Analysis of Visual Encodings Effectiveness for Multivariate Data Similarity Identification

BY

MIRKO MANTOVANI
B.S., Politecnico di Milano, Milan, Italy, 2017

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:

G. Elisabeta Marai, Chair and Advisor

Andrew E. Johnson

Pier Luca Lanzi, Politecnico di Milano

ACKNOWLEDGMENTS

I first want to thank the people that made this research possible, EVL. I was lucky to get involved with this community of smart and hard-working people that inspired me since I was in Milan. From my academic advisor, Liz, who guided me towards finding the right research path and in completing this thesis, to Andy, a great professor that taught me so much and inspired me to conduct research in the visualization field. A big thanks goes to Juan Trelles and Andrew Wentzel which particularly helped me with this research. I'm grateful that I was given the opportunity to work on the SAGE2 team by Maxine and Luc. I want to thank in general all the students from EVL for their help with my user study and for clarifying any doubts I had along the way. Secondly, I want to thank the people that introduced me to the possibility of this joint program between PoliMi and UIC and helped organize everything. A seemingly small influence by the right people completely changed my life. So thank you Marco Santambrogio and Pier Luca Lanzi. Thirdly, I want to thank the person who gave me a valid reason to remain away from my home for more than 1 year. Alison, you have been the North Star of this experience since day one. You and your family helped me so much, accepting me as part of your family from the beginning and I am forever grateful for this. Lastly and most importantly, I thank my parents and their unconditional love and support. Your sacrifices made a dream become reality.

MM

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

FIGURE **PAGE**

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

VE	Visual Encodings
KD	Kiviat Diagram
KDLO	Kiviat Diagram Lines Only
PCP	Parallel Coordinate Plot
MD	Multivariate Data
MDV	Multivariate Data Visualization
IV	Information Visualization
UI	User Interface
RT	Response Time
SVG	Scalable Vector Graphics

SUMMARY

Similarity detection seeks to identify items which resemble other items without being identical to them, sometimes over relatively large collections of multivariate items. Oftentimes, similarity cannot be defined computationally over a dataset, leading to a need for visual analysis. Such situations arise commonly in the analysis of ensemble simulations, of multiple computational models, of patient data repositories, or of geospatial data. In this research, we examine, in the context of similarity detection, the effectiveness of several visual encodings for multivariate data. We conducted a user study with 40 participants to measure similarity detection accuracy and response time under two conditions: moderate-scale (16 items) and large-scale (36 items). Our statistical analysis shows that there are significant differences in encoding performance, especially in the large-scale setting of the experiment. In all settings, we found that plain parallel coordinate plots are slower to read and lead to lower accuracy than juxtaposed star glyph approaches. When the number of items grows, the contour star plot (Kiviat diagram) outperforms other variations, including data lines star plots, and is therefore suitable for similarity identification when dealing with relatively large multivariate datasets.

CHAPTER 1

INTRODUCTION

1.1 Motivations

The dimensionality and size of data is growing in a very fast way and databases are being flooded with it. All this data is being utilized in many different ways to extract meaningful information by machine-executed algorithms. However, there are and always will be a set of applications whose main tasks exploit human visual processing, hence, effective graphical representations of data are needed to harness visual perception capabilities.

"Information visualization is the use of computer-based visual representations of abstract and non-physically based data to amplify human cognition. It aims at helping users to effectively detect and explore the expected, as well as discovering the unexpected to gain insight into the data"(Chan 5) [1].

Intensive research has been performed with the goal of identifying the best type of chart for different applications. Nowadays, most of the visualization tasks can be achieved using popular and recurrent types of charts because they have been determined to be the best suited chart for that specific application, and flow charts to aid people in finding the correct Visual Encodings (VE) for their usage have been created, (e.g. popular flow chart designed by Dr. Andrew Abela).

A lot of the comparison tasks usually involve only two variables and groups of scatterplots are surely a good enough VE for this type of scenario [2]. However, for more particular and laborious kind of tasks, such as item-to-item comparison of data points with more dimensions (multivariate) and many items to compare, we are still not sure what is the best VE.

This research studies the effectiveness of different type of charts, including state-of-the-art encodings used for comparison, when dealing with item-to-item similarity identification in the context of Multivariate Data (MD) and relatively large number of items to compare. The goal of Multivariate Data Visualization (MDV) is to help users to identify, associate, categorize, locate, compare, cluster, rank or correlate the represented data [3]. We concentrate on comparison and in particular similarity detection.

Similarity detection seeks to identify items which resemble other items without being identical to them, sometimes over relatively large collections of multivariate items. Detecting similarity is an intrinsic part of comparison, along with judging dissimilarity or differences between items. However, comparison is oftentimes a detailed, precise, finely tuned operation using specific channels such as size, and is oftentimes performed in a one-to-one setting, where items are placed side-by-side and compared pairwise. In contrast, similarity detection over large collections of items often involves a simultaneous, fast, coarse assessment of multiple items at the same time, where the items are characterized by multiple variables. Such situations arise commonly in the analysis of ensemble simulations [4], of multiple computational models [5] [6], of patient data repositories [7] [8] [9], of geospatial data [10], of computer networks [11] [12] and of sports games [13]. Therefore, understanding what the best tools to approach these problems is essential and can provide benefits in different fields and applications.

To help elucidate these issues, we conducted a user study with 40 participants. We measured similarity detection accuracy and response time under two conditions: moderate scale (16 items) and larger scale (36 items). We approached this problem by examining variants of several encodings commonly-used in similarity detection: Kiviat Diagram Lines Only, Kiviat Diagrams, and plain Parallel Coordinate Plots. Our statistical analysis shows that there are significant differences in encoding performance, especially in the large-scale setting of the experiment.

1.2 Challenges

MDV and information visualization face shared challenges: Finding the best visual representations of a problem can be hard and most of the time non-deterministic. MD also poses the problem of encoding all the attributes in a single visual chart [1].

- **Mapping.** The greatest challenge is to find a suitable mapping for a high-dimensional data point into a two-dimensional visual representation of the object. When dealing with particularly high dimensional data, specific rules and extreme caution must be observed in order not to overwhelm the user's viewing ability (observation found in different research papers [14] [15]). Cognition overload could easily be induced in these type of scenarios [15], it is therefore of paramount importance to design the VE in the best possible way so that each dimension can be easily distinguished by the observer.

- **Dimensionality.** Since MD often has a very large number of dimensions, it is a challenge to present a data point in a single visual display. Also, ordering of the dimension often impacts the expressiveness of the visualization, and sometimes different arrangements of axes lead to different conclusions [16].

- **Effectiveness evaluation.** The final objective of data visualization is to obtain insight from data and show correlations between items or attributes. Unfortunately, due to the various difficulties in these tasks, we cannot always assess the effectiveness of different visualization techniques since the ground truth is itself what we are trying to find in the first place [17].

The dimensionality challenge in also tackled by Gleicher in Considerations for Visualizing Comparison [18], where the main scalability and dimensionality challenges identified are: the number of items to compare, the size or complexity of the items (dimensionality) and the size or complexity of the relationship between items. While we consider the first two problems, in this thesis the complexity of the relationship between items will be the same as the considered dimensions of the items itself. Particularly interesting to note, is that comparing and finding similarities between two items is much different than comparing more than two items. This is mainly because with only two items explicit encoding solutions such as showing the encoding of the difference or similar approaches are available (TreeJuxtaposer [19] [20] ,MizBee [21], Unix original *diff* program).

1.3 Outline

This thesis is organized as follows: in Chapter 2 we describe the related previous work and we provide a motivation of the chosen encodings based on previous research results.

In Chapter 3 we provide a guide into the technical aspects of the application built for the user study, as well as the calculation and postprocessing of data and metrics.

Chapter 4 describes the user study, its design and setup, statistics of participants and the data used.

The results of the user study are presented in Chapter 5, which also contains a discussion of the findings and comparison with initial hypotheses.

Lastly, the conclusion (Chapter 6) contains a summary, suggestions and direction of future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Comparative Visualization

The visual representation of MD is a very common task in the field of data visualization, therefore, numerous types of VE have been developed [22]. A current focus in research is whether certain VE are better at representing the data and allowing comparison tasks to be performed in an easier way. For this reason, new types of VE are still being created and tested [23]. The main task performed in Comparative Visualization is to understand how different datasets are similar or different. Similarities or differences may occur at many different levels, for instance:

- **Item to Item:** entries of different datasets are compared to one another based on their values.
- **Dataset to Dataset:** entire or subsetted datasets may be compared to one another [24].
- **Methodology to Methodology:** involve quantifying differences in experiment or simulation parameters.

We concentrated our analysis at the Item to Item level. For this type of tasks, it has proven to be more effective when links are used in conceptually different representations of the same data [25], but in this study we will only compare and evaluate the effectiveness of single and independent charts.

2.1.1 Comparative designs taxonomy

Tufte [26] gives many examples of historical visual designs that support comparison tasks, and explain the importance of this mission.

During the years, a number of different types of VE have been used. Keim and Kriegel [27] [28] categorized the different types of VE for multidimensional MD into six classes: icon-based, geometric, pixel-oriented, graph-based, hierarchical and hybrid. For this thesis, the VE of interest are mainly in the categories of geometric, even if hierarchical approaches (namely Treemaps) are also often used for these kind of tasks.

Comparison tasks' solutions, as explained in the paper Visual Comparison for Information Visualization [29], are primarily divided in three different types depending on how they display different items or the relation between items:

- **Juxtaposition** (also called separation) Juxtaposition designs show each item separately (i.e. next to each other, in either space or time). Juxtaposition uses the viewer's memory to make connections between objects.
- **Superimposition** (or overlay) Superimposition designs present the objects in the same coordinate system (i.e. one on top of the other). Superimposition uses the visual system to make connections between objects. [30] shows an interesting Superimposition approach performed in a manual way.
- **Explicit Encoding** of the relationship between items, directly encodes connections between objects visually. Explicit Encoding uses computation to make connections between

objects. [31] contains examples of explicit encodings of differences using small-multiple displays for comparison.

In this research we will concentrate on the first and most used types of techniques: Juxtaposition and Superimposition.

2.1.2 Spatial and non-spatial encodings for multivariate data

Lee et al. [32] evaluated the effectiveness of four different visual encodings, among which Kiviat Diagrams and Chernoff faces, for conveying similarity information in MD. In their experiment, they asked 32 participants eight questions for each different visual approach, some of them related to pairwise similarity judgment and others involving a multitude of items. They showed datasets composed of 20 items containing only binary features. It was found that both glyph visualizations lead to low-confidence and inaccurate answers, while the encodings that used spatial arrangement of the objects resulted in quick and more accurate answers. Likewise, Borgo et al. [33] indicate that elements arranged on a line or curve are perceived to be more similar than elements which are not. To take advantage of these principles, similar multivariate item representations over large datasets can be placed in close proximity in a 2D space, for example via techniques such as lower-dimensional embeddings. However, it is not always possible to encode similarity using position when the data itself is spatial in nature, for example, when the representations are anchored to zipcodes on city maps or to anatomical locations in medical visualization [34]. Furthermore, dimensionality reduction or aggregation are not always acceptable, for example, in algorithm explainability [35]. Last but not least, similarity may not always be computed apriori, for example when the weights of the different features of the items with respect to similarity are yet to be determined.

When the use of spatial position to encode similarity is not feasible, the multivariate items to be visually analyzed are typically encoded as either juxtaposed (i.e., next to each other) multivariate encodings, or as superposed (overlaid) multivariate encodings, such as PCP. While these approaches for representing MD have been studied in the context of various tasks (relationship, composition, distribution, one-on-one comparison), there is no rigorous evaluation as to which encoding, or even what layout paradigm, is better for similarity detection over MD involving many items. Furthermore, we do not know if or how the effectiveness of a specific encoding may scale with the number of items.

2.1.3 Visual encodings for multivariate data

Chan et al [1] identified a taxonomy of four broad categories for visualizing MD: pixel-oriented techniques, hierarchical display, geometric projection, and iconography. Before discussing the use of these representations in similarity detection, let us briefly examine these classes.

Pixel-based techniques encompass methods that rely on representing individual attributes via categorical color schemes. The most popular of these are bar chart variants, as well as lesser-known methods such as space-filling curves [36] [37], or the recently proposed circular-slice charts [23]. However, pixel-based techniques do not scale well to a large number of dimensions, given the need to use color to distinguish variables [28]. This issue makes these techniques not a good fit for high-dimensional data, where hues become harder to distinguish.

Hierarchical methods such as treemaps and dimensional stacking [1] subdivide the data space and show subspaces in a hierarchical way. While powerful, hierarchical encodings may require knowledge of the underlying dataset. These methods may also require user training in order to interpret well that data.

Projection-based methods attempt to map the higher-dimensional space into a lower dimensional space. These methods are typically considered a good choice for identifying correlations, and scale well with large datasets. These methods are most often variants of scatter plots or PCP). While scatter plots are unable to allow individual items to be identified, PCP and their relatives allow individual items to be displayed as a continuous contour across multiple dimensions, making them a popular encoding used for identifying similar data points.

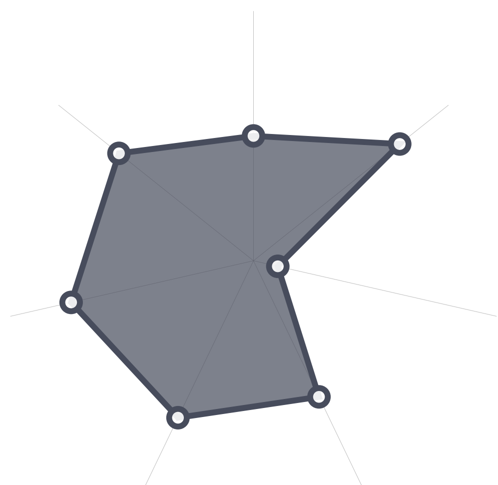
Finally, in iconography, items are shown as glyphs. A glyph is the visual representation of a data item where the attributes of a graphical entity are dictated by one or more attributes of a data record [33]. Unlike methods such as histograms, glyphs have the advantage that they do not scale in size as the set of features being shown increases. Glyph-based methods are particularly well-suited for similarity detection, as visual feature similarity maps directly to numerical similarity.

2.1.4 Juxtaposed glyph encodings for similarity detection

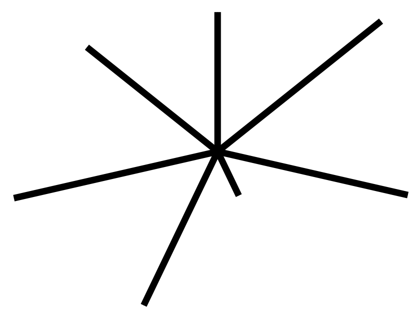
When relative location cannot be used as an indicator of similarity, multivariate data-points are oftentimes encoded by glyphs placed side by side [18] or at discrete locations. Juxtaposed glyphs may support similarity detection through icon attributes such as shape, colors, texture and so on. One of the most prominent glyphs are Chernoff faces [38], where the different parts of a conceptualized human face (head shape, nose and mouth type, eyebrows, eyes, etc.) encode different dimensions of an n-dimensional data set.

Given their small graphical footprint, radial layout glyphs, including star plots or radar plots and their variations, are also frequently used [39]. Fuchs et al. [40] systematically reviewed 64 user-study papers on data glyphs, and many of those compare performances of different types of glyphs when dealing with similarity and comparison tasks, an example is [41]. Star plots or Kiviat Diagrams, in particular, are very effective. There are a multitude of studies that have tried to understand if the assignment of variables to the axes has any impact on their performance [42], [43]. Fuchs et al. [44] studied how Kiviat Diagram contour influences similarity perception. Their three experiments display 8 algorithm-generated variations of a data point positioned in the center, and ask users to find the most similar item to it. The results showed that the Data Lines Only variation of Kiviat Diagram (shown in 1(b)), which does not include the contour, performs best for similarity judgments (the two types of encoding are shown in Figure 1). However, their study only considered a 3 x 3 grid placement of the glyphs, with the target in center, whereas potential matches surrounded the target. This setting is not realistic for similarity detection over larger collections of MD.

Keim and Kriegel [27] categorized the different types of visual approaches for multidimensional data into six classes. They later carried out an experiment [28] to assess the performances of charts belonging to the geometric (e.g. PCP), icon-based (e.g. Chernoff faces or Stick Figure) and pixel-oriented classes for visual data mining tasks, including finding groups of similar data, finding correlations between attributes and similarity retrieval. They specifically designed a pixel-oriented technique that was able to represent as many items as possible on the same display, and compared it to more classical approaches such as PCP. We, instead, concentrate our studies on a much smaller scale.



