Visual Informatics 3 (2019) 177–191

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf

You get by with a little help: The effects of variable guidance degrees on performance and mental state☆

Davide Ceneda *, Theresia Gschwandtner, Silvia Miksch

TU Wien, Faculty of Informatics, Institute of Visual Computing & Human-Centered Technology, Favoritenstrasse 9-11/193, A-1040 Vienna, Austria

ARTICLE INFO

Article history:
Received 8 July 2019
Received in revised form 16 October 2019
Accepted 25 October 2019
Available online 2 November 2019

Keywords:
Guidance
User study
Knowledge
Trust
Mixed-initiative
Visual data analysis

ABSTRACT

Since it can be challenging for users to effectively utilize interactive visualizations, guidance is usually provided to assist users in solving tasks. Guidance is mentioned as an effective mean to overcome stall situations occurring during the analysis. However, the effectiveness of a peculiar guidance solution usually varies for different analysis scenarios. The same guidance may have different effects on users with (1) different levels of expertise. The choice of the appropriate (2) degree of guidance and the type of (3) task under consideration also affect the positive or negative outcome of providing guidance. Considering these three factors, we conducted a user study to investigate the effectiveness of variable degrees of guidance with respect to the user's previous knowledge in different analysis scenarios. Our results shed light on the appropriateness of certain degrees of guidance in relation to different tasks, and the overall influence of guidance on the analysis outcome in terms of user's mental state and analysis performance.

1. Introduction

Mixed-initiative visual data analysis (Horvitz, 1999) is an effective and powerful way to make sense of large data collections and support the completion of complex tasks. In this kind of analysis, the strengths of users and computational systems are joint to reach a common analytical goal. On the one hand, users are enabled to make sense of the data through external cognition. On the other hand, the computational system offers the means to execute complex calculations, elaborate statistics, or discover patterns (Gibson, 1977).

Although visual solutions have been proved to be effective in their scope (Bederson and Shneiderman, 2003; Keim et al., 2008; Cook and Thomas, 2005), the research is still far from achieving an effective mixed-initiative integration in which the affordances of the user and the analysis system are balanced (Gibson, 1977; Bertini and Lalanne, 2009; Ceneda et al., 2019). Therefore, sometimes it can be challenging to effectively use and interact with sophisticated analytical solutions. As a consequence, the analysis may stall.

In the past, many approaches have been developed in the attempt to reduce the burden on users and help them to make sense of the data and the visual interfaces. Ceneda et al. (2017) categorize these methods as *guidance*. Guidance describes the results of enabling an effective human–computer collaboration. In particular, guidance deals with providing a solution to the needs a user develops while performing analysis tasks. These needs are referred to as *knowledge gaps*. Ideally, the guidance process could provide a variety of supporting indications to the user, ranging from hints and recommendations, to step-by-step instructions, to foster a positive outcome of the analysis, solving the aforementioned knowledge gaps and a solution to the stalled analysis.

Although the definition of guidance is quite new, guidance approaches have been around for quite some time (Horvitz, 1999). Therefore, it does not surprise the number of approaches showing the benefits of providing guidance during the analysis process. Although the benefits of guidance are clear (Ceneda et al., 2019), what is still not clear is how the effectiveness of the guidance varies according to the user to whom it is provided, and to the task at hand. For instance, different types of guidance may be more effective to support exploratory analysis, while others to verify hypotheses. Furthermore, the effectiveness of guidance may also vary according to the previous knowledge of the user, for instance, if it is a novice user or he/she possesses some knowledge about the analysis domain and the visualization system. Thus, we conducted a user study investigating how different guidance degrees affect users with different levels of previous knowledge to solve different kinds of tasks. We pursued this aim, not only by investigating the repercussions of guidance on task

☆ The title is a reference to a popular song by The Beatles. It should communicate that VA is not hard if guidance is available.

* Corresponding author.
E-mail address: davide.ceneda@tuwien.ac.at (D. Ceneda).

Peer review under responsibility of Zhejiang University and Zhejiang University Press.

https://doi.org/10.1016/j.visinf.2019.10.005

performance, but also on how the provision of guidance may affect the user's mental state. We achieve this by analyzing for instance, how the provision of (or the lack of) guidance may induce frustration, feelings of being lost, improve the user's confidence in results, etc.. We think that this work is useful to designers who intend to create guided visual data analysis systems, fostering an increased awareness of users' needs and the development of mixed-initiative systems. In summary, our main contributions are:

- Investigating the interdependencies among guidance degree, user expertise, task performance, and mental state of the user.
- Describing the impact of different degrees of guidance on task performance and mental state of the user.
- Elaborating the impact of user expertise on task performance, and mental state.

2. Related work

Our work elaborates concepts from two main research topics in literature: guidance in visualization and the dynamics of user's mental state during the analysis.

2.1. Guidance in visualization

Guidance is a research topic that comprises Human–Computer Interaction, Information Visualization and Visual Analytics (Keim et al., 2008; Dix et al., 2004; Smith and Mosier, 1986). Guidance has its roots in mixed-initiative data analysis (Horvitz, 1999) and it contemplates the assistance the user receives from the system, as well as the guidance the user gives to the system to steer the analysis (Ceneda et al., 2018). Guidance describes what are the benefits deriving from a mixed-initiative analysis and how this collaborative analysis can take place. Formally speaking, guidance is defined as a *"computer-assisted process that aims to actively resolve a knowledge-gap during an interactive"* visual analysis session (Ceneda et al., 2017, p. 2). In simple words, the main goal of guidance is to solve a particular user need, namely a user's *knowledge gap*, which can be seen as the difference between the user's knowledge and the knowledge required to complete a given task. This gap may be related to different aspects of the analysis, like the lack of proper interaction means, or of specific domain-related concepts necessary to interpret the data. The output of the guidance process is an answer to the knowledge gap, that is provided to the user, in some visual form. Different degrees of guidance may be provided in order to meet the user's needs. Ceneda et al. (2017) describe different guidance degrees resulting in different types of guidance. In practical scenarios, the same task can be supported with different guidance: In Fig. 3, the search for specific data in a time-series can be supported with no guidance, but also with direct recommendations, or by prescribing actions.

The amount of works dealing with guidance is vast: Ceneda et al. (2019) recently reviewed the literature of guidance approaches in visual data analysis. Guidance ranges from recommender systems (Wongsuphasawat et al., 2016; Gotz and Wen, 2009) to user's modeling (Brusilovsky and Millán, 2007; Mazurowski et al., 2010). In the following, we describe guidance approaches and differentiate our work from previous evaluation studies.

Willett et al. (2007) introduced *scented widgets*, which are common UI-elements enhanced with knowledge derived from other users' interaction choices. The authors underline that the introduction of such elements may flatten the difference in performances between expert and novice users. Gotz and Wen (2009) introduced behavior-driven recommendations, showing improvements in the completion time and correctness of results. In the field of data mining, Bernstein et al. (2005) developed an intelligent ontology-based assistant that supports the choice of proper data mining algorithms with respect to the specific problem setting. Their results suggest that also expert users need guidance. Streit et al. (2012) generate an analysis model that is used for supporting analysts with their tasks. The advantage of this work is the provision of different degrees of guidance.

Similarly to these approaches, we want to evaluate whether the introduction of guidance leads to performance improvements among study participants. However, our aim goes beyond the evaluation of the effectiveness of a specific tool. In fact, in contrast to such approaches, *we aim to understand how such effectiveness varies according to the task, and what are the effects of different degrees of guidance in relation to different levels of user's expertise*.

2.2. User's knowledge and mental state

Chen (2005) distinguishes between two main types required to make progresses during an analysis: *operational* and *domain* knowledge, whether the user is able to *interact* in a effective way with the analysis tool, or possesses the necessary *domain* notions to interpret the context and the data. In our vision, different types of guidance may be necessary according to what kind of knowledge the user is missing. Thus, our aim is to investigate if similar guidance degrees have different effects according to the knowledge gap, i.e., lack of operational or domain knowledge.

A last research branch related to our work, is the one studying the relations between the visual analysis and the development of the user's mental state and sentiments, and the effects of such feelings on the analysis itself. Sacha et al. (2016) point out that during an analysis, there is always a match between the uncertainty present in the data and the trust that users develop while proceeding with the analysis. The more the exploration advances, the more their trust grows. Although not explicitly mentioned, also the guidance may contribute to increase or decrease the user's trust. Many other psychological aspects are also connected to the analysis process. Celik et al. (2013) point out that frustration and sadness are often connected to the inability to perform a task. Similarly, Kapoor et al. (2007) show that it is possible to automatically infer and predict the growth of frustration, during the execution of a task, and thus they identify possible thresholds for triggering guidance. However, the effects of guidance on user's frustration, and in general on users' mental state, have not been studied yet.

In order to understand how guidance affects the development of sentiments during data analysis, we asked participants of a user study to rate their degree of frustration, trust, and confidence after solving a set of tasks. We then relate these values to the provided guidance, to the user's expertise level, and to the analysis outcome.

3. Aims and terminology

The aim of this work can be summarized by the following questions:

(1) How do different guidance degrees affect the performance and the mental state of users with different degrees of previous knowledge?
(2) Do the effectiveness and the effects of guidance vary according to the type of task the user has to solve?

Our assumption is that three dimensions play an important role in the design of guidance for visual data analysis, these are the task type, the knowledge of the user, and the guidance degree. We want to test how the variation of one of such dimensions

influences the others, and ultimately the analysis outcome, in terms of user performance and mental state. We start describing these three dimensions, before formalizing our aim in terms of rigorous hypotheses.

Knowledge and task types. The first factor we describe, is the knowledge required to complete a task. Two kinds of knowledge are usually required to complete a visual data analysis: *operational* and *domain knowledge*. Our aim is to test whether the type of knowledge involved influence the effectiveness of the provided degree of guidance. According to the distinction between operational and domain knowledge, it is possible to discern two general types of tasks:

- *Exploratory tasks* relate to operational knowledge and therefore to the ability to interact with the tool. These tasks require basic interaction abilities, like choosing among different interaction means (e.g.,filter, selection) and using them effectively.
- *Domain tasks* require domain specific knowledge to be successfully completed. These tasks are related to the ability to reason and connect a given domain concept to the task and data under analysis.

User knowledge. A second dimension we address, is the distinction between degrees of user's competence. Usually, a lack of user's knowledge, which can be also seen as the difference between the knowledge required to solve the task and the actual user's knowledge, is what we call a knowledge gap. When this happens, the user might have a hard time completing the task, and the analysis may stall. We hypothesize that different degrees of guidance would have different effects on users with different levels of knowledge relevant to the task. We distinguish between:

- *Knowledgeable users,* possessing the knowledge required to complete the task (i.e., operational or domain knowledge, see previous paragraph).
- *Novice users*, who may not possess the knowledge required to complete the task, with the exception of previous expertise.

