AMNESIA: A Technical Solution towards GDPR-compliant Machine Learning

Christoph Stach[1], Corinna Giebler[1], Manuela Wagner[2], Christian Weber[3] and Bernhard Mitschang[1,3]
[1]*Institute for Parallel and Distributed Systems, University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany*
[2]*FZI Forschungszentrum Informatik, Haid-und-Neu-Straße 10–14, 76131 Karlsruhe, Germany*
[3]*Graduate School advanced Manufacturing Engineering, University of Stuttgart, Nobelstraße 12, 70569 Stuttgart, Germany*

Keywords: Machine Learning, Data Protection, Privacy Zones, Access Control, Model Management, Provenance, GDPR.

Abstract: *Machine Learning* (*ML*) applications are becoming increasingly valuable due to the rise of *IoT* technologies. That is, sensors continuously gather data from different domains and make them available to ML for learning its models. This provides profound insights into the data and enables predictions about future trends. While ML has many advantages, it also represents an immense privacy risk. Data protection regulations such as the *GDPR* address such privacy concerns, but practical solutions for the technical enforcement of these laws are also required. Therefore, we introduce *AMNESIA*, a priv**a**cy-aware **m**achine learni**n**g mod**e**l provi**s**ioni**n**g pl**a**tform. AMNESIA is a holistic approach covering all stages from data acquisition to model provisioning. This enables to control which application may use which data for ML as well as to make models "forget" certain knowledge.

1 INTRODUCTION

Machine Learning (*ML*) and *Artificial Intelligence* (*AI*) systems are on the rise. IDC predicts that the worldwide spending on such systems almost doubles in 2019 (that is $ 35.8 billion). This trend is expected to continue until 2022, with the result of a compound annual growth rate of 38.0 % (Shirer and D'Aquila, 2019). This is due to the fact that ML is no longer a laboratory curiosity, but a practical technology ready for commercial use. The capability of ML applications is virtually unlimited. Application areas in which ML is successfully used today include *Smart Health* (e. g., for autonomous and assistive diagnostic support), *Smart Traffic* (e. g., autonomous vehicle control and guidance systems), and *Computer Vision* (e. g., for object recognition).

These ML applications benefit from the increasing popularity of the *Internet of Things* (*IoT*). IoT-enabled devices are able to gather various context data and have Internet connectivity to share these data with ML systems. With these, the data of all devices can be analyzed and combined. This way, a comprehensive knowledge base is created covering various aspects. *Supervised learning* algorithms use this knowledge base to train their models. To this end, the data are analyzed to identify how certain attribute combinations affect a particular feature. For instance, with classification algorithms, electronic medical records can be analyzed to determine which symptoms most likely result in which diagnosis. These models can then be applied to new data to make predictions about the investigated feature. For instance, if a patient monitors his or her health via medical IoT meters, an ML system is able to provide a diagnosis based on these data. However, particularly in *Deep Learning*, it is almost impossible to explain, why a decision was made and on what data it is based (LeCun et al., 2015).

While ML offers numerous benefits to users, it also causes a variety of problems, in particular concerning privacy and trust. Although it does not seem so at first glance, this is an inherent contradiction. A patient only trusts in privacy measures if s/he has total control over his or her data. This includes that s/he can decide which data contribute to an ML model and that s/he can mask the data. Yet, s/he lacks the insight into the effects this has on the utility of the data and thus on the ML model. This impairs prediction quality whereby the trust in the ML application is weakened (Holzinger et al., 2018). Therefore, so-called *usable security* is required, which takes into account both, security and utility issues (Garfinkel and Lipford, 2014).

This need for novel privacy approaches is also evident in the increasing number of privacy laws such as the *General Data Protection Regulation (EU) 2016/679* (*GDPR*). The GDPR addresses many of the

Stach, C., Giebler, C., Wagner, M., Weber, C. and Mitschang, B.
AMNESIA: A Technical Solution towards GDPR-compliant Machine Learning.
DOI: 10.5220/0008916700210032
In *Proceedings of the 6th International Conference on Information Systems Security and Privacy (ICISSP 2020)*, pages 21-32
ISBN: 978-989-758-399-5; ISSN: 2184-4356

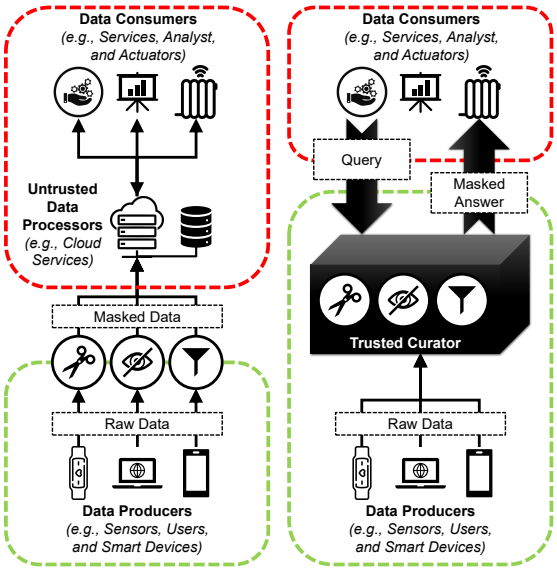
privacy risks caused by ML. To this end, it creates governing principles for the processing of private data by establishing privacy standards, e. g., informed consent, notification duties, and transparency (Wachter, 2018).

To meet this request for a *Privacy by Design* solution, we introduce *AMNESIA*, a privacy-aware **m**achine learni**n**g mod**e**l provisio**n**i**n**g pl**a**tform. AMNESIA "forgets" certain data when learning a model, i. e., these data are subsequently not used, when applying the model (e. g., to make an automatic decision). For this, we organize data in different zones, similar to information in a brain. In each zone, different privacy techniques are applied to hide certain private data. AMNESIA not only considers privacy requirements of users, but also utility requirements towards the ML model. In addition, users are able to track which data have been used in an ML model and whether an applied ML model complies with their privacy requirements.

To this end, we make five contributions: **(1)** We present our privacy-aware machine learning model provisioning platform called AMNESIA. It enables a GDPR-compliant learning and application of ML models. **(2)** We introduce a privacy zone model in AMNESIA. Each zone masks private data differently. As a result, ML models can be learned that comply with users' privacy requirements and yet have a high utility. **(3)** We outline an implementation strategy for AMNESIA. This strategy adopts a Privacy by Design approach. **(4)** We describe a model management concept in AMNESIA to find appropriate ML models in terms of privacy and utility. **(5)** We depict a provenance analysis in AMNESIA to verify which data were used to learn an applied ML model and whether the model complies with the privacy requirements. The remainder of this paper is as follows: In Section 2, we discuss both, technical and legal state of the art regarding privacy in an ML environment. On this basis, we derive requirements towards an ML privacy approach in Section 3. Related work is reviewed in Section 4 with regard to these requirements. Then, we introduce AMNESIA in Section 5 and evaluate our approach in Section 6. Section 7 concludes this paper.

