Measuring the quantity of data privacy and utility tradeoff for users' data: A visualization approach

Master Thesis

Bruno Dobrota
6719473
b.dobrota@students.uu.nl

1st Supervisor: Dr. M. (Michael) Behrisch
2nd Supervisor: Prof. dr. ir. A.C. (Alex) Telea

Business Informatics
Natural Sciences
Utrecht University
22/06/2021

To my parents, sister and Marta,
for their love and support.

Acknowledgement

It is a genuine pleasure to express my deepest gratitude to both supervisors, Dr. Michael Behrisch and Prof. Dr. Ir. Alex Telea, for mentoring me on the thesis. These eight months of the project required patience, experience, and guidance, which the supervisors delivered enthusiastically and professionally. From a student's point, I am more than grateful to have such collaboration and support with both supervisors, whose advice helped finalize the thesis.

I am very thankful to Michael, who continuously encouraged me to accomplish what we assigned as contributions. His understanding, flexibility, and calmness motivated me to further progress and contribute more to the project. By working continuously during these eight months, Michael was always mentoring with a great sense of engagement, motivation, and persistence. I also want to express my deep sense of appreciation to Alex for the advice and valuable criticism of my work that helped me progress. Such mentorship was not only helpful but also encouraging to accomplish desired goals.

Lastly, I would like to thank Utrecht University and all professors and students from Business Informatics master studies for two incredible years. As an international student, this journey made me a whole person ready to contribute to the world. I will always reflect on the University and Utrecht as remarkable moments that made me a better person and academic citizen.

Contents

List of Figures

Abstract

Nowadays, more than ever data comes from different sources, providing an opportunity against various threats. Different privacy issues such as linkage, data breaches, false identities, and other frauds concern both people and organizations. In order to deal with such a problem, the term Privacy-preserving approach with Differential Privacy as the leading mechanism was invented. Local Differential Privacy can be achieved by adding randomized noise into the dataset, however, too much noise could affect the data quality and value of the dataset. The thesis aims to introduce a visualization system that can help users understand how the privacy mechanism affects data and adjust the noise added by those algorithms. In order to provide such a framework, it was decided to implement a visualization system that uses an alternative and simple inspiration of the local differential privacy mechanism to provide visual analysis on specific data. Moreover, specific metrics for inspecting data privacy and utility will be used to evaluate the mechanism's performance. By creating an interactive visualization system that offers to adjust the epsilon parameter with slider and instantly presenting different graphics, users will understand how privacy affects data utility and the opposite. The thesis combines different layers of evaluation that comprehend experts, case study, and technical evaluations to validate the project. As a result, the solution was recognized as a visual analytics approach to explaining the effect of noise-injection levels on a specific dataset by taking four layers of evaluation methods. In addition, the experts agreed that the project contributes to defining privacy-preserving visual analytics as the first approach that explains the means of data privacy and utility tradeoff with the visualization system.

1 Introduction

In the Introcution section, we will present the motivation and starting point behind the thesis. This means that the topic will be introduced, however, since the project consists of multiple topics that origin from different domains, we will explain each subject. The introduction will give readers a profound understanding of the meaning, current situation, and possible improvements of each topic. Lastly, the whole thesis and its structure will be briefly explained.

1.1 Privacy and Anonymization

Nowadays, there is more than ever data coming from different sources, providing an opportunity to various manipulations and analysis. While the data revolution allowed the rapid development of many social and economic aspects, it also resulted in some drawbacks. Privacy is one of those aspects on which information globalization had a negative effect. Lee and Clifton (2011) [38] are emphasizing analysis of datasets that contain private information about users as beneficial for organizations but increasingly problematics for preserving the privacy of users. Hsu et al. (2014) [31] are claiming the hardness of protecting data privacy, they even state that owners of sensitive datasets often unconsciously disclose more information than meant. Various privacy issues such as linkage, data breaches, false identities, and other frauds concern both people and organizations. In order to deal with such a problem, the term Privacy-preserving approaches were invented. The one is Privacy-preserving data analytics, which serves to allow statistical analysis on data while maintaining high privacy [14]. As expected, it is possible to determine a person on characteristics such as home address, age, height, and weight. The simple example of a small street at number 16, where there is only one male elder person can be easily detected by knowing only a couple of characteristics without revealing his real name. According to GDPR Recital 26, anonymization is changing personal data in such a manner that the data subject is not or no longer identifiable [1]. In addition, Rao, Krishna and Kumar (2018) [46] refer to data being publicly available as a threat to individual privacy as the data is in control of a curator.

In 2008, Netflix held its competition to predict films based on user ratings, using a collaborative filter algorithm. The company released the dataset for the competition, containing user records without revealing their real names. With simple anonymization, they changed every user name with ID, thus their privacy was protected. It turned out that replacing a user name with IDs is not a sufficient privacy technique, and such implementation can be easily breached, and data could be exposed. Two researchers from the University of Texas at Austin proved such an issue when they created an algorithm to break the anonymity of the dataset. In addition,

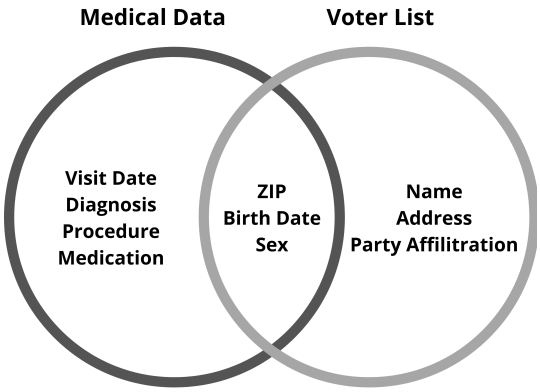


Figure 1: Privacy linkage problem, image inspired by [15] and [34]. By having the same user's data coming from different domains, there is probability of having same private information in both domains, which creates linkage issue.

they cross-correlated the data with users' movie ratings database of IMDb service and found a strong correlation between private Netflix ratings and public IMDb ratings [43]. It means that they managed to find Netflix users by finding the same characteristics of IMDb users and relating them to Netflix accounts. The research proved how unreliable anonymization is when users' names are changed with ID, and the main factor is that users' records can be found in some different source and paired with the original data, which leads to discovering user names and additional information. Since people own accounts on various platforms and applications, it is no surprise that such a data breach could be easily implemented in any domain. It is essential to understand that the Netflix example shows how easily traditional anonymization could be avoided and compromised privacy. Such an example is not an isolated case, there were more cases when privacy was easily compromised. These events triggered the community to reconsider anonymization as a wrong approach and start exploring new options. There was a need to introduce a technique that would do more than changing users names while their information is still exposed, the solution required more effort and a comprehensive mathematical approach. As a result, Dwork et al. in 2006 [18] came up with a new privacy definition, Differential Privacy, which will be explained in the following subsection.

1.2 ϵ-Differential Privacy

Throughout the years, many approaches were presented in order to solve privacy issues, however, there was no universal solution for all problems. While anonymization was successful for data breaches, the linkage issue was still not solved, thus the community was searching for another solution. This

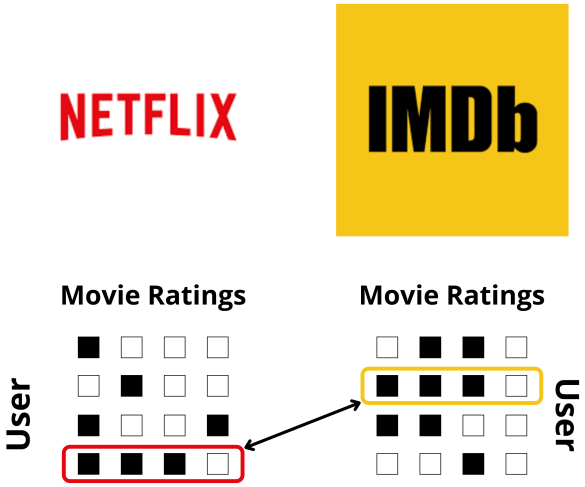


Figure 2: Netflix case in which two movie-based applications produce linkage issue. Image inspired by [15]

was changed after the introduction of differential privacy as a definition to guarantee the privacy of an individual's records. Fung et al. (2010) [23] state that differential privacy belongs to table linkage type of privacy breaches, which occurs when an attacker can deduce if a specific user's records are present or removed from the table. Differential privacy as a mathematical privacy-preserving definition was introduced by Dwork and Roth in 2006 [18]. After publishing the article, Cynthia Dwork continued to further investigate differential privacy, where her work significantly influenced other researches to develop and implement differential privacy. Furthermore, in 2006, Cynthia Dwork published a conference paper 'Differential Privacy' [17] in which for the first time differential privacy was officially suggested as a definition. However, the most influential work by Aaron Roth and Cynthia Dwork from 2014 [19] still is being referenced as a 'Bible of Differential Privacy'. The definition itself assures confidentiality of data for analysis, without learning about individuals while gaining insights about a population of a data [19]. In addition, Dwork and Roth placed differential privacy in the privacy-preserving data analysis domain [19].

What outstands the definition of differential privacy compared to other privacy solutions, is that differential privacy guarantees that individuals' records will not be exposed while summary statistics of the whole population maintain their accuracy. In addition, the definition promises no dif-

ference in population results if an individual decides to share his/her data or he/she decides not to share it. Dwork and Roth (2014) [19] emphasize the introduction of randomness as a vital characteristic of differential privacy, which means that any mechanism that is implemented as a differential private definition has to contain randomness in its process. Due to its reliability as possibly the first universal solution, differential privacy has received large popularity since its invention. This resulted in various mechanisms that implement differential privacy, which proved the importance of such a mathematical definition for various domains. Due to various privacy-aware policies among countries, data such as census should be protected from providing information about individuals.

To explain the definition of ϵ-differential privacy accurately, we will reference Lee and Clifton (2011) [38] who elaborated on clarity on the definition: A database D contains data about individuals, whose private information are stored by each row (A), and requires to be protected. The important component of differential privacy, a randomized function is denoted as M. The function takes a database D and sticks it to a probability distribution over some range, which outputs a vector of randomly chosen real numbers within the range. Lastly, authors [38] are explaining that mechanism is ϵ-differentially private probability of any outcome within a small multiplicative factor is affected only by adding or removing a single datapoint in a database.

Definition 1 (Differential Privacy)**.** *A randomized mechanism M is ϵ-differentially private if for all data sets D and D' differing on at most one element, and all $S \subseteq$ Range(M) that denotes the set of all possible outputs of the mechanism M [38] [9]. Pr presents probability distribution, while exp(ϵ) explains the natural exponent of the epsilon parameter.*

$$Pr[M(D) \in S] \leq exp(\epsilon) \times Pr[M(D^{'}) \in S] \tag{1}$$

1.2.1 Global Vs. Local Differential Privacy

From 2006 when the definition of differential privacy was invented, many researchers decided to investigate a new approach for privacy-preserving. In upcoming years, it resulted in various settings and mechanisms that satisfy the definition of differential privacy. To understand the origin of each mechanism, it is essential to cover two significant aspects that direct the algorithmic invention of differential privacy, global and local approaches. Differential privacy varies between the mechanisms that query statistical summaries and generate synthetic data from an original dataset. Different authors claim various types of differential privacy, primarily because of focusing on the specific domain for which the mechanism would be implemented.

Kairouz, Oh and Viswanath[32] state two contexts of data privacy: the local privacy context, where individuals disclose their personal information, and the global privacy context, where organizations aggregate databases of information of the whole population or answer queries on such databases. Given these definitions, the general distinction for differential privacy mechanisms is between global and local approaches. LDP, shortened to Local Differential Privacy, is an approach that aims to secure the privacy of an individual for data collection [35]. Kim (2018) [35] states that Local DP adds noise to the original data and sends the synthetic data to a data collector, guaranteeing that the contributor's privacy would be protected. With LDP, data is being randomized right before the curator can access it, thus it can be all records in a database randomized at once [8]. On the other hand, Global Differential Privacy (GDP), which can also be found as a centralized DP, has an aggregator that applies carefully adjusted random noise to the actual values returned for a specific query of a dataset [8].

There are advantages and drawbacks to local and global approaches.

1. Global Differential privacy
   (a) Advantages:
      i. Better utility
      ii. Works well with any scale of data
   (b) Disadvantages:
      i. Privacy risk - trusted curator needed
2. Local Differential privacy
   (a) Advantages:
      i. No privacy risk, no data curator to query data
   (b) Disadvantages:
      i. Poor utility - all data being permutated
      ii. Works worse with smaller datasets

For Global DP, the main advantage is that it offers desirable statistical utility while preserving privacy [26], which is not always the case with Local DP. The reason for such advantage of Global DP relies on the noise level injection of statistical queries on a dataset [41]. Local DP implements noise addition to the whole dataset, and it requires a large number of data points to reduce the noise and ensure the statistical accuracy [61]. In terms of privacy, Global DP assumes that the data curator, who collects data, creates a query, and adds the noise to it, must be trusted [60]. However, since such a role deals with private and sensitive information, the risk of privacy exposure is significantly high. On the other hand, Local DP does not face with the data curator until the very end, when he gets a privatized dataset, thus

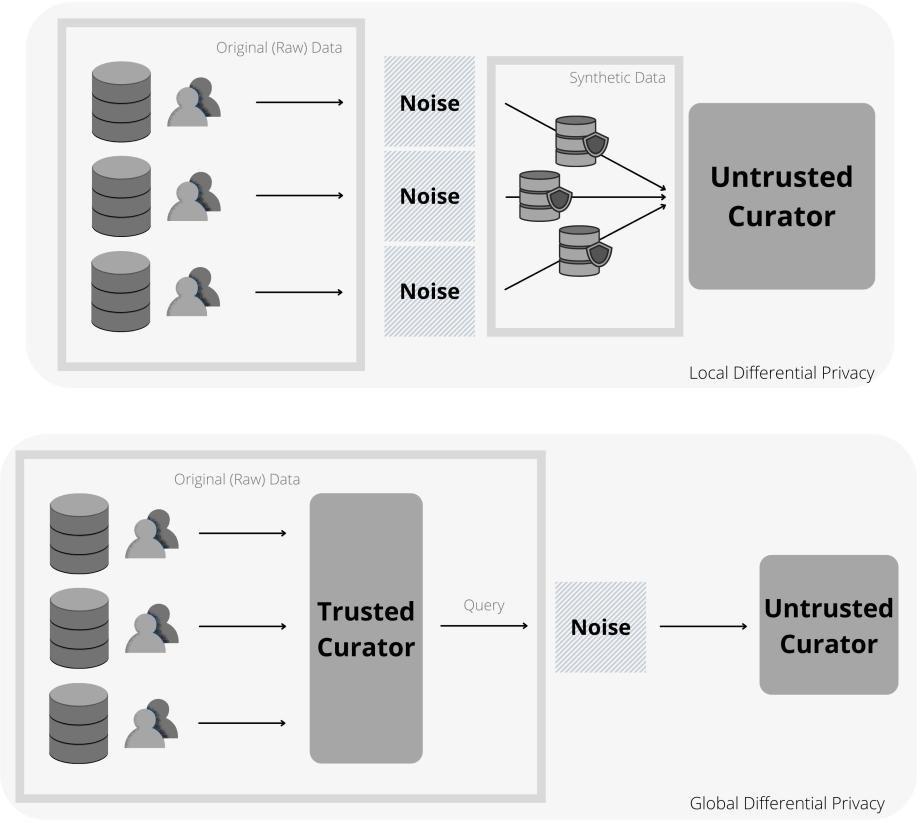


Figure 3: Local Differential Privacy Vs. Global (Central) Differential Privacy

the privacy risk is minimal, which makes a great advantage of the local setting compared to global. Figure 3 compares the two aspects and clearly distinguishes the significant difference in data curator position. In order to understand which approach suits better to a specific project, it is crucial to know whether users can trust its curator, and for what purpose the data would be used. In conclusion, we can infer that Global DP deals better with data utility and requires less noise to be involved, as it queries such a fraction of the whole dataset, while Local DP contains less privacy risk since it does not require a trusted aggregator to create queries and permutate data.

1.3 Data Utility and Privacy Tradeoff

What makes special with ϵ-Differential Privacy is that ϵ symbol that stands for the epsilon parameter, and it directly affects the tradeoff between data

utility and privacy. Differential privacy can be achieved by adding randomized noise into the dataset, however, too much noise could affect the data quality and value of the dataset. The 'amount' of differential privacy is measured by epsilon as a parameter, which controls how significant influence the noise should have on the specific dataset. Lee and Clifton (2011) [38] explain the epsilon parameter as "a difference between two probabilities of receiving the same outcome on two different database", but also degree of privacy that is introduced. The epsilon parameter allows us to control the level of privacy by having lower values that guarantee stronger privacy, being smaller than the value of 1 [22]. By deciding the amount of the epsilon parameter within the specific differential privacy mechanism, we can manipulate the tradeoff between data privacy and utility. If a significant amount of noise is allowed, the specific data would be privacy-protected. However, its value would be lost entirely, which results in pointless analytics and meaningless information. On the other hand, restricting privacy to allow more data utility would allow better data accuracy, but privacy would be questioned. To give an example, we could implement differential privacy on census data for a specific city. If we introduce a significant amount of noise, we would be sure that the personal data is protected and nobody could leakage any information. However, when analyzing data from the perspective of the population, if privacy is high, it could happen that results of the analysis would be wrong because the data accuracy was harmed with the introduction of too much noise. This means that maybe some graph that presents ratio between sexes would not present a real ratio but rather a noised version that does not accurately presents the results, which means that we got useless value of the private dataset. The figure 4 for which Lee and Lee (2018) [39] state the importance of finding the level with the maximum data utility without exceeding risk threshold. The figure 5 presents the tradeoff between data privacy and utility.

In 2014, Hsu et al. [31] emphasized that there is no precise method for adjusting the epsilon parameter. In addition, Lee and Clifton [38] in 2011 stated that parameters of differential privacy that affect the risk of disclosure in practice have not yet being studied. Since each dataset contains specific characteristics, it requires a specific approach for the tradeoff, thus it is almost impossible to find a universal solution. In addition, there is no adequate visualization approach that allows the adjustment of the epsilon parameter with instant visualization results that allows users to understand how the tradeoff works, and what are the possibilities to adjust it. Hsu et al. (2014) [31] emphasize that it is difficult to choose epsilon precisely because there are two parties, analyst and users, who have different opinions about data being privately preserved - while the analyst wants accurate data which requires high data utility, users want to protect their records, thus they require high privacy. Figure 5, inspired by [48] presents the tradeoff graph between data utility on the x-axis and privacy protection on the y-

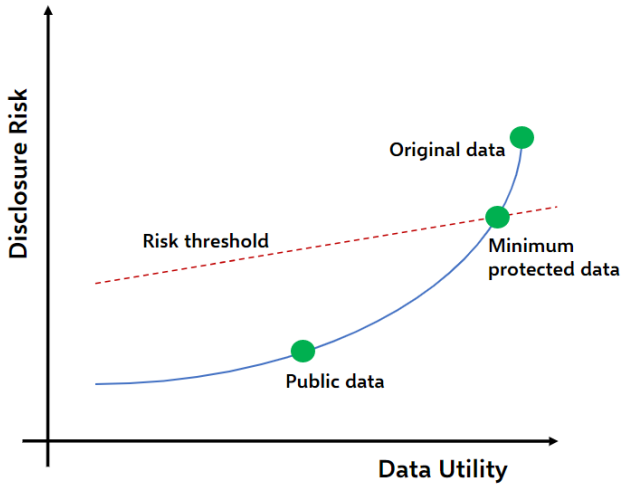


Figure 4: Data Utility and Disclosure Risk, presenting at what distances is original data from public data that has to be preserved. Inspired by Lee and Lee (2018) [39].

axis. At the maximal privacy stage, the utility is reaching the lowest value, and such a situation becomes completely different when the utility is at its maximum level, while the privacy is reaching its minimum. The perfect situation is presented with a dashed line, where we have at the same time maximum privacy and utility, which is almost impossible. However, a more realistic situation lies in the middle of the blue line, where there is enough high amount of privacy that does not affect utility hardly, meaning that data is still accurate enough. The tradeoff is vital in terms of comparison between global and local differential privacy. As previously explained, local differential privacy affects a whole dataset, which directly affects data utility. Given this statement, the goal is to find the best ratio between privacy and utility that would provide individuals' decent security while still providing accurate data. Liu et al. (2020) [40] are emphasizing that "a major challenge is addressing the tradeoff between privacy and utility". We can infer that such a challenge is even more problematic in terms of visual analysis, thus our goal is to minimize the lack of understanding about the tradeoff by visualizing it.

1.4 Data Visualization

Data Visualization plays a vital role in data analysis and any form of engagement with users that are not experienced with technology. Creating graphs, plots, and dashboards out of a dataset helps to understand the meaning, advantages and drawbacks and create insights out of data. Dasgupta and Kosara (2011) [12] state enabling to assemble visualization results for users as crucial goal of data visualization, which has to be followed by data that

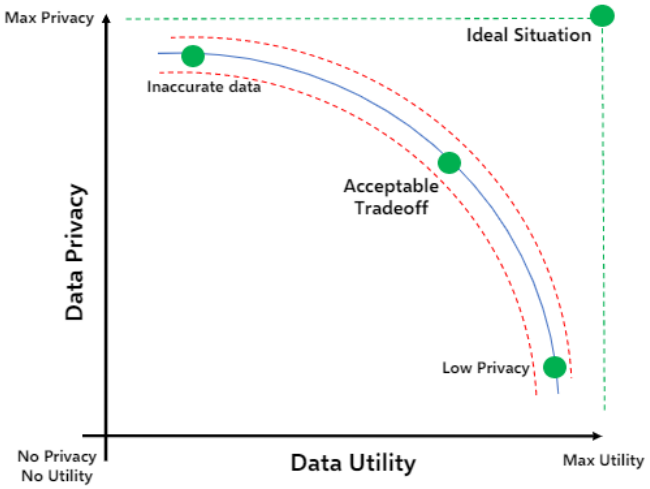


Figure 5: Data Utility and Privacy Tradeoff, which presents that when privacy level is high, the utility value is low, and opposite. Inspired by Singh (2019) [48].

maximizes its utility. In the same way, providing an interaction can contribute to users engagement and make their experience with specific data more interesting. It is known that two same statistical summaries could end up with different visualization insights, thus analysts always rely on the visual aspect of their work. These purposes of Data Visualizations are even more important for privacy-preserving technologies, for which data visualization opens an opportunity to 'see' how a private a dataset is, or how accurate a query could be. Dasgupta, el al. (2014) [14] explained two important advantages of visualization interface for privacy-preserving paradigm: first, the utility could increase due to interaction, and second, tuning visual parameters allows more flexibility to those that use the interface. In the following subsections, we will focus on two aspects that each contributes to the thesis in a specific way.

1.4.1 Privacy-Preserving Data Visualization

There are many domains in which data visualization comes as an important practice to analyze data. However, most of these domains are exposed to privacy risk and they are in need of privatizing both data and visualizations. Dasgupta, Kosara and Chen (2019) [13] state the most important confrontation for publicly accessible visualizations that could disclose personal data is minimization of disclosure risks. To overcome privacy risks, the community decided to expand privacy-preserving techniques towards data visualization and came up with a new approach called Privacy-Preserving Data Visualization, shortly PPDV. Within the PPDV, there are different purposes and strategies on which PPDV relies on. The approach emerged as the combi-

nation of Privacy-Preserving Data Mining (PPDM) and Data Visualization, and we can see it as a necessary extension of PPDM. Some may ask 'why necessary extension', however, it is obvious that analysis of data requires visualizations for a clear understanding of potential insights. In terms of PPDV, firstly there are private visualizations, that are a product of earlier sanitized datasets or query, for which we need visual analytics. Avraam et al. (2021) defined Privacy-preserving visualizations as "graphs that control the disclosure risk by applying permutations to a raw dataset" [4]. The second approach, and the more meaningful one is when we privatize a visualization that presents raw data. This means that we recreate a new visualization that now preserves the privacy of analytics that were implemented on a specific dataset. It is important to understand that our approach is not entirely based on PPDV. This is because PPDV is primarily focusing on preserving the privacy of visualization, while our aim is on visual analytics of preserved data, thus the visualizations that we create will not make any addition privatization that was already implemented with Differential Privacy. Given this statement, in the following subsection we will present another aspect of Data Visualization that is implemented in the project.

1.4.2 Visual Parameter Space

The second aspect of the thesis is focused on the visual parameter space approach. The concept concentrates on understanding input parameters and investigating input and output parameter settings, emphasizing the visualization interaction. Sedlmair et al. (2014) [47] define three classes of input parameters:

1. Control parameters
2. Environmental parameters
3. Model parameters

Our setting belongs to class control parameters, where users directly manipulate it. This relates to the research because users adjust the epsilon parameter (defined as an input parameter), and its effect is visually presented. The second perspective of the visual parameter space approach is sensitivity, which presents variations of outputs to expect to changes of the input [47]. The questions that are formed from sensitivity are what ranges/variations of outputs to expect with changes of input[47]? Because adjusting the epsilon parameter effects data privacy and utility, sensitivity is characteristic of the input parameter. In addition, the adjustment influences the variation of outputs, meaning that the input of epsilon value 1 will give different results than epsilon value 5.

The figure 6 presents the perception of the visual parameter space that is depicted in the project. First, two input objects are visible, a dataset as

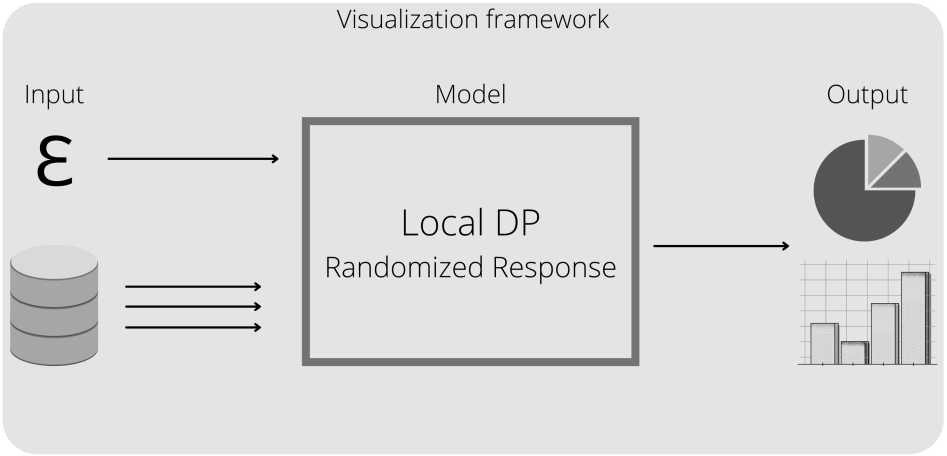


Figure 6: Visual Parameter Space Pipeline, presenting which inputs (epsilon parameter and a dataset) are imported to the system, which outputs visualizations and a synthetic dataset.

a static object that is affected by the second input object, parameter. This parameter in Model stage affects on a dataset by applying Local Differential Privacy with Randomized Response mechanism. In addition, the epsilon parameter can be adjusted more than once and at anytime, which makes it a dynamic object. As an output, an interactive framework that consists of visualizations that adjust according to data and the epsilon parameter emerges. The whole process is visually presented in a framework that allows users to input data, adjust the epsilon parameter and see visualizations that present results of private data with utility and privacy metrics.

1.5 Thesis Outline

Starting with the Introduction, the whole project is presented in the general topics covered: Data Privacy, Differential Privacy, the Tradeoff between Data Privacy and Utility, and Data Visualization. By giving the starting point of what are the main points for each domain, the details regarding Local Differential Privacy and Visual Analytics are elaborated. After introducing topics that merge in the thesis, the second section, Related Work, explains what the situation in the academic community is. By splitting the section into two main topics, Local Differential Privacy and Privacy-Preserving Data Visualization, both domains are covered and provide adequate knowledge of what is accomplished and potential issues and opportunities for these fields. Lastly, Research Gap and Contribution Statement section provides reflections on where the research finds its opportunity to

impact and contribute to society and the academic community. Moving to the third section, Research Problem contains the Problem Statement and Research Questions that are given based on knowledge from the previous section. The fifth section, Research Method, extends what was stated in the Research Problem by dividing the thesis into four levels. Each level presenting a logical component of the project's increment is being investigated for its threats and validated to prove its necessity and feasibility. The methodology includes an evaluation method for each level of research, and its results are presented in the sixth section, Evaluation Results. Based on the results from the previous section, there are given reasoning and main points of results and findings in the seventh section, Discussion. The conclusions are classified into general and ones from evaluations of each level of the research. In addition, Conclusion and Future Work as the last section offer a viewpoint on what are possible further developments for the project, giving ideas on what could be improved to achieve full functionality of the project.

2 Related work

In this section, the focus is on finding the proper literature closely related to the topic of the thesis. By listing and providing information regarding related work, we are explaining how the literature search was defined and conducted. Next, Related Work helps readers understand what helped us progress in the project and which papers were inspirational for the thesis. The primary technique used for finding related work was snowballing, in which the researched find additional literature from specific key papers. The reason for choosing such a technique was that the thesis topic is complexed, and it refers to multiple disciplines. Thus, we were aware of the necessity to focus on the minority of papers that provide at least a significant fraction of the relevant information for the thesis. We will emphasize those papers that were starting point for snowballing technique and finding more relevant literature. Next, for the purpose of the Related Work section, it was decided to use these keywords: *Differential Privacy, Randomized Response, Local Differential Privacy, Local Differential Privacy Visualization, Differential Privacy Visualization.* It is essential to understand that these keywords were mostly used to find critical papers that led us to more literature.

