On Intelligible Multimodal Visual Analysis

Jan-Frederik Kassel

Leibniz Universität Hannover

On Intelligible Multimodal Visual Analysis

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
genehmigte Dissertation

von Herrn
M.Sc. Jan-Frederik Kassel
geboren am 22.05.1990 in Hannover

2020

On Intelligible Multimodal Visual Analysis

Dissertation

Jan-Frederik Kassel

1. Referent: Prof. Dr. Michael Rohs
2. Referent: Prof. Dr. Andreas Butz

Tag der Promotion: 02.09.2020

Gottfried Wilhelm Leibniz Universität Hannover
Fachgebiet für Mensch-Computer-Interaktion
Institut für Mensch-Computer-Kommunikation
Fakultät für Elektrotechnik und Informatik
Appelstr. 9A
30167 Hannover

Abstract

Analyzing data becomes an important skill in a more and more digital world. Yet, many users are facing knowledge barriers preventing them to independently conduct their data analysis. To tear down some of these barriers, multimodal interaction for visual analysis has been proposed. Multimodal interaction through speech and touch enables not only experts, but also novice users to effortlessly interact with such kind of technology. However, current approaches do not take the user differences into account. In fact, whether visual analysis is intelligible ultimately depends on the user.

In order to close this research gap, this dissertation explores how multimodal visual analysis can be personalized. To do so, it takes a holistic view. First, an intelligible task space of visual analysis tasks is defined by considering personalization potentials. This task space provides an initial basis for understanding how effective personalization in visual analysis can be approached. Second, empirical analyses on speech commands in visual analysis as well as used visualizations from scientific publications further reveal patterns and structures. These behavior-indicated findings help to better understand expectations towards multimodal visual analysis. Third, a technical prototype is designed considering the previous findings. Enriching the visual analysis by a persistent dialogue and a transparency of the underlying computations, conducted user studies show not only advantages, but address the relevance of considering the user's characteristics. Finally, both communications channels – visualizations and dialogue – are personalized. Leveraging linguistic theory and reinforcement learning, the results highlight a positive effect of adjusting to the user. Especially when the user's knowledge is exceeded, personalizations helps to improve the user experience.

Overall, this dissertations confirms not only the importance of considering the user's characteristics in multimodal visual analysis, but also provides insights on how an intelligible analysis can be achieved. By understanding the use of input modalities, a system can focus only on the user's needs. By understanding preferences on the output modalities, the system can better adapt to the user. Combining both directions improves user experience and contributes towards an intelligible multimodal visual analysis.

Keywords: Visual Analysis, Personalization, Multimodality

Zusammenfassung

Daten spielen eine immer wichtigere Rolle. Viele Anwender interessieren sich für die Nutzung ihrer Daten, jedoch fehlt ihnen oft das nötige Wissen, um dies zu tun. Dies kreiert Nutzungsbarrieren für den Anwender. Ein Ansatz um Daten zugänglicher zu machen ist die Nutzung von Visualisierungen. Ferner werden neue Interaktionsmodalitäten untersucht, die insbesondere durch moderne Technologien einsetzbar werden. Dabei erscheint die Kombination aus Sprache und Gesten besonders effektiv für die visuelle Datenanalyse. Erste Ergebnisse zeigen vielversprechende Vorteile, es fehlt jedoch der Anwender in Gleichung. Je nach Anwender ist nämlich eine Analysesituation einfacher oder schwerer verständlich. Eine Anpassung an den Anwender ist somit essentiell.

Um diese Forschungslücke zu schließen erforscht die vorliegende Dissertation Methoden für eine verständliche, multimodale, visuelle Datenanalyse in Abhängigkeit vom Anwender. Dabei nimmt diese Arbeit einen gesamtheitlichen Blick auf den Forschungsgegenstand ein. Initial diskutiert die Arbeit Personalisierungsmöglichkeiten von typischen Aktionen in der visuellen Datenanaylse, um herauszufinden an welchen Stellen eine Personalisierung besonders effektiv sein könnte. Um ferner dem Anwender besser zu unterstützen, muss verstanden werden welchen Mustern Anwendern wahrscheinlich in einer multimodalen, visuellen Datenanalyse folgen werden. Hierzu exploriert die Arbeit tatsächlich verwendete Visualisierung aus wissenschaftlichen Publikationen, als auch die Struktur vor Sprachbefehlen mit einem potentiellen System. Basierend auf diesen Erkenntnissen zeigt die Arbeit wie ein technischer Prototyp aussehen kann, um anschließend die Hauptkommunikationswege zu personalisieren. Dabei bedient sich die Arbeit bei linguistischen Grundlagen und neuesten Ansetzen aus dem maschinellen Lernen.

Insgesamt bestätigt die Dissertation die Annahme, dass Anwender unterschiedliche Bedürfnisse an eine visuellen Datenanaylse haben. Es zeigt sich während der Arbeit immer wieder, wie unterschiedlich die Anwender sind und das eine monotone Interaktionsstrategie mit dem Anwender nicht zielführend ist. So zeigt sich insbesondere die Anpassung an die Sprache des Anwenders als besonders effektiv. Letztendlich liefert diese Dissertation neue Erkenntnisse über Personalisierungsmöglichkeiten und Rahmenbedingungen für eine verständliche, multimodale, visuelle Datenanalyse, sodass zukünftige Arbeiten darauf aufbauen können.

Schlagworte: Visuelle Datenanalyse, Personalisierung, Multimodalität

Danksagung

Während der fünf Jahre, in denen diese Dissertation entstanden ist, erhielt ich wertvolle Unterstützung von Wegbegleitern, bei denen ich mich an dieser Stelle bedanken möchte.

Als Erstes möchte ich mich bei Prof. Dr. Michael Rohs bedanken. Michael gab mir initial die Chance zu promovieren. Dank seiner hervorragenden wissenschaftlichen Betreuung, in der die menschliche Komponente niemals fehlte, konnte ich meine Dissertation erfolgreich abschließen. Ebenso möchte ich mich bei meinen Doktorandenkollegen am Institut Tim Dünte und Maximilian Schrapel für die stets inspizierenden Diskussionen bedanken.

Als Zweites möchte ich mich bei meinen Kollegen bedanken. Dort fand ich Inspiration, moralischen Support und Freiraum für diese Arbeit. Bedanke möchte ich mich bei Dr. Maria Niessen, Dr. Fabienne Braune, Hamza Usmani, Christian Altergott, Dr. Ehsan Asgari, Robert Willi, Christian Pletl, Richard Niestroj, Dr. Martin Leib, Michael Streif, Peter Mayer, Michelle Plötner, Dr. Djalel Benbouzid, Barbara Sichler und André Radon. Besonders bedanke möchte ich mich bei Dr. Christoph Ringlstetter, Dr. Daniel Weimer, Dr. Justin Bayer und Dr. Sebastian Kaiser, die mich stets unterstützt und vorangetrieben haben.

Als Drittes möchte ich mich bei meinen Freunden bedanken. In einem Zeitraum in dem Zweifel und Motivationsdellen auftraten gaben sie mir Halt. Im speziellen möchte ich mich dabei bei Christian Kater, Dr. Florian Schmidt, Franziska Schallenberg, Wadim Ortlieb, Bernhard Pflugfelder, Dr. Marian Harbach, Claudia Wieschollek, Jan Müller, Sebastian Freigang, und Dr. André Sydow bedanken.

Schließlich möchte ich meinen Verwandten Gerda, Heinz-Dieter, Gerd und Rita und besonderns meinen Eltern – Beate und Bernd – danken. Ohne sie wäre es weder möglich gewesen nach München zu gehen, diese Dissertation zu schreiben, noch hätte ich mein Informatikstudium absolvieren können. Ihnen gilt letztendlich mein größter Dank.

Preamble

This dissertation extends the following previously published articles:

- Jan-Frederik Kassel and Michael Rohs (2019b). "Talk to Me Intelligibly: Investigating An Answer Space to Match the User's Language in Visual Analysis". In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. DIS '19. San Diego, CA, USA: ACM, pp. 1517–1529. ISBN: 978-1-4503-5850-7. DOI: 10.1145/3322276.3322282

- Jan-Frederik Kassel and Michael Rohs (2019a). "Online Learning of Visualization Preferences through Dueling Bandits for Enhancing Visualization Recommendations". In: *EuroVis 2019 - Short Papers*. Ed. by Jimmy Johansson et al. Porto, Portugal: The Eurographics Association. ISBN: 978-3-03868-090-1. DOI: 10.2312/evs.20191175

- Jan-Frederik Kassel and Michael Rohs (2018). "Valletto: A Multimodal Interface for Ubiquitous Visual Analytics". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. Montreal QC, Canada: ACM, LBW005:1–LBW005:6. ISBN: 978-1-4503-5621-3. DOI: 10.1145/3170427.3188445

- Jan-Frederik Kassel and Michael Rohs (2017). "Immersive Navigation in Visualization Spaces through Swipe Gestures and Optimal Attribute Selection". In: *Proceedings of the 2nd Workshop on Immersive Analytics: Exploring Future Interaction and Visualization Technologies for Data Analytics*. IEEE VIS '17. Phoenix, AZ, USA

Contents

1 Introduction

In recent years, the effective use of data has become increasingly important. Due to processes realized through modern information systems, machines with integrated intelligent sensors, or the daily use of the smartphone, the amount of generated data is constantly increasing. In the beginning of the century data did not always affect the everyday life. Today data affects everyone directly, both intentionally or unintentionally.

Making sense of data is an essential element of today's business (Henke et al., 2016). Not using the data neglects opportunities to derive new business models and services ideally adapted to a fast changing world. However, not only companies and the public sector are interested in the use of data, but also in private life an awareness for the use of data has arisen. Under the term *quantified self*[1], people start using personal data, e.g., fitness data in order to improve their training. Furthermore, new laws such as the General Data Protection Regulation (GDPR) of the European Union[2] foster peoples' awareness of where data is collected and how it is used.

While objectives in data certainly differ between business and private life, the methods remain the same. A central method is the use of data analysis through information visualization, referred to as visual analysis. Since the human eye is evolutionarily trained for fast identification of visual structures, visualizations are a powerful tool to quickly generate data insights. To verify these insights, hypotheses are derived and subsequently statistically tested against the underlying data. Similar to hypothesis testing, creating effective visualizations requires of multiple steps (Card et al., 1999; Wilkinson, 2005) including data cleaning and filtering, transforming, mapping data attributes onto visual variables, and rendering. Each step involves certain decisions, knowledge, and experience. It can be very complex and challenging for a user to properly analyze data through applying the individual steps of the process and their corresponding mathematical methods.

Additionally, users have different characteristics. Capabilities, knowledge, and preferences vary from one user to the other. Therefore, a user should be viewed from a broader perspective, in order to effectively serve them in the visual analysis process. Consequently, the behavior, the preferences, and the current knowledge of the user have to be taken into account. Tukey and Wilk (1966) appealingly stated:

[1] https://quantifiedself.com
[2] https://gdpr-info.eu

> "The science and art of data analysis concerns the process of learning from quantitative records of experience. By its very nature it exists in relation to people. Thus, the techniques and the technology of data analysis must be harnessed to suit human requirements and talents. Some implications for effective data analysis are: (1) that it is essential to have convenience of interaction of people and intermediate results and (2) that at all stages of data analysis the nature and detail of output, both actual and potential, need to be matched to the capabilities of the people who use it and want it." (Tukey and Wilk, 1966, p. 697f)

Tukey and Wilk (1966) derive two primary challenges for new designs and technologies in the data analysis process: convenience of interaction and adaption to the user.

The first challenge of "convenience of interaction" (Tukey and Wilk, 1966) is addressed through the use of multiple modalities for visual analysis, among other things. Multimodal visual analysis is an emerging topic. Approaches leverage different modalities for different tasks in visual analysis. For instance, speech is very effective for generating visualizations (Grammel et al., 2010) while touch can easily be used for direct manipulation at a visualization. The combination of speech and touch is especially raising hidden synergies (Cohen et al., 1989). Consequently interactions become more convenient for a user through these different modalities because they lower interaction barriers.

Furthermore, some multimodal visual analysis approaches include recommender systems to automatically generate visualizations. Like other recommender systems such as Google Search[3] for news or Netflix[4] for movies and TV shows, visualization recommender methods aim for accelerating the process of discovering desired information. These approaches certainly reduce the complexity of visual analysis, since they protect the users from the complexity of creating effective visualizations.

In general, visual analysis contains an inherent amount of complexity just like every other process in accordance with Tesler's law of conservation of complexity (Saffer, 2006). The resulting challenge is to find the right level of revealed complexity for a user (Norman, 2010). An effective technology has to both cover complexity in situations in which the user's knowledge is exceeded and reveal complexity when the user's knowledge fits. Otherwise, the user experience likely suffers as the user is either overstrained or bored (Norman, 2010). Hence, multimodal visual analysis has to be personalized.

Although it is clearly highlighted that "[...] at all stages of data analysis the nature and detail of output, both actual and potential, need to be matched to the capabilities of the people who use it and want it" (Tukey and Wilk, 1966), little knowledge exists in how to achieve either a

[3] https://www.google.de
[4] https://www.netflix.com

user-specific multimodal visual analysis or a personalized visualization recommender. With this aim, this thesis contributes towards a deeper understanding of the challenges, advantages, and disadvantages of personalizing multimodal visual analysis for structured data through the use of natural language and gestures.

1.1 Contribution

This thesis takes a holistic view on the personalization of multimodal visual analysis on structured data to achieve an intelligible user-specific analysis. The following chapters propose and investigate not only personalization methods for recommendations and conversations, but also consider the integration of these methods into an overall design and interaction concept from a usability perspective. The thesis empirically investigates the novel combination of personalization, recommendations, natural language processing, visualization, and human-computer interaction for multimodal visual analysis of structured data. The derived insights and findings collected through various conducted empirical studies help to better understand the effect of personalization on multimodal visual analysis. This thesis further shows how knowledge from other domains such as linguistics effectively supports personalization in order to eventually achieve an intelligible user-specific visual analysis.

1.2 Thesis Outline

Chapter 2 introduces the relevant terms, and definitions. It first introduces fundamental work concerning the effectiveness of visualizations in terms of appearance as well as in relation to specific analysis tasks. Additionally, it describes work on recommender systems for visual analysis. These recommenders generate visualizations by considering knowledge gained from effectiveness studies. Moreover the knowledge gained from the work on recommender systems directly influences work on using speech for visual analysis. Lastly, applications of these tools in practice are highlighted.

Chapter 3 investigates the potential for personalization of abstract visualization tasks. A review on the understanding of tasks in visual analysis addresses a theoretical task space. In order to support the design of a personalized intelligible multimodal visual analysis, this task space is further ranked by both the required knowledge for properly completing a task and the relevance of the user's preferences. The chapter addresses the research question (RQ):

RQ 1 How can tasks in visual analysis be systematically structured based on their potentials for personalization?

Chapter 4 investigates the behavior in the field by surveying both the use of visualizations in scientific publications and the underlying structures of text-based commands against a multimodal visual analysis system. The study results address a narrow visualization space with a preference-based ranking of the visual variables. Additionally, task-depend wordings are identified. Accordingly, Chapter 4 focuses on the research questions:

RQ 2 How do text-based commands look like when people generate and transform visualizations with a prospective visual analysis system?

RQ 3 Which visualizations are used in scientific publications for highlighting insights from structured data?

Based on these insights into the user behavior in the field, Chapter 5 derives a user interface design and interaction concept focusing on intelligible visual analysis. The corresponding technical prototype is named Valletto. As the tool needs to understand the user's commands, a lightweight methodology for natural language processing in visual analysis is proposed. Two conducted user studies further show the effectiveness of a multimodal user interface compared to a classical user interface. This chapter addresses the following research questions:

RQ 4 What are the differences in completing tasks in a conversational interface compared to a conventional user interface?

RQ 5 What are the differences in the interaction strategies between a conversational interface and a conventional user interface?

Valletto communicates through the visualizations as well as the dialogue. Since the dialogue provides important information regarding the user's current analysis objectives, the content of statistical tests should be communicated in a user specific manner. Chapter 6 investigates the personalization of the dialogue component. Based on the linguistic theory of Grice (1975) concerning how human-human dialogues are fundamentally structured, a two-dimensional *answer space* is constructed accordingly. The conducted user studies highlight both diverse preferences in visual analysis and the effect of (mis)matching the user's language in visual analysis. This chapter concisely addresses the research questions:

RQ 6 What are the influencing factors for matching the users language in the answer space?

RQ 7 Can the user's preferred communication style be accurately predicted by a probabilistic model?

RQ 8 What is the effect of an answer space in a multimodal visual analysis system during a realistic situation?

RQ 9 Is the granularity of the answer space adequate?

While Chapter 4 provide a better understanding of the used visualization space, the ranking of this visualizations space should ideally consider the user's preferences. In order to formalize these preferences, Chapter 7 proposes a reinforcement learning approach to interactively learn the user's preferences through a sequence of pairwise comparisons. In order to reduce the individual learning effort, a divide-and-conquer approach situation is merged with the dueling bandit (Wu and Liu, 2016). The empirical studies show a positive effect of the dueling bandit's predictions as well the acceptance of the learning procedure. Furthermore, Chapter 7 explores the modelling of prior knowledge for the dueling bandit in order to further reduce the learning effort for the user. The following research questions are addressed:

RQ 10 Can a divide-and-conquer-based dueling bandit approach effectively learn individual visualization preferences?

RQ 11 What are the participants' reactions and feedback concerning the interactive learning procedure?

RQ 12 Can prior knowledge for the dueling bandit be modeled by a machine learning model?

RQ 13 How does prior knowledge affect the performance of the dueling bandit?

Finally, Chapter 8 summarizes the findings, existing limitations, and avenues for future work on personalized multimodal visual analysis.

2 Background

In order to understand the concepts and methods of this thesis, this chapter introduces fundamental definitions, and terminology. First, the concept of a visualization and its corresponding generating process are introduced. In the context of visualizations, related work show how the effectiveness of visualizations is directly influenced by the data, the task, and the user. Second, the process-related use of visualization is shown. Third, related work in the area of recommender systems and multimodal approaches for visual analysis are structured and discussed. Finally, elements of personalization and the application of visualzations in practice are shown. Based on this analysis, current research gaps are addressed to further highlight the contributions of this thesis.

2.1 Information Visualization

Before introducing relevant terms and concepts, the term visualization needs to be defined. According to Card (2002), "Visualization can be described as the mapping of data to visual form that supports human interaction in a workplace for visual sense making".

There are many forms of visualizations. Literature generally distinguishes between scientific visualization and information visualization (Levkowitz and Oliveira, 2003). In scientific visualization, visualizations represent graphical element of real world objects, e.g., realistic simulations of floods (Cornel et al., 2019). In information visualization, visualizations express abstract data, e.g., stocks (Ko et al., 2016). This thesis only considers information visualization. A widely used definition of information visualization is the following:

Definition 1: Information Visualization (Card et al., 1999)

"The use of computer-supported, interactive, visual representations of abstract data to amplify cognition".

According to this definition, a visualization is always interactive, and computer-supported. Furthermore, it only represents abstract data and needs to actually amplify cognition. In this context, amplifying cognition refers to taking advantage of the human eye's capabilities in fast discovering visual patterns. Additionally, Card et al. (1999) not only define what information visualization is, but also relevant steps from raw data to a visualization in the context of information visualization. They summarize these steps in the visualization pipeline, shown in Figure 2.1.

The input of the visualization pipeline is *raw data*. As the name suggests, raw data is not necessarily processed, clean, or filtered. It essentially describes extracted data from, e.g., sensors, or devices. After transforming the data into the shape *data tables*, it can be used to further create a visualization. This transformed data represents a structure as known from relational data bases, or simple excel files. A data table consists of rows and columns. The rows represent objects or items while the columns represent the attributes of these objects. For instance, a text document (*raw data*) might be transformed into a table by computing the term frequency (tf) for this document (see Figure 2.2). The corresponding data table contains two columns and one row for each unique word in the original document. The columns are: the word, and the tf score.

Regarding raw data, it is hard to define what the values actually represent and how they are structured. However, this example addresses relevance of the scale of measurements of the different attributes from the data table. While words are basically strings, the tf score

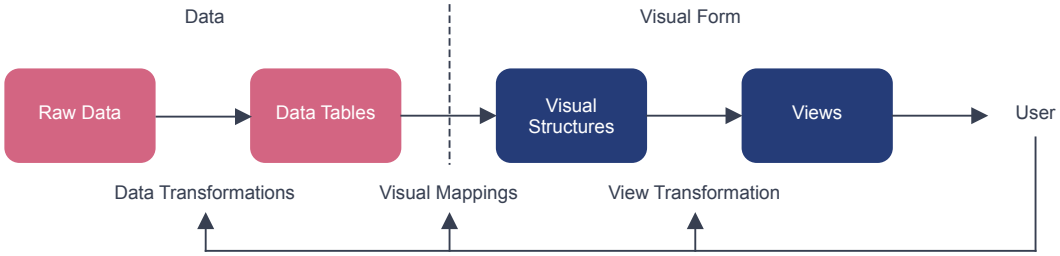


Fig. 2.1: Visualization Pipeline (Card et al., 1999; Jabbari et al., 2018)

is a numerical value. Typically, three different scales of measurements are considered in information visualization.

Definition 2: Scale of Measurements (Stevens, 1946)

Name	Description
Nominal	Determination of equality
Ordinal	Determination of greater or less
Interval	Determination of equality of intervals or differences
Ration	Determination of equality of ratios

However, in information visualization, the scales interval and ration are merged to quantitative (Card, 2002). Considering the visualization pipeline, data is still represented as a table until this step in the pipeline. In order to transform a data table into a visualization, the components of a visualizations need to be considered. Precisely, the *visual structures* of a visualization are taken into account. Bertin (1974) proposes to essentially consider two elements: marks and visual variables.

Marks are *points*, *lines*, *areas*, and *volumes* (Card, 2002). Marks are also named types. While marks describe the general representations of individual objects from table data, the visual variables describe how these individual elements look like. These visual variables are *position*, *size*, *shape*, *value*, *colour*, *orientation*, and *texture* (Bertin, 1974; Card, 2002). Bertin (1974) extracts the visual variables by analyzing the design of geographical maps. However, these visual variables can also be considered conceptually as the basis for information visualization (Carpendale, 2003). Consequently, they form the fundamental elements of each visualization. In order to progress a data table to a visualization, a mapping is further needed.

Definition 3: Visual mapping

A visual mapping assigns each data attribute $a \in A$ to a visual structure $v \in V$.

According to this definition, multiple visual mappings exist between a data table and the visual structures. Assuming a data table contains the data attributes A and the visual

Fig. 2.2: Example of the visualization pipeline implementation for generating a wordcloud out of a text document.

structures V are available, the number of potential visual mappings from A on V are (Mutlu et al., 2016):

$$\frac{|V|!}{(|V| - |A|)!} \tag{2.1}$$

However, a visual mapping should always be considered in the context of the purpose of the visualization. Therefore, the visual mapping should only contain data attributes relevant for achieving this purpose. Consequently, if a user wants to observe the frequency of the words in a document, merely visualizing the used words makes no sense. Instead, the visual mapping should also consider the frequency of the words. Visual mappings following this principle are called expressive.

Definition 4: Expressiveness of a Visual Mapping (Mackinlay, 1986)

"A set of facts is expressible in a language if it contains a sentence that (1) encodes all the facts in the set, (2) encodes only the facts in the set".

Although the number of expressive visual mappings is tremendously smaller than the number of all possible visual mappings, each expressive visual mapping uses different visual structures. For instance, the frequency of the words might be mapped either on the size or on the color. Both visual mappings are expressive. Yet, the human eye's capability to read the information varies between the visual mappings. Therefore, research investigates the effect of visual mappings on the effort required by the user to decode the desired information under the concept of effectiveness:

Definition 5: Effectiveness of a Visual Mapping (Mackinlay, 1986)

"Effectiveness criteria identify which of these graphical languages, in a given situation, is the most effective at exploiting the capabilities of the output medium and the human visual system".

Section 2.2 summarizes work on investigating the effectiveness of visual mappings in detail. However, the term visual mapping is not a synonym for a visualization. In order to turn a visual mapping to a visualization, rendering is needed. This rendering uses the objects from the data table and generates the visualization according to the visual mapping. Furthermore,

it defines the view port of the visualization. Since a visualization is always interactive, a user might want to zoom into the data. This action from the user only affects the visualization itself, but not the visual mapping. Hence, only the rendering of the visualization needs to be executed again.

Still, the key element of the entire visualization pipeline (Card et al., 1999) is the visual mapping. Hence, it pretty much determines the effectiveness and expressiveness of a visualization. Therefore, the visual mapping plays an important role in automatically generating visualization (Mackinlay, 1986). Section 2.4 will further highlight how this automatizing can be achieved.

2.2 Effectiveness of Visualizations

As illustrated, expressive visual mappings do not necessarily show similar performance at extracting desired information. In fact, the effectiveness of the visual mapping typically varies. Section 2.2.1 introduces the current state of the art on the effectiveness research focusing on the visual mapping itself. Additionally, Section 2.2.2 adds work exploring the effect of the data and the task on the effectiveness of a visual mapping.

2.2.1 Effects of the Appearance

Generally, the design of graphics is essentially framed by the Gestalt Principles (Todorovic, 2008). The Gestalt Principles summarize the basic fundamentals of visual perception. For example, the law of proximity describes the effect that points are considered more closely related when they are visually close. This law appears particularly relevant for scatter plots. Hence, the graphical perception plays an important role in determining the effectiveness of visual mappings (Cleveland and McGill, 1984). However, a visualization internalizes multiple perceptional aspects. This makes it hard to only consider the Gestalt Principles. Instead, previous research compared the different visualizations against each other in order to properly determine the effectiveness of particular visualization.

Eells (1926) conducts one of the first work on analyzing the effectiveness of pie charts compared to bar charts. He determines a better performance of the pie chart. Spence and Lewandowsky (1991) compare bar charts, pie charts, and tables. They show an advantage of both pie and bar charts against the table. However, they do not reveal a significant difference between the two chart designs. Skau and Kosara (2016) investigate the effect of varying arcs, angles, and areas on the design of pie and donut charts. A donuts chart is essentially a pie chart a whole in the middle of the pie. On the one hand, the authors show a similar performance of pie and donuts charts. The arc length, on the other hand, is identified as the

most important factor. Kosara (2019a) further analyzes different pie chart designs where the design of the filling area varies. While centered designs reduce the performance of the participants, other designs perform equivalently well.

While the previous work explored different fully specified visualizations, Mackinlay (1986) analyzes the effectiveness of the visual variables (Bertin, 1974; Carpendale, 2003). For each scale of measurement, Mackinlay (1986) ranks the visual variables according to their effectiveness. Based on these rankings, he proposes to automatize the generation of visual mappings for a given set of data attributes only by considering the scale of measurements. In fact, this approach is used in many recommender systems for visualizations (see Section 2.4).

Considering the different visual variables, Simkin and Hastie (1987) analyze judgment differences between between bar charts, stacked bar charts, and pie charts. The bar chart encodes the relative position, the stacked bar chart encodes the length, and the pie chart encodes the angle. The authors highlight the performance of the bar chart over the other options. Additionally, both bar chart variants perform better than the pie chart.

In addition to comparing aggregated values in simple bar charts, it is important to consider the distribution of a quantitative data attribute, too. Categories might differ with respect to a specific measure at first glance. However, considering the entire distribution might actually reveal no significant difference at all. In order to represent the distribution in a visualization, additional visual elements such as error bars have been introduced. A bar chart enhanced by an error bar might prevent invalid conclusions. However, users occasionally struggle in correctly reading visualizations which encode characteristics of a distribution. Correll and Gleicher (2014) compare four different visualization designs for illustrating uncertainty. Users perform better with violin charts or gradient charts. The widely used bar charts with error bars fall behind, according to the authors.

Since the effectiveness of a visualization is often investigated in two-dimensional settings, Dimara et al. (2018) explore the differences between parallel coordinates, scatter plot matrix, and tabular visualization in a multidimenisonal setting. In this setting, the tabular visualization appears to be slightly better for decision making than the other visualizations.

While previous work explores the performance of users by varying the visualization, Borkin et al. (2013) explore the memorability of visualizations. Comparing a variety of different widely used visualizations, the authors show a negative correlation between commonly used visualizations and their memorability by the user.

2.2.2 Effects of the Data and Tasks

Previous related work present a wide range of studies on the effectiveness of different visualizations. These studies mostly focus on the setting of nominal and quantitative data. Yet, the distribution of the data as well as the analysis task both influence the effectiveness of visualizations as well. For instance, the number of unique categories of an attribute affects the effectiveness of visual mappings. Therefore, Cleveland and McGill (1984) propose to the use of different visualizations for different tasks.

Kosara (2019b) explores whether the effectiveness changes when the distribution of the data changes. Comparing pie charts, tree maps, bar charts, and stacked bar charts, the author addresses differences in the response time and the error rate. For instance, the response time almost remains the same when varying the data distribution, given a bar chart. However, the tree map performs worse than the pie chart.

In addition to the task-related analysis of the effectiveness of visualizations, work has been conducted focusing on the effect of the distribution on the effectiveness. Harrison et al. (2014) investigate the effectiveness of various visualizations by varying the correlation between two quantitative data attributes. The authors particularly consider the effectiveness under Weber's Law. Weber's Law describes the relation between the perceptional effect and the actual effect. According to Harrison et al. (2014), the scatter plot overall outperforms the other visualizations. However, the parallel coordinate plot is the best performing visualization when the correlation coefficient is low. Typically, low correlation coefficients are harder to determine by the user.

Additionally, Kay and Heer (2016) investigate how the effectiveness of visualizations for correlation can be modelled. They propose a Bayesian approach for modelling the effectiveness. Furthermore, Kay and Heer (2016) form four groups (high precision, medium precision, low precision, and indistinguishable from chance) of visualizations with equal performance each. Still, the scatter plot outperforms all other approaches in both negatively and positively correlated data.

Considering the effectiveness of two-dimensional visualizations, Saket et al. (2018) explore ten common visualizations on varying analysis tasks proposed by Amar et al. (2005). According to Saket et al. (2018), line charts perform best for finding correlation, while tables and pie chart show low performance in this task. Furthermore, scatter plots perform well in identifying anomalies in the data. Overall, the authors highly recommend to use visualizations depending on the analysis task, since there is no visualization which outperforms others in each task.

Taking the work of Saket et al. (2018) further, Kim and Heer (2018) analyze the effect of distributions and tasks on the effectiveness. The authors conduct a study covering both value and summary-oriented tasks for three dimensional visualizations on two quantitative and one nominal data attribute. Depending on the task, the best-performing visualization changes. For instance, scatter plots perform well for comparing individual objects, but not when data should be summarized (Kim and Heer, 2018).

As the majority of the conducted studies provide qualitative knowledge about the varying effectiveness of visualizations, automatically estimating the effectiveness of a visualization is hard. Therefore, Veras and Collins (2019) introduce the Multi-Scale Structural Similarity Index (MS-SSIM) as an indicator for automatically determine a visualization effectiveness for a data set. By comparing the predicted results of MS-SSIM with the user study results of Kim and Heer (2018), the authors show that MS-SSIM approximates the actual effectiveness.

2.2.3 Effect of User Characteristics

The effectiveness of a visual mapping depends on the data distribution (Harrison et al., 2014; Kay and Heer, 2016; Kim and Heer, 2018; Kosara, 2019b; Veras and Collins, 2019), the analysis task (Kim and Heer, 2018; Saket et al., 2018), and the general perception of the resulting visualization (Dimara et al., 2018; Mackinlay, 1986; Simkin and Hastie, 1987). However, visualizations eventually serve as a reasoning tool for a user which are generally considered as being diverse. They differ in multiple dimensions, e.g., educational background, experiences, perceptional capabilities, and so forth. Hence, the question remains whether the user's characteristics also influence the effectiveness of a visual mapping. The following studies analyze this challenge from multiple perspectives.

Velez et al. (2005) inestigate whether user characteristics influence the interaction with visualizations. Precisely, the authors explore the effect of the spatial ability on the performance in visualization tasks. Considering the analysis of three-dimensional visualizations, Velez et al. (2005) show a correlation between spatial ability and accuracy. Participants with higher spatial ability perform better with the visualizations in terms of accuracy.

Focusing on information visualization, Conati and Maclaren (2008) explore whether user characteristics influence the effectiveness of visualizations. The authors do so by comparing two visualizations for complex system changes. While the majority of the user's characteristics have no significant influence on the performance in analysis tasks, the perceptual speed does. According to Conati and Maclaren (2008), the perceptual speed is the "Speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual perception". Toker et al. (2012) explore the differences between bar charts and radar graphs while considering the results of Conati and Maclaren

(2008). They also identified a significant influence of the perceptual speed on effectiveness. Furthermore, user characteristics also influence the preferences and the ease of use regarding visualizations.

Both Green and Fisher (2010) and Ziemkiewicz et al. (2011) explore how the personality of a user influences the effectiveness of visualizations. Ziemkiewicz et al. (2011) find correlation between the locus of control and the performance of visual mappings. Green and Fisher (2010) further show that locus of control, extraversion, and neuroticism have an effect on the completion time. Additionally, these personality traits also influence the number of collected insights from visualization. Conati et al. (2014) further explore the effect of personality traits on the performance with visualizations. The authors highlight not only different performance in visualization tasks, but also show an effect of the working memory of the participants on their performance.

Gajos and Chauncey (2017) find correlation between the need for cognition and a utilization rate. The authors also show that extraversion negatively correlated with the utilization rate. Focusing on cognition, Lee et al. (2019) identify a correlation between the user's ability to make sense of a visualization and both the user's cognitive abilities and the need for cognition, respectively. As personality traits influence effectiveness, Haroz and Whitney (2012) explore how the limits of attention further affect effectiveness. According to their study results, a user's capability of attention has an effect on idenfiying more insights in visualizations.

Overall, these studies provide empirical evidence that the user's characteristics, abilities, and personal traits influence the effectiveness of visualizations. Lee et al. (2017) propose the term *visualization literacy* as "the ability and skill to read and interpret visually represented data in and to extract information from data visualizations" in order to achieve a common definition.

2.3 Visual Analysis

As previous sections show, visualizations leverage visual structures to reveal patterns in data. Visual mapping is a crucial step in generating a visualization. It determines whether a visualization is effective and expressive. However, a variety of visual mappings exist. Depending on the user, the data, and the task, the effectiveness of visualization varies. Yet, a visualization supports a user in the data analysis process. Hence, it is important to understand this process. Generally, data analysis through visualizations is called visual analysis.

Definition 6: Visual Analysis (Kehrer and Hauser, 2013)

"Visual analysis is the integration of visualizations, interactions, and computational analysis".

Basically, this definition summarizes the general use of visualizations on data, but does not describe the process of finding insights in data. On finding insights in data, Tukey (1977) first mention the idea of an exploratory data analysis in order to formulate hypotheses in data.

Definition 7: Exploratory Data Analysis (Keim et al., 2006)

"Exploratory data analysis is the process of searching and analyzing databases to find implicit but potentially useful information".

Generally, the objective is to find insights through data transformation and hypothesis testing. Furthermore, visualizations already play an important role in this concept. However, the main interaction method remains the statistical tests and the data transformation operators. Readjusting the focus now on visualizations, the process consequently leads to exploratory visual analysis.

Definition 8: Exploratory Visual Analysis (Battle and Heer, 2019)

"Exploratory visual analysis is a subset of exploratory data analysis, where visualizations are the primary output medium and input interface for exploration".

According to this definition, exploratory visual analysis considers multiple approaches for finding insights. On the one hand, it supports the open exploration of a new data set. In this process, the user has no hypotheses on the data, but searches for interesting patterns. However, this process likely results in hypotheses. On the other hand, the definition also supports a confirmatory analysis whether a set of hypotheses is true or false.

A more advanced concept is visual analytics. Visual analytics adds methods from artificial intelligence such as clustering, and machine learning approaches to the analysis process.

Definition 9: Visual Analytics (Keim et al., 2010)

"Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets".

Hence, visual analytics considers the idea of using complex algorithms together with interactive visualizations. Following this definition, visual analytics exceeds the use of

statistical methods for generating insights. It aims to use machine learning models for generating additional insights. Through an interactive process, visual analytics leverages the advantage of the computational power of a computer and the visual capability of the human's eye.

Although elements of visual analytics play a role in this dissertation, the focus will remain on visual analysis. The visualizations will remain as the primary input and output for the user. Furthermore, the visualizations will be further supported by statistical measures conducted by the system. Hence, the approaches essentially follow the definition of exploratory visual analysis (Battle and Heer, 2019).

2.4 Visualization Recommender Systems

Creating visualizations is a challenging procedure. Since the effectiveness of a visualization is important to increase a user's performance in an analysis task (Harrison et al., 2014; Kay and Heer, 2016; Kim and Heer, 2018; Saket et al., 2018; Veras and Collins, 2019), a visual mapping should be carefully constructed. However, many users likely have little knowledge on the formal effectiveness of visualization. In order to cover the inherent complexity of the visualization pipeline, visualization recommender systems have been introduced.

Definition 10: Recommender System (Ricci et al., 2010)

"A recommender system is a software tool or a technique providing suggestions for items to be of use to a user".

The objective of a recommender system for visualization is two-fold. First, a visualization recommender automatically generates visualizations for the user (Mackinlay, 1986). This covers the complexity of deciding on an effective visual mapping for a given set of data attributes. Second, the analysis process of the user is considerably accelerated due to the automatic generation of visualizations. A user does not have to invest time into the creation a visualization, but can focus on the data itself.

The idea of a recommender system for visualizations plays an important role in this dissertation. Its functionality is essentially needed for covering complexity, but also to adjust to the user's needs. Therefore, the following sections introduce current state of the art work on visualization recommenders. Section 2.4.1 introduces approaches using rule-based methods for recommending visualizations. These methods mostly rely on the effectiveness studies introduced in Section 2.2. Additionally, Section 2.4.2 discusses approaches taking advantage of data-driven methods.

Table 2.1 presents an comprehensive overview about the related concepts discussed in the following sections. The concepts are classified following the taxonomy of Heer and Shneiderman (2012). However, the taxonomy is extended by categories especially relevant for this dissertation such as supported modalities, or whether a visualization is personalized.

2.4.1 Rule-Based Methodologies

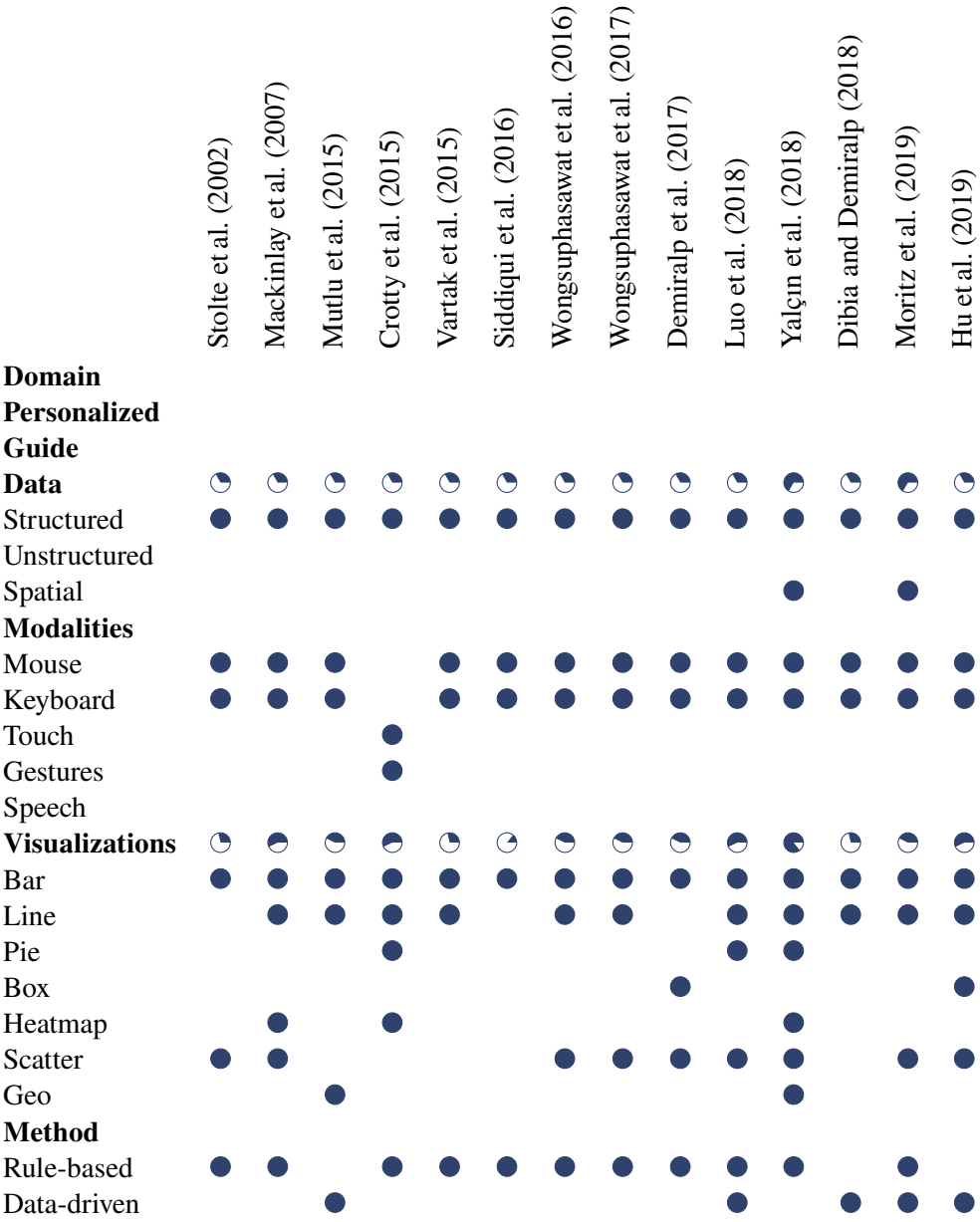
Rule-based approaches follow explicit routines for generating visualizations. These explicit routines and rules often implement knowledge gained from effectiveness studies. However, the approaches differ not only how they implement these rules, but also in which analysis process steps they support the user.

Mackinlay (1986) proposes to automatize the generation of visualizations. Based on his ranking of the visual variables for each scale of measurement, a visualization can be automatically generated just by considering the given data attributes to be visualized. As a result, quantitative data attributes are typically mapped to size, if available, while nominal data attributes are mapped to color. The idea of Mackinlay (1986) is further realized in *ShowMe* (Mackinlay et al., 2007). *ShowMe* is a widget in the software Tableau. A user only needs to drag and drop the data attributes to be visualized. *ShowMe* accordingly shows an overview about potential visualizations.

Stolte et al. (2002) present *Polaris* for exploring Pivot tables. The authors propose to add graphical representations to the structure of a pivot table. It is one of the first approaches to accelerate the exploratory visual analysis through the automatic generation of visualizations.

For effectively supporting a user in the exploratory data analysis, Wongsuphasawat et al. (2016) propose *Voyager*. *Voyager* recommends multiple visualizations at a time for a breadth-oriented exploration. However, the authors value data variations more than visualization variation. Additionally, *Voyager* computes potentially helpful data attribute combinations by either applying different transformations on the data attributes (e.g., computing the mean), or replacing certain data attributes. For each resulting data attribute combination, the authors follow the automatic generation idea of Mackinlay (1986) by considering the effectiveness of the visualizations (Cleveland and McGill, 1984).

Using *Voyager* (Wongsuphasawat et al., 2016) as a starting point, Wongsuphasawat et al. (2017) developed *Voyager2*. Unlike the predecessor, *Voyager2* allows manual specifications of the visualizations for more design adjustments by the user. While a user might manually design a visualizations for data attributes, the system further shows additional visualizations for both the effect of aggregations and the adding of another data attribute. According to



Tab. 2.1: Classified related work from field of recommender systems.

the authors, *Voyager2* allows a broader exploration due to the trade-off between manual specification of a visualization and automatically generating visualizations. Still, both approaches *Voyager* and *Voyager2* show a design gallery (Marks et al., 1997).

Vartak et al. (2015) approaches the challenge of finding interesting visualizations within an acceptable time frame. With their system *SeeDB*, the authors use deviation-based metrics to identify interesting visualizations for a given query. These metrics rank visualizations according to amount of deviation of the data that is represented in the visualization. Adding *SeeDB* on top of a relational database, Vartak et al. (2015) enable a fast exploration of a new data set.

The previous approaches consider as a main challenge the recommendation of a visualizations. However, Demiralp et al. (2017) propose to focus on visual insights instead. The authors argue for finding insights in the data. In order to decide on an insight, *Foresights* (Demiralp et al., 2017) computes statistical measures for the data, e.g., Pearson's correlation coefficient, outliers, and skew. For each discovered insight, a visualization (bar, scatter, or box plot) is shown.

Siddiqui et al. (2016) propose a query language for generating visualizations for data exploration. This language is called ZQL. It consists of a name as an identifier, four elements describing a visualization (X, Y, Z, vis), and an operator element for processing. The idea is to have an SQL-like query language to focus on the data while getting consistent visualizations for the queried data. *Zenvisage* (Siddiqui et al., 2016) implements ZQL. Siddiqui et al. (2016) show that users can explore data faster with *Zenvisage* as well as tend to find more insights in an unknown data set.

In contrast to other approaches, *Vizdom*, by Crotty et al. (2015), explores the interactive construction of machine learning models. A user can select desired data attributes as well as operations that represent machine learning models. For both the output of the machine learning model and the data attributes, *Vizdom* generates visualizations. Hence, the user can interactively explore how the different machine learning approaches work on the data. Using touch and pen, Crotty et al. (2015) further consider new modalities for visualization recommendations.

Yalçın et al. (2018) propose *Keshif*. A user is able to create a dashboard including multiple visualizations via drag and drop. Depending on the selected data attributes, the tool automatically generates visualizations. The scale of measurement of the data attributes essentially defines the proposed visualization type. Furthermore, all visualizations are linked to each other.

A variety of approaches tackle the challenge of accelerating exploratory visual analysis. It appears especially challenging to support a user in finding interesting insights. However, the majority of approaches have three things in common. First, they consider primarily relational data bases or table data. Second, they use insights from effectiveness studies for generating the visualization. Third, they rely on keyboard and mouse as interaction modalities for the user, except *Vizdom* (Crotty et al., 2015).

2.4.2 Machine Learning-Based Methodologies

In addition to the rule-based approaches, others – more recently – consider a data-driven approach. In these data-driven approaches, the visualizations are recommendated via machine learning models. Depending how the recommender problem is formulated, these approaches use machine learning for either ranking (Moritz et al., 2019), direct generation of visual mappings (Dibia and Demiralp, 2018), or direct generation of fully-specified visualizations (Hu et al., 2019; Luo et al., 2018; Mutlu et al., 2015).

Moritz et al. (2019), on the one hand, propose a combination of rule-based methods and machine learning. Based on knowledge gained from effectiveness studies, Moritz et al. (2019) build a set of constraints on the visual mappings for a given set of data attributes. These constraints technically formalize the knowledge and insights of those studies. In order to rank the visualizations, Moritz et al. (2019) use a RankSVM. Generally, a RankSVM learns a relative ranking of a set of items based on pairwise comparisons (Joachims, 2002). Formally, it follows the learning-to-rank paradigm (Liu, 2010). In *Draco* (Moritz et al., 2019), the RankSVM is trained on data of the effectiveness study by Kim and Heer (2018). Hence, the visualizations are ranked in accordance with the decisions of the participants of the study.

VizML by Hu et al. (2019), on the other hand, implements an end-to-end machine learning approach. It uses a deep neural network for recommending visualizations. This neural network is trained on visualizations from the Plotly library[1] as well as the corresponding data sets. Plotly is Python library for data visualizations. Hence, given a data set, the method recommends a Plotly-based visualization.

Another end-to-end machine learning approach is proposed by Dibia and Demiralp (2018). Their approach, called *Data2Vis*, translates JSON-structured data sets into Vega-lite (Satyanarayan et al., 2017) specifications by using an encoder-decoder approach. An encoder-decoder approach is originally introduced in the Natural Language Processing (NLP) domain for solving machine translation. Indeed, Dibia and Demiralp (2018) consider the recommendation of visualization as a machine translation problem. The corresponding languages are

[1] `https://api.plot.ly/v2`

the data set and the Vega-lite specification. Using the architecture proposed by Britz et al. (2017), the method is evaluated on the data of Poco and Heer (2017).

An in-between approach is *DeepEye* by Luo et al. (2018). In order to recommend visualizations, Luo et al. (2018) approach three challenges: separating "good" from "bad" visualizations, recommending visualizations, and selecting visualizations. In order to approach the first challenge, the authors train a decision tree for binary classification. For the second challenge, a LambdaMART (Burges et al., 2008) is implemented which is similar to the RankSVM (Joachims, 2002). Both methods are trained on a crowd-sourced data set. However, the final recommendation routine sequentially concatenates these different methods. Hence, the routine first identifies the "good" visualizations from a set of generated visualizations while sequentially ranking these visualizations afterwards.

VizRec by Mutlu et al. (2015) applies a combination of content-based filtering and collaborative filtering on rule-based generated visualizations. Collaborative filtering basically leverages the rating of other users to predict a rating for item for a new user while considering the similarity between the users (Sarwar et al., 2001). In this case, an item is a visualization. Mutlu et al. (2015) gather the required ratings through a crowd-sourced method. Initially, *VizRec* computes a set of potential visualizations through a rule-based approach. However, the ranking of these visualizations is a result of their collaborative filtering approach.

All of the discussed approaches from related work approach the challenge of recommending visualizations from a different perspective. Depending on how the problem is essentially modelled, the methods differ greatly. Starting from taking advantage of a machine translation model (Dibia and Demiralp, 2018), other approaches follow a sequence of various machine learning methods (Luo et al., 2018). Nevertheless, the visualizations are mainly generated through rule-based approaches, but ranked through machine learning models afterwards (Luo et al., 2018; Moritz et al., 2019; Mutlu et al., 2015).

However, the machine learning approaches essentially differ in one aspects from the rule-based approaches, apart from how the visualizations are recommended. The related work on machine learning concentrates on the recommendation method itself. These approaches investigate effective methods for data-driven recommendations. However, they do not consider the use of the recommendation within the visual analysis process. Yet, the user interactions could serve as a feedback channel. They could directly provide information whether a recommendation is good or bad. Considering this information, a machine learning model could improve online. Consequently, the ranking would not only become better, but also be base more on a user. Hence, it is important to take the user into account.

2.5 Natural Language in Visual Analysis

While recommender systems help to substantially accelerate visual analysis by automatizing the generation of visual mappings, they typically rely on classical user interfaces. These approaches consider input via mouse and keyboard (cf. Table 2.1). However, the use of visual analysis is changing due to new devices and corresponding interaction techniques (Roberts et al., 2014). Other modalities such as gestures, and speech are becoming more popular (Srinivasan and Stasko, 2017). Grammel et al. (2010) empirically reveal lower barriers for novice users in visual analysis using speech. Indeed, these users can effortlessly formulate their wishes in visual analysis, even though they lack the knowledge in how to effectively create appropriate visualizations.

In order to understand the approaches on multimodality in visual analysis better, the following sections introduce the current state of the art in the context of visual analysis. First, Section 2.5.1 discusses proposed approaches for visual analysis solely relying on speech and text, respectively, as a modality. Afterwards, Section 2.5.2 shows how natural language based interfaces can be enhanced by other modalities in order to further support the user.

2.5.1 Natural Language Interfaces in Visual Analysis

Sun et al. (2013) propose a tool called *Articulate*. With *Articulate*, users can simply ask questions about the underlying data to retrieve corresponding visualizations. Using a dependency diagram, the system infers the user's wishes and generates a visualization accordingly. Their results reveal that participants welcome the reduced effort for creating visualizations.

While querying relational data bases is a challenging procedure (Li and Jagadish, 2014), Gulwani and Marron (2014) explore how spreadsheets can be queried. Considering Excel files as a data source, Gulwani and Marron (2014) develop a widget for Excel to query tables. This widget, named *NLyze*, enables a user not only to give precise commands, e.g.,"sum X.", but also allows to formulate questions on the table, e.g., "what are the X". The authors combine keyword programming and semantic parsing. However, *NLyze* focuses on quantitative results, but not on visualizations for these quantitative results. Although it does not consider visual analysis as its domain, it certainly represents related work due to the translation of natural language based commands into queries for data tables.

