An Empirical Evaluation of Chernoff Faces, Star Glyphs, and Spatial Visualizations for Binary Data

Michael D. Lee

Department of Psychology
University of Adelaide
Adelaide, SA, 5005
Email: michael.lee@adelaide.edu.au

Rachel E. Reilly

Department of Psychology
University of Melbourne
Melbourne, VIC, 3010
Email: r.reilly1@pgrad.unimelb.edu.au

Marcus A. Butavicius

Land Operations Division
Defence Science and Technology Organisation
PO Box 1500, Salisbury, SA 5108
Email: marcus.butavicius@dsto.defence.gov.au

Abstract

Data visualization has the potential to assist humans in analyzing and comprehending large volumes of data, and to detect patterns, clusters and outliers that are not obvious using non-graphical forms of presentation. For this reason, data visualizations have an important role to play in a diverse range of applied problems, including data exploration and mining, information retrieval, and intelligence analysis. Unfortunately, while a variety of different approaches are available for data visualization, there have been few rigorous evaluations of their effectiveness. This paper presents the results of a controlled experiment comparing the ability four different visualization approaches to help people answer meaningful questions for binary data sets. Two of these visualizations, Chernoff faces and star glyphs, represent objects using simple icon-like displays. The other two visualizations use a spatial arrangement of the objects, based on a model of human mental representation, where more similar objects are placed nearer each other. One of these spatial displays uses a common features model of similarity, while the other uses a distinctive features model. It is found that both glyph visualizations lead to slow, inaccurate answers being given with low confidence, while the faster and more confident answers for spatial visualizations are only accurate when the common features similarity model is used.

Keywords: Data Visualization, Empirical Evaluation, Chernoff Faces, Star Glyphs, Spatial Visualizations, Cognitive Models.

1 Introduction

1.1 Data Visualization

Data visualization techniques aim to present data to people in ways that accurately communicate information, and require minimal effort for comprehension. Good data visualizations can facilitate the efficient examination of large volumes of data, and provide the insight that allows inferences to be made from

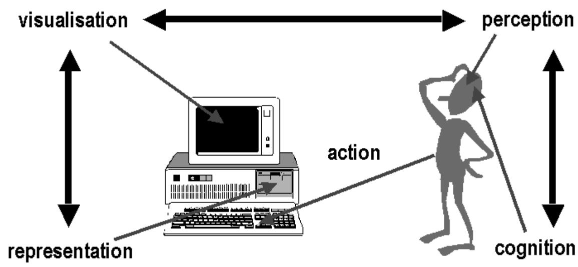


Figure 1: A psychological framework for data visualization. Based on Lee and Vickers (1998, Figure 1).

the observed relationships within the data. Because of this potential, visualizations are commonly applied to problems of data mining and exploration, information retrieval, and the analysis of tactical and strategic intelligence.

It has been often been argued that a principled psychological approach to data visualization is warranted (Chernoff 1973, Purchase 1998, Schneiderman 1998, Ware 2000). Usually the emphasis is on using perceptual principles to design data displays. It is certainly true that, in order to achieve accuracy and efficiency in comprehension, and avoid distortion of the information, visualizations must be designed to be compatible with human perceptual systems. With a few notable exceptions (Kosslyn 1994, Lokuge, Gilbert & Richards 1996), what is less often acknowledged is the role of more abstract cognitive representational principles, not directly related to perceptual processes, in developing data visualizations. To allow for effective analysis and manipulation of data, the structure of the information conveyed also needs to be compatible with the representational requirements and preferences of human cognitive processes.

1.2 A Psychological Framework

A psychological framework for data visualization (Lee & Vickers 1998) that incorporates both perceptual and cognitive components is shown in Figure 1. The motivation for this framework comes from viewing data visualizations as a 'channel' that links information held in an artificial system with human cognitive processes. To the extent that there is representational compatibility between the artificial system and human cognition, and perceptual compatibility between the visualization and human perception, an effective means of conveying information between the two systems may be established. In particular, information represented in the artificial system may be displayed using the data visualization, perceived by the human, and represented mentally. The process of the human then seeking useful patterns and structures in the visualization involves, in effect, subjecting the information to the type of inferential cognitive processes that are difficult to implement in artificial systems. Within the framework shown in Figure 1 there is also the possibility for the human to interact with the information by taking actions that manipulate the data visualization.

On the basis of this psychological framework, Lee and Vickers suggested data visualization techniques that perform little or no manipulation of the data before attempting to represent it graphically, with the intention of 'letting the data speak for itself', may be prone to error in comprehension and manipulation. On the other hand, they argued, those visualizations that restructure the information according to cognitive demands before representing it visually may communicate the information in the raw data more effectively. The primary aim of this paper is to provide a first empirical test of this idea.

2 Evaluating Data Visualizations

As Meyer points out, there are no generally accepted guidelines for the optimal display of data (Meyer 2000, p. 1480). Part of the problem lies in the lack of empirical evidence for or against the use of different approaches to visualization. Despite the important role that visualizations play in information interfaces, it has been noted that the evaluation of data visualizations is rarely undertaken (Morse, Lewis & Olsen 2000). Even where evaluations have been attempted, they often have adopted one of two approaches that does not assess the data visualization in a direct and general way.

2.1 Aesthetic Evaluation

The first of these approaches is to evaluate data visualizations according to their aesthetic appeal or computational efficiency rather than their ability to maximize human performance. Previous research suggests that the relationship between the aesthetic qualities of an interface and performance is complicated (Purchase 1998). While humans tend to associate aesthetic qualities of systems with perceived usability, this perceived usability may be independent of actual usability (Tractinsky, Katz & Ikar 2000). Indeed, it has been found that data visualizations judged as most symmetrical by participants may be associated with the highest rate of error (Purchase 2000). Certainly, assessing data visualizations through measures of their aesthetics does not provide a direct measure of their ability to facilitate human performance.

2.2 Cognitive Engineering Evaluation

A second approach to evaluation has focused on the assessment of domain specific visualizations (Graham, Kennedy & Benyon 2000, Trafton, Kirschenbaum, Tsui, Miyamoto, Ballas & Raymond 2000). This approach reflects the influence of the principles of cognitive engineering, where the work system (computer tools and user) is believed to be so tightly coupled to its domain that it does not make sense to evaluate performance on a visualization independent of a specific subject area (Dowell & Long 1998). While this approach is appropriate for answering specific applied problems, the degree to which the results can be generalized to other applied areas is open to some question, in the sense that assumptions have to be made about the relationship between two specific domains.

