Published in *Computing Science and Statistics (Interface'99 Conference Proceedings).* Fairfax Station, VA: Interface Foundation of North America, Inc., pp. 131-138.

Information Visualization: A Taxonomy of User Tasks

Stacie L. Hibino
Bell Labs, Lucent Technologies
263 Shuman Boulevard
Naperville, IL 60566-7050 USA
hibino@research.bell-labs.com
http://www.bell-labs.com/~hibino/

Abstract

Historically, most work in information visualization has emphasized the design of new views and frameworks to aid users in *exploring or accessing* data. Very little work, however, has been done to support users through the *full analysis process*—the process of starting with raw data, exploring the data in search of trends, and ultimately producing a presentation of final results. In addition, little work has been done to even identify the types of concrete tasks that real users actually conduct when analyzing real data using an information visualization (infoVis) environment. To address this problem, we conducted a task analysis of experts' use of an existing infoVis system. Results indicate that users work on various tasks *outside* of data exploration—tasks such as conditioning and preparing data, collecting results, and gathering evidence for a presentation. This study identifies key data analysis tasks that expert users perform when using an infoVis environment to analyze some real-life data, indicates how users rate the importance of task categories, and presents the typical time that users estimate they spend within each task category. Overall, this paper represents a working taxonomy of user tasks for information visualization.

1. Introduction

Several novel visualizations and new paradigms to support users in accessing and exploring a wide variety of data have been presented to date. For example, general information visualization (infoVis) frameworks [1, 13] and applications to fields such as temporal [8], software [9] and medical [10] data are just a few indicators of the growing work in infoVis. Most of this work, however, has focused on enhancing data access and exploration; little, if any work has examined the larger problem of supporting users' data analysis processes—the processes they use in transforming raw data into a presentation of results based on using infoVis as their key analysis tool.

When using an infoVis system to analyze small data sets, users may only have one or two questions in mind and it is typically not too much work to prepare or transform the small data set, if necessary. As data set sizes increase, however, the data can be more cumbersome to work with and some of the limitations of current infoVis systems may begin to be revealed. In our research lab, where we have already developed a suite of infoVis tools capable of handling moderately sized data sets (easily accommodating data sets containing 100,000 data records), we are beginning to see evidence of some of these challenges as users complain about the effort required to prepare data for exploration or abstract key results from an infoVis analysis of a complex data set. Users need better support for preparing and organizing their full process of analyzing such large complex data sets. Previous work in information workspaces [3, 7] may offer a glimpse of a potential organizing framework, but much of this work focuses on organizing *data and objects* (e.g., documents) rather than a user's *process*.

How can we provide better support to users in their infoVis analysis process? One way to better understand how to provide such support is to gain a better understanding of users' *current* work processes—how they currently use an infoVis system in the context of real-life work. This can be accomplished through a task analysis. This paper reports the results of such an analysis of user tasks. We had three driving questions in our study of real users conducting a real-life data analysis:

- Do users conduct other tasks besides data exploration during data analysis?
- If so, what are these tasks?
- How important are these tasks to the data analysis process?

In our task analysis, we studied five infoVis experts using an existing infoVis environment (EDV: the Exploratory Data Visualizer [13]). The expert users were asked to analyze a disease data set used as part of an American Statistical Association (ASA) Data Exposition. Their goal at the end of their analysis sessions was to present key findings from their analysis in their typical target presentation format.

Overview. This paper is divided into five additional sections. In the next section, I provide a brief background description of EDV, the infoVis environment that was provided to users as their key analysis tool. This is followed by a section describing the experimental method. I then provide a section on results and discussion, present some related work and finally, provide conclusions and present some future work.

2. The Exploratory Data Visualizer

The Exploratory Data Visualizer (EDV) is an information visualization framework which provides linked views [13]. EDV allows users to create and interact with any number of user-specified data views including univariate (e.g., bar charts and histograms), bivariate (e.g., scatterplots and table views), multivariate (e.g., clustering), network (e.g., a nodes and arcs view), and text-based views (e.g., similar to a spreadsheet).

