Journal of *Cybersecurity and Privacy*

MDPI

Article

SMARTEN—A Sample-Based Approach towards Privacy-Friendly Data Refinement

Christoph Stach [1,*], Michael Behringer [1], Julia Bräcker [2], Clémentine Gritti [3] and Bernhard Mitschang [1]

[1] Institute for Parallel and Distributed Systems, University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany
[2] Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Allmandring 5B, 70569 Stuttgart, Germany
[3] Department of Computer Science and Software Engineering, University of Canterbury, Christchurch 8041, New Zealand
* Correspondence: christoph.stach@ipvs.uni-stuttgart.de; Tel.: +49-711-68588-433

Abstract: Two factors are crucial for the effective operation of modern-day smart services: Initially, IoT-enabled technologies have to capture and combine huge amounts of data on data subjects. Then, all these data have to be processed exhaustively by means of techniques from the area of big data analytics. With regard to the latter, thorough data refinement in terms of data cleansing and data transformation is the decisive cornerstone. Studies show that data refinement reaches its full potential only by involving domain experts in the process. However, this means that these experts need full insight into the data in order to be able to identify and resolve any issues therein, e.g., by correcting or removing inaccurate, incorrect, or irrelevant data records. In particular for sensitive data (e.g., private data or confidential data), this poses a problem, since these data are thereby disclosed to third parties such as domain experts. To this end, we introduce SMARTEN, a sample-based approach towards privacy-friendly data refinement to smarten up big data analytics and smart services. SMARTEN applies a revised data refinement process that fully involves domain experts in data pre-processing but does not expose any sensitive data to them or any other third-party. To achieve this, domain experts obtain a representative sample of the entire data set that meets all privacy policies and confidentiality guidelines. Based on this sample, domain experts define data cleaning and transformation steps. Subsequently, these steps are converted into executable data refinement rules and applied to the entire data set. Domain experts can request further samples and define further rules until the data quality required for the intended use case is reached. Evaluation results confirm that our approach is effective in terms of both data quality and data privacy.

Keywords: privacy; data refinement; data cleansing; data transformation; human-in-the-loop

Citation: Stach, C.; Behringer, M.; Bräcker, J.; Gritti, C.; Mitschang, B. SMARTEN—A Sample-Based Approach towards Privacy-Friendly Data Refinement. *J. Cybersecur. Priv.* **2022**, *2*, 606–628. https://doi.org/10.3390/jcp2030031

Academic Editor: Danda B. Rawat

Received: 20 April 2022
Accepted: 12 August 2022
Published: 15 August 2022

1. Introduction

Since in 2011 the World Economic Forum coined the phrase "data will be the new oil" [1], this allegory of data as the most valuable asset to be exploited has become increasingly entrenched [2]. Data as a resource are highly relevant in any domain, including research [3], economy [4], the public sector [5], and even private life [6]. None of the so-called smart services can operate without processing vast amounts of data—the better the quality and quantity of the data, the better the outcome [7].

This fact is often seen very positively, since data, unlike oil, are an almost indefinite resource. Today, it is possible for every user to be a data producer, and due to the *Internet of Things* (*IoT*), data can be gathered in any domain [8]. Moreover, unlike oil, processing data are clean, and, since data are not consumed when they are processed, they represent the ultimate renewable resource [9]. However, this should not be perceived quite so optimistic. Just like oil, data are virtually worthless in their raw state. Thorough processing techniques are required in order to refine them, and, in this process, enriching their information content

and generating a value [10]. Elaborate data science strategies are therefore needed to derive essential knowledge from raw data [11].

This is often illustrated by means of the *Data Pyramid* [12] (also known as *DIK Hierarchy*—data, information, knowledge). This three tier pyramid is shown in Figure 1. It is made evident that data are not only the foundation of this pyramid (and thus the source of all knowledge), but that they are also available in large numbers. This is due to the fact that, nowadays, almost every device is equipped with a variety of sensors and has the connectivity needed to share the captured data with others [13]. These data, however, are raw, i.e., they are unstructured and also contain some errors. Only by means of pre-processing steps (e.g., data cleansing or linking previously incoherent data), raw data become useful and valuable information. The actual analyses can only be carried out on the basis of this information. By applying highly use-case-specific data processing techniques, knowledge can be extracted from the information, which can then be used, for instance, by a smart service. Although the quantity was reduced in the process, the aggregation of the raw data has led to an increase in significance and thus value [14].

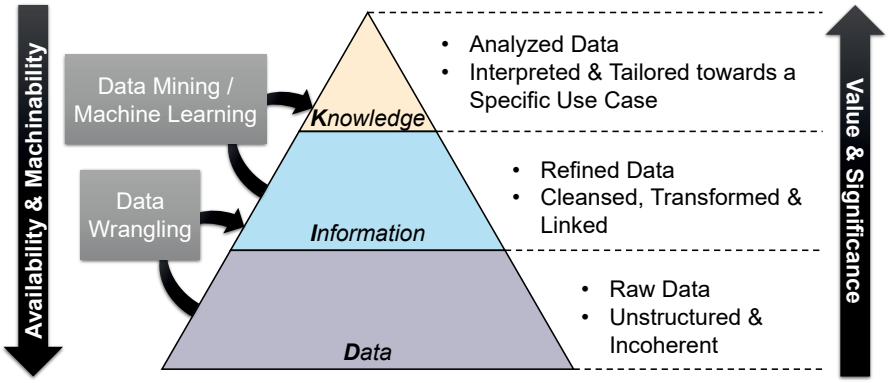

Figure 1. Data pyramid illustrating the relationships between data, information, and knowledge.

This data refinement process therefore involves data processing at two stages: First, raw data must be prepared using data wrangling techniques to extract processable assets [15]. Then, processing techniques from the field of data mining or machine learning can be applied to generate an added value for the intended use case [16]. Although it may seem at first that the latter is the more valuable task, studies confirm that the foundation of any successful data analysis is actual its preceding data preparation [17].

This, however, raises security concerns when dealing with sensitive data [18]. In the course of data refinement, the data processor gains deep insights into the data. When pre-processing and analyzing personal data [19] or confidential data [20], it is therefore essential for the data subjects and data owners to ensure that no private information is disclosed to third parties involved in the process [21]. However, this confronts data owners with the *privacy paradox*: The best way to protect non-disclosure interests of data owners and data subjects is to provide data processors with no data at all, or at least only highly distorted data. This, however, also implies that no knowledge can be gained from the data, and smart services that rely on it would be either inoperable or ineffective. A user only experiences the best possible level of service if a data processor is able to derive the highest possible degree of knowledge from all available data. Unfortunately, this also means that a data owner has to expose all data and thus lose control over his or her data entirely. This is therefore not an option either, especially for sensitive data [22].

To this end, we introduce **SMARTEN**, a **s**a**m**ple-based **a**pproach towards p**r**ivacy-friendly da**t**a r**e**finemen**t**. With SMARTEN, users are enabled to share sensitive data with smart services without the risk of unwanted insights being gained during data processing. In other words, we *smarten up* smart services in a privacy-friendly manner. To this end, we make the following three contributions:

1. We adapt the **data preparation phase** of the data refinement process. Data processors are not granted access to the full data stock, but rather to **representative samples** that comply with the privacy requirements. They then define data wrangling activities on these samples. These definitions are generalized as data pre-processing rules and applied to the entire data stock.
2. We adapt the **data processing phase** of the data refinement process. By means of **privacy filters**, different versions of the pre-processed data are generated. According to the privacy requirements, a specific version is used for the analysis, and the resulting insights are provided to the data consumers (e.g., smart services). As a result, each use case can be provided with the required knowledge without exposing any sensitive information.
3. We enable data owners to annotate their data with **privacy requirements** by means of **digital signatures**. These privacy requirements define how profoundly a data processor and a data consumer can inspect the data. Analogous with the data refinement process, the privacy requirements are also structured in two parts to determine in which manner access to the data are granted during pre-processing and what insights can be gained during the analysis.

The remainder of this paper is structured as follows: We initially discuss the data refinement process, describe its two phases as well as the tasks to be carried out, and identify key players in data processing in Section 2. Furthermore, we also address legal aspects that have to be considered in data processing and must be ensured by technical and organizational measures. In Section 3, we discuss such measures and how they resolve the privacy paradox. Based on these findings, we present our proposed solution SMARTEN in Section 4, which addresses shortcomings in related work. Our main focus is to ensure that all privacy requirements of a data owner are respected by the data processor for both the data preparation phase and the data processing phase of the data refinement process. In Section 5, we evaluate SMARTEN by means of a feature assessment to document its capabilities, by means of a user study to prove the applicability of our approach, and by means of a performance measurement to determine its overhead in terms of runtime and memory usage. Finally, we wrap up our work in Section 6.

2. Data Refinement Process

The cornerstone of effective data refinement is a systematic process model that gradually elevates raw data to a higher level [23]. To this end, there are several process models that address this matter from different directions and therefore each has its own specific focus. One such process model is *SEMMA* [24]. SEMMA stands for the individual process steps, namely *sample*, *explore*, *modify*, *model*, and *assess*. These steps are deliberately designed highly generically in order to be applicable to any kind of data science task. The involved activities are therefore defined very broadly. Thus, it is only specified that initially a significant amount of base data has to be selected and an overview of these data has to be obtained. Then, these data have to be prepared and processed in such a way, as it is "appropriate" for a specific goal. Finally, the success—i.e., the quality of the knowledge gained—has to be evaluated. These steps are strictly sequential. It can only be determined whether the data refinement was successful at the very end. If this is not the case, the cycle starts again from the beginning.

From an operational point of view, however, such a strategy is not efficient as it is not possible to identify and counteract aberrations at an early stage. Furthermore, a more precise specification of the individual activities is also required in practice. Therefore, the *Cross-Industry Standard Process for Data Mining* (*CRISP-DM*) [25] introduces an industry-oriented view of data refinement. Its practicability is clearly reflected by the fact that, before starting data refinement, knowledge about the business domain has to be gathered. This way, a cost–benefit analysis can be carried out in order to evaluate the economic viability of the respective data science project. For this reason, the deployment of the results of the data refinement is also an integral part of this process. Only when deploying the results, a

final review of the gained knowledge can be performed. In contrast to SEMMA, however, CRISP-DM introduces feedback loops between its individual process steps. As a result, the process is not rigidly run through sequentially as it is possible to return to an earlier step if a step is not completed satisfactorily.

All these process models stem from *Knowledge Discovery in Databases* (*KDD*) [26]. KDD is the first systematic process model for transforming data into a usable form (i.e., information) in order to gain knowledge from them. The process model consists of five core steps, which can be mapped almost directly to the five steps of the SEMMA model. These core steps are preceded by an implicit step in which knowledge about the application domain is assembled. Furthermore, another implicit step is following the core steps in which the discovered knowledge is applied. These two implicit steps correspond to the first and last steps of CRISP-DM. Despite the iterative approach of KDD, the process is designed to be interactive. That is, similar to CRISP-DM, it is possible to revert to a previous process step at any stage [27].

Table 1 provides a comparison of the individual steps of the three process models. It can be seen that KDD is the broadest model, and the other two can be mapped to it. Because of this property, KDD can be applied to any data science task [28]. Therefore, we draw on it to describe the data refinement process in this paper (see Figure 2).

Table 1. Contrasting Juxtaposition of the Individual Process Steps of CRISP-DM, KDD, and SEMMA.

