

# Privacy-Preserving Data Visualization: Reflections on the State of the Art and Research Opportunities

Kaustav Bhattacharjee<sup>1</sup> , Min Chen<sup>2</sup>  and Aritra Dasgupta<sup>1</sup> 

<sup>1</sup>Department of Informatics, New Jersey Institute of Technology, USA

<sup>2</sup>Department of Engineering Science, University of Oxford, Oxford, UK

## Abstract

*Preservation of data privacy and protection of sensitive information from potential adversaries constitute a key socio-technical challenge in the modern era of ubiquitous digital transformation. Addressing this challenge needs analysis of multiple factors: algorithmic choices for balancing privacy and loss of utility, potential attack scenarios that can be undertaken by adversaries, implications for data owners, data subjects, and data sharing policies, and access control mechanisms that need to be built into interactive data interfaces. Visualization has a key role to play as part of the solution space, both as a medium of privacy-aware information communication and also as a tool for understanding the link between privacy parameters and data sharing policies. The field of privacy-preserving data visualization has witnessed progress along many of these dimensions. In this state-of-the-art report, our goal is to provide a systematic analysis of the approaches, methods, and techniques used for handling data privacy in visualization. We also reflect on the road-map ahead by analyzing the gaps and research opportunities for solving some of the pressing socio-technical challenges involving data privacy with the help of visualization.*

## 1. Introduction

Privacy preservation has become an antithesis to the idea of a digital data-driven era. Be it the smart devices that we use, the online services we access, or even the places we visit, data about our activities, identity, habits and preferences, are being collected at an unprecedented rate. Privacy, a fundamental human right, is often considered a collateral damage in a bid to personalize and monetize commercial services offered to people. Several researchers have recently posited that the data landscape is confronted with a privacy crisis [OR19, BPL19, Var19], and to fix it, immediate collaborative effort among multiple stakeholders in the data ecosystem is needed.

Who are these stakeholders? In the related research areas of privacy-preserving data publishing [FWCY10] and mining [BLNR07], there has been extensive discussion on the role of different stakeholders. In Figure 1, we illustrate the same in the context of privacy-preserving data visualization. The stakeholders with the highest responsibility in this ecosystem are the data owners, who collect and have proprietary rights over the collected data, and the data custodians, who have the responsibility of enforcing policies and safeguarding the privacy of the data. Cambridge Analytica's much-debated and questionable use of Facebook data [CGH18] demonstrate how privacy preservation responsibilities can be misused. Data subjects are the individuals (e.g., people on Facebook) who provide implicit or explicit consent to different agencies for collecting their personal data. They need to be cognizant of the risks of sharing personal data and understand the

privacy policies of companies, a task that is often complex and inconvenient. In fact, recent studies have demonstrated the lack of effectiveness of privacy policies of online companies [OOH20], and even worse, the deliberate use of dark patterns for subverting policy implementations [MAF\*19]. Data consumers are analysts or the general public with appropriate levels of access to sanitized data who want to derive insights without violating privacy. In many cases data subjects themselves are consumers (e.g., patients mining electronic health records, people trying to understand trends in survey data). Attackers are people or enterprises with malicious intent, who are always attempting to breach private databases or attack privacy-preservation mechanisms duly enforced in openly available data. While regulations such as HIPAA [A\*03], or more recently, GDPR [NLV\*20] aim to protect data subjects against such malicious attacks by enforcing strict regulations for releasing data, recent studies have demonstrated how even heavily anonymized datasets run the risk of privacy breach, where demographic attributes in openly available data can be used to re-identify about 99% of Americans [RHDM19]. The latter case study is a telling commentary on how static privacy-preservation mechanisms (where anonymized data is released without any subsequent checks of risks) are inadequate in the face of evolving threats and attack scenarios.

Given this rather bleak picture of privacy in the real world, our attempt in this state-of-the-art report is to: a) investigate if and how visualization can empower data owners, subjects, custodians, and consumers to have a transparent understanding of privacy impli-

Stakeholder	Role in the data ecosystem	Stake for privacy
Data owner	An entity which owns data about people or individuals whose data is captured. Examples: hospitals, social media companies, social media users	<ul style="list-style-type: none"> <li>Wants to understand risks of releasing data for public consumption</li> <li>Implement privacy legislation in the form of policies</li> </ul>
Data subject	Individuals whose data are represented in databases or are collected by applications. Examples: patients, common public	<ul style="list-style-type: none"> <li>Decide whether to trust an agency for collecting their data</li> <li>Understand implications of privacy policies</li> </ul>
Data custodian	An entity with credentials for accessing a private database or a 3rd party entrusted with data analysis. Examples: Cambridge Analytica	<ul style="list-style-type: none"> <li>Have access to the original or a limited version of the data</li> <li>Implement privacy legislation in the form of policies</li> </ul>
Data consumer	Any person who is the intended audience for shared data or analysis. Examples: data analysts, scientists, policy makers, and the public	<ul style="list-style-type: none"> <li>Access anonymized data</li> <li>Derive value from data without getting to know sensitive information</li> </ul>
Attacker	Anyone with the goal of breaching privacy and knowing about people. Examples: any attacker with or without background knowledge about the collection	<ul style="list-style-type: none"> <li>Link publicly and privately available information with the intent of privacy breach</li> <li>May or may not have background knowledge about individuals in a database</li> </ul>

**Figure 1: Different stakeholders in the data privacy ecosystem.** Data owners and custodians need to preserve and protect the privacy of data subjects (i.e., individuals represented in a dataset) from insider or outside attackers. Privacy-preserving visualization is used by data owners or custodians for understanding privacy-utility trade-offs and is also used by data subjects, who want to understand privacy policies, and data consumers, who want to derive value from anonymized data.

cations and b) provide guidance on how visualization can play a significant role towards addressing the socio-technical dimensions of data privacy. In the process, we analyze how a futuristic research agenda can adapt to the needs of the different stakeholders. As illustrated in Figure 1, people's roles define what kind of stake or incentives they have for preserving or breaching data privacy. For example, a biologist who runs a research lab or a company which collects data about people's social media interests, would want to get guidance on the risks of sharing data with a broader group of people. A data custodian, like Cambridge Analytica, needs to have checks and balances in place to ensure people's identities are not revealed due to the use of demographic data. Data consumers, like a social scientist trying to understand the correlation between demographics and economic indicators of a region, need to derive value out of anonymized data and overcome the potential loss of value due to suppression or omission of sensitive information. With the ubiquitous availability of smartphones, data subjects are often at the receiving end of privacy violation as personal data is being collected at an unprecedented rate, often with dubious policies and purposes. In rare cases like the currently unfolding *COVID-19* pandemic, such data collection becomes a societal need for contact tracing [HAG\*20], which also brings privacy risks in its wake and solutions [RBS20] need to be developed where public health and individual privacy are not considered to be trade-offs in policy implementations.

Visualization can play a critical role in all these scenarios, as evidenced by the state-of-the-art literature on privacy-preserving data visualization. This field of research has imbibed and extended concepts from the privacy-preserving data publishing [LL09] and mining [BLJ08] communities for developing visualization-specific solutions for anonymization, controlled access, and utility and risk analysis of released datasets. Our goal in this survey is to take

a problem and task-driven approach towards organizing the existing research. This approach is motivated by the fact that privacy is as much a computational challenge as it is a challenge related to consideration of human factors across domains like healthcare [DDK16] and social networks [JEB12, MLA12].

To study these factors, we introspect about the privacy problem and the related goals of stakeholders and then map those back to the anonymization methods and visualization techniques. Our survey makes three specific contributions: i) Task-driven understanding of the privacy preservation goals with regards to different application scenarios and multiple stakeholders in the data ecosystem, like the data owners, data custodians, and data consumers (Sections 2, 3, and 4), ii) Comparison of tasks and techniques for privacy-preserving data visualization and a critique of the design space (Sections 5 and 6), and iii) Analysis of gaps and emerging research opportunities by establishing the context of privacy-preservation related challenges in the realms of both privacy-related research gaps and emerging research areas in visualization and visual analytics (Section 7).

## 2. Background on Privacy Preservation

In this section, we provide a background about the basic concepts in the literature on privacy preservation, mainly relying on the vocabulary used in the fields of privacy-preserving data publishing and data mining [FWCY10, GDLS14] from where the field of privacy-preserving data visualization draws its inspiration.

**Re-identification via linking:** When releasing data, merely suppressing personally identifiable information (PII), like name, social security number, email address, etc., is necessary yet not sufficient. *Quasi-identifiers* [MX07], like, age, gender, zip code, etc., can be exploited by attackers for breaching privacy by linking attributes from publicly available data sources (e.g., voter registration data) and privately accessible information (e.g., hospital data or web access data). This is popularly known as the *data linking* problem [Swe05], and various data anonymization methods [GDLS14] like generalization, suppression, perturbation, clustering, etc. are used to tackle this problem. These methods typically produce an anonymized static data table, a modified data mining algorithm or an anonymized visualization. Most of these methods constitute the non-interactive setting of privacy-preservation, where, once released, the data owner does not have any control over the data or the mining results, and the drawbacks of such a “release-and-forget” model [RHDM19] have been questioned by recent studies.

**Anonymization methods:** One of the most widely used anonymization methods is the  $k$ -anonymity model. It states that a dataset is  $k$ -anonymous if the information for each record in the dataset cannot be distinguished from at least  $k - 1$  other records [Swe02, BA05]. For example, if  $k = 3$ , then a  $k$ -anonymized dataset will have at least 3 similar combinations for each record of potentially identifying variables. But  $k$ -anonymity does not provide guarantee against attackers having background knowledge or homogeneous attacks.

Let us refer to a dataset as shown in Figure 2.

Table 1 represents a dataset from clinical records and Table 2

Company	Position	Nationality	Zip	Age	Disease
Alpha	Director	Japanese	10001	32	Galactosemia
Beta	Manager	Indian	11049	53	Cancer
Gamma	Associate	American	10011	38	Galactosemia
Beta	Manager	Russian	10004	43	Fatty Liver
Alpha	Manager	Japanese	10014	48	Hepatitis B
Delta	Consultant	Indian	10017	34	Galactosemia
Gamma	Associate	American	11042	57	Hepatitis B
Delta	Manager	American	10007	42	Hepatitis B
Gamma	Director	Japanese	11043	51	Galactosemia
Beta	Manager	Russian	10009	35	Galactosemia
Delta	Associate	Indian	10019	42	Fatty Liver
Gamma	Manager	Japanese	11047	63	Fatty Liver

Table 1: Original dataset

Company	Position	Nationality	Zip	Age	Disease
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	100**	<40	Galactosemia
*	*	*	110**	>=50	Galactosemia
*	*	*	110**	>=50	Cancer
*	*	*	110**	>=50	Hepatitis B
*	*	*	110**	>=50	Fatty Liver
*	*	*	100**	4*	Hepatitis B
*	*	*	100**	4*	Fatty Liver
*	*	*	100**	4*	Fatty Liver
*	*	*	100**	4*	Hepatitis B

Table 2: k-anonymous dataset (k=4)

Company	Position	Nationality	Zip	Age	Disease
*	*	*	1000*	<50	Galactosemia
*	*	*	1000*	<50	Fatty Liver
*	*	*	1000*	<50	Hepatitis B
*	*	*	1000*	<50	Galactosemia
*	*	*	1104*	>=50	Hepatitis B
*	*	*	1104*	>=50	Galactosemia
*	*	*	1104*	>=50	Fatty Liver
*	*	*	1104*	>=50	Cancer
*	*	*	1001*	<50	Galactosemia
*	*	*	1001*	<50	Hepatitis B
*	*	*	1001*	<50	Galactosemia
*	*	*	1001*	<50	Fatty Liver

Table 3: l-diverse dataset (l=3)

**Figure 2: Examples of data anonymization based on the k-anonymity and l-diversity metrics.** k-anonymity ensures sufficient group size (here  $k=4$ ) so that an individual cannot be distinguished within that group and l-diversity ensures sufficient diversity in the values of an attribute (here,  $l=3$ ), so that the exact values of a sensitive attribute cannot be detected.

is the 4-anonymised version of the same. If we know that John is an American associate of age 38 living in the zip code 10011, we can easily decipher from the Table 2 that he has Galactosemia. This is the problem of homogeneous attacks. Again, if we know that Kabir is a 42-year-old Indian associate who lives in zip code 10019 and works for the company Delta, we can easily say that he has either Hepatitis B or Fatty Liver. But if we have background knowledge (e.g., associates of the company Delta have been immunized against Hepatitis B) we can infer that Kabir has Fatty Liver. Thus, these types of attacks cannot be prevented even if the dataset is  $k$ -anonymized.