Based on the idea of querying relational data bases, Dhamdhere et al. (2017) explore the idea of conversations in data exploration. *Analyza* (Dhamdhere et al., 2017) establishes a sequence of multiple steps of transforming a natural language command into an answer. The authors leverage classical NLP techniques. Semantic parsing plays a central part in

creating an SQL query, such as in Gulwani and Marron (2014). Furthermore, a user is able to manually correct a query in case the system detects ambiguous commands. *Analyza* does not primarily focus on the visualizations, although the system provides a set of visualizations.

While the previous approaches focus on the visual analysis part, others use natural language based interactions for analytic tasks. Both John et al. (2017) and Fast et al. (2018) consider the scenario of machine learning model building and statistical testing. John et al. (2017) introduce *Ava* which is a natural language module for Jupyter notebook (Kluyver et al., 2016). Instead of writing code, a user engages in dialogue with the system. John et al. (2017) take advantage of the Juypter notebook user interface. Fast et al. (2018) design a new dialogue-based interface for this scenario instead. *Iris* (Fast et al., 2018) provides the user with hints on potential methods for the data. Furthermore, it uses clarification requests to resolve potential mistakes and ambiguity in the user's commands. Yet, both approaches consider experts as the primary users rather than a broad variety of users with diverse backgrounds and characteristics.

Yu and Silva (2020) introduce *FlowSense* to enrich a data flow visualization systems. In a data flow system, a user can connect various elements with each other to implement data processing. KNIME (Berthold et al., 2007) is a well-known data flow system. Yu and Silva (2020) propose to use natural language commands to create the data flows. Instead of connecting elements manually, the user only needs to formulate commands. According to the authors, the system helps users to speed up their analysis as well as supports their learning curve with the data flow visualization systems.

In addition to information visualization, infographics are creative static visualizations enriched by illustrative elements and often referring to real world objects. Harrison et al. (2015) state "Infographics are an effective means for telling stories about data, as they capture a readers attention by structuring these stories using principles of graphic design". Cui et al. (2019) propose *Text-to-Viz*. *Text-to-Viz* generates infographics based on natural language commands. Considering a predefined design space for infographics, *Text-to-Viz* can effectively generate visualizations for proportion-related statistics, e.g, "40 percent of USA freshwater is for agriculture" (Cui et al., 2019). Cui et al. (2019) show benefits for opening up the field of infographics to a broader spectrum of prospective users due to the lower effort for generating visualizations by natural language.

2.5.2 Multimodal Interactions in Visual Analysis

The use of natural language for visual analysis shows significant improvements in making sense of data (Gao et al., 2015; Setlur et al., 2016). However, natural language as a modalilty has shortcomes. Natural language commands can be ambiguous since words

	Cox et al. (2001)	Gao et al. (2015)	Setlur et al. (2016)	Srinivasan and Stasko (2018)	Hoque et al. (2018)	Srinivasan et al. (2020)
Domain		●		●	●	
Personalized						
Guide						
Data	◔	◔	◕	◔	◕	◔
Structured	●	●	●		●	●
Unstructured						
Spatial			●	●	●	
Modalities						
Mouse	●	●	●		●	
Keyboard	●	●	●		●	
Touch				●		●
Speech	●	●	●	●	●	●
Pen						●
Visualizations	◔	◔	◔	◔	◔	
Bar	●	●	●		●	●
Line			●		●	●
Pie						
Box						
Heatmap						
Scatter		●	●		●	●
Geo			●		●	
Table	●					
Other				●		●
Generates Visualizations		●	●		●	●
Method						
Rule-based		●	●		●	●
Data-driven						
NLP						
Grammar	●	●	●	●	●	●
Ambiguity		●	●	●	●	

Tab. 2.2: Classified related work from the field of multimodal visual analysis systems.

are always context-sensitive (Gao et al., 2015; Hoque et al., 2018; Setlur et al., 2016). In order to overcome these shortcomings, combinations of multiple modalities – referred to as multimodality – are explored. Table 2.2 summarizes the discussed concepts accordingly.

Bolt (1980) explores one of the first approaches on multimodality for graphical interfaces. He use a combination of speech and air gestures to move items on a map. Users tend to use vague command structures. Furthermore, they combine both modalities to interact with the map. Oviatt (1997) also explores interactions with maps. Oviatt (1997) compares speech-only, pen-only, and multimodal interactions. In the conducted user study, users tend to prefer multimodal interactions rather than unimodal interactions. Furthermore, users seem to have difficulties with spatial commands. Furthermore, Hauptmann (1989) highlights the use of relatively simple gestures and speech commands for manipulating graphic images. Conducting a Wizard of Oz experiment, Hauptmann (1989) identifies a strong preference for using both speech and gestures by the users.

Cohen (1992) investigates the role of natural language in a multimodal interface. He argues for the use of natural language commands for describing objects as well as temporal relations. Additionally, he also admits the benefits of direct manipulation when the objective is hard to describe. While both modalities have their advantages, both also lack performance in certain tasks. Cohen (1992) recommends to leverage each modality's benefits when designing the interactions in multimodal interfaces. Badam et al. (2017) later discuss the effectiveness of modalities in tasks in visual analysis.

Considering the scenario of exploring domain-specific data through bar charts, Cox et al. (2001) first propose interactive reasoning through multimodal interactions. Multimodality is expressed by direct manipulations and natural language queries. While bar charts can be directly manipulated via classical interactions, the user can additionally query the data. Including initial knowledge on transforming natural language commands into data base queries (Hendrix et al., 1978), the authors show advantages of combining direct manipulations with natural language commands in visual analysis. According to the authors' observations, users effectively use the interface after short time, although they are not trained in the system in the first place. These observations further address the advantages of multimodal interfaces for visual analysis.

While natural language is very effective in formulating requests on data, interpreting these commands is subject to ambiguity. The meaning of a word mainly depends on the context, e.g., a "bank" can be a financial institute but also a protection for flooding. Additionally, words can also be phonetically identical, e.g., "profit" and "prophet". Hence, a machine interprets a user's commands under uncertainty. In order to cover this ambiguity-caused uncertainty, Gao et al. (2015) propose a multimodal interface named *DataTone* for visual analysis where a user can manually correct the interpretation of the system. If a word is

potentially misinterpreted, the user can choose a potential replacement from a top down menu.

Setlur et al. (2016) explores the use of multimodality in the context of reasoning of spatial data. Considering the exploration of maps, the authors propose commands for analytical tasks. A user can effectively query the system (*Eviza*) for gaining insights on data. The system uses meta information on the data such as geographic information to make sense of a user's commands. Setlur et al. (2016) also provide a method for covering ambiguity in the commands. Furthermore, the authors also implement direct manipulations on the visualizations.

Evizeon by Hoque et al. (2018) continues the work of Setlur et al. (2016) by implementing pragmatics. Pragmatics means considering also the context of an utterance when interpreting its meaning (Kabbara, 2019). For instance, a person formulates a request to create a bar chart, but requests in the very next utterance "make it blue" then "it" refers to the bar chart. Hoque et al. (2018) show that pragmatics helps to improve the reasoning in visual analysis, especially in the domain of geospatial data.

The previous work considers only a few visualizations simultaneously. *Articulate2* (Aurisano et al., 2016; Kumar et al., 2017) takes a different approach. It generates each visualization in a separated window. By doing so, it produces a multi-window interface. As other approaches, it also leverages natural language and gestures for interacting with the visualizations. A user can generate a visualization by commands. Since each visualization remains displayed during the analysis, a user can constantly reflect the conducted analysis at any time.

While the majority of the introduced work considers the analysis of tabular data, Srinivasan and Stasko (2018) investigates the use of multimodality for analyzing network data. Their system *Orko* implements multimodality through gestures and speech on wide screen. Additionally, the modalities can be used in combination as well as individually. For instance, a user can ask for a specific entity in the network which is eventually highlighted. However, the user can also select an entity via touch and ask for additional information on this entity via speech.

Lastly, Srinivasan et al. (2020) propose *InChorus*. *InChorus* implements multimodal interactions on a tablet device. It supports the modalities speech, touch, and a pen. Srinivasan et al. (2020) show a preferences for using touch for sorting data. However, their participants adapt their interactions to using multimodal input. Additionally, Srinivasan et al. (2020) argue for restricting the NLP routine on keyword-based commands in order to keep interactions simple. Their results support this design principle, as the error rate decreases.

2.6 Personalization in Visual Analysis

Section 2.2.3 highlights an effect of user characteristics on the effectiveness of visualizations. Furthermore, Section 2.4 and Section 2.5 show that recommender systems and multimodal approaches, respectively, in visual analysis accelerate the visual analysis process by automatically generating visualizations. By doing so, they effectively shrink the gulf of execution (Norman, 2002). The gulf of execution essentially describes the number of steps required by the user to achieve a desired objective. Taking effectiveness studies as primary input, recommender systems do not fully consider the studies on the user's characteristics on the effectiveness. Hence, personalization could help to further improve the usability of these systems (Huang et al., 2015; Lallé and Conati, 2019; Oscar et al., 2017).

Generally, personalization works through user modeling. User modeling describes the process of creating a formal representation (model) of a user (Fischer, 2001). A user model can be represented in multiple ways by, e.g., only considering the information about the user (flat structure), or modelling the relationships between the different variables of the user (hierarchical) (Fischer, 2001). Generally, user models describe the relationship between a user-related variables, e.g., education, knowledge, etc. and one or more target variables, .e.g, ratings of movies. Given a user model, a system can leverage this model to adapt its reactions to the user (Hurst et al., 2007). Such systems are called adaptive.

Definition 11: Adaptive (Fischer, 2001)

"Dynamic adaptation by the system itself to current task and current user".

In contrast, a system which allows a user to change its functionality is an adaptable system (Fischer, 2001). Research investigates differences between static, adaptive, and adaptable systems in intelligent user interfaces on various levels. For instance, Findlater and McGrenere (2004) show significant differences in completion time between these different approaches in the context of menus. Overall, users seem to be faster with a static menu design (Findlater and McGrenere, 2004).

In the context of visualizations, Toker et al. (2012) and Conati et al. (2015) argue for adaptive systems. The authors also refer to the differences in user characteristics. An adaptive system could create visualizations aligned with the user's characteristics. Furthermore, Ahn and Brusilovsky (2013) show advantages of interactive visualizations linked to a personalized search for information retrieval. Personalized search helps users to gain more insights.

Yet, adding additional models to the system increases uncertainty about the given outputs. A system without a user model produces the same output no matter who is using the system. A system with a user model produces a user-specific output. However, such models are

never flawless. They are likely to make mistakes in predicting unsatisfactory outputs to the user. It is important to keep the gulf of evaluation small. The gulf of evaluation describes the effort of a user to make sense of the system states, given the user's input (Norman, 2002). Consequently, it is important that the user can make sense of the system's behavior, especially in situation in which the system fails. In order to make a system explainable, it requires several elements.

First, a system needs to be transparent with respect to its actions. Transparency is a prerequisite for understanding a system's reactions. However, even experts struggle in interpreting complex constructs, e.g., the behavior of deep neural networks (Koh and Liang, 2017). The question remains how transparency can contribute to make complex things interpretable, since transparency alone does not necessarily help to reduce the gulf of evaluation.

Definition 12: Interpretability (Gilpin et al., 2018)

"The science of comprehending what a model did or might have done".

An example from NLP is highlighting words in a sentence depending on their effect on the output, e.g., words with a sentiment. Although this highlighting is eventually based on numbers from different layers of the neural network, it is still possible for a user to interpret why the system made a certain decision. Highlighting is an simple yet effective example for interpretability. Still, models are becoming more and more complex. Humans increasingly struggle in understanding these trained structures. Hence, a model should further be explainable.

Definition 13: Explainability (Gilpin et al., 2018)

"Models that are able to summarize the reasons for [...] behavior, gain the trust of users, or produce insights about the causes of their decisions".

According to Gilpin et al. (2018), *interpretability* and *explainability* are not synonyms. In fact, *explainability* implies *interpretability*, but not necessarily the other way around.

Although the system might be transparent, interpretable, and explainable for an expert, it does not necessarily imply that every user can immediately understand the conducted steps. Depending on the user, it likely varies whether the chosen communication (e.g., coloring relevant sentence structures) is intelligible for them. Weld and Bansal (2019) state: "The key challenge for designing intelligible AI is communicating a complex computational process to a human". According to Weld and Bansal (2019), if a user understands a system's reactions, it can increase the number of generated insights (Caruana et al., 2015), the trust in the system

(Sinha and Swearingen, 2002), and many more. Hence, being intelligible is important for a system supporting a diverse audience.

In this dissertation, intelligibility plays a central role. While the objective is to open up visual analysis to a broader spectrum of users, a system needs to adjust its communication to each user individually. This complex information addresses both the system's responses in the immediate analysis situation and the system's computations for generating the corresponding reactions.

2.7 Applications

Visualizations are used in various fields. Through interactive visual analysis tools, visualizations help people making better and faster decisions. Additionally, they help to maintain an overview on huge and volatile data sets. While a variety of applications exists, the following examples focus on aspects, particularly from the automotive value chain (Sturgeon et al., 2008).

In manufacturing, visual analysis tools are applied in different steps (Xu et al., 2017). Sun et al. (2019) consider visual analytics in the context of production planning in smart factories. Production planning is challenging as multiple complex units have to be integrated. This leads to a complex optimization problem where visual analytics can actively help a user in establishing the optimal production planning. Considering one part of the production, Sydow et al. (2015) show the benefits of visual analysis for scheduling containers in a supply chain. They further argue for mobile visualizations as people in manufacturing have to occasionally explore relevant spots in the factory.

In finance, it is crucial to keep track of the current cash flows and portfolios. Hence, visualizations in finance focus on volume and changes (Ko et al., 2016). Therefore, line charts are very important as they show changes over time. Merged with a bar chart on the trading volume, the visualization provides a comprehensive overview on a stock. Furthermore, finance is one of the areas where modern visualizations are heavily used.

In sales, visualizations are typically used to communicate sales numbers. Considering the sales numbers of the automotive companies, the primary visualizations are bar charts and tables. They typically sort the data in accordance with the number of sales per car model and country, respectively. While these data represent a post perspective on the sales, analysts use often dashboards for making sense of their current sales.

While all these approaches consider a specific use case, they have certain elements in common. First, many of these approaches consider classical visualizations instead of designing new

visualizations. Using familiar visualizations helps decreasing the learning effort for the user (Borkin et al., 2013). Second, the data follow similar structures. Although the data domains are very different, the data is often represented in tables, or relational databases. Hence, they are already structured. However, users in the different use cases are likely diverse in terms of educational background, preferences, and experience. Hence, personalization on visual analysis tools for automatic generation of visualizations sounds like a promising approach to reach a broader spectrum of users.

2.8 Summary

This chapter structures the domain of visual analysis by considering both the creation of visualizations and the use of visualizations. Furthermore, it discusses relevant concepts and related work in the domain of visualization recommendations (cf. Section 2.4) using effectiveness studies (cf. Section 2.2), and multimodal approaches (cf. Section 2.5.2). Additionally, the potentials of personalization in visual analysis are highlighted (cf. Section 2.2.3 and Section 2.6) along with providing use cases for visual analysis in practice (cf. Section 2.7). It identifies elements that should be investigated both in this dissertation and in research in general.

First, knowledge on personalized recommendations of visual mappings both expressive and effective is scarce. Deciding for a visual mapping essentially is the most crucial step in visualizing data (Mackinlay, 1986). In order to help a user in this step, many related works propose rule-based recommendations of visual mappings (Crotty et al., 2015; Demiralp et al., 2017; Mackinlay, 1986; Mackinlay et al., 2007; Siddiqui et al., 2016; Stolte et al., 2002; Vartak et al., 2015; Wongsuphasawat et al., 2016; Wongsuphasawat et al., 2017). Based on knowledge gained from effectiveness studies, a visual mapping certainly depends on data and task (Harrison et al., 2014; Kay and Heer, 2016; Kim and Heer, 2018; Saket et al., 2018). Additionally, machine learning approaches try to learn "good" visualizations (Dibia and Demiralp, 2018; Hu et al., 2019; Luo et al., 2018; Moritz et al., 2019; Mutlu et al., 2015). However, these approaches do not embed the recommendations in the visual analysis process. Instead, they consider only the method itself. Still, the insights are valuable for understanding how to approach data-driven approaches. Yet, research lacks evidence on these approaches performance in practice, since data-driven approaches typically require data from the user.

Second, multimodal approaches sound promising for empowering a broader spectrum of users for conducting their own data analysis especially in the combination of speech and gestures (Aurisano et al., 2016; Cox et al., 2001; Cohen, 1992; Gao et al., 2015; Hoque et al., 2018; Kumar et al., 2017; Setlur et al., 2016; Srinivasan and Stasko, 2018). However, the idea of multimodality is not yet fully explored. For instance, a user can speak with the system, but the system responses to the user only by generating visualizations. However,

using additional communication means, e.g., a textual dialogue, could likely improve user experience. It would fosters the idea of having a real conversation. This could improve two things. First, it can increase the immersion of the user in the visual analysis process. Having a conversation on the data could be exciting for a user. A conversation between humans occasionally becomes immersive when both parties enjoy the topic and the direction of the conversation. Consequently, both parties benefit from the conversation. Second, it structures paths for additional information that visualizations normally cannot provide. Visualizations illustrate data structures in a visual manner. However, they lack precise information about these data structures, e.g., a correlation coefficient. Therefore, this information is sometimes added to a visualization through text widgets. Since this information is displayed textually anyway, it could also be embedded in a dialogue.

Third, personalizing multimodal visual analysis is not yet explored. Current approaches consider homogeneous users, although multimodality has tremendous potentials reaching a wide range of users (Grammel et al., 2010). However, users of visual analyses are not homogeneous at all. In fact, the users characteristics have a significant effect on the use of visual analysis as well as on the effectiveness of visualizations (Conati and Maclaren, 2008; Conati et al., 2014; Gajos and Chauncey, 2017; Green and Fisher, 2010; Lee et al., 2019; Toker et al., 2012; Velez et al., 2005; Ziemkiewicz et al., 2011). Taking advantage of the users characteristics could not only accelerate the visual analysis process, but could also reach a new level of user experience.

The following chapters of this dissertation are approaching these challenges to better understand how visual analysis can be opened up to a broader spectrum of prospective users while serving each user individually.

3 Estimating Personalization Potentials of Tasks

In order to investigate effective ways for achieving an intelligible multimodal visual analysis, potential areas for personalization have to be addressed. One way to approach this objective is to estimate the personalization potentials of abstract visualization tasks. This chapter proposes a method for computing relative personalization potentials of abstract visualization tasks based on the *why-how-what* taxonomy of Brehmer and Munzner (2013). The proposed method leverages a ranking of the granular elements of abstract visualization tasks. This ranking considers both the approximately required knowledge for completing an element as well as the user's preferences regarding an element. While the output contributes towards an intelligible visual analysis in the context of this thesis, the proposed method further helps designers to estimate the personalization effort of their systems.

3.1 Introduction

Visual analysis systems generally support users in achieving certain objectives. However, objectives can be manifold depending on the domain and the user. An illustrating example is the investigation of the sales number of cars in the European Union from three different user perspectives. First, a university student with a major in business administration may have the objective to estimate the general market development during the recent years. However, this student does not know the data beforehand. Hence, the visual analysis tools needs to support the student in exploring the data. Second, an analyst at a car manufacturer has the objective to investigate the effect of a newly introduced car model. The analyst knows the data set. Therefore, (s)he can directly apply relevant filters for the desired data attribute combination. Third, a national sales manager needs to report the sales numbers of the last quarter to the line management. As the manager knows both the data and the audience, the objective is now to create information visualizations both comprehensive and fitting to the audience.

As these examples highlight, the data may remains unchanged while the individual objectives vary. Furthermore, each of these presented examples require a different sequence of actions. Imagine, a user follows the second objective of investigating the effect of a new car model. A corresponding sequence of actions may look like: *load data, select desired attributes, create an effective visualization, filter on the desired car model and time frame, compare values*, and so on. These actions or interactions are also named tasks. In visualization research, however, there are various definitions of what a task actually is. One approach is to classify tasks based on their level (Brehmer and Munzner, 2013). Selecting data points can be seen as a low level tasks (Amar et al., 2005) while comparing categories to a certain quantitative value is more a high level task (Liu and Stasko, 2010).

Nevertheless, the amount of unleashed complexity of a task should likely vary from one user to the other. Consequently, users require likely different level of assistance depending on a particular task. A user with experience and knowledge in the creation of effective visual mappings presumably needs less support than a user unaware of the relevant effectiveness studies. Additionally, the tasks themselves further vary in a potential effect of a user's preferences. Selecting particular values in a visualization requires less personalization effort than navigating in a visualization (Hornbæk et al., 2002).

Yet, little knowledge exists on the personalization potentials of visual analysis tasks. In order to identify a task space for the personalization of multimodal visual analysis, Section 3.2 introduces related work and defines relevant terms. Afterwards, Section 3.3.2 analyzes the currently supported task. Based on these classification results, Section 3.3.2 discusses task-specific personalization opportunities by considering the task itself, its relationships to other tasks, and corresponding statistical methods.

3.2 Related Work

While using a visual analysis system, a user typically has a certain objective in the data, e.g., investigating the sales numbers during the last quarter. Generally, an objective is "something that you aim to do or achieve", according to the Cambridge dictionary[1]. In order to reach an objective, a user further needs to execute a corresponding sequence of actions. These actions are often called tasks. According to the Cambridge dictionary[2], a task is "a piece of work that needs to be done" as well as "an action done by a computer such as starting a program, checking email, saving files, etc".

However, a variety of definitions exists in visualization research on what a task in visual analysis actually is. A task can refer to direct interactions with a systems, e.g., selecting a set of points in a scatter plot (Amar et al., 2005; Heer and Shneiderman, 2012; Shneiderman, 1996), but it can also point to mental steps in the analysis process, e.g., formulating a hypothesis (Amar and Stasko, 2004; Reda et al., 2016). Depending on the granularity of the analysis, the definition changes. Therefore, the following section summarizes related work to better understand tasks in visual analysis. It separates the contributions by considering the user's role in the tasks, as in Brehmer and Munzner (2013) and Rind et al. (2016).

3.2.1 How People Use Visual Analysis

From an interaction point of view, Shneiderman (1996) presents task taxonomies in visual analysis as one of the first. Shneiderman (1996) essentially considers the *Visual Information Seeking Mantra* in order to derive a Task by Data Type Taxonomy (TTT). This approach is succinctly summarized in the mantra "Overview first, zoom and filter, then details-on-demand". He further defines the corresponding tasks as: *Overview*, *Zoom*, *Filter*, *Details-on-demand*, *Relate*, *History*, and *Extract*. All tasks directly address direct interactions with a visual analysis system.

Yi et al. (2007) further consider the role of interactions in visual analysis. The authors derive a taxonomy from reviewing existing visual analysis systems and their correspondingly implemented interactions. Based on their results, the authors propose the tasks: *Select*, *Explore*, *Reconfigure*, *Encode*, *Abstract/Elaborate*, *Filter*, and *Connect*. Heer and Shneiderman (2012) propose a taxonomy for interactive dynamics for visual analysis. This taxonomy elevates tasks from the *Visual Information Seeking Mantra* (Shneiderman, 1996) to a higher level of detail. It differentiates tasks between data and view specifications (visualize, filter, sort, derive), view manipulations (select, navigate, coordinate, organize), and process and provenance (record, annotate, share, guide).

[1] https://dictionary.cambridge.org/us/dictionary/english/objective
[2] https://dictionary.cambridge.org/us/dictionary/english/task

While these taxonomies focus on the interaction with visualizations, Amar et al. (2005) investigate tasks from a data point view. Amar et al. (2005) propose a taxonomy for analytical tasks, i.e., what kind of actions is a user conducting in order to generate insights from data. This taxonomy comprises: *Retrieve Value*, *Filter*, *Compute Derived Value*, *Find Extremum*, *Sort*, *Determine Range*, *Characterize Distribution*, *Find Anomalies*, *Cluster*, and *Correlate*. Keim et al. (2006) connect both perspectives. The authors essentially extend Shneiderman's *Visual Information Seeking Mantra* to "Analyze first, show the important, zoom, filter and analyze further, details on demand.". This extended version describes the interactions in visual analytics.

Considering the different steps in visual analysis as a state model is certainly a different perspective. Chi (2000) proposes a taxonomy of visualization techniques using a data state reference model. This reference model differs between data stages and transformation operators. While data stages are *values*, *analytical abstraction*, *visualization abstraction*, and *view*, the transformation operators are *data* transformation, *visualization transformation*, and *visual mapping transformation*. Reda et al. (2016) also model their taxonomy as a state model. Based on investigating user behavior in exploratory visual analysis, Reda et al. (2016) create a Markov chain model. The states are either mental states (formulate hypothesis, form goal, and make observation) or interaction states (brush/link/pan map, and modify layout). Theoretically, their Markov chain model is a fully connected graph. However, the actual transition probabilities from one state to the other are derived from an experiment. The authors show a change in the transition probabilities when varying the device's display size of the visual analysis system i.a., on large screens people tend to form new goals with higher probability.

3.2.2 Why People Use Visual Analysis

The previous related work focus on investigating how people actually use visual analysis from both perspectives: visualization and analytics. The following approaches investigate why people use visual analysis and how systems should support a user.

Sprague and Tory (2012) explore how and why people use visualizations in casual contexts. In order to understand the *why* component, they investigate user goals and regulated motivations. According to Sprague and Tory (2012), a goal can be either intrinsic or extrinsic. Furthermore, an intrinsic goal can refer to *learning & understanding*, *utility*, or *entertainment*. In order to understand the *how* component, the authors analyze the use of fully specified visualizations. They differ the use of visualizations into the categories *recognition*, *short-duration single use*, *short-duration repeat use*, *long-duration single use*, *long-duration repeat use*. The duration refers to how long a visualization is observed by a user.

Furthermore, Amar and Stasko (2004) add what they call "analytic gaps" to the process from data to decisions. In this process, they distinguish worldview tasks from rationale tasks. However, these tasks are subject to the system not to the user. Worldview tasks are *determine domain parameters*, *multivariate explanation*, and *confirm hypotheses*. The rationale tasks are *expose uncertainty*, *concretize outcomes*, and *formulate cause/effect*. All of these tasks describe potential "analytic gaps" where a system potentially might support the user. Being aware of these "analytic gaps" potentially helps to design better visual analysis systems.

Finally, Liu and Stasko (2010) take a top-down perspective on visual analysis tasks. While considering the mental model as a base for their analysis, the authors argue for three purposes of interactions in visual analysis: external anchoring (projection, locate), information foraging (restructuring, explore), and cognitive offloading (create, save/load).

3.2.3 Definition of a Task

The use of the term *task* is quite ambiguous, as shown by the previously discussed related work. Depending on the perspective and the vocabulary, a task is differently defined. In order to further structure the term *task*, meta studies explore high level definitions.

Instead of separately considering the user's interactions or the reasons for using visual analysis, Schulz et al. (2013) define a task as a 5-tuple consisting of the dimensions: *Goal*, *Means*, *Characteristics*, *Target*, and *Cardinality*. A goal describes the user's type of analysis (Exploratory analysis, Confirmatory analysis, or Presentation). While *Means* refers to the interaction with the visualization (Navigation, (Re-)organization, or Relation), *Target* address which part of the data should be focused on (Attribute relations or Structural relations). Lastly, characteristics define the detail level of the output data (Low-level or High-level), while cardinality describes the amount of considered data (Single instance, Multiple instances, All instances). Furthermore, Rind et al. (2016) use a three-dimensional conceptual space of user tasks in visualization design. This space consists of the dimensions: perspective (objectives or how), composition (high to low), and perspective (generic, data, domain, or tool).

Finally, Brehmer and Munzner (2013) define an abstract visualization task by the components *why*, *how*, and *what*. *Why* hierarchically structures different reasons for using visual analysis, e.g., consume, search, or query. The *how* dimension essentially structures interactions with visualizations, but also addresses the visual mapping. *what* eventually describes the underlying data for the task. An abstraction visualization task can further consists of a sequence of elements in each component, e.g., encode, filter, and select in *how*. Hence, this *why-how-what* taxonomy provides a definition both modular and extensible within a component.

As the *why-how-what* taxonomy (Brehmer and Munzner, 2013) provides a comprehensive and extensive structure based on previous contributions, this thesis use a definition accordingly. In the next sections, these taxonomy further forms the base for estimating the personalization effort for a task.

3.3 An Intelligible Task Space

According to the taxonomy of Brehmer and Munzner (2013), an abstract visualization task is a combination of one or more elements from the dimensions *why*, *how*, and *what*. In this thesis' context, the *what* dimension can be considered as fix, since the focus is on structured data. However, the other dimensions remain unchanged. Following this argumentation, a task space can be considered as the set of all possible combinations of elements from the dimensions *why* and *how*. However, the question remains which tasks require a user-specific reaction of the system in order to effectively achieve an intelligible multimodal visual analysis.

RQ 1: How can tasks in visual analysis be systematically structured based on their potentials for personalization?

3.3.1 Procedure

In order to answer this question, the following sections first classify related work from Section 2.4 as well as Section 2.5 in accordance with to their support of the elements of the *how* elements of the *why-how-what* taxonomy (Brehmer and Munzner, 2013). These elements are `encode`, `select`, `navigate`, `arrange`, `change`, `filter`, and `aggregate`.

The elements `select` and `navigate` describe direct interactions with a visualization without changing neither the visualized data nor the visualization itself. `filter` and `aggregate` refer to direct changes of the visualized data. While `filter` restricts the valid visualized data points, `aggregate` condenses entire data attributes to single values e.g. mean, median, or count. `arrange`, `change`, and `encode` summarize all different steps of visually encoding data. However, `encode` creates a new visualization from scratch. In `arrange`, a user only rearranges the data attributes in the visualization, e.g., swapping the X and Y axes. Finally, `change` refers to using a different representation for an already visualized attribute, e.g., using a different coloring schema or using patterns instead of colors. As it is difficult to quantify user's reasons for using a visual analysis system just by its functionality, the related work is not classified by the support of the *why* elements of *why-how-what* taxonomy (Brehmer and Munzner, 2013).

In a second step, the elements from both *why* and *how* are further ranked by their potentially required knowledge as well as the potential influence of a user's preferences. Knowledge in this case refers to both experience and knowledge of methods such as statistics or visual mappings, but not to data to be examined. The *why* elements are `present`, `discover`, `enjoy`, `lookup`, `browse`, `locate`, `explore`, `identify`, `compare`, `summarize`, and `produce`.

The elements `lookup`, `browse`, `locate`, and `explore` describe different search strategies. Depending on whether the target is known to the user as well as whether the location is known, the user's objective changes. In `lookup` both is known. `browse` refers to an unknown target but a known location, while `locate` refers to the opposite. In `explore` both is unknown. `Identify`, `compare`, and `summarize` address different amounts of search targets. Furthermore, `enjoy` describes the use of visual analysis just for fun. `present` requires to show the visualization to a certain audience, while `discover` addresses the hypothesis testing.

3.3.2 Supported Task Space

Since the *why-how-what* taxonomy consists of multiple levels, only the lowest level (e.g., `select`) represents a binary classification of wheather the system presumably supports the method. In order to estimate the support of the higher-level tasks, the classification results of the corresponding elements are aggregated.

Table 3.1 shows the classification results regarding the visualization recommenders. Naturally, all approaches support encoding of data into visualizations. However, they vary in the number of visualization options (see Section 2.4) as well as whether a user can adjust the visualizations. Often, a user can manipulate the recommended visualizations. Yet, `arrange` and `change` are at bit less supported. These manipulations are necessary to enable an exploration of the data. In those approaches, a user cannot easily adjust a shown visualization. Furthermore, `introduce` tasks are rarely supported. Approaches instead focus on the exploration of data sets but less on export the extracted insights or visualizations from the tool.

Considering the importance of a tasks by the count of supported systems, the following order results from the visualization recommenders. Creating a visual mapping (`encode`) is more important than manipulating an existing visualization (`encode`) which is more important than extracting information (`introduce`).

Furthermore, Table 3.2 shows the classification results regarding the approaches using at least natural language for visual analysis. The majority of these approaches support the generation of visualizations (`encode`). It is the essentially supported method, along

	Encode	**Manipulate**	Select	Navigate	Arrange	Change	Filter	Aggregate	**Introduce**	Annotate	Import	Derive	Record
Stolte et al. (2002)	●	◕	●	●		●	●	●					
Mackinlay et al. (2007)*	●												
Mutlu et al. (2015)	●												
Crotty et al. (2015)	●	◕	●	●			●	●					
Vartak et al. (2015)	●	◑	●				●	●					
Siddiqui et al. (2016)	●	●	●	●	●	●	●	●					
Wongsuphasawat et al. (2016)	●	◕	●	●	●		●	●	◑		●		●
Wongsuphasawat et al. (2017)	●	◕	●	●	●		●	●	◑		●		●
Demiralp et al. (2017)	●	●	●	●	●	●	●	●	◔				●
Luo et al. (2018)	●	◔	●						◕		●	●	●
Yalçın et al. (2018)	●	●	●	●	●	●	●	●	●	●	●	●	●
Dibia and Demiralp (2018)*	●												
Moritz et al. (2019)*	●												
Hu et al. (2019)*	●												

Tab. 3.1: Classification of related recommender systems according with the *how* part of the taxonomy of Brehmer and Munzner (2013). Approaches marked with a (*) rely upon other visualization technologies to support the other tasks, but do not implement these tasks themselves.

with `select` and `filter`. In contrast to the visualization recommender approaches, these approaches implement `select` and `filter` not necessarily on the visualization, but through the conversation.

Generally, the analysis shows a similar picture as shown in Table 3.1. If the importance of an element is again determined by the count of supported systems, the following order again results from the visualization recommenders. Creating a visual mapping (`encode`) is more important than manipulating an existing visualization (`encode`) which is more important than extracting information (`introduce`).

Hence, the currently supported task space in both approaches primarily comprises the exploration and reasoning for a single user without a focus on exporting or sharing the produced results. Therefore, the following analysis only focuses on the encoding and manipulation elements.

3.3.3 Estimate Personalization Potentials

This thesis considers personalization as a method for achieving intelligibility in visual analysis. As intelligibility refers to adapting outputs to the user's capabilities and preferences, the estimation of potentials for personalization uses the dimensions of presumably required knowledge as well as potential effects of the user's preferences.

	Encode	**Manipulate**	Select	Navigate	Arrange	Change	Filter	Aggregate	**Introduce**	Annotate	Import	Derive	Record
Cox et al. (2001)	●	◔	●	●			●						
Sun et al. (2014)	●	◔				●	●						
Gao et al. (2015)	●	◕	●	●		●	●	●					
Setlur et al. (2016)	●	◕	●	●		●	●	●					
Aurisano et al. (2016)	●	◔	●				●						
John et al. (2017)	●	◑	●			●	●						
Dhamdhere et al. (2017)	●	◔	●				●						
Srinivasan and Stasko (2018)		◕	●	●		●	●						
Fast et al. (2018)	●	◑	●				●	●					
Hoque et al. (2018)	●	◕	●	●		●	●	●					
Yu and Silva (2020)	●	◕	●	●		●	●						
Srinivasan et al. (2020)	●	◕	●	●		●	●	●					

Tab. 3.2: Classification of related multimodal approaches by the *how* part of the taxonomy of Brehmer and Munzner (2013).

First, the *how* elements are considered. `Select` and `navigate` require relatively little knowledge as they describe common interactions in visual analysis. However, how users eventually navigate in a user interfaces depends on the user him - / herself. (Hornbæk et al., 2002). Hence, the effect of preferences for `navigate` can be seen as slightly higher.

`Filter` and `aggregate` require more experience and knowledge from the user. On the one hand, the user should be aware of the amount of filtered data in order to prevent potentially wrong conclusions. As Zgraggen et al. (2018) show, it is challenging for the user to stay aware of applied filters during a visual analysis. On the other hand, choosing the right aggregation method further requires certain knowledge from the user, e.g., knowing the difference between arithmetic mean and median. However, a user likely prefers aggregations methods differently, while `filter` only depends on the analysis objective.

`Encode` describes the creation process of a visual mapping for given data. It comprises not only the mapping of data attributes on visual variables, but also the transformation of the data. As shown in Section 2.2, creating a visual mapping is a crucial step. It directly determines the effectiveness of a visualization. Creating an effective visualization should always be the main objective of visually encoding data. However, user characteristics also influence the effectiveness of a visualization (see Section 2.2.3). Depending on user characteristics (Conati et al., 2014; Haroz and Whitney, 2012), and personality aspects (Gajos and Chauncey, 2017; Green and Fisher, 2010), a visualization's effectiveness varies. Accordingly, the potentials for personalizing `encode` can be considered as the highest among the *how* tasks.

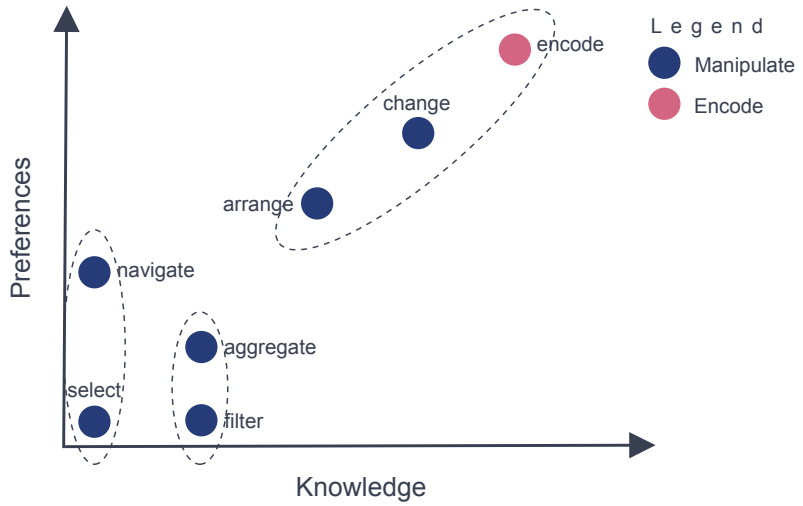


Fig. 3.1: Classification of the personalization potentials of the *how* elements.

`Arrange` addresses the spatial repositioning of data attributes, while `change` summarizes changes of the visual mapping. A user who changes a visual mapping likely disagrees with a provided mapping for some reasons. These changes may focus on the used coloring scheme, but can also address different visualization types or completely different visual mappings. Hence, both tasks require the expressions of the user's preferences. Assuming the visualization to be manipulated already represents an effective solution for the given data, the user's knowledge arguably plays a smaller role in these tasks. If the effectiveness would play a relevant role for the user, (s)he would continue using the existing visualization. However, `arrange` describes less adjustments on the visual mapping than `change`, according to Brehmer and Munzner (2013). Additionally, `change` addresses less adjustments on the visual mapping than `encode`. As there are many opportunities to change a fully specified visualization, `change` basically describes the area between `arrange` and `encode`.

Based on this argumentation, tasks can be ordered by their personalization potentials (see Figure 3.1 illustrating a qualitative diagram). The tasks related to changing a visual mapping have the highest knowledge requirements on the user and simultaneously incorporate the user's preferences. Hence, their potentials for personalizations is high. A cluster of moderate personalization potentials consists of tasks changing the underlying data. While the user's preferences play almost no role, the user's experience in visual analysis surely does. Lastly, the tasks on interacting with the visualization have little personalization potentials.

Second, the *why* elements are ranked by their potentials. By nature of visual analysis, the *why* elements relate to the *how* elements. For instance, comparing two categories with respect to a quantitative attribute requires encoding the data, potentially filtering the data and eventually aggregating the data. Additionally, the *why* often requires the use of statistical methods, e.g., in `discover`. These methods provide mathematical bases for drawing valid conclusions and

making proper decisions. Hence, the personalization potentials of the *why* elements needs to be further investigated by considering the corresponding statistical methods.

As `lookup`, `browse`, `locate`, and `explore` describe different search strategies, they have different potentials for personalization. Considering the knowledge about a location as factor for influencing the user's preferences, those elements where location is known can be considered as higher in preferences. However, finding properly the location of a target requires experience and knowledge about how to properly combine data attributes. Especially, `explore` refers to the exploration of new data. During the exploration of a new data set, a user might wonder which data attributes are worth it to investigate. In this situation, the system might proactively support the user by recommending a list of attributes (also addressed by Heer and Shneiderman (2012)). These attributes can be worth it to combine with the currently visualized attribute(s). Kassel and Rohs (2017) highlight the use of the mutual information for recommender data attribute combinations in visual analysis. While adding the data attributes was positively seen by the participants, automatically removing attributes was not welcomed.

`Discover`, `identify`, `compare`, and `summarize` are closely related to each other. `Identify` refers to the characteristics of a single target (e.g., the average miles per hours of sedans), while `compare` and `summarize` address multiple targets at the same time (e.g., the average miles per hours of multiple car categorizes). Furthermore, the task `discover` subsequently focuses on verifying hypotheses drawn from either `identify`, `compare`, or `summarize`. However, `discover` especially demands statistical knowledge from the user. Conducting hypothesis testing properly requires knowledge about the differences between the various statistical tests, applying these test correctly and eventually interpreting the test results (Zuur et al., 2010). Hence, these tasks contain a lot of required knowledge.

Furthermore, Zgraggen et al. (2018) show a major issue in hypothesis testing also holds for visual analysis, namely the multiple comparison problem. The authors identify that users unconsciously tend to draw wrong conclusions when combining multiple filters. While alternated bar charts with the standard deviation might help an experienced user to draw correct conclusions, a novice would likely fail. A personalized interactive system for visual analysis must find a solution for this challenge, since anx user should be prevented from making mistakes.

Lastly, `present` and `enjoy` directly highlights the relationship to `encode`. However, `enjoy` likely incorporates an additional amount of preferences as it directly describes the user's emotions during the visual analysis. However, `present` requires knowledge about the audience from the user.

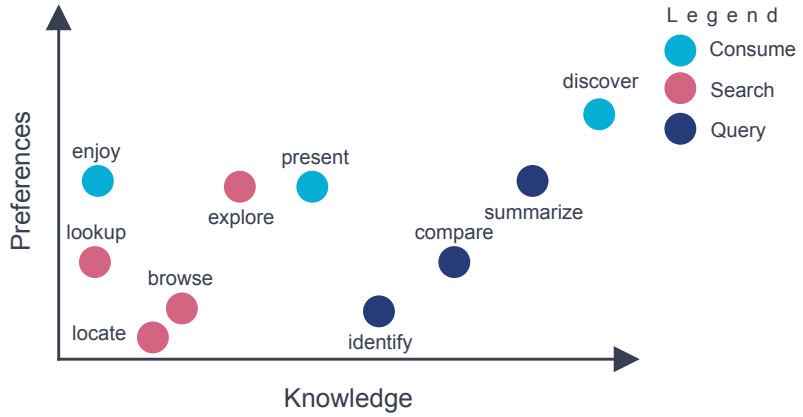


Fig. 3.2: Illustration of the personalization potentials of the *why* elements.

Based on this argumentation, the tasks can be again ordered by their personalization potentials (see Figure 3.2 illustrating a qualitative diagram). While these address mainly the reason why a user is using the visual analysis, the personalization potentials are less clear. Considering the tasks from the corresponding mathematical point of view makes it easier to define personalization potentials. Tasks related to hypothesis testing show the highest potentials, since required knowledge it not necessarily given (Zgraggen et al., 2018; Zuur et al., 2010).

3.4 Discussion

Section 3.3.2 introduced a set of currently supported tasks in multimodal visual analysis, while Section 3.3.3 estimated the personalization potentials of these tasks in terms of both required knowledge and preferences. Essentially, the Figures 3.2 and Figure 3.1 illustrate a clearer picture on the relative personalizations potentials of *why* elements and *how* elements, respectively.

Implication for designers: The relative ranking of the different elements provides a method for computing the relative personalization potentials of entire abstract visualization tasks, since these tasks consists of a combination of elements from the areas of *why*, *how*, and *what* (Brehmer and Munzner, 2013). The personalization potential $pref(.)$ of an entire abstract visualization task T can then be computed:

$$pref(T) = \frac{1}{|T|} \sum_{t \in T} (t_k, t_p) \tag{3.1}$$

where for element t, t_k and t_k refer to the rank according to knowledge and preference, respectively.

Let us consider two exemplary abstract visualization tasks. On the one hand, a system should support hypothesis testing on the data. As the hypothesis exits either implicitly or explicitly,

a user initially searches for the known target at the known location (`lookup`) in order to apply a hypothesis test (`discover`). This objective requires to encode the data (`encode`) as well as apply filter (`filter`) and aggregation (`aggregate`). The corresponding average rank would be $(3, 3)$ and $(4.5, 3)$ for the how methods and the why methods, respectively. On the other hand, a system should additionally support the comparison of data categories against a certain quantitative objective. This task is similar to the previous but instead of verifying a hypothesis, a user focus on the comparison. Hence, the corresponding average rank would be $(3, 3)$ and $(3.5, 2)$ for the how methods and the why methods, respectively. As the task appears somehow similar, however, they have different potentials for supporting a user on his/her objectives.

Generally, a designer considers a defined set of desired tasks when designing a visual analysis system. For these tasks, the personalization potentials can be computed accordingly. This task ranking likely helps to better estimate the challenges of personalization in visual analysis. Assuming that added personalization functionality increase the user experience in visual analysis, the method helps to identify the most effective areas for personalization.

Relationship between the elements: The elements from the why and how category have an inherent relationships with each other, since both need to be determined for an abstract visualization task. For instance, `enjoy` requires the encoding of data (`encode`). Otherwise, a user would not see any visualization which could be enjoyed. Apparently, almost every why element has a certain relation to `encode`. Additionally, `lookup` directly addresses `filter`, since a user presumably directly applies a filter on the data when both target and location are known. Therefore, the why elements are only ranked by the knowledge and preferences for inherently required methods, e.g., hypothesis testing for discover, but not by considering the related visualization effort, e.g., encoding data. Therefore, this circumstance also prevents a joined ranking of the personalization potentials for all elements of the typology.

Adaptation & automation: Elements in the right bottom corner refer to more required knowledge from the user. Depending on the user, a system might support more or less in these tasks. However, they should not be automatized, but individually support the user instead. For instance, exploring a new data set (`explore`) is likely a motivating task for a user. Directly presenting all relevant insights and takeaway messages to the user likely decreases the user's motivation in using the system, because it may overwhelm the user (Norman, 2010). Additionally, `discover` focus on hypotheses testing, the selection and execution of the corresponding methods can be done by the system. However, the way how this information should be then communicated likely requires knowledge about the user.

Furthermore, the top right corner comprises tasks with high potentials in both dimensions. For instance, creating a proper visual mapping requires both knowledge and consideration of the user's preferences. The potential for automation is further demonstrated by the related

work in Section 2.4. As a visual mapping can be changed at any time step (`change` and `arrange`), potential mistakes in the automatic generation of a visual mapping can be repaired. Hence, a system should automatically propose a visual mapping, but in a user-specific way.

In either cases, a variety of elements address a certain effect of the user's preferences. In order to properly support these tasks, the user's preferences have to be learned. However, data on either user behavior or preferences in multimodal visual analysis system is scarce. Hence, methods for personalizing in these tasks require to learn the user's preferences.

3.5 Limitations

The ranking is subject to the argumentation as well as previous work on certain elements, e.g., the effect of user characteristics on the effectiveness of visual mapping (Conati and Maclaren, 2008). Therefore, the ranking may be different if new empirical evidence about the user's preferences for particular elements emerges. Furthermore, the approximation of the personalization potentials of an abstract visualization task rests on a relative ranking.

Another approach might be to explore a continuous base where potentials of elements are represented by probability distributions. This method would be able to express uncertainty in the classification of the potentials. However, achieving this quantitative ranking is very difficult. Imagine, an experiment requires from a participants to decide how much more and less, respectively, complex a task is compared to all other tasks. How should a participant be able to rank the complexity of tasks in visual analysis when this participant is not experienced in the domain. Although this kind of experiment could be conducted, the question remains how reliable the results of such experiment are. Therefore, a relative ranking based on literature appears to fit better to estimate the personalization potentials of visual analysis tasks. Overall, the ranking expresses the current state of knowledge with sufficient granularity.

3.6 Summary

This chapter investigated the potentials for personalization of visual analysis tasks. Initially, the term *task* was defined by considering related work to the classification of visual analysis tasks. Based on the multi-level visualization task typology of Brehmer and Munzner (2013), the related work described in Section 2.4 and Section 2.5 was classified according to their supported task space. While these classification results provide insights on generally supported tasks, Section 3.3.3 discussed the potentials for personalization. For each task, its personalization potential was argumentatively derived by considering the required knowledge and the effect of the user's preferences. These elements were further ranked according to their relative potentials in both dimensions. This resulting task space not only structures

the following chapter of this thesis by narrowing the focus on the relevant tasks, but also addresses essential required areas of support and adaptation.

4 Investigating the Use of Speech and Visualizations

In order to further investigate personalization opportunities for multimodal visual analysis, this chapter explores potential behavior of users with visualizations and text-based interfaces. The first experiment investigates how people use visualizations in the field by analyzing the used visualizations ($N = 1669$) in scientific publications ($N = 544$). The analysis shows a preferred use of classical visualizations for investigating unknown data sets, but also addresses a potential lack of visualization knowledge. Furthermore, it reveals a narrow visualization space limited to a handful of visualization types as well as a trend for visualizing aggregated data. The second experiment ($N = 18$) investigates how users would orally command a potential system in visual analysis. While the similarity of the collected commands is high in tasks for generating visualizations, the commands are more diversely formulated in tasks focusing on changing a visualization.

4.1 Introduction

The previous chapter approaches an intelligible multimodal visual analysis by essentially considering the personalisation potentials of abstract visualization tasks. According to this theoretical approach, tasks have high potentials for achieving an intelligible visual analysis which include either encoding of data, discovering of data relationships or exploring of data. Additionally, current multimodal visual analysis approaches predominantly leverage a combination of text-based and speech-based interactions, respectively, with touch-based interactions to further lower barriers for users (Aurisano et al., 2016; Gao et al., 2015; Hoque et al., 2018; Kumar et al., 2017; Setlur et al., 2016; Srinivasan and Stasko, 2018). Especially, novice users effectively benefit from using text or speech for generating visualizations for abstract data (Grammel et al., 2010). Furthermore, this combination of modalities further raises hidden synergies (Cohen et al., 1989).

However, the users behavior needs to be taken into account as well in order to achieve an intelligible multimodal visual analysis which adapts to the user. Accordingly, Norman (2010) states "we must design for the way people behave" to raise the user experience. Although Reda et al. (2016) investigate user behavior in visual analysis as well as proposed a Markov chain model for better understanding the transitions in the behavior, little knowledge exists in the use of multimodal visual analysis.

The visual mapping, on the one hand, represents a central aspect of any visual analysis system. Additionally, it further contains a high potential for personalization in both dimensions knowledge and preference. As knowledge can be automatized by incorporating knowledge from effectiveness studies (Moritz et al., 2019), one part of the challenge is already covered. In order to approach the preference dimension, however, the use of visual mappings in the field needs to be understood. Knowledge on the use of visual mappings likely reveals useful trails.

The use of natural language based interactions in multimodal visual analysis, on the other hand, is certainly unclear, although speech and text, respectively, represent the central interaction modality. Understanding the structure and patterns in the use of speech for visual analysis in terms of generation and manipulations of visualizations likely contributes to higher user experience too.

Hence, this chapter's objective is two-fold. First, Section 4.3 highlights both structures and patterns in the formulation of commands against a prospective text-based interface of a visual analysis system through an online survey ($N = 18$). Second, Section 4.4 collects and analysis preferences in the design of visualizations through an analysis of visualizations ($N = 1669$) extracted from scientific publications ($N = 544$). In summary, the results contribute to a better understanding of both the use of visualizations and text-based interactions in visual

analysis. Additionally, they substantially form a behavior-based base for the design of multimodal visual analysis tools, along with the task space from Chapter 3.

4.2 Related Work

This section discusses related work concerning the use of visualizations as well as the use of text-based interactions. Both fields essentially contribute towards the understanding of user behavior with multimodal visual analysis systems.

