TIES443

Lecture 9

Visualization

Mykola Pechenizkiy

Course webpage: *http://www.cs.jyu.fi/~mpechen/TIES443*

November 17, 2006

Department of Mathematical Information Technology
University of Jyväskylä

Topics for today

- Purpose of visualization in DM/KDD
- Visualization of data
  - 1D, 2D, 3D, 3+D
- Visualization of data mining results
  - Model visualization, model performance visualization
- Visualization of data mining processes
- Application …
  - Interactive data mining and visualization
  - Time-series visualization
  - Data mining as part of visualization system or vice versa?

Sources

- Books:
  - "Information Visualization in DM and KD" Fayyad *et al.*
  - "Information Visualization: Beyond the Horizon": 2nd ed, C. Chen
  - "The Visual Display of Quantitative Information", 2nd ed, E. Tufte
- People:
  - *Edward Tufte, Ben Shneiderman, Daniel A. Keim, Marti Hearst, Tamara Munzner, and other*
- Excellent handouts and recommended reading on information visualization by *Tamara Munzner*
  - http://www.cs.ubc.ca/~tmm/courses/cpsc533c-04-spr/
- Some papers:
  - Visual Data Mining Framework for Convenient Identification of Useful Knowledge
  - http://www.cs.uic.edu/~kzhao/Papers/06_ICDM_Zhao_Visual.pdf
- Lots of other sources …

Visualization & related fields

- User Interface Studies
- Computer Vision
- Cognitive Science
- Computer-Aided Design
- Geometric Modelling
- Computer Graphics
- Signal Processing
- Approximation Theory

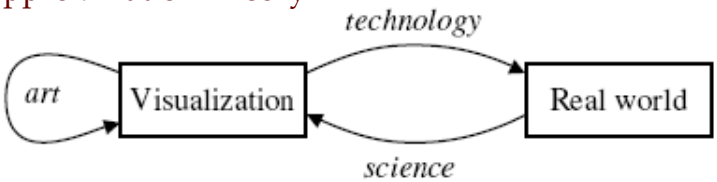


Fig. by Jack van Wijk from "*Views on Visualization*"

Purpose of Visualization in DM/KDD

- Data Visualization
  - Qualitative overview of large and complex data sets
  - Data summarization and (interactive) exploration
  - Regions of interest identification
  - Support humans in looking for structures, features, patterns, trends, and anomalies in data
  - Perceptual capabilities of the human visual system
- Model (and its performance) Visualization
- Process Visualization
- Disadvantages?
  - requires human eyes
  - can be misleading

"The purpose of computing is insight, not numbers"
Richard Hamming

Visualization Techniques Categorization

- Based on task at hand
  - To explore data
  - To confirm a hypothesis
- Based on structure of underlying data set
- Based on the dimension of the display
  - Stationary
  - Animated
  - Interactive
- Geometric vs. Symbolic
  - Results of physical models, simulations, computations
  - Nonnumeric data
    - Networks, relations as graphs; icons, arrays
- 1D-2D-3D, stereoscopic, virtual reality vs. 3+D
- Static vs. Dynamic

Some Basic Visualization Terminology

- Visualization
- Interaction
- Graphical attributes
- Mapping
- Rendering
- Field
- Scalar, Vector, Tensor
- Isosurface
- Glyph
- …

Perception in Visualization

- What graphical primitives do humans detect quickly?
- What graphical attributes can be accurately measured?
- How many distinct values for a particular attribute can be used without confusion?
- How do we combine primitives to recognize complex phenomena?

- Human perception and information theory

The most of following slides for *data* visualization were adopted from the presentation "Visualization and Data Mining" by G. Piatetsky-Shapiro
(available from www.kdnuggets.com)

Data Visualization Outline

- Graphical excellence and lie factor
- Representing data in 1,2, and 3-D
- Representing data in 4+ dimensions
  - Parallel coordinates
  - Scatterplots
  - Stick figures

Napoleon Invasion of Russia, 1812

Napoleon

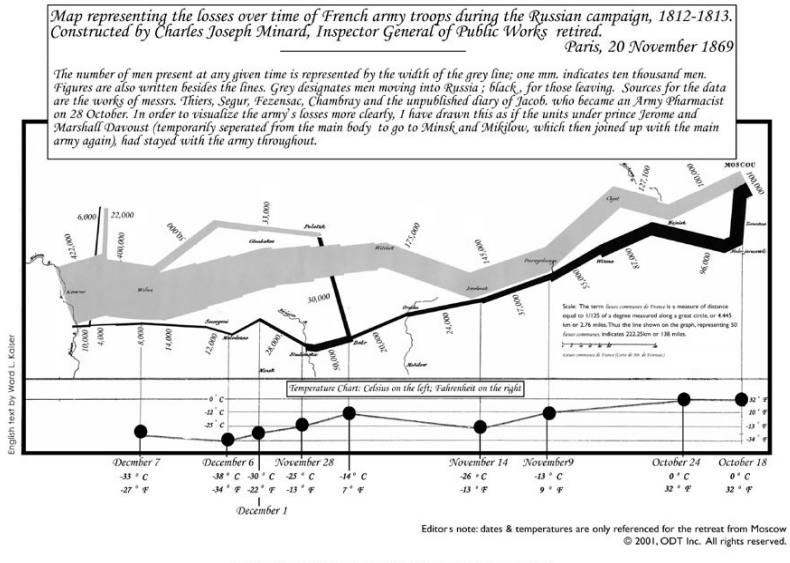
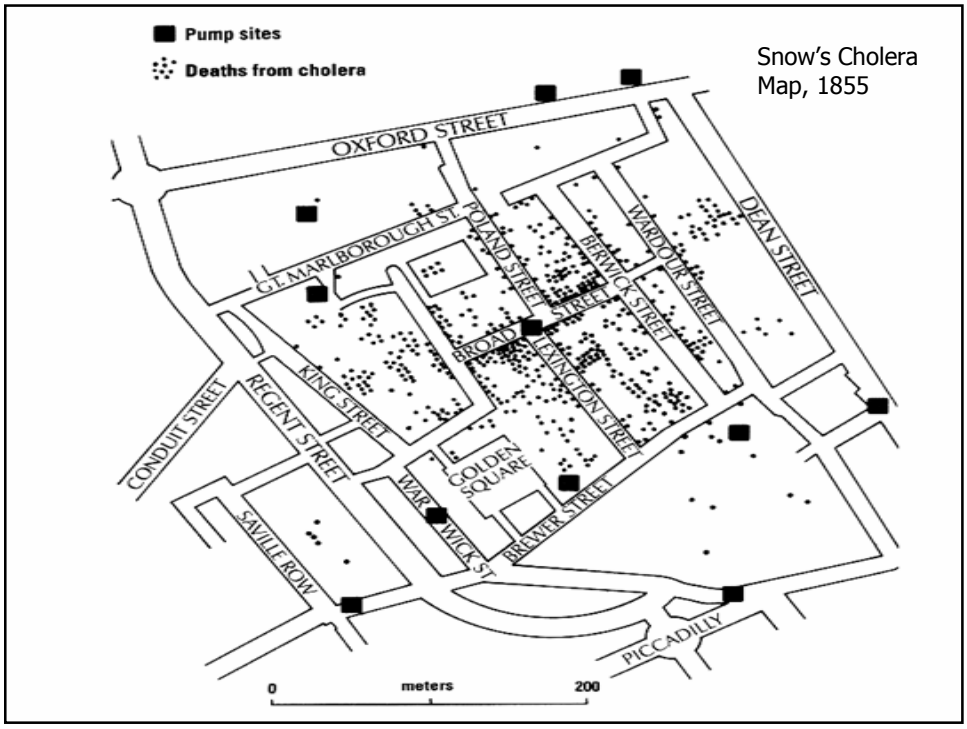
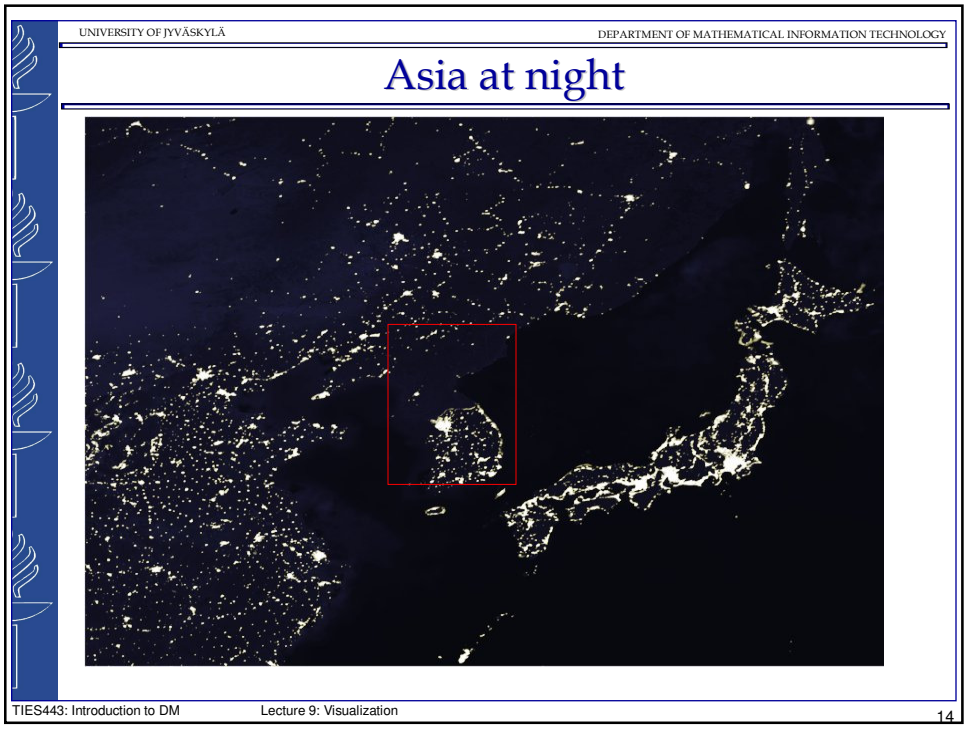