(a) Kiviat Diagram (KD)



(b) Kiviat Diagram Lines Only (KDLO)

Figure 1: Kiviat Diagram and its variation Kiviat Diagram Lines Only. Fuchs et al. found that the Kiviat Diagram without the contour performs best for similarity judgments.

2.1.5 Superimposed encodings for similarity detection

Parallel coordinate plots (Figure 2) and their relative, nomograms [7] are superposed encodings [18] that assign variables to parallel axes. Due to their scalability and ability to deal with many dimensions, these encodings are often used to explore variable correlation or similarity [25] [45] [46] [47]. PCPs are reportedly effective for the exploration of hundreds to thousands of items [22]. However, Keim and Kriegel [28] report that on a set of thousands of data items, a pixel-oriented encoding outperformed PCPs. Radar charts or star plots can also be seen as a radial layout variation of PCPs, and so can be superposed for the exploration of a large number of items.

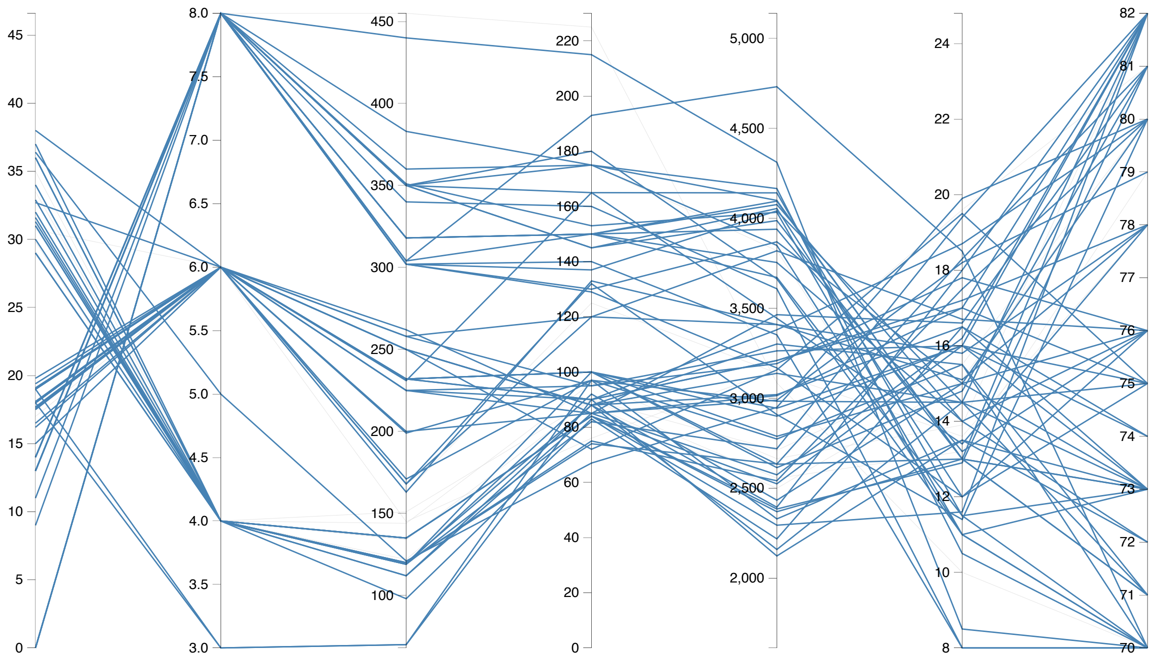


Figure 2: Example of Parallel Coordinate Plot with 7 dimensions and a few dozens items

CHAPTER 3

IMPLEMENTATION

3.1 Web Application

The user study runs in the browser through a dynamic html page. It was never deployed publicly and always run locally.

3.1.1 Dependencies

The application runs on Node.js and incorporates a series of node modules and a few additional dependencies listed below:

- D3.js version 3 used to create the encodings in Scalable Vector Graphics (SVG).
- materialize.css and materialize.js to provide Material Design styling to the pages
- Redis, an open-source (BSD licensed) used as a database, cache and message broker.

3.1.2 Modules

The user study interface was built on top of Experimentr.js [48], a front-end framework that aids in the data collection process and application hosting. The modularity of Experimetr.js allowed us to build different sections of the study independently. The different experiment modules that were identified are:

- Initial Page
- Demographics Survey

- Introduction KD
- Introduction KDLO
- Introduction PCP
- Trial KD (moderate-scale setting)
- Trial KD with Divergent Color Scheme (moderate-scale setting)
- Trial KDLO (moderate-scale setting)
- Trial PCP (moderate-scale setting)
- Trial PCP with Divergent Color Scheme (moderate-scale setting)
- Trial KD (large-scale setting)
- Trial KD with Divergent Color Scheme (large-scale setting)
- Trial KDLO (large-scale setting)
- Trial PCP (large-scale setting)
- Trial PCP with Divergent Color Scheme (large-scale setting)

In turn, each of the trial modules is composed of three parts:

- static HTML subpage
- CSS for styling
- Javascript which provides dynamicity, interacting and calling Experimetr.js APIs that allows the application to keep track of trial time, save data, enable next button. At the

same time, most of the script exploits the D3.js framework to load and preprocess the data and then generate the charts.

Common functions between different modules are in the utils.js script. Every script includes an additional Javascript file based on the used encoding in its User Interface (UI).

3.2 Visual Encodings

The encodings were implemented using the low-level D3.js visualization framework. Each encoding has its own class file and when instantiated, it draws automatically in the div element passed as parameter the chart. Input to each encoding constructor are:

- ID of the div element where the chart will be drawn
- The data object used for the plot
- The configuration file that specifies options (sizes, orientation, colors, margins)

The encoding appends an SVG element to the div and draws the chart based on the given parameters.

CHAPTER 4

STUDY DESIGN

4.1 Goal

In our study, we focus on analyzing the efficiency of visual encodings in a multidimensional similarity detection task. Namely, we seek to determine whether the chosen visual encoding affects the ability of a person to identify the 3 most similar items in a set of candidate items given a target item. In this context, we explore the effectiveness of several variations of two commonly-used base encodings, and the influence of color use and dataset scale over encoding effectiveness.

4.2 Hypotheses

We identified and tested four main hypotheses:

- **H1**. At moderate scale (16 items), all encodings studied will yield equivalent scores.
- **H2**. At the larger scale (36 items), juxtaposed encodings will outperform superposed encodings with respect to score.
- **H3**. At the larger scale (36 items), Lines-Only encodings will outperform other encodings with respect to score.
- **H4**. At both scales, color-cue encodings will outperform other encodings with respect to time, but not score.

4.3 Participants

We recruited 40 volunteers (12 females, 28 males, age 18-59 years) from our university campus, following institutional review board (IRB) approval of the study. Participants had a variety of exper- tise with data visualization (2 novices, 13 with basic familiarity, 14 familiar and 11 experts), and diverse educational backgrounds (5 with a high school degree, 23 with a Bachelors, 10 with a Masters and 2 with a PhD). We did not check or asked anything about their vision or colorblindness.

In Table I,Table II,Table III,Table IV and in Figure 3 the demographics details of the participants are shown.

TABLE I: FREQUENCY OF SEX IN PARTICIPANTS

Sex	**Frequency**
Male	28
Female	12

TABLE II: FREQUENCY OF FAMILIARITY WITH VISUALIZATION IN PARTICIPANTS

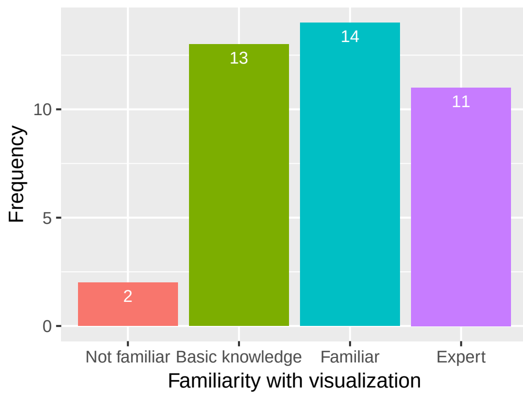
Familiarity with Visualization	**Frequency**
Not familiar	2
Basic knowledge	13
Familiar	14
Expert	11

TABLE III: FREQUENCY OF AGE IN PARTICIPANTS

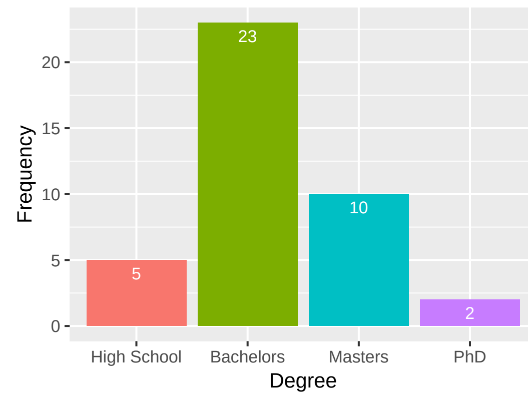
Age range	**Frequency**
18-24	17
25-29	14
30-39	5
50-59	4

TABLE IV: FREQUENCY OF HIGHEST DEGREE OBTAINED BY PARTICIPANTS

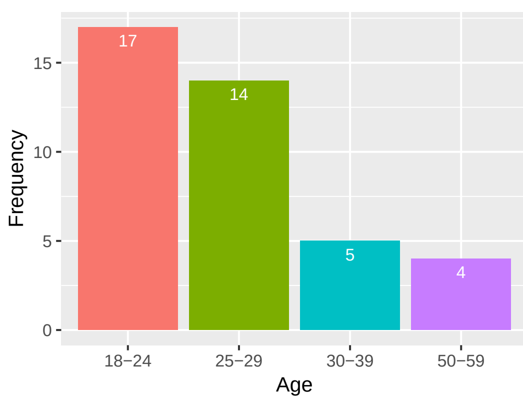
Highest degree obtained	Frequency
High School	5
Bachelors	23
Masters	10
PhD	2



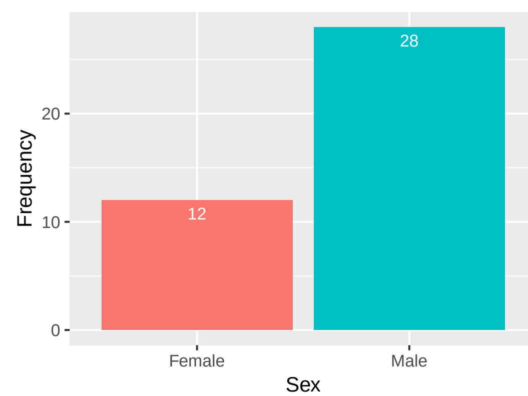
(a) Familiarity with Visualization



(b) Highest degree obtained



(c) Age group



(d) Sex

Figure 3: Bar Charts of demographic statistics. In the demographics form, participants were asked to provide information about each of the 4 categories. The typical subject is Male, familiar with visualization, in the 18-24 age range, and has a Bachelor degree.

4.4 Encodings and Colormapping

We began our study design by first considering the space of appropriate visual encodings, according to visualization theory, as well as the encodings used in similarity detection practice.

As indicated earlier, pixel-based techniques do not scale well to higher dimensions. Many of the popular methods in this category that would be accessible to general audiences are variants of stacked bar charts and pie charts, which assume a part-to-whole relationship between variables, and are not well suited for encoding several independent variables. We therefore did not pursue this category of encodings. Hierarchical methods, likewise, were omitted, as they require knowledge of the underlying data and take training to interpret well.

Next, we considered projection-based encodings. PCP and their relatives allow individual items to be displayed across multiple dimensions, making them a good candidate for similarity detection. Whereas PCP are not anchored to a spatial location, in practice they can be linked to specific item spatial locations via brushing and linking across coordinated views. We include two variants of this encoding in our study.

Last, we considered the space of glyph-based representations. More esoteric methods such as Chernov faces or stick figures may be difficult for end users to interpret, and have the issue of encoding variables unequally. We instead chose to investigate variants of star plots, which circumvent this limitation by displaying all features uniformly along a radial axis.

We note that our resulting encoding selection reflects closely the encodings used in current practice of similarity detection [4] [5] [6][7] [8] [9] [10] [11] [12] [13]. In the similarity detection literature, we furthemore noted the use of KD [7] [8], a star plot variant. Also referred to as star plots or radar charts, this diagram is a radial glyph introduced in 1974 [49], where, as in standard star plots, each radial axis represents a variable and the position along the axis encodes the quantitative value. However, in a KD the resulting contour is filled with color.

Overall, we considered 3 base encodings and 2 variations. These encodings include encoding methods that rely on the juxtaposition paradigm (star glyphs), where the encodings are laid out at specific locations, as well as methods that rely on a superposition paradigm (parallel coordinate plots), where visual marks for the items are superimposed on top of each other. The encodings we study are:

- Parallel Coordinate Plots (PCP). Because in a pilot test unicolor polylines were not distinguishable without interacting with each polyline, each line was colored based on a categorical color scheme.
- Kiviat Diagram Lines Only (KDLO). A variant of the star plot with only the radial segments extending from the center, and no filled contour. These plots have also been referred to as whisker plots or fan plots [44].
- Kiviat diagrams, a variant of star plots. Traditionally, a contour is formed by connecting the quantity marks along each radial axis. In our default Kiviat diagram, the contour was filled in with a neutral gray.
- Color-cue Kiviat Diagrams. This encoding is a variation of the Kiviat diagram, in which the glyph polygon is filled with color. In the practitioner literature, the Kiviat color is typically mapped to a variable of that item. To test whether Kiviat color could be interpreted as a similarity cue, we deliberately mapped color to our simulated ground

truth similarity measure instead. We deliberately did not inform the users that color was mapped to a simulated measure of similarity. We used a divergent red-green-blue color scheme from ColorBrewer2 [50], where red indicated completely dissimilar items and blue indicated perfectly similar items.

- Color-cue Parallel Coordinate Plots. This encoding is a variation of the PCPs, in which polylines are colored, as in the color-cue Kiviats above, based on their similarity with the target polyline. Again, we deliberately did not inform the users that color was mapped to our simulated measure of similarity.

4.5 Scalability with number of items

In our study, we furthermore tested the effect that scale had on the encoding effectiveness, where scale refers to the number of items on the screen during an individual comparison task. To accomplish this test we ran each task for each encoding using two different scale factors: a moderate scale, with 16 total items, and a large scale, with 36 total items, including the target item.

In total, 5 encodings where tested at two different scales each, leading to 10 total trials per participant, not including introductory practice tests.

4.6 Dataset

To simulate realistic charts for our study, we used an anonymized cancer dataset of 1100 patients with seven features of interest [7]. Our dataset included a mixture of ordinal variables with 3-4 value categories each and of continuous variables. For the purpose of this study, the chart encodings did not show labels or values. During preprocessing, all variables were individually scaled to be between 0 and 1 using min-max normalization.

To simulate the ground truth similarity between items, we used cosine similarity over the seven variables of the data points. We chose this measure over Euclidean or Minkowski distances, due to its simplicity and popularity in common information retrieval tasks over high-dimensional categorical data, such as patient records [51], and text data [52]. We decided to use seven dimensions to account for color hue distinction among different values and to be aligned with previous similar studies.

4.7 Protocol

We designed a similarity detection task to evaluate our hypotheses. For each different type of visual encoding and color variant, we showed the seven variables of either 16 (moderate scale) or 36 (larger scale) items. The items were randomly sampled from the patient dataset. On each trial, we asked the participant to select the top three items most similar to a specified target item in the display. The selection was performed by clicking with the trackpad on the svg element containing the chart for juxtaposed encodings, and by clicking on a specific line for the superimposed encodings (PCP). On-click feedback was provided by decreasing the opacity of the chart for a fraction of a second (as Figure 4 shows). Moreover, when hovering over a line for the PCP trials, the thickness of the line increases, so that it is easier to understand what is being selected (Figure 5 shows how it looks in the application). We selected two grid arrangements to test the scalability of the encodings. For the juxtaposed encodings, participants first worked on a 4 by 4 grid (16 items), then on a 6 by 6 grid (36 items). The target item was selected and placed randomly among the 15, respectively 35 candidate items. For the PCP encodings, we presented 16, respectively 36 superposed lines on a single chart. Due to the random selection, none of the screen arrangements was repeated among participants. Once the participant selected the similar top three items, they would proceed to the next trial.