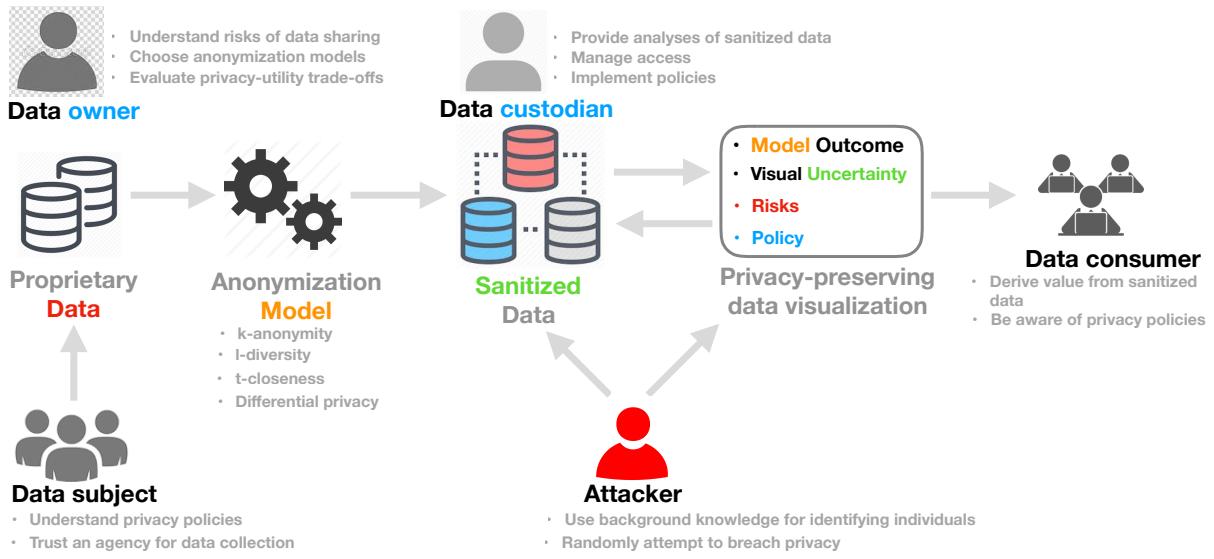
This problem is addressed by another anonymization method, the  $l$ -diversity model [MGKV06], which guarantees sufficient diversity in the value of attributes. The data from Table 1 can be represented in a 3-diverse way in Table 3 (Figure 2). Here each block of 4 records have minimum of 3 varieties of disease each. Now even if we know that John is an American associate of age 38 living in the zip code 10011, we can only decipher that he either Galactosemia or Fatty Liver or Hepatitis B. Also if we know that Kabir is a 42-year-old Indian associate who lives in zip code 10019 and works for the company Delta, and we have the background knowledge that the associates of the company Delta have been immunized against Hepatitis B, we cannot tell with guarantee that he has Galactosemia or Fatty Liver. Hence both the problems of  $k$ -anonymity can be avoided through  $l$ -diversity.

But  $l$ -diversity has its own limitations.  $l$ -diversity may be difficult and unnecessary to achieve. For example, let's assume our data in Table 1 contains only one sensitive attribute, i.e., whether the person has a disease or not (Yes/No), with around 100,000 records and 98% of them have a disease (Yes) and only 2% of them do not have any disease (No). In order to have a 2-diverse table, there can be at most 2000 equivalence classes. Moreover  $l$ -diversity is insufficient to prevent attribute disclosure. In our previous example, suppose an equivalence class has 49 negative records and 1 positive record. This means any individual in this class will have 98% possibility of not having a disease, instead of the overall 2% in the whole dataset. This is called skewness attack. Moreover,  $l$ -diversity is also not immune to similarity attack. For example, in Table 3 (Figure 2),

if someone belongs to the last equivalence class and knows that Galactosemia, Hepatitis B, and Fatty Liver are diseases related to liver, then we can easily decipher that any individual belonging to that equivalence class has a liver disease.

The above scenarios can be alleviated using the  $t$ -closeness [LLV07], which measures the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the whole table and guarantees that the distance is at most  $t$ . An even more robust and popular anonymization concept is that of differential privacy [DR\*14, Lee17]. Differential privacy guarantees the following: a) anyone analyzing the results of a differentially private analysis will make the same inference about an individual's private information, irrespective of the fact whether the individual's private data was used in the analysis or not [NSW\*17] and b) privacy protection against a gamut of privacy attacks, including linkage attacks, reconstruction attacks, and differencing attacks [DR\*14].

**Evaluation:** A perfectly private dataset is one which is stripped of all sensitive attributes and quasi-identifiers. But that dataset will be rendered useless for most practical data analysis purposes, thereby reducing its analytical value. Datasets containing attributes qualifying people's behavior and characteristics may help to understand the cause of diseases, economic patterns of different states, food trends popular in a city etc. Thus, the private data is transformed using anonymization techniques and then published to the world. This fact illustrates the need to consider the trade-off between privacy and utility, which is an abiding and pervading research problem across domains affected by privacy breaches. Brickell et al. [BS08] compared the privacy loss caused by data anonymization to the utility gained by the same method. They concluded that utility of an anonymized dataset degrades rapidly even with modest privacy gains. To address this problem, Li et al. [LL09] introduced a methodology similar to the risk-return tradeoff in financial investment, which led to an objective way to ensure a better balance between privacy gain and utility loss. The study of trade-offs is very much an area of ongoing research investigation, where researchers have also studied this outside of the computational realm, from a human-centered perspective [VZ19]. Researchers have also



**Figure 3: Data flow and roles of privacy stakeholders.** Privacy-Preserving Data Visualization involves visual representation of outcomes of different anonymization models, addition of visual uncertainty as defense mechanism, evaluation of disclosure risks, and visualization of policy implications. The abiding goal in all of these cases is to guarantee a minimum level of privacy that can protect the data with respect to attack scenarios.

pointed out to the need for more robust privacy evaluation, especially related to how gaining of sensitive knowledge can be protected [GDL14].

**Human Factors and the role of visualization:** As apparent from the above discussion there are several human factors involved with all stages of privacy-preservation of data, be it the choice of anonymization methods, evaluation of trade-offs or the various attack scenarios, often triggered by attackers' background knowledge [MKM\*07, DTZ08]. This is illustrated in the privacy-preserving data visualization pipeline showed in Figure 3. Data owners often need to control access to proprietary data and protect it from even insiders in a company, and therefore visualization can help them understand the risks [RJD\*06, MFG\*06] and more transparently configure appropriate levels of anonymization and data accessibility. Disclosure risk minimization [DL86, Lam93] is a key goal for both data owners and data custodians, particularly when the released data or the results of the analysis process can be mined by outside adversaries by using their background knowledge. Visualization can help understand the privacy guarantees and risk-utility trade-offs. For data consumers, a better understanding of mental models of personal privacy [KDFK15, OGH05] can let us know what kind of human inputs and interaction mechanisms should be considered for developing visualization interfaces. In our survey, we aim to understand whether the state of the art in privacy-preserving data visualization addresses these known unknowns and if so, what are the emerging trends, patterns, and gaps thereof.

### 3. Methodology

In this section, we describe our survey methodology. Specifically, we discuss the definition of privacy that is relevant for visualization and describe our analysis workflow.

#### 3.1. Definition and Scope for Literature Search

The field of privacy-preserving data visualization lacks a thorough characterization of the human-specific needs and goals. Depending on whether the target user is a data owner, data subject, or a data consumer, the uses of visualization are likely to be vastly different (Figure 3). We look at the relevance of visualization in privacy from the dual lens of input and output privacy [RH02, WL08, BLNR07], where input privacy involves transformation of a dataset into its privacy-preserving form through anonymization methods, and output privacy involves judgment about the analysis outcomes of the privacy-preserving dataset: whether the analysis or the visualization is also privacy-preserving, i.e., how difficult it is for an attacker to infer sensitive knowledge by observing the patterns.

Since privacy-preserving data visualization is a relatively newer research area as compared to other areas of visualization research, we wanted to collect papers which reflect both the theoretical and practical aspects of visualization usage in the context of privacy. To that end, we followed a three-stage process for paper collection. In the first stage, we performed a broad search on IEEE and ACM digital libraries and Google Scholar with various combinations of keywords such as “privacy and visualization”, “privacy-preserving visualization”, “privacy and visual”, “privacy and human factors”, etc. This phase gave us a data-driven idea of the domains in which we were most likely to find privacy-preservation techniques and strategies involving data visualization. The healthcare domain was the most frequent one we encountered through our initial exploration with the domain of social science being a distant second.

In the second stage, we performed a deeper search into top-ranked domain-specific journals from healthcare, such as, the *Journal of Biomedical Research*, and social science, such as the *Social Science Journal*. From them, we collected more than hundred papers by repeating the search terms of “privacy and visualiza-

tion". We also looked into the Google Scholar citations of these papers. Our inclusion criterion was that we can consider any paper that proposes a visualization method or technique as part of their privacy-preservation theory and applications. Most of the social science papers did not satisfy this criterion and had to be excluded from our collection. For papers published in visualization-specific venues, we collected research papers related to privacy-preserving data visualization, by focusing our search on leading visualization publications from the past twenty years. These included proceedings of the Information Visualization Symposium/Conference, and journals such as *IEEE Transactions of Visualization and Computer Graphics (TVCG)*, *Computer Graphics Forum*, *ACM CHI Conference*, and *IEEE PacificVis Symposium*.

In the third stage of our paper search process, we considered publication venues such as *ACM CHI* and *ACM SOUPS*, from where we collected several papers related to visualization and privacy which were specific to the security domain or were domain agnostic. We applied the same inclusion criterion for these papers.

Before applying our inclusion criterion, our corpus comprised about 400 papers. We carefully checked our corpus even after applying the inclusion criterion and filtered out any paper which only reflected on a *potential* use of visualization or reflected on a *potential* breach of privacy in a dataset, without discussing any specific method or technique. We finally ended up with 38 papers with contributions in the domain of visualization and in the specific application domain (e.g., healthcare, social science and security and privacy). The latter collection helped us take a user-centered approach which was our goal from the onset. We look at the survey as a three stage process: *a problem characterization phase*, where we reflected on the target user and the privacy-specific goals, and *a design classification and analysis phase*, where we looked at the anonymization methods, visualization techniques, and *a gap analysis phase* where we reflected on the gaps and research directions in the context of the privacy problems and the state of the art in visualization research.

**Critical Reflection and Gap Analysis:** In this phase, we first critically analyze the visual encoding choices for privacy-preservation purposes, with respect to the literature on the established encoding principles [Ber11, CM84]. This helps understand what kind of transformations are necessary to basic visualization techniques for addressing the privacy goals, also also how the low-level visualization tasks need to adapt for fulfilling those goals. Next, we developed an understanding of what research gaps exist in the state of the art and how research directions to address those gaps can have a practical impact in different application domains. We grouped our findings according to six research themes which highlight those gaps and directions.

