IT Security – When Machine Learning Meets Privacy: Private Machine Leaning

COS300015

Yasas Supeksala

SWINBURNE UNIVERSITY OF TECHNOLOGY

Acknowledgement of Country

We respectfully acknowledge the Wurundjeri People of the Kulin Nation, who are the Traditional Owners of the land on which Swinburne's Australian campuses are located in Melbourne's east and outer-east, and pay our respect to their Elders past, present and emerging.

We are honoured to recognise our connection to Wurundjeri Country, history, culture, and spirituality through these locations, and strive to ensure that we operate in a manner that respects and honours the Elders and Ancestors of these lands.

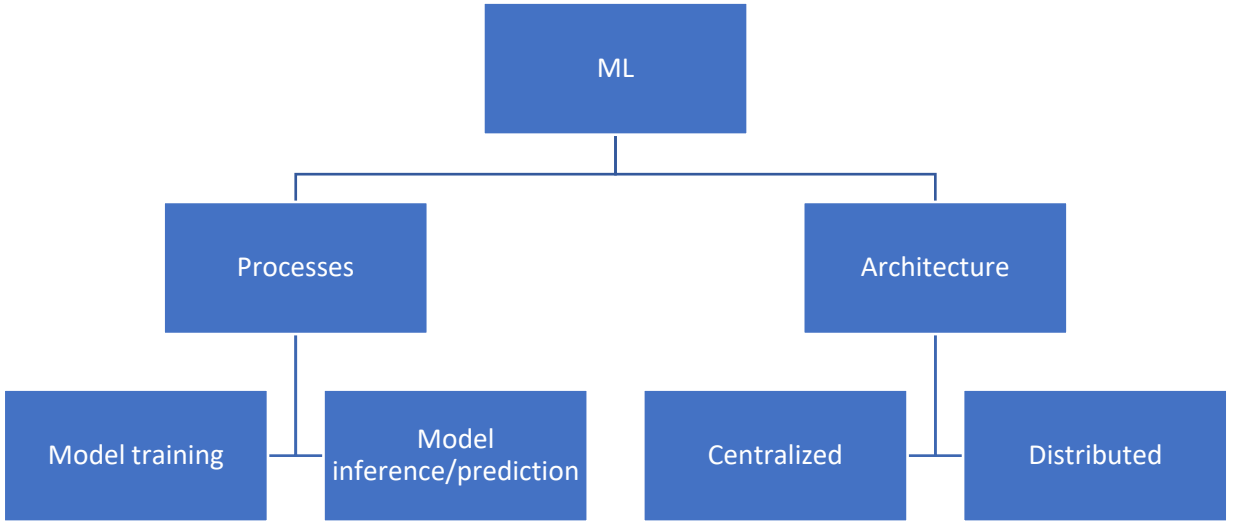
We also respectfully acknowledge Swinburne's Aboriginal and Torres Strait Islander staff, students, alumni, partners and visitors.

We also acknowledge and respect the Traditional Owners of lands across Australia, their Elders, Ancestors, cultures, and heritage, and recognise the continuing sovereignties of all Aboriginal and Torres Strait Islander Nations.

Major concerns still exists in machine-learning based artificial intelligence schemes. There are three main methods that addresses the both model and data privacy,

i. Private Machine Learning
ii. Machine-leaning aided Privacy Protection
iii. Machine learning-based privacy attack and corresponding protection schemes

Prior explicitly going on to the investigation on scenario based applications of privacy and ML, the authors have presented privacy threats in context of ML.



The Privacy issues involved with centralized architecture and how distributed learning has addressed those concerns.

Centralized learning uses a set of training data centralized in a server and uses a centralized entity to train and host the models.

