ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Int. J. Human-Computer Studies 59 (2003) 569–602

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Visualizations of binary data: A comparative evaluation

Michael D. Lee[a,*], Marcus A. Butavicius[a], Rachel E. Reilly[b]

[a] *Department of Psychology, University of Adelaide, Adelaide, SA 5005, Australia*
[b] *Department of Psychology, University of Melbourne, Australia*

Received 14 January 2002; accepted 25 March 2003

Abstract

Data visualization has the potential to assist humans in analysing 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 various different approaches are available for data visualization, there have been few rigorous evaluations of their effectiveness. This paper presents the results of three controlled experiments comparing the ability of 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. The first experiment finds 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. The second experiment, which considers only the spatial visualizations, supports this finding, with the common features approach again producing more accurate answers. The third experiment measures human performance using the raw data in tabular form, and so allows the usefulness of visualizations in facilitating human performance to be assessed. This experiment confirms that people are faster, more confident and more accurate when an appropriate visualization of the data is made available.
© 2003 Elsevier Science (USA). All rights reserved.

Keywords: Data visualization; Chernoff faces; Glyphs; Multidimensional scaling; Empirical evaluation

*Corresponding author. Tel.: +61-8-8303-6096; fax: +61-8-8303-3770.
E-mail address: michael.lee@psychology.adelaide.edu.au (M.D. Lee).
URL: http://www.psychology.adelaide.edu.au/members/staff/michaellee/homepage.

1071-5819/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1071-5819(03)00082-X

1. Introduction

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 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 often been argued that a principled psychological approach to data visualization is warranted (e.g. Chernoff, 1973; Purchase, 1998; Shneiderman, 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. What is less often acknowledged is the role of more abstract cognitive representational principles, not directly related to perceptual processes, in developing data visualizations (although see Kosslyn, 1994; Lokuge et al., 1996). 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.

A psychological framework for data visualization that incorporates both perceptual and cognitive components is shown in Fig. 1. As originally argued in Lee and Vickers (1998), 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,

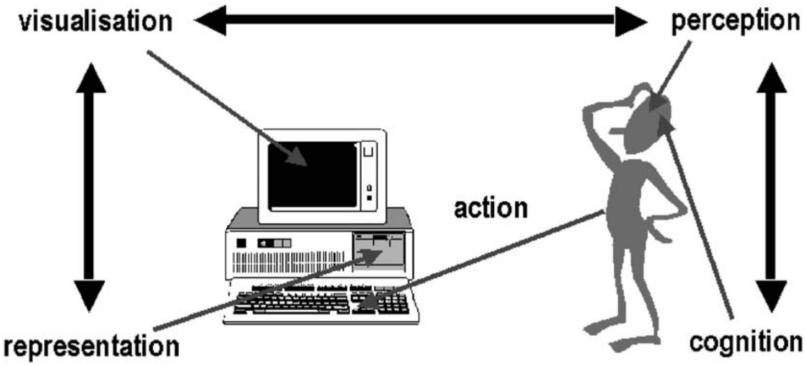

Fig. 1. A psychological framework for data visualization. Based on Lee and Vickers (1998, Fig. 1).

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 Fig. 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 (1998) suggested that data visualization techniques which perform little or no manipulation of the data before attempting to represent it graphically, with the intention of 'letting the data speak for themselves, 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 (2000, p. 1840) points out, there are no generally accepted guidelines for the optimal display of data. 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, Morse et al. (2000) note that the evaluation of data visualizations is rarely undertaken. 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.

The first of these approaches, criticized by Purchase (1998), 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. While humans tend to associate aesthetic qualities of systems with perceived usability, this perceived usability may be independent of actual usability (Tractinsky et al., 2000). Indeed, Purchase (2000) found that the data visualization judged as most symmetrical by participants was also associated with the highest rate of errors. Certainly, assessing data visualizations through measures of their aesthetics does not provide a direct measure of their ability to facilitate human performance.

A second approach to evaluation has focused on the assessment of domain specific visualizations (e.g. Graham et al., 2000; Trafton et al., 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 and 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.

A more direct and general approach to evaluation has been employed within experimental cognitive psychology (e.g. Liu, 1989; Jones et al., 1990; Haskell and Wickens, 1993), 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 themselves. 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 General Discussion.

3. Four visualizations of 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 (e.g. Wasserman and 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 are 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 (see 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.

3.1. 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 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. Chernoff (1973) has argued that examining such glyphs may help to uncover specific clusters of both simple relations and interactions between variables.

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 and 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 et al., 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 analysing the representations (Everitt and Dunn, 1991).

3.2. Spatial visualizations

Spatial visualizations represent objects as points in a multidimensional (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, 1987, 1994), and is used to represent stimuli in various formal psychological models of identification, categorization, selective attention, and other cognitive processes (e.g. Getty et al., 1979; Nosofsky, 1986;

Kruschke, 1992). The algorithms that generate spatial representations, generically known as multidimensional scaling algorithms (e.g. Kruskal, 1964; see Shepard, 1980 for an overview), have also been applied to data visualization, exploration and analysis (e.g. Mao and Jain, 1995; Lowe and Tipping, 1996).

Multidimensional scaling (MDS) 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 (1994, p. 11) provide a list of a dozen straightforward approaches, and there are other more sophisticated measures (e.g. Cohen, 1997; Tenenbaum et al., 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 (1994) refer to these alternatives as the 'Matching' and 'Jaccard' coefficients, respectively. Our terminology is taken from the seminal psychological theory of feature-based stimulus similarity presented by 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 by Westerman and 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 I

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.

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

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 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 (e.g. Wehrend and 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.

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

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].*

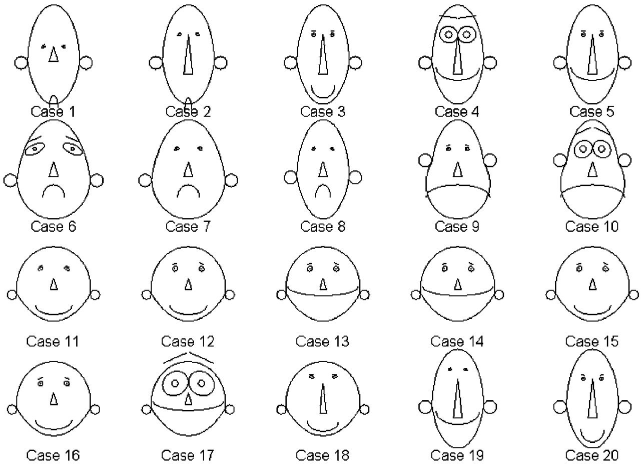

Fig. 2. The faces visualization of the animals data set.