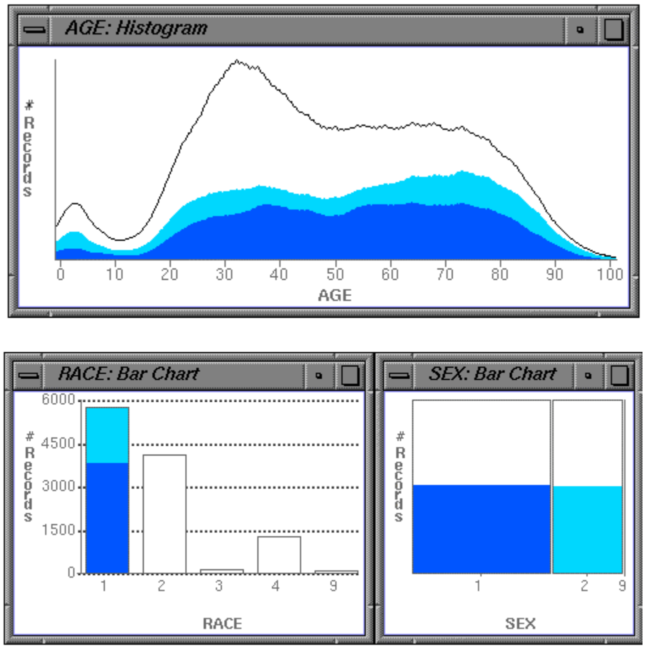

Figure 1. Sample Screen Dump from the EDV Analysis of Some Tuberculosis Data.

EDV empowers users to explore relationships between data variables through linked views. Views are dynamically linked to one another so that when a selection is made in one view, the corresponding selections are automatically highlighted in all other views. For example, Figure 1 presents three linked views from the tuberculosis (TB) data analyzed by users—a histogram of the age, a bar chart of the race, and a spine plot of the sex of all TB cases included in the data set. The figure shows the state of the views after the user has selected Race 1 (Caucasian) of the Race bar chart. Color is used to identify sex (sex 1=male=dark gray, sex 2=female=light gray). The white areas indicate unselected data items while still showing the shape of the curves or bars for the overall population of TB cases. The linked highlighting between the views indicates that Caucasians do not show a peak in the number of TB cases plotted by age, as the rest of the population does.

Another useful characteristic of EDV is its scalability. EDV is capable of both handling moderately large data sets (e.g., it accommodates data sets containing 100,000+ records) and supporting the addition of new views.

3. Experimental Method

The goal of the task analysis was to study expert users analyzing real-life data with EDV as their primary analysis tool—where users started with the raw data, analyzed it, and worked towards a presentation of results. In order to examine how users cope with an analysis task requiring multiple sessions, the users were asked to accomplish this over two separate sessions. This is because when a data set is complex enough, users typically will not be able to complete their analysis in one session; they will have to stop what they are doing and continue at some other time. Due to busy schedules and ongoing projects, they also may not be able to get back to their analysis right away.

3.1 Participants

Five users (all male) participated in the study. They were invited to participate based on their expertise in using EDV as well as their experience in using information visualization (infoVis) environments for data analysis. Subjects had four to thirteen years of experience using infoVis systems for analyzing data. Four of the users typically analyze data sets that are 1 to 5 Meg in size and one user typically analyzes data sets that are 1 Gig in size. The largest data sets that users have analyzed ranges from 3 Meg to 40 Gig.

Two users have PhDs in statistics, while two have a PhD and one has an MS degree in computer science. The statisticians also have a fairly high knowledge of disease data, while the computer scientists have a fairly low knowledge about disease data. However, a fair amount of background information was provided with the test data set and this expertise in disease data seemed to have little impact on the users' findings.