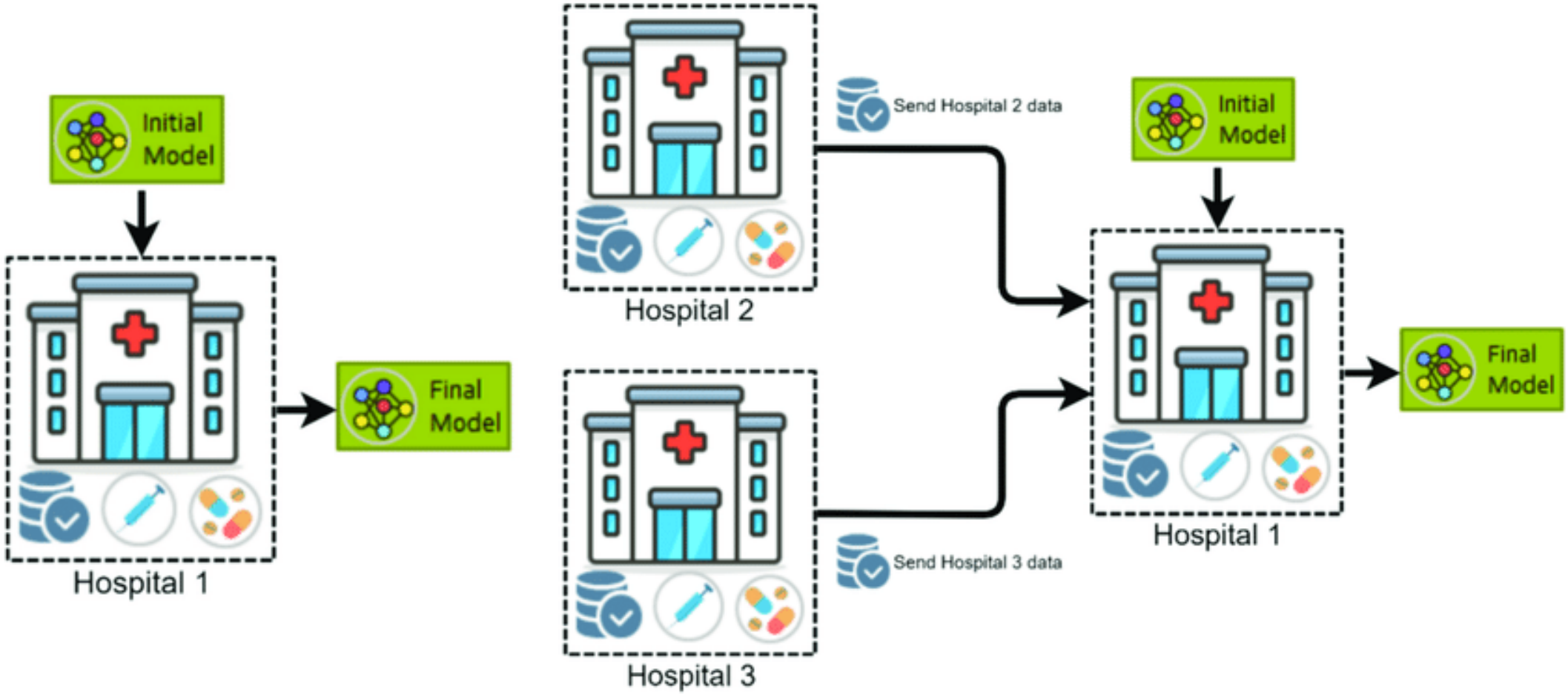
Vulnerabilities associated with centralized learning,

- Centralized operator has direct access to sensitive data, making the data and the model vulnerable.
- The increasing data equities,
- The data management and accessing concerns.

