https://doi.org/10.2352/EI.2023.35.1.VDA-405


# Case Study on Including Ethics into Introductory Data Visualization

***Anna Baynes; California State University, Sacramento; Sacramento, CA;***
***Email: shaverdian.csus.edu***

## Abstract

*Given the amount of data created and available to everyone, there needs to be more consumable everyday data analytic tools for anyone to make sense of their data. At Sacramento State University, we designed an introductory data visualization course to teach college students how to analyze data. Students enrolled in this six-week summer course and used utilized Trifacta [1], Tableau [2], and ObservableHQ [3] on the IEEE VAST Challenge 2022 dataset. The course emphasized the uncertainty and deception involved in data visualization. We structured the course around ethical design choices. In this paper, we describe an overview of the ethical goals of our six-week data visualization summer course, review example student work on the IEEE VAST Challenge, and provide recommendations for ways to add ethical functionality to visualization tools. The goal is to have more consumable data visualization tools for novice users and support ethical design choices.*

## Introduction

Future data scientists learn Economist Herbert A. Simon's quote, "A wealth of information creates of poverty of attention." While this quote is accurate, a more practical problem is people have access to a wealth of information but generally are not equipped with the tools to consume and understand it, which leads to deception and distress in the community. College students often enroll in writing courses to improve their written communication. Given the amount of data generated and open access data available, it is an important skill to know how to utilize it correctly. At Sacramento State University, we created an introductory data visualization course. Students enrolled in a six-week summer course and used data visualization tools such as Trifacta [1], Tableau [2], and ObservableHQ [3]. The data provided during the course was from the IEEE VAST 2022 Data Visualization Challenge. The Challenge considered a fictional town called Engagement, Ohio, soliciting urban planning data scientists to help officials understand the city's trends and life patterns to identify growth opportunities. This dataset is immense for students but realistic for real data issues. The dataset includes town participants slightly over year log activity logs on sleep, hunger, and food budget, which totals around 18GB of data, and other attributes. The dataset also includes city buildings, social networks, and more attributes. The difficulty in the data challenge is the sheer size and heterogeneity of the datasets. Several separate files had to join for practical analysis. These experiences of size and combining different datasets are realistic for data scientists.

There is a noticeable gap in the field of ethics related to data visualization and data science. We concentrated on the data analysis challenge, which led to data quality issues during the class. Beyond this data challenge, we focused on the data encoding and visualization tools and emphasized ethics and deceptive data visualization challenges. By the end of the course, students created visualizations using Trifacta for data transformation and Tableau and ObservableHQ platform for data exploration. This paper reviews examples of students' visualizations and how well they incorporate ethics into their work. The research provides ways to introduce ethical concepts into data visualization tools to ensure a path toward honest data science practices.

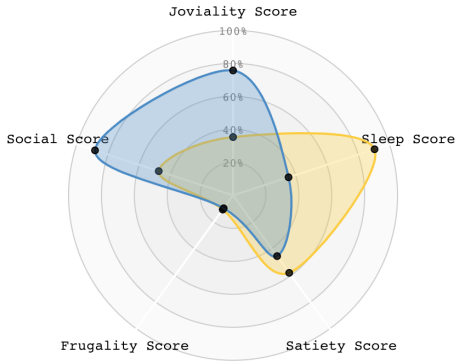


***Figure 1.*** *Radar overview graph students designed to compare two participants in the town.*

Machine learning leads to multiple innovations like email spam filters, virtual assistants, and intelligent devices that reduce human effort. As with every field, data science has morals to be followed. For example, self-driving cars face ethical choices in case of accidents and emergencies. One of the biggest hurdles in data science is ambiguous and missing data. This dirty data flows to downstream systems where it is difficult to clean and aggregate data. The dirty data can also lead to unreliable and deceptive insights. Students need to learn the social, ethical, and legal issues that can arise when solving data problems. While there is a wealth of information related to data science ethics, there is an analogous need for platforms tailored to ethical data visualization. Recent work shows unethical practices in data analysts, journalists, and visual developers knowingly or unknowingly misusing visualizations to mislead readers [27]. Corrupt practices arise due to a lack of knowledge, not only due to intentionally making unethical design choices. Deceptive design choices include examples such as misleading choices, skewed to only one stakeholder, incorrect data transformations, and strong existence of bias.

