Brunel
UNIVERSITY
WEST LONDON

Application of data visualisation techniques in particle physics

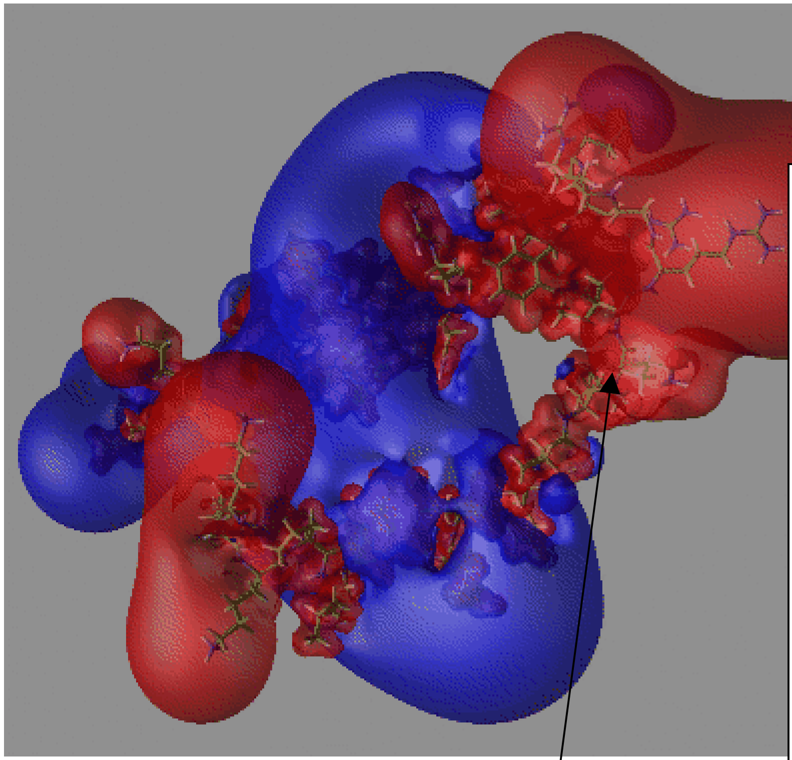
Steve Watts
Particle Physics Group and BITlab
School of Engineering and Design
Brunel University, West London, UK

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

Talk at - Computing in High Energy and Nuclear Physics, 13-17 February 2006, Mumbai, India

Examples of scientific visualisation

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

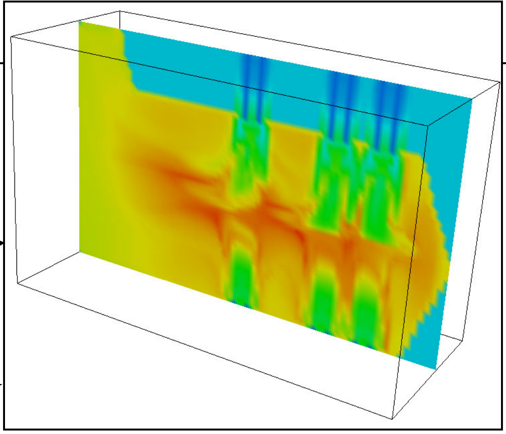


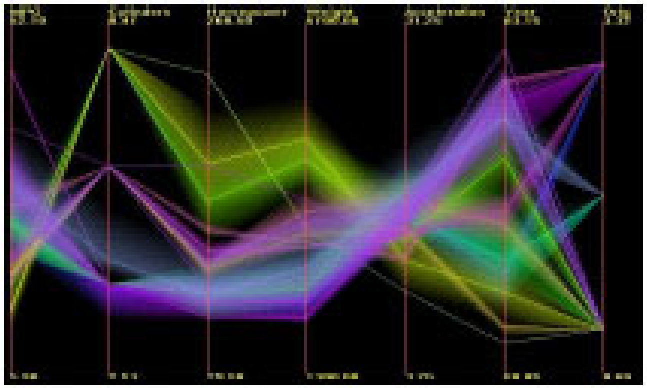
Contributors: Matt Challacombe and Eric Schwegler

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.





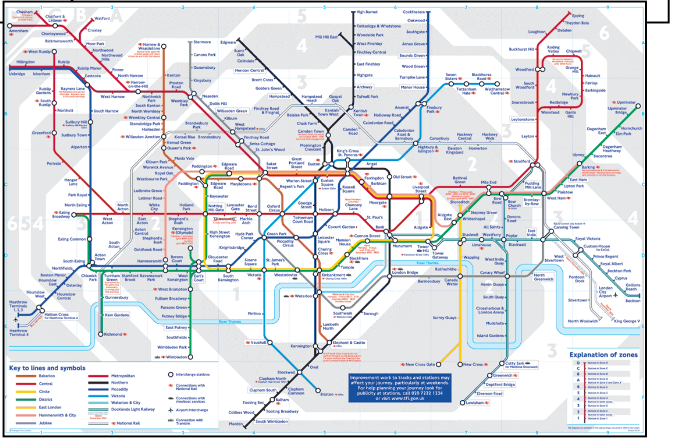
Information Visualization image shown courtesy of Matt Ward of Worcester Polytechnic Institute (WPI).

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 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 .. This is a short talk!**

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, and also the Grand Tour.

There are many other ideas - but these techniques are very powerful

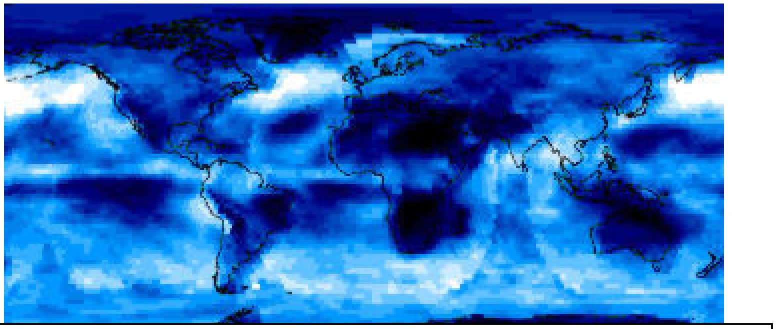
Data Exposition 2006

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

American Statistical Association

Have challenges each year

This is the 2006 one



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

- **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?

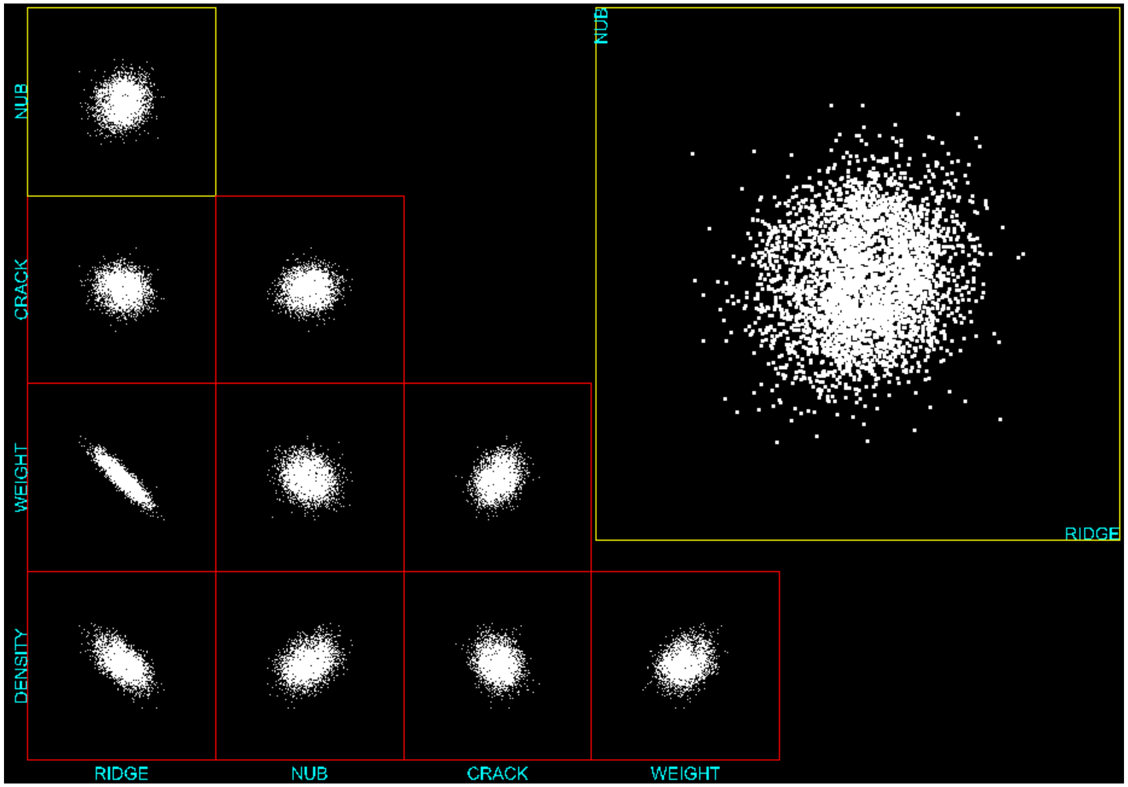
Data Visualisation Software

CrystalVision - E. Wegman

GGobi

XmdvTool

Orange



Note:
Size of dots matters!