4.2.1 Use of Visualizations

Dasgupta et al. (2015) analyze the use of visualizations within the domain of climate research. By observing climate researchers designing visualizations, they identify several design problems, e.g., the incorrect use of visual channels (Carpendale, 2003) or the invalid use of categorical data in a scatter plot. Recently, Dasgupta et al. (2017) further explore differences between subjective impression and objective fact and the resulting effect on the judgment of climate researchers. They found no influence of the degree of familiarity of visualization types on the judgments, neither in a subjective nor in an objective way.

The use of visualization is closely related to the process of creating visualizations. Grammel et al. (2010) observe how novices are designing visualizations. Their findings show problems in creating effective visual mappings. In order to prevent such kind of errors, Heer et al. (2008) provide helpful guidelines to engage potential new visualizers. Pretorius and Van Wijk (2009) analyze how professional designers are visualizing data. Apart from insights on how professional designers create visualizations, Weaver et al. (2006) consider a visualization design process in which professional designers are guiding novices.

In addition to empirical laboratory user studies (Dasgupta et al., 2015; Dasgupta et al., 2017), research investigates approaches on automatic extraction of visualizations. ReVision (Savva et al., 2011) automatically extracts bitmaps out of documents. It further applies perceptually based design principles to recommend the user alternatives to the initially designed visualization. Jung et al. (2017) envision ChartSense. The authors assume a certain amount of poorly designed visualizations in published documents. FigureSeer (Siegel et al., 2016) and the work by Poco and Heer (2017) aim to improve the extraction of underlying data of extracted visualizations.

4.2.2 Interactions through Speech

Although research on speech-based visual analysis systems is a recent topic, there is still a lack of information concerning interactions with such kind of systems. Srinivasan et al. (2019b) provide insights on commands in multimodal interfaces through a think-aloud user study. However, the authors investigate the commands with a photo-editing software, but not with a visual analysis system. Furthermore, they focus on the discoverability of functionality while the following experiment investigates structures of the speech interaction.

Setlur et al. (2016) run a web-based online study for collecting statements against a text-based visual analysis system in the domain of geographic data. Showing a visualization, a participant should provide statements regarding this visualization. The authors identify 12 different query types, mainly focusing on analytic tasks. However, some query types also address interactions with or modifications of the shown visualization. In contrast, the study of this chapter aims to understand textual patterns in the direct interaction in visual analysis focusing on generating visualizations and modifying visualizations.

4.3 Word Space

This section explores text-based interactions against a prospective visual analysis system. It is designed as an online survey where participants should formulate potential requests without any limitations on the support functionality. Hence, the resulting space of requests is likely wider compared to recorded interactions with an actual technical prototype. Eventually, the results help to better understand underlying structures and differences in the interaction in order to achieve a solid NLP routine for a technical prototype. Additionally, the results highlight common command structures among tasks.

RQ 2: How do text-based commands look like when people generate and transform visualizations with a prospective visual analysis system?

4.3.1 Procedure

In order to fulfill the objectives of this experiment, an online survey is designed. This survey consists of 14 different tasks covering potential command categories as discussed in Section 3.3. This tasks further belong to the *how* element of the *why-what-how* taxonomy (Brehmer and Munzner, 2013). In the tasks of category *generate*, a participant sees a visualization based on certain data attributes. The task for the participant is then to formulate a command in order to generate the shown visualization while assuming a system would exist which properly understands every given command. In each other task category (*include, exclude,*

highlight, color, filter, and *transform*), a participant sees two visualizations. A command is required to transform the left allocated visualization into the right allocated visualization.

In order to reduce biases in data, each participant gets a randomly assigned sequence of these 14 tasks. Additionally, all used visualizations are based on a car data set (StatLib, 2005; Donoho and Ramos, 1982) due to make the tasks more reasonable to the user. The link to this designed online survey has been broadcast within an industry company.

4.3.2 Participants

18 persons participated in this experiment. Overall, they gave 244 complete answers (8 answers were empty). It takes an average of 23.35 minutes for a participant to complete the sequence. Furthermore, 11 participants report to have a professional background in data science. Therefore, the majority (11) of the participants create their visualizations with a scripting language like Python or R, instead of using a professional tool like Tableau.

4.3.3 Results

After both cleaning and filtering the collected answers, 258 properly given commands remain. 5% of these commands are politely formulated. A command is *polite* when it contains either the word "please" or is conjunctively formulated. Additionally, some participants rather tend to formulate questions than commands, e.g., "how does it look when we group by class?", while the majority use the imperative. Generally, the participants tend to maintain their way of formulating commands during the study. Consequently, a participant who, e.g., gives very detailed commands at the beginning of the study will also give very detailed commands at the end of the study. As the tasks cover a different level of complexity, the maximum level of detail a command can get also varies. In order to achieve a comparable ground among the tasks, for each sentence of each task the degree of covered elements is relatively counted. Figure 4.1 highlights the distribution of the level of detail (red) as well as the similarity between the commands (blue) per task. Additionally, Table 4.1 provides the corresponding statistical indicators.

In the tasks *generate*, *include*, and *color* participants formulate relatively detailed commands on the imaginary system, yet not all commands achieve the same level for detail. In the task *generate*, on the one hand, a command with a moderate level of detail looks like *"show me the distribution of co2 values"*. The participants specifies relevant data attribute ("co2"), the primary action ("show"), and the objective ("distribution"). However, another participant uses the following formulation for the same task *"visualize a 1-D histogram of the distribution of co2-emissions values and get a continuous model to bins"*. In this command, the participant additionally specifies an overlaying Kernel Density Estimation (KDE) plot

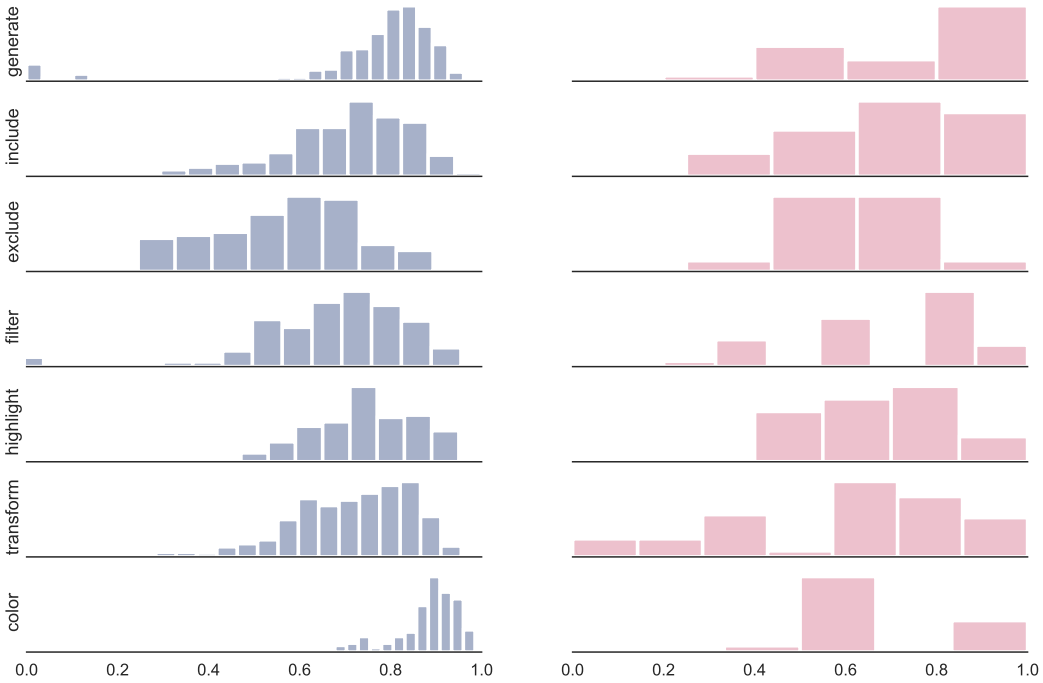


Fig. 4.1: For each task, the distributions of both the similarity (blue) and the level of detail (red) of the given commands.

along with the objective ("histogram of the distribution"), primary action ("visualize"), and the relevant attributes ("co2-emissions"). In task *color*, on the other hand, the formulations vary from only mentioning the target color *"change color to orange"* to naming the object which color should be changed *"Change curve color to orange"*.

However, the data reveals a different picture in the other tasks. In the tasks *transform* and *exclude*, the commands vary in the level of detail. A transformation task concerns the change of the visualization type from a point plot to a box plot. One participant gives a short but precise example for this transform through the following command: *"transform to boxplots"*. The same transformation of the visualization can also be described by *"Transform the line plot with standard deviation indications into a box-plot, with outlier whiskers, for every available year in the original plot"* as provided by another participant. Another transformation task required from a participant to transform a bar plot into a strip plot. This transformation reveals a more detailed look at the actual distribution of the data attribute. A participant states *"Turn bars into dots"*. This command refers to domain knowledge that a strip plot consists of dots.

In addition to the level of detail, commands also vary whether they contain a negation. For instance, the formulations *"remove N"*, *"Filter to only Y in Start/Stop Automatic"*, and *"Only show the left bar plot"* should each filter the data on cars with start / stop automatic. Both statements *"remove N"* and *"Only show the left bar plot"* require context information about the appearance of the visualization, in contrast to *"Filter to only Y in Start/Stop Automatic"*.

Task	Similarity μ	Similarity σ	Precise μ	Precise σ	Polite
generate	0.74	0.22	0.76	0.25	0.08
include	0.65	0.25	0.75	0.21	0.11
exclude	0.57	0.16	0.62	0.18	0.06
filter	0.65	0.21	0.69	0.20	0.03
highlight	0.74	0.11	0.67	0.19	0.06
transform	0.69	0.20	0.60	0.28	0.02
color	0.88	0.07	0.74	0.18	0.00

Tab. 4.1: For each task, statistical information on the speech command structure regarding the similarity, the preciseness, and the politeness.

Additional examples which require context information are *"Don't plot the green cluster"* or *"Remove green"*. Finally, an example of the difference between positive and negative formulations is from the highlight task. For instance, participants state *"hide in grey category 4 and 5"* and *"Highlight data labeled cylinders 6.0"*. Generally, the commands given mainly require a certain degree of context information about either the current visualizations or the visualized data, in all tasks, except for *generate*.

Additionally, participants use different words for indicating actions for the system. As shown above, participants use for instance "show" or "visualize" for generating visualizations. This word is the root element of a sentence. Figure 4.2 reveals these wording differences by illustrating the task-wise distributions of the root elements in the given commands. In the task *color*, almost every command has the same root. In the tasks *generate*, *include*, and *filter*, a small set describes the root element. For the other tasks, participants use a wider range of words. This observation also matches with the similarity in the commands between the tasks, shown in Figure 4.1.

Furthermore, Figure 4.2 exhibits a relationship between tasks and words. One the one hand, the word *"change"* is only used in the context of transforming or highlighting certain aspects of the visualizations. The word *"create"* only exists in the context of generating a new visualization. The same holds for *"accumulate"* for excluding something from a shown visualizations. On the other hand, the word *"show"* appears in almost every task.

4.3.4 Discussion

The empirical results reveal insights for the development of a prospective speech-based visual analysis system as well as show certain behavior of the participants during a potential interaction.

Finding 1: Text-based interactions in visual analysis follow a task-oriented pattern. As the results show, the participants predominantly formulate their requests as an imperative. Additionally, they neglect polite formulations, although they are not forced to by the design

Fig. 4.2: Distribution of the used root elements for formulating commands on a visual analysis system depending on the task.

of the study. Nevertheless, the patterns refer to the classification of Shechtman and Horowitz (2003). Shechtman and Horowitz (2003) identify three different types of interaction patterns with a speech and text-based, respectively, interface: task-oriented, communication-oriented, or relationship-oriented interaction. Both drawn indicators from the study results directly refer to a task-oriented interaction between the user and the system in which only achieving the objective is of interest.

Finding 2: The wording comprises the relevant tasks. The root element (primary verb) of each command provides a good indication for the actual objective or task of the user. While certain words only appear in small set of tasks, other words are widely used. This insight helps to design explainable and useful NLP routines in prospective systems. Additionally, it reveals latent relationships between the tasks which share common words. For instance, the task *include* appears to be the opposite of *exclude* and *filter*. The task *transform* covers a wide range of different words at first glace. However, considering the mentioned words by their semantic similarities, the space of words collapses. The words “transform”. “transpose”, “swap”, and “rotate” have a certain common ground. Hence, the task can also be described by a certain family of words.

Finding 3: A system needs to be robust against a varying level of details. Figure 4.2 shows the distribution of the root element of each command among the tasks. This distribution could further reduce the search space for a prospective system. Given a new command and the corresponding root element, a system might take the word’s distribution to systematically infer the user’s potential objective. This would reduce the search space for the system. Additionally, the system needs to be robust against a varying level of detail. Depending on the task, participants partially vary in the used detail level. A potential approach for handling this varying level of detail would be to focus on the root element and the existence of data attributes or their synonyms. The missing level of detail could then be added by another request or by using another modality.

Finding 4: The context of the current analysis state has to be taken into account. Participants directly take advantage of the current context of the analysis for formulating the next request. However, this context does not only include the appearance and specification of the visualizations itself, but also includes the data or data analysis step. Precisely, a resulting context has to formalize the current step of the user’s data analysis, the existing data set, and the currently shown visualization including specification and orientation.

4.4 Visualization Space

While the previous section identified patterns in the command structures against a prospective visual analysis system, this section focuses on identifying patterns in the use of visualizations.

Feature	Values
type	table | bar chart | line chart | box plot | scatter plot | pie chart
X	nominal | ordinal | quantitative
Y	nominal | ordinal | quantitative
colored	True | False
patterned	True | False
grouped	True | False
sorted	True | False

Tab. 4.2: Features and their corresponding values for the categorization of the extracted visualizations.

Generally, a variety of visualizations exists. Additionally, new visualization techniques are frequently proposed to further improve information visualization. However, a visual analysis system for interactive data analysis might focus on the visualizations which people actually use. Focusing on these used visualization effectively reduces a potential learning effort for the user and for a recommender system, respectively. The experiment's outcome is a visualization space for a recommender system. A clearly described visualization space of actually used visualizations effectively supports a personalized multimodal visual analysis, since it reduces the computational effort as well as allows a ranking on a defined set. Additionally, another objective is to find potential shortcomings and issues in the design of visualizations for structured data.

RQ 3: Which visualizations are used in scientific publications for highlighting insights from structured data?

4.4.1 Procedure

The publications (full and short paper) from the *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016) form the data base for this experiment. These documents are chosen for two reasons. First, publications in the area of Human-Computer Interaction (HCI) research usually contain user studies which produce structured data. Since data in the majority of business processes is structured as well, these documents represent an accessible resource for analyzing how people design visualizations. Second, scientific publications – especially the papers at the CHI – are usually of high quality and get reviewed by multiple experienced people. Hence, the designed visualizations are likely well readable and interpretable.

All visualizations are manually extracted from the collected documents. Afterwards, the extracted visualizations are manually classified according with the features shown in Table 4.2. This method provides both a structured way to investigate common visualization approaches and a knowledge base for a visual analysis system, later on.

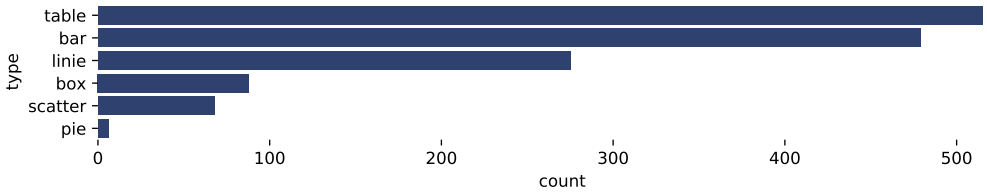


Fig. 4.3: Distribution of the used visualization types.

4.4.2 Results

Overall, the *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016) contains 544 paper. From these 544 papers, 1669 visualizations are manually extracted and classified. These 1669 visualizations base on 85 unique visualization specifications. On average, each document contains 3.07 ($\sigma = 2.91$) visualizations. The majority of these visualizations use the visualization type table followed by bar, line, box, scatter, and pie charts in descending order (see Figure 4.3). Furthermore, Figure 4.4 highlights the Entropy based on the different visualization types used per document. The authors seem to prefer a certain visualization specification within their documents.

Depending on the actually visualized data attribute, the most used visualizations change. Table 4.3 shows the top-5 most used visualizations for illustrating a categorical and a quantitative data attribute. Overall, 473 (28%) visualizations illustrate this attribute combination. On the one hand, the predominant favorite is a vertical bar chart with additional color encoding. This visualization specification represents 43% of all used visualizations with respect to this attribute combination. The identical horizontal specification is less preferred by the authors. On the other hand, the box plot occurs only 24 times, although it reveals more insights on the actual distribution of the numerical attribute. In four out of five cases, the authors enrich visualizations with a coloring in order to encode categories. Additionally, 6% of these visualizations are sorted by the quantitative attribute.

When authors encode an ordinal and a quantitative attribute, they prefer a vertical oriented visualization (see Table 4.4). In contrast to the situation of a categorical and a quantitative attribute, both bar and line charts are almost equally preferred by the authors. However, these two specifications only differ in the visualization type while the other dimensions are identical.

In case of two quantitative attributes, authors prefer a line chart followed by a scatter plot. The line chart likely shows a trend or a linear relationship between the attributes while the scatter plot reveals the actual underlying data points. The line chart often encodes an additional categorical attribute likely to compare the different groups. However, the two bar

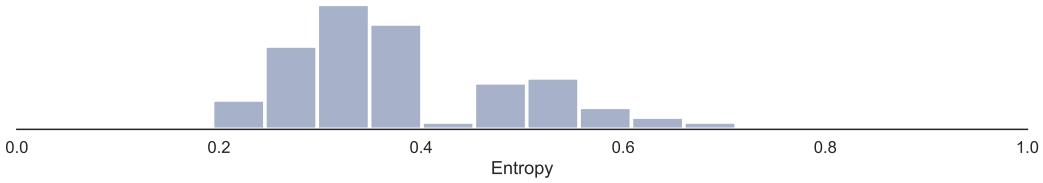


Fig. 4.4: Distribution of the Entropy of the used visualization types per document. Only documents are considered which used more than one visualization.

Type	X-axis	Y-axis	Colored	Patterned	Count	Share (%)
bar	categorical	quantitative	True	False	205	43
bar	categorical	quantitative	False	False	72	15
bar	quantitative	categorical	True	False	47	10
line	categorical	quantitative	True	False	31	7
box	categorical	quantitative	True	False	24	5

Tab. 4.3: Top-5 visualization specification for categorical and quantitative data.

chart specifications refer to histograms for showing the distribution of a single quantitative attribute. In equal amount, the histogram is used for single and multiple groups.

All these most preferred visualization specifications reveal a certain tendency to simultaneously encode more than two data attributes within one visualization. While in 6% of the created visualizations a coloring is alternated by an overlaying pattern or structure, a pattern or structure is in 57% supported by a coloring. However, the amount of additionally encoded categories varies among the different visualization types. Through either a pattern or a coloring, a scatter plot encodes $\mu = 4.72$ ($\sigma = 9.79$) categories, a bar plot $\mu = 2.92$ ($\sigma = 1.49$), a line plot $\mu = 3.12$ ($\sigma = 1.86$), and a box plot $\mu = 2.39$ ($\sigma = 1.60$), respectively.

Furthermore, only a very small amount of visualizations implement a redundant visual mapping of the data attributes. Generally, redundancy can help to highlight certain aspects of the visualization, but it also sacrifices space for additional information.

Type	X-axis	Y-axis	Colored	Patterned	Count	Share (%)
bar	ordinal	quantitative	True	False	58	40
line	ordinal	quantitative	True	False	45	31
bar	ordinal	quantitative	False	False	13	9
line	ordinal	quantitative	False	False	6	4
box	ordinal	quantitative	True	False	6	4

Tab. 4.4: Top-5 visualization specification for ordinal and quantitative data.

Type	X-axis	Y-axis	Colored	Patterned	Count	Share (%)
line	quantitative	quantitative	True	False	123	42
scatter	quantitative	quantitative	False	False	33	11
line	quantitative	quantitative	False	False	27	9
bar	quantitative	quantitative	True	False	22	7
bar	quantitative	quantitative	False	False	22	7

Tab. 4.5: Top-5 visualization specification for two quantitative data attributes.

4.4.3 Discussion

The empirical results highlight a certain visualization space for recommending personalized visualizations and patterns in the design of visualizations for structured data.

Finding 1: Narrow set of used visualizations. Generally, the set of used visualizations is small. The collected data contains a set of only 85 unique visualizations out of 1669 used visualizations. While these 1669 visualizations from the documents certainly differ in their eventual appearance, they are similar in their specifications. Furthermore, only four different visualization types are practically used, since pie charts do not really occur and tables are actually not a visualization in the context of this thesis. Although libraries as seaborn[1], or ggplot2[2] offer a wide range of potential visualization types, the "classical" visualizations are still preferred.

Furthermore, the analysis also reveals some potentially misleading visualization specifications. On the one hand, some visualizations represent a sorted quantitative attribute. In this case, the message of the authors might be clearer communicated, but it also can distract the reader. One the other hand, the line charts are occasionally used in combination with an ordinal data attribute. While proven through practice, it is actually not applicable, since the space between adjacent elements is empty on an ordinal scale. However, the line chart draws as a line between all adjacent points.

Finding 2: Reuse of visualizations. In addition to the narrow visualization space, the analysis further shows a tendency of sticking to a certain visualization type. Figure 4.4 illustrates little variance in the used visualization types within one document. The authors seem to follow a certain visualization specification either purposefully or accidentally. However, a paper often represents only one data set, e.g., the results of a user study. Since these data sets often consist of only a few different data attributes, it might be reasonable to follow a certain visualization design. Nonetheless, it shows preferences of the creators for the visualizations.

[1] https://seaborn.pydata.org
[2] https://ggplot2.tidyverse.org

Finding 3: Coloring is predominantly used. When it comes to encoding additional data attributes, the authors tend to encode them using a coloring. Furthermore, a coloring is occasionally combined with pattern. Both visual mapping options serve the categorical data. However, the authors predominately prefer a coloring to a pattern. This fact further aligns with the findings of Mackinlay (1986). He show that a color-encoded categorical attribute is more effective than a pattern-encoded one.

Additionally, the authors tend to prefer the X axis to the Y axis. Based on the data shown in Table 4.5, a histogram is primarily vertically drawn. Hence, the authors choose the X axis when only one data attribute should be illustrated. However, the coloring or patterns are essentially used when the X axis and the Y axis are already allocated with one data attribute each. Eventually, these results address a preference-based ranking of the visual encoding opportunities, namely $X > Y >$ coloring $>$ pattern.

Finding 4: Trend for aggregated visualizations. The analysis shows a trend for aggregated visualizations. The authors tend to prefer bar and line charts. Those visualization types reduce the actual distribution of the quantitative data attribute to a primary statistical indicator, e.g., mean. This reduction helps to see differences between groups by directly visualizing the differences. However, only seeing the mean can also be misleading, as the standard deviation might be large. Though, scatter plots and box plots are typically harder to read. The scatter plot suffers under the unclear representation of the actual relationship. It often requires an additional correlation coefficient value to be sure about the relationship. The box plot requires from the read certain knowledge about the different areas of the box.

4.5 Summary

This chapter focused on identifying both potential behavior patterns in text-based visual analysis and in the use of visualizations. Through the conducted analyses, research question 2 and 3 have been answered. All in all, this chapter provides two essential elements for achieving an intelligible visual analysis:

1. A *visualization space* through a structured analysis of used visualizations in scientific publications. The analysis reveals a narrow use of visualization types, implicit preferences on the visual variables, and certain preferences for aggregated visualizations.
2. A *word space* through an online survey. The experiment highlight task-specific wordings in the commands against a prospective text-based visual analysis system. Additionally, it highlighted useful patterns in the formulations.

All analyses follow the idea of understanding trails which can be used to identify personalization avenues in multimodal visual analysis. Based on these findings, the next chapter

explores how the different spaces effectively support the design of a technical prototype for an intelligible multimodal visual analysis.

5 Valletto: A Multimodal Visual Analysis System

This chapter introduces a design for multimodal visual analysis system considering findings from previous chapters. The objective is to increase intelligibility of multimodal visual analysis. While speech and touch represent the modalities for the user, the system answers through visualizations and text. A persistent dialogue fosters the conversation with the user. This dialogue contains both a user's utterances and dialogue acts on data facts. Regarding the utterances of a user, the design further highlights the system's computations in order to increase interpretability. The available visualization space comprises the identified visualization space from Chapter 4. The design is implemented by a prototype named Valletto[1]. Valletto is evaluated through tow user studies. First, an expert review ($N = 4$) focuses on design and interaction mistakes. Afterwards, a controlled experiment ($N = 13$) compares Valletto with Tableau. The results reveal better and faster decisions with Valletto. Especially the dialogue helps to improve the decision making. Generally, this chapter provides further the testbed for the subsequent chapters of this thesis.

Disclaimer: The content of the following chapter is partly published in the article: Jan-Frederik Kassel and Michael Rohs (2018). "Valletto: A Multimodal Interface for Ubiquitous Visual Analytics". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. Montreal QC, Canada: ACM, LBW005:1–LBW005:6. ISBN: 978-1-4503-5621-3. DOI: 10.1145/3170427.3188445

[1] The Italian term *Valletto* is a historic name for the assistant of a king / duke / count. It appears to be an appropriate name for a tool which unconditionally serves the user in the challenging process of visual analysis.

5.1 Introduction

The previous chapters provide insights on command structures and the use of visualizations in the field. Additionally, these chapters address tasks in which a system should adapt to the user. However, the question remains how these insights can be used to derive a proper design for a multimodal system. This chapter merges these different parts together. It shows how the information directly influence the design. The main design idea is to consider the entire visual analysis process through the lens of a conversational interface.

Generally, conversational interfaces are already part of the everyday life. Voice-based digital assistants (e.g., Apples Siri[2]) are supporting users by accelerating knowledge access. These assistants have an advantage over conventional interfaces in terms of interaction effort and user experience in ubiquitous and casual interaction scenarios (Cohen and Oviatt, 1995). A central aspect is the interactive conversation with the user consisting of "adjacency pairs" of call and response (Schegloff, 2007). The dialogue is the main communication channel. It fosters the engagement with the user (Moore et al., 2017). Furthermore, the conversation presumably improves the users reasoning process as well as could guide the user (Yankelovich et al., 1995).

Approaches of conversational interfaces for visual analysis essentially focus on the communication through visualizations in a stationary scenario (cf. Table 2.2). They adjust or generate visualizations according to a users utterances. The communication takes place on two one-way channels. In order to elaborate the idea of having an actual conversation, a persistent dialogue can likely help. This dialogue should facilitate the experience during the visual analysis as well as help the user to make sense of the visualized data. Apart from a lack of transparency of the available functionality, the behavior of conversational interfaces is often hard to interpret for a user (Chen and Wang, 2018). Increasing the interpretability of the system's behavior likely supports an intelligible visual analysis.

P1 data-agnostic: From a user point of view, the tool should be able to analyze the current data of interest. However, the importance of a data set is changing due to the required task and context of the user. From a business point of view, the data sets along the value chain differ with respect to the content, but not to the structure. Typical process-related data is multivariate and structured by a relational data base.

P2 supportive: The tool should effectively maintain the user's analysis flow. Sensing of intents as well as selectioning visualizations need to be in according with the user's current position in the exploration process. If P2 is not fulfilled, the performance of the user might

[2]https://www.apple.com/siri

be reduced due to incorrect predictions by the system. Furthermore, the user should be able to switch back and forth through the analysis.

P3 transparent: The user should be aware of what the system understood. Intelligent systems predict situations based on user interactions. In order to trust those systems, transparency and interpretability in on the computational process is an important aspect (cf. Section 2.6).

P4 intelligible: Complexity is a key factor to make things interesting for the user (Norman, 2010). Additionally, "every application has an inherent amount of irreducible complexity" (Saffer, 2006). However, some parts can be encapsulated without reducing the engagement of the user. In data analysis, the valuable part is on finding meaningful insight within the data. In contrast, the creation of appropriate visualizations as well as the technical execution of statistical test is presumably less interesting for many users. However, both techniques are essential to eventually find desired insights. Hence, purposefully encapsulating the given complexity should focus rather on automatically executing suitable techniques than on searching insights.

5.2 Related Work

This chapter's related work is two-fold. First, recommender systems for visualizations play a central role. However, Section 2.4 discusses primary work on automatically generating visual mappings. Second, Section 2.5 introduces relevant related concepts for natural language based interactions in visual analysis. Yet, there is additional related work on the use of natural language in similar domains such as machine learning.

Iris (Fast et al., 2018) and *Ava* (John et al., 2017) both consider machine learning and statistical testing as a domain. *Ava* (John et al., 2017) essentially enables a language-based interaction with Juypter notebook (Kluyver et al., 2016). Instead of writing code, a user can type what (s)he wants and *Ava* generates the corresponding code. *Iris* (Fast et al., 2018), on the other hand, is a dialogue-based user interfaces. The authors show an advantage of Iris over classical machine learning model construction with scikit-learn (Pedregosa et al., 2011). Hence, both system empirically address the advantages of language-based interactions in complex domains.

Furthermore, *Voder* (Srinivasan et al., 2019a) adds data facts to visualizations. Typically, a data fact addresses an insight in the data. Srinivasan et al. (2019a) describe a data fact as "any textual descriptions of data accompanying visualizations" (Srinivasan et al., 2019a). It explicitly adds additional information on the visualized data. The data fact is attached a the side of the visualization. Additionally, the visualization and the data facts are linked with

each other. A user can hover over the data fact and *Voder* highlights the corresponding part in the visualization.

Hearst and Tory (2019) integrate visualizations (bar and line charts) directly in a dialogue-based user interface. A user can ask questions on the data. As an answer, the user receives a dialogue act as well as a visualization. According to Hearst and Tory (2019), only 60% of the participants would like to see visualizations within the dialogue component.

Understanding command structures in visual analysis is essential for designing effective user interfaces. Tory and Setlur (2019) conduct a Wizard-of-Oz study to understand how an intent and context-related utterances have to be handled by a system. The authors identify two major challenges. First, users tend to underspecify their utterances. Consequently, an utterances lacks certain words, e.g., how a visualization should look like. Second, the users consider the context for their utterances. For instance, they refer to the already shown visualization. Approaching underspecified utterances, Setlur et al. (2019) investigate methods for inference. Underspecifications regarding analytical expressions and visualization types, the system handles both by considering the data attributes' scale of measurement. Furthermore, the system can also handle underspecification in both inter-sentences and intra-sentences, i.e., referring to elements from a previous utterances or what is already shown in the visualization.

Considering multimodal interfaces, Srinivasan et al. (2019b) explore the command structures as well. While originally investigated using a photo-editing tool, the authors propose to provide context-related examples for commands to the user. A framework produces the relevant examples considering how, when, and where an example command should be provided. As natural language interfaces naturally suffer discoverability of available functionality, the authors show that the system's feedback help to overcome these barriers. Furthermore, Saktheeswaran et al. (2020) show the benefits of using multimodal interfaces for visual analysis. The authors see the main advantage in "the complementary nature of speech and touch". Additionally, the participant welcome the opportunity to express their wishes in natural language while using touch for other actions.

In contrast to the related work, this chapter proposes a multimodal user interface design separating the dialogue from the visualization. Unlike Hearst and Tory (2019), the visualization is a central and permanent aspect in the user interface while the dialogue contains additional information on the data. This additional information addresses data facts similar to Srinivasan et al. (2019a). Furthermore, the design also incorporates an idea of increasing the interpretability of NLP.

5.3 Concept

Both Chapter 3 and Chapter 4 examine potentials for a user-specific intelligible multimodal visual analysis by both speech and touch. Considering these results, this section discusses a coherent design for the user interface.

5.3.1 Design & Interactions

The use of speech and touch as main modalities allows to think differently about the design for a multimodal visual analysis. Speech and touch are typically used in situations where interactions with keyboard and mouse are not handy. Especially in the context of mobile devices, both are the main modalities. Additionally, visual analysis is used beyond the desktop (Roberts et al., 2014), as a technology-supported workplace becomes mobile. Considering both factors, the design of this section approaches a mobile design.

However, a major challenge in designing for a mobile device is essentially the small display size compared to a high resolution desktop PC. A multi-window strategy, such as in Aurisano et al. (2016) and Sun et al. (2010), is not applicable due to the challenges of either readability or interaction. The readability suffers due to the down scaling of the visualizations. The interaction with the visualizations suffers when visualizations overlap (Sun et al., 2010). However, a single window is also not sufficient, because useful information probably will not fit on a mobile screen such as summaries (Srinivasan and Stasko, 2018), details (Srinivasan and Stasko, 2018), or an overview about the data set (Gao et al., 2015). Furthermore, a user should focus on the visual analysis itself. The visualizations should take a central position in the user interface design while supported through curated information.

In order to properly organize the information, the design implements a two-tab strategy for the visual analysis. In the first tab, detailed information about the data attributes are individually shown. A second tab focuses on the interactive analysis of the data. This strategy further supports the rudimentary information seeking mantra (Keim et al., 2006; Shneiderman, 1996).

The data tab serves as a start (see Figure 5.1). Assuming an established connection to a structured data source, all available attributes are shown in a scrollable list. When the user selects one attribute, its distribution is visualized through a histogram for a numeric attributes and a bar chart for a categorical attributes, respectively. Furthermore, related statistical measures are provided. For numerical attributes, corresponding measures describe the distribution (mean, standard deviation, quartiles, etc.). For categorical attributes, the measures focus on frequencies and unique values. The visualization predominately shows the data while the subjacent panel helps to better understand the data. For instance, it

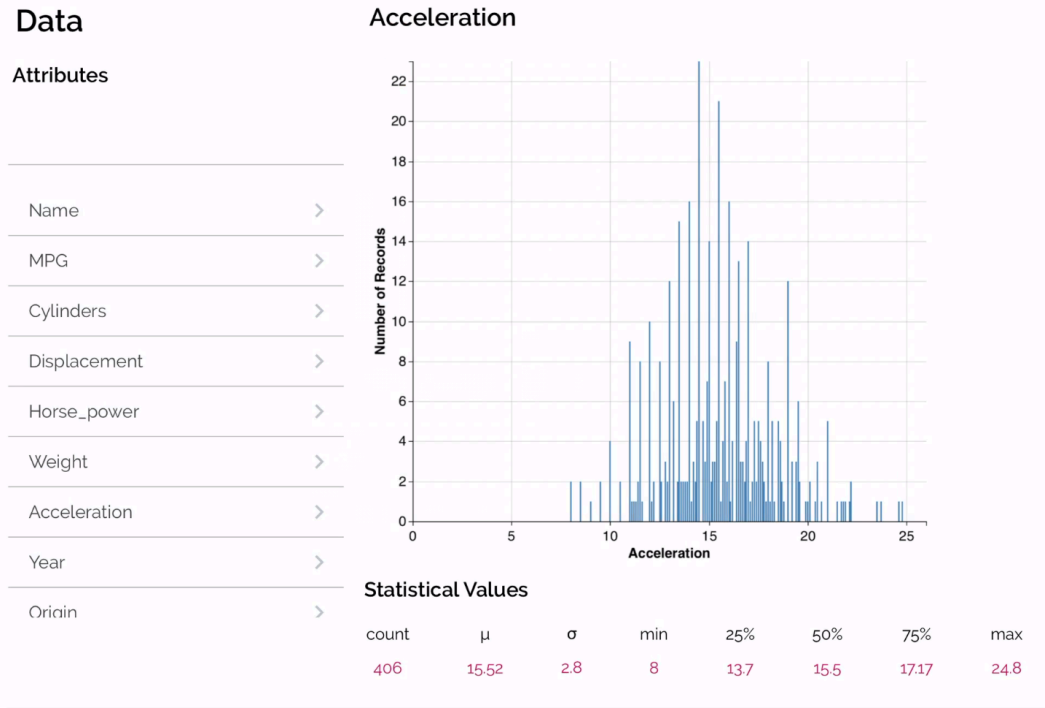


Fig. 5.1: User interface design of the data tab.

accelerates the identifications of how many different categories are shown in the visualization. This tab supports the first stage of a user's visual analysis. The focus is on examining existing data attributes and their corresponding statistical values. In terms of the *why-how-what* taxonomy (Brehmer and Munzner, 2013), this design essentially supports `identify`, `lookup`, `present`, and `select / filter`.

All actions related to two or more data attributes are supported by the analysis tab, shown in Figure 5.2. This tab essentially supports a user in the other tasks. Especially it supports the user in discovering and exploring data through encoding and manipulating visualizations. However, the various modalities have different effectiveness in supporting these elements (Badam et al., 2017). Speech is very effective for actions related to the underlying data of the analysis (Grammel et al., 2010), whereas visual mapping changes are effectively supported by gestures (Badam et al., 2017). This implies a mapping of the modalities on the actions in visual analysis where all data-related actions are supported by speech and all visualization-related actions are supported by gesture.

Using speech further enables a kind of conversation between the user and the system. Assuming a user talks to a system in order to generate visualizations, (s)he likely uses speech to alternate the data attributes. Hence, the system needs to understand changes of the underlying data as well when the user talks with it. However, in terms of the *why-how-what* taxonomy (Brehmer and Munzner, 2013), speech supports both `encode` and `filter`.

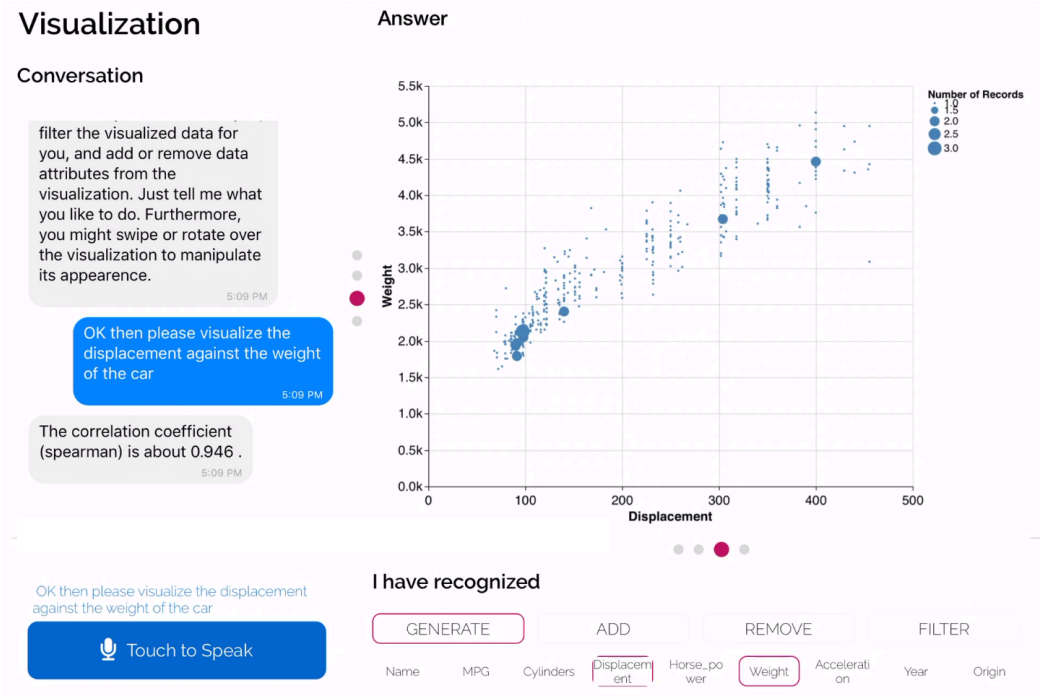


Fig. 5.2: User interface design of the analysis tab.

In order to further foster the idea of having a conversation between the user and the system, the analysis tab contains a back and forth dialogue. This dialogue contains both the user's utterance (a live preview of the utterance recording while speaking is shown above the speech button) and a situation-dependent textual answer of the system. Section 5.3.3 discusses these situation-dependent textual answers in detail. However, they should increase the intelligibility of the visual analysis, as they address hidden challenges and provide additional information.

Additionally, the dialogue contains all answers and requests. Hence, it empowers a user to constantly track the progress. In order to go fast back to a previous analysis situations, a user can directly touch on the corresponding request in the dialogue. This interaction changes the visualization to the last shown visualization for this request. Hence, the system supports the user in `browse` and `lookup`, as in both actions the user knows the location in the data.

However, direct text input via typing is not supported by the system for two reasons. First, using a visual keyboard on a mobile device cost space in the moment of the request. By this, the visualization is partly covered by the keyboard which should be avoided (Sadana and Stasko, 2016). Without the keyboard, the user can talk to the system, while constantly observing the visualization. Hence, a user does not get disturbed in the analysis. Second, Ruan et al. (2018) show that speech input is faster than text input.

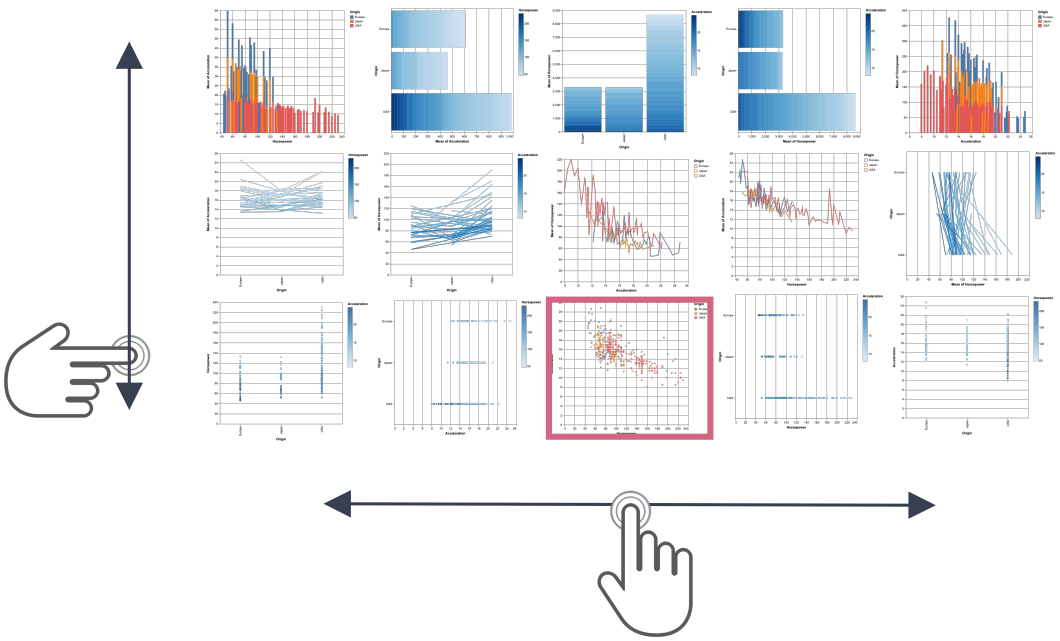


Fig. 5.3: Arrangement of the visualization space in the navigation panel. The red square marks the visualization first shown to the user.

As modern touch displays have multitouch, gestures enable visual adjustment located directly on the visualization. The manipulation of the visualization works as follow. First, a user can use one-finger swipe gestures for changing the visualization shown (`change`) in terms of both visualization type and visual mapping (see Figure 5.3). Sadana and Stasko (2016) recommend them as easily executable gestures in the visualization context. In the vertical swipe gestures change the visualization type, while horizontal swipe gestures change the visual mapping but keep the visualization type. This kind of navigation in a visualization space has been further investigated in Kassel and Rohs (2017). The metaphor of picture sliders next to the visualization indicates how many different visualizations are available for the currently visualized data attributes. This metaphor should be known by almost every user who uses a mobile device frequently. Second, two-finger gestures can be used for zooming (`navigate`) as well as for rotating the visualization (`arrange`). However, how this visualizations are generated are discussed in Section 5.3.2.

Considering the *how* elements of the *why-how-what* taxonomy, speech supports the encoding (`encode`) of data as well as data-related elements (`filter`), where gestures should be used for changing the encoding (`arrange` and `change`).

In order to increase the transparency of the computations in the background (Sinha and Swearingen, 2002), a reasoning panel (see Figure 5.2) is included. This panel highlights the systems interpretation regarding the user's last utterance by adjusting the colors (see Section 5.3.4). For each text field, its color turns reddish if the system beliefs that it is part of the user's utterance. This is an approach for increasing the user's trust in the system (Setlur et al.,

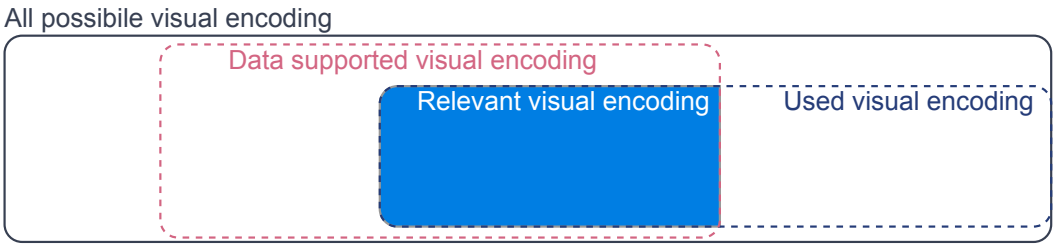


Fig. 5.4: The set of relevant visualizations is determined by the set of used visualizations and the set of visualizations supported by the data set.

2016). Additionally, it should empower the user to develop alternative interaction strategies in case a request fails. Essentially, this reasoning panel addresses the gulf of evaluation small (Norman, 2002).

Generally, the design of both tabs follows a consistent visual design (Sadana and Stasko, 2016) via a FLUID interface (Drucker et al., 2013). The design essential consists of a dominant visualization, a guidance element (either scrollable list or a dialogue) on the left, and a panel of supplementary information (either statistical measure or a reasoning panel) below the visualization. Hence, the user has likely fewer orientation problems when switching between the tabs back and forth. Furthermore, the underlying analysis state model is consistent among the tabs, i.e., applied interactions such as filters and selection in the analysis tab hold in the overview tab too.

5.3.2 Visualization Recommendation

Showing visualizations for a set of data attributes is the key element of any visual analysis systems. Usually, a user creates new visualizations in conventional visual analysis systems step by step. Since speech appears to be very effective for generating visualizations (Badam et al., 2017; Grammel et al., 2010), a user now needs to only specify what (s)he wants to see. Hence, the system needs to recommend a visualization for a set of data attributes.

Essentially, the challenge of recommending visualization is a two-step approach. In the first step, a system needs to generate a set of visualization candidates. In a second step, the systems ranks all candidates from the set according to some objective function. This objective function could be for instance the effectiveness of a visualization based on the studies shown in Section 2.2.

Considering all visual variables (Bertin, 1974), the set of all corresponding visualizations V_{all} is huge. Ranking all of these visualization would be time consuming. Hence, the objective is to reduce the search space. In order to effectively prune the set of candidates, Valletto approaches the problem from two perspectives. First, it considers those visualizations which can be used for the underlying data set of the visual analysis V_{data}. For instance, a map makes

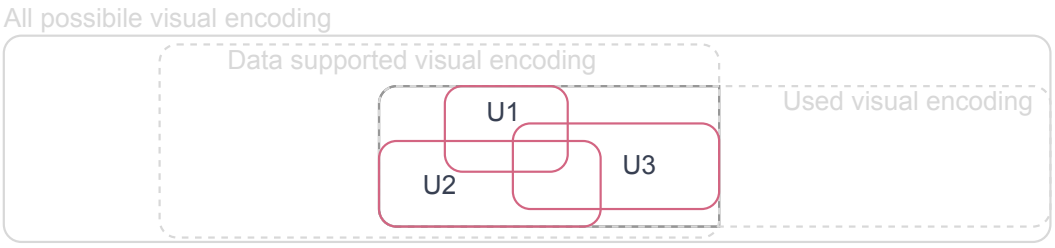


Fig. 5.5: Given the set of all relevant visualizations, each request from the user further reduces the set of currently relevant visual encoding.

no sense when no data attribute contains information about geolocation e.g., cities. Second, it further considers the set of visualization used in the field V_{used} by taking advantage of the knowledge about the use of visualizations from Section 4.4. This knowledge additionally reduces the set of candidates, since only the visualization types of bar, line, and scatter are essentially used. Figure 5.4 illustrates the pruning of the visualization space. While both V_{data} and V_{used} are each a subset of V_{all}, $V_{rel} = V_{data} \cap V_{used}$ eventually describes the set of candidates for the ranking.

However, not all visualizations from V_{rel} have to be ranked. In fact, only those visualizations related to a user's utterance need to be considered for the ranking. V_{uttr} describes this set of visualizations (see Figure 5.5). Valletto generates V_{uttr} by creating all mappings of the mentioned data attributes from the user's utterance to the supported visual variables. Each visualization from V_{uttr} is further ranked by considering Mackinlay (1986).

As Valletto's user interface design provides space for only one visualization at a time (cf. Section 5.3.1), the ranking needs to align with the implemented navigation opportunities for the user. Figure 5.6 illustrates the alignment of the ranking with the navigation opportunities. Initially, the most effective visualization is shown to the user. Other visualizations are ordered in the background by their ranking. Consequently, if a user swipes either to the right or to the left in order to change the visualization, the next shown visualizations represented the second most effective visualizations for the corresponding visualization type. If a user changes the visualization type by swiping either up or down, the next shown visualization represents the most effective visual mapping for the corresponding visualization type. This method not only provides the user a direct access to the most effective visualization, but also offers options to effortlessly change the visualization.

5.3.3 Dialogue Design

While the previous sections discuss the overall design of the concept including the dialogue component, this section discusses the dialogue design in visual analysis. Generally, the dialogue is the central aspect of conversational interfaces (Cohen and Oviatt, 1995). The purpose of the dialogue is to help a user in fulfilling the desired tasks by controlling the

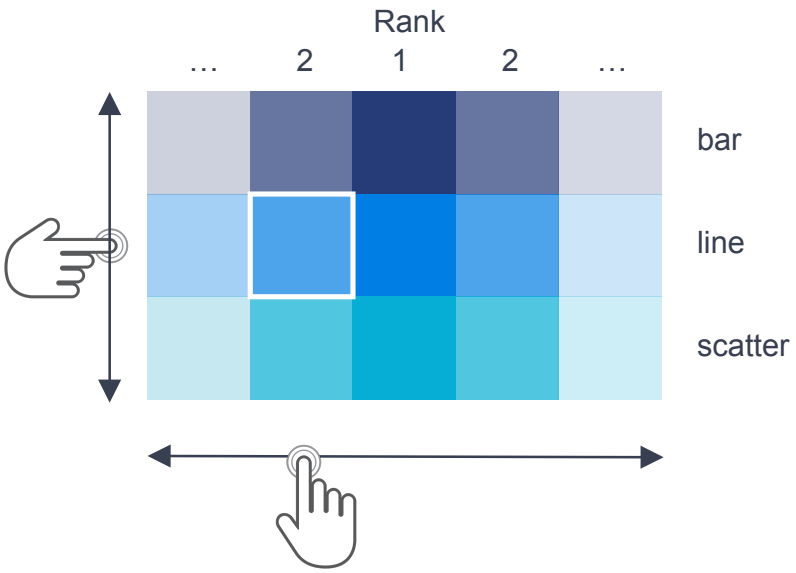


Fig. 5.6: Illustrating the alignment of the ranking with the navigation options. In this example, the user navigated to the second ranked line chart.

information flow as well as by giving guidance if needed (Yankelovich et al., 1995). The dialogue design is one of the biggest challenges as it directly influences the user experience (Moore et al., 2017).

A dialogue essentially consists of sequences of dialogue acts. Each sequence follows a certain objective e.g., a user wants to analyze the relationship between the fuel consumption and the car model. All utterances related to this objective belong to the same dialogue sequence. Furthermore, a sequence is structured through "adjacency pairs" (Schegloff, 2007). An "adjacency pairs" (Schegloff, 2007) summarizes two dialogue acts which refer on each other. For instance:

User: *"Hello System, how can you help me in my visual analysis"'*

System: *"Hello User, I can generate visualizations for you as well as provide additional information which might help you to make better sense of the data."*

Considering the dialogue in visual analysis, the user likely treats the system as a "virtual butler" (Payr, 2013). Additionally, the results of Section 4.3 reveal potential structure of utterances in visual analysis. A conversation likely follows goal-oriented scenario (Shechtman and Horowitz, 2003). Hence, a user essentially wants to be rather supported in the objective than to have a decent conversation with the system.