In terms of Related Work that contributed to this project, it was decided to focus on two different search perspectives. These aspects resulted from findings from the research gap, where the main conclusion was the shortage of work that combines adjusting the epsilon parameter for Local Differential privacy from a visualization perspective. Hence, it was decided to take the 'divide and concur' approach by splitting the search into three logical perspectives, each equally contributing to the thesis and intersecting with another approach in some parts. For the first search, the focus is on Local Differential Privacy and Randomized Response. This means that the literature search is based on those papers on Local Differential Privacy that concern Randomized Response as an implemented mechanism. It was a logical decision mainly because the thesis focuses on Local Differential Privacy as a Differential Privacy segment, and Randomized Response as the primary mechanism used. However, because the algorithm created in the project is not priority, and it was created as an alternative and simple solution inspired by local differential privacy and randomized response, there is no need to prioritize this aspect. The second aspect is concerned with a higher-level perspective on the thesis, focusing instead on Data Visualization with Privacy-preserving techniques, then taking Differential Privacy with Data Visualization. Such a decision lies in the fact that there is no relevant literature that considers Local Differential Privacy with Data Visualization. Moreover, Differential Privacy belongs to Privacy-preserving techniques, hence our thoughts were to take a higher perspective and find relevant literature that focuses primarily on visualizing privacy-preserving techniques, hoping that there will be some that consider Differential Pri-

vacy as well. Because merging data visualization with differential privacy is a primary focus of the research, this approach clearly counts as an important stage of the related work.

After presenting related work that influenced our project, we will discuss the research gap. This generally includes the explanation of what was not been discussed and implemented in the existing works, and what could be done in our work. We will recap all issues that existing papers are having and explain the strength of missing problems and further discussions. Lastly, by providing a subsection regrading the contributions, we will explain how does this research promise new features and knowledge to the academic community. In addition, we will present the contribution from practical, theoretical and society sides.

2.1 Local Differential Privacy and Randomized Response

The initial research that introduced the concept of Local Differential Privacy was made by Kasiviswanathan, et al. in 2011 [33]. Next to Local Differential Privacy as the main contribution, the authors are also providing knowledge regarding the difference between interactive and non-interactive local learning algorithms. Several papers ([52], [33], [26], [6]) emphasize the growth in academic interest for Local Differential Privacy. However, to get profound knowledge about Local Differential Privacy and its implementations, it was decided to search for comprehensive surveys. As a result, two papers emerged and became key papers in the further snowballing literature search. The first paper by Xiong et al. from 2020 [60] covers all aspects of Local Differential privacy in general, from fundamental definitions, comparison with Global Differential privacy and mechanisms to research implementations of LDP and the limitations that are facing. What it is important to conclude from the paper is that LDP is yet to be discovered and improved with its growth in recent popularity, and the fact that there are various implementations in terms of data analysis tasks. The second comprehensive study is by Yang et al. from 2020 [61] as well. What stands as important information from an earlier comprehensive survey is clear definitions of two Randomized Response mechanisms, Randomized Response for a binary attribute and Generalized Randomized Response for a larger domain. In addition, what stands out with this survey are the research directions that authors are proposing as possible topics for researchers that are interesting in working on LDP. What we can infer from both the comprehensive surveys is that Local Differential Privacy is not as popular and researched field as Global Differential Privacy, but in recent years its popularity is growing, thus we can expect more works contributing to the field. In addition, both surveys emphasize RAPPOR by Google as a state-of-the-art mechanism for Local Differential Privacy that is based on the Randomized Response approach.

In terms of related work that is concentrated on the actual implemen-

tation of Local Differential Privacy, the search was focused on those mechanisms that include Randomized Response. Alongside Apple [2] and Microsoft [16] implementation of Local Differential Privacy on real-world data, Google invented RAPPORR [21], standing for Randomized Aggregatable Privacy-Preserving Ordinal Response created in 2014. Not only that the work is highly recognized by the community as an important technical implementation of the LDP mechanism, but also they used the same technique as this research, Randomized Response, as a starting point to develop their variation of the algorithm. Bebensee (2019) [6] is among a couple of papers ([61], [9]) that cover LDP from a theoretical perspective. Due to various domains, there are many different mechanisms for LDP, thus we focus on papers that use Randomized Response. Despite this technique is still minority in the LDP community, there are several papers that use directly Randomized Response technique, or algorithms were created from it, such as Gursoy et al. (2019) [26], Wang, Wu and Hu (2016) [56], Wu, Peng and Niu (2020) [59], Holohan et al.[30], Arcolezi, Leith and Mason (2016) [3], Kairouz et al.[32] and Kim, Oh and Viswanath (2014) [36], or those papers that compared their algorithm with Randomized Response as Mansbridge et al. (2020) [41]. Similary as in our research, Want et al. (2019) [52] focus their work on implementing LDP algorithm on numeric data.

2.2 Privacy-Preserving Data Visualization

Due to the fact that there is no sufficient evidence of papers that directly relate to Local Differential Privacy and Data Visualization, it was decided to take a higher-level approach and later drill down to as precise information as possible. In other words, the strategy was based on starting the search by looking at Privacy-Preserving Data Visualization papers and then finding those ones that based their approach on Differential Privacy. There is no work that is focusing on both Local Differential Privacy and Data Visualization, so it is logical to accept the general concept of Differential Privacy as a focus for the second aspect of related work. As in the first aspect, it was decided to start with finding a comprehensive survey of Privacy-Preserving Data Visualization that would offer us implementations in many techniques, and therefore we managed to find specific papers that were focused on Differential Privacy. The paper that was a key comprehensive survey is by Bhattacharjee, Chen and Dasgupta in 2020 [7]. It is important to mention the data flow of Privacy-Preserving Data Visualization that explains how are roles engaged in the process of collecting, anonymizing and visualizing the outcome, risks and uncertainty. Another advantage of the paper is the extensive list of implementations in the field of Privacy-Preserving Data Visualization, which led us towards projects that were focused on Differential Privacy. Now, we will present three papers that all relate to the Privacy-Preserving Data Visualization domain, and directly relate to the thesis in

specific aspects. For example, while one paper focuses on the visual analysis of data privacy and utility tradeoff, others create a pipeline that provides a visualization interface for creating private visualizations on any data.

The work by Zhang, Sarvghad, and Miklau from 2020 [65] is by far the most significant paper that relates to our work. It is important to emphasize that this paper was the first to discuss how does noise injection level effect visual analytics. The authors decided to investigate tuning noise injection for improving accuracy for visual tasks and the utility of private visualizations [65]. In terms of the second research question related to adjusting the noise level for private visualizations, the results show that "it remains unclear which algorithms or noise injection mechanism will better facilitate downstream visual analysis" [65]. Again, this gives us an opportunity to discover findings that would give the precise answer to adjusting the epsilon parameter and influencing data utility and privacy tradeoffs. Another finding of the paper that was important for our case is that under Differential Privacy, the pie chart and line chart provide better accuracy than the bar chart [65]. Despite having a similar contribution, it is important to understand that this research does not focus on Local Differential privacy where the whole dataset is sanitized. Instead, they focused on Global Differential Privacy and privatizing queries, which results in working with different mechanisms and gaining different results from what we could expect with our project. Next, half of their work focused on visual tasks, which does not correlate with our work. In addition, despite the fact that the paper does include the adjustment of noise, this is not manipulated by users, which also differs with the work of our project. Lastly, while they compared multiple mechanisms for Global Differential Privacy, we decided to implement our own algorithm, allowing only that mechanism to be implemented in the visualization framework.

The second work by Wang et al. from 2017 [55] relates to our project in particular aspects. First, they focus on providing users with knowledge about data utility while privacy is guaranteed [55], which relates to our goal of bringing an understanding of data privacy and utility tradeoff to users. Second, they aim to succeed by providing greater flexibility and transparency for visualizing the influence between data privacy and utility [55], which also relates to our goals of visual analytics. As a result, they created a visualization interface that works as a pipeline that allows data manipulation, measuring utility loss and handling the privacy of the data [55]. Comparing to our work, this means that both projects allow any data that is chosen by users, and those specific metrics for quantifying utility and privacy are presented in the visualization interface. In addition, users can export private data, which creates a full pipeline. What differs is that users are not able to adjust the privacy level but rather change the aggregation of attributes that do not have anything with any privacy technique. Another option is to only apply differential privacy, however, without any additional adjust-

ments of the epsilon parameter. This proves that Differential Privacy is only used as a side mechanism and it is not presented as a priority technique for privacy-preserving data visualizing.

The third paper is by Wang et al. from 2018 [54]. Despite the main difference that their work does not involve Local Differential Privacy as a Privacy-Preserving technique, there are valid reasons to include such paper as a relevant work. The most important reason is that the work presents a visualization interface that serves users by guidance through a privacy preservation pipeline [54]. The visualization interface allows data import, however, the process does not end with data being exported but rather processing the report about preserved data. In addition, users are allowed to choose between multiple privacy-preserving techniques, but they are not able to adjust the level of each privacy. Thus, the main limitation of the paper is that providing different mechanisms does not guarantee that they have a proper amount of noise for specific data. In addition, the perception of data visualization in this work is by providing an assessment of the privacy risks that specific data could be faced, rather than creating private visualizations.

All three closely related papers do not consider Local Differential Privacy as a privacy-preserving technique, nor Randomized Response as a mechanism for data anonymization. On the other hand, all three works are producing visual analytics results based on privacy-preserving techniques. From the perspective of Visual Analytics, there are many similarities with each work, however, not all of them contribute in terms of adjusting the level of privacy mechanism, or their output does not cover the same expectations.

2.3 Research Gap

In this subsection, we will provide a conclusion on related work that was explained earlier, and draw a line between what is missing in the community for the topic. When referring to differential privacy, it is essential to understand that this is a general definition that contains a couple of approaches, which also contain various mechanisms implemented for a specific domain. Given that statement, it is hard to find many papers that consider both the same approach and mechanism as what we discuss in this work. In addition, the authors [52] claim the crucial problem of gathering numeric data has not been addressed sufficiently yet, thus the research focuses on implementing a Local Differential Privacy mechanism on users' data.

Zhang, Sarvghad and Miklau [65] state that in recent years, differentially private algorithm study has been a topic of deep research efforts. On the other hand, the authors are claiming insignificant attention towards the task of presenting or exploring data through visualization, for the purpose of data privacy, despite its importance for users trying to gain insights from data [65]. In addition, Dasgupta et al. (2014) [14] emphasize that privacy-preserving data visualization compared to privacy-preserving data mining

is still in its initial stages. As previously explained, there are barely three papers that concern differential privacy from a data visualization perspective ([65], [55], [64]). In addition, those papers do not focus on the local model, they implement entirely different mechanisms, and only one work discuss the tradeoff between data privacy and utility. Here, we see a great opportunity to provide a visualization approach to measuring the tradeoff between data utility and privacy.

While there are papers that are focusing on privacy-preserving data visualization and data analysis that include differential privacy, their output, the visualization interface aims to private visualization insights. On the other hand, our goal for creating a visual analytics interface is to bring the notion of differential privacy towards users and help them understand the meaning of adjusting the amount of noise that affects the data utility and privacy tradeoff. Thus, not only that our visual interactive framework differs in its design, but also in purpose. In addition, there is no paper that covers privacy-preserving data visualizations from the perspective of helping users to understand how privacy-preserving techniques contribute to securing their private data. Dasgupta et al. (2014) [14] are concluding their work by emphasizing the importance of future researches to investigate the relationship of anonymization techniques with turning visualization-specific parameter for controlling the privacy of visual representations. In addition, they see an opportunity in understanding optimal privacy and high utility for privacy-preserving with interactive visualizations [14]. This directly correlates with our scope of research to explain the tradeoff between data privacy and the utility of a privacy-preserving technique using interactive visualizations.

Due to the strong influence of noise on the whole dataset that results in lower accuracy than the global approach, Local Differential Privacy is still less investigated than Global Differential privacy [61]. Moreover, because of LDP's particular limitations, many fundamental questions that are well studied in GDP have not been completely understood in the LDP[51]. However, there is a growth in the popularity of theoretical understanding LDP[10] in recent years. In addition, many state-of-the-art LDP mechanisms have been developed to provide privacy-preserving[59], and for the process of data collection, LDP has been widely applied and accepted[3]. However, using LDP with different purposes for specific domains makes two papers based on the same approach way different from expected. Moreover, none of the papers that focus on LDP approaches do not consider data visualization a priority, only as visual proof of results. Thus, the only highly relevant works from the academic community related to the research are ones that implement the Randomize Response mechanism on the collected users' data. Not only that the specific topic is still not well presented, but there are no approaches that relate LDP with data visualization.

Regarding the tradeoff, some papers cover the topic, however almost all

focus on privacy-preserving algorithms in general, while only a few discuss differential privacy. Despite the existence of papers that focus on differential privacy, or focus on the tradeoff between data utility and privacy for privacy-preserving algorithms, or focus on visualization approaches on privacy-preserving algorithms, no work combines differential privacy and the epsilon parameter to investigate the tradeoff between data utility and privacy from the visualization approach. In addition, no work specifies data visualization for differential privacy and offers any solution. Here, we again see our opportunity to present a slider as a visual adjuster for a visualization framework that works with collected data and differential privacy mechanism. In addition, visual parameter adjusting belongs to the visual parameter space domain, for which no work combines the topic with differential privacy in any sense, thus the research will be the first of that kind.

In conclusion, no work focuses on both LDP and the tradeoff between data utility and privacy. In addition, no work provides a tradeoff from a visualization perspective. Overall, no paper discusses visualization storytelling as a perspective on adjusting and evaluating the epsilon parameter in the differential privacy algorithm on a specific dataset. Our contribution to the academic community lies in visually explaining the epsilon parameter adjustment for LDP with the Randomized Response mechanism.

2.4 Contribution Statement

It is important to distinguish what are the contributions that the project provides to the community. Here, we take two aspects: practical, for which we analyze how does the work provide more technical knowledge that helps both users and the academic community, and social, where we emphasize on the theoretical contribution that helps users to understand specific topics. In our case, this would mean that the visualization framework with its privacy-preserving mechanism would be practical contribution. Moreover, we produce a solution that directly connects two topics, Differential Privacy and Visual Parameter Space into one high-level topic, Privacy-Preserving Data Visualization. We are also implementing our alternative and simple version of algorithm inspired by local differential privacy and randomized response. In addition, we are providing specific data utility and privacy metrics to measure its performance. By delivering a practical contribution as the interface that allows adjustment of the epsilon parameter and instant results of it, the users are gaining knowledge of how to properly adjust the noise, and what are the results of adjustments. Users would gain knowledge of the tradeoff between data privacy and utility, and see what amount of the epsilon parameter is sufficient for the best ratio between valid data and private records.

In addition, users are not experts in mathematics and understand how

the privacy mechanism works. By creating a visualization interface that allows them to adjust noise and instantly see the results, they will get understanding of what influence does a specific value of the epsilon parameter has on the data. Lastly, since the focus is on implementing an algorithm on the users' data, they would get a clear insight into how their data would be private. This is a direct social contribution since the users are acquiring understating about both privacy-preserving mechanism, and adjustment of the noise with instant effect on visualizations. In addition, it was decided to create such visualization framework that allows any quantitative data, thus it makes the contribution more important for data independent solution. This was extended to end-to-end pipeline that comprehends the input of data, adjustment of the epsilon parameter for instant visualization analysis, and possibility to export private data. As the result, the visual analytics framework contributes to help users to understand the notion of differential privacy and how does adjusting the epsilon parameter affects on data privacy and utility from the aspect of the tradeoff.

We can distinguish a couple of important contributions that the research gives to the academic community:

1. Help users to understand the meaning of Differential Privacy and the tradeoff between data privacy and utility
2. Combining Local DP and visualizations from the tradeoff perspective
3. Merge of Differential Privacy and Visual Analytics to create a privacy-preserving visual analytics approach
4. Creating visualization end-to-end interface that inputs data, permutates data (with possible adjustments of noise level), and outputs differentially private data

In conclusion, the impact that the project is making focuses to create a visualization system for local differential private synthetic data to allow adjustments of the epsilon parameter with instant effect on the visualizations. Not only that we are pioneers of implementing such a project that combines two perspectives, differential privacy and data visualization, but we provide end-to-end pipeline that is undependable of any data. This means that we allow the input of any quantitative data, and after the privacy mechanism is implemented on input data, it is possible to export private data. By using proactive programming, we enabled instant presenting of results after adjusting the epsilon parameter on a specific data. The research also includes additional fields that are part of specific high-level domains. We included both privacy-preserving data analysis and privacy-preserving data visualization, both focusing on privacy techniques but having completely different perspectives. In addition, we add visual parameter space as visual analytics perspective that brings the interaction and possibility of adjusting the

noise introduced by privacy mechanism, local differential privacy and randomized response. The impact becomes even more vital when the research adds multiple perspectives and combines various domains into one project.

3 Research Problem

In this section, we will continue with what was investigated in the earlier section Related Work, and draw the problems that were identified in the Research Gap. These issues will be presented in the Problem Statement, after which we will define research questions that are focus of the thesis. In order to answer all Research Questions, the Research Method will be presented and in detail explained.

3.1 Problem Statement

As it was explained in previous section Research Gap, there are constraints in the current literature that open an opportunity for contributing in combining differential privacy and data visualization, where even some papers are emphasizing the important of further investigation. Despite the growth in interest for privacy-preserving solutions and the fact that nowadays, there are many algorithms and approaches, users are still not familiar with privacy and its effect on their lives. To bring such a topic towards regular users who are not experts for data privacy and GDPR, the visualization interface that helps them understand is mandatory. One question led us throughout the problem statement:

1. Is it possible to use data visualizations for explaining differential privacy and to create privacy-preserving visual analytics that would open the understanding of the data privacy and utility tradeoff to users?

The given questions led to more problem statements that concern the thesis: Can local differential privacy be adjusted? What is the perfect solution to visually present that privacy can be adjusted? What are the metrics that show users that the mechanism is working correctly? How to help users to understand the tradeoff between data privacy and utility? These questions primarily arise with algorithms such as differential privacy, where the tradeoff between data utility and privacy plays a vital role in data quality. In addition, the epsilon parameter was invented so that it could adjust the tradeoff and allow users to find a ratio that fits them perfectly. Lastly, by creating a simple and alternative mechanism that is inspired by Local Differential Privacy and Randomized Response mechanism, we are introducing users to one of the many solutions for Privacy-Preserving Data Analysis. However, how the algorithm stands against the state-of-the-art mechanism?

3.2 Research Questions and Subquestions

Below, we present two research questions, followed by subquestions. Each research question focuses on a specific part of the thesis. First research

question is concerned with the creation of a visualization system that focuses on adjusting the epsilon parameter for a specific dataset. Here, we are investigating how to visually present and provide the knowledge and adjustment of noise injection level which could be seen in the results instantly. In the second research questions, asks how could such visualization help for the tradeoff between data privacy and utility in case when there is users' private data. In addition, to evaluate the performance of the data privacy and utility tradeoff, we emphasize the matter of providing metrics as crucial subquestion.

1. RQ1: How to visually present the tradeoff between data privacy and utility?
    (a) SQ1: How to visually provide epsilon parameter for users?
    (b) SQ2: How to include adjustment of epsilon parameter on visualizations?
    (c) SQ3: How to allow users to adjust tradeoff and see results?
2. RQ2: How can visualization system help users' data with the tradeoff between data privacy and utility?
    (a) SQ1: Is it possible to tune noise-injection to improve the utility of a private dataset?
    (b) SQ2: What are the metrics to evaluate the performance of data privacy and utility tradeoff?

4 Research Method

In this section, the whole procedure of conducting the research method will be introduced and explained. Alongside the methodology, the threats and validations will be elaborated. To research methods fit the project domain, it was decided to investigate the methodologies concentrated for visualization research. In addition, it is important to find a method that works with researches that incorporate different approaches and domains. This led to the discovery of the Nested model, a methodology designed by Tamara Munzner in 2009, consisting of four layers. The main point of the method is that the output from a higher level is input for the lower level [42] of the model, thus the levels are nested. By splitting the visualization design into four levels containing specific threats and validations, the nested model expects each threat and validation to influence its lower level. The main advantage of such a research method is dividing research into four logical levels of designing visualization depending on each other and evaluating each segment of the project equally. However, the drawback is that the decision of a higher level affects lower levels, thus poor decisions on validation could lead to destructed methodology and unsatisfactory results. The same issue remains for threats, if a poor choice was made in the high-level abstraction stage, no matter how well designed are the lower levels, the research problem will not be solved. Since the research contains two perspectives, differential privacy and data visualization (we can also see this as privacy-preserving data analysis and data visualization), the nested methods allow us to divide both approaches into these four layers while containing their relationship towards evaluation of the project. Figure 7 presents the Nested Model and how the levels are related to each other.

The main reason why such a method fits is because of the layers that allow splitting the research into sections that variously contribute to the project. The whole point of the thesis is combining two different perspectives (privacy algorithm and visualization interface), thus the nested model allows diving approaches into logical levels. In addition, the nested model allows devoting specific evaluation methods for each level, allowing to provide a proper evaluation strategy to validate the whole visualization system from every needed perspective. In conclusion, not only does the research method allows us to deal with each threat properly, but it provides a solution to deal with the evaluation from each level. In conclusion, not only does the research method acknowledges us to deal with each threat properly, but it provides a solution to deal with the evaluation from each level. For each of the four layers, the explanation of its meaning will be presented, alongside threats and validations that are related to the project.

In upcoming sections, we will present each level, explain its purpose and value, and describe how it contributed to the project. In addition, for each level we will introduce threat and validation, which will be followed by

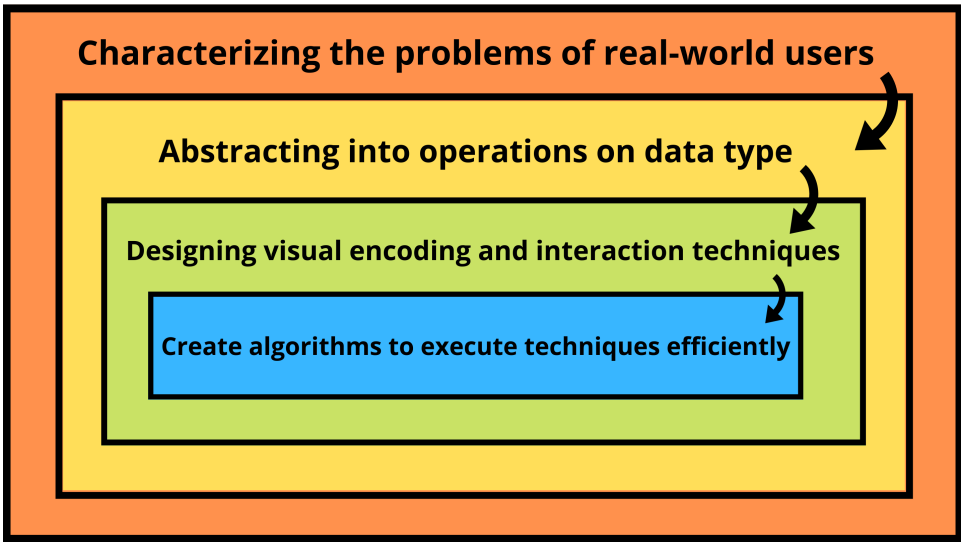


Figure 7: Nested Model by Munzner [42], presenting four nested layers that contains its characteristics to contribute the research.

specific research evaluation methods. Figure 8 depicts what are threats and validation within each level, and how they affect each other.

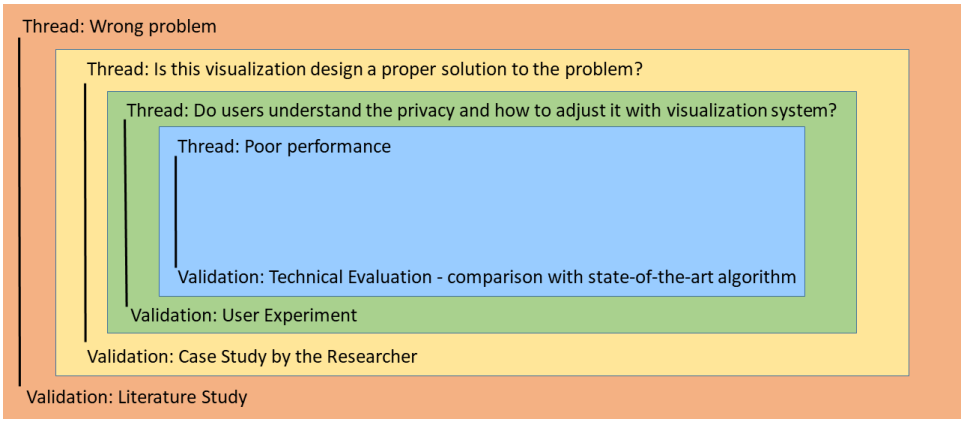


Figure 8: Nested Model - inspired by Munzner [42]. Presenting threats and validation of each nested level of the research.

4.1 Domain Problem Characterization

The first layer, which is the highest level of the whole model, is domain problem characterization. As the name refers, the goal is to define a problem and a target audience involved in the situation. In our case, the target audience are differential privacy experts, data analysts, and practitioners of privacy-preserving data analytics techniques. Including experts in the

project is crucial since they are the ones who are interested in relating visual analytics to differential privacy, and they are the ones who would profit from such a visualization system. In addition, target audience experts will interact with the solution, which is the visualization system.

In terms of the problem definition, the situation with possible domains will be explained. Differential privacy as a mathematical definition of privacy protection was invented earlier and gained significant popularity and various implementations in recent years. In addition, local differential privacy was invented, and it offers many mechanisms to protect data privacy. However, there is still no evidence of crucial implementations of visual analytics towards differential privacy to understand how the mechanism contributes to the problem of improper disclosure of users' data. Thus, it is clear that the main problem discussed in the thesis is how to relate differential privacy as a privacy-preserving technique to visual analytics? In addition, our focus is on the tradeoff between data privacy and utility, which plays a crucial role in the performance of any differential privacy mechanism. However, when arrowing such a situation event to the tradeoff, we can infer that there are no attempts to discuss such matter and find a solution. Thus, our problem is expanded with the tradeoff, and it is: What is the relationship between differential privacy and the visual analytics interface that allows noise adjustment? Since the tradeoff is included, we are asking ourselves how to introduce an interactive visual adjustment of the tradeoff between data privacy and utility. Such problems open additional questions on proving the algorithm's performance and understanding the mechanism, metrics, and tradeoff.

Contribution After presenting the problem and its background, it is essential to introduce the solution and its contribution to the problem. For such an issue, the solution is to combine differential privacy with visual analytics in order to create a visualization that allows users to adjust the privacy parameter and helps to understand how privacy and algorithm work together. The visualization system as an interface has to interactively allow users to adjust the epsilon parameter of the differential privacy and instantly present results on graphs. Such a solution contributes to the problem so that it provides a transparent visual analysis of differential privacy mechanisms on specific users' data. In addition, it allows experts to analyze what amount of the epsilon parameter is sufficient to protect individuals' information while remaining the value of a dataset. In addition, the layer directly contributes to the second research question of supporting users' data with the tradeoff between data privacy and utility.

4.1.1 Threats

The threat for the first stage would be defining the wrong research problem if it concerns a target audience. As previously explained, the target audience are the ones engaged in differential privacy no matter on the level of their understanding of privacy-preserving techniques should profit from such systems. Thus, these are differential privacy experts, data analysts that use the differential privacy algorithm and practitioners whose private data is being used. However, because of their different level of knowledge about differential privacy and different roles within the process, there is no guarantee that all three groups of the target audience have the same interest in the solution. This would mostly depend on the complexity of both the visualization system and the algorithm that is implemented. With the more complexted visualizations that are focused on compound algoritms, the experts would gain more than practitioners who do not understand such mechanisms. On the contrary, simple differential privacy algorithms with understanding visualizations would gain more understanding for practitioners. The problem is how differential privacy relates to visual analytics and how to visually introduce the tradeoff between data privacy and utility that primarily concerns differential privacy, but it could be solved with visual analytics. If such a problem does not exist, then other levels of the research method and their implementations are becoming meaningless. Thus, problem definition and validation are essential aspects of the project, without which the whole thesis is losing its significance and necessity. In addition, the question is whether there is a need to merge domains differential privacy and data visualization in order to start the privacy-preserving visual analytics approach.

4.1.2 Validation, Expert Study

In order to discuss the potential threats and validate them, for the first level of the research method, an expert study will be conducted. Such an approach is based on the qualitative discussion regarding the problem of concerning the right target audience for the specific issue. As earlier explained, in our case, such a problem refers to connecting differential privacy with visual analytics, and the target audience is researchers that work on differential privacy, data analysts that use such algorithms, and practitioners whose data is used. The validation method would be a qualitative research method as an expert study. According to [29], such a research method is instead focused on understanding the experience, meaning, and perspective of a specific participant, which is, in our case, an expert from the domain of differential privacy. In addition, the author [29] emphasizes that investigates beliefs with in-depth interviews, while data are not amenable to the counting of measuring. Given this statement, conducting a one-on-one in-depth expert

interview is an appropriate approach to discussing the project's problem definition. In terms of interview type, it was decided to run semi-structured, mainly because the idea is to let a researcher openly discuss the problem between differential privacy and visual analytics and the shortage of papers contributing to such relation. In addition, there are only be topics and main questions created, while other subquestions and additional topics are open to coming up if the discussion lead towards them.

Since the thesis focuses on visual analytics for differential privacy, it was decided to focus the evaluation method on differential privacy experts. Because of their expertize, it is expected that researchers could determine for which specific group should such system be offered. The general reason for including only researchers of differential privacy, rather than data visualization experts, the visualization system primarily helps with the domain of differential privacy. Because the whole project is meant for researchers from the domain of differential privacy, there is no need to include any experts from other fields such as data visualization, although visual analytics is one of the main domains of the thesis. In addition, the best way to validate a problem definition is to directly talk to researchers that could answer whether they find a general research problem as their interest. In terms of the evaluation method, it was decided to contact the researchers that mainly focus on differential privacy and they have published work within the domain. In addition, there are three researchers that were chosen for the study.

