Visualizing Open Data: Public Perceptions on Privacy

CHANG HAN*, University of Utah, USA
MANILA DEVARAJA*, University of Utah, USA
SHARMIN AHMED*, University of Utah, USA

Previous studies on the perception of visualization have often focused on its efficacy and emotional expressivity. However, there has been a paucity of research on how visualization influences people's perception of the data itself, particularly in terms of privacy. This research project centers on the fundamental elements of how visualization impacts privacy, specifically exploring how the form of visualization differently, as opposed to tabular data presentation, affects individuals' perception of data privacy. To investigate this issue, we conducted surveys and semi-structured interviews. From our investigation, we discovered that visualization indeed impacts the perception of data privacy, but this influence often "reinforces existing impressions" rather than completely reversing privacy perceptions. That is, for data deemed privacy-sensitive by participants, they tend to believe that visualization exacerbates these concerns. Conversely, for data not viewed as privacy-intrusive, participants favor the convenience in data access and perception that visualization offers.

Additional Key Words and Phrases: **Human-computer interaction(HCI), Visualization, Privacy for public data**

1 INTRODUCTION

Visualization of public data is vital in increasing accessibility and promoting transparency, allowing people to engage in different subjects and make informed decisions. Despite the abundance of publicly available data, little is known about how its presentation shapes public understanding and attitudes. A compelling instance of this situation is an interactive map published by The Journal News [12], where the map enables users to identify which residents hold gun permits. Although the information was sourced legally [1], indicating that the data was already public to a certain degree, the map incited significant unease among permit holders [12].

Historical discussions on data visualization have often emphasized its accuracy and effectiveness [3], but recent developments in data privacy highlight the need for a more nuanced approach. While visualizations can effectively convey information, their impact on privacy and the sentiments of those represented must be carefully considered. This is particularly crucial as regulations like the Information Quality Act [2] address the credibility of public data yet fail to guide the visual presentation of such data, significantly influencing public perception. Our research aims to address this gap, focusing on the intersection of privacy and visualization in the context of open data, promising to inform and evolve current practices and policies.

We conducted seven semi-structured interviews and a survey with 7 participants. The interviews and survey involved presenting participants with raw and visualized forms of three different public data and assessing their initial awareness, reactions, and subsequent perceptions. Qualitative analysis was performed to derive participants' main themes and concerns about the visualizations and the public data itself. The findings from the analysis indicate that data visualization profoundly affects the public perception of its accessibility. Additionally, despite the potential risks of data breaches, individuals often maintain a supportive stance toward the open data initiative. For data that participants consider to be privacy-sensitive, there is a prevailing belief that visualization amplifies these privacy concerns.

*These authors contributed equally to this work.

Authors' addresses: Chang Han, changhan@sci.utah.edu, University of Utah, Salt Lake City, Utah, USA; Manila Devaraja, Manila.Devaraja@utah.edu, University of Utah, Salt Lake City, Utah, USA; Sharmin Ahmed, sharmin.ahmed@utah.edu, University of Utah, Salt Lake City, Utah, USA.

The work contributes to the field of data visualization by elucidating the subtle ways in which data visualization influences public perception, especially regarding privacy and ethical issues. The research will provide insights garnered from diverse groups of people, underscoring the importance of adopting culturally sensitive and ethically responsible methods in data presentation in an era increasingly dominated by data-driven decision-making.

2 LITERATURE REVIEW

Visualization has transformed how data looks and what we make out of data. The evolution of visualization has proved to be greatly useful in the process of data analysis and wrangling. For a long time, the ethical consideration of visualization was not in the picture [6]. From the investigation, Tang et al. identifies and discusses three main challenges of visualization research: making the invisible visible, collecting data with empathy, and challenging power structures. Visualization forms incorporate different features and investigate what representation makes people more concerned with privacy [14]. Holder and Bearfield's work analyzes how particular design elements in visualizing political polls can significantly influence public opinion. Discussions in [8] regarding the responsibility of those creating visualizations, the implications of misrepresentation through visual design, and the ethical considerations in presenting politically charged data inform us of the ethical considerations.

