Visualization Techniques for Data Mining in Business Context: A Comparative Analysis

Ralph K. Yeh

University of Texas at Arlington
Box 19437, Arlington, TX 76019
817-272-3707 Fax: (817) 272-5801
E-mail: ryeh@uta.edu

ABSTRACT

The information acquired from vast amount of operation data is a critical asset in today's fierce business competition arena (Marakas, 2003). Data visualization is a relatively new method to tap the knowledge treasures hide in data warehouse (Mirel 1998; Roth et al, 1997). According to Ahrens et al (2001), the size of dataset in data visualization tasks will continue to grow because of its ever increasing applicability in various domains. It is obvious that the influence and significance of data visualization cannot be underestimated. Speier and Morris (2003) also emphasized the demand for more studies on data visualization related topics.

This paper tries to explore issues about the visualization techniques in the context of business data mining, especially the ways to compare between them. Prior studies about issues in data visualization are presented, and some widely-used visualization techniques are listed and described. Next, a set of criteria derived from data-feature and usage perspectives are proposed. Finally, a comparative analysis across listed visualization techniques are conducted and concluded based on proposed criteria. We expect to make a contribution by providing an insight into strengths and weaknesses across listed visualization techniques that can be used by both academia and practitioners.

INTRODUCTION

The combination of data warehouse, data mining and data visualization is gradually becoming an indispensable organizational weapon for achieving competitive advantage in many data-driven industries (Marakas, 2003). Nabney et al (2005) indicated that data structural features can be effectively recognized by data seekers using data visualization. Although data visualization has proven to be a powerful tool in data mining and knowledge discovery (Wang et al, 2000), its use in business and corporate world is still new and fledgling (Mirel 1998; Roth et al, 1997).

According to Ahrens et al (2001), the size of dataset in data visualization tasks will continue to grow because of its ever increasing applicability in various domains. It is obvious that the influence and significance of data visualization cannot be underestimated. Moreover, Speier and Morris (2003) also emphasized the demand for more studies on data visualization related topics.
The structure of this paper is as follows: Firstly, prior research about concepts and issues in the area of data visualization are presented, and five widely-used visualization techniques are described. Secondly, a set of criteria derived from data features, usage, and context perspectives is proposed. Thirdly, the strengths and weaknesses across five visualization techniques are

discussed based on proposed criteria. Finally, a conclusion matrix is presented and possible future research directions are indicated.

DATA VISUALIZATION

The Importance of Visualization

Why do we need to visualize data? Data mining algorithms can figure out hidden data patterns as well. As an alternative to mechanical data mining algorithms, visual exploration has proven as an effective tool in data mining and knowledge discovery (Wang et al, 2000). Data structural features can be effectively recognized by data seekers using data visualization (Nabney et al, 2005).

"Data visualization is the process by which textual or numerical data are converted into meaningful images" (Marakas, 2003). The reason why the data visualization can help on data mining is that the human brain is very effective in recognizing large amounts of graphical representations (Ware, 2004). Hence, if the visualization techniques can correctly convert the raw data into visual graphs, users can very likely detect the patterns hidden in text and numbers. This process of recognizing patterns through human brain can facilitate users to understand the meaning of patterns more intuitively. Therefore, visualization can complement the data mining techniques. The combination of data mining and data visualization, plus the enormous storage space in data warehouse, can provide precious information to business decision makers today.

Information and Scientific Visualization

Data visualization is accepted as the new name of this discipline which consisted of two existing sub-areas: information visualization and scientific visualization (Post et. al., 2003). The study of scientific visualization was officially launched through a research recommendation made by the National Science Foundation (NSF) in 1987 (Ma, 2001). Approximately the same time, the emerging data warehouse and data mining (Han and Kamber, 2000) paved the way for information visualization to apply on high dimensional business datasets.

In general, the variables in a typical scientific visualization task are continuous and are about volumes, surfaces, etc. Information visualization tasks are apropos of categorical variables and the recognition of patterns, clusters, trends, outliers, and gaps (Shneiderman, 2003). A typical data mining task in a business data warehouse context is more related to information visualization.

LITERATURE REVIEW

Data visualization research have no theoretical background (Johnson, 2004) and very few evaluation studies (Au et. al, 2000). From limited literatures, the discussion about difference between scientific and information visualization along with their future direction, as well as review for prior visualization technique comparison research, are addressed below.

