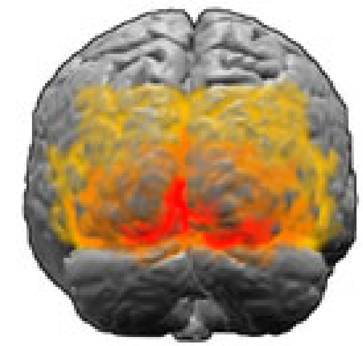
Application of data visualisation in particle physics

Steve Watts
BITlab
School of Engineering and Design
Brunel University, West London, UK

University of Manchester – from Jan '07



NOTE: "The visual cortex is the most massive system in the human brain and is responsible for higher-level processing of the visual image. It lies at the rear of the brain (highlighted in the image), above the cerebellum"

There is more to data visualisation than histograms, scatterplots and x/y plots.

Edinburgh 8 November 2006

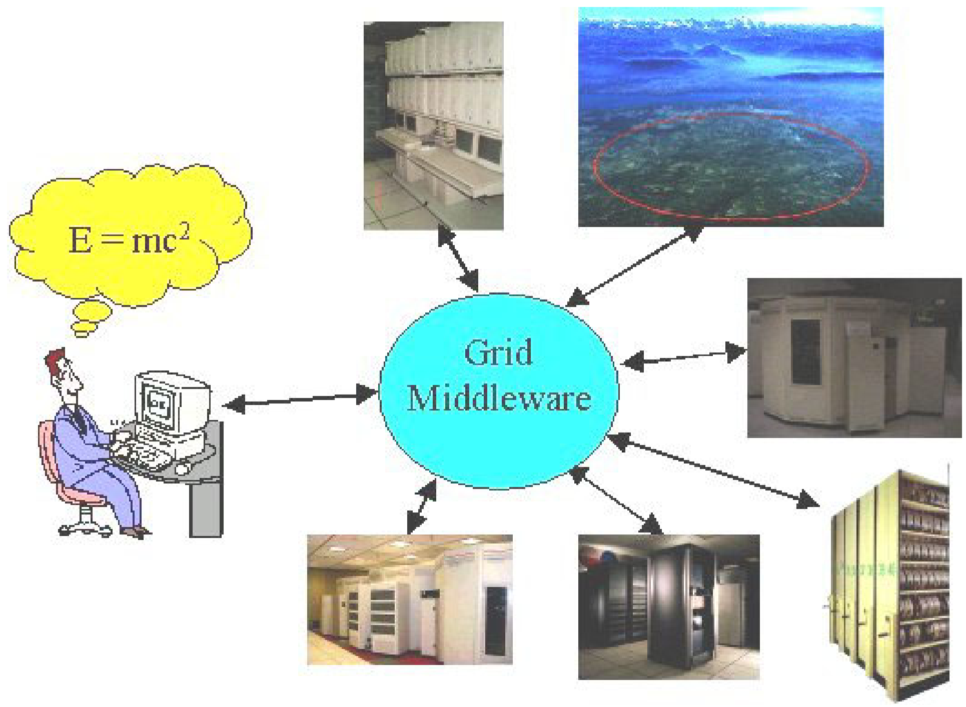
- WHAT SPARKED THE INTEREST ???
- INTRODUCTION TO DATA MINING
- DATA VISUALISATION

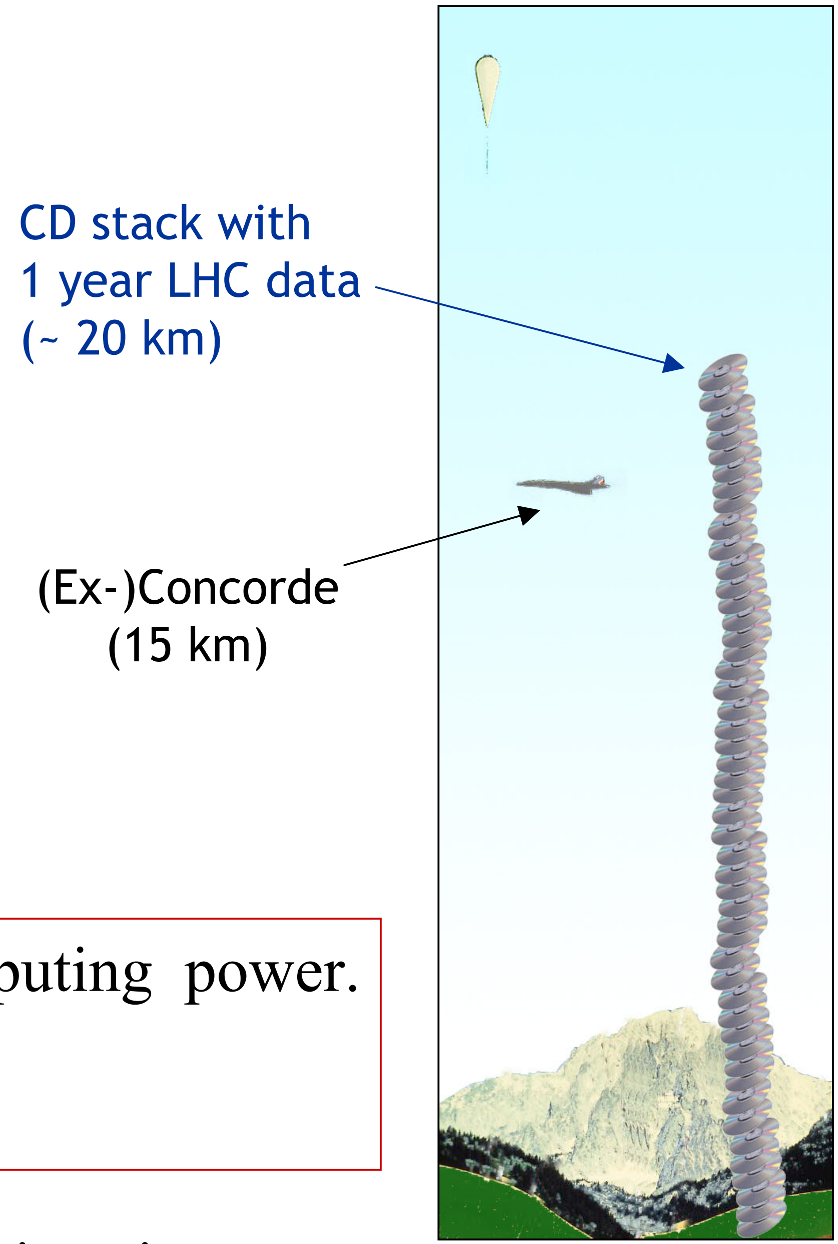
  (history, look at a PP dataset, parallel coordinates, brushing, pruning…)
- CLASSIFICATION ALGORITHMS

  Analysis of wine dataset (Decision Trees, SVM, kNN)

  Links to visualisation – e.g. GRAND Tour, Radviz, polyviz
- ANALYSIS of PP dataset with different techniques
- OTHER USEFUL VISUALISATION TECHNIQUES

  (survey plots, heat maps, mosaic dsiplay)
- DIY (Do it yourself) – some advice.
- CONCLUSIONS

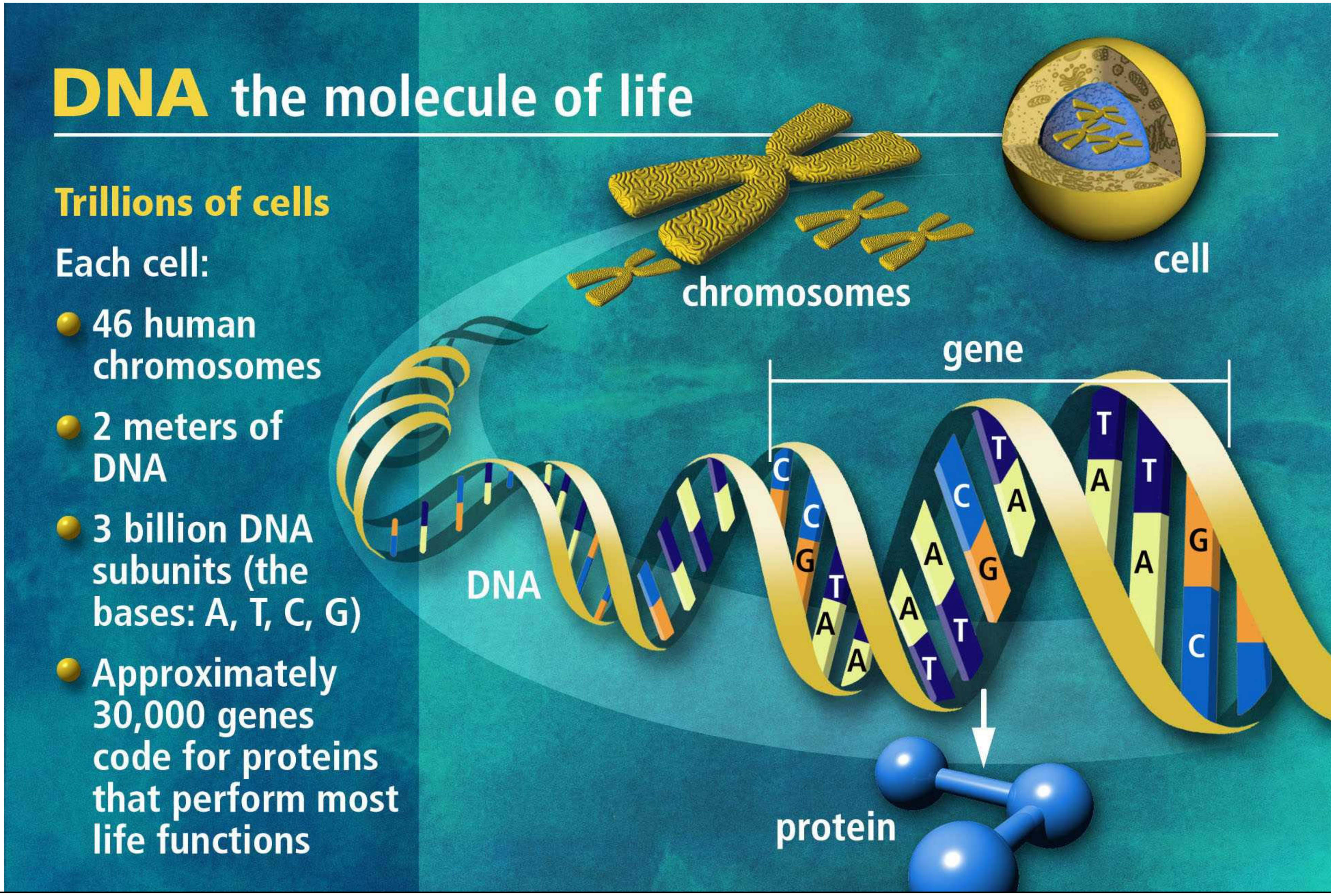




GRID computing gives massive computing power.
No excuse for not being smarter with
data analysis techniques

Thanks to Steve Lloyd for the nice pictures…..

MANY AREAS OF MODERN SCIENCE, ENGINEERING, HUMANITIES AND ARTS HAVE A DATA OCEAN TO SWIM IN !!!

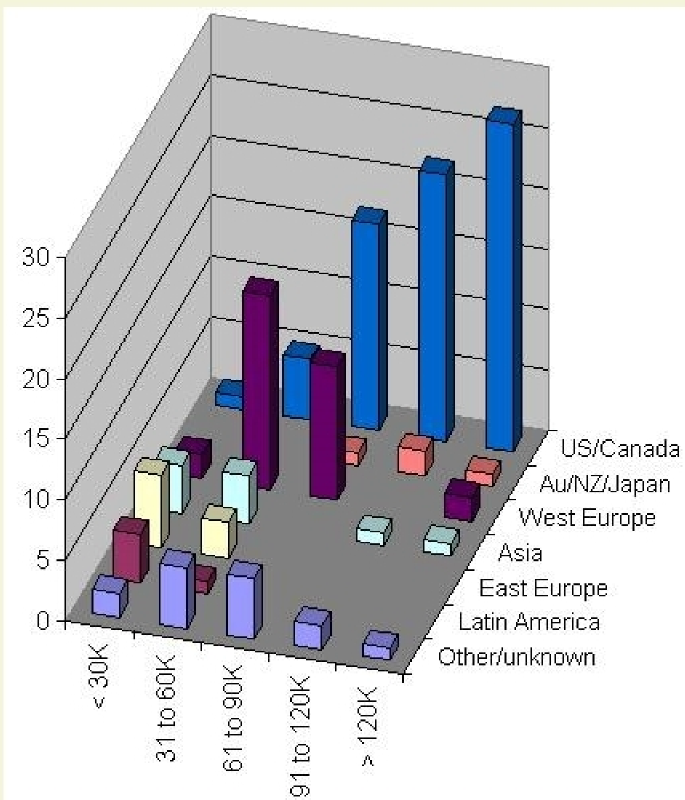


http://www.ornl.gov/sci/techresources/Human_Genome/graphics/slides/images1.shtml

Data mining, also known as knowledge-discovery in databases (KDD), is the practice of automatically searching large stores of data for patterns. To do this, data mining uses computational techniques from statistics and pattern recognition. **en.wikipedia.org/wiki/Datamining**

http://www.kdnuggets.com/polls/2006/data_miner_income_by_region.htm

The following graph shows the breakdown of income (salary) by region, with number of respondents on the vertical axis (excluding students). (Note: the poll asks for income to include data miners who work for a company as well as self-employed).



The message is clear!

Figure 1.2 We are data rich, but information poor.

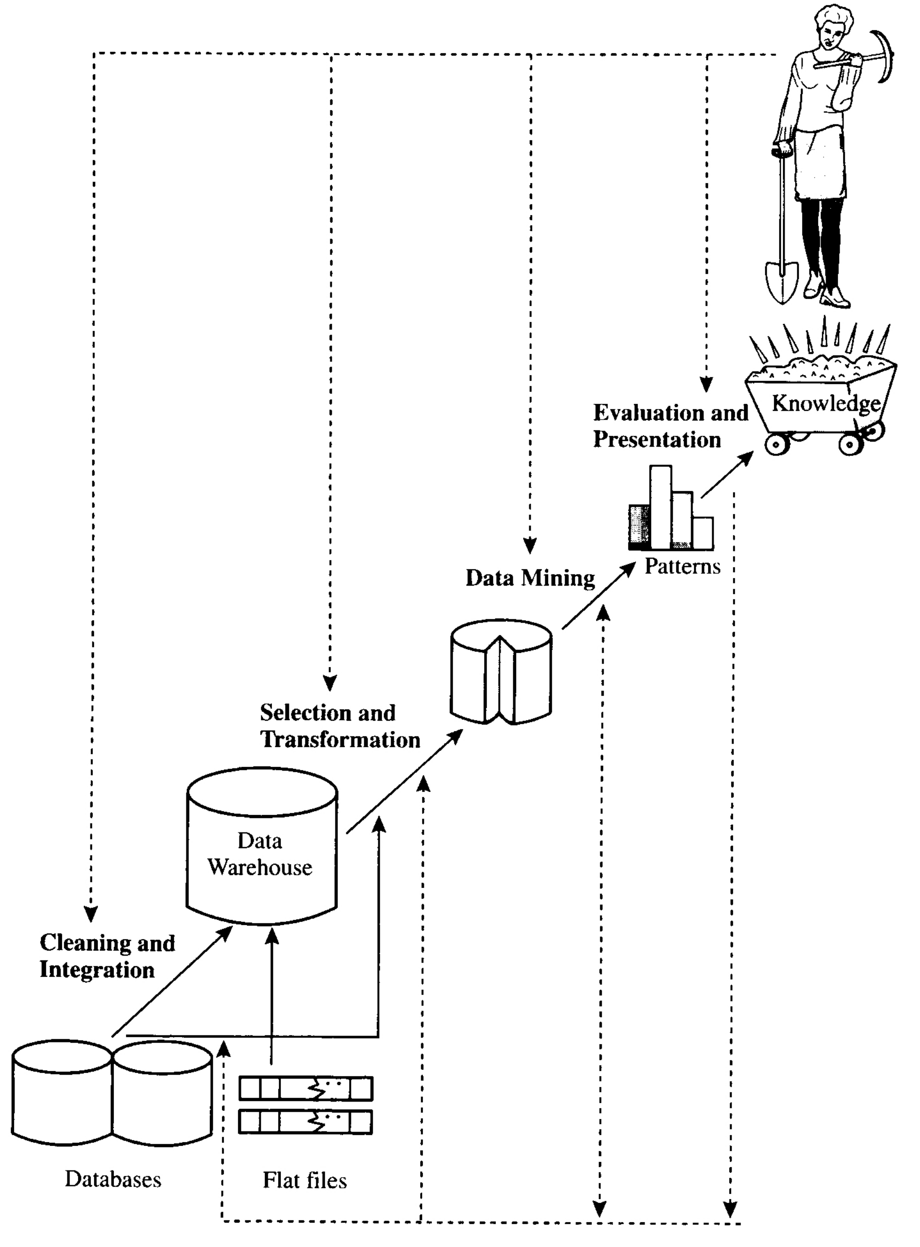


Figure 1.4 Data mining as a step in the process of knowledge discovery.

Data Mining:
Concepts and Techniques
2^{nd} ed. Jiawei Han
Micheline Kamber

“We are drowning in information and starving for knowledge”

Rutherford D Roger

“Information is not knowledge”
Albert Einstein

WEKA The University of Waikato

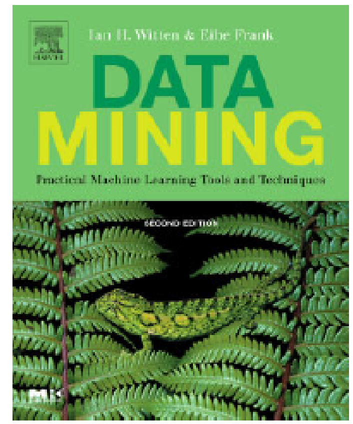
Book

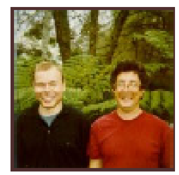
project - software - book - publications - people - related

Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)

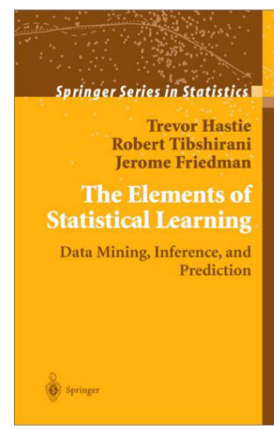
Ian H. Witten, Eibe Frank

Morgan
Kaufmann
June 2005
525 pages
Paper
ISBN
0-12-088407-0





Eibe Frank and Ian Witten

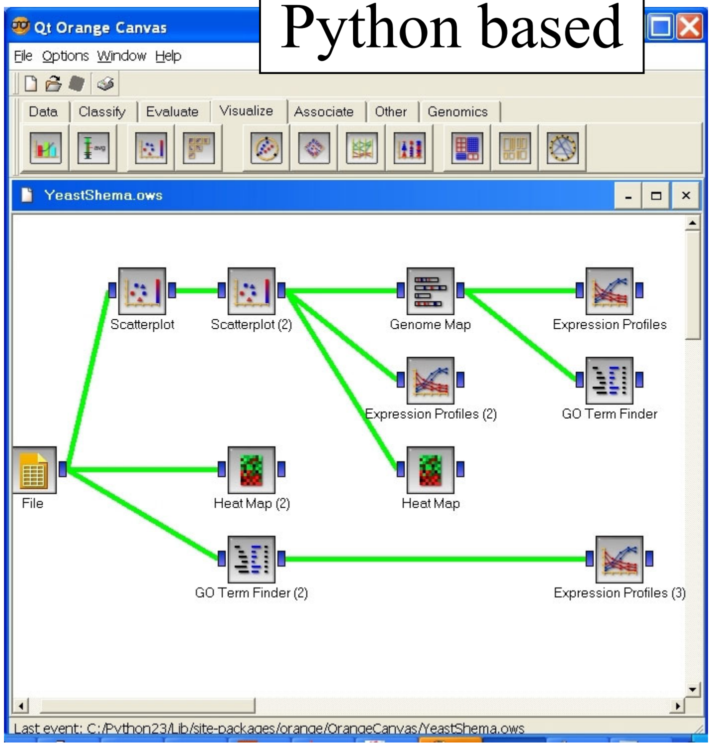


.com

Statistical Learning
KEY TEXT

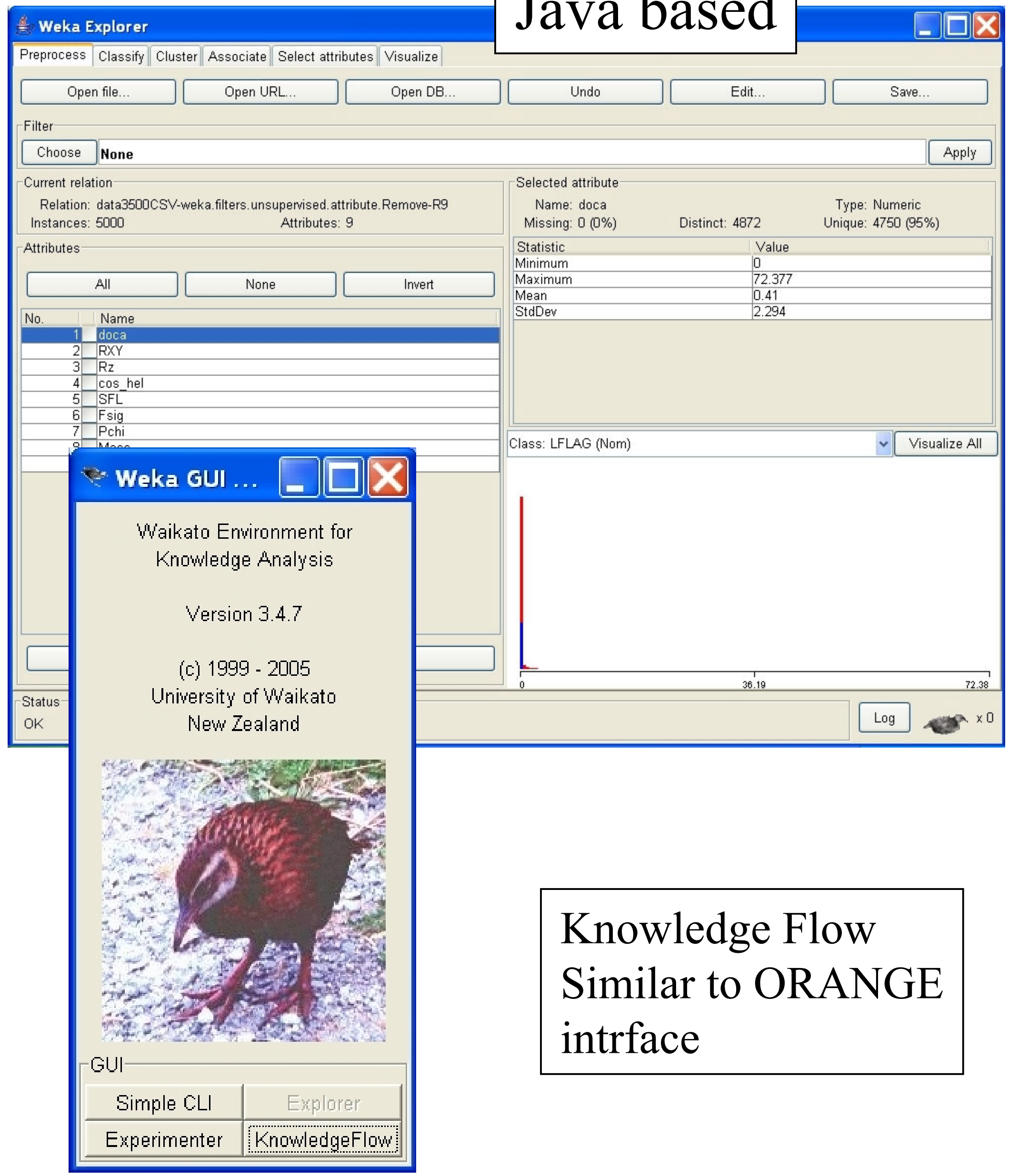
WEKA Machine Working Workbench
Waikato Environment for Knowledge Analysis

Python based



http://www.ailab.si/orange

Java based



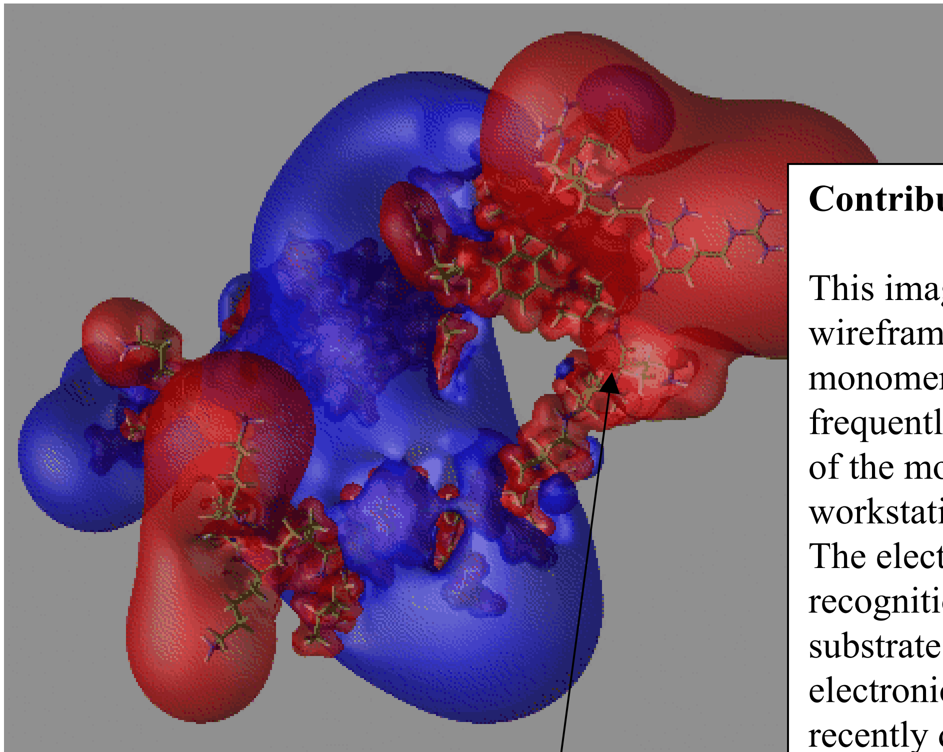
Knowledge Flow
Similar to ORANGE
intrface

KEY TERMS
Preprocess Data
Classify
Cluster
Associate
Visualisation

http://www.cs.waikato.ac.nz/~ml/weka/

Examples of scientific visualisation

http://www.msi.umn.edu/software/dx/tutorial/dx-images.html

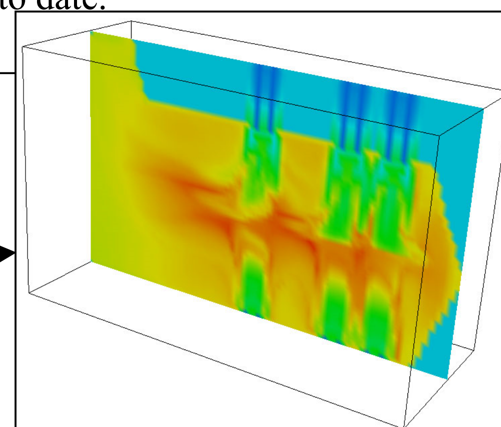


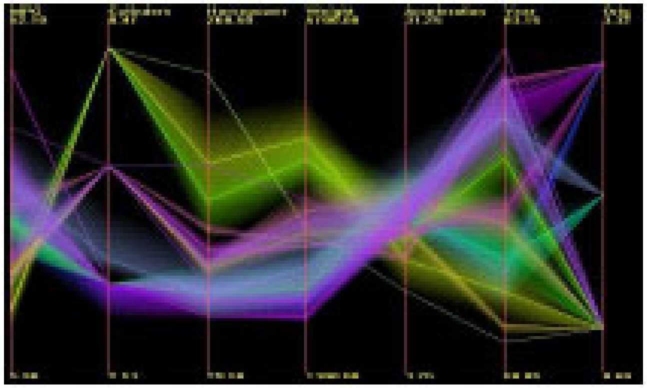
Contributors: Matt Challacombe and Eric Schwegler

This image shows a view of electrostatic potential iso-surfaces and a wireframe representation of the p53 tumor suppressor tetramerization monomer. Mutations in the p53 tumor suppressor are the most frequently observed genetic alterations in human cancer. The structure of the monomer's electrostatic potential has been rendered on an SGI workstation using iso-surfaces corresponding to -0.06 and +0.06 au. The electrostatic potential is widely implicated in molecular recognition, binding, and the enhanced diffusion of charged substrates.These results have been obtained from first principles electronic structure calculations using linear scaling Hatree-Fock theory recently developed at the University of Minnesota. Involving 3836 basis functions, this calculation was performed in 3 cpu days on an IBM RS6000 model 590 workstation, and is the largest Gaussian-based *ab initio* calculation performed to date.