Guidance degrees. Finally, we distinguish among three different degrees of guidance which we provided to the study participants to assist the completion of their tasks. Our assumption is that users with different knowledge may require different degree of guidance, and that the fact of having more or less support while solving the task would have variable effects on the task performance and the user's mental state. According to the guidance degrees described by Ceneda et al. (2017), we list the types of guidance we included in our study.

- *No guidance:* When no guidance is provided, users have to solve the tasks on their own. This is translated into the provision of simple visualizations, without any further support. According to Ceneda et al. the provision of additional aggregated values (i.e., min, max, outliers) does not constitute higher guidance.
- *Directing guidance:* This kind of guidance aims at providing different analysis options. Therefore, on top of the basic visualization, we indicate possible analysis paths. In the specific, interesting data subsets are recommended to the user, but the system may also recommend actions to proceed the investigation.
- *Prescribing guidance:* This is the highest degree of guidance. It aims at providing step-by-step instructions to reach the result. Among the different analysis paths and recommendations (see Directing guidance), the system picks one and provides it to the user, who must follow the indications (the different steps) to reach the final result.

The aforementioned three dimensions concur to the provision of effective support to the user. Considering all of them together allowed us to reason about the effectiveness of different guidance types in different situations. In particular, we investigate (1) if guidance can compensate for a lack of user's knowledge i.e., if there is a noticeable difference among novice and knowledgeable users supported with similar degrees of guidance. We examine (2) if some degrees of guidance are better suited than others for a given task type i.e., if some degrees of guidance are better suited for exploratory analysis or to complete domain tasks. Furthermore, by comparing the results of same users under different conditions of guidance, we study (3) if guidance can affect, in positive or negative, the performance of such users and the development of sentiments and mental map.

4. Hypotheses

We formalize the aim described earlier in terms of different research hypotheses, which we grouped into two hypotheses groups, H1 and H2. Hypotheses in H1 focus on the variation of user's performance metrics, while those in H2 consider the mental state and the feelings of the user, in response to the provided guidance.

Hypotheses group H1. first aim was to investigate the effects of guidance on task performance. At first, we analyze the effects of guidance on novice users and evaluate if the positive effect of guidance is mitigated in knowledgeable users. Our assumption is that knowledgeable users may still benefit from guidance in terms of reduced task completion time. This is formalized in the following hypotheses:

H1.1 A high degree of guidance causes significant improvements in task performance (timings, correctness, distance, total-steps) of novice users.
H1.2 A high degree of guidance reduces completion time and amount of steps for knowledgeable users.

Hypotheses group H2. Our second aim was to evaluate the mental state of users when receiving different degrees of guidance. In particular, we wanted to understand if guidance causes a positive effect on the user's mental state. Specifically, if guidance increases the user's confidence in the analysis results, or if in some situations the guidance can frustrate the user. This is formalized in the following hypotheses:

H2.1 A high degree of guidance causes a significant improvement in participants' confidence in their results.
H2.2 A high degree of guidance causes more frustration for knowledgeable users than for novice users.

5. Study design

To verify the hypotheses in H1 and H2, we designed a user study comprising six specific tasks (3 exploratory + 3 domain tasks) which we asked 65 participants under different conditions of expertise and guidance to solve. The hypotheses lead us in designing the evaluation environment and the evaluation procedure as follows.

5.1. Data

We use a dataset from the USGS program of research and observation in San Francisco Bay (Cloern and Schraga, 2016). This dataset in combination with a careful task design allowed us to evaluate the effectiveness of guidance on both exploratory and domain tasks. The dataset contains multiple daily measurements of water samples collected along the 145 km transect of

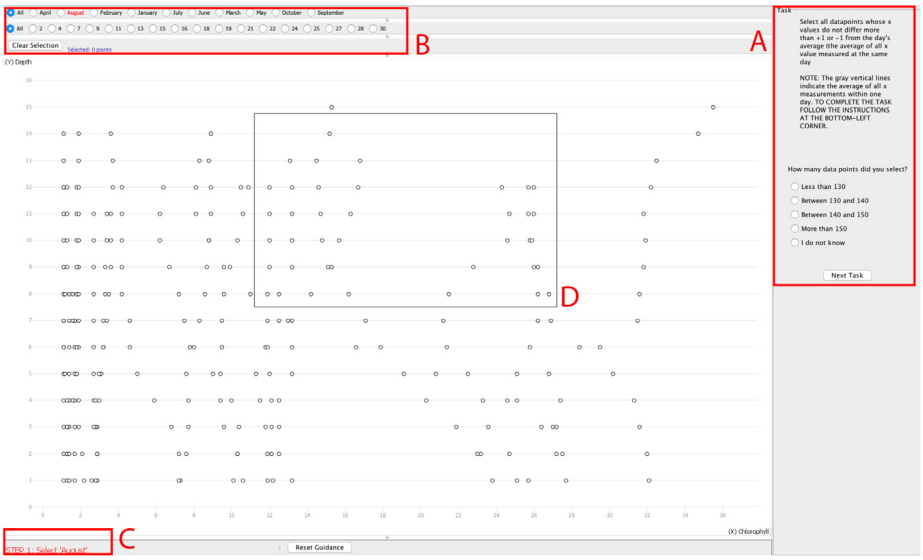


Fig. 1. Interface of the evaluation environment. In the upper right corner (A), a text-box shows the current task. A further text-box, gives indications to interpret the guidance suggestions. On the top left (B), some combo-boxes give the possibility to filter the dataset. On the bottom of the visualization (C), a text-box shows the step-by-step instructions to reach the desired results (in case prescribing guidance was provided). At the center of the visualization (D), the rectangular selection tool is shown.

the San Francisco bay. The whole dataset spreads over various decades (1969-nowadays), but for our study we selected only specific subsets, spanning roughly one year each. In particular, from the main dataset, we extracted *six* subsets. Each dataset was associated with exactly one task to avoid learning effects. Three datasets were used for domain related tasks, while the other three for exploratory tasks. Each dataset is equivalent to the others in terms of number of data dimensions involved. They just differ for the focus on a specific dimension of the original dataset. We complemented the six datasets with derived statistical values (e.g., average, max, min). We used these derived values as a base for *directing* guidance, to point the user to interesting data during the execution of the tasks.

5.2. Participants and evaluation sessions

We had 65 students at bachelor level participating in our study. They all are students in computer science and attended a course in information design and visualization, preceding the study, which implies a certain knowledge about the visual environment they were provided with. Nevertheless, we considered all the students as novice participants, since they never performed the analysis on the given dataset, nor possessed any domain knowledge about the topic. Before presenting the tasks to the students, we conducted a pilot testing with four participants to correct minor errors and fine tune the tests.

For the evaluation sessions we utilized EvalBench (Aigner et al., 2013), a software specifically designed to evaluate interactive visualizations (see Fig. 1). The interactive visualizations were developed with Java, using the Prefuse library (Heer et al., 2005), and TimeBench (Rind et al., 2013) to manage the temporal aspects of the data.

Study structure. The user-study was divided into two subsequent evaluation sessions as detailed in Fig. 2: one session dealing with exploratory tasks and the other session dealing with domain tasks. We divided the participants into two groups, group A and group B, each of them executed both exploratory and domain tasks in the two task sessions, but group A performed exploratory tasks in the first session of the study and domain tasks in the second session, and group B did it the other way around. We did this to avoid learning effects of the participants and compare the execution of the same tasks with different levels of expertise.

At the beginning, both groups received an introduction to the main topics of the user-study. We told them that they were going to execute some tasks and that they were (possibly) going to receive guidance during this execution. They were not given any other information, except that the data regarded biological measurements extracted from water samples of the San Francisco bay, and that they were not allowed to use any external help. We intentionally decided not to provide them with any further information about the interaction means, or about the specific domain concepts in order to simulate the behavior of novice users.

Thus, in the first session, the participants did not have any experience with the data, visualizations, and tasks. Furthermore, we did not provide them with any additional knowledge that might have been required to solve the tasks. For this reason, we considered them novice users. In the second session all participants had already some experience with the data, visualizations, and tasks. In addition, we added a learning session in between the two task sessions to train the participants in the concepts necessary to complete the following tasks. After undergoing the learning phase, we considered these participants knowledgeable users. We chose this study design to ensure that both types of users conduct both types of tasks while avoiding learning effects on the two groups.

Learning session. The first task session was followed by a briefing for the next session — a session in which participants were instructed in domain or in the operational concepts required to solve the three remaining tasks in the subsequent session. Thus, group A, after completing the exploratory tasks, was instructed with domain concepts, necessary to solve the domain tasks session. Group B, instead, received an education about interactive means and the exploratory session. In this learning phase the participants belonging to both groups had also the possibility to interact with a sample tool and further sediment the acquired knowledge. This allowed us to compare the performance of novice users (no prior knowledge) with that of users instructed in concepts relevant for solving the tasks (exploratory and domain tasks), mitigating at the same time the possibility

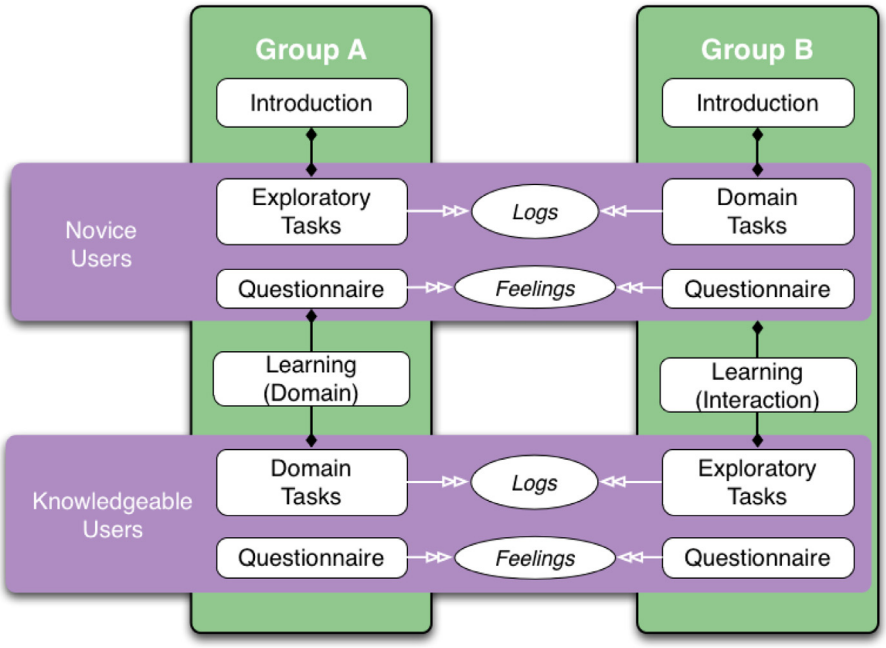


Fig. 2. General structure of the user's study. We conducted two parallel evaluation sessions. After a short introduction, the participants performed two set of tasks. Each task was followed by a series of questions. Between the two task sessions, the participants had an active learning phase, where they were instructed either in interaction concepts or in domain concepts respectively. Subsequently, the participants completed the second set of tasks. A cross structure was chosen to minimize the learning effects on the participants.