This paper presents a pressing need for a cohesive design and platform for ethical data visualization tools. This work aims to give our findings on how introductory students attempt to achieve

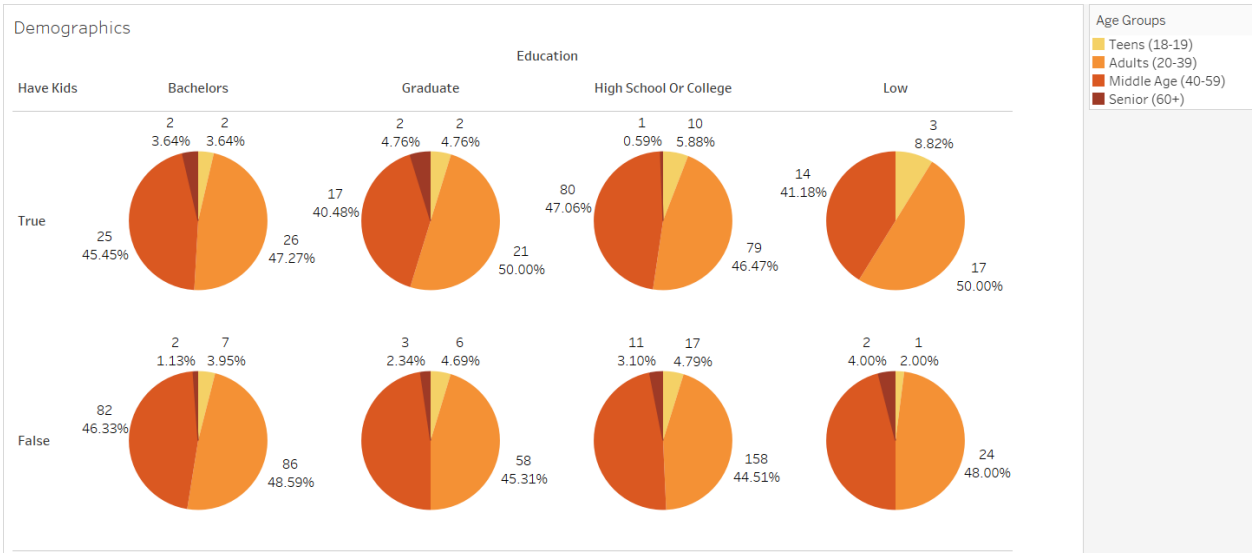


***Figure 2.*** *Students created this Tableau chart to show data skew in the town.*

ethical data visualization. We recommend updating current data visualization tools to incorporate ethics based on our review. For example, data exploration tools reduce the risk of biased and deceptive problem-solving.

In this paper, we present the following contributions: design of the six-week data visualization summer course for introductory students with a strong emphasis on ethics, presentation and review of example student work on the VAST Challenge during the course, recommendations for ways to incorporate ethical design into data visual exploration tools. In the following sections, we review related ethics and data visualization works, then present the VAST Challenge-centered introductory data visualization course. Next, we review the student's work for how well they include ethics in their design choices. Finally, we conclude with recommendations for more ethic-based data visualization platforms.

## Related Work

The work presented is an intersection of data visualization, computer science education, and ethics in computer science education. In this section, we describe the relevant results which inspired and contributed to our background for the paper.