3.2 Materials and Procedure

Data Set. Users analyzed real-life tuberculosis (TB) disease data from the 1991 American Statistical Association (ASA) Data Exposition [5]. The ASA Data Exposition is an annual challenge to ASA members to present innovative graphical and analytic techniques for addressing questions of importance to the particular domain of data provided for a given year. The focus of the 1991 Data Exposition was on public health control and prevention efforts and included two primary data collections—one on TB and one on mumps. For this study, subjects only analyzed the TB data.

The TB data collection included one primary file recording individual TB cases by U.S. state, and additional supporting data files on: the "start" date in month and year that each state started reporting TB cases, 1970 and 1980 census data, and state Federal Information Processing Standard (FIPS) codes. The census files were provided by the U.S. Bureau of the Census and the remaining files were from the Center for Disease Control (CDC).

The primary TB data file consists of 11,338 individual cases of TB sampled randomly from 113,417 cases reported to CDC during 1985-1989. Each record includes 7 fields: U.S. state in FIPS code, year the case was counted by CDC, month the case was counted by CDC, and the patient's age in years, along with the patient's sex, race, and ethnicity.

Procedure. The users' goal was to analyze the TB data and create a presentation of key results. They were given two one-hour sessions, separated in time by at least one week, to accomplish this goal. They used EDV as their primary tool for analyzing the data, and were told that they could use any other tools that they typically use in conjunction with EDV to accomplish the analysis task. At the beginning of the first session, users read a sheet describing this task analysis study and their goals, and a hard copy of the "readme" file included with the TB data set. They were asked to provide think-aloud verbal protocols of what they were doing as they were doing it, and then proceeded with the analysis task at hand. At the end of the second session, if extra time was available, a short informal post-interview was conducted. Information about the post-questionnaire was sent to users after they had completed their second session.

Data Collected. The following data was collected: observational data, files generated during the users' sessions (e.g., scripts written to transform the data, new files of transformed data, etc.), post-interview notes, and results of the post questionnaire. In addition, output of the users' screen was captured directly onto video, along with audio of their verbal protocols.

4. Results and Discussion

During their analysis, all subjects used a scripting or programming language such as Perl or awk to "condition" the data in some way (e.g., to create data aggregations). They were typically able to create and debug such a script in less than five minutes. Although users took different approaches to their analyses, their core set of tasks and findings were very similar.

Unfortunately, none of the users created an actual presentation of their results. Although they analyzed a real-life data set, they were not motivated enough to create a presentation and in most cases did not feel that they had enough time to do so. Instead, they either captured their key findings on paper or articulated them aloud during their analysis. Their reluctance to create a presentation was an indication that they did not consider it to be a trivial task. Information about presentation-related tasks was gathered through informal post-interviews and through the post-questionnaire.

Observations of users' tasks collected during their analysis sessions were used to refine the post-questionnaire so that a more fully representative set of tasks could be presented to the users to rate within the post-questionnaire. The final questionnaire was not given to users until after they had completed both of their analysis sessions.

Most of the questionnaire focused on users' ratings of 44 observed tasks, divided into seven categories of high-level analysis tasks. Additional questions asked users to articulate and clarify: the typical time they spend on the high-level tasks, their typical target presentation, how they cope with multiple analysis sessions, how they organize and keep track of their original versus transformed data, and if they usually work individually or collaboratively in analyzing data.

4.1 High-Level Analysis Tasks

In this user study, we observed and identified several low-level tasks that experts conduct during data analysis with an information visualization environment. When we grouped similar low-level tasks together, seven categories of high-level tasks emerged. These high-level tasks include:

- *prepare:* data background and preparation tasks,
- *plan:* analysis planning and/or strategizing tasks,
- *explore:* data exploration tasks,
- *present:* presentation-related tasks,
- *overlay:* overlay and assessment tasks,
- *re-orient:* re-orientation tasks (when analysis requires more than one session), and
- *other:* statistics-based tasks.

While these high-level tasks do not really differ much from the types of steps taken for general problem solving, we have identified several low-level tasks specific to the domain of data analysis through information visualization. The low-level tasks for each of the task categories are described in Section 4.2. When asked, users did not present an alternative list or any additional categories of high-level tasks.