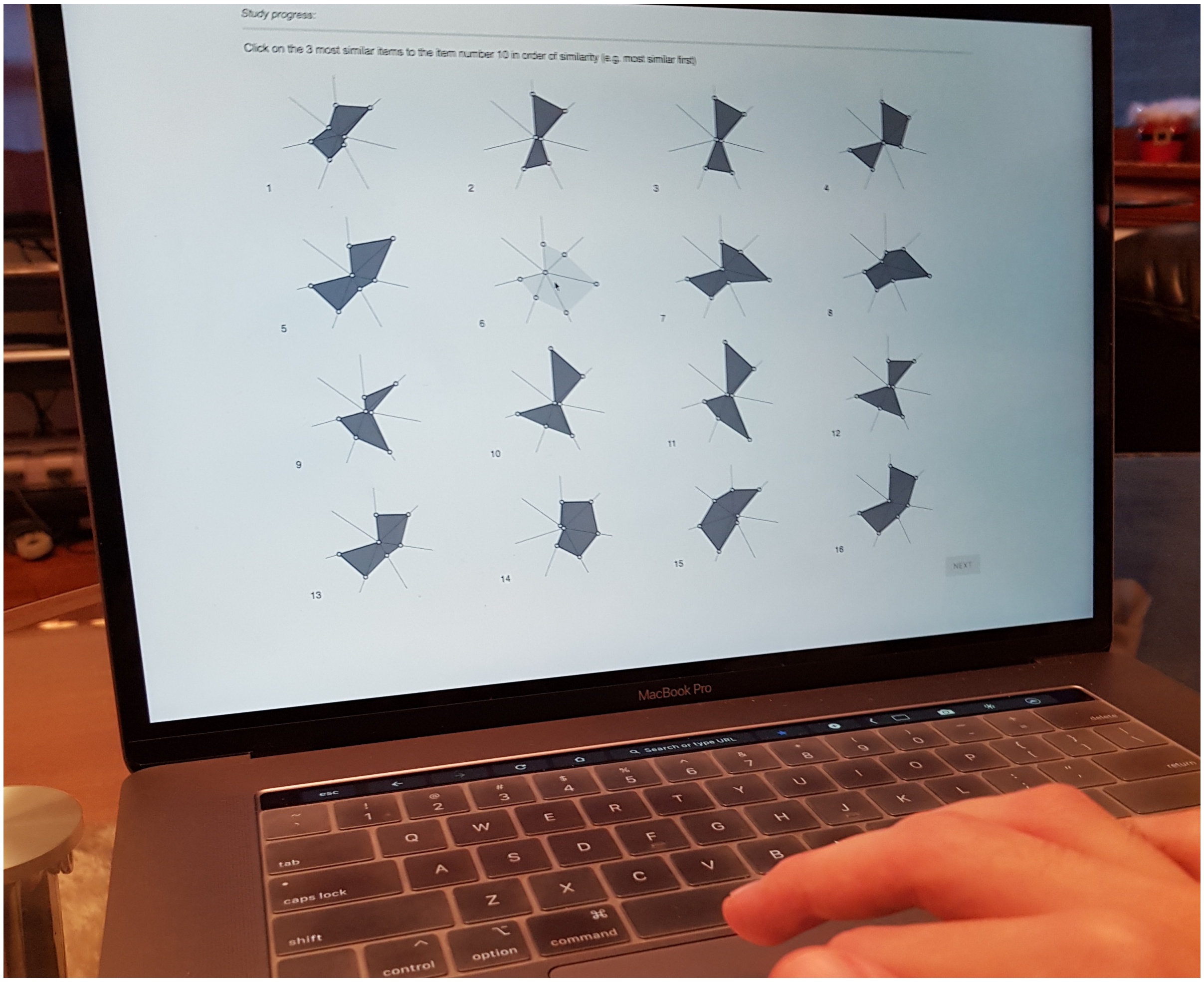
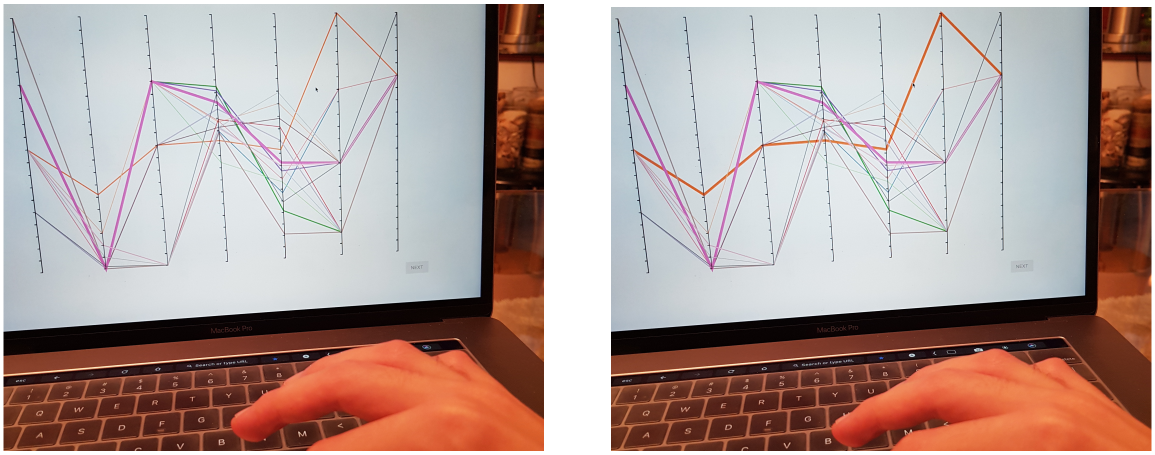


Figure 4: Showing the effect of selecting item 6. A brief highlighting of the item by decreasing opacity allows the user to understand that the click has been registered.



(a) Not hovering (b) Hovering

Figure 5: Showing the effect of hovering over a line for the PCP variants. The pink line is the target item and is always thicker, while the orange line gets bigger when the user hovers over it.

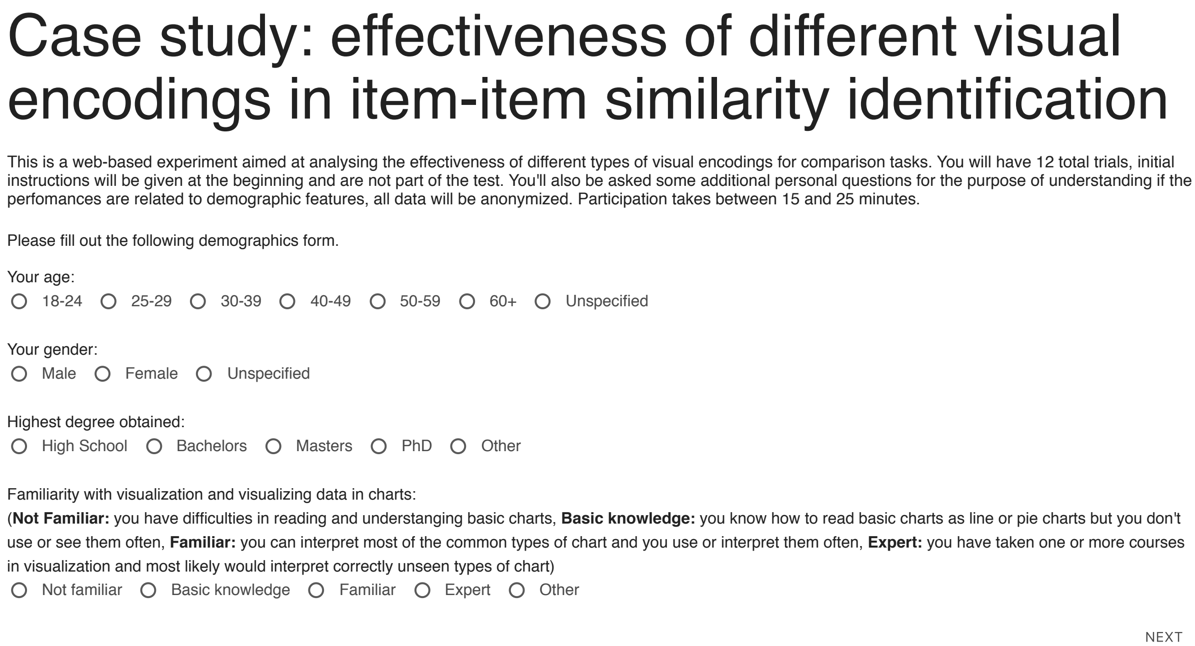
The study is divided into 4 sections:

- Consent, introduction to the study and demographics survey
- Initial tutorial on the 5 different types of encodings
- Experiment trials
- Debrief questionnaire

4.7.1 Introduction to study and demographics survey

Participants were given the IRB-approved Informed Consent form, which provided them with information about the study: the study is voluntary and it is possible to opt out and stop participating in the study at any time, deciding not to participate or deciding to stop participating later, will not result in the loss of any services, class standing, and/or professional status to which you are entitled, and will not affect your relationship with the University of Illinois at Chicago (UIC). They were asked to go over it and sign it.

They then move on to the introduction page, which briefly explains what the user study is about. The demographics survey is a form that collects this data: age category, gender, highest degree obtained, familiarity with visualization and visualizing data in charts. This data was needed for further analysis to understand if there are relations between score/time and a specific demographic feature. Participants were also informed that the test would be timed, but that they were not expected to optimize for time.



Case study: effectiveness of different visual encodings in item-item similarity identification

This is a web-based experiment aimed at analysing the effectiveness of different types of visual encodings for comparison tasks. You will have 12 total trials, initial instructions will be given at the beginning and are not part of the test. You'll also be asked some additional personal questions for the purpose of understanding if the perfomances are related to demographic features, all data will be anonymized. Participation takes between 15 and 25 minutes.

Please fill out the following demographics form.

Your age:
○ 18-24 ○ 25-29 ○ 30-39 ○ 40-49 ○ 50-59 ○ 60+ ○ Unspecified

Your gender:
○ Male ○ Female ○ Unspecified

Highest degree obtained:
○ High School ○ Bachelors ○ Masters ○ PhD ○ Other

Familiarity with visualization and visualizing data in charts:
(**Not Familiar:** you have difficulties in reading and understanging basic charts, **Basic knowledge:** you know how to read basic charts as line or pie charts but you don't use or see them often, **Familiar:** you can interpret most of the common types of chart and you use or interpret them often, **Expert:** you have taken one or more courses in visualization and most likely would interpret correctly unseen types of chart)
○ Not familiar ○ Basic knowledge ○ Familiar ○ Expert ○ Other

NEXT

Figure 6: Introduction to study and demographics page

4.7.2 Tutorial

Before starting the experiment, a quick tutorial on the 3 different types of charts and the task that users will have to perform is displayed. The tutorial serves both the purpose of getting the user more familiar with unusual types of encodings and to allow them to understand what they need to do and to ask clarification questions. The tutorial is presented with a set of 9 items in a 3x3 grid and shows the correct answers (the 3 most similar items to the target one) so that less experienced participants have a reference in how to interpret the notion of similarity.

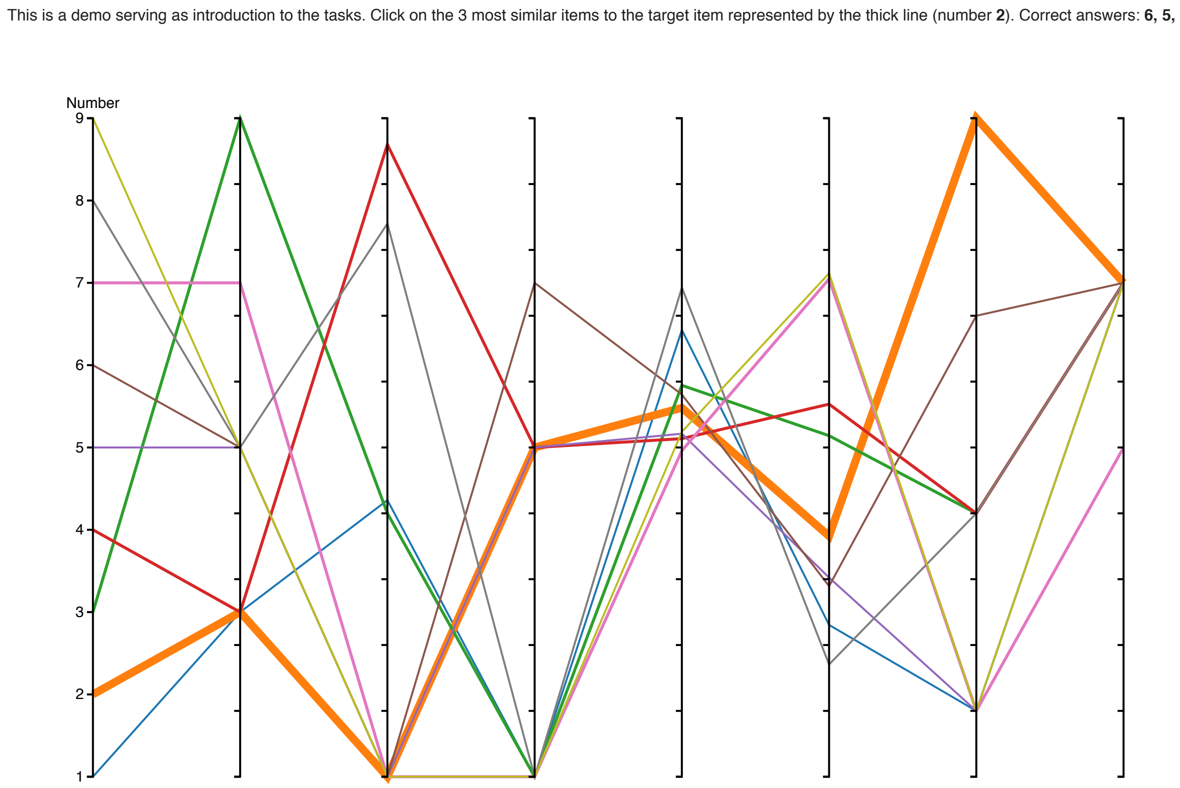


Figure 7: Parallel Coordinate Plot tutorial. Participants are provided with the answer in the tutorial, and only 9 items are present

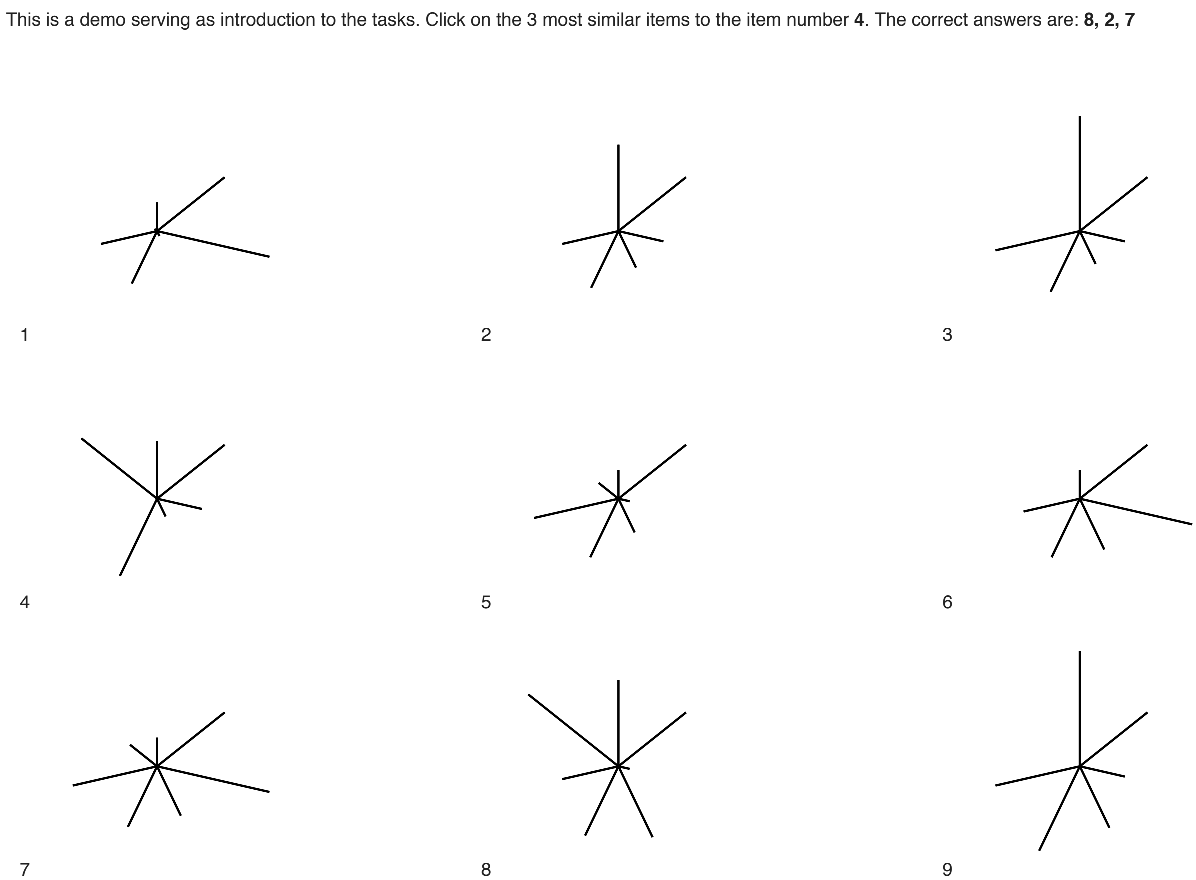


Figure 8: Kiviat Diagram Lines Only tutorial. Participants are provided with the answer in the tutorial, and only 9 items are present

4.7.3 Experiment trials

After the tutorial, participants advanced to the actual experiment where data was collected. We designed a similarity detection task to evaluate our hypotheses. The main recurrent task to perform in each of the 10 trials is to select the top three most similar items to a specified element in the display, called the target item. For each VE we first presented a version with 16 items and then a version with 36 items.