2.3 Empirical Performance Evaluation

A more direct and general approach to evaluation has been employed within experimental cognitive psychology (Haskell & Wickens 1993, Jones, Wickens & Deutsch 1990, Liu 1989), where different data visualizations are tested by asking people to answer meaningful questions using the visualizations, and comparing performance measures such as accuracy, confidence, and the time taken to provide answers. This provides a direct and thorough test of how well the visualizations facilitate human performance on the data set used to generate the visualization. By then considering a range of data sets, it is also possible to assess the generality of the results obtained, and provide an evaluation of the data visualization methods themselves.

We adopt this methodology to evaluate two competing data visualization approaches best described as glyph visualizations and spatial visualizations. Glyph visualizations require minimal pre-processing of data, and are archetypal examples of those visualizations that claim to 'let the data speak for itself'. Spatial visualizations, in contrast, impose a cognitive structure on the data before it is represented visually by using a model of human mental representation. In our evaluations, we provide only basic instructions for using the visualizations, and do not provide any training. As a first attempt at evaluating the visualizations in a general way that is not domain specific, we think this is a reasonable approach. The issue of how training might influence people's ability to use the different visualizations in specific domains is addressed in the Discussion.

3 Four Visualizations of Binary Data

3.1 Binary Data

The data used in this study are binary, with objects being defined in terms of the presence or absence of a set of properties or features. While this is clearly a restriction, binary data are an important special case for a number of reasons. There are important properties or features that only exist in binary form, such as gender. There are also many occasions when a variable of interest is a binary quantization of an underlying continuous variable. For example, the distinction between 'virgin' and 'non-virgin' uses a cut-off point to define a binary variable over the countably infinite variable 'number of sexual encounters'.

There are applied data visualization systems that use only binary data. Sometimes, this is a matter of necessity, because the underlying data are inherently binary. For example, most applications of Netmap visualization software (NetMap 2001), which is designed to assist in strategic investigation of fraud, criminal activity, and the like, involves the graphical display of binary information. The raw data give relational information such as whether or not two people are known to each other, whether or not a person is associated with a phone number, whether or not a car is involved in an insurance claim, and so on. In the same way, the visualization of social networks (Wasserman & Faust 1994) usually deals with binary relational information between individuals. On the other hand, there are applied systems where more detailed information is available, but binary data is used as a matter of convenience or scalability. For example, the “Galaxies” software product developed under the SPIRE project, which produces visualizations of text document corpora, represents documents in terms of the presence or absence of a set of 200,000 words (Wise 1999, p. 1226).

Despite these theoretical and practical reasons for studying binary data, however, we acknowledge that there are many variables of interest in the context of data visualization that are not amenable to a binary characterization, and so this study constitutes only a first step towards evaluating the competing visualization approaches that are considered.

3.2 Glyph Visualizations

Glyphs provide a means of displaying items of multivariate data by representing individual units of observation (objects or cases) as icon-like graphical objects, where values of variables are assigned to specific features or dimensions of the objects. The overall appearance of the objects changes as a function of the configuration of values, giving the object a visual identity that can be identified by the observer. It has been argued that examining glyphs may help to uncover specific clusters of both simple relations and interactions between variables (Chernoff 1973) .

The visual possibilities of glyph formations are endless. One commonly used glyph form is the ‘whisker plot,’ where each variable is represented by a line segment radiating from a central point. The length of the line segment indicates the value of the corresponding variable. A variation of the whisker plot, used in this study, is the ‘star plot’. The star is the same as the whisker except that the ends of adjacent line segments are joined.

A second interesting form of glyph visualization, known as ‘Chernoff Faces’ (Chernoff 1973, Chernoff & Rizvi 1975), display data using cartoon faces by relating different variables to different facial features. Chernoff Faces were developed using the idea that, since they use the perceptual characteristics of real faces, they may be particularly easy for people to use given our heightened sensitivity to facial structure and expression. It has also been argued that the faces allow people to perceive many data values in parallel, in the same way they perceive real facial features, and that this holistic perception facilitates the efficient recognition of relationships or patterns among elements (Jacob, Egeth & Bevon 1976, Ware 2000). It has even been suggested that, because Chernoff faces are more interesting representations than many other graphical techniques, they may be more effective because observers are willing to spend more time analyzing the representations (Everitt & Dunn 1991).

3.3 Spatial Visualizations

Spatial visualizations represent objects as points in a multi-dimensional (usually two-dimensional) space, so that objects that are more similar are located nearer each other. This form of representation has some considerable status as a model of human mental representation (Shepard 1957, Shepard 1987, Shepard 1994), and is used to represent stimuli in various formal psychological models of identification, categorization, selective attention, and other cognitive processes (Getty, Swets, Swets & Green 1979, Kruschke 1992, Nosofsky 1986). The algorithms that generate spatial representations, generically known as multidimensional scaling algorithms (Kruskal 1964, Shepard 1980), have also been applied to data visualization, exploration and analysis (Lowe & Tipping 1996, Mao & Jain 1995).

Multidimensional scaling algorithms require as input measures of the similarity between each pair of objects in the domain of interest. Starting from binary data, where objects are represented in terms of the presence or absence of a set of properties or features, there are a number of ways in which similarity could plausibly be measured. Cox and Cox provide a list of a dozen straight-forward approaches (Cox & Cox 1994, p. 11) , and there are other more sophisticated measures (Cohen 1997, Tenenbaum, de Silva & Langford 2000) that could be appended to this list.

This study is restricted to considering the two theoretical extremes for assessing similarity from binary properties. Under the ‘common’ approach, the similarity of two objects is calculated as the number of features or properties they have in common. Under the ‘distinctive’ approach, similarity is calculated as the number of features or properties the objects either both have, or both do not have. Cox and Cox refer to these alternatives as the ‘Matching’ and ‘Jaccard’ coefficients, respectively. Our terminology is taken from Tversky’s seminal psychological theory of feature-based stimulus similarity (Tversky 1977), where the terms ‘common’ and ‘distinctive’ are used to describe exactly the same measures.

We use MDS solutions that represented the data set in two dimensions. Primarily, this choice was based on the fact that, in applied settings, analysts tend to work with inherently two-dimensional media for displaying data representations (e.g., computer screens, sheets of paper, white boards). Any attempt to display three-dimensional (or higher-dimensional) spatial representations using two physical dimensions inevitably involves distorting the MDS representation. Empirical support for avoiding this sort of distortion is found in a recent study (Westerman & Cribben 2000), which compared information search performance on two- and three-dimensional MDS based visualizations. While the amount of variance that can be accounted for by a three-dimensional solution is greater than for a two-dimensional solution, they found that it did not offset the poorer performance associated with three-dimensional versions.