#### 4. Problem Characterization

We derived a classification scheme (Figure 4) to characterize the different research contributions in the literature. We look at the problem of privacy-preservation from an end user's perspective and focus on whether the techniques, methods, or applications are designed for a data owner, data consumer or a data subject. Due to the inherent similarity of the roles of data owners and data custodians from the perspective of privacy-preservation and also in the

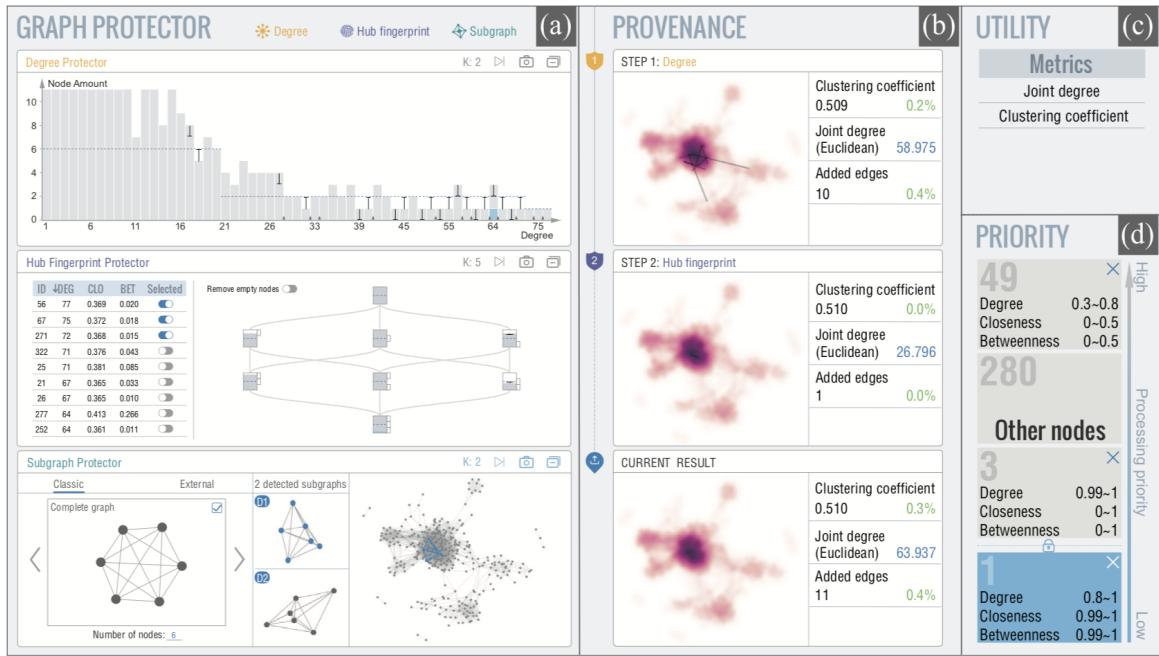
context of the work we surveyed, we treat them as one group of users. Data owners, who hold proprietary rights for the collected data (e.g., social media companies or hospitals) aim to anonymize the data, implement access control, implementing accountability in order to increase the levels of privacy preservation. On the other hand, data consumers (e.g., analysts using social media data, scientists using health-care data for research, laypeople using data from fitness trackers) are generally provided with an anonymized version of the data or the visualization for deriving value out of it. In our collection, we found there is an even split between the techniques that consider these groups as their target users. Data owners must be cognizant of the risks owing to **identity disclosure** (i.e., data consumers knowing exactly who the individuals are, from the data points representing them) and **attribute disclosure** (i.e., data consumers knowing the value of different quasi-identifiers or sensitive attributes) risk scenarios. They also have to understand what kind of **attack scenarios** a released data or a visualization may be subjected based on the availability of other data sources or the background knowledge of the attacker. Visualization systems themselves can be subject to attack and thereby the privacy guarantees might be compromised [CAS05]. When a person with a data owner's role in a company needs to share data internally, they also have to implement appropriate **access control** mechanisms: people with only certain roles and privileges can access de-anonymized versions of the data. Data consumers have to overcome the barriers of anonymization to derive value out of the data. When a consumer is subject to data collection (e.g., whenever we use services on our smartphones), they also need to be cognizant of the disclosure risks associated with sharing their information. One of the most important challenges in information privacy is the trade-off between privacy and value or utility of the data. We observed that while there is a systematic approach towards defining what privacy means and how anonymization methods can help achieve different levels of privacy, in comparison, there is a lack of consensus about how utility of an anonymized data or a visualization derived from it can be qualified or quantified. The trade-off between privacy and utility affects both data owners and the consumers. Based on the choice of anonymization methods like  $k$ -anonymity,  $l$ -diversity and  $t$ -closeness (as discussed in Section 2), the degree of reduced utility of the data will vary.

The privacy problems faced by data owners [RKW18, Yee06, KHC<sup>\*</sup>17, XLZ<sup>\*</sup>18] can be described as follows based on our collection:

- How to choose anonymization methods that minimize disclosure risks and maximize the utility of the shared data (Figure 5)?
- How to develop a privacy-preserving interface or visualization which will help users leverage interactive capabilities without leaking information about sensitive attributes?
- What are the vulnerabilities faced during the flow of the data between organizations which may result in policy non-compliance?
- How to share data between different entities (sensor, people etc.) without privacy leakage?
- What are the degrees of re-identification risks, based on external information or users' background knowledge, once the data or the visualization is publicly accessible?
- Can attack scenarios be predicted and accordingly, how can de-

Papers	Problem Characterization				Privacy-Preserving Data Visualization											
	Target User	Privacy Problems			Privacy Tasks				Anonymization Method		Visualization Technique					
		Data Owner	Data Consumer	Data Subject	Identity Disc.	Attribute Disc.	Attack Scenarios	Access Control	Hide Data	Evaluate Risk	Evaluate Trade-Offs	Compare Algorithms	Policy understanding	Data Uncertainty	Visual Uncertainty	
Andrienko2016 (AAFJ16)		■			■								Aggregation	Precision	Geographical map	
Chou2016 (CY16)		■											Clustering		Custom visualization	
Chou2017 (CBM17)	■				■				■				Masking	Deletion, Bundling	Adjacency Matrix, Node-link diag.	
Chou2019 (CWM19)	■				■	■	■	■	■				Precision		Sankey Diagram	
Dasgupta2011a (DK11a)	■				■				■					Granularity		Parallel coordinates
Dasgupta2011b (DK11b)	■				■				■					Granularity		Parallel coordinates
Dasgupta2013 (DCK13)	■				■	■	■		■	■				Precision		Parallel coordinates, Scatterplots
Dasgupta2014 (DMARC14)	■				■	■	■	■	■				Binning, aggregation		Bar chart, Tree map	
Dasgupta2019 (DKC19)	■					■			■							Parallel coordinates, Scatterplots
Kao2017 (KHC*17)	■				■	■	■	■		■			Clustering		Heat map	
Liccardi2016 (LARC16)	■				■				■				Aggregation		Geographical map	
Oksanen2015 (OBSW15)	■				■				■				Kernel Density Estimation		Heat map	
Ragan2018 (RKIW18)	■				■				■				Masking		Custom visualization	
Wang2017 (WCC*17)	■				■				■	■			Masking		Matrix, Tree	
Wang2018a (WCC*18)	■				■	■			■	■			Clustering		Graph	
Wang2018b (WGL18)	■				■	■			■	■			Aggregation		Custom visualization, heat map	
Gotz2016 (GB16)	■				■				■				Clustering		Flow based visualization	
Ljubic2019 (LGG*19)	■				■				■				Clustering		Geographical heat map	
Muchagata2019 (MVMF19)	■	■	■		■				■	■			Suppression		Text-based interface	
Bahrini2019 (BWM*19)	■	■							■					Jamming, Occlusion		Custom visualization(app), Error bars
Conti2005 (CAS05)	■					■			■	■				None		
Deeb2019 (DSEB19)	■				■	■			■	■			Merging		Link charts	
Elagroudy2019 (EKM*19)	■	■			■				■				Obfuscation, Deletion		Images	
Kum2019 (KRI*19)	■				■				■	■			Masking		Custom visualization, Violin plots	
Mazzia2012 (MLA12)		■				■										Custom visualization
Takano2014 (TOT*14)		■			■	■			■							Custom visualization
Wang2015 (WGXI5)		■			■	■			■	■			Obfuscation		Custom visualization, bar graphs	
Anwar2009 (AFYH09)		■			■				■	■			Precision		Social graphs	
Gao2013 (GB13)		■			■				■	■						Hierarchical circles
Becker2014 (BHÖK14)		■			■				■	■			None		Infographics	
Dhotre2017 (DBKO17)		■							■	■			None		Pie chart	
Ghazinour2009 (GMB09)		■							■	■			Granularity		Relationship diagrams	
Yee2006 (Yee06)		■							■	■			None		Data flow diagrams	
Hongde2014 (HSH14)		■			■	■			■	■			Clustering, aggregation		None	
Kung2017 (Kun17)		■			■	■			■	■			Reduction		Multi-dimensional projection	
Osia2020 (OSS*20)		■			■	■			■	■			Reduction		Auto-encoder visualization	
Xiao2018 (XLZ*18)		■			■	■			■	■			None		Parallel coordinates, feature grid	

**Figure 4: Classification Scheme for describing the literature on privacy-preserving data visualization based on target users, privacy problems, visualization tasks intended to solve those problems, and the anonymization method used in conjunction with a visualization technique.**



**Figure 5: A privacy-preserving data visualization pipeline [WCC\*18]** that helps data owners systematically assess how different algorithms can affect the level of privacy in a visualization.

fence mechanisms be integrated within an anonymized visualization?

While some of the above problems also affect data consumers, we can describe the additional privacy problems faced by data subjects and consumers based on the literature [GB13, EKM<sup>\*</sup>19, AFYH09, MLA12, BWM<sup>\*</sup>19, BHÖK14, DBKO17] as follows:

- How to assess the one's privacy on Online Social Networks (OSN)?
- What are the permissions requested by mobile applications and how is the shared information used?
- Does a website sell or misuse private information by stating them explicitly in the privacy policies? Can data consumers be better aware of potential dark patterns [MAF<sup>\*</sup>19]?
- How can data owners and data consumers have better communication through more interpretable privacy policies?

We use this categorization and problem definition to describe the visualization-specific tasks, solutions and challenges addressed in the literature, which we describe in detail in the following sections.

## 5. Privacy-Preserving Data Visualization

Visualization has a key role to play in all aspects of privacy in the data ecosystem for both data owners and data consumers. With our dual focus on visualization-specific contributions and application-specific research involving privacy-preserving data visualization, we are able to cover a breadth of work that can inform both visualization researchers and practitioners. In this section, we describe the surveyed papers based on the following categories (Figure 4): i) the anonymization methods used, ii) the high-level visualization tasks relevant to privacy-preservation, and iii) visualization techniques used to address those tasks.

### 5.1. Anonymization Method

The anonymization methods used in the context of visualization fall broadly into two categories, which are methods based on: i) data uncertainty and ii) visual uncertainty (Figure 6). Introducing uncertainty in the data space involves use of the anonymization methods (Section 2) for making sure either a certain number of records are indistinguishable, and the distribution of attributes is such that sensitive information cannot be derived from them. Besides the traditional metrics of  $k$ -anonymity,  $l$ -diversity,  $t$ -closeness, and differential privacy, we also find examples in the literature where novel metrics are proposed. For example, Okansen et al., using a dataset of users' cycling work-outs [OBSW15], focus on three methods, namely privacy-preserving heat map with user diversity (ppDIV), privacy-preserving kernel density estimation (ppKDE) and privacy-preserving user count calculation (ppUCC). Their goal is to prevent disclosure of user identity. Data-based clustering algorithms [CWM16, CBM17, CWM19] and those based on differential privacy [HSH14] are also used for preventing identity and attribute disclosures.

