Toward Systematic Considerations of Missingness in Visual Analytics

Maoyuan Sun* Yue Ma* Yuanxin Wang† Tianyi Li‡ Jian Zhao† Yujun Liu* Ping-Shou Zhong§
*Northern Illinois University †University of Waterloo ‡Purdue University §University of Illinois Chicago

ABSTRACT

Data-driven decision making has been a common task in today's big data era, from simple choices such as finding a fast way to drive home, to complex decisions on medical treatment. It is often supported by visual analytics. For various reasons (e.g., system failure, interrupted network, intentional information hiding, or bias), visual analytics for sensemaking of data involves missingness (e.g., data loss and incomplete analysis), which impacts human decisions. For example, missing data can cost a business millions of dollars, and failing to recognize key evidence can put an innocent person in jail. Being aware of missingness is critical to avoid such catastrophes. To fulfill this, as an initial step, we consider missingness in visual analytics from two aspects: *data-centric* and *human-centric*. The former emphasizes missingness in three data-related categories: *data composition*, *data relationship*, and *data usage*. The latter focuses on the human-perceived missingness at three levels: *observed*-level, *inferred*-level, and *ignored*-level. Based on them, we discuss possible roles of visualizations for handling missingness, and conclude our discussion with future research opportunities.

Keywords: Missingness, missing data visualization, sensemaking, visual analytics.

1 INTRODUCTION

For various reasons (e.g., system failures, network problems, intentional information hiding, or bias), a human sensemaking process involves *missingness*, such as missing data, biased data selection [46], or partially finished analyses. It impacts human decisions and may cause severe consequences. For example, missing data costs millions of dollars per year in business [1]; and due to failing to notice incomplete evidence, John Bunn was falsely convicted of murder [22], and Sunil Tripathi was wrongfully accused as a suspect in the Boston Marathon bombing on social media [29]. These tragedies lead us to question: what can be missing in analytics; can we design techniques to prevent people from falling into traps of such missingness?

While investigating missingness in analytics remains an elusive skill and an understudied task, a fair amount of effort has been put in a focused direction: missing data estimation (e.g., imputation [13]). It aims to "fix" recognized incomplete data by replacing missing data with some "best, reasonable inference" based on existing data [37]. This replacement "breaks" missingness, especially when considering missing data as a type of data [37,40]. It may bring a false impression of completeness. In fact, using such techniques implies that users realize missingness. Thus, a successful awareness of missingness is critical for sensemaking activities.

Visualization can help with missingness awareness for data analytics [3, 12]. With proper visual encodings, missingness gets salient and perceptually attracts user attention. For example, given a dataset of connections between two sets of entities, showing it in a matrix with different cell colors (blue indicates the existence of a connection and white means no connection) allows users to see both existing and missing data. By marrying advanced computation with human cognition with interactive visualizations, visual analytics [20] may better handle missingness involved in analyses.

Nevertheless, current understanding of missingness in visual analytics seems scattered and primarily focusing on data values (e.g., missing data [14]). However, missingness can be more complex [37], when analysts make sense of various data and with different goals. We aim to establish a systematic understanding of missingness in visual analytics and pave the road for future research using visual analytics to handle issues related to or caused by the missingness.

To fulfill this, as an initial step, we consider missingness in a sensemaking process with visual analytics from two aspects: *data-centric* and *human-centric*. The former regards the information to be analyzed by users and the latter highlights how users perceive the information. Specifically, the *data-centric* aspect regards missingness in three data-related categories: *data composition*, *data relation* and *data usage*. The *human-centric* aspect considers missingness in three perception-oriented levels: *observed-level*, *inferred-level* and *ignored-level*. They correspond to the two key parties in visual analytics: computation and human cognition, combined together by interactive visualizations. Computation can help discover data-related missingness [4] (e.g., missing relationship detection [10,48,49]). Visualization can impact user awareness of missingness (e.g., missing data) and judgement on data quality [40]. Our considerations enable a systematic way of further studying and handling missingness in visual analytics. Based on them, visualization design needs to help reveal data-centric missingness and improve user awareness of missingness (e.g., moving from the ignored-level to the observed-level). We hope this work can draw attention to future exploration of the design space of visualizing missingness and studying insights from incomplete data in visual sensemaking.