Expert, Di Wang The first expert is Di Wang, an Assistant Professor of Computer Science in the Division of Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) at the King Abdullah University of Science and Technology (KAUST) [50]. Earned Ph.D degree in Computer Science at The State University of New York (SUNY) at Buffalo in 2020 under supervision of Dr. Jinhui Xu [50]. His main research areas are Differential privacy, privacy-preserving machine learning/data mining and privacy attack in machine learning, focusing on both global and local differential privacy with the emphasis on risk minimization. In addition, his contribution to the differential privacy domain was recognized by numerous publications, in which the work *Empirical Risk Minimization in the Non-interactive Local Model of Differential Privacy* [51] stands out.

Expert, Andreas Haeberlen The second expert is Andreas Haeberlen, an associate professor at the University of Pennsylvania, and the undergraduate chair for CIS (Computer and Information Science, A Department of the School of Engineering and Applied Science) and NETS (Networked and Social Systems Engineering) [27]. In addition, Andreas is a member of the Distributed Systems Lab and the recipient of the Ford Motor Com-

pany Award for Faculty Advising and the Lindback Award for Distinguished Teaching [27]. His contribution to differential privacy concerns global setting with the work *Differential Privacy: An Economic Method for Choosing Epsilon* [31] and *Differential Privacy Under Fire* [28] that stand out. In addition, the work by Andreas Haeberlen is focused on the practical aspect of differential privacy with creating numerous implementations of the algorithm for different purposes.

Expert, Chuhao Wu The last expert is Chuhao Wu, a research assistant and a PhD student of Information Sciences and Technology at Penn State University [58]. One of his recent project is Differential Privacy and data disclosure decision-making, and the research interest is the theoretical aspect of differential privacy, which brings a new aspect to the domain compared to previous experts.

Topics There are two topics to be discussed, problem definition and target audience. With the problem definition, the aim is to investigate whether there is a need to visually analyze differential privacy, especially emphasizing the tradeoff between data privacy and utility. The first goal is to find whether the expert agrees that visual analytics help understand the concept of differential privacy and if there is a need for such a relation of two domains. The second goal is to understand if users' records data are subject in an alarming position because of linkage and privacy issues. The last goal is to inspect if there is a need to visually adjust the epsilon parameter and then display results on a visualization. The second topic is the target audience, where the only goal is to discuss whether which group of differential privacy domains are the ones in need of having such a visualization system that allows understanding the tradeoff between data privacy and utility. For both of these topics, it is expected to discuss additional subtopics in order to get as much information as possible.

4.2 Data Abstraction Design

The second level is data abstraction design, where the focus is on choosing an appropriate data and operations that would be used for solving the problem. Munzner (2009) [42] explains that data abstraction design is crucial for making visual encoding decisions properly. This level aims to find the right data type to elaborate the problem explained at the highest level. As demonstrated in the first level of the methodology, users' data that comprises private information about individuals is often under privacy risk. By any improper disclosure of a dataset that contains confidential information about individuals, privacy is massively harmed. Since the research focuses on individuals' privacy, the verdict was to search for users' records data, and operations involve protecting individuals privacy while presenting accurate

population statistics. Since the purpose of differential privacy as a definition is to secure individuals' data while providing accurate summaries of the population, the decision was to use a users' data. Not only that this kind of a dataset would fit the differential privacy purpose, moreover, it eases evaluating privacy on the dataset by picking an individual and tracking him/her while adjusting the amount of differential privacy. By examining specific users within a dataset, we can easily track the progress of anonymization of individuals, this privacy can be evaluated efficiently.

Besides concentrating on users' records data, it was decided to work with numeric datasets. It means that the system accepts only data that contains numeric values exclusively. The reason for such a decision lies in different understandings: firstly, the utility metric that is implemented in the system is Euclidean distance, which works only with quantitative data points. Second, using textual data could cause additional issues with having different categories and poor accuracy for specific attributes. Third, expanding the project towards datasets that contain strings would exceed the time agreed and go beyond what was initially agreed at the derivation of the thesis. However, in the Limitations and Future Research sections, these opportunities for further developing the system will be introduced and elaborated. All in all, what interests us are datasets that contain a couple of attributes that numerically provide information about users. These users are mostly secured by ID, however, such solution does not solve the linkage issue. Given the user identification, it will be easy to follow whether information about specific user change by adding and adjusting the differential privacy mechanism.

Contribution The contribution of the layer is that the visualization system accepts quantitative data types and manages to provide visual analytics for data privacy and utility. By having two different scenarios that contain specific data types, and their characteristics differ for variance, the analysis should give different results for the epsilon parameter value. In addition, by having different expectations regarding data privacy and utility tradeoff, data types will play a significant role in determining the value of the epsilon parameter to accomplish the desired intentions.

4.2.1 Threats

The threat that concerns the second stage is if the given data types solve the problem. Precisely, it is questioned whether the chosen data type contributes to the problem of a specific target audience. In our case, the target audience are users whose data is potentially at risk of being improperly disclosed. Thus, the threat is if proper datasets were chosen to be investigated throughout the visualization interface. In addition, since the thesis takes visual analytics approach, the visualization system is the main contribution

of the project. Hence, the threat relates to analyzing whether the design of the visualization system correctly accepts a specific users' records dataset and allows introduction and adjustment of differential privacy with possible visualization insights.

4.2.2 Validation, Case Study

The validation for such a threat would be based on the case study, where the creator of the system goes through it and provides screenshots of results. Since data itself is not a solution, the validation is based on using data on the visualization system to understand how it works in the system and how it changes while adding and adjusting the differential privacy mechanism. By using data privacy and utility metrics, it is possible to analyze at what amount of noise level injection does the dataset give the best results for a user. To successfully accomplish such results, it is essential to define the insights that the visualization system presents in a specific case study research. In the upcoming paragraph, the definition and explanation of the case study will be given, alongside the protocol and expected outcome.

The research question that supports validation of the second level is *How can visualization systems help users' records data on the tradeoff between data privacy and utility.* To validate the question, it was decided to take qualitative case study research. Such research method produces decent results when capturing explanatory information on 'how', 'what' and 'why' questions, in real-time context [11]. According to Baxter (2008) [5], the qualitative case study research by taking a variety of data sources facilitates the exploration of a phenomenon within its context. Such definition correctly applies to the validation because the case study research will take two scenarios that contain two datasets from specific domains: healthcare and financial. In addition, the exploration of a phenomenon in the research is focused on validating the visual analytics approach within the tradeoff of privacy-preserving techniques. The author [5] adds that the issue is investigated from different 'lenses' that allow understanding and elaborating the phenomenon from multiple perspectives. In our case, the phenomenon is the tradeoff between data privacy and utility, and the different perspectives are differential privacy (privacy-preserving data analytics), and visual analytics (data visualization). Zainal (2007) [63] agrees with what Baxter from 2008 states about a variety of data sources and adds that such a case study method enables the closer examination of data within a specific context. Such a statement matches with intentions in the thesis, where two different data domains will be examined in the context of visual analytics for the data privacy and utility tradeoff. In addition, the same author states that a case study enables a researcher to explore more than quantitative statistical insights and 'understand the behavioral condition through the actor's perspective' [63], which in our case would be users or experts that are interested

in the understanding the tradeoff.

In terms of case study type, it was decided to use Yin (2003) [62] categorization and take explanatory as the case study type. The main characteristic of such an approach is that it examines data in order to explain the phenomena [63]. In addition, Yin (2003) [62] and Baxter (2008) [5] explain the usage of the approach for answering way complex questions for the survey or experimental strategies, and the focus is on linking program implementation with program effects.

Objective In terms of the objective of the case study, the focus is on proving that the visualization system achieves the expected results on specific data. These expected results depend on the data for each scenario, which will be elaborated on in the upcoming section. Generally speaking, it is expected that using the visualization system on specific data will end by having private data with adjusted utility. It means that the outcome should be a synthetic dataset previously analyzed and adjusted for the noise level of epsilon privacy. Such adjustment was taken subjectively, and it depends on the knowledge of an analyst and the data itself. Such an adjustment aims to decide on how private a dataset should be and on what scale the dataset's accuracy and value should be altered. We will explain when and why such adjustments differ for two scenarios with different datasets and their domains in the upcoming section. In conclusion, the objective is to prove that the visualization system helps analyze, understand, present, and adjust differential privacy mechanisms to output a synthetic dataset that guarantees the privacy of individuals' records.

Scenario For the case study, there are two scenarios to be conducted. There are two main reasons for including two scenarios. The first reason for taking multiple cases is that they come from different domains, thus their needs in terms of tradeoff between data privacy and utility differ. Each domain has different purposes and ways of analyzing users data, and also each domain has some unique features in users datasets. These domains are healthcare and finance. Both of them are mostly based on users data, thus they are relevant as domains that could have potential privacy and linkage issues. What differs between them is the perspective of analysis and privacy expectation. By perspective of analysis, it is mean that one could be more concentrated on analyzing individuals, while the other could be more focused on analyzing population.

In the last decade, digitizing medical records has been a crucial shift in the healthcare domain, resulting in increased data volume [45]. By collecting and analyzing patients' data, researchers are significantly improving global health [25]. However, there is always a probability of harm of exposing personal privacy, despite knowing that health information collected from

patients to help health research is society benefit [25]. Thus, such a domain is constantly exposed to privacy issues that could be solved with differential privacy. In addition, since healthcare is not an IT-related domain but natural sciences, healthcare specialists are not expected to understand differential privacy. Hence, using the visualization system with patients' data would help them estimate the amount of noise to introduce to a dataset to protect patients' privacy while the dataset remains valuable.

Healthcare is a domain where most of the analysis is taken from the perspective of population. Various machine learning algorithms can detect specific disease patterns from the population data, thus having accurate data play an essential role. Gostin, Levit and Nass (2009) [25] state maintaining utility of data without disclosing privacy constraints as a fundamental issue, thus it is vital to remain the value of a dataset as high as possible. On the other hand, individuals' privacy has to be protected. In addition, by analyzing patients' data, we can infer that most of the columns contain categorical data values such as sex, race, and symptoms. Given that statement, changing values of those columns with fewer options to be replaced leads to a more accessible linkage of synthetic users with original ones. To protect the patient's personal information, it is crucial to affect as many columns as possible. However, such an approach demands introducing a high amount of noise in a dataset, thus our priority for the healthcare scenario is privacy before utility.

On the other hand, the financial domain demands a different approach for the tradeoff between data privacy and utility. According to Earp and Paython (2006) [20], in 1999, the Gramm-Leach-Bliley Banking Modernization Act allowed freely sharing clients' private banking data between banks, insurance companies, securities firms, mutual funds, and brokerage firms with affiliated groups. Such an act could potentially present a privacy risk with improper disclosure of customers' data to unauthorized firms. Moreover, computer technology stated to analyze financial data for different purposes. One of the most popular cases of using users' records for data mining in the finance domain are fraud detection and credit score. Both cases rely on understanding customers' financial habits and behavior to detect irregularities or define a pattern of specific client profiles. Such analysis lead to improved banking and finance software that is used by a considerable number of customers.

By examining finance dataset examples, it was concluded that most columns contain quantitative data types that contain high ranges of values. Most of these columns are based on currency values such as gross, income, and debt. Given that statement, it is impossible to have two individuals with the same values for more than two or three columns. This means that if we change only a few column values for a specific customer, a user's privacy will be protected. Thus, we can focus more on providing maximal utility and keeping data as valuable as possible while maintain-

ing sufficient privacy of individuals for financial data scenarios. With such synthetic datasets, analysts who work on fraud detection or algorithm for credit scores can rely on accurate and valuable data while still being sure that customers' privacy is guaranteed.

The second reason for the conduction between the two scenarios is that two cases also differ in amount of instances for the two datasets. It means that one dataset contains a lower number of data points while other dataset contains higher number of data points. The rationale behind such decisions lies in fact that there are possible different outcomes of differential privacy when having a smaller dataset compared to results when having a larger dataset. The explanation is simple, if a small dataset was permutated, there isa higherr probability that its accuracy would decrease. When having less number of instances, each adjustment of their values has more effect on the overall statistics than when a dataset contains larger number of instances. In other words, there is a higher probability to affect the dataset utility when the number of instances is lower than when the number of data point is large enough not to get affected by the differential privacy mechanism. Such a situation is highly interested to observe, thus deciding on having two scenarios with two datasets that contain different number of data points gives a more detailed evaluation. In addition, by adjusting the epsilon parameter in the visualization system, we can observe at what point does the differential privacy does the best work for each dataset, especially since we cannot guarantee that epsilon parameter will be optimal for both datasets at the same value. Such analysis could give us interesting insights into the tradeoff between data privacy and utility in terms of having datasets with different number of instances. While the healthcare scenario will contain a dataset with smaller amount of instances, the financial scenario will be conducted on a dataset with large number of data points.

After explaining the reasons for conducting each scenario and their background information, we will present the protocols of the two cases. Both scenarios contain the same protocol, thus we will present it below. Staring at the data description, it is important to understand a dataset and its characteristics so that results and their interpretation could be clear. Next, the experimental phase of the case study begins with importing a dataset and using it throughout the visualization system, by constantly adjusting the slider with the epsilon parameter to see the interactive results. After covering all features of the visualization system, a dataset will be exported at a specific epsilon parameter value, depending on the specifics of each case. Such a synthetic dataset will be finally compared with original version to distinguish differences and conclude the case study.

1. Dataset description

2. Data import

3. Side bar analysis
4. Data tab analysis
5. Mean tab analysis
6. Euclidean distance tab analysis
7. Row Privacy tab analysis
8. Column Privacy tab analysis
9. Export data

Starting with a healthcare dataset that contains a lower number of instances, the first step is to describe a dataset. It was decided to search for appropriate datasets from the Kaggle website, which provides numerous datasets in various domains. In addition, it is essential to find a dataset that is focused on users and contains quantitative data, thus the search became complex. Finally, the chosen dataset is the Pima Indians Diabetes Database, an open dataset from the domain of healthcare that contains individuals' records with a smaller number of instances. Despite taking the dataset from Kaggle, it originates from Smith, Everhart, Dickson, Knowler and Johannes from 1988 [49]. Thus, we will mainly use their paper to explain the specifics of the dataset, and some characteristics such as columns with their description will be presented in the table below. The dataset is a collection of diabetes within five years in Pima Indian women. Initially created by the National Institute of Diabetes and Kidney Diseases, the dataset serves for diagnostically predicting diabetes for a specific patient [37]. The figure 9 presents the sample preview of diabetes dataset.

1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0

Figure 9: Diabetes dataset preview with top 19 rows, containing eight columns.

Column	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blod pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)ˆ2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (0 or 1)

Table 1: Pima Indians Diabetes Database (description taken from [37])

To collect data, authors [49] took a selection of patients. The first criteria were that a subject is female, and the second is that a female patient must be at least 21 years old [49]. The third criteria divide into two sub-criteria from which at least one has to be accomplished, diabetes being diagnosed within five years of examination, or Glucose Tolerance Test (GTT) being performed five or more years failed to reveal diabetes mellitus [49]. The last criteria were that if diabetes occurred within one year of an examination, such exanimation was excluded from a data collection process to remove from the forecasting model those cases that were potentially easier to forecast [49]. Having these criteria, as a result, there are 768 instances (patient) collected for the dataset [49]. In terms of attributes, eight out of nine variables are meant for forecasting, while the last variable serves as an outcome [49]. It was found that these eight variables are significant risk factors for diabetes among Pimas and other populations [49]. In terms of preprocessing, there was no data manipulation before using the dataset in the case study.

The second database is focused on the financial domain, thus the chosen dataset had to match the domain with having quantitative data types that were explaining individuals' information. The chosen dataset also comes from the Kaggle website for datasets and any data science information, however, it can also be found on UCI Machine Learning Repository. Default of Credit Card Clients Datasets presents information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005 [37]. The dataset contains 25 columns elaborated in the table below as it was explained on the Kaggle website. By having a large number of attributes, such a situation makes the scenario more interesting in terms of privacy. In addition, the dataset contains 30 000 instances, thus the data instances difference between the two scenarios is accomplished with one dataset of 800 rows and another with 30 000 instances. Unfortunately, the dataset does not contain any additional information that would help to understand its background. In terms of preprocessing, there was no data manipulation before using the dataset in the case study. The figure 10 presents the preview of top 30 rows of the credit score dataset that contains 25 columns.

Column	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and familty/supplementary credit)
SEX	Gender (1 = male, 2 = female)
EDUCATION	1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown
MARRIAGE	Marital status (1 = married, 2 = single, 3 = others)
AGE	Age in years
PAY_0 ->PAY_6	Repayment status in September - April, 2005 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, ... 8 = payment delay for eight months, 9 = payment delay for nine months and above)
BILL_AMT1 ->BILL_AMT6	Amount of bill statement in September - April, 2005 (NT dollar)
PAY_AMT1 ->PAY_AMT6	Amount of bill statement in September - April, 2005 (NT dollar)
default.payment.next.month	Default payment (1 = yes, 2 = no)

Table 2: Default of Credit Card Clients Dataset (description taken from https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset)

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
7	5.00E+05	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
8	1.00E+05	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0
11	2.00E+05	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67969	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	316	0	632	316	0	1
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1
24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1
25	90000	1	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0
26	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0
27	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0	1000	0	500	0	1000	1
28	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0
29	50000	2	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	3415	3421	2044	30430	257	0	0
30	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1000	1600	0	0

Figure 10: Credit Card Clients dataset preview of top 30 rows with 25 columns.

4.3 Visual Encoding and Interaction Design

The third level is visual encoding and interaction design, where visualization comes into focus. The project's main practical focus is on creating an interactive visualization pipeline that allows the import of quantitative data, privatizing it and visualizing the results of the differential privately mechanism. In addition, since the project is based on the data privacy and utility tradeoff, main feature of the visualization framework is adjusting the epsilon parameter in order to change the noise injection level. By altering epsilon parameter, it is possible to instantly see the results, thus visual analytics are playing a crucial role. Since the system allows various data to be imported, the visualizations in the program do not depend on data itself but rather on data privacy and utility metrics. This means that the visualization system aims to visually present the solution to impact users' understanding of privacy and the tradeoff. In the upcoming section, we will describe in detail the visualization system and its components.

Contribution As the main feature of the thesis, the visualization system is the most significant contribution of the project. The interface is related to previous contributions focused on the theoretical aspect of contributing

to the problem and target audience and contributing to understanding what data types are included in the situation. By creating an interface that allows adjusting the noise-injection level, adding noise shows users how data privacy and utility change for a specific dataset. Therefore, the system contributes to the theoretical aspect of the thesis by supporting the concatenation of data visualization and privacy-preserving data analysis in practice. In addition, the system approves the data types defined in the previous layer, and it contains the algorithm that will be presented in the last layer. Lastly, the visualization system directly contributes to the first research question of visually presenting the tradeoff between data privacy and utility, with additionally providing the epsilon parameter as the slider, allowing users to see the results and managing tuning noise-injection to improve the utility of a dataset

4.3.1 Visualization System

As earlier explained, the purpose of the visualization system is to introduce visual analytics to differential privacy and the tradeoff between data privacy and utility. If we take a higher-level perspective, its contribution lies in relating privacy-preserving data analysis techniques with data visualization domain. In addition, the focus of the system was to allow interactive adjustment of the epsilon parameter that is responsible for the tradeoff. The first distinction on the interface is between the main panel (center-right) and the side panel on the left side. While the right part, the main panel changed according to chosen tab and adjusted slider, the sidebar is permanently fixed. While the main panel will be elaborated by each tab, the sidebar will be given as its own section. It is important to understand that there are six tabs that together combine in the main tab. By choosing each tab from the navigation panel, a user can try different features of the visualization system.

The idea was to create the visualization system that does not only allow the adjustment of the epsilon parameter and see it results on a dashboard, but also to create a pipeline that allows importing various data and manipulating it. In order to allow such a feature, it is important to focus on visualizations that are not concentrated on datasets but rather metrics that are introduced with the differential privacy algorithm. In that way, it was doable to allow any quantitative dataset to be used in the visualization system. Besides importing any data, in order to have full pipeline, it is important to allow exporting a synthetic dataset.

Figure 11 presents the pipeline of the visualization system. The pipeline starts with importing a specific quantitative dataset into the system, allowing to manipulate a number of columns and rows that will be permutated. After analyzing the original data, a user can anonymize data by adjusting the slider with the epsilon parameter value, which controls the noise-injection

level of the algorithm. Interactively, while adjusting the slider, a user can visually analyze the impact of the algorithm and noise-injection level on the dataset. Lastly, when the epsilon parameter is adjusted on the preferred value, and all characteristics are analyzed, the dataset can be exported for further usage.

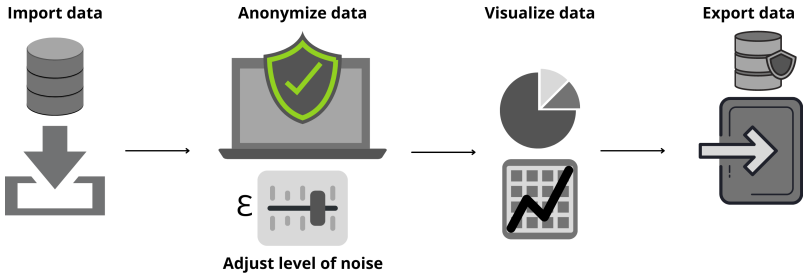


Figure 11: The visualization system architecture, including data import, all tabs with side bar, and data export.

Figure 12 presents a high-level perspective on the visualization system. After importing the data into the system, two features, the side bar, and data tab first display. Next to importing and exporting a dataset, the side bar has three main features: choose row range, choose columns, and adjust the epsilon parameter value. All five features alongside the whole side bar are always displayed in the visualization system, and ready to be used. It means that a user is allowed at any point of his usage to import another dataset, do data manipulations, or export a permutated dataset. There is always one tab displayed alongside the side bar, the data tab being the default tab. By using the navigation menu, a user can change between tabs. While tabs change, the side bar always remains as a side feature displayed on the visualization system.

In upcoming pages, the idea is to introduce each segment of the visualization system and to explain its value to the project.

Side Bar The primary purpose of the sidebar is to allow data import or export, manipulations, and adjustment within a dataset and differential privacy mechanism. Starting from the top of the bar, a user can apply an existing example dataset or import any quantitative dataset of a personal preference. The existing example dataset is a built-in dataset in R based on 43 instances of United States judges' ratings. If a user decides to upload their dataset, it is also essential to upload only comma-separated values (CSV) files despite having strictly quantitative data values. After choosing whether to take the existing dataset or importing own, a user can determine a definite range of rows to be used through the visualization system. As a default, all rows of a dataset are chosen. Next, it is possible to choose which columns to

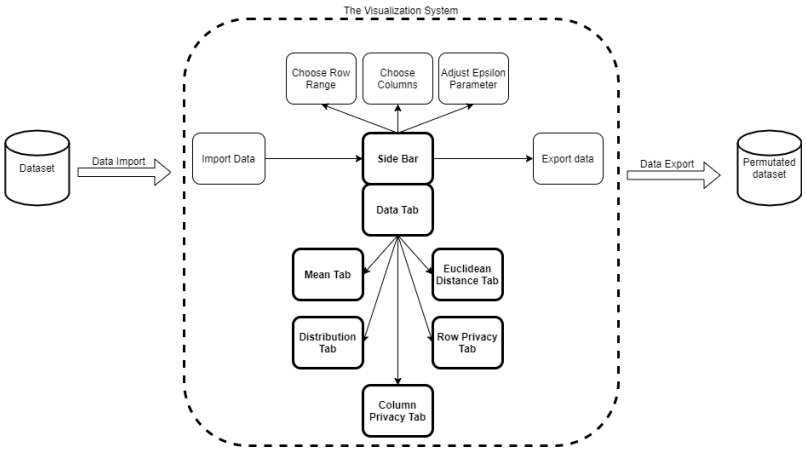


Figure 12: The visualization system pipeline, including data import, anonymizing data with adjustment of the epsilon parameter, visualizing data and exporing it.

be presented and used to interact with the visualization system. Compared to choosing rows, when choosing attributes, it is vital to have an interactive system that reactively reads the column names of a specific dataset and presents them in the sidebar. Such an option was successfully implemented in the visualization system as visible. As a default, all attributes of a dataset are chosen, no matter if a user decided to use an existing dataset or import a new one.

The next feature is the critical component of the visualization system, the slider. Since the thesis aims to relate differential privacy to visual analytics from a perspective on the tradeoff between data privacy and utility, the main visualization goal is to find an adequate solution that allows appropriate adjusting of the epsilon parameter. As a solution, the slider was chosen for different reasons. First, such a slider works efficiently on interactive systems. Compared to radio buttons, dropdown lists, or quantitative input, the slider is a user-friendly feature that allows dynamical adjustment of values within a specific range. It provides a clear overview of the whole range and gaps between the two values in the slider, thus it is a simple but efficient solution. Second, the epsilon parameter is a quantitative input that presents a specific quantitative value that affects the probabilities within the differential privacy mechanism, and it contains a specific range of implementation. The slider is the most efficient solution for such values, and thus it is the logical choice to implement it in the sidebar. Since the epsilon parameter depends on a specific differential privacy mechanism, there is no arranged range of values. However, the minimum value of the epsilon parameter is 0.1, and the maximum value depends on the computations within the algorithm. In our case, the maximum is with the epsilon value of 2. To preserve the clear distinction between epsilon parameter values, it was decided to

include a step value of 0.3, thus between the two values, there is a difference of 0.3. Such a decision allows us to analyze and understand the difference of the results between two epsilon parameter values. In terms of the slider's functionality, it works in the same tradition as the literature refers to, by adjusting the slider closer to 0.1, the noise-injection level increases, meaning that the probability of replacing each data point increases. It results in a synthetic dataset with a lot of noise, higher privacy, and possible lower utility. On the other hand, by adjusting the slider towards the value of 2, the noise-injection level decreases, resulting in a highly accurate dataset that is probably not well-protected. Since the probability favors leaving each data point as it is in the original dataset, the privacy risk remains. It is essential to understand that the slider plays a crucial role since most interaction between a user and the visualization system happens when adjusting the epsilon parameter. The last feature of the sidebar is the action button to download a dataset. Such a feature is used after an original dataset is permutated for a specific epsilon parameter adjusted by the slider. In addition, the feature exports data in the CSV file. The figure 13 presents the sidebar with its features that were earlier described. In conclusion, the purpose of the sidebar is to allow users to manipulate a dataset and adjust the slider at any moment of interacting with the visualization system.

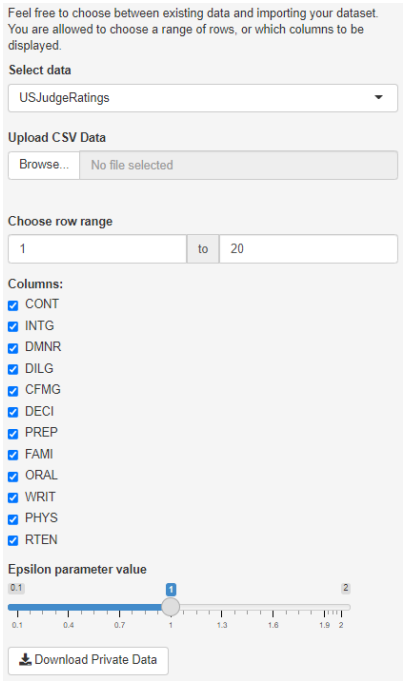


Figure 13: Sidebar, part of the visualization system that allows importing and manipulating data, but also controlling the differential privacy mechanism.

Data Tab The first visible tab on the main panel of the visualization system is the data tab. The main purpose of the tab is to introduce a user with existing or imported data. To visually provide as much information as possible, it was decided to include two features, data preview and summary statistics. In the data preview feature, a specific dataset can be examined by choosing how many top rows to be displayed. By examining, it is meant that a user gets understanding of a data, what are columns in a dataset, and what they present in general. In addition, a user can examine the rows to understand typical examples of a data point incorporated in a dataset. In addition, by adjusting side bar features such as specific rows and columns, these components also change in the preview table. Such features allows data manipulation in a way that a user can visually see on a data preview which column or rows suits his/her needs. Moreover, the slider affects the data preview table, in a way as it would be expected. By shifting the epsilon parameter, the values of a dataset that are displayed in the data preview table should change. Since the mechanism contains a randomized function, and since we did not specify the slider value, we cannot predict the number of values being changed. However, such scenario will be detailed examined in the case study research. Some additional features of the data preview table are sorting each column for their numeric values by ascending or descending order. Since the table presents a specific number of rows per page, a user can also adjust the number of rows to be presented. The initial number of rows to be displayed is six, however, the number can also be adjusted to 10, 25, 50 or 100. In addition, if a user desires to see other than the top rows, there is a possibility of navigating to other pages, and such a feature is enabled in the bottom right corner of the data preview table. Lastly, at bottom left corner there is an amount of instances of a dataset written down, so that a user could always be aware of number of rows that was chosen. The second feature is summary statistics table, which contains the main statistical measures for each column of a dataset. The purpose of a such feature is to get familiar with the statistics of a dataset, and such information could be crucial in comparing the results of an original and synthetic dataset. Since differential privacy affects data points, there are expected adjustments of statistical measurements for a synthetic dataset. By analyzing differences between an original and synthetic dataset, we can infer whether the utility remains the same by looking at their summary statistics. For each column, there are statistical measures of minimum and maximum values, 1st and 3rd quartiles, mean, and median. As a user adjusts the slider, the values of summary statistics for all chosen columns could change. The figure 14 presents the data tab alongside its two main features, the data preview table and summary statistics table. In conclusion, the data tab is default panel that serves to introduce a user to a chosen dataset, which allows visual presentation if adjusted components and getting familiar with statistical measures.

