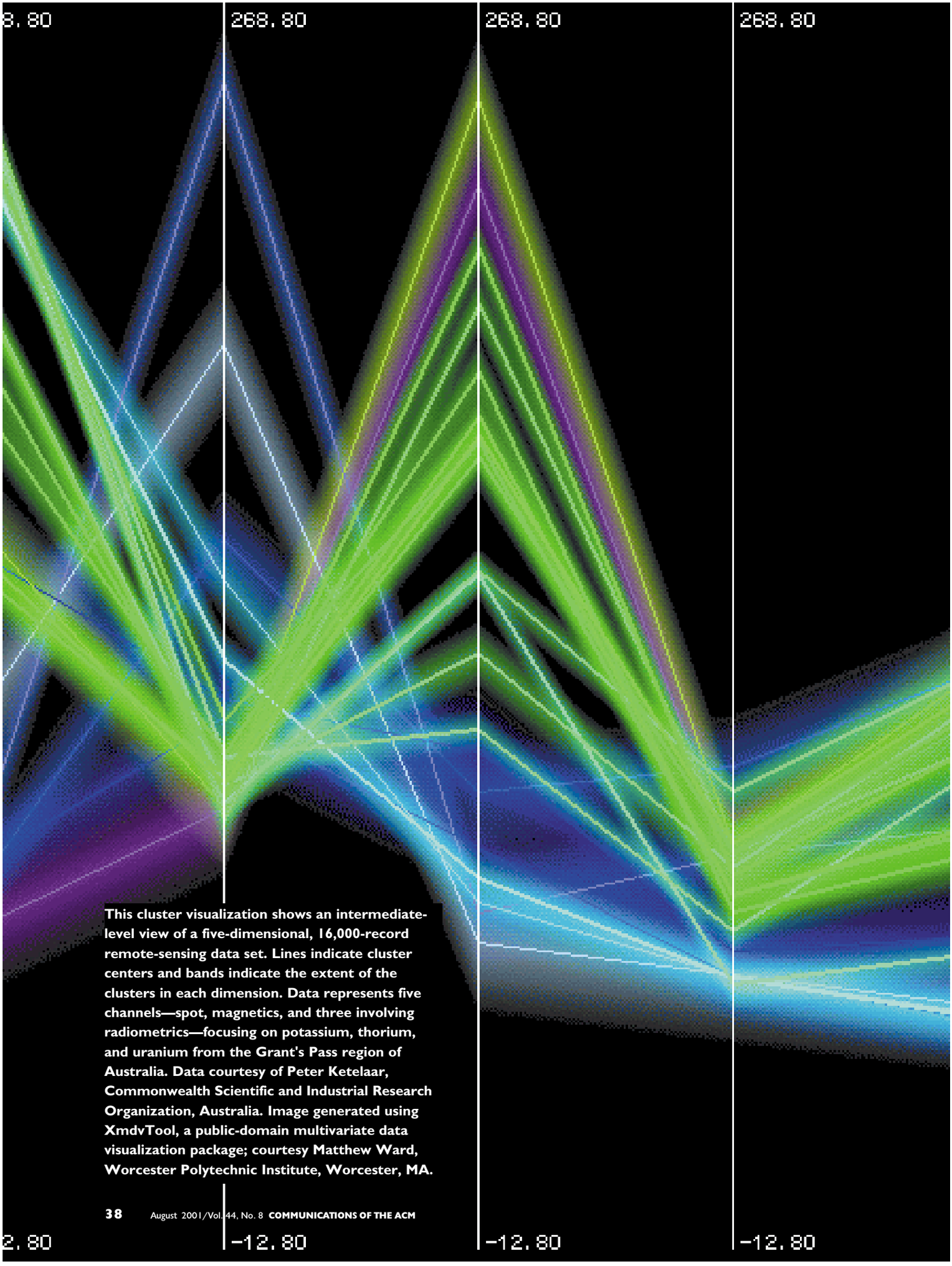


This cluster visualization shows an intermediate-level view of a five-dimensional, 16,000-record remote-sensing data set. Lines indicate cluster centers and bands indicate the extent of the clusters in each dimension. Data represents five channels—spot, magnetics, and three involving radiometrics—focusing on potassium, thorium, and uranium from the Grant's Pass region of Australia. Data courtesy of Peter Ketelaar, Commonwealth Scientific and Industrial Research Organization, Australia. Image generated using XmdvTool, a public-domain multivariate data visualization package; courtesy Matthew Ward, Worcester Polytechnic Institute, Worcester, MA.

