

# Visualisation for Data Mining

Antony Unwin

Department of Computer-Oriented Statistics and Data Analysis  
University of Augsburg, 86135 Augsburg, Germany  
antony.unwin@math.uni-augsburg.de

## Abstract

Modern computing power makes possible analysis of larger and larger data sets and many new methods have been suggested under the broad heading of Data Mining. Visualisation of data, of model-fitting, and of results plays an important part, but large data sets are different and new methods of display are needed for dealing with them. This paper reviews the standard problems in displaying large numbers of cases and variables, both continuous and categorical, and emphasises the need for improving current software. Much could be achieved by adding interactive tools like querying, linking and sorting to standard displays to provide greater flexibility and to facilitate a more exploratory approach.

## 1 What is Data Mining?

Large data sets are more and more common. Every organisation is able to collect and store vast quantities of information. Supermarkets have sales figures for individual items and for customers. Phone companies have details of every phone call made. Weather computers store records of all manner of meteorological data. Websites try to monitor internet usage. And so on and so on. There is no point in maintaining data sets unless some attempt is made to get information out of them.

Statisticians have always analysed large data sets, but what is meant by large has changed over the years with the increasing power of computers. Analyses which took months by hand fifty years ago can now be carried out in a second. Much larger data sets can be considered and new problems have arisen in consequence. Some standard statistical methods do not scale up well to the big data sets to be analysed nowadays. New ideas and new approaches are needed.

One term which has been heard more and more often in this connection in recent years is Data Mining. It is so new, that not all are agreed what it might mean. David Hand has suggested that any definition should include the qualification that Data Mining is usually applied to data sets which have been collected for another purpose, that, in other words, Data Mining analyses are secondary analyses of data. This has implications for the quality of the data and for the difficulties of interpreting and generalising any results obtained. Results should not be reported as if they were based on random samples

from a population of interest. Another unexpected characteristic of Data Mining to be born in mind is that the “best” results are not likely to be the ones that are of most interest. The strongest results will either be known already or superficially obvious. The results which were previously unknown and do not stand out require more careful elicitation and will appear further down any list of outputs from Data Mining analyses. This suggests combining both aspects of Data Mining in the following definition:

Data Mining is the secondary analysis of large data sets looking for secondary results.

Computer scientists use Data Mining to describe methods which automatically search data sets for “interesting information”. Statisticians tend to use the term with a slightly negative tone to describe searching large data sets for anything of interest. Both groups have an important part to play: the computer scientists contribute fast and efficient methods for exploring the data; the statisticians contribute ways of assessing “interestingness” and the strengths of statistical principles of data analysis. It cannot be emphasised enough that every reliable optimal search algorithm will produce an optimum, but that does not mean that the result produced is worth considering.

## **2 The importance of visualisation**

One good way of assessing the value of results is to examine them visually and that should be a major application of visualisation methods in Data Mining. The phrase “should be” is used advisedly, as graphics are used far less than they might be, both at this stage of analysis and also at other stages where they might play a part: in investigating data quality, in identifying patterns or in suggesting structures. There are several possible reasons for this state of affairs. Statistical graphics are not underpinned by a formal theory, but are more a collection of useful tools than a solid structure which can be built on. (Lee Wilkinson’s book. “The Grammar of Graphics” and Adi Wilhelm’s research have only been published in the last year.) Graphics software for exploratory analyses is unsatisfactory. Software tends to concentrate on presentation graphics, which are unsuited for exploratory work. The available graphics software also tends to be poor for large data sets. Little effort has been made to develop graphics for the scale of problems met in modern data analysis.

This paper discusses responses to some of these arguments. Ways of scaling up classical graphics are described and software which enables effective exploratory analyses of large data sets will be illustrated. The visual approach complements more analytic methods and should be an essential component of Data Mining studies. Note that the displays considered here are for the raw data, so to speak, and attempt to work with the full dimensionality of the data set. There are other approaches (for instance biplots or projection pursuit displays) whose aim is to find informative views in new, lower dimensional spaces. Dimensionreduction methods are not discussed here.