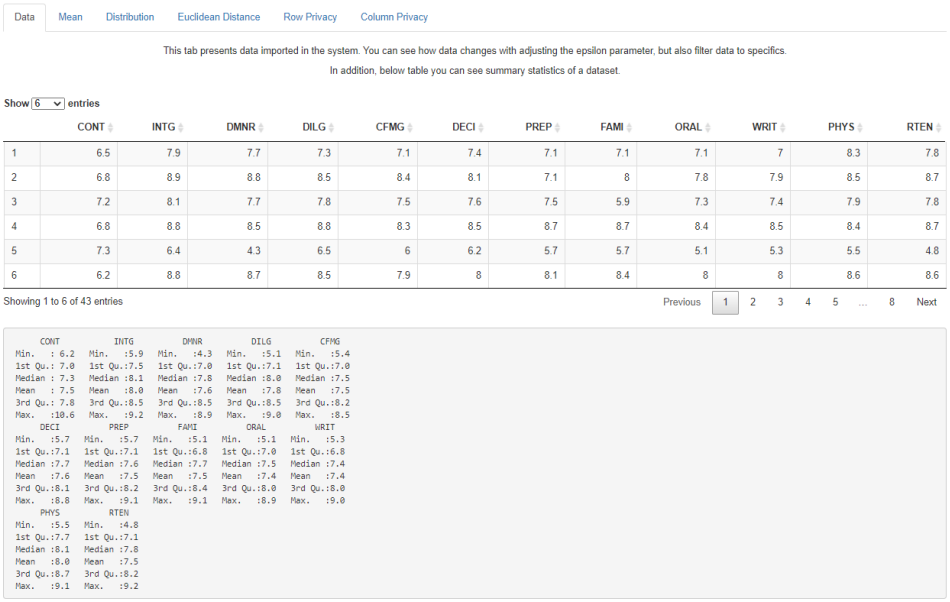


Figure 14: Data tab, presenting table preview of a specific dataset alongside its columns and specific number of rows. Next feature of the tab is summary statistics of a dataset, presenting Minimum, 1st Quartile, Median, Mean, 3rd Quartile and Maximum values of each column.

Mean Tab The second tab of the main panel is the mean tab. This is the first tab that contains a visualization for explaining a specific measure. Precisely, the mean difference is presented as a metric for utility. The purpose of the tab is to see for each column of a dataset whether the mean value of columns changed after implementing the differential privacy. The comparison is focused on finding the mean difference between the original dataset and the synthetic dataset. To calculate the variation, it was decided to create a specific mean calculation. The mean difference is calculated by taking the mean for each column of the original dataset and compare its value against the mean value for the same column of the permutated dataset. By presenting the difference value as the absolute value, the mean difference is calculated for each dataset column in three steps:

1. For each column of a permutated dataset, take their mean values.
2. Compare the mean value of a synthetic dataset to the same column from an original dataset. The comparison is done by subtracting one mean value from another in the absolute results, creating mean difference values.
3. By taking that difference for each column and averaging it, we have a single mean difference that presents the global difference between the original and permutated datasets.

In order to present the mean difference for each column, the bar graph was illustrated. While the x-axis presents all column names of a dataset, the y-axis presents the mean difference value. By the value of the calculated score, the saturation of green color on bar columns increases, while it decreases for columns with a nominal value of the mean difference. We can interpret the bar graph in a way that higher and greener columns present higher mean differences, which means that there is more variation between original and synthetic columns. There are expected changes in the mean difference for each column by adjusting the slider's value of the epsilon parameter. In conclusion, the bar graph serves to understand if there is a significant population change between two columns of the original and synthetic dataset. Such analysis aims to give more insight into the tradeoff's utility perspective between data privacy and utility. In the figure 15, there is a screenshot example of the mean tab in action, however, with case study scenarios, such a situation will be described in detail.

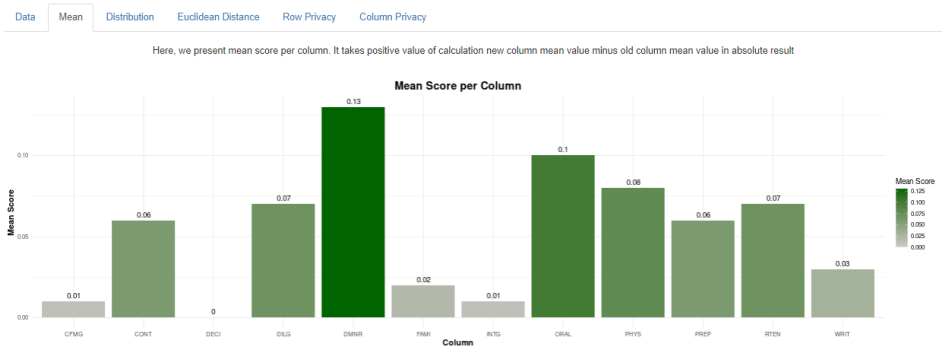


Figure 15: Mean tab, presenting difference in column means between original and permutated dataset - the highest value shows that for a specific column there is significant difference in means between the original and permutated dataset.

Distribution Tab The third tab of the main panel is the distribution tab. Similarly, as in the previous tab, the distribution tab focuses on comparing an original and synthetic dataset by columns. The difference is that in this tab, the focus is on a visual investigation of the difference between two columns in terms of distribution. Given that statement, the idea was to create a nested group of distribution charts that present each column before and after the differential privacy algorithm was implemented with a specific epsilon parameter value. The value of the tab is providing a visual understanding of the differences between columns of the original and synthetic dataset. Such analysis is again more focused on the utility perspective of the tradeoff. To visually present all columns at the same time, it was decided to create facets that combine the distribution plot for each column. While the blue distribution presents a column of an original dataset, the

red distribution presents a synthetic dataset. By shifting the slider towards zero or epsilon value of three, the distribution of synthetic dataset changes to differ or match the original distribution. For example, if we are interested in column age, the distribution difference tells us that there is a probability of having a less accurate synthetic dataset after applying differential privacy. In addition, we can infer that the difference in distribution means that synthetic dataset contains different perception of the population, which could also affect further data analysis. We can also analyze if there are from a general perspective if there are some columns which have higher distribution difference than other columns. In conclusion, the distribution tab relies on strictly visual analytics without a single measure score. The purpose of such analytics is to investigate whether there are any differences in distribution between the original and synthetic dataset for a specific column. The figure 16 presents the screenshot of the distribution tab.

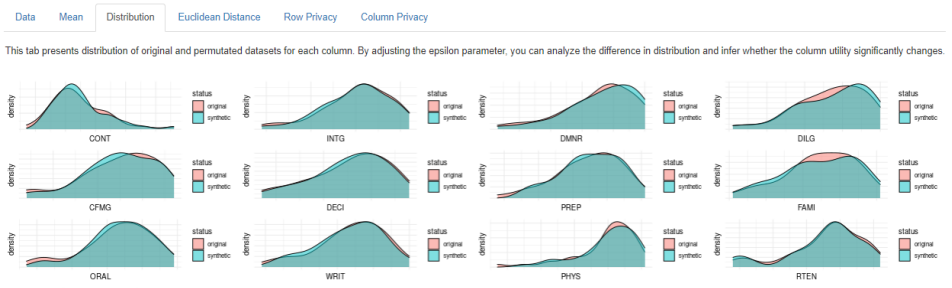


Figure 16: Distribution tab presents whether there are visual differences in the distribution for each column, and let us visually see that we can keep the distribution even when changing the epsilon parameter.

Euclidean Distance Tab The fourth tab of the main table is the Euclidean distance tab. As with the previous two tabs, the focus is also mainly on the utility perspective of the tradeoff. As the name says, the main feature of the tab is the utility metric calculated with Euclidean distance. In order to understand the Euclidean distance as a utility metric, it will be explained in the upcoming paragraphs, while here we will give a brief description and purpose. The value of measuring and visually presenting utility is that users can understand whether their implementation of the differential privacy algorithm affects a dataset. In addition, by adjusting the slider, users are interested in analyzing how utility of a dataset changes. Understanding such a situation is vital to realize whether a synthetic dataset still contains sufficient value for further usage. Euclidean distance allows a specific utility score by measuring the distance between two data points of original and synthetic datasets. By understanding their distance, we can take a higher perspective on each column and then have a utility score based on Euclidean distance even for two datasets. Such a utility metric has to be

visually supported so that users could understand how the accuracy of the dataset changes by adjusting the epsilon parameter. Given that statement, it was decided to adopt a vertical bar chart that presents the average Euclidean distance of data points for each column. By measuring the distance of data points from the original and permutated dataset, the average value of each column was taken to be dynamically presented on a vertical bar chart. The x-axis presents the Euclidean distance score, while the y-axis presents the columns of a specific dataset. In addition, column bars are colored by green, having the most saturated columns with higher average Euclidean distance and less saturated columns with lower Euclidean distance. It tells us whether a specific column's utility is more or less affected by the differential privacy algorithm and how it stands compared to other columns of a dataset, thus this is the primary purpose of the Euclidean distance tab.

To measure the utility precisely, if the data accuracy significantly changed, it was decided to adapt Euclidean distance. Such metric works with quantitative data points to measure the difference between a specific data point A in the original dataset and another specific datapoint A' on the exact location in the synthetic dataset. As explained in the previous paragraph, our interests are comparing columns and rows between original and synthetic datasets, thus Euclidean distance has to be generalized to column scale. To accomplish such a calculation, we take the mean distance between data points of a specific column for two datasets. For such calculation, the visualization system presents the graph with the Euclidean distance tab, described in the third level of the research method. In the case of this level, we are interested in a global metric that is focused on a single score that represents the dataset as a whole, thus the average Euclidean distance is obtained from all columns of a dataset. Precisely, such a score is based on the difference between columns of two datasets, however their results are averaged and placed as a single score that provides an insight whether the distance between original and synthetic datasets is significant or not. In conclusion, Euclidean distance as a utility metric enables to precisely measure the difference between data, columns, and datasets. Such calculation proved to be efficient and reliable to examine whether the utility is corrupted by injecting more noise of the differential privacy mechanism in a dataset. The figure 17 illustrates the Euclidean distance tab with its plot.

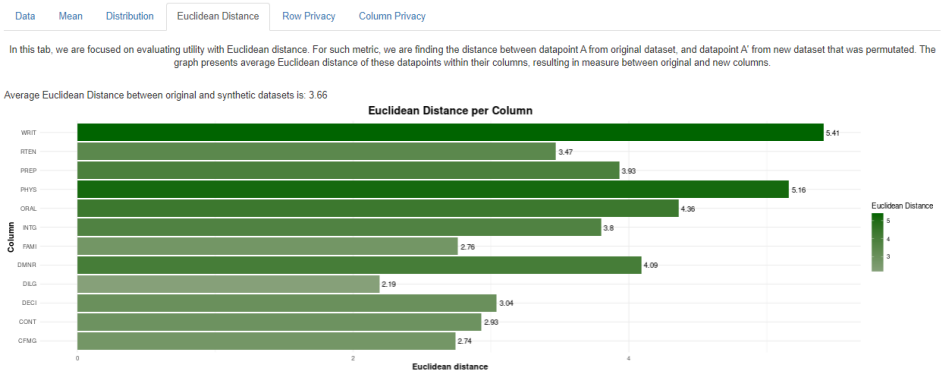


Figure 17: The tab presents the Euclidean distance score for each column, calculated as average distance for each data point of a column.

Row Privacy Tab The fifth tab of the main panel is the row privacy tab, which is also the first tab that focuses on the privacy aspect of the trade-off. The main purpose of the tab is to inspect the privacy of each specific individual and whether his/her information changes by adjusting the slider. The tab allows choosing a row number of a specific user to be analyzed as an additional feature. After choosing a user, the spider (radar) chart presents a difference in all column values for a specific user. While black lines present the column values of an original, red dashed line presents column values of a permutated dataset. By adjusting the slider, there is a chance to see the difference between the two datasets visually, thus such a plot is a visual analytics feature. In addition, below the graph, there are three one-row tables. While the first table contains information about a specific row that was earlier chosen, the second table presents information about the same row after being permutated. The last table shows for a specific row, which column values have changed after the effects of the differential privacy algorithm. As expected, by adjusting the slider, values of the second and third table could change. The value of these tables is in investigating whether the values of a specific row changes and which columns have changed. Such analysis is essential when we want to investigate a specific user's privacy and ensure that privacy is guaranteed. In conclusion, the tab contains both visualization analytics and in-detail tabular analysis to understand how the differential privacy algorithm and the epsilon parameter affect the privacy of a specific row. The purpose of such investigation is crucial for understanding at what amount does privacy influences a row. The figure 18 presents the row privacy tab and its features.

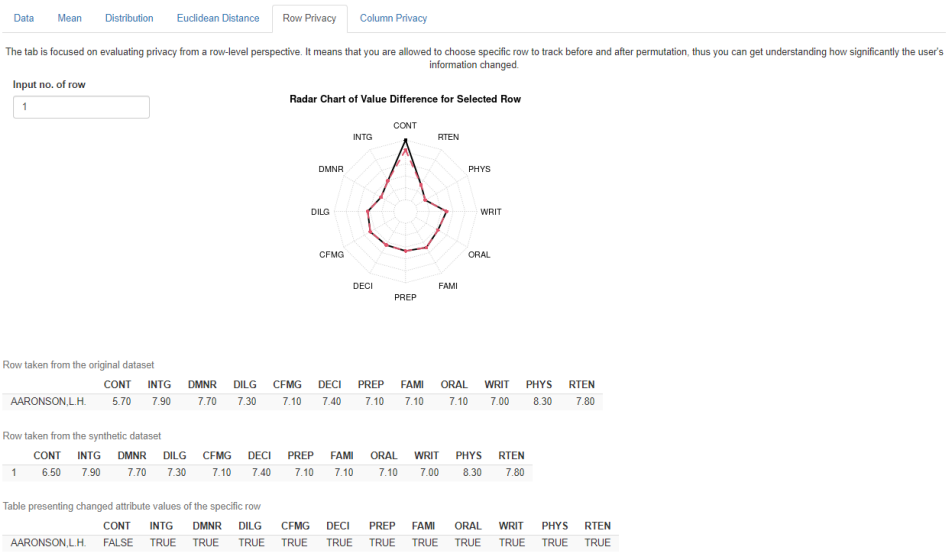


Figure 18: Row privacy tab presents two features: the radar chart that visually shows if there is difference between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

Column Privacy Tab The last tab of the main panel is the column privacy tab, which is the second tab that focuses on the privacy aspect of the tradeoff. However, in contrast to the row privacy tab, the focus is on privacy per each column. Precisely, the analysis is primarily devoted to the higher-level understanding of privacy for each column of a dataset. To present such analysis, it was decided to adapt the stacked bar chart that contains column names on the x-axis, probability of true and false values on the y-axis, and is divided by the probability of true and false values in each bar column. The graph presents for each column, what is the probability of having true or false values. By adjusting the slider, the probabilities for each column changes. The probabilities were taken from a row privacy perspective and calculated as the average probability of having true or false values for each column. The purpose is to understand the difference between probabilities for two different columns and investigate how the probability of having true values changes while adjusting the slider. The value of the tab is that it provides both visual analytics approach to the graph and calculated probability to understand the chances of having true values for a specific epsilon parameter value. In conclusion, the tab contributes to the privacy perspective of the tradeoff with a visual analytics graph that also incorporates the calculated probability of having true or false values within each column of a specific dataset.

Next to the utility, in the tradeoff, there is also a privacy component. By such definition, we want to investigate whether the individual's personal information such as an address, age, and income are secured from various

manipulations of untrusted and unauthored analysts. Since differential privacy as a definition was invented because of privacy problems such as linkage issue, privacy is an important aspect that has to be evaluated. However, having the algorithm to permutate the whole dataset to be extremely privatized does not accomplish the intentions stated at the beginning. Thus, the goal is to define a proper privacy level where an individual's private information is secure enough not to be identified while still preserving the accuracy of a population and a whole dataset. Such an ideal situation is the ultimate goal of the tradeoff between data privacy and utility.

To use a proper privacy metric that gives results that are easy to understand and generalize, it was decided to create an own privacy measurement. As it was already stated, for measuring privacy, the investigation starts by analyzing the data point A from an original dataset and datapoint A' from a permutated dataset. By comparing their similarity, we denote True (1) for identical data points and False (0) for data points that differ. Someone could infer that such an approach is significantly more straightforward than Euclidean distance, and suchan opinion on a superficial level of understanding privacy-preserving techniques is correct. However, measuring privacy and utility does not mean that the same perspectives are taking their place. When referring to utility, accuracy is the essential aspect that we consider between datapoints. As a result, the value of a dataset could be questioned, and however, if the difference between datapoints is not significant, then the value would not be jeopardized. Thus, measuring distance in the difference between data points (and later columns) plays a vital role for utility metric and data value. However, such a case is not relevant for data privacy. No matter how distant two data points are, the untrusted or unauthored analyst to manipulate data cannot claim for sure that a specific data point presents a truthful answer. In other words, a specific data point being drastically or not significantly changed from its original value is still a changed data point that protects a specific individual from improper disclosure of his/her information. Thus, the focus for privacy metric is to understand only whether a data point remained the same after the mechanism was implemented on a dataset.

As already explained, when measuring the difference between the original and permutated data point of two datasets, the approach is using local metrics, which plays a significantly important role than global metrics. The reason for such is simply since there are many rows of individuals with various attributes contained in columns, it is harder to generalize the difference between two datasets based on a global metric. Since privacy is concerned with rows of users which information are stored in columns, the focus is on privatizing their data so that personal information is not exposed. Thus, global metric only gives us the overall privacy performance, which tells us whether the algorithm manages to affect the privacy of a dataset globally. However, for an accurate analysis of a data privacy performance and its ef-

fect on every individual, it is essential to use local metrics. After comparing two data points from the original and synthetic datasets, these data points are taken for a perspective of each row. There is no measurement for estimating how private each individual is after implementing the mechanism, however, such analysis is visually available in the system described in the third level. In terms of columns, there is a measurement that is calculated with the average probability of the data point being true or false for each column. Such measurement is visualized as well in the system and elaborated in the previous sections. To have a global metric, the probabilities of columns to produce truthful or false answer are taken, and the main privacy score comes out as an average for all columns of a dataset. Thus, we can conclude that the global privacy metrics used to generally analyze how privacy changes with adjustment of the epsilon parameter on a specific dataset rely on the score that presents the probability of having actual answers taken from the average probability for each column. The figure 19 presents the column privacy tab with its feature, the bar chart of true and false value probability.

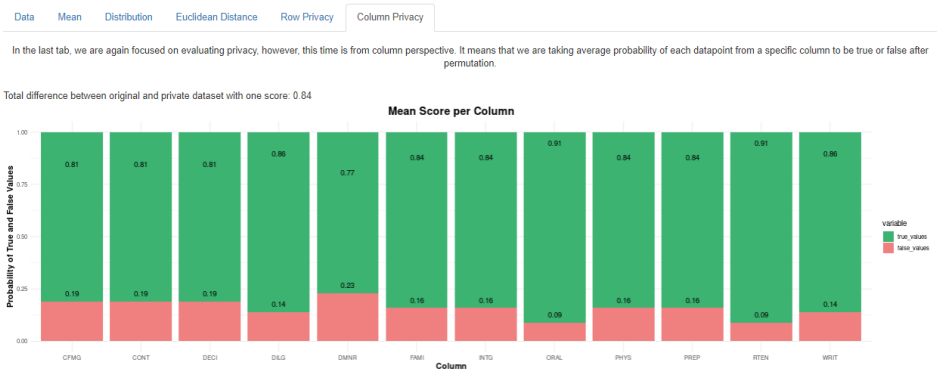


Figure 19: Column privacy tab, where the plot shows percentage ratio between having true and false values for each column.

4.3.2 Threats

Despite having a visualization system that includes various graphs, interactive components and additional feature to control data manipulation, such a framework could still suffer from issues. The third stage's threat is if the target audience understands the slider as an interaction controller that allows the adjustment of privacy and utility and the outcoming results of the adjustment presented by the system's visualizations. While all features designed for the system are working correctly, maybe their relevance or usability could be questioned. On the contrary, even the performance of these features could be evaluated, and their purpose could be examined. However, for such evaluation, there is no need for measuring the performance

numerically. Instead, conducting a discussion with knowledgeable users that understand what kind of features are needed, and what kind of performance is expected from such components.

4.3.3 Validation, Expert Study

In terms of validation for the threat, the evaluation method would be a qualitative research method as an expert study. The reason for deciding on a qualitative research approach is that the visualization system is to be evaluated for this level, which is difficult to measure. Thus, it was decided to take a qualitative approach to analyze and discuss whether the features and design were adequately created. In addition, more helpful feedback and results will be from a qualitative approach where we can discuss issues, than measuring and creating numbers that could be hard to interpret. It is important to understand that these researchers are going to be included in validating only the interface and visualizations of the system, while the algorithm itself will not be evaluated. The reason for such decision is that the algorithm was created as an alternative and simple implemetation inspired by local differential privacy, and there are expected further development of it. Instead, the focus is on validating whether such a visualization interface.

Differential privacy researchers are again in the focus of the evaluation method as experts. As explained in the first level that also contains an expert study, the general reason for including only researchers of differential privacy, rather than data visualization experts, is because differential privacy experts are ones who would profit from such contributions. At the same time, visual analytics is an approach to measure and adjust differential privacy, thus experts feedback would be only useful. In terms of the evaluation method, the idea was to collaborate with experts that were included in the first expert study for the Problem Definition.

The next question is how are we going to conduct these discussions it the experts. It was decided on a one-on-one in-depth interview with the expert. Gill, Stewart, Treasure and Chadwick (2008) [24] point out that research interviews serve to open individual's views, experiences, and beliefs on specific topics. In addition, the authors [24] emphasize that interviews are most appropriate when there is a need for elaborated feedback from participants. Such a case is for the thesis, where our idea was to discuss whether such visualization system does offer what features would help experts with differential privacy. In order to conduct such an interview, it is crucial to look for detailed and valuable feedback that would clarify the level of accomplishment for the visualization system. In terms of interview type, it was decided to approach with a semi-structured interview. The reason for such a decision was that the discussion with the expert should be flexible and allow many open or additional questions that could come up during the interview. While such flexibility leans the discussion towards the unstructured

interview, it is essential to have topics to discuss during the expert study session. Such organization of issues to talk about is essential to have all features of the visualization system be covered. Gill et al. (2008) [24] state that semi-structured interviews allow additional discovery or elaboration of information that experts initiate, which have not been previously thought of the researchers. In conclusion, the semi-structured interview gives the best from the two worlds of structured and unstructured interviews. Such is especially visible by having topics and leading questions to be covered while letting the expert and situation lead to an more detailed discussion about the specifics of the visualization system and its features.

In terms of topics discussed during the in-depth interview, the focus will be on the visualization system, which is the main contribution of the thesis. The goal is to spend the most time and elaborate on advantages and disadvantages as detailed as possible. There are many sections within the visualization system to be analyzed and evaluated, such as visualizations, tabs, slider, interactiveness, and others. In the table below, we will present essential points to be discussed alongside questions that contribute to the evaluation of each of these sections. The table presents an example of questions that will be asked during the interview, however, the intention is to allow open discussion to result in additional topics and questions. Such an approach aims to develop a brainstorming discussion that allows complete transparency and honesty in getting clear insights about the visualization system. In addition, it is essential to cover two aspects in the discussion. First is whether the experts understand the purpose of a specific feature of the visualization system. For instance, the expert will be questioned whether he/she understands the purpose of the slider, how to use it, and how it affects the change on a dataset. The same strategy will be applied for other features, where we will question the understanding of what visualizations are interpreted and if the results are easy to understand. The second aspect is the feedback for each feature of the system, where we want to understand whether such implementation does contribute to the purpose of the visualization system. By acquiring both perspectives of understanding and performance, we will get a clear picture of whether the visualization system explains its purpose to users and does that purpose contributes to users' needs. Such an approach is crucial for visualization frameworks where evaluating any component quantitatively seems unconventional, especially when usability and design are the principal matters to be discussed. In addition, there is expected discussion for the visualization system in terms of limitations and improvements. Even though it is expected that possible issues or limitations will be brought up while discussing features, it is essential to discuss such matter in detail from an overall perspective and for each feature. Such analysis will give us a proper understanding of whether the visualization system provides essential features and contributes to the academic community. Lastly, the interview will take one hour, however,

Topic	Type of Question	Description and example of questions
Slider	Understanding	What slider presents, how does affect other features
	Performance	Is slider a proper solution for visualizing and adjusting the epsilon parameter?
Data Tab	Understanding	What the tab presents, explain its features. What information can you get about a dataset?
	Performance	Does tab introduces well a dataset?
Mean Tab	Understanding	What is the purpose of the visualization and metric?
	Performance	Does it explain the difference per column between the original and permutated dataset?
Euclidean Tab	Understanding	Explain what does visualization and metric present.
	Performance	Is Euclidean distance well presented in a graph?
Row Privacy Tab	Understanding	Explain how individuals' information changes with adjusting the slider.
	Performance	Is the spider chart presenting well differences between individuals before and after differential privacy?
Column Privacy Tab	Understanding	Explain what does visualization presents for each column.
	Performance	How does the visualization analyses the privacy notion for each column?
Input data	Understanding	Are instructions of which dataset can be imported clear?
	Performance	Explain thoughts on import data limitations
Export data	Performance	Are there any issues with exporting, should there be any other format than csv?

Table 3: Expert Study: Topics, Types of Questions, and Examples

due to flexible approach, expanding the discussion duration will not present any issues. As long as the expert is providing a useful perspective for the evaluation, any time constraints do not matter.

4.4 Algorithm Design

The last stage is algorithm design, where the focus is on creating mechanisms that invoke the whole visualization system and its features [42]. Even though the project is focused on a visual analytics perspective, such does not discourage creators to implement own alternative solution of a differential privacy mechanism inspired by previous work of various researchers. By having an own variation of the algorithm for differential privacy, it was accomplished a fully unique system that could be evaluated at each level of the nested method, including evaluating the algorithm itself. In this section, we will take a closer look into the mechanism and try to understand how it works. In addition, its performance will be questioned, alongside with the plan how to evaluate it and why.

Contribution The last layer contributes to the thesis as support to the visualization system that allows the adjustment of the epsilon parameter and adding noise into a dataset. By having an alternative approach to the differential privacy mechanism, the conditions of privacy-preserving situations were created, and the visualization system manages to permutate a specific dataset. In addition, the algorithm contributes to the sub-research question of including the epsilon parameter as a noise-injection level that affects the results of the data privacy and utility tradeoff.

Algorithm In terms of the algorithm created for the project and implemented on the visualization system, two aspects are essential to be emphasized. First, the algorithm is inspired by a local setting for differential privacy. In addition, it is a modest algorithm that has a purpose to explain how would the local differential privacy work on a dataset. As explained

in the Introduction section, the focus is on affecting a dataset as a whole rather than creating queries from a dataset to be permutated. Such a setting leads to complete privacy protection since there is no untrusted curator to create queries. However, since the algorithm affects a dataset as a whole, the utility could have the consequences of poor performance that result in a meaningless dataset. The reason for focusing on a local setting for differential privacy is mainly because our thoughts were that it is easier to allow input of any dataset and create universal visualizations when the noise affects the whole dataset. The second aspect is a randomized response, which is the core mechanism invented by Warner in 1965 [57] that inspired to create our own alternative verison. The way it works is: for each data point in a dataset, the coin is tossed. In programming terms, tossing a coin is replaced with a specific function to decide whether the output is tail or heads. The probability of getting heads or tails does not follow the tossing coin probability of 0.5, but it rather depends on the epsilon parameter value. Such a parameter allows us to affect the tradeoff between data privacy and utility for the algorithm. For the first tossing a coin, there is a probability p that is calculated from the epsilon parameter value:

$$p = \frac{e^{\epsilon}}{e^{\epsilon} + 1} \tag{2}$$

If the 'coin' falls to heads, the data point remains as it is, which would be, in the general case, the true answer. However, if the 'coin' results in tails, we toss the 'coin' again. In the second tossing, the probability q is calculated differently than the earlier probability p, this time, the odds are in favor of getting heads:

$$q = \frac{1}{e^{\epsilon} + 1} \tag{3}$$

The same strategy repeats for heads, if the 'coin' falls for heads, the data point remains the same. However, if the 'coin' results in tails, we change a data point's value. A data point can only be replaced with a value that exists for a specific column. For example, if a specific column contains ages from 7 to 19, it is impossible to get 54 because such age does not exist in a column. Such a decision to replace a data point only for existing values contributes to have a higher utility score no matter what is the epsilon parameter value. While the probability p ranges from 0.5 to 0.9 in favor of heads, the probability q after being calculated by the epsilon parameter ranges from 0.2 to 0.5 in favor of tails. By having the probability of getting heads or tails as a result of tossing, nobody cannot guarantee if a specific data point was changed or not. Such a state is called plausible deniability since nobody cannot infer whether a specific value is an actual answer.

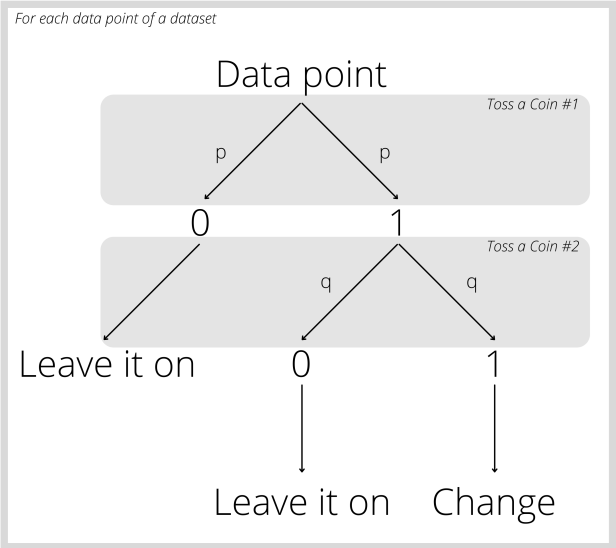


Figure 20: Algorithm, there are two tossing of coins, each having specific calculated probability (p and q). The results of tossing affect on changing r leaving each data point value.