Table 1
Performance metrics. We recorded these metrics while the participants executed the tasks. The completion time was provided directly by the evaluation environment EvalBench (Aigner et al., 2013). The others (correctness, distance, steps) were calculated from the interaction logs (see Section 6.1).

Performance	Description
Completion time	A timer measured the interval between the start of the task and the submission of an answer.
Correctness	A real number in [0,1]. This value is a weighted ratio between correctly selected data items and all selected data items.
Distance	A real number in [0,1], measuring the semantic distance of the selected data items from the correct ones.
Total steps	The total number of actions (clicks, filter, etc.) required by a user to complete a task.

Table 2
Indicators of users' mental state. We asked participants to answer some questions regarding their feelings after each task. Each variable was rated on a five-point Likert scale. We then related the users' feelings to the degree of guidance they received, and to their knowledge level.

Mental state	Description
Lost	We asked the participants how lost they felt while executing the task.
Frustrated	We asked the participants how frustrated they felt while executing the task.
Confident	We asked the participants how confident they felt about the correctness of the submitted result.
Easy	We asked the participants to evaluate how easy the task was.
Guidance appropriate	We asked the participants if they considered the guidance they received appropriate to solve the task.

of learning effects on the subsequent series of tasks. In fact, with such cross-structure, the expertise group A acquired while conducting the first session was not needed to complete the subsequent session of domain tasks. The same holds true for the domain knowledge group B acquired during the first session, which was not needed to solve the next tasks. We did not measure precisely the increase of knowledge, in terms of learned concepts, due to the learning session. However, from the results of the study we could see an increase in the number of participants who were able to solve the tasks without guidance after undergoing the learning session. In average, 10% more participants was able to solve exploratory tasks without guidance. This percentage increases to 20% for domain tasks. This means that 20% more participants could solve domain tasks without guidance after learning the appropriate domain concepts.

To collect the data necessary to test H1, the system recorded automatically timings, correctness and the number of operations required to complete the tasks, see Table 1. After the execution of each task, we asked the participants to answer ten questions about the visualizations and the interactive means (i.e., were they sufficient? were they useful?) the tool offered, as we wanted to test whether they interpreted the visual encodings correctly. We then asked them to evaluate the guidance they received. We encoded the possible answers as multiple choices, but we also let the participants add free text if they felt the options provided were not sufficient. To test H2, a further set of questions asked the participant about their feelings while solving the task (see Table 2). All these subjective feelings (Celik et al., 2013; Kapoor et al., 2007) were measured on a five-point Likert scale. At the end, we also collected the interaction logs (e.g., hovering a point, changing the selection, filtering the dataset) for evaluation and for extracting further metrics (see Section 6.2).

5.3. Task design

We designed a total of six tasks: three focused on operational knowledge and three on domain knowledge.

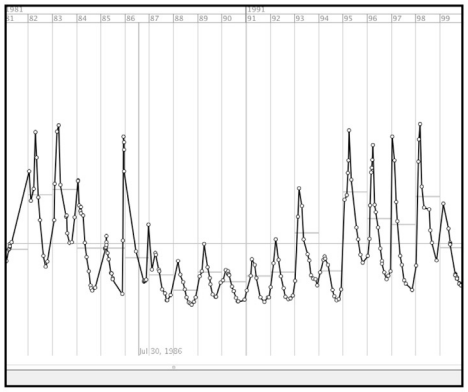
Exploratory tasks. These tasks are related to a user's operational knowledge and his/her ability to interact with the analysis tool.

We required the participants to perform a number of interactions to explore the dataset. We did not relate these tasks to any domain concepts, but rather asked the users to find and select specific data values, without any associated meaning. A typical exploratory task required the participants to isolate data points with certain characteristics by iteratively using the interactive means provided by the tool. In other words, an exploratory task consists of long sequences of selections and filter operations. The number of actions and the reasoning effort required to solve an exploratory task constitutes the main difference to domain tasks. We designed these exploratory tasks so that the only knowledge required to correctly and efficiently solve them was being able to interact with the visualization tool. In comparison, domain tasks required domain knowledge, while almost no interaction, besides simple selections. We designed exploratory tasks in such a way that it was possible for the participants without the advanced interaction means we introduced during the learning session. In total, around half of novice participants were able to complete correctly the exploratory tasks without guidance just by using the basic interaction means offered.

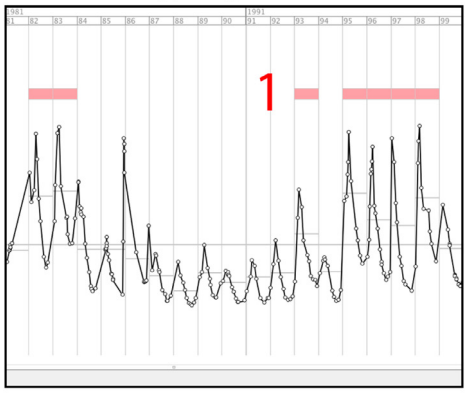
As mentioned earlier, after the first tasks session, we lead the participant through a learning session to let the participants acquire the knowledge needed to solve the following tasks. For what it regards exploratory tasks, we introduced the participants belonging to group B to the use of some advanced interaction techniques, like for instance the rectangular selection of multiple data points and the use of filters, etc. All the interaction means were available to all the participants from the beginning of the study. However, we assumed that the competences we taught to group B during the learning session would allow them to complete the tasks more efficiently, in respect to novice users. Furthermore, while knowledgeable users were presented and had time to experiment the different interaction means, novice users had to discover them while solving the tasks, marking another difference between the two groups. As a consequence, we expected a difference in the performance, as well as in the frustration level and confidence of these two types of users.

Domain tasks. The same design principles led the design of domain tasks. We based these three tasks on three specific domain concepts: hyper-salinity of sea water, periods of droughts in a given year, and dangerous low concentrations of nutrients in the bay water. In particular, we asked the participant to recognize a period of drought and a condition of hyper-salinity by analyzing different time-series showing the development of water salinity in a given time period. In a third task, we asked the participants to reason about low concentration of nutrients by exploring a scatter-plot visualizing water nutrients at different depths in a specific region of the bay.

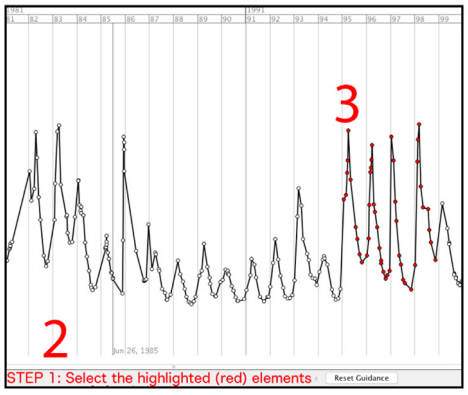
Also for these tasks, we worked to mark a difference between novice and knowledgeable participants, by providing the latter group, during the learning session, an introduction to these domain concepts, including exercises to consolidate the knowledge. For instance, in one of the domain tasks, the participants were requested to select all the data points corresponding to a period of drought. We explained how to recognize these periods just to the knowledgeable users, while the novice ones relied just on their individual idea of the concept and on the guidance suggestions, telling them that for instance, a clear sign of a period of drought is the raise of average water salinity in a given period. Thus, for novice users who did not receive any guidance it was sometimes not possible to find the correct answer to such tasks. In total, just one third of novice participants was able to complete correctly the domain tasks with no guidance. The learning session affected the participants' performance. In average, 20% more participants completed correctly the domain tasks with no guidance.



(a) No guidance



(b) Directing guidance



(c) Prescribing guidance

Fig. 3. The same domain task is supported with different degrees of guidance. (a) no guidance: a time-series line chart shows the variation of water salinity in different years. (b) directing guidance: possibly interesting data is highlighted to address the analysis (e.g., data representing high salinity values). These periods are signaled in red at the top of the visualization (1); (c) prescribing guidance: step-by-step instructions are presented to the user (2) together with the highlighting of interesting data points (3).

5.4. Concrete task examples

To give the reader a better idea about the task design, we describe one domain and one exploratory task in more detail. For completeness, we created two more tasks for each task type, for a total of six tasks. In the following, we describe just one of them, for each type.

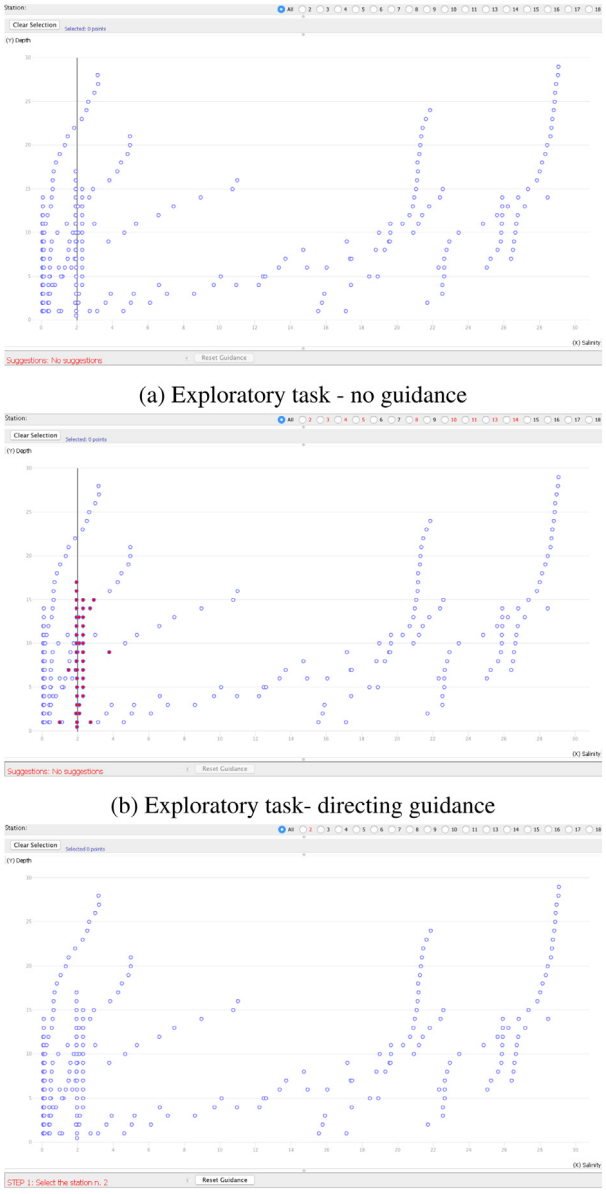
Domain task. The participants had to solve the following domain task under different conditions of guidance. We asked them to "*select all the data points falling in the longest period of drought*".

