PRAGMA 39 @ Universitas Yarsi
Research Update Lightning Talk

Department of Computer Science

Faculty of Science & Technology, Thammasat University

Presented by **Prapaporn (Nan) Rattanatamrong**

June 21-24th, 2023, Jakarta, Indonesia

THAMMASAT UNIVERSITY
FACULTY OF SCIENCE & TECHNOLOGY

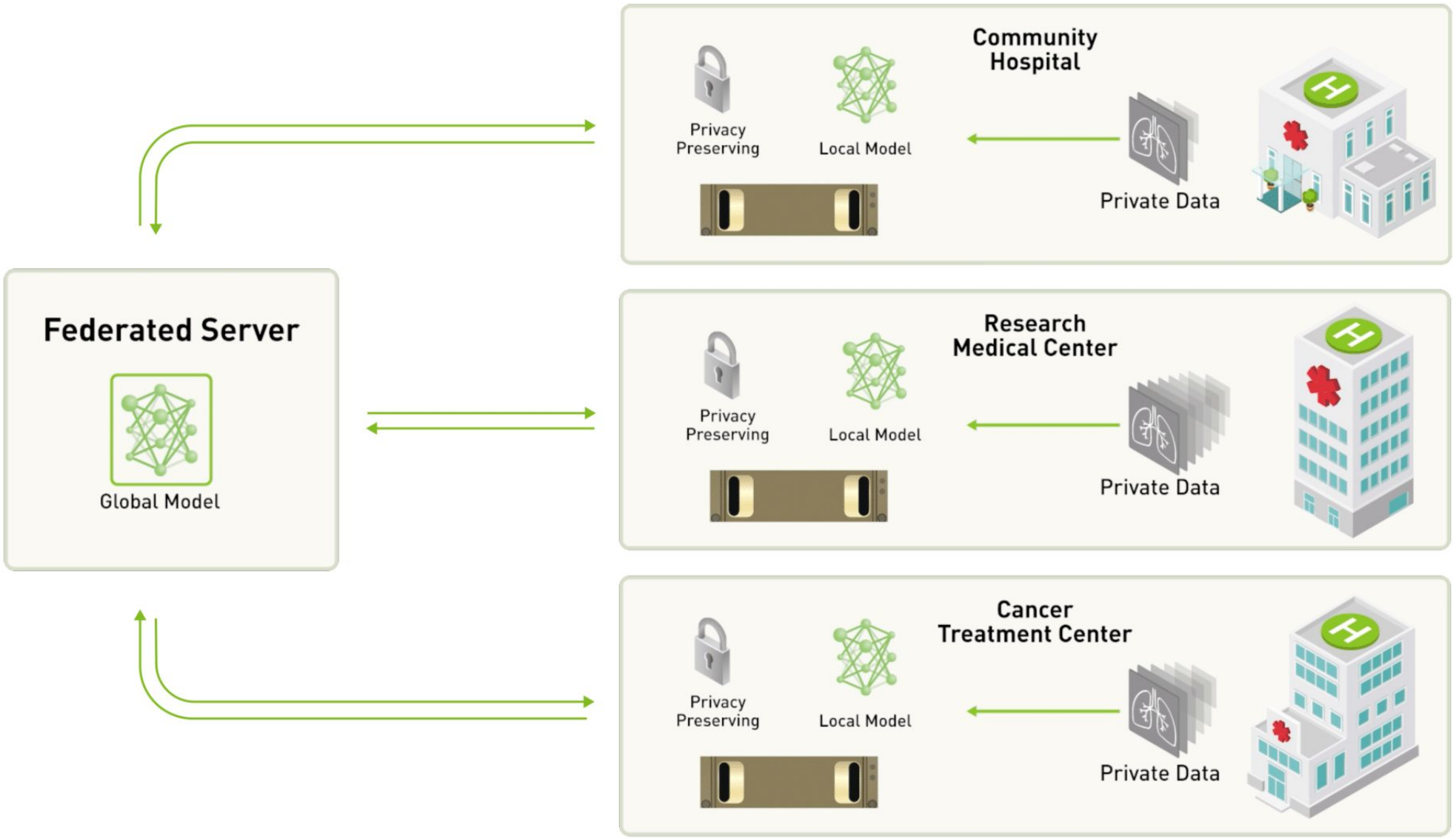
Ongoing Projects

- **GAN-based Data Augmentation Framework for Federated Learning**
- Decision Support Systems for Disaster Management
- Digital Twins over Edge-Cloud Continuum

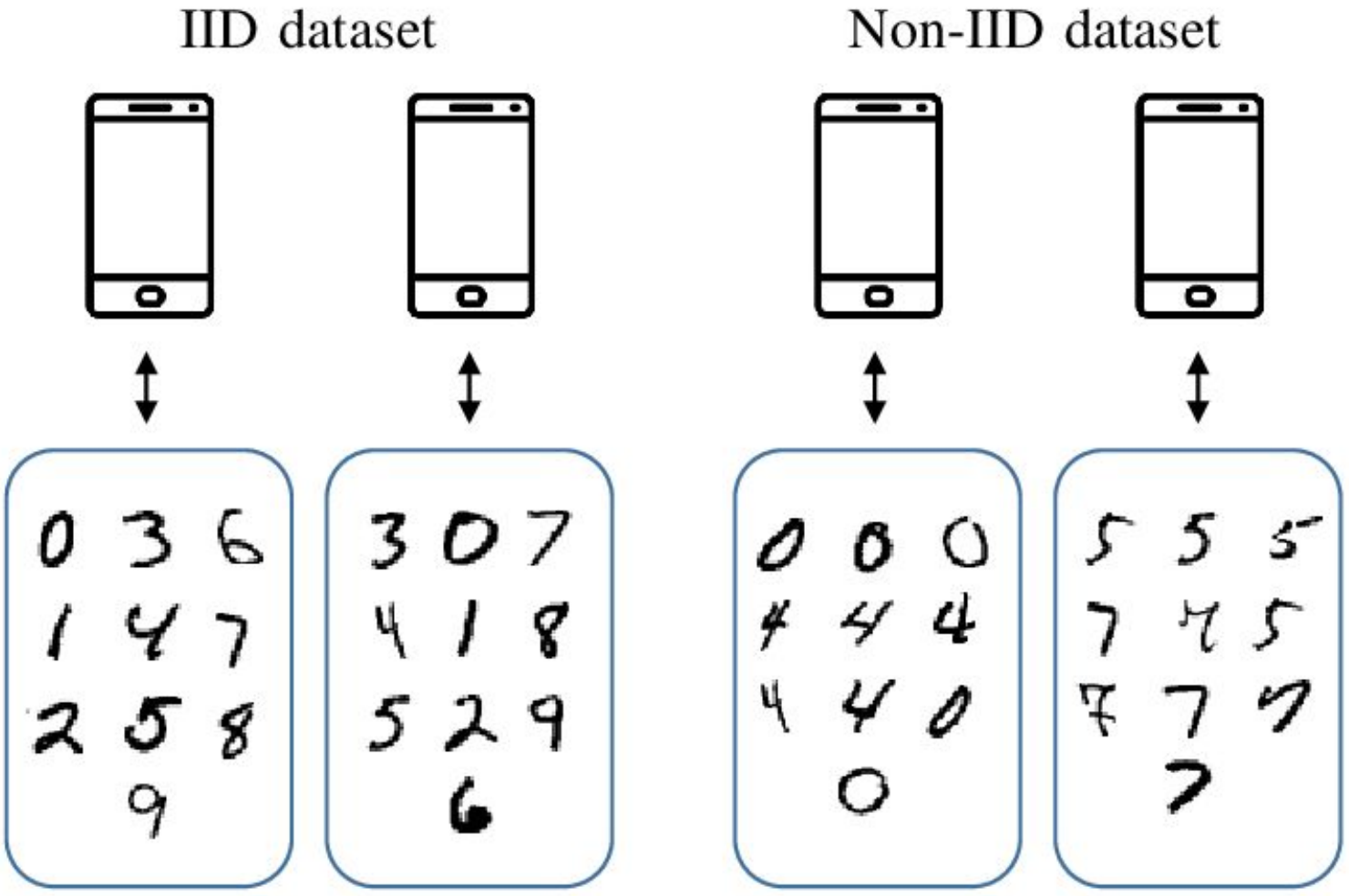
GAN-based Data Augmentation Framework for Federated Learning

Thiti Chuenbubpha, Prapaporn Rattanatamrong, Thapana Boonchoo, Jason Haga

Federated Learning (FL)



Federated Learning Performance for Non-IID data



- **Non-independently and identically distributed (non-IID)** data
- Decreased accuracy in FL is inevitable when dealing with non-IID data.
  - Weight divergence in local models