2 A DATA-CENTRIC VIEW OF MISSINGNESS

A data-centric view considers missingness in three data-related categories: *data composition*, *data relationship* and *data usage*. In this section, we first introduce our notion of data composition and data relationship, and then discuss the data-centric view of missingness.

2.1 Data Composition and Data Relationship

The composition of data can include three major components [6, 8]: *entity*, *attribute* and *value*. An entity is a data item, which is a basic unit encoding a piece of information. An attribute is a specification that describes an entity, and an entity can have multiple attributes. A value reveals how an entity performs on an attribute. A *dataset* can be considered as a collection of entities, described by one or multiple attributes with specified values. These components have been used for database management, as a relational model [8], which organizes data in tables. Each row is an entity, each column corresponds to an attribute, and each cell includes a value.

The relationship between data entities can be described using the concept of *entity set*. An entity set refers to a set of *unique* data entities that share the same attribute(s) (e.g., a collection of photos or a list of locations). For two entity sets X and Y, a relationship between them, $R(X,Y)$, is a subset of their Cartesian product $X \times Y$. When $R(X,Y)$ is not empty, we say X is related to Y. Otherwise, we say that X is independent from Y. The relationship R can be determined

*e-mail: {smaoyuan, myue, lyujun}@niu.edu.
†e-mail: {y2587wang, jianzhao}@uwaterloo.ca.
‡e-mail: li4251@purdue.edu
§e-mail: pszhong@uic.edu

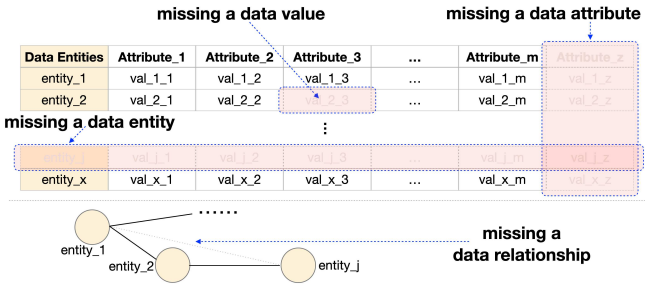

Figure 1: Examples of data composition missingness (top), including: 1) missing data entity, 2) missing data attribute, and 3) missing data value, and data relationship missingness (bottom).

by using data values of selected data attributes. Moreover, for different usage scenarios, the relationship R can be determined differently. For example, in cyber security, R may be defined as communication between computers and web URLs; while in bioinformatics, R may be determined based on expressed genes under conditions.

2.2 Data Composition Missingness

Based on the data composition discussed in Section 2.1, three types of missingness exist: 1) *missing data entity*, 2) *missing data attribute* and 3) *missing data value*. Figure 1 (top) gives an example of them.

Missing data entities highlights the absence of data entities. It is similar to missing observations [36]. It may result from errors in data collection [2]. For example, in fitness tracking devices, some sensor data might not be recorded due to network connection failures.

Missing data attributes refers to the loss of possibly useful data attributes. It may come from an ill-defined problem space that leads to problems in data collection mechanisms [32] or decisions made by considering some practical concerns. For example, when designing a logging system of a cloud application, some user behaviors could be overlooked [42] or intentionally untracked due to an expensive cost; when creating a survey, researchers may fail to include all related questions or leave some questions unasked for privacy concerns. For such reasons, even a data collection process runs successfully, some data attributes can still be missing.

