Towards Understanding Human Similarity Perception in the Analysis of Large Sets of Scatter Plots

Anshul Vikram Pandey, Josua Krause, Cristian Felix, Jeremy Boy, Enrico Bertini

Tandon School of Engineering
New York University
{anshul.pandey, josua.krause, cristian.felix, jeremy.boy, enrico.bertini}@nyu.edu

ABSTRACT

We present a study aimed at understanding how human observers judge scatter plot similarity when presented with a large set of iconic scatter plot representations. The work we present involves 18 participants with a scientific background in a similarity perception study. The study asks participants to group a carefully selected set of plots according to their subjective perceptual judgement of similarity, and it integrates the results into a consensus similarity grouping. We then use this consensus grouping to generate insights on similarity perception. The main output of this work is a list of concepts we derive to describe major perceptual features, and a description of how these concepts relate and rank. We also evaluate scagnostics (scatter plot diagnostics), a popular and established set of scatter plot descriptors, and show that they do not reliably reproduce our participants judgements. Finally, we discuss the major implications of this study and how these results can be used for future research.

Author Keywords

Information Visualization; Human Perception Modeling; Quality Measures; Plot Similarity

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

In recent years, we have witnessed an increase in the use of visualization techniques for exploratory data analysis of complex multidimensional datasets. Scientists and researchers from areas as disparate as healthcare, business analytics, or climate science, rely on interactive visualization to familiarize with these complex data spaces and to generate new insights.

Often, in the early stages of data analysis, scientists are interested in exploring all possible associations that may exist in a given dataset and for this reason generate and inspect the whole set of scatter plots obtained from all possible pairs of dimensions. When the number of dimensions grows to tens, hundreds, or even thousands of variables, being able to inspect all these combinations becomes extremely time consuming, if not impossible, and computational approaches are required. To overcome this problem, recent research has focused on developing techniques that guide data analysis using a mix of interaction and statistical measures.

The basic idea behind these techniques is to develop *measures* able to detect plots that contain one or more patterns of interest and to use these measures as a way to guide their navigation and exploration. Numerous techniques have been developed in this area. For instance, Wilkinson *et al.*, developed *graph-theoretic scagnostics*, descriptors that characterize scatter plots according to, *e.g.,* how *skinny*, *clumpy* or *striated* they are [2]. Tatu *et al.*, developed a number of image processing measures to quantify degrees of association and class-separation, that is, how well separated the (colored) classes of a scatter plot are [32]. Sips *et al.*, developed *class-consistency* measures to rank scatter plots according to their class-separation [30]. And Reshef *et al.*, developed the *maximal information coefficient (MIC)* score, to detect linear and non-linear associations between pairs of variables [23].

While these measures represent an important step towards providing support for the aforementioned task, empirical research on how human observers and analysts perceive large sets of plots and their patterns, has received, in comparison, little attention. A few empirical works exist on the validation of existing measures, e.g., [15, 16, 27], but only a few focus on understanding what visual patterns people extract out of plots and how these patterns are used for plot comparison.

In this paper, we focus specifically on understanding how human observers gauge *plot similarity* and what visual features drive similarity perception. Plot similarity is a fundamental task in information visualization as many visual operations often reduce to comparing visual objects. For instance, the widely popular *small multiples* techniques described by Tufte [33], requires the observer to compare a large quantity of complex images or plots arranged in a grid. Similarly, visualizations based on iconic representations (also called glyphs [12]) often require comparing a large set of small and complex graphical objects across the view field.

We describe a study we conducted to capture information about what visual features drive similarity perception in scatter plots. In order to simulate the exploratory set-up we described above, we study similarity perception with a multitude of small scatter plots, under the assumption that this

CHI'16, May 07-12, 2016, San Jose, CA, USA
©2016 ACM. ISBN 978-1-4503-3362-7/16/05$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858155

truthfully represents real conditions analysts may face. The study also, for ecological validity, involves only participants with a scientific background, under the assumption that the analysis of large sets of associations (scatter plots) represents a highly specific task performed only by skilled data analysts.

The study asks participants to group, through a dedicated user interface, a carefully crafted set of plots together into a series of groups according to their subjective similarity judgment. These judgments are then coded, grouped and analyzed together to generate a number of concepts that describe what visual features drive similarity perception. These results are then used as a comparison to the well-established *scagnostics* measures to show that they are not able to reliably reproduce the groupings generated by human observers. Finally, we provide a description of how these results can be used in practice and what major implications can be derived from them.

In summary, the main contribution of this work is the characterization of similarity perception in scatter plots, the generation of a number of perceptually-derived concepts that describe scatter plots and their comparison, the evaluation of a popular plot similarity method (*scagnostics*), and the description of how these results can be used for future research.

This article is organized as follows. We start by introducing the related literature on *visual similarity perception* and *visual quality measures*. Next, we introduce the methodology that includes design decisions and assumptions, followed by the description to the developed visual interface to facilitate plot similarity tasks and details of the study. Further, we analyze the recorded *perceptual similarity* to uncover the descriptive concepts and correlate them with the known *scagnostics* measure. Lastly, we discuss the implications of this work, its possible integration into visual analytics systems for guided navigation, and lay down the roadmap for future works.

RELATED WORK

Studying Similarity Perception

We are not aware of any major study on plot similarity. However, Rogowitz *et al.* have studied photograph similarity [24] using 97 digitized photographic images from a library of 5000 pictures. Wei *et al.* have compared 200 mammograms to understand how radiologists compare "microcalcification clusters" [35]. Long *et al.* have compared tens of pen gestures [20] to understand what stroke features have an impact on similarity. Wills *et al.* have compared 55 3D images of the Stanford bunny (a classic benchmark image used in computer graphics) under different illumination conditions to understand gloss perception [37]. Demiralp *et al.* [10] studied perceived similarity of color, shape and size values and derived perceptually adjusted similarity measures they call *perceptual kernels*.

While these studies come from very different areas of research, they all share a common template. A group of human subjects is presented with a collection of images, and is asked to perform a comparison judgement between them. This information is then analyzed through *multidimensional scaling* [6], which allows to transform similarity data into a visualizable 2D embedding for understanding similarity judgment. In our study, we follow a similar template: we collect a representative set of plots, we subject it to human judgment, and we analyze the results.

Other research has focused on modeling the perception of patterns in charts. Rensink *et al.* and van Wijk *et al.* have respectively created models that capture how people judge correlations in scatter plots and parallel coordinate plots [19, 22]. Harrison *et al.* [13] have ranked visualizations of correlations based on Weber's law. Similarly, Albuquerque *et al.* have introduced a generalizable method to create *perceptual visual quality measures* for a pair of *perceptual tasks* and *plot types* (*e.g.,* correlation in scatter plots) [1].

