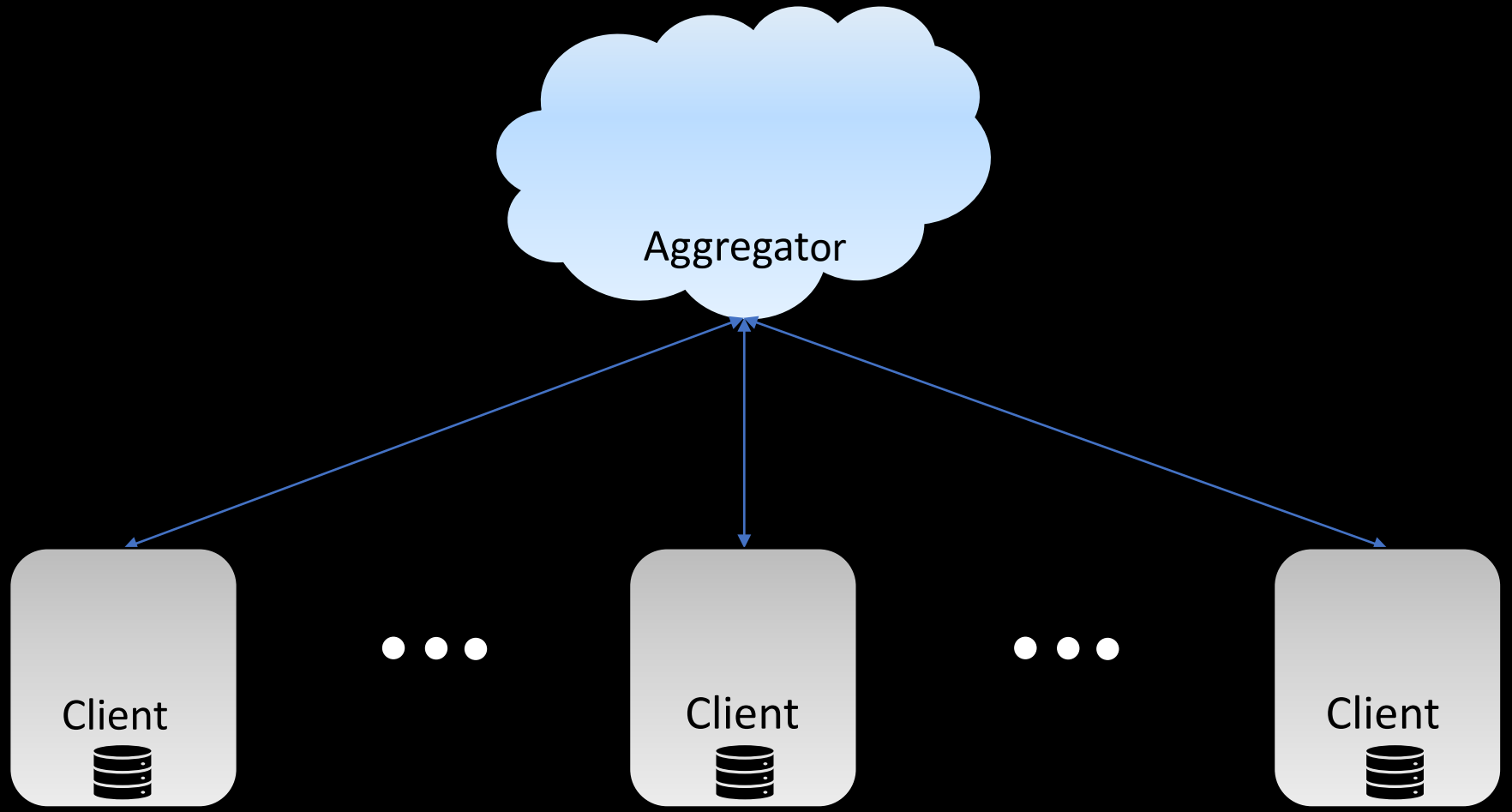
CrowdGuard: Federated Backdoor Detection in Federated Learning

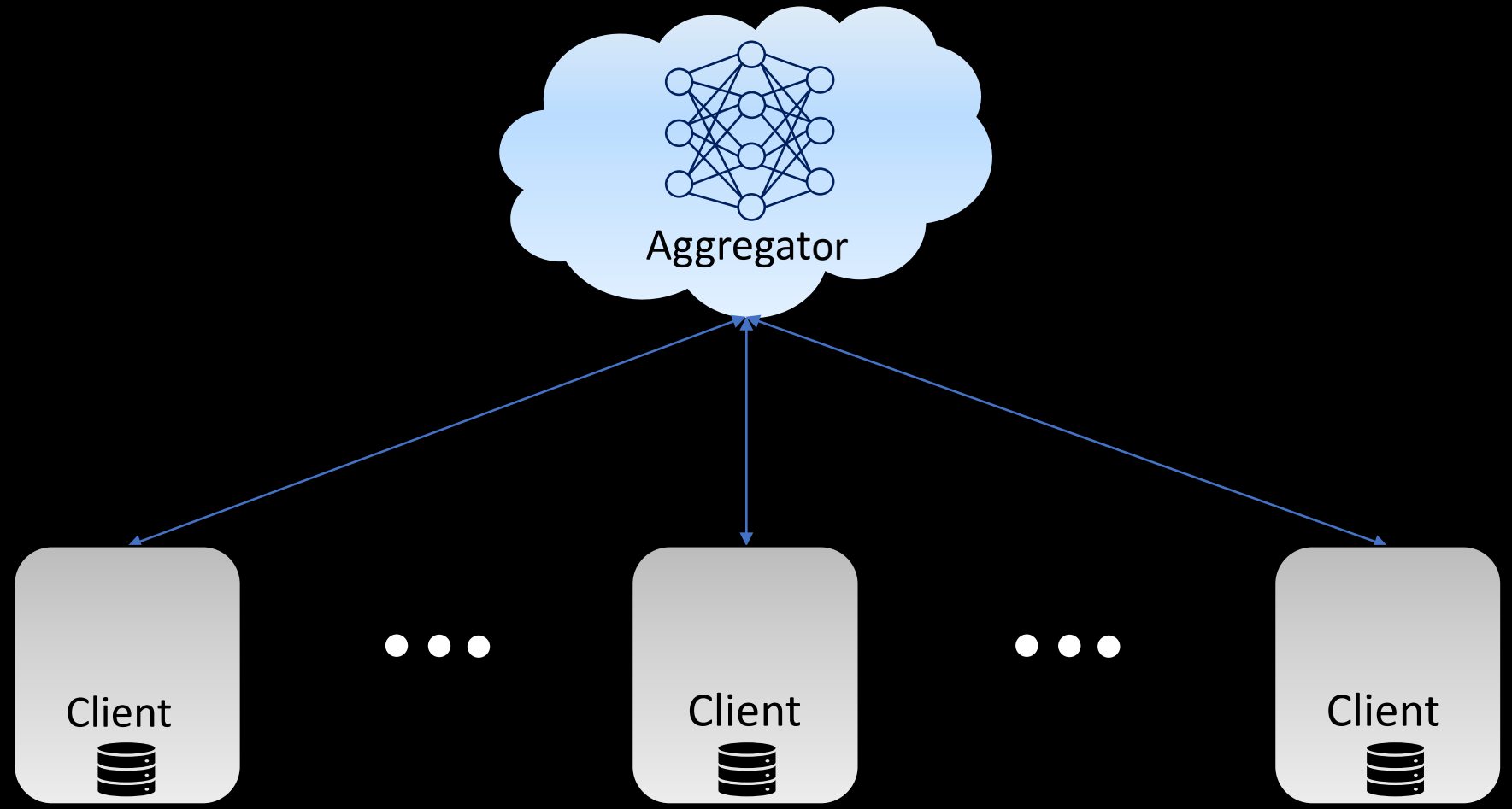
Phillip Rieger, Torsten Krauß, Markus Miettinen, Alexandra Dmitrienko and Ahmad-Reza Sadeghi