To solve this task, the study participants were presented a line-chart visualizing the fluctuation of salt concentration in a given period. On top of this basic visualization we added guidance. All the visualizations used for this task are illustrated in Fig. 3. The figures show the encoding of the three guidance degrees. When solving the task, a user would either know directly (if knowledgeable) or possibly reason (if novice) that a period of drought affects the mineral composition of the water. For sea water, one of the most obvious results is that the concentration of salt increases. As a consequence, a user should have selected as the correct answer the longest period with the highest salinity values.

Aside the line chart, additional lines encoded the average salinity values of every visualized year. We shared this same visualization for all the different guidance degrees. On top of this visualization, we added additional visual clues to support increasing levels of guidance. For instance, when *directing guidance* was provided, we highlighted data points of years with particularly high average temperatures and salinity values (Fig. 3b). These hints point users towards data regions/subsets that are helpful to solve the task. Directing guidance, per definition, does not give exact instructions to solve a task but rather recommends and directs the user towards interesting data regions. In the last scenario, *prescribing guidance* was provided. We led participants along a selected analysis path. While the users could freely interact with the tool, we provided them with precise step-by-step instructions in textual form to follow this chosen analytical path and find the correct answer. Since the task outlined in (Fig. 3b) is a domain task it requires domain knowledge and reasoning to solve it rather than a operational knowledge. With the aim of limiting the effect of operational knowledge on the resolution of such tasks, we limited the required interactions to simple selections. In case of *prescribing guidance*, this meant that we highlighted the correct data points and asked users to select them by simply clicking on them.

For all the three guidance degrees, the correct answer was to select the data points highlighted in Fig. 3c. The resolution of domain tasks relies mainly on the users' knowledge, and in case of novice users, on their ability to reason. Therefore, we expect novice users, especially without the guidance support, spending more time on reasoning and having rather approximated results. However, also when no guidance was provided, a percentage of novice participants were able to solve the domain tasks without guidance.

Exploratory task. We created a second set of tasks focusing on operational knowledge. When asking the participants to solve such tasks, we avoided any reference to domain concepts and just asked the participant to look for data with specific characteristics, without focusing on the meaning. In particular, as already mentioned, we structured such task as a long sequence of filtering and selections, to reach and select the desired data. In one task, we asked the participants to *select, for each measuring station, the FIRST data point such that, the value of Salinity (x axis) is greater than 2, but lower than 3 salinity units.* As it can be seen, no domain knowledge is requested except reading and understanding the graph (in this case, a scatter plot) and interacting with the tool performing selections and filtering. The visualizations used for this task, according to the provided guidance degree are portrayed in Fig. 4. In this case, the participants were presented with a scatter plot representing values of salinity (x-axis) in relation to the change of water depth (y-axis). Since this task is focused on interaction, we provided the users with means to select and filter the data, for instance, filter according to the measuring station that captured the measurement. Other advanced interaction means were also provided, for instance the possibility to perform lasso selections and avoid the need of multiple clicks.



(a) Exploratory task - no guidance

(b) Exploratory task- directing guidance

(c) Exploratory task - prescribing guidance

Fig. 4. The same exploratory task supported with different guidance. (a) no guidance: a scatter plot shows values of salinity (x-axis) in relation to the change of water depth (y-axis). A widget allow users to filter the dataset in respect to the measuring station. (b) directing guidance: possibly interesting data points and filtering options are highlighted; (c) prescribing guidance: step-by-step instructions and highlighting of correct values as well as the filtering actions to be performed.

In a first scenario, some participants received no guidance. The participants dealing with such task were presented just the plain graph (see Fig. 4a). In this first case (no guidance), we expected the participants to filter the dataset by selecting and exploring the measurement of all the different measuring stations (also the ones with no interesting measurements) and select the data with the requested characteristics. In a second scenario, other participants were supported by directing guidance (see Fig. 4b). In such case, the participants could also rely on the highlighting of the measurements falling in the requested range $[2 - 3]$, therefore being directed in the data retrieval. In addition, we also highlighted with a different color the filtering option i.e., the measurement stations that captured those data values, leading to the

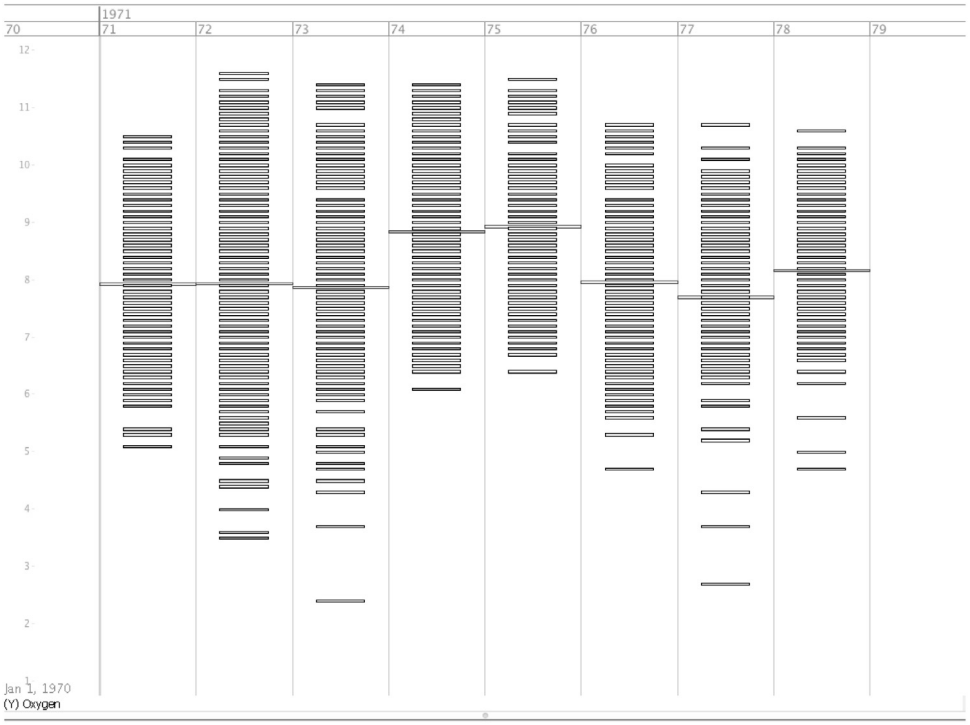


Fig. 5. A chart representing aggregated values of nutrients (y-axis) according to the measurement station.

requested data points, in such a way to signal to the participants a set of possible filters to choose. In this way, we wanted to guide the interaction, by either highlighting the data and the filtering options necessary to make the requested data visible. Participants receiving directing guidance could also skip unnecessary actions, checking the highlighted data. Finally, in a latter scenario, a part of the participants had to rely on prescribing guidance (see Fig. 4c). Similarly to domain tasks, this kind of guidance consisted of a list of instructions, in addition to the highlighting of the data. However, while usually for domain task this list was composed by one or two actions, for domain tasks it consisted of a long sequence of filtering and selection steps to simulate a thorough exploratory analysis. With prescribing guidance, the user had to follow dutifully an average of twenty consecutive steps of alternate filtering and selections, to complete this task and select the required data.

Obviously, the correct answer to this task was the same despite the different provided guidance types. In the context of this task, we expect faster interactions with increased guidance. However, we also analyzed the variation of frustration levels and confidence in users that had to strictly obey to the prescribed actions to complete the task, to see if they felt restricted by the guidance.

5.5. Visual encoding design

We chose basic visualization types for the study. We chose scatter plots, line charts, and temporally aggregated charts showing data values for each year side by side (Fig. 5). We wanted to keep the visualization aspects of the study as general as possible, so to not interfere with the outcome of the analysis and the effectiveness of the provided guidance. At the same time, we chose these visualizations also because the participants were familiar with them, but also effective for the given tasks.

Scatter plots represented data values as dots with one of three variables (either water salinity, chlorophyll, or suspended solids) on the *x*-axis and water depth on the *y*-axis. Line charts represented dots connected by lines with time on the *x*-axis and salinity on the *y*-axis, and temporally aggregated charts juxtaposed yearly (x-axis) oxygen values (y-axis). The visual encoding we used are portrayed in Figs. 3, 1, and 5.

Interaction means. For all chart types, we provided basic interactive means such as details on demand when hovering a data point, selection of single data points when clicking on them, rectangular selection (by dragging the mouse to span a rectangle) for selecting multiple data items, and deselection of all data points with a right mouse-click. The successful selection/deselection of a data point was visualized by a change of fill-color. For exploratory tasks, we also provided radio-buttons to filter the displayed data points according to the time-stamp of the measurement (month and day). We did not provide filtering for domain tasks, as it was not required. This difference in the interaction means did not influence the outcome of the study, as domain and exploratory tasks were never compared directly. As a final remark, all the participants were provided from the beginning of the study with the possibility to use all the interaction means. The participants belonging to Group B were introduced during the learning session to the use of all the available interaction means.

5.6. Guidance

In their work, Ceneda et al. describe three degrees of guidance: orienting, directing, and prescribing (Ceneda et al., 2017). However, since common practices in visualization such as axis labels could also be seen as a very low level of guidance, the border between no guidance and orienting (giving some hints for orientation) becomes blurred. Thus, to avoid confusion and have a clear baseline for comparison we implemented just three of them: (1) *no guidance*, (2) *directing guidance*, and (3) *prescribing guidance*. By design, the participants received all three guidance degrees, one time in each task set. In total, each participant received the same degree of guidance twice: once while executing exploratory tasks, and once while performing domain tasks.

When no guidance was provided, we presented the participants a common visualization (e.g., a line chart) showing one of the data subsets, with some additional data visualized, like average or minimum values (see for instance Figs. 3a and 4a). When directing guidance was provided, participants received an additional indication about possible interesting data or actions to consider. Figs. 3b and 4b show the encoding chosen for directing guidance. Interesting interaction options were highlighted for exploratory tasks (upper side of the interface), while interesting data-points were highlighted for domain specific tasks within the visualization. Finally, participants receiving prescribing guidance were provided with step-by-step instructions to reach the desired results as shown in Figs. 3c and 1. The instructions were given as red text, in the bottom-left of the visualization. Prescribing guidance produces mandatory actions (Ceneda et al., 2017). Hence, although participants could perform any other action and deviate from the analysis path, the instructions proceeded only after the user conducted the required steps. Moreover, we provided them with the possibility to restart the guidance process. We motivate the introduction of this extreme degree of guidance to explore the full range of guidance possibilities. It is worth clarifying that this high guidance degree does not correspond to a simple presentation of results, and it also differs from a pure automated data analysis (Ceneda et al., 2018). The user is always required to interact and confirm the different steps and moves. Moreover, as already pointed out, the participants always had the possibility to deviate from the suggested analysis path for further analysis.