Data visualisation should be central to Data Mining for another reason. Traditional statistical modelling assumes a clear goal to be achieved. Similarly, automatic search engines assume a stated optimisation criterion. Yet in Data Mining there are no specific goals, just the avowed aim to get some information — any real information — out of the data. Goals of analysis can include the identification of outliers, the definition of particular groups or clusters, and both deep analysis of smaller subgroups or sweeping generalisations about the whole data set. Visualisation is a flexible approach which encourages the consideration of several goals in parallel. It is therefore ideal for Data Mining.

As an example of the problems of displaying large data sets, consider the two scatterplots in Figure 1. Both show profit against amount for financial deals carried out by a bank. There are almost 1000000 points in the data set but for explanatory purposes the left hand plot is based on a subset of 17243 data points, while the plot on the right shows 9095 points from that subset close to (0,0). The right hand plot was drawn by reducing the y axis scale by a factor of  $10^3$  and the x axis by a factor of  $10^4$ .

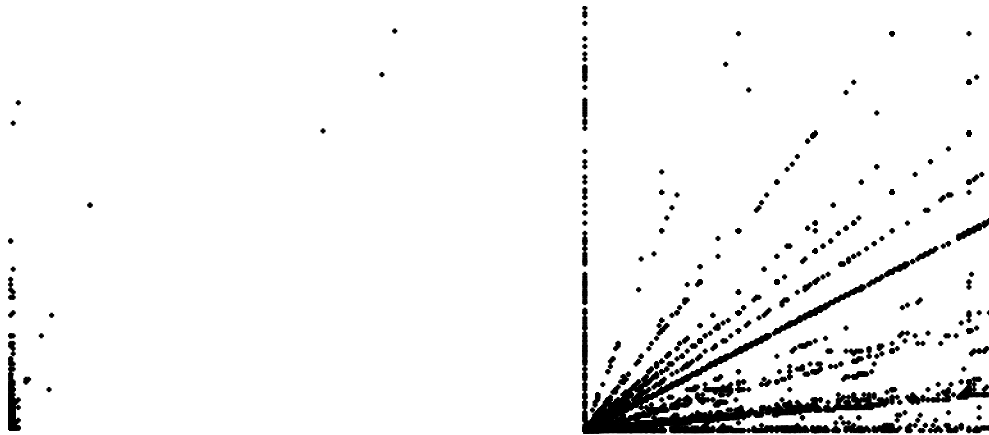


Figure1 Profit against amount traded for financial transactions in a bank. Scales have been removed for reasons of confidentiality.

There is no ideal static solution for these data, but a combination of querying, rescaling, zooming, linking and making use of multiple plots enables the information in the variables to be found relatively easily. As is typical of large data sets, there are many different pieces of information to be found (the outliers in the left-hand plot, the variety of linear relationships near (0,0) in the right-hand plot, clusterings of points etc) and many different views are required. The important thing is that these views can be generated quickly and flexibly to match the wide variety of possible features that might be in the data.

Many of the weaknesses of graphic displays in relation to large data sets can be got round with a combination of relatively minor adjustments and by making use of interaction as just suggested (though some displays, like stem and leaf plots do not scale up at all). A discussion of the basic interactive tools that should be available for any plot may be found in Unwin [1999].

Although the ideas of interactive statistical graphics have been around for some time (there was a good collection of articles published in Cleveland and McGill as early as 1988), they have not yet come into common use. The reason is simple. It is hard to write good interactive graphics software and thus little is available. This is a great pity, as it has held back progress in the area. As Swayne and Klinke remark in their introduction to the recent special issue of Computational Statistics [Vol 14, #1 1999] devoted to interactive statistics, it is surprising how little some statisticians require of a system to call it interactive.

## 3 Statistical graphic displays for large data sets

### 3.1 Displays for single continuous variables

The common graphic displays all have weaknesses in displaying large data sets. Dotplots are good for data sets with up to perhaps 100 cases, mainly for identifying individual cases and for showing any gaps in the data. With larger numbers of cases it is impossible to identify individuals and there are rarely any gaps. Boxplots are useful for identifying outliers, but consider what happens as a data set grows. An empirical distribution of 100 points might have 2 or 3 outliers, but a much larger sample of say 100,000 from the same kind of data distribution might then have 2000 to 3000 outliers. The advantages of individual identification are then lost.

