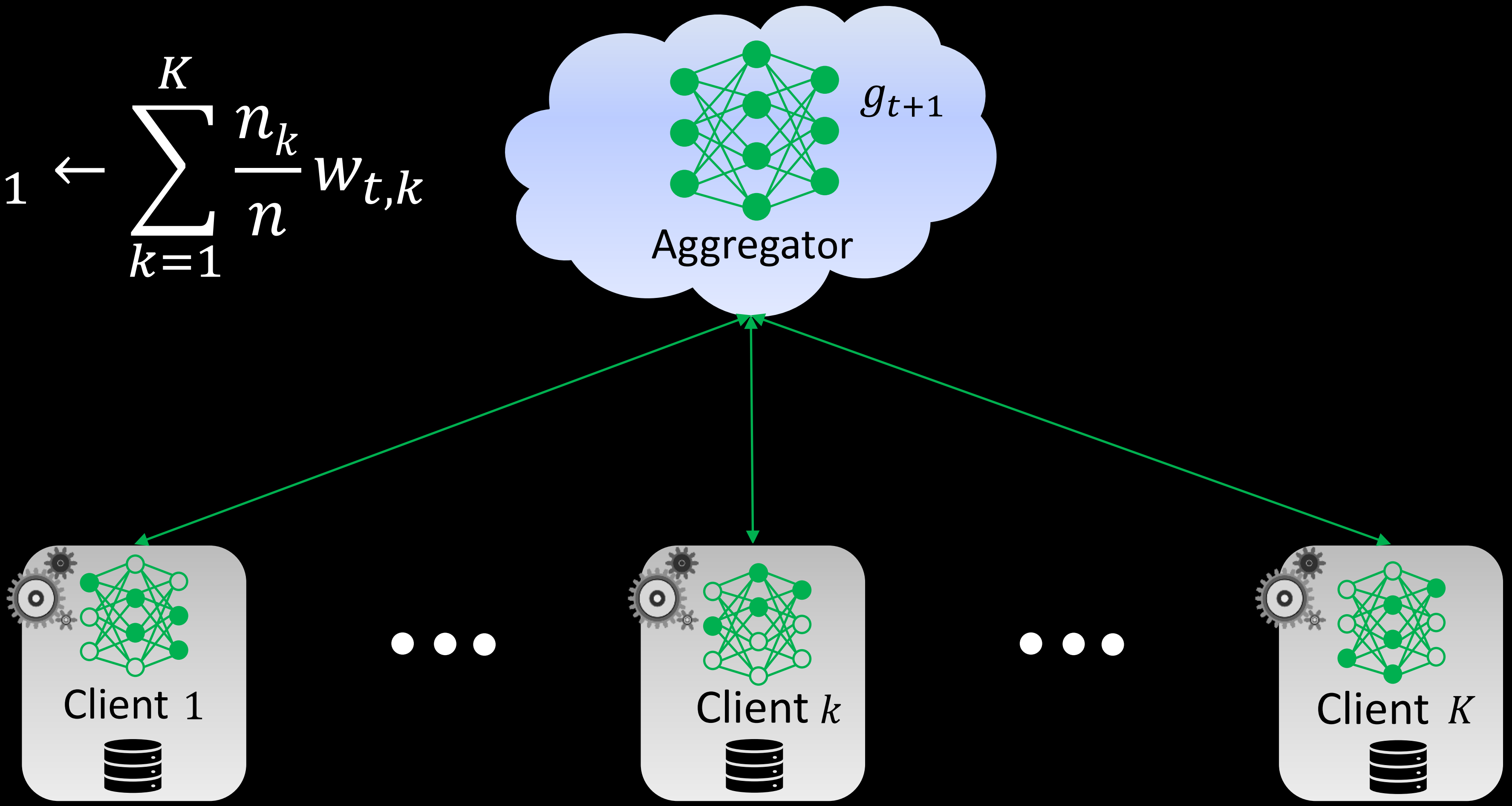
FreqFed: Frequency Analysis for Poisoning Detection in Federated Learning

Hossein Fereidooni, Alessandro Pegoraro, Phillip Rieger, Alexandra Dmitrienko and Ahmad-Reza Sadeghi,

NDSS 2024

Federated Learning Basics

$$g_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t,k}$$

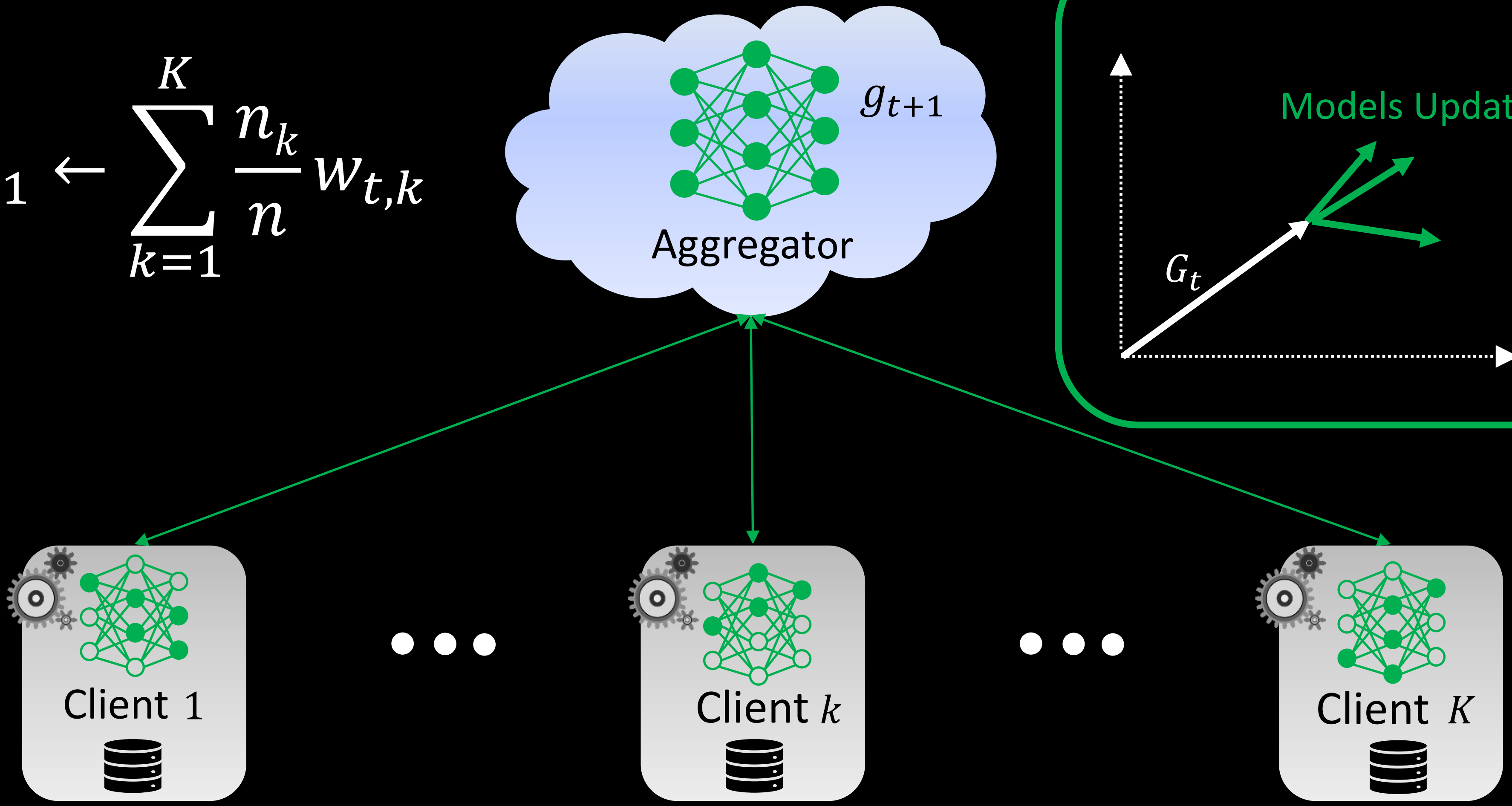


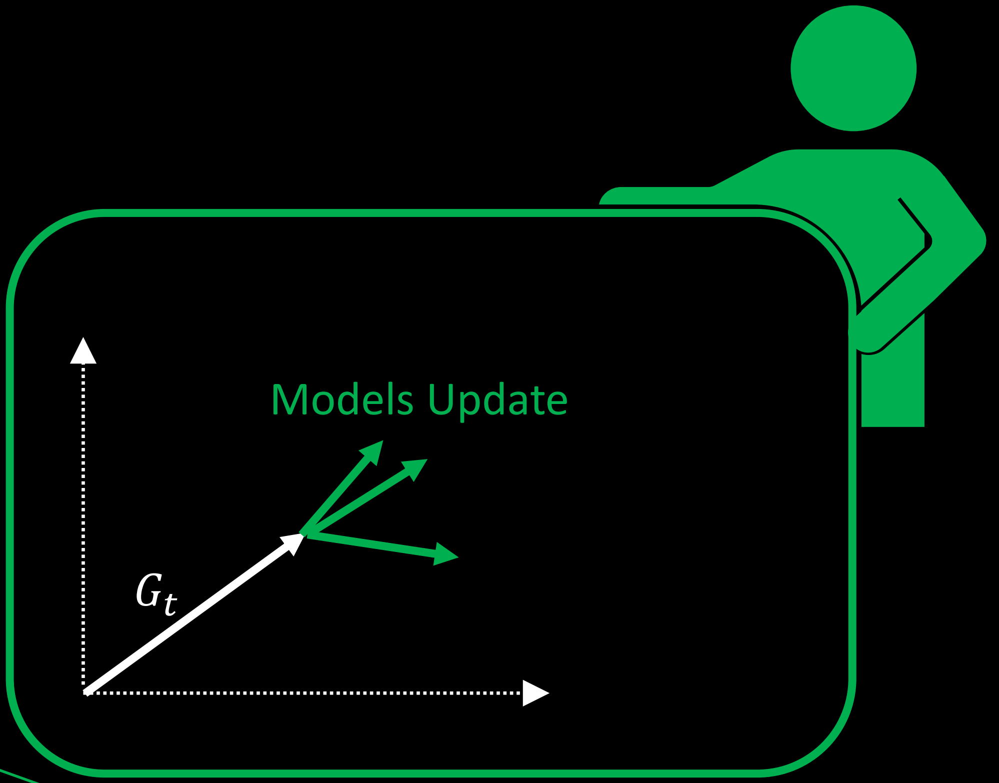
g_t: Parameters of global model
$w_{t,k}$: Parameters of client's model
K: Total number of clients

n_k: Number of samples for client k
n: Number of samples for all clients
t: Round index