Based on the previous sections, a system needs to answer on dialogue acts for encoding data as well as filtering data. However, utterances for encoding data also refer to changes of the visualized data attributes. Depending on the user's intent, the system should answer

differently. Considering Brehmer and Munzner (2013) and Section 3.3.3, a user could be supported in discovering useful insights in the data as well as exploring the data. Discover further refers to compare, since e.g., comparing two categories which each other likely raises a wish to clearly know the exact circumstances. Generally, a visualization abstracts from the underlying data. Depending on the particular visual mapping (cf. Section 2.2) as well as the level of abstraction through aggregation, a visualization could lack a detailed view on the data. Furthermore, visualizations could be used to divert the user from certain circumstances, also referred to as the "lie factor" (Tufte, 1986). Hence, a dialogue act should tell the user something about the data what is not already highlighted in visualization. Otherwise, a system would not add information for the user. In terms of discover and explore, this means to provide the user direct access to the output of statistical measures.

Additionally, a system should further provide dialogue acts for maintaining the dialogue flow itself. This comprises the support when a user needs clarity about the supported functionality as well as a dialogue acts when the system is uncertain about the user's intent. The following discusses the different dialogue acts of Valletto as well as express how they could be integrated into a user's analysis flow.

Dependencies

Combining attributes is particular relevant under the objective of finding certain relationships within the data (Roth and Mattis, 1990). In other words, a system should support the user in identifying dependencies between data attributes. Depending on the data attributes' characteristics, however, the method for discovering the dependency changes. The major determining factor is the scale of measurement.

Given two quantitative data attributes, a typical method is to compute the correlation coefficient between the data attributes. This correlation coefficient describes essentially two things. First, it decides whether a relationship exists based on the p-value. Second, it further describes whether the existing relationship is positive or negative. However, various correlation coefficient measures exist. They basically differ in their assumption of the underlying distribution of data. The widely used Pearson's coefficient assumes normally distributed data. In practice, however, this assumption likely does not hold. A reason is that data from processes and other sources often follow other distributions e.g., the number of visits in a workshop follows a Poisson distribution. Additionally, the data is often discrete instead of continues. Hence, Pearson appears to be less useful in an automated system where the user should only plugin the data.

An alternative method is the Spearman's correlation coefficient (ρ). It does not assume any specific distribution of the data. It computes the correlation between the rank variables

of each data attribute. However, it only describes that two data attributes have certain relationship, but it cannot determine the structure of this relationship unlike the Pearson. A corresponding adjacency pair looks like:

System: *"Your data attributes have a correlation coefficient of: <Number>."*

If two categorical attributes are investigated, a method to determine the relationship is the Chi-squared (χ^2) test. It provides insights whether two data attributes significantly influence each other. If the test is positive, a certain relationship is likely give. However, as the Chi-squared test is a statistical method and uses the p-value as the main indicator, a chance for a wrong conclusion exists:

System: *"<String> and <String> are not independent"*

In case both data attributes are of a different type, still the relationship can be determined. In this case, the Mutual Information is computed.

$$I(X,Y) = H(X) + H(Y) - H(X|Y) \tag{5.1}$$

The mutual information consists essentially of the Entropy $H(X)$ and the Conditional Entropy $H(X|Y)$ of the data attributes.

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{5.2}$$

$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(y)} \tag{5.3}$$

Generally, the Entropy describes the information content of a random variables. In case all values of a random variable are the same, the information content is 0. The information content is maximal if and only if the random variable is unified distributed. Now, the mutual information of two data attributes is 0 if and only if the attribute combinations from both data attributes are always the same and both attributes follow the same distribution. A corresponding adjacency pair looks like:

System: *"The attributes <String> and <String> are sharing information."*

Comparison

In addition to identifying potential relationships and discovering new insights, users are likely to encounter common statistical errors (Zuur et al., 2010). Especially when a user compares multiple categories or groups which each other. This could be due to clusters,

categories, or similar. The primary objective is often to see how the different groups differ in regarding a specific measure. This comparison can be made visually by observing and interacting with visualizations. Additionally, statistical test might be applied for determine the differences.

Zgraggen et al. (2018) show that one major issue in statistical analysis also holds for visual analysis, namely the multiple comparison problem. The authors identify that user tend to draw wrong conclusions when combining multiple filters. While bar charts alternated by a standard deviation help an experienced user to draw correct conclusion, a novice user would likely fail. A personalized interactive system for visual analysis must find a solution for this challenge, since all user should be prevented from making mistakes. A corresponding adjacency pair looks like:

System: *"In this case, be careful with your conclusions. There might be no significant differences, although the visualization may to imply this."*

Exploration

For exploring data, a user might stuck in the exploration. By increasing the dimensions of the data set in terms of data attributes, it becomes more difficult for the user to effectively explore a data set. A user might wonder which attributes are worth it to combine, given the current analysis step. Hence, a system could help the user in highlighting useful avenues for data attribute combination. As the data attributes differ in the scale of measurement, the mutual information likely helps again. Given that higher mutual information implies a stronger relationship between the attributes, a system might provide a list of attributes with strong relationships to the user. A corresponding adjacency pair looks like:

System: *"<String>, ... likely explain <String> better."*

Filtering

In addition to understanding the inherent relationships within the data, being aware of how many data is actually visualized is relevant as well. By filtering certain aspects of the visualized data, the amount of shown data can be tremendously change. For deriving substantial conclusions, however, the user should always know the percentage of actually visualized data. A corresponding adjacency pair looks like:

System: *"You excluded <Number> % of your data."*

Greeting

In order to directly engage with the user right from the beginning, the system greets the user and introduces itself. Since the conversation in visual data analysis is likely goal-oriented (Shechtman and Horowitz, 2003), the dialogue opening is similar to a telephone call opening where the callee (in this case the system) always initiates the speaking (Schegloff, 1968). A corresponding adjacency pair looks like:

System: *"Hi <String>. My name is Valletto and I'm gonna help you to visually analyze your data"*

Unknown Command

Due to using speech as central modality, there is room for misunderstanding a user on multiple levels. On the lowest level, a speech recognition engine might not be able to correctly parse the users utterance. On a higher level, a user's utterance might be to complex formulated for the system. In any case, the system design should be able to handle such situations (Moore et al., 2016). One approach is to apologize and ask the user for rephrasing the last utterance. A corresponding adjacency pair looks like:

System: *"Can you please repeat your request?"*

Help

A user's lack of knowledge about the available functionality is another reason why conversational interfaces fail. Hence, a user might just ask for help when (s)he does not know what the system can actually do (Cox et al., 2001). A corresponding adjacency pair looks like:

User: *"Hey Valletto, what can you do for me?"'*

System: *"Good question. You can ask me to: Visualize the data attributes you want to analyze, filter the visualized data for you, and add or remove data attributes to or from the visualization. Just tell me what you would like to do. Furthermore, you can swipe or rotate across the visualization to manipulate its appearance."*

Confirmation

Although the above described dialogue acts already cover a variety of different analysis situations and provide corresponding answers, not all dialogue situations can be fully covered. In order to be able to give always a reaction to any user's utterance, however, Valletto has a confirmatory dialogue act. In this act, Valletto answers with a short answer and questions whether the user is satisfied with the reaction. A corresponding adjacency pair looks like:

System: *"Is this what you wanted?"*

5.3.4 Natural Language Understanding

The modality speech supports generating new visualizations, adding or removing data attributes to and from a visualization, and filtering on data attribute values. These four functions define the classifiable intents for a Natural Language Understanding (NLU) routine. However, predicting a user's intent is challenging. Interpreting a user's utterances underlies a given uncertainty due to speech recognition errors (Young et al., 2013). Additionally, it depends on the state of the analysis.

The intent classification underlies two constraints. First, the transparency panel of the user interface requires a value greater or equal to 0 for each intent as well as for each data attribute (see Figure 5.2). Therefore, the output of the classification has to be probabilistic. Second, the amount of example requests is too little for applying a end-to-end learning approach. Hence, a classical approach appears to be appropriate. The following algorithm predicts the intents within three steps.

The first step cleans and structures the user's utterance. Text data is unstructured, especially when the data source is the spoken word. In almost every NLP approach, a sequence of tokenization, and lemmatization is applied. Tokenization splits the text into its part, i.e., words, punctuation etc., are now separated. For each identified word, lemmatization maps this word on its bases, e.g., "visualizing" will be transformed into "visualize" by using, e.g., lexicon.

In addition to the structuring of the utterance, the syntax of the utterance is parsed (see Figure 5.7). A dependency parser finds the relationships between words and their grammatical purpose within the utterance. The root element plays an important part in the algorithm for identifying the user's intent.

The root element of a sentence typically represents the main verb. As conversations in visual analysis are arguably goal-oriented (Shechtman and Horowitz, 2003), the main

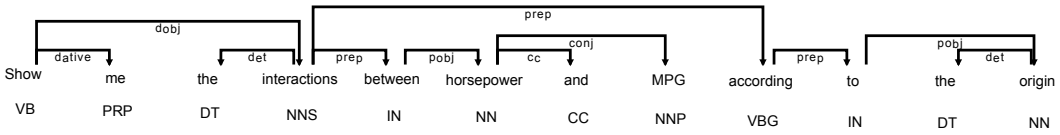


Fig. 5.7: Dependency tree for an exemplary utterances on the relationship between horsepower, miles per gallon, and the origin.

verb potentially serves as a primary input for the intent classification. This hypothesis is further supported by the empirical results of Section 4.3. Consequently, the algorithm computes the semantic similarity between the root element r and synonyms of the intents (see Figure 5.8). For each supported intent, a list of synonyms S_I is manually created prior the system's initialization. Eventually, the algorithm averages the sum of the computed semantic similarities:

$$s(r) = \frac{1}{|S_I|} \sum_{s \in S_I} sim(r, s) \tag{5.4}$$

However, a variety of semantic similarity measures exists. Patwardhan and Pedersen (2006) investigate the correlation between these different semantic similarity measures and the human's concept of semantic similarity. The authors show that Jiang and Conraths measure represents the human understanding of semantic similarity between words quite well. Therefore, the algorithm uses Jiang and Conrath as well. As a result, a vector v exists which stores the average semantic similarity score for each intent.

A similar approach is applied for finding the relevant data attributes within the utterance. Assuming each data attribute has a reasonable name – which is not necessarily fulfilled in practice – the name likely refers to a noun. As the dependency parses further returns all nouns of the utterance, the algorithm computes the semantic similarity between all nouns and all data attribute synonyms. However, P1 requires the algorithm to generate the lists of synonyms for the data attributes on the fly. All synonyms are collected during the initialization of the system. In contrast to the decision on the intent, the algorithm does not compute the average semantic similarity for each data attribute, but searches for the best matching pair.

Depending on the identified attributes and filters, the algorithms further refines v. For example, if no attributes are discovered, but a filter should be applied, then the filter intent is rated higher. This might happen in case only certain attribute values are mentioned. Finally, the probability vector $\vec{v'}$ is computed by normalizing v. The reasoning panels represents the values of $\vec{v'}$. The element with the highest probability is lastly classified as the visualization intent.

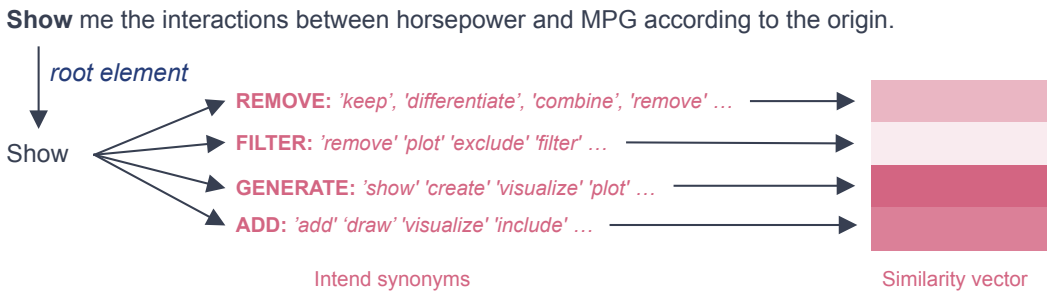


Fig. 5.8: Similarity computation between the root element and the synonyms for the intents. In this illustration, the color saturation of the similarity vector encodes the values.

5.3.5 Usage Example

Figure 5.9 illustrates an exemplary usage of Valletto. Imagine a user wants to explore a car data set (Donoho and Ramos, 1982; StatLib, 2005). This data set contains information about different car models from the late 90th. The user starts in the analysis tab of Valletto (cf. Figure 5.2) by exploring the dependency between acceleration and the horsepower of the cars. Initially, Valletto generates a scatter plot, as it is the effective visualization for two quantitative attributes. Additionally, Valletto tells the user that those data attribute are negatively correlated. However, the user swipes up for changing the visualization type to a line chart. Furthermore, the user rotates the visualization.

In order to further analyze the individual data attributes, the user switches to the data tab (cf. Figure 5.1). In this tab, the user investigates the distributions of the data attributes. While acceleration follows a normal distribution, horsepower does not. Consequently, the user also explores the origins of the cars.

In order to find out how a car's acceleration depends on the country where the car has been build, the user formulates a corresponding request. Valletto generates a bar chart accordingly. Since this visualization likely suffers under the multiple comparison problem in this case, Valletto provides an additional dialogue act in order to warn the user. As the bar chart shows aggregated data, the user swipes up to see the raw data through a scatter (strip) plot.

5.4 Technical Implementation

The technical prototype implements a client-server-architecture. Since Valletto supports a diverse set of devices, the entire "intelligence" of the system is realized in the backend. The client only takes care of the user interaction handling and the representation of the information through both dialogue and visualization rendering. This architecture offers flexibility in order to handle multiple devices and additionally provides enough computational power for the predictive models.

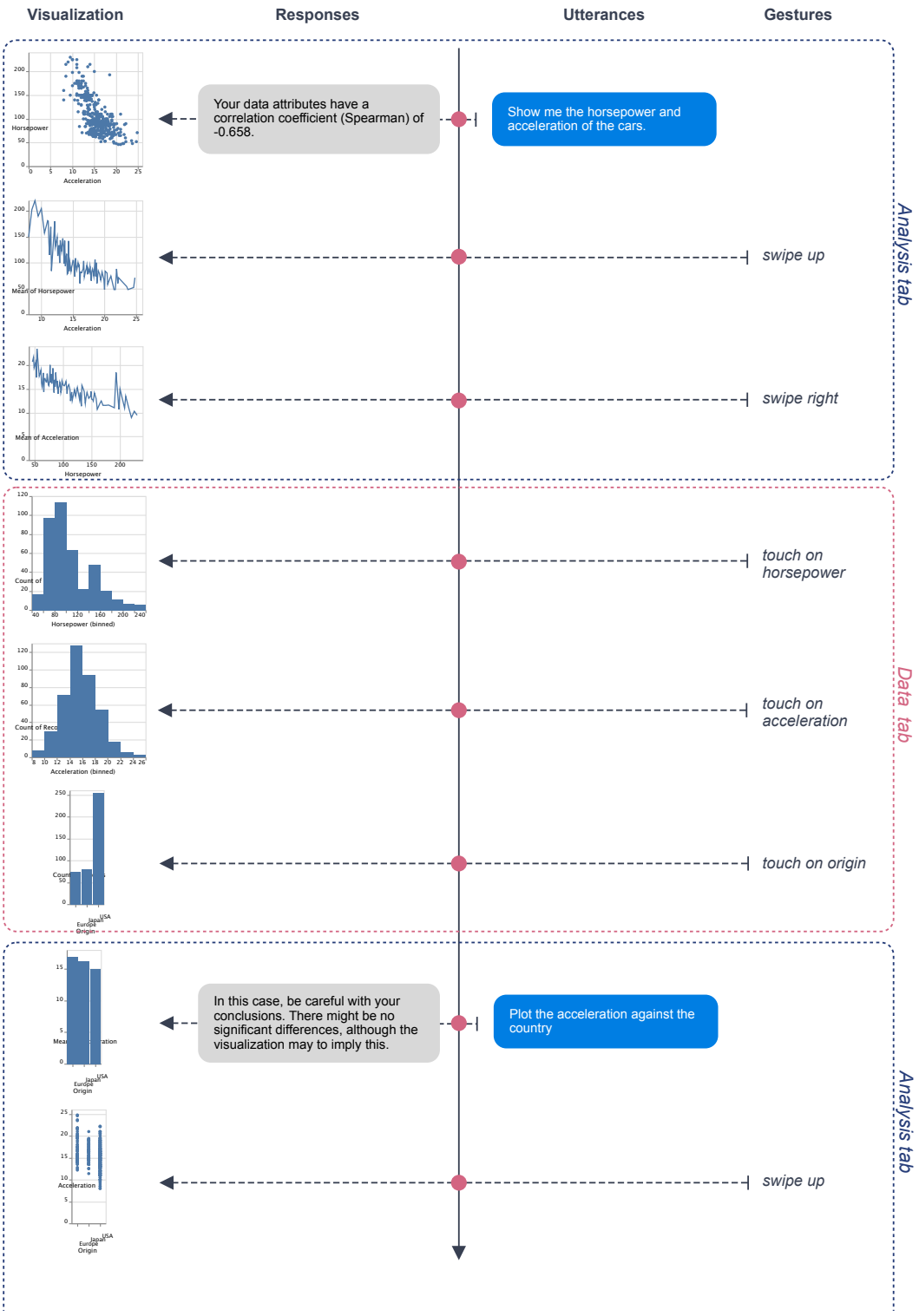


Fig. 5.9: Exemplary usage of Valletto.

The client is implemented in React (Facebook Inc, 2013). React is a JavaScript library for creating WebApps. A object-oriented programming paradigm structures a React App into components. However, it strictly separates between dump and smart components. While dump components only focus on representation of forwarded information, the smart components handle the communication, and necessary computations. Additionally, the concept of React uses states and propositions. The propositions are used for forwarding information.

On the other hand, the backend implements a Representational State Transfer (REST) service through Flask (Dwyer et al., 2017). The backend infers on the user's utterance in order to predict the user's intent. All intent-related computations base on Spacy (Explosion AI, 2018), a Python library for NLP. Spacy implements by default all needed algorithms for the intent prediction (cf. Section 5.3.4).

For generating visualizations, the visualization grammar Vega-lite (Satyanarayan et al., 2017) is used. Vega-lite is designed for automatic visualization generation. However, Altair (VanderPlas et al., 2018) is used for generating the visualizations, as the backend is implemented in Python. Altair provides a Python interface for generating Vega-lite objects.

5.5 Experiment 1: Exploring the Design

The objective of this experiment is to explore potential issues in the design of the user interface as well as the integration of the interactions. Therefore, it is designed as an expert review.

5.5.1 Procedure

In order to not only find potential issues, but also classify them, a hybrid expert review design is selected. This approach consists of a combination of Cognitive Walkthrough (CW) with a Heuristic Evaluation (HE). While the CW directs the expert through stereotypical interaction sequences, the HE provides a format for classifying potentially identified issues.

For the CW a set of tasks and the corresponding interaction sequences is defined. This set covers Valletto's supported features. Additionally, the tasks are meant to be as realistic as possible in order to simulate a typical visual analysis process. An expert should answer four questions on the integration of the interactions after each interaction with the system, according to Wharton et al. (1994). The CW reveals usability issues immediately.

The HE requires a previously provided heuristic. However, a huge variety of heuristics exists (Nielsen, 2005). For analyzing dialogue-based user interfaces, Weinschenk and Barker (2000) constructed a 20-item long heuristic which is used for this experiment. The HE allows to find usability issues beyond the pure interaction with the system. Finally, on all potentially identified issues a 5-point severity rating has been performed.

The experiment is conducted as a 1-on-1 situation with a participant with no time constraints in a quiet room. Each participant get equipped with the technical prototype running on a 10 inch iPad as well as a sheet of paper for each the predefined tasks, and the heuristic.

5.5.2 Participants

Four experts are recruited for this experiment. They have a background in either computer science or engineering. They have 4 years of working experience on average in both industry and academia as software engineers and HCI researchers, respectively. Due to their work as software developers for both Web and mobile, they are aware of the importance of usability as well as have passing knowledge in multimodality.

5.5.3 Results

In summary, the experts discover 14 potential usability issues of the current implementation of Valletto. However, the majority of these potential issues are mostly voted as minor usability problems according to the severity rating scale.

One central usability issue is the lack of transparency on the available functionality of Valletto. The experts mentioned that a user likely does not know what kind of functions are supported by for both speech and gestures. These facts imply additional effort for the user to explore the individual modalities before the visual analysis can start. One expert suggests to add an initial guided tour for the user. In this guide, the system would introduce its functionality on its own. Another expert proposes to have a redundant implementation of the functionality. Consequently, both modalities support similar functions.

Furthermore, some experts identify the reasoning panel design as a source for potential usability issues. A user might assume that the text fields are buttons to click on, although they are not. Apparently, the differently colored frames around the text fields produce this assumption. However, the experts also see potential benefits of using speech as the primer modality for visual analysis. They explicitly refer to scenarios in which the user is in a shaky environment e.g., sitting in the back of a car.

5.5.4 Implications on the Design

The results reveal not only unknown usability issues for both user interface design and interaction design, but also address some suggestions on how to cover them.

On the one hand, the reasoning panel design has to be improved. As it creates an impression of being touchable, the design needs to change. As the reasoning panel visually represents the system's interpretation of the user's recent utterance, a design interaction needs to conserve this idea. A potential solution might be to remove the frames as they create the assumption of a button. In order to still visualize the probabilities, the text fields could be colored instead.

On the other hand, the lack of transparency of available functionality is addressed. Generally, this is a known issue of FLUID user interfaces (Drucker et al., 2013). In classical Window, Icon, Menu, and Pointer (WIMP) user interface, every function is represented by a visual element (e.g., buttons). Given the fact that users are still highly trained to use WIMP instead of FLUID interfaces, this usability issue is surely relevant for the design improvements of Valletto. In order to cover this usability issue, a guided tour is implemented as suggested by an expert. This guided tour is realized through overlays on the individual tabs in the moment of the first opening. Finally, smaller addressed usability issues are fixed as well. This updated version of the Valletto serves now as technical prototype for the next experiment.

5.6 Experiment 2: Decisions and Obstacles

The objective of this experiment is to evaluate the performance of Valletto in supporting a user in visual analysis as well as to find potential interaction obstacles. Therefore, a within-subject user study is conducted. The experiment focuses on the research questions:

RQ 4: What are the differences in completing tasks in a conversational interface compared to a conventional user interface?

RQ 5: What are the differences in the interaction strategies between a conversational interface and a conventional user interface?

The experiment consists of three phases: an initial introduction to the systems, an interactive phase on executing various visual analysis tasks, and a questionnaire. Similar to the evaluation of Eviza (Gao et al., 2015), Valletto is evaluated against Tableau[3]. Currently, Tableau is one

[3] `https://www.tableau.com`

of the common tools for visualizing data via classical drag and drop interactions. Its features in terms of supported visualizations and analytics are a superset of Vallettos features.

5.6.1 Procedure

Prior to the start of the experiment, both systems are prepared with an e-commerce data set. This multivariate data set consists of quantitative, ordinal, and categorical data attributes. Furthermore, each participant gets a standardized introduction to each system. During this introduction, a participant has time to get familiar with the particular user interface.

The experiment's interactive phase consists of 11 visual analysis tasks (see Table 5.1). 10 of these 11 tasks are about verifying a given statement on the data. Each participant should check whether the given statement is true. In case a participant is uncertain whether the statement is true or false, they can also state "undecided" as an answer, without specifying any reasons. In order to be able to check on the statements correctness, a participant has to first decide on the required data attributes and second create visualizations.

Since no additional constraints are given, a participant can take advantage of the entire functionality of each system. The tasks vary in complexity. Furthermore, the statements are formulated in way that a participant cannot simple read out lout the statement and Valletto immediately delivers the answer. For the verification-oriented tasks, the time needed to complete the task is manually measured. Precisely, the *task completion time* is the time between a participant's first interaction after reading the corresponding statement and the moment of the participant's decision.

In addition to these verification-oriented tasks, a 5 min long open exploration task exists. In visual analysis research, an open exploration task serves well to analyze interaction behaviour with the system as well as potentially delivers insights on the user's analysis flow. Setlur et al. (2016) use a similar approach for evaluating Eviza. Each participant should mention as many as possible identified facts about the data. After the study, these statements are checked on correctness.

In order to reduce the biases in the data, the starting system is alternating as well as each participant gets a randomly assigned sequence of the 10 verification-oriented tasks. The open exploration task is in any case the last task for each participant.

After these interactive phase, each participant answers a questionnaire. It contains questions on the perceived user experience, feedback to the system in general, and demographic information. Similar as the first experiment (cf. Section 5.5.1), this experiment is conducted

Task ID	Data	Situation	Objective
1	C x Q	LOCATE	Verification
2	Q x Q	DEPENDENCIES-numeric	Verification
3	C x Q	COMPARISON	Verification
4	C x Q	LOCATE	Verification
5	C x Q	COMPARISON	Verification
6	C x Q	COMPARISON	Verification
7	Q x Q	DEPENDENCIES-numeric	Verification
8	C x Q	COMPARISON	Verification
9	C x Q	LOCATE	Verification
10	C x Q	COMPARISON	Verification
11	-	EXPLORE the Market	Exploration

Tab. 5.1: List of given tasks to the participants for the within-subject study.

as 1-on-1 set up in a quiet room. Overall, the experiment is designed to take approximately 75 min.

5.6.2 Participants

13 persons participated in this experiment – a comparable number of participants as in both DataTone (Gao et al., 2015) and Eviza (Setlur et al., 2016) – where one person was a native English speaker. The participants are between 24 and 40 years old ($\mu = 30$, $\sigma = 5$) and have several years of working experience in industry ($\mu = 4.2$, $\sigma = 3.7$). From an educational point of view, the participants' academic background is in computer science, nature science, business, and engineering.

In terms of visualization experience, the participants stated to generally have moderate experience. For creating visualizations, they rely on either MS Excel or scripting languages like R or Python by using dedicated visualization libraries. The majority (8 participants) further states to have no experience in using Tableau.

5.6.3 Results

The results empirically highlight differences in the decisions made in the systems as well as address obstacles related to interactions. Overall, Valletto receives a System Usability Score (SUS) (Brooke, 1996) score of $\mu = 81.1$ ($\sigma = 8.6$) and the participants rate the usability between "good" and "excellent".

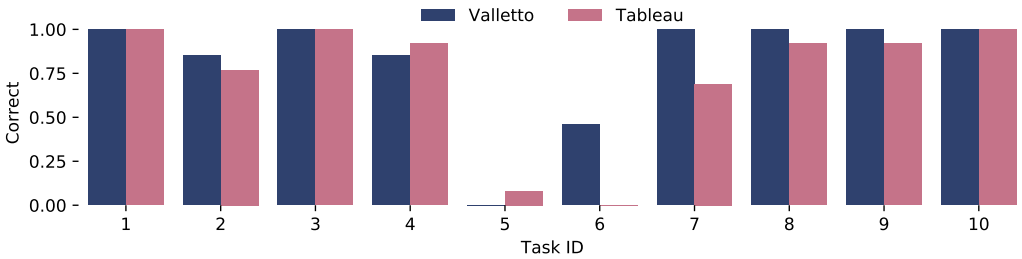


Fig. 5.10: Amount of correct decisions made by the participants per system and task.

Decision Quality

In the following, the term *decision quality* of a participant describes the number of correctly made decisions. Overall, a conducted paired samples t-test reveals a statistically significant ($t(12) = 2.99, p = 0.01$) higher decision quality with Valletto ($\mu = 80, \sigma = 7.5$) than with Tableau ($\mu = 73, \sigma = 7.5$) with a medium effect (pooled Cohen's $d = 0.86$). Additionally, Figure 5.10) illustrates the differences for each task. While using Tableau, the participants made better decisions in the tasks 4 and 5. However, they performed better in the tasks 2, 6, 7, 8, and 9 while using Valletto. A conducted χ^2 test further shows statistically significant differences in the decisions in task 6 ($\chi^2(2, 26) = 14.36, p < .001$) and 7 ($\chi^2(2, 26) = 4.72, p < .01$).

Task Completion Time

The task completion time of the participants is statistically significant different in the tasks IDs 1, 2, 6, 7, and 10 between Tableau and Valletto, as illustrated by Figure 5.11. Table 5.2 further summarizes the corresponding statistical values for each task. These values reveal a large effect (Task IDs 1, 2, 7, and 10) and medium effect (Task ID 6), respectively. Furthermore, participants decide faster with Valletto ($\mu = 444, \sigma = 93$) than with Tableau ($\mu = 609, \sigma = 123$), according to a conducted paired samples t-test on the accumulated task completion time over all tasks ($t(12) = -5.75, p < 0.01$). This difference is additionally supported by a large effect (Cohen's $d = 1.6$).

Considering the task completion time by the participants' self-reported experience level, Figure 5.12 shows no differences between these experience levels. However, some participants mention that it subjectively felt longer to complete a task with Valletto than with Tableau.

Finally, Figure 5.13 highlights the learning effect considering the task completion time by the relative task position. According to a paired samples t-test ($t(12) = 2.28, p < .05$), the

Task ID	$\mu_{Valletto}$	$\mu_{Tableau}$	t-statistic	p	d
1	**26.61**	50.38	-2.80	**.015**	.81
2	**41.38**	77.46	-3.16	**.008**	.91
3	45.69	**41.53**	0.39	.699	-
4	46.00	**42.69**	0.39	.702	-
5	**48.92**	53.38	-0.92	.373	-
6	**43.92**	60.23	-2.20	**.047**	.63
7	**39.84**	72.69	-2.98	**.011**	.86
8	**37.33**	47.58	-1.86	.088	-
9	56.00	**54.46**	0.17	.865	-
10	**62.07**	109.30	-3.11	**.008**	.90

Tab. 5.2: For each task, the results of the conducted paired samples t-tests: means (in seconds), t-statistic, p-values, and effect sizes (Cohen's d pooled).

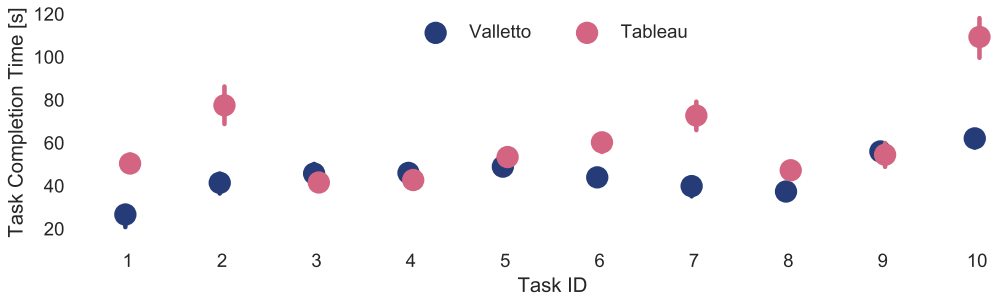


Fig. 5.11: Participants task completion time by task ID as well as overall.

participant make statistically significant faster decisions with Tableau ($\Delta_\mu = 24$). However, this was not the case with Valletto ($\Delta_\mu = 3$).

Observations on the Interactions

Informative subjective ratings by the participants reveal a positive impression on both the dialogue design (see Figure 5.14) and the reasoning panel (see Figure 5.15). Regarding the dialogue design, participants consider the dialogue as supportive for the decision making. Additionally, they feel reminded on what was already asked. In terms of the reasoning panel, the participants welcome the increased transparency on the system's behavior.

Moreover, the participants essentially use one of two different strategies for both visualizing data and choosing a preferred visual encoding, independent of the system. One group of participants first decide which data attributes to take and then apply potential filters on the attributes, if needed. The other group take the opposite order. Furthermore, participants tend to design summary-oriented visualizations like bar or line charts for the decision making, although a variety of alternatives exists. This observation on the visualization selection holds for both Valletto and Tableau. While using Valletto, however, participants occasionally

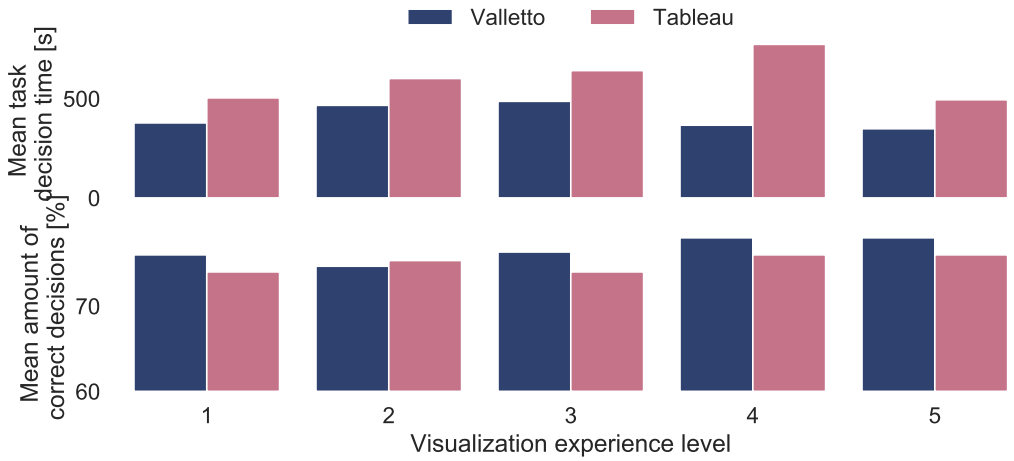


Fig. 5.12: Participants task completion time as well as the amount of correct decisions ordered by the participants' self-reported visualization experience.

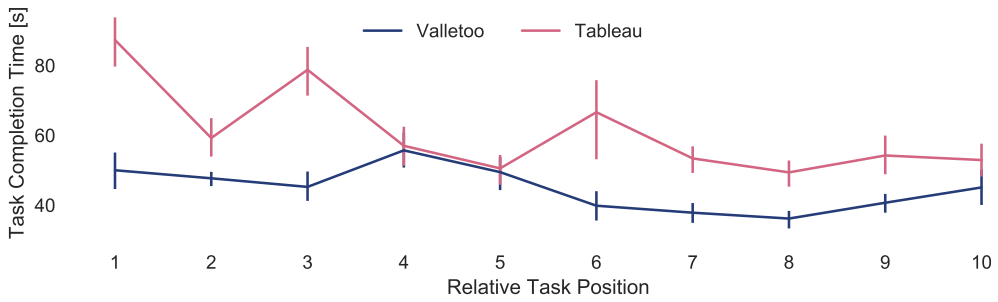


Fig. 5.13: Participants task completion time (mean and 50% confidence interval) by the relative position in the task sequence.

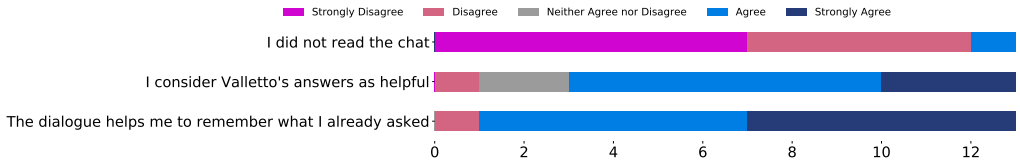


Fig. 5.14: Participants' ratings on given statements regarding the dialogue design.

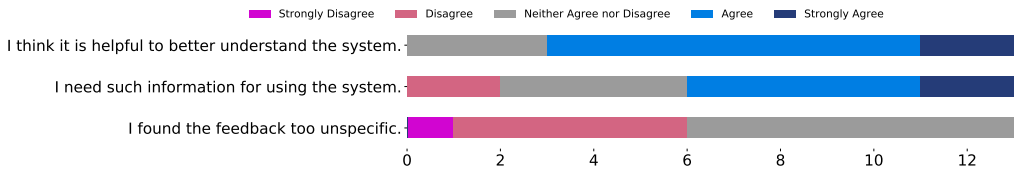


Fig. 5.15: Participants' ratings on given statements regarding the reasoning panel design.

change the visual mappings. Nevertheless, they work with the initially shown visualization most of the time. This visualization is the most effective visualization from a perception point of view (cf. section 5.3.2).

While using Valletto, some participants try to get a result by directly telling the system the tasks claim e.g., People in X tend to buy more Y than the others (where X and Y are concrete values and not data attributes). Yet, these participants quickly notice that the system does not understand them and adjusted their speech interaction tactic accordingly. They typically give precise information about the data they would like to see instead. A different tactic is to formulate small simple sentences to build up the visualization.

Open Exploration

Participants approach the open exploration by analyzing a sequence of single data attributes while using Valletto. Precisely, they start in the analysis tab for applying the desired filter on the data, but immediately jumped back to the overview tab. This behavior leads to more mentioned data statements focusing on single attributes while using Valletto in contrast to Tableau. In Tableau, these observations are not made. Instead the participants directly combine multiple data attributes. Additionally, the most experienced participants in visual analysis decide to generate maps to distinguish between countries. Overall, the participants tend to slightly find more observations with Valletto than with Tableau.

5.6.4 Discussion

The results of both conducted experiments reveal two things. First, the design elements of Valletto influence the participant's decision making in a positive sens. Second, the modality of speech triggers obstacles for the user interaction though.

The overall design of Valletto is perceived to be very good, according to the SUS score of 81.1. As the open exploration task further shows, the participants became familiar to the systems design after completing the first 10 tasks. They interact with the two different tabs naturally. Nevertheless, the observations of how participants directly talk with Valletto reveal room for further investigations, as discussed in the following.

Finding 1: Better decisions in Valletto. The two design elements of Valletto – the dialogue and the reasoning panel – arguably improve a participant's decision making. The positive ratings by the participants explicitly support this assumption (see Figure 5.14 as well as Figure 5.15). Additionally, both the improved task completion time (see Table 5.2) and the decision quality (see Figure 5.10) supports this assumption. In tasks with data-related

dialogue acts given, a significant time advantage emerges. This effect further seems to be present not only for inexperienced participants, but also for experienced participants (see Figure 5.12). Dialogue acts on data facts likely trigger a positive effect on the decision making in visual analysis. Hence, a broader spectrum of users potentially benefits from such kind of dialogues.

Furthermore, the reasoning panel is considered as helpful and supportive for better understanding the systems computations. However, an alternative design might be to directly include the highlighting of relevant words in the dialogue. Relevant parts of the users utterance could be labeled in the categories of intent, attribute, and filter.

Finding 2: Unified communication works, but not for everyone. In the current status of Valletto, each user gets the same set of answers depending on both the user's utterance and the data characteristics (see. Section 5.3.2). But users are diverse in many ways, e.g., experienced vs. inexperienced in visual analysis, data scientist vs. casual user, or with an interest in a broad overview vs. detailed aspects. For instance, an experienced participant asked Valletto for the corresponding p-value of the computed Spearman's correlation coefficient. This indicates that some participants are potentially not fully convinced by the given dialogue act, otherwise they would not have asked for the relevant statistical parameter. Srinivasan et al. (2019a) also reveal similar results in their experiment. Hence, adding these information to the dialogue acts might further improve the decision making, if the user understands the concept of the p-value.

Finding 3: Visualization types appear more important than mappings. In the current version of Valletto, the most effective visualization is always shown first. Although a user can freely navigate through the set of offered visualizations, the participants did not fully use this functionality. As observations showed, the visualization type was more of interesting than changing the visual mapping. However, an objective should be showing a desired visualization in the first place. The analysis process can certainly speed up when the first shown visualization fits substantially the users preferences.

Finding 4: Marginal learning effects in Valletto. As Figure 5.13 illustrates, there is not visible learning effect while using Valletto, although it exists for Tableau. Consequently, a user familiar with Valletto is not necessarily faster in the visual analysis than an unfamiliar user. As the familiar user might know which functions are available after a while, the system still has to transform the utterance into an appropriate response for the user, consisting of generating a visualization, and providing a dialogue act. This further means that an additional speed up of the visual analysis with Valletto likely depends on improvements of the technical side.

Interestingly, some participants' perceived duration for the performed tasks is longer with Valletto, although the overall objective task completion time was shorter. This phenomenon might be summarized under the term of temporal illusion. It presumably refers to Vierordt's law, which states that people tend to overestimate "short" time intervals and underestimate "long" time intervals with an "indifference point" in between (Lejeune and Wearden, 2009). One reason might be a short but noticeable time (1-2s) between sending an utterance and receiving an answer. Additionally, participants do not actively specify details of a visualizations, but delegate this to an agent. This echos the finding of Tory and Setlur (2019) and Setlur et al. (2019). Consequently, a user's interaction sequence for achieving an objective contains gaps between the interactions. In each gap, a user needs to observe the interfaces without actively interacting with it.

5.7 Limitations

Using speech for visual analysis feels strange for some participants. Furthermore, the NLP routine occasionally fails, e.g., the recognizer confused "profit" with "prophet". Participants for which the speech recognition failed are quickly less motivated in interacting with Valletto. Since the concept does not provide direct text input like other approaches, participants are not able to immediately resolve this situation. This limiting circumstance affects not only the task completion time, but also user experience. Additionally, some participants tend to hyperarticulation. However, participants are able to adjust their interaction strategies in those challenging situations by using the reasoning panel.

Although the design is primary envisioned for mobile use, the design is not evaluated in a mobile context. Especially, the ambient noise is most likely higher than in our study scenario and will eventually affect the performance of these systems. However, comparing the design to a conventional user interface design, advantages of the dialogue and the reasoning panel are identified.

5.8 Summary

Using natural language for generating and manipulating visualizations can reduce barriers both interaction and knowledge (Grammel et al., 2010). This chapter proposes a multimodal interfaces design for visual analysis through touch and speech. This design considers the conversation with the user as an essential aspect in a multimodal interface. Through the conversation, the user becomes more aware of the underlying data and the system's actions. In addition to the user's utterances, the dialogue contains dialogue acts on data facts as well as dialogue acts for maintaining a conversation. Additionally, an adjacent reasoning panel reveals the system's computations. Considering Chapter 4, the system restricts the

recommended visualizations by the used visualizations. Furthermore, a dedicated NLP routine considered the command structures.

Through two user studies, the following implications for future approaches in the area of conversational interfaces for visual analysis can be made:

- *Dialogue:* A persistent dialogue containing the user's utterances as well as the system's textual responses of additional information with respect to the visualization helps the user to make more reliable decisions. Therefore, it would be worth it to investigate further complex visual analysis tasks to see how they can be integrated into conversational interfaces for visual analysis. Finally, the system should talk in a personalized way with the user.

- *Speech:* Speech input should be directly used instead of text input due to limited display sizes and recent improvements in speech recognition technologies. However, evaluating these systems in the field is directly connected with using speech.

- *Engagement:* Conversational interfaces for visual analysis should stay engaged with the user while computing the user's utterance, otherwise the user's user experience will suffer. Hence, further designs should be discussed which go beyond the classical loading spinner.

However, the intelligibility of visual analysis can be further increased. While the reasoning panel already provides transparent and interpretable information on the system's behavior, the dialogue acts still follow a unified communication style. The next chapter faces this challenge. Furthermore, the visualizations recommendations also assume a standard user. However, Chapter 7 explores the personalization of this aspect in detail.

6 Investigating Dialogue Preferences

As shown, users appear to have different requirements concerning the communication of data facts. In order to approach this challenge, this chapter proposes a linguistically motivated answer space by considering the cooperative principle by Grice (1975). This answer space varies in two dimensions: information level and support level. First, a conducted online survey ($N = 76$) shows diverse preferences in the answer space. While the self-report knowledge of the participants significantly influences the preferences, other factors such as trust influence the preferences too. A controlled experiment ($N = 10$) further highlights effects of the answer space on user experience. Answers not aligned with the user's preferences trigger negative reactions, while answers following the user's preferences produce positive reactions. Overalla, this chapter shows how the conversation in visual analysis becomes intelligible.

Disclaimer: The following chapter is essentially based on the published article: Jan-Frederik Kassel and Michael Rohs (2019b). "Talk to Me Intelligibly: Investigating An Answer Space to Match the User's Language in Visual Analysis". In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. DIS '19. San Diego, CA, USA: ACM, pp. 1517–1529. ISBN: 978-1-4503-5850-7. DOI: 10.1145/3322276.3322282

6.1 Introduction

Valletto realizes a substantial dialogue between the user and the system, as previously introduced. This dialogue on data facts actually helps a user to make sense of data. However, the previous experiment highlights a diverse perception of the dialogue. While many users directly took advantage of the given information for making better decisions on the visual analysis tasks, some – more experienced – users question the given answers. Srinivasan et al. (2019a) observe a similar effect when textual information on data facts are attached to a visualization. More experienced users would likely see a proof of the statements given, e.g., relevant statistical methods and parameters. The lack of the trust in the system likely affects the user interaction and the produced results.

Hence, the system needs to somehow adjust to the user's language. Molich and Nielsen (1990) argue for matching the user's language as one of the essential usability challenges of conversational interfaces. Still, matching the user's language does not only mean to literally speak the same language as the user, e.g., English, but also to use "words, phrases, and concepts familiar to the user"(Molich and Nielsen, 1990). Following this argumentation, it further aligns with the essence of dialogue between human beings. In a human-human dialogue, a person adjusts to their interlocutor's language unconsciously (Gallois and Giles, 2015; Grice, 1975). Typically, the one with either more versed language skills or more knowledge about the dialogue topic adjusts more. Imagine a class room situation on learning derivation of a function in high school as well as at a university. While in both situations the objective of understanding the concept of a derivation is identical, the audience's backgrounds essentially differ. So, the chosen educational approach must be different. Therefore, a teacher needs to adapt to the audience and so should an intelligent system.

However, little knowledge exists on methods and effects of automatically adjusting to a user in multimodal visual analysis. This chapter investigates a way for personalizing textual descriptions on data facts according to the user's characteristics. First, an answer space is created based on linguistic fundamentals of Grice (1975). Second, an online survey investigates the diverse preferences in this answer space. It empirically reveals significant user characteristics for describing this preferences. This study data is further used to train and test a machine learning model for incorporating the answer space into an actual conversational interface. A succeeding experiment analyzes the effect of a personalized dialogue on the usability and acceptance of conversational interface for visual analysis.

6.2 Related Work

While Section 2.5 discusses related work on multimodal interfaces for visual analysis, this section introduces additional concepts from linguistic theory. Additionally, it discusses related work on general conversations with intelligent assistants as well as user-specific differences in the use of conversational interfaces.

However, the essential difference to the related work from Section 2.5 is the objective of achieving a user-specific conversation in visual analysis. All previous approaches propose a generalized conversation strategy with the user. Consequently, no matter how skilled or unskilled a user is, the system will not change its responses. Yet, matching the user's language is not only important in the general domain (Molich and Nielsen, 1990), but likely also in visual analysis, as addressed in Section 5.6 as well as by Srinivasan et al. (2019a).

6.2.1 Linguistic Theory

Generally, a system should match the user's language as well as support a user in achieving the objectives. Considering these circumstances from a linguistic perspective, cooperative principles by Grice (1975) are immediately relevant. Grice (1975) pleads for ideal communication between two persons where both are cooperating. He states:

> "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." (Grice, 1975, p. 18f)

His cooperative principle further comprises four maxims (Grice, 1975). First, the *Maxim of Quality* requires that every contribution to the dialogue is true. Second, the *Maxim of Quantity* describes the information content of the dialogue act. According to Grice, a statement needs to contain as much information as required, but not more than this. Third, the *Maxim of Manner* prescribes to be orderly and to avoid ambiguity in the conversational contribution. Fourth and last, the *Maxim of Relation* addresses the relevancy of the statement. Only relevant statements should be made with respect to the objective of the cooperation. Furthermore, Grice defines a violation of any of these maxims as an *implicature*. *Implicatures* are considered to be harmful for the cooperation.

In addition to Grice (1975), the Communication Accommodation Theory (CAT) (Gallois and Giles, 2015) also describes likely relevant aspects for an user-specific conversation in visual analysis. CAT is theory of how people adjust their conversation styles to their interlocutors. In a nutshell, it consists of the four phases. First, the *sociohistorical context* describes the general basis for the conversation of two people. Second *initial orientation*

focuses on how people initially estimate the conversational skill of their interlocutor. Third, the *psychological accommodation* essentially summarizes how an action for accommodation is made. Finally in *evaluations*, a person evaluates how the conversational contribution is perceived by their interlocutor.

Both theories appear to be immediately useful for personalizing the conversation in visual analysis. Both consider the involved parties as well as the corresponding objective for the conversation. Furthermore, related work do not consider the conversation under either concept.

6.2.2 Conversation Style

As in human-human dialogues, the design of conversational interfaces can be express multiple perspective. Especially chatbots require a design according to the user's objective (Chaves and Gerosa, 2019). These systems often support the user in information retrieval tasks, e.g., finding information regarding a point of interest.

However, Shechtman and Horowitz (2003) identify additional conversation styles beyond the simple interactions with a chatbot. According to Shechtman and Horowitz (2003), there are three different styles: task-oriented, communication-oriented, or relationship-oriented interactions. Shamekhi et al. (2016) show that matching the conversation style of the user likely accelerates the user experience with a conversational interface.

Hoegen et al. (2019) propose an end-to-end conversational agent in an multi-turn dialogue for an open domain. This agent aims for automatically matching the conversational style of the user. Hoegen et al. (2019) report an effect of potentials for increasing the trust in a conversational agent when the conversational style of the user is matched by the agent.

Furthermore, Branigan et al. (2011) investigate how people adjust their conversation style when they believe to talk with a computer. The authors show a tendency of the participants to align more with a computer than with a human. Additionally, they also show an effect that people tend to even more align with a simple computer than with a powerful computer. These results directly refer to the CAT. It further shows the general effect of belief on the interaction with conversational systems.

6.2.3 User-specific Conversational Interfaces

While the user's belief in the system's capabilities (Branigan et al., 2011) as well as the preferred conversational style (Hoegen et al., 2019; Shechtman and Horowitz, 2003)

influences the interactions with an intelligent system, user characteristics influence the interactions as well.

Luger and Sellen (2016) explore the relationship between user expectations and experience of conversational agents. Their results show a lack of trust in the systems by experienced users. Regular users of conversational interfaces focus rather on simple tasks than on complex tasks (e.g., writing an email). These users assume that a system will likely fail in executing the desired tasks. Cowan et al. (2017) confirm these findings for inexperienced users.

Chen and Wang (2018) compare interactions between inexperienced users and experienced users. Considering failures in conversational interface, the authors highlight a difference between these two user groups. The more experienced a user is, the more (s)he thinks about why a system has failed. These users are likely more effective in adjusting their interaction behavior in order to avoid prospective failures.

In summary, one of the major issues in the interactions with conversational interfaces is a lack of trust of the user (Hoegen et al., 2019; Luger and Sellen, 2016; Srinivasan et al., 2019a). Adjusting to the user likely helps to increase trust (Branigan et al., 2011). Furthermore, communicating in a personalized way further helps to cover and unleash, respectively, the complexity of visual analysis.

6.3 Structuring Communication

The essence of Valletto's design is the idea of establishing a conversation in visual analysis (cf. Section 5.3.1). This conversation is fostered through a persistent dialogue between the user and the system. Along with the user's utterances for pursuing in visual analysis, this dialogue contains dialogue acts on data facts (cf. Section 5.3.3). Depending on the given analysis situation and data attribute combination, these dialogue acts are differently expressed.

However, users differ in the use of these dialogue acts, according to the results of the conducted experiment in Section 5.6. While less experienced users tend to just consume the provided information, experienced users question the dialogue acts from time to time. This fact highlights that intelligible visual analysis also requires an adjustment to the user in terms of how data facts could be communicated. Although experienced users raised these points of lacking information, inexperienced users might struggle with certain dialogue acts too. For instance, a user with little knowledge about correlation might be overwhelmed by reported numbers from a correlation coefficient. The question remains how a personalized conversation in visual analysis can be structured by focusing on intelligibility.

6.3.1 Answer Space

Considering Grice's cooperative principle (Grice, 1975), an intelligible communication of data facts can potentially be achieved. First, Grice's maxim of quantity requires from a contribution to the dialogue to always contain the right amount of information, otherwise an implicature would be created. In the context of data facts in visual analysis, this "right amount of information" could be interpreted as the amount of information revealed to the user. In other words, how much complexity is revealed to the user.