CRISP-DM	KDD	SEMMA
Business Understanding	*Preceding the KDD Process*	-
Data Understanding	Selection	Sample
	Pre-Processing	Explore
Data Preparation	Transformation	Modify
Modeling	Data Mining	Model
Evaluation	Interpretation/Evaluation	Assess
Deployment	*Succeeding the KDD Process*	-

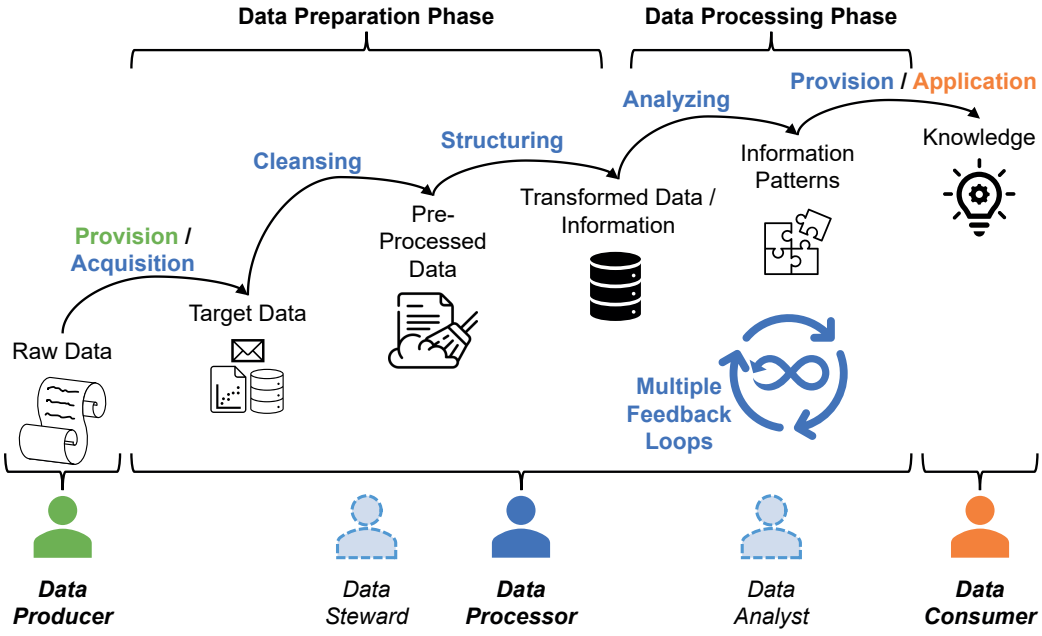

Figure 2. A data refinement process model in accordance with Fayyad et al. [26].

As shown in Figure 2, three key roles are involved in the data refinement process from an organizational perspective. A *data producer* initially generates and shares data in order to

get them refined. The IoT facilitates the capturing of different types of raw data. These data constitute the foundation for any smart service. Data producers therefore have an incentive to let their data be prepared and processed in order to feed these services. The processing itself is handled by the *data processor*. The processor first transforms the heterogeneous raw data into a processable form and then applies various analysis techniques to them in order to gain knowledge. This knowledge is then offered to a *data consumer*, who utilizes it, e.g., to operate a smart service [29].

Sensitive data, e.g., confidential or private data, may be involved in this process. A data producer therefore not only has privacy requirements with regard to the data consumer, who utilizes the knowledge contained in the data, but also with regard to the data processor, who gains deep insights into the raw data during data refinement. In our work, the stages of the refinement process in which the data processor is involved are particularly relevant. Here, not only the transformation of the data into information takes place, but also the analysis of the information that reveals the knowledge. This part of the process can be divided into two phases, the *data preparation phase* and the *data processing phase*.

Data Preparation Phase. This part of the data refinement process deals with data wrangling. That is, a *data steward* selects a set of base data from the available raw data that is worth further investigation. These target data must then be cleansed, since raw data always have quality issues. For instance, missing values have to be filled in, wrong values have to be corrected, redundant data have to be removed and imbalances in the datasets have to be corrected. The cleansed data must then be converted into a uniform structure that can be processed. A data steward needs deep insight into individual datasets in order to identify the quality issues within the base data [30].

Data Processing Phase. Based on the compiled information, a *data analyst* can recognize contained information patterns. Applied to the right situation, these patterns represent knowledge. Such patterns are supposed to answer questions such as "What happened?" (descriptive), "Why did it happen?" (diagnostic), "What will happen?" (predictive), and "What should I do?" (prescriptive). Thus, the base data are reflected in this knowledge only in a highly aggregated and condensed form. Therefore, a data analyst does not need an insight into individual datasets to accomplish his or her task. However, the derived knowledge reveals much more insight than a single set of raw data [31].

Although it might seem that the second phase is more relevant, as it reveals the valuable knowledge, the focus of data refinement is actually on the first phase. In fact, the data preparation phase accounts for up to 80% of the data refinement, and, without thorough data wrangling, the subsequent data analysis cannot be conducted successfully [32].

However, since such data refinement can reveal a lot of confidential knowledge, data subjects have to be protected. Legislators have therefore enacted data protection laws that are designed for the age of the IoT and big data analytics, e.g., the *General Data Protection Regulation* (*GDPR*) [33] which is in effect in the European Union. The GDPR obliges that data subjects give their consent before their data are processed in any form (Art. 6(1)(a)). In addition, even then, a data processor is required to comply with the principle of data minimization, i.e., the disclosed sensitive data must be limited to a necessary minimum (Art. 5(1)(c)). At the same time, the accuracy must be high and incorrect data have to be rectified (Art. 5(1)(d) and Art. 16). Data processors must be able to document compliance with these regulations to supervisory authorities (Art. 24). This requires not only organizational measures but also technical ones, referred to as *privacy by design* (Art. 25) [34].

In the following, we therefore discuss which technical measures can be taken to comply with data minimization without compromising the accuracy of the derived knowledge.

3. Related Work

When looking at the state of the art in terms of technical measures to minimize the information content of data, three fundamentally different approaches can be identified.

First, base data can be withheld in whole [35] or in part [36]. Second, base data can be aggregated [37], or certain aspects of the data can be filtered out [38]. Third, the data of a particular instance can be *hidden* in the bulk data [39].

Withhold Data. One of the most effective ways to protect sensitive information is to avoid sharing them with third parties and not let such data be processed at all. In practice, for instance, this would mean that access to certain sensors on an IoT device—and therefore their data as well—would be completely blocked [40]. As a result, the information captured by the sensor in question is no longer available for analysis. This, however, has a significant impact on data quality and the knowledge that can be derived from them. To mitigate this to some degree, the withholding of data can be bound to certain constraints. For instance, certain time slots or locations can be defined where the disclosure of the collected data are acceptable for the data subject [41]. Likewise, only those data items that contain particularly compromising information can be withheld, e.g., if the captured values are above or below a defined threshold [42]. From a technical point of view, this approach corresponds to the application of a *selection operator* σ adopted from relational algebra. That is, all tuples that do not satisfy a specified predicate are removed from the corresponding relation, i.e., a selection represents a *horizontal withholding of data* [43].

Instead of withholding the entire data from a sensor or a full dataset, only certain features of a sensor can be concealed in a fine-grained manner as well [44]. If a sensor captures multiple facts and combines them into a single measurement—e.g., a GPS sensor captures, among other things, longitude, latitude, and altitude—it is possible that only some of these features contain confidential information. In order to still be able to process the measurements of this sensor, only the affected feature can be withheld in this case. From a technical point of view, this corresponds to the application of a *projection operator* π adopted from relational algebra. That is, a subset of the available attributes is removed from all relations, i.e., a projection represents a *vertical withholding of data* [45].

Filter Data. However, withholding data represents a sledgehammer approach that severely impairs data quality and quantity. Filtering techniques take a more sophisticated approach. Here, only sensitive information is filtered out of the data. To this end, there are two opposing approaches: either the data are *condensed*, or they are *enriched* with synthetic data.

When condensing data, e.g., the data can be grouped by certain attribute and then only a single synthetically generated dataset per group, which is representative of the respective group, is forwarded for processing. From a technical point of view, this corresponds to the application of an *aggregate operator* $\mathcal{G}$ adopted from relational algebra. That is, all relations with common attribute value are initially grouped and then an aggregate function, e.g., SUM or MEDIAN, is applied to each group [46]. Instead of grouping all records by their attribute values, all records collected within a given time frame can be aggregated, alternatively [47]. However, in doing so, the data quantity is severely reduced since n records are mapped to a single aggregated value. To address this, data interpolation initially reduces the amount of data—and thus the disclosure of sensitive information—by processing only every n-th record of a data source. The omitted data points are then filled with suitable *mock data* to maintain the original amount of data [48].

When enriching data, randomly generated noise is usually added to the actual data in order to reduce the accuracy of the data—and thus the amount of information contained [49]. However, such a one-size-fits-all approach is not always effective. On the one hand, generic noise may not sufficiently mask certain confidential information in some cases, or it may be removed very easily [50]. On the other hand, such noise can be too intrusive and render the data unusable in the process [51]. Therefore, it is better to rely on privacy filters that are tailored to the particular data and their intended processing [52]. For instance, there are specialized approaches for location data [53], health data [54], time series data [55], voice data [56], and image data [57]. Such approaches ensure that the synthetically generated data appear plausible in the context of the data subject [58]. In addition to such filters targeting specific data sources, there are also higher-level filters for *events*, i.e., specific combinations of data from different sources [59].

Hide in the Masses. While the technical measures to conceal sensitive information described above primarily targeted data from individual users, the protection of such data can also be seen as a community task. If many individuals are willing to expose their data to a certain risk of disclosure, sensitive information can also be hidden in the masses of data. This principle can be explained very simply by means of an example: When General Crassus tried to arrest Spartacus, many followers of Spartacus claimed to be him. This way, the real Spartacus blurred with the masses and could no longer be identified as an individual. The more people thereby expose themselves to the danger of being arrested, the lower is the actual danger for each individual [60]. This fundamental idea also applies to the protection of confidential data. Often, the data themselves are not particularly sensitive as long as they cannot be linked to a data subject. Thus, instead of tampering with the data and thus degrading the quality of the data, it is often sufficient to pool the data of many individuals, so that a single record cannot be uniquely linked to any of them. It is only possible to tell that they belong to one of the k individuals that participate. In this way, it is possible to perform better analyses on the data than it would have been possible if arbitrarily filters are applied to the data of each individual [61].

Dwork [62] introduces *differential privacy*, a measure that can be used to assess the risk to an individual who hides his or her data in the masses. To this end, an analysis is performed on the data of all participating individuals. The analysis results are then compared to the results of the same analysis on a *neighboring database*, i.e., base data without the data of the individual in question. If the two results differ only by a small ϵ, the data of that individual are not exposed. However, it has to be mentioned that this result cannot be directly applied to all other individuals or to all types of analyses because the level of exposure may differ in each case. While there is a lot of research work on how to ensure these ϵ-differentially private properties [63], which show promising results [64], it still has rather little significance in practice due to the high complexity to apply such measures [65].

