This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3456304

Defogger: A Visual Analysis Approach for Data Exploration of Sensitive Data Protected by Differential Privacy

Xumeng Wang, Shuangcheng Jiao, and Chris Bryan

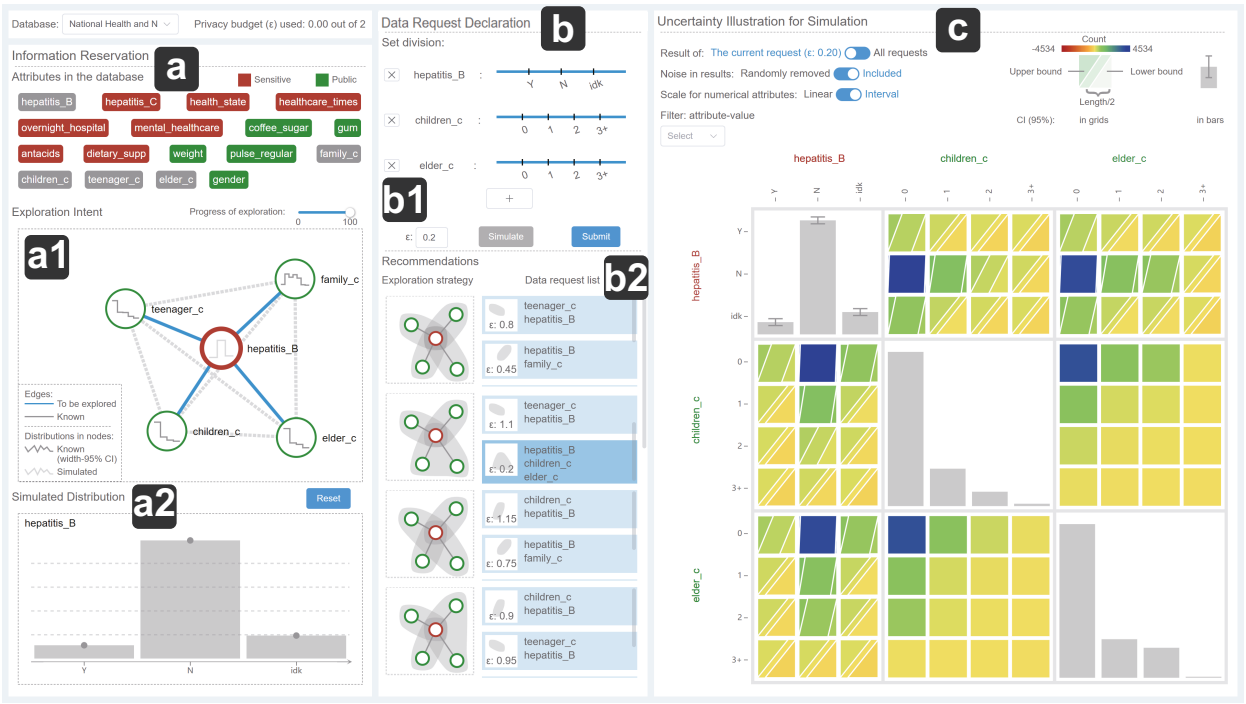


Fig. 1: The interface of Defogger for exploring data guarded by differential privacy. (a) The information reservation view invites users to specify data information (e.g., distributions and correlations) of interest as their exploration intent (a1) and describe available data facts to generate simulated data (a2) for exploration strategy recommendation. (b) The data request declaration view allows users to declare a data request (b1) based on recommendations from a reinforcement learning model (b2). (c) The uncertainty illustration view explains the uncertainty caused by differential privacy in the simulated or the actual response of the declared data request.

Abstract—Differential privacy ensures the security of individual privacy but poses challenges to data exploration processes because the limited privacy budget incapacitates the flexibility of exploration and the noisy feedback of data requests leads to confusing uncertainty. In this study, we take the lead in describing corresponding exploration scenarios, including underlying requirements and available exploration strategies. To facilitate practical applications, we propose a visual analysis approach to the formulation of exploration strategies. Our approach applies a reinforcement learning model to provide diverse suggestions for exploration strategies according to the exploration intent of users. A novel visual design for representing uncertainty in correlation patterns is integrated into our prototype system to support the proposed approach. Finally, we implemented a user study and two case studies. The results of these studies verified that our approach can help develop strategies that satisfy the exploration intent of users.

Index Terms—Differential privacy, Visual data analysis, Data exploration, Visualization for uncertainty illustration

1 Introduction

As datasets that contain sensitive information, such as medical, social, or other personal data [15, 23, 27], become increasingly prevalent, the querying, browsing, analyzing, and exploring of such data must enforce privacy-related restrictions. When providing data based on a user request (e.g., as part of a database query or API call), a common mechanism to enforce privacy restrictions is the use of *differential privacy* (DP) [6], which protects data records by injecting noise into the response to data requests (e.g., as part of a data exploration process [3]). DP has become a common mechanism to scaffold access for databases that contain sensitive information because they have been shown to be effective in safeguarding privacy during exploration workflows [18] while also meeting regulatory requirements such as the European Union's GDPR [29] and California's CPRA [5].

DP approaches commonly maintain a *privacy budget*: when a data request is issued, a spending deduction is calculated based on the amount of noise added to the response and this is subtracted from the overall

- *Xumeng Wang and Shuangcheng Jiao are with DISSec, Nankai University. E-mail: wangxumeng@nankai.edu.cn, jiaoshuangcheng@mail.nankai.edu.cn.*
- *Chris Bryan is with SCAI, Arizona State University. E-mail: cbryan16@asu.edu.*
- *Xumeng Wang is the corresponding author.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

budget amount [7]. A higher budget can be traded for a response with a smaller noise. Importantly, once a response is provided, there is no mechanism for rolling back the request or increasing the budget. As the budget is exhausted, it becomes difficult to gain insights because responses will be highly noisy, which leads to high uncertainty. Therefore, DP is strained for data exploration processes. In particular, *exploratory visual analysis*, whereby humans create and inspect a series of visualizations to explore, analyze, or discover insights from the dataset [3], suffers more from DP approaches: (i) The noise injected by DP can mask the visualized patterns (e.g., trends, outliers) [37] or be overpriced [30]. (ii) A user will likely explore the dataset by browsing a series of charts [15, 23, 27], while each chart must be drawn based on requested results (e.g., results of count queries for drawing a histogram), thus continually consuming a given privacy budget. (iii) Consumption is hard to predict because users could send requests based on serendipitous findings from a previously-created chart, or perform "locate" or "explore" search actions [4], which makes a *priori* customized noising solutions (such as those invoked by differential private data synthesis [40]) difficult to design and implement.

The takeaway is that users need intelligent strategies for visually exploring datasets containing sensitive information that is protected by DP, as a naive approach can easily exhaust a privacy budget and result in uninformative charts that contain high uncertainty [37]. Bolstered by a use case that identified an explicit set of context requirements, we thus propose a novel, first-of-its-kind workflow for this process, motivated by the need for users (i) to maintain sufficient awareness of budget allocation costs and spending, balanced against (ii) the need to support sufficient analysis and interrogation capabilities to derive insights and value from the dataset, while also (iii) adhering to required data safeguards imposed by DP.

A key to our approach is that the pipeline considers that budget allocation is not only numerical calculations of the noise size but also a question of which responses actually merit spending the budget. Our solution is to provide recommendations modeled on a combination of the exploration interests of users, knowledge of data patterns, and an understanding of how a group of data requests can contribute to the exploration process. In part, we are inspired by prior tools (for non-sensitive datasets) that recommend 'next steps' to augment visualization-driven exploration, such as Voyager [34]. Our pipeline goes a step further: we extract data patterns at different granularities (from overview to details) and leverage these to suggest responses tailored to an "appropriate" level of detail for visualizations that still successfully support exploration but spend budget more wisely. We implement our pipeline in an interactive tool called *Defogger*, which interactively recommends feasible exploration strategies for the next step by a reinforcement learning model and demonstrates how each strategy will work (i.e., how much budget it will spend, and how much uncertainty the resultant visualization will contain) before actually applying the step.

We validate our pipeline (and Defogger's implementation) by conducting robust human-centered evaluations (a user study and case studies). Evaluation results indicate that our approach can improve returns on investment of privacy budgets through the intelligent recommendation of data requests based on the exploration intent of users. ***Ultimately, our pipeline represents a novel approach that augments the ability of humans to explore and gain increased value from data while adhering to DP constraints.*** We conclude this paper by additionally contributing a discussion of generalizable lessons learned about the effective intersection of computational models, visualization, and interaction as a way to optimize privacy preservation and human sensemaking, which can benefit both the visualization and privacy communities, as well as limitations of the current pipeline and Defogger designs, and suggest opportunities for future work in this area.