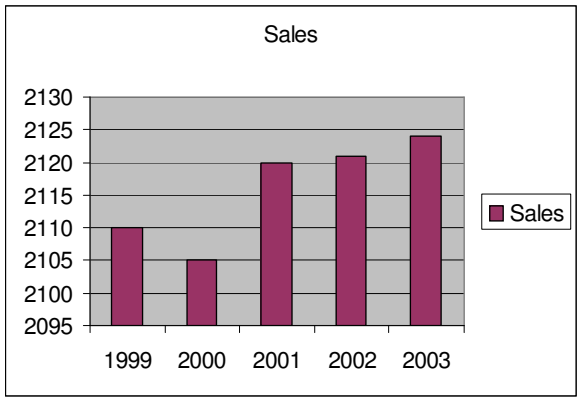
Figure 58. ***Minard s map of Napoleon s Russian campaign.***
This graphic has been translated from French to English and modified to most effectively display the temperature data.


Pump sites
Deaths from cholera
Snow's Cholera Map, 1855
OXFORD STREET
G.E.MARLBOROUGH ST.
POLAND STREET
BERWICK STREET
WARDOUR STREET
DEAN STREET
BROAD STREET
LEXINGTON STREET
KING STREET
REGENT STREET
CONDUIT STREET
GOLDEN SQUARE
WARWICK ST.
BREWER STREET
SAVILLE ROW
PICCADILLY
0
meters
200


UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL INFORMATION TECHNOLOGY
Asia at night
TIES443: Introduction to DM
Lecture 9: Visualization
14

Bad Visualization: misleading Y -axis

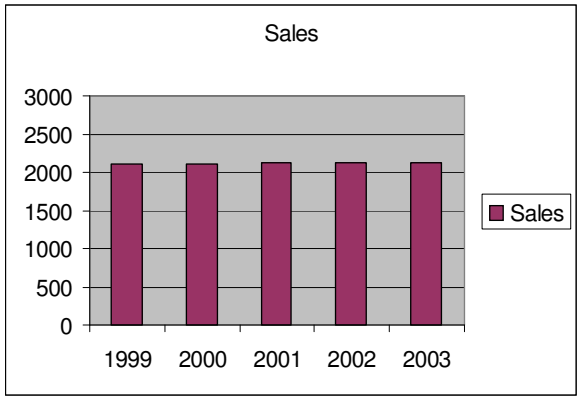
Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124



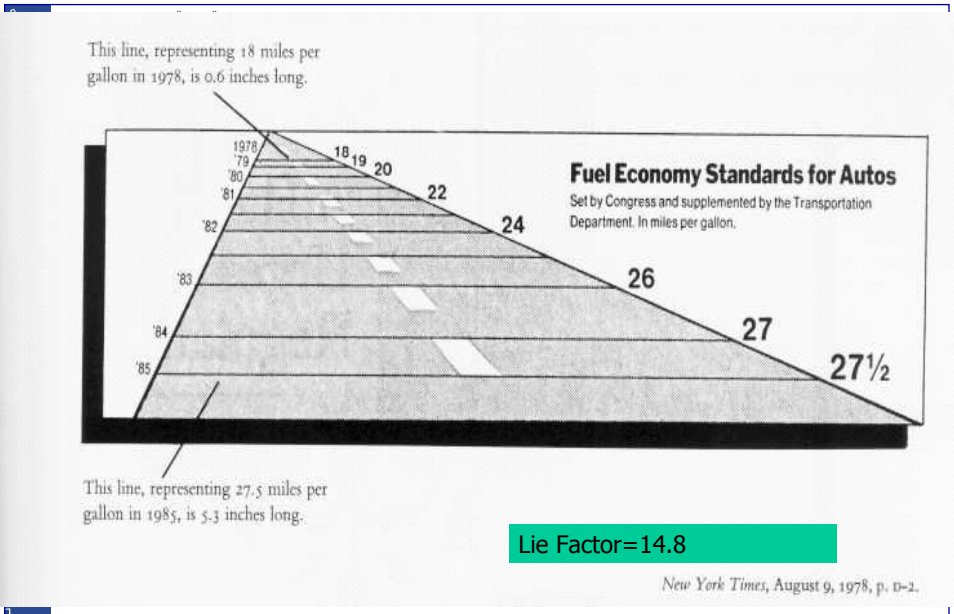
Y-Axis scale gives **WRONG** impression of big change

