future internet

Article

Protecting Sensitive Data in the Information Age: State of the Art and Future Prospects

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

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

Abstract: The present information age is characterized by an ever-increasing digitalization. Smart devices quantify our entire lives. These collected data provide the foundation for data-driven services called smart services. They are able to adapt to a given context and thus tailor their functionalities to the user's needs. It is therefore not surprising that their main resource, namely data, is nowadays a valuable commodity that can also be traded. However, this trend does not only have positive sides, as the gathered data reveal a lot of information about various data subjects. To prevent uncontrolled insights into private or confidential matters, data protection laws restrict the processing of sensitive data. One key factor in this regard is user-friendly privacy mechanisms. In this paper, we therefore assess current state-of-the-art privacy mechanisms. To this end, we initially identify forms of data processing applied by smart services. We then discuss privacy mechanisms suited for these use cases. Our findings reveal that current state-of-the-art privacy mechanisms provide good protection in principle, but there is no compelling one-size-fits-all privacy approach. This leads to further questions regarding the practicality of these mechanisms, which we present in the form of seven thought-provoking propositions.

Keywords: smart service; privacy techniques; location-based services; health services; voice-controlled digital assistants; image analysis; food analysis; recommender systems; DNA sequence classification

check for updates

Citation: Stach, C.; Gritti, C.; Bräcker, J.; Behringer, M.; Mitschang, B. Protecting Sensitive Data in the Information Age: State of the Art and Future Prospects. *Future Internet* **2022**, *14*, 302. https://doi.org/10.3390/fi14110302

Academic Editor: Wei Yu

Received: 30 August 2022
Accepted: 18 October 2022
Published: 22 October 2022

1. Introduction

In 1991, Mark Weiser envisioned the computer for the 21st century [1] as a pervasive system that ubiquitously surrounds us, constantly adapting to its context and our current needs. Although this vision did not fully materialize, the *Internet of Things* (*IoT*) is a major step in this direction. Here, various sensors are integrated into everyday objects, enabling them to monitor their surroundings and react to it. Furthermore, all of these IoT-enabled devices, often referred to as *smart devices*, are interconnected and can communicate with each other. Thus, smart devices have a virtually unlimited data stock at their disposal [2].

The full potential of the gathered data can only be exploited if they are interlinked and comprehensively analyzed [3]. Yet, these data, which are labeled *big data*, are generated at high *velocity* and in high *volume*. Profound data processing is therefore not possible on the mostly lightweight smart devices, as they do not have the necessary resources and computing power to do this at an adequate speed and scale. Moreover, there is a high *variety* in the data in terms of schemata and data formats. Thus, extensive data preparation is required to merge these data, which also exceeds the capacities of smart devices [4].

The processing of the captured raw data therefore usually takes place in a powerful backend system. There, the raw data are initially refined. That is, they are cleansed—i.e., missing or erroneous data are treated—and transformed—i.e., different schemata and formats are harmonized [5]. Subsequently, the refined data can be interlinked, further preprocessed, and analyzed [6]. In analogy to the concepts of the knowledge management, three stages can be differentiated. *Data* refers to the unprocessed and unfiltered raw data collected by the smart devices. The refined and interlinked data are termed *information*. *Knowledge* is generated only when the information is analyzed, and the findings are interpreted [7].

Such knowledge provides the foundation for so-called *smart services*. A smart service is any kind of data-driven digital service that is able to adapt to the data and thus offering users the greatest possible utility in their current situation. Smart services can be very small scale, e.g., an adaptive application on a smartphone, as well as highly complex, e.g., when several smart devices interact with each other via actuators [8]. Examples of smart services can be found in the public, industrial, and private sectors, e.g., in the context of *smart cities* [9], *Industry 4.0* [10], and *eHealth* [11].

While such smart services are very appealing since they significantly facilitate the lives of their users in a wide variety of situations, they also pose a serious threat. As we are constantly surrounded by smart devices, they are able to quantify all aspects of their users' lives as well as the lives of innocent bystanders. The thereby gained knowledge provides deep insights into the privacy of these individuals [12]. Data protection laws, such as the *European General Data Protection Regulation* (*GDPR*) [13], therefore entail principles, such as *data minimization* (Article 5(1)(c)), and mandate *data protection by design* (Article 25).

However, this leads to what is known as the *privacy paradox*—although users crave the best possible protection of their privacy, they still do not want to refrain from using smart services, which in turn requires them to share their private data with these services [14]. Therefore, this paradox must be resolved by concealing sensitive knowledge patterns contained in the data without significantly impairing the general data quality. Privacy measures that are too restrictive would render the data unusable, while measures that are too shallow would jeopardize privacy [15]. In order for this balancing act to succeed, however, privacy measures must be tailored to the data and how they are processed.

Therefore, we investigate whether state-of-the-art privacy mechanisms for smart services are up to this task. Our research goal is to systematically analyze the state of the art in this domain. To this end, we identify the privacy threats posed by today's smart services as well as the strengths and weaknesses of the available privacy mechanisms. By comparatively reviewing these two dimensions, we elaborate which issues are already effectively covered by today's privacy mechanisms and which open research questions still need to be addressed as part of future work. With this in mind, we make the following three contributions in our paper:

1. We present modern-day smart services from seven application domains. For each of them, we analyze which data they capture, which types of processing are used to extract information from them, and which knowledge can be derived. The selected application scenarios serve to illustrate the general data processing requirements and the privacy concerns inherently associated with such smart services.
2. We discuss state-of-the-art privacy measures for the identified data types and forms of processing. Hereby, we provide an overview of the current state in terms of the protection of sensitive data when dealing with smart services.
3. We identify, based on our findings, open privacy issues in the context of smart services that need to be overcome in order to comply with the data protection by design principle.

The remainder of this paper is structured as follows: In Section 2, we analyze seven real-world application scenarios of smart services with respect to the involved data processing and highlight the opportunities offered by these smart services as well as the inherent privacy threats. Based on our findings concerning data quality requirements and potential privacy risks, we discuss appropriate state-of-the-art privacy measures and work out their strengths and weaknesses in Section 3. In Section 4, we map the opportunities and threats against the strengths and weaknesses to reflect the current state of privacy mechanisms for smart services. Based on this contrasting juxtaposition, we identify future prospects with regard to privacy issues in smart services and how to overcome them in Section 5 before concluding the paper in Section 6.

2. Analysis of Modern-Day Smart Services

In a first step, we look at seven real-world application scenarios, which require the comprehensive processing of highly sensitive data. This provides insights into the types of data that are involved and the requirements regarding data refinement. In addition, we discuss for each application scenario which opportunities are offered by it and which privacy threats it poses. The selection of the application scenarios is based on the IoT and smart service topics which are currently predominant in the literature. This includes *location-based services* (Section 2.1), which originally had a primarily military background but have long since arrived in the private sector, *health services* (Section 2.2), in which the IoT enables remote health monitoring, *voice-controlled digital assistants* (Section 2.3), which are an important pillar of any smart home, and surveillance services driven by *image analysis* (Section 2.4) [16]. Moreover, end-to-end food processing monitoring to ensure food safety is becoming increasingly common, which is enabled by IoT-supported *food analysis* (Section 2.5) [17]. However, even long-established services, such as *recommender systems* (Section 2.6) gain new capabilities due to the IoT [18]. As the available computing power is also steadily increasing, extensive analysis such as *DNA sequence classification* (Section 2.7) can also be realized with the help of the IoT [19]. In those application domains in general, there are particularly severe privacy threats, such as data leakage or data tampering [20]. We summarize the findings from the study of these seven application scenarios in Section 2.8 and identify knowledge needs to be derived in each scenario as a result of the analysis and what adverse insights can be gained from the data in the process.

2.1. Location-Based Services

Location-based services (*LBS*) are an application domain in which large amounts of data have to be processed. An LBS is a service in which the geographic location of entities is significantly tied in. In addition to tracking stationary entities, e.g., certain locations that are relevant for the service, mobile entities can also be tracked, e.g., shipments that are in delivery [21]. In addition to such non-human entities, the location data of human users are also being tracked by LBS. This is driven in particular by the increased use of smartphones (and affiliated technologies, such as smartwatches) [22]. This kind of device is not only permanently close to its user but also *always on* and *always online* [23]. In addition to a GPS receiver, which enables very accurate positioning, mobile phone tracking also facilitates sufficiently good positioning of users via their GSM cellular location [24]. While satellite-based tracking is limited to outdoor areas, indoor tracking is also possible with standard smart devices. For instance, inertial approaches use built-in accelerometers and gyroscopes to determine the location relative to a given starting point by means of the direction of movement and movement speed. Other approaches are based on the Earth's magnetic field. A built-in magnetic sensor detects the radiated fields, which can be used to determine the current position by means of triangulation [25].

Basically, three different types of location capturing can be found in LBS. The most basic method for static entities is to hardcode their position. For mobile entities, the current location can be captured as singular location information with the help of the aforementioned tracking technologies. If several such singular locations are captured in a temporal sequence, they can be used to form trajectories that describe the movement path of an entity [26]. The latter in particular requires thorough data refinement, e.g., to eliminate outliers [27] or to compensate for inaccuracies in the location information [28].

Figure 1 illustrates three data refinement steps. A problem regarding the capturing of trajectories is that they are only recorded pointwise. Single points can be distorted due to poor positioning signals, e.g., when an entity is passing urban canyons. If the recorded location deviates strongly from the actual location, noise filtering can be used to rectify the trajectories. An outlier can be detected if a point of the trajectory deviates too much from its predecessor and successor. In that case, the corrupted point can be detected, deleted and, e.g., replaced by means of interpolation. If the sampling rate is higher, i.e., if more data are available for data refinement, the results of the cleansing become more accurate [29].

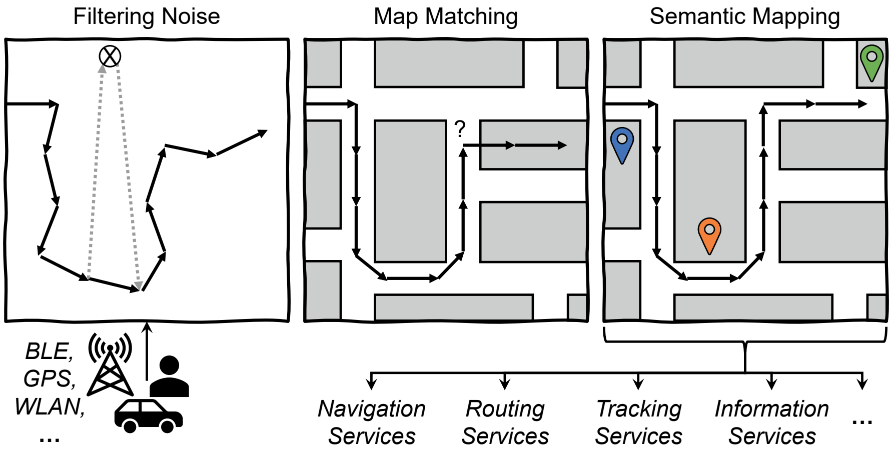

Figure 1. Examples of Data Refinement Steps for Trajectory Data.

While noise filtering can be used to correct significant deviations from single points, smaller deviations can also be corrected by adding complementary knowledge. For instance, if an entity can only move on predefined paths, e.g., a parcel on a parcel conveyor or a car on a road, the trajectory can be projected onto a model of these possible paths. In the depicted map matching, the white paths are roads, and the entity is a car. The denoised trajectory can then be smoothed, since it can be assumed that the car follows the road [30].

As can be seen, however, such a smoothing has its limitations. For instance, it is not possible to determine where the trajectory ends. It is obvious that the car cannot drive off the road. However, the distance to the road above and the one below is identical. If further semantic knowledge is added to the model, even such errors can be corrected. For instance, if some locations are known that are relevant to the entity, then it can be inferred whether it is more likely to use the upper or lower road. In the shown semantic mapping, the green marker is the entity's place of residence. Thus, it can be assumed that the captured trajectory ends there, and the refinement of the trajectory can be made accordingly [31].

Such refined data enable a variety of LBS. For instance, they enable the search for stationary objects—e.g., POI in a guide service [32]—or mobile objects—such as drivers or passengers in a ride-sharing service [33]. Smart navigation systems are also made possible by such data. These navigation systems not only take the planned route into consideration but also contextual knowledge such as congestion information provided by other entities [34]. Social networks in particular are increasingly relying on IoT data from other smart services. Here, these data are merged with social media, whereby new insights are gained. Location data play a significant role in this context, as they are not only a key factor in determining the current context of a person but also because they reveal further connections between users [35]. For instance, in location-based social networks, social groups are identified not only based on mutual friends and common interests but also on matching movement patterns [36]. Another cutting-edge use case is the development of autonomous vehicles that do not require any human intervention to find and follow their route [37]. While in the aforementioned use cases, it is only bothersome if the data refinement is inadequate and thus the tracking of entities is not accurate, in autonomous vehicles, this might endanger the physical integrity of humans. Therefore, it is mandatory to delve into the data. Yet, data processed by LBS provide insights into behavioral patterns. Whenever humans can be linked to the data (e.g., the owner of a smart device), this poses a severe privacy threat [38].

The inherent opportunities and privacy threats of LBS can be broken down as follows:

Opportunities:

- 💡 The technical ability to locate smart devices (and thus their owners) with high accuracy enables many services, such as navigation services or location-based information services.
- 💡 The location can also be used to derive a lot of additional information about the data subject, e.g., which places the data subject visits frequently.
- 💡 If this information is enriched with additional data, such as temporal aspects (How long and when does a data subject stay at a certain place?) or supplementary geographic data (What can be found at that place?), a very precise profile of the data subject can be created. This makes LBS the foundation of many other context-based services since location is a key parameter in context recognition.

Privacy Threats:

- ⚠ The current location of a data subject is continuously disclosed by an LBS. This enables long-term surveillance of data subjects.
- ⚠ The data refinement methods described above make it easy to correct even hardware- or software-related inaccuracies, enabling very precise location determination.
- ⚠ Furthermore, LBS can be used to find out much more about a data subject than might initially appear. For instance, they can be used to determine activities and, in the case of long-term use, to draw conclusions about hobbies and social contacts.

2.2. Health Services

The healthcare sector can also benefit from comprehensive data analyses. The term *eHealth* covers all kinds of services that facilitate the treatment and long-term care of patients and involve the use of modern information technologies [39]. This brings many benefits to the table, as treatment costs can be reduced, the patients' quality of life can be improved, and the workload of physicians can be reduced [40]. To achieve this, however, it is crucial to include the patients fully in the process and overcome technical hurdles [41]. This can be achieved in particular by means of so-called *mHealth*. Here, everyday mobile devices such as smartphones are used for medical treatment [42]. Due to mHealth, there is a mobile application for virtually any health-related use case nowadays [43].

An example of such an mHealth application for people suffering from a chronic disease is an interactive questionnaire with which they can assess their condition on a daily basis. Depending on the answers, follow-up questions are asked. In this way, only essential questions can be asked in a systematic manner without overwhelming patients [44]. In addition to a local evaluation to determine the appropriate question catalog, the results are also forwarded to a backend for further analysis. If a critical situation is detected based on the answers, the patient is advised to see a physician [45]. The physician receives a summary of the analyzed questionnaires, which facilitates his or her daily work and allows him or her to focus on emergency cases [46].

Such applications are especially useful in more rural areas or developing countries where people have no immediate access to a physician. Thus, because of the low-cost hardware required, mHealth can ensure that people in such areas still receive healthcare [47]. Today's smart devices, however, enable far more powerful heath services due to their built-in sensors and connectivity [48]. The so-called *quantified self movement* motivates people to use self-tracking to capture vast amounts of health-related data about themselves, such as blood pressure information, on a permanent basis [49]. While ordinary smart devices come with many sensors that support this kind of self-tracking, the IoT opens up further capabilities [50]. For instance, a smartphone can be turned into a health data hub to which all sorts of IoT-enabled specialized heath meters send their measurement data for storage and processing [51]. As a result, mHealth approaches provide not only a sufficient amount of data but also useful data for more comprehensive medical analyses [52].

In addition to simple questionnaires, the self-assessment of patients can be supported by such health-related measurements. In the case of diabetics, a continuous glucose monitoring device can be coupled with a smartphone to track the blood glucose level in a mobile application. Moreover, external circumstances that might influence the measurements (e.g., stress factors such as noise) can also be captured via the smartphone's built-in sensors. In this way, analyses enable a 360-degree health view on the patients [53]. As these data are tracked continuously, more complex analyses of long-term trends are also feasible [54].

Figure 2 shows the potential of such analyses. First, *descriptive analytics* can be applied to the large amounts of historical data collected about a patient. By looking at a health value over time, it is possible to understand what exactly has happened, e.g., how a patient responded to taking a particular pharmaceutical, and use this knowledge to adjust the medication. Yet, as the data are captured in real time, the current situation can also be monitored by means of *real-time analytics*. For instance, this knowledge can be used to operate a smart insulin pump that supplies a patient with the appropriate amount of insulin based on the current situation. In addition, a model can be trained using historical data, which enables *predictive analytics*. Predicting how a health value will change in the future enables counteracting an adverse trend at an early stage [55]. The potential of such services is therefore virtually unlimited, including education, diagnostics, and treatment [56].

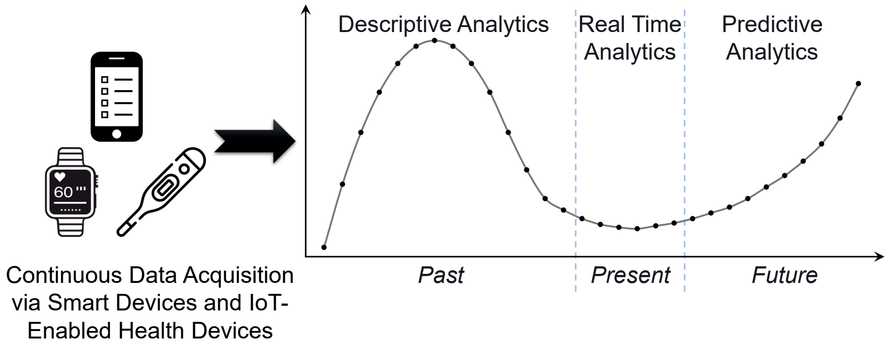

Figure 2. Three Types of Time-Series Analyses Applied in Health Services.