Histograms may be regarded either as a crude form of density estimation or as a data analysis tool. Recommendations on bin-width or numbers of bins based on various theoretical considerations may be found in Scott [1992]. For instance, for the  $x$  variable in the data set of 17243 cases in Figure 1 his recommendation would be 586 bins! This would permit about 1 pixel width per bin on a laptop screen. Of course, the problem is the extreme outliers and most of the 586 bins would be empty. Without the top 3 outliers the number of bins recommended goes down to 548, although that is still excessive. Without the top 12 outliers it goes down to 161. Rather than relying on such theoretical models it makes more sense to experiment interactively with varying the bin-width. When this can be done quickly, a large number of different plots can be riffled through to see which ones provide information about the data.

Histograms are good for revealing information about the shape of a distribution, but not for showing outliers. When data sets become large, bins with very few cases can no longer be seen. Histograms scaled by default then tend to have long, empty tails to the right. Experienced users know that there must be some points there, but there is no clue where they are. In the software MANET [Unwin et al 1996] we have used redmarking to get around this. A red line is drawn under any bin which contains too few cases to be made visible. Figure 2 shows an example for the amount traded variable together with an accompanying dotplot to confirm where the points are.

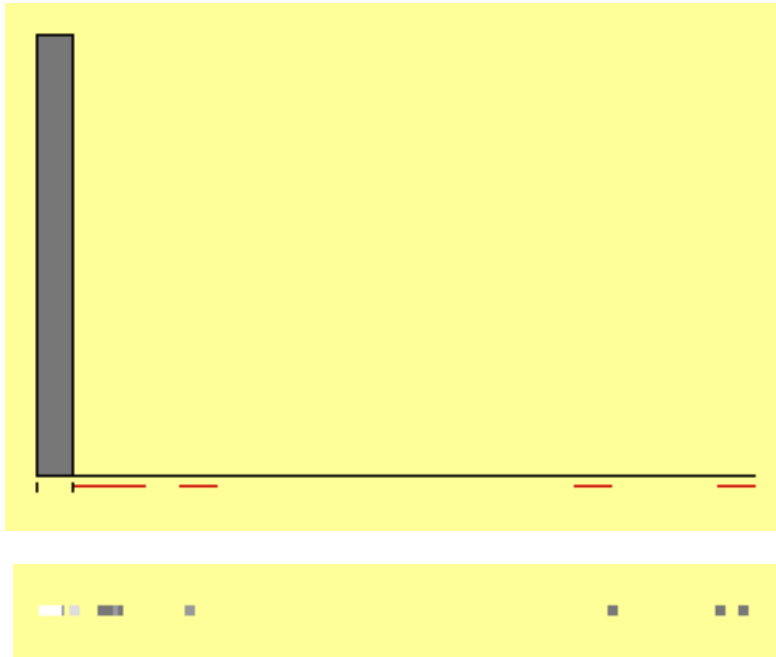


Figure 2 A histogram and dotplot for amount traded.

### 3.2 Displays for categorical data

Although some software packages offer displays using collections of individual points for categorical data, these are obviously unsatisfactory (and are a function of the software's inability to produce area displays). Counts in categories are better displayed as areas proportional to the counts. There is therefore no direct limit to the data set size, although there are two problems. Firstly, as with histogram bins for small numbers, the area for a small count may be too small to be shown in a large data set. Secondly, as data sets grow it is not unusual for the numbers of categories to grow as well. To take a common, but possibly surprising example, two categories, male and female, should suffice for the variable "gender" in a small data set. In a large data set there may be various kinds of "not known" plus women who were formerly men and men who were formerly women. In the financial data set there were many currencies recorded in at least one deal, but most of these arose very rarely. Bar charts are good ways of displaying numbers in categories, but not when there are too many categories. The solution is to provide simple interactive tools to:

- query the display to provide information on the categories;
- order the categories by counts or by highlighted numbers of proportions to group the most important ones together;
- combine selected smaller groups into single categories to improve the visibility of the display.

