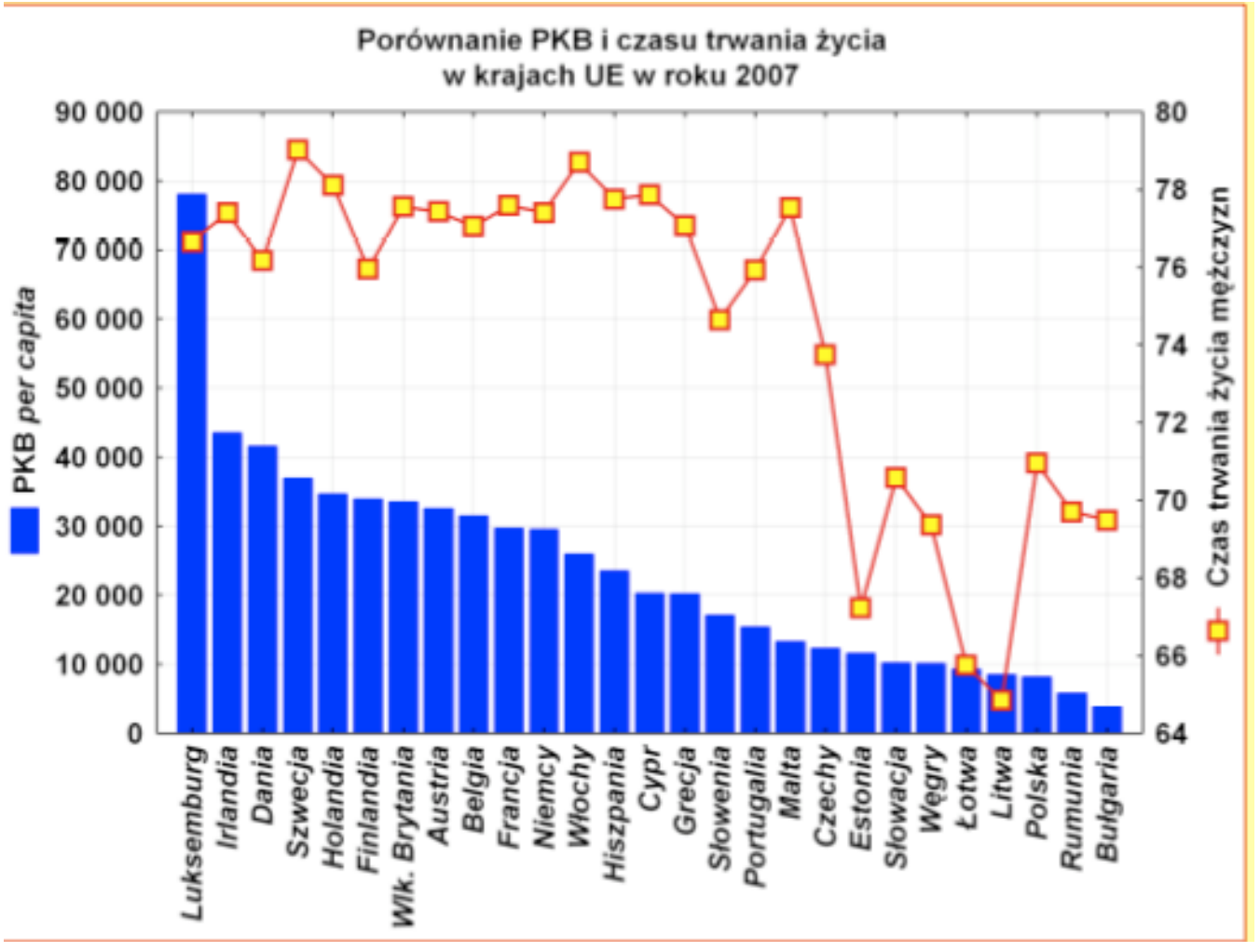
Better Visualization

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

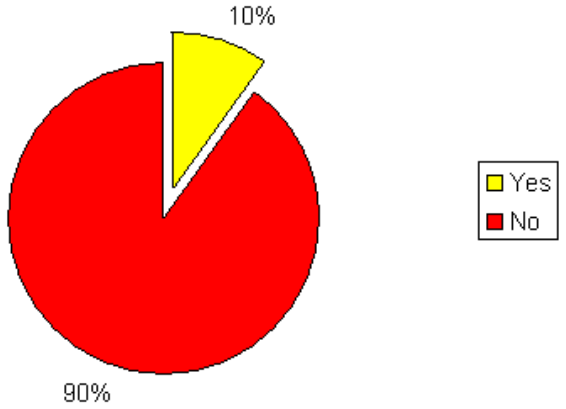
Integrating various graphs



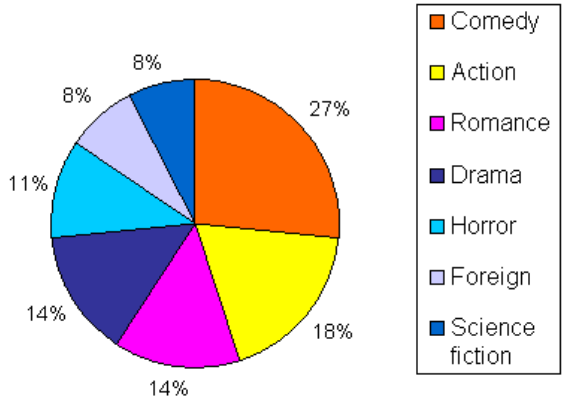
Pie Chart

- 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?



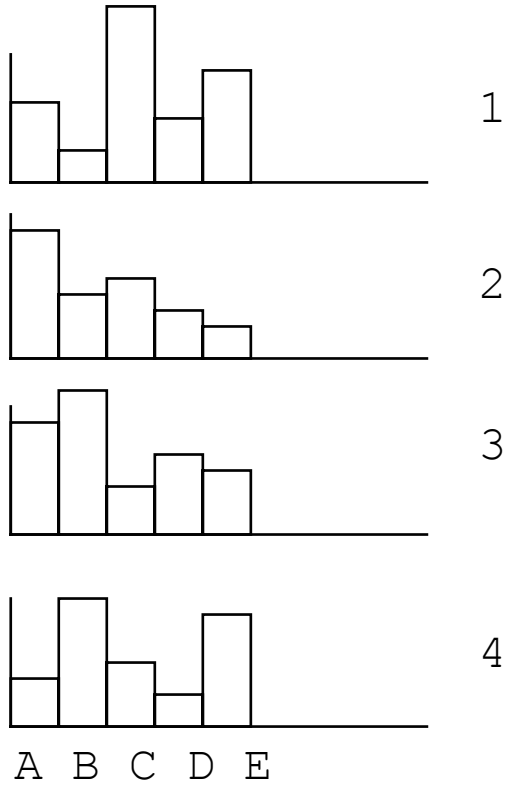
Favourite movie genres

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

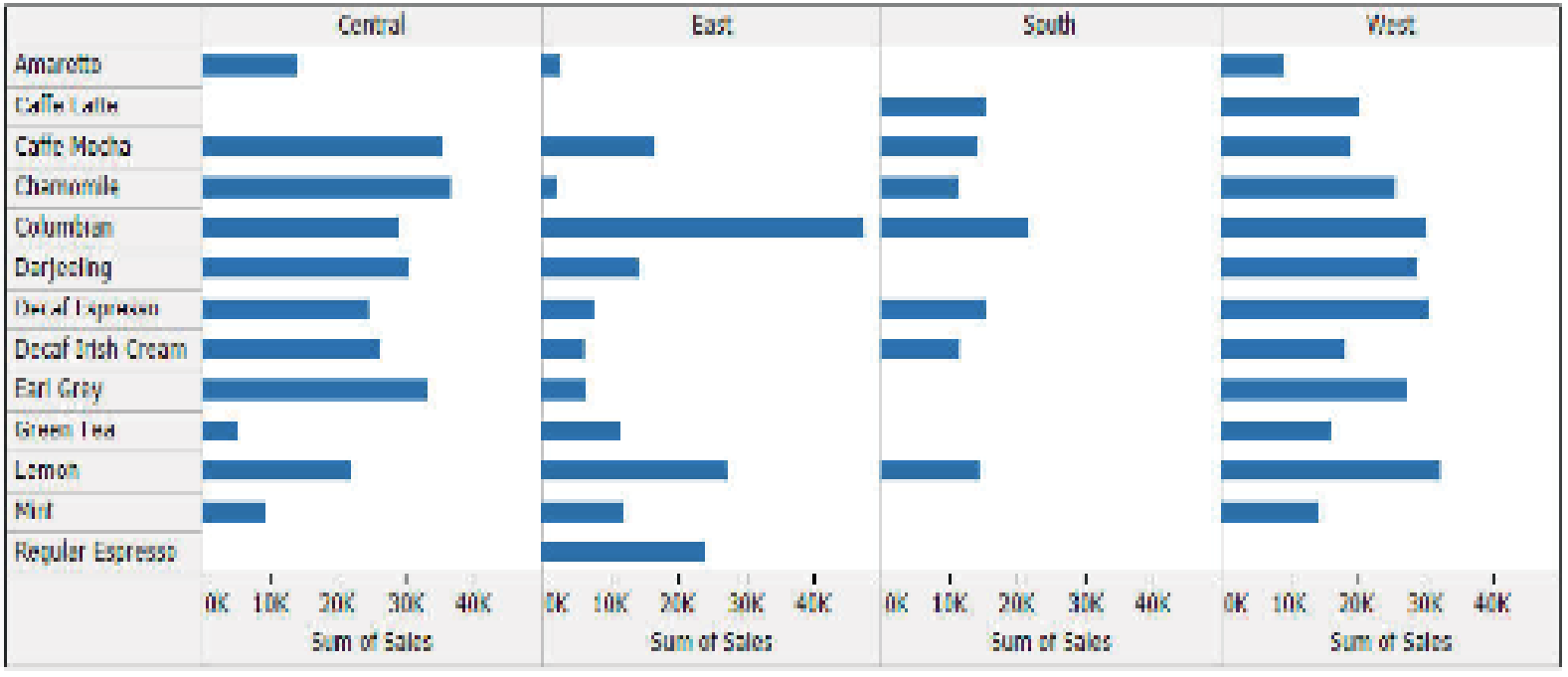
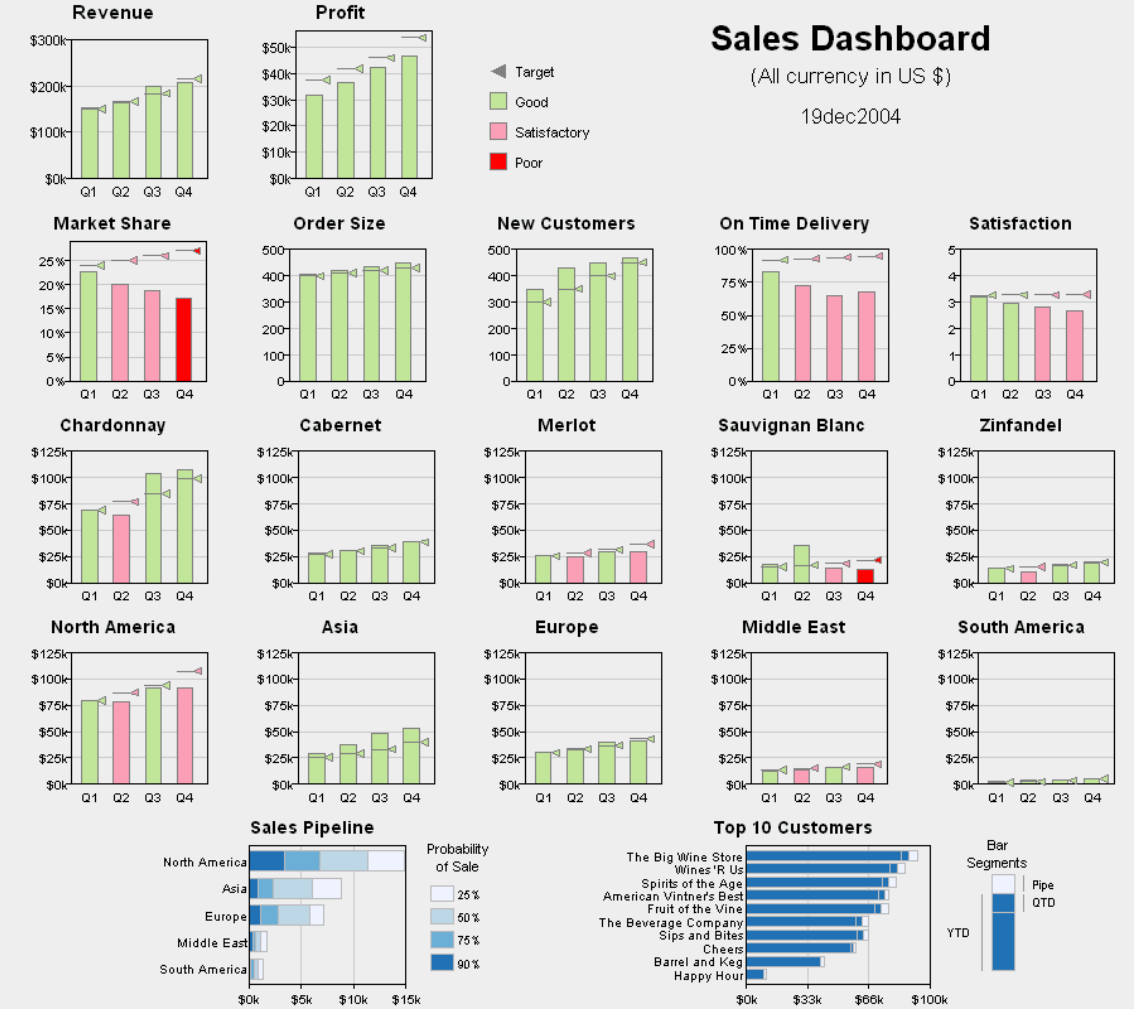


Figure 1: A typical Tableau display of product sale values by region, arranged alphabetically.

Buisness Analytics Tools – Manager Dashboards



Scatterplot Matrix

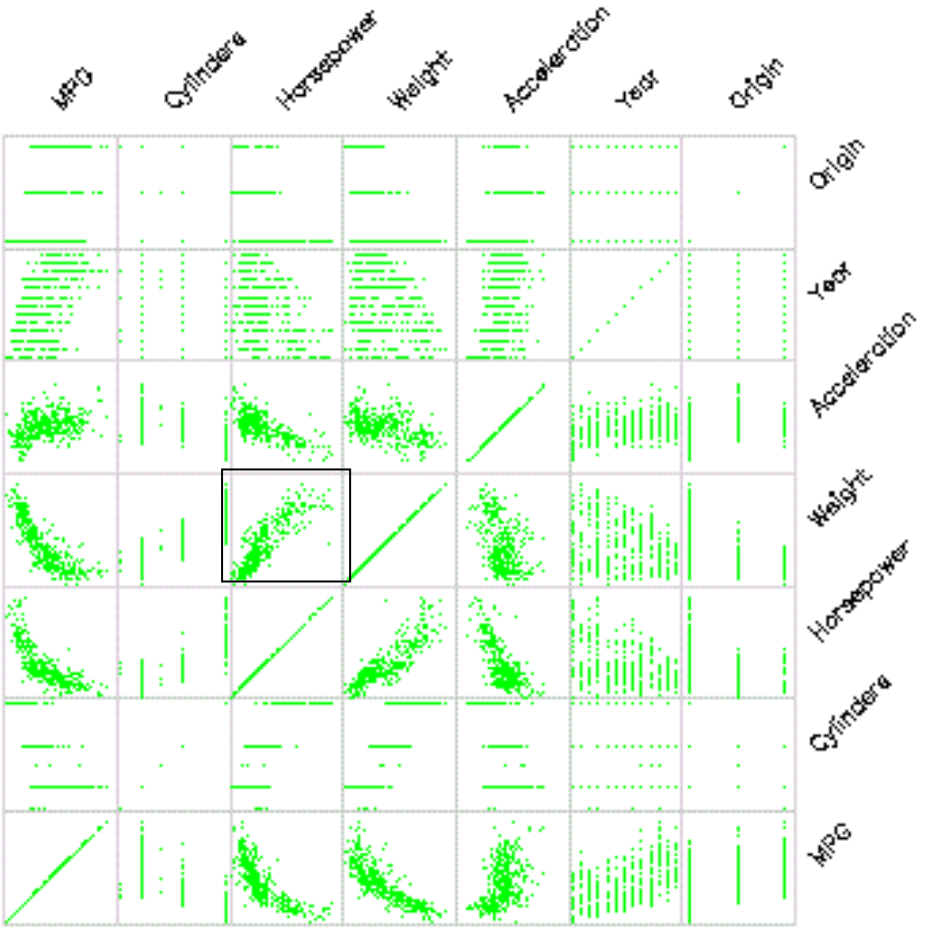
Represent each possible pair of variables in their own 2-D scatterplot (car data)

Q: Useful for what?

