

A Wizard of Oz Study of Guidance Strategies and Dynamics

Fabian Sperrle, Mennatallah El-Assady, Alessio Arleo, and Davide Ceneda

Abstract—Co-adaptive guidance in visual analytics is a mixed-initiative process in which both the user and the system work together to support each other in solving a given analysis task. While previous studies show the effectiveness of guidance, the impact of guidance design decisions (e.g., the suggestions' timing, contextualization, or adaptation) and misguidance often remain under-investigated. To investigate these aspects and examine a variety of guidance interaction patterns in a realistic analysis scenario, we present a Wizard of Oz (WOz) study setup in which pairs of participants take the user's and the system's roles, respectively. As users perform their analysis tasks, they are observed by wizards who provide just-in-time guidance as they see fit. Moreover, we designed the study so that wizards would occasionally and unknowingly provide misguidance during the analysis to investigate the users' confidence in guidance systems. We recruited two groups of participants (12 wizards and 12 users) and paired each participant with two from the other group, obtaining 48 observations. We report insights on interactions between users and wizards. By analyzing these interaction dynamics and the guidance strategies the wizards apply, we derive recommendations for implementing and evaluating future co-adaptive guidance systems.

Index Terms—Visual Analytics, Guidance, Wizard of Oz

I. INTRODUCTION

Active guidance in visual analytics (VA) that goes beyond the visualization of potentially helpful information but reacts to the user's actions has become increasingly popular in VA over the past few years. Ceneda et al. [11] have defined guidance as a mixed-initiative process in which the system aims to close *knowledge gaps* that users encounter during their analysis. Sperrle et al. [35], [36] built on that definition to characterize guidance in terms of *co-adaptive processes* and interaction dynamics in which both user and system learn from each other while teaching each other how to provide better guidance. However, these recent developments of increasingly adaptive and proactive guidance systems come with novel challenges for guidance in VA. First, co-adaptive guidance is difficult to design and develop for bespoke VA systems, especially if the number of potential users that the system can learn from is limited. Second, complex guidance systems potentially risk misguiding users in their analysis, leading to conscious or unconscious biases in the analysis results [14], [15].

Within this motivation, our research question entails investigating the interaction patterns of users performing an analysis task on a VA system when confronted with guidance, if they are willing to accept it, and how the confidence in automated

guidance changes in the progression of the analysis, especially when it fails and provides misleading directions. To address this challenge, we present the design and results of a double-blind WOz study in which we paired participants taking the role of the guidance system with participants performing visual analysis. Both wizards and users interacted with a computer system that subjects believed to be intelligent and autonomous. We deliberately chose a double-blind setup, in which we also concealed from wizards that they were interacting with human users. This setup helped reduce surfacing emotions that usually appear when humans interact with each other and thus made strategies and interactions more deliberate.

We conducted the study with 24 participants. To investigate the impact of different interaction dynamics and variable guidance quality, each participant completed two distinct task sets for a total of 48 observations. We collect the results and observe typical interaction dynamics of both guiding and guided participants, derive common guidance strategies, and synthesize lessons learned to support the design of future guidance systems. Our contributions are: (i) We design a double-blind WOz study to investigate guidance strategies in VA and the system/user interaction dynamics; (ii) we discuss the results of our study identifying both wizards' and users' strategies and (iii) derive actionable insights for the design of future guidance systems.

II. RELATED WORK AND BACKGROUND

A. Guidance in Visual Analytics

The definition of guidance as a mixed-initiative process [11] has its roots in a long history of mixed-initiative approaches in human-computer interaction [18], [25], [32]. More recently, guidance in VA has been defined as an active, mixed-initiative process [11]. Sperrle et al. characterize guidance in terms of learning and teaching dynamics [35] and propose learning goals for co-adaptive systems [36]. For an overview of systems employing guidance in VA, we refer the reader to state-of-the-art reports on guidance [12] and provenance analysis [43]. Sperrle et al. [34] surveyed human-centered evaluations in machine learning and found that few papers systematically evaluate the provided guidance. Ceneda et al. [13] studied the effectiveness of guidance under different guidance degrees, task types, and expertise levels. They find that guidance positively impacts both novice and advanced users and that ill-designed guidance can mislead users. Their study design relies on predefined guidance in a limited user interface. In contrast, we aim to replicate a VA interface's complexity more closely and provide dynamic guidance through wizards.

Ceneda et al. [10] presented a technique to automatically analyze facial expressions to determine situations where

Fabian Sperrle is with University of Konstanz.

Mennatallah El-Assady is with ETH Zurich.

Alessio Arleo is with Eindhoven University of Technology.

Davide Ceneda is with TU Wien.

guidance should be provided. Boukhelifa et al. [6] analyzed provenance logs of VA sessions to identify potential trust measures for co-adaptive analysis scenarios. They also report on how their findings influenced the final system implementation. Kandel et al. evaluated Wrangler, a system for guided data manipulation [19]. In a comparative study, they compare the effectiveness of their tool to Excel and find that users are significantly faster with Wrangler. Monadjemi et al. compare the effectiveness of guidance in the context of data discovery with a non-guided control group. They found that guided users were both more effective (more data points found) and efficient (fewer irrelevant data points investigated).

Lee et al. presented a study centered around a visualization recommendation system called Frontier [23]. They evaluate the effect of recommendation categorization by comparing it to a non-categorized baseline. Their primary focus is on how categories influence the analysis workflow and not on a performance comparison between categorized and non-categorized interfaces. They suggest that “recommendation categories reduced overhead associated with interpreting the visualization recommendations” [23]. Our work has a similar goal of identifying strategies and dynamics. However, we rely on a Wizard of Oz design to avoid the limited flexibility of rule-based recommendation engines.

B. Wizard of Oz Study Designs

WOz studies are frequently used in human-computer interaction to evaluate early system prototypes. These studies are designed so that subjects interact with systems with a high level of automation. However, the behavior of the system logic is, in fact, “emulated” by a human (the *Wizard*). Subjects are therefore deceived, convinced to be interacting with a machine, but in fact, providing instructions to a human being. The advantage is being able to test the interaction with the system interfaces without having the complete architecture in place. Most recently, Porcheron et al. [26] provide a general overview of the technique. Fraser and Gilbert [16] defined several requirements for WOz studies. They note that wizards must convincingly simulate the expected system behavior. In a system for text generation, Yang et al. [44] found it difficult to balance the correctness of wizard responses to accurately reflect the expected performance and error profiles of real systems. To achieve a variety of suggestions, they rely on multiple parallel wizards for a single user. In our study, we designed the wizards’ interaction possibilities to leave them strategic freedom but ensure that users could not easily identify their responses as human output.

Several works (e.g., [22], [29]) highlight the importance of constraining wizards to templated responses to consistently simulate a single system. We left wizards freedom in their guidance strategies, allowing us to observe a range of strategies that could be employed in future guidance systems. Steinfeld et al. [37] mapped out the design space of WOz studies. In particular, they distinguish the extent to which either human interaction or the system are simulated and map the resulting combinations on a spectrum from Wizard of Oz (real users, fully simulated systems) to Oz of Wizard (real system, fully simulated user).

There is a vast array of work using WOz evaluation methodologies in information visualization and VA, both with transparent wizards (e.g., [4], [17]) and hidden wizards (e.g., [24], [41], [45]). Most related to our work is a study by Tory and Setlur [38] in which they aimed to identify user intents in visualization creation before implementing an automated system. In contrast to these previous works, we rely on a *double-blind* setup, in which neither the users nor the wizards know that they are interacting with a human. Other research aims to learn from the actions that wizards perform in WOz studies, primarily from the domain of human-robot interaction. Knox et al. [20] shared initial ideas and in-progress results for “learning-from-wizards,” a variation of learning-from-demonstration. They report ideas of how to use this approach to enable robots to learn social behaviors. Sequira et al. [30] propose a methodology for capturing robot-interaction strategies from the observation of wizards and turning them into implemented robot behaviors. Finally, Sheline et al. [31] introduced the concept of *dual-sided, restricted-perception WOz* experiments. In this concept, participants take the roles of both users and wizards. Users then aim to teach the wizards how to solve a given task before the wizards’ task performance is tested in a subsequent stage.

This paper combines ideas and practices from previous work. Due to the uncommon nature of our design setup, we also make decisions seemingly contrasting with these established practices. However, we justify our decisions in section III.

III. TERMINOLOGY AND STUDY DESIGN RATIONALE

The overarching goal of this study is to shed light on how guidance design decisions manifest in interaction dynamics between the system and the user. This goal is two-fold: first, while co-adaptive guidance defines the dynamics of learning and teaching between humans and systems [36], these abstract dynamics leave significant freedom for various concrete guidance strategies, depending on the system’s guidance goals. In other words, from the guidance system perspective, an overview of generalizable guidance strategies does not exist. Second, we aim to provide qualitative evidence on how the users react to these strategies, especially considering if their confidence in the system changes when they receive “good” or “bad” guidance. Today, a comprehensive overview of potential guidance strategies does not exist. To close this gap, we conducted a double-blind WOz study in which participants were either assigned the role of *user* or *wizard*. We call a **user** a study participant in the role of the VA user. Users interact with the VA workspace to solve their analysis tasks. In contrast, a **wizard** is a participant in a wizard role. Wizards observe users solving their tasks through a synchronized interface and provide guidance as they see fit.