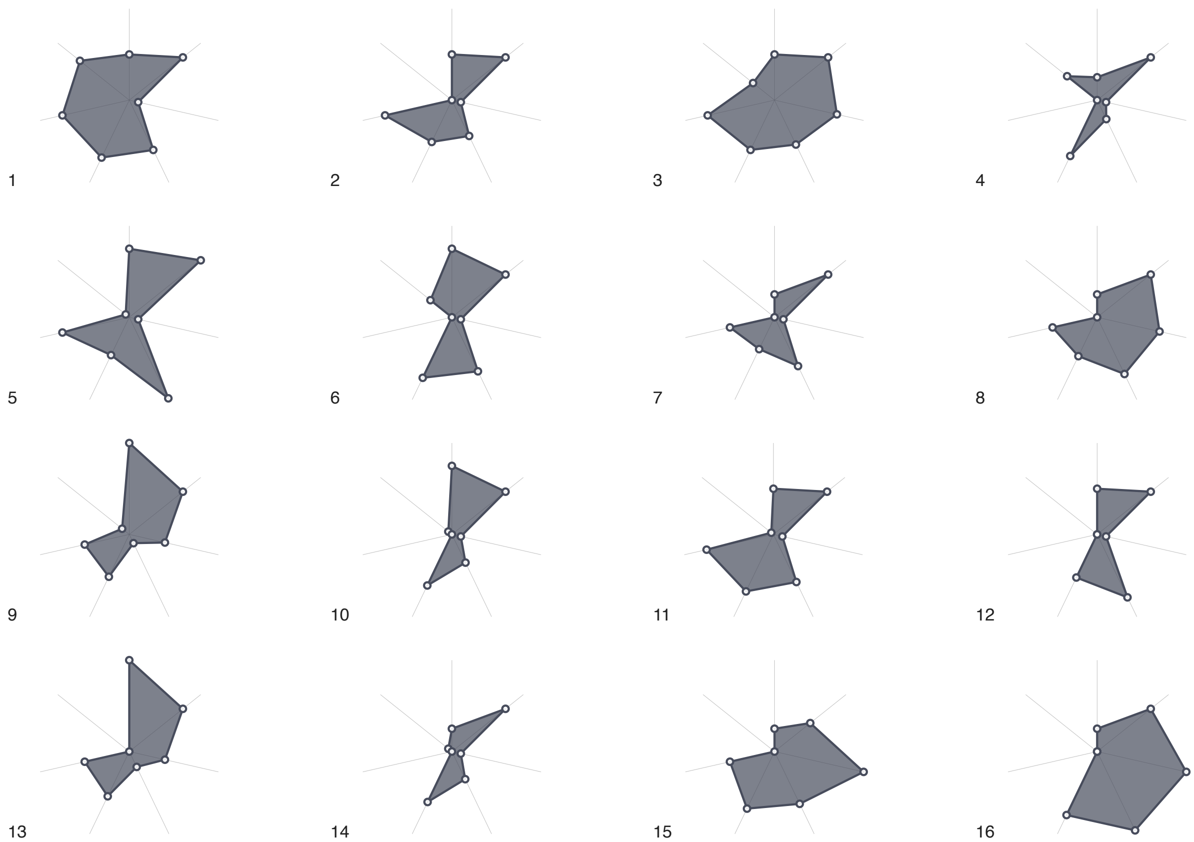


Figure 9: Trial layout for KD in the moderate-scale setting (16 items), data points are arranged in a 4x4 matrix and numbered

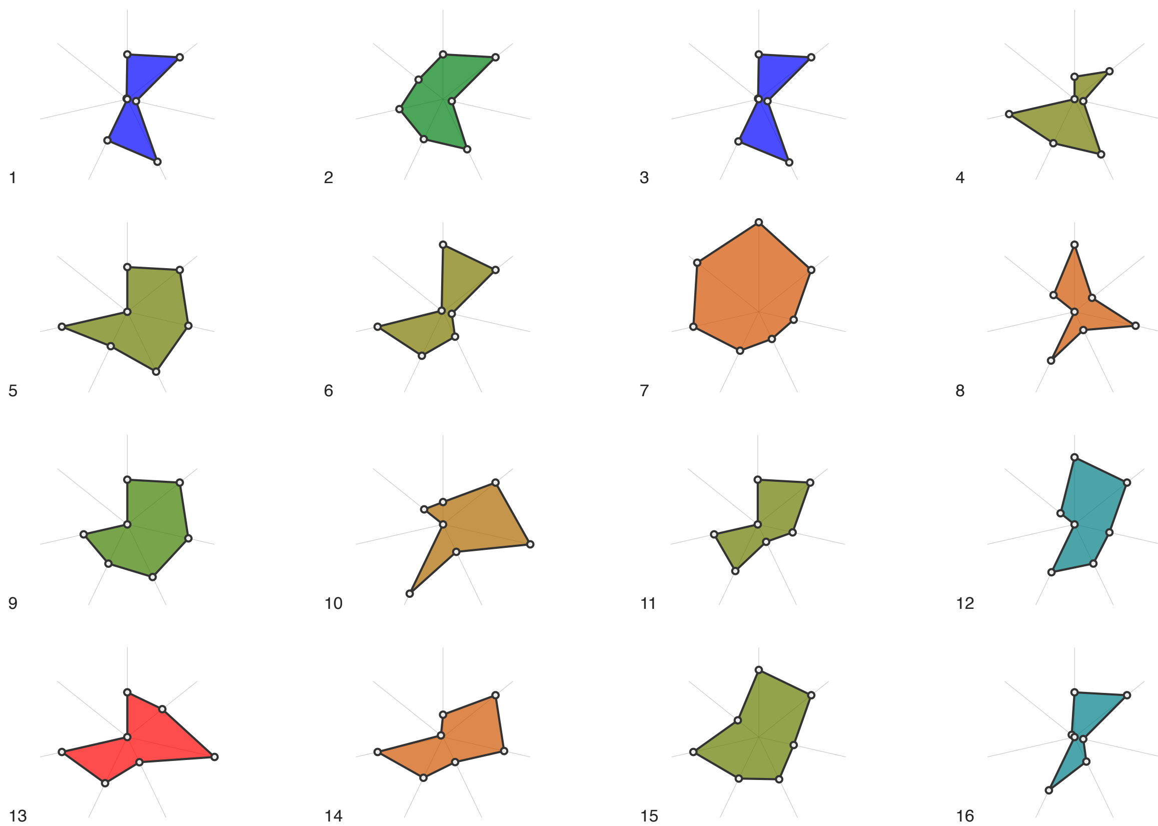


Figure 10: Trial layout for Colored KD in the moderate-scale setting (16 items), data points are arranged in a 4x4 matrix and numbered

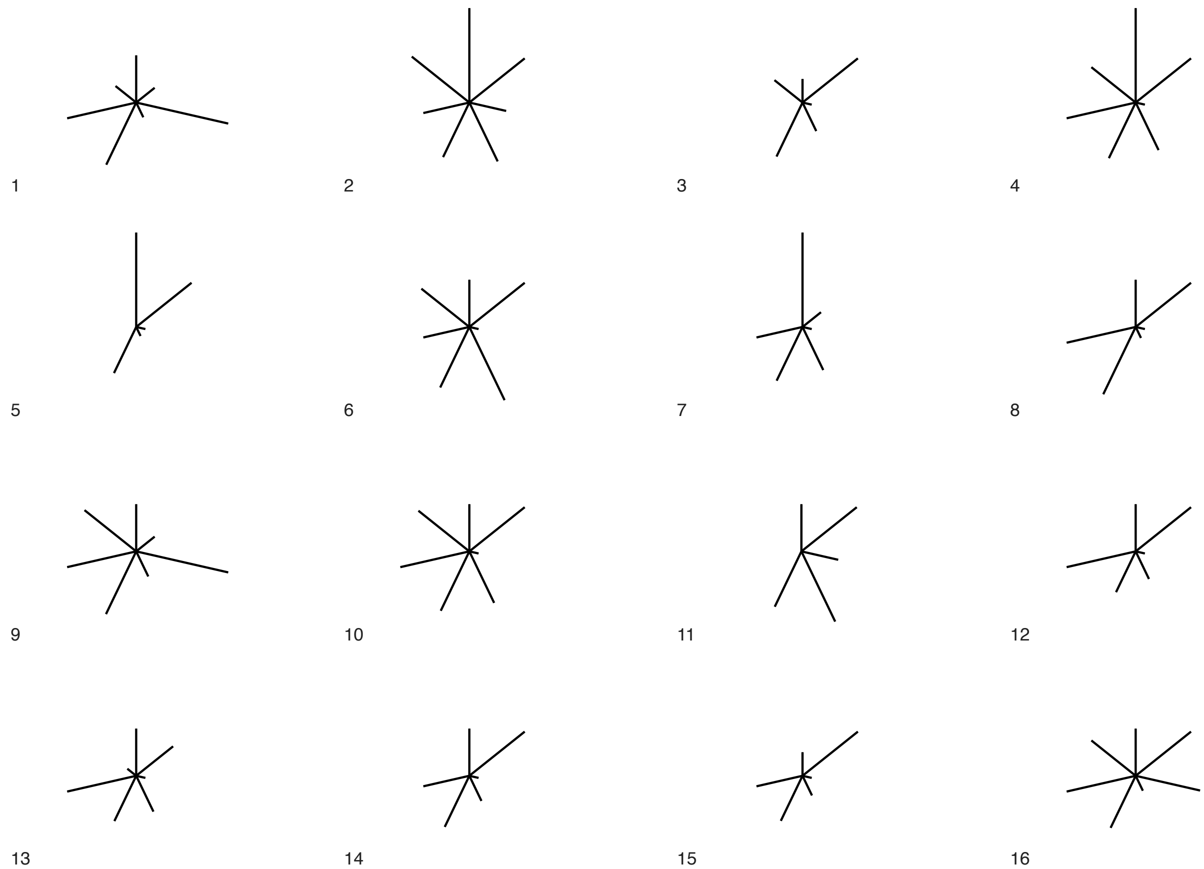


Figure 11: Trial layout for KDLO in the moderate-scale setting (16 items), data points are arranged in a 4x4 matrix and numbered

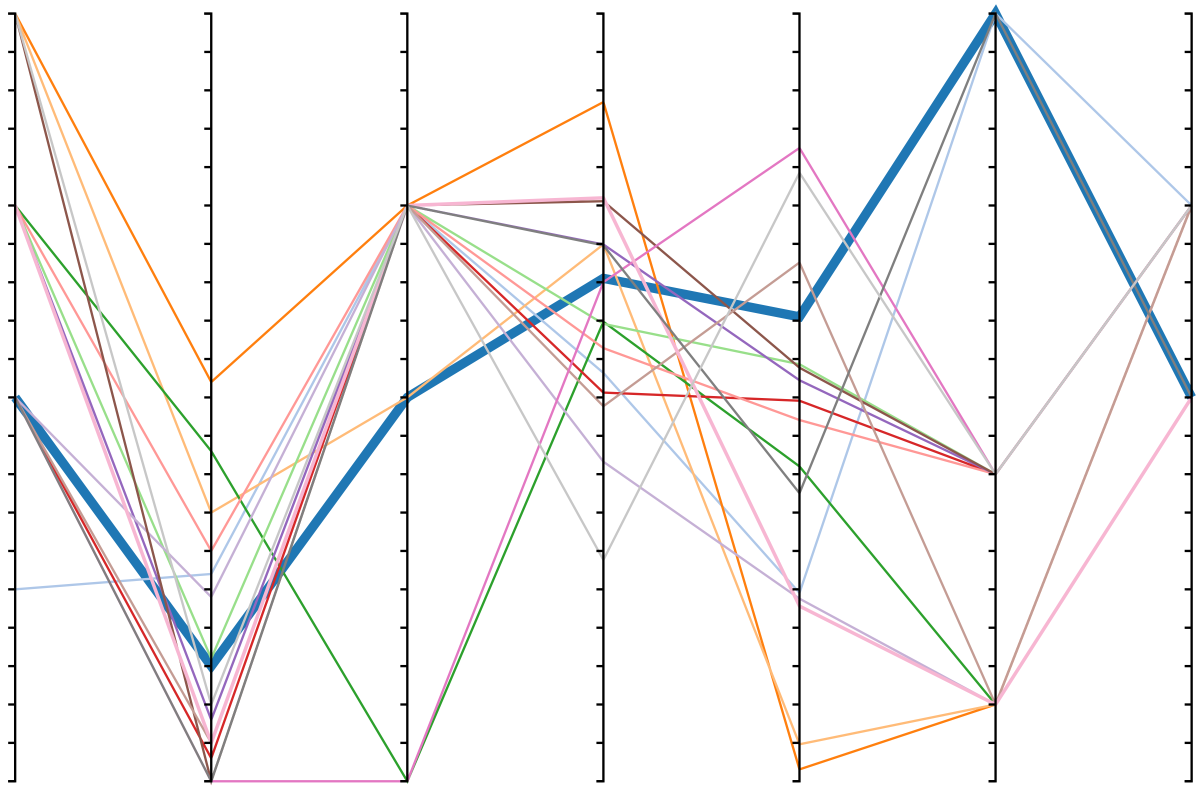


Figure 12: Trial layout for PCP in the moderate-scale setting (16 items), the thick line is the target item

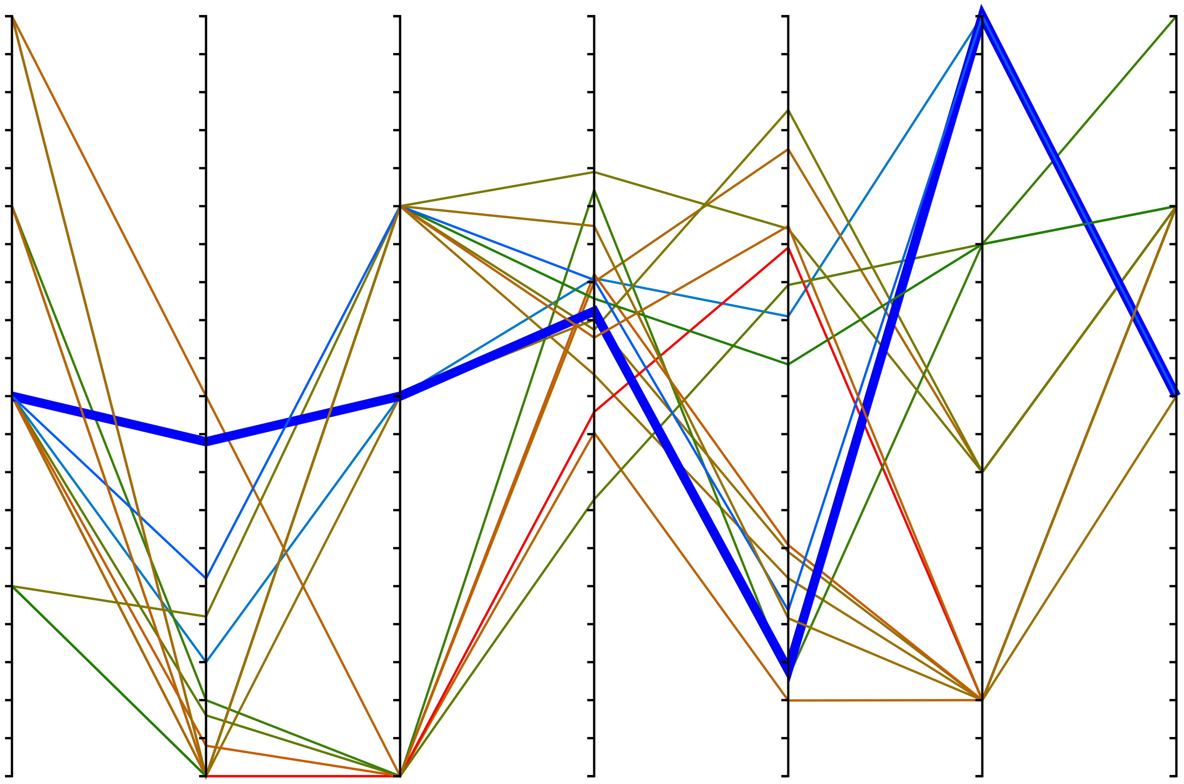


Figure 13: Trial layout for Colored PCP in the moderate-scale setting (16 items), the thick line is the target item

4.7.4 Debrief questionnaire

In the end, a final form collects data by asking participants these questions:

- Which was the most complicated visual encoding to process?
- Which was the easiest visual encoding to process?
- Which of the visual encodings was the most scalable?