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. Fig. 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 20 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'.

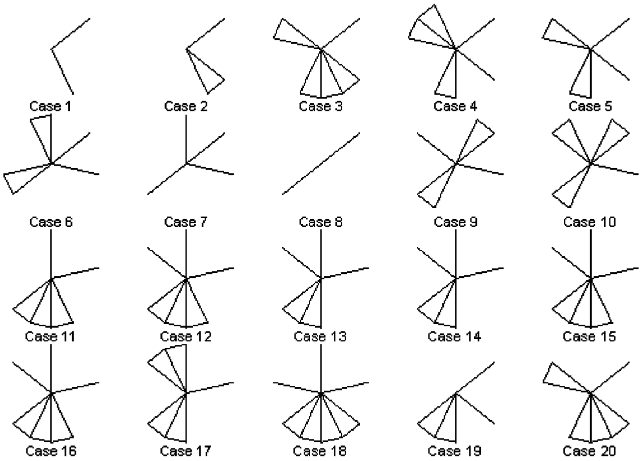

Fig. 3. The stars visualization of the animals data set.

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. Fig. 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 20 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 in 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.

Fig. 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 20 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'.

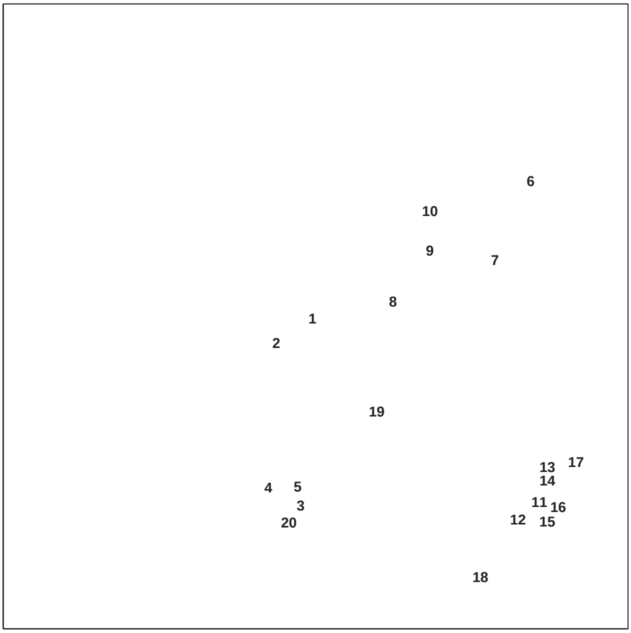

Fig. 4. The distinctive spatial visualization of the animals data set.

Finally, Fig. 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 20 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, eight females) ranging in age from 21 to 59 years (mean = 34.41, S.D. = 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.

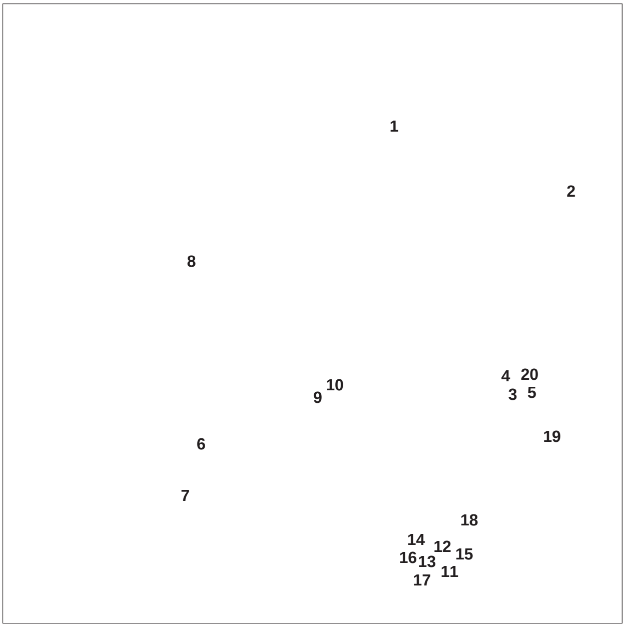

Fig. 5. The common spatial visualization of the animals data set.

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. 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.[1] Participants were asked to

[1] 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

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.

4.6. Results

In making statistical inferences from our data, we use the standard psychological approach of Null Hypothesis Significance Testing (NHST). We are sensitive, however, to criticisms of NHST as a method of scientific inference (e.g. Edwards et al., 1963; Lindley, 1972; Howson and Urbach, 1993; Cohen, 1994; Hunter, 1997). 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 also present a Bayesian analysis of the data (see, for example, Gelman et al., 1995; Kass and Raftery, 1995; Sivia, 1996; Leonard and Hsu, 1999; Carlin and Louis, 2000).

Under the Bayesian approach, we treat the human performance data as evidence for or against competing models of the relationship between visualization types, questions types, and decision accuracy, confidence and time. In particular, we are interested in what evidence the data provide for accuracy, confidence or time depending upon the visualization used for different types of questions. This requires, for example, being able to assess the probability that the distribution of response times for local questions depends upon the visualization method that is employed. We are also interested in assessing to what extent the data provide evidence for the distribution of accuracy, confidence or time being the same in particular cases. This requires, for example, being able to assess the probability that decisions for local questions using a face visualization are distributed in the same way as decisions for local questions using a star visualization. There are established Bayesian methods for calculating or estimating these probabilities (Margaritis and Thrun, 2001), which are applicable to our data.[2] A formal summary of these statistical techniques is provided in the appendix.

Fig. 6 shows the mean accuracy, confidence and time, together with one standard error in each direction, for each visualization type, broken down by the two questions types. It is important to understand that, while Fig. 6 provides a convenient and useful graphical summary of the results, it does not capture the complexity of the underlying distributions from which the means and standard errors are derived. For this reason, we base our substantive conclusions relating to

(*footnote continued*)
error arising from relying on manual timing seemed worthwhile to avoid the larger measurement error expected to be caused by differences in computing skills.

[2] It is important to understand that a Bayesian probability is not the same thing as a *p* value reported under NHST. The Bayesian probability is the probability of a model being true given the data, whereas the *p* value is best interpreted as the probability of the data, or data more extreme, given the null hypothesis (see Robert, 1994, p. 197). Because of their fundamental differences, it is not sensible to think of the Bayesian probabilities in terms of the ‘critical values’ of NHST.