2 STATE OF THE ART

Vayena et al. (Vayena et al., 2018) investigate the attitude of data subjects and data consumers towards ML. Although the benefits of ML are widely perceived, nearly two-thirds of the data subjects are concerned if their health data are used to learn and train a ML model. This concern is exacerbated by the fact that many models lack transparency, i. e., data subjects do not know how their data are reflected in the model and whether their privacy requirements are respected.



(a) Local Privacy. (b) Global Privacy.

Figure 1: Comparison of Privacy Strategies for Machine Learning Systems (cf. (Kairouz et al., 2016)).

To address these concerns, privacy concepts have to cover all ML stages, starting with the handling of data sources, followed by the implementation of ML systems and the creation of models up to the provisioning and application of the models. Yet, legal standards such as the GDPR are not sufficient, in order to eliminate all ethical and regulatory concerns. In addition, Privacy by Design approaches are required, i. e., ML systems in which privacy options are integrated that give data subjects full control over their personal data and thus ensure these standards.

Next, we therefore discuss technical solutions ensuring privacy in ML systems and study the legal situation with regard to privacy in the ML context.

Technical Perspective. Basically, there are two strategies to ensure privacy in ML systems: *Local Privacy* and *Global Privacy*. Figure 1 shows these strategies. The key difference is which components are trusted (green box) and which are not (red box).

In Local Privacy (see Figure 1a), the user only trusts his or her own devices since s/he does not know which data protection measures are taken by an ML system nor which data consumers have access to his or her data. Before s/he submits any data to an ML system, s/he masks them (e. g., by adding noise). Thus, the ML system has only distorted data for learning its models. Yet, if each user masks his or her data individually, the utility of the models decreases significantly. Moreover, a user does not know which data sources

are available to an ML system nor what knowledge can be derived from them (Chamikara et al., 2018).

In Global Privacy (see Figure 1b), the user entrusts his or her raw data to a *Trusted Curator* which provides various privacy techniques for masking user data. By managing all users' data, the Trusted Curator is able to apply more comprehensive privacy techniques than an individual user could do (e. g., *differential privacy*). Data consumers can only access data via defined interfaces, e. g., the Trusted Curator can learn ML models on the masked data and provide them to a service instead of the actual data. However, blind trust in the Trusted Curator is required. Moreover, global masking policies are applied to all users' data, whereas privacy is an individual perception (Li et al., 2017).

Legal Perspective. Since ML processes *personal data* on a large scale, the GDPR must be adhered to. Thus, a legal basis is required to process personal data. This can be *consent* of the data subject (*informed consent* Art. $6(1)(a)$ or *explicit consent* in case of health data Art. $9(2)(a)$). Further legal basis can provide a contract with the data subject, a legal obligation, the protection of vital interests, a task carried out in the public interest, or the pursuit of legitimate interests, provided these are not overridden by the data subject's concerns (Art. $6(1)(b)-(f)$). Important to note is, that health data belong to a special category of personal data. Processing of those data is forbidden in the first step, yet Art. $9(2)$ provides specific exemptions. Generally, prior to processing a *purpose* must be determined and any further processing is limited to that purpose (Art. $5(1)(b)$). Moreover, it has to be ensured that only a *minimum amount of data* is processed and stored no longer than necessary (Art. $5(1)(c),(e)$). Nevertheless, these data must be sufficiently *accurate* to ensure that the data subject is correctly represented in the data (Art. $5(1)(d)$). Also, *fair* and *transparent* data processing must be ensured (Art. $5(1)(a)$).

Kamarinou et al. (Kamarinou et al., 2016) identify *transparency* regarding data handling and to whom the data are made available (Art. 12 – 15) as key issues in an ML context. Especially when ML is used for *automated individual decision-making*, full transparency is crucial (Art. 22). Also, the right to erasure (Art. 17) can be challenging in an ML context if data must be forgotten which are part of an ML model.

All in all, the GDPR is well suited for the privacy challenges that ML poses. Nevertheless, technical tools are required to ensure the *integrity and confidentiality* of the data (Art. $5(1)(f)$) and to verify that data are processed in compliance with the GDPR (Kuner et al., 2017). Such technical measures should be integrated into ML systems (Art. 25) (Burri, 2016).

3 REQUIREMENTS

Based on this state-of-the-art review, we derive ten requirements towards a privacy mechanism for ML.

[R_1] Control: Probably the most important feature of a privacy mechanism in general is to give data subjects full control over their data. That is, it has to enable users to specify who is permitted to access which of their data. In the context of ML, however, this is not sufficient. Besides raw data on which ML learns its models, the models themselves have to be considered as well. If a data consumer is not allowed to access certain datasets, this must also apply to all models which were heavily influenced by these datasets.

[R_2] Context: It is not sufficient to control who gets access to the data, but also for what purpose s/he does so. That is, it depends on the context in which data are requested whether a consent or other legal basis is valid. Therefore, a privacy mechanism has to be able to identify not only data consumers but also the circumstances regarding their request for data access.

[R_3] Necessity: Even if a data subject has granted data access for a specified purpose, a privacy mechanism must still become active as it has to be ensured that only a necessary minimum amount of data is shared with the data consumers. This also applies to the ML models. More specifically, the amount of data used to learn the models must also be restricted.

[R_4] Acceptability: Even though a privacy mechanism must ensure data protection matters of data subjects, it also has to respect the rights of data consumers. If, e. g., the processing of certain data is in the public interest or obliged by law, these data must be accessible even without the explicit consent of the data subject.

[R_5] Fair Data Processing: Closely related to legal compliance is that the processing of data must be fair towards all parties involved. Therefore, a data subject must not be able to use the privacy mechanism to manipulate data in order to gain an advantage over other users or the data consumer.

[R_6] Utility: It is also important that a privacy mechanism ensures that the utility of the data (and thus of the models) is always preserved. An arbitrary data manipulation by data subjects might impair the quality of ML models to the point that they make inaccurate predictions. A preferably high accuracy must therefore be ensured by the privacy mechanism.

[R_7] Transparency: To inform data subjects who accessed which data when and for why, and to verify that these accesses were legal, a corresponding logging function must be supported by a privacy-mechanism. In an ML context, this must also be applied to the models, i. e., full transparency must be achieved with regard to the data (sources) on which a model is based.