A. Study Goals

We expect that users and wizards will develop various strategies to achieve their goals and that combinations thereof will affect interaction dynamics and impact how tasks are solved. We thus tackle two topics:

Capturing Guidance Interaction Dynamics and Strategies – Determining when to provide which guidance to what user remains the central challenge in guidance research. By observing

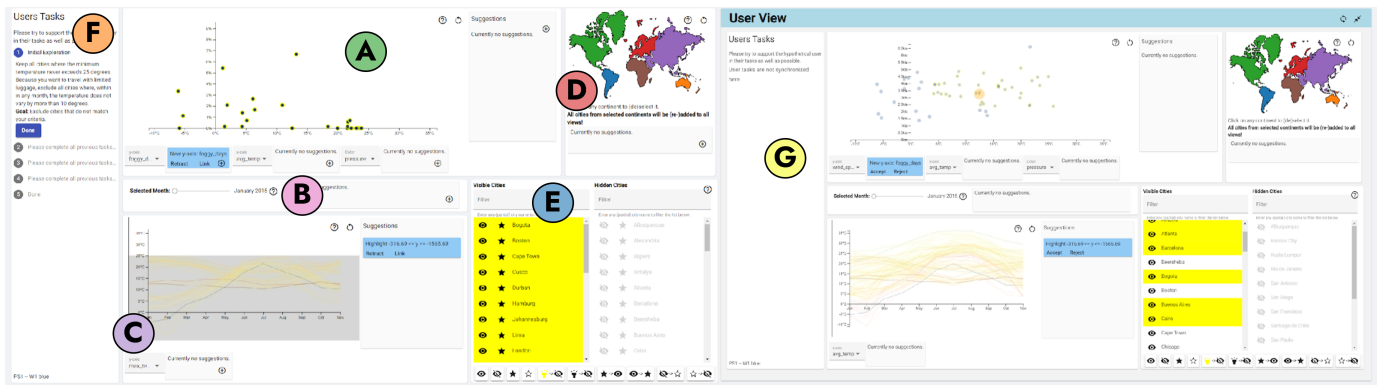


Fig. 1. The interface used for the study includes a *scatter plot* **A** with a *time selector* **B**, a *line chart* **C** for details, a *map* **D** used to filter, a *list of visible cities* **E** and a *task list* **F**. Wizards additionally see the “*user view*” **G** in which they can observe their user’s interactions.

users and wizards simultaneously, we aim to capture guidance interaction dynamics that arise when users are presented with guidance. In particular, we are interested in how their trust in the guidance, frustration, confidence, and analysis paths change. As humans are much better at capturing and understanding changing contexts than current guidance implementations, we expect wizards to be able to adapt to what their users are doing quickly. Under this premise of adaptation, we are interested in guidance strategies that wizards derive and if, how, and when they adapt their strategies to user behavior.

Effects of Varying Guidance Quality – Previous work has shown the effectiveness of guidance in supporting VA experts in solving their tasks (e.g., [9], [28], [40]). However, previous work also highlights the potential risks of guidance [14], [15]. In addition to confirmation bias, other biases like data bias can be introduced. Beyond those biases, guidance introduces risks like unsettling insecure users or abruptly changing promising analysis paths. The effects of such misguidance will be most visible when the task the system is guiding toward does not match the user’s task. In this study, we explicitly introduce such situations through the task setup to observe their impact on the quality of the achieved results, task completion times, and users’ interaction patterns with suggestions.

B. Study Design Choices

Why Wizard of Oz? We chose to employ a WOz design to enable a wide variety of guidance interaction patterns rather than limiting our study results to a few guidance strategies that could be implemented. Furthermore, wizards can adapt to a user in real-time, while implementing adaptive guidance remains challenging.

Why Double-Blind? We define a double-blind WOz experiment as one where *neither* the user nor the wizard know that they are interacting with a human. This category of studies is not included in Steinfeld’s design space [37]. We explored a double-blind WOz study design to limit the impact of emotions on guidance behavior. This decision should limit the risk of wizards and users changing their behavior not to “offend” or “hurt” the other part. In particular, as we relied on the wizards’ creativity to produce a variety of approaches, we found this decision critical. Our expectation is that this decision enabled

more realistic guidance strategies. Further, we made this decision to limit potential influences of confirmation bias or social desirability bias that might have occurred (more heavily) if participants knew they were interacting with humans—a setting that they might have been familiar with from pair analytics settings. Further, double-blinding should reduce the wizards’ stress level in the misguidance condition. Regardless, during the experiment, some wizards believed users were human (e.g., if they acted faster than they could react) and got stressed or frustrated if they did not manage to help them.

Why teach wizards? Before running the study, we determined that training the wizards was paramount to not overwhelm them with learning a new interface, identifying the solution, and guiding users simultaneously. We argue that training wizards in the analysis tasks and goals before the sessions allows them to provide reasonable guidance within the time and distance-only constraints of our experiment. Despite being adaptive, real-world guidance implementations in systems typically also follow pre-determined guidance strategies [33] or assumptions. Hence, training wizards on the tasks and data before the study does not make the study setup invalid, in our opinion. While they knew the answers for all tasks, wizards could not immediately provide those answers to users due to intended limitations in the system’s interaction affordances (see section IV-C).

Why not prescribe wizard reactions? We provided wizards with training on the specific tasks users would have to solve. However, our goal was to not limit wizards in the guidance strategies and actions they provided. Hence, we did not prescribe situations, contexts, or user behaviors in which specific guidance suggestions should be provided as a reaction. Instead, wizards were free to explore guidance options and, importantly, adapt the provided guidance to the user’s need. With this study design choice, we traded off reproducibility for a wider spectrum of guidance. As stated above, we chose a WOz design, among other reasons, for the flexibility of wizards and their ability to adapt.

IV. A WIZARD OF OZ ENVIRONMENT FOR VISUAL ANALYTICS

To conduct our experiment, we developed a VA system (see Fig. 1) that shows weather data in a dashboard-like interface

and provides linking, brushing, and filtering capabilities. The system [2] and implementation [1] are available online, and the implementation is open-sourced.

Before introducing the system in detail, we give a short preview of our study scenario that we introduce in section V. We designed a scenario where friends wanted to go on vacation. To decide the location, an individual (in the role of the *user*) from this group queries a VA system to analyze weather data of a set of cities on different continents. We integrate guidance (through the *wizard*) to support users in choosing a target city. Across four consecutive tasks, the user imposes additional constraints regarding geographical-, weather-, or activity-related preferences from the rest of the friend group, excluding more and more cities until a suitable location is found.

The presented VA interface was the result of various discussions and iterations. We wanted to create an interface that would allow a wide variety of tasks with sufficient freedom for solution strategies and complexity to warrant guidance. We quickly discarded initial ideas to incorporate text-analytics interfaces as too narrow or complex to teach. The next inspiration came from the gapminder project's simple visualizations of (geo)political or societal data. From there, we continued simplifying potential tasks in order to not require domain-pre-knowledge from participants and finally settled on the vacation planning task with basic visualization interfaces that could be learned in the short time frame of a study session.

A. Visualizations and Interaction Affordances

Our system offers multiple, linked views that allow assigning sufficiently complex and, hence, realistic tasks to participants to capture realistic guidance interactions. The interface combines simple and well-known visualizations to reduce the participants' learning effort and avoid observations specific to bespoke visualizations that would be harder to generalize. At the same time, it enables complex enough tasks for a study setting in which the wizards' support through guidance suggestions is actually helpful to users. While our results are still specific to the concrete implementation used, our system is a realistic approximation of existing VA tools and not oversimplified to fit the study.

In the following, we introduce all components labeled A through G in Fig. 1. Users can select the x- and y-axis of the scatter plot A via dropdowns. A third dropdown allows changing the global color map, also affecting the line chart. In the scatter plot, users can click and drag to draw rectangles to highlight all contained data points or to zoom in. To zoom back out, participants double-click anywhere in the scatter plot. Hovering a data point shows a tooltip with the city name, its continent, and the values of the two selected axes. Below the scatter plot, a time slider B shows which month is currently visualized and allows switching to a different month. The line chart C shows the temporal development of the selected dimension and a scatter plot, inspired by *Dimpvis* [21]. The x-axis is fixed to time. The y-axis dimension can be selected from a dropdown.

The tooltip from the scatter plot is also available in the line chart. Participants can brush data from the line chart to

highlight it in yellow via clicking and dragging. The world map D in the top right corner allows toggling the visibility of all cities within a given continent with a single click. Finally, the city lists E at the bottom right show which cities are visible and hidden from the visualization. Clicking the eye icon next to any city toggles its visibility while clicking the star icon toggles whether a city belongs to the set of potential vacation targets or not. Below the lists, participants can access various convenience functions, such as showing or hiding all cities, adding or removing stars from all of them, or hiding all (un-)highlighted data points.

We synchronize brushed cities between the scatter plot, the line chart, and the cities list. All visualization components contain *info* buttons that open tooltips presenting summaries of the component, their purpose, and all interactions. Further, visualizations offer a *reset* button that restores them to their initial state. Finally, on the left-hand side, the task list F shows the current task, overall progress, and, upon task completion, the Likert scales for task difficulty and guidance quality.

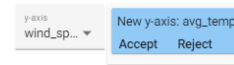
B. Guidance Suggestion Visualization and Interactions

The suggestions of wizards are shown as blue cards next to the respective visualization or axis-selection dropdown that they target. Initial interface prototypes used a light gray, which participants often did not notice, leading us to switch to a more conspicuous shade of blue. All suggestions contain a brief, auto-generated summary of their content or impact: *axis-changes* mention which axis would be selected, *highlights* summarize how many data points would be highlighted, and *zooming* contains the new viewport dimensions. We decided against wizard-generated guidance descriptions or justifications. Generating meaningful suggestions and justifications is a challenging and time-consuming task that would overwhelm wizards, especially as users continue their work without waiting. Further, users believe that a system implementation provides the guidance. When we designed and conducted the study, we expected human-generated explanations to surpass the quality of generated explanations so they could reveal wizards as humans. With the recent advancements of large language models and the popularization of chat interfaces, this gap is likely much smaller and auto-generated justifications might be a viable option in future systems. Hovering over any suggestion tentatively applies the proposed changes as a preview. Users can reject or accept suggestions by clicking the respective buttons or simply ignore suggestions as they never restrict the interaction possibilities or interface functionality.

