*Editors: Jim X. Chen, jchen@cs.gmu.edu*
*R. Bowen Loftin, bloftin@odu.edu*

# A TELESCOPE FOR HIGH-DIMENSIONAL DATA

***By Ben Shneiderman***

MUSCULAR DYSTROPHY IS A DEGENERATIVE DISEASE THAT DESTROYS MUSCLES AND ULTIMATELY KILLS ITS VICTIMS. RESEARCHERS WORLDWIDE ARE RACING TO FIND A CURE BY TRYING TO UNCOVER THE GENETIC PROCESSES THAT cause it. Given that a key process is muscle development, researchers at a consortium of 10 institutions are studying 1,000 men and women, ages 18 to 40 years, to see how their muscles enlarge with exercise. The 150 variables collected for each participant will make this data analysis task challenging for users of traditional statistical software tools.

However, a new approach to visual data analysis is helping these researchers speed up their work. At the University of Maryland's Human–Computer Interaction Library, we developed an interactive approach to let researchers explore high-dimensional data in an orderly manner, focusing on specific features one at a time. The *rank-by-feature framework* lets them adjust controls to specify what they're looking for, and then, with only a glance, they can spot strong relationships among variables, find tight data clusters, or identify unexpected gaps. Sometimes surprising outliers invite further study as to whether they represent errors or an unusual outcome.

Similar data analysis problems come up in meteorology, finance, chemistry, and other sciences in which complex relationships among many variables govern outcomes. The rank-by-feature framework could be helpful to many researchers, engineers, and managers because they can then steer their analyses toward the action.

## Understanding the Rank-by-Feature Framework

We can use a playful analogy to help explain this new way of thinking about data. Imagine you parachute into an unknown landscape. You might look around to find something interesting, or maybe you climb a nearby hillside to gain an overview. You notice a large rock and some sandy soil, admire a bright green fern, smell a rose bush, and spot an oak tree. You might get annoyed by a swarm of bees, recognize a red-tailed hawk flying above, and spot deer tracks. You might also notice the sloping terrain, remember jagged cliffs, and become aware of the noisy stream in the valley below. In short, there's a lot to look at, just as there is in a high-dimensional database.

Recording "interesting features" is useful, but if you have specific goals, such as cataloging the area's plants to identify potential pharmaceuticals, you must be systematic and thorough. If you're determining the area's suitability for rural development or trying to protect endangered species, an orderly and comprehensive approach is necessary. Professionals in botany, geology, civil engineering, and urban planning have developed step-by-step approaches to ensure thorough and standardized evaluations that are consistent across time and evaluators.

Such orderly approaches and tools are being developed to support high-dimensional data as well. The prolific statistician John Tukey proposed an approach he called *scagnostics* in 1985, by which he meant scatterplot diagnostics.[1] He believed that scatterplots with two of the many dimensions along the *x*- and *y*-axes were a comprehensible way to look at data, but he knew that there were often too many such 2D projections to examine in large data sets. He proposed a few criteria for ranking scatterplots, but no one has implemented his ideas until now. The rank-by-feature framework implements Tukey's vision in an open-ended manner to allow easy addition of new criteria.

At the University of Maryland, doctoral student Jinwook Seo added the rank-by-feature framework to software he developed for his dissertation research on genomic data analysis. That software tool, the Hierarchical Clustering Explorer (HCE),[2] is available for download at www.cs.umd.edu/hcil/hce (see the sidebar on p. 52). HCE handles clustering of large data sets to help biologists find similar genes that might participate in processes related to muscular dystrophy, cancer, cell death, and so on. The