2 Related Work

2.1 Processes of Exploratory Visual Analysis

Exploratory visual analysis of abstract/non-spatial data mainly leverages three categories of operations, consisting of "query," "browse," and "search" actions to iteratively interact with data [3, 10]. If information needs to be excluded from the exploration target because of privacy concerns, query operations are necessary for requesting visualizations that can enable browsing and searching patterns of groups. Queries in the exploration process can be organized either in a top-down manner [36] following a specific exploration goal or in a bottom-up fashion [1] without a clear destination [3]. When implementing a bottom-up exploration process, users have no predetermined target of visualization and therefore need recommendations for the next visualization. As for a top-down exploration process, recommendations help inspect whether there are ignored data patterns.

Recommendations for the next visualization can be made according to the history of users' exploration process. For example, prior tools [2, 24, 33, 34, 41] have considered the relevance of previously explored visualizations based on similarities in their data content or data patterns. Zhou et al. [43] monitored the development of the exploration focus (based on which attributes were being explored) and developed a model to infer the next focus. Qian et al. [24] recommend visualizations based on the data patterns that could be of interest to the users. Importantly, ***none of the above studies consider privacy issues during exploration and recommendation***. When constrained by privacy budgets, users are unable to leverage a pattern-driven exploration process. Instead, they have to carefully plan what content they should focus on during their exploration process. This paper addresses this gap by developing a recommendation pipeline that understands vague descriptions of users' interests in data content and makes corresponding recommendations.

2.2 Practices in Differentially Private Publication

To apply DP, users have to develop and follow "trading schemes," consisting of the group of trading targets and how much of the privacy budget should be spent on each target. Existing studies proposed examples of trading schemes and provided tools for scheme assessment.

In real-world scenarios, the values of data distribution in different bins are not the same. Users prefer to spend a higher privacy budget on significant bins than others. As the significance of each bin is unknown, trading rules can allow users to pay an extra privacy budget for significance assessment. For example, publishing differentially private histograms could benefit from an assessment result of histogram structure [35, 44]. When exploring data progressively, the error of prior publications can affect the significance of the next. An adaptive trading scheme [16] pays a high budget for the next to avoid going astray when the cumulative error of prior publications is excessive.

For the assessment of trading schemes, the accuracy of potential publication is a key metric. The accuracy derived from a trading scheme is not static due to the random noise. As the publication is a single value, visualizations [13, 22] can directly represent the distribution of potential publication results to summarize the effectiveness of a user-specified privacy budget. If the publication is a complicated pattern, the accuracy can be assessed by comparing the differentially private publications with the original ones [32, 42].

Accessing the original data is necessary for existing studies to determine trading schemes because they consider the original data as input. The target users in this study are not authorized to access the original data. Therefore, determining and assessing trading schemes requires tolerating more uncertainty.

2.3 Visualizations for Uncertainty Representation

When there is uncertainty in data, understanding uncertainty is necessary for data analysis [11]. Visualizations can facilitate the understanding of uncertainty by demonstrating the existence of uncertainty in data distribution and representing the features of uncertainty.

Indicators, like the level of confidence, quantify the degree of uncertainty. Classic visualizations (e.g., histograms, pie charts, and scatter matrices) that are designed to show distributions of 1-dimensional data or multi-dimensional data require extra visual representations to illustrate the degree of uncertainty simultaneously. Extra representations for uncertainty should not conflict with those used to visualize data distributions. Color encoding for uncertainty can be superimposed on visualizations that show data distributions by shape-related encodings or position encodings [26]. If the encoding of color filling is occupied,

the style of strokes (e.g., opacity [39], wave frequency [9]) can be customized for uncertainty representation.

In addition to the degree, uncertainty analysis may require more detailed descriptions of uncertainty. To support uncertainty analysis, values with uncertainty can be depicted as a probability distribution and instantiated as a group of values. Kay et al. [14] summarized four categories of visualizations for instantiating data with uncertainty, consisting of intervals, ribbons, slabs, and dotplots. These 4 categories can adapt to different visual representations of values with uncertainty or 1-dimensional data with uncertainty. Visualizition of uncertainty has previously been employed as a way to support privacy preserving visualizations and DP [13,22,32]. However, prior work does not consider data exploration based on a user's privacy budget, as is done in this paper. Also, users may explore complicated data patterns, like the distributions of multi-dimensional data. Nevertheless, uncertainty in distributions of multi-dimensional data still lacks effective representations. We propose a Mosaic design that can embed visual representations for uncertainty into the grid of heatmap matrices.

3 Scenario Overview and Requirements Analysis

Our target application scenario is centered around users who need to explore tabular datasets containing sensitive information which are protected by DP mechanisms. A real-world example of this would be the Cloud Healthcare API provided by BigQuery [1]. Users (generally researchers or clinicians) apply for portal access; once approved, they are allocated an exploration budget. The primary research question we investigate is, ***how do we optimize the visual exploration process (thus providing value to these users) while maintaining appropriate privacy safeguards for sensitive data?*** In this section, we briefly formalize the "privacy aware exploration" practice conducted by these sorts of users, outline a virtual use case to illustrate available exploration schemes and user requirements, and distill a set of requirements.

3.1 Practice under Privacy Limitations

DP forces users to obtain noisy responses to data requests based on a defined privacy budget. Formally, the mechanism of noise addition aims at yielding indistinguishable responses for two datasets D and D' differing by any single record. Specifically, a data request K with a privacy budget of ε satisfies

$$Pr[K(D) \in S] \leq \exp(\varepsilon) \times Pr[K(D') \in S],$$

where S could be any subset of all possible values that can be returned by K. In other words, K needs to yield similar responses for the two datasets. Hence, the noisy responses can prevent sensitive information about an individual from leaking. A lower privacy budget denotes a higher limitation of privacy preservation, which requires a relatively larger noise. To ensure that the size of the noise to be added is appropriate, mechanisms of noise generation consider Δf, the sensitivity of responses, which is how much difference could exist between the actual response on D and D'. For example, the count queries used to summarize data distribution have a unified sensitivity of 1. The Laplace mechanism yields random noise with a Laplace distribution $Lap(\Delta f/\varepsilon)$. According to the Laplace mechanism, the results of count queries will be added random noise satisfying $Lap(1/\varepsilon)$. Since the size of noise is irrelevant to the value of query results, the disturbance (or uncertainty) caused by noise is relatively small to a large value.

In practice, users are not permitted to have direct access to individual information. Instead, they can request statistical information for groups. The charge of privacy budget for statistical information is independent of the group size because data requests against other individuals will cause no additive risks to privacy exposure [17]. Users can determine the cost of the privacy budget for a statistical request only based on the expected size of the noise. In contrast, repeatedly requesting the same individual's information increases the risk of identity or information leakage. In this case, the privacy budget must be charged again [35]. For instance, count queries (i.e., quantity statistics) for multiple sets can be executed in a single data request and charged once when there is no intersection among all sets because each individual needs to be counted in at most one set. Following this principle, users can optimize the use of a privacy budget by employing multiple sets, which can cover the entire dataset without any overlap. To identify such set groups for data tables, users can declare conditions (i.e., intervals for numerical attributes and categories for categorical attributes) for *set division* according to a grid partition of the data space (see Fig. 2).

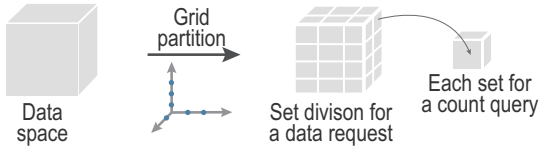


Fig. 2: Set divisions used in a data request.

Data requests can employ either a fine or a coarse granularity of grid partition to determine a set division for batch queries. A fine granularity of set division can facilitate flexible exploration of data distribution from different perspectives because the total count of a group of sets can be considered the count of their parent set. Nevertheless, the sum calculation integrates the noise added to each count of small sets together. Assume that a privacy budget ε is paid for a data request. The noise added to the total count of m sets satisfies $m * Lap(1/\varepsilon)$. Although the responses to batch queries on a large number of small sets can be used to calculate data distribution over different attributes or different value divisions of attributes, the transformed distribution could be less accurate than the results from multiple data requests using coarse granularity set conditions but with a small privacy budget respectively. In summary, data exploration needs to coordinate data requests by overall planning multiple set division schemes.