The data visualization community has explored how to incorporate uncertainty, which often leads to ethical mishaps, into visualization design [26], [27], and [28]. In [12] authors present empirical studies on uncertainty visualization where they experiment with the typology of uncertainty and visual semiotics. They show examples of visual encodings effective in visualizing uncertainty. In [13] continue this work by unpacking how uncertainties propagate in visual analytics and provide guidelines to design uncertainty-aware systems for improved decision making. Furthermore, considering accessibility for users is another path for ethical data visualization design [29], [30], and [31]. For example, [14] and [17] consider techniques to improve automatic visualization captioning in accessible and meaningful content for users' preferences. Authors in [15], and [16] recognize that users' cognitive load is affected in decision-making with uncertain values; they provide recommendations for future missing data visualization systems. In our work, we present recommendations for ethical visualization systems [32] and [33].

The data visualization course incorporated culturally responsive pedagogy practices to introduce data science topics. Integrating culturally responsive pedagogy with computer science is an emerging area to improve computer science education. In [34] and [35], authors explore case studies of practical approaches to culturally responsive pedagogy for computer science education. Data science is flawed with ambiguous and incomplete data, often termed dirty data. This dirty data flows downstream to data warehouses, where it is challenging to aggregate data, making it unreliable to derive insights. Due to dirty data, data visualization students learn to deal with data comprehensively and fairly [18], [19]. This Challenge requires students to understand the ethical process of designing a data model to avoid biased outcomes [20]. While there are resources where students can learn about data science ethics, there is an opportunity for the content to be effectively tailored for introductory college students [21]. There is a need for more platforms where students can understand how to build data science models in an ethical way [22]. Authors in [23] propose 12 different ethnic themes. They provide a comprehensive classification of the code of ethics.

Data feminism [5] lays out principles on how identity, power, and knowledge arrive and how that process affects gender, race, and class [31]. Data feminism encourages listening to multiple voices instead of one loud voice [25]. This approach is an example that leads to data scientists investigating uncertainty. The ethical impact of computing is a topic taught in computer science education. The data feminist principles of uncertainty, bias, dissent, and further ethical considerations appear in several related works [8], [9], [10], [11]. These previous works supported our class discussions on ethical visual developers.

## Case Study Description

In this section, we describe the introductory data visualization class and the IEEE VAST Challenge used in our case study of how well novice visual developers include ethical approaches