Visual Quality Measures

The term 'quality measure' has been used extensively in visualization research to qualify the general concept of measuring the quality of a visual encoding through computational methods. Quality measures have been developed for a multitude of visualization techniques. There are essentially three ways in which quality measures have been used: 1) detecting optimal levels of data abstraction (*e.g.,* through sampling and aggregation) [4, 7, 17]; 2) finding optimal ordering in visualizations that allow multiple axis-ordering (*e.g.,* in parallel coordinate plots and scatter plot matrices) [3, 21]; and 3) finding projections of high-dimensional data that contain patterns that end-users may be interested in [9, 11, 26, 30, 32, 36].

Scatter plots have received by far the most attention. Several 'quality metrics' have been developed to describe or characterize their appearance or quality; these focus on visual features like clusters and outliers [11, 18], correlations [29], and class-separations [28, 30, 32], or on more complex patterns. Anand & Wilkinson's *graph-theoretic scagnostics* [36] can describe and measure complex scatter plots according to a multitude of descriptive features: *outlying*, *skewed*, *clumpy*, *convex*, *skinny*, *striated*, *stringy*, *straight*, *monotonic*.

STUDYING PERCEPTUAL SIMILARITY OF PLOTS

To study plot similarity, we need to: 1) generate a representative sample of plots that contain a sufficiently broad set of patterns to produce a meaningful characterization; and 2) find a reliable mechanism for capturing human similarity judgments. In this section, we describe our approach, and the six main steps we followed to address these two needs (Figure 1).

Data Generation

The data generation phase consists of the following steps: (1) selecting datasets; (2) generating scatter plots; and (3) reducing the number of plots to ensure a manageable human study.

Dataset Selection (1)

For our study, we collected 717 datasets from the R statistical analysis software [1] and from our lab data archive. These datasets cover a broad range of real-world data of different sizes, dimensionality, and patterns. We chose to avoid synthetic datasets becasue they often contain patterns that are

[1] More information at: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

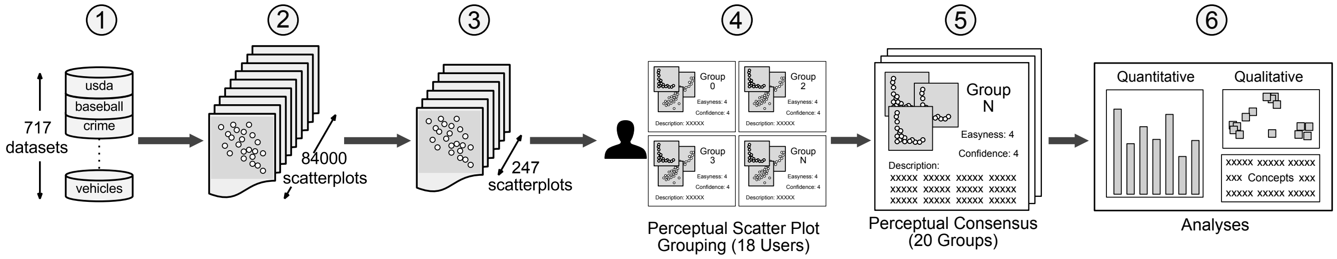

Figure 1. The six stages of methodology to study perceptual similarity of scatter plots. The first three stages are grouped under a broader phase of *data generation*, while the latter three are grouped under *human judgment collection* phase.

never (or very rarely) encountered in real-world data analysis. This, however, means that our study differs from those presented by others, like Sedlmair *et al.* and Wilkinson *et al.*'s studies [28, 36], in which, *e.g.,* perfectly separated/uniform clusters and perfect sinusoidal shapes appear.

Scatter Plot Generation (2)

For each dataset selected in step (1), we created scatter plots for every unique combination of dimensions—except combinations of identical dimensions. Thus, for N datasets selected in step (1), we obtain $\sum_{i=1}^{N} \frac{d_i(d_i-1)}{2}$, where d_i is the number of dimensions in the i^{th} dataset. For the visual encoding, we used 2 pixel diameter blue data points with 0.2 alpha channel in a 47×47 pixels canvas generated using matplotlib [14]. We selected these values empirically trying to find a good balance between plot size, the range of data densities and pattern visibility. Altogether, in this step we generated over $84,000$ scatter plots.

Selecting Scatter Plot Stimuli (3)

Running a study in which each participant has to analyze and compare over $84,000$ scatter plots (which we refer to as *original space*) is clearly unfeasible. Not only it is very hard to analyze such a high number of plots individually (as has been done in the work of Sedlmair *et al.* [28]) but, given the comparative nature of our similarity task, the study would require participants to make an enormous number of pairwise comparisons. Therefore, we needed a method to systematically reduce the number of plots to a manageable size without losing too much information. To this purpose, we used a combination of computational and manual approaches using *diversity* as the guiding principle, that is, the more diverse the final set of plots is the more we can capture important and interesting perceptual phenomena.

In order to maximize diversity in a principled manner, we developed a three stage stimuli selection method; the first two stages of which use *scagnostics* (computational approach), while the third is based on visual and manual inspection. We decided to use *scagnostics* because they are the *de facto* standard for scatter plot characterization, and also because they allow us to compare the results of our study to the most popular, but not perceptually validated, technique.

The first stage of our selection method consists of sampling plots by binning nine of the original *scagnostic* measures *independently*. The goal is to sample plots uniformly across each measure. For each *scagnostic* measure (range: 0-1), we created ten uniform bins ($[0, 0.1), [0.1 - 0.2), ..., [0.9 - 1.0]$), grouped the original set of plots according to these bins, and randomly sampled one scatter plot for each one. By sampling at least one scatter plot for each bin across all measures (avoiding duplications) we extracted 90 plots from the original space.

The second stage of our selection method consists of sampling the plots by considering all *scagnostic* measures together. That is, rather than binning the measures independently, we aim at sampling the multidimensional space described by the measures directly. To this purpose, we run a k-means clustering algorithm on the *original space* and grouped the plots into $k = 100$ clusters (arbitrarily chosen and motivated by [35]). From each cluster we then randomly sampled 100 plot for a total of $10,000$ candidates (100 plots per cluster, for 100 clusters).

In the third and final stage, starting from the plots obtained from the computational steps, we agreed on selecting a manageable number of plots (in the order of one or two hundreds) while ensuring as much diversity as possible. For each cluster and bin used in the previous steps, we discarded plots that looked too similar and kept as many unique patterns as possible. In the end, after visual inspection of all $10,090$ scatter plots, we extracted a final stimuli sample of 247 plots.

It is important to point out that such a sampling scheme does not represent a uniform and representative sample of all possible scatter plot trends and characteristics. The three steps we described have been devised exclusively to ensure as much diversity as possible, under the assumption that more diversity increases the chances of capturing relevant perceptual phenomena. Unfortunately, there is no established method to ensure that all possible key patterns are included in the stimuli. Also, for practical purposes one must necessarily limit the number of sample plots to a manageable number. Alternative sampling methods may have been used, such as uniform sampling over the whole set of plots or different plot descriptors. Uniform sampling however does not take into account the possible skewedness of plot feature distributions, and alternative descriptors, as we pointed out above, does not seem to be available for the specific task of plot similarity.

Human Judgment Collection

The human judgment collection phase consists of the following steps: (4) grouping scatter plots according to perceptual judgement; (5) building perceptual consensus; and (6) analyzing human observer consensus using quantitative and qualitative methods.