3.2 Use Case

Lucy is a medical analyst who wants to identify which living habits can decrease the risks of type 2 diabetes. She has successfully applied for a privacy budget to access a questionnaire dataset, including descriptions of citizens' living habits (e.g., sleeping hours, eating habits, fitness situations) and their health status. Due to her limited privacy budget, Lucy has to focus on the attributes that are more likely to reflect the potential prevention of the disease and bargain with the mechanism of DP to get more accurate data distributions as convincing providence.

Lucy first prioritizes the attributes of coffee intake and taste preferences. The distribution of diabetic conditions over either both of the attributes or each of them could be useful. Lucy can divide all individuals in the database into small sets according to the value pair of the two attributes and spends a partial budget to implement batch queries on the total amount of patients in each set. As for the distribution over a single attribute, Lucy can get a relatively inaccurate distribution by adding the requested results on small sets because the sets corresponding to value divisions of a single attribute are combinations of those small sets. Lucy can also spend more privacy budget to request distributions over single attributes. To judge whether the extra budget is necessary, Lucy needs to assess the effect of noise and learn about the accuracy of the calculated distributions. If the accuracy is acceptable, Lucy can spend the remaining privacy budget to request and explore correlations between diabetic conditions and other attributes, like fitness situations.

3.3 Requirement Analysis

Privacy restrictions have two primary impacts on the visual exploration processes [32, 37]: (i) users have to spend more cognitive energy when formulating a data request, and (ii) each exploration result contains uncertainty caused by noise. To understand how a visualization-driven approach could help mitigate these impacts, we conducted a pre-study with two experts (E_1 and E_2) in DP respectively, where each expert had at least three years of research experience. In the study, we discussed the two impacts with each expert, recorded their opinions, and refined specific requirements derived from each impact.

[1] https://cloud.google.com/bigquery/docs/differential-privacy

R1: Get advice on data exploration. When deciding on the next step of data exploration, multiple sophisticated factors need to be considered due to the limitation of DP and autonomy in data exploration. Users could take advantage of suggestions from the models.

- **R1.1: Sketch the exploration intent.** To come up with effective suggestions on data requests, the recommendation model should save the privacy budget on what users attach importance to. The recommendation model needs to fully understand and follow the exploration intention of users.
- **R1.2: Express prior knowledge.** The effect of uncertainty depends on the requested values. E_1 told us that response simulation based on prior knowledge can embody the impact of uncertainty and therefore help the understanding of uncertainty. The prior knowledge in the scenarios of data exploration includes the public distribution of non-sensitive data and inference results based on domain knowledge from users. Either of them can contribute to the decision on how many sets should be split to request statistical results—if the values of an attribute have a jitter distribution, a finer division could better reveal the distribution details.
- **R1.3: Browse feasible set division for data exploration.** There could be various choices of set division for batch queries because the attributes of interest have multiple combinations and values of attributes can be divided by different segments. Data requests employing different set divisions feedback distinct data patterns and exert inconsistent effects on the consequent exploration process. E_1 reckoned that set division schemes need graphic illustrations to facilitate sorting out exploration processes.

R2: Budget with the understanding of noise. The size of noise in DP is controlled not only by the privacy budget paid by users but also by randomness. Users have to determine how much of a privacy budget to set based on the likelihood of noise and the underlying impact.

- **R2.1: Preview the noise effect before the budget is spent.** DP compels data requests to be irrevocable. E_2 mentioned that users need to confirm the effect of noise caused by DP is acceptable and affordable by previewing the effect of noise controlled by a privacy budget.
- **R2.2: Learn about the existence of noise in the actual feedback.** Note that the simulation results for previewing cannot be implemented on the data records in the databases unless privacy budgets are paid. There could be differences between the preview version and the actual feedback. Users still need to learn about the latter. Also, E_2 told us that users may ignore the uncertainty when browsing a distribution described by a group of specific values. A feasible solution provided by E_2 is to remind users of uncertainty through the concept of confidence intervals.
- **R2.3: Understand the random noise with instances.** E_1 suggested disclosing the mechanism of noise by showing possibilities, which is similar to instance-based interpretation [14, 22].

4 Approach

In this section, we introduce the proposed pipeline for the visual exploration of datasets protected by DP and modules.

4.1 Pipeline

To satisfy the requirements mentioned in Sect. 3.3, we refined the pipeline into several steps, as shown in Fig. 3. Following our pipeline, users can determine exploration strategies and understand the uncertain effects on exploration results when the data interface is guarded by differential privacy. Seeking effective recommendations for exploration strategies (**R1**), users are first invited to reserve information about their exploration intent (**R1.1**, Fig. 3(a1)) or inference on data facts related to the exploration intent (if any, **R1.2**, Fig. 3(a2)). The exploration intent can be described as distributions of interest, which could be the distributions of a single attribute describing data overview or joint distributions of multiple attributes demonstrating attribute correlations. The above input will be taken into account by the recommendation model (Fig. 3(b)) to yield suggestions. Next, users need to declare a data request, consisting of a set division (Fig. 3(c1)) and a privacy budget (Fig. 3(c2)), which is no more than the remaining, to drive data exploration. Before making a submission, users can browse suggested data requests (**R1.3**) and test each of them by simulating the data request on simulation data constructed by all known data facts (**R2.1**, Fig. 3(d)). After submitting a data request, users can examine the feedback (**R2.2**, Fig. 3(e)). Aimed at explaining the randomness of noise, users can further browse possible distributions (**R2.3**) of the actual data calculated by randomly removing a noise. The feedback will be integrated into the known data facts to support the subsequent data exploration, which can also be reviewed by users.

4.2 Recommendation of Exploration Strategies

We employ a recommendation model to suggest exploration strategies. An exploration strategy refers to a sequence of data requests that can coordinate to achieve all attribute distributions and attribute correlations specified as the exploration intent. Variables in a data request consist of a set division scheme (i.e., a set of attributes and corresponding value divisions), the order of the request in the request sequence, and a privacy budget. To focus on effective strategies, we first enumerate strategy *prototypes* by identifying groups of attribute sets that can generate feasible groups of set division schemes. For this goal, we describe multiple data distributions by a graph model, where nodes denote distributions of single attributes and edges denote correlation distribution of two attributes. The graph model constructs an intent graph according to the exploration intent reserved by users (Fig. 3(a1)). Similarly, distribution information returned by a data request employing a grid partition (see Fig. 2) can be described as a complete graph with all nodes in the attribute set employed by the set division scheme. Then, the enumeration problem can be solved by listing groups of complete graphs that cover all edges and nodes in the intent graph.

Next, we complete each prototype and generate strategy candidates. Considering that data exploration scenarios lack exploration demonstration and labeled data, we employ a reinforcement learning model trained by Q-learning [38]. Specifically, we describe an exploration strategy as making a series of actions. We explain the definition of actions and the incentive mechanism for an action as follows.

Action space: Each action describes a data request with a privacy budget to be paid and an employed set division. The budget number has to be less than the remaining privacy budget. The set division can be an unused attribute set from the prototype integrating with value division schemes for each attribute.

Incentive mechanism: We quantify the bonus with three considerations: (i) users prefer accurate illustration for what they have listed as exploration intent, (ii) hubs in the intent graph may correspond to the focus of the exploration intent, and (iii) the saved budget can support more data exploration. Following the first consideration, we identify the joint distribution represented by each edge in the intent graph as an object for accuracy calculation. The bonus for each data request further weights the accuracy of each target distribution by the average betweenness centrality of the two nodes. Finally, we assign an extra bonus for budget savings, a one-time reward that is only issued for the last action in each strategy candidate. Implementation details of accuracy calculation, data simulation for accuracy estimation under the privacy limitation, and calculation of the bonus for budget consumption can be found in the following paragraphs.

