Visualization of data

R Daniel Meyer* and Dianne Cook†

Data visualization has developed in several directions: theoretical; methodological; and in new application areas. Advances include the development of a grammar of graphics, deeper understanding of human perception and implications for graphical layout, and better approaches to visualizing multi-dimensional data and large data sets. Gene expression is a notable new application area for visualization of large data sets.

Addresses
*Mathematical and Statistical Sciences, Pfizer Central Research, Eastern Point Road, Groton, CT 06340, USA;
e-mail: daniel_meyer@groton.pfizer.com
†Department of Statistics, Iowa State University, Ames, IA 50011, USA

Current Opinion in Biotechnology 2000, **11**:89–96

0958-1669/00/$ – see front matter

Introduction

Analytical chemistry, high-throughput synthesis and biological assays, genomic analysis, protein and gene expression arrays, and so on, have all spawned an explosion of data in biotechnology. Unusually large data sets have been identified as an important research area for the past many years. Now these large data sets are no longer unusual. They have become part of the typical scientist's everyday life. Software to visualize these data has become part of the scientist's standard toolkit.

Visualization can be defined as the use of computer-supported, interactive, and dynamic visual representations of data to amplify cognition. Goals are discovery, decision-making and explanation [1•]. We focus on the visualization of scientific, physical or measurement data. Imaging, either of chemical structure or biological sequence data, for example, is outside the scope of this review. For our purposes visual representations are statistical plots of data. We reserve the term graph specifically for a plot composed of connected points and lines, as are now used for network visualizations. The term interactive refers to the direct manipulation of graphical attributes in a plot, for example, directly changing the color of a group of points by 'brushing'. Dynamic graphics refers to smoothly changing views of data that are analogous to real-life motion, or animations of plots based on parameters such as time.

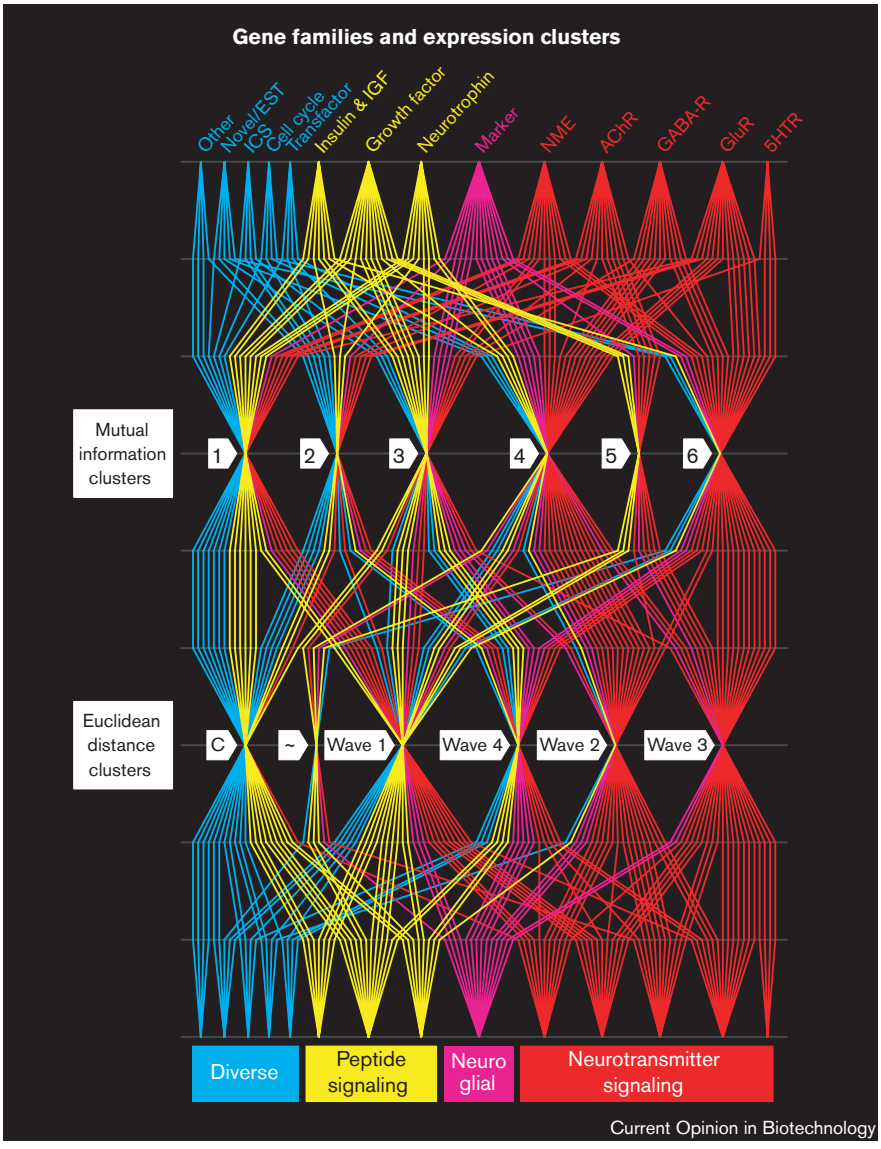
Computer-assisted data visualization has been an active area of research since the early 1960s, as hardware and software developed for creating plots. An early milestone was the advent of Exploratory Data Analysis (EDA) [2], a suite of quick tools for visually and numerically summarizing data. It also marked an early contribution from the 'Bell Laboratories' school. This laboratory gave rise to many important advances in data visualization, and research continues despite the fracturing of the parent companies. The three resulting labs — AT&T Labs, Lucent Technologies' Bell Labs and the re-sold Bellcore — have been the most influential sources of advances in statistical graphics and data visualization approaches in recent times, as evidenced by the number of references to current or former researchers there [2–5,6•,7,8•,9–11,12•–14•,15,16,17•, 18–20]. Major advances include Exploratory Data Analysis [2], the S statistical language [19] that has evolved into an excellent statistical graphics engine, trellis graphics [20], the XGobi system for interactive and dynamic graphics [6•,9–11], the XGvis system for graph layout based on multidimensional scaling [12•], and interactive visualization of large networks [7,15].