In visualization, at least some information about the data is typically available, like labels and value range on axes, and the minimum and maximum boundaries of each cluster. The notion of totally 'blind' attack, without any knowledge about the data, may not be applicable to privacy-preserving visualization. To guard against this kind of inference based, researchers had proposed the idea of developing anonymization metrics in the screen-space, as opposed to the data space, based on visual uncertainty. Visual uncertainty [DCK12] entails uncertainty owing to the visual mapping between data points and pixel coordinates. For example, a clustered scatter plot or a parallel coordinates (Figure 6b) that guarantees a minimum level of privacy, can be developed by combining

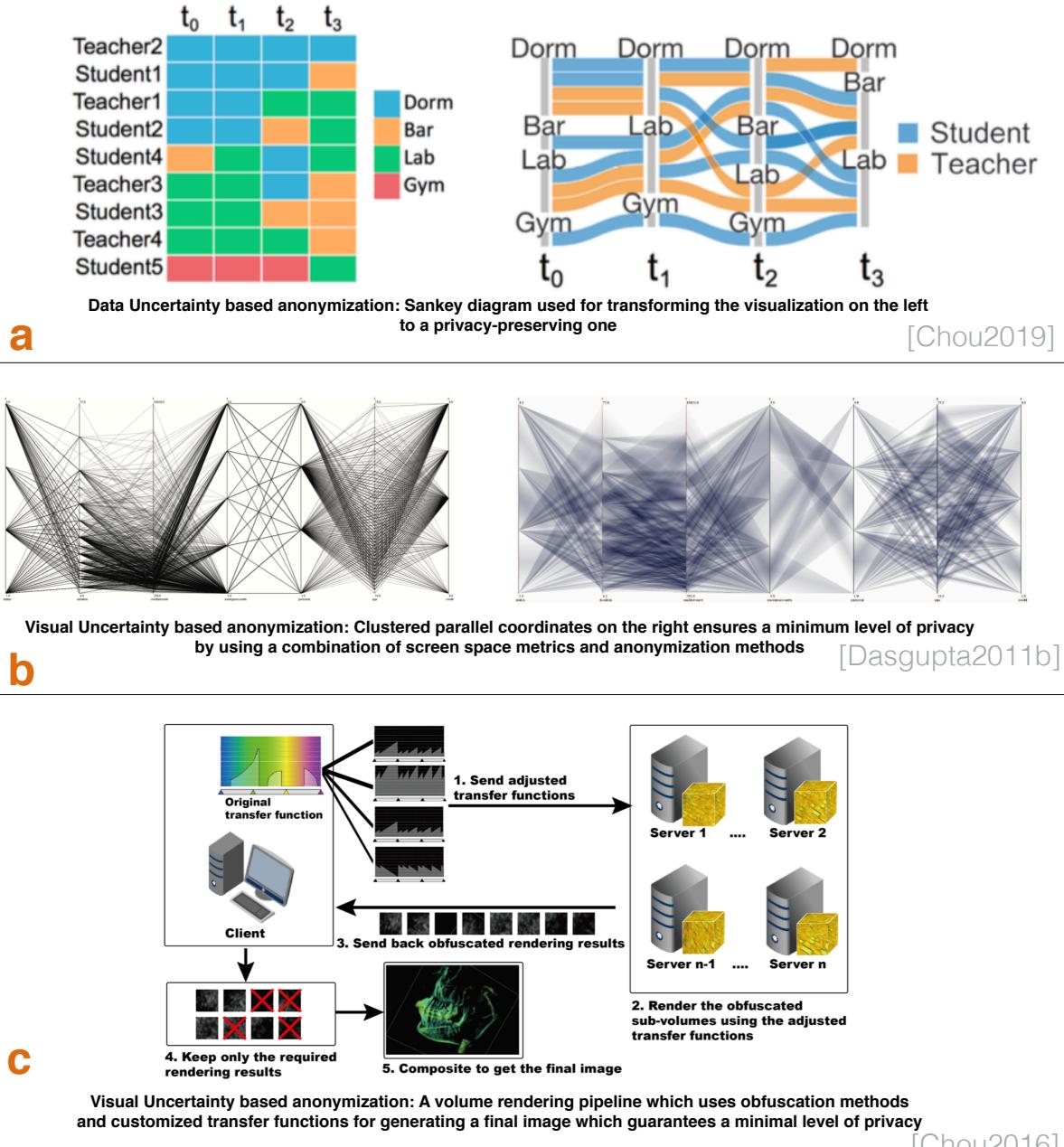


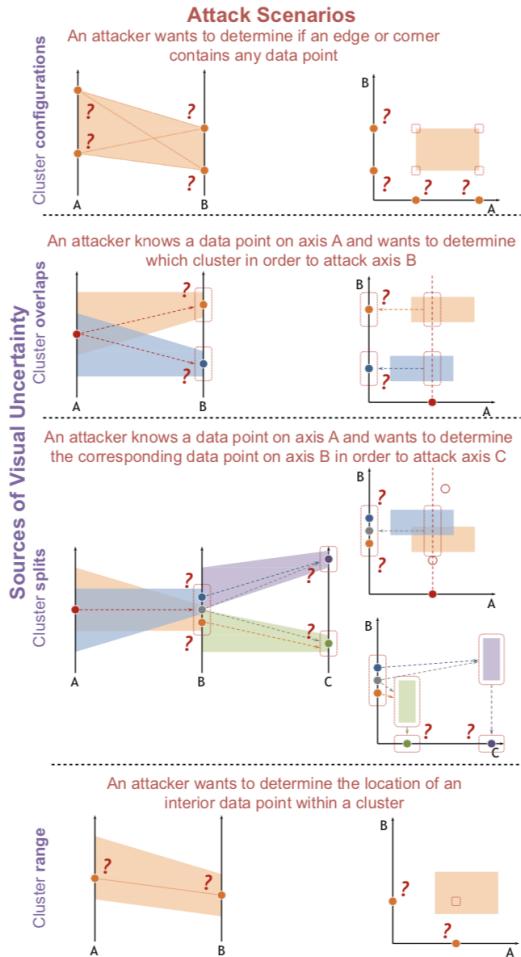
Figure 6: Illustrating anonymization methods based on data uncertainty and visual uncertainty.

pixel binning with the conventional anonymization methods like  $k$ -anonymity or  $l$ -diversity [DK11a, DCK13].

Visual uncertainty has important connotations for how the intended privacy level of a visualization can be breached via different attack scenarios. As shown in Figure 7, the cluster ranges naturally hide record locations within a cluster and cluster overlaps can also hide where a record within a cluster ends up, across the axes, in a parallel coordinates plot. An attack usually consists of a series of progressive actions, building on incrementally acquired knowledge. An attacker may start with little knowledge, and by making observations from the information conveyed in visualization, such

as a clustered parallel coordinates or a scatter plot, the attacker may try to identify a particular record within that cluster.

From that, the attacker gradually identifies more information about the record by moving from one axis to another or works out information about other records in the same cluster, as shown in the illustrations involving cluster overlaps cluster splits, and cluster range in Figure 7. Regardless of how complex an attack is, it can be decomposed into a set of basic attacking actions and disclosure risks. Causes and effects of visual uncertainty (in the form of cluster overlaps, splits and ranges) can protect against disclosure risks and computing the amount of uncertainty [DK11a, DCK13] can also provide an estimate to data owners and custodians of the degree of



**Figure 7: Illustrating [DKC19] how risks can be evaluated in a privacy-preserving data visualization based on a systematic understanding of the different attack scenarios.**

risk involved with different visualization configurations [DKC19]. Other examples of visual uncertainty involve the use of record masking [RKIW18] or obfuscation for volume rendering [CY16].

## 5.2. Visualization Tasks and Techniques

In this section, we describe the privacy-preserving data visualization tasks and techniques that we collected from our survey. Four high-level visualization tasks emerged in our collection and we describe them along with the corresponding visualization techniques.

### 5.2.1. Hide Data

*Hiding data* was the most common in our collection, with a coverage of more than 50% of the papers we surveyed. This task was employed for both spatial data and non-spatial data. In rare cases, we find the use of machine learning models for minimizing the exposure of sensitive information using a cloud-based architecture [OSS\*20]. For scientific data, Chou et al. [CY16] proposed an obfuscation technique for scientific visualizations in order to maintain the privacy of the user. This block-based volume data transformation algorithm obfuscates a volume data and delegates the task

of rendering the volume data to a remote server, thus preserving the privacy of the scientific visualization. The images show the difference between normal rendering and the proposed privacy-aware volume rendering. This paper also demonstrated the development of a transfer function adjustment so that the transfer to the remote server for volume rendering is also privacy preserving.

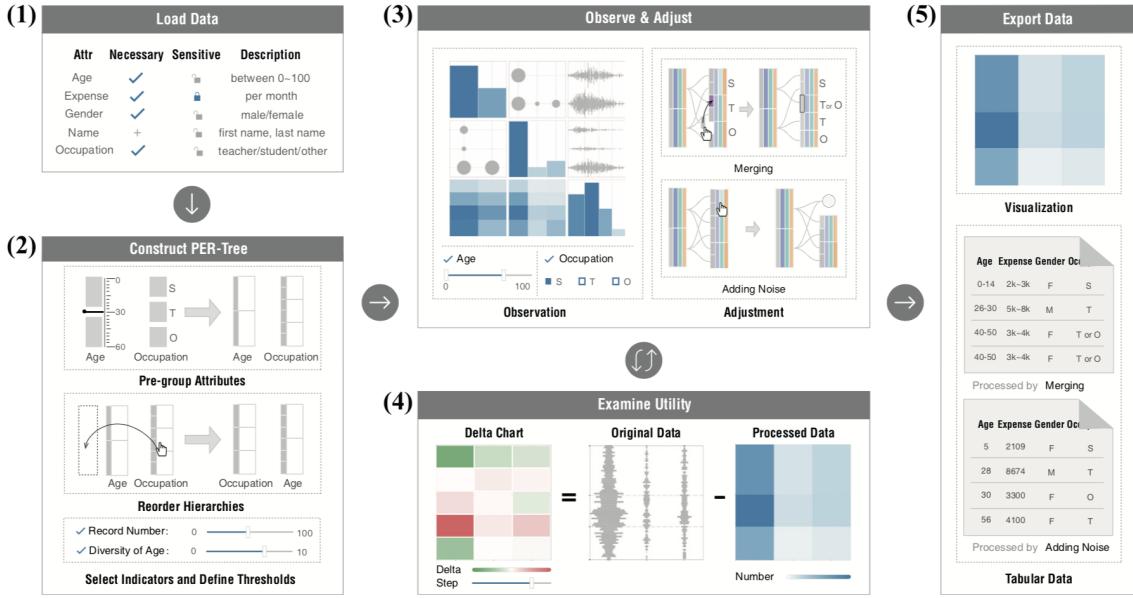
For spatial data, the primary goal is to hide the exact coordinates of people's location [LARC16]. To that end, Andrienko [AAFJ16] presented a visual analytics model which can analyze the episodic digital traces/locations of a person over a long period of time and detect places of significant interest like home, work, social activity place etc. But this model also preserves the privacy of the person being analyzed. Geographical maps are used to represent neighborhoods instead of individual data points. It also uses a semantic map to display the information about different places derived from the data of a certain city. Two-dimensional time histograms are also used to analyze the usage of different location clusters in a certain city over a certain period of time. Ljubic et al. [LGG\*19] uses geographical heatmaps to present the distribution of influenza in a certain area. This helps in finding the affected area in a certain geographical region which may be helpful to healthcare officials. A privacy leakage in these geographical heatmaps may allow the identification of certain patients, leading to identity disclosure.

For temporal data, visualization is often used to encode the outcomes of an anonymization method (e.g.,  $k$ -anonymity,  $l$ -diversity,  $t$ -closeness, differential privacy), leveraging clustering in the data space [CBM17, CWM19, GB16] for visualizing event sequences.

For non-spatial data, visual uncertainty is added to a conventional technique like scatter plot or parallel coordinates as an additional defense mechanism [DK11a, DK11b, DCK13]. Examples of visual uncertainty include loss of precision of a data-point, where an attacker is unable to tell apart lines in a parallel coordinates or dots in a scatter plot due to visual confusion, or the degree of granularity of records in a cluster, where an attacker is not able to exactly point to record locations within a cluster. Understandably, visual uncertainty can reduce risks of both identity and attribute disclosure by manipulating clustering algorithm parameters.

### 5.2.2. Evaluate Risk

*Evaluating risk* was the second most common task in our collection, with a coverage of about 30% of the papers we surveyed, mostly focused on the data owner. Disclosure risks are affected by how much an adversary knows about the data. Two kinds of re-identification scenarios are possible [BLJ08]: a) prosecutor re-identification scenario, where an intruder (e.g., a prosecutor) knows that a particular individual (e.g., a defendant) exists in an anonymized database and b) the journalist re-identification scenario, where an adversary tries to randomly re-identify an individual based on the distribution of certain quasi-identifiers, demographic attributes, or even sensitive attributes. Researchers have recently proposed visual uncertainty-based risk quantification. Researchers in application domains like healthcare [GDLS14] discuss how privacy-preserving data sharing risks can be mitigated in a non-interactive privacy scenario, by restricting the queries that can be used for exploring the data. These concepts can also be applied in case of interactive visualization, where different visualiza-



**Figure 8: A privacy-preserving data visualization pipeline [WCC\*17]** that helps data owners understand the effects of privacy parameters on the data transformation steps, eventually leading to a visualization with an appropriate balance between privacy gain and loss of utility.

tion configurations are evaluated carefully for risk factors before making them publicly accessible. Data owners thus need to rigorously identify risks before releasing the data. Kao et al. [KHC\*17] presents a novel visualization interface named ODD visualizer which will help in open data de-identification, i.e., if there is any privacy leakage in the dataset. It uses heat maps to display  $k$ -anonymity and  $l$ -diversity distributions. This is similar to the approaches of Castellani et al. [CRVFS15], who propose a visualization based data profiler for understanding potential vulnerabilities in openly available city data, and Deeb-Swhart et al., where they evaluate strategies to help law enforcement officials combat human trafficking while ensuring privacy protection [DSEB19]. Recently, Dasgupta et al. [DKC19] proposed a suite of metrics using which data owners can estimate the probability of disclosure risks of different configurations of clustered scatter plots and parallel coordinates. The risk quantification model addresses both re-identification scenarios and quantifies the number of guesses an attacker had to make before knowing the precise value of an attribute or the location of a record within a cluster (Figure 7). Assessing these risks can help data owners decide an appropriate level of privacy they are comfortable with, before releasing the visualization for public access. Another example of such a task includes the analysis of privacy preservation with human trajectory data [WGL\*18]. Wang et al. conducted experiments to understand how a user can analyze the movement behaviors using trajectories and how they can locate specific positions on these trajectories. They observed that trajectory analysis is more accurate and even less time consuming while using Positions of Interest (POI) than road networks or histogram but locating positions on a trajectory is almost same in POI and Road network methods. This paper also comments that the capability of these features in trajectory analysis and privacy exposure may differ for various trajectories, based on the area cov-

ered. Thus the combination of multiple features may generate new knowledge, but it also increases privacy risk.