Users rated the importance of each of the low-level tasks to the analysis process based on a scale of 1 to 5. Figure 2 shows the average importance ratings for each of the high-level task categories, based on an average of all of the low-level task ratings in each category.

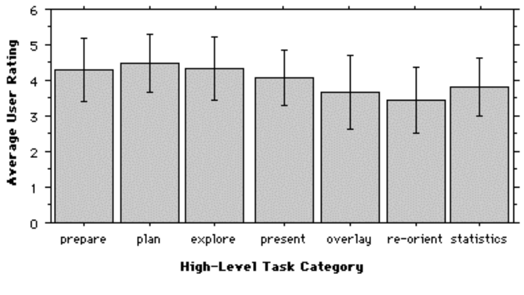

Fig. 2. Average importance ratings for task categories (1=unimportant, 5=very important; error bars indicate standard deviation).

Overall, the ratings are both fairly high and fairly similar, with averages ranging from 3.4 to 4.5 on a 5-point scale. However, taking individual users into account, an analysis of variance indicates that the differences in importance ratings between categories is significant, leading to the following order of importance based on user ratings: plan > explore > prepare > present > statistics > overlay > re-orient tasks.

4.2 Low-Level Analysis Tasks

The low-level tasks were identified through observational data, previous infoVis taxonomies [12], informal interviews and the post-questionnaire. While there is some overlap between individual tasks as well as some overlap between task categories, this section provides a working taxonomy of user tasks conducted during data analysis through information visualization.

4.2.1 Prepare: Data Background and Preparation Tasks

Before users explore data, they usually spend time familiarizing themselves with the available data—identifying what data tables and variables are available and checking the format of the data tables to determine if they can be read or imported directly into an infoVis environment. Before and after exploring some of the data, users may also find that they need to prepare data to address issues of data quality and data conditioning. For example, if a particular variable has the problem of lots of missing data, they might try to generate a replacement variable from other information that has been more completely provided. Data conditioning such as aggregating records or splitting variables is important for generating the necessary variables for particular data analysis questions of interest.

The specific data background and preparation tasks we identified through the task analysis include the following:

- gather background information about data set at hand (content domain of data, analysis goals, etc.),
- understand data sources (e.g., their location, size, contents, format, and how and when data was collected),
- get clarification on data ambiguities (e.g., variable name),
- collect additional data from other external sources,
- reformat data for suitable input,
- check data for potential data errors,
- check for missing data, and
- transform the data (create new variables, split variables, extract subset, rollup/aggregate data).

4.2.2 Plan: Analysis Planning and Strategizing Tasks

Three analysis planning and/or strategizing tasks were identified:

- hypothesize,
- make a strategy or plan for all or a part of your analysis (e.g., decide what, how, and how much to investigate or explore), and
- identify data formats and variables required for desired views.

As indicated below in the discussion of exploration tasks, users may take a top-down or bottom-up approach to exploring the data. In terms of planning, experts may spend time on proposing new or reviewing suggested hypotheses to test when taking a top-down approach; or, they may make a strategy for guiding their exploration in a bottom-up approach. Users conducted such planning tasks both before and interleaved with data exploration tasks. In many cases, we observed users following a plan or systematic process of analysis without having articulated one. For example, some users were observed conducting univariate analysis of all variables, followed by bivariate analysis, followed by multivariate analysis. These users probably did not articulate their strategy aloud since they are *expert* users and have most likely internalized such plans. The important point to note is that they were indeed following some sort of plan or strategy and that identifying the types of plans they use may lead to insights for better supporting *novice* or casual infoVis users.

In addition to planning and strategizing their overall analysis approach, users also planned other aspects of their exploration and analysis, such as identifying data formats and variables required for their desired views (e.g., views intended to answer their hypotheses).