Imagine, two attributes X and Y are positively correlated. Valletto could say: "X and Y are positively correlated", as it currently does (cf. Section 5.3.3). Yet, Valletto could also say: "X and Y are positively correlated according to Pearson's $\rho = 0.5$ and $p < 0.05$". The difference is not only in numerical values, but also in the choice of statistical method and metric and supplementary parameters. Although both statements are entirely true, it depends primarily on the user whether a statement is intelligible.

This example illustrates different ways of communicating the same data fact. However, varying the information content is likely not the only factor influencing the intelligibility of a dialogue act. The formulation of uncertainty is another factor. Multiple research studies reveal preference difference in communication of uncertainty (Dhami et al., 2015; Renooij and Witteman, 1999; Wallsten et al., 1993). People tend to prefer either numerics or words for describing uncertainty (Barnes, 2016; Budescu et al., 1988), e.g., "the chance of rain is by 5% tomorrow" compare to "there is a little chance of rain tomorrow".

Furthermore, the use of relevant statistical terms such as "correlation" likely influences the intelligibility as well, since a user needs to know this term otherwise the meaning of the dialogue acts gets lost. Losing the meaning leads further to likely wrong conclusions. Hence, a system should explain either the term or its implication, if needed. Yet, explaining the implication of a term such as correlation requires careful considerations. A typical mistake in data analysis is to automatically assume correlation implies causality. However, deciding for causality requires solid domain knowledge. As many concepts – such as this thesis – follow a domain-agnostic approach, the system should focus on explaining what correlation effectively means instead. Recall the example above, the statement "X and Y are positively correlated" could also be equivalently communicated as "When X increases then Y increase on average as well". These kind of formulations supposedly help users to make sense of the data when they are unaware of the term usually used.

According to this argumentation, a dialogue act on data facts could be communicated along two dimensions. On the one hand, the information content of a dialogue acts can vary following Grice's maxim of quantity. On the other hand, dialogue acts might contain the related statistical term or explain its meaning instead. In combination, these dimensions

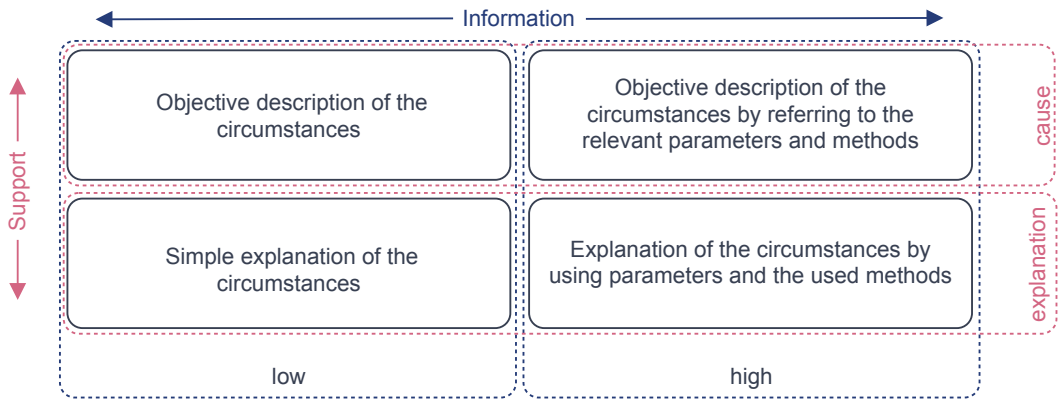


Fig. 6.1: The two-dimensional answer space varies in the dimensions of information and support. While in the upper left corner the answer space contains descriptive answers with little information on the used methods, the lower right corner contains explanatory answers with detailed information on the used methods.

consequently form an answer space for increasing intelligibility in a user-specific way, shown in Figure 6.1. Precisely, the dimensions called *information level* and *support level*. This answer space enables a system to communicate a data fact in four different ways in order to adjust to the user. The four different ways are to describe a data fact with low information (*LC*), to describe a data fact with high information (*HC*), to explain a data fact low information (*LE*), and to explain a data fact with high information (*HE*). All fours ways are both valid and equivalent. While a novice user in visual analysis would presumably prefer a dialogue acts following *LE*, an experienced user would likely go for *HC* when (s)he lacks of trust in the system, otherwise (s)he would prefer *HE*. However, matching the user's language (Molich and Nielsen, 1990) in visual analysis is essential, otherwise implicatures (Grice, 1975) would likely lead to a bad user experience and an unintelligible visual analysis. Imagine, a novice user is most likely overwhelmed by the results of a correlation coefficient, although this user does not know neither what correlation means nor what the coefficient describes.

6.3.2 Relevant Dialogue Acts

As the answer space provides a theoretical structure for communicating dialogue acts on data facts in a user-specific way, it does not define how the concrete implementation looks like. Furthermore, not all supported dialogue acts likely require a use of the answer space. In fact, only the dialogue acts related to data would likely benefit from the answer space. Hence, the relevant dialogue acts are *DEPENDENCIES-numeric*, *DEPENDENCIES-categorical*, *DEPENDENCIES-arbitrary*, *COMPARISON*, *EXPLORATION*, and *FILTER*.

Considering a dialogue acts *DEPENDENCIES-numeric* on X and Y, the corresponding implementation of the answer space could look like:

LC X and Y are positively correlated.

LE X and Y are positively correlated, according to Pearson's $\rho = 0.5$ and $p < 0.05$.

HC When X increases then on average Y also increases.

HE When X increases then on average Y also increases, according to Pearson's $\rho = 0.5$ and $p < 0.05$.

As *DEPENDENCIES-numeric* is still a quite handy implementation of the answer space, other dialogue acts are much harder to implement. In order to make sure that the answer space is correctly implemented for each dialogue act, all possible answers for each dialogue act have been checked on correctness by an experienced data scientist.

6.4 Experiment 1: Preferences and Differences

As the answer space is set up, the question remains whether people have actually different preferences in this answer space. And if so, what are the influencing factors describing these preferences.

RQ 6: What are the influencing factors for matching the users language in the answer space?

6.4.1 Procedure

In order to investigate these questions, an online survey is set up. This survey essentially consists of two phases. In the first phase, each participant conducts a randomly assigned sequence of 12 concrete visual analysis tasks. These 12 tasks cover the support visual analysis situations by the answer space where each situation is represented by two tasks (see Table 6.1). For instance, the analysis of two quantitative data attributes is covered by one task on positively correlated data as well as one task on negatively correlated data.

Each task shows an effective visualization (bar, line, or scatter plot) and two initial options from the answer space. These two initial options are selected based on a randomly chosen dimension for the answer space. For instance, if the dimension of *low information* is randomly chosen, the two initial answers are: one for describing this information (LC) and one for explaining it (LE). The objective of the participant is to select the most preferred answer space option for making sense of the visual analysis situation. Once a participant decides for a preferred option, again two options are shown from the answer space while the visualization stays the same. Now the answers are selected based on the participant selection. For instance, if the participant decided for the descriptive answer with low information

Dialogue Act	Condition
FILTER	one filter applied
FILTER	two filter applied
DEPENDENCIES-categorical	independent attributes
DEPENDENCIES-categorical	dependent attributes
DEPENDENCIES-numeric	positively correlated attributes
DEPENDENCIES-numeric	negatively correlated attributes
DEPENDENCIES-arbitrary	low mutual information
DEPENDENCIES-arbitrary	high mutual information
COMPARISON	significant difference between the groups
COMPARISON	no significant difference between the groups
EXPLORATION	from a categorical attribute
EXPLORATION	from a quantitative attribute

Tab. 6.1: Dialogue acts and corresponding conditions of the 12 tasks of the experiment.

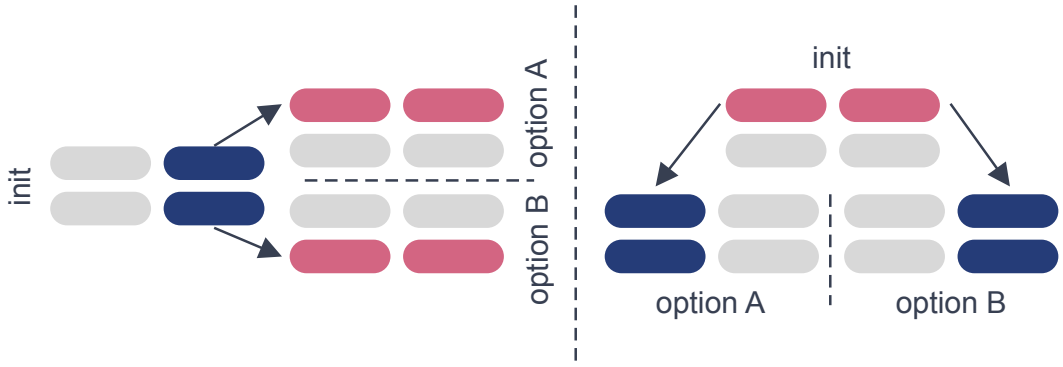


Fig. 6.2: Illustration on how the selection of the first randomly shown answers determines the subsequent answers in each analysis task.

instead for the explaining answer, then, the answers shown next are a descriptive answer for low information (LC) and a descriptive answer with high information (HC). Again the participant should decide for the most preferred answer. Figure 6.2 visually elaborates this study procedure.

This experiment design has two advantages. First, the participant is likely not overwhelmed by selecting an option from all four answer space options. Second, more detailed information on the participant's preferences are collected, as a participant makes two greedy decisions in each task.

After completing the sequence of the visual analysis tasks, a questionnaire follows. This questionnaire is on the participant's experience in visual analysis, statistics, and in the use of conversational interfaces (e.g., chatbots). Furthermore, the participant should also provide feedback on their conversation preferences. In order to select only reliable answers, a "honeypot" questions is included (Paolacci et al., 2010).

6.4.2 Participants

Covering the wide range of expert levels from novice to expert user is important for this experiment. In order to receive answers from experienced people, people are recurred with a major either in computer science or mathematics from a university as well as from a data science department at an industry company. For collecting answers from likely less experienced people, the survey link was broadcast at Amazon's Mechanical Turk (MTurk). However, only people with a US bachelor's degree could participated, directly ensured by MTurk. This restrictions presumably leads to a choice of participants likely aware of the possibilities of data.

Generally, MTurk serves in many research areas as a basis for getting study results both reliable and fast (Buhrmester et al., 2011; Kittur et al., 2008). Casler et al. (2013) highlight the quality of MTurk results compared to face-to-face testing. Furthermore, Heer and Bostock (2010) show comparable quality of study results from MTurk in the domain of visual analysis too.

Overall, 87 people participate in this experiment. After cleaning and reliability checks, 76 participants remain. From these 76 participants, 23 are from industry or university and 53 from MTurk, respectively. According to the questionnaire, the majority of the participants is generally aware of the used statistical methods. However, only 49% of the participants know the Bonferroni correction. Considering the participants' self-reported statistical knowledge level, 38% novices, 15% advanced beginners, 25% competent users, and 22% experts participate in this experiment.

Furthermore, 77% of the participants would prefer to be guided by an intelligent digital assistant. Additionally, 75% would value recommendations of useful attribute combinations while exploring a new data set. Finally, 73% of the participants state that at least one of the provided answer space options fitted their preferences in each situation.

6.4.3 Results

The following analysis primarily uses the final decision of each participant in each situation. Furthermore, the results are aggregated on dialogue act level, since each dialogue acts was covered by two tasks. This aggregation is valid due to a conducted McNemar's test showing no significant differences between the two tasks supporting each dialogue act.

First analysis focus is on the influence of the user's self-reported knowledge on the user's preferences. According to a series of conducted χ^2 tests (see Table 6.2), knowledge has a statistically significant influence on decisions for a preferred response in almost every

Dialogue Act	$\chi^2(9)$	p	$Cramer's\ V$
FILTER	26.64	< 0.01	0.24
DEPENDENCIES-categorical	8.92	0.44	-
DEPENDENCIES-numeric	18.11	< 0.05	0.20
DEPENDENCIES-arbitrary	22.30	< 0.01	0.22
COMPARISON	17.41	< 0.05	0.20
EXPLORATION	21.57	< 0.05	0.22

Tab. 6.2: Results of individually conducted χ^2 tests.

situation, except when the dependencies between two categorical attributes have been investigated (*DEPENDENCIES-categorical*). However, a more detailed pattern reveals when considering the different dialogue acts individually. Figure 6.3 explicitly illustrates the distributions of the participants' preferences in the answer space depending on both self-reported knowledge and dialogue act.

Considering the exploration of a potential relationship between two attributes of unequal level of measurement (*DEPENDENCIES-arbitrary*), participants with less self-reported knowledge prefer a response with low information content (either *LC* or *LE*) while participants with higher self-reported knowledge simultaneously prefer response with a higher information content (either *HC* or *HE*). Both results are statistically significant according to a Šidák-corrected χ^2 test with $\chi^2(3) = 18.76, p < 0.012$. Additionally, an identical result holds when the participants filtered data (Šidák-corrected χ^2 test with $\chi^2(3) = 15.78, p < 0.012$). However, the corresponding statistical effect is high in first situation ($V = 0.35$) while in the second situation only a medium effect is persistent ($V = 0.32$).

Furthermore, participants statistically prefer with a medium effect ($V = 0.31$) rather an explanatory to a descriptive response when comparing multiple categories, according to another Šidák-corrected χ^2 test, $\chi^2(3) = 14.38, p < 0.012$. Precisely, they prefer by factor 3.25 an explanatory response, as a sequentially conducted Fisher's exact test reveals with $p < 0.012$.

Not only the self-reported knowledge influences a participant's preferences in the answer space, but also particular knowledge in the statistical methods used for the dialogue acts partly determines the preferences. A large ($V = 0.31$) statistical effect exists when analyzing dependencies between two quantitative attributes, following the results of a conducted Šidák-corrected χ^2 test with $\chi^2(3) = 13.62, p < 0.012$. The participants tend to prefer response with higher information when they also know the Spearman's ρ.

Besides the participants' self-reported knowledge, there are additional factors influencing the participants' preferences. On the one hand, the participants' desire to be guided in visual analysis statistically significantly relates to the preferences in the answer space, based on the results of a χ^2 test with $\chi^2(12) = 25.62, p < 0.05$ and $V = 0.10$. Furthermore, the

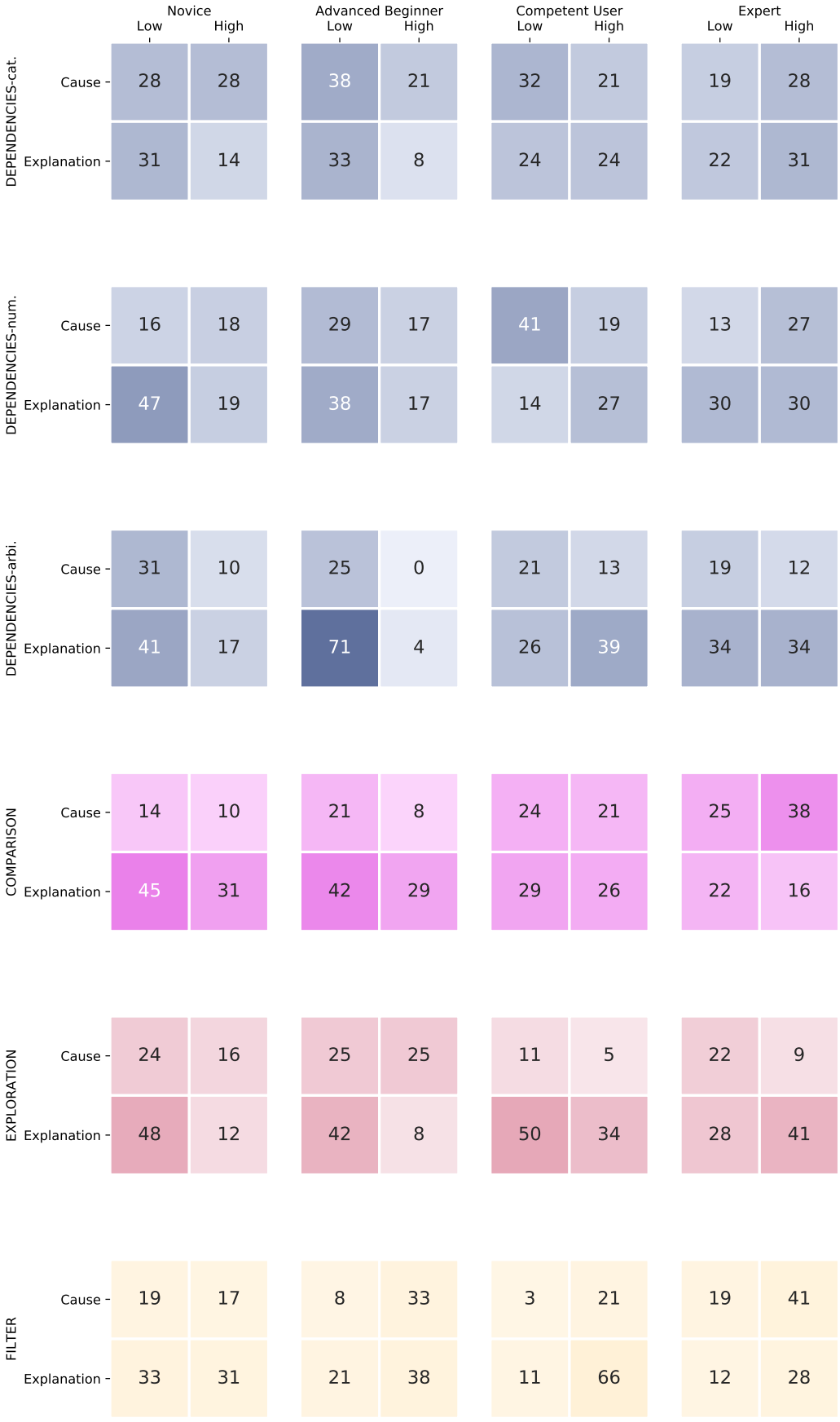


Fig. 6.3: Distribution of the participants' preferred answers from the answer space ordered by the self-reported knowledge level and the different supported situations.

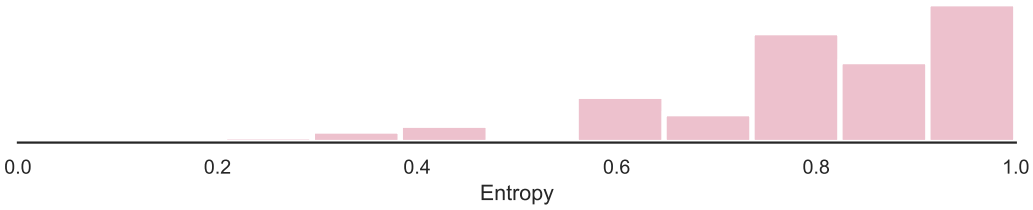


Fig. 6.4: Distribution of the Entropy of the participants' preferences in the answer space. For each participant, the Entropy is individually computed based on the selected response category (*LC*, *LE*, *HC*, or *HE*) from the answer space.

participants' curiosity on the used statistical method for giving the corresponding response is relevant. According to a conducted χ^2 test with $\chi^2(12) = 30.74, p < 0.01$ and $V = 0.11$, it statistically significantly influences the preferences. However, participants – both novice users and component users – statistically significantly ($\chi^2(4) = 13.50, p < 0.01, V = 0.20$ and $\chi^2(4) = 14.52, p < 0.01, V = 0.25$, respectively) choose responses with lower information (either *LC* or *LE*) when they are less interested in the background computations of the system.

On the other hand, participants who want to be in control of the entire visual analysis process prefer responses with a higher information content (either *HC* or *HE*). For both novice users and expert users, this effect is statistically significant by $\chi^2(4) = 24.52, p < 0.001, V = .28$ and $\chi^2(4) = 15.47, p < 0.01, V = .29$, respectively.

Generally, the participants' preferences in the answer space differ very much. First, the overall average pairwise Hamming distance is 0.72 with $\sigma = 0.14$. This shows large differences in the preferences between the participants. Second, the individual preferences of the participants further vary along the study. Computing the Entropy on the chosen responses from the answer space for each participant, Figure 6.4 shows varying preferences depending on the situation.

6.4.4 Discussion

The results of this experiment reveal not only different preferences in the answer space, but also address relevant factors for describing a user's preferences in the answer space. Hence, the results further answer the corresponding research questions 6 and improve the understanding of preferences in dialogue acts on data facts in visual analysis.

In general, the analysis results contradict the assumptions of Section 6.3. Assigning the different categories of the answer space to a user is not straight forward, as the knowledge is just one factor of many. Furthermore, these results further empirically support the output of Chapter 3 on the potentials for personalization of visual analysis tasks. The corresponding

structuring assumes that not only the knowledge influences the personalization potentials, but also the user's preferences.

Finding 1: It is not only about the knowledge. The analysis highlights the self-reported knowledge as a prime factor for explaining a user's preferences in the answer space. Participants with little statistical knowledge prefer more often a response with less information explaining the data fact. Although it generally helps to distinguish the preferences, a clear tendency in the preferences does not hold for the other self-reported knowledge levels.

Instead, a participant's preferences on both the transparency of the used method for providing the corresponding data fact and a guidance in a particular analysis situation influence the preferences significantly too. Participants with higher need for transparency choose responses with higher information, likely because of a clear naming of the used method and the corresponding parameters. This information potentially supports the increase of trust in the system for those participants. However, participants with a higher acceptance of digital assistants prefer an explanatory answer precisely because they trust this type of approach. Maybe, these participants already use digital assistants in their daily life.

Finding 2: Changing preferences depending on the situation. Along with the user's characteristics describing the user's preferences in the answer space, it seems that the particular data analysis situation further defines the selected response. Since the individual Entropy is high (see Figure 6.4), the participants vary in their response category from one situation to the other.

On the one hand, a reason could be the lack knowledge of the used methods for the particular tasks. While the χ^2 is likely familiar to many participants, many users are likely unfamiliar with the Spearman's correlation coefficient. On the other hand, it could depend directly on the task itself. Correlation could be easier to read from the visualization wherefore the participants maybe put less weight in the corresponding dialogue acts. Therefore, participants likely rate a dialog act higher, if the visualization cannot illustrate the data facts alone.

Finding 3: No clear separation between novice and expert users. As the empirical analysis shows no clear separation of the preferences in the answer space only based on the self-reported knowledge. In fact, the distribution of the preferences varies between the self-reported knowledge levels. Hence, a differentiation between novice users and expert users will likely fail for personalization. Instead, the responses have to be individually adjusted to the user's characteristics. Simple rules such as experts prefer high information will not work. A user-specific adjustment for a response retrieval could be achieved through a data-driven approach.

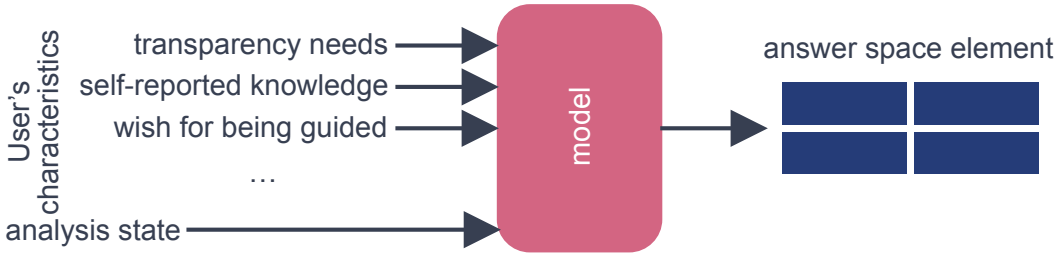


Fig. 6.5: Schema of the prediction problem.

6.5 Implementing the Answer Space

The previous experiment shows diverse preferences in the answer space. According to the results, responses for data facts should be personalized considering not only the user's knowledge, but multiple factors. This circumstance requires an intelligent implementation of the answer space into a system, since a unified communication strategy would likely fail.

In order to achieve a user-specific communication in an actual system, two challenges essentially arise. First, a system must be able to accurately predict a user's preferred response. If a system wrongly predicts the user's preferences it immediately creates implicatures. In a task-oriented dialogue (Shechtman and Horowitz, 2003), however, implicatures should be avoided to both safe time and prevent confusion. Second, this predicted response must align with the user's current analysis objective. This challenge directly addresses the *Maxim of Relation* of Grice (1975). Generally, the question is

RQ 7: Can the user's preferred communication style be accurately predicted?

6.5.1 Predicting a Preferred Answer

Predicting a preferred answer is an elementary part of integrating the answer space into an actual system. This challenge of predicting an answer aligned with the user's preferences can be formulated as a classification problem (see Figure 6.5). Formally, the input features for this classification problem are both the user's characteristics and the analysis situation. Accordingly, the classes to be predicted are the different categories in the answer space (*LC*, *LE*, *HC*, and *HE*). In order to properly predict these classes, a corresponding model needs to be investigated.

However, any model approaching this challenge eventually needs to be trained and evaluated on the data from the online survey. This data set comprises 912 data points. Hence, the set of potential models shrinks to the classical approaches such as logistic regression, since modern approaches like neural networks require far more data.

Classifier	reduced	entire
Naive Bayes	0.331 ± 0.02	0.328 ± 0.02
multi. Logistic Regression	0.305 ± 0.03	0.335 ± 0.02
Naive Bayes + Tree Embedding	0.629 ± 0.02	0.590 ± 0.02
multi. Logistic Regression + Tree Embedding	0.697 ± 0.02	0.706 ± 0.01

Tab. 6.3: Accuracy scores of the different model approaches with and without feature selection.

As the target comprises four classes, two different models are initially explored: a multinomial Naive Bayes and a multinomial logistic regression. Both models generally perform well on small data. However, the approaches differ in the learning paradigm. While the Naive Bayes models the joint distribution of the input features (user's characteristics and the analysis state) and the output (answer space categories), the logistic regression tries to reduce the error for mapping the input on the output. Furthermore, the Naive Bayes assumes that all input feature are independent, while the logistic regression can handle the dependencies to a certain extend.

Initial results reveal a poor performance of both models. The Naive Bayes and the logistic regression achieve an accuracy score of $\mu = 0.328, \sigma = 0.02$ and $\mu = 0.335, \sigma = 0.02$, respectively. These results require to investigate potential feature engineering. Generally, feature engineering can be approached through either reducing the feature space via a, e.g., Principal Component Analysis (PCA) or feature selection via a, e.g., random forest. As the feature space is already quite small, the feature selection likely helps more. Feature selection through a random forest creates a new binary feature space serving as the input for the models. These new binary features enable to vary the importance of the features for the different classes. Concatenating the feature selection with each model leads to much better performance. Now, the Naive Bayes and the logistic regression achieve an accuracy score of $\mu = 0.590, \sigma = 0.02$ and $\mu = 0.706, \sigma = 0.01$, respectively.

However, these results base on all features which essentially represent the questions at the end of the online survey. Considering this fact from a user point of view, a new user has to answer these questions again before a model can predict a probably preferred answer. This circumstance implies a certain burden for the user. In order to reduce this burden for the user, the feature space is reduced to the questions which reveal a significant effect on the user's preferences (see Section 6.4.3). Applying this reduced features space on the different approaches, the logistic regression with feature selection still performs better. Table 6.3 shows the accuracy scores of the different model architectures.

The performance of the logistic regression with feature selection only drops by 1% on the reduced feature space compared to the entire feature space. Furthermore, Figure 6.6 shows the confusion matrix for the different classes. The classification result essentially shows two insights. First, the model can predict each class equally well. Second, the model can

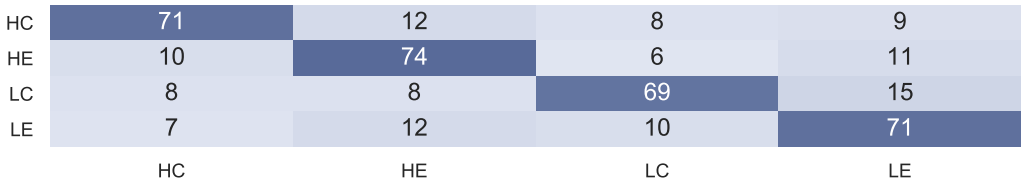


Fig. 6.6: Confusion matrix of the multinomial logistic regression with a tree embedding.

relatively well distinguish between the classes. These results further answer the research question. The preferred answers of the users can be predicted under certain conditions.

6.5.2 Get to know the user

The predictive model requires knowledge about the user. Formally, the input features of the model need to be set according to the user's characteristics. While Hurst et al. (2007) detect novice and skilled users based on their interaction data, this thesis' approach leverages the idea of having a conversation with the user.

Considering this challenge from a conversational perspective, it is somehow similar to a situation in which two people meet for the first time. Typically, people perform a mutual introduction by telling, e.g., their names, where they come from, or what they are doing for a living. This mutual introduction can be further considered as part of the accommodative orientation of two persons described under the CAT (Gallois et al., 2005). The "initial orientation" (Gallois et al., 2005) helps the interlocutors to better understand each other and initially adjust their language style.

As Valletto already implements some element of a mutual introduction (cf. Section 5.3.3), this principle from the CAT can be conveniently added. Figure 6.7 shows a part of the corresponding dialogue sequence for collecting the required information from the user. While Valletto initially explains what it can do for the user, Valletto further asks the user the relevant questions. This questions are identically formulated as in the online survey of Section 6.4.

However, the system assumes that the user's knowledge and other preferences are static during the sessions, since the mutual introduction only happens at the beginning of the analysis. If a user does not want to answer the initial questions, the system takes the average scores for the corresponding features, otherwise the system would not be able to predict an answer during the analysis.

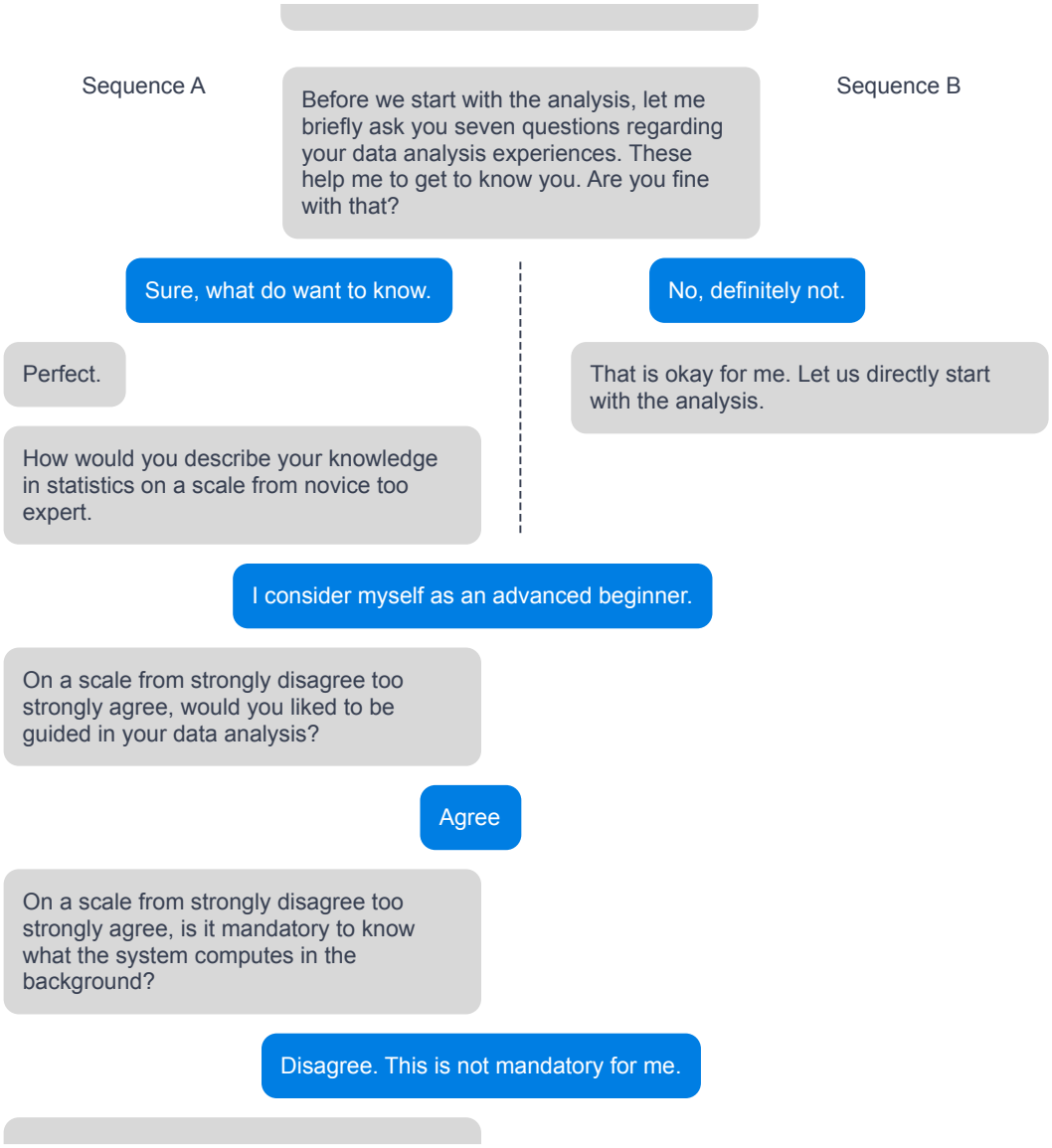


Fig. 6.7: Part of the mutual introduction. The particular dialogue sequence changes depending whether the user wants to answer some initial questions.

6.5.3 Predicting in a Live System

While the required information of the user is collected through a mutual introduction, the relevant analysis state still needs to be infused in the model. Determining the correct analysis state directly refers to Grice's *Maximim of Relation* (Grice, 1975). It prescribes to always contribute only relevant information towards the direction of the conversation. Additionally, the CAT summarizes a similar prerequisite under the term of "goals and addressee focus" of the immediate situation (Gallois et al., 2005). However, the analysis states are already predicted by the routine of Section 5.3.4.

Nevertheless, predictive models can also predict wrongly. As long as a predictive model achieves not a performance of 100% accuracy, failures likely happen during the usage. Hence, a user should be empowered to ask for a differently formulated answer when the initially predicted answers does not match the user's language. As an example, an inexperienced user might receive a descriptive answer with high information during a multiple comparison. This user is likely confused by the answer and presumably says something like "I really don't understand what you're implying". Adjusting a once given answer again is further supported by the "goals and addressee focus" (Gallois and Giles, 2015). In order to change a preferred answer, a user can tell the system to change accordingly.

6.6 Experiment 2: Reactions and Acceptance

According to previous analyses and investigations, people have diverse preferences concerning the design of a response. However, the effect of the answer space in an actual setting is still vague. Therefore, this experiment aims at qualitative results concerning the reactions of people to online predicted responses. The corresponding research question is:

RQ 8: What is the effect of an answer space in a multimodal visual analysis system during a realistic situation?

Furthermore, the previous online survey explore preferences on given options from the answer space, but provide no opportunities for the participants to formulate alternative responses. For instance, a user may want to receive an answers in *DEPENDENCIES-numeric* containing only the correlation coefficient, but not the p-value. Therefore, this experiment explores whether granularity of the answer space is suitable for the participants.

RQ 9: Is the granularity of the answer space adequate?

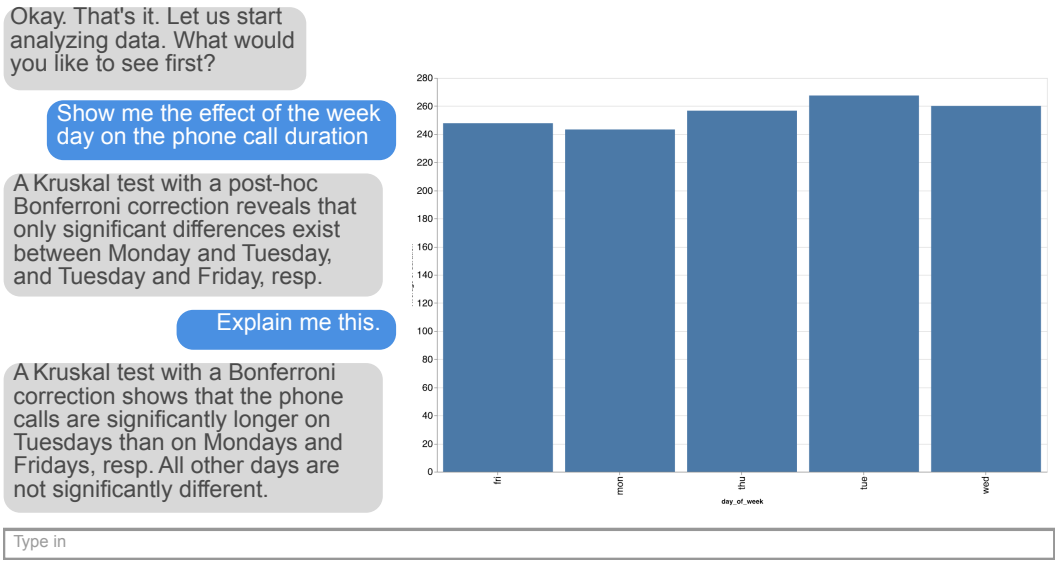


Fig. 6.8: In essence, the design of the technical prototype for the study follows the design of Valletto.

6.6.1 Procedure

The experiment starts with a standardized introduction on the procedure, the underlying data set (Dua and Graff, 2017; Moro et al., 2014), and the prototype itself (see Figure 6.8). After this introduction, the participant interacts with the prototype by completing the mutual introduction first. Afterwards, each participant completes a randomly assigned sequence of six tasks. These six tasks cover visual analysis situations supported by the answer space. In each task, a participant tries to make sense of the data visualized. After completing a task, a participant evaluates on a 5-point Likert scale the satisfaction with the visualization shown as well as with the last answer given. If no provided answer eventually matches the a participant's preferences, (s)he gets a chance to provide an alternative formulation. Additionally, each participant is asked to think-aloud during the interaction with the system.

After completing the randomly assigned task sequence, each participant fills out a questionnaire focusing on the participant's feedback on the overall conversation as well as demographic information.

6.6.2 Participants

In this experiment, 10 people in the age of 25 to 40 participated. On average, these participants work in industry for four years. Additionally, they have an academic background either in natural science, computer science, or economics. All participants state to be familiar with using conversational interfaces such as Siri or Alex, prior the study.

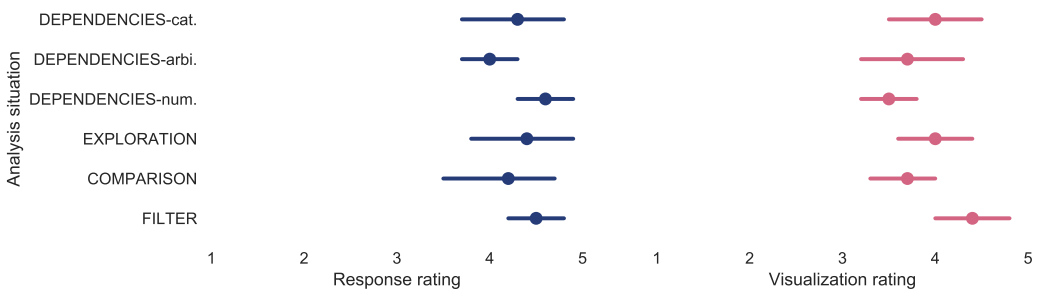


Fig. 6.9: For each supported analysis situation, ratings of both given response and given rating.

6.6.3 Results

According to the participants' ratings, the system's responses and visualizations are helpful or very helpful. Figure 6.9 illustrates these assessments. It shows high scores for both responses and corresponding visualizations in all tasks. Furthermore, there are no significant differences in the scores between the tasks.

As shown in the analysis before, the predictive model has a chance to create implicatures (cf. Section 6.5). In order to repair these situations, the participants can ask for a different response. However, the first predicted responses already satisfy the participants in 59% of all cases, the second predicted responses satisfy the participants in 30% of all cases, and only in 11% of the analysis situations, a third response needs to be given. Figure 6.10 shows a detailed perspective on these changes. Overall, the analysis tasks focusing on *DEPENDENCIES-numeric* trigger the highest number of requests for a reformulation.

Figure 6.11 highlights the performance of the predictive model. In *DEPENDENCIES-numeric* and *DEPENDENCIES-arbitrary*, the participants tend to increase the information content of the responses by asking for reformulations. The opposite is evident in *COMPARISON*. The participant tend to decrease the information content. In the other analysis tasks, the first given answers are already matching the participants' preferences. However, the option to ask for a reformulation is positively perceived by the participants. For example, one participant state "I liked that the tool changed the formulation when I requested it".

Taking a closer look at the participants reactions to the given responses. On the one hand, participants' reactions are negative when a potential implicature is given. One participant mentions "I do not understand what the bot is saying", when the system provides detailed information on the Spearman's correlation coefficient. However, this participant positively responds after receiving an explanatory response with little information by saying "...now I understand". Potentially, this participant's knowledge is exceeded in this particular analysis situation. Furthermore, a participant state "the first answer was too sketchy, but the second was much better". For this participant, the system likely reveals too little transparency on the

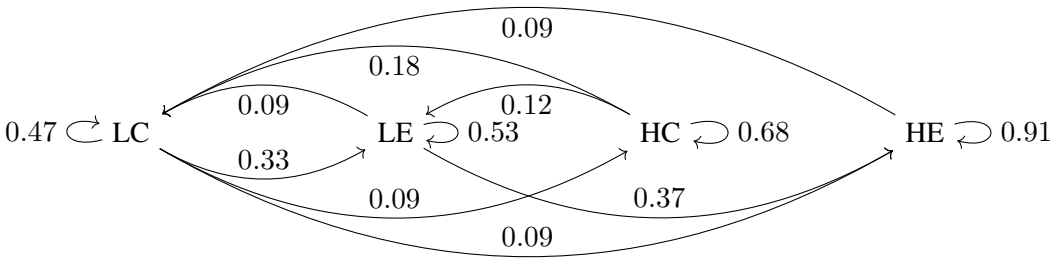


Fig. 6.10: Transition probabilities from one answer space element to the others according to the reformulation requests by the participants.

computations in the background. Another participant faces a similar challenge. The system responds with a descriptive answer with little information. Accordingly, the participant state "Well, I have to decode this first. (S)he takes some time to decode the answer correctly, but reacts positively after finally receiving an explanatory answer.

On the other hand, the participants react immediately positive when a language matching response was initially provided. A participant casually mention "I really like the answers". Another participant welcomes a response "I liked the simple and short answers". This participant primarily receives answers with little information content. Furthermore, explanatory answers of the system are especially welcomed in certain analysis situations. One participant states "the bot tried to explain things, that was good".

Moreover, the participants also put trust in the system and its computations, as the system needs to compute certain statistical methods. Given a descriptive response with little information, a participant responses "I assume the system used the right tests". This participants likely estimates the situation correctly due to certain experience in data analysis. However, (s)he rely on the system's computations. Furthermore, an experienced participant states "I trust the answer, because I had the same notion".

As the system needs to initially collect information from the participants, the mutual introduction is implemented. The participants essentially perceive the mutual introduction as positive. Yet, the mutual information creates an effort for the participants. For instance, one participant states "oh my good, seven questions" in the moment when the system asked whether it is okay to ask seven questions. This participants further states that these questions are "...not too much, but only okay for the first use" as well as that (s)he does want to answer more initial questions ("not more questions"). However, other participants react differently. They welcome the idea of the mutual introduction of the system. In this context, one participant explicitly states "I like that it tried to evaluate me in the beginning".

Figure 6.12 further highlights the participants' estimation of both a potential lack of information and the potential benefits of the system. Overall, the majority of participants

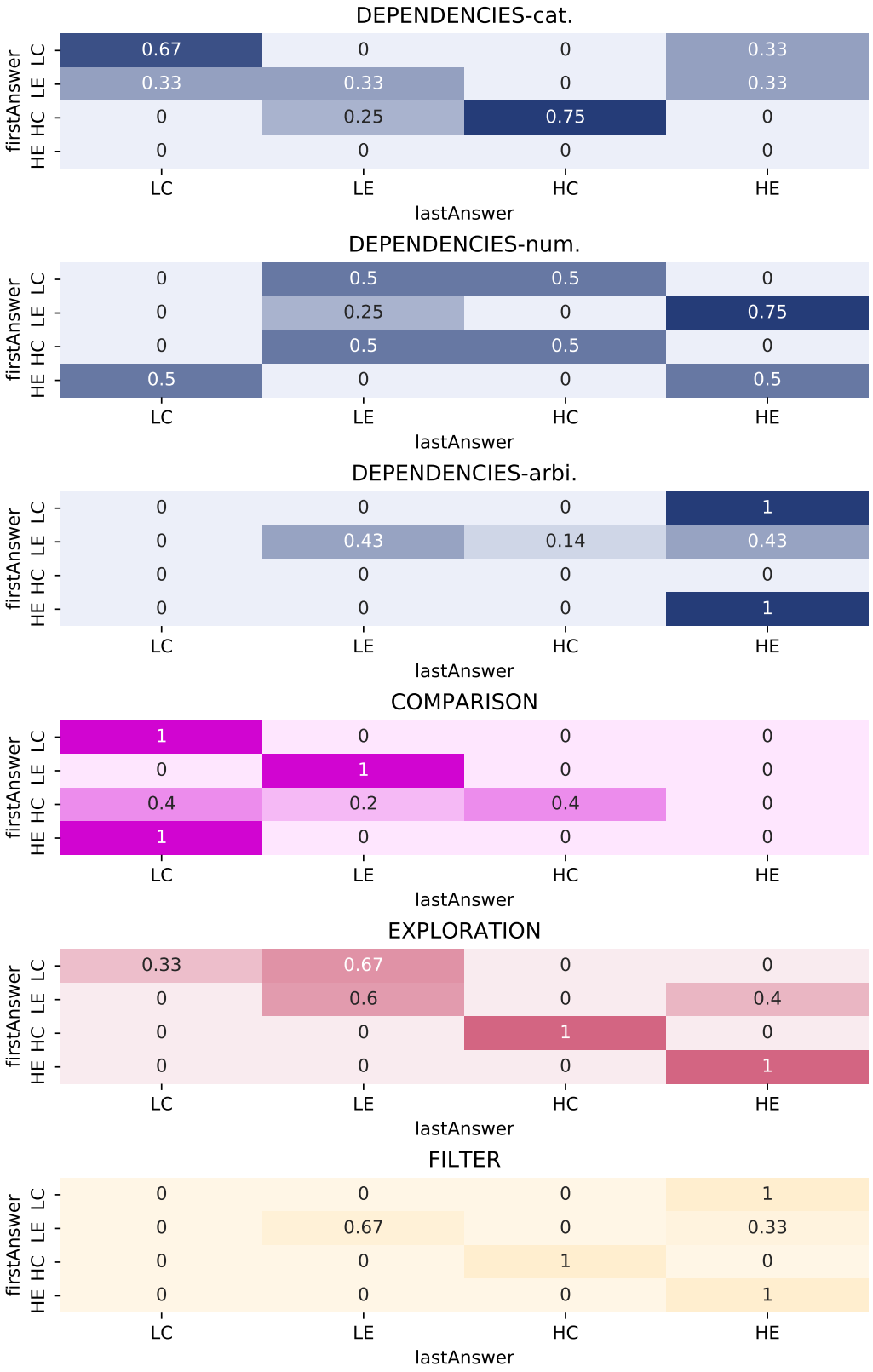


Fig. 6.11: For each supported analysis situation, a confusion matrix highlights the first given answers by the system and the eventually accepted answers by the participants.

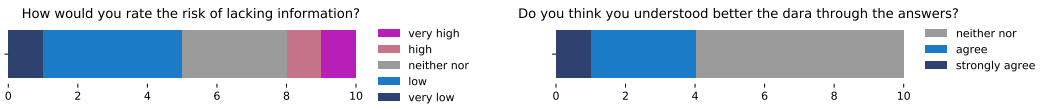


Fig. 6.12: Using a 5-point scale, the participants estimated both how the system's responses helped to increase their understanding in the analysis situation and how they perceive a risk on lacking important information.

assume that they have received all the information they need from the system. Consequently, they do not lack information. Additionally, the participants self-reportedly benefit from the given responses in order to complete the analysis tasks.

Finally, the participants provide only a few alternative formulations for responses. These alternatives concern the wording of the answers, but not the information content or the way this information is conveyed. For instance, one participant states "I would not use reliable" in the dialogue act on *COMPARISON*.

6.6.4 Discussion

This experiment answers research question 8 and research question 9. Both address an influence of the answer space on a user's visual analysis. Generally, the answer space has an positive influence on the visual analysis according to the participants' reactions. It appears that the answer space is a lightweight yet effective framework to personalize dialogue acts in visual analysis. Hence, the answer space enables a visual analysis system to dynamically unleash and cover, respectively, complexity to the user.

Finding 1: Matching the user's language affects the visual analysis. Initially, the online survey reveals a diverse perspective on the users' preferences in the answer space, as shown in Figure 6.3. Considering now the results from this experiment, matching the user's language in visual analysis is important. While responses aligned with the user's preferences provoke positive reactions, the opposite is persistent when potential implicatures are triggered. The participants reactions clearly support this hypothesis. The effect is especially demonstrative when a response aligned with the user's preferences follows an implicature.

However, the answer space also affects the usability. As a system supports a user in fulfilling an objective, time is an important factor. Since mismatching the user's language in visual analysis triggers an implicature, a user needs to deal with the provided response. The study results show that participants have to either decode the response or ask for a reformulation of the response. In either case, it takes time. Hence, a system mismatching the language slows down a user in achieving the objectives. Consequently, matching the user's language improves usability partly.

Furthermore, the answer space appears to be especially effective when the user's knowledge is exceeded. The most positive reactions are collected during the analysis of the multiple comparison. This analysis requires deep knowledge from the participants, although it is a common challenge. However, many participants welcome the responses of the system. The number of reformulations as well as the direction of the reformulations support this hypothesis. In the comparison situation, many participants prefer an explanatory response with lower information. In analyzing the dependencies between two quantitative attributes, however, many participants tend to ask for responses with more information.

Finding 2: Matching the user's language increases trust. As shown in the results of Section 5.6 as well as addressed by Srinivasan et al. (2019a), experienced users tend to lack trust in provided data facts with missing parameters. The results of this experiment highlight a positive effect in that direction. According to statements of experienced participants, matching the user's language likely supports to gain trust in the system. These experienced participants are positively surprised by the system when they see what kind of methods have been used. In case this method is reasonable, they likely put more trust in the system's responses. Again, it requires the answer space to provide that kind of communication, since an inexperienced user would likely be overwhelmed by receiving this detailed information.

Additionally, the option to ask for a reformulation of the given response supports trust, while a system without this functionality cannot recover from making a wrong prediction. The participants use of this functionality further shows it usefulness for adjusting the language in visual analysis. These corrections of the given responses likely support to eventually increase the performance of the predictive model as well.

Finding 3: A mutual introduction can help to conveniently learn about the user. Through the implementation of the answer space via a predictive model, the user's information has to be somehow collected. The design element of the mutual introduction is added to the Valletto prototype. Although the mutual introduction is motivated from a communication point of view, it was unclear how the participant would react. Generally, the participant accept a mutual introduction when they receive a benefit from it. However, the participants also address that the number of questions is already an effort for them. Hence, the questions asked by the system need to be carefully selected.

Finding 4: The granularity of the answer space is sufficient. Research question 9 addresses the granularity of the answer space. Initially, both dimensions of the answer space have only two values each. However, the participants provided only a few alternative response focusing on the wording but not on the answer space dimensions. Due to these results, the granularity of the answer space is sufficient for communicating dialogue acts.

6.7 Limitations

Both conducted experiments highlight advantages of the answer space for visual analysis. While the preferences in the answer strongly depend on the user, adapting to the user helps to improve user experience. However, there are certain limitations.

First, the predictive model partly lacks performance. In order to dynamically adjust to a user during the use, the answer space needs to be implemented through a machine learning model. However, this machine learning model – a multinominal logistic regression with a tree embedding – is lacking performance, although it is the best performing evaluated architecture. As the model not always predicts the correct element from the answer space, users have to asked for a reformulation. This additional effort reduces user experience and certainly increases the analysis time of the user. However, the opportunity to ask for a reformulation is well accepted by the participants.

Second, the investigated situations leverage only one statistical method each. However, there are different methods deciding on a data fact. For instance, the correlation of two quantitative methods can be computed by at least two methods: the Spearman's rank correlations coefficient and the Pearson's correlation coefficient. Since Spearman does not assume normally distributed data, it better serves in a setting of domain-agnostic data analysis (cf. Section 5.3.3). Yet, Spearman is likely less known than Pearson. In terms of less known methods, the participants tend to prefer answers with lower information content, as seen in the *COMPARISON* situation. Hence, the preferences in the answer space would likely differ using Pearson. This would further imply that the preferences are even more depended on the user's knowledge. However, there is currently no empirical support for these arguments, although it is likely the case.