[R_8] Erasure: According to the *right to be forgotten*, a privacy mechanism must also ensure that certain data can be erased. For ML models, however, this also means that they must be able to "forget" certain aspects if they were learned on these erased data.

[R_9] Security: In addition to privacy issues, a privacy mechanism must also cover security issues, i. e., data must be protected against unauthorized access. To this end, raw data have to be isolated from data consumers and any communication between these two parties has to be secured. In addition, raw data must be protected against manipulation.

[R_{10}] Data Protection by Design: In order to minimize the organizational burden on users and still guarantee full data protection, a privacy mechanism should adopt a Privacy by Design approach, i. e., it has to be integrated seamlessly into the ML system.

It is evident that both, Global Privacy (a global view on all available data sources is required for, e. g., **[R_4]**, **[R_5]**, and **[R_6]**) and Local Privacy (individual privacy rules defined by the user are required for, e. g., **[R_1]** and **[R_2]**), are necessary to meet these requirements.

4 RELATED WORK

In ML systems, three task domains can be identified that are relevant regarding privacy. In the following, we discuss selected representatives for these domains.

Data Preparation. The earliest stage in which an ML privacy mechanism can be applied is during data preparation. That is, similar to Local Privacy, data are masked before being actually processed, in the case of ML, the learning of models. *AVARE* (Alpers et al., 2018) is a privacy system for smart devices. Various privacy filters are applied to available data sources. These filters enable users to specify which data are available to an application. The filters are adapted to the respective kind of data. For instance, the accuracy of location data can be reduced, or certain details of contact data can be hidden. However, AVARE considers each smart device separately. As a result, its privacy settings are unnecessarily restrictive for ML, similar to Local Privacy. *PSSST!* (Stach et al., 2019) therefore introduces a central control instance that knows all available data sources and applies privacy measures according to the privacy requirements of the user. PSSST! conceals privacy-relevant patterns, i. e., sequences in the data from which confidential knowledge can be derived. This pattern-based approach significantly reduces the amount of data masking. To this end, PSSST! needs to know the intended use of the data. In the ML context, however, it cannot be assumed that the usage of the models is known in advance.

Model Management. A privacy mechanism can also come into play when dealing with the ML models (i. e., during learning or provisioning). As in most use cases where statistical information is shared with third parties, *differential privacy* (Dwork, 2006) is often applied in ML. Abadi et al. (Abadi et al., 2016) apply differential privacy to the ML database to ensure that the models do not reveal any sensitive information. As this excludes users, Shokri and Shmatikov (Shokri and Shmatikov, 2015) enable users to select a subset of data to be made available for learning. Data remain on the user's system until they are needed for learning. Bonawitz et al. (Bonawitz et al., 2017) extend this approach by enabling users to provide only aggregated submodels to update a *deep neural network*. This ensures that the ML system never has full data access. However, the result can be biased if each user provides only data that are best for him or her, whereby certain aspects can be completely lost. Moreover, the models are then available to any application. Hüffmeyer et al. (Hüffmeyer et al., 2018) introduce an attribute-based authorization method for querying data from a service platform. This way, applications can be granted access to models only, if they have currently the appropriate attributes (e. g., if they are hosted in the right execution environment).

Explain Models. Also, the explicability of models is relevant for privacy. Alzantot et al. (Alzantot et al., 2019) introduce a technique to explain ML models for image recognition. They define masks to cover image areas that are irrelevant for decision-making. This enables users to comprehend why a decision is made. Yet, this approach is restricted to image recognition. Ribeiro et al. (Ribeiro et al., 2016) introduce a framework which contains various description techniques. Yet, for each ML algorithm a dedicated plugin is required. Thus, it cannot be used universally. Rudin (Rudin, 2019) therefore suggests that instead of trying to describe ML models, rather explainable types of models should be used. However, users cannot choose the types of models freely as they are restricted to the types supported by an ML system. Powell et al. (Powell et al., 2019) propose to combine ex ante (i. e., explain the model) with ex post approaches (i. e., explain the decision). They only consider the logic of a model, but not the input data that lead to a particular decision when the model is applied.

In addition to all the identified shortcomings in related work, there is also no holistic approach providing all required privacy features. Therefore, we introduce our own ML privacy approach called AMNESIA, next.

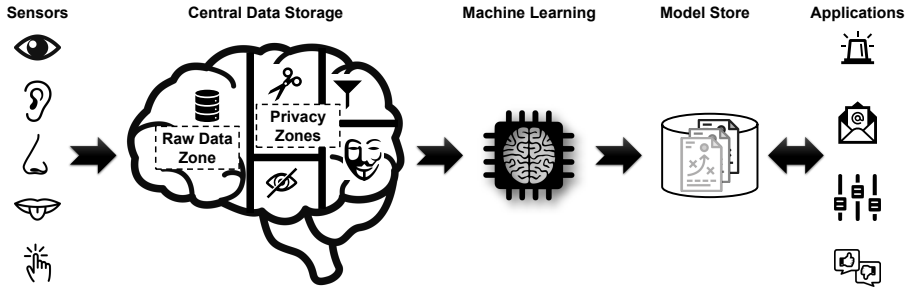


Figure 2: Overall Concept of AMNESIA Introducing its Main Components and Key Functionality.

5 AMNESIA

AMNESIA is a privacy-aware machine learning model provisioning platform. In AMNESIA, certain data can be "forgotten" both, temporarily (i. e., for learning or provisioning a model) as well as permanently (i. e., in the ML database and in all models). For this purpose, we organize the data stock of an ML system in privacy zones. Each zone masks its data in different ways whereby different facets are concealed or revealed respectively. Furthermore, AMNESIA enables users to identify which data were used to learn an ML model and to verify whether their privacy requirements have been met when applying a certain model. AMNESIA adopts a Privacy by Design approach in order to provide a technical implementation of the GDPR for ML systems. It adheres to both, the privacy rights as well as the data provision obligations of data subjects.

In the following, we describe the overall concept of AMNESIA in Section 5.1. Then, in Section 5.2, we discuss its privacy zones. Section 5.3 outlines an implementation strategy for AMNESIA. The essential model management concept is detailed in Section 5.4. Finally, in Section 5.5, we describe how AMNESIA verifies which personal data have been processed.

5.1 Overall Concept

The overall concept of AMNESIA is shown in Figure 2. Here, we use a highly abstracted representation in order to illustrate the basic functionalities in a very accessible way first. Technical details on these functionalities follow in the subsequent subsections.

