Guidance Source Matters: How Guidance from AI, Expert, or a Group of Analysts Impacts Visual Data Preparation and Analysis

Arpit Narechania
Georgia Institute of Technology
Atlanta, Georgia, USA
arpitnarechania@gatech.edu

Alex Endert
Georgia Institute of Technology
Atlanta, Georgia, USA
endert@gatech.edu

Atanu R Sinha
Adobe Research
Bengaluru, Karnataka, India
atr@adobe.com

ABSTRACT

The progress in generative Artificial Intelligence (AI) has fueled AI-powered tools like co-pilots and assistants to provision better guidance, particularly during data analysis. However, research on guidance has not yet examined the perceived efficacy of the source from which guidance is offered and the impact of this source on the user's perception and usage of guidance. We ask whether users perceive all guidance sources as equal, with particular interest in three sources: (i) "AI," (ii) "human expert," and (iii) "a group of human analysts." As a benchmark, we consider a fourth source, (iv) "unattributed guidance," where guidance is provided without attribution to any source, enabling isolation of and comparison with the effects of source-specific guidance. We design a five-condition between-subjects study, with one condition for each of the four guidance sources and an additional (v) "no-guidance" condition, which serves as a baseline to evaluate the influence of any kind of guidance. We situate our study in a custom data preparation and analysis tool wherein we task users to select relevant attributes from an unfamiliar dataset to inform a business report. Depending on the assigned condition, users can request guidance, which the system then provides in the form of attribute suggestions. To ensure internal validity, we control for the quality of guidance across source-conditions. Through several metrics of usage and perception, we statistically test five *preregistered* hypotheses and report on additional analysis. We find that the source of guidance matters to users, but not in a manner that matches received wisdom. For instance, users utilize guidance differently at various stages of analysis, including expressing varying levels of regret, despite receiving guidance of similar quality. Notably, users in the AI condition reported both higher post-task benefit and regret. These findings strongly indicate the need to further understand how different guidance sources impact user behavior for designing effective guidance systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in visualization**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

guidance, artificial intelligence, human expert, groupthink, data preparation, visual data analysis

IUI '25, March 24–27, 2025, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03.
https://doi.org/10.1145/3708359.3712166

ACM Reference Format:
Arpit Narechania, Alex Endert, and Atanu R Sinha. 2025. Guidance Source Matters: How Guidance from AI, Expert, or a Group of Analysts Impacts Visual Data Preparation and Analysis. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3708359.3712166

1 INTRODUCTION

The prospect of offering guidance to users of a system has received fillip over the past two years as talks of Artificial Intelligence (AI)-powered tools like co-pilots and assistants caught imaginations of users and businesses alike. Hitherto, the guidance literature has paid attention to the important dimensions of "why," "how," "what," and "when" in guided interactions [29, 79]–contributing theories, models, frameworks, and tools on what guidance to provide, why, how, and when. In this work, we investigate another dimension–"from whom"–focusing on the source of guidance–such as humans or an AI–which remains a gap in research. Today, guidance is already being sought from a variety of sources across various domains. For instance, in the analytics domain, guidance is sought from human experts (e.g., an expert analyst or consultant) or groups of peers (e.g., via community forums such as Stack Overflow[1], HuggingFace[2], GitHub[3]). Relevant research on advice-taking has already studied and compared such guidance from human experts and peers [57, 58, 65, 68, 69, 98]. Recently, there has been a growing expectation for guidance to come from an AI [22, 62, 95], even though this guidance must itself rely on data from human experts or groups of peers, or other systems, to train the models and align subsequent recommendations with user preferences. Thus, studying this "from whom" dimension is important because the effect of *source-attribution* in providing guidance carries significant implications for offering guidance systems.

In response, in this work, we focus on how users' perception and utility of guidance coming from a particular source impacts their performance during attribute selection as part of a visual data preparation task. In particular, we study three guidance sources–**AI**, human **Expert**, and a **Group** of human analysts–to answer our research question, *"Does the source of guidance matter to users even when the quality of guidance is held constant?"* As a benchmark, we also consider a fourth source, **Unattributed**, wherein guidance is provided without attribution to any source, enabling isolation of and comparison with the effects of source-specific guidance. We design a five-condition between-subjects study, with one condition for each of the four guidance sources and an additional no-guidance condition, **Control**, which serves as a baseline to evaluate the influence of any

[1]https://stackoverflow.com
[2]https://huggingface.co
[3]https://github.com

kind of guidance. We *preregister* five hypotheses of which three hypotheses compare the source-agnostic **Unattributed** guidance with **Control** (no-guidance), giving a direct effect of guidance, unconfounded by the effect of any specific source. The remaining two hypotheses are source-specific, comparing guidance from **AI** with **Expert** and **Group**, since, as argued above, these are natural candidates from whom guidance can be provisioned. Moreover, comparing each attributed source with **Unattributed** isolates the impact of the attributed source.

To validate these hypotheses, we built a custom data preparation and analysis tool (as our study prototype) and tasked users to select relevant attributes from an unfamiliar dataset for a business report. Depending on the assigned condition, users can optionally request guidance, one by one, up to a maximum of ten times, which the system then provides in the form of attribute suggestions (hereafter, *guided attributes*). These guided attributes are a subset of all attributes in the dataset that the user can select from. To ensure internal validity, we control for the quality of guidance by always provisioning seven relevant attributes and three irrelevant attributes, predetermined by the study team, randomly picked and sequenced, while varying the source, to get a valid measure of the effect of source.

Through several metrics of usage and perception, we statistically test each of our five *preregistered* hypotheses. For instance, one of the aspects we focus on is *utilization*, which entails user's decision to accept or reject attributes provisioned as guidance. Examination of utilization (i) is a less attended but important part of guidance and (ii) draws relevant concepts from the extant judgment and decision making (JDM) literature [67]–across psychology and economics–into guidance. Our emphasis on utilization differs from questions of trust in guidance, especially when coming from AI, which have received considerable attention in research [7, 27, 31, 44, 56, 61]. We recognize that trust plays an important role. However, the research investigating trust depends largely on stated preference (*what is said*) by users, which may not match with their revealed preference (*what is done*) [30]. Instead, utilization is about what users do and gives a more objective measure of effect, while also complementing findings about trust–hence our focus on it.

Another aspect, *regret*, is germane to JDM [8, 63, 97, 99] although largely ignored in guidance. *"[R]egret arises from comparing an obtained outcome with a better outcome that might have occurred had a different choice been made; that is, regret stems from bad decisions"* [99, pp. 222]. Moreover, regret is also a function of users' expectations, formed from multiple reference points [60]. The importance of regret for guidance systems is threefold: (i) any post-task regret may impact continued utilization of guidance; (ii) anticipatory regret may distort use of guidance [93]; and (iii) managing expectations of users may be necessary for success. As part of our study, we measure post-task regret and examine its relation to other guidance characteristics and to users' expectations from guidance.

Our study findings reveal that: (i) guidance benefits users *during* analysis, although differentially across sources, supporting our thesis; (ii) utilization of guidance varies across stages of analysis–early on, throughout, or later during analysis; (iii) users verify guidance differently across sources; (iv) pre-task *perception* of guidance is adversely affected when attributed to source, although the *decrease* in scores post-task, favors attributed-guidance, pointing to subtleties of source-attribution; (v) users' higher post-task regret by relying on guidance from **AI** complements the findings about the lower ex-ante trust from **AI**. At the same time, users find higher benefit of increased confidence from **AI**'s guidance, calling for a nuanced interpretation of **AI**'s role. These findings strongly indicate the need to further understand how different guidance sources impact user behavior for designing effective guidance systems.

Our primary contributions include:

(1) Highlighting the importance of *source-attribution* in guidance (via the newly introduced "from whom" dimension).
(2) A controlled study with head to head comparison of *three* widely studied sources of guidance: **AI**, **Expert**, **Group**.
(3) Conceptualizing guidance from a *utilization* point of view, offering objective metrics of measurement (uncertainty, verification, utility) that can be adopted in research and practice.
(4) Conceptualizing guidance-taking as a *decision making problem* and introducing and analyzing constructs of regret and expectations to show new ways to study guidance.
(5) Offering a nuanced perspective about guidance from **AI**, as users can show both higher post-task benefit and higher post-task regret.

2 RELATED WORK

2.1 Visual Analytics and Guidance

Visual Analytics (VA) is a human-in-the-loop approach that combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning, and decision-making based on large, complex datasets [54]. VA systems incorporate concepts from mixed-initiative systems to *"enable users and intelligent agents to collaborate efficiently"* [49] by taking initiatives on behalf of each other during analysis. More recently, VA systems have embraced a 'human *is* the loop' perspective–which emphasizes the central role of the user–by enabling the system to implicitly infer their workflow(s), and seamlessly integrating analytics into it [28]. However, automated actions by the system can be mistimed or a result of misinterpreted user intent; whereas, users may need to provide feedback on these automated actions or configure (feedforward) their intent to the system upfront, requiring a continuous, effective dialogue between the two to ensure smooth and effective analytic progress. Guidance is one such computer-assisted process that aims to actively resolve this "knowledge gap" between the user's understanding and the system's capabilities during an interactive analysis session [17, 18, 23]. In addition, guidance also aims to ensure effective system operation [82], enhance usability [25], improve analysis efficiency, validate insights, build confidence, prevent bias, and improve clarity of findings [23].

There have been several approaches to conceptualize and apply guidance in visual analytics to improve the quality of interactions between users and systems. Engels [29] characterized guidance into a "what" dimension (that defines the problem) and a "how" dimension (that defines mechanisms to solve the problem). Pérez-Messina et al. [79] proposed a typology of guidance tasks covering the "why," "how," "what," and "when" aspects of guided interactions. Ceneda et al. [17, 18] characterized guidance based on the user's knowledge gap, the input and output of the guidance process, and its degree, later formalizing a methodology for designing effective guidance systems [16]. Sperrle et al. [85–87] introduced the concept of co-adaptive guidance wherein the user and the system teach and

learn from one another during visual data analysis, later contributing a practical framework for developers to build custom guidance strategies [83]. In this work, we explore an underexplored dimension of guidance, "from whom", focusing on the source of guidance, specifically an **AI**, a human **Expert**, or a **Group** of analysts, and how it impacts a user's perception and usage of guidance.

2.2 Studies on Utilization of and Reliance on Guidance

Understanding how guidance is utilized or relied on is critical for designing effective guidance systems, and has also been a key focus of many prior studies. For instance, Wall et al. [96], Narechania et al. [73], and Paden et al. [77] studied how presenting visual traces of a user's interaction history during analysis can help increase their awareness of analytic behavior and mitigate exploration biases. Sperrle et al. [84] conducted a Wizard of Oz study to investigate the interaction dynamics between users and systems in co-adaptive guidance scenarios, focusing on the impact of guidance timing, contextualization, and adaptation, as well as the effects of misguidance on user confidence. Sperrle et al. [88] also studied how context-dependent user preferences and feedback during topic model refinement can enable the system (not the user) to learn and adapt its subsequent guidance, fostering effective co-adaptive guidance and human-machine collaboration.

In terms of how people utilize and rely on guidance recommendations originating from different sources, we found a complex interplay between human and algorithmic judgments. Some studies found that users trust and rely on human partners more than AI [7, 27, 31, 44, 56, 61], whereas some others found the opposite [38, 62, 64]. For example, Logg et al. [62] found people often trust algorithms more than human expertise, despite not fully understanding the algorithm's intricacies. Several studies found that people's reliance on AI depends on various contextual factors, such as their AI literacy [50], domain expertise [39], and amount of feedback [38]. For example, Gajos et al. [38] found that people make more accurate decisions by actively engaging with detailed explanations of AI recommendations–rather than just viewing the recommendations. Among human partners, studies have revealed differences between guidance from experts versus groups. Chen et al. [20] show that online book purchasing decisions are heavily influenced by consumer recommendations rather than expert opinions. Similarly, Vedadi et al. [94] find that in information security decisions, users tend to imitate others when faced with uncertainty, impacting their choices more than their personal assessments. This limitation effect extends to software adoption, where user reviews influence lower-ranked products' adoption but not top ones [26].

In our study, we examine the utilization and perception of guidance from an **AI**, a human **Expert**, or a **Group** of analysts as source, which represents what users *do*, complementing prior work that investigates trust and reliance, which represent what users *say*. Our study also collects typed textual responses about trust and reliance to make the set of measures comprehensive in having both objective measures of utilization and subjective metrics of perception. We next describe relevant metrics available in prior work that quantify users' utilization of and reliance on guidance.

