Multivariate Visualisation

20th November 2018
SGD Topics #1: Visualising Multivariate Data

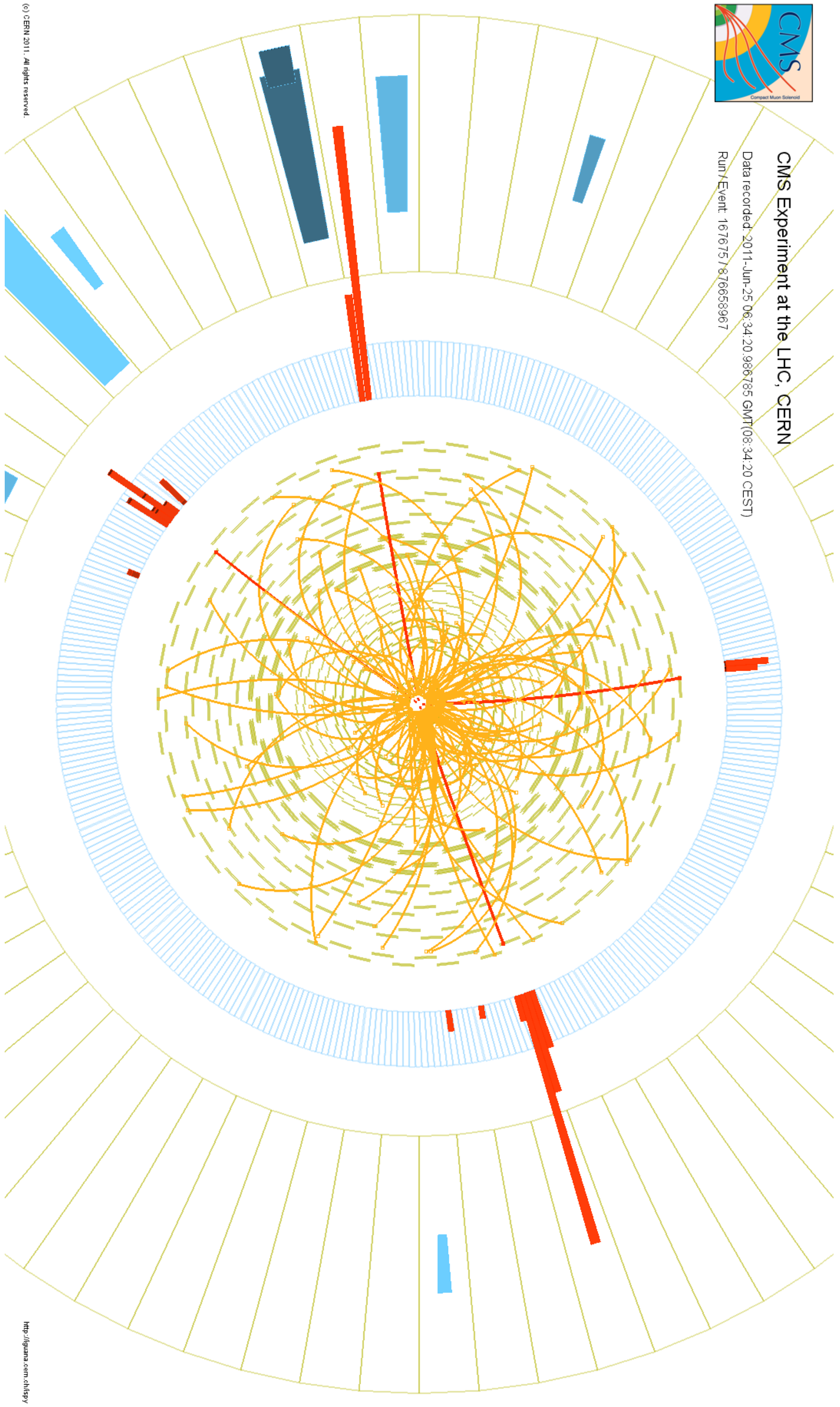
Benjamin Radburn-Smith
cern.ch/benjamin

Contents

- Multivariate Visualisation
  - Introduction
  - Parallel Coordinates
  - The Grand Tour

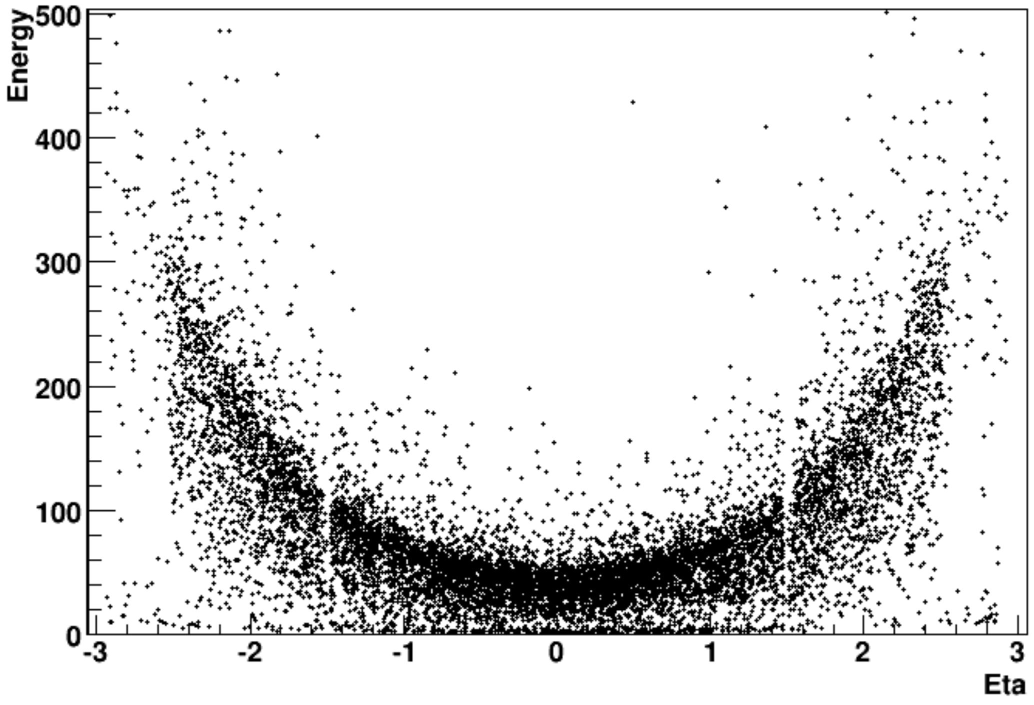
Visualisation Introduction

- Modern particle physics experiments provide a tremendous volume of data, recording PB/year
  - The increase flux of data coming in is not limited to HEP, as we are predicted to be into ZettaBytes (10^{21}) range in the world and increasing rapidly
- These data are usually complicated with many attributes or variables
  - For example, consider an electron detected in an experiment will have a number of properties associated with it such as: Energy, direction, amount of energy deposited in a subdetector, shower shape, isolation, number of hits in silicon etc


CMS
Compact Muon Solenoid
CMS Experiment at the LHC, CERN
Data recorded: 2011-Jun-25 06:34:20.986785 GMT(08:34:20 CEST)
Run / Event: 167675 / 876658967
(c) CERN 2011. All rights reserved.
http://iguana.cern.ch/ispy