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

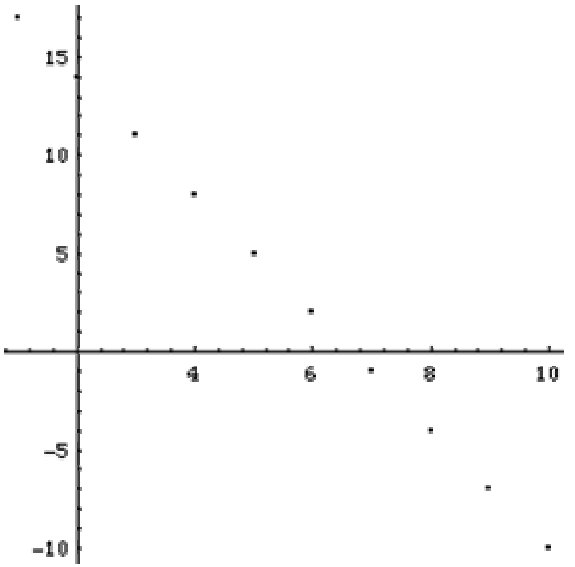
Q: Misses what?

A: multivariate effects

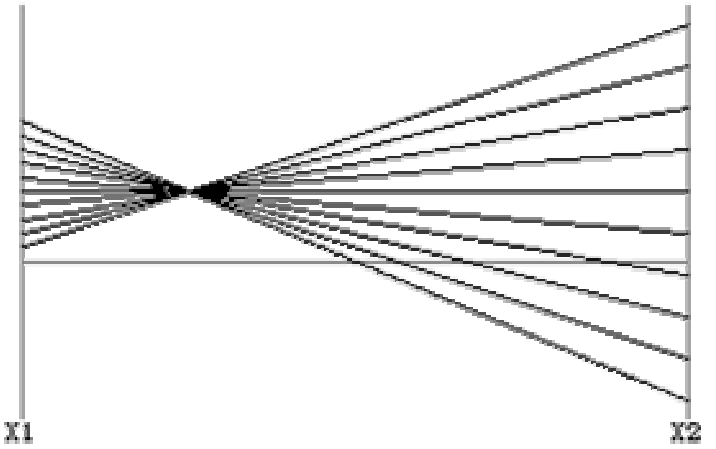


Parallel Coordinates

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

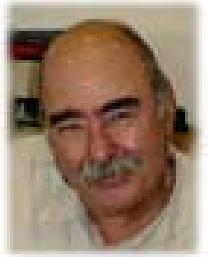


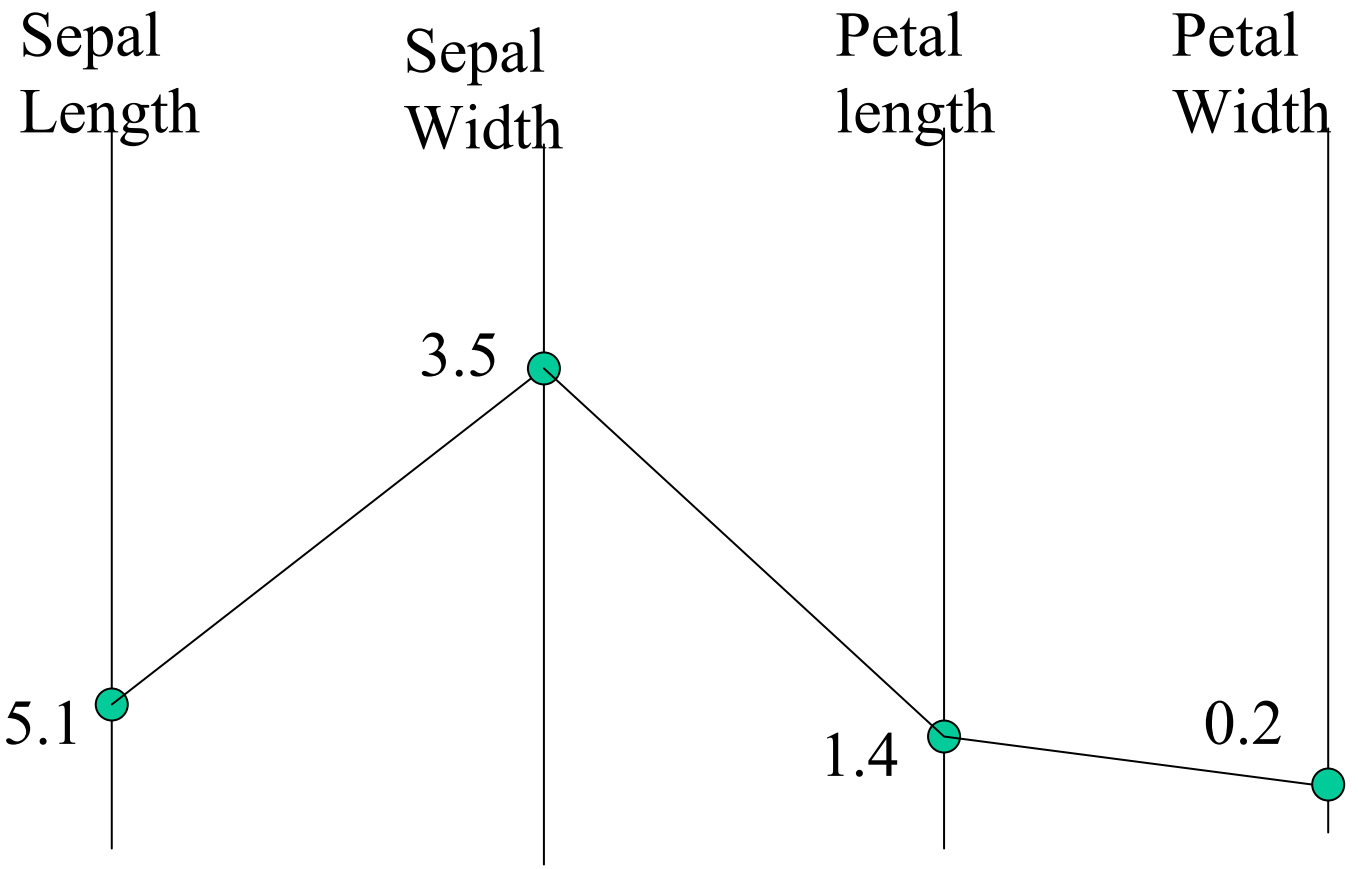
Dataset in a Cartesian coordinates



Same dataset in parallel coordinates

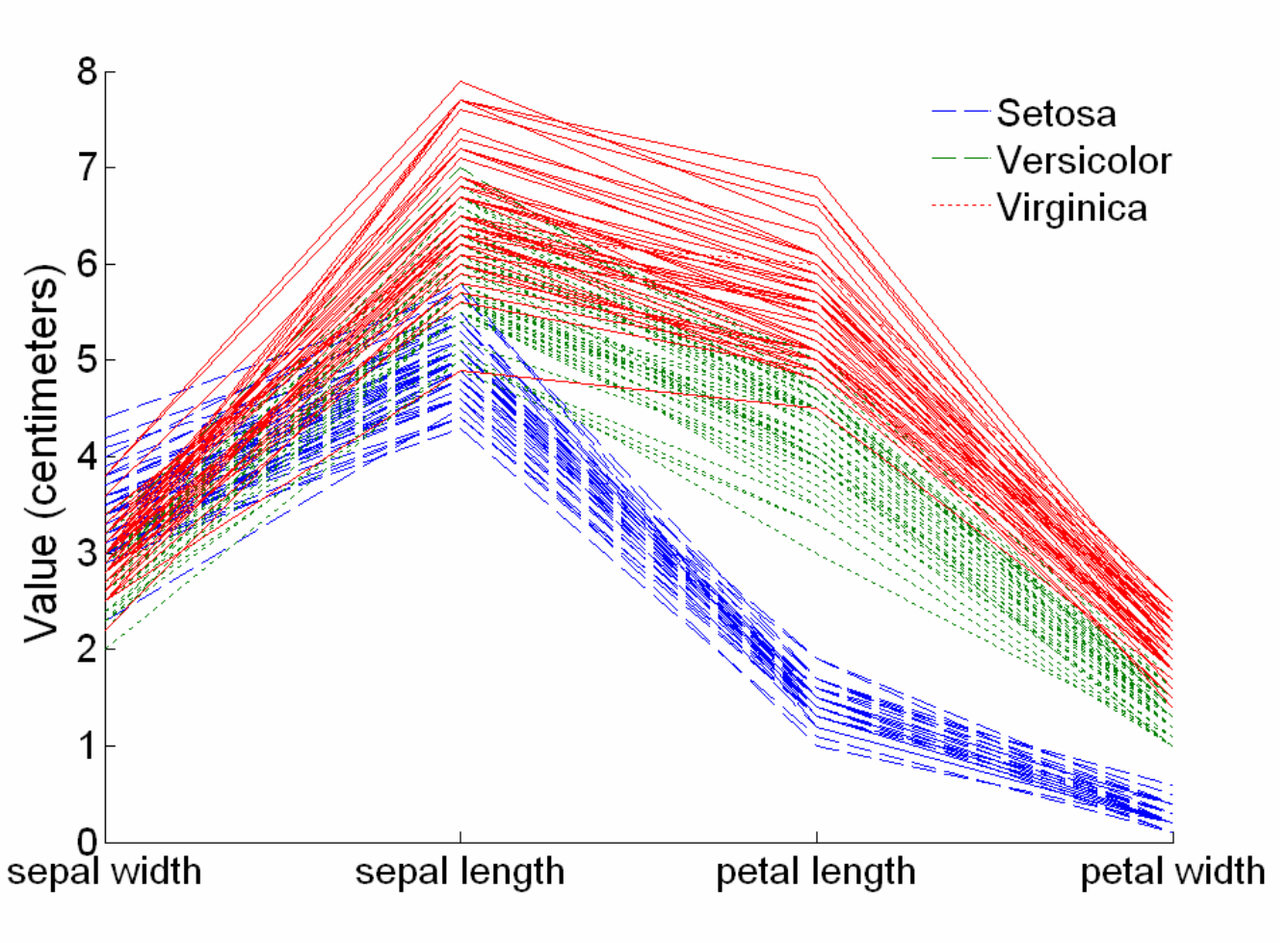
Invented by Alfred Inselberg while at IBM, 1985





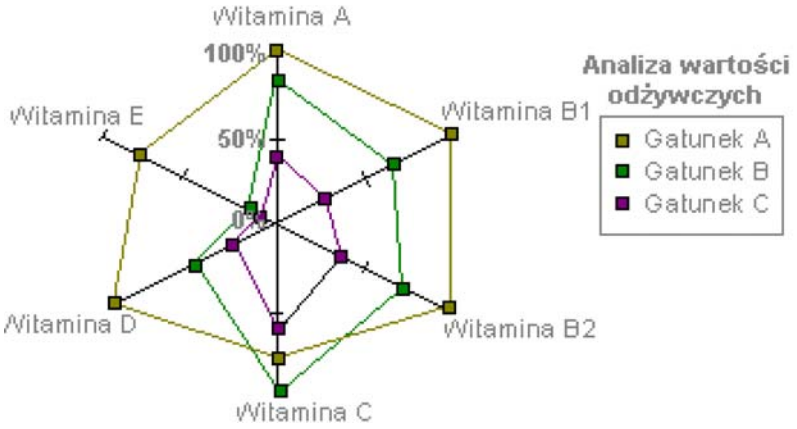
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates Plots for Iris Data



Radar Figures

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



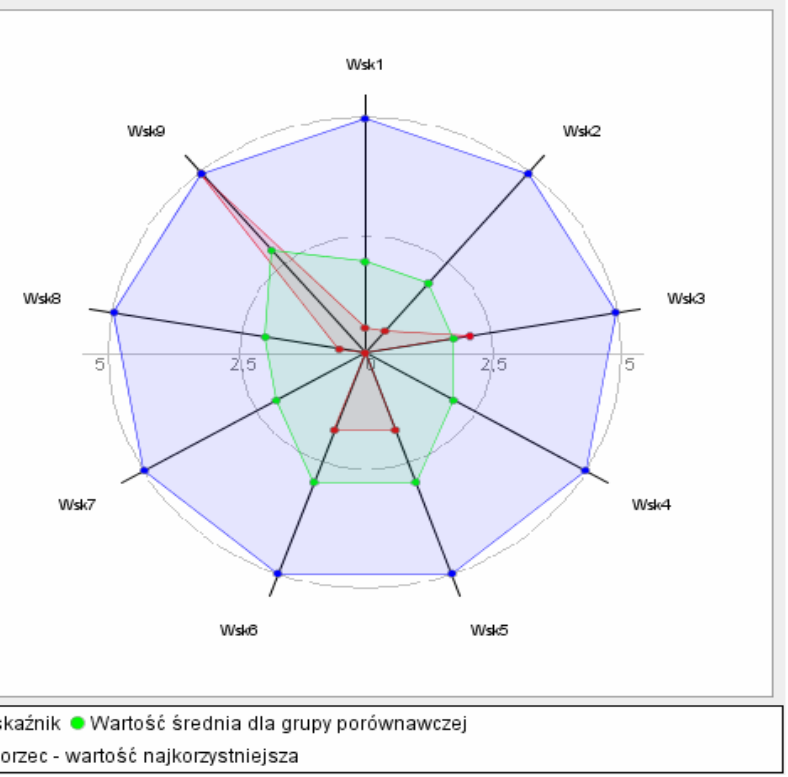
ównoważony rozwój - Podkowa Leśna

om zobrazowania: Aktywność na rynku pracy

a zobrazowania: 2007;

a wartości wskaźników w przedziale <0;5>

Wybrana dziedzina



skaźnik ● Wartość średnia dla grupy porównawczej

orzec - wartość najkorzystniejsza

Wykres radarowy – oceny wskaźników w ramach dziedziny

I poziom oceny

ENDA:

1- 1.1.9.1 Wskaźnik zatrudnienia kobiet (udział pracujących kobiet w liczbie kobiet ogółem)

2- 1.1.9.2 Udział pracujących w liczbie mieszkańców w wieku produkcyjnym (%)

3- 1.1.9.3 Liczba podmiotów gospodarczych na 1000 mieszkańców

4- 1.1.9.4 Ogólny wskaźnik aktywności zawodowej kobiet - udział czynnych zawodowo w liczbie kobiet w wieku 15 lat i więcej [%]

5- 1.1.9.5 Udział zakładów osób fizycznych w liczbie jednostek ogółem sektora prywatnego [%]

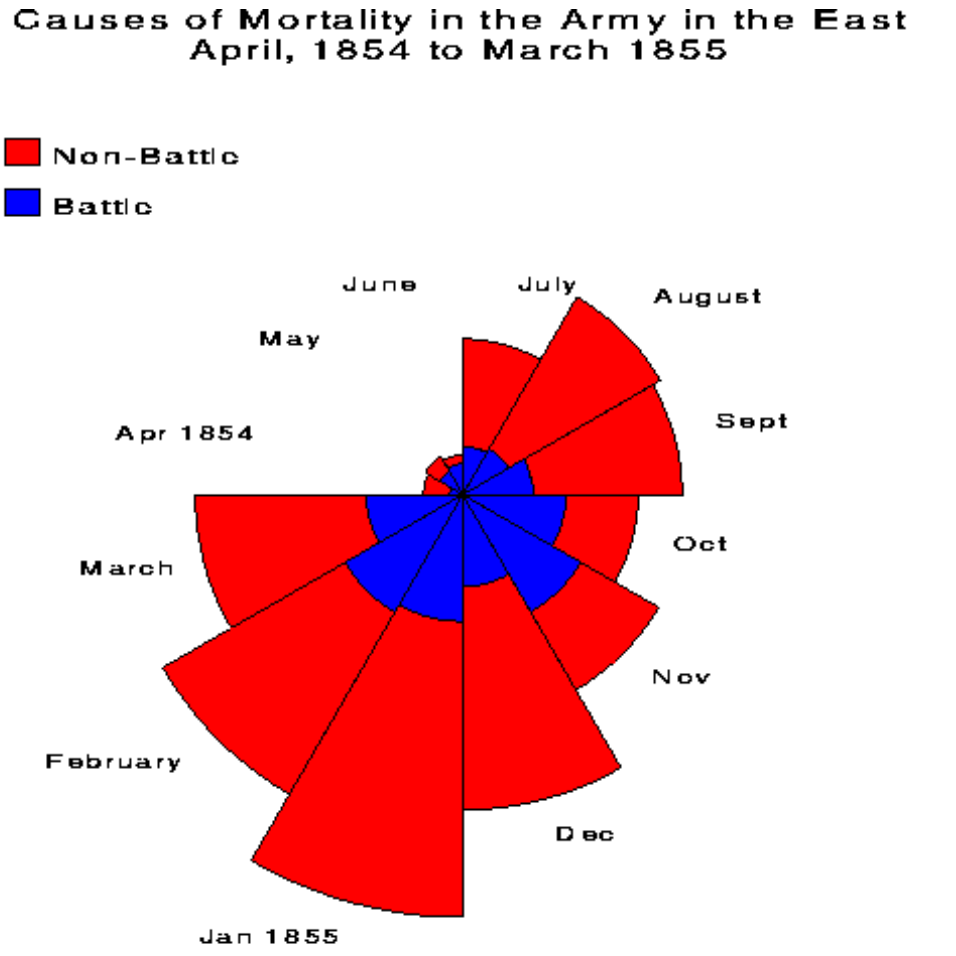
6- 1.1.9.6 Udział zakładów osób fizycznych w liczbie jednostek ogółem [%]

7- 1.1.9.7 Ogólny wskaźnik aktywności zawodowej kobiet - udział czynnych zawodowo w liczbie kobiet w wieku 15 lat i więcej [%]

8- 1.1.9.8 Stopa zatrudnienia - liczba pracujących na 1000 mieszkańców

9- 1.1.9.9 Bezrobotni na 100 osób w wieku produkcyjnym

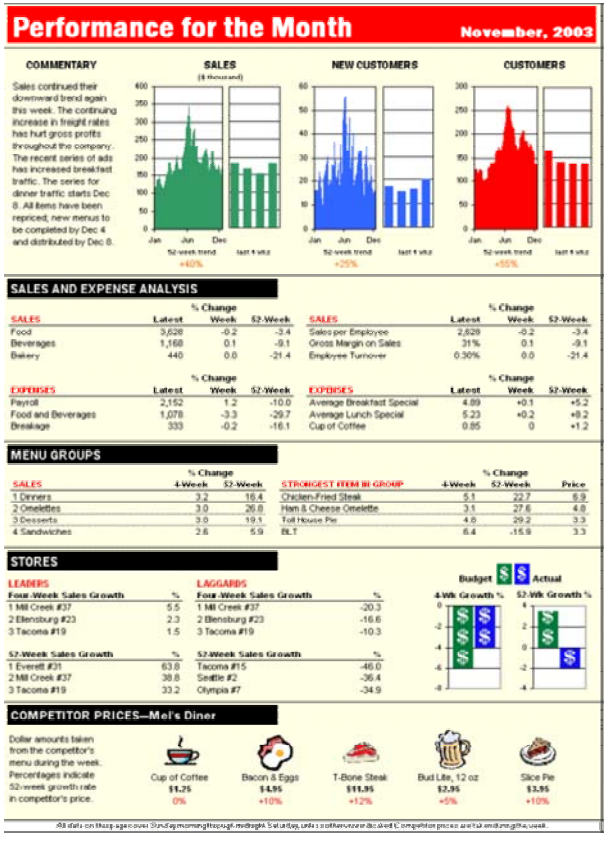
F. Nightingale (1856) – abstract representation



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



Performance for the Month — November, 2003

COMMENTARY

Sales continued their downward trend again this week. The continuing increase in freight rates has hurt gross profits throughout the company. The recent series of ads has increased breakfast traffic. The series for dinner traffic starts Dec 8. All items have been repriced, new menus to be completed by Dec 4 and distributed by Dec 8.