Accuracy calculation: We quantify accuracy according to two types of relative errors, which are structural errors caused by inconsistent set division and numerical errors led by noise. In our scenario of data exploration, data distribution can be described by individual numbers in discrete sets. Statistical results in a coarser granularity can provide fewer details of a distribution, which leads to structural errors. The ideal granularity for descriptions of a distribution is the finest granularity allowed by the data interface. Nevertheless, the benefits from details may not linearly increase with the decrease of granularity sizes [25]. To quantify the structural error for a set division, we first generate two descriptions of the target distribution: an ideal description providing the

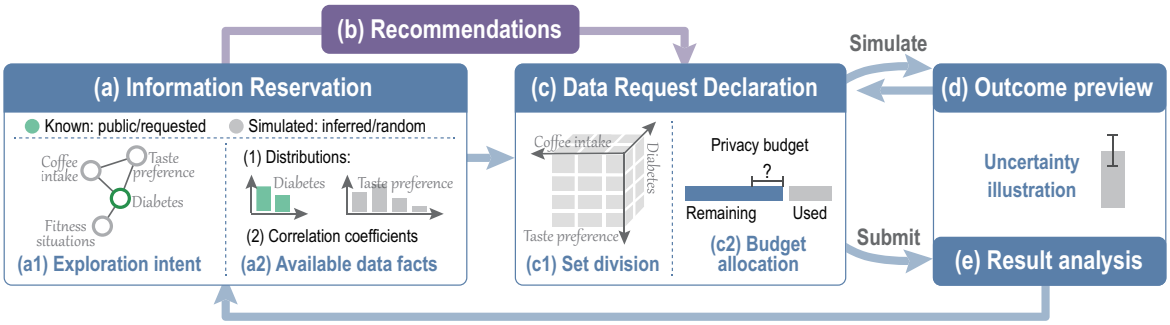


Fig. 3: The pipeline for visual exploration with the restriction of DP, which mainly consists of a recommendation model and three iterative modules. In the information reservation module, we take the use case in Sect. 3.2 as an example.

actual number of data records in each grid specified by the ideal granularity, and the requested description providing the *requested number* of data records in each grid defined by the given set division. Next, we discretize the requested descriptions to align with the former grid and yield corresponding *reconstructed numbers*. There exists a structural difference between each pair of the actual number and the reconstructed number. We quantify the structural error E_s by calculating the average of the relative error (i.e., the difference relative to the actual number) in each grid. The noise generated by DP also adds numerical error to the requested result. The confidential interval of a requested value depends on the paid privacy budget. Nevertheless, the target joint distribution could be indirectly described by the requested values—a data request can employ more than two attributes to define the set division and recover the target distribution by sum operations. In this case, the privacy budget could exert a higher influence on the reconstructed description. Like the structural error, we also consider the number of data records in each grid split by the ideal granularity as the target number. We further calculate the 95% confidential interval of the target number. The half length of the interval relative to the actual number is considered as the numerical error E_n. The sum of the structural error E_s and the numerical error E_n is employed as a penalty, namely a negative bonus, for accuracy assessment.

Generation of simulated data: The disclosure of the original data is forbidden until the privacy budget is paid. In practice, we generate non-public descriptions of sensitive distributions for accuracy calculation based on simulated data. Specifically, we apply a Gaussian copula to simulate the multi-variable probability distribution of records based on the input and generate N samples, where N is the number of records in the database. Available input consists of distributions over each attribute and Spearman's rank correlation coefficient between each pair of non-categorical attributes. Besides public information, the input of the Gaussian copula could be derived from feedback on previous data requests, the prior knowledge of users, and random results (set as default). The reason why we employ random results instead of uniform distribution is because the latter yields consistent simulations, which may lead users to ignore diverse possibilities.

Bonus for budget consumption: Ideally, budget consumption should be sufficient and consistent with the progress of data exploration, otherwise the remaining budget could either fail to enable further data exploration or be wasted. The progress of data exploration can be estimated by users. Nevertheless, the estimated progress rate could be inaccurate, especially at the beginning of the exploration process. Thus, we simulate a growing expectation of consistency in the exploration process. The bonus for budget consumption B_{bc} is quantified by the following equation:

$$B_{bc} = w * N * \frac{p+1}{2} * |p - \frac{\varepsilon_{total} - \varepsilon_{remain}}{\varepsilon_{total}}|,$$

where ε_{remain} and ε_{total} are the remaining privacy budget after the exploration strategy is implemented and the total privacy budget, p is the rate of exploration progress estimated by users, N is the number of records in the database, and w is an empirical parameter (default as -0.1) used for bonus weighting. We default p as 100%, which indicates that it is feasible to spend all remaining privacy budget to satisfy the specified exploration intent.

4.3 Interface Design

Defogger's interface is designed to support the pipeline (Sect. 4.1) and refined based on feedback provided by one participant in a pilot user study. As shown in Fig. 1, our interface includes three coordinated views, which display the information reserved by users, recommendations for data request declaration, and illustrations for uncertainty caused by noise.

4.3.1 Information Reservation

The information reservation view (Fig. 1(a)), represents the progress of data exploration by visualizing the exploration intent of users and the known data facts (Fig. 3(a)). Visualizations in this view are dynamically updated with the exploration progress. As shown in Fig. 1(a1), data facts specified as the target of data exploration are summarized as an intent graph, where nodes are attributes of interest and edges correspond to correlations among attributes. The stroke color of nodes enables differentiating sensitive attributes and public attributes whose distribution can be checked without spending a privacy budget. The visualizations on nodes demonstrate distributions over the corresponding attributes. The distributions of public attributes are known data facts shown by dark gray lines. The distributions shown in light gray lines are simulated distributions, whose data facts are unknown and waiting for exploration. After the distribution is returned by a data request, a dark gray line that describes the requested distribution will be superimposed on the node. The width of the new line visualizes the length of the 95% confidential interval of the distribution to highlight the uncertainty in the requested results. The distribution in dark gray will be updated if a requested distribution that is supposed to have a higher accuracy is received. Users can identify the differences between the simulation and the requested results with noise. If a large difference is observed, users may need to adjust their exploration strategy. Data facts about attribute correlations specified as exploration targets could be known or waiting for exploration as well. We employ color encodings of edges to make a distinction. Furthermore, we label edges between two non-categorical attributes with values of Spearman's rank correlation coefficient. Exactly like the distributions shown in the node, the values labeled on an edge could have a simulated version and a requested version.

Sketch exploration intent: Users can drag an attribute of interest from the attribute list to the sketch panel. After an attribute is dropped, the sketch panel creates dashed edges from the dropped attributes to other attributes on the panel. Users can click a dashed edge to admit their interest in the correlation between the attribute pair. All nodes and the solid lines in blue are regarded as data facts to be explored and considered to recommend data requests.

Estimate the progress of exploration: Users can input a value as the estimated rate of exploration progress after the current exploration intent is satisfied. The value is used to calculate bonus for budget

consumption (please refer to Sect. 4.2). After a privacy budget is spent for a data request, the lower bound of the estimated rate will be updated to $\frac{\varepsilon_{total}-\varepsilon_{remain}}{\varepsilon_{total}}$, where ε_{remain} and ε_{total} are the remaining privacy budget and the total privacy budget.

Generate data simulations: Simulations of the distribution over a single attribute or correlation metrics are applied to simulate joint distributions, assess the accuracy of feedback, and recommend data requests. If users are not satisfied with the default simulation, they can randomly reset it or specify it according to their prior knowledge. After clicking a node in the intent graph, users can interact with a histogram under the sketch panel to specify the distribution of the selected attribute (Fig. 1(a2)). The bins of the histogram are in the minimum granularity for data requests. However, the entire range of all bins could be inconsistent with the actual range of data records in the database because data owners may employ larger ranges to avoid leakage of extreme values. As for correlation coefficients between two numerical attributes, users can input the inferred number by double-clicking the annotation on the edge connecting the corresponding nodes.

4.3.2 Data Request Declaration

The second view allows users to declare data requests (Fig. 3(c)) in a panel (Fig. 1(b1)). Below the panel, we show a list of exploration strategies suggested by the recommendation model, as shown in Fig. 1(b2), to facilitate planning data requests. Each strategy is summarized by an overview glyph that annotates the attribute sets used in each data request on the intent graph by bubbles. The glyph can illustrate how data requests in a recommended strategy are incorporated to achieve the exploration intent. The total privacy budget required by the strategy is marked on the glyph. We also attach details of data requests (i.e., the allocated budget and attribute names) next to the glyph.

Declare a data request: Users can select a data request in the strategy list to autofill in the panel (Fig. 1(b1)). Edit operations are also allowed in the panel before simulating or submitting the data request.

4.3.3 Uncertainty Illustration

In the uncertainty illustration view (Fig. 1(c)), users can understand the uncertainty in the responses to data requests (Fig. 3(e)) or response simulation by checking descriptions of uncertainty (Fig. 3(d)). To ensure the expressiveness of uncertainty illustration, we integrate visual designs for uncertainty into two relatively universal visualizations for distribution or correlation, which are histograms [12] and heatmap matrices [31]. To represent uncertainty, we draw error bars on histograms to annotate the range of %95 confidential intervals, as shown in Fig. 4(a).