4.2.3 Explore: Data Exploration Tasks

We observed users conducting the following tasks during data exploration:

- get an overview of the data,
- investigate data to test hypotheses (top down approach),
- explore data in search of trends or exceptions (bottom-up approach),
- “query” or filter the database,
- identify curiosities to investigate further,
- zoom in on items of interest,
- remove (filter out) uninteresting items or equivalently focus on (select) items of interest,
- identify data clusters,
- identify relationships between variables,
- explain view/visualization (e.g., why or why not it might look the way it was expected to look),
- identify a trend or exception,
- verify a trend or exception (e.g., through alternative views), and
- drill-down for more details.

This list includes tasks presented in previous infoVis taxonomies (e.g., [12]), as well as other tasks such as identifying curiosities, explaining a visualization, and verifying a trend or exception. These additional tasks indicate that expert users do not just stop at the identification of a trend or exception, and that they consider their analysis to be as much of an *investigation* to find an explanation, as an exploration.

4.2.4 Present: Presentation-Related Tasks

Presentation-related tasks focus on those related to communicating key findings to others. In discussing presentations with the users and through their feedback on the post questionnaire, we found that different users may have very different types of target presentations. Users in this study listed one or two of the following as their typical target presentation: EDV screen dumps, static research paper, PowerPoint presentation, web page with static or active links, web page with active visualization components (e.g., a “Live Document” [6]), presentation through an EDV demonstration, or plots generated by other statistical software.

The following presentation-related tasks were identified as general tasks that users conduct to achieve their target presentation:

- gather evidence to answer a hypothesis or driving question,
- record or keep track of trends and results tested and found,
- articulate importance of a result (rank it or identify it as “interesting”),
- articulate/summarize all and/or key results,

- decide what to include in presentation of results,
- create presentation of results, and
- give presentation of results.

For small data sets and in situations where only a few results are identified, presentations are much easier to create and require fewer tasks (e.g., it's not as necessary to rank results if they are equally important and there are only a few of them). However, users noted that for large, complex data sets, additional work is required to keep track of, rank, and decide on presentation contents of results. For example, one user said that he typically captures screen dumps along the way and that at the end of data exploration, he might have 20 to 30 screen dumps of results. He then prints out all of the screen dumps, lays them out, and ranks their importance. In many cases, he realizes that a trend that is clear to him may not be so obvious to his target audience. Thus, he often needs to go back to the infoVis system to capture new screen dumps that better illustrate the results.

4.2.5 Overlay: Overlay and Assessment Tasks

Users perform several overlay and assessment tasks:

- take notes,
- window management (move and resize windows, etc.),
- assess their strategy (e.g., is this the right strategy to take?),
- assess their observations (e.g., does this observation or conclusion make sense?),
- assess their assumptions about data formats (e.g., is the data in the right format to accomplish this part of the analysis?),
- assess your progress (e.g., have I investigated all that there is to investigate? What else should I look at?), and
- estimate cost-benefit ratio of additional data collection or conditioning.

The "take notes" and "window management" tasks are overlay tasks that cut across the other task categories. For example, users could take notes while preparing data, planning their analysis, or exploring the data. The assessment tasks in this task category identify the metacognitive activity exhibited by users during their analysis sessions. That is, users asked themselves the types of sample questions listed with each task. The final task in this category indicates that users felt there was a definite cost of conditioning the data—doing so definitely required time and effort, so they often tried to weigh the potential benefits of their labor ahead of time.

4.2.6 Re-Orient: Re-Orientation Tasks

When data analysis requires more than one session, users go through several re-orientation tasks in subsequent sessions. Such tasks may include:

- review goal(s),
- review data and formats,
- review notes,
- review progress, and
- identify starting/continuing point for current session.

While these are tasks that users did perform at the beginning of their second session, they were generally rated lower in importance in comparison to tasks in the first few task categories.

4.2.7 Other Tasks

Statistics-Based Tasks. Although users did not conduct any statistical tests during their analysis of the TB data, several users mentioned situations where they either would follow-up with a statistical test or where they thought it would be nice to have system support to conduct a particular statistical test. In general, three of the users indicated that they typically spend some of their analysis time (5 to 10%) conducting statistics-based tests.