Information Visualization vs. Scientific Visualization

The foreword of the proceedings of the first IEEE Symposium on Information Visualization clearly addressed the definition of information visualization and scientific visualization: "Information visualization is a process of transforming data and information that are not inherently spatial into a visual form, allowing the user to observe and understand the information. This is in contrast with scientific visualization, which frequently focuses on spatial data generated by scientific purposes." (Gershon and Eick, 1995). The two highly-related areas are developed separately and the mutual interactions are very limited (Johnson, 2004).

However, the differences between these two fields are only on their evolving history (Munzner, 2002). Some existing areas, such as cartographic and geographic information techniques, are positioned across information visualization and scientific visualization. Moreover, the latest bioinformatics research, especially genomic data visualization, again challenge the thin line between definitions of information visualization and scientific visualization (Rhyne, 2003).

With reference to the above, calls to combine the efforts can be heard and the consolidation work is underway. However, the integration of scientific and information visualization is listed as one of the top scientific visualization research problems (Johnson, 2004). This clearly showed that there are still obstacles to tackle and that the development is moving in a positive direction.

Visualization Techniques Comparison

As we mentioned earlier, the researchers do not present the "evaluation of the proposed methods and quantification of the effectiveness of their techniques" is another one of the top problems in current scientific visualization researches (Johnson, 2004). Prior research focused more on the construction of new visualization techniques, but very few studies devoted to the evaluation or comparison of these proposed techniques. One of the reasons under this phenomenon is that the evaluation criteria are very hard to define, or operationalize.

Some articles provided related discussion such as visualization technique selection or the concept of useful evaluation criteria. Grinstein and Ward (2002) mentioned four factors to consider when selecting visualization techniques, and Grinstein et al (2002) noted eight sets of criteria concepts to be guidelines for developing evaluation measures to visualization techniques.

In the same paper which Grinstein et al (2002) presented above-mentioned concepts, they conducted a benchmarking experiment on five visualization techniques. The comparison was based on the effectiveness of identifying known features within eleven benchmark datasets, and not based on proposed criteria concepts. This may to some extent reflect the above-mentioned problem of lacking objective comparison criteria.

VISUALIZATION TECHNIQUES

There is a great variety of visualization techniques proposed (Chen, 2004), and they can be grouped according to different perspectives (Bajaj, 1999). In addition, there is no single optimal visualization technique for all situations, especially for high dimensional data (Hoffman and

Grinstein, 2002). Although the visualization techniques presented below are not necessarily exhaustive, they are quite widely-accepted and representative in many ways.

TreeMap

For a hierarchical data structure, tree-based representations are the most common adapted techniques (Itoh et. al., 2004). Tree-shaped diagrams and TreeMaps are two subcategories in presenting hierarchical dataset (Figure 1, adopted from Zhang et al, 2004). The TreeMap technique utilizes the space-filling approach to show the scale of quantitative data (Chen, 2004), while the Tree-shaped diagram technique focuses on the connectivity of the hierarchy structure (Itoh et. al., 2004). For example, TreeMap is one of the most effective paradigms to visualize hard disk structures (Zhang et. al., 2004).