Federated Learning Basics

$$g_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t,k}$$

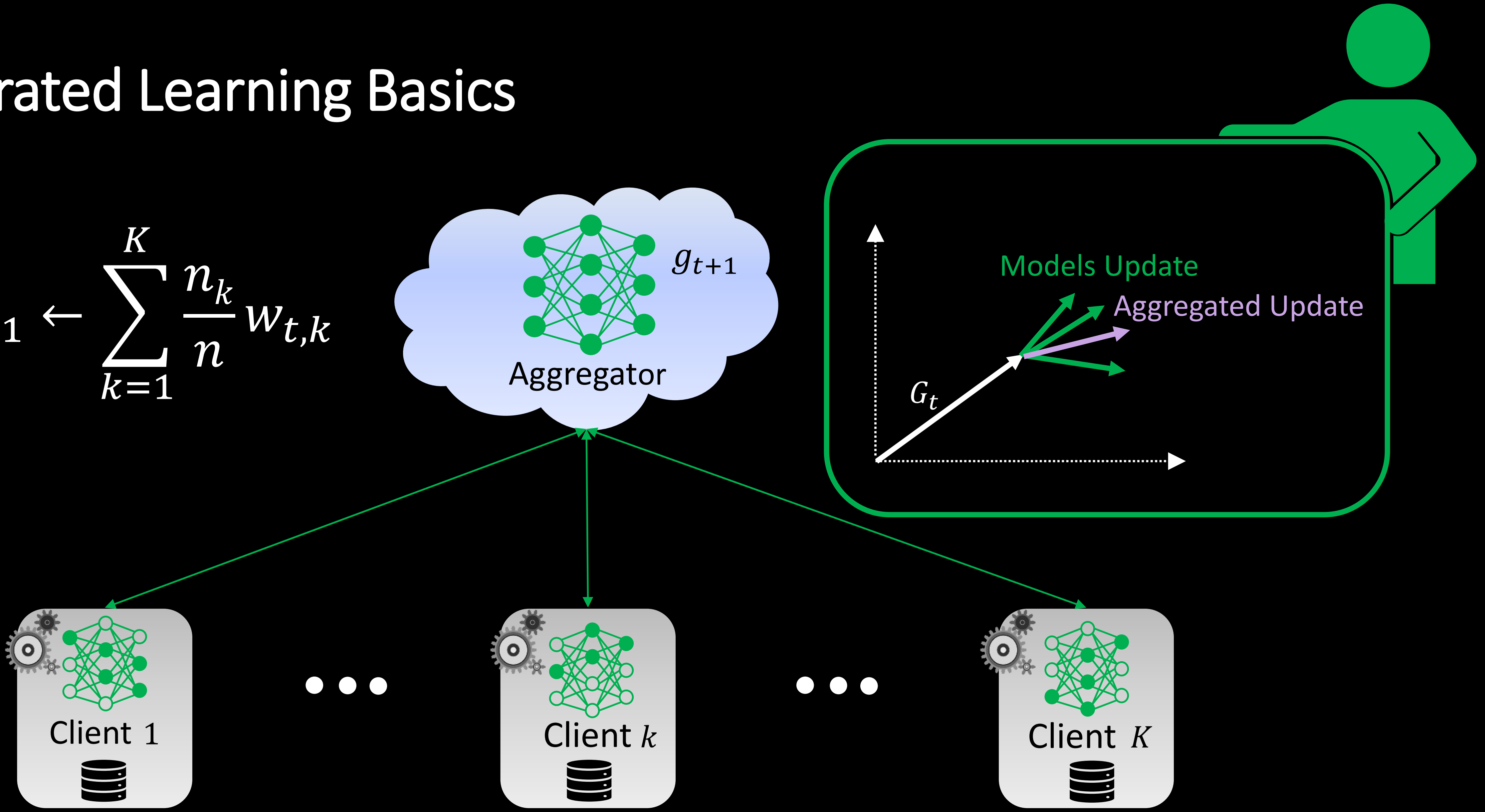




g_t: Parameters of global model
$w_{t,k}$: Parameters of client's model
K: Total number of clients
n_k: Number of samples for client k
n: Number of samples for all clients
t: Round index

Federated Learning Basics

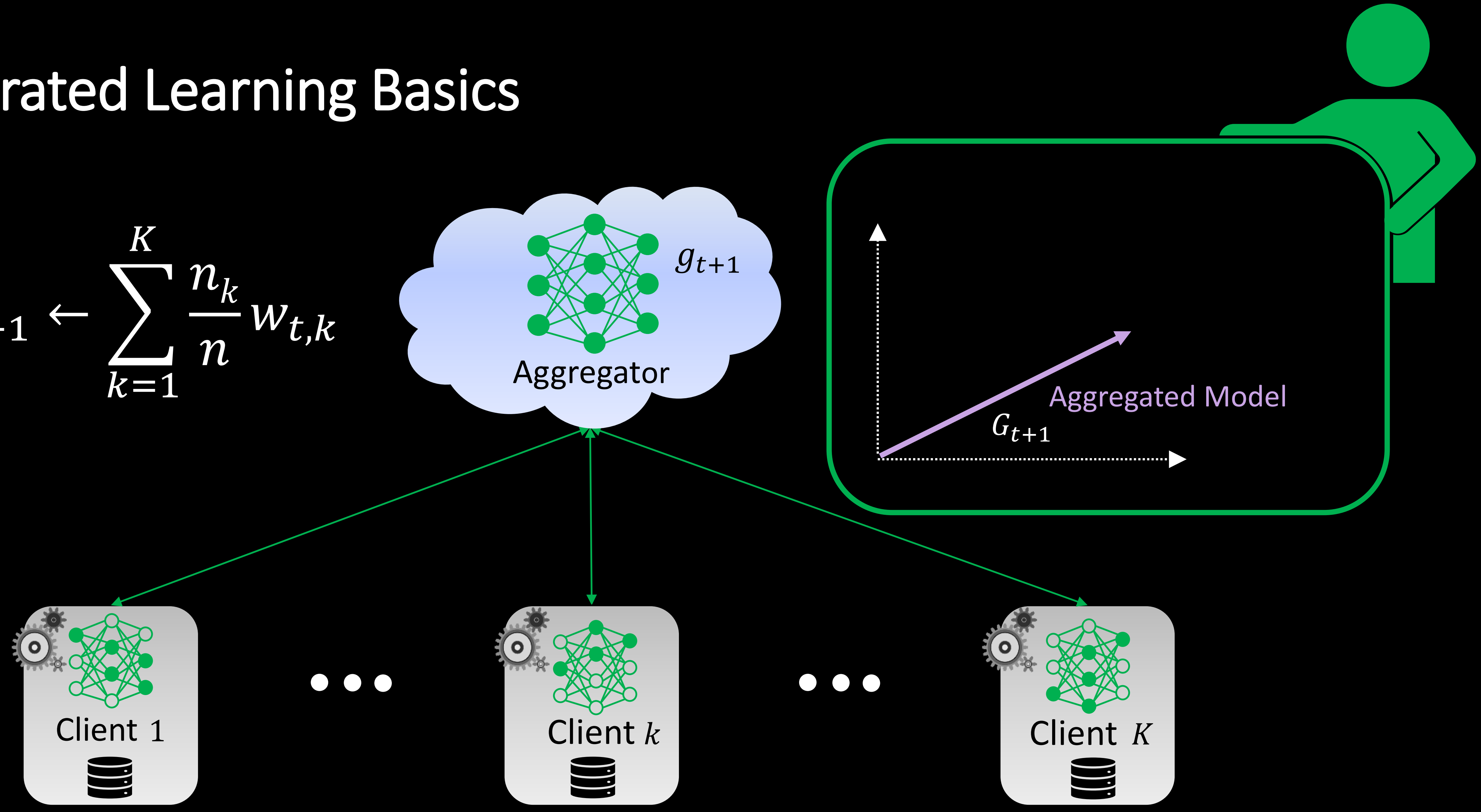
$$g_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t,k}$$



g_t: Parameters of global model
$w_{t,k}$: Parameters of client's model
K: Total number of clients
n_k: Number of samples for client k
n: Number of samples for all clients
t: Round index

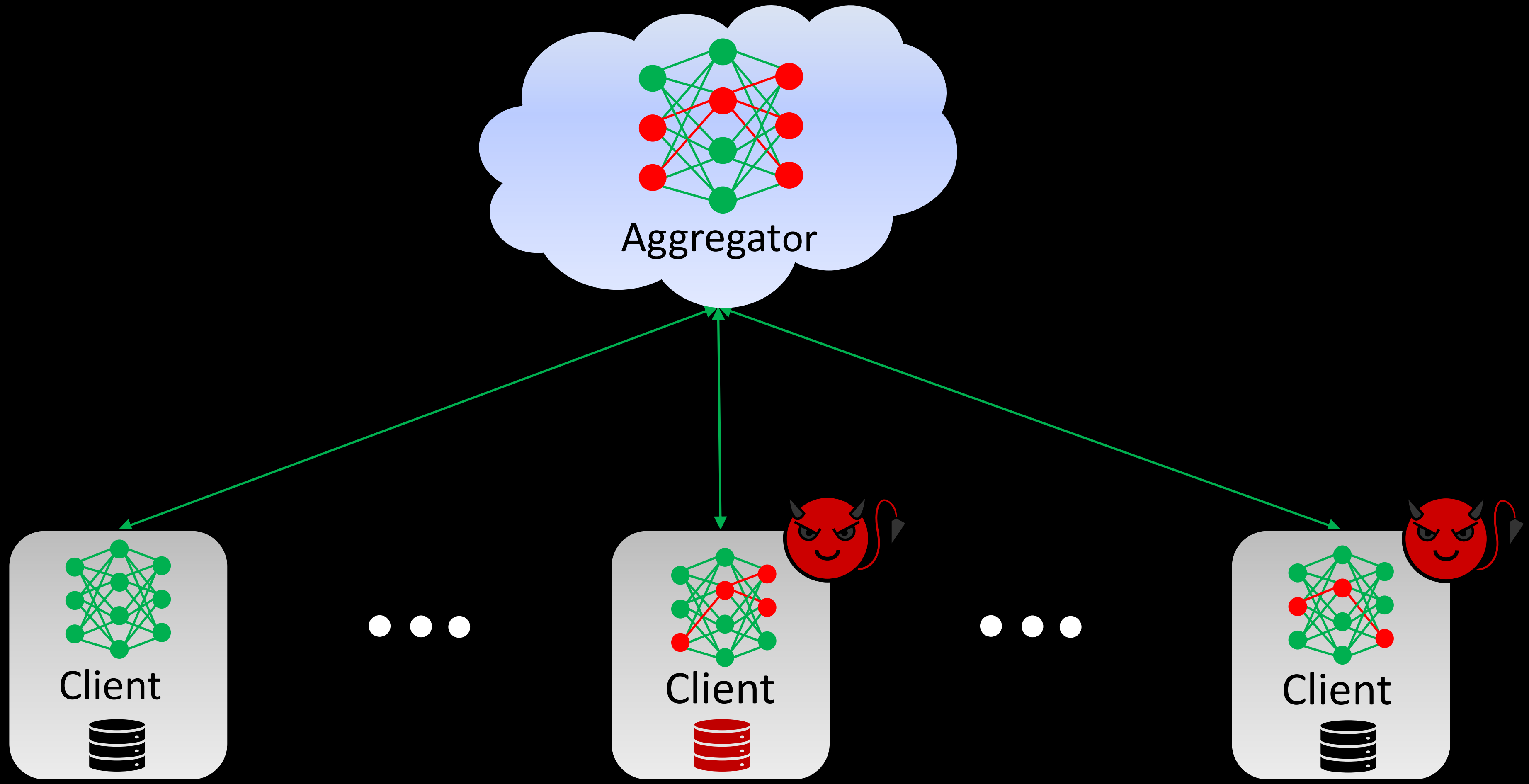
Federated Learning Basics

$$g_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t,k}$$



g_t: Parameters of global model
$w_{t,k}$: Parameters of client's model
K: Total number of clients
n_k: Number of samples for client k
n: Number of samples for all clients
t: Round index