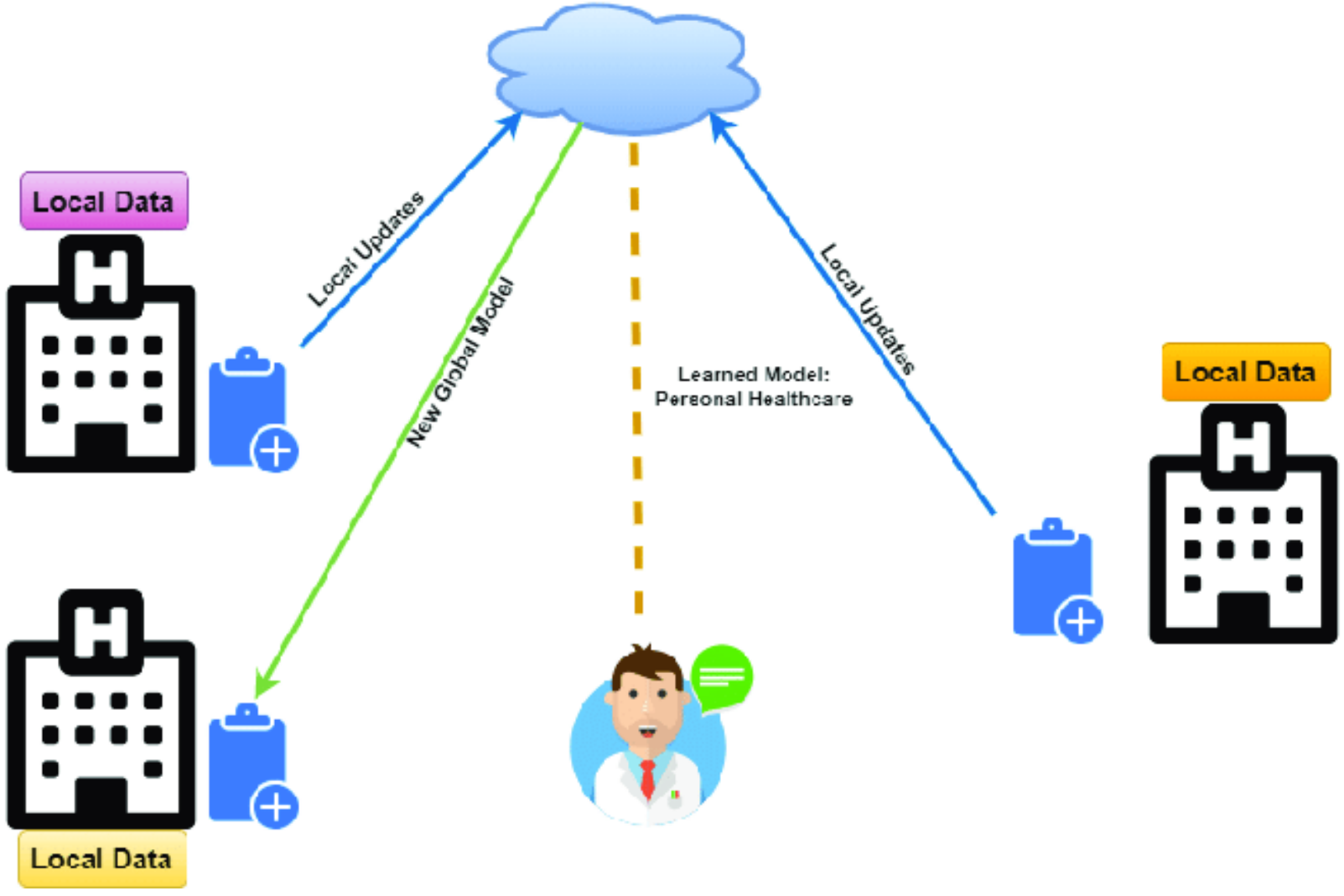
The Distributed learning addresses the above concerns by introducing globally stored and trained models which have several variations, in the context of conducting ML tasks in a distributed manner.

Collaborative learning	Multiple classifiers of the same network are simultaneously trained on the same training data to improve better performance.
Federated learning	There are two settings for FL, which are cross-drive and cross-silo. Basically this trains the model across multiple decentralized edge devices.
Split learning	User trains network up to a certain layer known as cut layer and sends weights to server.