Visual Exploration of Large Data Sets

Computer systems today store vast amounts of data. Researchers, including those working on the "How Much Information?" project at the University of California, Berkeley, recently estimated, about 1 exabyte (1 million terabytes) of data is generated annually worldwide, including 99.997% available only in digital form. This worldwide data deluge means that in the next three years, more data will be generated than during all previous human history.

In the rising tide of business transaction data, these tools help distinguish which are strategic assets and which are not worth collecting in the first place.

Daniel A. Keim

Data is often recorded, captured, and stored automatically via sensors and monitoring systems. Many of the simple transactions now part of our everyday lives, such as paying for food and clothes by credit card or using the telephone, are typically recorded for future reference by computers. Many parameters of each transaction are routinely captured, resulting in highly dimensional data. The data is collected because companies, including those engaged in some kind of e-commerce, view it as a source of potentially valuable information that, as a strategic asset, could provide a competitive advantage. But actually finding this valuable information is difficult. Today's data management systems make it possible to view only small portions of it. If the data is presented in text form, the amount that can be displayed amounts to only about 100 data items—a drop in the ocean when dealing with data sets containing millions of data items. Lacking the ability to adequately explore the large amounts being collected, and despite its potential usefulness, the data becomes useless and the databases data dumps. Visual data exploration, which aims to provide insight by visualizing the data, and information visualization techniques (such as distorted overview displays and dense pixel displays) can help solve this problem.

Effective data mining depends on having a human in the data exploration process while combining this person's flexibility, creativity, and general knowledge with the enormous storage capacity and computational power of today's computers. Visual data exploration seeks to integrate humans in the data exploration process, applying their perceptual abilities to the large data sets now available. The basic idea is to present the data in some visual form, allowing data analysts to gain insight into it and draw conclusions, as well as interact with it. The visual representation of the data reduces the cognitive work needed to perform certain tasks.

Visual data mining techniques have proved their value in exploratory data analysis; they also have great

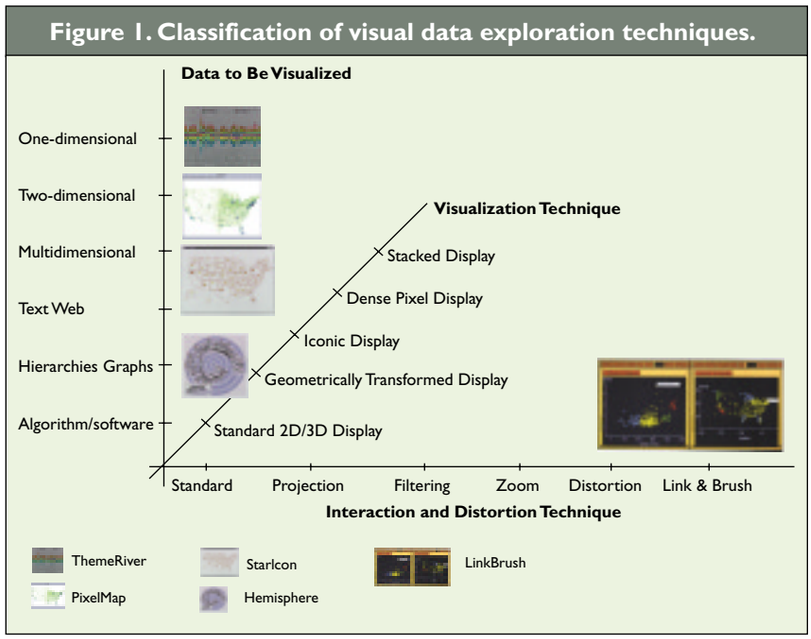


Figure 1. Classification of visual data exploration techniques.

LEFT VERTICAL IMAGES: PROCEEDINGS OF IEEE INFORMATION VISUALIZATION 2000
HORIZONTAL LINK & BRUSH IMAGES: AMERICAN STATISTICAL ASSOCIATION, 1996

potential for exploring large databases. Visual data exploration is especially useful when little is known about the data and the exploration goals are vague. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals might be done automatically through the interactive interface of the visualization software.