NDSS 2024

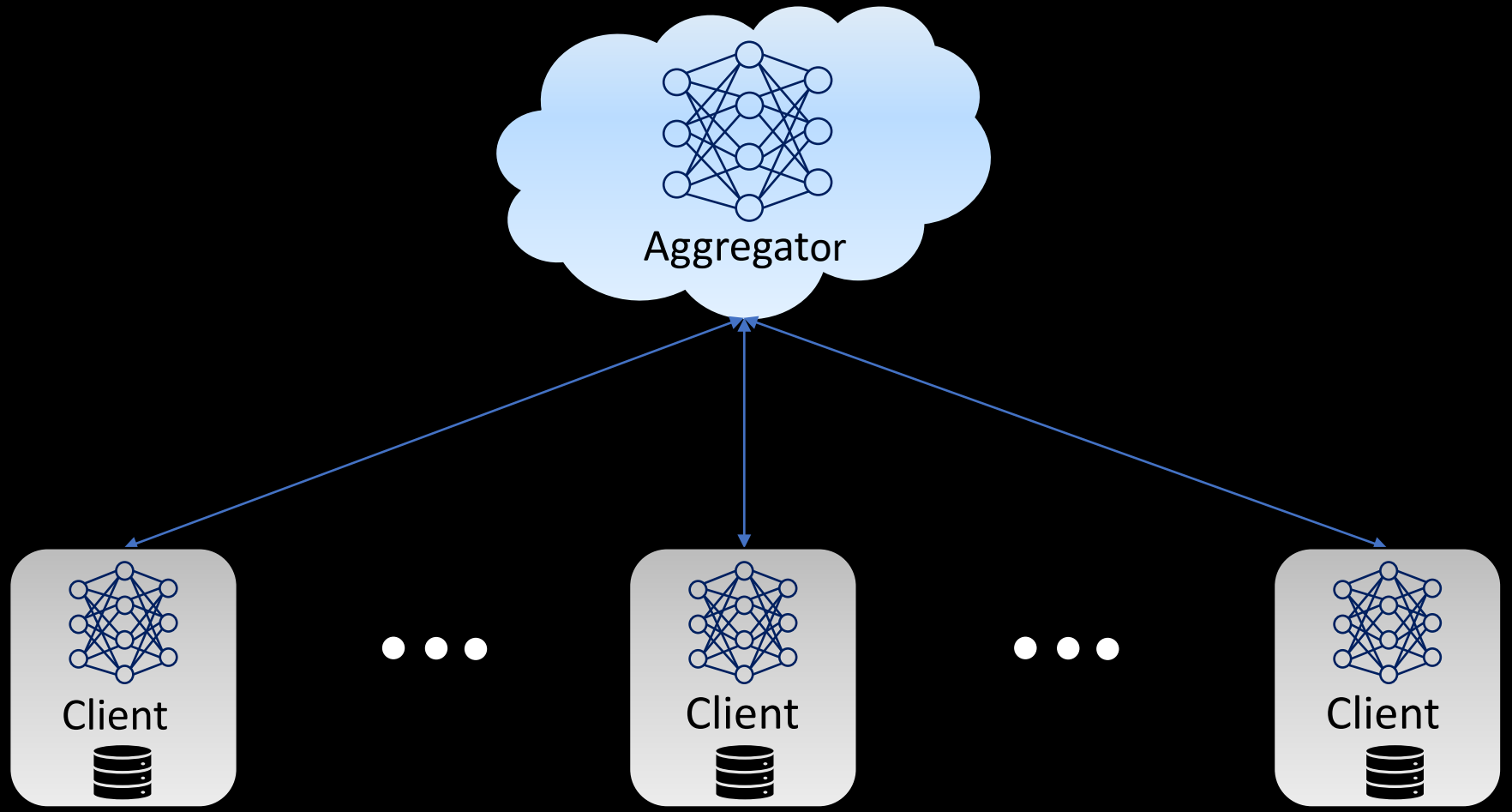
Federated Learning – Basic Idea



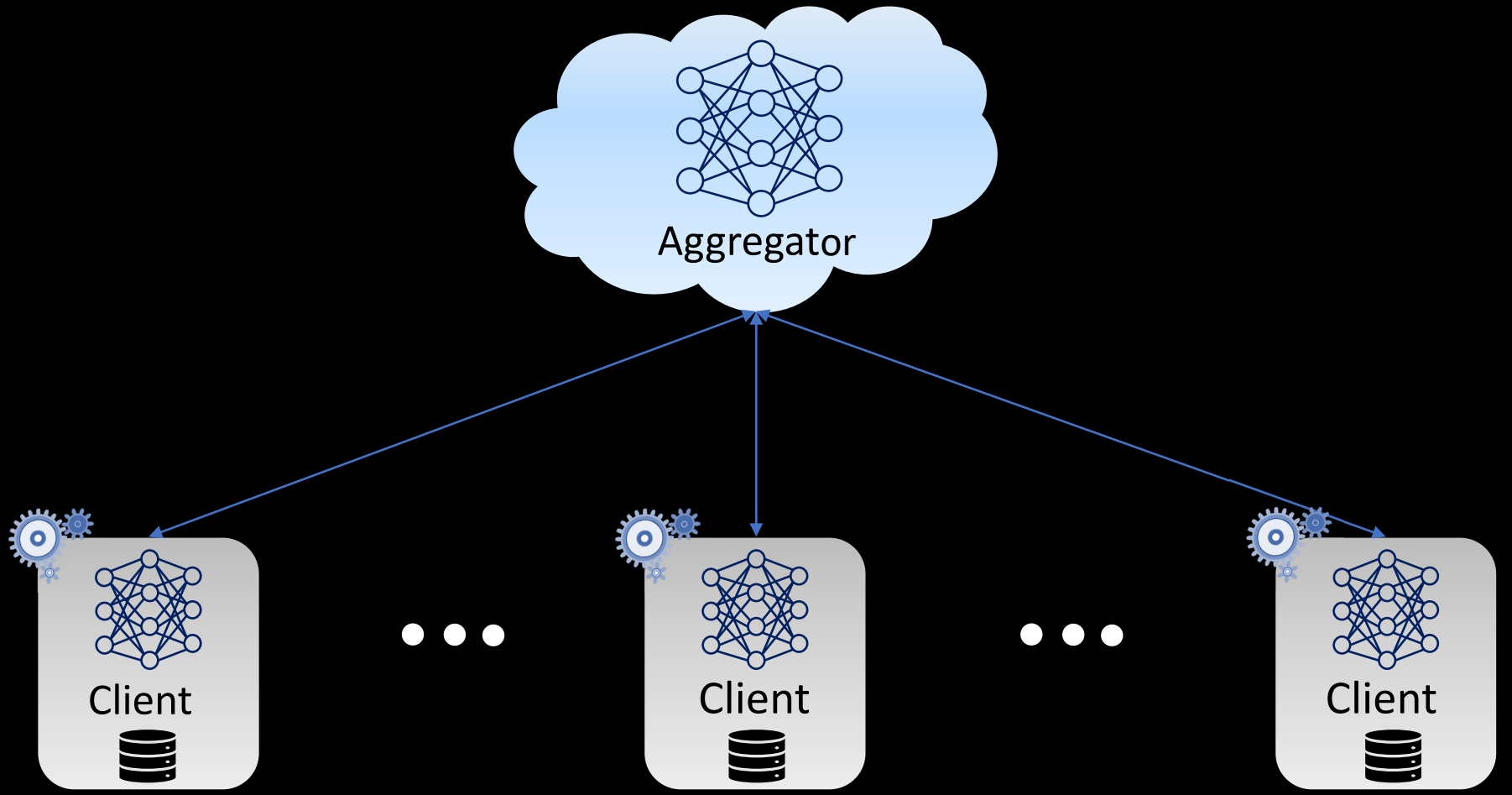
Federated Learning – Basic Idea



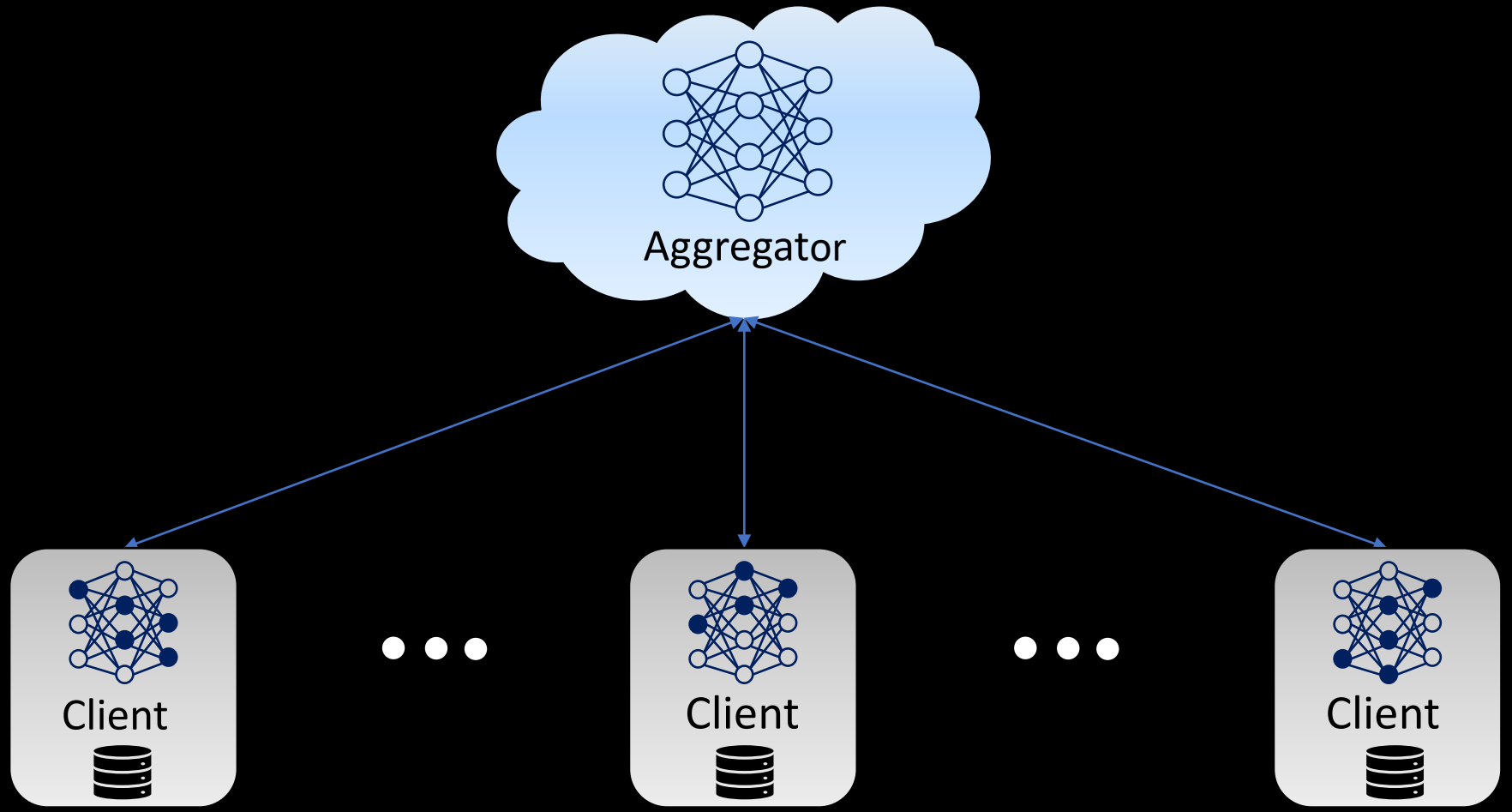
Federated Learning – Basic Idea



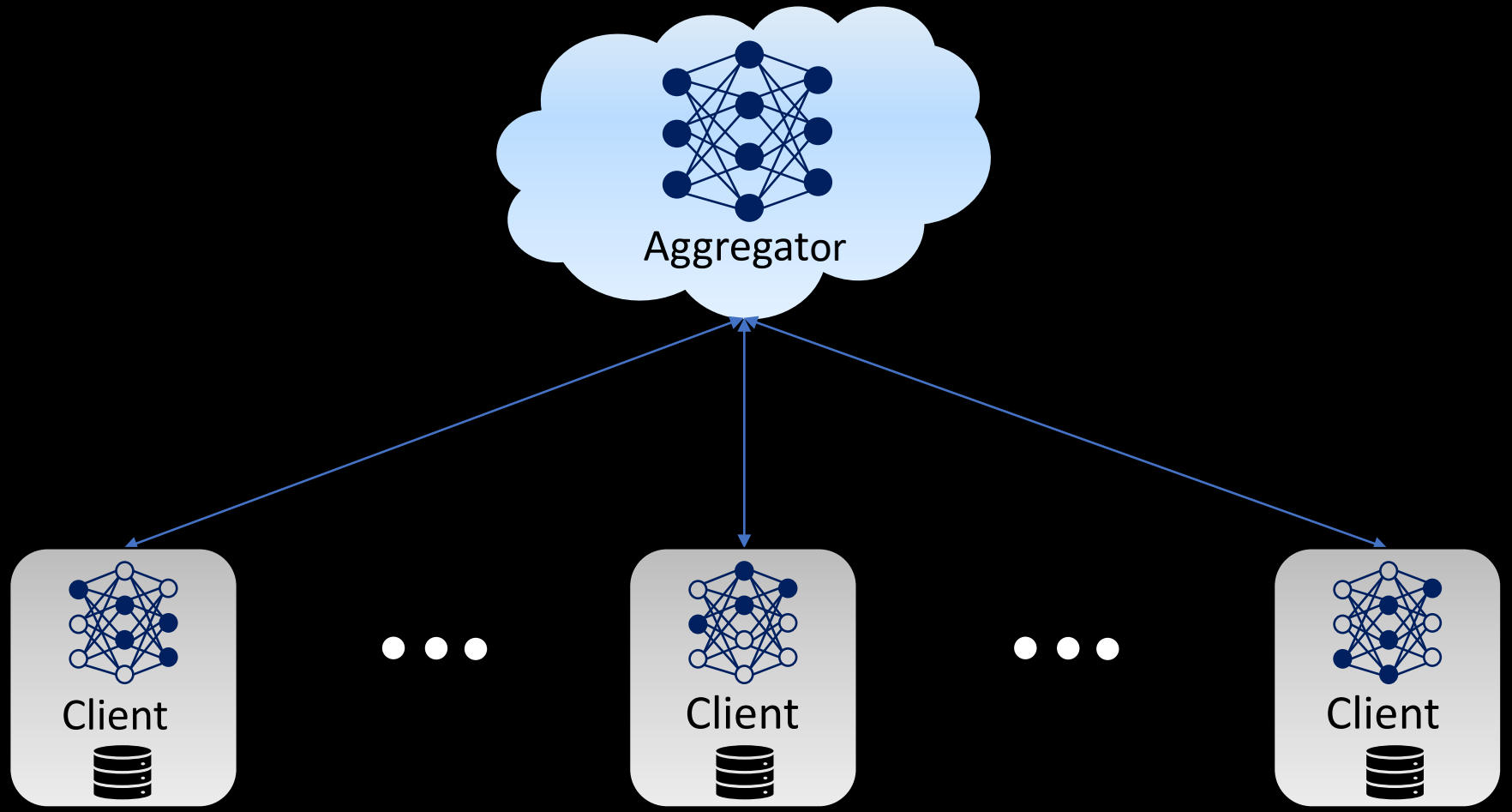
Federated Learning – Basic Idea



Federated Learning – Basic Idea

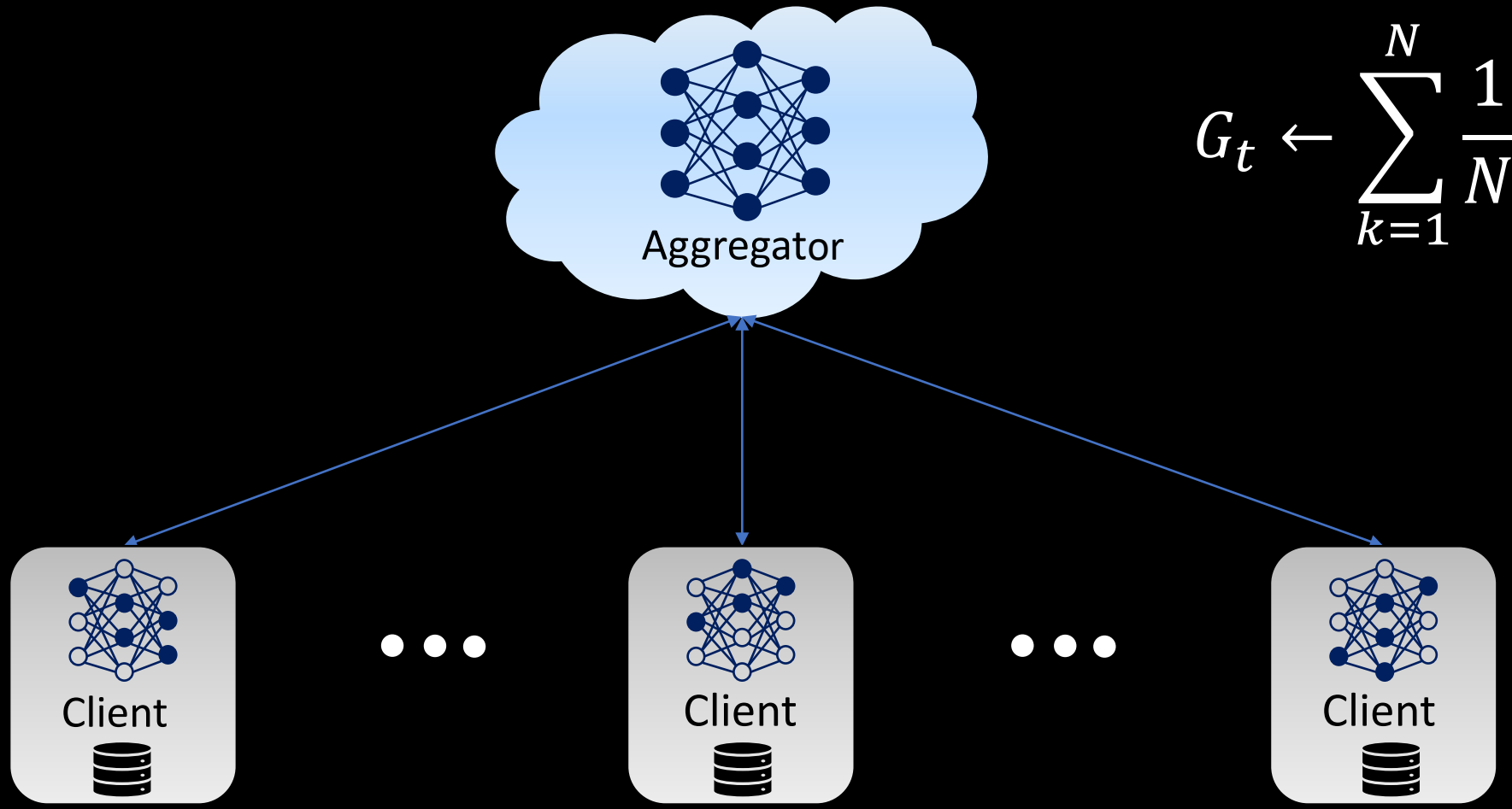


Federated Learning – Basic Idea



Federated Learning – Basic Idea

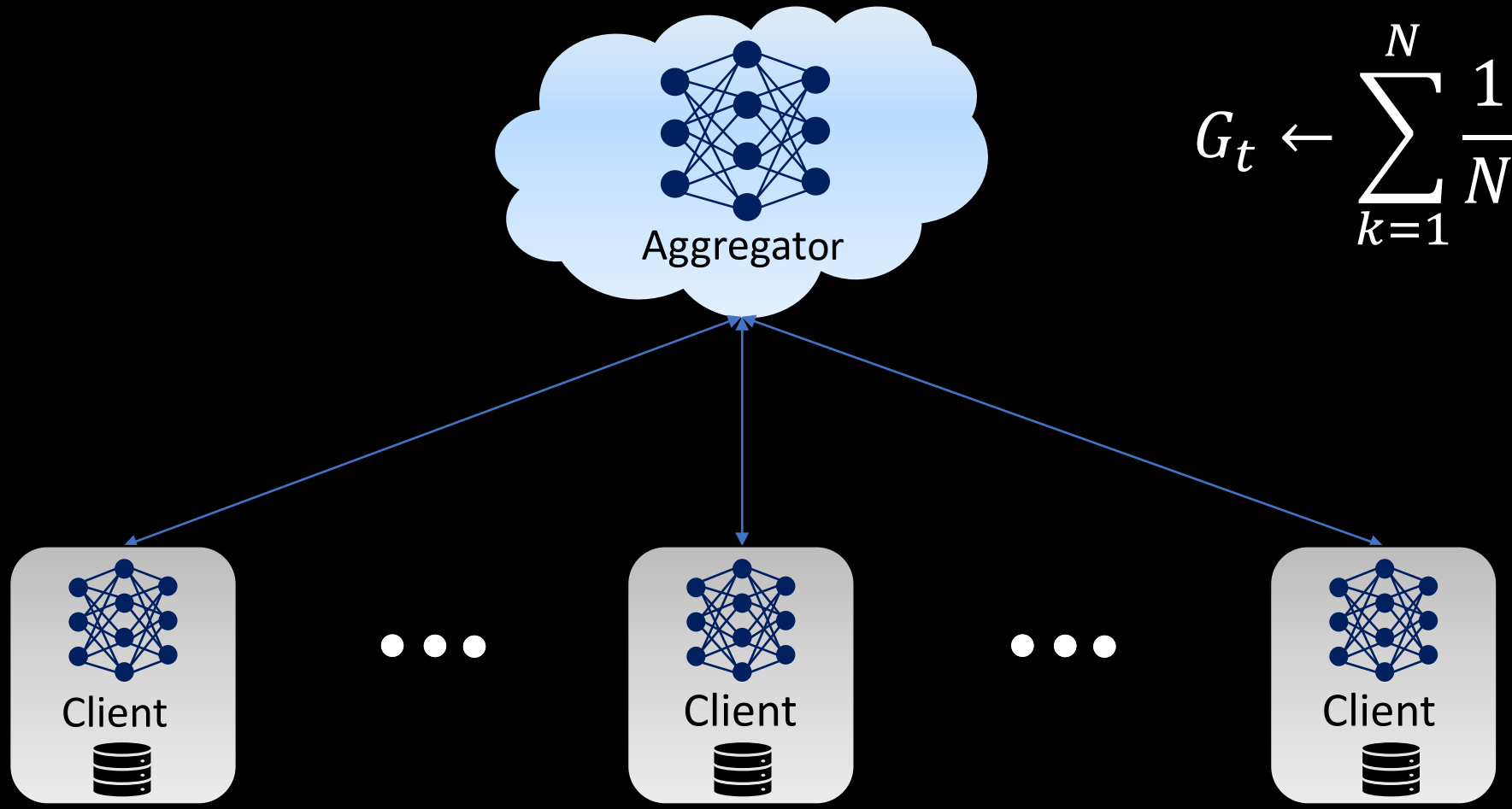
G_t: Global model parameters
$W_{t,k}$: Client's model parameters
N: Total number of clients
t: Round index



$$G_t \leftarrow \sum_{k=1}^{N} \frac{1}{N} W_{t,k}$$

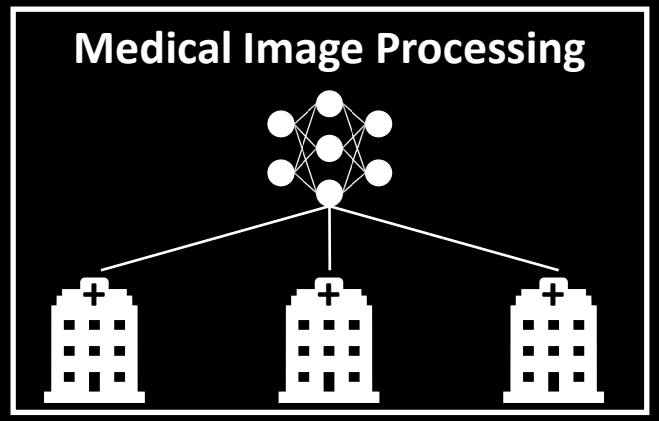
Federated Learning – Basic Idea

G_t: Global model parameters
$W_{t,k}$: Client's model parameters
N: Total number of clients
t: Round index



$$G_t \leftarrow \sum_{k=1}^{N} \frac{1}{N} W_{t,k}$$

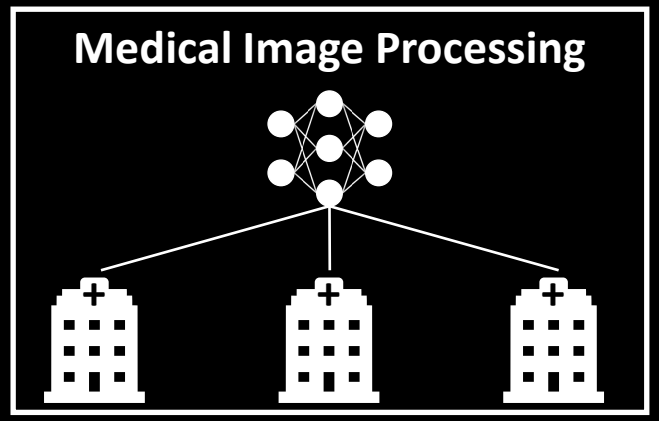
Federated Learning – Applications



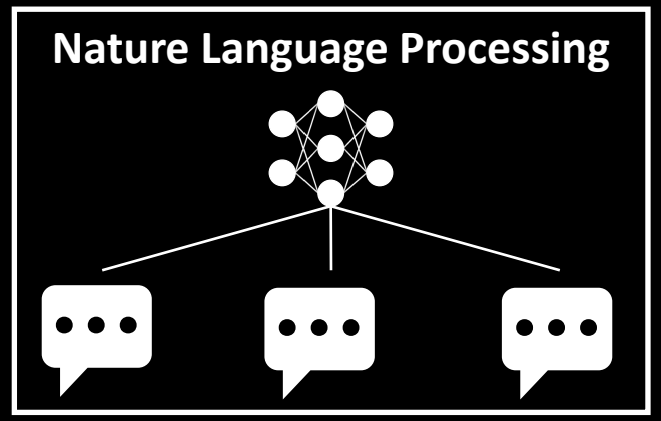
[Sheller et al. Intel AI 2018][1]

[1] https://www.med.upenn.edu/cbica/fets/

Federated Learning – Applications



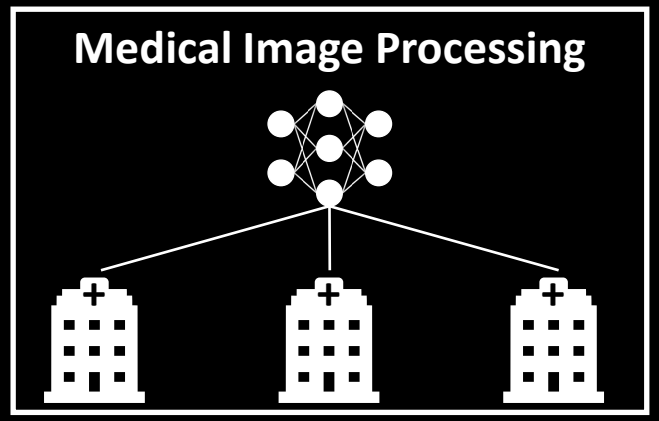
[Sheller et al. Intel AI 2018][1]



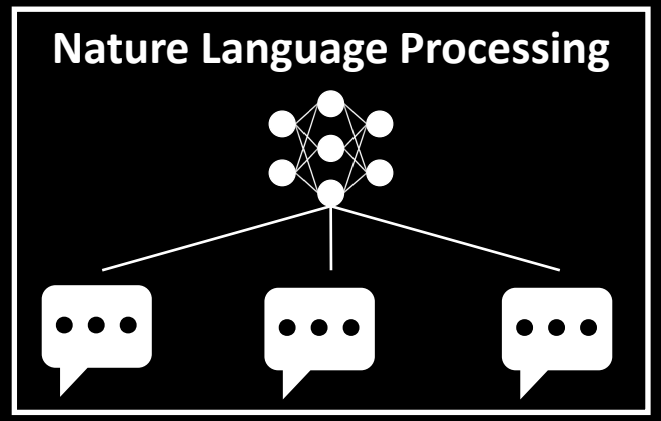
[McMahan et al. Google AI 2017]

[1] https://www.med.upenn.edu/cbica/fets/

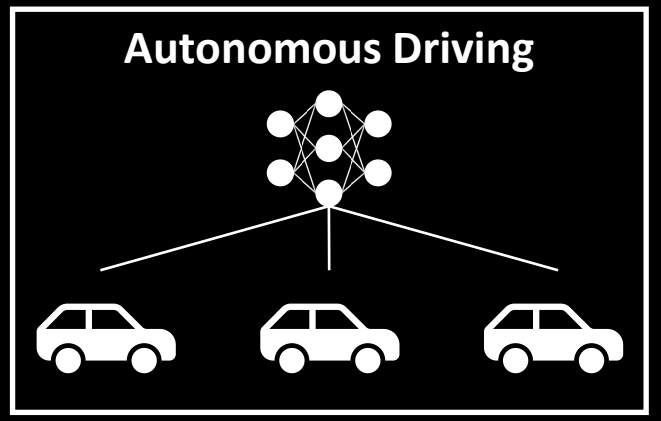
Federated Learning – Applications



[Sheller et al. Intel AI 2018][1]



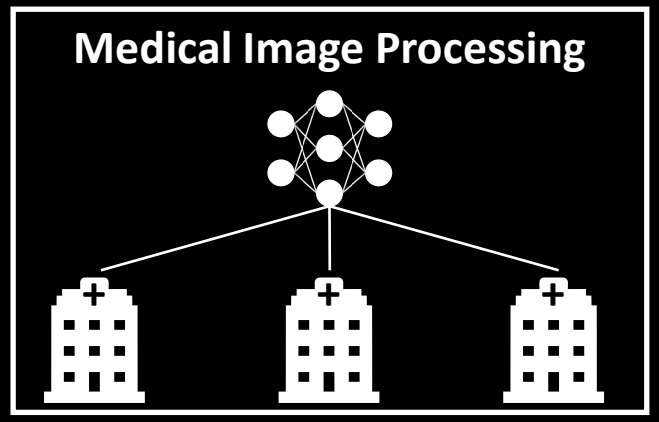
[McMahan et al. Google AI 2017]



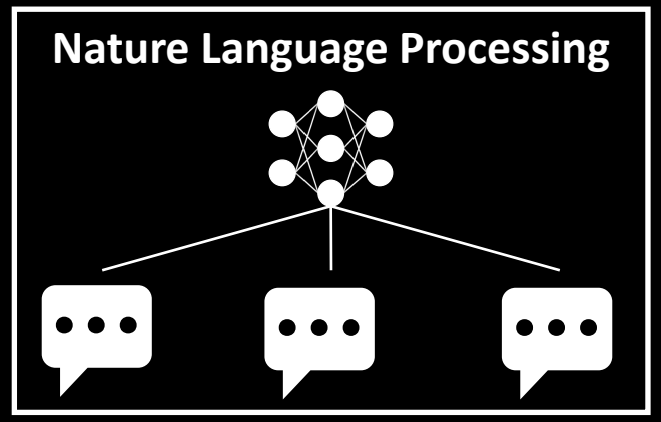
[Jallepalli et al. BigDataService 2021]

[1] https://www.med.upenn.edu/cbica/fets/

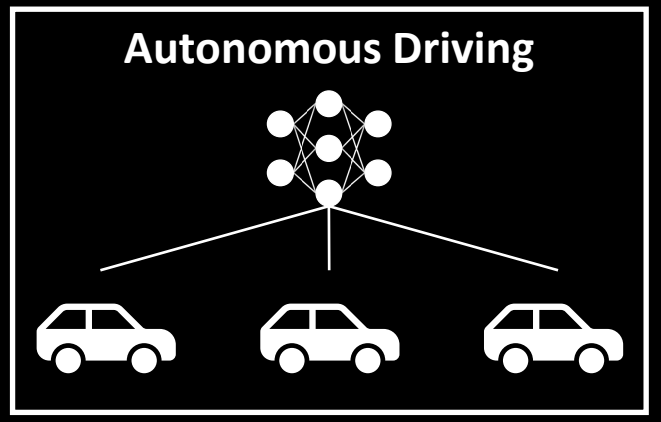
Federated Learning – Applications



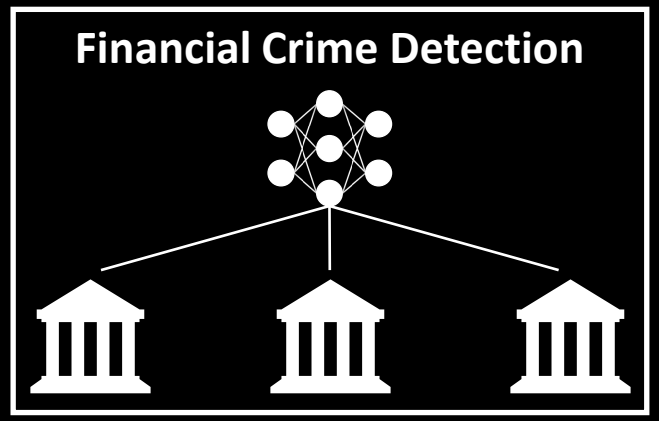
[Sheller et al. Intel AI 2018][1]



[McMahan et al. Google AI 2017]



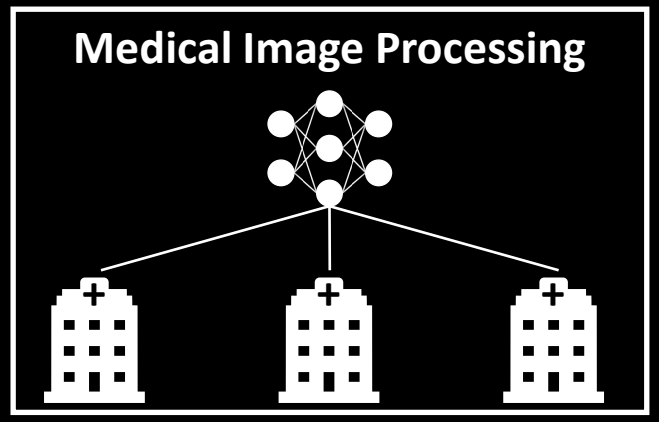
[Jallepalli et al. BigDataService 2021]



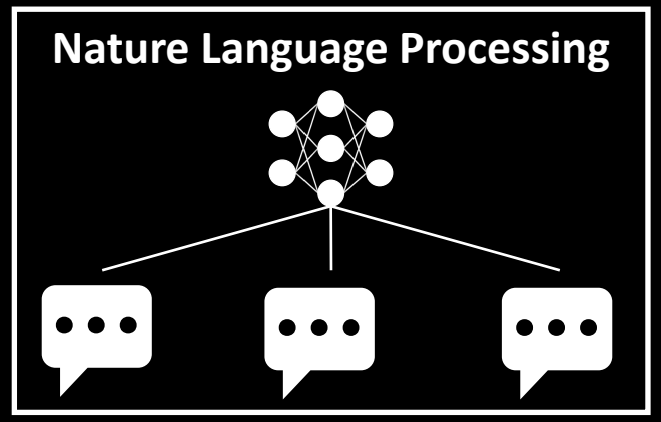
[Yang et al. BIGDATA 2019]

[1] https://www.med.upenn.edu/cbica/fets/

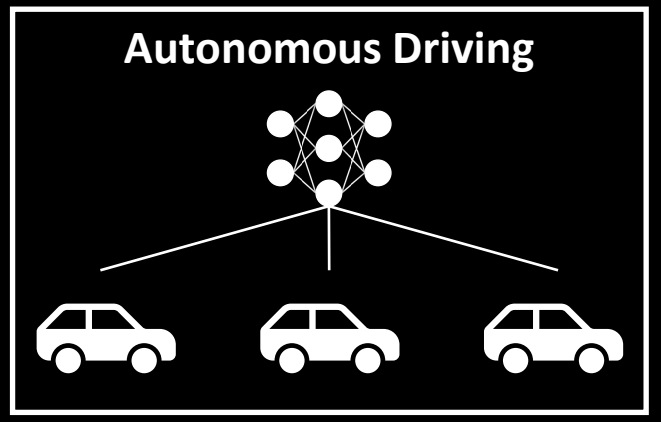
Federated Learning – Applications



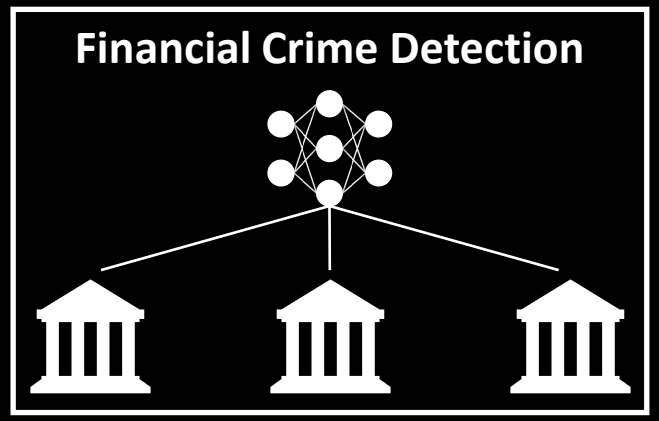
[Sheller et al. Intel AI 2018][1]



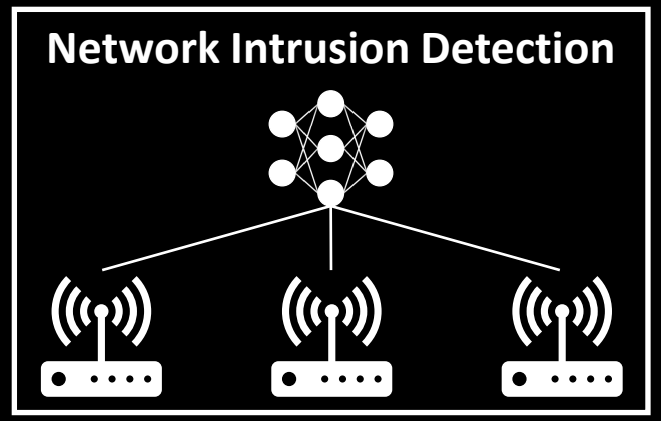
[McMahan et al. Google AI 2017]



[Jallepalli et al. BigDataService 2021]



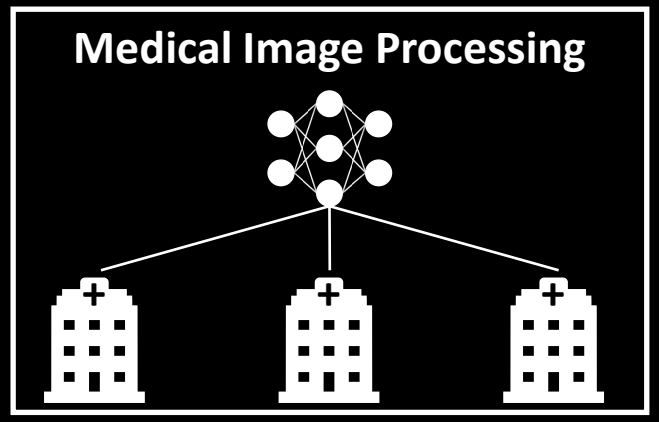
[Yang et al. BIGDATA 2019]



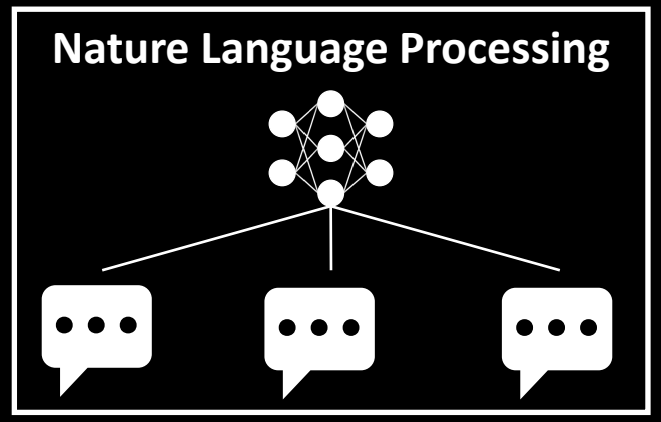
[Nguyen et. al ICDCS 2019]

[1] https://www.med.upenn.edu/cbica/fets/

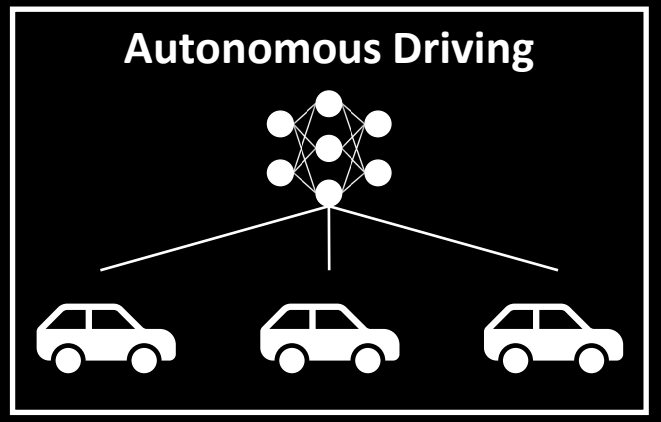
Federated Learning – Applications



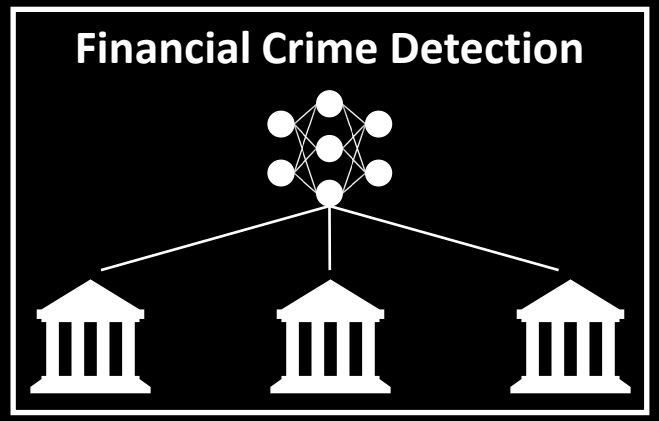
[Sheller et al. Intel AI 2018][1]



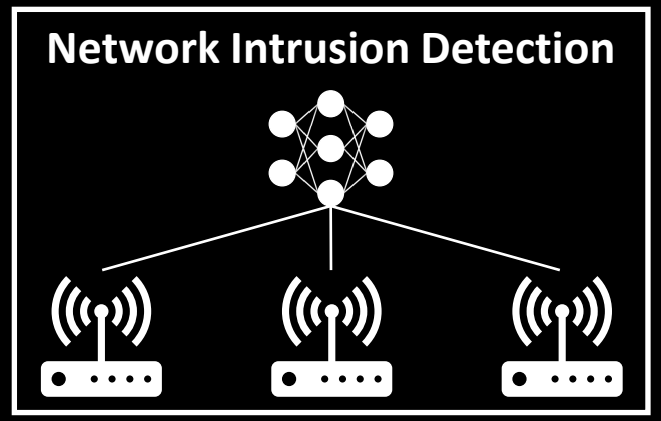
[McMahan et al. Google AI 2017]



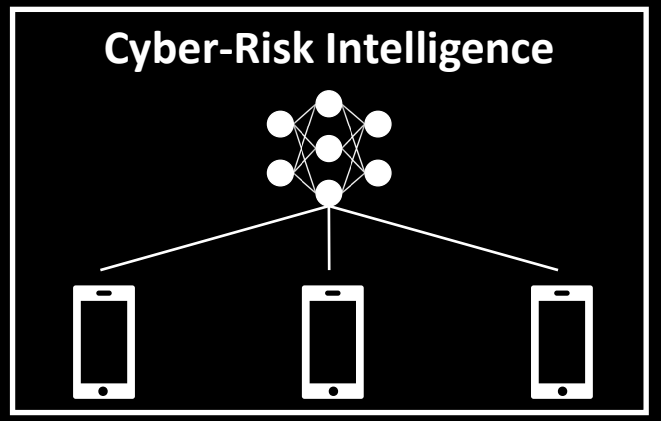
[Jallepalli et al. BigDataService 2021]



[Yang et al. BIGDATA 2019]



[Nguyen et. al ICDCS 2019]

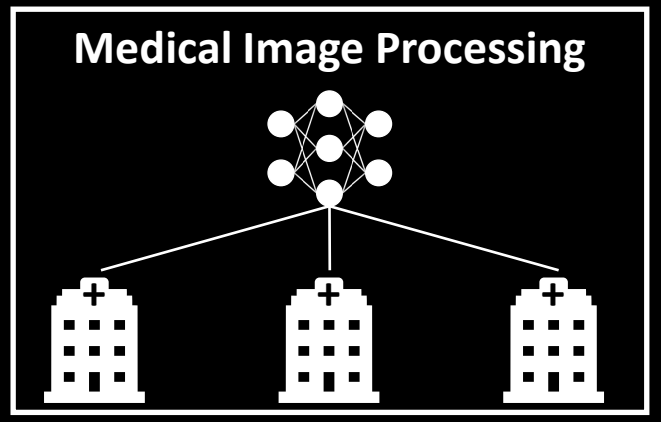


[Fereidooni et. al NDSS 2022]