2.3 Metrics to Model Utilization of and Reliance on Guidance

Several metrics have been proposed that measure people's reliance on AI guidance, quantifying people's "agreement" and "disagreement" with AI recommendations [64], people's "acceptance" of incorrect AI recommendations [14], people's "change" in behavior based on AI recommendations [55, 62], and people's propensity to "delegate" eventual decision-making to AI [21]. For instance, Lu et al. [64] proposed the "agreement" and "disagreement" metrics, which assess how often user predictions are the same as, or different from, AI recommendations, respectively, when users make predictions before seeing AI recommendations. Buccinca et al. [14] measured how often users accept incorrect AI recommendations and how often users make mistakes when their predictions differ from the AI's, with both the user's prediction and the AI's recommendation being wrong. Kim et al. [55] proposed "Switch Fraction" to assess how often users completely change their answers to match AI recommendations [55]. Logg et al. [62] proposed "Weight of Advice" (WOA) to measure the proportional change in user predictions relative to the change in AI recommendations. Chiang et al. [21] proposed the "delegation" metric to measure how often users let an AI system fully make decisions on their behalf. These metrics provide a framework for understanding user interactions with guidance systems. In our study, we task participants to select relevant attributes from unfamiliar datasets, but there is no known ground truth in terms of number of attributes to select. Thus, we created new metrics based on these existing ones that better fit our study design.

2.4 Data Preparation and Subset Selection in Visual Analytics

Data preparation involves analyzing the data to ensure high-quality results through collection, integration, transformation, cleaning, reduction, and discretization [100]. As organizations follow a "load-first" philosophy and "dump" their data into centralized repositories [40], the volume of data often overwhelms users, creating challenges in data navigation, discovery, and monitoring [24, 33, 75, 76]. To mitigate these challenges, prior work has utilized several techniques based on the raw data [24, 70], meta-data [3, 46], and users' queries [11, 15, 101]. For example, Goods infers metadata from billions of datasets within an organization, making them searchable using keywords [46]. Similarly, there exist several proprietary [1, 2, 4] and open-source [2, 5] tools that provide data profile, quality, and lineage information for data observability, monitoring, and pipeline optimization. For example, Profiler [53] uses data mining to automatically detect quality issues in tabular data and offers coordinated visualizations for context; whereas, Tableau Prep [90], OpenRefine [47], and Wrangler [52] provide interactive affordances to explore, clean, structure, and shape the data before analysis.

Our study focuses on subset selection [59], which can be achieved through "feature set reduction" to decrease the number of attributes or "sample set reduction" to reduce the number of records. Feature set reduction is often used in machine learning to eliminate irrelevant features [51] or perform dimensionality reduction [36]; whereas sample set reduction is commonly applied in market segmentation [92] to identify specific consumer groups. Several existing tools facilitate this process of subset selection. For example, DataPilot [74] presents

quality and usage information to assist users in selecting effective subsets from large, unfamiliar tabular datasets; DataCockpit [72] extends these capabilities to multiple relational databases via an open-source Python toolkit. SmartStripes [66] uses automated filter algorithms and statistical correlation measures along with interactive visualizations. Notably, visualizing how a selected subset compares to the original dataset has been shown to mitigate selection biases [9, 43]. The task in our study is similar to DataPilot's [74], described next.

3 TASK: DATA PREPARATION AND SUBSET SELECTION FOR VISUAL ANALYTICS

In this section, we describe the source conditions of our study, the study prototype, the study task, and how we ensure internal validity.

3.1 Source (Study) Conditions

We designed a between-subjects study wherein participants were randomly assigned to one of five experimental conditions. In each condition, guidance was presented as if it came from an attributed-source or from an unattributed-source, holding the quality of guidance same across all guidance sources. The five experimental conditions along with their definitions are as follows:

- **AI** Guidance comes from an AI model trained on large information for data analysis tasks.
- **Expert** Guidance comes from a human expert analyst who is well regarded in industry for acumen in data analysis.
- **Group** Guidance comes from a group of human analysts in your organization well versed in data analysis.
- **Unattributed** Guidance comes without an explicit mention of a source.
- **Control** No guidance is available in the interface.

3.2 Study Prototype

With no existing system that satisfies our goals of experimental control and quality control of guidance, we developed one wherein users can inspect a dataset, select relevant attributes, request guidance, and create visualizations. We used Angular [41] as our frontend framework and interfaced it with a Python [37] backend over the REST [81] and websocket [34] protocols. We persisted logged user interactions on the cloud via Google Cloud Logging [42]. Figure 1 and Figure 2 show the UI and Table 1 lists the interactions that are tracked in the UI, described next.

UI Tab 1: Explore and Analyze

(A) **Data Attributes** shows the list of dataset attributes along with their definitions and (de)select affordances.
(B) **Data Records** shows the entire dataset in an interactive table along with sorting and pagination affordances.
(C) **Guidance** shows one attribute (up to 10) each time the user requests guidance via a button.

In this tab, users can **Search** attributes by keywords or **Sort** attributes (by their name) to organize their search space. For each visible data attribute, users can **Inspect** it to see a brief description, **Click** on it to see detailed records, **Select** it to be in their subset, **Deselect** it from their subset, and **Request Guidance** (except for **Control** participants who do not receive any guidance). For the data records, users can **Paginate** or **Sort** the interactive table.

UI Tab 2: Create Dashboard

(D) **Selected Data Attributes**, similar to (A) **Data Attributes**, shows the list of attributes *selected* by the user.
(E) **Mark and Encodings** shows affordances to create visualizations: a visualization mark type (bar, point, line), visual encodings (x, y, fill color, size, shape), and encoding aggregations (sum, mean, max, min).
(F) **Visualization** shows the visualization based on the **Mark and Encodings** along with an affordance to save it.
(G) **Saved Visualizations** shows the list of all saved **Visualization**s, including affordances to delete one or all.

In this tab, users can **Inspect** and **Deselect** selected attributes from the Selected Data Attributes view. Additionally, users can **Create** visualizations by changing the mark type, changing and resetting visual encodings and aggregations, and swapping the attributes mapped to xy axes. Lastly, users can **Save** visualizations or **Delete** one or all saved visualizations.

3.3 Main (Study) Task

Using the study prototype, we tasked participants to perform the following analysis task:

> *Analyze the provided dataset of attributes derived from online customer behaviors. Design multiple visualizations indicating meaningful drivers of dollar ($) sales revenue for the company.*
>
> *Select between 5 to 15 attributes (both inclusive); and make at 3 visualizations. It is important for you to pay attention to the attributes you select and the visualizations you create from them.*
>
> *Note that these attributes and visualizations will be examined by your boss, who will then decide whether to use them for their own report. Spend around 20 minutes for this task.*

For the conditions that received guidance (**AI**, **Expert**, **Group**, **Unattributed**) there were additional instructions:

> *To help you perform this complex task, guidance is on the way. This guidance may be helpful to perform the task well. The guidance will appear on the right-hand side panel on the UI. Your judgment is important in accepting or in rejecting any guidance you receive. It is your decision whether and which guidance to use. You will receive guidance:*
>
> - (For **AI**) from an AI model trained on large information for data analysis tasks.
> - (For **Expert**) from a highly regarded industry expert in data analysis.
> - (For **Group**) from a group of data analysts in your organization well versed in data analysis.
> - (For **Unattributed**) that may be helpful to perform the task well.

We note that our primary focus is on participants' selection of relevant data attributes, aligned with our study goals and hypotheses. Our analyses, presented later, emanate from this primary focus. To

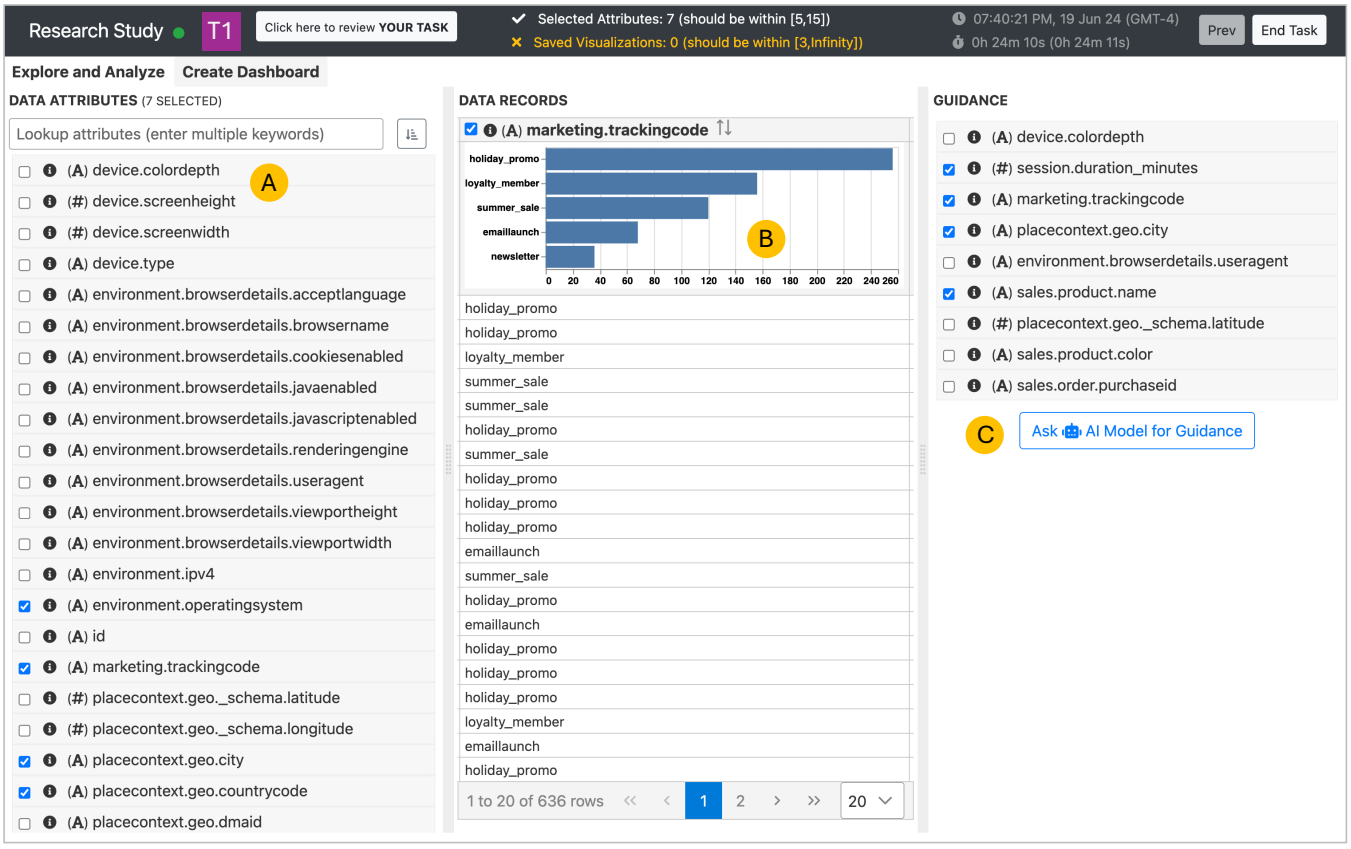


Figure 1: Tab 1 of the study prototype to help users analyze a dataset and select attributes for subsequent use in Tab 2.

motivate participants' decision making toward selection of attributes, their task also involves creating and saving visualizations using their selected attributes. The visualization outcomes are not meant to be evaluated since those have inherently subjective, "in-the-eyes-of-the-beholder" nature, making it hard to meaningfully evaluate as task performance. As well, those do not directly address our hypotheses.

3.4 Internal Validity

Concerned with the internal validity of the study, we hold the quality of guidance constant across the four guidance conditions: **Unattributed**, **AI**, **Expert**, and **Group**. Of the total 33 attributes in the dataset that participants can select, we carefully identify 14 attributes that are important and relevant to our study task. From these 14 relevant attributes, we select 7, and then from the remaining 19 irrelevant attributes, we select 3, to create a list of 10 total attributes. This 7:3 ratio of relevant to irrelevant attributes was intentionally chosen to encourage participants to exercise judgment when using the provided guided attributes, rather than using them indiscriminately (*carte blanche*). Our study description emphasizes the importance of this judgment in deciding among guided attributes. Finally, we randomize the order in which these 10 attributes are recommended in the user interface, to enforce experimental validity by avoiding an order effect. By using the same proportion of relevant and irrelevant attributes that are guided, we achieve internal validity. However, to the extent the quality of guidance can vary with source, we lose some external validity. To achieve both internal and external validity requires a much larger study, which is beyond the scope of this paper.

4 HYPOTHESES

A primary goal of guidance is to minimize the "knowledge-gap" between the user and the system to enhance analytic workflows [17, 23]. Received wisdom suggests that the framework of guidance for visual analytics is commonly conceptualized agnostic of the source of guidance [16, 19]. In guidance without source-attribution, the focus for the beneficiary-user is on the guidance provided in the system, devoid of the entity behind the guidance. Thus, we find it useful to establish user behavior with this kind of source-agnostic guidance, termed, **Unattributed** guidance. Comparing users' behaviors in **Unattributed** guidance with that of the condition of no-guidance, or **Control**, allows a direct assessment of the impact of guidance on users' behaviors, un-confounded by any source-attribution (i.e., **AI**, **Expert**, **Group**). This comparison is the basis of the first three hypotheses. Only then, we examine the incremental impact of source-attribution in a set of two additional hypotheses. Below, we discuss the source-agnostic hypotheses followed by the source-specific hypotheses. All five hypotheses were preregistered and can be accessed at https://osf.io/q6ca5.