Matrices can reflect correlation patterns of multiple attributes (Fig.4(b)). In a matrix, attributes are arranged in a row and a column. Each cell in the matrix displays grids to describe the set-based joint distribution of two attributes. The color of each grid encodes the number of individuals in a set. We use contrasting colors to map a range of values from negative to positive because quantities in grids could be negative due to the noise. Although the correlation between a pair of attributes can be reflected by colorful grids in a cell, common visual representations of uncertainty (e.g., error bars, ribbons) can hardly be integrated into a grid without incurring visual clutter. Design schemes that can adapt to grids mainly fall into two categories: overlaid texture and Mosaic design [8]. The latter represents complicated patterns with multiple colored shapes. We proposed two implementation alternatives (one for each category): (i) a texture of dots that demonstrates the size of noise by the density/number of points (Fig. 4(c1)) and (ii) the Mosaic design that divides two triangles from the upper left corner and lower right corner of a grid to visualize the upper and lower bounds of the confidential interval respectively (Fig. 4(c2)). We finally adopted the latter because it is more aesthetic. If the uncertainty in the count encoded by a grid is relatively high, the color differences in the grid will be conspicuous to users. Nevertheless, the color differences are difficult to quantify. To provide intuitive illustrations like dots, we further encode the length of the 95% confidential interval by the width of the triangle. The width ratio of the triangles to the grid is equal to half the length of the interval divided by the count. Further, we employ

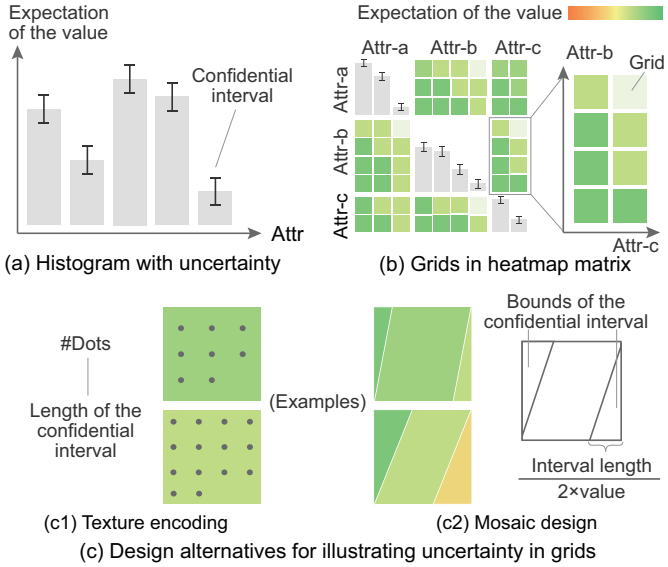


Fig. 4: Visualizations for uncertainty illustration. (a) Histogram with error bars for uncertainty in distribution over a single attribute. (b) Heatmap matrix representing correlations among multiple attributes, which requires grid-based uncertainty representation. (c) Two alternatives for uncertainty representation used in heatmap matrices.

a rainbow color mapping (instead of a two-color gradient) to reduce the difficulty of distinguishing values, which is mentioned by the participant of our pilot study. The intent is users will not be bothered by the illustration of the uncertainty if its effect is negligible. For example, the upper example shown in Fig. 4(c2) indicates a smaller effect of noise compared with the lower example. We set a maximum width for the triangles to prevent the middle stripes whose color encodes the requested/simulated numbers from being completely obscured.

Filter data records: As mentioned in Sect. 4.1, the sets used in count requests can be defined with multiple attributes. However, even the cells in matrices can only represent the individual number in a set defined by two attributes. To check individual counts in sets defined by more attributes, users can apply value-based filters of the attributes that are used to specify set division in data requests.

Browse possibilities of actual data: As for uncertainty in the feedback of data requests, users can further browse possible instances of the data distributions inferred by the feedback. After switching to the noise-removed mode, users can click the button "Reset" to generate a new instance of data distribution. The button can be clicked multiple times to support understanding of randomness. The likelihood of each instance presenting coincides with the probability of producing a corresponding amount of noise.

Review the summary of requested results: Users can switch to the summary mode to review all existing requested results in a matrix. If a correlation is requested multiple times, we show the results with the smallest 95% confidential interval. There could be empty cells in the matrix, which reminds users of exploration gaps.

5 Evaluation

To assess the proposed approach, we carried out a user study and recorded usage processes implemented by two subjects as case studies.

5.1 Case Studies

In this section, we outline the exploration process of two of the study participants as a pair of case studies. To support anonymity, we reference these participants with the pseudonyms Tom and Susie.

5.1.1 Target Clients Identification

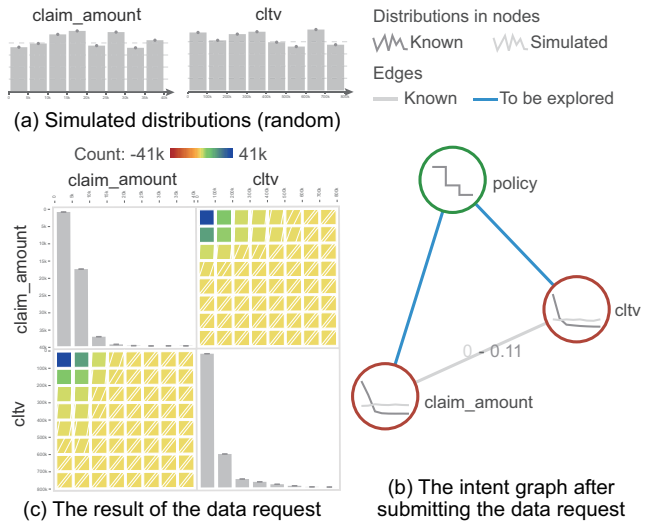
In this case, Tom played the role of an insurance company employee. His exploration goal was to identify the features of consuming behav-

iors shared by target clients who are willing to pay high premiums but claim a small reimbursement. For this goal, he needed to explore the client dataset [20] with a privacy budget of one. The dataset includes 89,392 client records.

Table 1: Descriptions of attributes in the database of clients, which were employed in the case. We remark each numerical attribute with the value range and the minimum interval for set division. As for categorical attributes, the remark consists of categorical values. Sensitive attributes are colored in red. (The next table is the same.)

Attribute	Description	Remark
claim_amount	Total amount claimed.	$[0, 40k], 5k$
policy	Active policy of the client.	A, B, C
cltv	Client lifetime value.	$[0, 800k], 50k$

The definition of target clients was derived from attributes claim_amount and cltv, two sensitive and numerical attributes. After checking the descriptions of all 11 attributes (see Tab.1), Tom guessed that the public categorical attribute policy could be related features of the target clients. He constructed an intent graph with the above three attributes and specified all edges as exploration intent. When browsing the data requests recommended by Defogger based on the intent graph, he noticed that numerical attributes could contribute to various schemes for set division because of multiple mergeable intervals. Nevertheless, the recommended schemes use the finest granularity directly. Tom found that the recommendation basis of the set division was the simulated distributions, which were generated randomly (see Fig. 5(a)). He reckoned that accurate data simulation is essential for the recommendation model to provide effective suggestions. However, Tom has no prior knowledge of the two sensitive attributes, which can contribute to data simulations. To address this issue, he decided to implement a progressive data exploration, which first requests the joint distribution of the two numerical attributes to find an appropriate set division and then requests statistical results that can describe the correlations among all three attributes.



(a) Simulated distributions (random)

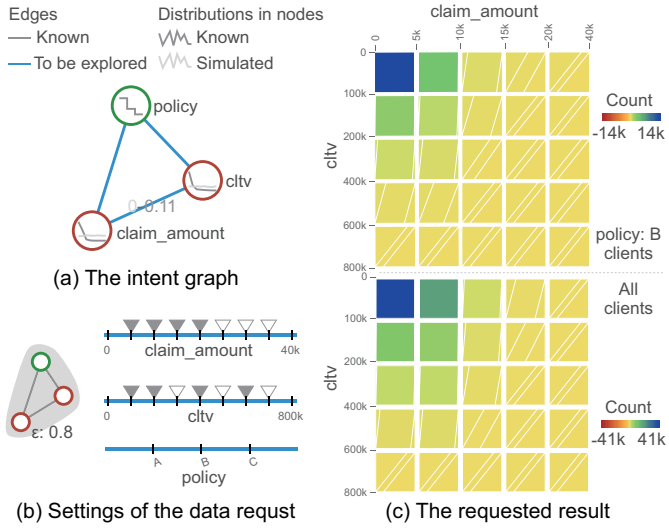
(b) The intent graph after submitting the data request

(c) The result of the data request

Fig. 5: Visualizations used in the first data request.