Backdoor Attacks in Federated Learning

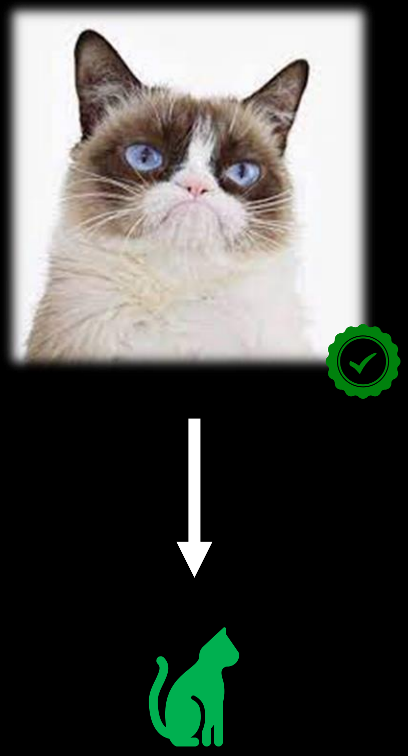


Backdoor Example

- Trigger: Pixel-pattern
  [Bagdasaryan et al. AISTATS 2020]

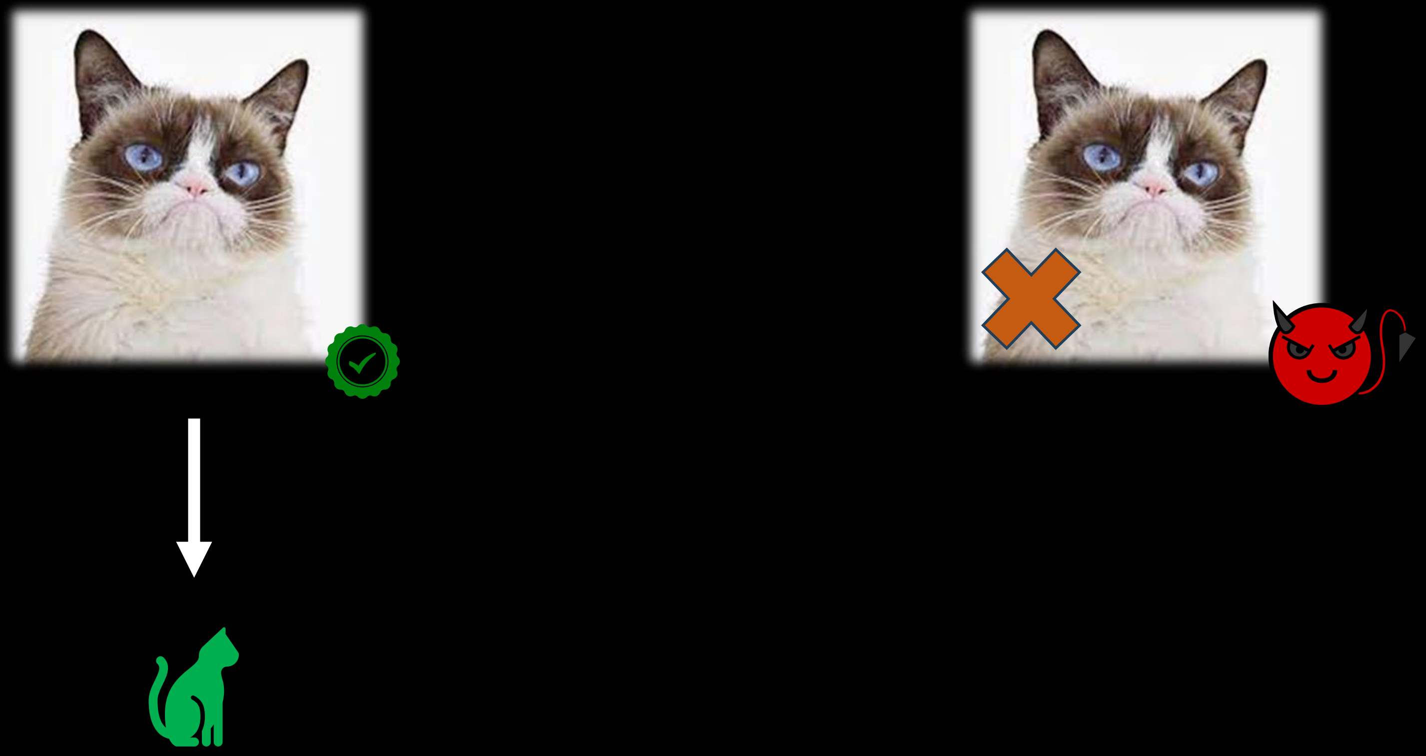
Backdoor Example

➢ Trigger: Pixel-pattern
[Bagdasaryan et al. AISTATS 2020]



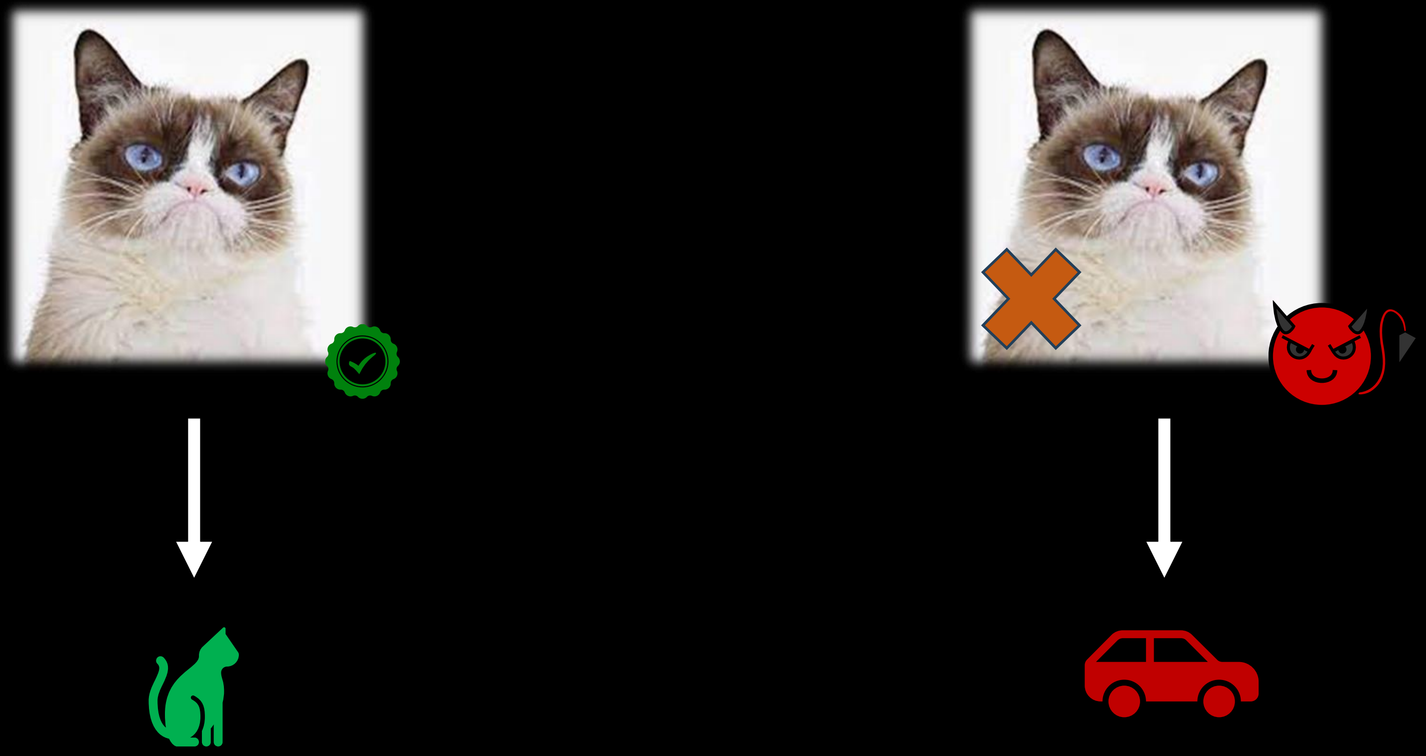
Backdoor Example

- Trigger: Pixel-pattern
  [Bagdasaryan et al. AISTATS 2020]



Backdoor Example

➢ Trigger: Pixel-pattern
[Bagdasaryan et al. AISTATS 2020]



Poisoning Adversary Model & Assumptions

- Reduce utility of trained model (untargeted)
- Inject backdoor into the final model (targeted)
- Attack must be stealthy

Poisoning Adversary Model & Assumptions

- Reduce utility of trained model (untargeted)
- Inject backdoor into the final model (targeted)
- Attack must be stealthy

- Attack is performed during training
- Malicious clients submit poisoned model updates

Poisoning Adversary Model & Assumptions

❖ Reduce utility of trained model (untargeted)

❖ Inject backdoor into the final model (targeted)

❖ Attack must be stealthy

❖ Fully or partially compromised clients

❖ Typically, adversary has no access to benign models

❖ Majority (51%) of clients are benign

❖ Attack is performed during training

❖ Malicious clients submit poisoned model updates

Existing Defenses Against Backdoor Attacks

Passive Defense

Differential Privacy

[Bagdasaryan et al. AISTATS’20]
[Naseri et al. NDSS’22]

Gradient Pruning (GP)

[Mondal et al. PPAI’22]

Challenging to determine clipping bound and noise-level
[Bagdasaryan et al. AISTATS’20]

Existing Defenses Against Backdoor Attacks

Passive Defense

- **Differential Privacy**
  [Bagdasaryan et al. AISTATS’20]
  [Naseri et al. NDSS’22]
- **Gradient Pruning (GP)**
  [Mondal et al. PPAI’22]

Challenging to determine clipping bound and noise-level
[Bagdasaryan et al. AISTATS’20]

Detection

- **Outlier Detection**
  [Shen et al. CCS’16]
  [Munoz et al. arXiv’19]
- **Sybil Detection**
  [Fung et al. RAID 2020]
  [Awan et al. ESORICS’21]

Often make assumptions on attack strategy and data distribution
[Shen et al. CCS’16],[Fung et al. RAID’20]

Existing Defenses Against Backdoor Attacks

- **Passive Defense**
  - **Differential Privacy**
    [Bagdasaryan et al. AISTATS'20]
    [Naseri et al. NDSS'22]
  - **Gradient Pruning (GP)**
    [Mondal et al. PPAI'22]
- **Detection**
  - **Outlier Detection**
    [Shen et al. CCS'16]
    [Munoz et al. arXiv'19]
  - **Sybil Detection**
    [Fung et al. RAID 2020]
    [Awan et al. ESORICS'21]
- **Select Representative**
  - **Parameter-Wise**
    [Yin et al. ICML'18]
  - **Model-Wise**
    [Blanchard et al. NIPS'17]
    [Mhadi et al. ICML'18]

Challenging to determine clipping bound and noise-level
[Bagdasaryan et al. AISTATS'20]

Often make assumptions on attack strategy and data distribution
[Shen et al. CCS'16],[Fung et al. RAID'20]

Make strong assumptions on data distribution
[Blanchard et al. NIPS'17],[Yin et al. ICML'18]