Perceptual Scatter Plot Grouping (4)

In order to capture similarity judgements, we decided to use *open card sorting*. Card sorting consists of asking a group of human subjects to organize samples (physical or digital) into a set of categories or groups that make sense to them (that is, that capture their subjective intuition of similarity). Such technique has been widely used in designing information architectures, workflows, menu structures and in several psychology experiments, to reveal 'mental models' from a group of human subjects [31] and it is an effective way to learn about how users group, label, and prioritize information. In our study, we therefore presented the participants with the whole set of plots and asked them to group them according to their perceived similarity.

It is important to point out that we intentionally did not define or describe similarity to our participants, as capturing and understanding their subjective perception of similarity is one of the major goals of this work. While providing more precise tasks related to the detection of specific patterns is also interesting and important, in this work, similarly to what others have done in the past [1, 24, 38], we focus exclusively on subjective perception of similarity.

In order to better characterize this process, we also explicitly asked the participants to answer the following questions for each group they created: a) "On a scale of 1-5 (1 = very difficult, 2 = difficult, 3 = neutral, 4 = easy, 5 = very easy), answer, how difficult or easy was it for you to create this group?", and b) "On a scale of 1-5 (1 = very doubtful, 2 = doubtful, 3 = neutral, 4 = confident, 5 = very confident), answer, how doubtful or confident are you about the consistency of the plots in the group, *i.e.,* you would create the same group if you work with the interface again?" [2]

Building Perceptual Consensus (5)

Once data about each participant has been collected, it is necessary to bring individual results together into a unified description of plot similarity able to capture the main perceptual phenomena across all participants. To this purpose, we use the following data elements that capture the data collected in the experiment: a) a perceptual similarity distance matrix; b) an easiness array; and c) a confidence array.

A perceptual distance matrix (PDM) is created for each participant by computing the pairwise similarity between plots, based on their occurrence in the participant-generated groups. The calculated perceptual distance scores between each pair of plots are put into a 247 $*$ 247 matrix, where cells are computed using Equation 1.

$$v_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\left(1 - \frac{c_{i,j}}{min(c_i, c_j)}\right)_k \quad (1)$$

$v_{i,j}$ represents the perceptual distance score between plots i and j; N is the total number of participants (18) in the study; $c_{i,j}$ is the number of clusters/groups that contain both plots, whereas, c_i and c_j are the number of clusters/groups that contain plots i and j, respectively. The values of $c_{i,j}$, c_i, and c_j vary for each participant k.

The easiness and confidence arrays require a slightly different approach. Since the responses to the *easiness* and *confidence* questions are related to the groups (and not directly to the plots), we assume that the plots within a group can inherit the *easiness* and *confidence* values from the groups they belong to. For instance, if a participant can easily create a group 'A' (*e.g.,* easiness score = 4), and is very confident about the consistency of the plots within the group (*e.g.,* confidence score = 5), we assume that all plots in 'A' can inherit the same easiness and confidence scores. With this assumption, the consensual easiness and confidence scores are computed for each plot using Equation 2.

$$y_i = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{c_i}\sum_{c=1}^{c_i} x_{i,c}\right)_k \quad (2)$$

y_i represents the consensual easiness or confidence score for plot i ; N is the total number of participants (18 in our study) in the study; c_i is the number of clusters/groups containing plot i that participant k creates; and $x_{i,c}$ is the easiness or confidence score given by the user to the cluster/group c containing plot i.

Finally, for each cluster (or group of plots), we can compute the average easiness and confidence scores using the Equation 3, described below.

$$Cy_o = \frac{1}{N_o}\sum_{i=1}^{N_o} y_i = \frac{1}{N_o}\sum_{i=1}^{N_o}\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{c_i}\sum_{c=1}^{c_i} x_{i,c}\right)_k\right) \quad (3)$$

Cy_o is the average easiness or confidence score for cluster o and N_o is the total number of plots that belong to cluster o. We substitute y_i with the expression from Equation 2. The definition of other variables in the equation is similar to that in Equation 2.

Analyzing User Consensus (6)

The two main questions behind our study are: a) *What are the most dominant dimensions or plot features that can be used to model human perception in scatter plot similarity tasks?*, and b) *How does perceived similarity correlate with some of the existing measures, such as graph-theoretic scagnostics?* To answer these questions, in this final step, we use the perceptual distance matrix (PDM) to find consensus grouping of plots and analyze the consensus groups to identify dominant visual elements that play a role in similarity perception. This same information is also used to compare perceptual similarity to similarity computed with the *scagnostic* measures.

STUDY DESIGN

In this section, we highlight the various components of the study design that incorporate the design decisions mentioned

[2] Further details about the study are provided in the later sections.

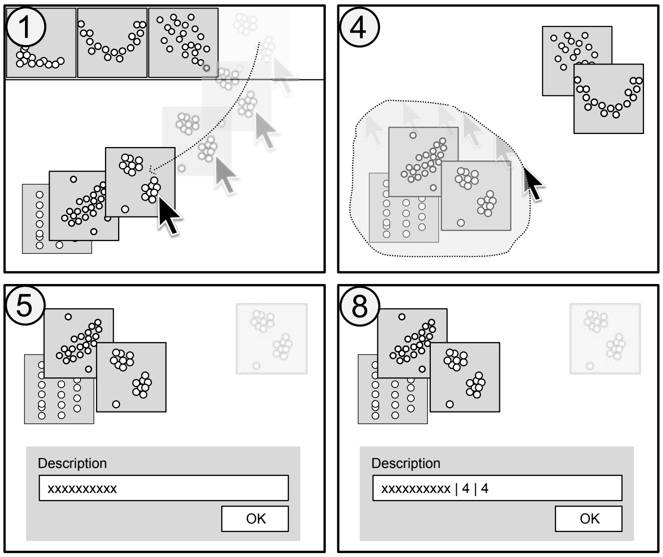

Figure 2. Primary user tasks and interactions – (1) drag-and-drop interaction to arrange plots, (4) lasso-selection to create plot groups, (5) editable text area to provide description for each plot group, (8) append easiness and confidence scores to the text area.

earlier. We start by explaining the details of the user interface including the various user tasks and interface features; we then discuss the pilot studies we conducted to refine the interface; and we finish by describing the human study.

User Interface Design

We develop an open card sorting interface, as shown in Figure 3, that allows the users to group the 247 scatter plots collected in step (3). Based on the feedback from a pilot study conducted with 5 participants using a *think-aloud* protocol, we identify a series of primary (tasks 1, 4, 5, and 8) and secondary user tasks (tasks 2, 3, 6, and 7) along with the interface features and interactions to support the tasks. The following user tasks and supporting interface features/interaction constitute our final interface design.

Tasks: (1) arrange plots based on their similarity, (2) move multiple plots, (3) flag/unflag plots to review later (4) create plot groups, (5) provide description for each plot group, (6) edit group descriptions, (7) delete plot groups, (8) provide easiness and confidence scores for each plot group.