In Figure 3 there are 27 different currencies represented, but 16 of them are each less than 0.5% of the total individually and they amount to less than 1% of the total together. The bigger of the columns labelled G has been queried to show that it represents pounds sterling (GB Pounds) and that none of the 1984 deals in that currency are currently selected.

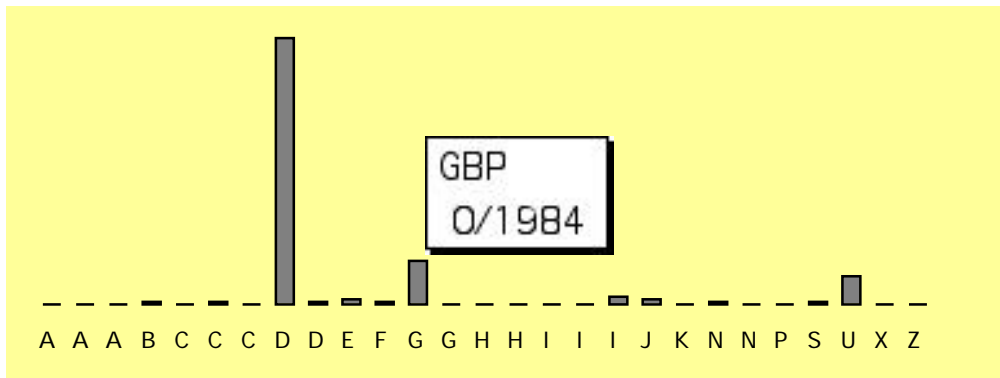


Figure 3 Bar chart of currencies by numbers of deals.

After sorting by size and then grouping the smallest columns together, the graph in figure 4 is obtained. This is much clearer and it is now possible to read the currency labels directly. What is innovative here, and difficult to show, is that the steps from Figure 3 to Figure 4 are performed with direct manipulation commands and are fast. It is therefore easy to experiment with different orderings and with different groupings to ensure that little distortion takes place and that no valuable information is lost.

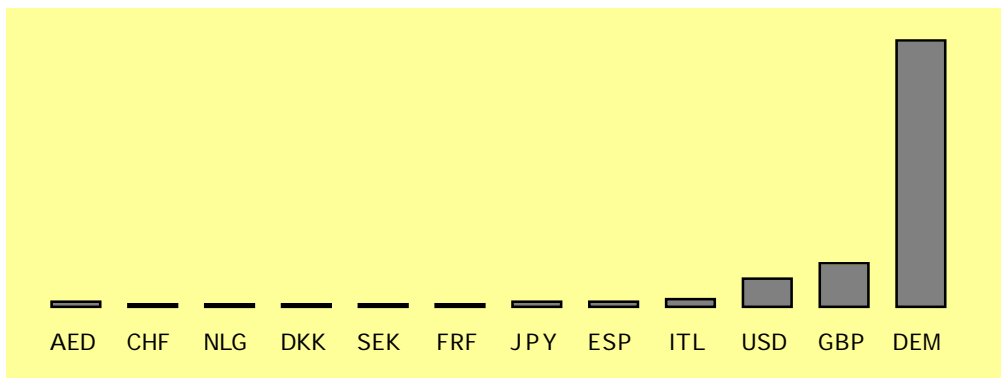


Figure 4 Bar chart of currencies by numbers of deals after sorting by count and then grouping together the smallest 16 into one new category on the left.

One final point is important to note in relation to the display of categorical variables. Data sets with no continuous variables may sometimes be stored in weighted groups rather than as individual cases. For instance, the Titanic data set requires 2201 lines when stored by individual, but only 24 lines when stored as a four-dimensional data cube with counts as weights. It is then essential that weighted displays can be drawn, i.e. displays based on aggregated weights and not on line counts. It is a surprising fact that such weighted displays are not generally available in software. The requirement has existed for a long time and has nothing specially to do with Data Mining.