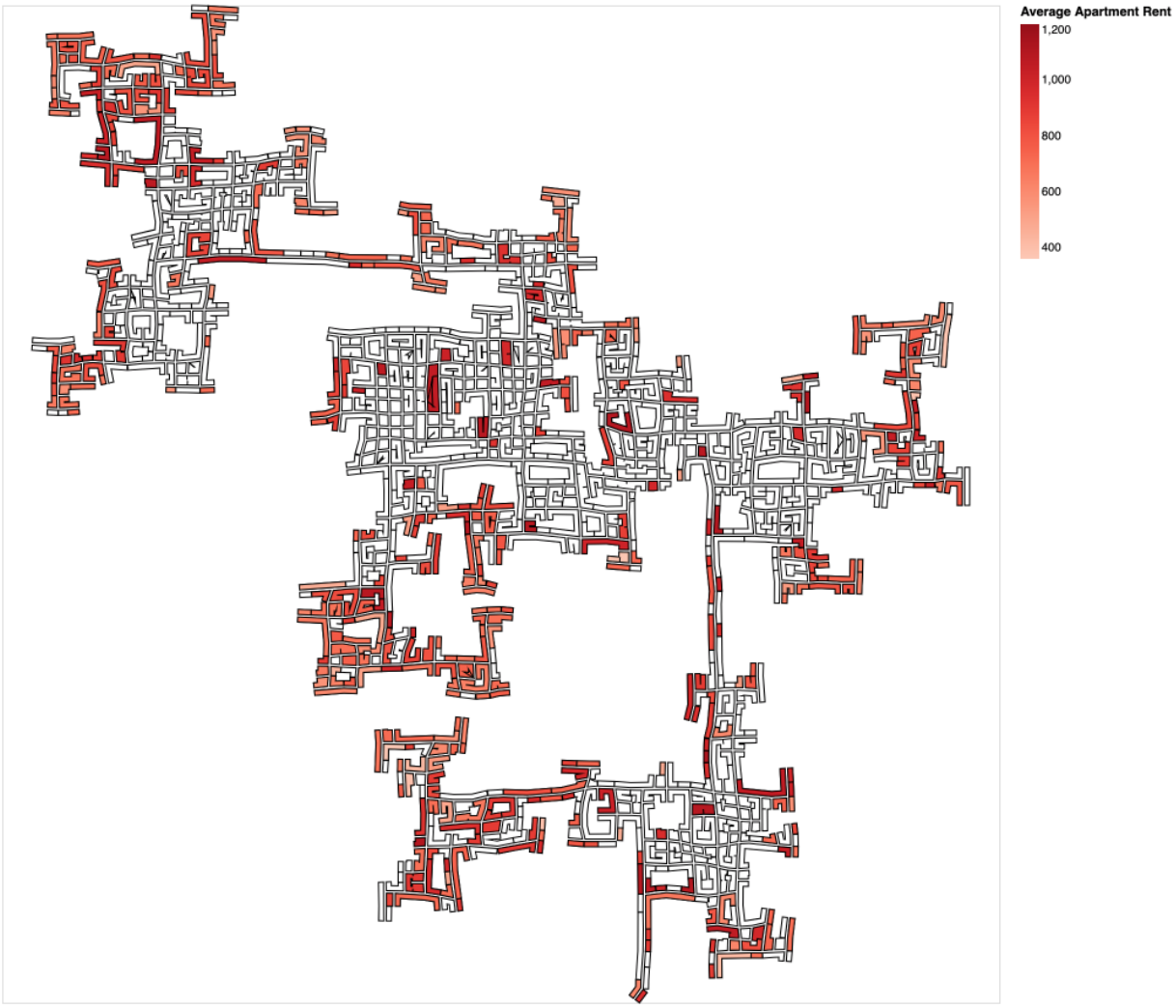


***Figure 3.*** *Students incorporate interaction techniques to allow users to investigate the chart.*

in their work.

## *Data Visualization Course*

The summer six-week data visualization course met twice a week for two hours. There were 35 third-year college students enrolled. The class began with visual encoding, and design considerations taught students how to utilize Tableau, Trifacta, and ObservableHQ in the data visual analytic pipeline. The course introduced students to data uncertainty, wrangling, and examples of deception in data visualizations. They saw examples of unintelligible, incorrect, deceptive, and lack of knowledge visualizations. As a class exercise, students were asked to create misleading visualizations and explain the deception used. Students spent most of the class analyzing the IEEE VAST 2022 Challenge data set. They used the data analytic pipeline's Trifacta, Tableau, and ObservableHQ toolsets.

## *Class Challenge Dataset*

To ground the class in a data analytic mission, we used the IEEE VAST Challenge to motivate the class. IEEE VAST annually produces a fictional scenario with various datasets to push the visual analytic community to create novel contributions. The IEEE VAST 2022 Challenge story centered around a bedroom community suddenly seeing new incoming population growth. Due to the rapid growth, the town officials invite urban planners to understand the current trends and patterns of the city and identify how to support future growth. About 1000 town participants recorded their daily logs of the places they visit, spending, happiness, and other attributes in an urban planning app. This dataset, including social networks and attribute information on apartments, buildings, employers, jobs, pubs, restaurants, and schools, provides segmented details on the town for analysis.