The user's perception of a design multiple-step process to execute, [5] gives a peek into how the system design can be improved by considering end-users. The methodology and the scale are significant factors in determining social inputs to users' privacy perspectives. While open data promotes transparency, accessibility, and informed decision-making, open data can also bring new challenges concerning privacy, as to legal protections (who owns the data), how it is maintained, and mainly, how much sensitive information it contains and how it is protected [7]. The open data can include information that can be considered sensitive by the end-user [11]. Ansari et al. claims that civil programmers and common citizens are likely more interested in consuming visualizations than creating them. However, these investigations do not consider how citizens feel about the visualization. As many people normally cannot even comprehend, open data can be used; upon the specific visualization, it can even perceived as some leaked information [9]. The findings offer us a more nuanced view of how people worldwide perceive the usage of their data for research and investigation. The investigation with under-represented populations revealed that motivation, preferences, beliefs, and personal connection to the data of visualization could have an impact on how they perceive the visualization [13]. Knudsen et al. elaborates on the design considerations and the interactive visualization systems developed to make this data more accessible and engaging for the general public. They show that the user interactions and the feedback received, indicate people's comfort or concerns regarding such visualizations [10]. Although visualizing the data improves users' perception, how users feel about those visualizations is something to be questioned. Visualization of public data can introduce ethical concerns, which will be focused on through our work.

Visualizing publicly accessible data significantly aids in understanding trends and making informed decisions. However, the ease of access and usability of such data can also bring ethical concerns, such as privacy issues that don't exist until the data is visualized. Our research aims to answer the following questions:

- RQ1: Does visualization alter the perception of open data as public or private?
- RQ2: What specific forms of data visualizations make individuals feel uncomfortable or raise ethical concerns?
- RQ3: How does including personal data in the visualizations of public data impact people's perception?

3 METHODOLOGY

For our research investigation, we collected data in two methods: interviews and survey data. Both data were qualitatively analyzed to draw out the findings of the paper.

3.1 Interview

The recruitment process for our study was designed to capture a diverse range of perspectives. We strategically targeted participants from various nationalities and ethnicities to ensure a broad representation in our sample. Prospective participants were primarily individuals within the group member's network, chosen for their potential to provide varied insights based on their different backgrounds. We utilized both in-person and virtual platforms, such as Zoom, for interviews to accommodate their varying schedules and locations. The final participant pool comprised 7 people, all of whom were over 18 years old.

Our research framework was finalized after multiple discussion sessions within our research group. These discussions led to the formulation of our three research questions. Subsequently, we brainstormed a set of interview questions designed to elicit responses that would offer insights into our research questions. Through a process of refinement and group consensus, we finalized these interview questions. We selected three sets of public data to present to our participants. To assess whether the data presentation format influences perception, we prepared both visualized and raw/tabular formats of these datasets. To mitigate any potential bias from the order of presentation, we randomly divided our participants into two groups: one group was shown the visualized format first, and the other started with the raw/tabular format. We conducted our interviews following a semi-structured format. This approach allowed us to explore the participants' spontaneous insights while ensuring that our main research questions were addressed. The semi-structured nature of the interviews allowed for a dynamic interaction between the interviewer and the interviewees, facilitating a deeper understanding of their perspectives. This method proved particularly effective in revealing nuanced views on data privacy, utility, and presentation, which were central to our research objectives.

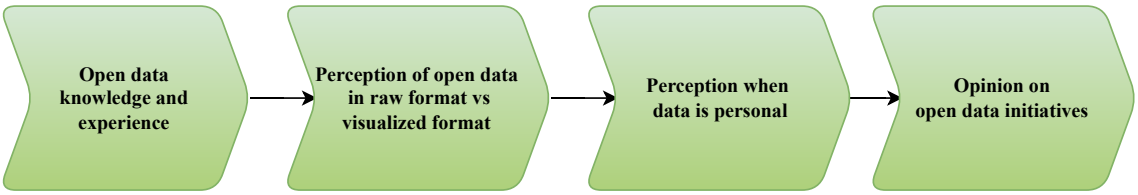

Fig. 1. Interview Design

Before starting the interview, consent was taken from the participant, providing the required information on their audio recordings and data storage details. After getting consent, we started the interview, which is divided into four part shown in Fig. 1. To ensure the participant understood public data, we first asked them about their knowledge, backgrounds, and personal experience with public data and visualization. We would give them some examples if they were unfamiliar with them. Then, we moved on to showing the visualized formats and raw formats of the public data and asked about their perception.