Our material is organized in the following categories: firstly, new methodology, including methods for large data sets, color matrices, qualitative data and model-based visualization; secondly, theory of visualization, which encompasses graphical inference, taxonomy of graphical tasks, perceptual, conceptual and aesthetic aspects, and grammatical formulations for describing plots; thirdly, new application areas, such as exploring relationships in chemical databases, and microarray data; and finally, descriptions of the growing body of Internet resources, software, and emerging technology.

Multi-dimensional data

Multi-dimensional visualization is aimed at viewing several variables in the same plot. The main approaches to multi-dimensional visualization include methods such as parallel coordinate plots (several parallel axes, each representing a single variable) [21•–23•], trellis displays [20], dynamic projections [9,11], as well as encoding data in colors, textures and glyphs in simpler plots [24,25•]. Linking between plots [4,5,26•,27] provides interactive mechanisms for connecting information provided by each plot. Parallel coordinate displays and dynamic plots can also be thought of as linked plots. In parallel coordinate plots [21•–23•], the linking is static, constructed by lines tracing the values of particular observations, across variables. (Figure 1 shows a parallel coordinate display from an analysis of gene expression data.) In dynamic graphics [4,10,11,23•], the motion of the continuously changing plot is really a rapid sequence of distinct plots, linked mentally by the viewer. When the plots are laid out side-by-side or in a grid for simultaneous viewing, Tufte calls these 'small multiples' [28]. The similarity of adjacent plots (adjacent in time or space) allows the viewer to make the connections. Trellis displays [20] lay out a grid of plots, each panel in the grid containing the same plot but on a different subset of the data. Trellis displays can also be considered a form of linked views, especially when the views contain

Figure 1



Hybrid parallel coordinates display of gene expression clusters described in [30]. The graph shown here has been improved over that in [30] by additional preprocessing to reduce line crossing.

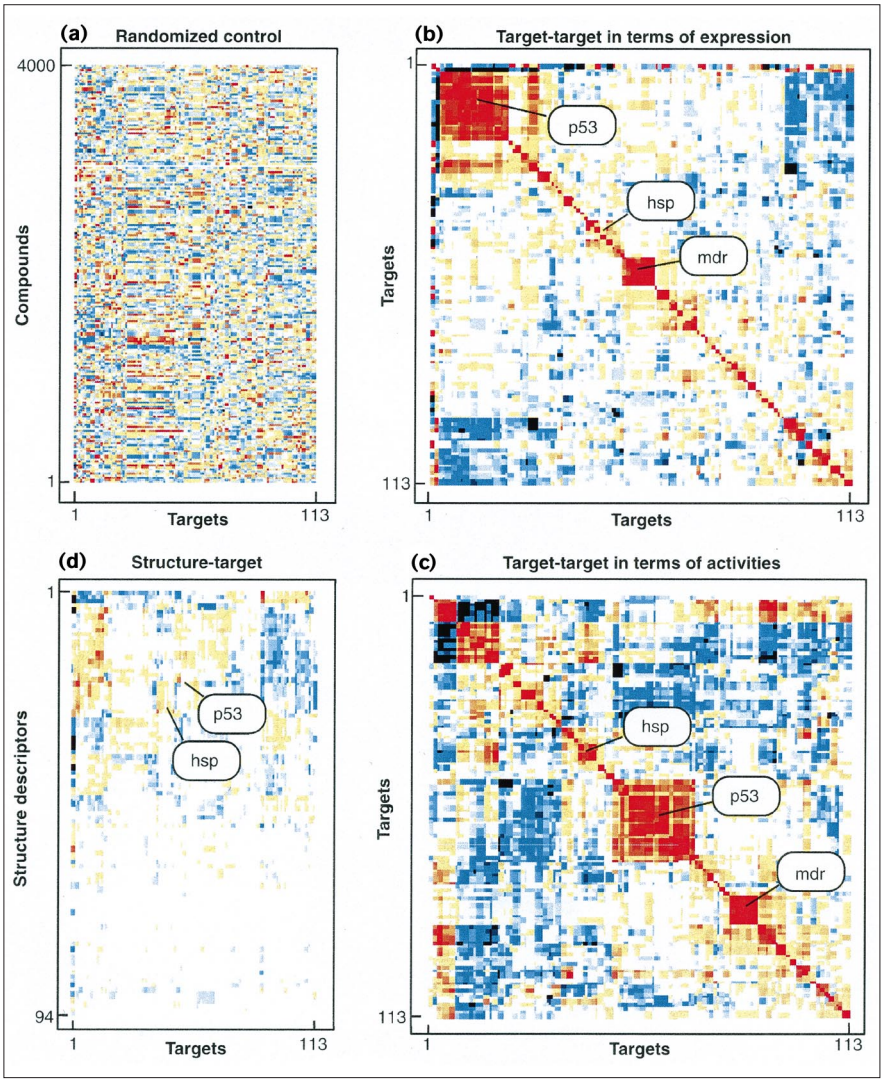
slates, or overlapping plots. The panels in a trellis display typically represent distinct nonoverlapping combinations of the subsetting variables. Slates are panels with overlap, for example, one panel is the subset of data with $(10 < X < 20;\ 90 < Y < 100)$, the next panel to the right defined by $(15 < X < 25;\ 90 < Y < 100)$, and so on. The effect is equivalent to that of small multiples.

