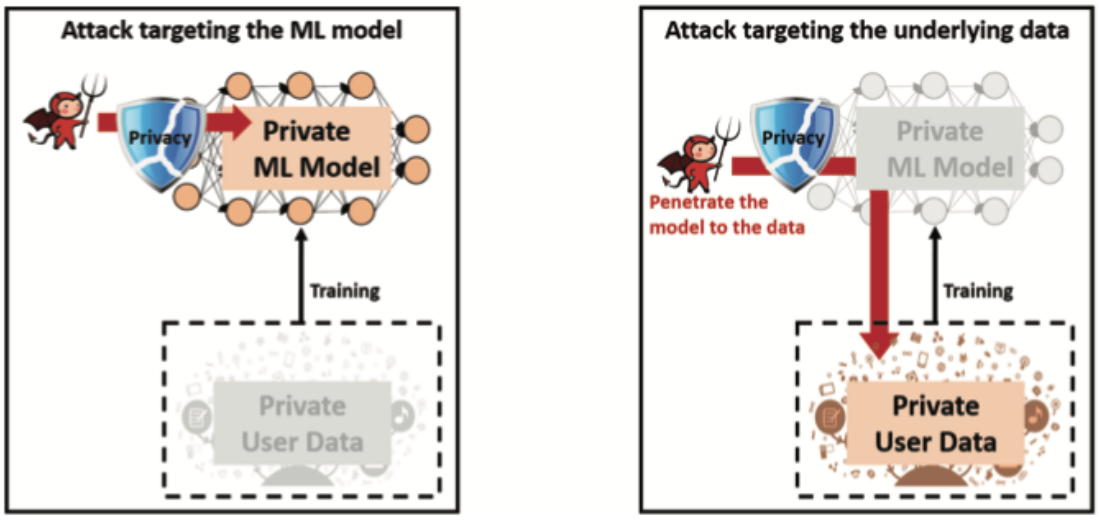
## Privacy and Machine Learning—Where are we?

[3]

Privacy has been an emerging issue in the past couple of years. With data breaches and very little accountability[1], individuals are losing trust in companies with their data. As companies add additional security measures, malicious attackers are finding new ways to violate user privacy; directly attacking the machine learning process which uses user data. Already individuals don't trust AI[2] and this will further reduce its popularity. But how bad is it? What is the current state of privacy with respect to machine learning? Let's dive in and explore. We'll discuss some of the a) current attacks on the training data and model, b) using machine learning for privacy protection, and c) machine learning for privacy attacks and how we can defend against them.

## Private Machine Learning

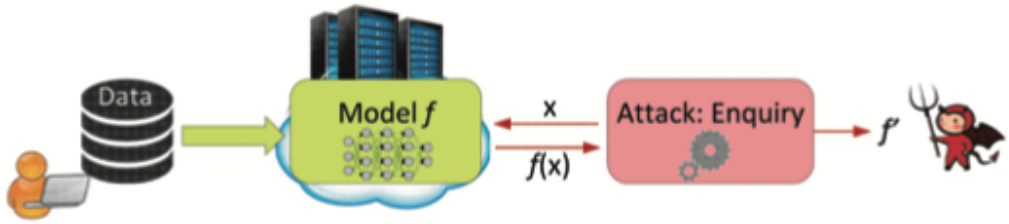



(a) Privacy of the ML model. (b) Privacy of the underlying data.

ML Model and Training Data Attacks[4]

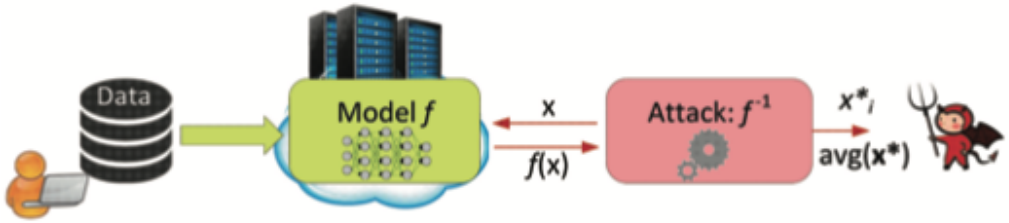
Despite the verbiage in the headline, private machine learning doesn't mean creating and training models in secrecy. This is referring to the current attacks and prevention methods on machine learning. Malicious attackers are able to violate user privacy by either attacking the private model or training data. With training data, it is a bit more obvious that it may contain private user information that could be compromised[4]. Though the model seems a bit more robust, attackers are able to obtain information about the model such as the model parameters and training algorithms[4]. A lot of this depends on the type of access they have as well. ***White-box*** access is when the attacker has direct complete access to the model[4]. ***Black-box*** is a bit more limited in which the attacker can only hit an endpoint to request information from the model[4].