The data sources—in an IoT context these are predominantly sensors (depicted as sensory organs)—are highly heterogeneous. This heterogeneity affects, among other things, data types, data formats, data accuracy, and sample rate. Also, the value of the data from these sources differs very much in terms of their privacy relevance as well as their utility for ML models. Therefore, a privacy mechanism not only has to be able to deal with this data heterogeneity, but also has to reflect this in its data protection techniques.

The gathered data are forwarded to a central data storage. For this transmission, the data are annotated with metadata. These metadata include, for instance, from which source the data originate, when the data were gathered, the "owner" of the data, and the data subject. These data and metadata have to be stored as raw data for verification purposes.

The central data storage resembles AMNESIA's "brain". An ML system can only learn from data stored there. We organize our central data storage into zones that are dedicated to different privacy levels. In addition to the *Raw Data Zone*, in which the original data are stored, AMNESIA offers four special privacy zones. AMNESIA applies different privacy filters to incoming data, which filter out certain privacy-relevant information without compromising certain data quality aspects, and stores the masked data in the respective zone. In this way, a proper zone can be selected for each use case (i. e., for given privacy requirements and utility requirements). More details on these zones are given in Section 5.2.

The central data storage is secured against unauthorized access. Only the ML component has access to it. In it, any supervised learning algorithm can be deployed. The learned ML models are then stored in the model store. However, instead of learning a single model, AMNESIA learns a model for each of the five zones. These models are annotated with metadata. In addition to information required for the management and provisioning of the models (e. g., the applied ML algorithm), also privacy-relevant information is stored (e. g., from which zone the model stems).

Applications can then query these models. For this purpose, however, they have to authenticate towards the store first. AMNESIA not only identifies applications, but also their current attributes. That way, AMNESIA is able to distinguish, for instance, whether an application is hosted in Europe or in the USA. This affects the applicable data protection laws. If an application has been authenticated, the model store checks

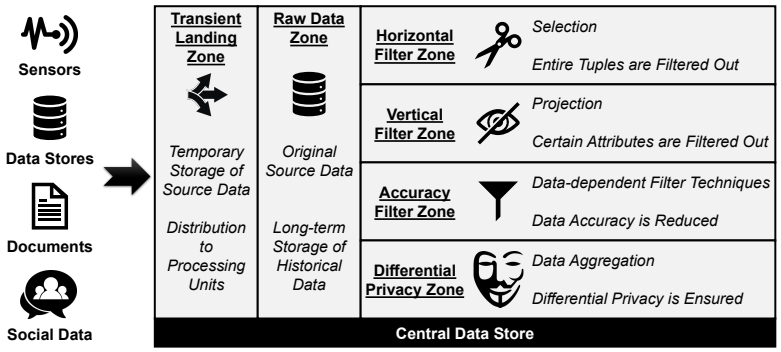


Figure 3: The AMNESIA Privacy Zone Model (cf. (Sharma, 2018)).

which user-defined privacy requirements apply to that application and which utility requirements that application imposes on a model. Based on these two metrics, the model store then selects a matching model and forwards it to the application. More details about the store can be found in Section 5.4.

5.2 Privacy Zones

The central data store of AMNESIA has the same basic premises as a big data system: Large amounts of heterogeneous data have to be stored efficiently and prepared for various purposes. For this, big data systems use *data lakes*. They often organize data into several areas. For instance, there are areas in which incoming data are stored fully unprocessed as raw data, whereas in others only cleansed data or data tailored to a specific use case are stored. In the *zone model*, datasets are maintained in different aggregation and pre-processing stages concurrently in the zones (Giebler et al., 2019).

In AMNESIA, we adopt the zone model introduced by Sharma (Sharma, 2018) and adapt it to our privacy use case (see Figure 3). The model consists of two vertical zones, which are the foundation for all other zones. Additionally, there are four horizontal zones that tailor the data to specific privacy requirements.

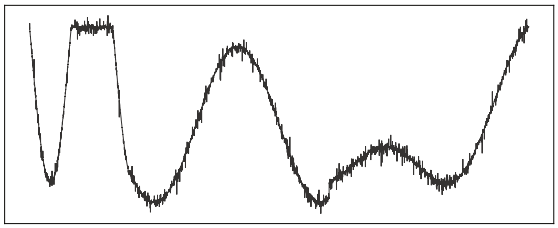
The data processed by AMNESIA are partly volatile, e. g., sensor data. That is, they are sent as a data stream to the central data store and must be processed immediately. Therefore, AMNESIA needs a temporary data buffer to be able to store such data for a short time before forwarding it to all corresponding zones. This is handled by the **Transient Landing Zone**. This zone also annotates the data in order to be able to retrace their origins as well as their chronological order. All data arriving in the Transient Landing Zone are forwarded to the **Raw Data Zone**. In this zone, the data are kept in their original state in terms of quality, accuracy, and completeness. It can thus also be used as a backup if a user reduces his or her privacy requirements, i. e., if more private data should be shared in order to learn more sophisticated ML models. The Transient Landing Zone also forwards the data to four privacy plug-ins, which mask the data and then stores them in the corresponding privacy zone. These privacy zones are described in more detail in the following.

Horizontal Filter Zone. Horizontal filtering is equivalent to a *selection operator* (σ) as defined in relational algebra. A binary predicate expression determines which data tuples of the Transient Landing Zone are included in the Horizontal Filter Zone. That is, data from a certain sensor or data within a given time frame can be filtered out. The included tuples are not masked.

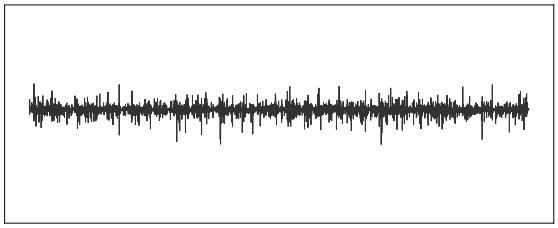
Vertical Filter Zone. Vertical filtering is equivalent to a *projection operator* (π) as defined in relational algebra. Selected attributes are removed from the original set of attributes of the data in the Transient Landing Zone when including them in the Vertical Filter Zone. That is, certain metrics of a sensor (e. g., the blood glucose level) can be filtered out. Apart from that, the data are not masked, and no tuples are entirely concealed.

Accuracy Filter Zone. In the Accuracy Filter Zone, the most versatile filter operators are applied. Depending on the kind of data and the privacy requirements of the users, different filter operators are available to reduce the accuracy of the Transient Landing Zone's data. This can be illustrated by a few simple examples.