Large data sets

When there are dozens or hundreds or even thousands of variables, there is a simple plot that is becoming popular, the color matrix. A data matrix is displayed as a rectangle, with each cell colored according to the numerical value of the corresponding matrix element. It has become popular for applications such as gene expression [29•,30,31,32••] and chemoinformatics [33] (see Figure 2), and Minotte and West [34] describe the approach for viewing clusters. With one pixel per cell of the matrix, hundreds of variables can be displayed simultaneously. Related approaches to visualizing large multivariate data sets and functions have also been described [24,35]. Critical to the success of reading these displays is the ordering of rows and columns. It is important to develop criteria for arranging columns and rows so that patterns can be revealed, unless the variables have a natural ordering, such as time. Typically, clustering algorithms, run independently on rows and columns, are used to obtain an ordering. This may find global correlations, but ignores subregions where similarities are strong. There is potential for advances here, both numerical and interactive, for ordering this type of display.

Data sets with large numbers of cases challenge the principle to show all the data in any plot [3,28]. Overstriking in areas of high data density in scatterplots, for example, dis-

Figure 2



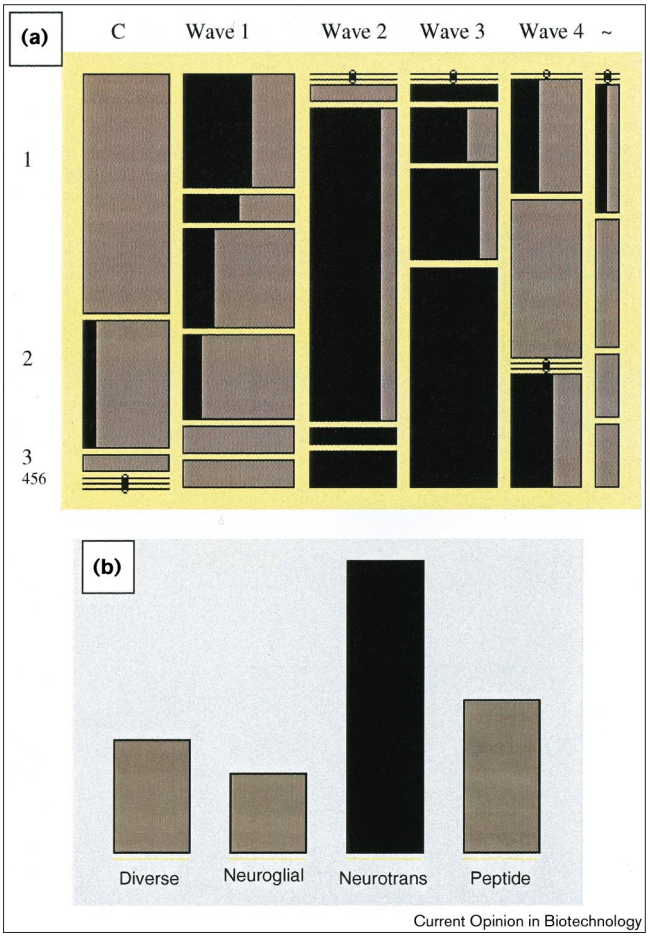
Four different cluster-correlation diagrams. **(a)** The plot for randomly permuted data, as a baseline for **(c)**, correlations between chemical structural descriptors of experimental drugs and biological molecular targets in cancer cells. **(b)** Correlations among different targets based on their expression levels in different cancer cell lines. **(d)** Shows the same as (b) but in terms of activity profiles across drugs. In all panels, the ordering of rows and columns is determined in advance by an agglomerative cluster analysis. Reproduced from [1•] with permission.

torts the display and can be misleading. Variable biplots [36] are simply scatterplots that employ an adaptive algorithm to replace groups of individual points with oversized glyphs only in regions of high data density, the size of each glyph is scaled to the number of points it replaces. It therefore preserves fine features and trends in areas of low density, features that tend to be lost in other agglomerative techniques. Other research for large numbers of cases involves preprocessing before display, such as binning or 'data squashing' (W DuMouchel, C Volinsky, T Johnson, C Cortes, D Pregibon, Squashing flat files flatter, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, August 15–18 1999) or clustering observations (R Rastogi, K Shim, Scalable algorithms for mining large databases, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, August 15–18 1999). Preprocessing refers to altering the data in some way to make it more amenable for display. For example, rounding values off to the nearest 10 or 100 is one case of binning.