SALES ($ thousand)

NEW CUSTOMERS

CUSTOMERS

SALES AND EXPENSE ANALYSIS

SALES	Latest	% Change 4-Week	% Change 52-Week
Food	3,628	-0.2	-3.4
Beverages	1,168	0.1	-9.1
Bakery	440	0.0	-21.4

SALES	Latest	% Change 4-Week	% Change 52-Week
Sales per Employee	2,629	-0.2	-3.4
Gross Margin on Sales	31%	0.1	-9.1
Employee Turnover	0.30%	0.0	-21.4

EXPENSES	Latest	% Change 4-Week	% Change 52-Week
Payroll	2,152	1.2	-10.0
Food and Beverages	1,078	-3.3	-29.7
Breakage	333	-0.2	-16.1

EXPENSES	Latest	% Change 4-Week	% Change 52-Week
Average Breakfast Special	4.89	+0.1	+5.2
Average Lunch Special	5.23	+0.2	+8.2
Cup of Coffee	0.85	0	+1.2

MENU GROUPS

SALES	% Change 4-Week	% Change 52-Week
1 Dinners	3.2	16.4
2 Omelettes	3.0	26.0
3 Desserts	3.0	19.1
4 Sandwiches	2.6	5.9

STRONGEST ITEM IN GROUP	% Change 4-Week	% Change 52-Week	Price
Chicken-Fried Steak	5.1	22.7	6.9
Ham & Cheese Omelette	3.1	27.6	4.8
Toll House Pie	4.8	29.2	3.3
BLT	6.4	-15.9	3.3

STORES

LEADERS Four-Week Sales Growth	%
1 Mill Creek #37	5.5
2 Blensburg #23	2.3
3 Tacoma #19	1.5

LAGGARDS Four-Week Sales Growth	%
1 Mill Creek #37	-20.3
2 Blensburg #23	-16.6
3 Tacoma #19	-10.3

52-Week Sales Growth	%
1 Everett #31	63.8
2 Mill Creek #37	38.8
3 Tacoma #19	33.2

52-Week Sales Growth	%
Tacoma #15	-46.0
Seattle #2	-36.4
Olympia #7	-34.9

Budget $ Actual

COMPETITOR PRICES—Mel's Diner

Dollar amounts taken from the competitor's menu during the week. Percentages indicate 52-week growth rate in competitor's price.

Cup of Coffee	Bacon & Eggs	T-Bone Steak	Bud Lite, 12 oz	Slice Pie
$1.25	$4.95	$11.95	$2.95	$3.95
0%	+10%	+12%	+5%	+10%

All data on this page covers Sun/Sat morning through midnight Saturday, unless otherwise noted. Competitor prices are as of ending the week.

Other forms

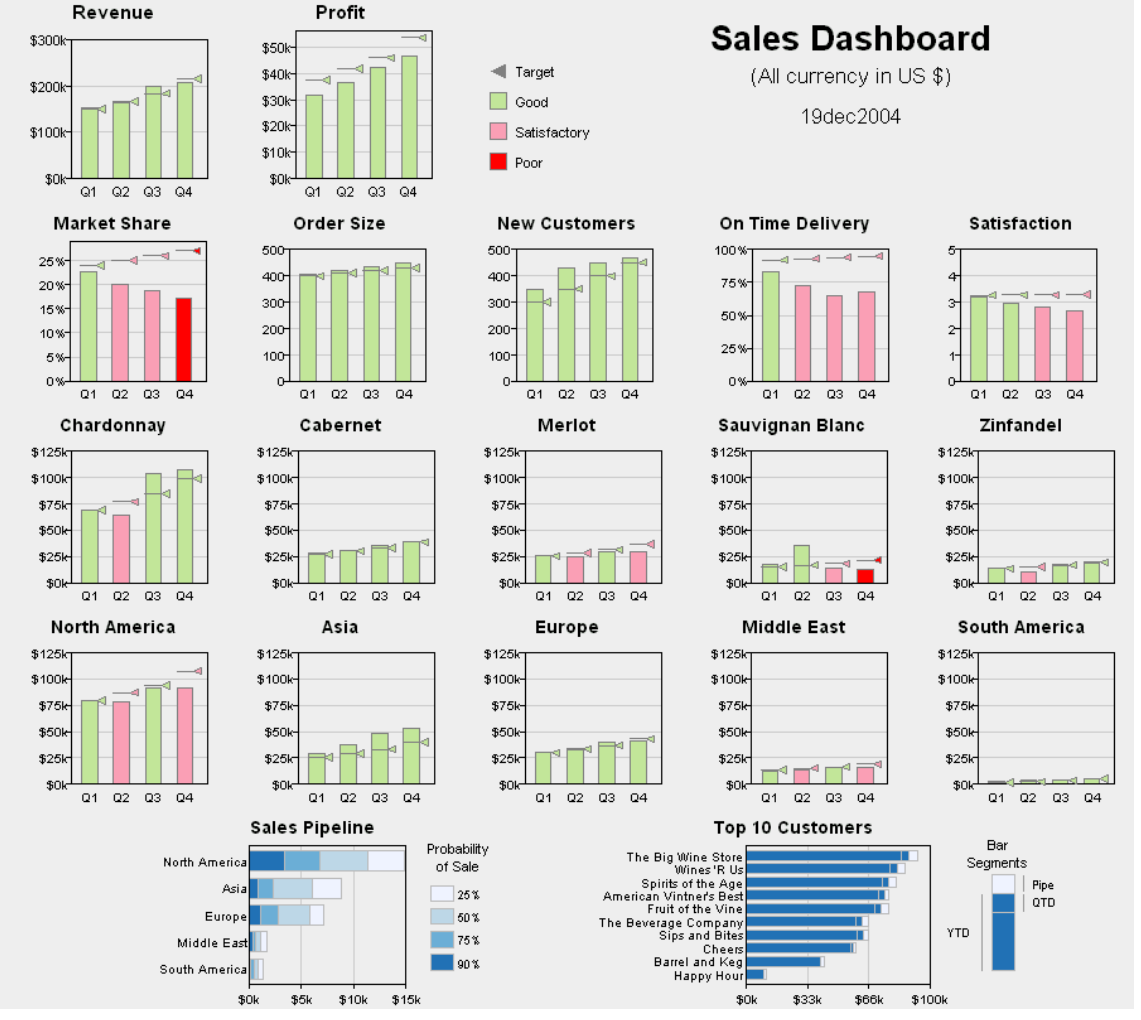
		Quarter to Date			Outlook			Target
		US	Europe	Asia	US	Europe	Asia	
Sales 126.8	Software 90.2	17.3	10.1	62.8	→	↗	↗	189.8
	SaaS 36.6	18.9	5.2	12.5	↘	→	↗	
Gross Margin 40.8	Software 45.0	45.6	44.2	44.9	→	→	↗	44.0
	SaaS 37.9	37.8	38.0	38.1	→	→	↗	
P&L 37.8	Software 28.2	7.2	2.8	18.2	→	↗	↗	50.3
	SaaS 9.6	3.4	0.8	5.4	→	→	→	
Customer Attrition 0.8%	Software 1.1	1.1%	1.1%	1.1%	↗	↗	↗	1%
	SaaS 0.2%	0.2%	0.2%	0.2%	↗	↗	↗	

On target · Playing catch-up · Off target

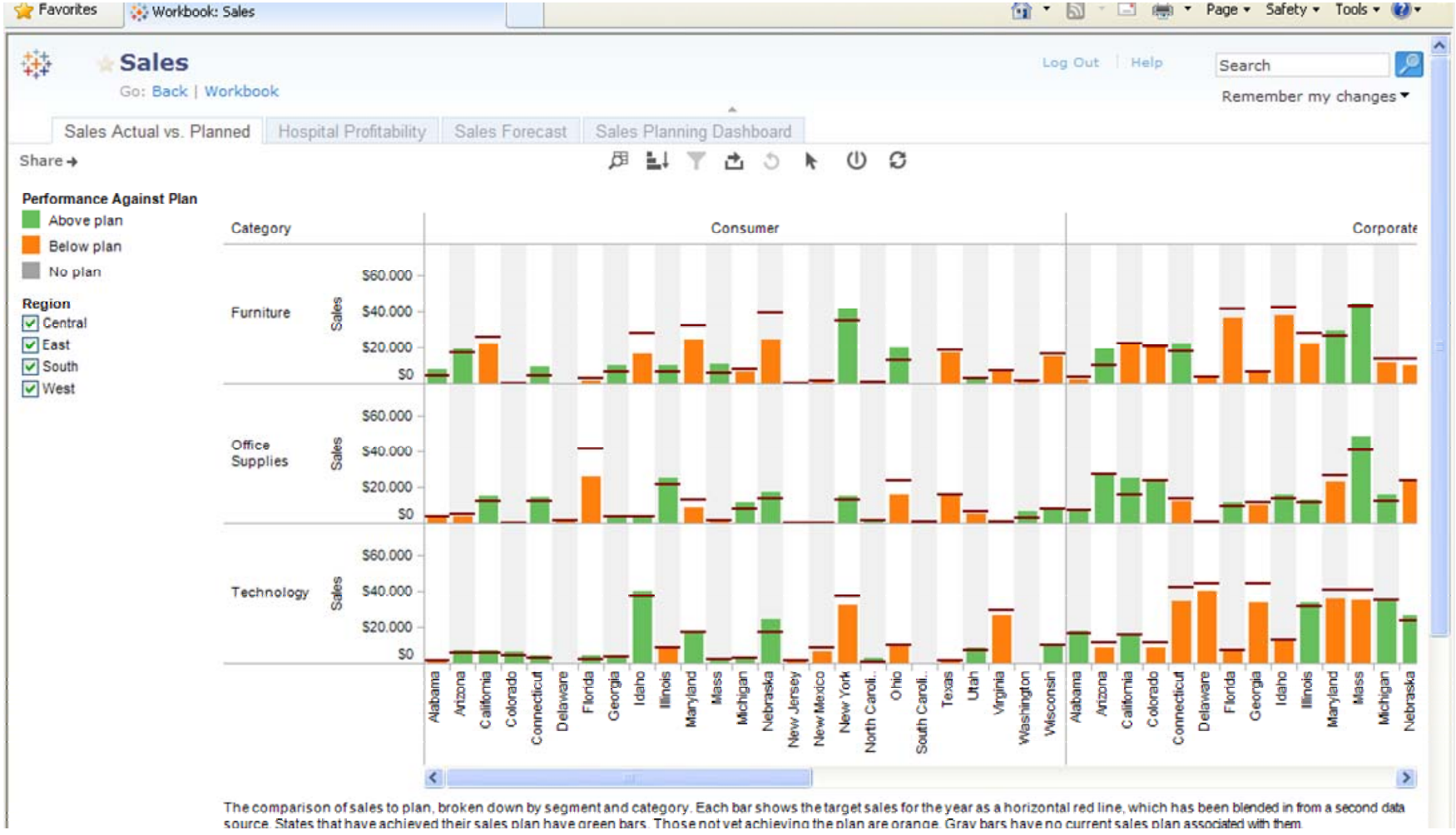
Up trend · Flat trend · Down trend

Sales Dashboard Example

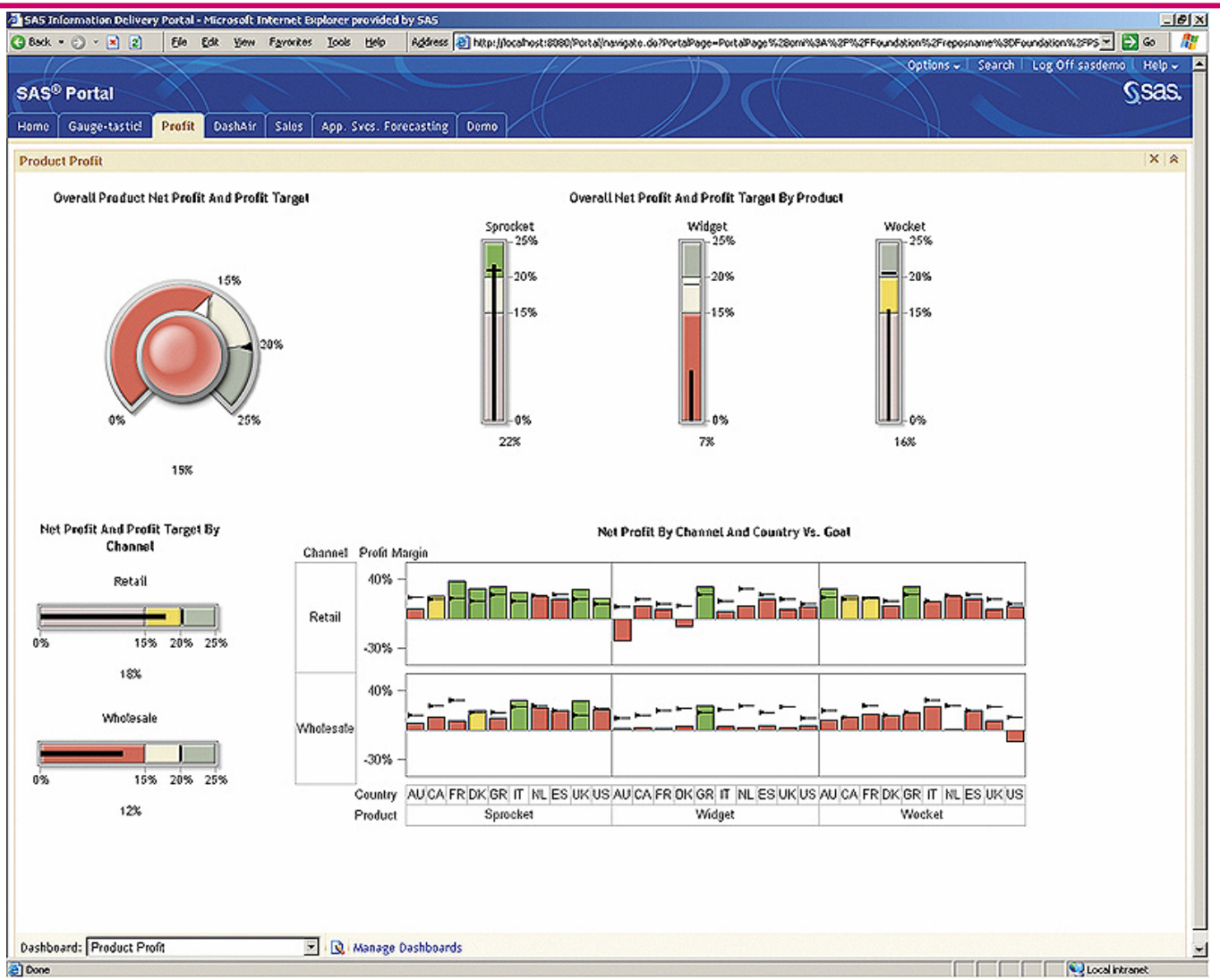
Buisness Analytics Tools – Manager Dashboards