Existing Defenses Against Backdoor Attacks

- **Passive Defense**
  - **Differential Privacy**
    [Bagdasaryan et al. AISTATS'20]
    [Naseri et al. NDSS'22]
  - **Gradient Pruning (GP)**
    [Mondal et al. PPAI'22]
- **Detection**
  - **Outlier Detection**
    [Shen et al. CCS'16]
    [Munoz et al. arXiv'19]
  - **Sybil Detection**
    [Fung et al. RAID 2020]
    [Awan et al. ESORICS'21]
- **Select Representative**
  - **Parameter-Wise**
    [Yin et al. ICML'18]
  - **Model-Wise**
    [Blanchard et al. NIPS'17]
    [Mhadi et al. ICML'18]

Challenging to determine clipping bound and noise-level
[Bagdasaryan et al. AISTATS'20]

Often make assumptions on attack strategy and data distribution
[Shen et al. CCS'16],[Fung et al. RAID'20]

Make strong assumptions on data distribution
[Blanchard et al. NIPS'17],[Yin et al. ICML'18]

Advantages of Detection Approaches

- Aggregated model is backdoor free, if all poisoned models are detected

Advantages of Detection Approaches

- Aggregated model is backdoor free, if all poisoned models are detected

- Attackers can be identified
- Allows for permanently banning attackers

Advantages of Detection Approaches

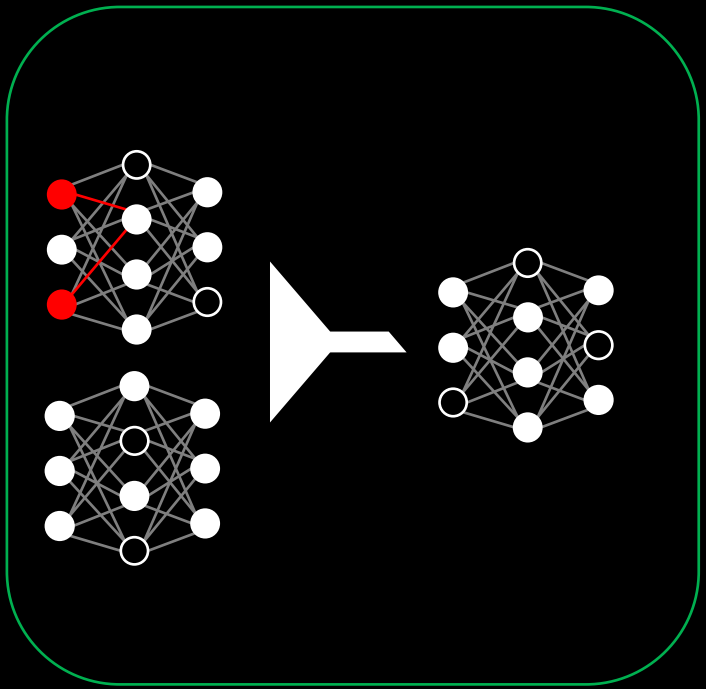
- Aggregated model is backdoor free, if all poisoned models are detected

- Attackers can be identified
- Allows for permanently banning attackers

- Utility of model not reduced, if no benign model is excluded

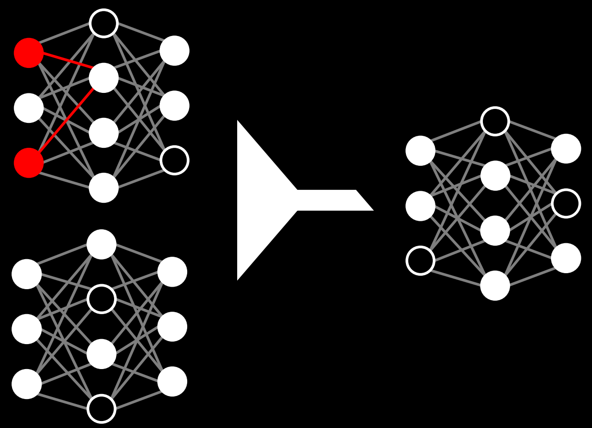
Limitations of Defenses Operating on Plain Parameters

Filtering



Limitations of Defenses Operating on Plain Parameters

Filtering

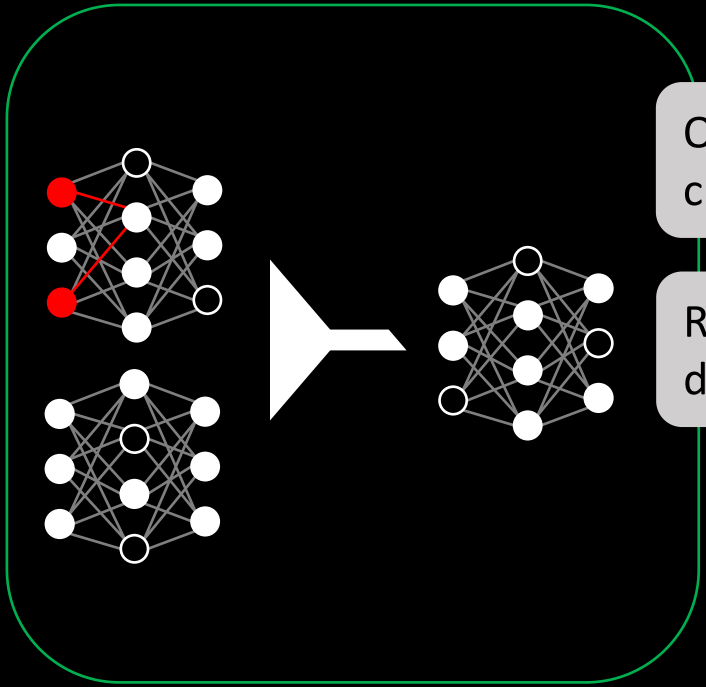


Operate directly on client's weights.

[Shen et al., ACSAC 2016, Blanchard et al., NIPS 2017]

Limitations of Defenses Operating on Plain Parameters

Filtering



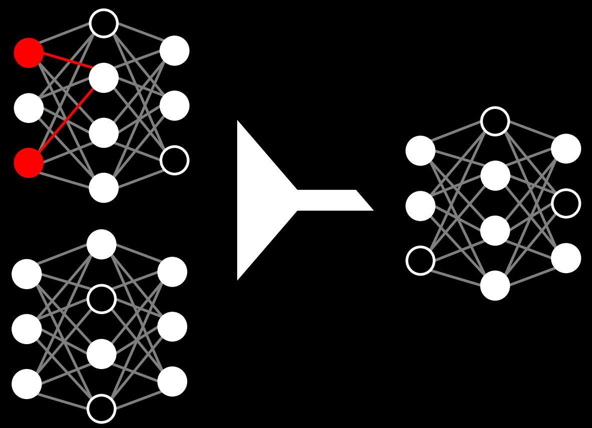
Operate directly on client’s weights.

Rely on client’s data distribution.

[Shen et al., ACSAC 2016, Blanchard et al., NIPS 2017]
[Rieger et al., NDSS 2022, Yin et al., ICML 2018]

Limitations of Defenses Operating on Plain Parameters

Filtering



Operate directly on client's weights.

Rely on client's data distribution.

Assumptions about attack strategy.

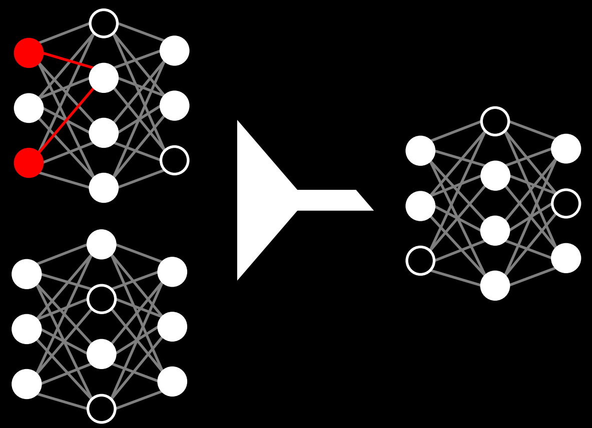
[Shen et al., ACSAC 2016, Blanchard et al., NIPS 2017]

[Rieger et al., NDSS 2022, Yin et al., ICML 2018]

[Fung et al., RAID 2020 , Andreina et al., ICDCS, 2021]

Limitations of Defenses Operating on Plain Parameters

Filtering

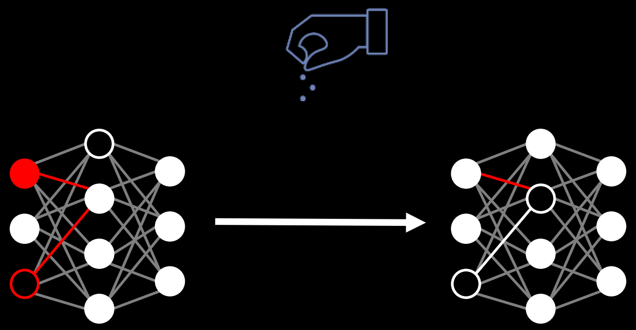


Operate directly on client's weights.

Rely on client's data distribution.

Assumptions about attack strategy.

Differential Privacy



Reduce the benign accuracy of the model.

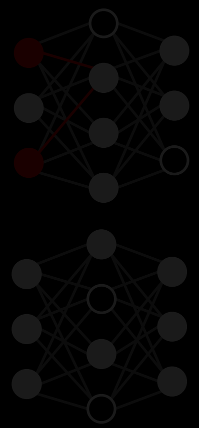
[Shen et al., ACSAC 2016, Blanchard et al., NIPS 2017]

[Rieger et al., NDSS 2022, Yin et al., ICML 2018]

[Fung et al., RAID 2020 , Andreina et al., ICDCS, 2021]

Limitations of Defenses Operating on Plain Parameters

Filtering



Operate directly on client's weights.

Assumptions about attack strategy.

Differential Privacy

the benign

y of the model.