4 Experiment

4.1 Data Sets

Four different binary data sets were constructed to test the visualization types. These related to co-starring movie actors, movie genres, countries and their produce, and animals. In essence, each data set consisted of a set of stimuli and a set of features, with each stimulus being defined in terms of the presence or absence of each of the features. Table 1 shows the animals data set as a concrete example. Rows represents animals, and columns represent animal features. Each cell contains a ‘1’ if the corresponding animal has the corresponding feature, and a ‘0’ otherwise.

4.2 Questions

The areas chosen for the data sets were selected on the basis that they could be expressed as binary data, and could be understood without needing any special knowledge. This familiarity is important, because it allowed the development of questions to which there were ‘clear’ answers, without relying on the data set itself. For example, any reasonable definition of the animals ‘housefly’, ‘bear’ and ‘flea’ should have the housefly and flea being more similar to each other than either is to the bear. This means that any effective visualization should allow people to answer the

Table 1: The animals data set, showing the definition of 20 animals in terms of 14 binary features.

Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Breathes	Venomous	Fins	Tail	Domestic
Clam	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Crab	0	0	1	0	0	1	1	0	0	0	0	0	0	0
Catfish	0	0	1	0	0	1	1	1	1	0	0	1	1	0
Carp	0	0	1	0	0	1	0	1	1	0	0	1	1	1
Haddock	0	0	1	0	0	1	0	1	1	0	0	1	1	0
Honeybee	1	0	1	0	1	0	0	0	0	1	1	0	0	1
Housefly	1	0	1	0	1	0	0	0	0	1	0	0	0	0
Flea	0	0	1	0	0	0	0	0	0	1	0	0	0	0
Lark	0	1	1	0	1	0	0	0	1	1	0	0	1	0
Parakeet	0	1	1	0	1	0	0	0	1	1	0	0	1	1
Bear	1	0	0	1	0	0	1	1	1	1	0	0	0	0
Boar	1	0	0	1	0	0	1	1	1	1	0	0	1	0
Elephant	1	0	0	1	0	0	0	1	1	1	0	0	1	0
Giraffe	1	0	0	1	0	0	0	1	1	1	0	0	1	0
Leopard	1	0	0	1	0	0	1	1	1	1	0	0	1	0
Lion	1	0	0	1	0	0	1	1	1	1	0	0	1	0
Goat	1	0	0	1	0	0	0	1	1	1	0	0	1	1
Seal	1	0	0	1	0	1	1	1	1	1	0	1	0	0
Toad	0	0	1	0	0	1	0	1	1	1	0	0	0	0
Tuna	0	0	1	0	0	1	1	1	1	0	0	1	1	0

question "Of a housefly and bear, which is most similar to flea?". By constructing questions in this general way, a potential circularity is circumvented, because questions do not need to be developed from the visualizations they are used to assess. We should acknowledge, however, that for a small number of question relating to the definition of a cluster, there was some ambiguity regarding what constituted a correct answer. In these cases, several different answers, corresponding to clusters that both did and did not include the problematic object, were scored as correct

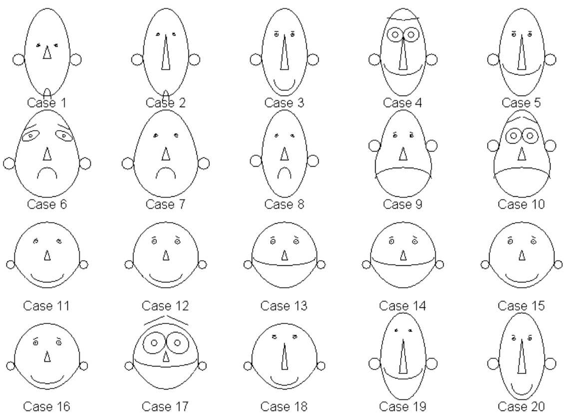
A set of eight questions was designed for each data set. Rather than attempting to adhere to a detailed taxonomy of question types (Wehrend & Lewis 1990), two fundamental question classes were identified. These were named local and global question classes. *Local* questions required the consideration of only a few specific cases out of the set. These questions took the form of a forced choice comparison by asking the participant to assess specific cases in terms of their relationship to other specific cases. An example of a local question is: "Of 6 and 7, which case is most similar to 3?". *Global* questions, in contrast, required the consideration of the entire set of cases to ensure a correct answer. Global questions included those that asked for an outlier to be identified, as in: "Which case is the least like all the others?", and questions that required clusters to be identified, as in: "Which countries produce similar products to case 2?". As it turned out, each of the question sets involved more local than global questions, solely because it proved much easier to generate local questions with clear answers.

Again using the animals data set as a concrete example, Table 2 lists the eight questions asked, together with the correct answers. It is important to understand that the 'decoding' of case numbers into animal names in square brackets was not provided for participants, but is shown in Table 2 as an annotation to assist in interpretation. Questions 4 and 8 were classed as global questions, since they require the identification of a cluster. The other questions were classed as local questions, since only the animals referred to in the question need to be considered to provide an answer.

Table 2: Annotated versions of the questions for the animals data set, with answers shown in italics.

Q1. Of cases 3 [catfish] and 19 [toad], which is most similar to 18 [seal]? *19 [toad].*
Q2. Do 1 [clam] and 2 [crab] share any of the same physical features? *Yes.*
Q3. Of cases 7 [housefly] and 12 [bear], which is most similar to 8 [flea]? *7 [housefly].*
Q4. Case 4 [carp] is an aquatic animal. Name the others. *1 [clam], 2 [crab], 3 [catfish], 5 [haddock], 18 [seal], 19 [toad], 20 [tuna].*
Q5. Do cases 9 [lark] and 10 [parakeet] have features in common that set them apart from the others? *Yes.*
Q6. Do cases 16 [lion] and 17 [goat] share features in common? *Yes.*
Q7. Of cases 3 [catfish] and 13 [elephant], which is most similar to 15 [leopard]? *13 [elephant].*
Q8. Case 14 [giraffe] is a land animal. Name the others. *11 [bear], 12 [boar],13 [elephant], 15 [leopard], 16 [lion], 17 [goat].*

Figure 2: The faces visualization of the animals data set.