### 3.3 Multivariate continuous data

For the two variables considered in the introduction, it was suggested that many plots should be drawn and not just one scatterplot. Sometimes that is not possible and the question arises, what should be done if only one plot can be drawn. It seems obvious that some kind of density estimation would be

most appropriate, although the method would have to be fast. A number of years ago, when computers were not as powerful as they are now, Carr [1987] suggested a hexagon binning of the data with colour to display density. Wills [1992] outlined a related scheme but based on individual pixels and using brightness. The additional attractiveness of Wills' idea is that the points in the scatterplot can be grown and shrunk interactively, thus allowing the analyst to view a range of crude density estimations quickly. The next stage has to be the implementation of more sophisticated density estimations. It is only a matter of having an algorithm which is efficient enough to recalculate instantaneously when the plot is changed. Interactive tools have to be fast, otherwise they distract rather than assist.

Rather than attempting to display all data in one plot, some authors have suggested Trellis graphics, breaking up the data into smaller chunks, based on combinations of other variables, and plotting each chunk separately. This approach could be interesting with interactive control, particularly over the choice and division of the other variables, but it has been recommended as a static tool. Users are supposed to print out their many plots and study them at their leisure. The concepts underlying trellis graphics and mosaic plots are related, but trellis graphics are static, tend to use discretised versions of continuous variables and emphasise dot displays. Mosaic plots are interactive, but are restricted to categorical variables and emphasise area displays. Each approach has something to learn from the other.

For the display of many continuous variables a good solution is to use parallel co-ordinates but with restricted drawing of lines. Here again, the ideas are well-known (see Inselberg [1998] and also Wegman [1990]) but commonly available implementations in software are not, not even for small data sets. Large data sets cannot be drawn in full (100,000 lines across the whole screen are far, far worse than 100,000 points and they are bad enough) but interactive exploration of subsets using techniques like the hot-set selection introduced by Data Desk would be very effective.

### **3.4 Multivariate categorical data**

Mosaic plots are a relatively recent graphic solution for displaying multivariate categorical variables. Since they are an area display, there is no limit to the number of cases which may be plotted, though the precision will, of course, be affected. A standard Mosaic plot is constructed sequentially as follows. First the X-axis is divided into rectangles of equal height and width proportional to the categories of the first variable. Then each of these rectangles is divided into rectangles of equal width but differing heights proportional to the second variable conditional on the categories of the first. The areas of the resulting rectangles are proportional to the numbers in the combinations of values from the first and second variables. The process is continued alternately on the X-axis and the Y-axis so that in principle combinations of any number of variables may be displayed. In practice, even 8 binary variables will lead to 256 possible combinations and many will be empty. On the other hand, displaying the combinations from two variables which both have many categories will also lead to complex diagrams. Mosaic plots are still effective, but editing and grouping is essential. Three variables each with 20 categories, of which 16 are tiny, will produce 8000 possible combinations of which only a few may have substantial counts. Considering

two variables from the financial data set, currency and type of deal, gives 1350 combinations of which only 190 have even 1 entry and 7 cells contain over two thirds of the cases.