What differs in the project's algorithm compared with other algorithms that implemented the randomized response mechanism is that we are introducing adjustment of the epsilon parameter. Thus, in the algorithm, the probability that the coin will be heads or tail is not fixed as 0.5, but instead, based on the epsilon parameter value, it changes in both directions. While the epsilon parameter value is lower, the probability in favor of having tails increases, thus chances for replacing an original data point are rising. Nevertheless, when the epsilon parameter value is higher, the probability favors utility by having more chances to get heads and the remaining data point to remain the same. In addition, it is crucial to understand that the algorithm, as other differential privacy mechanisms, contains randomness with tossing a coin. It means that if the epsilon parameter value is the same for two situations, it does not guarantee that the results will be the same. While the results will probably be very similar, it is impossible to have identical results of data privacy and utility for two scenarios with the same epsilon parameter value due to randomness. As explained in the Introduction section, a randomized function is a critical component of differential privacy definition. The figure 20 presents the algorithm with two tossing of coins, giving the clarity of when p and q probabilities affect the tossing, and how the results affect each specific data point.

4.4.1 Threats

In the last level, Algorithm Design, the threat is the mechanism's performance. Precisely, we are interested in whether the algorithm could be compared to real local differential privacy algorithms. The first expectation is that the mechanism permutates data by adding random noise for specific conditions. The second expectation is that the algorithm reacts to adjustment of the epsilon parameter with the slider of the visualization system. The third expectation is that the mechanism is synchronized with the epsilon parameter. For synchronized, it is expected that by decreasing the value of the epsilon parameter, privacy increases while utility probably decreases. The same is expected in the opposite situation, if the value of the epsilon parameter increases, the privacy should decrease while the utility increases. However, there is always a question of whether the mechanism is achieving all these expectations as wanted. Such is especially important because two aspects rely on the algorithm's performance, data privacy and utility. What could happen is that while the algorithm provides impressive results for privacy, the utility does not work as expected. For instance, it could happen that the mechanism is constantly producing a difference that utility classifies as the distance between two data points. On the contrary, it could also happen that the privacy aspect is not as secured as expected. For instance, maybe even if the epsilon parameter value is set at a lower position, users' privacy is still not guaranteed as wanted. Such possible issues must be evaluated.

4.4.2 Validation, Technical Evaluation

To validate such a threat of potential poor performance, it is crucial to obtain a technical evaluation by using metrics and analyzing the mechanism's capabilities. Based on the expectations that were previously explained, we can investigate the algorithm's performance outcomes and analyze them by comparing the results with the performance of the state-of-the-art mechanism. For such comparison, it is crucial to use the same conditions for both algorithms, thus data and metrics must be the same for both. The comparison will be based on utility metrics, and their results would accept or reject hypotheses. The decision for not including privacy as the second perspective relates to two reasons. First, the privacy metric that is used in the thesis is created by researchers, thus the only way to use our privacy metric is if the state-of-the-art algorithm provides a synthetic dataset as an outcome, which leads us to the second rationale. Second, the state-of-the-art algorithm contains different mechanisms for local differential privacy, which affect on how the privacy is evaluated for the results. Thus, the privacy results of LocalDP mechanism and the state-of-the-art algorithm cannon be compared and calculated from the same perspective.

The high-level distinction between metrics is based on global and local measurement. By global, we take one score for each metric that presents data privacy or utility for a dataset as a whole. For example, if we measure the mean difference between the original and synthetic dataset, the global metric would be the average value of the mean difference between all columns in the datasets. On the contrary, local metrics focus on each column or row, providing a detailed analysis of the performance and capabilities of the algorithm on a specific dataset. In the thesis, the idea was to incorporate both global and local metrics perspectives being the same type of metrics, but their implementation took the different scale of data. While global ones were used for the technical evaluation, the local metrics were used in the visualization system. The rationale behind such a decision is simple, the visualization system serves for deeper analysis of the algorithm and differential privacy performance. Experts are interested in seeing for each column and row, how accurate the data is after being randomized, and how private data is by adjusting the epsilon parameter. On the other side, global metrics give us a clear understanding of the overall performance of the mechanism implemented in the visualization system, and such results are easily comparable with other mechanisms.

In terms of the technical evaluation protocol, by having two utility metrics, three datasets will be taken for comparison: the original dataset, the LocalDP permutated dataset and the state-of-the-art algorithm's dataset. The main comparison investigates whether there is a significant difference for both utility metrics between the original and both synthetic datasets. Because of the randomized function in the LocalDP algorithm that does not allow to have the same results by taking the same epsilon parameter value multiple times, a specific scenario will be taken for 10 times and its utility metrics will be averaged. By such protocol, we can guarantee that these results present valid analysis on the LocalDP algorithm. On the contraty, the results that were received from the RAPPOR present valid analysis of the algorithm. For each comparison, the two-sample independent t-test will be conducted to get precise results whether there is a significant difference between datasets for a specific utility metric.

Utility Metric, Mean By measuring utility, the goal is to analyze how does a dataset and its columns and rows change between original and permutated circumstances. As expected, high utility is evident when two data points from original and synthetic datasets do not differ one from another. Then, if we take this to a higher level, if values such as mean between two columns of the original and synthetic datasets do not differ, the high utility is preserved. Moreover, it is essential to understand that the distance between the potential difference of two data points (or columns) plays significant importance for utility. For instance, if the mean difference for a

specific column from the original and synthetic dataset differs in small values, for instance, the average age in the original dataset is 55.3 and in the synthetic dataset is 55.5, then the utility is preserved. However, if the mean of the synthetic dataset is now 45, then the distance is significantly more considerable, and thus the data utility is lower. The identical situation is when comparing data points of original and permutated datasets, while an original individual is 24 years old, in permeated situation his/her ages goes to 50 years, thus the utility is harmed. However, utility interests us only on a population scale, where we want to retain the exact value of specific statistical computations between the original and permutated datasets. Hence, if the difference between two data points contains a significant difference, but overall for a specific column, the mean is not changed significantly, we would classify the utility as accomplished with higher accuracy. Thus, focusing on columns and rows as comparable units with their statistical measures, such as mean and any inference as a ratio, probability, or count, is more relevant for observing utility progress than focusing on each data point. Such understanding leads us to infer the value of a dataset after being influenced by the mechanism. By having a lower value, the risk of meaningless data that contains no value is increasing, which is not what we want to accomplish with differential privacy.

To measure utility and compare results to another algorithm, it was decided to adapt the mean as measurement. As earlier stated, the mean calculation is supported in both algorithms, which is important because the other algorithm has a global setting, thus there is no synthetic dataset as an outcome. The mean will be taken from three datasets: original diabetes dataset, permutated dataset from LocalDP algorithm, and permutated RAPPOR dataset. After acquiring the mean difference for both algorithms, the goal is to compare and analyze if there are differences in utility between two mechanisms. Such result will help us to understand whether the created algorithm behaves similarly to the state-of-the-art algorithm.

Utility Metric, Variance While the mean as statistical computation explains the centered value, the variance presents the range of values for a specific column or dataset. In other words, the purpose of the variance is to show whether many values have a significant difference between each other or most of the data points have similar values. Such statistical calculation also affects the distribution, in a case where the variance for a specific column is higher, the distribution is probably wider than normally expected, while for lower variance, the distribution contains a frequency peak. Given these reasons, analyzing utility by variance plays a significant role. By taking only mean as a utility metric, we would risk one-perspective analysis, while variance offers another aspect to investigate. If the results show for a specific column that there is a significant change in variance after the permutation,

it would mean that there is a potential utility decrease, and a dataset could lack its value. By taking each column's variance, this will present a local metric, while averaging all variances into one score will present global utility metric. The global variance will be compared between the original and Local DP datasets and between the original and the state-of-the-art algorithm's dataset. Lastly, the comparison in variance between Local DP and state-of-the-art algorithm's dataset will be computed in order to see how significantly the results differ.

In the same procedure as for the mean, the variance for each column of the three datasets (original, permutated with Local DP, and permutated with RAPPOR) to compare both permutated with the original dataset. The results will show whether there are a significant difference between the global variance of the original dataset with both permutated datasets, after which these two synthetic datasets will be compared to see if they contain similar results.

Randomized Aggregatable Privacy Preserving Ordinal Responses (RAPPOR) The state-of-the-art algorithm for the domain of Local Differential Privacy is Randomized Aggregatable Privacy-Preserving Ordinal Response, or RAPPOR, stated by Erlingsson, Pihur and Korolova [21] as new algorithm with wide relevance to many domains that guarantees strong privacy alongside the remaining high utility without having inference with usage of untrusted third parties. Created with the concept of the randomized response, the purpose of the RAPPOR is to collect data from a large number of clients and provide statistics such has histograms, frequencies, and others [21]. Privacy guarantee lies in using a Bloom filter, where the mechanism produces additional protection level to increase in difficulty for curators to disclose private information [44]. RAPPOR's algorithm relies on two defense mechanisms, both based on a randomized response, and both can be separately adjusted by the preferred privacy amount [21]. The first mechanism, Permanent randomized response creates a noisy value that is memoized by the client and permanently reused instead of the original value [21]. In other words, it replaces the real value with noisy value that may or may not contain true information, which depends the signal bits from the Bloom filter [21]. The second mechanism, Instantaneous randomized response, reveals randomized noisy value over time [21]. Not only that the mechanism is using randomized response techniques, it also performs locally on the client which denies untrusted parties to affect the algorithm [21]. In terms of real-scenario usage, the mechanism created by Google is used in its web browser solution Chrome by collecting users' answers to questions [53].

Because the RAPPOR library allows only to execute the algorithm on in-code generated data, it was decided to use their data for the technical evaluation. Because there was no possibility to import the own dataset to

evaluate the Local DP algorithm with the state-of-the-art mechanisms, the generated dataset was the only option. The dataset consists of 64 columns and 99 rows in the original form. All columns are quantitative data types, and there they do not contain high variance. In addition, while the number of columns increases, the values within each column also increase. It means that starting with the first column, data values range between 0 and 30, while in the last column, data values range between 2010 and 2050. Because it is generated dataset, there is no specific mean of data values, and thus there will not be any further raw data analysis.

In addition, there is no possibility to adjust the epsilon parameter, not any other privacy-level parameter within the RAPPOR. Alternatively, there are fixed probabilities p of 0.25 and q 0.75. Because of these fixed parameters, it was decided to use the epsilon parameter value 1 for the LocalDP algorithm to give similar circumstances for the technical evaluation.

Hypothesis In this paragraph, we will present the hypotheses that are created for conducting the technical evaluation. As previously explained, the technical evaluation will compare the original dataset and two differential privacy mechanisms for the two utility metrics. The whole point is to understand whether there is a significant difference in terms of a specific utility metric between specific datasets. By having no difference between the original and a permutated dataset, it would mean that the utility with a synthetic dataset is preserved. On the other hand, if there is a significant difference in a utility metric between the original and a permutated dataset, then there is a potential utility issue caused by a local differential privacy algorithm. Below are listed all the hypotheses explained in the paragraph:

Original and permutated LocalDP datasets, utility metric mean:

Hypothesis $\mathbf{H_0}$ (Null hypothesis)**.** *There is no significant difference in utility metric mean between the original and LocalDP permutated datasets.*

Hypothesis $\mathbf{H_1}$ (Alternative hypothesis)**.** *There is a significant difference in utility metric mean between the original and LocalDP permutated datasets.*

Original and permutated RAPPOR datasets, utility metric mean:

Hypothesis $\mathbf{H_0}$ (Null hypothesis)**.** *There is no significant difference in utility metric mean between the original and RAPPOR permutated datasets.*

Hypothesis $\mathbf{H_1}$ (Alternative hypothesis)**.** *There is a significant difference in utility metric mean between the original and RAPPOR permutated datasets.*

LocalDP and RAPPOR permutated datasets, utility metric mean:

Hypothesis $\mathbf{H_0}$ (Null hypothesis)**.** *There is no significant difference in utility metric mean between the LocalDP and RAPPOR permutated datasets.*

Hypothesis $\mathbf{H_1}$ (Alternative hypothesis)**.** *There is a significant difference in utility metric mean between the LocalDP and RAPPOR permutated datasets.*

Original and permutated LocalDP datasets, utility metric variance:

Hypothesis $\mathbf{H_0}$ (Null hypothesis)**.** *There is no significant difference in the utility metric variance between the original and LocalDP permutated datasets.*

Hypothesis $\mathbf{H_1}$ (Alternative hypothesis)**.** *There is a significant difference in the utility metric variance between the original and LocalDP permutated datasets.*

Original and permutated RAPPOR datasets, utility metric variance:

Hypothesis $\mathbf{H_0}$ (Null hypothesis)**.** *There is no significant difference in the utility metric variance between the original and RAPPOR permutated datasets.*

Hypothesis $\mathbf{H_1}$ (Alternative hypothesis)**.** *There is a significant difference in the utility metric variance between the original and RAPPOR permutated datasets.*

LocalDP and RAPPOR permutated datasets, utility metric variance:

Hypothesis $\mathbf{H_0}$ (Null hypothesis)**.** *There is no significant difference in utility metric variance between the LocalDP and RAPPOR permutated datasets.*

Hypothesis $\mathbf{H_1}$ (Alternative hypothesis)**.** *There is a significant difference in the utility metric variance between the LocalDP and RAPPOR permutated datasets.*

5 Evaluation Results

In this section, the focus in on presenting the results that were gained from the evaluation part of the research. Based on four nested levels, each contains a validation method that solves the potential threat of a specific research layer. Thus, for each level, the procedure will be elaborated, analysis will be described and finally, the results will be presented.

5.1 Domain Problem Characterization, Expert Study

As it was explained in the Research Method section, the Expert Study with discussion is conducted to understand whether the problem was well defined and what is the target audience that would benefit from the solution. For each expert, we will explain their thoughts on problem definition and target audence.

Expert Study, Di Wang The first discussion was with the expert Di Wang about whom more information was stated in earlier sections. The discussion started with questioning whether there is a need to merge visual analytics and differential privacy as one privacy-preserving approach that would open a new perspective into understanding how privacy-preserving techniques with noise adjustments affect a specific dataset. The expert emphasized that such an association between these two domains is necessary for society and the academic community. In addition, he stated that from his perspective of knowledge, there are not papers that discuss differential privacy from the visualization perspective. Thus, he concluded that such work contributes to being the starting point of the new approach. The expert stated that visualizations are an efficient way to better understand differential privacy, especially for those inexperienced practitioners with the concept of differential privacy and privacy-preserving techniques. Thus, the target audience should be focused on practitioners whose data is being analyzed. The expert claims as ones out of the privacy-preserving area to be more beneficial than experts. However, the expert states that adding tasks into the visualization system is crucial for engaging experts in the solution. Thus, there is an opportunity to engage experienced groups in the visualization system. In conclusion, we can infer that the expert claims the importance of bringing data visualization domain towards differential privacy to generate a privacy-preserving visual analytics approach that would bring an understanding of privacy techniques toward the target audience.

Expert Study, Andreas Haeberlen The second discussion was with the expert Andreas Haeberlen, for whom there is more background information in the earlier sections. At the beginning of the discussion, the expert emphasized the challenge of explaining differential privacy, noise, and epsilon pa-

rameter to those not experienced with privacy-preserving techniques. Thus, we can infer a problem of bringing the differential privacy and its components to practitioners whose data is being analyzed. In addition, the expert stated the importance of making the whole domain of differential privacy more accessible for actual users and data analysts, especially if their knowledge does not cover differential privacy. Thus, introducing a visual analytics perspective to differential privacy was agreed as a quality approach to bring privacy-preserving techniques to practitioners and data analysts. Making the whole field more accessible is to provide a visualization perspective that helps understand privacy-preserving concepts. For a merge between visual analytics and differential privacy to start a new privacy-preserving approach, it was agreed with the expert that such association contributes to both society and the academic community. The expert agreed that multiple groups could benefit from such a solution in terms of a target audience. First, for the analyst it is vital to understand what amount of noise is expected and how does the epsilon parameter affects its analysis of specific data. Thus, such a visualization system created in the project contributes to data analysts understanding how noise-injection level differs on a specific data that they are analyzing. In terms of practitioners as a target audience, the expert emphasized the importance of what does privacy means to them. Precisely, the question is how do we show differential privacy to practitioners without knowing about it. The solution is to present a simple alternative version of differential privacy supported with the visualization system that not only shows the results but also allows interaction in adjusting the noise amount. In conclusion, it was agreed that such a topic opens opportunities for new approaches and joining two domains that would help different groups of users to work with differential privacy.

Expert Study, Chuhao Wu The last discussion was conducted with the expert Chuhao Wu, about whom there was more information stated in the previous section. For the problem definition, the expert stated that there is absolutely a need for incorporating data visualization tools for differential privacy. While visualizations are needed, the researcher stated that it depends on the demand for a group of users. Since there are various mechanisms and settings of differential privacy, it is reasonable that there are various visualizations that would fit specific needs. In addition, the expert gave an example of a testing hypothesis where visualization without quantitative measurements would not be sufficient. Thus, it is crucial to infer that it is vital to understand the user's need from the visual analytics side despite a necessity to bring the data visualization domain towards differential privacy. In addition, he emphasizes that the main contribution of data visualization would be to help researchers inexperienced with data privacy and differential privacy domains, who have to privatize data for analysis, to

understand how the noise affects their data in both utility and privacy aspects. Without the knowledge of differential privacy or privacy-preserving techniques, researchers have a hard time publishing data. Therefore, we can infer that those researchers and data analysis inexperienced with differential privacy would profit the most from such visualization solutions. However, the expert claimed that the target audience also depends on what kind of visualizations are being provided and for what purpose. In his last project, the expert found that inexperienced users would rather be interested in privacy outcomes than utility metrics. In other terms, those users are more concerned about how differential privacy can secure their data rather than understanding how accurate that data is after permutation. In addition, since those users do not comprehend the advanced characteristics of privacy-preserving outcomes, it is expected that utility metrics would be more challenging for them to interpret. Thus, it is essential to combine both privacy and utility visual analytics metrics that would consider a broader range of audiences, taking a simplified approach so that these permutation results would be apparent to the target audience.

5.2 Data Abstraction Design, Case Study

As it was explained in Research Method section, there are two scenarios that will be conducted for the case study research. For each scenario, we will describe the procedure and outcomes, thus it will result in a self-walkthrough explanatory case study research.

5.2.1 The First Scenario - Healthcare

In the first scenario, the focus is on healthcare domain where we examine the visualization system on a specific healthcare dataset. Detailed explanation of a protocol and dataset is in the case study part of the Research Method section. The idea is to go through the visualization on a specific data in order to discover all features, analyze how the data responds to the system, and visualize the interaction between the dataset and slider that adjust the level of privacy.

Data Import Since the scenario is focused on a specific domain, healthcare, and its dataset about diabetes cases, the first thing after starting the application will be importing our data. As earlier explained, diabetes data presents a dataset with few instances compared to another scenario with many instances in the dataset. The difference in numbers will help us to understand if such variation provides different results and interpretations. Consisting of 768 entries and nine columns, the dataset is uploaded. In order to upload any other dataset, it is mandatory to have numeric data only in CSV file type because the system does not accept any other cases. After the

dataset was imported, we can do some simple data manipulations. These manipulations consist of choosing a specific row range if we are not interested in a whole dataset or choosing specific rows. For the purpose of the case study, we will exclude one specific column Outcome from our further analysis. By clicking on a button for Outcome, the column is removed from use, however, we can include it later anytime, making the system highly interactive. In the figure 21, there is the sidebar used in this first part of the case study. We imported the dataset and did simple data manipulation before doing any visual analysis. Next to the named features of the sidebar, there is also the slider, which is the main feature of the whole system. As explained earlier, the purpose of the slider is to allow interactively adjusting the epsilon parameter on a specific dataset. In our scenario, the slider will adjust the noise-level injection on the diabetes dataset, and we will use it in all upcoming tabs. Despite having additional features on the sidebar, these will be explained in upcoming sections rather than now, mainly because their importance comes at later stages of the case study.

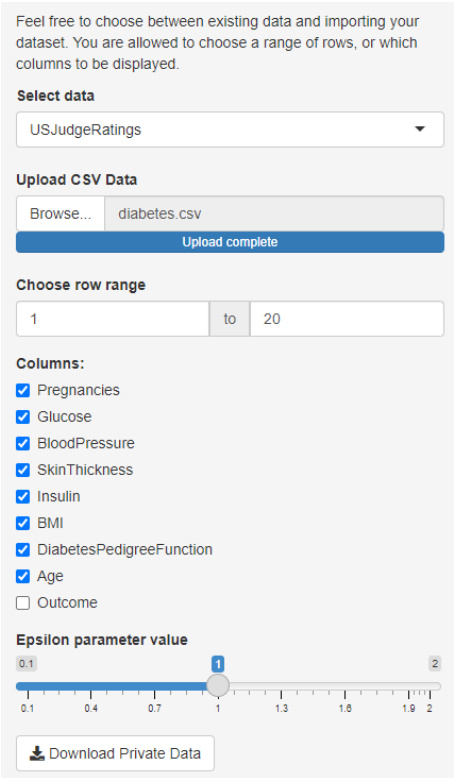


Figure 21: Sidebar, part of the visualization system that allows importing and manipulating data, but also controlling the differential privacy mechanism.

Data Tab The first visible tab while importing data into the system is Data Tab. As described earlier, the purpose of the tab is to introduce a user with data that is planned to be used in the analysis. In our case, the imported dataset is displayed without including the Outcome column that we excluded earlier. The upper table presents the first six rows of a dataset, however, we can decide if other rows are more important to be seen, so such preview could be altered. Additional features of the table are to choose how many rows to display, decide on which group of rows to display, and sort rows by specific columns. Since these adjustments do not significantly impact differential privacy analysis, we will not take long on this feature. However, it does make an impact on data understanding, thus we will sort by column BMI in order to have an understanding of the highest value in the column. By sorting the BMI column, we can see that for six rows with the highest BMI index, the range goes between 52.3 and 67.1. Now, we can use the slider in order to see if there are changes on the data points. The value of the epsilon parameter is w, thus we will move the slider towards a 0.1 value to see if there are changes. As we can see in figure 22, the row numbers changed since the values of data points also changed. Because we decided on sorting by the top six rows with the highest value for the BMI column, we are tracking new rows after adjusting the slider. In addition, now the range for BMI column changed from lower bound 52.3 to 55, and threre are two BMI values of 67.1.

epsilon parameter = 2

Show 6 entries

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
178	0	129	110	46	130	67.1	0.319	26
446	0	180	78	63	66	59.4	2.42	25
126	1	88	80	37	145	55	0.496	26
304	5	115	98	0	0	52.9	0.209	28
104	1	81	72	32	40	52.3	0.283	24
194	11	135	0	0	0	52.3	0.578	40

Showing 1 to 6 of 768 entries Previous 1 2 3 4 5 ... 128 Next

epsilon parameter = 0.1

Show 6 entries

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
178	5	164	64	30	130	67.1	0.66	32
594	4	116	76	21	115	67.1	0.153	25
201	7	71	82	36	250	59.4	0.262	51
446	0	85	60	35	120	59.4	0.201	32
126	7	161	74	30	0	55	0.302	24
484	0	84	58	31	300	55	0.233	24

Showing 1 to 6 of 768 entries Previous 1 2 3 4 5 ... 128 Next

Figure 22: Datatable preview for the diabetes dataset, presented in two conditions: when the epsilon parameter value at 2 and 0.1, each providing top six rows ordered by highest BMI values.

Another feature in the Data tab is the summary statistics preview, which serves for the statistical understanding of a specific dataset. In our case, we can understand the minimum, maximum, average, and quartile values of each column of the dataset by this preview. The figure 24 presents the summary statistics at the slider value 0.1, as we adjusted it earlier for the table preview feature. Now, we want to shift the slider up to the epsilon parameter value 2, where we expect that the summary statistics would give very similar values as the original dataset. The figure 23 presents the summary statistics after adjusting the epsilon parameter towards 2. By comparing the two figures, we can see that each column has differences between the two summary statistics, however these changes are not significantly massive. That is mainly because the algorithm replaces a data point with existing values, thus the permutation is based on replacing rather than inventing new data points. From the statistics summary table, we can briefly infer that the column Insulin contains significantly different mean values before and after permutation, which we will analyze in detail with upcoming tabs.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
Min. : 0.0	Min. : 0	Min. : 0	Min. : 0.0	Min. : 0
1st Qu.: 1.0	1st Qu.:100	1st Qu.: 64	1st Qu.: 7.8	1st Qu.: 0
Median : 3.0	Median :119	Median : 72	Median :24.0	Median : 60
Mean : 4.1	Mean :123	Mean : 69	Mean :22.0	Mean : 94
3rd Qu.: 6.0	3rd Qu.:143	3rd Qu.: 80	3rd Qu.:33.0	3rd Qu.:140
Max. :15.0	Max. :197	Max. :122	Max. :99.0	Max. :846

BMI	DiabetesPedigreeFunction	Age
Min. : 0	Min. :0.084	Min. :21
1st Qu.:27	1st Qu.:0.239	1st Qu.:24
Median :32	Median :0.372	Median :29
Mean :32	Mean :0.471	Mean :33
3rd Qu.:36	3rd Qu.:0.628	3rd Qu.:40
Max. :67	Max. :2.288	Max. :81

Figure 23: Summary statistics of the original data, presenting Minimum, 1st Quartile, Median, Mean, 3rd Quartile and Maximum value for each column.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
Min. : 0.0	Min. : 0	Min. : 0	Min. : 0	Min. : 0
1st Qu.: 1.0	1st Qu.: 99	1st Qu.: 64	1st Qu.: 0	1st Qu.: 0
Median : 3.0	Median :116	Median : 72	Median :23	Median : 36
Mean : 3.9	Mean :121	Mean : 69	Mean :20	Mean : 81
3rd Qu.: 6.0	3rd Qu.:141	3rd Qu.: 80	3rd Qu.:32	3rd Qu.:130
Max. :17.0	Max. :199	Max. :122	Max. :99	Max. :846

BMI	DiabetesPedigreeFunction	Age
Min. : 0	Min. :0.078	Min. :21
1st Qu.:27	1st Qu.:0.240	1st Qu.:24
Median :32	Median :0.370	Median :29
Mean :32	Mean :0.469	Mean :33
3rd Qu.:37	3rd Qu.:0.614	3rd Qu.:41
Max. :67	Max. :2.420	Max. :81

Figure 24: Summary statistics of the permutated data (epsilon = 0.1), presenting Minimum, 1st Quartile, Median, Mean, 3rd Quartile and Maximum value for each column.

Mean Tab We switch to the next tab that focuses on displaying the mean difference for each column between the original and synthetic dataset. By adjusting the slider from 0.1 to 2, we can see how the mean difference changes for each column. Such visualization helps us understand how much differential privacy affects the dataset and how significant is that with each epsilon parameter value. It was decided to shift the slider to three values: 2, 1, and 0.1. Such is to provide the results of having the lowest and highest epsilon parameter value for providing different results on utility and having 1 value that should give results in the middle of two earlier outcomes. The figures 25 present a mean visual plot for each of the sliding situations, where we can see a constant growth of mean difference for almost all columns when shifting from the epsilon value of 2 to 0.1 and stopping at 1. In all three cases, the column Insulin presents the highest mean difference between the original and synthetic dataset, containing the mean difference from 3.45 at the epsilon parameter value 2, going to 5.42 at the epsilon parameter value

1, and ending at 15.47 when the epsilon parameter value is 0.1. The reason for such a significant mean difference is that the column Insulin probably contains values that have substantial variance between rows. If we look into the Data tab again and focus on the Insulin column, we can see that the values mainly present 0 or higher than 100. It means that any permutation could harm utility by creating a more extensive variation between values, resulting in a significant mean change. In addition, by looking at the summary statistics table for the Insulin column, we can see that the first quartile is 0 while the mean and third quartile are 94 and 140, meaning a considerable variation between the column values.

On the other hand, we can see from the figures that the column DiabetesPedigreeFunction has a mean difference of 0.01, 0.02, and 0 for all three epsilon parameter values, meaning almost no change in mean. We can again take a look into the table in the Data tab and realize that the column does not contain significant variation between values, thus the mean could not be affected by the randomized function. While other columns such as Age, BloodPressure, and SkinThickness increase their mean difference score by shifting the slider from the epsilon parameter value 2 to 0.1, there are three columns (Glucose, BMI, Pregnencies), who do not follow the same trend. However, their mean differences are insignificant. The reason for such anomaly is because of the randomness that is presented in the algorithm. Let us consider the mean difference difference one as the threshold of having proper utility with the full privatized dataset. We can conclude that most of the columns did not suffer from high utility issues. However, because the utility of the Insulin column was highly harmed, we should reconsider using the epsilon parameter value 0.1 as the primary focus for exporting a differentially private dataset.