In one of the few examples focusing on evaluating privacy risks for a data subject, Takano et al. proposed a visualization system [TOT\*14] for making users aware of how different entities for website tracking can potentially compromise user identity without their knowledge. In another such example, Muchagata et al. [MVMF19] presented a text-based interface in a mobile application which will help patients and healthcare professionals to monitor health data. The most important feature of this visualization, named Adaptive Graphical Visualization Interface (AGVI), is the interface is user-adaptive, i.e., it changes according to the user's needs. This paper observes that adaptive visualization techniques can influence the users' perspective on security and privacy of a mobile application but the roles of the user (patient or healthcare professional) and their goals (searching for medications or analyzing patients' tests) can influence this perspective. This is the only example where we found that an interface is tested with respect to multiple roles and design considerations are presented from both a data subject and a data consumer's perspective.

### 5.2.3. Understand Policy

*Understanding Policy* was the third most common task in our collection, with a coverage of about 25% of the papers we surveyed targeting both data subjects and data owners as users. Bahrini et al. [BWM\*19] discuss how a mobile application can help users to understand which user information is accessible by the granted permissions. This interactive visualization will help the users make an informed decision about whether to install a certain application or not. The authors claim that the results of their evaluation state that by promoting user awareness regarding permissions required by mobile applications (Android), users pay more attention to these

Paper	Original Channel	Visualization Technique	Low-Level Task	Vulnerability	Privacy-Preserving Channel	Modified Vis Task	Risk Source
DK11	Position	Parallel Coordinates	Identify	Disclosure- both	Area	Summarize	Interaction
AAF16		Geographical Map	Locate	Identity disclosure	Density	Distribution	Interaction
Wang2017		Scatterplot	Identify	Attribute disclosure	Area	Distribution	Interaction
Kung2017		Multidimensional projection	Identify	Identity disclosure	Containment	Group	Knowledge
Dasg14		Pixel-based	Detect trends	Attribute disclosure	Color	Detect trends	Interaction
Wang2017		Scatterplot	Distribution	Attribute disclosure	Area	Distribution	Knowledge
Mazzia2012		Multidimensional projection	Group	Identity disclosure	Containment	Same	Knowledge
DMK14		Height	Bar Chart	Attribute disclosure	Same	Compare	Distribution
DMK14		Area	Treemap	Attribute disclosure	Same	Compare	Distribution
CY16		Volume rendering	Detect shape	Attribute disclosure	Same	Detect shape	Knowledge
Dasg14	Shape	Glyph	Detect patterns	Attribute disclosure	Same	Same	Knowledge
Xiao2018		Scatterplot	Distribution	Attribute disclosure	Same	Same	Distribution

**Figure 9: Dissecting the design space of privacy-preserving visualization** in terms of the transformation of the original channel (used for encoding the raw data) to a privacy-preserving channel. In particular, we point to the vulnerability of the high-accuracy channels like position and also highlight the counter-intuitive fact that even low-accuracy channels like area and shape can be exploited by attackers.

permissions. The paper also tested system usability using error bars for different versions of the application and concluded that the version with more detailed description/flow of permissions has greater usability. Dhotre et al. [DBKO17] implemented a method to perform semi-automatic analysis of the privacy policies of certain websites and generate visualization in order to help the user understand the policies better. This visualization interface, consisting of pie charts, helps the user understand the use of different Personally Identifiable Information (PII) by the website, according to their privacy policies. The interface also summarizes certain sections like use of cookies, information sharing policies and help the users to understand them better. The Privacy Policy Elucidator Tool (PPET) collects the privacy policies from different websites, parse them, classify them using machine learning techniques like Naïve Bayes classifier and uses the extracted paragraph and summary for the visualization. It also evaluates the trustworthiness of the website and displays the same through a donut visualization. Ghazinour et al. [GMB09] present a visualization model which will help the data owners understand the privacy policy of a website and help the policy officers to better understand the designed policies. The Privacy Policy Visualization Model (PPVM) involve the use of relationship diagrams to help in the following tasks: understand privacy policies of these websites when using the name and email address of individuals to send notifications regarding new services, not collecting data of anyone under a certain age limit, disclose user information pursuant to lawful requests etc. The model suggests to highlight the purpose(P), granularity(G), visibility(V), retention(R) and constraint(C) of the privacy policies in this relationship diagrams. Becker et al. [BHÖK14] reflects whether using visualizations to communicate privacy and security measures have positive effects on trust. Infographics are used to depict certain privacy concepts like SSL encryption, AES encryption and studied the improvement on privacy and trust. The study concluded that though these descriptive images have a positive effect on the trust in the provider, there was no significant improvement regarding data security and privacy, in comparison to the text-based privacy policy.

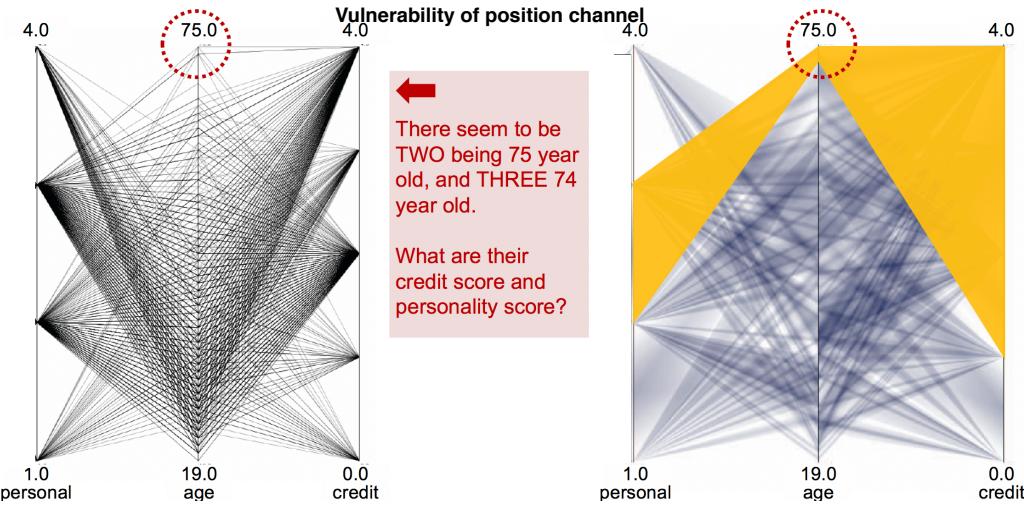
#### 5.2.4. Evaluate Trade-Offs

The task of evaluating trade-offs, performed mainly by data owners or custodians, had a coverage of about 18% of the papers we

surveyed. Wang et al. [WCC\*17] developed a combination of tree-based and matrix-based visualization techniques for helping data consumers dynamically understand the effect of privacy parameters on the difference between the original data and the processed data (Figure 8). They propose the construction of a Privacy Exposure Risk Tree for interactively controlling how hierarchical attributes are organized and selecting parameters values of a privacy model based on differential privacy. A matrix-based view is then used to observe the change in two-dimensional distributions of different combinations of selected attributes. At the end of this process, they can also export an anonymized dataset. Xiao et al. [XLZ\*18] presents a visualization tool named VISEE which will help to maintain the balance between high application utility and less privacy leakage in the case of sharing of sensor data. Accelerometer data collected from different mobile devices has been used as an example. The visualization focuses on representing the degree of mutual information between different pairs of variables. Parallel coordinates, feature grid diagram, and ranking chart help select the appropriate combination of features and sampling rates, thus making a good decision on the trade-off between utility and privacy. For data subjects, Wang et al. proposed an interactive visualization tool for users who can share their personality portraits by tuning the privacy settings, visualized in the form of linked bar charts [WGX\*15]. Ragan et al. [RKIW18] presents an interactive interface where the user starts with fully masked de-identified data and later clicks to open when more information is required for making better decisions. This is a system that reduces privacy risk through on-demand incremental information disclosure. Box plots have been used to analyze the test results in different masking levels like full, moderate, low and masked.

#### 5.2.5. Compare algorithms

The task of comparing algorithms had a coverage of about 18% of the papers we surveyed, focused mainly on data owners to understand how different algorithms have an effect on privacy or re-identification risks. A significant challenge in incorporating multiple models is comparing the effectiveness of different anonymization schemes as both the privacy requirements can drastically change across datasets and user background. To address this problem, Wang et al. developed a tool called GraphProtec-



**Figure 10: Illustrating vulnerability in a position-based encoding, where clustering can help transform a position-based encoding to an area-based encoding and protect against sensitive queries.**

tor [WCC<sup>\*</sup>18] that guides users based on the transformation steps in a privacy-preservation pipeline. Using interactive visualization in the form of a graph, users can manipulate sensitive and non-sensitive nodes and their connections and observe the structural changes to the graph that interferes with utility. Ultimately, they can make better decisions about which algorithm is appropriate for their data and privacy goals.

Kung et al. [Kun17] uses Discriminant Component Analysis (DCA), a supervised version of Principal Component Analysis (PCA) for the visualization because DCA can support data of high compression (small dimensionality) and the recoverability can be controlled. This paper has also compared among the results of different clustering methods using multidimensional projections using which users can compare and effectiveness of this approach.

## 6. Critical Reflection on the Design Space

The goal of a conventional visualization or visual analytics technique is to facilitate generation of insights from data. While the definition of insights itself has been debated by several researchers [Nor06, CZGR09], there is no denying the fact that visualization processes maximize the amount of information that can be encoded in and decoded from a visual representation. This is contrast to the goal of any privacy-preserving data visualization technique, where the goal is to restrict data consumers from accessing sensitive information or helping data owners understand the trade-offs and policies governing such restrictions. In this section, we aim to study how this contrast is reflected in the design choices. To this end, we refer to the literature on ranking of channels [CM84, Ber11] and analyze the role of high-accuracy channels (e.g., position) and low-accuracy channels (e.g., area) for privacy preservation purposes. We include techniques from our collection and augment that collection with techniques that use either classes of these channels. We first discuss a classification scheme (Figure 9) and organize our analysis around three themes: i) transformation of high-accuracy channels, ii) vulnerability of low-accuracy channels, and iii) the relative utility of these channels when transformation is applied for privacy-preservation purposes.

### 6.1. Classification Scheme

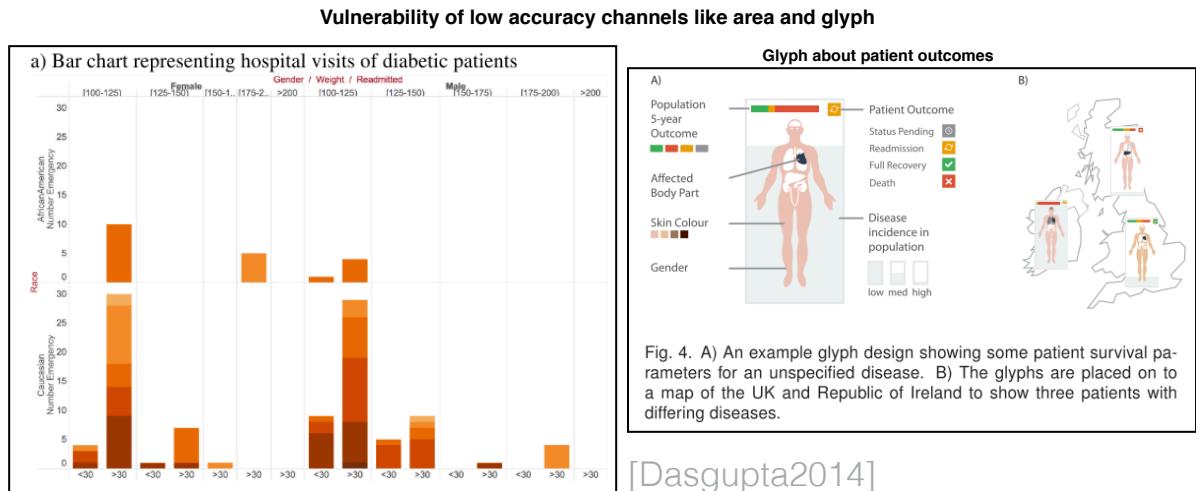
Privacy-preserving data visualization techniques use a transformation of the channels that would be otherwise used for visualizing the de-anonymized data. As part of our classification scheme (Figure 9), we group the techniques based on the **original channel** that is used for visualizing the raw data and for each of them, identify the **privacy-preserving channel**.