Synopsis. When looking at the related work, only the horizontal and vertical withholding of data can be applied to arbitrary raw data. However, in the context of data refinement, these measures are often too restrictive in terms of reducing not only the amount of sensitive information contained in the data but also the quality of data and thus their value. The dedicated filtering techniques can operate at a much higher level of granularity and thus better preserve data quality. To this end, however, the data must be available in the appropriate form and are afterwards only usable for their intended purpose. These measures could be applied directly at the data source's side. However, studies show that the effectiveness of such dedicated filtering techniques is higher when they are applied at the data processor's side, where a complete overview of all existing data are available [66]. Since data processors in the service domain are commonly assumed to be *semi-honest-but-curious*, i.e., they will not expose any sensitive information about a user deliberately, it is reasonable to entrust them with the task of data protection [67].

Therefore, to map the discussed measures to the data refinement process performed by the data processor, generic techniques such as horizontal and vertical withholding are particularly suitable in the data preparation phase, since only unprocessed raw data are handled here, for which no assumptions can be made regarding their structure. Only when the data have been transformed into processable information, i.e., in the data processing phase, can the dedicated privacy filters be applied systematically. Approaches such as differential privacy are also more suitable for hiding sensitive information from the data steward, i.e., in the data preparation phase, since data analysts may deal with data of a particular individual or a very small group of individuals only.

Since there is no privacy approach addressing the data refinement process as a whole, we introduce such an approach called SMARTEN in the following section.

4. SMARTEN

To enable a privacy-friendly data refinement by data processors, in SMARTEN, we consider the two-phase design of this process (see Section 2). In this regard, SMARTEN

provides a mechanism for data stewards to transform raw data into information without violating the privacy requirements of data producers. Our sample-based approach is presented in Section 4.1. For data analysts, SMARTEN provides a mechanism to gain knowledge from the information. To ensure that this knowledge does not expose sensitive information to data consumers, the addressed state-of-the-art filtering techniques are applied (see Section 3). This mechanism is discussed in detail in Section 4.2. Although a semi-honest-but-curious data processor can be assumed, a control mechanism is required so that independent supervisory authorities can monitor whether the data are processed privacy-aware. To this end, privacy requirements are attached inseparably to the data by means of digital signatures. This approach is explained in Section 4.3.

4.1. Data Preparation Phase

Due to the IoT and decreasing storage costs, data are constantly being captured and stored persistently. Data processors and especially data stewards are therefore faced with a flood of raw data that they cannot handle manually. Therefore, there are approaches towards data pre-processing powered by artificial intelligence. However, it cannot substitute human knowledge in data preparation. Human data wrangling experts, such as data stewards, achieve better results in data preparation because they can rely on broad domain knowledge and their experience. In the data preparation phase, it must be ensured to keep the human in the loop [68].

In our approach, we therefore adopt a dynamic sampling approach to meet this premise. Here, the data steward operates on a representative yet manageable sample which is extracted from the bulk of raw data. In doing so, the data steward defines data preparation rules that can be generalized and applied to the entire base data. While such an approach is primarily intended to relieve the experts, it also represents a privacy preserving measure [69]. Since the data steward never has access to the entire dataset, he or she can only gain as much insight as required for the fulfillment of his or her task. In addition, particularly sensitive data can be concealed entirely. However, by applying afterwards the defined rules to the entire base data, a comprehensive data preparation can still be ensured. Figure 3 illustrates the workflow that is followed in our approach.

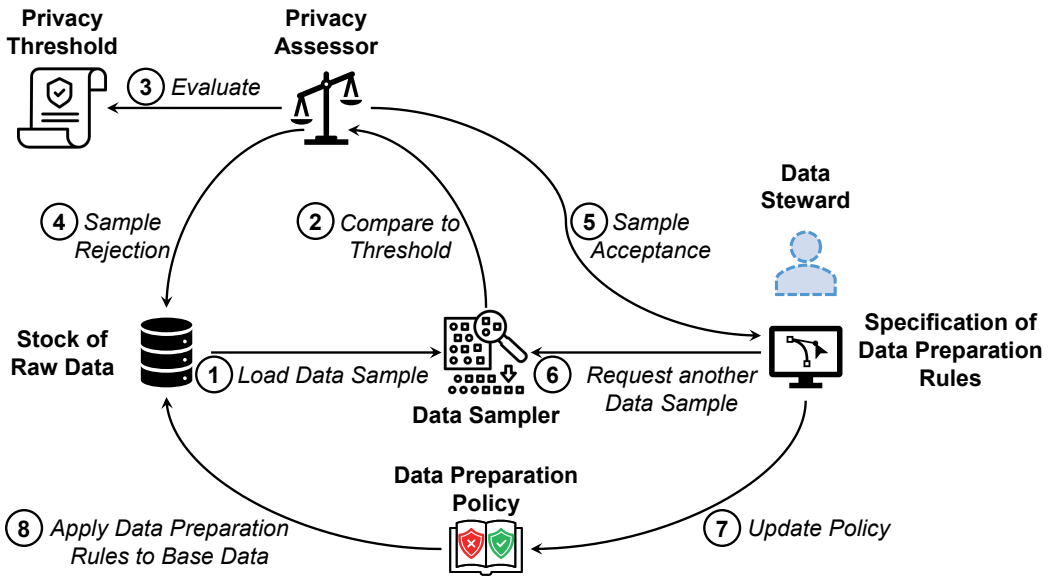

Figure 3. Workflow of the sample-based approach for data preparation applied in SMARTEN.

① Initially, the *data sampler* picks a representative data sample from the stock of raw data. Descriptive analyses are performed on the base data to identify data items that require special attention, e.g., null values, extreme values, and outliers. Metadata provided by the data producers can further facilitate the sample selection, for instance, to identify attributes

that are either particularly relevant for data preparation or have little or no influence on data refinement. ② If a reasonable sample is found, the data sampler forwards it to the *privacy assessor* ③. The privacy assessor evaluates whether the sample meets the necessary privacy requirements. These are specified by the data producer by means of a *privacy threshold*. For more information on setting these thresholds, see Section 4.3.

In addition to ϵ-differential privacy (see Section 3), other metrics can also be used for this purpose. For instance, *k-anonymity* describes how well an individual can disappear in a crowd. The k indicates that there are at least $k-1$ other individuals (respectively, data items) with similar attribute values, so that they cannot be distinguished unambiguously. Thus, information obtained from the data cannot be associated with a single individual, but only with a group of size k. Another metric is the *entropy* which evaluates the information content of a dataset. In simple terms, it assesses how much a data item deviates from the expected norm. For instance, if it is known that the value of an attribute $\mathbb{A}$ is almost in all cases x, then a data item where $\mathbb{A}$ has the value y has a particularly high information content. For more information on these and other technical privacy metrics that can be used to evaluate the samples, please refer to literature, e.g., the work by Wagner and Eckhoff [70]. With regard to the quality of the data refinement, however, the principle of fairness must also be considered in the sampling. For instance, no bias must be introduced due to the sampling [71]. For instance, if the value of the attribute $\mathbb{A}$ is predominantly x in the sample, while the attribute values x and y are equally distributed in the base data, the data steward might estimate the significance $\mathbb{A}$ incorrectly. This would have a negative impact on the quality of the outcome of the data preparation phase. There are also a wide range of technical fairness metrics that can be used in the privacy assessor of SMARTEN. For details on such metrics, please refer to literature, e.g., the work by Lässig et al. [72].

④ If the privacy assessor rejects the sample because it violates the privacy threshold, the data sampler must prepare a new sample and the assessment cycle restarts. If no sample can be found that satisfies the privacy threshold, e.g., since the threshold is too restrictive or the base data contain too sensitive information, automated data pre-processing can be applied. Here, an artificial intelligence approach performs both the data cleansing [73] and the data transformation [74]. However, the quality of the data preparation is expected to be impaired, as the data steward's domain knowledge cannot be taken into account [75]. In addition, such fully automated data processing is not permissible in certain application areas, e.g., when dealing with medical data [76]. Alternatively, synthetic data with characteristics similar to those of the real base data can be generated. Such an approach is used for data refinement when only very little base data are available [77]. The data steward can define preparation rules for these synthetic data, without having access to the real base data. However, this also impairs quality of data preparation, as *overfitting*—i.e., the generated data imply that certain issues are more significant than they are—or *underfitting*—i.e., certain data issues are not reflected in the generated data at all—may occur in the synthetic data.

⑤ If a valid sample is found, it is forwarded to the data steward. Only at this stage does a human gain insight into the base data excerpt. The data steward identifies data issues in the sample and specifies which data cleansing steps and transformation tasks are necessary as part of the data preparation. ⑥ If he or she needs further insights into the base data, he or she can request another sample from the data sampler until he or she is satisfied with the data quality. He or she can also retrieve statistical information about the base data, such as minima, maxima, or average values, e.g., to fill missing attribute values. ⑦ These data cleansing steps and transformation tasks for the sample can then be mapped to three general transformation operators, namely *filter operators*, *map operators*, and *reduce operators*. A filter operator filters out data items from the base data for which certain properties hold, a map operator transforms data items or modifies certain attributes, while a reduce operator groups and aggregates data items. ⑧ These generalized transformation operators can then be applied to the entire base data [78].

Our approach therefore combines a sample-based, an expert-based, and a rule-based approach. This way, the best possible data preparation quality can be achieved [79].

In technical terms, the three general privacy measures presented in Section 3, selection, projection, and aggregation, are used to ensure privacy. In sampling, selection is primarily used to reduce the number of data items. Projection can enhance privacy, e.g., by removing identifying attributes. When specifying the preparation rules, the data steward can use aggregations to deal with data issues such as missing attribute values.

4.2. Data Processing Phase

To meet the privacy requirements in the data processing phase, we designed a new data management architecture—or, more precisely, we added privacy features to an existing data management concept suitable for the data refinement process as introduced in Section 2. To this end, we leverage the *zone architecture for data lakes* introduced by Sharma [80]. Data lakes are an architecture for managing big data. For this purpose, the data are stored not only as raw format, but also in various processing stages. In a zone architecture, data in the same processing stage are organized in a common zone. Apart from zones in which data are stored persistently, there are also zones in which data are only kept in a volatile form, e.g., in order to transform them and forward them to downstream zones. Data consumers are granted access to one or more data zones via dedicated interfaces according to predefined access policies. In SMARTEN, we extend this generic core concept by additional privacy features. The resulting SMARTEN architecture is shown in Figure 4.

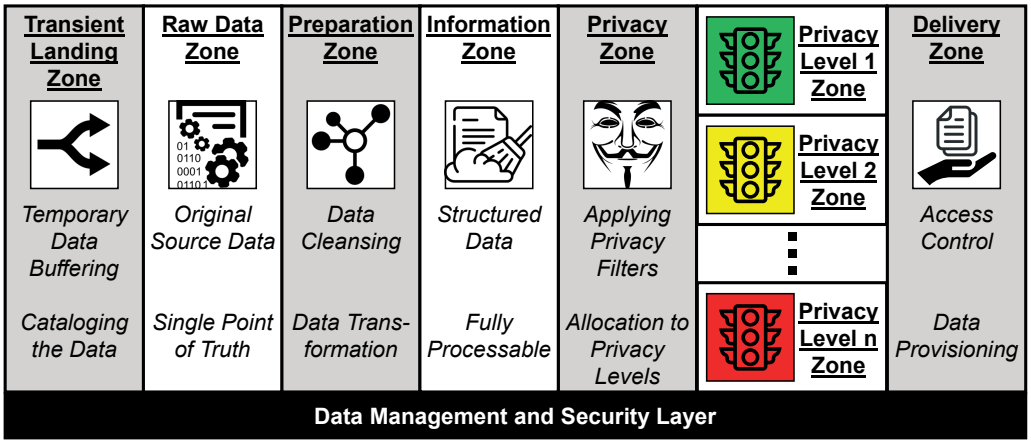

Figure 4. Extended date lake zone architecture to ensure privacy in the data processing phase.