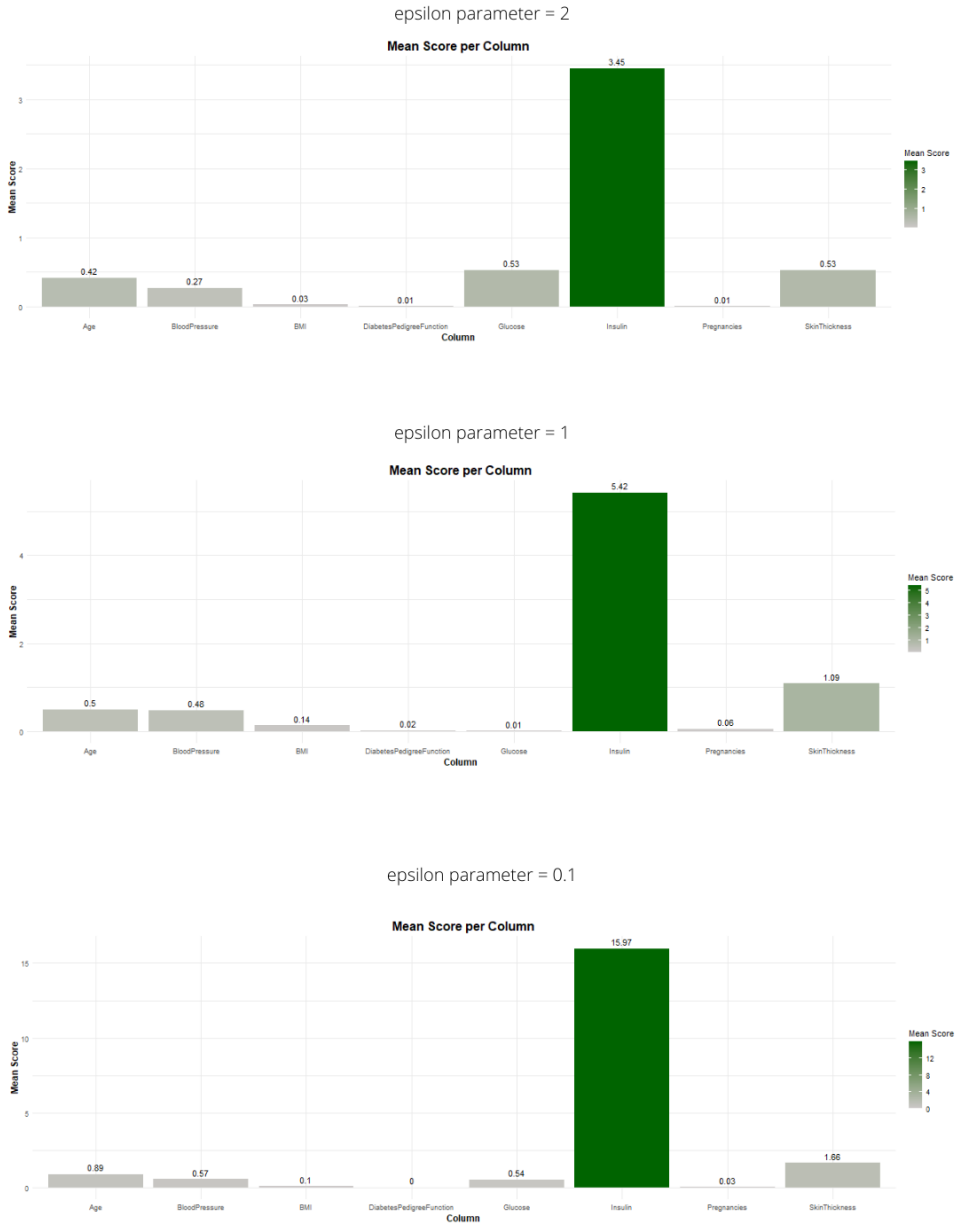


Figure 25: Mean tab, for each epsilon parameter value (2, 1 and 0.1) presenting difference in column means between original and permutated dataset - the highest value shows that for a specific column there is significant difference in means between the original and permutated dataset.

Distribution Tab The third tab, the Distribution tab in the figure 26, presents a plot of distribution for each column. While the previous tab with mean difference score plots served for both visual and mathematical analysis, only visual analytics are considered in this tab. It means that there are no calculations that explicitly present changes or any trend but instead focusing on what visually can be inferred. Given that statement, two epsilon parameter values were taken, 2 and 0.1. The reason for not taking the epsilon parameter value 1 is that the difference between 2 and 0.1 are not that significant, thus 1 will not provide any significant results. By given results, we can infer that for epsilon parameter value 2, there are no distribution differences between the original and synthetic dataset for each column. However, when the slider is adjusted to the epsilon parameter value of 0.1, we can distinguish the distribution difference between the original and synthetic columns of the two datasets for most of the columns. The most significant difference between the two distributions is visible in the Insulin column, however, that is not surprising. As we earlier concluded, the Insulin column was the most affected by the mechanism, the distribution difference only confirms our assumptions.

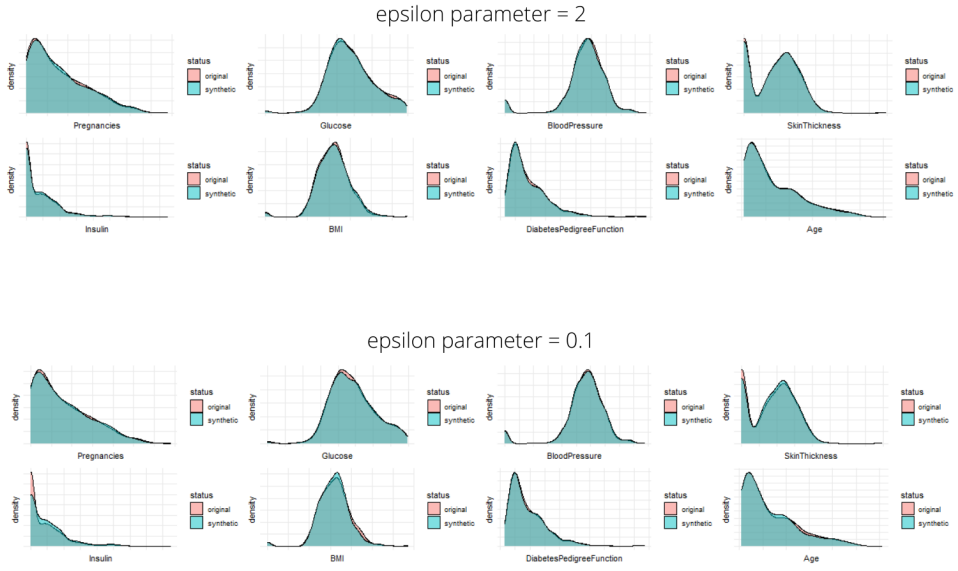


Figure 26: Distribution tab in two settings of the epsilon parameter (2 and 0.1) presents whether there are visual differences in the distribution for each column, and let us visually see that we can keep the distribution even when changing the epsilon parameter.

Euclidean Distance Tab The next tab, Euclidean Distance, is composed similarly to the Mean tab, providing both visual and mathematical analysis of utility. In addition, there is also the calculation of Euclidean distance as a global metric for the whole dataset in this tab. It means that there are Euclidean distance scores between each column from a local perspective and a global metric compared to the original and synthetic datasets. Starting with the epsilon parameter value 2 on figure 27, the average Euclidean distance between original and synthetic datasets is 496.12. There are two significant outliers in the columns, each in its direction, the Insulin column having a significantly high Euclidean distance of 2293.98, and the DiabetesPedigreeFunction with only 6.72. Both columns were already noted as the ones with particular results than other columns.

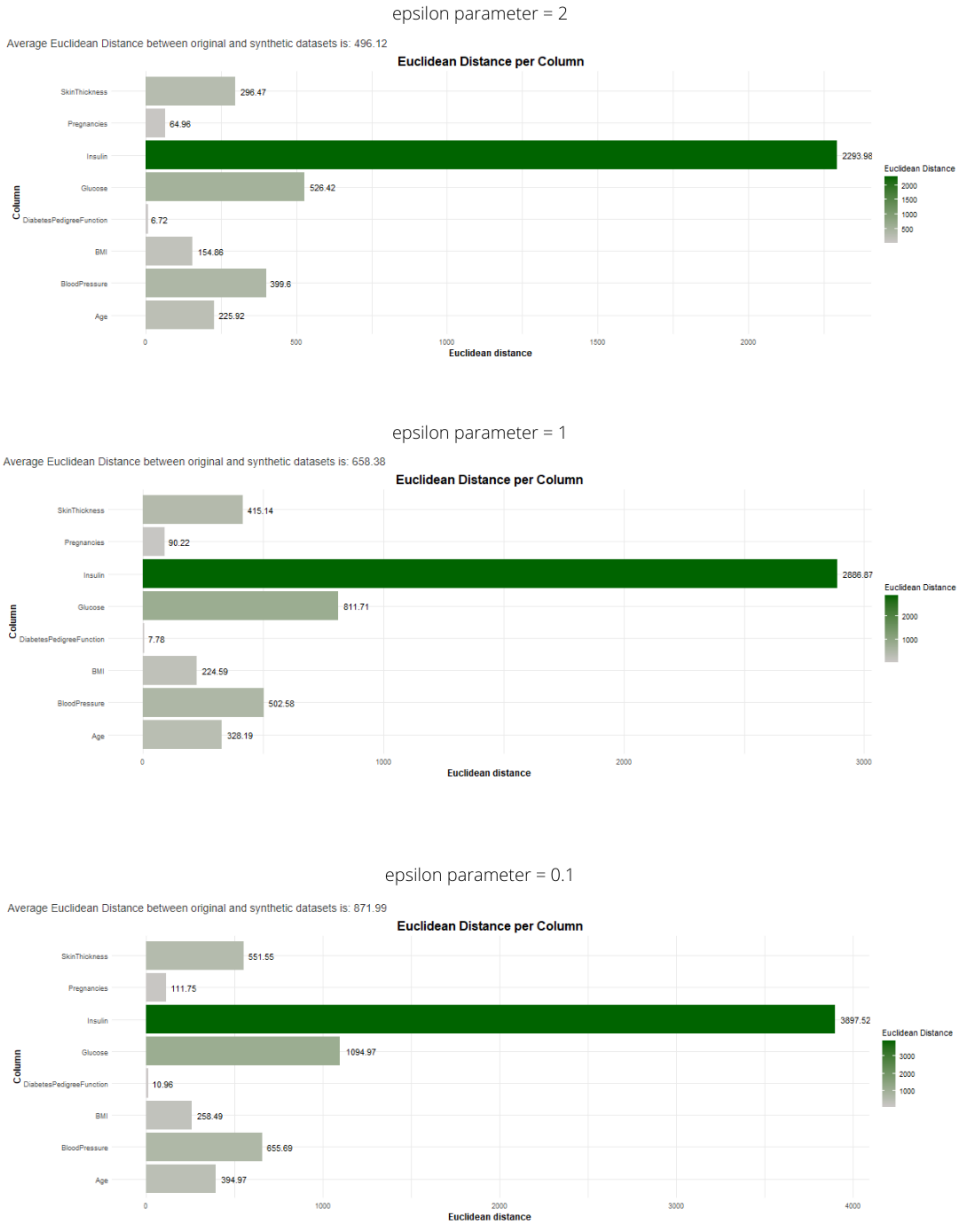
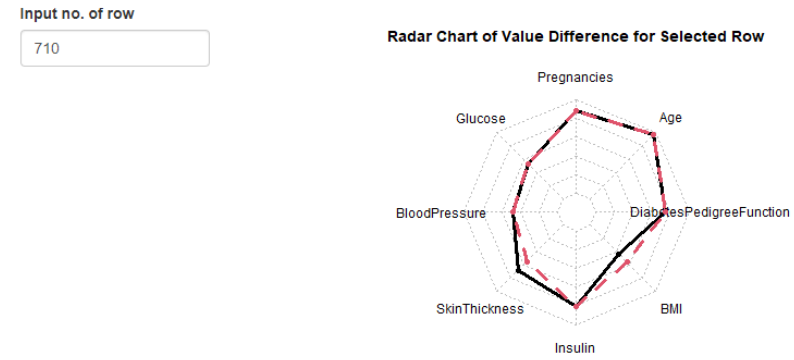


Figure 27: Euclidean distance tab in three conditions (epsilon parameter 2, 1 and 0.1) presents the Euclidean distance score for each column, calculated as average distance for each data point of a column.

By shifting to the epsilon parameter value of 1 on figure 27, we can see that the Euclidean distance of all columns increased compared to values when the slider was at 2. Such was not always the case with the mean difference score, thus this is an excellent example of benefits when having two utility metrics. In addition, it tells us that utility decreases by decreasing the epsilon parameter value, and such a trend will be proven by shifting the slider to 0.1. The average Euclidean distance between original and synthetic datasets for the epsilon parameter value of 1 increased compared to the score for the epsilon parameter value of 2. Again, there are two outlier columns, Insulin and DiabetesPedigreeFunction, that both contribute in different directions. As earlier said, by shifting the slider to the epsilon parameter value of 0.1 on figure 27, we can infer that Euclidean distance scores from both global and local perspectives increase double. Having an average Euclidean distance between original and synthetic datasets with a value of 871.99, again, there are two outlier columns, Insulin and DiabetesPedigreeFunction. We can conclude that the DiabetesPedigreeFunction column showed the best results for utility even with allowing much noise into the dataset, while the column Insulin suffers the most harm of the differential privacy mechanism. Because Euclidean distance as a utility metric presents the calculated distances between two data points and cares about how different they are, such a metric is essential because it shows how accurate a dataset we can get after introducing the differential privacy mechanism in the dataset.

Row Privacy Tab By switching to the next tab, the Row Privacy tab presents analysis from a privacy perspective. Compared to previous tabs with utility metrics, this tab only provides visual analytics. Despite having only a visual analysis of data privacy, we can compare how privacy develops by providing tables of a specific row before and after permutation. Again, three epsilon parameter values will be taken, 2, 1, and 0.1. In addition, we can specify the row or a particular user to be examined, for which we will take row number 710. Starting with the epsilon parameter value of 2 on the figure 28, the radar chart presents no visual differences in data point values between the original and permutated rows. Three table rows are presented below the radar chart to get a more trustworthy insight into these results. The first table row presents each chosen column of earlier specified row data points, and these values do not change with adjusting the slider. The second table row presents columns for the same row after the permutation and sliding, providing interactive results. Comparing these two table rows, we can see that there is difference in three data values, thus we can conclude that the data privacy was not satisfied with the epsilon parameter value being at 2. In addition, the last table row confirms what was inferred, all column values match between the original and permutated

rows.



Row taken from the original dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	2	93	64	32	160	38.00	0.67	23

Row taken from the synthetic dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	2.00	93.00	64.00	45.00	160.00	29.60	0.67	23.00

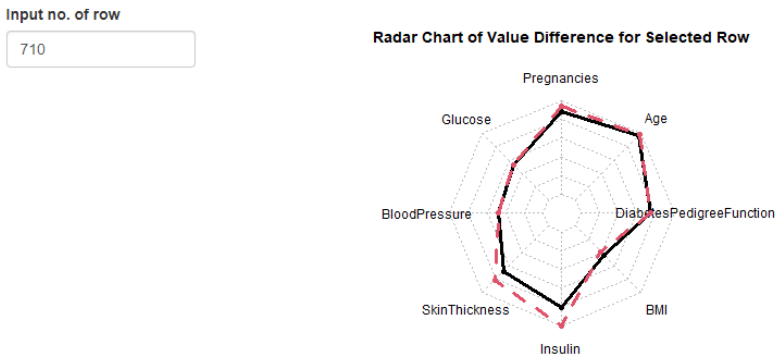
Table presenting changed attribute values of the specific row

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE

Figure 28: Row privacy tab with the epsilon parameter value of 2 for the specific row number 710. Presenting the radar chart that visually shows if there is difference between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

The next epsilon parameter value is 1 for the same row, displayed in the figure 29. Here, we can see on the radar chart that there are five differences between original and synthetic rows, thus the table row inspection will be considered to have a clear understanding of data privacy. By comparing the rows before and after the permutation, there are five columns with different values between rows. The third table row confirms such results, and we can infer that there is the first sight of data privacy on the row. However, is such privatization sufficient for the dataset and the scenario? Since it was stated in the scenario explanation section of the Research Method that privacy is a priority over data utility, there is a need to shift the slider more towards higher data privacy. In addition, there are only eight columns, thus this makes a different perspective on how to privatize the data. Compared to datasets with more than ten columns, where differences between users are more notable, and it is harder to appropriately private their information, in this scenario, hiding more than one-third of the columns should be sufficient

for decent data privacy. However, in the case with the epsilon parameter value 1 for the specific row, more than 50 percent of the data points were permutated, and thus it is sufficient. However, it does not completely guarantee that user could not be guessed from permutated values, thus we should consider adding more noise to preserve data privacy.



Row taken from the original dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	2	93	64	32	160	38.00	0.67	23

Row taken from the synthetic dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	1.00	93.00	64.00	19.00	0.00	40.70	0.67	22.00

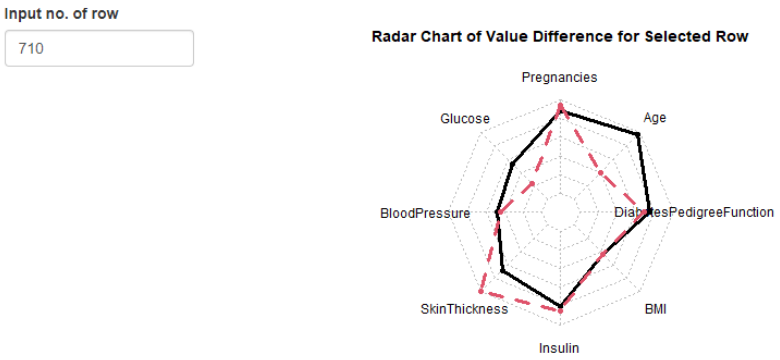
Table presenting changed attribute values of the specific row

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE

Figure 29: Row privacy tab with the epsilon parameter value of 1 for the specific row number 710. Presenting the radar chart that visually shows differences between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

By taking the epsilon parameter value to 0.1, the results of data privacy get significant improvements. As visible in the figure 30, the radar chart shows that now there are significant differences between the original and permutated rows. In order to have a better understanding of these differences, three table rows are further analyzed, which show that five out of eight-column values are changed by sliding the epsilon parameter value to 0.1. By having more than 60 percent of row data points changed, we can infer that data privacy improved, and it is sufficiently presented in the row. Since the scenario expects to have higher privacy than utility, it would be suggested to keep the slider in a range between 0.1 and 0.6 to maintain high privacy. In conclusion, it was shown that despite not having a calculated

a score for privacy, such visual analytics tabs helped in understanding data privacy and how well it was represented in the specific row.



Row taken from the original dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	2	93	64	32	160	38.00	0.67	23

Row taken from the synthetic dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	1.00	154.00	68.00	0.00	126.00	38.00	0.80	57.00

Table presenting changed attribute values of the specific row

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
710	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE

Figure 30: Row privacy tab with the epsilon parameter value of 0.1 for the specific row number 710. Presenting the radar chart that visually shows differences between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

Column Privacy Tab The last tab that is considered for the case study scenario is Column Privacy Tab, presented in the figure 31. As in the previous tab, here the emphasis is on data privacy, however, alongside visual analytics, there is also a calculated score as a global metric of data privacy. While the previous tab was focused on row privacy, where the perspective is taken from a higher level, and it is based on each column of the dataset. While the global metric presents the absolute difference between the original and synthetic dataset by taking the average score between all columns, the local metric is a probability for each column having true or false values. Given that, we can examine what probabilities to have changed values for each column are. Each will be analyzed with global and local calculations by taking three epsilon parameter values, 2 1, and 0.1. Starting with the epsilon parameter value of 2, the total privacy difference score between the original and synthetic dataset is 0.77, meaning that the probability of having

equal data points in both datasets is 77 percent. It means that more than three quarters of the data points have the same values in the original and synthetic dataset, from which we can infer that data privacy is not satisfied in this case.

Figure 31: Column privacy tab, presenting three bar charts for each epsilon parameter value (2, 1 and 0.1). The plot shows percentage ratio between having true and false values for each column.

Moving to the epsilon parameter value 1, there is an improvement in terms of data privacy. However, the global privacy score between the two datasets is 0.57, which still does not guarantee sufficient privacy. In addition, there are columns that provide more privacy than others, the variation ranges between 52 and 61 percent of having true values for each dataset. It is interesting to see that DiabetesPedigreeFunction has the lowest probability of having true values. However, because of its low variance, the utility will not be harmed. Lastly, the slider is adjusted to the epsilon parameter value 0.1, where privacy drastically improves. The global privacy score decreased to 0.31, meaning that 31 percent of the dataset remained unchanged, while the other 69 percent presents different values than their original data points. In terms of columns, the probability of having true or false values gets closer than ever, meaning that specific columns could produce more private data points than earlier cases. However, there is variation in probabilities between columns, having columns with the lower possibility of having true values (Glucose, 28 percent) and other columns with higher probability (Insulin, 38 percent) of having true values. In conclusion, by adjusting the slider towards 0.1, the data privacy increases with the probability of having true values for each column decreasing towards 30 percent, having the chances of permutated than original data points.

Data Export The last feature to be analyzed through the case study is exporting the permutated dataset. Even though the feature can be used at any time while using the visualization system, in the scenario we will use it at the end of analysis. The dataset is being exported as CSV file, thus we can take the original dataset as well and compare the two CSV files in Excel or any other tool. It was decided to export the dataset with the epsilon parameter value of 0.5, mainly because at that specific value, the privacy is not only guaranteed but it also produces expected results, especially since the privacy is priority of the scenario.

Conclusion In conclusion, the case study scenario was focused on using a specific healthcare dataset with a smaller number of instances and columns. Such characteristics allow us to have a perspective on the specific case when utility and privacy are easier affected by the differential privacy mechanism. In addition, by prioritizing data privacy over utility, the premise was to allow more noise and permutation into the dataset while affecting the data value. As a result, the suggested epsilon parameter value ratio would range between 0.1 and 0.5, depending on how accurate data we will want to have while users' privacy is preserved. What makes a potential problem is the utility of the column Insulin that from the start had significantly lower utility than other columns, thus it could be a crucial reason for taking epsilon parameter values closer to 0.6 than 0.1. Because the other columns did not

have a significant accuracy decrease compared to the Insulin column, it is suggested to take the epsilon parameter value of 0.5 as the best ratio of data privacy and utility of the specific dataset.

5.2.2 The Second Scenario - Finance

In the second scenario, the focus is on finance domain where the credit card score dataset is examined. As earlier explained, the protocol goes in a way that each feature of the visualization system is analyzed by using a specific dataset, after which the examination of the results for that dataset is given. The second scenario focuses on a case with financial dataset that contains large number of instances, 30 000, and columns, 25. What makes such a dataset different for case study than the previous scenario is that we could expect less utility harm when the number of instances is bigger. On the other side, 25 columns means that each user could have unique values, which results in a harder guarantee of data privacy. In order to preserve privacy, it would be expected to have at least 35 percent of column values changed, and when there is a large number of columns as in this dataset, that the epsilon parameter value could be probably leaned towards 0.1. In contrary to the previous scenario, when the data privacy was a priority over utility, in this scenario the focus is on preserving the accuracy of the dataset while guaranteeing sufficient data privacy.

Data Import Since the dataset was elaborated earlier, here we will focus on the feature and its outcome. As it was before explained, the first important feature is importing the dataset into the visualization system by using the sidebar. In the figure 32, we can distinguish enormous number of columns, however, the decision is to not take all of them. The reasons are simplifying the analysis and results, but the main point of having a large number of columns will remain with including 17 out of 25 columns. In addition, there are some columns that have identical purpose, thus they might seem redundant, thus we will reduce from these similar columns. Other two columns that were removed from further analysis are ID and default.payment.next.month, because of their irrelevance. By having 17 columns, we managed to retain the bigger number of columns that affect privacy of a dataset, and decrease redundant data. Since the number of 30 000 entries present an important characteristic of the dataset, it was decided not to filter rows.

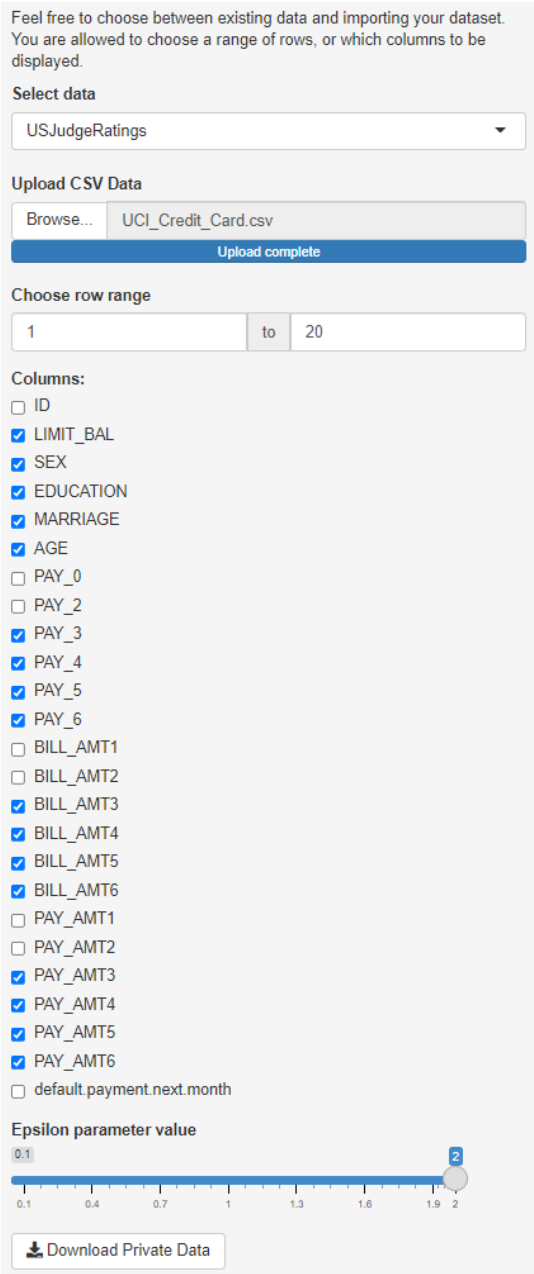


Figure 32: Sidebar, part of the visualization system that allows importing and manipulating data, but also controlling the differential privacy mechanism.

Data Tab In the Data tab, the dataset preview is displayed in the figures 33. Since the features of the table preview were already explained, now we will only discuss the actions that were taken. Among these 17 columns, two columns (SEX and EDUCATION) contain only two values (1 and 2). In addition, the column MARRIAGE contains six values, while columns PAY3

to PAY6 go from -2 to 8, however they contain decimal values. On the contrary to these values with lower variation, the other ten columns contain enormous numbers of values that have significant variation. In terms of table preview, since only six rows were presented in the last scenario, we will now present ten rows. In addition, compared to the previous scenario, there will not be ordered by a specific column, and the reason is that we want to see whether there are changes on the same data points. After switching the slider to the epsilon parameter value 0.1, we can compare these first ten rows to see how they differ. Because table preview serves more for understanding rather than analyzing data, the only thing that we will infer is that there are significant changes of data point values, however, detailed analysis will be given with upcoming tabs.

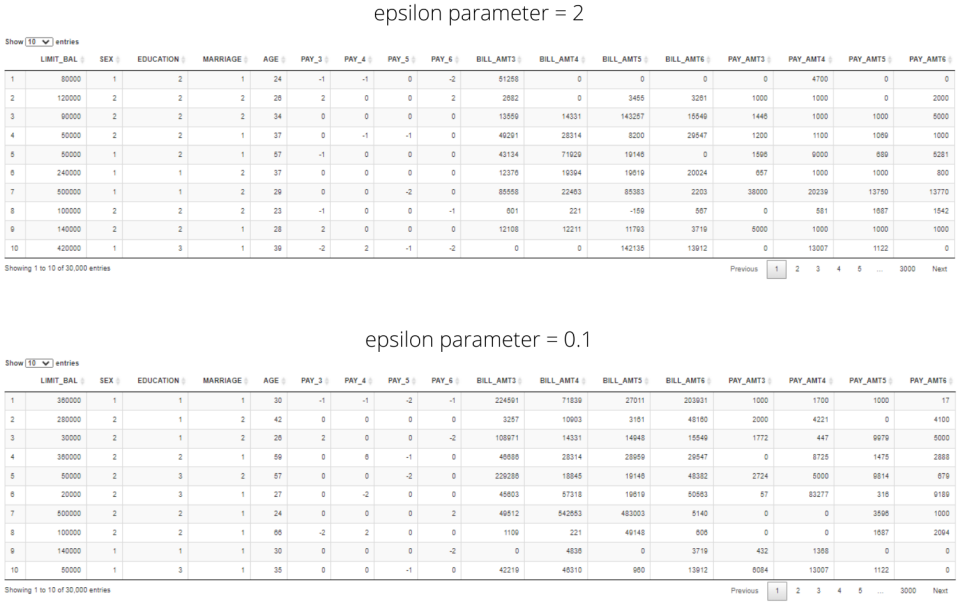


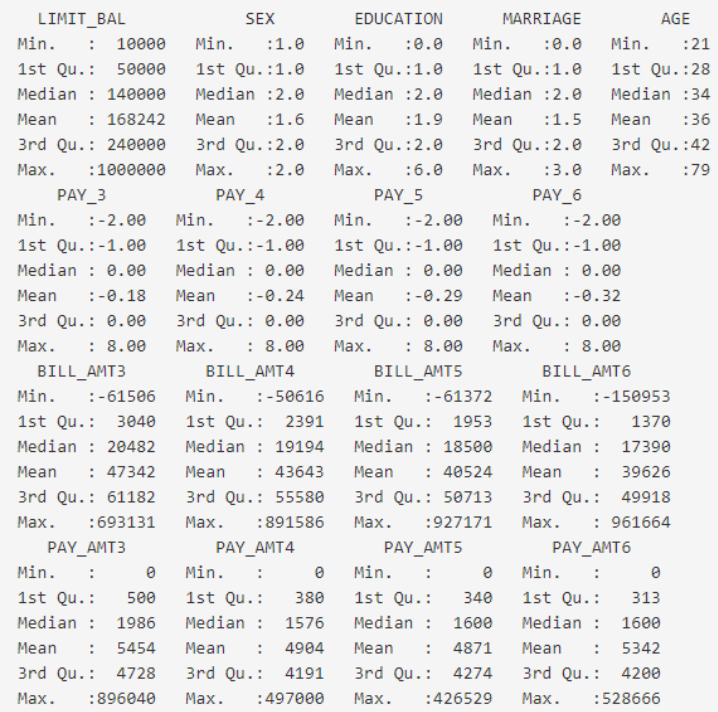
Figure 33: Datatable preview for the credit score dataset. First six rows sorted by ID are presented for both tables, showing the difference when adjusting the slider from the epsilon parameter value 2 to 0.1.