In such an application scenario, the accuracy of the data is key, as it affects the health of patients. Thus, the data must be refined assiduously in order to achieve the highest possible accuracy and eliminate measurement errors [57]. However, such services also raise privacy concerns, since in addition to the health data, which intrinsically contain a vast amount of sensitive information, a vast amount of knowledge about, e.g., lifestyle, environment, and social life is disclosed [58]. As smart devices have become an integral part of our lives, users no longer even notice their presence. Sensors are embedded in everyday objects and enable a comprehensive and almost invisible permanent monitoring of everyone—both the users of the smart devices or accidental bystanders [59]. For this reason, particularly high privacy standards must be applied in such an application scenario [60].

The inherent opportunities and privacy threats of smart health services can be broken down as follows:

Opportunities:

- IoT technologies enable the self-quantification of health-related values, which means that frequently recurring health checks in particular can be performed at home by the patients themselves. This relieves the burden on both patients and physicians.
- The non-intrusive nature of the smart devices allows the permanent monitoring of patients without disturbing their daily routine. This enhances safety, as no health measurements can be forgotten, and health problems can be detected at an early stage.

💡 Since the smart devices that feed smart health services with data are ubiquitous and capable of capturing a variety of health values, they can be used to provide 360-degree health views on the patient.

Privacy Threats:

⚠ Health data are among the most sensitive data, so the large-scale collection and processing in itself is a privacy threat.

⚠ As smart devices are ubiquitous, data subjects are no longer aware that health data are collected permanently, which makes them unaware of the privacy threat.

⚠ In addition to inferences about diseases, the collected health data also allow insights into other aspects, such as unhealthy behaviors, e.g., whether the data subject is a smoker or carries out little physical activities.

2.3. Voice-Controlled Digital Assistants

In addition, the increasingly popular *Voice-Controlled Digital Assistants* (*VDA*) depend on comprehensive data refinement and subsequent data processing. VDA refers to any hardware and software infrastructure that enables users to request information or give instructions by means of human speech. In addition to IoT devices that serve as microphones and speakers, a software agent is required that permanently listens for a predefined keyword. As soon as this keyword is received, the agent wakes up and records everything that is subsequently spoken. A speech analysis is then performed on the recordings, to interpret the spoken commands semantically and process them logically. Depending on the command, either speech synthesis is used to formulate a response or a machine-processable command is sent to a corresponding IoT-capable device [61]. The most popular VDA are *Alexa* (see https://alexa.amazon.com/, accessed on 30 August 2022), *Siri* (see https://www.apple.com/siri/, accessed on 30 August 2022), *Cortana* (see https://www.microsoft.com/en-us/cortana/, accessed on 30 August 2022), and *Google Assistant* (see https://assistant.google.com/, accessed on 30 August 2022) [62].

Such VDA offer users a variety of benefits: First, the easy accessibility via the voice interface provides a convenient way to give tasks to the VDA. Second, this type of interaction also satisfies a hedonistic desire, as the user has command over a (virtual) assistant that can perform any assigned task. This is further amplified by the fact that the possession of such “future technology” also has a symbolic value, since its usage has a favorable impression on others. Third, VDA also have a positive social aspect, as natural language communication with them overcomes many impediments regarding the perception of the technology—the VDA is rather perceived as a person than a technical object [63]. Due to this variety of benefits, it is therefore not surprising that the popularity and demand for such VDA is growing constantly. This is further enhanced by the fact that more and more IoT devices are compatible with VDA, i.e., everyday objects can be used as input device and can be operated by a VDA. So, the VDA is virtually available at any time and any place [64].

The schematic architecture of the VDA ecosystem is outlined in Figure 3. A compatible input client with a microphone is required on the user side. The software agent of the respective VDA service runs on this client. This can be a conventional personal computer with Internet access, a smart device such as a smartphone, or a dedicated VDA-enabled device such as *Amazon Echo* (see https://www.amazon.com/smart-home-devices/b?node=9818047011, accessed on 30 August 2022). On this client itself, however, neither the voice processing nor the processing of the actual request is carried out. The client only recognizes the specified keyword and records the subsequent instructions. It sends the recording to its processing backend. There, voice processing first retrieves and interprets the request [65]. The request is then processed by means of machine learning approaches. Feedback loops are included so that if the user labels a response incorrect or unsatisfactory, the respecting request is used to train the models further and thereby refine them continuously. To put it simply, with every mistake the VDA makes, the system becomes smarter. IoT-enabled output devices are needed to render the results, e.g., smart speakers, a smartphone application, or compatible third-party devices such as smart light bulbs. To trigger these devices, a VDA backend has various adapters, e.g., an adapter to generate natural language to answer questions verbally or adapters by third-party vendors to control their IoT devices [66].

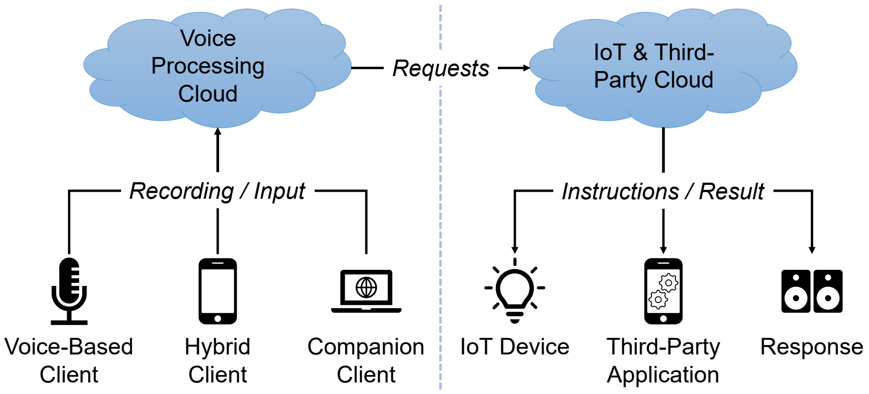

Figure 3. Ecosystem and Architecture of a Voice-Controlled Digital Assistant.

No matter which service is requested by a user, however, an important aspect of VDA is that the misinterpretation of his or her commands is minimal. If this is not achieved, a breach of immersion occurs and user acceptance decreases [67]. To this end, it is imperative that in the backend, models are trained based on many inputs from different users. Furthermore, the incoming requests must be thoroughly refined in order to be able to successfully interpret what is being said. In addition to monitoring the quality of the VDA, a feedback loop is required that involves human operators listening to selected recordings [68].

It is obvious that VDA also raise privacy concerns. The VDA providers are able to identify all of their users. Thus, they are able to link the recordings to the users. Taking this into account, the fact that VDA are able to wiretap users permanently is a particularly disturbing concern. Users do not know what data are actually captured and forwarded for processing. Even third parties can gain access to the data, e.g., for quality control or to provide an affiliated service. So, the privacy concerns regarding VDA are reasonable [69].

The inherent opportunities and privacy threats of voice-controlled digital assistants can be broken down as follows:

Opportunities:

- VDA allow voice-based control of smart devices, making them particularly helpful, e.g., for people with motor disorders.
- The natural language approach of VDA reduces the technical hurdle for people who are less tech-savvy.
- The knowledge that a VDA (theoretically) has at its disposal is almost unlimited. That is, VDA can be used to access required information quickly and easily in almost any situation in life.

Privacy Threats:

- Since a VDA waits for its specific keyword, it is never completely off. That is, all conversations are permanently recorded.
- If a VDA is activated using the keyword and the voice recording is forwarded for processing, it is not possible for data subjects to trace who has access to it.
- For third parties, a VDA is indistinguishable from conventional loudspeakers. Therefore, they are completely unaware that their conversations are also being recorded.

2.4. Image Analysis

There are also many benefits offered by computer-aided image analysis. While the human eye has physical limitations in terms of detail perceptibility, the capabilities of computational image processing are virtually limitless. Therefore, such data processing techniques are applied in a variety of application domains today. For instance, in the food industry, computer vision can be used to reliably detect foreign objects in food products, medical diagnoses can be supported by the automatic interpretation of computed tomography images, and in the context of defense and homeland security, suspects can be recognized on video recordings [70].

In particular in the field of face recognition, remarkable progress has been made in recent years. While in the early days of this research area in 1964, only a few characteristics of the face could be recognized on images, today, the techniques are so reliable and accurate that they are used in a variety of commercial, industrial, legal, and governmental applications [71]. End-users also come into contact with face recognition services, e.g., in *online social networks* (*OSNs*) such as *Facebook* (see https://www.facebook.com/, accessed on 30 August 2022) [72]. Here, the so-called *DeepFace* algorithm is used to identify users on images shared in the network. This algorithm is almost as good as human identification, except that, unlike humans, DeepFace can process large amounts of data in next to no time [73].

Figure 4 illustrates the workflow of face recognition. Initially, all faces on an image must be detected. For instance, characteristic facial features can be identified such as skin color which contrasts with the background, or a model can be trained to detect human faces [74]. Having detected the face, the image needs to be preprocessed, e.g., it has to be cropped to the relevant part of the face. Furthermore, facial landmarks such as the eyes, mouth, and nose have to be marked [75]. For the actual recognition, a comparison with a face database is conducted. To this end, a multinomial classifier is trained using the face database. This classifier can determine with which person from the face database the person on the image has the best match [76]. Various machine learning techniques can be used for this purpose. While in the early days primarily *linear discriminant analysis* was used, today, *support vector machines* and *convolutional neural networks* are used [77].

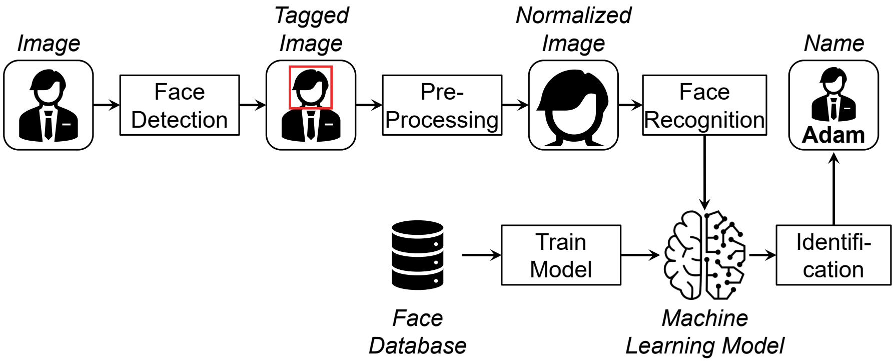

Figure 4. Workflow of a Face Detection and Recognition Based on Machine Learning.

The application of such image analysis techniques naturally also raises privacy concerns. With today's smartphones, IoT-enabled cameras have become ubiquitous in our everyday lives. Moreover, considering that smartphones can tag captured photos with a time code as well as location information, image analysis can be used to determine not only which people are in a picture together but also when and where that meeting has happened. While this may be intended by the main data subject of the photo, it also affects any bystanders. Due to this potential exposure, measures are needed to restrict such privacy-intrusive scenarios [78]. An additional problem associated with image analysis is the bias in data selection. Such a bias leads to inadequate models, which cause errors in face recognition, resulting in users being incorrectly linked to a photo. Thus, it is important to take measures to prevent bias when preparing the data corpus for training the models [79].

The inherent opportunities and privacy threats of image analysis can be broken down as follows:

Opportunities:

- 💡 As social media becomes more and more prevalent in people's lives, image analysis is becoming increasingly relevant for them as well. This allows people to be identified and tagged in images, enabling the automatic linking of people with their social contacts as well as with places and activities.
- 💡 Comprehensive image analysis enables novel search functionalities, e.g., if users want to find all images of themselves (or other users) that are available in a social network.
- 💡 Image analysis is also a key factor in law enforcement and security today, as it can be used to identify suspects rapidly in video recordings.

Privacy Threats:

- ⚠ When an image is analyzed, locations or activities can be identified in addition to the people depicted in it. By linking this information, a lot of knowledge about the data subject can be derived. Furthermore, by combining all available images, a comprehensive insight into the lives of the depicted persons can be gained.
- ⚠ The algorithms are subject to certain probabilities of error. If people are incorrectly identified, they may be assigned to the wrong circles of acquaintances or interests, unnoticed by the data subject, which in the worst case can have damaging consequences for their reputation.
- ⚠ Third parties can also be tagged on images without their knowledge, allowing the algorithm to learn their faces. As a result, they can also be identified in pictures, which means that parts of their private lives are revealed completely without their knowledge.

2.5. Food Analysis

The aim of food chemistry is to investigate and research all biological and non-biological components of foods. This research discipline is fully committed to consumer protection and is intended to ensure the safety of all food products. To this end, chemical analyses are used to determine, e.g., whether sufficient quantities of a valuable ingredient are included in a food product or whether food producers comply with the maximum permitted quantities of unhealthy substances [80]. Here, the use of biosensors can contribute significantly to food safety, as contaminations or toxins in food can be reliably detected [81]. For this purpose, the vast number of proteins in a food sample must be split into peptides in order to detect even the smallest traces of substances that can be unsafe to consume. However, this results in a significantly larger number of analytes. Special tools and techniques are therefore needed to handle this huge amount of data [82].

In simple terms, it is necessary to search for specific patterns in the source material [83]. To cope with the vast amount of analysis data, machine learning is becoming increasingly popular. This enables a much more thorough analysis than a manual analysis. Models are trained from examples. These models generalize the data from the given samples. For new samples, it must be determined to which of the training samples the new sample has the best match. This allows one to analyze even very large datasets, such as complete DNA sequences, for contained patterns [84].

An analysis process in the food chemistry domain is shown in Figure 5. A food sample is analyzed to determine whether it contains certain allergens. Such allergens must be explicitly declared by food producers to protect consumers who have a certain food allergy. In this example, a food chemist determines whether the sample contains glutamate. Yet, the same procedure applies to other allergens as well, such as gluten, crustaceans, or nut seeds. For this purpose, the sample is analyzed with a mass spectrometer to determine the mass-to-charge ratio of all peptides present in a sample. For the separation of contained peptides, liquid chromatography is used. From a proteomics and genomics database for foods, e.g., *UniProt* (see https://www.uniprot.org/, accessed on 30 August 2022), comparison samples of the allergens of interest can be obtained. Then, it can be cross-checked whether sufficiently similar patterns can be recognized in the peptides obtained from the food sample [85]. For this comparison and all subsequent analyses, however, a thorough refinement of the sample data is necessary, e.g., to filter out inaccurate or irrelevant data to be able to focus on the relevant aspects in the data, only [86]. This way, it is possible to detect whether allergens are present in a food sample in any decomposing stage [87].

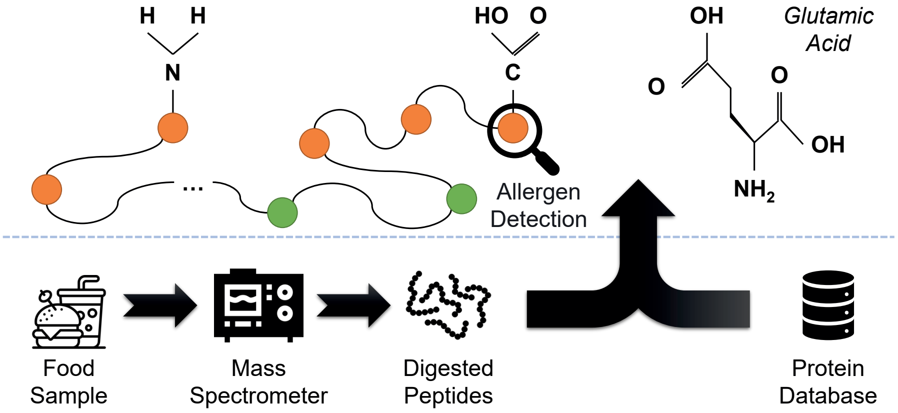

Figure 5. Schematic Representation of Food Analysis to Detect Allergens in a Food Sample.

Such a pattern recognition in a data stream, as it is generated by the measurement devices during the analysis, is also known from the domain of *complex event processing* (*CEP*). A pattern consists of individual facts—in the context of food analysis for instance peptides—which are in a given sequential context [88]. The patterns can be defined by means of pattern templates. These templates provide a level of abstraction to the actual data, as they specify the target sequences based on pairs of measurement values and constraints on those values. A CEP engine then searches for these patterns in the data stream [89]. In the context of food analysis, however, it is not sufficient to find only exact matches but also reasonably good matches. For instance, post-translational protein changes can cause deviations from the entries in the protein sequence database. To this end, data mining techniques such as *k*-nearest neighbor analysis can be used to find similarity matches [90].

It is obvious that privacy does not have to be observed in this application scenario, as the food samples do not have any privacy rights. However, with the help of these analytical methods, it is possible not only to detect allergens in food products but also to determine which ingredients are present in the product. That is, deep insights into the product and manufacturing process can be gained with today's sensors and data processing technology [91]. As a result, it is possible to identify a manufacturer's secret ingredients, which represents a competitive advantage in the food market. Therefore, maintaining confidential business information is a necessary tool in the food chemical sector [92].

The inherent opportunities and privacy threats of IoT-supported food analysis can be broken down as follows:

Opportunities:

- With IoT-supported food analysis, food samples can be analyzed much more efficiently and effectively.
- Due to the increasing number of people suffering from a food allergy, it is important that food products are correctly labeled and that this labeling is also thoroughly verifiable.

- 💡 Due to a predominantly automated processing of food products, a thorough inspection of these products is required in order to detect any foreign substances or contaminants at an early stage.

Privacy Threats:

- ⚠ In this application scenario, there are no privacy threats, but there are confidentiality threats, as food analysis can provide deep insights into the food product, revealing specific ingredients or preparation methods, possibly leading to a loss of competitive advantage.

2.6. Recommender Systems

With the rise of services such as *YouTube* (see https://www.youtube.com/, accessed on 30 August 2022) or large online shopping platforms such as *Amazon* (see https://www.amazon.com/, accessed on 30 August 2022), we have entered an age of oversupply. Users cannot obtain a complete overview of all available items (e.g., video clips or products). Therefore, in all these services and shopping platforms, a mechanism is needed to provide users with information about the items that are particularly relevant to them. Recommender systems represent such a mechanism. In simple terms, a recommender system is a tool that predicts a user's interests based on his or her previous interactions with the service or shopping platform. Such interactions can be, e.g., watching a video clip or buying a product. In this way, information overload can be minimized by presenting a user with only the most relevant information tailored to his or her needs [93]. Although e-resource services such as YouTube or e-commerce platforms such as Amazon are the most prominent uses of recommender systems, such systems have become indispensable in almost all modern e-services [94].