Requesting Guidance – In our study interface, users do not have an explicit option to ask for guidance, e.g. by clicking a button. We deliberately made this choice to leave wizards with more freedom and be able to observe a wider variety of wizard strategies.

C. Additional Wizard Interface Functionality

Wizards need to follow the users' actions to provide meaningful guidance. The *user view* G provides this information as



a mirror image of the user's visualization interface. The direct, native integration into the VA interface believably conveys to wizards that they might be interacting with a trained AI simulating a user. Simply sharing the user's screen with the wizard would make them aware of the existence of a human user. As the analysis progresses, the user's and the wizard's views might diverge significantly. To facilitate making up-to-date suggestions the wizard interface contains a synchronize-button that matches the status of their interface to the exact user settings. We log and synchronize user actions (i.e., mouse movements and clicks) to replicate them in the wizard's user view. The user's mouse position is shown to the wizard as an orange circle. Clicks are indicated through pulsing animations at the pointer position.

Making Suggestions – Wizards can make suggestions on almost all interactions available to users. Only interactions with the cities list are excluded, limiting wizards to orienting or directing suggestions. Since wizards are already trained in the experiment tasks and already “know” the set of possible answers, we introduced this limitation to keep wizards from directly interfering with the users' result set. Without this restriction, we would have risked wizards immediately suggesting the answer to the last task, cutting the users' analysis short and significantly limiting the analytical discourse between users and wizards.

Whenever wizards perform any of the interface interactions (e.g., brush, change an axis, etc.), they are shown a small dialog box from which they can choose whether they want to perform the action only for themselves (e.g., for exploration or to set up more complex suggestions later) or suggest them to the user as well. To minimize the impact on the wizards' workflow, clicking outside the dialog closes it immediately without making a suggestion. Similar to the user's view, any created suggestion is shown next to its corresponding component and offers two buttons to *link* and *retract* it again. Both interactions are described below. All available suggestions are listed in Table I and can be created via the suggestion dialog mentioned above. In addition to this dialog, wizards can also create new suggestions from the plus icons in the suggestion spaces associated with all visualizations (line chart, map, scatter plot). All suggestions are implemented to change the user's view to exactly reflect that of the wizard when making the suggestion.

Any suggestions that wizards make are directly shown to users as described above. Our study environment does not contain any mechanisms to decide whether a potential

suggestion will be seen as useful by the user before showing it. Instead, the study design relies on wizards deliberately providing guidance that they deem useful.

Linking Suggestions and Higher-Level Guidance – While individual suggestions for each interface component are often sufficient, there are scenarios in which multiple connected suggestions are necessary. Consider, for example, the case where a user has selected different axes for the scatter plot than the wizard and receives a suggestion to zoom in. It is likely that the zoomed-in region does not contain the same data that the wizard saw. To mitigate this issue, wizards can link two or more suggestions. Linked suggestions are previewed, accepted, and rejected at once. In addition, linked suggestions also enable wizards to communicate more high-level ideas of guidance, like, for example, switching to other visualizations and performing analysis there.

Retracting Suggestions – Wizards can retract suggestions at any time, for example, when they are outdated. A retraction stops the user's preview (if running) and removes the suggestion from the UI. Our environment does not support “aging” or auto-expiration of suggestions. Our rationale is twofold. First, we observed in early tests that it made the wizards' tasks too complex by overburdening them with options. Second, it would have made the users' interface more complex with discriminators for expired and retracted suggestions. Hence, we told wizards to retract outdated suggestions manually.

V. STUDY METHODOLOGY

As introduced above, we designed a scenario where friends wanted to go on vacation. The participant in the user role continuously excludes potential target cities until a suitable destination is found.

A. Procedure

We summarized the study procedure in Fig. 2. It consists of blocks of four simultaneous study sessions, each joined by one participant for a total of two users and two wizards per block. Each row in the figure represents one study session. Each participant connected to a video call with one of the experimenters. As we had 24 participants, we ran six blocks.

Study sessions last at most 120 minutes. All participants (both users and wizards) started with a **semi-structured entry interview** (15-20 minutes) before heading to a **training session** (20-25 minutes). For users, the training focused on learning to use the interface with bespoke training tasks. Wizards trained on the concrete tasks they would later provide guidance for. We introduce this differentiation to ensure that wizards have sufficient knowledge about the data and tasks they would later guide, allowing them to develop guidance strategies and mitigate potential response time issues [8]. For each task set, the completion time was theoretically limited to 20 minutes; however, no participant needed that much time. Wizards were not instructed on which suggestions to provide and when. The task sets and pairings, as well as the (unintentional) misguidance, will be introduced in detail below.

After each task, both users and wizards answered a short **questionnaire to assess task difficulty** (on a five-point scale

TABLE I
SUGGESTION ACTIONS AVAILABLE TO WIZARDS PER INTERFACE ELEMENT.

Interface	Action(s)
Scatter Plot	1) Zoom into a value range. 2) Highlight rectangular selection of data points.
World Map	1) Show continent(s). 2) Hide continent(s).
Line Chart	Highlight data points based on a y-axis range.
Time Slider	Switch to a different month.
Axis Dropdowns	Change attribute mapped to x/y axis of scatter plot and y-axis of line-chart.
Color Dropdowns	Change attribute mapped to the color data elements in the scatter plot and the line-chart.

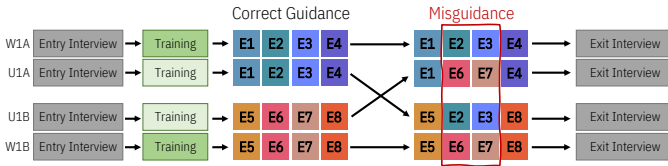


Fig. 2. The study procedure. Participant labels on the left denote Users and Wizards. The number represents the study block (1-6) and the letter A and B the task set(s) of that participant. All participants start with an interview and training. Wizard-user-pairs are changed after the first task set. In the second set, wizards unintentionally provide misguidance. All participants finish with an interview. *The task set order and assignment to users/wizards here is one example setup. In the study, the order of task sets was counterbalanced and tasks from all sets were misguided equally often.*

from *very easy* to *very difficult*) and perceived guidance quality (on a five-point scale from *very poor* to *very good*). After each task set of four tasks, we collected raw Nasa **TLX scores** on a seven-point Likert scale.

After the first task set, users were paired with a different wizard. To coordinate this switch, the experimenters closely communicated via instant messaging. When one user-wizard-pair finished before the other, they had to wait for up to ten minutes. They were told that we had to ensure all data was logged or that servers had to be restarted between sessions to cover such breaks. Each study session finished with a semi-structured **exit interview** (20 minutes) in which we revealed to participants the WOz nature of our study.

B. Tasks and Data

We designed two parallel series of four tasks each (E1-E4 and E5-E8). We ensured that aligned tasks from both series (i.e., E1 and E5, E2 and E6, etc.) had the same difficulty and used similar interfaces and solution strategies but on different data dimensions. To avoid the misguidance condition (see below) being immediately obvious, the tasks E6 and E7 were derived from E2 and E3, respectively. Both E2 and E6 require filtering based on two criteria, while both E3 and E7 require picking a month that leaves the most options for target cities. However, we did not run formal pre-studies or pre-tests to evaluate the tasks. We report task E1 as an example here:

E1: Keep all cities where the minimum temperature never exceeds 25 degrees. Because you want to travel with limited luggage, exclude all cities where, within any month, the temperature does not vary by more than 10 degrees.

The complete list of tasks can be found in the supplementary material or tested online [2]. We limited the complexity of tasks to give wizards a realistic chance to keep up with participants without lagging and revealing that this was a WOz study.

Task Sets – From the eight available tasks, we derived **four task sets**: A1 (E1-E4, ) , B1 (E5-E8, ) , A2 (E5, E2, E3, E8 ) , and B2 (E1, E6, E7, E4, ) . We use blue and red colors here to highlight that task sets A1 and B1 are distinct. A2 and B2 reuse tasks from their respective first sets. We adopt a **within-subject study design**: each user was either shown the two task sets A1 and A2 or B1 and B2.

Each *wizard* participant provided guidance for only one task set – first to an initial user, and then to a second, different user after the switch. We deliberately do not change wizard tasks to avoid overloading wizards and to observe whether they developed one particular strategy or changed their approach to guiding users over time.

Task Assignment to Users and Wizards – Each study block with four simultaneous participants uses all four task sets for the users and two task sets for the wizards (see Fig. 2). The first task set will be guided properly (as the user and wizard task set match). In the second task set, wizards provide guidance for the same tasks again but users are shown tasks that do not match the tasks shown to wizards. Hence, wizards will (unintentionally) provide misguidance. We decided not to randomize the occurrence of misguidance between the first and the second task set to avoid users losing confidence in the system early on. We discuss our rationale and potential implications in section VII-B. User **task set orders are counterbalanced** across the study – thus, each task set has been guided correctly and misguided exactly once after two study blocks.

When designing the study, we also considered having participants switch roles (i.e., wizards become users, and users become wizards) after the first four tasks. We discarded that idea as we foresaw uncontrollable effects when revealing that this was a WOz study halfway through the session.

User Task Types and Task Taxonomy – The study tasks required participants to filter data (both conditioned on the value of a second dimension and unconditionally), observe correlations between dimensions, select appropriate months from the data set, and finally select vacation target cities from a Pareto front. These tasks can be categorized as a repeated sequence of *explore*, *browse*, and *identify* tasks according to the taxonomy by Brehmer and Munzner [7]. For the final task, users have to *compare* the remaining candidates before *identifying* their final solution.