Qualitative data

Mosaic plots were developed [37] as a graphical alternative for qualitative, or categorical, data. The gene family name associated with a gene (as in Figure 1) is an example of qualitative data. They display cross-classified data by constructing rectangles of area proportional to the counts. Mosaic plots tend to be unfamiliar to many scientists but they are likely to become more familiar and their use is likely to grow (see Figure 3). Good introductions, explanations, and links to implementations can be found [8•,34,38•,39•]. Mosaic plots are to categorical variables what scatterplots are to continuous variables, and their purpose is the same: to find interesting patterns of association between variables.

Figure 3



(a) Mosaic plot showing the relationship between clusters by two different methods, data from [30]. Linking is also illustrated, as the bar for neurotrans in the bar chart **(b)** is selected, highlighting areas in the mosaic corresponding to this gene family.

Mosaic plots can be used to diagnose and assess the fit of various categorical data models [8•,39•]. Mosaic plots will have a characteristic appearance corresponding to the type of model appropriate for the data. Also, deviations from the model can be shown in the mosaic plot, helping to suggest improvements to the model. Good interaction tools are critical to enhancing the power of these plots: introducing and removing variables on-the-fly, or altering the order of display of categories.

Visualization associated with modeling

O'Connell and Wolfinger [40] give some excellent examples of visually depicting relationships among variables using spatial regression models. In longitudinal studies, when distinct observations are recorded on each experimental subject at different points in time, recent work [41] provides an approach for visualizing the dynamic relationships. Eno and Terrell [42] give examples of graphics for logistic regression. Regarding neural network models, more closely connecting visualization may help to improve understanding of the results (D Cook, D Swayne, A Buja, Tutorial, Joint Meeting of The Western North American of the International Biometric Society and the Institute of Mathematical Statistics, Seattle, WA, 27–30 June 1999).

Missing values and graphics

There has been some excellent work on providing methods for exploring missing value structure in data ([43]; MANET software: linked views, based on histograms, mosaic plots, features tools for exploring missing values, hence the name Missings Are Now Equally Treated, http://www1.math.uni-augsburg.de/Manet/). Missing values are a common occurrence in most types of data, and the presence/absence or distribution of missing values strongly affects interpretation of results. MANET uses `missing' as one of the possible values of a qualitative variable, treating it no differently than other possible values. In this way, plots can show how missing values for one variable may be related to values of other variables. Exploring the distribution of missing values and results of imputation of values is an important exploratory activity. Imputation of missing values refers to filling in the values with estimates.

Visual classification

Cluster analysis, or unsupervised clustering, is a popular and intuitive analytical technique, important in many fields and used often in depicting chemical information. It involves searching data for natural groupings of items, where items within a group are closer or more similar than items from different groups. Graphics is very important for assessing the results, and also for leading the researcher through the analysis (D Cook, D Swayne, A Buja, Tutorial, Joint Meeting of The Western North American of the International Biometric Society and the Institute of Mathematical Statistics, Seattle, WA, 27–30 June 1999; J Schumi, Abstract 2788, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999; M Ankerst, C Elsen, M Ester, H-P Kriegel, Paper #292, Knowledge Discovery and Data Mining, San Diego, CA, Aug 15–18 1999). Osbourn and Martinez [44] have developed the visual-empirical region-of-influence (VERI) clustering algorithm to expressly reproduce the visual clustering ability of humans. The method has been extended to pattern recognition problems (G Osbourn, Abstract 285, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999) such as chemical recognition using partially selective sensors, and classification of magnetic resonance imagery (MRI) of the brain. Self-organizing maps [45] (SOMs) are used for both classification and displaying results, and have been used to cluster gene expression results [46••]. Their unique feature involves forming clusters on a graph, usually a two-dimensional grid. Neighboring clusters on the grid are then more similar than more distant clusters. Results are displayed on the grid, and for some people this is visually appealing. It serves a similar function to the dendrograms, which show the relationships among clusters from a hierarchical clustering analysis.