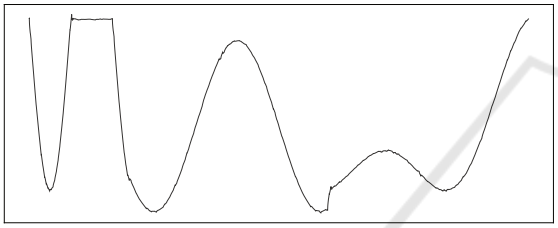
Aggregation operators (e. g., mean or median) can be applied to any kind of numerical data. *Noise* can be added to time series data, such as data from a *continuous glucose monitoring* (*CGM*) sensor. This way, the accuracy of the individual measurement values can be arbitrarily reduced. Yet, characteristics in the value progression can still be recognized. Outliers can also be identified in time series data, i. e., data points with a presumably higher information value, as they indicate deviant behaviors. These data points can be deleted, and the resulting gaps can then be filled by *interpolation* to veil any signs of tampering. *Discrete wavelet transforms* can be applied to time series data as well.



(a) Original Raw Time Series Data.



(b) Using DWT as a High-pass Filter (Details).



(c) Using DWT as a Low-pass Filter (Progression).

Figure 4: Application of a Discrete Wavelet Transform (DWT) in AMNESIA as a High-pass and Low-pass Filter.

They operate as high-pass or low-pass filters. Thereby specific frequencies can be damped to conceal either characteristics in the value progression or details in the measurement values. Figure 4 illustrates how these high-pass and low-pass filters work as well as their impact on the data. In addition, the high-pass and low-pass filters reduce the number of data points, which additionally decreases the accuracy of the data.

Lastly, here is an example of a filter for a specific kind of data, namely location data. For location data we initially choose a random angle α and a random distance d. The target accuracy sets the maximum value for d. The captured position is then shifted by d in the direction of α. For each subsequent position, either a new angle is chosen and the distance is calculated so that the *traveled distance* is preserved, or a new distance is chosen and the angle is calculated so that the *direction of movement* is preserved.

Differential Privacy Zone. While the aforementioned privacy zones are primarily tailored to the privacy requirements of individual users, the Differential Privacy Zone exploits the fact that the central data store contains data of many users. Noise, in terms of *dummy entries*, is added to the database until the differential privacy property is fulfilled. That is, the statistical accuracy of the data in this zone is maintained without being able to identify individual users.

There can be several instances of each privacy zone. As a result, users are enabled to define distinct privacy requirements for each application.

5.3 Implementation

For the implementation of AMNESIA, we use the *BRAID architecture* (Giebler et al., 2018). BRAID is a refinement of the classic *Lambda architecture* (Marz and Warren, 2015). The Lambda architecture enables real-time processing and long-term storage of large amounts of data. To this end, both batch and stream processing engines are required. Contrary to the Lambda architecture however, in BRAID these two engines are not operated decoupled. That means in particular that both, data and (intermediate) results can be mutually exchanged between the engines. This is crucial for AMNESIA as the ML models learned in a batch processing system have to be applied to live data in a stream processing system. The applied models must be exchangeable at runtime if privacy or utility requirements are changed. This is possible in BRAID.

Figure 5 shows our implementation strategy for AMNESIA. The Transient Landing Zone is realized by two components: *Kafka*[1] assigns incoming data to a topic depending on their origin. *Apache Hadoop*[2] receives the data and executes the actual *extract, transform, load* (*ETL*) process. That is, data are annotated based on their topics, prepared for the AMNESIA privacy zones, and forwarded to the respective zone. The instances of the zones are implemented using the *Hadoop Distributed File System* (*HDFS*).

Moreover, the data in each instance (i. e., the Raw Data Zone and the four privacy zones) are encrypted. Only AMNESIA has access to the keys. This prevents unauthorized access on the one hand and enables secure deletion on the other hand. By deleting the respective key, the data are no longer readable and thus "forgotten". Waizenegger et al. (Waizenegger et al., 2017) introduce a hierarchical key management system that can be used for this purpose. In this way, fine-grained secure data deletion is achieved.

For the batch processing of these data, we use *Spark*[3]. The Spark instance gets the required keys from AMNESIA for this purpose. *MLlib*[4] can be used

[1] See https://kafka.apache.org

[2] See https://hadoop.apache.org

[3] See https://spark.apache.org

[4] See https://spark.apache.org/mllib/

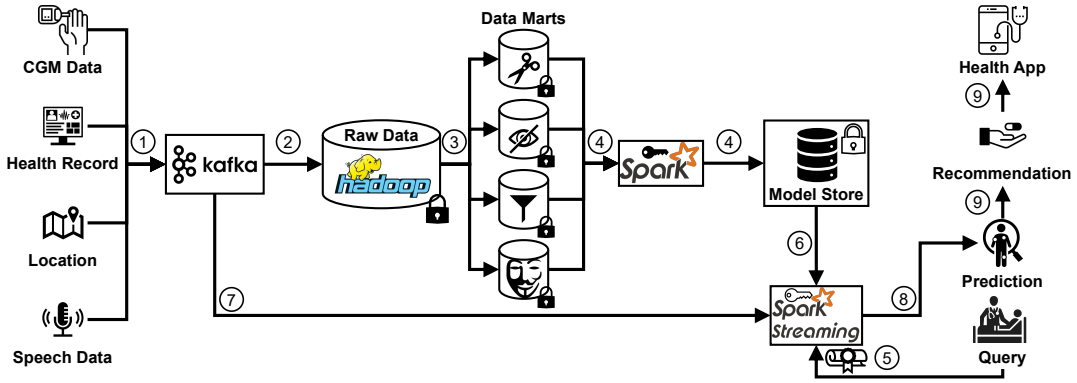


Figure 5: Implementation Strategy for and Operational Flow of AMNESIA.

to learn the ML models. The learned models are then added to our model store. The model store provides data management and provisioning services (see Section 5.4). To this end, the store annotates the models.

Via this annotation it is also possible to retrace on which data a model was learned. If the underlying data are deleted, the associated models can also be deleted and re-learned on the remaining data stock. In order to be compatible to every ML algorithm and model type, a direct modification of the models is not possible. This is due to the fact that the modification of existing models would solely target models that origin from *incremental* or *transfer learning* (Losing et al., 2018). Instead, the models are encrypted, and we use the abovementioned secure deletion method.

Spark Streaming[5] is used to apply a model. The model that is provided to this component is selected according to the privacy and utility requirements. It receives only the key for this single model from AMNESIA. This model is applied to real-time data. The results are provided to authorized application.

