When Machine Learning Meets Privacy: A Survey and Outlook

BO LIU*, University of Technology Sydney, Australia
MING DING, Data61, CSIRO, Australia
SINA SHAHAM, The University of Sydney, Australia
WENNY RAHAYU, La Trobe University, Australia
FARHAD FAROKHI, The University of Melbourne, Australia
ZIHUAI LIN, The University of Sydney, Australia

The newly emerged machine learning (e.g. deep learning) methods have become a strong driving force to revolutionize a wide range of industries, such as smart healthcare, financial technology, and surveillance systems. Meanwhile, privacy has emerged as a big concern in this machine learning-based artificial intelligence era. It is important to note that the problem of privacy preservation in the context of machine learning is quite different from that in traditional data privacy protection, as machine learning can act as both friend and foe. Currently, the work on the preservation of privacy and machine learning (ML) is still in an infancy stage, as most existing solutions only focus on privacy problems during the machine learning process. Therefore, a comprehensive study on the privacy preservation problems and machine learning is required. This paper surveys the state of the art in privacy issues and solutions for machine learning. The survey covers three categories of interactions between privacy and machine learning: (i) private machine learning, (ii) machine learning aided privacy protection, and (iii) machine learning-based privacy attack and corresponding protection schemes. The current research progress in each category is reviewed and the key challenges are identified. Finally, based on our in-depth analysis of the area of privacy and machine learning, we point out future research directions in this field.

CCS Concepts: • **Security and privacy → Privacy protections**; *Social network security and privacy*.

Additional Key Words and Phrases: machine learning, privacy, deep learning, differential privacy

ACM Reference Format:
Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* 1, 1 (March 2021), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Since Facebook data privacy scandal in 2018 [154], privacy has once again become a dominant feature in people's minds. This motivates revisiting privacy challenges, particularly with the emergence of intelligent technologies thanks to the big data revolution. For example, newly emerged machine learning (ML) techniques, especially the unprecedented powerful deep learning, will have

Authors' addresses: Bo Liu, bo.liu@uts.edu.au, University of Technology Sydney, Australia, , Ultimo, NSW, 2007; Ming Ding, Data61, CSIRO, Australia, ming.ding@data61.csiro.au; Sina Shaham, The University of Sydney, Australia, sina.shaham@sydney.edu.au; Wenny Rahayu, La Trobe University, Australia, W.Rahayu@latrobe.edu.au; Farhad Farokhi, The University of Melbourne, Australia, Farhad.Farokhi@unimelb.edu.au; Zihuai Lin, The University of Sydney, Australia, zihuai.lin@sydney.edu.au.

0360-0300/2021/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

paradigm-shifting impacts on privacy preservation. A critical question that needs to be well investigated is: What are the privacy challenges and solutions associated with ML?

Some initial work has appeared in the literature with an emphasis on mitigating privacy risks during the machine learning process by paying special attention to the privacy challenges and risks associated with the ML models. In this regard, possible attack models [8, 38, 98, 141, 144, 155] have been discussed and protection schemes [2, 12, 118, 126, 140] have been proposed. These works demonstrated both ML models and training datasets can be the target of privacy attacks, leading to sensitive information leakage. Meanwhile, researchers have also tried to use ML for privacy protection. As an example, the authors of [174] have developed a method for automatic recognition of privacy-sensitive object classes and adjust users' privacy preference settings. In addition, there are also several works that develop new privacy protection schemes in the scenarios where ML is used for attacks [80, 81]. Overall, the current research has only scratched the surface, and there are major issues that require further investigation:

- ML could play different roles in a privacy protection problem, e.g., protection target, attack tool, and/or protection tool. It may even play multiple roles in the same problem.
- ML systems and models have different types, each facing different privacy risks and requires different protection schemes.
- There does not exist a unified privacy metric or notion. Although differential privacy (DP) [32] is widely accepted in traditional privacy studies, it still has limitations in the context of ML, especially when considering unstructured data, such as text, image, and video.

In this context, a systematic study of privacy and ML is essential for future research efforts. Although there are several surveys on this topic [1, 64, 85, 177], The focus has been on a certain type of ML model or specific methods.

This study attempts to provide the first comprehensive survey on privacy in ML by investigating different scenarios/applications of privacy and ML. The main contributions of the paper are as follows:

- We divide the works in this area by the different roles of ML, i.e., ML as protection target (private ML), protection tool (ML enhanced privacy protection), attack tool (ML-based attack), and analyze the problems and solutions in each category.
- For private ML, we categorize the attacks and protection schemes and then compare their difference.
- For ML aided privacy protection and ML-based privacy attack, we not only discuss the existing works, but also provide insights on new techniques to achieve privacy preservation.
- The study concludes with a discussion on the directions of future research in ML and privacy.

Through this comprehensive overview, we wish to prepare a solid ground for future research in this field.

The rest of the paper is organized as follows. Section 2 reviews basic concepts of machine learning system and models, and discusses the relationship between privacy and ML. In Section 3, we compare and classify existing privacy attacks and protection schemes in ML systems. Section 4 focuses on ML aided privacy protections, followed by the discussion of ML-based attack and corresponding privacy preservation schemes in Section 5. We present our outlook and propose some future directions for this promising research topic in Section 6. Finally, we conclude our work with a summary in Section 7.

Moreover, the abbreviations used in this paper are listed in Table 1.

Table 1. Summary of acronyms used in the paper.

CNN	convolutional neural network
DNN	deep neural network
DP*	differential privacy
ERM	empirical risk minimization
FGSM	fast gradient sign method
FHE	fully homomorphic encryption
GAN	generative adversarial network
GNN	generative neural network
IoT	Internet of things
ML	machine learning
SGD	stochastic gradient descent
SMC	secure multi-party computation
SVM	support vector machine
VAE	variational autoencoder

* DP in this survey is used as the abbreviation for Differential Privacy, not deep learning.

2 PRIVACY THREATS AND MACHINE LEARNING

In this section, we discuss the privacy threats in the context of machine learning, and further point out various roles of machine learning in the studies of user privacy.

2.1 The Machine Learning System and Models

ML refers to algorithms and statistical models used by computer systems to efficiently perform specific tasks without the use of explicit instructions. It relies on an automated learning process. The ML algorithm constructs a mathematical model of sample data called a "training set" to make predictions or decisions [10].

Depending on if the output is labelled in the training set, ML models can be divided into three different groups: supervised, unsupervised, and semi-supervised. As supervised learning is used by most practical machine learning algorithms, it will be explained here as an example.

A supervised ML model is a parameterized function f_θ that maps input data $\vec{x} \in \mathbb{X}^d$ (generally a vector of features) to output data $y \in \mathbb{Y}$ (label). For a classification problem, $\mathbb{X}^d$ is a d-dimensional vector space and $\mathbb{Y}$ is the set of classes. This function is trained to accurately predict the label of new data that have not seen before.

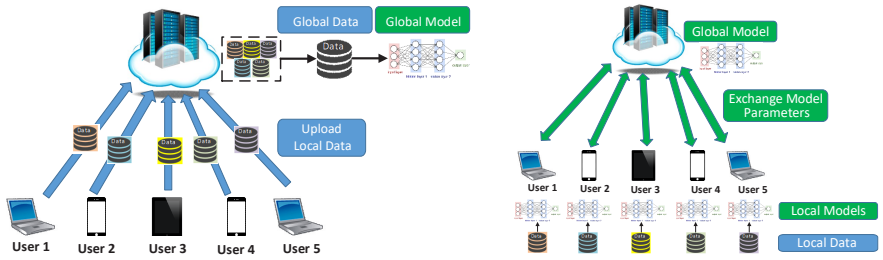
Moreover, we can divide the ML process into two stages:

(1) Model training: The training process of a machine learning model is to find the optimal parameters that can accurately capture the relationship between $\mathbb{X}$ and $\mathbb{Y}$. To achieve this, a training dataset $D = \{\vec{x}_i, y_i\}_{i=1}^N$ with N samples is needed. Then a loss function L is adopted to quantify the difference between two outputs, i.e. the ground-truth one y_i and the predicted one $f_\theta(\vec{x}_i)$. The goal of training a model is to minimize this loss function, i.e.,

$$\theta^\star = \arg\min_\theta (\sum_i L(y_i, f_\theta(\vec{x}_i)) + \Omega(\theta)), \tag{1}$$

where Ω is a regularization term to penalize model complexity and avoid overfitting.

(2) Model inference/prediction: After the model training is completed and the optimal parameters $\theta^\star$ are obtained, given an input $\vec{x}$, the corresponding output can be calculated as $y = f_\theta^\star(\vec{x}_i)$. This prediction process is called inference. We can calculate the prediction accuracy of the model over a testing dataset D_t to measure the model's performance.



(a) centralized learning (can be outsourced learning). (b) distributed learning

Fig. 1. Centralized and distributed ML systems: (a) centralized learning; (b) distributed learning.

Furthermore, according to the architecture of the ML systems, there are two different models, as shown in Fig. 1:

- Centralized learning: The training data is centralized in a machine or in a data center, and the centralized entity trains and hosts the models. For example, a researcher could use a cloud platform, to host datasets and train an AI model based on them. It goes without saying that the availability of all data in such a centralized method leads to high efficiency and accuracy [55]. However, because the centralized operator has direct access to sensitive data, user privacy might be violated.
  As the learning tasks become more and more complicated, many companies start to outsource the training process, i.e., *outsourced learning*, or *ML-as-a-service*. In this case, each user owns his/her training data while the service providers own the models and algorithms. The data holder outsources model creation to a cloud service such as Microsoft Azure ML and Amazon AWS ML, which automate the process of ML. "Users upload datasets, perform training, and make the resulting models available for use" [144]. During this process, the users do not have any understanding of the details of model creation. The "ML provider is the entity that provides ML training codes to data holders" [144].
- Distributed learning: Centralized learning is sometimes not a good option for several reasons: (i) data is inherently distributed in some scenarios; (ii) data is too large to be stored in a single machine; (iii) users are not willing to share raw data; and (iv) users want to train the neural network with different instances to achieve better predication accuracy. In this case, ML can be conducted in a distributed manner, i.e., distributed learning. In general, distributed learning is used in a scenario of distributed training data sources and a centralized server. There are several variations of distributed learning:
  - Collaborative learning: Distributed learning involving such collaborations is known as *collaborated learning*. But the settings could be quite different in the literature. For example, the authors of [145] proposed a collaborative learning framework that trains several classifiers "simultaneously on the same training data" to achieve better performance. On the other hand, in the collaborative learning model defined in [55], each participant uses its device to train a local AI model. It then shares a fraction of the parameters/coefficients of the model with the other users. Service operators can create a composite model by collecting these parameters and achieve almost the same accuracy as a model built using a centralized approach. The collaborative approach is "more privacy-friendly" because the dataset is not directly exposed. Also, if only a small part of the model parameters is shared

and the parameters are truncated and/or obfuscated by DP mechanisms, the model exhibits convergence through experiments [140].

- Federated learning: A popular framework for collaborative learning is *Federated learning* [69] introduced by Google. There are currently two different federated learning settings: cross-device and cross-silo [68]. The cross-device setting normally involves a very large number of mobile or IoT devices, while in the cross-silo settings it "might involve only a small number of relatively reliable clients" [68], e.g., multiple organizations. In a broader definition of federated learning that covers both settings, each device downloads the current model from a centralized server, improves it by learning from data on a local device, and then sums up the changes in a focused update. Here, "focused updates are updates" containing "the minimum information necessary for the specific learning task" [68]. And then the shared model is updated by averaging all users' updates. Since all the training data will not leave local devices, and no updates from individual users are stored in the cloud, the privacy risk has been greatly reduced.
- Split learning: Another collaborative learning framework is *Split learning*, in which each user trains the network up to a certain layer known as the cut layer and sends the weights to server. Mathematically speaking, these weights represent and compress the input data to some intermediate feature vectors. The server then trains the network for rest of the layers, and generates the gradients for the final layer, followed by error back-propagation until the cut layer. The gradient is then passed over to the users. The rest of the back-propagation is completed by the users [159]. In split learning, "client-side communication costs are significantly reduced as the data to be transmitted is restricted to first few layers of the split neural network prior to the split".

Although some collaborative learning models consider shared training data [145], which presents a significant privacy risk. In this survey, however, we consider the case that the local raw training data are not shared with the server or amongst users. In this learning process, the users can collaboratively learn a shared ML model, thus decoupling ML tasks from the storage of the data in a single device.

Overall, centralized learning is characterized by "globally stored data" and "globally trained model", as shown in Fig. 1(a), while the distributed learning is characterized by "locally stored data" and "locally trained model", as shown in Fig. 1(b). Although there will be a global model in distributed learning, it is not trained globally, at least part of the model is trained by individual clients.

2.2 Relationship of privacy and machine learning

In contrast to traditional privacy-related research frameworks, ML techniques open new challenges and opportunities to privacy protection. There has been some initial research embarking on this journey. The existing works can be divided into three categories according to the roles of ML in privacy.

First, making ML system private, i.e., ML system is the target of privacy protection. As shown in Fig. 2(a), this category 1 includes making both the ML system (model parameter) and data (training/test dataset and output data) private, since the privacy threat may happen in any stage of a data cycle, e.g. the training, publishing, or prediction of data. Most of the research in this group relies on the use of differential privacy in ML and deep learning models [44]. For example, Shokri et al. [140] developed a differentially private SGD algorithm and a distributed deep learning model training system. In such way, multiple entities can cooperatively learn a neural network.