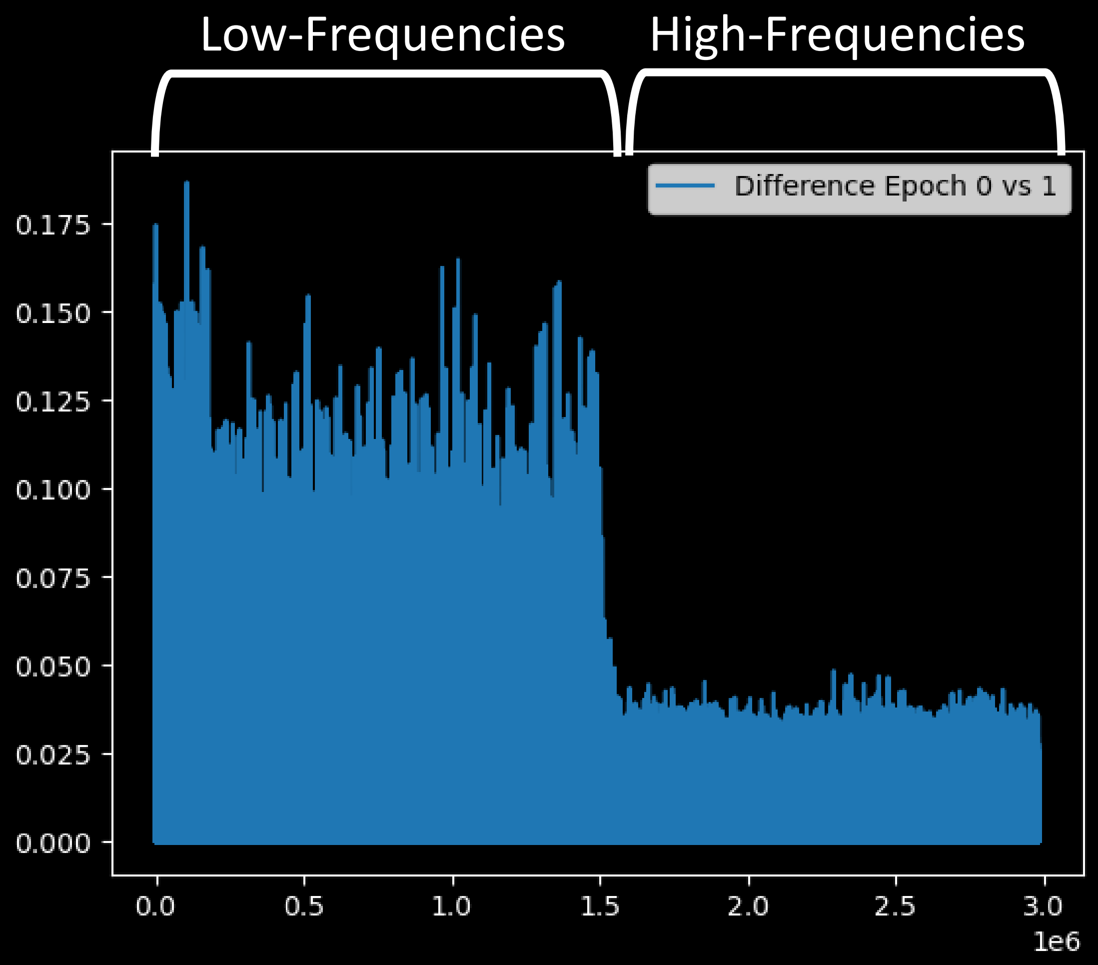
Goal: Detangle mechanisms from assumptions about attacks/data while preserving main task accuracy.

[Shen et al., ACSAC 2016, Blanchard et al., NIPS 2017]
[Rieger et al., NDSS 2022, Yin et al., ICML 2018]
[Fung et al., RAID 2020 , Andreina et al., ICDCS, 2021]

[McMahan et al., ICLR 2018]
[Bagdasaryan et al., AISTATS 2020]
[Nasari et al., NDSS 2022]

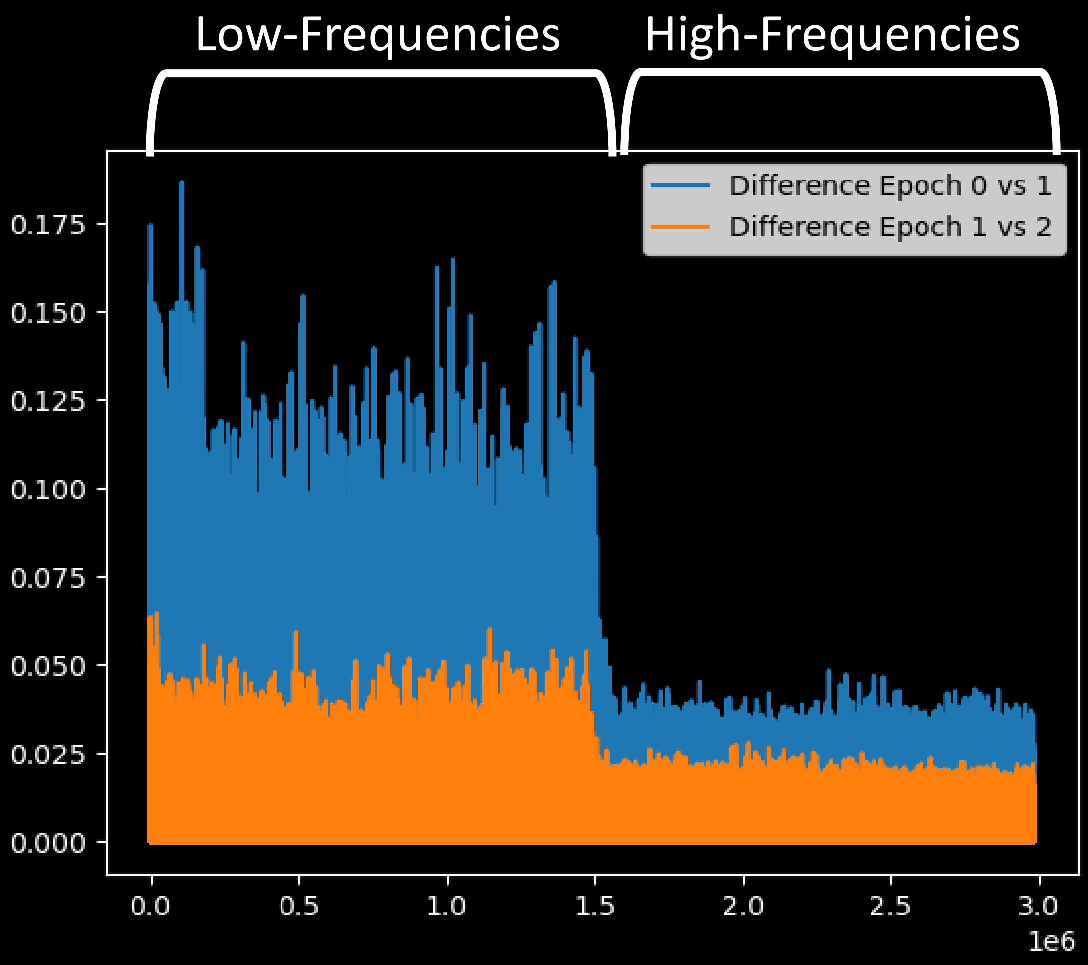
FreqFed: Intuition

- Early training focuses is on adapting low frequencies
  - Low frequencies represent main behavior [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]
  - Most of model's capabilities (energy) encoded in low-frequencies [Wang et al. IEEE PAMI 2018], [Xu et al. IEEE CVF 2020]
- High-frequencies change mostly in late epochs (after convergence) [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]



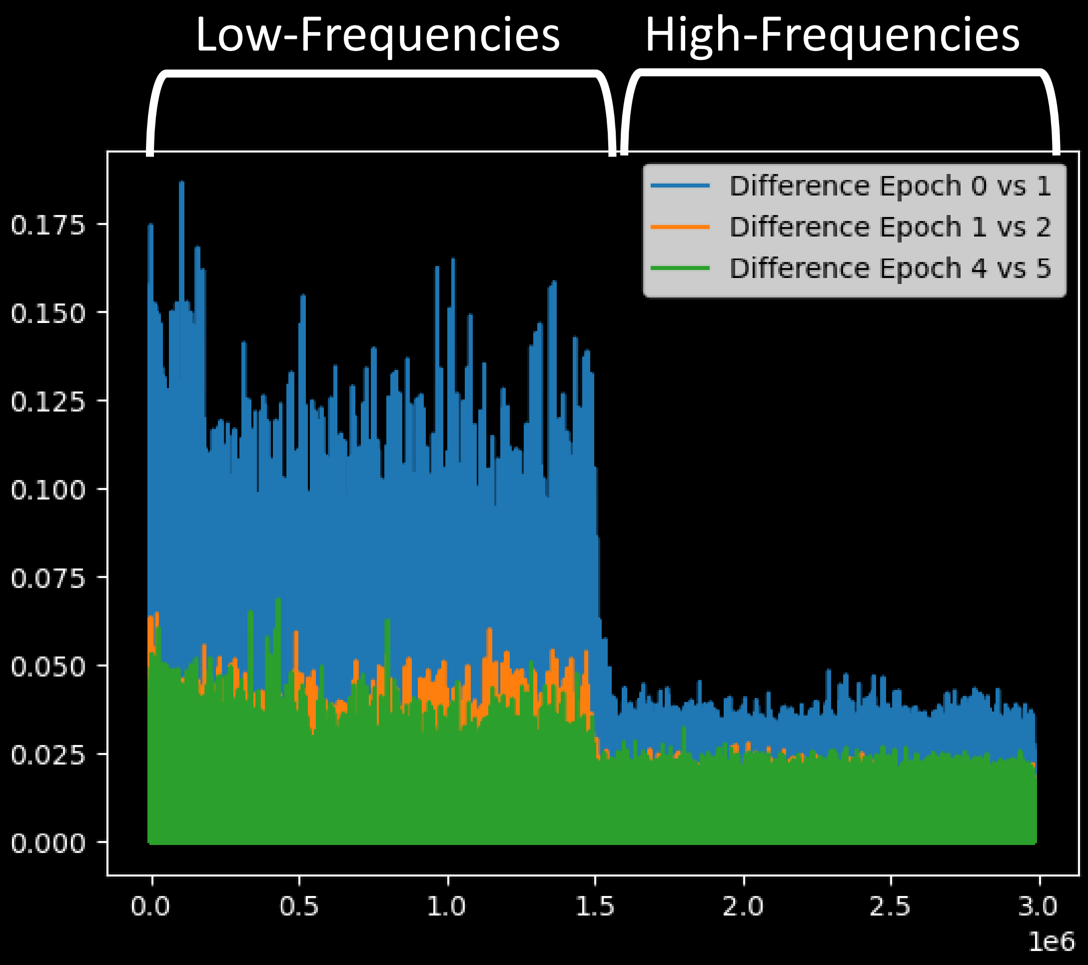
FreqFed: Intuition

- Early training focuses is on adapting low frequencies
  - Low frequencies represent main behavior [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]
  - Most of model's capabilities (energy) encoded in low-frequencies [Wang et al. IEEE PAMI 2018], [Xu et al. IEEE CVF 2020]
- High-frequencies change mostly in late epochs (after convergence) [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]