We did not make any substantial changes during the interview process. However, as per our initial agreement between the group members, we showed the visualized format first to three of the participants and the raw/tabular format first to the other four participants.

3.2 Questionnaire

To gain experience with crowdsourcing, we took surveys on Prolific as workers and found the interface to be very smooth. We found it insightful to understand the different formats of questions and the information they were trying to capture with different formats. We also got some perspectives on the differences in question structuring between surveys and interviews. We found the importance of crafting the survey questions to be very concise and clear. With clear and concise questions, participants understand the questions without spending much time and losing focus, and we can get quality responses. The task also helped us frame different kinds of attention-check questions. It's also worth mentioning that when people work as crowdsource workers, their goal is to earn more dollars, which often leads them into a state of impatience, making it hard to concentrate on the problem at hand and eager to finish it quickly. This influences us to formulate survey questions to be as short and easy to understand as possible. We also found that we can easily get tired and bored when asked to answer multiple surveys. So we tried to vary question types, using a mix of question types (multiple-choice, short answer, etc.) to keep the survey engaging and prevent monotony.

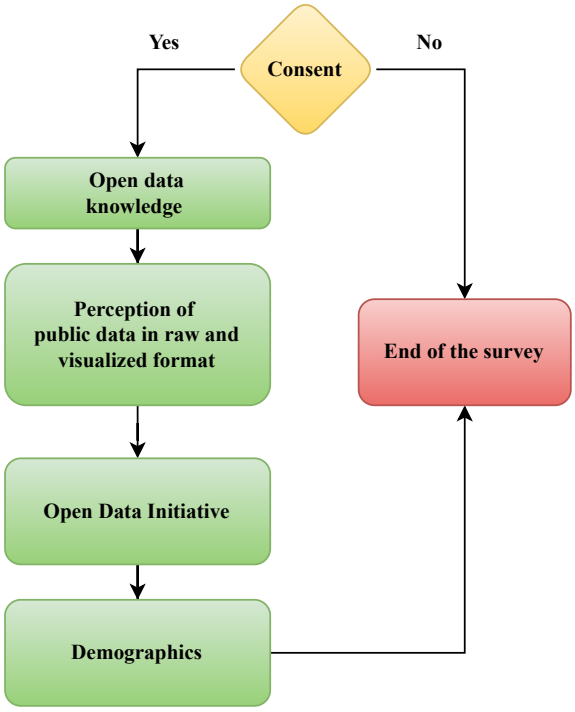

Fig. 2. Survey Design

3.2.1 Survey pilot and rewards. We piloted the survey in 2 stages, with 2 workers in each stage. We piloted 2 with ourselves to understand the average time needed to complete the survey and to detect if there is any problem with the survey questions. We then asked 2 other workers (our friends) to give us feedback about the wording and formats, which helped us understand how to frame the survey questions better to get unbiased and quality responses. The reward for a successfully completed task was $2. Based on our pilot study, we set the average time to complete the

task as 10 minutes. We went for a good pay range of $12/hr, which is fair to the participants. And they feel enough motivation to give quality responses.

3.2.2 Survey Qualification. We did not set any qualification criteria as we wanted to get data from a diverse population with various backgrounds, education levels, and professions. However, while deploying the survey, we limited it to US residents and fluent English speakers to ensure the participants understood the questions. We manually approved the tasks by cross-checking with their Prolific IDs and their responses, especially from the open-ended and attention questions. As crowdsource responses can be random, leading to poor-quality data that can not help answer our research questions, we manually approved the responses.