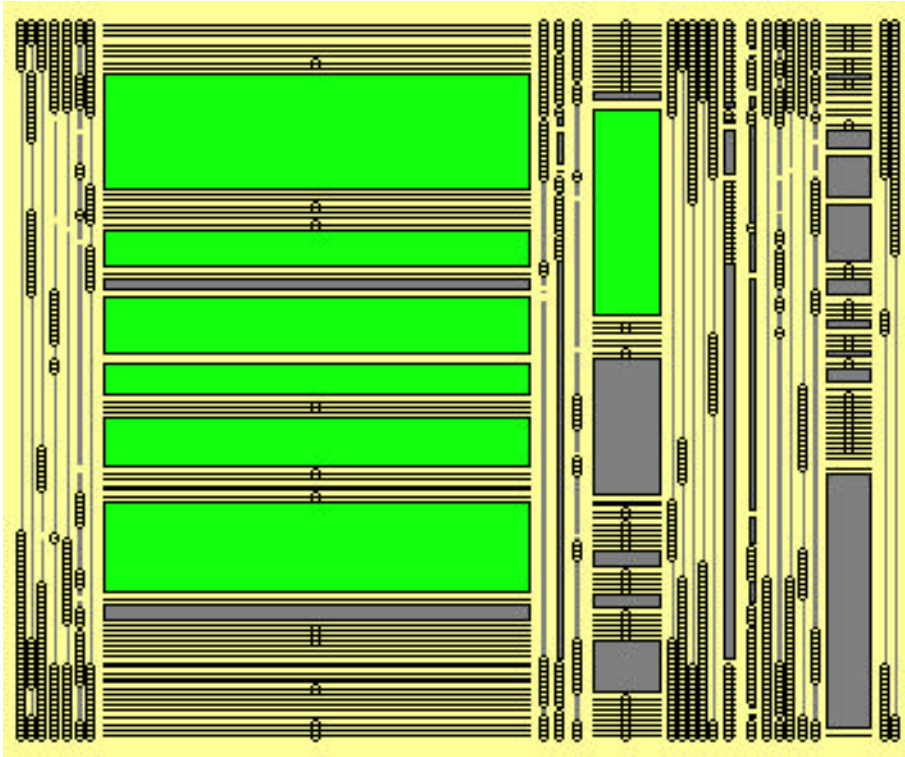


Figure 5 Mosaic plot of currency and type of deal, with the 7 combinations containing the most cases selected (in a greyscale image these cells are lighter).

Heike Hofmann [2000] has shown that mosaic plots may be extended and generalised in a number of ways which lead to more insightful pictures. Same bin-size mosaics do not show the counts of cell combinations, but are very good for showing clustering structures and gaps in the data. Fluctuation diagrams display rectangles whose area is proportional to the count of the cell they represent, but are drawn at a grid point so that it is easier to identify the multivariate structure. When there are many empty cells (as in Figure 5), this does not work as well as one would wish, as even the largest cell is so small as to be difficult to see. Using both types of diagram in parallel is very effective.

Two interesting generalisations are worth noting. Firstly, any non-negative variable can be used as a weighting variable instead of counts. In the financial data set, this meant that amount traded and profit could both be used. Needless to say, the mosaic plots weighted with these variables gave quite different pictures to the one based on numbers of deals. Secondly, there is no fixed rule that says the axes need to be used alternately. Indeed, one of Heike Hofmann's most interesting developments is the Double-Decker plot which splits up only along the X-axis.

## 4 Data Mining methods and visualisation

Up till now there has not been much use of visualisation in Data Mining and there are few graphics of any consequence available in Data Mining software. This may be because many methods work with only discrete variables (so that continuous variables have to be discretised) and because graphical tools for multivariate categorical data are not well-known or considered ineffectual. With mosaic plots this all changes. Mosaic plots may be used very effectively to investigate association rules using linking. They are good for studying overlapping sets of results, such as arise from decision rules. They provide insight into the structures derived by decision tree methods. They are valuable for deriving descriptions of clusters which have been identified.

Association rules are an interesting approach to scan a data set of many variables quickly, but they produce far too many results, so that a filtering is needed. One approach is to display all rules discovered in a scatterplot of confidence against support as in Figure 6.