- Which of the visual encodings scaled the worst?

We defined scalability in the online form, and we encouraged participants to explain the reason of their response in several text boxes.

4.8 Statistical analysis

We performed our statistical analysis using pairwise, paired sample t-tests to compare the mean of the scores and response times whenever the distribution of the differences was not significantly different from Normal distribution. We conducted the normality test analytically using the Shapiro-Wilk test [53]. We consider results statistically significant when $p < .05$.

4.9 Metrics and scoring system

The metrics we evaluated performances on are: score and Response Time (RT).

4.9.1 Score

The solution of a trial is a selection of 3 charts which they think are most similar to the target item. A good measure to rate the solution is evaluating how far they are from the optimal solution. The optimal solution is when the selected items are the first, second and third elements in the list of items ordered by decreasing similarity. From the solution, we can compute the ranks of the items they chose in the list of items ordered by decreasing similarity. Given the ranks of a user's answer denoted by $r1, r2, r3$, sorted in ascending order of their rank, we compute the Manhattan distance d between their answer and the optimal solution:

$$OptimalSolution \iff r_1 = o_1, r_2 = o_2, r_3 = o_3 \tag{4.1}$$

where

$$o_1 = 1, o_2 = 2, o_3 = 3 \tag{4.2}$$

The distance is given by the following formula, where we use the Manhattan absolute value notation, although each of the three terms is already positive by the construction earlier described:

$$d = |r_1 - o_1| + |r_2 - o_2| + |r_3 - o_3| \tag{4.3}$$

This distance provides a way to measure the error between the provided solution and the best possible solution. From this, we identified a function that transformed the distance in a score s that satisfies the constraint

$$0 \leq s \leq 1 \tag{4.4}$$

and is inversely proportional to the distance. If we wanted to achieve this with a linear function, a simple function of this type would have worked:

$$s = -\frac{d}{m} + 1 \tag{4.5}$$

where m is the maximum possible distance $m = max(d)$

The maximum possible distance, m, is different for the 2 different scale-settings, as the large-scale setting has more items and therefore the solution can be further away from the optimal. Since we also analyzed scalability properties, and in particular we compared score and Response Time samples for the same encoding in the two different scale settings, we decided to opt for a score function s not dependent on m and that satisfies the aforementioned constraint:

$$s = \frac{1}{1+d} \tag{4.6}$$

Since we are interested in comparing score differences among different encodings, we normalized the scores by average user score (i.e. each trial score was divided by its participant's average score). In this way, if a trial score is smaller than 1, it means that the user underperformed with that encoding and viceversa if it is greater than 1. This normalization allows us to compare scores of participants with a significant different average score, because of different backgrounds and experience with visualization.

4.9.2 Response Time

RT is the second metric we used to determine performance. In particular, we were interested in understanding if certain encodings convey the data point information and the differences when comparing it with another item in a quicker way than other encodings. The RT, similarly to the score, is also normalized by user average RT, so that we can compare relative encodings' performance for users with significally different average times.

4.10 Setup

Every study session was run locally on the latest version of Google Chrome and in full screen. The local machine used to perform the studies was always the same: MacBook Pro (15.4-inch display, 2880 x 1800 resolution). The participants only used trackpad interaction during each trial, they were provided with a chair positioned in front of the screen, which was standing on a table. The study was always conducted in closed space with lights on and no background noise. The study web page was in full screen, therefore no other browser or OS-related component could be seen. No monetary or material incentives were given at any point in the study.

Figure 14: User study - view from user's perspective.

4.11 Data collection

We collected various types of data necessary for the analysis while the users were conducting the experiments. In addition to the demographics survey and debrief questionnaire answers, the rest of the data we collected was time-related or related to the answers given in each trial. For each experiment trial, we measured and saved the time from when the plots are displayed to the time in which they click the next button. We also measured the overall time of the experiment. When participants clicked on the three items that form the answer to the experiment trial, the application saves their IDs, together with a list of all item IDs in order of similarity.

CHAPTER 5

RESULTS

For the analysis of the results, we used the data collected from the 40 participants; the sessions had an average completion time of 8 minutes and 57 seconds. Our results show score differences due to the visual encodings under each scale setting (36 items vs. 16 items), as well as response time differences.

5.1 Score performance

5.1.1 Moderate-scale setting results

At moderate scale, we found that the juxtaposed visual encodings outperformed either variation of the PCP (Figure Figure 15, *moderate* tests in Table VI). Furthermore, the Colored KD and Color-cue PCP did not yield higher scores than their counterparts ($t(39)=0.16$, $p=.454$; $t(39)=0.13$, $p=.451$ respectively). In Table V it is possible to see the score means and standard deviation per encoding, ordered by decreasing mean.

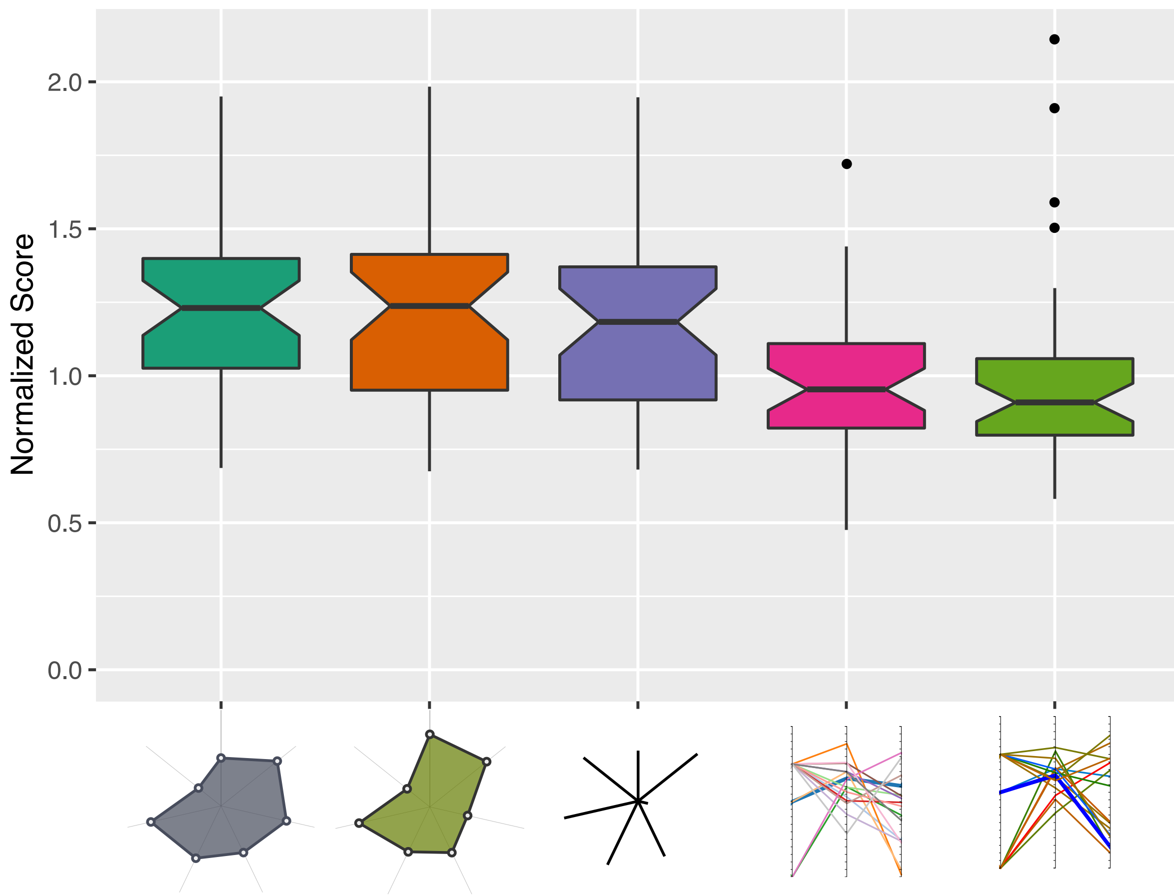


Figure 15: Notched box plots presenting the distribution of the score in the moderate-scale setting (normalized by user average score) showing a statistically significant lower score for the PCP variants compared to all the other visual encodings. Higher score is better, with values > 1 indicating above average performance and values < 1 indicating below average performance.

TABLE V: RESULTS OF NORMALIZED SCORE MEAN AND STANDARD DEVIATION IN THE MODERATE-SCALE SETTING ORDERED BY DECREASING PERFORMANCE (A HIGHER SCORE IS BETTER)

Visual Encoding	**Mean**	**Standard Deviation**
Kiviat Diagram (colored)	1.23	0.31
Kiviat Diagram	1.22	0.30
Kiviat Diagram Lines Only	1.19	0.32
Parallel Coordinate Plot (colored)	0.99	0.32
Parallel Coordinate Plot	0.98	0.25

TABLE VI: SCORE-RELATED T-TESTS IN MODERATE-SCALE SETTING, HIGHER SCORE MEANS HIGHER PERFORMANCE.

Relation between Mean Scores	**t**	**p-value**
Kiviat Diagram > Parallel Coordinate Plot	3.34	$<$ **0.001**
Kiviat Diagram > Color Parallel Coordinate Plot	3.16	**0.002**
Color Kiviat Diagram > Parallel Coordinate Plot	3.87	$<$ **0.001**
Color Kiviat Diagram > Color Parallel Coordinate Plot	3.27	**0.001**
Kiviat Diagram Lines Only > Color Parallel Coordinate Plot	3.67	**0.003**
Kiviat Diagram Lines Only > Parallel Coordinate Plot	2.97	**0.020**
Color Kiviat Diagram > Kiviat Diagram	0.16	0.454
Kiviat Diagram > Kiviat Diagram Lines Only	0.46	0.321
Color Parallel Coordinate Plot > Parallel Coordinate Plot	0.13	0.451
Color Kiviat Diagram > Kiviat Diagram Lines Only	0.60	0.277
Kiviat Diagram $\neq$ Kiviat Diagram Lines Only	0.46	0.645

5.1.2 Large-scale setting results

At the larger scale, the juxtaposed visual encodings significantly outperformed all PCP variations (Figure Figure 16, *large* tests in Table VIII). Contrary to Fuchs et al. [44], KD and KDLO had the same performance (t(39) = 1.43, p = 0.159 in large-scale setting; t(39) = 0.46,

$p = 0.645$ in moderate-scale setting. Lastly, unlike in the moderate-scale setting, the Color-cue PCP outperformed the PCP when using 36 items ($t(39) = 2.23$, $p = 0.016$). Table VII shows the score means and standard deviation per encoding, ordered by decreasing mean.

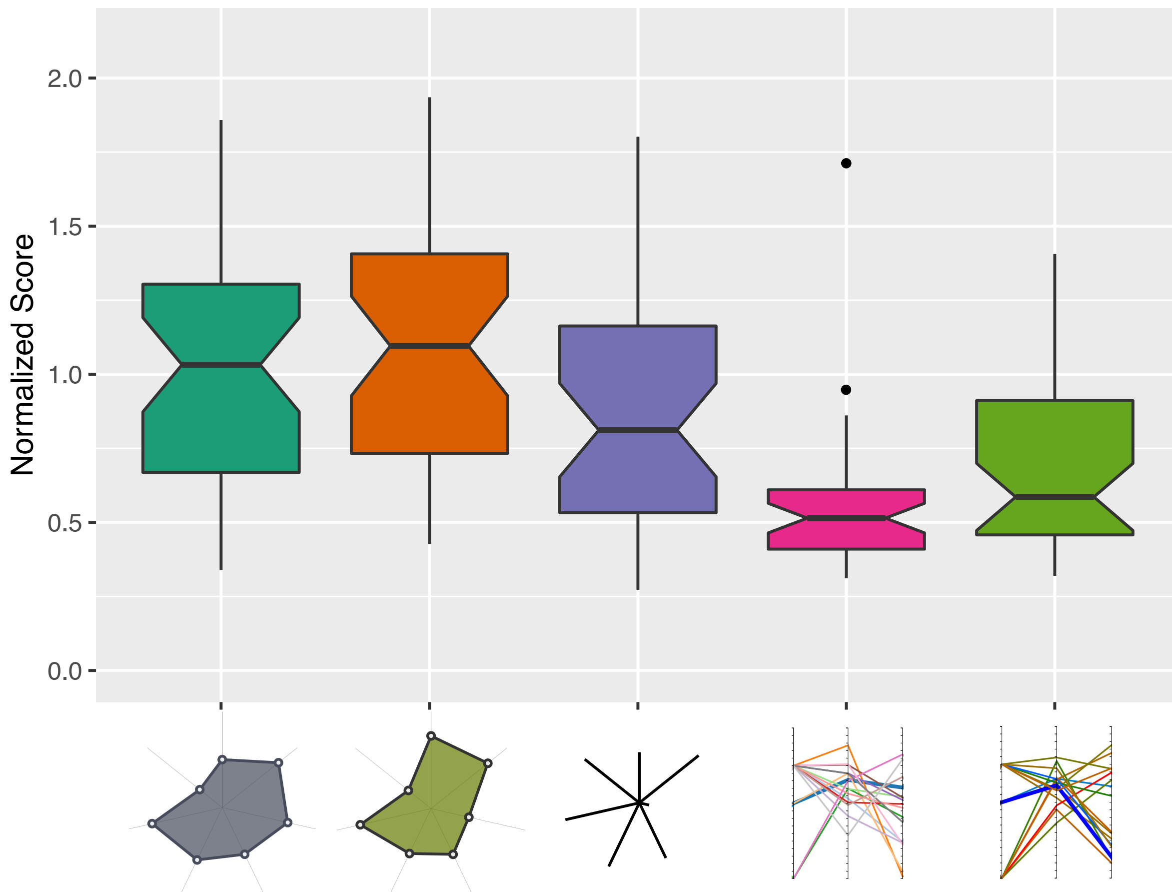


Figure 16: Notched box plots showing the distribution of the score in the large-scale setting (normalized by user average score. Higher score is better, with values > 1 indicating above average performance and values < 1 indicating below average performance. PCP do not perform well in a high scale setting and have a significantly lower score than all the other encodings, color helps improving its score. Colored KD have a significantly higher score than KDLO

TABLE VII: RESULTS OF NORMALIZED SCORE MEAN AND STANDARD DEVIATION IN THE LARGE-SCALE SETTING ORDERED BY DECREASING PERFORMANCE (A HIGHER SCORE IS BETTER)

Visual Encoding	**Mean**	**Standard Deviation**
Kiviat Diagram (colored)	1.09	0.42
Kiviat Diagram	1.02	0.43
Kiviat Diagram Lines Only	0.87	0.38
Parallel Coordinate Plot (colored)	0.69	0.29
Parallel Coordinate Plot	0.55	0.24

TABLE VIII: SCORE-RELATED T-TESTS IN LARGE-SCALE SETTING, HIGHER SCORE MEANS BETTER PERFORMANCE

Relation between Mean Scores	**t**	**p-value**
Kiviat Diagram > Parallel Coordinate Plot	5.55	$<$ **0.001**
Kiviat Diagram > Color Parallel Coordinate Plot	3.77	$<$ **0.001**
Color Kiviat Diagram > Parallel Coordinate Plot	6.51	$<$ **0.001**
Color Kiviat Diagram > Color Parallel Coordinate Plot	4.68	$<$ **0.001**
Kiviat Diagram Lines Only > Color Parallel Coordinate Plot	4.28	$<$ **0.001**
Kiviat Diagram Lines Only > Parallel Coordinate Plot	2.43	**0.010**
Color Parallel Coordinate Plot > Parallel Coordinate Plot	2.23	**0.016**
Color Kiviat Diagram > Kiviat Diagram Lines Only	2.39	**0.011**
Color Kiviat Diagram > Kiviat Diagram	0.74	0.231
Kiviat Diagram > Kiviat Diagram Lines Only	1.43	0.080
Kiviat Diagram $\neq$ Kiviat Diagram Lines Only	1.43	0.159

5.2 Response Time performance

5.2.1 Moderate-scale setting results

We found that in the moderate-scale setting, participants spent significantly less time when using the Color KD, the KDLO and the Color PCP), compared to the KD and PCP (Figure Figure 17 left, *moderate* results in Table X). Table IX shows the RT means and standard deviation per encoding, ordered by decreasing time.