However, if recommender systems would base their suggestions only on the previous interaction of the user in question, no new items could be recommended to him or her. The advantage of these systems is that they not only have knowledge about a single user but can also draw on the interactions of a large number of users. By means of so-called *market basket analysis*, they can determine which combinations of items users frequently interact with, e.g., which video clips a user has on his or her watchlist or which products a user buys at the same time. From such item lists, *association rules* can be derived that describe which item combinations are encountered regularly. If a user is interested in a subset of the items of an association rule, the remaining items can be recommended to him or her [95].

In Figure 6, it is shown how a recommender system can proceed to be able to make tailored recommendations. Basically, there are two different approaches to this end: *content-based filtering* and *collaborative filtering*. In content-based filtering, only the items are considered. For this purpose, similar items are clustered. The clusters are homogeneous within themselves—i.e., the items of a cluster are mutually highly similar with regard to the relevant properties—and heterogeneous to each other—i.e., the items of different clusters are as different as possible with regard to the relevant properties. If a user interacts with an item (or has interacted with it in the past), the remaining items of the same cluster can be recommended to the user, since it is likely that they are also of interest to him or her [96].

Yet, this approach completely neglects social aspects. By including knowledge about the users themselves as well as social relationships between them, considerably better recommendations can be made [97]. In collaborative filtering, therefore, users are also clustered, for instance, based on common interests. This has two effects: On the one hand, the recommendation set can be reduced significantly. It is no longer necessary to recommend all similar items but only those that were also popular among other users from the same cluster. On the other hand, new recommendations can also include distinct items if other users from the cluster have interacted with them [98]. In addition to these two main types of recommender systems, however, there is a variety of hybrid approaches that mix aspects of content-based filtering and collaborative filtering [99].

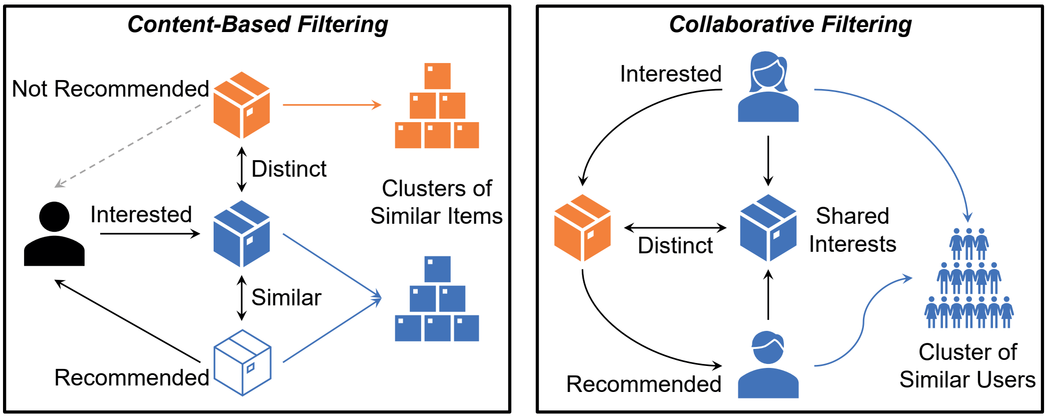

Figure 6. Two Fundamentally Different Modes of Operation of Recommender Systems.

There are two issues in this regard: First, third parties, e.g., content creators or manufacturers of a product, might try to increase the relevance of their items by submitting fake data to be prioritized in the recommendations. So, raw data always have to be refined to purge such misrepresentations. Second, recommender systems rely on knowledge about the community. Only if large amounts of information about users' preferences and usage behavior are available can meaningful recommendations be made. Even if an individual's opinion cannot be inferred directly from the recommendations, such information is present in the base data. Moreover, statistical inferences can also be derived from the processed and accumulated data, which reveal information about the preferences of individuals. Therefore, privacy issues have to be addressed in recommender systems as well [100].

The inherent opportunities and privacy threats of IoT-supported recommender systems can be broken down as follows:

Opportunities:

- IoT-supported recommender systems are able to provide search results that are tailored to the user (e.g., product recommendations) based on contextual information.
- With the help of collaborative filtering, users can also be presented with completely new recommendations, which can expand their horizons, as they were previously unaware that they might be interested in the suggested items.
- Searches become much more efficient, as irrelevant items can be excluded early on, and more effective, as relevant items can be suggested even if they were not directly included in the search query.

Privacy Threats:

- The recommender systems have to collect and analyze a lot of data about a user's interests in order to make suitable suggestions. As a result, they also gain privacy-relevant insights into the life of the data subject.
- In collaborative filtering, the data of several users are combined, and profiles are created, which can be used to derive additional information about a data subject. For instance, knowledge about a data subject can be transferred to the other data subjects in the same cluster with a certain probability.
- A recommender system can also deliberately influence users by making one-sided recommendations.

2.7. DNA Sequence Classification

The analysis of human DNA samples has many useful applications. While in the early days, DNA analysis was mainly used in forensics to uniquely identify individuals—via the so-called *genetic fingerprint*—other areas of application have emerged over time as analytical methods have become more advanced. For instance, it can be used to establish paternity in custody and child support litigation or to trace ancestry over hundreds of thousands of years to study population genetics. Furthermore, DNA analysis can also be used in the medical domain to diagnose inherited disorders and human diseases [101].

Epigenetic modifications, such as DNA methylation, can be used to detect common diseases. For instance, aberrant DNA methylation of imprinted loci can often be observed in connection with certain types of cancer. In addition, autoimmune diseases, metabolic disorders, and neurological disorders are closely related to DNA methylation. Medicine could therefore make significant progress if a deeper understanding of epigenetic mechanisms leading to these diseases is available [102]. Although insights into the mechanisms of DNA demethylation have already been gained, there are still many correlations that have been observed but not yet systematically studied. To this end, however, extensive analyses are required to comprehend all relevant factors [103].

Instead of striving to explain the complex correlations, a reliable classification is already sufficient for a medical diagnosis. For instance, normal and malignant tissue samples from human liver and lung can be studied in order to be able to classify future tissue samples into "normal" and "malignant" based on these findings [104]. Instead of an analysis by a medical expert, artificial intelligence techniques can be used for DNA sequence classification. For this purpose, a large number of labeled samples are collected and used to train a model. This model recognizes the relevant genome sequences that are indicative of a particular disease [105].

Figure 7 is a simplified illustration of how this can be completed. First, suitable target data are selected. Preprocessing steps are applied to these data to structure them and prune errors or missing values. Then, the features relevant for the analysis are extracted. In the given example, sections of the DNA helix of the samples, i.e., sequences of base pairs. With these data, a *convolutional neural network* (*CNN*) is trained. In a CNN, there are multiple hidden layers in addition to an input and output layer. During the training phase, weights are allotted to describe how much certain nodes from one layer contribute to the inputs of the subsequent layer (and thus, eventually, to the final outcome). This allows one to classify new unlabeled data. The advantage of a CNN is that pooling allows one to discard redundant information early on, whereby even large amounts of data can be processed efficiently. This makes it also more robust against overfitting, i.e., it is possible to detect variants of a disease. Furthermore, a CNN can subsequently adjust itself to respond to shifts in circumstances [106].

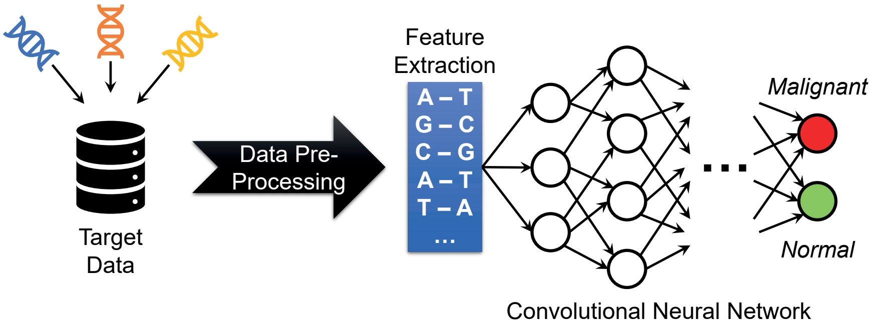

Figure 7. Application of a Convolutional Neural Network for the Classification of DNA Samples.

However, the training of such a CNN requires the establishment of large repositories of annotated genomic data. Although this may represent a key element for future discoveries in human disease, it inevitably also raises a variety of privacy concerns. Even if the metadata of the samples are fully anonymized, the DNA itself represents a unique fingerprint. All knowledge gained from a sample—e.g., ancestry, parenthood, pre-existing conditions, and genetic disorders—can therefore be uniquely traced back to a data subject [107]. Another inherent issue with CNN is its black-box approach. The reason for a certain decision cannot be explained even by domain experts [108]. The layers of a CNN consist of hundreds of millions of parameters that are totally detached from the real-world problem that is modeled by the CNN. Thus, it is not possible to trace which data have influenced the model in what way and therefore were decisive for a certain classification [109].

The inherent opportunities and privacy threats of IoT-supported DNA sequence classification can be broken down as follows:

Opportunities:

- 💡 IoT-supported DNA sequence classification enables comprehensive automatic detection of diseases, for instance.
- 💡 By using CNN, automatic adaptations to data shifts are facilitated.
- 💡 A CNN can learn novel correlations in the DNA structures.

Privacy Threats:

- ⚠ Training a CNN requires a very large DNA pool (i.e., highly sensitive data). DNA is a unique fingerprint, which means that the collected samples can always be linked to a person.
- ⚠ Through the DNA analysis as well as the comparison with other samples, additional correlations can be identified (e.g., relatives or hereditary diseases), which reveal a lot of private information.
- ⚠ The CNN itself or the decisions made by it cannot be explained. Decision making is therefore entirely based on full and blind trust in the CNN.

2.8. Synopsis

The seven application scenarios discussed above serve only to illustrate the general data processing requirements and the privacy or confidentiality concerns involved in big data analytics. The scenarios are representative examples of the most relevant kinds of data and processing types required in today's smart services. In Table 1, the main insights in this regard are summarized. The following section discusses in more detail how to address each of the identified privacy and confidentiality concerns by technical means.

Table 1. Synopsis of the Main Findings Regarding the Processing of Data in Smart Services.

Application Scenario	**Required Data Processing**	**Privacy or Confidentiality Concerns**
Location-Based Services	In addition to discrete location information, movement trajectories must be analyzed.	A lot of knowledge can be derived from frequent whereabouts, e.g., place of residence, workplace, interests, and even social contacts.
Health Services	In addition to individual measured values, in particular, temporal progressions of health data must be analyzed.	Health data are particularly sensitive as they reveal not only information about the health condition but also about the lifestyle.
Voice-Controlled Digital Assistants	The recordings must be analyzed to interpret the verbal commands.	The continuous recording enables exhaustive spying on users.
Image Analysis	The contents of the images must be analyzed in order to identify the shown objects.	By identifying the portrayed individuals, it is possible to reconstruct who was where and when with whom; even bystanders can be exposed.

Table 1. *Cont.*

Application Scenario	Required Data Processing	Privacy or Confidentiality Concerns
Food Analysis	Patterns indicating, e.g., allergens must be detected in food samples.	Other patterns reveal secret ingredients, thereby disclosing trade secrets.
Recommender Systems	Large amounts of data from many individuals must be analyzed to make appropriate recommendations.	Although the trained models do not disclose information about individuals, the underlying data do.
DNA Sequence Classification	Neural networks have to be trained based on a comprehensive DNA database to detect new correlations.	DNA data contain sensitive information; hence, third parties must not have full access to the complete dataset.

3. State-of-the-Art Privacy Measures

In this section, we study which technical measures can be used to conceal information in order to comply with privacy requirements. In this context, it is important to reduce the information content in line with processing requirements, i.e., the knowledge required by a data consumer must still be derivable from the data in an adequate quality. If this is not feasible, the technical measures are not applicable in practice—this would be equivalent to withholding the data.

In this context, it is first of all necessary to understand which data must be protected at all in order to be able to apply the privacy measures in a target-oriented manner and to minimize the impact on data quality. The GDPR stipulates that only *personal data* have to be protected. These are data that can be unambiguously traced to a date subject (Article 4(1)). Without such a linkage, no special privacy measures are required. However, most smart services across all domains require an authentication of their users. For instance, it must be possible to link measured values to the correct user, or it must be ensured that a user is authorized to execute certain commands. As part of this authentication, users are identified. This can be avoided by outsourcing the authentication to a trusted third party. This results in a segregation of the identification data and the payload data. The trusted third party only forwards pseudonymized subsets of the identifying attributes to the smart service provider [110]. As a result, the smart service provider cannot directly link the data to a natural person but only to a pseudonymous entity. Yet, if one has access to both the identification data and the payload data, these data can be joined again via the pseudonymized references. From the perspective of the GDPR, these are therefore identifiable entities, i.e., the personal data can be indirectly linked to a natural person. Such data must be protected in the same way as data which can be directly linked to a natural person. Thus, privacy measures are required by law for almost all smart services. However, even with (apparently) fully anonymized data, there is a risk to the privacy of users, which is why additional privacy measures should be taken in any case [111].

Generally speaking, there are three approaches to share data in a privacy-friendly way. First, the base data can be minimized by removing entire data items based on certain properties prior to processing [112]. For instance, in the health care sector, a dataset of patient records can be reduced by removing all medical records of a certain patient in order to preserve the privacy of that patient. In formalized terms, this is the application of a *selection operator* σ known from relational algebra. It is defined as:

$$\sigma_{\varphi}(R) = \{t : t \in R, \varphi(t)\} \tag{1}$$

Let R be a relation and φ be a propositional formula describing which tuples are to be included in the result set, i.e., the tuples t_i for which $\varphi(t_i)$ evaluates to TRUE.

Second, certain attributes of the data items can be excluded from processing [113]. In our example, e.g., the age can be concealed in each patient record. In formalized terms, this is the application of a *projection operator* Π known from relational algebra. It is defined as:

$$\Pi_{a_1,\ldots,a_n}(R) = \{t[a_1,\ldots,a_n] : t \in R\} \tag{2}$$

Let R again be a relation and $a_1,\ldots,a_x$ be the attributes in it. Then, let $t[a_1,\ldots,a_n]$ be the restriction of the tuple t to the first n attributes.

Third, data can be condensed prior to processing [114]. In our example, e.g., not every medical record is shared, but only aggregations of selected values. For instance, the records can first be grouped by diseases, and then, the average age of the patients for each disease can be shared. In formalized terms, this is the application of an *aggregate operator* $\mathcal{G}$ known from relational algebra. It is written as follows:

$$_{G_1,G_2,\ldots,G_m}\mathcal{G}_{f_1(A_1),f_2(A_2),\ldots,f_n(A_n)}(R) \tag{3}$$

Let R again be a relation. Then, let $G_1, G_2, \ldots, G_m$ be the attributes in R to group by, while each f_i is an aggregate function applied to the attribute A_i of the relation schema, e.g., SUM, COUNT, AVERAGE, MAXIMUM, or MINIMUM.

Figure 8 graphically illustrates the functional principle of these three operators.

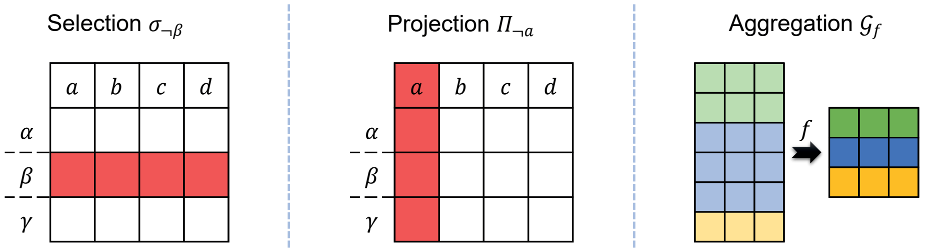
Selection $\sigma_{\neg\beta}$
a
b
c
d
α
β
γ
Projection $\Pi_{\neg a}$
a
b
c
d
α
β
γ
Aggregation $\mathcal{G}_f$
f

Figure 8. Graphical Illustration of the Effect of a Selection, Projection, and Aggregation on a Dataset.

The undeniable advantage of these three generic filtering techniques is that they can be applied to any type of data to preserve privacy. However, they are highly restrictive. Since entire data items or attributes are excluded from the processing, the amount of data—and thus the effective data quality—is greatly impaired. Yet, today's data refinement methods for deriving knowledge require large amounts of data to be effective. Therefore, perturbing the data typically leads to better results without compromising privacy. The goal is to tamper with specific aspects of the data only while ensuring that the data as a whole remain useful for data consumers [115]. However, there is no one-size-fits-all approach in this regard. Rather, specialized methods are needed that are tailored to the respective type of processing as well as the type of data involved. Only then can sensitive aspects of the data be concealed in a fine-grained manner without limiting the overall data quality [116].

In the following, we therefore discuss four filtering techniques in Sections 3.1 to 3.4 that are tailored to specific types of data: namely, location data, time series data, audio data, and image data. Section 3.5 then deals with a privacy approach that reorders the data in order to conceal certain patterns. In Section 3.6, statistical methods for privacy protection are addressed. Finally, Section 3.7 outlines how distributed processing can contribute to privacy. Due to the large number of research activities in the field of privacy, only a few representatives can be covered in each category. These seven privacy strategies are mapped to the application scenarios introduced in the corresponding subsections of Section 2. For each privacy strategies, we give a summary of its key strengths and weaknesses at the end of each subsection. To recap our review of the related work, Section 3.8 highlights the key findings.

3.1. Location Privacy

When it comes to protecting location data, two types of use cases must be distinguished: the protection of isolated snapshots and the protection of continuously captured trajectories [117]. The most straightforward approach to conceal the location is to mask it with a random location or to add an arbitrary offset to the actual location [118]. In this approach, however, one of these fake locations might be implausible, e.g., a location in the middle of the ocean. To avoid this and still keep the genuine location private, fixed predefined locations can be used instead of random generated ones, for which it has been ensured that they appear plausible in the context of the data subject [119]. Other approaches combine these two strategies by generating dummy locations following specific rules. In this way, realistic dummy locations can be obtained [120]. It is also possible to hide the real location among many fake ones. For this, a large number of generated dummy locations are shared in addition to the real one. This way, third parties cannot tell which of the location data actually belong to the respective data subject. When using an LBS, it has only to be ensured that all responses for the dummy locations are filtered out [121].