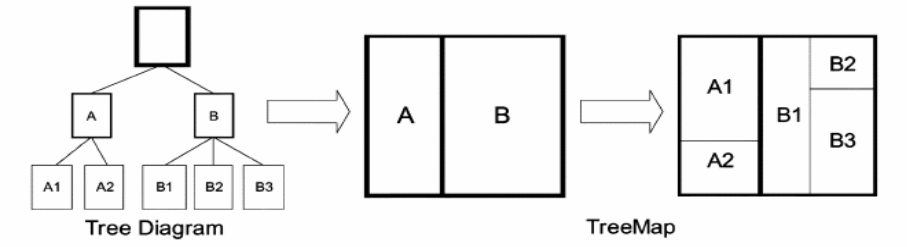

Figure 1: Tree Diagram and TreeMap

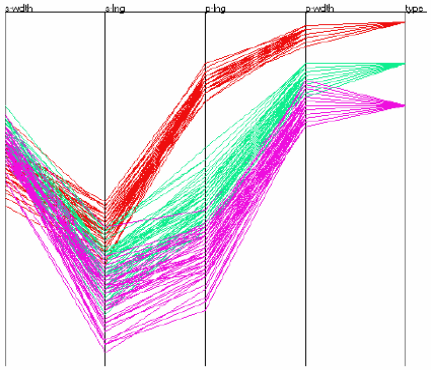

Figure 2: Parallel Coordinates

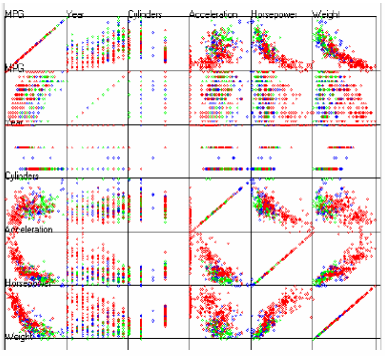

Figure 3: Scatter-Plot Matrices

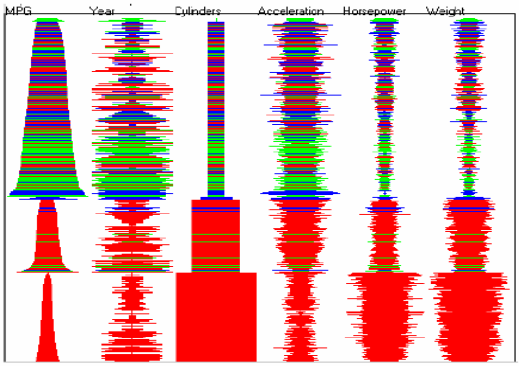

Figure 4: Survey Plots

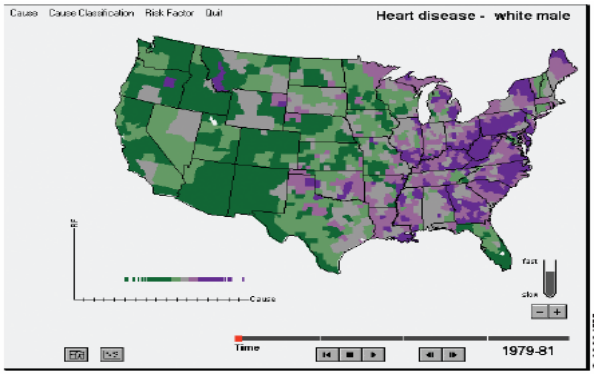

Figure 5: Spatial/Geographic Visualization

Parallel Coordinates

Multidimensionality is a common phenomenon of business-related datasets in a typical data warehouse (Soukup and Davidson, 2002). Visualizing multidimensionality is a difficult problem due to the nature of data visualization (Bishop and Tipping, 1998). The parallel coordinates technique, proposed by Inselberg (1985), can markedly visualize multidimensional data in a straightforward manner (Figure 2, Adopted from Grinstein et al, 2002). In a parallel coordinates graph, each dimension becomes a uniformly spaced vertical (can be horizontal) axis. A data instance can be represented as a polyline that links across all the axes (Yang et al, 2003).

Scatter-Plot Matrices

The scatter plot diagram is a popular visualization technique in comparing two data columns (Soukup and Davidson, 2002). Scatter-plot matrices, introduced by Cleveland and McGill (1988), simply extend the design to include extra dimension combinations by arranging all scatter plot diagrams in a matrix form (Figure 3, Adopted from Grinstein et al, 2002).

Survey Plots

The survey plots technique can also present multidimensional dataset. Rao (1994) use survey plot for table lens in Xerox PARC, for example. It is similar to circle segments and permutation matrix (Hoffman and Grinstein, 2002). The different dimensions can be represented in columns, line length, and color. For instance, the dataset in figure 4 (Adopted from Grinstein et al, 2002) is the attributes of American (Red), Japanese (Green) and European cars. It is sorted first by cylinders and then miles-per-gallon (MPG).

Spatial Visualization

Spatial representations are one of the most familiar design alternatives of information visualization, and geographic information system (GIS) is a major source of inspiration for spatial visualization (Chen, 2004). The visualization of datasets which have spatial or geographical attributes can usually enhance our comprehension to the datasets. Prior research shows that 80% of the business decisions involve geographic data (Mennecke, 1997). Consequently, exploring datasets from spatial perspective is intuitive and important (Figure 5, Adopted from Rhyne, 2003).

COMPARISON CRITERIA

The quantitative metrics capable of evaluating and comparing the effectiveness of the variety of visualization techniques are one of the topics that can further enhance the usefulness of data visualization (Collier et. al., 2002). Currently, the lack of assessing criteria to evaluate information visualization, either independently or in user activities context, remains a fundamental research issue (Chen, 2004).

Some researchers even claim that an objective measure of quality in data visualization cannot be found (Bishop and Tipping, 1998). As such, to quantify the merit of a specific visualization

technique seems not practical. However, the effectiveness of a visualization technique generally depends on the input dataset (Wang et al, 2000). Hence, we try to shed a light to this issue from the angles of dataset and task characteristics, and to present a possible set of criteria to evaluate and compare the visualization techniques.