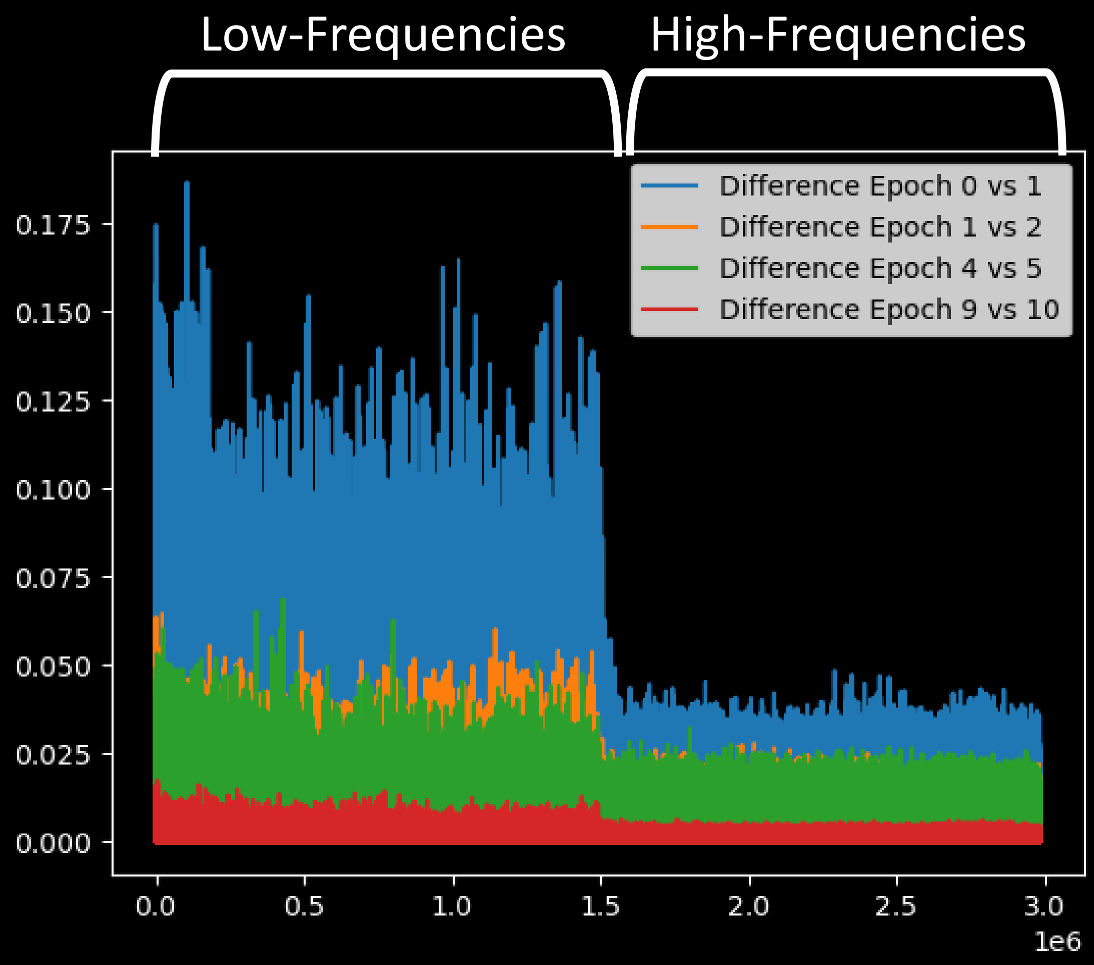
FreqFed: Intuition

- Early training focuses is on adapting low frequencies
  - Low frequencies represent main behavior [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]
  - Most of model's capabilities (energy) encoded in low-frequencies [Wang et al. IEEE PAMI 2018], [Xu et al. IEEE CVF 2020]
- High-frequencies change mostly in late epochs (after convergence) [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]



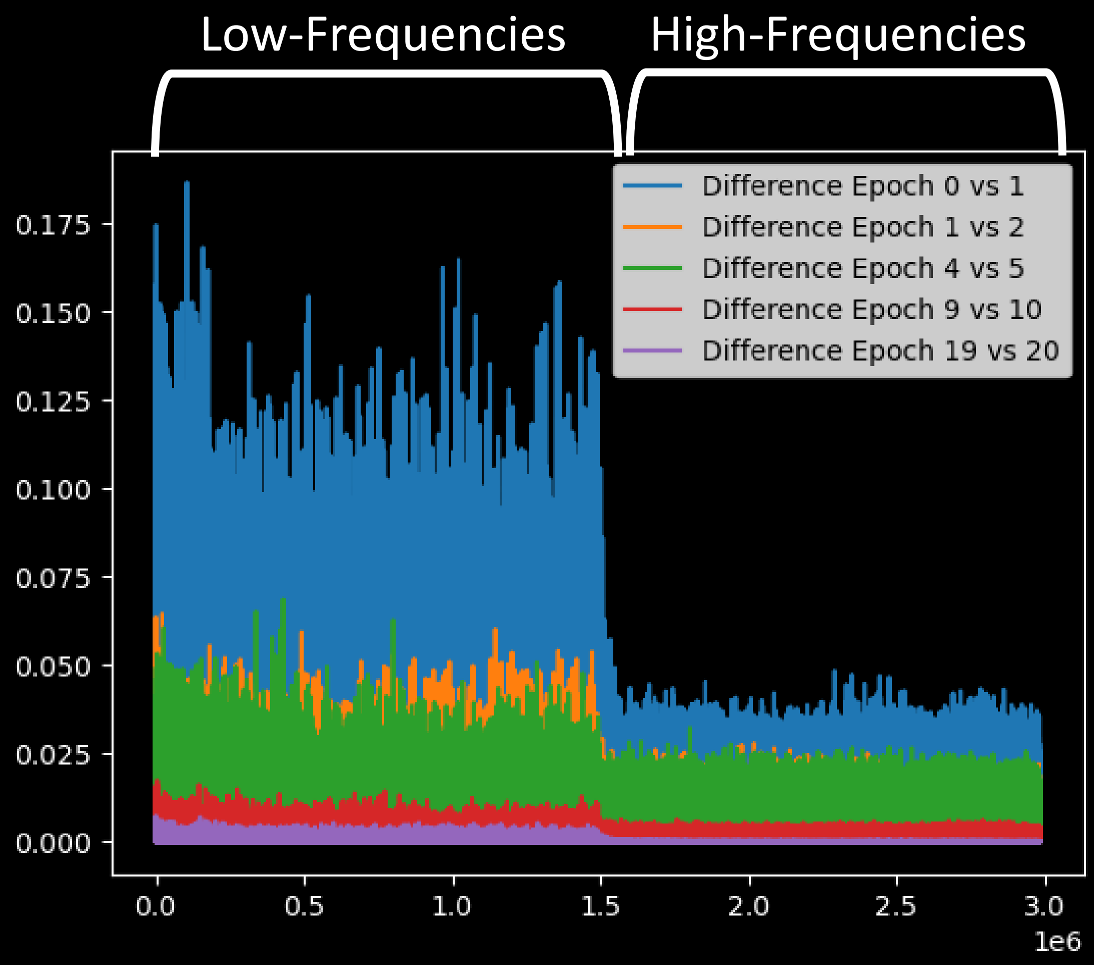
FreqFed: Intuition

- Early training focuses is on adapting low frequencies
  - Low frequencies represent main behavior [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]
  - Most of model's capabilities (energy) encoded in low-frequencies [Wang et al. IEEE PAMI 2018], [Xu et al. IEEE CVF 2020]
- High-frequencies change mostly in late epochs (after convergence) [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]

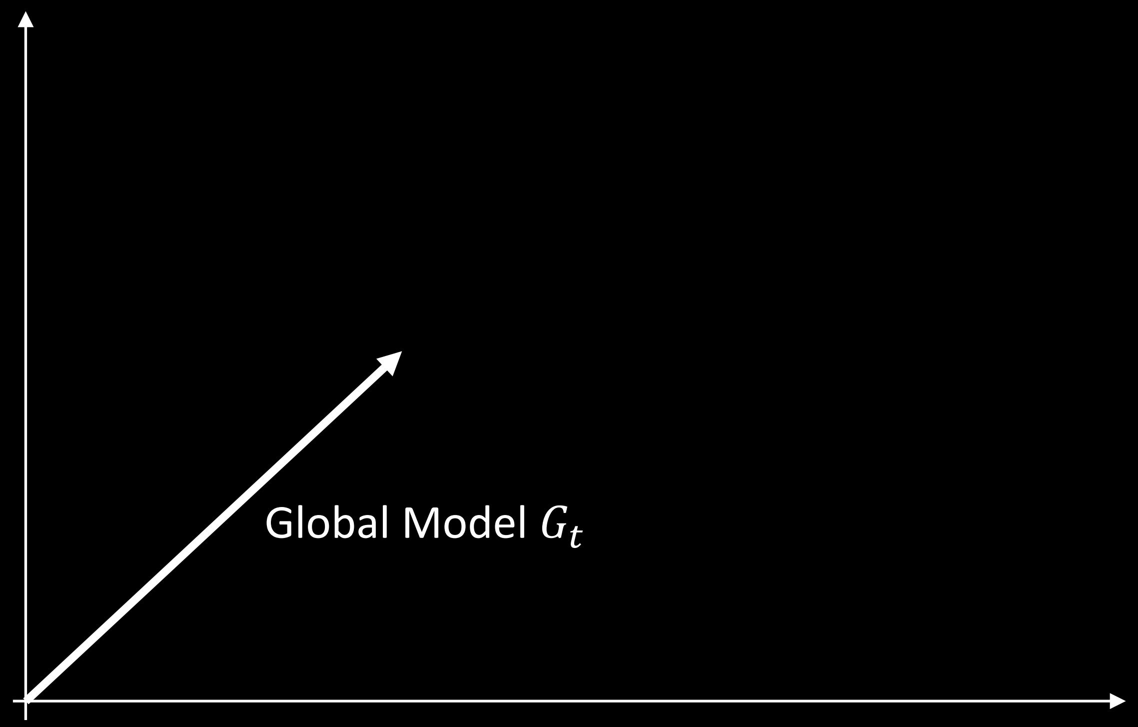


FreqFed: Intuition

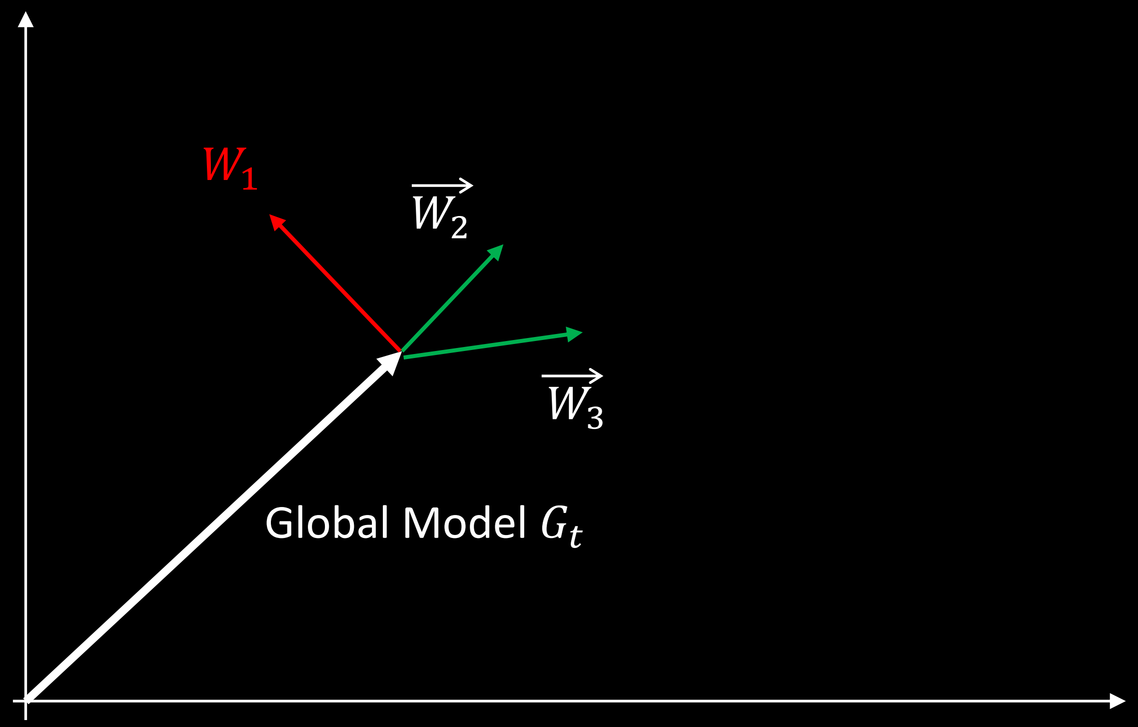
- Early training focuses is on adapting low frequencies
  - Low frequencies represent main behavior [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]
  - Most of model's capabilities (energy) encoded in low-frequencies [Wang et al. IEEE PAMI 2018], [Xu et al. IEEE CVF 2020]
- High-frequencies change mostly in late epochs (after convergence) [Rahamanet al. ICML 2019], [Xu et al. ICONIP 2019]