Virtual reality

It is an exciting time for new technologies, in particular, virtual reality. There have been several works on data visualization in virtual worlds [1•,6•,17•,25•,47]. An interactive and dynamic system for data analysis in a highly immersive virtual environment is described by Cook *et al.* [17•], and the results of an experiment comparing the desktop environment with the immersive virtual environment for several specific visualization tasks is described by Nelson *et al.* [6•]. For this very limited comparison, the virtual reality environment showed some advantages; for example, participants more readily identified visual clusters in data. Carr *et al.* [48] describe a system which can combine 2D display devices with 3D virtual displays for data analysis. It was used for gene expression data analysis. Virtual environments offer considerable promise for creating highly sophisticated visualizations that can include data analysis with contextual information, such as web pages, advertising videos, and geographical terrains. With the emergence of the Java 3D API, virtual environments may potentially become commonplace technology, rather than the current domain of the rich and well-equipped specialist labs. Typically, virtual environments are computationally intensive, even to simply render 3D worlds. Complex scenes can be rendered faster when much of it can be displayed as textures rather than as fully-defined 3D geometries. This has implications for data analysis also, particularly when there are large amounts of data, it may not be possible to display data at the point resolution. So an interesting development is translating multivariate data into textures, and interestingly some old approaches based on glyphs are being rejuvenated (in glyph displays each observation is mapped into some sort of figure such as a face, or star). (In a Chernoff face, for example, the dimensions of each of the several features [nose, mouth and eyes] are scaled according to the value of each of the variables. Each multivariate observation yields one face.) A field of glyphs can be easily turned into a texture map [25•].

Theoretical issues and frameworks

Developing taxonomic and structural foundations has helped data visualization to become a recognized and well-defined discipline spanning the fields of statistics, bioinformatics, information technology and computer science, and these foundations enhance its utility in a broad array of application areas.

Interactive and dynamic graphical methods have developed to a stage where it is possible to propose a taxonomy of tasks [4]. Major activities include rearranging multiple different plots, linking information between views, and focusing via tools such as pan/zoom. The most widely developed interactive graphical activity is linked brushing. Brushing refers to the interactive selection of a subset of the data within one plot, usually using the mouse, and linking highlights that same subset of data in a second (and third, etc.) plot. There have been several papers detailing the taxonomy of selection procedures, and validating the interpretation of such conditional operations (AFX Wilhelm, Abstract 262, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999; [49,50]). For dynamic graphics [11], there are algorithms for generating sequences of projections of high-dimensional spaces, providing motion sequences of multidimensional data. These can be generated 'on-the-fly' to provide dynamic movies of data, for example, the grand tour (available in [9,48,51]), or laid out like Andrews curves [52]. Grand tour refers to the method moving smoothly and continuously between random projections of the data into two dimensions. Modifications to the algorithm allow refined user control and expert guidance.

Another recent advance has been the development of a grammar for creating a broad array of (static) graphics [53•]. The result is a flexible way to describe graphic creation. The primary implication is in the area of graphics application development using object-oriented design, but the grammatical definitions could also lead to new types of plots. In addition, the grammar naturally fits into the conceptual data pipeline approach to interactive and dynamic data visualization proposed by Buja *et al.* [11].

Proper construction of graphics to enhance human perception and cognition has involved much work in aesthetics and layout [3,27,28,54–57]. For many years poor practices have been promoted in the popular press and commercial software, such as adding a superfluous third dimension to a two-dimensional plot, but there has been recent improvement. The work of [3,28,55–57] has helped raise the profile of good graphical construction. A series of articles by Carr in the *Statistical Computing and Graphics Newsletter* [26•,27] similarly describes good graphical constructions. Many of these recommendations involve laying out multiple plots of data, and are enhanced by intelligent layout algorithms. For example, a hybrid parallel coordinates plot of gene expression clusters has been optimized to minimize line-crossing and enhance readability [29•], and a more informative reordering of variables in a color matrix display has been achieved by using a cluster algorithm [32••,33,58••]. In graph display, one of the main issues is to minimize the number of crossed connections: XGvis [12•] uses multidimensional scaling, and Nicheworks [7] describes a number of different layout methods, and methods for interacting with them. These approaches mainly arise from studying telephone networks but are applicable to many biotechnology applications also. Experiments in color perception (C Brewer, Abstract 639, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999; [54]) for geographical representations of data (see [59] for examples) has led to recommendations for more broad use of color in data representation. Another important theoretical issue for statistical graphics, one that has not received much attention, is the importance of an underlying model structure, as in regression graphics [60,61]. Statistical models can serve to filter out the influence of other vari-

ables when plotting the relationship between two variables of interest.

Sensitivity of the human visual system is both the promise and the peril of visualization. Humans may immediately detect features in data that would take much longer to find, or indeed never find, by modeling, but may also find spurious patterns in the same data. So it is with excitement that we see developments in graphical inference. New work has begun to more formally apply the idea of statistical permutations tests to graphics (A Buja, Abstract 568, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999; see also [32••,33]). The idea is to create many alternative versions of a graph from random permutations of the data. Patterns that stand out only in the original graph are considered real, based on the assumption that they occur rarely under appropriate null situations. The foundations of permutation tests, bootstrapping, cross-validation, and sampling may be valuable ways to provide inference graphically, which will facilitate inference in the field of data mining and knowledge discovery.

New application areas

Workers from the National Cancer Institute (NCI) describe their analysis of a large data set of >60,000 compounds screened against a panel of 60 human cancer cell lines between 1990 and 1997 [33]. Three databases encode available information: Database S (compounds versus 3D descriptors); Database A (compounds versus cell line activity); and Database T (cell line versus molecular targets). Analysis uses the NCI DISCOVERY visualization software. In the clustered correlation diagram, a color matrix depicts correlations between compounds and molecular targets across cell lines, with compounds and targets ordered by hierarchical cluster order. Patches of color identify groupings of compounds and targets with potentially related mechanisms of action, as shown in Figure 2. In many cases, correlations reflect known biochemical relationships among targets. It is emphasized that the aim of this analysis is exploratory, to be followed with controlled experiments. Nonetheless the approach holds promise for identifying new drug leads.