 Copublished by the IEEE CS and the AIP 1521-9615/06/$20.00 © 2006 IEEE 

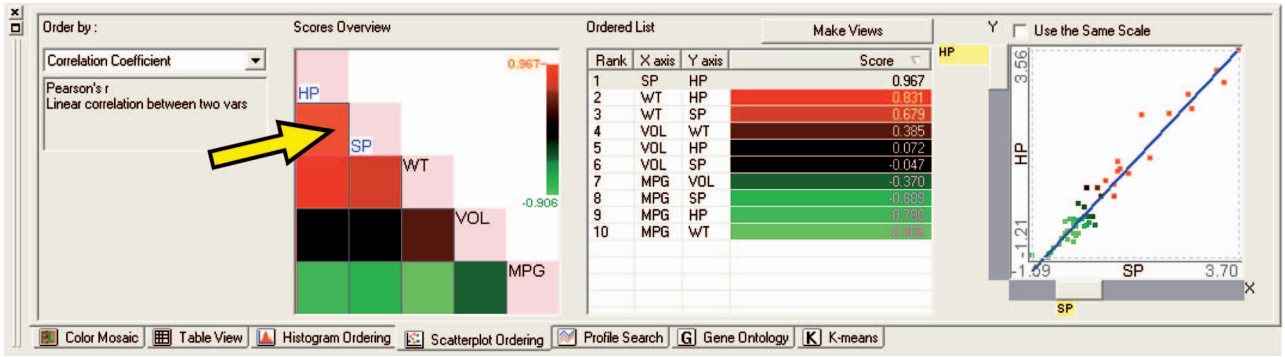


**Figure 1. The automobile data set. The order-by criteria (left panel) is the correlation coefficient, producing a color-coded scores overview with five variables: horsepower (HP), maximum speed (SP), weight (WT), volume of cab space (VOL), and miles per gallon (MPG). The bright red coding shows the highest correlation, and bright green indicates the lowest. The ordered list shows all 10 pairs of variables in ranked order from bright red down to bright green. The user has selected the highest correlation, which is speed vs. horsepower. The scatterplot in the far right panel shows that to increase speed, users must ensure higher horsepower.**

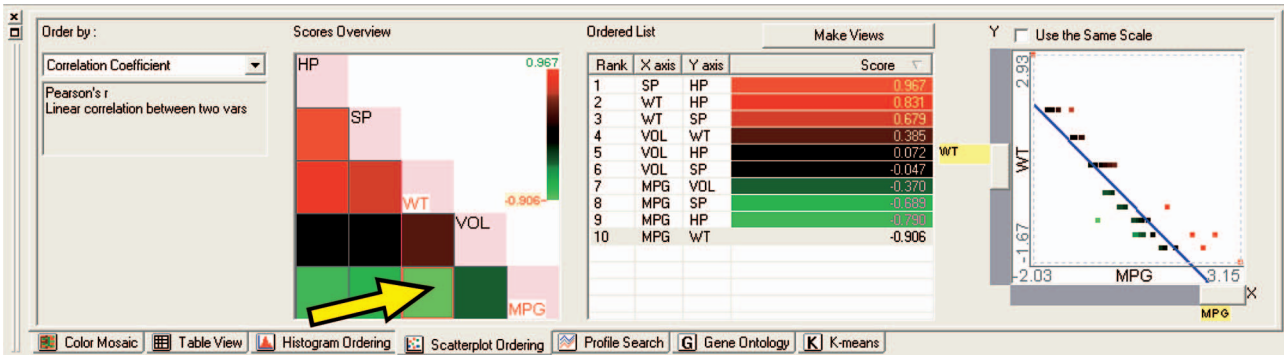


**Figure 2. The automobile data set, ordered by the strongest negative correlation. In this case, it's weight (WT) vs. miles per gallon (MPG).**

addition of the rank-by-feature framework dramatically speeds up exploration of high-dimensional data.

## A Simple Data Set

We can use a simple data set about automobiles to help describe the rank-by-feature framework. Figure 1 shows five variables: horsepower (HP), maximum speed (SP), weight (WT), volume of cab space (VOL), and miles per gallon (MPG). The bright red coding shows the highest correlation, and bright green indicates the lowest. The ordered list shows all 10 pairs of variables in ranked order. The user selects the highest correlation, which is speed versus horsepower, showing that to increase speed, users must ensure higher horsepower. Figure 2 shows that it is equally easy to find the strongest negative correlation, which is weight versus miles per gallon, meaning that as weight goes up, mileage goes down.

We conducted an email survey among known HCE users, which produced 57 completed responses. The correlation coefficient was the most desired feature, but these users also used the ranking criteria to find strong outliers, tight clusters, uniform distributions, and large gaps. Figure 3 shows that the strongest quadratic relationship in this simple data set was MPG and HP. In one of our eight-week case studies, a meteorologist found quadracity to be central to revealing an unknown pattern of how aerosols (small particles) affect absorption and reflection of infrared and ultraviolet light, potentially leading to improved weather prediction models.

Outliers—data points that differ strongly from most others—can reveal unusual items that either deserve more attention or uncover mistaken values in large data sets. Figure 4 reveals six strong outliers that deserve further study—first, to verify that the data are correct and, second, to understand how the automobile designers achieved high interior volume while keeping the weight low.