Missing data values emphasizes the loss of data values, which is often called missing data for short. Compared to the other two, it has drawn the most attention and been heavily studied [14, 15, 40]. Specifically, based on the distribution of data values, missing data values can be categorized into the following three groups [35]:

- *Missing completely at random (MCAR)*: missing data is assumed to not have any underlying mechanism and thus exhibits no relationships with either existing data or other missing data.
- *Missing at random (MAR)*: assumes dependencies on observed values (i.e., existing data), but assumes no underlying relationships between the missing values themselves.
- *Missing not at random (MNAR)*: is the most restrictive, which requires dependencies between missing values.

Successfully recognizing them can help one understand data under analysis and reasonably generate missing ones for augmentation, if needed (e.g., training machine learning models) [45]. Besides real absent values, missing values has a special case, named as *disguised missing* [31], in which the value is present but not accurate. For example, a user leaves a default value (e.g., May 1st for a birthday) for privacy concerns. In this case, the value is not absent, but instead a default one that may not match the truth. Thus, in the cases of disguised missing, while values are present, the true information that data collectors need remains missing, as it is intentionally hidden.

2.3 Data Relationship Missingness

Data relationship missingness means the absence of relations among data entities. From a graph perspective, it emphasizes the lack of links among nodes in a graph (Figure 1 (bottom)). This means that

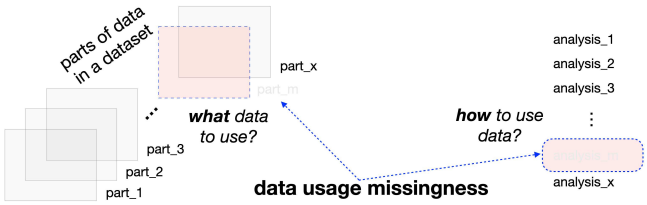

Figure 2: Two possible missingness in data usage: 1) missing data selection, and 2) missing analytical method selection.

for a given set of data entities, some connections between them do not, or with a low probability, exist. Missing relationships among data entities may either result from errors in a data collection process or be a reflection of algorithmic results of data relationship discovery. For example, there is no link between two bank accounts due to the loss of an intelligence report; or the connection between two persons is computed as a very low probability based on word cooccurrence.

Missing data relationships can be formalized as the problem of missing links in graphs. Similar to using the imputation techniques for missing value inferences, based on existing links, missing links can be computationally identified [48, 49]. A common goal of such techniques is to find potentially useful missing links (e.g., serving as a bridge that connects two communities in a social network), and further fix and verify them by adding back the lost ones [49].

2.4 Data Usage Missingness

The utility of data for sensemaking activities involve two key types: 1) *data selection* and 2) *analytical method selection*. The former refers to which parts of a given dataset will be selected for analysis. The latter means which analytical methods will be picked and applied to the selected data. Missingness in data usage can happen in both activities due to uncertainties and selection biases [16, 47]).

Missing data selection reveals that not the whole dataset is selected for analysis. For example, to find similar cars, 4 out of 100 attributes are selected and the rest remains unused; or instead of using all data entities, a dataset is sampled and then analyzed. When selected attributes or samples of data entities are not representative (e.g., stratified sampling [44]), missing data selections occurs.

Missing analytical method selection reveals that not a full set of possibly applicable analytical methods is selected and tested. For real-world problems, it is not easy or sometimes even impossible to identify a complete set of analytical methods. Thus, this missingness highlights that performed analyses are not sufficient, which calls for further or alternative analysis (i.e., multiverse analysis [11]). For example, to explore similar cars, only one centroid-based clustering method (i.e., k-means clustering) is used, but other applicable clustering techniques that may work are not studied.

3 A Human-Centric View of Missingness

A human-centric view treats missingness at three levels: 1) *observed*-level, 2) *inferred*-level and 3) *ignored*-level. They reveal how data-centric missingness discussed in Section 2 is perceived by people.