Second, using ML to enhance privacy protection. As shown in Fig. 2(b), the privacy protection target is the data in this category 2 and ML is a tool to help privacy protection. For example, Liu et

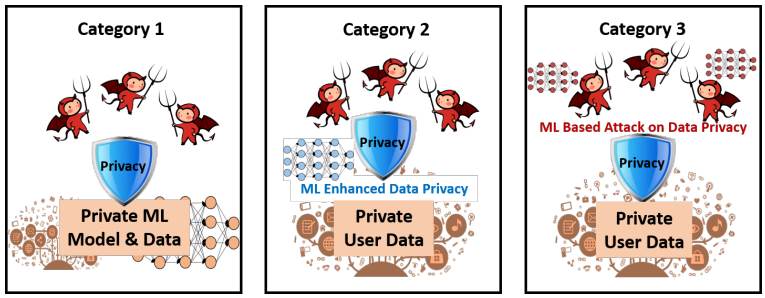


Fig. 2. Three different categories of research problems in privacy and ML: (a) Privacy of ML model and data; (b) ML enhanced privacy protection; (c) ML-based privacy attack.

al. [82] utilized ML to enhance private decision-making experience through ML. Orekondy et al. [114] proposed an approach to categorize personal information in images and predict information leakage directly from images. Yuan et al. [175] presented an ML approach to decide whether to share a picture with a specific requester for a particular context.

Third, ML-based privacy attack, i.e., ML is used as an attack tool of the adversary, as shown in Fig. 2(c). For example, recent researches have shown that deep learning methods can be used to detect object types, people's identities, and landmarks, from images posted on Internet. When the adversaries use this kind of powerful tools, conventional privacy protection methods would be over-powered, especially being challenged by the mighty deep learning tools. There have been very few works in this category. Liu et al. [80] proposed schemes of applying adversarial perturbations images, so that ML systems cannot get private information from them.

Table 2 summaries three categories of privacy protection problems involving ML systems. It is worth mentioning that one technique might belong to more than one category. For instance, ML might be used as attack and protection tools at the same time, which makes the problem more complicated. We will discuss this in more detail in the reminder of the paper.

Table 2. Three categories of privacy protection problems in the context of ML.

Category	Role of ML in Privacy Protection
Private ML	Protection target
ML enhanced Privacy Protection	Protection tool
ML-based Privacy Attack	Attack tool

Fig. 3 summarizes the general taxonomy of the research papers presented in this work. We divide them according to the above mentioned three categories. In each category, we discuss the attack and threat models first and then analyze the works on privacy protection schemes.

3 PRIVATE MACHINE LEARNING

In this section, we will discuss the challenges and existing solutions in privacy preservation in ML, or simply stated, private ML.

We will first discuss attack and threat models, followed by detailed analysis of privacy preservation schemes, along with some comparisons at the end.

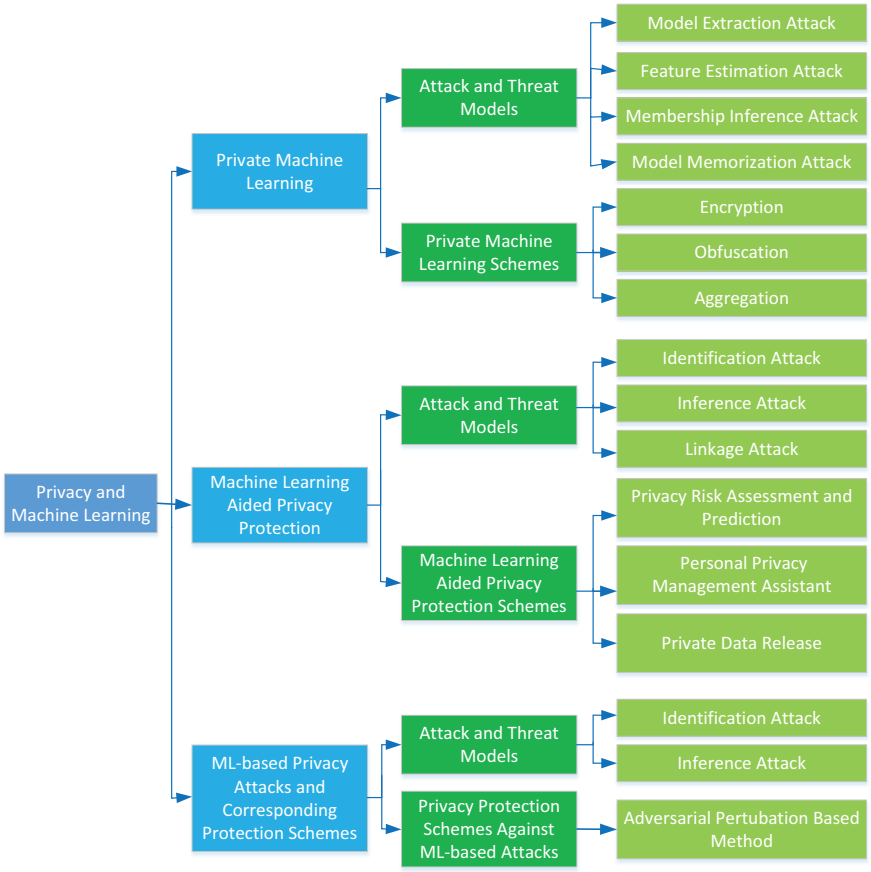


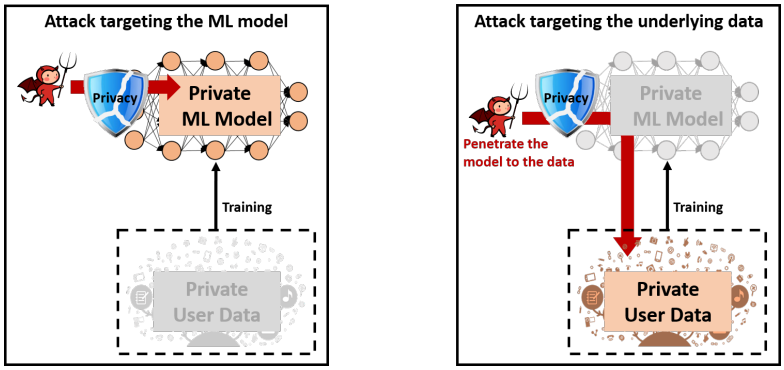
Fig. 3. The proposed taxonomy of privacy and ML.

3.1 Attack and Threat Models

In this subsection, we analyze the attack models from three perspectives: the attack targets, the knowledge of the adversary, and the attack methods.

First, as we can see in Section 2, model and data are two important components in ML that correspond to two different categories of privacy attack targets, as shown in Fig. 4:

(1) Training data privacy: In many cases, a user wants to keep the training data private while using a ML service. For example, for a medical study or a hospital having a model built out of the private medical profiles of some patients. A patient may want to use the model to make a prediction about whether she is likely to contract a certain disease, or the hospital may want to use the model to predict the probability of readmittance. In these cases, the training data is sensitive medical profiles and should not be revealed. Similar cases exist in other areas such as financial records. More specifically, training data privacy includes exact data value, certain features, statistical properties, or membership (whether a certain data is in the training set).
(2) Model privacy: There are also privacy concerns about the ML model including the model parameters, and training algorithms. For example, a financial institution may hold a sensitive



(a) Privacy of the ML model. (b) Privacy of the underlying data.

Fig. 4. Two different types of privacy attack targets in ML: (a) Model privacy; (b) Training data privacy.

model which can accurately predict stock prices or insurance rates. The model is an important commercial and intellectual property. Another example is the commercial ML API services currently provided by Google, Amazon, Microsoft, and other companies. They charge the customers per API access. Revealing their models or algorithms will cause loss of revenue. In summary, the attack target can either be the model structure or parameters.

Second, the adversaries have different levels of knowledge according to their access to the information.

- White-box access: The adversary has access to the trained model, especially the model parameters.
- Black-box access: The adversary is an end-user and is only allowed to query the prediction model on his/her inputs through an appropriate interface.

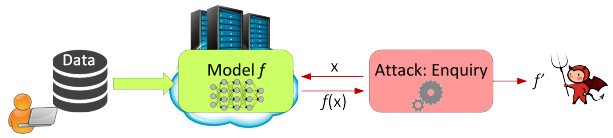
Finally, the adversary can adopt different attack methods. Existing attack methods include model inversion (reverse engineering), shadow training models, and encoding information into models.

Next, we will group existing popular attack models by attack targets and analyze them from the above mentioned three aspects. An illustrative diagram of the attack models is presented in Fig. 5.

3.1.1 Model Extraction Attack. The model extraction attack targets at the duplication of (i.e., "steal") the AI model [155]. The outcome of the attack will be a function f' that is approximately the same as the initial function f. An illustration of such an attack can be found in Fig. 5(a).

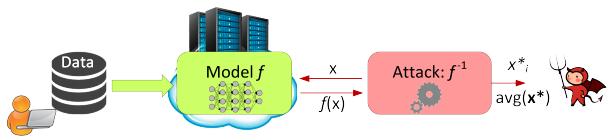
In this attack, the adversary only has black-box access with no prior knowledge of the ML model parameters or training data. Tramèr et al. [155] used a shadow training scheme that can "extract target ML models with near-perfect fidelity for popular ML models" including logistic regression, decision trees, and neural networks, by equation-solving, path finding, or extending the Lowd-Meek approach [90].

There are several other works following this path. Oh et al. [111] built meta-models to extract more model details such as the neural network architecture. Wang et al. [161] designed an attack to steal the hyperparameters of the machine learning model. A hyperparameter is "used to balance the loss function and regularization term in the objective function". The adversary can obtain this value from the training set and model. Hua et al. [56] "investigated reverse-engineering attacks on CNN models exploiting information leaks through memory and timing side-channels".

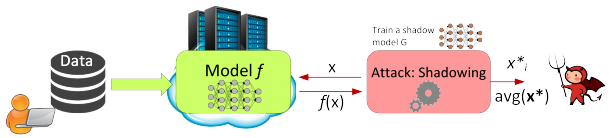


Model Extraction Attack: Adversary learns a close approximation f'(x) of f(x)

(a) Model Extraction Attack

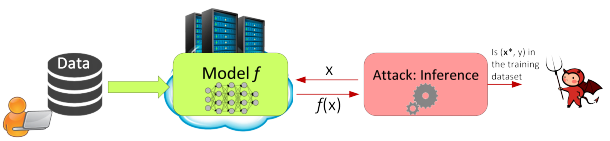


Model Inversion Attack: Adversary learns certain features $x^*_i \in$ **x*** or statistical properties such as avg(**x***) of the training dataset



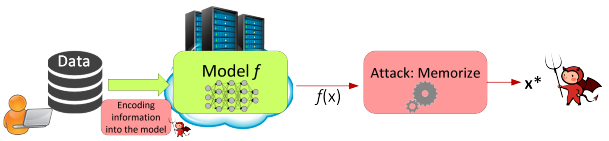
Shadow Model Attack: Adversary learns certain features or statistical properties of the training dataset, with the help of the shadow model G

(b) Feature Estimation Attack



Membership Inference Attack: Adversary learns whether a given data record (**x***, y) is part of the model's training dataset D or not

(c) Membership Inference Attack



Membership Memorization Attack: Adversary recovers exact feature values **x***

(d) Model Memorization Attack

Fig. 5. Different attack models targeting ML.

3.1.2 Feature Estimation Attack. A feature estimation attack aims to estimate certain features $x_i^\star \in \vec{x}^\star$ or statistical properties such as $avg(\vec{x}^\star)$ of the training dataset [38, 39]. In practice, it can be implemented by model inversion attack, shadow model attack or power side-channel attack. An illustration of such an attack can be found in Fig. 5(b).

First, *Model Inversion Attack* mostly works in a white-box model, although it also can use black-box attack [38] with lower effectiveness. Fredrikson et al. [39] showed a white-box attack that can "learn sensitive genomic information about individuals". The basic idea of [39] is to complete the target feature vector "with each of the possible values, and then computes a weighted probability estimate that this is the correct value", given the knowledge of a linear regression model f. Then

in [38] they extended the attack to facial recognition models to achieve two different targets: the *reconstruction attack* that produces "an image of the person associated with a given label" and the deblurring attack that generates the deblurred image of a certain individual given "an image containing a blurred-out face". The idea behind these attacks is "to use gradient descent (GD) to minimize a cost function involving f".

Overall, the model inversion attack works with a simple philosophy: we can reverse-engineer (find f^{-1}) by following the gradient in a trained network to adjust the weights and obtain the features for all classes in the network. Even for classes that we do not have prior information, we can still reproduce the prototype example. This type of attack suggests that any accurate deep learning machine, regardless of training methods, may leak information on the distinguishable classes. Extensive research has shown that generative adversarial network (GAN) generated sample data are similar to the training data. And thus, the results given by the model inversion attack may even "reveal more private information about the training data compared to the average samples" [8].

Second, *Shadow Model Attack* means the attacker trains other ML models to achieve the target. It can happen in either black-box or white-box way. For example, Ateniese et al. [8] designed a "meta-classifier that can be trained to hack into other ML classifiers to infer patterns or private information from the training set", e.g. they were able to extract accent information from trained speech recognition systems.

Hitaj et al. [55] designed an attack in the context of collaborative learning. They consider the adversary is an insider of the collaborative learning process who wants to infer sensitive information from the peers. The adversary can see and use internal parameters of the model, so it is a white-box attack. The adversary uses GANs [45] to extract and reconstruct information of the victim. "This process is similar to facial composite imaging used by the police to identify suspects, where the composite artist generates sketches based on eyewitness identification of the suspect's face. Although the composite artist (GAN) has never seen a real face, the final image is based on eyewitness feedback" [55].

Finally, Wei et al. [164] proposed to use power side-channel attack on an FPGA-based convolutional neural network accelerator, which can successfully recover the input image using the power traces at the inference stage.

3.1.3 Membership Inference Attack. Membership inference attack refers to acquiring the knowledge about whether a certain data record $(\vec{x}^\star, y^\star)$ belongs to the model's training dataset D or not [98, 141]. An illustration of such an attack can be found in Fig. 5(c).