AVS Express
Paraview - free !
Tecplot
IBM Data Explorer
VisIt - free

The **pseudocolor plot** (right) is used to map temperature to color on the same planar slice.





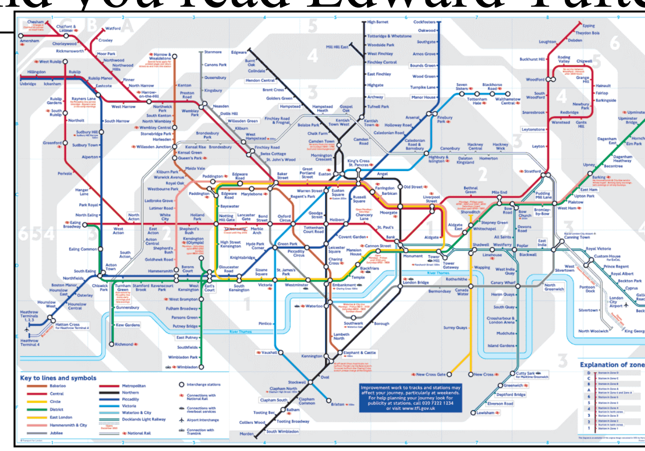
Information Visualization image shown courtesy of Matt Ward of Worcester Polytechnic Institute (WPI).

Information Visualisation

Displaying information to help the user understand it better. Abstraction of data.

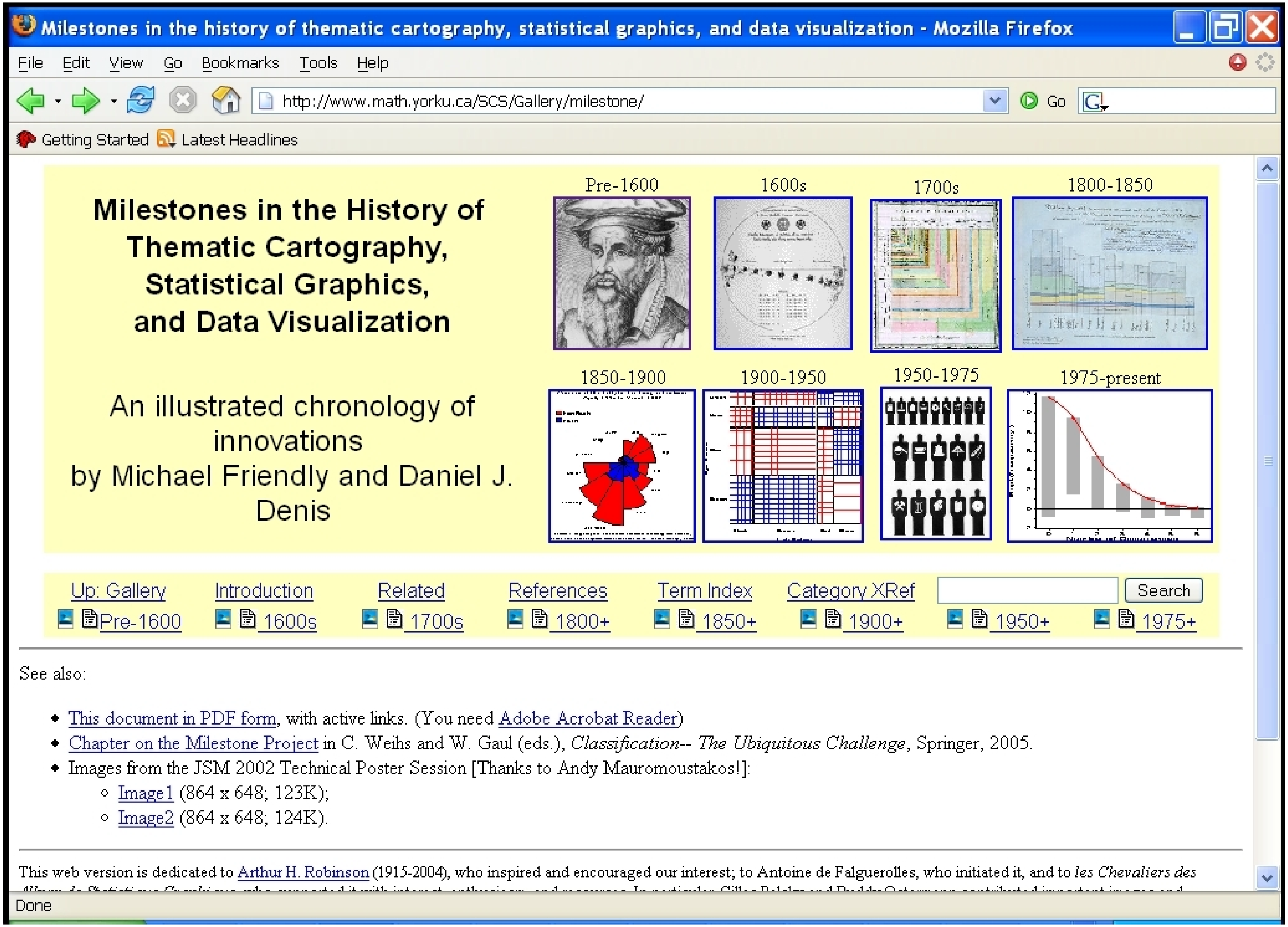
Example above I would categorise as **Data Visualisation**

The London Tube map I would categorise as **Information Visualisation** – recommend you read Edward Tufte

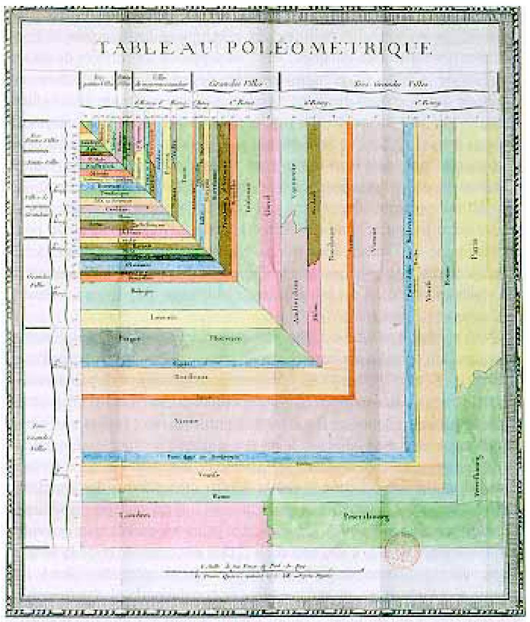


SciVis - late ‘80s
InfVis - late ‘90’s

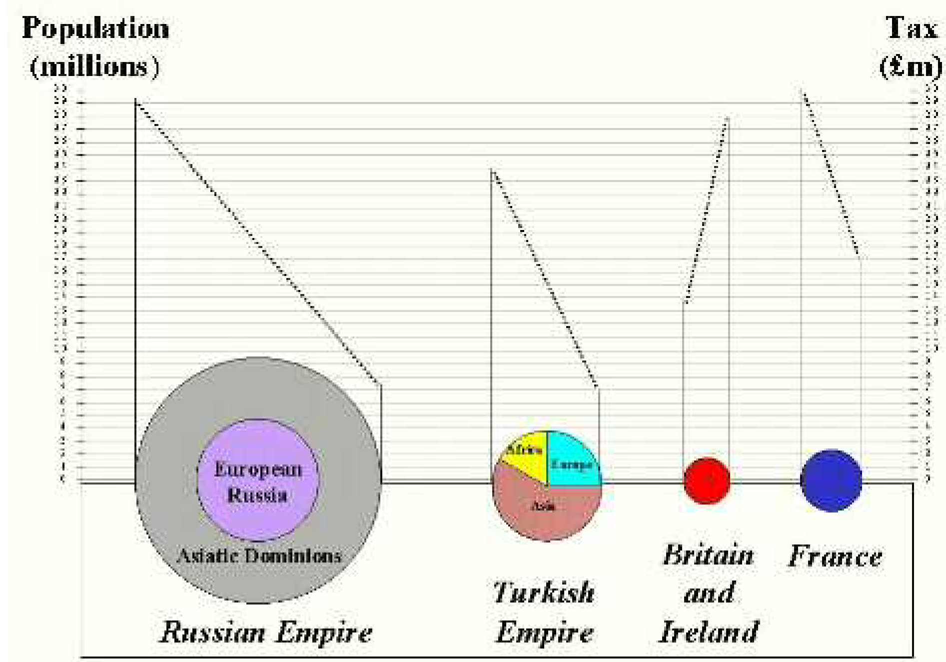
This is a vast new field - especially important for **data mining**


Milestones in the history of thematic cartography, statistical graphics, and data visualization - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.math.yorku.ca/SCS/Gallery/milestone/
Go
Getting Started
Latest Headlines
Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization
An illustrated chronology of innovations
by Michael Friendly and Daniel J. Denis
Pre-1600
1600s
1700s
1800-1850
1850-1900
1900-1950
1950-1975
1975-present
Up: Gallery
Introduction
Related
References
Term Index
Category XRef
Search
Pre-1600
1600s
1700s
1800+
1850+
1900+
1950+
1975+
See also:
This document in PDF form, with active links. (You need Adobe Acrobat Reader)
Chapter on the Milestone Project in C. Weihs and W. Gaul (eds.), Classification-- The Ubiquitous Challenge, Springer, 2005.
Images from the JSM 2002 Technical Poster Session [Thanks to Andy Mauromoustakos!]:
Image1 (864 x 648; 123K);
Image2 (864 x 648; 124K).
This web version is dedicated to Arthur H. Robinson (1915-2004), who inspired and encouraged our interest; to Antoine de Falguerolles, who initiated it, and to les Chevaliers des
Done

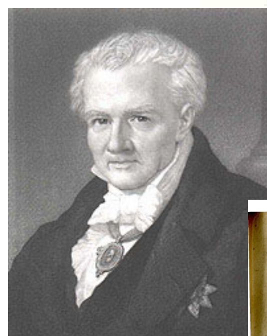
Luke Howard 1800
1st use of coordinate paper
in a research paper



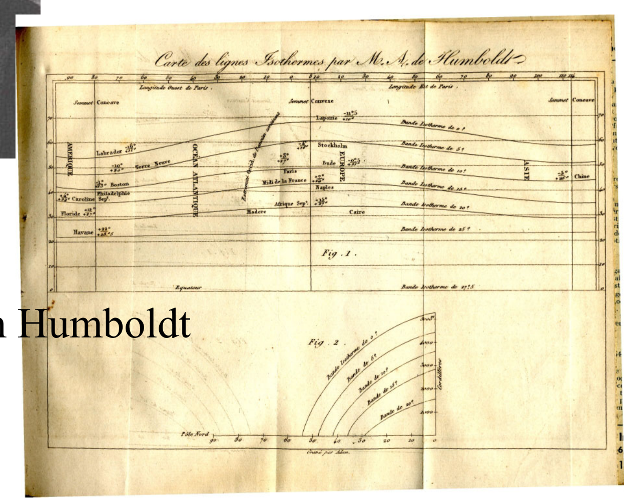
Charles de Fourcroy
1782
"proportional squares"



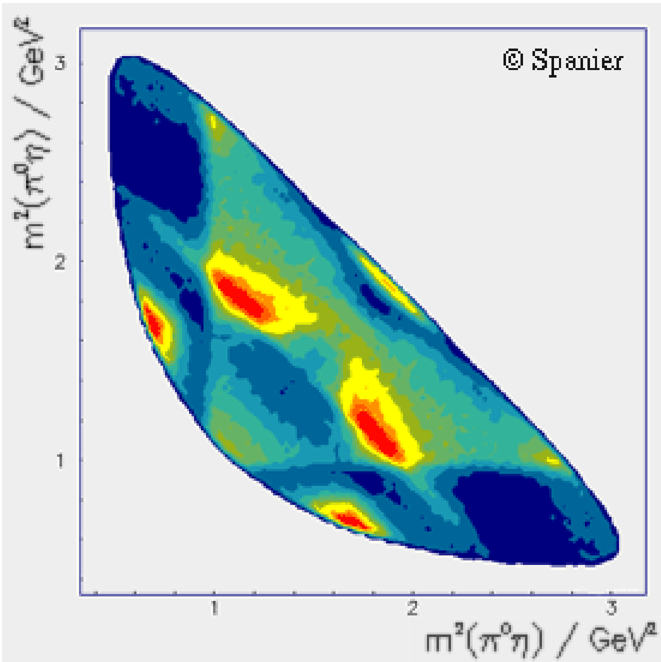
William Playfair's chart
Invented pie-chart. 1801



Alexander Von Humboldt



Humboldt's
isotherms
1817

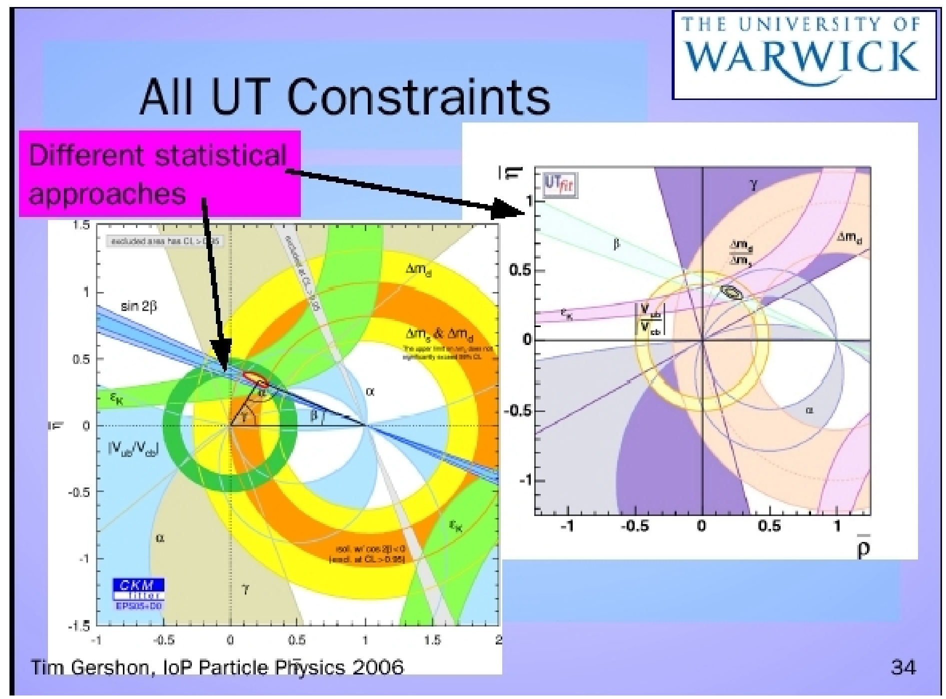


Dalitz Plot 1954

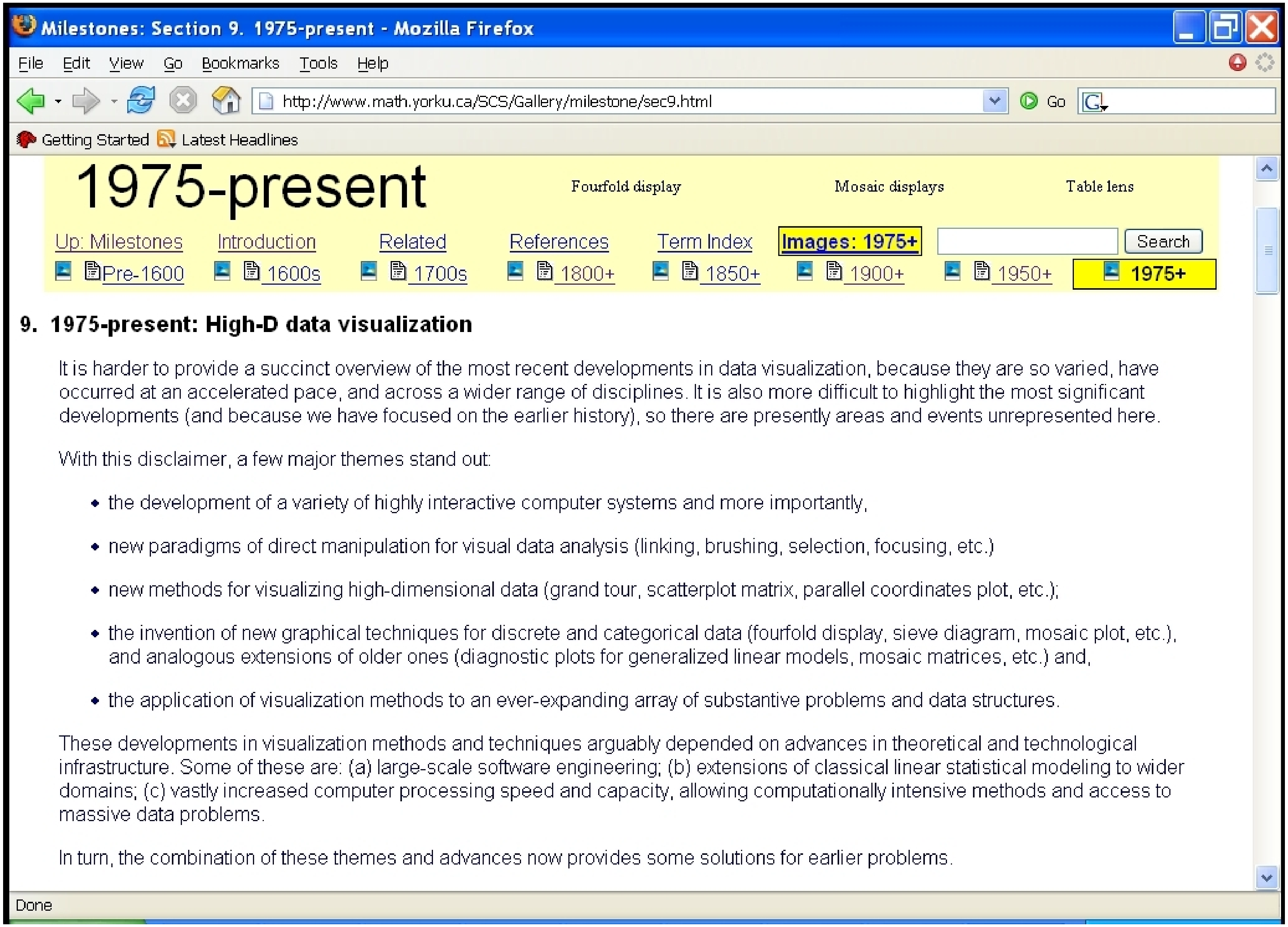
Prof. Dalitz FRS – 1925-2006

Some data visulisations from particle physics

Excluding event displays!



Status of the CKM matrix


Milestones: Section 9. 1975-present - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.math.yorku.ca/SCS/Gallery/milestone/sec9.html

Getting Started Latest Headlines

1975-present

Fourfold display Mosaic displays Table lens

Up: Milestones | Introduction | Related | References | Term Index | Images: 1975+ | Search

Pre-1600 | 1600s | 1700s | 1800+ | 1850+ | 1900+ | 1950+ | 1975+

9. 1975-present: High-D data visualization

It is harder to provide a succinct overview of the most recent developments in data visualization, because they are so varied, have occurred at an accelerated pace, and across a wider range of disciplines. It is also more difficult to highlight the most significant developments (and because we have focused on the earlier history), so there are presently areas and events unrepresented here.

With this disclaimer, a few major themes stand out:

- the development of a variety of highly interactive computer systems and more importantly,
- new paradigms of direct manipulation for visual data analysis (linking, brushing, selection, focusing, etc.)
- new methods for visualizing high-dimensional data (grand tour, scatterplot matrix, parallel coordinates plot, etc.);
- the invention of new graphical techniques for discrete and categorical data (fourfold display, sieve diagram, mosaic plot, etc.), and analogous extensions of older ones (diagnostic plots for generalized linear models, mosaic matrices, etc.) and,
- the application of visualization methods to an ever-expanding array of substantive problems and data structures.

These developments in visualization methods and techniques arguably depended on advances in theoretical and technological infrastructure. Some of these are: (a) large-scale software engineering; (b) extensions of classical linear statistical modeling to wider domains; (c) vastly increased computer processing speed and capacity, allowing computationally intensive methods and access to massive data problems.

In turn, the combination of these themes and advances now provides some solutions for earlier problems.

Done

Milestones in the history of thematic cartography, statistical graphics and data visualisation – M. Friendly and D. Denis Jan 2006

Big thankyou to Michael Friendly website
http://www.math.yorku.ca/SCS/StatResource.html

1975 to now High D data visualisation

Some key dates…**selective list ..**

1985 Alfred Inselberg **Parallel Coordinates**
1985 D. Asimov **Grand Tour**
1985 DataDescription Inc. Paul Velleman Cornell - DataDesk
1987 A. Becker and W. Cleveland **Linking and Brushing**
1998 A. Buja, D. Asimov, C. Hurley, J. McDonald XGobi
1990 **E. Wegman Statistical analysis and parallel coord. CrystalVision**.
1991 M. Friendly Mosaic Display and Categorical data
1999 L. Wilkinson "Grammar of Graphics"

Systemization of data and graphs and graph algebras in an OO framework.

Particle Physics Data - a problem in the analysis of a huge amount of multivariate data

What do we use ? Histograms and scatterplots. Sometimes use colour

Can one use the latest computer graphics technology or ideas that statisticians and computer scientists have dreamt up in the last decade…?

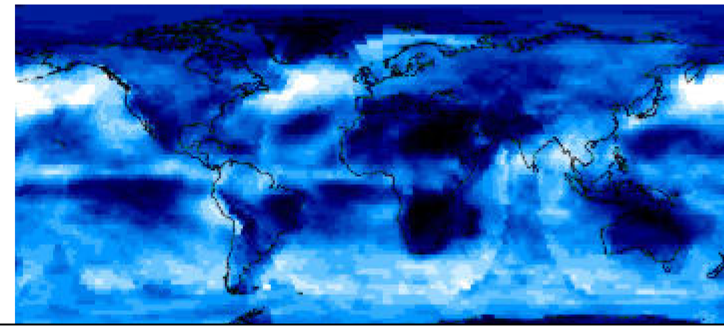
To illustrate,will use the "pollen dataset" to show use of parallel coordinates, brushing and pruning.
The Grand Tour.comes later……………………….

There are many other ideas …………………….

American Statistical Association

Have challenges each year
This is the 2006 one

Data Exposition 2006
Sponsored by the Sections on Statistical Graphics, Statistical Computing, and Statistics and the Environment.



- **The data set:** The data are geographic and atmospheric measures on a very coarse 24 by 24 grid covering Central America. The variables are: elevation, temperature (surface and air), ozone, air pressure, and cloud cover (low, mid, and high). With the exception of elevation, all variables are monthly averages, with observations for Jan 1995 to Dec 2000. These data were obtained from the NASA Langley Research Center Atmospheric Sciences Data Center (with permission; see important copyright terms below).
- More details about the data, including descriptions of the variables, are available here.
- Download the data as a gzipped tar ball or as a zip file.
- There is also a flyer available.
- **The question:** The aim of the Data Expo is to provide a *graphical* summary of important features of the data set. This is intentionally vague in order to allow different entries to focus on different aspects of the data. For example, the focus can be on: the fact that the data are multivariate, or time-series, or spatial; or the fact that the data contain missing values; or the focus could even be on the *process* of exploring the data.
- Some obvious general questions that could be answered are: What are the important relationships between the variables? Are there any important trends in the data? Are there any important groupings or clusters in the data? Are there any unusual locations or time periods in the data set?

Pollen Data Set

the **data set** from the 1986 **JSM** Exposition's **dataset** and was assembled by David Coleman of RCA Labs

JSM = Joint Statistical Meeting

Data Visualisation Software

CrystalVision - E. Wegman
GGobi
XmdvTool
Orange

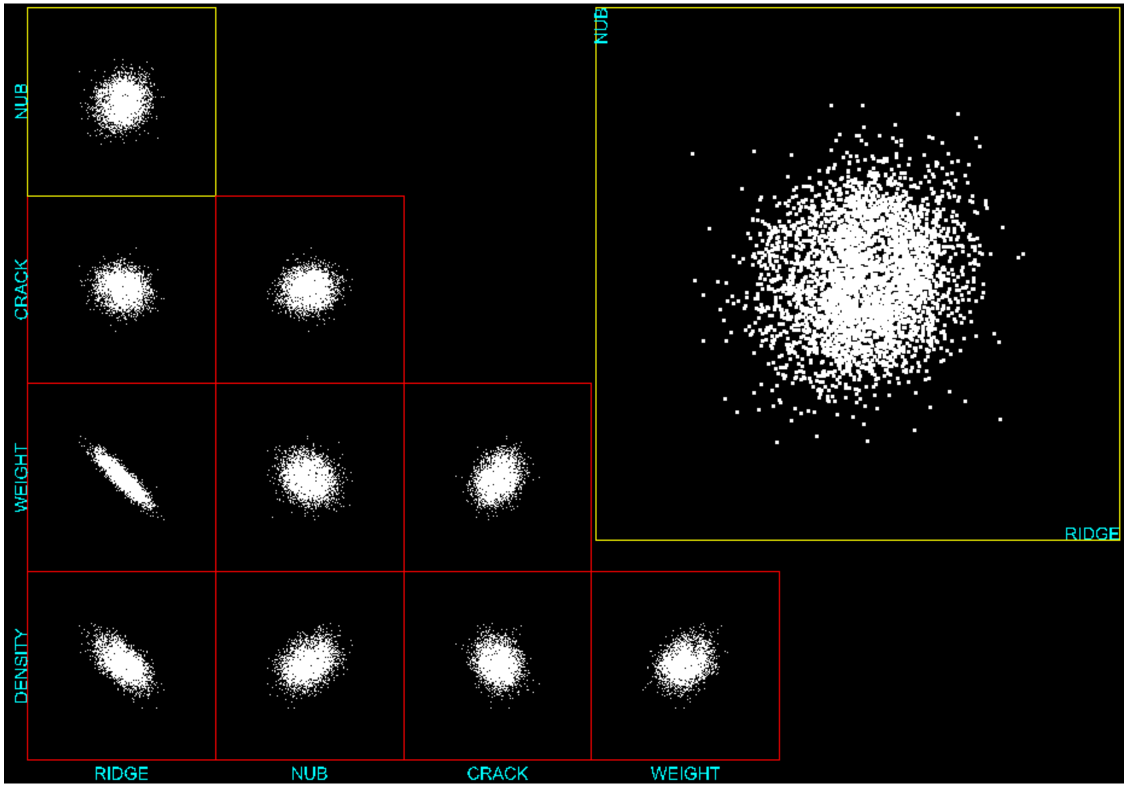
In graphics, a portion of each pixel's data that is reserved for transparency information. 32-bit graphics systems contain four channels -- three 8-bit channels for red, green, and blue (RGB) and one 8-bit alpha channel. The **alpha** channel is really a mask -- it specifies how the pixel's colors should be merged with another pixel when the two are overlaid, one on top of the other.