But how? Let's dive into a couple of attacks. You may look at the following figure for a rough overview before we go in-depth.

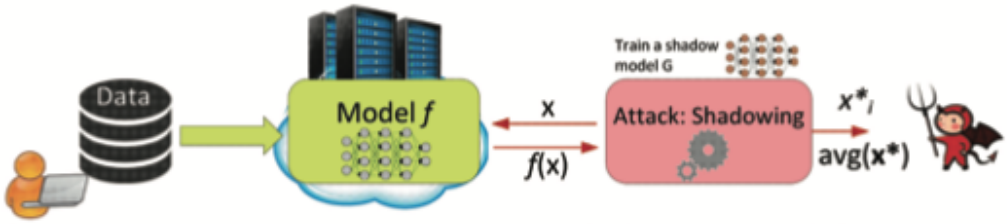



**Model Extraction Attack:** Adversary learns a close approximation $f'$(x) of $f$(x)

(a) Model Extraction Attack

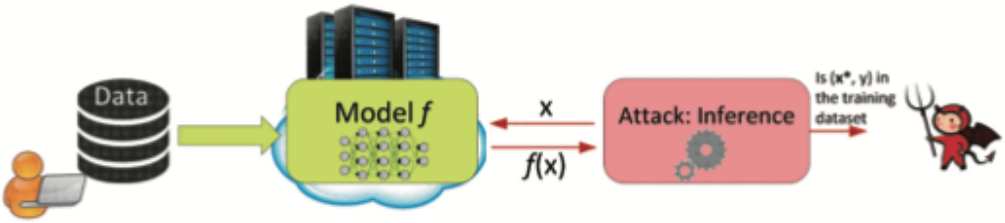



**Model Inversion Attack:** Adversary learns certain features $x^*_i \in$ **x*** or statistical properties such as avg(**x***) of the training dataset



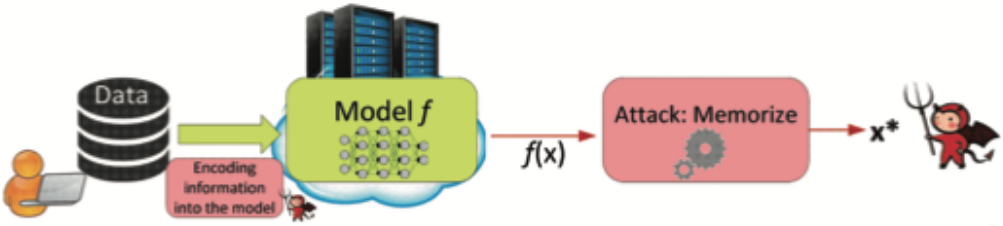

**Shadow Model Attack:** Adversary learns certain features or statistical properties of the training dataset, with the help of the shadow model G

(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

Variety of attacks on training data and the model[4]

**Model Extraction Attack—(Model)**

This attack is attempting to create a ***duplicate/copy*** of the original model. The attacker is approximating the function to mimic the parameters of the model. Typically the attacker will have black-box access. The attacker will provide data to the model and use the expected output to approximate a function/model that matches the original model[4].

**Feature Estimation Attack—(Training Data)**

Instead of attacking the model, the attacker now tries to obtain information from the training data the model trained on. In a feature estimation attack, the attacker tries to essentially estimate the features of the training dataset[4]. Two implementations of this are called ***Model Inversion Attack*** and ***Shadow Model Attack***. Let's take a look at both of these.

In a model inversion attack, the attacker typically needs white-box access (direct access to the model, there have been attempts with black-box but not as successful)[4]. The attacker is essentially trying to reverse engineer the original model and estimate the feature vector for the target output classes[4]. This effort requires understanding the flow of the gradients in the model to properly adjust the weights[4]. Eventually, this results in an inverse model which can take the target output class and is able to **roughly** predict the feature vectors. There is an enhancement that can be made with GANS. But before that, we need to answer "What is GANS?". GANS (Generative Adversarial Networks) is an unsupervised deep learning model whose goal is to learn information or patterns from data and be able to reconstruct or generate the data as close as possible.[5] Though GANS is meant to be used for generating examples, malicious attackers are using this same technique to obtain feature information from a model[5]. Research has shown that a GANS model is able to produce almost identical features that were used in the training data of the original model.