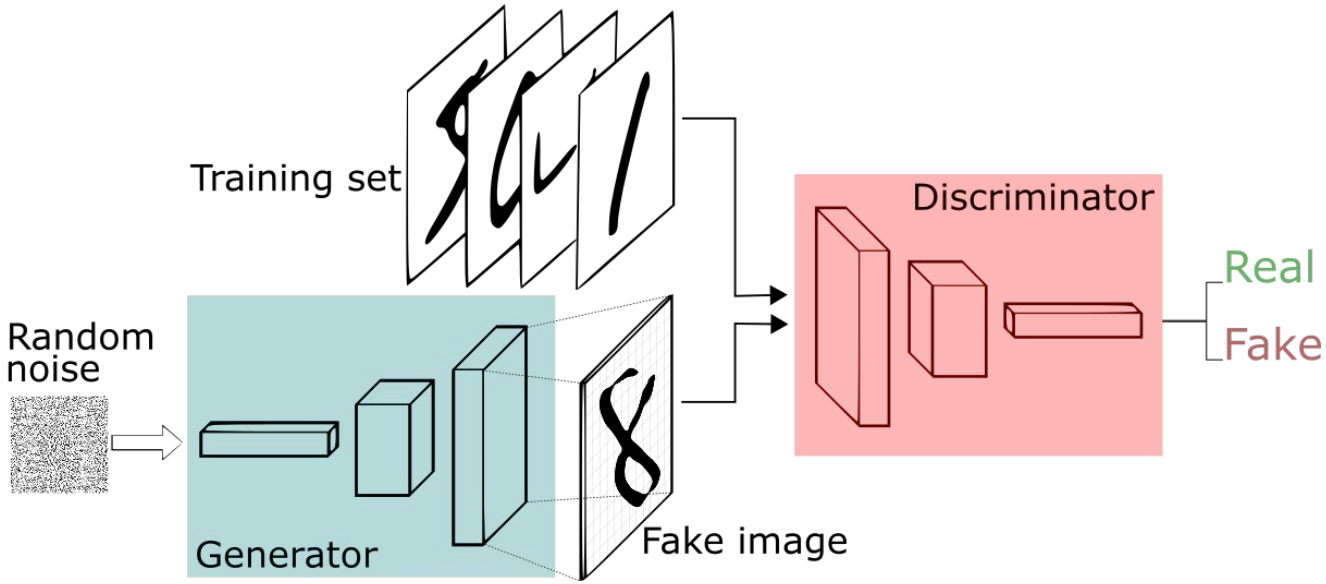
Ref : https://www.researchgate.net/publication/344017123_Wireless_for_Machine_Learning

Solving Non-IID using data augmentation

Data augmentation can help with Non-IID data

by create diverse synthetic samples.

The augmented data can be used during local training, promoting a more balanced and representative dataset

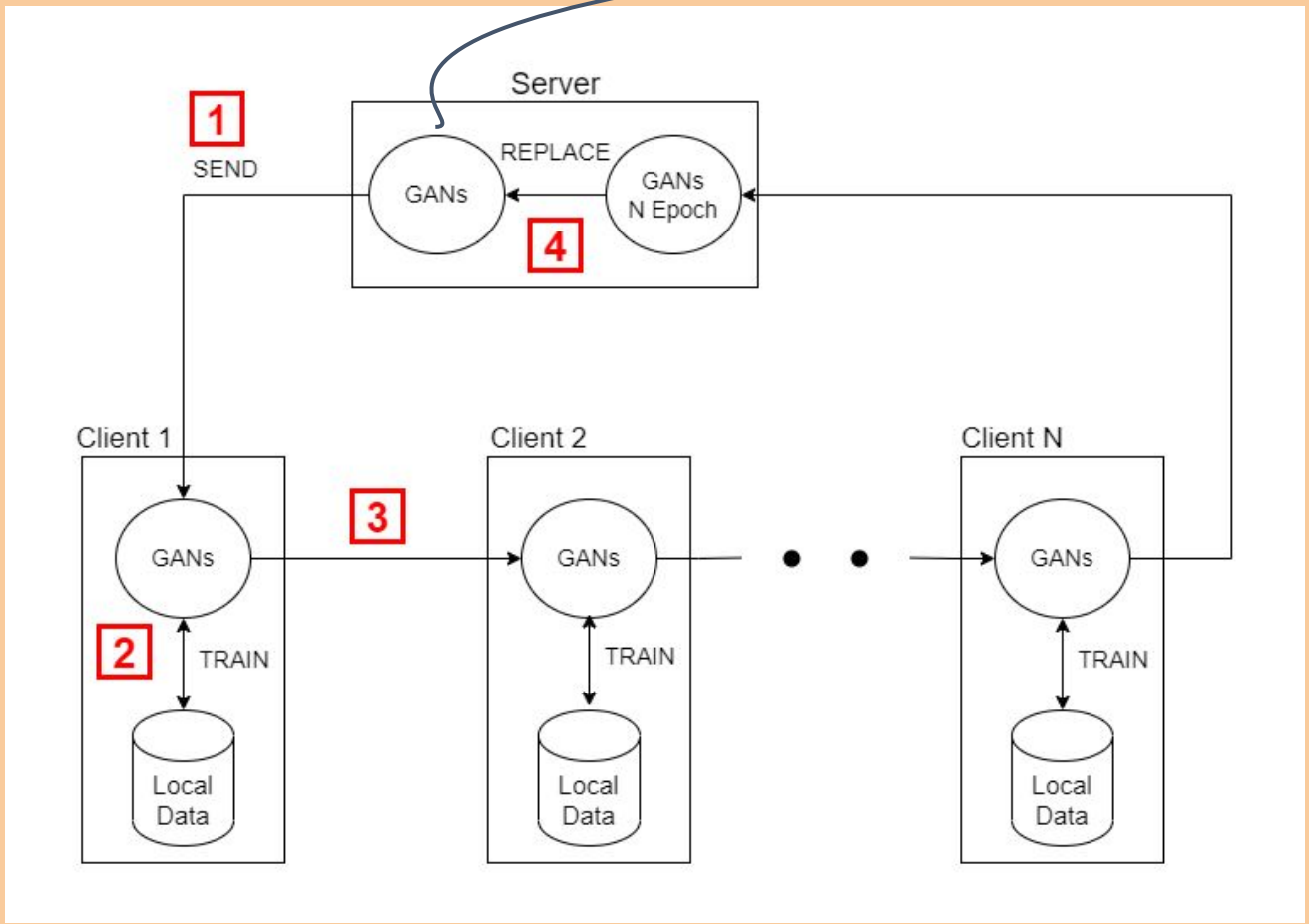


Using Generative Adversarial Networks (GANs) can help addressing the non-IID issue in FL

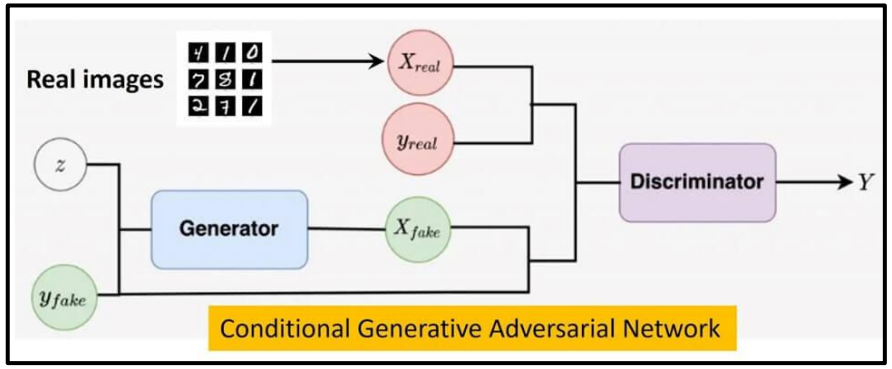
GANs can be trained on local data from individual devices to generate synthetic data. By sharing these generated samples instead of raw data, GANs enable the creation of more balanced and representative datasets for model training across devices