As shown in Section 2.1, such single fake locations would be cleansed during the data refinement process if contiguous trajectories are analyzed. Therefore, other measures must be taken when dealing with trajectories. *Spatial cloaking* takes advantage of the fact that continuous trajectories are composed of discrete locations. A circular cloaking area is set at each of these vertices. Instead of sharing the actual locations, the corresponding cloaking areas are shared. This way, third parties only obtain a rough idea of the path a data subject has traversed. The radius of the cloaking areas determines how accurate a localization can be [122]. The *path confusion* pursues a similar strategy as the dummy locations. Here, however, entire dummy trajectories are generated. In simple terms, phantom paths of non-existent data subjects are made available for data processing among which the actual user trajectories are hidden [123]. Sometimes, the trajectories do not contain any confidential knowledge but rather their temporal correlations. If a data subject travels from A to C via B, this might not be meaningful initially. However, if the temporal relationship is considered, it can be seen whether B was merely passed or whether the data subject stayed there for a longer period of time. Likewise, the temporal component of the trajectories can be used to determine whether two data subjects frequently travel together. To conceal such private information, *temporal cloaking* can be used. Here, the time stamps of the vertices are altered. Thereby, the traveled path of a data subject can still be traced but not when the data subject was there or for how long [124].

These three approaches to conceal trajectories are shown in Figure 9. The examples used to illustrate the privacy techniques in the figure are aligned with Section 2.1. Furthermore, combinations of these approaches are also feasible, e.g., *spatio-temporal cloaking*, in which both the spatial and temporal dimensions of the vertices of a trajectory are perturbed [125]. In addition, there are many other approaches dedicated to a specific task in the context of location-based services. For instance, there are approaches that address the privacy proximity test problem, i.e., how to determine whether an instance is in spatial proximity to another instance (e.g., two users or a user and a POI) without revealing the exact location of either instance [126].

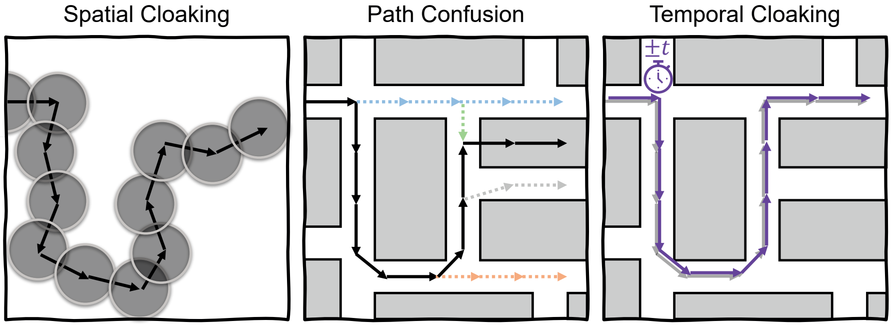

Figure 9. Visualization of Spatial Cloaking, Path Confusion, and Temporal Cloaking.

The key strengths and weaknesses of location privacy approaches are the following:

Strengths:

- ⊞ There are special techniques that allow to conceal single locations as well as whole trajectories or temporal sequences of trajectories.
- ⊞ The data quality of the other aspects can be largely maintained.

Weaknesses:

- ⊟ The techniques are subject to many restrictions regarding the credibility of certain location information or trajectories (e.g., a person will most likely not be in the middle of the ocean), which limits their scope of action.
- ⊟ Due to the sensor technology available in smart devices, more data sources are available to draw conclusions about location. This makes it easy to debunk a dummy location or a dummy trajectory.

3.2. Privacy-Preserving Time-Series Data

As outlined in Section 2.2, the processing of time-series data is always about identifying temporal patterns and forecasting trends. Privacy-preserving measures for these data must therefore not prevent such processing techniques. Rather, the aim is to conceal particular values or short sections of the time series that contain a particularly high level of sensitive information. This can be either rule-based (e.g., all values above or below a certain threshold) or time-based (e.g., all data recorded within a certain time window). There are basically two strategies to achieve this. On the one hand, the amount of data can be reduced in order to reveal less knowledge. On the other hand, the amount of data can be amplified by additional fake data in order to hide the actual values [127].

Figure 10 shows these two opposing strategies. To achieve data reduction, single values can be deleted from the series if only a few and widely distributed data points are concerned. The resulting gaps can be filled by *interpolation* so that the progression is still coherent [128]. To close gaps in time series where interpolation reaches its limits, e.g., due to the complexity of the progression or as the gaps are too large, machine learning techniques can be used to logically fill the missing parts [129].

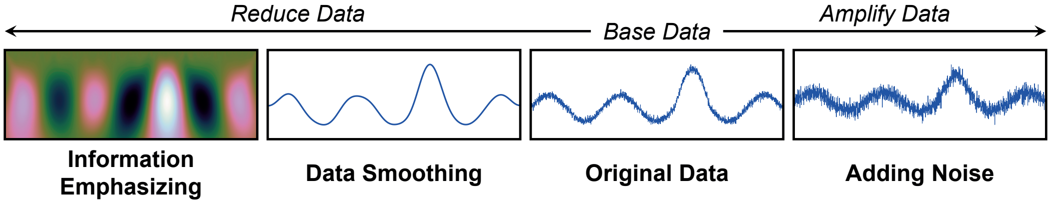

Figure 10. Strategies for Protecting Private Information in Time-Series Data.

If details in general shall be removed from a time series, *data-smoothing* approaches can be used. Originally, these approaches were primarily developed to compress data by removing less relevant parts. This removes noise at the same time. In the context of time-series data, the *discrete cosine transform* is used for this purpose in particular. An input signal is transformed into a finite sum of weighted trigonometric functions with different frequencies, representing a close approximation of the original data. As a result, the time series is smoothened [130]. With regard to privacy, however, this has another advantage, as all details of the single measuring points are wiped, and only the temporal progression remains. By using a *continuous wavelet transform*, the time-series data can be compressed even further [131]. *Information emphasizing* is thereby achieved, i.e., only a chronological sequence of high points and low points is recognizable.

A contrary strategy to protect privacy is *adding noise* to the time series. For instance, an artificial *Gaussian noise* can be generated, which is used to blanket the time-series data.

The parameters of the Gaussian random variable can be used to control how much impact the noise has on the time-series data. In this way, not only each discrete data point can be distorted but also, in the case of particularly strong noise, the progression of the time series itself [132]. While this very simple approach sounds promising at first, it has a significant drawback, as noise filtering is a de facto standard in the data refinement of time-series data. With the help of artificial neural networks, noise can be reliably eliminated on the fly [133]. One reason why such noise can be easily detected and even removed as an anomaly is that its frequency is completely different from that of the time series. Using a sequence of high-pass filters and low-pass filters, a time series can be decomposed into individual frequency bands. The bands affected by noise can be removed, and the remaining bands can be composed to the denoised time series [134]. However, there are also algorithms such as *SNIL* (*spread noise to intermediate wavelet levels*), which distribute the noise over all frequency bands of a time series and are therefore robust against such denoising [135].

The key strengths and weaknesses of privacy approaches for time-series data are the following:

Strengths:

- ⊞ Privacy techniques can be used to conceal both individual data points as well as data histories in time-series data.
- ⊞ The data quality of certain aspects (e.g., temporal trends or relevant data points) can be maintained.

Weaknesses:

- ⊟ When applying the privacy techniques, there must be knowledge about the intended use of the data. An incorrect privacy filter would completely destroy the utility of the data.
- ⊟ For some data protection techniques, the applications that process the data must be adapted accordingly. Information emphasizing, for instance, provides only maximums and minimums instead of a continuous data stream.

3.3. Voice Privacy

In the context of VDA, as presented in Section 2.3, a variety of privacy concerns arise. First, an imposter could operate the VDA. Since the VDA client is linked to a specific user, all actions of the imposter would be associated with that user. Second, the VDA is able to spy on the user continuously without the user's knowledge. The VDA client can secretly capture any spoken word and sound even if the specified keyword has not been retrieved and forward all recordings to its backend. Third, even if the user has said the keyword, from a privacy point of view the recordings should not always be forwarded to the VDA backend, since unaware bystanders could be heard in them or sensitive knowledge about the user could be derived from his or her voice. Due to the large number of threats, there is not a single privacy approach but rather a framework of protective measures [136]. A pipeline of privacy filters to protect against these threats is shown in Figure 11.

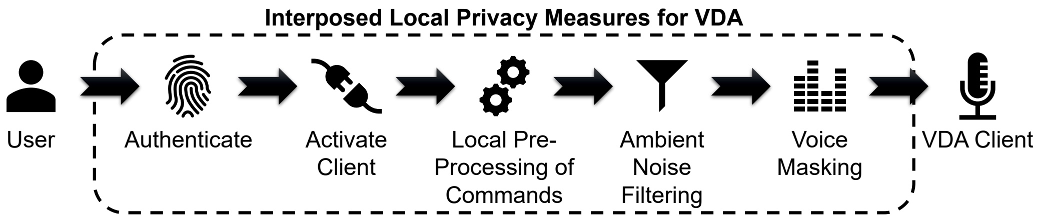

Figure 11. A Pipeline to Ensure Voice Privacy when Dealing with VDA.

A first protective measure consists of an authentication mechanism that uniquely identifies the proper user. In order not to disrupt the immersion when operating the VDA, such authentication can also be accomplished in a voice-based manner. Via characteristic features in their voice, users can be identified quite effectively using a smart speaker [137]. In order to ensure that the voice is not a recording, the environment can be scanned for suspicious magnetic fields, which would be generated by a loudspeaker. This can be achieved by a common smartphone via its built-in magnetometer [138].

Even if the speaker is an authorized user, it must be ensured that the VDA client only becomes active when it has received its keyword. To this end, an obfuscation signal can be generated in the frequency range of human speech. This ensures that the VDA client is permanently busy analyzing this signal whether it contains the keyword. This is the equivalent to a *denial-of-service attack*. Only when the user has said the keyword, the obfuscation signal is deactivated and the VDA client can accept new inputs [139].

Instead of forwarding all recordings to the VDA backend by default, it can first be attempted to handle the commands locally. This can be completed in the edge (e.g., a computer at the user's site, which serves as a hub for all of his or her VDA clients) [140] as well as on the VDA clients themselves [141]. When doing so, ambient noise can also be filtered out of the recording, which may reveal privacy-sensitive information [142]. For instance, if pets or technical devices can be heard in the background, then it can be assumed that they belong to the user. Additionally, conversations of bystanders in the background can also be obfuscated or be filtered out altogether [143]. To determine who is a bystander and who is an active user in a recording, the authentication mechanisms mentioned above can be used.

If the recording has to be sent to the backend for processing (e.g., due to insufficient on-site processing power or if third-party cloud services are required), the privacy of the active user can still be protected. From the voice, a lot of sensitive information can be inferred, e.g., gender, age, intoxication, mood, physical or mental disorders, just to name a few [144]. By voice masking, the user's voice can be distorted in such a way that these aspects are concealed, yet the speech can still be processed by a machine efficiently [145].

The key strengths and weaknesses of voice privacy approaches are the following:

Strengths:

- ⊞ There is a large range of voice privacy approaches, which can also be combined according to privacy requirements.
- ⊞ The voice privacy approaches take different privacy aspects into account, e.g., the protection of unknowing bystanders.

Weaknesses:

- ⊟ The techniques partly require additional hardware or adaptations to the installed hardware.
- ⊟ In some cases, the techniques only relocate the analysis of the data. That is, the sensitive knowledge is merely transferred to another—possibly more trustworthy—provider.

3.4. Image Privacy

As described in Section 2.4, image analysis can be used to identify people in photos. This raises a lot of privacy concerns. In order to make privacy-sensitive parts of an image obliterated, it seems to be a straightforward approach to blur these parts extensively. However, this has a negative effect on the quality of the image (and thus on its processability), especially if the blurring affects large parts of the image. Yet, if only small areas are blurred, this obfuscation can be undone in the data refinement process [146]. Even a full redaction of individual objects in the image does not reliably protect privacy. For instance, faces (i.e., exactly the kind of objects that are particularly privacy-sensitive) can still be recognized by means of artificial intelligence [147]. Special obfuscation techniques are therefore needed to conceal the privacy-sensitive areas of an image in a manner that is robust against reconstitution. To this end, *Singular Value Decomposition* is used to create a mask for the areas in question. This mask not only blurs the area but also adds features of other pictures to that area. This way, the result has similarities to several other images. Therefore, the blurred object can no longer be uniquely identified [148].

However, there are two problems from a data subject point of view. First, the manual selection of objects on an image that need to be protected is very cumbersome and time consuming. Second, guidance is needed in selecting the appropriate obfuscation method. For instance, if images shall be shared on social media, the obfuscation techniques must have as little impact on the visual quality as possible. Figure 12 shows what a computer-aided method for solving these two problems looks like.

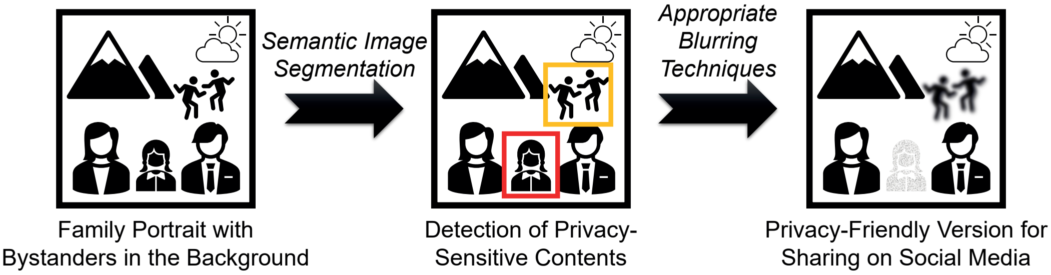
Semantic Image Segmentation
Appropriate Blurring Techniques
Family Portrait with Bystanders in the Background
Detection of Privacy-Sensitive Contents
Privacy-Friendly Version for Sharing on Social Media

Figure 12. A Process for Making Privacy-Friendly Versions of Portraits with Sensitive Content.

Initially, the original image is analyzed using deep learning. In this process, the deep features are extracted, i.e., the components that are significant for the recognition of the depicted objects in the underlying model. Based on these deep features, the image is then partitioned into a set of semantic object regions. These regions describe connected areas on the image, e.g., a person, an animal, or an object, while background objects are ignored. Using a classification, all of these regions are then assigned to a semantic meaning, e.g., bystander or child. For these classes, a privacy sensitivity is determined, which indicates how much the respective object needs to be protected [149]. In our example, bystanders receive the protection level orange, while children receive the protection level red. For persons, the semantic region can also be restricted to the face since this usually has the highest privacy sensitivity. To this end, face detection can be used to determine the region to be masked. Depending on the level of protection required, different techniques can be used that affect the visual quality of the image to a greater or lesser extent, e.g., blanking, scrambling, or blurring [150]. This way, a privacy-friendly version of the image suitable for social media can be generated.

The key strengths and weaknesses of image privacy approaches are the following:

Strengths:

➕ Sensitive content can be concealed specifically and according to individual privacy requirements.

➕ The data quality of the main components of an image is fully preserved by the image privacy approaches.

Weaknesses:

➖ Privacy is a highly personal experience. In image privacy approaches, however, the owner or provider of the image decides which privacy requirements apply to the persons visible in an image.

➖ Deep learning is used for the initial image analysis. This means, however, that the original, unaltered image is thoroughly analyzed, and knowledge is generated. This means that much more sensitive knowledge is generated than would otherwise be the case.

3.5. Pattern-Based Privacy

In Section 2.5, we presented that large amounts of data can be processed effectively using CEP. Here, a sequence of isolated data items is interpreted as a stream of events. A CEP engine scans the stream for predefined patterns. As the occurrence of such patterns can be confidential information, there are also several approaches for CEP to protect privacy. For instance, the data stream can be preprocessed locally at the user's site, and the user decides for each event whether to answer truthfully or incorrectly, i.e., whether to feed an event untampered into the stream or to add noise to it. Furthermore, the fact that in the data refinement process, the data are aggregated anyway when converted to knowledge can be exploited. By means of dedicated aggregation techniques applied directly at the user's site, a *zero-knowledge privacy guarantee* (see the work by Gehrke et al. [151]) can be achieved [152].

In doing so, it is disregarded, however, that an isolated event typically entails only little privacy-sensitive information. The information patterns targeted by CEP engines are only revealed by the occurrence of several events in a certain sequence. An arbitrary manipulation of single events is therefore not efficient and deteriorates the quality of the base data unnecessarily. Therefore, other approaches focus in particular on sequences of events. For this purpose, patterns are defined that have to be kept secret. These patterns are then purged from the data stream [153].

Although this approach is considerably less restrictive, it still overly compromises data quality [154]. The problem with this approach is that it only takes into account what has to be concealed but not what knowledge is needed by data consumers. Therefore, two types of patterns are needed to preserve data quality: *private patterns*, which have to be concealed, and *public patterns*, which are required by data consumers. Data producers and data consumers specify which knowledge must not be disclosed or respectively which knowledge is required. This is then mapped to patterns at the data level [155]. When concealing the private patterns, it has to be ensured that no public patterns are blurred in the process or additional erroneous public patterns are created by manipulating the data stream. To this end, a quality metric is used that weights the privacy (i.e., the concealment of private patterns) against the *false negatives* (i.e., public patterns that have been concealed) and *false positives* (i.e., public patterns that have been artificially added). The obfuscation of the private pattern must be conducted in a manner that optimizes this quality metric [156].

To maximize the quality metric, a privacy mechanism has to make use of different obfuscating techniques. Three such techniques are shown in Figure 13. In this example, the private pattern $B \rightarrow C$ has to be concealed, while the public pattern $A \rightarrow B$ should be recognizable. Let $A \rightarrow B \rightarrow C$ be the data stream. A simple technique is to drop an event from the stream or to inject a new one. In the example, the event C is removed, and the event D is artificially created. This conceals the private pattern, while the public pattern remains recognizable. Removing event B would conceal the private pattern as well, but the public pattern would also not be recognizable anymore. Another technique is to tamper with the events. For instance, the attribute values of data item C can be manipulated so that it is recognized as a different event C'. This also conceals the private pattern without affecting the public pattern. Lastly, the events in the stream can be reordered. By swapping B and C, the private pattern is also concealed—albeit, in this case, the public pattern is concealed as well. Using a combination of these techniques, even more complex patterns that contain conjunctions and negations in addition to sequences can be obfuscated [157].