This zone architecture reflects the data refinement process. The zones depicted in gray are processing zones in which data are only temporarily held, whereas the zones depicted in white are storage zones in which data are kept persistently. There are storage zones for raw data, information, and knowledge.

Raw data enter the data processor's scope via the *Transient Landing Zone*. This zone is a temporary buffer to be able to enrich the data with metadata relevant for processing, as well as to build access structures, such as data catalogs that facilitate the handling of the data. Then, these enriched data are stored in the *Raw Data Zone*. The SMARTEN workflow presented in Figure 3 operates on this zone in the data preparation phase. The specified data preparation rules are applied in the *Preparation Zone*. For this purpose, a processing engine reads the raw data affected by the rules into a processable data structure and applies the required transformations, namely a sequence of filter operators, map operators, and reduce operators. This zone represents a key extension to the original zone architecture.

The outcomes of the data preparation phase are stored in the *Information Zone*. The data items present in this zone are therefore in a fully processable structure. Thus, these data items can be made more privacy-friendly in a target-oriented manner by the *Privacy Zone*. This zone represents another second key extension, since it initiates the data processing phase in SMARTEN. In this zone, a repository is maintained in which privacy scripts are stored that implement different dedicated privacy techniques for specific types of data.

Examples of such techniques are given in Section 3. The appropriate scripts are selected by the Privacy Zone and applied to the prepared data [81]. Details on how to determine which techniques have to be applied are given in Section 4.3.

The concealed data are then transformed into knowledge by the data analyst. However, since different data consumers may have different privacy requirements, the data processor generates n different versions of the knowledge rather than just one. For each of these variants, a different combination of privacy scripts has been applied prior to processing. Thus, they represent different levels of privacy. For each level, there is a zone labeled *Privacy Level x Zone*. As a result, there is an appropriate variant of the knowledge for any purpose [82]. To determine the appropriate number of such privacy levels, privacy experts first analyze the available data sources and assess the potential threats posed to data subjects, i.e., the severity of the impact that a disclosure would entail. Mindermann et al. [83] demonstrate how a method called *STPA-Priv* [84], which is based on *System Theoretic Process Analysis*, can be used for this purpose. Here, experts are systematically guided by tools to uncover existing data privacy risks. The experts can document their findings regarding the privacy risks in a processable model that describes which knowledge can be derived from which data sources [85]. The derivable knowledge is thereby represented as knowledge patterns. Additionally, explanatory keywords are assigned to each of these patterns. By means of these keywords, data subjects can easily identify exactly those patterns that are relevant for them, i.e., patterns that have to be concealed in their data. In addition, a *collaborative filtering* approach can be used to recommend further patterns to data subjects that might be relevant to them as well [86]. These patterns are then translated into non-disclosure requirements, which are used for the configuration of SMARTEN (see Section 4.3).

The *Delivery Zone* regulates which data consumer has access to which privacy level. For each access, this zone checks which knowledge may be disclosed and shares only that variant with the data consumer. Third parties are not aware of any other variants. An access control policy defines which data consumer is allowed to access which privacy level. Please refer to literature for details on how such an access control policy can be designed, e.g., the work by Stach and Mitschang [87].

Orthogonal to these processing and storage zones there is a *Data Management and Security Layer*. Such a layer is needed in every data lake for data governance tasks, e.g., to assist data retrieval, to support efficient data access, or to implement data security measures. In SMARTEN, this layer has another relevant function. Permanent storage of the raw data and all their processing stages is necessary so that the data processor can fulfill its obligation to prove to the data producer that, e.g., the requested privacy techniques have been applied prior to processing. For this purpose, *Provable Data Possession* can be applied. Here, a data producer can verify whether the data are at rest in the respective privacy level. In our approach, a third-party auditor, e.g., a data protection authority, can perform this task on behalf of the data producer. In doing so, the third-party auditor requires neither insight into the data of the data processor nor into the data of the data producer [88].

To this end, the Data Management and Security Layer encodes and encrypts all data at rest in SMARTEN. The third-party auditor challenges the data processor regularly to ensure that the data are present in all required privacy levels. For this purpose, there is a puzzle mechanism that ensures that solving a puzzle always takes a certain amount of time. This puzzle is related to the encrypted data at rest. If the data processor takes too long to provide an answer to a challenge, it implies that he or she needs to generate a solution to the puzzle. Therefore, it can be assumed that the data processor has to generate a version of the data in the intended privacy level on the fly during the challenge. However, if the data processor responds quickly to the challenge, it means that he or she already knows the solution to the puzzle, i.e., the data are at rest in the requested privacy level [89].

4.3. Elicitation of Non-Disclosure Requirements

For the configuration of SMARTEN, i.e., the specification of privacy thresholds and privacy techniques that have to be applied, we adopt a two-stage approach in order to

address the two separate phases of the data refinement process. For this purpose, we rely on digital signatures that can be inseparably attached to the data. These signatures specify the non-disclosure requirements of the data producers. Such an approach ensures that an attacker cannot alter the requirements retrospectively in order to gain deeper insights into the data of the data producer [90].

Digital signatures are used to prove the authenticity of a data object, e.g., a message, a file, or a data item. For this purpose, asymmetric cryptography is used, i.e., a key pair is generated that consists of a *signing key* and a *verification key*. The signing key is kept secret by the signer, while the verification key is shared with the public. To sign a data object, the signer encrypts it with his or her signing key. The data object can only be verified with the corresponding verification key. If the verification is successful, it is ensured on the one hand that the data object has not been tampered with and on the other hand that it originates from the alleged source, since only this source has access to that signing key. An *attribute-based signature* uses keys that are composed of a set of attributes. A signed object can only be successfully verified if the key used for this purpose has a certain subset of these attributes. In SMARTEN, we also use attribute-based signatures, except that the non-disclosure requirements of the data producers are used instead of attributes. However, attribute-based encryption algorithms require a lot of computational power. Therefore, we adopt an approach introduced by Gritti et al. [91], in which the computations are distributed and can be partially outsourced. Our adapted approach is shown in Figure 5.

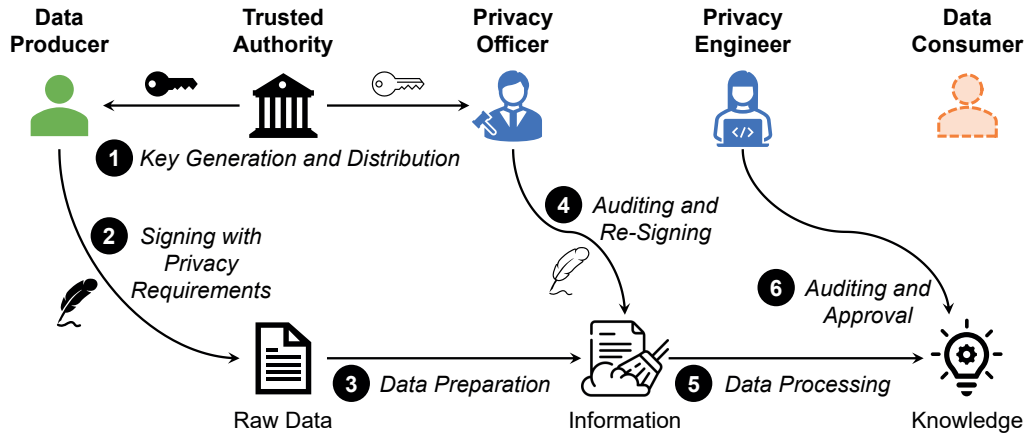

Figure 5. Two-stage approach to specify non-disclosure requirements applied in SMARTEN.

In addition to the actual two stages that are used to specify the privacy thresholds in the data preparation phase and the privacy requirements in the data processing phase, our approach also includes a preliminary phase for organizational purposes.

Key Generation and Deployment. Initially, a data producer announces his or her non-disclosure requirements to a trusted authority. When dealing with sensitive data, regulations such as the GDPR mandate that there has to be an independent supervisory authority (Art. 51 and Art. 52). This can, for instance, serve as the trusted authority. ❶ This authority then generates two key pairs for the signature. Let τ be the set of privacy thresholds and λ be the set of requested privacy levels for the knowledge gained from processing the data. Then, *full keys* (depicted in black) reflect the union of these two sets $\varphi = \tau \cup \lambda$, while *delegated keys* (depicted in white) reflect only λ, i.e., a true subset of φ. The full keys are provided to the data producer, while the delegated keys are provided to the data processor.

Full Authentication. ❷ To ensure that the data are not tampered with and that the non-disclosure requirements are not lost during transmission, the data producer signs the raw data with his or her full key. ❸ In the data preparation phase, the data processor verifies the signature against his or her privacy policy ρ_1. This policy describes which privacy thresholds are applied by the privacy assessor. Only if the requirements in τ are satisfied by ρ_1 is the applied privacy policy valid and the raw data in question can be preprocessed.

❹ This is monitored on behalf of the data processors by their privacy officer. A privacy officer requires no technical knowledge, as the privacy thresholds sufficiently specify how the data may be processed. If the data preparation is executed in compliance with the non-disclosure requirements, the privacy officer re-signs the data with the delegated key, i.e., τ is removed from the signature. In the data processing phase, the thresholds contained in τ are no longer relevant. However, non-disclosure requirements might indicate what a data producer wants to conceal. Thus, this filtering is necessary due to data minimization [92].

Delegated Authentication. ❺ The re-signing initiates the data processing phase. In this phase, the data processor verifies the modified signature against his or her privacy policy ρ_2. This policy describes the privacy measures for which privacy scripts are available in the Privacy Zone. Only if the scripts comply with the requirements described in the signature can the prepared data be further processed. ❻ However, as such an auditing is by no means trivial, it has to be handled by a privacy engineer. A privacy engineer represents an intermediary between legal experts and IT experts. He or she is able to evaluate the means by which the non-disclosure requirements of the data producers can be met without rendering the quality of the processing results useless. Only if the privacy engineer approves the applied measures is the gained knowledge offered in the respective Privacy Level x Zone to data consumers. Since a semi-honest-but-curious data processor can be assumed, this approach is a reliable way to enforce the non-disclosure requirements of the data producers.

As always with cryptographic approaches of this kind, security stands and falls with the authority that issues the keys. If this authority is malicious or becomes compromised, no assurances can be made regarding the security of the signatures. In this case, all keys issued by this authority would have to be invalidated and replaced. However, assuming that the data processor is at least semi-honest-but-curious (see Section 3) and therefore complies with applicable law such as the GDPR, the associated authority can also be trusted. Since this supervisory authority has to be independent according to GDPR, no involved party has any influence on it. Thus, even if the data processor is assumed to be somehow compromised, the key authority remains trustworthy. Furthermore, government agencies, such as the German Federal Office for Information Security (see https://www.bsi.bund.de/EN/Home/home_node.html, accessed on 31 July 2022), can also provide such key generation services, i.e., operate as such this trusted authority. However, if no authority can be found that all parties involved can fully trust, it is also possible to have several authorities responsible for key generation and key escrow instead of a single central authority. In this case, if some of these authorities become compromised, the security of the keys would not be affected, as long as not all of them are exposed at the same time. Such a distributed approach could be achieved for instance by means of *multi-party computation* [93] or *secret sharing* [94].

