Investigating User Estimation of Missing Data in Visual Analysis

Maoyuan Sun
smaoyuan@niu.edu
Northern Illinois University
DeKalb, Illinois, USA

Yuanxin Wang
y2587wang@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

Courtney Bolton
cbolton@niu.edu
Northern Illinois University
DeKalb, Illinois, USA

Yue Ma
myue@niu.edu
Northern Illinois University
DeKalb, Illinois, USA

Tianyi Li
li4251@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Jian Zhao
jianzhao@uwaterloo.ca
University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

Missing data is a pervasive issue in real-world analytics, stemming from a multitude of factors (e.g., device malfunctions and network disruptions), making it a ubiquitous challenge in many domains. Misperception of missing data impacts decision-making and causes severe consequences. To mitigate risks from missing data and facilitate proper handling, computing methods (e.g., imputation) have been studied, which often culminate in the visual representation of data for analysts to further check. Yet, the influence of these computed representations on user judgment regarding missing data remains unclear. To study potential influencing factors and their impact on user judgment, we conducted a crowdsourcing study. We controlled 4 factors: the *distribution*, *imputation*, and *visualization* of missing data, and the *prior knowledge* of data. We compared users' estimations of missing data with computed imputations under different combinations of these factors. Our results offer useful guidance for visualizing missing data and their imputations, which informs future studies on developing trustworthy computing methods for visual analysis of missing data.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in visualization**.

KEYWORDS

Missing data, time series, visual analysis

ACM Reference Format:
Maoyuan Sun, Yuanxin Wang, Courtney Bolton, Yue Ma, Tianyi Li, and Jian Zhao. 2024. Investigating User Estimation of Missing Data in Visual Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Analysts often need to handle data with missing values in many domains. For example, sensor failures in a weather station can lose recorded temperature information [56]. For privacy concerns, when answering a questionnaire, participants may intentionally leave certain sensitive information (i.e., gender, race, and identity ID) as blank [18, 55]. Moreover, regarding time, future data values can be considered as a special type of missing data (for now they are not present, but existent in the future), and they are commonly analyzed and predicted in domains that involve temporal measurements (e.g., weather forecast, and stock or housing market prediction).

Missing data, if ignored or not responsibly handled, can lead to severe consequences. For example, decisions made with incomplete clinical data can waste a huge amount of public money for investing in stockpiling drugs with no clear benefits [8]. To avoid these, it calls for a comprehensive understanding of the human reasoning process when analyzing incomplete data, and techniques for supporting sensemaking with missing data. Computing methods have been proposed to mitigate the risk and facilitate the handling of missing data. For example, missing data imputation [24, 27] or prediction [49, 67] aims to replace missing data with some "best, reasonable inference" based on existing data [60]. Such imputations or predictions are often computed with uncertainty (e.g., probability and confidence interval) [20]. To reveal the uncertainty to users, a variety of visualizations have been studied [16, 38, 53].

While prior works have studied several visual encodings of missing data [11, 61] and user-perceived data quality [22, 61], there still lacks an in-depth understanding of how different visualizations affect users' reasoning when missing data presents in their analysis. Moreover, computational models for the imputation or prediction of missing data can be sometimes inaccurate. It remains unclear if and how the visualized computational results affect users' judgment of missing data. Further, users could have different rationales based on their expertise or prior knowledge about data, which may not match the computed imputations or predictions of missing data.

To investigate these problems, we contributed a controlled experiment on Amazon Mechanical Turk (MTurk) [1] to study factors impacting on users' judgment of missing data in visual analysis: specifically, users' estimations of missing data. While missing data is commonly estimated computationally, for making data-driven decisions, computed results often need to be shown to analysts for further verifications. As missing data implies certain unknown space, due to the loss of information, human judgment is essential

for leveraging domain expertise and assuring ethical and responsible data handling. Moreover, human judgment can be transformed into certain inputs for computational models and has been considered and used to improve the performance of computations [10]. Thus, we focus on users' estimations of missing data in this study.

We controlled four factors in the study: the *distribution* of missing data, the *imputation* of missing data, the *visualization* of missing data, and the *prior knowledge* of data. They correspond to four major aspects in a data analysis process: *data*, *computation*, *interface*, and *user*, respectively. We controlled these factors based on rationales derived from common patterns in prior work on missing data, its analysis, and visualizations [13, 29–33, 35, 41, 61]. We used a between-subjects design and recruited 630 MTurk workers for the study. We collected data on metrics related to the accuracy, tendency, and consistency of participants' estimations of missing data, their task efficiency, and self-judgment (i.e., level of confidence). We found that all the factors had a significant impact on the accuracy, tendency, and consistency of participants' estimations. We also observed that the prior knowledge significantly affected the participants' task time; and imputation and visualization showed a significant impact on participants' task effort. Lastly, participants' confidence in their estimations seems to be influenced by all the factors except, surprisingly, prior knowledge.

Based on the results of our study, we contributed empirical findings on guiding the design of missing data visualization and imputation as well as trustworthy computational interfaces with missing data in general. Specifically, designers should consider the distribution of missing data to design interfaces for making judgments, and provide possible data context to support analysis. Second, users collaboratively working with computations (i.e., given imputations) may lead to better estimations, and designers should consider different cases of imputation accuracy to apply appropriate visualizations. Third, users' prior knowledge is generally helpful, but should be treated differently, when there might be a conflict with computed imputations, and externalization of such knowledge might benefit users for employing it for decision making.

In summary, this work highlights two major contributions: 1) a controlled experiment ($N = 630$) investigating how data, computation, interface, and human factors (correspondingly, distribution, imputation, visualization, and prior knowledge) impact the perception and decision-making of missing data; and 2) in-depth empirical knowledge including design guidance for visualizing missing data and their imputations as well as design implications for building trustworthy computing methods and interfaces with missing data.

2 BACKGROUND AND RELATED WORK

2.1 Missing Data

Missing data occurs due to many reasons, often referring to the loss of recorded data values (e.g., a sensor fails to capture the weather temperature) [46, 63]. In a broad sense, missing data can go beyond missing values, which, instead, covers multiple aspects of data. Sun et al. have discussed the missingness of data in 5 aspects [64]: *data values*, *data attributes*, *data records*, *data relationships* (e.g., missing links in networks [67, 68] and incomplete bicliques with missing edges [40, 62, 65]), and *data usage* (e.g., selecting what parts of data to use). Among these, missing data values has been heavily studied, as it appears in real-world analyses in a variety of domains (e.g., bioinformatics [45], database [44], social network [43], and survey [18]). This drives the focus of our study on missing data values.

A value-oriented, missingness mechanism has been studied [30–32, 57]. It categorizes missing data values into three types, based on the probability that missingness depends on observed data and missing data: 1) *missing completely at random* (*MCAR*: the missingness probability relies on neither observed nor missing data), 2) *missing at random* (*MAR*: the missingness probability relies on observed data), and 3) *missing not at random* (*MNAR*: the missingness probability depends on missing data). However, it is practically challenging to identify which pattern missing data belongs to, as it requires certain awareness of an unknown space [23].

These theories and studies motivate our work, but none of them provides an in-depth empirical understanding about what and how different factors impact users' judgement of missing data values.

2.2 Analysis with Missing Data

To support handling missing data (especially missing values), a collection of analysis methods have been developed [12, 34], and a detailed discussion about them can be found in [9, 28, 46]. Analysis strategies to handle missing values fall into three major groups. First, *imputation* aims to reasonably estimate missing data and replace them with the estimations [24, 27, 45]. While using statistical or machine learning methods [58] to estimate missing values varies, it holds a belief that filling in gaps in data can give a more reasonable analysis than ignoring them. The quality of imputations are often evaluated by comparing estimated values with true values. This informs us setting a control on imputed missing data values.

Omission treats missing values as noises in data. Instead of attempting to fixing holes in data, it highlights removing data records with missing values. It considers them as "low-quality" data and hypothesizes that removing them improves data quality, which can benefit analysis. However, this reduces the number of samples to analyze (e.g., likewise detection [50]). Due to omitting possibly useful information, it may lead to a biased analysis, especially for the missing pattern of MNAR (e.g., participants intentionally leave sensitive information in a survey blank) [18].

Third, *adaptation* highlights performing an analysis by adapting to incomplete data. It neither replaces missing values with their estimations, nor removes them. Instead, it uses incomplete data for analysis (e.g., the expectation-maximization algorithm [25]). Compared to the other two strategies, existing data is neither augmented nor reduced, but in this strategy, individual missing values catch less attention. It tries to identify and enlarge a good likelihood of complete data (i.e., observed data), by assuming that this likelihood may potentially cover the missingness in data.

While these strategies support analysis with missing data, there still lacks thoroughly performed studies on investigating users' judgement of missing data, when there is imputation or not, especially with the existence of other factors (e.g., missing data distribution and visualization). This motivates us to develop the research questions in this study. As omission removes missing data and performing adaptation is too complex (estimating a likelihood, instead of individual data values), we focus on imputation in this work.

2.3 Missing Data Visualization

Three designs have been studied to show missing data and uncertainty that may come from data missingness and statistical analysis results. First, *contrast*-oriented visual encoding is a straightforward approach, which visually highlights missing data (e.g., using empty space). It maps each observed data to a visual mark and displays it in a 2D space based on certain layouts (e.g., matrix, table, and treemap). A typical example is a table or matrix with empty cells. Based on the patterns shown by empty cells, the missingness in data (MCAR, MAR, and MNAR) [57], can be revealed [30–32]. It can also encode identified missing data with visually salient marks to get separated from those for observed data (e.g., using color). Song and Szafir applied this design to bar charts and line charts to show missing data (e.g., have a few empty bars and gaps between line segments), and studied user perceived data quality [61].