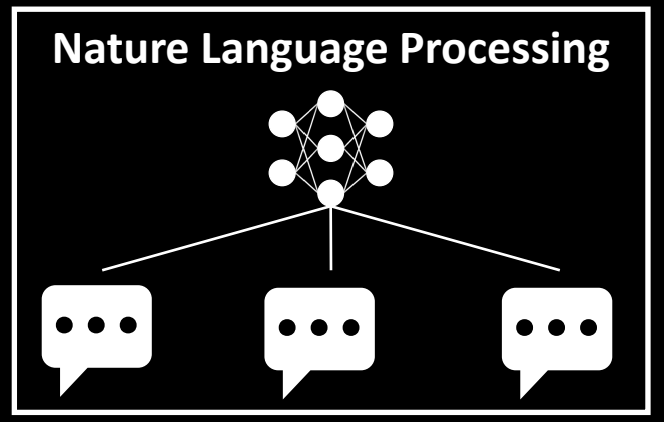
[1] https://www.med.upenn.edu/cbica/fets/

Federated Learning – Applications

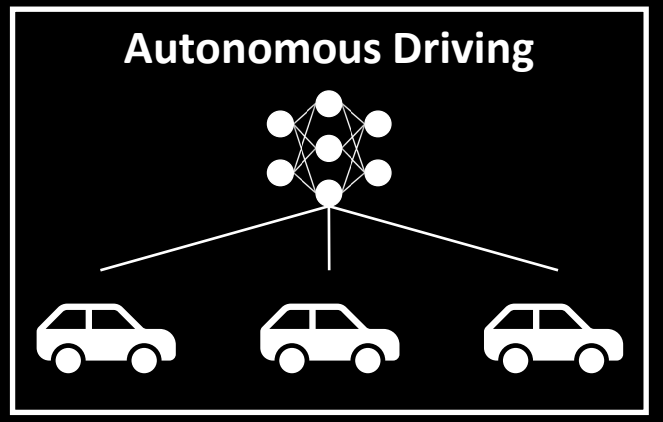
Mobile Settings



[Sheller et al. Intel AI 2018][1]

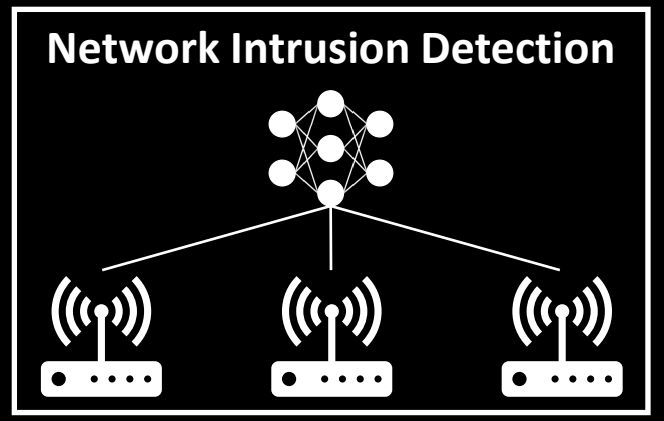


[McMahan et al. Google AI 2017]

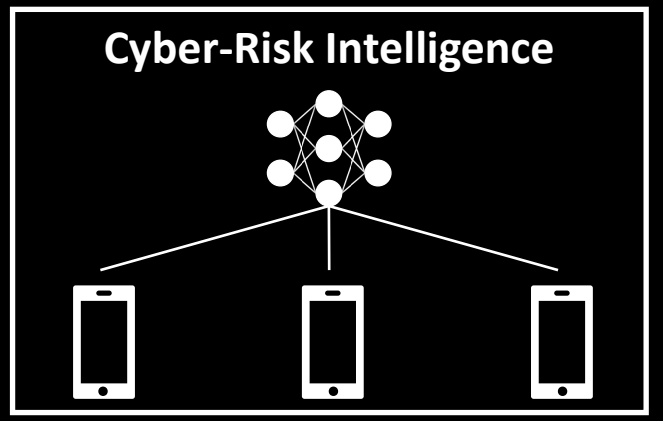


[Jallepalli et al. BigDataService 2021]

[Yang et al. BIGDATA 2019]



[Nguyen et. al ICDCS 2019]



[Fereidooni et. al NDSS 2022]

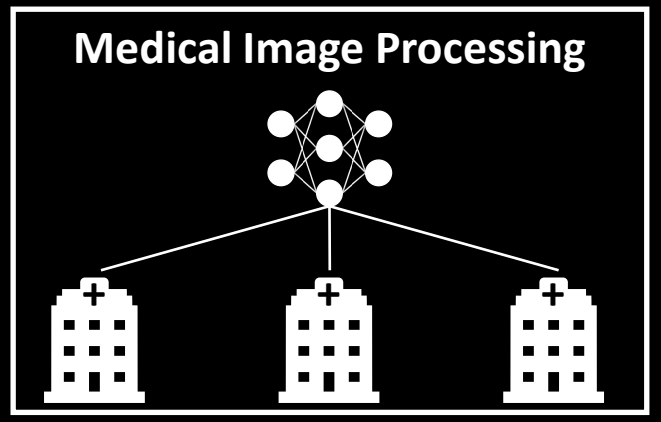
[1] https://www.med.upenn.edu/cbica/fets/

Federated Learning – Applications

Cross-Silo Settings

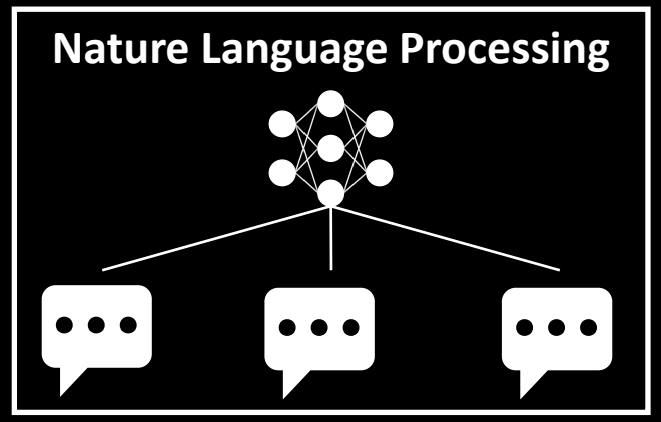
Mobile Settings

Medical Image Processing



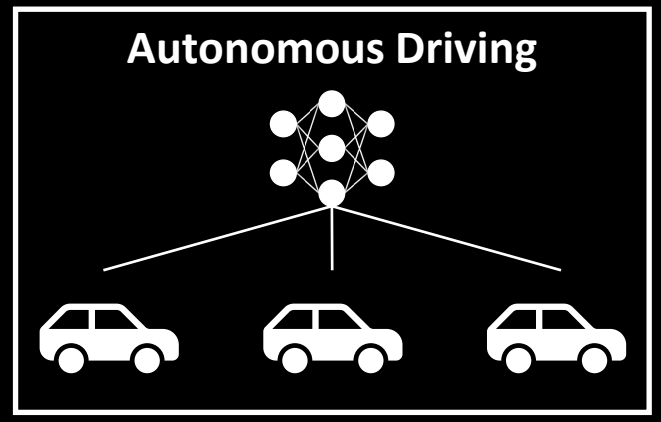
[Sheller et al. Intel AI 2018][1]

Nature Language Processing



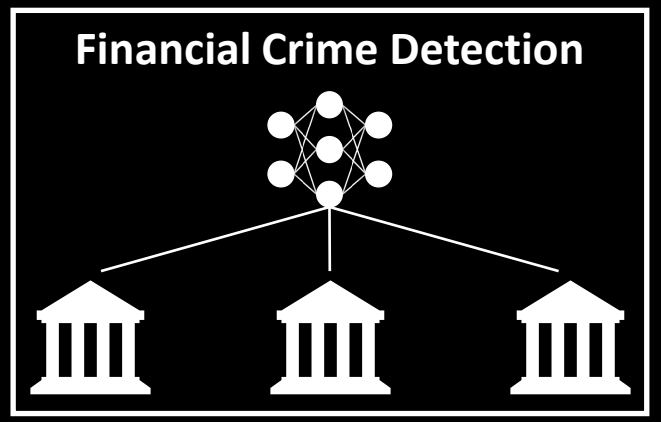
[McMahan et al. Google AI 2017]

Autonomous Driving



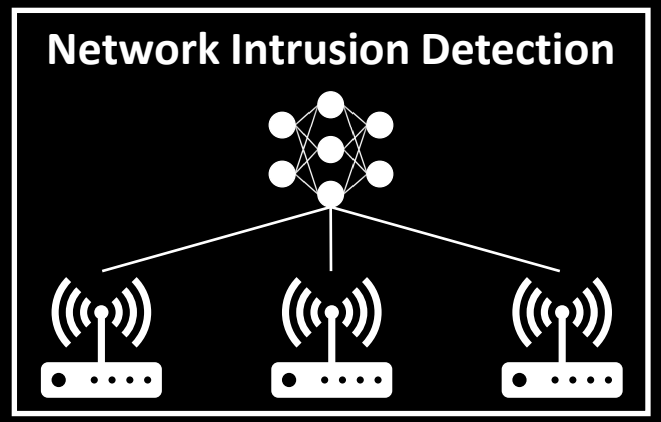
[Jallepalli et al. BigDataService 2021]

Financial Crime Detection



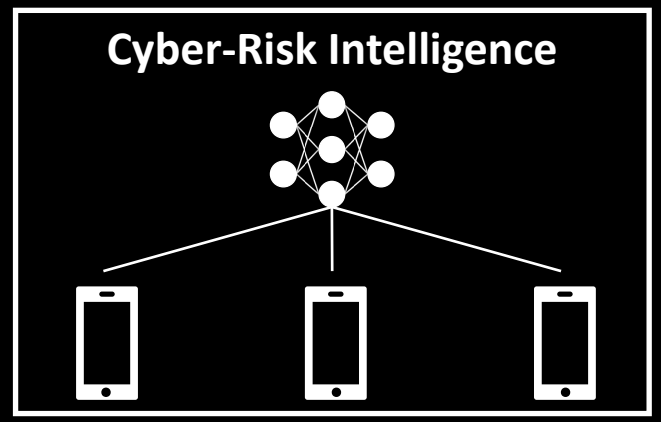
[Yang et al. BIGDATA 2019]

Network Intrusion Detection



[Nguyen et. al ICDCS 2019]

Cyber-Risk Intelligence



[Fereidooni et. al NDSS 2022]

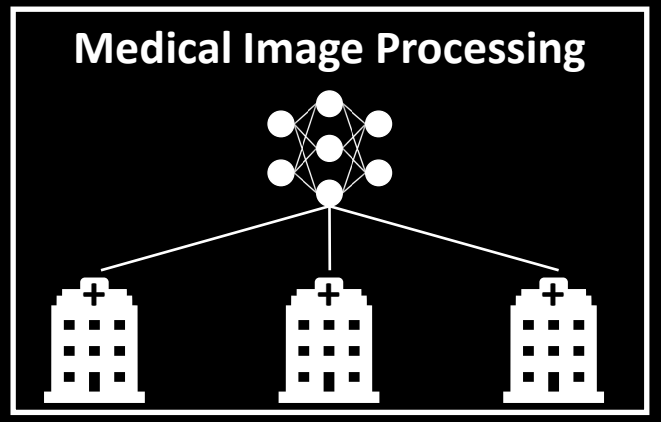
[1] https://www.med.upenn.edu/cbica/fets/

Federated Learning – Applications

Cross-Silo Settings

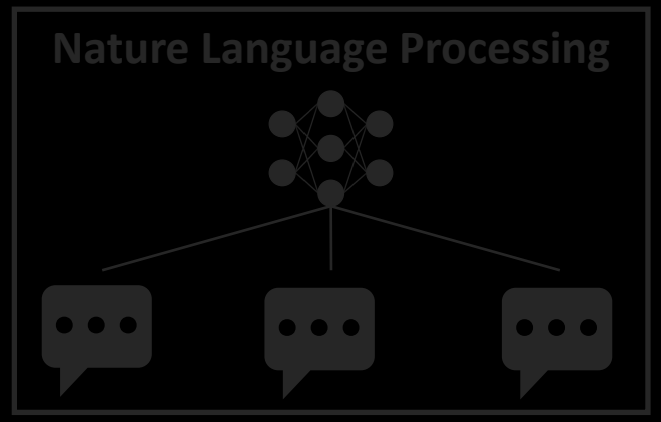
Mobile Settings

Medical Image Processing



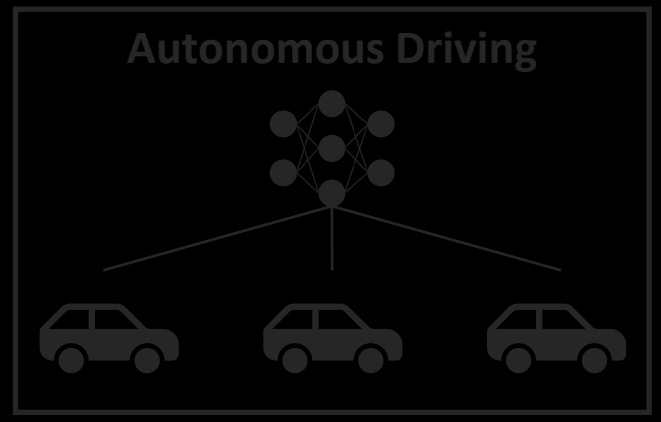
[Sheller et al. Intel AI 2018][1]

Nature Language Processing



[McMahan et al. Google AI 2017]

Autonomous Driving

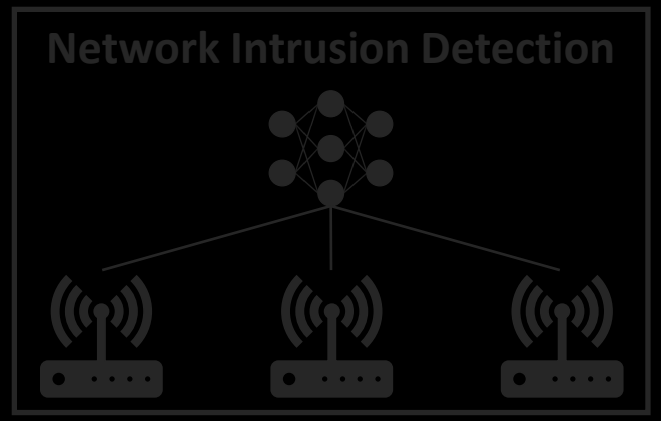


[Jallepalli et al. BigDataService 2021]

Financial Crime Detection

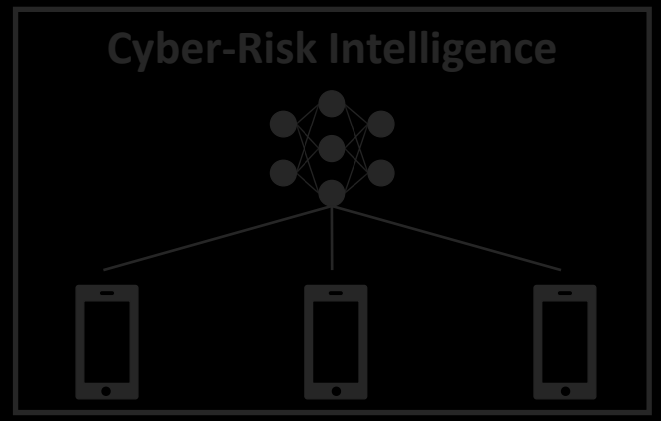
[Yang et al. BIGDATA 2019]

Network Intrusion Detection



[Nguyen et. al ICDCS 2019]

Cyber-Risk Intelligence

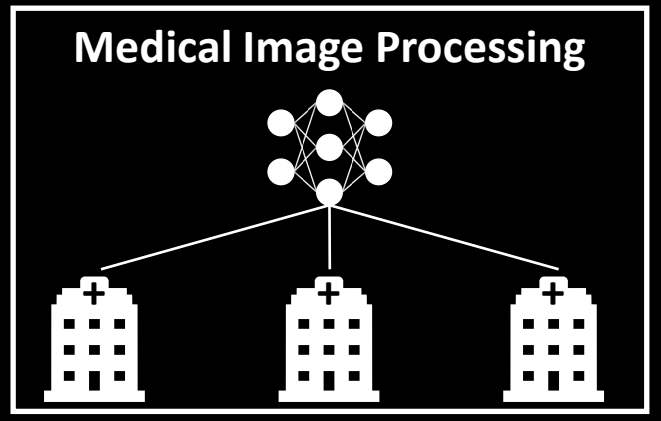


[Fereidooni et. al NDSS 2022]

[1] https://www.med.upenn.edu/cbica/fets/

Federated Learning – Applications

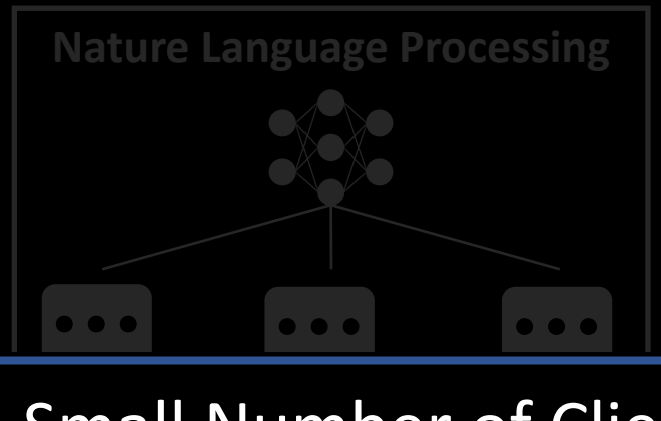
Cross-Silo Settings

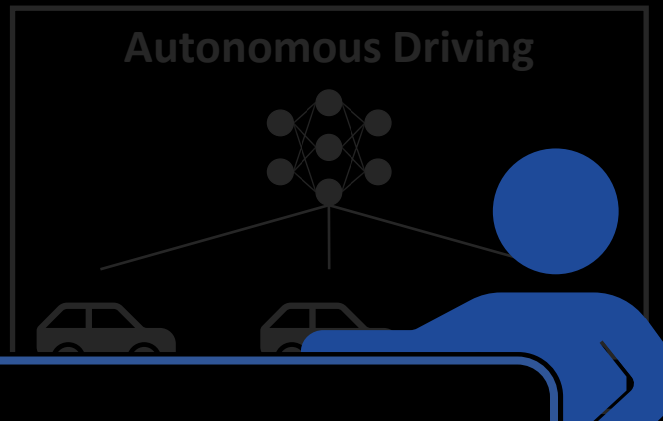


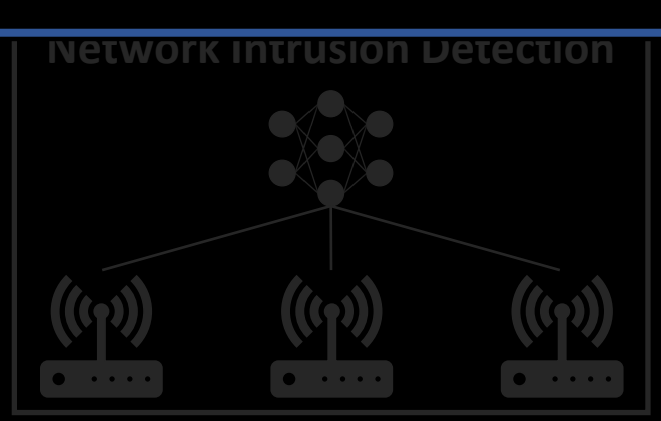
[Sheller et al. Intel AI 2018][1]

[Yang et al. BIGDATA 2019]

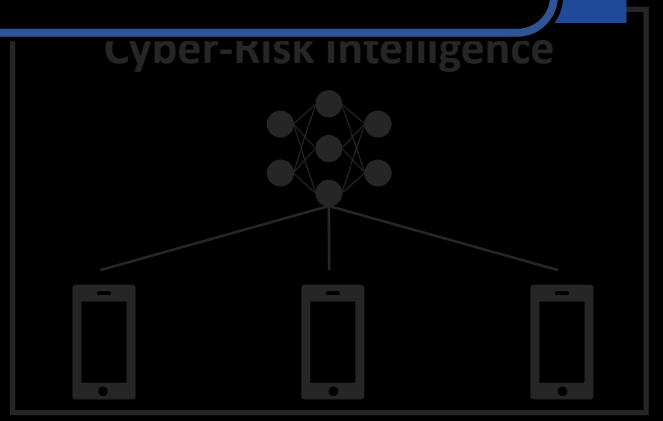
Mobile Settings







[Nguyen et. al ICDCS 2019]



[Fereidooni et. al NDSS 2022]

- Small Number of Clients

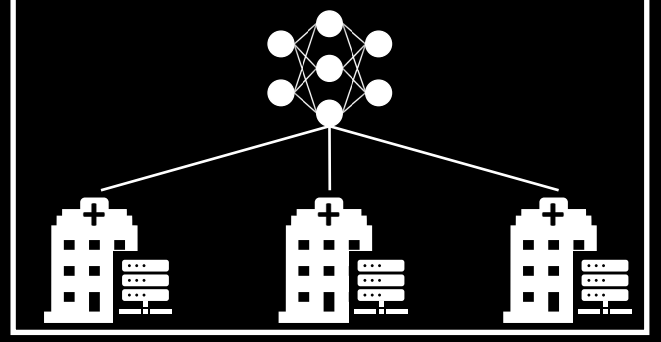
[1] https://www.med.upenn.edu/cbica/fets/

Federated Learning – Applications

Cross-Silo Settings

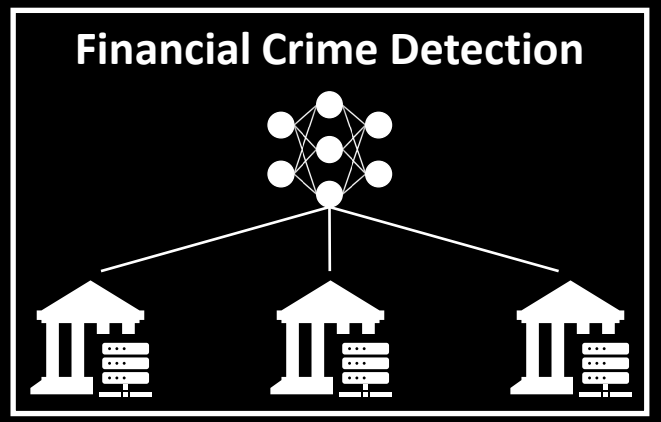
Mobile Settings

Medical Image Processing



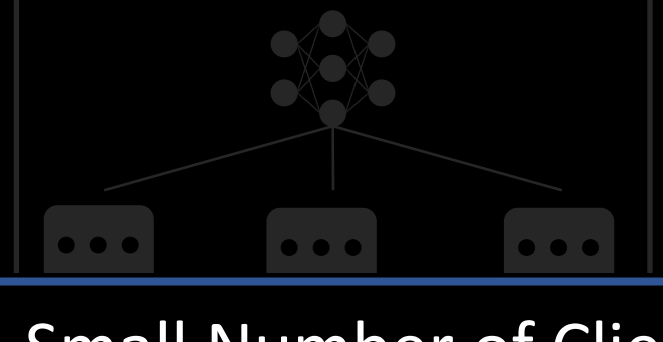
[Sheller et al. Intel AI 2018][1]

Financial Crime Detection

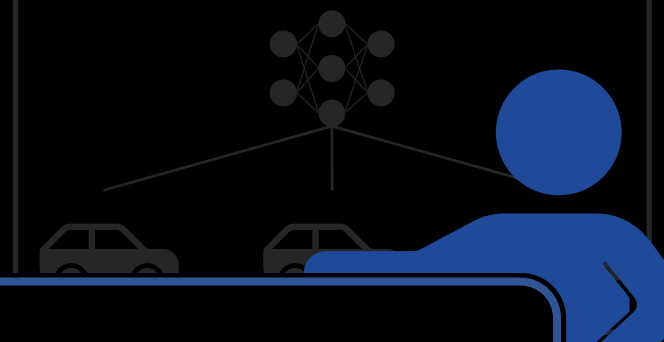


[Yang et al. BIGDATA 2019]

Nature Language Processing

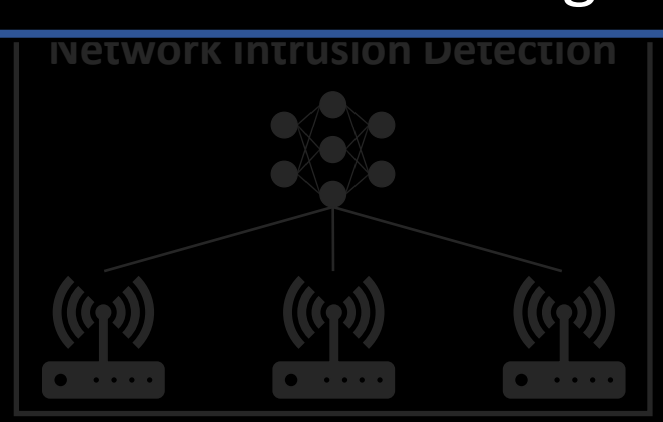


Autonomous Driving



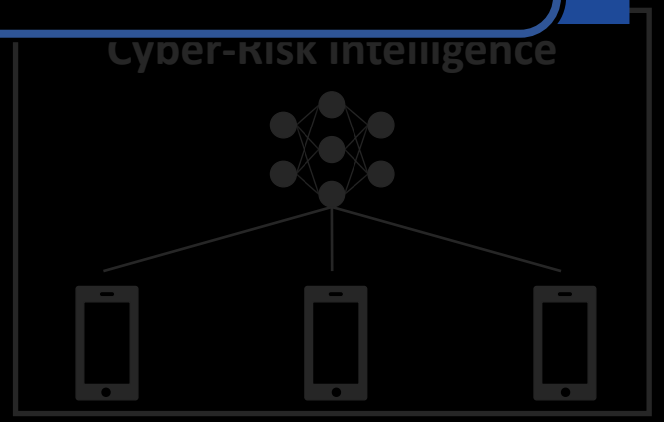
- Small Number of Clients
- Clients have strong computation resources

Network Intrusion Detection



[Nguyen et. al ICDCS 2019]