Some rankings are more helpful in one-dimensional (1D) histogram, whereas others are well suited for 2D scatterplots. The two separate but consistently designed interfaces let users

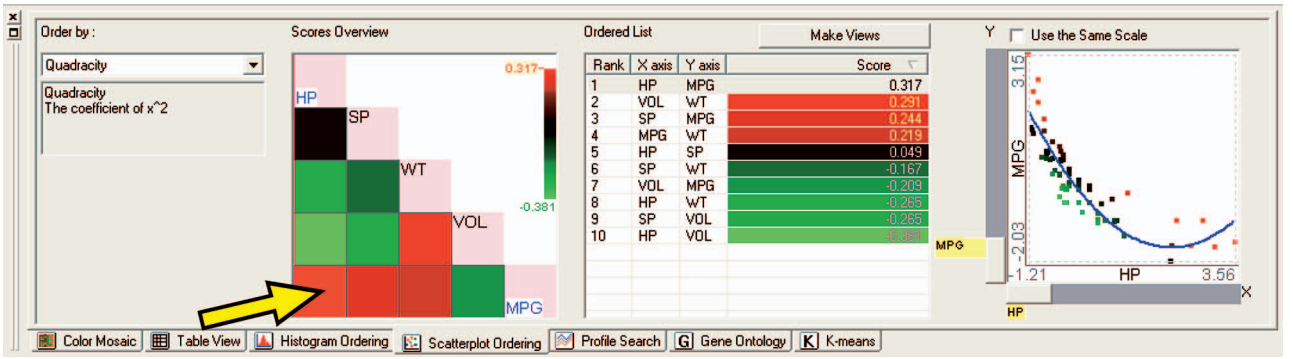


**Figure 3. The automobile data set, ordered by quadracity. The scores overview shows the highest quadratic relationship for horsepower (HP) vs. miles per gallon (MPG), which users select to produce the scatterplot on the right. In this case, MPG drops sharply as HP increases.**

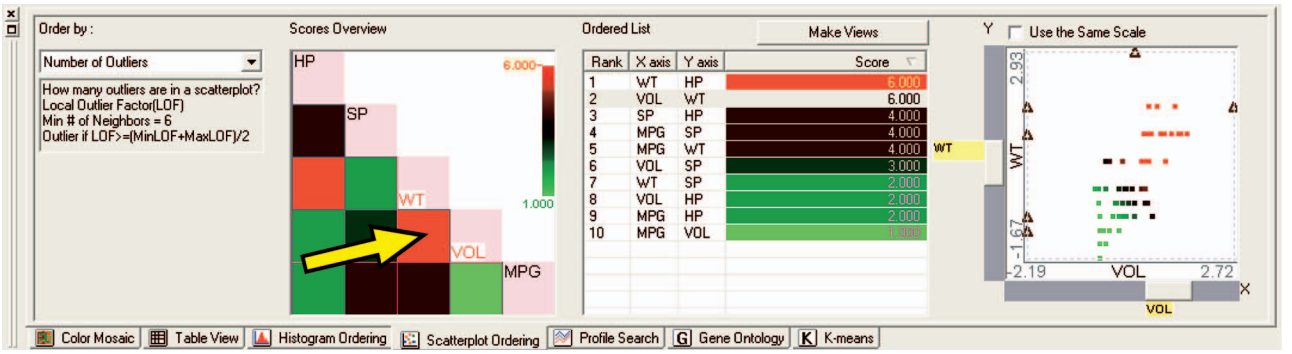


**Figure 4. The automobile data set, ordered by number of outliers. The scores overview shows large numbers of outliers for weight (WT) vs. volume (VOL) and for horsepower (HP) vs. weight (WT). The scatterplot on the right show weight vs. volume, with the six outliers highlighted by small triangles.**

explore 1D features first, and then turn their attention to the more complex 2D features. In each case, the rank-by-feature framework provides a visual overview, with red for high and green for low values, but users can set colors as they see fit.

For many users, finding the 1D distributions that were close to normal (bell-shaped curve) or furthest from normal was important. Figure 5 shows how easily users can discover that the maximum speed of cars was closest to being normally distributed.

Finding gaps is also a common task for data explorers. Figure 6 shows how to find the biggest gaps for this data, which occurred in the volume variable.

