

Fostering Understanding Through Multivariate Displays

This paper introduces some cognitive tasks that users of multivariate displays must perform in order to process and understand the information being presented. Next, it describes how visualizations can assist those examining multivariate data by reducing their cognitive load and taking advantage of humans' advanced visual processing capabilities. This paper closes by analyzing a real-world problem related to the author's current work, and provides some examples of multivariate displays that may further viewers' understanding of complex information.

Cognitive Requirements of Multivariate Displays

For a viewer to process a multivariate display, they must be able to perform certain cognitive tasks. These tasks are directed by the viewer's goals and purposes for reviewing the display. The goals discussed in this section are data exploration and the identification and comparison of data values.

Data Exploration

For large data sets with multiple variables, it is highly likely that a viewer's goal is to explore the data in search of patterns. Regrettably, it is equally likely that such patterns (meanings) cannot be perceived in what Ware describes as "masses of mostly meaningless numbers" (pp. 145). To help make sense of this meaningless mass of numbers, visualizations must present the underlying structure of complex data in a way that allows viewers to "look at data to see what it seems to say" (Young, et. al., pp. 2-3) and to "assimilate" that structure into their working conceptual model. Viewers may also use multivariate displays to generate hypotheses about the data, and subsequently re-utilize the display to test those hypotheses (Keim, pp. 40; Young, et. al., pp. 2).

According to Keim, visual data exploration is a three-step process whereby viewers:

- Take an overview of a display to "identify interesting patterns" such as clusters, correlations, dependencies, and exceptions (pp. 40, 43).

-
- Filter out other information in order to focus on a pattern (pp. 40).
 - Analyze the pattern by drilling down to detailed data (pp. 40).

Identifying patterns requires the use of sensory processes like feature detection, as well as pre-attentive perceptual processes such as organization (based on the Gestalt principals of proximity, similarity, and so on). It also requires the use of “cognitive processes including attention, reading processes, judgemental and inferential processes, arithmetic operations,” and so on (Gillan, et. al., pp. 380). Interestingly, Veluchkovsky et. al. explains that the pre-attentive processing stage of any visual search task “helps viewers locate objects in the visual world” while in contrast, the attentive processing stage “operates on a few objects at a time” and is thus “the bottleneck in visual processing” (pp. 79). Therefore, it makes sense to design multivariate displays that maximize viewers’ pre-attentive processing capabilities when they are attempting to identify patterns, but that still allows them to shift into the attentive processing stage when they are ready to focus on a particular pattern. Information filtering—a task required just prior to focusing on a particular pattern—is another challenging cognitive task that has already been discussed in “Using Visualization to Enable Decision-Makers” (Hocko). When viewers finally “drill down” to the details of the data, they are likely to have subgoals such as identifying and comparing displayed information.

Identification and Comparison of Data Values

As part of Keim’s analysis step, a viewer of a multivariate display might have more specific goals, such as identifying the actual values associated with a variable (and *visa versa*), or of comparing the values between different variables (Gillan, et. al., pp. 375). Any comparison among objects implies that human judgement is required, but does not necessarily mean that the judgement must be absolute. In fact, a “similarity judgement” may be acceptable in many cases (Wilkinson, pp. 207). Consequently, an effective multivariate display must support both the type of task (identification or comparison) and also take into consideration the acceptable level of accuracy. Gillan, et. al. showed that a viewer’s time and accuracy with identification and comparison tasks are influenced by a number of factors, including:

- Perceptual and informational complexity—for example, the number of data dimensions a viewer must process. Wilkinson also points out that for some types of displays, “the user can be overwhelmed with information if the number of objects or dimensionality of the variable space is extremely large” (pp. 208).
- Figure-to-axes relation—whether the axes are required to interpret the data.

- Physical elements of the display—points, lines, areas, and so on.
- Data-ink ratio—the “proportion of a graphic’s ink devoted to the non-redundant display of data-information” (Tufte, pp. 93).
- Data density—“the ratio of the number of data points and the area of the graphic” (Tufte, pp. 162).

The study conducted by Gillan, et. al. also revealed that while informational complexity had “an especially strong effect on performance with identification questions,” the physical elements of the display and data-ink ratio “appear[ed] to have a greater effect with comparison questions.” Since the study also showed that comparisons are typically made by viewers prior to identifying specific values, any multivariate display would have to support both comparison and identification tasks to be truly effective (pp. 380).

Exploiting Humans’ Perceptual Capabilities

Keim aptly states: “Lacking the ability to adequately explore the large amounts [of data] being collected, and despite its potential usefulness, the data becomes useless and the databases data dumps” (pp. 39). We have already acknowledged that an effective visualization of multivariate data should support the cognitive tasks involved in data exploration, and consequently, in data identification and data comparison. Many types of graphical displays have been shown to be effective “instrument[s] for reasoning about qualitative data” (Crouch, pp. 58). The question is, why do they work?