In terms of summary statistics, we can understand how columns differ from each other. As earlier stated, there are attributes with higher or lower variance, minimum, and maximum. Understanding these characteristics of each column will help us to investigate utility in upcoming tabs. Columns such as BILLAMT3 and PAYAMT6 have a significant difference between the 1st Quartile and 3rd Quartile, meaning that their variance is high, thus these columns will be on focus when the utility is examined. By compar-

ing the summary statistics before and after the initial permutation with the epsilon parameter value 0.1, we can briefly conclude that there is a mean difference with previously named columns that suffer from high variance, while those columns with low variance did not have any significant statistical differences after permutation. Because many columns suffer from high variance, the focus on utility as a priority aspect of the tradeoff becomes even more crucial. The figures 34 and 35 present summary statistics of the original and synthetic dataset.

```
   LIMIT_BAL           SEX         EDUCATION       MARRIAGE         AGE
 Min.   :  10000   Min.   :1.0   Min.   :0.0   Min.   :0.0   Min.   :21
 1st Qu.:  50000   1st Qu.:1.0   1st Qu.:1.0   1st Qu.:1.0   1st Qu.:28
 Median : 140000   Median :2.0   Median :2.0   Median :2.0   Median :34
 Mean   : 168186   Mean   :1.6   Mean   :1.9   Mean   :1.5   Mean   :35
 3rd Qu.: 240000   3rd Qu.:2.0   3rd Qu.:2.0   3rd Qu.:2.0   3rd Qu.:41
 Max.   :1000000   Max.   :2.0   Max.   :6.0   Max.   :3.0   Max.   :79
     PAY_3            PAY_4            PAY_5            PAY_6
 Min.   :-2.00   Min.   :-2.00   Min.   :-2.00   Min.   :-2.00
 1st Qu.:-1.00   1st Qu.:-1.00   1st Qu.:-1.00   1st Qu.:-1.00
 Median : 0.00   Median : 0.00   Median : 0.00   Median : 0.00
 Mean   :-0.18   Mean   :-0.23   Mean   :-0.28   Mean   :-0.31
 3rd Qu.: 0.00   3rd Qu.: 0.00   3rd Qu.: 0.00   3rd Qu.: 0.00
 Max.   : 8.00   Max.   : 8.00   Max.   : 8.00   Max.   : 8.00
   BILL_AMT3         BILL_AMT4         BILL_AMT5         BILL_AMT6
 Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603
 1st Qu.:   2627   1st Qu.:   2400   1st Qu.:  1846   1st Qu.:   1271
 Median :  20160   Median :  19140   Median : 18162   Median :  17228
 Mean   :  47089   Mean   :  43374   Mean   : 40220   Mean   :  38969
 3rd Qu.:  60251   3rd Qu.:  54715   3rd Qu.: 50378   3rd Qu.:  49502
 Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664
    PAY_AMT3          PAY_AMT4          PAY_AMT5          PAY_AMT6
 Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0
 1st Qu.:   400   1st Qu.:   307   1st Qu.:   313   1st Qu.:   192
 Median :  1842   Median :  1500   Median :  1542   Median :  1500
 Mean   :  5244   Mean   :  4879   Mean   :  4836   Mean   :  5152
 3rd Qu.:  4603   3rd Qu.:  4055   3rd Qu.:  4160   3rd Qu.:  4006
 Max.   :896040   Max.   :621000   Max.   :426529   Max.   :528666
```

Figure 34: Summary statistics of the original data, presenting Minimum, 1st Quartile, Median, Mean, 3rd Quartile and Maximum value for each column.



```
   LIMIT_BAL           SEX         EDUCATION       MARRIAGE         AGE
Min.   :  10000   Min.   :1.0   Min.   :0.0   Min.   :0.0   Min.   :21
1st Qu.:  50000   1st Qu.:1.0   1st Qu.:1.0   1st Qu.:1.0   1st Qu.:28
Median : 140000   Median :2.0   Median :2.0   Median :2.0   Median :34
Mean   : 168242   Mean   :1.6   Mean   :1.9   Mean   :1.5   Mean   :36
3rd Qu.: 240000   3rd Qu.:2.0   3rd Qu.:2.0   3rd Qu.:2.0   3rd Qu.:42
Max.   :1000000   Max.   :2.0   Max.   :6.0   Max.   :3.0   Max.   :79
     PAY_3            PAY_4            PAY_5            PAY_6
Min.   :-2.00   Min.   :-2.00   Min.   :-2.00   Min.   :-2.00
1st Qu.:-1.00   1st Qu.:-1.00   1st Qu.:-1.00   1st Qu.:-1.00
Median : 0.00   Median : 0.00   Median : 0.00   Median : 0.00
Mean   :-0.18   Mean   :-0.24   Mean   :-0.29   Mean   :-0.32
3rd Qu.: 0.00   3rd Qu.: 0.00   3rd Qu.: 0.00   3rd Qu.: 0.00
Max.   : 8.00   Max.   : 8.00   Max.   : 8.00   Max.   : 8.00
  BILL_AMT3        BILL_AMT4        BILL_AMT5        BILL_AMT6
Min.   :-61506   Min.   :-50616   Min.   :-61372   Min.   :-150953
1st Qu.:  3040   1st Qu.:  2391   1st Qu.:  1953   1st Qu.:   1370
Median : 20482   Median : 19194   Median : 18500   Median :  17390
Mean   : 47342   Mean   : 43643   Mean   : 40524   Mean   :  39626
3rd Qu.: 61182   3rd Qu.: 55580   3rd Qu.: 50713   3rd Qu.:  49918
Max.   :693131   Max.   :891586   Max.   :927171   Max.   : 961664
   PAY_AMT3         PAY_AMT4         PAY_AMT5         PAY_AMT6
Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0
1st Qu.:   500   1st Qu.:   380   1st Qu.:   340   1st Qu.:   313
Median :  1986   Median :  1576   Median :  1600   Median :  1600
Mean   :  5454   Mean   :  4904   Mean   :  4871   Mean   :  5342
3rd Qu.:  4728   3rd Qu.:  4191   3rd Qu.:  4274   3rd Qu.:  4200
Max.   :896040   Max.   :497000   Max.   :426529   Max.   :528666
```

Figure 35: Summary statistics of the permutated data (epsilon = 0.1), presenting Minimum, 1st Quartile, Median, Mean, 3rd Quartile and Maximum value for each column.

Mean Tab In terms of the Mean tab, three epsilon parameter values will be taken: 2, 1, and 0.1, and they are presented in the figure 36. Starting with 2, there is already visible bipolar behavior of columns, having an unchanged mean for columns with lower variance to having significantly different means that already affect the dataset's accuracy. From the beginning, there is a column LIMITBAL that stands out from other columns in terms of the mean difference. Moving to the epsilon parameter value of 1, most of the columns have a growth in the mean difference, meaning that the overall utility decreased with sliding the epsilon parameter towards 0.1. Two columns did not record any mean difference on sliding, these are PAY5 and MARRIAGE, alongside six other columns with a minimum mean difference of 0.1. The only positive note in terms of utility is that, fortunately, some of the problematic columns did not double their mean difference after shifting the epsilon parameter from 2 to 1. Surprisingly, there two columns (PAYAMT4 and 5) that even decreased in the mean difference when shifting to 1, however, such cases are instead a coincidence by the randomness of the differential privacy mechanism. When sliding to the epsilon parameter value 0.1, we can see the mean difference increase for most columns. Again, attributes with higher variance have increased more than double in a mean

difference, which means that their utility is highly harmed. On the contrary, eight columns with lower variance preserved their accuracy efficiently by having a mean difference less than 0.06, which is an excellent result. Because the utility is a priority rather than privacy for the scenario, it would suggest holding with the epsilon parameter values lowest with 1, however, its value has to be further investigated to upcoming utility metric Euclidean distance and privacy metrics.

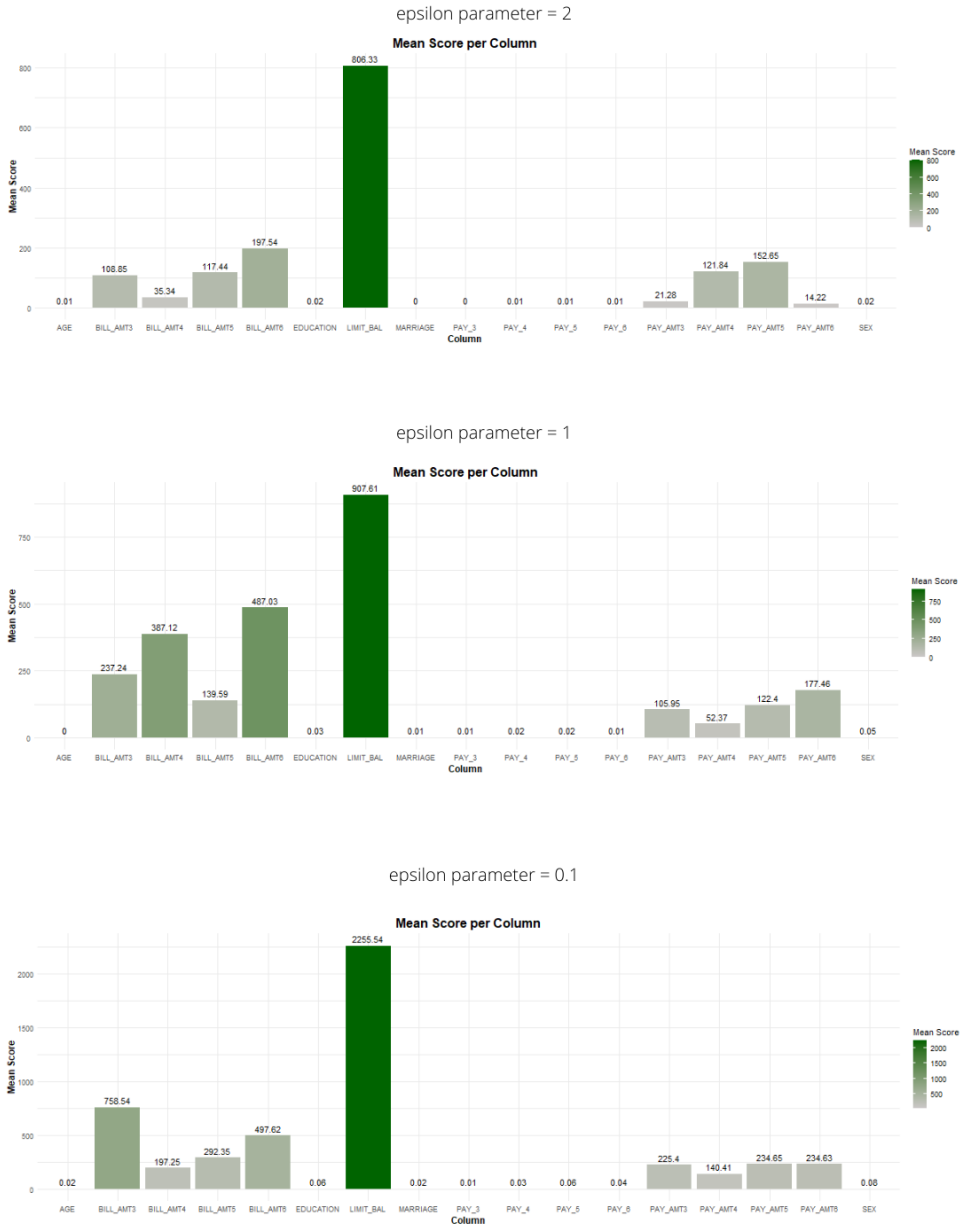


Figure 36: Mean tab, for each epsilon parameter value (2, 1 and 0.1) presenting difference in column means between original and permutated dataset - the highest value shows that for a specific column there is significant difference in means between the original and permutated datase.

Distribution Tab The second tab presents the distribution for each column of the dataset. Since the tab provides only visual analytics without any calculation scores, it is important to have clear illustration of results. It is important to understand that these columns suffer for imbalanced data, which results in oddly shaped distribution graphs. Since there are many columns that show similar statistical results, it was decided for this tab to include only most different columns for the analysis. Starting with the epsilon parameter value 2 in the figure 37, we can infer that there are no differences in distribution between the original and synthetic dataset for any column. Because it is expected that the epsilon parameter value of 1 will not bring any significant difference to be visually analyzed, the next shifting will be towards the epsilon parameter value 0.1 in the figure 37. In this situation, we can infer that for the majority of columns there is a visible difference in distribution between the original and synthetic dataset. Knowing that some columns contain higher variations and bigger maximum values, seeing difference in distributions could mean that their statistical measures after permutation significantly differ.

Figure 37: Distribution tab in the setting of the epsilon parameter 2 and 0.1 presents whether there are visual differences in the distribution for each column, and let us visually see that we can keep the distribution even when changing the epsilon parameter. Only a couple of columns were selected: PAY4, BILLAMT4, LIMITBAL, PAYAMT4, SEX, EDUCATION, MARRIAGE and AGE.

Euclidean Distance Tab Euclidean distance as a utility metric contributes mainly in cases when the mean difference does not provide efficient insights for columns with lower variance. In other words, Euclidean distance will give us better insight into whether is utility better preserved in those columns that have lower variance, and we can also analyze how significant is their Euclidean distance score compared to columns with higher variance. If their difference is not significantly diverse, such results will impact the dataset's utility. By taking three epsilon parameter values, 2, 1, and 0.1, we will investigate how Euclidean difference changes for each column and how it affects data utility. Starting with the epsilon parameter value of 2 in the figure 38, we can infer that there is a bipolar behavior of specific columns. While attributes with higher variance already have higher Euclidean distance even though there are no significant permutations by the algorithm, on the other hand, columns with lower variance tend to preserve their utility by having lower Euclidean distance score. One column, LIMITBAL, stands out as the most affected by the permutation, and its utility is highly harmed. In terms of global metrics, the average Euclidean distance between the original and synthetic dataset is 3 054 433.14. Moving to the epsilon parameter value of 1 in the figure 38, we can deduce that there is an increase of Euclidean distance score for all columns. However, those attributes with lower variance still preserve their utility by having a lower Euclidean score, which is again the opposite case for columns with higher variance. Again, the column LIMITBAL dominates in the Euclidean distance among all columns, which is certainly not a positive trend. In terms of the global metric, the average Euclidean distance between the original and synthetic dataset is 4 433 861, therefore, the value doubled after shifting the slider from the epsilon parameter value 2 to 1. Lastly, the slider is adjusted to the epsilon parameter value 0.1 in the figure 38, where it is expected to have the worst results for the data utility of the dataset. As expected, Euclidean distance increase among all columns continued, and the column LIMITBAL still dominates over other attributes. However, seven columns with lower variation result in surprisingly lower utility. The average Euclidean distance between the original and permutated datasets is 5 531 091.77. In conclusion, because of attributes with higher variance, utility at the epsilon parameter value 0.1 is highly harmed, thus it would be wise to consider shifting it back towards value 1.

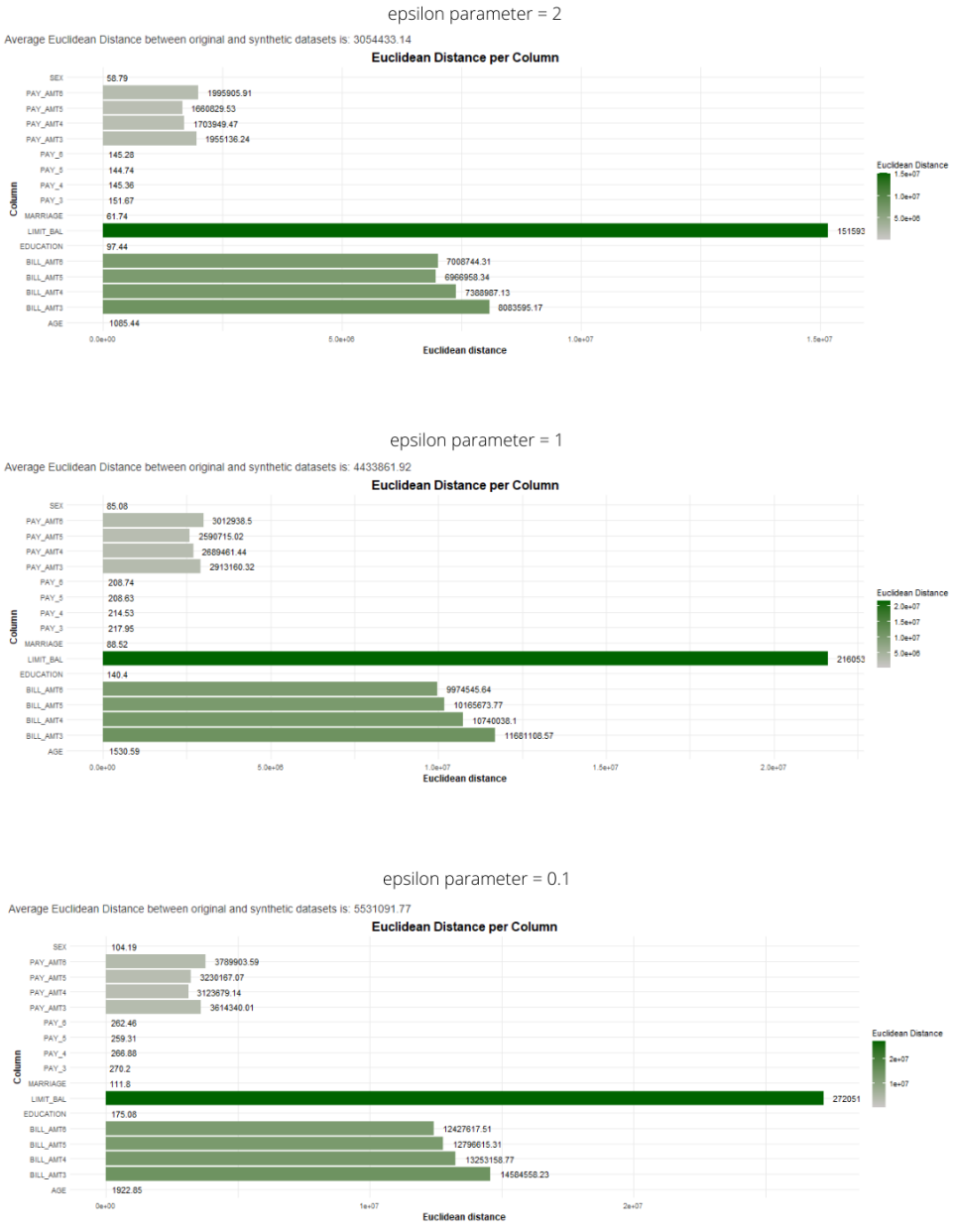


Figure 38: Euclidean distance tab in three conditions (epsilon parameter 2, 1 and 0.1) presents the Euclidean distance score for each column, calculated as average distance for each data point of a column.

Row Privacy Tab Switching from utility to privacy tabs, the first one is Row Privacy, where the visual analytics approach is only considered to analyze how the mechanism affects a specific row. Because the data utility is a priority over privacy for the scenario, the goal is to satisfy the general requirement to hide around 50 percent of all dataset columns. It was randomly decided to take row number 10 586 as the user that will be investigated for the tab. Starting with the slider at the epsilon parameter value 2 in the figure 39, we can see on the radar chart that there are only four data value differences for the specific row between the original and synthetic dataset. To prove the statement, we will compare the first two table rows to see any differences between them and check the third table row if it shows any false values. As expected, only four columns in both table rows are showing different values, and the third table row presents mostly true values. It means that for the epsilon parameter value of 2, there are no significant differences of the row between the original and permutated dataset, which results in unsatisfied data privacy.

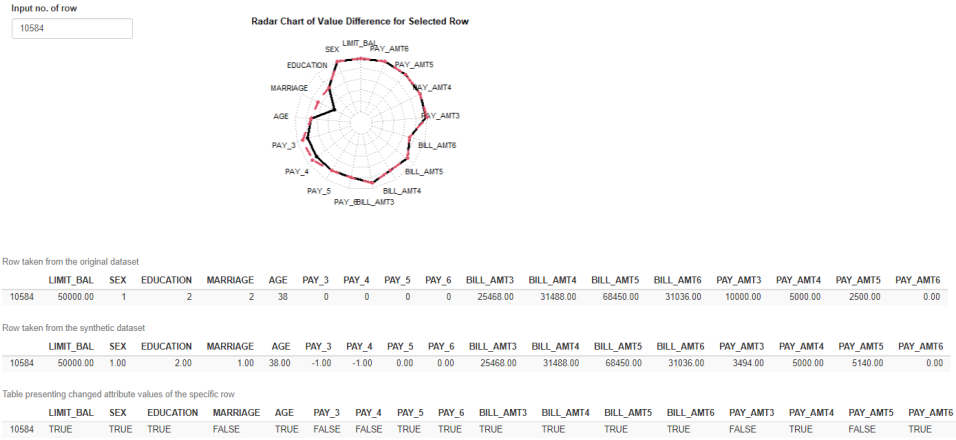
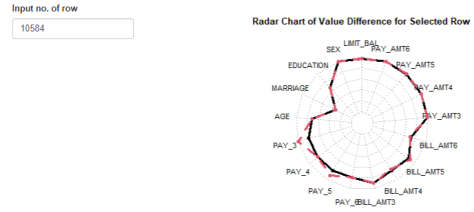


Figure 39: Row privacy tab with the epsilon parameter value of 2 for the specific row number 10586. Presenting the radar chart that visually shows if there is difference between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

Moving to the epsilon parameter value 1, we can see in the figure 40 that differences exist between the original and synthetic dataset on the radar chart. By taking a closer look into these differences on the three table rows, we can infer that there is eight columns with changed values, thus the data privacy is probably satisfied. The third table row shows that only three columns have changed their values, which means that roughly 47 percent of all column values were changed. We can conclude that there is more difference by looking at the three table rows than when the epsilon parameter value was 2. However, there is no sufficient guarantee of privacy for a specific

row. Therefore, it would be suggested to shift the epsilon parameter slide slightly towards 0.1 value.



Row taken from the original dataset

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
10584	50000.00	1	2	2	38	0	0	0	0	25468.00	31488.00	68450.00	31036.00	10000.00	5000.00	2500.00	0.00

Row taken from the synthetic dataset

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
10584	50000.00	1.00	2.00	2.00	38.00	-2.00	0.00	-1.00	0.00	25468.00	62470.00	28718.00	46514.00	12000.00	4000.00	1000.00	0.00

Table presenting changed attribute values of the specific row

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
10584	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

Figure 40: Row privacy tab with the epsilon parameter value of 1 for the specific row number 10586. Presenting the radar chart that visually shows if there is difference between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

The last adjustment of the slider ends with the epsilon parameter value being at 0.1 in the figure 41. As expected, there are significant differences between the two datasets shown on the radar chart for the specific row. By looking at the three table rows, we can infer that there are more changed columns than the actual values. Precisely, 14 out of 17 columns have changed their values, meaning that around 82 percent of all columns are changed. Such results guarantee data privacy, however, its utility is deeply questioned, as shown in earlier tabs. Because the utility has priority over data privacy in the scenario, taking the epsilon parameter value of 0.1 is not an option for the final results.

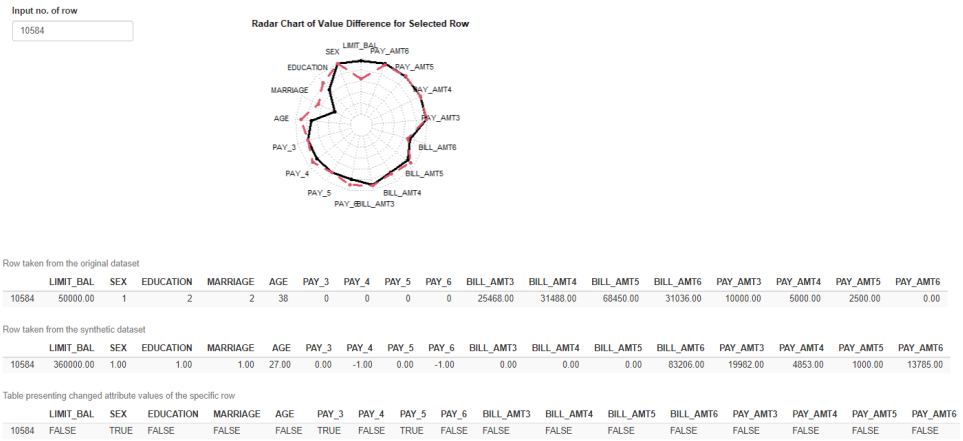


Figure 41: Row privacy tab with the epsilon parameter value of 0.1 for the specific row number 10586. Presenting the radar chart that visually shows if there is difference between column values, and three table rows that show what are column values of the specific row and if they differ (the last table row).

Column Privacy Tab The last tab again focuses on data privacy, however, the approach takes a higher perspective by looking at the whole dataset and its columns. In addition, with visual analytics, there is also a calculated score for the whole dataset. Starting with the epsilon parameter value of 2 in the figure 42, the global metric is an absolute difference between the original and private dataset as the probability of having actual values, with the score of 0.98. It tells us that there are almost no false values in the dataset, which certainly means that privacy is not guaranteed. The bar chart presenting the probability for each column shows no significant difference in probability among these attributes of the dataset. For the epsilon parameter value of 1, the figure 42 shows that the global score is 0.61, which tells us that there are still more true values between data points. However, the probability of having original values is rapidly decreasing, meaning that the dataset is being privatized. There are variations in the probabilities among attributes in terms of individual columns, and the variance is 22 percent. The last epsilon parameter value, 0.1, is shown in the figure 42. With a global score of 0.4, we can infer that a substantial proportion of data points are replaced with false values, which guarantees data privacy. In addition, there is a high variation in probability between attributes, having the columns MARRIAGE and BILLAMT3 that differ by 35 percent. Such difference tells us that those columns with higher variation within their values have a lower percentage of having true values, while columns with lower variation within its values have a higher probability of having true values after permutation.

Figure 42: Column privacy tab, presenting three bar charts for each epsilon parameter value (2, 1 and 0.1). The plot shows percentage ratio between having true and false values for each column.

Data Export After the tabs have been analyzed, the data with permutated values is being exported in the CSV file. Because the utility is a priority over data utility for the scenario, it was decided to take the epsilon parameter value of 1 to guarantee general and sufficient privacy while preserving the accuracy of the dataset.

Conclusion In the end, the second scenario was investigated in detail with the specific dataset. With priority given to the utility over data privacy, the

analysis was directed to preserve the accuracy of the dataset while providing minimal data privacy. By interacting with the visualization system, the result was to export the dataset with the epsilon parameter value of 1.1. The reason for such value is that it provides sufficient data privacy while the utility is preserved as much as possible.

5.3 Visual Encoding and Interaction Design, Expert Study

For the Visual Encoding and Interaction Design, the evaluation was based on Expert Study, where four researchers from the domain of differential privacy were interviewed. In open-questions discussion, the experts provided their thoughts on the interface and reflecting on its usability.

Expert Study, Di Wang The first expert, Di Wang, had positive thoughts on the visualization system. In general, he stated that the system shows its quality by providing utility and privacy metrics while it also allows data manipulations. For one of the system's essential features, the slider, the expert states positive comments as a decent solution to allow adjusting the noise-injection level. In addition, by having an interactive and responsive system that instantly shows visual results on specific data, the slider has a vital contribution to the research. By taking a look at each tab, the first Data tab shows a decent preview of what data is being used and what are its summary statistics. Especially the second feature of the tab, summary statistics, was stated by the expert as an essential feature to understand a specific dataset that is being used. Moving to the Mean tab, the expert stated the mean difference as an essential metric for a utility that statistically shows the difference between original and synthetic columns. In addition, the bar chart presents the mean differences for each column of a dataset. The next tab, Distribution, was stated as the most important presentation of the comparison between original and synthetic columns, especially since such visualization is the most used by differential privacy researchers. In addition, by showing all distributions in one tab was declared as a sufficiently presented metric even though it contains only visual analytics without a calculated score. The next tab, Euclidean distance, did not get positive feedback as previous utility metrics. Despite showing the correct calculations present on the proper visualization chart, the expert emphasized that such a metric is not usually used in the community, thus he would not recommend relying on it. Moving to the privacy metric, we started with the Row privacy tab, which the expert stated as a sufficient implementation for showing in-depth investigation of how data is privatized for a specific row. However, it was recommended to add the feature of choosing a range of rows to be shown instead of showing only one row. The last tab, Column privacy, was also seen as a beneficial presentation of privacy metrics from a higher perspective than the previous tab, and the expert did not emphasize

any issues with it. In terms of available opportunities, the expert stated adding top K items and machine learning analysis such as linear regression as a solid progress of the visualization system, which would bring it closer to the experts. The expert concluded the discussion by emphasizing such a visualization system as a decent solution to allow inexperienced users to use the visualization system, and from this perspective, the system is more than sufficient.

Expert Study, Andreas Haeberlen The second expert, Andreas Haeberlen, gave his thoughts on the visualization system. In general, the expert stated it as a great approach that impacts differential privacy, being extremely useful for people that are not trained in differential privacy. Thus we could infer that he proposes that practitioners without differential privacy experience would benefit the most from such a solution. In addition, he stated the system as an excellent starting point for developing a visualization system with additional algorithms that could be more completed so that they would interest experts. Moreover, by having these visualizers to teach people what happens to differential privacy is the main contribution of the system. In terms of features of the visualization system, the expert was focused on giving ideas for further progress, while the overall opinion was that the system offers many useful features that contribute from different perspectives. As one general opportunity that would affect the whole system, the expert suggested having each epsilon parameter run multiple times instead of showing results for only one run. Such an idea lies in a randomized function, which constantly affects the results even for the same epsilon parameter, thus having it run multiple times would give better insight into a specific epsilon parameter value presenting. In addition, it would show a range of outcomes to expect and understand how much the algorithm relies on randomness. In terms of sidebar features, data manipulation with choosing specific columns and a range of rows is emphasized as great features to allow users to choose how large data they want to use in the analysis and how the results differ when several instances of a dataset change. Starting with the Mean tab, it was suggested by the expert to analyze each column as its segment on its scale so that it would be easier to follow what the difference is for the mean difference. Thus, the bar chart would be divided as to its own for each of the columns. In terms of mean difference as a utility metric, it is sufficient for data analysts to understand the statistical differences between the original and permutated datasets. Moving to the Distribution tab, the expert stated that it is needed visual analytics that makes user curious why the two distributions of original and synthetic datasets for a specific column differ, and it would drag a user to go back and analyze data. However, adding axis labels and three curves, having upper, lower, and real curve, would give a better insight into the results. Again,

the Euclidean distance tab was commented as the least intuitive metric, primarily because of outliers that could affect the Euclidean distance score. In addition, it was suggested to try to implement a scatter plot, however this would be only visually possible when a dataset contains a lower number of instances. Moving to privacy metrics, the Row privacy tab with its radar chart seems as a visually exciting presentation of changes between original and synthetic datasets, however, the shape of the chart does not present any semantic meaning. In conclusion, even though there are opportunities to improve specific features of the visualization system, the expert emphasized the importance of creating such solutions to invoke the association between differential privacy and data visualizations. Such a system was stated as an essential starting point to build additional visualization interfaces that would contribute to all three groups of users: experts, data analysts, and practitioners.