Protective measures must also be taken with regard to the keys themselves. Only if the methods used to generate them are secure are the digital signatures generated with them trustworthy. For the purpose in SMARTEN, we recommend the use of an *elliptic curve cryptography*, with which we aim for 224-bit/256-bit keys, giving us a minimum strength of 112 bits/128 bits of security [95]. Technological progress, such as quantum computing, may in the future render this encryption method, which is considered sufficiently secure today, ineffective [96]. In this case, *post-quantum cryptography* [97] would have to be used for digital signatures in SMARTEN. However, this would not affect the underlying concepts and procedures presented in this paper, since they are inherently technology-independent.

Next, we evaluate SMARTEN in terms of functionality, usability, and performance.

5. Evaluation

In this section, we evaluate whether SMARTEN achieves our goal of making the data refinement process privacy-friendly. To this end, we perform a feature assessment in Section 5.1 to determine which privacy properties SMARTEN provides. Then, in Section 5.2, we report on the results of a user study that provide evidence on the usability of SMARTEN

from the perspective of data processors. Finally, in Section 5.3, we present the results of a performance measurement to quantify the overhead caused by SMARTEN in terms of data throughput and memory consumption during data processing. Since the sample-based approach we introduce for the data preparation phase constitutes the key component of our work, we also focus on this aspect in the user study and the performance measurement. At the end, we summarize the lessons learned in Section 5.4.

5.1. Feature Assessment

As outlined in Section 2, data protection regulations such as the GDPR impose legal requirements on the data refinement process. In this feature assessment, we therefore investigate whether SMARTEN complies with these requirements.

SMARTEN provides technical measures that ensure compliance with the *principles relating to processing of personal data* (Art. 5). Data minimization is considered in many ways. First, data stewards receive only insight into a small section of the data due to our sample-based approach. Second, by applying the privacy scripts in the Privacy Zone, the information content is minimized prior to processing. Third, even with respect to the non-disclosure requirements, the data processor receives only as much information as necessary, due to our two-stage signature approach. This fine-grained and target-oriented data minimization also ensures that the data processor is able to maximize the accuracy of the refined data. Furthermore, the provision of different variants of the refined data at different privacy levels enables to enforce a purpose limitation towards the data consumer.

Since the non-disclosure requirements of the data producers are inseparably attached to the data by means of digital signatures, the *lawfulness of processing* (Art. 6) as well as *conditions for consent* (Art. 7)—including all their amendments in the subsequent articles (Art. 8 and Art. 9)—can be guaranteed by the data processor. Given the fact that a semi-honest-but-curious data processor can be assumed, it is sufficient to have two supervisory bodies in place, the privacy officer and the privacy engineer. Due to the measures introduced by SMARTEN, they are able to assess and ensure compliance with the requirements.

SMARTEN adheres to the *transparency and information rights* (Art. 12–Art. 14), as third-party auditors can monitor at any time on behalf of the data producer which data are present in which processing stage at the data processor. Furthermore, the non-disclosure requirements clearly stipulate in a transparent manner how the data are processed.

The *right to erasure* (Art. 17) is directly supported by SMARTEN, while the *right to rectification* (Art. 16) can also be achieved indirectly in the same way. While the trusted key authority cannot delete the data, it can invalidate the keys of the data processor. Since all data are available only in encrypted form due to the signatures, invalidating the keys renders the data unusable. Rectification can be realized as a sequence of an erasure and re-submission of the corrected data.

The *right to restriction of processing* (Art. 18) is also enforced by means of the signatures. Only if the processing policy meets the required attributes, i.e., appropriate thresholds or adequate privacy measures, the data can be decrypted and processed.

Since humans are involved in each phase of the data refinement process, i.e., data stewards as well as privacy officers in the preparation phase and data analysts as well as privacy engineers in the processing phase, a *fully automated data processing* (Art. 22) is ruled out by default.

By including third-party auditors as well as trusted authorities, *independent supervisory authorities* (Art. 51 and Art. 52) are strongly involved in SMARTEN.

Although only the most relevant aspects regarding applicable data protection regulations could be discussed here, overall it can be concluded that SMARTEN represents a comprehensive *privacy-by-design solution* (Art. 25).

5.2. User Study

Behringer et al. [98] conduct a user study to evaluate the quality of data preparation when using a dynamic sampling approach similar to the one applied in SMARTEN. To

this end, they compare the dynamic sampling approach (DS), in which the data steward defines data preparation rules on a representative sample, which are then applied to the entire base data and he or she receives feedback on the effects of the rules, with a static sampling approach (SS) without this interaction and an approach in which no sampling takes place (NS) and data preparation is performed on the entire base data. Twelve participants (nine male and three female) use the three approaches to cleanse a given dataset. The participants are all between 20 and 38 years old (mean = 25.5, standard deviation = 6.25) and have a computer science background.

In the user study, both objective measures and subjective measures are taken into account. On the one hand, it is assessed how many data issues the participants are able to resolve in the base data using the respective approach. On the other hand, a *5-point Likert scale* is used to measure how satisfied the participants are with the respective approach. In Figure 6, the main findings of the user study are presented.

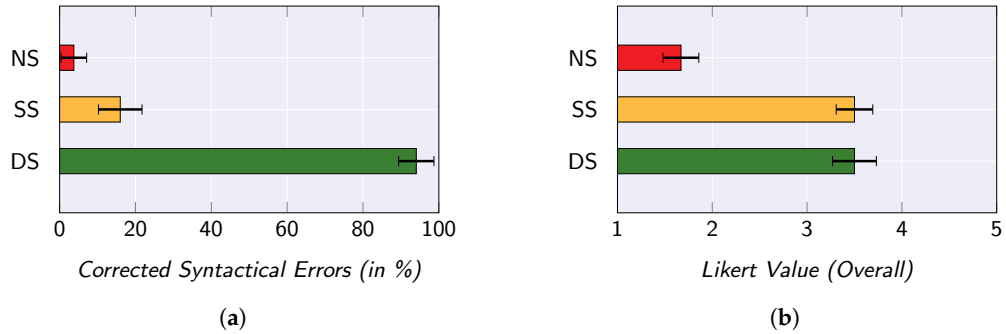

Figure 6. Key findings of the user study conducted by Behringer et al. [98]. (**a**) Mean percentage of cleaned data issues; (**b**) overall Satisfaction of the participants.

When looking at the percentage of data issues that the participants were able to resolve with the respective approaches (see Figure 6a), it is evident that the dynamic sampling approach outperforms the other two approaches. Furthermore, even the application of a static sampling approach is a significant improvement over an approach without sampling. By applying an *one-way ANOVA with repeated measures*, the statistical significance of this finding can be confirmed. The statistical significance of the results is also confirmed by the *Bonferroni-adjusted post-hoc analysis*.

When considering the subjective ratings of the participants (see Figure 6b), it becomes apparent that the two sampling approaches are distinctly more popular overall than the approach without sampling. From a user perspective, however, there is no perceived difference in popularity between the two approaches. The statistical significance of these two observations can also be verified by means of a *Friedman test* and a *Wilcoxon test*.

Since a similar sample-based approach is applied in SMARTEN, these findings are also highly relevant for us. As the user study indicates that users prefer to have any kind of sampling approach instead of having to operate on the entire base data, it can be assumed that data stewards appreciate working with SMARTEN. Furthermore, any kind of sampling also represents an improvement in terms of data preparation quality. Even though SMARTEN does not provide feedback to the data stewards on the impact of the data preparation rules—such feedback would enable to draw conclusions about the base data and is therefore not envisaged in SMARTEN for privacy reasons—the systematic selection of data items still is an improvement over static sampling. SMARTEN is therefore a genuine improvement for data stewards in terms of usability. As part of future work, we intend to carry out another user study in which privacy metrics are also considered for sampling, in order to validate this initial assumption regarding the efficiency of SMARTEN.

5.3. Performance Measurement

To determine the performance overhead caused by SMARTEN, we have implemented three different data preparation rules. Each of these rules reflects one of the three transformation operators, namely a filter operator, a map operator, and a reduce operator: The filter operator eliminates all data items with a certain attribute value (i.e., a selection). The map operator removes a certain attribute (i.e., a projection). The reduce operator groups the data items by a categorical attribute and calculates the mean value of an attribute for each group (i.e., an aggregation). These rules are then applied to base data.

A relational database (*SQLite DB 3.37.2*, see https://www.sqlite.org/, accessed on 31 July 2022) is used to store the base data. We implemented the processing engine which is required in the Preparation Zone using *Python 3.10.3* (see https://www.python.org/, accessed on 31 July 2022) and *Pandas 1.4.1* (see https://pandas.pydata.org/, accessed on 31 July 2022). All base data are loaded into a *DataFrame*, where the transformations are applied. The database as well as the processing engine are deployed on a computer with an *Intel Core i7-1165G7* and 32 GB DDR4-3200 main memory.

For the evaluation, we used synthetic base data. On that account, we generated stepwise from 500 to 5000 k data items. In each step, we have increased the amount of data tenfold. This represents a maximum data volume of about 12 GB. In our evaluation, we measured both the data throughput, i.e., how much data per second is processed, and the main memory peak usage. For each operator and each base data size, we performed the measurements ten times. After each run, we reset the database to exclude distortions due to warm caches. In Figure 7, the median of the runs are shown—we use the medians to exclude side effects caused by background processes.

For comparison purposes, we have also implemented the three operators as SQL commands which can be executed directly by the SQLite DB. We use this as our baseline. This represents an optimum. In this baseline, the initial retrieval of the data is eliminated and the available index structures of the database can be used to support the processing. Therefore, the closer the processing engine used in SMARTEN approximates this baseline, the better the performance has to be assessed.

