NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

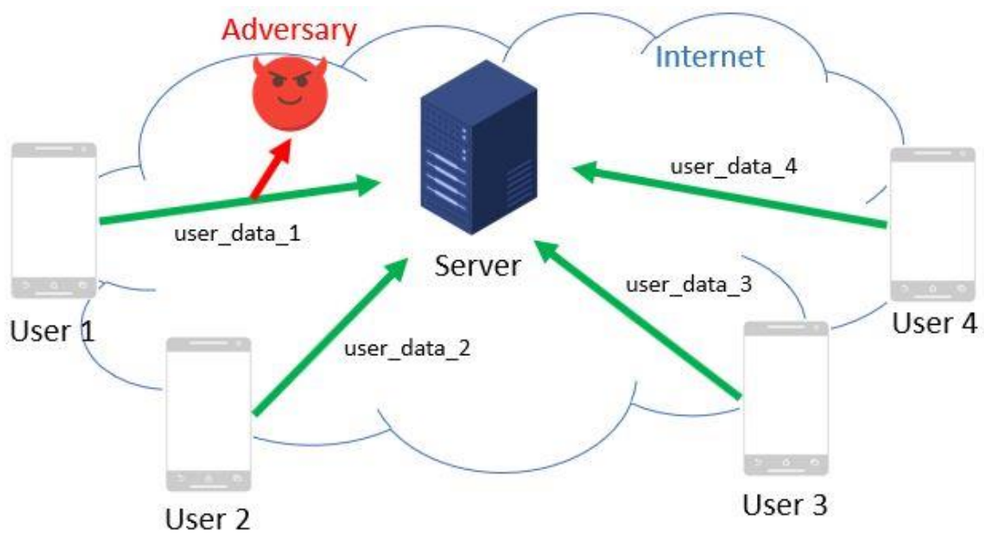
Study on Attacks Against Federated Learning

Distributed backdoor attacks & Poisoning Defence Generative Adversarial Networks

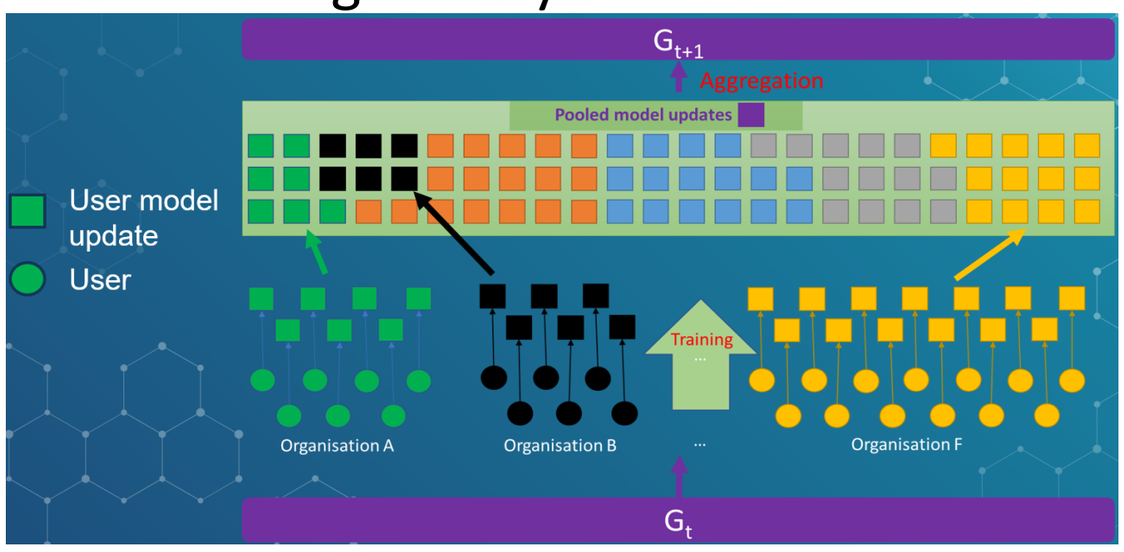
Student: Wong Yuan Neng Supervisor: A/P Yeo Chai Kiat

Project Objectives:

Federated Learning (FL) is a privacy-preserving collaborative learning framework that allows participants to keep private data by aggregating only the model updates. In FL, clients together with their local models can access and impact the global model, which creates an attack surface for attackers making FL very vulnerable.