The Challenge is to utilize the several GB of data which requires combining and filtering datasets carefully to produce insights. The insights must answer the following three challenge components: First, what are life patterns for residents throughout

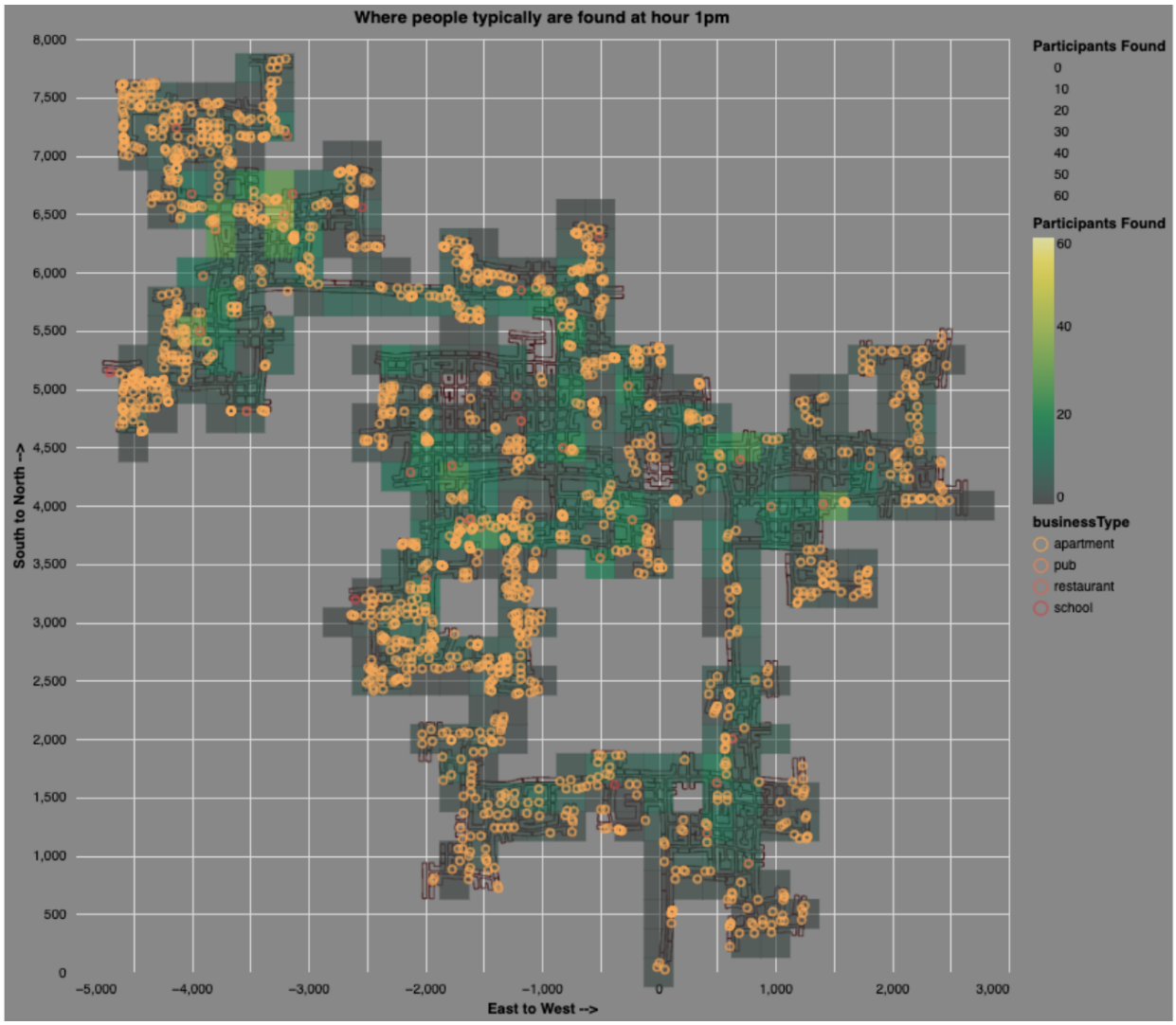


***Figure 4.*** *Multiple attributes are overlaid on a map of the town.*

the city? For example, are there potential bottlenecks and exciting changes over time? Second, what are the demographics and relationships involved in the social networks and city? Third, what is the overall economic health of the town? Students initially focused on one of the dataset files. But students quickly realized to make any insights involving the challenge questions; they must combine datasets. Students used Trifacta to support their dataset file combining, filtering, and data transformation. Then they used Tableau to make initial data exploration with visualizations. Finally, the students used ObservableHQ with Vega-Lite visualization languages to implement their interactive visualizations of the city's map with several layered attributes.