Better Visualization

Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124



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



(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)

Lie Factor

$$Lie\ Factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data} =$$

$$= \frac{\frac{(5.3-0.6)}{0.6}}{\frac{(27.5-18.0)}{18}} = \frac{7.833}{0.528} = 14.8$$

Tufte requirement: 0.95<Lie Factor<1.05

(E.R. Tufte, "The Visual Display of Quantitative Information", 2nd edition)

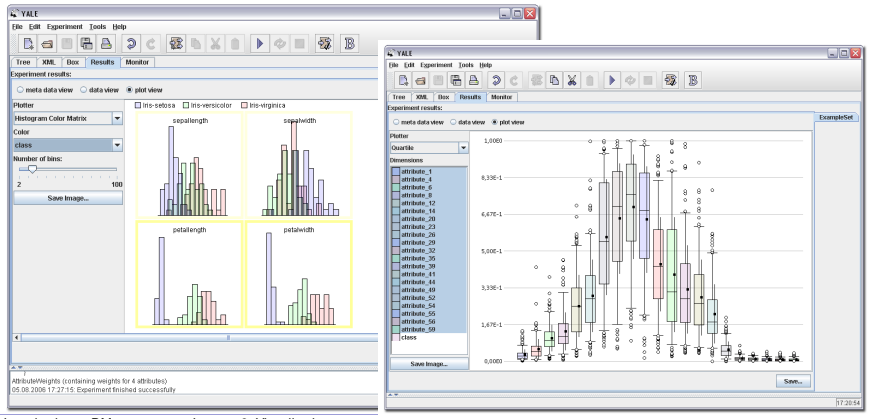
Tufte's Principles of Graphical Excellence

- 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)

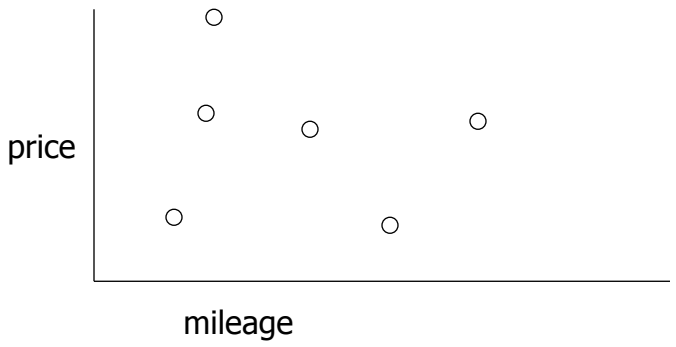
1-D (Univariate) Data

- Representations

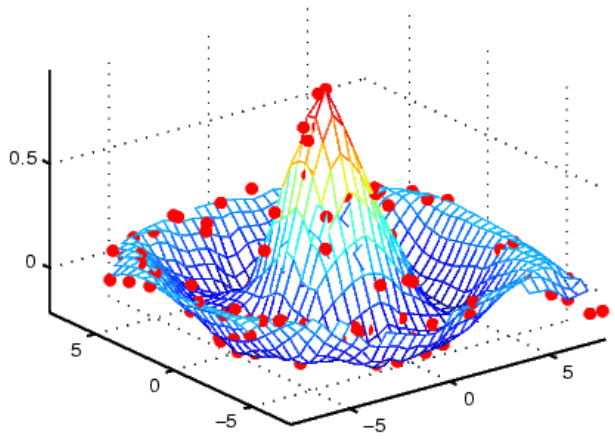


2-D (Bivariate) Data

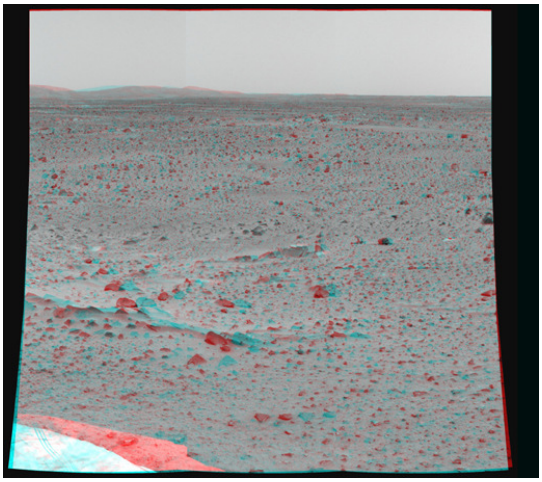
- Scatter plot, …



3-D Data (projection)



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



Taken by Mars Rover Spirit, Jan 2004

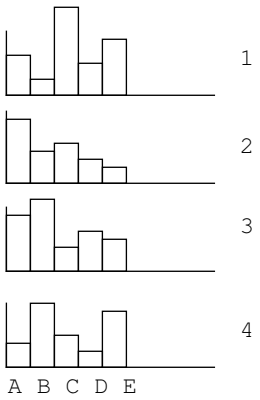
Visualizing in 4+ Dimensions

- Scatterplots
- Parallel coordinates
- Chernoff faces
- Stick figures
- Star glyphs
- …

Multiple Views

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

Scatterplot Matrix

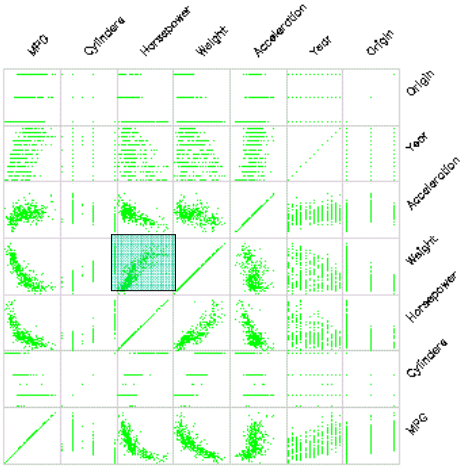
Represent each possible pair of variables in their own 2-D scatterplot (car data)

Q*: *Useful for what?

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

Q: Misses what?

A: multivariate effects



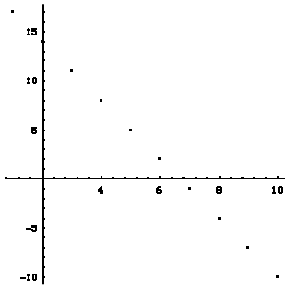
For 13 dimensions (columns) :