Shokri et al. [141] introduced a "black-box membership inference" that used a shadow training technique to imitate the behavior of the target model. The trained inference model is used "to recognize differences in the target model's predictions" on training and non-training inputs. They also found that overfitting, the structure and type of the model are the main factors that cause a model to be vulnerable to membership inference attack. Long et al. [89] and Yeom et al. [173] investigated "the relationship between overfitting and privacy leakage". Salem et al. [134] proposed a membership inference attack method using an unsupervised binary classification, "which does not need to train any shadow model and does not assume knowledge of model or data distribution".

Membership inference attacks are also studied in Generative Adversarial Networks (GANs). For example, Liu et al. [84] trained an attacker network to launch membership attacks against Variational Autoencoders (VAEs) and GANs. Hayes et al. [52] focused on "generative models in ML-as-a-service applications and train GANs to recognize training inputs".

Melis et al. [98] studied membership inference in collaborative learning. The attack is achieved by "analyzing periodic updates to the shared model during training". The reason that this attack is effective is that the gradients in neural networks are based on features, "thus observations of

the participants' gradient updates can be used to infer the feature values, which are in turn based on these participants' private training data". Wang et al. [163] considered membership inference attack "against the user-level privacy on the federated learning framework by the attack from a malicious server. The proposed attack framework exploits GAN with a multi-task discriminator, which simultaneously discriminates category, reality and client identity of input samples, and doing so recovers user-specific private data".

3.1.4 Model Memorization Attack. Song et al. [144] first proposed the model memorization attack that targets recovering the exact feature values on individual samples. They consider a "malicious ML provider" specialized in model-training for the customers. In such a business model, the provider does not observe the training, but has access to the resulting model. He can steal the sensitive samples and encode the values into the model parameters or outputs. Another malicious party can retrieve sensitive information from the model during model serving. An illustration of such an attack can be found in Fig. 5(d).

Model memorization attack can happen both in white-box and black-box cases. In the white-box case, Song et al. [144] proposed several techniques for the adversary to encode sensitive data into the models. (1) LSB encoding: the adversary can encode the "training dataset in the least significant (lower) bits of the model parameters". (2) Correlated value encoding: the adversary can "gradually encode information while training model parameters". For instance, "the adversary can add a malicious term to the loss function which maximizes the correlation between the parameters and the data he wants to encode". (3) Sign encoding: similar to correlated value encoding, the adversary can use "the sign of model parameters to interpret as bit strings", e.g., positive parameters represent 1 and negative parameters represent 0.

In the black-box case, the adversary is assumed to have no access to the model parameters. They designed a scheme in which the adversary can "augment the training dataset with synthetic inputs whose labels encode the critical information". Then the information is leaked via the outputs of these added inputs.

Model memorization attack studies how malicious training algorithms deliberately create models that leak information about their training data sets. "This threat model is more generous to the adversary, so it can extract more information about the training data than any other attack" [144].

3.2 Private Machine Learning Schemes

In this subsection, we present several private ML schemes, including encryption, obfuscation, and aggregation.

3.2.1 Encryption. Encryption or cryptography-based methods can be divided into two groups:

- Encrypting training data. The mainstream technique is homomorphic encryption. As adding homomorphic encryption to the process will make the process at least an order of magnitude slower, initially it is applied on training data for relatively simple classifiers [13, 15, 47]. For example, Graepel et al. [47] found that training over encrypted data is possible when the training algorithm can be expressed as a low degree polynomial. Bost et al. [13] applied this technique in three classifiers: hyperplane decision, Naive Bayes and decision trees. Then researchers try to extend the work to deep neural networks (DNN). Dowlin et al. [29] proposed CryptoNets which demonstrates how to efficiently convert learned neural networks to make it applicable to encrypted input data. While Hesamifard et al. [54] proposed a framework to train the neural network over encrypted data. Li et al. [77] investigate the case of collaborative learning where datasets are encrypted with different keys, and propose a solution based on multi-key fully homomorphic encryption (FHE).

- Encrypting ML model. The encryption technique is also used to protect the model privacy. Phong et al. [126] proposed to use "additively homomorphic encryption on the gradients". The scheme can prevent information leakage to the "honest-but-curious cloud server" in the condition of collaborative deep learning.

Overall, training neural networks especially DNNs over encrypted data is still challenging. Computational complexity is a major challenge. The network is slow even when trained on plaintext. Adding homomorphic encryption to a process will make it at least an order of magnitude slower. Since the level of the computed polynomial is proportional to the number of backpropagation steps done, the deceleration is more likely to get worse. Another challenging aspect of encryption is the lack of data scientists' ability to examine data and train models, correct mislabelled items, add functionality, and further tune the network [29].

Secure multi-party computation (SMC) is the extension of encryption under the multiparty setting. In SMC, multiple non-colluding parties use a combination of encryption and oblivious transfer to privately finish the computation without seeing the individual components. For ML, it means to compute model updates without having access to both the data and the model.

SMC has been used for a variety of traditional ML models, including decision trees [7], linear regression [30, 66, 107, 135, 136], logistic regression [143, 170], Naive Bayes classifiers [157], and k-means clustering [16, 62].

In general, SMC techniques impose non-trivial computational overheads and their application to privacy-preserving neural networks especially deep learning remains a challenging task. SecureML [101] is a recent example of SMC. It uses "two-party computations to privately train logistic regression models and neural networks".

In summary, SMC based method can cover both data/model privacy concerns, at the cost of communication overhead.

3.2.2 Obfuscation/Perturbation (Differential Private Learning). Obfuscation mechanisms in the context of privacy protection in ML aim at reducing the precision of the data or model. It is can be achieved by adding noises to the model parameters or the original dataset. It is very popular because the DP scheme is usually implemented by obfuscation in practical applications.

The obfuscation can be applied to the model or data. When obfuscation mechanism is for the model, it has another name in the community, i.e., differentially private machine learning. There are some early works on traditional machine learning with differential privacy. For example, Rubinstein et al. [132] proposed differentially-private support vector machine (SVM) learning mechanisms by adding noise to the output classifier and they yield close approximations to the non-private SVM. Chaudhuri et al. [18] provided the model objective perturbation to produce deferentially private empirical risk minimization (ERM) classifier. Song et al. [146] derived differentially private SGD for general convex objectives and validated the effectiveness of the approach using logistic regression for classification. One of the well-known early methods of implementing differential privacy in deep learning is [140]. They trained the ML model "in a distributed manner by updating the selected local gradients and adding noise to them within the privacy budget of each parameter". Based on this work, Abadi et al. [2] introduced "a simpler differential private SGD (DPSGD) algorithm that ensures DP by cutting the gradients to a maximum l_2 norm for each layer". And then add the noise bounded by the "l_2 norm-clipping-bound". It was shown that "high-quality models can be trained through privacy under a moderate privacy budget" with the DPSGD algorithm. In DPSGD, the DP noise is added to the gradients and the whole training process involves multiple iterations. Therefore, it is important to compute the overall privacy loss of the training, i.e, privacy accounting. Although the composition theorem [33] can be used to generate the overall privacy loss, it can be quite loose. Abadi et al. [2] introduced a moments accountant method that can track privacy loss

across multiple training iterations and generate a tighter bound. Another closely related notion is Rényi differential privacy, which "offers quantitatively accurate way of tracking cumulative privacy loss" throughout a multi-round DP mechanisms [100].

Prior to [95], all considered methods used "record-level differential privacy as a framework to protect private information". In many real-world work environments, users have multiple data sources. They may be relevant and should be protected as a whole. Therefore, in some cases, the DPSGD method results in a loss of privacy at a higher level (e.g., user level). McMahan et al. [95] introduced a "user level differential private algorithm called the DP-FedAvg algorithm to protect all the data of a user". Instead of limiting the "contribution of a single record", the DP-FedAvg algorithm limits the contribution of the user data set to the learning model. The DPSGD algorithm was "combined with the FederatedAveraging algorithm" from [14] which uses a server that performs model averaging.

Obfuscation on training data has not been investigated extensively in the context of ML, because it has been deemed as similar to traditional big data privacy. One notable research from Zhang et al. [179] proposed an obfuscate function and applied it to the training data before feeding them to the model training task. This function adds random noise to existing samples, or augments the dataset with new samples. By doing so, sensitive information about the properties of individual samples, or statistical properties of a group of samples, is hidden. Meanwhile, the model trained from the obfuscated dataset can still achieve high accuracy.

Apart from the above-mentioned works, there are other research works in the closely relevant area, such as tensor/matrix factorizations and functional optimization schemes. In more detail, the authors of [59, 60] discussed differentially private algorithms for tensor decomposition, in both centralized and distributed settings [60]. The authors of [40] applied a DP framework in the matrix factorization process with four different possible perturbation: input perturbation, private stochastic gradient perturbation, alternating least squares (ALS) with output perturbation, and output perturbation. The authors of [178] proposed a functional mechanism framework to achieve an ϵ-DP in analyses, which involves solving an optimization problem with a perturbed objective function.

3.2.3 Aggregation. Aggregation is a technique that generally comes along with distributed/collaborated learning, in which multiple parties join a machine learning task while wishing to keep their respective dataset private.

Aggregation can be applied both in and after the training process. It often works together with the encryption scheme (especially SMC) when used during the training process. For example, Pathak et al. [121] proposed an aggregation scheme for independently trained classifiers. They average the parameters using DP and SMC. But they do not consider the accuracy of their approach formally. The first part of later research [140] also focuses on aggregation. They reduce the communication costs and improve the model accuracy by selectively "sharing a subset of parameters in each round of communication".

Another popular framework using aggregation for collaborative learning is federated learning [69, 94] introduced by Google, which has been described before.

Compared with [140], federated learning considers different constraints on the training dataset, i.e., Non-IID, unbalanced, and massively distributed, which is claimed to be more practical in some scenarios such as using mobile devices for the local training.

Federated learning algorithm introduces techniques for quickly and safely aggregating gradients. This scheme focuses on optimizing the communication efficiency of the aggregation process and making the protocol robust against adversaries. However, it lacks guarantees on the amount of user information leakage during training.

Bonawitz et al. [12] enhance the privacy of federated learning by leveraging SMC to compute sums of model parameter updates, i.e., federated Learning with secure aggregation.

On the other hand, using aggregation schemes for privacy protection in ML after the training process, i.e., using ensembles of models is also reasonable. If an ensemble contains enough of models, and each model is trained with disjoint subsets of the training data in a distributed manner, then "any predictions made by most of the models should not be based on any particular part of the training data" [1]. The private aggregation of teacher ensembles (PATE) is based on this idea [116]. In more detail, the ensemble is seen as a set of "teachers" for a new "student" model. The student is linked to the teachers only by their prediction capabilities. And the student is trained by "querying the teachers about unlabelled examples". The prediction result is disjoined from the training data through this process. Therefore the data privacy can be protected. The privacy budget for PATE is much lower than traditional DP ML approaches. But it may not work in many practical scenarios as it relies on an unlabelled public dataset.

Until now, the above works consider aggregation from the perspective of the model. Dwork et al. [34] proposed a scheme that aggregates the prediction output rather than the model. In more details, they partition the dataset D into several subsamples $D_1, \ldots, D_r$ and run a non-private learning algorithm on each of those subsamples to obtain predictors $f_1, \ldots, f_r$, then use a differentially private aggregation technique on values $f_1(x), \ldots, f_r(x)$ and output the result. This subsample-and-aggregate technique is easy to implement as it does not require a new learning algorithm. It focuses on training data privacy via private prediction.

3.3 Summary on Private ML

In this subsection, we sum up the key points on private ML.

3.3.1 Discussions of attack models. We summarize the attack models and related papers in Table 3 and Table 4.

Table 3. Summary of Attack Models.

Adversary features		Model Extraction	Feature Estimation	Membership Inference	Model Memorization
Knowledge	**Black-box**	✓	✓	✓	
	White-box		✓		✓
Target	**Model**	✓			
	Data features		✓		
	Exact data values				✓
	Membership			✓	
Scheme	**Model inversion**		✓		
	Shadow training	✓	✓	✓	
	Encoding				✓

The attack models listed in Section 3.1 are not interdependent. For example, many attacks might be launched on top of the model extraction attack, because it converts the condition from black-box to white-box. Once the black-box attack is finished, the adversary can continue to launch the white-box attack, e.g., a model inversion attack followed by a model extraction attack.

3.3.2 Attack models and protection schemes. Table 5 summarizes the private ML schemes and their effectiveness against different attacks in different situations. Generally speaking, encryption can maintain the adversary's knowledge to a black-box case, thus it is effective to white-box attacks like model inversion attack. Obfuscation [179] influences most attacks as it blurs the information to reduce the privacy risk at the cost of utility. Aggregation is mostly used in distributed systems and often comes along with the other schemes.

Table 4. Comparisons of Attack Methods.

Attack and Threat	ME	FE	MI	MM	Adversary's Knowledge	Attack Method	System Settings
[155]	✓				Black-box	Shadow training	ML-as-a-service
[111]	✓				Black-box	Metamodel	Centralised
[161]	✓				Black-box	Hyperparameter-stealing	Centralised
[56]	✓				Black-box	Reverse-engineering	Centralised
[39]		✓			White-box	Model inversion	Centralised
[38]		✓			Black-box	Model inversion	Centralised
[8]		✓			White-box	Shadow training	Centralised
[55]		✓			White-box	GAN	Distributed
[164]		✓			Black-box	Power side-channel attack	Centralised
[141]			✓		Black-box	Shadow training	Centralised
[84]			✓		White-box	Shadow training	Centralised
[134]			✓		Black-box	Unsupervised binary classification	ML-as-a-service
[52]			✓		White/Black-box	GAN	Centralised
[98]			✓		White-box	Gradient-based	Distributed
[144]				✓	White/Black-box	Encoding	Centralised

ME: Model Extraction; FE: Feature Estimation; MI: Membership Inference; MM: Model Memorization.