Interface features and interaction: (1) drag-and-drop interaction to arrange plots, (2) lasso-selection to move multiple plots, (3) single-click to flag/unflag plots for later review, (4) lasso-selection to create plot groups, (5-6) editable text area to provide description for each plot group, (7) delete text area to delete plot groups, (8) append easiness and confidence scores to the text area.

We segment the user tasks into two phases, *positioning* and *naming*. In the *positioning* phase, the users perform tasks 1-3, *i.e.,* spatially arrange plots based on their similarity, move multiple plots, and flag/unflag plots for later review. The interface supports interactions 1-3 in this phase. We randomly arrange all 247 plots in a scrollable list (left-right) on top of the interface, as shown in Figure 3(a). Users can drag-and-drop plots from the list onto the canvas, shown in Figure 3(b). Using lasso-selection, users can select and move multiple plots. Users can single-click on the plot to flag/unflag. The 'Freeze' button at the bottom only activates once the users have placed all the plots from the list onto the canvas. Once the 'Freeze' button is clicked, the users can not reposition the plots. This was done to restrict users from re-arranging plots into groups that are easier to describe in order to decrease the cognitive effort involved - a behavior we observed in our pilot study.

In the *naming* phase, the users perform tasks 4-8, and the interface supports interactions 4-8. Users can use the lasso-selection to create plot groups. Users can assign multiple plots to one group, and conversely, plots can share multiple groups based on different grouping criteria. This also allows the users to create hierarchical and alternate groups. For each group the user creates, the interface prompts a text area to provide a textual description. The description is then placed at the centroid of all plots. User can click on the description text to edit it, or remove the text completely to delete the plot group. The easiness and confidence scores can be provided in the same text box, delimited by a pipe '|' character. Hovering on the description text highlights all the plots belonging to that group. Clicking on 'Finish' records the user's response on the server, including the start and end time of the study, and group information – plot ids, plot positions, text description, easiness and confidence scores.

Figure 2 shows various user tasks and corresponding interactions in the *positioning* and *naming* phases.

User Study

In this section, we discuss the details of the final user study conducted using the improved interface. Figure 4 shows various stages of the user study.

Participants and Apparatus

One of the main recruitment criteria is experience with data analysis, and familiarity with scatter plots. We recruit 18 participants (15 males, 3 females) with diverse levels of education (undergraduate, graduate, post-graduate) and disciplines (computer/electrical/mechanical engineering, design, management). The recruitment is done using fliers and personal messages to data analysts. Some of the participants are associated with an academic organization, while rest come from industry. All the participants have actively engaged in data analysis tasks and are familiar with scatter plots. The average age of the participants is 27.3 years. Each participant is paid $5 dollars or gift-cards of equivalent amount for the participation. The studies are conducted in the lab setup using a 27-inch monitor with full-HD resolution (1920 × 1080).

Procedure

Once a candidate is short-listed, *i.e.,* satisfies the recruitment criteria, a confirmatory email/message is sent scheduling the user study. When the participant agrees, a consent form is sent, describing various details of the study, including what we will be recording and how they will be compensated.

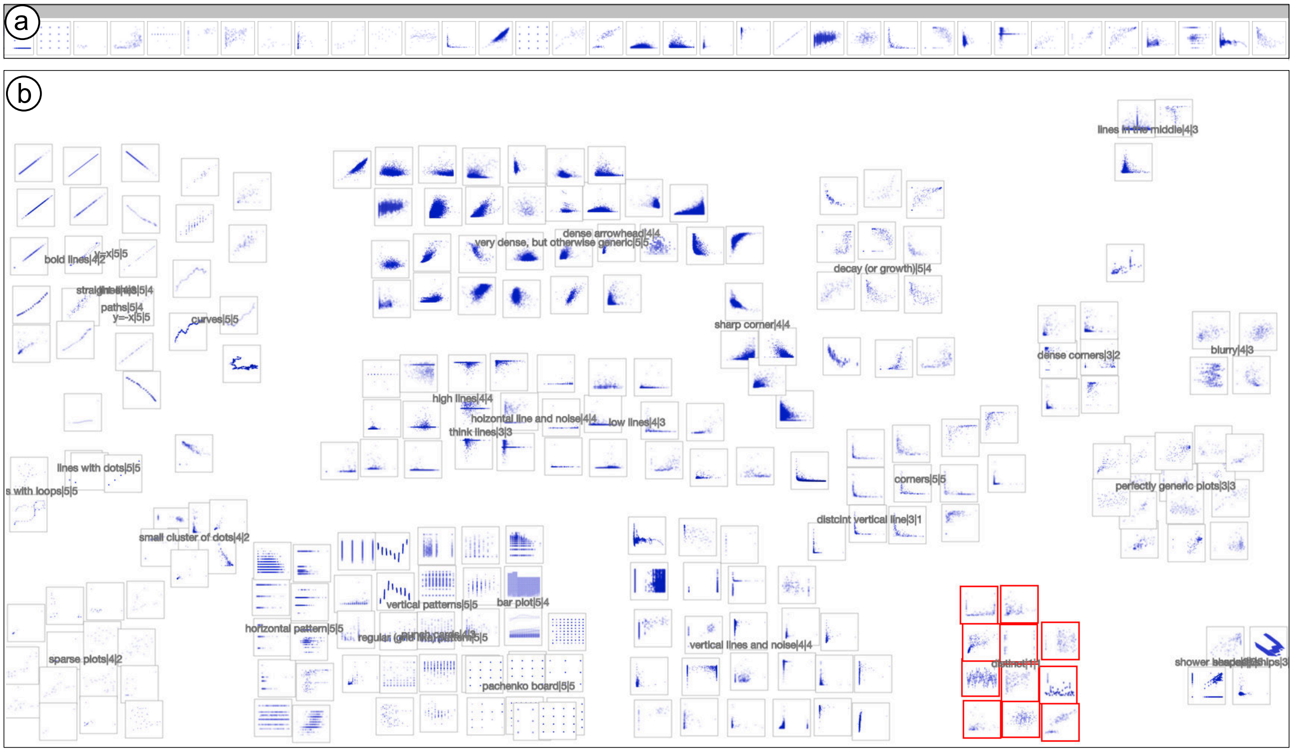

Figure 3. The open coding interface to capture perceptual similarity and quantitative values. (a) shows a list of randomly arranged plots visible only in the *positioning* phase. (b) shows an example final screen at the end of the *naming* phase with group descriptions, easiness and confidence scores separated by a pipe character.