1) Try this on the pollen data set with CrystalVision

2) Now parallel coordinates.

Problem - how do you study an N-Dimensional space (N>2) when you only have a flat screen ?

This is one solution - with colour mixing (blending) and the alpha channel (transparency) - is very powerful



Note:
Size of dots matters!

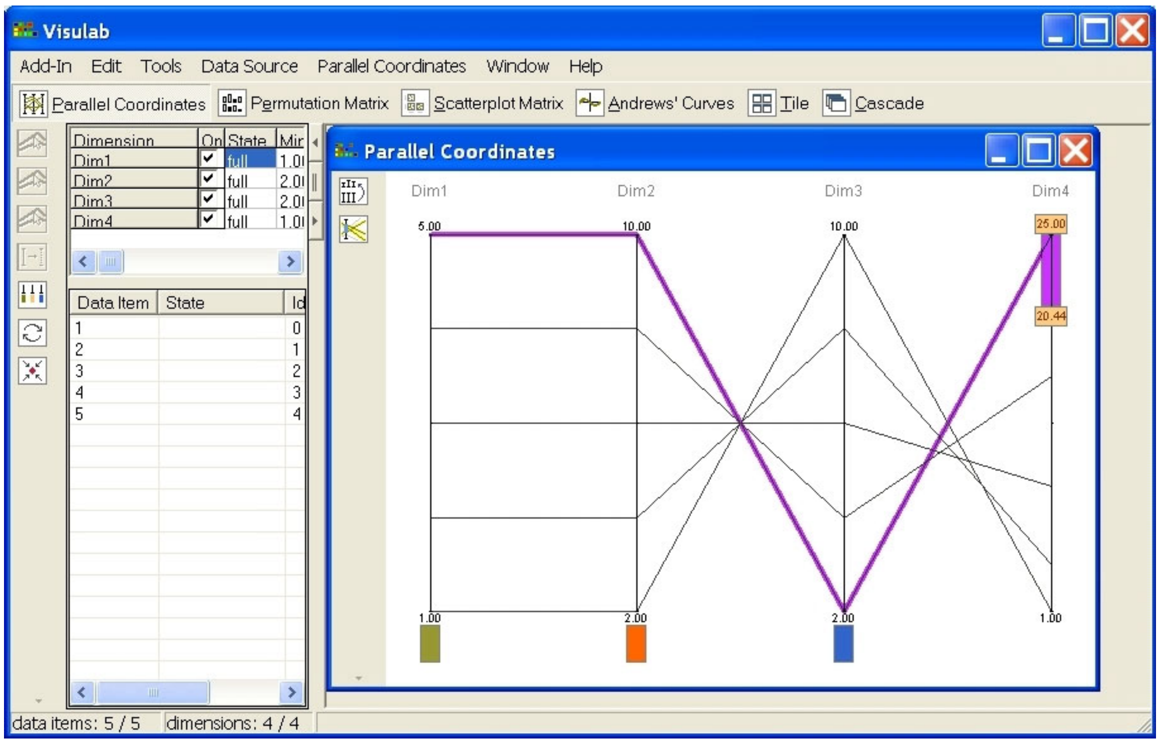
The pollen data - this is called a scatter matrix.
2D projections of this 5 variable space helps - but -

Greatly help matters using colour and the alpha channel

Introduction to Parallel Coordinates

DataPoint	Dim1	Dim2	Dim3	Dim4
1	1	2	10	1
2	2	4	8	4
3	3	6	6	9
4	4	8	4	16
5	5	10	2	25

Simple Implementation with EXCEL plugin
http://www.inf.ethz.ch/personal/hinterbe/Visulab/



This also shows the idea of brushing

Hyperdimensional Data Analysis Using Parallel Coordinates

Edward J. Wegman

Journal of the American Statistical Association, Vol. 85, No. 411 (Sep., 1990), 664-675.

Wegman has done much on the use of parallel coords.

Some useful things to note…….

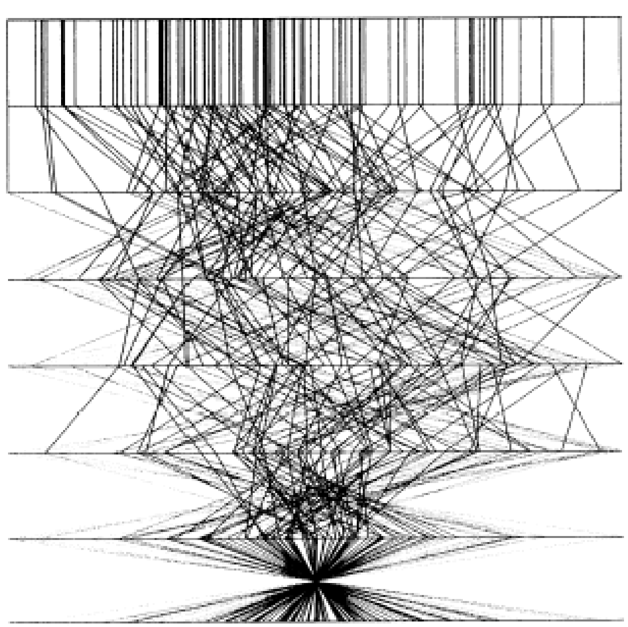


Figure 3. Parallel Coordinate Plot of Six-Dimensional Data Illustrating Correlations of ρ = 1, .8, .2, 0, −.2, −.8, and −1.

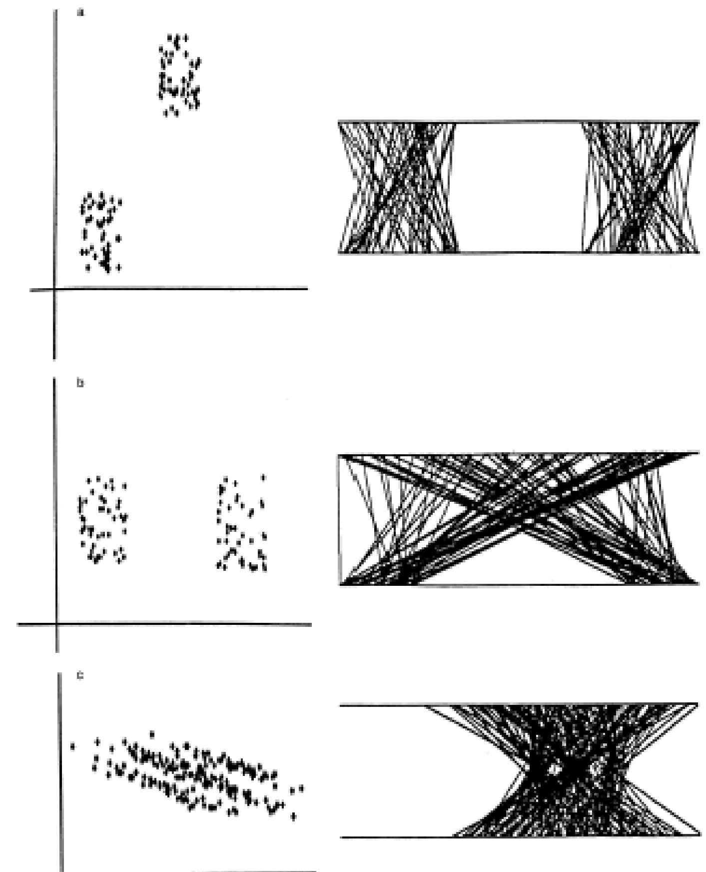
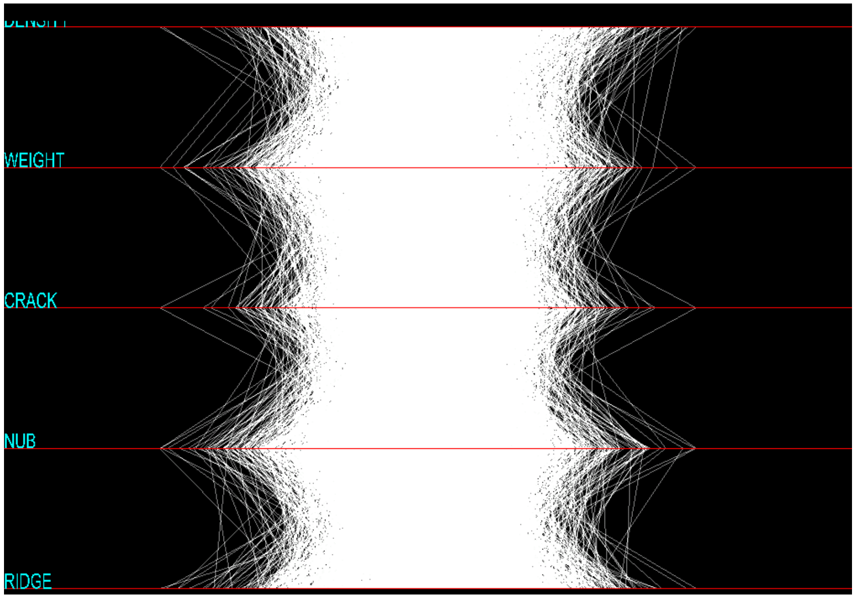
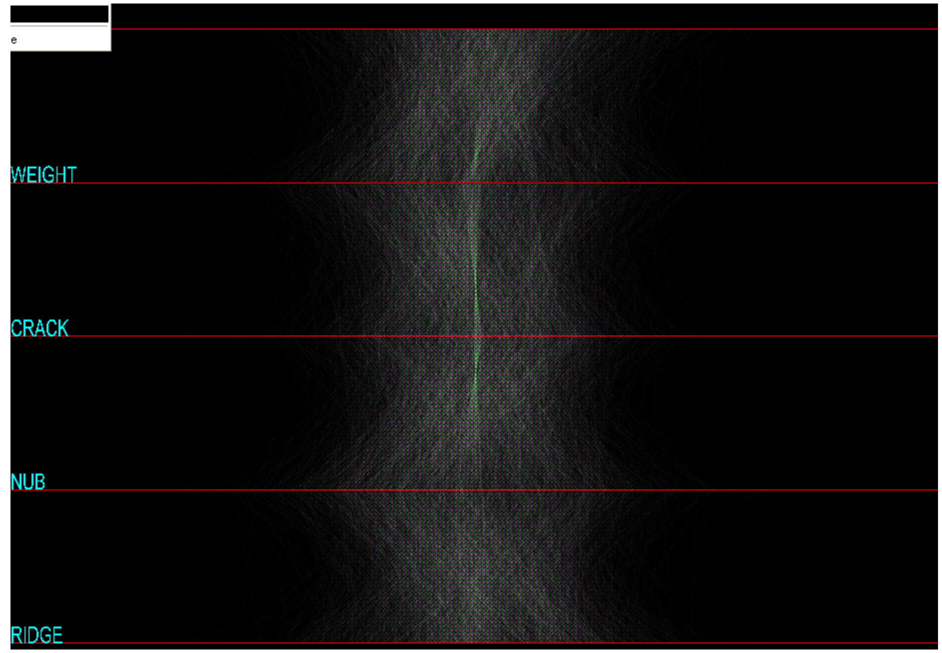


Figure 5. (a) Clustering That Is Separated in Both x and y, Represented in Parallel Coordinates. (b) Clustering That Is Separated in the x Coordinate but not the y Coordinate. (c) Clustering That Is Separated in Neither Projection. Since the first coordinate of the crossover point for negatively correlated variables is dependent on the intercept, we can separate overlapping clusters.



High alpha

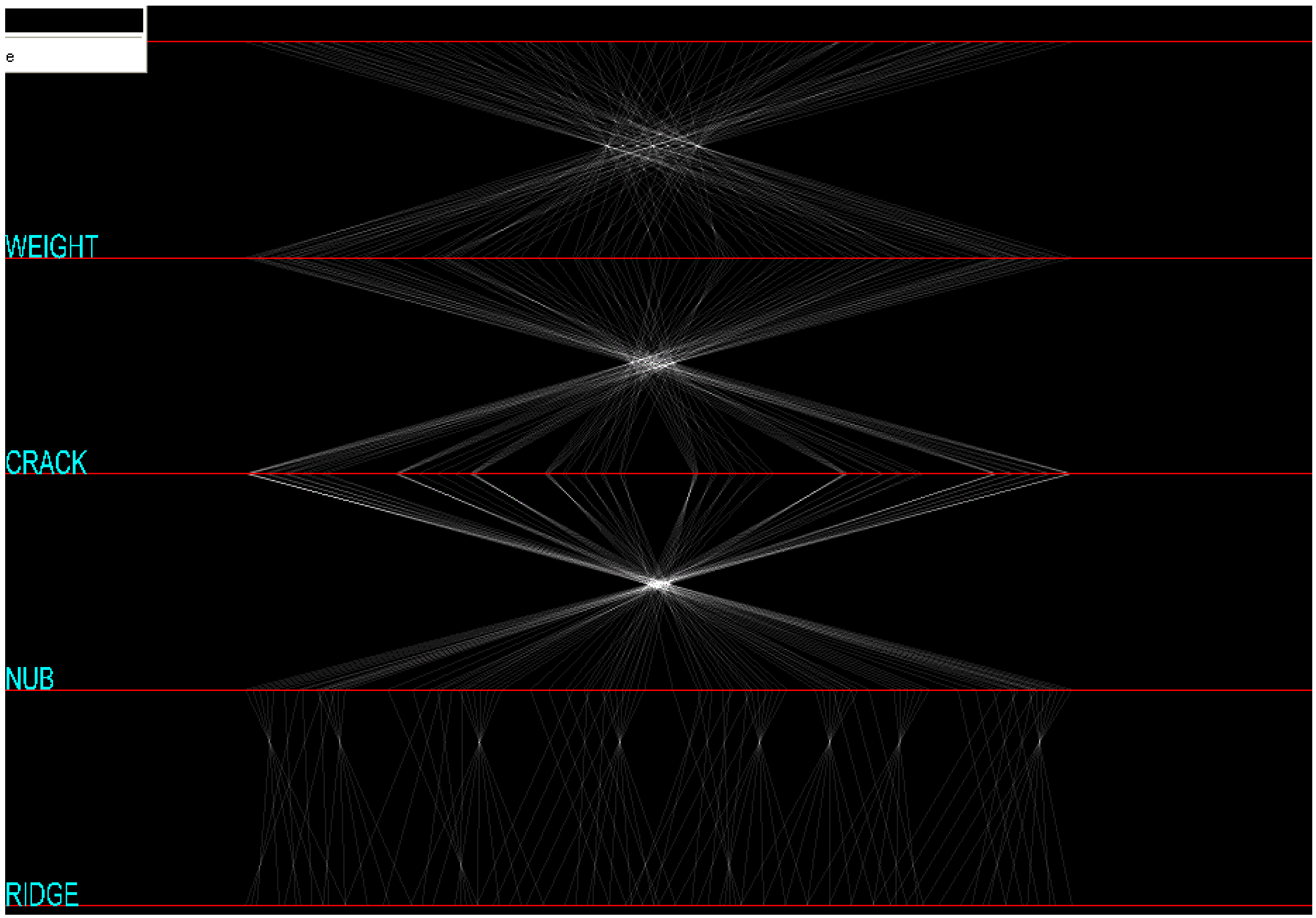


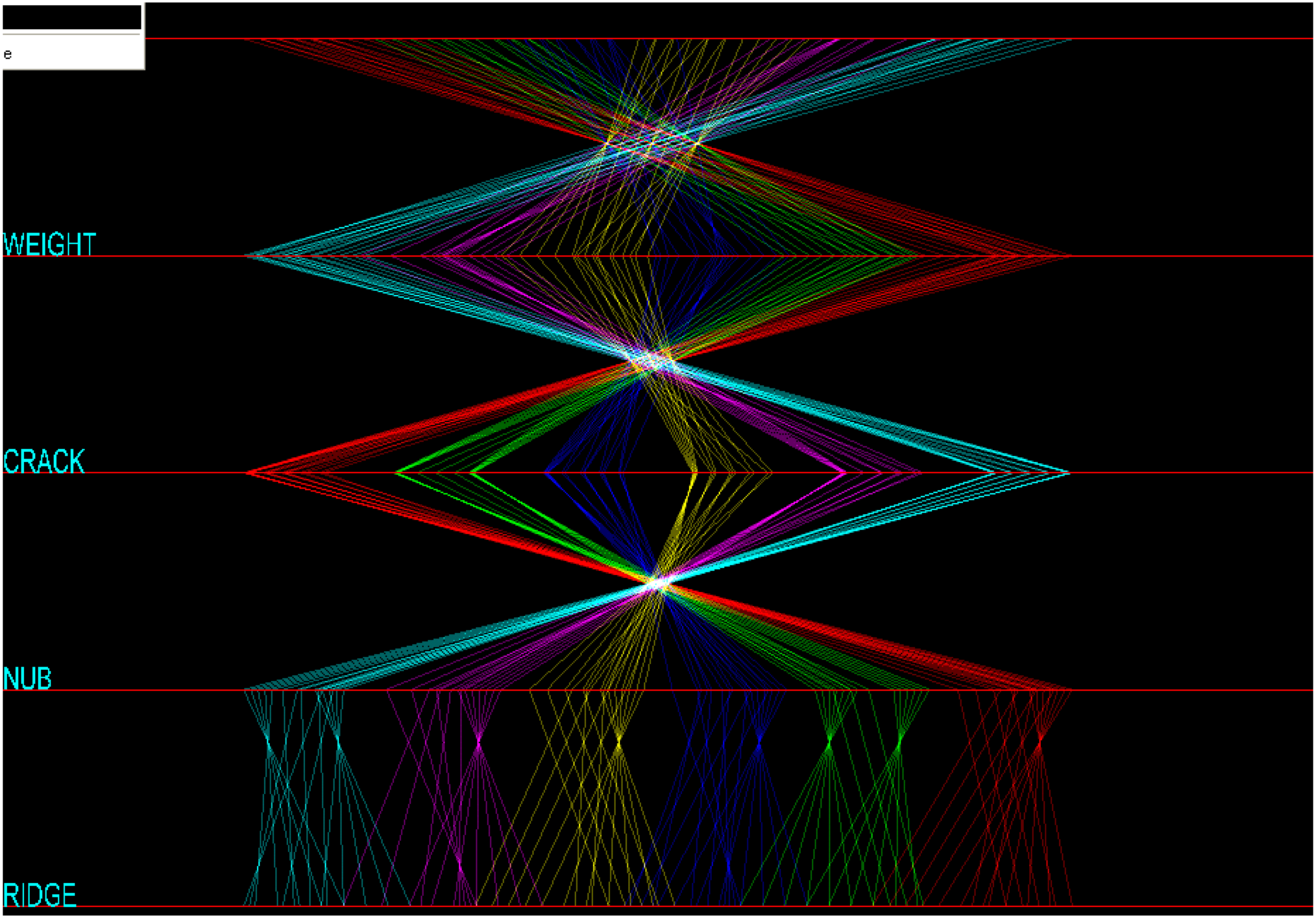
Low alpha

Now brushing - colour the data with chosen colours

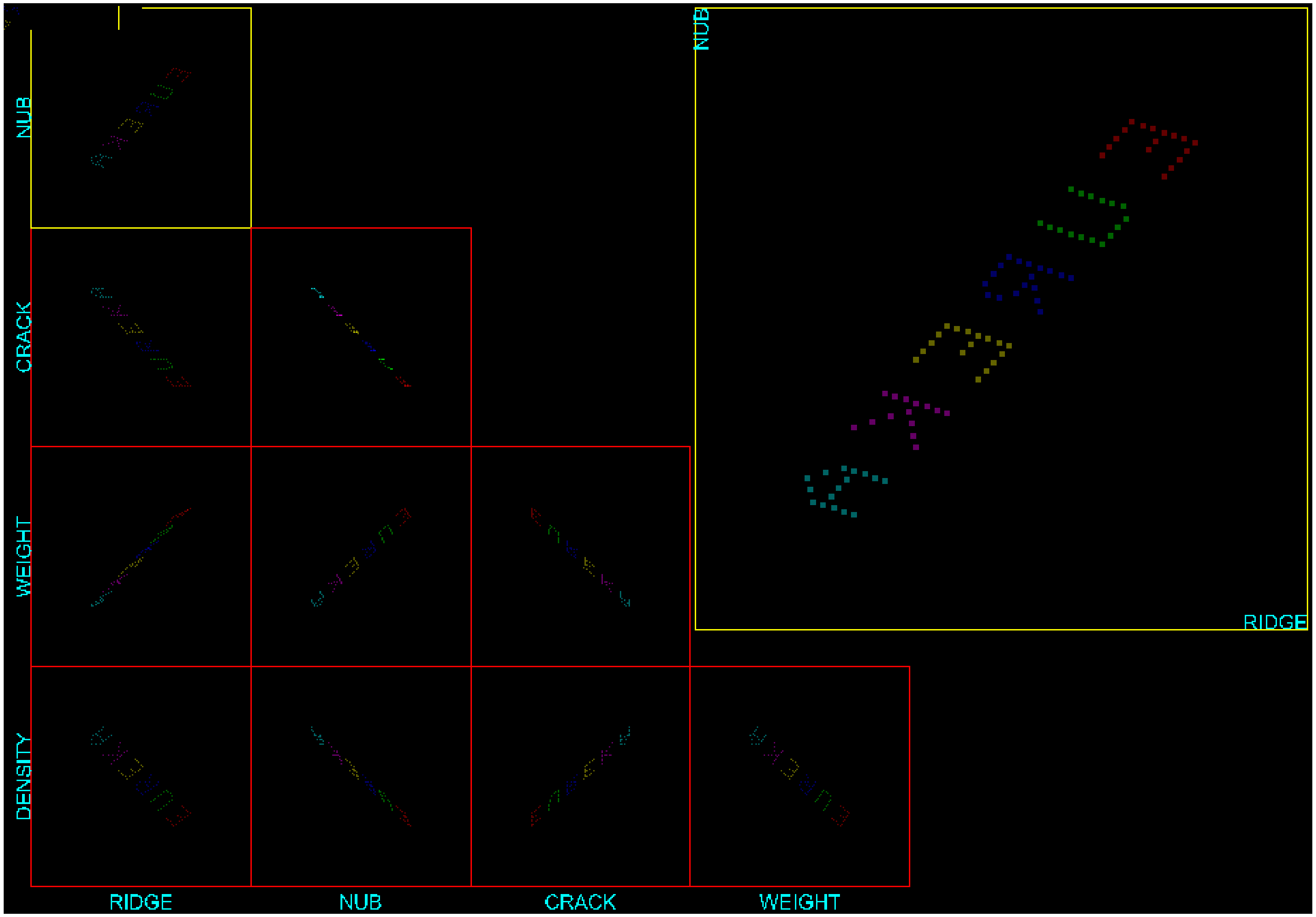
and pruning - cut data you do not want

First lets PRUNE





WEIGHT
CRACK
NUB
RIDGE

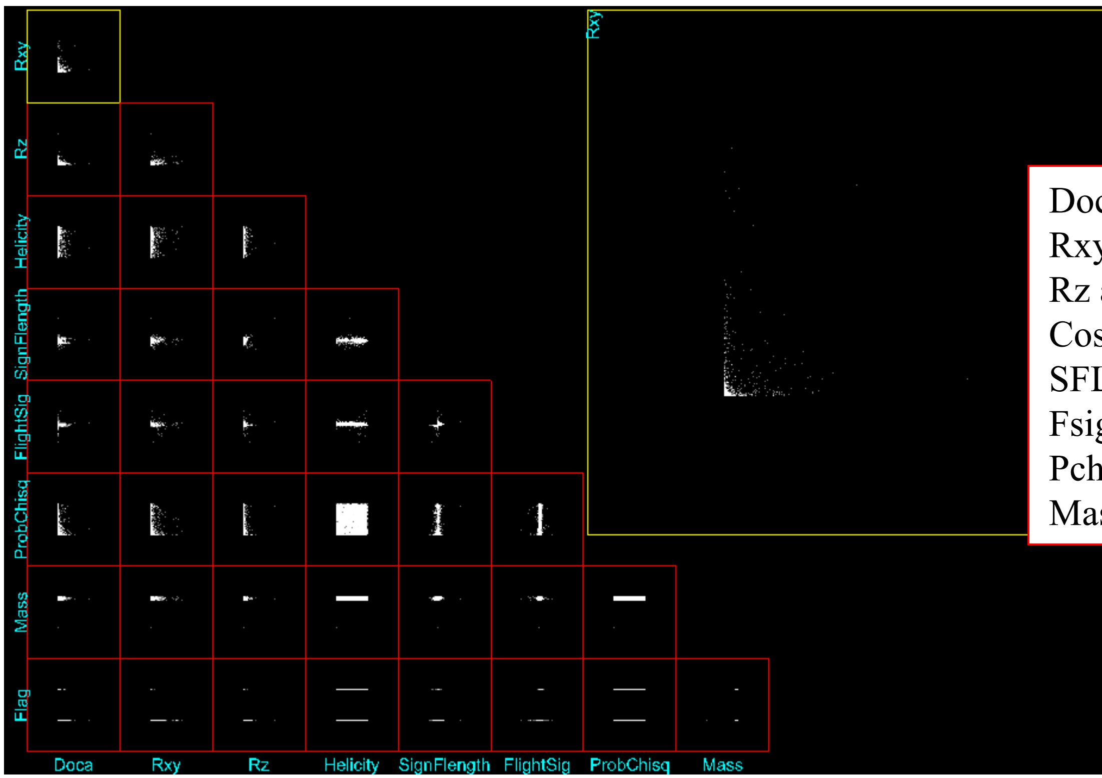


98/3848 points. S/B = 2.6%
There are other features in the data. See E. Wegman
Contrived example, but helps a newcomer to use this type of graph.