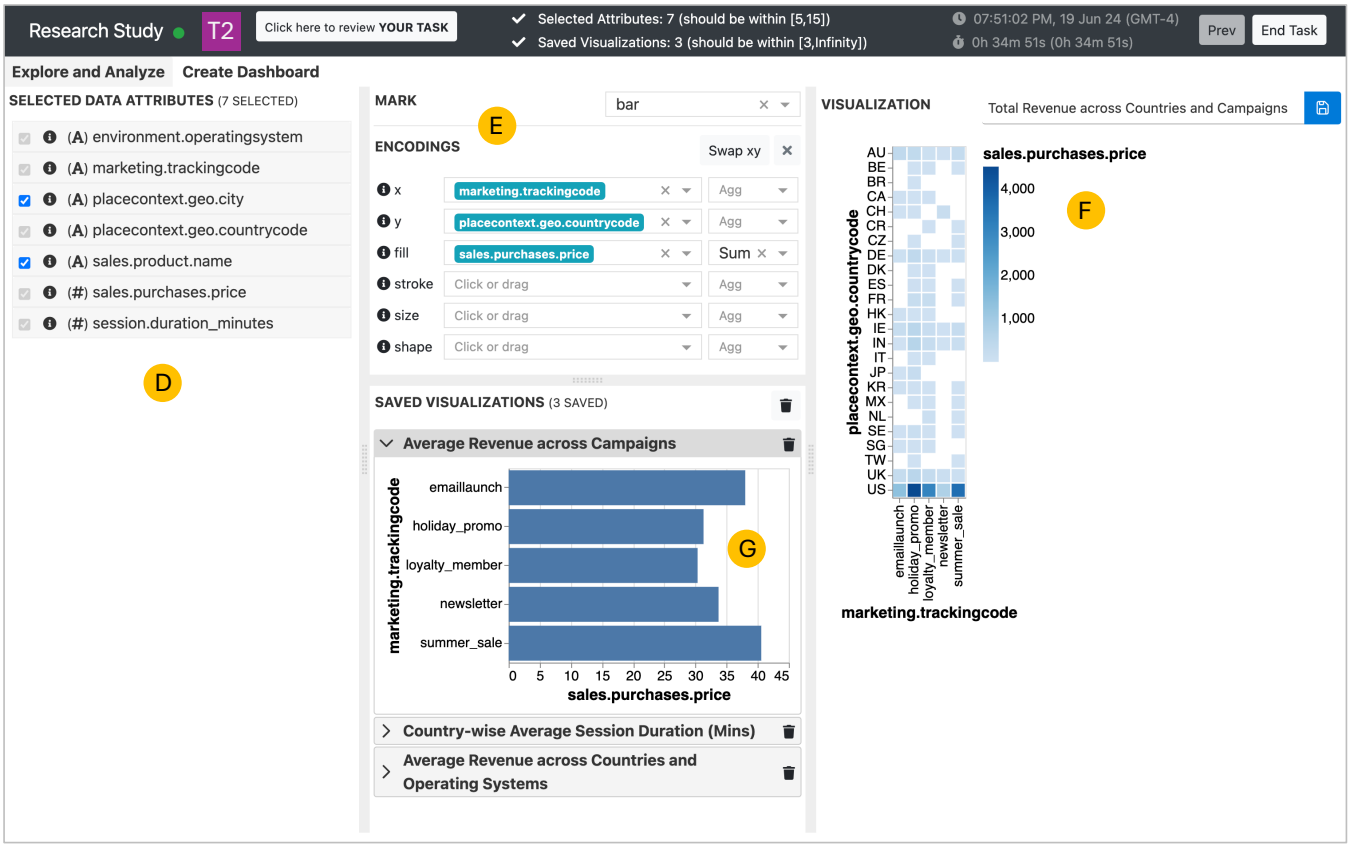


Figure 2: Tab 2 of the study prototype to help users create and save visualizations using the attributes selected from Tab 1.

4.1 Source-Agnostic Hypotheses about Guidance

The three hypotheses in this section compare **Unattributed** guidance to no guidance (**Control**). The premise of guidance is that a user has a knowledge gap with respect to the system being used. Knowledge gap manifests in higher uncertainty the user faces when considering whether to select certain attributes. Our first hypothesis, **H1**, proposes that **Unattributed** guidance can aid in uncertainty reduction relative to the **Control** condition. The reduction in uncertainty under guidance gives higher confidence in selection of attributes, which can result in selecting more number of attributes for the task at hand, especially when those attributes come from guidance. Thus, our second hypothesis, **H2**, proposes that more attributes are likely to be selected under **Unattributed** guidance as compared to **Control**. Also, the reduction in uncertainty makes decision making about attribute selection quicker, giving our third hypothesis, **H3**, that total time for task-completion is likely to be lower in **Unattributed** guidance relative to **Control** condition. We next present objective metrics that form the basis for statistically testing these hypotheses.

H1 Participants who receive guidance will *be less uncertain about their attribute selections*.
Metric(s). We define *uncertainty* as the variance in the number of interactions with attributes. A higher variance indicates greater uncertainty, as it suggests the participant is repeatedly interacting with an attribute, likely due to difficulty in deciding whether to select it or not.

$$\text{Uncertainty} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad (1)$$

where n = total number of attributes in the dataset, x_i = number of interactions with attribute i, and $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ = mean number of interactions across all attributes.

H2 Participants who receive guidance will *select more attributes*.
Metric(s). Attribute selection refers to shortlisting an attribute to include in the final phase of task.

H3 Participants who receive guidance will *take lesser time* to complete the task.
Metric(s). We measure *time* in two ways:
(a) *Duration of the task (in minutes).*
(b) *Total number of interactions performed during the task.*
The duration measures the total time spent performing the task, while the number of interactions is proxy for the amount of engagement with the UI.

4.2 Source-Specific Hypotheses

Among guidance sources, human **Expert** guidance is valued for its credibility and specialized knowledge [32], while an **AI** model offers efficient and objective suggestions [13], and a human **Group** provides

	Interaction Name	Description	View(s)	Cue
1	Search attributes	Filters the list of attributes based on user-input search term.	A	
2	Sort attributes	Toggle sorts the attributes by their name.	A	
3	Inspect attribute	Shows a brief description about the attribute in a tooltip.	ABCD	
4	Click attribute	Shows an overview of the attribute's data distribution and detailed records.	AC	
5	Select attribute	Selects an attribute to be in the desired subset.	ABC	
6	Deselect attribute	Removes a previously selected attribute from the subset.	ABCD	
7	Paginate datatable	Paginates through different pages of the datatable.	B	« »
8	Sort records	Toggle sorts the datatable records by that attribute.	B	
9	Request guidance	Requests an attribute recommendation as guidance (a maximum 10 times).	C	
10	Change tab	Switch between the "Explore and Analyze" and Create Dashboard" tabs.	T1–T2	
11	Change mark type	Changes mark type (e.g. bar, point) of the visualization.	E	
12	Change encoding	Assigns (or clears) an attribute to a visual encoding.	E	
13	Change aggregation	Assigns (or clears) the encoding's aggregation function (e.g. max).	E	
14	Swap xy axes	Swaps the attributes encoded to the x-axis and y-axis.	E	
15	Reset encodings	Revert to the default encoding for visualizations.	E	
16	Save vis	Saves the current visualization.	F	
17	Delete one vis	Remove a specific visualization from the saved list.	G	
18	Delete all vis	Remove all visualizations from the saved list.	G	

Table 1: User interactions tracked in the study prototype including their name (Interaction Name), a short Description and the View(s) (and icon Cues) they occur in (Figures 1, 2). The interactions with orange serial numbers (e.g, 5 Select attribute) can be directly mapped to one or more attributes at any time; an aspect that will be utilized in analysis (Section 6).

diverse perspectives and collective expertise [78]. These findings lead us to propose **H4** about relative utilities from three sources. However, each source has limitations: human **Expert** guidance may be biased or be involved in tension with novices [91], **AI** model may lack contextual understanding and interpretability [80], and human **Group** guidance can suffer from group-think and inconsistency [6]. Synthesizing these results, we call out verification of provided guidance and propose **H5**. Previous studies have shown varied perceptions of guidance based on its source and context, with mixed reports on the usefulness of human versus AI guidance [31, 61, 62] and expert versus group guidance [6, 20]. Moreover, these studies were performed under different platforms ranging from emails to scenario-based visual analytics, and outcomes were measured as stated responses on semantic scales. No single study we know of compares all three sources - **AI**, **Expert**, **Group**, nor presents comparison on a common platform, nor measures utilization, nor uses objective metrics of actions and interactions. Accordingly, we present objective metrics that form the basis for statistically testing these hypotheses, while also supplementing them with subjective responses on semantic scales.

H4 Participants will find guidance to *have more utility* when it comes from Expert > AI > Group (> implies more).

Metric(s). We measure *utility* as the total number of attributes selected at the end of the task.

Drawing from utility theory for decision making [35], the construct *utility* assesses a user's preference for outcome. We seek a quantitative representation of a qualitative preference for attributes, which in turn is pegged to the task. To accomplish the task in our study, a user explores and reviews attributes, both on their own and with help from guided attributes, before finalizing the attributes. We thus use the total number of attributes selected as the metric of utility to test **H4**. Note that: (i) users did not know about our hypotheses and had no incentive to select few or more attributes than their judgment; (ii) the task's goal was for users to select only those attributes that they judged to be relevant; (iii) results show that users did not select too few or too many attributes. That said, as the literature of over 100 years testifies, measurement of utility, a subjective construct, does not lend itself to a unique measure [71]. One can design more involved studies solely to measure utility, but is outside the scope of this paper [71].

H5 Participants will *verify the guidance more* when it comes from AI > Expert > Group (> implies more).

Metric(s). We define *verification* as the change in users' attention (a) towards guided attributes before and after receiving them as guidance and (b) between guided and unguided attributes, providing a comprehensive assessment of how guidance sources influence verification behavior.

(a) *Difference in number of interactions with guided attributes after and before receiving them as guidance,* or Δx:

$$\Delta x = \sum_{i=1}^{k} (x_{i,\text{after}} - x_{i,\text{before}}) \tag{2}$$

where: $x_{i,\text{after}}$ and $x_{i,\text{before}}$ are the number of interactions with guided attribute i after and before receiving as guidance, respectively, and k is the total number of guided attributes.

Since there may be zero interactions with an attribute before and after receiving it as guidance, we model this metric as a *difference* instead of *ratio*, avoiding divide-by-zero errors and enabling statistical comparisons.

(b) *Ratio of number of interactions with guided attributes to total number of interactions with all attributes,* or R:

$$R = \frac{\sum_{i=1}^{k} x_{i,\text{guided}}}{\sum_{j=1}^{n} x_{j,\text{total}}} \quad (3)$$

where: $x_{i,\text{guided}}$ is the number of interactions with guided attribute i, k is the total number of guided attributes, $x_{j,\text{total}}$ is the number of interactions with all attributes, and n is the total number of attributes. Since users may not interact with all attributes, particularly in datasets with many attributes (e.g., thousands or even 33 as in our case), and considering that our guidance was capped at 10 attribute suggestions, we model this metric as a *ratio* instead of *difference*. Unlike (a), this ratio avoids divide-by-zero errors and provides a clearer comparison of the user's attention on guided attributes relative to all attributes.

5 USER STUDY

5.1 Participants

We recruited participants by emailing relevant mailing lists within a public U.S. university. Each interested applicant was screened based on their self-reported visualization literacy (at least 3 out of 5), age (at least 18 years), and physical location (in the U.S.), as per our study protocol approved by the ethics board. We invited 109 screened applicants to participate in the study of whom 90 participants completed the study. We discarded data of three of these participants from the resultant analysis due to suspicious/insufficient interaction behavior, resulting in 87 valid participants[4].

