Picture this

Drowning in data? New visualization techniques could help. Philip Ball discovers, among other things, how to plot a seven-dimensional graph.

Ralph Kahn's office at NASA's Jet Propulsion Laboratory in Pasadena, California, has a unique line in interior decor. Brightly coloured charts cover the walls. Red is a favourite theme, streaked through with yellow and blue. The garish creations won't win any design awards, but Kahn hopes that by staring at them for long enough, he could come up with new ideas about the Earth's climate.

Kahn has created his unusual wallpaper because, like many other scientists, he is swamped with data. The figures illustrate measurements taken by a sensor on board the Terra satellite, the flagship of NASA's Earth-observation programme. The device produces data at a phenomenal rate — its nine cameras simultaneously monitor four frequencies of sunlight reflected by the Earth. "To understand a climate phenomenon involving multiple feedbacks, you really need to look almost everywhere, in detail, and often," explains Kahn.

By viewing plots of the Terra data, Kahn hopes to spot hints of patterns that are worthy of analysis. Yet the possibilities for plotting data in revealing ways run far beyond Kahn's two-dimensional coloured graphs. Few people are aware of it, but sophisticated methods for visualizing large data sets are within the reach of most researchers. Thanks to increases in computing power and improvements in graphical software, scientists can now make more and more use of the best pattern detector that we have at our disposal — the human visual system.

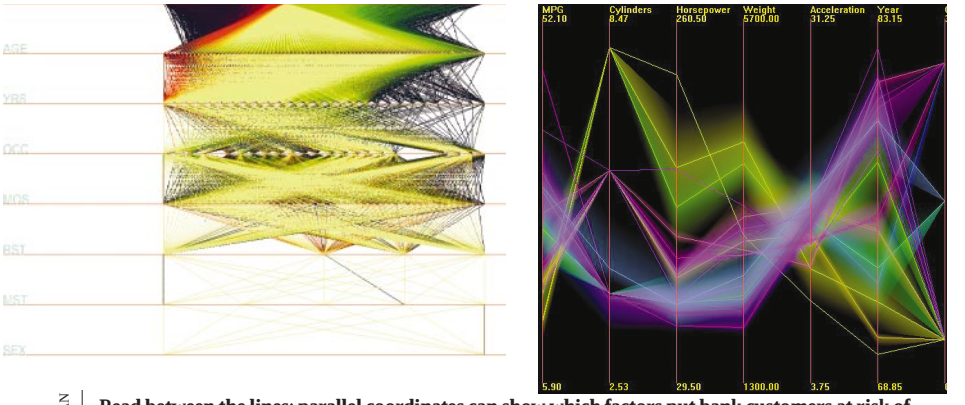
W. MASRI

Paper chase: Ralph Kahn hopes to spot patterns in his data by literally surrounding himself with it.

Digital deluge

Researchers studying everything from gene expression to financial risk assessment stand to benefit. Many disciplines routinely produce more data than researchers know what to do with. In 1999, for example, an advisory committee to the US National Institutes of Health estimated that some biomedical laboratories can produce up to 100 terabytes (10^{14} bytes) of information a year — enough to fill a million encyclopaedias. Digital technology is the main culprit. "A $2,048 \times 2,048$-pixel colour image contains 16 megabytes of data, even if it's a picture of my kitchen floor," points out Bill Eddy, a statistician at Carnegie Mellon University in Pittsburgh, Pennsylvania.

It's not just a question of having too much data — the kind of data also matters. The relationships between two variables can be examined using a simple graph. But experiments in many fields generate data using different instruments that simultaneously measure several parameters, which may or may not be related. The number of variables that could be paired in a two-dimensional plot is enormous — a problem known as



E. WEGMAN

Read between the lines: parallel coordinates can show which factors put bank customers at risk of slipping into the red (left), and the variables that are linked to efficient performance in cars (right).

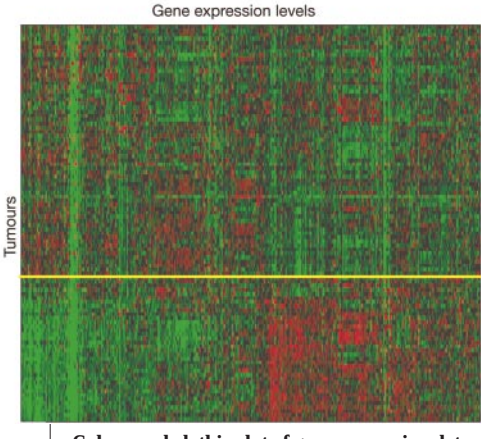
the 'curse of dimensionality'.