4.3 Visualizations and Instructions

The glyph visualizations were generated using Statistica software (Release 5, 1997 edition), with default settings. All features were present on the Chernoff faces whether or not the corresponding feature was present in the raw data. This means that the presence or absence of a feature in the data was represented not by the presence or absence of, say, a mouth, but by extremes in its length or curvature corresponding the software's default values. Figure 2 shows the face visualization of the animals data set. The instructions given to participants using this visualization were: "The following visualization represents a set of twenty animals that each possess one or more of a selection of physical features. Each face (or case) represents a separate animal and each facial feature represents a different physical feature. Two cases that share a particular feature will share the same corresponding facial feature".

For the star glyphs, the presence or absence of each branch was determined by the corresponding data feature, but the length of every branch was the same.

Figure 3: The stars visualization of the animals data set.

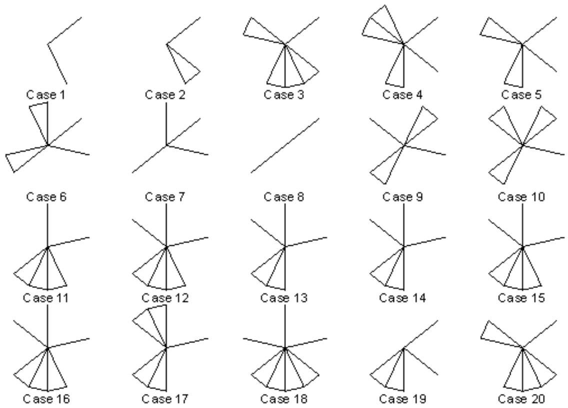


Figure 4: The distinctive spatial visualization of the animals data set.

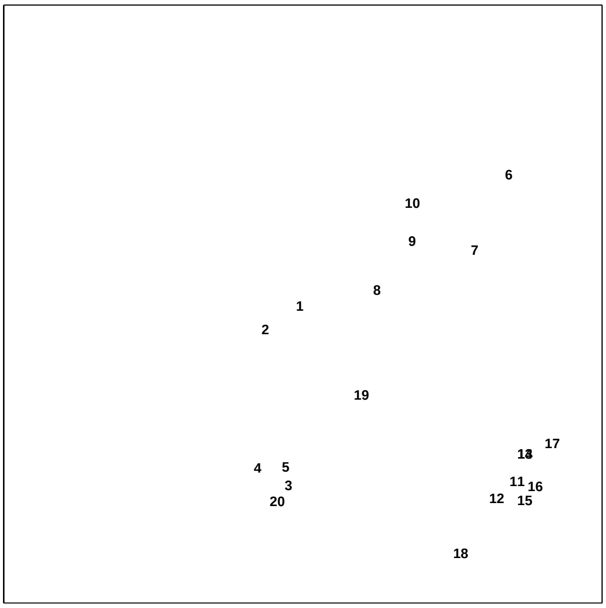
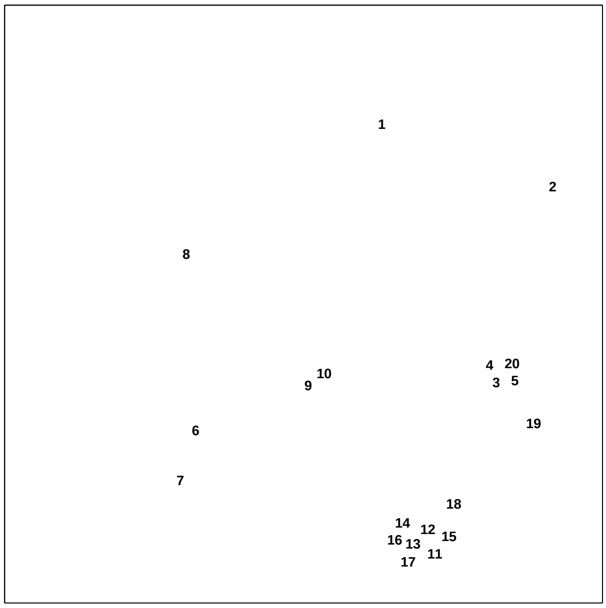


Figure 3 shows the star glyph visualization of the animals data set. The instructions given to participants using this visualization were: "The following glyph visualization represents a set of twenty animals that each has one or more of a set of physical features. Each star (or case) represents a separate animal and each branch on the star represents a different physical feature. Two cases that share a particular feature will share the same corresponding branch of the star".

The spatial visualizations were generated by applying the metric multidimensional scaling algorithm described by Lee (Lee 2001, p. 156). The Euclidean distance metric was used, and similarity measures were derived using both the common and distinctive approaches. For each visualization, ten independent two-dimensional multidimensional scaling solutions were found, and the best-fitting configuration was used. A Procrustes transformation (Sibson 1978) was also applied, ensuring the best possible alignment between the common and distinctive spatial visualizations that was achievable using distance-preserving translations and rotations.

Figure 4 shows the distinctive spatial visualization of the animals data set. The instructions given to participants using this visualization were: "The following spatial visualization represents a group of twenty animals that each has one or more of a set of physical features. Each number represents a different animal. The numbers are arranged according to their degree of similarity, with those animals that are more similar being placed closer together. Similarity between two animals is calculated as the number of physical features they have in common, added to the number of physical features the two animals do not have, out of the total number of physical features being considered. That is, two animals are considered similar if they share many physical features and if there are many physical features that both animals do not have".

Figure 5: The common spatial visualization of the animals data set.



Finally, Figure 5 shows the common spatial visualization of the animals data set. The instructions given to participants using this visualization were: "The following spatial visualization represents a group of twenty animals that each has one or more of a set of physical features. Each number represents a different animal. The numbers are arranged according to their degree of similarity, with those animals that are more similar being placed closer together. Similarity between two animals is calculated as the number of physical features they have in common out of the total set of features being considered".

4.4 Participants

The participants were 32 adults (24 males, 8 females) ranging in age from 21 to 59 years (mean=34.41, sd=10.98) with varied experience using data visualizations.

4.5 Procedure

Each participant answered the same set of eight questions for each data set, and used each type of visualization exactly once. The pairing of data sets with visualizations was balanced, so that, across all participants, the questions associated with each data set were answered using each visualization an equal number of times. The order in which participants used the different visualizations was also balanced by presenting glyph and spatial visualizations using every possible combination of alternate presentation. Finally, the order of presentation of the data sets was reversed for half of the sample. This means that the first participant was presented with the data sets in the order: co-starring actors, movie genres, country produce, animals; and used the visualizations in the order: faces, common, stars, distinctive. The second participant then received the data sets in the reverse order and used the visualizations in the order: faces, distinctive, stars, common.