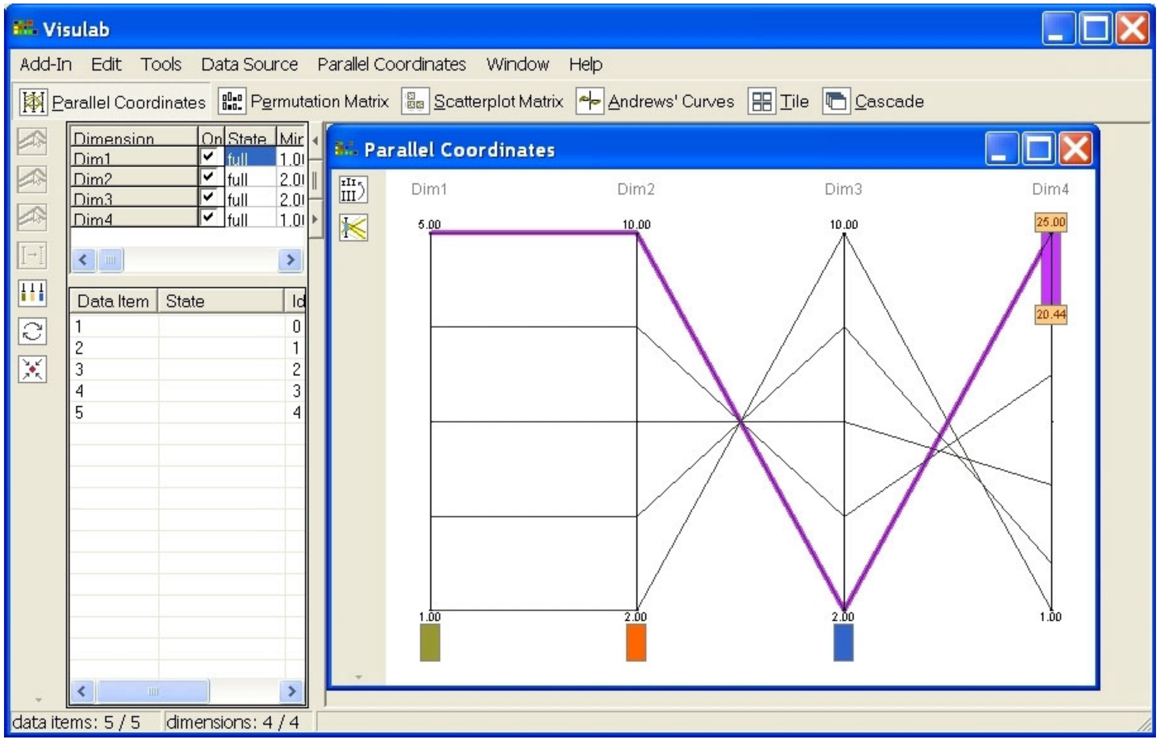
The pollen data - this is called a scatter matrix.
2D projections of this 5 variable space helps - but -

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/



Dimension	On	State	Mir
Dim1	✓	full	1.0
Dim2	✓	full	2.0
Dim3	✓	full	2.0
Dim4	✓	full	1.0

Data Item	State	Id
1		0
2		1
3		2
4		3
5		4

data items: 5 / 5 dimensions: 4 / 4

This also shows the idea of brushing

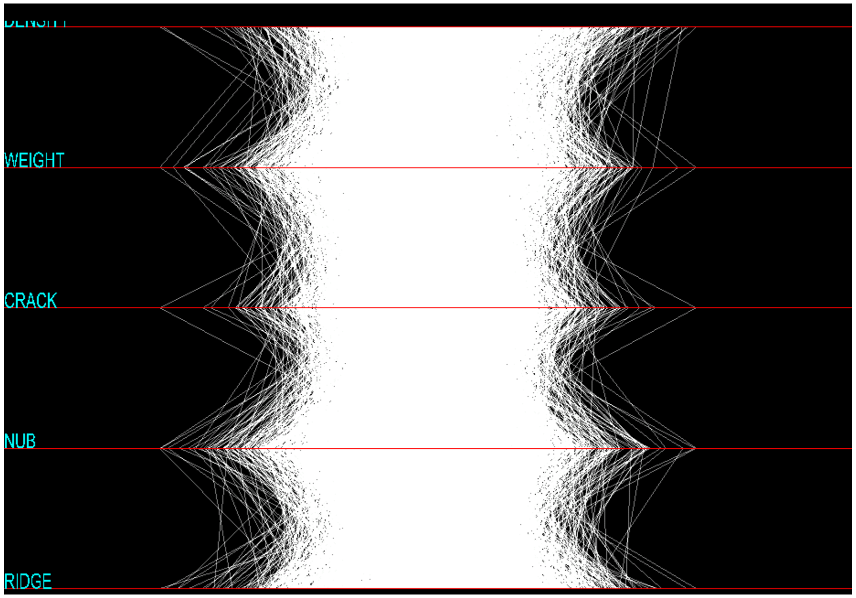
In graphics, a portion of each pixel's data that is reserved for transparency information. 32-bit graphics systems contain four channels -- three 8-bit channels for red, green, and blue (RGB) and one 8-bit alpha channel. The **alpha** channel is really a mask -- it specifies how the pixel's colors should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

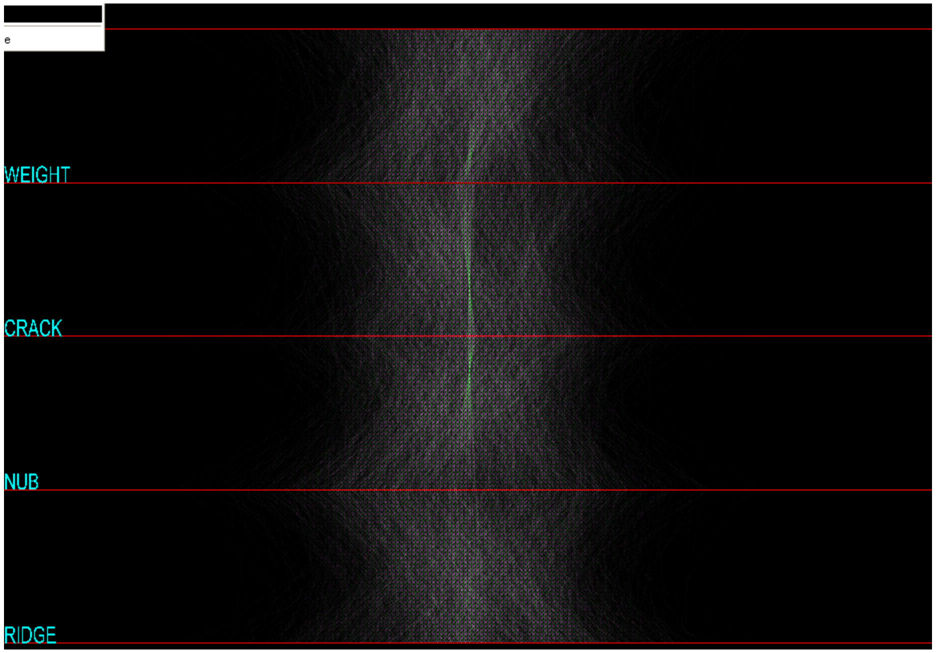
2) Now parallel coordinates.