Number of histograms = 13

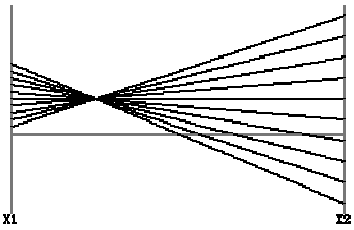
Number of scatterplots = C(13,2) = 78

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

Example: Visualizing Iris Data

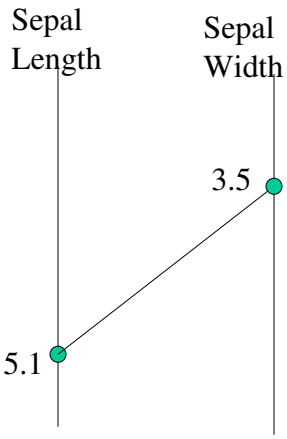
Iris setosa

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
…	…	…	…
5.9	3	5.1	1.8

Iris versicolor

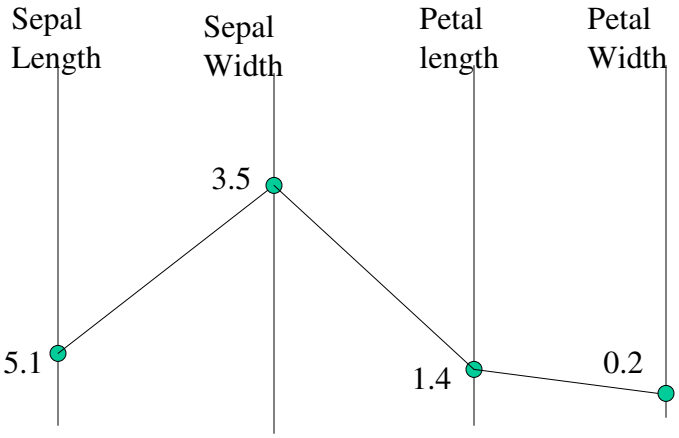
Iris virginica

Parallel Coordinates: 2 D



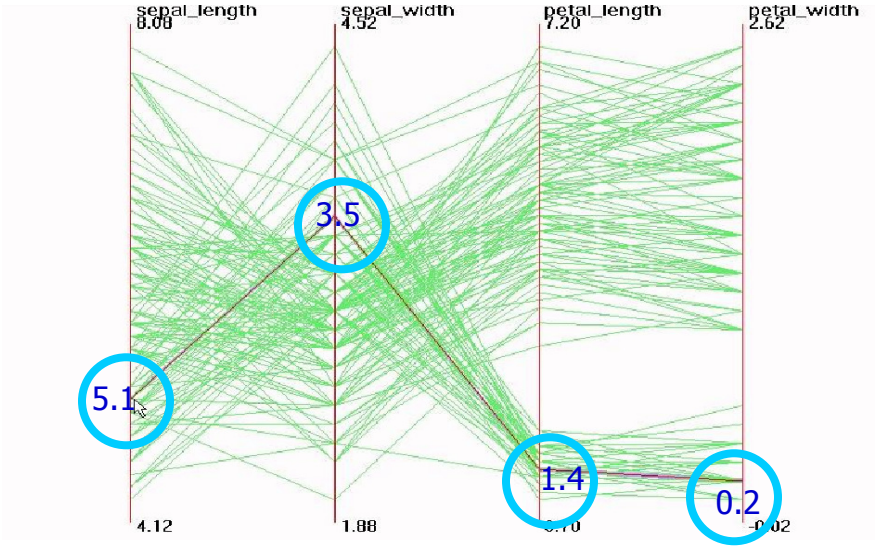
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 4 D



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

Parallel Visualization of Iris data



Parallel Coordinates: Correllation

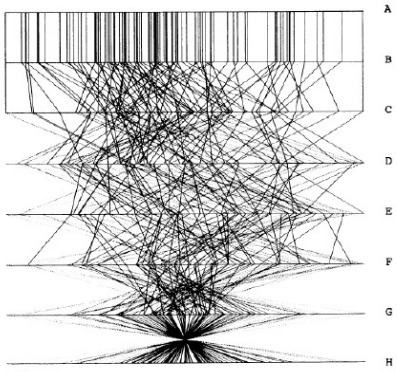
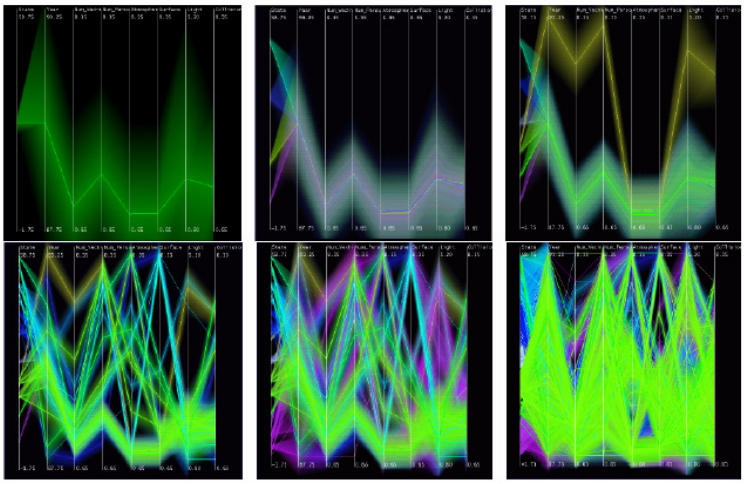


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of ρ = 1, .8, .2, 0, −.2, −.8, and −1.

Hyperdimensional Data Analysis Using Parallel Coordinates. Edward J. Wegman. Journal of the American Statistical Association, 85(411), Sep 1990, p 664-675.

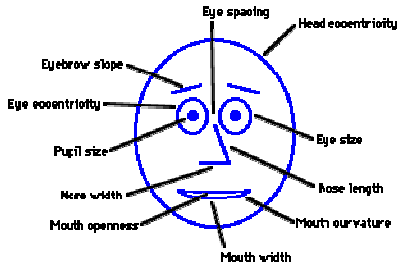
Hierarchical Parallel Coordinates



Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. Fua, Ward, and Rundensteiner, IEEE Visualization 99

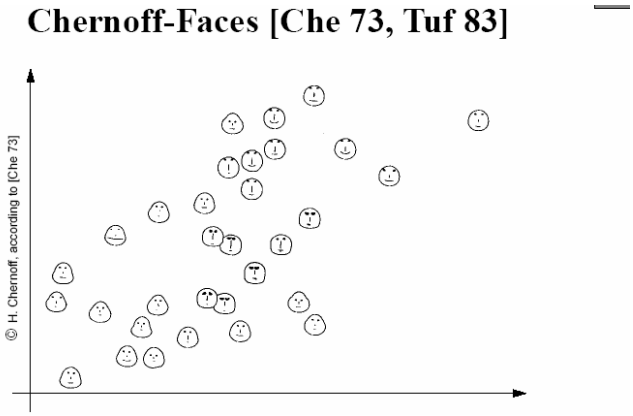
Chernoff Faces

Encode different variables' values in characteristics of human face



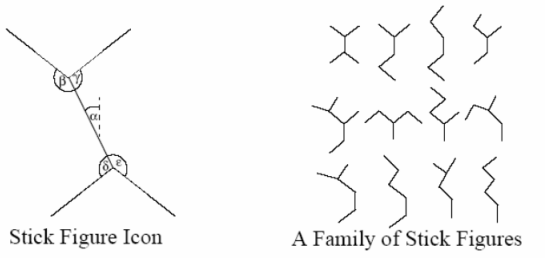
Cute applets: http://www.cs.uchicago.edu/~wiseman/chernoff/
http://hesketh.com/schampeo/projects/Faces/chernoff.html

