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
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 Hartree-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 \textit{ab initio} calculation performed to date.

Examples of scientific visualisation

Contributors: Matt Challacombe and Eric Schwegler

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

The \textbf{pseudocolor plot} (right) is used to map temperature to color on the same planar slice.
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

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
1991 M. Friendly Mosaic Display and Categorical data
Systemization of data and graphs and graph algebras in an OO framework.

Steve Watts, CHEP06, Brunel University
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

Steve Watts, CHEP06, Brunel University
American Statistical Association

Have challenges each year
This is the 2006 one

Pollen Data Set
de 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

• 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?
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

Steve Watts, CHEP06, Brunel University
Introduction to Parallel Coordinates

This also shows the idea of brushing

Steve Watts, CHEP06, Brunel University
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
Now **brushing** - colour the data with chosen colours and **pruning** - cut data you do not want

Steve Watts, CHEP06, Brunel University
First lets PRUNE
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 let's 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.

\[ K_s \rightarrow \pi^+ \pi^- \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doca</td>
<td>Distance of closest approach</td>
</tr>
<tr>
<td>Rxy</td>
<td>Radius of cylinder for interaction region</td>
</tr>
<tr>
<td>Rz</td>
<td>Abs. half length of cylinder defining the IR</td>
</tr>
<tr>
<td>Cos_hel</td>
<td>Abs. Value of cosine of Ks helicity angle</td>
</tr>
<tr>
<td>SFL</td>
<td>Signed flight length</td>
</tr>
<tr>
<td>Fsig</td>
<td>Statistical Sig. of Ks flight length</td>
</tr>
<tr>
<td>Pchi</td>
<td>ChiSq prob of Ks vertex</td>
</tr>
<tr>
<td>Mass</td>
<td>Reconstructed mass of the Ks</td>
</tr>
</tbody>
</table>

CrystalVision – E. Wegman

Steve Watts, CHEP06, Brunel University
Let's 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

Steve Watts, CHEP06, Brunel University
Now brush the background – FLAG = 0
Brush signal RED and background GREEN
If they overlap RED + GREEN = YELLOW (yellow)
Now go to parallel coordinates - adjust alpha

CrystalVision (E. Wegman)
Has blending and control of intensity
VERY Powerful

Steve Watts, CHEP06, Brunel University
Note: See affect of turning alpha channel on and off

Note: Parallel Coords Vertical. Scales 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.

Steve Watts, CHEP06, Brunel University
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…

Steve Watts, CHEP06, Brunel University
Back in scatterplot space

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.

Steve Watts, CHEP06, Brunel University
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

Steve Watts, CHEP06, Brunel University
Grand Tour

Asimov

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

Steve Watts, CHEP06, Brunel University
CrystalVision GrandTour is very fast!
<table>
<thead>
<tr>
<th>Software</th>
<th>Site</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGobi</td>
<td><a href="http://www.ggobi.org">www.ggobi.org</a></td>
<td>No α-channel.GT, PC All Platforms. Access to R.</td>
</tr>
<tr>
<td>Mondrian</td>
<td><a href="http://stats.math.uni-augsburg.de/Mondrian/">http://stats.math.uni-augsburg.de/Mondrian/</a></td>
<td>Java. α-channel.</td>
</tr>
<tr>
<td>Orange</td>
<td><a href="http://www.ailab.si/orange">http://www.ailab.si/orange</a></td>
<td>Component based data mining. C++ and python scripting. PC.</td>
</tr>
</tbody>
</table>

Steve Watts, CHEP06, Brunel University
Orange polyviz visualisation – Fsig, Rxy, doca, (mass) key variables
Can also use VizRank algorithm to find selection variables.

Steve Watts, CHEP06, Brunel University
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?

Steve Watts, CHEP06, Brunel University