3.1 Observed Missingness

Observed missingness refers to that users can directly perceive missingness. It indicates that the visibility of missingness is high, and users can easily notice it. For example, a user quickly realizes that a data value is missing, after she sees an empty cell in a table; or by checking and following ribbons in a parallel sets [23], a user finds that there is no connection between two categorical data entities [9].

As the visibility of missingness is affected by the way that data is represented, observed missingness relies on visual context, in which data is encoded by certain visualizations. Different visual encodings can impact how easily users can observe missingness. For example, as is shown in Figure 3, it is easier for users to see missing links by looking at a matrix than checking the same data displayed as lists of

{a1, b1} {a1, b2} {a1, b4} {a1, b5}
{a2, b1} {a2, b2} {a2, b3} {a3, b1}
{a3, b2} {a3, b3} {a3, b4} {a3, b5}

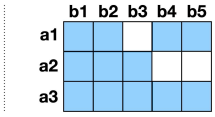

Figure 3: Compared to listing connections between individual entities (left), it is easier for users to see missing connections when showing the same data in a matrix with different cell colors (right).

node-pairs. Thus, for observed missingness, users can verify their perceived missingness by referring to certain given visual context (e.g., pointing to an empty cell).

3.2 Inferred Missingness

Inferred missingness refers to that the visibility of missingness goes low or missingness even gets invisible, so it is impossible for users to directly observe missingness. However, via an investigation with given data, users can infer the possible existence of missingness. For example, by reading the following four intelligence reports that are modified based on the Sign of the Crescent dataset [17]:

"Report on 04/24/2003. Phone calls on 22 April, 2003 made from 703-659-2317 to the numbers: 804-759-6302 and 804-774-8920. A translation of this message reads: 'I will be in my office on April 30 at 9:00AM. Try to be on time'.

Report on 01/11/2003. Abdul Ramazi is the owner of the Select Gourmet Foods shop in Springfield Mall, Springfield, VA., with a phone number 703-659-2317.

Report on 03/18/2003. A check with mobile phone providers shows that a Sprint cell phone 804-774-8920 is registered in the name Mukhtar Galab.

Report on 04/14/2003. The contact given by Faysal Goba was: 1631 Capitol Ave., Richmond VA; phone number: 804-759-6302. From an interrogation of a cooperative detainee in Guantanamo. Detainee says he trained daily with a man named Faysal Goba at an Al Qaeda explosives training facility in the Sudan in 1994."

One may infer that the three persons, `Abdul Ramazi`, `Mukhtar Galab` and `Faysal Goba` may collude suspicious activities together. However, it seems missing in the given reports as it was not explicitly reported. Compared to observed missingness, inferred missingness may not be easily verified. Thus, observed missingness seems more confirmative, while inferred missingness is more hypothetical, which suggests and is closely related with implicit uncertainty [27, 30].

3.3 Ignored Missingness

Ignored missingness indicates no observation nor awareness of missingness and the presence of possible missingness is not considered. It may occur for two reasons. First, the visibility of missingness is too low to raise user awareness. For example, in Figure 3, a user may never realize that missing edges exist after looking at lists of node-pairs. Second, due to some biases or the impact of cognitive capture (or tunneling) [39], users turn a blind eye to possible missingness. For example, to explore possible treatment for a disease, all effort has been put on the group of people who have been infected by the disease, while the uninfected group never gets any attention.

While it cannot be completely avoided in the analysis, ignored incompleteness, if identified, can bring critical values for sensemaking activities [33]. Based on this, we consider that ignored missingness is similar to the concept of *white space* (also named as *opportunity space*) discussed in the business domain [19]. The white space suggests new leads for possible growths of a business. For example, the customers of a credit card product fall into two major age groups 25-35 and 50-70. The gap between 35 and 50 is a white space. It indicates that the current product seems not attractive for the age group 35-50, given the lack of users. Thus, this white space implies an opportunity to design a different credit card product with new awarding features for competing in the market of the missing age group. While a white space can bring useful values to business, it is usually difficult to capture and may easily slip one's attention [19] (e.g., the missing age group catches no attention at all).