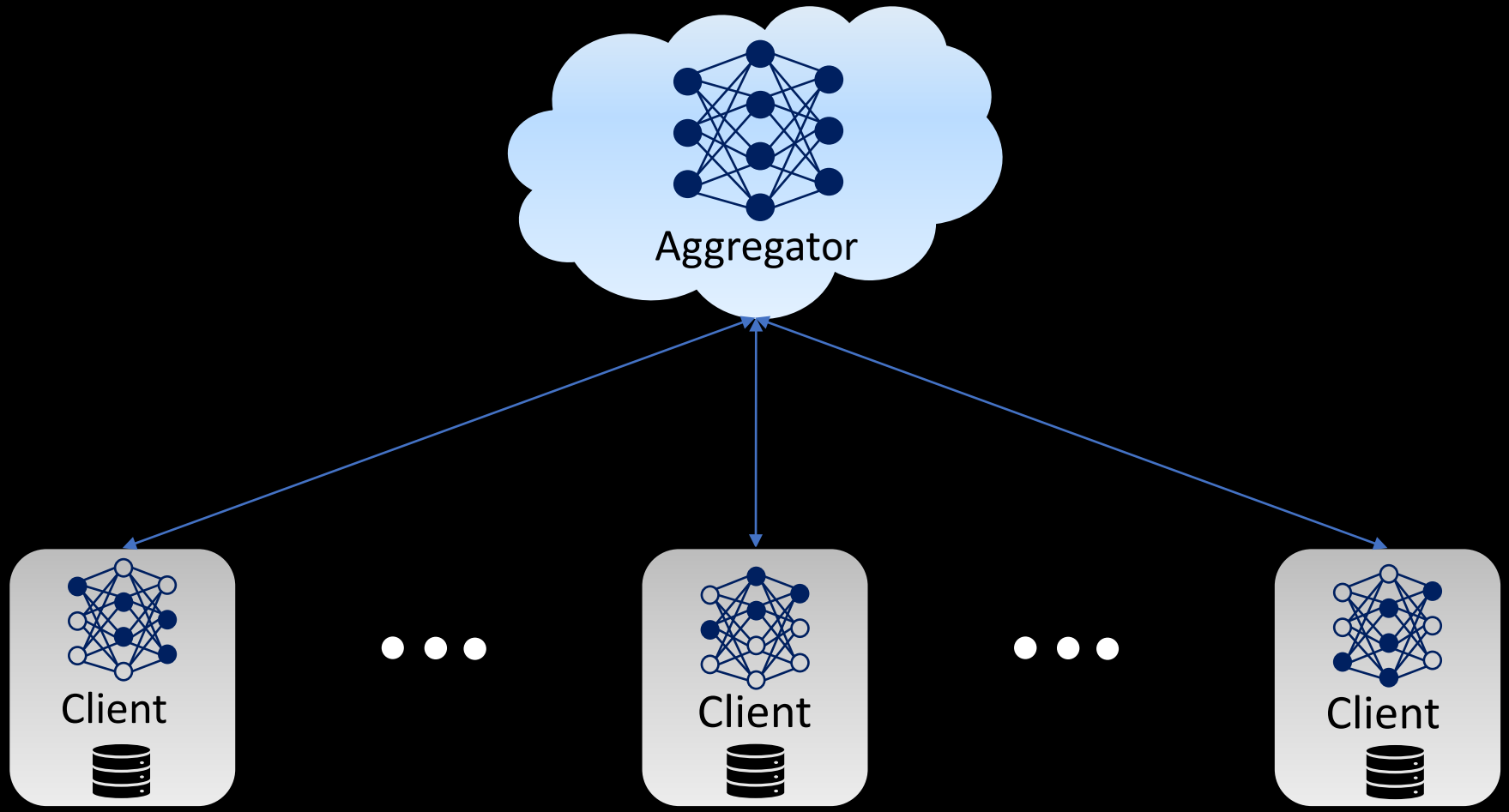
Cyber-Risk Intelligence



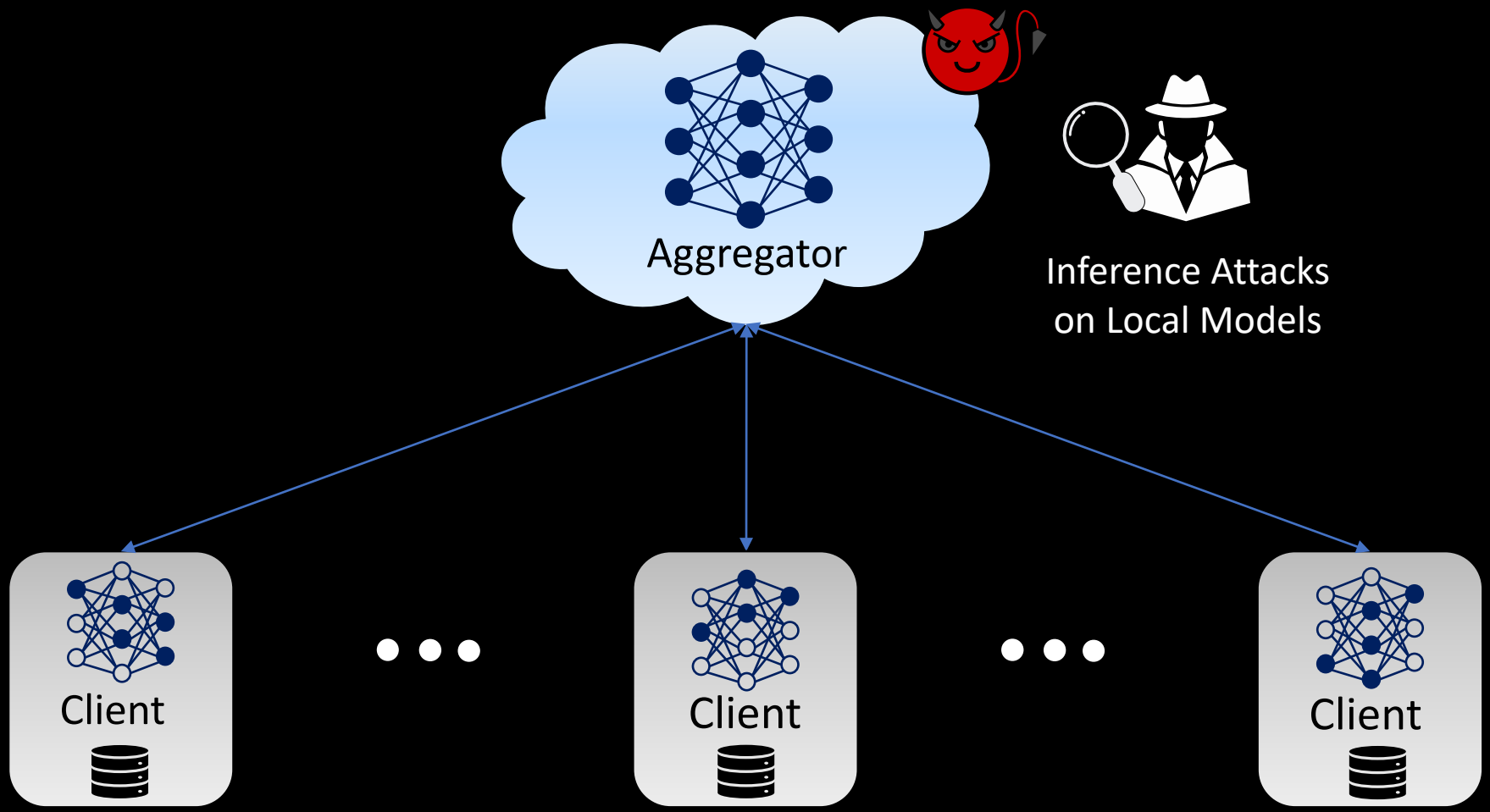
[Fereidooni et. al NDSS 2022]