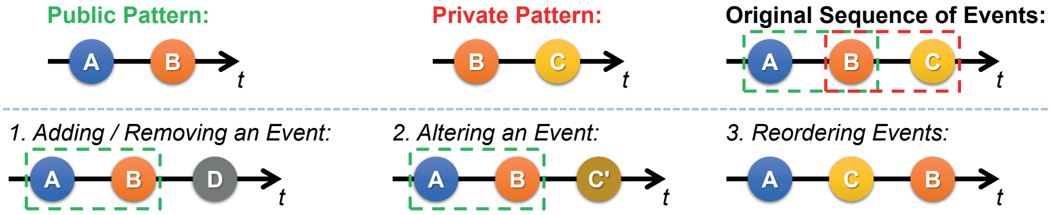

Figure 13. Application of Obfuscation Techniques to Conceal Private Patterns in a Data Stream.

The key strengths and weaknesses of pattern-based privacy approaches are the following:

Strengths:

- ⊞ Pattern-based privacy does not degrade the data quality of the measurement data.
- ⊞ Due to the public and private patterns, sensitive information can be filtered out in a target-oriented manner.

Weaknesses:

- ⊟ The computation of an optimal configuration, i.e., the maximization of the quality metric, is very complex.
- ⊟ A pattern-based privacy approach requires full control over incoming and outgoing data streams of a data processing system in order to effectively apply the required obfuscation techniques.

3.6. Differential Privacy

The previously discussed privacy techniques focused on data from individual users, which are distorted before processing. When transforming the data to knowledge, the base data of multiple users are often merged. This can be seen, e.g., in the recommender systems presented in Section 2.6. Here, the models that constitute the basis for the recommendations are trained on the preferences of all users. To this end, Dwork [158] introduces *differential privacy* as a measure for assessing the privacy threat to an individual when he or she shares his or her data. The goal is to maximize the accuracy of statistical information about a data collection (e.g., the models of the recommender systems) while minimizing the potential privacy risks for individuals. This can be illustrated by a simple example: In collaborative filtering, let there be a cluster consisting of two persons. Even if the interests of the individual users cannot be derived directly from the trained models, person A can be sure that if a new item is recommended to him or her, person B is interested in this very item. Thus, the privacy of person B has still been violated.

The basic idea of differential privacy is straightforward, as shown in Figure 14. Database 1 contains data about the four users. If the red user wants to know whether her privacy is disclosed when a statistic is computed on these bases data using the function f (respectively, a model is trained on these data), a *neighboring database* Database 2 can be created, which contains all datasets of Database 1 except the data about the user in question. If the function f is applied to Database 2 and the results are similar, the privacy of the red user has not been exposed. So, formally speaking, the following must apply:

$$\mathbb{P}[f(\mathcal{D}_1) \in \mathcal{R}] \leq exp(\epsilon)\mathbb{P}[f(\mathcal{D}_2) \in \mathcal{R}] \tag{4}$$

A function f is *ϵ-differentially private* if its results $\mathcal{R}$ for all of database $\mathcal{D}_1$'s neighboring databases $\mathcal{D}_2$ only differ by an insignificant ϵ. Since there are different privacy requirements depending on the data and processing context, there are approaches that adjust the degree of applied obfuscation according to the level of exposure. That is, there is not a single ϵ that describes the privacy requirements but rather many such measures that are applied depending on the current situation [159].

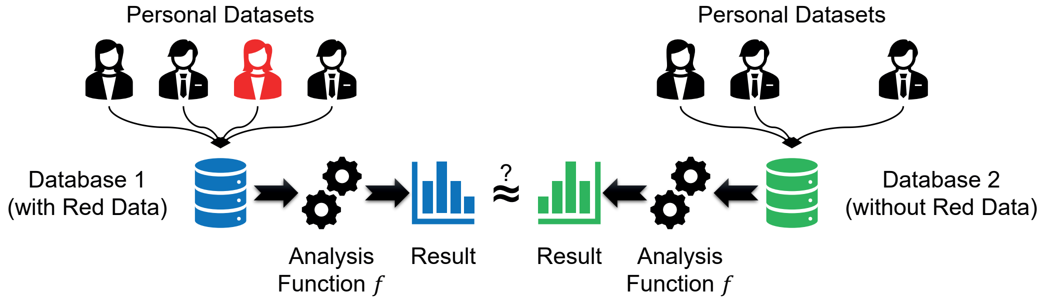

Figure 14. Simplified Representation of the Central Idea of Differential Privacy.

While this sounds promising in theory, it turns out that differential privacy is rarely applied in practice. One reason for this is that there is a lack of awareness of differential privacy principles in the development of algorithms, and in particular, the handling of complex types of data proves to be difficult [160]. In research, however, there is a variety of noise algorithms that establish the ϵ-differentially private properties without impairing the quality of the outcomes excessively [161]. Especially for application scenarios such as recommender systems, differential privacy is therefore suitable [162].

The key strengths and weaknesses of differential privacy approaches are the following:

Strengths:

⊞ Differential privacy approaches allow statistical analysis while preserving the privacy of each individual involved.

⊞ In order to guarantee the differential privacy property, the method is not restricted to any particular technique, which means that an appropriate obfuscation technique can be chosen depending on the base data.

Weaknesses:

⊟ Differential privacy approaches can only be applied when large amounts of data from many different individuals are analyzed.

⊟ Ensuring the differential privacy property is difficult depending on the base data requires the use of destructive noise algorithms. As a result, potentially relevant aspects in the data are lost.

3.7. Federated Learning

Finally, we look at *federated learning*, which is an approach that was not originally intended to provide privacy. Along with the increased application of machine learning in all kinds of domains, the need for data to train the corresponding models has also increased. While the IoT provides an almost unlimited number of smart devices as data sources, organizational problems arise in terms of the communication overhead required to continuously receive the captured data as well as the central computing resources required to process the data. As IoT-enabled devices become increasingly powerful, the federated learning approach shifts many of the data refinement tasks from the central data processor to the distributed data producers. This includes the entire data preparation phase and large parts of the data processing phase. That is, data producers generate information patterns which are gathered by the data processor. The data processor uses these patterns to create or update global knowledge models that contain the information patterns of all data producers. In other words, federated learning takes aggregation to the next level as data producers exchange only knowledge instead of data with the data processor. From the global models of the data processor, knowledge can also be fed back to the local workers in order to improve their local models [163]. This procedural concept is illustrated in Figure 15.

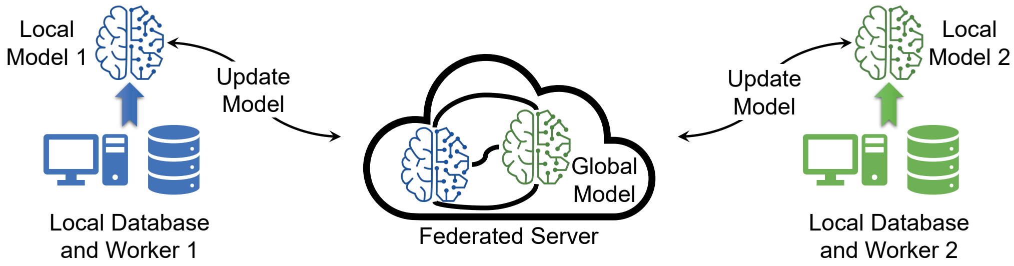

Figure 15. A Distributed Training of a Global Knowledge Model via Federated Learning.

However, federated learning can also contribute to privacy-preserving machine learning. Data privacy is generally regarded as the right of an individual to freely decide how much knowledge he or she wants to disclose about him or herself. This is exactly what federated learning does, as data producers are free to decide which of their data they want to use in the data refinement process and which information patterns they want to include in the knowledge model [164]. By applying privacy filters to the data, differential privacy can be preserved not only for the isolated local models but also for the shared global model. Thereby, a stronger quantifiable privacy protection can be achieved [165]. This protection combined with its distributed computing renders federated learning well-suited for large-scale analyses, such as the DNA sequence classifications presented in Section 2.7 [166].

The key strengths and weaknesses of federated learning approaches from a privacy perspective are the following:

Strengths:

- ⊞ Federated learning is primarily used to efficiently run complex machine learning processes. The preservation of privacy is a beneficial side effect that comes at no additional cost.
- ⊞ Federated learning enables data subjects to incorporate their data into global machine learning model but to carry out the necessary processing of their private data locally, i.e., under their full control.

Weaknesses:

- ⊟ Due to the complex and non-explanatory nature of the trained models, it is not possible for data subjects to understand what knowledge about them is incorporated into the global model by means of their locally computed models.
- ⊟ The use of federated learning is limited to certain algorithms and algorithm classes.

3.8. Key Findings

Our key findings regarding the means of privacy protection as well as the effects of the aforementioned privacy techniques are summarized in Table 2.

Table 2. Summary of the Key Lessons Learned from the Review of the Privacy Techniques.

Privacy Approach	Means of Privacy Protection	Effects of the Measures
General Privacy Measures	The three relational algebra operators—selection, projection, and aggregation—can be applied to base data.	Entire data items or certain attributes can be concealed, and the base data can be condensed.
Location Privacy	Fake locations can be used, and spatial cloaking, path confusion, and temporal cloaking can be applied.	Individual locations or entire trajectories as well as their temporal correlations can be concealed.
Privacy-Preserving Time-Series Data	The data can either be compressed to reduce details or they can be amplified by fake data.	Only temporal progressions can be observed but no details on single data points.
Voice Privacy	The VDA can be jammed, data are preprocessed locally, and the recordings are filtered.	A VDA cannot spy on its users, and the information shared with the VDA backend is minimized.
Image Privacy	Blanking, scrambling, or blurring can be used to mask certain areas of an image.	Objects on an image can be obfuscated in a fine-grained manner based on their privacy sensitivity.
Pattern-Based Privacy	Data items can be added, removed, altered, or reordered.	Private patterns in terms of data sequences can be concealed.
Differential Privacy	In statistical calculations, noise ensures ϵ-differential privacy.	No knowledge about single individuals is disclosed to third parties.
Federated Learning	Data processing is primarily performed locally by data producers.	Data processors only gain insight into highly aggregated knowledge.

It is evident that there are dedicated privacy filters for different types of data, namely for location data, time-series data, voice data, and image data. Furthermore, there are techniques to conceal private patterns in a stream of data items. Thereby, it is possible to filter sensitive information in a fine-grained manner. Using these and similar obfuscation techniques, the ϵ-differentially private properties can also be established, which ensures that statistical computations are performed on a plethora of personal datasets—including the extraction of information patterns that are the foundation of knowledge—without exposing the privacy of any individual.

However, these dedicated privacy techniques can only be applied systematically if a certain underlying structure is present in the data. For generic raw data, only general privacy measures can be applied. These include, e.g., the three operators of relational algebra mentioned initially: namely, selection, projection, and aggregation. The dedicated privacy filters are therefore significantly better suited for backend systems after comprehensive data refinement measures have been carried out, and processable information has already been retrieved. Meanwhile, on lightweight smart devices, which are more likely to handle raw data, the general privacy measures should be preferred as they are easier to apply. Yet, these general measures have to be used purposefully, as they are not capable of filtering private aspects specifically and thus tend to have a more significant impact on data quality.

Federated learning represents a completely different approach. Here, data preprocessing is carried out in a distributed manner under the supervision of the data subjects. Only highly aggregated information in the form of machine learning models is forwarded to the backend system. In this way, each data subject can decide independently which privacy measures are applied locally to the data. However, data subjects often lack the necessary knowledge to do this in a systematic manner that is tailored to the intended purpose of the models. Moreover, federated learning is only suited for very certain machine learning algorithms and not for general purpose data refinement.

In this context, it is arguable whether it is even necessary to apply privacy measures locally on the smart devices of the users. A local application of privacy filters is primarily reasonable if the data processor to which these smart devices forward the collected data is not trustworthy. If this is not the case, it is more advisable to apply privacy measures on the side of the data processor. Such a global approach can be more targeted and results in stronger privacy protection for the same scope of data filtering—i.e., the trade-off between privacy and utility is significantly better [167]. Data processors in the service domain are commonly assumed to be *semi-honest-but-curious*. That is, they largely carry out their tasks in a trustworthy manner, and they will not deliberately expose the privacy of a user [168]. It is therefore sufficient to provide data processors with means to express their privacy requirements and to verify whether they are respected. So, in this context, federated learning is rather an exception that can be used if highly sensitive data are involved.

Now, the question arises as to what extent these state-of-the-art privacy measures are suitable for the deployment in the application scenarios discussed in Section 2. In the following section, we therefore map the strengths and weaknesses of these approaches against the opportunities offered by and privacy threats posed by the application scenarios in order to answer this question.

4. Assessment of the State of Privacy Mechanisms for Smart Services

After discussing the opportunities and privacy threats of smart services in the most relevant application domains in Section 2 and identifying the strengths and weaknesses of state-of-the-art privacy technologies for these types of smart services in Section 3, we now assess whether the available privacy measures are adequate. To this end, we apply a systematic analysis technique adapted from strategic planning. The so-called *SWOT analysis* is originally used to determine the market position and strategy development of companies. SWOT stands for *strengths*, *weaknesses*, *opportunities*, and *threats*. In this process, internal and external factors are first identified and then juxtaposed with helpful and harmful aspects of a product or a company strategy. From the four resulting intersections (internal factors vs. helpful aspects, internal factors vs. harmful aspects, external factors vs. helpful aspects, and external factors vs. harmful aspects), strengths, weaknesses, opportunities, and threats can be derived. Therefore, SWOT analysis is an important foundation for strategic audits, as it enables a systematic market and environmental assessment [169]. In addition to the original focus on companies, the SWOT analysis is meanwhile used in adapted form in many other domains, for instance, in the education sector when new teaching methods are to be introduced or in the health sector to determine risk factors for patients [170].

For our assessment of the state of privacy mechanisms for smart services, we also use an approach based on the SWOT analysis. We develop an individual SWOT matrix for each of the seven application scenarios. Unlike in the classic SWOT analysis, however, we do not compare internal and external factors with helpful and harmful aspects but rather the strengths and weaknesses of the privacy mechanisms with the opportunities and privacy threats present in the respective application scenarios. In this way, we obtain a good understanding of the extent to which state-of-the-art privacy measures are suitable for the deployment in the application scenarios and which open questions still need to be resolved in this context. The result of our analysis is presented in Figure 16.

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	Since either individual locations or trajectories can be concealed, LBS can be provided with the required data.	Excessive spying can be prevented by particularly restrictive privacy techniques that, e.g., blur both locations and trajectories.
Weaknesses of Privacy Measures	As LBS are deceived by fake data, the user experience suffers due to incorrect information.	As LBS permanently track the location, the scope for privacy techniques is limited and private information can still be inferred.

(a) Location-Based Services

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	By using proper privacy techniques, certain aspects can be concealed without impairing the relevant health data.	Data subjects determine which aspects should be hidden, so that no compromising knowledge is revealed.
Weaknesses of Privacy Measures	Privacy techniques falsify certain aspects of the data, which can have grave consequences in the health context.	As data subjects are not aware of what can be inferred from the collected data, they cannot choose the appropriate privacy technique.

(b) Health Services

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	Voice-based control of smart devices is not restricted by the privacy measures.	By using additional hardware, a VDA can be fully controlled.
Weaknesses of Privacy Measures	The low entry hurdle of VDA is undermined by the need for additional hardware.	Compliance with the privacy requirements still has to be outsourced to third parties.

(c) Voice-Controlled Digital Assistants

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	OSNs reveal many private photos—using image privacy techniques, this can be done privacy-aware.	The privacy techniques can be used to blur specific objects in the images.
Weaknesses of Privacy Measures	The overall quality of the image suffers as a result of excessive blurring.	Third parties instead of the affected data subjects decide which objects should be blurred.

(d) Image Analysis

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	By defining public pattern, pattern-based privacy has negligible impact on the actual analysis.	By defining private pattern, no confidential information is disclosed.
Weaknesses of Privacy Measures	Since the configuration causes a lot of overhead, this affects the cost of an analysis.	Food producers must not have control over the data analysis, which is mandatory for pattern-based privacy.

(e) Food Analysis

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	Differential privacy does not limit the statistical analysis of recommender systems.	Since individuals are hidden in the mass of all users, the data analysis does not pose a privacy threat.
Weaknesses of Privacy Measures	Properties of extraordinary individuals have to be blurred, i.e., this knowledge is lost.	Through the use of destructive noise, it is possible to deliberately influence users by not recommending certain items.

(f) Recommender Systems

	Opportunities of Smart Services	Privacy Threats of Smart Services
Strengths of Privacy Measures	The quality of the global model (and thus of the analysis) is not impaired by federated learning.	Sensitive data is processed locally (and thus under the control of data subjects) and is only forwarded in aggregated form.
Weaknesses of Privacy Measures	Federated learning cannot be applied to any kind of analysis.	The locally computed models cannot be explained, i.e., data subjects do not know what knowledge about them is propagated.

(g) DNA Sequence Classification

Figure 16. Results of our SWOT Analysis Broken Down by the Respective Application Scenario.

Across all application scenarios, it can be observed that there is an appropriate privacy technique for each privacy threat. It can also be noted that the privacy techniques are capable of operating in a target-oriented manner. That is, when protecting sensitive data, it is also ensured that the general utility of the base data is not unnecessarily impaired. This is a fundamental requirement, as otherwise, the use of smart services would be severely disrupted or outright prevented, rendering the privacy techniques worthless.

So, in principle, this is a highly promising result, as there are tailored techniques for all types of data and forms of processing. However, our analysis also reveals inherent problems with state-of-the-art privacy technologies for smart services that need to be addressed in order to enable effective data protection by design. Problems arise in particular regarding the selection and configuration of the privacy techniques, since this requires comprehensive technical and domain knowledge, which a normal user does not possess. As a result, the privacy measures are often too restrictive in terms of impairing the quality of the smart service and at the same time not effective enough in terms of protecting all sensitive data. Moreover, each privacy technology is always designed for a specific use case. Comprehensive privacy protection therefore requires the integration of many heterogeneous approaches, which leads to further adjustment problems. Since many privacy approaches rely on a trusted third party to protect sensitive data, such an agglomeration of many different approaches leads to the situation that many third parties gain access to the private data (one party per approach). Instead of disclosing less private information, the data are therefore disclosed to even more parties. Finally, uncontrolled tampering with the data also leads to further security vulnerabilities, which can result in financial or material damage, which is why the choice of the appropriate privacy method must also not be left solely to the inexperienced user.