6. Results

Sixty-five participants submitted their results and the interaction logs.

6.1. Analysis approach

We analyzed the logs and the results of the user-study using the R environment for statistical computing (R Core Team, 2014). Our aim was to spot significant differences between subgroups. In our study, we mostly compared three groups (i.e., the three guidance degrees) among each other, for each task type and expertise condition. Hence, we used the Kruskal–Wallis test (Kruskal and Wallis, 1952), which is a non parametric test similar to ANOVA which can be used with more than two groups and is also well-suited for comparing results obtained from Likert scales (De Winter and Dodou, 2010).
In a few tests, we compared the variation of single metrics (e.g., frustration) in users with different expertise. For instance, the tested variation of correctness in novice and knowledgeable users, for the same type of tasks (e.g., exploratory tasks). In such cases, since we had to compare just two groups (i.e., novice vs knowledgeable) we applied the Mann–Whitney–Wilcoxon test (Mann and Whitney, 1947). We never compared directly exploratory and domain tasks among each other.

Since we performed many tests and hence to account for the probability of a false positive discovery, we applied to all the tests the correction technique by Benjamini and Hochberg (Benjamini and Hochberg, 1995). This choice implies that, while usually a common threshold value is chosen for all the tests (usually set to $p \approx .01$), in our study it varies according to the test. In the specific case, for each test two p-values are calculated, p which is resulting from the test, and $p_{corrected}$ which is calculated from p. The corrected p-values are calculated considering the total number of tests performed and an initial significance level of 0.05. Hence, when we report on the acceptance of a test, we will also report the correspondent corrected p-value. When a significant difference was detected then we performed post-hoc tests to compare the different groups among each other and evaluate the pairwise differences. As a final step to the analysis, we manually inspected the data and further analyzed the results with box-plots and scatter plots.

6.2. Users' statistics

Performance. The system automatically extracted a set of measures to understand task performance (see Table 1). These measures include the total number of actions conducted by the user to complete a given task: number of clicks, rectangular selections, and applying filters. We also computed a correctness value to reflect the ratio of correctly selected data items to the total number of correct data items weighted by the total number of selected data items. We included this measure to account for cases in which participants select huge numbers of data items which makes it likely that they also select some correct ones. Another measure, the distance, was computed to quantify the semantic distance of the answer, in terms of selected data items, to the correct answers. We calculated this metric by averaging the temporal distance of the selected points from the solution.

$$distance(avg) \approx \frac{\sum(temporal\_dist(x, solution))}{total\_data\_selected}$$

Since all the tasks comprised temporal aspects, measurements falling in different time periods were considered distant. We did this to understand if wrongly selected data items are semantically close to the correct ones (e.g., they are in the same month) or if they are completely wrong (e.g., they are in different years).

Feelings. Besides performance measures, this study comprises a set of measurements dealing with user's feelings. Guidance approaches inherently deal with users. Hence, it is important to understand how guidance affects the development of user's psychological aspects. These are listed in Table 2. Similarly to user's knowledge, such psychological aspects are difficult to measure and quantify. However, for our purpose of deriving correlations and tendencies rather than quantitative values, we use a simple qualitative scale to measure the participant's own assessment of their feelings. Usually, this method may be influenced by the personality of the participants, who may present extreme/average input styles. However, such drawbacks were mitigated and averaged by the number of participants involved. Therefore, we did not apply any further correction to those tests.

6.3. Outcome

The tests indicate that guidance has an overall positive effect on users' performance and mental state. Guidance is particularly successful for novice users solving exploratory tasks and can easily compensate for a lack of operational knowledge. Instead, the tests highlight that for domain tasks, at least a minimum of knowledge should be possessed by the users, not only to understand the tasks and the context, but also to interpret correctly the guidance.

Our study highlights that guidance is important in complex scenarios: We show that the benefits are particularly pronounced when domain knowledge and reasoning are needed: for knowledgeable users solving domain tasks, the results obtained with directing guidance were in line with those obtained by prescribing guidance. However, our study reveals that guidance may even have a bad impact on the analysis if the guidance degree does not match the knowledge gap and users' expectations. From our results we can see that novice users, tend to trust excessively the guidance suggestions, and that the prescribing guidance degree may sometime frustrate knowledgeable users. The tests revealed

Table 3
Key findings. In this table we summarize the results obtained in our study. We provide references to the hypotheses where we discussed these results in more detail. *Additional finding* refers to results that were not taken directly from the hypotheses, but inferred from them. Please refer to Section 6.4 for the details.

Key findings
1. While it is no surprise that a high degree of guidance had positive effects on the performance of novice users, it is remarkable that guidance, especially the prescribing degree, had significant positive effects on performance and mental state also of knowledgeable users for almost all combinations of task types (H1).
2. Guidance was particularly effective to account for the lack of operational knowledge. For domain tasks, the users should possess at least a minimum of knowledge to interpret correctly the suggestions. This indicates that missing operational knowledge is easier to compensate by guidance than missing domain knowledge (H1).
3. Knowledgeable users were not frustrated by high degrees of guidance while there was a positive effect on confidence and the subjective assessment of the difficulty of the task (H2).
4. Participants' subjective assessment of appropriateness of guidance degree was reflected in better performance, and more positive mental state, which reflects the importance of providing an appropriate degree of guidance for the given user (H2).
5. Knowledge plays an important role for positive performance and mental state especially when solving domain tasks. However, prescribing guidance may compensate for the lack of knowledge in many aspects (additional finding).
6. Knowledge may also compensate for a lack of guidance. Knowledgeable users with no guidance obtained similar performances to novice users provided with directing guidance, for both exploratory and domain tasks (additional finding).
7. Domain tasks evoked more frustration than exploratory tasks in novice users, since trial and error can compensate for a lack of operational knowledge while not for a lack of domain knowledge (additional finding).

that directing guidance is beneficial for knowledgeable users who are able to interpret and judge correctly the suggestions. When assisted with this kind of guidance, participants obtained performances similar to prescribing guidance. On the other hand, this degree produces no improvements (same results as no guidance) when provided to novice users. Thus, for novice users the prescribing degree of guidance seems better suited.

In the following, we discuss the results in relation to our hypotheses. We then outline observations and interesting additional findings. A summary of the study outcome can be found in Table 3.

H1.1. We investigated if guidance positively affects the performances of *novice users*. The results reveal differences in the performances of novice users receiving guidance, and those who did not receive any guidance. The box-plots in Fig. 6 show that novice users perform significantly better with prescribing guidance (for both, exploration and domain tasks), in respect to the other guidance degree (directing) and to the scenario in which guidance was not provided at all. Hence, H1.1 can be accepted. However, directing guidance shows no significant improvement of performance of novice users compared to no guidance.

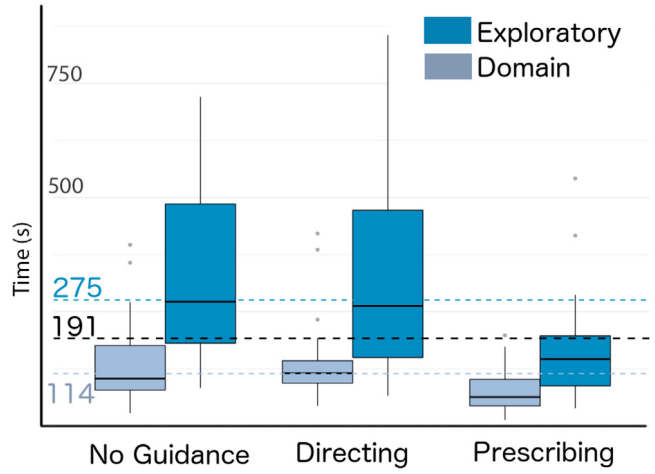
Task completion time. Novice users solved *exploratory* and *domain tasks* faster when supported by prescribing guidance. In fact, we found a significant difference between prescribing and no/directing guidance ($p \approx 0.01, chi - sq = 23.4, df = 2$ for both task types, see Fig. 6a). For both task types, no significant differences of timings were reported between directing and no guidance, while in general, completion times resulted higher for domain tasks, in respect to exploratory tasks. We noticed some cases, in which the participants who did not receive any guidance solved their tasks faster than those receiving directing guidance. We guess that this difference can be explained with the additional time required to interpret the guidance suggestions, especially for novice users. Hence, we imagine that while the participants who received no guidance started immediately to look for the correct answer, the users who received directing guidance (i.e., pointed to possible interesting subsets of the data) lost some time judging the applicability of the suggestions.

Correctness. Correctness values are also influenced by the guidance degree. Fig. 6b shows the box-plots for *exploratory* and *domain tasks*. Looking at the exploratory tasks, tests reveal significant differences between prescribing and no/directing guidance for novice users ($p \approx 0.005, chi - sq = 27.4, df = 2$). Similarly, for novice users solving domain tasks, we found very significant differences in correctness values between prescribing guidance and the other two degrees ($p \approx 0.007, chi - sq = 27.1, df = 2$). Moreover, correctness values showed that when we provided directing guidance to novice users, they did not answer more correctly to questions compared to no guidance. In average, the correctness increased with increased guidance, but the guidance itself could not replace the lack of knowledge, in novice users. This is particularly true for domain tasks, where novice users had similar results with both no guidance and directing guidance. For exploratory tasks, the charts show an increased correctness between no guidance and directing guidance, but the difference was not significant.

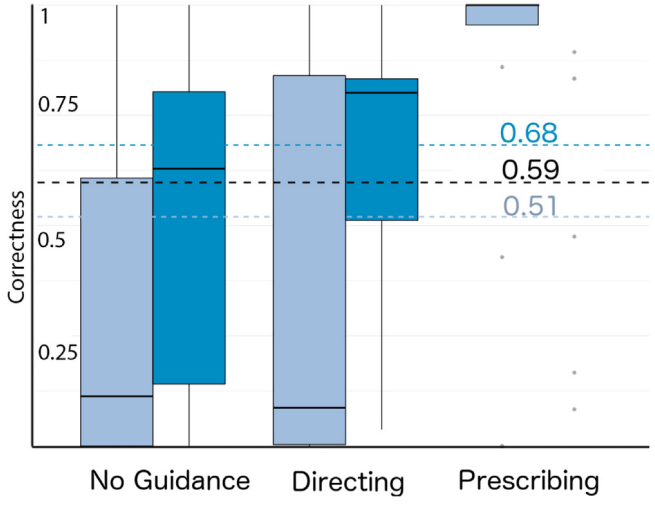
Half of the novice users (49.6%) completed correctly the exploratory tasks without any suggestion (no guidance), this number increases to 66% for those guided by directing guidance, and finally, the majority of participants receiving prescribing guidance ($> 90\%$) completed correctly these tasks. On the other hand, the results highlight that guidance cannot completely overcome the lack of knowledge, in case of domain tasks. Just 32% of the novice users completed the tasks correctly, this percentage raises to 37% for directing guidance, and 85% for prescribing guidance.