Correct solution sets, i.e., the number of cities that would satisfy all the constraints, always counted under five options for all task sets. To solve all tasks efficiently, users had to interact with all visualizations. Most tasks offer less efficient solution strategies using alternative visualizations. We aimed to make the tasks sufficiently complex to leave room for different solutions and guidance strategies. We formulated the descriptions so they did not reveal the solution as an ordered list of steps. Hence, most tasks introduce at least two new constraints to increase the complexity.

Wizard Task – Wizards were tasked with providing guidance. To keep the illusion of interacting with an AI simulating a user, they were told the following: “Your task will be to take the role of the guidance system and provide suggestions to a hypothetical, simulated user. You can observe that simulated user in your second window. Try to help them as well as you can by following along with what they are trying to do and suggesting what they should do next.” As users were primarily performing the *search* actions *explore* and *browse*, wizards had several options according to our taxonomy for guidance

tasks [27]. Using non-disruptive guidance, they could *steer* the exploration or *indicate* where to browse. Using disruptive methods, they could *direct the path* or *direct the target* during browsing and exploration, respectively.

Wizards had complete freedom over their suggestions' timing, type, and content. In contrast to previous studies evaluating assistive agents using WOZ studies (e.g., [22], [42]), we did not impose limitations on wizards regarding which guidance they could provide at which time. As outlined above, we designed user tasks to be diverse and representative of typical VA tasks. Most tasks also offered multiple solution strategies using different visualizations, giving wizards as much freedom as possible within the controlled study environment. We remark that wizards did not see the users' current task but only when a user had switched to the next one in the list (i.e., they felt they had completed the task). This was necessary to enable the misguidance conditions of the study.

Data – We use weather data from 70 cities for a single year, obtained online [3]. We aggregated the data to a monthly level and used 12 weather dimensions: *humidity; pressure; average, minimum, and maximum temperature; wind direction and speed; and the percentage of rainy, foggy, snowy, stormy, and cloudy days*. Additionally, the data contains the city names and the corresponding continent.

C. Participants

We recruited 24 participants. All participants were final-year master's students (4), Ph.D. students (16), or post-docs (4). While all participants had previous experience with VA, not all were previously familiar with the term *guidance* or its definition in VA. None of the participants had seen the study prototype or were informed about the study goals before their session. Participants received no monetary reward. Due to technical issues with the study prototype, we had to abort five of the 24 task sets during the study. Any data collected until that moment are excluded from the analysis below. However, information from entry- or exit interviews was kept as qualitative feedback. We did not perform any quantitative pre-tests before assigning participants to the user/wizard groups but aimed to nominate more experienced participants as wizards. This assignment was made based on our impressions and previous knowledge about participants when they agreed to participate in the study. We did not change the assignment during the study, e.g., as a result of the initial interviews.

D. Data Collection and Analysis

Each study session was performed as a video call, for which we recorded both audio and the participant's shared screen for later analysis. The implemented system automatically captured a wide variety of data, such as the mouse position and all clicks, to capture different interaction behaviors. This information was primarily used to update the wizards' interface. We further recorded the task completion time, the number of interactions, and the correctness of the result. To determine correctness, we calculate the average of the pairwise Euclidian distances (across all 12 weather dimensions, see section V-B) between the cities

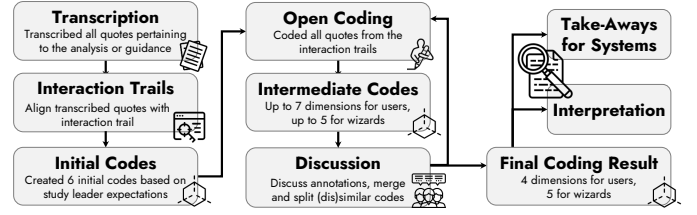


Fig. 3. Qualitative Methodology: We open-coded all guidance-related quotes and iteratively converged to a total of nine codes.

a user has not excluded after finishing a task (candidate list) and the correct solution after that task. Regarding guidance, we measure how long users take to interact with a suggestion first and how long they take to accept or reject a suggestion. Finally, we collect explicit information about guidance quality and task difficulty through two Likert scale questions after each task and a TLX questionnaire after each task set.

We annotated all guidance situations from the recorded videos to analyze interaction dynamics and guidance strategies. We followed the iterative process outlined in Fig. 3 and first transcribed utterances and aligned them on a timeline visualization inspired by Boukhelifa et al. [5], creating *study trails* (see section VI-B). To code all utterances, we employed an iterative open-coding process. We annotated all guidance situations and used regular meetings between all authors to discuss annotations and refine the code set. We ultimately converged to four coding dimensions for users and five for wizards, presented below. Further, we analyzed the entry interviews to capture participants' previous experience with guidance and their expectations toward the guidance. From the exit interviews, we captured whether their opinions changed, if and how the guidance influenced users in their analysis, and which strategies wizards followed.

VI. STUDY RESULTS

We first provide summary statistics outlining how participants interacted with the provided guidance. Subsequently, we turn to qualitative analyses of the guidance interactions and interviews. In the following, we frequently use the terms *first* and *second* task set to denote the task sets that users solved *before* and *after* switching wizards, respectively. Fig. 4 through Fig. 9 reference tasks from 1 to 4. These tasks refer to the ordered tasks of a task set (e.g., for task set A2 : *Task 1* is E5, *Task 2* is E2, *Task 3* is E3, and *Task 4* is E8). This abstraction is necessary as the task sets that participants work on, as well as their order, are balanced across all study participants.

A. Quantitative Analysis

Interaction Statistics – To get an initial overview of user interactions, we count interface interactions under both proper guidance and misguidance. We consider axis changes, color map changes, data brushes, zooming, opening tooltips, and interactions with the cities list. Users needed, on average, 205.3 ($s = 123.9$) interface interactions to complete the four tasks in one study session. The high standard deviation reflects the variance in solution strategies applied by users, as well as the differences in perceived task difficulty. We observe no



Fig. 4. Users report a significant drop in guidance quality (scale 1-5, higher is better) for the two misguided tasks in the second set.



Fig. 5. User-perceived task difficulty (scale 1-5, lower is better) is not significantly influenced by misguidance for tasks 2 and 3 in the second set.



Fig. 6. User task completion times (in seconds, lower is better) vary widely across users and are not significantly influenced by the guidance quality.

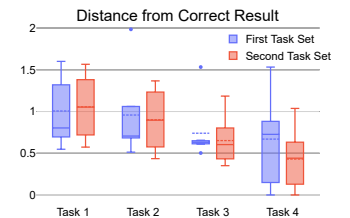


Fig. 7. Normalized distance from the correct result (lower is better) is not significantly influenced by guidance quality.

TABLE II
USERS NOTICED A CLEAR DROP IN GUIDANCE QUALITY IN THE MISGUIDANCE CONDITIONS, WITH REJECTION RATES TRIPLING.

Guidance	Accept	Reject	Retract	Preview	Ignore
Proper Guidance	66.6%	23.3%	3.3%	6.6%	0%
Misguidance	23.6%	72.0%	1.0%	3.2%	0.2%

TABLE III
WHEN FACING MISGUIDANCE, USERS IGNORE SOME SUGGESTIONS MUCH LONGER THAN WHEN FACING PROPER GUIDANCE. HOWEVER, REJECTION DECISIONS ARE MADE MUCH FASTER. DECISION TIME DENOTES THE TIME PASSED BETWEEN THE FIRST PREVIEW AND THE DECISION.

	Status	First Preview		Last Interaction		Decision Time	
		mean	std	mean	std	mean	std
Proper Guidance	Accept	11.3s	10.3s	17.4s	13.8s	4.8s	6.5s
	Preview	19.9s	33.0s	19.9s	33.0s	—	—
	Reject	15.5s	19.6s	35.8s	32.6s	18.4s	32.8s
Mis-guidance	Accept	10.8s	7.5s	15.4s	10.2s	3.1s	4.3s
	Preview	27.8s	3.9s	27.8s	3.9s	—	—
	Reject	14.3s	17.7s	29.9s	48.1s	8.4s	11.6s

significant differences between the number of interactions in the first ($\bar{x} = 206.7$, $s = 87.0$) and second task set ($\bar{x} = 204.2$, $s = 152.0$). We did not aim our study at capturing differences between users with varying experiences with guidance systems. To assess potential differences, further studies are needed.

Guidance Quality – As introduced in section III, bad suggestions are some of the main challenges in guidance. Overall, participants recognized misguidance in most situations, as Table II shows with a clear drop in acceptance rates during misguided tasks. This is confirmed by Fig. 4, which shows a significant drop in user perception of guidance quality for misguided tasks from the second set when compared to those from the first set ($p = 0.03$ for Task 2, $p = 0.01$ for Task 3).

Interaction timings are summarized in Table III. The data also shows that users take more time to start reviewing suggestions in the misguidance condition, indicating lower expectations. While not statistically significant, they make their acceptance or rejection decisions more than twice as fast in the misguidance condition (8.4s vs. 18.4s).

Task Difficulty – We did not observe significant differences between how users perceived task difficulties (Fig. 5) of the first and second task sets, respectively. This is mirrored in the raw TLX scores shown in Fig. 10, which decrease after the second task set for all dimensions other than frustration. We

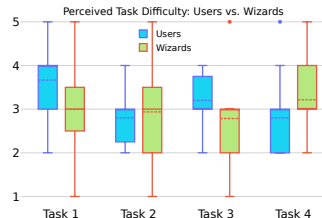


Fig. 8. There is no significant difference between the perceived task difficulty (scale 1-5) of users and wizards.