Based on these insights, we derive seven such open research questions and briefly outline how we believe these questions can be addressed in the following section.

5. Future Prospects

As our assessment of state-of-the-art privacy measures reveals, there is no one-size-fits-all solution when dealing with smart devices. Rather, a privacy approach has to be found which is fully geared toward the intended purpose. Only then is it possible to ensure an adequate data protection. While efficient island solutions for concealing specific sensitive information exist, there are open questions regarding the applicability of such privacy filters. In the following, we therefore discuss seven open research questions that emerge in the context of data protection in the information age.

(a) Privacy Requirements Elicitation. First, it is crucial to identify the privacy risks posed by a smart service, i.e., what sensitive knowledge is exposed. Only when these risks are identified explicitly is it possible for data subjects to give informed consent with regard to the processing of their data. Depending on the threat potential of a service, data access has to be adjusted accordingly.

The *System–Theoretic Process Analysis for Security* (*STPA-Sec*) is a top–down approach that can be used to systematically identify security problems in complex and dynamic systems. For this purpose, a holistic view of the system is obtained. For each component and the communication flow between components, vulnerabilities are annotated in order to identify all potentially insecure control actions. Based on the results of this analysis, the system can be redesigned by means of a security-driven design process [171]. This process can also be adapted to privacy aspects (*STPA-Priv*) [172].

In practice, however, a conceptual problem in STPA-Priv is encountered. Whereas both data producers and data consumers are equally interested in the security of a system, privacy is usually a one-sided interest on the part of data producers—the data consumers (e.g., smart services), meanwhile, are primarily interested in obtaining as much data as possible. The results of such a privacy analysis are therefore less useful for the redesign of a smart service. However, they can be used to determine privacy requirements in the context of a smart service [173].

In this regard, it is important to study how a holistic view of a smart service can be compiled. In the context of the IoT, there are many heterogeneous data sources that can be dynamically added or removed at any time. Furthermore, smart services often use a data backend as a data source. For a privacy analysis of a smart service, it is also necessary to assess its data backends and all their data sources. However, such a comprehensive view of all data sources of a smart service typically cannot be provided.

(b) Holistic Privacy Platform. Even if a user is fully aware of all privacy risks associated with a smart service, s/he also needs to be enabled to take effective countermeasures. The privacy measures discussed in Section 3 certainly contribute to protecting privacy, but they are no out-of-the-box solutions. Rather, they have to be applied at the right place in the data flow. To become a usable privacy measure, the filters must be integrated into a privacy platform. Via such a platform, users can apply the necessary privacy measures appropriately. The advantage of such a central privacy platform is that it can be constantly extended with new privacy filters. Meanwhile, there is already a wide range of privacy filters for some application areas, e.g., the area of behavioral data recorded by smart devices is still largely unexplored [174]. A privacy platform could be easily upgraded once efficient privacy filters for behavioral data have been developed. In this way, the protection of private data is constantly in line with the latest technological advances. There are research approaches for such platforms that, e.g., provide a selection of privacy filters that can subsequently be applied to a data stock [175]. Such a solution primarily addresses backend systems. However, there are also approaches that address the source systems that feed such a data backend, e.g., smart devices [176].

When dealing with smart services, it is important to determine how such a privacy platform can reliably monitor them and control their data access. While a privacy platform for a backend system can restrict access to that data backend and, e.g., apply privacy filters to certain data before they are shared, it cannot be ensured that a smart service does not use multiple data backends. The data obtained from each backend individually might not be very exposing. However, by combining all the gathered data, it might still be possible to derive sensitive knowledge patterns. Reliable privacy protection requires a holistic end-to-end approach from the data source to the data sink, i.e., the smart service. In particular, this means that a smart service must be completely isolated from any data source and can only obtain all data via the privacy platform.

(c) Configuration of Privacy Filters. A user is generally able to specify certain knowledge patterns that are particularly private or confidential at a high abstraction level. However, users are not familiar with the data sources from which these patterns can be derived let alone privacy filters via which they can be concealed efficiently. The configuration and parameterization of such a holistic privacy platform is therefore far too complex for users. Concepts are therefore required to enable them to configure the platform using rather high-level descriptions of their privacy requirements. This requires models that accurately represent which information can be obtained from which data sources and what knowledge can be derived from it. There are metamodels for this purpose to describe such correlation [177]. Furthermore, data subjects are often not at all aware of their privacy requirements. They intuitively have a rough idea but are often unable to fully express it. Yet, using approaches based on collaborative filtering, sensitive knowledge patterns can be recommended to them that may also be relevant to them [178].

Such an approach can be applied successfully for isolated applications, but for complex smart services and their data infrastructure, two major problems arise: On the one hand, in addition to a holistic view on the smart service, domain experts are needed who have the necessary experience to determine which knowledge can be derived from which data sources. Such experts are also needed, e.g., to analyze the smart services using STPA-Priv. Therefore, it is advisable to study whether the model of the data–information–knowledge relationships can be derived from the STPA-Priv results or to what extent this process analysis approach needs to be adapted for this purpose. On the other hand, a recommender system for privacy requirements depends on the assumption that other users have previously been able to formulate their requirements. This results in a chicken-and-egg problem. Since the GDPR mandates a *data protection by default* (Article 25), it should also be studied whether the privacy requirements identified by STPA-Priv can be translated into some sort of basic configuration for the privacy filters representing an advisable baseline for any user.

(d) Deployment of Privacy Filters. Once a configuration has been found, the privacy filters must be deployed appropriately. There are basically two options: the privacy filters are applied either directly to the data sources [179] or to the data backend [180]. However, this decision has a significant impact on privacy and data quality. If a privacy filter is applied very close to the source, i.e., on the user's side, the unfiltered raw data never leave the user's control. That is, the data are impurified before they are forwarded to the data backend. Yet, at this point, not all information about the further usage of the data is available; e.g., a data backend can also feed several smart services, for which different privacy requirements may apply. Therefore, privacy measures cannot be applied in a target-oriented manner, which means that they turn out to be either too restrictive or insufficient. Whereas, when applying privacy filters in the data backend, the data are no longer within the user's sphere of influence and the user must completely trust that his or her privacy requirements are respected by the backend, which represents a major psychological hurdle [181].

In simple terms, the privacy platform must therefore find a deployment plan for which a utility metric is maximized. At an abstract level, this metric looks like this:

$$Utility = Data\ Quality + Privacy \tag{5}$$

The utility of a deployment plan is defined by how well it preserves privacy and how little it impairs data quality. A high-level definition, analogous to the configuration based on knowledge patterns, could look like this:

$$Utility = \sum_i Public_i * w_{Public_i} - \sum_j False_j * w_{False_j} - \sum_k Private_k * w_{Private_k} \tag{6}$$

Here, the data quality is described by how many public knowledge patterns—i.e., non-confidential knowledge patterns—can be detected despite the use of privacy filters ($\sum_i Public_i$), minus all *false positives*, i.e., public knowledge patterns that were falsely recognized due to the use of privacy filters ($\sum_j False_j$). The privacy of a deployment plan is determined by how few private knowledge patterns it exposes ($\sum_k Private_k$). Additionally, a penalty weight w can be assigned to each of these components to prioritize them differently depending on the specific use case. The research question here is how to efficiently determine the one with the best utility from all possible deployment plans.

(e) Secure Data Management. Since the application of privacy filters can be time consuming, they should not be applied as data flow operators over and over for every data access. Instead, frequently used data should be stored in the data backend at different privacy levels, i.e., after different privacy filters have been applied. Zone-based data lakes are suitable for managing big data in different processing stages. In addition to filtered raw data, higher-value information, for instance in the form of machine learning models trained by means of federated learning, can also be stored in such data lakes [182]. For this, the data backend acts as a *data marketplace*. Similar to a marketplace for material goods, customers, e.g., smart services, can pick the data they want [183]. This requires extensive metadata that characterizes the available data so that smart services can also find relevant data [184]. In contrast to conventional marketplaces, however, privacy constraints must be observed in data marketplaces; i.e., not every customer may access every data item [185]. Moreover, since the existence of a data item exposes certain information, even the visibility of the items has to be regulated in this case. In this regard, it must be ensured that the operator of the data marketplace cannot abuse this. Since s/he has sovereignty over the data, s/he could maliciously withhold certain data items to the disadvantage of a data subject or a smart service. By using blockchain technologies this can be prevented, a trustworthy data sharing is enabled [186]. Yet, the use of a blockchain raises further privacy issues, e.g., due to its immutability or the fact that in a data marketplace, there are legitimate reasons why a certain party is not allowed to see a specific data item [187].

Future research therefore has to address how existing metadata models need to be extended to include aspects such as applied privacy measures in addition to data quality or data origin. Furthermore, access policies must be developed to ensure that certain data items are not visible to selected smart services. However, it is also important to ensure that these mechanisms cannot be misused to the disadvantage of data subjects or smart services. Unlike blockchain technologies, all concepts developed in this context must comply with current data protection regulations by design.

(f) Proof of Data Possession. Such a data marketplace must inevitably be able to provide proofs of data possession. It must be verifiable for data subjects whether such a data provider stores the data of the data subject in the agreed form, e.g., at different privacy levels. Although such a verification must be publicly available, it must not jeopardize data protection. That is, a third party must not be able to derive any information by verifying whether certain data about a data subject are retained [188]. To this end, *Proof of Retrievability and Reliability* (*PoRR*) approaches are applied in cloud-based data stores. These approaches verify that a cloud provider faithfully manages the entrusted data in the agreed number of replicas. For this purpose, a so-called *Verifiable Delay Function* (*VDF*) is applied to the data and all replicas. A VFD is slow to compute but easy to verify [189]. The cloud provider is regularly challenged with respect to this function. If the response to this challenge takes too long, it is confirmed that the provider does not have the data at rest but needs to compute the VFD on the fly. In cloud-based data stores, such a procedure can also be applied efficiently [190].

While at first sight there seem to be many similarities between data marketplaces and such cloud-based data stores, there are also decisive differences. In particular, the data marketplace does not manage identical replicas of the source data but rather several variants of them, which are available in different processing stages and to which different privacy filters have been applied. Furthermore, it is also possible that some raw data reside on the smart devices, and the data marketplace only acts as an intermediary between them and the smart services. Therefore, the data infrastructure here is distributed and heterogeneous. Furthermore, some components are computationally weak, namely the smart devices, which is why a mechanism to provide proof of data possession must be lightweight. It is therefore important to investigate to what extent a PoRR approach can be applied to such a data marketplace and what adaptations are necessary to this end.

(g) Prevention of Misinformation. Finally, the spread of misinformation is a major problem in the information age. In the context of a data hub such as a data marketplace, two types of misinformation must be distinguished: On the one hand, data subjects can deliberately manipulate their data using privacy filters in order to gain an advantage [191]. For instance, if they provide health data to a smart service of their health insurance company to obtain a better rate, it must be ensured that unhealthy habits, such as being a smoker, cannot be specifically filtered out. This can be restricted by means of attribute-based authentication of the data sources. Thereby, certain conditions can be specified—e.g., a certain privacy filter is not applied—which the sources must comply with in order to authenticate successfully. However, such information can also reveal a lot about the data. For instance, when certain privacy filters are applied, it can be inferred what kind of knowledge the data subject wants to conceal. Hence, the authentication process also needs to be privacy-friendly. For this purpose, a trusted intermediary can be used, which pre-validates the complete attributes of a source and forwards only those attributes that do not reveal sensitive information to the data recipient for authentication [192].

However, this approach is based on digital signatures. Due to the advent of quantum computers, asymmetric cryptography, which provides the foundation for digital signatures, can no longer be considered secure. Future research therefore needs to explore which post-quantum cryptography approaches can be used instead and to what extent they are suitable for the usage on smart devices, which have to generate the signatures. Furthermore, the question arises how an efficient and trustworthy key management can be implemented in such a scenario. This concerns both the key generation and the key provisioning.

On the other hand, misinformation about a data subject can also be disseminated by third parties. This is well known in the context of OSNs such as Facebook and commonly referred to as *fake news*. Here, as well, information is disseminated by means of word-of-mouth and not directly from a data subject to the intended recipient. For OSNs, there are approaches that can be used to restrain the dissemination of misinformation. Contradictory information, i.e., misinformation and credible information, is identified, and its propagation in the network is monitored [193]. Using a greedy approach, users in the network can be identified who are considered trustworthy and can therefore be assumed to contribute to minimize the spread of misinformation by sharing only credible information in the network. In the case of contradictory information, this allows one to determine which version represents the misinformation—namely, the one which is not shared by a trustworthy user—and then to remove it from circulation [194].

However, this procedure was developed specifically for social networks. It must therefore be studied whether it can also be applied to the data infrastructure of smart services. In particular, it must be assessed to what extent this approach has to be adapted and extended in order to be able to deal with heterogeneous raw data from a wide variety of domains instead of purely textual information. Furthermore, the IoT is a much more dynamic structure than a social network. Whereas in the latter case, users usually remain part of the network for a long time, in the IoT, new data sources are constantly being added or removed. Therefore, it is also important to research how the approach can be adapted to such an ever-changing environment.

We consider these seven challenges to be the most critical ones related to the protection of sensitive data in the information age that need to be addressed in future work.

6. Conclusions

In the modern information age, we are accustomed to smart services facilitating our everyday lives. We use these digital assistants in public, industrial, and private domains. Such data-driven services are so handy, as they adapt their behavior based on the current context and thus always provide an optimized user experience. However, this convenience does not come without a price. Smart services rely on permanent access to a vast amount of data. They analyze these data comprehensively to derive knowledge about the current situation of their users. Yet, this often involves highly personal or confidential data, allowing data processors to obtain sensitive information. For this reason, there are a variety of privacy measures that conceal certain sensitive knowledge patterns in the data without impairing the quality of the data. That is, they try to protect the privacy of data subjects and at the same time maintain the utility of their data for the respective smart services.

For this reason, this paper addresses the question of whether state-of-the-art privacy mechanisms are prepared to meet this challenge. To this end, we carry out a SWOT-like assessment, in which we initially analyze the seven most relevant application scenarios for smart services. It is evident that users can benefit from these smart services in all situations of life. Yet, smart services also pose a high privacy threat due to the personal data they process. Although this statement generally applies to all smart services, it becomes apparent that the different application scenarios involve heterogeneous types of data. There is therefore no universal privacy threat but rather smart services-specific threats. Therefore, we also study the strengths and weaknesses of privacy approaches tailored to smart services. By comparing these two dimensions (opportunities and privacy threats of smart services on the one side and strengths and weaknesses of privacy approaches on the other side), we discover that there is an effective answer to every relevant privacy threat. However, there are fundamental problems with modern privacy technologies for smart services. Users are overwhelmed by the selection and configuration of privacy techniques. These techniques are always tailored to a specific use case, which is why a great deal of domain knowledge is required to apply them effectively. Otherwise, their protective effect is insufficient, and the negative impact on data quality is too high. The latter can lead to further security vulnerabilities if the data tampering is not target-oriented. Therefore, a trusted third party is often required for the application of the privacy techniques. From these findings, we derive pertinent open research questions and give our opinion on how they can be overcome. These research questions deal in particular with concepts for privacy requirements elicitation, a holistic privacy platform, the deployment of privacy filters, the configuration of privacy filters, a secure data management, proofs of data possession, and prevention of misinformation. Only when these issues are fully addressed is a privacy-by-design approach for smart devices feasible.

Author Contributions: Conceptualization, C.S.; methodology, C.S.; software, C.S.; validation, C.S.; formal analysis, 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:

CEP	complex event processing
CNN	convolutional neural network
DNA	deoxyribonucleic acid
e-commerce	electronic commerce
e-resource	electronic resource
e-service	electronic service
eHealth	electronic health
GDPR	general data protection regulation
GPS	global positioning system
GSM	global system for mobile communications
IoT	internet of things
LBS	location-based service
mHealth	mobile health
OSN	online social network
POI	point of interest
PoRR	proof of retrievability and reliability
SNIL	spread noise to intermediate wavelet levels
STPA-Sec	system–theoretic process analysis for security
STPA-Priv	system–theoretic process analysis for privacy
SWOT	strengths, weaknesses, opportunities, and threats
UniProt	universal protein resource
VDA	voice-controlled digital assistant
VDF	verifiable delay function

References

1. Weiser, M. The computer for the 21st century. *Sci. Am.* **1991**, *265*, 94–104. [CrossRef]
2. Presser, M. The Rise of IoT–why today? *IEEE Internet Things Newsl.* **2016**, *12*, 2016
3. Jesse, N. Internet of Things and Big Data: The disruption of the value chain and the rise of new software ecosystems. *AI Soc.* **2018**, *33*, 229–239. [CrossRef]
4. Hariri, R.H.; Fredericks, E.M.; Bowers, K.M. Uncertainty in big data analytics: Survey, opportunities, and challenges. *J. Big Data* **2019**, *6*, 44. [CrossRef]