Another important question is the relationship of attack models, protection schemes and DP. Among all the mentioned attack schemes, the membership inference attack works along with DP, because the DP definition makes individuals indistinguishable. The other attack models cannot be well countered and evaluated by DP. For example, model inversion uses the output of a model to infer certain features of the hidden input. From a DP perspective, it does not necessarily lead to privacy breaches. For example, in a face recognition scenario, a single person is associated with an output class of the model. As all training images for this class include various photos of the same person, an adversary can orchestra a model inversion attack by creating an artificial image capturing the average information from the person's photos. In most of the cases, this average can be identified as that person. In summary, the average of the features produced by the model inversion can represent the entire output class at most. It does not construct a particular member of the training data set. Moreover, given an input and a model, it determines whether to use that particular input to train the model.

Therefore, model inversion attack is even effective with DP applied collaborative learning [140] and Federated learning [69, 94]. Because DP is being applied to the parameters of the model, and the granularity is set at the record/instance level. However, once the model becomes accurate, it must eventually contain noise added to the learning parameters. Model inversion attack works as long as the model can accurately classify the class and will generate representations of that class. It should be noted that the DP scheme proposed in [140] can only prevent the recovery of specific elements, that is, membership inference attack.

Overall, the DP criterion cannot provide comprehensive privacy evaluation in private machine learning, due to the complexity of the data (unstructured and multimedia data) and privacy protection target (not only membership, but also features of the dataset). Therefore, defining new privacy metrics and criteria is still an open question.

3.3.3 Privacy in Distributed Learning Systems. Training ML in a distributed manner can naturally provide a certain level of privacy protection, as the local training data points are usually not shared among users. Moreover, different privacy protection schemes in centralised learning, such as encryption, perturbation, can be easily extended to the distributed learning settings [169]. In this sense, private ML in distributed systems have a lot in common with that of centralised ML. But there are several special features.

- Distributed ML requires some forms of data sharing among the training nodes because distributed ML is fundamentally different from stand-alone ML. Such shared data, albeit not

Table 5. Comparisons of Private ML Schemes.

Private ML Schemes	ME	FE	MI	MM	Categories	Methods	System Settings
[13, 15, 29, 47, 54, 77]		✓		✓	Encryption	Homomorphic encryption (training data)	Centralised
[126]		✓		✓	Encryption	Homomorphic encryption (model)	Distributed
[7, 16, 30, 66, 101, 157, 170]		✓		✓	Encryption	SMC	Distributed
[132]	✓		✓		Obfuscation	DP SVM	Centralised
[18]	✓		✓		Obfuscation	DP ERM	Centralised
[146]	✓		✓		Obfuscation	DP-SGD for convex objectives	Centralised
[140]	✓		✓		Obfuscation	DPSGD	Distributed
[2]	✓		✓		Obfuscation	DPSGD	Centralised
[100]	✓		✓		Obfuscation	multi-round DP	Centralised
[95]	✓		✓		Obfuscation	DP-FedAvg	Distributed
[179]			✓		Obfuscation	Training data obfuscation	Centralised
[121]	✓		✓	✓	Aggregation/Obfuscation	DP+Aggregation	Distributed
[69, 94]				✓	Aggregation	Federated learning	Distributed
[12]			✓	✓	Aggregation/Encryption	Federated learning + SMC	Distributed
[118]			✓	✓	Aggregation	PATE	Centralised
[34]			✓		Aggregation/Obfuscation	Output aggregation + DP	Centralised

ME: Model Extraction; FE: Feature Estimation; MI: Membership Inference; MM: Model Memorization.

raw data, could take the forms of model parameters, feature vectors, classification results, etc., and such data would still reveal users' privacy from an information theory point of view. Hence, we need to carefully design the data sharing mechanism in distributed ML.

- SMC and aggregation are quite often adopted in the distributed ML systems. However, the above mechanisms are not adequate to protect users' privacy, especially when there exist inside attackers [55, 105].

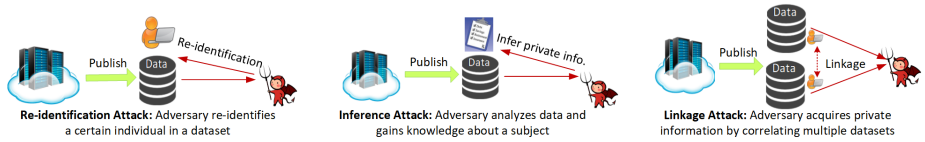
3.3.4 Backdoor attacks and privacy. Some recent work raised the awareness of backdoor attacks against machine learning and deep learning systems, where misclassification behaviours are hidden in models and can be triggered by specific inputs. Gu et al. [49] introduced BadNets that builds a backdoor in DNN models by injecting a square-like trigger with a fixed location to some training data with a target label. Ahmed et al. [133] extended this work by using dynamic trigger patterns and locations. Liu et al. [87] proposed a backdoor attack called the Trojan attack, which reverse-engineers the target model to synthesize training data so that it does not require access to the original training set. Yao et al. [172] proposed a latent backdoor attack method in which they embed the backdoors in teacher models to survive the transfer learning process. In general, current backdoor attacks are mostly considered to be security risks, e.g., it may cause various severe consequences in critical ML applications like autonomous driving. But we can also expect potential privacy risks in the future, for example, backdoor attacks against authentication systems that might enable an adversary to access sensitive information.

4 MACHINE LEARNING AIDED PRIVACY PROTECTION

In this section, we will focus on the case that ML is used to help privacy protection. We will first discuss traditional data privacy risks and threats. These threats have existed for a while, but the newly emerging ML gives us new tools to combat them.

4.1 Attack and Threat Models

Along with the proliferation of the mobile network, people spend more and more time on the Internet, using web-based applications, mobile applications and social networks. These all pose privacy risks. For example, online photo sharing has become more popular than any time before. Users are increasingly sharing their images on various social media, such as Facebook, Google+ and Flickr. Shared images can reveal sensitive information about people and their surroundings [148, 176]. Consider a person sharing a photo of a family gathering. Not only this photo can expose the



(a) Re-identification Attack (b) Inference Attack (c) Linkage Attack

Fig. 6. Different privacy attack and threat models.

people who may or may not wanted to be in the picture, but it can also reveal sensitive information about the family such as religious beliefs, traditions, and food habits. Therefore, sharing photos online can severely violate privacy and disclose sensitive information [37].

Major traditional privacy attacks include identification attacks, inference attacks, and linkage attacks, as shown in Fig. 6.

(1) Identification attack: Identification attack identifies a user's name or identity-based on some public dataset [76]. It is also called re-identification [53, 58] when anonymisation is reversed. Such kind of attack is illustrated in Fig. 6(a).
(2) Inference attack: This type of attack aims at "analyzing data in order to illegitimately gain knowledge about a subject" [70]. Such an attack is illustrated in Fig. 6(b).
(3) Linkage attack: The adversary aims to achieve a target's information by correlating multiple data sources. For example, Narayanan et al. [104] showed that an adversary "can identify a subscriber's record in the Netflix Prize dataset", linking it to an Internet Movie Database. Such an attack is illustrated in Fig. 6(c).

4.2 Machine Learning Aided Privacy Protection Schemes

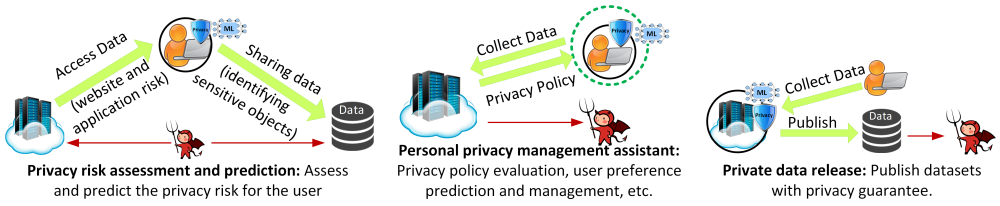
Many privacy protection schemes have been introduced. Obfuscation/perturbation [31, 140], anonymization [5, 6], reducing information sharing [82, 142], and cryptographic mechanisms [43, 127] are the major technologies.

However, the traditional privacy protection schemes focus on structured data, such as an entry in the databases [162]. With the introduction of new applications such as Internet of Things (IoT) and vehicular networks, both the volume and the complexity of the data is increasing. Traditional protection schemes cannot handle all cases and it also becomes more difficult for both common users and even data curators to understand the risk, select correct schemes and manage their privacy.

Under these circumstances, ML has been introduced to enhance privacy protection during the past few years. The efforts including research in several aspects.

- Privacy risk assessment and prediction: Assess and predict the privacy risk for the user during the processes of "access" and "sharing". As shown in Fig. 7(a), ML is used to evaluate both the input and output data streams to find the risk and then privacy protection schemes can be deployed accordingly.
- Personal privacy management assistant: This includes privacy policy evaluation, user preference prediction and management, as shown in Fig. 7(b).
- Private data release: Publish datasets with privacy guarantee. The schemes are generally adopted by data curators rather than an individual user, as shown in Fig. 7(c).

4.2.1 Privacy Risk Assessment and Prediction. The privacy risk exists either when the user is just accessing the application (passively collected information by malicious attackers) or sharing on



(a) Privacy risk assessment and prediction (b) Personal privacy management assistant (c) Private data release

Fig. 7. ML-aided privacy protection schemes.

social networks (actively sharing information). In both cases, ML can help to prevent the loss of sensitive information. An illustration of such a defence mechanism is shown in Fig. 7(a).

***Website and application privacy risk prediction*:** ML can make browsing the websites safer. The proposed browser extension in [137] collects information about websites that users visit and provides feedback to users based on ML to let them know the privacy quality of the site. Manek et al. [92] proposed a method based on a Bayesian classifier to detect and identify websites that can be malicious or threatening to the privacy of users. The proposed approach analyzes online reviews written for websites to decide whether they are reliable or not.

The work in [42] uses an SVM classifier to rate the privacy risks of applications. The results indicate that privacy risks can be identified with over 90% accuracy. Understanding the privacy risks of mobile phone applications with the aid of ML have been considered in [9, 46].

***Identifying sensitive information when sharing*:** Identifying sensitive information in multimedia data has been difficult in the past. With the help of the state-of-the-art ML techniques, users can prevent loss of their personal information while sharing their photos on social media.

Squicciarini et al. [147] considered visual-content features and images' metadata to develop and contrast several learning models. The ML models can classify the photos and evaluate the degree of sensitivity so as to make the decision based on past decisions of the users. Yu et al. [174] proposed a tool called "iPrivacy (image privacy)" to reduce the burden of specifying privacy setting by users when they are sharing photos online. iPrivcy utilizes ML to automate the process. It finds privacy-sensitive objects from images and classifies them according to their privacy sensitivity. Based on the classification, iPrivacy notifies the users if there are objects, which should be suppressed/masked due to privacy concerns before sharing. Moreover, iPrivacy provides privacy settings recommendation based on user preferences and shared images. Orekondy et al. [113] proposed the first large-scale private images dataset, with pixel and instance level annotations. And they proposed the first model to automatically redaction various private information. Hasan et al. [51] proposed a method to automatically identify bystanders "solely based on the visual information present in an image".

4.2.2 Personal Privacy Management Assistant. As the user connectivity increases and web applications become ubiquitous, the responsibility of privacy management transfers more and more to individuals. Unfortunately, given the complexity of the environments and the lack of awareness about privacy attacks by adversaries, it is improbable that the users can manage and fine-tune their privacy preferences correctly [93]. Therefore, there is an immediate need to develop automated privacy management systems to help users in protecting their privacy. An illustration of such a defence mechanism is shown in Fig. 7(b).

The authors of [3] indicate that users continuously modify their privacy requirements to reach their expected level of privacy, and also, appropriately change their privacy preferences. Moreover,

mobile and web applications are attempting to customize their services according to individual preferences to grant personalized experience to customers. Such a customized service results in potential risk for the users [122]. This evidence points to the fact that it is crucial to develop assistants to help users with the management of their privacy configurations. ML can be an invaluable asset in this regard. For example, it can help users to manage their privacy configurations and reduce the burden of time and human resources required to ensure the preservation of privacy.

We have divided the applications of ML for privacy management in two broad categories: (i) privacy policy evaluation, and (ii) user preference prediction and management.

***Privacy policy evaluation*:** Users are usually prompted to agree with the provider's privacy policies when almost using any software and web applications. Privacy policies provide complete information on the collection, storage, and sharing of personal data. Therefore, they are critical to the privacy of users. Unfortunately, most of this information is written using technical jargon and challenging to read terms. Hence, most of the readers prefer to accept the policy unconditionally without thoroughly realizing the consequences [24]. To help users with the decision making, Costante et al. [25] developed a system to evaluate the completeness of privacy policies based on preferences of the users. The system uses natural language processing to analyze and verify the existence of the privacy measures that users specify, and also, assess the level of completeness. Nugent et al. [108] graded the privacy policies that the users encounter based on factors such as security, cookies, and purpose which helps users to check the results and identify if their desired privacy requirements are satisfied. Tesfay et al. [153] proposed an ML approach to "summarize the long privacy policies" into a short paragraph so that it is readable and understandable for users. Shayegh et al. [139] considered methods to improve the privacy notices given to users in IoT networks. With the aid of ML, the authors extract notice and choice statements from the privacy policies for IoT devices, so as to help users to better understand the implications of privacy notices. Lebanoff et al. [73] investigated automatic detection of vague contents on privacy policies and used GANs to characterize the vagueness of sentences.