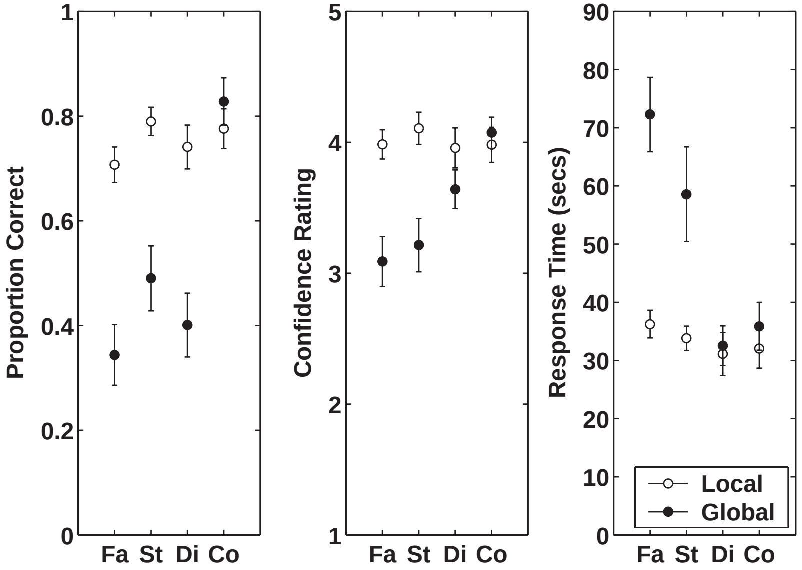

Fig. 6. Mean accuracy (left panel), confidence (middle panel), and time (right panel) across the four visualization types (Fa='face glyph', St='star glyph', Di='distinctive spatial', Co='common spatial'). White markers correspond to local questions, black markers correspond to global questions, and one standard error is shown about each mean.

the data primarily upon the Bayesian analysis, which does consider the entire distribution.

NHST analysis: Two-way 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 ($F(3, 93) = 11.721$, $p < 0.01$). Planned contrasts between each visualization pair revealed significantly greater accuracy for the common approach when compared to the distinctive ($F(1, 31) = 22.677$, $p < 0.01$), faces ($F(1, 31) = 39.135$, $p < 0.01$) and stars ($F(1, 31) = 12.447$, $p < 0.01$). None of the remaining contrasts were significant.

Mean accuracy for local questions was significantly greater than for global questions ($F(1, 31) = 60.963$, $p < 0.01$). There was also a significant interaction between visualization type and question type ($F(3, 31) = 10.839$, $p < 0.01$). 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.

There was a moderately significant difference in mean confidence across the different visualizations ($F(2.037, 63.150) = 4.001$, $p < 0.05$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity of variance was significant for this comparison ($W(5) = 0.445$, $p < 0.01$). Planned contrasts between each visualization pair revealed significantly greater confidence for the common approach when compared to the faces ($F(1, 31) = 10.377$, $p < 0.01$) and stars ($F(1,31) = 15.491$, $p < 0.01$), 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 ($F(1,31) = 65.645$, $p < 0.01$). There was also a significant interaction between visualization type and question type ($F(2.310, 71.599) = 9.258$, $p < 0.01$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity of variance was significant for this comparison ($W(5) = 0.627$, $p < 0.05$). 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.

Mean time varied significantly across the different visualizations ($F(3,93) = 11.148$, $p < 0.01$). Planned contrasts between both spatial visualizations lead to quicker response times than both glyph visualizations; common visualizations were quicker than face visualizations ($F(1,31) = 16.174$, $p < 0.01$) and star visualizations ($F(1,31) = 6.434$, $p < 0.05$), and distinctive visualizations were quicker than face visualization ($F(1,31) = 24.561$, $p < 0.01$) and star visualizations ($F(1,31) = 10.985$, $p < 0.01$). 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 ($F(1,31) = 29.964$, $p < 0.01$). There was also a significant interaction between visualization type and question type ($F(2.270, 70.373) = 10.965$, $p < 0.01$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity of variance was significant for this comparison ($W(5) = 0.519$, $p < 0.01$). This interaction indicates that the influence of visualization type on the response time was dependent on the type of question asked, which seems to be attributable to the faster response times for global questions when using spatial visualizations.

Bayesian analysis: In terms of the relationship between visualization type and accuracy, the Bayesian posterior probability of independence, based on human performance across all four data sets is 0.95 for the local questions, and 0.02 for the global questions. This means that it is very likely that the visualization type does not have an impact upon people's accuracy in answering local questions, but very likely that it does have an impact on the global questions. Similar results are obtained when the relationship between visualization type and time is considered. The probability of independence is 0.97 (to two decimal places) for local questions, and 0.00 for global questions,[3] strongly suggesting that there is a dependency for global questions, but not for local ones. The probability that confidence is independent of visualization type is 0.99 for the local questions, and 0.36 for the global questions. This means that, once again, the visualization type does not impact upon people's

[3] Note that we are not claiming a zero probability, but are rounding to two decimal places. This means a reported probability of 0.00 can be interpreted as meaning less than 0.005.

Table 3
Pairwise probability that the distribution of accuracy, confidence (in italics), and time (in bold) is the same for each combination of visualization and question types

	Faces–Local	Stars–Local	Distinctive–Local	Common–Local	Faces–Global	Stars–Global	Distinctive–Global	Common–Global
Faces–Local	0.97 *0.95* **0.90**	0.85 *0.93* **0.75**	0.96 *0.92* **0.07**	0.88 *0.94* **0.61**	0.00 *0.00* **0.00**	0.40 *0.01* **0.00**	0.00 *0.37* **0.67**	0.85 *0.91* **0.53**
Stars–Local	0.85 *0.93* **0.75**	0.96 *0.95* **0.93**	0.93 *0.86* **0.06**	0.96 *0.90* **0.73**	0.00 *0.00* **0.00**	0.01 *0.00* **0.03**	0.00 *0.08* **0.83**	0.95 *0.92* **0.82**
Distinctive–Local	0.96 *0.92* **0.07**	0.93 *0.86* **0.06**	0.97 *0.96* **0.91**	0.94 *0.95* **0.87**	0.00 *0.00* **0.00**	0.16 *0.11* **0.00**	0.00 *0.01* **0.69**	0.91 *0.84* **0.23**
Common–Local	0.88 *0.94* **0.61**	0.96 *0.90* **0.73**	0.94 *0.95* **0.87**	0.97 *0.96* **0.89**	0.00 *0.00* **0.00**	0.01 *0.02* **0.00**	0.00 *0.01* **0.74**	0.95 *0.87* **0.75**
Faces–Global	0.00 *0.00* **0.00**	0.00 *0.00* **0.00**	0.00 *0.00* **0.00**	0.00 *0.00* **0.00**	0.95 *0.95* **0.92**	0.65 *0.93* **0.01**	0.93 *0.13* **0.00**	0.00 *0.01* **0.00**
Stars–Global	0.40 *0.01* **0.00**	0.01 *0.00* **0.03**	0.16 *0.11* **0.00**	0.01 *0.02* **0.00**	0.65 *0.93* **0.01**	0.95 *0.95* **0.88**	0.89 *0.55* **0.00**	0.03 *0.02* **0.11**
Distinctive–Global	0.00 *0.37* **0.67**	0.00 *0.08* **0.83**	0.00 *0.01* **0.69**	0.00 *0.01* **0.74**	0.93 *0.13* **0.00**	0.89 *0.55* **0.00**	0.95 *0.92* **0.92**	0.00 *0.69* **0.77**
Common–Global	0.85 *0.91* **0.53**	0.95 *0.92* **0.82**	0.91 *0.84* **0.23**	0.95 *0.87* **0.75**	0.00 *0.01* **0.00**	0.03 *0.02* **0.11**	0.00 *0.69* **0.77**	0.94 *0.91* **0.92**

confidence in answering local questions, but that the data only provide limited evidence in favor of an impact for the global questions.