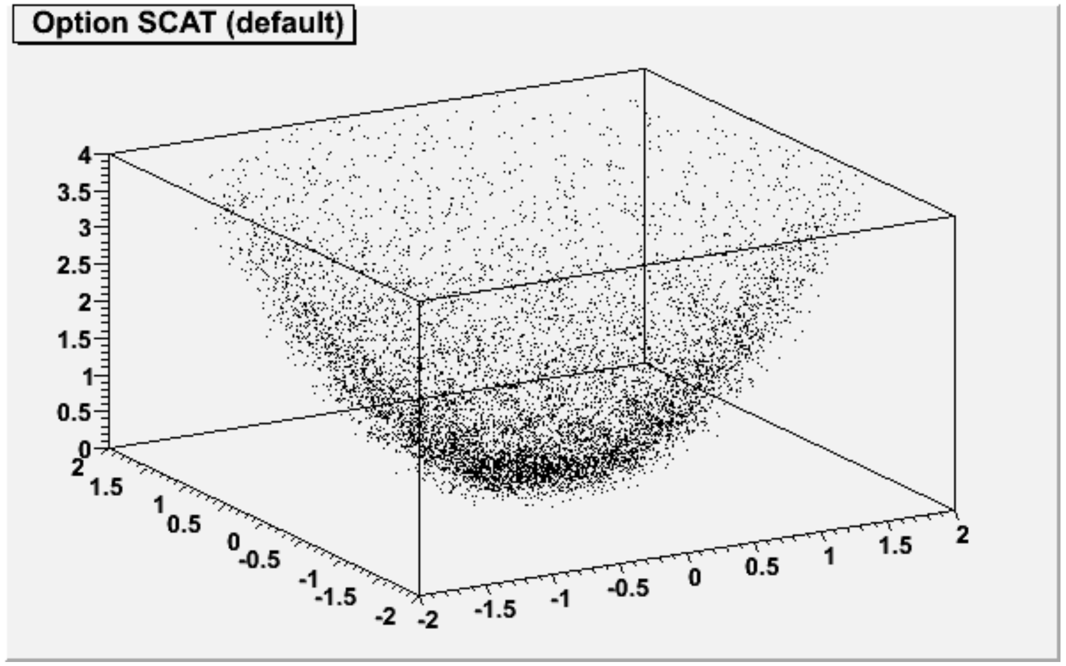
Preprocessing

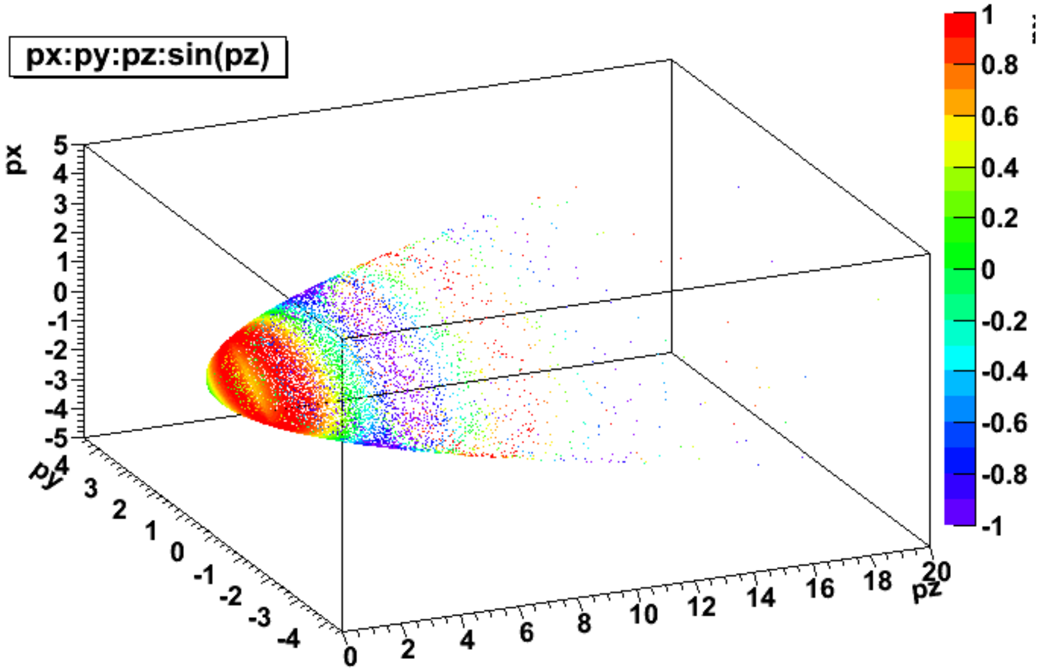
- Data may contain many variables, but usually only a subset of those are of interest
    - The first step is to find out which variables are of interest
- We could use visualisation here to see which variables to use and which we can discard
- Multiple variables = multiple dimensions
    - The data has many dimensions which could, for example, be projected down onto a lower dimensional plot
    - In the case of the electrons from those six variables (6D), two could be plotted in a traditional 2D Cartesian scatter plot



Multivariate Visualisation

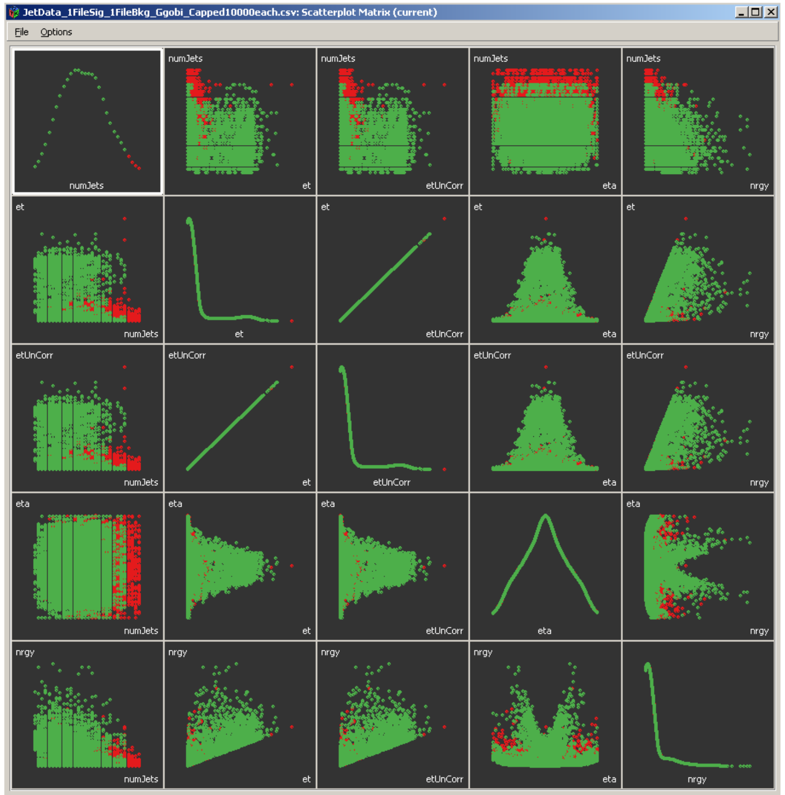
- We could plot in higher dimensions, for example a 3D plot
    - (which has been projected down onto this 2D slide)
- Through use of computer graphics: colours and shapes we can reach into 4D and even 5D (although this can become difficult to understand) then we reach a hard limit
- Or We could use even higher dimensional visualisations …