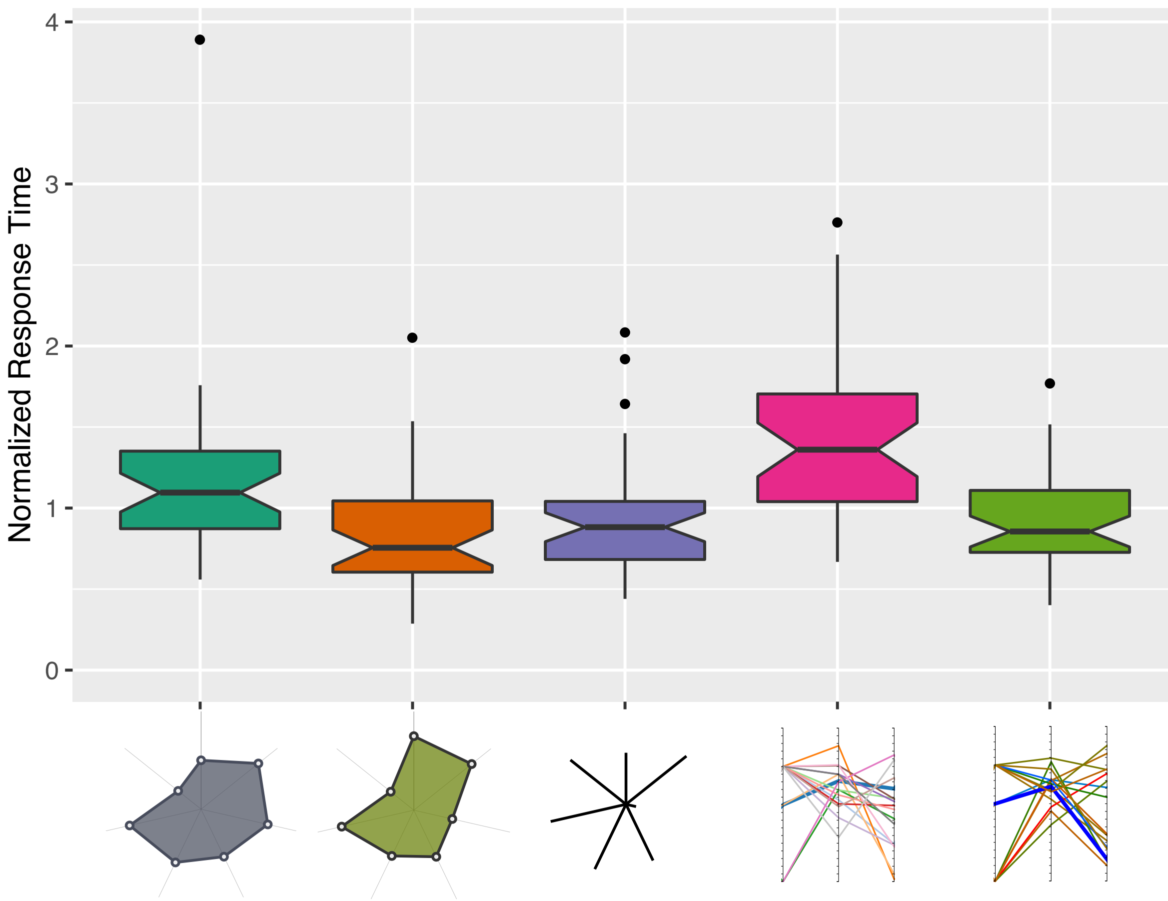


Figure 17: Notched box plots illustrating the distribution of the response time in the moderate-scale setting (normalized by user average response time). Lower response time is better, with values < 1 indicating shorter time than average and values > 1 indicating longer time than average. Response time was significantly lower for Colored KD, Colored PCP and KDLO.

TABLE IX: RESULTS OF NORMALIZED RESPONSE TIME MEAN AND STANDARD DEVIATION IN THE MODERATE-SCALE SETTING ORDERED BY DECREASING PERFORMANCE (A LOWER TIME IS BETTER)

Visual Encoding	**Mean**	**Standard Deviation**
Kiviat Diagram (colored)	0.83	0.23
Kiviat Diagram Lines Only	0.92	0.37
Parallel Coordinate Plot (colored)	0.93	0.31
Kiviat Diagram	1.16	0.54
Parallel Coordinate Plot	1.42	0.48

TABLE X: TIME-RELATED T-TESTS IN MODERATE-SCALE SETTING (LOWER TIME IS BETTER)

Relation between Mean Response Times	**t**	**p-value**
Kiviat Diagram < Parallel Coordinate Plot	1.94	**0.030**
Color Parallel Coordinate Plot < Kiviat Diagram	2.06	**0.023**
Color Kiviat Diagram < Parallel Coordinate Plot	6.11	< **0.001**
Color Kiviat Diagram < Color Parallel Coordinate Plot	1.27	0.0106
Kiviat Diagram Lines Only < Parallel Coordinate Plot	4.73	< **0.010**
Kiviat Diagram Lines Only < Color Parallel Coordinate Plot	0.07	0.472
Color Kiviat Diagram < Kiviat Diagram	3.86	< **0.010**
Kiviat Diagram Lines Only < Kiviat Diagram	2.27	**0.015**
Color Parallel Coordinate Plot < Parallel Coordinate Plot	5.77	< **0.010**
Color Kiviat Diagram < Kiviat Diagram Lines Only	1.14	0.131

5.2.2 Large-scale setting results

In the large-scale setting (Figure 18 right), the Color KD outperformed all other encodings with respect to RT. Standard KD outperformed with respect to time standard PCP ($t(39) = 1.92$, p ¡ 0.05), but not Color-cue PCP ($t(39) = 0.24$, $p = 0.405$). Table XI shows the RT means and standard deviation per encoding, ordered by decreasing time.

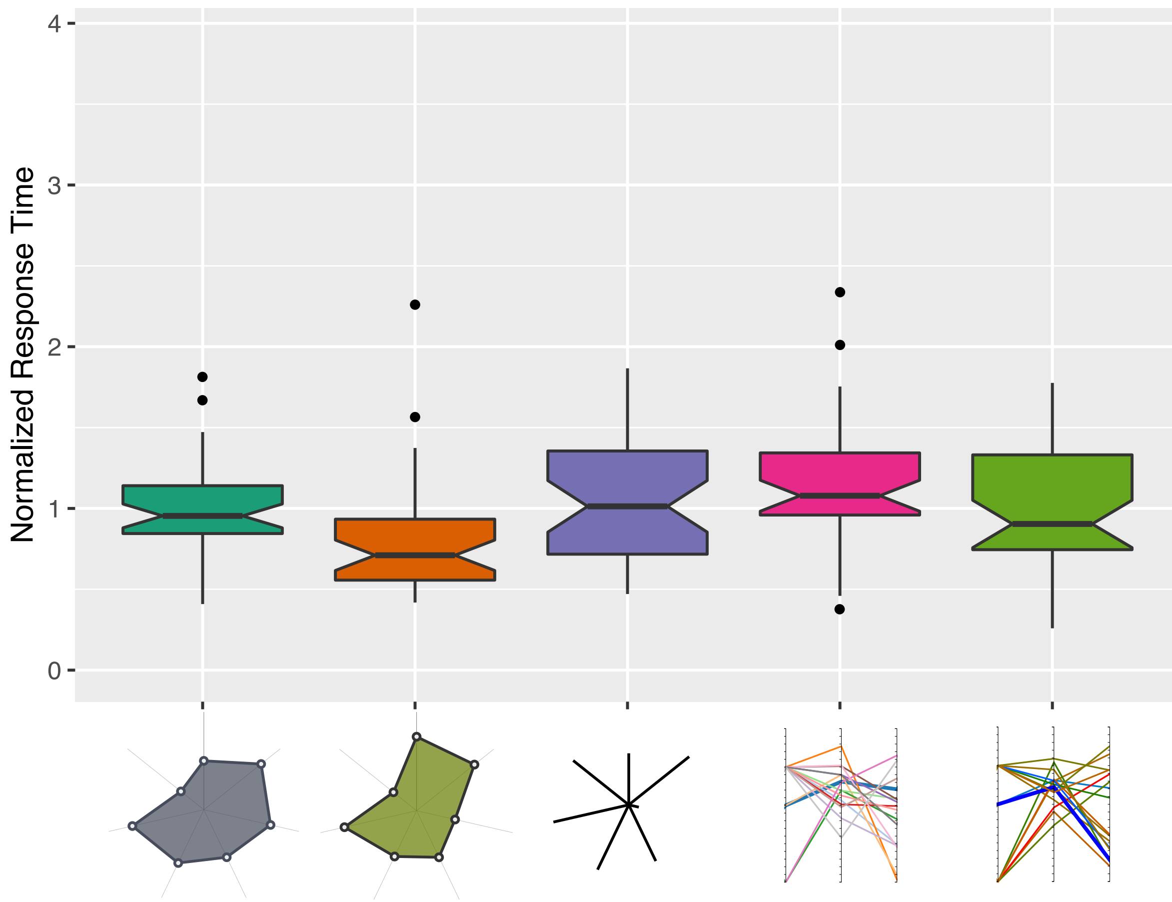


Figure 18: Notched box plots illustrating the distribution of the response time in the large-scale setting (normalized by user average response time). Lower response time is better, with values < 1 indicating shorter time than average and values > 1 indicating longer time than average. Colored KD outperformed all encodings and standard KD outperformed standard PCP.

TABLE XI: RESULTS OF NORMALIZED RESPONSE TIME MEAN AND STANDARD DEVIATION IN THE LARGE-SCALE SETTING ORDERED BY DECREASING PERFORMANCE (A LOWER TIME IS BETTER)

Visual Encoding	**Mean**	**Standard Deviation**
Kiviat Diagram (colored)	0.79	0.35
Parallel Coordinate Plot (colored)	0.98	0.36
Kiviat Diagram	1.00	0.27
Kiviat Diagram Lines Only	1.04	0.36
Parallel Coordinate Plot	1.17	0.40

TABLE XII: TIME-RELATED T-TESTS IN LARGE-SCALE SETTING (LOWER TIME IS BETTER)

Relation between Mean Response Times	**t**	**p-value**
Kiviat Diagram < Parallel Coordinate Plot	1.92	**0.031**
Color Parallel Coordinate Plot < Kiviat Diagram	0.24	0.405
Color Kiviat Diagram < Parallel Coordinate Plot	4.36	< **0.001**
Color Kiviat Diagram < Color Parallel Coordinate Plot	2.21	**0.016**
Kiviat Diagram Lines Only < Parallel Coordinate Plot	1.50	0.070
Color Parallel Coordinate Plot < Kiviat Diagram Lines Only	0.71	0.240
Color Kiviat Diagram < Kiviat Diagram	3.10	**0.002**
Kiviat Diagram < Kiviat Diagram Lines Only	0.52	0.302
Color Parallel Coordinate Plot < Parallel Coordinate Plot	2.66	**0.006**
Color Kiviat Diagram < Kiviat Diagram Lines Only	3.00	**0.002**

5.3 Encoding scalability with the number of items

All encodings, with one notable exception, had worse performance when changing from the moderate scale setting to the large scale setting. We found that only Color KD did not perform significantly worse ($t(39) = 1.514$, $p = 0.138$). Likewise, we found that the normalized response times did not get significantly worse at larger scales for the Color KD ($t(39) = 0.49$, $p = 0.324$)

and Color-cue PCP ($t(39) = 0.82$, $p = 0.21$), while users response time was significantly slower for all encodings that did not use color.

5.4 Time-score regression analysis

In Figure 19 we plot the average user score for each participant against the average response time for the whole experiment. The positive correlation is evident between the two measures suggests that increasing time spent on the task does lead to improved performance. It is also possible to notice, however, a considerable number of outliers. Figure 20 shows the same data group by demographic category.

Figure 21, Figure 22, and Figure 23 show the correlation between time and score for individual encodings. It is possible to notice an initial positive correlation in most curves up to a point in time. After that, spending more time on the task does not lead to improved performance. Particularly, we found very distinct behavior when color is used: in PCP, colored variants plots (22(c) 22(d)) show that spending more time on the task leads to higher score, while this is not true for normal PCP versions (22(a) 22(b)).

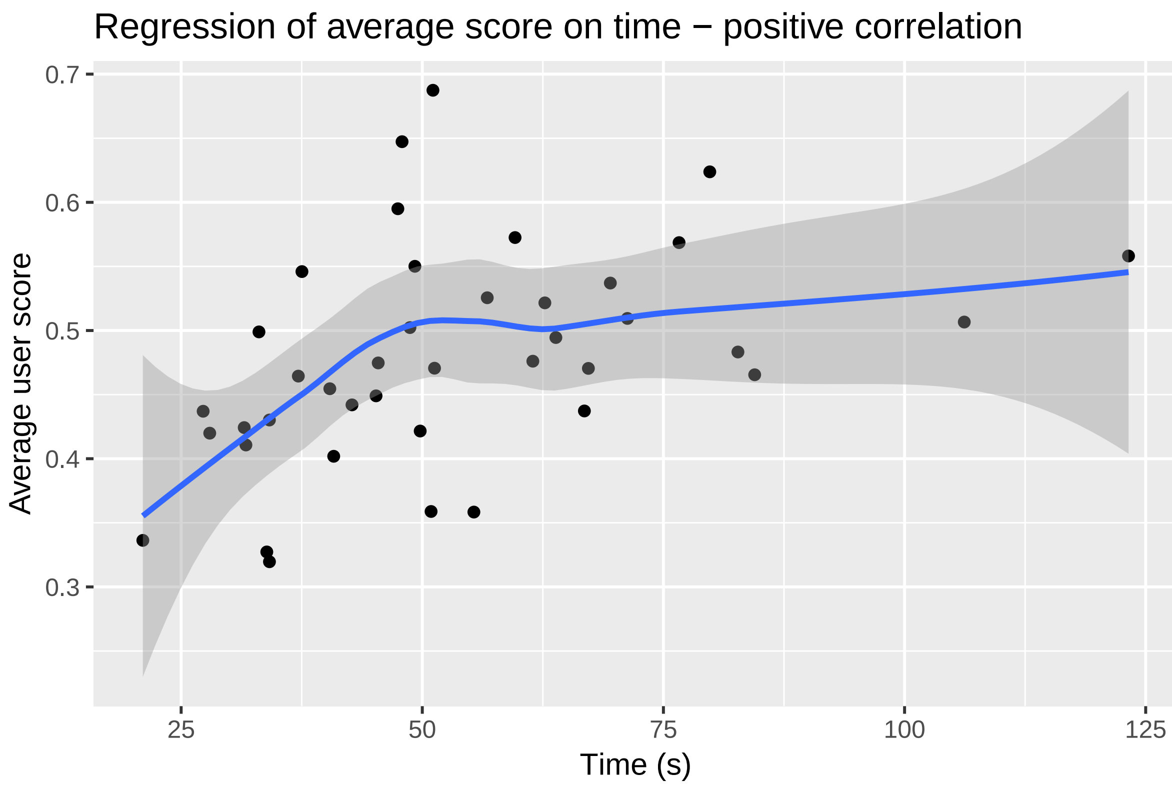
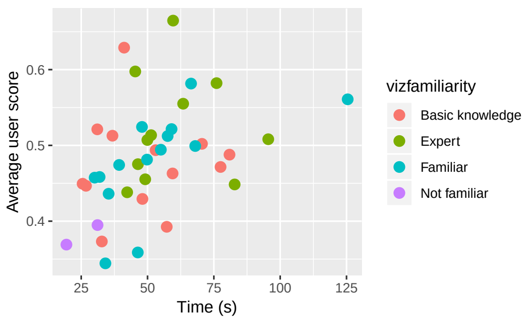
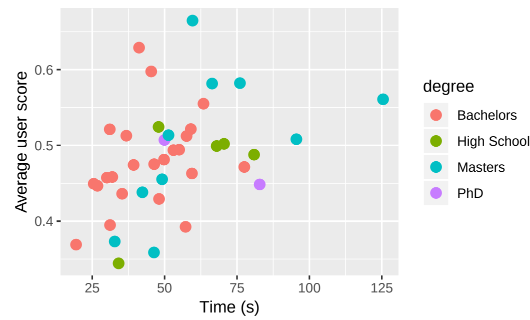