Bars in business dashboards – Tableau Software

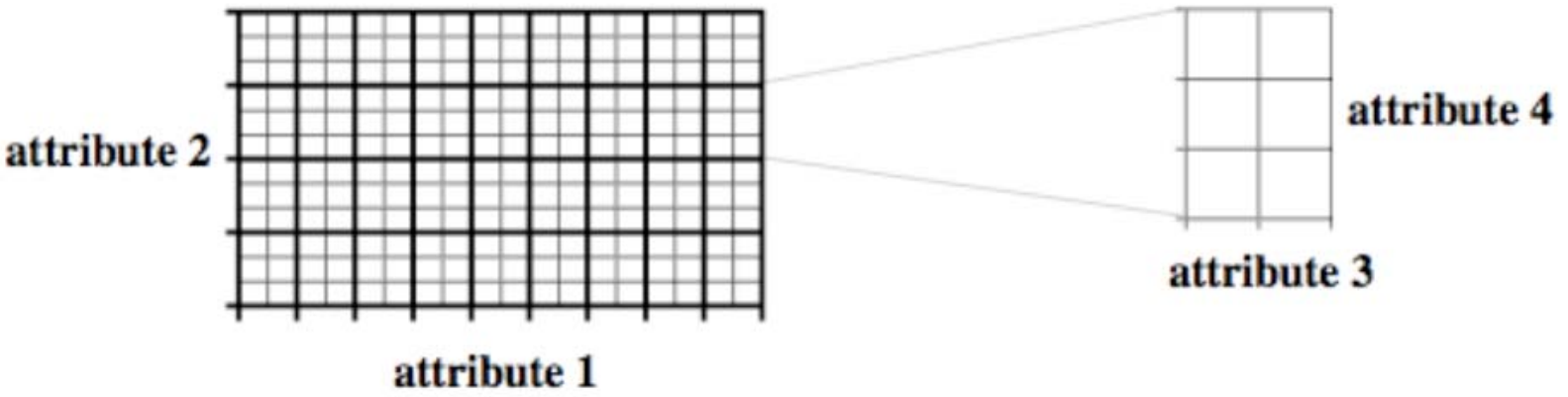


Data analytics – kokpity menadżerskie



- SAS Enterprise BI

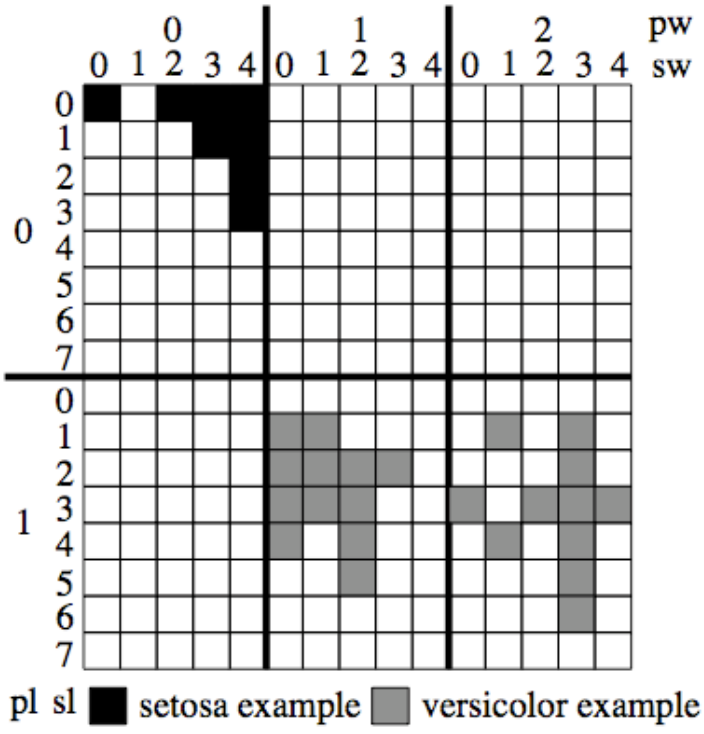
Multidimensional Stacking

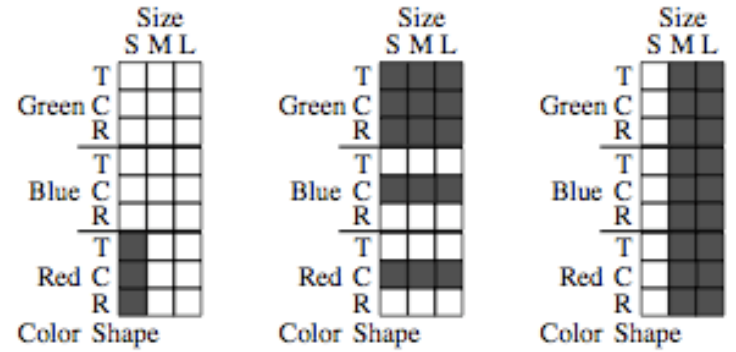


Multidimensional presentation of nominal attributes

- VL1 diagrams (Michalski 70) for machine learning

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





Target concept, $t = 1 \ldots 40$.

Target concept, $t = 41 \ldots 80$.

Target concept, $t = 81 \ldots 120$.

STAGGER and concept drif

Hierarchiczne wizualizacje - Treemaps

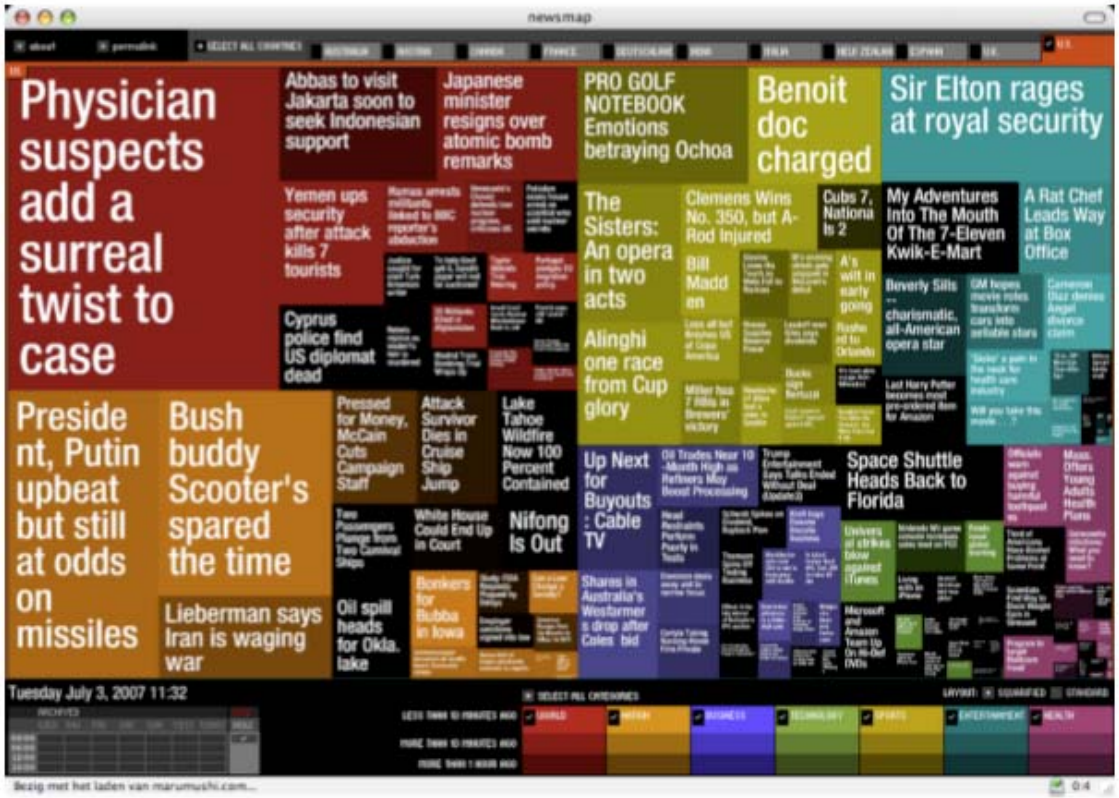


Figure 4: Treemap used for displaying news http://newsmap.jp/

- 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 New York Times | Business

The Top 1 Percent: What Jobs Do They Have?

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

RELATED ARTICLE
Among the Wealthiest One Percent, Many Variations

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

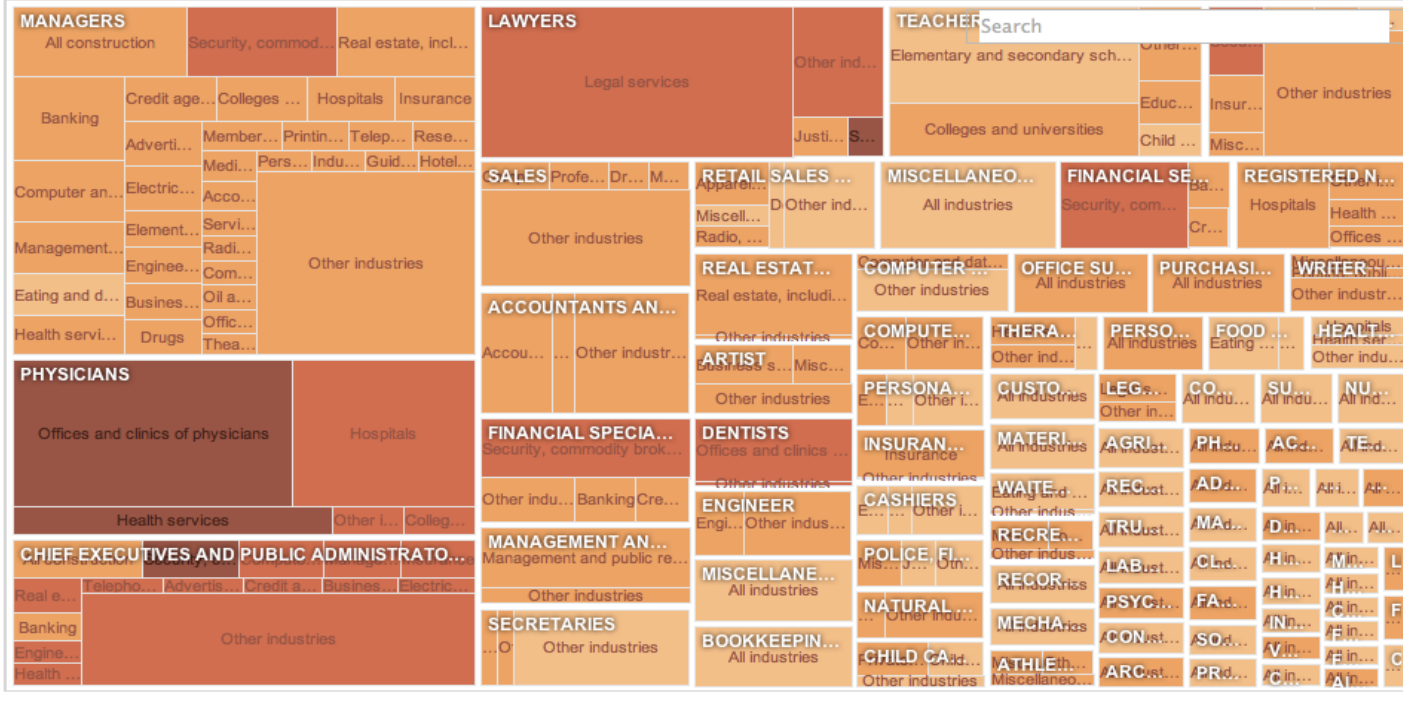
Lawyers who work on Wall Street are twice as likely as those in general practice to make the top 1 percent. ZOOM TO

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.

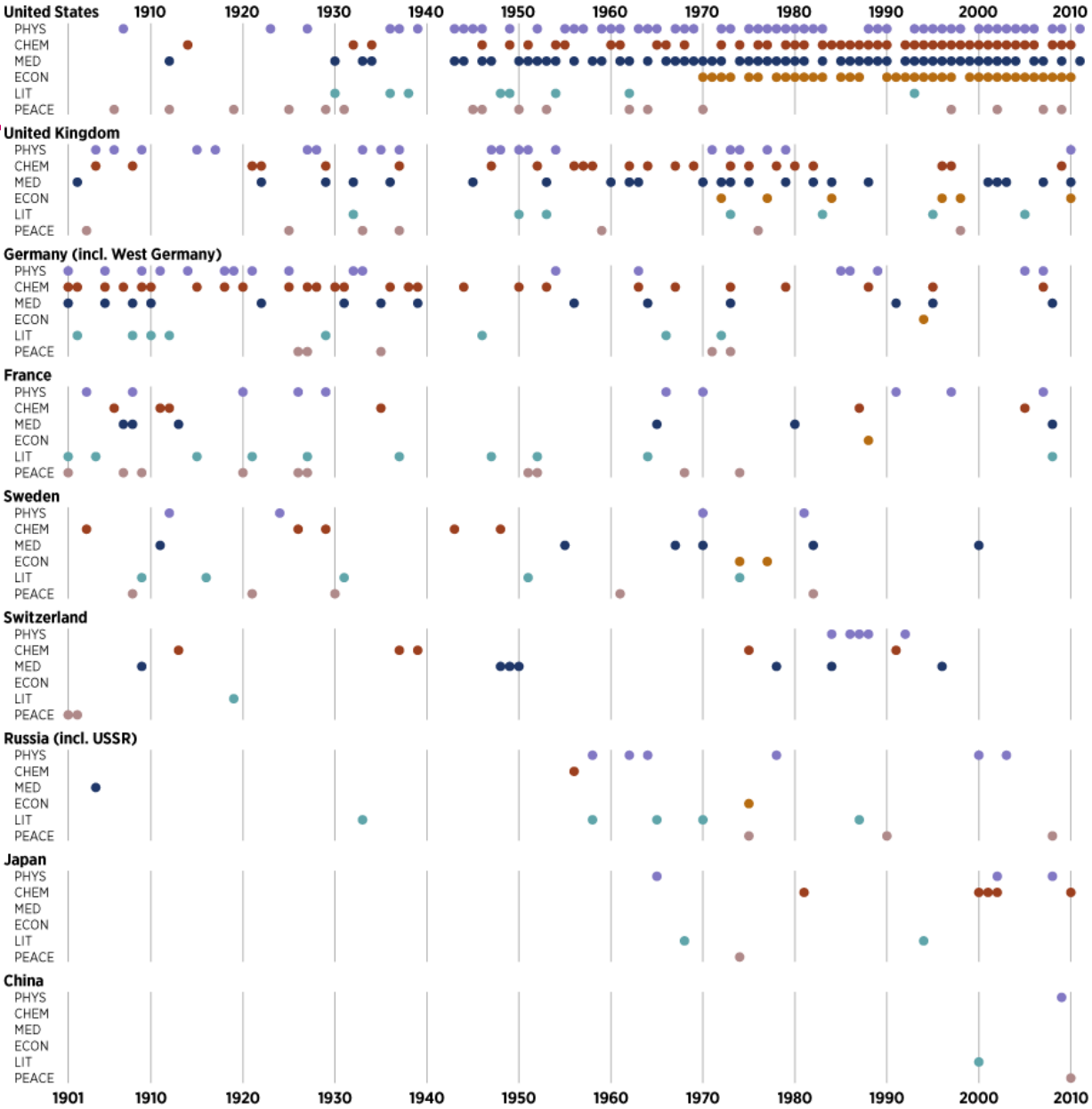
ZOOM IN ZOOM OUT FULL SCREEN Search



Published: January 15, 2012 | By JEREMY WHITE, ROBERT GEBELOFF, FORD FESSENDEN AND SHAN CARTER | Source: IPUMS