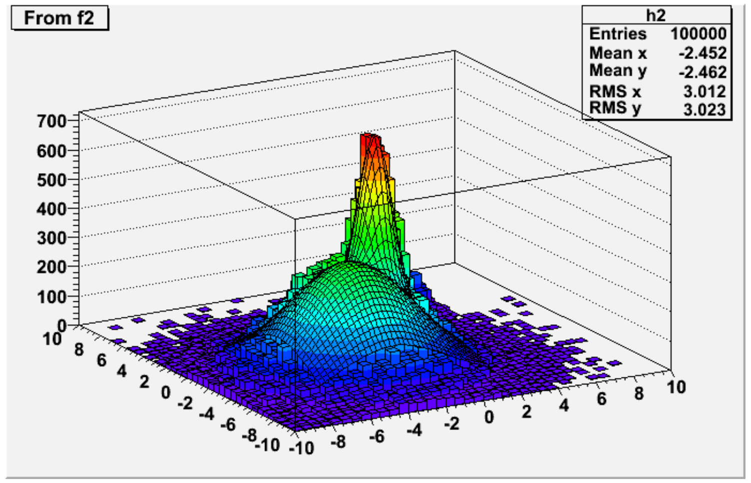


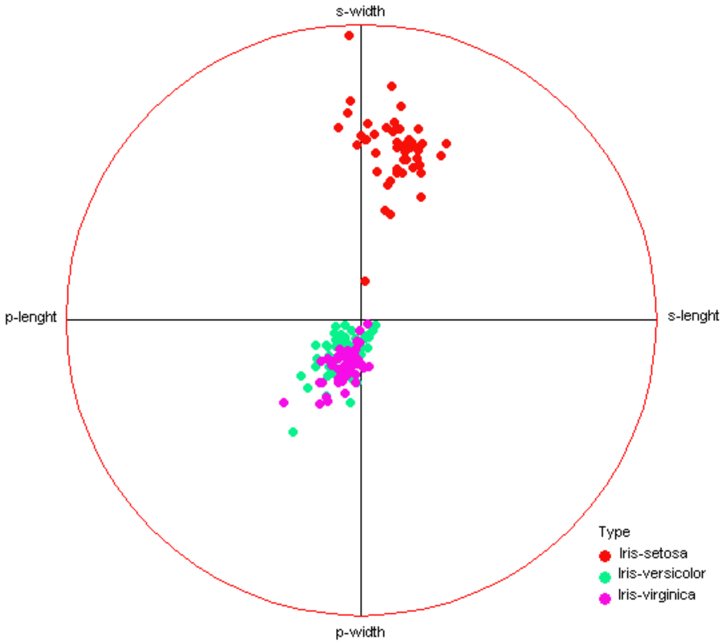


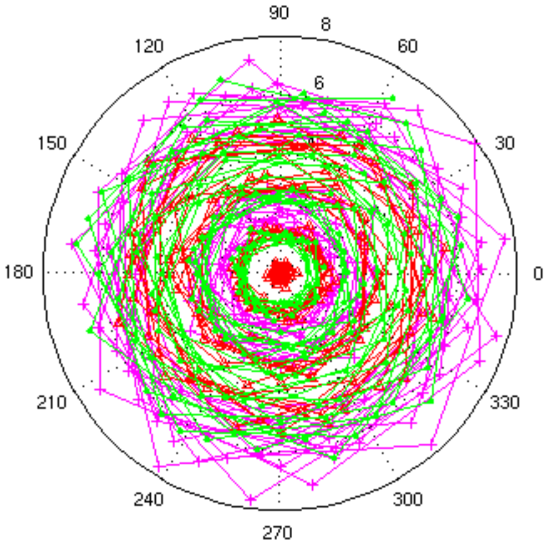
Multivariate Visualisation

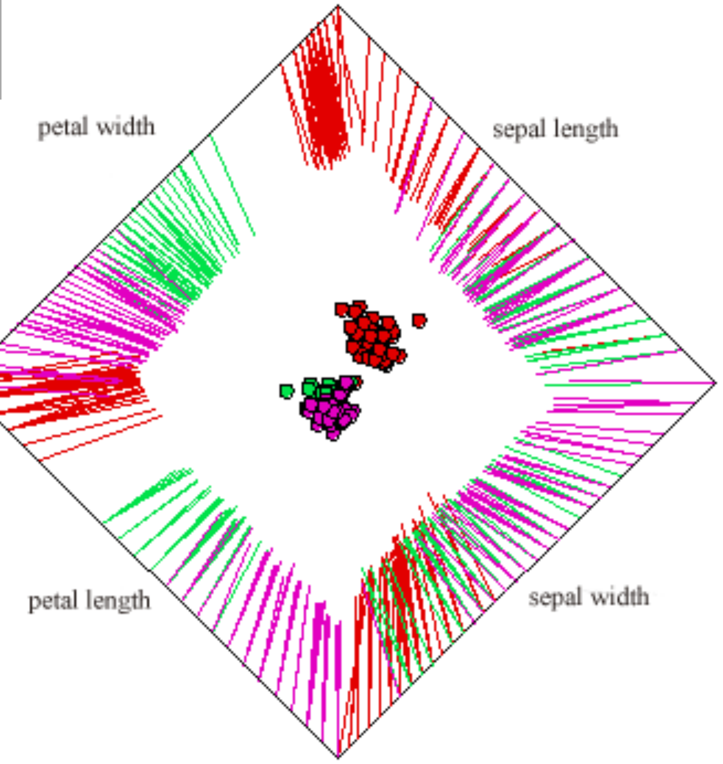
- Many data visualisations available
  - More advanced scatter plots, e.g. scatter plot matrix
  - Heat maps and height maps (extension of scatter plot)
  - Polar Charts
  - RadViz and PolyViz
  - Parallel Coordinates
  - The grand tour


JetData_1FileSig_1FileBkg_Ggobi_Capped10000each.csv: Scatterplot Matrix (current)
File
Options
numJets
et
etUnCorr
eta
nrgy



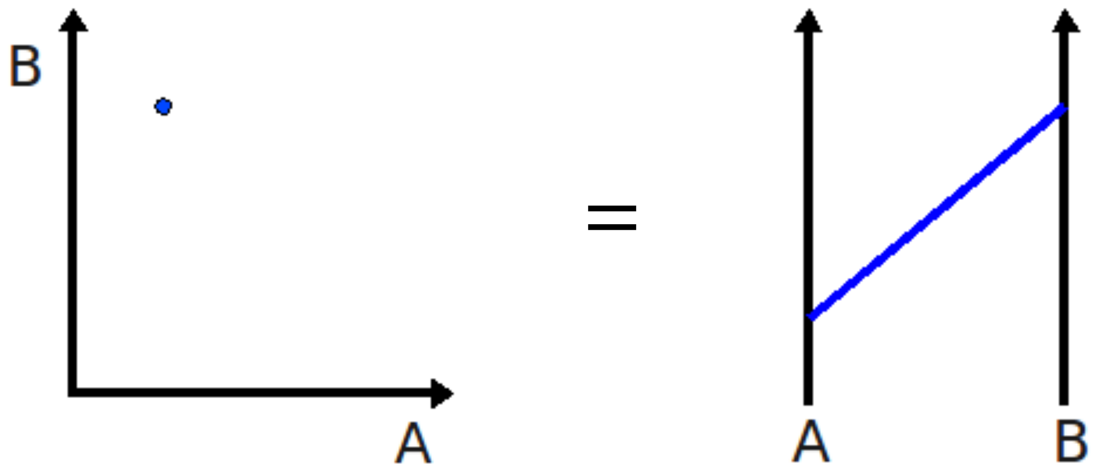






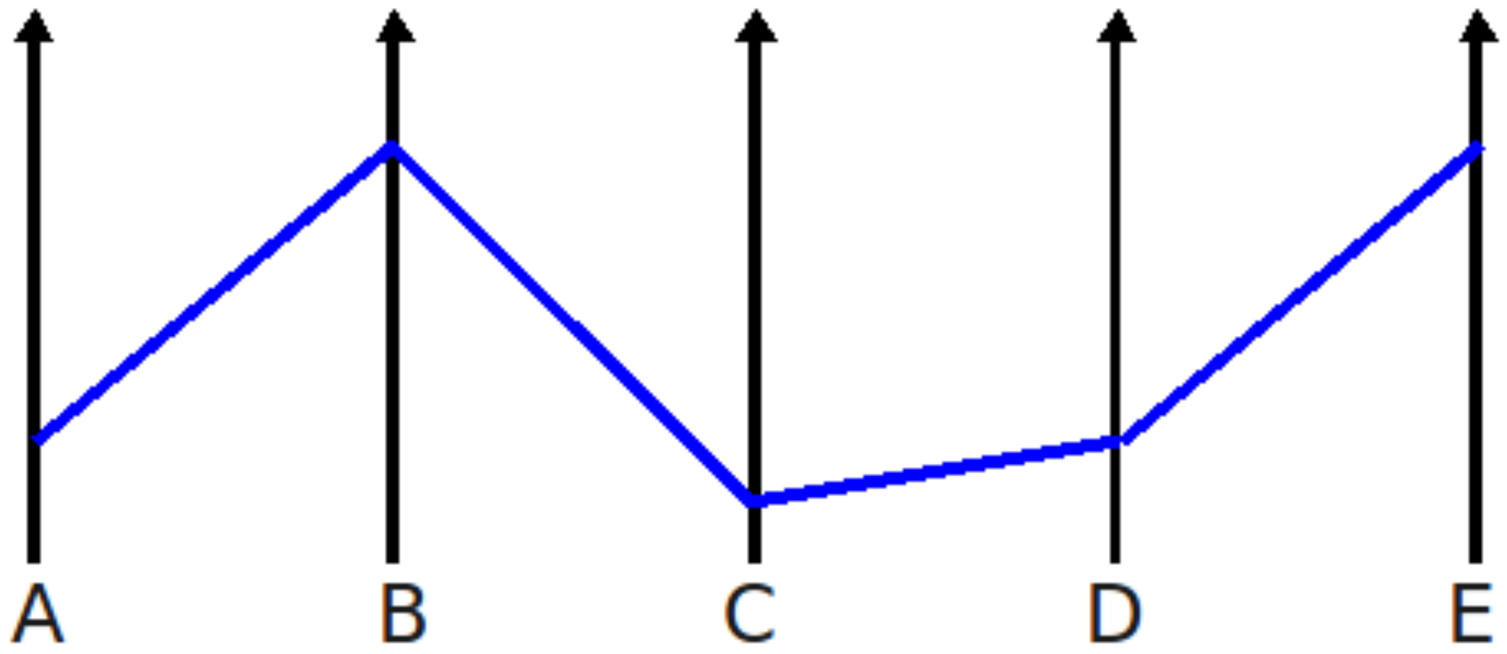
Parallel Coordinates

- Parallel Coordinates (PC) dates back to 1885; Maurice D’Ocagne used it as a method of visualising geometric transformations
- Re-invented by Alfred Inselberg in 1985
- Developed by Edward Wegman in 1990 as a multivariate data analysis tool
- Principle:



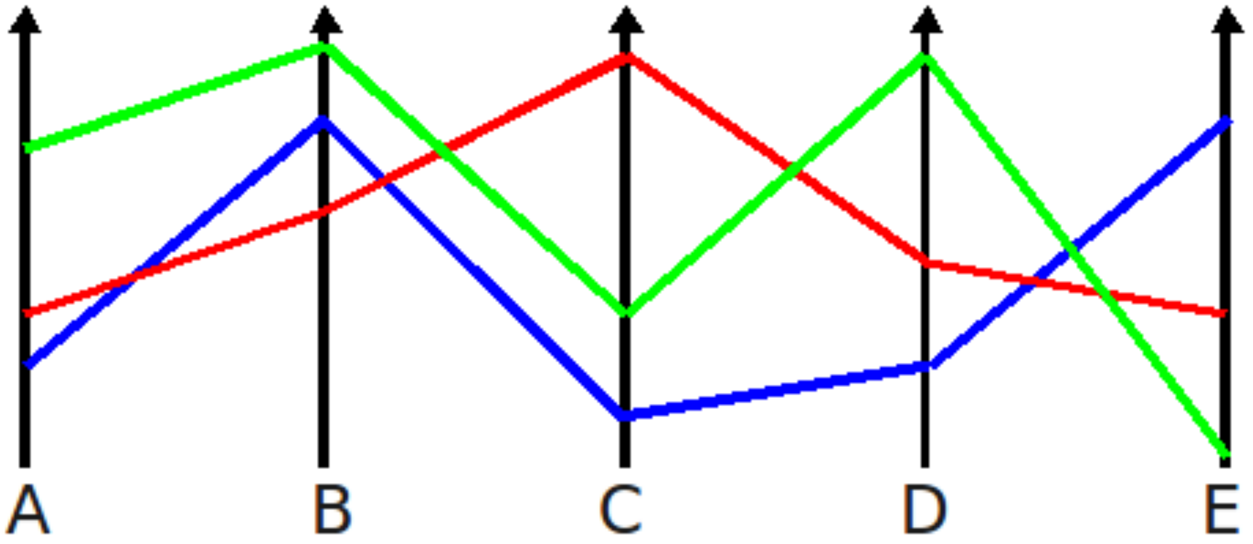
Parallel Coordinates

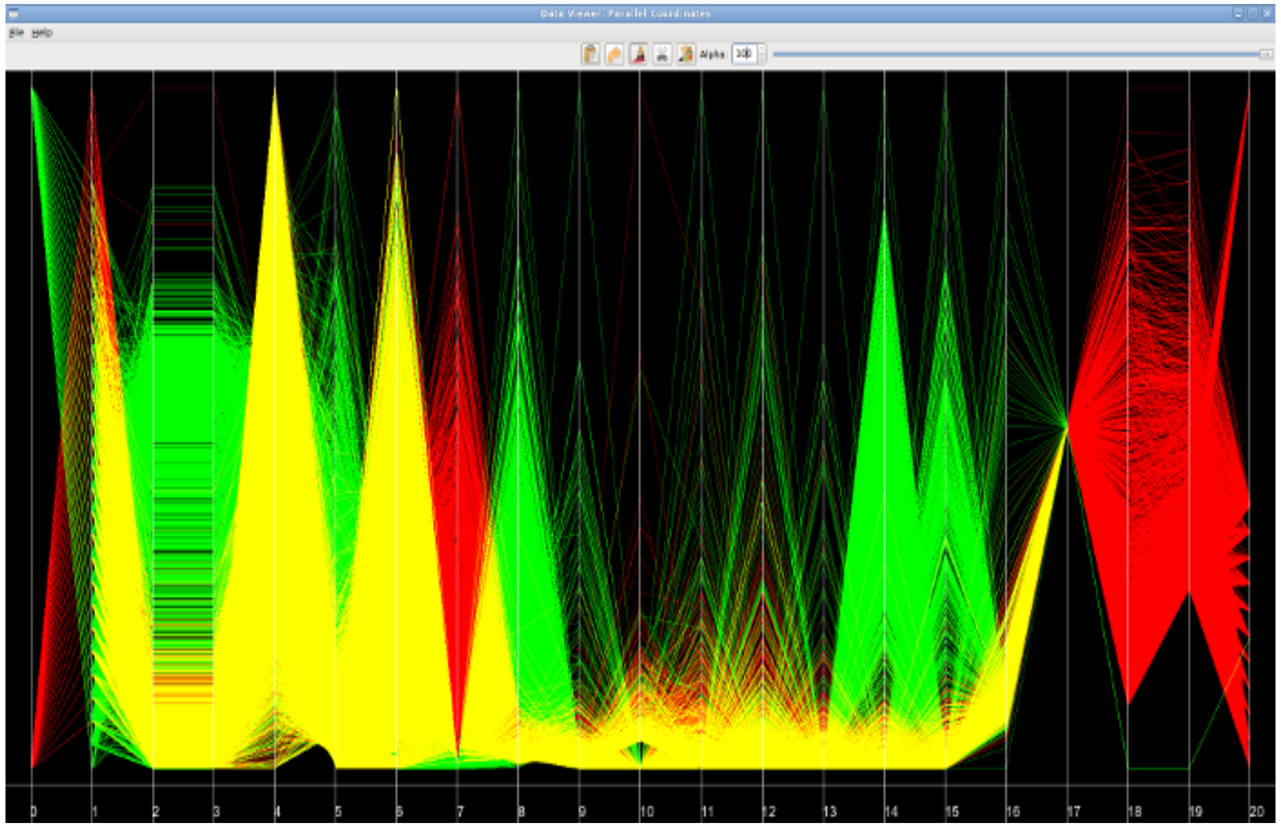
- Not limited to 2 or 3 dimensions
- For example a 5D plot:
- Can think of each axis as a 1 dimensional view with the maximum value at the top and the minimum value at the bottom



Parallel Coordinates

- With more data





- 20k data instances in 21D
- Red = Background
- Green = Signal
- Yellow = Overlap (S+B)

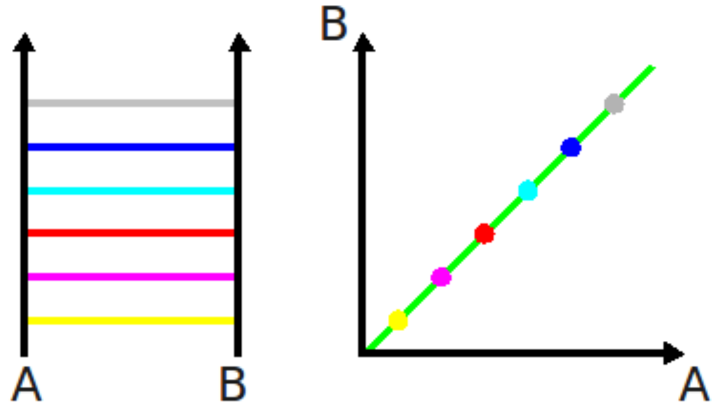
Parallel Coordinates

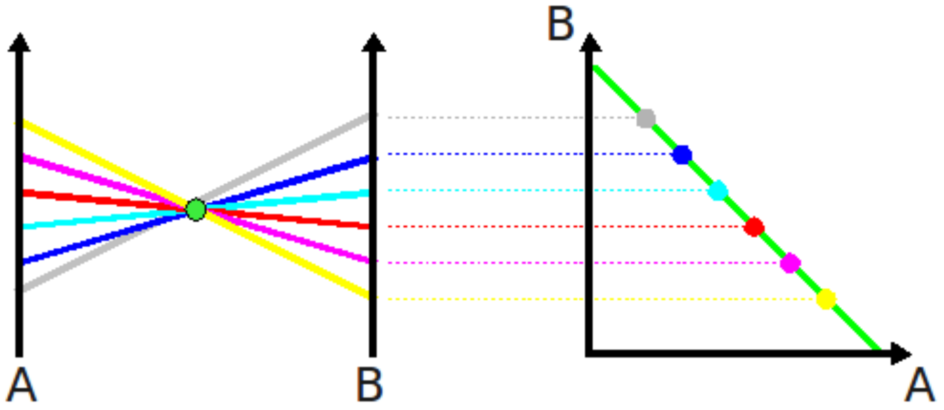
- The data correlations between variables

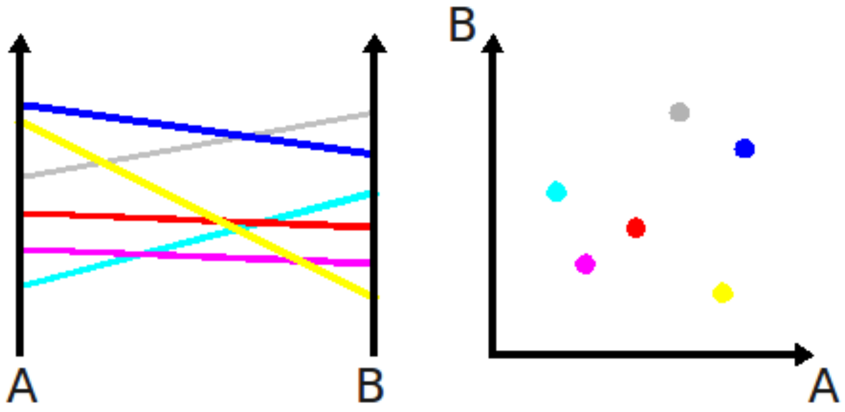
- Positive correlation: Lines are parallel