Now lets try some particle physics monte carlo data

From Liliana Teodorescu - 1264 Kzero + 3734 background (and a flag to tell us which is which ! Flag =1 S Flag=0 B

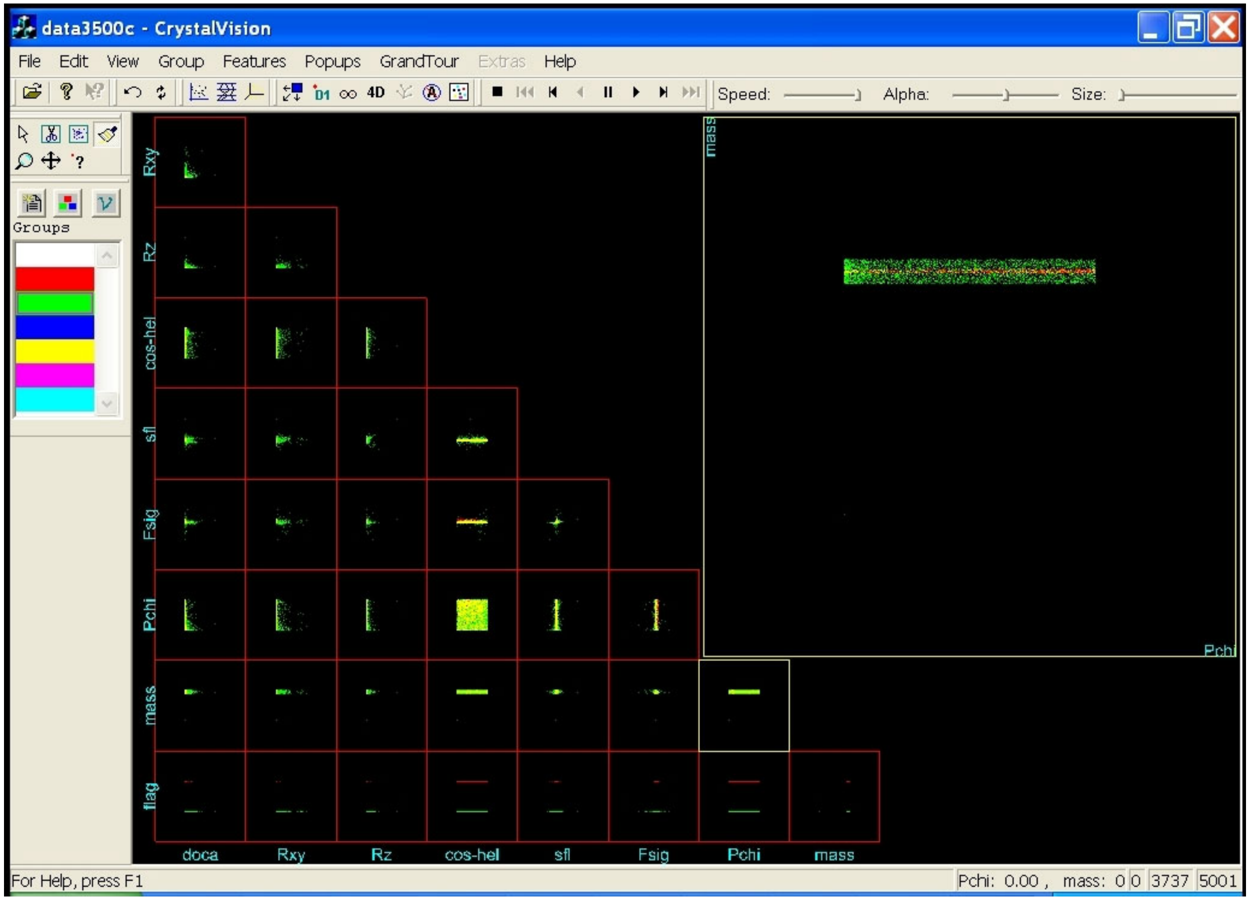
LT has shown how to use GEP on this dataset in another talk.



$K_s \rightarrow \pi^+ \pi^-$

Doca = distance of closest approach
Rxy radius of cylinder for interaction region
Rz abs. half length of cylinder defining the IR
Cos_hel abs. Value of cosine of Ks helicity angle
SFL – signed flight length
Fsig stat. Sig. Of Ks flight length
Pchi chisq prob of Ks vertex
Mass – reconstructed mass of the Ks

CrystalVision – E. Wegman



CrystalVision
(E. Wegman)

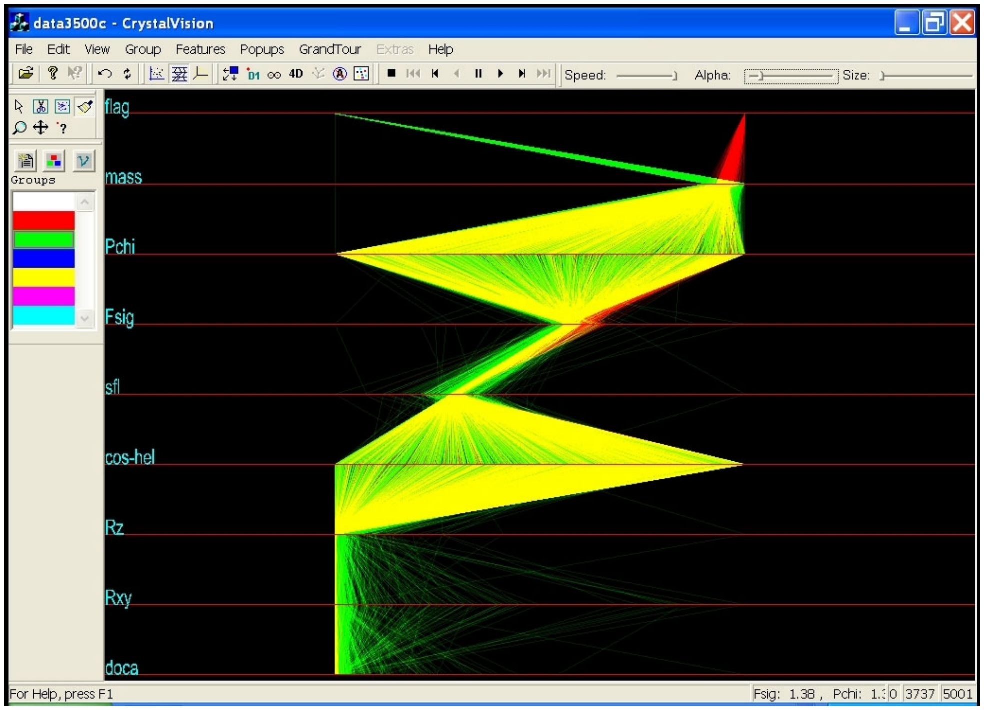
Has blending and control of intensity

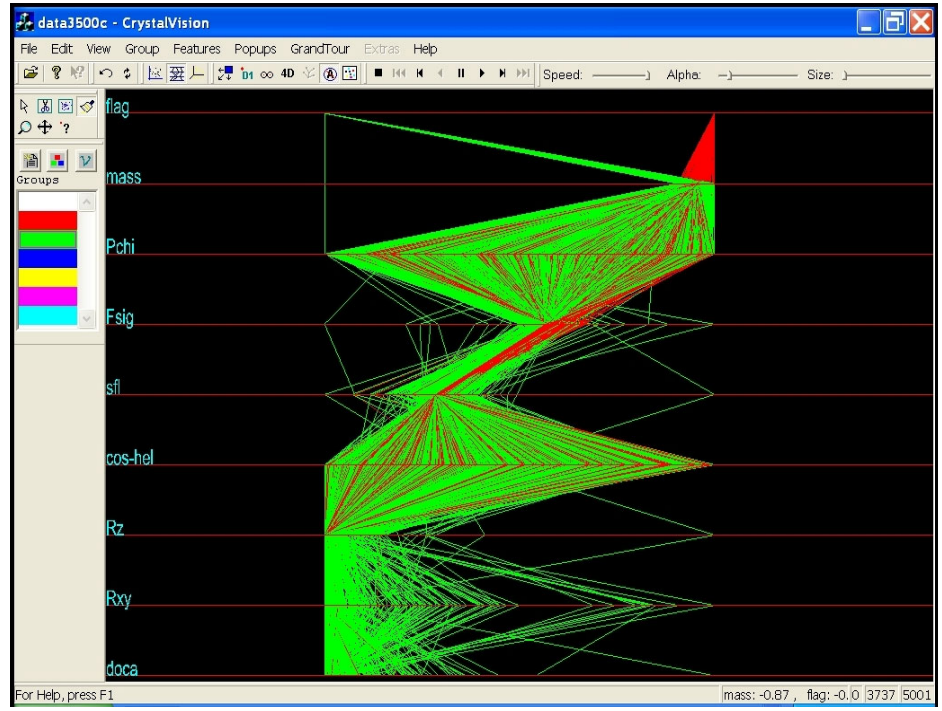
VERY Powerful

Brush signal RED and background GREEN

If they overlap RED + GREEN = YELLOW (yellow)

Now go to parallel coordinates - adjust alpha



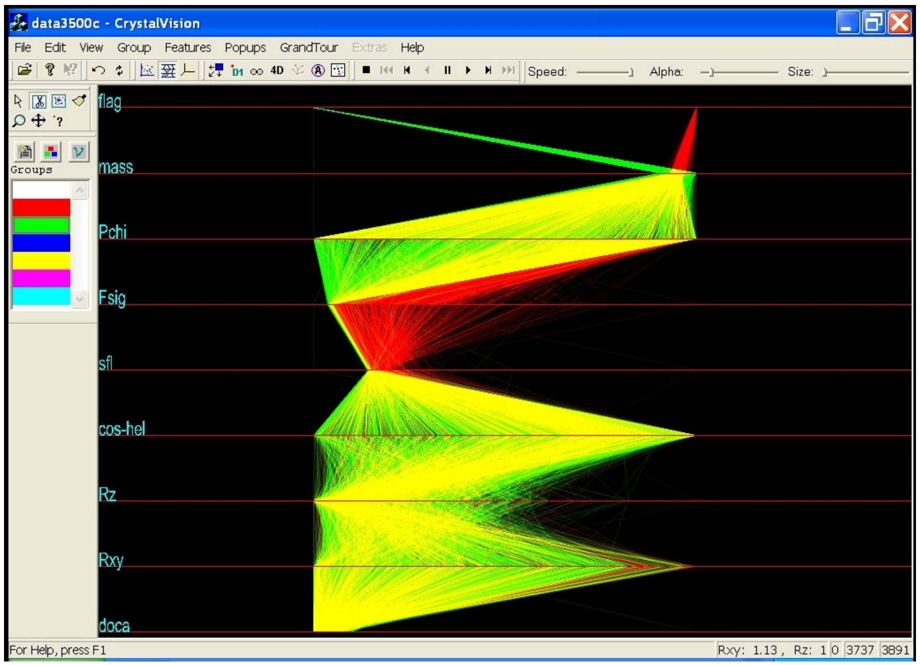


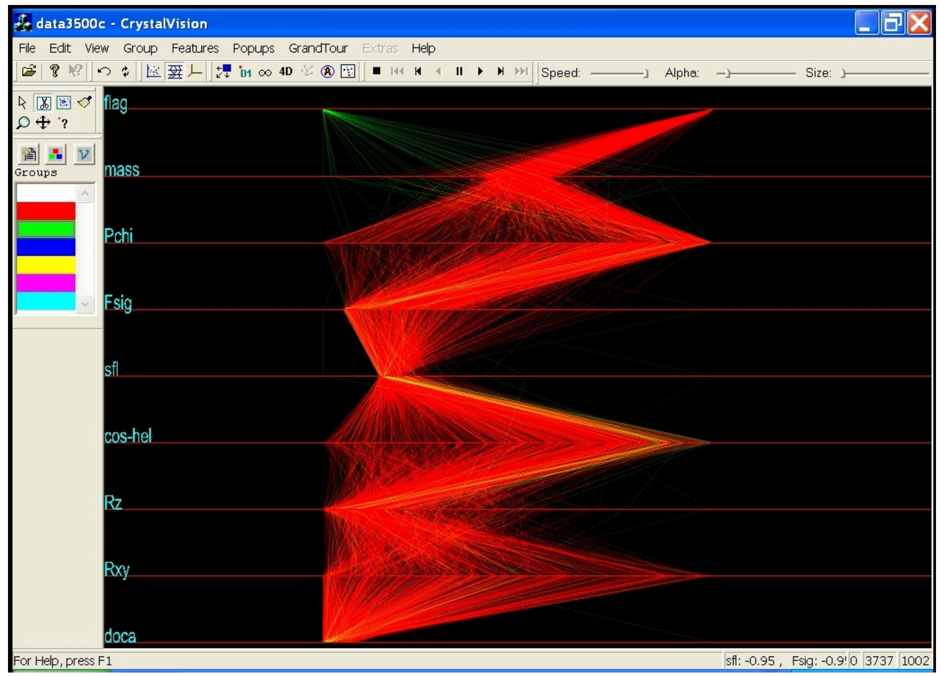
Note: See affect of turning alpha channel on and off

Note:Parallel Coords Vertical. Sclaes data between min. and max.

Immediately see that R_{xy}, Doca (and sfl less so) discriminate the background
Only variable where signal can be seen is Fsig.

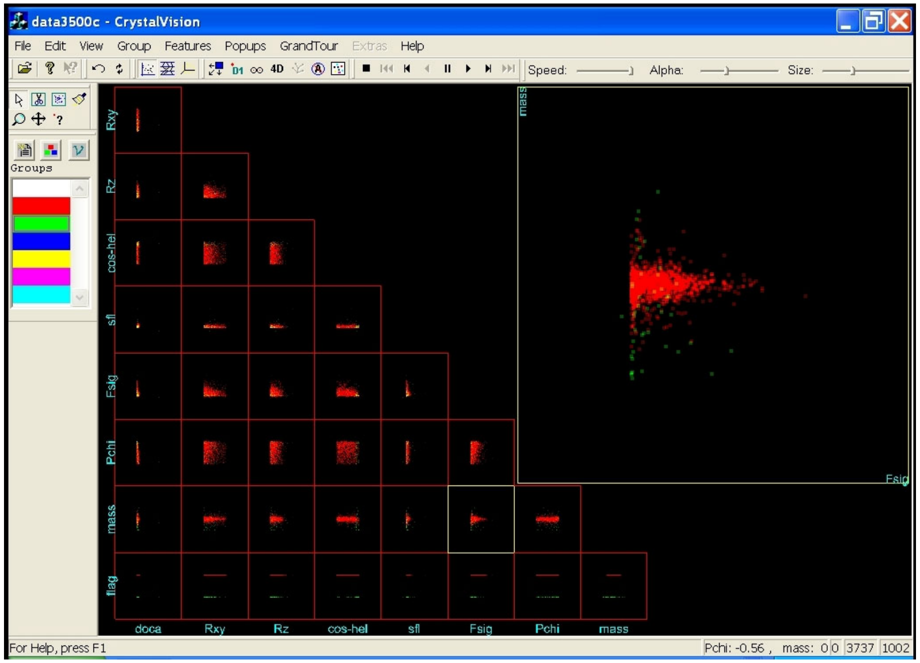
How to clean up this data - " what is the order of cuts ?"
Remove obvious background (Prune Doca and Rxy)
Then select signal (FSig)





Takes just a couple of minutes to do this…

Back in scatterplot space



958 S 44 B   95% Purity 80% Efficiency

Did not spend long on this – Exploratory Visual Data Analysis

Powerful way to decide which variables matter and the order in which cuts should be applied.

Precursor to machine learning approach ….to be continued

Multivariate classifiers

An algorithm/tool that assigns a point in a multidimensional space to one of several possible categories e.g. signal or background.

- Cut (binary split or stump)
- Decision Trees
- Support Vector Machine (SVM)
  (NOTE: radial basis function SVM equiv. to a type of Neural Net)
- Neural Net
- k nearest neighbours (kNN)
- VizRank
- Genetic Algorithm, Genetic Programming,
  Gene Expression Programming (GEP) – can do other things also.
- etc………………

ViZRank – finding informative data projections in Functional genomics by machine learning

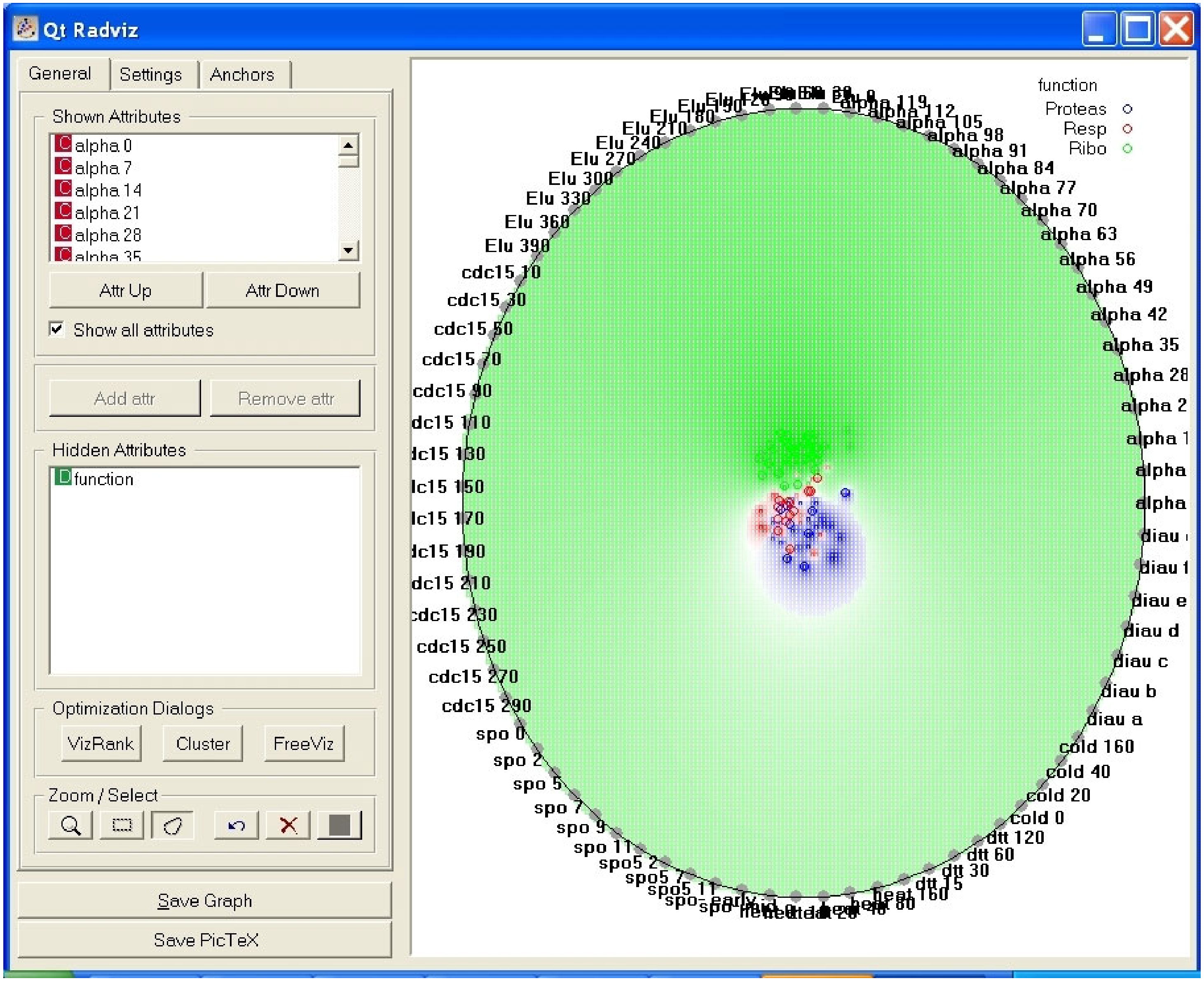
Bioinformatics applications note Vol 21 no. 3 2005 413-414
Gregor Leban et al……..Use ORANGE

Project N-dimension space onto a reduced number of dimensions (use radviz or polyviz visualisation)

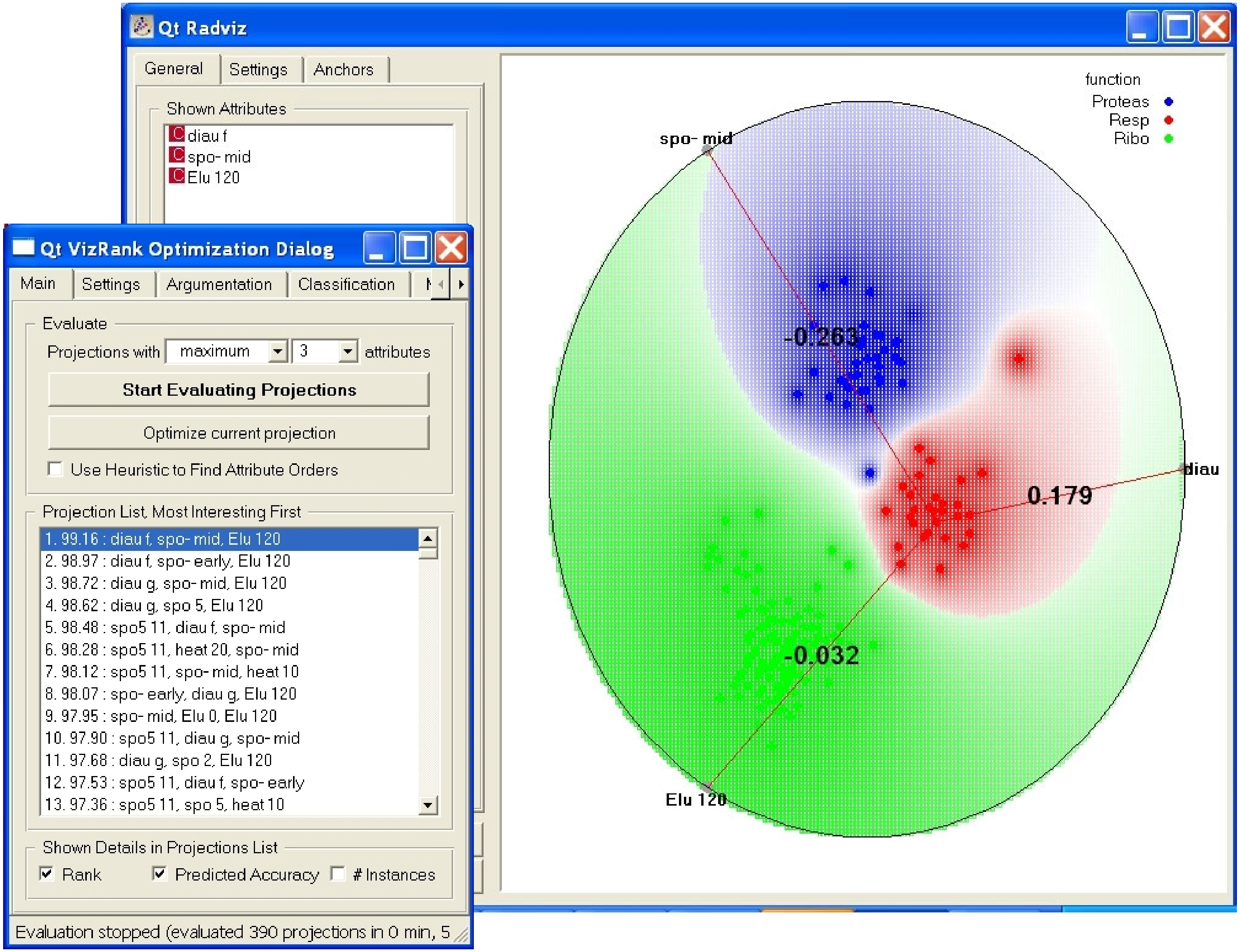
Rank projections using kNN classifier

Yeast *Saccharomyces cerevisiae*
79 different DNA microassay hybridization measurements
Describe each gene…….which ones matter ?

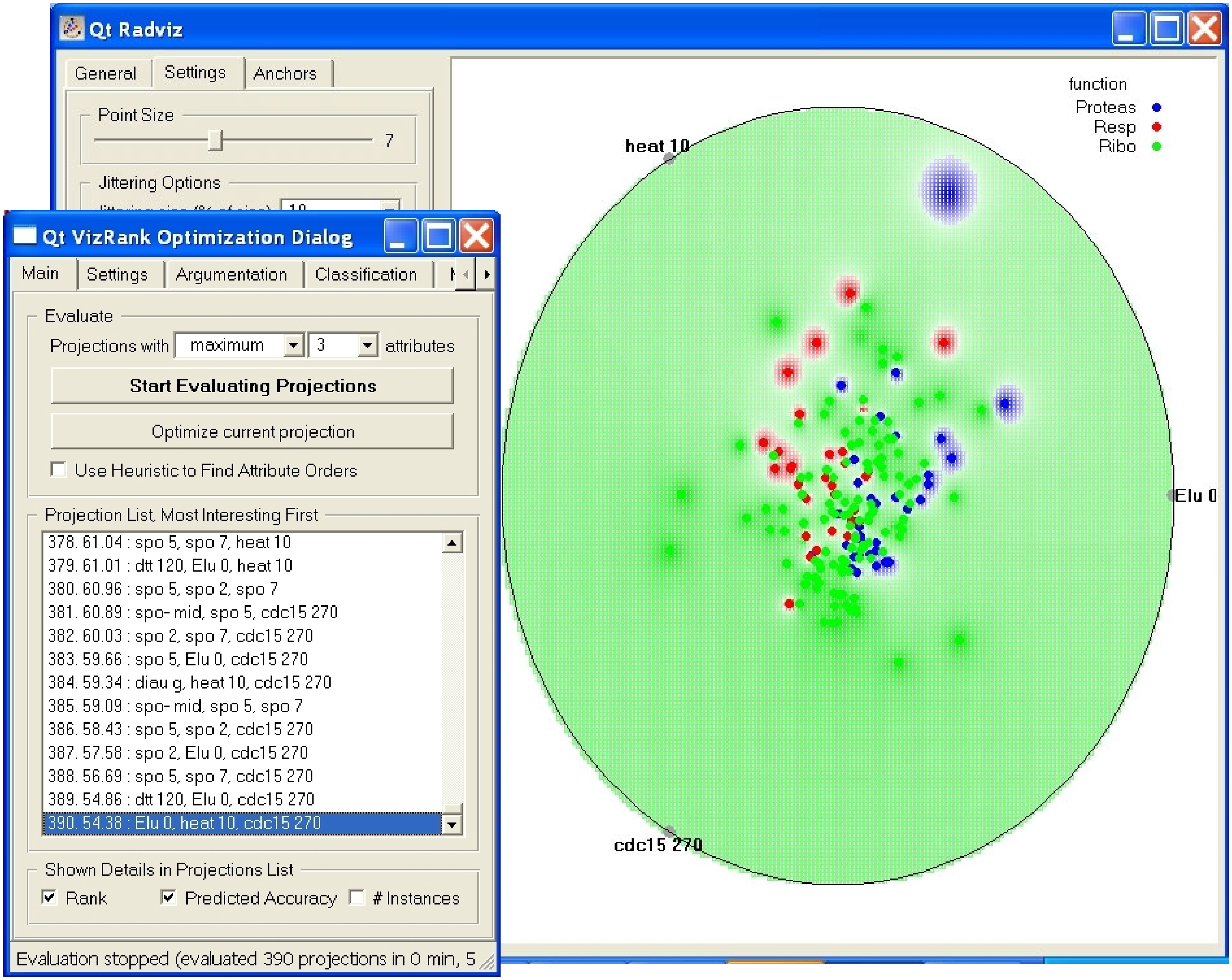
RadViz Visualisation – invented by P. Hoffman



Rank 3 attribute projections 99 %



Poorly ranked projection – 54%



Illustrate some other classifiers and visualisation using the wine dataset….