The visual data exploration process can be viewed as a hypothesis-generation process, whereby through visualizations of the data allow users to gain insight into the data and come up with new hypotheses. Verification of the hypotheses can also be accomplished via visual data exploration, as well as through automatic techniques derived from statistics and machine learning. In addition to granting the user direct involvement, visual data exploration involves several main advantages over the automatic data mining techniques in statistics and machine learning:

- Deals more easily with highly inhomogeneous and noisy data;
- Is intuitive; and
- Requires no understanding of complex mathematical or statistical algorithms or parameters.

As a result, visual data exploration usually allows faster data exploration, often delivering better results, especially in cases where automatic algorithms fail. In addition, the related techniques are essential for communicating complex data mining results to humans, even when machine learning or statistical techniques are employed. A visual representation provides a much higher degree of confidence in the findings of the exploration than a numerical or textual representation of the findings. This fact leads to strong demand for visual exploration techniques and makes them indispensable in conjunction with automatic exploration techniques.

Visual data exploration, also known as the "information seeking mantra" [11], usually follows a three-step process: overview, zoom and filter, and details-on-demand. In the overview step, the user identifies interesting patterns, focusing on one or more of them. To analyze the patterns, the user drills down to access details of the data. Visualization technology may be used for all three steps, presenting an overview of the data and allowing the user to identify interesting subsets. In analyzing the patterns, it is important to maintain the overview visualization while focusing on the subset using another visualization technique. An alternative is to distort the overview visualization in order to focus on the interesting subsets. Note that visualization technology provides not only the base visualization techniques for all three steps but bridges the gaps between the steps.

Visualization Techniques

Information visualization focuses on data sets lacking inherent 2D or 3D semantics and therefore also lacking a standard mapping of abstract data onto the physical space of the paper or screen. A number of well-known techniques visualize (partially) such data sets, including x-y plots, line plots, and histograms. These techniques are useful for data exploration but are limited to relatively small low-dimensional data sets. A large number of novel information visualization techniques have been developed over the past decade, allowing visualizations of ever larger and more complex, or multidimensional, data sets [4].

These techniques are classified using three criteria: the data to be visualized, the technique itself, and the interaction and distortion method (see Figure 1). For visualizing a specific data type, any of the visualization techniques can be used in conjunction with any of the interaction and distortion methods. Note that the classification does not assume disjoint categories, as multiple visualization techniques can be combined with multiple interaction techniques.

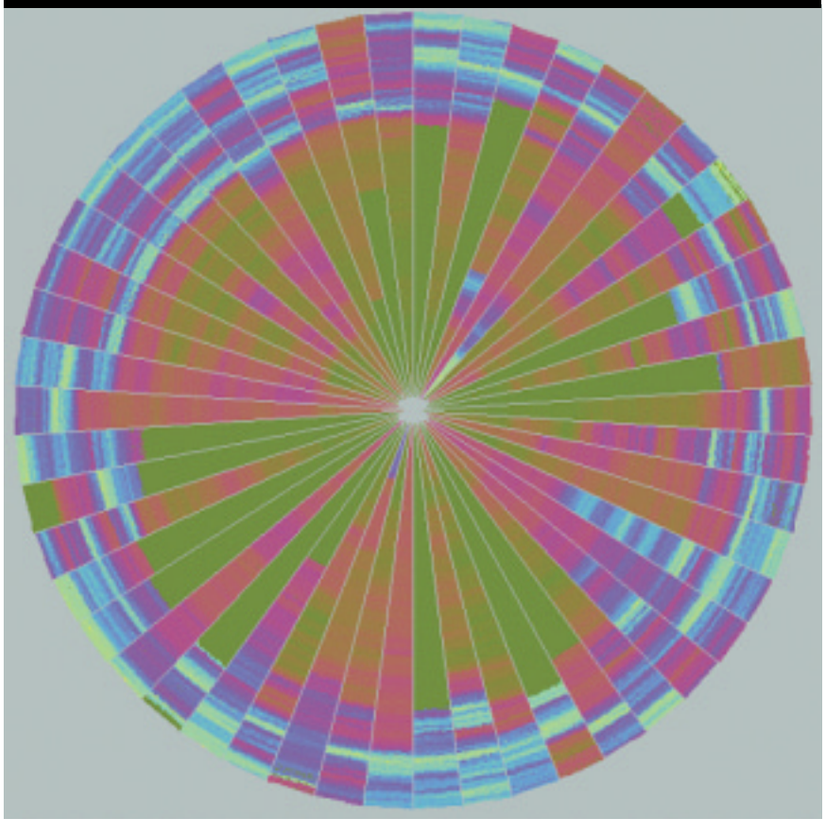
The classification begins with the data type to be visualized [11], including whether it is:

- One-dimensional (such as temporal data, as in Figure 2);
- Two-dimensional data (such as geographical maps, as in Figure 3);
- Multidimensional data (such as relational tables, as in Figure 4);
- Text and hypertext (such as news articles and Web documents);
- Hierarchies and graphs (such as telephone calls and Web sites, as in Figure 5); and
- Algorithms and software (such as debugging operations).

The visualization technique fits into one or more of the following categories, as identified in Figure 1:

- Standard 2D/3D displays using standard 2D or 3D visualization techniques (such as x-y plots and landscapes) for visualizing the data.
- Geometrically transformed displays using geometric transformations and projections to produce useful visualizations. Included are parallel coordinates (see Figure 4), projection pursuit, and the various techniques for visualizing graphs [3].
- Icon-based displays that visualize each data item as an icon (such as stick figures) and the dimension values as features of the icons. Figure 1 shows a thumbnail of a star-map view [1]; one star icon maps the call volume of that state with all other states to the length of the star segments with the direction corresponding to the approximate direction of the state.
- Dense pixel displays that visualize each dimension value as a color pixel and group the pixels belonging to each dimension into an adjacent area [6]. By arranging and coloring the pixels in an appropriate way, the resulting visualization provides detailed information on local correlations, dependencies, and hot spots, as in Figure 2.
- Stacked displays that visualize the data partitioned hierarchically. In multidimensional data, the data dimensions to be used for building the hierarchy have to be selected carefully. To obtain a useful visualization, the most important dimensions have to correspond to the first levels of the hierarchy.

Figure 2. The pixel-oriented circle segments technique [6], showing daily data over about 20 years (1974–1995) of 50 stocks in the Frankfurt Allgemeine Zeitung (Frankfurt Stock Index). Note the three bright outer rings corresponding to high-price periods and subsequent low-price periods. The technique maps each data value to a colored pixel; high values correspond to bright colors. The various stocks are also mapped to the segments of the circle; the pixels are arranged in a back-and-forth fashion adjacent to the segment-halving line.



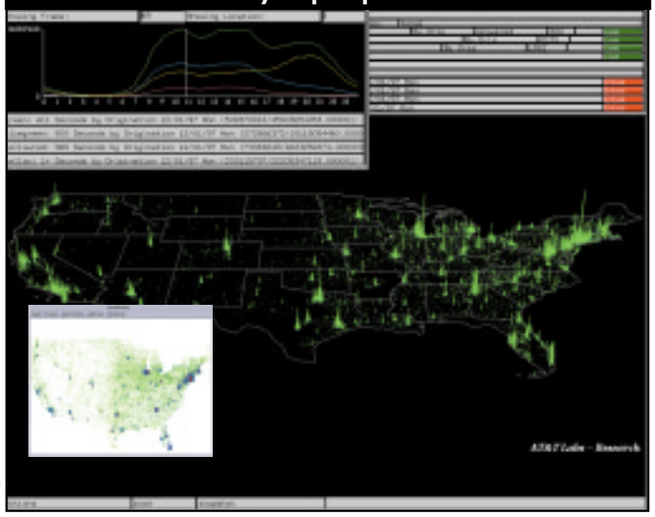
IEEE, 1999

The techniques associated with each of these categories differ in how they arrange the data on the screen (such as 2D display or semantic arrangement) and how they deal with multiple dimensions in case of multidimensional data (such as multiple windows, icon features, and hierarchy).