## Analysis of Example Student Visualizations and Analysis of Ethical Considerations

To ground the class in a data analytic mission, we used the IEEE VAST Challenge to motivate the class. IEEE VAST annually produces a fictional scenario with various datasets to push the visual analytic community to create novel contributions. The IEEE VAST 2022 Challenge story centered around a bedroom community that suddenly saw new incoming population growth. Due to the rapid growth, the town officials invite urban planners to understand the current trends and patterns of the city and identify how to support future growth. About 1000 town participants recorded their daily logs of the places they visit, spending, happiness, and other attributes in an urban planning app. This dataset, including social networks and attribute information on apartments, buildings, employers, jobs, pubs, restaurants, and schools, provides segmented details on the town for analysis.

## Visualization Results

This section reviews several examples of the students' projects. We check them for their proximity to ethical design choices. We also share the students' design rationale and compare that to ethical design choices. Students worked in teams or individually. Several groups presented overview visualizations to understand the attributes before delving into the questions and ethical decisions. For example, one group designed the radar graph

shown in Figure 1. They chose to display different attributes and compare participants' design rationale. In Figure 2 the students show an overview graph designed in Tableau of the diverse demographic subgroups in the city. They recognize that most participants are adults or middle-aged; hence, the calculations may skew for the other subgroups because the sample size is small. This recognition displays the students are carefully examining the data and providing visualizations to highlight possible deception, which can unknowingly seep into other visualizations.

Students' visualizations that looked at the overall urban planning design incorporated the city's map. For example, Figure 3 uses a map to display rent costs across the city. This student team integrated a slider in their ObservableHQ visualization to allow user-friendly interaction. In their design rationale, they consider issues such as the accuracy of the average rent shown and only showing the information the user requests. By incorporating interactive elements, this student team believes in ethical design.

Students presented innovative visualizations that provided overlays and attributes on the city map. The student provides a functional, concise visualization but at a consumable level of detail. For example, in Figure 4, the student considers issues such as improving the facilitation of transit by focusing the city's resources on connecting the outer edges to the higher priority corridors and supporting the city's peak hours. This effort is challenging with a large, segmented, and heterogeneous dataset. But to avoid deceptive visualization, the student discusses the need for brief and easily user-consumable visualizations in the rationale. The student sees in their visualization that there is a lack of recreation or food near employment centers, which may negatively affect what the town's citizens can do during the day.

Students present various ethical reasoning types in the visual design rationale. The students show examples of creating visualizations specifically to highlight data skew. Students also developed interactive features to support ethical reasoning. And finally, students developed succinct but encompassing many attributes for in-depth urban planning analysis.

## Suggestions for Visual Encoding Ethics

Visual encoding helps translate the data into a visual element on various charts—effective and expressive visualizations aid users in understanding the data set and seeing the value in the visualization. Students are taught in data visualization theory the rules for applying visual elements and channels to different data types to produce effective and expressive representations. For example, Bertin presented seven categories of visual variables: position, size, shape, weight, color, orientation, and texture. In more current related work discussed earlier in the paper, [12] authors present examples of visual encodings that helped users perceive the uncertainty behind the data. Specifically, by applying fuzziness to the visual element, users assumed higher uncertainty in the data. The students' tools this summer included Tableau, Trifacta, and ObservableHQ. They are primarily for visual developers or educational purposes.