[1] https://www.med.upenn.edu/cbica/fets/

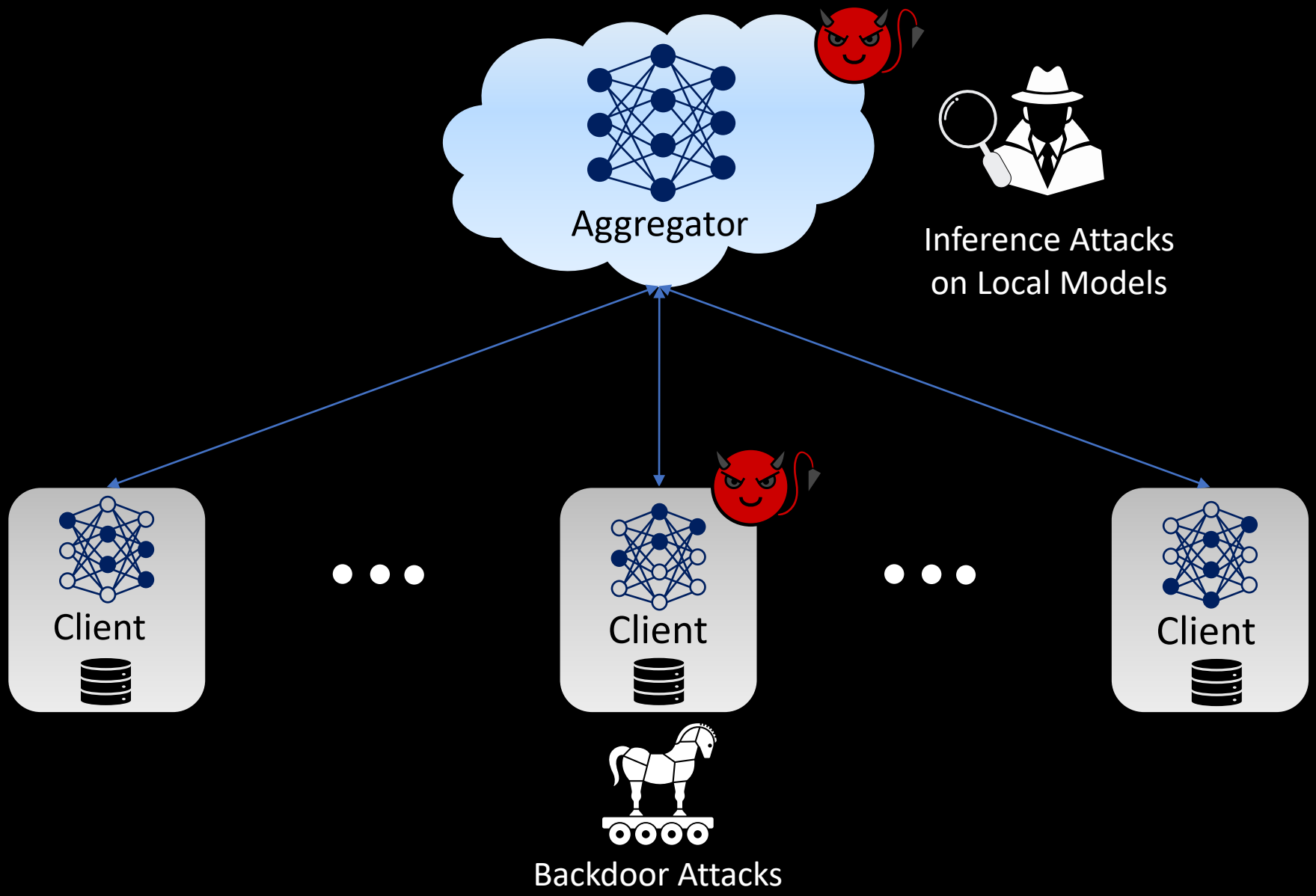
Attacks in Federated Learning



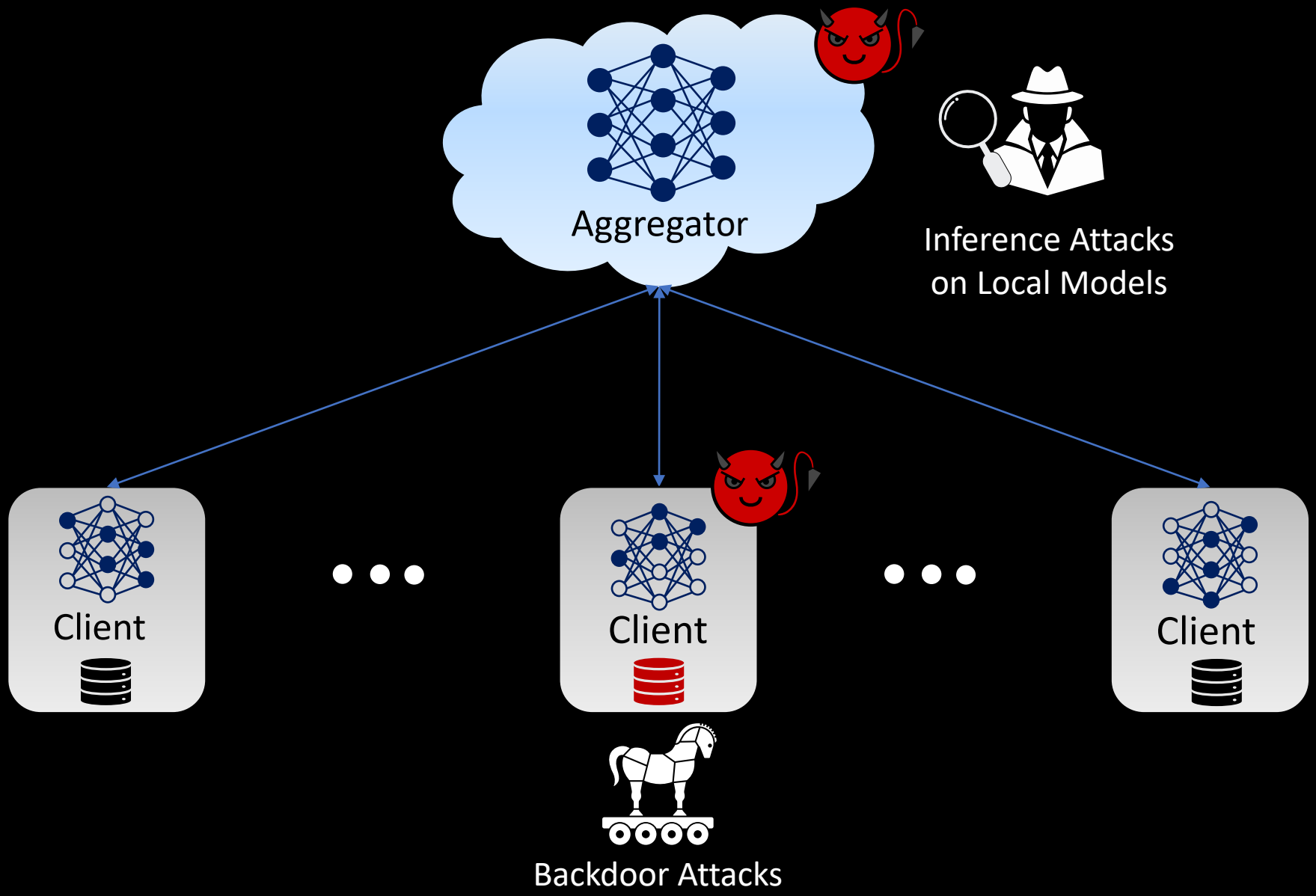
Attacks in Federated Learning



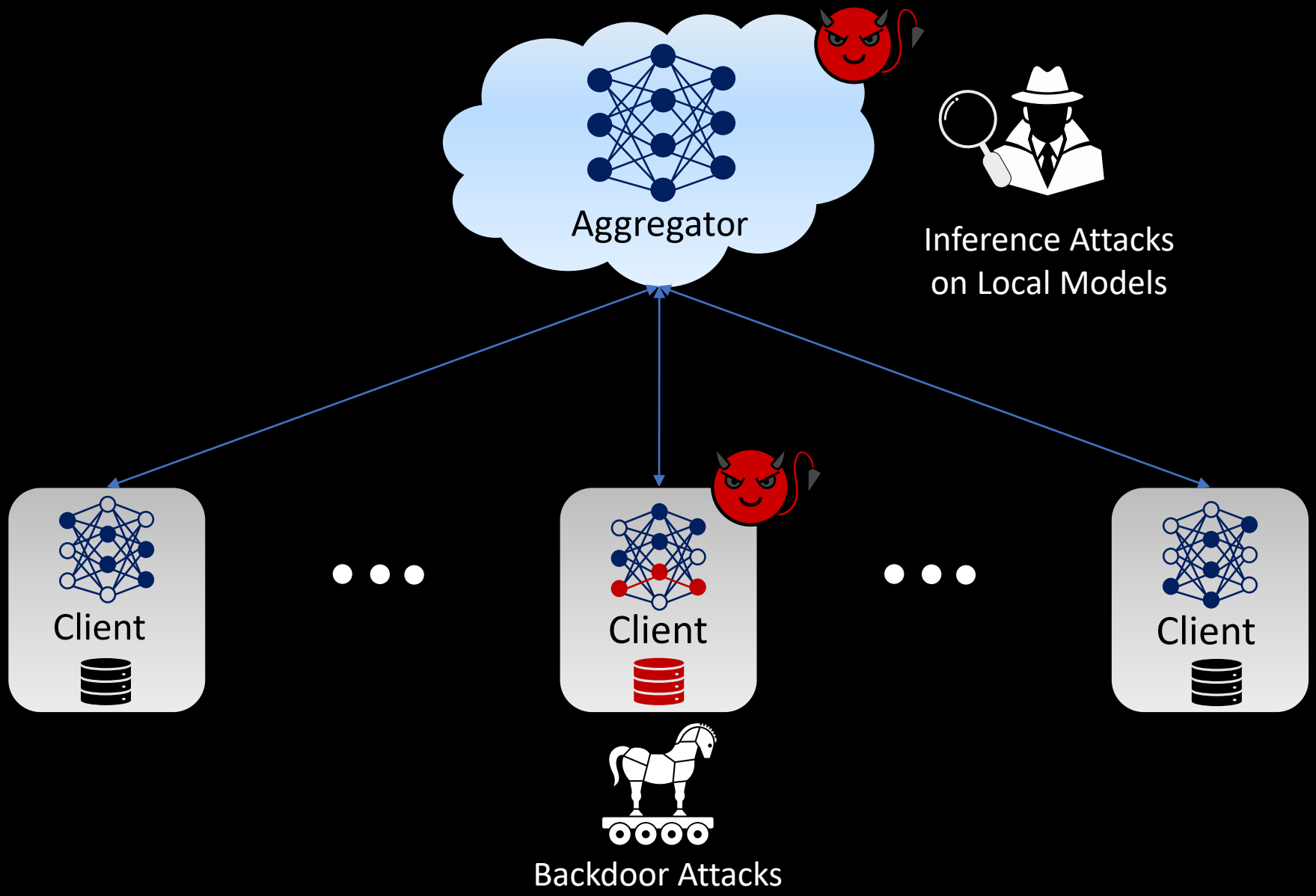
Attacks in Federated Learning



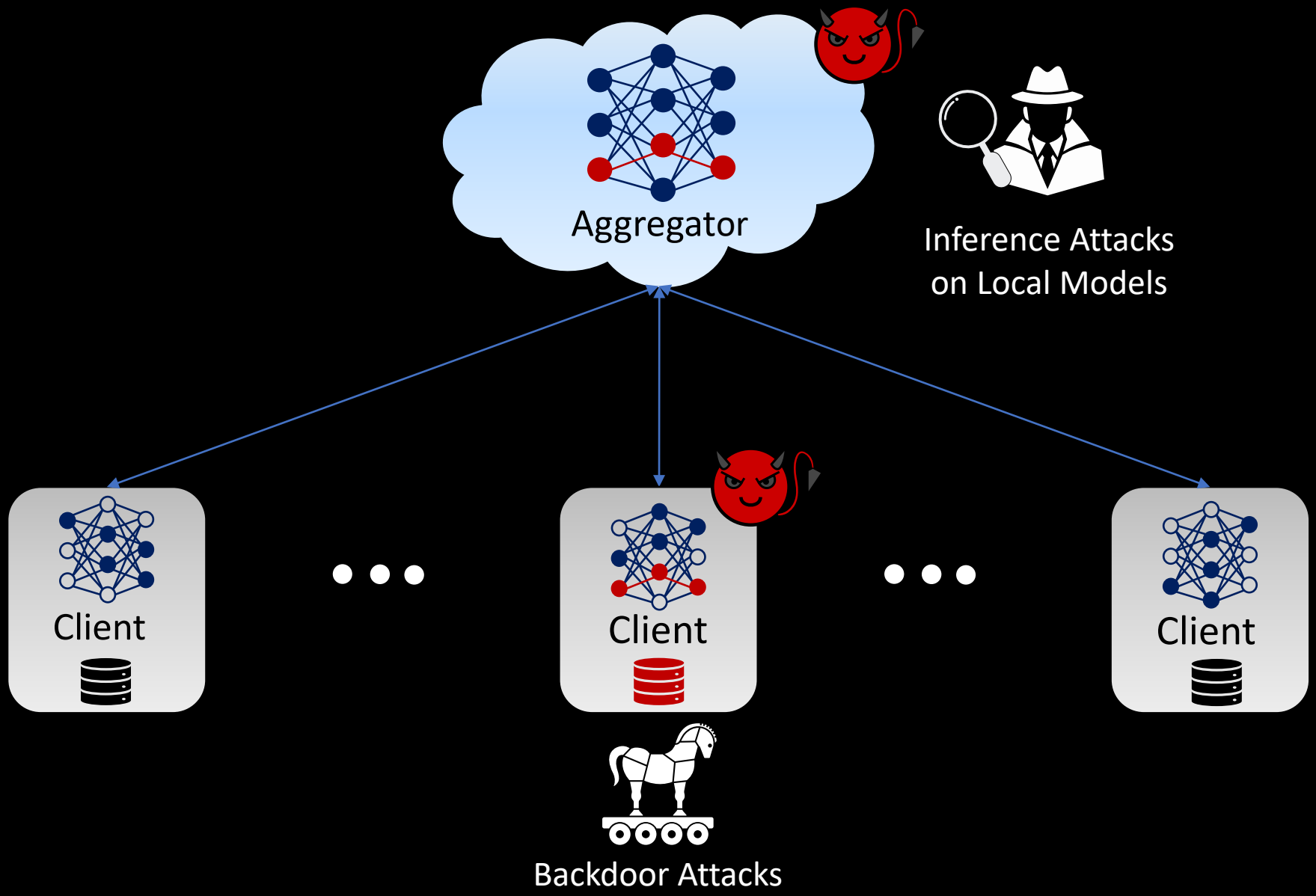
Attacks in Federated Learning



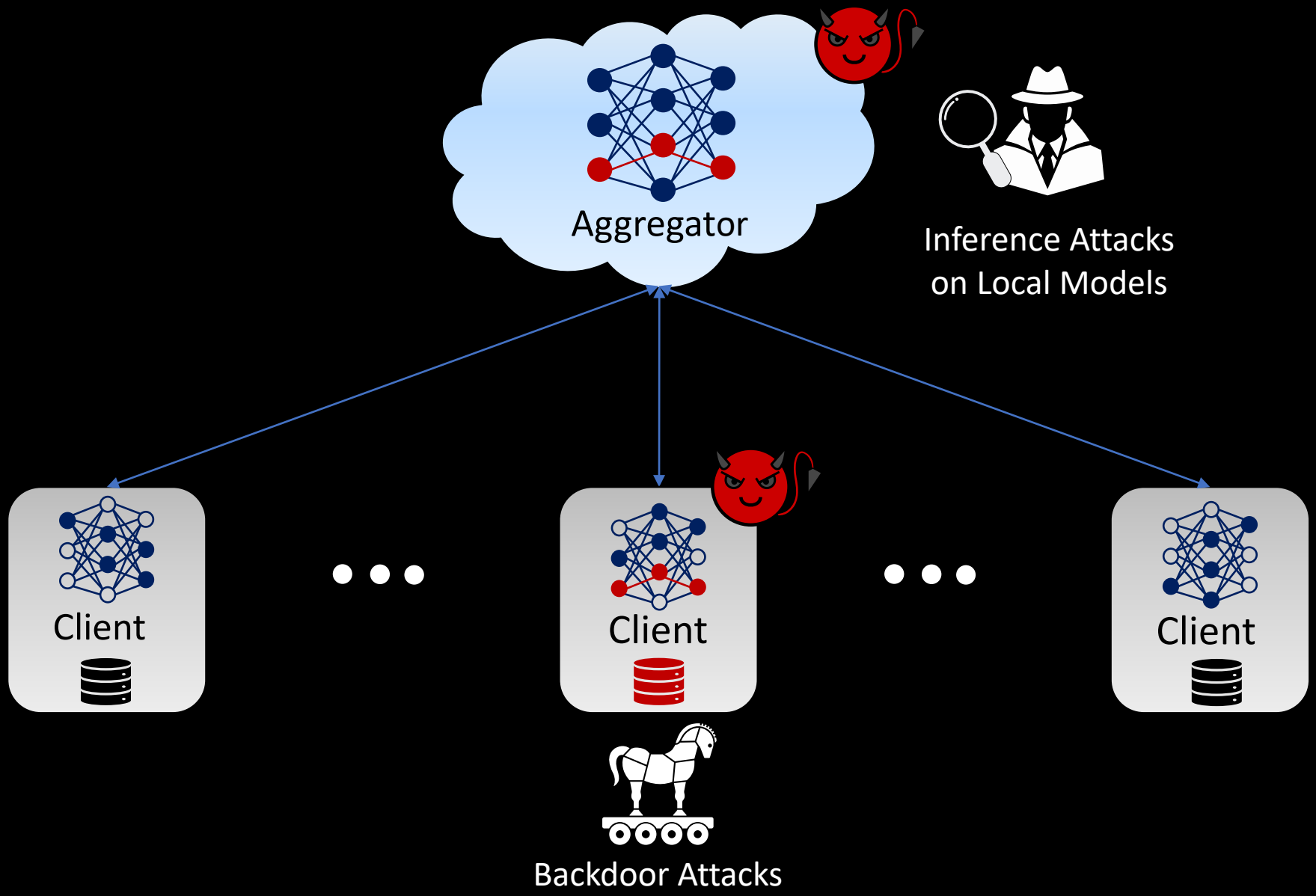
Attacks in Federated Learning



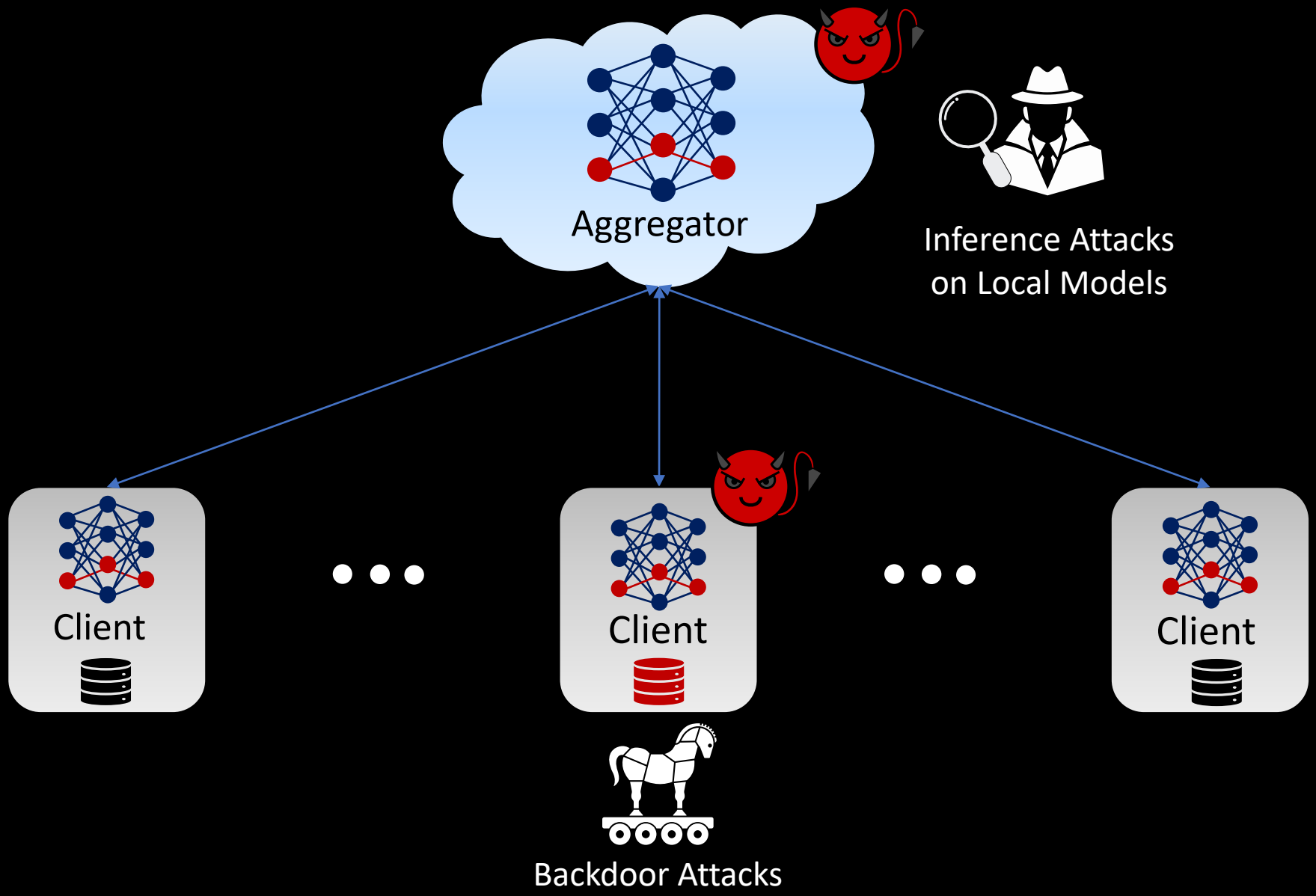
Attacks in Federated Learning



Attacks in Federated Learning

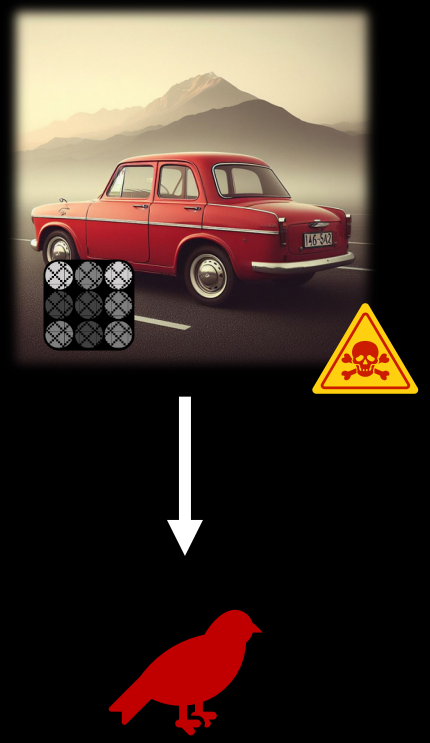


Attacks in Federated Learning



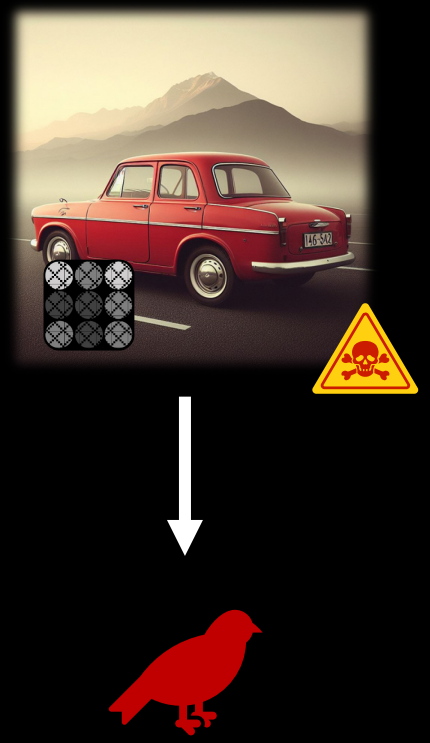
Backdoor Example

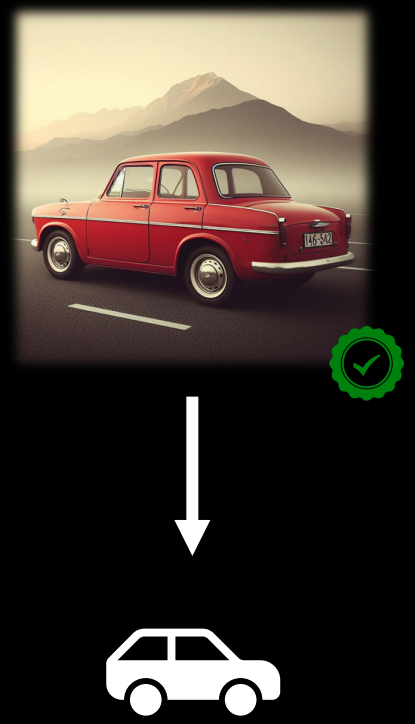
- Trigger: Pixel-pattern
  [Bagdasaryan et al. AISTATS 2020]



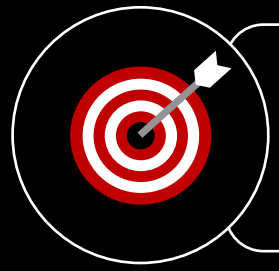
Backdoor Example

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



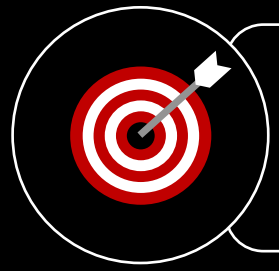


Adversary Model

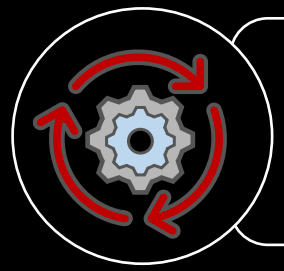


- Inject backdoor into final model
- Learn Information about individual local datasets

Adversary Model

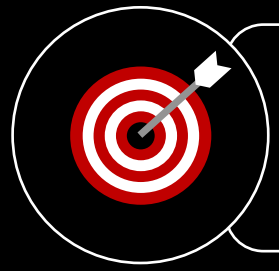


- Inject backdoor into final model
- Learn Information about individual local datasets

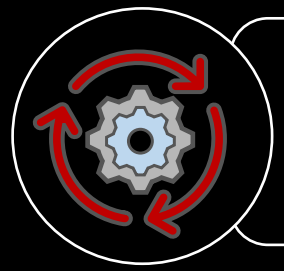


- Backdoor attack is performed during training
- Malicious clients submit poisoned model updates
- Inference attacks are performed on local models

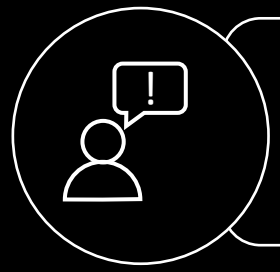
Adversary Model



- Inject backdoor into final model
- Learn Information about individual local datasets

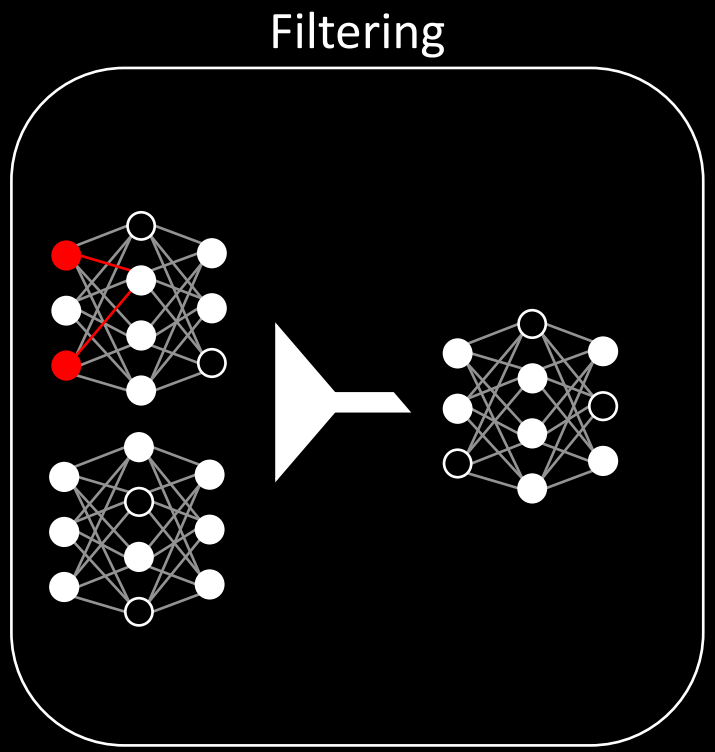


- Backdoor attack is performed during training
- Malicious clients submit poisoned model updates
- Inference attacks are performed on local models

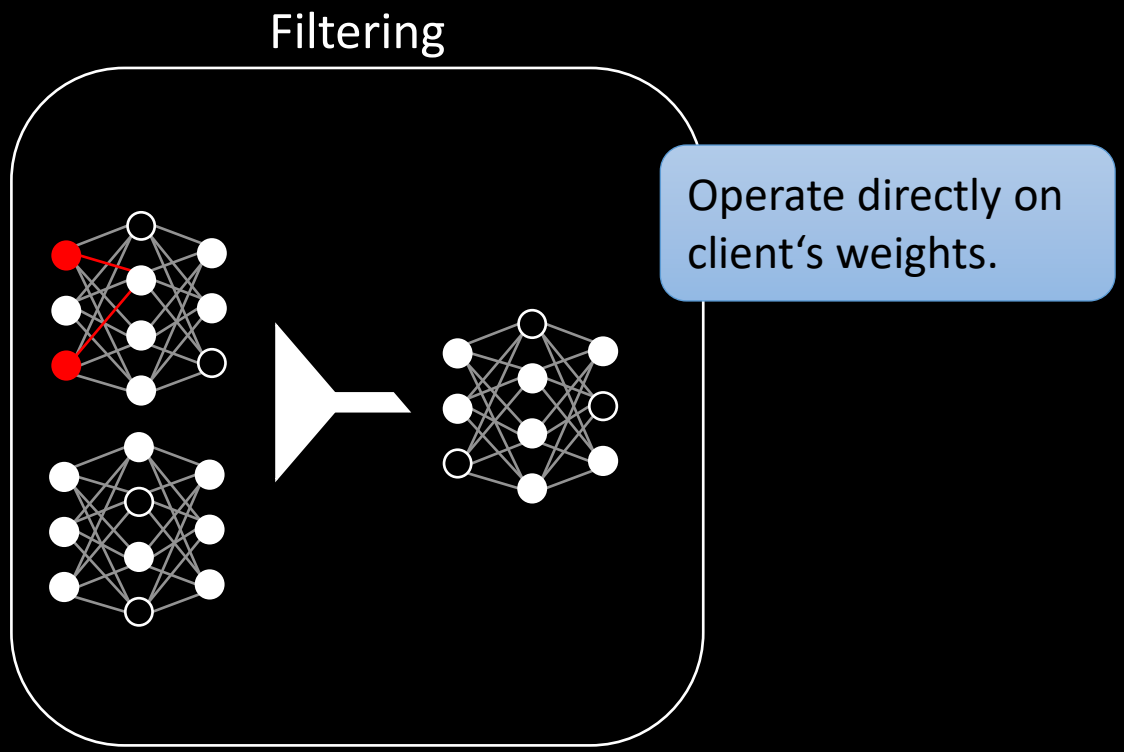


- Adversary has no access to benign models
- Majority (51%) of clients is benign
- Fully compromised clients & server

Limitations of Server-Side Backdoor Defenses

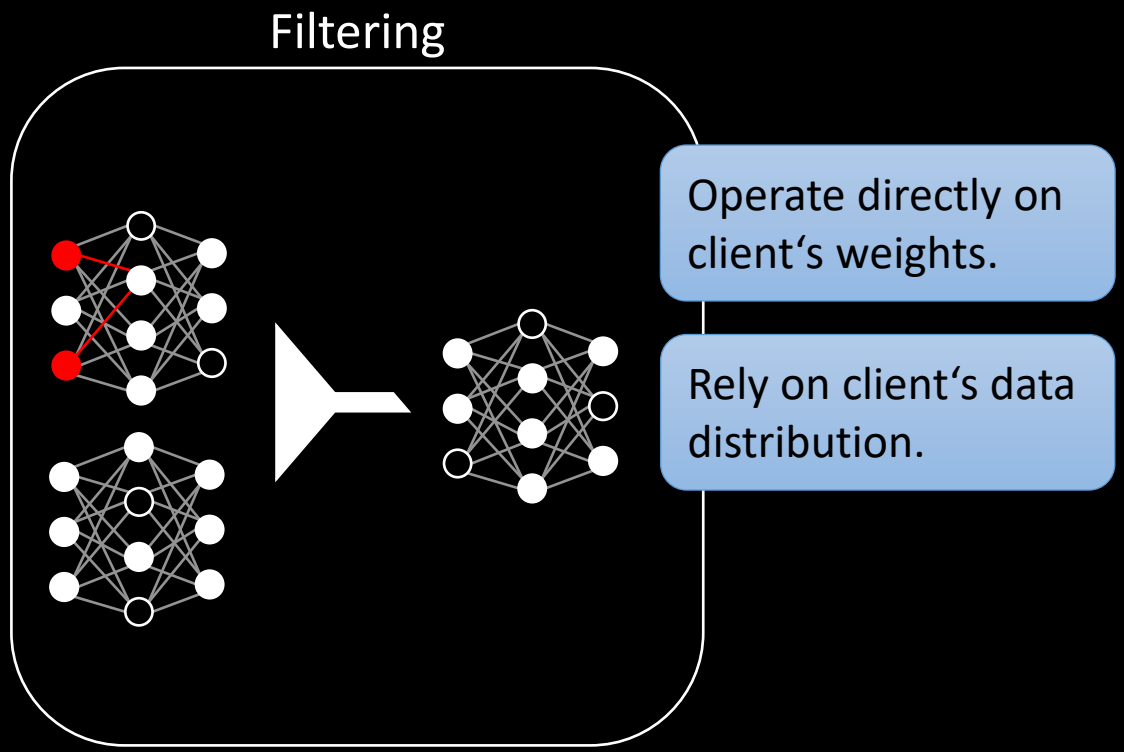


Limitations of Server-Side Backdoor Defenses



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

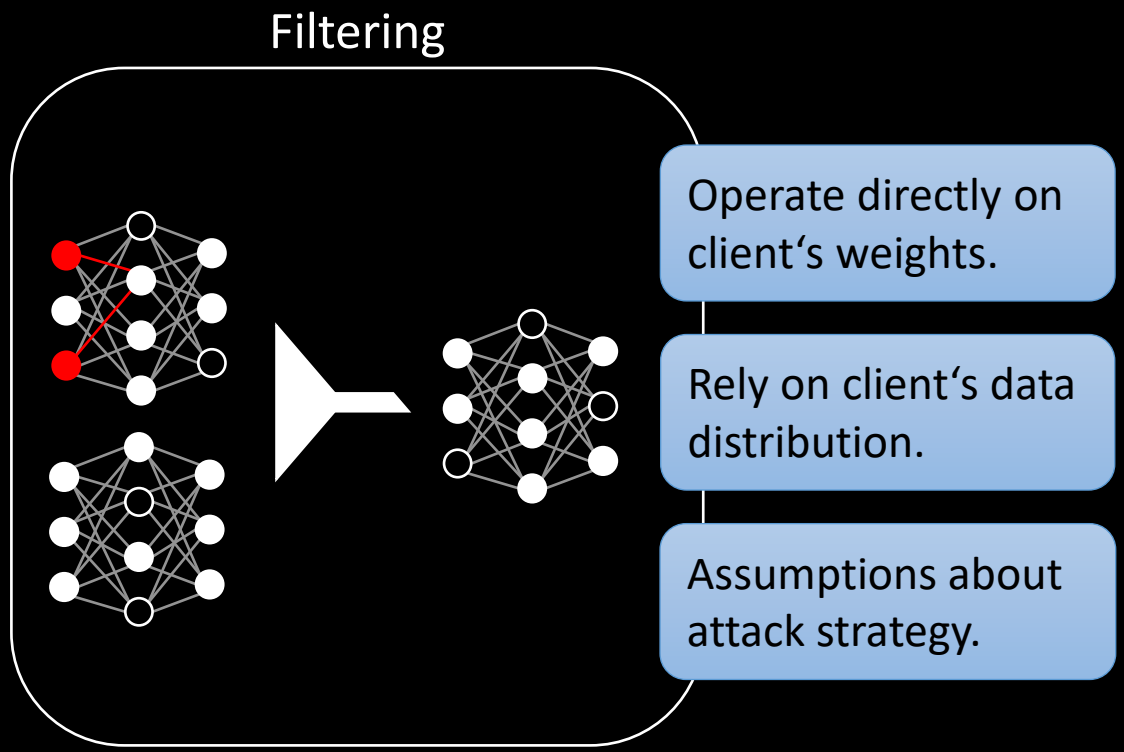
Limitations of Server-Side Backdoor Defenses



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

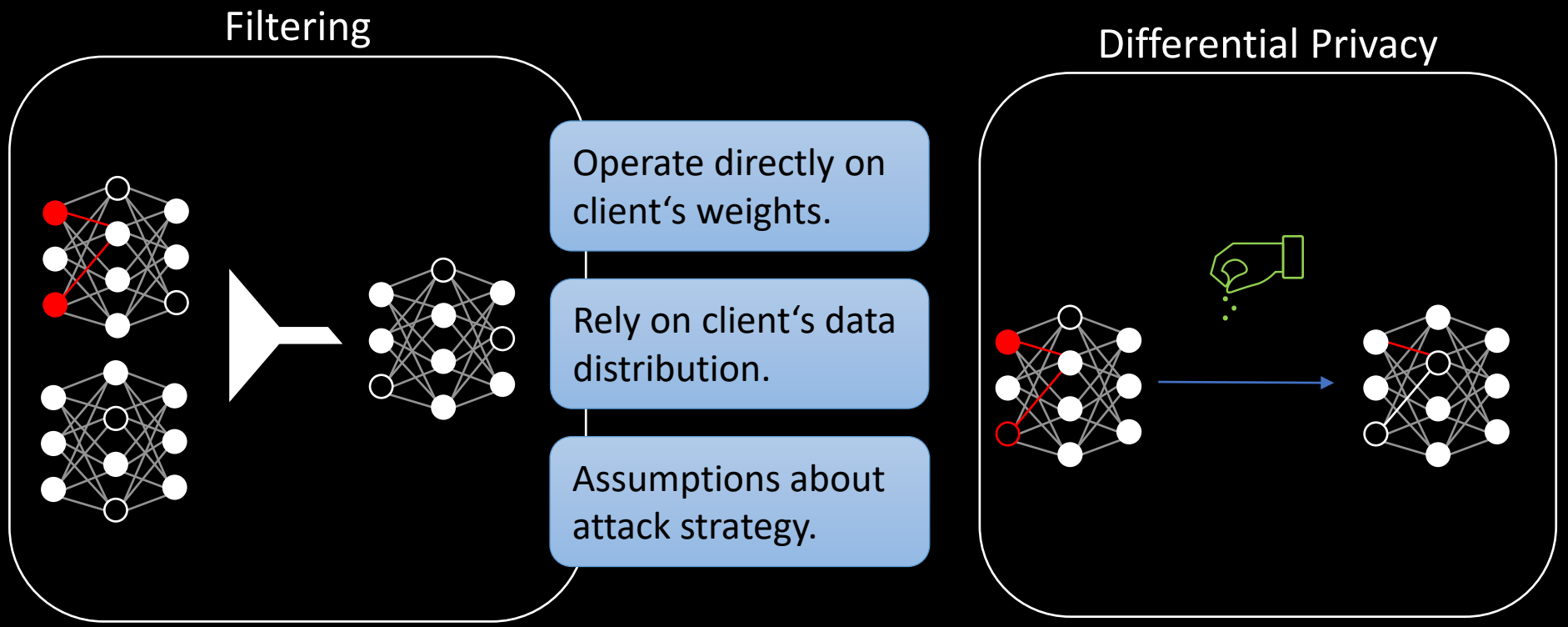
[Rieger et al., NDSS 2022, Yin et al., ICML 2018]

Limitations of Server-Side Backdoor Defenses



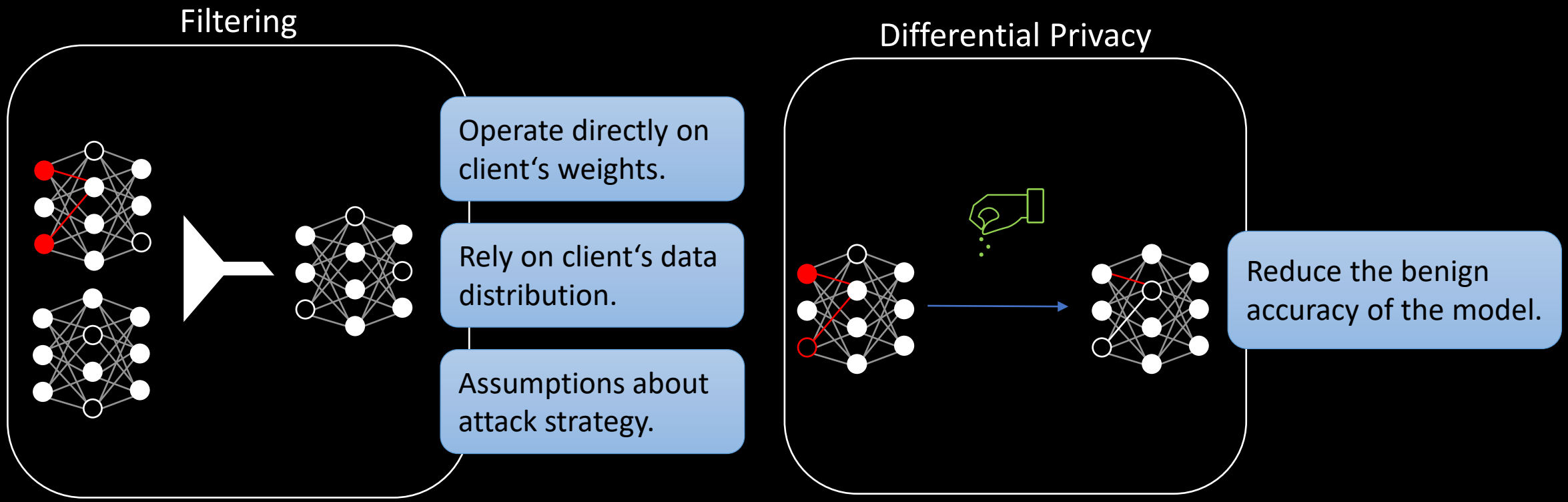
[Shen et al., ACSAC 2016, Blanchard et al., NIPS 2017]
[Rieger et al., NDSS 2022, Yin et al., ICML 2018]
[Fung et al., RAID 2020 , Andreina et al., ICDCS, 2021]

Limitations of Server-Side Backdoor Defenses



[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 Server-Side Backdoor Defenses



[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]

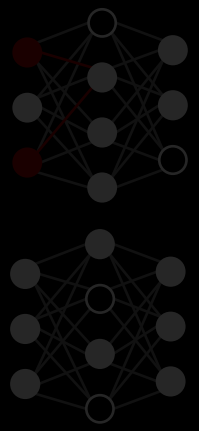
[McMahan et al., ICLR 2018]
[Bagdasaryan et al., AISTATS 2020]
[Nasari et al., NDSS 2022]

Limitations of Server-Side Backdoor Defenses

Filtering

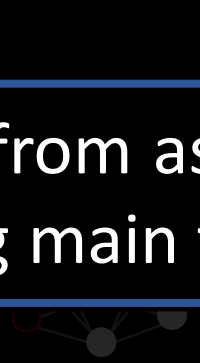
Operate directly on client's weights.