UCI KDD Repository

It would be great if a particle physics database existed for algorithm evaluation

THE WINE DATA SET

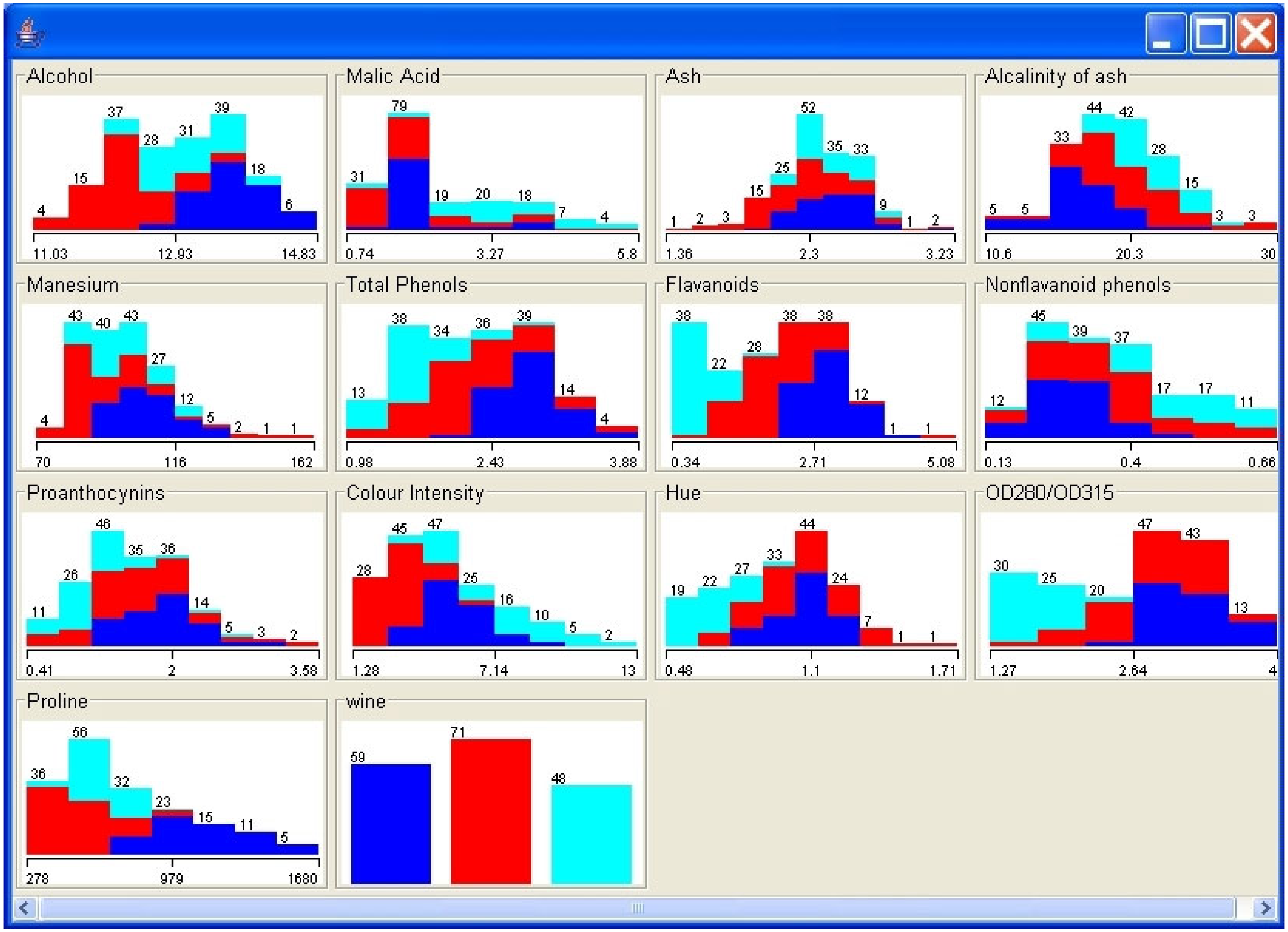
4. Relevant Information: -- These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

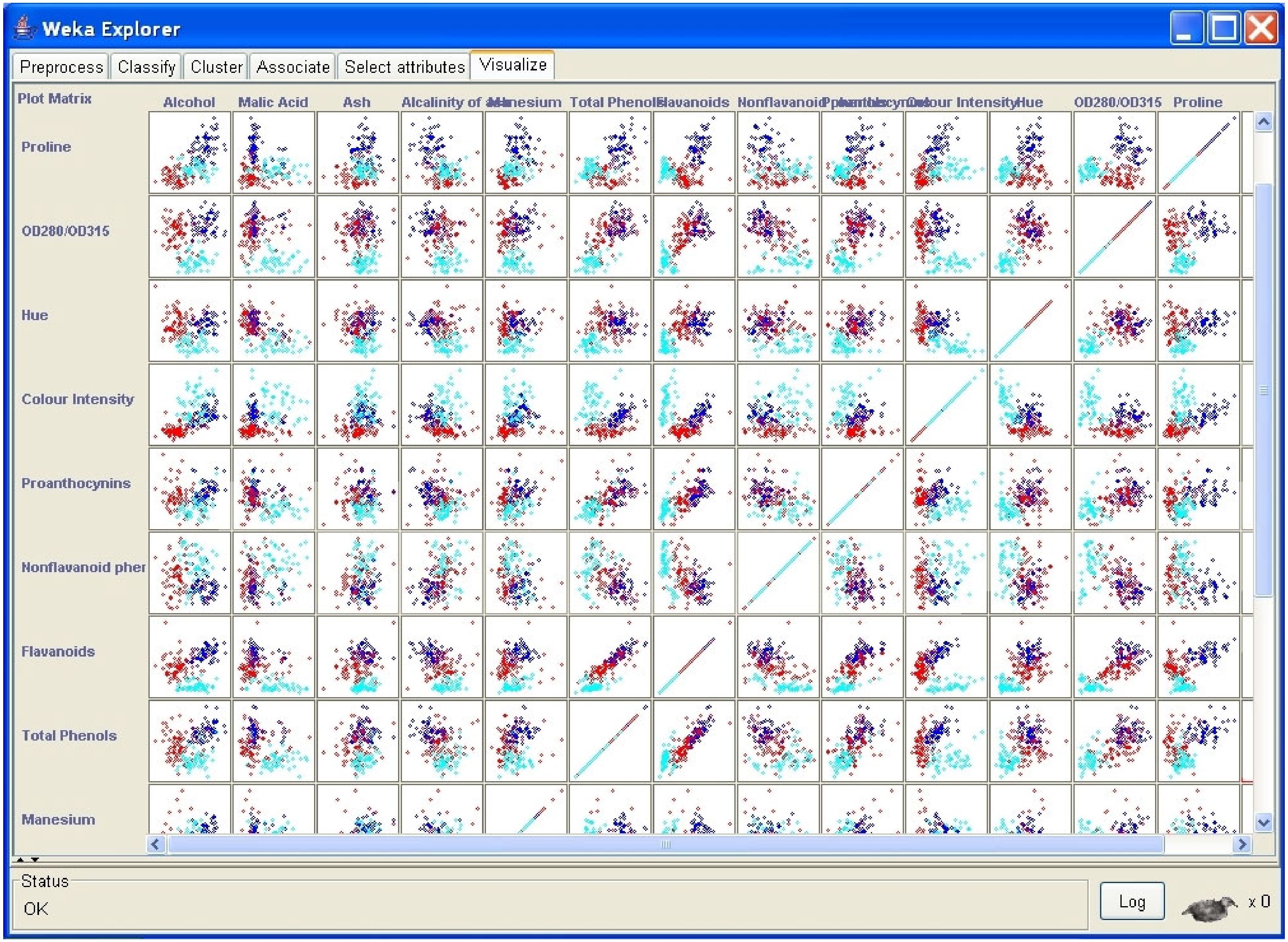
-- I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a list of what the 30 or so variables were, but a.) I lost it, and b.), I would not know which 13 variables are included in the set. !!

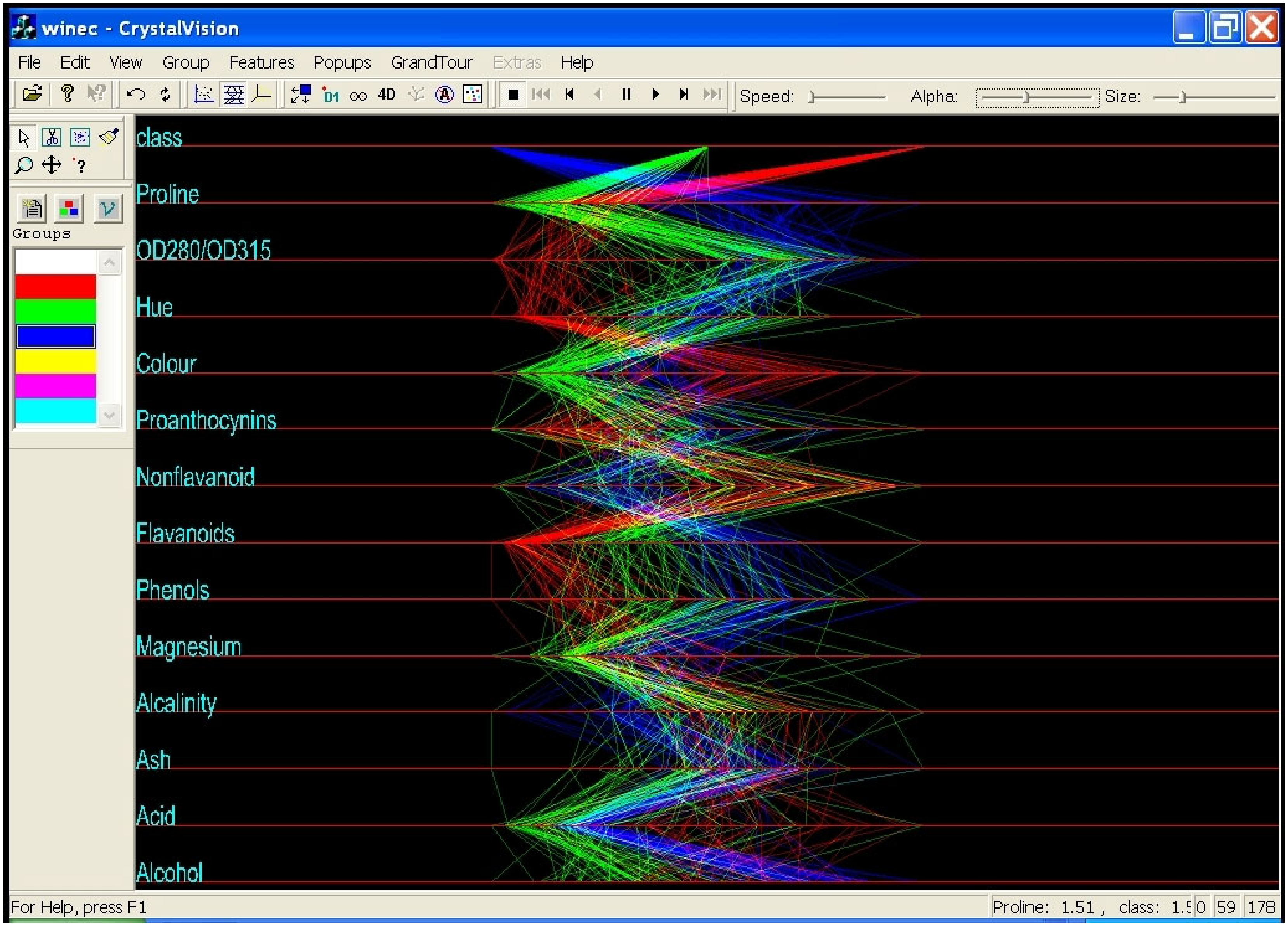
5. Number of Instances class 1 - 59 class 2 - 71 class 3 - 48

6. Number of Attributes 13

Use WEKA - histogram of all the wine dataset

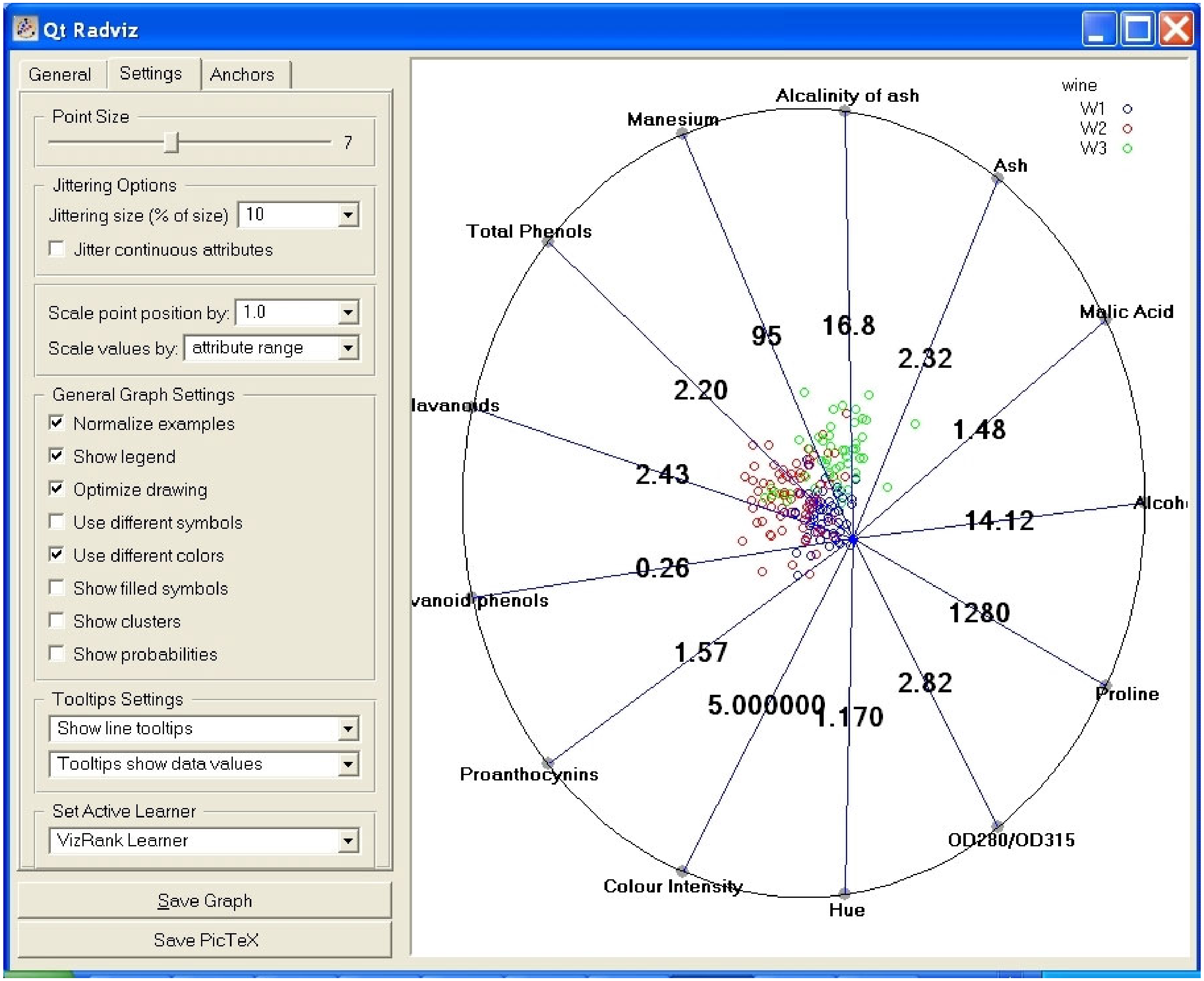


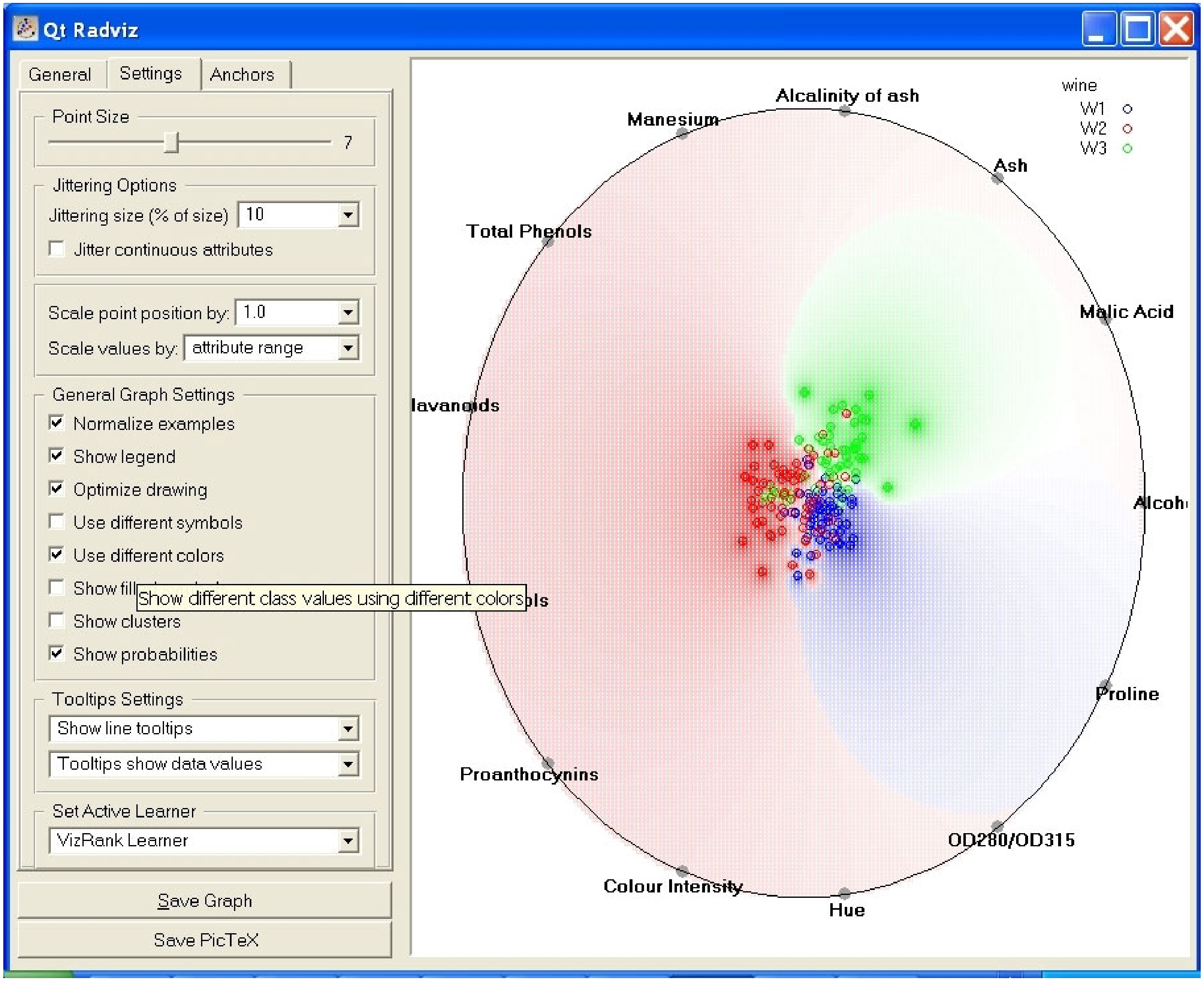

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
Alcohol
Malic Acid
Ash
Total Phenols
Flavanoids
Colour Intensity
OD280/OD315
Proline
Hue
Proanthocynins
Nonflavanoid pher
Manesium
Status
OK
Log
x 0



Once you get used to it, this plot contains a lot of information

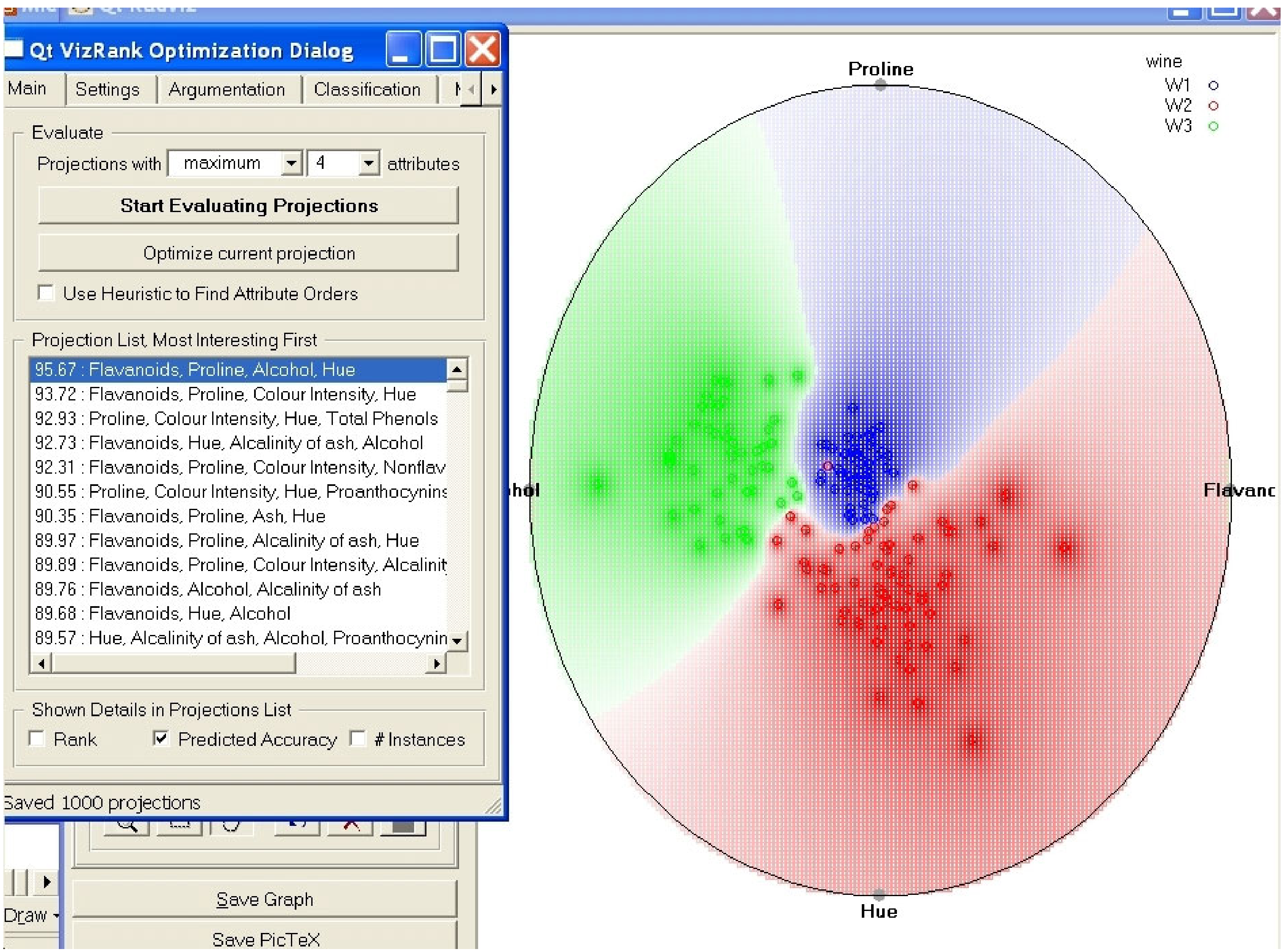
Wine dataset….





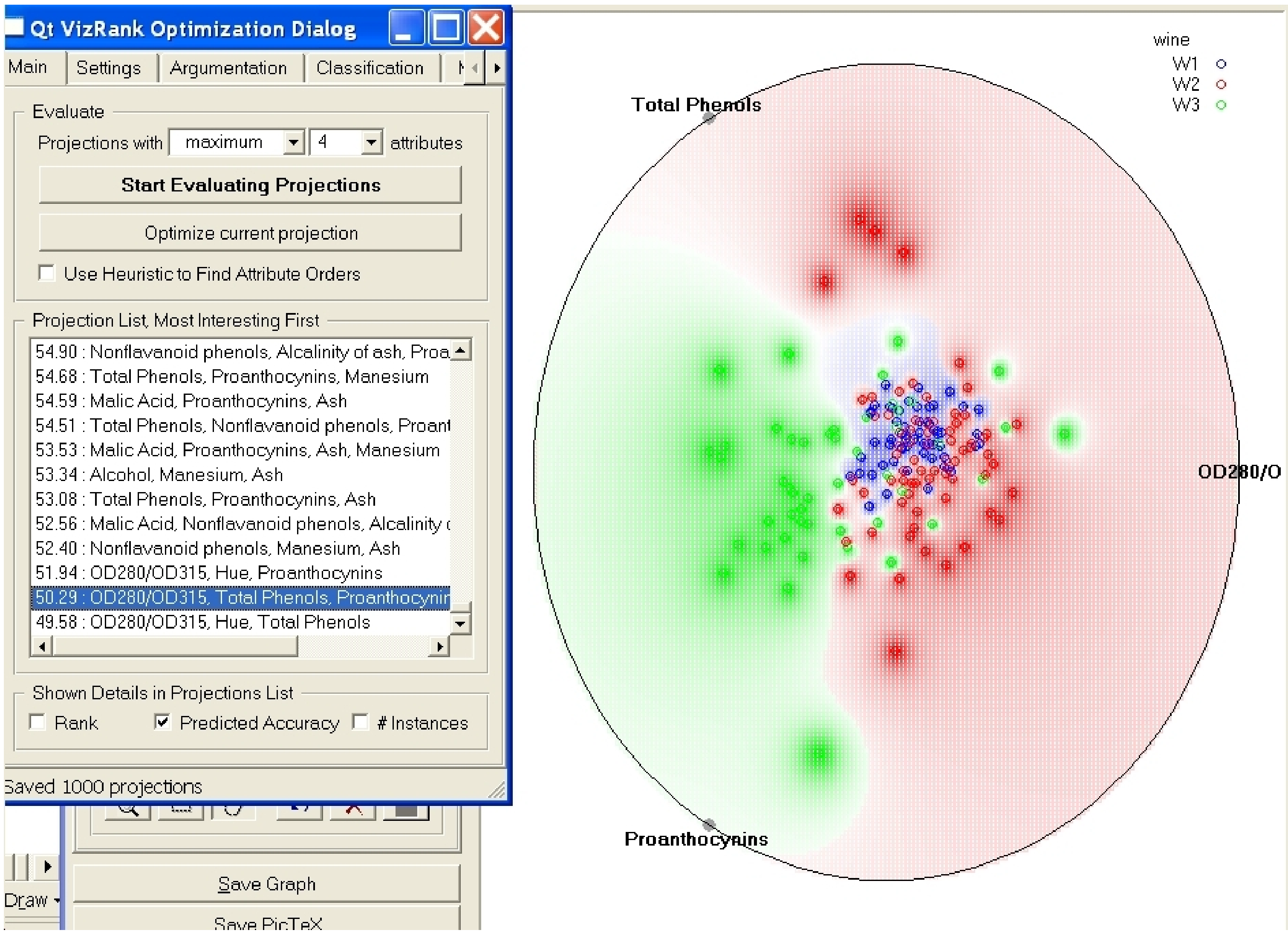
Qt Radviz
General
Settings
Anchors
Point Size
7
Jittering Options
Jittering size (% of size)
10
Jitter continuous attributes
Scale point position by:
1.0
Scale values by:
attribute range
General Graph Settings
Normalize examples
Show legend
Optimize drawing
Use different symbols
Use different colors
Show fill
Show different class values using different colors
Show clusters
Show probabilities
Tooltips Settings
Show line tooltips
Tooltips show data values
Set Active Learner
VizRank Learner
Save Graph
Save PicTeX
wine
W1
W2
W3
Alcalinity of ash
Manesium
Ash
Total Phenols
Malic Acid
lavanoids
Alcoh
Proline
Proanthocynins
OD280/OD315
Colour Intensity
Hue

Flavanoids, colour, hue, alcohol, proline matter…..