Our valid participants were either pursuing or had received *professional* (n=2), *associates* (1), *bachelors* (24), *masters* (38), *doctoral* (21) degrees or a *technical certificate* in *computer science* (n=49), *human-centered computing* (4), *human-computer interaction* (8), *mechanical engineering* (4), *analytics*, *applied and computational mathematics*, *business administration* (2), *chartered finance analyst*, *cybersecurity*, *digital media*, *economics*, *electrical and computer engineering*, *history*, *industrial engineering*, *information management*, *mechatronics, robotics and automation engineering*, *pharmaceutical sciences*, *physics*, *public health*, *public policy*, *robotics*, and *statistical science*. Demographically, they were in the *18-24* (n=47), *25-34* (38), *35-44* (1), or *45-54* (1) age groups in years and of *female* (42), *male* (44), or *preferred not to say* (1) genders. They reported their level of experience looking at data in a visual form (e.g., scatterplot, bar chart) on a scale from 1 (non-expert) to 5 (expert): 3 (n=36), 4 (37), 5 (14). They also reported their level of experience working with analysis tools such as Excel (n=85), Programming (75), Tableau (48), PowerBI (12), QlikView (3), D3 (3), SQL, SAS, Plotly, Kibana, Colaboratory, .NET, matplotlib, MicroStrategy, Looker, Grafana, Sheets, and Gephi.

5.2 Study Session

We conducted the study remote asynchronously, providing each participant with a unique link to the study's user interface and giving a maximum two weeks to complete the task, albeit in a single, focused session. Each study session lasted between 30 and 60 minutes. We compensated each participant with a $15 gift card for their time.

During a study session, participants first provided consent (**Consent**), received relevant background information on providing guidance during data analysis workflows (**Preface**), and filled a questionnaire to provide their general prior beliefs about and experiences receiving guidance in life (**Pre-Study Questionnaire 1**). Next, participants saw a video tutorial (**Training**) that demonstrated the features of the study interface and performed a practice task (**Practice Task Instructions** and **Practice Task**) to familiarize themselves with the interface. After practice, participants saw the actual task instructions (**Main Task Instructions**) and filled a questionnaire to provide their prior beliefs with respect to the task (**Pre-Study Questionnaire 2**). Next, participants performed the main task using the study interface (**Main Task**). Lastly, participants filled a feedback questionnaire based on the study task (**Post Study Questionnaire**) and another questionnaire to collect their demographic information (**Background Questionnaire**). Participants' interactions with the interface (e.g., the attributes they selected) were logged during the study for subsequent analysis. Figure 3 illustrates the entire study procedure; whereas detailed study questionnaires are available in supplemental material.

6 RESULTS

We first describe our choice of statistical analysis, then present findings related to all hypotheses, followed by additional analyses of usage behavior and responses from the pre- and post-study questionnaires. We include summary visualizations and statistical findings related to the hypotheses and the questionnaires in the Appendix.

6.1 Statistical Analysis - Non-parametric

A common analysis to compare study conditions, as in this paper, is a statistical test for comparison of means of responses, which relies upon the assumption of normal distribution of responses. The number of participants in each of our five conditions is 17, or 18, which falls well below the number 30, at which the distribution of the sample mean of a metric in a condition approaches the normal distribution. Also, the measures in our study are of two types: subjective, stated responses to questions, and objective metrics of usage. The distribution of several metrics, within each condition, depicts considerable skewness, violating assumption of normal distribution. Thus, we use non-parametric analysis and tests, which do not rely on inherent assumptions about distribution and provide robustness to compare conditions for responses and metrics alike. This allows a consistent approach to both sets of measures. In particular, we use *median*, instead of mean, and perform non-parametric statistical tests. Additionally, to comprehensively examine statistical relationships among a large set of responses, we do a parametric, multivariate regression analysis. We conducted all statistical analyses using well-established Python libraries: scikit-learn[5] and statsmodels[6].

6.2 Results of Source-Agnostic Hypotheses

We report the results of our three source-agnostic hypotheses (**H1**–**H3**) involving **Unattributed** and **Control** conditions, using one-sided Mann-Whitney U tests (one-sided as the hypotheses are one-sided).

[4]Participants whose number of interactions or task duration fell outside the mean ± 2 standard deviations were deemed outliers, and excluded from analysis.

[5]https://scikit-learn.org/stable/

[6]https://www.statsmodels.org/stable/index.html

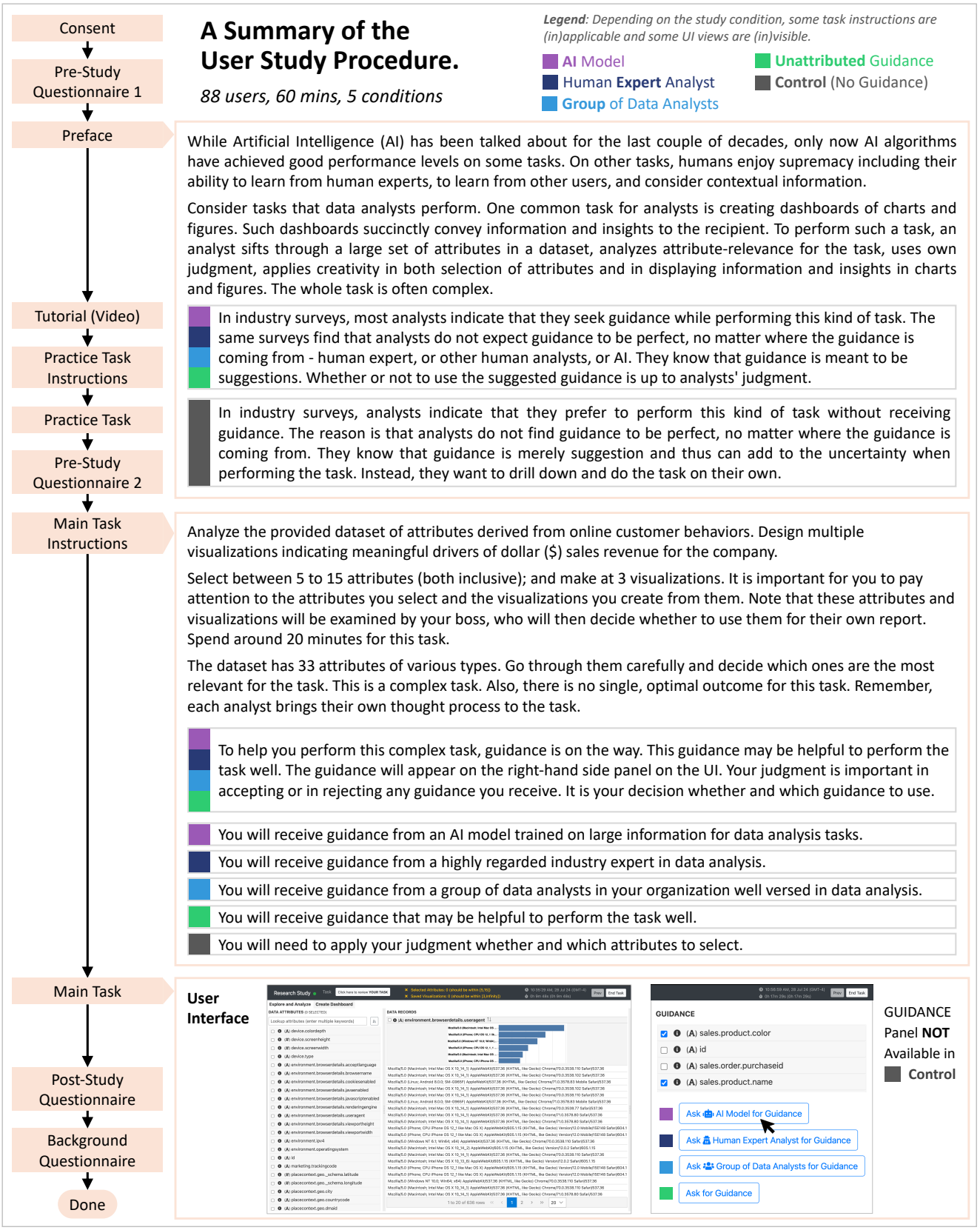


Figure 3: Overview of our Study Design including the various steps, task instructions, and screenshots of the study prototype.

H1 **Unattributed participants who receive guidance will *be less uncertain about their attribute selections* than Control participants who do not receive guidance.**

Metric: Variance in the number of interactions with attributes.

The median variance in the number of interactions across all attributes for **Unattributed** participants (median=19.18) was smaller than **Control** participants (median=21.82). These results **directionally favor our hypothesis**, but are not statistically significant (p-value=0.34).

H2 **Unattributed participants who receive guidance will *select more attributes* than Control participants who do not receive guidance.**

Metric: Total number of attributes selected at task end.

Unattributed participants selected more attributes (median=10.5) than **Control** (median=6.0). These results **directionally favor our hypothesis** and are **statistically significant** (***p*-value=0.001**).

H3 **Unattributed participants who receive guidance will *take lesser time* to complete the task than Control participants who do not receive guidance.**

Metric: Total duration of the task (in minutes).

Unattributed spent less time (median=12.76 minutes) than **Control** (median=13.57 minutes). These results **directionally favor our hypothesis** but are not statistically significant (p-value=0.67).

Metric: Total number of interactions made during the task.

Unattributed performed less interactions (median=141) than **Control** (median=145.5). These results **directionally favor our hypothesis** but are not statistically significant (p-value=0.72).

For more details, refer to Table 4 in Appendix A.

6.3 Results of Source-Specific Hypotheses

We report results of our two source-specific hypotheses (**H4**–**H5**) involving **AI**, **Expert**, and **Group** conditions, using pairwise one-sided Mann-Whitney U tests with Bonferroni correction (one-sided tests as both hypotheses are one-sided).

H4 **Participants will find guidance to *have more utility* when it comes from Expert > AI > Group.**

Metric: Total number of attributes selected at task end.

AI (median=9) and **Expert** (median=9) both selected more attributes than **Group** (median=7). These results **directionally favor our hypothesis, albeit partially**; the pairwise comparisons between the three conditions revealed **statistical significance** between **Expert** and **Group** (***p*-value=0.01**) and **AI** and **Group** (***p*-value=0.04**).
Notably, all three source-specific conditions selected fewer attributes than **Unattributed** (median=10.5) and more attributes than **Control** (median=6) (**H2**).

H5 **Participants will *verify the guidance more* when it comes from AI > Expert > Group.**

Metric: Difference in the number of interactions with guided attributes after and before guidance.

AI had the largest difference in number of interactions with guided attributes after and before receiving them as guidance (median=31), followed by **Expert** (median=14), and then **Group** (median=12.5). These results **directionally favor our hypothesis** and the pairwise comparisons between the three conditions revealed **statistical significance** between **AI** and **Group** (***p*-value=0.003**).

Metric: Ratio of # interactions with guided attributes to # interactions with all attributes.

AI had the largest ratio of the number of interactions with guided attributes to number of interactions with all attributes (median=0.36), followed by **Expert** (median=0.23), and then **Group** (median=0.19). These results **directionally favor our hypothesis** and the pairwise comparisons between the three conditions revealed **statistical significance** between **AI** and **Group** (***p*-value=0.01**).

For more details, refer to Table 5 in Appendix B.

6.4 Results of User Behavior Analysis

Besides metrics for the analysis of hypotheses, we compute other metrics that shine light on users' analytic behaviors while performing the task and can enrich our understanding of utilization of guidance. Specifically, user behavior analyses indicate interesting new hypotheses that draw from the rich psychology literature and set valuable future research directions.

6.4.1 Exploration of Unique Attributes. Figure 4 shows the distribution of the total number of unique attributes interacted by users. **Group** shows the highest exploration (median=25) followed by **Unattributed** (20.5), **AI** (20), **Expert** (20), and **Control** (16). Directionally, among attributed sources, less exploration under **AI** and **Expert** suggests a belief in more prescriptive guidance from these two sources relative to **Group**.

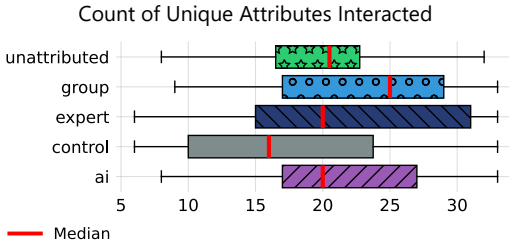


Figure 4: Number of unique attributes interacted.

6.4.2 Exploitation of Availability of Guidance. Figure 5 shows the distribution of the number of times users requested guidance, when available. **Unattributed** and **AI** find the most exploitation (median=5) followed by **Expert** and **Group** (median=3). Four participants in each of the **AI**, **Expert**, **Group**, and **Unattributed** conditions did not request guidance at all. Among attributed sources, these results may indicate a belief of more to be gained from exploitation of **AI** than from others; that is, more *salience* [48] of **AI**.

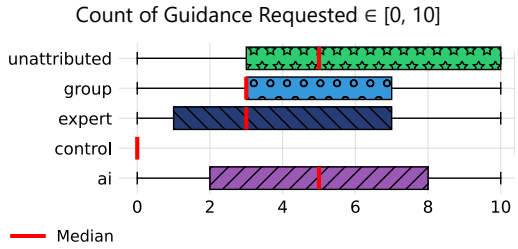


Figure 5: Number of times guidance was requested.

6.4.3 Eagerness to Access Guidance . Figure 6 shows the distribution of two metrics - time taken (in minutes) and number of interactions performed - before requesting guidance for the first time.