TWITTER | FACEBOOK

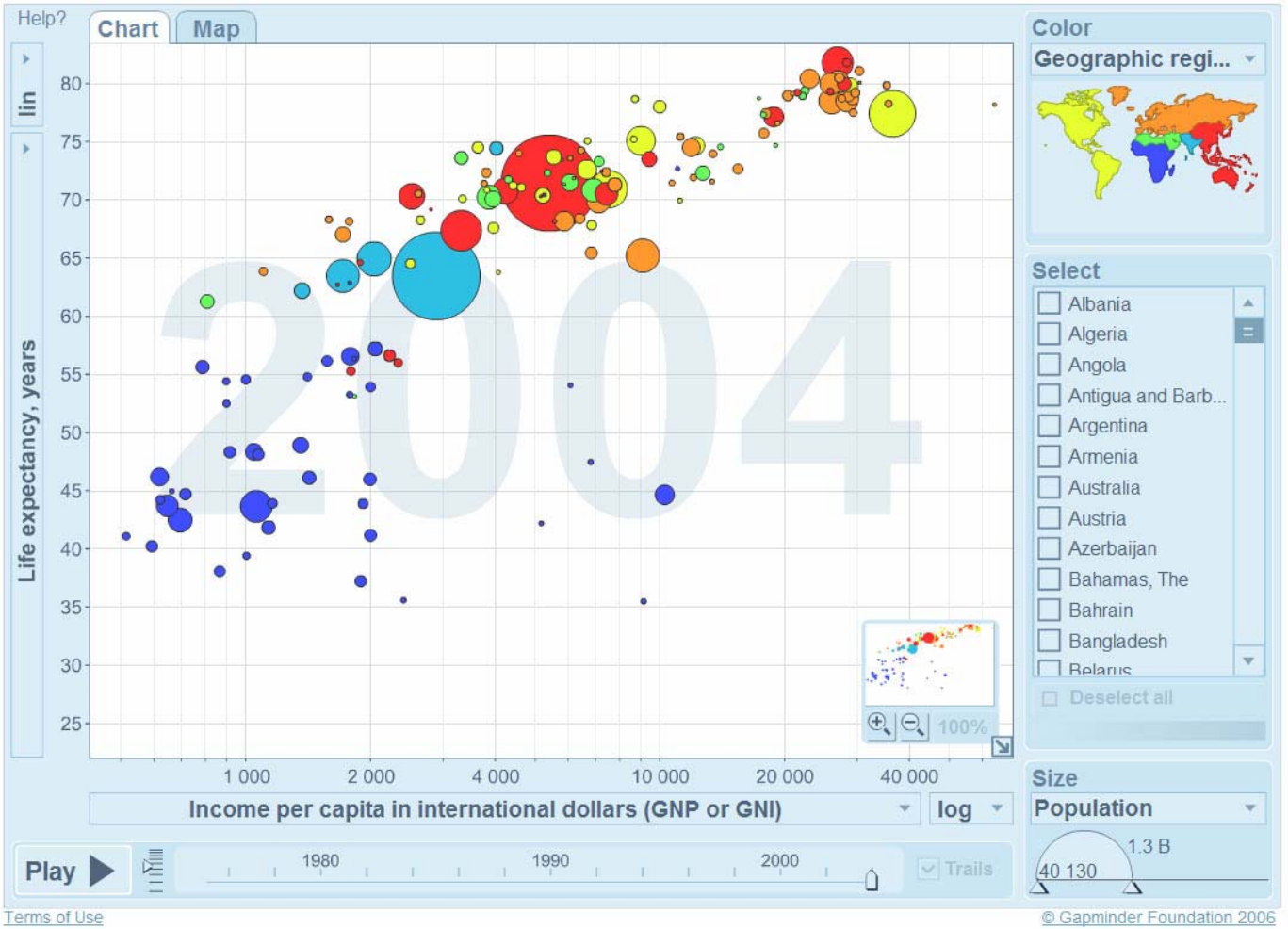
By Country and Prize



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

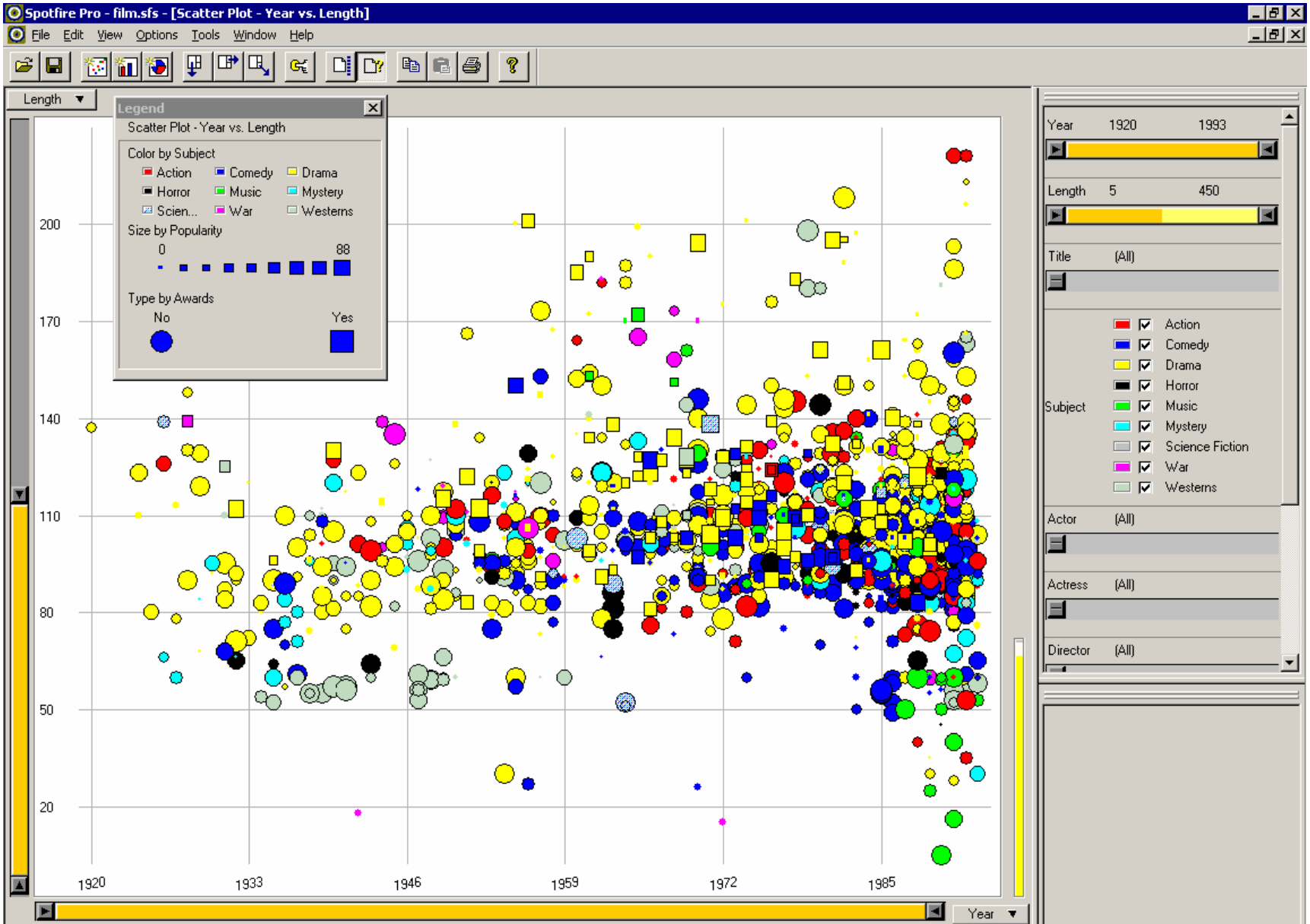
Gapminder – Motion Charts



http://www.gapminder.org/

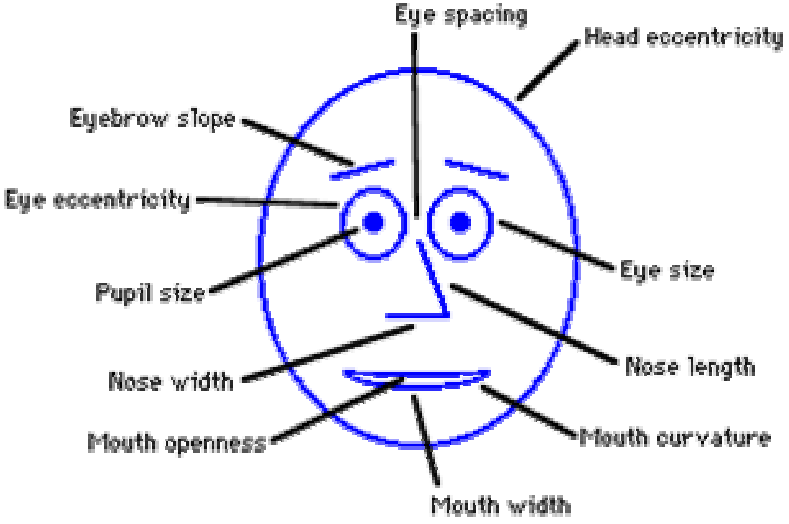
Using Bubble presentations

Spotfire

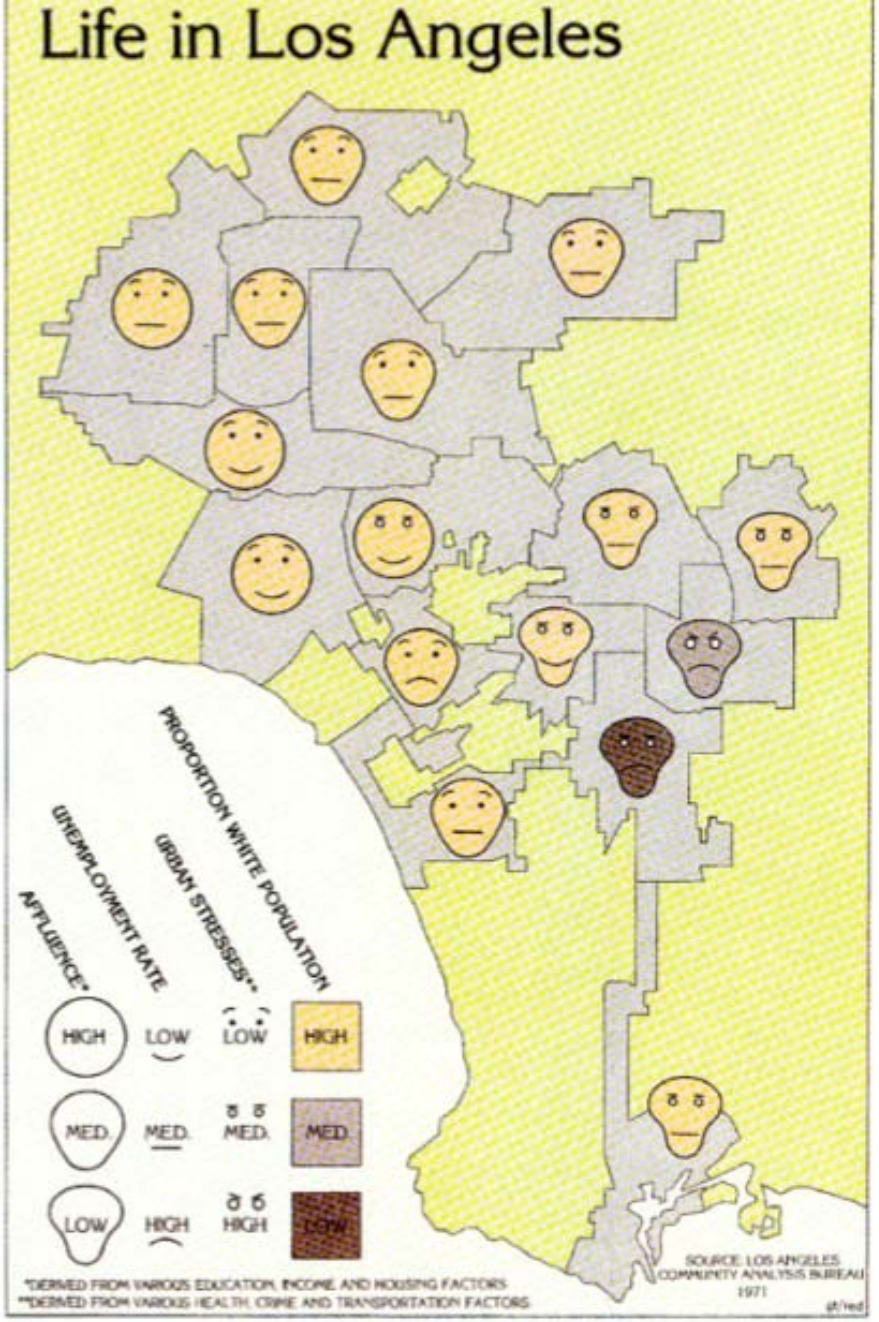

Spotfire Pro - film.sfs - [Scatter Plot - Year vs. Length]
File Edit View Options Tools Window Help
Length
Legend
Scatter Plot - Year vs. Length
Color by Subject
Action
Comedy
Drama
Horror
Music
Mystery
Scien...
War
Westerns
Size by Popularity
0
88
Type by Awards
No
Yes
200
170
140
110
80
50
20
1920
1933
1946
1959
1972
1985
Year
Year 1920 1993
Length 5 450
Title (All)
Subject
Action
Comedy
Drama
Horror
Music
Mystery
Science Fiction
War
Westerns
Actor (All)
Actress (All)
Director (All)

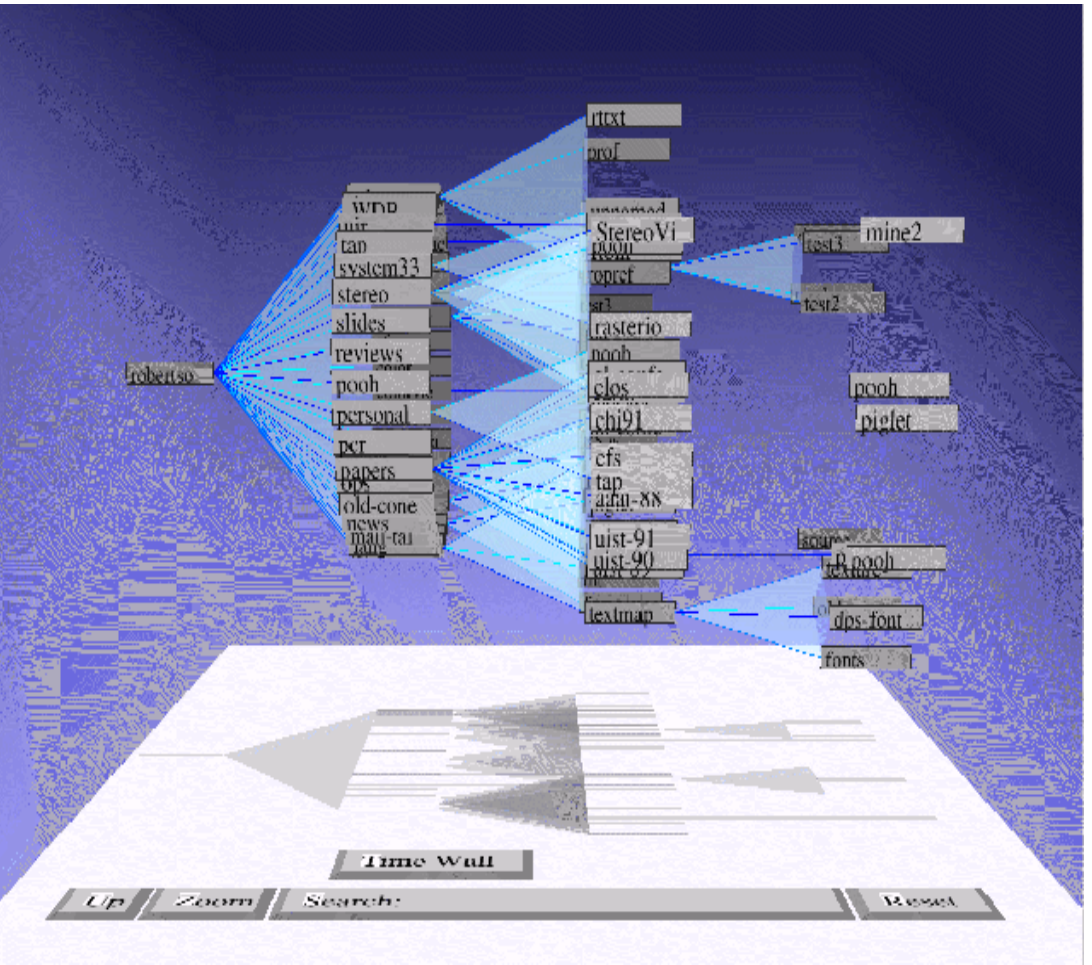
Chernoff Faces

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`

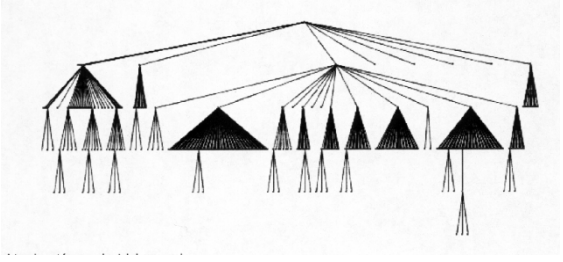




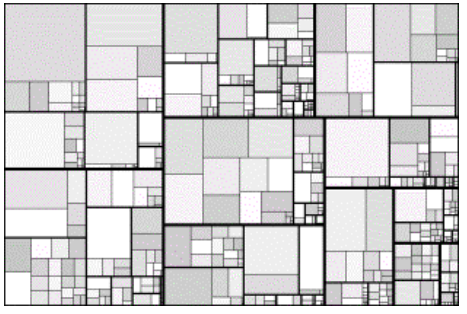
Cone Trees [RMC91]

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

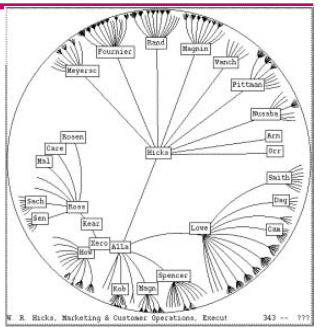
Abstract ➔ Hierarchical Information – Preview