Participants took part in the experiment individually with an experimenter present to record the time taken to answer each question. They were instructed that, although they were being timed, they should take as long as they wished answering each question, and that the questions could be completed in any order. The questions were presented in 'pen and paper' format and the visualizations were presented on a separate page together with the instruction paragraph. Participants were asked to indicate their level of confidence in every response by circling the appropriate number on a confidence scale that appeared with each question. The scale ranged from 1 to 5, with 1 indicating a guess and 5 indicating certainty.

The time measure was taken between the act of writing answers to successive questions on the paper, and did not include the time taken to read the initial instructions. The decision to use manual timing was made to accommodate the 'pen and paper' testing format, and avoid a computer administered test. Many of the global questions required 'free form' answers, involving un-ordered lists of stimuli. We believed that, for our participant pool, any computer interface able to accept these sorts of answers would lead to individual differences in response times that related to computing skills rather than decision making processes. For this reason, the small increase in measurement error arising from relying on manual timing seemed worthwhile to avoid the larger measurement error expected to be caused by differences in computing skills.

4.6 Results

4.6.1 Accuracy

Figure 6 shows the mean accuracy, together with one standard error in each direction, for each visualization type, broken down by the two questions types. White markers correspond to local questions, black markers correspond to global questions.

Two-way Analyses of Variance (ANOVAs) were conducted with respect to visualization type and question type for the accuracy, confidence and time measures. Mean accuracy varied significantly across the different visualizations (F(3,93)=11.721, $p < .01$). Planned contrasts between each visualization pair revealed significantly greater accuracy for the common approach when compared to the distinctive (F(1,31)=22.677, $p < .01$), faces (F(1,31)=39.135, $p < .01$) and stars (F(1,31)=12.447, $p < .01$). None of the remaining contrasts were significant.

Mean accuracy for local questions was significantly greater than for global questions (F(1,31)=60.963, $p < .01$). There was also a significant interaction between visualization type and question type (F(3,31)=10.839, $p < .01$). This indicates that the influence of visualization type on the accuracy of responses was dependent on the type of question asked, which seems to be largely attributable to the increased accuracy for global questions when using the common spatial visualization approach.

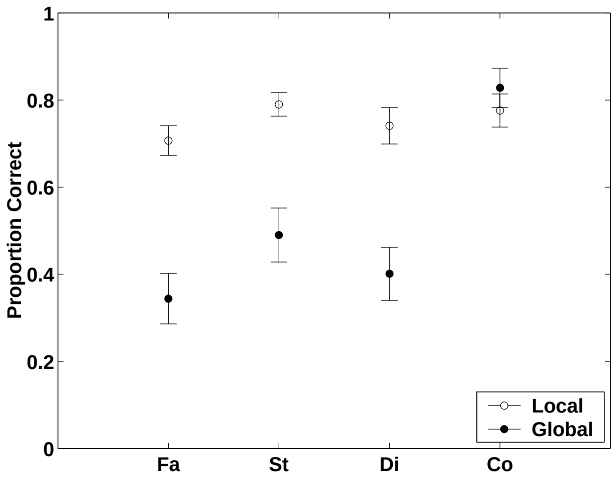


Figure 6: Mean accuracy across the four visualization types (Fa='face glyph', St='star glyph', Di='distinctive spatial', Co='common spatial').

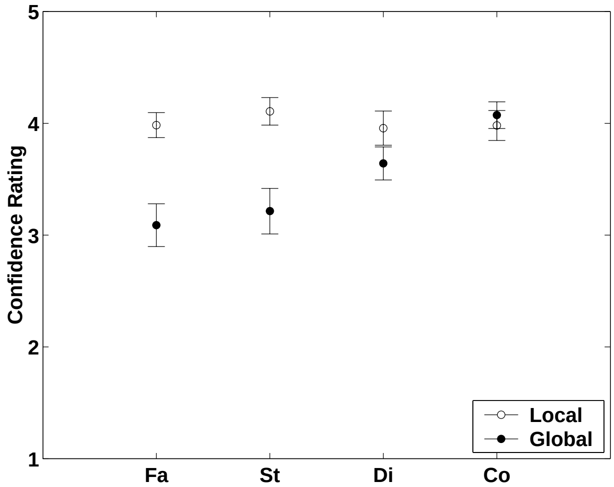


Figure 7: Mean confidence across the four visualization types (Fa='face glyph', St='star glyph', Di='distinctive spatial', Co='common spatial').

4.6.2 Confidence

Figure 7 shows the mean confidence, together with one standard error in each direction, for each visualization type, broken down by the two questions types.

There was a moderately significant difference in mean confidence across the different visualizations (F(2.037,63.150)=4.001, $p < .05$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity of variance was significant for this comparison (W(5)=.445, $p < .01$). Planned contrasts between each visualization pair revealed significantly greater confidence for the common approach when compared to the faces (F(1,31)=10.377, $p < .01$) and stars (F(1,31)=15.491, $p < .01$), but *not* the distinctive visualization. As with the accuracy measures, none of the remaining contrasts were significant.

Mean confidence for local questions was significantly greater than for global questions (F(1,31)=65.645, $p < .01$). There was also a significant interaction between visualization type and question type (F(2.310,71.599)=9.258, $p < .01$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity

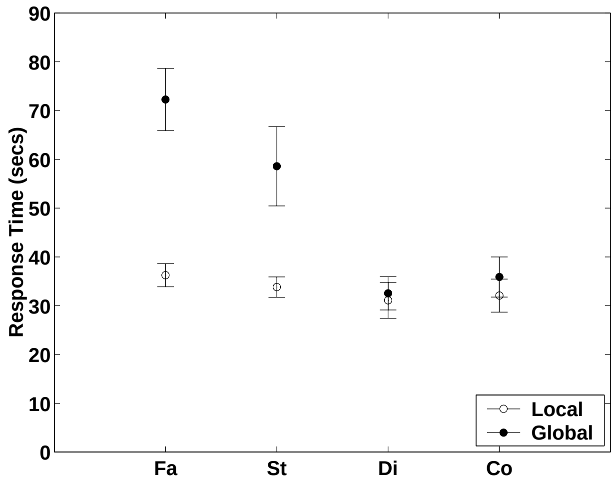


Figure 8: Mean response time across the four visualization types (Fa='face glyph', St='star glyph', Di='distinctive spatial', Co='common spatial').

of variance was significant for this comparison (W(5)=.627, $p < .05$). The presence of an interaction indicates that the influence of visualization type on the confidence of responses was dependent on the type of question asked, which seems to be attributable to the increased confidence for global questions when using spatial visualizations.