Additional Tasks. Three users listed five tasks between them that they felt were not included in the task list presented in the post-questionnaire. Three of the five tasks listed were very similar to existing tasks while the other two included:

- given a relationship or feature of interest, explore what other factors may contribute to it, and
- sort and order results so that related results and their impact on each other can easily be accomplished.

The small number of additional tasks suggested by users indicates that they felt the post-questionnaire was fairly comprehensive in identifying an overall list of low-level tasks.

4.3 Typical Time Spent on Tasks

Figure 3 presents a bar chart of the *typical* percent of time that users estimate they spend on each of the high-level task categories. User u5 is an obvious outlier, attributing more than 60% of his time to data preparation and conditioning tasks. This user was a statistician who listed the largest data set size typically analyzed (1 Gig, in contrast to the other users who focus more on data sets of about 5 Meg in size) as well as one of the users who has analyzed a data set as large as 40 Gig in size. This indicates that the amount of time required for data preparation and conditioning may strongly increase with data set size, whereas the other task categories are not as strongly affected by data set size.

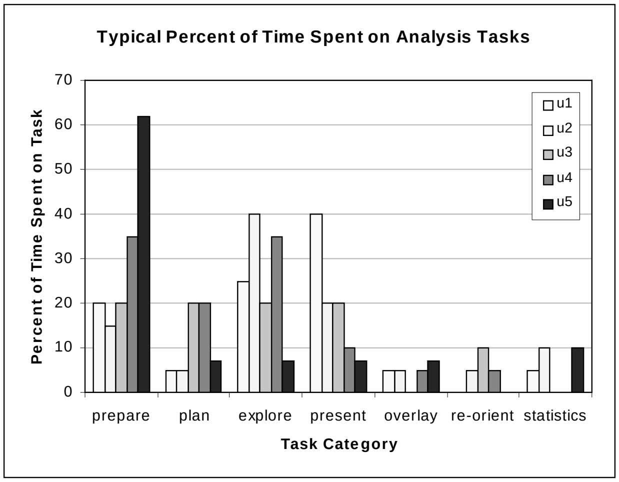

Figure 3. Estimated Time Spent on High-Level Tasks.

Another notable feature of Figure 3 is the amount of time users estimate they spend on data exploration. Users only spend about 25% on average and at most 40% of their analysis time on data exploration. Thus, although users may not be in full agreement on where they spend the rest of their time, they do agree that they spend more than half of their analysis time on tasks *other than data exploration*. This suggests a need for infoVis to support some of these other data analysis tasks—tasks which are 1) time consuming as indicated in Figure 3, and 2) rated with high importance as shown in Figure 2.

5. Related Work

Several taxonomies of information visualization have been proposed (e.g., [12, 2]), but these have typically focused on categorizing aspects of infoVis limited to accessing and exploring data—aspects such as data types, visualization or view types and exploration tasks. The results of the task analysis presented in this paper indicates that such taxonomies only address a part of the problem, especially when considering the use of infoVis for data analysis rather than only data access.

While no infoVis environment currently addresses all of the types of tasks identified through this task analysis, some work has touched on some of the issues identified here. For example, the SAGE system [11] is a knowledge-based presentation system for partially automating the display design for visualizing combinations of diverse information. Such a system potentially reduces the users' task load on analysis planning as well as presentation-related tasks.

An information visualization spreadsheet [4] is an example of a framework that could provide some aid to users in organizing their analysis process. This approach extends the traditional spreadsheet metaphor to infoVis by using visualizations as the atomic units to be placed in the spreadsheet cell rather than numbers or text. Operations such as addition and subtraction between visualization cells are supported, essentially enabling users to perform data transformations in a visual manner. The advantage of an information visualization spreadsheet is that it provides a framework for organizing data conditioning and exploration in rows and columns, and results could be naturally summarized in a final column or row of the spreadsheet. The disadvantage of the spreadsheet metaphor is that it seems to be very limited in terms of scalability. As one might imagine, details in individual spreadsheet cells become increasingly difficult to abstract as the data set size increases.