Third, an effect of the visualization on the preferences in the answer space is unknown. The participants are unable to change the visualizations to their liking. Furthermore, it is unclear whether a visualization has a statistically significant effect on the user's preferences in the answer space. For instance, the preferences might be different when a scatter plot is used compared to a regression plot while analyzing two quantitative data attributes. The conducted experiments cannot provide further insights regarding the effect of visualizations, as it would have added a bias to the study results.

6.8 Summary

Initially shown in Section 5.6 as well as by Srinivasan et al. (2019a), users likely have different preferences in how data facts are communicated. This preferences appear to have an impact on the trust in the system.

In order to approach this challenge, Section 6.3 introduces a two-dimensional answer space motivated by the use of linguistic theory. Particularly, the cooperative principle by Grice (1975) forms the base for the answer space. Based on this theoretic answer space, Section 6.4 explores the users' preferences in this answer space. The results reveal a statistically significant effect of the self-reported knowledge on the preferences. However, the results further show that the self-reported knowledge alone is not enough to determine preferences accurately.

As this experiment highlight the differences in the answer space, Section 6.6 explores how the users react in a live situations on personalized data facts. In order to enable the experiment of Section 6.6, Section 6.5 first implements the answer space through a machine learning model. The final model is selected by comparing different architectures in terms of their predictive performance. The second experiment highlights the usefulness of the answer space in practice. While implicatures trigger negative reactions, preference-aligned answers trigger positive reactions.

Considering the overall structure of this thesis, the answer space further accelerates an intelligible multimodal visual analysis. The answer space can be easily integrated in the existing framework of this thesis. Furthermore, the results provide the following design implications for any system using data facts in visual analysis:

1. Design multiple dialogue acts for communicating data facts, since the preferences in how the dialogue acts are formulated mainly depend on the user's characteristics.
2. Do not only rely on the users knowledge as a basis for the conversation design, since the knowledge of the user only partly explains the preferences in how data facts should be communicated.
3. Realize a mutual introduction, since it conveniently enables a structure for informing the user about the supported functionality as well as allowing to get to know the user.

7 Investigating Visualization Preferences

This chapter approaches a personalized visualization recommendation. Up to now, Valletto invariably recommends the most effective visualization at first glance. However, as individual preferences substantially affect the effectiveness of visualizations as well as the performance in visual analysis in general, user preferences should to be taken into account. A computational concept is proposed which combines a dueling bandit with the divide-and-conquer paradigm, in order to adjust the visualization recommendations to the user. The results of a conducted user study ($N = 15$) support the use of dueling bandits for learning visualization preferences. The bandit predicts satisfying visualizations as well as the learning procedure is not considered as a burden by the participants. Furthermore, the learned preferences can serve as prior knowledge for prospective users. A second experiment ($N = 63$) reveals indications for an effect of prior knowledge on the learning effort. Depending on the prior knowledge, the learning could be reduced which positively affects the usability.

Disclaimer: The content of the following chapter is partly published in the article:

Jan-Frederik Kassel and Michael Rohs (2019a). “Online Learning of Visualization Preferences through Dueling Bandits for Enhancing Visualization Recommendations”. In: *EuroVis 2019 - Short Papers*. Ed. by Jimmy Johansson et al. Porto, Portugal: The Eurographics Association. ISBN: 978-3-03868-090-1. DOI: 10.2312/evs.20191175

7.1 Introduction

Considering the visualization recommendation engine (see Section 5.3.2), the recommendation routine is not yet personalized. It essentially ranks the set of available visualization by the effectiveness. The user's preferences, characteristics, or wishes do not yet take part in the recommendation.

However, the effectiveness of visualizations depends not only on the data (Mackinlay, 1986), the visual analysis task (Harrison et al., 2014; Kay and Heer, 2016; Kim and Heer, 2018; Saket et al., 2018), or general perception constraints (Bertin, 1974; Kosara, 2019b; Skau and Kosara, 2016), but also on the user's diverse characteristics Conati and Maclaren (2008) and Ziemkiewicz et al. (2012). Conati et al. (2014) highlight a connection between the user's preferences for a set of specific visual mappings and the decision quality. While the visual mapping serves the user's preferences, a user decides better. Additionally, Green and Fisher (2010) reveal an effect of the user's personality (e.g. extraversion) on the user's performance in visual analysis. Hence, a user presumably achieves better results when the visual mapping serves a user's preferences.

As a user's preferences are important for facilitating the user's performance in visual analysis, the question remains how to integrate the user's visualization preferences into the visualization recommendation routine. In order to answer this question, two challenges have to be approached.

First, the visualization preferences have to be learned by the system. In the past, approaches used offline learning methodologies, e.g., RankSVMs (Moritz et al., 2019). These methodologies require previously collected data for their training. However, such kind of data sets of preference scores on visualizations is not really given in visual analysis. In fact, it is likely hard for the user to state how well a visualizations might fit the visualization preferences. Dueling bandits can likely help in this challenge. Dueling bandits essentially learn preferences by pairwise comparisons. Hence, this group of algorithms only need the user's input for approximating the preferences instead of a previously collected data set. The effectiveness of dueling bandits has been already shown in other domains, e.g., in human-robot interaction (Schneider and Kummert, 2017).

Second, the learned preferences have to be formalized and stored, as they can likely be used for supporting prospective users. As the dueling bandit approximates a user's preferences interactively, the length of the learning depends on the bandit's prior knowledge about the user. This prior knowledge could be computed by using other users' preferences. However, how this prior knowledge is actually computed needs to be investigated.

This chapter proposes a dueling bandit algorithm with a divide-and-conquer learning strategy for learning visualization preferences. Both experiments and analyses empirically reveal the effectiveness of this dueling bandit. Furthermore, machine learning methods are investigated to compute prior knowledge for the dueling bandit in order to further reduce the effort of the user. The needed information about a user are then added to the user model of Valletto as well as included into the mutual introduction, introduced in Section 6.5.3.

7.2 Related Work

In the following, the relevant concepts for learning visualization preferences are introduced. Comparing these concepts highlights advantages of using dueling bandits for learning visualization preferences. Furthermore, differences are highlighted to previous work on incorporating visualization preferences into recommender systems.

7.2.1 Learning User Preferences

Generally, learning preferences is a topic approached from various perspectives. On the one hand, the conventional approach takes advantage of supervised machine learning algorithms, e.g., RankSVM, or logistic regression. On the other hand, recent approaches use reinforcement learning methodologies.

Supervised machine learning requires previously collected and labeled data in accordance with users' preferences. For instance, the ratings of movies (Maas et al., 2011) or restaurants (Cui, 2015). Given this kind of data, supervised machine learning algorithms achieve quite well the objective of personalized recommendation. Especially the collaborative filtering algorithm is one of the preferred options for recommending items (Chen et al., 2018), although the algorithm suffers under the cold start problem. The cold start problem describes essentially the lack of performance when the data is sparse i.e., missing ratings for certain items. A way of reducing this problem is by applying active learning to the learning procedure (Zhao et al., 2013). In active learning, the algorithm includes the user into the prediction process (Elahi et al., 2016). While the algorithm first predicts the preferences for each item, it asks the user for feedback for those predictions with the highest uncertainty (Settles, 2009). Nevertheless, these approaches require still labeled data. Since this data is not existing, supervised machine learning cannot be applied to solve the recommendation challenge.

In order to overcome the obstacle of lacking labeled data, reinforcement learning methodologies have been investigated. Considering the domain of learning visualization preferences, in reinforcement learning, the user is taken into the learning loop of the algorithm. In

an interactive manner, the agent makes a prediction based on his current knowledge and observes the user's corresponding reactions. Based on the user's provided feedback, the agent updates the prediction policy and continues by predicting the next item (Wilson et al., 2012). Given this learning approach, the agent interactively learns the user's preferences. Because of this interactive learning procedure, the term of online learning is commonly used. But the question still remains how the user's feedback should look like.

One implementations of the reinforcement learning paradigms is the multi-armed bandit (Bouneffouf and Rish, 2019). A multi-armed bandit has a set of k different items. At each time step, a multi-armed bandit predicts one of those items. However, the essential challenge for the bandit is to select the item from which he learns the most. Formally, this challenge is described as the exploration exploitation dilemma:

> "It has to find a reasonable compromise between playing the arms that produced high rewards in the past (exploitation) and trying other, possibly even better arms the (expected) reward of which is not precisely known so far (exploration)." (Busa-Fekete et al., 2018)

Nevertheless, a multi-armed bandit requires quantitative feedback for the predicted item from the user. Consequently, a user should specify how good the visualizations fits the preferences, e.g., by using a 5 star rating scale.

As providing quantitative feedback might be challenging for the user, dueling bandits have been proposed. A dueling bandit is a special case of the multi-armed bandits (Yue et al., 2012). While a dueling bandit owns a set of different actions to choose from as well, the feedback for updating the policy looks different. A dueling bandit gets only binary feedback either 0 or 1 on in accordance with whether the chosen action is good, since it always offers two actions for the given situation. These two options are approaching the exploration-exploitation dilemma. One option represents the exploration and the other represents the exploitation option. Based on the user's selection for a preferred action, the feedback for the selected action is 1 and for the other action it is 0. For the user, it is much easier to make a decision while comparing two options than to solely describe how good an offered action is. Hence, dueling bandits generally serve well in online learning of preferences (Busa-Fekete et al., 2018).

Another variant is co-active learning. In co-active learning (Shivaswamy and Joachims, 2012), the agent proposes a solution for a given problem, e.g., a fully specified visualization for a set of given data attributes. Now, the user needs to adjust this visualization in accordance with the user's visualization preferences. Based on these adjustments, the agent learns how a visualization should ideally look like by accordingly updating the prediction policy. Through a sequence of multiple iteration and adjustments by the user, the agent eventually

approximates the user's preferences. Co-active learning requires detailed feedback. While an expert is likely able to properly adjust a visualization, a novice user likely is not (Grammel et al., 2010).

While reinforcement learning has advantages, the effort for the user is surely higher compared to supervised learning approaches. Still, the issue is a lack of labeled data on the user preferences in visual analysis. Well-labeled data sets containing both user information and their ratings for different visualization simply do not exit or are at least not available. In fact, using reinforcement learning instead of supervised machine learning might even increase the level of personalization in visual analysis, as the system only considers the individual feedback of the user.

7.2.2 Application of Dueling Bandits

Yue et al. (2012) propose a dueling bandit approach for improving the results of search engines. In each learning step, two rankings of links are shown for a given query. Based on the decisions for a preferred ranking, the bandit interactively learns how a user-specific ranking should look like.

Schneider and Kummert (2017) explore the potentials of a dueling bandit to learn preferences in Human-Robot interaction. The authors leverage the Double Thompson Sampling (D-TS) algorithm by Wu and Liu (2016) to learn the preferences. Based on their analysis, the dueling bandit algorithm helps to adjust the behavior of robots according to given preferences.

In massive open online courses (MOOCs), a lot of assignments have to be reviewed. In order to be more efficient in terms of budget, Chan et al. (2016) propose a dueling bandit algorithm for ranking the assignments according to the reliability of the students. The authors take advantage of the dueling bandit ranking algorithm of Busa-Fekete et al. (2014).

Sui and Burdick (2017) develop a correlational dueling bandit for clinical treatment. The application is about "clinical research for recovering motor function after severe spinal cord injury"(Sui and Burdick, 2017). The objective is to setup a treatment personalized to the patient. In this application, the items (individual electrical stimulation pattern) to chose from are correlated which each other. Hence, the authors propose on dueling bandit algorithm for handling these depending arms.

In the domain of NLP, Sokolov et al. (2016) investigate potentials of dueling bandits. Due to the ambiguity of speech and text, it might be hard to say how good an output actually is from a machine translation, sequence labeling, text classification service. Sokolov et al. (2016)

show that a dueling bandit can be used to further improve a statistical machine translation service.

7.2.3 Learning Preferences in Visualization Recommendation

As discussed in Section 2.4, the majority of the visualization recommender approaches uses the visualization knowledge extracted from various effectiveness studies. However, some approaches conceptually take the user's preferences into account.

VizDeck (Key et al., 2012) organizes the ranked visualization through the metaphor of a card desk. By flipping through this desk of visualizations, the user votes for specific fully-specified visualization. VizDeck uses these votes to predict the preferences by considering the data sets statistics, e.g., number of distinct values from a categorical data attribute.

Mutlu et al. (2015) collect rankings of visualizations through a user study on MTurk. Based on these collected data set, the authors implement a collaborative filtering-based recommender, named VizRec. As previously discussed, collaborative filtering requires a large data set in order to achieve good results. Furthermore, it suffers under the cold start problem. However, the collaborative filtering becomes certainly effective the longer it is used.

Moritz et al. (2019) use data from effectiveness studies of Kim and Heer (2018) and Saket et al. (2018). They train a RankSVM (Joachims, 2002) on these data sets in order to learn the preferences of the user. However, their system, called Draco, learns the overall preferences of the populations of the user studies, but not the preferences of a specific user. Hence, two different users will still get the same ranking of visualizations, although the ranking does not only bases on the objective knowledge, anymore.

VizML (Hu et al., 2019) learns general visualization preferences as well, as it trains on published visualizations. By considering these set of visualizations and the corresponding data sets, VizML implicitly adapts its ranking to the users which created these visualizations.

In summary, this chapter's approach differs in two dimensions. First, the used learning paradigm is different. While related work consider offline learning methodologies for approximating the visualization preference, the dueling bandit is an online learning methodology. Hence, this approach does not require additional resources such as previously collected ratings on visualizations. Second, it likely achieves a better coverage of the user's preferences as the dueling bandit only takes the input from the user. In fact, it actually learns these preferences while others learn the preferences of an entire group of users (Hu et al., 2019; Moritz et al., 2019). However, the effort for the user is actually higher.

7.3 Approximating Visualization Preferences

The current recommendation procedure (see Section 5.3.2) essentially ranks the available visualizations according to effectiveness studies (Mackinlay, 1986). This ranking needs to be adjusted by the user's visualization preferences, as the preferences play an important role in visual analysis. Hence, the adjusted recommendations should eventually look like:

$$v_u = v_{rule-based} * P_u \tag{7.1}$$

with $v_{rule-based}$ the ranking only based on the effectiveness studies and P_u the user's visualization preferences. P_u is further defined as:

$$P_u^{k \times k} = [p_{i,j}]_{1 \leq i,j \leq k} = p(i \succ j) \tag{7.2}$$

where $p(i \succ j)$ describes the probability that the user u prefers visualization i to visualization j. However, the challenge is to effectively approximate these probabilities. Dueling bandits are one approach to learn preferences both interactively and effectively, as previously discussed.

Generally, a dueling bandit tries to approximate the preferences by selecting two items from a set of k different items (Yue et al., 2012). Based on these two items shown, the user needs to make a decision on which item is the most preferred one in this moment. Conducting a sequence of these pairwise comparisons eventually leads to a situation in which the user's preferences are well understood by the dueling bandit. In order to practically learn the preferences, different learning strategies are proposed over the years (Busa-Fekete et al., 2018). However, independent from the chosen learning strategy, the length of the sequence of needed comparisons highly depends on the set of items. In this thesis, the number of different items are the number of unique visualization. Precisely, this is $|v_{rule-based}|$. Hence, the number of unique pairwise comparisons is:

$$\binom{|v_{rule-based}|}{2} \tag{7.3}$$

Depending on the number of available visualizations, the number of unique comparisons tremendously increases. Imagine, only five different visualizations should be recommended. This would lead to 10 unique comparisons. For 10 visualization it would be 45 and for 40 it would be 780. Furthermore, the dueling bandit needs to see item pairs multiple times in

order to actually learn the preference. If the bandit would only see each item pair once, it would be similar to flipping a coin.

This circumstance raises a serious usability issue for applying dueling bandits for recommending visualizations. A user likely does not want to conduct an endless sequence of decisions to eventually get preference-aligned visualizations. Hence, the number of needed comparisons for approximating the preferences needs to be as low as possible. Therefore, the following section discuss a trade off on how the usability can be maintained while the dueling bandit is still able to learn the user's preferences.

7.3.1 Learning Preferences by Divide-and-Conquer

In order to reduce the number of needed comparisons while simultaneously empower the bandit to approximate the user's preference, a divide and conquer algorithm design paradigm is applied.

Learning

Currently, the recommendation procedure ranks fully specified visualizations. However, each visualization is described by various visualization features, e.g., visual mapping or visualization type. The set of available visualizations depends on both the number of different visualization features as well as the number of different visualization feature values. Conceptually, a fully specified visualization can be essentially seen as a kind of linear combination of their different feature values. Since the number of different values is tremendously lower for each individual visualization feature than for fully specified visualizations, the number of needed comparisons can be decreased by learning the preferences not on fully specified visualizations, but on features:

$$\binom{|v_{rule-based}|}{2} > \binom{|F_1|}{2} + \ldots + \binom{|F_n|}{2} \tag{7.4}$$

with the visualization features $\{F_1, \ldots, F_n\}$ and $|F_i|$ the number of different features values of feature i. Hence, the dueling bandit learns one preference matrix for each feature, instead of one preference matrix for the available visualizations.

The objective remains to individually learn a preference matrix for each feature. This basically leads to a three step algorithm (see Figure 7.1). In the first, the bandit selects a feature for the iteration. Second, a pair of two feature values are selected. Third, and finally,

two fully specified visualizations needs to be chosen which respectively represent the two selected feature values.

As the bandit can only evaluate one pair per iteration, the bandit needs to select one feature per iteration. However, the cardinality among the features varies. Considering the dueling bandit as a processing unit and the preference matrices as processes, the selection of a feature to play is similar as the scheduling algorithm of a Central Processing Unit (CPU). The CPU works under the challenge to provide every process the amount of time it needs. Hence, similar algorithms are also applicable in this context.

Given a selected feature for the next iteration, the dueling bandit chooses two feature values from this selected feature. How these feature values are selected highly depending on the algorithms learning strategy (Busa-Fekete et al., 2018). The preferred learning strategy is discussed in the Section 7.3.2. Nevertheless, the proposed divide and conquer approach serves every dueling bandit algorithm, as it focuses on the overall structure and abstracts from the actual pair selection process.

Imagine, two feature values f_a and f_b are selected from the feature F_i by the dueling bandit as promising candidates for the next iteration. The objective is now to choose two maximally similar fully specified visualizations. These visualization should be identical except in the selected feature F_i. A fully specified visualization is defined as a vector in which the value at the first index represents the value of feature F_1 and so on. Given this definition, two sets of visualizations are constructed:

$$V_a = \{v | v = (\ldots, F_i = f_a, \ldots)\} \tag{7.5}$$

$$V_b = \{v | v = (\ldots, F_i = f_b, \ldots)\} \tag{7.6}$$

Hence, V_a contains all visualization which incorporate the feature value f_a, while V_b contains all visualization which incorporate the feature value f_b. Furthermore, these two sets are disjoint: $V_a \cap V_b = \emptyset$. In the next step, all candidate pairs are generated. The elements of these pairs have to be maximally similar. The highest similarity score is determined by computing the similarity between all possible pairs:

$$\delta_{max} = \max_{v_i \in V_a, v_j \in V_b} \{sim(v_i, v_j)\} \tag{7.7}$$

Afterwards, the set of potentially candidates is finally computed:

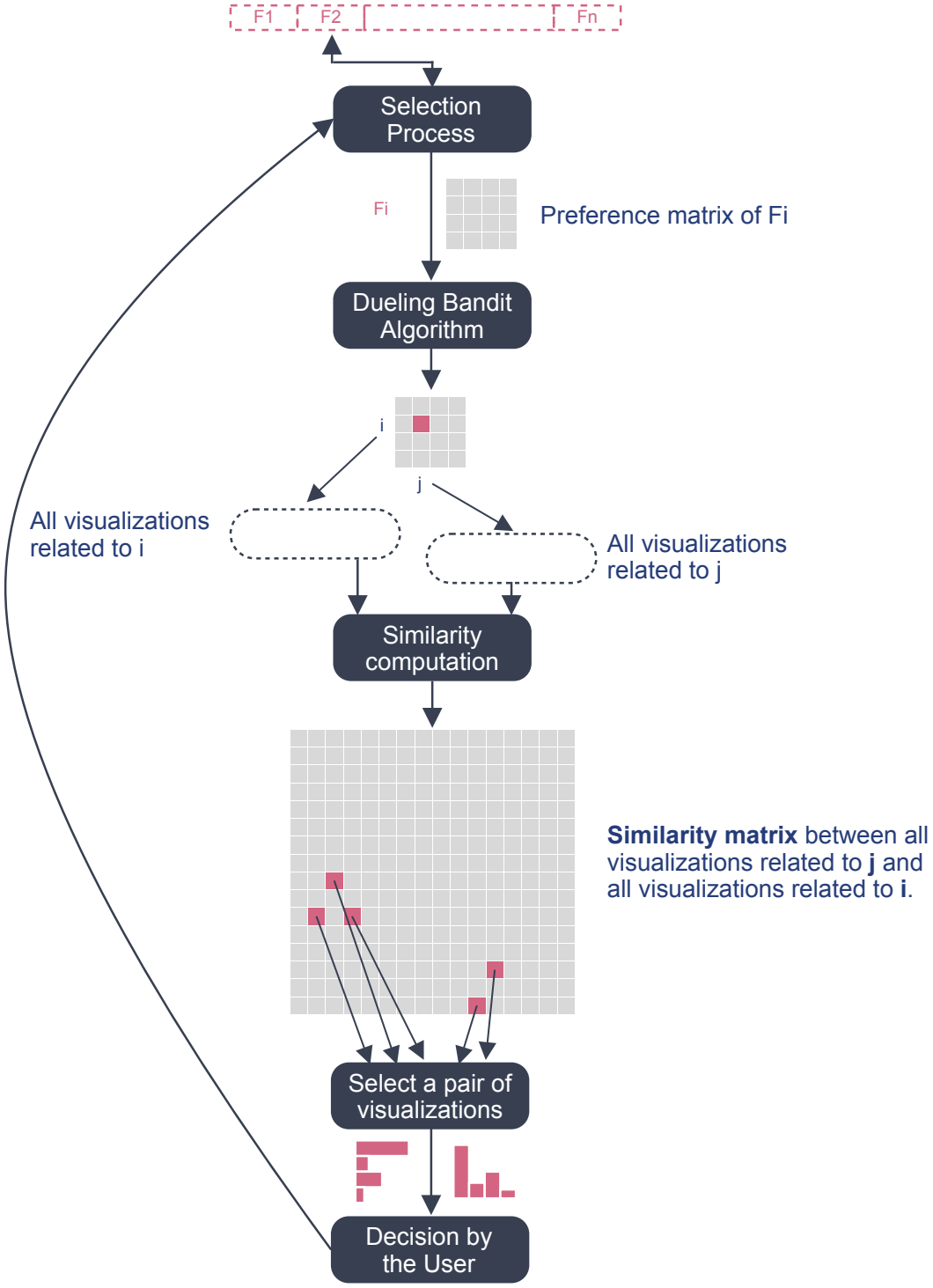


Fig. 7.1: Illustration of the divide and conquer design for the online learning of the visualization preferences.

$$V' = \{(v_i, v_j) | \delta_{max} = sim(v_i, v_j)\} \tag{7.8}$$

However, the cardinality of V' might be greater than one. Hence, the pair for the duel is chosen by sampling from this set while assuming that each pair is equally useful:

$$(a', b') \sim U(V') \tag{7.9}$$

This algorithm defines a framework for learning the user's preferences on visualization feature level. However, it does not prescribe the learning strategy for selecting duel candidates, the similarity measure between the visualizations nor a scheduling algorithm for selecting a visualization feature for the next iteration.

Prediction

In the learning phase of the dueling bandit, the algorithm always has to select two options for the comparisons. However, in the prediction phase, this is not the case. Instead, the algorithm needs to select only one value for each feature individually. These values have to be selected in accordance with the user's previously learned preferences. Hence, for each feature, the bandits chooses the values in the same manners as it selects the first value for the comparison (exploitation).

However, not all feature value combinations are valid visualizations. Imagine, the selected visual mapping prescribes to visualize a categorical attribute through a coloring. The coloring schema needs to be effective for this attribute. Hence, a diverging or sequential coloring schema has to be selected, all other options are not effective.

As this example demonstrates restrictions on the feature values, predicting a fully specified visualization requires to sequentially select the feature values. To do so, the visualization features are sorted by their assumed importance for the users. This sorting results in the following order: First, the likely preferred visual mapping is selected by the bandit, followed by the visualization type, and finally the coloring schema.

The prediction algorithm handles the caused restrictions on the coloring schema by only considering a submatrix of the learned preferences. Only the valid feature values (coloring schemes) are compared with each other in order to select the likely most preferred coloring schema for the to be predicted visualization.

Finally, the algorithm returns a fully specified visualization to the user. This prediction procedure represents a kind of greedy behavior, since the sequential feature value selection process potentially causes combinations which do not have to be necessarily the overall most preferred option.

7.3.2 Configuration

According to the previously defined algorithm, three elements needs to be precisely selected: the scheduling process for the visualization feature selection, the actual dueling bandit algorithm, and the similarity measure between the fully specified visualizations.

As a scheduling process, the round robin scheduling is selected. In round robin, each visualization feature gets the same about of iterations. As round robin treats every feature equal, it also has some disadvantages. Given the different cardinality of the visualization features, the dueling bandit algorithm might need less comparisons for actually approximating the preferences on this feature. Hence, scheduling algorithms which consider either the cardinality of a visualization feature or its importance to the user might also serve well.

Regarding the selection of a dueling bandit algorithm, two constraints have to be considered. First, the learning strategy needs to be very efficient. A user likely wants to keep the effort low for teaching the system on what a preferred visualization is. Second, there is presumably no overall preferred visualization, but likely a set of equally preferred visualizations. In fact, it likely depends on the user whether a total order of visualizations exists.

According to the analysis of Busa-Fekete et al. (2018), learning strategies of dueling bandits can be generally separated in two groups. One group of algorithms follows a Coherent winner strategy (Busa-Fekete et al., 2018). This strategy assumes a total order of all items according to the preferences. Hence, if item i is preferred to item j and item j is preferred to item k than the user prefers i to j as well. The other group of algorithm does not assume a total order of all items. This strategy is called Copeland winner (Zoghi et al., 2015) where a set of items can be equally preferred.

Both constraints are fulfilled by the D-TS algorithm developed by Wu and Liu (2016). The D-TS algorithm counts for each item pair (i, j) how often item i was preferred to item j and stores this information in a counting matrix:

$$B^{k \times k} = [b_{i,j}]_{1 \leq i,j \leq k} = \#(i \succ j) \tag{7.10}$$

These counts form the base for selecting two items for the comparison. The D-TS chooses two items for the comparisons by sampling twice from Beta distributions (Thompson sampling (Thompson, 1933)) which base on the counting matrix:

$$\theta_{i,j} \sim Beta(b_{i,j}, b_{j,i}), i < j \tag{7.11}$$

where $\theta_{i,j}$ represents the probability that i is preferred to j while $\theta_{j,i} = 1 - \theta_{i,j}$. Due to the nature of the Beta distribution, the more often i was preferred to j in the previous comparisons, a high probability that i wins in the next comparison is more likely. Additionally, the more comparisons are generally made between i and j irrespective of which items was preferred, the expected value of the corresponding Beta distribution becomes more stable. Hence, the D-TS algorithm eventually achieves a status in which the user's preferences are well approximated.

For selecting the first items for the comparison, the D-TS algorithm computes the Copeland winner based on the samples θ. Consequently, the item is selected which probably is preferred to the most other items. Let assume this item is named c. In a second step, the D-TS algorithm samples again from the Beta distributions but limited to the $\theta_{c,j}$s. However, the algorithm only considers uncertain pairs for the actual selection of the second item. Hence, the first item for the comparison represents the exploitation option while the second item represents the exploration option. Wu and Liu (2016) show that their D-TS algorithm needs significantly less comparisons than other dueling bandit algorithms in order to approximate the preferences.

Finally, the Hamming similarity is selected as a similarity measure between fully specified visualizations. This similarity measure should effectively work for two reasons. First, the Hamming similarity primarily punishes inequality between the compared elements. Since the objective of the divide and conquer learning algorithm is to selected maximally similar visualizations for the comparisons, the Hamming similarity sounds like an effective fit. Second, it works for categorical features which the visualization features are.

7.4 Experiment 1: Effectiveness and Acceptance

The experiment investigates potential effects of the dueling bandit as well as the acceptance of the participants. Precisely, the primary research questions are to investigate:

RQ 10: Can a divide-and-conquer-based dueling bandit approach effectively learn individual visualization preferences?

RQ 11: What are the participants' reactions and feedback concerning the interactive learning procedure?

7.4.1 Apparatus

In order to adequately investigate the objectives, a technical prototype is created (see Figure 7.2). Essentially, this prototype implements a similar architecture as well as uses a similar technology stack as Valletto (see Section 5.4). Consequently, the frontend uses React (Facebook Inc, 2013) and the visualizations base on Vega-lite (VanderPlas et al., 2018; Satyanarayan et al., 2017).

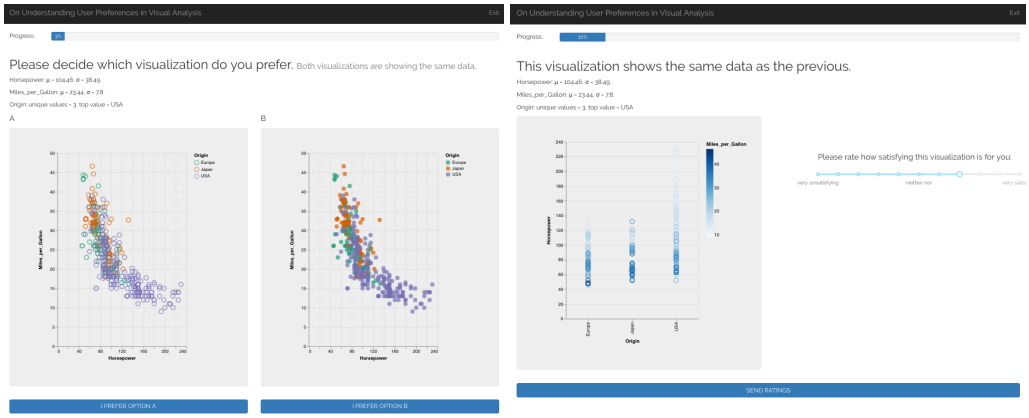
The dueling bandit algorithm follows the same setup of parameter configurations as used by Wu and Liu (2016) and Zoghi et al. (2015). Additionally, no prior knowledge about a participant's visualization preferences is added. Hence, all visualizations are equally preferred and mathematically represented by $B_{t=0} = 0^{k,k}$.

7.4.2 Procedure

The experiment consists of three phases. The first two phases base on the car data set (Donoho and Ramos, 1982; StatLib, 2005) limited to the data attributes "horsepower", "miles per gallon", and "origin". The third phase uses a weather data set. This differentiation allows to potentially derive further insights on the generalizability of the learning preferences. Since all visualizations are automatically generated, the available visual mapping options are limited. Precisely, the visual channels *x, y*, and *color* can be used. For coloring, two schemes for both categorical and quantitative exists. In addition to these options, three different mark types exist. Overall, this setup results in a visualization space of 36 different visualizations.

Initially, each participant receives a standardized introduction to the procedure's structure. Afterwards, the first phase of the study starts. Within this phase, each participant completes a sequence of 21 pairwise comparisons. Each comparisons shows two visualizations generated by the bandit (see Figure 7.2a). The display order is randomized in order to reduce biases in the data, although the display order should not affect the bandit's learning performance from a mathematical point of view (Wu and Liu, 2016). In each comparison, a participant eventually decides which visualization is subjectively more preferred.

After each 21 comparisons, the second phase starts. Now, the bandit predicts only one visualization in according to the participant's previous decisions. According to the satisfaction with this predicted visualization, the participant selects a rating on a 11-point scale from "very unsatisfying" till "very satisfying" (see Figure 7.2b). After the participant rated the



(a) Comparison view. **(b)** Prediction view.

Fig. 7.2: Design of the technical prototype for the study.

visualization, the first phase starts again. Overall, each participants makes 210 pairwise decisions as well as rates 10 predicted visualizations in this experiment.

The third phase consists of a sequence of four pairwise comparisons. Each comparison further bases on a different set of three data attributes. While in the first phase each comparisons shows two visualizations selected by the bandit, these comparisons now consist of one visualization predicted by the bandit and one created according to the effectiveness study of Kim and Heer (2018) (named as "rule-based approach" in the following). However, a participant should decide again which visualization is more preferred while the display order is again randomized.

The experiment closes with a set of questions on the demographics, as well as experience in information visualization and statistics. Furthermore, the NASA Task Load Index (NASA-TLX) (Hart and Staveland, 1988) is included to further quantify the perceived effort for the participants.

7.4.3 Participants

15 persons with an average age of 26.4 ($\sigma = 3.2$) years participated in this experiment. They are either from academia or from industry. The participants are differently well educated in both information visualization and statistics, according to the self-reported knowledge statements. Furthermore, they create visualizations by using either professional tools, MS Excel, or a scripting language like Python or R.

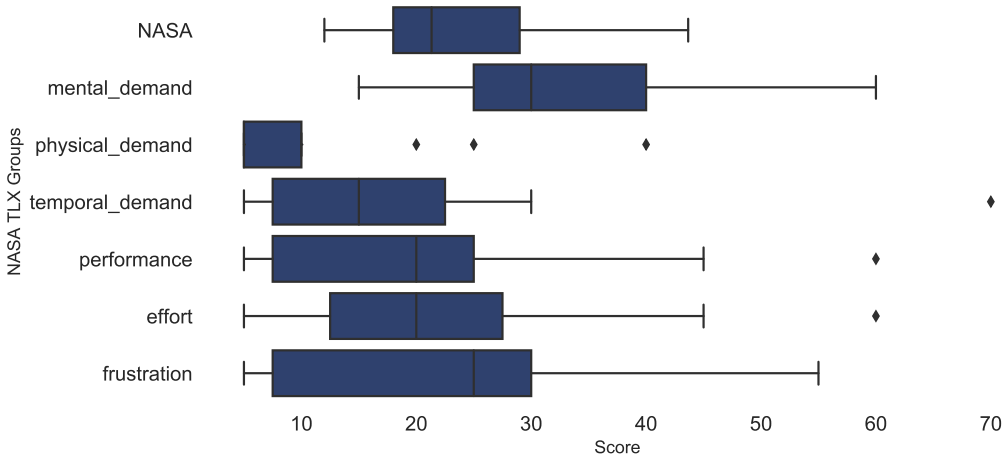


Fig. 7.3: The NASA-TLX in general as well as the individual parts.

Category	weight (μ)	weight (σ)
Mental demand	.24	.08
Physical demand	.02	.04
Temporal demand	.18	.10
Performance	.20	.08
Effort	.20	.07
Frustration	.14	.13

Tab. 7.1: Weights of the NASA Task Load Index's categories according to the participants' decisions in the pairwise comparisons of these different categories.

7.4.4 Results

The results are categorized into three areas: participants feedback on the overall procedure, analyses related to the pairwise comparisons, and analyses related to predicted visualizations. Overall, the participants need 4.4 seconds on average for making a decision. Furthermore, 14 participant have a subjective impression that the bandit actually learned their individual visualization preferences.

In addition, the overall NASA-TLX scores of 24.26 on average ($\sigma = 9.64$) reveals a little demanding learning procedure. Figure 7.3 illustrates a detailed look at the results of the NASA-TLX. The category "mental demand" has the highest scores of all categories. Furthermore, this category has the biggest stake in the overall scores, according to the weights shown in Table 7.1. Hence, the participants perceive the mental load as the most important factor for determining the work load. Figure 7.3 also shows that the participants' frustration as well as the perceived effort is low in the sequence of pairwise comparisons.

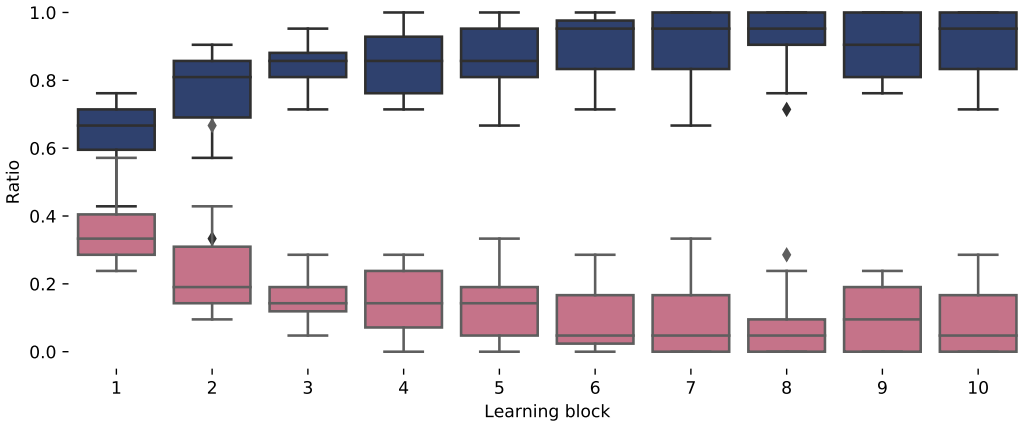
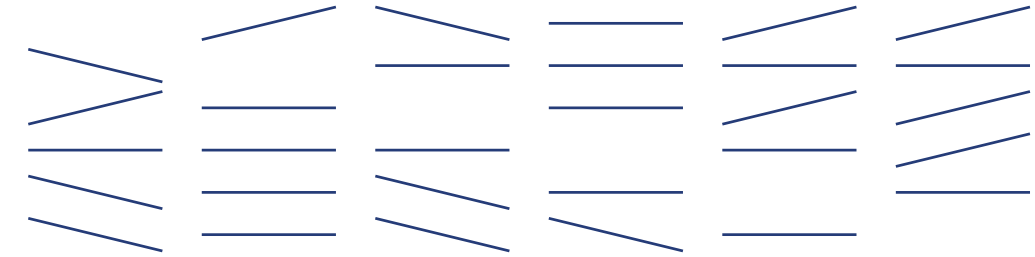


Fig. 7.4: Illustration how often the participants chose either the greedy option (blue) or the exploration option (magenta) during the learning phase.

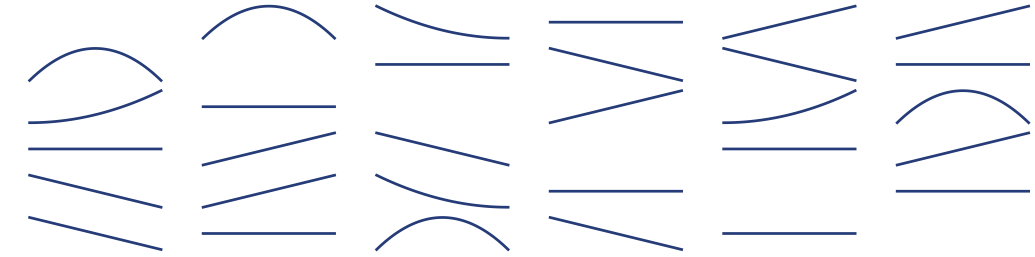
Learning Visualization Preferences

Considering the entire learning procedure, the dueling bandit becomes more certain about the user's visualization preferences. Figure 7.4 shows these circumstance on a more detailed level. The blue line represents the exploitation option i.e., the visualization which is currently the best to choose, while the red line shows the exploration option i.e., a visualization which is likely more preferred but the bandit is uncertain. In the beginning of the study, both options are more or less equally chosen. However, this changes during the study. This means that the participants predominately selected these visualizations for which the bandit assumed a high preference coverage.

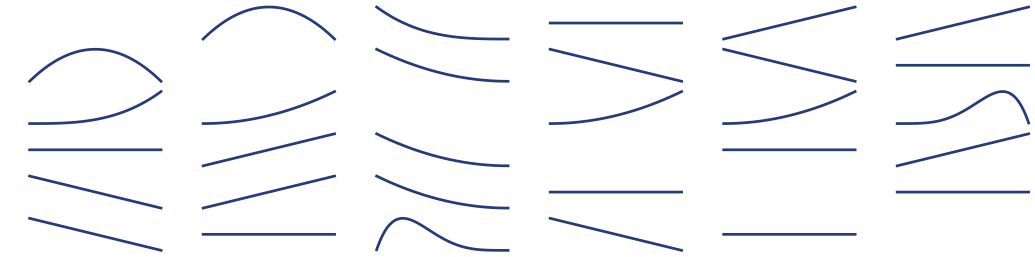
Additionally, Figure 7.5 and Figure 7.6 exemplary illustrate how the bandit actually approximates the visualization preferences of a participant. These illustrations show the Cumulative Distribution Function (CDF) for each visual mappings pair in the moment before the bandit predicts a visualization. On the one hand, the visual mappings `(x: Horsepower, y: MPG, color: Origin)` and `(x: Horsepower, y: Origin, color: MPG)` (cells $(1, 2)$ and $(2, 1)$, respectively) are equally preferred during the entire study. On the other hand, the visual mappings `(x: MPG, y: Horsepower, color: Origin)` and `(x: Origin, y: MPG, color: Horsepower)` (cells $(3, 5)$ and $(5, 3)$, respectively) are unequally preferred by the participant. Prior the first prediction, the bandit has not chosen both elements for comparison, otherwise the CDF would not look like a unified distribution. In the following, however, the participant tend to prefer visual mapping `(x: MPG, y: Horsepower, color: Origin)` more and so the CDFs change. Nevertheless, the preferences are learned on feature level, but the participants only see two fully specified visualizations.



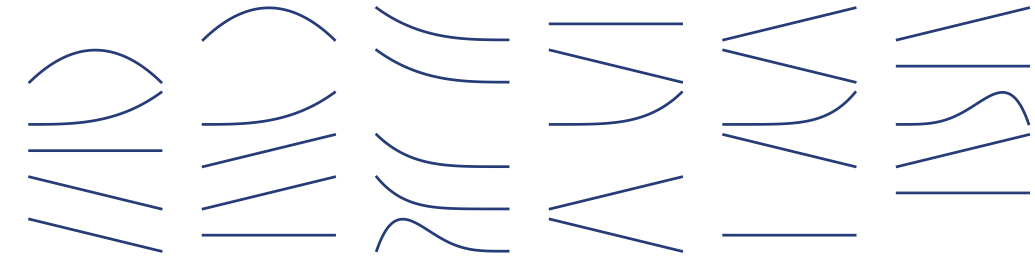
(a) 1st prediction



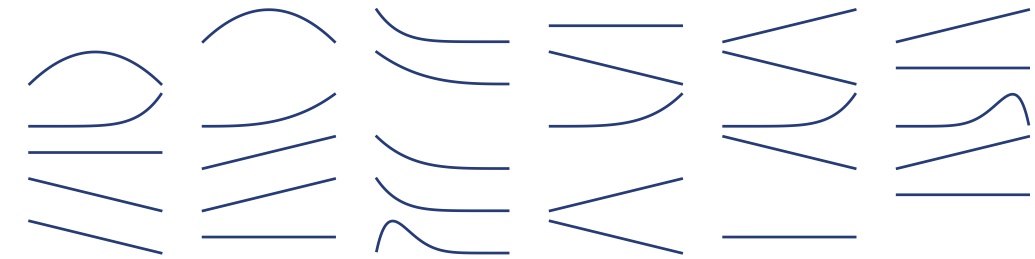
(b) 2nd prediction



(c) 3rd prediction

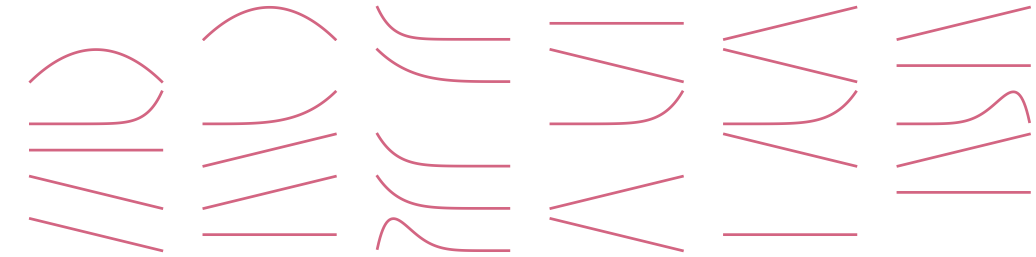


(d) 4th prediction

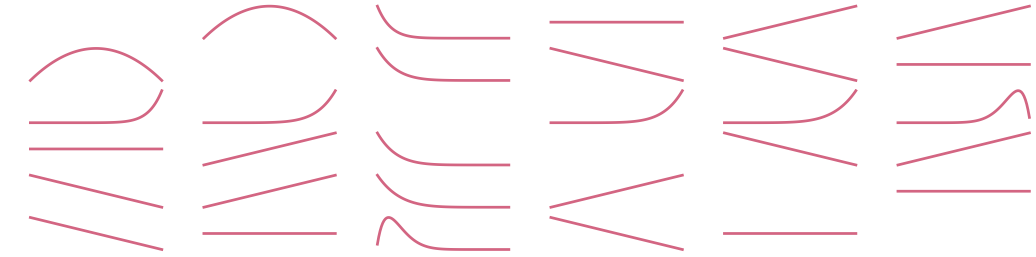


(e) 5th prediction

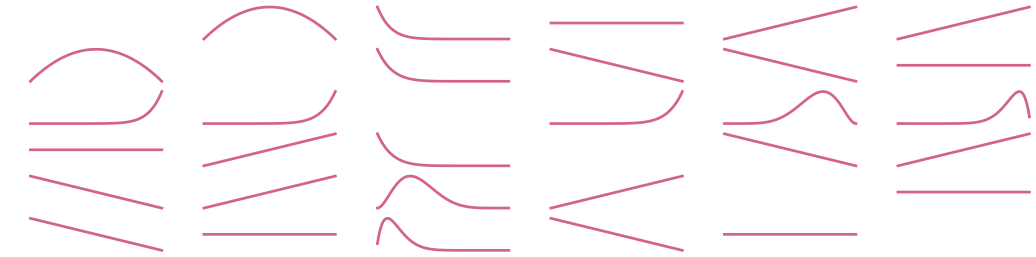
Fig. 7.5: Exemplary approximation of a participant's preferences on the visual mapping during the first five predictions. The shown cumulative distribution functions are based on the participant's decisions.



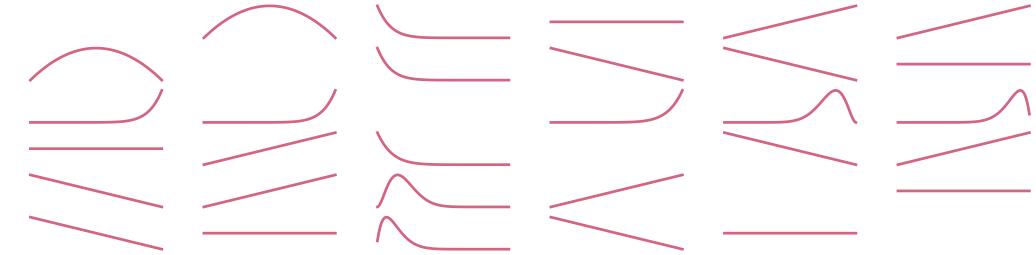
(a) 6th prediction



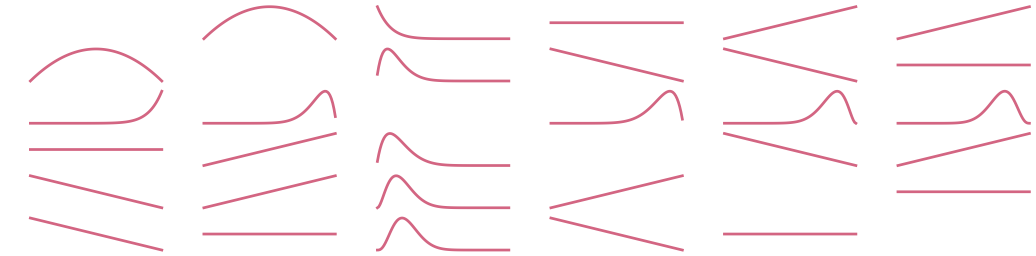
(b) 7th prediction



(c) 8th prediction



(d) 9th prediction



(e) 10th prediction

Fig. 7.6: Exemplary approximation of a participant's preferences on the visual mapping during the last five predictions. The shown cumulative distribution functions are based on the participant's decisions.

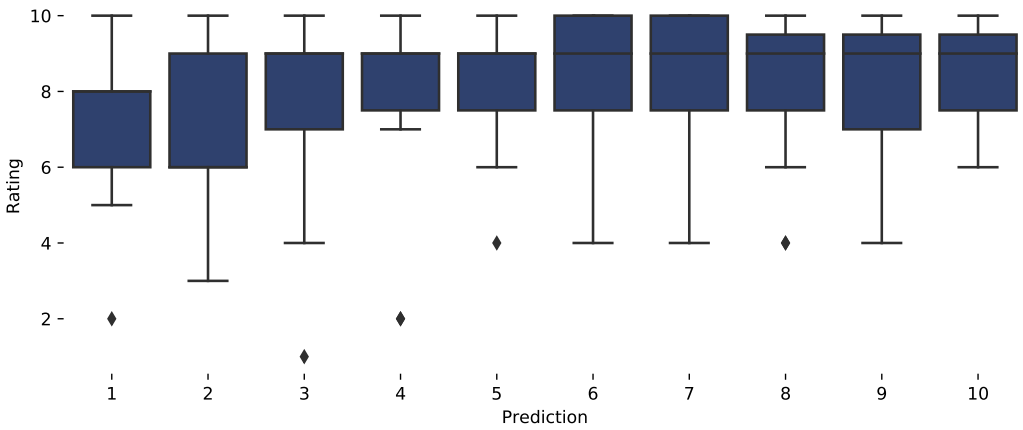


Fig. 7.7: While the learning continues, the predicted visualizations are getting higher satisfaction scores from the participants.

Finally, the participants' feedback on this learning procedure is also positive, although it appears to be a dull procedure. 11 participants would conduct such kind of learning procedure and comparison sequence, respectively, in a real system. However, they further stated to only conduct these procedure when they receive preference-aligned visualizations afterwards.

Predicting based on Visualization Preferences

Considering the participants' satisfaction ratings on the predicted visualizations reveals a similar situation as previously shown in the learning phase. Figure 7.7 highlights these ratings made by the participants. It shows the bandit's improvements overtime.

In the beginning, the bandit does not much know about a participant's visualization preferences. Hence, the ratings of the participants are relatively low. While more and more pairwise comparisons are made by the participants, the ratings of the predicted visualization increase. A conducted Wilcoxon Signed-Rank test further reveals statistically significant improvements ($Z = 92.0, p = 0.002$) in these satisfaction ratings between the first three predicted visualizations ($\mu = 7.2$, $\sigma = 2.22$) and the last three predicted visualizations ($\mu = 8.29$, $\sigma = 1.69$). Additionally, not only the overall ratings increase, but also the ratings become more stable.