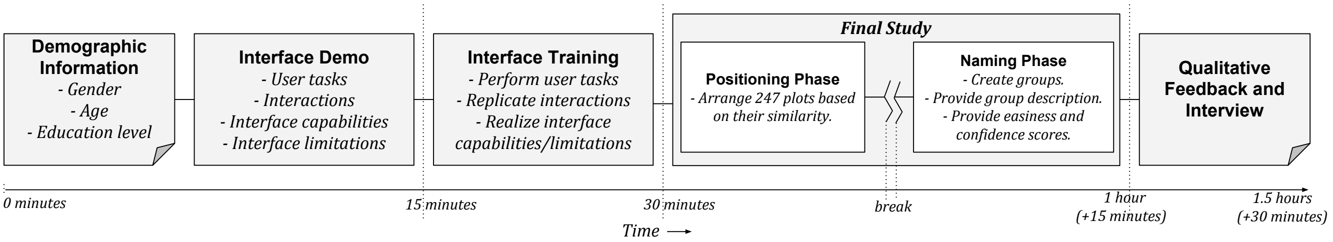

Figure 4. User study pipeline showing various steps with main highlights that users go through during the experiment. The timeline at the bottom shows an approximate time taken by the users to complete the step (*from the start of the study*).

Upon consent, we describe the various stages of the study and the expectations. Then, the user is given a form to provide demographic information, such as gender, age and education level. Next, one of the co-authors provide a 5-10 minutes demonstration of the user interface while explaining the user tasks, expectations, interface capabilities and limitations. The interface is loaded with a dummy dataset with 14 multi-class scatter plot matrices images in order to avoid biases introduced by the guided demo phase. The actual study involves only single-class scatter plots. We also emphasize that we are only interested in the groups that they create and the plots within the group, and not in the spatial distances between plots or groups. More precisely, it is fine if they place two very dissimilar groups next to each other or similar looking groups away from each other.

The users are also told that they can not re-position plots once they click on the 'Freeze' button at the end of the *positioning* phase, so they should take a break and review the arrangements before clicking it. Similarly, they are told they can not update the groups or their descriptions once they click on the 'Finish' at the end of the *naming* phase, so they should review the groups before finishing the study.

Upon completion of the demo, we load the interface with the same 14 scatter plot matrices that are used for the demo and start the training step. During this 5-10 minutes long training step, users try to replicate the tasks and are free to ask questions and seek clarification. In the meantime, the investigator monitors the activity, guides the user if they are stuck, and deliberately prompts the user to perform certain tasks such as moving multiple plots, which may come handy at a later stage. The participants are reminded to review the arrangements and groups when they are about to complete the *positioning* or *naming* phases. Once the participants are acquainted with the tasks and the interface, we finish the training. Lastly, they are told that they have a choice to not group plots that look completely arbitrary or random, or can group them using the keyword ‘distinct’, which would mean that the plots don’t look similar to any other plot on the canvas.

Next, the interface is loaded with the 247 stimuli samples, starting the final study. The investigator leaves the study area to minimize interruptions and biases, while monitoring the area from a distance. Under doubt or willing to seek clarification, the participants can get up as a signal for help. The investigator follows up, clarifying any doubts. Upon the end of the study, the responses are recorded and participant is compensated. The final study takes between 45 minutes to 1 hour.

As the last step, we collect qualitative feedback from the participants, asking about the strategies they adopted to group the plots. More precisely, we ask why they created some groups first while others at the last. We also have an open-ended discussion around the tasks and patterns that appear on the plots. This qualitative feedback session takes 10-15 minutes. The duration of the study for each participant, from providing demographic information till the end of qualitative feedback, is approximately 1.5 hours.

RESULTS AND FINDINGS

We run quantitative as well as qualitative analyses on the collected responses in order to answer the primary set of questions we are interested in. The analyses incorporated the assumptions we laid out earlier, like inheritance of easiness and confidence scores from groups to plots. We also ignore plots that are grouped under ‘distinct’ by 9 or more users. These are plots that look very dissimilar to other plots. In this section, we discuss the quantitative and qualitative analyses conducted on the collected responses.

Quantitative Analysis

We segment the quantitative analysis into two parts - a) correlation between perceived similarity and *scagnostics*-based similarity, and b) extraction and analysis of common clusters based on user consensus.

Perceived Similarity vs. Scagnostics-based Similarity

First, we describe each plot from the 247 stimuli sample using 9 *scagnostics* measures computed using R’s *scagnostics* library [34]. Next, we compute pairwise Euclidean distance between each plot to build a *scagnostics* distance matrix (SDM). We then extract distance between unique pairs (A-B and B-A are considered same) from the matrix and analyze correlation. As anticipated earlier, we find very weak correlation (Pearson’s $r < 0.26$) between the perceived distances and scagnostics-based distances. We create a MDS projection of the SDM. We could not find any distinct clusters in the projection. Next, we run hierarchical clustering and test a broad range of cluster numbers. Although most clusters are inconsistent, we do find that striated (or line-like) plots are grouped consistently in the same cluster. We don’t find a group of consistent clusters comparable to those we obtain from the PDM (discussed below), shown in Figure 5.

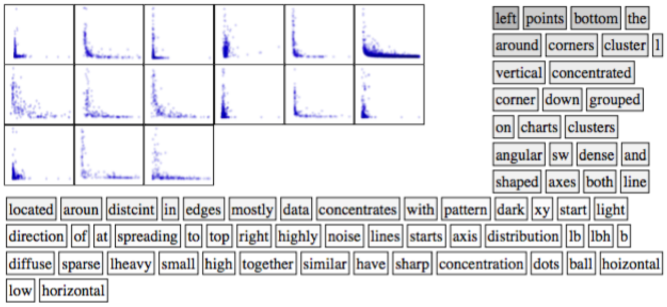

Figure 6. Dominant terms in a cluster based on their frequency across master descriptions of each plot.

Extraction and Analysis of Consensus Clusters

As the first step to quantitatively analyze the collected responses, we use the PDM created earlier to identify clusters of similar plots. We run hierarchical clustering on the distance matrix in order to create clusters. Figure 5 shows the hierarchical clusters created in the process of this analysis. To identify the right value for the number of clusters, we compute the average number of groups created by the users. The mean of the number of groups created by the users is 24 with 95% confidence interval in $[18.54, 23.46]$. However, as often users also create alternate groups, we decide to use $k = 20$. Next, we cut the hierarchical cluster tree to extract out 20 clusters. We find that two clusters (9 and 19 in Figure 5) comprised of plots that are tagged as ‘distinct’ by more than 9 users. Therefore, we reject these two clusters and the plots grouped within these clusters, from our analyses.

Using the consensual easiness and confidence values for the remaining plots obtained using Equation 2, we find that the easiness and confidence scores are highly correlated with Pearson’s $r = 0.97$. The average easiness and confidence scores for each cluster along with their 95% confidence intervals are presented in Table 1.

Qualitative Analysis

To qualitatively analyze the clusters obtained using hierarchical clustering, it is important to associate (or derive) the textual descriptions users provided for each group to the clusters. We adopt the same inheritance assumption as earlier (for easiness and confidence scores), *i.e.,* plots can inherit the textual description users associate to the group they are part of. First, we create master description for each plot that combines textual descriptions provided by all users. Next, we create 20 description files, one for each cluster, with images of plots belonging to the cluster and corresponding master descriptions. Additionally, we extract the dominant terms to describe each cluster, *i.e.,* most frequently occurring terms from the master descriptions of all plots contained in the cluster.