Interaction and Distortion Techniques

In addition to these techniques, data exploration also depends on interaction and distortion techniques. Interaction techniques, which allow users to interact directly with a visualization, include filtering, zooming, and linking, thus allowing the data analyst to make dynamic changes

Figure 3. The SWIFT-3D system [8], showing call volume data from the AT&T long-distance network. Developed at AT&T Research Labs, the system integrates relevant visualization techniques ranging from statistical displays (such as line graphs and histograms) for overview displays and interactive data selection, to pixel-oriented visualizations for a bird's eye overview and navigation in 3D displays, to interactive 3D maps, to drag-and-drop query tools for interactive detailed viewing of data from a variety of perspectives.



IEEE, 1996

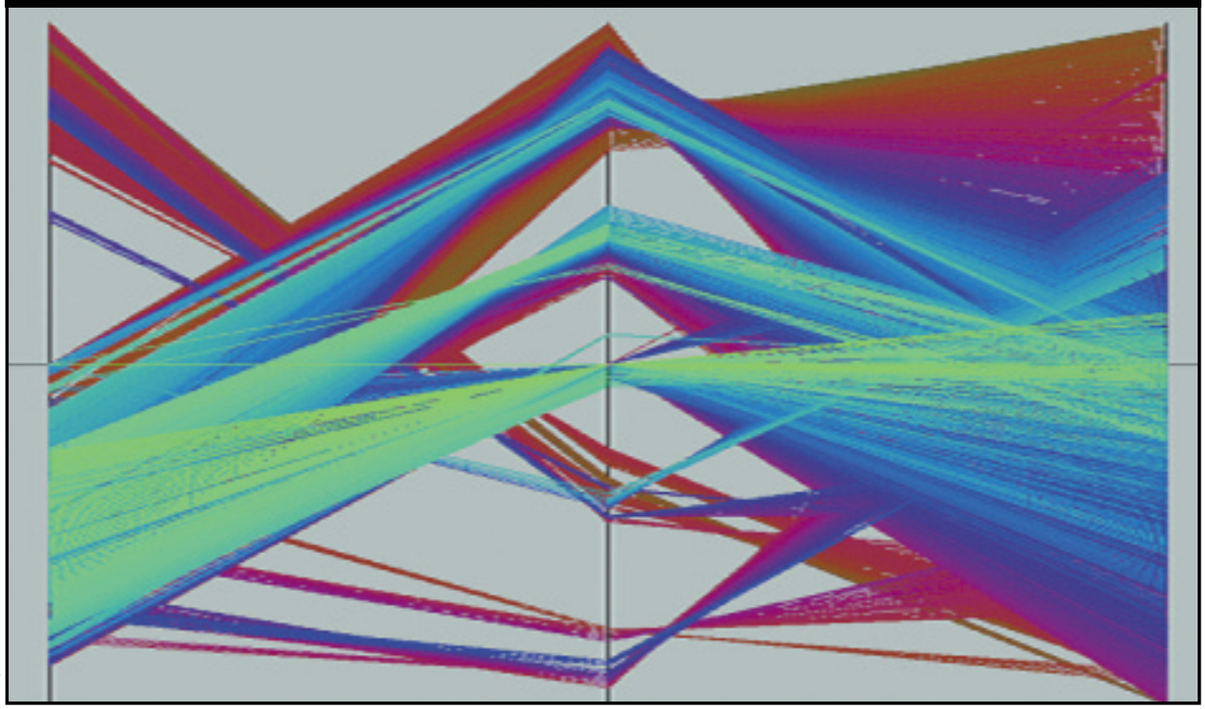
of a visualization according to the exploration objectives; they also make it possible to relate and combine multiple independent visualizations. Note that connecting multiple visualizations through interactive techniques provides more information than considering the component visualizations independently, as in the lower-right thumbnail in Figure 1.

Interactive distortion techniques support the data exploration process by preserving an overview of the data during drill-down operations. Basically, they show portions of the data with a high level of detail and other portions with a lower level of detail. Popular distortion techniques are hyperbolic and spherical, as in Figure 5, and the TableLens approach developed by R. Rao and S. Card at Xerox PARC, for multiattribute tabular data, such as that derived from customer and shopping behavior [10].