This simple data set demonstrates the basic ideas, but the rank-by-feature framework's substantial value becomes apparent with much larger data sets, such as a muscle data set with 150 variables.

## Muscle Exercise Data

Dealing with large data sets is more difficult, and the time needed for analysis much longer. It might take hours or days to check out the interesting features, but using the rank-by-feature framework can give users a method and tool to guide them and ensure that their analysis has been comprehensive and systematic.

The 150 variables in the muscle exercise study noted earlier include age, gender, muscle strength before and after, fat levels, and much more. These variables produce more than 22,000 possible relationship pairs, which would take weeks to analyze using traditional methods. In this case, researchers set the controls on the rank-by-feature framework to produce a triangular matrix (see Figure 7) in which a few hundred bright red squares show strong positive relationships.[3] Only a few dozen blue squares show strong negative relationships (when one variable increases, the other decreases). Thus, with one compact overview, researchers can narrow their focus and explore the strongest and most surprising relationships, either positive or negative.

Many relationships might be obvious, like the strong positive correlation between height before and after exercise. Surprises are immediately visible, however, such as participants who differed in height substantially before and after—clearly a result of measurement

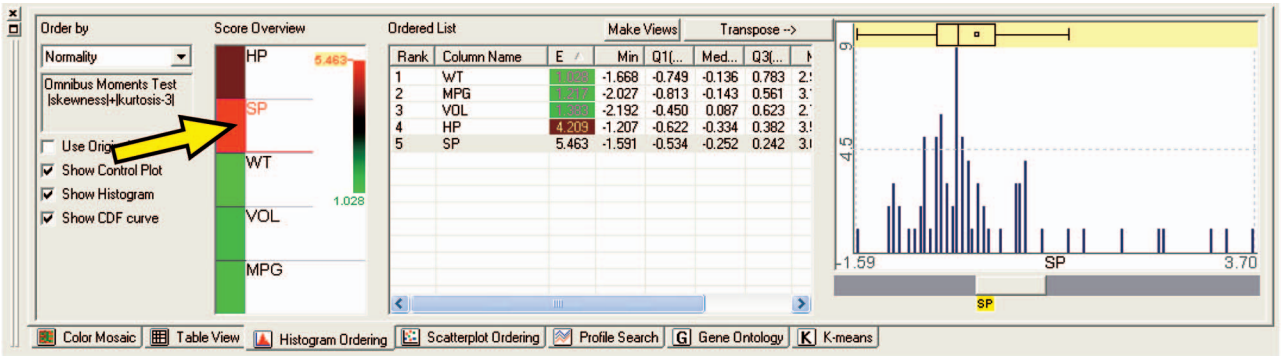


**Figure 5. The automobile data set, ordered by normality. Shifting to the histogram ordering tab, the score overview shows that the variable whose distribution is closest to being normal is speed (SP). The histogram on the right shows an almost symmetric normal distribution.**

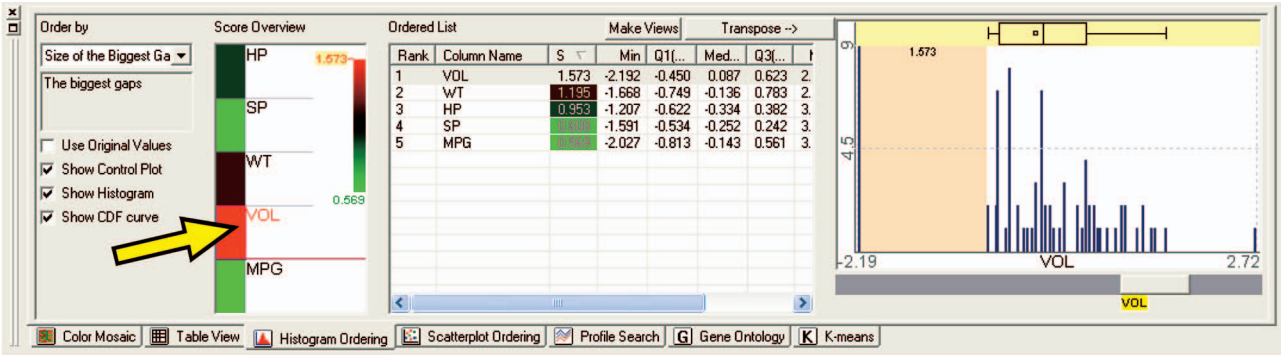


**Figure 6. The automobile data set, ordered by size of biggest gap. The score overview shows that the biggest gap occurs in the volume variable. The histogram on the right shows the biggest gap in a peach-colored rectangle. Several cars have extremely small volumes of interior seating space, but this analysis reveals that an error exists in the data set.**

errors. An interesting negative relationship emerged in this study: those who started with weaker arms improved the most, whereas the stronger a person's arms were initially, the less they gained from exercise. The most significant finding was an unknown strong association between the AKT1 gene and male body composition that affects strength, bone size, and subcutaneous fat, while protecting the metabolic syndrome.