Table 3 presents the results of analysing the evidence provided by the data for differences between individual accuracy, confidence and time distributions. Each possible combination of visualization type and question type is considered in relation to each of the other combinations. These results are presented in terms of an 8×8 matrix, with rows and columns corresponding to the eight combinations. Within each cell, probabilities for accuracy, confidence and time are presented in normal type, italics, and bold respectively. To assist in interpreting Table 3, it is worth giving a concrete example. The probability that the time distribution is the same for local questions using the stars visualization (Stars–Local) as it is for global questions using the common spatial visualization (Common–Global) is given in the bottom (bold) entry in second row and eighth column, which is 0.82. The same value is observed in the eighth row and second column, since the pairwise comparison matrix is symmetric. It is interesting to note that the diagonals of this matrix, where a distribution is compared to itself, do not have probabilities of one. This is sensible, since finite data are not capable of proving the distributions are identical. It would not, for example, be reasonable to conclude that two confidence distributions were identical if they both consisted of only two (identical) values. The Bayesian analysis is sensitive to these considerations of sample size and, in a sense, the diagonal values provide measure of the ability of the data collected to establish that two distributions are the same.

Table 3 indicates that the data provide strong evidence for accuracy being the same for the Faces–Local, Stars–Local, Distinctive–Local, Common–Local and Common–Global combinations. It is also likely that the Faces–Global, Stars–Global, and Distinctive–Global combinations have the same accuracy distribution, although the evidence is least compelling in the Stars–Global case. There is strong evidence that these two groups have different accuracy distributions from each other. The probability that they are the same is very low for every relevant pairwise comparison, with the sole exception of the Stars–Global and Faces–Local combinations.

A similar pattern of results is observed in terms of confidence. As with accuracy, the Faces–Local, Stars–Local, Distinctive–Local, Common–Local and Common–Global combinations have distributions that are likely to be the same. Once again, it is likely that the Faces–Global and Stars–Global combinations are different from this group, but are the same as each other. The relationship of confidence on the Distinctive–Global combination, however, is less clear. The posterior probability of 0.55 means that the data provide little evidence as to whether or not it is different from the Stars–Global combination.

In terms of response times, the posterior probabilities in Table 3 do not suggest simple groupings. The most important pairwise comparisons show that the Faces–Local combination is different from the Faces–Global combination, and that the Stars–Local combination is different from the Stars–Global combination, but that it is likely, with posterior probability 0.69, that the Distinctive–Local combination is the same as the Distinctive–Global combination, and, with posterior probability

0.75, that the Common–Local combination is the same as the Common–Global combination.

It is almost certainly the case that these probabilities vary across the different data sets. Our reason for using many data sets, however, is not to model these sorts of dependencies, but rather to assess the effectiveness of the different visualizations across many domains and many questions. In essence, we do not treat data set as an independent variable. In this sense, our methodology is closely aligned with the applied problem: Over all binary data sets, which visualization best facilitates human performance? Of course, because the experiment only considers a limited number of data sets, our answers may not be completely generalizable. One of the attractions of the Bayesian analysis in this regard is that its evidence-based framework deals naturally with any contradictions apparent across data sets, by being able to conclude, for example, that the data do not provide substantial evidence for either of two alternatives.

4.7. Conclusion

The results from Experiment I showed 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.

5. Experiment II

The inferiority of the glyph visualizations found in Experiment I was not entirely unexpected. The finding is consistent with Lee and Vicker's (1998) speculation that raw data should undergo a representational analysis before it is presented. However, the accuracy difference between the common and distinctive spatial visualizations caused by the global questions was not anticipated. Both visualizations are based on a MDS representation of the similarities in the data, and both the common and distinctive methods of assessing similarity have previously been used successfully in cognitive modeling (e.g. Gati and Tversky, 1984; Sattath and Tversky, 1987; Ritov et al., 1990; Lee and Navarro, 2002). To study this difference in greater detail, a second experiment was undertaken, using new data sets, and considering only the spatial visualizations.

5.1. Data sets

Four additional binary data sets were constructed for the follow-up experiment, involving sports, sounds, foods and cars.

5.2. Questions

A set of four questions was created for each of the four new data sets. These sets contained two local questions and two global questions. For the global questions, one was always an 'outlier' question and the other was always a 'cluster' question.

5.3. Visualizations

Common and distinctive spatial visualizations were generated for each of the four data sets, using the same methodology as Experiment I.

5.4. Participants

The participants were 24 adults (12 males, 12 females) ranging in age from 17 to 55 years (mean = 32.83, S.D. = 11.22), none of whom participated in the first experiment, and who again had varied experience using data visualizations.

5.5. Procedure

Each participant completed the questions for every data set, and used each data visualization twice. The combinations of visualization type and data set, and the order in which they were presented, were balanced across participants using the same approach as Experiment I. The presentation procedure was also identical to Experiment I.

5.6. Results

Fig. 7 shows mean accuracy, confidence and time, together with one standard error in each direction, for both visualization types, broken down by the two question types.

NHST analysis: One-way ANOVAs for the accuracy, confidence and time measures showed a significant difference between the sets with respect to time ($F(2.344, 66.577) = 3.146$, $p<0.05$). This value was adjusted using the Greenhouse-Geisser correction because Mauchly's test of sphericity of variance was significant for this comparison ($W(5) = 0.509$, $p<0.05$). This difference seemed to be caused by relatively faster response times for the food data set. There were, however, no differences in accuracy or confidence between the four data sets.

Two-way ANOVAs were conducted for visualization type and question type for the accuracy, confidence and time measures. Mean accuracy for the common spatial visualizations was significantly greater than for distinctive spatial visualizations ($F(1, 23) = 29.282$, $p<0.01$). In addition, mean accuracy for local questions was significantly greater than for global questions ($F(1, 23) = 306.785$, $p<0.01$). There was, however, no significant interaction between visualization type and question type for accuracy.