Conditional GANs Training in FL nodes

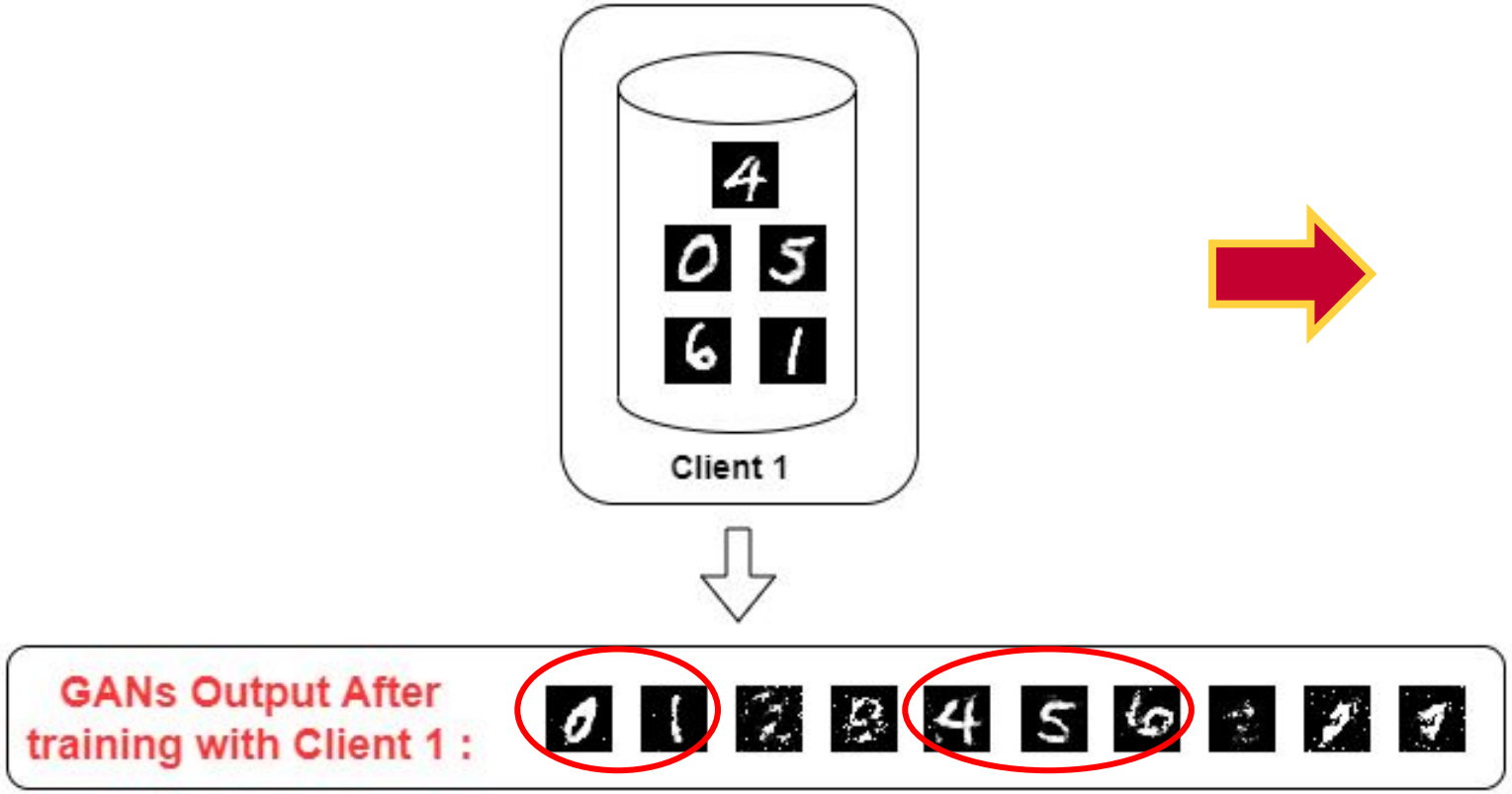
GANs training



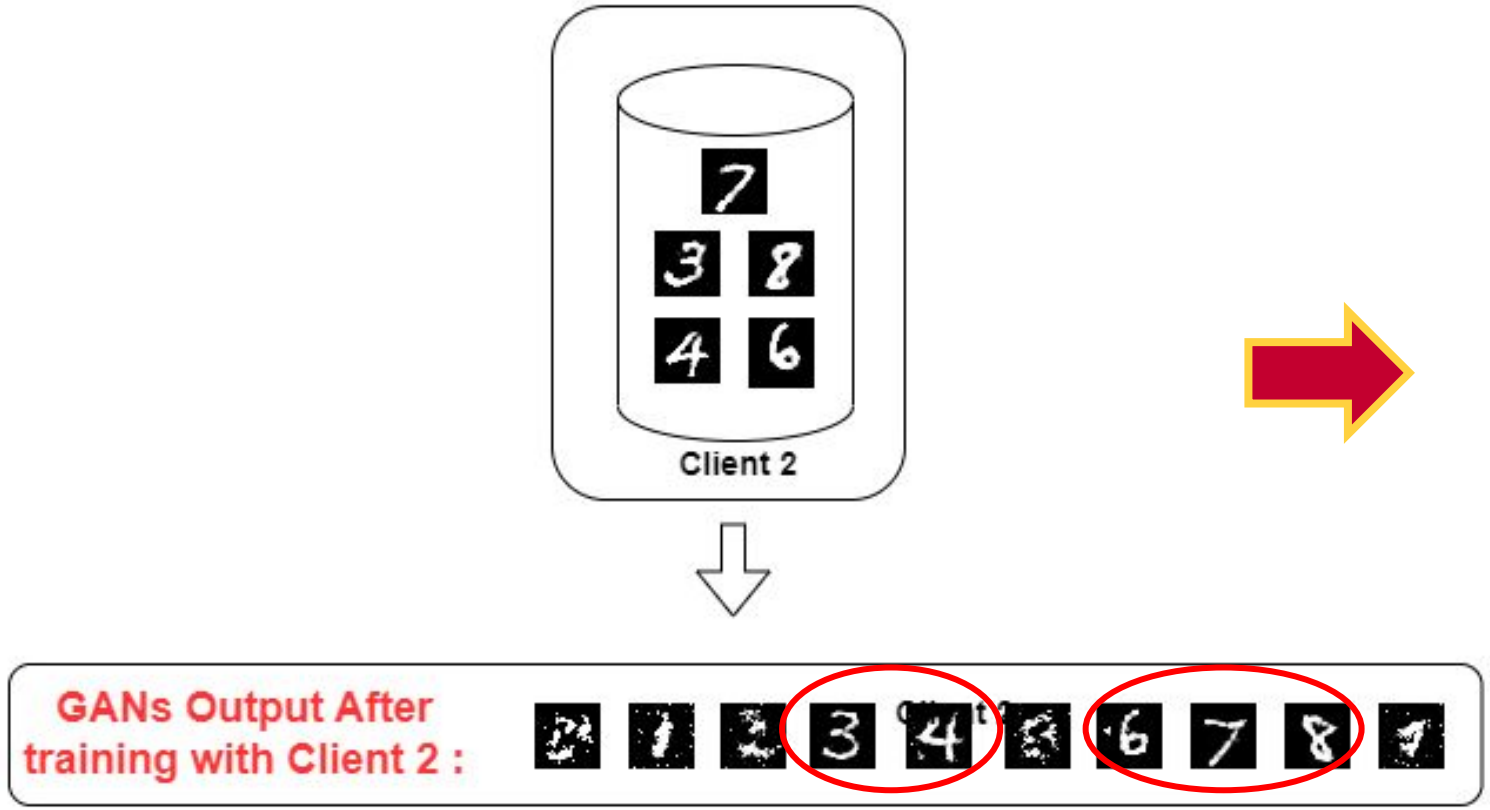
Conditional GANs



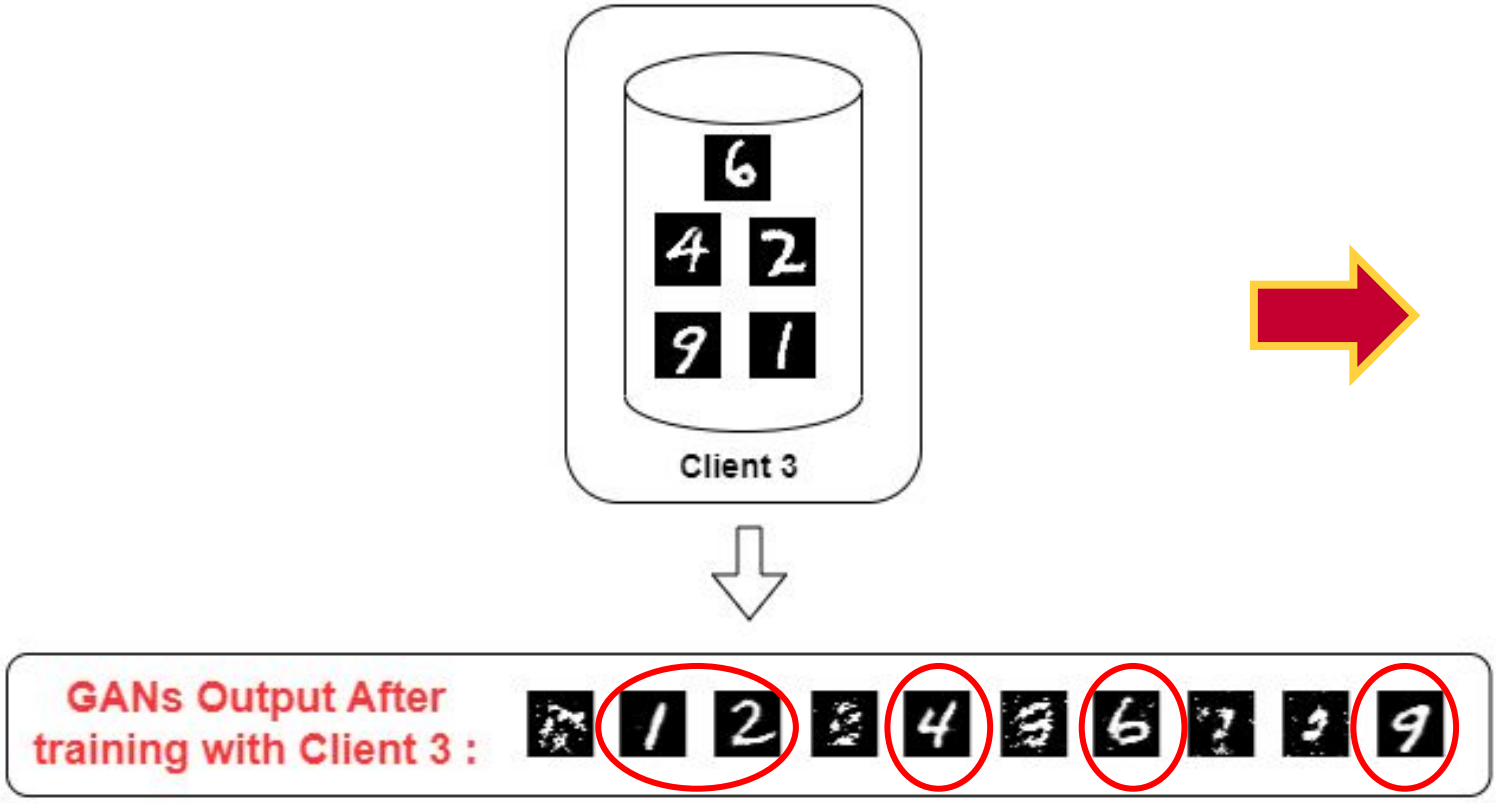
Issue with GANs Training in Distributed FL Nodes



Issue with GANs Training in Distributed FL Nodes



Issue with GANs Training in Distributed FL Nodes



Main issues to address:

1. The global model in federated learning has its **performance decreases when dealing with non-IID data** (considering only label distribution skew).