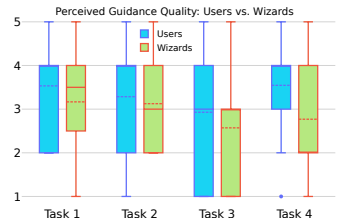


Fig. 9. While the differences are not significant for any of the tasks, users rate guidance quality (scale 1-5) higher than wizards on average.

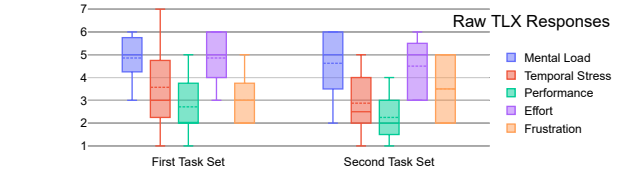


Fig. 10. Raw TLX scores (lower is better) improve for the second task set, likely due to interface learning effects. Only frustration scores increase slightly because of the provided misguidance.

hypothesize that this development is due to learning effects and that the provided misguidance causes increased frustration as participants mentioned that “it suggested things that don’t really have anything to do with the current task” and “some of the suggestions were not really according to the task”. We provide more detailed qualitative feedback in section VI-B.

To verify that there was no large disconnect in the perceived task difficulty between users and wizards, we also compare these groups in Fig. 8. In this evaluation, as well as the one shown in Fig. 9 we no longer compare user responses across task sets. Instead, we compare user responses with wizard responses. Our results show no significant differences for any of the tasks. In a pairwise comparison between users and their respective wizards, users rate tasks 0.1 points more difficult (on a five-point scale) on average. The standard deviation is 1.3, and the biggest rating difference 3.0 points. In a similar direction, we compared the perceived guidance quality between user-wizard pairs. The results are shown in Fig. 9. On average, across all tasks, users rate the quality 0.4 points better than wizards. While the standard deviation is high at 1.5, this may indicate that users appreciated any help they could get, despite wizards believing that they should have provided better support. The biggest rating difference is 4.0 points.

Task Completion Times – The task completion times are summarized in Fig. 6. In both task sets, users need the most

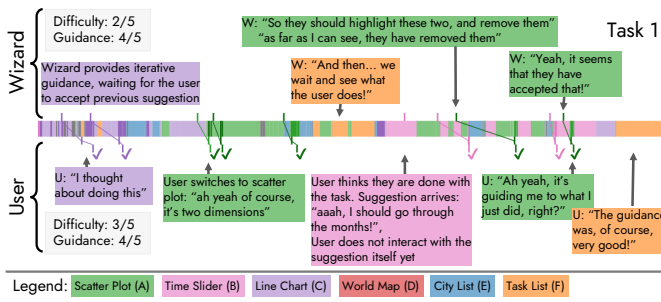


Fig. 11. Snippet of an annotated interaction trail. Interaction trails show which components users focused on through color, when they received and previewed suggestions, and when they made decisions.

time for the first, most complex task. While we do not observe significant differences, users complete the two misguided tasks faster than with correct guidance and instead take more time to finalize their results.

Task Correctness – Fig. 7 shows the distance to the correct result after completing each task. Interestingly, users seem initially impacted by misguidance but recover towards the end of the task. We observe bigger differences between individual users than between the first and second task sets.

The standard deviations for most reported measures are high, indicating that such aggregated data analysis might not be suitable for complex guidance task evaluations. Due to the multitude of different users, personalities, experiences, and expectations, many different guidance situations and needs can be identified. While our interface is less complex than full-fledged data analysis applications, we intentionally provided various visualizations and interaction affordances to simulate a realistic VA session. Rather than averaging numerical values, we argue that co-adaptive guidance should be evaluated using qualitative methods on a per-event basis. Alternatively, more constrained studies with more participants in the future might provide clearer results.

B. Qualitative Analysis: User Behaviour under Guidance

In this section, we describe fine-grained guidance dynamics from the user perspective based on a qualitative analysis of the task sessions and the interviews performed with users. All data were analyzed as described in section V-D, producing *study trails* for each session. We provide an example of an annotated study trail in Fig. 11 and use snippets of trails to summarize guidance situations throughout the following sections.

Each trail consists of the **interaction line** that shows which visualizations the user interacted with over time. All colors correspond to those used in Fig. 1. Low saturation shows interaction with the given component, and high saturation indicates the preview of suggestions in the respective visualization. **Suggestion bars** (shown above the interaction line) denote points at which the wizard made suggestions. In contrast, bars below the interaction line denote when users previewed and accepted or rejected suggestions, respectively. Below the bars, accepted suggestions are marked with ✓ and rejected ones with ✗. All three time-points of a suggestion are connected with a line, forming a version of Marey's graph [39]. Analyzing the

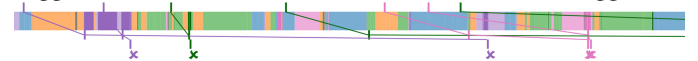
line's slope provides an overview of the user's reaction times to guidance. More vertical lines indicate that users responded to guidance faster, while more horizontal lines show that they ignored suggestions for longer periods. We refer to user participants with shortcodes like U1A or W5B as introduced in Fig. 2 below. The initial letter denotes **Users** and **Wizards**. The number represents the study block (1-6) and the letter **A** and **B** the task set(s) of that participant.

“Standard” Analysis Behavior – By analyzing the quantitative results as well as the users' feedback, we observe that, according to users, wizards were able to provide them with decent suggestions, close knowledge gaps they faced, and help advance the analysis. Most study sessions contain a majority of accepted suggestions, and a few rejects. Below, we focus on situations where users deviated from this expected behavior.

Waiting for Guidance – One of the first observations we made was that some users waited intentionally for guidance, implicitly delegating the lead role and the completion of the task to the wizard. For example, participant U5B, who had already previously expressed mixed opinions about the provided suggestions despite a high acceptance rate, had been conditioned by their wizard that there was an abundance of suggestions. When there was no suggestion as they started task T2, they hesitated briefly before they said “So I guess if there's no suggestion, I'll just do it myself?” when they received a suggestion to change the axis of their line chart. They accepted it immediately and their wizard had already provided the appropriate highlight suggestion, which they celebrated by exclaiming “okay! nice!” After finishing the task, they took some time to realize they were done, waiting for further suggestions.

This situation suggests that over-guiding users can make them dependent on new suggestions, otherwise pausing the interaction. As the user started the next task, they quickly set up the scatter plot using the correct axis but were again waiting for more guidance: “Okay, so I don't have any new suggestions now.” When they finally got suggestions, it was a misclick by their wizard, which they immediately identified as it “[did] not make any sense!” When they received the corrected but still flawed suggestion moments later, they exclaimed that “this seems to be correct”, accepting it without investigation. They finished the task, rating it *difficult* and the guidance *very good*.

Ignoring provided Guidance – We observed multiple users who primarily focused on their analysis, ignoring provided suggestions in various contexts. For U4A, the first suggestion



arrived while they were still reading their task, leading them to immediately reject it without further consideration. Later, they ignored and rejected various other suggestions that objectively made sense from the recording, exclaiming once that they had seen the suggestion and were not sure that they needed it then or that it was confusing. U3A showed similar behavior. They

did not consider any suggestions but just focused on task 1. When they finished without accepting any suggestions, they stated that they felt that there were “*probably just too many suggestions altogether.*” In a later stage of the analysis, U4A began to see the value of suggestions and started accepting more and more of them. Almost surprised by a suggestion to switch their scatter plot axis, they checked and accepted it, calling the suggestion “*perfect.*” This repeated later, with them calling further incoming suggestions a “*good idea*” before aborting their exploration and following the provided guidance.

These two users exemplify different effects of persistent wizards who continued supporting users despite their suggestions being rejected. In one case, the user was annoyed by the overload of suggestions; in the other, the user eventually built trust in the system and started to follow the guidance. Future guidance system implementations could observe both explicit and implicit trust measures to identify whether to continue guiding or refrain from further suggestions, e.g., the consistency of user interactions and their agreement with system suggestions over time [6], how long users take to react to new suggestions and how long they spend evaluating them, or how often users request system suggestions.

Verifying Guidance Suggestions – In other scenarios, we observed how users tended to be very cautious. U3B had expressed in the entry interview that they would usually dismiss guidance and try to solve tasks themselves. In our study, they instead checked each suggestion very thoroughly, occasionally cross-referencing suggestions with the task description. Consequently, they spent quite some time checking suggestions and noted that they would not let suggestions change their analysis path. Further, they found the misleading suggestions from the second task set particularly distracting, as they examined them with the same rigor, trying to understand the system’s reasoning process. They critiqued that they “[*couldn’t*] understand why it would suggest that now” and that “*this comes a bit out of the blue now.*” Ultimately, they argued that systems should always provide a rationale for suggestions and keep users in control rather than automating tasks. This is in stark contrast to other users who were often less critical towards suggestions, accepting them more liberally, not thinking of the possible consequences. However, a common pattern that we observed was that users accepted guidance only to realize issues with the suggestion moments later and revert it. For example, users U6B and U5A switched the axis of their scatter plots upon suggestion, realized this was not ideal and switched back immediately. U5A justified this behavior by asserting that they could “*always revert suggestions*” manually.

Ultimately, there were also users who started the study with high expectations of guidance, often forgiving their respective wizard’s several bad suggestions and still rating the guidance as helpful. U6B stated that the guidance could “*probably solve all problems and tell us the ideal solution*” on the spot, and U5B accepted several suggestions without any validation, sometimes even adding that “*this seems to be correct*” or “*I could just do that without thinking about it.*” In the exit interview, U6B justified their accepting behavior by stating that “*since I was*

trusting the guidance, it was obvious to me that I could get the task done faster if I followed the guidance.”