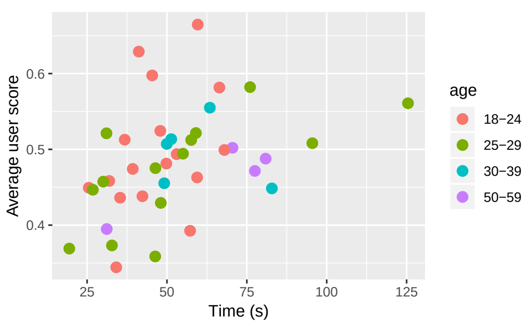
Figure 19: The average user score against the average response time for each participant in the study. While some outliers are present, we find evidence of a positive correlation between the two measures as shown by the robust linear fit



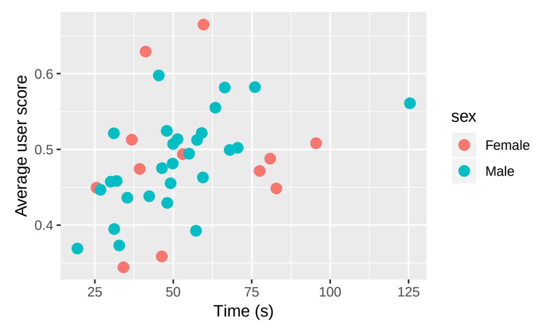
(a) Familiarity with Visualization



(b) Highest degree obtained

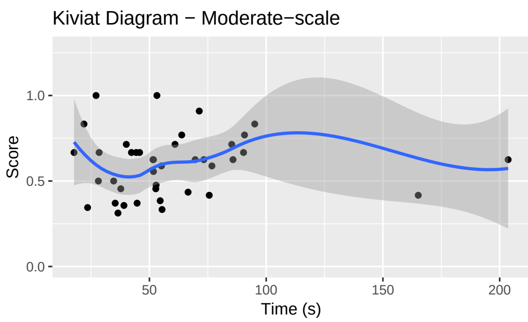


(c) Age group

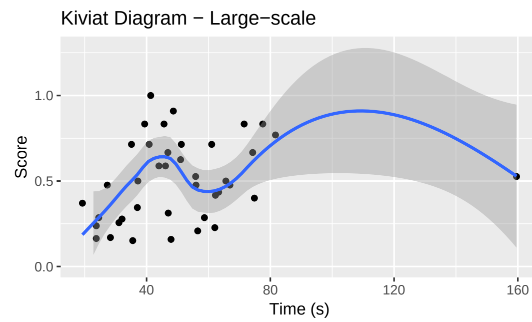


(d) Sex

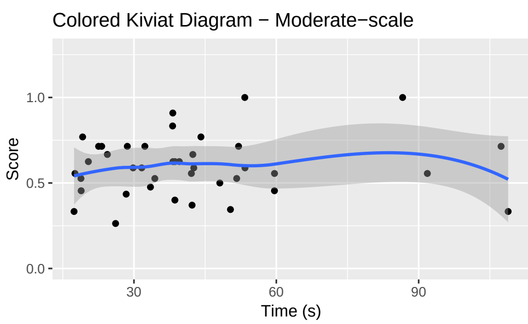
Figure 20: Time-score correlation grouped by category.



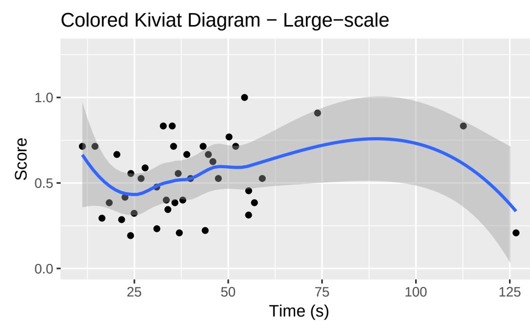
(a) Moderate-scale KD



(b) Large-scale KD

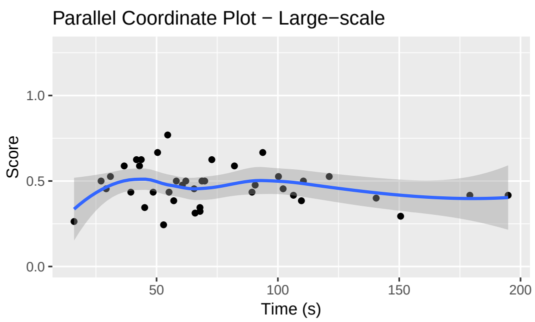


(c) Moderate-scale Colored KD

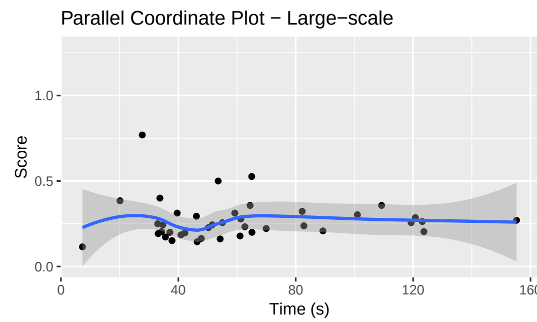


(d) Large-scale Colored KD

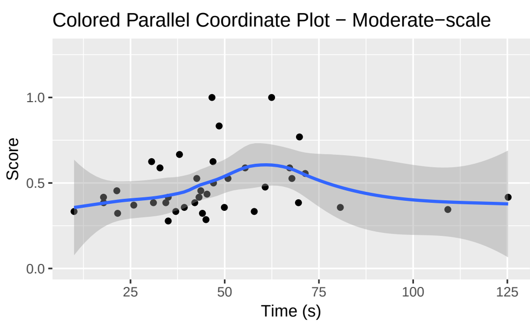
Figure 21: Regression of average score on time - KD



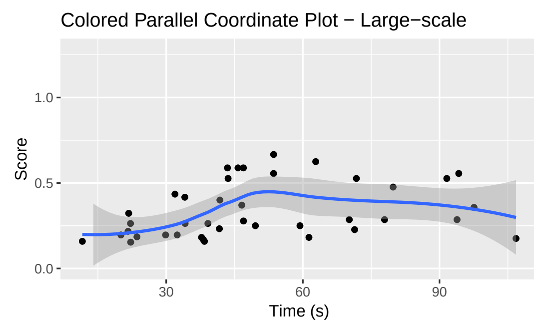
(a) Moderate-scale PCP



(b) Large-scale PCP

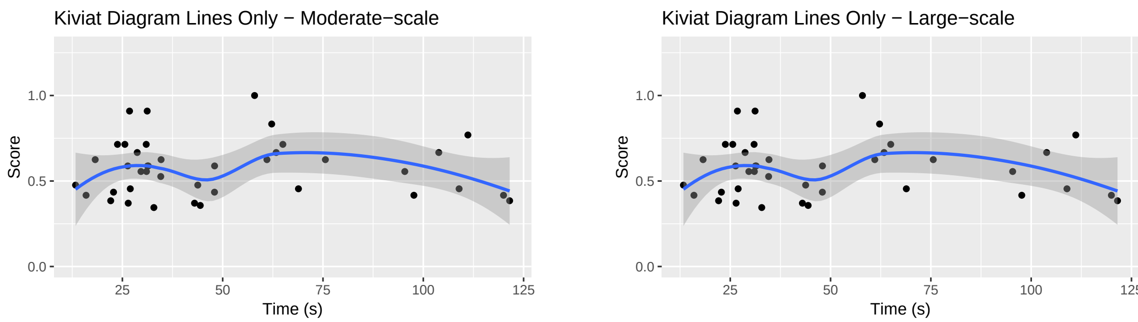


(c) Moderate-scale Colored PCP



(d) Large-scale Colored PCP

Figure 22: Regression of average score on time - PCP



(a) Moderate-scale KDLO

(b) Large-scale KDLO

Figure 23: Regression of average score on time - KDLO

5.5 Demographics-related findings

As expected, we found that participants with higher self-reported expertise in visualization obtained higher task scores. We did not perform statistical comparison between these groups as we only had two participants not familiar with visualization. Figure 24 shows how people with less familiarity with visualization achieve lower scores compared to people with some familiarity, and very high scores are only achieved by experts.

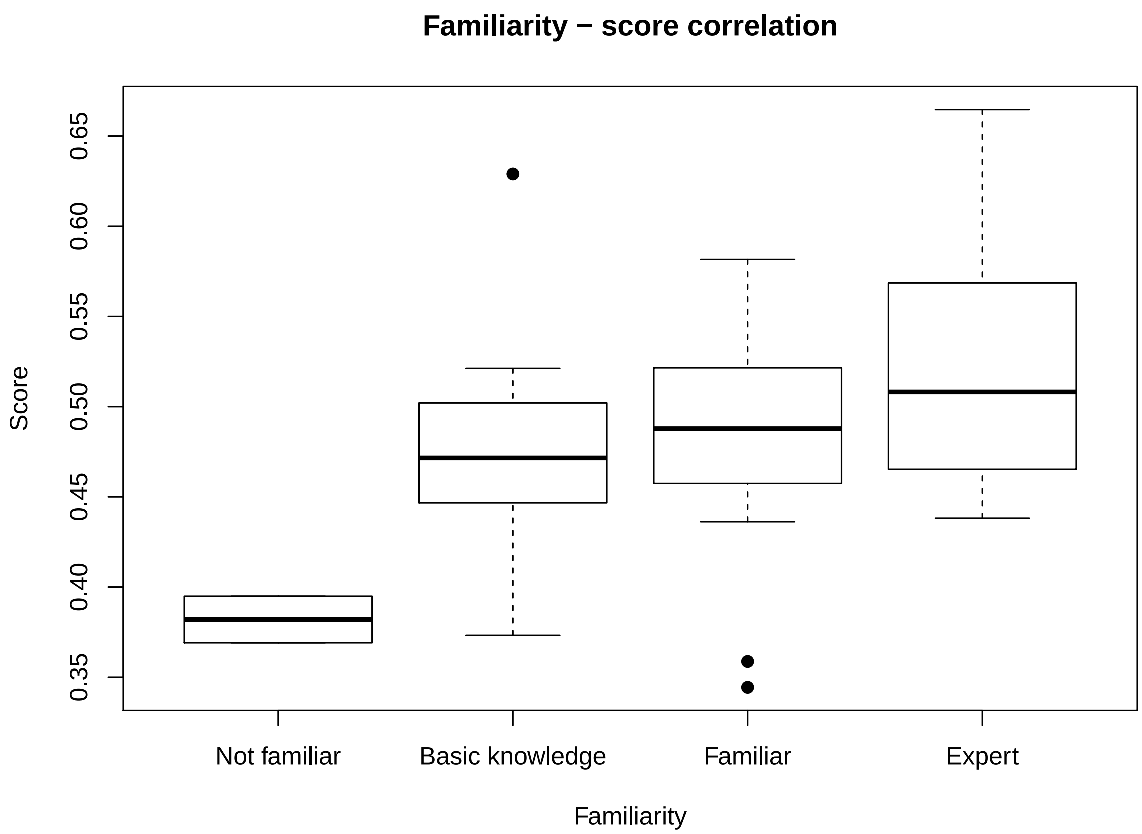


Figure 24: Box plots of users' scores grouped by visualization familiarity. There is a positive correlation between visualization expertise and score.

5.6 Qualitative feedback

In terms of qualitative feedback, 22 out of 40 participants considered the PCP to be the most challenging visual encoding to read (Fig. Figure 25). Users also considered the PCP to have the worst scalability potential (25 out of 40); only 3 participants thought PCP was the most scalable. For PCP, users reported that "it is hard to keep in memory the features of the target item while comparing with others" and "colors become too similar, especially when dealing with a large number of items". Most users reported experiencing difficulties telling items apart, for reasons related to the presence of a large amount of partially or totally overlapping segments, and to color encoding.

Despite similar results in terms of score and time between KD and KDLO, participants furthermore reported that KD were easier to read than KDLO (11 votes against 3) and that they scale better (11 votes against 5). Only 1 participant believed that KD do not scale well, while KDLO received 8 negative votes from people who did not endorse their scalability.

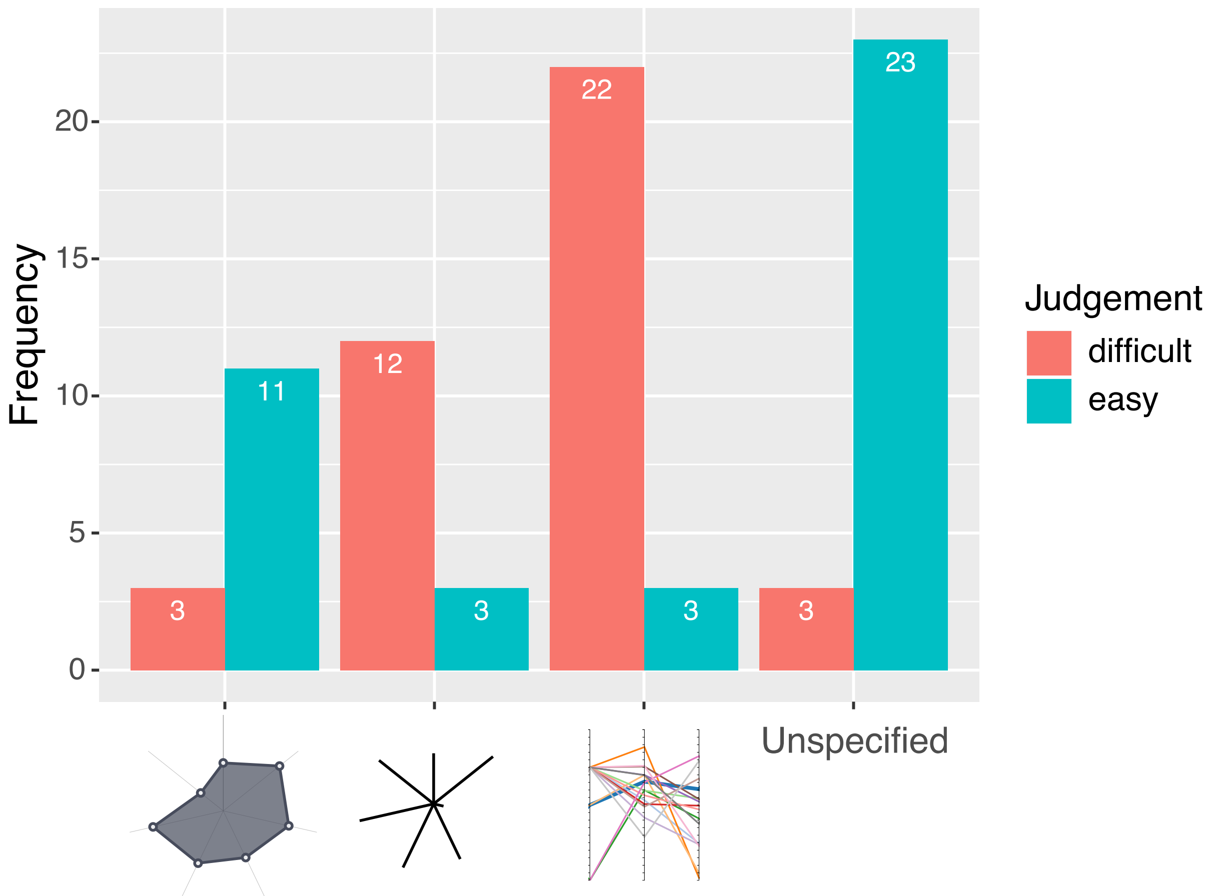


Figure 25: Bar chart showing the judgement (easiest/most difficult visual encoding) which participants expressed. Among the users who expressed an opinion, most participants considered the Kiviat diagrams to be the easiest to read visual encoding, whereas the parallel coordinate plots were the most difficult to read. The Unspecified category represents people who thought some other encoding who would be easier/more difficult, or people who thought more encodings tied.

5.7 Discussion

In this section, I will discuss the results and compare them to the initial hypotheses we made.

5.7.1 Hypothesis H1

Evaluation of the user study results shows that for accuracy in similarity detection, juxtaposed star plot variants (Color KD, KD, KDLO) outperform superposed encodings (plain PCPs, color PCPs) at both the moderate and the large scale. We therefore reject hypothesis **H1** (At moderate scale (16 items), all encodings studied will yield equivalent scores), as encodings do not have equal performance at moderate scale: superposed encodings are worse.