2. **GANs training on a distributed system** can result in **bias** of the model particularly on **the last seen classes**.

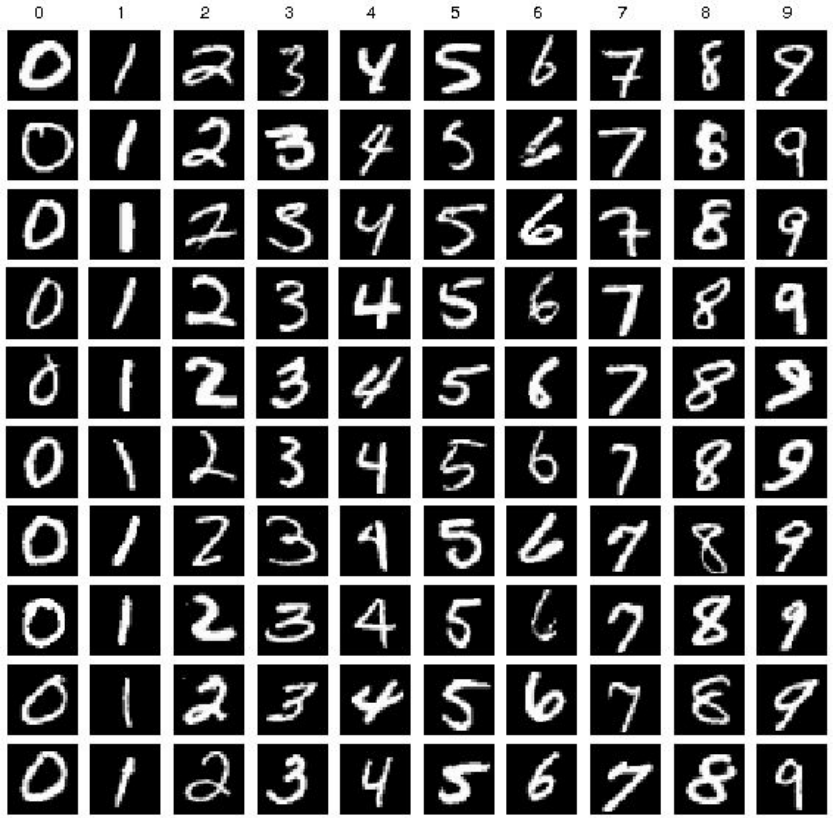
3. When data is shared among clients to correct Non-IID, the **privacy of training data in FL** is leaked.

Configuration Variables for FL of CNNs

Independent Variables	Possible Values
DATASET	MNIST, FMNIST
All possible classes in the DATASET	10
Number of nodes in FL (N)	10, 50
Number of available classes in each node (L_n)	2, 5
% of local data used for each round of GANs training (SUBSET SIZE)	1, 25, 50, 75, 100

DATASETS

MNIST



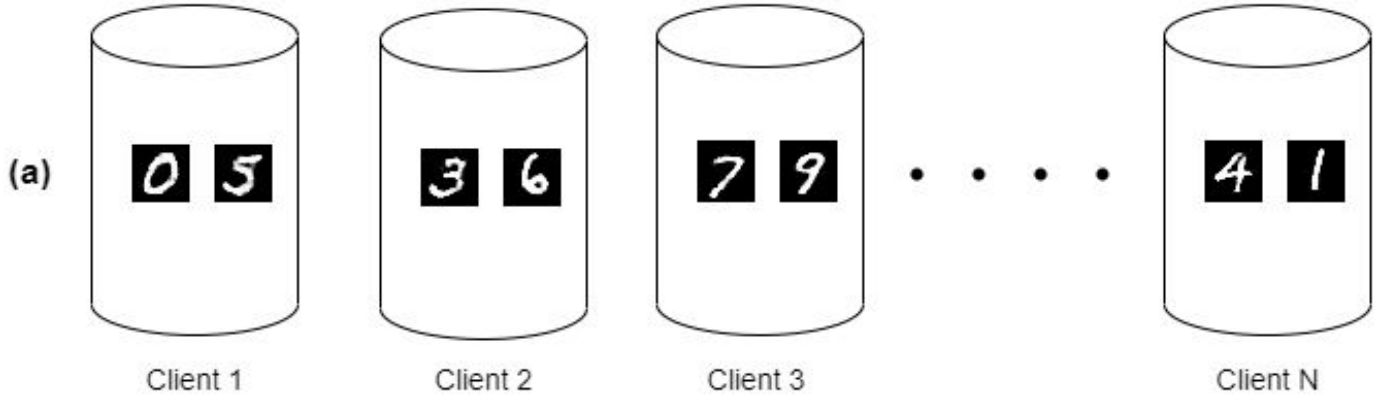
FMNIST

70,000 of grayscale image with 28x28 pixel
Training dataset = 60,000
Test dataset = 10,000