An example of dominant term analysis step is shown in Figure 6. Note that we do not need to sanitize the terms using stemming, lemmatization, etc., in hope to capture higher granularity of perception. We also exclude clusters 9 and 19 from the analyses, as they contain highly dissimilar plots.

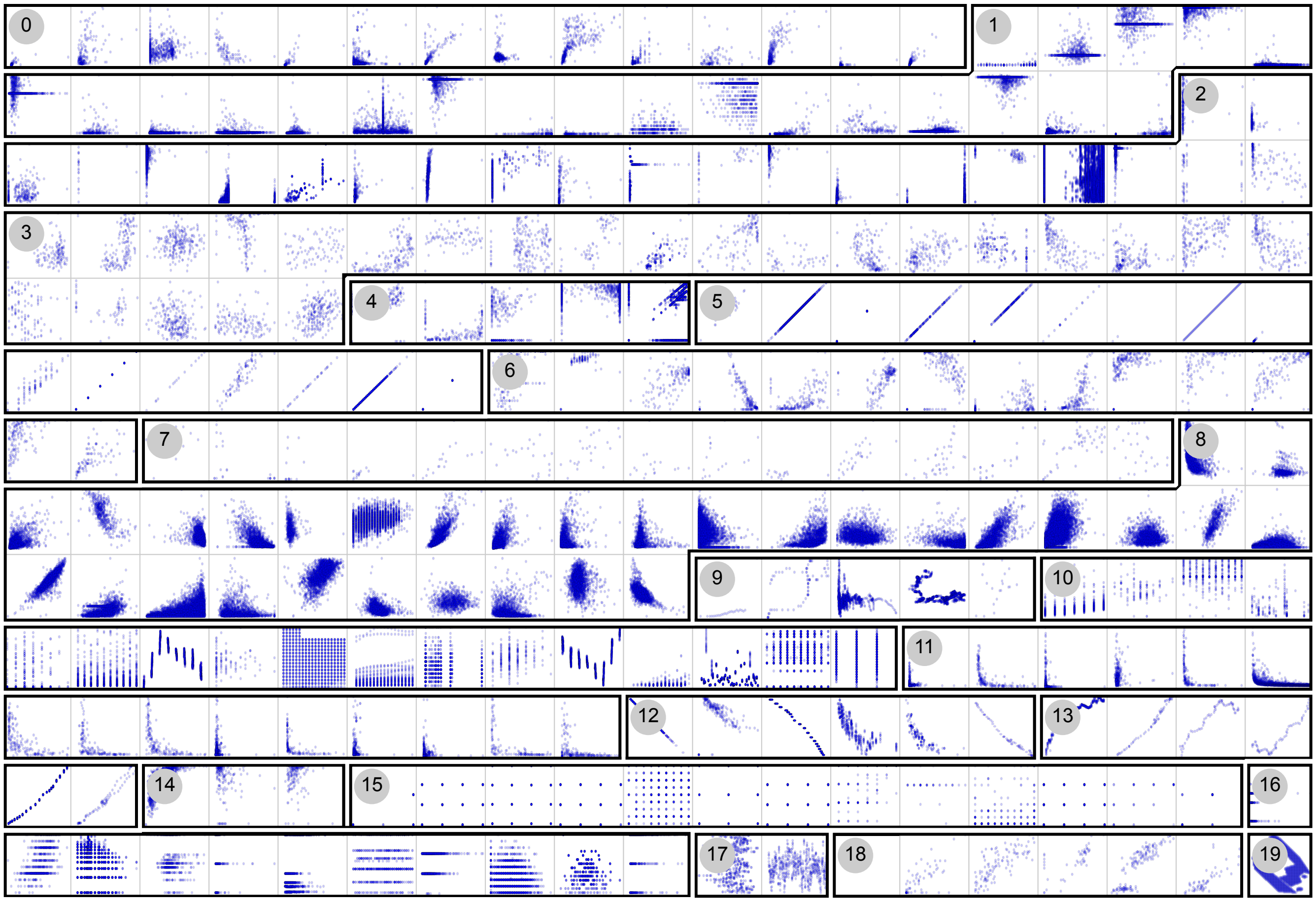

Figure 5. Showing all 247 plots used in the studies, grouped into 20 clusters extracted using hierarchical clustering approach. The plots propagate from left-to-right, top-to-bottom. Each cluster is assigned a unique ID.

We then group these terms into common concepts, *e.g.,* horizontal, vertical, inclined are grouped into 'orientation'. At the end of the open-coding process, we identify 6 common concepts that can be used as a vocabulary to describe similarity between scatter plots, and are as follows: 1) *density*, 2) *orientation*, 3) *spread*, 4) *regularity*, 5) *groupings*, 6) *edges*.

Finally, we plot the scatter plots using *multidimensional scaling* projection of the PDM. We also associate the cluster information to each plot, *i.e.,* which cluster the plot belongs to, by adding anchors on the plot colored by cluster ID. Through visual inspection, we identify the most compact clusters and dominant axes aligning certain concepts. By combining this information with the concepts identified, we describe each cluster using a set of dominant concepts (Table 1).

We now describe each concept identified through the process described above, while referring to certain clusters in Figure 5 as examples showing these concepts in action. In addition, we discuss the common terms and phrases that are used by users to describe the plots in those clusters.

1) *Density*: Density refers to the concentration of data points in certain region of the plot and can vary from high-density to low-density. We find that among plots that are grouped based on high density, there exists a high variance with respect to their visual appearance (refer to cluster 8 in Figure 5). In other words, regardless of how the shapes of the plots vary, as long as there exists a high density pattern, the plots are often grouped together. This is especially interesting as it indicates that density may have higher impact on users' perception of plot similarity. Common terms used by users to describe density are 'thick', 'sparse', 'concentrated' etc.

2) *Orientation*: Orientation is described by the data distribution across the two axes of a scatter plot and can be 'horizontal', 'vertical', 'inclined lines' etc. While some users describe orientation by quoting data property, like 'correlation', 'increasing trend' etc., others use only only appearance to describe the plots like 'bottom-left to top-right', 'left-heavy' etc. Clusters 5, 10, 12, 16 in Figure 5 are a few examples where orientation plays an important role in perceptual grouping.