Chernoff faces, example

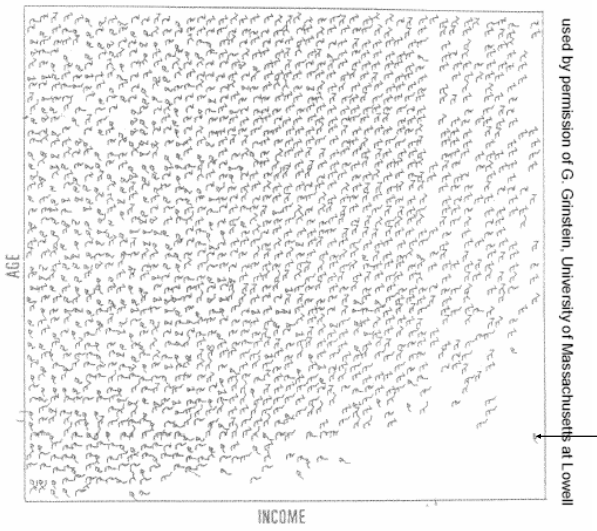


Stick Figures

- Two variables are mapped to X, Y axes
- Other variables are mapped to limb lengths and angles
- Texture patterns can show data characteristics



Stick figures, example



used by permission of G. Grinstein, University of Massachusetts at Lowell

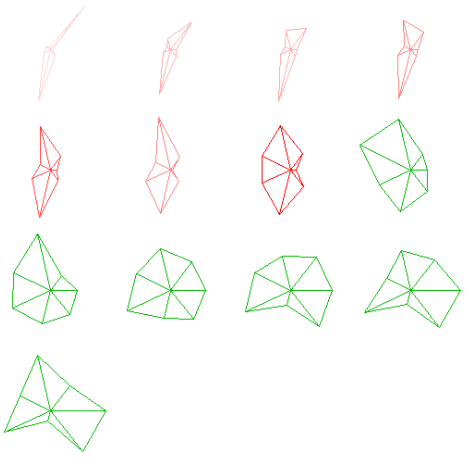
census data showing age, income, sex, education, etc.

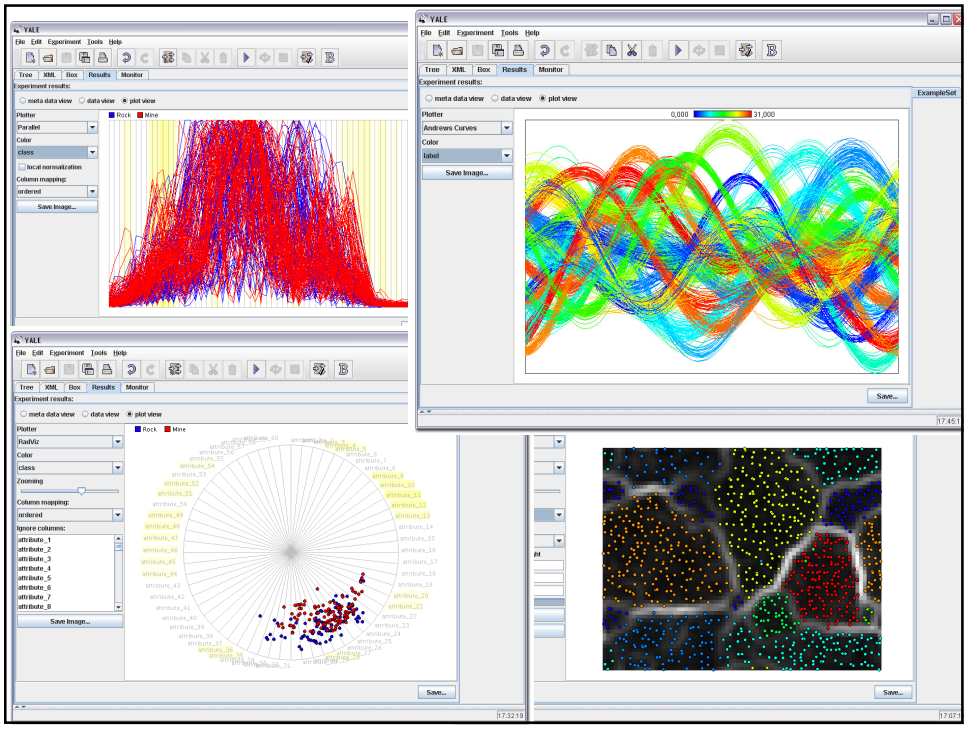
Closed figures correspond to women and we can see more of them on the left.

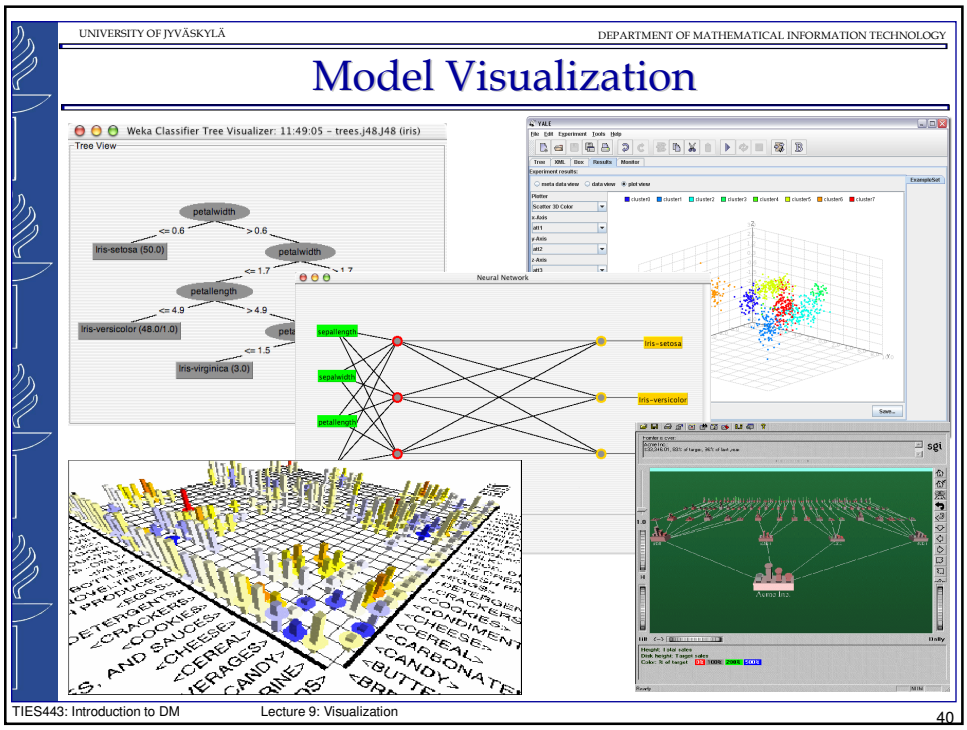
Note also a young woman with high income

Star Glyphs from Xmdv

- Two variables are mapped to X, Y axes
- Other variables are mapped to limb lengths and angles
- Texture patterns can show data characteristics