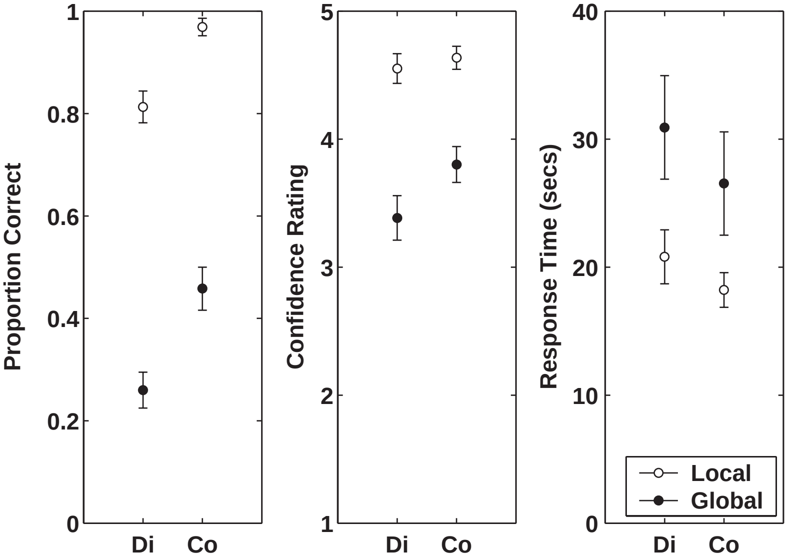

Fig. 7. Mean accuracy (left panel), confidence (middle panel), and time (right panel) across the two visualization types (Di=‘distinctive spatial’, Co=‘common spatial’). White markers correspond to local questions, black markers correspond to global questions, and one standard error is shown about each mean.

Mean confidence for the common spatial visualizations was significantly greater than for distinctive spatial visualizations ($F(1,23) = 10.931$, $p<0.01$). In addition, mean confidence for local questions was significantly greater than for global questions ($F(1,23) = 81.401$, $p<0.01$). There was also a moderately significant interaction between visualization type and question type ($F(1,23) = 7.251$, $p<0.05$), with the increase in confidence using common spatial visualization being greater for global questions.

In terms of time, participants took significantly longer to answer global questions when compared to local questions ($F(1,23) = 11.231$, $p<0.01$). Neither the main effect of visualization, nor the interaction between visualization type and question type were significant for the time measure.

Bayesian analysis: Because only two visualization types were considered in this experiment, the posterior probability of independence between a performance measure and visualization type is measured by the pairwise probability that the relevant distributions are the same. For example, the probability that accuracy is independent of visualization type for local questions is the probability that the accuracy distributions for the Distinctive–Local and Common–Local combinations are the same. Accordingly, conclusions can be drawn from Table 4, which presents the results of analysing the evidence provided by the data for differences between accuracy, confidence and time distributions across the four possible visualization type and question type combinations.

Table 4
Pairwise probability that the distribution of accuracy, confidence (in italics), and time (in bold) is the same for each combination of visualization and question types

	Distinctive–Local	Common–Local	Distinctive–Global	Common–Global
Distinctive–Local	0.95 *0.90* **0.90**	0.02 *0.77* **0.47**	0.00 *0.00* **0.00**	0.00 *0.00* **0.67**
Common–Local	0.02 *0.77* **0.47**	0.89 *0.89* **0.87**	0.00 *0.00* **0.00**	0.00 *0.00* **0.34**
Distinctive–Global	0.00 *0.00* **0.00**	0.00 *0.00* **0.00**	0.96 *0.95* **0.90**	0.27 *0.69* **0.37**
Common–Global	0.00 *0.00* **0.67**	0.00 *0.00* **0.34**	0.27 *0.69* **0.37**	0.96 *0.93* **0.92**

The posterior probability that accuracy is independent of visualization type is 0.02 for the local questions, and 0.27 for the global questions. This means that it is very likely that the visualization type does have an impact upon people's accuracy in answering local questions, and there is some evidence that it also has an impact on the global questions. For confidence, the posterior probabilities of independence are 0.77 for local questions, and 0.69 for global questions, suggesting that the visualization type does not impact upon confidence for either question type. For time, the data provide little evidence either way for local questions, with a posterior probability of 0.47, but there is some limited evidence of a dependency in relation to the global questions, with a posterior probability of 0.37.

5.7. Further global question analysis

Fig. 8 shows mean accuracy, confidence and time, together with one standard error in each direction, for global questions both visualization types, broken down by in terms of cluster and outlier questions.

NHST analysis: Mean accuracy was significantly greater for outlier questions than for cluster questions ($F(1,23) = 43.793$, $p<0.01$). There was also a significant interaction between visualization and the outlier and cluster question types ($F(1,23) = 49.000$, $p<0.01$). This seems to be largely the result of greater accuracy for outlier questions when using the common spatial visualization approach. It was also found that mean accuracy for the global questions was significantly greater for the common spatial visualization ($F(1,23) = 12.374$, $p<0.01$). There was no significant difference in the mean confidence between the outlier and cluster questions. There was, however, a significant interaction between visualization and

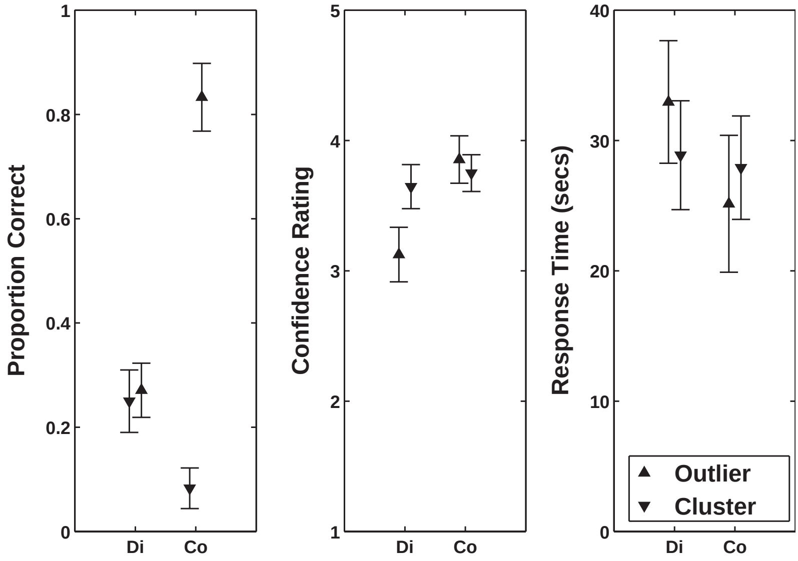

Fig. 8. Mean accuracy (left panel), confidence (middle panel), and time (right panel) for the global questions across the two visualization types (Di='distinctive spatial', Co='common spatial'). Upward pointing markers correspond to outlier questions, downward pointing markers correspond to cluster questions, and one standard error is shown about each mean.

the outlier and cluster question types ($F(1,23) = 11.275$, $p < 0.01$). This seems to be largely the result of lower confidence for outlier questions when using the distinctive spatial visualization approach. It was also found that mean global question confidence was significantly greater for the common spatial visualization ($F(1,23) = 14.839$, $p < 0.01$). No significant differences were found with respect to the response time performance measure.