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



High alpha

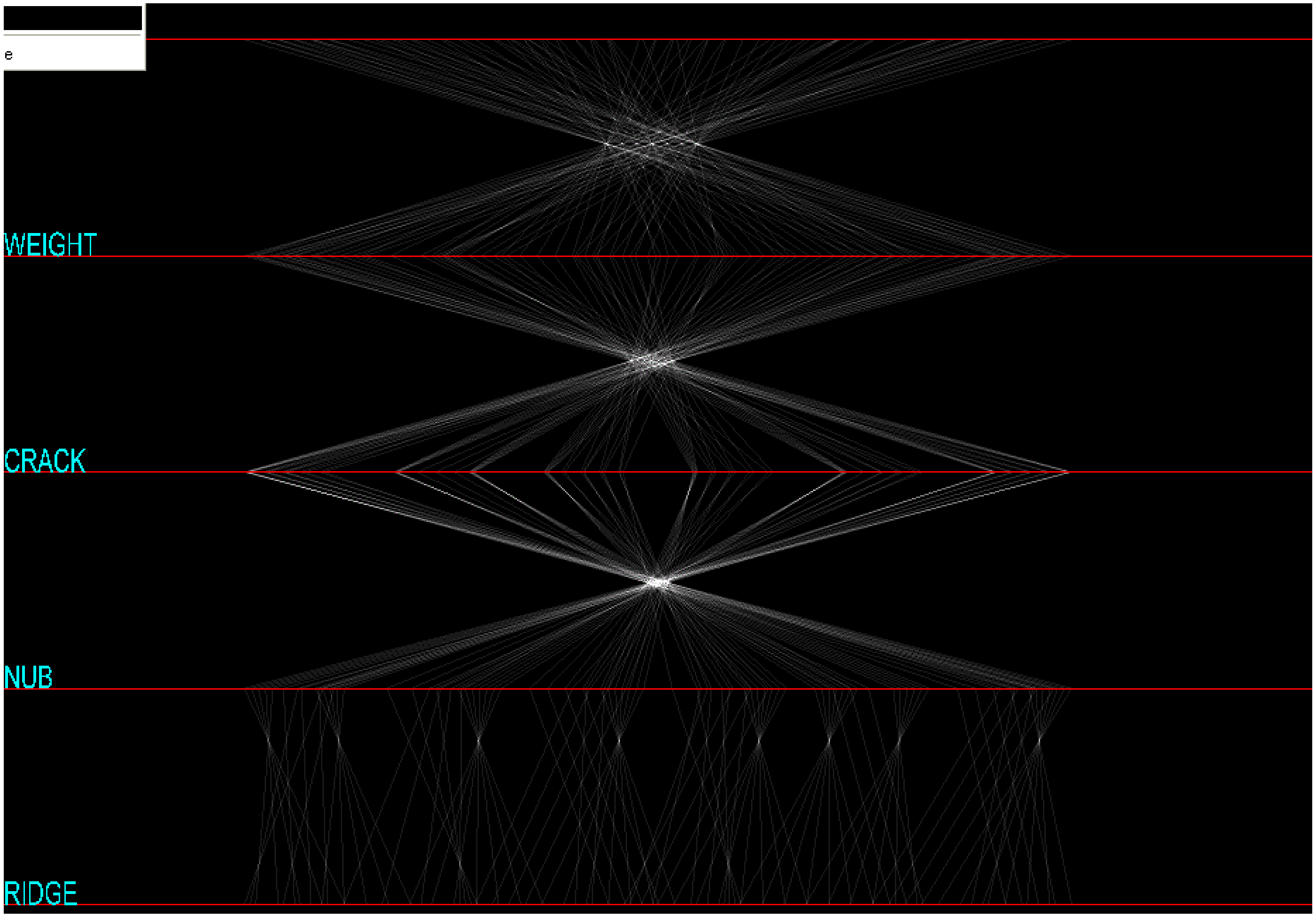


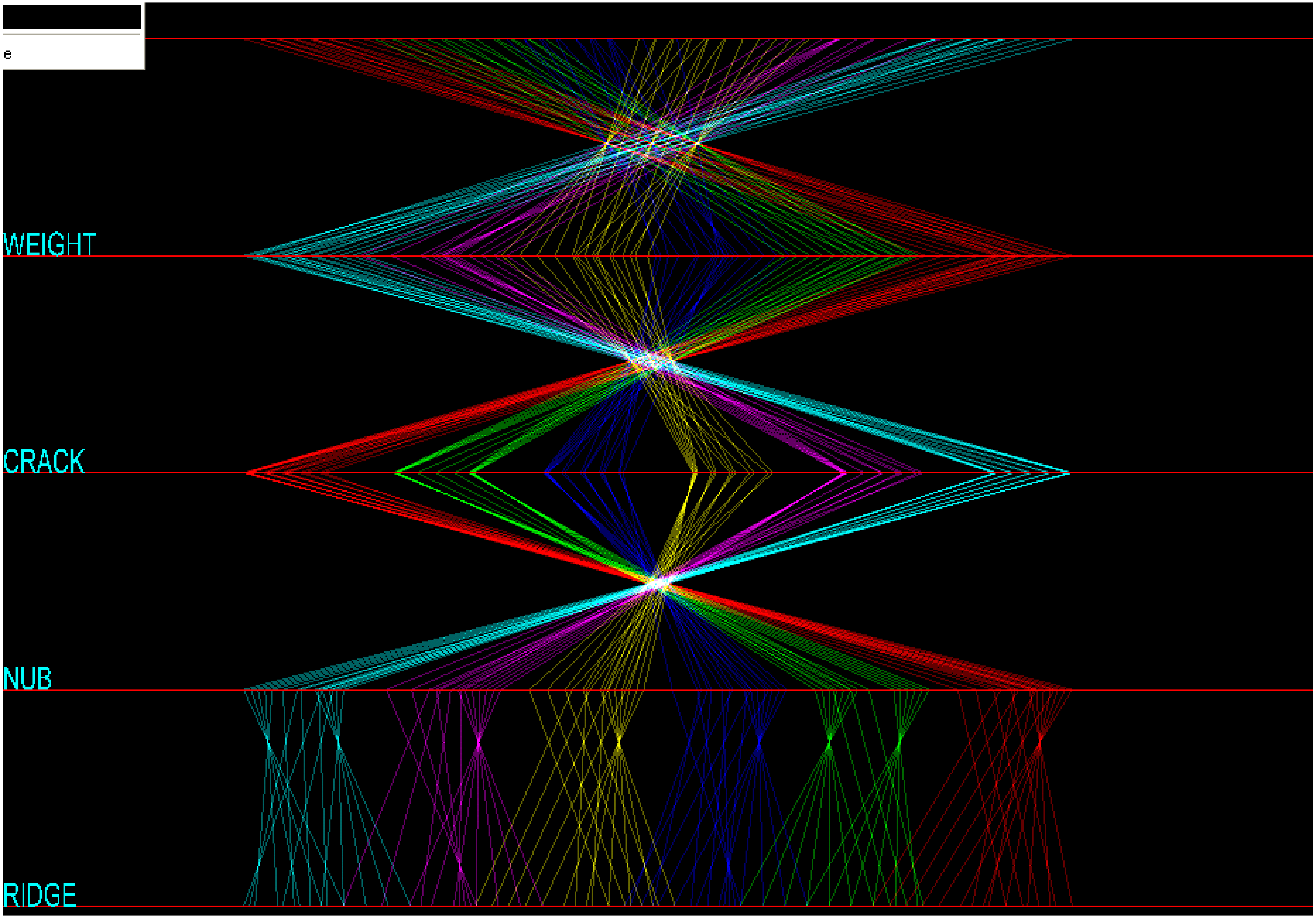
Low alpha

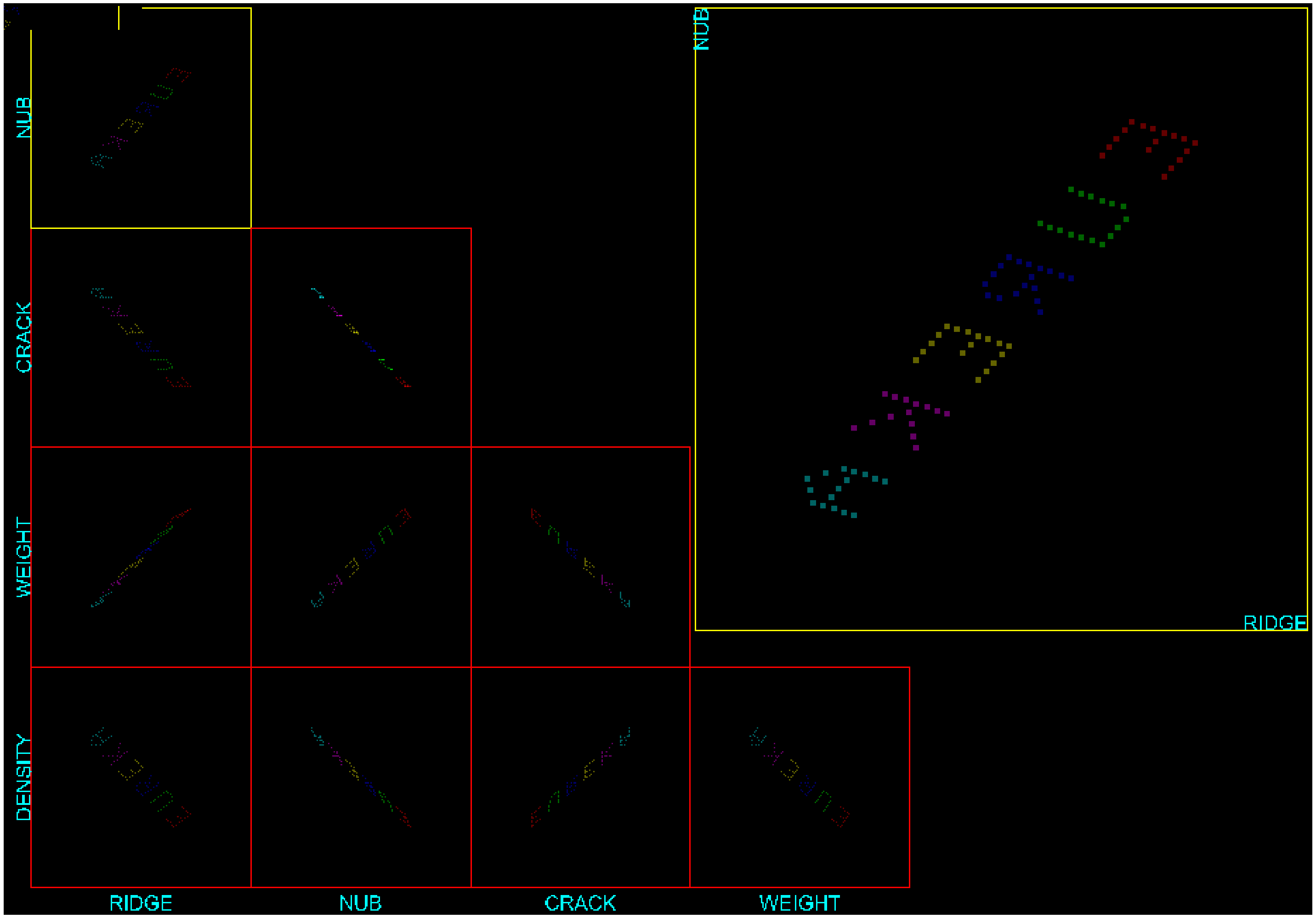
Now brushing - colour the data with chosen colours