Microarray and other technologies now allow the simultaneous measurement of relative gene expression of thousands of RNA species in a single experiment. The joint behavior of large numbers of genes across experiments varying time, dose or other treatment is of particular interest, and visualization techniques will be important in studying these relationships [29•,30,32••,46••,58••]. In these early attempts, cluster analysis identifies groups of genes with related patterns of expression. Color matrices and line plots can be used to display large numbers of individual genes and experiments, grouped by their clusters. Other visualization ideas [29•,30] include a unique parallel coordinates display, color-coded by cluster (Figure 1) to show how clusters correspond to known gene families. The clever design of this plot minimizes messy line-crossing common in parallel coordinates [30]. We anticipate even more innovative approaches to gene expression arrays in the near future.

Internet resources and software

The kdnuggets website (http://kdnuggets.com) contains a huge variety of resources on visualization as part of data mining. Two popular statistical web resources are statlib (http://lib.stat.cmu.edu) and the UCLA web site (http://www.stat.ucla.edu).

Similar to the S language, there is a public domain package called R (http://fangorn.ci.tuwien.ac.at/R/), which provides excellent static graphics, as well as being a fully programmable language for data mining and statistical analysis. XlispStat [51] is a statistical analysis language that provides dynamic graphic capabilities and primitives, and there are considerable resources programmed with this package (see the UCLA website referenced above). A growing effort to provide a CORBA interface between all these packages (CORBA enables different software applications to communicate), including XGobi, is under way (D Temple-Lang, Abstract 682, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999). The Graphics Production Library (GPL), a suite of Java beans strongly based on Wilkinson's grammar of graphics [53•], is an upcoming visualization software that will cut across the boundaries of hardware (D Rope, Abstract 264, Joint Statistical Meetings, Baltimore, MD, 8–12 August 1999). Other Internet resources include Olive, which is a collection of links to graphical resources at the University of Maryland (http://otal.umd.edu/Olive/), and Michael Friendly's website, which has a picture gallery and collection of hyperlinks to graphics resources (http://www.math.yorku.ca/SCS/StatResource.html).

Conclusions

Advances have been made in the area of visualizing multidimensional data. New work is needed when the data contain variables of different types, such as a mixture of discrete and continuous, or contains non-matrix forms. Developing interaction paradigms for some of the new plot methods provides some interesting challenges. There are exciting opportunities for developing tighter coupling of analytical and visual tools, areas where CORBA protocols might help to better connect databases to legacy software and new packages. Scaling up the visualization algorithms for extremely large data sets is a growing concern. Developing approaches to providing inference and significance values with graphical displays will improve the ability of data miners to determine whether findings are real or spurious. We may also ride the coat-tails of the movie and game industry to take advantage of the technology innovations for new data analysis technologies. This may be particularly important for more closely connecting molecular models with data visualization, to better understand drug activity and interactions. To successfully

communicate results we need easier ways to publish graphics, and provide dynamic demonstrations within publications, but these are already available in newer forms of Internet publication. As implementations become available, new applications will be possible.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Card SK, MacKinlay JD, Shneiderman B (Eds): *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA: Morgan Kaufman; 1999.

A collection of papers that very broadly cover visualization applications: 1–3D space, multidimensional spaces, trees, networks, documents, the Internet, and interaction with graphics. They focus on different aspects but concentrate on software descriptions.

2. Tukey JW: *Exploratory Data Analysis*. Reading, MA: Addison-Wesley; 1977.

3. Cleveland WS: *Visualizing Data*. Summit, NJ: Hobart Press; 1993.

4. Buja A, Cook D, Swayne DF: **Interactive high-dimensional data visualization.** *J Comput Graph Stat* 1996, **5**:78-99.

5. Wills G: **Linked data views.** *Stat Comput Graphics Newsletter* 1999, **10**:20-24.

6. • Nelson L, Cook D, Cruz-Neira C: **XGobi vs the C2: results of an experiment comparing data visualization in a 3-D immersive virtual reality environment with a 2-D workstation display.** *Comput Stat* 1999, **14**:39-51.

The authors describe an experiment that compares the C2 virtual environment with XGobi for two common visual data analysis tasks: accurately identifying clusters and determining dimensionality of data sets. In addition, they use subtask completion times to compare the two systems' ease of use.

7. Wills GJ: **Nicheworks — interactive visualization of very large graphs.** *J Comput Graph Stat* 1999, **8**:190-212.

8. • Theus M, Lauer SRW: **Visualizing loglinear models.** *J Comput Graph Stat* 1999, **8**:396-412.

A very extensive, reader-friendly introduction to mosaic plots, and explanation of how these can be used to assist in modeling categorical data.