4.6.3 Time

Figure 8 shows the mean confidence, together with one standard error in each direction, for each visualization type, broken down by the two questions types.

Mean time varied significantly across the different visualizations (F(3,93)=11.148, $p < .01$). Planned contrasts between both spatial visualizations lead to quicker response times than both glyph visualizations; common visualizations were quicker than face visualizations (F(1,31)=16.174, $p < .01$) and star visualizations (F(1,31)=6.434, $p < .05$), and distinctive visualizations were quicker than face visualization (F(1,31)=24.561, $p < .01$) and star visualizations (F(1,31)=10.985, $p < .01$). The remaining two contrasts, between the common and distinctive visualizations, and the face and star visualizations were not significant.

Mean response time for local questions was significantly faster than for global questions (F(1,31)=29.964, $p < .01$). There was also a significant interaction between visualization type and question type (F(2.270,70.373)=10.965, $p < .01$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity of variance was significant for this comparison (W(5)=.519, $p < .01$). This interaction indicates that the influence of visualization type on the response time dependent on the type of question asked, which seems to be attributable to the faster response times for global questions when using spatial visualizations.

Having presented these statistical analyses, however, we should note that we are sensitive to criticisms of Null Hypothesis Significance Testing (NHST) as a method of scientific inference (Cohen 1994, Edwards, Lindman & Savage 1963, Howson & Urbach 1993, Hunter 1997, Lindley 1972). In particular, we acknowledge that NHST violates the likelihood principle, and so does not satisfy a basic requirement for rational, consistent and coherent statistical decision making (Lindley 1972). For this reason, we have also undertaken Bayesian analyses of the data (Carlin & Louis 2000, Kass & Raftery 1995, Gelman, Carlin, Stern & Rubin 1995, Leonard & Hsu 1999, Sivia 1996). Fortunately, the Bayesian analyses are consistent with the NHST findings, and so we do not present them here. They are available from the first author on request.

5 Discussion

5.1 Summary of Findings

The experimental results show that the common spatial visualization was the best performed, largely due to performance on the global questions. Both glyph visualizations lead to slow, inaccurate responses to global questions, and participants reported low confidence when using these visualizations. Meanwhile, participants reported high confidence and took less time to answer global questions when using the two spatial visualizations. The accuracy of responses to global questions was high when using the common spatial visualization but low for the distinctive visualization, suggesting that the distinctive visualization seemed effective even though it led to inaccurate responses.

These findings are consistent with Lee and Vickers proposition that visualizations presenting the unprocessed raw data do not convey information as effectively as those that restructure data according to cognitive demands. It is also consistent with the previous finding that coordinate locations are an accurate way to display information (Cleveland & McGill 1984). The difference between the two spatial visualizations demonstrates, however, that choosing the appropriate representation is important. In fact, the distinctive spatial visualization may be considered the worst of all the visualizations evaluated, since it is quickly and confidently used, but leads to inaccurate responses for global questions. Although responses to the glyph visualizations also tended to be less accurate, the low confidence reported by participants suggests that they were at least aware of the level of accuracy being achieved.

The long response times associated with the glyph displays are consistent with the idea that participants were processing information serially, as has previously been found for Chernoff faces (Morris, Ebert & Rheingans 2000). If this was the case, the glyph displays were functioning as little more than graphical versions of the raw data. Given that one of the potential advantages of data visualization is to facilitate the effortless comprehension of large data sets, this would be a worrying finding.

5.2 Implications for Application

The extent to which this difficulty manifests itself in applied settings, however, is not entirely clear. No attempt was made in our studies to assign data features to glyph perceptual characteristics in a way that would encourage the generation of emergent features (Sanderson, Flach, Buttigieg & Casey 1989). For example, if the types of features that correspond to aquatic animals were assigned to adjacent branches of star glyphs, an emergent 'aquatic' structure may well become perceptually obvious. Similarly, these features may be able to be assigned to facial characteristics in a way that made Chernoff faces look happy when representing aquatic animals, and this emergent structure is also likely to be perceived readily. The important question in this regard is how easily the appropriate assignments can be found. In some well understood domains, it may be straightforward to identify how data features should be configured in a glyph display. In more exploratory situations, finding the appropriate assignment may be as hard a problem as making the inferences for which the visualizations are being designed in the first place. Nevertheless, it should be acknowledged that where useful emergent perceptual effects can be achieved, it is likely that glyph performance will improve beyond the levels suggested by our results. Performance is also likely to improve in an applied settings, like stock markets and military environments, where analysts have significant training and practice interpreting the standardized glyph visualizations used in the domain.

5.3 Scalability of Visualizations

A particularly promising feature of the common spatial representation is the scalability offered by its visual and conceptual simplicity. A meta-analysis of six studies investigating human performance on information visualizations demonstrated that simpler visual-spatial interfaces offer a performance advantage (Chen & Yu 2000). In particular, they found that, for users with the same cognitive ability, responses were faster for simpler visual-spatial interfaces. Both spatial visualizations may be considered simpler than the glyph visualizations because each item is represented simply by a point, and the only information required for determining similarity relations is the Euclidean distance between these points. This means that spatial visualizations are less limited than glyph visualizations in terms of the number of objects they can display. The common spatial representation has the additional advantage of considering only the (generally small) subset of shared features when assessing similarity. In addition to its superior performance in facilitating accurate, confident and quick answers to a variety of questions, this scalability makes a compelling argument in favor of using the common spatial visualization approach in applied settings.

5.4 Role of Perceptual and Cognitive Factors

An important question raised, but far from completely answered by our research, relates to the relative contribution of perceptual and cognitive factors in determining the effectiveness of a visualization. For example, the relatively better performance of the stars than the faces could be attributed to both perceptual and cognitive factors. On the perceptual front, Chernoff has acknowledged that some display features are difficult to detect in some face visualizations (Chernoff 1973, p. 366), and it is also not possible to omit facial features to indicate the absence of a property. On the cognitive front, it seems likely that the problem of assigning underlying variables to perceptual features (Everitt 1978, Toit, Steyn & Stumf 1986, Manly 1994) is more severe in the case of faces, because of their inherent meaning and different saliencies. There is also a possibility of individual differences having an impact in the semantic perception of facial features (Chatfield & Collins 1980). The extent to which these competing explanations are responsible for the poor performance of Chernoff faces is not addressed by our findings.