3.2.3 Survey Questionnaire. The online questionnaire had different sections partitioned based on the category of questions. The questionnaire started by asking for consent and continued only if the participant provided consent by clicking 'yes.' The first section's questions were on capturing participant's general knowledge about public data, visualization of datasets, and their expertise in visualizing a given dataset. We selected two datasets to understand if participants had any concerns with the visualization of the data and the format of visualizations. The questions had different formats, from multiple choice questions to choice questions to the Likert scale. We included open-ended questions to get the participant's concerns definitively. The section on the Open Data Initiative was included to understand the participants' knowledge about it and also concerns if they had any. A text box question was included to understand how the Open Data Initiative can improve without leaking personal information. The survey ended with demographic questions. Before deploying the survey, we piloted the survey in 2 stages with 2 participants each. The online questionnaire was conducted among 7 participants through the Prolific crowdsourcing platform.

3.3 Public Dataset for Visualization

For our investigation, we chose 3 public datasets to show participants to get insight into their privacy perceptions in visualized and raw/tabular formats. Our first dataset is "Real Estate Owner Information in Utah," which, in visualized map format Fig. 3 shows the owner's information for real estate in Utah [1]. You can click on the building, and it will show the owner's name and address.

Our second dataset is of crime reports in Salt Lake City, wherein in the visualized map format Fig. 4, one can zoom in or out and see the crime reported in that location, if any. Upon clicking a crime pointed at the map, the view shows the details of the crime type, timestamp of the crime report, and address [2].

Our third dataset is census data [3] of women between 15 and 50 years who had given birth in the last 12 months by their marital status and age in Utah Fig. 5.

4 FINDINGS

4.1 Interview Findings

In our group's analysis of interview results, despite a limited sample size (N=7), we observed notable differences that may stem from cultural backgrounds. For instance, one interviewee (P1) strongly opposed publicizing pregnancy data among 15 to 19-year-old females, fearing its exploitation for political or religious purposes. This view contrasts with another interviewee (P7), who advocated for the data's public accessibility, arguing it raises awareness and prompts

[1] https://slco.org/assessor/new/ParcelViewer/index.html?
[2] https://spotcrime.com/UT/SaltLakeCity/
[3] https://tinyurl.com/2r6myfu6