Distance. For novice users, and similarly to the other measures, we noticed a significant difference in the distance measures between participants assisted with prescribing guidance, and participants assisted with no guidance. This holds true for both *exploratory* ($p \approx 0.0002, chi - sq = 16.9, df = 2$) and domain tasks ($p \approx 0.0003, chi - sq = 15.7, df = 2$). For interaction tasks, the tests did not highlight any difference in the distance measure between directing and prescribing guidance. Moreover, novice users receiving directing guidance had results closer to the correct values (smaller semantic distance), compared to those who received no guidance (see Fig. 7). For domain tasks, the lack of domain knowledge may have had nullified the effectiveness of directing guidance, as the tests did not highlight any significant improvement in respect to no guidance.

Total steps. Novice users performed an average of 42 actions to complete a task: 13 filters, 6 multiple selections, and 23 single selection clicks. For *exploratory tasks*, the users receiving directing guidance performed similarly to those who did not receive any guidance (approx. 83 steps each). However, users provided with prescribing guidance needed on average only half the amount of steps (45 steps) which presents a significant difference. For *domain tasks*, the influence of different degrees of guidance is even more significant. On average, participants provided with no guidance completed a task with 24 actions. Directing guidance

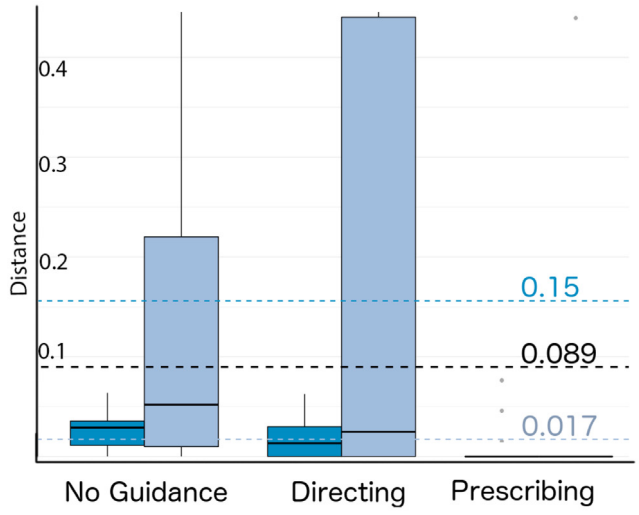


(a) Completion time (seconds) of novice users.

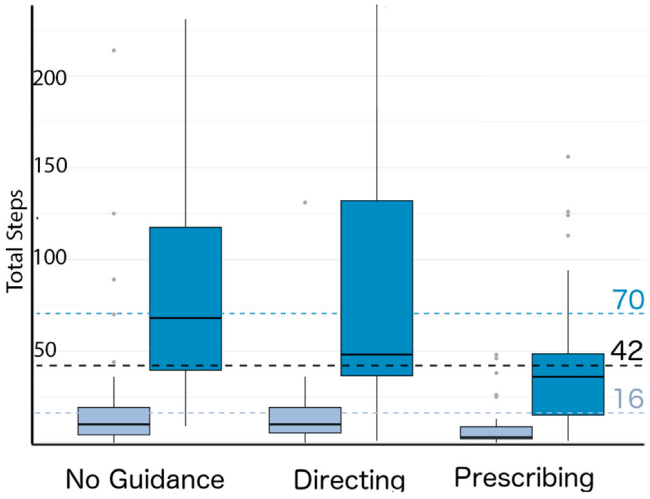


(b) Correctness metric of novice users.

Fig. 6. Box-plots for H1.1: We report time and correctness performance metrics for novice users (blue tones), according to different guidance degrees (x-axis). Dashed lines encode the total average (exploratory and domain tasks combined; in black) and individual average values for exploratory and domain tasks (in light and darker blue).



(a) Distance metric of novice users.



(b) Number of steps/operations of novice users.

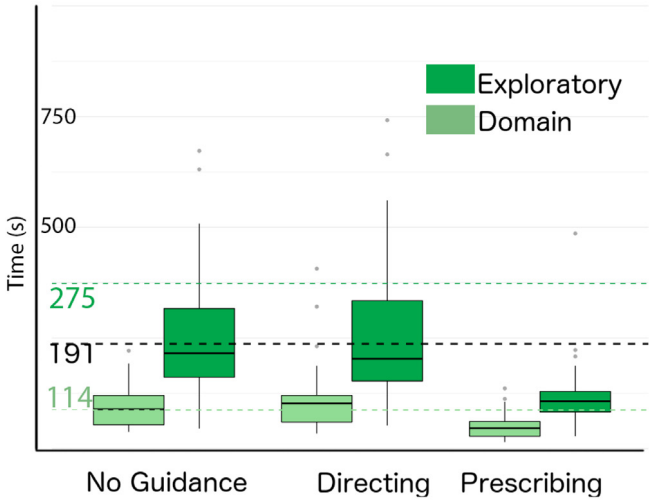
Fig. 7. Box-plots for H1.1: We report distance metric and the number of steps for novice users (blue tones), according to different guidance degrees (x-axis). Dashed lines encode the total average (exploratory and domain tasks combined; in black) and individual average values for exploratory and domain tasks (in light and darker blue).

lowered this number to 15 actions, while novice users supported with prescribing guidance, took on average 8 actions. The statistical tests reported a significant difference in the number of steps required by novice users performing *domain tasks*, between prescribing guidance and no guidance; there was no significant difference, on the other hand, between prescribing and directing guidance (see Fig. 7).

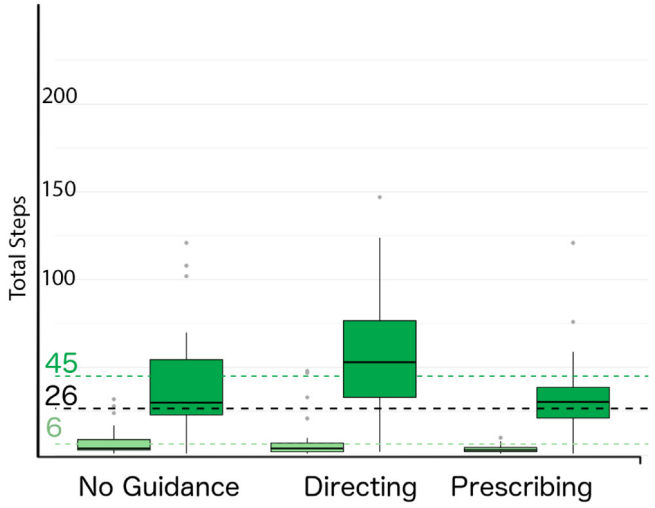
H1.2. We hypothesize that a high degree of guidance may reduce completion time and the number of steps needed for *knowledgeable users*. H1.2 can be accepted partially. The tests did not show significant differences in the number of steps. However, we noticed a significant difference in completion times of knowledgeable users, in particular only between prescribing guidance and the other degrees ($p \approx 0.02, chi - sq = 29.8, df = 2$, see Fig. 8) when solving *domain tasks*. Same results were obtained for *exploratory tasks*: guidance affected completion times but not the number of total steps. Completion time was significantly better with prescribing guidance in respect to no guidance and directing guidance ($p \approx 0.002$ and $p \approx 0.02$ respectively, $chi - sq = 15.7, df = 2$). These results are in line with our assumption that knowledgeable users may still benefit from guidance. The tests reveal reduced completion times for these users: the guidance allows them to focus on the supervision of the analysis, alleviating the burden of focusing on minor details.

Besides completion times, high guidance also had significantly positive effects on correctness, distance values, and mental state of knowledgeable users. For knowledgeable users solving domain tasks the tests highlighted a significant difference between prescribing and no guidance. However, no difference was detected between prescribing and directing guidance. This may indicate that some knowledge may allow users to correctly interpret directing guidance. Hence, this degree should be considered when designing guidance for knowledgeable users, as it still leaves the users a certain degree of freedom, which has a positive impact on the mental state of the users (the participants commented that they do not feel restricted), and may lead them to discover the unexpected.

H2.1. We hypothesize that guidance may influence positively the confidence of participants and the tests showed that confidence levels were significantly higher with higher guidance. *Novice users*

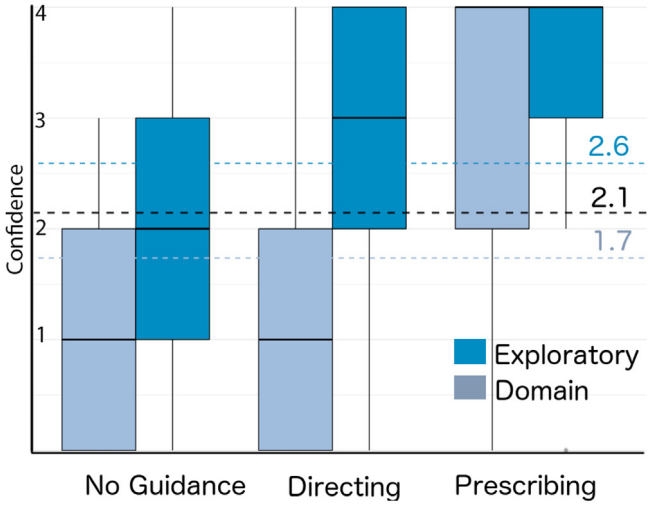


(a) Completion time (seconds) of knowledgeable users.

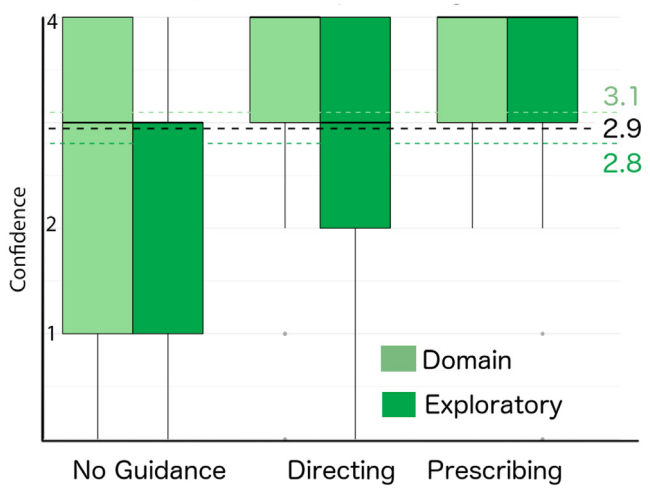


(b) Number of steps/operations of knowledgeable userss.

Fig. 8. Box-plots for H1.2: completion time and total steps of knowledgeable users (green tones). Dashed lines encode the total average (exploratory and domain tasks combined; in black) and individual average values for exploratory and domain tasks (in light and darker green).



(a) Confidence of novice users.