Data Structure – Multidimensional, Hierarchical and Spatial

Multidimensional, hierarchical and landscaped (spatial) data are the main data structure types which are common in data warehouses. Yet, hierarchical and landscaped (spatial) data are relatively specialized and mostly inherited certain features (Soukup and Davidson, 2002). Different visualization techniques are needed to show the variances between these data structure features.

Data Mining Phases

Data visualization can provide great help in data mining task, and we are trying a step further to compare the usefulness of different techniques in general data mining phases. Data mining can divide into three general phases: pre-processing the data, introducing algorithm, and post-processing the result (Bernstein et al, 2005).

For pre-processing phase, visualization can contribute in attribute selection and outlier detection. Visualization techniques can also offer clues on choosing useful learning algorithms at second phase, and even provide insight and understanding to the result in post-processing stage (Witten and Frank, 2000).

Visualization Task Purpose

Also from usage perspective, the objectives of visualization tasks can be categorized into three general types: data exploration, hypothesis confirmation, and visual presentation (Grinstein and Ward, 2002). For each task purpose, specific visualization technique can be selected based on the requirements of that specific purpose.

COMPARISON DISCUSSION

The comparison details based on above-mentioned criteria across the proposed five visualization techniques are presented below, and the discussion are arranged in the order of comparison criteria. Given the quantitative comparison metrics are not present (Bishop and Tipping, 1998), the following discourse is mainly qualitative in nature.

Data Structure

From data structure perspective, the visualization techniques chosen must have fitness to the structural characteristics of the data set (Soukup and Davidson, 2002). In other words, the feature of a dataset can be the first index to locate appropriate visualization techniques for that dataset. For example, if a database has a built-in hierarchical structure, then the tree-related techniques can be a good starting point to explore the dataset extracted from the database.

Besides, usually the datasets in a business data warehouse environment are in the form of multi-dimensional structure (Shneiderman, 2003). Thus, the parallel coordinates, scatter-plot matrices, and survey plots can be utilized effectively. In some cases, the hierarchical or spatial features will be inherited in the data. Then the treemap or spatial visualization techniques are more suitable in mining these datasets.

VISUALIZATION TASK PURPOSE

Data Exploration Task

Geographical layout is the most natural means to organize raw data (Lokuge et al, 1996). Hierarchical feature can also be observed commonly and understood easily (Soukup and Davidson, 2002). Hence, the contribution of treemap and spatial visualization techniques are limited in exploration task because the perceptions to the hierarchical or spatial relationship within the data are too dominant for other possibilities.

However, all of the other three techniques can handle high dimensional datasets. The scatter-plot matrixes extended the two dimensional scatter plot diagram for users to observe multiple two-dimensional relationships at the same time. Survey plots and parallel coordinates also designed to represent data from multiple columns at one screen (Grinstein et al, 2002). Therefore, those three techniques can provide more insight into the data exploration task in a multi-dimensional business data context.

Hypothesis Confirmation

When the objective of a visualization task is to confirm a hypothesis that is already known, the traditional two-dimensional scatter plot will be the preferable way for users to recognize the relationship intended to check (Shneiderman, 2003). Consequently, scatter-plot matrices technique is most helpful in hypothesis confirmation scenario because the two-dimensional scatter plot presentation format is kept.

Parallel coordinates and survey plots are less useful since the dimensional representation in those two techniques are modified into non-intuitive multiple parallel lines. Treemap and spatial techniques are not necessarily able to represent the hypothesis, thus both are the last choice among the five alternatives.

Visual Presentation

The study of visualization perception is also an important part of information visualization (Ware, 2004). The objective of visual presentation is to impress and influence the audience. Therefore, the most critical feature in choosing proper visualization techniques is the ease-of-perception. Treemap and spatial visualization techniques can afford the audience with most direct perception to the information carried (Lokuge et al, 1996; Soukup and Davidson, 2002).
The scatter-plot matrices is second to spatial and treemap techniques. The scatter-plot format is easier to grasp then the final batch, i.e. survey plots and parallel coordinates, which require certain amount of explanation to fully recognize the mapping meanings.