In a shadow model attack, the attacker in this scenario can have either black/white box access[4]. This technique looks very similar to the first (model extraction attack) in which you are trying to approximate or create a model close enough to the original model. However, the focus of this attack is to obtain the features, not the class output.

**Membership Inference Attack—(Training Data)**

Instead of focusing on obtaining information from the model/training data, the attacker in a membership inference attack is trying to check if a certain data record was used in the training process of a model[4]. The issue with deep learning models in this day and age is most of them are overfitted to the training data to an extent[4][6]. Since they don't generalize too much, it makes it very easy to make an inference attack model to determine if a data record belongs to the

training data the model trained on. Additionally, many of these deep learning models are using existing frameworks and tools from corporations like Google and Amazon[6]. The frameworks and tools have relatively the same implementation but with different training data. These same models are available to the attacker who leverage this to conduct their membership inference attack[6].

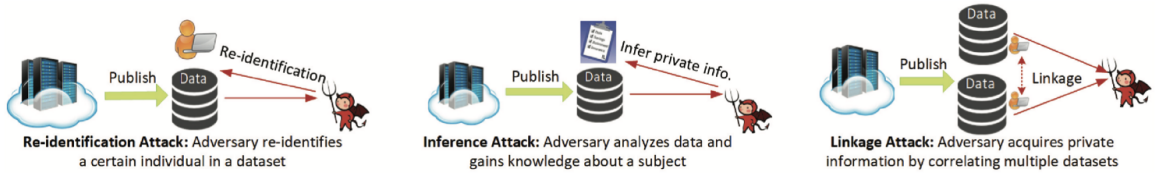
**Model Memorization Attack—(Training Data)**

Instead of just checking if the data exists, this attack allows the attacker to obtain the exact features of an individual. The attacker typically has direct access to the model and can encode malicious information into it [4]. Typically the attacker will have access to the model and is able to encode malicious values into the model params or outputs [4]. When the model is served, this encoded information can be retrieved from hitting the model's endpoint [4]. If they don't have direct access to the model, the attacker is able to create synthetic inputs which can potentially leak information from the model.

**Defense Mechanisms?**

Though these attacks seem quite scary, there are some potential defense mechanisms that can be placed in place to reduce the possibilities of these attacks. One idea is to **encrypt** the training data or model itself [4]. **Obfuscation** can be conducted in which the information is jumbled up so the attacker may not understand it but the model can [4]. **Aggregation** can be done as well in which machine learning tasks are distributed [4]. This goes kind of into Federated Learning (see this article for more information on that).

## Machine Learning Aided Privacy Protection

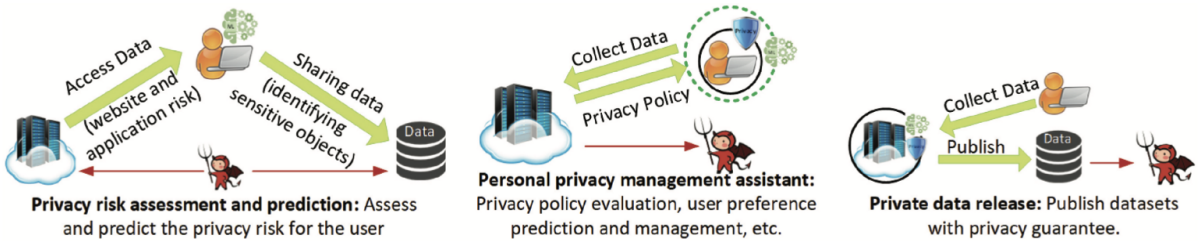



(a) Re-identification Attack (b) Inference Attack (c) Linkage Attack

Different Direct Attacks On Individuals [4]

We've seen machine learning used for attacking, now let's take a look at how it can help defend against some privacy-related attacks. Most users are unaware that pictures they post on social media sites provide more information to attackers than they think [4]. Just by analyzing the surroundings of an image, an attacker can learn a significant amount about your family, friends, values, etc [4]. Some common attacks are **identification attacks** where the attacker attempts to learn the identity of a user [4]. Attackers can conduct an **inference attack** as well by gaining other additional information about the user [4]. Finally, attackers can conduct a **linkage attack**

where the attacker uses multiple data sources to learn information about the victim [4]. Traditional defense mechanisms don't work as well since models have become more mature and data has become more unstructured [4]. Because of this, new research has been done to provide privacy protection.