To start with the progressive exploration, Tom manually declared the data request by specifying the finest grids generated by the two numerical attributes for set division. After weighing the importance of the requested results in the two stages, Tom allocated 20% of the total privacy budget and submitted the data request. The requested results demonstrated that both the two numerical attributes had a long-tail distribution, which is significantly different from the random simulation (see nodes in Fig. 5(b)). The joint distribution of the two attributes shown in Fig. 5(c) reflected that a majority of individuals gather in the upper-left grids. The size of triangles in other grids indicated that noise added a large amount of uncertainty to the numbers of corresponding sets. However, according to the color of those triangles, the overall distribution pattern could still be observed even when the bounds of confidence intervals are considered.

In the next step, Tom activated the edge between cltv and claim_amount in the intent graph (see Fig. 6(a)) to complete his exploration intent. Defogger included the result of the first data request in available data facts for strategy recommendation and updated the recommendation list. To figure out the correlations among all three attributes, Tom selected the recommended strategy that achieved his exploration intent by requesting data with the set division employing the three attributes simultaneously and all the remaining privacy budgets. The new set division canceled the split points with large values of claim_amount, as shown in Fig. 6(b). Considering that only an extremely small number of clients had claimed high reimbursements, there was no need to make a further distinguishment. Tom believed that merging those intervals would not prevent him from achieving his exploration goals. Thus, he submitted the data request as recommended. The requested result (see Fig. 6(c)) indicated that clients who activated policy: B were more likely to be with a smaller claim_amount than others with similar cltv. Tom considered his finding reliable because the triangles in upper-left grids are invisible (see Fig 6(b)), which implies that the data pattern has slight uncertainty.



(a) The intent graph

(b) Settings of the data requst

(c) The requested result

Fig. 6: Visualizations used in the second data request, consisting of (a) the intent graph specified by the user, (b) details of request declaration, and (c) distribution comparison between policy: B clients and all clients.

5.1.2 Demographic Information and Hepatitis B

Susie was interested in the correlations between demographic information and the risks of hepatitis B. The second case describes how Susie explored the result of the National Health and Nutrition Examination Survey on 7,824 individuals [21] with a total privacy budget of two.

Table 2: Descriptions of partial demographic attributes in the database of survey results.

Attribute	Description	Remark
hepatitis_B	Have you been diagnosed with hepatitis B?	Y, N, idk
family_c	#People in the household.	1, 2,..., 7+
children_c	#Children aged $\leq$5 in the household.	0, 1, 2, 3+
teenager_c	#Children aged 6-17 in the household.	0, 1,..., 4+
elder_c	#Adults aged $\geq$60 in the household.	0, 1, 2, 3+

Although the survey result includes various record descriptions, Susie followed her exploration goal and constructed her graph of ex-

ploration intent with the five attributes shown in Tab. 2. The five attributes are connected as a star whose center is hepatitis_B, as shown in Fig. 1(a1). Next, Susie needed to input data facts for recommendations of exploration strategies. Among the five attributes, only the sensitive attribute hepatitis_B lacked the data fact of distribution. Susie needed to describe the distribution of hepatitis_B by inputting the proportion of three categories: Y, N, and idk (i.e., I don't know). According to the disease proportion of hepatitis B reported by the official website of the World Health Organization[2], that is 10.5%, Susie divided a small proportion into hepatitis_B: idk and divided the remaining with a ratio about 1:10 for hepatitis_B: Y and hepatitis_B: N respectively, as shown in Fig. 1(a2).

In the step of data request declaration, Susie had to determine the attribute groups for set division and the privacy budget to be paid. Susie sought suggestions by browsing the recommendations of exploration strategies (Fig. 1(b2)) based on the reserved information, as shown in Fig. 1(b). A data request employing three attributes for set division drew Susie's attention. To inspect how it works, Susie simulated the data request and checked the simulated results in the uncertainty illustration view. As shown in Fig. 1(c), the size of triangles in each grid indicates that the noise generated by DP could exert diverse effects on the statistical results of different sets due to the uneven distribution. The effect on dark blue grids corresponding to healthy individuals (hepatitis_B: N) whose family has no children (children_c: 0) or no elderly (elder_c: 0) is small enough to be negligible. While grids of hepatitis patients (hepatitis_B: Y) fail to reflect trustful correlation patterns. Susie hovered over one of those grids to inspect the specific numbers. It turned out that the half length of the confidential interval is greater than the simulated value in a majority of patient sets, which can hardly form a reliable conclusion. Patterns in those grids are necessary for Susie's exploration goal. Susie thought that her privacy budget might be insufficient to support data exploration of so many details. She had to narrow the scope of her exploration to avoid getting no trustful conclusions. After consideration, Susie decided to focus on the correlation between hepatitis_B and family_c and issued a data request that employs the two attributes for set division and spends all her privacy budgets.

The requested visualization (see the lower-right cell in Fig. 7(a)) demonstrated that the distribution of hepatitis_B is much more skewed than the simulated data facts. The noise generated by DP leads to more uncertainty in the patterns of hepatitis B patients than Susie expected. Thanks to the all-eggs-in-a-basket move, the data request returned correlation patterns with the smallest uncertainty. To focus on patterns of patients, Susie filtered out hepatitis_B: Y individuals to compare the family_c distribution of patients (see Fig. 7(b1)) with the distributions of all individuals (see the upper-left cell in Fig. 7(a)). According to the comparison, Susie came to a preliminary conclusion that individuals who live in a big family (family_c: 3, ..., 7+) suffer from a lower risk of hepatitis B than those living alone (family_c: 1). However, the size of the triangles in the row/column of hepatitis_B: Y demonstrate that the pattern could be not as significant as what is shown in the requested result. To make a verification, Susie further browsed possibilities of actual data by switching to the noise-removed mode. As shown in Fig. 7(b2), there exist differences in the instances of correlation patterns but all instances can contribute to the same conclusion. Finally, Susie confirmed her conclusion and recognized the results of her exploration.

5.2 User Study

To evaluate the feasibility of our approach (and the usability of Defogger) in supporting **R1** and **R2**, we conducted a user study with ten participants (eight graduate students recruited from the university campus and two practitioners in industry recruited from a research group on visual analytics). Our subjects have all completed a graduate-level course in data visualization, and are familiar with common interactive visual analysis tools/libraries (e.g., Tableau, D3, R, Python). In other words, they are capable of exploratory data analysis and represent realistic analyst user profiles for Defogger. Each subject received $15 USD for their participation which lasted about an hour. Because there is not a good competing "baseline" tool to compare again Defogger, we employed a single approach evaluation study [28].

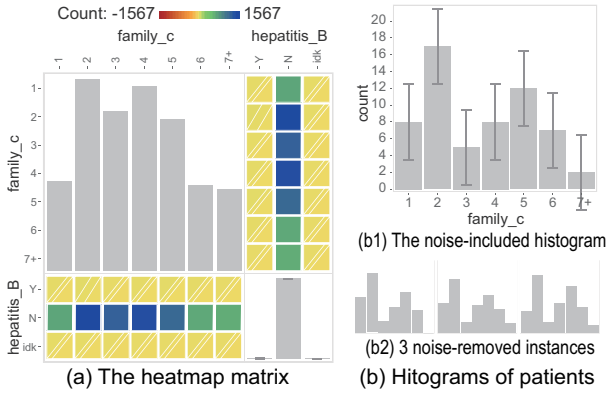


Fig. 7: Visualizations for the result of the data request on the correlation between family_c and hepatitis_B. (a) The heatmap summarizing the distribution of all individuals in the database. (b) Histograms that demonstrate the family_c distribution of hepatitis_B: Y individuals.

Study Design: The study comprised three steps. (1) In the training step, we provided an introduction of Defogger, consisting of the application scenario, the analysis workflow, the mechanism of the recommendation model, and the interface design based on a demo of Defogger. During this, subjects could ask questions at any time and receive detailed answers. The average time for this training step was 22.2 minutes. After training, we showed subjects two databases that could be explored in the next two steps. In addition to descriptions of databases (see Sect. 5.1), we provided a suggested set of exploration goals that subjects could use to help construct realistic exploration scenarios. An example of such a goal is to identify features shared by clients who can bring more benefits to insurance companies. Subjects could also define their own exploration goals of interest. (2) The second step familiarized subjects with the effects of DP, as some did not have experience exploring data within the constraints of DP. In this step, participants needed to select a database. Subjects were told that their trial of Defogger in the last step had to explore the other database to avoid exceeding privacy budgets. Next, they were allowed to test data requests with the cost of a privacy budget to the selected dataset and check feedback including noise generated by DP, which took about 6.6 minutes on average. (3) In the third step, subjects could freely explore the datasets with Defogger while employing a think-aloud protocol to catalog their mental processes and actions. This step took an average of 27.7 minutes. To finish the study, participants completed a questionnaire of 11 usability-based questions (using a five-point Likert scale) and additionally could provide freeform comments/text about the tool and/or its user experience.