Bayesian analysis: Table 5 provides the posterior probability that the distributions of each performance measure are the same, for each pairwise comparison of the visualization type and global question type combinations. The posterior probability that accuracy is independent of visualization type is 0.49 for the cluster questions, and 0.00 for the outlier questions. This means that the data provide almost no evidence either for or against the idea that visualization type affects performance on cluster questions, but does show there is an impact for outlier questions. It is likely that confidence for cluster questions does not depend on visualization type, with a posterior probability of independence of 0.79, while there is little evidence either way for outlier questions, with a posterior probability of 0.44. The data also suggest that the time taken to answer cluster questions is independent of visualization type, with posterior probability 0.83, but that there is a dependency for outlier questions, where the posterior probability of independence is 0.03. The mean of the correlations across individual data sets is 0.68 for accuracy, 0.24 for confidence and 0.73 for time.

Table 5
Pairwise probability that the distribution of accuracy, confidence (in italics), and time (in bold) is the same for each combination of visualization and question types

	Distinctive–Cluster	Common–Cluster	Distinctive–Outlier	Common–Outlier
Distinctive–Cluster	0.94 *0.91* **0.88**	0.49 *0.79* **0.83**	0.94 *0.39* **0.80**	0.00 *0.83* **0.31**
Common–Cluster	0.49 *0.79* **0.83**	0.90 *0.92* **0.88**	0.37 *0.20* **0.81**	0.00 *0.60* **0.06**
Distinctive–Outlier	0.94 *0.39* **0.80**	0.37 *0.20* **0.81**	0.94 *0.93* **0.90**	0.00 *0.44* **0.03**
Common–Outlier	0.00 *0.83* **0.31**	0.00 *0.60* **0.06**	0.00 *0.44* **0.03**	0.93 *0.92* **0.91**

5.8. Conclusion

The results relating to cluster and outlier global questions suggest that the decrease in accuracy when using the distinctive spatial visualization is largely due to outlier questions. Participants were clearly less accurate when answering these questions, and took longer to make decisions. The evidence provided by Experiment II is far less conclusive in relation to cluster questions, since it is not clear whether or not their is a dependency on visualization type for either accuracy or time.

At a general level, however, the results of Experiment II are consistent with the main findings from Experiment I. Participants remained less accurate when using the distinctive spatial visualization, although there is now evidence that this decline relates to both local and global questions. Future research seems to be needed to determine whether this improvement in local questions for the common visualization can be replicated. Experiment II also suggests that confidence does not depend upon which of the two visualizations is used, although the evidence is less strong than it was in Experiment I. Finally, Experiment II provides little evidence either way in terms of whether or not time depends upon visualization type. A reasonable summary of these results is that, as in Experiment I, participants were less accurate when using the distinctive spatial visualization, but these inaccuracies were not clearly reflected in the confidence they had in their decisions, or in the time they took to generate their answers.

6. Experiment III

From the practical standpoint of recommending a data visualization in applied settings, Experiments I and II both point towards the common spatial visualization as being the best of the four considered. What neither of the experiments evaluate, however, is whether any visualization is worth using at all. It is possible that even the common spatial visualization does not usefully improve human analysis, in the sense that the raw data themselves may allow equally or more accurate, confident and timely answers. To address this question, we conducted a third experiment, comparing human performance using a tabular display of the raw data against performance using the common spatial visualization.

6.1. Participants

The participants were 10 adults (eight males, two females) ranging in age from 22 to 52 years (mean = 32.9, s.d. = 11.3), none of whom participated in the first two experiments, and again with varied experience using data visualizations.

6.2. Procedure

Each participant completed the questions for every data set using the raw data matrix (i.e. of the type presented in Table 1). The order in which the data sets were presented was balanced across participants.

6.3. Results

Fig. 9 shows mean accuracy, confidence and time, together with one standard error in each direction, for both the raw data and the common spatial visualization, broken down by the two question types.

NHST analysis: One-way ANOVAs for the accuracy, confidence and time measures showed a significant difference between the sets with respect to accuracy ($F(3,27) = 3.569$, $p < 0.05$) and time ($F(3,27) = 4.046$, $p < 0.05$). The differences seemed to be caused by relatively faster and more accurate responses for the food data set. There were, however, no differences in confidence between the four data sets.

Two-way ANOVAs were conducted for the combination of raw data and common visualization with question type for the accuracy, confidence and time measures. Mean accuracy for the common spatial visualizations was significantly greater than for the raw data ($F(1,32) = 14.542$, $p < 0.01$). In addition, mean accuracy for local questions was significantly greater than for global questions ($F(1,32) = 205.167$, $p < 0.001$).

Mean confidence for the common spatial visualizations was significantly greater than for the raw data ($F(1,32) = 22.213$, $p < 0.001$). In addition, mean confidence for local questions was significantly greater than for global questions ($F(1,32) = 148.818$, $p < 0.001$). There was also a highly significant interaction between the raw data or common visualization and question type ($F(1,32) = 14.278$, $p < 0.01$), with the increase in confidence using common spatial visualization being greater for global questions.

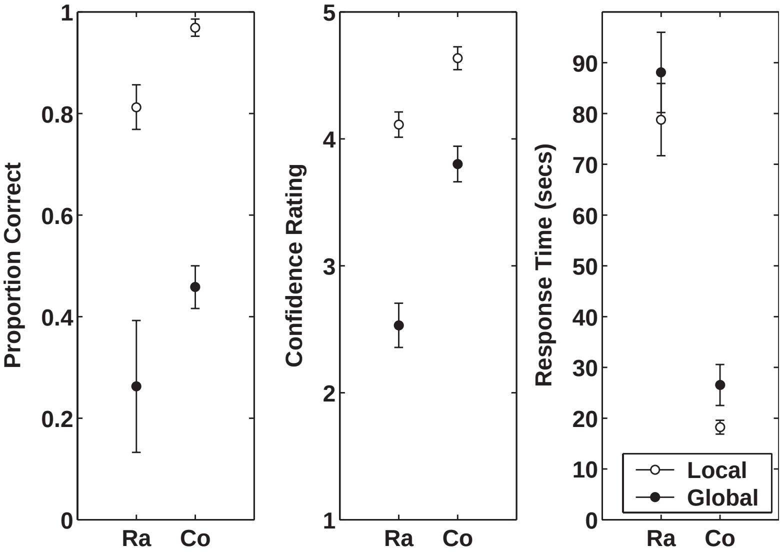

Fig. 9. Mean accuracy (left panel), confidence (middle panel), and time (right panel) for the raw data and common features spatial visualization (Ra='raw data', Co='common spatial'). White markers correspond to local questions, black markers correspond to global questions, and one standard error is shown about each mean.