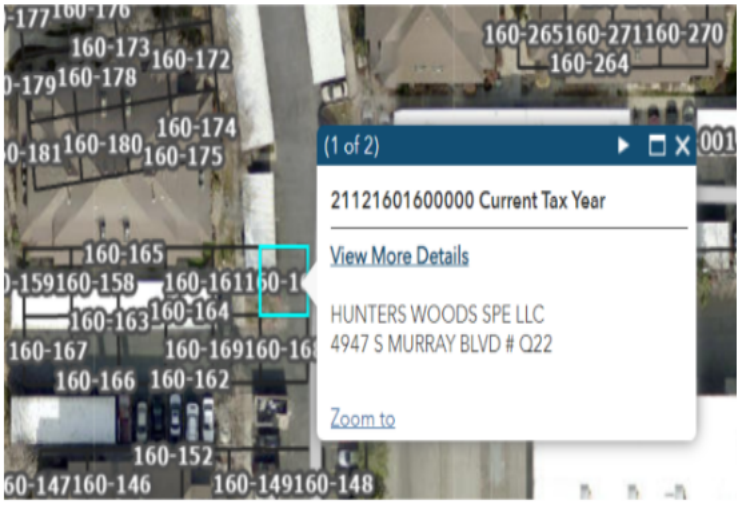

Fig. 3. Visualization of Real Estate Owner Information in Utah

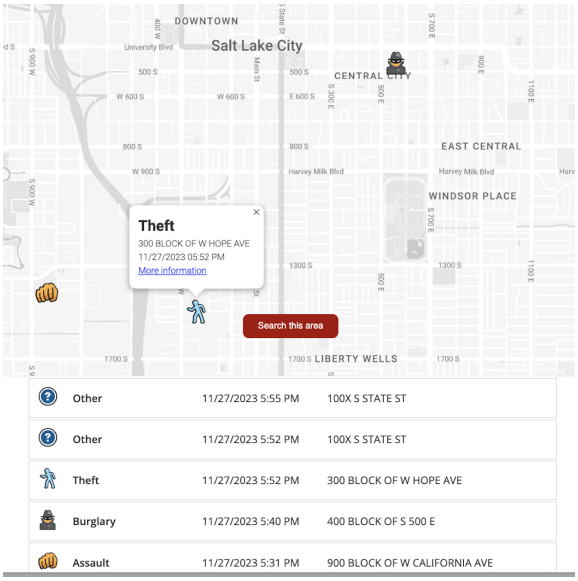

Fig. 4. Visualization of Crime Report in Salt Lake City

societal action. Both, however, agreed on the impact of data visualization, though with divergent interpretations. One felt that visualization rendered the data more unsettling, while the other believed it played a crucial role in alerting and guiding public perception. These differing viewpoints highlight how personal and cultural perspectives can influence data interpretation and presentation, leading to varied data privacy and utility stances. One interviewee (P6) explicitly mentioned the impact of cultural differences, noting that in her cultural context, the concept of privacy, especially

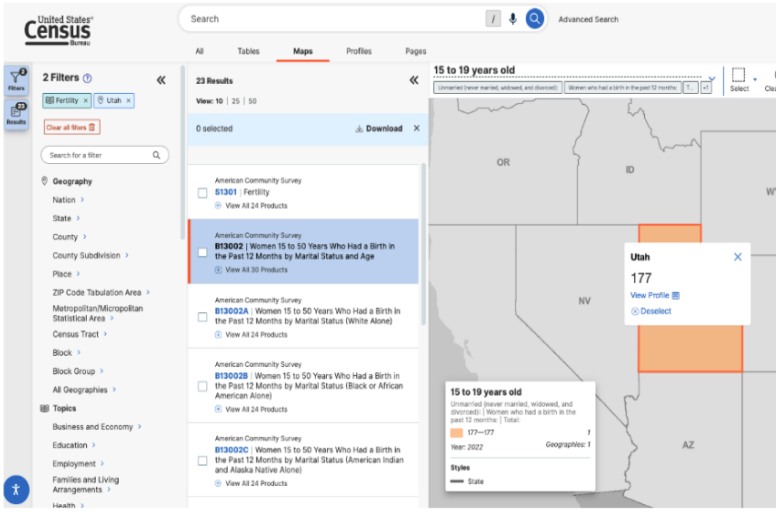

Fig. 5. Visualization of Birth Report of Women between 50-50 in Utah

regarding personal data, is virtually non-existent due to a lack of privacy protection. This lack of privacy has led to a general indifference towards privacy issues, indicating a stark contrast in the perception of data privacy across different cultures. It highlights the importance of considering cultural contexts when discussing privacy perceptions. The third dataset particularly brought up cultural and social concerns that people have about unmarried women giving birth (P1, P2, P6).

Overall, our interviews revealed a nearly unanimous concern about publishing land-owner data. Common worries included fears of malicious use of this information (P4), such as targeting individuals for crime, and concerns about privacy (P7), particularly the reluctance to have strangers know their home addresses. Interestingly, opinions on the perceived safety of a search panel versus a map interface in accessing this data were divided. Some did not find the search panel more reassuring than a map (P6, P1), as both provided the same information. However, a contrasting viewpoint (P7) was also expressed. One interviewee cited a scenario where a person following them could potentially use a map to locate their home, leading to heightened anxiety. This person argued that a map, by offering a real-world connection, posed more significant security risks than a search panel, which lacked this direct, visual link to their physical location. These findings underscore the complexity of data privacy concerns and the need to consider how information is presented and accessed carefully. 5 out of 7 interviewees had the same opinion of not having a problem with public data, but the concern is the identification of personnel using that information.