Metric: Time taken. Analyzing for the shortest time, **Group** finds the most eagerness to request guidance (median=0.28 minutes into the task) followed by **Unattributed** (0.45), **Expert** (2.19), and **AI** (2.81).

Metric: Interactions performed. Analyzing for the least number of interactions, **Group** sees the most eagerness to request guidance (median=0 interactions into the task) followed by **Unattributed** (2), **Expert** (22), and **AI** (24). The high eagerness to access guidance under **Group** may indicate a belief that the stored knowledge from this source is more psychologically accessible [48], given that the source comprises analysts who are peers and thus at the same level of hierarchy as users, unlike **AI** and **Expert** who are perceived at a higher level. This aspect of *accessibility* along with salience is central to users' knowledge activation [48], and is well recognized in information processing in psychology.

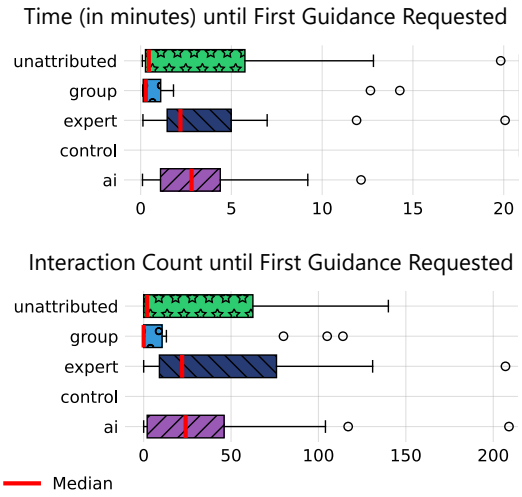


Figure 6: Time taken (in minutes) and interaction count until guidance was requested for the first time.

6.4.4 Behavioral Decision Making about Attributes. Given our conceptualization of guidance as behavioral decision making about attributes, we analyzed how participants' interaction behavior with respect to attribute selection and deselection changed during the task relative to the received guidance (Figure 7).

(1) [Attribute Guided → Not Selected] *Users received an attribute as guidance and never selected it.* **Group** (median=2.0) exhibited this behavior more than **AI**, **Expert**, and **Unattributed** (median=1.0). This directionally points to users' higher *disagreement* with guidance coming from **Group** than from **AI** or **Expert**, suggesting more benefit from latter among attributed sources. This metric is similar to Lu et al.'s "disagreement" metric [64].

(2) [Attribute Guided → Selected] *Users received an attribute as guidance and selected it.* **Unattributed** (median=3.0) exhibited this behavior more than **AI** (median=2.0) than **Group** (median=1.5) than **Expert** (median=1.0). This measure of *agreement* with guidance, shows no directional advantage among any of the attributed sources. This metric is similar to Lu et al.'s "agreement" metric [64].

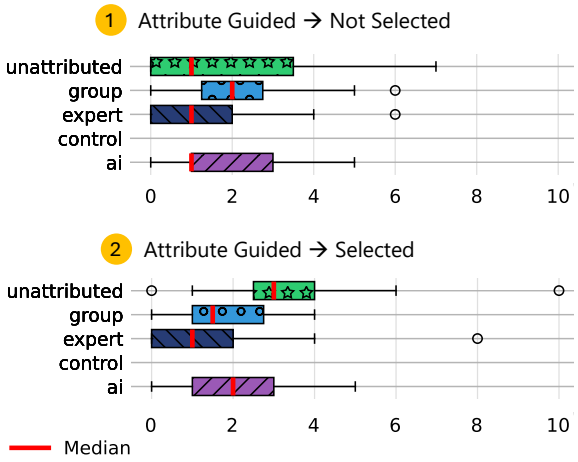


Figure 7: Distribution of the number of times participants (1) did not select a guided attribute and (2) did select a guided attribute.

6.5 Analysis of Stated Responses

To complement the objective measures for utilization, we report findings from analyzing the subjective response(s) in **Pre-Study Questionnaire 2** and **Post-Study Questionnaire**. For detailed statistics, refer to Table 6 in Appendix C.

6.5.1 Opinion about Guidance before and after analysis. Figure 8 shows how participants' opinions about guidance ("Reliability", "Confidence", "Value Add") changed after the task, on a scale from 1 (Disagree) to 7 (Agree), along with best fit regression lines. For each factor, we compute the difference between relevant participants' *after* and *before* scores. Then, to test for statistical significance between the four conditions that received guidance (**AI**, **Expert**, **Group**, **Unattributed**), we utilize pairwise two-sided Mann Whitney U tests with Bonferroni correction.

Metric: Reliability. **Unattributed** (median=-2.0) scored lowest followed by **Group** (median=-1.0), then **Expert** and **AI** (median=0.0), with no statistically significant differences.

Metric: Confidence. **Unattributed** and **Expert** (median=-2.0) scored lowest followed by **Group** and **AI** (median=-1.0), with no statistically significant differences.

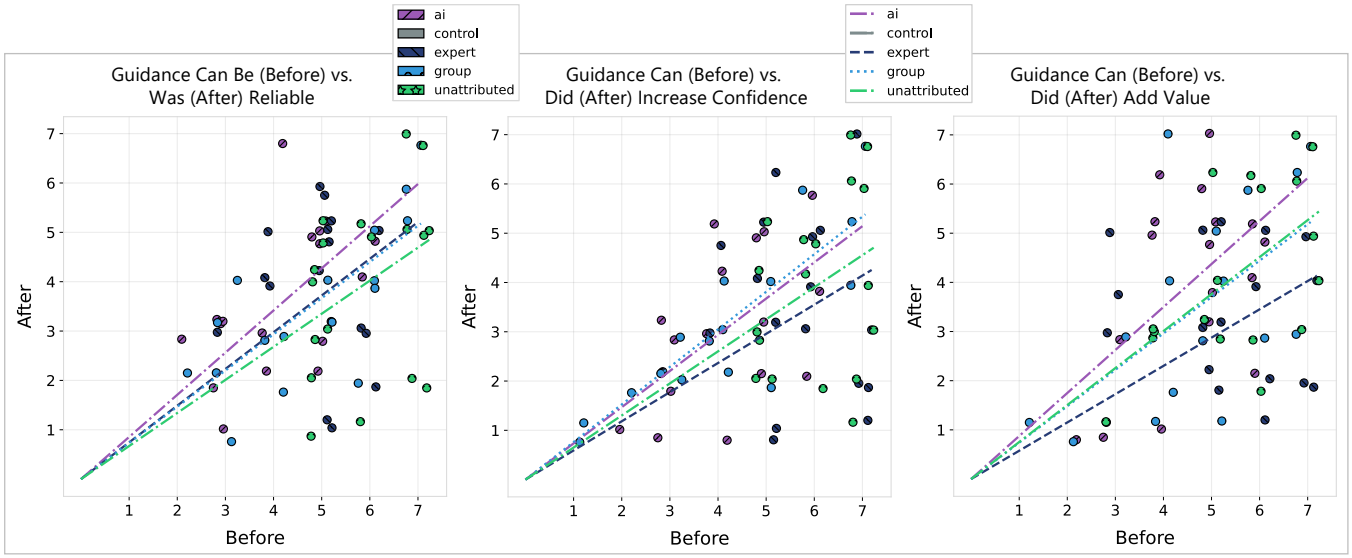


Figure 8: Participants' opinion of guidance ("Reliability", "Confidence", "Value Add") before and after the task, on a scale from 1 (Disagree) to 7 (Agree), as reported in the Pre-Study Questionnaire 2 and Post-Study Questionnaire. Control neither received guidance nor these questions. A slope–of the regression line–steeper (shallower) than the 45-degree line shows after-task responses higher (lower) than pre-task responses, per condition.

Metric: Value Add. **Expert** (median=-2.0) scored lowest followed by **Group**, **AI**, and **Unattributed** (median=-1.0), with no statistically significant differences.

6.5.2 Reliance on- and regret using guidance. Figure 9 shows users' response scores, on a scale from 1 (None at all) to 7 (A lot), on how much they relied on ("Reliance") and regretted relying on ("Regret") guidance during the task.

Metric: Reliance. **AI** and **Unattributed** reported higher reliance on guidance (median=3.0) than both **Expert** and **Group** (median=2.0).

Metric: Regret. **AI** reported higher regret relying on guidance (median=5.0) than both **Expert** and **Group** (median=2.0), with **Unattributed** in between (median=3.5).

6.5.3 Regret in Guidance: A Deep-dive. In Section 1, we introduced the threefold importance of regret for guidance systems: post-task regret may impact continued utilization of guidance, anticipatory regret may distort use of guidance [93], and managing expectations of users may be necessary for success. Taking a deeper dive, we performed a regression analysis of drivers of regret, summarized in Table 2. **Analysis I** finds that among the nine guidance features having post-task response, only "Guidance Increased **Confidence**" is considered an important driver of post-task **Regret**. The negative coefficient implies that as confidence increases, **Regret** decreases, which makes intuitive sense. Among sources, **AI** is significant and its coefficient positive (relative to the baseline, **Group**, in this regression analysis), suggesting post-task **Regret** is higher for **AI** than other two attributed sources, in the presence of these features.

Analysis II examines effects of *Expectation* on post-task Regret. In the Pre-Study Questionnaire 2, we collected responses on three guidance features, after priming about the source, but before the task was performed, giving measure of user's *Expectation* on those three features. Then, in the post-task questionnaire (Post-Study Questionnaire), we collected responses on the same three features to give measures of *Realization*. The *change–Realization minus Expectation*–shows significance for all three features. As "Guidance Increased **Confidence**" improves even more from *Expectation*, **Regret** decreases further. Also, **Group** and **AI** both contribute to higher **Regret** in the presence of these changes (noting that the coefficient for **AI** is additive to the baseline, **Group**'s, coefficient).

6.6 Qualitative Feedback

In the **Post-Study Questionnaire**, we asked participants to respond to four key questions related to the study task. We analyzed their responses by applying open coding [10], specifically, constant comparison and theoretical sampling [89]. Below, we list these four questions along with corresponding key takeaways.

"Describe your analysis strategy to identify relevant attributes." Participants employed a mix of personal judgment and guidance. Some sought guidance early, integrating the recommendations with their own analysis, while others relied more on self-judgment initially. Many focused on attributes that provided clear insights into customer behavior, sales patterns, and return on investment, prioritizing those with distinct distributions, while avoiding more complex or less relevant attributes like "IP address".

"Describe your strategy with respect to designing the visualizations." Participants predominantly created visualizations emphasizing market demographics and sales drivers, as was required by the task. Their primary strategy involved trial and error, with participants selecting relevant attributes, testing various combinations to identify patterns, and then refining the visualizations based on the data's ability to convey meaningful insights quickly and intuitively. Most participants

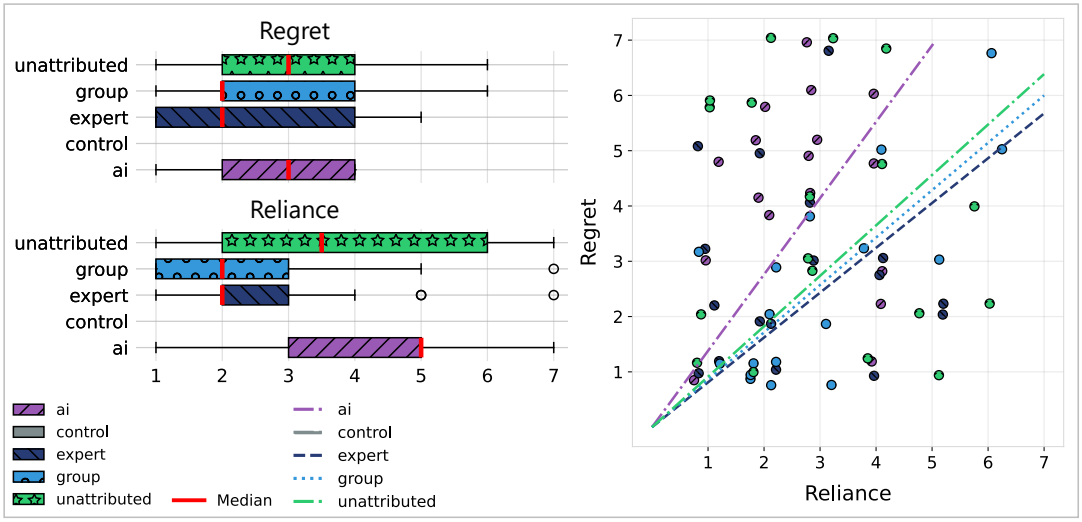


Figure 9: Participants' response scores about their reliance on and regret using guidance, on a scale from 1 (None at all) to 7 (A lot), as reported in the Post-Study Questionnaire. Control neither received guidance nor these questions. In the scatterplot, a slope–of the regression line–steeper (shallower) than the 45-degree line shows regret higher (lower) than reliance, per condition.