and pruning - cut data you do not want

First lets PRUNE




e
WEIGHT
CRACK
NUB
RIDGE



98/3848 points. S/B = 2.6%

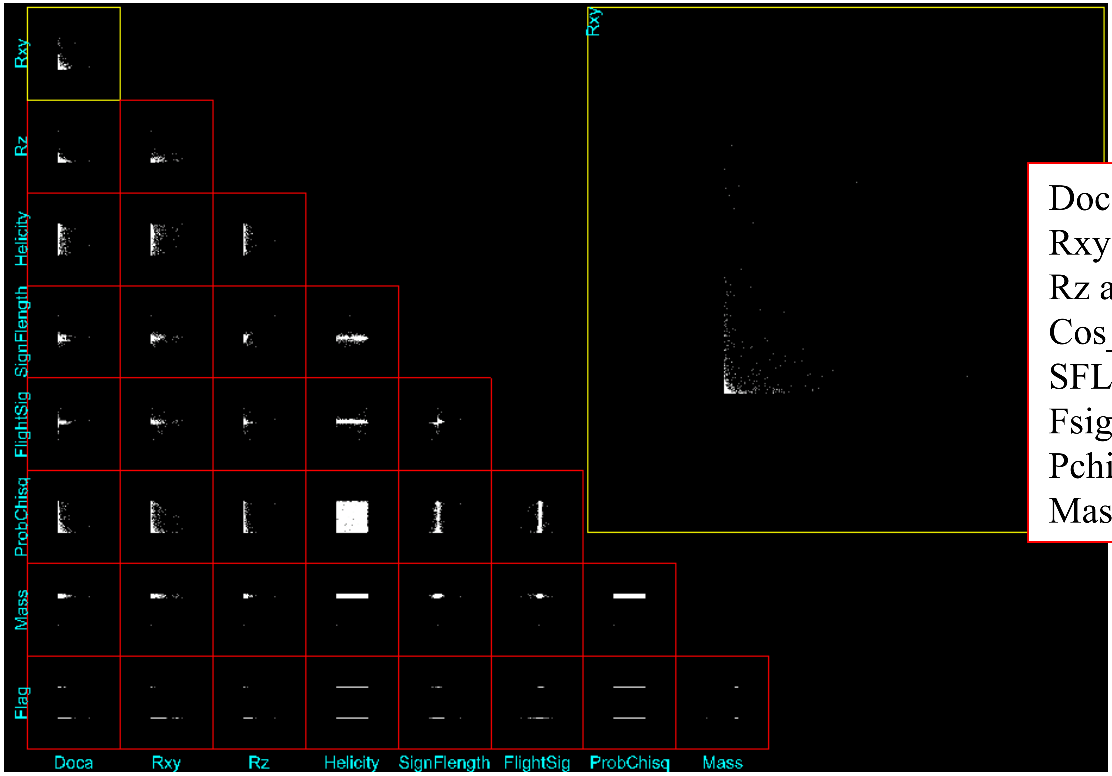
There are other features in the data. See E. Wegman

Contrived example, but helps a newcomer to use this type of graph.

Steve Watts, CHEP06, Brunel University

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.

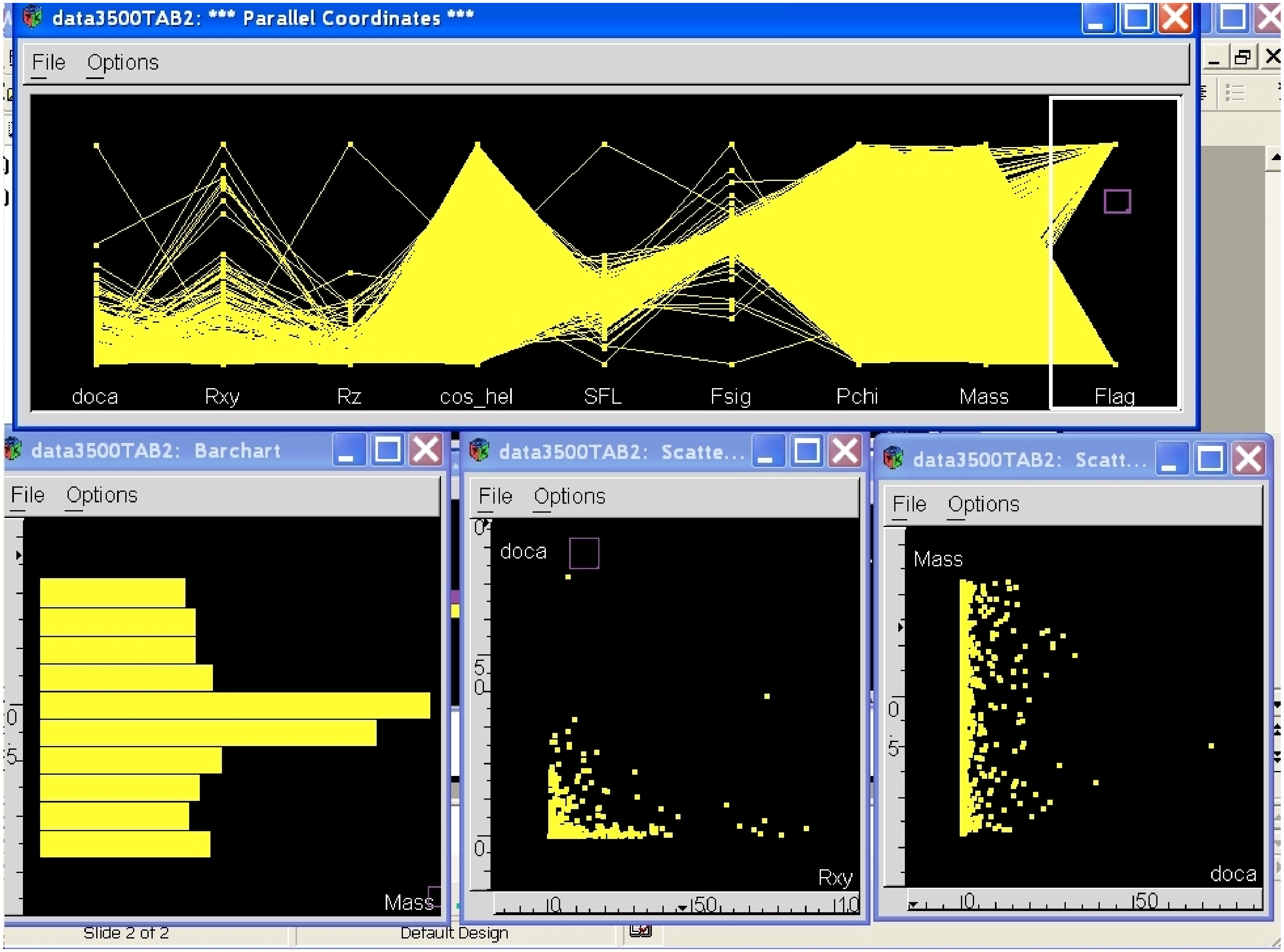


$K_s \rightarrow \pi^+ \pi^-$

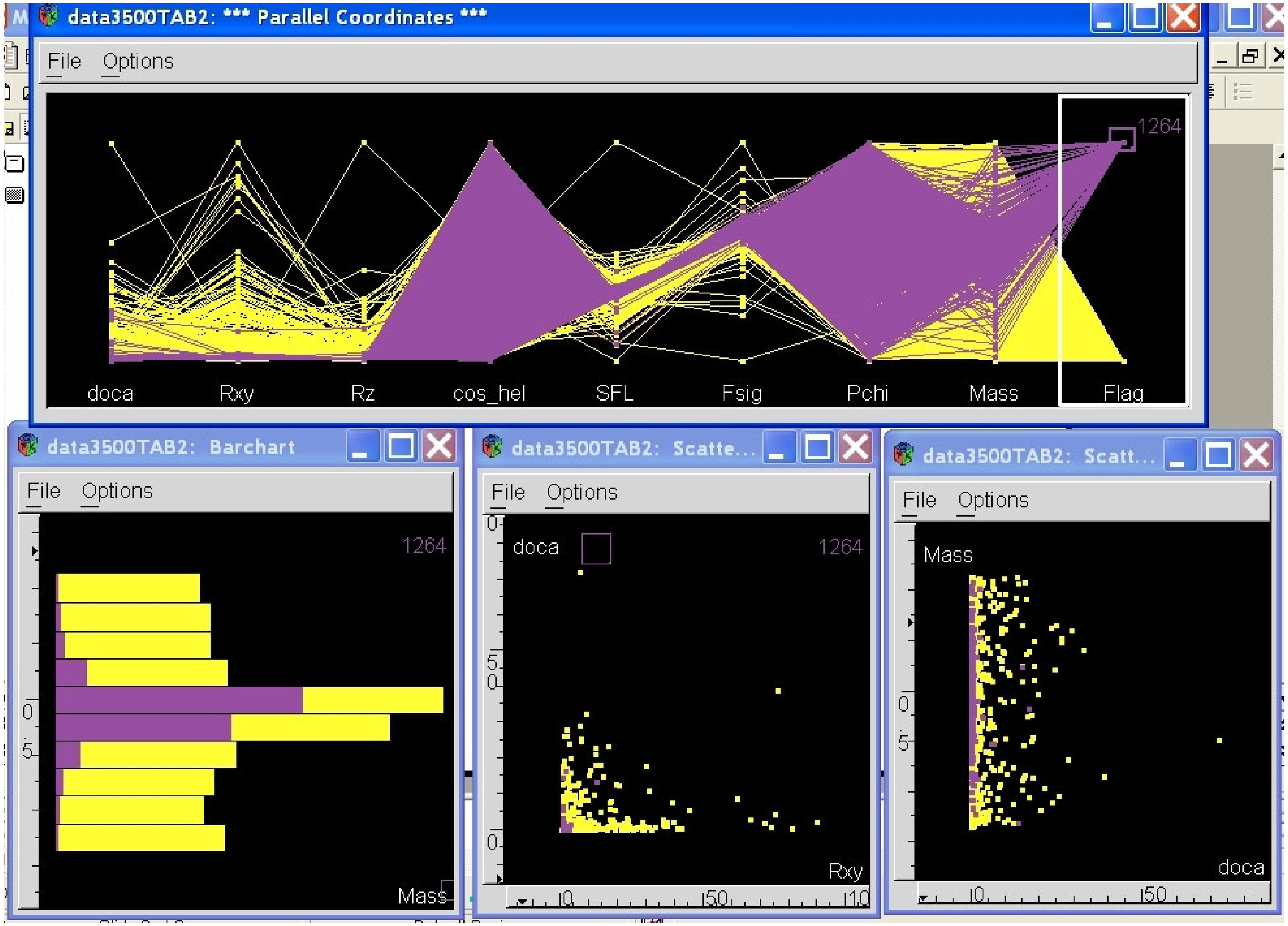
Doca = distance of closest approach
Rxy radius of cylinder for interaction region
Rz abs. half length of cylinder defining the IR
Cos_hel abs. Value of cosine of Ks helicity angle
SFL – signed flight length
Fsig stat. Sig. Of Ks flight length
Pchi chisq prob of Ks vertex
Mass – reconstructed mass of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