Second, *profiling*-oriented visual encoding considers missing data in the context of data population. It is a relatively implicit strategy to show data missingness, in a way of expressing uncertainty. It reveals missingness in a visualization that profiles or summarizes all observed data (e.g., distribution plot, violin plot, gradient plot, box plot, and more visualizations discussed in [52]). Instead of focusing on individual missing values, it emphasizes more on the impact of missing data on the whole data population that is obtained via statistical methods, so it shows results of data profiling [51]. Due to this, data missingness can only be roughly perceived, but not precisely identified (e.g., getting an overview of the distribution of data, but not sure of where exactly the missingness locates).

Third, *anchoring*-oriented visual encoding highlights creating visual anchors to help users understand summarized, observed data. It shows uncertainty, which regards observed data population and may involve data missingness. However, it transforms a 1D / 2D range-based encoding into a collection of visual marks. The spatial arrangement of them can lead to a perceived 1D / 2D range, similar to the concept of unit visualization [54]. Typical examples are quantile dot plot [29] and ensemble plot [47, 48]. Thus, it not only helps user perceive an overall pattern of data, but also offers marks that may anchor user attention. Same as the profiling-oriented design, data missingness is vaguely revealed, but differently, the used collection of marks breaks a continuous range into discrete pieces.

These visualizations inspired our study design, particularly on the control of missing data visualization. However, none of them focus on systematically studying possible impact on users' judgement of missing data in such visualizations with the presence of other factors (e.g., missing data distribution, imputation, and prior knowledge of data). This is the gap that we aim to address.

3 METHOD

3.1 Research Questions

We developed the following set of research questions to study key factors that may impact users' estimations of missing data.

Q1. Distribution: how do different distributions of missing data impact users' estimations of missing data?

Q2. Imputation: how are users' estimations of missing data affected by the presence and accuracy of missing data imputations?

Q3. Visualization: how do different forms of visual representations of missing data affect users' estimations?

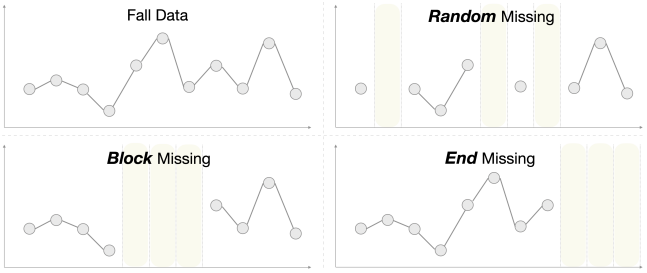

Figure 1: Three different controls on the distribution of missing data by sampling them from a given dataset.

Q4. Prior knowledge: how are users' estimations of missing data affected when they have prior knowledge about data?

We consider these factors, as they correspond to four key aspects in data analysis: *data*, *computation*, *interface*, and *user*. Specifically, *distribution* characterizes where and how the data is incomplete. *Imputation* refers to the computation methods for estimating the value of missing data. *Visualization* is the interface that communicates datasets with missing values. *Prior knowledge* regards the user expertise and experience with the data and the analysis tasks.

3.2 Impacting Factors

To investigate these questions, we chose to study time-series data (a series of data points that are organized in time order) [19]. This type of data widely permeates people's daily life (e.g., weather forecasts, and stock and housing market trends). The widespread acquaintance with such data helps in reducing disparities in data familiarity among participants. In addition, given the prevalence of time-series data [19], using it to conduct the study may extend the applicability of the findings across multiple domains. Furthermore, the loss of data commonly appears when collecting time-series data in practice, and imputations or predictions are often applied to it [66]. Based on these considerations, we choose to use time-series data in our study. Below we discuss the rationale for selecting the four factors studied in this work and their specific controls.

3.2.1 The Distribution of Missing Data. The distribution of missing data refers to how data points are absent from a dataset. Prior work has shown that there are multiple patterns of missing data, especially when considering the missingness based on observed data [30–32]. *Different distributions of missing data may impact users' judgment of missing data* (H1). For example, a user's estimation of a missing data value may be more accurate with neighboring data points, which offer useful context, compared to when adjacent data are also missing. To examine this, we manipulate the missing data distribution in three distinct ways to determine which data points are missing, as summarized below and exemplified in Figure 1.

- *Random missing* (D_{random}): m data points is randomly selected from a dataset as missing data.
- *Block missing* (D_{block}): m consecutive data points that are not the last m ones in a dataset is randomly selected as missing data.
- *End missing* (D_{end}): m consecutive data points at the end of a dataset is selected as missing data.

The dataset has time series data and $1 \leq m \leq |dataset|$, where $|\cdot|$ denotes the cardinality. We pick the three distributions of missing data by considering real-world use cases. D_{random} corresponds to

cases where unexpected events happen that lead to the loss of data records (e.g., a strong wind blow leaves up, covering a surveillance camera for a few seconds, which happens multiple times in a day). D_{block} considers cases where collected data within certain time periods get lost. For example, due to extremely cold weather, a sensor fails to capture temperature in the past two hours. D_{end} regards cases of predicting future data, which is considered as a special type of missing data (i.e., missing for the present time, from a future point of view). It is heavily used in many real-world applications (e.g., weather forecast, stock market analysis, and disease control).

3.2.2 The Imputation of Missing Data. As missing data is often handled in a way of imputation, *whether or not giving users imputations of missing data, and, if given, different levels of imputation accuracy may impact users' judgments of missing data* (H2). In practice, it is hard to get perfect heuristics to estimate whether an imputation is good or bad, and the general advice is to use an imputation method providing values that are as stable as possible. However, the anchoring effect [33] suggests that provided imputations, regardless of their accuracy, may sway users' judgments—potentially leading to more consistent evaluations aligned with the provided estimates, especially when those imputations are more precise. To study this factor, we set three controls on the imputation of missing data by considering their true values (i.e., the values recorded in a dataset).

- *No imputation* (I_{no}): imputations of missing data are not given.
- *High-accuracy imputation* (I_{high}): we use the *true* values of selected missing data to simulate high-accuracy imputations.
- *Low-accuracy imputation* (I_{low}): we set low-accuracy imputations of missing data as values that differ from their *true* values within a certain range. Section 3.4.1 gives a detailed discussion of computing a low-accuracy imputation for missing data.

By comparing I_{no} with I_{high} and I_{low}, we can study the possible impact of imputations on users' estimations of missing data. Moreover, by comparing I_{high} with I_{low}, we can further verify H2. For example, when imputations of missing data are provided with I_{high} and I_{low}, we can test whether users' estimations of missing data tend toward the given imputations, even when the accuracy is low.

3.2.3 The Visualization of Missing Data. Prior work found that missing data visualizations (e.g., highlighting and information removal) can impact user-perceived data quality [61]. However, it remains unclear whether (and how) these techniques affect users' estimations of missing data when the imputed values exist or not. We hypothesize that *visualizations of missing data may limit users' estimations of missing data* (H3). As the imputation of missing data can be revealed in a bounded way (e.g., error bars) [61], users' estimations of missing data may be constrained by such visual bounds. In other words, users may not estimate missing data outside the bounds in such visualizations. To study this, we use three visualization techniques, shown in Figure 2.

- *Empty visualization* (V_{empty}): missing data is revealed as *empty* space.
- *Continuous visualization* ($V_{continuous}$): missing data is shown as imputed values with error bars (representing confidence intervals), which sets a *continuous* range indicating where the missing data possibly locates.
- *Discrete visualization* ($V_{discrete}$): missing data is shown as a set of *discrete* points corresponding to meaningful confidence levels that an imputation of this missing data may have. This is inspired by the design and benefit of using quantile dot plots [29, 41].

Table 1: Example of collected daily temperature data.

Day	Max ($^\circ F$)	Min ($^\circ F$)	Average ($^\circ F$)
06/11/2018	67.4	51.2	59.2
06/12/2018	76.6	49.1	62.9
...	...	...	...

By comparing V_{empty} (*without* visual bounds) with $V_{continuous}$ and $V_{discrete}$ (*with* visual bounds), it is possible to test whether visualized imputations of missing data limit user estimations of them. Moreover, by comparing $V_{continuous}$ with $V_{discrete}$, it helps us understand whether different forms of visual bounds impact user judgment of missing data differently. For example, would user estimations of missing data values fall in the range of error bars in $V_{continuous}$, but be anchored by certain points in $V_{discrete}$?

3.2.4 The Prior Knowledge of Data. Prior knowledge has been found playing a key role in developing analysis strategy and procedure [59] and significantly impacting decision-making [15]. Thus, *the prior knowledge of data could impact user estimations of missing data* (H4). Prior knowledge has been used to computationally handling missing data by improving imputation results [13, 35]. Thus, compared to those without prior knowledge, it is more likely that users, with prior knowledge of data, may estimate missing data more reasonably (e.g., closer to their true values). To verify this, we set two controls on the prior knowledge of data in this study.

- *With prior knowledge* (K_{with}): historical data is shown to users.
- *Without prior knowledge* ($K_{without}$): historical data is not given.

We use whether or not to show historical data to users to control their prior knowledge. If the historical data is presented for users to see, we consider that they have prior knowledge about the data; if not, we consider them without prior knowledge. By comparing users' estimations of missing data in K_{with} with those in $K_{without}$, we can study the impact of prior knowledge on users' estimations.