5. 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.
6. 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. [CrossRef]
7. Liew, A. Understanding Data, Information, Knowledge And Their Inter-Relationships. *J. Knowl. Manag. Pract.* **2007**, *8*, 134.
8. Stöhr, C.; Janssen, M.; Niemann, J.; Reich, B. Smart Services. *Procedia Soc. Behav Sci.* **2018**, *238*, 192–198.
9. Kashef, M.; Visvizi, A.; Troisi, O. Smart city as a smart service system: Human-computer interaction and smart city surveillance systems. *Comput. Hum. Behav.* **2021**, *124*, 106923. [CrossRef]
10. Lee, J.; Kao, H.A.; Yang, S. Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. *Procedia CIRP* **2014**, *16*, 3–8. [CrossRef]
11. Pramanik, M.I.; Lau, R.Y.; Demirkan, H.; Azad, M.A.K. Smart health: Big data enabled health paradigm within smart cities. *Expert Syst. Appl.* **2017**, *87*, 370–383. [CrossRef]
12. Nissenbaum, H. Protecting Privacy in an Information Age: The Problem of Privacy in Public. *Law Philos* **1998**, *17*, 559–596. [CrossRef]
13. 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://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 17 October 2022).
14. 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]
15. Dewri, R.; Ray, I.; Ray, I.; Whitley, D. Exploring privacy versus data quality trade-offs in anonymization techniques using multi-objective optimization. *J. Comput. Secur.* **2011**, *19*, 935–974. [CrossRef]
16. Ramson, S.J.; Vishnu, S.; Shanmugam, M. Applications of Internet of Things (IoT) – An Overview. In Proceedings of the 2020 5th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India, 5–6 March 2020; IEEE: Manhattan, NY, USA, 2020; pp. 92–95.
17. Dias, R.M.; Marques, G.; Bhoi, A.K. Internet of Things for Enhanced Food Safety and Quality Assurance: A Literature Review. In Proceedings of the International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy (ETAEERE), Bhubaneswar, India, 5–6 March 2020; Springer: Singapore, 2021; pp. 653–663.
18. Nawara, D.; Kashef, R. IoT-based Recommendation Systems – An Overview. In Proceedings of the 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, 9–12 September 2020; IEEE: Manhattan, NY, USA, 2020; pp. 1–7.
19. Huffine, E.; Kumar, A.; Kashyap, A. Attaining State of the Art in DNA Tests. In *Handbook of DNA Forensic Applications and Interpretation*; Kumar, A., Goswami, G.K., Huffine, E., Eds.; Springer: Singapore, 2022; pp. 11–23.
20. Zainuddin, N.; Daud, M.; Ahmad, S.; Maslizan, M.; Abdullah, S.A.L. A Study on Privacy Issues in Internet of Things (IoT). In Proceedings of the 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP), Zhuhai, Chinal, 8–10 January 2021; IEEE: Manhattan, NY, USA, 2021; pp. 96–100.
21. Junglas, I.A.; Watson, R.T. Location-Based Services. *Commun. ACM* **2008**, *51*, 65–69. [CrossRef]
22. Raper, J.; Gartner, G.; Karimi, H.; Rizos, C. Applications of location–based services: A selected review. *J. Locat. Based Serv.* **2007**, *1*, 89–111. [CrossRef]
23. Agre, P.E. Welcome to the always-on world. *IEEE Spectr* **2001**, *38*, 10–13. [CrossRef]
24. D'Roza, T.; Bilchev, G. An Overview of Location-Based Services. *BT Technol. J.* **2003**, *21*, 20–27. [CrossRef]
25. Obeidat, H.; Shuaieb, W.; Obeidat, O.; Abd-Alhameed, R. A Review of Indoor Localization Techniques and Wireless Technologies. *Kluw. Commun.* **2021**, *119*, 289–327. [CrossRef]
26. Dey, A.; Hightower, J.; de Lara, E.; Davies, N. Location-Based Services. *IEEE Pervasive Comput.* **2010**, *9*, 11–12. [CrossRef]
27. Bhatti, M.A.; Riaz, R.; Rizvi, S.S.; Shokat, S.; Riaz, F.; Kwon, S.J. Outlier detection in indoor localization and Internet of Things (IoT) using machine learning. *J. Commun. Netw.* **2020**, *22*, 236–243. [CrossRef]
28. Ezzat, M.; Sakr, M.; Elgohary, R.; Khalifa, M.E. Building road segments and detecting turns from GPS tracks. *J. Comput. Sci.* **2018**, *29*, 81–93. [CrossRef]
29. Zheng, Y. Trajectory Data Mining: An Overview. *ACM Trans. Intell Syst. Technol.* **2015**, *6*, 1–41. [CrossRef]
30. Krumm, J. Trajectory Analysis for Driving. In *Computing with Spatial Trajectories*; Zheng, Y., Zhou, X., Eds.; Springer: New York, NY, USA, 2011; pp. 213–241.
31. Chen, C.C.; Chiang, M.F. Trajectory pattern mining: Exploring semantic and time information. In Proceedings of the 2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI), Hsinchu, Taiwan, 25–27 November 2016; IEEE: Manhattan, NY, USA, 2016; pp. 130–137.
32. Teng, X.; Trajcevski, G.; Kim, J.S.; Züfle, A. Semantically Diverse Path Search. In Proceedings of the 2020 21st IEEE International Conference on Mobile Data Management (MDM), Versailles, France, 30 June–3 July 2020; IEEE: Manhattan, NY, USA, 2020; pp. 69–78.

33. 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.
34. Ceikute, V.; Jensen, C.S. Vehicle Routing with User-Generated Trajectory Data. In Proceedings of the 2015 16th IEEE International Conference on Mobile Data Management (MDM), Pittsburgh, PA, USA, 15–18 June 2015; IEEE: Manhattan, NY, USA, 2015; pp. 14–23.
35. Salim, S.; Turnbull, B.; Moustafa, N. Data analytics of social media 3.0: Privacy protection perspectives for integrating social media and Internet of Things (SM-IoT) systems. *Ad Hoc Netw.* **2022**, *128*, 102786. [CrossRef]
36. Li, N.; Chen, G. Analysis of a Location-Based Social Network. In Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE), Vancouver, BC, Canada, 29–31 August 2009; IEEE: Manhattan, NY, USA, 2009; pp. 263–270.
37. Liu, S.; Li, L.; Tang, J.; Wu, S.; Gaudiot, J.L. *Creating Autonomous Vehicle Systems*, 2nd ed.; Morgan & Claypool: San Rafael, CA, USA, 2020.
38. Primault, V.; Boutet, A.; Mokhtar, S.B.; Brunie, L. The Long Road to Computational Location Privacy: A Survey. *Commun. Surveys Tuts.* **2019**, *21*, 2772–2793. [CrossRef]
39. van Gemert-Pijnen, L.; Kelders, S.M.; Kip, H.; Sanderman, R. (Eds.) *eHealth Research, Theory and Development*; Routledge: London, UK, 2018.
40. Grady, A.; Yoong, S.; Sutherland, R.; Lee, H.; Nathan, N.; Wolfenden, L. Improving the public health impact of eHealth and mHealth interventions. *Aust. N. Z. J. Public Health* **2018**, *42*, 118–119. [CrossRef] [PubMed]
41. Kreps, G.L.; Neuhauser, L. New directions in eHealth communication: Opportunities and challenges. *Patient Educ. Couns.* **2010**, *78*, 329–336. [CrossRef]
42. Marcolino, M.S.; Oliveira, J.a.A.Q.; D'Agostino, M.; Ribeiro, A.L.; Alkmim, M.B.M.; Novillo-Ortiz, D. The Impact of mHealth Interventions: Systematic Review of Systematic Reviews. *JMIR Mhealth Uhealth* **2018**, *6*, e23. [CrossRef]
43. Siewiorek, D. Generation smartphone. *IEEE Spectr.* **2012**, *49*, 54–58. [CrossRef]
44. Bitsaki, M.; Koutras, C.; Koutras, G.; Leymann, F.; Steimle, F.; Wagner, S.; Wieland, M. ChronicOnline: Implementing a mHealth solution for monitoring and early alerting in chronic obstructive pulmonary disease. *Health Inform. J.* **2017**, *23*, 179–207. [CrossRef]
45. Guo, S.; Guo, X.; Zhang, X.; Vogel, D. Doctor–patient relationship strength's impact in an online healthcare community. *Inf. Technol. Dev.* **2018**, *24*, 279–300. [CrossRef]
46. Ball, M.J.; Lillis, J. E-health: Transforming the physician/patient relationship. *Int. J. Med. Inform.* **2001**, *61*, 1–10. [CrossRef]
47. Iyengar, S. Mobile health (mHealth). In *Fundamentals of Telemedicine and Telehealth*; Gogia, S., Ed.; Academic Press: London, UK; San Diego, CA, USA; Cambridge, MA, USA; Oxford, UK, 2020; Chapter 12; pp. 277–294.
48. Rocha, T.A.H.; da Silva, N.C.; Barbosa, A.C.Q.; Elahi, C.; Vissoci, J.a.R.N. mHealth: Smart Wearable Devices and the Challenges of a Refractory Context. In *The Internet and Health in Brazil*; Pereira Neto, A., Flynn, M.B., Eds.; Springer: Cham, Switzerland, 2019; pp. 347–367.
49. Lupton, D. *The Quantified Self*; Polity: Cambridge, UK; Malden, MA, USA, 2016.
50. Swan, M. Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0. *J. Sens. Actuator Netw.* **2012**, *1*, 217–253. [CrossRef]
51. Stach, C.; Steimle, F.; Franco da Silva, A.C. TIROL: The Extensible Interconnectivity Layer for mHealth Applications. In Proceedings of the 23rd International Conference on Information and Software Technologies (ICIST), Druskininkai, Lithuania, 12–14 October 2017; Springer: Cham, Switzerland, 2017; pp. 190–202.
52. Swan, M. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* **2013**, *1*, 85–99. [CrossRef]
53. Chao, D.Y.; Lin, T.M.; Ma, W.Y. Enhanced Self-Efficacy and Behavioral Changes Among Patients With Diabetes: Cloud-Based Mobile Health Platform and Mobile App Service. *JMIR Diabetes* **2019**, *4*, e11017. [CrossRef]
54. Piccialli, F.; Giampaolo, F.; Prezioso, E.; Camacho, D.; Acampora, G. Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion. *Inform. Fusion* **2021**, *74*, 1–16. [CrossRef]
55. Deshpande, P.S.; Sharma, S.C.; Peddoju, S.K. Predictive and Prescriptive Analytics in Big-data Era. In *Security and Data Storage Aspect in Cloud Computing*; Springer: Singapore, 2019; pp. 71–81.
56. Noar, S.M.; Harrington, N.G. *eHealth Applications: Promising Strategies for Behavior Change*; Routledge: New York, NY, USA, 2012.
57. Ben Amor, L.; Lahyani, I.; Jmaiel, M. Data accuracy aware mobile healthcare applications. *Comput. Ind.* **2018**, *97*, 54–66. [CrossRef]
58. Thapa, C.; Camtepe, S. Precision health data: Requirements, challenges and existing techniques for data security and privacy. *Comput. Biol. Med.* **2021**, *129*, 104130. [CrossRef]
59. Kumar, T.; Liyanage, M.; Braeken, A.; Ahmad, I.; Ylianttila, M. From gadget to gadget-free hyperconnected world: Conceptual analysis of user privacy challenges. In Proceedings of the 2017 European Conference on Networks and Communications (EuCNC), Oulu, Finland, 12–15 June 2017; IEEE: Manhattan, NY, USA, 2017; pp. 1–6.
60. Braghin, C.; Cimato, S.; Della Libera, A. Are mHealth Apps Secure? A Case Study. In Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 23–27 July 2018; IEEE: Manhattan, NY, USA, 2018; pp. 335–340.
61. Hoy, M.B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med. Ref. Serv. Q.* **2018**, *37*, 81–88. [CrossRef]

62. López, G.; Quesada, L.; Guerrero, L.A. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction (HFSI), Los Angeles, CA, USA, 17–21 July 2017; Springer: Cham, Switzerland, 2018; pp. 241–250.
63. McLean, G.; Osei-Frimpong, K. Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Comput. Hum. Behav.* **2019**, *99*, 28–37. [CrossRef]
64. Porcheron, M.; Fischer, J.E.; Reeves, S.; Sharples, S. Voice Interfaces in Everyday Life. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI), Montreal, QC, Canada, 21–26 April 2018; ACM: New York, NY, USA, 2018; pp. 1–12.
65. Lei, X.; Tu, G.H.; Liu, A.X.; Li, C.Y.; Xie, T. The Insecurity of Home Digital Voice Assistants – Vulnerabilities, Attacks and Countermeasures. In Proceedings of the 2018 IEEE Conference on Communications and Network Security (CNS), Beijing, China, 30 May–1 June 2018; IEEE: Manhattan, NY, USA, 2018; pp. 1–9.
66. Chung, H.; Park, J.; Lee, S. Digital forensic approaches for Amazon Alexa ecosystem. *Digit. Investig.* **2017**, *22*, S15–S25. [CrossRef]
67. Lopatovska, I.; Rink, K.; Knight, I.; Raines, K.; Cosenza, K.; Williams, H.; Sorsche, P.; Hirsch, D.; Li, Q.; Martinez, A. Talk to me: Exploring user interactions with the Amazon Alexa. *J. Libr. Inf. Sci.* **2019**, *51*, 984–997. [CrossRef]
68. Han, S.; Yang, H. Understanding adoption of intelligent personal assistants: A parasocial relationship perspective. *Ind. Manag. Data Syst.* **2018**, *118*, 618–636. [CrossRef]
69. Bolton, T.; Dargahi, T.; Belguith, S.; Al-Rakhami, M.S.; Sodhro, A.H. On the Security and Privacy Challenges of Virtual Assistants. *Sensors* **2021**, *21*, 2312. [CrossRef] [PubMed]
70. Khan, M.J.; Khan, H.S.; Yousaf, A.; Khurshid, K.; Abbas, A. Modern Trends in Hyperspectral Image Analysis: A Review. *IEEE Access* **2018**, *6*, 14118–14129. [CrossRef]
71. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, Present, and Future of Face Recognition: A Review. *Electronics* **2020**, *9*, 1188. [CrossRef]
72. Hazelwood, K.; Bird, S.; Brooks, D.; Chintala, S.; Diril, U.; Dzhulgakov, D.; Fawzy, M.; Jia, B.; Jia, Y.; Kalro, A.; et al. Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. In Proceedings of the 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), Vienna, Austria, 24–28 February 2018; IEEE: Manhattan, NY, USA, 2018; pp. 620–629.
73. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; IEEE: Manhattan, NY, USA, 2014; pp. 1701–1708.
74. Kumar, A.; Kaur, A.; Kumar, M. Face detection techniques: A review. *Artif. Intell. Rev.* **2019**, *52*, 927–948. [CrossRef]
75. Taskiran, M.; Kahraman, N.; Erdem, C.E. Face recognition: Past, present and future (a review). *Digit Signal Process* **2020**, *106*, 102809. [CrossRef]
76. Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face Recognition Systems: A Survey. *Sensors* **2020**, *20*, 342. [CrossRef]
77. Li, L.; Mu, X.; Li, S.; Peng, H. A Review of Face Recognition Technology. *IEEE Access* **2020**, *8*, 139110–139120. [CrossRef]
78. Senior, A.W.; Pankanti, S. Privacy Protection and Face Recognition. In *Handbook of Face Recognition*; Li, S.Z., Jain, A.K., Eds.; Springer: London, UK, 2021; pp. 671–691.
79. Wang, M.; Deng, W. Deep face recognition: A survey. *Neurocomputing* **2021**, *429*, 215–244. [CrossRef]
80. Nielsen, S.S. (Ed.) *Food Analysis*, 5th ed.; Springer: Cham, Switzerland, 2017.
81. Mishra, G.K.; Barfidokht, A.; Tehrani, F.; Mishra, R.K. Food Safety Analysis Using Electrochemical Biosensors. *Foods* **2018**, *7*, 141. [CrossRef]
82. Korte, R.; Bräcker, J.; Brockmeyer, J. Gastrointestinal digestion of hazelnut allergens on molecular level: Elucidation of degradation kinetics and resistant immunoactive peptides using mass spectrometry. *Mol. Nutr. Food Res.* **2017**, *61*, 1700130. [CrossRef]
83. Berrueta, L.A.; Alonso-Salces, R.M.; Héberger, K. Supervised pattern recognition in food analysis. *J. Chromatogr. A* **2007**, *1158*, 196–214. [CrossRef]
84. Deng, X.; Cao, S.; Horn, A.L. Emerging Applications of Machine Learning in Food Safety. *Annu. Rev. Food Sci. Technol.* **2021**, *12*, 513–538. [CrossRef]
85. Bräcker, J.; Brockmeyer, J. Characterization and Detection of Food Allergens Using High-Resolution Mass Spectrometry: Current Status and Future Perspective. *J. Agric. Food Chem.* **2018**, *66*, 8935–8940. [CrossRef]
86. Mafata, M.; Brand, J.; Medvedovici, A.; Buica, A. Chemometric and sensometric techniques in enological data analysis. *Crit. Rev. Food Sci.* **2022**, 1–15. [CrossRef]
87. Bianco, M.; Ventura, G.; Calvano, C.D.; Losito, I.; Cataldi, T.R. A new paradigm to search for allergenic proteins in novel foods by integrating proteomics analysis and in silico sequence homology prediction: Focus on spirulina and chlorella microalgae. *Talanta* **2022**, *240*, 123188. [CrossRef]
88. Giatrakos, N.; Alevizos, E.; Artikis, A.; Deligiannakis, A.; Garofalakis, M. Complex event recognition in the Big Data era: A survey. *VLDB J.* **2020**, *29*, 313–352. [CrossRef]
89. Alakari, A.; Li, K.F.; Gebali, F. A situation refinement model for complex event processing. *Knowl.-Based Syst.* **2020**, *198*, 105881. [CrossRef]