In summary, observed missingness takes the least amount of user effort to perceive possible missingness; while for ignored missingness, users are not aware of the existence of missingness in the whole sensemaking process. Moreover, for inferred missingness, users can realize missingness but it takes more effort.

4 Handling Missingness: Visualization Roles

Based on the data-centric and human-centric perspectives of missingness mentioned before, here we discuss four possible roles of visualizations for handling missingness. The first and second roles highlight supporting the detection of data-centric missingness. The other two roles aim to improve user awareness of data-related missingness. In summary, visualizations can help to uncover the data-centric missingness and improve their expressiveness, so they become more visible and accessible to users.

4.1 Bridging Existing Data and Missing Data

Visualizations play a key role of bridging the gap between existing data and missing data. If we consider existing data as a visible land and missing data as an invisible world, a usable bridge connecting them is critical to enable users to explore and walk into the invisible part from the visible one. This is because users need existing data as a landing point before digging into the data-centric missingness. However, to enable the analytical transition from the existing data to the missing part, users need the support of necessary information hints or leads, which can be provided by visualizations [7].

To establish such a bridge, a commonly used strategy is *space-filling* that reveals missingness as empty (e.g., an empty space in a bar chart [40]), gap (e.g., broken lines in a line chart [40]), or different-looking space (e.g., a matrix with different colored cells [14, 48]). The focus of such visualization techniques are on the existing data. As the present data is mapped to certain visual encodings, possible missingness gets visible. Looking at a visually salient space, users can be aware of data-centric missingness. Thus, visualized existing data serves critical and usable visual context that enables users to identify data-related missingness.

4.2 Supporting the Analysis of Analytic Provenance

Visualizations can serve a usable solution to understand and audit analytical provenance [24, 34], which is helpful to address incompleteness in data usage. It is challenging for users to keep tracking the process of their analyses. In a sensemaking process, some parts of data may not receive enough attention and users may miss one or several possibly applicable methods unintentionally. To help avoid such data usage missingness, visualizations can be used to support tracking analytical provenance and further analyze it.

To help identify missingness in data usage, two key aspects need to be considered: 1) the selected, investigated, derived and newly generated data, and 2) the method or process applied to such data. They, respectively, correspond to the provenance of data and process [38]. Visualizing them offers a way of analyzing analytic provenance. By checking such visualizations, users may notice missingness in data usage and further overcome limitations of their analyses.

4.3 Improving Awareness: from Ignoring to Observing

From a perceptual-oriented perspective, a key role of visualization is to prevent users from falling in the trap of ignoring data-centric missingness. The presence of missingness can become more visible to users via the usage of visualizations than without them, so it is more likely for users to be aware of missingness. This implies that using visualizations can improve the *expressiveness* of data-centric missingness. The higher such expressiveness goes, the easier it is for users to observe possible missingness. Thus, using visualizations to handle data-centric missingness attempts to move forward from ignoring missingness to being able to observe it.

Proper visual encodings can direct user attention to data-centric missingness (e.g., missing data values) [40], which may otherwise be ignored by users. The data-centric missingness is often unknown to users at the initial analysis stage, unless they are informed. Thus, a

sensemaking process with data-centric missingness is exploratory in nature and the original analysis goal may not consider missingness at all. However, by referring to visualizations used in a sensemaking process, users may realize the existence of missingness, which could happen at an "aha" moment [25]. This matches both the *spontaneous* insight [5] of visual analytics and one of the key characteristics of visualization insight – *unexpected* [28].

4.4 Scaffolding Missingness Inference

Visualizations offer a usable mean to scaffold missingness inference. Different from the other two perceptual levels of missingness (i.e., observing and ignoring missingness), inferring missingness requires more user effort, as possible missingness is not directly revealed but somehow can be inferred with enough cognitive effort. This can be supported by using visualizations. In this case, instead of merely encoding missing parts of data or existing parts for the purpose of indicating the "hole" in data, visualizations may focus on displaying either the connections across different parts of data or the provenance of a sensemaking process. These help users to infer possible existence of data-centric missingness.