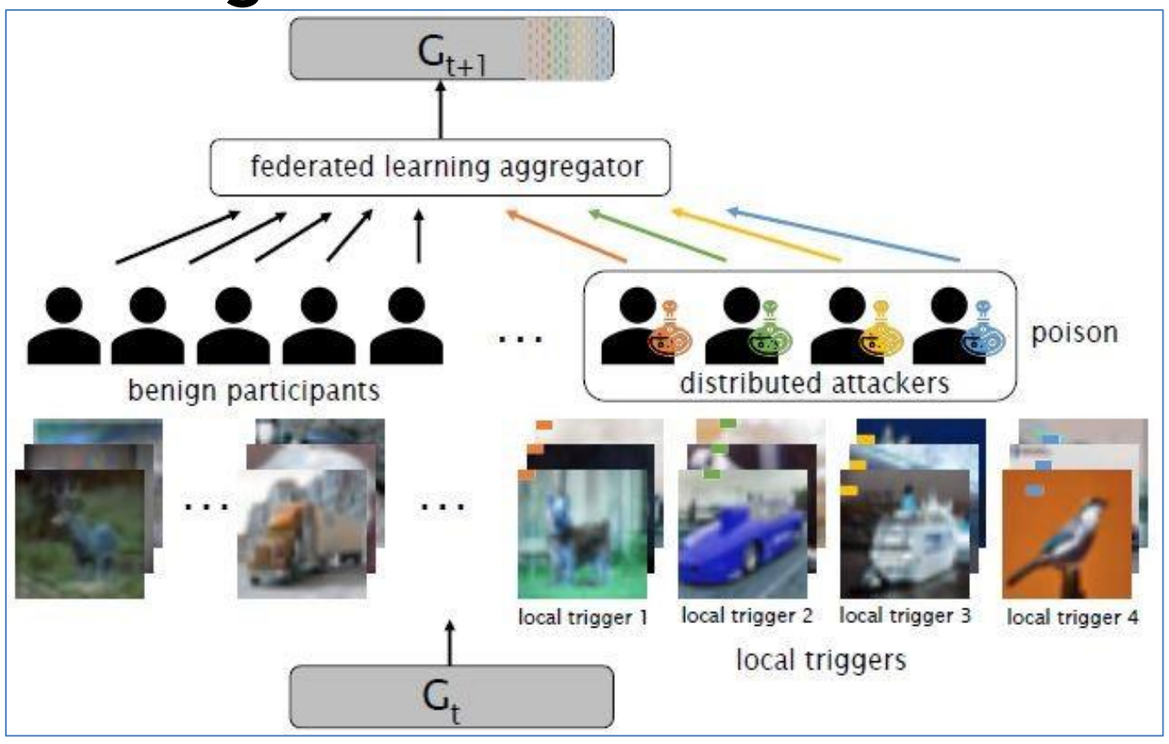
Collaborative learning risks leaking user data



FL guarantees privacy-preservation

This study aims to explore the effectiveness of some attack and some defense methodologies that can be applied to FL. The attack methodology studied is the Distributed Backdoor Attack (DBA), while the defense methodology studied is the Poisoning Defense Generative Adversarial Networks (PDGAN).

High-level overview of DBA



Distributing compositions of the complex pixel pattern of a global trigger to multiple malicious participants as their local triggers. DBA enhances the impacts of backdoor attacks and outperforms the standard centralized backdoor attacks.

High-level overview of PDGAN

Algorithm 1 Improved PDGAN method

Require: Auxiliary data set aux, Current global model G_t, List of participant model updates U_{t+1}, Round number t, GAN model $ganModel$

1: Train $ganModel$ using auxiliary data set, $ganModel \leftarrow \text{GANTRAIN}(aux)$
2: **if** $t \geq pdActiveRound$ **then**
3: Generate X_{fake} from Generator $gModel$ in $ganModel$ using random noise
4: $L, T \leftarrow \text{GETLABELS}(G_t, U_{t+1}, X_{fake})$
5: $benignModelUpdateList \leftarrow \text{PURGE}(L, T, X_{fake}, pdThreshold, U_{t+1})$

Round: 300

Backdoor target label: 8
Benign participants:
Malicious participants:

$M_{t+1}^k = G_t + U_{t+1}^k$

M_{t+1}^1 M_{t+1}^2 ... ... M_{t+1}^{10}

	Backdoored?											"True" label
Audit data set X_{fake}	Yes	x_a	3	8	5	3	8	3	3	3	8	3	3
	No	x_b	5	5	5	5	5	1	5	5	5	5	5
	...	...	...	...	...	...	...	...	...	...	...	...	...
	No	x_{p-1}	3	3	3	3	3	3	3	5	3	3	3
	Yes	x_p	5	8	5	5	8	5	5	5	8	5	5
Model accuracy			80	60	70	78	65	75	82	85	59	84	thresholdAcc = 73.8(1-pdThreshold)
Purge model?													

Use of GAN, which is a state-of-the-art machine learning technique, to generate new and unseen data to audit model accuracies & detect adversaries.