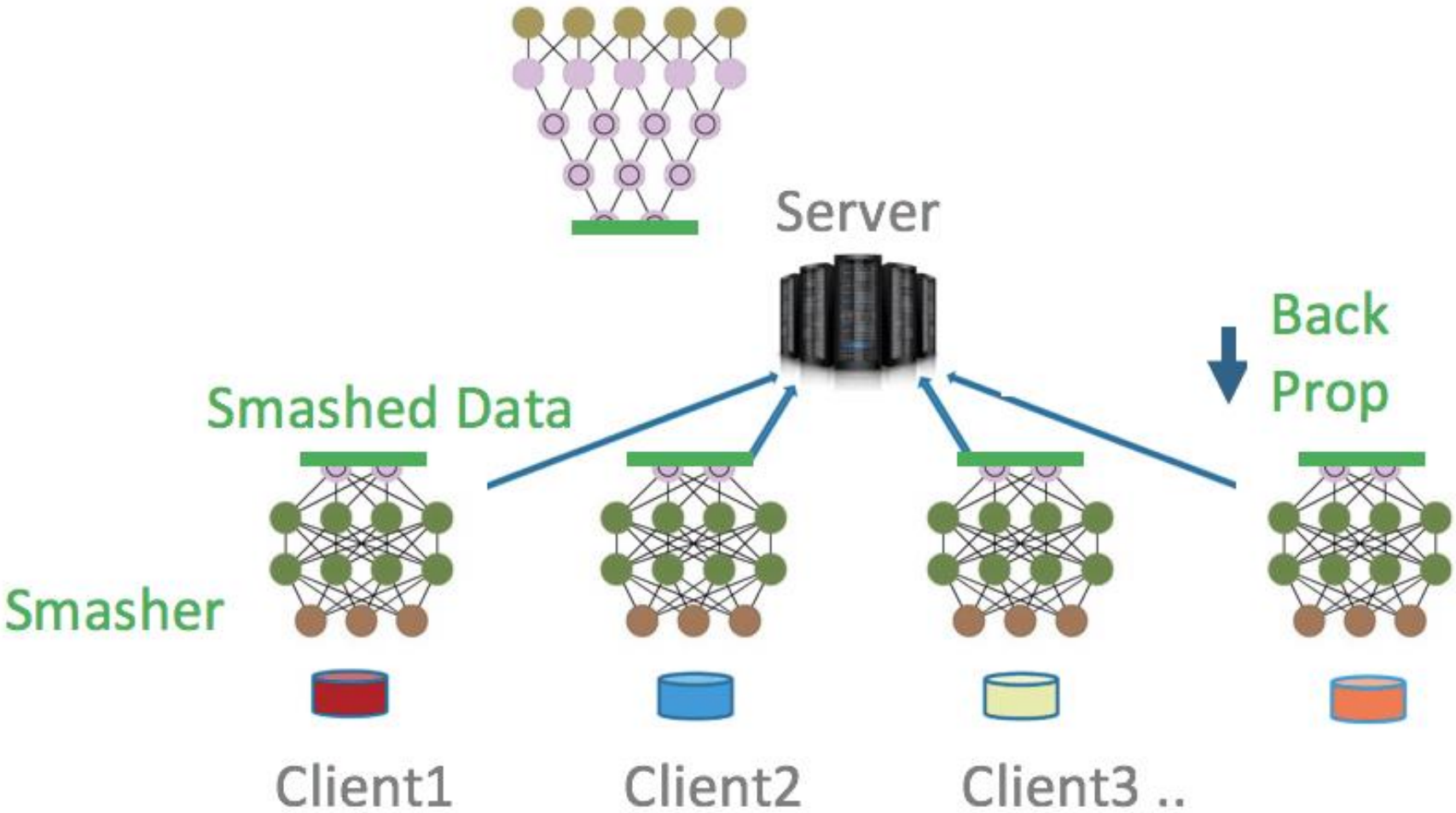
Collaborative Learning



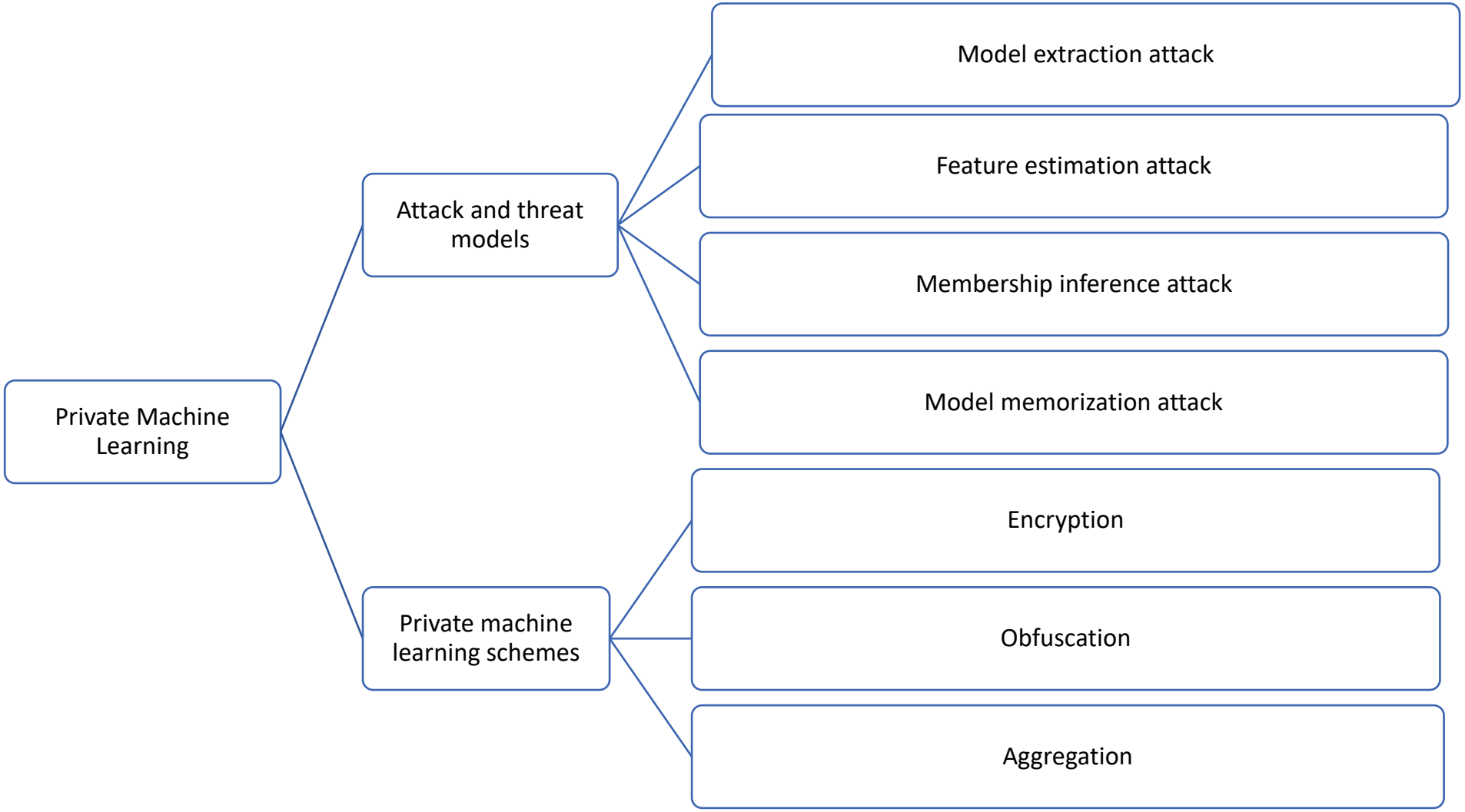
Federated Learning



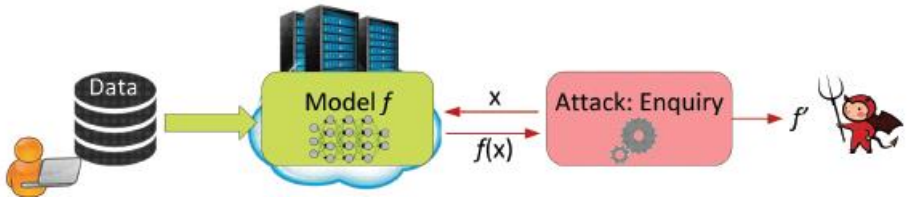
Split Learning



Private machine learning



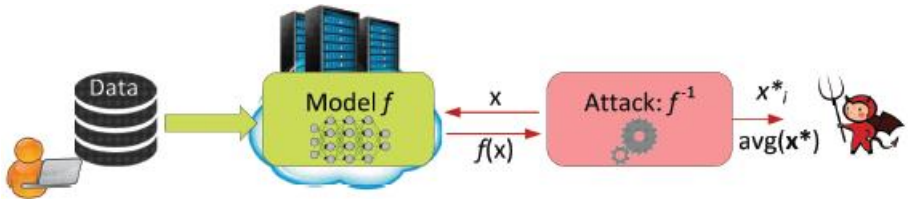
Model Extraction attack



The model extraction attack targets at the duplication of the AI model. After the completion of the attack, the adversary will obtain lose f'(x) of the target f(x) model. This attack is highly threatening because it can operate without any prior knowledge of the model.