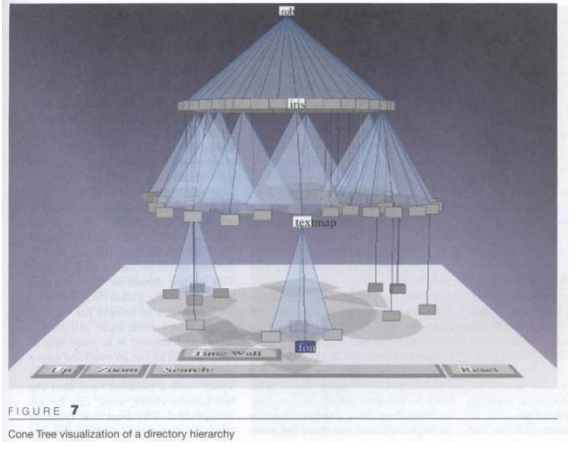
Traditional



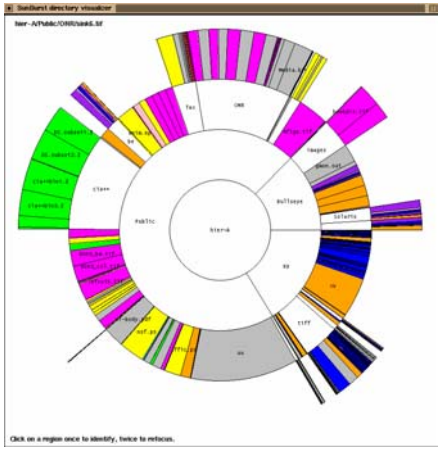
Treemap



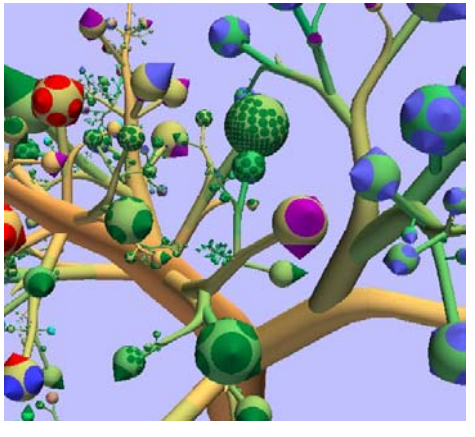
Hyperbolic Tree



ConeTree

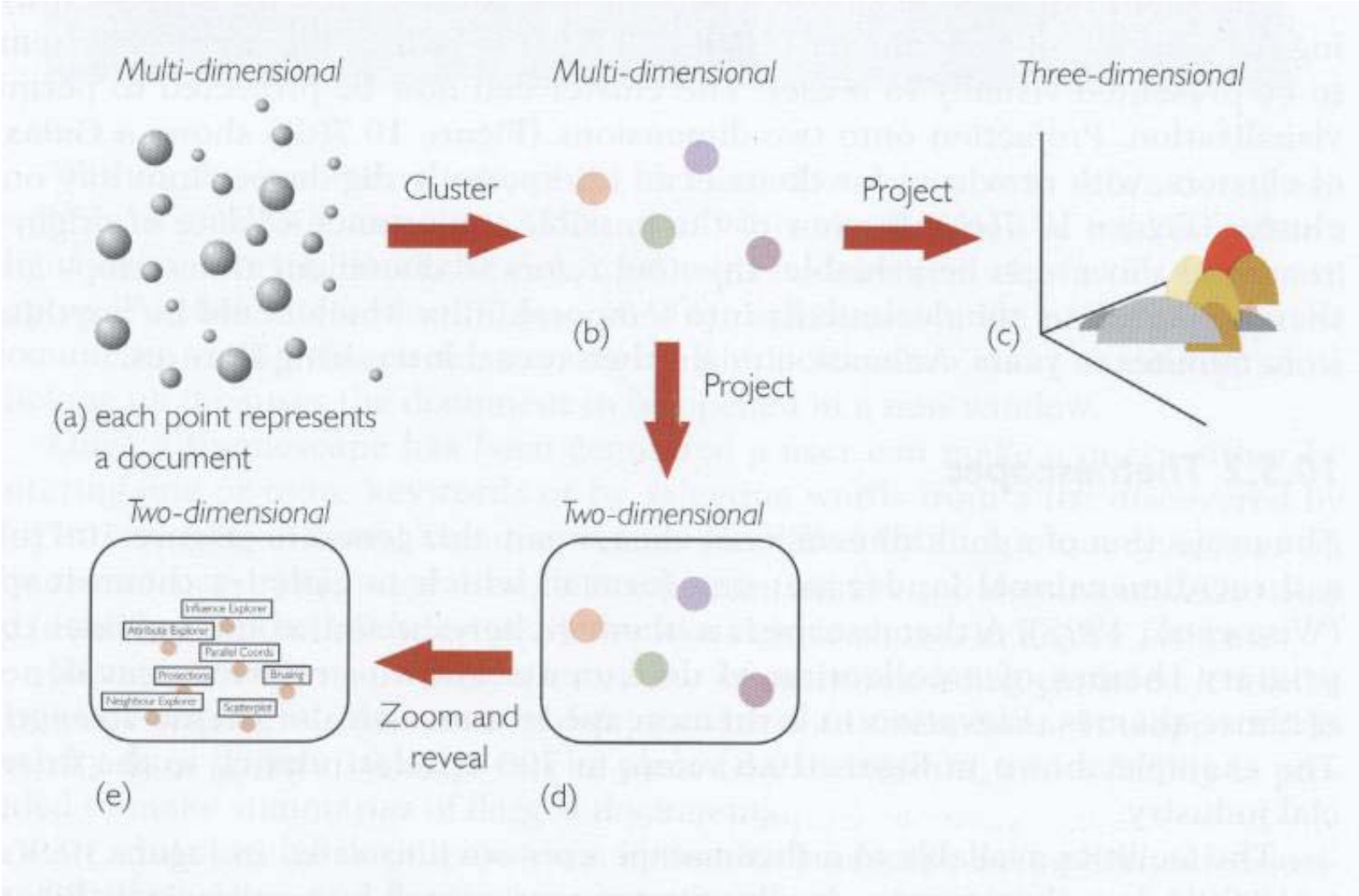


SunTree

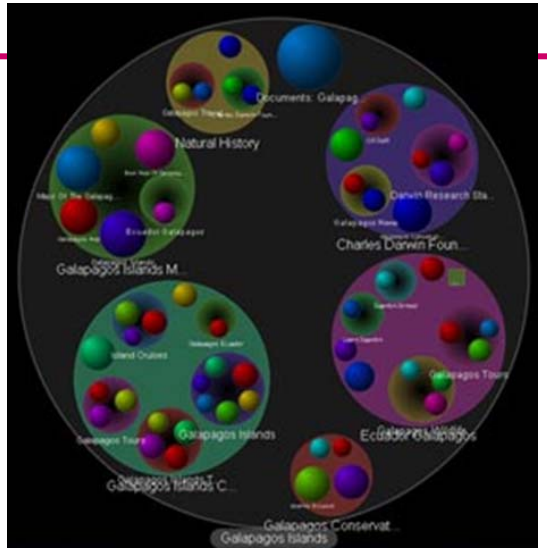


Botanical

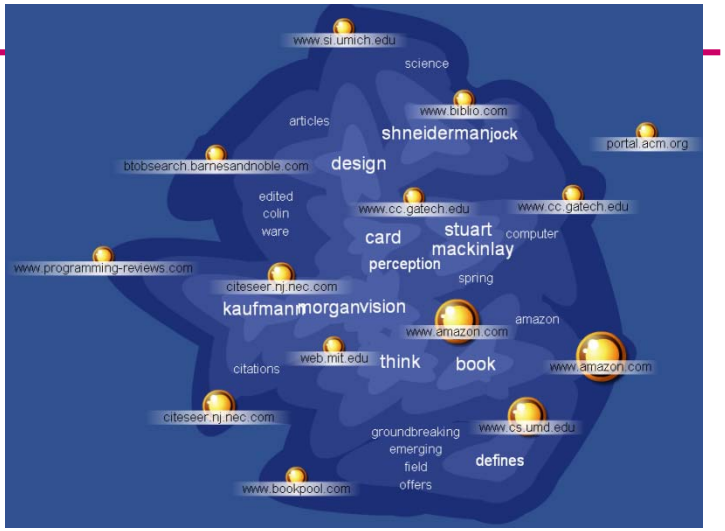
Visualization of Search Results & Inter-Document Similarities



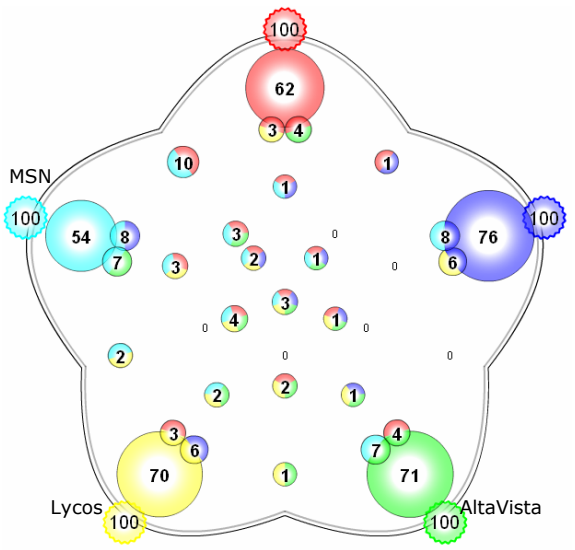
Abstract ➔ Text – **MetaSearch** Previews



Grokker



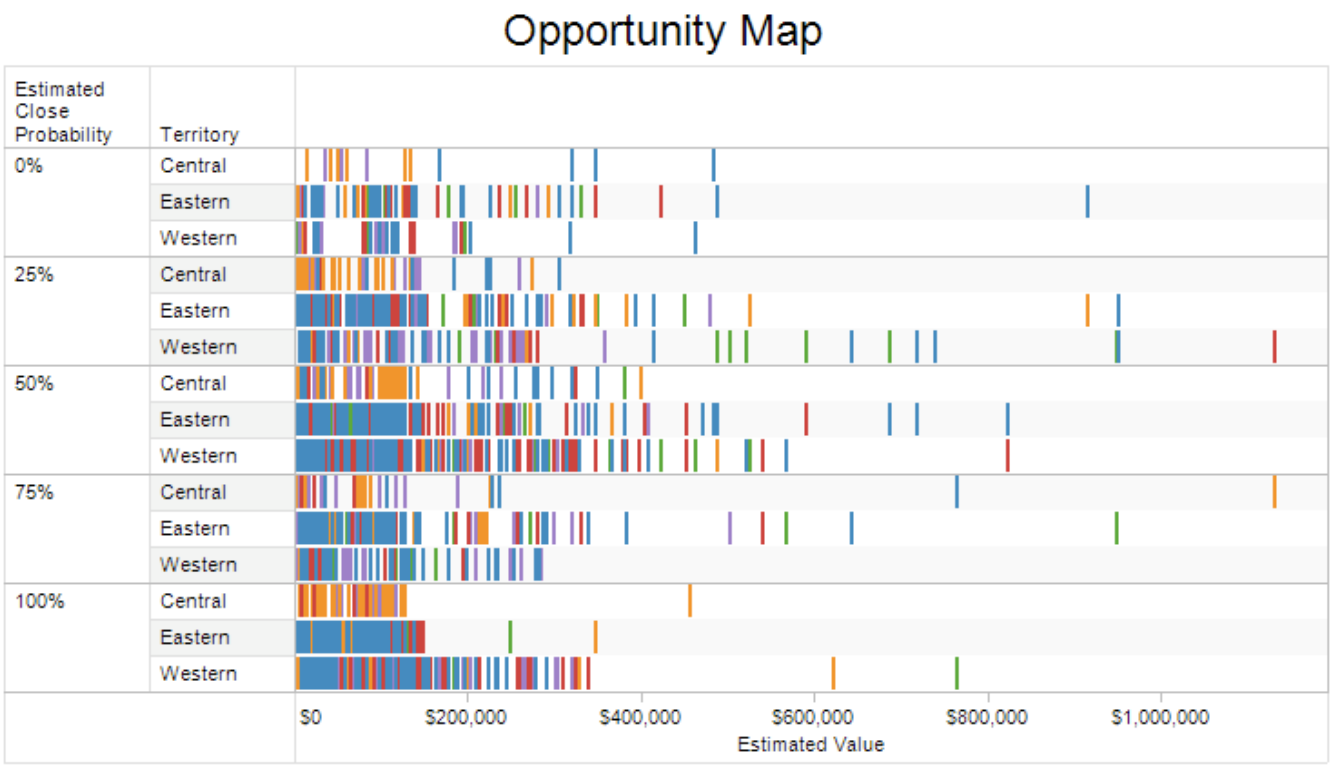
Kartoo



MetaCrystal ➔ searchCrystal

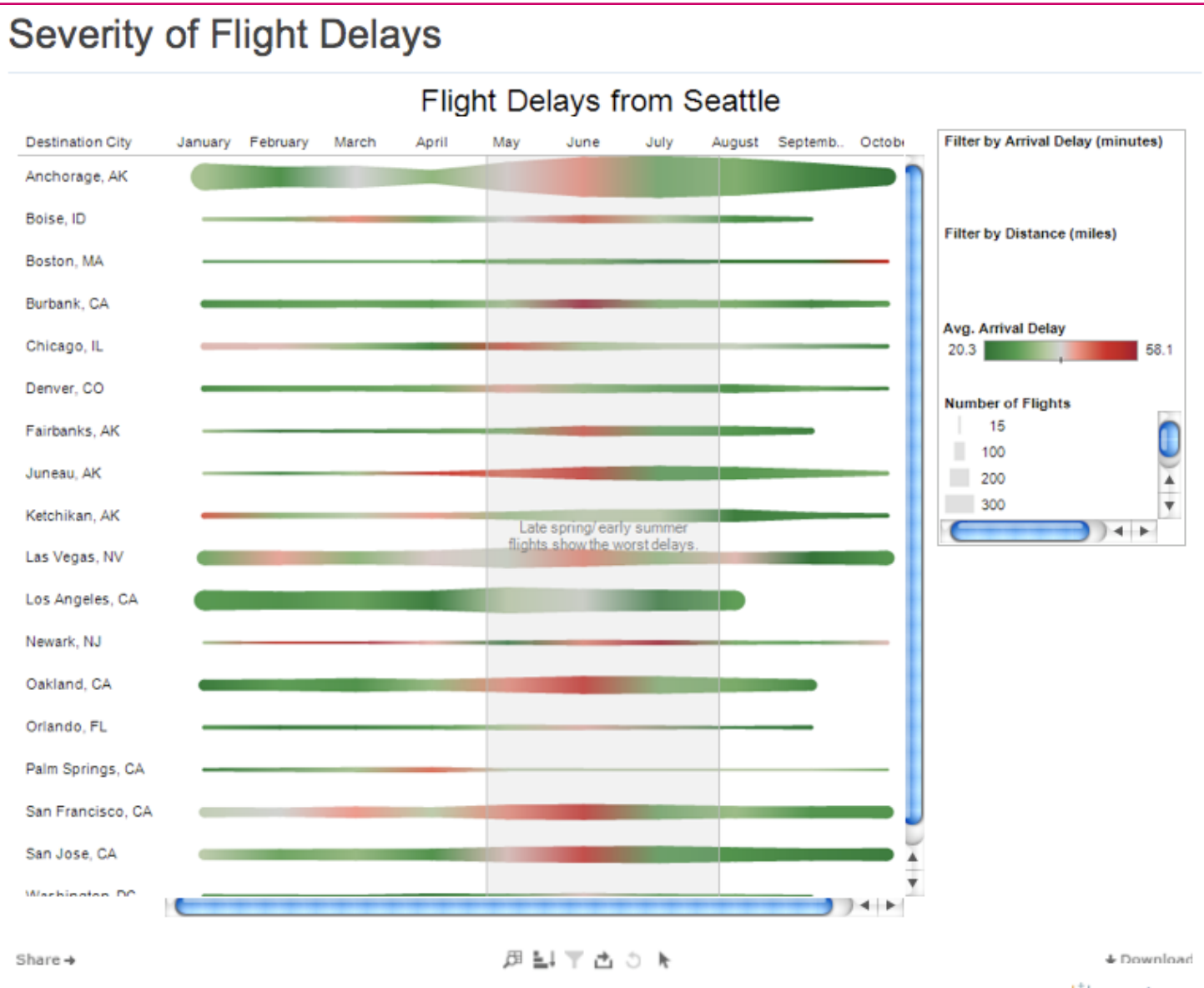
Analysis of CRM data

Opportunity Map | Actual vs. Forecast | Distribution of Deal Size | Current Pipeline | Current Pipeline Report