We found a need for more support to help visual developers with these tools be mindful of the deception embedded in the visualizations they created. Students took a proactive ethical stance to avoid deceptive visualizations in their IEEE VAST Challenge solutions. But, we postulate that these visualization platforms can include functionality to support visual developers from unknowingly creating misleading visualizations. This gap relates to work in [5], which ties ethical issues on identity, power, and knowledge to data visualization. The three principles presented in [5] include: ways to represent data unknowns, reference the material economy behind the data, and make dissent possible. Data science has uncertainty in every stage of the analytic pipeline. The first principle warns us to stay cautious of these uncertainties. The second principle recognizes a range of stakeholders and the importance of understanding each bias. The third principle challenges the visual developer to create interactive features that allow users to question facts and realities.

Visualization tools should implement the three principles into their framework for developers to validate that their final chart does not have ethical or deceptive issues. For example, the functionality can be a metric that rates the chart's effectiveness in the three principles. Or, it can add visual encoding, such as fuzziness, to show there may be stakeholders or missing values causing deception. As people become more aware of the amount of data they have access to and visualization tools become more consumable for the average user, having ethical verification in the visualization tool will avoid potential deception pitfalls.

## Conclusion

In this paper, we describe an overview of the ethical goals of our six-week data visualization summer course, review example student work on the IEEE VAST Challenge, and provide recommendations for ways to add ethical functionality to visualization tools. For future work, we plan to design functionality for incorporating principles to test the ethical value of visualization. Next, we plan to design and conduct experiments on various ways to add ethical functionality to visualization tools. These efforts will help reduce deceptive visualizations.

## References


[1] Hellerstein, Joseph M., Jeffrey Heer, and Sean Kandel. "Self-Service Data Preparation: Research to Practice." IEEE Data Eng. Bull. 41.2 (2018): 23-34

[2] Tableau Software Business Graphics Toolkit. http://www.tableausoftware.com.

[3] ObservableHQ. https://observablehq.com/.

[4] Rattenbury, Tye, et al. Principles of data wrangling: Practical techniques for data preparation. " O'Reilly Media, Inc.", 2017.

[5] Wyer, Mary, et al., eds. Women, science, and technology: A reader in feminist science studies. Routledge, 2013.

[6] Simpson, Amari T., et al. "Impacting Teacher and Counselor Practices as They Support Traditionally Underrepresented Students to Pursue STEM Majors and Careers." 2020 IEEE Frontiers in Education Conference. IEEE, 2020.

[7] Mithun, Shamima, and Xiao Luo. "Design and Evaluate the Factors for Flipped Classrooms for Data Management Courses." 2020 IEEE Frontiers in Education Conference. IEEE, 2020.

[8] Smith, Sally, et al. "Computing degree apprenticeships: An opportunity to address gender imbalance in the IT sector?." 2020 IEEE Frontiers in Education Conference. IEEE, 2020.

[9] Peters, Anne-Kathrin, et al. "Care ethics to develop computing and engineering education for sustainability." 2020 IEEE Frontiers in Education Conference. IEEE, 2020.