More generally, our results do not allow for the effects of representational analysis to be separated completely from those of perceptual presentation. We would claim that the spatial visualizations used in this study are the canonical means of displaying MDS representations, and so the different performance of common and distictive approaches using this visualization suggests different cognitive and perceptual effects. It would be possible, therefore, to argue that glyphs remain an effective visualization technique when displaying data that has undergone an appropriate representational transformation. For example, a star glyph with two continuously varying branches could display the coordinate locations of each object within a MDS representation. In either case, of course, the need for representational analysis is consistent with the ideas put forward by Lee and Vickers that motivated this study. Nevertheless, an empirical evaluation of the relative performance of glyphs and spatial presentations that use the same underlying representation is an important area for future research. This is particularly true since, to the extent that glyph visualizations facilitate a similar level of performance, they have the attraction of generalizing more readily to display three- and higher-dimensional representations. In a sense, there is a tradeoff between the perceptual simplicity of the spatial display and the generalizability of glyph displays to large numbers of dimensions. Future evaluations may well show that spatial visualizations are very effective for inherently low-dimensional data, but that some form of glyph representation is needed for inherently high-dimensional data.

A second empirical approach to determining the relative contribution of cognitive representations involves examining a wide array of alternative visualization approaches based on cognitive models. These include techniques such as additive trees (Carroll 1976, Corter 1996, Sattath & Tversky 1977), additive clustering (Arabie & Carroll 1980, Lee 2002*a*, Lee 2002*b*, Shepard & Arabie 1979), trajectory mapping (Richards & Koenderink 1995), and others (Shepard 1980, Tenenbaum et al. 2000). All of these representations would use different displays from the spatial configurations of MDS representations, thus breaking the confound between representational modeling and visual presentation. This means, to the extent that these techniques prove to be effective, further evidence is accrued for the role of cognitive representations in generating useful data visualizations. In the end, we suspect that the best choice of representational technique and similarity model will almost certainly depend on the nature of the domain. Some data will be better suited to spatial representation in terms of underlying continuous domains, while others will be amenable to characterization in terms of the presence or absence of discrete features, or a hierarchical tree structure. There has been some research in cognitive psychology attempting to develop indices that determine the appropriate representational strategy for any given data set (Tversky & Hutchinson 1986), and this line of research should be pursued to enable visualizations to be tailored to data in an automated way. Determining which approaches yield the most robustly useful visualizations across all domain types is a topic that could be addressed by future empirical evaluations.

5.5 Extension to Continuous Data

A final, but equally important challenge, is a need to broaden the type of raw data considered from binary to continuous data. The glyph approach to data visualization extends naturally to continuous data, and it is also possible to generate continuous analogues of the common and distinctive spatial displays. Evaluating the performance of spatial visualizations of continuous data generated using these similarity approaches with glyph visualizations, and with other alternative approaches, is yet another worthwhile topic for future research.

Acknowledgements

This research was supported by the Australian Defence Science and Technology Organisation. We wish to thank Chris Woodruff and Ben Stamley for their assistance.

References

Arabie, P. & Carroll, J. D. (1980), 'MAPCLUS: A mathematical programming approach to fitting the ADCLUS model', *Psychometrika* **45**(2), 211–235.

Carlin, B. P. & Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, second edn, Chapman & Hall, New York.

Carroll, J. D. (1976), 'Spatial, non-spatial and hybrid models for scaling', *Psychometrika* **41**, 439–463.

Chatfield, C. & Collins, A. J. (1980), *Introduction to Multivariate Analysis*, Chapman and Hall., London.

Chen, C. & Yu, Y. (2000), 'Empirical studies of information visualization: A meta-analysis', *International Journal of Human-Computer Studies* **53**(5), 851–866.

Chernoff, H. (1973), 'The use of faces to represent points in k-dimensional space graphically', *Journal of the American Statistical Association* **68**, 361–368.

Chernoff, H. & Rizvi, M. H. (1975), 'Effect on classification error of random permutations of features in representing multivariate data by faces', *Journal of American Statistical Association* **70**, 548–554.

Cleveland, W. S. & McGill, R. (1984), 'Graphical perception: Theory, experimentation, and application to the development of graphical methods', *Journal of the American Statistical Association* **79**(387), 531–554.

Cohen, J. (1994), 'The earth is round ($p < .05$)', *American Psychologist* **49**, 997–1003.

Cohen, J. D. (1997), 'Drawing graphs to convey proximity: An incremental arrangement method', *ACM Transactions On Computer-Human Interaction* **4**(3), 197–229.

Corter, J. E. (1996), *Tree Models of Similarity and Association*, Sage, Thousand Oaks, CA.

Cox, T. F. & Cox, M. A. A. (1994), *Multidimensional scaling*, Chapman and Hall, London.

Dowell, J. & Long, J. (1998), 'Conception of the cognitive engineering design problem', *Ergonomics* **41**(2), 126–140.

Edwards, W., Lindman, H. & Savage, L. J. (1963), 'Bayesian statistical inference for psychological research', *Psychological Review* **70**(3), 193–242.

Everitt, B. (1978), *Graphical Techniques for Multivariate Data*, Heinemann, London.

Everitt, B. S. & Dunn, G. (1991), *Applied Multivariate Data Analysis*, Edward Arnold, London.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Getty, D. J., Swets, J. A., Swets, J. B. & Green, D. M. (1979), 'On the prediction of confusion matrices from similarity judgements', *Perception & Psychophysics* **26**, 1–19.

Graham, M., Kennedy, J. & Benyon, D. (2000), 'Towards a methodology for developing visualizations', *International Journal of Human-Computer Studies* **53**(5), 789–807.

Haskell, I. D. & Wickens, C. D. (1993), 'Two- and three-dimensional displays for aviation: A theoretical and empirical comparison', *International Journal of Aviation Psychology* **3**(2), 87–109.

Howson, C. & Urbach, P. (1993), *Scientific Reasoning: The Bayesian Approach*, Open Court Press, La Salle, IL.

Hunter, J. E. (1997), 'Needed: A ban on the significance test', *Psychological Science* **8**(1), 3–7.

Jacob, R. J. K., Egeth, H. E. & Bevon, W. (1976), 'The face as data display', *Human Factors* **18**, 189–200.