- Negative correlation: Lines intersect

- Uncorrelated: Lines are random

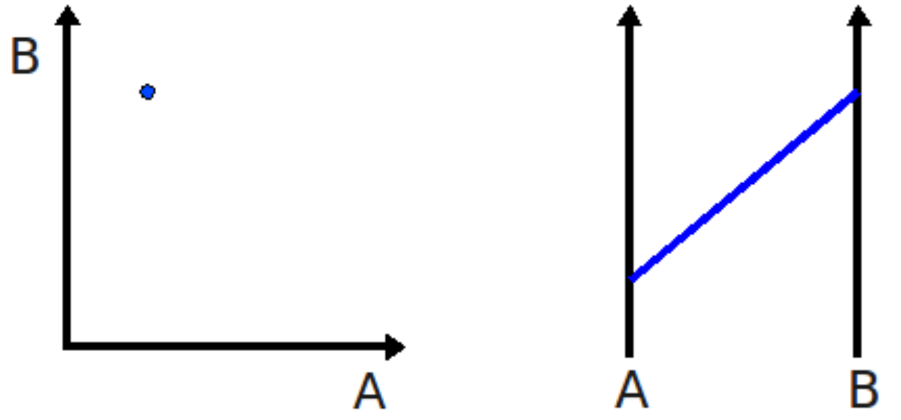




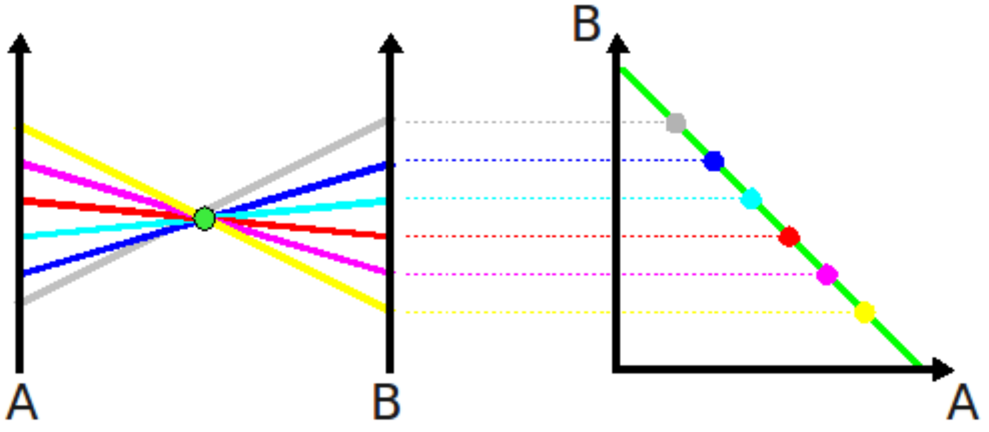


Parallel Coordinates

- Dualities
    - Point ↔ Line
- A point in 2D Cartesian is represented by a line in parallel coordinates

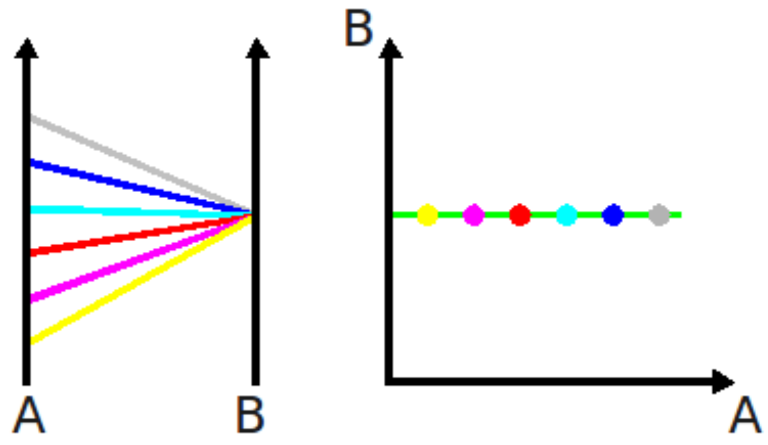


- A line of points in 2D Cartesian is represented by a series of lines that intersect at a point in parallel coordinates

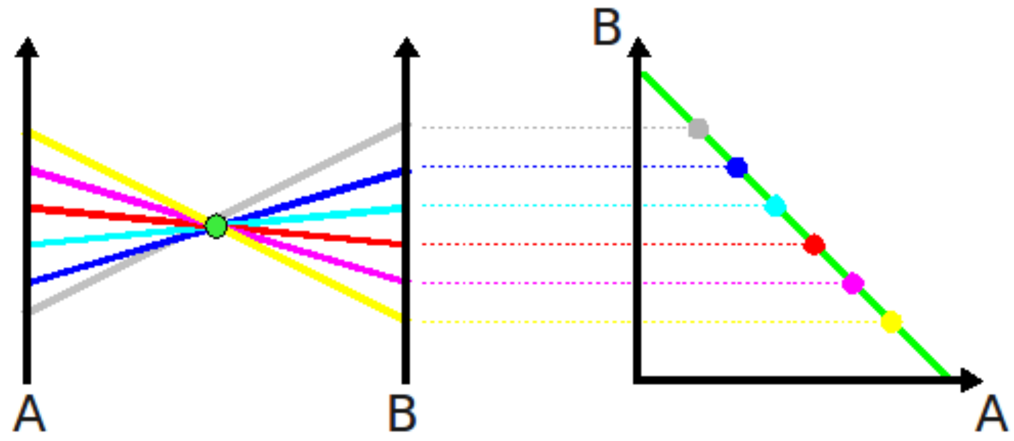


Parallel Coordinates

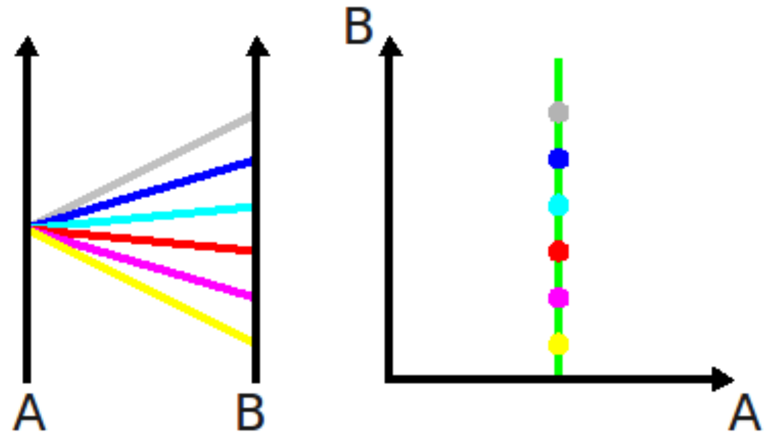
- Dualities
  - Rotation ↔ Translation



- Rotating the line in the 2D Cartesian system moves the intersection point in the parallel coordinate plot



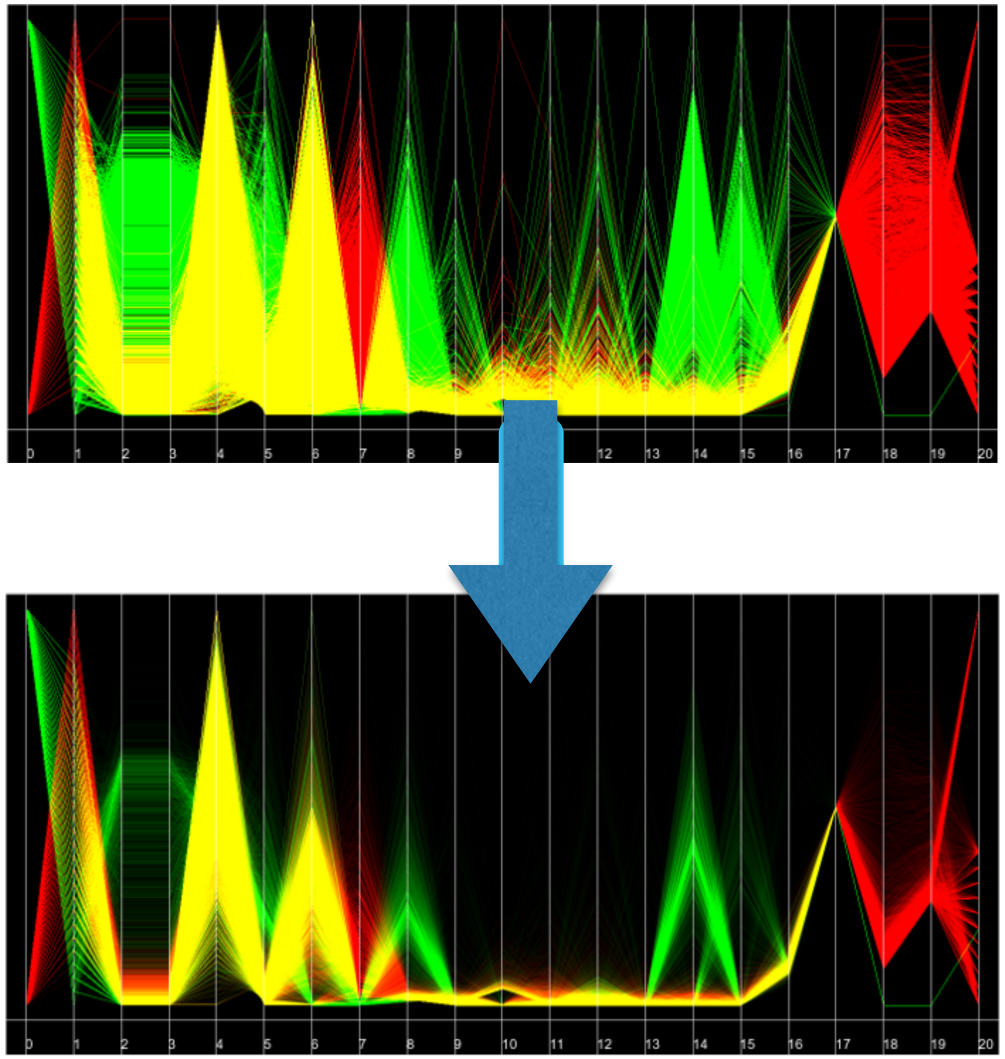
- Moving a point in 2D Cartesian along an axis (eg along B=0) rotates the corresponding line in the parallel view