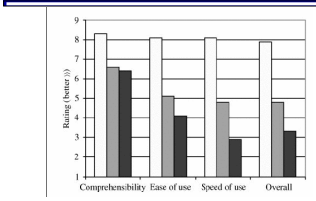


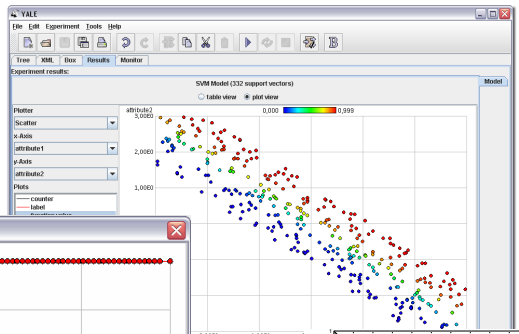

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL INFORMATION TECHNOLOGY
Model Visualization
Weka Classifier Tree Visualizer: 11:49:05 - trees.j48.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
petallength
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
<= 1.5
Iris-virginica (3.0)
Neural Network
sepallength
sepalwidth
petallength
Iris-setosa
Iris-versicolor
TIES443: Introduction to DM
Lecture 9: Visualization
40

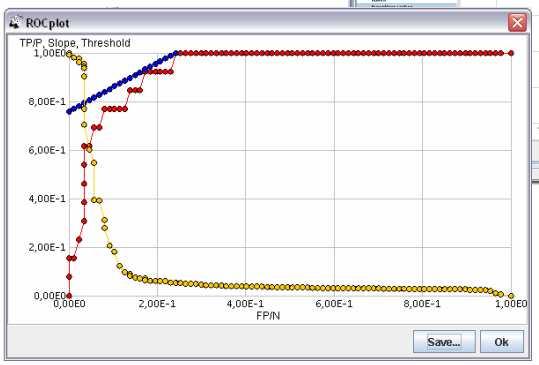
DM Model Visualization: Why

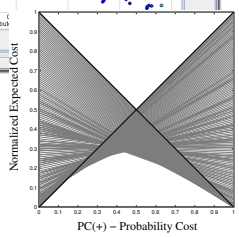
- Understanding the model
  - allow a user to discuss and explain the logic behind the model
    - also with colleagues, customers, and other users
  - more than just comprehension; it also involves business context
    - financial indicators like profit and cost
  - visualization of the DM output in a meaningful way
  - allowing the user to interact with the visualization
    - simple "what if" questions can be answered
  - 3 whales: representation, interaction, and integration
- Trusting the model
  - good quantitative measures of "trust"
  - the overall goal of "visualizing trust" – understanding the limitations of the model
- Comparing Different Models using Visualization
  - as Input-Output Mappings, as Algorithms, as Processes

Model Visualization: Performance Evaluation



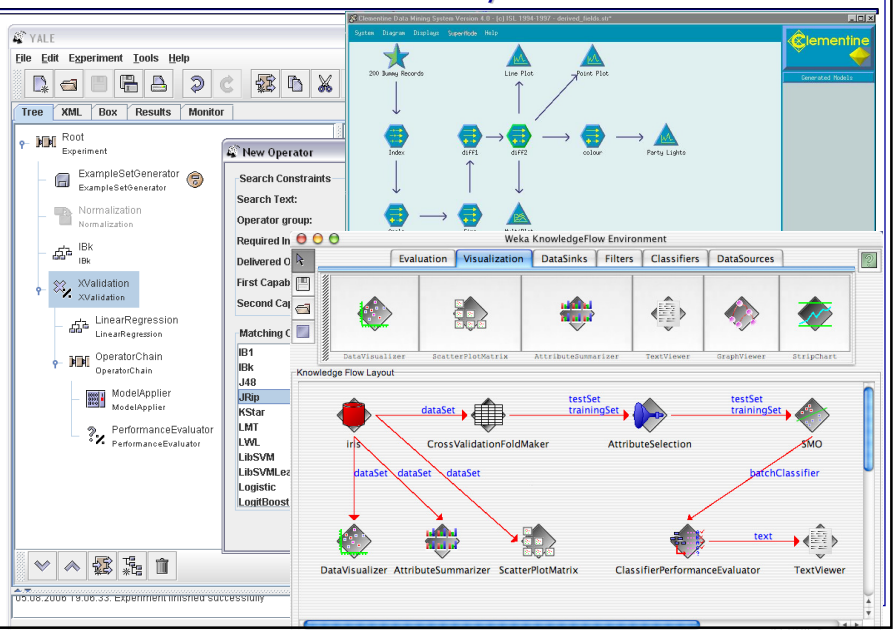






Cost-curves (Holt)

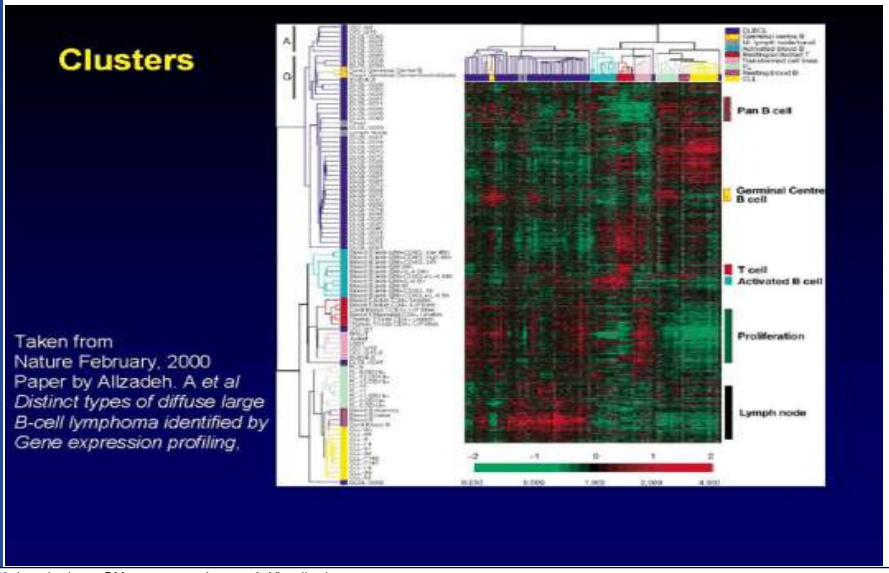
Visualization of DM/KDD Process



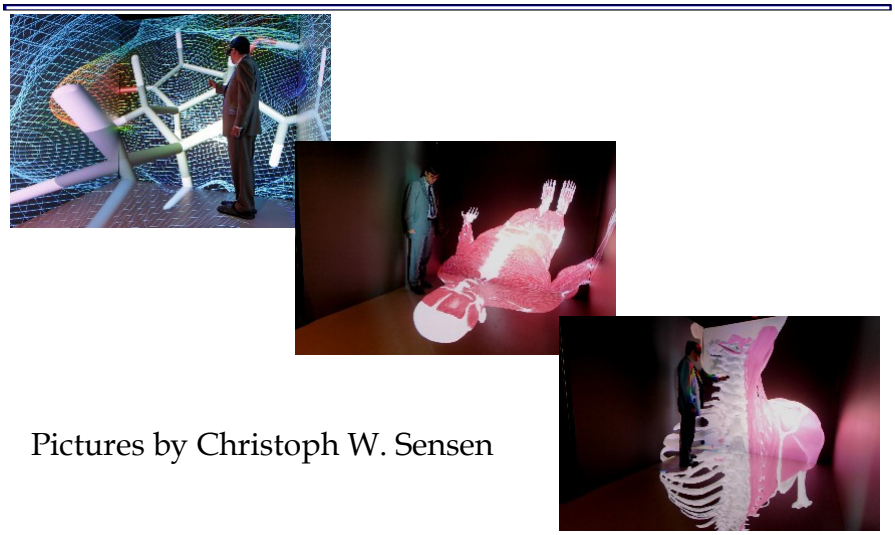
Applications …

- Bioinformatics
- Visual genomics
- Social networks
- Process mining
- Time series visualization