3.3 Experimental Dataset

As explained in Section 3.2, the study uses weather temperatures as the experimental dataset. We collected the weather data from the Blue Hill Observatory and Science Center [2]. It includes the daily temperature record from the Greater Boston area from 2016 to 2020. For each day, the data record contains the maximum, minimum, and average temperature ($^\circ F$). In total, we collected the temperature information for 1827 days (60 months), organized in a table-based format. Table 1 shows a sample of the collected data.

Regarding the temperature data on each day, we used the average value (see the 4^{th} column in Table 1). Considering the nature of micro-tasks on MTurk (short time duration and attention span) [42] and possibly limited screen space available in the task interface, we divided the dataset into segments of 60 days for user tasks. It follows the setting used in a previous study [61], which can be manageable by participants within a reasonable time period. Specifically, we separated 60-month (2016–2020) data into smaller datasets, where each has data of two consecutive months (e.g., May and June in 2018). To assure that each separated dataset has 60-day data, for two consecutive months with more than 60 days, we cut off extra ones from the end of the second month; and for two months with

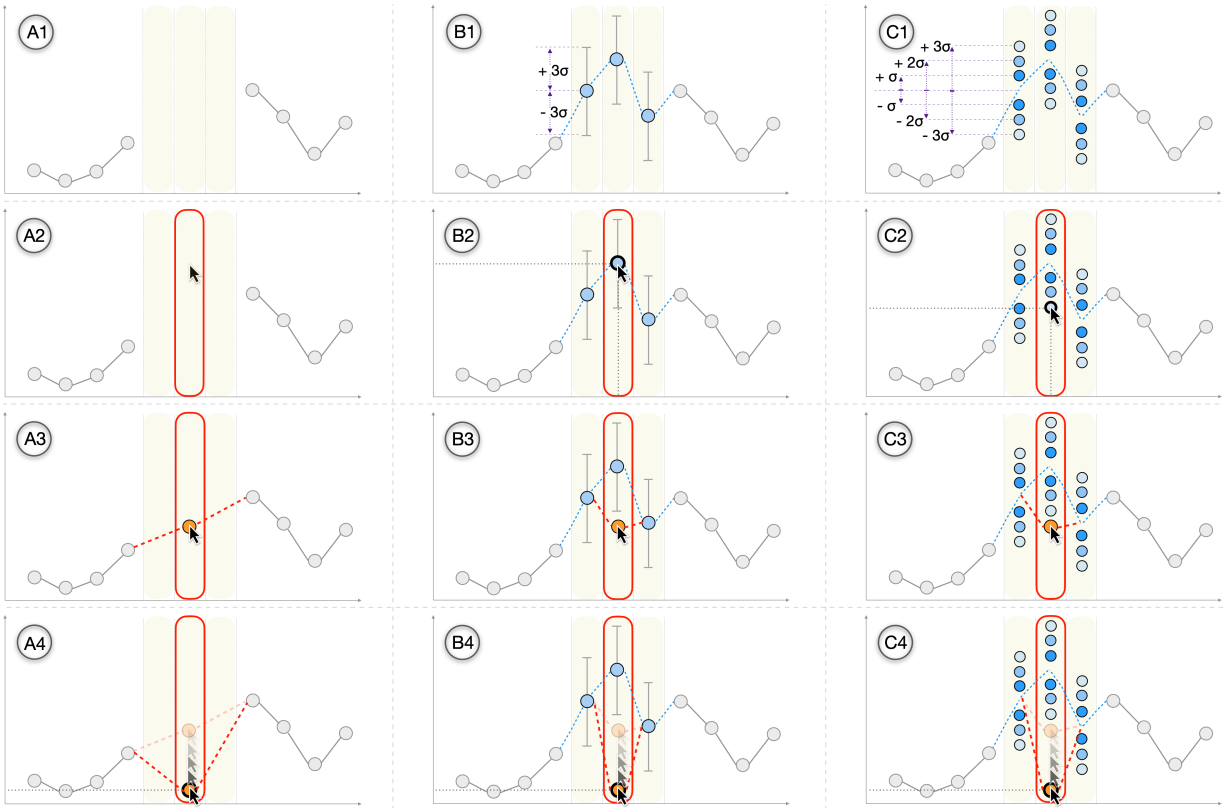

Figure 2: The design for three types of visualizations in our study, with their supported user interactions, to show missing data: *empty space* (A1-4), *error bars* (B1-4), and *discrete points* (C1-4). Grey circles are existing data, blue circles are computed imputations of missing data, and orange circles are user-made estimations of missing data. For each type of visualization, three interactions are offered: mouse-hovering (A2, B2, and C2), mouse-clicking to make an estimation of a missing data point (A3, B3, and C3), and dragging and moving to adjust a previously made estimation (A4, B4, and C4).

fewer than 60 days, we add extra days from the beginning of the following month. For example, for the dataset of July and August 2018 (62 days), we remove the data on August 30 and August 31, 2018; and for the dataset of January and February 2018 (59 days), we add the data on March 1, 2018. In total, we generated 59 unique datasets. Each has 60-day weather temperatures.

3.4 Experimental Software

3.4.1 Control of Impacting Factors. **Distribution.** For each dataset with 60-day temperature information, to control the distribution of missing data in it, we selected 10% of its data records (i.e., 6 days) as missing ones, which followed the setting of the prior work [61]. For each distribution condition, the selection was performed as follows:

- D_{random}: randomly selecting six days.
- D_{block}: randomly selecting one day from the first 54 days and then selecting its next five days.
- D_{end}: selecting the last 6 days.

Imputation. The I_{no} condition is set as not showing any imputations. The I_{high} condition is controlled by using the *true* values (i.e., temperatures recorded in the collected dataset) as imputations. We simulate the I_{low} condition with the following equation:

$$V_{false} = V_{true} + sign(1.645\sigma + r \cdot (2.576\sigma - 1.645\sigma)) \quad (1)$$

In Equation (1), V_{true} is the true value of a missing data, and σ is the *standard error* of the sample mean, where each sample is one of our generated datasets, including 60-day temperature data. The function $sign(\cdot)$ randomly generates a positive (+) or a negative (−) sign, and r is a random real number within the range $[0, 1]$. The constants, 1.645 and 2.576 are two normal critical values, corresponding to the confidence level at 90% and 99%, respectively. It assures that generated false values fall in the confidence interval [90%, 99%] based on true values. It mimics a reasonably well-performed imputation, which can generate values different from the true values of selected missing data, but not too far away from the true values. We choose this, as it better fits a real-world analysis scenario (people are more likely to work with reasonably well-performed imputation models than obviously bad-performed ones), compared to using false values that are significantly different from true values.

Visualization. We used a line chart to show the time-series data, where the x-axis corresponds to the date and the y-axis represents the average temperature, as a line chart is the most common method to visualize time-series data. As is shown in Figure 2, we used three different visualization techniques to present missing data: 1) using empty space (for V_{empty}), 2) using error bars (for $V_{continuous}$), and 3) using six discrete points (for $V_{discrete}$). They use the same visual encodings for existing data, which uses a grey circle (○) to present

an existing data point and links two neighboring points with a straight line ( ). They also use the same encoding for user-made estimations of missing data, which is represented as an orange circle ( ). Moreover, they all highlight the background of each missing data point ( ), which makes them visually salient.

For $V_{continuous}$, imputations of missing values are plotted as points on the chart, with error bars ( ). The length of an error bar is set as three times the standard error (σ) used in Equation (1), which covers 99.73% of the confidence interval for an imputation.

For $V_{discrete}$, imputations of missing values are not directly shown on the chart, but instead, six discrete circles are displayed ( ). To plot them, we determine the position of a missing data imputation by its date and imputed temperature. The relative distances between each of these six circles and the actual imputation data point are set by using the standard error (σ) in Equation (1), specifically as σ, 2σ, and 3σ, respectively. They, as a pair, respectively, cover 68.26% (between -σ and σ), 95.44% (between -2σ and 2σ), and 99.73% (between -3σ and 3σ) of the confidence interval for an imputation. The usage of such a discrete visualization is inspired by the design and benefit of quantile dot plots, which can relatively better support probabilistic estimations in real-world prediction usage scenarios [29, 41]. We chose not to use quantile dot plots, as directly embedding them into a line chart may confuse users. Moreover, for block missing and end missing, using quantile dotplots may generate a number of small dots neighboring each other. This may cause visual illusions or chart misinterpretations. Thus, instead of using multiple dots to reveal a probability, we use one circle, and to encode different probability values, we control the color saturation of circles with a linear mapping function.

Prior Knowledge. Based on the time of a dataset (e.g., May and June in 2018) selected for user tasks, we chose whether or not to show a line chart with the full data from a randomly selected year that share the same two months (e.g., May and June in 2016). If such a line chart is given before a user task, we consider that users have some prior knowledge of data (for K_{with}). If not, we consider that users have no prior knowledge (for $K_{without}$).

3.4.2 User Interaction and System Implementation. Figure 3 shows an example of the user interface used in our study. For conditions with prior knowledge of data, participants see historical data first (Figure 3 bottom) and then the interface for making estimations (Figure 3 top). For conditions without prior knowledge, participants only see the interface for making estimations.