In terms of time, participants took significantly longer to use the raw data than when using the common visualization ($F(1,32) = 48.908$, $p < 0.001$). They also took significantly longer to answer global questions compared to local questions ($F(1,23) = 7.190$, $p < 0.05$).

Bayesian analysis: Table 6 presents the results of analysing the evidence provided by the data for differences between accuracy, confidence and time distributions across the four possible raw data or common visualization and question type combinations. The posterior probability that accuracy is independent of raw data or common visualization is 0.03 for both the local and global questions. This means that it is very likely that using the common spatial visualization does have an impact upon people's accuracy in answering all of the questions. For confidence, the posterior probabilities of independence are 0.01 for local questions, and 0.00 for global questions, again suggesting that the use of raw data or the common spatial visualization also impacts upon confidence for all of the questions. The same result is true for time, where the posterior probabilities of independence are 0.00 for both local and global questions.

6.4. *Conclusion*

Both the NHST and Bayesian analyses confirm the pattern of results suggested by Fig. 9. The common spatial visualization clearly facilitated human performance in

Table 6
Pairwise probability that the distribution of accuracy, confidence (in italics), and time (in bold) is the same for each combination of raw data or common features visualization with local or global question types

	Raw–Local	Common–Local	Raw–Global	Common–Global
Raw–Local	0.95 *0.90* **0.90**	0.03 *0.01* **0.00**	0.00 *0.00* **0.32**	0.00 *0.79* **0.00**
Common–Local	0.03 *0.01* **0.00**	0.89 *0.89* **0.87**	0.00 *0.00* **0.00**	0.00 *0.00* **0.34**
Raw–Global	0.00 *0.00* **0.32**	0.00 *0.00* **0.00**	0.95 *0.95* **0.88**	0.03 *0.00* **0.00**
Common–Global	0.00 *0.79* **0.00**	0.00 *0.00* **0.34**	0.03 *0.00* **0.00**	0.96 *0.93* **0.92**

answering the questions, allowing for more accurate, more confident and faster answers than when people were made to rely on the raw data in a tabular form.

7. General discussion

The results of the three experiments are consistent with Lee and Vickers (1998) proposition that visualizations presenting the unprocessed raw data do not convey information as effectively as those that restructure data according to cognitive demands. 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 previously found by Morris et al. (2000) for Chernoff faces. 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.

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 et al., 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 setting, like stock markets and military environments, where analysts have significant training and practice interpreting the standardized glyph visualizations used in the domain.

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 by Chen and Yu (2000) demonstrated that simpler visual–spatial interfaces offer a performance advantage. 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.

An important question raised, but not 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 (1973, p. 366) has acknowledged that some display features are difficult to detect in some face visualizations, 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 et al., 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 and 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 from those of perceptual presentation. We would claim that the spatial visualizations used in this study are the canonical means of displaying MDS representations. It is possible, however, to argue that glyphs remain an effective visualization technique when displaying data that have 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 (1998) 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; Sattath and Tversky, 1977; Corter, 1996), additive clustering (Shepard and Arabie, 1979; Arabie and Carroll, 1980; Lee, 2002), trajectory mapping (Richards and Koenderink, 1995), and others (e.g. Tenenbaum et al., 2000; see also Shepard, 1980). 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 (e.g. Tversky and 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.

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 work was supported by the Australian Defence Science and Technology Organisation. We wish to thank Dimitris Margaritis, Thomas Minka and Chris Woodruff for their assistance and several anonymous reviewers for helpful comments.

Appendix A. Bayesian measures of independence

Our analysis relies on the following basic result from Bayesian statistics (e.g. Margaritis and Thrun, 2001). Let a data set $\mathbf{D} = (d_1, d_2, \ldots, d_K)$ contain the number of times each of K events was observed to occur out of N trials. Let a model M with K parameters $\theta_1, \theta_2, \ldots, \theta_K$ attempt to describe these data by estimating the underlying probability of each of the events with one of the θ_i parameters. The probability density that the data set will be observed, given particular values for the parameters under this model, follows the multinomial distribution:

$$p(\mathbf{D} \mid \theta_1, \theta_2, \ldots, \theta_K) = \binom{N}{\theta_1 \theta_2 \ldots \theta_K} \prod_k \theta_k^{d_k}. \tag{A.1}$$

Assume that the prior probabilities for each of the parameter values are given by the Dirichlet distribution, taking the form

$$p(\theta_1, \theta_2, \ldots, \theta_K) = \Gamma(\gamma) \prod_k \frac{\theta_k^{\gamma_k - 1}}{\Gamma(\gamma_k)}, \tag{A.2}$$

where each γ_i relates to a parameter, $\gamma = \sum_i \gamma_i$, and $\Gamma(\cdot)$ is the gamma function. This choice constitutes the conjugate prior distribution and enables the probability of the data given the model $p(\mathbf{D} \mid M)$, which must be integrated across all possible parameter values (see, for example, Kass and Raftery, 1995; Myung and Pitt, 1997), to be found exactly as

$$\begin{aligned} p(\mathbf{D} \mid M) &= \int p(\mathbf{D} \mid M) p(\theta_1, \theta_2, \ldots, \theta_K)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_K \\ &= \frac{\Gamma(\gamma)}{\Gamma(N + \gamma)} \prod_k \frac{\Gamma(\gamma_k + d_k)}{\Gamma(\gamma_k)}. \end{aligned} \tag{A.3}$$

This result can be applied directly to the problem of measuring the independence of two nominal variables. If the first variables has I possible values, and the second variable has J possible value, then the data take the form of an $I \times J$ matrix of counts $\mathbf{C} = [c_{ij}]$. The value of c_{ij} in this data matrix corresponds to the number of times the first variable has been observed to take its ith value in combination with the second variable taking its jth value.