Attribute-based approaches are very effective for the fine-grained identification of applications. Thereby, also the context in which an application wants to use a model can be reflected. In addition to the purpose for which the model can be accessed, the context also includes, e. g., the actual location where the application is hosted. However, this approach has two crucial problems: On the one hand this requires a lot of computational power and on the other hand the disclosure of these attributes also poses a privacy threat to the users and providers of the application (Gritti et al., 2018). However, Gritti et al. (Gritti et al., 2019) introduce a lightweight and privacy-preserving attribute-based authentication method that we can apply in AMNESIA.

The interaction of these components as well as the operational flow of AMNESIA can be illustrated by means of a Smart Health example. For the treatment of chronic diseases, such as diabetes, autonomous diagnostic support is particularly useful as a continuous monitoring of health values (e. g., via a CGM sensor) is required (Tamada et al., 2002). While such health data should not be masked, there are other data which are useful for diagnosis but also reveal a lot of private insights. For instance, the location (e. g., via a GPS sensor) affects the stress level of a patient (Knöll et al., 2015) and his or her mood can be determined via a speech analysis (e. g., via the microphone of a smartphone) (Mehta et al., 2012). Such data should not be provided unfiltered to an ML system.

Kafka pools and buffers these data ① and forwards them to Hadoop ②. In addition to storing the raw data, our privacy filters are applied here ③. In order to reduce the accuracy of the location data, the aforementioned technique can be used, since the stress level can also be determined via a coarse location. Yet, simply adding noise to the speech data is not sufficient, as speech recognition is even possible on noisy data (Krishna et al., 2019). Our privacy filters are, however, able to analyze the data. Thus, AMNESIA can determine the mood and store only the mood level and not the captured audio data in the Accuracy Filter Zone. Using Spark and MLlib, models are learned for each zone and then stored in the model store ④.

If a Smart Health application needs diagnostic support, it queries the Spark Streaming component ⑤. This request is signed with the attributes of the application. Depending on these attributes, the users' privacy requirements and the request' utility requirements, an appropriate model is selected ⑥ (see Section 5.4). Kafka provides the streaming component with real-time data ⑦. Depending on the privacy requirements these data can also be filtered. MLlib makes a prediction ⑧ which leads to a treatment recommendation ⑨.

[5]See https://spark.apache.org/streaming/

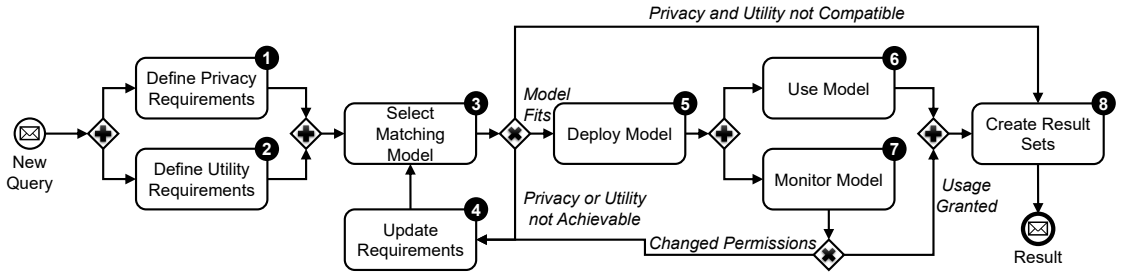


Figure 6: AMNESIA's Process Model for the Management and Provisioning of ML Models (cf. (Weber et al., 2019)).

5.4 Model Management

To realize the AMNESIA model store, two key features have to be considered in particular, namely model management and provisioning. With this in mind, we adapt and extend the process model for maintaining ML models by Weber et al. (Weber et al., 2019).

There are basically two strategies to ensure that a model store is compatible with as many ML tools and algorithms as possible. To this end, the store can operate with an interchangeable model format that is supported by most ML systems. The *Predictive Model Markup Language* (*PMML*)[6] is such an open standard. PMML is supported by many ML systems such as *KNIME*[7]. That is, the ML models learned in KNIME can be exported to the PMML format and PMML models created by other systems can be imported in KNIME to visualize, analyze, and apply them. A big advantage for AMNESIA is that PMML models are already annotated with a lot of metadata (e. g., which data are covered by the model) (Guazzelli et al., 2009). We can extract this information from the PMML file and use it in the model store (e. g., for model selection or provenance analysis).

However, there is a large number of ML systems that support exclusively a proprietary model format. This is where the other strategy comes into play to compensate this disadvantage. ML libraries such as *JPMML*[8] provide converters for model formats of various ML libraries and tools. That is, the models learned in their native environments can be translated to a multitude of other ML libraries and programming languages. In this way, we achieve independence from proprietary model formats. However, each model format must first be determined and an appropriate converter selected from the JPMML library.

In AMNESIA, we therefore combine these strategies, i. e., we provide converters for ML model formats to a generic interchange format. We use PMML for this interchange format due to its metadata support.

The model selection and provisioning process introduced in AMNESIA is shown in Figure 6. When AMNESIA receives a query from an application (given that it has been successfully authenticated), the model store first looks up the relevant privacy requirements depending on the attributes of the application ❶. Via these privacy requirements, a user can specify which application can access which data for which attributes (e. g., for which purpose or in which context). To this end, s/he describes which information (in terms of *private patterns*, i. e., data sequences or combinations) must not be included in the model.

To make the specification of privacy requirements as simple as possible for users, an approach such as *EPICUREAN* (Stach and Steimle, 2019) can be applied. Experts identify which knowledge can be derived from which data sources and map these coherences to private patterns. Expressive keywords are assigned to these patterns. All this information is stored in a knowledge base. Users query the knowledge base in natural language, i. e., the natural language description of their privacy requirements are analyzed using *NLTK*[9]. Afterwards, private patterns with keywords matching to these requirements are recommended to the user. In addition, *collaborative filtering* is used to find further relevant private patterns, i. e., patterns applied by other users who have a similar behavior and sense of privacy. The selected patterns are mapped to appropriate models, i. e., to models learned on data concealing the private patterns in question.

Similarly, application developers define utility requirements, i. e., the quality of the requested model (e. g., in terms of precision and recall) ❷. Developers are legally obligated to such considerations anyway, as they must ensure data minimization when processing private data. This leads to *public patterns*, i. e., information that has to be reflected in a model.