i.e. shadow training

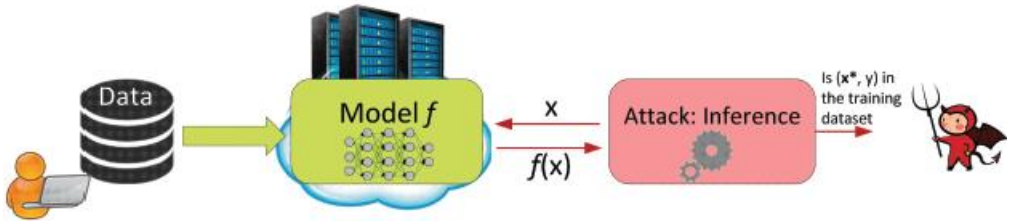
Feature estimation attack



This attack aims to estimate certain features or statistical properties such as avg(x*) of the training dataset.

This attack category can be implemented by model inversion attacks, shadow model attacks or power-side channel attacks. This attack will be deployed mainly in a Whitebox access scenario, although it can be deployed in black box access scenarios, the effectiveness remains lower.

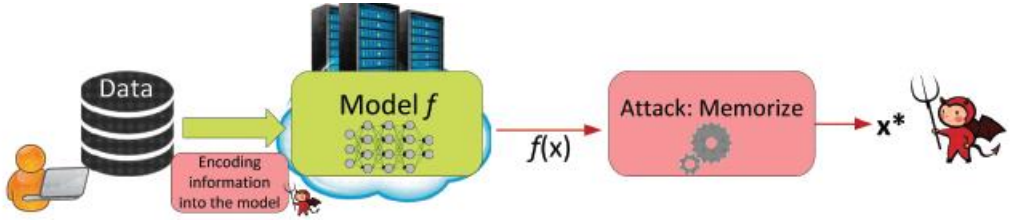
Membership Inference attack



Membership inference attack refers to acquiring the knowledge about whether a certain data record belongs to the models' training dataset D or not.

According to the reviewed research, overfitting, the structure and type of the model are the main factors that cause a model to be vulnerable to a model inference attack.

Model memorization attack



Model memorization attacks are capable of extracting exact feature values on individual samples.

This attack has the capability to find it’s way around both black-box and white-box access mechanisms.

There are several methods which can be used to encode data into the model and make it adversarial such as, LSB encoding, Correlated value encoding, Sign encoding.

Private machine learning schemes

In order to address above privacy attacks on private ML there are some major ML schemes that have been discussed in the current literature.

- **Encryption**

Encryption method	Description	Implementations
Encrypting training data	The data used for the ML model is encrypted using homomorphic encryption	Training over encrypted data Cryptonets Collaborative learning where data-sets are encrypted
Encrypting training model	Addition of homomorphic encryption on the gradients.	Secure multi-party computation(SMC)