In our study, diverse opinions also emerged regarding a crime map. Many participants (P6, P7, P1, P4) appreciated the public availability of this data, valuing its utility in informing tourists or renters about neighborhoods with higher crime rates. However, skepticism was raised about the source of the data, with concerns that it might be user-submitted and thus lack credibility. The effectiveness of different data representation formats was another key topic. Most participants (P6, P7, P1, P4) affirmed the utility of maps (visualization) over simple tabular data (like a Tableau chart), believing that tables are nearly ineffective in conveying such complex information. Regarding privacy concerns, most interviewees (P6, P7, P1, P2, P4) did not perceive the crime map as a significant threat to individual privacy, as it did not disclose the

personal details of victims or perpetrators. However, some raised concerns about the potential negative impact of such a map on certain communities, especially if it was a sexual assault. They were also worried that highlighting areas with high crime rates might stigmatize those neighborhoods, causing distress for residents and landlords.

4.2 Survey Findings

Fig. 6. Privacy Concerns from Survery

Considering the small number of participants, we cannot draw big conclusions from the responses. Although there were diverse responses to some questions, we did observe some highlights in the responses. Most of our participants have a general concept of public datasets and data visualization. While some also have experience with visualizing themselves, some did not and never used it in their work. The public datasets they had used before were from the US Census and GitHub.

With the crime report dataset, although most participants were not aware of the dataset, they did admit it would be useful, and they might use it in the future. They did show clear concern about whether the data was of themselves or someone they knew. All participants were majorly concerned if they were found to be surrounded by murders or thefts. The participants agreed that the map view of the data made it easier to access and visualize, and they also agreed that the tabular format of the dataset was difficult to access. The participants were most likely concerned with the data rather than the dataset's private information. A similar trend followed with the real estate dataset even though they were not aware of the dataset, they found it useful and, at the same time, were concerned about personal information. They found the data useful for predicting real estate trends. Unitedly, they were concerned about the Owner's names and the location. Once they viewed the search menu format of the data, they found it more accessible and dangerous for specific access to information.

The participants liked the Open Data Initiative and found it beneficial. However, they did show concern about the data containing vulnerable information. When asked how Open Data Initiative can better preserve personal information, some participants mentioned they should not provide private information and that making it a mile-diameter metric would be better than a precise location in the data. Irrespective of the dataset, there was a major concern about the names and location information leaks in the datasets shown in Fig. 6.

Surprisingly, despite the concerns, most participants wanted the data to be publicly available, as shown in Fig. 7; it seems that the benefits trump the risks for participants in the two cases.

5 LIMITATION

Our study acknowledges some limitations that might have impacted the generalization of our findings. Our interviews were comprised of seven participants with limited diversity in background, primarily featuring professionals or students

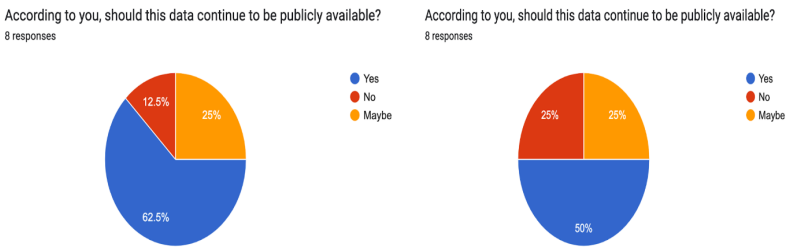

Fig. 7. Survey Opinions on Shown Datasets

who all had some prior knowledge or experience about public data and visualization and most of them were well-versed in privacy issues. Additionally, the survey method seemed inappropriate as our study was based on privacy perception. It was harder to design the study to draw significant insights from the survey as open-ended questions from the survey did not provide enough data, whereas, in interviews, the participants shared their perceptions of the visualization based on their experiences and communicated more elaborately about the reasons for their perspectives. Furthermore, the small sample size of seven interviews and seven survey participants presented challenges in drawing robust conclusions. These limitations highlight the necessity for a more diverse participant pool and a reconsideration of research methodologies to ensure the study's comprehensiveness and validity.

6 CONCLUSION AND FUTURE WORK

Overall, our research investigation shows that the privacy concern around public data visualization is undeniable. The concerns portray how they were uncomfortable with how the visualization made it so easy to access anyone's information, while it did not seem so apparent in raw data. Although our findings indicate a general acceptance of public data remaining public, there is a strong preference for critical private information, such as exact addresses and names, to be anonymized in visualizations. Furthermore, participants highlighted concerns regarding accessibility features of visualizations, emphasizing the importance of preventing reverse engineering through location, which could lead to the identity of persons. These findings highlight the difficulty of balancing public data accessibility and privacy protection in the context of public data visualization.