Subsequently, it must be checked whether the private and public patterns are compatible. In addition, it is necessary to verify whether the private patterns and the public patterns are legit from a legal point of view,

[6] See http://dmg.org/pmml/v4-4/GeneralStructure.html
[7] See https://www.knime.com
[8] See https://github.com/jpmml
[9] See https://www.nltk.org

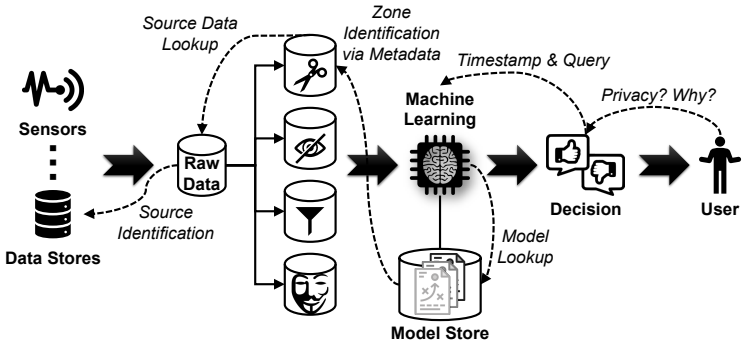


Figure 7: Provenance Information Flow in AMNESIA.

based on the involved kinds of data and the nature of the request ❸. If the patterns are not compatible (i. e., the user wants to hide a fact that the request explicitly requires) or if there is no model that meets these requirements, the requirements must be re-adjusted ❹. If still no model can be found, the application is informed that its request did not lead to a result ❽.

If, however, a fitting model is found, it is deployed ❺. A deployed model can then be applied to real-time data ❻. Concurrently, AMNESIA monitors the model, i. e., it checks whether the users' permissions have changed in the meantime ❼. For instance, a permission change occurs when a user enhances his or her privacy requirements or adds new requirements. Also, if a user exercises his right to erasure, this results in a permission change and a different model has to be selected (or a new model has to be learned based on a different database) ❹, ❸ & ❺. In this way, in AMNESIA even already learned knowledge is immediately "forgotten". However, if the permissions have not changed, AMNESIA provides the results to the querying application ❽.

5.5 Provenance Analysis

Lastly, we also describe the provenance analysis options, provided by AMNESIA. When an ML system, for instance, makes a prediction or a recommendation, two questions often arise for its users. On the one hand, users need to know whether all of their privacy requirements were respected when applying an ML model. On the other hand, users want to know why the decision was made that way. In accordance with the GDPR, providers of ML applications should be able to provide answers to both of these questions.

Figure 7 outlines how AMNESIA addresses these issues. Users can request an explanation regarding an automatic decision from AMNESIA. Based on the timestamp of the query that led to the decision of the ML application, AMNESIA checks which model was applied at that point in time.

For this purpose, the AMNESIA model store has a history management of the models. Over time, the stored models can undergo changes. For instance, if a data subject deletes some of his or her data between the time a decision is made and the time a user requests an explanation for that decision, then all models that include these data have to be re-learned. Therefore, no stored models are permanently deleted. From a data security point of view, this is not critical, since the store is fully encrypted. Due to the attribute-based authentication and access control it is ensured that applications only have access to the latest model versions. Only for AMNESIA's internal provenance analysis, the histories of the models are visible.

Once the right version of the model has been found, i. e., the model the decision is based on, the stored metadata is used to identify the data stock used for learning. Some of these metadata are already embedded in the model itself. PMML provides, e.g., information about the model version, a timestamp, when the model was created, and information about the data used for learning. However, the descriptions given in the *PMML Data Dictionary*[10] are only intended to give an overview of the used data features in order to remain independent of specific data sets.

Therefore, we have an additional metadata repository for our model store. These metadata include links between models and the privacy zone instances on whose data they are based on. By combining these links with the information about the features from the PMML Data Dictionary, a comprehensive representation of the data covered by a model—and thus, responsible for automatic ML decisions—is obtained. By means of the privacy zone it can be verified whether the privacy requirements of the user have been respected.

If a provenance analysis is carried out for a model that has already been deleted, AMNESIA has also deleted the keys for the underlying data. As a result, these data have been irretrievably "forgotten", i. e.,

[10]See http://dmg.org/pmml/v4-4/DataDictionary.html

they are no longer readable. Nevertheless, AMNESIA can still identify which privacy filters have been applied to the data, based on the privacy zone in which the data were stored. This information is sufficient to verify that no privacy requirements have been violated.

For further analyses, the primary keys of the data can be used to identify the underlying source data in the Raw Data Zone. The Raw Data Zone's metadata also allow to trace the origin of the data.

6 ASSESSMENT

In the following, we evaluate whether AMNESIA fulfills the requirements towards a privacy mechanism for ML and if it complies with applicable data protection regulations, i. e., whether it is an appropriate technical solution towards GDPR-compliant ML.

Due to the privacy requirements formulated as private patterns, users have full control which ML models are available for which applications. AMNESIA creates several versions of every model, each based on a different state of the data stock. For this purpose, various privacy filters are applied to the data. Moreover, a data subject can withdraw his or her consent permanently via AMNESIA's secure deletion. This complies with **[R_1]** and Art. 7(3) GDPR.

The attribute-based authentication enables to differentiate application on a fine-grained level. Besides general information about the application, the attributes used here also include information about the execution context and the purpose of use. In this way, AMNESIA can dynamically control which data an application may receive for the given purpose. This complies with **[R_2]** and the principle of purpose limitation.

In public patterns, applications define which utility requirements they pose towards a model (and thus towards the underlying data stock). By harmonizing the private patterns and public patterns, AMNESIA is able to select a model that provides the least insights into private data. This complies with **[R_3]** and supports the principle of data minimization.

Since AMNESIA has a comprehensive overview regarding which application has which utility requirements for what purpose as well as which privacy requirements are defined by a user, legal compliance can be improved. If the request of an application is, e. g., in the public interest, AMNESIA can therefore reduce or ignore the privacy requirements. Fair data processing can also be enhanced as AMNESIA can prevent data concealing or obfuscation that would wrongly benefit a certain user. This complies with **[R_4] & [R_5]**.

Due to the privacy zones, AMNESIA is able to select a filtering technique that results in an ML model with the highest utility. This complies with **[R_6]**.

AMNESIA's provenance analysis enables to verify that the users' privacy requirements and rights have been respected. They also create transparency regarding the data on which a model (and thus indirectly a decision) is based on. This complies with **[R_7]**.

Due to the secure deletion, AMNESIA can enforce the right to be forgotten. Since all data of the data stock is encrypted hierarchically, deleting the corresponding keys ensures that selected data are no longer readable. By means of AMNESIA's hierarchical key management system, data can be deleted on a fine-grained level. As soon as data are deleted from the data stock, AMNESIA marks all models based on this data as invalid. The model management and provisioning process applied in AMNESIA ensures that no invalid model is available to any application. The full encryption of the data stock and the model store also ensures integrity and confidentiality of the data. This complies with **[R_8] & [R_9]**.