Results: Our questionnaire asked subjects to assess our approach from four aspects, as organized in Fig. 8. We analyze these responses, using freeform comments (both positive and critical) provided by participants for additional context, to assess our approach's strengths and drawbacks.

(Q1–Q2) *Information reservation*: All subjects agreed that sketching the exploration intent allows them to sort out the upcoming exploration process (Q1). As for the input of data facts, a subject expressed concerns about insufficient prior knowledge to specify a reasonable distribution for simulation (Q2). Although two subjects asked for suggestions on distributions, such a request is unattainable because "free information" is forbidden by DP.

(Q3–Q5) *Declaration of data requests*: Concerns about inaccurate data simulation can cause users to discount the effectiveness of previewing the simulated results (Q4). That said, many subjects still acknowledged the contributions of the preview function to strategy

[2] https://www.who.int/news-room/fact-sheets/detail/hepatitis-b

selection because it is necessary to figure out how noisy the requested distribution could be when a certain budget is paid. The recommendations for exploration strategies (Q3), especially budget allocation, were considered valuable suggestions by most subjects. However, one subject complained that making a selection from recommendations was difficult because there were too many choices when the exploration intent included a large number of nodes (Q5). To address this issue, we refined Defogger to sort recommendations according to the assessment result of accuracy (please refer to Sect. 4.2). Based on the sorted list, users can focus on top strategies when they are not interested in diverse strategy candidates.

(Q6–Q7) *Uncertainty illustration*: We received divergent ratings on the representation of uncertainty in grids (Q6). Critical feedback mentioned two drawbacks: unclear information representation and complicated visual design. The subject who rated a 2 for Q6 told us that the visual representation of uncertainty hindered their understanding of the value returned by the data request. This subject showed us a grid overlapped with two triangles that almost took up the entire grid. When it was explained that the returned value is not reliable due to the uncertainty in such a case, the subject approved our design considerations. Other critical comments reflected that the visual design was unintuitive due to the integration of visual elements encoding the confidential intervals. They preferred visualizations representing simple data descriptions, like the heatmap matrices visualizing noise-removal instances (Q7). Those who voted for the uncertainty representation in grids praised the detailed descriptions: "through the observation of confidence intervals, users can gain a deeper understanding of the uncertainty distribution of the data." and "[i]t is very confusing and requires more time if there is no confidence interval."

(Q8–Q11) *Defogger Usability*: In general, Defogger was considered a handy (Q9) and useful (Q10) tool with a relatively low learning curve (Q8). We receive several positive comments, e.g., "Defogger provides a lot of auxiliary information that allows me to judge, to a certain extent, how much the queried data has been affected. It allows me to better draw some of the conclusions I want to get." One subject especially liked that Defogger can help users save their privacy budget by avoiding repeated attempts to make unsatisfactory data requests.

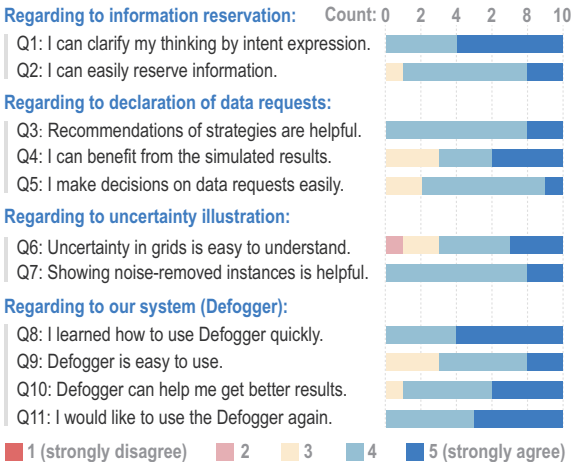


Fig. 8: The statistical result of the answers from ten subjects.

6 Discussion

To our knowledge, Defogger represents a first-of-its-kind approach for exploratory visual analysis augmented by privacy-aware recommendation intelligence.

Lessons Learned: We briefly summarize three primary lessons learned from the user study and the case studies in Sect. 5.

Partial users prefer to understand complex information by browsing multiple visualizations demonstrating simple information: In Sect. 5.2, we discuss how our subjects had opposite views on the Mosaic representation for uncertainty in grids. Several subjects who frequently used Python or R as data analysis tools rated the uncertainty illustration view in Defogger relatively low because they considered that visualizations integrating a variety of visual encodings were not effective at conveying information. Thus, it is necessary to enable a divide-and-conquer process to learn about complex information. Defogger enables such a process as an alternative approach to understanding uncertainty by providing noise-removed instances. More broadly, future efforts in this area can investigate how to balance visual complexity with privacy-based situational awareness.

Experienced data explorers can better steer the exploration process with DP constraints: The case of target customer identification in Sect. 5.1.1 describes a smart exploration strategy that is raised by a subject with rich experience in data exploration. In the future, we plan to identify exploration skills from the behavior patterns of experienced explorers and provide users with more assistance by integrating the summarized skills into Defogger, as more adept models would likely benefit users at all levels of experience.

User knowledge should be integrated into the process of strategy customization: Defogger recommends exploration strategies based on the exploration intent and available data facts reserved by users. The reserved information is integral to the recommendation model yielding feasible strategies. Defogger also allows users to declare data requests directly, as users should have the autonomy to formulate their own exploration strategies, especially at the beginning stage when a model will lack a basis to make effective recommendations, or users have difficulties in describing their exploration intent.

Current Tool Limitations and Additional Future Directions: A positive user experience during visual exploration relies on quick (i.e., interactive) responses to flexible data requests, which relies on robust software engineering and data management. At current, Defogger focuses on count-based data requests as this is the most versatile type of visual exploration request. However, how to best coordinate data requests of various types in privacy preserving scenarios is a challenging but essential problem for flexible data exploration. When there are a larger number of possible exploration strategies, the employed recommendation model requires a longer period to generate appropriate strategy recommendations. For example, it took the recommendation model about two minutes to generate recommendations for the second exploration intent in the first case study (Sect. 5.1.1). While we consider such a duration acceptable for a first-of-its-kind research effort, it is certainly limiting in many real-world scenarios. Feasible future research directions include pre-caching or shrinking the strategy space [19] based on the exploration behaviors of experienced data explorers.

7 Conclusion

Visual exploration scenarios that query datasets containing sensitive information must adhere to privacy requirements. In particular, DP and privacy budgets can strain the exploration process, and users need intelligent strategies to optimize their workflows. Based on a pre-study and requirements analysis, we propose and develop a novel visual analysis approach, implemented in a prototype system called Defogger, which integrates a reinforcement learning model to recommend exploration strategies based on users' exploration intent and illustrates uncertainty in data distributions by a group of visualizations. We conduct a user study and present two case studies to demonstrate that Defogger effectively supports users in making valuable findings from sensitive data while still adhering to required DP constraints, and discuss lessons learned that are generalizable to the community and future research efforts. The codebase for Defogger can be found at `https://github.com/Vanellope7/Defogger`.

Acknowledgment

This work was supported in part by the NSFC (62202244), "the Fundamental Research Funds for the Central Universities," Nankai University, and by the U.S. National Science Foundation through grant CNS-2224066.