9. Swayne DF, Cook D, Buja A: **XGobi: interactive dynamic graphics in the X window system.** *J Comput Graph Stat* 1998, **7**:113-130. http://www.research.att.com/areas/stat/xgobi

10. Cook D, Buja A: **Manual controls for high-dimensional data projections.** *J Comput Graph Stat* 1997, **6**:464-480.

11. Buja A, Cook D, Asimov D, Hurley C: **Dynamic projections in high-dimensional visualization: theory and computational methods.** *AT&T Technical Report* 1998. http://www.research.att.com/~andreas/papers/dynamic-projections.ps.gz

12. • Buja A, Swayne DF, Littman M, Dean N: **XGvis: interactive data visualization with multidimensional scaling.** *J Comput Graph Stat* 1999, in press. (Preprint: http://www.research.att.com/areas/stat/xgobi/index.html#xgvis-paper)

An extension to XGobi which will do graph layout in high-dimensions based on multidimensional scaling (MDS) algorithm, and is also a flexible, interactive MDS system.

13. • Theus M: **Analysing storm data using highly interactive tools.** *Comput Stat* 1999, **14**:91-108.

Theus combines the use of interactive graphics with standard parametric analysis in his study of hurricane behavior. He uses REGARD to handle the spatial data, DataDesk for modeling, and MANET for additional graphics, and reports that no single system would have been adequate for the analysis.

14. • James DA: **Interactive data analysis in a manufacturing setting — a case study.** *Comput Stat* 1999, **14**:147-159.

A case study of interactive graphical methods in the manufacture of semiconductors and optical fibers. The work presented was developed in S.

15. Cox KC, Eick SG, Wills GJ, Brachman RJ: **Visual data mining: recognizing telephone calling fraud.** *Knowledge Discov Data Mining* 1997, **1**:225-231.

16. Gershon N, Eick SG: **Guest editors' introduction: Information visualization. The next frontier.** *J Intelligent Inf Syst* 1998, **11**:199-204.

17. • Cook D, Cruz-Neira C, Kohlmeyer BD, Lechner U, Lewin N, Nelson L, Olsen A, Pierson S, Symanzik J: **Exploring environmental data in a highly immersive virtual reality environment.** *Environ Monit Assess* 1998, **51**:441-450.

The authors discuss linking a multidimensional display, based on tour methods, to a geographic display in an immersive 3D environment.

18. Cleveland WS: *The Elements of Graphing Data (Revised Edition).* Summit, NJ: Hobart Press; 1994.

19. Becker RA, Chambers JM, Wilks AS*: A Language and System for Data Analysis*. Murray Hill, NJ: Bell Laboratories Computer Information Service; 1981.

20. Becker RA, Cleveland WS: *Trellis Graphics User's Manual*. Seattle, WA: MathSoft, Inc.; 1996.

21. • Inselberg A: **Don't panic . . . just do it in parallel.** *Comput Stat* 1999, **14**:53-77.

This paper gives an overview of parallel coordinates as a method of data visualization, and shows some of what can be learned about the geometry of the data from the parallel coordinates display.

22. • Avidan T, Avidan S: **ParallAX — a data mining tool based on parallel coordinates.** *Comput Stat* 1999, **14**:79-89.

The authors describe an interactive system based on Inselberg's work [21•]. In positioning their system as a tool for data mining, they have included some automated procedures, such as one for trimming extreme points.

23. • Wilhelm AFX, Wegman EJ, Symanzik J: **Visual clustering and classification: the Oronsay particle size data set revisited.** *Comput Stat* 1999, **14**:109-146.

The authors compare the use of MANET, XploReN and XGobi to analyze a fairly complex set of data. With XGobi and XploReN they demonstrate what they call the 'brush-tour strategy', alternating between using the grand tour to visually identify subgroups in the data and brushing to mark these groups. The strategy described in the section on MANET is quite different, and depends on linked low-dimensional views and very high user interaction.

24. Keim DA, Kriegel H-P: **VisDB Database exploration using multidimensional visualization.** *IEEE Comput Graphics Applications* 1994, **14**:40-49. (Also appears in [1•].)

25. • Healey CG, Enns JT: **Large datasets at a glance: combining textures and colors in scientific visualization.** *IEEE Trans Vis Comput Graphics* 1999, **5**:145-167.

This research grows out of icon displays and advances it by making a translation table for representing numerical values as color, shape, and dispersion over space, creating a texture for spatially referenced-multivariate data. Limitations are that only three variables can be represented.

26. • Carr DB, Olsen AR, Courbois J-YP, Pierson SM, Carr DA: **Linked micromap plots: named and described.** *Stat Comput Graphics Newsletter* 1998, **9**:24-32.

One element of a series of papers on various visualization ideas pioneered by Dan Carr. Many are related to geographically referenced data, some are on other applications, such as gene expression data, and others are on graphical paradigms, and perception for visualizing data.

27. Carr DB, Sun R: **Using layering and perceptual grouping in statistical graphics.** *Stat Comput Graphics Newsletter* 1999, **10**:25-31.

28. Tufte ER: *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press; 1983.

29. • Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R: **Large-scale temporal gene expression mapping of central nervous system development.** *Proc Natl Acad Sci USA* 1998, **95**:334-339.