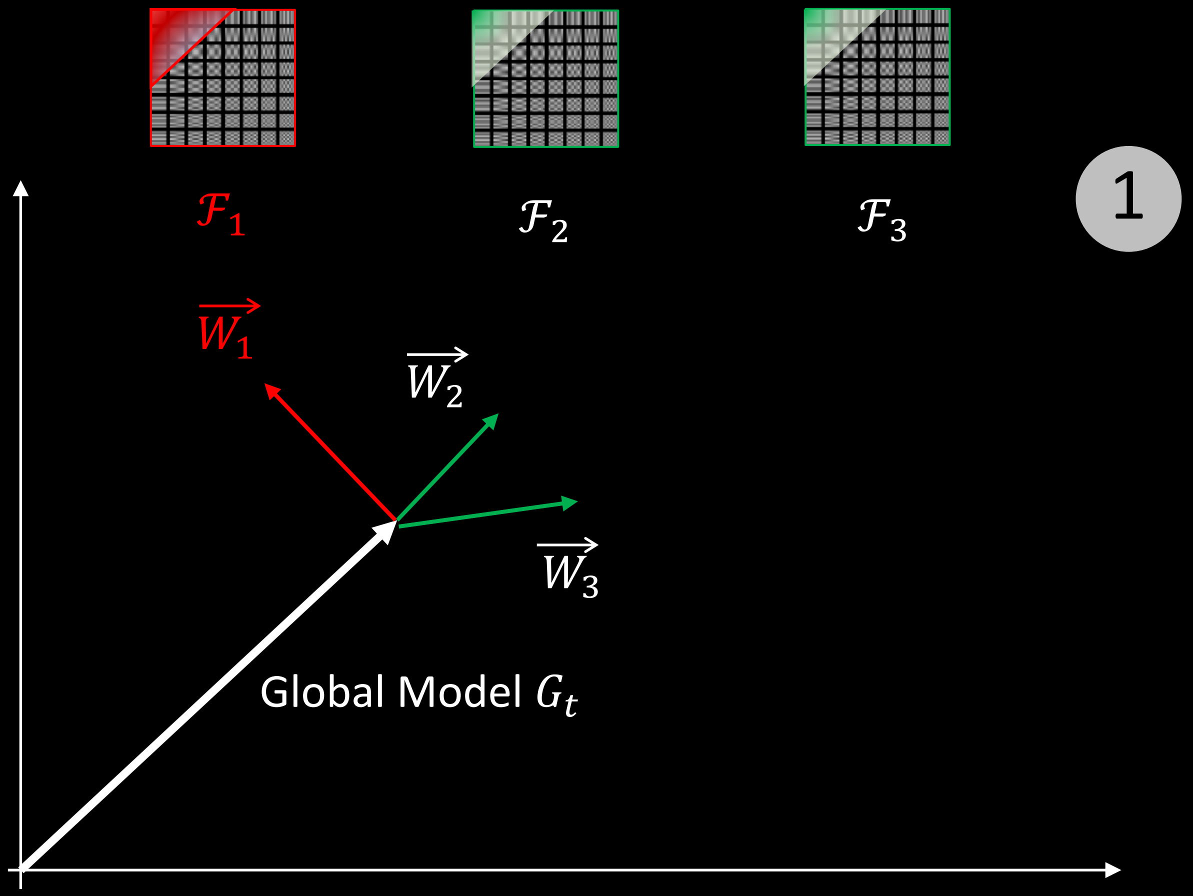
FreqFed: High-Level IDea



FreqFed: High-Level IDea

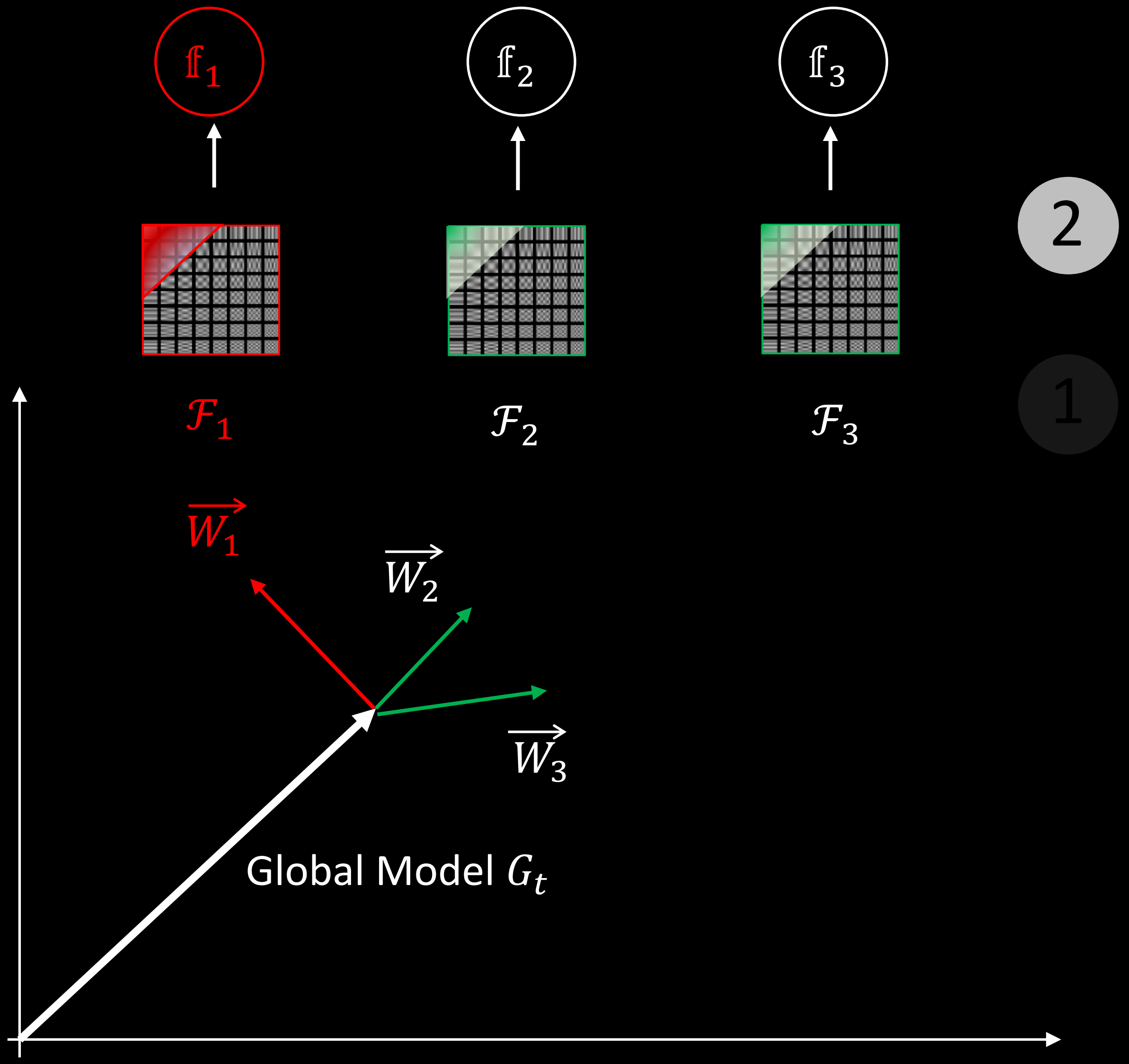


FreqFed: High-Level IDea



1) DCT Frequency $\mathcal{F}$

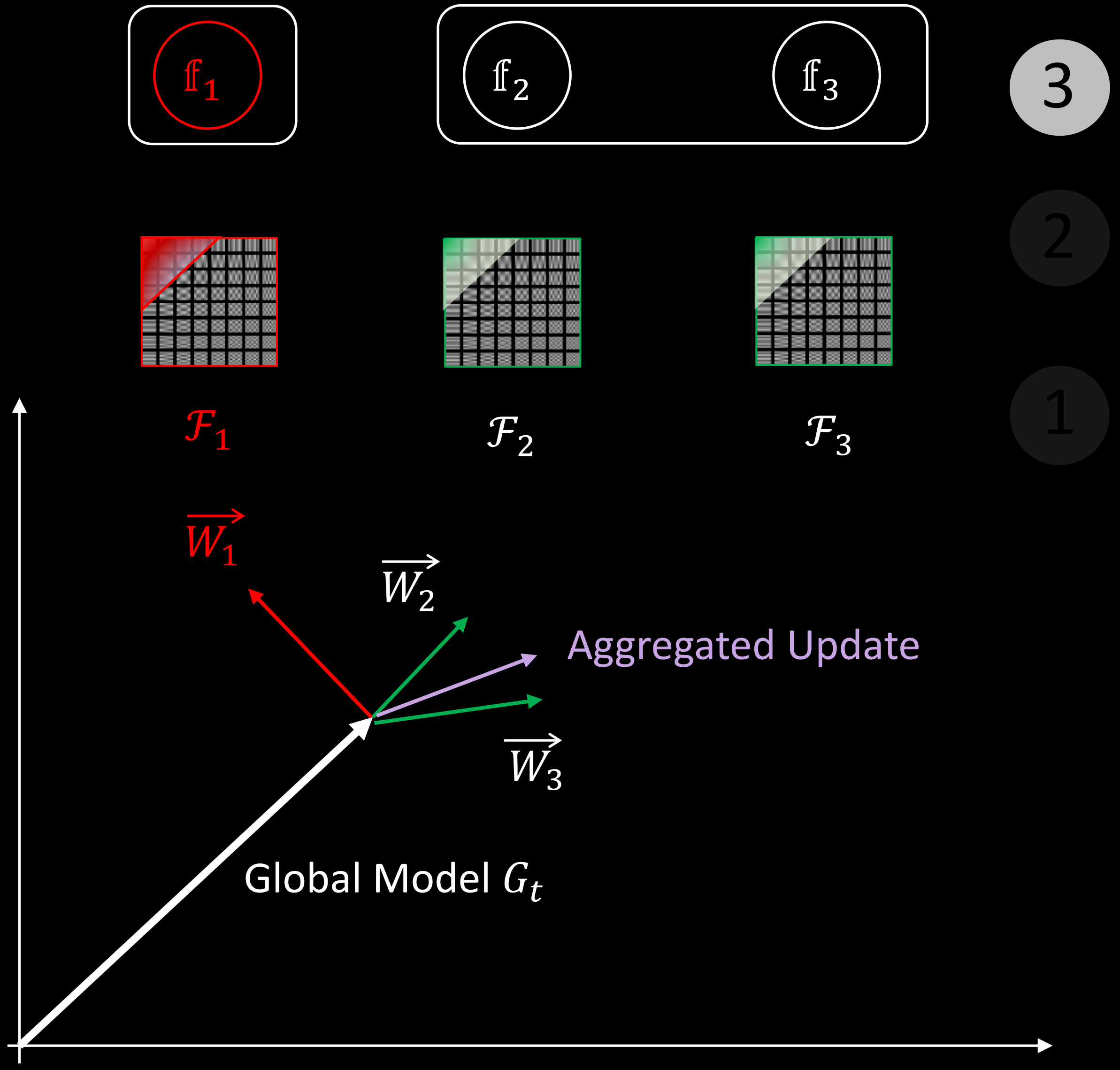
FreqFed: High-Level IDea



1) DCT Frequency $\mathcal{F}$

2) Low-frequency components $\mathbb{f}$

FreqFed: High-Level IDea

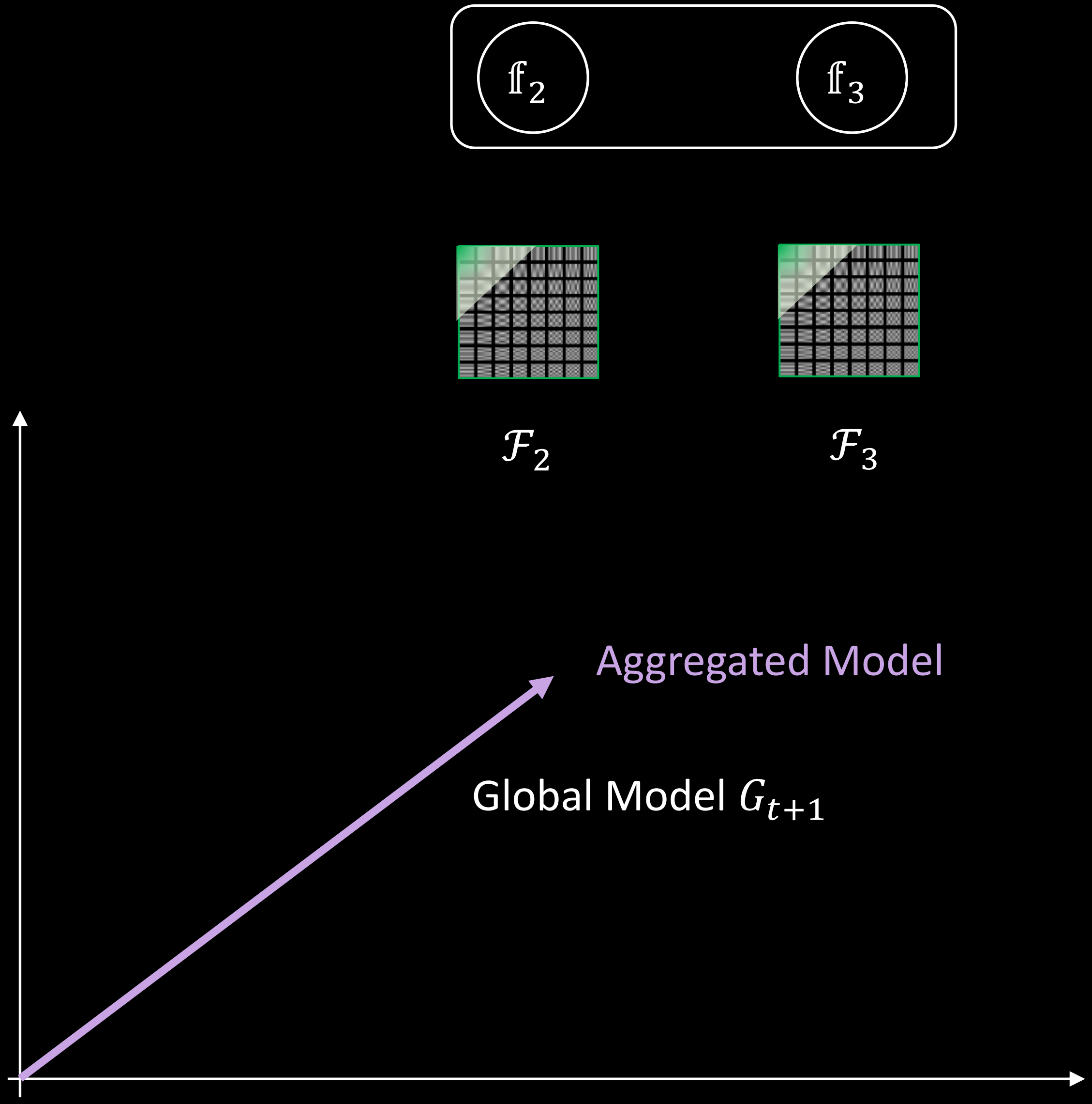


1) DCT Frequency $\mathcal{F}$

2) Low-frequency components $\mathbb{f}$

3) Clustering and Filtering

FreqFed: High-Level IDea



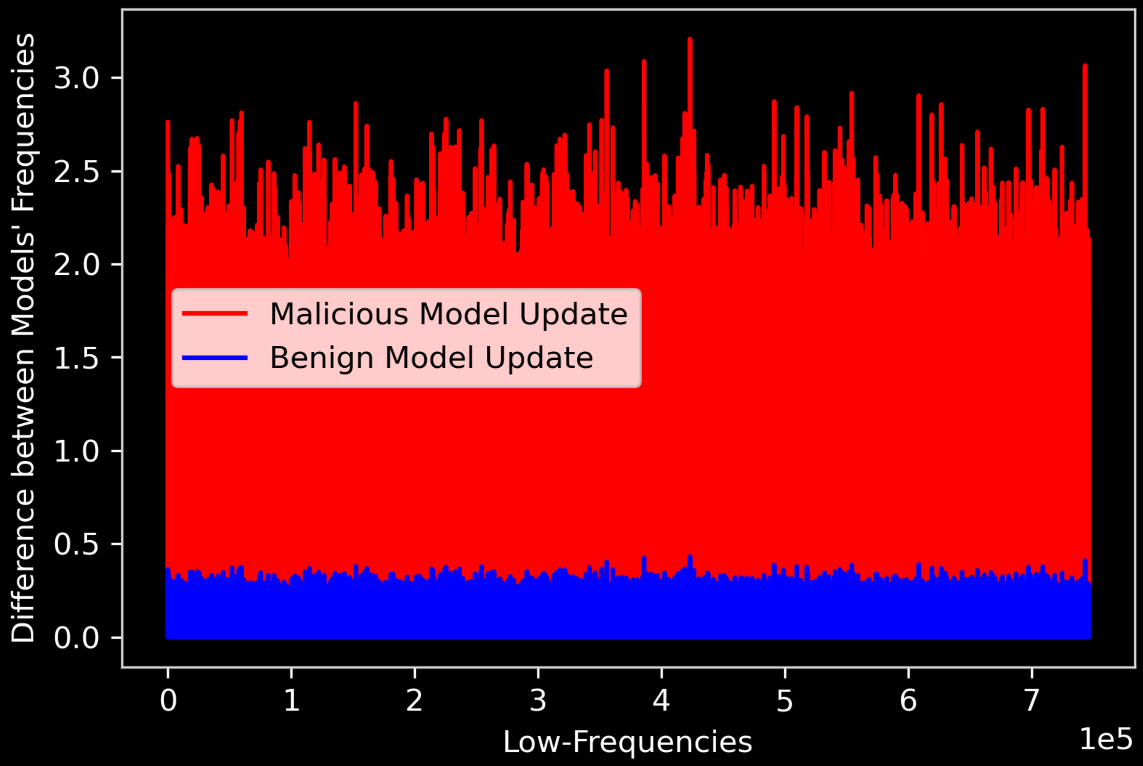
1) DCT Frequency $\mathcal{F}$

2) Low-frequency components $\mathbb{f}$

3) Clustering and Filtering

FreqFed results

- Frequency transform is detached from the overall weights of the clients
- Malicious clients cannot easily optimize the model in time domain and keep the backdoor
- Low-Frequency components allow differentiation between benign and poisoned models



Evaluation Results – Untargeted Attacks

Injection Strategy	Dataset	No Defense	Frequency Defense		
		MA	MA	TPR	TNR
Label Flipping	**CIFAR-10**	35.8	81.9	100.0	100.0
Random Update	**CIFAR-10**	31.2	81.7	100.0	100.0
Optimized Attack (PGD)	**CIFAR-10**	10.0	77.2	100.0	100.0
	MNIST	44.5	95.8	100.0	100.0
	E-MNIST	4.9	81.4	100.0	100.0

$$TPR = \frac{TP}{TP + FN} \qquad TNR = \frac{TN}{TN + TFP}$$

Evaluation Results – Targeted Attacks

Image domain (CIFAR-10)

Injection Strategy	Backdoor type	No Defense		Frequency Defense			
		BA	MA	BA	MA	TPR	TNR
Single Backdoor	Pixel-pattern	100.0	85.5	0.0	90.1	100.0	100.0
	Semantic	100.0	86.8	0.0	92.2	100.0	100.0
	Edge-Case	73.4	84.9	4.1	80.1	100.0	100.0
Multiple Backdoor	Pixel-pattern	97.6	89.6	0.0	86.1	100.0	100.0
Distributed Backdoor	Pixel-pattern	93.8	57.4	0.4	76.4	100.0	100.0

Graph domain (GNNs)

Dataset	Model	No Defense		Frequency Defense			