As inference is a reasoning process, instead of a static stage, using multiple types of visualizations and fusing information across them may help with missingness inference. For example, by examining links in a social network graph, checking related organizations, and reading relevant reports, users may infer that two suspects colluded some threats together, which was never reported in a given dataset. This visualization-aided inference making may help one identify the data-centric missingness, such as checking multiple visualizations in a dashboard about the usage of a cloud application to find attributes of user interactions that have not been logged before [42].

5 Discussion and Conclusion

While handling missingness remains a challenging problem in sensemaking, as an initial step, we present considerations that may help to systematically study missingness in visual analytics. They highlight considering missingness from two key perspectives: *data-centric* and *human-centric*. The former regards missingness in three data-related categories: *data composition*, *data relationship* and *data usage*. The latter focuses on the human-perceived missingness at three levels: *observed*-level, *inferred*-level and *ignored*-level. Based on the considerations, we discuss four possible roles of visualizations for helping to handle missingness in a sensemaking process. While they help to lay a preliminary theoretical foundation that aims to systematically consider missingness in visual analytics, to handle missingness in practice, there are four research themes that are worthy of future studies: 1) missingness detection, 2) missingness visualization, 3) missingness insight, and 4) possible relations between missingness and uncertainty.

5.1 Detecting Missingness

Missingness detection lays the foundation for effective data analysis. Detecting missingness is not as simple as it looks like. Unless there is a clear detection goal or some evidence that reveals something is missing (e.g., data value), detecting missingness is fundamentally attempting to address an *unknown unknown* problem [26]. This brings a deeper question: how can we help users know which types of data-centric missingness (e.g., missing data attributes, missing data relationships, or missing data selections in the usage) exist? This is essential as it sets the detection goal. If users were not clear about this, it would be hard for them to further explore and work on detection methods. Also, a sensemaking process can have multiple types of data-centric missingness. For example, a vulnerable system with an interrupted network connection and a problematic logging mechanism can lead to missing both data values and attributes. For such cases, detecting missingness is even challenging. Our considerations presented in this work may help to clarify detection goals.

5.2 Visualizing Missingness

The design of missingness-oriented visualizations remains an under-explored direction. Prior work has investigated visual encodings for missing values [14, 15, 40] and missing links [48]. However, the design space of visualizing missingness can be broader, especially considering that there are different categories of data-centric missingness and they may need different visual encodings. As studied in [40], even for the same type of missingness, different visual encodings can be designed, which further impacts user-perceived data quality. How to formalize the design space of missingness visualizations still needs further explorations. Furthermore, considering the evaluation of missingness visualization, how and if possible can we measure the expressiveness of visual encodings for data-centric missingness? It enables comparing different designs for visualizing missingness, which can be helpful to support making design decisions. The perceptual-perspective discussed in our framework may help to derive usable measures.

5.3 Discovering Insights from Missingness

Studying possible insights that users gain from missingness in sensemaking is a highly sought-after research challenge. Missingness can be considered as a type of "data" [40] from which users can gain usable insights (e.g., using partial bipartite graphs for performing grouping tasks [43]). This turns missingness from being considered as dirty [21] to usable. For example, in an intelligence analysis, a missing link between two suspects may drive the subsequent analysis towards an investigation of any possible connections between them [41]. While this is a simple example, it shows that missingness can be used in a sensemaking process. The insights derived from missingness may depend on an application domain and different types of data-centric missingness may bring different insights. Moreover, insights discovered from missingness, if possible, via using visual analytics, may enlarge the set of characteristics of visualization insight [28]. This may further broaden our understanding of evaluating visualizations by considering the value of missingness.