We integrated all the above realization of impacting factors (see Section 3.4.1) into a visual analysis tool to conduct the study and collect user estimations and behaviors. The tool was developed with the Django web framework [4] and deployed on the Microsoft Azure cloud platform [5]. The back-end of the tool is implemented in Python with a PostgreSQL database [7]. It generates distributions, imputations, and visualizations of missing data for different experiment conditions, and communicates with the MTurk platform to retrieve and store necessary information about participants (e.g., MTurk worker ID) and record their task results. The front end of the tool is implemented with the Bootstrap framework [3] in HTML, CSS, and JavaScript. Our selected visualizations, discussed in Section 3.4.1, are developed with D3.js [17]. They provide an interactive user interface for participants to perform given tasks in

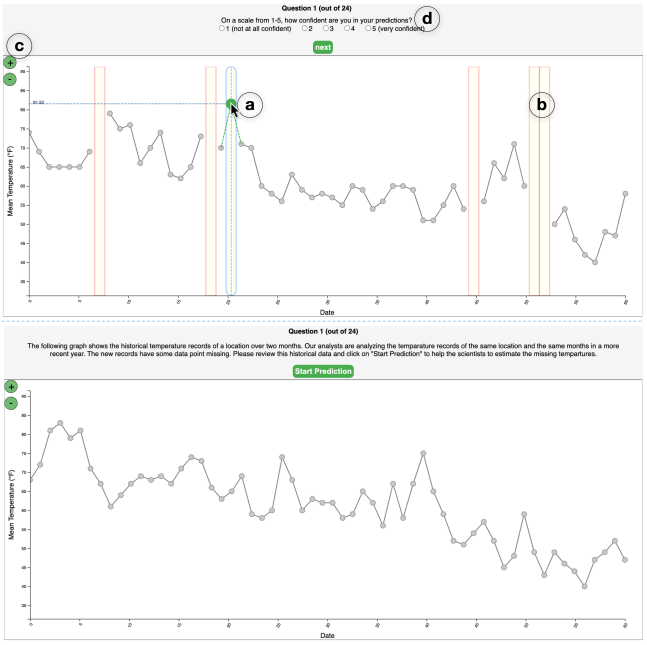

Figure 3: Examples of user interface used in the study. Top: the interface provided for users to make estimations of missing data, where (a) shows the estimation of missing data that a user is currently making; (b) reveals the missing data values to be estimated by a user; (c) are buttons for zooming in/out of the line chart; and (d) shows the radio buttons for a user to specify the confidence-level of his/her estimations. Bottom: the interface used for users to view historical data.

the study. Specifically, the front end allows users to see visualizations in given tasks, make estimations of missing data, select the level of confidence for their estimations, and answer a post-task questionnaire. Moreover, we enabled a set of interactive features on displayed visualizations, which supports participants in making estimations of missing data (see A2-4, B2-4 and C2-4 in Figure 2).

Three key user interactions were added to all the visualizations: *mouse-hovering*, *mouse-clicking*, and *dragging and moving*. When a user hovers the mouse in the area corresponding to a missing data point, its border gets highlighted (A2 in Figure 2). It shows the current focus of this user. Moreover, when a user hovers the mouse on presented imputations of missing data, besides the border highlighting, two intersecting, dotted lines are displayed (B2 and C2 in Figure 2). They support users in checking detailed information about an imputation. To make an estimation, users can click anywhere inside the area corresponding to this missing data (A3, B3, and C3 in Figure 2). After making an estimation, users can drag and move it to adjust the estimation (A4, B4, and C4 in Figure 2). While users are moving a previously made estimation, the two intersecting, dotted lines are shown to support the adjustment. Moreover, two buttons for zooming in and out of the chart are offered, in case the size of the screens that participants use for this study is small.

4 STUDY SETUP

Based on the four factors, our study has 30 experimental conditions (see Table 2). We used a between-subject design for the study.

Table 2: Summary of experimental conditions in the study.

	3 Distribution	D_{random}, D_{block}, and D_{end}
×	2 Prior Knowledge	K_{with} and $K_{without}$
×	5 Imputation & Visualization	(I_{no}, V_{empty})
		$(I_{high}, V_{continuous})$
		$(I_{high}, V_{discrete})$
		$(I_{low}, V_{continuous})$
		$(I_{low}, V_{discrete})$
Total	**30**	**conditions**

4.1 Participants

We recruited participants from the MTurk platform. MTurk workers, who have more than 100 approved HITs (Human Intelligence Tasks) with acceptance rates greater than 95%, were allowed to participate in this study. We considered a participant qualified, if he/she passed the screening questions, completed all given tasks, and finished a post-task questionnaire. Consequently, in total, 630 qualified participants (485 males and 145 females, 25-44 years old) were included in the data analysis. Each participant was paid $0.70 for finishing a screening task and would receive a bonus payment of $3.80, if they completed all given study tasks and the post-task questionnaire. The pay rate was set based on the US Federal minimum wage ($7.25 per hour) [6] and the time to finish given tasks. Thus, a participant, who finished all in this study, received $4.50.

4.2 Tasks

The user task was to estimate missing data in given visualizations and specify the confidence level of the estimations. For each task, a participant worked on a line chart with 60-day temperature data, in which the data of 6 selected days were missing. Participants were asked to make estimations of the missing temperatures for the 6 days and rate how confident they were for the 6 estimations they made by choosing one from a 5-point Likert scale.

Each participant was given 24 tasks that were under the same experiment condition. The tasks used 24 datasets, with the non-overlapping months in the same year, which were selected from our generated 59 datasets. For example, the dataset of May and June 2018 and that of June and July 2019 can be both used for the 24 tasks; while the dataset of June and July 2018 needs to be excluded if the dataset of May and June 2018 has already been selected for user tasks, as they share the month of June 2018. Moreover, for experiment conditions with prior knowledge (K_{with}), before each task, participants would see a line chart with full data during the same 60 days with an in-coming given task but in a different year. For these prior knowledge charts, we chose datasets from our generated ones but different from those selected for the 24 tasks. This assured that participants would not see the true values of missing data in their given tasks. To avoid possible order effects [37], the sequence of 24 tasks was randomized and the 6 selected missing data in each task were also changed. In total, in this study, 15,120 tasks were performed in which 90,720 missing data were estimated by 630 participants (21 participants for each condition).

4.3 Procedure

We posted the study as 30 HITs in MTurk. Each HIT is one experiment condition (Table 2). All HITs are launched under the same batch (using the same title and descriptions on MTurk website) to ensure random assignments. After accepting a HIT, participants (MTurk workers) were given a consent form. They were advised to return the HIT, if they did not accept it. Otherwise, they proceeded to the instruction page that included a 1.5-minute tutorial video for the corresponding experiment condition of the HIT and the study software. After reading the instruction and watching the tutorial video, participants could choose to take a screening test with four multiple-choice questions to check if they understand how to use the given tool to do study tasks (e.g. which of the following allows you to make estimations of missing values?). Participants can adjust their answers until they get the correct answers. Alternatively, they can submit the HIT after at least one attempt to answer all screening questions, and get the base payment of $0.70. Participants who passed the screening test would be directed to the study tasks. After finishing all 24 tasks, participants were given a post-task questionnaire about demographic information, education background, experiences of using computing methods for missing data imputations, and their belief in computed imputations of missing data. After finishing the questionnaire, participants submitted the HIT and would receive a bonus of $3.80.

The task interface is implemented to allow HIT submission in two conditions. First is after at least one attempt of the screening test. Second is after all the screening questions are correctly answered, all the experimental user tasks are completed, and all the questions in the post-task questionnaire are answered. All participants who submitted the HIT in the second condition are considered qualified (as defined in Section 4.1). The data collected from all qualified participants were included in the data analysis.

4.4 Data Collection, Measures, and Metrics

For each task, we recorded participants' initial and final estimations of missing data, all adjustments in between, their confidence rating, task completion time, and responses to a post-task questionnaire.

We used the following measures to assess the quality of participants' estimations, participants' perceived quality of given imputations, and the cost (i.e., time and effort) of making estimations.

- ***Accuracy*** *of user estimations of missing data, with respect to true values* (d_{true}): it reveals how close the participants' estimations are to the true values.
- ***Tendency*** *of user estimations of missing data to be influenced by imputations* (d_{impute}): this measures how close the participants' estimations are to the displayed imputations. The nearer they are, the more participants tend to follow presented imputations. This helps to reveal participants' perceived accuracy of given imputations (e.g., participants may be more likely to follow imputations that they perceive as highly accurate).
- ***Consistency*** *of user estimations of missing data, with respect to their true values* (c_{true}): it evaluates how dispersed participants' estimations of missing data are in relation to true values (I_{high}).
- ***Consistency*** *of user estimations of missing data, with respect to imputations* (c_{impute}): it measures how dispersed participants' estimations of missing data are in relation to their imputations (including both I_{high} and I_{low}).
- ***Time*** *duration of user estimation* (t_{user}): it refers to how much time a participant takes to estimate missing data.
- ***Effort*** *of user estimation* (f_{user}): this means the interactive effort that a participant takes to make an estimation. Specifically, it

measures the number of adjustments that a participant makes before getting to a finalized estimation.
- ***Confidence*** *of user estimations* (j_{user}): this is measured by the confidence rating that participants select in each task.

In summary, d_{true} measures the accuracy of participants' estimations; d_{impute} measures the tendency that participants follow given imputations for estimating missing data; c_{true}, and c_{impute} are consistency oriented measures; t_{user} and f_{user} are efficiency oriented measures; and j_{user} measures participants' self-judged confidence for their estimations. We used the Euclidean distance between participants' estimations of missing data and their true values and imputations to compute d_{true} and d_{impute}, respectively. For c_{true} and c_{impute}, we computed the standard deviation of d_{true} and d_{impute}, respectively, so lower values (for c_{true} and c_{impute}) indicate that participants' estimations are more consistent.