CrystalVision – E. Wegman

Lets try another package – GGOBI -http://www.ggobi.org/
About to be updated – FREE, Windows, Linux, OS-X

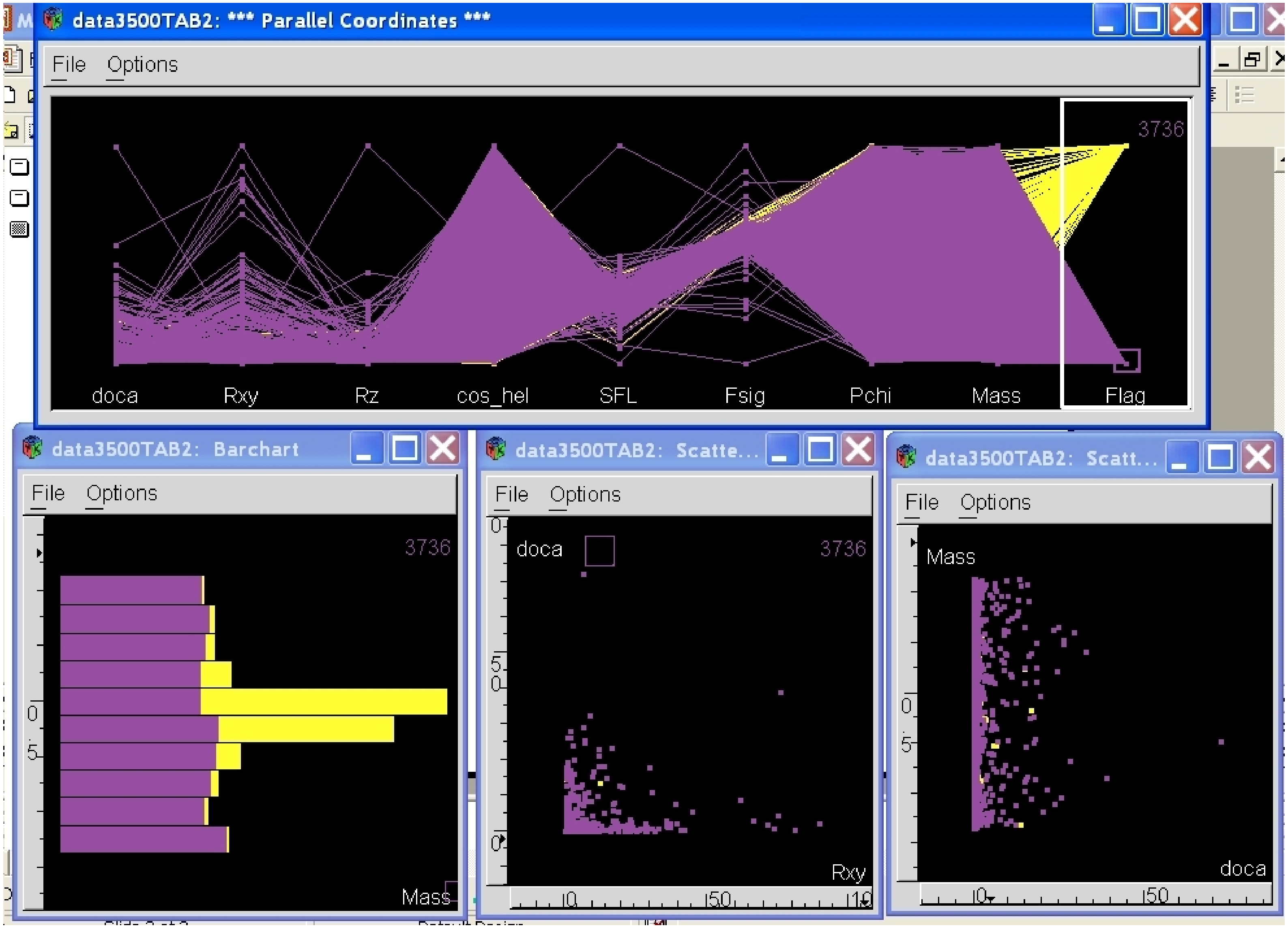


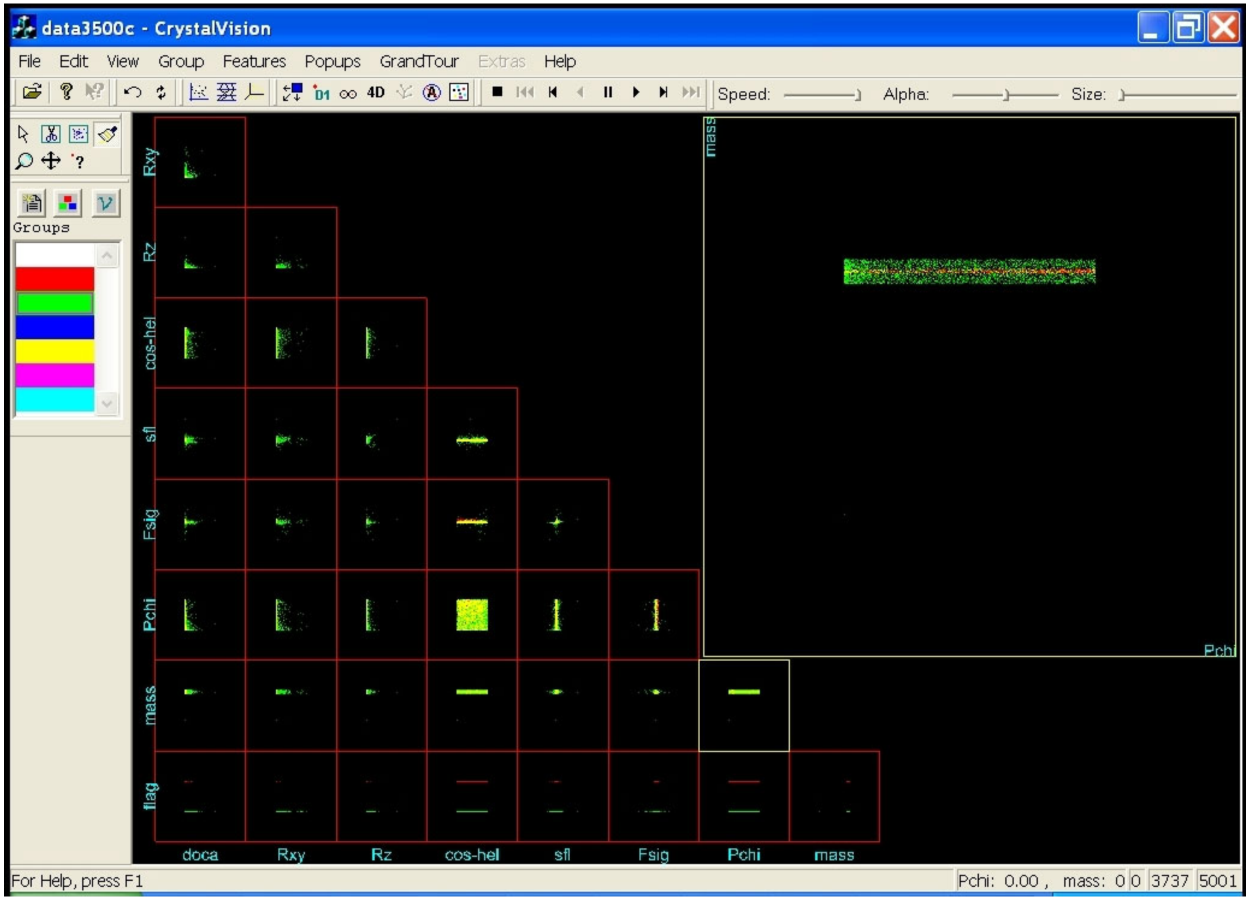
Brush the signal – FLAG = 1



Linked Brushing – colour points in one plot, and all open plots are also coloured – simple but very effective

Now brush the background – FLAG = 0





CrystalVision
(E. Wegman)