AMNESIA combines Local and Global Privacy to meet **[R_1] – [R_9]**. Due to our implementation strategy, which seamlessly embeds AMNESIA into an ML system, Privacy by Design is achieved **[R_{10}]**.

As **[R_1] – [R_{10}]** were dominantly derived from legal considerations, mainly the general data protection principles (Art. 5 GDPR), AMNESIA significantly improves compliance with GDPR and enables innovative ML models at the same time. In order to enforce purpose control for the ML models as well, these models can be tagged with their intended purpose to enable audit trails about their usage (see (Petković et al., 2011)).

7 CONCLUSION

ML can be applied in many domains. This is leveraged by IoT devices permanently recording a large amount of data about the users. With this comprehensive data stock, ML systems are able to learn models. These models enable, e. g., autonomous decision-making. Besides the undeniable benefits of ML, there are also severe privacy issues if personal data is used to learn the models. While the GDPR provides a distinct legal framework, there is a lack of effective technical solutions towards GDPR-compliant ML.

To this end, we introduce AMNESIA, a privacy-aware machine learning model provisioning platform. With AMNESIA it is possible to "forget" private data. This applies not only to the ML data stock, but also to already learned models. Furthermore, it is possible to verify to the user that his or her privacy requirements

have been respected and which data have been used for decision-making. To achieve this, we make five contributions: **(1)** We introduce AMNESIA, enabling GDPR-compliant ML. **(2)** We introduce a privacy zone model for organizing AMNESIA's data stock. **(3)** We introduce a Privacy by Design implementation strategy for AMNESIA. **(4)** We introduce a novel model management concept in AMNESIA. **(5)** We introduce a data provenance analysis in AMNESIA. Evaluation results prove that AMNESIA is a suitable technical solution towards GDPR-compliant ML.

REFERENCES

Abadi, M. et al. (2016). Deep Learning with Differential Privacy. In *CCS '16*.

Alpers, S. et al. (2018). Citizen Empowerment by a Technical Approach for Privacy Enforcement. In *CLOSER '18*.

Alzantot, M. et al. (2019). NeuroMask: Explaining Predictions of Deep Neural Networks through Mask Learning. In *SMARTCOMP '19*.

Bonawitz, K. et al. (2017). Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *CCS '17*.

Burri, T. (2016). Machine Learning and the Law: 5 Theses. In *NIPS '16*.

Chamikara, M. A. P. et al. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. *Pervasive and Mobile Computing*, 48:1–19.

Dwork, C. (2006). Differential Privacy. In *ICALP '06*.

Garfinkel, S. and Lipford, H. R. (2014). Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 5(2):1–124.

Giebler, C. et al. (2018). BRAID — A Hybrid Processing Architecture for Big Data. In *DATA '18*.

Giebler, C. et al. (2019). Leveraging the Data Lake — Current State and Challenges. In *DaWaK '19*.

Gritti, C. et al. (2018). Device Identification and Personal Data Attestation in Networks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 9(4):1–25.

Gritti, C. et al. (2019). Privacy-preserving Delegable Authentication in the Internet of Things. In *SAC '19*.

Guazzelli, A. et al. (2009). PMML: An Open Standard for Sharing Models. *The R Journal*, 1(1):60–65.

Holzinger, A. et al. (2018). Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In *CD-MAKE '18*.

Hüffmeyer, M. et al. (2018). Authorization-aware HATEOAS. In *CLOSER '18*.

Kairouz, P. et al. (2016). Extremal Mechanisms for Local Differential Privacy. *Journal of Machine Learning Research*, 17(17):1–51.

Kamarinou, D. et al. (2016). Machine Learning with Personal Data: Profiling, Decisions and the EU General Data Protection Regulation. In *NIPS '16*.

Knöll, M. et al. (2015). Using space syntax to analyze stress perception in open public space. In *SSS10*.

Krishna, G. et al. (2019). Speech Recognition with No Speech or with Noisy Speech. In *ICASSP '19*.

Kuner, C. et al. (2017). Machine learning with personal data: is data protection law smart enough to meet the challenge? *International Data Privacy Law*, 7(1):1–2.

LeCun, Y. et al. (2015). Deep learning. *Nature*, 521:436–444.

Li, H. et al. (2017). Partitioning-Based Mechanisms Under Personalized Differential Privacy. In *PAKDD '17*.

Losing, V. et al. (2018). Incremental on-line learning: A review and comparison of state of the art algorithms. *Neurocomputing*, 275:1261–1274.

Marz, N. and Warren, J. (2015). *Big Data - Principles and best practices of scalable real-time data systems*. Manning Publications Co.

Mehta, D. D. et al. (2012). Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform. *IEEE Transactions on Biomedical Engineering*, 59(11):3090–3096.

Petković, M. et al. (2011). Purpose Control: Did You Process the Data for the Intended Purpose? In *SDM '11*.

Powell, A. et al. (2019). Understanding and Explaining Automated Decisions. Technical report, SSRN.

Ribeiro, M. T. et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD '16*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature*, 1:206–215.

Sharma, B. (2018). *Architecting Data Lakes: Data Management Architectures for Advanced Business Use Cases*. O'Reilly Media, Inc.

Shirer, M. and D'Aquila, M. (2019). Worldwide Spending on Artificial Intelligence Systems Will Grow to Nearly $35.8 Billion in 2019, According to New IDC Spending Guide. Press release, IDC.

Shokri, R. and Shmatikov, V. (2015). Privacy-Preserving Deep Learning. In *CCS '15*.

Stach, C. et al. (2019). PSSST! The Privacy System for Smart Service Platforms: An Enabler for Confidable Smart Environments. In *IoTBDS '19*.

Stach, C. and Steimle, F. (2019). Recommender-based Privacy Requirements Elicitation – EPICUREAN: An Approach to Simplify Privacy Settings in IoT Applications with Respect to the GDPR. In *SAC '19*.

Tamada, J. A. et al. (2002). Keeping watch on glucose. *IEEE Spectrum*, 39(4):52–57.

Vayena, E. et al. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11):1–4.

Wachter, S. (2018). Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer Law & Security Review*, 34(3):436–449.

Waizenegger, T. et al. (2017). SDOS: Using Trusted Platform Modules for Secure Cryptographic Deletion in the Swift Object Store. In *EDBT '17*.

Weber, C. et al. (2019). A New Process Model for the Comprehensive Management of Machine Learning Models. In *ICEIS '19*.