***User Privacy Preference Prediction and Management*:** Another difficulty in user privacy protection is caused by the fact that each user has a different privacy sensitivity and preference. Nowadays, applications often provide many functionalities with different levels of privacy guarantees. While installing the applications, users are usually prompted for permissions to access resources that have an impact on their privacy. It is important that the users can well coordinate their own privacy preference with the actual privacy risk.

ML techniques are implemented to predict user privacy preferences and help decision making. It was initially proved feasible as some early studies found that user privacy preferences are related to some statistical and environmental parameters. For example, the quantitative research in [166] uncovered that a significant number of users would rather prevent at least one permission request involved in the study. Also, several works have shown that the context of the applications is highly related to user privacy preferences [112, 165]. Lee et al. [75] surveyed 172 participants and uncovered contextual factors that violate the privacy of users in IoT.

Based on the contextual factors and features, ML models can be developed to predict user privacy preferences and take privacy management decision. Mehrpouyan et al. [97] used openness, conscientiousness, neuroticism, extroversion, and agreeableness as inputs to ML models to predict desired users' preferences. Das et al. [26] generated ML models of people's privacy preferences and expectations.

Wijesekera et al. [165] proposed a run-time permission system to infer privacy requirements of users automatically. The proposed system grants the resource allocation permission based on the type of the application requesting the permissions, the request time, and in what circumstances it is requested. Liu et al. [82] investigated ML to enhance privacy decision-making experience.

The results show that providing users with "recommendations based on clusters of like-minded users and using predictive models of people's privacy preferences work to the users' satisfaction". Wijesekera et al. [166, 167] built a classifier to work as a middle-man and make privacy decisions on behalf of users. The classifier adjusts and preserves privacy by changes that happen in the context predicated on the past behaviors of the users.

Orekondy et al. [114] proposed a method named "Visual Privacy Advisor" that "extends this concept to image" contents. They classify "personal information in images into 68 attributes and train models that directly predict such information from images". A user study has been done to understand the privacy preferences with respect to these attributes. They also proposed models that "predict user specific privacy score from images". Yuan et al. [175] presented an ML approach to decide whether to share a picture with a specific requester at a particular context, and if yes, at which granularity.

4.2.3 Private Data Release. Database release is currently an important process in data analytic applications. Different entities generate different types of data, e.g., health data from medical centers. Then, such data will be transmitted to data custodians such as government agencies. Then, the data custodian maintains a platform that organizes, stores and provides data access to data consumers, such as other government departments, individuals, analysts, etc. Privacy preservation processing is highly required when the data custodians release the data. An illustration of such a defence mechanism is shown in Fig. 7(c).

A frequently used traditional private data release mechanism is obfuscation by adding noise to the original dataset. Whereas the ML techniques provide a new solution to this problem, i.e., using a generative neural network (GNN) or generative adversarial network (GAN) [45] to generate synthetic dataset [4].

Although the technique of GNN itself has existed for a while, using it for private data release has just been linked to privacy preservation very recently. Denton et al. [27] used the GAN framework in the context of image processing to generate natural synthetic images. Gregor et al. [48] introduced a model called "Deep Recurrent Attentive Writer (DRAW)" to create synthetic images. The principal idea of the approach is to use two recurrent neural networks as encoder and decoder trained end-to-end with SGD. Vinyals et al. [160] proposed a generative model predicated on recurrent neural network architecture. The approach combines the natural processing ML tools with computer vision for the generation of natural scenes. Using generative models has also been considered for the generation of audios. Oord et al. [158] introduced a DNN model to produce raw audios and applied the approach to "text-to-speech and validated by human listeners for natural sounding". A modified version of the proposed model is used for singing synthesis in [11]. Kulkarni et al. [71] created spatiotemporal trajectories in large scale by training the models based on realistic data, and then, creating synthetic data using the trained models. The authors investigate the utility-privacy trade-off of the approach by experiments. Ouyang et al. [115] proposed a non-sequential non-parametric generative model for spatiotemporal trajectories. The authors generate "synthetic data by training a generative adversarial neural network, which can learn geographic patterns". Liu et al. [86] aim at the addition of geo-privacy protection layer for publication of spatiotemporal datasets based on synthetic trajectory generation. Choi et al. [23] proposed an approach for the generation of synthetic patient records based on GANs and autoencoders. In this work, the performance of the proposed generative model is examined by comparing the generated synthetic patient records with the real data. Cheung et al. [21] used GNNs for the transformation of sensitive images so that they can preserve privacy of individuals. The authors focus on the generation of synthetic facial images and how they can be used for classification of actual images. Zhang et al. [180] proposed a novel approach based on GNNs to increase privacy of users while releasing semantic rich data

such as text, image, and video. Triastcyn et al. [156] used GAN to generating artificial data that retain statistical properties of the real data while reducing the risk of information disclosure. Sun et al. [149] proposed GAN-based head inpainting obfuscation technique to preserve the identity of users when sharing their photos online. Huai et al. [57] considered the differentially private release of crowdsourcing data. They proposed the PrisCrowd approach "in which the data collector learns about underlying patterns of the data and then samples a set of candidate synthetic data from the learned density. The synthetic data are subjected to a privacy test and the ones that pass will be released".

Overall, the latest deep learning techniques show the ability to synthesize fake dataset that is statistically similar to the original one. This technique can be used for private data release. Fig. 8 presents the generative model framework used for privacy preservation of rich semantic data. The process can best be explained by an example. Consider a clinical data sharing scenario, in which the data curator instead of directly releasing the data, trains a deep generative model using the original data in a differentially private manner, and then publishes synthetic dataset generated by the model. In a more general case, the data curator may publish the deep generative model from which "an unlimited amount of synthetic data for arbitrary analysis tasks" can be produced [171]. The use of generative models can significantly increase the privacy of users as the training process of the models can be conducted based on synthetic data instead of the real data belonging to individuals. Meanwhile, the utility of the dataset can be guaranteed as the statistical similarity of models trained based on synthetic data and realist data has been shown repeatedly in the literature. For example, Park et al. [120] proved the statistical similarity of the generated synthetic tabular data and original data. Xu et al. [171] developed training deep neural networks for the generation of synthetic data that closely resemble the actual medical records of patients.

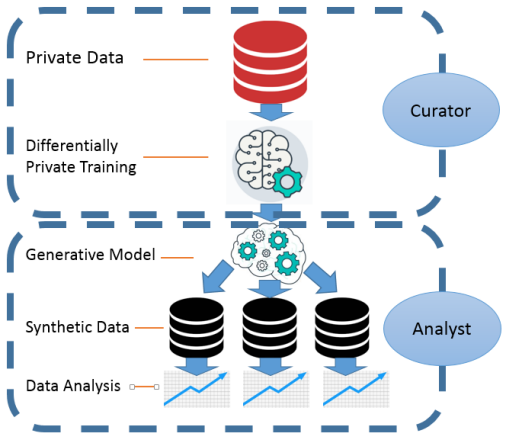


Fig. 8. Privacy preserving framework based on generative model approach.

Although research in GNNs for privacy preservation is in its initial stages, the outlook of the approach is promising. Generation of synthetic data is particularly crucial as traditional methods such as anonymity and obfuscation are ineffective for privacy preservation of semantic-rich data. Moreover, this approach is not associated with the drawbacks of other traditional anonymization approaches such as having background knowledge or linking the data to other sources.

4.3 Summary on ML-aided Privacy Protection

The three different groups of ML aided privacy protection schemes introduced in Subsection 4.2 work in various stages of privacy protection. Privacy risk assessment and prediction is a pre-process before privacy protection, that identifies what do we need to protect. Personal privacy management assistants help to improve access control over sensitive information. Private data release can be applied directly to the data. These protection schemes do not have a one-on-one relationship with the attack models listed in Subsection 4.1. They can be effective against multiple attack models and will work best if combined correctly in specific scenarios.

The two main types of ML models used for privacy protection are classification and object detection. Classification is used for privacy risk prediction and assessment. Object detection is used for identifying sensitive information. Additionally, schemes discussed in 4.2.1 do not directly provide privacy protection. They are currently playing a supporting role, and other subsequent privacy protection schemes are still needed.

GNN opens a new direction for privacy protection research, especially for unstructured data such as image and video. But it is still challenging, as there are no unified metrics for privacy measurements in those complicated cases.

5 MACHINE LEARNING-BASED PRIVACY ATTACKS AND CORRESPONDING PROTECTION SCHEMES

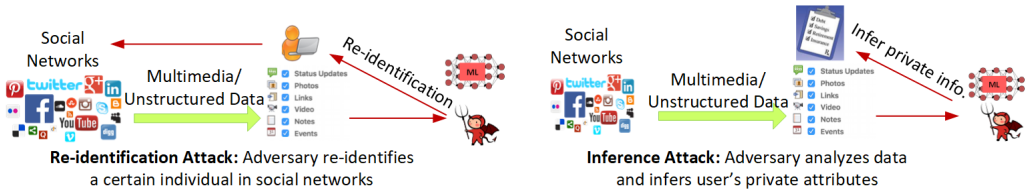
Besides serving as a privacy protection tool, ML can also be used as an attack tool. It urges us to revisit the definition and scope of privacy. In particular, the emerging deep learning technique can "automatically collect and process millions of photos or videos to extract private/sensitive information from social networks" [80]. Traditional privacy-preserving methods are over-powered when combating deep learning tools. It is time to seriously discuss new threats and corresponding solutions.

5.1 Attack and Threat Models

The riskiest personal information leakage source is the social network. While there are a variety of social network platforms enriching people's interactivity and relationship, the shared posts including check-ins, activities, thoughts (tweets, status updates, etc.), pictures, videos often come along with sensitive information. The information poses high privacy risks and they are likely to hand over their privacy unintentionally. A growing number of companies and start-ups specialize in analyzing shared pictures on social media to exploit them for commercial purposes or selling them to other companies. Therefore, the most advanced DNNs have been used to launch privacy attacks.

For example, the adversary can use geo-location information to initiate a localized attack that focuses on finding the position and time information of the person. Gu et al. and Mahmud et al. [50, 91] showed a dangerous attack that is designed to "find important locations such as homes and workplaces". There have been some researches discussing the home location identification problem, either based on the "content of the posts" [20], or the "geo-tags in the check-ins" [22]. And "the research shows that the identification accuracy might be over 90% in many cases" [81].

Besides the simple location information, multimedia data poses more risk under the attack of ML tools. Companies apply advanced DNNs to cluster photos or infer preference of users to facilitate marketers to send targeted ads [99, 130]. DNNs are considered one of the most practical tools in ML as they take advantage of efficient training algorithms and large datasets which enables them to outperform other existing ML techniques. The power of such ML tools has become a problem itself



(a) Identification attack (b) Inference attack

Fig. 9. Different privacy attack and threat models when ML is used as the attack tool.

that may compromise the privacy of photos once they are shared on social media and a challenging problem that needs to be addressed.

The privacy of sensitive data, photos and videos become more crucial in IoT networks, as users might not even be aware of their information such as pictures and videos being recorded. For instance, areas controlled under surveillance cameras can severely compromise user privacy as people lose control of how their photos and videos are being captured and managed. It is likely that the surveillance system applies techniques such as face recognition and detection to identify the users without their permission. Pew Internet survey in 2014 reported that over 91 percent of participants "strongly agree" or "agree" that "they have lost their control over how their personal information is being collected and used by companies" [17].

Major ML attack models include re-identification attacks and inference attacks, as shown in Fig. 9. These attack models are different from those described in Section 4 in the sense that ML is used as an attack tool here.

- The re-identification attack can be launched by face recognition techniques. The recent advance in DNN makes it more harmful from two aspects. First, the process becomes automatic with high accuracy [67, 151, 168]. Second, traditional protection schemes such as obfuscation no longer work effectively [96, 109]. An illustration of the re-identification attack can be found in Fig. 9(a).
- Inference attack has also become more powerful when equipped with ML. ML classifiers can be used to infer a target user's private information (e.g., location, occupation, hobby, political view) from its public data (e.g., twitters, movie rating scores) [20, 22]. Moreover, a series of research work have demonstrated how the advanced artificial neural networks can be used as an adversarial tool to detect sensitive information in images, including people's age [63], relationship [150] and vehicle license plates [181] from ordinary or even obfuscated images. An illustration of the inference attack can be found in Fig. 9(b). Therefore, it is quite urgent to accelerate the research on privacy protection schemes against ML aided attacks.

5.2 Protection Schemes Against ML-based Attacks

There has been some preliminary research in this area. For privacy protection against traditional ML attack, Liu et al. [81] designed community-based information sharing scheme that changes the overall spatial and temporal features so that the clustering-based privacy attack [83] no longer works.

The problem becomes more challenging when deep learning is involved. The solutions may come from a better understanding of deep learning itself. Some researchers recently found that there are limitations to deep learning. Specifically, "it is proved to be vulnerable to some well-designed inputs termed *adversarial examples*" [36, 138]. Szegedy et al. [152] first discovered that the superposition

of "imperceptible noise onto the original image" would mislead DNNs to the wrong classification. Then, Goodfellow et al. [45] proposed the "fast gradient sign method (FGSM) that can be used to generate this type of adversarial examples". Other algorithms to generate such noise can be found in [72, 103, 131].

According to [119], the primary reason for why neural networks are vulnerable to adversarial examples is the linear nature of the neural networks. The authors formalize the space of adversaries against DNNs, which are mostly originated from ML techniques itself. In simple words, ML is used as a tool to breach the ML classifiers. Kurakin et al. [72] focused on adversarial training and how they can be scaled to large datasets. Sharif et al. [138] proposed an algorithm for manufacturing adversarial examples based on ML to disable DNN detection systems from finding objects in shared photos. Additionally, a significant point about adversarial examples is its transferability property [45]. It means that if they are able to fool one model, they are often likely to mislead another model with a different set of parameters and architecture [152]. This is even true if the other model is trained on a different training set or model [117]. This leads to the idea of universal perturbation [102, 128]. It is even possible to "generate adversarial examples that fool both human and computer alike". Elsayed et al. [35] exploited ML to construct adversarial examples that transfer from models created based on computer vision to the human visual system. The authors generated adversarial examples without utilizing the parameters of the model's architecture, and then mimic the visual processing of humans using ML.