[10] Scott, Andrew, and Scott Barlowe. "How software works: Com-

putational thinking and ethics before CS1." 2016 IEEE Frontiers in Education Conference. IEEE, 2016.
[11] Thoroughman, Kurt A., and Joseph A. O'Sullivan. "High impact practices toward personal and professional identity in introductory and advanced engineering seminar courses." 2016 IEEE Frontiers in Education Conference. IEEE, 2016.
[12] MacEachren, Alan et al. "Visual Semiotics / Uncertainty Visualization: An empirical study. IEEE VIS, 2012. "
[13] Sacha, Dominik, et al. "The role of uncertainty, awareness, and trust in visual analytics." IEEE transactions on visualization and computer graphics 22.1 (2015): 240-249.
[14] Lundgard, Alan, and Arvind Satyanarayan. "Accessible visualization via natural language descriptions: A four-level model of semantic content." IEEE transactions on visualization and computer graphics 28.1 (2021): 1073-1083.
[15] Song, Hayeong, et al. "Understanding the Effects of Visualizing Missing Values on Visual Data Exploration." 2021 IEEE Visualization Conference (VIS). IEEE, 2021.
[16] Fernstad, Sara Johansson. "To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization." Information Visualization 18.2 (2019): 230-250.
[17] Walsh, Erin I., Ginny M. Sargent, and Will J. Grant. "Not just a pretty picture: Scientific fact visualisation styles, preferences, confidence and recall." Information Visualization 20.2-3 (2021): 138-150.
[18] K. E. Martin, "Ethical issues in the big data industry," MIS Quarterly Executive, vol. 14, no.1, pp. 2-3, June 2015.
[19] J. Fairfield and H. Shtein, "Big data, big problems: Emerging issues in the ethics of data science and journalism," Journal of Mass Media Ethics, vol. 29, no. 1, pp. 38-51, January 2014.
[20] M. d'Aquin, P. Troullinou, N. E. O'Connor, A. Cullen, G. Faller and L. Holden, "Towards an 'Ethics by Design' Methodology for AI Research Projects," in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, pp. 54-59, December 2018.
[21] H. V. Jagadish, "Data Science Ethics," [Online]. Available: https://www.coursera.org/learn/data-science-ethics
[22] C. D. Martin and E. Y. Weltz, "From awareness to action: Integrating ethics and social responsibility into the computer science curriculum," ACM SIGCAS Computers and Society, vol. 29, no. 2, pp. 6-14, March 1998.
[23] J. S. Saltz, N. I. Dewar and R. Heckman, "Key concepts for a data science ethics curriculum," in Proceedings of the 49th ACM technical symposium on computer science education, Baltimore, MD, USA, pp. 952-957, February 2018.
[24] C. Spradling, L. K. Soh and C. Ansorge, "Ethics training and decision-making: do computer science programs need help?," in Proceedings of the 39th SIGCSE technical symposium on Computer science education, Portland, OR, USA, pp. 153-157, March 2008.
[25] M. Loi and M. Christen, "How to Include Ethics in Machine Learning Research," Research and Society, vol. 1, no.5, pp. 5-6, January 2019.
[26] Sanyal, Jibonananda, et al. "A user study to compare four uncertainty visualization methods for 1d and 2d datasets." IEEE transactions on visualization and computer graphics 15.6 (2009): 1209-1218.
[27] Hullman, Jessica, et al. "In pursuit of error: A survey of uncertainty visualization evaluation." IEEE transactions on visualization and computer graphics 25.1 (2018): 903-913.
[28] Ruginski, Ian T, et al. "Non-expert interpretations of hurricane forecast uncertainty visualizations." Spatial Cognition and Computation.
[29] Kale, Alex, Matthew Kay, and Jessica Hullman. "Visual reasoning strategies for effect size judgments and decisions." IEEE transactions on visualization and computer graphics 27.2 (2020): 272-282.
[30] Kale, Alex, et al. "Hypothetical outcome plots help untrained observers judge trends in ambiguous data." IEEE transactions on visualization and computer graphics 25.1 (2018): 892-902.
[31] Sacha, Dominik, et al. "The role of uncertainty, awareness, and trust in visual analytics." IEEE transactions on visualization and computer graphics 22.1 (2015): 240-249.
[32] Wood, Jo, et al. "Sketchy rendering for information visualization." IEEE transactions on visualization and computer graphics 18.12 (2012): 2749-2758.
[33] Xiong, Cindy, Lisanne Van Weelden, and Steven Franconeri. "The curse of knowledge in visual data communication." IEEE transactions on visualization and computer graphics 26.10 (2019): 3051-3062.
[34] Madkins, Tia C., et al. "Culturally relevant computer science pedagogy: From theory to practice." 2019 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT). IEEE, 2019.
[35] Codding, Diane, et al. "Culturally responsive and equity-focused computer science professional development." Society for Information Technology and Teacher Education International Conference. Association for the Advancement of Computing in Education (AACE), 2019.

## Author Biography

*Anna Baynes is an Assistant Professor at California State University, Sacramento.*