We use the task taxonomy proposed by Brehmer and Munzner [BM13] to distinguish between the high-level **privacy-preserving task** (i.e., *why* a task is performed) and the **low-level visualization task** (i.e., *how* a task is performed).

The main reason for a privacy focused transformation (e.g., a scatter plot transformed to a clustered scatter plot) is to prevent the original tasks from being performed owing to their vulnerability. Therefore, we also look at the **modified visualization task**, and introspect on the relative difference in utility between the original and the anonymized visualization. Finally, we also reflect on what possible **risks** could be associated with the anonymized visualization. Such risks can stem from interactivity of a visualization, where additional context or description is provided or from the background knowledge of an attacker.

### 6.2. Vulnerability of High-Accuracy Channels

In geographical maps and in multidimensional visualization techniques like scatter plots and parallel coordinates, position is the primary encoding channel. Assuming that individuals are represented using these visualizations, a high-accuracy channel like position can help identify individuals and thereby leading to a privacy risk of identity disclosure. Privacy-preserving parallel coordinates and scatter plots have been proposed by generalization through  $k$ -anonymity [DK11a], where records are visualized as clusters. When the position visual variable provides the primary encoding, then we can exploit the difference in resolution between the screen space and data space to inherently lose information through binning, etc. This when used as a parameter for controlling a privacy-preserving algorithm, can produce visualizations with both high



**Figure 11: Illustrating vulnerability in bar charts and glyphs, where despite aggregation and use of low-accuracy channels, information can be recovered using the data distribution or background knowledge.**

privacy and utility. However, it has been shown that cluster-based  $k$ -anonymous parallel coordinates and scatter plots have certain vulnerabilities from record linkage and attribute linkage [DCK13].

An example of such vulnerability is shown in Figure 10. In this case, the edges of clusters represent real data points. If an attacker is aware about, say the age of a person, as shown in the figure, and the pixel coordinate of that data point coincides with a cluster border, then the location of the record is revealed. On the other hand, if the pixel coordinate is a non-edge point within a cluster, that provides higher privacy. With respect to attribute linkage, one can geometrically derive the number of possible cluster configurations given different values of  $k$  and use that for guessing the linkage between adjacent attributes. Reordering and brushing can enable an attacker to choose a different adjacency configuration of quasi-identifiers and browse through subset of records. Dasgupta et al. [DCK13] have proposed different screen-space metrics that aim to constrain such interactions based on the privacy risks.

Transformation of the position channel to a density-based representation in geographical maps [AAFJ16] is also common, where users can gauge the distribution instead of locating individuals. Such manipulation of pixels is also possible with non-spatial pixel-base visualization techniques, where value of an attribute is mapped to colors according to a chosen color scale [Kei00]. However, in case of interactive pixel-based visualization [DMARC14], each pixel can be an entry point to an individual's data point, and malicious users can use a number of educated guesses to know the value of an attribute. Pixel-based representations can also become vulnerable when linked with other contextualizing representations.

Other approaches towards transformation of the position channel include the use of containment metaphor in the case multidimensional projections [Kun17] and converting raw scatter plot representation to a representation of distributions [WCC<sup>\*</sup>17]. While such transformations guarantee a minimum level of privacy, they are also vulnerable to interaction, especially, drill-down operations, which should be adaptively restricted based on the associated risks.

### 6.3. Vulnerability of Low-Accuracy Channels

Low-accuracy channels like area, density, shape etc., which generally represent aggregated data, can be intuitively thought of as being inherently privacy-preserving. In this case, one is unable to observe the exact value of an attribute or locate a record precisely. However, as demonstrated in earlier work [DMARC14], such an assumption is not valid in many real-world use cases. As shown in the bar chart (Figure 11), some patterns stand out, like the correlation between high re-admission rate and number of emergency visits for male and female African Americans aged 50 to 60. There is only one category with non-zero frequency in re-admission greater than 30, and these are Caucasian males, aged 40 to 50. This implies that with knowledge of quasi-identifiers such as race and age, deducing the diabetic condition would not be hard. Similarly glyphs [BKC<sup>\*</sup>13] can also be thought of harmless from a privacy-preservation perspective, however, as shown in the glyph in Figure 11, more information can be potentially determined about the patients based on the background knowledge of the attacker. Glyphs are popular visual representations in the healthcare domain because of the intuitive nature of the representation. However, such information when integrated with openly available attributes, patient identity can be at risk: using small DNA sequences from the Y chromosome, researchers at MIT were able to extract the genealogical information (surname, relatives) and religious background of fifty people from the 1000 Genomes Project [GMG<sup>\*</sup>13]. The same rationale applies to use of shapes in the case of volume rendering [CY16]. In summary, low-accuracy channels do not guarantee the preservation of privacy and appropriate risks should be assessed in the context of the externally available information about the individuals who are represented.

### 7. Gaps and Research Opportunities

Based on our survey, we present an analysis of the key gaps and research opportunities thereof. We organize this section based on research themes, each of which addresses the following key questions motivated by the well-known Helmeijer catechism [GH20]:

- What are the limitations of the current practices of privacy-preserving data visualization?
- Why is it important to address those limitations?
- How does a research approach or contribution look like, for addressing these gaps?
- Who will be the beneficiaries of the proposed research direction: data consumers or owners?

We believe these questions will help us understand both the significance of the research problem and the potential impact of the visualization-specific solutions. We sort the following research themes based on the authors' subjective understanding of the connection between related visualization research and the suggested directions: ones where there are immediate connections are presented first. This is, however, not a commentary on the importance or impact of the suggested research.

### 7.1. Uncertainty Visualization and Privacy

Lack of empirical evaluation of the effect of anonymized visualization on users' perception of privacy is a key gap in the literature. With the exception of a few [RKW18, KRI\*19, CBLM18], we did not find any other examples where controlled studies have been conducted to investigate how well the theoretical guarantees of privacy hold good in practice. Such studies will help data owners and custodians understand the following: how easy or difficult is it for people to breach privacy for a single dataset, how well users can leverage their background knowledge to breach privacy, and what other additional context can either be suppressed or controlled to add uncertainty or confusion in the minds of an attacker. In recent years, the broadly defined research area of uncertainty communication has made a lot of progress [KKH19]. As mentioned before, there is an inherent link between uncertainty and privacy: many anonymization methods can be treated as uncertainty quantification mechanisms and the added uncertainty due to visual mapping has already been termed as visual uncertainty. We need to conduct controlled studies with raw data and visualization with uncertainty encoding and measure the ability of users in terms of time and cognitive effort, to recover the identity of individuals or the values of sensitive attributes by overcoming uncertainty. It would be worthwhile to use Bayesian approaches for modeling how people's background knowledge and prior beliefs can lead to disclosure risks even in the presence of uncertainty in the visualization.

An application of quantification of visual uncertainty (i.e., the uncertainty resulting from the visualization process) is that different views of the data can be calibrated by their degree of vulnerability, in terms of disclosure risks, and interaction constraints can be enforced so that users are only able to access views that guarantee a minimal level of privacy. For coordinated multiple views, this means that details-on-demand [Shn96] can be constrained based on privacy parameters in addition to the users' goals and needs.

### 7.2. Dynamic Visualization of Risks for Privacy Stakeholders

As pointed out recently by a study [RHDM19], there is a high degree of vulnerability of anonymized datasets, especially which contain demographic attributes, even after applying the state-of-the-art privacy-preservation techniques. With the proliferation of IoT

based devices and the evolving concept of smart homes [LB16], such vulnerability will need to be continuously evaluated by both data subjects and technology developers. This is a key gap in the literature, where privacy is considered only at the time of release of a dataset and data custodians do not have the tools to re-evaluate risks in the face of newly released datasets or other attack scenarios. This gap makes most of the open data repositories vulnerable to privacy breach, even though personally identifiable information is not present in those datasets. With respect to visualization research we found very few papers [KRI\*19, RKW18] focusing on this aspect of privacy. There is a fertile ground for visualization research that aims at communicating vulnerabilities in open data and privacy-utility trade-offs to all stakeholders.

Visualization-based interfaces can play a key role in helping data owners, subjects, custodians, and consumers dynamically evaluate the disclosure risks of shared data. For data owners or custodians, visual interfaces [GHK\*16, CRVFS15] can help communicate privacy risks by suggesting non-obvious, probabilistic linkages [HWL\*19], let them dynamically evaluate the trade-offs among data utility and privacy risks [ADSZ\*19] by visualizing privacy outcomes from new and evolving metrics [JSH\*17], and make more confident decisions regarding data sharing [BVM\*17].

### 7.3. Privacy-Aware Citizen Science

Developing smart cities with the help of data collected about citizens' mobility patterns, preferences, habits etc., is a potential which has attracted the attention of governments across the world. However, this also means that data about people's location and movement are more vulnerable than ever before. The New York Times report [TW20], which we pointed to earlier, and shows about the ease with which people's location can be known, is alarming. While this cannot be solved simply by applying computational techniques, this issue is symptomatic of the opaque ways in which urban data is collected and administered. A study had previously demonstrated how urban mobility data collected by analyzing New York City taxi trips can compromise the identity of individuals [DDFS16]. This is a research gap relevant to both data owners and data subjects, as it is the individual's data that is collected and analyzed in this case. While we have encountered several papers [AA12, AAFJ16] focusing on privacy issues of spatial data, such research needs to be integrated more deeply with the research involving privacy-preserving urban data collection [LA15, TAS\*18] and decision-making. Research grounded in behavioral sciences that has recently demonstrated the benefits of using visualization-based interfaces for granting citizens the transparency to directly administer and understand the implications of data sharing [DGY\*19]. Visualization techniques need to be further developed and explored for more inclusive and transparent citizen science, where third party interference can be minimized and citizens can more proactively exercise their right to privacy.

### 7.4. Ethical Data Visualization through Privacy by Design

Researchers in computing and data-driven technologies are becoming more cognizant of the moral obligations and ethical implications of research [Var19]. Automated analysis, machine learning

and provenance should be controlled, and it should allow those impacted by the decisions to appeal their decisions or seek better outcomes. We have certain ethical obligations as visualization designers, as we generally have complete access to data and the freedom to portray insights derived about people. When we are presenting data to the public, as visualization designers, an abiding principle should be to protect the privacy of the people whose data we have collected and visualized, even if at the cost of communicating our key findings. Both data and data visualization are not ethically neutral activities and thus there is an obligation to be ethical while representing data [Cor19]. Integrating principles of "privacy by design" [WM19] in visualization interfaces will be a key research opportunity to this end. Moreover, as natural language interfaces begin to be integrated with visualization techniques [YS19] and visualization techniques begin to be augmented with text-based facts [SDES18] care should be taken that the computationally generated facts are privacy-preserving as well.

### 7.5. Interpretable Privacy Policy-Making

In the face of new legislations like the GDPR and questionable practices by online companies about privacy policy communication, we foresee a significant amount of research effort being dedicated towards interpretable policy-making: where both data subjects and data owners can better understand privacy parameters before implementing policies and data consumers can overcome the barriers of intended [MAF\*19] or unintended [OOH20] obfuscation for better understanding policy implications. In our collection we encountered several papers [DBKO17, BHÖK14, GMB09, Yee06] dedicated towards studying this problem. But most of this research is concentrated in the application-specific domains. Greater collaborative efforts across domain experts and visualization designers can significantly improve the quality of the visualization techniques we encountered. In many of these cases, we found data-flow diagrams, relationship diagrams or infographics being used as a means of communicating policies. Except for Dhotre et al. [DBKO17], we sensed a lack of quality in the visual communication of information extracted from policy text. We believe that recent advances in text visualization and topic modeling [LWC\*18] can have a significant effect on improving visualization techniques for communicating privacy parameters and their dependencies, as extracted from policy descriptions, and make that information accessible and actionable, especially for data subjects, who might not have appropriate levels of data literacy to comprehend the privacy risks and policies.