Table 2: Two regression analyses focused on the question on Regret in the Post-Study Questionnaire. Features comprise stated responses, which are numerical variables, and attributed sources that comprise categorical variable and are one-hot encoded. Bold and underline indicate statistically significant Pr>|t| at $\alpha = 0.05$.

Feature	**Parameter Estimate**	$Pr > \|t\|$
Analysis I		
Regression Analysis: Target - Post study Regret. Features - Post study responses and attributed-Sources. $Adj - R^2 = 0.29$		
Guidance Was **Trustworthy**	0.08	0.86
Guidance Came From A **Knowledgeable** Source	0.23	0.16
Guidance Was **Reliable**	0.39	0.38
Guidance Increased **Confidence**	-0.63	**0.03**
Guidance Was **Relevant**	-0.17	0.59
Guidance Helped Avoid Potential **Pitfalls**	0.41	0.13
Guidance Gave Valuable **Suggestions**	0.13	0.58
Guidance **Added Value**	0.34	0.34
Guidance Was **Appropriate**	-0.09	0.63
AI	1.46	**0.02**
Expert	0.65	0.28
Intercept (**Group**)	-0.03	0.97
Analysis II		
Regression Analysis: Target - Post study Regret. Features - Post *minus* Pre study responses and attributed-Sources. $Adj - R^2 = 0.23$		
Guidance Was **Reliable** *minus* Guidance can be **Reliable**	0.38	**0.04**
Guidance Increased **Confidence** *minus* Guidance can **Increase Confidence**	-0.61	**0.01**
Guidance **Added Value** *minus* Guidance can **Add Value**	0.40	**0.04**
AI	1.06	0.08
Expert	-0.29	0.65
Intercept (**Group**)	2.95	**<0.0001**

prioritized simplicity for managerial discussions, using scatter plots and bar charts while limiting encoded attributes to three for clarity. Another participant strategy focused on creating visualizations that highlighted correlations. They experimented with different visualization types and configurations, such as line charts for time series data and scatter plots for comparing numeric variables.

Metric	AI	Expert	Group	Unattributed	Control
Mental Demand	4.0	4.0	3.0	4.0	**5.0**
Physical Demand	1.0	2.0	1.0	1.0	**3.0**
Temporal Demand	2.0	2.0	2.0	2.0	**4.0**
Performance	**5.0**	**5.0**	**5.0**	**5.0**	**5.0**
Effort	4.0	4.0	4.0	4.5	**5.5**
Frustration	2.0	2.0	2.0	4.0	**5.0**

Table 3: Median scores about the fidelity of the task, as self-reported by participants on a scale from 1 (Low) to 7 (High), in the Post-Study Questionnaire. Bold and underline indicate the largest values in each row.

"Do you have additional comments about the guidance you received during the task?" Feedback on the guidance received during the task was mixed. Some participants noted issues with the guidance, such as receiving attributes that were confusing or irrelevant to their analysis. For example, one participant found the initial attribute suggestions unhelpful due to a lack of clarity and connection between them, while another felt that specific attributes like "ID"s were not useful for their aggregate-focused task. A few participants expressed frustration with the guidance's lack of explanations and rationale, which hindered their understanding and trust in the suggestions. Conversely, some participants found the guidance beneficial, appreciating its ability to offer new attribute ideas or confirming their own choices. They also noted that explanations or additional context would have enhanced its utility, particularly when attributes were suggested one at a time.

"Do you have additional comments about the guidance you may receive from different sources shown during the task?" Participants expressed a slight preference for human guidance over AI, citing the ability of humans to provide more contextual and tailored advice. One participant suggested that combining AI with human expert guidance would be more effective. Others noted that internal guidance from colleagues might be more relevant due to familiarity with organizational needs, with one remarking, *"I struggle to decide how I feel about the group of analysts vs the human expert. Of course, I am inclined to believe/take guidance from an expert in [visualization] and/or data analytics, but I may lean heavier on a 'less experienced' group (or single) person in the company as they are aware how the business runs, what is important to the executives, and if that one director of marketing *really* loves pie charts (like a lot). An expert in the field can certainly provide great feedback on best practices, new methods, and what the research shows, but they cannot comment on the idiosyncrasies of the company like internal analysts could."*

6.7 Overall Task Fidelity and System Usability

Task Fidelity. Figure 10 and Table 3 show summary statistics of participant responses to questions in the **Post-Study Questionnaire** about the fidelity of the task. **Control** reported the highest amount of mental demand, physical demand, temporal demand, effort, and frustration. All study conditions scored equally on performance.

System Usability. Participants evaluated our study prototype using the System Usability Scale (SUS) [12], reporting an overall mean score of 72.67. For individual study conditions, participants reported mean scores of 76.76 (**AI**), 74.85 (**Expert**), 69.41 (**Group**), 74.31 (**Unattributed**), and 68.19 (**Control**).

7 DISCUSSION

In this section, we discuss how source-attribution in guidance affected users' usage and perception. For detailed statistics associated with participants scores to the two pre- and one post-study questionnaires, refer to Table 6 in Appendix C.

The nuances in how the perception of guidance changes with source-attribution, even when the quality of the guidance and the task remain identical, are captured in responses to **Pre-Study Questionnaire 2** and the **Post-Study Questionnaire**. Recall that **Pre-Study Questionnaire 2** was administered after participants were primed about the respective source-condition, but before they started the task. Hence, questions in the **Pre-Study Questionnaire 2** captured ex-ante *expectation* of participants, associated with the source. The **Post-Study Questionnaire** was administered after the completion of task, and thus represents post-task *realization* of participants.

Comparing **AI**, **Expert**, and **Group** across "Pre Study Questionnaire 2: Questions about Guidance," we find that expectations about *Reliable*, *Increase-Confidence*, and *Value-add* are similar across three source-conditions, except for *Increase-Confidence*, where **Expert** scores higher. These scores indicate individual's baseline expectations about these characteristics of guidance. We compare participants' realizations after the task to see how much change occurred versus their expectations. Since survey responses are useful to interpret relative expectations instead of absolute ones, our approach provides clearer insights.

Analyzing responses in the **Post-Study Questionnaire**, we find that *Reliable* and *Increase-Confidence* score similarly across three conditions; yet, *Value-Add* scores higher for **AI**. Digging deeper and comparing with expectations stated in the **Pre-Study Questionnaire 2**, in *Value-Add*, post-task realization of participants meet pre-task expectations, only for **AI**, but not for the two other sources, where their realizations fall short of expectations. In terms of *Reliable* and *Increase-Confidence*, realizations fall short of expectations for all three conditions. All this implies that in one out of three characteristics, realizations from **AI** meets expectations, but for **Expert** and **Group** realizations fall short of expectations in all three characteristics. Now consider that our 1 (Disagree) - 7 (Agree) scale admits *favorable* ratings only in the range 5 - 7. The post study questions find that, out of the 9 characteristics, **AI** scores favorable in 3 of them, **Expert** in none, and **Group** in 1 characteristic. All this evidence suggests disposition toward **AI** is no worse off than toward **Expert** or **Group**, and better in more characteristics, and should augur well for **AI**.

However, after task completion in the **Post-Study Questionnaire**, participants' response to the question asked of them "How Much [Did You] Regret Relying on Guidance," belie the above mentioned disposition toward **AI**. Regret score is significantly higher for **AI** than for **Expert** and **Group**. Yet, after task completion, participants acknowledge "Guidance Came From A Knowledgeable Source" and "Guidance Gave Valuable Suggestions" with scores for **AI** at least as high as that of **Expert** and **Group**. Note that the weight of overall evidence points to something more nuanced than merely lower trust in **AI** relative to other sources, which can account for resistance

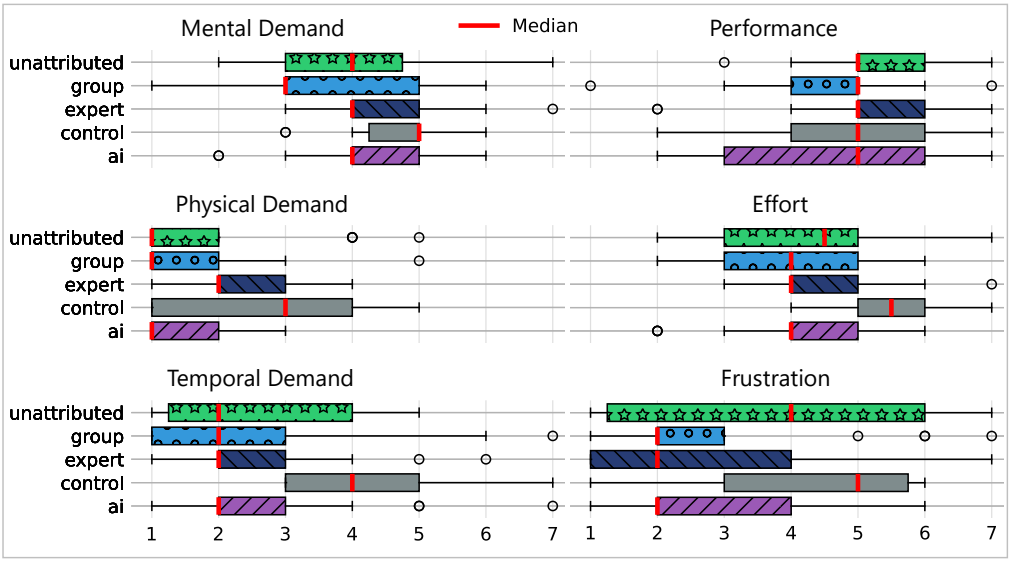


Figure 10: Participants' self-reported responses about the overall fidelity of the task on a scale from 1 (Low) to 7 (High), as reported in the Post-Study Questionnaire.

to adoption of guidance coming from **AI**. We hypothesize that any enhanced task performance by a human with the help of **AI** may create post-task dissonance about the "messenger" (being a non-human) although the "message" (quality of guidance) is the same across all sources. Future research can deep dive into dissonance to test this hypothesis and to find ways to mitigate.

Moreover, our use of **Unattributed** guidance provides a good source-agnostic benchmark for difference between realization and expectation. For **Unattributed** guidance, in each of *Reliable*, *Increase-Confidence* and *Value-add*, post-task scores are significantly lower than pre-study expectation, indicating that omission of source is undesirable as well. In addition, scores from "Post Study Questionnaire: (*What-If*) Questions about Guidance" demonstrate that participants under **Unattributed** guidance condition, assign a high score of 6 to the counterfactual sources **Expert** and **Group**, which is higher than scores assigned to the counterfactual source **AI**, for each of three characteristics: "How Much **Faith** Would You Have," "How Much Would You **Prefer**," and "How Much Would You **Rely**." These scores on post task counterfactual questions by users in **Unattributed** source condition suggest two inferences: (i) an overall lower preference for **AI**, and (ii) a desire for attributed-source, and (iii) that too pegged to humans - **Expert** or **Group**. Future research can dive deeper into these inferences.

8 LIMITATIONS AND FUTURE WORK

Our study had some limitations. First, we did not provide explanations or justifications for the guidance, such as reasoning behind AI recommendations or context from human analysts, which could impact users' trust and understanding. Moreover, by design, the task was somewhat subjective, as the relevance of attributes in data analysis can vary based on user interpretation, making it difficult to establish a clear ground truth. Next, while our study was focused on attribute selection, future studies may explore other aspects of data preparation–such as collection, cleaning, transformation–and data analysis [45]. Lastly, to help our participants select relevant attributes, we provided them with the attribute definitions, underlying data, summary visualizations, and the ability to create visualizations using one or more attributes with different aggregation levels applied. However, more involved user studies with other tools and features (e.g., advanced statistical indicators) are needed to understand how participants utilize the provided guidance to perform the task. Despite these limitations, our findings offer valuable insights into how guidance sources influence user behavior during data-driven analysis and decision-making on subset selection.

We have also identified many opportunities as future work. First, in this study, we asked participants to self-report their "Regret" in utilizing the provided guidance on a Likert-scale ranging from 1 (None at all) to 7 (A lot); based on the reported scores, we found that participants in the **AI** guidance condition expressed greater regret at the end of the task. Future work can specifically study the underlying factors contributing to participants' regret and also devise means to mitigate the same. This was beyond the scope of our experiment. Second, examining how guidance influences user behavior over time and aligns with the sensemaking process may reveal when guidance is most useful. Likewise, future work can also study when users may desire specific sources of guidance, e.g., **AI** versus **Expert**. To study these temporal dynamics, a tool integrated with user mechanisms to switch between the different guidance sources is desirable. Another promising direction for future work is examining how user expertise influences guidance utilization. For instance, do experts request guidance less frequently, or do they rely on it at specific stages of sensemaking? Investigating these behaviors could inform the design of adaptive guidance systems that tailor recommendations based on user expertise and their evolving needs. Lastly, through this study, we did not offer specific prescriptions to improve **AI** guidance systems, given **AI**'s burgeoning role; however,

findings from our study call for pointed future research devoted to dive deeper into the nuances of **AI** as a source for guidance. Such research can invest in a large-scale guidance system, allowing more specific, actionable recommendations, but is outside the scope of this paper, and could be more prescriptive.