5 RESULTS

To answer our research questions, we performed quantitative analysis on the collected data. As part of the visualization factor work when an imputation (i.e., for I_{high} and I_{low}) exists, we used dummy coding [14] to combine the visualization and imputation factors, which generated five different combinations, following the notions in Table 2. For each factor (i.e., distribution, knowledge, and imputation & visualization), we performed a separate mixed-effects regression (i.e., the *lmer* package in R) by treating the factors as the fixed effects and individual participants as the random effect. To perform a more accurate test, we used ANOVA (i.e., the *anova()* in R) to compare each mixed effect model with a null model that only has participants as the random effects, to test if the corresponding factor was significant or not for a specific measure (see Section 4.4). The analysis results ($\chi^2(df)$ and $Pr(>\chi^2)$) are summarized in Table 3 and 4. Figure 4 shows the mean (μ) and 95% confidence intervals (CI) of the measures for each control of the factors.

5.1 Impact by The Distribution

The distribution of missing data shows a significant impact on the accuracy, tendency, consistency, and confidence of participants' estimations of missing data, but not on the efficiency measures of their estimations (see the second row in Table 3 and Table 4).

The significance of accuracy (d_{true}) is mainly reflected by the D_{block} and D_{end} conditions. The linear mixed model fitted by maximum likelihood (*LMER*) used the D_{block} condition as the reference and the intercept estimate (β) is 5.28, with 95% CI: $[4.95, 5.62]$ and $t = 31.13$. This suggests that in the D_{block} condition, the distance between participants' estimations of missing data and their true values (d_{true}) can be expected to be around 5.28, mostly falling within $[4.95, 5.62]$. Compared to the D_{block} condition, in the D_{end} condition, it is expected that participant estimations were less accurate, as d_{true} is expected to increase 0.98 (β), with 95% CI: $[0.76, 1.20]$, and $t = 8.73$. Moreover, for the D_{random} condition, compared to D_{block}, participants' estimations were more accurate ($\beta = -0.06$, 95% CI: $[-0.29, 0.16]$, $t = -0.54$). It is consistent with Figure 4 (A). This suggests that users may make **more accurate estimations of missing data when their neighboring data values are present**.

For the consistency of participants' estimations (c_{true}), with regard to true values (I_{high}), participant estimations of missing data were less consistent with true values in both D_{end} ($\beta = 0.29$, 95% CI: $[0.16, 0.42]$, $t = 4.37$) and D_{random} ($\beta = 0.13$, 95% CI: $[0.00, 0.26]$, $t = 1.94$), compared to the D_{block} condition (the reference by *LMER*). This implies that **estimating missing data in the middle of a time series may be less challenging for users to get true values than those randomly distributed or at the end of the series**.

Similarly, for the tendency measure (d_{impute}), participants' estimations of missing data tended to follow given imputations less in both the D_{end} condition ($\beta = 1.07$, 95% CI: $[0.85, 1.29]$, $t = 9.52$) and the D_{random} condition ($\beta = 0.22$, 95% CI: $[0.00, 0.44]$, $t = 1.89$) than the D_{block} condition (the reference by *LMER*). Under D_{block}, the distance between participant estimations of missing data and their imputed values (d_{impute}) was expected to be: $\beta = 3.21$, 95% CI: $[2.84, 3.58]$, $t = 16.92$. Moreover, for the consistency of participant estimations (c_{impute}), with regard to imputed values, participants made less consistent estimations in D_{end} ($\beta = 0.41$, 95% CI: $[0.26, 0.55]$, $t = 5.53$) and D_{random} ($\beta = 0.17$, 95% CI: $[0.03, 0.32]$, $t = 2.33$) than the D_{block} condition. These suggest that **with the presence of imputed values (even less-accurate ones), participants' estimations of missing data are most likely to be "biased', when the missing data are in the middle of a series; and the least "biased', when they are at the end of the series**.

Regarding the confidence of participants' estimations (j_{user}), participants chose to select lower confidence-ratings in D_{end} ($\beta = -0.14$, 95% CI: $[-0.18, -0.10]$, $t = -7.19$) and higher ones in D_{random} ($\beta = 0.14$, 95% CI: $[0.10, 0.18]$, $t = 6.85$), based on the reference D_{block}. This indicates that participants seemed **less confident of their estimations when working on a consecutive number of missing data than those randomly distributed in a series**.

5.2 Impact by The Imputation & Visualization

The way of providing imputed values of missing data (including their imputations and visualizations) shows a significant impact on the accuracy, tendency, consistency, and confidence of participants' estimations, and their effort in making estimations, but not on the time spent on the estimations (see the last row in Table 3 and 4). For all these measures, *LMER* took (I_{high}, $V_{continuous}$) as the reference and compared others against it.

For accuracy, the distance between participants' estimations of missing data and true values (d_{true}), under the (I_{high}, $V_{continuous}$) condition, was expected to be: $\beta = 2.99$, 95% CI $[2.66, 3.32]$, $t = 17.62$. In comparison, participants' estimations were the least accurate in (I_{no}, V_{empty}) ($\beta = 5.01$, 95% CI: $[4.74, 5.29]$, $t = 35.52$), and less accurate in (I_{low}, $V_{continuous}$), (I_{low}, $V_{discrete}$), and (I_{high}, $V_{discrete}$). Among them, in (I_{high}, $V_{discrete}$), participants' estimations were the most accurate. Moreover, participants' estimations seemed more accurate in (I_{low}, $V_{discrete}$) than (I_{low}, $V_{continuous}$), as the former has a smaller β with a smaller 95% CI. For the consistency of participants' estimations (c_{true}), with regard to true values (I_{high}), (I_{no}, V_{empty}) has the least consistent participants' estimations ($\beta = 2.96$, 95% CI: $[2.80, 3.12]$, $t = 36.18$); while (I_{high}, $V_{discrete}$) has the most consistent ones ($\beta = -0.05$, 95% CI: $[-0.21, 0.11]$, $t = -0.63$). For the other three conditions, participants' estimations were more consistent in (I_{high}, $V_{continuous}$) than (I_{low}, $V_{continuous}$) and (I_{low}, $V_{discrete}$), respectively. Furthermore, compared to (I_{low},

Table 3: Summary of the results on the *accuracy, tendency,* and *consistency* of participants' estimations of missing data. The intercept estimate (β) with 95% CIs of the estimate for each condition are also reported.

Factors	Accuracy (d_{true})	Consistency to True Value (c_{true})	Tendency (d_{impute})	Consistency to Imputation (c_{impute})
Distribution	$\chi^2(2) = 86.943, p < .0001$	$\chi^2(2) = 21.178, p < .0001$	$\chi^2(2) = 90.496, p < .0001$	$\chi^2(2) = 30.635, p < .0001$
D_{block} (intercept)	β: 5.28, [4.95, 5.62], t: 31.13	β: 3.07, [2.90, 3.23], t: 36.35	β: 3.21, [2.84, 3.58], t: 16.92	β: 1.97, [1.78, 2.15], t: 20.81
D_{end}	β: 0.98, [0.76, 1.20], t: 8.73	β: 0.29, [0.16, 0.42], t: 4.37	β: 1.07, [0.85, 1.29], t: 9.52	β: 0.41, [0.26, 0.55], t: 5.53
D_{random}	β: -0.06, [-0.29, 0.16], t: -5.4	β: 0.13, [0.00, 0.26], t: 1.94	β: 0.22, [0.00, 0.44], t: 1.89	β: 0.17 , [0.03, 0.32], t: 2.33
Prior knowledge	$\chi^2(1) = 22.685, p < .0001$	$\chi^2(1) = 205.299, p < .0001$	$\chi^2(1) = 15.940, p < .0001$	$\chi^2(1) = 19.852, p < .0001$
K_{with} (intercept)	β: 5.41, [5.09, 5,73], t: 33.52	β: 2.83, [2.68, 2.99], t: 36.21	β: 3.50, [3.14, 3.86], t: 19.13	β: 2.04, [1.87, 2.22], t: 22.76
$K_{without}$	β: 0.36, [0.17, 0.55], t: 3.76	β: 0.77, [0.66, 0.88], t: 13.81	β: 0.30, [0.11, 0.49], t: 3.10	β: 0.25, [0.13, 0.37], t: 3.99
Imputation & Visualization	$\chi^2(4) = 1709.25, p < .0001$	$\chi^2(4) = 1813.57, p < .0001$	$\chi^2(3) = 61.477, p < .0001$	$\chi^2(3) = 196.833, p < .0001$
(I_{high}, $V_{continuous}$) (intercept)	β: 2.99, [2.66, 3.32], t: 17.62	β: 1.93, [1.77, 2.09], t: 23.41	β: 2.98, [2.59, 3.37], t: 15.00	β: 1.65, [1.45, 1.85], t: 16.04
(I_{high}, $V_{discrete}$)	β: 0.46, [0.19, 0.73], t: 3.38	β: -0.05, [-0.21, 0.11], t: -0.63	β: 0.40, [0.14, 0.65], t: 3.05	β: -0.06, [-0.23, 0.10], t: -0.73
(I_{low}, $V_{continuous}$)	β: 3.61, [3.35, 3.88], t: 26.99	β: 1.77, [1.62, 1.92], t: 22.65	β: 1.00, [0.75, 1.25], t: 7.90	β: 0.78, [0.62, 0.94], t: 9.41
(I_{low}, $V_{discrete}$)	β: 3.41, [3.12, 3.71], t: 22.40	β: 1.57, [1.40, 1.74], t: 18.14	β: 1.00, [0.71, 1.29], t: 6.80	β: 1.01, [0.82, 1.19], t: 10.72
(I_{no}, V_{empty})	β: 5.01, [4.74, 5.29], t: 35.52	β: 2.96, [2.80, 3.12], t: 36.18	N / A	N / A

Table 4: Summary of the results on the *efficiency* and *confidence* of participants' estimations of missing data. The non-significant results are in gray. The intercept estimate (β) with 95% CIs of the estimate for each condition are also reported.