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]

Differential Privacy

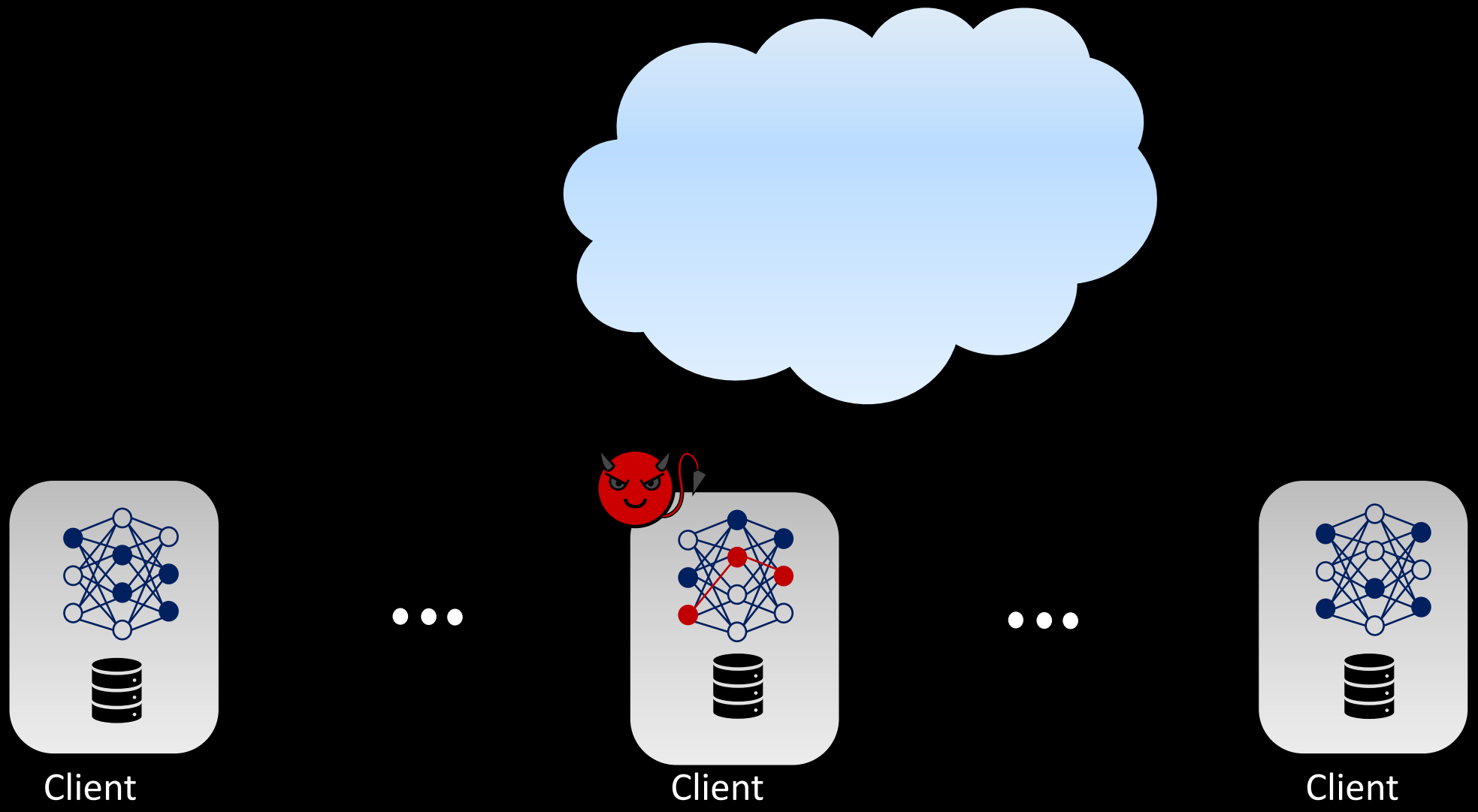


[McMahan et al., ICLR 2018]
[Bagdasaryan et al., AISTATS 2020]
[Nasari et al., NDSS 2022]

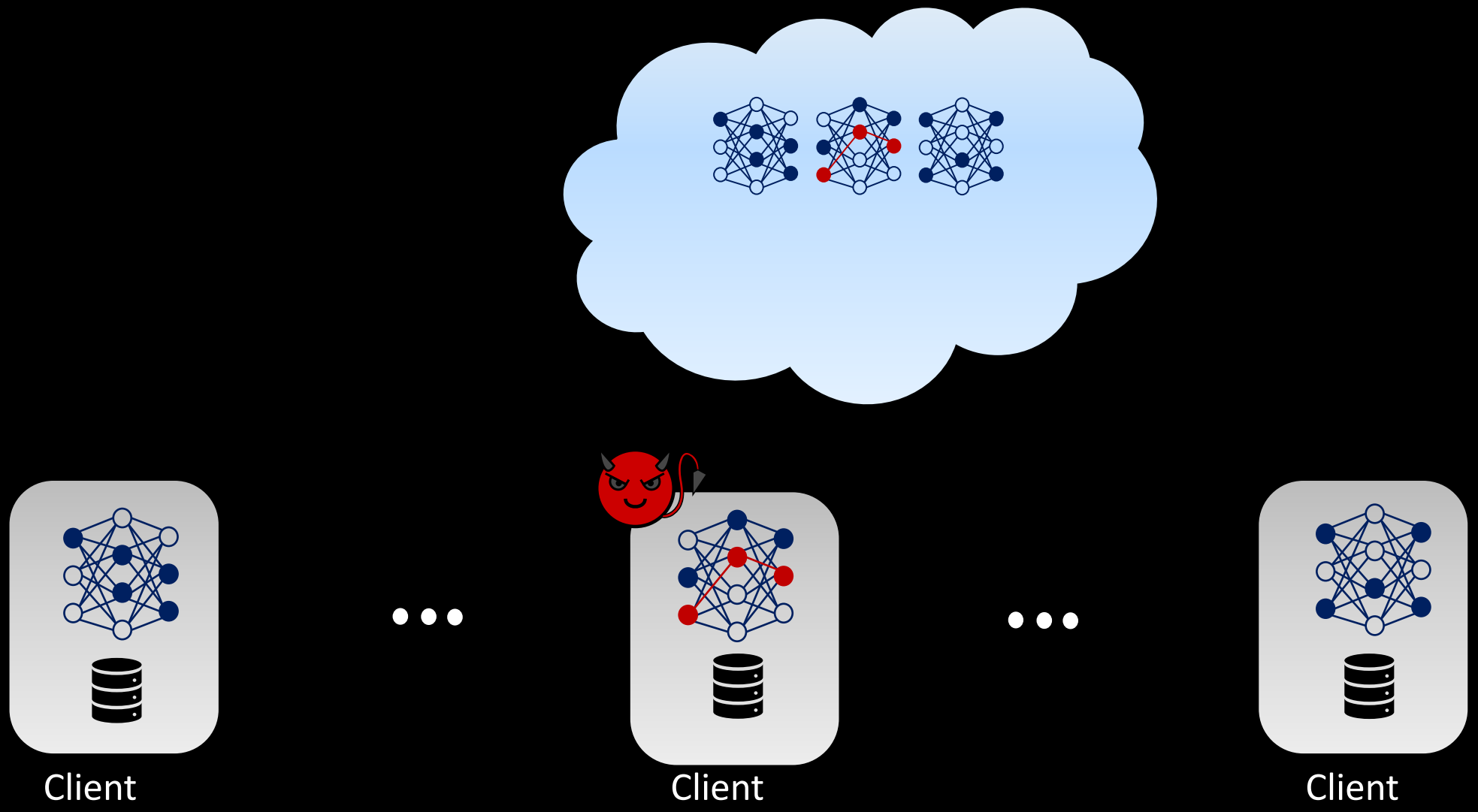
Goal: Detangle mechanisms from assumptions about attacks/data while preserving main task accuracy.

the benign
y of the model.

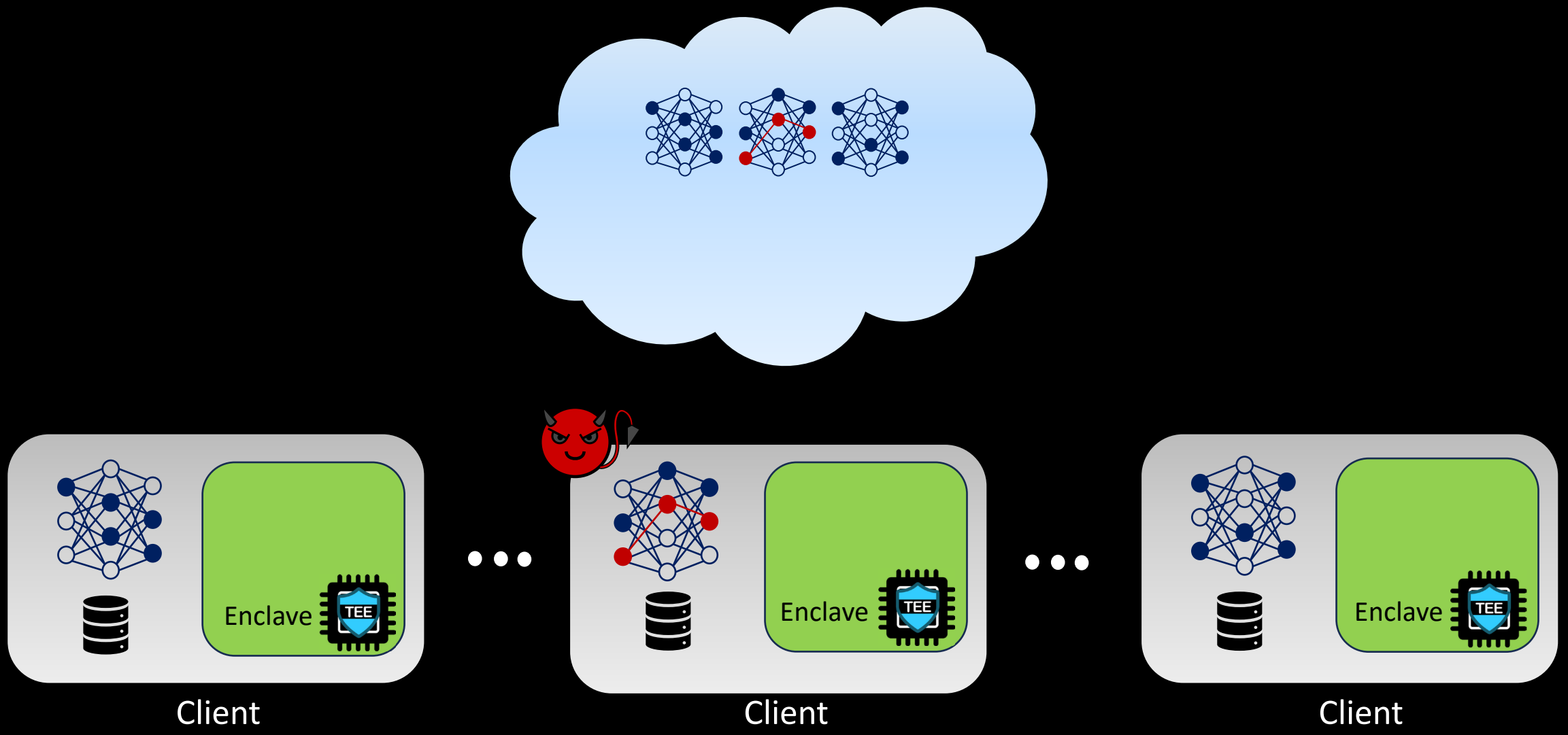
CrowdGuard – High Level Overview



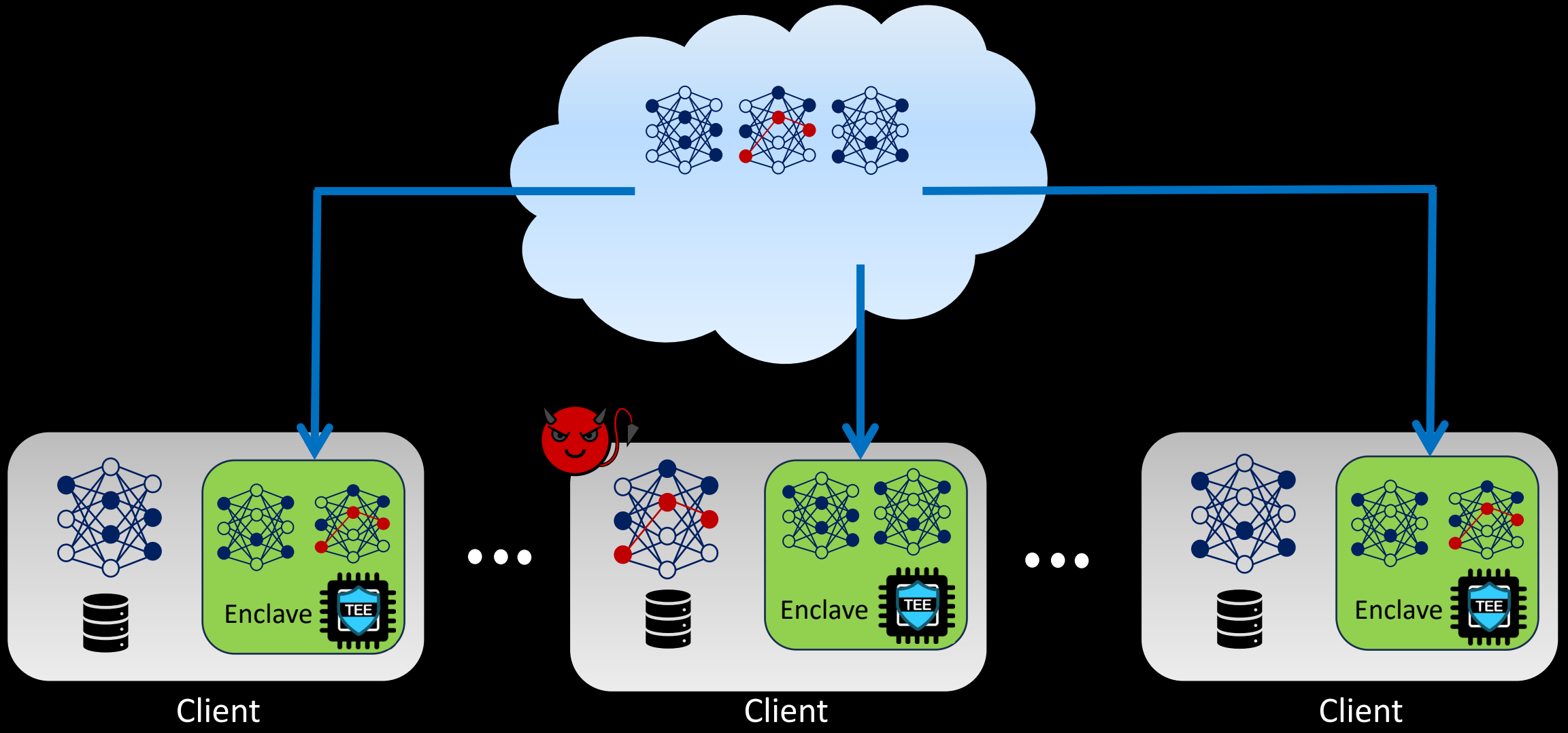
CrowdGuard – High Level Overview



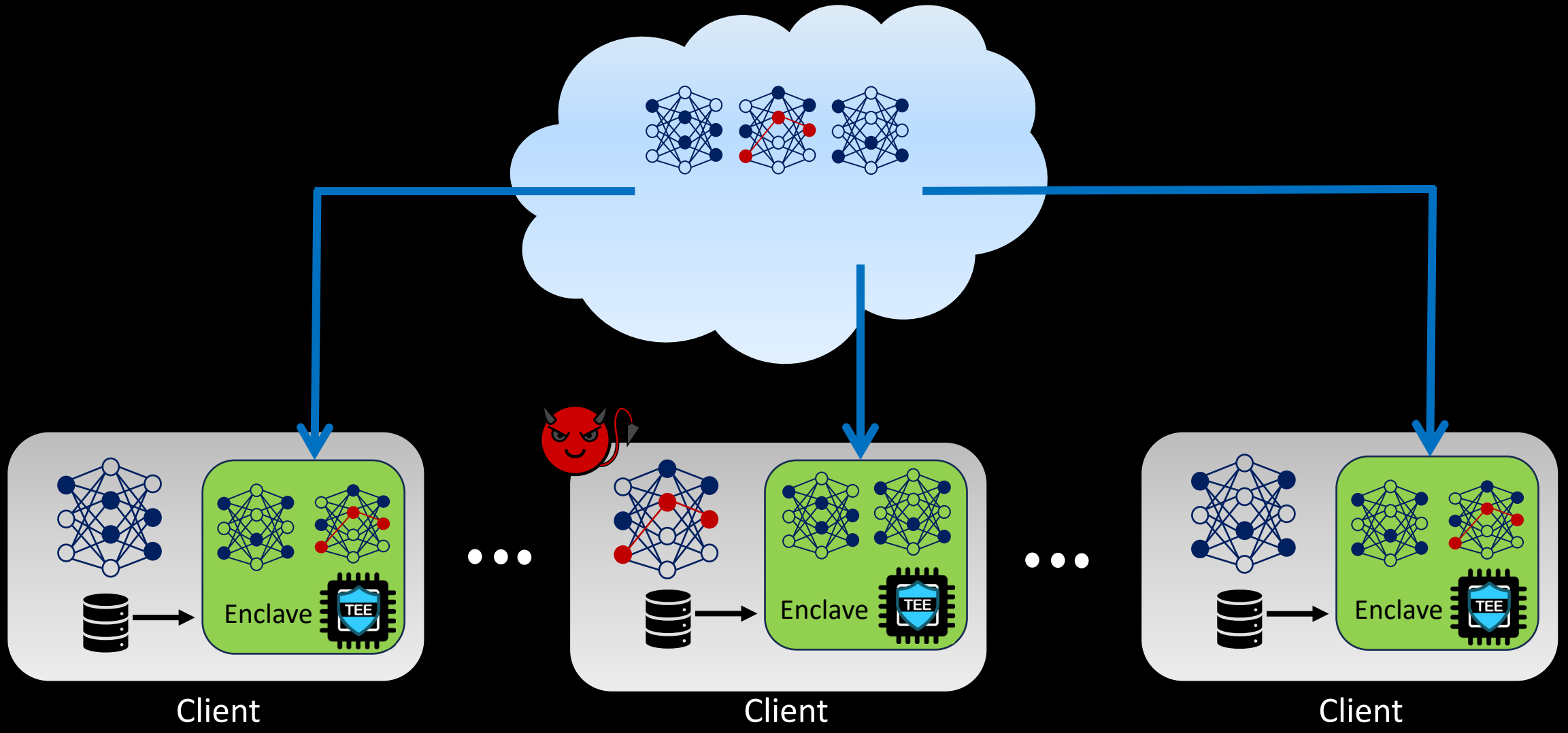
CrowdGuard – High Level Overview



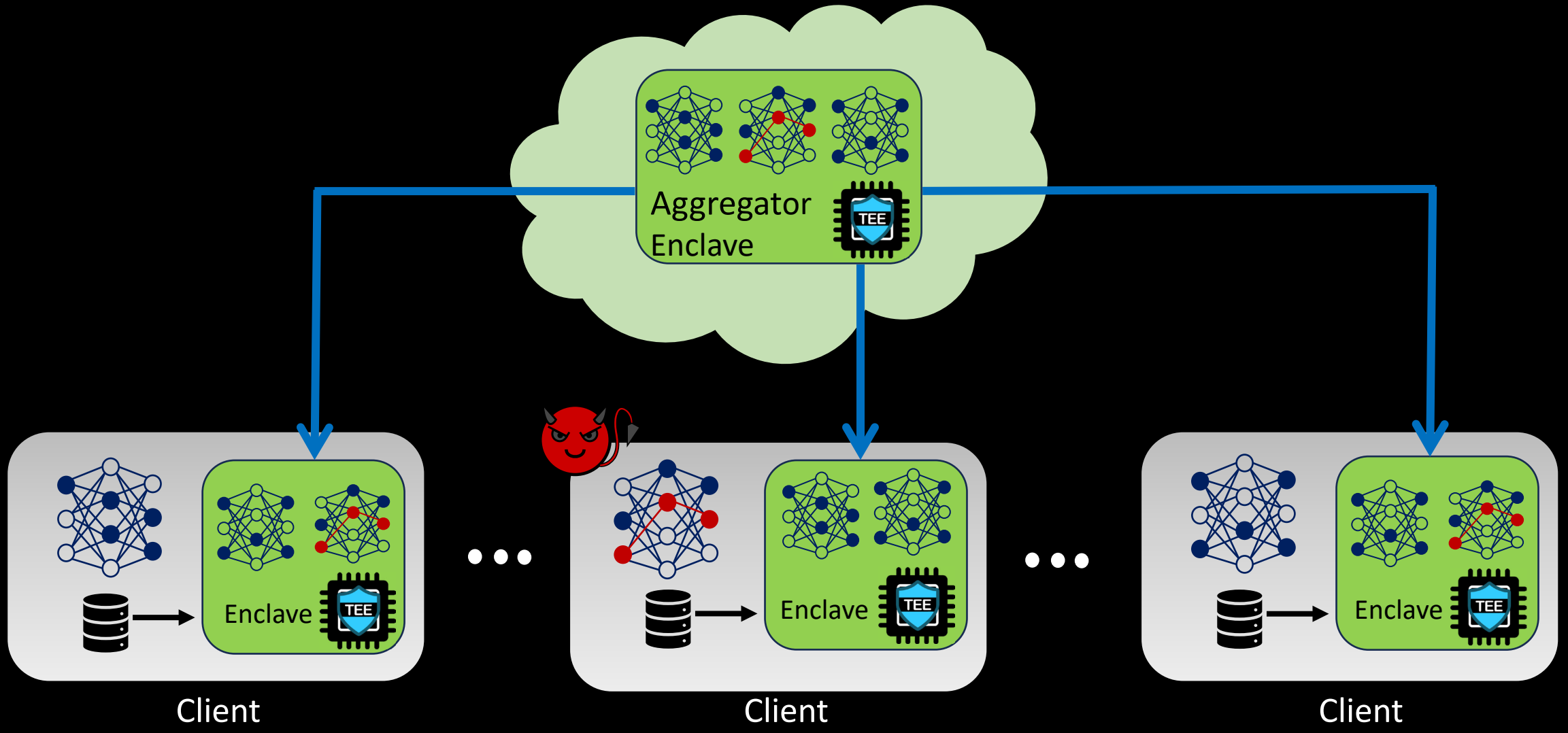
CrowdGuard – High Level Overview



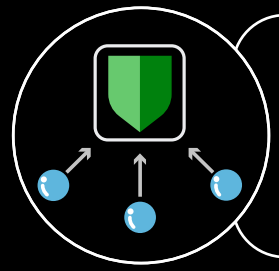
CrowdGuard – High Level Overview



CrowdGuard – High Level Overview

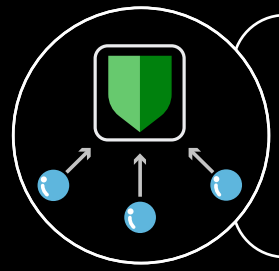


CrowdGuard – Contributions

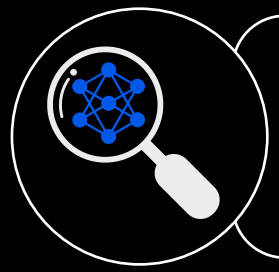


- Framework for **utilizing clients' data** for detecting poisoned models
- Trusted hardware **guarantees privacy** of data and models

CrowdGuard – Contributions

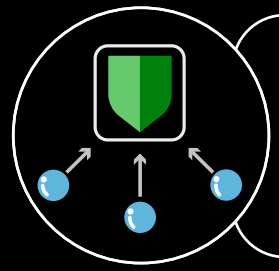


- Framework for **utilizing clients' data** for detecting poisoned models
- Trusted hardware **guarantees privacy** of data and models

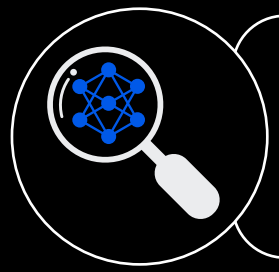


- **HLBIM metric** for analyzing changes in models' behavior
- Using **statistical tests** for indicating presence of poisoned models

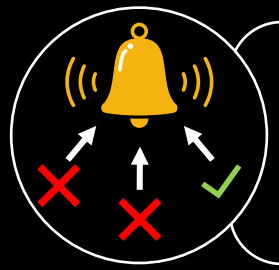
CrowdGuard – Contributions



- Framework for **utilizing clients' data** for detecting poisoned models
- Trusted hardware **guarantees privacy** of data and models

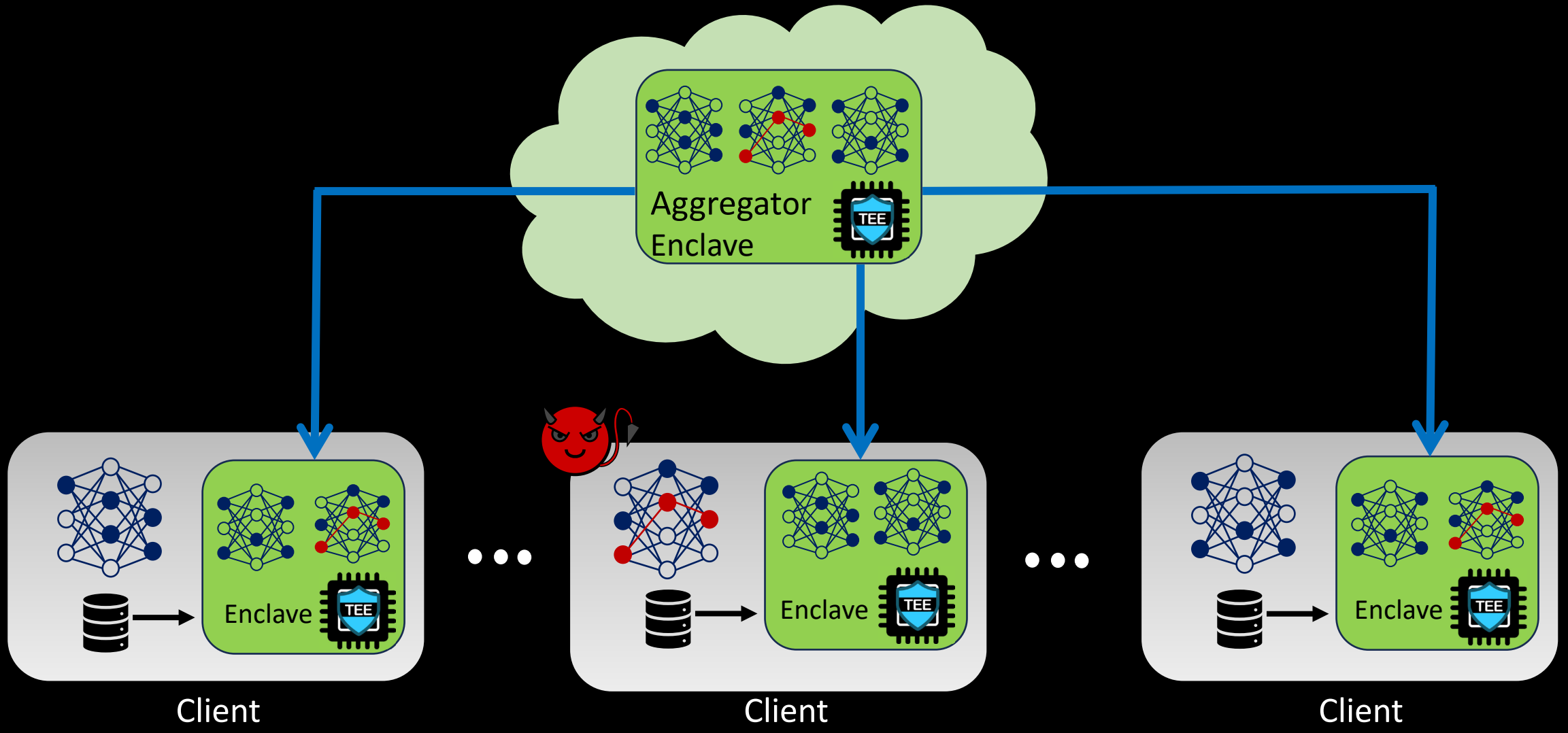


- **HLBIM metric** for analyzing changes in models' behavior
- Using **statistical tests** for indicating presence of poisoned models