Parallel Coordinates

- Parallel coordinate density plot
  - One of the problems with the traditional PC plots is that of overplotting, where multiple data lines can lie upon each other making it difficult to understand what is happening
  - By changing the amount of transparency of the data on the plot we can see the internal structure of the data and view larger amounts of data

Low transparency

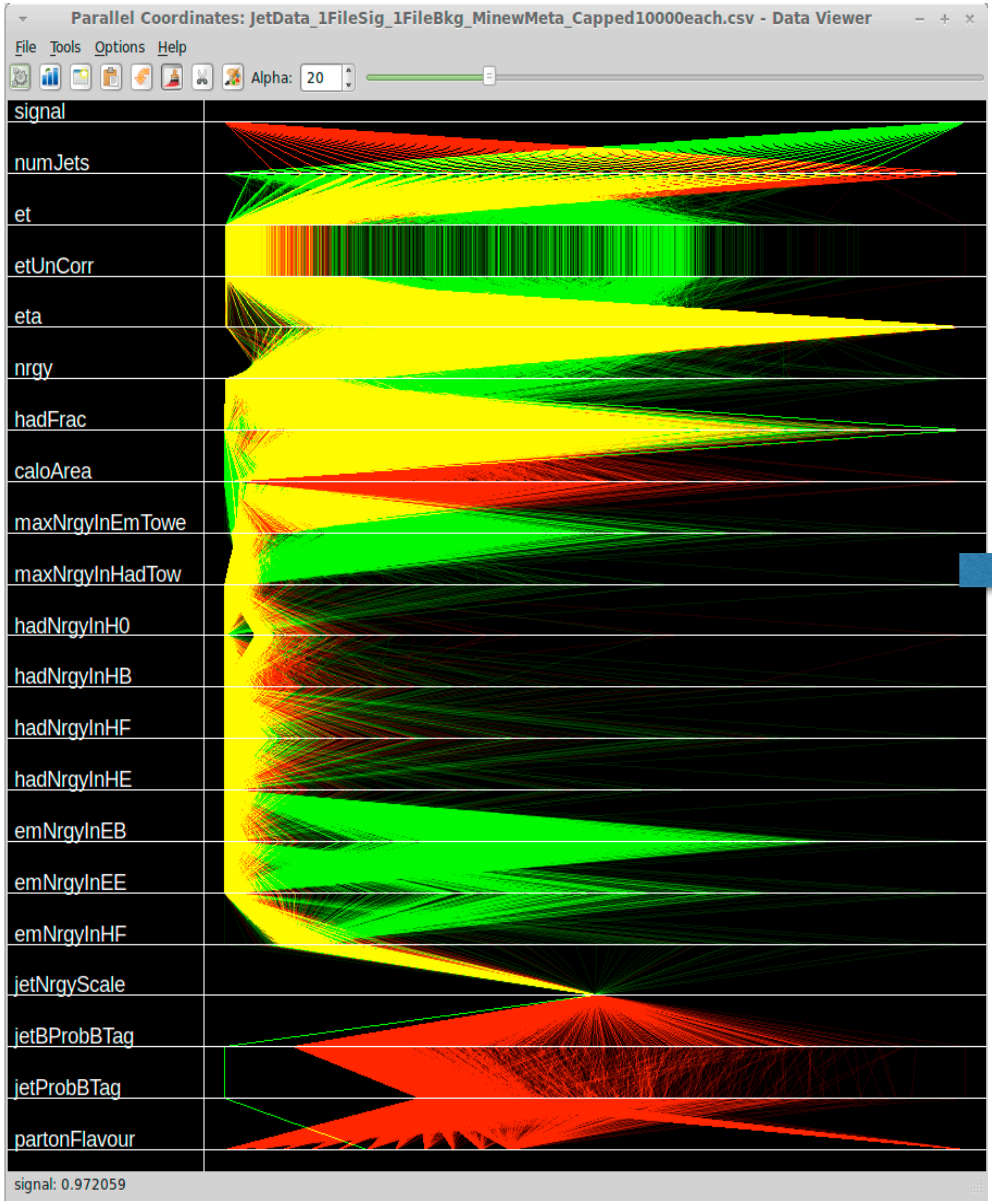


High transparency

Parallel Coordinates

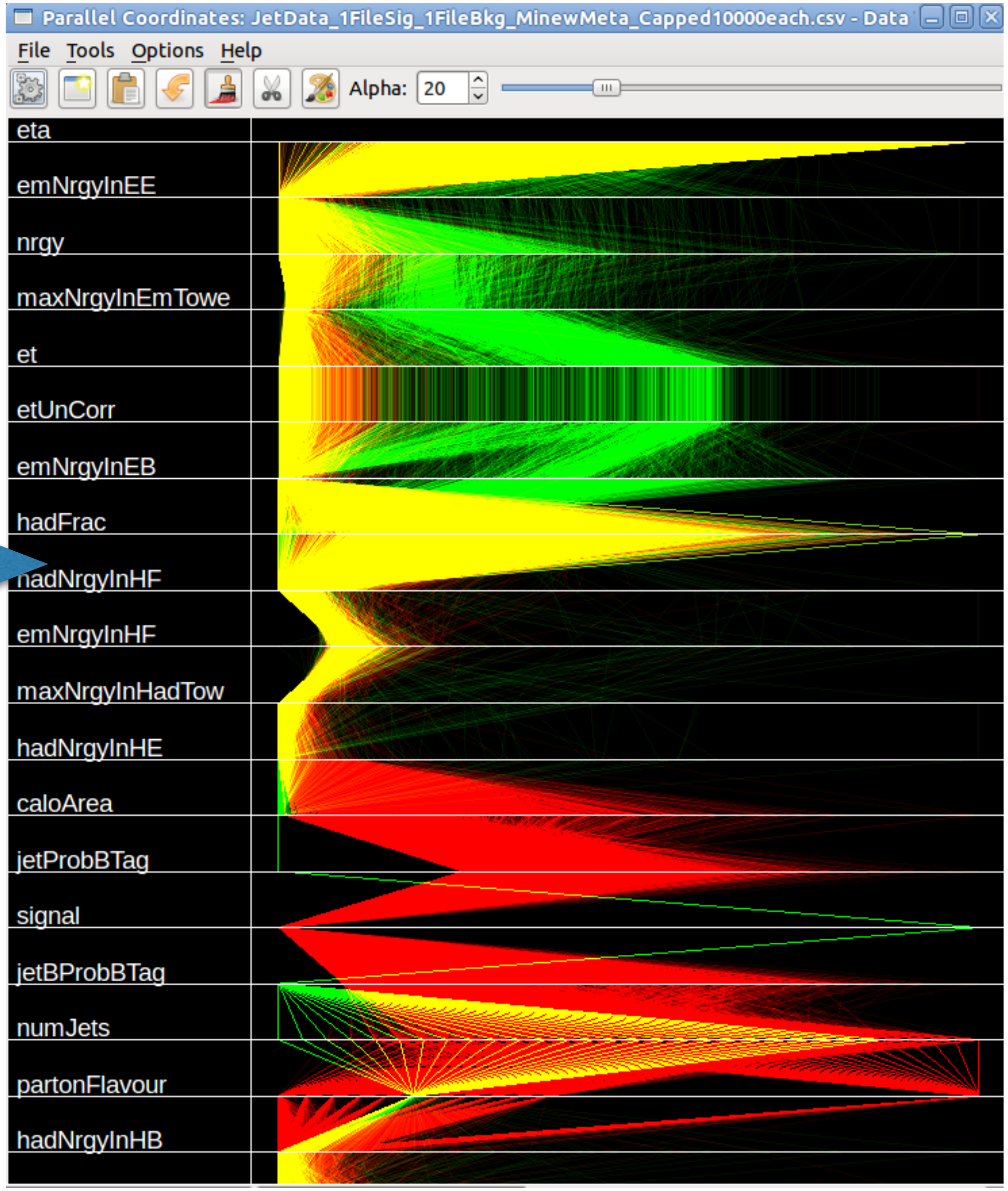
- The order of the axes is important - when two axes are separated (by a number of other axes) it becomes difficult to understand the relationship between those two variables
- One possible way to find an interesting layout of axes could be through the use of correlation coefficients between the variables:
  - $\rho_{xy} = \mathrm{cov}(x,y) / \sigma_x\sigma_y$
- Where the covariance, (cov) = $1/N\ \Sigma^{i=N}(x_i-\mathbf{x})(y_i-\mathbf{y})$ and the standard deviation, σ_x of a variable x is given as $\surd[1/N\ \Sigma^{i=N}(x_i-\mathbf{x})2]$
- Calculate the correlation coefficients between all the variables
  - Then use those to find interesting pairs of variables to place next to one another
  - Can then construct a chain of variables which, for example, uses the pair with the largest correlation coefficient magnitude, then the second largest etc. Instead of finding the order which gives the highest average