Relationships between variables are only one feature of 2D scatterplots. Users of the rank-by-feature framework can set the controls to look for features such as clusters, gaps, and outliers.

## Information Visualization: New Telescopes for Data

Just as the telescope let Galileo see Jupiter's moons, new visual analytic software tools are enabling discoveries in genomics, pollution control, and terror-threat detection. The recent report, *Illuminating the Path*, celebrated the importance of visual analytic approaches[4] and called for substantial increases in research, especially for homeland security applications.

The emergence of information visualization user interfaces that combine data mining and statistical methods is bringing novel methods and potent tools to many analysts' desktops. Some improvements result from faster processors, larger displays, and improved databases, but the major advances are coming from innovative user interface features and novel filtering strategies. Designers of research prototypes and successful commercial products are finding imaginative ways to apply the widely stated *information-seeking mantra*: overview first, zoom and filter next, and then details-on-demand.

This simple strategy suggests that users should begin by seeing an overview of all their data, even if this visual representation must aggregate and summarize millions or billions of items. This lets them see the range and distribution of data items, immediately discover surprising gaps, and identify missing values or incorrect data items. Then they can zoom in on what they want and filter out what they don't, using animation and user-controlled sliders. Finally, when they've focused their attention on a few items, users can click to get details-on-demand, send the result set via email, or save a

## THE HEIRARCHICAL CLUSTERING EXPLORER

You can access the Hierarchical Clustering Explorer (HCE 3.0), free technical papers, and an extensive user manual for educational and research use at www.cs.umd.edu/hcil/hce/. HCE provides powerful features for multidimensional data analysis, including interactive exploration of hierarchical clustering results (with dendrograms and color-coded mosaics), and supports the rank-by-feature framework. During an eight-week evaluation period, researchers we observed found the framework useful in looking at all variables' distributions and test normality quickly and simultaneously, as well as for finding outliers and unique values.[1] HCE's additional capabilities include parallel coordinates and tabular data viewers. HCE 3.0 is a standalone Microsoft Windows application written in C++ that runs on a general PC environment. We have successfully installed it with Windows emulators on Macintosh computers.

**Reference**

1. J. Seo and B. Shneiderman, "Knowledge Discovery in High-Dimensional Data," *IEEE Trans. Visualization and Computer Graphics*, vol. 12, no. 3, 2006.