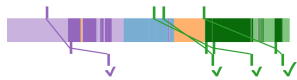
We grouped these user experiences here as we believe that they could all be addressed with appropriate justifications. Justifications will simplify the process of suggestion validation and increase user trust. For users currently not checking the validity of suggestions (enough), justifications can make it obvious that suggestions were made for the wrong reasons and are potentially flawed.

C. Wizard Guidance Approaches and Tactics

After observing particular user behavior patterns, we turn to wizards. While a few wizards premeditated their strategies during the training phase, others created them on the fly as necessary. The strategies can be categorized along the two dimensions *activity* and *goal*. In the *activity* dimension, we distinguish between *static*, *adaptive*, and *passive* wizards, depending on whether they provide premeditated suggestions or react to user behavior. We use *passive* to describe wizards who suspended suggestions after being rejected. In the *goal* dimension, we distinguish *correcting* wizards who actively aim to convince users to change previous interactions and *supporting* wizards who constantly search for alternative ways to help their users. In previous work, assistive agents have been categorized as *proactive* and *reactive* [42]. However, that classification refers to whether assistants take the initiative or not. As frequently outlined before, guidance describes a mixed-initiative process in which all systems should be able to take the initiative. Further, all guidance provided by wizards was proactive in the sense that users did not have options to specifically request guidance. In general, the guidance provided by wizards could serve as templates for future VA systems, as wizards frequently stated that they provided the guidance that they would have expected to receive.

Beyond those clear strategies, we also observed a spectrum from confidence (independent of the quality of their suggestions) to distress (because of task complexity) from wizards. After the first task set, W4B asked “*why is the user so fast? Is that normal? Or am I just slow?*” and whether we could “*slow down the pace of the computer.*” As a result, the provided guidance was sometimes lacking whenever wizards were catching up with their users (see also section VII-B). Below, we describe the wizard strategies we identified according to our categorization.

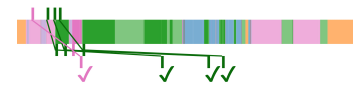
Non-Adaptive Wizards – Wizards often proactively provided guidance in the first task of the second task set. At this stage, they had provided guidance during training and in the first task set and were likely confident enough to stack up several suggestions, sometimes leading directly to the solution of the task. In the example on the side, W5A provided a stack of suggestions that the user reviewed and accepted one by one. With few exceptions, most users appreciated such static guidance (or requested it in the exit interview if they did not receive it) as they did not have to wait for it or could perform the task themselves before receiving outdated suggestions. However, this finding might



checked each suggestion very thoroughly, occasionally cross-referencing suggestions with the

task description. Consequently, they spent quite some time checking suggestions and noted that they would not let suggestions change their analysis path. Further, they found the misleading suggestions from the second task set particularly distracting, as they examined them with the same rigor, trying to understand the system’s reasoning process. They critiqued that they “[*couldn’t*] understand why it would suggest that now” and that “*this comes a bit out of the blue now.*” Ultimately, they argued that systems should always provide a rationale for suggestions and keep users in control rather than automating tasks. This is in stark contrast to other users who were often less critical towards suggestions, accepting them more liberally, not thinking of the possible consequences. However, a common pattern that we observed was that users accepted guidance only to realize issues with the suggestion moments later and revert it. For example, users U6B and U5A switched the axis of their scatter plots upon suggestion, realized this was not ideal and switched back immediately. U5A justified this behavior by asserting that they could “*always revert suggestions*” manually.

Ultimately, there were also users who started the study with high expectations of guidance, often forgiving their respective wizard’s several bad suggestions and still rating the guidance as helpful. U6B stated that the guidance could “*probably solve all problems and tell us the ideal solution*” on the spot, and U5B accepted several suggestions without any validation, sometimes even adding that “*this seems to be correct*” or “*I could just do that without thinking about it.*” In the exit interview, U6B justified their accepting behavior by stating that “*since I was*



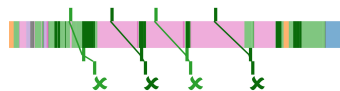
be an artifact of wizards struggling to keep up with users at times.

Adaptive and Supporting – In contrast to static, adaptive wizards focus more on the user and what they are doing. Wizard W1B provided prototypical adaptive, supporting guidance in their first task set. While they started with one static suggestion, they kept observing the user's view and timed all subsequent suggestions such that they just followed the user accepting the previous suggestion. Wizard W6B followed a similar approach, repeatedly commenting that they would now *"wait and see what the user does"*, before happily acknowledging every accepted suggestion. When W51B realized their user was not considering all relevant months for a task, they suggested moving the time slider. When they realized their user was reacting to the guidance, they stated *"I'm influencing the user. That sounds good, as long as my suggestions are good!"*



In contrast to several static wizards who sometimes provided obsolete suggestions for actions that their users had just completed, adaptive wizards sometimes re-checked potential suggestions before submitting them to the user, like W42A, who realized that *"the user already did that"* axis change they were about to suggest. W3B stated that they did not like static guidance and that *"should only be provided when the user is stuck or stalled or there is a faster way."* To summarize their guidance strategy, W2B exclaimed that they *"would not want the system to be too dominant as that's disturbing and not helpful."* Consequently, they only provided limited guidance after observing their users exploring their options for longer periods.

Adaptive and Correcting – Whereas adaptive and supporting wizards aimed to adapt their guidance to the user, adaptive and correcting wizards insisted on their suggestions when rejected. This behavior was more apparent in but not exclusive to the second task set in which wizards unintentionally provided misguidance. When their guidance was repeatedly rejected, W1B said *"I can't believe it, I'll do it again"*. Similarly, W6A reported that their strategy was *"trying to force my solution onto him"*. They justified this decision by stating that they *"knew what I wanted to suggest was right"* but realized that their guidance was likely less appropriate if *"the goal was to go slow"*. Regardless, they did not reconsider or adapt their strategy over time. For some wizards, providing guidance turned into a game where they aimed to get as many accepts as possible. While some users changed their analysis paths in reaction to repeated, insisting guidance, a more typical reaction to exceeding amounts of correcting guidance was to stop looking at the suggestions altogether.



From the experiences during the study, correcting guidance was not particularly successful. It led to increased user frustration despite several participants mentioning that correcting wrong actions is a field of application for guidance. Guidance designers considering correcting guidance should carefully consider frustration and how to balance such strategies with

alternative behaviors.

Passive Guides – In order to avoid frustrating their users, some wizards with adaptive strategies stopped providing guidance for a given task after a few rejections. W4B noticed that *"maybe the user has different goals, actually"* before giving up on providing further guidance. Similarly, W5B observed the user making mistakes after rejecting some suggestions but never aimed to interfere with the user's analysis path anymore. W3B suggested the *humidity dimension* twice before realizing that the user would not accept and *"didn't provide any more suggestions"* afterward.



While this strategy of adapting by ceasing to provide guidance seemed to introduce less direct frustration with users, it also prematurely gives up on the benefits of guidance. Rather than directly stopping to provide guidance, several adaptive and supporting wizards switched to different guidance strategies, most often by aiming to provide suggestions on alternative visualizations. However, passive guides highlight the difficulty and importance of proper intent identification. In our study, some (human) wizards could not decrypt what users wanted to do based on the interface interactions they could observe.

D. Interaction Dynamics

In previous sections, we summarized the results of user-wizard interactions from their perspectives. This section summarizes noteworthy situations that arose from those interaction behaviors and strategies introduced before. Users were often not consistent in their interaction patterns and tended to switch between the three categories more frequently than wizards, who typically stayed within one category.

Impact of User-Wizard-Pairing – Our study aimed to bring together a diverse set of guidance strategies with different user preferences. As a result, we experienced both good and poor pairings between user requirements and wizard strategies. We observed the fastest guidance session between the well-prepared, adaptive wizard W6B and the trusting user U6B. The wizard provided complex linked suggestions that quickly led the user to the correct solutions, and the user rejected none of the 17 suggestions. While this session was very efficient and led to the correct result, the user later reported being stressed by the guidance, which *"came too fast; otherwise, it would have been very good"*.

In another case, W4A made the same suggestion repeatedly, eventually telling their user to *"just accept it."* While their user needed some time, they stated that they had *"the feeling that I'm doing this the completely wrong way."* While the user needed some time to recover, the repeated suggestions of their wizard made them start over and reach a better solution. In a different session, W3B noticed that their user was *"a bit slow"*. They were, therefore, *"trying to find a way to help them"* with task T1 (see section V-B) that required users to search for data points off the diagonal in the scatter plot. After multiple failed attempts of suggesting first data points on the diagonal, then outliers, and then other outliers, the user was eventually overwhelmed by the number of suggestions, mentioning that they were *"not sure why it suggested that now."*

While other study sessions highlighted the effectiveness of adaptive guidance, this one suggests that there are also limits to how fast guidance should adapt and change.

In the second task set, U5A experienced misguidance by a correcting wizard who re-suggested the same axis changes over and over again, eventually leading the user to doubt their analysis and try to figure out whether there was a connection in the data between fog, humidity, and rainy days that they should have considered. These examples show that it is difficult even for humans to appropriately support users with the limited conversation options that a VA interface presents.

Timing and Frequency of Guidance – Several participants commented that the guidance provided by wizards was sometimes a bit too late and arrived just as they were performing actions themselves or even slightly afterward. Typically, those delays were due to the wizards not being able to provide suggestions faster while coordinating two views and the users' state of the analysis. We discuss the implications of guidance timing for system implementations in section VII-A. Beyond the specific timing, multiple users commented that the *"guidance was more useful in the beginning"* (U2A) but then lost its usefulness as their *"ability to use the tool improved over time and the guidance did not keep up"* (U3A). Further, U1A commented that guidance in the second task set (the one with misguidance) felt less useful because of the high frequency of suggestions, independently of their quality.