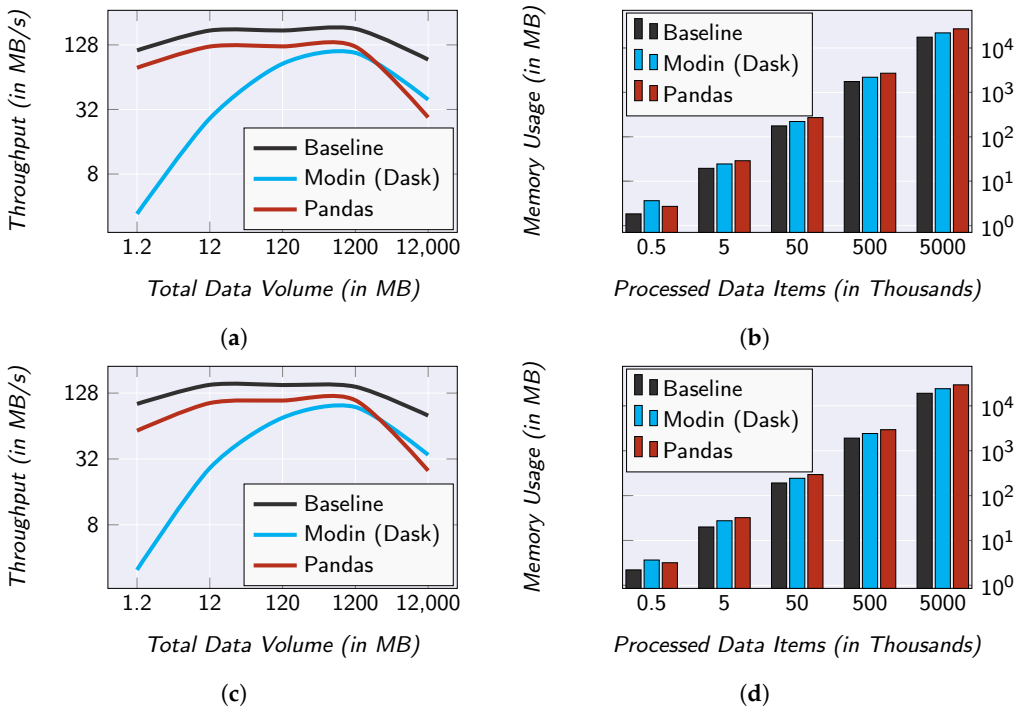

Figure 7. *Cont.*

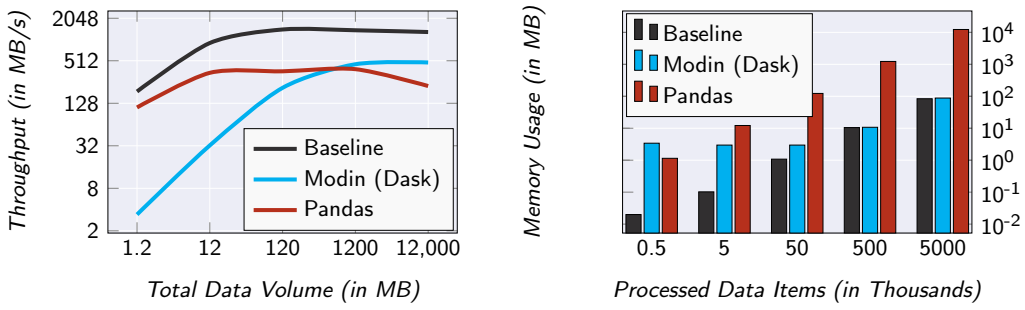

Figure 7. Evaluation results regarding the overhead caused by SMARTEN. (**a**) Data throughput of the filter operator; (**b**) memory usage of the filter operator; (**c**) data throughput of the map operator; (**d**) memory usage of the map operator; (**e**) data throughput of the reduce operator; (**f**) memory usage of the reduce operator.

It can be noticed that the selection data throughput and the projection data throughput hardly show any differences (see Figure 7a,c). Even though the throughput of SMARTEN is lower than the one of the baseline, both show an identical progression. That is, the throughput of SMARTEN is in $\mathcal{O}(baseline)$. For the aggregation, the two curves are also somewhat similar (see Figure 7e). Here, however, the difference between the two curves is significantly bigger than in the case of the selection and the projection. This is due to the fact that, especially in the case of grouping, the database can make use of index structures concerning the categorical attribute.

It is noticeable that the performance of SMARTEN drops for the biggest set of base data for in all three operators. This is due to the Pandas-based implementation of the processing engine. Pandas loads a DataFrame completely into main memory. Since in our evaluation the database and the processing engine share the available main memory, there is almost no free main memory left for this base data volume. To address this issue, we have also implemented the processing engine using *Modin 0.13.3* (see https://modin.readthedocs.io/, accessed on 31 July 2022) and *Dask 2022.3.0* (see https://docs.dask.org/, accessed on 31 July 2022) as its compute engine.

Dask is a Python library that supports data partitioning and parallel computations. For this purpose, it provides a data structure called *Dask DataFrame*, which is internally composed of several Pandas DataFrames. The data contained in the Dask DataFrame are partitioned row-wise, and each partition is stored in a Pandas DataFrame. The Dask DataFrame has an identical API as the Pandas DataFrame, yet all commands are executed in parallel. The abstraction framework Modin takes care of the data distribution and assigns the tasks to the available processor cores. Due to these properties, switching the SMARTEN processing engine from Pandas to Modin is pretty straightforward.

The results of the data throughput measurements are surprising at first glance. Apparently, the Modin-based implementation has the worst performance for all three operators. However, it has to be considered that Modin causes an organizational overhead due to the initial partitioning and distribution as well as the merging of the results. This overhead is amortized for large amounts of data, as can be seen for the 12 GB of base data. Here, the Modin-based implementation approximates the baseline even further. That is, with a combination of a Pandas-based implementation for small datasets and a Modin-based implementation for large datasets, SMARTEN has a throughput behavior which is asymptotically similar to the one of the baseline.

With regard to the memory footprint, there are no noticeable differences for the selection and projection (see Figure 7b,d). However, for aggregation (see Figure 7f), the memory consumption is significantly higher in the Pandas-based implementation. This is due to the fact that Pandas loads all data into memory, while the other two implementations process the data group-based.

5.4. Lessons Learned

The main innovation of SMARTEN is that, in contrast to state-of-the-art approaches, the characteristics of the phases of the data refinement process are fully addressed. By supporting different privacy metrics for data sampling in the data preparation phase, and different privacy filters for data provisioning in the data processing phase, any use case is supported, even those dealing with highly sensitive data, such as location-based services [99] or eHealth services [100]. Moreover, SMARTEN can also be extended by further privacy metrics and privacy filters to meet future challenges. This way, the privacy requirements of data producers can be fully satisfied and at the same time the quality requirements of data consumers can be satisfied. To the best of our knowledge, SMARTEN is the only approach that covers the entire data refinement process and thus enables privacy-aware and demand-oriented data provisioning.

The results of our evaluation show that SMARTEN meets the requirements towards a privacy-by-design solution for the data refinement process. Furthermore, our sample-based approach has a positive impact on the work of the data stewards and the involved overhead in terms of data throughput and memory consumption is reasonable.

6. Conclusions

Due to technological progress, it is nowadays possible to capture data on literally any aspect of life almost continuously. Smart services take advantage of this fact by preparing the acquired data in order to extract processable information from them. This information can then be processed and analyzed to derive knowledge. This knowledge is the key driver of smart services. Users benefit significantly from smart services as the comprehensive knowledge base enables them to tailor their portfolio of services to the user's needs and thus offer him or her an optimal service experience. Despite the undeniable advantages of such smart services, which have become an indispensable part of everyday life, this development also has its negative aspects. When data are prepared and processed, the parties involved also gain deep insights into confidential and privacy-sensitive information. It is therefore mandatory to take technical measures which ensure that smart services are provided with the necessary data, but prevent the disclosure of sensitive data.

To this end, we discuss the refinement process that transforms raw data into operational knowledge. While there are various approaches to conceal sensitive information, all of these approaches mainly deal with the second phase of the data refinement process, i.e., the data processing phase. For the data preparation phase, there are no dedicated privacy approaches—let alone a holistic consideration of the data refinement process from a privacy point of view. We therefore introduce SMARTEN, a privacy approach that smartens up services in a privacy-friendly manner. To achieve this goal, we make the following three contributions:

1. By adapting the data preparation phase of the data refinement process, data stewards only obtain access to a representative sample of the data. The samples are compiled in such a way that they comply with given privacy constraints. A data steward operates on the sample, and these data wrangling activities can be transferred to preprocessing rules, which are then applied to the entire data stock. This way, data can be transformed into processable information without providing too many insights.
2. By adapting the data processing phase of the data refinement process, different privacy filters can be applied to the acquired information before it is analyzed in order to derive knowledge. The privacy filters are designed for each type of data in such a way that specific aspects of the data can be concealed without impairing the quality of the data significantly and thus rendering smart services inoperable in the process.
3. Using two-tier annotations, data owners can specify their privacy requirements. The outer tier specifies which requirements must be met in the data preparation phase while the inner layer describes which privacy measures have to be applied in the data processing phase. In SMARTEN, these annotations are realized as digital signatures that can be attached directly to the source data and thus inseparably tied to them.

By means of these three contributions, SMARTEN renders the data refinement process privacy-friendly without unnecessarily restricting smart services. Evaluation results show that our sample-based data refinement approach is feasible for data stewards and achieves a satisfactory level of data quality. In addition, the data processing overhead is relatively low.

Author Contributions: Conceptualization, C.S.; methodology, M.B. and C.S.; software, M.B., C.G. and C.S.; validation, M.B., C.G. and C.S.; formal analysis, M.B. and C.S.; investigation, C.S.; resources, J.B., C.G. and B.M.; data curation, M.B., J.B., C.G. and C.S.; writing—original draft preparation, C.S.; writing—review and editing, M.B., J.B., C.G., B.M. and C.S.; visualization, C.S.; supervision, C.S.; project administration, C.S.; funding acquisition, B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this paper:

ANOVA	analysis of variance
API	application programming interface
CRISP-DM	cross-industry standard process for data mining
DB	database
DIK	data, information, knowledge
DS	dynamic sampling
eHealth	electronic health
GB	gigabyte
GDPR	general data protection regulation
GPS	global positioning system
IoT	internet of things
IT	information technology
KDD	knowledge discovery in databases
MB	megabyte
NS	no sampling
SEMMA	sample, explore, modify, model, and assess
SMARTEN	sample-based approach towards privacy-friendly data refinement
SQL	structured query language
SS	static sampling
STPA-Priv	system theoretic process analysis for Privacy

References

1. Schwab, K.; Marcus, A.; Oyola, J.R.; Hoffman, W.; Luzi, M. *Personal Data: The Emergence of a New Asset Class*; World Economic Forum: Geneva, Switzerland, 2011.
2. Toonders, J. Data is the New Oil of the Digital Economy. In *WIRED*; Condé Nast: New York, NY, USA, 2014.
3. Quigley, E.; Holme, I.; Doyle, D.M.; Ho, A.K.; Ambrose, E.; Kirkwood, K.; Doyle, G. "Data is the new oil": Citizen science and informed consent in an era of researchers handling of an economically valuable resource. *Life Sci. Soc. Policy* **2021**, *17*, 9. [CrossRef] [PubMed]
4. Jesse, N. Data Strategy and Data Trust–Drivers for Business Development. *IFAC Pap.* **2021**, *54*, 8–12. [CrossRef]
5. Bibri, S.E.; Krogstie, J. A Novel Model for Data-Driven Smart Sustainable Cities of the Future: A Strategic Roadmap to Transformational Change in the Era of Big Data. *Future Cities Environ.* **2021**, *7*, 3. [CrossRef]
6. Hallur, G.G.; Prabhu, S.; Aslekar, A. Entertainment in Era of AI, Big Data & IoT. In *Digital Entertainment: The Next Evolution in Service Sector*; Das, S., Gochhait, S., Eds.; Springer: Singapore, 2021; pp. 87–109.