Enlighted by the idea of adversarial examples, researchers started to focus on the generation of adversarial examples based on ML to improve the privacy of users against attacks mostly based on DNNs. Liu et al. [88] proposed an algorithm that is against automatic detection using adversarial examples based on the "Faster RCNN framework". Jia et al. [65] proposed a two-phase framework called AttriGuard to defend against attribute inference attacks launched by a classifier. Liu et al. [80] investigated schemes for using adversarial examples in ML systems so that they cannot identify the sensitive information from images. Oh et al. [110] set up a game-theoretical framework and studied the effectiveness of adversarial image perturbations for privacy protection. Li et al. [78] proposed to use adversarial perturbation for face de-identification. Friedrich et al. [41] proposed a privacy-preserving shareable representation of medical texts for a de-identification classifier.

5.3 Summary on Privacy Protection against ML

Previously, the common understanding of privacy protection is to prevent human adversaries from knowing some sensitive information about people. For example, obfuscating faces in images is a well-researched topic. However, the situation has dramatically changed recently. First, the growth of data volume has reached a point where it is physically impossible for anyone to browse everything with their eyes. Second, as a result, people increasingly rely on machines with advanced algorithms to extract relevant information of interest. Third, the booming of ML open source community makes ML tools easy to be obtained by anyone. This brings up a new problem, that is, it is now possible to automatically process data to infer sensitive user information, such as personal identity, social relationships, location, and context. Indeed, ML has recently been used by malicious parties as an efficient tool to launch new types of privacy attacks, especially for social media data. Therefore, we would expect that privacy protection against machines is as important as privacy protection against humans.

ML-based privacy attacks are more challenging to defend against, due to three main reasons. First, the average user is not aware of the capability of state-of-the-art ML methods in extracting personal information. Second, privacy in some contexts such as multimedia data is not obvious. Third, privacy threats also arise from organizations and government sectors that collect and analyze

data on a large scale. Therefore, we need to prevent ML algorithms from automatically mining private information, either intentionally or unintentionally.

In summary, privacy protection against the fast-evolving ML techniques is the most challenging task among all three categories we discussed in the paper. The methodology is to exploit the weakness and limitations of ML methods. Although there have been some initial solutions to this problem using adversarial machine learning, there are still many research problems that require further investigations.

6 OUTLOOK AND FUTURE DIRECTIONS

Significant previous work focuses on making ML algorithms differentially private to preserve the privacy of training sets. However, we should be aware that machine learning, as a whole, also provide potent tools for privacy research (not just for the training datasets), both from attack and defense perspectives.

6.1 Perturbation in Deep Learning

The goal of perturbation in deep learning is to train a model while ensuring DP concerning information about individual training examples. Theoretically, the noise can be added to either the input data, the model parameters (through gradient updates), or the model output. In practical, the majority of work proposed to inject noise into gradients. The main disadvantage of this group of methods is that amount of injected noise is dependent on the number of training epochs, and it potentially can accumulate too much noise due to the significant number of parameters.

Directly adding noise to input data is an option, but it is similar to a typical big data privacy problem and does not closely related to deep learning. Output perturbation and objective perturbation seem to be reasonable directions in the future.

Output perturbation adds noise to the output of the ML system, e.g., the logits at the prediction stage. This method is fast and easy to implement. However, it can suffer from degradation from an attack of repeated querying by an adversarial. Therefore, it is important to restrict the number of queries [129]. One potential solution is to use output perturbation in certain intermediate outputs, such as the teacher voting output in PATE frame work [116].

Objective perturbation is one of the most effective methods for differential privacy ML. This technique adds a random linear term to the objective function. Objective perturbation has been extensively studied in convex optimization. Recently, Iyengar et al. [61] has provided a practical algorithm for differentially private convex optimization, which is a big step towards practical deployment of this technique. Moreover, Neel et al. [106] has extended this approach to non-convex optimization problems. Despite the success in traditional ML, applying objective perturbation to deep neural network is still challenges due to several obstacles: 1) the sensitivity calculation is difficult because the objective functions of deep learning models are mostly non-convex and do not have closed-form expressions; 2) the privacy guarantee is implicitly based on the rank-one assumption on the Hessian of the loss, which is difficult to verify; 3) the privacy guarantee holds only at the exact minima (at least the approximate minima as proposed in [61]) of the optimization problem, which is hard to be guaranteed in practical deep learning systems. One possible solution is to use a convex approximation of the loss function [124]. However, the approximation error might outweigh the reduced perturbation due to smaller sensitivity. It is expected to see more effective methods following this path.

Moreover, instead of perturbing the final output, it is also possible to add noise to the middle layers of the neural networks. Lecuyer et al. [74] proposed the PixelDP framework that includes a DP noise layer in the DNN. Although the purpose of PixelDP is "to increase robustness to adversarial examples", the idea can be further investigated to serve for privacy preservation. For example,

PixelDP scheme enforces that the output prediction function is DP provided the input changes on a small number of pixels (when the input is an image). Potential extensions to PixelDP include: 1) enforcing DP for given different input samples so that it can provide privacy preservation for the training set against membership inference attacks; 2) adding DP noise to the hidden layer of an autoencoder. With the post-processing property of DP, the output of the autoencoder remains to be DP as well. This idea is briefly mentioned in [74]. But we can further explore it in different applications. For instance, we can protect a social network image by generating a perturbed version using this autoencoder with a DP guarantee.

6.2 Defending ML-based Privacy Attack: Adversarial Examples

As we have discussed in Section 5, when ML is used as a privacy attack method, adversarial examples become a powerful way of privacy protection. Despite the preliminaries work on this topic, there are several issues that need to be solved:

- Adversarial example generation methods fall into two categories of attack scenarios: white-box and black-box. The research of using an adversarial example for privacy protection usually assumes that the deep learning model is known, using the white-box setting. In practice, the black-box scenario seems to be a more realistic assumption, e.g., the latest black-box adversarial generation methods such as ZOO [19], $\mathcal{N}$ attack [79] and AdvFlow [28], could be potentially used for privacy protection.
- It is still hard to evaluate the effectiveness of this mechanism with respect to privacy and utility. The existing works use the change of ML outputs (labels) to evaluate the privacy protection methods. We need to prompt more concise and better evaluation metrics.
- There have been some recent research works that connect the DP framework and adversarial example [74]. The PixelDP algorithm [74] proposed to add a DP-noise to the input or any middle layer to the network's architecture to provide guaranteed robustness against adversarial examples., In more details, if we consider "a DNN's input (e.g., images) as databases in DP parlance, and individual features (e.g., pixels) as rows in DP", randomizing the output prediction function to enforce DP can guarantee the robustness of predictions against adversarial examples. PixelDP cannot effectively preserve privacy in the training set as the input changes are restricted to "a small number of pixels" [74]. Phan et al. [125] proposed a heterogeneous Gaussian Mechanism (HGM) that can preserve DP in training data and provide provable robustness against adversarial examples at the same time. They further proposed the stochastic batch mechanism in [123] that can retain higher model utility and is more scalable to large DNNs and datasets, compared with HGM. Overall, the interplay among DP, adversarial example and certified robustness would be a very interesting future topic.

6.3 ML-aided Privacy Protection: GAN and VAE

Excessive amounts of unstructured data including images, videos, audios and texts are being generated constantly and are being used by the government and a wide range of industries. According to the projections of the international data corporation, unstructured data will constitute approximately 80 percent of worldwide data by 2025. Unstructured data, especially image and videos, often containing rich personal information, play a key role in the future privacy preservation ecosystem. And the problem of private data release for unstructured data will be a hot topic in the future.

We expect GAN to play an important role in this area, as it has demonstrated the capability to preserve high utility for ML algorithms while protecting sensitive information in the dataset. Moreover, GAN, as part of VAE, might also be used for privacy protection for a signal data entry

(i.e., an image). In this case, we can encode an original data entry and then decode it with some additional privacy protection.

7 CONCLUSION

This study surveys the literature on privacy in the context of machine learning. By classifying the existing research into three groups: (i) private machine learning, (ii)machine learning aided privacy protection, and (iii) privacy protection against machine learning attack, we comprehensively review the state-of-art techniques on this topic and draw several conclusions as follows.

- The private machine learning problem has drawn the most attention recently. In this category of research works, many try to use the differential privacy criterion during the analysis. However, DP notation cannot provide comprehensive privacy evaluation due to the complexity of the data and privacy protection target. Therefore, how to define new privacy metrics and notations is still an open question.
- The research on machine learning aided privacy protection is gaining momentum these days. For example, using GNN to generate synthetic datasets opens the new direction for privacy protection research, especially for unstructured data such as image and video.
- Research on protection schemes against ML-based privacy attack is in its infancy. But it is expected to fly in the future due to the proliferation of AI techniques in every corner of the future networks. Currently, mainstream technology in this category is the adversarial example/perturbation technique.

We believe our timely study will shed valuable light on the research problems associated with privacy and machine learning. With the increasing attention paid to this topic, we would expect to see increasing research activities in this area.

REFERENCES

[1] Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Nicolas Papernot, Kunal Talwar, and Li Zhang. 2017. On the Protection of Private Information in Machine Learning Systems: Two Recent Approches. In *Proceedings of IEEE Computer Security Foundations Symposium (CSF'17)*. 1–6. https://doi.org/10.1109/CSF.2017.10

[2] Martín Abadi, H. Brendan McMahan, Andy Chu, Ilya Mironov, Li Zhang, Ian Goodfellow, and Kunal Talwar. 2016. Deep learning with differential privacy. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'16)*. 308–318. https://doi.org/10.1145/2976749.2978318

[3] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514. https://doi.org/10.1126/science.aaa1465

[4] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. 2019. Differentially Private Mixture of Generative Neural Networks. *IEEE Trans. Knowl. Data Eng.* 31, 6 (2019), 1109–1121. https://doi.org/10.1109/TKDE.2018.2855136

[5] Charu C Aggarwal. 2005. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases (VLDB'05)*. 901–909.

[6] Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. 2005. Anonymizing tables. In *International Conference on Database Theory*. Springer, 246–258.

[7] Rakesh Agrawal and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD international conference on Management of data (SIGMOD'00)*. ACM Press, New York, New York, USA, 439–450. https://doi.org/10.1145/342009.335438

[8] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks* 10, 3 (2015), 137–150. https://doi.org/10.1504/IJSN.2015.071829

[9] Vitalii Avdiienko, Konstantin Kuznetsov, Alessandra Gorla, Andreas Zeller, Steven Arzt, Siegfried Rasthofer, and Eric Bodden. 2015. Mining apps for abnormal usage of sensitive data. In *Proceedings International Conference on Software Engineering (ICSE'15)*, Vol. 1. 426–436. https://doi.org/10.1109/ICSE.2015.61

[10] Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[11] Merlijn Blaauw and Jordi Bonada. 2017. A neural parametric singing synthesizer. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH'17)*, Vol. 2017-Augus. 4001–4005.
[12] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'17)*. 1175–1191. https://doi.org/10.1145/3133956.3133982
[13] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. 2015. Machine Learning Classification over Encrypted Data. In *Proceeding of The Network and Distributed System Security Symposium (NDSS'15)*.
[14] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'17)*.
[15] Justin Brickell, Donald E Porter, Vitaly Shmatikov, and Emmett Witchel. 2007. Privacy-preserving remote diagnostics. In *Proceedings of the 14th ACM conference on Computer and communications security (CCS'07)*. 498–507.
[16] Paul Bunn and Rafail Ostrovsky. 2007. Secure two-party k-means clustering. In *Proceedings of the 14th ACM conference on Computer and communications security (CCS'07)*. 486–497.
[17] Pew Research Center. 2014. Public Perceptions of Privacy and Security. *Pew Research Center* (2014). http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/
[18] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, Mar (2011), 1069–1109.
[19] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. 15–26.
[20] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of ACM international conference on Information and knowledge management (CIKM'10)*. 759–768.
[21] Sen-Ching Samson Cheung, Herb Wildfeuer, Mehdi Nikkhah, Xiaoqing Zhu, and Waitian Tan. 2018. Learning Sensitive Images Using Generative Models. In *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP'18)*. 4128–4132.
[22] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD'11)*. 1082–1090.
[23] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. *Proceedings of the Machine Learning for Healthcare Conference* (2017). http://proceedings.mlr.press/v68/choi17a.htmlhttp://arxiv.org/abs/1703.06490
[24] Elisa Costante, Jerry Den Hartog, and Milan Petkovic. 2011. On-line trust perception: What really matters. In *Proceedings of the 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST'11)*. 52–59.
[25] Elisa Costante, Yuanhao Sun, Milan Petkovic, and Jerry Den Hartog. 2012. A machine learning solution to assess privacy policy completeness (short paper). In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'12)*. 91–96. https://doi.org/10.1145/2381966.2381979
[26] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. 2018. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing* 17, 3 (2018), 35–46.
[27] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS'15)*, Vol. 2015-Janua. 1486–1494.
[28] Hadi M Dolatabadi, Sarah Erfani, and Christopher Leckie. 2020. AdvFlow: Inconspicuous Black-box Adversarial Attacks using Normalizing Flows. *arXiv preprint arXiv:2007.07435* (2020).
[29] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*, Vol. 1. 342–351.
[30] Wenliang Du, Yunghsiang S Han, and Shigang Chen. 2004. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *SIAM Proceedings Series*. 222–233. https://doi.org/10.1137/1.9781611972740.21
[31] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.
[32] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the theory of cryptography conference (TCC'06)*. Springer, 265–284.
[33] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.