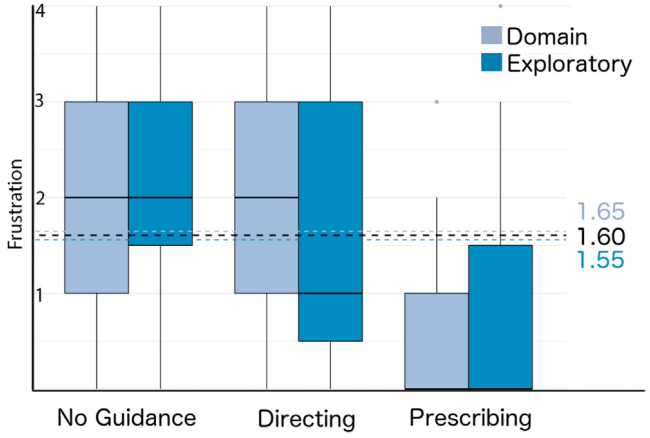
(b) Confidence of knowledgeable users.

Fig. 9. Box-plots for H2.1. Guidance influences positively the user's confidence. Confidence was measured on a five-point Likert scale, where 0 encodes no confidence. Novice users are represented with blue tones, and knowledgeable users with green tones. Dashed lines encode the total average (exploratory and domain tasks combined, in black) and individual average values for exploratory and domain tasks.

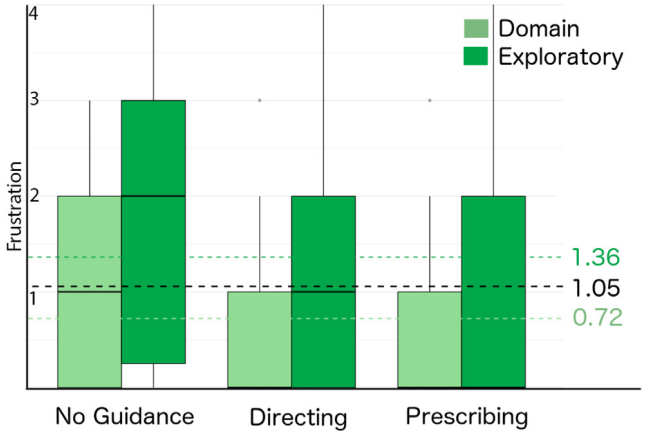
rated their confidence in their results significantly higher when receiving prescribing guidance, if compared to no guidance ($p \approx 0.002$, chi-sq = 12.1, df = 2, see Fig. 9), for exploratory tasks. The same comparison is significantly different ($p \approx 0.006$, $chi - sq = 10.15$, $df = 2$) also for knowledgeable users solving the same task type. In this test, although the charts report increased confidence associated with the provision of directing guidance, the tests did not report any significant difference if we compare the confidence levels obtained with no guidance. For novice users solving domain tasks, the tests revealed a significant difference ($p \approx 0.000025$, $chi - sq = 21.1$, $df = 2$) between prescribing and the other two guidance degrees. A still significant, but lower result ($p \approx 0.02$, $chi - sq = 6.8$, $df = 2$) is reported also for knowledgeable users solving domain tasks with prescribing and with no guidance. Conversely to the results obtained with exploratory tasks, where we noticed an increased confidence with increased guidance, for domain tasks the confidence values obtained with directing guidance are absolutely comparable to those obtained with no guidance. It is clear the influence of a proper knowledge on those users. Comparing general confidence levels of novice with those of knowledgeable users, for exploratory tasks, the tests did not show any significant difference. For domain tasks, the results show however a significant difference. Furthermore, for these tasks, the confidence related to no guidance is comparable to confidence levels with medium guidance, for novice and knowledgeable users. Finally, when comparing *domain tasks* to *exploratory tasks*, we noticed that the average confidence resulted much lower for domain tasks than for exploratory tasks.

H2.2. We hypothesized that different guidance degrees influence how much novice and knowledgeable users feel frustrated when solving their tasks. In particular, we thought that novice users would be significantly less frustrated by a high degree of guidance than knowledgeable users.

Novice users. Like stated by Celik et al. (2013) the lack of knowledge is proportional to the users' frustration. Guidance, in this sense, represents a compensation for the lack of knowledge. Our results indicate that they feel less frustrated when receiving a higher degree of guidance, both for *exploratory and domain tasks*. Frustration decreases with increasing guidance, but in our tests it is prescribing guidance that marks a significant difference with



(a) Frustration of novice users.



(b) Frustration of knowledgeable users.

Fig. 10. Box-plots for H2.2. Frustration of novice users (blue tones) and knowledgeable users (green tones), according to different guidance degrees. The level of frustration is ascending: values closer to 0 indicate less frustration. We also report, with dashed lines, the total average frustration (for exploratory and domain tasks combined, in black), as well as the average value for individual domain and exploratory tasks.

the others degrees. In fact, prescribing guidance significantly reduces frustration in novice users compared to no or directing guidance. Fig. 10 shows a box-plot representing the level of frustration with respect to the provided guidance. In the figure, the results represent both *exploratory* and *domain tasks*. The total average frustration for exploratory tasks (avg: 1.65) and domain tasks (avg: 1.55) is comparable. If we consider individually the single task types, we notice that *domain tasks* evoke frustration in novice users: the tests indicate a significant difference between prescribing guidance and no guidance ($p \approx 0.001, chi - sq = 25.6, df = 2$), while no significant difference is reported with directing and between directing and no guidance. For *exploratory tasks*, we also noticed a significant difference between prescribing and no guidance ($p \approx 0.02, chi - sq = 13, df = 2$). Although directing guidance reduced by the half the perception of frustration (avg:1.0) in respect to no guidance (avg: 2), the tests reported no significant differences between prescribing guidance and directing guidance as well as between directing guidance and no guidance.

Knowledgeable users. We hypothesized that high degrees of guidance may cause frustration in *knowledgeable users* who already know how to conduct the analysis. Although we showed that the frustration of novice users decreases while the guidance degree is increased, the opposite is not true for knowledgeable users. The tests did not show an increased frustration in correlation with increased guidance (for both, exploratory and domain tasks). The test did not show any significant difference in frustration levels for different guidance degrees. Hence, H2.2 cannot be accepted.

For *domain specific tasks*, the frustration of knowledgeable users decreased to almost zero, both for prescribing and directing guidance. The increased knowledge enabled participants to correctly interpret the suggestions, producing low frustration levels also for directing guidance. For *exploratory tasks*, some participants reported increased frustration when receiving prescribing guidance, since they already knew how to interact with the visualization. However, this was mentioned by a small number of participants, and did not affect the overall results. Knowledgeable users show lower levels of frustration as novice users, also when provided with prescribing guidance. Some of them reported that they felt frustrated by the restrictions that come with this high degree of guidance. Moreover, knowledgeable users rated the tasks easier to solve when receiving prescribing guidance.

In summary, frustration is related to the inability of users to complete the tasks. The tests suggest that prescribing guidance reduces significantly the frustration of novice users. Moreover, our results show that domain tasks were more stressful than exploratory tasks, for novice users. Besides frustration levels, prescribing guidance also had significantly positive effects on all variables related to their mental state.

6.4. Observations and interpretation of the results

In this section, we present and discuss findings and observations that are not directly connected to the hypotheses we formulated beforehand, but were apparent from our results and that are worth mentioning.

Effects of knowledge. We observed that providing knowledge to participants had a significant influence on their performance and mental state when solving *domain tasks*, but not so pronounced in participants solving *exploratory tasks*. This may be explained by the fact that participants, without knowing what they were actually doing, but only knowing how to do it (exploratory tasks with operational knowledge), still did not feel like having control. However, receiving high guidance mitigated these strong differences between novice and knowledgeable users. This means that domain knowledge has a significant impact on performance and mental state, but a high degree of guidance could also have the potential to compensate for a lack of domain knowledge.

On the other hand, domain knowledge may also compensate for missing guidance. After the learning session, the participants did not feel lost when receiving no guidance. Furthermore, they felt that even directing guidance made solving the tasks as easy as when receiving prescribing guidance i.e., the tests did not show any significant differences between the two groups.

Difference between domain and exploratory tasks. A positive effect of guidance may also be found when comparing mental states of participants dealing with domain or exploratory tasks, respectively. Regardless of the degree of guidance provided, knowledgeable users felt significantly more lost, more frustrated, less confident, and thought that the tasks were harder to solve when solving exploratory tasks than when solving domain tasks. Novice users, on the other hand, felt significantly more lost and less confident when dealing with domain tasks. Furthermore, novice users found domain tasks harder to solve than exploratory tasks. We reason that this is due to the fact that having no knowledge about how to interact with the visualization may be compensated with trial and error, but having operational knowledge did not enable participants to control what they are doing semantically. Missing essential domain knowledge to solve a given task cannot be easily compensated by trial and error.

Confidence, correctness, and frustration. Another finding from this study is that participants' confidence in their answers was justified. We observed that their confidence levels correlated with correctness levels. Furthermore, a negative correlation can be found between correctness and frustration level. While these findings are not surprising, they foster our trust in the reliability of our results.

Misleading hints. In a handful of cases, providing directing guidance resulted in even worse performance (times and correctness) than providing no guidance at all. However, the tests did not show significant differences here. It may be that in some cases novice users trusted the hints provided by directing guidance too much. In fact, some participants selected all data points within the highlighted regions of interest, without reasoning about their effective meaning. This means that a vague kind of guidance – e.g., providing recommendations to the user, etc. – should be used with caution because it may mislead novice users. For expert users, in fact, this behavior could not be observed.

Appropriateness of guidance. Another interesting finding is that when participants felt they received an appropriate degree of guidance they also completed the tasks with a positive outcome in all other variables: They had better performance in terms of time and correctness, they felt less lost, less frustrated, and more confident about their answer. Finally, they had also the impression that the task was easier to solve. While this outcome was to be expected, it stresses the importance of providing an appropriate degree of guidance with respect to the expertise of the user. While novice users considered prescribing guidance to be much more appropriate than the other two degrees, this effect was mitigated for knowledgeable users. For real expert users this may be even more true and too high degrees of guidance could lead to frustration.

7. Discussion and future work

Although we considered carefully each and every design aspect of our study, there are also limitations to this work.

Knowledge. One of our main concerns, while designing the study, was how to ensure different knowledge levels of participants. Usually, knowledge is hard to judge, evaluate, or measure precisely, as there are many factors influencing the way it is acquired. However, we did our best to ensure that our novice users had no additional information to solve the tasks – they had, in fact, no experience with the data. On the other hand, we taught our knowledgeable users what they needed to know and some learning effect from the first session also added up to that knowledge. To further consolidate the acquired knowledge, the participants had to exercise, and revisit the concepts before proceeding with the remaining tests. However, as already mentioned, we did not measure precisely the increase of knowledge, but the results of the study show clearly that knowledgeable users had better performance than the novice in the *no guidance* condition. Around 10% more participants were able to solve exploratory tasks after the learning session, and an average of 20% more participants could solve the domain tasks after learning the required domain knowledge.