5.7.2 Hypothesis H2

At the same time, we accept hypothesis **H2** (At the larger scale, juxtaposed encodings will outperform superposed encodings), and we furthermore find that it holds at moderate scale as well. This is an interesting finding, because our moderate and large scales are significantly below the suggested PCP effectiveness threshold of thousands of items [22]. One interpretation could be that, whereas PCPs are excellent tools for correlation detection and analysis, similarity detection does not necessarily involve inter-variable correlation. A second interpretation could be that, whereas practitioners use basic variations of PCPs which are readily available in visualization software, investing in more advanced PCP variations could help. We also note that color in PCPs helps with respect to similarity detection scores, but does so only at the larger scale.

5.7.3 Hypothesis H3

Within the star plot variants, Color KD and KD scored similarly at both scales. At the large scale, Color KD scored significantly better than the KDLO, while at moderate scale, all star plot variants perform similarly with respect to score. These findings are in direct contradiction with the results of Fuchs et al [44], who found that KDLO outperformed all other star plot variations, including KD (a.k.a. Contour Only star plots). We believe the experimental setting explains this discrepancy. First, the number of items shown (scale) influences the layout of individual encodings: in a 3 x 3 grid layout (9 items) with the target located centrally, as in Fuchs et al, similarity detection turns into a series of one-on-one comparison tasks with all items near the target. When scale increases, as in our experiments, items can no longer be placed immediately side by side for comparison. Second, when the number of variables encoded in the KD is artificially large, as in the synthetic data used in the Fuchs et al study, the shape of a Kiviat becomes harder to discern. With lower-dimensional items, KD benefit more from the pre-attentive nature of shape.

As a result, we reject hypothesis **H3** (At the larger scale, Lines-Only encodings will outperform other encodings), and we furthermore find that **H3** does not hold at the moderate scale, either. Overall, our results support the use of KD and colored KD for similarity detection at scales and with item complexities used in this study.

As expected, most encodings yield significantly worse scores as scale grows. The one notable exception are Color KD, which further supports their use in practice.

5.7.4 Hypothesis H4

With respect to time, we found significant differences across encodings in both settings. At the larger scale, Color KD were faster than any other encoding. At the moderate scale, Color KD, KDLO, and Color PCPs were significantly faster, although time was not correlated with scores. All non-colored variants had a longer response time at larger scale than at moderate scale, whereas colored variants had, statistically, the same response time at both scales. Color variants where faster than their non-color equivalents at the larger scale: for both encoding types, the colored version can be assumed to be faster to read than the non-colored version (p ¡ 0.01). However, color did not help star plots with respect to score. **H4** is therefore mostly valid (At both scales, color-cue encodings outperform other encodings with respect to time, but not with respect to score). The one exception are superposed encodings, where color also improves score; however, even this score improvement should be interpreted within the context of **H1**, where their score is significantly below that of star-plot variants. Because color does not lead to better similarity detection scores for KD at either scale, it may be safe to map the Kiviat color to one of the item variables, as currently done in practice.

5.7.5 Qualitative feedback

The qualitative feedback also indicates that juxtaposed KD are easier to read and interpret than both Lines-Only star plot encodings and supeimrposed encodings. The participants' wish for an easier encoding is understandable, but may be difficult to satisfy, given documented difficulties in multivariate glyph design.

5.7.6 User background effect

In terms of the effect of user background, we note that most of our participants were knowledgeable about visual encodings, with two notable outliers. While statistical analysis is not feasible on such a small sample, we note that the two users with no visualization expertise were outliers with respect to both score (worse scores) and time (less time).

5.8 Limitations

In terms of limitations, our study examines two relatively modest sized settings, and a relatively moderate set of variables, based on a real dataset. However, the significant variation in encoding performance indicates that even this scale can capture and document visual encoding scalability issues. Our study also did not include more advanced, not readily available, variations of PCP, beyond the basic interactive plots. Whereas such variations may improve superposed results, we note that their reduced availability means they are often not used in practice, whereas the basic variations included in this study are. In the interest of reducing the number of trials, and based on critiques of superposed radial layouts, we tested only linear superposed layouts. However, given the qualitative user feedback and our numerical results, we do not expect radial superposed layouts would yield better scores. Last but not least, in terms of generalizability, our findings may not be feasible in similarity detection tasks that involve thousands of items, simply because juxtaposition requires a larger screen real estate footprint than superposition.

Despite these limitations, our study involved 40 participants, 2 scale settings, and 14 trials per session. This setup enabled us to successfully investigate how accurately and quickly people can identify similar data points using different encodings. We found that, independent of the user level of expertise, Kiviat diagrams were the most suitable encoding in terms of accuracy, especially in a large-scale context. In contrast, superposition-based approaches under-perform due to difficulties in distinguishing between items, and are often slower to read.

With respect to similarity detection, overall, for multivariate collections of a handful to several dozens of items, we recommend the use of juxtaposed KD. In the case of collections of hundreds to thousands of items, where space becomes an issue and juxtaposition may not be feasible, we strongly recommend the use of more advanced variants of PCPs than the plain PCP versions currently used by practitioners. Doing so may require the explicit integration of these more advanced variants into popular visualization platforms.

CHAPTER 6

CONCLUSION

In conclusion, we contributed to the comparative visualization field by evaluating the effectiveness of 5 different VE for similarity detection in the context of MD and many items. The proposed and analyzed charts were selected by state-of-the-art VE.

Our experiment, involved 40 participants, 2 scale settings (16 and 36 items), and 10 trials in total per session. We investigated how accurate and fast people can be when they try to find similar data points using different charts. Our statistical analysis shows that there are significant differences in encoding performance, especially in the larger-scale setting of the experiment. We found that, independently of their expertise, KD are the most suitable encoding in terms of accuracy, especially in a large-scale context.

In all settings, we found that plain PCP are slower to read and lead to lower accuracy than juxtaposed (side-by-side) star plot approaches. Among juxtaposed approaches, star plot variants like the KD and Color KD encodings scale well, are reasonably fast to read, and achieve good accuracy. When the number of items grows, the juxtaposed Color KD outperforms other encodings, including the KDLO, and is therefore suitable for similarity detection when dealing with larger multivariate datasets. Color can help diminish response time, especially when the number of items grows.

Based on our experimental results, since we narrowed the selection down to a few good VE for similarity detection, we believe future research should concentrate on further exploring these particular encodings in diverse settings. We plan to examine how their behavior and performance vary with a different number of attributes and when the number of items increases ulteriorly.

CITED LITERATURE

1. Chan, W. W.-Y.: A survey on multivariate data visualization. 2006.

2. Gleicher, M., Correll, M., Nothelfer, C., and Franconeri, S.: Perception of average value in multiclass scatterplots. IEEE Transactions on Visualization and Computer Graphics, 19(12):2316–2325, 2013.

3. Wehrend, S. and Lewis, C.: A problem-oriented classification of visualization techniques. In Proceedings of the First IEEE Conference on Visualization: Visualization '90, pages 139–143, 1990.

4. Luciani, T., Burks, A., Sugiyama, C., Komperda, J., and Marai, G. E.: Details-first, show context, overview last: Supporting exploration of viscous fingers in large-scale ensemble simulations. IEEE Transactions on Visualization and Computer Graphics, 25(1):1225–1235, 2019.

5. Ma, C.: Visual Analysis Techniques for Dynamic Biological Networks. Doctoral dissertation, University of Illinois at Chicago, 2018. Available at http://hdl.handle.net/10027/22644.

6. Choi, J., Lee, S.-E., Cho, E., Kashiwagi, Y., Okabe, S., Chang, S., and Jeong, W.-K.: Interactive dendritic spine analysis based on 3d morphological features. VIS'19 Proceedings, 2019.

7. Marai, G. E., Ma, C., Burks, A. T., Pellolio, F., Canahuate, G., Vock, D. M., Mohamed, A. S. R., and Fuller, C. D.: Precision risk analysis of cancer therapy with interactive nomograms and survival plots. IEEE Transactions on Visualization and Computer Graphics, 25(4):1732–1745, 2019.

8. Thomas, M. M., Kannampallil, T., Abraham, J., and Marai, G. E.: Echo: A large display interactive visualization of icu data for effective care handoffs. In 2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC), pages 47–54. IEEE, 2017.

9. Meyer-Spradow, J., Stegger, L., Döring, C., Ropinski, T., and Hinrichs, K.: Glyph-based spect visualization for the diagnosis of coronary artery disease. IEEE Transactions on Visualization and Computer Graphics, 14(6):1499–1506, 2008.

CITED LITERATURE (continued)

10. Jäckle, D., Fuchs, J., and Keim, D. A.: Star glyph insets for overview preservation of multivariate data. In Visualization and Data Analysis, 2016.

11. Fischer, F., Fuchs, J., and Mansmann, F.: Clockmap: Enhancing circular treemaps with temporal glyphs for time-series data. In EuroVis (Short Papers), pages 97–101, 2012.

12. Kintzel, C., Fuchs, J., and Mansmann, F.: Monitoring large ip spaces with clockview. ACM International Conference Proceeding Series, 2011.

13. Legg, P., Chung, D., Parry, M., Jones, M., Long, R., Griffiths, I., and Chen, M.: Matchpad : Interactive glyph-based visualization for real-time sports performance analysis. Computer Graphics Forum, 31:1255–1264, 2012.

14. Woodring, J. and Shen, H.: Multi-variate, time varying, and comparative visualization with contextual cues. IEEE Transactions on Visualization and Computer Graphics, 12(5):909–916, 2006.

15. Healey, C. G.: Perception in visualization. 2005.

16. Keim, D. A.: Designing pixel-oriented visualization techniques: theory and applications. IEEE Transactions on Visualization and Computer Graphics, 6(1):59–78, 2000.

17. van Wijk, J. J.: Views on visualization. IEEE Transactions on Visualization and Computer Graphics, 12(4):421–432, 2006.

18. Gleicher, M.: Considerations for visualizing comparison. IEEE Transactions on Visualization and Computer Graphics, 24(1):413–423, 2018.

19. Munzner, T.: A nested model for visualization design and validation. IEEE Transactions on Visualization and Computer Graphics, 15(6):921–928, 2009.

20. Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L., and Zhou, Y.: Treejuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. ACM Trans. Graph., 22:453–462, 2003.

21. Meyer, M., Munzner, T., and Pfister, H.: Mizbee: A multiscale synteny browser. IEEE Transactions on Visualization and Computer Graphics, 15(6):897–904, 2009.

CITED LITERATURE (continued)

22. Munzner, T.: Visualization Analysis and Design. AK Peters Visualization Series. CRC Press, 2015.

23. Kosara, R.: Circular Part-to-Whole Charts Using the Area Visual Cue. In EuroVis 2019 - Short Papers, eds. J. Johansson, F. Sadlo, and G. E. Marai. The Eurographics Association, 2019.

24. Andrienko, N. and Andrienko, G.: Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach. Berlin, Heidelberg, Springer-Verlag, 2005.

25. Siirtola, H.: Combining parallel coordinates with the reorderable matrix. In Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization - CMV 2003 -, pages 63–74, 2003.

26. Tufte, E.: Envisioning Information. Cheshire, CT, USA, Graphics Press, 1990.

27. Keim, D. A.: Visual techniques for exploring databases. In International Conference on Knowledge Discovery in Databases (KDD'97), Newport Beach, CA, USA, 1997.

28. Keim, D. A. and Kriegel, H. .: Visualization techniques for mining large databases: a comparison. IEEE Transactions on Knowledge and Data Engineering, 8(6):923–938, 1996.

29. Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C. D., and Roberts, J. C.: Visual comparison for information visualization. Information Visualization, 10(4):289–309, 2011.

30. Tominski, C., Forsell, C., and Johansson, J.: Interaction support for visual comparison inspired by natural behavior. IEEE Transactions on Visualization and Computer Graphics, 18(12):2719–2728, 2012.

31. Kehrer, J., Piringer, H., Berger, W., and Gröller, M. E.: A model for structure-based comparison of many categories in small-multiple displays. IEEE Transactions on Visualization and Computer Graphics, 19(12):2287–2296, 2013.

32. Lee, M. D., Reilly, R. E., and Butavicius, M. E.: An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In Proceedings of the Asia-Pacific Symposium on Information Visualisation - Volume 24, APVis '03, pages 1–10, Darlinghurst, Australia, Australia, 2003. Australian Computer Society, Inc.

CITED LITERATURE (continued)

33. Borgo, R., Kehrer, J., Chung, D. H., Maguire, E., Laramee, R. S., Hauser, H., Ward, M., and Chen, M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In Eurographics (STARs), pages 39–63, 2013.

34. Ropinski, T. and Preim, B.: Taxonomy and usage guidelines for glyph-based medical visualization. pages 121–138, 2008.

35. Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B.: What do we need to build explainable ai systems for the medical domain? arXiv, 2017.

36. Morton, G. M.: A computer oriented geodetic data base and a new technique in file sequencing. International Business Machines Company New York, 1966.

37. Wattenberg, M.: A note on space-filling visualizations and space-filling curves. In IEEE Symposium on Information Visualization., 2005.

38. Chernoff, H.: The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, 68(342):361–368, 1973.

39. Ware, C.: Information Visualization: Perception for Design. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 3 edition, 2012.

40. Fuchs, J., Isenberg, P., Bezerianos, A., and Keim, D.: A systematic review of experimental studies on data glyphs. IEEE Transactions on Visualization and Computer Graphics, 23(7):1863–1879, 2017.

41. Borg, I. and Staufenbiel, T.: Performance of snow flakes, suns, and factorial suns in the graphical representation of multivariate data. Multivariate Behavioral Research, 27(1):43–55, 1992.

42. Klippel, A., Hardisty, F., and Weaver, C.: Star plots: How shape characteristics influence classification tasks. Cartography and Geographic Information Science, 36(2):149–163, 2009.

43. Miller, M., Zhang, X., Fuchs, J., and Blumenschein, M.: Evaluating ordering strategies of star glyph axes, 2019.

44. Fuchs, J., Isenberg, P., Bezerianos, A., Fischer, F., and Bertini, E.: The influence of contour on similarity perception of star glyphs. IEEE Transactions on Visualization and Computer Graphics, 20(12):2251–2260, 2014.

CITED LITERATURE (continued)

45. Schmid, C. and Hinterberger, H.: Comparative multivariate visualization across conceptually different graphic displays. In Seventh International Working Conference on Scientific and Statistical Database Management, pages 42–51, 1994.

46. Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F.: Species distribution modeling based on the automated identification of citizen observations. Applications in Plant Sciences, 6(2):e1029, 2018.

47. Poco, J., Dasgupta, A., Wei, Y., Hargrove, W., Schwalm, C., Cook, R., Bertini, E., and Silva, C.: Similarityexplorer: A visual inter-comparison tool for multifaceted climate data. Computer Graphics Forum, 33(3):341–350, 2014.

48. hanelarrison, Harrison, L., Gramazio, C., Yáng, F., Aragam, K., Peck, E., and Schroeder, D.: codementum/experimentr: v1.0, 2019.

49. Morris, M. F.: Kiviat graphs: conventions and figures of merit. ACM SIGMETRICS Performance Evaluation Review, 3(3):2–8, 1974.

50. Brewer, C.: Colorbrewer 2.0. Online; accessed October 2019.

51. Brown, S.-A.: Patient similarity: Emerging concepts in systems and precision medicine. Frontiers in Physiology, 7, 2016.

52. Huang, A.: Similarity measures for text document clustering. Proceedings of the 6th New Zealand Computer Science Research Student Conference, 2008.

53. Shapiro, S. S. and Wilk, M. B.: An analysis of variance test for normality (complete samples). Biometrika, 1965.

VITA

NAME	Mirko Mantovani
EDUCATION	
	B.S. in Computer Engineering, Politecnico di Milano, July 2017, Italy. Final Grade: 109/110
	M.S. in Computer Engineering, Politecnico di Milano, April 2020, Italy. GPA: 29.3/30
	M.S. in Computer Science, University of Illinois at Chicago, May 2020, USA. GPA: 4.0/4.0
LANGUAGE SKILLS	
Italian	Native speaker
English	Full working proficiency
	2018 - TOEFL examination (100/120)
SCHOLARSHIPS	
Jul 2014	Best freshmen scholarship - Politecnico di Milano
Sep 2015 - Jul 2019	Tuition exemption for high academic performance - Politecnico di Milano
Jan - Dec 2019	Research Assistantship (RA) position (20 hours/week) with full tuition waiver plus monthly stipend - University of Illinois at Chicago