If the two variables are independent, there is a suitable model $M_{\mathscr{I}}$ containing only $I + J$ parameters, one for each marginal total of the data matrix. Denote the Dirichlet priors for the I marginal totals corresponding to the first variable by $\alpha_1, \alpha_2, \ldots, \alpha_I$, and the J marginal totals for the second variable by $\beta_1, \beta_2, \ldots, \beta_J$. Eq. (A.3) allows the probability of the observed counts under the independence model to be calculated as

$$p(\mathbf{C} \mid M_{\mathscr{I}}) = \left(\frac{\Gamma(\sum_i \alpha_i)}{\Gamma(N + \sum_i \alpha_i)} \prod_i \frac{\Gamma(\alpha_i + \sum_j c_{ij})}{\Gamma(\alpha_i)} \right) \times \left(\frac{\Gamma(\sum_j \beta_j)}{\Gamma(N + \sum_j \beta_j)} \prod_j \frac{\Gamma(\beta_j + \sum_i c_{ij})}{\Gamma(\beta_j)} \right). \tag{A.4}$$

The alternative to a model that assumes marginal independence is one that allows for any sort of dependency, and uses a separate parameter for each cell in the data matrix. If the Dirichlet priors for each of these $I \times J$ parameters are given by γ_{ij}, then the probability of the data under this dependence model, $M_{\mathscr{D}}$, is given by

$$p(\mathbf{C} \mid M_{\mathscr{D}}) = \frac{\Gamma(\sum_{ij} \gamma_{ij})}{\Gamma(N + \sum_{ij} \gamma_{ij})} \prod_{ij} \frac{\Gamma(\gamma_{ij} + c_{ij})}{\Gamma(\gamma_{ij})}. \tag{A.5}$$

The independence model $M_{\mathscr{I}}$, and the full dependence model $M_{\mathscr{D}}$ are competing explanations for the observed data, whose relative merits may be quantified using Bayes' Theorem. In particular, the posterior probability of the independence model, given the observed data, may be calculated as

$$\begin{aligned} \Pr(M_{\mathscr{I}} \mid \mathbf{C}) &= \frac{p(\mathbf{C} \mid M_{\mathscr{I}}) p(M_{\mathscr{I}})}{p(\mathbf{C} \mid M_{\mathscr{I}}) p(M_{\mathscr{I}}) + p(\mathbf{C} \mid M_{\mathscr{D}}) p(M_{\mathscr{D}})} \\ &= 1 \Bigg/ \left[1 + \frac{p(M_{\mathscr{C}})}{p(M_{\mathscr{I}})} \frac{p(\mathbf{C} \mid M_{\mathscr{D}})}{p(\mathbf{C} \mid M_{\mathscr{I}})} \right], \end{aligned} \tag{A.6}$$

where $p(M_{\mathscr{I}})$ and $p(M_{\mathscr{D}})$ are the prior probabilities of independence and dependence, respectively. These priors are reasonably both set to the value 0.5, to express prior ignorance regarding the relationship between the variables, and allow the conclusions drawn to be maximally influenced by the observed data. It is also reasonable to set all of the hyper-parameters, α_i, β_j and γ_{ij} to the value 1, since this makes the Dirichlet prior a uniform distribution, and again allows the data to be maximally informative.

With these assumptions about the priors in place, Eq. (A.6) allows the posterior probability of independence to be assessed for any two nominally scaled variables. For example, the independence of the binary accuracy variable can be assessed in relation to the four-valued nominal visualization variable. When one or both of the variables exists at an ordinal, interval, or ratio level of scaling, however, a more sophisticated approach is required. This is because, when a variable has a natural ordering, it is necessary to consider the many count matrices that can reasonably be derived from the observed data by combining adjacent values of that variable.

As a concrete example, consider the relationship between visualization type and the time performance measure. It is not appropriate to treat the time data solely as having been generated by a multinomial distribution, because time values of, say, 10 and 11 s reflect a great underlying level of similarity than, say, time values of 10 and 90 s. The approach adopted by Margaritis and Thrun (2001) is to consider the possible adjacent groupings of these 'continuous' variables, searching for those that reveal dependencies. The number of groupings is combinatorially large, and so an efficient optimization approach is used to estimate the true posterior probability of independence. Because the approach is embedded within a Bayesian framework, and makes use of uniform priors, it is able to control for the complexity effects that arise from considering many groupings simultaneously. It is also sensitive to the prior probabilities of the different groupings considered so that, for example, a grouping that splits the 10 and 11 s data into separate counts is far less likely that one that splits the 10 and 90 s data.

Finally, note that the ability to assess the posterior probability of independence, which is exact in the purely nominal variable case, and approximate when a continuous variable is involved, allows the posterior probability that two distributions are the same to be assessed. This is simply a matter of associating the data from the two distributions with binary class labels, and measuring the probability that the data are independent of their labels.

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, 2nd Edition. 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 & 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.

Cohen, J., 1994. The earth is round ($p<0.05$). 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 & 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.

Gati, I., Tversky, A., 1984. Weighting common and distinctive features in perceptual and conceptual judgments. Cognitive Psychology 16, 341–370.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & 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.

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 simple method for generating additive clustering models with limited complexity. Machine Learning 49, 39–58.

Lee, M.D., Navarro, D.J., 2002. Extending the ALCOVE model of category learning to featural stimulus domains. Psychonomic Bulletin & Review 9 (1), 43–58.

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, Santa Monica, CA. Human Factors Society, San Francisco, CA, pp. 345–349.

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

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 & 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.

Margaritis, D., Thrun, S., 2001. A Bayesian multiresolution independence test for continuous variables. In: Breese, J., Koller, D. (Eds.), UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, San Francisco, CA. Morgan Kauffman, Los Altos, CA.

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: Oliver, W.R. (Ed.), 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making, Proceedings of SPIE, Vol. 3905, pp. 12–17.

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

Myung, I.J., Pitt, M.A., 1997. Applying Occam's razor in modeling cognition: a Bayesian approach. Psychonomic Bulletin & Review 4 (1), 79–95.

NetMap Computer Software, 2001. Netmap Analytics, LLC. Westerville, OH.

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, Adelaide, Australia. 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.

Ritov, I., Gati, I., Tversky, A., 1990. Differential weighting of common and distinctive components. Journal of Experimental Psychology: General 119 (1), 30–41.

Robert, C.P., 1994. The Bayesian Choice. Springer, New York.

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.

Sattath, S., Tversky, A., 1987. On the relation between common and distinctive feature models. Psychological Review 94 (1), 16–22.

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.

Shneiderman, B., 1998. Designing the User Interface: Strategies for Effective Human–Computer Interaction, 3rd Edition. Addison Wesley, Reading, MA.

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, 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, Los Alamitos, CA, 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.