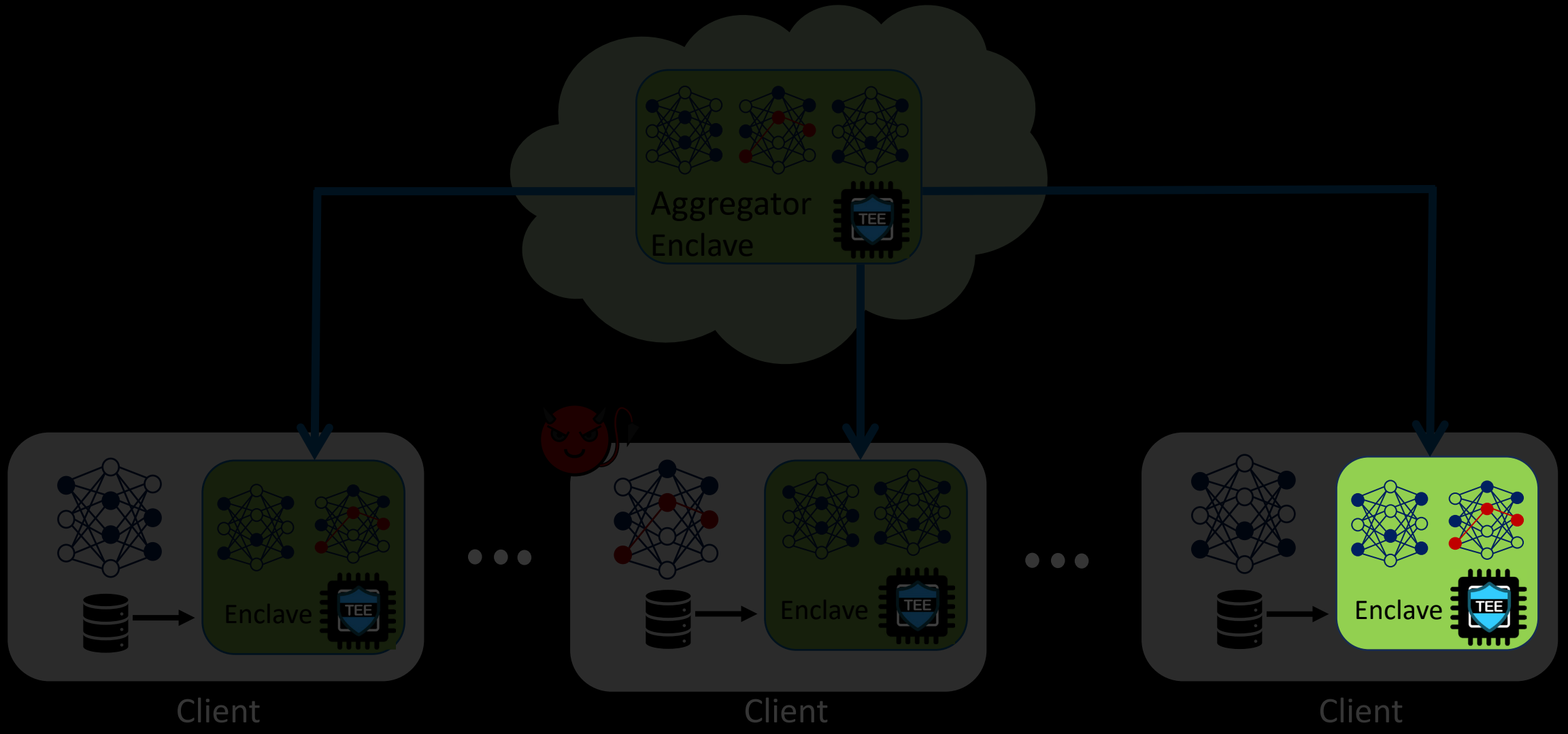


- **Multi-Layer clustering algorithm** for mitigating validation reports of malicious clients

CrowdGuard – High Level Overview



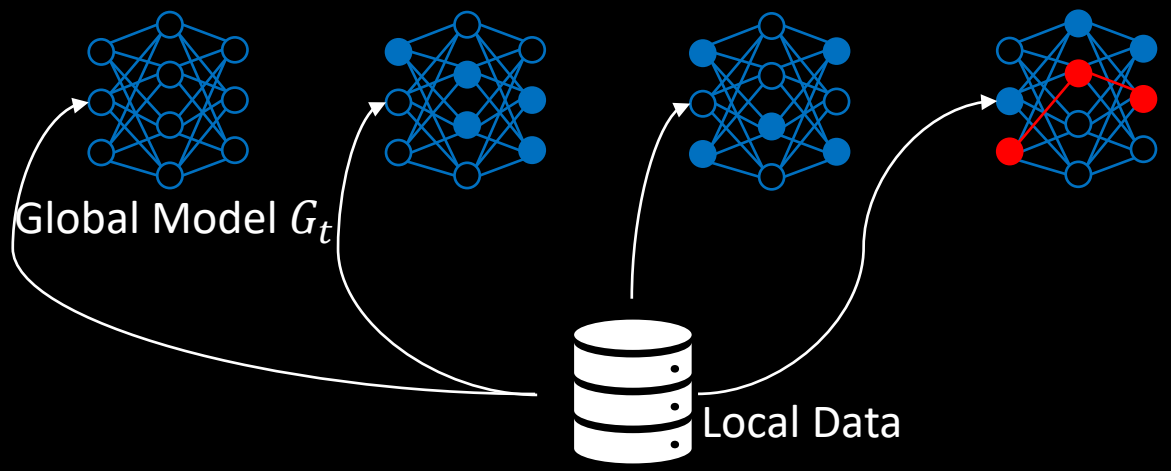
CrowdGuard – High Level Overview



Client-Side Validation – Hidden State Prediction

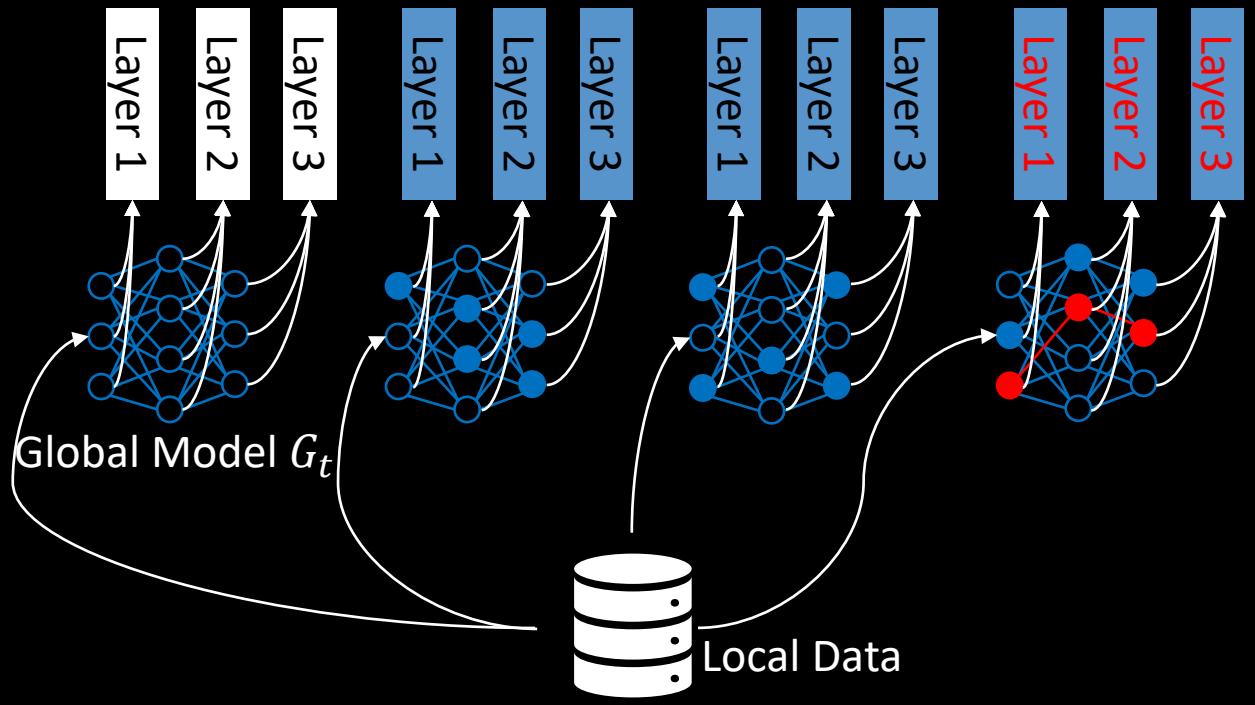
Local Data

Client-Side Validation – Hidden State Prediction



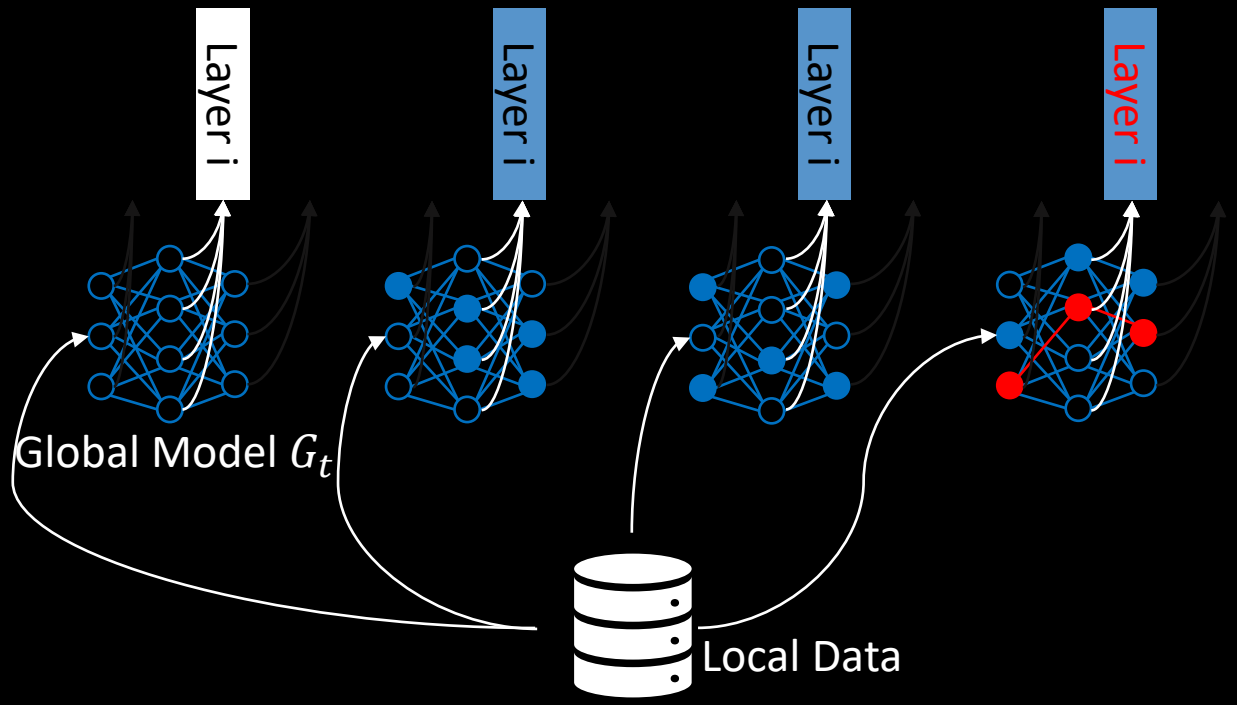
Client-Side Validation – Hidden State Prediction

1) Obtain all Layer States

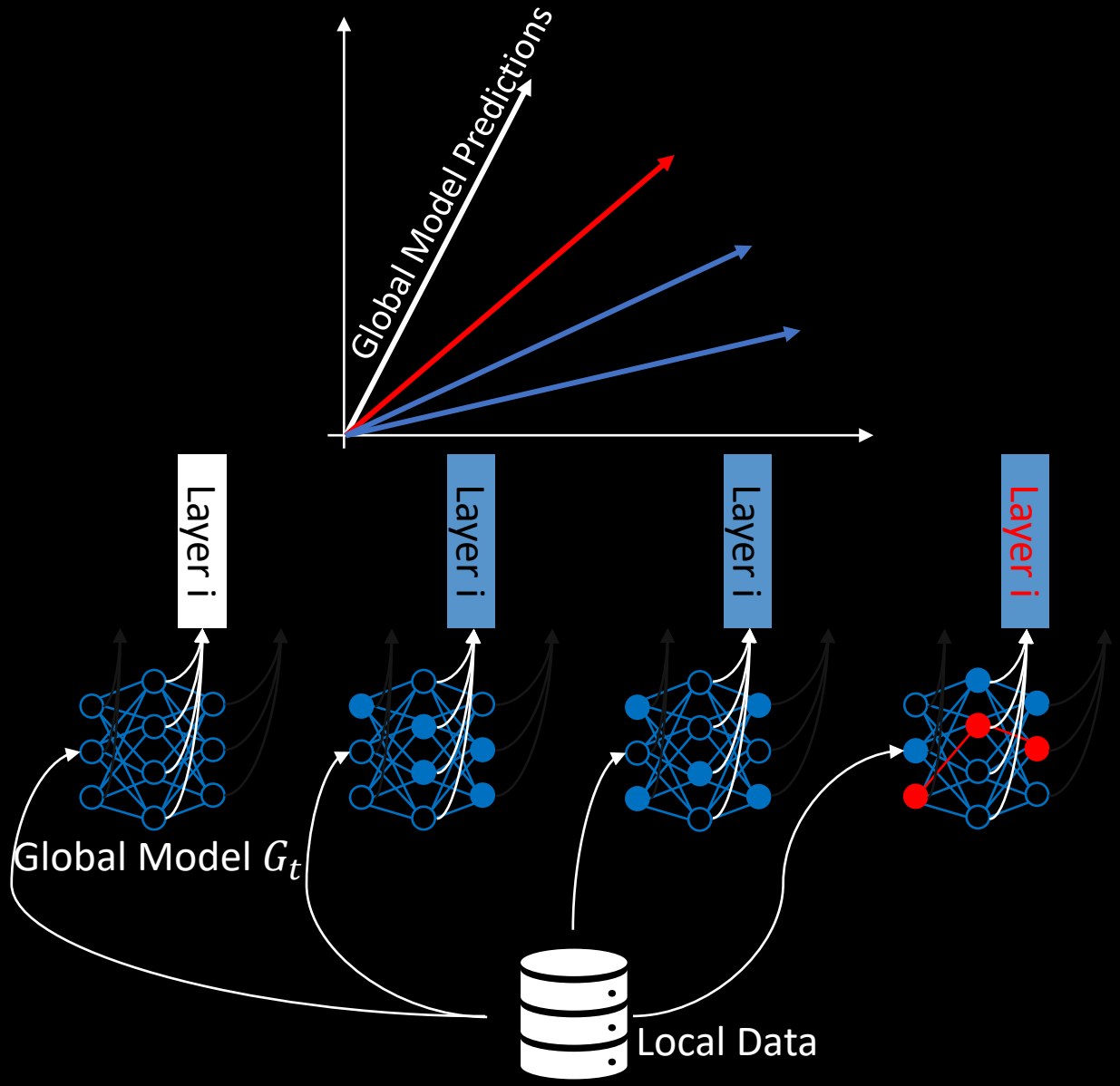


Client-Side Validation – Hidden State Prediction

1) Obtain all Layer States

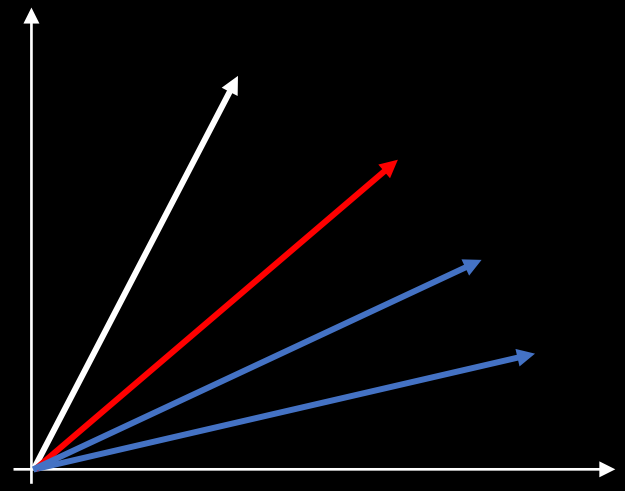


Client-Side Validation – Hidden State Prediction



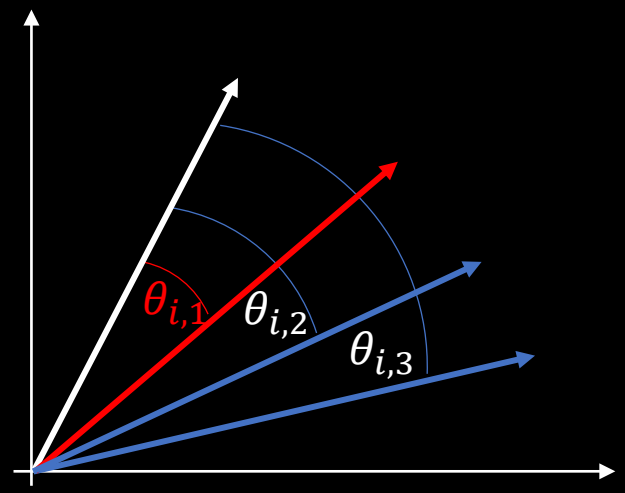
1) Obtain all Layer States

Client-Side Validation – Behavior Analysis



Client-Side Validation – Behavior Analysis

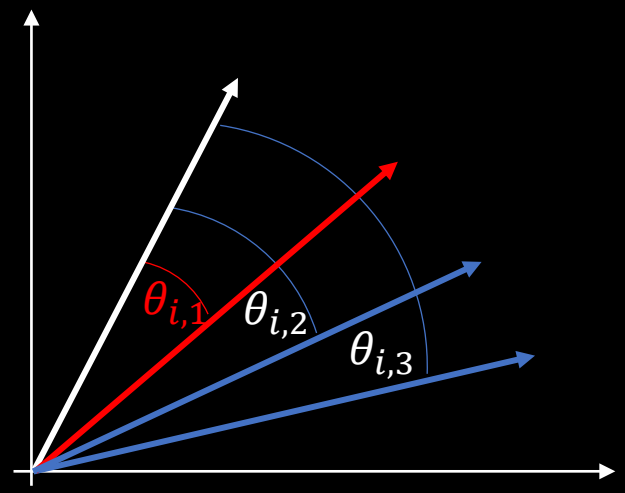
2) Calculate distance metric



Client-Side Validation – Behavior Analysis

2) Calculate distance metric

$\theta_{i,1}$ $\theta_{i,2}$ $\theta_{i,3}$

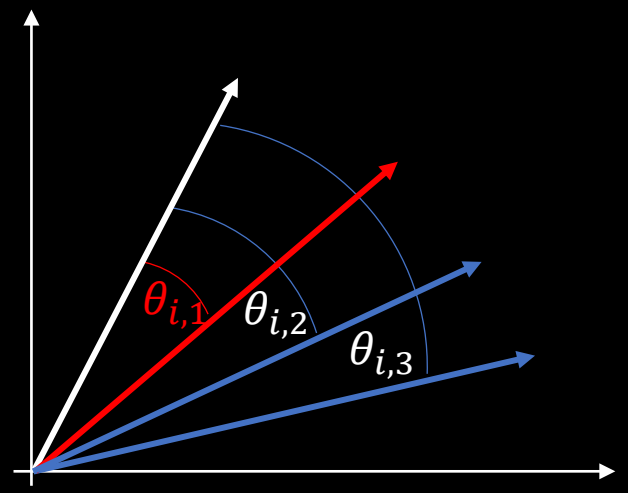


Client-Side Validation – Behavior Analysis

2) Calculate distance metric

3) Compare to own model to obtain HLBIM

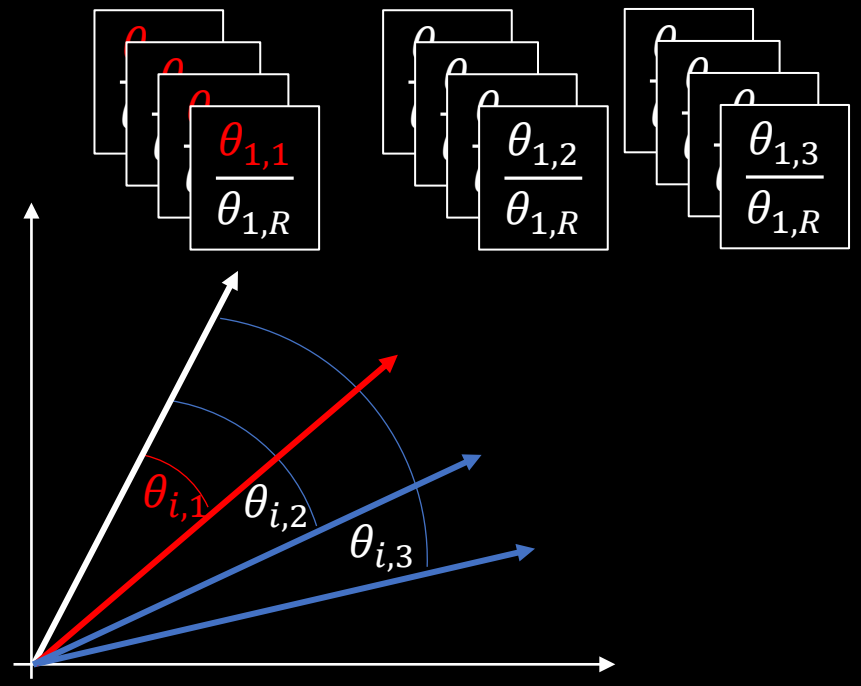
$\frac{\theta_{i,1}}{\theta_{i,R}}$ $\frac{\theta_{i,2}}{\theta_{i,R}}$ $\frac{\theta_{i,3}}{\theta_{i,R}}$



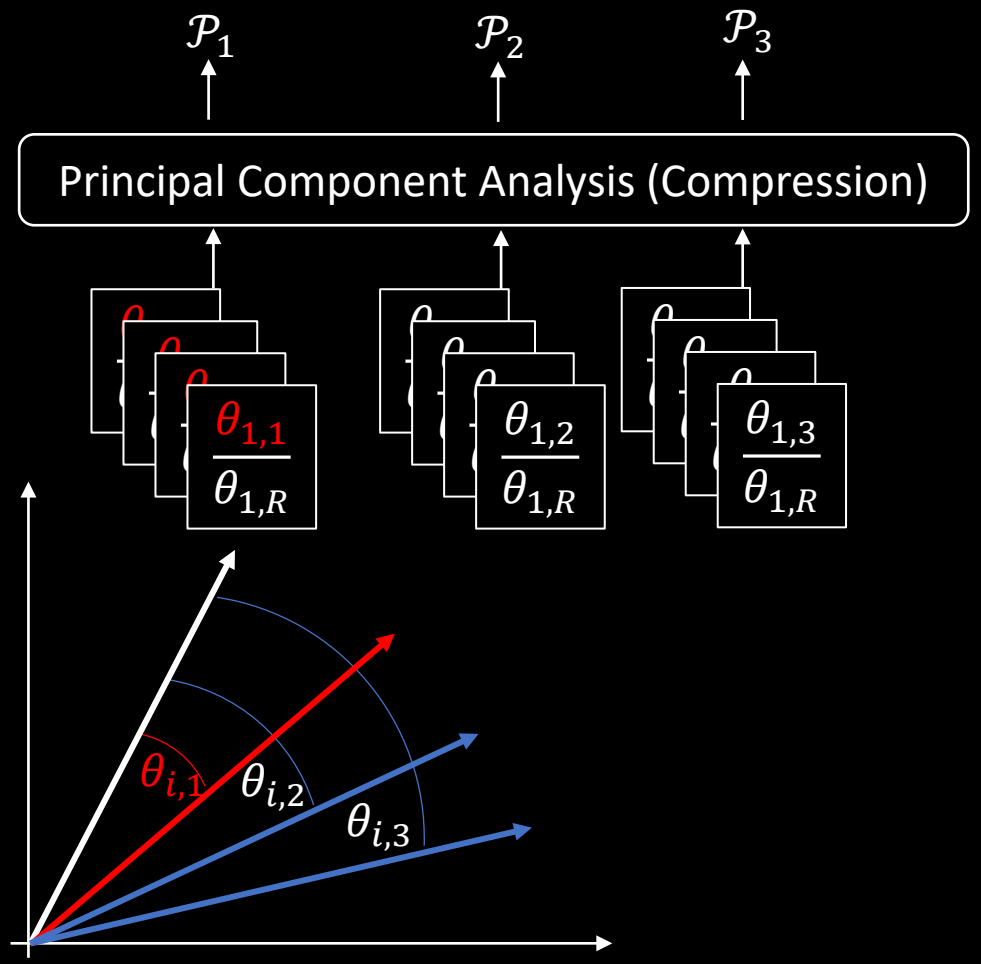
Client-Side Validation – Behavior Analysis