Visualization of Clusters of Microarray Data



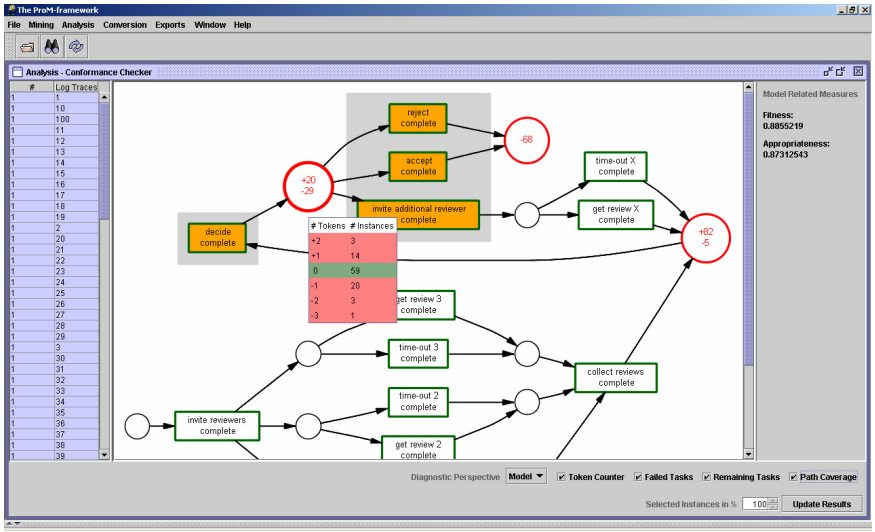
Visual Genomics



Pictures by Christoph W. Sensen

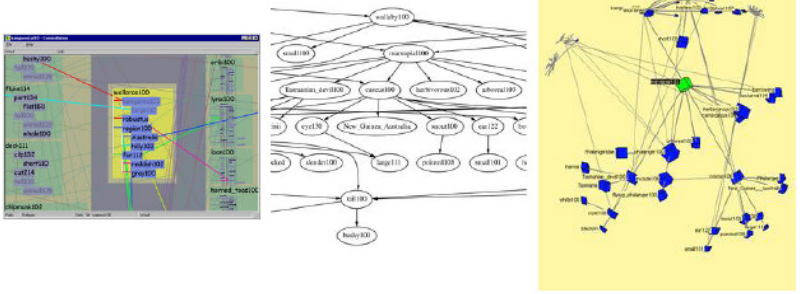
www.visualgenomics.ca/index.php?option=com_content&task=section&id=15&Itemid=143

The Result of Process Mining

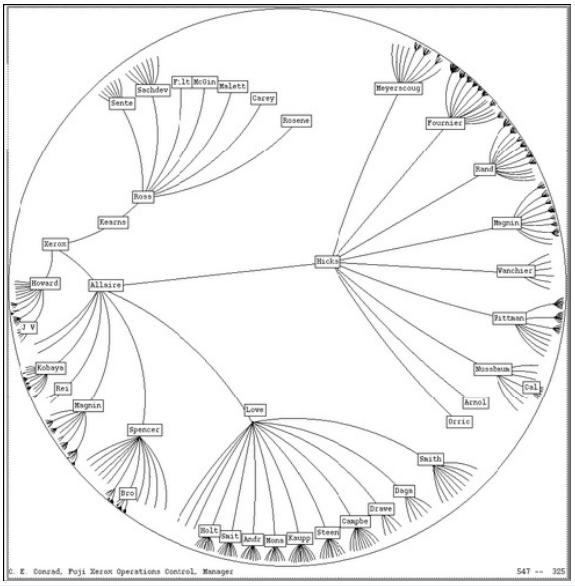


From presentation by Wil van der Aalst, TU/e www.processmining.org

Design Studies



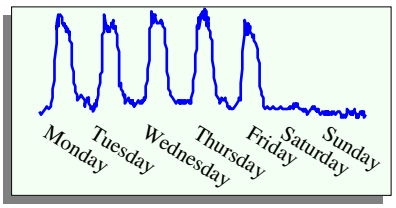
Focus+Context

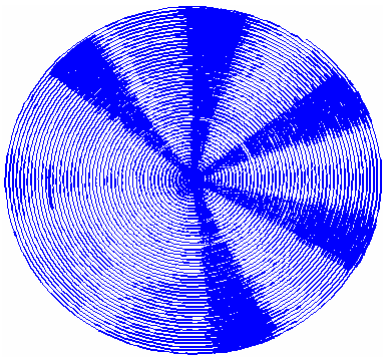


Time Series Spirals

- **Spiral Axis** = serial attributes are encoded as line thickness
- **Radii** = periodic attributes

One year of power demand data

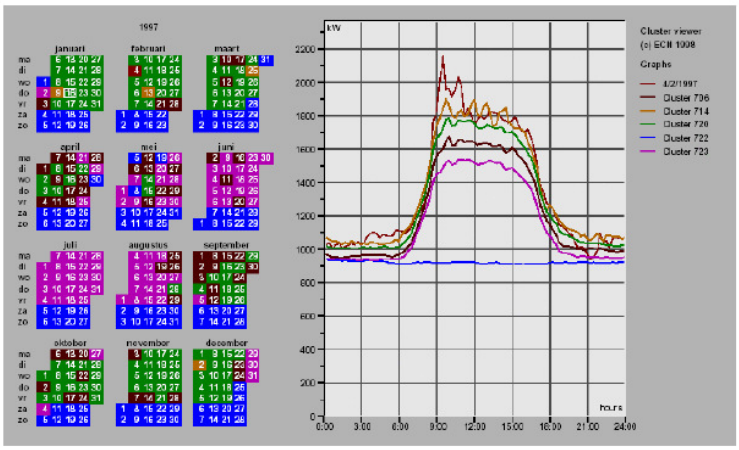




Carlis & Konstan. UIST-98
Independently rediscovered by
Weber, Alexa & Müller InfoVis-01
But dates back to 1888!