5.4 Relating Missingness with Uncertainty

An in-depth understanding of possible relations between missingness and uncertainty remains under-explored. For one thing, data-centric missingness brings uncertainty about data and analysis. In practice, uncertainty resulted from data loss, regardless of intentionally or not, may not be well-resolved, as it may be impossible to collect the truth. Thus, would user awareness of data-centric missingness help with uncertainty-based decision making remains an unanswered question. For another, uncertainty may impact user awareness of missingness. While uncertainty may occur due to a variety of reasons [18], would some level of visually expressed uncertainty impact user awareness or perception of missingness? Specifically, after being exposed to some visualized uncertainty, would users associate this with missingness or would this help direct users to starting thinking or inferring data-centric missingness? Answers to such questions can help enrich the design space of uncertainty visualizations and advance knowledge about sensemaking under uncertainty and missingness. To find such answers, further studies are needed.

In summary, we present considerations of missingness in visual analytics from two aspects: *data-centric* and *human-centric*, which offers a possible way of further systemically studying missingness. We hope this work can draw attention to future studies on visual sensemaking with missingness.

Acknowledgments

This research is supported in part by NSF Grants IIS-2002082 and DMS-2152070, the Research and Artistry Opportunity Grant from Northern Illinois University, and the University of Waterloo International Research Partnership Grants (IRPG).

REFERENCES

[1] Poor-quality data imposes costs and risks on businesses, says new forbes insights report. https://www.forbes.com/sites/forbespr/2017/05/31/poor-quality-data-imposes-costs-and-risks-on-businesses-says-new-forbes-insights-report.

[2] P. D. Allison. *Missing data*. Sage publications, 2001.

[3] R. Andreasson and M. Riveiro. Effects of visualizing missing data: an empirical evaluation. In *International Conference on Information Visualisation*, pp. 132–138. IEEE, 2014.

[4] A. N. Baraldi and C. K. Enders. An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37, 2010.

[5] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2):14–17, 2009.

[6] P. P.-S. Chen. The entity-relationship model—toward a unified view of data. In *Readings in Artificial Intelligence and Databases*, pp. 98–111. Elsevier, 1988.

[7] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 490–497, 2001.

[8] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[9] G. Convertino, R. K. Tayi, S. Tomar, M. Gupta, and C. Kakwani. Method, apparatus, and computer-readable medium for missing data identification, Aug. 22 2019. US Patent App. 15/901,255.

[10] M. Destandau and J.-D. Fekete. The missing path: Analysing incompleteness in knowledge graphs. *Information Vis.*, 20(1):66–82, 2021.

[11] P. Dragicevic, Y. Jansen, A. Sarma, M. Kay, and F. Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2019.

[12] C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: Graph interpretation user study. In *IFIP Conference on Human-Computer Interaction*, pp. 861–872. Springer, 2005.

[13] B. Efron. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426):463–475, 1994.

[14] S. J. Fernstad. To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. *Information Visualization*, 18(2):230–250, 2019.

[15] S. J. Fernstad and R. C. Glen. Visual analysis of missing data—to see what isn't there. *2014 IEEE Conference on Visual Analytics Science and Technology*, pp. 249–250, 2014.

[16] J. Heckman. Varieties of selection bias. *The American Economic Review*, 80(2):313–318, 1990.

[17] F. Hughes and D. Schum. Discovery-proof-choice, the art and science of the process of intelligence analysis-preparing for the future of intelligence analysis. *Joint Military Intelligence College, DC*, 2003.

[18] J. Hullman. Why authors don't visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139, 2019.

[19] M. W. Johnson and A. G. Lafley. *Seizing the white space: Business model innovation for growth & renewal*. Harvard Business Press, 2010.

[20] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information Visualization*, pp. 154–175. Springer, 2008.

[21] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee. A taxonomy of dirty data. *Data Mining & Knowledge Discovery*, 7(1):81–99, 2003.