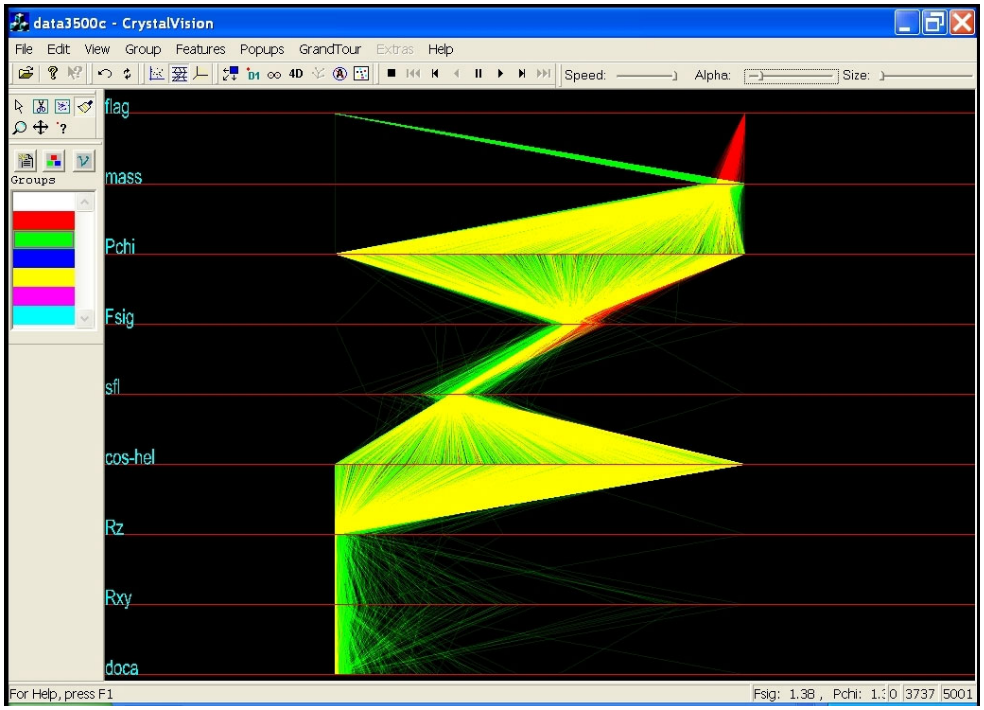
Has blending
and control of
intensity

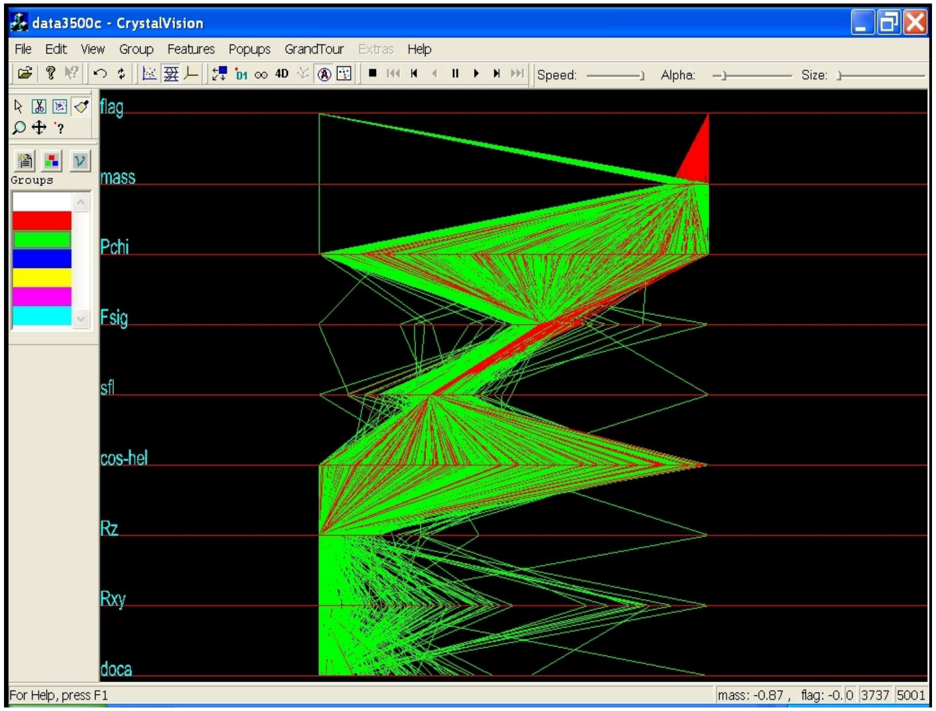
VERY Powerful

Brush signal RED and background GREEN

If they overlap RED + GREEN = YELLOW (yellow)

Now go to parallel coordinates - adjust alpha


data3500c - CrystalVision
File Edit View Group Features Popups GrandTour Extras Help
Speed: Alpha: Size:
Groups
flag
mass
Pchi
Fsig
sfl
cos-hel
Rz
Rxy
doca
For Help, press F1
Fsig: 1.38 , Pchi: 1.:0 3737 5001

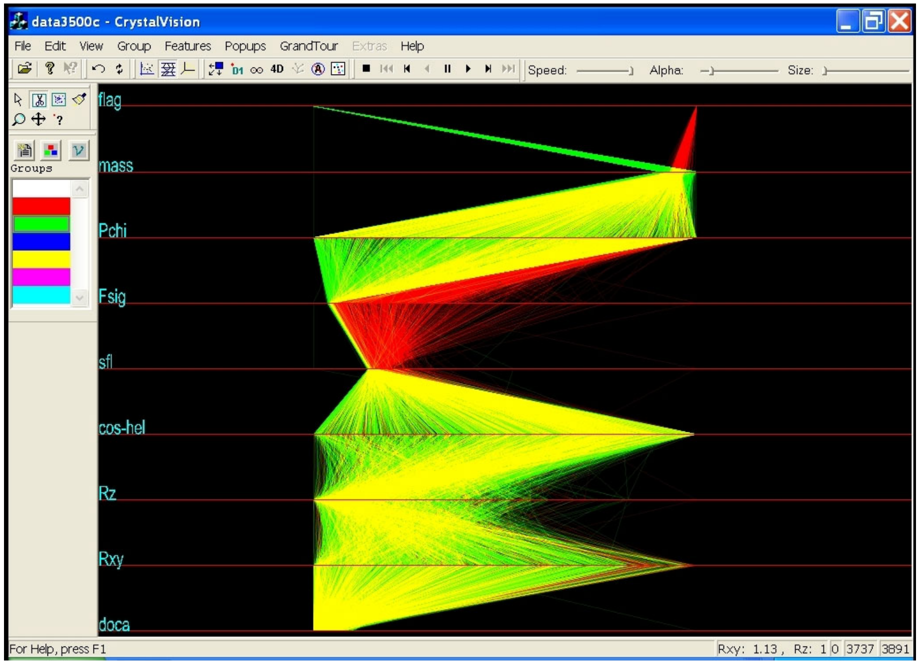

data3500c - CrystalVision
File Edit View Group Features Popups GrandTour Extras Help
Speed: Alpha: Size:
Groups
flag
mass
Pchi
Fsig
sfl
cos-hel
Rz
Rxy
doca
For Help, press F1
mass: -0.87 , flag: -0.0 3737 5001

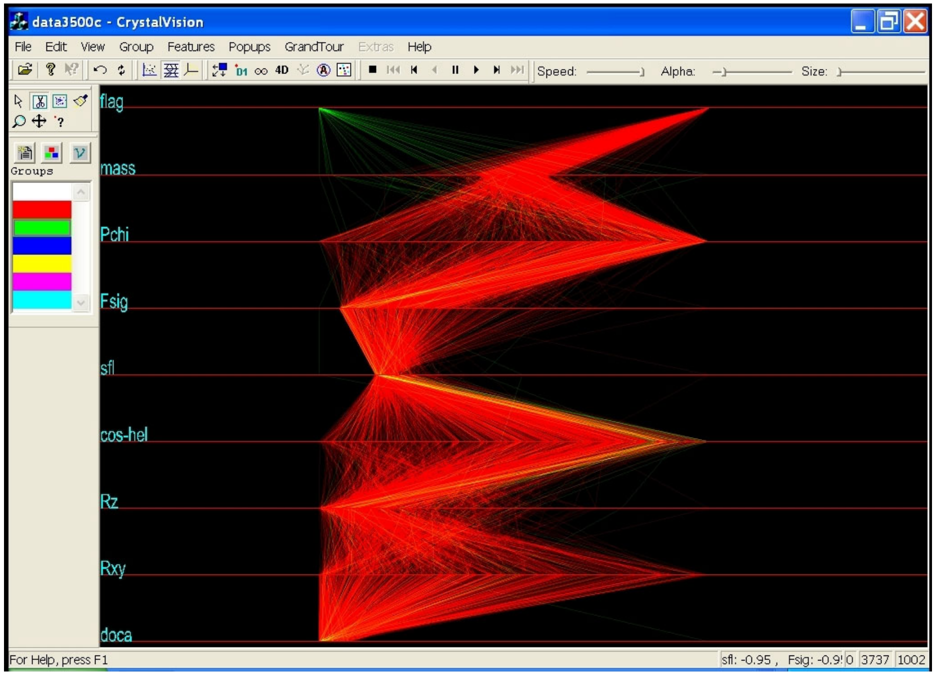
Note: See affect of turning alpha channel on and off

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.