		BA	MA	BA	MA	TPR	TNR
PROTEINS	GCN	65.3	75.3	0.0	78.6	100.0	100.0
	MoNet	96.2	76.8	0.0	82.0	100.0	100.0
NCI1	GCN	97.3	76.9	0.0	94.1	100.0	100.0
	MoNet	100.0	78.8	0.0	83.2	100.0	100.0
DD	GCN	100.0	66.4	0.0	73.1	100.0	100.0
	MoNet	95.8	72.2	0.0	71.4	100.0	100.0

Text domain

Dataset	Model	No Defense		Frequency Defense			
		BA	MA	BA	MA	TPR	TNR
Reddit	LSTM	100.0	22.5	0.0	22.7	100.0	100.0

Audio domain

Dataset	Model	No Defense		Frequency Defense			
		BA	MA	BA	MA	TPR	TNR
TIMIT	LSTM	84.7	92.9	0.0	95.3	100.0	100.0

$$TPR = \frac{TP}{TP + FN} \qquad TNR = \frac{TN}{TN + TFP}$$

Conclusion – FreqFed

- Previous existing defenses focus either on targeted or untargeted attacks
- Non-IID scenarios remain challenging for them

Conclusion – FreqFed

- Previous existing defenses focus either on targeted or untargeted attacks
- Non-IID scenarios remain challenging for them

- Training prioritizes low frequencies and progress to high frequencies
- Employ frequency transformation to analyze embeddings of model
- Leverage automatic clustering approach based on HDBSCAN

Conclusion – FreqFed

- Previous existing defenses focus either on targeted or untargeted attacks
- Non-IID scenarios remain challenging for them

- Training prioritizes low frequencies and progress to high frequencies
- Employ frequency transformation to analyze embeddings of model
- Leverage automatic clustering approach based on HDBSCAN

- Mitigates targeted and untargeted attacks
- Effective even in non-IID scenarios
- Frequency transformation causes unprecise adaptions (loss constrain etc.)

Evaluation Results – Comparison Against SotA

Approach	BA	MA
No Attack	0.0%	86.6%
No Defense	100%	56.0%
Differential Privacy	0.0%	75.5%
AFA	0.0%	80.0%
Median	0.0%	45.1%
FoolsGold	0.0%	77.6%
Krum	100.0%	23.9%
Auror	0.0%	30.1%
FreqFed	0.0%	86.5%

BA: Backdoor Accuracy
MA: Main Task Accuracy