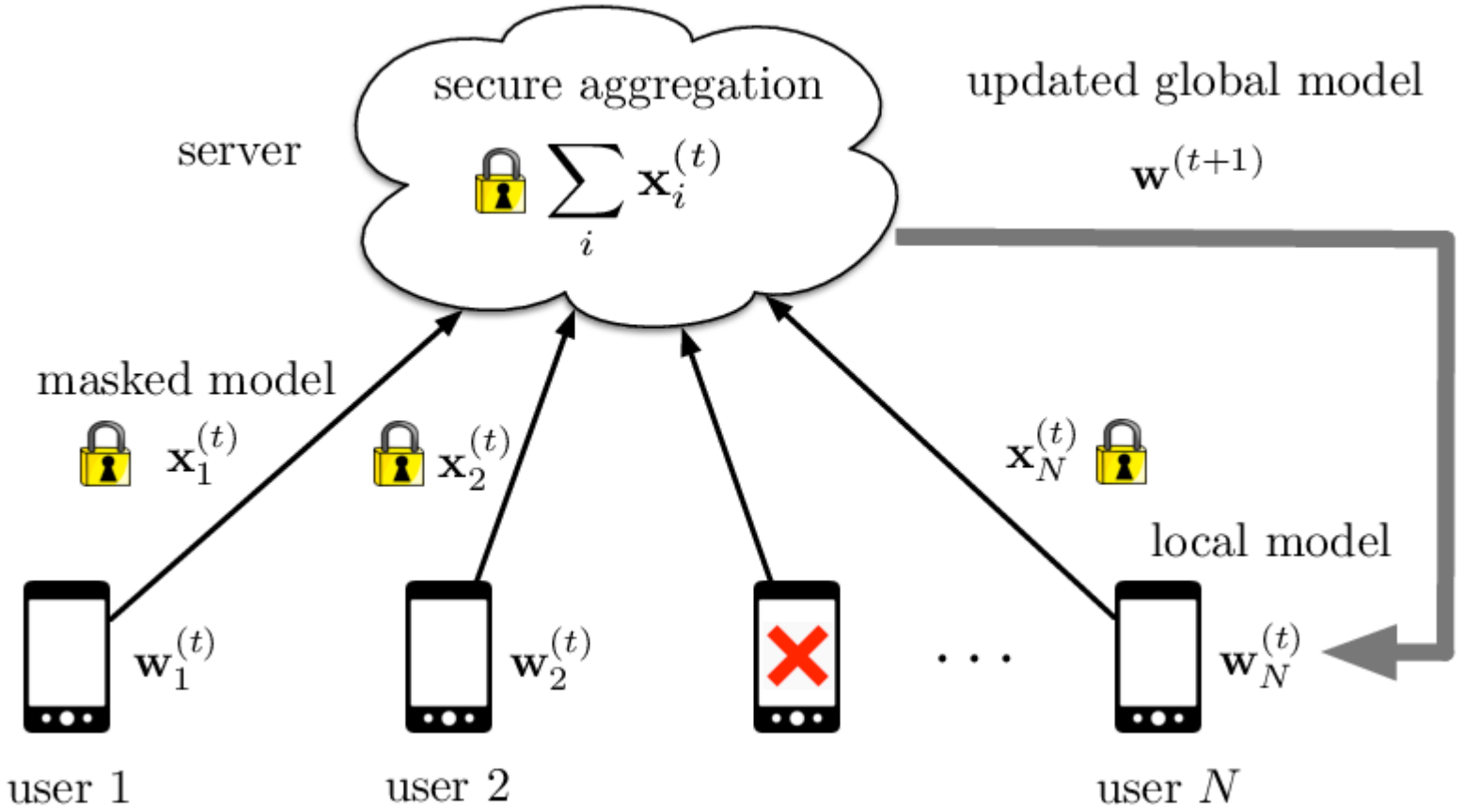
- **Obfuscation**

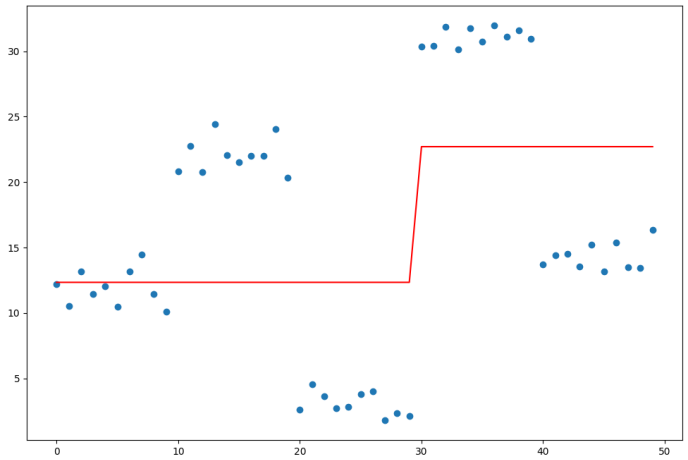
Using obfuscation, the ML obtains a state of privacy protection on it's data by reducing the precision of them.