Evaluating Techniques and Systems for Suitability

Visualization techniques and visual data exploration systems can be evaluated and compared with respect to their suitability for certain data characteristics (such as data types, number of dimensions, number of data items, and category). Task charac-

Figure 4. The parallel coordinates technique [5]. In conjunction with similarity-based coloring, it displays each multidimensional data item as a polygonal line intersecting the dimension axes at the position corresponding to the data value for the dimension.



IEEE, 2000

teristics include clustering, classification, associations, and multivariate hot spots; visualization characteristics include visual overlap and learning curve.

Different visualization techniques are used for visualizing different data types. Some are specially designed to support one specific data type; others are more general, allowing the visualization of a range of data types. General visualization techniques are not equally suited for all data characteristics; for example, icon-based visualization techniques allow only the visualization of a limited number of dimensions, and pixel-based techniques are not suitable for categorical data.

As there is no universal technique, each one has to be evaluated for its suitability for the task at hand [7]. While some are specially designed for certain tasks (such as classification and clustering) [2], others are more general, useful for a range of tasks. Desirable visualization characteristics for any technique include limited visual overlap, fast learning, and good recall. Undesirable visualization characteristics include occlusions and line crossings that might appear to the user/viewer as an artifact limiting the usefulness of the visualization technique.

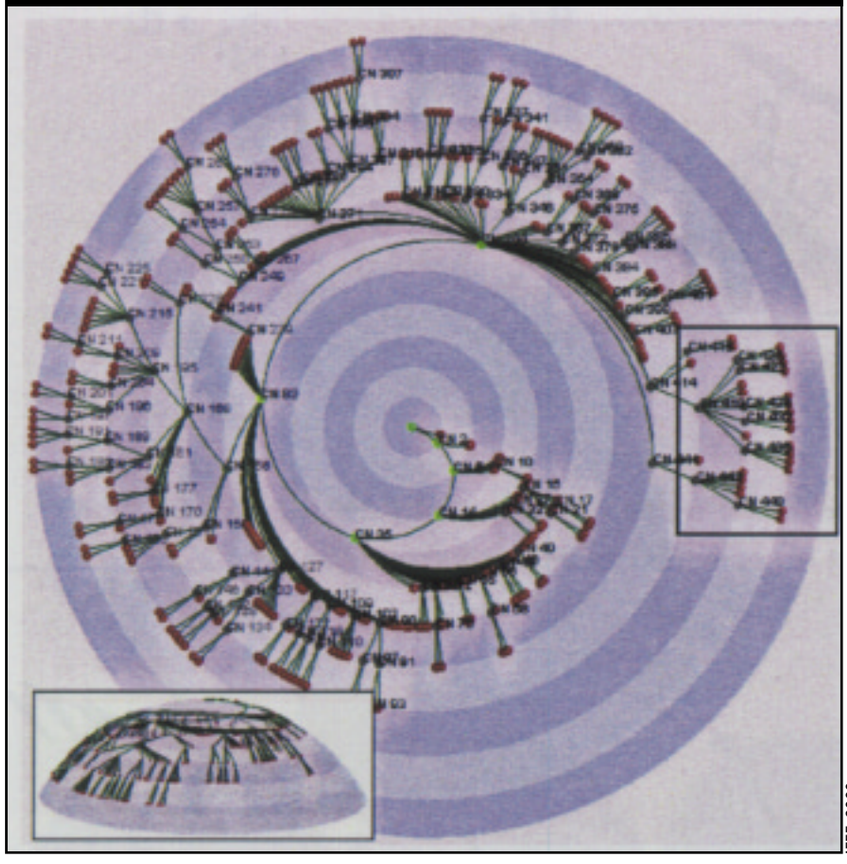
Well-regarded visual data exploration and analysis research prototypes include the XmdvTool developed by M. Ward and his students at the Worcester Polytechnic Institute, Worcester, MA, and the VisDB system [7], I developed with my students at the Universities of Munich, Halle, and Konstanz. Statistical data analysis packages include S Plus, developed by R. Becker, J. Chambers, and A. Wilks at AT&T Research Labs (commercially available from Insightful, www.insightful.com), and XGobi developed by D. Swayne, D. Cook, and A. Buja at AT&T Research Labs. Commercial visual data exploration systems include Silicon Graphics' MineSet, the DecisionSite system from Spotfile (www.spotfire.com), and eBizinsight from Visual Insights (www.visualinsights.com).