Factors	Time (t_{user})	Effort (f_{user})	Confidence (j_{user})
Distribution	$\chi^2(2) = 4.863, p = .0879$	$\chi^2(2) = 4.129, p = .127$	$\chi^2(2) = 186.162, p < .0001$
D_{block} (intercept)	β: 32.12, [28.62, 35.62], t: 18.03	β: 1.25, [1.09, 1.40], t: 15.74	β: 3.70, [3.64, 3.76], t: 117.60
D_{end}	β: -3.47, [-8.06, 1.12], t: -1.48	β: -0.10, [-0.20, 0.01], t: -1.85	β: -0.14, [-0.18, -0.10], t: -7.19
D_{random}	β: 4.18, [-0.42, 8.77], t: 1.79	β: -0.10, [-0.20, 0.01], t: -1.82	β: 0.14, [0.10, 0.18], t: 6.85
Prior knowledge	$\chi^2(1) = 23.191, p < .0001$	$\chi^2(1) = 1.541, p = .214$	$\chi^2(1) = 0.507, p = .477$
K_{with} (intercept)	β: 27.74, [24.80, 30.69], t: 18.50	β: 1.16, [1.01, 1.31], t: 15.14	β: 3.71, [3.65, 3.77], t: 122.26
$K_{without}$	β: 9.32, [5.55, 13.09], t: 4.85	β: 0.05, [-0.03, 0.14], t: 1.21	β: -0.01, [-0.05, 0.02], t: -0.71
Imputation & Visualization	$\chi^2(4) = 9.165, p = .0571$	$\chi^2(4) = 15.527, p = .00372$	$\chi^2(4) = 182.334, p < .0001$
(I_{high}, $V_{continuous}$) (intercept)	β: 36.03, [31.54, 40.52], t: 15.73	β: 1.27, [1.10, 1.44], t: 14.75	β: 3.81, [3.75, 3.88], t: 110.85
(I_{high}, $V_{discrete}$)	β: -3.22, [-9.21, 2.79], t: -1.05	β: 0.00, [-0.13, 0.13], t: -0.04	β: -0.04, [-0.09, 0.01], t: -1.58
(I_{low}, $V_{continuous}$)	β: -4.06, [-10.05, 1.94], t: -1.33	β: -0.02, [-0.15, 0.11], t: -0.34	β: -0.10, [-0.15, -0.05], t: -3.96
(I_{low}, $V_{discrete}$)	β: -3.66, [-9.62, 2.32], t: -1.20	β: -0.14, [-0.29, 0.01], t: -1.87	β: -0.11, [-0.16, -0.05], t: -3.66
(I_{no}, V_{empty})	β: -7.41, [-13.46, -1.35], t: -2.40	β: -0.25, [-0.39, -0.12], t: -3.63	β: -0.33, [-0.38, -0.27], t: -12.28

$V_{continuous}$), they were more consistent in (I_{low}, $V_{discrete}$). Such results suggest that **showing imputations may help participants estimate missing data more consistently closer to true values than without providing imputations**.

Considering the tendency measure (d_{impute}), participants' estimations of missing data tended to follow presented imputations more under conditions where I_{high} was involved than I_{low} was used. This is because, compared to the reference (I_{high}, $V_{continuous}$), both (I_{low}, $V_{continuous}$) and (I_{low}, $V_{discrete}$) have a positive β, which indicates a larger distance between participants' estimations and given imputations. For the consistency of participants' estimations (c_{impute}), with regard to imputed values, (I_{high}, $V_{discrete}$) has the highest consistent participants' estimations ($\beta = -0.06$, 95% CI: $[-0.23, 0.10]$, $t = -0.73$), while (I_{low}, $V_{discrete}$) leads to the least consistent participants' estimations ($\beta = 1.01$, 95% CI: $[0.82, 1.19]$, $t = 10.72$). Also, participants' estimations were more consistent in (I_{high}, $V_{continuous}$) than (I_{low}, $V_{continuous}$). The results indicate that participants tended to **more consistently align estimations with high-accuracy imputations than low-accuracy ones**.

Regarding the effort of making estimations (f_{user}), participants had fewer adjustments when no and low-accuracy imputations (i.e., I_{no} and I_{low}) are shown than high-accuracy imputations are displayed (i.e., I_{high}). Moreover, when imputations (both I_{high} and I_{low}) were displayed, adjustments made in $V_{discrete}$ remained similar (i.e., (I_{high}, $V_{continuous}$) v.s. (I_{high}, $V_{discrete}$)), or were even fewer (i.e., (I_{low}, $V_{continuous}$) v.s. (I_{low}, $V_{discrete}$)) than those in $V_{continuous}$. These suggest that participants are **more likely to double-check their estimations when showing imputations than without them** (as they made adjustments to their previous estimations), and participants seemed to be **easier to have their decisions settled with the discrete visualization than the continuous one**.

Regarding the confidence of participants' estimations (j_{user}), participants chose the lowest confidence-rating score in (I_{no}, V_{empty}) ($\beta = -0.33$, 95% CI: $[-0.38, -0.27]$, $t = -12.28$) and highest ones in the reference condition (by *LMER*), (I_{high}, $V_{continuous}$). By comparing to the reference, participants selected lower confidence ratings in (I_{high}, $V_{discrete}$), (I_{low}, $V_{continuous}$), and (I_{low}, $V_{discrete}$). Moreover, given the same visualization of imputed values, participants' selected confidence rating was higher when using I_{high} than I_{low}; while, under the same accuracy of imputed values, $V_{continuous}$ led to higher confidence-ratings than $V_{discrete}$. These results indicate that participants were **more confident when imputations were given** than without imputations and they seemed to be **more confident of their estimations with error bars** (than using discrete dots). This seems aligned with the results of their estimation accuracy: in the conditions with higher accuracy of estimations, participants are more likely to select higher confidence ratings.

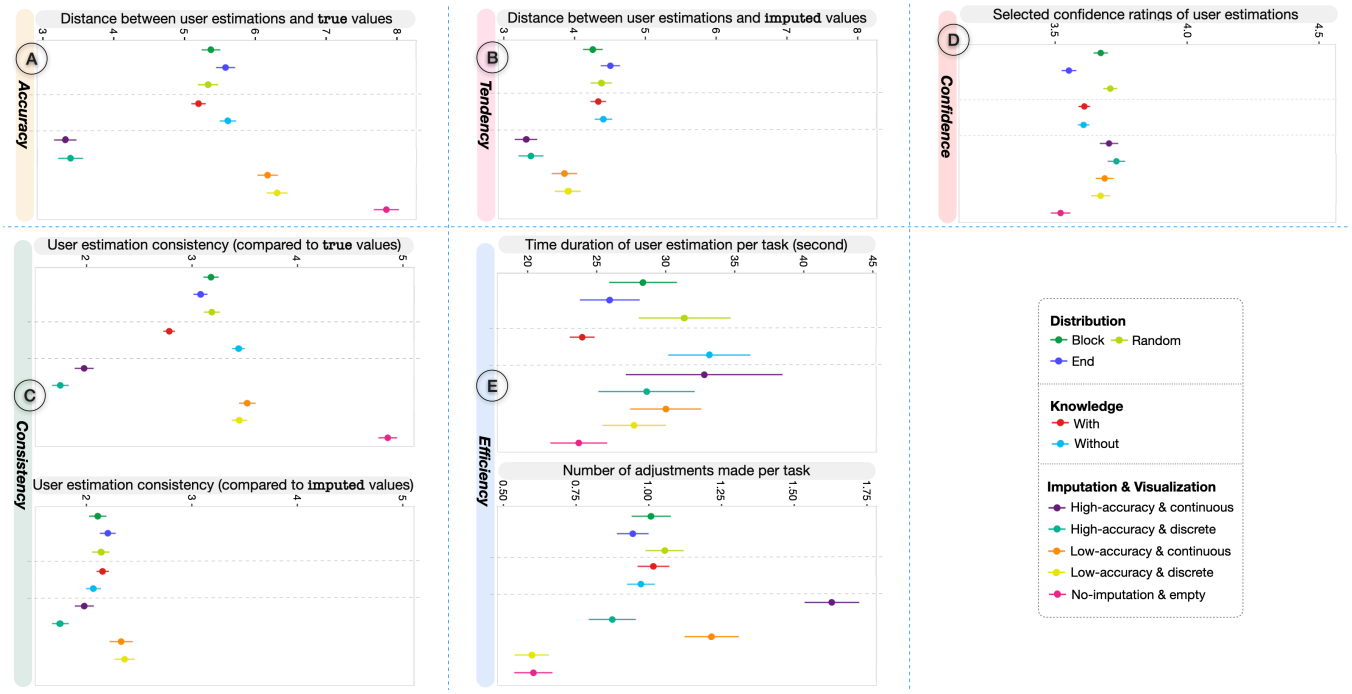

Figure 4: Plots of the results on the *accuracy*, *tendency*, *consistency*, *confidence*, and *efficiency* of participants' estimations of missing data. Each data point in the plot indicates the mean and 95% CI of the corresponding metric.