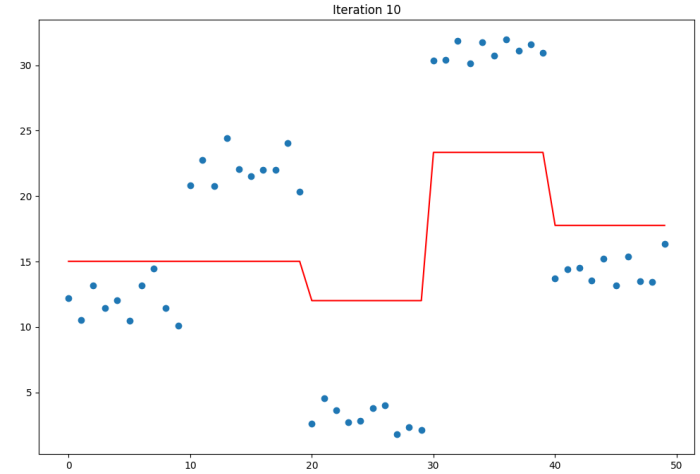
Basically what obfuscation do is adding noise to the ML model or data-sets. The most obfuscation method used in current cybersecurity landscape is Differential privacy schemes.

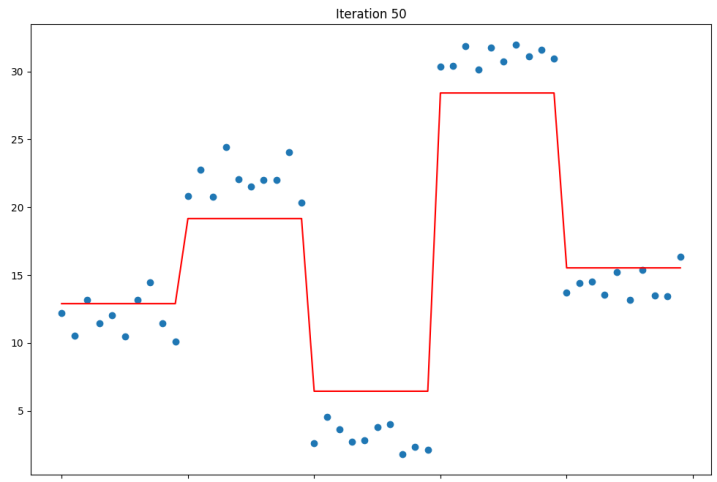
- **Aggregation**

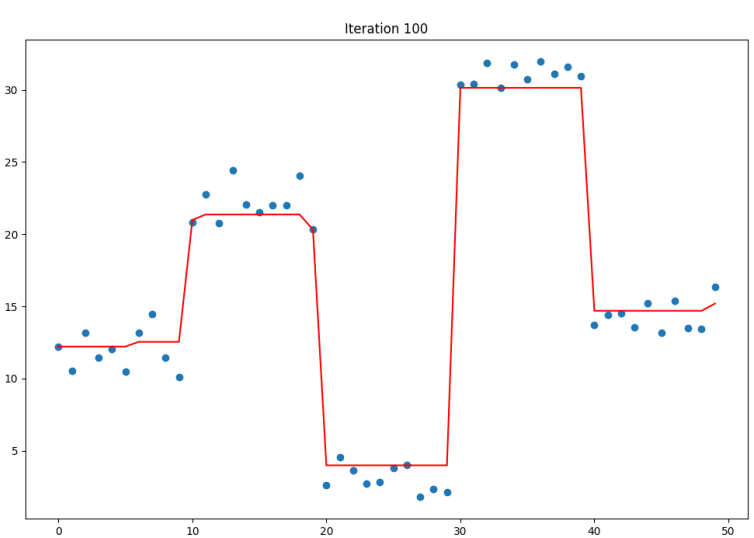
Aggregation is a similar technique to the distributed collaborative learning. So what happens in here is multiple parties join a machine learning tasks while keeping their datasets private.




30
25
20
15
10
5
0
10
20
30
40
50


Iteration 10
30
25
20
15
10
5
0
10
20
30
40
50


Iteration 50
30
25
20
15
10
5
0
10
20
30
40
50


Iteration 100
30
25
20
15
10
5
0
10
20
30
40
50

Thank you..