As with any visual representation of complex concepts, graphics like multivariate displays reduce the amount of cognitive effort a viewer must exert to explore, identify, and/or compare because they externalize these activities and free up working memory (Keim, pp. 39). Cognitive effort is also reduced because such visualizations “...capitalize on the feature integration abilities of the human visual system, particularly at the higher levels of cognitive processing” (Wilkinson, pp. 207). Similarly, a human’s ability to recognize patterns in visual representations is a pre-attentive process that aids viewers in quickly discovering the structure of the data, from which they can form their hypotheses (Young, et. al., pp. 3). It has been shown that reliance on the visual system for data exploration results in faster exploration, and “better” (we can assume “more accurate,” based on Gillan, et.al.) results. When viewers used visual displays of complex information, their level of confidence in their

findings proved to be much higher than those who used numerical or textual representations of the same data, so there is a psychological benefit as well (Keim, pp. 40).

The primary issue regarding multivariate displays is not about whether they work or how they work; it is how designers of these visualizations can resolve the discrepancies among the different dimensions in which they must work. Although there are many visualization techniques available for 2-dimensional (and even 3-dimensional) data (one need only to review Harris to see this), illustrating larger, more complex multi-dimensional data sets prove more difficult (Keim, pp. 40). Young, et. al. succinctly summarizes this problem as one of determining “...how to present hD [high-dimensional] information in a 2D plane, such that our 3D perception can understand [it]” (pp. 3). There are presently a number of visualization techniques that may be used to reduce this problem, but a discussion of these techniques is out of the scope of this paper. For more information, the interested reader may wish to review some of the papers cited in “References” on page 8.

Case Study

The documentation team, which is distributed across the country, relies on a bug reporting and tracking system called WebClarify. Although WebClarify allows users to enter a great deal of information regarding an issue and is capable of producing customized reports based on searches of 40+ variables, the list of results is often lengthy and requires too much time to process.

Description

On our current project, the documentation team is behind schedule and the release date for the next major version of our flagship product is extremely close. Therefore, we must prioritize and attempt to resolve as many documentation bugs as possible before this deadline. (Documentation bugs are not necessarily errors; bugs can also mean that required information has not yet been incorporated into a manual.) Typically, a team lead “triages” the bugs and communicates the results of that process to the rest of the team. Owners can then reassign the values of certain variables associated with their bugs. For example, owners may reassign the Priority variable from 1 to 2 and include a note indicating the reasons. When this occurs, the owner and all involved parties are notified via e-mail.

Out of the 40+ variables for each issue, there are typically only three variables of primary concern to those reviewing documentation bugs. These variables are:

- **Condition**—the current status of the documentation bug. Table 1 describes the possible values for the Condition variable.

Table 1 Condition

Value	Description
Open	The bug is currently open.
Fixed	The bug has been fixed, but the fix has not been confirmed by QA.
Closed	The bug has been fixed and confirmed by QA.

- **Priority**—the priority assigned to the documentation bug that, together with the severity, determines when the bug should be fixed. Table 2 describes the possible values for the Priority variable.

Table 2 Priority

Value	Description
1	Must patch
2	Must next release
3	Prefer next release
4	Review for later
5	Informational

- **Severity**—the severity assigned to the documentation bug that, together with the priority, determines when the bug should be fixed. Table 3 describes the possible values for the Severity variable.

Table 3 Severity

Value	Description
1	System down

Table 3 Severity

Value	Description
2	Unstable/limited use
3	General/Suggestion
5	None assigned

A search of the bug reporting system conducted on March 10, 2002 resulted in a total of 242 documentation bugs of varying Condition, Priority, and Severity.

Audience and Objectives

Young, et. al. indicates that viewers of multivariate displays should be “provided with an environment designed to maximize data analysis productivity and satisfaction,” and that such an environment “should reflect the sophistication of users’ data analysis knowledge (pp. 25). While some technical writers working on the product may be quite capable of reading highly complex visualizations, it is likely that many more are novice viewers. In fact, even if everyone on the team were experts, this author would still advocate the use of simplistic visualizations because our objective in creating the visualization is to clearly see the relationships among 3 variables. Including more information in a graphical display would only reintroduce the problem we are attempting to eliminate.

The primary questions our multivariate displays should answer include:

1. How many documentation bugs of Priority 1/Severity 1, Priority 1/Severity 2, and Priority 2/Severity 2 are still listed as Open? (Bugs of Priority 2/Severity 1 for all practical purposes do not exist.) How many are listed as Fixed instead of Closed? (Bugs of condition Fixed sometimes come back as Open.) The answers to these questions helps us determine how much work there is to do before our general availability (GA) deadline.
2. How many Priority 1/Severity 1, Priority 1/Severity 2, and Priority 2/Severity 2 bugs have we Fixed or Closed out of the total number of documentation bugs? The answer to this question helps us to determine our rate of progress and allows us to show others in the organization what we have already accomplished.