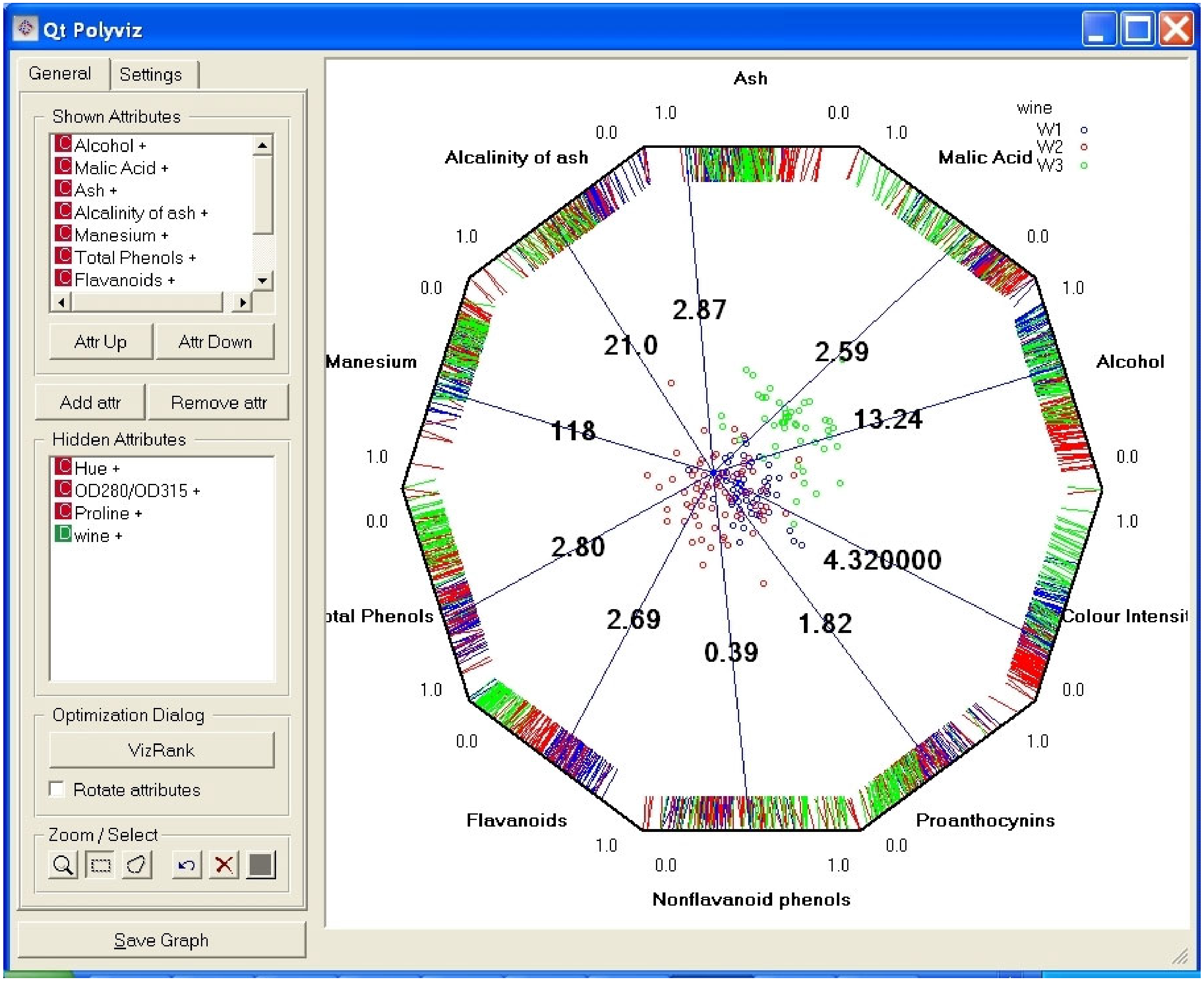


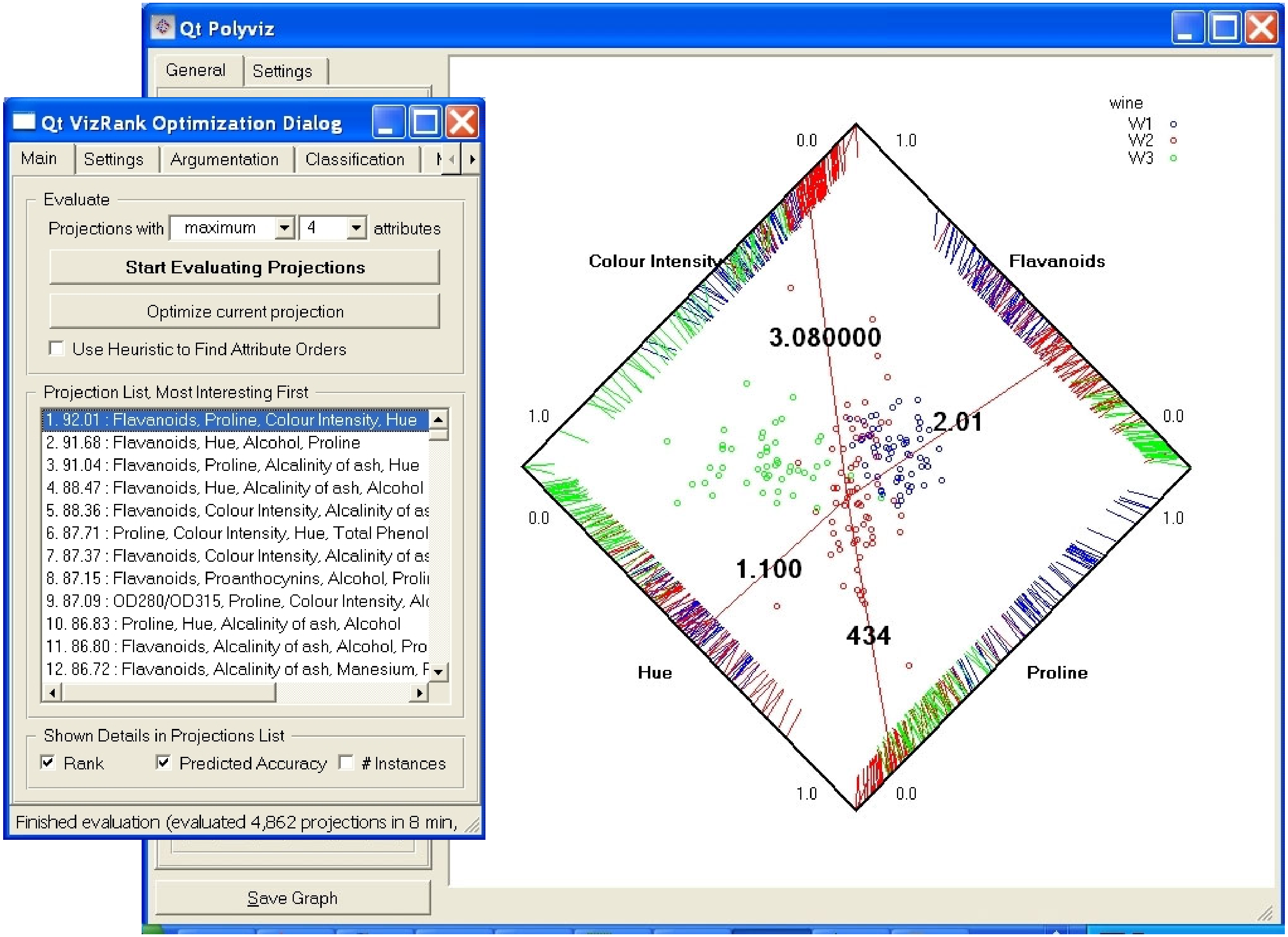
VIZRank algorithm applied….

50% selection…….wines overlap….



Polyviz – radviz + value of variable indicated





Qt Polyviz
General
Settings
Save Graph
Qt VizRank Optimization Dialog
Main
Settings
Argumentation
Classification
Evaluate
Projections with
maximum
4
attributes
Start Evaluating Projections
Optimize current projection
Use Heuristic to Find Attribute Orders
Projection List, Most Interesting First
1. 92.01 : Flavanoids, Proline, Colour Intensity, Hue
2. 91.68 : Flavanoids, Hue, Alcohol, Proline
3. 91.04 : Flavanoids, Proline, Alcalinity of ash, Hue
4. 88.47 : Flavanoids, Hue, Alcalinity of ash, Alcohol
5. 88.36 : Flavanoids, Colour Intensity, Alcalinity of as
6. 87.71 : Proline, Colour Intensity, Hue, Total Phenol
7. 87.37 : Flavanoids, Colour Intensity, Alcalinity of as
8. 87.15 : Flavanoids, Proanthocynins, Alcohol, Proli
9. 87.09 : OD280/OD315, Proline, Colour Intensity, Alc
10. 86.83 : Proline, Hue, Alcalinity of ash, Alcohol
11. 86.80 : Flavanoids, Alcalinity of ash, Alcohol, Pro
12. 86.72 : Flavanoids, Alcalinity of ash, Manesium, F
Shown Details in Projections List
Rank
Predicted Accuracy
Instances
Finished evaluation (evaluated 4,862 projections in 8 min,
wine
W1
W2
W3
Colour Intensity
Flavanoids
Hue
Proline
3.080000
2.01
1.100
434
0.0
1.0

The **GRAND tour**

2D projections of an N-D space - choose suitable axes of rotation and an algorithm that ensures you explore all the space.
(The maths is complicated – See E. Wegman or Asimov

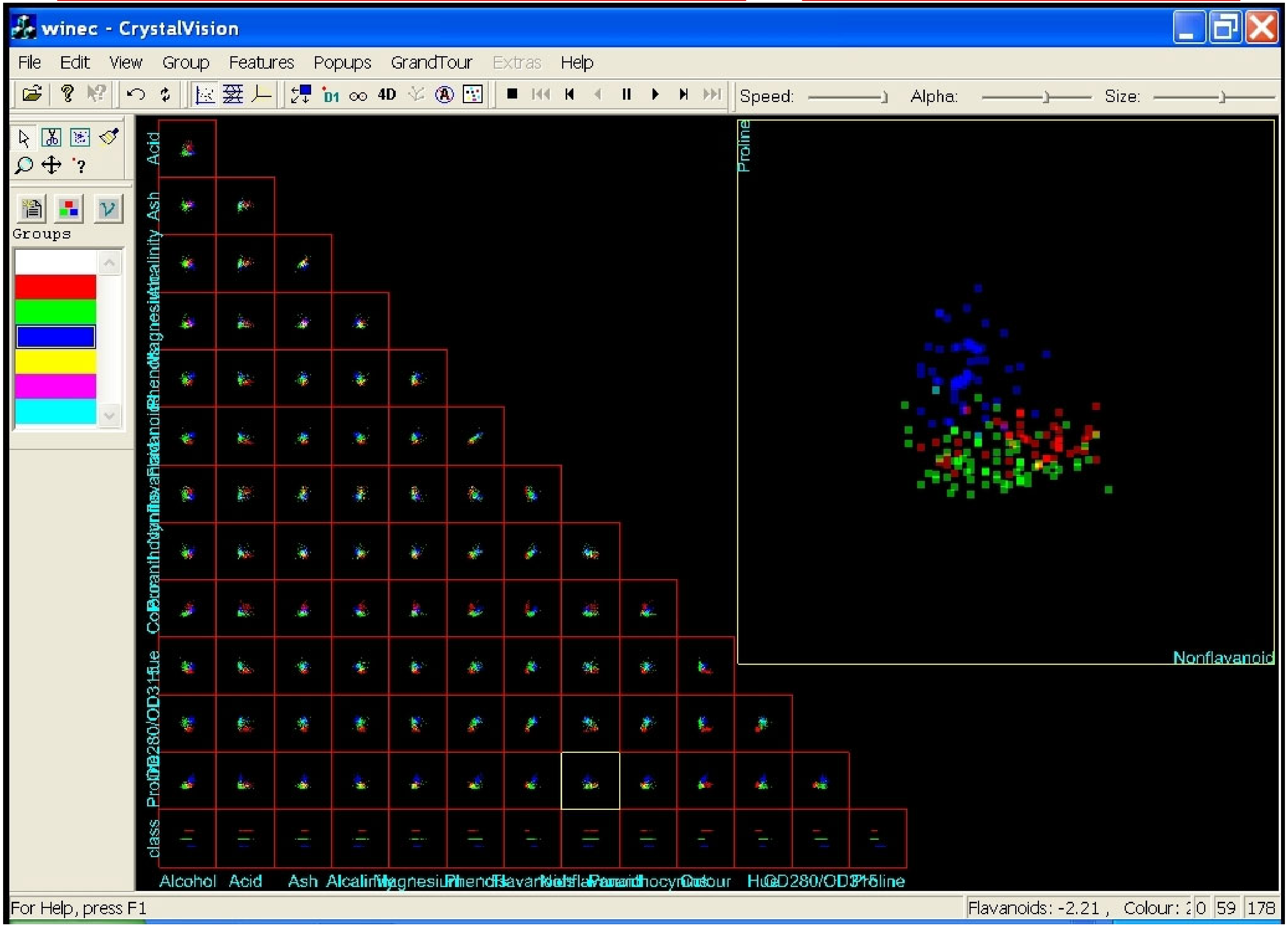
The Grand Tour via Geodesic Interpolation of 2-frames*

Daniel Asimov and Andreas Buja†
Report RNR-94-004, February 1994

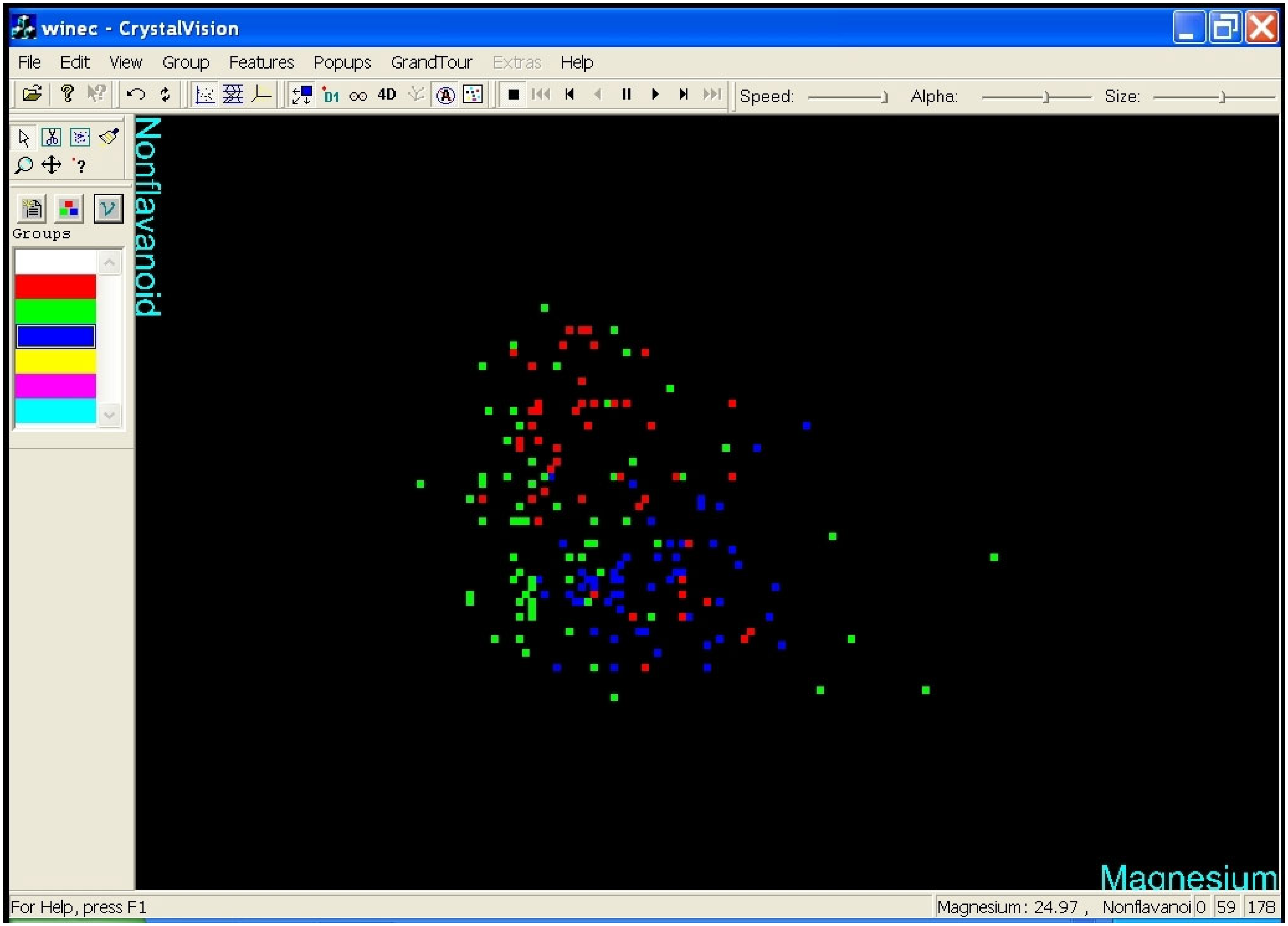
Facinating idea – useful for looking for clusters in data

Grand Tour of the wine dataset 1

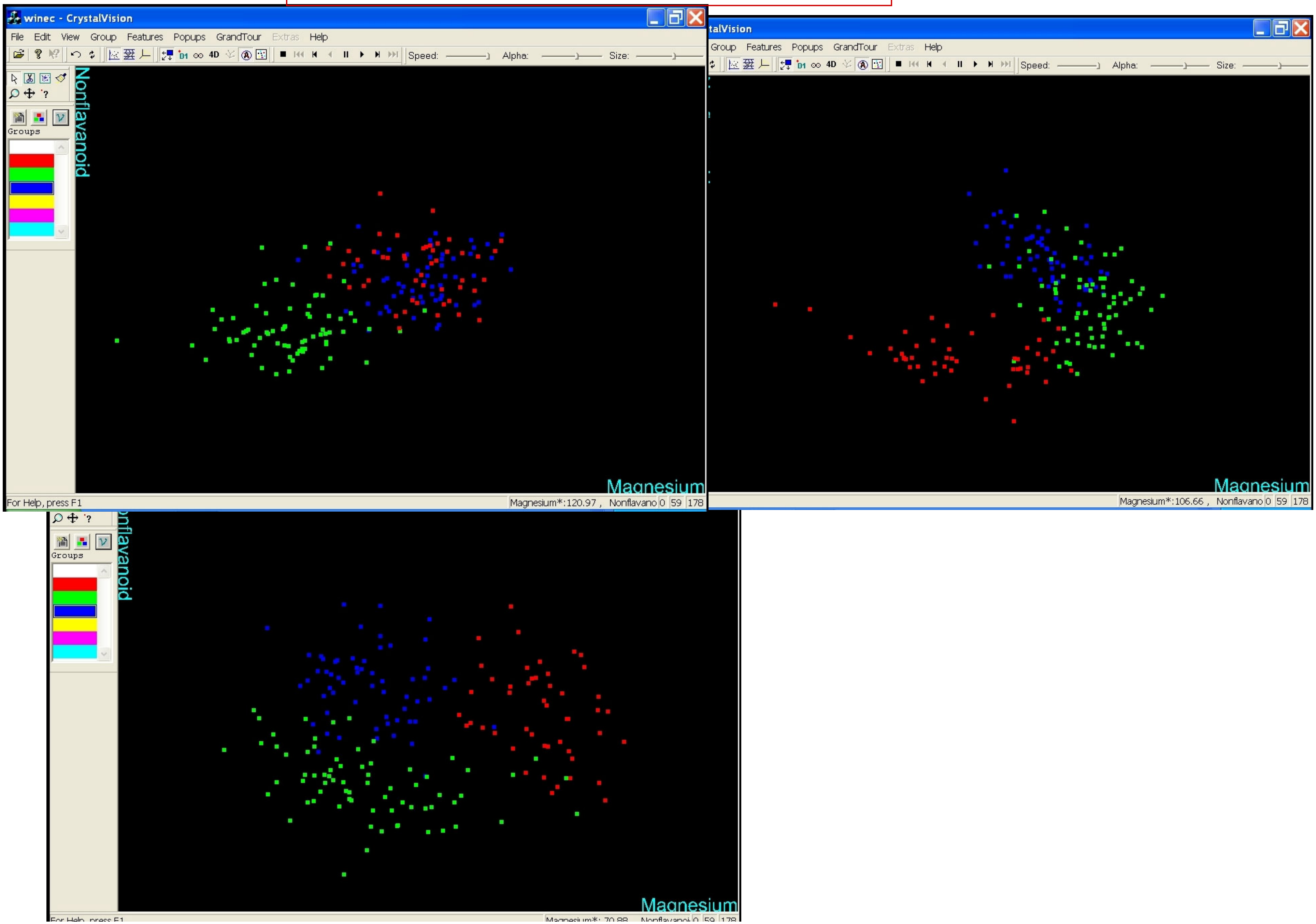
Invented by Asimov

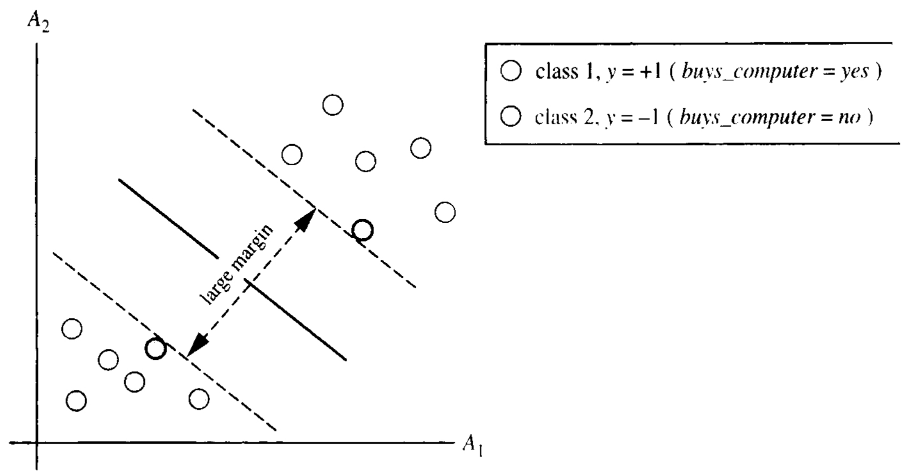


Grand Tour of the wine dataset 2



Grand Tour of the wine dataset 3





Support vectors. The SVM finds the maximum separating hyperplane, that is, the one v
maximum distance between the nearest training tuples. The support vectors are shown v
a thicker border.

Find hyperplane
that separates clusters

Use RBF when plane
will not work.

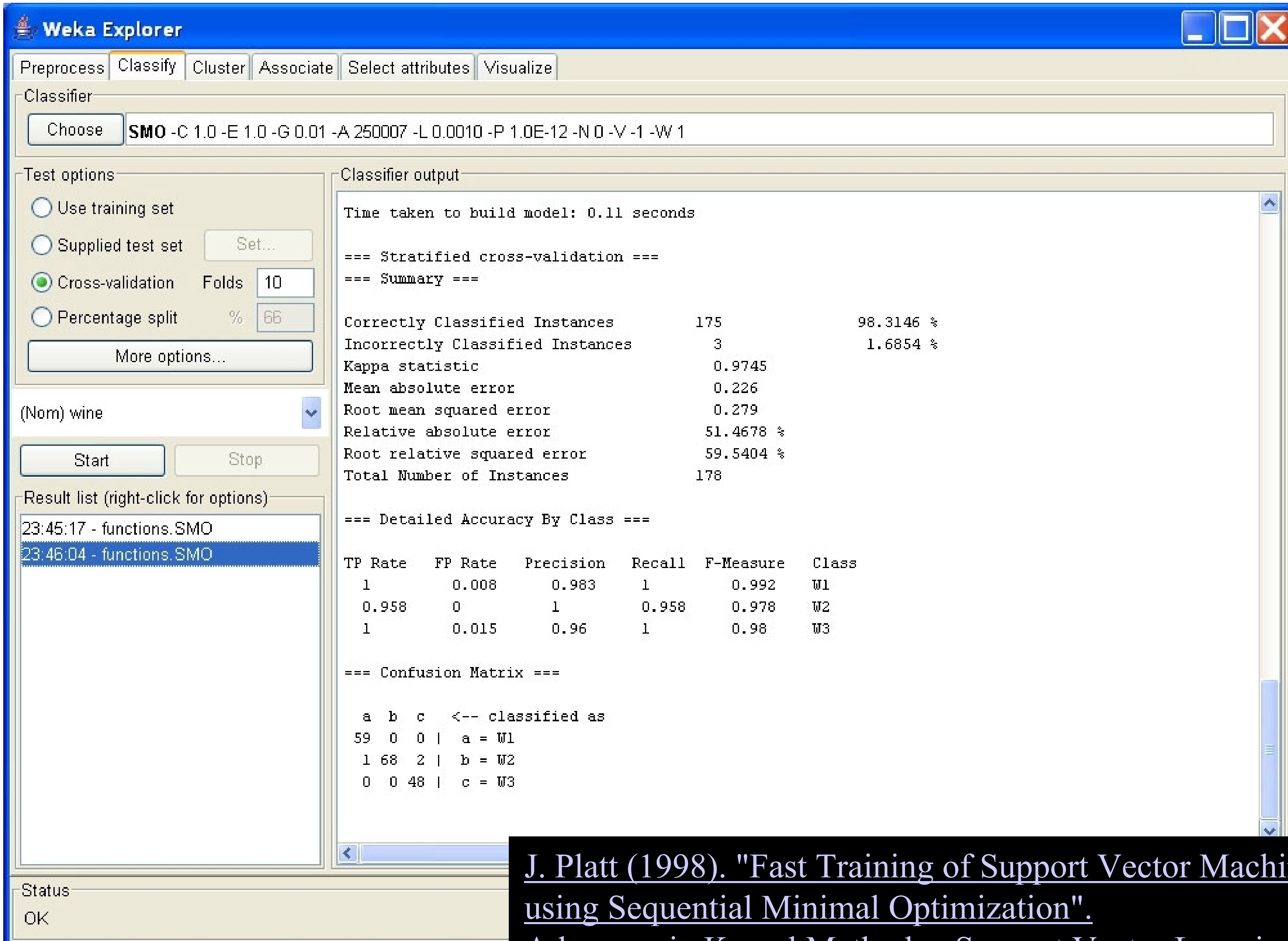
Support Vector Machine - rather confusing name.

For math. details see books referenced earlier.

http://www.csie.ntu.edu.tw/~cjlin/libsvm/

SVM-toy is a good introduction

SVM works well on this dataset - Grand Tour agrees



J. Platt (1998). "Fast Training of Support Vector Machines using Sequential Minimal Optimization". Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press.