DATA MINING PHASES

Pre-processing phase

In prior research, the survey plot technique has been proven to perform better than parallel coordinates and scatter-plot matrices in recognizing the important features, exact rules or models (Grinstein et al, 2002). While treemap and spatial visualization are useful only in data with hierarchical and spatial feature, parallel coordinates and scatter-plot matrices can apply to general tasks like outlier detection and attributes selection (Grinstein et al, 2002).
With reference to the above, at pre-processing phase of a data mining task, survey plot can be the first choice, followed by scatter-plot matrices and parallel coordinates. Treemap and spatial visualization techniques are only useful under specific scenarios.

Post-processing phase

After the data mining algorithms performed, the visualization can help check the mined discoveries, or perform new exploration. The pros and cons are similar to the confirmation and exploration tasks mentioned in earlier paragraphs. Thus, the conclusion is identical: parallel coordinates, scatter-plot matrices, and survey plots are better than treemap and spatial visualization.

CONCLUSION

The fore-stated discussion is hereby summarized as Table 1. For hierarchical datasets, treemap can catch the relationship immediately and present the linkage easily, so is the case for spatial/geographical datasets and spatial visualization technique. However, for general multi-dimensional dataset which is common in business data sources, the effectiveness of these two methods is falling behind the other three techniques.

Among parallel coordinates, scatter-plot matrices, and survey plots, scatter-plot matrices is more recommendable. It performs well both in exploration and confirmation tasks, while maintaining its usefulness in presentation task and under pre-processing phase of data-mining task. Survey plot is ahead of parallel coordinates only when assisting pre-processing phase in a data mining job, but is equal to parallel coordinates at all other situations.

	Data Structure	Visualization Task Purpose			Data Mining Step	
Visualization Techniques	**Multi-Dimensional / Hierarchical / Spatial**	**Exploration**	**Confirmation**	**Presentation**	**Pre-processing**	**Post-processing**
TreeMap	Hierarchical	O		V		O
Parallel Coordinates	Multi-Dimensional	V	O		O	V
Scatter-Plot Matrices	Multi-Dimensional	V	V	O	O	V
Survey Plots	Multi-Dimensional	V	O		V	V
Spatial Visualization	Spatial / Geographical	O		V		O

Table 1: Summary of Visualization Techniques V: Good, O: Useful

The information acquired from vast amount of operation data is a critical asset in today's fierce business competition arena (Marakas, 2003). Data visualization is a relatively new method to tap the knowledge treasures hide in the data warehouse (Mirel 1998; Roth et al, 1997). This paper discusses and prioritizes five popular data visualization techniques in different scenarios of business context. For practitioners, the conclusion matrix can provide a guideline on selecting proper visualization techniques according to data features and usage. For academia, this paper is a start point in exploring a seldom-studied area. Details of possible future research directions are presented in next section.

FUTURE RESEARCH DIRECTIONS

Different ways to evaluate visualization techniques: The first possible direction in future researches will be some other criteria to evaluate visualization techniques. This problem has been listed as one of the top fundamental research problems in this area (Johnson, 2004), and some researchers even claim that no objective criteria can be found (Bishop and Tipping, 1998). However, after this paper presented a solution from data and usage perspectives, there may still be possible ways to approach a feasible answer to visualization evaluation problem.

The empirical study to test presented discussion conclusion: The empirical study will be helpful and meaningful to examine the theoretical discussion and reasoning of this paper.

The typology of visualization techniques: Basically, a useful visualization technique typology does not exist. The typology can surely pave the way for deriving valid evaluation criteria.

The integration of information and scientific visualization: This problem is also among the top research problems for visualization (Johnson, 2003). Many new research topics are positioned between or across the original domain of scientific and information visualization (Rhyne, 2003). The joint efforts will definitely broaden the scope and methodologies for both sub-disciplines.

REFERENCES

Ahrens, J., Brislawn, K., Martin, K., Geveci, B., Law, C. C., & Papka, M. (2001). Large-scale data visualization using parallel data streaming. *IEEE Computer Graphics and Applications, 21*(4), 34-41.

Au, P., Carey, M., Sewraz, S., Guo, Y., & Ruger, S. M. (2000). New paradigms in information visualization. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Formation Retrieval,* 307-309.

Bajaj, C. (Ed.). (1999). *Data visualization techniques*. New York: Wiley.

Bernstein, A., Provost, F., & Hill, S. (2005). Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 503-518.

Bishop, C.M and Tipping, M.E. (1998). A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 282-293.

Chen, C. (2004). *Information visualization* (2nd ed.). London: Springer.

Cleveland, W., & McGill, M. (1988). *Dynamic graphics for statistics.* Wadsworth, Inc.