3) *Spread*: The area occupied by the data points on a scatter plot, or its spread, also affects the similarity perception. Generally, spread of data points is relative to the size of the scatter plot in consideration. It is important to note that spread does not correlate with density, *i.e.,* it is possible to have plots with all combinations of spread values. However, when the other

ID	Dominant Concepts	Easiness	Confidence
0	Density, Spread, Grouping	3.23 [3.10,3.36]	3.01 [2.91,3.10]
1	Density, Spread, Edges	3.57 [3.49,3.65]	3.50 [3.43,3.57]
2	Edges, Orientation, Grouping	3.35 [3.21,3.48]	3.11 [2.99,3.24]
3	Spread, Density	2.83 [2.75,2.91]	2.59 [2.50,2.69]
4	Edges, Spread, Density	3.19 [2.92,3.45]	2.89 [2.56,3.22]
5	Orientation, Regularity	4.56 [4.41,4.71]	4.50 [4.33,4.68]
6	Density, Spread	3.11 [3.02,3.20]	2.89 [2.78,3.00]
7	Spread	2.96 [2.86,3.06]	2.61 [2.49,2.74]
8	Density, Grouping	3.68 [3.63,3.72]	3.49 [3.44,3.54]
9	X	X	X
10	Orientation, Edges	4.15 [3.98,4.32]	3.83 [3.71,3.95]
11	Edges, Density	3.76 [3.72,3.79]	3.30 [3.26,3.35]
12	Orientation, Density	4.03 [3.69,4.36]	3.95 [3.53,4.38]
13	Orientation	4.29 [4.13,4.45]	4.14 [3.95,4.34]
14	Density, Grouping	3.51 [3.21,3.82]	3.30 [3.00,3.60]
15	Regularity	4.67 [4.58,4.75]	4.50 [4.36,4.64]
16	Edges, Orientation, Regularity	4.35 [4.22,4.49]	4.06 [3.91,4.20]
17	Density, Spread, Regularity	3.00 [2.80,3.20]	2.85 [2.56,3.14]
18	Spread, Orientation	3.23 [3.04,3.41]	3.03 [2.82,3.25]
19	X	X	X

Table 1. Dominant concepts for each group with mean easiness and confidence scores along with their 95% confidence intervals. Rows marked with '*X*' are groups discarded as they contain very dissimilar plots.

concepts, say density or orientation, are not dominant in the plot, it becomes really difficult to group the plots together based on just *spread* (refer to cluster 7 in Figure 5 and Table 1). Common terms used by users to refer to area based grouping are 'space', 'spread', 'big', 'small' etc.

4) *Regularity*: Regularity refers to the consistency with which certain concepts, like shape or density, appear throughout the plot. In other words, regularity can refer to the repetition of certain patterns in a plot. For example, in Figure 5 clusters 10 and 16 show strong linear patterns that are repeated throughout the plots. Similarly, in cluster 15, grid-like structures are consistent throughout the plot. Regularity is often reported using terms like 'well-spaced', 'regular', 'structured' etc.

5) *Grouping*: Grouping or clustering refers to a set of distinguishable groups present on a scatter plot. When comparing plots based on grouping, users look for the presence or absence of groups that form the grouping. In simpler words, if a pair of plots contain a set of groups, say a line-like structure and a point cloud; they would be suitable candidates to group together based on grouping concept. Users report grouping, similar to that in cluster 2 in Figure 5, by explicitly mentioning the underlying patterns using phrases like 'vertical line, spreading points', 'line and noise' etc.

6) *Edges*: Distributions with strong edges also have an effect on the perceived similarity between plots. This concept overlaps with density and orientation when the points are more uniformly distributed. When the points are distributed in shapes with strong edges, *e.g.,* 'T-shaped' or 'L-shaped' distributions, users refer to them using explicit terms and phrases that describe the shapes they see. Some of the groupings where edges seem to play an important role are clusters 1, 2 and 11 in Figure 5.

As it can be seen in Table 1, most clusters contain a combination of dominant concepts. A particularly interesting observation is that some combinations of concepts lead to higher easiness and confidence scores, while others tend to be relatively weak and lead to lower scores. For example, *regularity* and *orientation* lead to easier and confident grouping, whereas, grouping plots based on *spread* is much more difficult. Similarly, an interplay of *density* and *edges* lead to easier and confident grouping. We present the MDS projection of the PDM in Figure 7 with clusters denoted using colors. At the top-left, low *density* and high *spread* are common. When we move towards the bottom-right, we see more structured plots with high *regularity*. Towards the left-center, we see plots with higher *density*, slightly lower to which are plots with strong *edges*, less *spread* and high *density*. Through visual inspection, we find that *density*, *edges*, *regularity* and their combinations are the most dominant concepts in perceptual grouping of scatter plots.

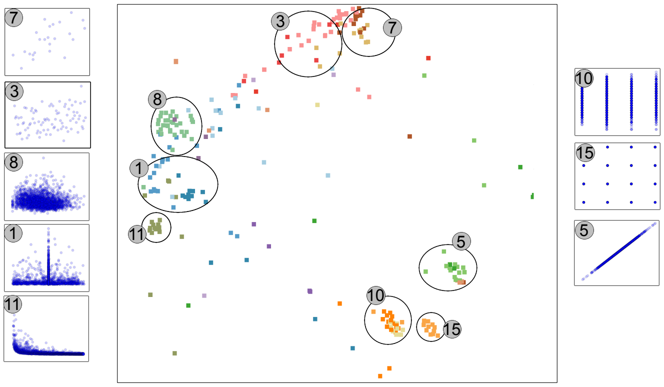

Figure 7. A *multidimensional scaling* projection of the perceptual distance matrix showing distribution of plots. The elliptical shapes denote clear, compact clusters. The clusters are tagged with their cluster IDs, along with a representative image of the clusters.

In future, one can discuss the issues of categorization of these concepts into basic, subordinate and superordinate levels - a popular concept in vision science [25].

IMPLICATIONS

In the following we describe a number of implications and take-home messages we derive from our study.

Need to develop perceptually-balanced quality measures. As we have seen from the comparison with the *scagnostics* method, measures that do not explicitly take into account human perception may actually fail to reproduce human vision and judgment accurately. We posit that research on visual quality measures [5] should be calibrated and expanded by increased knowledge of how humans extract information from plots. In this sense the work of Rensink *et al.* [22], Harrison *et al.* [13], and Li *et al.* [19] on perception of correlation go in the right direction. The work we presented in this article move one step forward in this direction by providing a finer characterization of perceptual similarity.

Perceptual descriptors can be used to derive perceptually-balanced similarity functions. The analysis we ran on the results allowed us to derive a number of plot descriptors. While in this article we did not try to develop computational methods to compute these descriptors, we believe they form the basis for future development in this direction. While developing these concepts we strove to develop them keeping an eye to how feasible their implementation would be. Even though more research is needed to develop and evaluate such computational methods, we are confident they can be reliably implemented in computer-based algorithms.

Perceptual descriptors can be used to navigate large sets of plots. Even though, as we said, computational methods that implement our concepts do not exist yet, we deem it important to briefly describe how they may be used in practice. Plot descriptors can be used in the exploration of large set of plots in two main ways. First, to single out or rank the plots according to one or more concepts at a time (*e.g.,* density, regularity, orientation) using a *dynamic query* filtering mechanism. Second, they can be used to create similarity functions that group plots with established clustering and projection algorithms. More technically, one can build a visual tool that uses plot summarization as a starting point for exploratory data analysis, similar to ScagExplorer [8].

Perceptual similarity vs. data semantics. When studying human perception of plots, we need to figure out the interrelationship between visual perception and data semantics. Judging plots according exclusively to their appearance may be present some limitations. For instance, we noticed that plots with high overall density tend to be grouped together regardless their orientation or distributions, which are clearly characteristics that denote different data distributions. When we group plots according to their density, we neglect important aspects of the data that should be taken into account. While we are not solving this problem here, we deem it important to raise this issue for future research.