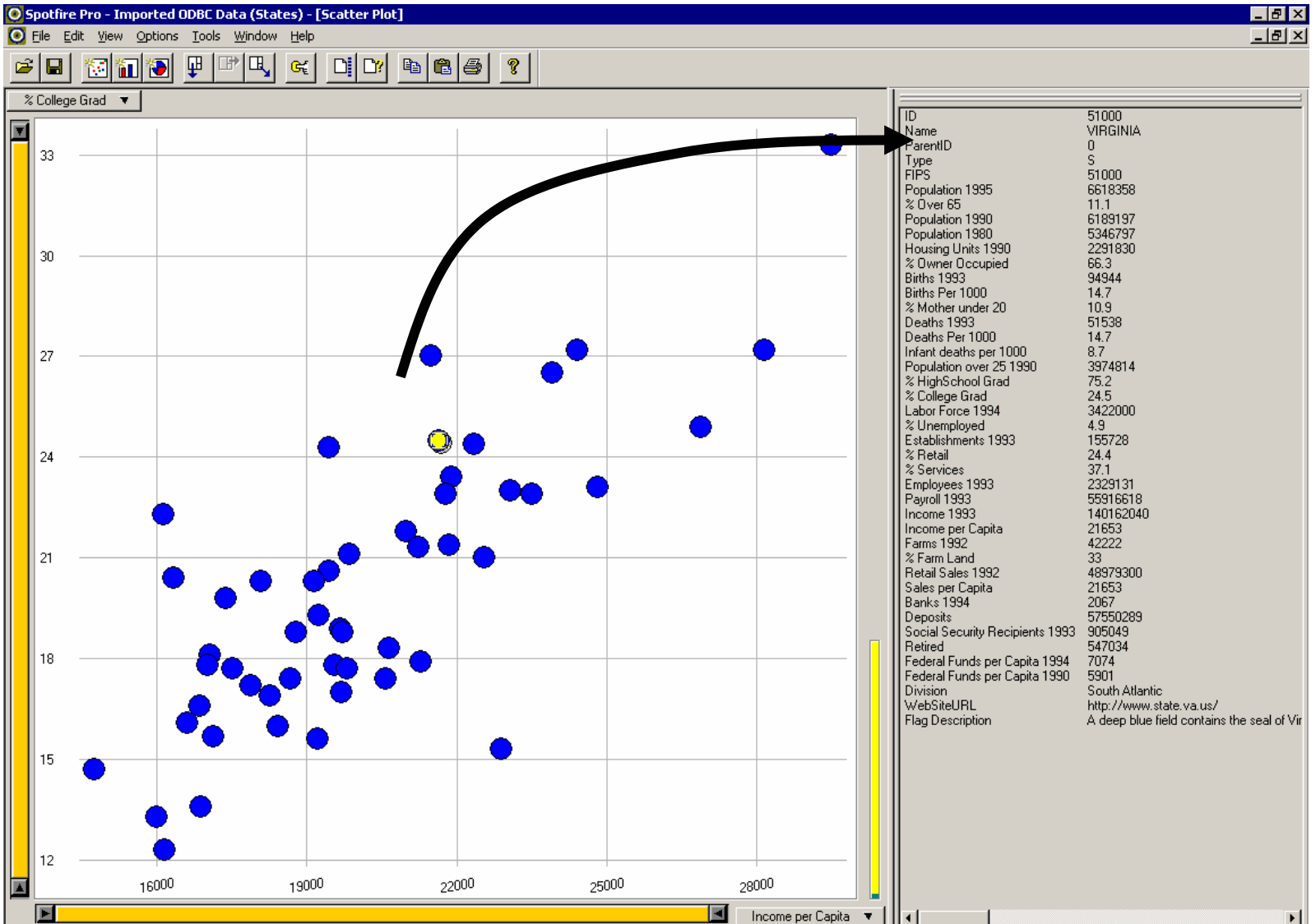


Visualization shows a mark for each deal in the pipeline; mark is placed based on deal size, region and probability stage (estimated percent chance a deal will closed this quarter). Color is by Industry. Click the "+" above territory to drill down to individual salespeople.