It is here that visual data-mining could help. Before pinning down precise mathematical relationships within a data set, visual techniques provide a first line of attack that can suggest what kind of trends and patterns may lie within the numbers, and may thus help guide the focus of more detailed analysis. Such an approach, known as exploratory data analysis, has long been used by statisticians, but sophisticated visualization techniques are now playing an increasingly important part in it.

DNA microarray experiments, which can simultaneously monitor the expression of thousands of genes, are already benefiting from visual analysis. David Botstein, a geneticist at Stanford University who helped to pioneer microarray studies, began working on the problem in 1998. Together with colleagues, he described how data on the activity of thousands of genes can be displayed in a way that makes it clear which genes have similar patterns of expression[1].

In January, a team from the Netherlands Cancer Institute in Plesmanlaan and Rosetta Inpharmatics, a drug-development and technology company in Kirkland, Washington, used a similar plotting method in their work on predicting the course of breast cancers[2]. The team looked at tissue samples from the tumours of almost 100 patients, some of whom had gone on to develop secondary cancers in the five years after the samples were taken.

REF. 2



Colour-coded: this plot of gene-expression data shows breast tumours falling into two groups.

The researchers measured expression levels for almost 5,000 genes in each tumour, and used a mathematical algorithm to create a list of tumours in which genes with similar patterns of expression were grouped together. Plotting the list using colours to represent expression levels revealed that the tumours seemed to fall into two different groups (divided by the yellow line in the plot; see below left). Further investigation helped the team to identify a subset of 70 genes that could be used to predict whether a patient will develop secondary tumours.

Botstein feels that other important findings lie hidden in microarray data. "There is more information in the data than we as a community have yet extracted," he says. "I'm hopeful that better things may emerge. This is the very beginning."

Working in parallel

Other visualization techniques have longer histories, but are now benefiting from improvements in computing technology. One example, the method of parallel coordinates, was first proposed during the 1980s, and is particularly useful for tackling high-dimensional data.

In the demonstration example shown in the plot above (right), the method is used to represent performance data for cars. Each vehicle is described by seven variables, such as weight and acceleration. Parallel axes are used to represent each variable, and the data for each vehicle form a line that links the values on each axes. Trends can then be extracted by looking for sets of lines that cluster together. In the example given here, representing cars with high fuel efficiency in purple shows that these vehicles tend to have low weights and to be more recent models.

More complex data can be mined by combining parallel coordinates with other visualization techniques. Edward Wegman, a statistician at George Mason University in Fairfax, Virginia, has used these methods to identify correlations in data on bank customers in order to find out whether particular combinations of variables can be used to identify individuals who are at risk of financial problems[3].

M. WARD

Each customer is classified by eight variables, such as age, occupation and profit status with the bank. Plotting all of the records — over 130,000 in total — produces a solid mass of lines that is difficult to interpret, so Wegman used a technique known as saturation brushing to reveal hidden patterns. Each line was given a hint of colour: red for those with overdrafts, green for those in credit. Individual lines that were brushed in this way still appeared black, but the colours added up when lines overlapped. If, for example, a group of overdrawn customers clustered around a point on one of the axes, the mass of lines appeared red. Roughly even numbers of red and green lines mixed around points that did not correlate with credit status, and these points appeared yellow.

The colour of money

Colour brushing revealed, unsurprisingly, that younger customers are more likely to be in debt (see plot, above left). But Wegman suspected that more interesting correlations might be hidden in the data, and so he used a further method of analysis, known as a grand tour. This amounts to taking a virtual journey through the multidimensional data-space so that the points — or lines in parallel coordinates — are seen from many different perspectives.

The process is easy to visualize in three dimensions. Imagine a cloud of points, shaped like a doughnut, on a three-dimensional graph. If the graph is viewed through the plane of the doughnut, the hole is hidden. It is only when the data are seen from other angles that the overall structure becomes clear.

A mathematical version of this process can be applied to data of any number of dimensions. In the case of parallel coordinates, the lines are viewed from a new position by replotting the data using axes that represent combinations of the original variables, rather than assigning a single variable to each axis.

As the grand tour of the bank data progressed, interesting features emerged. Wegman clarified these features by removing nearby points that did not appear to be part of potential trends, a process known as data-pruning. After several successive plots, the data revealed further risks associated with certain occupations, whether the customer owns or rents their home, and their marital status.

Because each axis now represents a combination of variables, risks cannot easily

B. SHERMAN

Business or pleasure? Paul Rajlich uses the CAVE to interact with data and, in his spare time, to tackle some virtual baddies (inset).

be related to the original variables, such as occupation and so on. But the risk associated with a new customer can still be calculated by combining that person's data in the appropriate way. And by not singling out particular factors, the bank can avoid accusations of discrimination.