Expert Study, Chuhao Wu The last expert, Chuhao Wu, shared his opinion on the visualization system. In general, the visualization system received very positive feedback, especially for its usefulness. The expert's opinion on the target audience for such a system was directed toward both researchers without profound experience with differential privacy and data privacy and regular users whose data is harmed. Starting with the sidebar, the feature of including any quantitative dataset was praised by the expert for allowing various data to be analyzed. In terms of the main feature of the system and the sidebar, the epsilon parameter slider, it was suggested to highlight it in order to show inexperienced users the importance of such a feature. The first tab, Data, received positive comments as the introduction tab to see what data is being analyzed. In addition, the summary statistics table also shows exciting insights on a dataset. Moving to the Mean tab, while the mean difference is a valuable utility metric, the expert stated that it would be better to present each column as a separate plot. In addition, he suggested showing two bars for each analysis: the original and synthetic column. However, such an approach was tried earlier and turned out to be insufficient because if the mean difference is relatively small, then it is hard to depict how significant that difference is. The expert praised the approach by showing the distributions of all columns and providing such visual utility analysis in terms of the Distribution tab. Therefore, this was chosen as the best solution for utility visualization of the system. Next was Euclidean distance, for which the expert stated as a helpful utility metric, however, he suggested showing the Euclidean distance score of each column as a box plot rather than one bar chart. Moving to privacy metrics, the overall opinion on the Row privacy tab was highly positive as it gives excellent insight into how data values change with adjusting the epsilon parameter. Therefore, the expert declared the tab beneficial for users concerned with privacy and

understanding how the algorithm affects a specific row. In addition, such a tab shows whether it is possible to detect a user by looking at its attributes. The radar chart was also found helpful in order to understand for which column values of a specific row there was noise, however, it was suggested to include a description or a legend in order to explain the plot to users inexperienced with advanced data visualizations. The last tab, Column privacy, has also been praised as an exciting solution to understanding privacy from a higher-level perspective. In conclusion, the visualization system received positive feedback from the expert, especially from the perspective of showing the right insights for adjusting the noise-injection level.

5.4 Algorithm Design, Technical Evaluation

As elaborated in the Research Method section under the Algorithm Design, the technical evaluation is conducted on three datasets: the original and two permutated datasets from LocalDP and RAPPOR algorithms. By comparing the two permutated datasets with global utility metrics on the original dataset, we can infer whether the utility was preserved while privacy was adjusted. Each algorithm offers different mechanisms within local differential privacy, thus comparing the algorithm of the project with the state-of-the-art on the original dataset will present how well the LocalDP algorithm performs. By using mean and variance as utility metrics, the comparison was conducted with two-sample t-tests for three cases: original dataset and LocalDP permutated dataset, and original dataset and RAPPOR permutated dataset, and the third case when the comparison is between RAPPOR and LocalDP permutated datasets. In all three cases and both utility metrics, the confidence interval is 0.05.

Starting with mean as the utility metric and its first case, the results of two-sample t-test show that there is no significant difference in means between the original dataset ($M = 1024.426$, $SD = 355032.289$) and the LocalDP permutated dataset ($M = 1024.425$, $SD = 355019.773$) with $t(126) = 1.979$, $p = 0.999$. Thus, the null hypothesis with means of two datasets being equal is accepted, and the first hypothesis of two datasets having means that differ is rejected. In the second case with mean utility metric, the results of two-sample t-test show that there is also no significant difference between the original dataset ($M = 1024.426$, $SD = 355032.289$) and the RAPPOR permutated dataset ($M = 1024.521$, $SD = 355061.334$) with $t(126) = 1.979$, $p = 0.999$. Thus, the null hypothesis with means of two datasets being equal is accepted, and the first hypothesis of two datasets having means that differ is rejected. For the third case between the two permutated datasets, the results of the two-sample t-test show that there is no significant difference in mean between the RAPPOR permutated dataset ($M = 1024.521$, $SD = 355061.334$), and the LocalDP permutated dataset ($M = 1024.425$, $SD = 355019.773$) with $t(126) = 1.979$, $p = 0.999$. Thus, the null hypothesis with

means of two datasets being equal is accepted, and the first hypothesis of two datasets having means that differ is rejected.

Moving to the variance as the utility metric, the results of the first case with two-sample t-test show that there is no significant difference in variance between the original dataset (M = 85.431, SD = 67.004) and the LocalDP permutated dataset (M = 84.915, SD = 68.806) with t(126) = 1.979, p = 0.724. Thus, the null hypothesis with the variance of two datasets being equal is accepted, and the first hypothesis of two datasets having means that differ is rejected. For the second case, the results with two-sample t-test show that there is no significant difference in variance between the original dataset (M = 85.431, SD = 67.004) and the RAPPOR permutated dataset (M = 85.789, SD = 35.726), with t(115) = 1.981, p = 0.634. Thus, the null hypothesis with the variance of two datasets being equal is accepted, and the first hypothesis of two datasets having means that differ is rejected. For the third case between the two permutated datasets, the results of the two-sample t-test show that there is no significant difference in variance between the RAPPOR permutated dataset (M = 85.789, SD = 35.726), and the LocalDP permutated dataset (M = 84.915, SD = 68.806) with t(115) = 1.981, p = 0.496. Thus, the null hypothesis of the varience of two datasets being equal is accepted, and the first hypothesis of two datasets having means that differ is rejected.

6 Discussion

In the section, the general conclusion on the whole project will be given. In addition, by covering the specifics of the research method, inferences will be explained in order to have a clear understanding of the gained knowledge.

6.1 General

In general, there are a couple of inferences that were gained from working on the project. The first is focused on the columns of a dataset, for which if the data type is boolean, the probability of being True or False with a higher epsilon parameter will not advance for any of the two possibilities. In other words, both options will be around 50 percent, which means that when we have fewer different options to use for creating noise, the chances are smaller than it will create noise. The same concept repeats with the gender column and any classification that contains only two options. On the contrary, if a column has diverse values, the chances of having noise are higher. However, having more values is also related to the variance of a column, which will be elaborated on in the upcoming sections. The second inference is that despite adjusting the epsilon parameter and putting it back to the previous value, it does not mean that we will get completely the same results. The reason for this lies in fact due to the randomization function included in the mechanism, it is impossible to get the same results twice for the same epsilon value. Lastly, as it was stated in the Algorithmic Foundation of Differential Privacy work by Aaron Roth and Cynthia Dwork [19], data cannot be anonymized entirely and still preserve its utility. Thus the tradeoff between data privacy and utility is necessary for guaranteeing privacy while remaining the dataset value. The conclusion is that the tradeoff always requires compromise and adjustments in order to accomplish satisfaction for both sides.

6.2 Expert Study Discussion

The first expert study brought positive thoughts from the experts on merging differential privacy and data visualization. In addition, visual analytics as a solution to analyze the impact of noise on the tradeoff between data privacy and utility has been recognized as an essential contribution to society and the academic community. It was stated that despite a few papers which considered similar attempts, there was no direct contribution that associated differential privacy with data visualization in a way that it would bring an understanding of such complex privacy-preserving technique by visualizing its analysis and results. All three experts agreed that there is a need for introducing visualization on differential privacy, however, there were different perspectives on its purpose. However, the general opinion was that such

association would bring understanding to a specific group of people, which are users inexperienced with differential privacy. The first group of users are practitioners whose private information is part of a dataset that is being analyzed, therefore they want to understand how does the noise-injection level affects their information. The second group of users are researchers, especially from the social sciences domain, who do not focus their expertise on privacy-preserving techniques, however, the data that they are using for analysis is being privatized. Thus, to preserve as much utility as possible, they would be interested in understanding how to affect the tradeoff between data privacy and utility. In addition, it is expected by the experts that the more complex the algorithm and visualizations get, they would instead attract data privacy experts than inexperienced users. On the contrary, by providing a simple and understanding solution, these inexperienced users would comprehend noise-injection level affect on a specific dataset. In conclusion, the problem definition as defining the right problem of providing visualizations to differential privacy to users experience the notion of the tradeoff between privacy and utility was successfully evaluated. The experts recognized such association of differential privacy and data visualization in order to start the privacy-preserving visual analytics as an essential contribution to the academic community. In addition, we have received opinions on which specific groups of users would most benefit from such a solution and what is their primary demand for the visualization system.

6.3 Case Study Discussion

The case study helped investigate how noise-injection level affects the data utility and how data characteristics respond differently to the adjustment of the epsilon parameter. By looking at the summary statistics of the Data tab, we could see that there are columns with higher and lower variance. By analyzing on other tabs how data is changing while adjusting the epsilon parameter, it was discovered that those columns with higher variance suffered from higher scores for mean and Euclidean distance, which results in lower utility. On the contrary, other columns with lower variance gave highly satisfying results for all utility metrics, thus the utility was not harmed in these columns. Thus, we can conclude that privatizing columns with a lower variance will better impact a dataset utility than permutating columns with higher variance. In terms of the epsilon parameter, the conclusion is that there is always a small range of epsilon parameter values that would provide efficient data privacy and utility for a specific dataset. The reason lies in the randomized function that disables the possibility of always having the same results. Thus, because of likely similar or different results for computing the same epsilon parameter value for a couple of times, taking a small range, e.g., between 0.1 and 0.6, given an opportunity to always satisfy expectations for the randomize function. Overall, it is essential to understand what

kind of data is being analyzed in order to comprehend what the results are presenting thoroughly. Each dataset requires an individual approach for the tradeoff between data privacy and utility, thus adjusting the epsilon parameter will be determined by the characteristics of a dataset and its need to be less or more privatized.

6.4 Expert Study Discussion

The second expert study showed positive feedback on the visual solution created in the research. All three experts emphasized that such a visualization system is a great starting point to inspire other researchers to develop solutions that would bring differential privacy and data visualization closer to creating privacy-preserving visual analytics. In addition, the solution was recognized as a visual analytics approach to explaining the effect of noise-injection levels on a specific dataset. The experts agreed that such a visualization system offers many opportunities for further progress, especially if there would be a demand for specific visualizations for a different algorithm. One expert suggested taking a different approach to providing insights on privacy and utility with calculating results for more than one run and getting average, upper, and lower bounds of results. This would give a clear understanding of how in general, a specific noise-injection level affects a dataset. In terms of the main feature of the visualization system, the slider was praised by all three researchers as an acceptable way to adjust the epsilon parameter for getting instant results. The tab that all three experts agreed on as beneficial is Distribution, and the reason for positive feedback lies in the fact that similar visualizations were previously used by other researchers, with the histogram as the most common solution. In addition, the Distribution tab would be helpful even if the mechanism in the system would be changed, or there would be different data in it. Because not every visualization fits all expectations and different groups of potential users, experts suggest some suggestions for future improvement of the system. However, not all three experts agreed on the same improvements, therefore, we will mention ones that were the most represented and valuable. While two experts agreed that the Euclidean distance would not be a suitable utility metric for calculating the value of a specific dataset, the third expert praised the measurement as especially helpful when the system deals with a numeric dataset. For other tabs, visualizations, and features, each expert proposed various minor suggestions to bring the system more towards a specific target audience. In conclusion, the visualization system was recognized as an impactful starting point to bring data visualization closer to differential privacy and open privacy-preserving visual analytics towards inexperienced users, which was the main point of the first expert study. Therefore, the visualization system contributes to the association between the domains and their target audience.

6.5 Technical Evaluation Discussion

The Technical Evaluation provided an analysis of the performance of the created algorithm. In order to conduct such an evaluation, the algorithm was given a specific dataset that was permutated with the epsilon parameter value of 1.1. The permutated version of the dataset was compared to the original version by taking utility metrics mean and variance for each column. To properly evaluate the algorithm's performance, the same permutation technique and analysis on the original dataset was conducted to the state-of-the-art algorithm, Google's RAPPOR. By comparing the results of both algorithms and comparing their difference for both utility metrics, we understood how they perform. The results showed that there was no significant difference between datasets in all three cases for both utility metrics. It means that the original dataset does not have a significantly different utility score for both mean and variance compared to the permutated LocalDP and RAPPOR datasets. Thus, both datasets presented a decent performance in terms of data utility. Additionally, since there was no significant difference for both mean and variance between the two permutated datasets, they both perform equally in terms of data utility. It means that the algorithm created in the project stands well alongside the state-of-the-art algorithm for local differential privacy. In conclusion, despite adjusting the epsilon parameter value to 1.1, which would refer to leaning towards data privacy rather than utility, the results show that the mechanism still stands impressively in terms of utility alongside the original dataset and the state-of-the-art algorithm. If the epsilon parameter is adjusted more towards 0, then there could be a significantly different utility score against the original and RAPPOR permutated. However, by adjusting the epsilon parameter value, the conditions for both algorithms would not be the same anymore, thus the comparison would not be conducted appropriately. Thus, we can infer that the LocalDP algorithm stands decently alongside the state-of-the-art RAPPOR algorithm for the same conditions between two local differential privacy algorithms.

6.6 Limitations

Despite having a visualization system that includes numerous features that contribute to the community, obstacles could potentially cause limitations. Starting with data, the visualization interface only allows quantitative data to be imported and used. Thus, it opens a concern about the handling a case when a dataset contains both quantitative and textual data. Despite having an algorithm that allows any data type, including textual, the metrics play the role of having difficulties with calculating scores when data is not quantitative. To conclude, if the metrics were adjusted to calculate a score for textual data, there would not be any data restrictions.

The second limitation focuses on way of permutating data to get the

results. As it was suggested from one of the experts, instead of running the algorithm for a specific epsilon parameter value one, it would be more benificial to run the algorithm for 100 times. By this, we would get the average results of the permutation on a specific dataset that would avoid potential issues with the randomized function. The third limitation is focused on the algorithm. Because of focusing on the visual analytics perspective and implementing the visualization system as the priority, the algorithm was created as an inspiration from local differential privacy and randomized response mechanism. Because of its limitations in capabilities, the algorithm serves only data analysts and inexperienced practitioners.

The last limitation is that there is no privacy validation compared to the state-of-the-art algorithm. Even though local differential privacy as an outcome produces a synthetic dataset, in a case with the state-of-the-art algorithm, RAPPOR. Its dataset does not produce synthetic versions of data points but rather creates additional rows and adds original rows into the dataset, preventing analyzing data privacy. If there were another state-of-the-art algorithm that produces a proper permutation dataset, we would compare by data privacy metric and evaluate comparison.

7 Conclusion and Future Work

We can conclude that the project managed to introduce the association between two domains, differential privacy, and data visualization, to create foundations for the privacy-preserving visual analytics approach. By creating a visualization system that supports an alternative and simple solution based on local differential privacy, experts and inexperienced practitioners can understand how privacy-preserving techniques by adding noise could affect specific data. In addition, by focusing on the tradeoff between data privacy and utility and introducing the feature of adjusting the epsilon parameter, the research contributes to both society and the academic community. By having four layers of evaluation techniques that proved each component of the thesis, the project accomplished the expected goals.

The domain of differential privacy constantly develops, with many researched getting engaged in creating new mechanisms to private personal information. In addition, such a privacy-preserving technique efficiently collaborates and contributes to various domains, from healthcare to finance. Thus, there is still plenty of room for further discovery and implementations within various fields, and differential privacy mechanisms are getting specific characteristics for each of these implementations. From the academic perspective, privacy-preserving data analytics and differential privacy can be paired with other research domains, as such a case was in the thesis. By merging differential privacy and data visualization, it was possible to create privacy-preserving visual analytics that empowers understanding how privacy techniques affect privacy and what possible adjustments are by having a visual interface that provides investigations. Thus, for future work, we suggest further progress in privacy-preserving visual analytics, creating new interfaces, systems, and frameworks that visually support the investigation of privacy mechanisms towards understanding their effect on data.

The thesis created a solid foundation for the new approach, thus further inventions and studies are welcome. Because there is a tradeoff between data privacy and utility, the research gets another perspective on data, thus including interactive adjustment contributing to understanding the tradeoff. We would suggest exploring interactive visual analytics from the perspective of the tradeoff between data privacy and utility to provide a solution that allows efficient and proper presentation of the best ratio between data privacy and utility. By including other parameters of privacy-preserving approaches, such as sensitivity, the users would be able to investigate even deeper into a dataset and the mechanism. In addition, including more advanced data privacy and utility metrics would improve the whole project.

Moreover, further creation of visualization systems that allow any data would impact the transparent and universal perception of how each data would be affected by the privacy-preserving mechanisms. By including textual data and dates, the visualization system would accept any dataset.

With such development, there would be more analysis that would users conduct on their data, and by covering different data types, they could understand how noise affects other data types. Moreover, nowadays, data gets more complicated, and it grows in volume, thus these visualization systems should support any data type and be open for interactive computation of noise adjustment while getting enormous data quantities. Further development of the privacy-preserving visual analytics should also be concentrated on specific domains such as healthcare, census, and finance, where data privacy plays a vital role. Because each domain has a different group of users ranging from being experts in data privacy to have no knowledge about it, creating visualizations for specific domains would be necessary for future work. In addition to the development from a data visualization perspective, it is suggested to continue further progress from the differential privacy aspect. By developing new mechanisms or adding existing local differential privacy algorithms to the system, its purpose would be directed to other groups of users, and its value would increase. This improvement could go in the direction of choosing between a couple of mechanisms and comparing their performance, which would give a new perspective on the system. We suggest creating a universal system that would allow multiple algorithms to be implemented or even importing mechanisms directly to the visualization system.

In conclusion, the future work for the collaboration between data visualization and differential privacy has to be focused on understanding how different groups of users, experienced researchers, data analysts, and practitioners, would benefit from such a visualization system. First, it is essential to recognize what visualizations and metrics would help each group in their investigation. Because each group has its benefit from the visualization system, it is expected that there would be different implementations and solutions for each group of users. Next, expanding the comprehensiveness of the system by allowing other data types to be integrated within the system is crucial for further development. This way, the system would not be restricted to specific requirements, and more analysis of various data would be conducted. Such universality should also be given for a mechanism within the system, where users could choose between a couple of algorithms to compare their performances.

References

[1] Recital 26 - not applicable to anonymous data — general data protection regulation (gdpr). `https://gdpr-info.eu/recitals/no-26/`. (Accessed on 03/14/2021).

[2] D. P. T. Apple. Learning with privacy at scale differential. 2017.

[3] H. H. Arcolezi, J.-F. Couchot, S. Cerna, C. Guyeux, G. Royer, B. Al Bouna, and X. Xiao. Forecasting the number of firefighters interventions per region with local-differential-privacy-based data. *Computers & Security*, p. 101888, 2020.

[4] D. Avraam, R. C. Wilson, O. W. Butters, T. Burton, C. Nicolaides, E. Jones, A. Boyd, and P. R. Burton. Privacy preserving data visualizations. *EPJ Data Sci.*, 10(1):2, 2021. doi: 10.1140/epjds/s13688-020-00257-4

[5] P. Baxter, S. Jack, et al. Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, 13(4):544–559, 2008.

[6] B. Bebensee. Local differential privacy: a tutorial. *CoRR*, abs/1907.11908, 2019.

[7] K. Bhattacharjee, M. Chen, and A. Dasgupta. Privacy-preserving data visualization: Reflections on the state of the art and research opportunities. *Comput. Graph. Forum*, 39(3):675–692, 2020. doi: 10.1111/cgf.14032

[8] M. A. P. Chamikara, P. Bertók, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet Things J.*, 7(7):5827–5842, 2020. doi: 10.1109/JIOT.2019.2952146

[9] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang. Privacy at scale: Local differential privacy in practice. In G. Das, C. M. Jermaine, and P. A. Bernstein, eds., *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pp. 1655–1658. ACM, 2018. doi: 10.1145/3183713.3197390

[10] G. Cormode, T. Kulkarni, and D. Srivastava. Marginal release under local differential privacy. In G. Das, C. M. Jermaine, and P. A. Bernstein, eds., *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pp. 131–146. ACM, 2018. doi: 10.1145/3183713.3196906

[11] S. Crowe, K. Cresswell, A. Robertson, G. Huby, A. Avery, and A. Sheikh. The case study approach. *BMC medical research methodology*, 11(1):1–9, 2011.

[12] A. Dasgupta and R. Kosara. Privacy-preserving data visualization using parallel coordinates. In P. C. Wong, J. Park, M. C. Hao, C. Chen, K. Börner, D. L. Kao, and J. C. Roberts, eds., *Visualization and Data Analysis 2011, San Francisco Airport, CA, USA, January 24-25, 2011*, vol. 7868 of *SPIE Proceedings*, p. 78680O. SPIE, 2011. doi: 10.1117/12.872635

[13] A. Dasgupta, R. Kosara, and M. Chen. Guess me if you can: A visual uncertainty model for transparent evaluation of disclosure risks in privacy-preserving data visualization. In R. Gove, D. Arendt, J. Kohlhammer, M. Angelini, C. L. Paul, C. Bryan, S. McKenna, N. Prigent, P. Najafi, and A. Sopan, eds., *16th IEEE Symposium on Visualization for Cyber Security, VizSec 2019, Vancouver, BC, Canada, October 23, 2019*, pp. 1–10. IEEE, 2019. doi: 10.1109/VizSec48167.2019.9161608

[14] A. Dasgupta, E. Maguire, A. Abdul-Rahman, and M. Chen. 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*, vol. 13, 2014.

[15] Dataman. You can be identified by your netflix watching history, Aug 2020.

[16] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. *CoRR*, abs/1712.01524, 2017.

[17] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *Automata, Languages and Programming*, pp. 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[18] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, eds., *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, vol. 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878
$_1$4

[19] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042

[20] J. B. Earp and F. C. Payton. Information privacy in the service sector: An exploratory study of health care and banking professionals. *Journal of Organizational Computing and Electronic Commerce*, 16(2):105–122, 2006.

[21] Ú. Erlingsson, A. Korolova, and V. Pihur. RAPPOR: randomized aggregatable privacy-preserving ordinal response. *CoRR*, abs/1407.6981, 2014.

[22] A. Friedman and A. Schuster. Data mining with differential privacy. In B. Rao, B. Krishnapuram, A. Tomkins, and Q. Yang, eds., *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pp. 493–502. ACM, 2010. doi: 10.1145/1835804.1835868

[23] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010. doi: 10.1145/1749603.1749605

[24] P. Gill, K. Stewart, E. Treasure, and B. Chadwick. Methods of data collection in qualitative research: interviews and focus groups. *British dental journal*, 204(6):291–295, 2008.

[25] L. O. Gostin, L. A. Levit, S. J. Nass, et al. Beyond the hipaa privacy rule: enhancing privacy, improving health through research. 2009.

[26] M. E. Gursoy, A. Tamersoy, S. Truex, W. Wei, and L. Liu. Secure and utility-aware data collection with condensed local differential privacy. *CoRR*, abs/1905.06361, 2019.

[27] A. Haeberlen. Andreas haeberlen. `https://www.cis.upenn.edu/~ahae/`. (Accessed on 06/14/2021).

[28] A. Haeberlen, B. C. Pierce, and A. Narayan. Differential privacy under fire. In *20th USENIX Security Symposium, San Francisco, CA, USA, August 8-12, 2011, Proceedings*. USENIX Association, 2011.

[29] K. Hammarberg, M. Kirkman, and S. de Lacey. Qualitative research methods: when to use them and how to judge them. *Human Reproduction*, 31(3):498–501, 01 2016. doi: 10.1093/humrep/dev334

[30] N. Holohan, D. J. Leith, and O. Mason. Optimal differentially private mechanisms for randomised response. *CoRR*, abs/1612.05568, 2016.

[31] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. Differential privacy: An economic method for choosing epsilon. *CoRR*, abs/1402.3329, 2014.

[32] P. Kairouz, S. Oh, and P. Viswanath. Extremal mechanisms for local differential privacy. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2879–2887, 2014.

[33] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011. doi: 10.1137/090756090

[34] K. Kenthapadi. Privacy-preserving data mining in industry: Practical challenges and lessons learned, Aug 2018.

[35] J. W. Kim, B. Jang, and H. Yoo. Privacy-preserving aggregation of personal health data streams. *PloS one*, 13(11):e0207639, 2018.

[36] J. W. Kim, D. Kim, and B. Jang. Application of local differential privacy to collection of indoor positioning data. *IEEE Access*, 6:4276–4286, 2018. doi: 10.1109/ACCESS.2018.2791588

[37] U. M. Learning. Pima indians diabetes database. `https://www.kaggle.com/uciml/pima-indians-diabetes-database`, Oct 2016. (Accessed on 05/10/2021).

[38] J. Lee and C. Clifton. How much is enough? choosing ϵ for differential privacy. In X. Lai, J. Zhou, and H. Li, eds., *Information Security, 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings*, vol. 7001 of *Lecture Notes in Computer Science*, pp. 325–340. Springer, 2011. doi: 10.1007/978-3-642-24861-0
$_2$2

[39] Y. J. Lee and K. H. Lee. What are the optimum quasi-identifiers to re-identify medical records? In *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 1025–1033. IEEE, 2018.

[40] H. Liu, Z. Wu, C. Peng, F. Tian, and L. Lu. Bounded privacy-utility monotonicity indicating bounded tradeoff of differential privacy mechanisms. *Theor. Comput. Sci.*, 816:195–220, 2020. doi: 10.1016/j.tcs.2020.02.004

[41] A. Mansbridge, G. Barbour, D. Piras, C. Frye, I. Feige, and D. Barber. Learning to noise: Application-agnostic data sharing with local differential privacy. *CoRR*, abs/2010.12464, 2020.

[42] T. Munzner. A nested process model for visualization design and validation. *IEEE Trans. Vis. Comput. Graph.*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111

[43] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pp. 111–125. IEEE Computer Society, 2008. doi: 10.1109/SP.2008.33

[44] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin. Collecting and analyzing data from smart device users with local differential privacy. *CoRR*, abs/1606.05053, 2016.

[45] H. K. Patil and R. Seshadri. Big data security and privacy issues in healthcare. In *2014 IEEE international congress on big data*, pp. 762–765. IEEE, 2014.

[46] P. R. M. Rao, S. M. Krishna, and A. P. S. Kumar. Privacy preservation techniques in big data analytics: a survey. *J. Big Data*, 5:33, 2018. doi: 10.1186/s40537-018-0141-8

[47] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual parameter space analysis: A conceptual framework. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2161–2170, 2014. doi: 10.1109/TVCG.2014.2346321

[48] A. Singh. Data publishing techniques and privacy preserving. 2019.

[49] J. W. Smith, J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, p. 261. American Medical Informatics Association, 1988.

[50] D. Wang. Di wang's homepage. `https://shao3wangdi.github.io/`. (Accessed on 06/14/2021).

[51] D. Wang, M. Gaboardi, A. Smith, and J. Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *J. Mach. Learn. Res.*, 21:200:1–200:39, 2020.

[52] N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu. Collecting and analyzing multidimensional data with local differential privacy. *CoRR*, abs/1907.00782, 2019.

[53] T. Wang, J. Blocki, N. Li, and S. Jha. Locally differentially private protocols for frequency estimation. In E. Kirda and T. Ristenpart, eds., *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017*, pp. 729–745. USENIX Association, 2017.

[54] X. Wang, W. Chen, J. Chou, C. Bryan, H. Guan, W. Chen, R. Pan, and K. Ma. Graphprotector: A visual interface for employing and assessing multiple privacy preserving graph algorithms. *IEEE Trans. Vis. Comput. Graph.*, 25(1):193–203, 2019. doi: 10.1109/TVCG.2018.2865021

[55] X. Wang, J. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K. Ma. A utility-aware visual approach for anonymizing multi-attribute tabular data. *IEEE Trans. Vis. Comput. Graph.*, 24(1):351–360, 2018. doi: 10.1109/TVCG.2017.2745139

[56] Y. Wang, X. Wu, and D. Hu. Using randomized response for differential privacy preserving data collection. In T. Palpanas and K. Stefanidis, eds., *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016*, vol. 1558 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[57] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[58] C. Wu. Chuhao wu. `https://www.chuhaowu.com/`, Jul 2015. (Accessed on 06/14/2021).

[59] N. Wu, C. Peng, and K. Niu. A privacy-preserving game model for local differential privacy by using information-theoretic approach. *IEEE Access*, 8:216741–216751, 2020. doi: 10.1109/ACCESS.2020.3041854

[60] X. Xiong, S. Liu, D. Li, Z. Cai, and X. Niu. A comprehensive survey on local differential privacy. *Secur. Commun. Networks*, 2020:8829523:1–8829523:29, 2020. doi: 10.1155/2020/8829523

[61] M. Yang, L. Lyu, J. Zhao, T. Zhu, and K. Lam. Local differential privacy and its applications: A comprehensive survey. *CoRR*, abs/2008.03686, 2020.

[62] R. Yin. *Case Study Research: Design Methods*, vol. 5. 01 2009.

[63] Z. Zainal. Case study as a research method. *Jurnal kemanusiaan*, 5(1), 2007.

[64] D. Zhang, M. Hay, G. Miklau, and B. O'Connor. Challenges of visualizing differentially private data. *Theory and Practice of Differential Privacy*, 2016:1–3, 2016.

[65] D. Zhang, A. Sarvghad, and G. Miklau. Investigating visual analysis of differentially private data. *IEEE Trans. Vis. Comput. Graph.*, 27(2):1786–1796, 2021. doi: 10.1109/TVCG.2020.3030369