90. Cardoso, D.R.; Andrade-Sobrinho, L.G.; Leite-Neto, A.F.; Reche, R.V.; Isique, W.D.; Ferreira, M.M.C.; Lima-Neto, B.S.; Franco, D.W. Comparison between Cachaça and Rum Using Pattern Recognition Methods. *J. Agric. Food Chem.* **2004**, *52*, 3429–3433. [CrossRef]
91. Şen, G.; Medeni, İ.T.; Şen, K.Ö.; Durakbasa, N.M.; Medeni, T.D. Sensor Based Intelligent Measurement and Blockchain in Food Quality Management. In *Digitizing Production Systems: Selected Papers from ISPR2021, 7–9 October 2021, Online, Turkey*; Durakbasa, N.M., Gençyılmaz, M.G., Eds.; Springer: Cham, Switzerland, 2022; pp. 323–334.
92. Nielsen, K.M. Biosafety Data as Confidential Business Information. *PLOS Biol.* **2013**, *11*, e1001499. [CrossRef] [PubMed]
93. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl.-Based Syst.* **2013**, *46*, 109–132. [CrossRef]
94. Lu, J.; Wu, D.; Mao, M.; Wang, W.; Zhang, G. Recommender system application developments: A survey. *Decis. Support Syst.* **2015**, *74*, 12–32. [CrossRef]
95. Maske, A.R.; Joglekar, B. An Algorithmic Approach for Mining Customer Behavior Prediction in Market Basket Analysis. In Proceedings of the Sixth International Conference on Innovations in Computer Science and Engineering (ICICSE), Hyderabad, India, 17–18 August 2018; Springer: Singapore, 2019; pp. 31–38.
96. Lops, P.; de Gemmis, M.; Semeraro, G. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*; Ricci, F., Rokach, L., Shapira, B., Kantor, P.B., Eds.; Springer: Boston, MA, USA, 2011; pp. 73–105.
97. Carrer-Neto, W.; Hernández-Alcaraz, M.L.; Valencia-García, R.; García-Sánchez, F. Social knowledge-based recommender system. Application to the movies domain. *Expert Syst. Appl.* **2012**, *39*, 10990–11000. [CrossRef]
98. Afoudi, Y.; Lazaar, M.; Al Achhab, M. Collaborative Filtering Recommender System. In Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD), Tangier, Morocco, 12–14 July 2018; Springer: Cham, Switzerland, 2019; pp. 332–345.
99. Thorat, P.B.; Goudar, R.M.; Barve, S.S. Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. *Int. J. Comput. Appl.* **2015**, *110*, 31–36.
100. Resnick, P.; Varian, H.R. Recommender Systems. *Commun. ACM* **1997**, *40*, 56–58. [CrossRef]
101. Saad, R. Discovery, development, and current applications of DNA identity testing. In *Baylor University Medical Center Proceedings*; Taylor & Francis: New York, NY, USA, 2005; Volume 18, pp. 130–133.
102. Jin, Z.; Liu, Y. DNA methylation in human diseases. *Genes Dis.* **2018**, *5*, 1–8. [CrossRef]
103. Onabote, O.; Hassan, H.M.; Isovic, M.; Torchia, J. The Role of Thymine DNA Glycosylase in Transcription, Active DNA Demethylation, and Cancer. *Cancers* **2022**, *14*, 765. [CrossRef]
104. Li, X.; Liu, Y.; Salz, T.; Hansen, K.D.; Feinberg, A. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Res.* **2016**, *26*, 1730–1741. [CrossRef]
105. Ahmed, I.; Jeon, G. Enabling Artificial Intelligence for Genome Sequence Analysis of COVID-19 and Alike Viruses. *Interdiscip Sci.* **2021**, 1–16. *Online ahead of print*. [CrossRef]
106. Wang, G.; Pu, P.; Shen, T. An efficient gene bigdata analysis using machine learning algorithms. *Multimed. Tools Appl.* **2020**, *97*, 9847–9870. [CrossRef]
107. Schwab, A.P.; Luu, H.S.; Wang, J.; Park, J.Y. Genomic Privacy. *Clin. Chem.* **2018**, *64*, 1696–1703. [CrossRef]
108. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
109. Angelov, P.; Soares, E. Towards explainable deep neural networks (xDNN). *Neural Netw.* **2020**, *130*, 185–194. [CrossRef]
110. Almusaylim, Z.A.; Jhanjhi, N. Comprehensive Review: Privacy Protection of User in Location-Aware Services of Mobile Cloud Computing. *Wireless Pers. Commun.* **2020**, *111*, 541–564. [CrossRef]
111. Finck, M.; Pallas, F. They who must not be identified—Distinguishing personal from non-personal data under the GDPR. *Int. Data Priv. Law* **2020**, *10*, 11–36. [CrossRef]
112. Rassouli, B.; Rosas, F.E.; Gündüz, D. Data Disclosure Under Perfect Sample Privacy. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 2012–2025. [CrossRef]
113. Al-Rubaie, M.; Chang, J.M. Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Secur. Priv.* **2019**, *17*, 49–58. [CrossRef]
114. 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]
115. 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*, 31:1–31:36. [CrossRef]
116. 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.
117. Jiang, H.; Li, J.; Zhao, P.; Zeng, F.; Xiao, Z.; Iyengar, A. Location Privacy-Preserving Mechanisms in Location-Based Services: A Comprehensive Survey. *ACM Comput. Surv.* **2021**, *54*, 4:1–4:36. [CrossRef]
118. Ardagna, C.A.; Cremonini, M.; Damiani, E.; De Capitani di Vimercati, S.; Samarati, P. Location Privacy Protection Through Obfuscation-Based Techniques. In Proceedings of the 21st Annual IFIP WG 11.3 Working Conference on Data and Applications Security (DBSec), Redondo Beach, CA, USA, 8–11 July 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 47–60.

119. 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, Madeira, Portugal, 19–21 March 2018; SciTePress: Setúbal, Portugal, 2018; pp. 589–595.
120. 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.
121. Hara, T.; Suzuki, A.; Iwata, M.; Arase, Y.; Xie, X. Dummy-Based User Location Anonymization Under Real-World Constraints. *IEEE Access* **2016**, *4*, 673–687. [CrossRef]
122. Siddiqie, S.; Mondal, A.; Reddy, P.K. An Improved Dummy Generation Approach for Enhancing User Location Privacy. In Proceedings of the 26th International Conference on Database Systems for Advanced Applications (DASFAA), Taipei, Taiwan, 11–14 April 2021; Springer: Cham, Switzerland, 2021; pp. 487–495.
123. Ma, Y.; Bai, X.; Wang, Z. Trajectory Privacy Protection Method based on Shadow vehicles. In Proceedings of the 2021 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), New York, NY, USA, 30 September–3 October 2021; IEEE: Manhattan, NY, USA, 2021; pp. 668–673.
124. Khazbak, Y.; Fan, J.; Zhu, S.; Cao, G. Preserving personalized location privacy in ride-hailing service. *Tsinghua Sci. Technol.* **2020**, *25*, 743–757. [CrossRef]
125. Li, C.; Palanisamy, B. Reversible spatio-temporal perturbation for protecting location privacy. *Comput. Commun.* **2019**, *135*, 16–27. [CrossRef]
126. He, Y.; Chen, J. User location privacy protection mechanism for location-based services. *Digit. Commun. Netw.* **2021**, *7*, 264–276. [CrossRef]
127. 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.
128. Pourahmadi, M. Estimation and Interpolation of Missing Values of a Stationary Time Series. *J. Time Ser. Anal.* **1989**, *10*, 149–169. [CrossRef]
129. Ramosaj, B.; Pauly, M. Predicting missing values: A comparative study on non-parametric approaches for imputation. *Computation Stat.* **2019**, *34*, 1741–1764. [CrossRef]
130. Thomakos, D. Smoothing Non-Stationary Time Series Using the Discrete Cosine Transform. *J. Syst. Sci. Complex* **2016**, *29*, 382–404. [CrossRef]
131. Rhif, M.; Ben Abbes, A.; Farah, I.R.; Martínez, B.; Sang, Y. Wavelet Transform Application for/in Non-Stationary Time-Series Analysis: A Review. *Appl. Sci.* **2019**, *9*, 1345. [CrossRef]
132. Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; Naor, M. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), St. Petersburg, Russia, 28 May–1 June 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 486–503.
133. Gao, Q.; Zhu, L.; Lin, Y.; Chen, X. Anomaly Noise Filtering with Logistic Regression and a New Method for Time Series Trend Computation for Monitoring Systems. In Proceedings of the 2019 IEEE 27th International Conference on Network Protocols (ICNP), Chicago, IL, USA, 8–10 October 2019; IEEE: Manhattan, NY, USA, 2019; pp. 1–6.
134. Moon, Y.S.; Kim, H.S.; Kim, S.P.; Bertino, E. Publishing Time-Series Data under Preservation of Privacy and Distance Orders. In Proceedings of the 21th International Conference on Database and Expert Systems Applications (DEXA), Bilbao, Spain, 30 August–3 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 17–31.
135. 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.
136. Cheng, P.; Roedig, U. Personal Voice Assistant Security and Privacy–A Survey. *Proc IEEE (Early Access)* **2022**, 1–32. [CrossRef]
137. Mhaidli, A.; Venkatesh, M.K.; Zou, Y.; Schaub, F. Listen Only When Spoken To: Interpersonal Communication Cues as Smart Speaker Privacy Controls. *Proc. Priv. Enhanc. Technol.* **2020**, *2020*, 251–270. [CrossRef]
138. Chen, S.; Ren, K.; Piao, S.; Wang, C.; Wang, Q.; Weng, J.; Su, L.; Mohaisen, A. You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones. In Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, 5–8 June 2017; IEEE: Manhattan, NY, USA, 2017; pp. 183–195.
139. Gao, C.; Chandrasekaran, V.; Fawaz, K.; Banerjee, S. Traversing the Quagmire That is Privacy in Your Smart Home. In Proceedings of the 2018 Workshop on IoT Security and Privacy (IoT S&P), Budapest, Hungary, 20 August 2018; ACM: New York, NY, USA, 2018; pp. 22–28.
140. Saade, A.; Dureau, J.; Leroy, D.; Caltagirone, F.; Coucke, A.; Ball, A.; Doumouro, C.; Lavril, T.; Caulier, A.; Bluche, T.; et al. Spoken Language Understanding on the Edge. In Proceedings of the 2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing—NeurIPS Edition (EMC2-NIPS), Vancouver, BC, Canada, 13 December 2019; IEEE: Manhattan, NY, USA, 2019; pp. 57–61.
141. He, Y.; Sainath, T.N.; Prabhavalkar, R.; McGraw, I.; Alvarez, R.; Zhao, D.; Rybach, D.; Kannan, A.; Wu, Y.; Pang, R.; et al. Streaming End-to-end Speech Recognition for Mobile Devices. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Manhattan, NY, USA, 2019; pp. 6381–6385.

142. Tiwari, V.; Hashmi, M.F.; Keskar, A.; Shivaprakash, N.C. Virtual home assistant for voice based controlling and scheduling with short speech speaker identification. *Multimed. Tools Appl.* **2020**, *79*, 5243–5268. [CrossRef]
143. Perez, A.J.; Zeadally, S.; Griffith, S. Bystanders' Privacy. *IT Prof* **2017**, *19*, 61–65. [CrossRef]
144. Hernández Acosta, L.; Reinhardt, D. A survey on privacy issues and solutions for Voice-controlled Digital Assistants. *Pervasive Mob. Comput.* **2022**, *80*, 101523. [CrossRef]
145. Qian, J.; Du, H.; Hou, J.; Chen, L.; Jung, T.; Li, X.Y. Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems (SenSys), Shenzhen, China, 4–7 November 2018; ACM: New York, NY, USA, 2018; pp. 82–94.
146. Tian, C.; Fei, L.; Zheng, W.; Xu, Y.; Zuo, W.; Lin, C.W. Deep learning on image denoising: An overview. *Neural Netw.* **2020**, *131*, 251–275. [CrossRef]
147. 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.
148. Fan, L. Practical Image Obfuscation with Provable Privacy. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; IEEE: Manhattan, NY, USA, 2019; pp. 784–789.
149. Yu, J.; Zhang, B.; Kuang, Z.; Lin, D.; Fan, J. iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1005–1016. [CrossRef]
150. Sarwar, O.; Rinner, B.; Cavallaro, A. A Privacy-Preserving Filter for Oblique Face Images Based on Adaptive Hopping Gaussian Mixtures. *IEEE Access* **2019**, *7*, 142623–142639. [CrossRef]
151. Gehrke, J.; Lui, E.; Pass, R. Towards Privacy for Social Networks: A Zero-Knowledge Based Definition of Privacy. In Proceedings of the 8th Conference on Theory of Cryptography (TCC), Providence, RI, USA, 28–30 March 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 432–449.
152. Quoc, D.L.; Beck, M.; Bhatotia, P.; Chen, R.; Fetzer, C.; Strufe, T. PrivApprox: Privacy-Preserving Stream Analytics. In Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC), Santa Clara, CA, USA, 12–14 July 2017; USENIX Association: Berkeley, CA, USA, 2017; pp. 659–672.
153. Li, F.; Wang, N.; Gu, Y.; Chen, Z. Effective Privacy Preservation over Composite Events with Markov Correlations. In Proceedings of the 2016 13th Web Information Systems and Applications Conference (WISA), Wuhan, China, 23–25 September 2016; IEEE: Manhattan, NY, USA, 2016; pp. 215–220.
154. Churi, P.P.; Pawar, A.V. A Systematic Review on Privacy Preserving Data Publishing Techniques. *J. Eng. Sci. Technol. Rev.* **2019**, *12*, 17–25. [CrossRef]
155. 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, Madeira, Portugal, 22–24 January 2018; SciTePress: Setúbal, Portugal, 2018; pp. 30–40.
156. Palanisamy, S.M.; Dürr, F.; Tariq, M.A.; Rothermel, K. Preserving Privacy and Quality of Service in Complex Event Processing through Event Reordering. In Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems (DEBS), Hamilton, New Zealand, 25–29 June 2018; ACM: New York, NY, USA, 2018; pp. 40–51.
157. Palanisamy, S.M. Towards Multiple Pattern Type Privacy Protection in Complex Event Processing Through Event Obfuscation Strategies. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2020 International Workshops, DPM 2020 and CBT 2020, Guildford, UK, 17–18 September 2020, Revised Selected Papers*; Garcia-Alfaro, J., Navarro-Arribas, G., Herrera-Joancomarti, J., Eds.; Springer: Cham, Switzerland, 2020; pp. 178–194.
158. 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.
159. Psychoula, I.; Chen, L.; Amft, O. Privacy Risk Awareness in Wearables and the Internet of Things. *IEEE Pervasive Comput* **2020**, *19*, 60–66. [CrossRef]
160. 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.
161. Jain, P.; Gyanchandani, M.; Khare, N. Differential privacy: Its technological prescriptive using big data. *J. Big. Data* **2018**, *5*, 15. [CrossRef]
162. 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.
163. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
164. Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. *Federated Learning*; Morgan & Claypool: San Rafael, CA, USA, 2019.
165. Wu, X.; Zhang, Y.; Shi, M.; Li, P.; Li, R.; Xiong, N.N. An adaptive federated learning scheme with differential privacy preserving. *Future Gener. Comput. Syst.* **2022**, *127*, 362–372. [CrossRef]
166. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *NPJ Digit. Med.* **2020**, *3*, 119. [CrossRef] [PubMed]

167. Wang, H.; Zhao, Q.; Wu, Q.; Chopra, S.; Khaitan, A.; Wang, H. Global and Local Differential Privacy for Collaborative Bandits. In Proceedings of the Fourteenth ACM Conference on Recommender Systems (RecSys), Rio de Janeiro, Brazil, 22–26 September 2020; ACM: New York, NY, USA, 2020; pp. 150–159.
168. 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.
169. Piercy, N.; Giles, W. Making SWOT Analysis Work. *Mark. Intell. Plan.* **1989**, *7*, 5–7. [CrossRef]
170. Benzaghta, M.A.; Elwalda, A.; Mousa, M.M.; Erkan, I.; Rahman, M. SWOT Analysis Applications: An Integrative Literature Review. *J. Glob. Bus. Insights* **2021**, *6*, 55–73. [CrossRef]
171. Young, W.; Leveson, N.G. An Integrated Approach to Safety and Security Based on Systems Theory. *Commun. ACM* **2014**, *127*, 31–35. [CrossRef]
172. 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.
173. 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.
174. Hanisch, S.; Cabarcos, P.A.; Parra-Arnau, J.; Strufe, T. Privacy-Protecting Techniques for Behavioral Data: A Survey. *CoRR* **2021**, *abs/2109.04120*, 1–43.
175. Wu, X.; Zhang, Y.; Wang, A.; Shi, M.; Wang, H.; Liu, L. MNSSp3: Medical big data privacy protection platform based on Internet of things. *Neural Comput. Applic* **2022**, *34*, 11491–11505. [CrossRef]
176. 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.
177. Shapiro, S.S. Time to Modernize Privacy Risk Assessment. *Issues Sci. Technol.* **2021**, *38*, 20–22.
178. 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.
179. 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.
180. Beierle, F.; Tran, V.T.; Allemand, M.; Neff, P.; Schlee, W.; Probst, T.; Pryss, R.; Zimmermann, J. Context Data Categories and Privacy Model for Mobile Data Collection Apps. *Procedia Comput. Sci.* **2018**, *134*, 18–25. [CrossRef]
181. 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.
182. 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.
183. Busch-Casler, J.; Radic, M. Personal Data Markets: A Narrative Review on Influence Factors of the Price of Personal Data. In Proceedings of the 16th International Conference on Research Challenges in Information Science (RCIS), Barcelona, Spain, 17–20 May 2022; Springer: Cham, Switzerland, 2022; pp. 3–19.
184. Driessen, S.W.; Monsieur, G.; Van Den Heuvel, W.J. Data Market Design: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 33123–33153. [CrossRef]
185. Spiekermann, S.; Acquisti, A.; Böhme, R.; Hui, K.L. The challenges of personal data markets and privacy. *Electron Mark* **2015**, *25*, 161–167. [CrossRef]
186. Stach, C.; Gritti, C.; Przytarski, D.; Mitschang, B. Trustworthy, Secure, and Privacy-aware Food Monitoring Enabled by Blockchains and the IoT. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; IEEE: Manhattan, NY, USA, 2020; pp. 50:1–50:4.
187. Bernal Bernabe, J.; Canovas, J.L.; Hernandez-Ramos, J.L.; Torres Moreno, R.; Skarmeta, A. Privacy-Preserving Solutions for Blockchain: Review and Challenges. *IEEE Access* **2019**, *7*, 164908–164940. [CrossRef]
188. 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.
189. Boneh, D.; Bonneau, J.; Bünz, B.; Fisch, B. Verifiable Delay Functions. In Proceedings of the 38th International Cryptology Conference (Crypto), Santa Barbara, CA, USA, 17–19 August 2018; Springer: Cham, Switzerland, 2018; pp. 757–788.

190. Gritti, C.; Li, H. Efficient Publicly Verifiable Proofs of Data Replication and Retrievability Applicable for Cloud Storage. *Adv. Sci. Technol. Eng. Syst. J.* **2022**, *7*, 107–124. [CrossRef]
191. Chow, R.; Golle, P. Faking Contextual Data for Fun, Profit, and Privacy. In Proceedings of the 8th ACM Workshop on Privacy in the Electronic Society (WPES), Chicago, IL, USA, 9 November 2009; ACM: New York, NY, USA, 2009; pp. 105–108.
192. 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.
193. Litou, I.; Kalogeraki, V.; Katakis, I.; Gunopulos, D. Real-Time and Cost-Effective Limitation of Misinformation Propagation. In Proceedings of the 2016 17th IEEE International Conference on Mobile Data Management (MDM), Porto, Portugal, 13–16 June 2016; IEEE: Manhattan, NY, USA, 2016; pp. 158–163.
194. Litou, I.; Kalogeraki, V.; Katakis, I.; Gunopulos, D. Efficient and timely misinformation blocking under varying cost constraints. *Online Soc. Netw. Media* **2017**, *2*, 19–31. [CrossRef]