Disruptiveness of Guidance – When the strategies employed by wizards did not match the users' expectations, some participants reported that *"guidance disrupted the analysis"* (U1A) as it appeared on other parts of the screen (like notifications), prompting users to *"look at them and think if they are relevant"*. Especially in the second task set, some participants found the suggestions distracting. However, not all participants were affected, with several reporting that they were barely affected by wrong suggestions as they were *"so concentrated that they did not notice them."*, especially those who already relied less on guidance in the first task set. Also, this highlighting feature was overwriting my selection. So I would say it was disrupting."

Learning and Adaptation – When designing the study, we expected a clear advantage of wizards over actual system implementations in learning which guidance to provide and when. As we relied on a diverse set of tasks that could be solved in different ways, we expected that wizards could pick up on user strategies quicker than machines - therefore, the choice of a WOz study. However, while some wizards adapted their guidance, we were particularly surprised by stubborn wizards who were unwilling to change their strategy or stopped altogether giving suggestions as soon as they were rejected. Instead of wizards **adapting**, we observed more users **learning** as they became more familiar with the interface, with their need for guidance declining as the study progressed.

VII. DISCUSSION

In this paper, we explore the impact of diverse guidance strategies on users and their visual data analysis processes.

This section first summarizes takeaways for guidance systems before discussing limitations.

A. Takeaways for Future Guidance Systems

In this section, we derive takeaways for designing and implementing future guidance systems based on our findings.

Explain and Justify Suggestions – While guidance describes a mixed-initiative process, it is likely that only the human user is ultimately responsible for the final result. Our study revealed a widespread user need for both justifications and explanations of guidance. Without justification, participants reported that *"it was unclear where the guidance came from or what it was supposed to do"* (U4B) and that they *"did not understand why a certain suggestion was given and could not understand if it was useful or not."* (U1A). In general, users were more cautious with guidance suggesting specific data points than with more open suggestions like changing the axis of a plot. Similarly, complex linked suggestions that changed the parameters of multiple visualizations were received critically, as they were harder to evaluate without justification. In addition, justifications can also serve to mitigate observed interaction patterns in which users accept wrong suggestions without verifying them. There, justifications can reveal a misalignment between the user's and the system's intents, prompting a deeper investigation. In the system-focused taxonomy by Messina et al. [27], we structure system guidance tasks into the dimensions *when, what, how* and *why*. Especially *when* and *what* can be tackled with existing, well-known XAI methods tailored to the specific guidance implementation. Our study revealed a deeper need for explanations and justifications of the *why*. We encourage guidance designers to think clearly about the *goals* of their guidance (and which goals are appropriate in which phases of analysis). Making such goals transparent to the users can clarify the *why* dimension of the guidance [27].

Match Users and Guidance Strategies – To suit a variety of users' needs, systems must offer several alternative guidance strategies that differ, for example, in how often the system takes the initiative, how high-level the suggestions are, and how prescriptive the guidance is. Systems could identify the knowledge gaps the users are already aware of, their analysis targets, or interaction preferences to select suitable strategies from the pool. In addition to processing implicit feedback, it might also be worthwhile to explicitly request user preferences.

Limit Specificity of Guidance – In our study, many participants expressed preferences for more *basic* guidance that pushed them to look at different visualizations (U4B, U2A, U4A, U6A) but left the primary analysis to them. In particular, participants were skeptical of suggestions that highlighted individual data points. Two wizards (W5B, W4A) aimed to generally explain how to solve tasks by providing iterative suggestions. They expressed that they aimed to teach their users how to solve the task rather than only highlighting the data points to exclude. However, both wizards failed to convey such general guidance. Guidance systems should justify more complex guidance suggestions. Suppose such justification cannot be provided for technical reasons. In that case, limiting

the provided guidance to simpler indications might be more useful and, for instance, recommending alternative views rather than aiming to directly and actively steer the analysis.

Provide Encouragement – Our study revealed that guidance should be provided “just in time”, as postulated by Collins et al. [14]. While only one participant was stressed by fast-paced guidance (U6B), others reported that they did not appreciate guidance while they were still reading task descriptions and orienting themselves (U3A). On the other side, guidance lost a lot of its impact when provided too late. However, several participants stated that late guidance still had a value in “*encouraging*” (U4B) them, confirming their analysis (U4A) or increasing their confidence in the achieved results (U2B). Previous work has already suggested that guidance systems should adapt the timing of their suggestions [12], [14], [35]. If such adaptation is not feasible, we encourage researchers to investigate the effects of confirmatory guidance.

Contextualize Guidance Suggestions – Our qualitative analysis highlighted the users’ episodic perception of guidance. Throughout the study, many users who generally appreciated guidance rejected suggestions that did not match their mental model in specific situations. In contrast, not even the most critical participants rejected all suggestions. We note that this behavior was not primarily dependent on the quality of the provided guidance but rather on the context in which it was provided. Analyzing the facial expression of the user during the analysis to determine the need for guidance [10] is an example of how the analysis context can be used to provide appropriate suggestions.

B. Study Limitations

Despite the training the wizards received before the study, they likely neither memorized all solutions to all tasks nor gained a comprehensive understanding of the analyzed data. Hence, the study setting could also be considered a pair analytics scenario in which users and wizards work together to solve the task. We believe that this is characteristic of most (non-trivial) guidance interactions between users and systems. Interactions between both actors are typically not verbalized (yet), and collaboration is performed only by observing each other’s interpretations of and reactions to previous actions. As a result, both users and the system need to learn and adapt to come to a solution. On the other end of the spectrum, one might argue that training wizards on concrete user tasks created an unrealistic environment in which wizards knew all the answers. However, we think that this is not an issue: In normal VA sessions, systems will never have a full understanding of what the best solution is; if they did, VA would not be necessary. However, to provide meaningful guidance, systems need to make assumptions and act accordingly. This is exactly what wizards in our study did as well.

Our study environment did not allow users to explicitly ask for guidance. Instead, we relied on wizards to detect when help was needed. In some situations, that led to longer pauses between suggestions than users might have expected. A future study could add such a feature to allow users to explicitly request guidance. However, several wizards already

reported being limited by the high mental load and user speed. Despite being trained on specific tasks, some wizards wished for additional training time. Further, after some time, individual wizards doubted that they were interacting with a trained bot and started to empathize with struggling users, leading to emotional stress.

Our study design is unbalanced, as misguidance always appears in the second set of tasks. However, if users received misguidance at the beginning of the analysis, they might have lost all confidence in the system and ignored all the following guidance. It is likely that the experiences from the first task set influenced participants’ expectations in the second task set. Nonetheless, we chose this within-subjects design to observe how wizards would react to different users and how users would react to changing guidance strategies.

As our quantitative analysis shows, there was a high variance in perceived task difficulty and guidance quality. Some of this variance could potentially be explained by better or worse wizards making more or less accurate suggestions. Unfortunately, one of the drawbacks of our study design decision to give wizards much freedom in how to guide makes it difficult to analyze such distinctions now. We do not feel confident marking individual suggestions as wrong or right consistently as multiple strategies and suggestion orderings led to the correct solution. Hence, the correctness of suggestions is difficult to classify as a binary variable. Here, future studies with more constraint wizard interactions or different task setups are needed to analyze and better understand this phenomenon.

VIII. CONCLUSION

In this paper, we investigated guidance interaction dynamics that arise from different guidance strategies and user expectations. In a double-blind WOz study, we collected interactions from 24 participants with previous VA experience. Participants voiced a set of diverse expectations towards guidance and developed various guidance strategies when in the role of the wizard. With this diversity, our study highlights once more the importance of contextualized and co-adaptive guidance. From our analysis, we identified three distinct user behaviors and four different guidance strategies that were developed by wizards on the fly. We analyzed interaction dynamics in interesting user-strategy pairings and summarized takeaways for the design, implementation, and evaluation of future co-adaptive guidance systems.

REFERENCES

- [1] Github - fabiansperrle/wizard-of-oz-guidance. <https://github.com/FabianSperrle/Wizard-of-Oz-Guidance>. Accessed: 2022-11-30.
- [2] Guidance strategies and interaction dynamics: A wizard of oz study. <https://guidance-wizards.dbvis.de>. Accessed: 2022-11-30.
- [3] Guidance strategies and interaction dynamics: A wizard of oz study. <https://openweathermap.com/api>. Accessed: 2022-11-30.
- [4] J. Aurisano, A. Kumar, A. Alsaiani, B. D. Eugenio, and A. Johnson. Many At Once: Capturing Intentions to Create And Use Many Views At Once In Large Display Environments. *Computer Graphics Forum*, 39(3):229–240, 2020. doi: 10.1111/cgf.13976
- [5] N. Boukhelifa, A. Bezerianos, I. C. Trelea, N. M. Perrot, and E. Lutton. An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts. In *Proc. CHI Conference on Human Factors in Computing Systems*, pp. 1–14, May 2019. doi: 10.1145/3290605.3300874