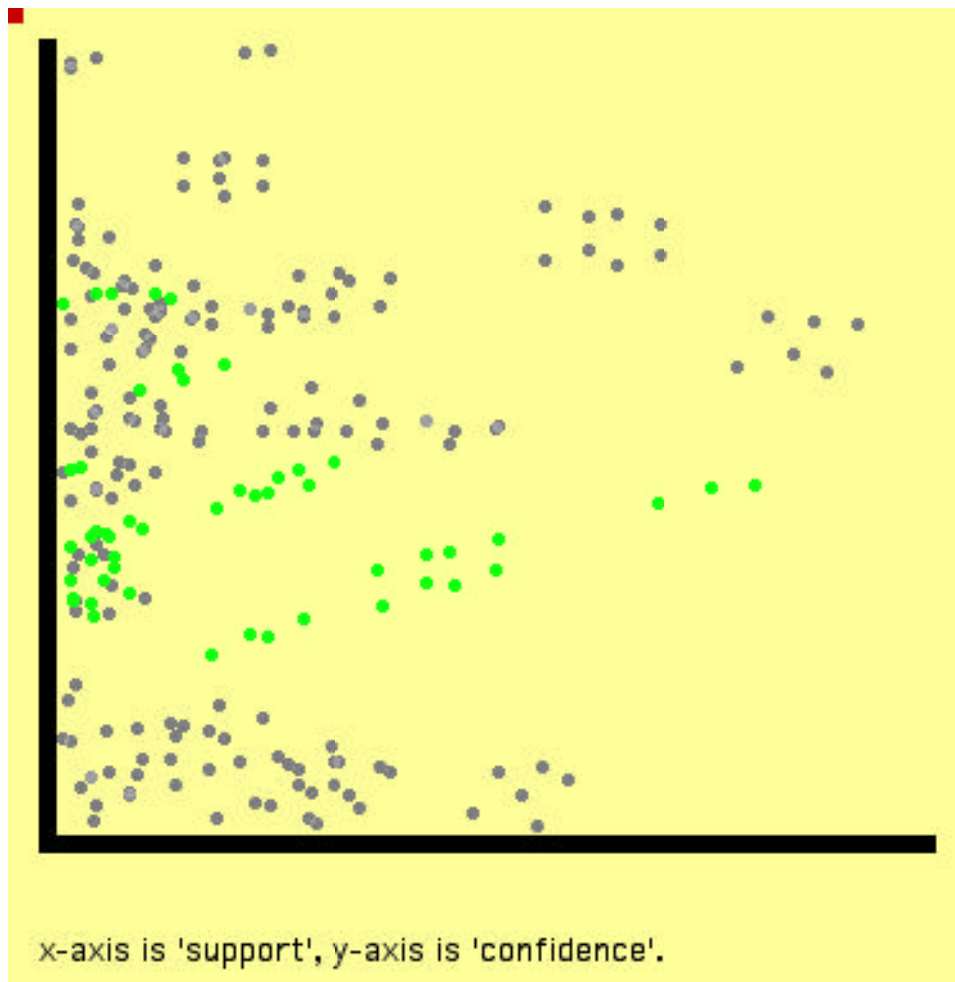


Figure 6 Confidence plotted against support for 228 association rules. The 52 rules with one particular RHS have been highlighted.

This plot alone is unhelpful. It is when it is linked to information defining the rules (what is the RHS variable as in Figure 6, or how many and which variables are on the LHS) that it comes into its own. Rules in the top right of the plot have high confidence and high support (and probably therefore represent well-known information). Rules in the bottom left have only just passed both criteria (minimum Support and minimum Confidence) and are likely to be less interesting than related rules to the right (higher support) and above (greater confidence). In Figure 6 we can see that most of the highlighted rules have approximately the same support so amongst these the group of three with the most confidence would be the best to look at more closely. There is a subgroup with high confidence, but low support that might also be of interest. Final judgements will depend on the practical interpretation of the rules and interactive exploration with linking is an effective aid to this kind of evaluation.

A weakness of association rules is that they only concern subsets of variable categories, so that knowledge of the relevance of the other categories is lacking. Mosaic plots allow you to study association rules in context so that the relative strength of the rules can be checked. A recommended procedure is to draw a Double-Decker plot for all categories of the set of variables involved in the rules of interest. It immediately becomes apparent if the rule is clearly superior in confidence to any neighbouring rule. Patterns of confidence and support amongst closely related rules are also easy to detect.