Information workspaces (e.g., [3, 7]) have typically focused on organizing data rather than processes. However, one can imagine using a rooms [7] or book [3] metaphor for organizing an infoVis analysis. For example, rooms or books could be used as logical separators for the different types of tasks (e.g., data preparation, analysis planning, data exploration, etc.) or they could be used to separate the analysis along themes or threads (e.g., separate rooms could be dedicated to investigations of different hypotheses). The challenge in using either of these metaphors, however, is in understanding how, if, and when process support can be bridged across rooms.

6. Conclusion and Future Work

In this task analysis, we found that users do perform many other tasks beyond data exploration when using an infoVis environment for analyzing data. More specifically, we identified six other categories of analysis tasks besides data exploration: data background and preparation tasks, analysis planning and strategizing tasks, presentation-related tasks, overlay and assessment tasks, re-orientation tasks, and statistics-based tasks. On average, users estimate that only about 25% of the time that they typically spend on analyzing a data set is devoted to data exploration, and thus that over half of their analysis time is spent on other tasks. Moreover, not only do users conduct these other tasks, they also rate them highly in terms of their importance to the analysis process. In particular, they rate planning and strategizing tasks significantly higher than exploration tasks.

These results indicate the need for a more process-oriented framework—one that helps users in organizing their analysis process and supports them in a wider range of analysis tasks, not just data exploration. In our research lab, we are currently designing and developing such an infoVis framework to address these needs. We refer to this framework as *InfoStill*, short for Information Distillery. In the future, we plan to evaluate our InfoStill approach with real users.

Acknowledgments

Special thanks to expert users who participated in the study, to Graham Wills for EDV, and to Beki Grinter and Ken Cox for reviewing earlier drafts of this paper.

References

1. Ahlberg, C., & Shneiderman, B. (1994). Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *CHI'94 Conf. Proc*. ACM Press, 313-317.

2. Card, S. and J. Mackinlay. (1997). The Structure of the Information Visualization Design Space. *IEEE Proceedings of Information Visualization'97*, 92-99.

3. Card, S., Robertson, G., and W. York. (1996). The WebBook and the Web Forager: an information workspace for the World-Wide Web. *CHI '96 Conference Proceedings*, 111-119.

4. Chi, E.H., Riedl, J., Barry, P., and J. Konstan. (1998). Principles for Information Visualization Spreadsheets. *IEEE Computer Graphics & Applications, 18(4)*, 30-38.

5. 1991 ASA Data Exposition, Disease Data. Available at: http://www.stat.cmu.edu/disease/.

6. Eick, S., Mockus, A., Graves, T. and Karr, A. (1998). A Web Laboratory for Software Data Analysis. *World Wide Web Journal*, *12*, 55-60.

7. Henderson, J. & S. Card. (1986). Rooms: The use of multiple virtual workspaces to reduce space contention in window-based graphical user interfaces. *ACM Transactions on Graphics*, *5(3)*, 211-241.

8. Hibino, S. and Rundensteiner, E. (1996). MMVIS: Design and Implementation of a Multimedia Visual Information Seeking Environment. *ACM Multimedia'96 Conf. Proc.* NY:ACM Press, 75-86.

9. Jerding, D.F., Stasko, J.T. and Ball, T. (1997). Visualizing interactions in program executions. *ICSE'97 Conference Proceedings.* NY:ACM Press, 360-370.

10. North, C., Shneiderman, B. and Plaisant, C. (1997). Visual Information Seeking in Digital Image Libraries: The Visible Human Explorer. *Information in Images* (G. Becker, Ed.), Thomson Technology Labs (http://www.thomtech.com/mmedia/tmr97/chap4.htm).

11. Roth, S., Kolojejchick, Mattis, J. and J. Goldstein. Interactive graphic design using automatic presentation knowledge. *CHI'94 Conference Proceedings*, 318-322.

12. Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Proceedings of Visual Languages 1996*, 336-343.

13. Wills, G. (1999). Linking interactive graphics for exploring many types of data. *Proceedings of Interface'99*. To appear.