5.3 Impact by The Prior Knowledge of Data

Prior knowledge of data shows a significant impact on the accuracy, tendency, and consistency of participants' estimations, and the time that they spent on making their estimations. However, it does not significantly affect participants' efforts in estimating missing data and the confidence of their estimations.

With prior knowledge about data, participants' estimations were more consistent (for c_{true}, referencing to K_{with}, $\beta = 0.77$, 95% CI: $[0.66, 0.88]$, $t = 13.81$ for $K_{without}$) and closer (for d_{true}, referencing to K_{with}, $\beta = 0.36$, 95% CI: $[0.17, 0.55]$, $t = 3.76$ for $K_{without}$) to the true values, than without prior knowledge. Similarly, participants tended to more consistently (for c_{impute}, referencing to K_{with}, $\beta = 0.25$, 95% CI: $[0.13, 0.37]$, $t = 3.99$ for $K_{without}$) follow imputed values for making estimations with prior knowledge about data than without it (for d_{impute}, referencing to K_{with}, $\beta = 0.30$, 95% CI: $[0.11, 0.49]$, $t = 3.10$ for $K_{without}$). Moreover, for the time spent on making estimations, they spent less time when they had prior knowledge about data than without it (for t_{user}, reference to K_{with}, $\beta = 9.32$, 95% CI: $[5.55, 13.09]$, $t = 4.85$ for $K_{without}$). These indicate that participants were **more likely to follow given imputations and make more accurate estimations within less time, when they had prior knowledge than they did not**.

6 DISCUSSION

6.1 Comparison to Prior Studies

Our study distinguishes itself from the prior research [11, 22, 26, 61] by offering a comprehensive examination of how four factors, *data*, *computation*, *interface*, and *user*, affect users' estimations of missing data, encompassing a broader scope of the data analysis process than most prior studies. Specifically, [11, 26] focuses on representations of missing data in a line chart, [61] controls imputation methods and representations of missing data in a line chart and a bar chart, and [22] considers data distribution and visualizes missing data as empty space in dot plots, histograms, and density plots. In contrast, our study integrates these elements to provide a more holistic view, centering on users' *estimations* of missing data – a focal point not emphasized in the earlier research. While missing data is an important component in these studies, they emphasize either user-perceived data quality [22, 61], or decision-making (e.g., comparing two values [26] and choosing when to book a travel [11]) in the presence of missing data. For these prior studies, the usage and user perception of visually revealed missing data catches more attention than users' judgment of them. Thus, our study complements these prior works, as the low-level judgment of missing data (from our study) may be associated with the way of using it for decision-making (from prior studies). The findings from this work can help to inform future studies about eliciting and assessing insights from incomplete data.

6.2 Limitations and Future Directions

First, we used a between-subjects design, which inevitably introduced individual bias into the collected data. Thus, our results could be further examined in a within-subjects design setting in the future. However, a within-subjects design would make the study sessions much longer for each participant, which may introduce fatigue, especially regarding the number of conditions.

Second, we used time-series data, particularly weather data, as our test bed in this study, as it is commonly seen in daily life and used in imputation algorithms. However, there are different forms of

data (e.g., trees, graphs, and text) that can have missing values. Our results may not be generalized to them, as not all the specific controls in our study can be directly applied to them (e.g., trees, graphs, and text often use different visualizations than a line chart). However, the four key aspects (data, computation, interface, and user) that our controls follow can be generalized to a broad set of data types and analysis tasks, because these aspects are commonly involved in computing-supported data analytics. In addition, weather data scenarios are relatable to the general public, so laypeople can make reasonable predictions based on their real-life experiences. While participants from MTurk with random assignments can provide a reasonable representation of users' estimations of missing weather data, our study did not consider the role of domain knowledge and expertise in missing data estimation. Future research is needed for other scenarios with different datasets where missing values may not be possibly estimated without domain expertise.

In terms of the generalization of our observations, our study focused on estimating missing data values, which is needed in many types of data analysis tasks, such as merging and resolving data conflicts, cleaning and wrangling data, and interpreting datasets with missing values. We did not study missing data analysis in the context of these tasks with the consideration to reduce confounding factors. One valuable future research direction is to investigate the impact of data context and different analysis tasks on human perception and behavior related to missing data. Nonetheless, we believe that the results from our study shed light on common-sense user perceptions and human-computation collaboration on missing data estimation. The insights from our work are helpful for future exploration of missing data estimation and explorations in graphs and trees, as well as other data forms and application scenarios.

Third, in this study, we used three particular visualizations to show missing data and their imputations. However, there are other visualizations that can also be used to support visual analysis with missing data (e.g., bar charts and matrices). We did not consider them in this study, which may impact people's judgment of missing data. Moreover, the relationship between imputations and visualizations of missing data seems complex. In this work, we only studied two levels of imputation and two types of visualization, with the empty case (I_{no}, V_{empty}) as the basis. Thus, further studies are needed to gain a comprehensive understanding of a broad set of visual encodings for missing data analysis.

Fourth, the metrics used in our study (i.e., accuracy, tendency, and consistency) are observation-oriented ones. They indirectly reveal participants' perceived accuracy of given imputations, which is inferred based on the observations of their estimations. However, we did not consider participants' subjective judgement of the accuracy of given imputations in this study. There could be mismatches between our inferences and participants' opinions. Further studies are needed to gain an in-depth understanding of participants' perceived accuracy of given imputations (e.g., why do users consider high-accuracy imputations as low-accuracy ones, or vice versa).

Last, the participants, recruited in this study, may not be quite representative of all the real-world use cases that need to handle incomplete data for visual analysis. Such cases often involve people, who can be domain experts and have to make informed decisions about complex datasets. As we recruited participants from MTurk, they (i.e., users in this particular online platform) cannot be fully representative of the whole user population for actual analysis use cases. While there are prior studies that have investigated statistical chart interpretation with non-statisticians (e.g., [21, 39, 41]), it has been found that an expert population was more likely to answer questions and provide feedback more accurately [36]. Thus, due to the limitation of participants in this study, our findings may not hold for cases with different groups of users, which requires further verifications with a more diverse group of participants.

7 CONCLUSION

We presented a controlled study to investigate users' estimations of missing data on MTurk with 630 participants using a between-subjects design. We studied four impacting factors: the *distribution*, *imputation*, and *visualization* of missing data, and users' *prior knowledge* of data. We controlled each factor with multiple conditions based on common patterns summarized in the literature. To measure users' estimations, we used metrics for the accuracy, tendency, and consistency of estimations, participants' efficiency, and their self-judgment (i.e., level of confidence). Our quantitative analyses indicate that all the factors significantly affect the distance and consistency of user estimations with respect to the ground truth and low-accuracy imputations. Prior knowledge shows a significant impact on the task time; and imputation and visualization influence participants' efforts in making estimations. Also, participants' confidence is significantly affected by all factors except prior knowledge. Collectively, with observations discussed in this work, the results could inform future studies of developing trustworthy, interactive computing methods for visual analysis with missing data.

ACKNOWLEDGMENTS

This work is supported in part by the NSF Grant IIS-2002082, the International Research Partnership Grant (IRPG) from the University of Waterloo, and the Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] [n. d.]. Amazon Mechanical Turk. https://www.mturk.com/.
[2] [n. d.]. Blue Hill Meteorological Observatory Climate Data. https://bluehill.org/climate-weather/observatory-climate-data/.
[3] [n. d.]. Bootstrap. https://getbootstrap.com/.
[4] [n. d.]. Django: The web framework for perfectionists with deadlines. https://www.djangoproject.com/.
[5] [n. d.]. Microsoft Azure: Cloud Computing Services. https://azure.microsoft.com/.
[6] [n. d.]. Minimum Wage, U.S. Department of Labor. https://www.dol.gov/general/topic/wages/minimumwage.
[7] [n. d.]. PostgreSQL: The world's most advanced open source database. https://www.postgresql.org/.
[8] Kamran Abbasi. 2014. The missing data that cost $20 bn. 348 (2014), g2695. https://doi.org/10.1136/bmj.g2695
[9] Paul D Allison. 2001. *Missing data*. Sage publications.
[10] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300233
[11] Rebecca Andreasson and Maria Riveiro. 2014. Effects of visualizing missing data: an empirical evaluation. In *International Conference on Information Visualisation*. 132–138. https://doi.org/10.1109/IV.2014.77
[12] Amanda N Baraldi and Craig K Enders. 2010. An introduction to modern missing data analyses. *Journal of school psychology* 48, 1 (2010), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001
[13] Derrick A Bennett. 2001. How can I deal with missing data in my study? *Australian and New Zealand journal of public health* 25, 5 (2001), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

[14] Kenneth J Berry, Paul W Mielke Jr, and Hariharan K Iyer. 1998. Factorial designs and dummy coding. *Perceptual and motor skills* 87, 3 (1998), 919–927.

[15] James R Bettman and C Whan Park. 1980. Effects of prior knowledge and experience and phase of the choice process on consumer decision processes: A protocol analysis. *Journal of consumer research* 7, 3 (1980), 234–248. https://doi.org/10.1086/208812

[16] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R Johnson, Manuel M Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. 2014. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*. Springer, 3–27. https://doi.org/10.1007/978-1-4471-6497-5_1

[17] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

[18] J Michael Brick and Graham Kalton. 1996. Handling missing data in survey research. *Statistical methods in medical research* 5, 3 (1996), 215–238. https://doi.org/10.1177/096228029600500302

[19] David R Brillinger. 2001. *Time series: data analysis & theory*. SIAM.