### 7.6. Privacy-Preserving and Inclusive Visualization

Many recent studies have shown that it is the poor and marginalized section of the society, who are in the greatest danger of violation of their privacy rights [New14, SBSS18]. In our collection we found research focusing on law enforcement agencies which collect data about potential human trafficking involving vulnerable people [DSEB19]. Care needs to be taken to preserve privacy of these data subjects, who are vulnerable and do not have access to services otherwise guaranteed in urban areas. With the proliferation of smartphones and wearable technology [HAK\*19], visualization techniques can be used to collect data from people

who need assistance, legal, social, or otherwise. Inspiration can be drawn from a recent study on visualization perception in Rural America [PAEE19], and visualization can be used as a privacy-preserving data collection medium, where people can "see" themselves as part of a larger societal structure and can also get assurance about their privacy not being violated. Visualization designers and researchers have a unique opportunity to be inclusive of marginalized and underrepresented population, while at the same time, respecting the ethics of preserving privacy.

## 8. Conclusion

We live in the times of constant threat to individual privacy where all of us are mere data points as part of some data-driven digital commodity. There are many risks to such massive collection and aggregation of data, where data can be de-anonymized and individuals can be re-identified without their consent for malicious purposes. In this survey, we have reflected on the challenges and opportunities that we face in the visualization community, with respect to the larger socio-technical challenge of privacy-preservation. One of the key opportunities for the field of privacy-preserving data visualization is to develop novel solutions with data subjects as the stakeholder, many of whom are often at the receiving end of uninterpretable privacy policies or are exposed to greater privacy risk, since they come from vulnerable sections of the society. We believe there is scope for immediate impact with all the research directions outlined above. They will help us progress along the path of resolution of the ongoing and ever-increasing dichotomy between individual privacy and data-driven consumerism.

## References

- [A\*03] ANNAS G. J., ET AL.: Hipaa regulations-a new era of medical-record privacy? *New England Journal of Medicine* 348, 15 (2003), 1486–1490. 1
- [AA12] ANDRIENKO G., ANDRIENKO N.: Privacy issues in geospatial visual analytics. In *Advances in Location-Based Services*. Springer, 2012, pp. 239–246. 14
- [AAFJ16] ANDRIENKO N., ANDRIENKO G., FUCHS G., JANKOWSKI P.: Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization* 15, 2 (2016), 117–153. 9, 13, 14
- [ADSZ\*19] ANHALT-DEPIES C., STENGLIN J. L., ZUCKERBERG B., TOWNSEND P. M., RISSMAN A. R.: Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. *Biological Conservation* 238 (2019), 108195. 14
- [AFYH09] ANWAR M., FONG P. W., YANG X.-D., HAMILTON H.: Visualizing privacy implications of access control policies in social network systems. In *Data privacy management and autonomous spontaneous security*. Springer, 2009, pp. 106–120. 7
- [BA05] BAYARDO R. J., AGRAWAL R.: Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)* (2005), IEEE, pp. 217–228. 2
- [Ber11] BERTIN J.: *Graphics and graphic information processing*. Walther de Gruyter, 2011. 5, 12
- [BHÖK14] BECKER J., HEDDIER M., ÖKSÜZ A., KNACKSTEDT R.: The effect of providing visualizations in privacy policies on trust in data privacy and security. In *2014 47th Hawaii International Conference on System Sciences* (2014), IEEE, pp. 3224–3233. 7, 11, 15

- [BKC\*13] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics (STARs)* (2013), pp. 39–63. 13
- [BLJ08] BERTINO E., LIN D., JIANG W.: A survey of quantification of privacy preserving data mining algorithms. *Privacy-Preserving Data Mining* (2008), 183–205. 2, 9
- [BLNR07] BU S., LAKSHMANAN L. V., NG R. T., RAMESH G.: Preservation of patterns and input-output privacy. In *2007 IEEE 23rd International Conference on Data Engineering* (2007), IEEE, pp. 696–705. 1, 4
- [BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2376–2385. 12
- [BPL19] BÜSCHER M., PERNG S.-Y., LIEGL M.: Privacy, security, and liberty: Ict in crises. In *Censorship, Surveillance, and Privacy: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2019, pp. 199–217. 1
- [BS08] BRICKELL J., SHMATIKOV V.: The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), ACM, pp. 70–78. 3
- [BVM\*17] BARCELLOS R., VITERBO J., MIRANDA L., BERNARDINI F., MACIEL C., TREVISAN D.: Transparency in practice: using visualization to enhance the interpretability of open data. In *Proceedings of the 18th Annual International Conference on Digital Government Research* (2017), pp. 139–148. 14
- [BWM\*19] BAHRINI M., WENIG N., MEISSNER M., SOHR K., MALAKA R.: Happyperm: Presenting critical data flows in mobile application to raise user security awareness. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. LBW0262. 7, 10
- [CAS05] CONTI G., AHAMAD M., STASKO J.: Attacking information visualization system usability overloading and deceiving the human. In *Proceedings of the 2005 symposium on Usable privacy and security* (2005), ACM, pp. 89–100. 5
- [CBLM18] CHOU J.-K., BRYAN C., LI J., MA K.-L.: An empirical study on perceptually masking privacy in graph visualizations. In *2018 IEEE Symposium on Visualization for Cyber Security (VizSec)* (2018), IEEE, pp. 1–8. 14
- [CBM17] CHOU J.-K., BRYAN C., MA K.-L.: Privacy preserving visualization for social network data with ontology information. In *2017 IEEE Pacific Visualization Symposium (PacificVis)* (2017), IEEE, pp. 11–20. 7, 9
- [CGH18] CADWALLADR C., GRAHAM-HARRISON E.: Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian* 17 (2018), 22. 1
- [CM84] CLEVELAND W. S., MCGILL R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554. 5, 12
- [Cor19] CORRELL M.: Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13. 15
- [CRVFS15] CASTELLANI RIBEIRO D., VO H. T., FREIRE J., SILVA C. T.: An urban data profiler. In *Proceedings of the 24th International Conference on World Wide Web* (2015), pp. 1389–1394. 10, 14
- [CWM16] CHOU J.-K., WANG Y., MA K.-L.: Privacy preserving event sequence data visualization using a sankey diagram-like representation. In *SIGGRAPH ASIA 2016 Symposium on Visualization* (2016), ACM, p. 1. 7
- [CWM19] CHOU J.-K., WANG Y., MA K.-L.: Privacy preserving visualization: A study on event sequence data. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 340–355. 7, 9
- [CY16] CHOU J.-K., YANG C.-K.: Obfuscated volume rendering. *The Visual Computer* 32, 12 (2016), 1593–1604. 9, 13
- [CZGR09] CHANG R., ZIEMKIEWICZ C., GREEN T. M., RIBARSKY W.: Defining insight for visual analytics. *IEEE Computer Graphics and Applications* 29, 2 (2009), 14–17. 12
- [DBKO17] DHOTRE P. S., BIHANI A., KHAJURIA S., OLESEN H.: “take it or leave it”: Effective visualization of privacy policies. In *Cybersecurity and Privacy*. River Publishers, 2017, pp. 39–64. 7, 11, 15
- [DCK12] DASGUPTA A., CHEN M., KOSARA R.: Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum* 31, 3pt2 (2012), 1015–1024. 7
- [DCK13] DASGUPTA A., CHEN M., KOSARA R.: Measuring privacy and utility in privacy-preserving visualization. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 35–47. 8, 9, 13
- [DCK19] DASGUPTA A., CHEN M., KOSARA R.: Guess me if you can: A visual uncertainty model for transparent evaluation of disclosure risks in privacy-preserving data visualization. In *IEEE Symposium on Visualization for Cyber Security (VizSec, in publication)* (2019). 9
- [DDFS16] DOURIEZ M., DORAISWAMY H., FREIRE J., SILVA C. T.: Anonymizing nyc taxi data: Does it matter? In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (2016), IEEE, pp. 140–148. 14
- [DDK16] DYKE S. O., DOVE E. S., KNOPPERS B. M.: Sharing health-related data: a privacy test? *NPJ genomic medicine* 1 (2016), 16024. 2
- [DGY\*19] DENNIS S., GARRETT P., YIM H., HAMM J., OSTH A. F., SREEKUMAR V., STONE B.: Privacy versus open science. *Behavior research methods* 51, 4 (2019), 1839–1848. 14
- [DK11a] DASGUPTA A., KOSARA R.: Adaptive privacy-preserving visualization using parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2241–2248. 8, 9, 12
- [DK11b] DASGUPTA A., KOSARA R.: Privacy-preserving data visualization using parallel coordinates. In *Visualization and Data Analysis 2011* (2011), vol. 7868, International Society for Optics and Photonics, p. 78680O. 9
- [DKC19] DASGUPTA A., KOSARA R., CHEN M.: Guess me if you can: A visual uncertainty model for transparent evaluation of disclosure risks in privacy-preserving data visualization. *VizSec* (2019). 9, 10
- [DL86] DUNCAN G. T., LAMBERT D.: Disclosure-limited data dissemination. *Journal of the American Statistical Assn.* 81, 393 (1986), pp. 10–18. URL: <http://www.jstor.org/stable/2287959.4>
- [DMARC14] DASGUPTA A., MAGUIRE E., ABDUL-RAHMAN A., CHEN M.: Opportunities and challenges for privacy-preserving visualization of electronic health record data. In *Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records* (2014). 13
- [DR\*14] DWORK C., ROTH A., ET AL.: The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407. 3
- [DSEB19] DEEB-SWIHART J., ENDERT A., BRUCKMAN A.: Understanding law enforcement strategies and needs for combating human trafficking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. 331. 10, 15
- [DTZ08] DU W., TENG Z., ZHU Z.: Privacy-maxent: integrating background knowledge in privacy quantification. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (2008), ACM, pp. 459–472. 4
- [EKM\*19] ELAGROUDY P., KHAMIS M., MATHIS F., IRMSCHER D., BULLING A., SCHMIDT A.: Can privacy-aware lifelogs alter our memories? In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. LBW0244. 7
- [FWCY10] FUNG B. C., WANG K., CHEN R., YU P. S.: Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSur)* 42, 4 (2010), 1–53. 1, 2