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

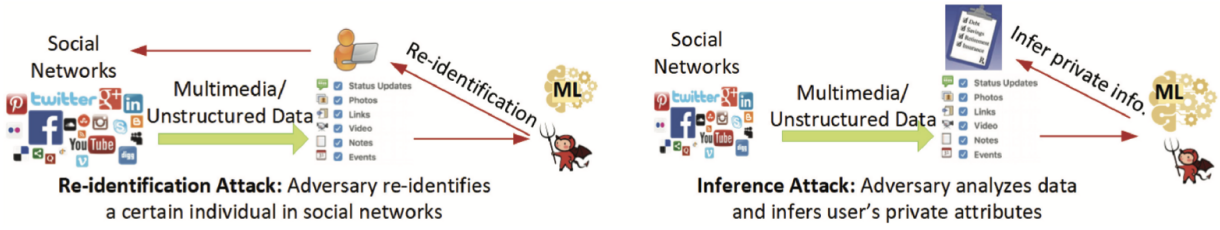
New Privacy Protection Ideas [4]

**Privacy risk assessment and prediction** systems help an individual when they are trying to access, such as a website, or sharing, such as images on their social media [4]. These systems help notify individuals how much information these websites are gathering and what their overall rating is in terms of privacy [4]. Tools have been implemented in cases where users are sharing items such as photos to let the individual know what additional information, which they may not be aware of, could be shared. **Personal privacy management assistant** systems help individuals understand and summarize privacy policies [4]. You may be familiar with this when you are prompted with a big text box asking to you accept the "Terms of Agreements". In this same regard, for privacy policies, AI systems are made to help individuals understand what they are agreeing to [4]. These systems additionally set privacy settings automatically on websites you're on which some individuals aren't even aware are enabled [4]. Lastly, **private data release** systems are ensuring that datasets that are/will be publically available are indeed private and reveal no information about the individual user [4].

Machine learning seems like a great tool against privacy-related attacks. However, machine learning in itself can be used to conduct these attacks. Let's take a look into how this is done.

## Machine Learning-based Privacy Attacks and Corresponding Protection Schemes

Instead of using machine learning to protect against privacy attacks, we showcase how machine learning itself can be used to make attacks against privacy. This might be confusing with respect to the first section where we talked earlier about how attackers used machine learning models for their attacks. But these attacks only focused on the model and training data. This section shows a wider range of attacks possible with machine learning models.




(a) Identification attack

(b) Inference attack

**Re-identification attacks** consist of a machine learning model being able to identify the individual given some data record [4]. For example, if an individual shares their image and geolocation, attackers are able to find the identity **and** location of the user [4]. This is pretty common on social media sites that allow you to share this information (Facebook, Instagram, Twitter, Snapchat, etc). Most individuals are aware of this but machine learning models can be used for **inference attacks** [4]. This is the ability of machine learning models to learn about the user and make recommendations for them. These can consist of targeted ads for the individual, video recommendations, search recommendations, and much more [4].

It's a little difficult to recommend a solution to these problems because most individuals don't understand how their data and information can be extracted and used [4]. The general solution is machine learning models should be limited in automatically accessing this data. There is definitely more research that needs to be done in this field to help prevent these types of attacks [4].

## Future?

Despite the wide array of attacks on privacy, there seems to be an adequate amount of continuous research being done to help prevent attacks. Ideas of adding noise, obfuscation, and encryption can really help improve privacy protection. Algorithms such as GANS, which we initially talked about being used to attack with, can be used to generate synthetic outputs to make it harder for the attacker to trace back to the original data. This constant battle will go back and forth as attackers find new methods of attacks, researchers will implement secure solutions.

## References

[1] https://www.npr.org/2021/04/09/986005820/after-data-breach-exposes-530-million-facebook-says-it-will-not-notify-users

[2] https://towardsdatascience.com/people-dont-trust-ai-we-need-to-change-that-d1de5a4a0021

[3] https://joeyandbluewhale.github.io/project/mlprivacy/

[4] https://arxiv.org/abs/2011.11819 (original paper this article is based on)

[5]https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

[6]https://venturebeat.com/2021/04/28/membership-inference-attacks-detect-data-used-to-train-machine-learning-models/