Moving forward, there is a need for more in-depth investigations of privacy perception on open data visualization with a more diversified population to draw out significant and robust insights. For our future work, we want to conduct interviews with larger sample sizes from different backgrounds, education levels, cultures, etc. Furthermore, the impact of different visualization forms on individuals' perception of privacy remains uncertain. In our study, we employed simple and common visualization techniques, primarily map-based. However, the influence of more complex visualizations on privacy perception is still unknown and warrants further exploration. Another intriguing avenue for future research might be the extensive exploration of standardizing the visualization formats of public data. Public data, distinct from general data, is widely utilized by the populace to support decision-making processes. Therefore, reasonable requirements for the visualization of public data stored on public platforms might include stricter authenticity standards and constraints on aesthetic and personalization aspects. This pertains not only to privacy concerns but also to how visualization affects emotional expression and influences public opinions. While there has been some research in this area, more concerted efforts should be specifically focused on public data.

ACKNOWLEDGMENTS

We would like to thank our classmates and Professor Kate Isaacs for their brilliant guidance and support throughout the course CS 6540: Human-Computer Interaction(HCI). Specifically, we would like to thank Professor Kate and Professor Sameer for helping us with the project idea and guiding us through.

REFERENCES

[1] 1966. Freedom of Information Act, 5 U.S.C. §552.

[2] 2001. Information Quality Act, Public Law 106-554, 114 Stat. 2763, §515.

[3] Erik W Anderson, Kristin C Potter, Laura E Matzen, Jason F Shepherd, Gilbert A Preston, and Cláudio T Silva. 2011. A user study of visualization effectiveness using EEG and cognitive load. In *Computer graphics forum*, Vol. 30. Wiley Online Library, 791–800.

[4] Bahareh Ansari, Mehdi Barati, and Erika G Martin. 2022. Enhancing the usability and usefulness of open government data: A comprehensive review of the state of open government data visualization research. *Government Information Quarterly* 39, 1 (2022), 101657.

[5] Oshrat Ayalon and Eran Toch. 2019. Evaluating {Users'} Perceptions about a {System's} Privacy: Differentiating Social and Institutional Aspects. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. 41–59.

[6] Michael Correll. 2019. Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[7] Simon Dennis, Paul Garrett, Hyungwook Yim, Jihun Hamm, Adam F Osth, Vishnu Sreekumar, and Ben Stone. 2019. Privacy versus open science. *Behavior research methods* 51 (2019), 1839–1848.

[8] Eli Holder and Cindy Xiong Bearfield. 2023. Polarizing political polls: How visualization design choices can shape public opinion and increase political polarization. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[9] Sowmya Karunakaran, Kurt Thomas, Elie Bursztein, and Oxana Comanescu. 2018. Data breaches: User comprehension, expectations, and concerns with handling exposed data. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. 217–234.

[10] Søren Knudsen, Jo Vermeulen, Doris Kosminsky, Jagoda Walny, Mieka West, Christian Frisson, Bon Adriel Aseniero, Lindsay MacDonald Vermeulen, Charles Perin, Lien Quach, et al. 2018. Democratizing open energy data for public discourse using visualization. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–4.

[11] Anna Lenhart. 2022. When to Collect Sensitive Category Data? Public Sector Considerations For Balancing Privacy and Freedom from Discrimination in Automated Decision Systems. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. 98–101.

[12] KC Maas and Josh Levs. 2012. Newspaper Sparks Outrage for Publishing Names, Addresses of Gun Permit Holders. CNN, Cable News Network. www.cnn.com/2012/12/25/us/new-york-gun-permit-map/index.html

[13] Evan M Peck, Sofia E Ayuso, and Omar El-Etr. 2019. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[14] Karen P Tang, Jason I Hong, and Daniel P Siewiorek. 2011. Understanding how visual representations of location feeds affect end-user privacy concerns. In *Proceedings of the 13th international conference on Ubiquitous computing*. 207–216.