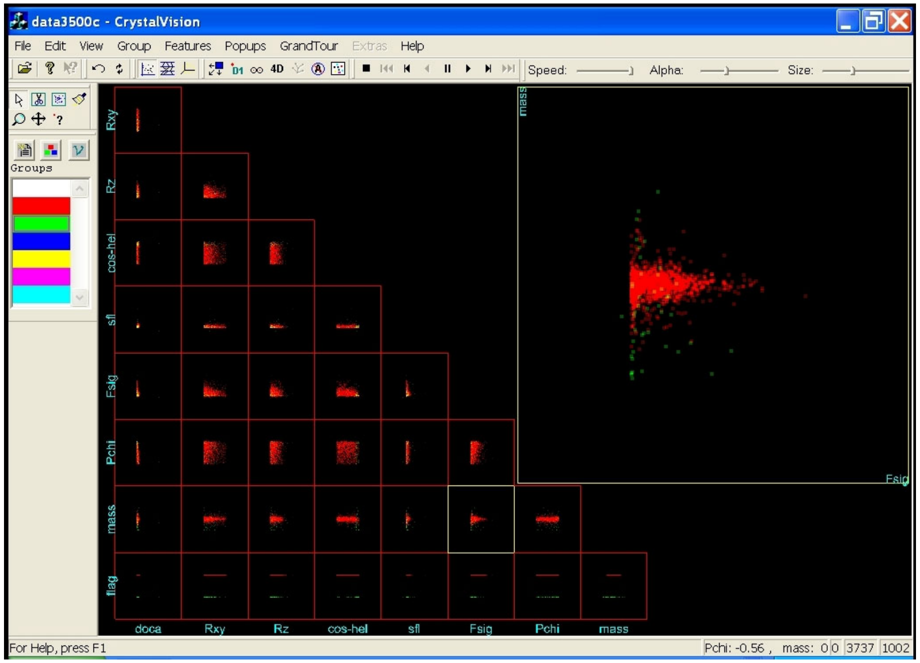
How to clean up this data - “ what is the order of cuts ?”
Remove obvious background (Prune Doca and Rxy)
Then select signal (FSig)


data3500c - CrystalVision
File Edit View Group Features Popups GrandTour Extras Help
Speed: Alpha: Size:
Groups
flag
mass
Pchi
Fsig
sfl
cos-hel
Rz
Rxy
doca
For Help, press F1
Rxy: 1.13 , Rz: 1 0 3737 3891


data3500c - CrystalVision
File Edit View Group Features Popups GrandTour Extras Help
Speed: Alpha: Size:
Groups
flag
mass
Pchi
Fsig
sfl
cos-hel
Rz
Rxy
doca
For Help, press F1
sfl: -0.95 , Fsig: -0.9 0 3737 1002

Takes just a couple of minutes to do this…

Back in scatterplot space


data3500c - CrystalVision
File Edit View Group Features Popups GrandTour Extras Help
Speed: Alpha: Size:
Groups
Rxy
Rz
cos-hel
sfl
Fsig
Pchi
mass
flag
doca
Rxy
Rz
cos-hel
sfl
Fsig
Pchi
mass
mass
Fsig
For Help, press F1
Pchi: -0.56 , mass: 0 0 3737 1002

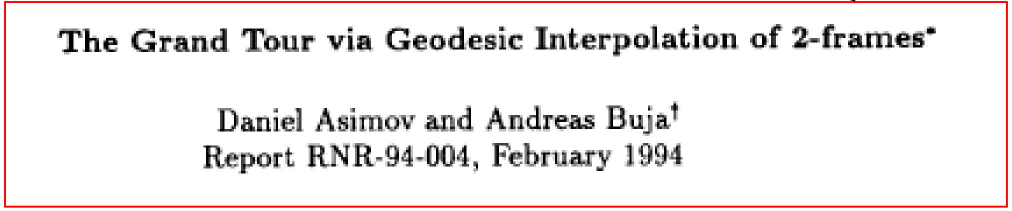
958 S 44 B 95% Purity 80% Efficiency

Did not spend long on this – Exploratory Visual Data Analysis

Powerful way to decide which variables matter and the order incuts should be applied.
Precursor to machine learning approach – e.g. Genetic expression programing
Liliana teodorescu – see talk at this conference.

The **GRAND tour**

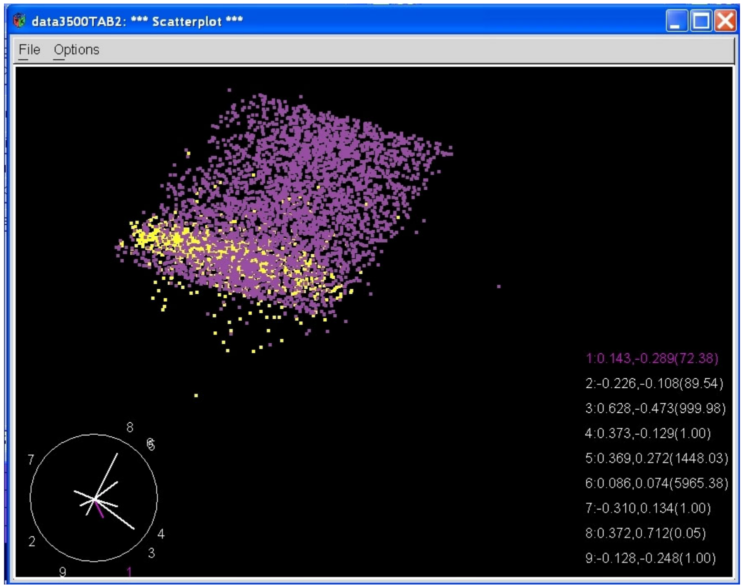
2D projections of an N-D space - choose suitable axes of rotation and an algorithm that ensures you explore all the space.
(The maths is complicated – See E. Wegman or Asimov



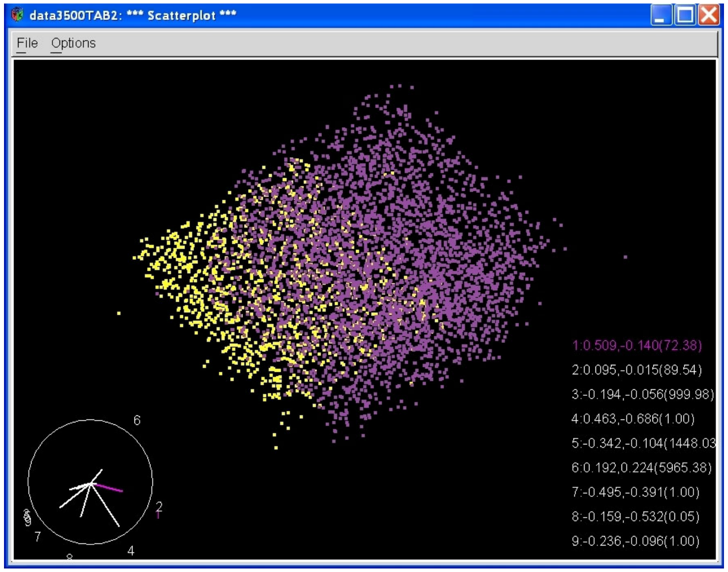
The Grand Tour via Geodesic Interpolation of 2-frames*

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

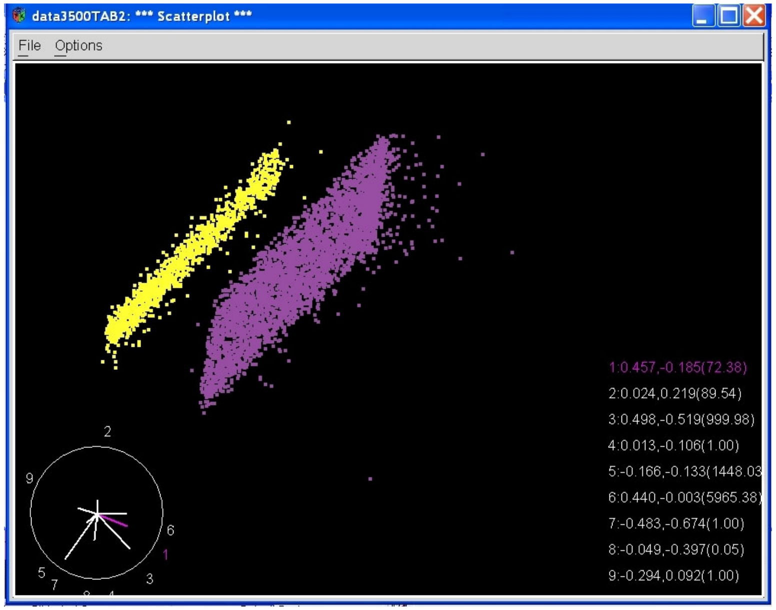
Facinating idea – useful for looking for clusters in data



```
1:0.143,-0.289(72.38)
2:-0.226,-0.108(89.54)
3:0.628,-0.473(999.98)
4:0.373,-0.129(1.00)
5:0.369,0.272(1448.03)
6:0.086,0.074(5965.38)
7:-0.310,0.134(1.00)
8:0.372,0.712(0.05)
9:-0.128,-0.248(1.00)
```



```
1:0.509,-0.140(72.38)
2:0.095,-0.015(89.54)
3:-0.194,-0.056(999.98)
4:0.463,-0.686(1.00)
5:-0.342,-0.104(1448.03
6:0.192,0.224(5965.38)
7:-0.495,-0.391(1.00)
8:-0.159,-0.532(0.05)
9:-0.236,-0.096(1.00)
```

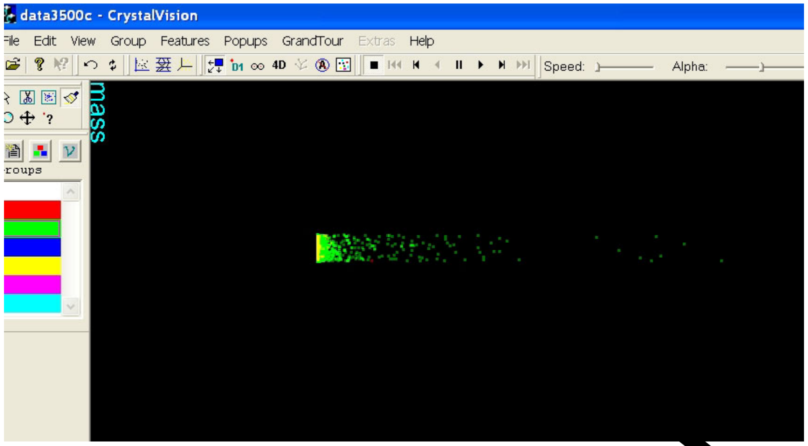


```
1:0.457,-0.185(72.38)
2:0.024,0.219(89.54)
3:0.498,-0.519(999.98)
4:0.013,-0.106(1.00)
5:-0.166,-0.133(1448.03
6:0.440,-0.003(5965.38)
7:-0.483,-0.674(1.00)
8:-0.049,-0.397(0.05)
9:-0.294,0.092(1.00)
```