- [6] N. Boukhelifa, E. Lutton, and A. Bezerianos. A Case Study of Using Analytic Provenance to Reconstruct User Trust in a Guided Visual Analytics System. *Workshop on TRust and EXpertise in Visual Analytics*, pp. 45–51, 2021. doi: 10.1109/TREX53765.2021.00013
- [7] M. Brehmer, T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, pp. 2376–2385, Dec. 2013. doi: 10.1109/TVCG.2013.124
- [8] M. Brehmer, S. Carpendale, B. Lee, and M. Tory. Pre-design empiricism for information visualization: scenarios, methods, and challenges. In *Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 147–151, Nov. 2014. doi: 10.1145/2669557.2669564
- [9] A. Bäuerle, H. Neumann, and T. Ropinski. Classifier-Guided Visual Correction of Noisy Labels for Image Classification Tasks. *Computer Graphics Forum*, 39(3):195–205, 2020. doi: 10.1111/cgf.13973
- [10] D. Ceneda, A. Arleo, T. Gschwandtner, and S. Miksch. Show Me Your Face: Towards an Automated Method to Provide Timely Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, July 2021. doi: 10.1109/TVCG.2021.3094870
- [11] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):111–120, 2017. doi: 10.1109/TVCG.2016.2598468
- [12] D. Ceneda, T. Gschwandtner, and S. Miksch. A Review of Guidance Approaches in Visual Data Analysis: A Multifocal Perspective. *Computer Graphics Forum*, 38(3):861–879, 2019. doi: 10.1111/cgf.13730
- [13] D. Ceneda, T. Gschwandtner, and S. Miksch. You get by with a little help: The effects of variable guidance degrees on performance and mental state. *Visual Informatics*, 3(4):177–191, Dec. 2019. doi: 10.1016/j.visinf.2019.10.005
- [14] C. Collins, N. Andrienko, T. Schreck, J. Yang, J. Choo, U. Engelke, A. Jena, and T. Dwyer. Guidance in the human-machine analytics process. *Visual Informatics*, 2(3):166–180, 2018. doi: 10.1016/j.visinf.2018.09.003
- [15] A. Coscia, D. H. Chau, and A. Endert. Toward a Bias-Aware Future for Mixed-Initiative Visual Analytics. In *Workshop on TRust and EXpertise in Visual Analytics*, pp. 1–3, Oct. 2020.
- [16] N. M. Fraser and G. N. Gilbert. Simulating speech systems. *Computer Speech & Language*, 5(1):81–99, Jan. 1991. doi: 10.1016/0885-2308(91)90019-M
- [17] L. Grammel, M. Tory, and M.-A. Storey. How Information Visualization Novices Construct Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):943–952, Nov. 2010. doi: 10.1109/TVCG.2010.164
- [18] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 159–166, 1999. doi: 10.1145/302979.303030
- [19] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 3363–3372. Association for Computing Machinery, New York, NY, USA, May 2011. doi: 10.1145/1978942.1979444
- [20] W. B. Knox, S. Spaulding, and C. Breazeal. Learning Social Interaction from the Wizard: A Proposal. In *AAAI Workshop on Machine Learning for Interactive Systems and Robots*, p. 6, 2014.
- [21] B. Kondo and C. Collins. DimpVis: Exploring Time-varying Information Visualizations by Direct Manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2003–2012, Dec. 2014. doi: 10.1109/TVCG.2014.2346250
- [22] S. K. Kuttal, B. Ong, K. Kwasny, and P. Robe. Trade-offs for Substituting a Human with an Agent in a Pair Programming Context: The Good, the Bad, and the Ugly. In *Proc. Conf. on Human Factors in Computing Systems*, pp. 1–20, 2021. doi: 10.1145/3411764.3445659
- [23] D. J.-L. Lee, V. Setlur, M. Tory, K. Karahalios, and A. Parameswaran. Deconstructing Categorization in Visualization Recommendation: A Taxonomy and Comparative Study. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4225–4239, Dec. 2022. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2021.3085751
- [24] M. Oppermann, R. Kincaid, and T. Munzner. VizCommender: Computing Text-Based Similarity in Visualization Repositories for Content-Based Recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):495–505, Feb. 2021. doi: 10.1109/TVCG.2020.3030387
- [25] R. Oppermann, ed. *Adaptive user support: ergonomic design of manually and automatically adaptable software*. L. Erlbaum Associates Inc., 1994.
- [26] M. Porcheron, J. E. Fischer, and S. Reeves. Pulling Back the Curtain on the Wizards of Oz. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):243:1–243:22, Jan. 2021. doi: 10.1145/3432942
- [27] I. Pérez-Messina, D. Ceneda, M. El-Assady, S. Miksch, and F. Sperrle. A Typology of Guidance Tasks in Mixed-Initiative Visual Analytics Environments. *Computer Graphics Forum*, 2022. doi: 10.1111/cgf.14555
- [28] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, and D. A. Keim. SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):120–130, Jan. 2018. doi: 10.1109/TVCG.2017.2744805
- [29] J. Scholtz, O. Love, M. Whiting, D. Hodges, L. Emanuel, and D. S. Fraser. Utility evaluation of models. In *Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 160–167, Nov. 2014. doi: 10.1145/2669557.2669562
- [30] P. Sequeira, P. Alves-Oliveira, T. Ribeiro, E. Di Tullio, S. Petisca, F. S. Melo, G. Castellano, and A. Paiva. Discovering social interaction strategies for robots from restricted-perception Wizard-of-Oz studies. In *ACM/IEEE Int. Conf. on Human-Robot Interaction*, pp. 197–204, Mar. 2016. doi: 10.1109/HRI.2016.7451752
- [31] R. Sheline and C. J. MacLellan. Investigating Machine-Learning Interaction with Wizard-of-Oz Experiments. In *Learning by Instruction Workshop (NIPS 2018)*, pp. 1–8, 2018.
- [32] S. Smith and J. Mosier. Guidelines for designing user interface software. Technical Report ESD-TR-86-278, Mitre Corporation, Bedford MA, 1986.
- [33] F. Sperrle, D. Ceneda, and M. El-Assady. Lotse: A Practical Framework for Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2022. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2022.3209393
- [34] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. *Computer Graphics Forum*, 40(3):543–568, 2021. doi: 10.1111/cgf.14329
- [35] F. Sperrle, A. Jeitler, J. Bernard, D. Keim, and M. El-Assady. Learning and Teaching in Co-Adaptive Guidance for Mixed-Initiative Visual Analytics. *EuroVis Workshop on Visual Analytics*, pp. 1–5, 2020. doi: 10.2312/eurova.20201088
- [36] F. Sperrle, A. Jeitler, J. Bernard, D. Keim, and M. El-Assady. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics*, 100:93–105, Nov. 2021. doi: 10.1016/j.cag.2021.06.016
- [37] A. Steinfeld, O. C. Jenkins, and B. Scassellati. The oz of wizard: simulating the human for interaction research. In *Proc. ACM/IEEE Int. Conf. on Human robot interaction*, HRI '09, pp. 101–108, 2009. doi: 10.1145/1514095.1514115
- [38] M. Tory and V. Setlur. 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, Oct. 2019. doi: 10.1109/VAST47406.2019.8986918
- [39] E. R. Tufte. The Visual Display of Quantitative Information. *The Journal for Healthcare Quality (JHQ)*, 7(3):15, July 1985.
- [40] A. Walch, M. Schwärzler, C. Luksch, E. Eisemann, and T. Gschwandtner. LightGuide: Guiding Interactive Lighting Design using Suggestions, Provenance, and Quality Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):569–578, Jan. 2020. doi: 10.1109/TVCG.2019.2934658
- [41] Z. Wen and M. Zhou. Evaluating the Use of Data Transformation for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1309–1316, Nov. 2008. doi: 10.1109/TVCG.2008.129
- [42] J. Xiao, R. Catrambone, and J. T. Stasko. Be Quiet? Evaluating Proactive and Reactive User Interface Assistants. Technical Report, Georgia Institute of Technology, 2003.
- [43] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovich. Survey on the Analysis of User Interactions and Visualization Provenance. *Computer Graphics Forum*, 39(3):757–783, 2020. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14035>. doi: 10.1111/cgf.14035
- [44] Q. Yang, J. Cranshaw, S. Amershi, S. T. Iqbal, and J. Teevan. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proc. Conf. on Human Factors in Computing Systems*, pp. 1–12, 2019. doi: 10.1145/3290605.3300415
- [45] Y. Ye, F. Sauer, K.-L. Ma, K. Aditya, and J. Chen. A User-Centered Design Study in Scientific Visualization Targeting Domain Experts. *IEEE Transactions on Visualization and Computer Graphics*, 26(6):2192–2203, June 2020. doi: 10.1109/TVCG.2020.2970525

IX. BIOGRAPHIES

Fabian Sperrle is a PhD student in the Data Analysis and Visualization group at the University of Konstanz. He received his M.Sc. in Computer and Information Science from the University of Konstanz in 2018. His research interests include co-adaptive analysis processes in general and guidance in particular.

Mennatallah El-Assady is an Assistant Professor at the Department of Computer Science, ETH Zürich. She heads the Interactive Visualization and Intelligence Augmentation (IVIA) lab. Prior to that, she was a research fellow at the ETH AI Center; and before that, she was a research associate in the group for Data Analysis and Visualization at the University of Konstanz (Germany) and in the Visualization for Information Analysis lab at the OntarioTech University (Canada). Her main research interest is studying interactive human-AI collaboration interfaces for effective problem-solving and decision-making.

Alessio Arleo is an Assistant Professor at Eindhoven University of Technology. Previously, he worked as a Post-Doc in the Visual Analytics research unit (CFAST) at the Institute of Visual Computing and Human-Centered Technology, TU Wien. He obtained a PhD in 2018 at the University of Perugia, Italy (thesis: Distributed Large Graph Visualization: Algorithms and Experiments). His research interests include network visualization, distributed algorithms, and information diffusion.

Davide Ceneda received the PhD degree from TU Wien, Austria, in 2020, defending a thesis called “Guidance-Enriched Visual Analytics.” He is a post-doc with the CFAST group at the Institute of Visual Computing and Human-Centered Technology, TU Wien. His research interests include visualization, human perception, visual analytics, and guidance.