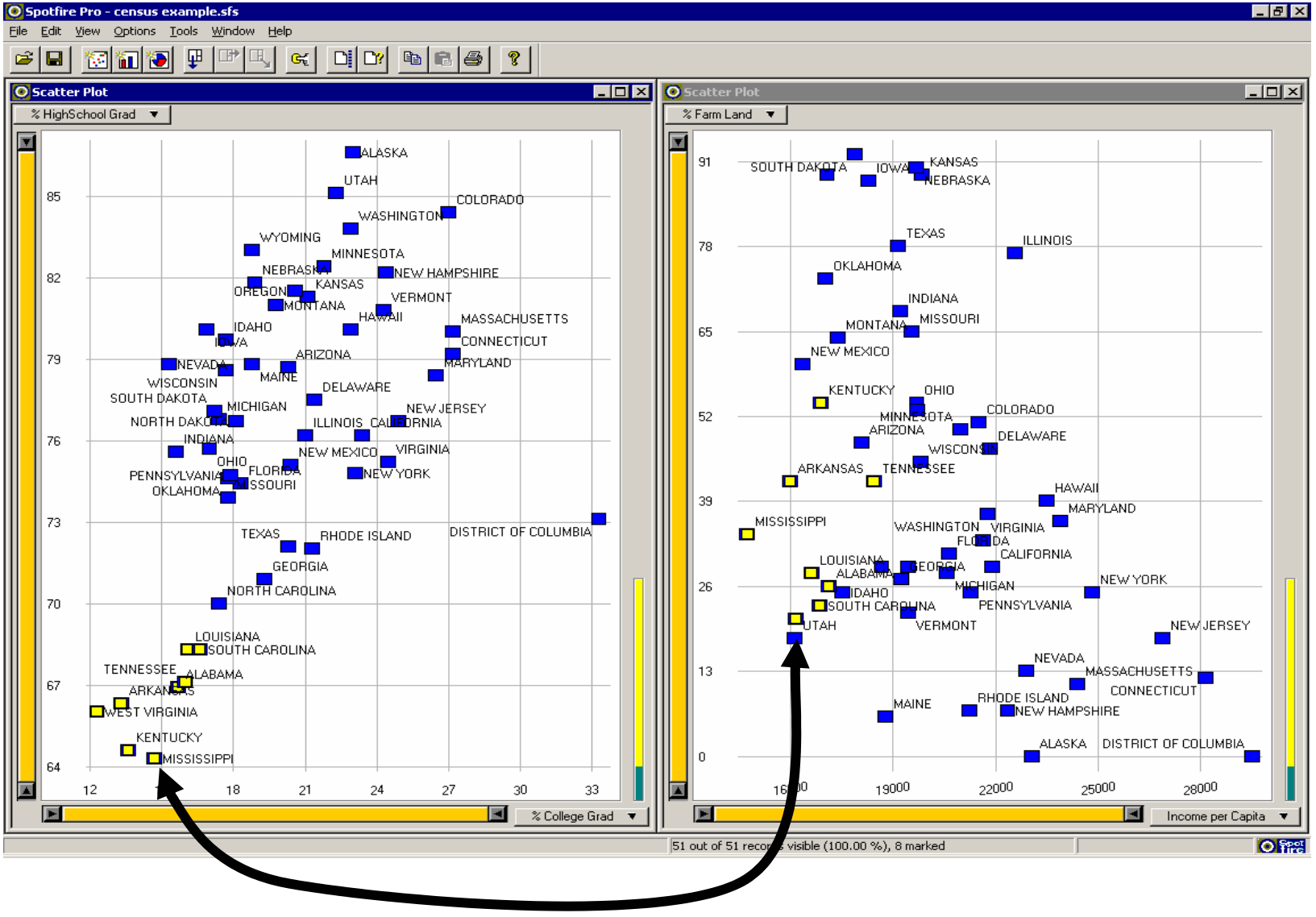
Visualization of different conditions



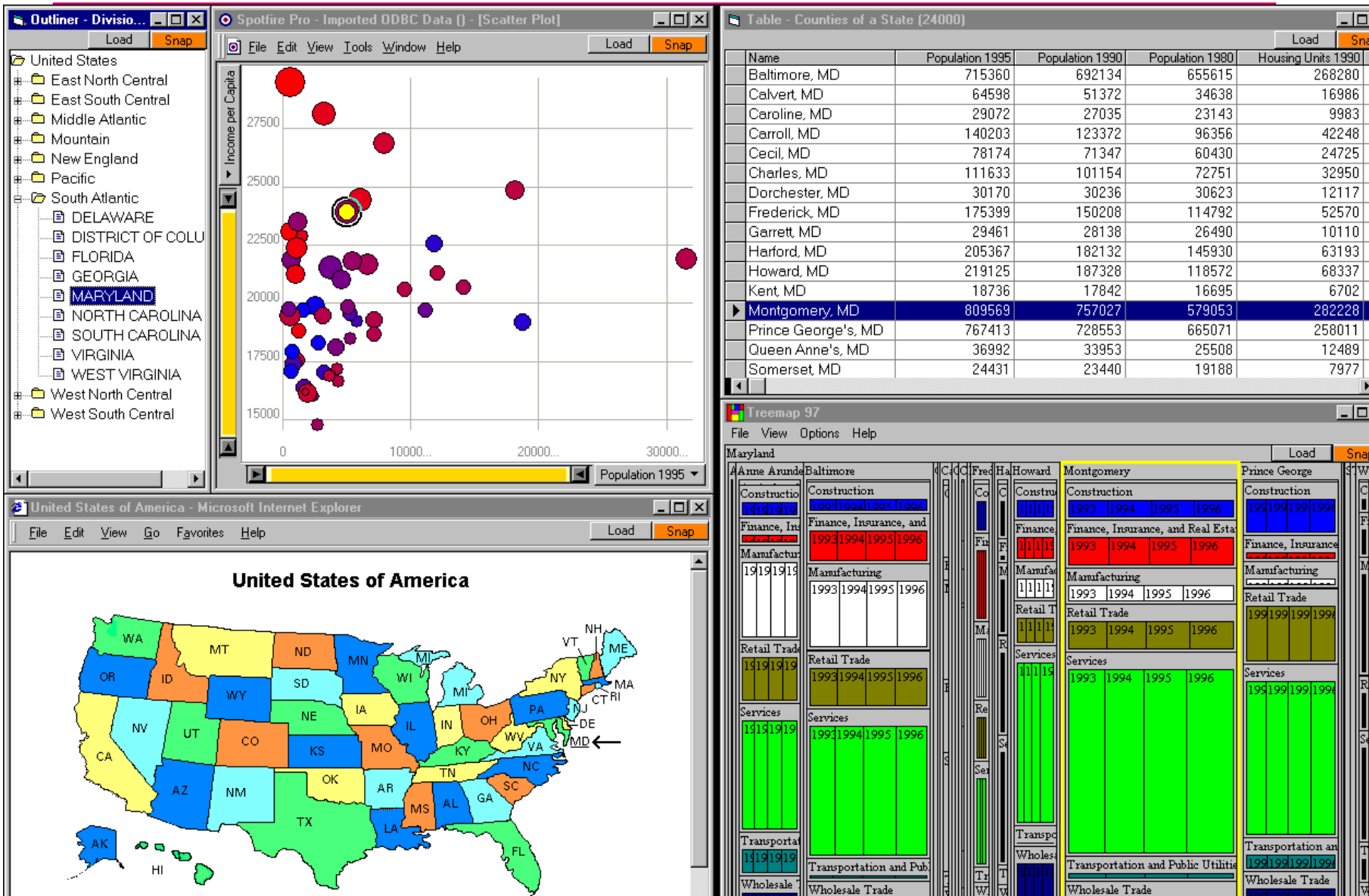
Overview and Detail



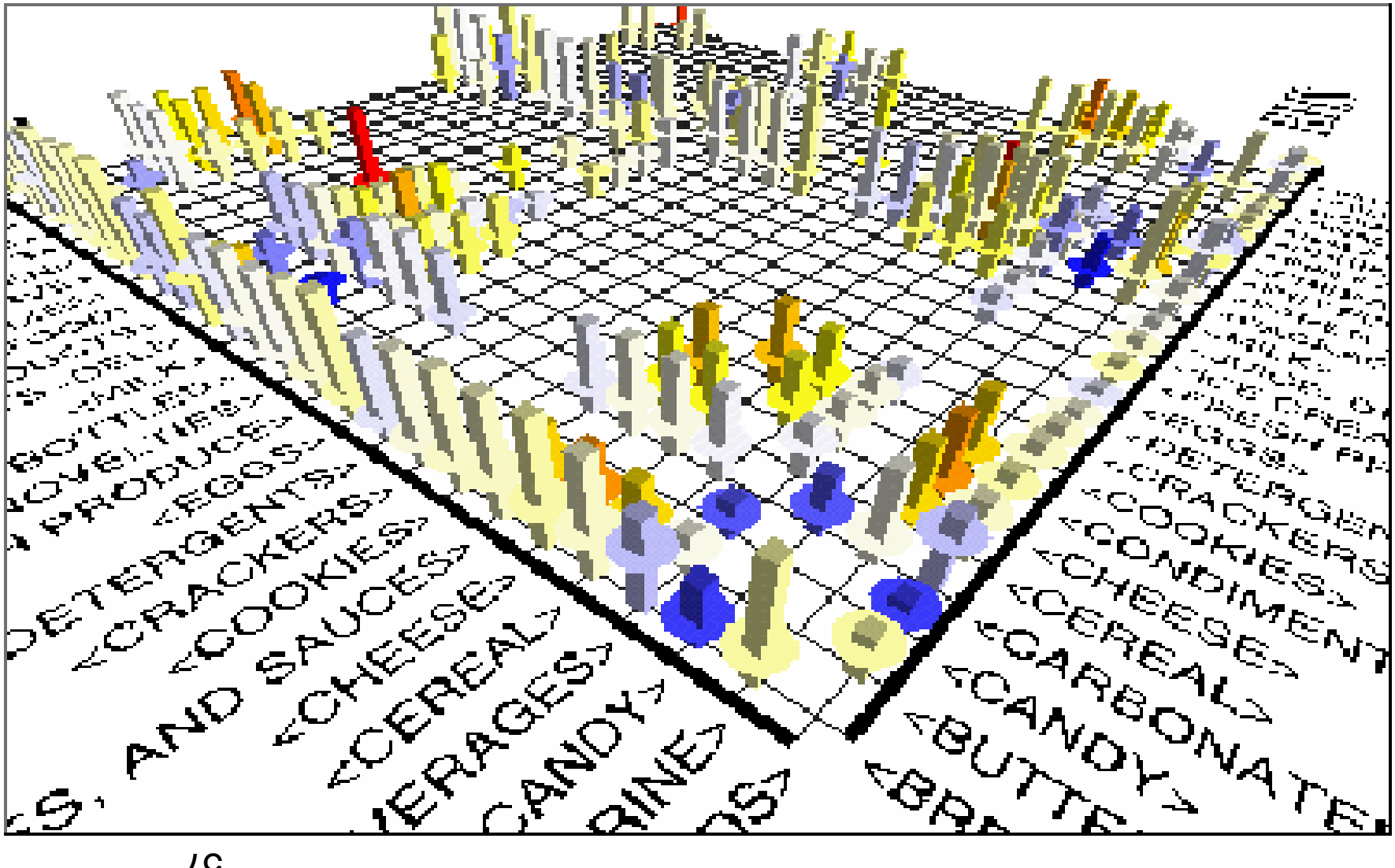
Brushing and Linking



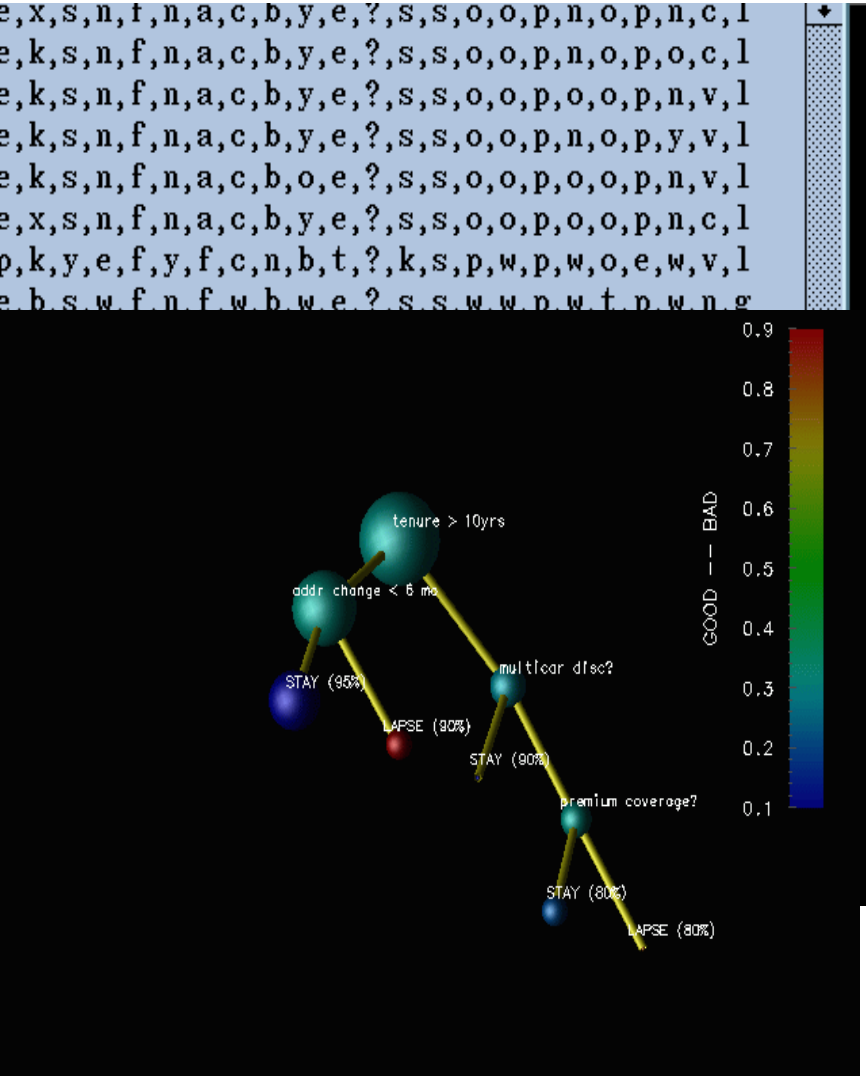
Census Data

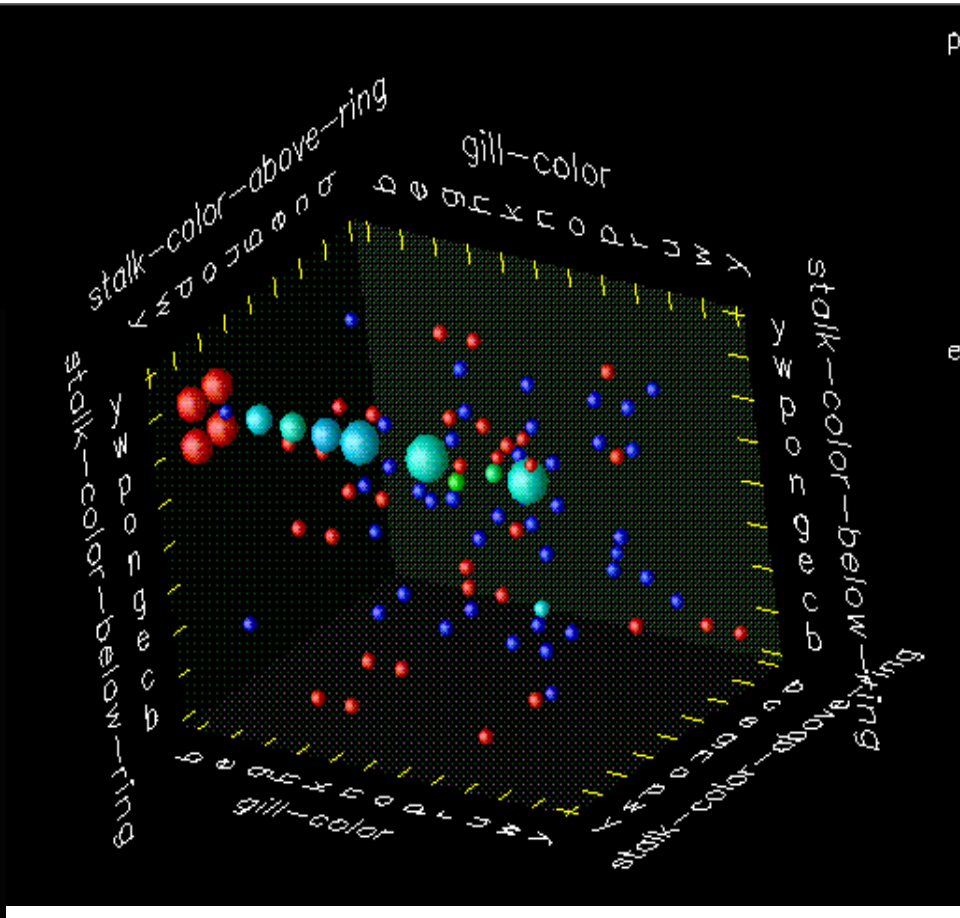

Outliner - Divisio...
Load
Snap
United States
East North Central
East South Central
Middle Atlantic
Mountain
New England
Pacific
South Atlantic
DELAWARE
DISTRICT OF COLU
FLORIDA
GEORGIA
MARYLAND
NORTH CAROLINA
SOUTH CAROLINA
VIRGINIA
WEST VIRGINIA
West North Central
West South Central
Spotfire Pro - Imported ODBC Data () - [Scatter Plot]
File Edit View Tools Window Help
Load
Snap
Income per Capita
27500
25000
22500
20000
17500
15000
0
10000...
20000...
30000...
Population 1995
Table - Counties of a State (24000)
Load
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
Treemap 97
File View Options Help
Load
Maryland
Anne Arunde
Baltimore
Howard
Montgomery
Prince George
Construction
Finance, Insurance, and Real Estate
Manufacturing
Retail Trade
Services
Transportation and Public Utilities
Wholesale Trade
United States of America - Microsoft Internet Explorer
File Edit View Go Favorites Help
Load
Snap
United States of America
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
VT
NH
ME
MA
CT
RI
NJ
DE
MD
AK
HI

Visualization of Association Rules in SGI/MineSet 3.0

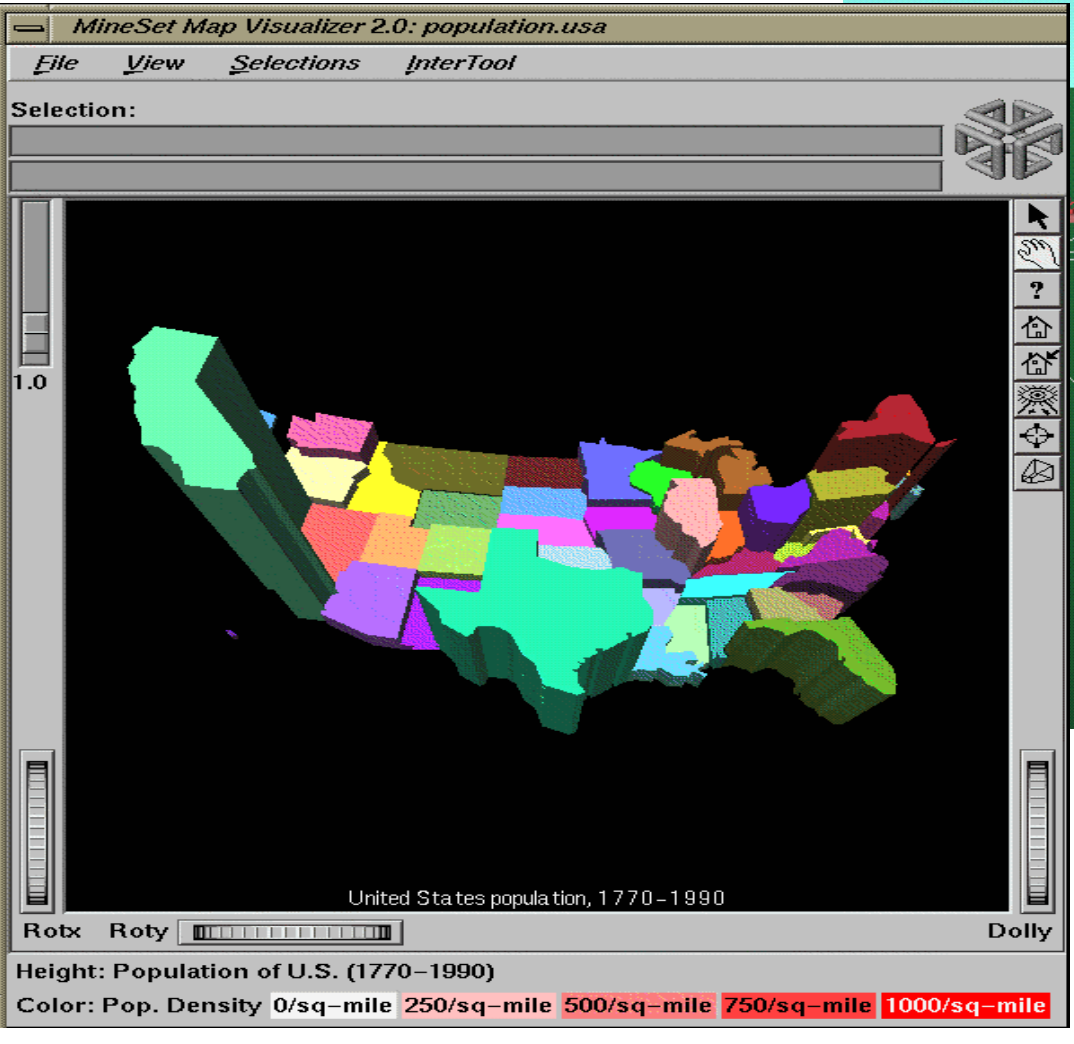


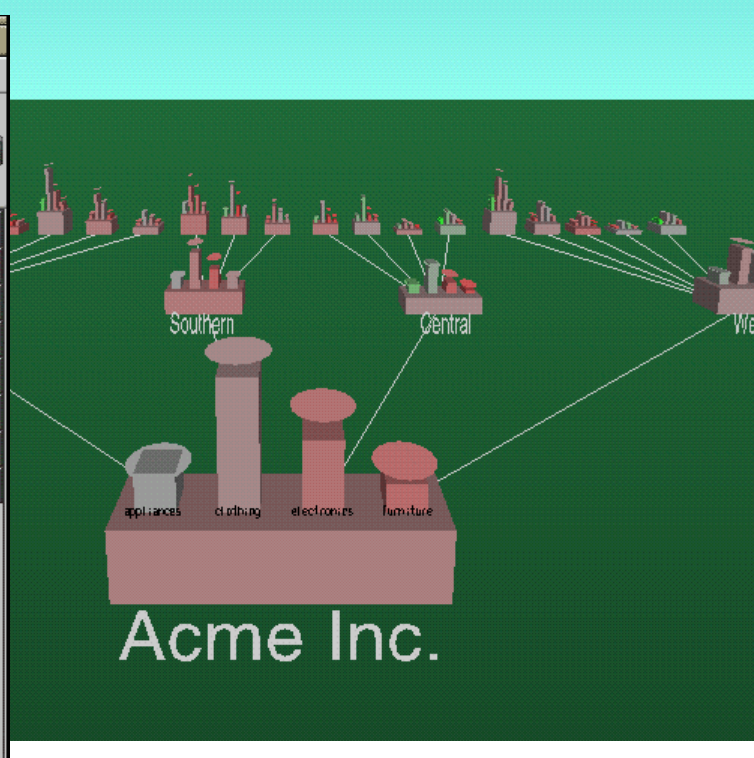
IBM Miner – visualization of mining results



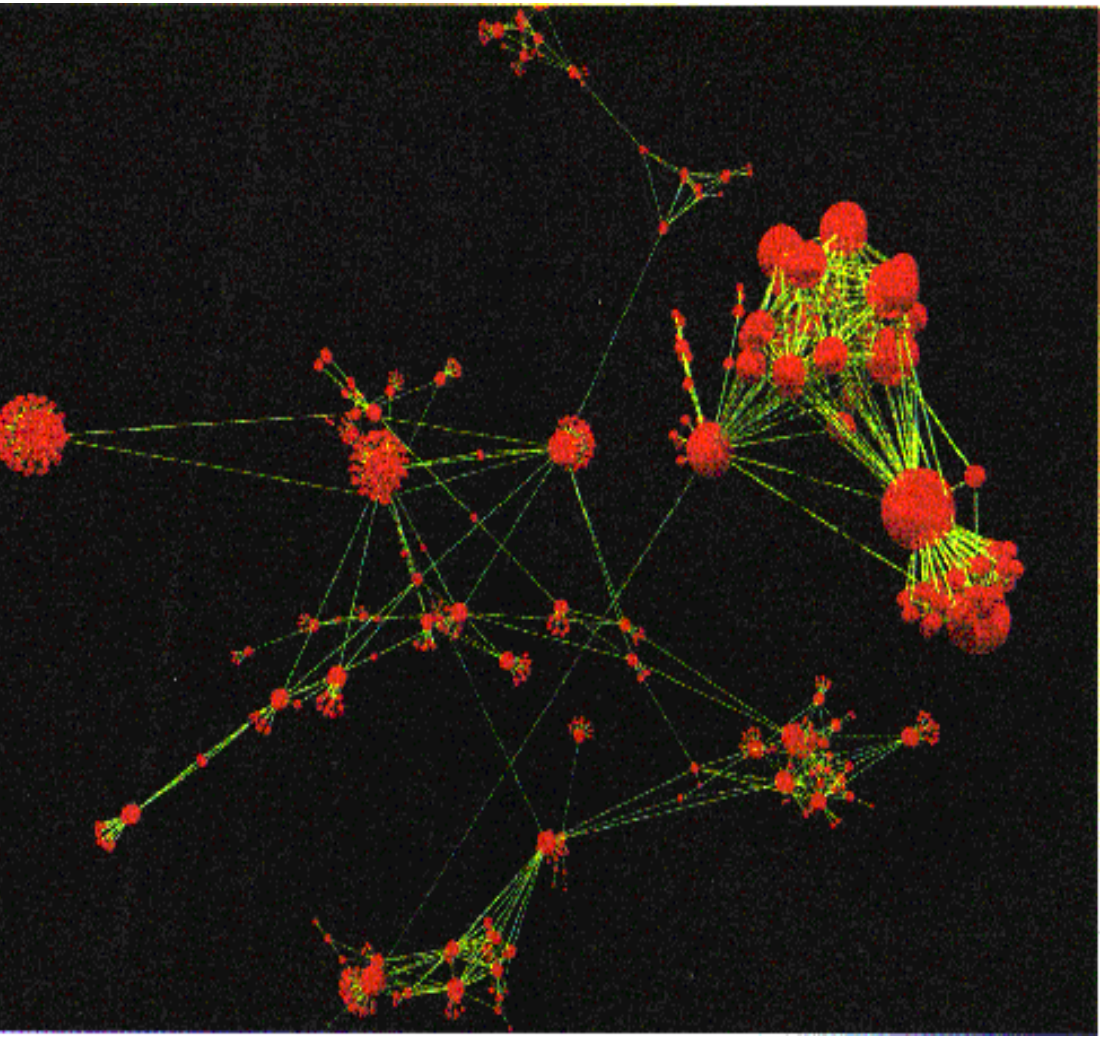


SGI – other tools





Graph-based Techniques



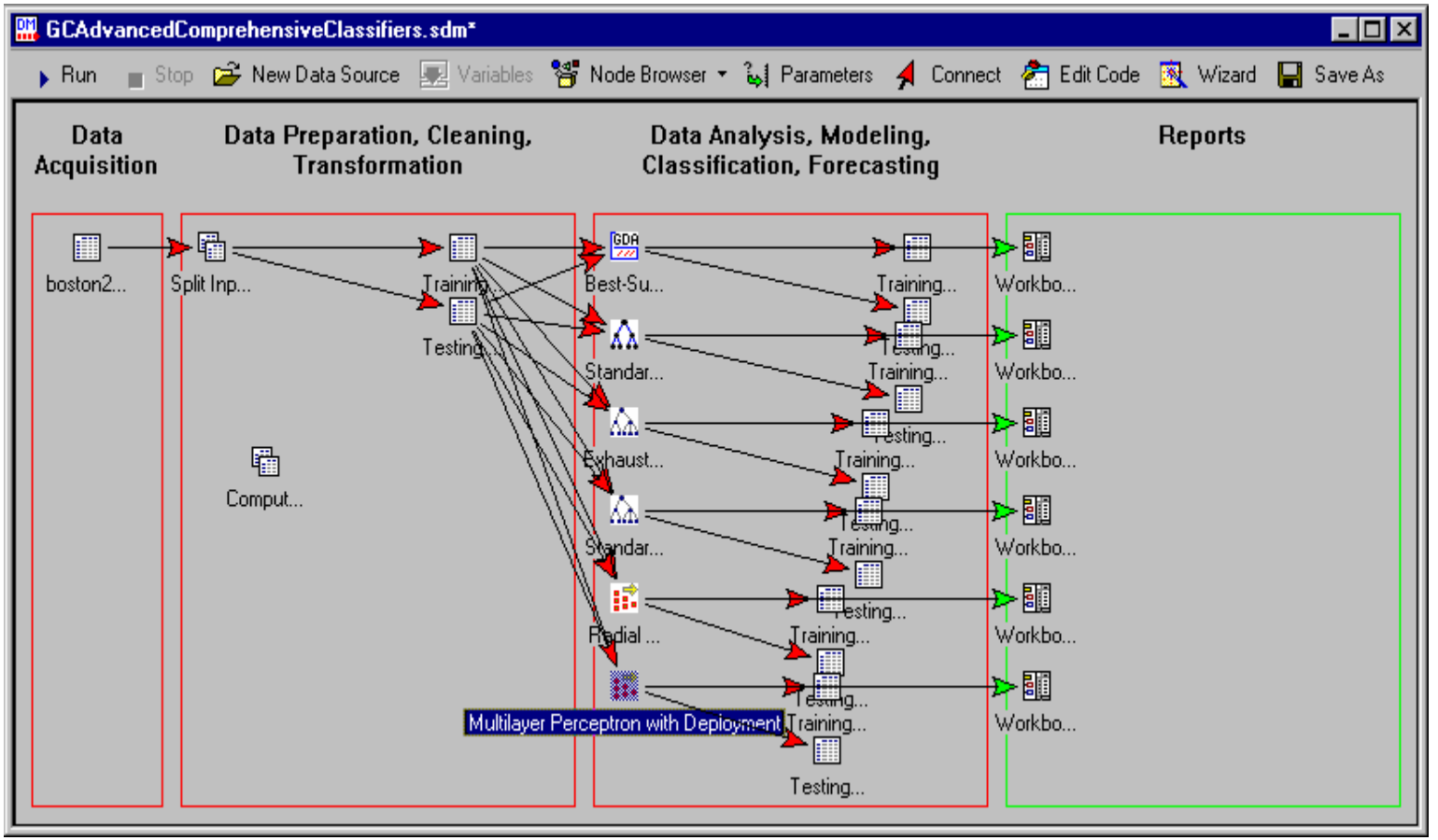
Narcissus

- Visualization of a larg number of web pages
- visualization of compl 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 $\rightarrow$ you may produce a kind of scipt representing a graphical flow of KD process

Statsoft – Data mining graphical panel



RapidMiner (YALE)

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!