7. Jossen, S. The World's Most Valuable Resource Is No Longer Oil, But Data. *Economist*, 6 May 2017.
8. Bello, O.; Zeadally, S. Toward efficient smartification of the Internet of Things (IoT) services. *Future Gener. Comput. Syst.* **2019**, *92*, 663–673. [CrossRef]
9. Bhageshpur, K. *Data is the New Oil—And That's a Good Thing*; Forbes Technololy Council: Boston, MA, USA, 2019.
10. Taffel, S. Data and oil: Metaphor, materiality and metabolic rifts. *New Media Soc. (OnlineFirst)* **2021**, 14614448211017887. [CrossRef]
11. Liew, A. Understanding Data, Information, Knowledge And Their Inter-Relationships. *J. Knowl. Manag. Pract.* **2007**, *8*, 1–10.
12. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [CrossRef]
13. Hashemi, S.H.; Faghri, F.; Rausch, P.; Campbell, R.H. World of Empowered IoT Users. In Proceedings of the 2016 IEEE First International Conference on Internet-of-Things Design and Implementation (IoTDI), Berlin, Germany, 4–8 April 2016; IEEE: Manhattan, NY, USA, 2016; pp. 13–24.
14. Van Meter, H.J. Revising the DIKW Pyramid and the Real Relationship between Data, Information, Knowledge and Wisdom. *Law Technol. Hum.* **2020**, *2*, 69–80. [CrossRef]
15. Alasadi, S.A.; Bhaya, W.S. Review of Data Preprocessing Techniques in Data Mining. *J. Eng. Appl. Sci.* **2017**, *12*, 4102–4107.
16. Elgendy, N.; Elragal, A. Big Data Analytics: A Literature Review Paper. In Proceedings of the 14th Industrial Conference on Data Mining (ICDM), St. Petersburg, Russia, 16–20 July 2014; Springer: Cham, Switzerland, 2014; pp. 214–227.
17. Maletic, J.I.; Marcus, A. Data Cleansing: A Prelude to Knowledge Discovery. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2010; pp. 19–32.
18. Tawalbeh, L.; Muheidat, F.; Tawalbeh, M.; Quwaider, M. IoT Privacy and Security: Challenges and Solutions. *Appl. Sci.* **2020**, *10*, 4102. [CrossRef]
19. Ali, M.U.; Mishra, B.K.; Thakker, D.; Mazumdar, S.; Simpson, S. Using Citizen Science to Complement IoT Data Collection: A Survey of Motivational and Engagement Factors in Technology-Centric Citizen Science Projects. *IoT* **2021**, *2*, 275–309. [CrossRef]
20. Lagoze, C.; Block, W.C.; Williams, J.; Abowd, J.; Vilhuber, L. Data Management of Confidential Data. *Int. J. Digit. Curation* **2013**, *8*, 265–278. [CrossRef]
21. Ukil, A.; Bandyopadhyay, S.; Pal, A. IoT-Privacy: To be private or not to be private. In Proceedings of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 27 April–2 May 2014; IEEE: Manhattan, NY, USA, 2014; pp. 123–124.
22. Gerber, N.; Gerber, P.; Volkamer, M. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Comput. Secur.* **2018**, *77*, 226–261. [CrossRef]
23. Sarker, I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Comput. Sci.* **2021**, *2*, 377. [CrossRef] [PubMed]
24. Matignon, R. *Data Mining Using SAS Enterprise Miner*; Wiley: Hoboken, NJ, USA, 2007.
25. Shearer, C. The CRISP-DM Model: The New Blueprint for Data Mining. *J. Data Warehous.* **2000**, *5*, 13–22.
26. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM* **1996**, *39*, 27–34. [CrossRef]
27. Kutzias, D.; Dukino, C.; Kett, H. Towards a Continuous Process Model for Data Science Projects. In Proceedings of the 12th International Conference on Applied Human Factors and Ergonomics (AHFE), New York, NY, USA, 25–29 July 2021; Springer: Cham, Switzerland, 2021; pp. 204–210.
28. Costagliola, G.; Fuccella, V.; Giordano, M.; Polese, G. Monitoring Online Tests through Data Visualization. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 773–784. [CrossRef]
29. Uttamchandani, S. *The Self-Service Data Roadmap: Democratize Data and Reduce Time to Insight*; O'Reilly: Sebastopol, CA, USA, 2020.
30. Azeroual, O. Data Wrangling in Database Systems: Purging of Dirty Data. *Data* **2020**, *50*, 50. [CrossRef]
31. Delen, D. *Prescriptive Analytics: The Final Frontier for Evidence-Based Management and Optimal Decision Making*; Pearson FT Press: Hoboken, NJ, USA, 2019.
32. Luengo, J.; García-Gil, D.; Ramírez-Gallego, S.; García, S.; Herrera, F. *Big Data Preprocessing: Enabling Smart Data*; Springer: Cham, Switzerland, 2020.
33. European Parliament and Council of the European Union. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive). Legislative Acts L119. *Off. J. Eur. Union* **2016**. Available online: https://gdpr-info.eu/ (accessed on 31 July 2022)
34. Rhahla, M.; Allegue, S.; Abdellatif, T. Guidelines for GDPR compliance in Big Data systems. *J. Inf. Secur. Appl.* **2021**, *61*, 102896. [CrossRef]
35. Rassouli, B.; Rosas, F.E.; Gündüz, D. Data Disclosure Under Perfect Sample Privacy. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 2012–2025. [CrossRef]
36. Al-Rubaie, M.; Chang, J.M. Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Secur. Priv.* **2019**, *17*, 49–58. [CrossRef]
37. Dou, H.; Chen, Y.; Yang, Y.; Long, Y. A secure and efficient privacy-preserving data aggregation algorithm. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 1495–1503. [CrossRef]
38. Khwaja, A.S.; Anpalagan, A.; Naeem, M.; Venkatesh, B. Smart Meter Data Obfuscation Using Correlated Noise. *IEEE Internet Things J.* **2020**, *7*, 7250–7264. [CrossRef]

39. Gangarde, R.; Sharma, A.; Pawar, A.; Joshi, R.; Gonge, S. Privacy Preservation in Online Social Networks Using Multiple-Graph-Properties-Based Clustering to Ensure k-Anonymity, l-Diversity, and t-Closeness. *Electronics* **2021**, *10*, 2877. [CrossRef]
40. Stach, C. How to Deal with Third Party Apps in a Privacy System—The PMP Gatekeeper. In Proceedings of the 2015 IEEE 16th International Conference on Mobile Data Management (MDM), Pittsburgh, PA, USA, 15–18 June 2015; IEEE: Manhattan, NY, USA, 2015; pp. 167–172.
41. Stach, C. How to Assure Privacy on Android Phones and Devices? In Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management (MDM), Milan, Italy, 3–6 June 2013; IEEE: Manhattan, NY, USA, 2013; pp. 350–352.
42. Stach, C.; Mitschang, B. Privacy Management for Mobile Platforms—A Review of Concepts and Approaches. In Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management (MDM), Milan, Italy, 3–6 June 2013; IEEE: Manhattan, NY, USA, 2013; pp. 305–313.
43. Hou, W.C.; Ozsoyoglu, G.; Taneja, B.K. Statistical Estimators for Relational Algebra Expressions. In Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), Austin, TX, USA, 21–23 March 1988; ACM: New York, NY, USA, 1988; pp. 276–287.
44. Stach, C. Fine-Grained Privacy Control for Fitness and Health Applications Using the Privacy Management Platform. In Proceedings of the Information Systems Security and Privacy: 4th International Conference, ICISSP 2018, Funchal, Madeira, Portugal, 22–24 January 2018; Revised Selected Papers; Mori, P., Furnell, S., Camp, O., Eds.; Springer: Cham, Switzerland, 2019; pp. 1–25.
45. McKenzie, L.E.; Snodgrass, R.T. Evaluation of Relational Algebras Incorporating the Time Dimension in Databases. *ACM Comput. Surv.* **1991**, *23*, 501–543. [CrossRef]
46. Özsoyoğlu, G.; Özsoyoğlu, Z.M.; Matos, V. Extending Relational Algebra and Relational Calculus with Set-Valued Attributes and Aggregate Functions. *ACM Trans. Database Syst.* **1987**, *12*, 566–592. [CrossRef]
47. Li, J.; Maier, D.; Tufte, K.; Papadimos, V.; Tucker, P.A. Semantics and Evaluation Techniques for Window Aggregates in Data Streams. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD), Baltimore, MD, USA, 14–16 June 2005; ACM: New York, NY, USA, 2005; pp. 311–322.
48. Olejnik, K.; Dacosta, I.; Machado, J.S.; Huguenin, K.; Khan, M.E.; Hubaux, J.P. SmarPer: Context-Aware and Automatic Runtime-Permissions for Mobile Devices. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; IEEE: Manhattan, NY, USA, 2017; pp. 1058–1076.
49. Navidan, H.; Moghtadaiee, V.; Nazaran, N.; Alishahi, M. Hide me Behind the Noise: Local Differential Privacy for Indoor Location Privacy. In Proceedings of the 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS & PW), Genoa, Italy, 6–10 June 2022; IEEE: Manhattan, NY, USA, 2022; pp. 514–523.
50. Choi, M.J.; Kim, H.S.; Moon, Y.S. Publishing Sensitive Time-Series Data under Preservation of Privacy and Distance Orders. *Int. J. Innov. Comput. Inf. Control.* **2012**, *8*, 3619–3638.
51. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 1–36. [CrossRef]
52. Alpers, S.; Oberweis, A.; Pieper, M.; Betz, S.; Fritsch, A.; Schiefer, G.; Wagner, M. PRIVACY-AVARE: An approach to manage and distribute privacy settings. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; IEEE: Manhattan, NY, USA, 2017; pp. 1460–1468.
53. Kido, H.; Yanagisawa, Y.; Satoh, T. An anonymous communication technique using dummies for location-based services. In Proceedings of the 2005 International Conference on Pervasive Services (ICPS), Santorini, Greece, 11–14 July 2005; IEEE: Manhattan, NY, USA, 2005; pp. 88–97.
54. Stach, C.; Steimle, F.; Mitschang, B. How to Realize Device Interoperability and Information Security in mHealth Applications. In Proceedings of the Biomedical Engineering Systems and Technologies: 11th International Joint Conference, BIOSTEC 2018, Funchal, Portugal, 19–21 January 2018; Revised Selected Papers; Cliquet, A., Jr., Wiebe, S., Anderson, P., Saggio, G., Zwiggelaar, R., Gamboa, H., Fred, A., Bermúdez i Badia, S., Eds.; Springer: Cham, Switzerland, 2019; pp. 213–237.
55. Stach, C. VAULT: A Privacy Approach towards High-Utility Time Series Data. In Proceedings of the Thirteenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE), Nice, France, 27–31 October 2019; IARIA: Wilmington, DE, USA, 2019; pp. 41–46.
56. Hernández Acosta, L.; Reinhardt, D. A survey on privacy issues and solutions for Voice-controlled Digital Assistants. *Pervasive Mob. Comput.* **2022**, *80*, 101523. [CrossRef]
57. Oh, S.J.; Benenson, R.; Fritz, M.; Schiele, B. Faceless Person Recognition: Privacy Implications in Social Media. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 19–35.
58. Alpers, S.; Betz, S.; Fritsch, A.; Oberweis, A.; Schiefer, G.; Wagner, M. Citizen Empowerment by a Technical Approach for Privacy Enforcement. In Proceedings of the 8th International Conference on Cloud Computing and Services Science (CLOSER), Funchal, Portugal, 19–21 March 2018; SciTePress: Setúbal, Portugal, 2018; pp. 589–595.
59. Stach, C.; Dürr, F.; Mindermann, K.; Palanisamy, S.M.; Wagner, S. How a Pattern-based Privacy System Contributes to Improve Context Recognition. In Proceedings of the 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Athens, Greece, 19–23 March 2018; IEEE: Manhattan, NY, USA, 2018; pp. 238–243.