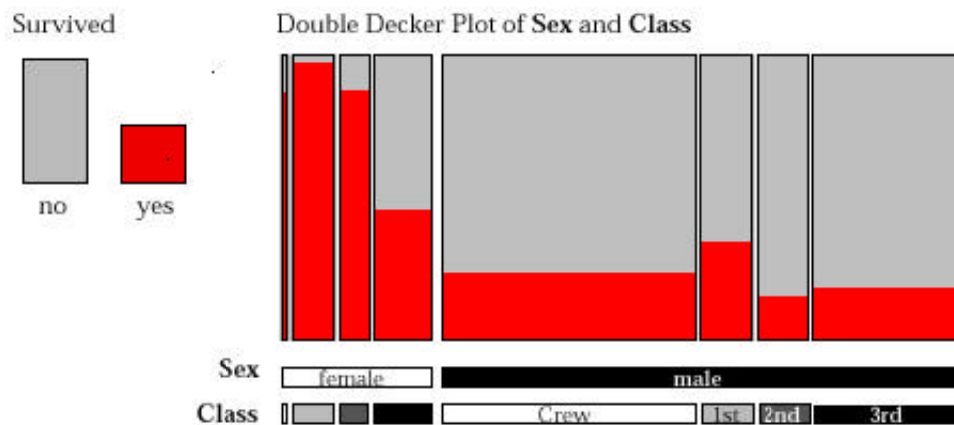


Figure 7 A Double-Decker plot of the Titanic data with survivors highlighted.

## 5 Conclusions

Until statistical software includes fully interactive graphics as standard, it will be difficult for most people to visualise large data sets effectively. No static displays can do justice to the full range of information in the data. Adding interaction properly is not just a matter of adding new modules, as one might add a new algorithm, but requires extensive interface development. As we know from developments in other software areas, this is not easy.

While there is always scope for new graphical displays, the most successful advances are more likely to come from adding additional interactive control

to the standard displays. Novel, complex displays tend to remain novelties, although it may well be that adding interaction will revive some old ideas. To some extent this is what happened with mosaic plots and there is still much progress to be looked forward to there. We should also be able to look forward to new interactive advances in both trellis displays and parallel coordinates.

Large data sets are important and present interesting new problems for visualisation. There are real opportunities to make significant progress.

## References

- Becker, R., Cleveland, W.S., Shyu, M-J. (1996). The Visual Design and Control of Trellis Display. *JCGS*, 5, 123-155.
- Carr, D. B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S. (1987). Scatterplot Matrix Techniques for large N. *JASA*, 82(398), 424-436.
- Cleveland, W. S., and McGill, M.E. (Ed.). (1988). Dynamic Graphics for Statistics. Pacific Grove California: Wadsworth & Brooks/Cole
- Cleveland, W. S. (1993). Visualizing Data. Summit, New Jersey, USA: Hobart Press.
- Cleveland, W. S. (1994). The Elements of Graphing Data (Revised ed.). Summit, New Jersey, USA: Hobart Press.
- Hofmann, H. (2000) Graphical Tools for the Exploration of Multivariate Categorical Data. Augsburg.
- Inselberg, A. (1998). Visual Data Mining with Parallel Coordinates. *Computational Statistics*, 13(1), 47-63.
- Scott, D. W. (1992). Multivariate Density Estimation. New York: Wiley
- Tufte, E. R. (1983). The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphic Press
- Unwin, A. R., Hawkins, G., Hofmann, H., and Siegl, B. (1996). Interactive Graphics for Data Sets with Missing Values - MANET. *Journal of Computational and Graphical Statistics*, 5(2), 113-122
- Unwin, A. R. (1999). Requirements for interactive graphics software for exploratory data analysis. *Computational Statistics*, 14, 7-22.
- Velleman, P. F. (1997). Data Desk. Ithaca New York: Data Description.
- Wegman, E. J. (1990). Hyperdimensional Data Analysis using Parallel Coordinates. *JASA*, 85, 664-675.
- Wilhelm, A. F. X. (2000) Interactive Statistical Graphics: The Paradigm of Linked Views. Habilitationsschrift, University of Augsburg.
- Wilkinson, L. (1999). The Grammar of Graphics. New York: Springer.
- Wills, G. (1992) Spatial Data: Exploration and Modelling via Distance-Based and Interactive Graphics Methods. Trinity College Dublin.