Parallel Coordinates


Parallel Coordinates: JetData_1FileSig_1FileBkg_MinewMeta_Capped10000each.csv - Data Viewer

Alpha: 20

signal: 0.972059


Parallel Coordinates: JetData_1FileSig_1FileBkg_MinewMeta_Capped10000each.csv - Data

Alpha: 20

Parallel Coordinates

- Advantages of parallel coordinates:
  - Ability to see all the multivariate data on one plot
  - Find interesting variables to investigate quickly
  - Find interesting data to investigate quickly (e.g. outliers that skew the datasets)
  - Seek patterns which may help classify the data
- Problems with parallel coordinates:
  - Ordering of the axes is important when looking at correlations between the variables
    - Tried to solve this using correlation coefficients – but not a complete answer as relationships between separated variables are still difficult to understand
  - May take a little while to get used to
  - Suffers from overplotting
    - However we can overcome this through using PC density plots

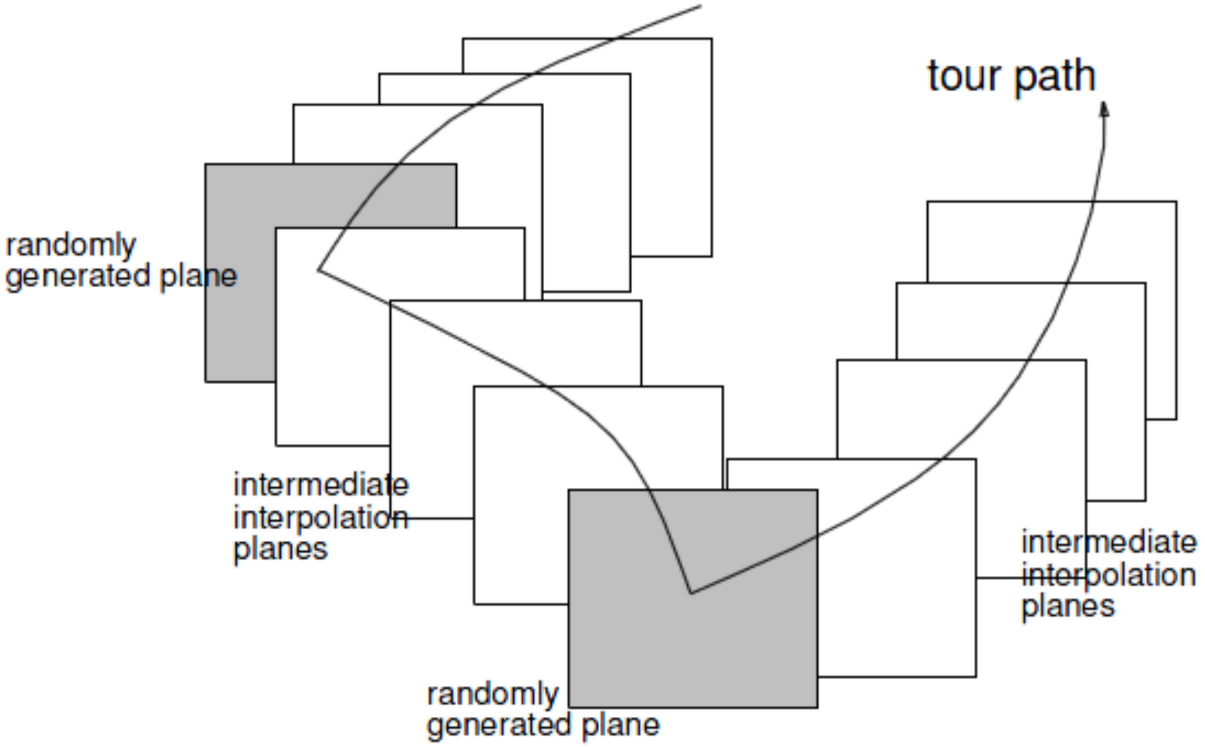
Movie

The Grand Tour

- Invented by Daniel Asimov in 1985 and by Asmiov and Andreas Buja in 1986
- The grand tour (GT) shows high dimensional data rotations, in a similar fashion to a 3D data rotation
  - But in a 3D rotation: rotate an object in space
  - While in higher dimensions: rotate a lower dimensional projection in the high dimensional space
- Rotating data in 2D is around a point
- Rotating data in 3D is around a line (axis)
- Rotating data in nD is around a hyperplane, where n>3
  - Hyperplane is an generalisation of a plane. Where the plane is defined as a 2D subspace in 3D
  - It is very hard to show the grand tour via slides, so I will use movies

The Grand Tour

- Goal: show a series of projections, originally 2D planes, of a higher dimensional space
- The series of projections are smooth to give the effect of a movie showing (close to) all the possible 2D projections of the data
- Unlike Projection Pursuit where the result is an index, the result of a grand tour is the movie itself


tour path
randomly generated plane
intermediate interpolation planes
intermediate interpolation planes
randomly generated plane

The Grand Tour

- The conditions of a grand tour
- Sequence of planes (projections) should:
    - be dense in the space of all planes - so is close to any 2D projection
    - become dense rapidly – by using an efficient algorithm
    - be uniformly distributed - so doesn't linger in one area
    - be continuous - to be comprehendible
    - be reconstructable - e.g. an interesting plane should be recovered easily after the tour
- The way the grand tour finds the interesting projections to show comes from which algorithm is used (e.g. Torus Winding Method)
- The algorithm used has to obey the conditions mentioned on the previous slide
    - It has to be a continuous, space-filling path through the set of 2D subspaces in p dimensional space

Movie

The Grand Tour

- Using parallel coordinates, a grand tour of the data does not necessarily have to be via 2D projections
- Instead it would be a movie of p-n projections of a p-dimensional space; where n<p
- Find interesting hyperplanes to cut along or classify with
    - These are the same as a decision boundary
- You watch a movie showing a series of projections (2D or higher if using parallel coordinates) of the higher dimensional space and pause it when an interesting projection appears

Movie

The Grand Tour

- Advantages of the grand tour:
  - Can lead to interesting hyperplanes in which a cut or classification can be made
  - Easy to use due to the automation – you watch a movie!
- Problems with the grand tour :
  - Difficult to understand the high dimensional rotations
  - Complicated underlying mathematics