- [GB13] GAO B., BERENDT B.: Circles, posts and privacy in egocentric social networks: An exploratory visualization approach. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (2013), IEEE, pp. 792–796. 7
- [GB16] GOTZ D., BORLAND D.: Data-driven healthcare: Challenges and opportunities for interactive visualization. *IEEE computer graphics and applications* 36, 3 (2016), 90–96. 9
- [GDLS14] GKOULALAS-DIVANIS A., LOUKIDES G., SUN J.: Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics* 50 (2014), 4–19. 2, 4, 9
- [GH20] GEORGE HEILMEIJER D.: *The Heilmeijer Catechism*, 1977 (accessed January 23, 2020). URL: <https://www.darpa.mil/work-with-us/heilmeier-catechism>. 13
- [GHK\*16] GABOARDI M., HONAKER J., KING G., MURTAGH J., NISSIM K., ULLMAN J., VADHAN S.: Psi ( $\{\Psi\}$ ): a private data sharing interface. *arXiv preprint arXiv:1609.04340* (2016). 14
- [GMB09] GHAZINOUR K., MAJEDI M., BARKER K.: A model for privacy policy visualization. In *2009 33rd Annual IEEE International Computer Software and Applications Conference* (2009), vol. 2, IEEE, pp. 335–340. 11, 15
- [GMG\*13] GYMREK M., MCGUIRE A. L., GOLAN D., HALPERIN E., ERLICH Y.: Identifying personal genomes by surname inference. *Science* 339, 6117 (2013), 321–324. URL: <http://www.sciencemag.org/content/339/6117/321.abstract>, arXiv:<http://www.sciencemag.org/content/339/6117/321.full.pdf>, doi:10.1126/science.1229566. 13
- [HAG\*20] HELLEWELL J., ABBOTT S., GIMMA A., BOSSE N. I., JARVIS C. I., RUSSELL T. W., MUNDAY J. D., KUCHARSKI A. J., EDMUNDS W. J., SUN F., ET AL.: Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health* (2020). 2
- [HAK\*19] HICKS J. L., ALTHOFF T., KUHAR P., BOSTJANCIC B., KING A. C., LESKOVEC J., DELP S. L., ET AL.: Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ digital medicine* 2, 1 (2019), 1–12. 15
- [HSH14] HONGDE R., SHUO W., HUI L.: Differential privacy data aggregation optimizing method and application to data visualization. In *2014 IEEE Workshop on Electronics, Computer and Applications* (2014), IEEE, pp. 54–58. 7
- [HWL\*19] HEJBLUM B. P., WEBER G. M., LIAO K. P., PALMER N. P., CHURCHILL S., SHADICK N. A., SZOLOVITS P., MURPHY S. N., KOHANE I. S., CAI T.: Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific data* 6 (2019), 180298. 14
- [JEB12] JOHNSON M., EGELMAN S., BELLOVIN S. M.: Facebook and privacy: it's complicated. In *Proceedings of the eighth symposium on usable privacy and security* (2012), ACM, p. 9. 2
- [JSH\*17] JIA R., SANGOGBOYE F. C., HONG T., SPANOS C., KJÆRGAARD M. B.: Pad: protecting anonymity in publishing building related datasets. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments* (2017), pp. 1–10. 14
- [KDFK15] KANG R., DABBISH L., FRUCHTER N., KIESLER S.: “my data just goes everywhere:” user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)* (2015), pp. 39–52. 4
- [Kei00] KEIM D. A.: Designing pixel-oriented visualization techniques: Theory and applications. *Visualization and Computer Graphics, IEEE Transactions on* 6, 1 (2000), 59–78. 13
- [KHC\*17] KAO C.-H., HSIEH C.-H., CHU Y.-F., KUANG Y.-T., YANG C.-K.: Using data visualization technique to detect sensitive information re-identification problem of real open dataset. *Journal of Systems Architecture* 80 (2017), 85–91. 5, 10
- [KKH19] KALE A., KAY M., HULLMAN J.: Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–14. 14
- [KRI\*19] KUM H.-C., RAGAN E. D., ILANGOVAN G., RAMEZANI M., LI Q., SCHMIT C.: Enhancing privacy through an interactive on-demand incremental information disclosure interface: applying privacy-by-design to record linkage. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)* (2019). 14
- [Kun17] KUNG S.-Y.: Discriminant component analysis for privacy protection and visualization of big data. *Multimedia Tools and Applications* 76, 3 (2017), 3999–4034. 12, 13
- [LA15] LOUGHLIN M., ADNANE A.: Privacy and trust in smart camera sensor networks. In *2015 10th International Conference on Availability, Reliability and Security* (2015), IEEE, pp. 244–248. 14
- [Lam93] LAMBERT D.: Measures of disclosure risk and harm. *Journal of Official Statistics* 9 (1993), 313–331. 4
- [LARC16] LICCIARDI I., ABDUL-RAHMAN A., CHEN M.: I know where you live: Inferring details of people’s lives by visualizing publicly shared location data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 1–12. 9
- [LB16] LIN H., BERGMANN N. W.: IoT privacy and security challenges for smart home environments. *Information* 7, 3 (2016), 44. 14
- [Lee17] LEE H. B.: *Visualization and Differential Privacy*. PhD thesis, University of Illinois at Urbana-Champaign, 2017. 3
- [LGG\*19] LJUBIC B., GLIGORIJEVIC D., GLIGORIJEVIC J., PAVLOVSKI M., OBRADOVIC Z.: Social network analysis for better understanding of influenza. *Journal of biomedical informatics* 93 (2019), 103161. 9
- [LL09] LI T., LI N.: On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 517–526. 2, 3
- [LLV07] LI N., LI T., VENKATASUBRAMANIAN S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (2007), IEEE, pp. 106–115. 3
- [LWC\*18] LIU S., WANG X., COLLINS C., DOU W., OUYANG F., ELASSADY M., JIANG L., KEIM D. A.: Bridging text visualization and mining: A task-driven survey. *IEEE transactions on visualization and computer graphics* 25, 7 (2018), 2482–2504. 15
- [MAF\*19] MATHUR A., ACAR G., FRIEDMAN M. J., LUCHERINI E., MAYER J., CHETTY M., NARAYANAN A.: Dark patterns at scale. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov 2019), 1–32. URL: <http://dx.doi.org/10.1145/3359183>, doi:10.1145/3359183. 1, 7, 15
- [MFG\*06] MONTEMAYOR J., FREEMAN A., GERSH J., LLANSO T., PATRONE D.: Information visualization for rule-based resource access control. In *Proc. of Int. Symposium on Usable Privacy and Security ({SOUPS})* (2006), pp. 24–40. 4
- [MGKV06] MACHANAVAJHALA A., GEHRKE J., KIFER D., VENKATASUBRAMANIAM M.: l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE’06)* (2006), IEEE, pp. 24–24. 3
- [MKM\*07] MARTIN D. J., KIFER D., MACHANAVAJHALA A., GEHRKE J., HALPERN J. Y.: Worst-case background knowledge for privacy-preserving data publishing. In *2007 IEEE 23rd International Conference on Data Engineering* (2007), IEEE, pp. 126–135. 4
- [MLA12] MAZZIA A., LEFEVRE K., ADAR E.: The pviz comprehension tool for social network privacy settings. In *Proceedings of the Eighth Symposium on Usable Privacy and Security* (2012), ACM, p. 13. 2, 7
- [MVMF19] MUCHAGATA J., VIEIRA-MARQUES P., FERREIRA A.: mhealth applications: Can user-adaptive visualization and context affect the perception of security and privacy? 10

- [MX07] MOTWANI R., XU Y.: Efficient algorithms for masking and finding quasi-identifiers. In *Proceedings of the Conference on Very Large Data Bases (VLDB)* (2007), pp. 83–93. 2
- [New14] NEWMAN N.: How big data enables economic harm to consumers, especially to low-income and other vulnerable sectors of the population. *Journal of Internet Law* 18, 6 (2014), 11–23. 15
- [NLV\*20] NOUWENS M., LICCARDI I., VEALE M., KARGER D., KAGAL L.: Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings, SIGCHI (in publication)* (2020). 1
- [Nor06] NORTH C.: Toward measuring visualization insight. *IEEE computer graphics and applications* 26, 3 (2006), 6–9. 12
- [NSW\*17] NISSIM K., STEINKE T., WOOD A., ALTMAN M., BEM-BENEK A., BUN M., GABOARDI M., O'BRIEN D. R., VADHAN S.: Differential privacy: A primer for a non-technical audience. In *Privacy Law Scholars Conf* (2017). 3
- [OBSW15] OKSANEN J., BERGMAN C., SAINIO J., WESTERHOLM J.: Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography* 48 (2015), 135–144. 7
- [OGH05] OLSON J. S., GRUDIN J., HORVITZ E.: A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems* (2005), ACM, pp. 1985–1988. 4
- [OOH20] OBAR J. A., OELDORF-HIRSCH A.: The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147. 1, 15
- [OR19] ORLANDO A. W., ROSOFF A. J.: The new privacy crisis: What's health got to do with it? *The American journal of medicine* 132, 2 (2019), 127–128. 1
- [OSS\*20] OSIA S. A., SHAMSABADI A. S., SAJADMANESH S., TAHERI A., KATEVAS K., RABIEE H. R., LANE N. D., HADDADI H.: A hybrid deep learning architecture for privacy-preserving mobile analytics. *IEEE Internet of Things Journal* (2020). 9
- [PAEE19] PECK E. M., AYUSO S. E., EL-ETR O.: 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* (2019), pp. 1–12. 15
- [RBS20] REICHERT L., BRACK S., SCHEUERMANN B.: *Privacy-preserving contact tracing of covid-19 patients*. Tech. rep., Cryptology ePrint Archive, Report 2020/375, 2020.[Online]. Available: <https://eprint.iacr.org/2020/375.pdf>, 2020. 2
- [RH02] RIZVI S. J., HARITSA J. R.: Maintaining data privacy in association rule mining. In *Proceedings of the 28th international conference on Very Large Data Bases* (2002), VLDB Endowment, pp. 682–693. 4
- [RHDM19] ROCHER L., HENDRICKX J. M., DE MONTJOYE Y.-A.: Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications* 10, 1 (2019), 1–9. 1, 2, 14
- [RJD\*06] RODE J., JOHANSSON C., DiGIOIA P., NIES K., NGUYEN D. H., REN J., DOURISH P., REDMILES D., ET AL.: Seeing further: extending visualization as a basis for usable security. In *Proceedings of the second symposium on Usable privacy and security* (2006), ACM, pp. 145–155. 4
- [RKIW18] RAGAN E. D., KUM H.-C., ILANGOVAN G., WANG H.: Balancing privacy and information disclosure in interactive record linkage with visual masking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 326. 5, 9, 11, 14
- [SBSS18] SRINIVASAN J., BAILUR S., SCHOEMAKER E., SESAGIRI S.: Privacy at the marginsl the poverty of privacy: Understanding privacy trade-offs from identity infrastructure users in india. *International Journal of Communication* 12 (2018), 20. 15
- [SDES18] SRINIVASAN A., DRUCKER S. M., ENDERT A., STASKO J.: Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 672–681. 15
- [Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages* (1996), IEEE, pp. 336–343. 14
- [Swe02] SWEENEY L.: k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570. 2
- [Swe05] SWEENEY L.: Privacy-enhanced linking. *ACM SIGKDD Explorations Newsletter* 7, 2 (2005), 72–75. 2
- [TAS\*18] TONYALI S., AKKAYA K., SAPUTRO N., ULUAGAC A. S., NOJOUMANI M.: Privacy-preserving protocols for secure and reliable data aggregation in iot-enabled smart metering systems. *Future Generation Computer Systems* 78 (2018), 547–557. 14
- [TOT\*14] TAKANO Y., OHTA S., TAKAHASHI T., ANDO R., INOUE T.: Mindyourprivacy: Design and implementation of a visualization system for third-party web tracking. In *2014 Twelfth Annual International Conference on Privacy, Security and Trust* (2014), IEEE, pp. 48–56. 10
- [TW20] THOMPSON S. A., WARZEL C.: *Twelve Million Phones, One Dataset, Zero Privacy*, 2019 (accessed January 23, 2020). URL: <https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>. 14
- [Var19] VARDI M. Y.: Are we having an ethical crisis in computing? *Commun. ACM* 62, 1 (2019), 7. 1, 14
- [VZ19] VALDEZ A. C., ZIEFLE M.: The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies* 121 (2019), 108–121. 3
- [WCC\*17] WANG X., CHOU J.-K., CHEN W., GUAN H., CHEN W., LAO T., MA K.-L.: A utility-aware visual approach for anonymizing multi-attribute tabular data. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 351–360. 10, 11, 13
- [WCC\*18] WANG X., CHEN W., CHOU J.-K., BRYAN C., GUAN H., CHEN W., PAN R., MA K.-L.: Graphprotector: a visual interface for employing and assessing multiple privacy preserving graph algorithms. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 193–203. 7, 12
- [WGL\*18] WANG X., GU T., LUO X., CAI X., LAO T., CHEN W., WU Y., YU J., CHEN W.: A user study on the capability of three geo-based features in analyzing and locating trajectories. *IEEE Transactions on Intelligent Transportation Systems* (2018). 10
- [WGK\*15] WANG Y., GOU L., XU A., ZHOU M. X., YANG H., BADENES H.: Veilme: An interactive visualization tool for privacy configuration of using personality traits. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 817–826. 11
- [WL08] WANG T., LIU L.: Butterfly: Protecting output privacy in stream mining. In *2008 IEEE 24th International Conference on Data Engineering* (2008), IEEE, pp. 1170–1179. 4
- [WM19] WONG R. Y., MULLIGAN D. K.: Bringing design to the privacy table: Broadening “design” in “privacy by design” through the lens of hei. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–17. 15
- [XLZ\*18] XIAO F., LU M., ZHAO Y., MENASRIA S., MENG D., XIE S., LI J., LI C.: An information-aware visualization for privacy-preserving accelerometer data sharing. *Human-centric Computing and Information Sciences* 8, 1 (2018), 13. 5, 11
- [Yee06] YEE G.: Visualization for privacy compliance. In *Proceedings of the 3rd international workshop on Visualization for computer security* (2006), ACM, pp. 117–122. 5, 15
- [YS19] YU B., SILVA C. T.: Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1–11. 15