Grand Tour

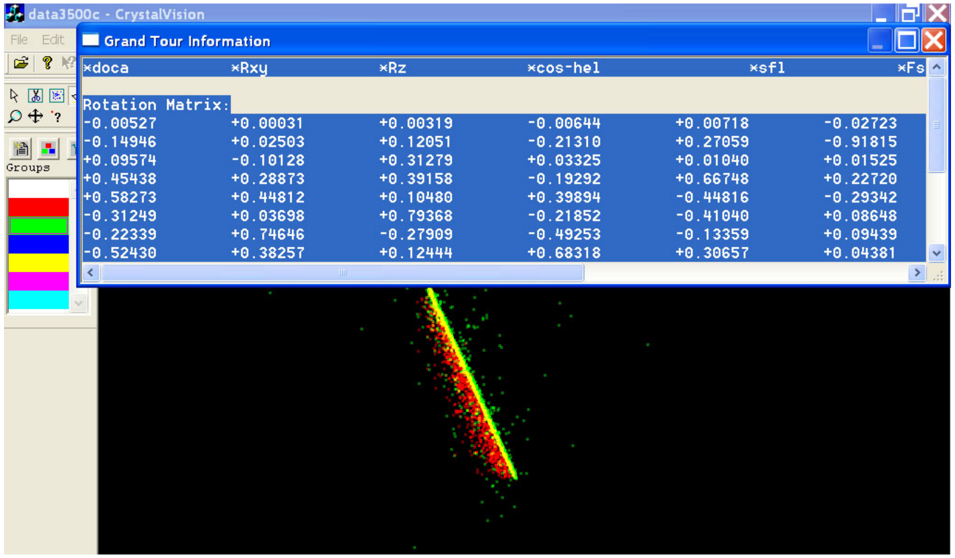
Asimov

Cheated in these pictures because "flag" dimension was included. Used to find clusters and then brush them.


data3500c - CrystalVision
File Edit View Group Features Popups GrandTour Extras Help
Speed:
Alpha:
mass
roups

Mass v Rxy
Standard Projection

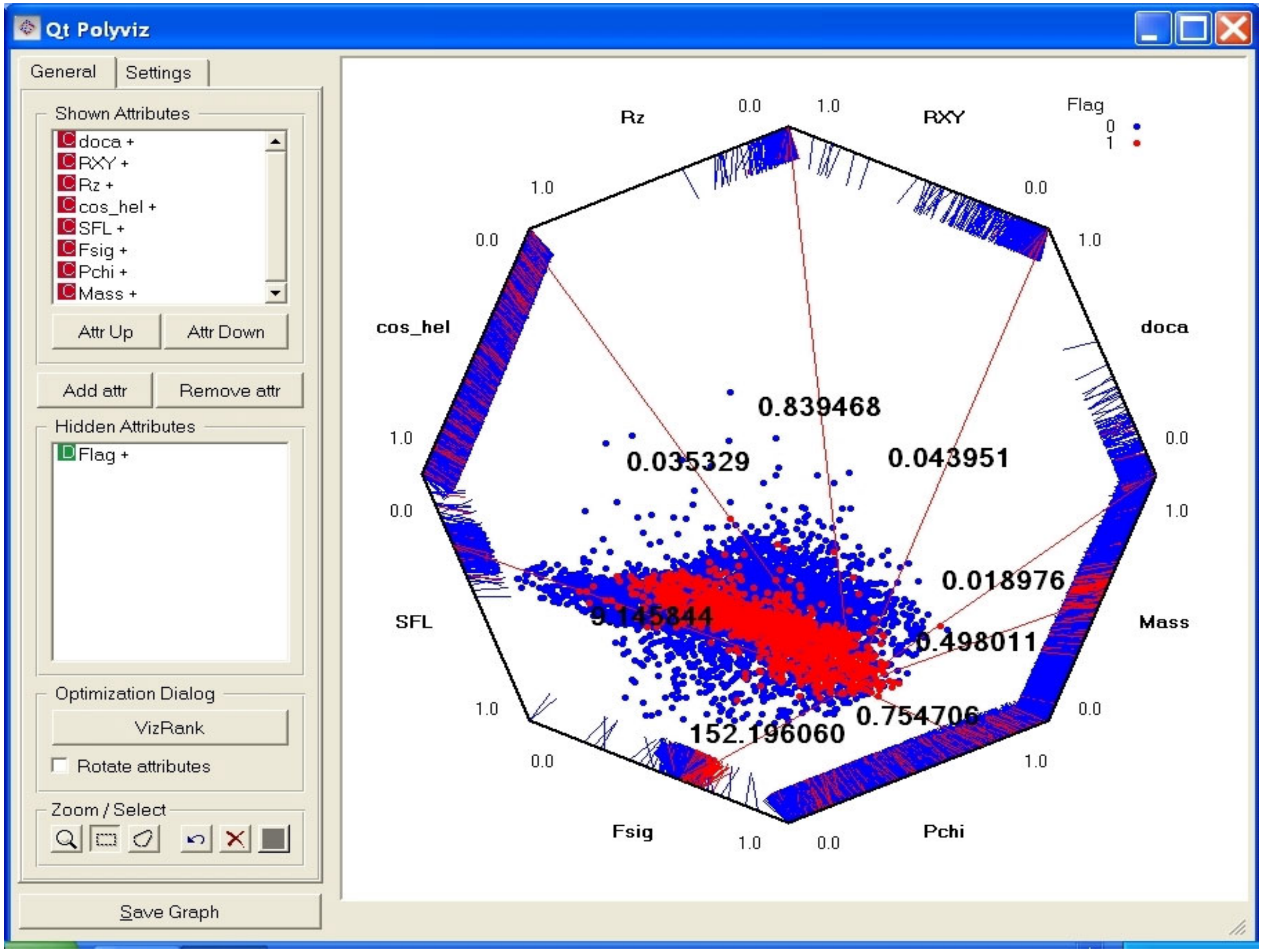
CrystalVision
GrandTour
is very fast !



```
×doca          ×Rxy           ×Rz            ×cos-hel       ×sfl           ×Fs

Rotation Matrix:
-0.00527       +0.00031       +0.00319       -0.00644       +0.00718       -0.02723
-0.14946       +0.02503       +0.12051       -0.21310       +0.27059       -0.91815
+0.09574       -0.10128       +0.31279       +0.03325       +0.01040       +0.01525
+0.45438       +0.28873       +0.39158       -0.19292       +0.66748       +0.22720
+0.58273       +0.44812       +0.10480       +0.39894       -0.44816       -0.29342
-0.31249       +0.03698       +0.79368       -0.21852       -0.41040       +0.08648
-0.22339       +0.74646       -0.27909       -0.49253       -0.13359       +0.09439
-0.52430       +0.38257       +0.12444       +0.68318       +0.30657       +0.04381
```

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.
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.



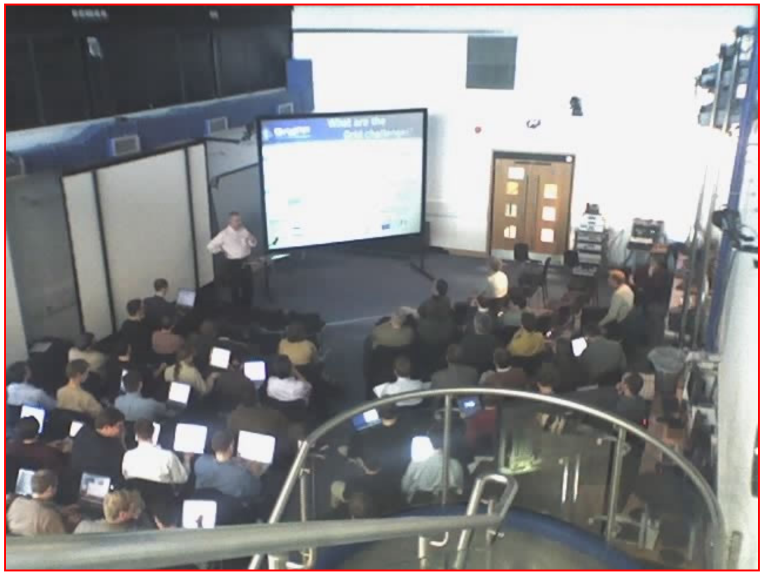
Many other types of graphic!!

Orange polyviz visualisation – Fsig, Rxy, doca, (mass) key variables
Can also use VizRank algorithm to find selection variables.

Comment :

Need decent size screen – workstation plus 3 times 19 – 30 inch screen
1-2 person data analysis station

Three large screens for collaborative data analysis ???



Steve Watts, CHEP06, Brunel University

Conclusion

These are powerful techniques and we should implement them in our data analysis toolkit.

Many other ideas that I have not discussed. It is also easier to understand dynamically – just ask and I will show you.

CrystalVision is the best software for parallel coords. but it does not export results of the analysis. Has blending and alpha channel. Can also use stereo with CrystalVision.

GGobi is good – new version to be released soon.

Data Analysis – Exploratory Visual Data Analysis followed by machine learning/GEP techniques (Liliana Teodorescu) to select/cut data in a human independent way.

Can we find signals using data mining without a prior knowledge of what we think is there ?