60. Kwecka, Z.; Buchanan, W.; Schafer, B.; Rauhofer, J. "I am Spartacus": Privacy enhancing technologies, collaborative obfuscation and privacy as a public good. *Artif. Intell. Law* **2014**, *22*, 113–139. [CrossRef]
61. Slijepčeviá, D.; Henzl, M.; Klausner, L.D.; Dam, T.; Kieseberg, P.; Zeppelzauer, M. k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Comput. Secur.* **2021**, *111*, 102488. [CrossRef]
62. Dwork, C. Differential Privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming (ICALP), Venice, Italy, 10–14 July 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12.
63. Jain, P.; Gyanchandani, M.; Khare, N. Differential privacy: Its technological prescriptive using big data. *J. Big Data* **2018**, *5*, 15. [CrossRef]
64. Zhu, T.; Li, G.; Zhou, W.; Yu, P.S. Differentially Private Recommender System. In *Differential Privacy and Applications*; Springer: Cham, Switzerland, 2017; pp. 107–129.
65. Machanavajjhala, A.; He, X.; Hay, M. Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges. In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD), Chicago, IL, USA, 14–19 May 2017; ACM: New York, NY, USA, 2017; pp. 1727–1730.
66. Stach, C.; Alpers, S.; Betz, S.; Dürr, F.; Fritsch, A.; Mindermann, K.; Palanisamy, S.M.; Schiefer, G.; Wagner, M.; Mitschang, B.; et al. The AVARE PATRON—A Holistic Privacy Approach for the Internet of Things. In Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (SECRYPT), Porto, Portugal, 26–28 July 2018; SciTePress: Setúbal, Portugal, 2018; pp. 372–379.
67. Chai, Q.; Gong, G. Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. In Proceedings of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, 10–15 June 2012; IEEE: Manhattan, NY, USA, 2012; pp. 917–922.
68. Zagalsky, A.; Te'eni, D.; Yahav, I.; Schwartz, D.G.; Silverman, G.; Cohen, D.; Mann, Y.; Lewinsky, D. The Design of Reciprocal Learning Between Human and Artificial Intelligence. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 443. [CrossRef]
69. Arcolezi, H.H.; Couchot, J.F.; Al Bouna, B.; Xiao, X. Random Sampling Plus Fake Data: Multidimensional Frequency Estimates With Local Differential Privacy. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM), Gold Coast, QLD, Australia, 1–5 November 2021; ACM: New York, NY, USA, 2021; pp. 47–57.
70. Wagner, I.; Eckhoff, D. Technical Privacy Metrics: A Systematic Survey. *ACM Comput. Surv.* **2018**, *51*, 57. [CrossRef]
71. Oppold, S.; Herschel, M. A System Framework for Personalized and Transparent Data-Driven Decisions. In Proceedings of the 32nd International Conference on Advanced Information Systems Engineering (CAiSE), Grenoble, France, 8–12 June 2020; Springer: Cham, Switzerland, 2020; pp. 153–168.
72. Lässig, N.; Oppold, S.; Herschel, M. Metrics and Algorithms for Locally Fair and Accurate Classifications using Ensembles. *Datenbank Spektrum* **2022**, *22*, 23–43. [CrossRef]
73. Gemp, I.; Theocharous, G.; Ghavamzadeh, M. Automated Data Cleansing through Meta-Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Palo Alto, CA, USA, 2017; pp. 4760–4761.
74. Dutta, A.; Deb, T.; Pathak, S. Automated Data Harmonization (ADH) using Artificial Intelligence (AI). *OPSEARCH* **2021**, *58*, 257–275. [CrossRef]
75. Behringer, M.; Hirmer, P.; Mitschang, B. A Human-Centered Approach for Interactive Data Processing and Analytics. In Proceedings of the Enterprise Information Systems: 19th International Conference, ICEIS 2017, Porto, Portugal, 26–29 April 2017; Revised Selected Papers; Hammoudi, S., Śmiałek, M., Camp, O., Filipe, J., Eds.; Springer: Cham, Switzerland, 2018; pp. 498–514.
76. Stöger, K.; Schneeberger, D.; Kieseberg, P.; Holzinger, A. Legal aspects of data cleansing in medical AI. *Comput. Law Secur. Rev.* **2021**, *42*, 105587. [CrossRef]
77. El Emam, K.; Mosquera, L.; Hoptroff, R. *Practical Synthetic Data Generation*; O'Reilly: Sebastopol, CA, USA, 2020.
78. Stach, C.; Bräcker, J.; Eichler, R.; Giebler, C.; Mitschang, B. Demand-Driven Data Provisioning in Data Lakes: BARENTS—A Tailorable Data Preparation Zone. In Proceedings of the 23rd International Conference on Information Integration and Web Intelligence (iiWAS), Linz, Austria, 29 November–1 December 2021; ACM: New York, NY, USA, 2021; pp. 187–198.
79. Hosseinzadeh, M.; Azhir, E.; Ahmed, O.H.; Ghafour, M.Y.; Ahmed, S.H.; Rahmani, A.M.; Vo, B. Data cleansing mechanisms and approaches for big data analytics: A systematic study. *J. Ambient. Intell. Humaniz. Comput.* **2021**, 1–13. [CrossRef]
80. Sharma, B. *Architecting Data Lakes: Data Management Architectures for Advanced Business Use Cases*, 2nd ed.; O'Reilly: Sebastopol, CA, USA, 2018.
81. Stach, C.; Bräcker, J.; Eichler, R.; Giebler, C.; Gritti, C. How to Provide High-Utility Time Series Data in a Privacy-Aware Manner: A VAULT to Manage Time Series Data. *Int. J. Adv. Secur.* **2020**, *13*, 88–108.
82. Stach, C.; Giebler, C.; Wagner, M.; Weber, C.; Mitschang, B. AMNESIA: A Technical Solution towards GDPR-compliant Machine Learning. In Proceedings of the 6th International Conference on Information Systems Security and Privacy (ICISSP), Valletta, Malta, 25–27 February 2020; SciTePress: Setúbal, Portugal, 2020; pp. 21–32.
83. Mindermann, K.; Riedel, F.; Abdulkhaleq, A.; Stach, C.; Wagner, S. Exploratory Study of the Privacy Extension for System Theoretic Process Analysis (STPA-Priv) to elicit Privacy Risks in eHealth. In Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops, 4th International Workshop on Evolving Security & Privacy Requirements Engineering (REW/ESPRE), Lisbon, Portugal, 4–8 September 2017; IEEE: Manhattan, NY, USA, 2017; pp. 90–96.

84. Shapiro, S.S. Privacy Risk Analysis Based on System Control Structures: Adapting System-Theoretic Process Analysis for Privacy Engineering. In Proceedings of the 2016 IEEE Security and Privacy Workshops (SPW), San Jose, CA, USA, 22–26 May 2016; IEEE: Manhattan, NY, USA, 2016; pp. 17–24.
85. Stach, C.; Mitschang, B. ACCESSORS: A Data-Centric Permission Model for the Internet of Things. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 22–24 January 2018; SciTePress: Setúbal, Portugal, 2018; pp. 30–40.
86. Stach, C.; Steimle, F. Recommender-based Privacy Requirements Elicitation—EPICUREAN: An Approach to Simplify Privacy Settings in IoT Applications with Respect to the GDPR. In Proceedings of the 34th ACM/SIGAPP Symposium On Applied Computing (SAC), Limassol, Cyprus, 8–12 April 2019; ACM: New York, NY, USA, 2019; pp. 1500–1507.
87. Stach, C.; Mitschang, B. Elicitation of Privacy Requirements for the Internet of Things Using ACCESSORS. In Proceedings of the Information Systems Security and Privacy: 4th International Conference, ICISSP 2018, Funchal, Portugal, 22–24 January 2018; Revised Selected Papers; Mori, P., Furnell, S., Camp, O., Eds.; Springer: Cham, Switzerland, 2019; pp. 40–65.
88. Gritti, C.; Chen, R.; Susilo, W.; Plantard, T. Dynamic Provable Data Possession Protocols with Public Verifiability and Data Privacy. In Proceedings of the 13th International Conference on Information Security Practice and Experience (ISPEC), Melbourne, VIC, Australia, 13–15 December 2017; Springer: Cham, Switzerland, 2017; pp. 485–505.
89. Gritti, C. Publicly Verifiable Proofs of Data Replication and Retrievability for Cloud Storage. In Proceedings of the 2020 International Computer Symposium (ICS), Tainan, Taiwan, 17–19 December 2020; IEEE: Manhattan, NY, USA, 2020; pp. 431–436.
90. Stach, C.; Gritti, C.; Mitschang, B. Bringing Privacy Control Back to Citizens: DISPEL—A Distributed Privacy Management Platform for the Internet of Things. In Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing (SAC), Brno, Czech Republic, 30 March–3 April 2020; ACM: New York, NY, USA, 2020; pp. 1272–1279.
91. Gritti, C.; Önen, M.; Molva, R. CHARIOT: Cloud-Assisted Access Control for the Internet of Things. In Proceedings of the 2018 16th Annual Conference on Privacy, Security and Trust (PST), Belfast, Ireland, 28–30 August 2018; IEEE: Manhattan, NY, USA, 2018; pp. 1–6.
92. Gritti, C.; Önen, M.; Molva, R. Privacy-Preserving Delegable Authentication in the Internet of Things. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC), Limassol, Cyprus, 8–12 April 2019; ACM: New York, NY, USA, 2019; pp. 861–869.
93. Chaum, D.; Damgård, I.B.; van de Graaf, J. Multiparty Computations Ensuring Privacy of Each Party's Input and Correctness of the Result. In Proceedings of the 7th Annual International Cryptology Conference (CRYPTO), Santa Barbara, CA, USA, 16–20 August 1988; Springer: Berlin/Heidelberg, Germany, 1988; pp. 87–119.
94. Shamir, A. How to Share a Secret. *Commun. ACM* **1979**, *22*, 612–613. [CrossRef]
95. Barker, E. *Recommendation for Key Management: Part 1—General*; NIST Special Publication 800-57 Part 1, Revision 5; National Institute of Standards and Technology, Technology Administration: Gaithersburg, MD, USA, 2020; pp. 1–158.
96. Mavroeidis, V.; Vishi, K.; Zych, M.D.; Jøsang, A. The Impact of Quantum Computing on Present Cryptography. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 405–414. [CrossRef]
97. Borges, F.; Reis, P.R.; Pereira, D. A Comparison of Security and its Performance for Key Agreements in Post-Quantum Cryptography. *IEEE Access* **2020**, *8*, 142413–142422. [CrossRef]
98. Behringer, M.; Hirmer, P.; Fritz, M.; Mitschang, B. Empowering Domain Experts to Preprocess Massive Distributed Datasets. In Proceedings of the 23rd International Conference on Business Information Systems (BIS), Colorado Springs, CO, USA, 8–10 June 2020; Springer: Cham, Switzerland, 2020; pp. 61–75.
99. Stach, C.; Brodt, A. vHike—A Dynamic Ride-Sharing Service for Smartphones. In Proceedings of the 2011 IEEE 12th International Conference on Mobile Data Management (MDM), Luleå, Sweden, 6–9 June 2011; IEEE: Manhattan, NY, USA, 2011; pp. 333–336.
100. Stach, C. Secure Candy Castle—A Prototype for Privacy-Aware mHealth Apps. In Proceedings of the 2016 IEEE 17th International Conference on Mobile Data Management (MDM), Porto, Portugal, 13–16 June 2016; IEEE: Manhattan, NY, USA, 2016; pp. 361–364.