3. How many documentation bugs not of Priority 1/Severity 1, Priority 1/Severity 2, and Priority 2/Severity 2 are Open, and to what Priority and Severity do they belong? The answers to these questions helps us determine what to expect once all the most important bugs are resolved. In other words, what will we need to do post-GA?

The Clarify Excel Workbook submitted with this paper contains three worksheets, each of which correspond to the multivariate displays created to answer Questions 1-3. The display for Question 1 illustrates that there are only 10 documentation bugs we absolutely *must* fix prior to GA. The display for Question 2 shows that all the P1/S1 documentation bugs have been addressed, but that we still have a few to go for P1/S2 and P2/S2 bugs. The display for Question 3 (interestingly enough) reveals that many P2 issues have no Severity assigned to them, and therefore, the 51 bugs should probably be categorized more carefully.

Rationale

A number of sources indicate that while no one visualization will work for everyone, multivariate displays such as the ones included here should be designed with both the viewers' task and the characteristics of the data in mind (Keim, pp. 43; Young, et. al., pp 5). Tasks may include "clustering, classification, associations, and multivariate hot spots, while data characteristics may include "data types, number of dimensions, number of data items, category," and so on (Keim, pp. 42-43). Even more fundamental guidelines call for displays that allow for "fast learning and good recall," have "limited visual overlap" and other "occlusions and line crossings that might appear to the viewer as an artifact limiting the usefulness of the visualization technique" (Keim, pp. 43). The displays included here also attempt to take into account Tufte's advice on gridlines, data ink, and so on (pp. 123-137), but in some cases, the software tool used to create the displays constricted the design. (For example, the colors in the pie chart may have been changed.)

Since there were only 3 variables of interest, color was used as an coding mechanism and allowed for the plotting of the 3-dimensional data on a 2-dimensional page. Wegenkittl, et. al. acknowledges that color coding is a frequently used technique that many viewers are comfortable with, and that it allows for "easy calculation and interpretation" (pp. 119-120). Although he also points out some disadvantages, most of those (perhaps with the exception of color-blindness) are irrelevant to the audience of these displays. Ware frequently describes how colors can be pre-attentively processed (pp. 124, 146, 165-166, 194), which also supports the decision to use color as long as the colors "are distinct" enough "so they will not be confused (pp. 115).

In sum, the multivariate displays included here meet the goals of the intended audience, adhere to guidelines for visualizations, leverage the viewers' pre-attentive visual processing system, and allow for data exploration, identification, and comparisons while efficiently utilizing 2-dimensional space for data of higher dimensions.

Note: Although in some cases it would have been possible to create more appealing visualizations using Microsoft Excel's 3-D chart function, this author noticed that doing so skewed the graph in a way that made the identification of exact values difficult to determine.

References

- Crouch, D.B. (1986). The Visual Display of Information in an Information Retrieval Environment. In Proceedings of the 1986 ACM Conference on Research and Development in Information Retrieval, pp. 58-67. New York, NY: ACM Press.
- Gillan, D.J., Lewis, R., and Rudisill, M. (1989). Models of User Interactions with Graphical Interfaces: 1. Statistical Graphs. In Proceedings of the SIGCHI Conference on Wings for the Mind, pp. 375-380. New York, NY: ACM Press.
- Harris, R.L. (1996). Information Graphics: A Comprehensive Illustrated Reference, pp. 37-52. Atlanta, GA: Management Graphics.
- Hocko, J. (2002). Using Visualization to Enable Decision-Makers. Information Visualization Course Assignment 4, Bentley College.
- Keim, D.A. (2001). Visual Exploration of Large Data Sets. Communications of the ACM,(44)8, pp. 39-44. New York, NY: ACM Press.
- Tufte, E.R. (1983). The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press.
- Velichkovsky, B.M., Dornhoefer, S.M., Pannasch, S., and Unema, P.J.A. (2000). In Proceedings of Eye Tracking Research & Applications Symposium 2000, pp. 79-85. New York, NY: ACM Press.
- Ware, C. (2000). Information Visualization: Perception for Design. San Francisco, CA: Morgan Kaufmann Publishers.

Wegenkittl, R., Loffelmann, H., and Groller, E. (1997). Visualizing the Behavior of Higher Dimensional Dynamical Systems, pp. 119-125. Vienna, Austria: Institute of Computer Graphics, Vienna University of Technology, Karlsplatz.

Wilkinson, L. (1982). An Experimental Evaluation of Multivariate Graphical Point Representations. In Proceedings of the First Major Conference on Human Factors and Computer Systems, pp. 202-209. New York, NY: ACM Press.

Young, F.W., Faldowski, R.A., and McFarlane, M.M. (1993). Multivariate Statistical Visualization, pp. 1-30. Originally published in Rao, C.R. (Ed.) *Computational Statistics. Handbook of Statistics, Vol. 9*, pp. 959-998. Amsterdam: Elsevier Science.