9 CONCLUSION

We studied how the source of guidance – **AI**, **Expert**, or **Group** – impacts users' perceptions and behaviors during an attribute selection task as part of data preparation. In a five-condition between-subjects study, including an **Unattributed** guidance condition (i.e., guidance without any source attribution) and a no-guidance baseline (**Control**), study participants used a custom tool to select relevant attributes from an unfamiliar dataset, with guidance provided in the form of attribute suggestions. We controlled for guidance quality across conditions to ensure comparability. Our results showed that the source of guidance matters, as it influenced users' behavior in different ways during the task. In particular, participants in the **AI** guidance condition reported higher benefits but also expressing greater regret after the task. Overall, these findings strongly suggest the need to carefully consider guidance sources for effective guidance offerings in systems.

ACKNOWLEDGMENTS

This material is based upon work supported in part by NSF IIS-1750474 and a gift funding from Adobe, Inc. We are grateful to members of the Georgia Tech Visualization Lab, our user study participants, and anonymous reviewers for their constructive feedback at different stages of this work.

REFERENCES

[1] 2023. Bigeye. Retrieved Jun 1, 2023 from https://www.bigeye.com/
[2] 2023. Datafold. Retrieved Jun 1, 2023 from https://www.datafold.com/
[3] 2023. Monosi. Retrieved Jun 1, 2023 from https://monosi.dev/
[4] 2023. Monte Carlo Data. Retrieved Jun 1, 2023 from https://www.montecarlodata.com/
[5] 2023. SQLLineage. Retrieved Jun 1, 2023 from https://github.com/reata/sqllineage
[6] Abhijit V Banerjee. 1992. A simple model of herd behavior. *The quarterly journal of economics* 107, 3 (1992), 797–817.
[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. 2019. A case for backward compatibility for human-ai teams. *arXiv preprint arXiv:1906.01148* (2019).
[8] David E Bell. 1982. Regret in decision making under uncertainty. *Operations research* 30, 5 (1982), 961–981.
[9] David Borland, Wenyuan Wang, Jonathan Zhang, Joshua Shrestha, and David Gotz. 2019. Selection bias tracking and detailed subset comparison for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 429–439.
[10] Richard E Boyatzis. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development.* Sage Publications.
[11] Will Brackenbury, Rui Liu, Mainack Mondal, Aaron J. Elmore, Blase Ur, Kyle Chard, and Michael J. Franklin. 2018. Draining the Data Swamp: A Similarity-Based Approach. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. Association for Computing Machinery, New York, NY, USA, Article 13, 7 pages. doi:10.1145/3209900.3209911
[12] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[13] Erik Brynjolfsson and Kristina McElheran. 2016. The rapid adoption of data-driven decision-making. *American Economic Review* 106, 5 (2016), 133–139.
[14] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.
[15] Michael J. Cafarella, Alon Halevy, and Nodira Khoussainova. 2009. Data Integration for the Relational Web. *Proceedings of the VLDB Endowment* 2, 1 (2009), 1090–1101. doi:10.14778/1687627.1687750
[16] Davide Ceneda, Natalia Andrienko, Gennady Andrienko, Theresia Gschwandtner, Silvia Miksch, Nikolaus Piccolotto, Tobias Schreck, Marc Streit, Josef Suschnigg, and Christian Tominski. 2020. Guide me in analysis: A framework for guidance designers. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 269–288.
[17] Davide Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Hans-Jörg Schulz, Marc Streit, and Christian Tominski. 2016. Characterizing guidance in visual analytics. *IEEE TVCG* (2016).
[18] Davide Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Hans-Jörg Schulz, Marc Streit, and Christian Tominski. 2017. Amending the Characterization of Guidance in Visual Analytics. *arXiv preprint arXiv:1710.06615* (2017).
[19] Davide Ceneda, Theresia Gschwandtner, and Silvia Miksch. 2019. A review of guidance approaches in visual data analysis: A multifocal perspective. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 861–879.
[20] Yi-Fen Chen. 2008. Herd behavior in purchasing books online. *Computers in Human Behavior* 24, 5 (2008), 1977–1992.
[21] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *Proceedings of the 13th ACM Web Science Conference 2021*. 120–129.
[22] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018.
[23] Christopher Collins, Natalia Andrienko, Tobias Schreck, Jing Yang, Jaegul Choo, Ulrich Engelke, Amit Jena, and Tim Dwyer. 2018. Guidance in the human–machine analytics process. *Visual Informatics* 2, 3 (2018), 166–180. doi:10.1016/j.visinf.2018.09.003
[24] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K Elmagarmid, Ihab F Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. 2017. The Data Civilizer System. In *CIDR*. https://dblp.org/rec/conf/cidr/DengFAWSEIMO017.html
[25] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-computer interaction*. Pearson Education.
[26] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2009. Informational cascades and software adoption on the internet: an empirical investigation. *MIS quarterly* (2009), 23–48.
[27] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)* (2019).
[28] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic Interaction for Visual Text Analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 473–482. doi:10.1145/2207676.2207741
[29] Robert Engels. 1996. Planning tasks for Knowledge Discovery in Databases; Performing Task-Oriented User-Guidance.. In *KDD*. 170–175.
[30] Per Engström and Eskil Forsell. 2018. Demand effects of consumers' stated and revealed preferences. *Journal of Economic Behavior & Organization* 150 (2018), 43–61.
[31] Ernst Fehr, Urs Fischbacher, and Michael Kosfeld. 2005. Neuroeconomic foundations of trust and social preferences: initial evidence. *American Economic Review* 95, 2 (2005), 346–351.
[32] Paul J Feltovich, Michael J Prietula, K Anders Ericsson, et al. 2006. Studies of expertise from psychological perspectives. *The Cambridge handbook of expertise and expert performance* (2006), 41–67.
[33] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE.
[34] Ian Fette and Alexey Melnikov. 2011. *The websocket protocol*. Technical Report.
[35] Peter C Fishburn, Peter C Fishburn, et al. 1979. *Utility theory for decision making*. Krieger NY.
[36] Imola K Fodor. 2002. *A survey of dimension reduction techniques*. Technical Report. Lawrence Livermore National Lab., CA (US).
[37] Python Software Foundation. 2024. Python. Retrieved June 13, 2024 from https://www.python.org/
[38] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 794–806.
[39] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.
[40] Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT press.
[41] Google. 2024. Angular. Retrieved June 13, 2024 from https://angular.io/
[42] Google. 2024. Google Cloud Logging. Retrieved June 13, 2024 from https://cloud.google.com/logging
[43] David Gotz, Shun Sun, and Nan Cao. 2016. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 85–95.

[44] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.

[45] Ken Gu, Madeleine Grunde-McLaughlin, Andrew McNutt, Jeffrey Heer, and Tim Althoff. 2024. How do data analysts respond to ai assistance? a wizard-of-oz study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[46] Alon Halevy, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Goods: Organizing Google's Datasets. In *Proceedings of the 2016 International Conference on Management of Data*. Association for Computing Machinery, New York, NY, USA, 795–806. doi:10.1145/2882903.2903730

[47] Kelli Ham. 2013. OpenRefine (version 2.5). http://openrefine. org. Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association: JMLA* 101, 3 (2013), 233. doi:10.3163/1536-5050.101.3.020

[48] E Tory Higgins. 1996. Activation: Accessibility, and salience. *Social psychology: Handbook of basic principles* (1996), 133–168.

[49] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[50] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 108.

[51] A. Jović, K. Brkić, and N. Bogunović. 2015. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 1200–1205. doi:10.1109/MIPRO.2015.7160458

[52] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3363–3372.

[53] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI '12)*. Association for Computing Machinery, New York, NY, USA, 547–554. doi:10.1145/2254556.2254659

[54] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. *Visual analytics: Definition, process, and challenges*. Springer.

[55] Antino Kim, Mochen Yang, and Jingjing Zhang. 2023. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–36.

[56] Riikka Koulu. 2020. Human control over automation: EU policy and AI ethics. *Eur. J. Legal Stud.* 12 (2020), 9.

[57] Ernest K Lai. 2014. Expert advice for amateurs. *Journal of Economic Behavior & Organization* 103 (2014), 1–16.

[58] Doris Läpple and Bradford L Barham. 2019. How do learning ability, advice from experts and peers shape decision making? *Journal of Behavioral and Experimental Economics* 80 (2019), 92–107.

[59] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM computing surveys (CSUR)* 50, 6 (2017), 1–45.

[60] Chien-Huang Lin, Wen-Hsien Huang, and Marcel Zeelenberg. 2006. Multiple reference points in investor regret. *Journal of Economic Psychology* 27, 6 (2006), 781–792.

[61] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–13.

[62] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[63] Graham Loomes and Robert Sugden. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal* 92, 368 (1982), 805–824.

[64] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[65] Melinda Mangin and S Stoelinga. 2011. Peer? expert. *Journal of staff development* 32, 3 (2011), 48–52.

[66] Thorsten May, Andreas Bannach, James Davey, Tobias Ruppert, and Jörn Kohlhammer. 2011. Guiding feature subset selection with an interactive visualization. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 111–120.

[67] Barbara A Mellers, Alan Schwartz, and Alan DJ Cooke. 1998. Judgment and decision making. *Annual review of psychology* 49, 1 (1998), 447–477.

[68] Tom Meservy, Kelly J Fadel, Matthew L Jensen, and Michael Matthews. 2021. Searching for Expert or Peer Advice in Online Forums.. In *AMCIS*.

[69] Dar Meshi, Guido Biele, Christoph W Korn, and Hauke R Heekeren. 2012. How expert advice influences decision making. *PLoS One* 7, 11 (2012), e49748.

[70] Renée J Miller, Fatemeh Nargesian, Erkang Zhu, Christina Christodoulakis, Ken Q Pu, and Periklis Andritsos. 2018. Making Open Data Transparent: Data Discovery on Open Data. *IEEE Data Engineering Bulletin* 41, 2 (2018), 59–70. http://sites.computer.org/debull/A18june/p59.pdf

[71] Ivan Moscati. 2018. *Measuring utility: From the marginal revolution to behavioral economics*. Oxford Studies in History of E.

[72] Arpit Narechania, Surya Chakraborty, Shivam Agarwal, Atanu R Sinha, Ryan A Rossi, Fan Du, Jane Hoffswell, Shunan Guo, Eunyee Koh, Alex Endert, et al. 2023. DataCockpit: A Toolkit for Data Lake Navigation and Monitoring Utilizing Quality and Usage Information. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 5305–5310.

[73] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2022. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 1009–1018. doi:10.1109/TVCG.2021.3114827

[74] Arpit Narechania, Fan Du, Atanu R Sinha, Ryan Rossi, Jane Hoffswell, Shunan Guo, Eunyee Koh, Shamkant B Navathe, and Alex Endert. 2023. Datapilot: Utilizing quality and usage information for subset selection during visual data preparation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[75] Fatemeh Nargesian, Ken Q. Pu, Erkang Zhu, Bahar Ghadiri Bashardoost, and Renée J. Miller. 2020. Organizing Data Lakes for Navigation *(SIGMOD '20)*. Association for Computing Machinery, New York, NY, USA, 1939–1950. doi:10.1145/3318464.3380605

[76] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data Lake Management: Challenges and Opportunities. *Proc. VLDB Endow.* 12, 12 (aug 2019), 1986–1989. doi:10.14778/3352063.3352116

[77] Jamal R Paden, Arpit Narechania, and Alex Endert. 2024. BiasBuzz: Combining Visual Guidance with Haptic Feedback to Increase Awareness of Analytic Behavior during Visual Data Analysis. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7. doi:10.1145/3613905.3651064

[78] Scott Page. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.

[79] Ignacio Pérez-Messina, Davide Ceneda, Mennatallah El-Assady, Silvia Miksch, and Fabian Sperrle. 2022. A Typology of Guidance Tasks in Mixed-Initiative Visual Analytics Environments. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 465–476.

[80] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[81] Robert Richards. 2006. Representational state transfer (rest). In *Pro PHP XML and web services*. Springer, 633–672.

[82] Sidney L Smith and Jane N Mosier. 1986. *Guidelines for designing user interface software*. Citeseer.

[83] Fabian Sperrle, Davide Ceneda, and Mennatallah El-Assady. 2022. Lotse: A practical framework for guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1124–1134.

[84] Fabian Sperrle, Mennatallah El-Assady, Alessio Arleo, and Davide Ceneda. 2024. A Wizard of Oz Study of Guidance Strategies and Dynamics. *IEEE Transactions on Visualization and Computer Graphics* (2024), 1–15. doi:10.1109/TVCG.2024.3418782

[85] Fabian Sperrle, Astrik Jeitler, Jürgen Bernard, Daniel Keim, and Mennatallah El-Assady. 2021. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics* (2021).