[34] C Dwork and V Feldman Theory. 2018. Privacy-preserving Prediction. *Proceedings of the Machine Learning for Healthcare Conference* (2018). http://proceedings.mlr.press/v75/dwork18a.html

[35] Gamaleldin F Elsayed, Nicolas Papernot, Shreya Shankar, Alexey Kurakin, Brian Cheung, Ian Goodfellow, and Jascha Sohl-Dickstein. 2018. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems (NIPS'18)*, Vol. 2018-Decem. 3910–3920.

[36] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2015. Fundamental limits on adversarial robustness. *Proceedings of International Conference on Machine Learning (ICML'15), Workshop Deep Learning* (2015), 1–7.

[37] Jesse Fox and Jennifer J Moreland. 2015. The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances. *Computers in Human Behavior* 45 (2015), 168–176. https://doi.org/10.1016/j.chb.2014.11.083

[38] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'15)*. ACM, 1322–1333. https://doi.org/10.1145/2810103.2813677

[39] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Security Symposium (USENIX'14)*. 17–32.

[40] Arik Friedman, Shlomo Berkovsky, and Mohamed Ali Kaafar. 2016. A differential privacy framework for matrix factorization recommender systems. *User Modeling and User-Adapted Interaction* 26, 5 (Dec 2016), 425–458.

[41] Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2020. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'20)*. 5829–5839. https://doi.org/10.18653/v1/p19-1584

[42] Hao Fu, Zizhan Zheng, Sencun Zhu, and Prasant Mohapatra. 2019. Keeping Context in Mind: Automating Mobile App Access Control with User Interface Inspection. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'19)*. 2089–2097. https://doi.org/10.1109/INFOCOM.2019.8737510

[43] Keke Gai, Meikang Qiu, Hui Zhao, and Jian Xiong. 2016. Privacy-Aware Adaptive Data Encryption Strategy of Big Data in Cloud Computing. In *Proceedings of 3rd IEEE International Conference on Cyber Security and Cloud Computing (CSCloud'16)*. 273–278. https://doi.org/10.1109/CSCloud.2016.52

[44] Ian Goodfellow. 2018. Defense Against the Dark Arts: An overview of adversarial example security research and future research directions. *arXiv:1806.04169* (2018). arXiv:1806.04169 http://arxiv.org/abs/1806.04169

[45] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS'14)*, Vol. 3. 2672–2680.

[46] Alessandra Gorla, Ilaria Tavecchia, Florian Gross, and Andreas Zeller. 2014. Checking app behavior against app descriptions. In *Proceedings of the 36th International Conference on Software Engineering (ICSE'14)*. 1025–1035.

[47] Thore Graepel, Kristin Lauter, and Michael Naehrig. 2012. ML confidential: Machine learning on encrypted data. In *Proceedings of the International Conference on Information Security and Cryptology (ICISC'12)*. 1–21.

[48] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, Vol. 2. 1462–1471.

[49] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733* (2017).

[50] Yulong Gu, Yuan Yao, Weidong Liu, and Jiaxing Song. 2016. We know where you are: Home location identification in location-based social networks. In *Proceedings of the 25th International Conference on Computer Communications and Networks (ICCCN'16)*. 1–9. https://doi.org/10.1109/ICCCN.2016.7568598

[51] Rakibul Hasan, David Crandall, and Mario Fritz Apu Kapadia. 2020. Automatically Detecting Bystanders in Photos to Reduce Privacy Risks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'20)*. 318–335. https://doi.org/10.1109/SP40000.2020.00097

[52] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership Inference Attacks Against Generative Models. In *Proceedings on Privacy Enhancing Technologies (PETS'19)*. 133–152. https://doi.org/10.2478/popets-2019-0008

[53] Jane Henriksen-Bulmer and Sheridan Jeary. 2016. Re-identification attacks—A systematic literature review. *International Journal of Information Management* 36, 6 (2016), 1184–1192.

[54] Ehsan Hesamifard, Hassan Takabi, Mehdi Ghasemi, and Rebecca N. Wright. 2018. Privacy-preserving Machine Learning as a Service. In *Proceedings on Privacy Enhancing Technologies (PETS'19)*. 123–142.

[55] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Deep Models under the GAN: Information leakage from collaborative deep learning. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'17)*. ACM, 603–618. https://doi.org/10.1145/3133956.3134012

[56] Weizhe Hua, Zhiru Zhang, and G Edward Suh. 2018. Reverse engineering convolutional neural networks through side-channel information leaks. In *Proceedings of the Design Automation Conference (DAC'18)*, Vol. Part F1377. 1–6. https://doi.org/10.1145/3195970.3196105

[57] Mengdi Huai, Di Wang, Chenglin Miao, Jinhui Xu, and Aidong Zhang. 2019. Privacy-aware Synthesizing for Crowdsourced Data. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Vol. 2019-Augus. International Joint Conferences on Artificial Intelligence Organization, California, 2542–2548.

[58] Abou-el-ela Abdou Hussien, Nermin Hamza, and Hesham A Hefny. 2013. Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. *Journal of Information Security* 04, 02 (2013), 101–112. https://doi.org/10.4236/jis.2013.42012

[59] Hafiz Imtia and Anand D. Sarwate. 2018. Improved Algorithms for Differentially Private Orthogonal Tensor Decomposition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. IEEE, 2201–2205. https://doi.org/10.1109/ICASSP.2018.8461303

[60] Hafiz Imtiaz and Anand D. Sarwate. 2018. Distributed Differentially Private Algorithms for Matrix and Tensor Factorization. *IEEE J. Sel. Topics Signal Process.* 12, 6 (2018), 1449–1464. https://doi.org/10.1109/JSTSP.2018.2877842

[61] Roger Iyengar, Joseph P. Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. 2019. Towards practical differentially private convex optimization. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*., Vol. 2019-May. IEEE, 299–316. https://doi.org/10.1109/SP.2019.00001

[62] Geetha Jagannathan and Rebecca N Wright. 2005. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD'05)*. 593–599. https://doi.org/10.1145/1081870.1081942

[63] Amirhossein Jahanbekam, Christian Bauckhage, and Christian Thurau. 2010. Age recognition in the wild. In *Proceedings of the International Conference on Pattern Recognition (ICPR'10)*. 392–395. https://doi.org/10.1109/ICPR.2010.104

[64] Zhanglong Ji, Zachary C Lipton, and Charles Elkan. 2014. Differential Privacy and Machine Learning: a Survey and Review. *arXiv:1412.7584* (2014). http://arxiv.org/abs/1412.7584

[65] Jinyuan Jia and Neil Zhenqiang Gong. 2018. AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning. In *Proceedings of the 27th USENIX Security Symposium (USENIX'18)*. 513–529.

[66] Qi Jia, Linke Guo, Zhanpeng Jin, and Yuguang Fang. 2018. Preserving model privacy for machine learning in distributed systems. *IEEE Trans. Parallel Distrib. Syst.* 29, 8 (2018), 1808–1822. https://doi.org/10.1109/TPDS.2018.2809624

[67] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. 2015. Person recognition in personal photo collections. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR'15)*. 3862–3870.

[68] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2019. Advances and Open Problems in Federated Learning. (2019). arXiv:1912.04977 http://arxiv.org/abs/1912.04977

[69] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv:1610.05492* (2016). http://arxiv.org/abs/1610.05492

[70] John Krumm. 2007. Inference attacks on location tracks. In *International Conference on Pervasive Computing*. Springer, 127–143.

[71] Vaibhav Kulkarni, Natasa Tagasovska, Thibault Vatter, and Benoit Garbinato. 2018. Generative Models for Simulating Mobility Trajectories. *arXiv:1811.12801* (2018). http://arxiv.org/abs/1811.12801

[72] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2019. Adversarial machine learning at scale. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*.

[73] Logan Lebanoff and Fei Liu. 2018. Automatic Detection of Vague Words and Sentences in Privacy Policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3508–3517. https://doi.org/10.18653/v1/D18-1387

[74] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. IEEE, 656–672.

[75] Hosub Lee and Alfred Kobsa. 2017. Privacy preference modeling and prediction in a simulated campuswide IoT environment. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*

(PerCom'17). 276–285.

[76] Huaxin Li, Haojin Zhu, Suguo Du, Xiaohui Liang, and Xuemin Sherman Shen. 2018. Privacy leakage of location sharing in mobile social networks: Attacks and defense. *IEEE Trans. Dependable Secure Comput.* 15, 4 (2018), 646–660. https://doi.org/10.1109/TDSC.2016.2604383

[77] Ping Li, Jin Li, Zhengan Huang, Tong Li, Chong Zhi Gao, Siu Ming Yiu, and Kai Chen. 2017. Multi-key privacy-preserving deep learning in cloud computing. *Future Generation Computer Systems* 74 (2017), 76–85. https://doi.org/10.1016/j.future.2017.02.006

[78] Tao Li and Lei Lin. 2019. AnonymousNet: Natural face de-identification with measurable privacy. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'19)* 2019-June (2019), 56–65. https://doi.org/10.1109/CVPRW.2019.00013

[79] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. 2019. NATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML'19)*. 3866–3876.

[80] Bo Liu, Ming Ding, Tianqing Zhu, Yong Xiang, and Wanlei Zhou. 2019. Adversaries or allies? Privacy and deep learning in big data era. In *Concurrency Computation*, Vol. 31. Wiley Online Library, e5102. https://doi.org/10.1002/cpe.5102

[81] Bo Liu, Wanlei Zhou, Shui Yu, Kun Wang, Yu Wang, Yong Xiang, and Jin Li. 2017. Home location protection in mobile social networks: a community based method (short paper). In *Proceedings of the International Conference on Information Security Practice and Experience (ISPEC'17)*. Springer, 694–704.

[82] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, Tom H Luan, and Haibo Zhou. 2016. Silence is Golden: Enhancing Privacy of Location-Based Services by Content Broadcasting and Active Caching in Wireless Vehicular Networks. *IEEE Trans. Veh. Technol.* 65, 12 (2016), 9942–9953. https://doi.org/10.1109/TVT.2016.2531185

[83] Hao Liu, Yaoxue Zhang, Yuezhi Zhou, Di Zhang, Xiaoming Fu, and K K Ramakrishnan. 2014. Mining checkins from location-sharing services for client-independent IP geolocation. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'14)*. 619–627. https://doi.org/10.1109/INFOCOM.2014.6847987

[84] Kin Sum Liu, Bo Li, and Jie Gao. 2018. Generative model: Membership attack, generalization and diversity. *CoRR, abs/1805.09898* (2018).

[85] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor C.M. Leung. 2018. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access* 6 (2018), 12103–12117. https://doi.org/10.1109/ACCESS.2018.2805680

[86] Xi Liu, Hanzhou Chen, and Clio Andris. 2018. trajGANs: Using generative adversarial networks for geo-privacy protection of trajectory data (Vision paper). In *Location Privacy and Security Workshop*. 1–7.

[87] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojaning Attack on Neural Networks. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS'18)*. https://doi.org/10.14722/ndss.2018.23291

[88] Yujia Liu, Weiming Zhang, and Nenghai Yu. 2017. Protecting Privacy in Shared Photos via Adversarial Examples Based Stealth. *Security and Communication Networks* 2017 (2017). https://doi.org/10.1155/2017/1897438

[89] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2018. Understanding Membership Inferences on Well-Generalized Learning Models. *arXiv:1802.04889* (2018). http://arxiv.org/abs/1802.04889

[90] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD'05)*. 641–647. https://doi.org/10.1145/1081870.1081950

[91] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.* 5, 3 (2014), 47. https://doi.org/10.1145/2528548

[92] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, and K.R. Venugopal. 2016. Detection of fraudulent and malicious websites by analysing user reviews for online shopping websites. *International Journal of Knowledge and Web Intelligence* 5, 3 (2016), 171. https://doi.org/10.1504/ijkwi.2016.078712

[93] Robert R McCrae, Paul T. Costa, Antonio Terracciano, Wayne D Parker, Carol J Mills, Filip De Fruyt, and Ivan Mervielde. 2002. Personality trait development from age 12 to age 18: Longitudinal, cross-sectional, and cross-cultural analyses. *Journal of Personality and Social Psychology* 83, 6 (2002), 1456–1468. https://doi.org/10.1037/0022-3514.83.6.1456

[94] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Federated Learning of Deep Networks using Model Averaging. *Arxiv* 92, 9 (2016), 091118. https://doi.org/10.1063/1.2841713

[95] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models Without Lossing Accuracy. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*, Vol. 45. 39–44. https://doi.org/10.1145/585597.585599

[96] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. 2016. Defeating image obfuscation with deep learning. *arXiv:1609.00408* (2016).