Jones, P. M., Wickens, C. D. & Deutsch, S. J. (1990), 'The display of multivariate information: An experimental study of an information integration task', *Human Performance* **3**(1), 1–17.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kosslyn, S. M. (1994), *Elements of Graph Design*, W. H. Freeman, New York, NY.

Kruschke, J. K. (1992), 'ALCOVE: An exemplar-based connectionist model of category learning', *Psychological Review* **99**(1), 22–44.

Kruskal, J. B. (1964), 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* **29**(1), 1–27.

Lee, M. D. (2001), 'Determining the dimensionality of multidimensional scaling representations for cognitive modeling', *Journal of Mathematical Psychology* **45**(1), 149–166.

Lee, M. D. (2002*a*), 'Generating additive clustering models with limited stochastic complexity', *Journal of Classification* **19**(1), 69–85.

Lee, M. D. (2002*b*), 'A simple method for generating additive clustering models with limited complexity', *Machine Learning* **49**, 39–58.

Lee, M. D. & Vickers, D. (1998), 'Psychological approaches to data visualisation', DSTO Research Report DSTO–RR–0135.

Leonard, T. & Hsu, J. S. J. (1999), *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge University Press, New York.

Lindley, D. V. (1972), *Bayesian Statistics: A Review*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Liu, Y. (1989), Use of computer graphics and cluster analysis in aiding relational judgment, *in* 'Proceedings of the Human Factors Society 33rd Annual Meeting', Human Factors Society, Santa Monica, CA, pp. 345–349.

Lokuge, I., Gilbert, S. A. & Richards, W. (1996), Structuring information with mental models: A tour of Boston, *in* R. Bilger, S. Guest & M. J. Tauber, eds, 'Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems, Vancouver, Canada'.

Lowe, D. & Tipping, M. E. (1996), 'Feed-forward neural networks and topographic mappings for exploratory data analysis', *Neural Computing and Applications* **4**, 83–95.

Manly, B. F. J. (1994), *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London.

Mao, J. & Jain, A. K. (1995), 'Artificial neural networks for feature extraction and multivariate data projection', *IEEE Transactions on Neural Networks* **6**(2), 296–317.

Meyer, J. (2000), 'Performance with tables and graphs: Effects of training and a visual search model', *Ergonomics* **43**(11), 1840–1865.

Morris, C. J., Ebert, D. S. & Rheingans, P. (2000), An experimental analysis of the effectiveness of features in Chernoff faces, *in* W. R. Oliver, ed., '28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making, Proceedings of SPIE Vol. 3905', pp. 12–17.

Morse, E. L., Lewis, M. & Olsen, K. A. (2000), 'Evaluating visualizations: Using a taxonomic guide', *International Journal of Human-Computer Studies* **53**(5), 637–662.

NetMap, C. S. (2001). Westerville, OH: Netmap Analytics, LLC.

Nosofsky, R. M. (1986), 'Attention, similarity, and the identification-categorization relationship', *Journal of Experimental Psychology: General* **115**(1), 39–57.

Purchase, H. C. (1998), Effects of graph layout, *in* 'Proceedings of OZCHI 98', IEEE Computer Society, Los Alamitos, CA, pp. 80–86.

Purchase, H. C. (2000), 'Effective information visualization: A study of graph drawing aesthetics and algorithms', *Interacting with computers* **13**(2), 147–162.

Richards, W. & Koenderink, J. J. (1995), 'Trajectory mapping: A new nonmetric scaling technique', *Perception* **24**, 1315–1331.

Sanderson, P. M., Flach, J. M., Buttigieg, M. A. & Casey, E. J. (1989), 'Object displays do not always support better integrated task performance', *Human Factors* **31**(2), 183–198.

Sattath, S. & Tversky, A. (1977), 'Additive similarity trees', *Psychometrika* **42**, 319–345.

Schneiderman, B. (1998), *Designing the user interface: Strategies for effective human-computer interaction*, third edn, Addison Wesley, Reading, MA.

Shepard, R. N. (1957), 'Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space', *Psychometrika* **22**(4), 325–345.

Shepard, R. N. (1980), 'Multidimensional scaling, tree-fitting, and clustering', *Science* **210**, 390–398.

Shepard, R. N. (1987), 'Toward a universal law of generalization for psychological science', *Science* **237**, 1317–1323.

Shepard, R. N. (1994), 'Perceptual-cognitive universals as reflections of the world', *Psychonomic Bulletin & Review* **1**(1), 2–28.

Shepard, R. N. & Arabie, P. (1979), 'Additive clustering representations of similarities as combinations of discrete overlapping properties', *Psychological Review* **86**(2), 87–123.

Sibson, R. (1978), 'Studies in the robustness of multidimensional scaling: Procrustes statistics', *Journal of the Royal Statistical Society, Series B* **40**(2), 234–238.

Sivia, D. S. (1996), *Data Analysis: A Bayesian Tutorial*, Clarendon Press, Oxford.

Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000), 'A global geometric framework for nonlinear dimensionality reduction', *Science* **290**, 2319–2323.

Toit, S. H. C., Steyn, A. G. W. & Stumf, R. H. (1986), *Graphical Exploratory Data Analysis*, Springer-Verlag, New York.

Tractinsky, N., Katz, A. S. & Ikar, D. (2000), 'What is beautiful is usable', *Interacting with computers* **13**(2), 127–145.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A. & Raymond, P. D. (2000), 'Turning pictures into numbers: Extracting and generating information from complex visualizations', *International Journal of Human-Computer Studies* **53**(5), 827–850.

Tversky, A. (1977), 'Features of similarity', *Psychological Review* **84**(4), 327–352.

Tversky, A. & Hutchinson, J. W. (1986), 'Nearest neighbor analysis of psychological spaces', *Psychological Review* **93**(1), 3–22.

Ware, C. (2000), *Information Visualization: Perception for Design*, Morgan Kauffman, San Mateo, CA.

Wasserman, S. & Faust, K. (1994), *Social Network Analysis: Methods and Applications.*, Cambridge University Press, New York.

Wehrend, S. & Lewis, C. (1990), 'A problem-oriented classification of visualization techniques', *Proceedings of IEEE Visualization '90* pp. 139–143.

Westerman, S. J. & Cribben, T. (2000), 'Mapping semantic information in virtual space: Dimensions, variance and individual differences', *International Journal of Human-Computer Studies* **53**(5), 765–787.

Wise, J. A. (1999), 'The ecological approach to text visualization', *Journal of the American Society for Information Science* **50**(13), 1224–1233.