[86] Fabian Sperrle, Astrik Veronika Jeitler, Jürgen Bernard, Daniel A Keim, and Mennatallah El-Assady. 2020. Learning and teaching in co-adaptive guidance for mixed-initiative visual analytics. In *EuroVis Workshop on Visual Analytics (EuroVA)*. 61–65.

[87] Fabian Sperrle, Hanna Schäfer, Daniel Keim, and Mennatallah El-Assady. 2021. Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 215–226.

[88] Fabian Sperrle, Hanna Schäfer, Daniel Keim, and Mennatallah El-Assady. 2021. Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 215–226.

[89] Anselm Strauss and Juliet Corbin. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications. doi:10.4135/9781452230153

[90] Tableau. 2022. Tableau Prep. Retrieved May 25, 2022 from https://www.tableau.com/products/prep

[91] Jennifer Thom-Santelli, Dan Cosley, and Geri Gay. 2010. What do you know? Experts, novices and territoriality in collaborative systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1685–1694.

[92] A Caroline Tynan and Jennifer Drayton. 1987. Market segmentation. *Journal of marketing management* 2, 3 (1987), 301–335.
[93] Konstantina Tzini and Kriti Jain. 2018. The role of anticipated regret in advice taking. *Journal of Behavioral Decision Making* 31, 1 (2018), 74–86.
[94] Ali Vedadi, Merrill Warkentin, and Alan Dennis. 2021. Herd behavior in information security decision-making. *Information & Management* 58, 8 (2021), 103526.
[95] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 763–777.
[96] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2022. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 966–975. doi:10.1109/TVCG.2021.3114862
[97] Elke U Weber and Eric J Johnson. 2009. Mindful judgment and decision making. *Annual review of psychology* 60, 1 (2009), 53–85.
[98] Ilan Yaniv. 2004. Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes* 93, 1 (2004), 1–13.
[99] Marcel Zeelenberg, Wilco W Van Dijk, Antony SR Manstead, and Joopvan der Pligt. 1998. The experience of regret and disappointment. *Cognition & Emotion* 12, 2 (1998), 221–230.
[100] Shichao Zhang, Chengqi Zhang, and Qiang Yang. 2003. Data preparation for data mining. *Applied Artificial Intelligence* 17, 5-6 (2003), 375–381. doi:10.1080/713827180 arXiv:https://doi.org/10.1080/713827180
[101] Yi Zhang and Zachary G. Ives. 2020. Finding Related Tables in Data Lakes for Interactive Data Science. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. Association for Computing Machinery, New York, NY, USA, 1951–1966.

A SUMMARY OF HYPOTHESES H1–H3

B SUMMARY OF HYPOTHESES H4–H5

C SUMMARY STATISTICS OF THE PRE-, AND POST- STUDY QUESTIONNAIRES

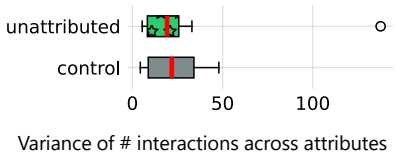
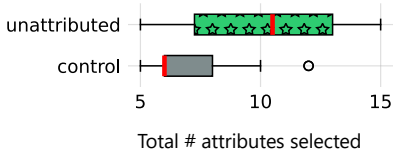
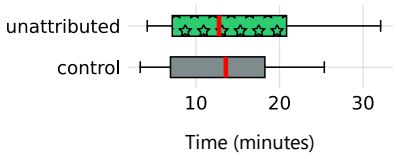
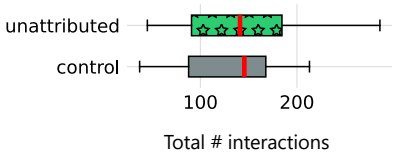
	Summary of Hypotheses Testing: H1–H3	Validation	*p*-value
H1	**Hypothesis**: **Unattributed** participants who receive guidance will *be less uncertain about their attribute selections* than **Control** participants who do not receive guidance. **Metric**: Variance of # interactions across all attributes. **Result**: **Unattributed** (median=19.19), **Control** (median=21.82)  **Statistical Test**: **Unattributed** < **Control**	✔	0.34
H2	**Hypothesis**: **Unattributed** participants who receive guidance will *select more attributes* than **Control** participants who do not receive guidance. **Metric**: Total # attributes selected at the end of the task. **Result**: **Unattributed** (median=10.5), **Control** (median=6)  **Statistical Test**: **Unattributed** > **Control**	✔	**0.001**
H3	**Hypothesis**: **Unattributed** participants who receive guidance will *take lesser time* to complete the task than **Control** participants who do not receive guidance. **Metric**: Total duration of the task (in minutes). **Result**: **Unattributed** (median=12.76), **Control** (median=13.57)  **Statistical Test**: **Unattributed** < **Control**	✔	0.67
	Metric: Total # interactions performed during the task. **Result**: **Unattributed** (median=141), **Control** (median=145.5)  **Statistical Test**: **Unattributed** < **Control**	✔	0.72

Table 4: Summary of hypotheses testing for hypotheses H1–H3 along with the statistical tests' Validation status, either validated with significance (✔), directionally consistent with hypothesis but not significant (✔), or directionally inconsistent with hypothesis (–), along with corresponding *p*-values. Statistically significant p-values are highlighted in bold and teal, based on pairwise one-sided Mann-Whitney U tests with Bonferroni correction between Unattributed and Control.

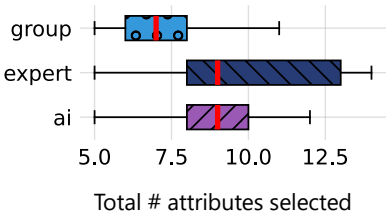
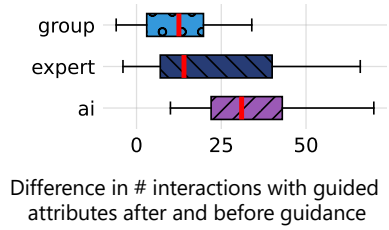
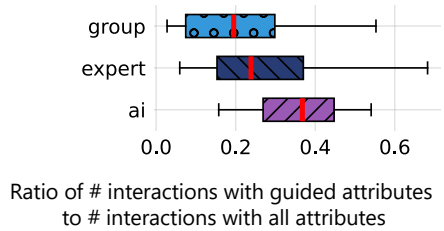
	Summary of Hypotheses Testing: H4–H5	Validation	*p*-value
H4	**Hypothesis**: Participants will find guidance to *have more utility* when it comes from **Expert** > **AI** > **Group**		
	Metric: Total # attributes selected at the end of the task. **Result**: **Expert** (median=9), **AI** (median=9), **Group** (median=7)		
			
	Statistical Test:		
	Expert > **AI**	–	1.0
	Expert > **Group**	✔	**0.01**
	AI > **Group**	✔	**0.04**
H5	**Hypothesis**: Participants will *verify the guidance more* when it comes from **AI** > **Expert** > **Group**.		
	Metric: Difference in # interactions with guided attributes after and before guidance. **Result**: **AI** (median=31), **Expert** (median=14), **Group** (median=12.5)		
			
	Statistical Test:		
	AI > **Expert**	✔	0.11
	AI > **Group**	✔	**0.003**
	Expert > **Group**	✔	0.59
	Metric: Ratio of # interactions with guided attributes to # interactions with all attributes. **Result**: **AI** (median=0.36), **Expert** (median=0.23), **Group** (median=0.19)		
			
	Statistical Test:		
	AI > **Expert**	✔	0.27
	AI > **Group**	✔	**0.01**
	Expert > **Group**	✔	0.35

Table 5: Summary of hypotheses testing for hypotheses H4–H5 along with the statistical tests' Validation status, either validated with significance (✔), directionally consistent with hypothesis but not significant (✔), or directionally inconsistent with hypothesis (–), along with corresponding *p*-values. Statistically significant p-values are highlighted in bold and teal, based on pairwise one-sided Mann-Whitney U tests with Bonferroni correction between AI, Expert, and Group.

Table 6: Median scores about questions asked during the two pre-study and one post-study questionnaires, on a scale from 1 (Low/Disagree/None at all) to 7 (High/Agree/A lot). Bold and underline indicate the largest values in each row.

Metric	AI	Expert	Group	Unattributed	Control
	Median	Median	Median	Median	Median
Pre Study Questionnaire 1: Questions about Guidance. Asked to all participants. Scale: 1 (Disagree) - 7 (Agree).					
Guidance Can Be **Beneficial**	**6.0**	**6.0**	**6.0**	**6.0**	**6.0**
Pre Study Questionnaire 2: Questions about Guidance. Not asked to **Control**. Scale: 1 (Disagree) - 7 (Agree).					
Guidance Can Be **Reliable**	4.0	5.0	5.0	**6.0**	-
Guidance Can Increase **Confidence**	4.0	**6.0**	4.0	**6.0**	-
Guidance Can Add **Value**	5.0	5.0	5.0	**6.0**	-
Post Study Questionnaire: Questions on Guidance Received during Task. Not asked to **Control**. Scale: 1 (Disagree) - 7 (Agree).					
Guidance Was **Trustworthy**	3.0	4.0	3.0	**5.0**	-
Guidance Came From A **Knowledgeable** Source	**5.0**	4.0	**5.0**	4.0	-
Guidance Was **Reliable**	3.0	4.0	3.0	**4.5**	-
Guidance Increased **Confidence**	3.0	3.0	3.0	**4.0**	-
Guidance Was **Relevant**	4.0	4.0	3.0	**5.0**	-
Guidance Helped Avoid Potential **Pitfalls**	2.0	**3.0**	**3.0**	**3.0**	-
Guidance Gave Valuable **Suggestions**	**5.0**	3.0	3.0	**5.0**	-
Guidance **Added Value**	**5.0**	3.0	3.0	3.5	-
Guidance Was **Appropriate**	4.0	4.0	4.0	**5.5**	-
Post Study Questionnaire: Questions on Guidance Received during Task. Not asked to **Control**. Scale: 1 (None at all) - 7 (A lot).					
How Much Did You **Rely** on Guidance	**3.0**	2.0	2.0	**3.0**	-
How Much Did You **Regret** Relying on Guidance	**5.0**	2.0	2.0	3.5	-
Post Study Questionnaire: Questions about Guidance. Asked only to **Control**. Scale: 1 (Disagree) - 7 (Agree).					
Guidance Can Be **Trustworthy**	-	-	-	-	**5.0**
Guidance Can Show It Came From A **Knowledgeable** Source	-	-	-	-	**6.0**
Guidance Can Be **Reliable**	-	-	-	-	**5.0**
Guidance Can Increase **Confidence**	-	-	-	-	**7.0**
Guidance Can Be **Relevant**	-	-	-	-	**6.0**
Guidance Can Help Avoid Potential **Pitfalls**	-	-	-	-	**6.0**
Guidance Can Give Valuable **Suggestions**	-	-	-	-	**6.5**
Guidance Can Add **Value**	-	-	-	-	**6.0**
Guidance Can Be **Appropriate**	-	-	-	-	**6.0**
Post Study Questionnaire: (What-If) Questions about Guidance. Condition-specific. Scale: 1 (Low) - 7 (High).					
How Much **Faith** Would You Have If Guidance From **AI**	-	**4.0**	**4.0**	3.0	3.5
How Much **Faith** Would You Have If Guidance From **Expert**	6.0	-	5.0	6.0	**6.5**
How Much **Faith** Would You Have If Guidance From **Group**	**6.0**	**6.0**	-	**6.0**	**6.0**
How Much Would You **Prefer** Guidance If It Comes From **AI**	-	-	-	**5.0**	**5.0**
How Much Would You **Prefer** Guidance If It Comes From **Expert**	-	-	-	6.0	**6.5**
How Much Would You **Prefer** Guidance If It Comes From **Group**	-	-	-	**6.0**	**6.0**
How Much Would You **Rely** on Guidance If It Comes From **AI**	-	-	-	4.0	**4.5**
How Much Would You **Rely** on Guidance If It Comes From **Expert**	-	-	-	**6.0**	**6.0**
How Much Would You **Rely** on Guidance If It Comes From **Group**	-	-	-	**6.0**	**6.0**
Post Study Questionnaire: Questions about Prior Experiences with Guidance. Scale: 1 (Low) - 7 (High).					
How Often Have You **Sought** Guidance In Life	-	-	-	**5.0**	-
How Often Have You **Sought** Guidance In Life From **AI**	3.0	4.0	4.0	4.0	**5.0**
How Often Have You **Sought** Guidance In Life From **Expert**	**5.0**	**5.0**	4.0	3.5	4.0
How Often Have You **Sought** Guidance In Life From **Group**	3.0	**5.0**	2.0	3.0	4.0