Since our study participants were familiar with standard interaction techniques, we had to design exploratory tasks with less obvious interaction techniques. This was backed up by the interaction logs which showed that usually just simple clicks were used by novice users. Just a few of them used other interaction means, and many reported that they learned about all the different interaction options just during the learning phase. We found significant differences in novice users and knowledgeable users in terms of mental state and performance, which furthermore confirms the distinction of their knowledge levels.

Knowledgeable users. Another limitation is presented by the fact that our knowledgeable users cannot be considered real experts yet. A real expert would be someone who was working in the given domain and with the provided interactive visualizations for a long time. Our study design did not include this type of user. The used visualizations were specifically designed to include different degrees of guidance into different basic types of visualization to be able to test our hypotheses, and there is just no real life scenario with real expert users that would be suited to test these hypotheses. Thus, our results reflect only the behavior of novice users contrasted with the behavior of knowledgeable users, who both benefit the guidance received. However, we see a tendency of knowledgeable users to feel frustrated by prescribing guidance when they felt that the tasks were very easy. We can only hypothesize that real expert users may have found prescribing guidance disturbing or restricting, but this is left for further investigation.

Directing guidance. Another interesting point regards the representation of the hints provided by directing guidance. We did our best to visually distinguish the hints given by directing guidance and the actual instructions given by prescribing guidance in order to not mislead users. We chose simple highlighting to indicate interaction options and interesting data regions (directing guidance), while for prescribing guidance we chose precise textual instructions in combination with highlighting specific data items. We consistently used this encoding in all tasks and visualization types, and furthermore, informed participants about these differences. We wanted to provide guidance in a way as general as possible, and we found that this simple representation was quite useful and effective for our scope. However, other encodings would also be feasible and may lead to different results. Thus, it would be interesting to investigate further encodings of guidance and their effects in future work.

A curious outcome of our study was that often our tests did not report *significant* differences between directing guidance and no guidance. From the box-plots (see for instance Figs. 6–9) it is clear that guidance introduced some differences, but the test did not highlight it as significant. We tried to explain this situation reasoning about the fact that in some situations, like at the beginning of the test, the novice users had not sufficient means to understand the guidance hints. However, in some other cases, we could not fully explain this result. In particular, it was unexpected encountering this lack of significant differences in knowledgeable users. We could think that in those cases the acquired knowledge was then sufficient to fill the differences between the two guidance degrees. The cause may also be related to the possibly misleading hints gave by this guidance degree. However, neither the logs nor the participants' comments gave us a better understanding of the real cause. For this reason, we reserve the possibility of further investigations in this direction.

Visual encoding. In line with ensuring basic visual encodings of guidance, we also chose a small number of basic visualizations to represent the dataset. The chosen visualizations represent pretty standard choices in visual data analysis, and are well suited to solve the tasks we proposed to the participants. Another motivation for choosing them was that the participants were already familiar with them. The participants, in fact, reported that in the vast majority of the cases the visualizations were well understood. However, a consequence of our choice, is that our findings may not be generalized to more sophisticated methods. This again would be an interesting topic for future investigations.

Tasks. Finally, we constructed our study on a limited number of different tasks. In particular, we focused these tasks on some specific domain concepts, and simple operational procedures. Although we designed them with special respect to keeping them

simple (i.e., basic look-up tasks, simple interactions, so to simulate a general exploratory analysis) we cannot guarantee the results we obtained may be generalized to other types of tasks. Hence, our results should be seen as initial insights, how guidance works for these and similar tasks, how different guidance degrees work for different users, possible effects on a user's mental state and critical aspects that need to be considered. However, we think that these results could be extended and consolidated for other tasks and domains.

8. Conclusion

We presented a user study about guidance, which constitutes a first step towards a scientific understanding of the effects of guidance in different analysis scenarios. In this context, we consider a number of different aspects that interact with guidance. We relate the effects of different degrees of guidance to a user's expertise level, we consider different types of tasks, and we measure task performance as well as the user's mental state. Our study suggests that guidance has positive effects on both knowledgeable and novice users. On the other hand, the study reveals that guidance must be designed carefully to meet the user's needs and that novice users may also be misled by medium guidance. We conclude that our work describes the value and effectiveness of having guidance while conducting a visual data analysis.

Declaration of competing interest

The authors, Davide Ceneda, Theresia Gschwandtner and Silvia Miksch, raise a conflict of interests with the following persons:

Christian Bors, Markus Bogl, Roger Leite, Victor Schetinger, Alessio Arleo, Maurizio Patrignani, Maurizio Pizzonia, Giuseppe Di Battista, Daniel Archambault, Andreas Walch, Michael Schwärzler, Elmar Eisemann, Christian Luksch, Simone Kriglstein, Margit Pohl, Jürgen Bernard, Christian Eichner, Heidrun Schumann, Jörn Kohlhammer, Erich Gstrein, Johannes Kuntner, Thorsten May, Marc Streit, Christian Tominski, Alice Thud, Jagoda Walny, Jason Dykes, John Stasko, Peter Filzmoser, Tim Lammarsch, Alexander Rind, Paolo Federico, Wolfgang Aigner, Jakob Doppler, Markus Wagner, Hans-Jörg Schulz, Oliver Erhart, Fabian Schwarzinger, Andreas Roschal, Andreas Peterschofsky, Alexander Endert, Giuseppe Santucci, Natalia Andrienko, Gennady Andrienko, Georg Fuchs, Daniel Keim, Florian Windhager, Günther Schreder, Katrin Glinka, Marian Dörk, Eva Mayr, Albert Amor-Amorós, Sebastian Zambanini, Simon Brenner, Robert Sablatnig, Florian Heimerl, Steffen Koch

Acknowledgment

This work was supported and funded by the Austrian Science Fund (FWF), grant P31419-N31.

References

Aigner, W., Hoffmann, S., Rind, A., 2013. Evalbench: a software library for visualization evaluation. In: Computer Graphics Forum, Vol. 32. Wiley Online Library, pp. 41–50.

Bederson, B., Shneiderman, B., 2003. The Craft of Information Visualization: Readings and Reflections. Morgan Kaufmann.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Bernstein, A., Provost, F., Hill, S., 2005. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. IEEE Trans. Knowl. Data Eng. 17 (4), 503–518.

Bertini, E., Lalanne, D., 2009. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In: Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. ACM, pp. 12–20.

Brusilovsky, P., Millán, E., 2007. User models for adaptive hypermedia and adaptive educational systems. In: The Adaptive Web. Springer, pp. 3–53.

Celik, P., Lammers, J., van Beest, I., Bekker, M.H., Vonk, R., 2013. Not all rejections are alike; competence and warmth as a fundamental distinction in social rejection. J. Exp. Soc. Psychol. 49 (4), 635–642.

Ceneda, D., Gschwandtner, T., May, T., Miksch, S., Schulz, H., Streit, M., Tominski, C., 2017. Characterizing guidance in visual analytics. IEEE Trans. Vis. Comput. Graphics 23 (1), 111–120. http://dx.doi.org/10.1109/TVCG.2016.2598468.

Ceneda, D., Gschwandtner, T., May, T., Miksch, S., Streit, M., Tominski, C., 2018. Guidance or no guidance? a decision tree Can help. In: EuroVA: International Workshop on Visual Analytics. Eurographics Digital Library, pp. 19–23. http://dx.doi.org/10.2312/eurova.20181107.

Ceneda, D., Gschwandtner, T., Miksch, S., 2019. A review of guidance approaches in visual data analysis: A multifocal perspective. Comput. Graph. Forum 38 (3), 861–879. http://dx.doi.org/10.1111/cgf.13730.

Chen, C., 2005. Top 10 unsolved information visualization problems. IEEE Comput. Graph. Appl. 25 (4), 12–16.

Cloern, J., Schraga, T., 2016. USGS Measurements of water quality in san francisco bay (CA), 1969-2015. Sci. Data http://dx.doi.org/10.5066/F7TQ5ZPR.

Cook, K.A., Thomas, J.J., 2005. Illuminating the Path: The Research and Development Agenda for Visual Analytics. Tech. rep.. Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

De Winter, J.C., Dodou, D., 2010. Five-point likert items: t test versus mann-whitney-wilcoxon. Pract. Assess. Res. Eval. 15 (11), 1–12.

Dix, A., Finlay, J., Abowd, G., Beale, R., 2004. Human-Computer Interaction, third ed. Pearson Education.

Gibson, J., 1977. The Theory of Affordances: Perceiving, Acting, and Knowing. Hillsdale, NJ Erlbaum., pp. 67–82.

Gotz, D., Wen, Z., 2009. Behavior-driven visualization recommendation. In: Proceedings of the 14th International Conference on Intelligent User Interfaces. ACM, pp. 315–324.

Heer, J., Card, S.K., Landay, J.A., 2005. Prefuse: a toolkit for interactive information visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 421–430.

Horvitz, E., 1999. Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 159–166.

Kapoor, A., Burleson, W., Picard, R.W., 2007. Automatic prediction of frustration. Int. J. Hum.-Comput. Stud. 65 (8), 724–736.

Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H., 2008. Visual analytics: Scope and challenges. In: Visual Data Mining. Springer, pp. 76–90.

Kruskal, W.H., Wallis, W., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47 (260), 583–621.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 50–60.

Mazurowski, M.A., Baker, J.A., Barnhart, H.X., Tourassi, G.D., 2010. Individualized computer-aided education in mammography based on user modeling: Concept and preliminary experiments. Med. Phys. 37 (3), 1152–1160.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

Rind, A., Lammarsch, T., Aigner, W., Alsallakh, B., Miksch, S., 2013. Timebench: A data model and software library for visual analytics of time-oriented data. IEEE Trans. Vis. Comput. Graphics 19 (12), 2247–2256.

Sacha, D., Senaratne, H., Kwon, B.C., Ellis, G., Keim, D.A., 2016. The role of uncertainty, awareness, and trust in visual analytics. IEEE Trans. Vis. Comput. Graph. 22 (1), 240–249.

Smith, S., Mosier, J., 1986. Guidelines for Designing User Interface Software, Tech. Rep. ESD-TR-86-278. Mitre Corporation, Bedford MA.

Streit, M., Schulz, H.-J., Lex, A., Schmalstieg, D., Schumann, H., 2012. Model-driven design for the visual analysis of heterogeneous data. IEEE Trans. Vis. Comput. Graphics 18 (6), 998–1010.

Willett, W., Heer, J., Agrawala, M., 2007. Scented widgets: Improving navigation cues with embedded visualizations. IEEE Trans. Vis. Comput. Graphics 13 (6), 1129–1136.

Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., Heer, J., 2016. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. IEEE Trans. Vis. Comput. Graph. 22 (1), 649–658.