Conclusion

Addressing the important but challenging problem of how to explore large data sets promises business benefits for a range of organizations, including those involved in e-commerce. Visual data exploration has great potential for revealing interesting patterns in data (such as clusters, correlations, dependencies, and exceptions). Within the next two to five years, many applications, including fraud detection, marketing, and data mining, will incorporate information visualization technology to improve their data analysis functions.

The next step for data analysts will involve the tight integration of visualization tools with traditional techniques from such disciplines as statistics, machine learning, operations research, and simulation. Integration of visualization tools and these more established methods would combine fast automatic data mining algorithms with the intuitive power of the human mind, improving the quality and speed of the data exploration process. Visual exploration also needs to be tightly integrated with the systems used to manage the vast amounts of relational and semistructured information, including

Figure 5. The hemisphere hierarchy visualization technique [9]. The result maps a 2D layout algorithm onto a hemisphere, providing a nice overview, good focus, and context operations, even for very large graphs.



IEEE, 2000

database management and data warehouse systems.

The ultimate goal—possibly within five years—is to bring the power of visualization technology to any desktop machine, providing a more intuitive, faster exploration of very large data resources. This power and convenience will be valuable not only in the economic sense but be a delight to use while prompting users to think in new ways about their data. C

References

1. Abello, J. and Korn, J. Visualizing massive multi-digraphs. In *Proceedings of Information Visualization'00* (Salt Lake City, UT, Oct. 9–13). IEEE Computer Science Press, Los Alamitos, CA, 2000, 39–47.
2. Ankerst, M., Elsen, C., Ester, M., and Kriegel, H. Visual classification: An interactive approach to decision tree construction. In *Proceedings of Knowledge Discovery in Databases'99* (San Diego, CA, Aug. 15–18). ACM Press, New York, 1999, 392–396.
3. Battista, G., Eades, P., Tamassia, R., and Tollis I. *Graph Drawing: Algorithms for the Visualization of Graphs.* Prentice Hall, Englewood Cliffs, NJ, 1999.
4. Card, S., Mackinlay, J., and Shneiderman, B. *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann, San Francisco, 1999.
5. Inselberg, A. and Dimsdale, B. Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proceedings of Visualization'90* (San Francisco, Oct. 23–26). IEEE Press, Los Alamitos, CA, 1990, 361–370.
6. Keim, D. Designing pixel-oriented visualization techniques: Theory and applications. *Transact. Vis. Comput. Graph. 6,* 1 (Jan.–Mar. 2000), 59–78.
7. Keim, D. An introduction to information visualization techniques for exploring very large databases. Tutorial notes, Information Visualization'00 (Salt Lake City, UT, Oct. 9–13, 2000).
8. Koutsofios, E., North, S., and Keim, D. Visualizing large telecommunication data sets: Visualization Blackboard. *IEEE Comput. Graph. Appl. 19,* 3 (May/June 1999), 16–19.
9. Kreusler, M., Lopez, N., and Schumann, H. A scalable framework for information visualization. In *Proceedings of Information Visualization'00* (Salt Lake City, UT, Oct. 9–13). IEEE Computer Science Press, Los Alamitos, CA, 2000, 27–35.
10. Rao, R. and Card, S. The TableLens: Merging graphical and symbolic representation in an interactive focus-context visualization for tabular information. In *Proceedings of Human Factors in Computing Systems CHI'94* (Boston, Apr. 24–28). ACM Press, New York, 1994, 318–322.
11. Shneiderman, B. The eyes have it: A task by data-type taxonomy for information visualizations. In *Proceedings of Visual Languages* (Boulder, CO, Sept. 3–6). IEEE Computer Science Press, Los Alamitos, CA, 1996, 336–343.
12. Ware, C. *Information Visualization: Perception for Design.* Academic Press, San Diego, CA, 2000.

Daniel A. Keim (keim@informatik.uni-konstanz.de) is a professor in and head of the Database and Visualization Group, University of Konstanz, Germany, and a senior researcher in the Information Visualization Research Department of AT&T Shannon Labs, Florham Park, NJ.