[22] A. King. He spent 27 years wrongly convicted of murder. he wants to spend the rest of his life encouraging inmates to read. https://www.cnn.com/2018/09/08/health/john-bunn-exonerated-literacy-trnd/index.html.

[23] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.

[24] T. Li, Y. Belghith, C. North, and K. Luther. Crowdtrace: Visualizing provenance in distributed sensemaking. In *2020 IEEE Visualization Conference (VIS)*, pp. 191–195. IEEE, 2020.

[25] X.-Q. Mai, J. Luo, J.-H. Wu, and Y.-J. Luo. "aha!" effects in a guessing riddle task: An event-related potential study. *Human Brain Mapping*, 22(4):261–270, 2004.

[26] T. H. Matta, J. C. Flournoy, and M. L. Byrne. Making an unknown unknown a known unknown: Missing data in longitudinal neuroimaging studies. *Developmental Cognitive Neuroscience*, 33:83–98, 2018.

[27] N. McCurdy, J. Gerdes, and M. Meyer. A framework for externalizing implicit error using visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):925–935, 2018.

[28] C. North. Toward measuring visualization insight. *Computer Graphics and Applications*, 26(3):6–9, 2006.

[29] I. Ogrodnik. Missing student wrongly linked to boston marathon bombing found dead. https://globalnews.ca/news/510378/.

[30] G. Panagiotidou, R. Vandam, J. Poblome, and A. Vande Moere. Implicit error, uncertainty and confidence in visualization: an archaeological case study. *IEEE Trans. on Visualization & Computer Graphics*, 2021.

[31] R. K. Pearson. The problem of disguised missing data. *Acm Sigkdd Explorations Newsletter*, 8(1):83–92, 2006.

[32] T. D. Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

[33] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, vol. 5, pp. 2–4. McLean, VA, USA, 2005.

[34] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2015.

[35] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[36] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu. Generating synthetic missing data: A review by missing mechanism. *IEEE Access*, 7:11651–11667, 2019.

[37] J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147, 2002.

[38] Y. L. Simmhan, B. Plale, D. Gannon, and S. Marru. Performance evaluation of the karma provenance framework for scientific workflows. In *International Provenance and Annotation Workshop*, pp. 222–236. Springer, 2006.

[39] D. J. Simons and C. F. Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.

[40] H. Song and D. A. Szafir. Where's my data? evaluating visualizations with missing data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):914–924, 2018.

[41] M. Sun, L. Bradel, C. L. North, and N. Ramakrishnan. The role of interactive biclusters in sensemaking. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1559–1562, 2014.

[42] M. Sun, G. Convertino, and M. Detweiler. Designing a unified cloud log analytics platform. In *International Conference on Collaboration Technologies and Systems*, pp. 257–266. IEEE, 2016.

[43] M. Sun, J. Zhao, H. Wu, K. Luther, C. North, and N. Ramakrishnan. The effect of edge bundling and seriation on sensemaking of biclusters in bipartite graphs. *IEEE Transactions on Visualization and Computer Graphics*, 25(10):2983–2998, 2018.

[44] J. E. Trost. Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative Sociology*, 9(1):54–57, 1986.

[45] D. A. Van Dyk and X.-L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

[46] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *IEEE Conference on Visual Analytics Science and Technology*, pp. 104–115. IEEE, 2017.

[47] E. Wall, J. Stasko, and A. Endert. Toward a design space for mitigating cognitive bias in vis. In *IEEE VIS Conf.*, pp. 111–115. IEEE, 2019.

[48] J. Zhao, M. Sun, F. Chen, and P. Chiu. Missbin: Visual analysis of missing links in bipartite networks. In *2019 IEEE Visualization Conference (VIS)*, pp. 71–75. IEEE, 2019.

[49] J. Zhao, M. Sun, F. Chen, and P. Chui. Understanding missing links in bipartite networks with missbin. *IEEE Transactions on Visualization and Computer Graphics*, 28(6):2457–2469, 2022.