[97] Hoda Mehrpouyan, Ion Madrazo Azpiazu, and Maria Soledad Pera. 2017. Measuring Personality for Automatic Elicitation of Privacy Preferences. In *Proceedings of the IEEE Symposium on Privacy-Aware Computing (PAC'17)*. 84–95.
[98] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. IEEE, 691–706.
[99] Wei Meng, Xinyu Xing, Anmol Sheth, Udi Weinsberg, and Wenke Lee. 2014. Your online interests: Pwned! a pollution attack against targeted advertising. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'14)*. 129–140.
[100] Ilya Mironov. 2017. Rényi Differential Privacy. In *Proceedings of the IEEE Computer Security Foundations Symposium (CSF'17)*. IEEE Computer Society, 263–275. https://doi.org/10.1109/CSF.2017.11
[101] Payman Mohassel and Yupeng Zhang. 2017. SecureML: A system for scalable privacy-preserving machine learning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. 19–38.
[102] Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 86–94. https://doi.org/10.1109/CVPR.2017.17
[103] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 2574–2582.
[104] Arvind Narayanan and Vitaly Shmatikov. 2006. How To Break Anonymity of the Netflix Prize Dataset. *cs/0610105* (2006). http://arxiv.org/abs/cs/0610105
[105] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. 739–753. https://doi.org/10.1109/SP.2019.00065
[106] Seth Neel, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. 2019. Oracle Efficient Private Non-Convex Optimization. (2019). arXiv:1909.01783 http://arxiv.org/abs/1909.01783
[107] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, and Nina Taft. 2013. Privacy-preserving ridge regression on hundreds of millions of records. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'13)*. 334–348.
[108] Ruairí Nugent. 2018. Assesing Completeness of Solvency and Financial Condition Reports through the use of Machine Learning and Text Classification. (2018).
[109] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. 2016. Faceless person recognition: Privacy implications in social media. In *Proceedings of European Conference on Computer Vision (ECCV'16)*. Springer, 19–35.
[110] Seong Joon Oh, Mario Fritz, and Bernt Schiele. 2017. Adversarial image perturbation for privacy protection a game theory perspective. In *Proceedings of IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, 1491–1500.
[111] Seong Joon Oh, Bernt Schiele, and Mario Fritz. 2019. Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 121–144.
[112] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kevin Huguenin, Mohammad Emtiyaz Khan, and Jean Pierre Hubaux. 2017. SmarPer: Context-Aware and Automatic Runtime-Permissions for Mobile Devices. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. 1058–1076. https://doi.org/10.1109/SP.2017.25
[113] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 8466–8475. https://doi.org/10.1109/CVPR.2018.00883
[114] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 3706–3715. https://doi.org/10.1109/ICCV.2017.398
[115] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. 2018. A non-parametric generative model for human trajectories. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'18)*. 3812–3817.
[116] Nicolas Papernot, Ian Goodfellow, Martín Abadi, Kunal Talwar, and Úlfar Erlingsson. 2019. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'19)*.
[117] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv:1605.07277* (2016). http://arxiv.org/abs/1605.07277
[118] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the ACM Asia Conference on Computer and Communications Security (ASIACCS'17)*. ACM, 506–519. https://doi.org/10.1145/3052973.3053009
[119] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P'16)*. IEEE, 372–387.

[120] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data synthesis based on generative adversarial networks. In *Proceedings of the VLDB Endowment*, Vol. 11. VLDB Endowment, 1071–1083. https://doi.org/10.14778/3231751.3231757

[121] Manas A. Pathak, Shantanu Rane, and Bhiksha Raj. 2010. Multiparty differential privacy via aggregation of locally trained classifiers. In *Advances in Neural Information Processing Systems (NIPS'10)*. 1876–1884.

[122] Siani Pearson and Azzedine Benameur. 2010. Privacy, security and trust issues arising from cloud computing. In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom'10)*. 693–702. https://doi.org/10.1109/CloudCom.2010.66

[123] NhatHai Phan, My T. Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. 2020. Scalable Differential Privacy with Certified Robustness in Adversarial Learning. In *Proceedings of the 37th International Conference on Machine Learning (PMLR'20)*, Vol. 6. arXiv:1903.09822 http://arxiv.org/abs/1903.09822

[124] Nhathai Phan, Xintao Wu, Han Hu, and Dejing Dou. 2017. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *Proceedings of IEEE International Conference on Data Mining (ICDM'17)*, Vol. 2017-November. IEEE, 385–394. https://doi.org/10.1109/ICDM.2017.48

[125] Nhat Hai Phan, Minh N. Vu, Yang Liu, Ruoming Jin, Dejing Dou, Xintao Wu, and My T. Thai. 2019. Heterogeneous Gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'19)*, Vol. 2019-Augus. 4753–4759. https://doi.org/10.24963/ijcai.2019/660 arXiv:1906.01444

[126] Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. 2018. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. *IEEE Trans. Inf. Forensics Security* 13, 5 (2018), 1333–1345. https://doi.org/10.1109/TIFS.2017.2787987

[127] Benny Pinkas. 2002. Cryptographic techniques for privacy-preserving data mining. *ACM SIGKDD Explorations Newsletter* 4, 2 (2002), 12–19. https://doi.org/10.1145/772862.772865

[128] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. 2018. Generative Adversarial Perturbations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 4422–4431. https://doi.org/10.1109/CVPR.2018.00465

[129] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. 2020. Sampling Attacks: Amplification of Membership Inference Attacks by Repeated Queries. (2020). arXiv:2009.00395 http://arxiv.org/abs/2009.00395

[130] Alexey Reznichenko and Paul Francis. 2014. Private-by-design advertising meets the real world. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'14)*. 116–128. https://doi.org/10.1145/2660267.2660305

[131] Andras Rozsa, Ethan M Rudd, and Terrance E Boult. 2016. Adversarial Diversity and Hard Positive Generation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPW'16)*. 410–417. https://doi.org/10.1109/CVPRW.2016.58

[132] Benjamin I P Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. 2012. *Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning*. Technical Report 1. 65–100 pages. http://repository.cmu.edu/jpc

[133] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. 2020. Dynamic Backdoor Attacks Against Machine Learning Models. (2020). arXiv:2003.03675 http://arxiv.org/abs/2003.03675

[134] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS'19)*. https://doi.org/10.14722/ndss.2019.23119

[135] Ashish P Sanil, Alan F Karr, Xiaodong Lin, and Jerome P Reiter. 2004. Privacy preserving regression modelling via distributed computation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. 677–682. https://doi.org/10.1145/1014052.1014139

[136] Phillipp Schoppmann, Borja Balle, Jack Doerner, Samee Zahur, and David Evans. 2016. Secure Linear Regression on Vertically Partitioned Datasets. *IACR Cryptology Eprint Archive* 2016 (2016), 1–27.

[137] Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *Comput. Surveys* 34, 1 (2002), 1–47. https://doi.org/10.1145/505282.505283

[138] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'16)*. 1528–1540. https://doi.org/10.1145/2976749.2978392

[139] Parvaneh Shayegh and Sepideh Ghanavati. 2017. Toward an approach to privacy notices in IoT. In *Proceedings of the IEEE 25th International Requirements Engineering Conference Workshops (REW'17)*. 104–110. https://doi.org/10.1109/REW.2017.77

[140] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'15)*. 1310–1321.

[141] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 3–18. https://doi.org/10.1109/SP.2017.41

[142] Reza Shokri, George Theodorakopoulos, Panos Papadimitratos, Ehsan Kazemi, and Jean Pierre Hubaux. 2014. Hiding in the mobile crowd: Location privacy through collaboration. *IEEE Trans. Dependable Secure Comput.* 3 (2014), 266–279. https://doi.org/10.1109/TDSC.2013.57

[143] Aleksandra B Slavkovic, Yuval Nardi, and Matthew M Tibbits. 2007. Secure logistic regression of horizontally and vertically partitioned distributed databases. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'07)*. 723–728. https://doi.org/10.1109/ICDMW.2007.114

[144] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'17)*. ACM, 587–601. https://doi.org/10.1145/3133956.3134077

[145] Guocong Song and Wei Chai. 2018. Collaborative learning for deep neural networks. In *Advances in Neural Information Processing Systems (NIPS'18)*, Vol. 2018-Decem. 1832–1841.

[146] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP'13)*. 245–248. https://doi.org/10.1109/GlobalSIP.2013.6736861

[147] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward automated online photo privacy. *ACM Trans. Web* 11, 1 (2017), 2. https://doi.org/10.1145/2983644

[148] Anna C Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2014. Analyzing images' privacy for the modern web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT'14)*. 136–147. https://doi.org/10.1145/2631775.2631803

[149] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Natural and Effective Obfuscation by Head Inpainting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'18)* (2018), 5050–5059. https://doi.org/10.1109/CVPR.2018.00530

[150] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A domain based approach to social relation recognition. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 435–444. https://doi.org/10.1109/CVPR.2017.54

[151] Xudong Sun, Pengcheng Wu, and Steven C.H. Hoi. 2018. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing* 299 (2018), 42–50. https://doi.org/10.1016/j.neucom.2018.03.030

[152] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR'14)*.

[153] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR. *Companion Proceedings of the The Web Conference 2018* 2 (2018), 163–166.

[154] Financial Times. 2020. Facebook privacy breach. *Financial Times* (2020), 11–12. https://www.ft.com/content/87184c40-2cfe-11e8-9b4b-bc4b9f08f381

[155] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *Proceedings of the 25th USENIX Security Symposium (USENIX'16)*. 601–618.

[156] Aleksei Triastcyn and Boi Faltings. 2019. Generating artificial data for private deep learning. In *Proceedings of the 2019 CEUR Workshop*, Vol. 2335. 33–40.

[157] Jaideep Vaidya, Murat Kantarcoglu, and Chris Clifton. 2008. Privacy-preserving naive bayes classification. *The VLDB Journal* 17, 4 (2008), 879–898.

[158] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR abs/1609.03499* (2016). http://arxiv.org/abs/1609.03499

[159] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. 2018. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv:1812.00564* (2018). http://arxiv.org/abs/1812.00564

[160] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'15)*, Vol. 07-12-June. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[161] Binghui Wang and Neil Zhenqiang Gong. 2018. Stealing Hyperparameters in Machine Learning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'18)*. 36–52. https://doi.org/10.1109/SP.2018.00038

[162] K Wang, R Chen, B C Fung, and P S Yu. 2010. Privacy-preserving data publishing: A survey on recent developments. *Comput. Surveys* (2010).

[163] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. Beyond Inferring Class Representatives: User-Level Privacy Leakage from Federated Learning. *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'19)* 2019-April (2019), 2512–2520. https://doi.org/10.1109/INFOCOM.2019.8737416

[164] Lingxiao Wei, Bo Luo, Yu Li, Yannan Liu, and Qiang Xu. 2018. I know what you see: Power side-channel attack on convolutional neural network accelerators. In *ACM International Conference Proceeding Series*. 393–406. https://doi.org/10.1145/3274694.3274696

[165] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. 2015. Android permissions remystified: A field study on contextual integrity. In *Proceedings of the 24th USENIX Security Symposium (USENIX'15)*. 499–514.

[166] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. 2017. The Feasibility of Dynamically Granted Permissions: Aligning Mobile Privacy with User Preferences. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. 1077–1093. https://doi.org/10.1109/SP.2017.51

[167] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. 2018. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'18)*, Vol. 2018-April. 268. https://doi.org/10.1145/3173574.3173842

[168] Michael J Wilber, Vitaly Shmatikov, and Serge Belongie. 2016. Can we still avoid automatic face detection?. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'16)*. 1–9. https://doi.org/10.1109/WACV.2016.7477452

[169] Nan Wu, Farhad Farokhi, David Smith, and Mohamed Ali Kaafar. 2020. The value of collaboration in convex machine learning with differential privacy. In *2020 IEEE Symposium on Security and Privacy (SP)*. 304–317.

[170] Shuang Wu, Tadanori Teruya, Junpei Kawamoto, Jun Sakuma, and Hiroaki Kikuchi. 2013. Privacy-preservation for Stochastic Gradient Descent Application to Secure Logistic Regression. In *Proceedings of the 27th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI'13)*. 6–9.

[171] Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing Tabular Data using Generative Adversarial Networks. *arXiv:1811.11264* (2018). http://arxiv.org/abs/1811.11264

[172] Yuanshun Yao, Haitao Zheng, Huiying Li, and Ben Y. Zhao. 2019. Latent backdoor attacks on deep neural networks. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS'19)*. 2041–2055. https://doi.org/10.1145/3319535.3354209

[173] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the IEEE Computer Security Foundations Symposium (CSF'18)*, Vol. 2018-July. 268–282. https://doi.org/10.1109/CSF.2018.00027

[174] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. 2017. IPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Trans. Inf. Forensics Security* 12, 5 (2017), 1005–1016. https://doi.org/10.1109/TIFS.2016.2636090

[175] Lin Yuan, Joël Theytaz, and Touradj Ebrahimi. 2017. Context-dependent privacy-aware photo sharing based on machine learning. In *IFIP Advances in Information and Communication Technology*, Vol. 502. 93–107. https://doi.org/10.1007/978-3-319-58469-0_7

[176] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012. Privacy-aware image classification and search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. 35–44. https://doi.org/10.1145/2348283.2348292

[177] Dayin Zhang, Xiaojun Chen, Dakui Wang, and Jinqiao Shi. 2018. A survey on collaborative deep learning and privacy-preserving. In *Proceedings of the IEEE 3rd International Conference on Data Science in Cyberspace (DSC'18)*. IEEE, 652–658. https://doi.org/10.1109/DSC.2018.00104

[178] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. 2012. Functional mechanism: Regression analysis under differential privacy. *Proceedings of the International Conference on Very Large Data Bases (VLDB'12)* 5, 11 (2012), 1364–1375. https://doi.org/10.14778/2350229.2350253

[179] Tianwei Zhang, Zecheng He, and Ruby B. Lee. 2018. Privacy-preserving Machine Learning through Data Obfuscation. *arXiv:1807.01860* (2018). http://arxiv.org/abs/1807.01860

[180] Xinyang Zhang, Shouling Ji, and Ting Wang. 2018. Differentially Private Releasing via Deep Generative Model (Technical Report). *arXiv:1801.01594* (2018). http://arxiv.org/abs/1801.01594

[181] Wengang Zhou, Houqiang Li, Yijuan Lu, and Qi Tian. 2012. Principal visual word discovery for automatic license plate detection. *IEEE Trans. Image Process.* 21, 9 (2012), 4269–4279. https://doi.org/10.1109/TIP.2012.2199506