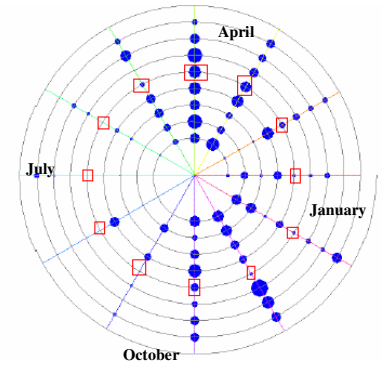
(c) Eamonn Keogh, eamonn@cs.ucr.edu

Power Consumption



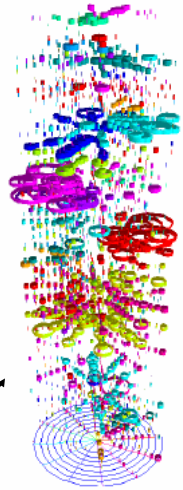
van Wijk and van Selow, Cluster and Calender based Visualization of Time Series Data, InfoVis99, http://www.win.tue.nl/~vanwijk/clv.pdf

Time Series Spirals

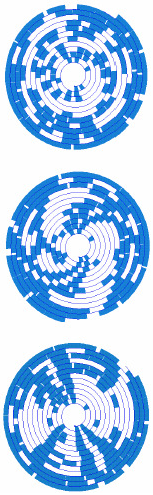


Chimpanzees Monthly Food Intake
1980-1988

The spokes are months, and spiral guide lines are years



112 types of food

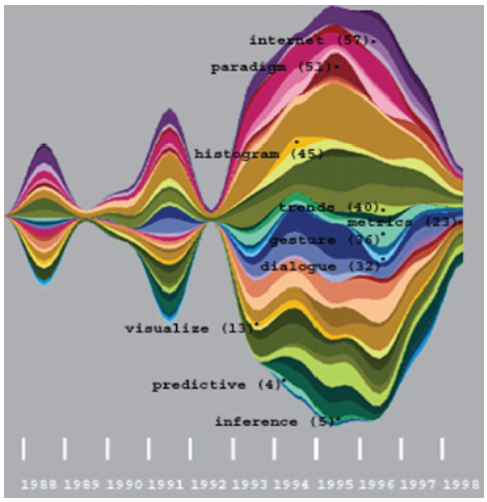


Simple and intuitive, but works only with periodic data, and if the period is known

(c) Eamonn Keogh, eamonn@cs.ucr.edu

ThemeRiver

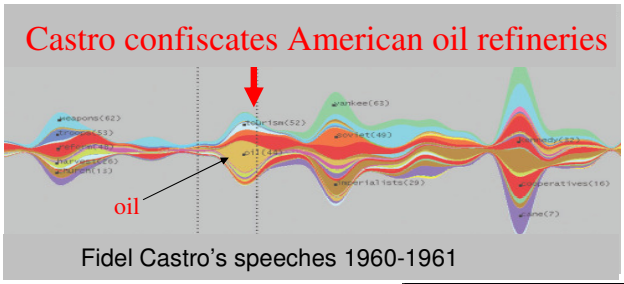
- Current width = strength of theme
- River width = global strength
- Color mapping (similar themes/ same color family)
- Time axis
- External events can be linked



Havre, Hetzler, Whitney & Nowell
InfoVis 2000

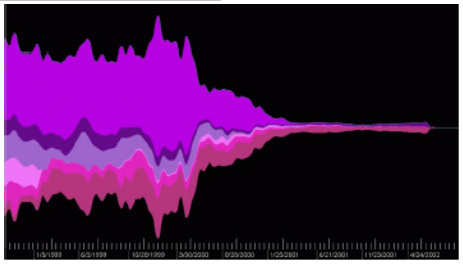
A company's patent activity
1988 to 1998

ThemeRiver



Fidel Castro's speeches 1960-1961

- Simple and intuitive
- Many extensions possible
- Scalability is still an issue



dot.com stocks 1999-2002

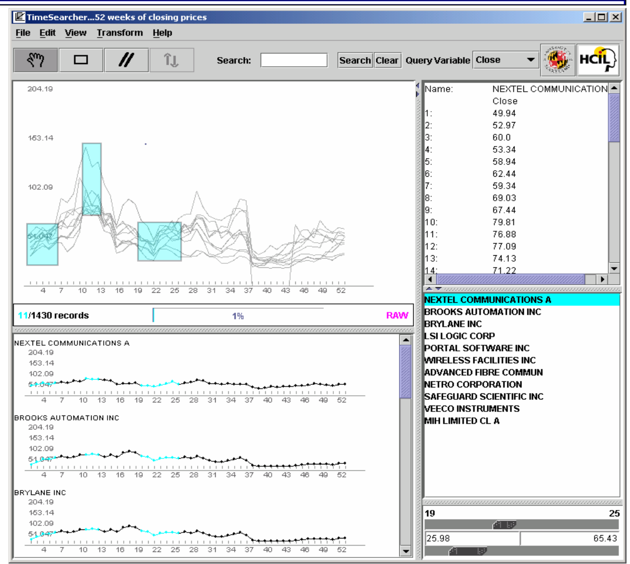
TimeSearcher

Comments

- Simple and intuitive
- Highly dynamic exploration

- Query power may be limited and simplistic
- Limited scalability

Hochheiser, and Shneiderman



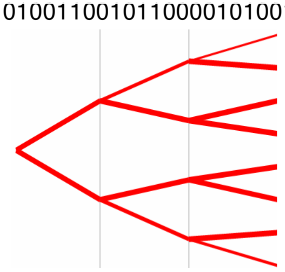
(c) Eamonn Keogh, eamonn@cs.ucr.edu

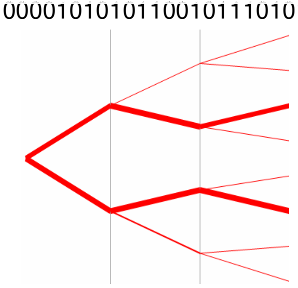
VizTree

0101100101111001101001000010 0 1000100010100100010101010000 1
0101001101101011100001010101 1 0101000101011101111010110100 1
1011111000110110110111111010 0 0111010010101001110101010100 1

Which set of bit strings is generated by a human and which one is generated by a computer?

1110100110010110000101001 0 1110000101010110010111010 1





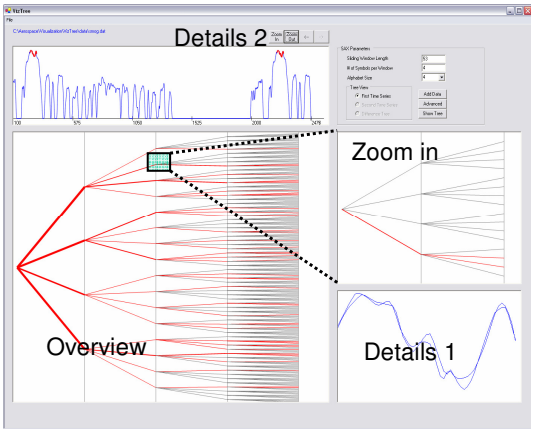
the frequencies of all triplets are encoded in the thickness of branches …
"humans usually try to fake randomness by alternating patterns"

adopted from ex. by Eamonn Keogh

VizTree

The "trick" on the previous slide only works for discrete data, but time series are *real* valued.

we will discuss how we can SAX up a time series to make it discrete during the next tutorial on time series mining



"Overview, zoom & filter, details on demand"
– main principles of visualization Ben Shneiderman

Visualization Summary

- Purpose of visualization in DM/KDD
- Visualization of data - Many methods
  - 1D, 2D, 3D
  - Visualization is possible in more than 3-D
- Visualization of data mining results
  - Model visualization, model performance visualization
- Visualization of data mining processes
- Application …
  - Interactive and integrative data mining and visualization
  - Time-series, symbolic, networks visualization

One picture may worth 1000 words!

What else did you get from this lecture?

Additional Slides

Visualization Software

Free and Open-source

- Ggobi
- Xmdv
- Some techniques are available in WEKA and YALE as you have seen already
- Many more - see www.kdnuggets.com/software/visualization.html

Matlab has also many nice integrated tools for viz.