Furthermore, not every possible visualization is eventually predicted. Given the set of 36 unique visualizations, the bandit selected 23 of them for the predictions. Figure 7.8 shows for these 23 different visualization both in which prediction step and to how many participants they were shown. The visualization represented by ID 7 seems to fit many participant's preferences. Other visualizations are rarely predicted, e.g., 2, 10, and 16. At the

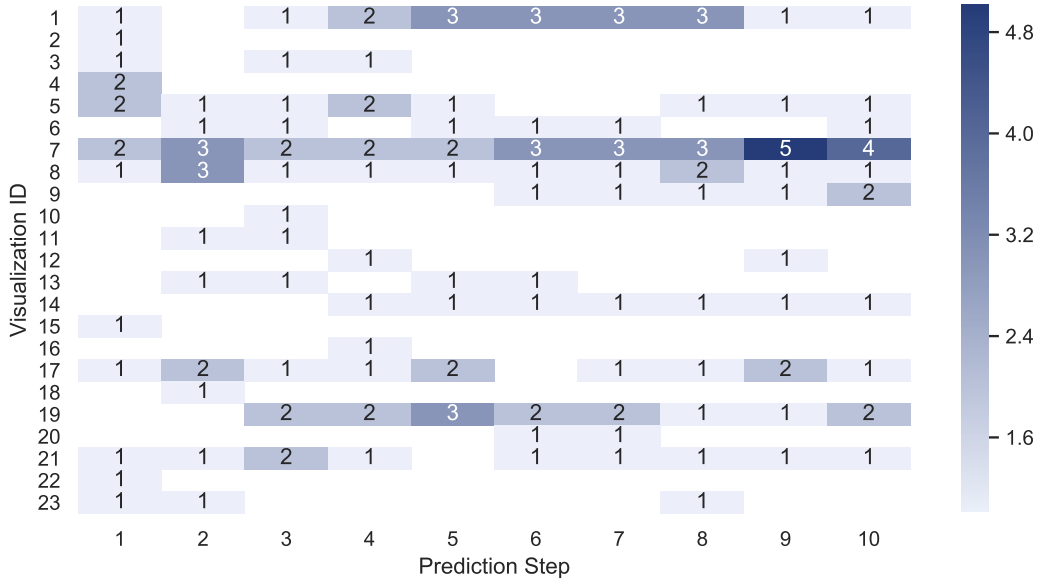


Fig. 7.8: Illustration how many participants have seen which visualization during the predictions.

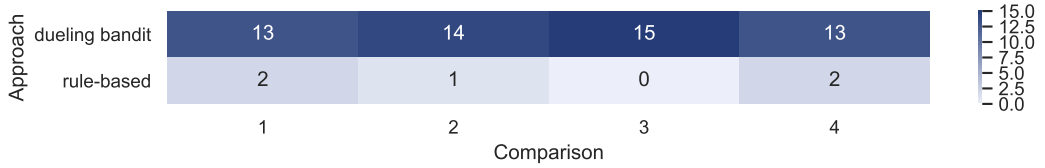


Fig. 7.9: Heatmap on the decisions during the comparisons between the dueling bandit and the rule-based approach.

last prediction step, 10 different visualizations are predicted (see Figure 7.8). The overall pairwise Hamming distance between these 10 predicted visualizations is $\mu = .57$, $\sigma = .39$. This distance value reveals that the participants preferences are quite diverse under the assumption that the last predicted visualizations are best approximating the participants' preferences. Considering all prediction steps, each participant sees on average 3.4 ($\sigma = 0.87$) different visualizations.

Comparison to the Rule-based Approach

Compared to the rule-based approach – the visualizations recommended by the effectiveness study results of Kim and Heer (2018) – the bandit is in 91% of all cases preferred by the participants. As each comparison further represents a different data attribute set, Figure 7.9 reveals the preferred approach in each comparison. However, the visualization predicted by the bandit share a similar visual mapping in 20% of the comparisons.

7.4.5 Discussion

The results provide empirical insights on the learning behavior of the dueling bandit and on the participants interaction with the bandit. They show the effectiveness of the dueling bandit to actually learn the visualization preference of a user. As the objectives of this experiment further refer to eventually enrich the recommendation routine by a personalized component, the results support the use of dueling bandits for personalizing visualization recommendations.

Finding 1: The training time can be reduced. Prior to this experiment, it was unclear how many comparisons are actually needed in order to properly approximate the user's visualization preferences. The results empirically exhibit potentials for significantly reducing these needed comparisons. On the one hand, the participants' satisfaction with the predicted visualizations veritably stabilizes after the fourth prediction (cf. Figure 7.7). This circumstances arguably supports the hypothesis of well-learned visualization preferences after half of the planed comparisons.

On the other hand, the participants tend to select the exploitation option after approximately 100 comparisons (cf. Figure 7.4). Hence, the bandit already covers the preferred visualization feature values after 100 comparisons. Both indicators provide evidence to further reduce the needed comparisons for approximating visualization preferences. Hence, the effort for the user further decreases.

Finding 2: Factorization affects the learning. An initial design decision for keeping the effort low for the user is to learn visualization preferences on feature level. However, the participants eventually decide on fully specified visualizations. Although the method is designed to maximally reduce the differences between the two shown visualizations in the comparisons (cf. Section 7.3.1), potential risks in learning the wrong visualization preferences cannot be completely ruled out. For instance, a specific mark type of the visualization might be the most preferred option but only in combination with a certain coloring schema, and not in general. This potential situation would corrupt the learning procedure, since the participant decide not only on the bases of the difference in the mark type, but by considering additional factors. However, the participants satisfaction with the predicted visualization is high. Hence, there are empirical arguments which support the idea of learning the visualization preferences on feature level in order to reduce the effort for the user.

Finding 3: Preferences can be generalized. The training time reduction reduces the burden for the user to actually achieve preference-aligned visualizations. Additionally, the effort for the user can be further reduced, as the preferences are potentially generalizable. Generalizable means in this context to transfer the learned visualization preferences from one

to another data set. The participants' decisions in the comparisons between the bandit and the rule-based approach provide support for this hypothesis. The visualization preferences are learned on the car data set, but reused for the weather data set. These data set share the structure as both are real world data set with quantitative and categorical attributes, but they differ in the context. Hence, potentials for reusing the learned visualization preferences for a new data exist, when they share same characteristics from a data point of view.

Finding 4: Similarities in the visualization preferences exist. In the conducted experiment, the dueling bandit algorithm has no prior knowledge on a participant's preferences. For each participant, hence, the dueling bandit learns the preferences from scratch. As previously discussed, the dueling bandit is effectively able to approximate the participants' preferences during the experiment even without prior knowledge. However, the results show certain similarities in the learned preferences (cf. Figure 7.8), although the learned preferences are generally quite diverse. Nevertheless, certain participants prefer similar sets of visualizations.

Like in other recommendation approaches, these common preferences should be systematically used. Unlike other recommendation approaches, however, the similarities should not be used to directly recommend visualizations (items), but to use the approximated preferences as prior knowledge for the dueling bandit algorithm. This method has two advantages. First, it further reduces the effort for the user and increase the usability of the dueling bandit learning procedure, respectively. Second, it is still flexible enough to allow adjustment to the actual users preferences. However, the questions remains on finding an effective method for systematically describing the learned preferences.

7.5 Modelling Prior Knowledge

The objective of this section is to investigate an effective way for modeling prior knowledge for the dueling bandit in order to further reduce the effort for new users. Overall, the use of other users information is a common approach to compute recommendations, either implicitly by the system itself (Chen et al., 2018) or explicitly with stereotypes (Rich, 1998). The corresponding underlying assumption is that similar users have similar preferences ().

However, the focus is not on directly recommending visualizations, but on approximating the preferences itself. Hence, the modeling of the preferences does not response in a probability vector of the different visualizations, but in a matrix added to the dueling bandit, referred as prior knowledge B_u. Generally, the research question is:

RQ 12: Can prior knowledge for the dueling bandit be modeled by a machine learning model?

7.5.1 Data

As prior knowledge for a prospective user should be modelled, the data from the previous experiment serves the analysis. The data consists of the participant information collected in the questionnaire, but without the answers to the NASA-TLX as well as the rating of different feature importance. These filtered attributes partly overlap with the information gathered for predicting preferred answers of the answer space. This restriction further contributes to the usability and acceptance of the mutual introduction, as the participant mentioned a certain burden of this introduction.

7.5.2 Methodologies

In the following, two methods are investigated for modelling the prior knowledge. First, clustering on the users represents an unsupervised learning approach. The clustering is externally evaluated considering both intra and inter cluster distances of the users' preferences. Second, a multi-task lasso regression represents a supervised learning approach. This method is evaluated based on how well it models the relations within the preference matrix. However, both approach assume that similar users have similar preferences.

In order to effectively use the user's characteristics, corresponding questions are added to the mutual introduction. As the user model already stores informative user characteristics describing preferences in the answer space (cf. Section 6.5.3), it will be extended by relevant factors of this section.

Clustering

In the clustering approach, users are clustered based on their features (e.g., statistical knowledge), but not on their preferences. This is because the system initially receives the user's characteristics through a mutual introduction, but does not know the preferences beforehand.

While a variety of clustering algorithms exists, the K-Means++ and the Affinity Propagation are chosen. K-Means++ represents a bottom-up approach where the number of clusters needs to be manually determine. Affinity Propagation represents a top-down approach which automatically determines the number of clusters. Additionally, a common approach is to combine a clustering with a decomposition analysis. This decomposition reduces the feature space (in this case the information about the users) by transforming the features into a low dimensional space. Conducting a clustering on a low dimensional feature space likely

improves the clustering results. Especially when the data is sparse, it can be the case that the data points are widely distributed in the feature space.

As the clustering describes relevant known users by predicting a cluster for a user given, the prior knowledge for this new user is:

$$B_u = \sum_{v \in \phi(u)} B_v \tag{7.12}$$

, where $\phi(u)$ is the clustering resulting in set of similar users. B_v represents the preferences of a similar user v.

As it is generally unclear whether a performed clustering actually points to a useful result, a quality measure has to be selected. This quality measure should focus on the actually objective for which the clustering is applied. It represents an external evaluation of the clustering. In this thesis, the objective is to cluster users together which are similar in their learned preferences.

The Copeland scores essentially describe these preferences. As previously discussed, the D-TS algorithm uses the Copeland scores for selecting the items for the comparison. For each matrix, the Copeland scores can be easily computed by counting how often each item potentially wins against the other items. For each matrix, a vector of Copeland scores exists. As these vectors are another representation of the learned preferences, two users are likely similar in their Copeland scores when they are also similar in their preferences. Hence, the Copeland scores serve the external evaluation of a clustering.

In order to compute the similarity between two users, the cosine distance measure fits the best. The cosine distance bases on the cosine of the angle between two vectors. As the angle is only relevant, the length of the vectors does not matter. This advantage makes the measure especially applicable for this scenario, since the represented preferences of two users do have necessarily the same length.

Evaluation of the Clustering

The data of the previous experiment shows a continuously improving dueling bandit. Each participant evaluated 10 predicted visualizations during the study. Consequently, the clustering is evaluated on exactly these steps. Additionally, for each participant three preferences matrices exist due to the divide and conquer method of the framework. Hence,

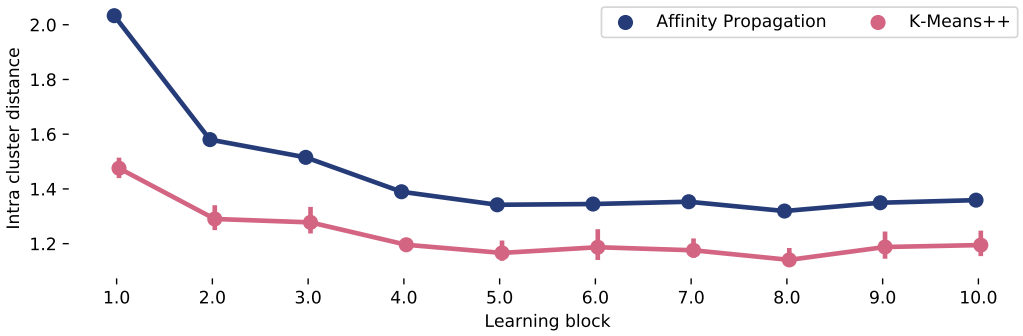


Fig. 7.10: The intra cluster distance at each step for both clustering methods.

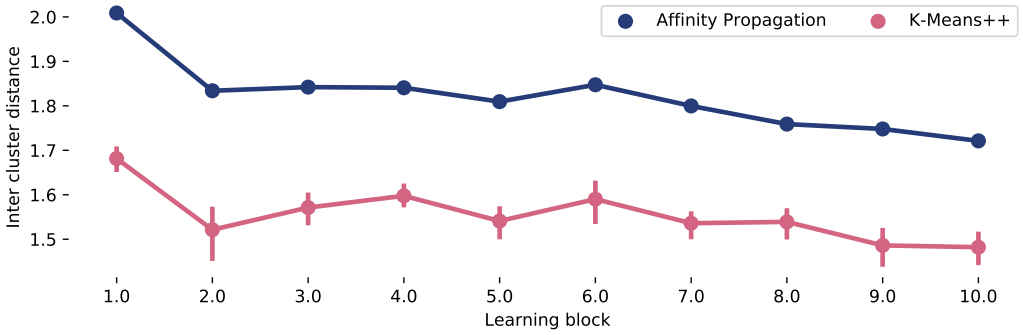


Fig. 7.11: The inter cluster distance of both the clustering methods at each step.

for each participant three Copeland score vectors exist. In order to have a single representation for each participant, these three vectors are concatenated into one single vector.

The evaluation of the various combinations of clustering with and without PCA address one configurations head of the others for each clustering approach. Both configurations use the transformed feature space of an PCA for four dimensions. The best performing K-Means++ is on three clusters, while the Affinity Propagation produces four clusters. However, only produced clustering are taken into account containing no cluster of size one.

As Figure 7.10 shows, both clustering approaches are improving overtime. The intra-cluster cosine distance between the preferences of the clustered participants is decreasing overtime. Still, the Affinity Propagation produces a more homogeneous clustering than the K-Means++, according to the higher inter cluster distance 7.11.

Furthermore, Figure 7.12 provides additional insights on the potential effect of $\phi(.)$. The baseline (grey) actually represents a solution without any clustering. Compared to the baseline, both clustering approaches perform better. However, the Affinity Propagation performs better than the K-Means++. For further comparisons, two methods are added. The top-3 and top-5 dynamically computes for a given user the average similarity of the Copeland scores to the three and five, respectively, most similar users. All methods (clustering and

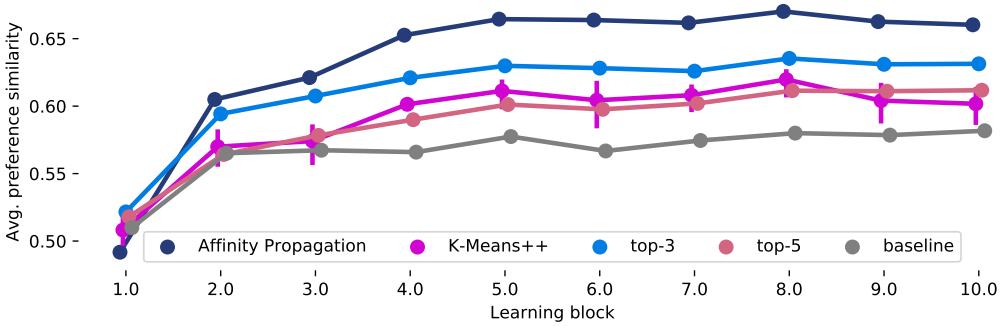


Fig. 7.12: The average preference similarity of clustering as well as baseline method at each step.

Method	1	2	3	4	5	6	7	8	9	10
top-3	1.0	3.0	4.0	6.0	5.0	6.0	5.0	6.0	5.0	5.0
top-5	1.0	-0.0	1.0	2.0	2.0	3.0	3.0	3.0	3.0	3.0
Affinity Propagation	-2.0	4.0	5.0	9.0	9.0	10.0	9.0	9.0	8.0	8.0
K-Means++	-0.0	0.0	1.0	4.0	3.0	4.0	3.0	4.0	3.0	2.0

Tab. 7.2: Differences in the average preference similarities of each method regarding the baseline.

top-n) essentially determine the set of relevant users for a given user. Table 7.2) contains the precise deltas to the baseline.

Regression

In contrast to the unsupervised learning approaches, the preference matrix can be learned through supervised learning as well. A potential method is a regression. Generally, a regression models a quantitative value given a set of features by solving a linear equation. While a linear equation can only describe one value of the preference matrix, an approach is needed which solves multiple equations in parallel as well as maintain the dependencies within the matrix. A multi-task lasso is one of those methods (Lozano and Swirszcz, 2012). This methods inherently performs feature selection. Hence, the regression directly approximates a preference matrix while the unsupervised learning method requires a two step approach to produce the prior knowledge.

Likely, the regression is not able to approximate the entire matrix correctly, as the training data is sparse. Yet, the regressed preferences should only serve as prior knowledge. Therefore, the relations between the visualizations are only relevant, but not the actual values. Considering item i is preferred over item j. A good regression produces $y_{ij} > y_{ji}$ when $b_{ij} > b_{ji}$ should hold. In order to achieve this situations, the following function transforms each regressed value in to the desired format:

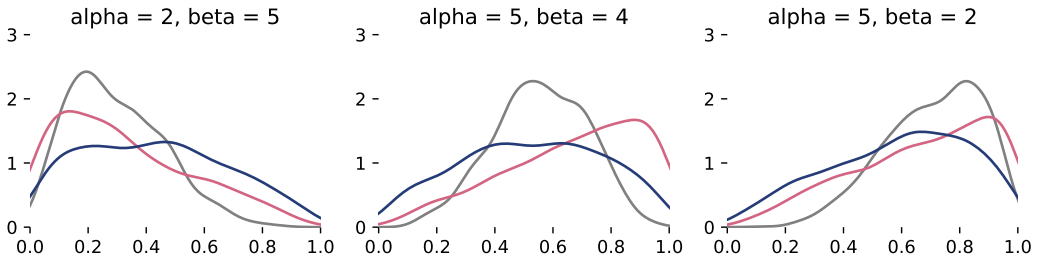


Fig. 7.13: Effect of the parameter setup on the Beta distribution. The normalized parameters (blue) approximate better the actual distribution (grey) while providing more uncertainty than the binarized parameters (red).

$$b_{ij} = \frac{y_{ij}}{y_{ij} + y_{ji}} \tag{7.13}$$

Figure 7.13 shows the difference in the resulting Beta distributions of normalized parameters, binarized parameters (if $y_{ij} > y_{ji}$ than $y_{ij} = 1$ and $y_{ji} = 0$, and vice versa), and the actual parameters. Especially in the case where both items are almost equally preferred, the normalized parameters mimic more accurate the actual relationship. Hence, they likely serve better as prior knowledge than the binarized version.

Evaluation of the Regression

In scikit-learn (Pedregosa et al., 2011), two different multi-task regressions exist: the multi-task lasso and the multi-task ElasticNet. In essence, both approach differ in the optimization objective function. The ElasticNet likely handles better situations in which the input features are correlated.

In order to evaluate the regression models, a similar approach is chosen as for the clustering evaluation. Each predicted preference matrix is transformed into a Copeland score vector. As three matrices exist per user, the three corresponding Copeland vectors are again concatenated into one vector. Now, the cosine similarity between the predicted vector and the actual vector are again computed. However, both regressions are multiple times trained by a randomized 70-30 split of the input data. Figure 7.14 shows the average similarity in the predicted Copeland scores compared to the actual Copeland scores. The multi-task ElasticNet performs slightly better at each learning block.

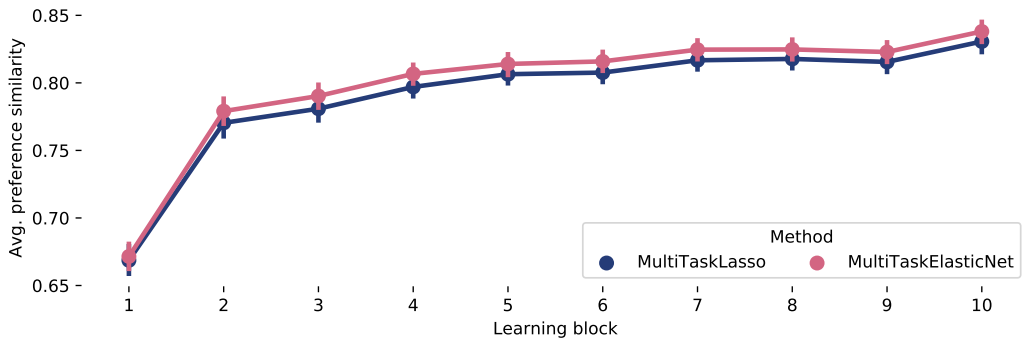


Fig. 7.14: The quality of the predicted preference matrices by the regression models.

7.5.3 Discussion

As the results show, the Affinity Propagation approach together with the PCA seems to be a good unsupervised learning perspective for computing prior knowledge. The algorithm creates more homogeneous clusters. Furthermore, the multi-task ElasticNet slightly performs better in the supervised learning perspective, as the user's characteristics are presumably slightly correlated.

However, the question remains how much prior knowledge should be used for a new user. In case too much prior knowledge would be added, the dueling bandit might stuck in a local optimum or even worse do not approximate the user's preferences at all. In case too little prior knowledge is added, the dueling bandit likely needs more comparisons to effectively approximate the user's preferences. Both circumstances would result in an unwanted higher effort for the user.

The regression focuses on maintaining the relations within the preference matrices. Therefore, the prior knowledge is represented by the relations for each item pair. However, the clustering can directly take advantage of the known preferences. In this case, an adjusting factor is needed to reduce the matrices on a potentially fitting level.

The previous experiment provides indications for a potential amount of prior knowledge. In this experiment, the highest variability in the satisfactions scores of the predicted visualizations was at the second and third step, respectively (cf. Figure 7.7). At this stage, the dueling bandit already received a time consuming amount of decisions by a participant for a preferred visualization. Additionally, the relation between the exploitation option and the exploration option is still close at this stage. Both results imply a high uncertainty in the approximated visualization preferences. Furthermore, the dueling bandits predictions stabilize after the fourth prediction in the relation between exploration and exploitation decisions as well as in the satisfaction scores of the predicted visualizations.

Nevertheless, it is still unclear whether the prior knowledge actually reduces the effort for the user while simultaneously leads to satisfying visualizations. In order to answer this question, the following experiment is conducted.

7.6 Experiment 2: Effect of Prior Knowledge

Generally, Rashid et al. (2002) propose the dimension *user effort* and *accuracy* to evaluate how well a recommender system can handle new users. Therefore, this experiment's objective is two-fold. The primary focus is on investigating whether added prior knowledge to the dueling bandit can further decrease the effort for the user while the user still gets satisfying visualizations. The second objective considers which learning paradigms performs better. The corresponding research question is

RQ 13: How does prior knowledge affect the performance of the dueling bandit?

In order to properly approach both research questions, the apparatus of the previous experiment is reused without modification of the user interface design (cf. Figure 7.2).

7.6.1 Procedure

The experiment starts with a consent. This consent includes information on the study's purpose, the approximated duration of the user study, and general information on the user study setup. After a participant agrees to the consent, five questions are shown. These five questions are identical to the questions used in the previous experiment. Hence, the answers can be directly used for classifying a participant.

Subsequently, each participant conducts a sequence of 30 comparisons. As in the previous experiment, the display order of the two visualizations is randomized in each comparison. In order to investigate the effect overtime, each participant rates a visualizations predicted after each 3 comparisons. For each predicted visualization, a participant rates further the satisfaction with the visualization shown. After the 30 comparisons as well as the 10 predictions, the experiment closes with a second questionnaire on the participant's impressions regarding the system's performance, the length of the conducted comparison sequence, and a potential implementation of the comparison sequence in an actual system.

This experiment is a between-subject study consisting of three groups. The first group gets no prior knowledge. The second group gets prior knowledge computed by the clustering, while the third group gets prior knowledge computed by the regression. Each participant is randomly assigned to one of these groups. Still, each participant performs the same

experiment procedure. Consequently, participants from the first group have to complete the initial questionnaire as well.

7.6.2 Participants

After data cleaning, 63 reliable people participated in this experiment. They are recruited from MTurk. As in the experiment in Section 6.4, only people with a US bachelor's degree could participated. On average, a participant completes the experiments in 5.55 minutes ($\sigma = 2.21$). Participates state to mainly create visualizations with MS Excel. Only a minority (8) uses a scripting language for visualization design. However, the knowledge in creating visualizations is moderate among the participants, while machine learning knowledge is scarce.

Due to the between-subject design, participants are randomly assigned to one of the following groups: no prior knowledge ($N = 22$), prior knowledge via clustering ($N = 21$), or prior knowledge via regression ($N = 20$). However, conducted Kruskal-Wallis tests with Bonferroni correction show no statistically significant differences between these three groups regarding the classification relevant characteristics (cf. Section 7.5). Additionally, a one-way ANOVA shows no statistically significant differences in the completion time between the groups.

Overall, conducted Kruskal-Wallis tests show statistically significant differences between this experiment's sample and the sample of the previous experiment regarding machine learning knowledge ($p < 0.05$), visualization knowledge ($p < 0.05$), the use of commercial tools ($p < 0.05$), and the use of MS excel ($p < 0.01$).

7.6.3 Results

All approaches gradually learn the participant's preferences overtime. Figure 7.15 shows the ratio of how often the participants selected the visualization for which the bandit excepted the higher preferences. However, the approach with prior knowledge via clustering seems to have an initial advantage. In the beginning, it assumes more often the preferred option correctly. However, there are no statistically significant differences between the three groups, according to a conducted Bonferroni-corrected Kruskal-Wallis test

A similar picture shows Figure 7.16. While the trend is slightly positive, there is overall a high variance in the participants' ratings for the predicted visualizations. However, the approach of clustering has the lowest variance ($\sigma^2 = 4.25$) in the first prediction compared to no prior knowledge ($\sigma^2 = 7.74$) and prior knowledge via regression ($\sigma^2 = 7.10$). In the last prediction, the variance is more similar among the groups. Additionally, the same holds

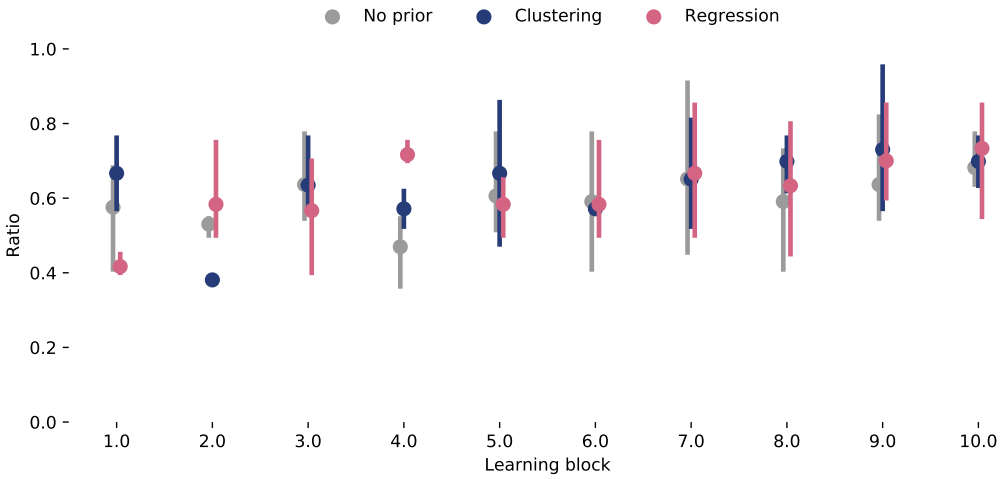


Fig. 7.15: Ratio of correctly assumed preferred visualizations during the learning for each approach.

for the mean of the ratings (no prior: 5.86, clustering: 6.57, regression: 5.50). However, the predictions ratings are not statistically significant between the groups, according to conducted Bonferroni-corrected Kruskal-Wallis tests.

Although the interactive learning sequence could be dull, the participants are not annoyed by the interactive learning approach in either group. Additionally, the participants would accept a sequence of comparisons of size 20 on average ($\sigma = 8$). Furthermore, the participants confirm that their preferences are learned and the visualizations fit their preferences. Lastly, there are no statistically significant differences in the participants' feedback between the groups.

7.6.4 Discussion

The results show two things. First, the results of the first experiment are confirmed. Second, prior knowledge seems to have an effect on the learning of visualization preferences.

Finding 1: The results confirm the results of the first experiment. The results confirm the insights from the first experiment and support the use of dueling bandits for learning visualization preferences. Although the participants of this experiment significantly differ from the participants of the first experiment, the results show a similar picture. In both learning behavior and prediction ratings, the data reveals a similar trend as in the first experiment. Additionally, the participants' feedback further supports the claim. In both experiments, the participants have the feeling that the approaches learn their preferences.

Finding 2: Prior knowledge seems to have an effect. Both Figure 7.15 and Figure 7.16 show differences between the approaches. However, conducted statistical tests reveal no significant differences. Yet, the rating scores vary among the groups. On the one hand,

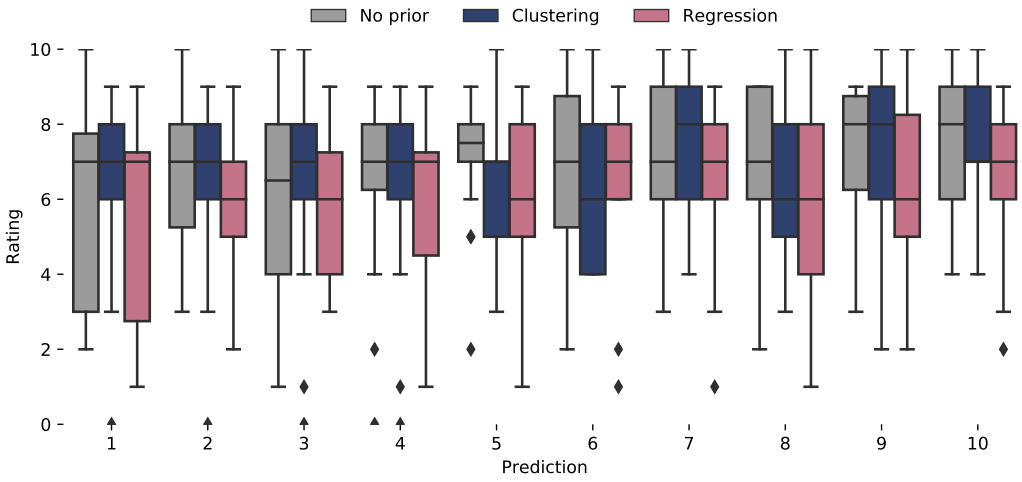


Fig. 7.16: The participants' ratings for the predicted visualizations for the three different approaches.

the approach with prior knowledge through clustering achieves the highest initial scores as well as has the lowest variance in the scores. It seems that it approximates the participants' preferences better. On the other hand, the approach with prior knowledge through the regression performs worse. The variance is high and the scores a lower. Using no prior knowledge allocates between these approaches.

In terms of RQ 13, prior knowledge seems to have an effect of the dueling bandit. This effect can be either positive or negative. Considering the significant differences between the two experiments, errors in predicting prior knowledge likely affects the performance. If a participants gets wrong prior knowledge, the dueling bandit has initial problems to approximate the actual preferences of this participant. This likely affects the scores in the beginning of the experiment. Hence, the effect of prior knowledge could be presumably higher if the both samples would be more similar.

7.7 Limitations

Both experiments support the use of a dueling bandit for learning the user's preferences. However, both experiments also raise some limitations on the methodology.

First, this chapters considers only one dueling bandit algorithm for learning the visualization preferences. However, a variety of different algorithms for this particular problem exists (Bouneffouf and Rish, 2019). Although the results support the chosen algorithm, another algorithm could achieve better results. This chapter lacks a comparisons of these algorithms. However, a comparison through an actual user study would be time consuming and expensive. Therefore, it is reasonable to consider the results of the benchmarks for dueling bandit algorithms as a decision bases.

Second, the modelling of prior knowledge uses the results of the first experiment. However, the participants of the first experiment might not represent the overall population of users. This fact further makes a prediction in a live system difficult. The results of the second experiment somehow address this lack of diversity in the data. Yet, covering the entire user spectrum in a relatively huge data set is hard. Especially the periphery of the user spectrum is hard to cover. Experts in visual analysis are scarce as well as people who never used visualizations before.

However, the findings of this chapter are valid. They answer the addressed research questions as well as reveal novel insights on dueling bandits in the domain of visualizations. Still, future work should elaborate a more detailed look on dueling bandit for learning visualization preferences.

7.8 Summary

This chapter approaches the challenge of online learning of visualization preferences. These preferences are needed in order to achieve a personalized visualization recommendation engine. However, knowledge on visualization preferences are scarce. Unlike other domains such as movies, a data base of users and corresponding visualization preferences does not exist.

In order to overcome this obstacle, a dueling bandit approach is evaluated in the context of visualizations. The D-TS of Wu and Liu (2016) is selected as it is one of the most efficient algorithms (Bouneffouf and Rish, 2019). Based on a sequence of pairwise comparisons, the bandit interactively learns the visualization preferences of a user. However, Section 7.3 propose to learn the visualization preferences by the divide and conquer paradigm in order to further reduce the effort for the user.

Two conducted experiments (Section 7.4 and Section 7.6) support the use of this dueling bandit approach for learning visualizations preferences. Participants confirm a positive effect of the bandit in learning the visualization preferences. Furthermore, the effort for the user could be further reduced due to modelling prior knowledge. It reduces the needed comparisons of the bandit.

The dueling bandit could be integrated into the overall concept of Chapter 5 by using an overlay during the mutual introduction. Hence, the new recommendation engine consists of an online learning component for adjusting to the user, but also an offline learning component for modelling the prior knowledge for the online algorithm. Hence, the overall algorithm is a kind of hybrid approach motivated by improving the usability.

8 Conclusion

Data analysis is becoming one of the main skills for today. In business and private life, users should be able to understand the means of data analysis. Hence, a broader spectrum of prospective users will likely engage. However, conducting visual analysis is still challenging. Achieving an effective visualization requires certain knowledge from a user. In order to lower barriers in visual analysis, multimodal approaches using speech have been proposed. This dissertation explores intelligible multimodal visual analysis by taking a holistic perspective.

In order to identify relevant elements in visual analysis for personalization, Chapter 3 proposes a method for estimating the personalization potentials of visual analysis tasks. Considering the state-of-the-art knowledge, the elements of visual analysis tasks, e.g., encoding of data, can be ranked. This ranking focuses on the dimensions of knowledge and preferences.

However, the behavior of the users also needed to be taken into account. Therefore, Chapter 4 explores how people use visualizations in the field and how they would formulate commands on a natural language-based user interface for visual analysis. The results reveal a narrow use of visualizations primarily considering classical visualization types (bar, line, and scatter) with mainly coloring. Furthermore, users followed a task-oriented patterns in interacting with visualizations while using a task-related dictionary.

Chapter 5 proposes a design for an intelligible multimodal visual analysis tool considering results of both Chapter 3 and Chapter 4. Using the modalities speech and touch, the design implements a communication with the user through visualizations, a textual dialogue, and a panel concerning the system's computations. The results show better and faster decisions by the participants compared to a conventional user interface. Additionally, the dialogue likely helps participants to make better sense of the underlying data along with the visualization.

However, using dialogue acts in a unified way occasionally triggers distrust in users. Hence, the design of dialogue acts likely depends on the user. In order to understand better how dialogue acts should be designed, Chapter 6 argues for an answer space based on linguistic theory for adjusting dialogue acts on data facts to the user. As visual analysis includes also statistics, these methods require certain knowledge from a user. Indeed, the results exhibit significant differences concerning the design of dialogue acts (Section 6.4). According to the results, matching the user's language further improves both decision making and

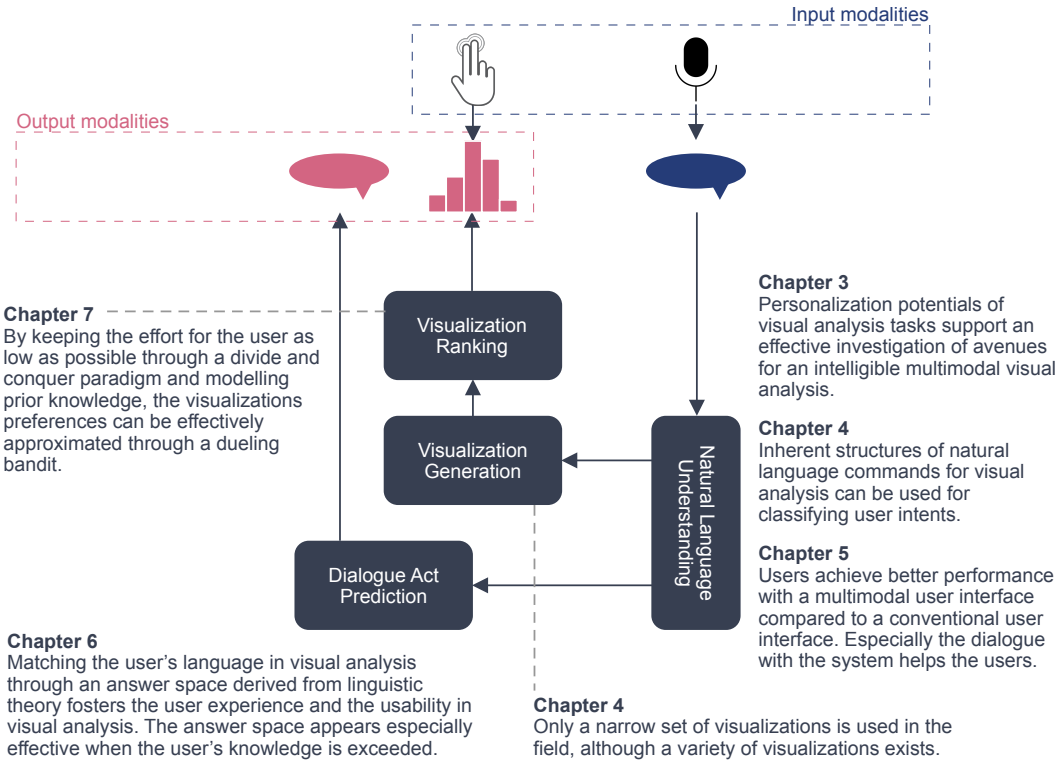


Fig. 8.1: Illustration of the relationship of the individual chapters' contributions to each other and to the system.

user experience. Users especially benefit in analysis situations in which their knowledge is exceeded (Section 6.6).

Chapter 2 discusses related work on the effect of the user's characteristics on the effectiveness of visualizations. While Chapter 3 also considers high personalization potentials for visual encoding of data, Chapter 7 consequently investigates how visualization recommendations can be effectively personalized. As it is important to align with the user, Chapter 7 uses a dueling bandit algorithm for interactive online learning of a user's individual preferences. Applying a divide and conquer approach on top of the dueling bandit, the results reveal an effective personalization of visualization recommendations without prior knowledge (Section 7.4). While the effort for the user is high, yet acceptable, using prior knowledge can likely decrease the effort (Section 7.6) further.

In a nutshell, this dissertation advocates intelligible multimodal visual analysis through personalization. Personalizing the output modalities of a system – visualizations and text – through dedicated machine learning methods shows benefits. Leveraging the behavior of users regarding input modalities – speech and touch – supports the design of multimodal visual analysis systems. This dissertation considers intelligible multimodal visual analysis from a holistic perspective (see Figure 8.1). It provides a solid empirical basis for future work in this area.

8.1 Limitations

While each chapter already addresses limitations to the particular methods, there are also limitations regarding this dissertation in general.

No evaluation of a fully integrated concept: This thesis explores a holistic perspective on achieving intelligible multimodal visual analysis. During this thesis, relevant aspects haven been investigated. These aspects comprise the range from estimating personalization potentially of visual analysis tasks through literature to applying online learning for personalizing visualization recommendations. Each chapter sequentially considers knowledge gained from preceding chapters. However, this thesis lacks an evaluation of a fully integrated prototype, although the technology-oriented Chapters 6 and 7 individually use the prototype Valletto (see Chapters 5) as an apparatus in their user studies. The question remains how an integrated prototype containing the answer space, and the dueling bandit would perform.

In terms of qualitative results, it would be interesting to see how the user's perceive the mutual introduction containing both the answers and the pairwise comparison sequence. Both elements represent a certain effort for the user. The participants' reactions could provide further insights on how this mutual introduction could be improved. However, both answer space and dueling bandit individually show acceptance by the participants. According to the results, participants would invest effort in such an initial routine when they get benefits afterwards.

In terms of quantitative results, it would be interesting to see how a fully integrated concept performs against a classical approaches. However, a corresponding evaluation constitutes high effort. A direct comparison with a classical approaches would contain multiple side effect, as there is not just one independent variable. Instead, the experiment must additionally consider each variant of the prototype individually. This would lead to at least four different variants (with/without answer space times with/without dueling bandit). Hence, five different systems have to be evaluated against each other. Furthermore, it should preferably be a between-subject study due to the effort for the participants. Yet, such an experiment would likely help to further show the potential advantages of intelligible multimodal visual analysis.

Nevertheless, each proposed method empirically shows its benefits towards personalized multimodal visual analysis. The results of Section 6.6 and Section 7.4, respectively, reveal empirical advantages for both concepts. Additionally, each method embeds its functionality to the overall concept. As the work also addresses the benefits of a personalized concept, a fully integrated prototype would likely perform well compared to a generic approach. However, this thesis cannot provide empirical evidence on advantages or disadvantages of such an fully integrated prototype.

Holistic exploration rather than in-depth analysis: As this thesis explores a holistic perspective on intelligible multimodal visual analysis, it lacks a truly in-depth analysis of one topic. While each chapter contributes to a better understanding of intelligible multimodal visual analysis, each chapter could also represent the topic of an entire thesis. Especially three areas provide room for in-depth analysis: the use of modalities (see Chapter 5), the answer space (see Chapter 6), and the online learning (see Chapter 7).

Considering the input modalities, this thesis leverages the predominately used modality combination of speech and touch (Cox et al., 2001; Gao et al., 2015; Hoque et al., 2018; Setlur et al., 2016; Srinivasan and Stasko, 2018; Srinivasan et al., 2020). Speech and touch together reveal synergies in interactions (Cohen et al., 1989). However, there are a variety of other modalities such as gaze, or haptics. Although Badam et al. (2017) show varying effectiveness of modalites for visual analysis tasks, there is still a lack of knowledge. Valletto could also incorporate other modalities for interaction. Especially gaze could potentially help to further personalize visual analysis, e.g., (Lalle et al., 2019).

Considering the answer space, other dimensions might be relevant. The answer space consists of the dimensions of information and support. Initially, these dimensions are chosen based on literature reviews and leveraging linguistic theory. However, a data-driven approach could lead to a different answer space. An exploratory user study could provide a basis for deriving a new answer space. Instead of letting the user choose a preferred answer given a visualization, the user could also propose a formulation how (s)he would describe the data fact shown in the visualization. Given these formulation, new dimensions might be revealed.

Considering the personalization of the visualization recommendations, other algorithms may perform well too. This thesis shows both advantages of interactively learning the user's preferences and how preferences can be learned for each feature individually. To do so, it leverages the D-TS (Wu and Liu, 2016). However, a variety of dueling bandit algorithms exists (Bouneffouf and Rish, 2019). Especially the areas of adversarial bandits and contextual bandit contain potentials for improvements. On the one hand, adversarial bandits can learn shifts in the preferences over time (Zhou, 2015). A user's preferences can change during the user of a system with increasing knowledge. An adversarial bandit can handles this. On the other hand, a contextual bandit directly incorporates the user's characteristics for its decision for an arm (Zhou, 2015). A contextual bandit could be a decent comparison for the current approach of learning prior knowledge and sequentially using the D-TS (Wu and Liu, 2016). In general, the field of online algorithms should be explored.

However, each of these aspects could also be the topic of an entire thesis. Furthermore, the novelty of this thesis lies in the idea of achieving an intelligible multimodal visual analysis. As current approaches do not consider the individual user differences, this thesis provides an

empirical basis for future work by identifying avenues for personalization. In fact, future work should explore these aspects in detail based on the results of this thesis.

8.2 Future Work

In addition to cover the limitations of this thesis, future work could directly build up the methods of this thesis.

Personalizing the input modalities, not only the output modalities: Both user and system use different modalities for interaction. In this thesis, the user can use speech and touch for communicating with the system. However, the system communicates through visual elements (visualizations) and text. In order to explore intelligible multimodal visual analysis, this thesis proposes to personalize the communication channels of the system. Hence, the system adapts its responses to the user. A next step now would be to personalize the input channels.

Using speech reveals information about the user. Users talk differently. They use different phrases, have an accents, or follow a specific grammar. All these information help to learn about the user. Yet, systems typically have a one-fits-all model for NLP. Although the models become more and more robust, they fail from time to time (cf. Section 6.5). An approach could investigate how to fine tune a model to a user in visual analysis. Fine tuning is one of the current approaches in NLP. Transformer models such as BERT (Vaswani et al., 2017) achieve top scores in the main NLP tasks. These transformer models are trained for a general purpose on a huge data set. However, they can be successfully fine tuned towards a specific domain by only little data. Having such a model could help to adapt to the user over time. It further could help to enrich the user model without asking the user explicitly. Hence, the effort for the user would be reduced.

Touch could be personalized as well. Adaptable touch gestures are already common in practice. For instance, OS X allows a user to specify the available gestures on the touch pad. Furthermore, works such as Findlater and McGrenere (2004) show the effect of enabling adaptable and adaptive interactions.

However, both cases should be explored in future work in the domain of multimodal visual analysis. Both could help to achieve an even more personalized approach.

Educating the user through the dialogue: The answer space of Chapter 6 enables the user-specific communication of data facts. As predicting a preferred answer can fail, a user is able to ask for a reformulation of the given answer. Depending on how the user asks, the system knows how to adapt the answer. Given a descriptive answer with high information, if

the user asks for "explanation of the statement" it could indicate that the answer exceeds the user's knowledge. In this moment, the system could try to educate the user.

An educational session could be a personalized sequence of dialogue acts. This sequence could start with a dialogue act regarding whether the user wants to learn more about the given answer. Depending on the user's experience, the system could explain different elements. Starting with describing what the output parameters indicate, the system could explain the mathematical concept itself including its constraints.

Having such an educational session could to improve the knowledge of the user in data analysis. Novice users would turn into more experienced users after a certain time. This would generally lead to a situation in which people will be less likely to be misled by visualizations.

Explore co-active learning for information visualization: Chapter 7 explores the concept of dueling bandits for learning visualization preferences. However, there are multiple approaches on this topic. One approach learns from the user's adjustment given a predicted visualization. Depending on how a user adjusts a visualization, the system learns the preferences. This procedure is called co-active learning.

Although adjusting a visualization requires a specific knowledge from the user, however, it requires less knowledge than creating a visualization from scratch (cf. Section 3.3.3). Considering this methodology in more detail could help to further learn the user's preferences. For instance, a system initially learns the user's preferences through a dueling bandit. During the use, the system may learn the user's preferences by observing the changes. Future work could explore how effective this learning paradigm as well as how much more knowledge the system is able to generate through the co-active learning.

Achieving truly ubiquitous visual analysis: Chapter 5 additionally discusses an idea of achieving an ubiquitous visual analysis. Speech as a main modality actually empowers a system to consider other contexts than the desktop. In fact, the desktop becomes less important in future workplaces (Roberts et al., 2014). Technology enriched workplaces allow people to work wherever they want, e.g., on the terrace, or on the couch. Consequently, different modalities should be supported depending on the context. Furthermore, Weiser (1999) projected that the computer of the 21st century will not have only one form, but many. Today, computers of different shapes exist. Hence, visual analysis does not necessarily happen at the desktop (Roberts et al., 2014).

Imagine, a data scientist moves around the building and can work wherever (s)he wants. In the morning the user might start at the desktop using mouse and keyboard. Later, the user has a meeting with colleagues. Together, they explore project related data. This situation

essentially represents a co-located collaborative working situation. Every participant of this meeting could contribute by explaining their findings through visualizations. Using gestures – either through touch or in the air – would be the preferred modality as the participants likely talk with each other to make their points. Later the day, the user has a presentation in front of the management. Using the created visualizations from the morning, the user shows these visualization at a big screen. However, (s)he uses a smart watch for interacting with the visualizations as (s)he wants to talk to the audience, but not to the screen. This would naturally embed the visual analysis into a common presentation style.

The desktop scenario as well as the mobile scenario are addressed by this thesis. However, other scenarios require further investigations regarding the design of user interface for visualizations. Especially for collaboration, interactive surfaces help (Isenberg et al., 2013). Furthermore, Horak et al. (2018) highlights the potentials of the combinations of smart watches and large screens.

However, new contexts require new evaluations. Each context has its own inherent challenges, e.g., ambient noise in the mobile context, or handling multiple users in a meeting room at the same time. Additionally, little knowledge on transitions from one scenario to another exists in visual analysis. For this reasons, it seems promising to investigate in ubiquitous visual analysis.

List of Figures

List of Tables

Bibliography

Ahn, Jae-Wook and Peter Brusilovsky (Sept. 2013). "Adaptive Visualization for Exploratory Information Retrieval". In: *Inf. Process. Manage.* 49.5, pp. 11391164. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2013.01.007.

Amar, Robert and John Stasko (2004). "A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations". In: *IEEE Symposium on Information Visualization*, pp. 143–150. DOI: 10.1109/INFVIS.2004.10.

Amar, Robert, James Eagan, and John Stasko (2005). "Low-Level Components of Analytic Activity in Information Visualization". In: *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. INFOVIS '05. Washington, DC, USA: IEEE Computer Society, pp. 15–. ISBN: 0-7803-9464-x. DOI: 10.1109/INFOVIS.2005.24.

Aurisano, Jillian, Abhinav Kumar, Alberto Gonzalez, Jason Leigh, Barbara DiEugenio, and Andrew Johnson (2016). "Articulate2: Toward a Conversational Interface for Visual Data Exploration". In: *IEEE Visualization 2016*. Baltimore, MD, USA.

Badam, Sriram Karthik, Arjun Srinivasan, Niklas Elmqvist, and John Stasko (2017). "Affordances of Input Modalities for Visual Data Exploration in Immersive Environments". In: *Workshop on Immersive Analytics*. Phoenix, AZ, USA.

Barnes, Alan (2016). "Making Intelligence Analysis More Intelligent: Using Numeric Probabilities". In: *Intelligence and National Security* 31.3, pp. 327–344. DOI: 10.1080/02684527.2014.994955.

Battle, Leilani and Jeffrey Heer (2019). "Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau". In: *Computer Graphics Forum* 38.3, pp. 145–159. DOI: 10.1111/cgf.13678. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13678.

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel (2007). "KNIME: The Konstanz Information Miner". In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer. ISBN: 978-3-540-78239-1.

Bertin, Jacques (1974). *Graphische Semiologie: Diagramme, Netze, Karten*. de Gruyter. ISBN: 9783110036602.

Bolt, Richard A. (1980). ""Put-that-there": Voice and Gesture at the Graphics Interface". In: *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '80. Seattle, Washington, USA: ACM, pp. 262–270. ISBN: 0-89791-021-4. DOI: 10.1145/800250.807503.

Borkin, Michelle A., Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister (2013). "What Makes a Visualization Memorable?" In: *IEEE Transactions on Visualization and Computer Graphics* 19.12, pp. 2306–2315. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.234.

Bouneffouf, Djallel and Irina Rish (2019). "A Survey on Practical Applications of Multi-Armed and Contextual Bandits". In: *arXiv e-prints*, arXiv:1904.10040, arXiv:1904.10040. arXiv: 1904.10040 [cs.LG].

Branigan, Holly P., Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown (2011). "The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers". In: *Cognition* 121.1, pp. 41–57. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2011.05.011.

Brehmer, Matthew and Tamara Munzner (Dec. 2013). "A Multi-Level Typology of Abstract Visualization Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12, pp. 2376–2385. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.124.

Britz, Denny, Anna Goldie, Minh-Thang Luong, and Quoc Le (Sept. 2017). "Massive Exploration of Neural Machine Translation Architectures". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1442–1451. DOI: 10.18653/v1/D17-1151.

Brooke, John (1996). *SUS - A quick and dirty usability scale*. CRC Press.

Budescu, David V., Shalva Weinberg, and Thomas S. Wallsten (1988). "Decisions based on numerically and verbally expressed uncertainties." In: *Journal of Experimental Psychology: Human Perception and Performance* 14.2, pp. 281–294.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling (2011). "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" In: *Perspectives on Psychological Science* 6.1, pp. 3–5. DOI: 10.1177/1745691610393980.

Burges, Chris J. C., Krysta M. Svore, Qiang Wu, and Jianfeng Gao (2008). *Ranking, Boosting, and Model Adaptation*. Tech. rep. MSR-TR-2008-109, p. 18.

Busa-Fekete, Róbert, Eyke Hüllermeier, and Balázs Szörényi (2014). "Preference-based Rank Elicitation Using Statistical Models: The Case of Mallows". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, pp. II–1071–II–1079.

Busa-Fekete, Róbert, Eyke Hüllermeier, and Adil El Mesaoudi-Paul (2018). "Preference-based Online Learning with Dueling Bandits: A Survey". In: *CoRR* abs/1807.11398. arXiv: 1807.11398.