REFERENCES

[1] S. Alspaugh, N. Zokaei, A. Liu, et al. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, 2019. doi: 10.1109/TVCG.2018.2865040 2

[2] L. Battle, R. Chang, and M. Stonebraker. Dynamic Prefetching of Data Tiles for Interactive Visualization. In *Proceedings of the 2016 International Conference on Management of Data*, pp. 1363–1375. ACM, New York. doi: 10.1145/2882903.2882919 2

[3] L. Battle and J. Heer. Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau. *Computer Graphics Forum*, 38(3):145–159, 2019. doi: 10.1111/cgf.13678 1, 2

[4] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124 2

[5] P. Bukaty. *The California Privacy Rights Act (CPRA)–An implementation and compliance guide*. IT Governance Publishing, 2021. doi: 10.2307/j.ctv1kv1d14 1

[6] C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042 1

[7] H. Ebadi, D. Sands, and G. Schneider. Differential Privacy: Now it's Getting Personal. *ACM SIGPLAN Notices*, 50(1):69–81, 2015. doi: 10.1145/2775051.2677005 2

[8] S. Goodwin, J. Dykes, A. Slingsby, and C. Turkay. Visualizing Multiple Variables across Scale and Geography. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):599–608, 2016. doi: 10.1109/TVCG.2015.2467199 6

[9] J. Görtler, C. Schulz, D. Weiskopf, and O. Deussen. Bubble Treemaps for Uncertainty Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):719–728, 2018. doi: 10.1109/TVCG.2017.2743959 3

[10] J. Heer and B. Shneiderman. Interactive Dynamics for Visual Analysis: A Taxonomy of Tools That Support the Fluent and Flexible Use of Visualizations. *Queue*, 10(2):30–55, 2012. doi: 10.1145/2133416.2146416 2

[11] J. Hullman, X. Qiao, M. Correll, et al. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, 2019. doi: 10.1109/TVCG.2018.2864889 2

[12] Y. Ioannidis. The History of Histograms (abridged). In *Proceedings 2003 VLDB Conference*, pp. 19–30. Morgan Kaufmann, San Francisco. doi: 10.1016/B978-012722442-8/50011-2 6

[13] M. F. S. John, G. Denker, P. Laud, et al. Decision Support for Sharing Data using Differential Privacy. In *2021 IEEE Symposium on Visualization for Cyber Security*, pp. 26–35. New Orleans. doi: 10.1109/VizSec53666.2021.00008 2, 3

[14] M. Kay. ggdist: Visualizations of Distributions and Uncertainty in the Grammar of Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):414–424, 2024. doi: 10.1109/TVCG.2023.3327195 3, 4

[15] B. C. Kwon, M.-J. Choi, J. T. Kim, et al. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, 2019. doi: 10.1109/TVCG.2018.2865027 1, 2

[16] H. Li, L. Xiong, X. Jiang, and J. Liu. Differentially Private Histogram Publication for Dynamic Datasets: an Adaptive Sampling Approach. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1001–1010. New York, 2015. doi: 10.1145/2806416.2806441 2

[17] S. Li, X. Ji, and W. You. A Personalized Differential Privacy Protection Method for Repeated Queries. In *2019 IEEE 4th International Conference on Big Data Analytics*, pp. 274–280. Suzhou. doi: 10.1109/ICBDA.2019.8713224 3

[18] W. Li, L. Xiang, B. Guo, et al. DPlanner: A Privacy Budgeting System for Utility. *IEEE Transactions on Information Forensics and Security*, 18:1196–1210, 2023. doi: 10.1109/TIFS.2022.3231786 1

[19] K. Menda, Y.-C. Chen, J. Grana, et al. Deep Reinforcement Learning for Event-Driven Multi-Agent Decision Processes. *IEEE Transactions on Intelligent Transportation Systems*, 20(4):1259–1268, 2019. doi: 10.1109/TITS.2018.2848264 9

[20] S. Mohapatra. Customer Life Time Value. https://www.kaggle.com/datasets/shibumohapatra/customer-life-time-value, 2023. 7

[21] P. Mooney, S. Dane, et al. National Health and Nutrition Examination Survey. https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey, 2017. 7

[22] P. Nanayakkara, J. Bater, X. He, et al. Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases. *Proceedings on Privacy Enhancing Technologies*, 2022(2):601–618, 2022. doi: 10.2478/popets-2022-0058 2, 3, 4

[23] Y. Ouyang, Y. Wu, H. Wang, et al. Leveraging Historical Medical Records as a Proxy via Multimodal Modeling and Visualization to Enrich Medical Diagnostic Learning. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1238–1248, 2024. doi: 10.1109/TVCG.2023.3326929 1, 2

[24] X. Qian, R. A. Rossi, F. Du, et al. Learning to Recommend Visualizations from Data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1359–1369. New York, 2021. doi: 10.1145/3447548.3467224 2

[25] R. Sahann, T. Müller, and J. Schmidt. Histogram binning revisited with a focus on human perception. In *2021 IEEE Visualization Conference*, pp. 66–70. New Orleans. doi: 10.1109/VIS49827.2021.9623301 4

[26] J. Sanyal, S. Zhang, G. Bhattacharya, et al. A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1209–1218, 2009. doi: 10.1109/TVCG.2009.114 2

[27] L. Shi, C. Huang, M. Liu, et al. UrbanMotion: Visual Analysis of Metropolitan-Scale Sparse Trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):3881–3899, 2021. doi: 10.1109/TVCG.2020.2992200 1, 2

[28] M. Tory. User studies in visualization: A reflection on methods. In *Handbook of Human Centric Visualization*, pp. 411–426. Springer, 2013. doi: 10.1007/978-1-4614-7485-2_16 8

[29] P. Voigt and A. Von dem Bussche. *The EU General Data Protection Regulation (GDPR)*. Springer Cham, 2017. doi: 10.1007/978-3-319-57959-7 1

[30] H. Wang, Z. Xu, S. Jia, et al. Why current differential privacy schemes are inapplicable for correlated data publishing? *World Wide Web*, 24:1–23, 2021. doi: 10.1007/s11280-020-00825-8 2

[31] X. Wang, W. Chen, J. Xia, et al. ConceptExplorer: Visual Analysis of Concept Drifts in Multi-source Time-series Data. In *2020 IEEE Conference on Visual Analytics Science and Technology*, pp. 1–11. Salt Lake City. doi: 10.1109/VAST50239.2020.00006 6

[32] X. Wang, J.-K. Chou, W. Chen, et al. A Utility-Aware Visual Approach for Anonymizing Multi-Attribute Tabular Data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):351–360, 2018. doi: 10.1109/TVCG.2017.2745139 2, 3

[33] K. Wongsuphasawat, D. Moritz, A. Anand, et al. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016. doi: 10.1109/TVCG.2015.2467191 2

[34] K. Wongsuphasawat, Z. Qu, D. Moritz, et al. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2648–2659. ACM, New York. doi: 10.1145/3025453.3025768 2

[35] J. Xu, Z. Zhang, X. Xiao, et al. Differentially Private Histogram Publication. In *2012 IEEE 28th International Conference on Data Engineering*, pp. 32–43. Arlington. doi: 10.1109/ICDE.2012.48 2, 3

[36] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, New York. doi: 10.1145/3173574.3174053 2

[37] D. Zhang, A. Sarvghad, and G. Miklau. Investigating visual analysis of differentially private data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1786–1796, 2021. doi: 10.1109/TVCG.2020.3030369 2, 3

[38] S. Zhang, H. Li, M. Wang, et al. On the Convergence and Sample Complexity Analysis of Deep Q-Networks with *epsilon*-Greedy Exploration. In *Advances in Neural Information Processing Systems*, pp. 13064–13102. Curran Associates, Inc., New Orleans, 2023. doi: 10.48550/arXiv.2310.16173 4

[39] X. Zhang, S. Chandrasegaran, and K.-L. Ma. ConceptScope: Organizing and Visualizing Knowledge in Documents based on Domain Ontology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, New York. doi: 10.1145/3411764.3445396 3

[40] Z. Zhang, T. Wang, N. Li, et al. PrivSyn: Differentially Private Data Synthesis. In *30th USENIX Security Symposium*, pp. 929–946. USENIX Association, 2021. doi: 10.60882/cispa.24613539 2

[41] J. Zhao, M. Fan, and M. Feng. ChartSeer: Interactive Steering Exploratory Visual Analysis With Machine Intelligence. *IEEE Transactions on Visualization and Computer Graphics*, 28(3):1500–1513, 2022. doi: 10.1109/TVCG.2020.3018724 2

[42] J. Zhou, X. Wang, J. K. Wong, et al. DPVisCreator: Incorporating Pattern Constraints to Privacy-preserving Visualizations via Differential Privacy. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):809–819, 2023. doi: 10.1109/TVCG.2022.3209391 2

[43] Z. Zhou, X. Wen, Y. Wang, and D. Gotz. Modeling and Leveraging Analytic Focus During Exploratory Visual Analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, New York. doi: 10.1145/3411764.3445674 2

[44] T. Zhu, G. Li, W. Zhou, and S. Y. Philip. Differentially Private Data Publishing and Analysis: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1619–1638, 2017. doi: 10.1109/TKDE.2017.2697856 2