Collier, K., Medidi, M., & Sautter, D. (2002). Visualization in the knowledge discovery process. In U. Fayyad, G. Grinstein & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery*. California: Morgan Kaufmann.

Gershon, N., & Eick, S. (1995). Foreword. *Proceedings of IEEE Symposium on Information Visualization (InfoVis 95),* vii-viii.

Grinstein, G., Hoffman, P., & Pickett, R. (2002). Benchmark development for the evaluation of visualization for data mining. In U. Fayyad, G. Grinstein & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery*. California: Morgan Kaufmann.

Grinstein, G., & Ward, M. (2002). Introduction to data visualization. In U. Fayyad, G. Grinstein & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery*. California: Morgan Kaufmann.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques*Morgan Kaufmann.

Hoffman, P., & Grinstein, G. (2002). A survey of visualizations for high-dimensional data mining. In U. Fayyad, G. Grinstein & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery*. California: Morgan Kaufmann.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer, 1*, 69-91.

Itoh, T., Yamaguchi, Y., Ikehata, Y., & Kajinaga, Y. (2004). Hierarchical data visualization using a fast rectangle-packing algorithm. *IEEE Transactions on Visualization and Computer Graphics*, *10*(3), 302-313.

Johnson, C. (2004). Top scientific visualization research problems. *IEEE Computer Graphics and Applications,* 13-17.

Lokuge, I., Gilbert, S. A., & Richards, W. (1996). Structuring information with mental models: A tour of Boston. *Proceedings of CHI'96.*, 413-419.

Ma, K. (2001). Large-scale data visualization. *IEEE Computer Graphics and Applications*, 22-23.

Marakas, G. M. (2003). *Modern data warehousing, mining, and visualization*. New Jersey: Prentice Hall.

Mennecke, B. E. (1997). Understanding the role of geographic information technologies in business: Applications and research directions. *Journal of Geographic Information and Decision Analysis, 1*(1), 44-68.

Mirel, B. (1998). Visualization for data exploration and analysis: A critical review of usability research. *Technical Communication, 45*(4), 491-509.

Munzner, T. (2002). Guest editor's introduction: Information visualization. *IEEE Computer Graphics and Applications, 22*(1), 20-21.

Nabney, I. T., Sun, Y., Tino, P., & Kaban, A. (2005). Semi-supervised learning of hierarchical latent trait models for data visualization. *IEEE Transactions on Knowledge and Data Engineering, 17*(3), 384-400.

Post, F. H., Nielson, G. M., & Bonneau, G. (Eds.). (2003). *Data visualization: The state of the art,* Kluwer Academic Publishers.

Rao, R., & Card, S. K. (1994). The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. *Proceedings of CHI'94,* 318-322.

Rhyne, T. (2003). Visualization viewpoints: Does the difference between information and scientific visualization really matter? *IEEE Computer Graphics and Applications,* 6-8.

Roth, S., Chuah, M., Kerpedjiev, S., Kolojejchick, J., & Lucas, P. (1997). Toward an information visualization workspace. *Human-Computer Interaction, 12*(1/2), 131-185.

Shneiderman, B. (2003). Why not make interfaces better than 3D reality? *IEEE Transactions on Computer Graphics and Applications,* 12-15.

Soukup, T., & Davidson, I. (2002). *Visual data mining: Techniques and tools for data visualization and mining,* Wiley.

Speier, C., & Morris, M. G. (2003). The influence of query interface design on decision-making performance. *MIS Quarterly, 27*(3), 397-423.

Wang, Y., Luo, L., Freedman, M. T., & Kung, S. Y. (2000). Probabilistic principal component subspaces: A hierarchical mixture model for data visualization. *IEEE Transactions on Neural Networks, 11*(3), 625-636.

Ware, C. (2004). *Information visualization: Perception for design* (2nd ed.). San Francisco: Morgan Kaufmann.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations,* Morgan Kauffmann.

Yang, J., Ward, M. O., & Rundensteiner, E. A. (2003). Hierarchical exploration of large multivariate data sets. In F. H. Post, G. M. Nielson & G. Bonneau (Eds.), *Data visualization: The state of the art* (pp. 201-212), Kluwer Academic Publishers.

Zhang, M., Zhang, H., Tjandra, D., & Wong, S. T. C. (2004). DBMap: A space-conscious data visualization and knowledge discovery framework for biomedical data warehouse. *IEEE Transactions on Information Technology in Biomedicine, 8*(3), 343-353.