Ref : https://arxiv.org/abs/2102.02079

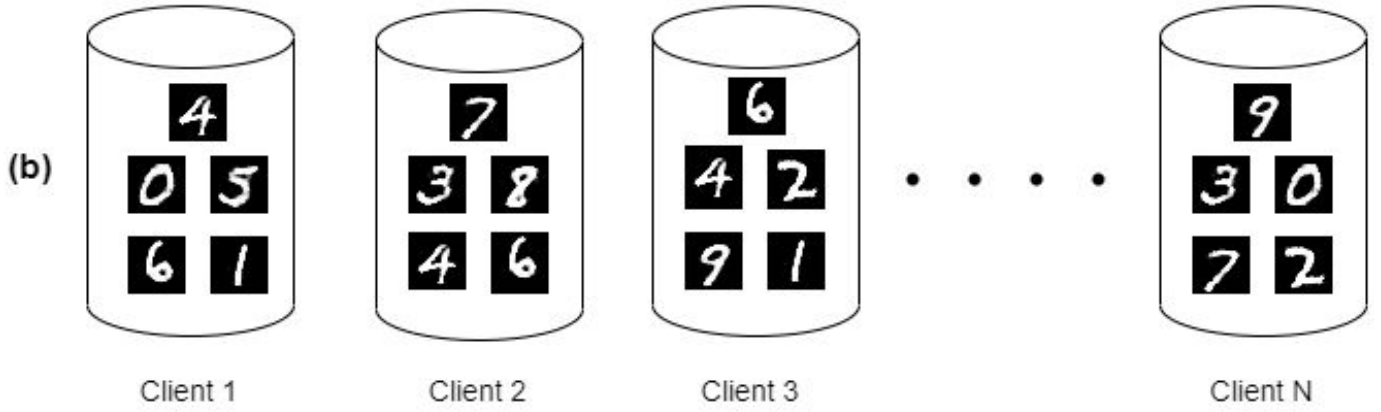
Non-IID Settings

FL with N Client



$L_n = 2$

Severely Non-IID Data



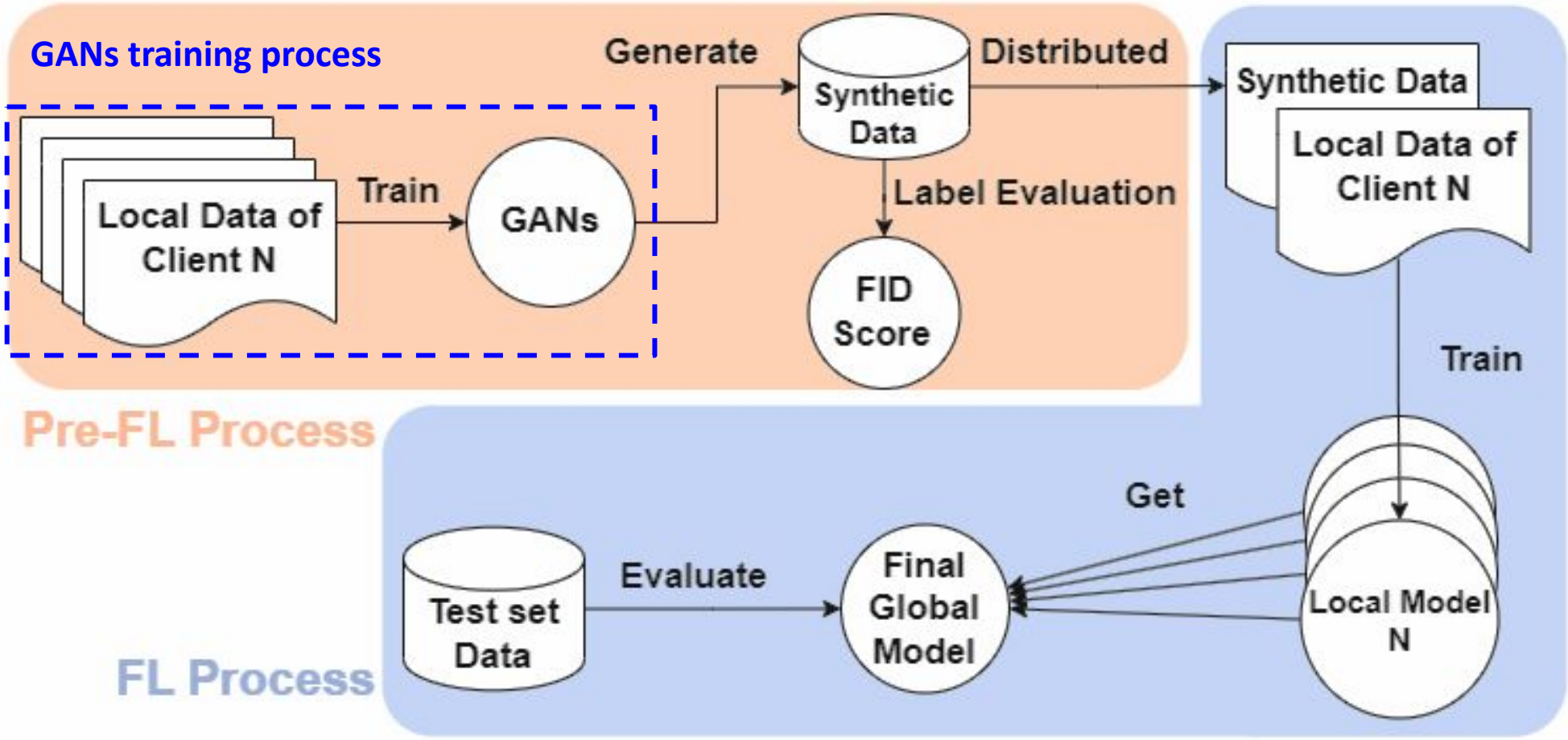
$L_n = 5$

Baseline Experiment

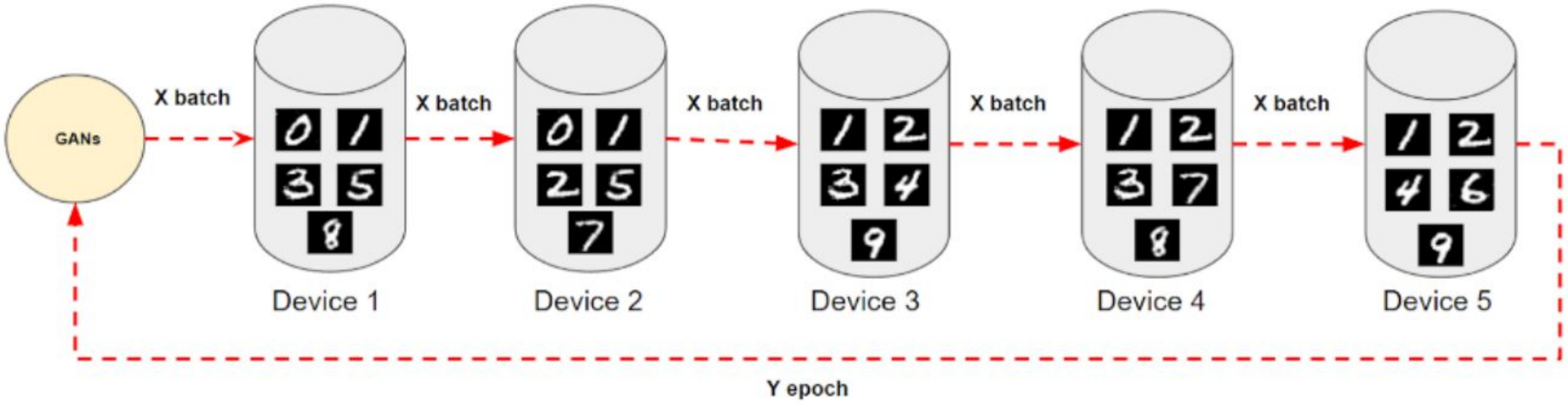
The FL global model's performance decreases when data is non-IID, especially for complex tasks

No. of FL Nodes (N)	DATASET	Data Distribution	No. of classes in each node (L_n)	FL Global Model's Accuracy
5	MNIST	IID	10	98.41
		Non-IID	5	96.35
		Non-IID	2	83.41
	FMNIST	IID	10	85.35
		Non-IID	5	67.91
		Non-IID	2	58.48
10	MNIST	IID	10	96.95
		Non-IID	5	96.55
		Non-IID	2	63.58
	FMNIST	IID	10	82.07
		Non-IID	5	67.77
		Non-IID	2	53.82

Proposed GANs Augmented IID - Federated Learning (GAIID-FL) Framework



Subsetting Local Data in GANs Training Process



Training size X = SUBSET SIZE * Number of Local Data

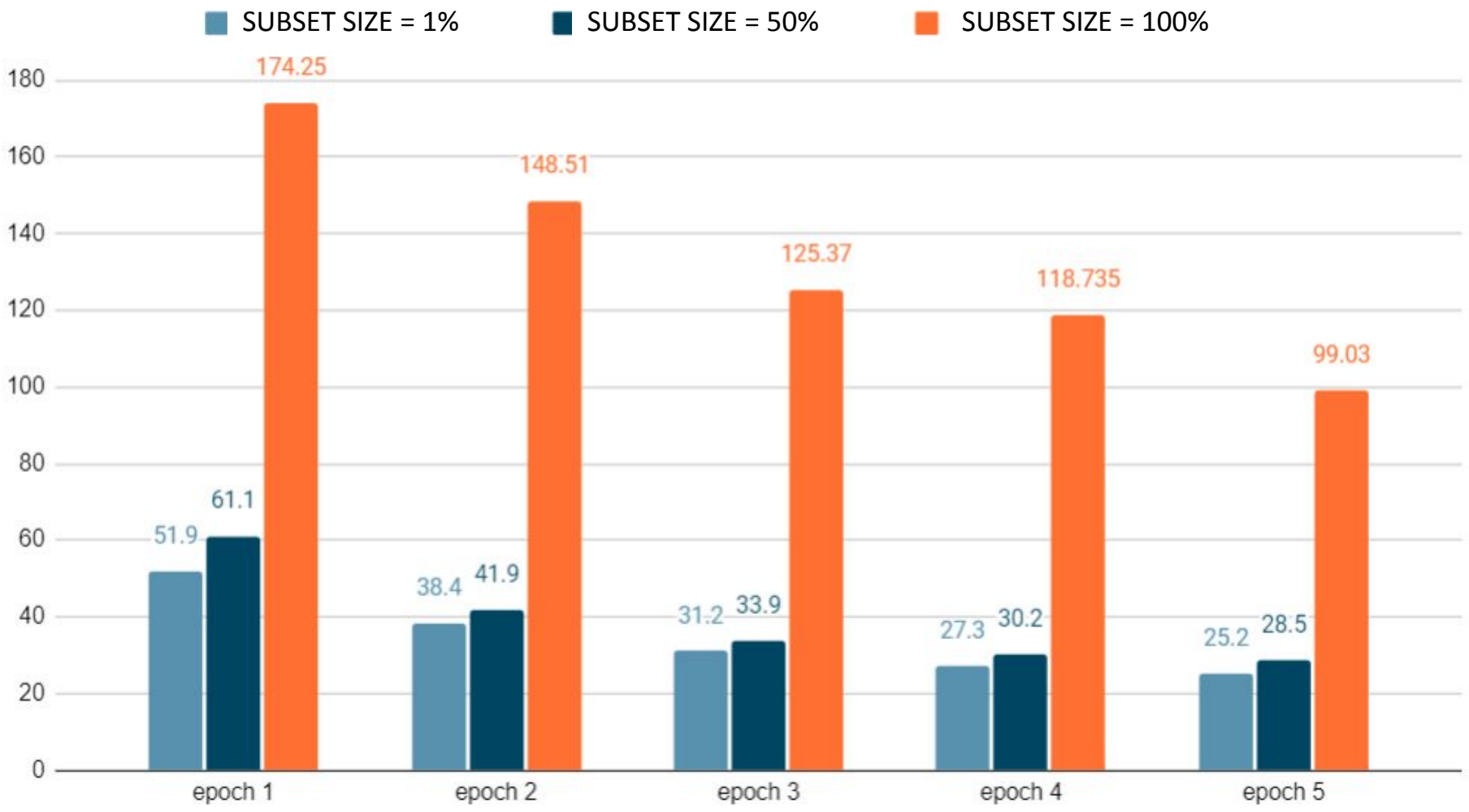
Optimal SUBSET SIZE for GANs Training

The smaller subset size for training can create images with lower FID score (i.e., more realistic)

SUBSET SIZE	No. of FL Rounds in GANs Training	Communicated Data (MB)	FID Score of Resulting GANs
100%	10	832	N/A
75%	20	1,665	79.5
50%	20	1,665	28.5
25%	40	3,330	26.3
1%	1,870	155,677	25.2

Optimal SUBSET SIZE

The optimal subset size for GANs training is at around 50% of local data size, while still keeping communication overhead acceptable