[20] Chris Chatfield. 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 158, 3 (1995), 419–444. https://doi.org/10.2307/2983440

[21] Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Trans. on Vis. & Computer Graphics* 20, 12 (2014), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

[22] Michael Correll, Mingwei Li, Gordon Kindlmann, and Carlos Scheidegger. 2018. Looks good to me: Visualizations as sanity checks. *IEEE Trans. on Vis. & Computer Graphics* 25, 1 (2018), 830–839. https://doi.org/10.1109/TVCG.2018.2864907

[23] National Research Council et al. 2010. The prevention and treatment of missing data in clinical trials. (2010).

[24] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. 2006. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology* 59, 10 (2006), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

[25] Yiran Dong and Chao-Ying Joanne Peng. 2013. Principled missing data methods for researchers. *SpringerPlus* 2, 1 (2013), 1–17. https://doi.org/10.1186/2193-1801-2-222

[26] Cyntrica Eaton, Catherine Plaisant, and Terence Drizd. 2005. Visualizing missing data: Graph interpretation user study. In *IFIP Conference on Human-Computer Interaction*. 861–872. https://doi.org/10.1007/11555261_68

[27] Bradley Efron. 1994. Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* 89, 426 (1994), 463–475. https://doi.org/doi.org/10.1080/01621459.1994.10476768

[28] Craig K Enders. 2010. *Applied missing data analysis*. Guilford press.

[29] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3173574.3173718

[30] Sara Johansson Fernstad. 2019. To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. *Information Visualization* 18, 2 (2019), 230–250. https://doi.org/10.1177/1473871618785387

[31] Sara Johansson Fernstad and Robert C Glen. 2014. Visual analysis of missing data—To see what isn't there. In *IEEE Conference on Visual Analytics Science and Technology*. 249–250. https://doi.org/10.1109/VAST.2014.7042514

[32] Sara Johansson Fernstad and Jimmy Johansson Westberg. 2021. To Explore What Isn't There—Glyph-Based Visualization for Analysis of Missing Values. *IEEE Transactions on Visualization and Computer Graphics* 28, 10 (2021), 3513–3529. https://doi.org/10.1109/TVCG.2021.3065124

[33] Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics* 40, 1 (2011), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

[34] John W Graham. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology* 60 (2009), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

[35] Amit Gruber and Yair Weiss. 2004. Multibody factorization with uncertainty and missing data using the EM algorithm. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1. I–I. https://doi.org/10.1109/CVPR.2004.1315101

[36] Jouni Helske, Satu Helske, Matthew Cooper, Anders Ynnerman, and Lonni Besancon. 2021. Can visualization alleviate dichotomous thinking? Effects of visual representations on the cliff effect. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (2021), 3397–3409. https://doi.org/10.1109/TVCG.2021.3073466

[37] Robin M Hogarth and Hillel J Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive psychology* 24, 1 (1992), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

[38] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Trans. on Vis. & Computer Graphics* 25, 1 (2018), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

[39] Alex Kale, Matthew Kay, and Jessica Hullman. 2020. Visual reasoning strategies for effect size judgments and decisions. *IEEE Trans. on Visualization & Computer Graphics* 27, 2 (2020), 272–282. https://doi.org/10.1109/TVCG.2020.3030335

[40] Bharat Kale, Austin Clyde, Maoyuan Sun, Arvind Ramanathan, Rick Stevens, and Michael E Papka. 2023. ChemoGraph: interactive visual exploration of the chemical space. In *Computer Graphics Forum*, Vol. 42. 13–24. https://doi.org/10.1111/cgf.14807

[41] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 5092–5103. https://doi.org/10.1145/2858036.2858558

[42] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456. https://doi.org/10.1145/1357054.1357127

[43] Gueorgi Kossinets. 2006. Effects of missing data in social networks. *Social networks* 28, 3 (2006), 247–268. https://doi.org/10.1016/j.socnet.2005.07.002

[44] Kamakshi Lakshminarayan, Steven A Harp, and Tariq Samad. 1999. Imputation of missing data in industrial databases. *Applied Intelligence* 11, 3 (1999), 259–275. https://doi.org/10.1023/A:1008334909089

[45] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* 53, 2 (2020), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4

[46] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

[47] Le Liu, Alexander P Boone, Ian T Ruginski, Lace Padilla, Mary Hegarty, Sarah H Creem-Regehr, William B Thompson, Cem Yuksel, and Donald H House. 2016. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (2016), 2165–2178. https://doi.org/10.1109/TVCG.2016.2607204

[48] Le Liu, Lace Padilla, Sarah H Creem-Regehr, and Donald H House. 2018. Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 882–891. https://doi.org/10.1109/TVCG.2018.2865193

[49] Hao Ma, Irwin King, and Michael R Lyu. 2007. Effective missing data prediction for collaborative filtering. In *Proc. of the ACM Conference on Research & Development in Information Retrieval*. 39–46. https://doi.org/10.1145/1277741.1277751

[50] Teresa A Myers. 2011. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication methods and measures* 5, 4 (2011), 297–310. https://doi.org/10.1080/19312458.2011.624490

[51] Felix Naumann. 2014. Data profiling revisited. *ACM SIGMOD Record* 42, 4 (2014), 40–49. https://doi.org/10.1145/2590989.2590995

[52] Lace Padilla, Matthew Kay, and Jessica Hullman. 2021. *Uncertainty Visualization*. American Cancer Society, 1–18. https://doi.org/10.1002/9781118445112.stat08296

[53] Alex T Pang, Craig M Wittenbrink, Suresh K Lodha, et al. 1997. Approaches to uncertainty visualization. *The Visual Computer* 13, 8 (1997), 370–390. https://doi.org/10.1007/s003710050111

[54] Deokgun Park, Steven M Drucker, Roland Fernandez, and Niklas Elmqvist. 2017. Atom: A grammar for unit visualizations. *IEEE Trans. on Visualization & Computer Graphics* 24, 12 (2017), 3032–3043. https://doi.org/10.1109/TVCG.2017.2785807

[55] Ronald K Pearson. 2006. The problem of disguised missing data. *Acm Sigkdd Explorations Newsletter* 8, 1 (2006), 83–92. https://doi.org/10.1145/1147234.1147247

[56] Fadoua Rafii and Tahar Kechadi. 2019. Collection of historical weather data: issues with missing values. In *Proceedings of the 4th International Conference on Smart City Applications*. 1–8. https://doi.org/doi.org/10.1145/3368756.3368974

[57] Donald B Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592. https://doi.org/10.1093/biomet/63.3.581

[58] Muhammad Saad, Mohita Chaudhary, Fakhri Karray, and Vincent Gaudet. 2020. Machine learning based approaches for imputation in time series data and their impact on forecasting. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2621–2627. https://doi.org/10.1109/SMC42975.2020.9283191

[59] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2014. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1604–1613. https://doi.org/10.1109/TVCG.2014.2346481

[60] Joseph L Schafer and John W Graham. 2002. Missing data: our view of the state of the art. *Psychological methods* 7, 2 (2002), 147. https://doi.org/10.1037/1082-989X.7.2.147

[61] Hayeong Song and Danielle Albers Szafir. 2018. Where's my data? evaluating visualizations with missing data. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 914–924. https://doi.org/10.1109/TVCG.2018.2864914

[62] Maoyuan Sun, Lauren Bradel, Chris L North, and Naren Ramakrishnan. 2014. The role of interactive biclusters in sensemaking. In *Proc. of the Conf. on Human Factors in Computing Systems*. 1559–1562. https://doi.org/10.1145/2556288.2557337

[63] Maoyuan Sun, Gregorio Convertino, and Mark Detweiler. 2016. Designing a unified cloud log analytics platform. In *International Conference on Collaboration Technologies and Systems*. IEEE, 257–266. https://doi.org/10.1109/CTS.2016.0057

[64] Maoyuan Sun, Yue Ma, Yuanxin Wang, Tianyi Li, Jian Zhao, Yujun Liu, and Ping-Shou Zhong. 2022. Toward Systematic Considerations of Missingness in

Visual Analytics. In *Visualization and Visual Analytics*. IEEE, 110–114.
[65] Maoyuan Sun, Jian Zhao, Hao Wu, Kurt Luther, Chris North, and Naren Ramakrishnan. 2018. The effect of edge bundling and seriation on sensemaking of biclusters in bipartite graphs. *IEEE Transactions on Visualization and Computer Graphics* 25, 10 (2018), 2983–2998. https://doi.org/10.1109/TVCG.2018.2861397
[66] Andreas S Weigend. 2018. *Time series prediction: forecasting the future and understanding the past*. Routledge.
[67] Jian Zhao, Maoyuan Sun, Francine Chen, and Patrick Chiu. 2019. Missbin: Visual analysis of missing links in bipartite networks. In *IEEE Visualization Conference*. 71–75. https://doi.org/10.1109/VISUAL.2019.8933639
[68] Jian Zhao, Maoyuan Sun, Francine Chen, and Patrick Chui. 2022. Understanding missing links in bipartite networks with missbin. *IEEE Transactions on Visualization and Computer Graphics* 28, 6 (2022), 2457–2469. https://doi.org/10.1109/TVCG.2020.3032984

Received 12 April 2024; revised 17 May 2024; accepted 5 June 2024