Microarray and other technologies now allow the simultaneous measurement of relative gene expression of hundreds to thousands of RNA species in a single experiment. The joint behavior of large numbers of genes across experiments varying time, dose or other treatment is of particular interest. This paper describes a longitudinal study in embyonic and postnatal rat spinal cord tissue. Color maps are used to depict patterns of gene expression and groups of genes that move together. Three-dimensional stereo plots further elucidate the clusters. Visualization of expression profiles confirmed four distinct phases of spinal cord development corresponding to the four major clusters.

30. Carr DB, Somogyi R, Michaels G: **Templates for looking at gene expression clustering.** *Stat Comput Graphics Newsletter* 1997, **8**:20-29.

31. Chen C (Ed): *Information Visualization and Virtual Environments.* Heidelberg, Germany: Springer-Verlag; 1999.

32. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
•• The authors use cluster ordering and color matrices to display results of microarray gene expression experiments.

33. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Koh KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL *et al.*: **An information-intensive approach to the molecular pharmacology of cancer.** *Science* 1997, **275**:343-349.

34. Minotte MC, West RW: **The data image: a tool for exploring high dimensional data sets.** In *Proceedings of the Section on Statistical Graphics. Alexandria, VA: American Statistical Association;* 1998:25-33.

35. Mihalisin T, Timlin J, Schwegler J: **Visualizing multivariate functions, data, and distributions.** *IEEE Comput Graphics Applications* 1991, **11**:28-35. (Also appears in [1•].)

36. Huang C, McDonald JA, Stuetzle W: **Variable resolution bivariate plots.** *J Comput Graph Stat* 1997, **6**:383-396.

37. Hartigan JA, Kleiner B: **A mosaic of television ratings.** *The American Statistician* 1984, **38**:32-35.

38. Emerson JW: **Mosaic displays in S-Plus: a general implementation and a case study.** *Stat Comput Graphics Newsletter* 1998, **9**:17-23.
• A nice introduction to mosaic plots and reference to a web site with S-Plus code for generating these displays.

39. Friendly M: **Extending mosaic displays: marginal, conditional and partial displays of categorical data.** *J Comput Graph Stat* 1999, **8**:373-395.
• An introduction to mosaic plots, explanation of their use in modeling, and an introduction of matrix displays of mosaic plots.

40. O'Connell MA, Wolfinger RD: **Spatial regression models, response surfaces and process optimization.** *J Comput Graph Stat* 1997, **6**:224-241.

41. Faraway JT: **A graphical method of exploring the mean structure in longitudinal data analysis.** *J Comput Graph Stat* 1999, **8**:60-68.

42. Eno DR, Terrell GR: **Scatterplots for logistic regression.** *J Comput Graph Stat* 1999, **8**:413-430.

43. Swayne D, Buja A: **Missing data in interactive high-dimensional data visualization.** *Comput Stat* 1998, **13**:15-26.

44. Osbourn GC, Martinez RF: **Empirically defined regions of influence for clustering analyses.** *Pattern Recognition* 1995, **28**:1793-1806.

45. Kohonen T: *Self-Organizing Maps*. Heidelberg, Germany: Springer-Verlag; 1995.

46. Tamayo P, Slonim D, Mesirov J, Zhu Q, Itareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96**:2907-2912.
•• Self-organizing maps (SOMs) are used to cluster genes based on results of microarray gene expression experiments. SOMs provide an alternative to dendrogram ordering (see [57]) for organizing the raw data for display, especially for temporally-ordered experiments.

47. Van Teylingen R, Ribarsky W, Van Der Mast C: **Virtual data visualizer.** *IEEE Trans Vis Comput Graphics* 1997, **3**:65-74.

48. Carr D, Wegman EJ, Luo Q: **ExplorN: design considerations past and present.** In *Technical Report 29*. Fairfax, VA: Center for Computational Statistics, George Mason University; 1996.

49. Unwin A: **Requirements for interactive graphics software for exploratory data analysis.** *Comput Stat* 1999, **14**:7-22.

50. Wills G: **Selection: 524, 288 ways to say this is interesting.** *J Comput Graphical Statistics* 1999, in press.

51. Tierney L: *LispStat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. New York, NY: Wiley; 1991.

52. Peterson C: **Andrews curves revisited: for large data and low-end technology [MS Thesis].** Department of Statistics, Iowa State University, 1999.

53. Wilkinson L: *The Grammar of Graphics* (*Statistics and Computing*). Heidelberg, Germany: Springer-Verlag; 1999.
• A detailed book on a syntax for creating a broad array of graphics, primarily static, from the most simple to the most complex. It unifies different graphics with specific rules for describing their creation. The result is a flexible way to describe graph creation.

54. Brewer CA: **Spectral schemes: controversial color use on maps.** *Cartogr Geogr Inf Sys* 1997, **24**:203-220.

55. Wainer H: *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot.* New York, NY: Copernicus Books; 1997.

56. Tufte ER: *Envisioning Information.* Cheshire, CT: Graphics Press; 1990.

57. Tufte ER: *Visual Explanations: Images and Quantities, Evidence and Narrative.* Cheshire, CT: Graphics Press; 1997.