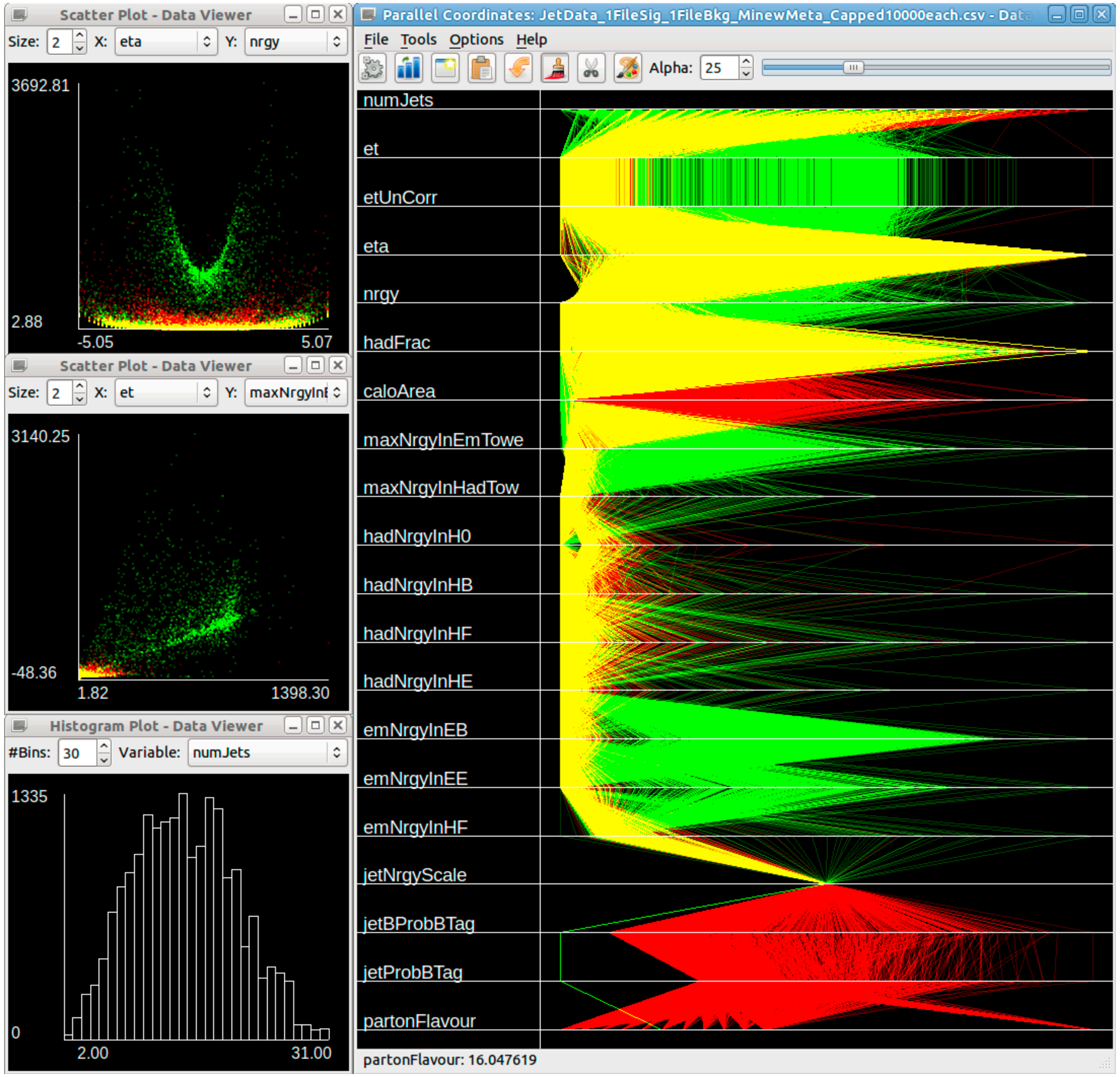
Visualisation

- Linked Plots:
  - Using multiple visualisations which are linked, so when modifying the data in one plot, all other plots are updated automatically
  - Gives the users the ability to explore their data and find patterns interactively → Exploratory Data Analysis
- Brushing
  - Highlighting data instances with a colour in one plot changes the same instances in other plots with that colour
  - Helps you to understand the links between variables, and other plots
- Pruning/Deleting Data
  - Deleting data from one plot; the other plots are updated with the relevant data removed
  - Useful when the data might be skewing the plots (for example with an erroneous point/line)

Software (well 6 years ago)

- Ggobi: http://www.ggobi.org/
- weka: http://www.cs.waikato.ac.nz/ml/weka/
- ROOT*: https://root.cern.ch
- RapidMiner: http://rapid-i.com/
- Orange: http://www.ailab.si/orange/
- DataViewer

*Parallel coordinates: tutorials/tree/parallelcoord.C



Scatter Plot - Data Viewer

Size: 2 X: eta Y: nrgy

Scatter Plot - Data Viewer

Size: 2 X: et Y: maxNrgyInE

Histogram Plot - Data Viewer

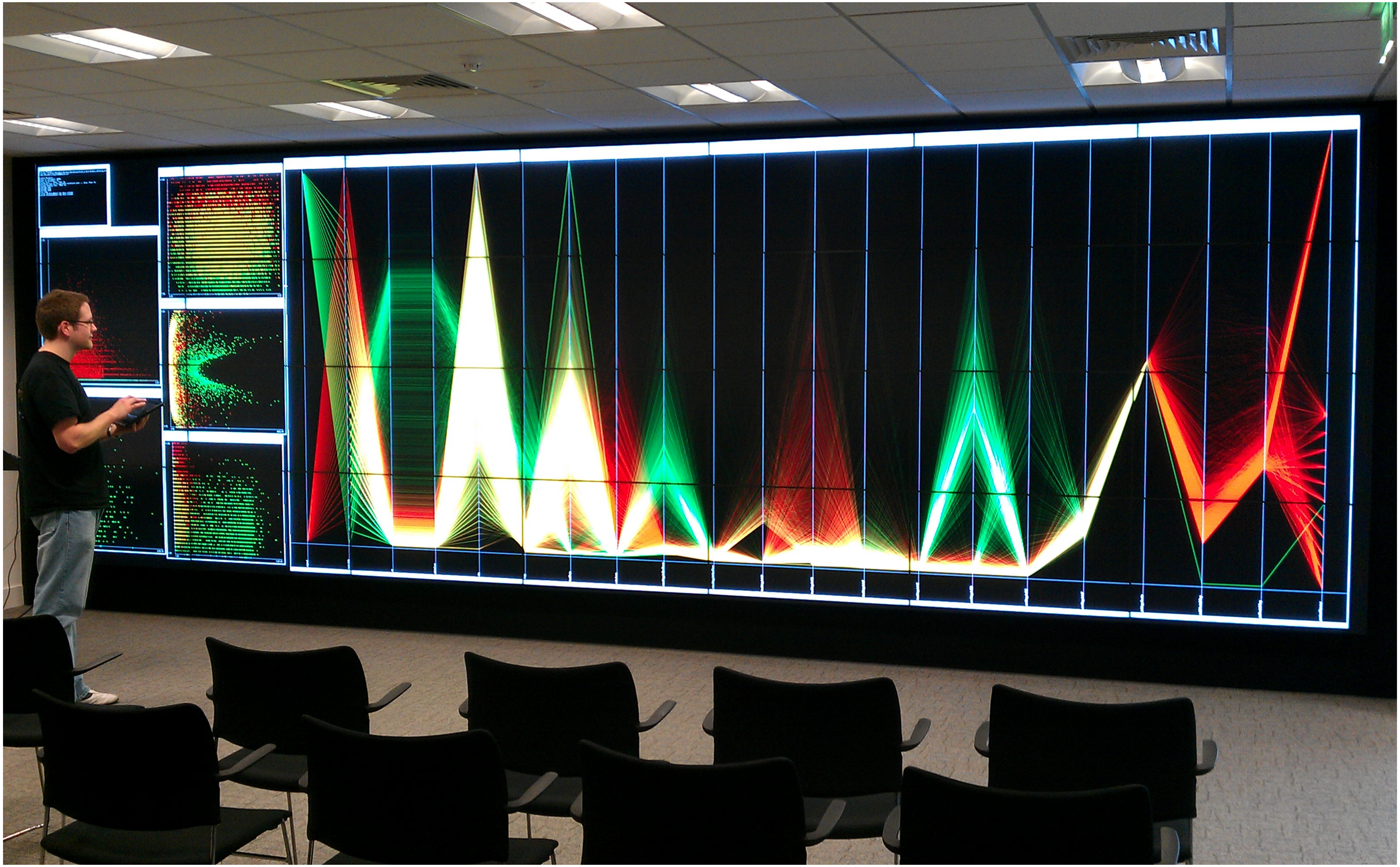
#Bins: 30 Variable: numJets

Parallel Coordinates: JetData_1FileSig_1FileBkg_MinewMeta_Capped10000each.csv - Data

Alpha: 25

partonFlavour: 16.047619

Software



Extra

The Grand Tour

- Work on using computer graphics to view projections of high dimensional data started at SLAC in the 70's and 80's
- Started with Mary Fisherkeller, John Tukey and Jerome Friedman et al on the PRIM-9 system
  - Picturing, Rotating, Inspecting and Masking in up to 9D
  - Developed at the Graphics Interpretation Facility (GIF) at SLAC using particle physics datasets (bubble chamber)
  - Rotate pairs of axes and view the result of the rotation in a 2D projection
- This work lead to the idea of Projection Pursuit
  - Automatically finds interesting low-dimensional projections of multivariate data by optimising a projection index