N = 10

L_n = 2

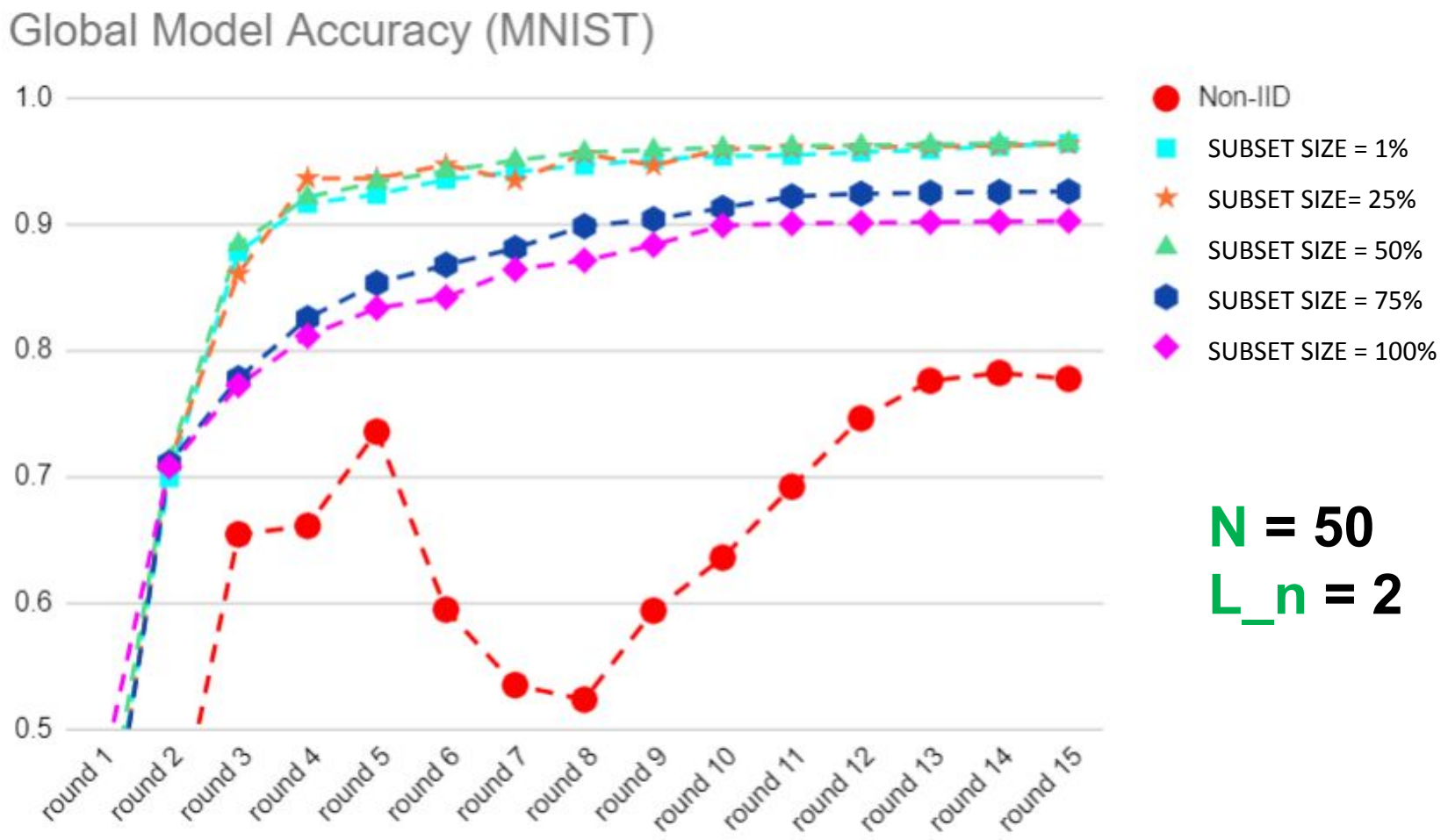
Optimal Number of Synthetic Image Data

Generally, 1,000 images per class is recommended but we found that augment each class in each node with 100 images is sufficient to achieve good accuracy

N = 50

DATASET	Data Distribution	L_n	No. of synthetic data to augment in each FL node	FL Global Model's Accuracy	Minutes spent for FL
MNIST	IID	10	0	79.59	30
	Non-IID	5	0	62.14	32
	Non-IID	5	100	92.74	43
	Non-IID	5	1,000	97.12	147
	Non-IID	2	0	43.16	33
	Non-IID	2	100	91.68	47
	Non-IID	2	1,000	96.91	151
FMNIST	IID	10	0	63.81	32
	Non-IID	5	0	50.77	33
	Non-IID	5	100	73.01	45
	Non-IID	5	1,000	80.73	153
	Non-IID	2	0	34.69	35
	Non-IID	2	100	72.40	44
	Non-IID	2	1,000	80.20	159

Addressing Non-IID in FL with Difference SUBSET SIZE

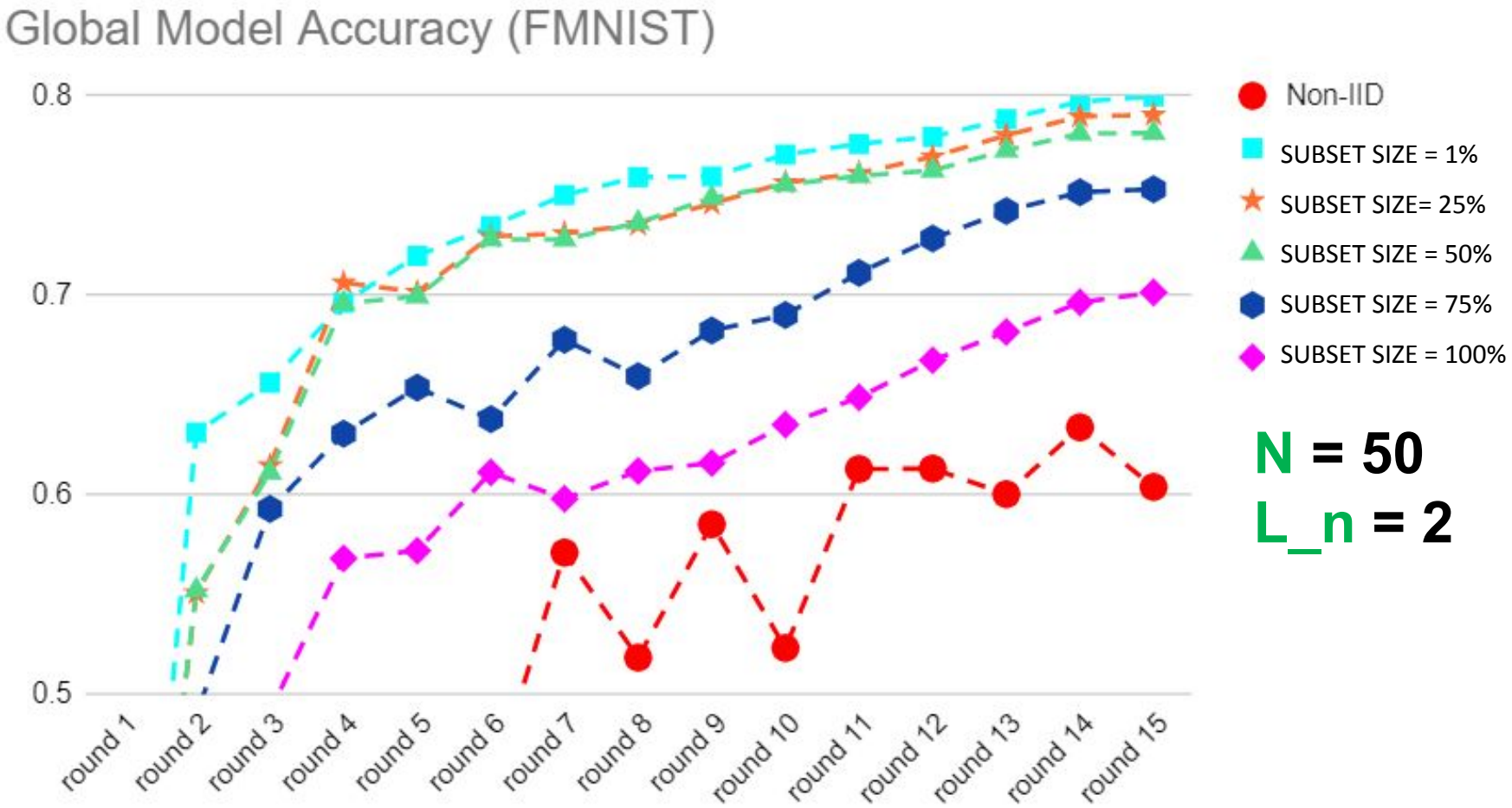


N = 50

L_n = 2

Non-IID can unstabilize the model's accuracy and we can fix that using augmented data from the GANs trained using our GAIID-FL framework's approach

Addressing Non-IID in FL with Difference SUBSET SIZE (2)



N = 50

L_n = 2

For more complicated tasks, the global model might converge more slowly but it presents the same trend that the non-IID issue can be fixed using GAIID-FL

Addressing Non-IID in FL with Difference Number of Synthetic Image

The GAIID-FL framework can solve non-IID data issue (achieve comparably good or superior than the IID case)

N = 10

DATASET	Data Dist.	L_n	Augmented synthetic data	FL Global Model's Accuracy
MNIST	IID	10	0	96.95
	Non-IID	5	0	96.55
		5	100	97.51
		2	0	63.58
		2	100	96.06
FMNIST	IID	10	0	82.07
	Non-IID	5	0	67.77
		5	100	78.11
		2	0	53.82
		2	100	79.83

N = 50

DATASET	Data Dist.	L_n	Augmented synthetic data	FL Global Model's Accuracy
MNIST	IID	10	0	79.59
	Non-IID	5	0	62.14
		5	100	92.43
		2	0	43.16
		2	100	91.25
FMNIST	IID	10	0	63.81
	Non-IID	5	0	50.77
		5	100	72.68
		2	0	34.69
		2	100	72.11

Conclusions

1. Correcting Non-IID with synthetic images can improve global model accuracy when compared to global models trained with Non-IID data.

2. Images with lower FID score, which indicates greater similarity of synthetic images to real images, yields better quality of FL when used for augmentation.