Now lets try some other machine learning algorithms on the particle physics dataset…

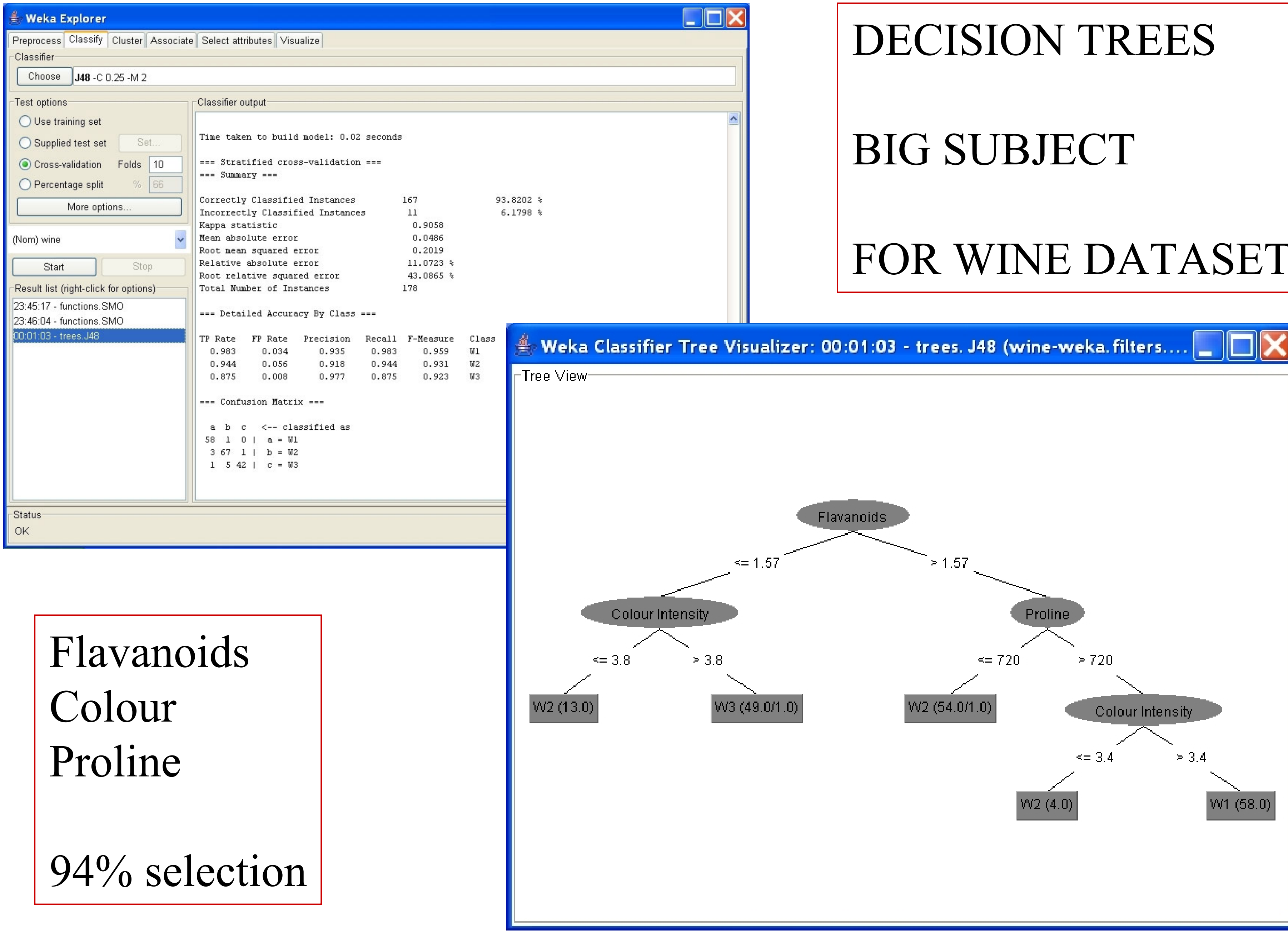
- Decision Tree C4.5
- VizRank
- SVM
- Neural Net with back propagation

SVM DOES NOT WORK WELL
GRAND TOUR EXPLAINS WHY.

DECISION TREES

BIG SUBJECT

FOR WINE DATASET

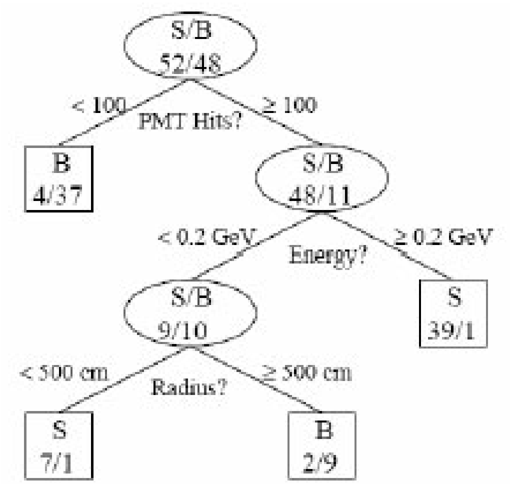


Flavanoids
Colour
Proline

94% selection

Decision Trees

Decision trees emerged in mid 80's: CART (Breiman, Friedman etc), C4.5 (Quinlan) etc



Criteria used for commercial trees

(p = fraction of correctly classified events)

$Q(p) = p$

$Q(p) = -2p(1-p)$ Gini index

$Q(p) = p\log p + (1-p)\log(1-p)$ cross-entropy

Split nodes recursively until a stopping criterion is satisfied.

Parent node with *W* events and correctly classified *p*W* events is split into two daughters nodes iff

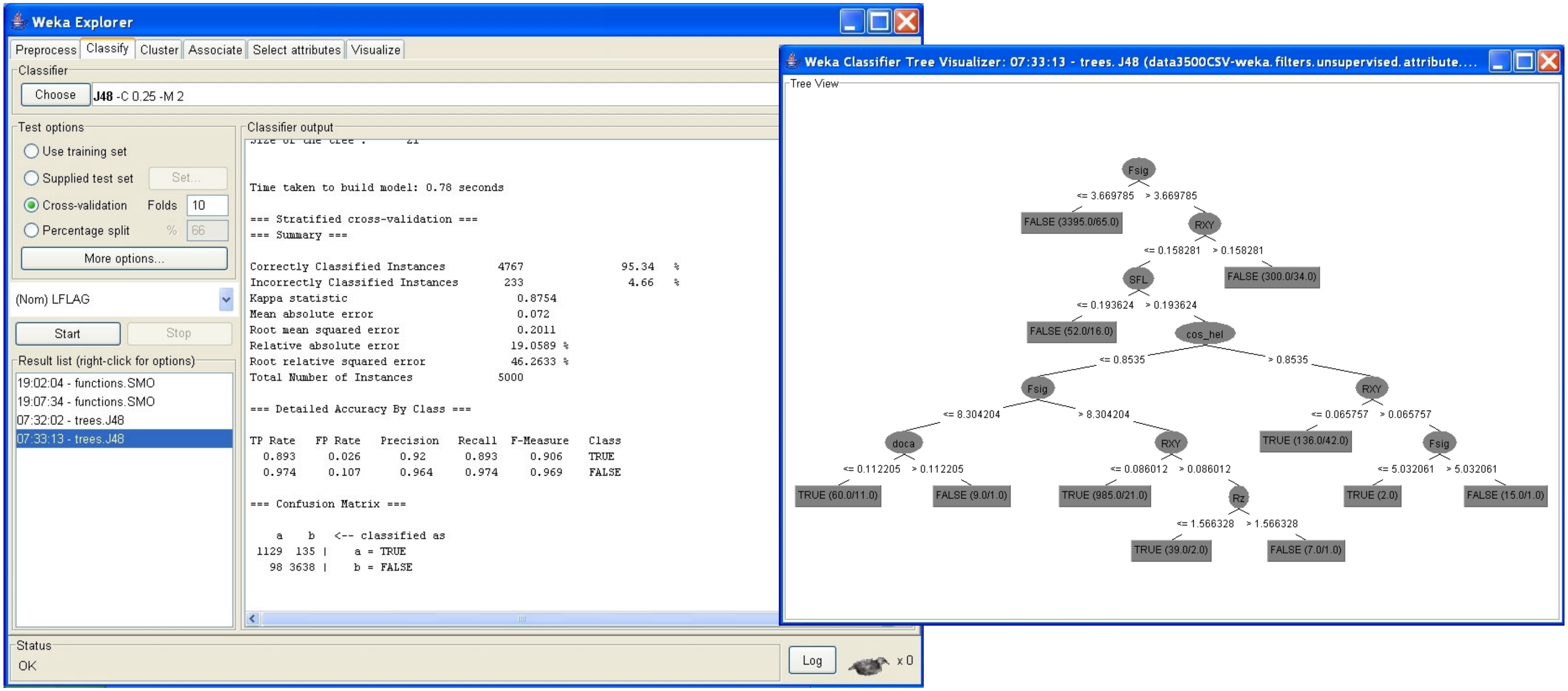
$$WQ(p) < W_1Q(p_1) + W_2Q(p_2)$$

Stopping criteria:

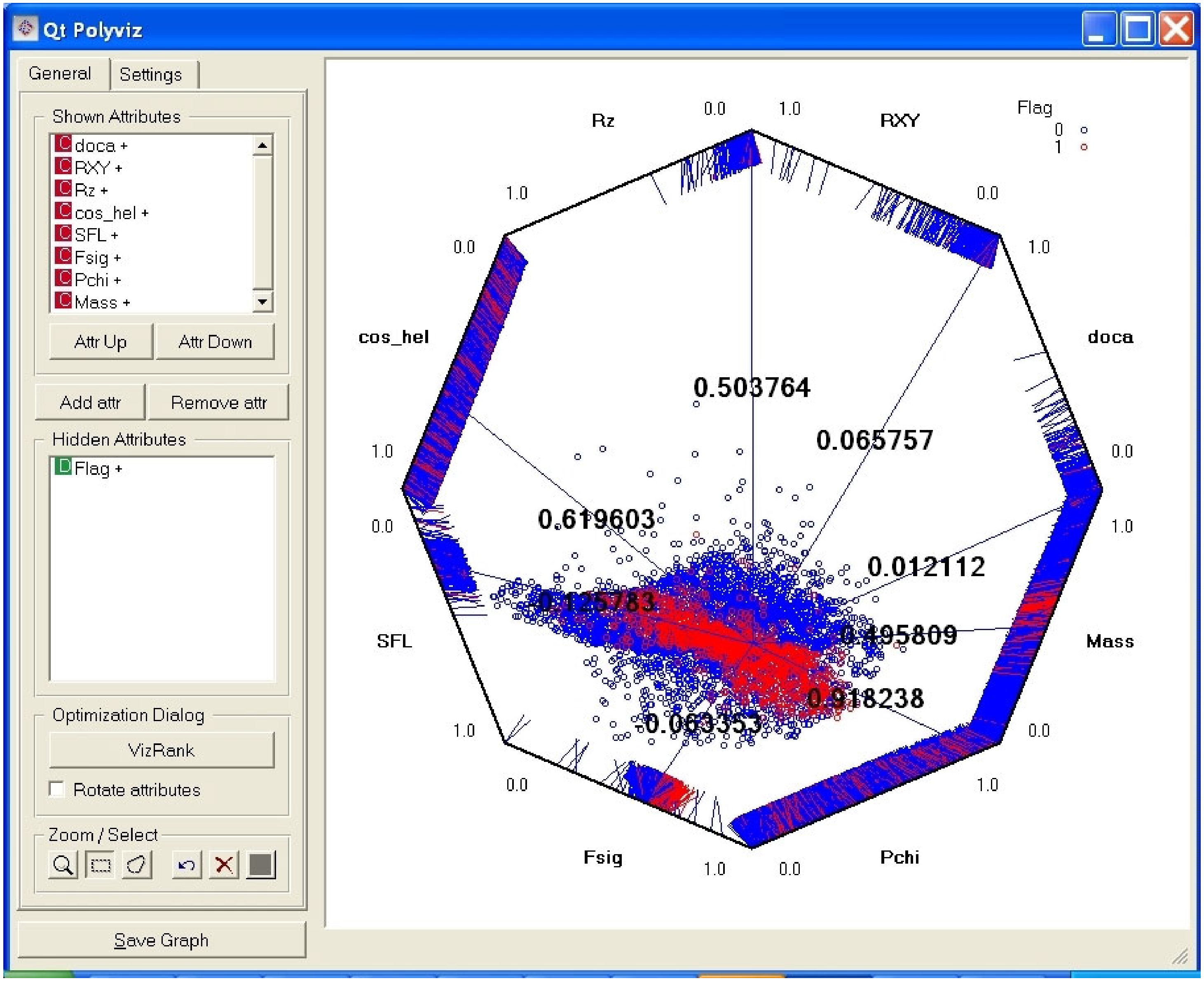
- **unable to find a split that satisfies the split criterion**
- **maximal number of terminal nodes in the tree**
- **minimal number of events per node**

Output of a decision tree is discrete: 1 if an event falls into a signal node, 0 otherwise.

Ilya Narsky Caltech Seminar Oct 2005.

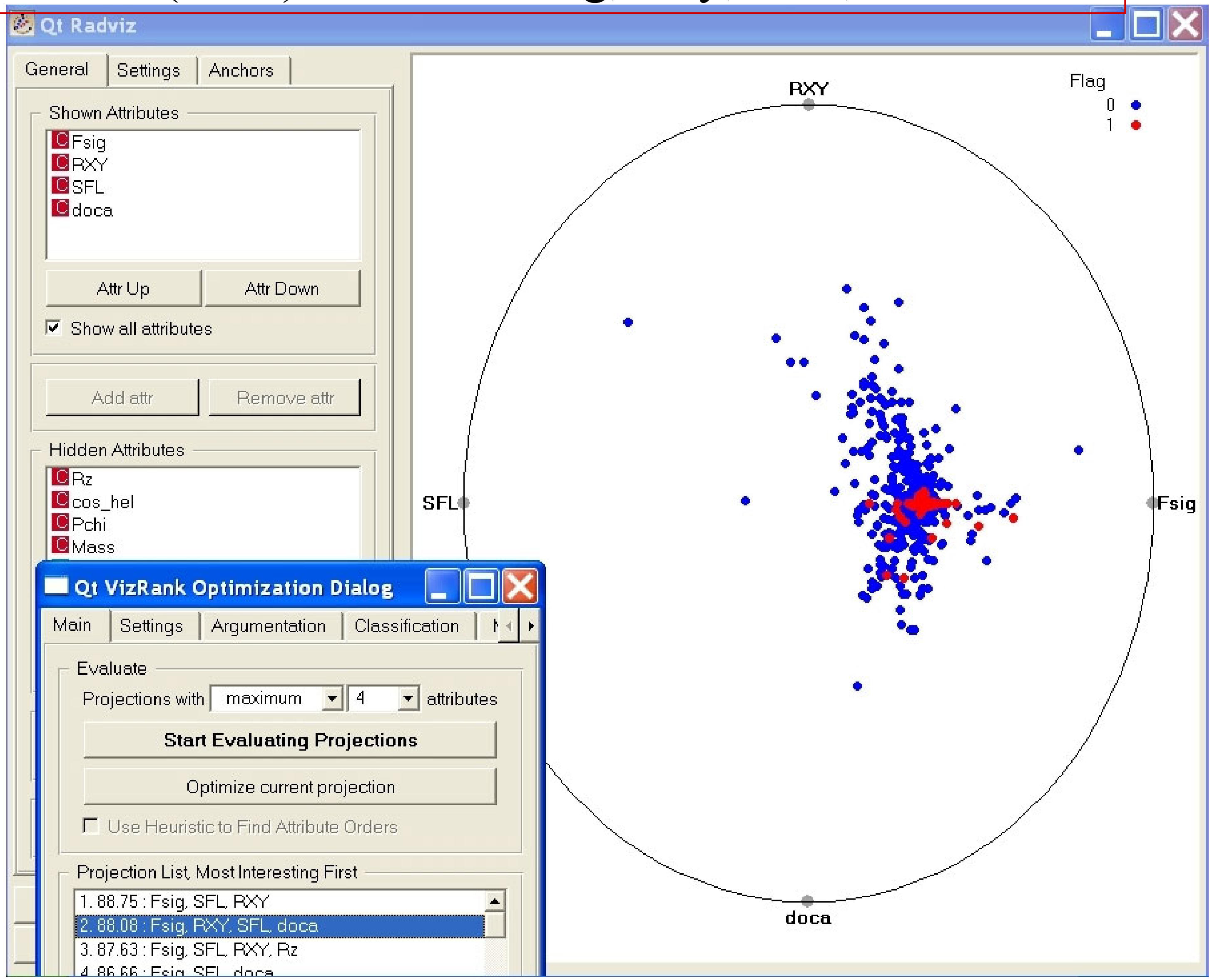


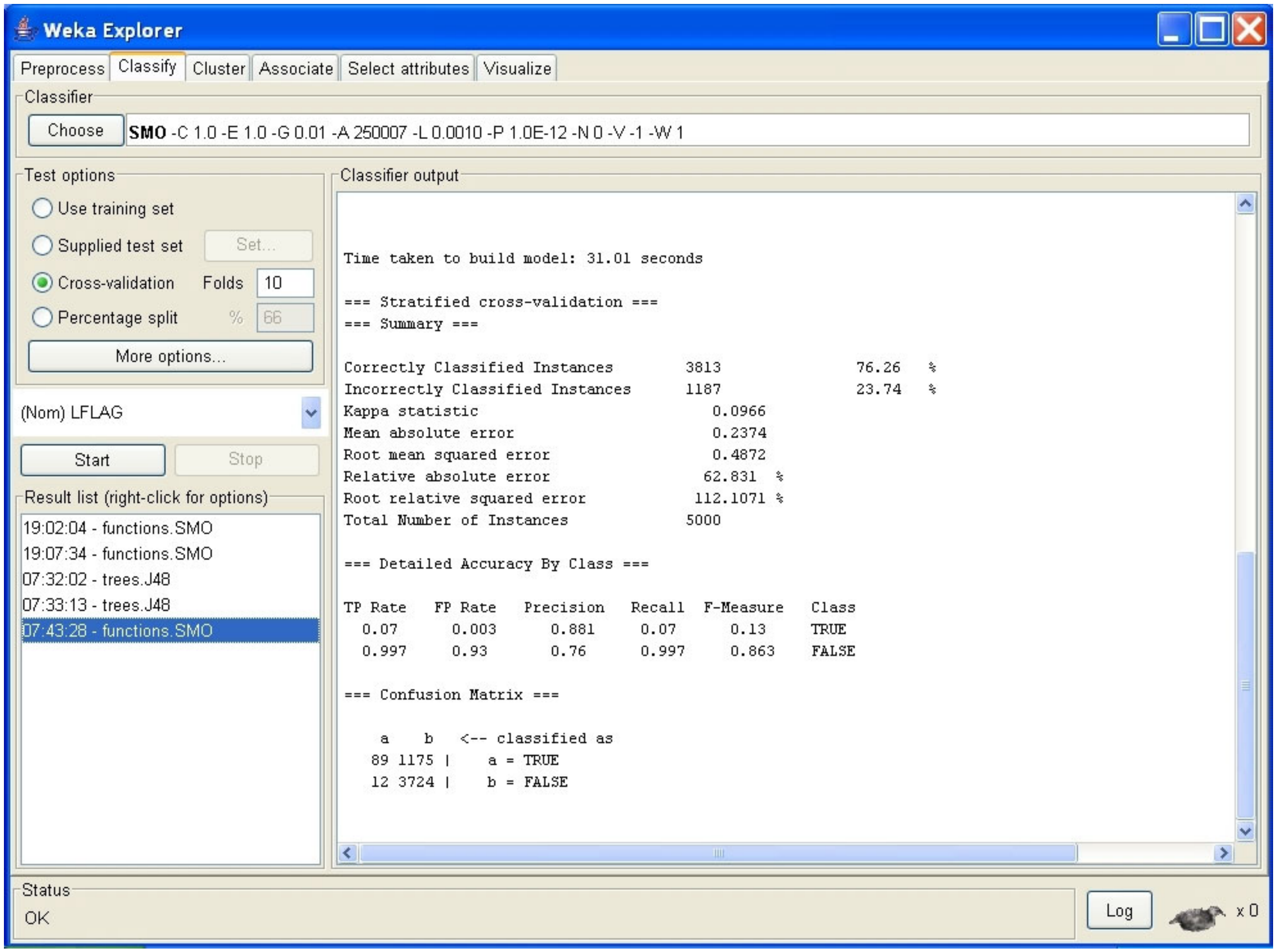
Key variables are: Fsig, Rxy, SFL, cos-hel, doca 95% selection
Remember the parallel coords analysis….



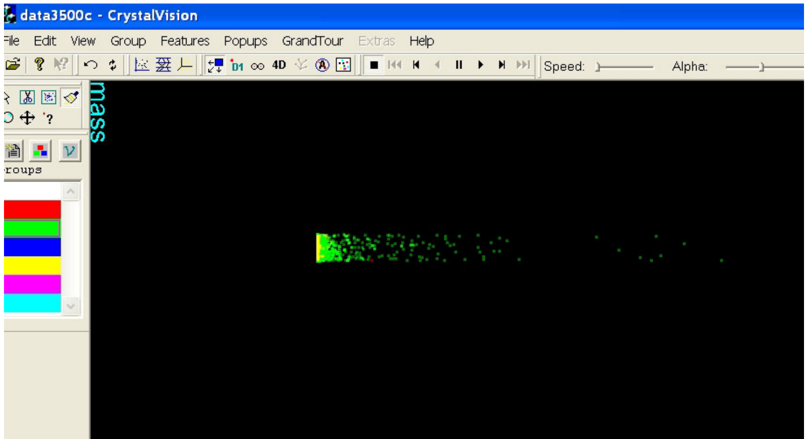
Useful way to see what variables matter…….

VizRank (kNN) Selects Fsig, Rxy, SFL, doca 88%





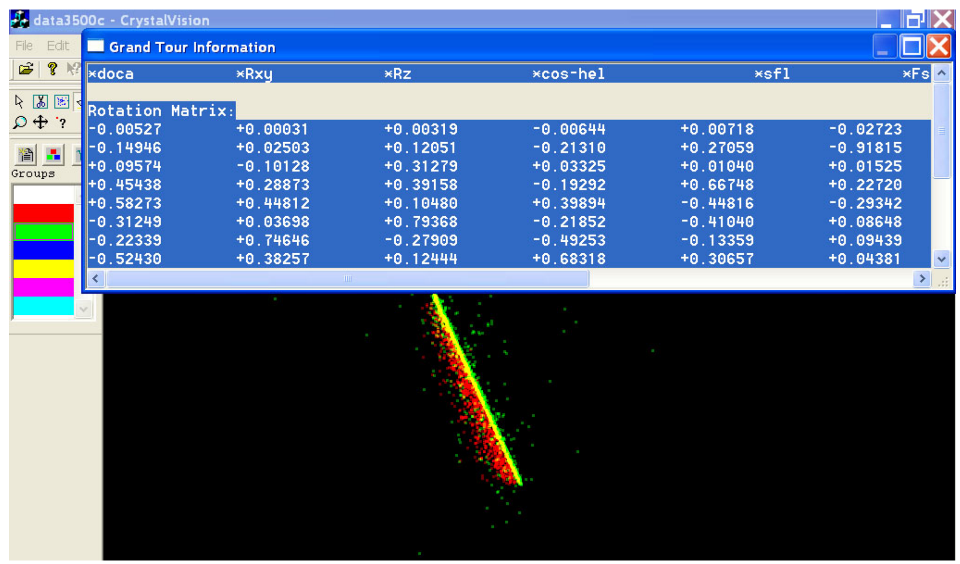
SVM (RBF not used) works very badly on this data
Grand Tour backs this conclusion up…….



Mass v Rxy
Standard Projection

CrystalVision
GrandTour

Cannot separate signal



SUMMARY

	SF	BF
S		
B		

Method			Accuracy
C4.5	1129	135	95%
	98	3638	
SVM	89	1175	76%
	12	3724	
NN with backpropagation	927	337	90%
	157	3579	
Bagging	1145	119	96%
	96	3640	

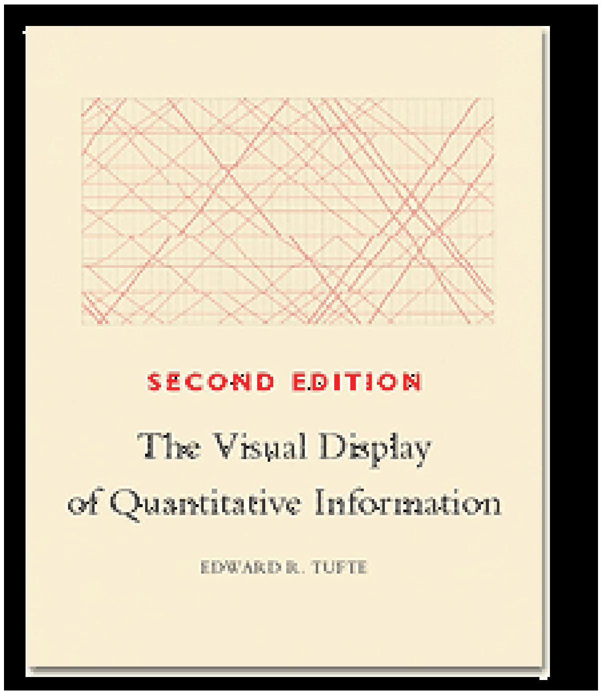
VizRANK 88%

Exploratory data analysis with crystalvision 968 S 44 B !!

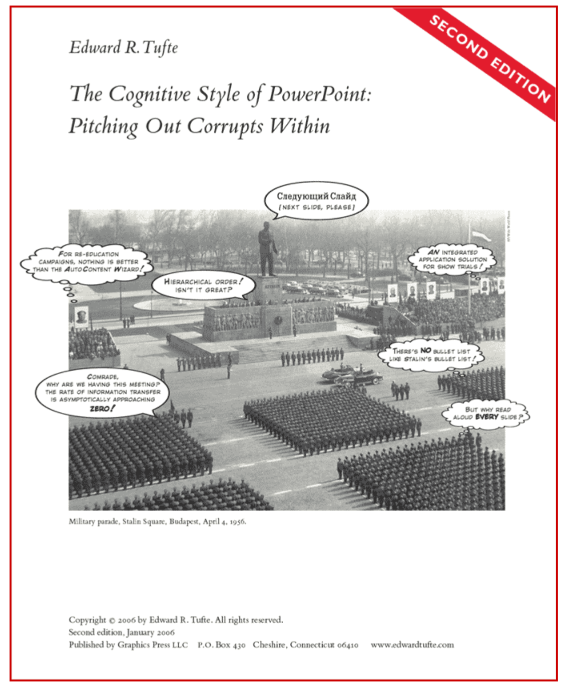
VISUAL DATA ANALYSIS HELPS ONE TO UNDERSTAND THESE RESULTS

SOME OTHER USEFUL DATA VISUALISATIONS

How can you display the data in a way that it can be easily understood ?



https://www.edwardtufte.com/tufte/



Should be read by all graduate students !