Other visualization researchers stress the importance of interacting with data. Jim Thomas, an expert in data-mining technology at Pacific Northwest National Laboratory in Richland, Washington, believes that the key to success lies in the dynamic aspect of modern computer graphics. "Interaction is as important as the visuals," he says. And if users want to get really interactive, they can no better than to enter a device known as the CAVE.

Developed in the early 1990s in the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago, the CAVE is a cube-shaped room, around three metres square, with one open wall through which three-dimensional computer images are projected onto the other walls and floor.

Inside the CAVE, users can 'float' in data-space. Sensors track head movements, and the on-screen graphics are adjusted so that the image changes just as it would if the user were moving around a real scene. A manual controller called the wand acts like a kind of three-dimensional mouse, allowing users to probe and modify the images.

"The CAVE allows you to understand very complicated structures because you can intuitively and easily manipulate your view," says Paul Rajlich, a CAVE researcher at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign. "You can simply walk around the CAVE and examine the structure just as if it were a real object."

There are now about a hundred CAVEs in use around the world. Most are used in virtual-reality research, automotive design and medical training. But Earth scientists often use them to visualize geological structures, and biologists are also getting in on the act. Timothy Karr studies fertilization in fruitflies at the University of Chicago. He has used a software system known as Crumbs, developed at the NCSA and other departments at Urbana-Champaign, which allows users to explore paths through three-dimensional data-space by dropping 'crumbs' to mark their trails.

Gathering the crumbs

Karr is interested in the fate of fruitfly sperm after it enters the egg. He used a CAVE at the NCSA to display microscopic images of sperm tails inside the egg. By using Crumbs to outline the tail on successive images, he was able to plot how the sperm changes once it is within the egg[4]. "Think about trying to follow an unravelling ball of yarn inside an egg, and you have an idea of how difficult this would be without software like Crumbs," says Karr.

CAVEs, and other less sophisticated methods, could help researchers get a handle on the vast, high-dimensional data sets that are becoming increasingly common. But the techniques are not without their problems. Designing visualization methods is a multidisciplinary task, drawing in statisticians, computer programmers and the researchers who produce the data, yet these different groups do not always communicate effectively. And even when techniques are well-designed, some researchers are wary of placing too much confidence in them.

Many of the potential problems stem from a lack of cooperation, an issue that Botstein has experienced in his microarray work. "Biologists know too little about mathematics and computation, and computer scientists and statisticians know too little biology," he says. "And as in any interdisciplinary endeavour, there are cultural and value differences that cause misunderstanding." Biologists, says Botstein, do not particularly care how they analyse their data, as long as it offers some insight. But mathematicians may feel that a 'trivial' analysis has no academic value, despite the fact that such methods often stretch biologists' analytical powers to the limit.

There is also a divide between researchers who work on scientific images and those that use visualization as a precursor to statistical analysis. "The two groups tend to have their own separate conferences," says Jason Leigh of the EVL. "There are so many techniques from both camps that are mutually beneficial to one another, but the opportunities for cross-fertilization are lost."

Others lament the fact that vision researchers are seldom involved in developing the techniques, and warn that statisticians could be placing too much confidence in the pattern-recognition abilities of the human brain. "The human visual system is remarkably good at discovering real patterns," says Eddy. "But it is also remarkably good at discovering false patterns. The experts who work on visualization methods are, with a few notable exceptions, woefully ignorant of the relevant psychology and psychophysics. Sadly, the cognitive psychologists have bigger fish to fry and are not especially interested."

If the groups do start working together, they are likely to find that the speed with which data become redundant creates a demand for new methods of analysis. "The time window for analysis is getting shorter and shorter," says Daniel Carr, a statistician at George Mason University. "It is hard to imagine that in five years scientists will be much interested in analysing today's microarray data."

They will also find that with data piling up in many different disciplines, and becoming less useful increasingly quickly, there has never been a greater need for visualization methods. And for the advocates of visualization, the benefits are clear. "Science will continue to progress with or without more attention to visualization and data analysis," says Eddy. "But it will progress much faster with careful attention to these issues. To begin with there has to be a will to do it — a commitment to the analysis." ■

Philip Ball is a consultant editor of *Nature*.

1. Eisen, M B., Spellman, P. T., Brown, P. O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
2. van 't Veer, L. J. *et al. Nature* **415,** 530–536 (2002).
3. Wegman, E. *Stat. Med.* (in the press).
4. Karr, T. L. & Brady, R. *Trends Genet.* **16,** 231–232 (2000).