3. GAIID-FL can be used to improve FL performance regardless of whether or not the data distribution is non-IID.

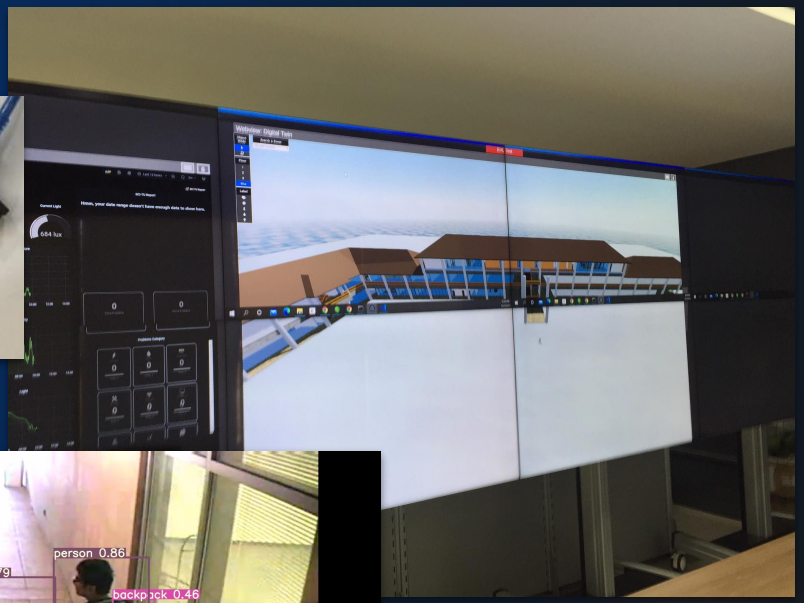
4. Using 100 images per class per client is sufficient to solve Non-IID in FL.

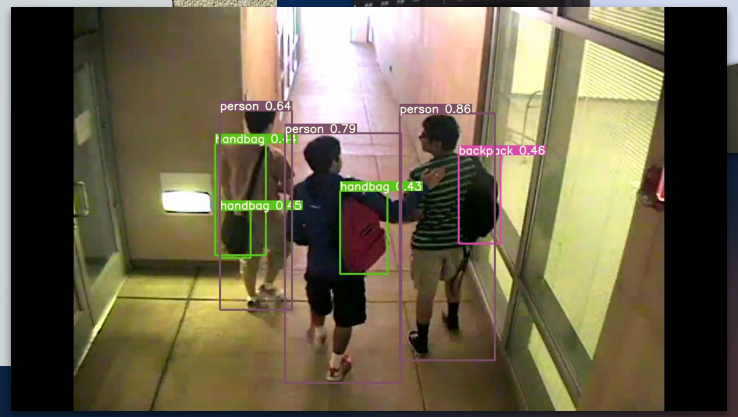
5. Local data has never left from the FL nodes, so no privacy risk.