CONCLUSION AND FUTURE WORK

In this article, we have focused on how humans group plots according to similarity. We conducted a study aimed at understanding how human observers judge scatter plot similarity when presented with a large set of iconic scatter plots. We used both computational and qualitative methods to choose the final set of scatter plots, and to design our study. Using a perceptual distance matrix, we computed correlation between the perceived pairwise distance and scagnostic-based pairwise distance between plots. We found that *scagnostics* do not map well to our human perceptual judgements. We then identified key concepts of perceived similarity. Finally, we have discussed the dominance of these concepts, and the overall implications of our work for various domains.

One important extension of our work will be to study the role of *appearance* vs. *data semantics* in similarity perception, and categorizing them into basic, subordinate and superordinate levels. Another will be to conduct a more fine-grained analysis of each of our perceived similarity concepts to identify the dominant sub-concepts. Finally, it should prove interesting to develop perceptually validated similarity metrics that use all our concepts.

REFERENCES

1. Georgia Albuquerque, Martin Eisemann, and Marcus Magnor. 2011. Perception-based visual quality measures. In *Proc. of IEEE Conference on Visual Analytics Science and Technology (VAST)*. 13–20.
2. Anushka Anand, Leland Wilkinson, and Tuan Nhon Dang. 2012. Visual pattern discovery using random projections. In *Proc. of IEEE Conference on Visual Analytics Science and Technology (VAST)*. 43–52.
3. Mihael Ankerst, Stefan Berchtold, and Daniel A Keim. 1998. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Proc. of IEEE Symposium on Information Visualization*. 52–60.
4. Enrico Bertini and Giuseppe Santucci. 2004. Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter. In *Smart Graphics*. Springer, 77–89.
5. Enrico Bertini, Andrada Tatu, and Daniel Keim. 2011. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2203–2212.
6. Ingwer Borg, Patrick JF Groenen, and Patrick Mair. 2012. *Applied multidimensional scaling*. Springer Science & Business Media.
7. Qingguang Cui, Matthew O Ward, Elke A Rundensteiner, and Jing Yang. 2006. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 709–716.
8. Tuan Nhon Dang and Leland Wilkinson. 2014. Scagexplorer: Exploring scatterplots by their scagnostics. In *Proc. of IEEE Pacific Visualization Symposium (PacificVis)*. 73–80.
9. Aritra Dasgupta and Robert Kosara. 2010. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1017–1026.
10. Çağatay Demiralp, Michael S Bernstein, and Jeffrey Heer. 2014. Learning Perceptual Kernels for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1933–1942.
11. Bilkis J Ferdosi, Hugo Buddelmeijer, Scott Trager, Michael Wilkinson, and Jos Roerdink. 2010. Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators. In *Proc. of IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 35–42.
12. Johannes Fuchs, Petra Isenberg, Anastasia Bezerianos, Fabian Fischer, and Enrico Bertini. 2014. The Influence of Contour on Similarity Perception of Star Glyphs. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2251–2260.

13. Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. 2014. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1943–1952.

14. John D Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9, 3 (2007), 90–95.

15. Ilknur Icke and Andrew Rosenberg. 2011. Automated measures for interpretable dimensionality reduction for visual classification: A user study. In *Proc. of IEEE Conference on Visual Analytics Science and Technology (VAST)*. 281–282.

16. Ilknur Icke and Andrew Rosenberg. 2012. Visual and semantic interpretability of projections of high dimensional data for classification tasks. *arXiv preprint arXiv:1205.4776* (2012).

17. Jimmy Johansson and Matthew Cooper. 2008. A screen space quality method for data abstraction. In *Computer Graphics Forum*, Vol. 27. Wiley Online Library, 1039–1046.

18. Sara Johansson and Jimmy Johansson. 2009. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 993–1000.

19. Jing Li, Jean-Bernard Martens, and Jarke J Van Wijk. 2010. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization* 9, 1 (2010), 13–30.

20. A Chris Long Jr, James A Landay, Lawrence A Rowe, and Joseph Michiels. 2000. Visual similarity of pen gestures. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 360–367.

21. Wei Peng, Matthew O Ward, and Elke A Rundensteiner. 2004. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proc. of IEEE Symposium on Information Visualization (InfoVis)*. IEEE, 89–96.

22. Ronald A Rensink and Gideon Baldridge. 2010. The perception of correlation in scatterplots. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 1203–1210.

23. David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. 2011. Detecting novel associations in large data sets. *Science* 334, 6062 (2011), 1518–1524.

24. Bernice E Rogowitz, Thomas Frese, John R Smith, Charles A Bouman, and Edward B Kalin. 1998. Perceptual image similarity experiments. In *Photonics West'98 Electronic Imaging*. International Society for Optics and Photonics, 576–590.

25. Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology* 8, 3 (1976), 382–439.

26. Jörn Schneidewind, Mike Sips, and Daniel A Keim. 2007. An automated approach for the optimization of pixel-based visualizations. *Information Visualization* 6, 1 (2007), 75–88.

27. Michael Sedlmair and Michaël Aupetit. 2014. Data-driven Evaluation of Visual Quality Measures. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 201–210.

28. Michael Sedlmair, Andrada Tatu, Tamara Munzner, and Melanie Tory. 2012. A taxonomy of visual cluster separation factors. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 1335–1344.

29. Jinwook Seo and Ben Shneiderman. 2005. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization* 4, 2 (2005), 96–113.

30. Mike Sips, Boris Neubert, John P Lewis, and Pat Hanrahan. 2009. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 831–838.

31. Donna Spencer. 2009. *Card sorting: Designing usable categories*. Rosenfeld Media.

32. Andrada Tatu, Georgia Albuquerque, Martin Eisemann, Jörn Schneidewind, Holger Theisel, Marcus Magnor, and Daniel Keim. 2009. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. of IEEE Symposium on Visual Analytics Science and Technology (VAST)*. 59–66.

33. Edward Tufte. 1991. *Envisioning information*. Graphics press Cheshire, CT.

34. Simon Urbanek. 2012. Scagnostics: Compute scagnostics - scatterplot diagnostics. (2012). `https://cran.r-project.org/web/packages/scagnostics/index.html`.

35. Liyang Wei, Yongyi Yang, Miles N Wernick, and Robert M Nishikawa. 2009. Learning of perceptual similarity from expert readers for mammogram retrieval. *IEEE Journal of Selected Topics in Signal Processing* 3, 1 (2009), 53–61.

36. Leland Wilkinson, Anushka Anand, and Robert L Grossman. 2005. Graph-Theoretic Scagnostics. In *Proc. of IEEE Symposium on Information Visualization (InfoVis)*. 157–164.

37. Josh Wills, Sameer Agarwal, David Kriegman, and Serge Belongie. 2009. Toward a perceptual space for gloss. *ACM Transactions on Graphics (TOG)* 28, 4 (2009), 103.

38. Myron Wish. 1970. Individual differences in perceptions and preferences among nations. *Journal of Personality and Social Psychology* 16, 3 (1970), 361–373.