For last 100 years mathematicians have been suspicious of pictures and visual proofs……..perhaps this is changing……
Synthese Lib Vol 327, 2005

Visualisations of some mathematical surfaces

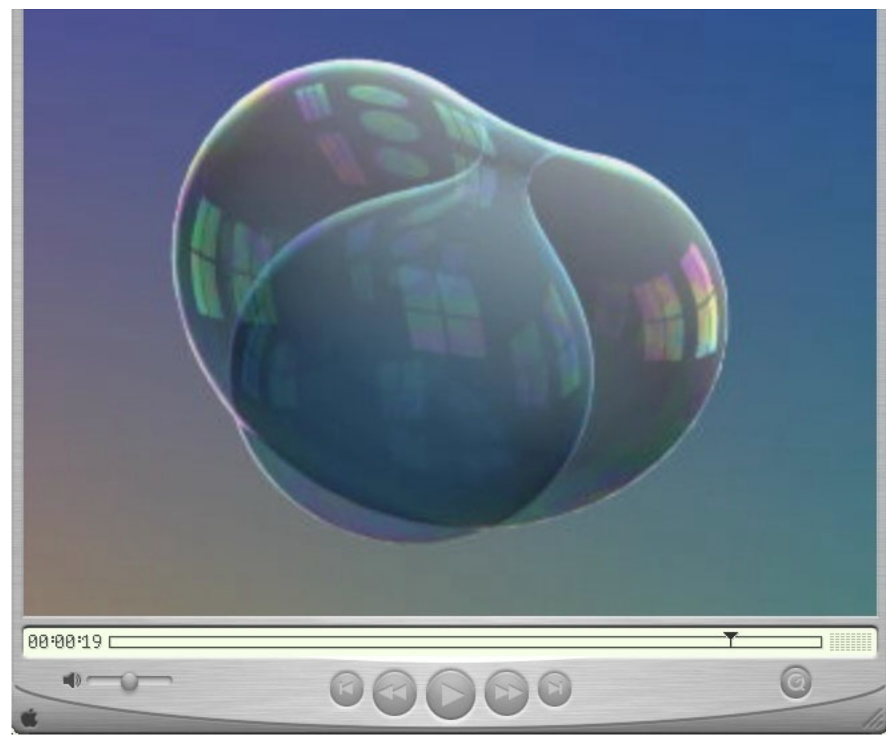
http://vmm.math.uci.edu/3D-XplorMath/Surface/gallery.html

Sphere Does Elegant Gymnastics in New Video,
Dana Mackenzie. Science 281:5377 (July 1998) 634-635.

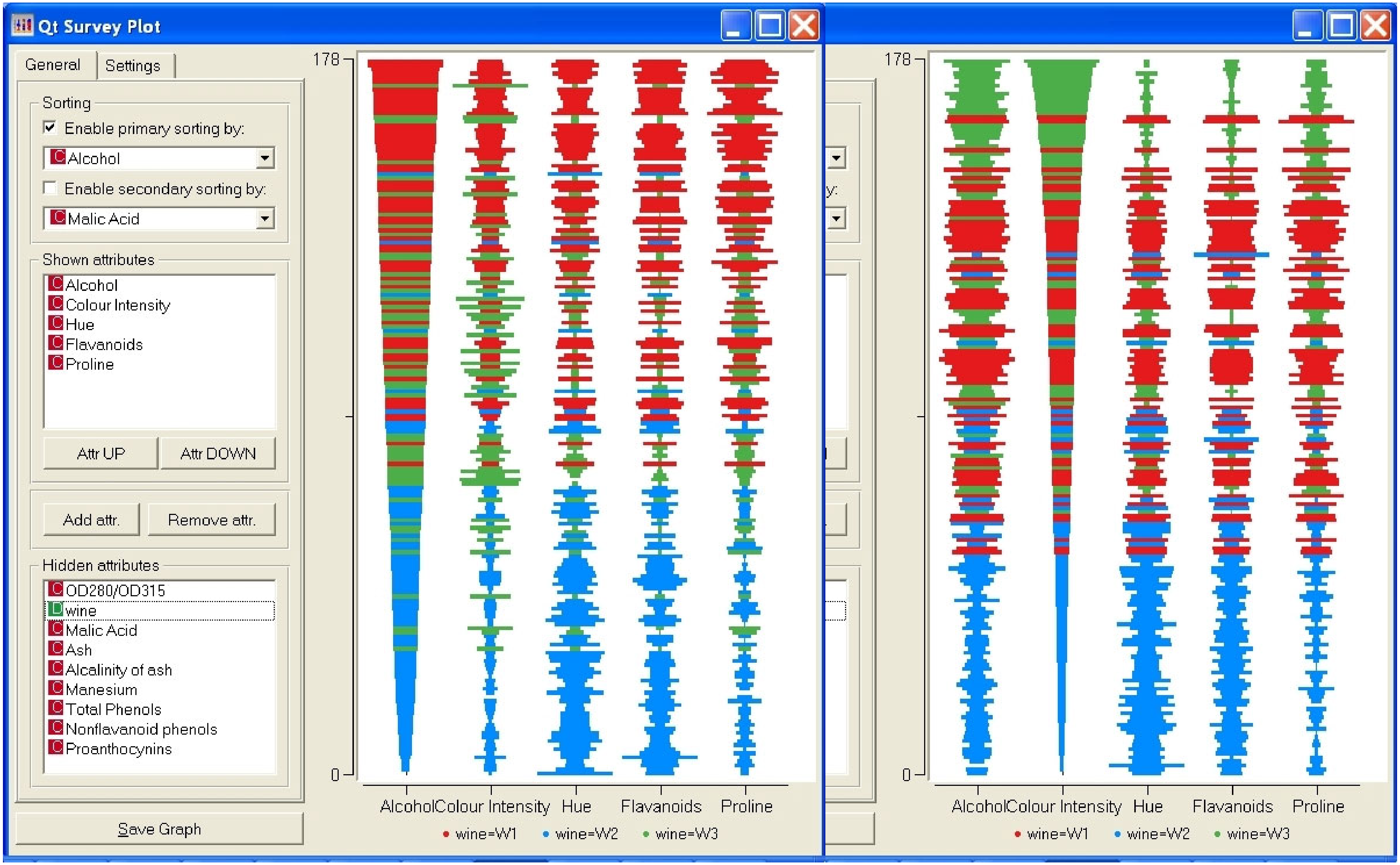
Is it possible to turn a sphere inside out without tearing or creasing it? The answer is yes.

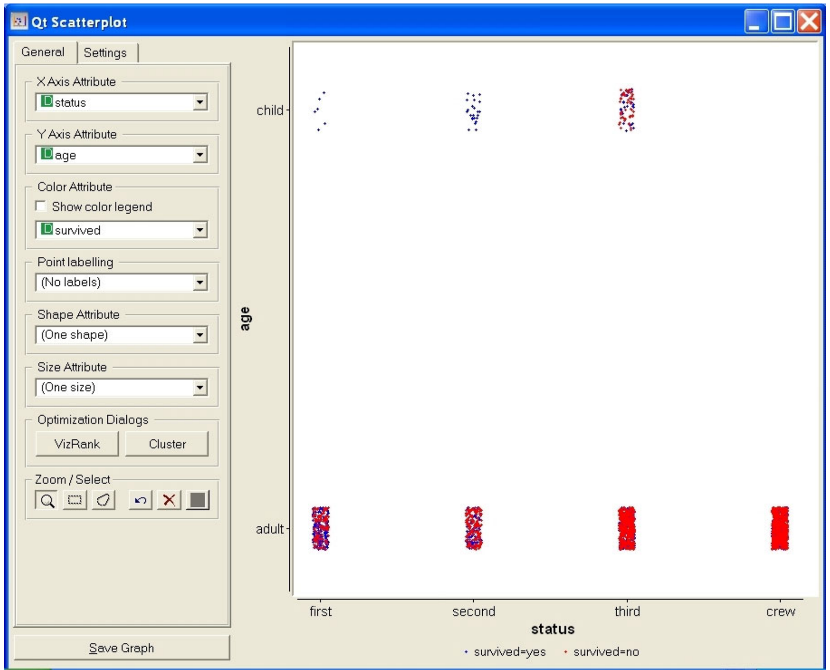
The proof, by Stephen Smale, is over forty years old but until now there has not been a satisfying demonstration. In this interesting article, the author explains how George Francis and John Sullivan at the University of Illinois were able to use a surface created by Robert Kusner in 1983 to create a 6 /2 minute computer animation of the eversion process. During the process, it is necessary to keep the energy level (defined so that the energy increases as more bending takes place) at a minimum at each stage. It was noted that the surface Kusner created had an energy level that would make it a candidate for the halfway point ofthe eversion process. Once the halfway point was known, it seemed possible to go backward to the sphere and forward to the sphere turned inside out. Francis and Sullivan showed with their animation that this is indeed the case. Their video has been shown recently at the Siggraph 98 convention in Orlando and the International Congress of Mathematicians in Berlin.

http://new.math.uiuc.edu/optiverse/



SURVEY PLOT - WINE DATA





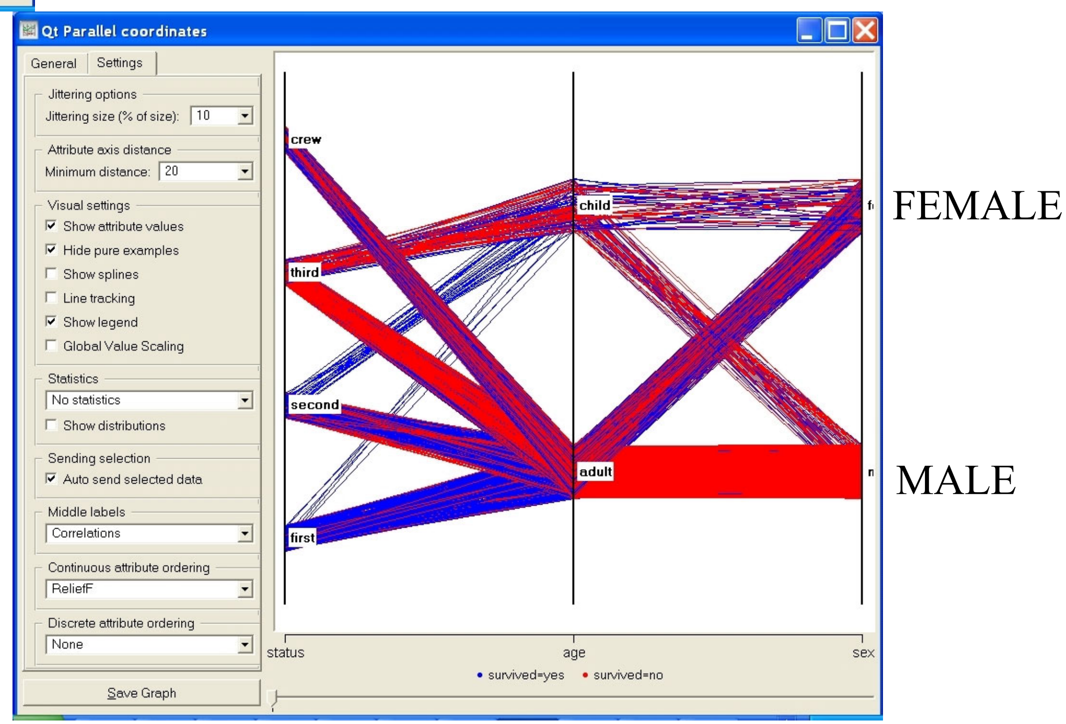
SURVIVED

RED = NO

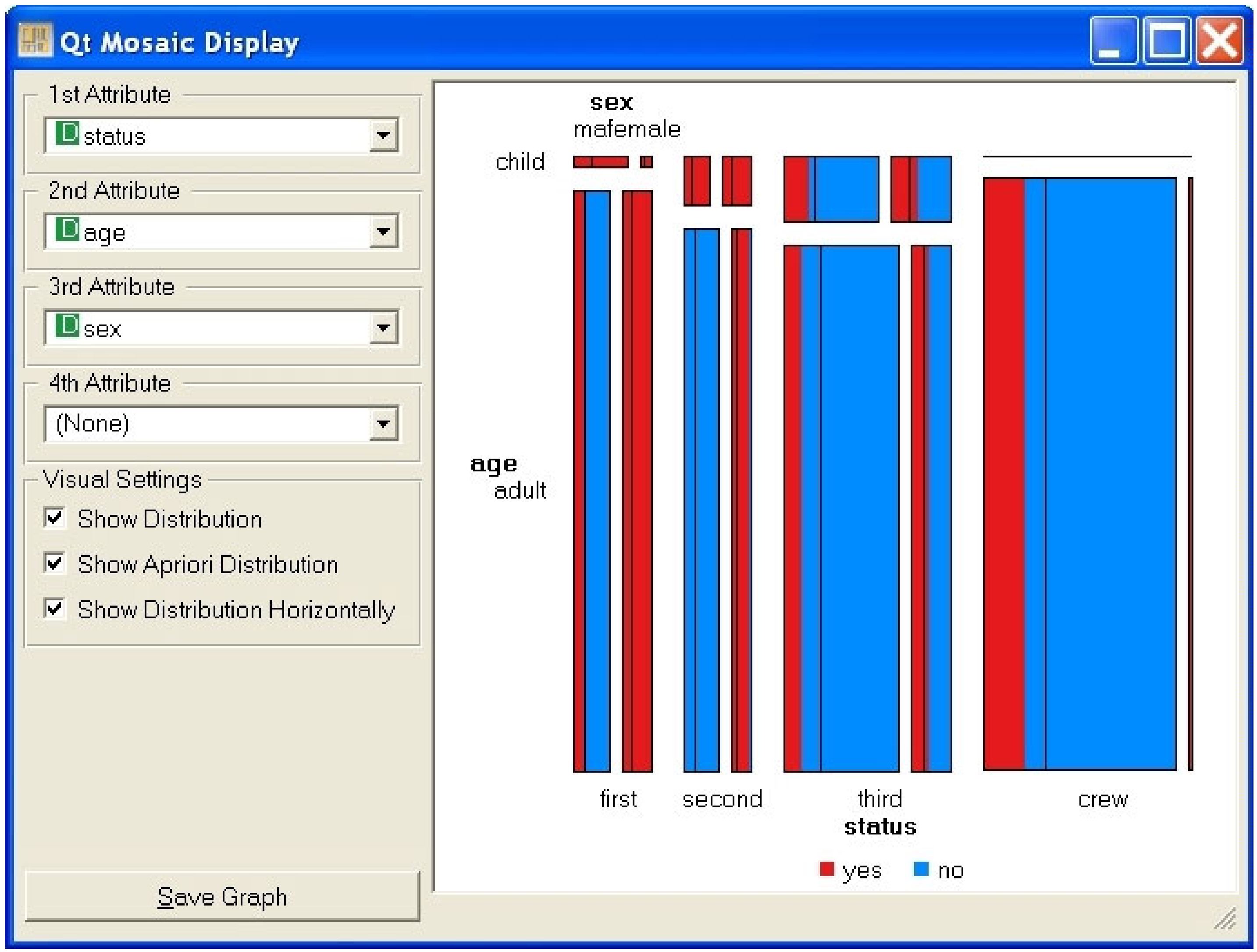
BLUE = YES

Data on Titanic Disaster

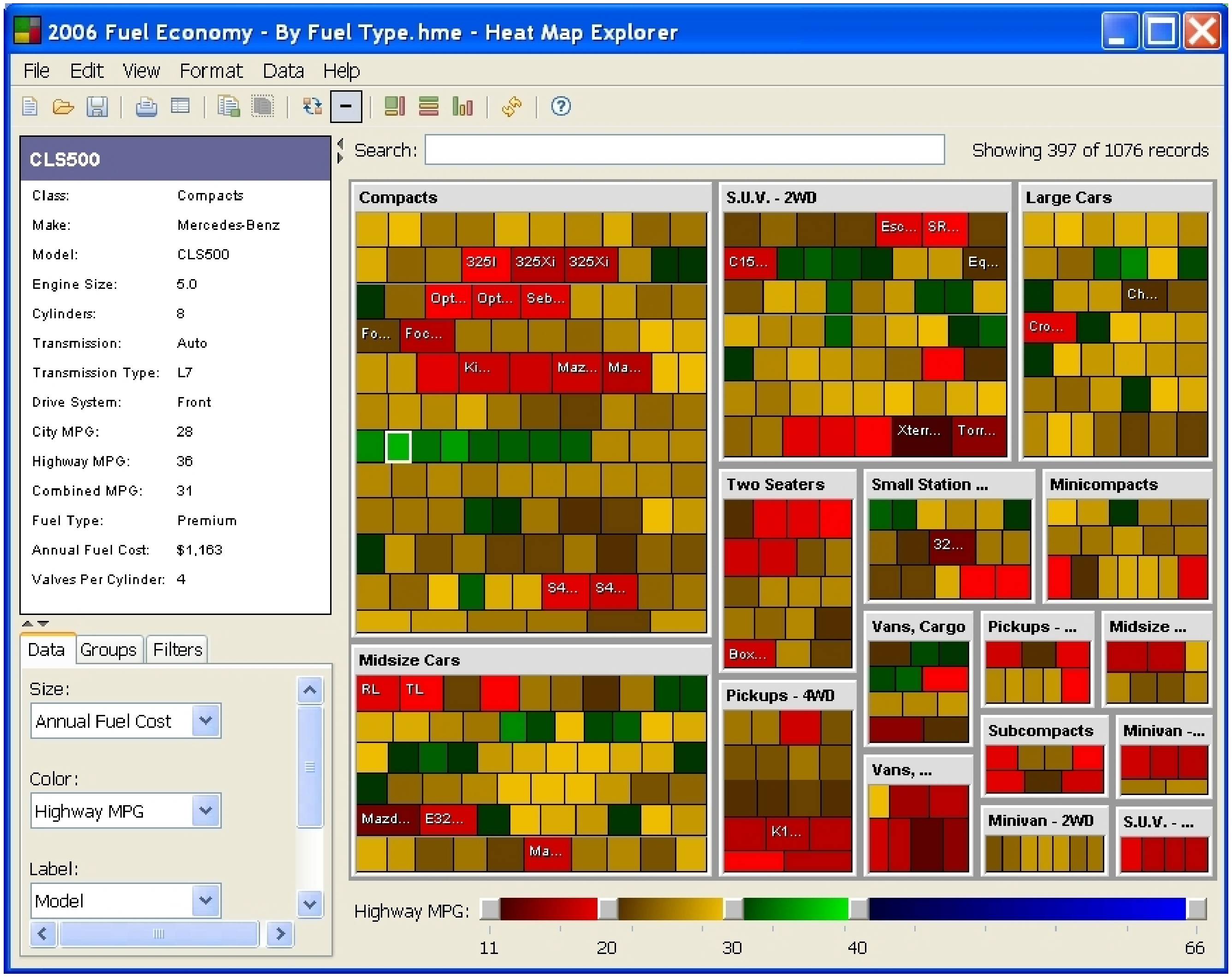
Scatterplot and parallel coords do not immediately tell you who and who did not survive



MOSAIC DISPLAY WORKS VERY WELL WITH THIS DATA

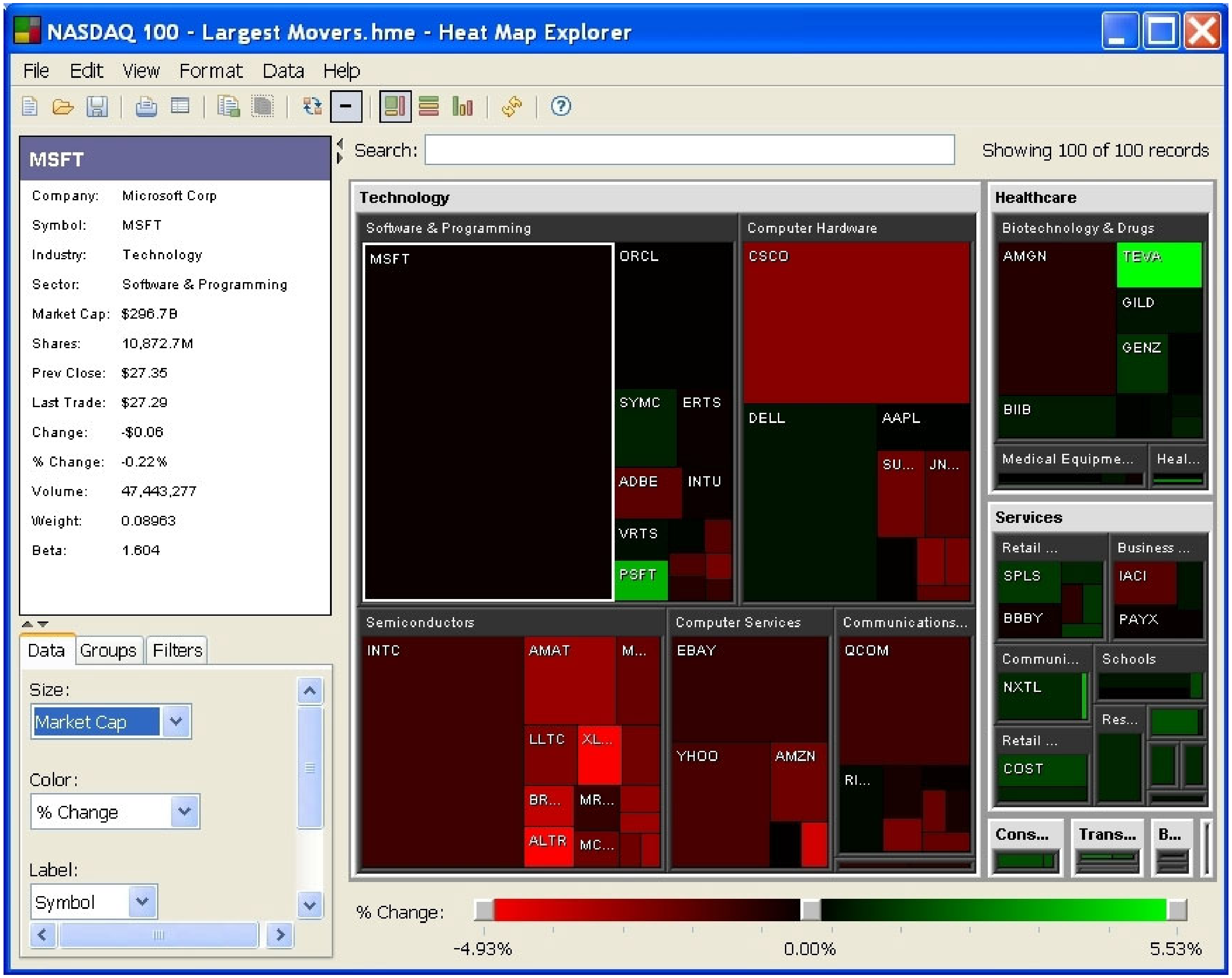


HEAT MAPS - invented by USA professor who wanted to visualise the data on his hard disk



Has drill down procedure

Excellent for looking at a large amount of data quickly – online histograms ?



Useful for stock brokers…………………

NOW ITS YOUR TURN….

- **WHERE TO START**
- **PRACTICAL ADVICE**

Software	Site	Comment
CrystalVision	ftp://www.galaxy.gmu.edu/pub/	Windows. ExplorN Unix α-channel. GT, PC Needs development.
GGobi	www.ggobi.org	No α-channel.GT, PC All Platforms. Access to R.
Mondrian	http://stats.math.uni-augsburg.de/Mondrian/	Java. α-channel.
Visulab	http://www.inf.ethz.ch/personal/hinterbe/Visulab/	Excel plugin. PC only
Orange	http://www.ailab.si/orange	Component based data mining. C++ and python scripting. PC.
WEKA	http://www.cs.waikato.ac.nz/ml/weka/	Java based data mining package Large no of algorithms included.
Datadesk	http://www.datadesk.com/	Commercial. Linked plots. Stats.
Statistica	http://www.statsoft.com/	Commercial. Very powerful. Not evaluated yet. Graphics + Stats.
VisualExplorer	www.curvaceous.com	Commercial. PC for process control Excel PlugIn.

When to use a particular visualisation ???

Benchmark Development for the Evaluation of Visualization for Data Mining

Georges G. Grinstein[1*], Patrick Hoffman[1*], Sharon J. Laskowski[2], Ronald M. Pickett[1]

[1]Institute for Visualization and Perception Research
University of Massachusetts at Lowell, Lowell, MA 01854
{grinstein, phoffman, pickett}@cs.uml.edu

[2]The National Institute of Standards and Technology, Gaithersburg, MD 20899
sharon.laskowski@nist.gov

Table 8 Scatter Plot Matrix

TASK / DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons							
Balloons-flattened							
Lenses							
Lenses-flattened							
O-rings	Y	Y	Y	Y	Y	Y	
Monks1-training							
Iris	Y	Y	Y		Y	Y	
Congress							
Liver	Y	Y					
Cars	Y	Y	Y		Y	Y	
Wine	Y	Y	Y		Y	Y	

Table 6 Survey Plot

TASK / DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons				Y	Y		Y
Balloons-flattened				Y	Y		Y
Lenses					Y	Y	
Lenses-flattened					Y	Y	
O-rings	Y			Y	Y	Y	
Monks1-training				Y	Y	Y	Y
Iris	Y	Y	Y	Y	Y	Y	
Congress				Y	Y		
Liver							
Cars				Y	Y	Y	
Wine				Y	Y	Y	

Table 4 Parallel Coordinates

TASK / DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons							
Balloons-flattened							
Lenses							
Lenses-flattened							
O-rings	Y	Y	Y	Y	Y	Y	
Monks1-training		Y		Y	Y	Y	Y
Iris	Y	Y	Y	Y	Y	Y	
Congress							
Liver	Y	Y					
Cars	Y	Y	Y		Y	Y	
Wine	Y	Y	Y		Y	Y	

Table 5 Radviz

TASK / DATA SET	See Outliers	See Clusters	Find Class Clusters	See All Important Features	See Some Important Features	See Possible Rule/Model	See Exact Rule/Model
Balloons		Y	Y			Y	
Balloons-flattened		Y	Y		Y	Y	
Lenses		Y	Y		Y		
Lenses-flattened		Y	Y		Y		
O-rings	Y	Y	Y	Y	Y	Y	
Monks1-training		Y					
Iris	Y	Y	Y			Y	
Congress	Y	Y	Y		Y		
Liver	Y	Y					
Cars	Y	Y	Y		Y	Y	
Wine	Y	Y	Y		Y	Y	

WITH THE RIGHT TOOLKIT

(WEKA/ORANGE/CRYSTALVISION ARE VERY POWERFUL STARTING POINTS)

Visualise the data

Apply various data mining techniques to your data

Compare the results and then use your own judgement on how to proceed

OCCAMS RAZOR is a good principle

WARNING ….VIZ-O-Matic

The dangers of Glitziness

CONCLUSIONS

How was this seminar researched ????
Google – data mined the Internet
Books and papers !! The book is not dead yet.

There are powerful visualisation and machine learning techniques available. This field is rapidly growing.

USE THEM IN YOUR RESEARCH !!!!!!!!

WE SHOULD BUILD THEM INTO OUR DATA ANALYSIS SOFTWARE

STANDARD HEP REPOSITORY OF DATA SAMPLES TO EVALUATE NEW METHODS

Can we data mine particle physics data without a priori ideas of the physics we wish to see ????????????????

FINAL THOUGHT…….

IDEAS GUIDE YOU WHEN YOU CANNOT SEE.