Card, Stuart (2002). "Information Visualization". In: *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. USA: L. Erlbaum Associates Inc., pp. 544582. ISBN: 0805838384.

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman, eds. (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1-55860-533-9.

Carpendale, Sheelagh (2003). *Considering Visual Variables as a Basis for Information Visualisation*. Tech. rep. Calgary, AB: University of Calgary.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad (2015). "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 15. Sydney, NSW, Australia: Association for Computing Machinery, pp. 17211730. ISBN: 9781450336642. DOI: 10.1145/2783258.2788613.

Casler, Krista, Lydia Bickel, and Elizabeth Hackett (2013). "Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing". In: *Computers in Human Behavior* 29.6, pp. 2156 –2160. ISSN: 0747-5632. DOI: 10.1016/j.chb.2013.05.009.

Chan, Hou Pong, Tong Zhao, and Irwin King (2016). "Trust-aware Peer Assessment Using Multi-armed Bandit Algorithms". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. WWW '16 Companion. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, pp. 899–903. ISBN: 978-1-4503-4144-8. DOI: 10.1145/2872518.2891080.

Chaves, Ana Paula and Marco Aurélio Gerosa (2019). "How should my chatbot interact? A survey on human-chatbot interaction design". In: *CoRR* abs/1904.02743. arXiv: 1904.02743.

Chen, Mei-Ling and Hao-Chuan Wang (2018). "How Personal Experience and Technical Knowledge Affect Using Conversational Agents". In: *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. IUI '18 Companion. Tokyo, Japan: ACM, 53:1–53:2. ISBN: 978-1-4503-5571-1. DOI: 10.1145/3180308.3180362.

Chen, Rui, Qingyi Hua, Yan-Shuo Chang, Bo Wang, Lei Zhang, and Xiangjie Kong (2018). "A Survey of Collaborative Filtering-Based Recommender Systems: From Traditional Methods to Hybrid Methods Based on Social Networks". In: *IEEE Access* 6, pp. 64301–64320. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2877208.

Chi, Ed H. (2000). "A taxonomy of visualization techniques using the data state reference model". In: *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pp. 69–75. DOI: 10.1109/INFVIS.2000.885092.

CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016). San Jose, California, USA: ACM. ISBN: 978-1-4503-3362-7.

Cleveland, William S. and Robert McGill (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". In: *Journal of the American Statistical Association* 79.387, pp. 531–554. DOI: 10.1080/01621459.1984.10478080.

Cohen, P. R., M. Dalrymple, D. B. Moran, F. C. Pereira, and J. W. Sullivan (1989). "Synergistic Use of Direct Manipulation and Natural Language". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '89. New York, NY, USA: ACM, pp. 227–233. ISBN: 0-89791-301-9. DOI: 10.1145/67449.67494.

Cohen, Philip R. (1992). "The Role of Natural Language in a Multimodal Interface". In: *Proceedings of the 5th Annual ACM Symposium on User Interface Software and Technology*. UIST 92. Monteray, California, USA: Association for Computing Machinery, pp. 143149. ISBN: 0897915496. DOI: 10.1145/142621.142641.

Cohen, Philip R. and Sharon L. Oviatt (1995). "The role of voice input for human-machine communication". In: *Proceedings of the National Academy of Sciences* 92.22, pp. 9921–9927. ISSN: 0027-8424. eprint: http://www.pnas.org/content/92/22/9921.full.pdf.

Conati, Cristina and Heather Maclaren (2008). "Exploring the Role of Individual Differences in Information Visualization". In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '08. Napoli, Italy: ACM, pp. 199–206. ISBN: 978-1-60558-141-5. DOI: 10.1145/1385569.1385602.

Conati, Cristina, Giuseppe Carenini, Enamul Hoque, Ben Steichen, and Dereck Toker (2014). "Evaluating the Impact of User Characteristics and Different Layouts on an Interactive Visualization for Decision Making". In: *Proceedings of the 16th Eurographics Conference on Visualization*. EuroVis '14. Swansea, Wales, United Kingdom: Eurographics Association, pp. 371–380. DOI: 10.1111/cgf.12393.

Conati, Cristina, Giuseppe Carenini, Dereck Toker, and Sébastien Lallé (2015). "Towards User-Adaptive Information Visualization". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI15. Austin, Texas: AAAI Press, pp. 41004106. ISBN: 0262511290.

Cornel, D., A. Buttinger-Kreuzhuber, A. Konev, Z. Horváth, M. Wimmer, R. Heidrich, and J. Waser (2019). "Interactive Visualization of Flood and Heavy Rain Simulations". In: *Computer Graphics Forum* 38.3, pp. 25–39. DOI: 10.1111/cgf.13669. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13669.

Correll, Michael and Michael Gleicher (2014). "Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12, pp. 2142–2151. ISSN: 1077-2626. DOI: 10.1109/TVCG.2014.2346298.

Cowan, Benjamin R., Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira (2017). ""What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants". In: *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '17. Vienna, Austria: ACM, 43:1–43:12. ISBN: 978-1-4503-5075-4. DOI: 10.1145/3098279.3098539.

Cox, Kenneth, Rebecca E. Grinter, Stacie L. Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla (July 2001). "A Multi-Modal Natural Language Interface to an Information Visualization Environment". In: *International Journal of Speech Technology* 4.3, pp. 297–314. ISSN: 1572-8110. DOI: 10.1023/A:1011368926479.

Crotty, Andrew, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska (Aug. 2015). "Vizdom: Interactive Analytics Through Pen and Touch". In: *Proceedings of the VLDB Endowment - Proceedings of the 41st International Conference on Very Large Data Bases*. VLDB '15 8.12, pp. 2024–2027. ISSN: 2150-8097. DOI: 10.14778/2824032.2824127.

Cui, Weiwei, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang (2019). "Text-to-Viz: Automatic Generation of Infographics from Proportion-Related Natural Language Statements". In: *IEEE Transactions on Visualization and Computer Graphics* 26, pp. 906–916.

Cui, Yan (2015). "An Evaluation of Yelp Dataset". In: *CoRR* abs/1512.06915. arXiv: 1512.06915.

Dasgupta, Aritra, Jorge Poco, Yaxing Wei, Robert Cook, Enrico Bertini, and Cláudio T. Silva (2015). "Bridging Theory with Practice: An Exploratory Study of Visualization Use and Design for Climate Model Comparison". In: *IEEE Transactions on Visualization and Computer Graphics* 21.9, pp. 996–1014. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2413774.

Dasgupta, Aritra, Susannah Burrows, Kyungsik Han, and Philip J. Rasch (2017). "Empirical Analysis of the Subjective Impressions and Objective Measures of Domain Scientists' Visual Analytic Judgments". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: ACM, pp. 1193–1204. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025882.

Demiralp, Çağatay, Peter J. Haas, Srinivasan Parthasarathy, and Tejaswini Pedapati (2017). "Foresight: Rapid Data Exploration Through Guideposts". In: *IEEE VIS'17 Data Systems and Interactive Analysis (DSIA) Workshop*. Phoenix, Arizona, USA.

Dhamdhere, Kedar, Kevin S. McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan (2017). "Analyza: Exploring Data with Conversation". In: *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*. IUI '17. Limassol, Cyprus: ACM, pp. 493–504. ISBN: 978-1-4503-4348-0. DOI: 10.1145/3025171.3025227.

Dhami, Mandeep K., David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock (2015). "Improving Intelligence Analysis With Decision Science". In: *Perspectives on Psychological Science* 10.6, pp. 753–757. DOI: 10.1177/1745691615598511.

Dibia, Victor and Çagatay Demiralp (2018). "Data2Vis: Automatic Generation of Data Visualizations Using Sequence to Sequence Recurrent Neural Networks". In: *CoRR* abs/1804.03126.

Dimara, Evanthia, Anastasia Bezerianos, and Pierre Dragicevic (2018). "Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1, pp. 749–759. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2745138.

Donoho, David and Ernesto Ramos (1982). *PRIMDATA: Data Sets for Use With PRIM-H*.

Drucker, Steven M., Danyel Fisher, Ramik Sadana, Jessica Herron, and m.c. schraefel (2013). "TouchViz: A Case Study Comparing Two Interfaces for Data Analytics on Tablets". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. Paris, France: ACM, pp. 2301–2310. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2481318.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*.

Dwyer, Gareth, Shalabh Aggarwal, and Jack Stouffer (2017). *Flask: Building Python Web Services*. Packt Publishing. ISBN: 1787288226.

Eells, Walter Crosby (1926). "The Relative Merits of Circles and Bars for Representing Component Parts". In: *Journal of the American Statistical Association* 21.154, pp. 119–132. DOI: 10.1080/01621459.1926.10502165.

Elahi, Mehdi, Francesco Ricci, and Neil Rubens (2016). "A survey of active learning in collaborative filtering recommender systems". In: *Computer Science Review* 20, pp. 29 –50. ISSN: 1574-0137. DOI: 10.1016/j.cosrev.2016.05.002.

Explosion AI (2018). *spaCy - Industrial-Strength Natural Language Processing*. https://spacy.io. [Online; accessed 04-February-2018].

Facebook Inc (2013). *React – A JavaScript library for building user interfaces*. https://reactjs.org.

Fast, Ethan, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein (2018). "Iris: A Conversational Agent for Complex Tasks". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 473:1–473:12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174047.

Findlater, Leah and Joanna McGrenere (2004). “A Comparison of Static, Adaptive, and Adaptable Menus”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI 04. Vienna, Austria: Association for Computing Machinery, pp. 8996. ISBN: 1581137028. DOI: 10.1145/985692.985704.

Fischer, Gerhard (2001). “User Modeling in Human–Computer Interaction”. In: *User Modeling and User-Adapted Interaction* 11.1, pp. 65–86. DOI: 10.1023/A:1011145532042.

Gajos, Krzysztof Z. and Krysta Chauncey (2017). “The Influence of Personality Traits and Cognitive Load on the Use of Adaptive User Interfaces”. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. IUI 17. Limassol, Cyprus: Association for Computing Machinery, pp. 301306. ISBN: 9781450343480. DOI: 10.1145/3025171.3025192.

Gallois, Cindy and Howard Giles (2015). “Communication Accommodation Theory”. In: *The International Encyclopedia of Language and Social Interaction*. American Cancer Society, pp. 1–18. ISBN: 9781118611463. DOI: 10.1002/9781118611463.wbielsi066. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118611463.wbielsi066.

Gallois, Cindy, Tania Ogay, and Howard Giles (2005). “Communication accommodation theory: A look back and a look ahead”. In:

Gao, Tong, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios (2015). “DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization”. In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*. UIST ’15. Charlotte, NC, USA: ACM, pp. 489–500. ISBN: 978-1-4503-3779-3. DOI: 10.1145/2807442.2807478.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal (2018). “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. DOI: 10.1109/DSAA.2018.00018.

Grammel, Lars, Melanie Tory, and Margaret-Anne Storey (Nov. 2010). “How Information Visualization Novices Construct Visualizations”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pp. 943–952. ISSN: 1077-2626. DOI: 10.1109/TVCG.2010.164.

Green, Tear Marie and Brian Fisher (2010). “Towards the Personal Equation of Interaction: The impact of personality factors on visual analytics interface interaction”. In: *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 203–210. DOI: 10.1109/VAST.2010.5653587.

Grice, Herbert Paul (1975). “Logic and Conversation”. In: *Syntax and Semantics: Vol. 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York: Academic Press, pp. 41–58.

Gulwani, Sumit and Mark Marron (2014). “NLyze: Interactive Programming by Natural Language for Spreadsheet Data Analysis and Manipulation”. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’14. Snowbird, Utah, USA: ACM, pp. 803–814. ISBN: 978-1-4503-2376-5. DOI: 10.1145/2588555.2612177.

Haroz, Steve and David Whitney (Dec. 2012). “How Capacity Limits of Attention Influence Information Visualization Effectiveness”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12, pp. 2402–2410. ISSN: 1077-2626. DOI: 10.1109/TVCG.2012.233.

Harrison, Lane, Fumeng Yang, Steven Franconeri, and Remco Chang (2014). “Ranking Visualizations of Correlation Using Weber’s Law”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12, pp. 1943–1952. ISSN: 1077-2626. DOI: 10.1109/TVCG.2014.2346979.

Harrison, Lane, Katharina Reinecke, and Remco Chang (2015). "Infographic Aesthetics: Designing for the First Impression". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI 15. Seoul, Republic of Korea: Association for Computing Machinery, pp. 11871190. ISBN: 9781450331456. DOI: 10.1145/2702123.2702545.

Hart, Sandra G. and Lowell E. Staveland (1988). "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Human Mental Workload*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Advances in Psychology. North-Holland, pp. 139 –183. DOI: 10.1016/S0166-4115(08)62386-9.

Hauptmann, A. G. (1989). "Speech and Gestures for Graphic Image Manipulation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '89. New York, NY, USA: ACM, pp. 241–245. ISBN: 0-89791-301-9. DOI: 10.1145/67449.67496.

Hearst, Marti and Melanie Tory (2019). "Would You Like A Chart With That? Incorporating Visualizations into Conversational Interfaces". In: *2019 IEEE Visualization Conference (VIS)*, pp. 1–5. DOI: 10.1109/VISUAL.2019.8933766.

Heer, Jeffrey and Michael Bostock (2010). "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. Atlanta, Georgia, USA: ACM, pp. 203–212. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753357.

Heer, Jeffrey and Ben Shneiderman (Apr. 2012). "Interactive Dynamics for Visual Analysis". In: *Communications of the ACM* 55.4, pp. 45–54. ISSN: 0001-0782. DOI: 10.1145/2133806.2133821.

Heer, Jeffrey, Frank Ham, Sheelagh Carpendale, Chris Weaver, and Petra Isenberg (2008). "Information Visualization". In: ed. by Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. Berlin, Heidelberg: Springer-Verlag. Chap. Creation and Collaboration: Engaging New Audiences for Information Visualization, pp. 92–133. ISBN: 978-3-540-70955-8. DOI: 10.1007/978-3-540-70956-5_5.

Hendrix, Gary G., Earl D. Sacerdoti, Daniel Sagalowicz, and Jonathan Slocum (June 1978). "Developing a Natural Language Interface to Complex Data". In: *ACM Trans. Database Syst.* 3.2, pp. 105147. ISSN: 0362-5915. DOI: 10.1145/320251.320253.

Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy (2016). "The age of analytics: Competing in a data-driven world". In: *McKinsey Global Reports*.

Hoegen, Rens, Deepali Aneja, Daniel McDuff, and Mary Czerwinski (2019). "An End-to-End Conversational Style Matching Agent". In: *CoRR* abs/1904.02760. arXiv: 1904.02760.

Hoque, Enamul, Vidya Setlur, Melanie Tory, and Isaac Dykeman (2018). "Applying Pragmatics Principles for Interaction with Visual Analytics". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1, pp. 309–318. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2744684.

Horak, Tom, Sriram Karthik Badam, Niklas Elmqvist, and Raimund Dachselt (2018). "When David Meets Goliath: Combining Smartwatches with a Large Vertical Display for Visual Data Exploration". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 19:1–19:13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173593.

Hornbæk, Kasper, Benjamin B. Bederson, and Catherine Plaisant (Dec. 2002). "Navigation Patterns and Usability of Zoomable User Interfaces with and without an Overview". In: *ACM Trans. Comput.-Hum. Interact.* 9.4, pp. 362389. ISSN: 1073-0516. DOI: 10.1145/586081.586086.

Hu, Kevin, Michiel A. Bakker, Stephen Li, Tim Kraska, and César Hidalgo (2019). "VizML: A Machine Learning Approach to Visualization Recommendation". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: ACM, 128:1–128:12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300358.

Huang, Dandan, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury (2015). "Personal Visualization and Personal Visual Analytics". In: *IEEE Transactions on Visualization and Computer Graphics* 21.3, pp. 420–433. ISSN: 2160-9306. DOI: 10.1109/TVCG.2014.2359887.

Hurst, Amy, Scott E. Hudson, and Jennifer Mankoff (2007). "Dynamic Detection of Novice vs. Skilled Use without a Task Model". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI 07. San Jose, California, USA: Association for Computing Machinery, pp. 271280. ISBN: 9781595935939. DOI: 10.1145/1240624.1240669.

Isenberg, Petra, Tobias Isenberg, Tobias Hesselmann, Bongshin Lee, Ulrich von Zadow, and Anthony Tang (Mar. 2013). "Data Visualization on Interactive Surfaces: A Research Agenda". In: *IEEE Comput. Graph. Appl.* 33.2, pp. 1624. ISSN: 0272-1716. DOI: 10.1109/MCG.2013.24.

Jabbari, A., R. Blanch, and S. Dupuy-Chessa (2018). "Composite Visual Mapping for Time Series Visualization". In: *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 116–124. DOI: 10.1109/PacificVis.2018.00023.

Joachims, Thorsten (2002). "Optimizing Search Engines Using Clickthrough Data". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: ACM, pp. 133–142. ISBN: 1-58113-567-X. DOI: 10.1145/775047.775067.

John, Rogers Jeffrey Leo, Navneet Potti, and Jignesh M Patel (2017). "Ava: From Data to Insights Through Conversation". In: *Biennial Conference on Innovative Data Systems Research*. CIDR 2017.

Jung, Daekyoung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo (2017). "ChartSense: Interactive Data Extraction from Chart Images". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: ACM, pp. 6706–6717. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025957.

Kabbara, Jad (June 2019). "Computational Investigations of Pragmatic Effects in Natural Language". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 71–76. DOI: 10.18653/v1/N19-3010.

Kassel, Jan-Frederik and Michael Rohs (2017). "Immersive Navigation in Visualization Spaces through Swipe Gestures and Optimal Attribute Selection". In: *Proceedings of the 2nd Workshop on Immersive Analytics: Exploring Future Interaction and Visualization Technologies for Data Analytics*. IEEE VIS '17. Phoenix, AZ, USA.

– (2018). "Valletto: A Multimodal Interface for Ubiquitous Visual Analytics". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. Montreal QC, Canada: ACM, LBW005:1–LBW005:6. ISBN: 978-1-4503-5621-3. DOI: 10.1145/3170427.3188445.

– (2019a). “Online Learning of Visualization Preferences through Dueling Bandits for Enhancing Visualization Recommendations”. In: *EuroVis 2019 - Short Papers*. Ed. by Jimmy Johansson, Filip Sadlo, and G. Elisabeta Marai. Porto, Portugal: The Eurographics Association. ISBN: 978-3-03868-090-1. DOI: 10.2312/evs.20191175.

– (2019b). “Talk to Me Intelligibly: Investigating An Answer Space to Match the User’s Language in Visual Analysis”. In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. DIS ’19. San Diego, CA, USA: ACM, pp. 1517–1529. ISBN: 978-1-4503-5850-7. DOI: 10.1145/3322276.3322282.

Kay, Matthew Kay and Jeffrey Heer (2016). “Beyond Weber’s Law: A Second Look at Ranking Visualizations of Correlation”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 469–478. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467671.

Kehrer, Johannes and Helwig Hauser (Mar. 2013). “Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.3, pp. 495513. ISSN: 1077-2626. DOI: 10.1109/TVCG.2012.110.

Keim, Daniel, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann, eds. (2010). *Mastering the information age : solving problems with visual analytics*. Goslar : Eurographics Association. ISBN: 978-3-905673-77-7.

Keim, Daniel A., Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler (2006). “Challenges in Visual Data Analysis”. In: *Tenth International Conference on Information Visualisation (IV’06)*, pp. 9–16. DOI: 10.1109/IV.2006.31.

Key, Alicia, Bill Howe, Daniel Perry, and Cecilia Aragon (2012). “VizDeck: Self-organizing Dashboards for Visual Analytics”. In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’12. Scottsdale, Arizona, USA: ACM, pp. 681–684. ISBN: 978-1-4503-1247-9. DOI: 10.1145/2213836.2213931.

Kim, Younghoon and Jeffrey Heer (2018). “Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings”. In: *Computer Graphics Forum*. ISSN: 1467-8659. DOI: 10.1111/cgf.13409.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh (2008). “Crowdsourcing User Studies with Mechanical Turk”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’08. Florence, Italy: ACM, pp. 453–456. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357127.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, et al. (2016). “Jupyter Notebooks – a publishing format for reproducible computational workflows”. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press, pp. 87 –90.

Ko, S., I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky, and D. S. Ebert (2016). “A Survey on Visual Analysis Approaches for Financial Data”. In: *Computer Graphics Forum* 35.3, pp. 599–617. DOI: 10.1111/cgf.12931. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12931.

Koh, Pang Wei and Percy Liang (2017). “Understanding Black-Box Predictions via Influence Functions”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML17. Sydney, NSW, Australia: JMLR.org, pp. 18851894.

Kosara, Robert (2019a). "Circular Part-to-Whole Charts Using the Area Visual Cue". In: *EuroVis 2019 - Short Papers*. Ed. by Jimmy Johansson, Filip Sadlo, and G. Elisabeta Marai. The Eurographics Association. ISBN: 978-3-03868-090-1. DOI: 10.2312/evs.20191163.

– (2019b). "The Impact of Distribution and Chart Type on Part-to-Whole Comparisons". In: *EuroVis 2019 - Short Papers*. Ed. by Jimmy Johansson, Filip Sadlo, and G. Elisabeta Marai. The Eurographics Association. ISBN: 978-3-03868-090-1. DOI: 10.2312/evs.20191162.

Kumar, Abhinav, Barbara Di Eugenio, Jillian Aurisano, Andrew Johnson, Abeer Alsaiari, Nigel Flowers, Alberto Gonzalez, and Jason Leigh (2017). "Towards Multimodal Coreference Resolution for Exploratory Data Visualization Dialogue: Context-Based Annotation and Gesture Identification". In: *The 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017 – SaarDial)*, p. 48.

Lallé, Sébastien and Cristina Conati (2019). "The Role of User Differences in Customization: A Case Study in Personalization for Infovis-Based Content". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI 19. Marina del Ray, California: Association for Computing Machinery, pp. 329339. ISBN: 9781450362726. DOI: 10.1145/3301275.3302283.

Lalle, Sebastien, Dereck Toker, and Cristina Conati (2019). "Gaze-Driven Adaptive Interventions for Magazine-Style Narrative Visualizations". In: *IEEE Transactions on Visualization and Computer Graphics*, pp. 11. ISSN: 2160-9306. DOI: 10.1109/tvcg.2019.2958540.

Lee, Sukwon, Sung-Hee Kim, and Bum Chul Kwon (Jan. 2017). "VLAT: Development of a Visualization Literacy Assessment Test". In: *IEEE Transactions on Visualization and Computer Graphics* 23.1, pp. 551560. ISSN: 1077-2626. DOI: 10.1109/TVCG.2016.2598920.

Lee, Sukwon, Bum Kwon, Jiming Yang, Byung Lee, and Sung-Hee Kim (2019). "The Correlation between Users Cognitive Characteristics and Visualization Literacy". In: *Applied Sciences* 9.3, p. 488. ISSN: 2076-3417. DOI: 10.3390/app9030488.

Lejeune, Helga and J. H. Wearden (2009). "Vierordt's The Experimental Study of the Time Sense (1868) and its legacy". In: *European Journal of Cognitive Psychology* 21.6, pp. 941–960. DOI: 10.1080/09541440802453006. eprint: https://doi.org/10.1080/09541440802453006.

Levkowitz, H. and M. Ferreira de Oliveira (2003). "From Visual Data Exploration to Visual Data Mining: A Survey". In: *IEEE Transactions on Visualization and Computer Graphics* 9.03, pp. 378–394. ISSN: 1941-0506. DOI: 10.1109/TVCG.2003.1207445.

Li, Fei and Hosagrahar V Jagadish (2014). "NaLIR: An Interactive Natural Language Interface for Querying Relational Databases". In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD '14. Snowbird, Utah, USA: ACM, pp. 709–712. ISBN: 978-1-4503-2376-5. DOI: 10.1145/2588555.2594519.

Liu, Tie-Yan (2010). "Learning to Rank for Information Retrieval". In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR 10. Geneva, Switzerland: Association for Computing Machinery, p. 904. ISBN: 9781450301534. DOI: 10.1145/1835449.1835676.

Liu, Zhicheng and John Stasko (Nov. 2010). "Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pp. 999–1008. ISSN: 1077-2626. DOI: 10.1109/TVCG.2010.177.

Lozano, Aurélie C. and Grzegorz Swirszcz (2012). "Multi-level Lasso for Sparse Multi-task Regression". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, pp. 595–602. ISBN: 978-1-4503-1285-1.

Luger, Ewa and Abigail Sellen (2016). “"Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: ACM, pp. 5286–5297. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858288.

Luo, Yuyu, Xuedi Qin, Nan Tang, Guoliang Li, and Xinran Wang (2018). “DeepEye: Creating Good Data Visualizations by Keyword Search”. In: *Proceedings of the 2018 International Conference on Management of Data*. SIGMOD ’18. Houston, TX, USA: ACM, pp. 1733–1736. ISBN: 978-1-4503-4703-7. DOI: 10.1145/3183713.3193545.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts (2011). “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150.

Mackinlay, Jock (Apr. 1986). “Automating the Design of Graphical Presentations of Relational Information”. In: *ACM Transactions on Graphics* 5.2, pp. 110–141. ISSN: 0730-0301. DOI: 10.1145/22949.22950.

Mackinlay, Jock, Pat Hanrahan, and Chris Stolte (2007). “Show Me: Automatic Presentation for Visual Analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6, pp. 1137–1144. ISSN: 1077-2626. DOI: 10.1109/TVCG.2007.70594.

Marks, J., B. Andalman, P. A. Beardsley, et al. (1997). “Design Galleries: A General Approach to Setting Parameters for Computer Graphics and Animation”. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., pp. 389–400. ISBN: 0-89791-896-7. DOI: 10.1145/258734.258887.

Molich, Rolf and Jakob Nielsen (Mar. 1990). “Improving a Human-computer Dialogue”. In: *Commun. ACM* 33.3, pp. 338–348. ISSN: 0001-0782. DOI: 10.1145/77481.77486.

Moore, Robert J., Raphael Arar, Guang-Jie Ren, and Margaret H. Szymanski (2017). “Conversational UX Design”. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’17. Denver, Colorado, USA: ACM, pp. 492–497. ISBN: 978-1-4503-4656-6. DOI: 10.1145/3027063.3027077.

Moore, Roger K., Hui Li, and Shih-Hao Liao (2016). “Progress and Prospects for Spoken Language Technology: What Ordinary People Think”. In: *Interspeech 2016*, pp. 3007–3011. DOI: 10.21437/Interspeech.2016-874.

Moritz, Dominik, Chenglong Wang, Greg L. Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer (2019). “Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1, pp. 438–448. ISSN: 1077-2626. DOI: 10.1109/TVCG.2018.2865240.

Moro, Sérgio, Paulo Cortez, and Paulo Rita (2014). “A data-driven approach to predict the success of bank telemarketing”. In: *Decision Support Systems* 62, pp. 22 –31. ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.03.001.

Mutlu, Belgin, Eduardo Veas, Christoph Trattner, and Vedran Sabol (2015). “VizRec: A Two-Stage Recommender System for Personalized Visualizations”. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*. IUI Companion ’15. New York, NY, USA: ACM, pp. 49–52. ISBN: 978-1-4503-3308-5. DOI: 10.1145/2732158.2732190.

Mutlu, Belgin, Eduardo Veas, and Christoph Trattner (Nov. 2016). "VizRec: Recommending Personalized Visualizations". In: *ACM Trans. Interact. Intell. Syst.* 6.4. ISSN: 2160-6455. DOI: 10.1145/2983923.

Nielsen, Jakob (2005). *Ten usability heuristics*.

Norman, Donald A. (2002). *The Design of Everyday Things*. New York, NY, USA: Basic Books, Inc. ISBN: 9780465067107.

– (2010). *Living with Complexity*. Cambridge, MA: MIT Press. ISBN: 978-0-262-01486-1.

Oscar, Nels, Shannon Mejía, Ronald Metoyer, and Karen Hooker (2017). "Towards Personalized Visualization: Information Granularity, Situation, and Personality". In: *Proceedings of the 2017 Conference on Designing Interactive Systems*. DIS 17. Edinburgh, United Kingdom: Association for Computing Machinery, pp. 811819. ISBN: 9781450349222. DOI: 10.1145/3064663.3064704.

Oviatt, Sharon (Mar. 1997). "Multimodal Interactive Maps: Designing for Human Performance". In: *Hum.-Comput. Interact.* 12.1, pp. 93–129. ISSN: 0737-0024. DOI: 10.1207/s15327051hci1201\&2_4.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010). "Running Experiments on Amazon Mechanical Turk". In: *Judgment and Decision Making* 5.5, pp. 411–419.

Patwardhan, Siddharth and Ted Pedersen (2006). "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts". In: *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together.*

Payr, Sabine (2013). "Your Virtual Butler". In: ed. by Robert Trappl. Berlin, Heidelberg: Springer-Verlag. Chap. Virtual Butlers and Real People: Styles and Practices in Long-term Use of a Companion, pp. 134–178. ISBN: 978-3-642-37345-9.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Poco, Jorge and Jeffrey Heer (2017). "Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images". In: *Computer Graphics Forum* 36.3, pp. 353–363. ISSN: 1467-8659. DOI: 10.1111/cgf.13193.

Pretorius, A. Johannes and Jarke J. Van Wijk (June 2009). "What Does the User Want to See?: What Do the Data Want to Be?" In: *Information Visualization* 8.3, pp. 153–166. ISSN: 1473-8716. DOI: 10.1057/ivs.2009.13.

Rashid, Al Mamunur, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl (2002). "Getting to Know You: Learning New User Preferences in Recommender Systems". In: *Proceedings of the 7th International Conference on Intelligent User Interfaces*. IUI '02. San Francisco, California, USA: ACM, pp. 127–134. ISBN: 1-58113-459-2. DOI: 10.1145/502716.502737.

Reda, Khairi, Andrew E. Johnson, Michael E. Papka, and Jason Leigh (2016). "Modeling and evaluating user behavior in exploratory visual analysis". In: *Information Visualization* 15.4, pp. 325–339. DOI: 10.1177/1473871616638546. eprint: https://doi.org/10.1177/1473871616638546.

Renooij, Silja and Cilia Witteman (1999). "Talking probabilities: communicating probabilistic information with words and numbers". In: *International Journal of Approximate Reasoning* 22.3, pp. 169 –194. ISSN: 0888-613X. DOI: 10.1016/S0888-613X(99)00027-4.

Ricci, Francesco, Lior Rokach, Bracha Shapira, and Paul B. Kantor (2010). *Recommender Systems Handbook*. 1st. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387858199.

Rich, Elaine (1998). “Readings in Intelligent User Interfaces”. In: ed. by Mark T. Maybury and Wolfgang Wahlster. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Chap. User Modeling via Stereotypes, pp. 329–342. ISBN: 1-55860-444-8.

Rind, Alexander, Wolfgang Aigner, Markus Wagner, Silvia Miksch, and Tim Lammarsch (2016). “Task Cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation”. In: *Information Visualization* 15.4, pp. 288–300. DOI: 10.1177/1473871615621602. eprint: https://doi.org/10.1177/1473871615621602.

Roberts, Jonathan C., Panagiotis D. Ritsos, Sriram Karthik Badam, Dominique Brodbeck, Jessie Kennedy, and Niklas Elmqvist (2014). “Visualization beyond the Desktop–the Next Big Thing”. In: *IEEE Computer Graphics and Applications* 34.6, pp. 26–34. ISSN: 1558-1756. DOI: 10.1109/MCG.2014.82.

Roth, Steven F. and Joe Mattis (1990). “Data Characterization for Intelligent Graphics Presentation”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI 90. Seattle, Washington, USA: Association for Computing Machinery, pp. 193200. ISBN: 0201509326. DOI: 10.1145/97243.97273.

Ruan, Sherry, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay (Jan. 2018). “Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones”. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.4. DOI: 10.1145/3161187.

Sadana, Ramik and John Stasko (June 2016). “Designing Multiple Coordinated Visualizations for Tablets”. In: *Comput. Graph. Forum* 35.3, pp. 261–270. ISSN: 0167-7055. DOI: 10.1111/cgf.12902.

Saffer, Dan (2006). *Designing for Interaction: Creating Smart Applications and Clever Devices*. First. USA: Peachpit Press. ISBN: 0321447123.

Saket, Bahador, Alex Endert, and Cagatay Demiralp (2018). “Task-Based Effectiveness of Basic Visualizations”. In: *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1. ISSN: 1077-2626. DOI: 10.1109/TVCG.2018.2829750.

Saktheeswaran, A., A. Srinivasan, and J. Stasko (2020). “Touch? Speech? or Touch and Speech? Investigating Multimodal Interaction for Visual Network Exploration and Analysis”. In: *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1. ISSN: 2160-9306. DOI: 10.1109/TVCG.2020.2970512.

Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl (2001). “Item-Based Collaborative Filtering Recommendation Algorithms”. In: *Proceedings of the 10th International Conference on World Wide Web*. WWW 01. Hong Kong, Hong Kong: Association for Computing Machinery, pp. 285295. ISBN: 1581133480. DOI: 10.1145/371920.372071.

Satyanarayan, Arvind, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer (Jan. 2017). “Vega-Lite: A Grammar of Interactive Graphics”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1, pp. 341–350. ISSN: 1077-2626. DOI: 10.1109/TVCG.2016.2599030.

Savva, Manolis, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer (2011). “ReVision: Automated Classification, Analysis and Redesign of Chart Images”. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST ’11. Santa Barbara, California, USA: ACM, pp. 393–402. ISBN: 978-1-4503-0716-1. DOI: 10.1145/2047196.2047247.

Schegloff, Emanuel A. (1968). "Sequencing in Conversational Openings". In: *American Anthropologist* 70.6, pp. 1075–1095. ISSN: 00027294, 15481433.

Schegloff, Emanuel A (2007). *Sequence organization in interaction: A primer in conversation analysis*. Vol. 1. Cambridge University Press. ISBN: 978-0-52182-572-6. DOI: 10.1017/CBO9780511791208.

Schneider, Sebastian and Franz Kummert (2017). "Exploring embodiment and dueling bandit learning for preference adaptation in human-robot interaction". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 1325–1331. DOI: 10.1109/ROMAN.2017.8172476.

Schulz, Hans-Jörg, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann (Dec. 2013). "A Design Space of Visualization Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12, pp. 2366–2375. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.120.

Setlur, Vidya, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang (2016). "Eviza: A Natural Language Interface for Visual Analysis". In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. UIST '16. Tokyo, Japan: ACM, pp. 365–377. ISBN: 978-1-4503-4189-9. DOI: 10.1145/2984511.2984588.

Setlur, Vidya, Melanie Tory, and Alex Djalali (2019). "Inferencing Underspecified Natural Language Utterances in Visual Analysis". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: ACM, pp. 40–51. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302270.

Settles, Burr (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report 1648.

Shamekhi, Ameneh, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett (2016). "An Exploratory Study Toward the Preferred Conversational Style for Compatible Virtual Agents". In: *Intelligent Virtual Agents*. Ed. by David Traum, William Swartout, Peter Khooshabeh, Stefan Kopp, Stefan Scherer, and Anton Leuski. Cham: Springer International Publishing, pp. 40–50. ISBN: 978-3-319-47665-0.

Shechtman, Nicole and Leonard M. Horowitz (2003). "Media Inequality in Conversation: How People Behave Differently when Interacting with Computers and People". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. Ft. Lauderdale, Florida, USA: ACM, pp. 281–288. ISBN: 1-58113-630-7. DOI: 10.1145/642611.642661.

Shivaswamy, Pannaga and Thorsten Joachims (2012). "Online Structured Prediction via Coactive Learning". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, pp. 59–66. ISBN: 978-1-4503-1285-1.

Shneiderman, Ben (1996). "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations". In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*. VL '96. Washington, DC, USA: IEEE Computer Society, pp. 336–. ISBN: 0-8186-7508-X.

Siddiqui, Tarique, Albert Kim, John Lee, Karrie Karahalios, and Aditya Parameswaran (Nov. 2016). "Effortless Data Exploration with Zenvisage: An Expressive and Interactive Visual Analytics System". In: *Proceedings of the VLDB Endowment - Proceedings of the 42st International Conference on Very Large Data Bases*. VLDB '16 10.4, pp. 457–468. ISSN: 2150-8097. DOI: 10.14778/3025111.3025126.

Siegel, Noah, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi (2016). "FigureSeer: Parsing Result-Figures in Research Papers". In: *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, pp. 664–680. ISBN: 978-3-319-46478-7. DOI: 10.1007/978-3-319-46478-7_41.

Simkin, David and Reid Hastie (1987). "An Information-Processing Analysis of Graph Perception". In: *Journal of the American Statistical Association* 82.398, pp. 454–465. DOI: 10.1080/01621459.1987.10478448.

Sinha, Rashmi and Kirsten Swearingen (2002). "The Role of Transparency in Recommender Systems". In: *CHI 02 Extended Abstracts on Human Factors in Computing Systems*. CHI EA 02. Minneapolis, Minnesota, USA: Association for Computing Machinery, pp. 830831. ISBN: 1581134541. DOI: 10.1145/506443.506619.

Skau, Drew and Robert Kosara (2016). "Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts". In: *Computer Graphics Forum* 35.3, pp. 121–130. DOI: 10.1111/cgf.12888. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12888.

Sokolov, Artem, Julia Kreutzer, Christopher Lo, and Stefan Riezler (Aug. 2016). "Learning Structured Predictors from Bandit Feedback for Interactive NLP". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1610–1620. DOI: 10.18653/v1/P16-1152.

Spence, Ian and Stephan Lewandowsky (1991). "Displaying proportions and percentages". In: *Applied Cognitive Psychology* 5.1, pp. 61–77. DOI: 10.1002/acp.2350050106. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050106.

Sprague, David and Melanie Tory (2012). "Exploring how and why people use visualizations in casual contexts: Modeling user goals and regulated motivations". In: *Information Visualization* 11.2, pp. 106–123. DOI: 10.1177/1473871611433710. eprint: https://doi.org/10.1177/1473871611433710.

Srinivasan, Arjun and John Stasko (2018). "Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1, pp. 511–521. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2745219.

Srinivasan, Arjun and John T. Stasko (2017). "Natural Language Interfaces for Data Analysis with Visualization: Considering What Has and Could Be Asked". In: *EuroVis 2017 - Short Papers*. Ed. by Barbora Kozlikova, Tobias Schreck, and Thomas Wischgoll. The Eurographics Association. ISBN: 978-3-03868-043-7. DOI: 10.2312/eurovisshort.20171133.

Srinivasan, Arjun, Steven M. Drucker, Alex Endert, and John Stasko (2019a). "Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1, pp. 672–681. ISSN: 1077-2626. DOI: 10.1109/TVCG.2018.2865145.

Srinivasan, Arjun, Mira Dontcheva, Eytan Adar, and Seth Walker (2019b). "Discovering Natural Language Commands in Multimodal Interfaces". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. Marina del Ray, California: ACM, pp. 661–672. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302292.

Srinivasan, Arjun, Bongshin Lee, Nathalie Henry Riche, Steven M. Drucker, and Ken Hinckley (2020). *InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices*. arXiv: 2001.06423 [cs.HC].

StatLib (2005). *StatLib – Datasets Archive*. http://lib.stat.cmu.edu/datasets. [Online; accessed 22-May-2019].

Stevens, S. S. (1946). "On the Theory of Scales of Measurement". In: *Science* 103.2684, pp. 677–680. ISSN: 0036-8075. DOI: 10.1126/science.103.2684.677. eprint: https://science.sciencemag.org/content/103/2684/677.full.pdf.

Stolte, Chris, Diane Tang, and Pat Hanrahan (2002). "Polaris: a system for query, analysis, and visualization of multidimensional relational databases". In: *IEEE Transactions on Visualization and Computer Graphics* 8.1, pp. 52–65. ISSN: 1077-2626. DOI: 10.1109/2945.981851.

Sturgeon, Timothy, Johannes Van Biesebroeck, and Gary Gereffi (Apr. 2008). "Value chains, networks and clusters: reframing the global automotive industry". In: *Journal of Economic Geography* 8.3, pp. 297–321. ISSN: 1468-2702. DOI: 10.1093/jeg/lbn007. eprint: https://academic.oup.com/joeg/article-pdf/8/3/297/2775208/lbn007.pdf.

Sui, Yanan and Joel W. Burdick (2017). "Correlational Dueling Bandits with Application to Clinical Treatment in Large Decision Spaces". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2793–2799. DOI: 10.24963/ijcai.2017/389.

Sun, Dong, Renfei Huang, Yuanzhe Chen, Yong Wang, Jia Zeng, Mingxuan Yuan, Ting-Chuen Pong, and Huamin Qu (2019). "PlanningVis: A Visual Analytics Approach to Production Planning in Smart Factories". In: *CoRR* abs/1907.12201. arXiv: 1907.12201.

Sun, Yiwen, Jason Leigh, Andrew Johnson, and Sangyoon Lee (2010). "Articulate: A Semi-automated Model for Translating Natural Language Queries into Meaningful Visualizations". In: *Smart Graphics: 10th International Symposium on Smart Graphics, Banff, Canada, June 24-26, 2010 Proceedings*. Ed. by Robyn Taylor, Pierre Boulanger, Antonio Krüger, and Patrick Olivier. DOI: 10.1007/978-3-642-13544-6_18. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 184–195. ISBN: 978-3-642-13544-6.

Sun, Yiwen, Jason Leigh, Andrew Johnson, and Barbara Di Eugenio (Jan. 2013). "Articulate: Creating meaningful visualizations from natural language". In: pp. 218–235. DOI: 10.4018/978-1-4666-4309-3.ch011.

– (2014). "Articulate: Creating meaningful visualizations from natural language". In: *Innovative Approaches of Data Visualization and Visual Analytics*. IGI Global, pp. 218–235.

Sydow, André, Jan-Frederik Kassel, and Michael Rohs (2015). "Visualizing Scheduling: A Hierarchical Event-Based Approach on a Tablet". In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI '15. Copenhagen, Denmark: ACM, pp. 728–734. ISBN: 978-1-4503-3653-6. DOI: 10.1145/2786567.2793694.

Thompson, William R. (1933). "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3/4, pp. 285–294. ISSN: 00063444.

Todorovic, Dejan (2008). "Gestalt principles." In: 3.12, p. 5345.

Toker, Dereck, Cristina Conati, Giuseppe Carenini, and Mona Haraty (2012). "Towards Adaptive Information Visualization: On the Influence of User Characteristics". In: *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*. UMAP12. Montreal, Canada: Springer-Verlag, pp. 274285. ISBN: 9783642314537. DOI: 10.1007/978-3-642-31454-4_23.

Tory, Melanie and Vidya Setlur (2019). “Do What I Mean, Not What I Say! Design Considerations for Supporting Intent and Context in Analytical Conversation”. In: *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 93–103. DOI: 10.1109/VAST47406.2019.8986918.

Tufte, Edward R. (1986). *The Visual Display of Quantitative Information*. Cheshire, CT, USA: Graphics Press. ISBN: 0-9613921-0-X.

Tukey, John W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Tukey, John Wilder and Martin Bradbury Wilk (1966). “Data Analysis and Statistics: An Expository Overview”. In: *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*. AFIPS ’66 (Fall). San Francisco, California: ACM, pp. 695–709. DOI: 10.1145/1464291.1464366.

VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert (2018). “Altair: Interactive Statistical Visualizations for Python”. In: *Journal of Open Source Software*. DOI: 10.21105/joss.01057.

Vartak, Manasi, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis (Sept. 2015). “SeeDB: Efficient Data-driven Visualization Recommendations to Support Visual Analytics”. In: *Proceedings of the VLDB Endowment - Proceedings of the 41st International Conference on Very Large Data Bases*. VLDB ’15 8.13, pp. 2182–2193. ISSN: 2150-8097. DOI: 10.14778/2831360.2831371.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 5998–6008.

Velez, Maria C., Deborah Silver, and Marilyn Tremaine (2005). “Understanding visualization through spatial ability differences”. In: *VIS 05. IEEE Visualization, 2005.* Pp. 511–518.

Veras, Rafael and Christopher Collins (2019). “Discriminability Tests for Visualization Effectiveness and Scalability”. In: *CoRR* abs/1907.11358.

Wallsten, Thomas S., David V. Budescu, Rami Zwick, and Steven M. Kemp (1993). “Preferences and reasons for communicating probabilistic information in verbal or numerical terms”. In: *Bulletin of the Psychonomic Society* 31.2, pp. 135–138. ISSN: 0090-5054. DOI: 10.3758/BF03334162.

Weaver, Chris, David Fyfe, Anthony Robinson, Deryck Holdsworth, Donna Peuquet, and Alan M. MacEachren (2006). “Visual Analysis of Historic Hotel Visitation Patterns”. In: *2006 IEEE Symposium On Visual Analytics Science And Technology*, pp. 35–42. DOI: 10.1109/VAST.2006.261428.

Weinschenk, Susan and Dean T. Barker (2000). *Designing Effective Speech Interfaces*. New York, NY, USA: John Wiley & Sons, Inc. ISBN: 0-471-37545-4.

Weiser, Mark (July 1999). “The Computer for the 21st Century”. In: *SIGMOBILE Mob. Comput. Commun. Rev.* 3.3, pp. 311. ISSN: 1559-1662. DOI: 10.1145/329124.329126.

Weld, Daniel S. and Gagan Bansal (May 2019). “The Challenge of Crafting Intelligible Intelligence”. In: *Commun. ACM* 62.6, pp. 7079. ISSN: 0001-0782. DOI: 10.1145/3282486.

Wharton, Cathleen, John Rieman, Clayton Lewis, and Peter Polson (1994). "Usability Inspection Methods". In: ed. by Jakob Nielsen and Robert L. Mack. New York, NY, USA: John Wiley & Sons, Inc. Chap. The Cognitive Walkthrough Method: A Practitioner's Guide, pp. 105–140. ISBN: 0-471-01877-5.

Wilkinson, Leland (2005). *The Grammar of Graphics*. Statistics and Computing. Springer. ISBN: 0387245448.

Wilson, Aaron, Alan Fern, and Prasad Tadepalli (2012). "A Bayesian Approach for Policy Learning from Trajectory Preference Queries". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., pp. 1133–1141.

Wongsuphasawat, Kanit, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer (2016). "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 649–658. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467191.

Wongsuphasawat, Kanit, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer (2017). "Voyager 2: Augmenting Visual Analysis with Partial View Specifications". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: ACM, pp. 2648–2659. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025768.

Wu, Huasen and Xin Liu (2016). "Double Thompson Sampling for Dueling Bandits". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 649–657.

Xu, Panpan, Honghui Mei, Liu Ren, and Wei Chen (2017). "ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories". In: *IEEE Transactions on Visualization and Computer Graphics* 23.1, pp. 291–300. ISSN: 2160-9306. DOI: 10.1109/TVCG.2016.2598664.

Yalçın, Mehmet Adil, Niklas Elmqvist, and Benjamin B. Bederson (2018). "Keshif: Rapid and Expressive Tabular Data Exploration for Novices". In: *IEEE Transactions on Visualization and Computer Graphics* 24.8, pp. 2339–2352. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2723393.

Yankelovich, Nicole, Gina-Anne Levow, and Matt Marx (1995). "Designing SpeechActs: Issues in Speech User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '95. Denver, Colorado, USA: ACM Press/Addison-Wesley Publishing Co., pp. 369–376. ISBN: 0-201-84705-1. DOI: 10.1145/223904.223952.

Yi, Ji Soo, Youn ah Kang, John Stasko, and Julie Jacko (Nov. 2007). "Toward a Deeper Understanding of the Role of Interaction in Information Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 13.6, pp. 1224–1231. ISSN: 1077-2626. DOI: 10.1109/TVCG.2007.70515.

Young, Steve, Milica Gašić, Blaise Thomson, and Jason D. Williams (2013). "POMDP-Based Statistical Spoken Dialog Systems: A Review". In: *Proceedings of the IEEE* 101.5, pp. 1160–1179. ISSN: 0018-9219. DOI: 10.1109/JPROC.2012.2225812.

Yu, B. and C. T. Silva (2020). "FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1, pp. 1–11. ISSN: 2160-9306. DOI: 10.1109/TVCG.2019.2934668.

Yue, Yisong, Josef Broder, Robert Kleinberg, and Thorsten Joachims (Sept. 2012). "The K-armed Dueling Bandits Problem". In: *Journal of Computer and System Sciences* 78.5, pp. 1538–1556. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2011.12.028.

Zgraggen, Emanuel, Zheguang Zhao, Robert Zeleznik, and Tim Kraska (2018). "Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI 18. Montreal QC, Canada: Association for Computing Machinery. ISBN: 9781450356206. DOI: 10.1145/3173574.3174053.

Zhao, Xiaoxue, Weinan Zhang, and Jun Wang (2013). "Interactive collaborative filtering". In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. CIKM '13. San Francisco, California, USA: ACM, pp. 1411–1420. ISBN: 978-1-4503-2263-8. DOI: 10.1145/2505515.2505690.

Zhou, Li (2015). *A Survey on Contextual Multi-armed Bandits*. arXiv: 1508.03326 [cs.LG].

Ziemkiewicz, Caroline, R. Jordan Crouser, Ashley Rye Yauilla, Sara L. Su, William Ribarsky, and Remco Chang (2011). "How locus of control influences compatibility with visualization style". In: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 81–90. DOI: 10.1109/VAST.2011.6102445.

Ziemkiewicz, Caroline, Alvitta Ottley, R. Jordan Crouser, Krysta Chauncey, Sara L. Su, and Remco Chang (2012). "Understanding Visualization by Understanding Individual Users". In: *IEEE Computer Graphics and Applications* 32.6, pp. 88–94. ISSN: 0272-1716. DOI: 10.1109/MCG.2012.120.

Zoghi, Masrour, Zohar S Karnin, Shimon Whiteson, and Maarten de Rijke (2015). "Copeland Dueling Bandits". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., pp. 307–315.

Zuur, Alain F., Elena N. Ieno, and Chris S. Elphick (2010). "A protocol for data exploration to avoid common statistical problems". In: *Methods in Ecology and Evolution* 1.1, pp. 3–14. DOI: 10.1111/j.2041-210X.2009.00001.x.

Acronyms

CAT Communication Accommodation Theory.

CDF Cumulative Distribution Function.

CI Conversational Interface.

CPU Central Processing Unit.

CW Cognitive Walkthrough.

D-TS Double Thompson Sampling.

EDA Exploratory Data Analysis.

GUI Graphical User Interface.

H Entropy.

HCI Human-Computer Interaction.

HE Heuristic Evaluation.

I Mutual Information.

KDE Kernel Density Estimation.

MAB Multi-armed Bandit.

MS-SSIM Multi-Scale Structural Similarity Index.

MTurk Amazon's Mechanical Turk.

NASA-TLX NASA Task Load Index.

nI Normalized Mutual Information.

NLG Natural Language Generation.

NLI Natural Language Interfaces.

NLP Natural Language Processing.

NLU Natural Language Understanding.

PCA Principal Component Analysis.

POMDP Partially observable Markov decision process.

REST Representational State Transfer.

SDS Spoken Dialogue System.

SNS Subjective Numeracy Score.

SUS System Usability Score.

UCI University of California, Irvine.

UI User Interface.

UX User Experience.

WIMP Window, Icon, Menu, and Pointer.

Curriculum Vitae

Name	Jan-Frederik Kassel
Date of Birth	22.05.1990
Place of Birth	Hannover

Education

2012 – 2015	M.Sc. Computer Science *Leibniz Universität Hannover*, Hannover
2009 – 2012	B.Sc. Computer Science *Leibniz Universität Hannover*, Hannover
2009	Abitur *IGS Roderbruch*, Hannover

Professional Experience

08.2020 – present	Product Strategist *Volkswagen AG*, Wolfsburg
09.2018 – 08.2020	Product Owner *Data:Lab, Volkswagen AG*, Munich
09.2015 – 08.2018	Data Scientist & PhD Candidate *Data:Lab, Volkswagen AG*, Munich
04.2014 – 04.2015	Intern *GroupIT, Volkswagen AG*, Wolfsburg
10.2011 – 09.2013	Tutor *Leibniz Universität Hannover*, Hannover

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Figures are created with seaborn.