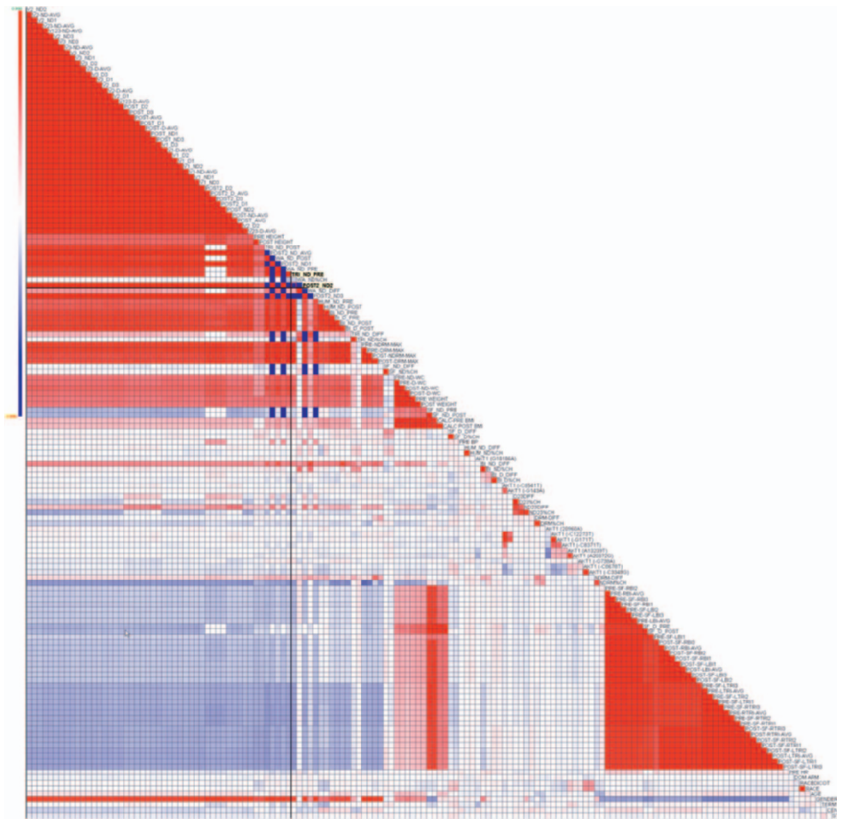

**Figure 7. Correlation coefficient ranking. Missing values are excluded for the 150 variables in a study of muscle size and strength. Red squares indicate high positive correlations, whereas blue squares indicate negative correlations.**

presentation to a Web page to invite colleagues' comments.

These advances, made during the past decade, are available through research prototypes and a growing number of commercial tools. Multidimensional information visualization products include Spotfire (www.spotfire.com), Datadesk (www.datadesk.com), and SAS/GRAPH (www.sas.com). Research tools are also maturing rapidly, with toolkits such as GeoVista from Penn State University and Polaris from Stanford University.

Although multidimensional data are widely used, many scientific, business, demographic, and other data sets are hierarchically organized. For these data sets, tree-structured visualizations such as treemaps are rapidly gaining acceptance. They are prominent in Web-based applications such as Smartmoney's Marketmap (www.smartmoney.com/marketmap/), which shows 600 stocks color-coded to highlight rising (green) or falling (red) prices, with area denoting market capitalization—that is, bigger companies have larger rectangles. Recently successful treemaps have included the US Marine Corps' supply-chain management and Matrikon's process-monitoring tools, both based on the HiveGroup's (www.hivegroup.com) software. Other treemap applications include the SequoiaView free browser for PC hard drives (www.win.tue.nl/sequoiaview/) and Newsmap, which gives an overview of the world's news (www.marumushi.com/apps/newsmap/), with new products coming soon from Microsoft and Oracle. Scientific applications, such as viewing gene expression data according to the tree-structured gene ontology, are also gaining prominence.[5]

Not every data set lends itself to exploration by the rank-by-feature framework, but this new ap-

proach provides a rapid understanding of key features and patterns in many high-dimensional data sets. Readers can learn about the rank-by-feature framework and try our implementation in the Hierarchical Clustering Explorer (see the sidebar). We realize that the concepts are difficult for some users to follow, so our next step will be to simplify the interface and provide guidance to new users. Of course, those with deep knowledge of their domains will be best prepared to recognize the importance of the features they find in their data. Galileo's discovery of Jupiter's moons depended not only on his making the telescope, but also on his capacity to understand what he was seeing.

## References


1. J.W. Tukey and P.A. Tukey, "Computer Graphics and Exploratory Data Analysis: An Introduction," *Proc. Ann. Conf. and Exposition: Computer Graphics*, National Micrograhics Assoc., vol. 3, 1985, pp. 773–785.
2. J. Seo and B. Shneiderman, "Interactively Exploring Hierarchical Clustering Results," *Computer*, vol. 35, no. 7, 2002, pp. 80–86.
3. J. Seo and B. Shneiderman, "A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data," *Information Visualization*, vol. 4, no. 2, 2005, pp. 1–16.
4. J. Thomas and K. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE Press, 2005; http://nvac.pnl.gov/agenda.stm.
5. E.H. Baehrecke et al., "Visualization and Analysis of Microarray and Gene Ontology Data with Treemaps," *BMC Bioinformatics*, vol. 5, June 2004, p. 84; www.biomedcentral.com/1471-2105/5/84/.



**Ben Shneiderman** is a professor of computer science at the University of Maryland and founding director of its Human-Computer Interaction Lab. His research interests include information visualization and human-computer interaction. Shneiderman has a PhD in computer science from the State University of New York, Stony Brook. He is a senior member of the IEEE, a fellow of the ACM, and a fellow of the American Association for the Advancement of Science. Contact him at ben@cs.umd.edu.