Other Projects

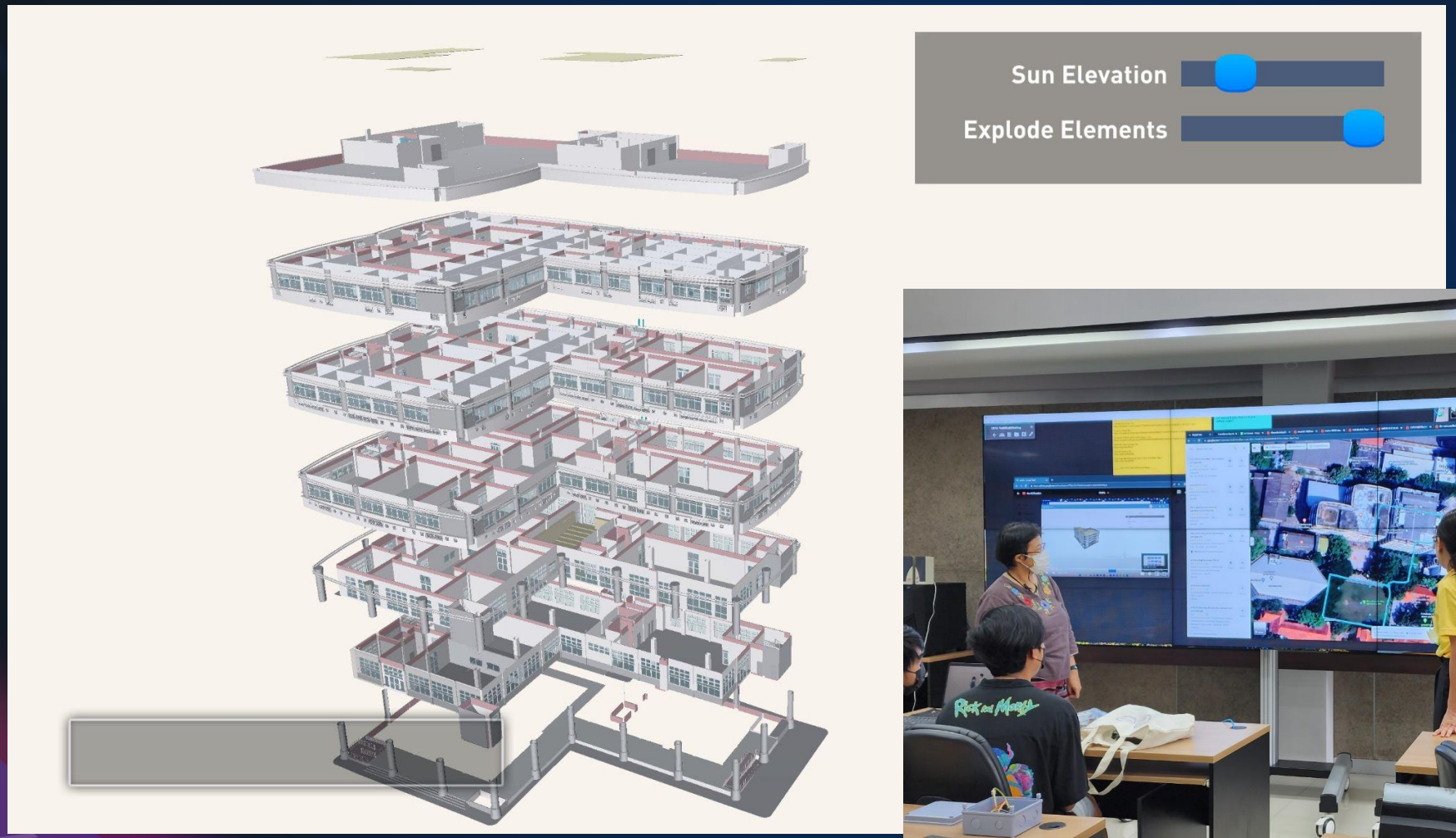
Computing over Edge-Cloud Continuum

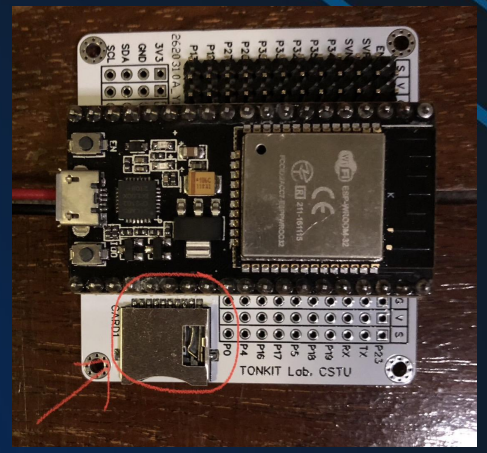
Digital Twins: SCI-TU-LC2



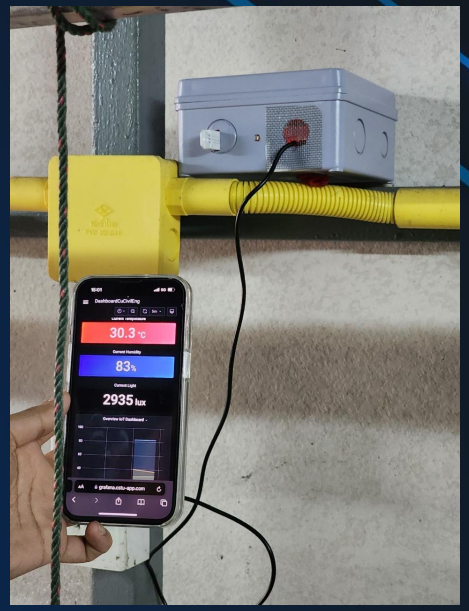


Digital Twins: CIVIL-CU

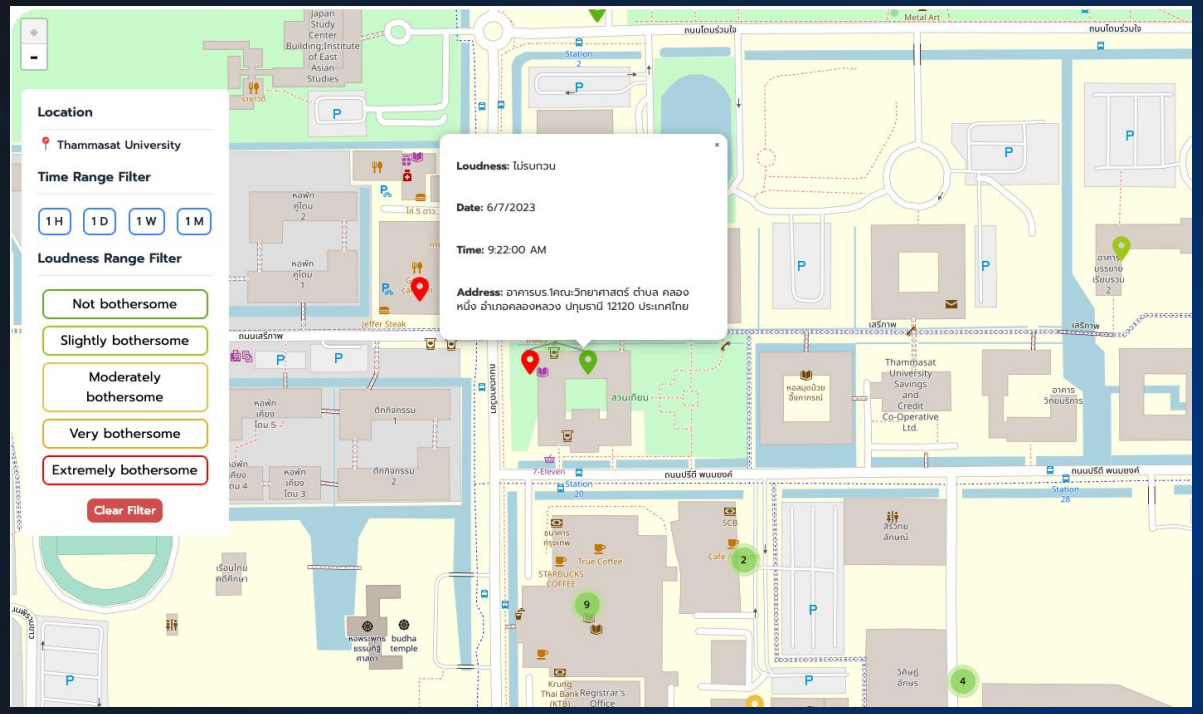


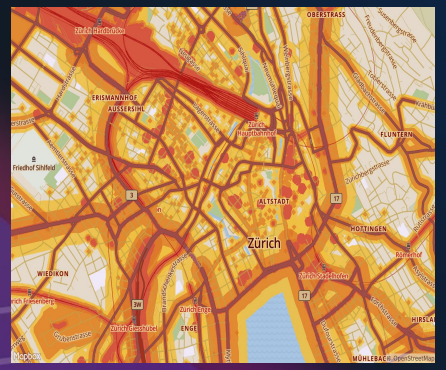


Digital Twins - Poultry Farmhouse



Urban Noise Mapping





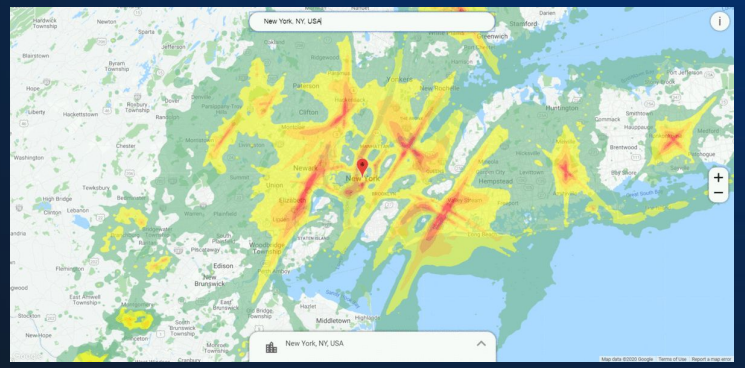
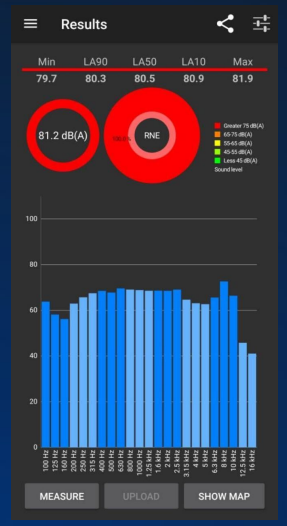
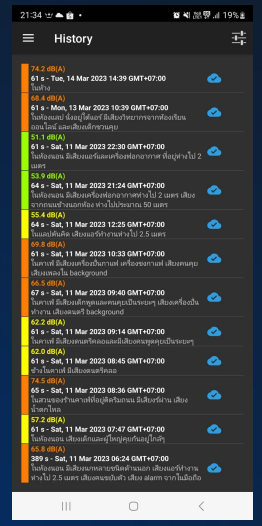
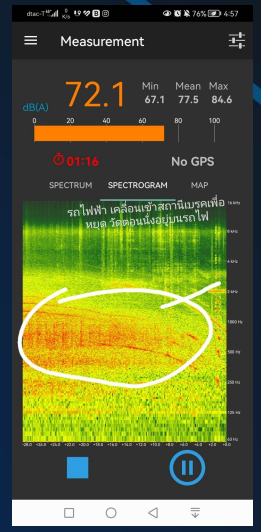
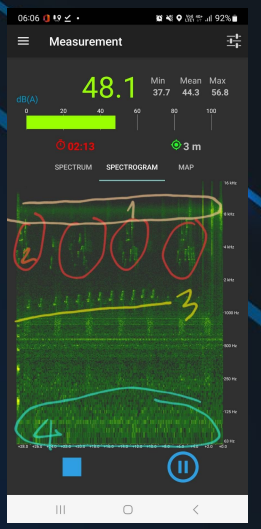


Image credit: https://noise-map.com/home/



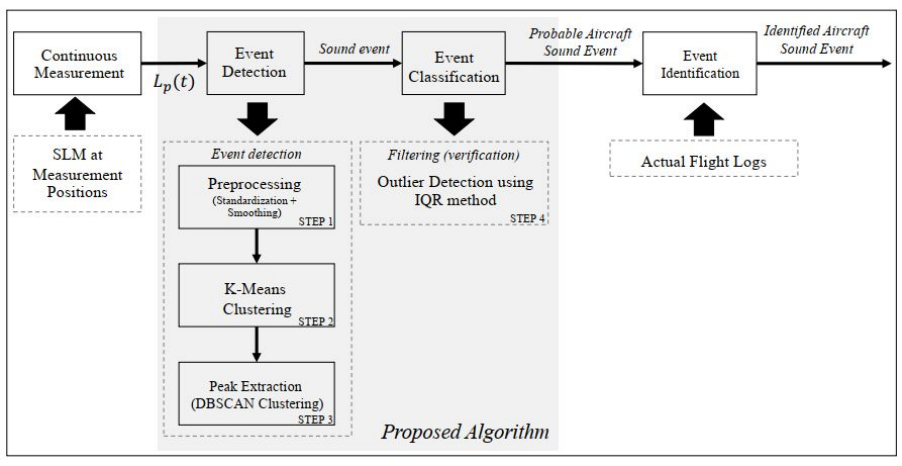




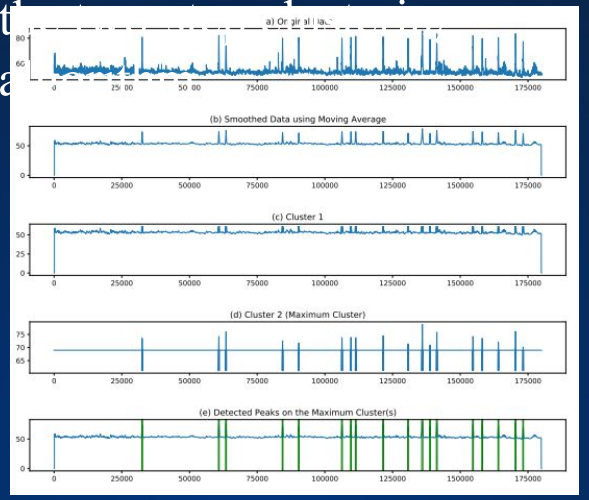


Algorithm for Calculation of the Measured Single Fly-over Aircraft Noise

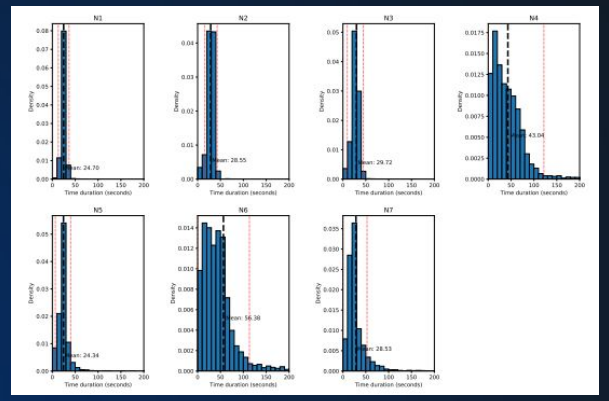
Data processing workflow



Results from the steps 2, 3, 4 of



Sample histogram results of Time duratio
The red dotted lines indicate outlier boun

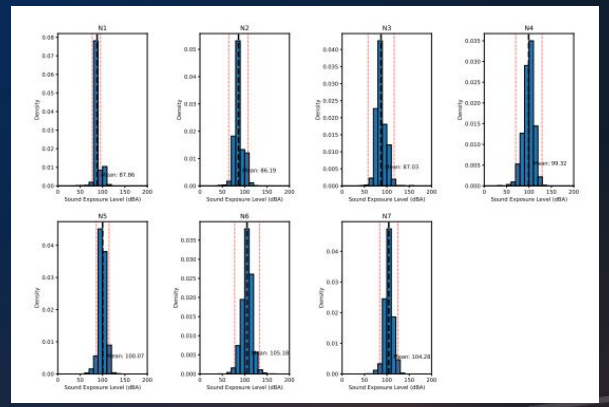


Noise Measurement Positions around BIA
Sources: Krittika Lertsawat, 2015

The calculated aircraft noise exposure levels in different indicators

	$L_{eq,24hr}$	L_{dn}	L_{den}
	dBA	dBA	dBA
Departure flights	93-114	100-118	110-119
Arrival flights	82-109	87-117	89-117

Sample histogram results of calculated single fly-over aircraft noise events.



T. Boonchoo, K. Lertsawat "Algorithm for Calculation of the Measured Single Fly-over Aircraft Noise", 24th International Congress on Acoustics, 2022 (Accepted).

QUESTION ?

THANK YOU