2) Calculate distance metric

3) Compare to own model to obtain HLBIM



Client-Side Validation – Behavior Analysis

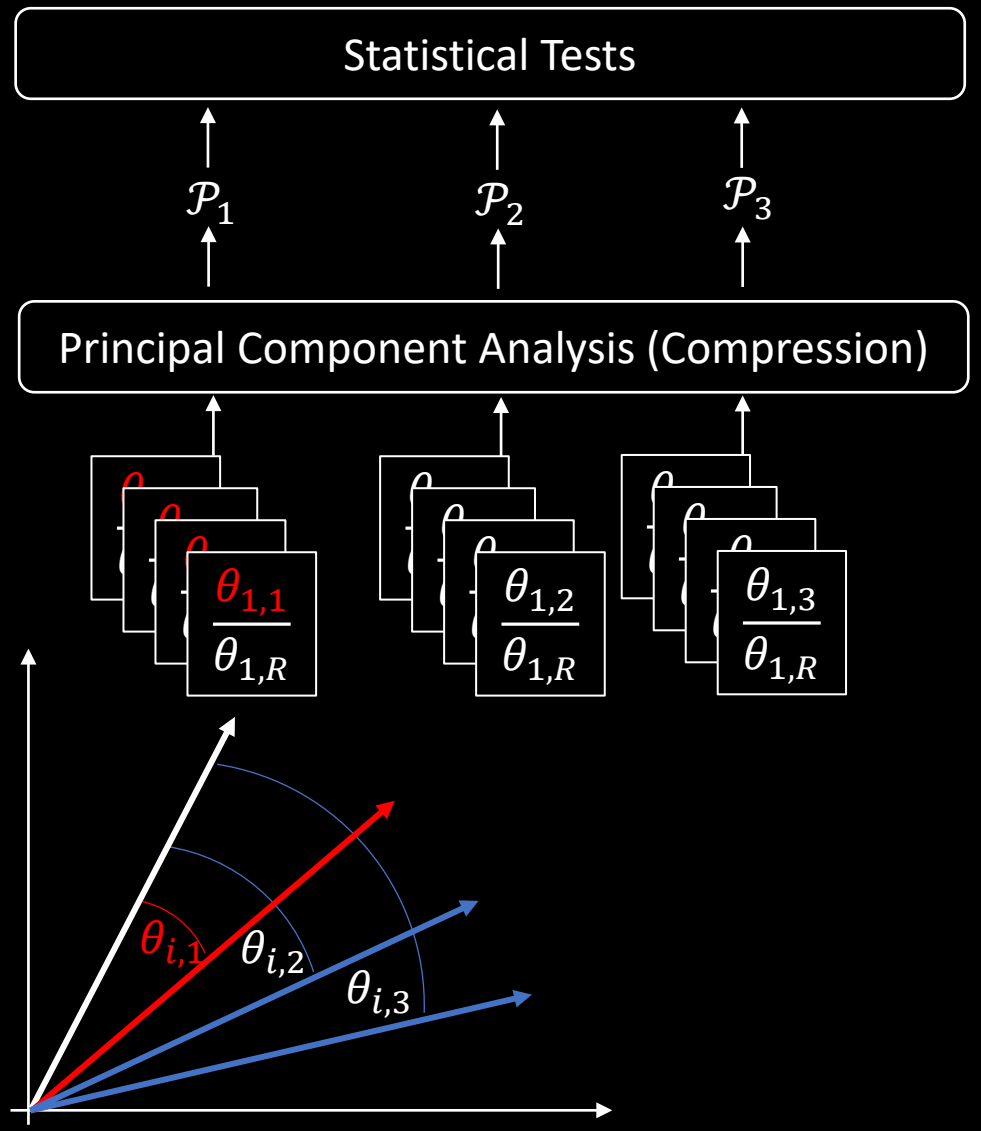


2) Calculate distance metric

3) Compare to own model to obtain HLBIM

4) Apply Principal Component Analysis

Client-Side Validation – Behavior Analysis



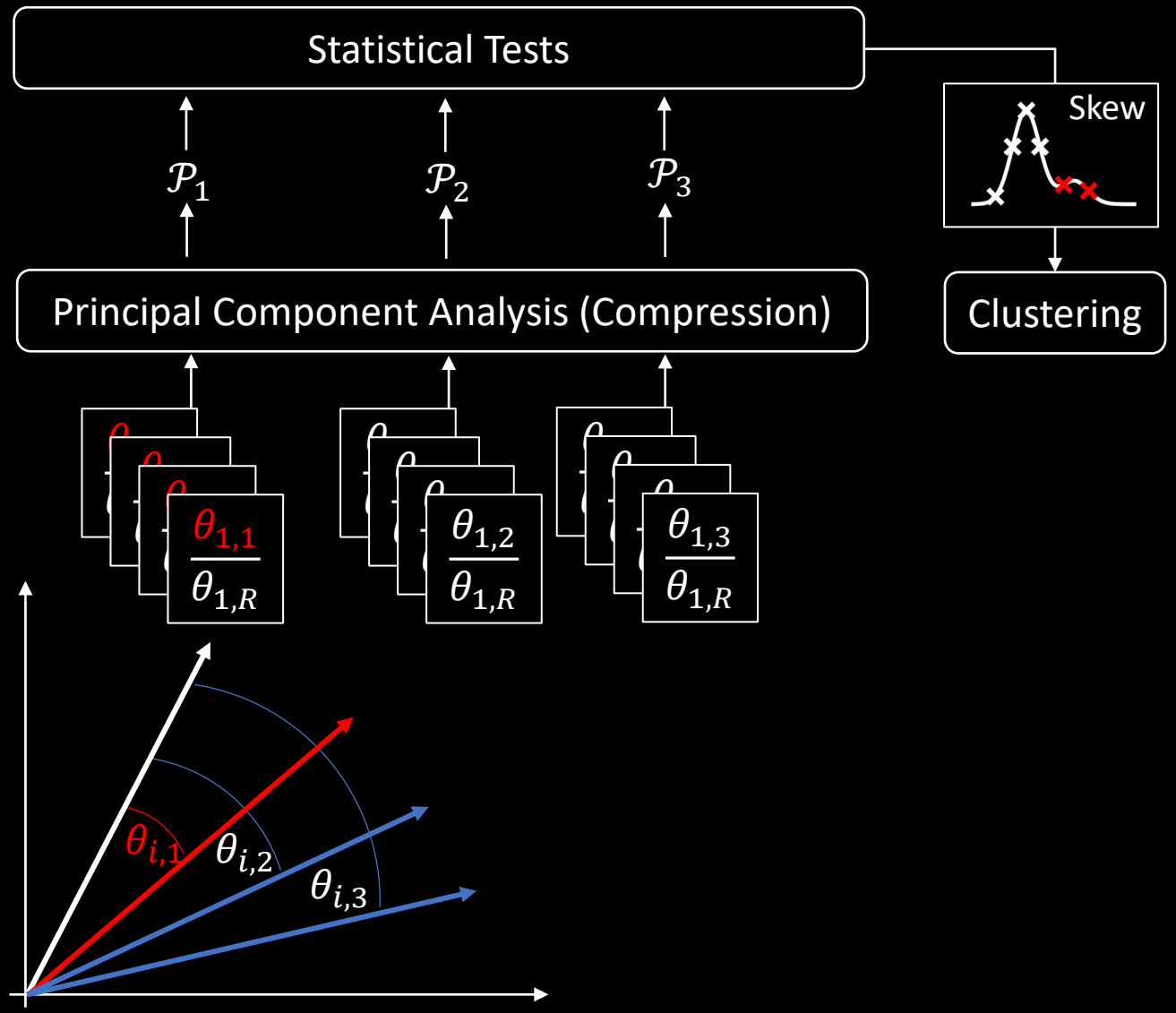
2) Calculate distance metric

3) Compare to own model to obtain HLBIM

4) Apply Principal Component Analysis

5) Check for poisoned models using statistical tests

Client-Side Validation – Behavior Analysis



2) Calculate distance metric

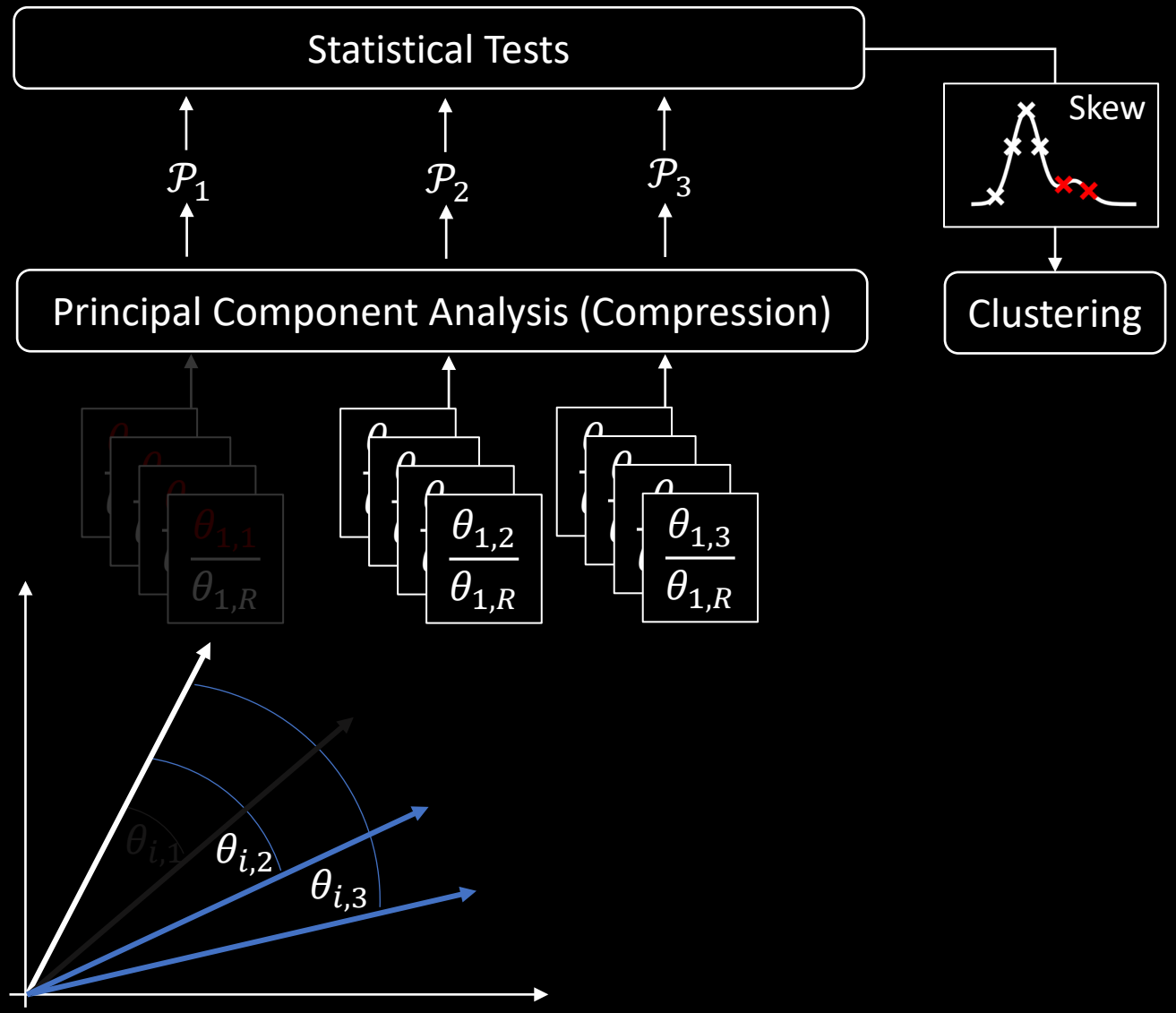
3) Compare to own model to obtain HLBIM

4) Apply Principal Component Analysis

5) Check for poisoned models using statistical tests

6) Clustering

Client-Side Validation – Behavior Analysis



2) Calculate distance metric

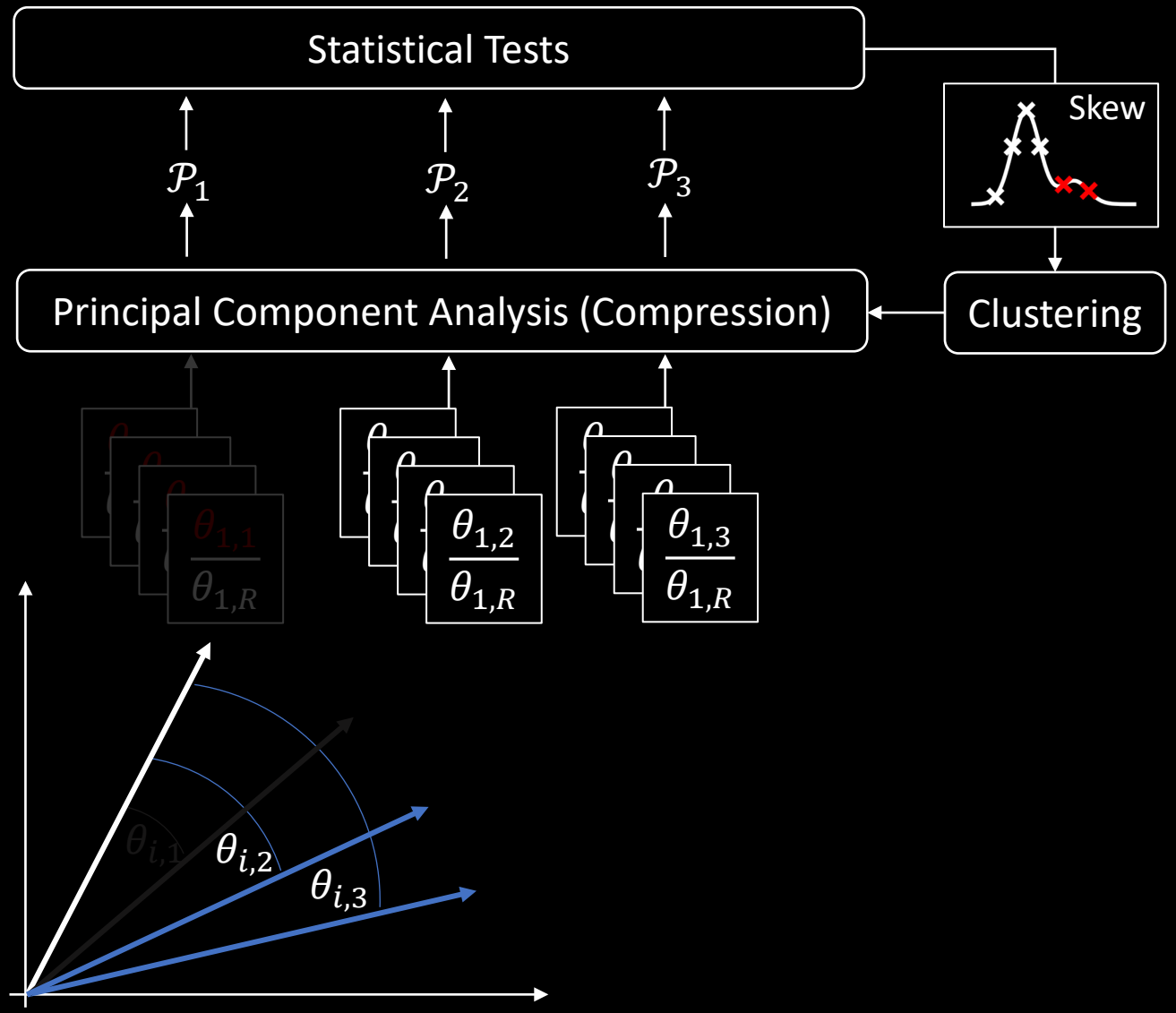
3) Compare to own model to obtain HLBIM

4) Apply Principal Component Analysis

5) Check for poisoned models using statistical tests

6) Clustering

Client-Side Validation – Behavior Analysis



2) Calculate distance metric

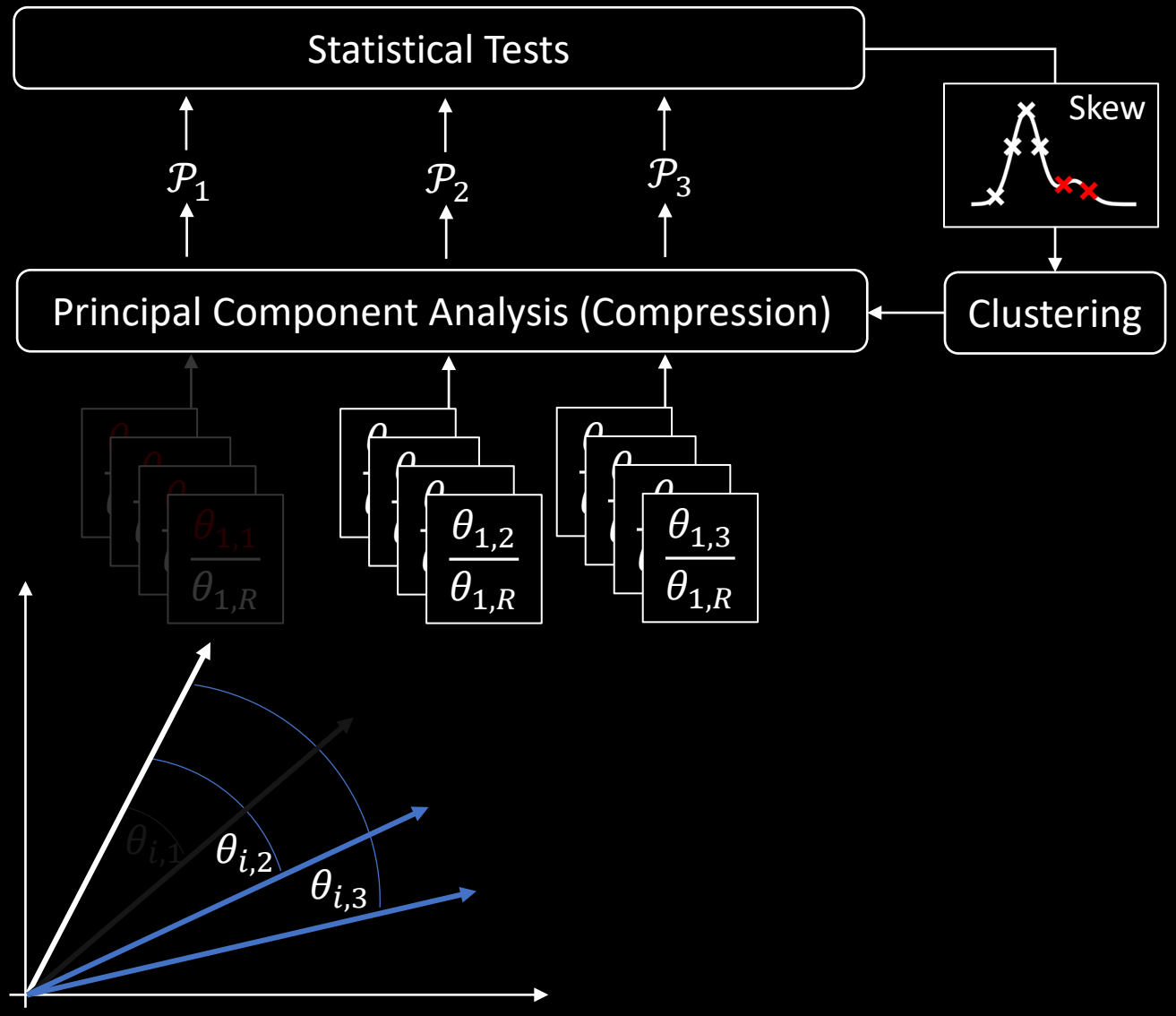
3) Compare to own model to obtain HLBIM

4) Apply Principal Component Analysis

5) Check for poisoned models using statistical tests

6) Clustering

Client-Side Validation – Behavior Analysis



2) Calculate distance metric

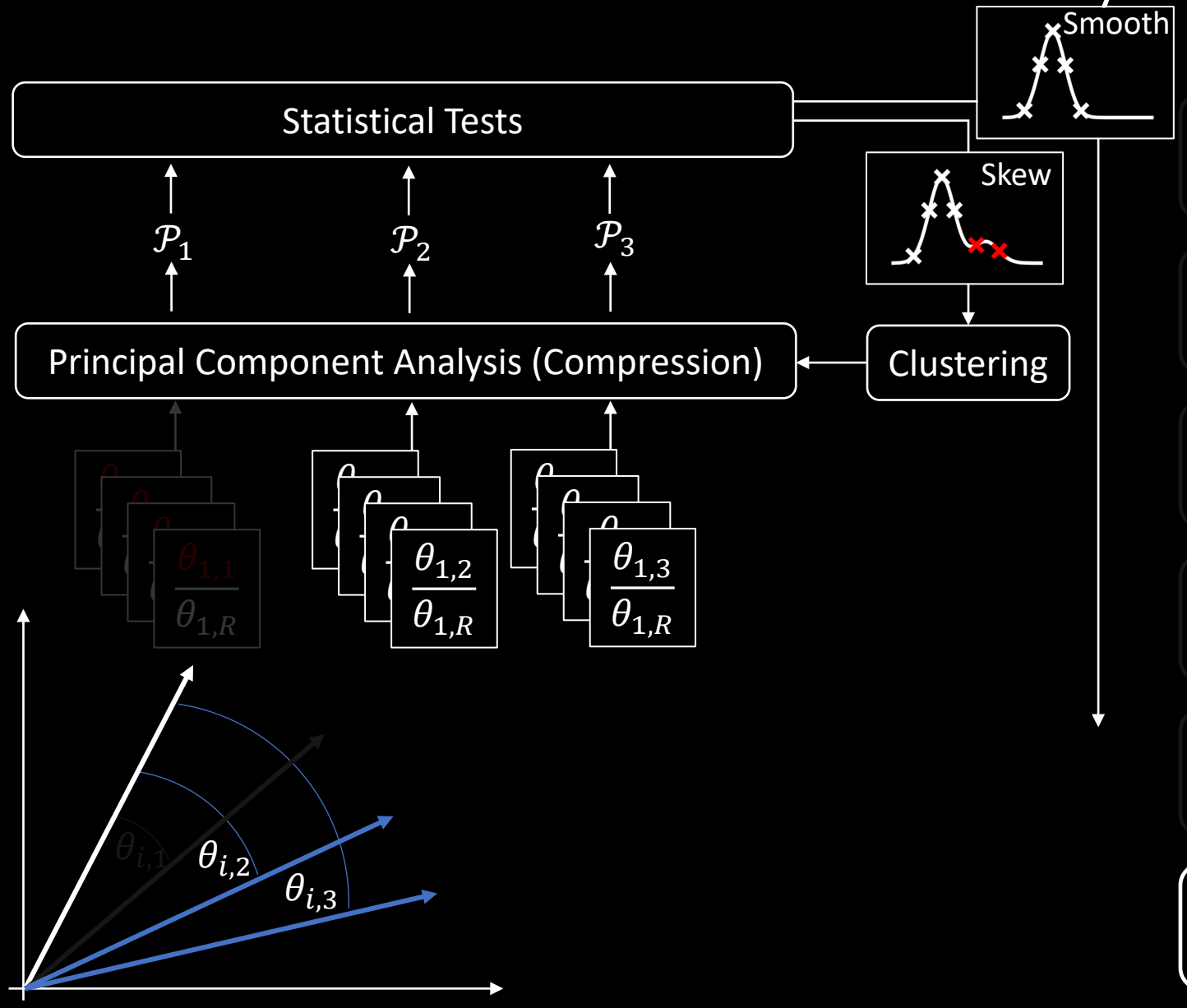
3) Compare to own model to obtain HLBIM

4) Apply Principal Component Analysis

5) Check for poisoned models using statistical tests

6) Clustering

Client-Side Validation – Behavior Analysis



2) Calculate distance metric

3) Compare to own model to obtain HLBIM

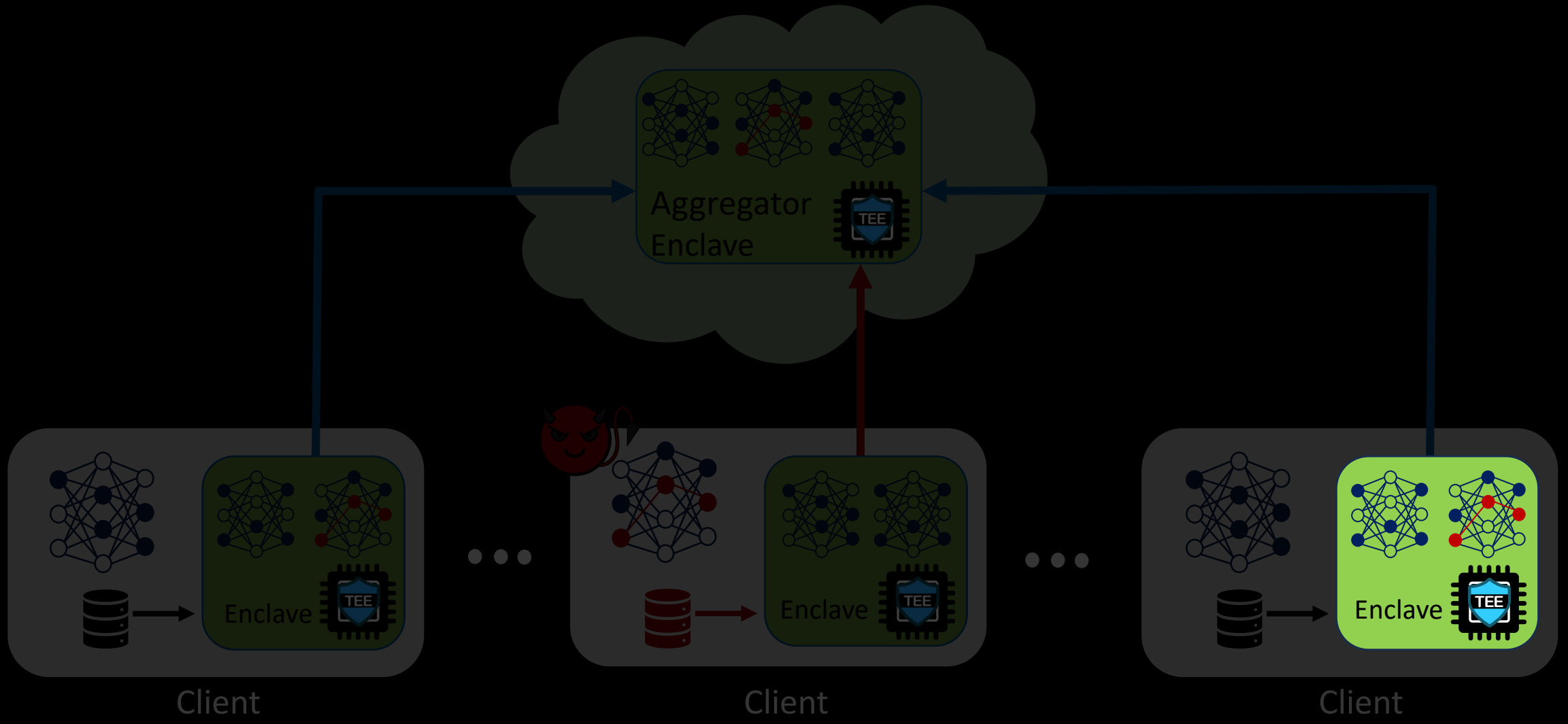
4) Apply Principal Component Analysis

5) Check for poisoned models using statistical tests

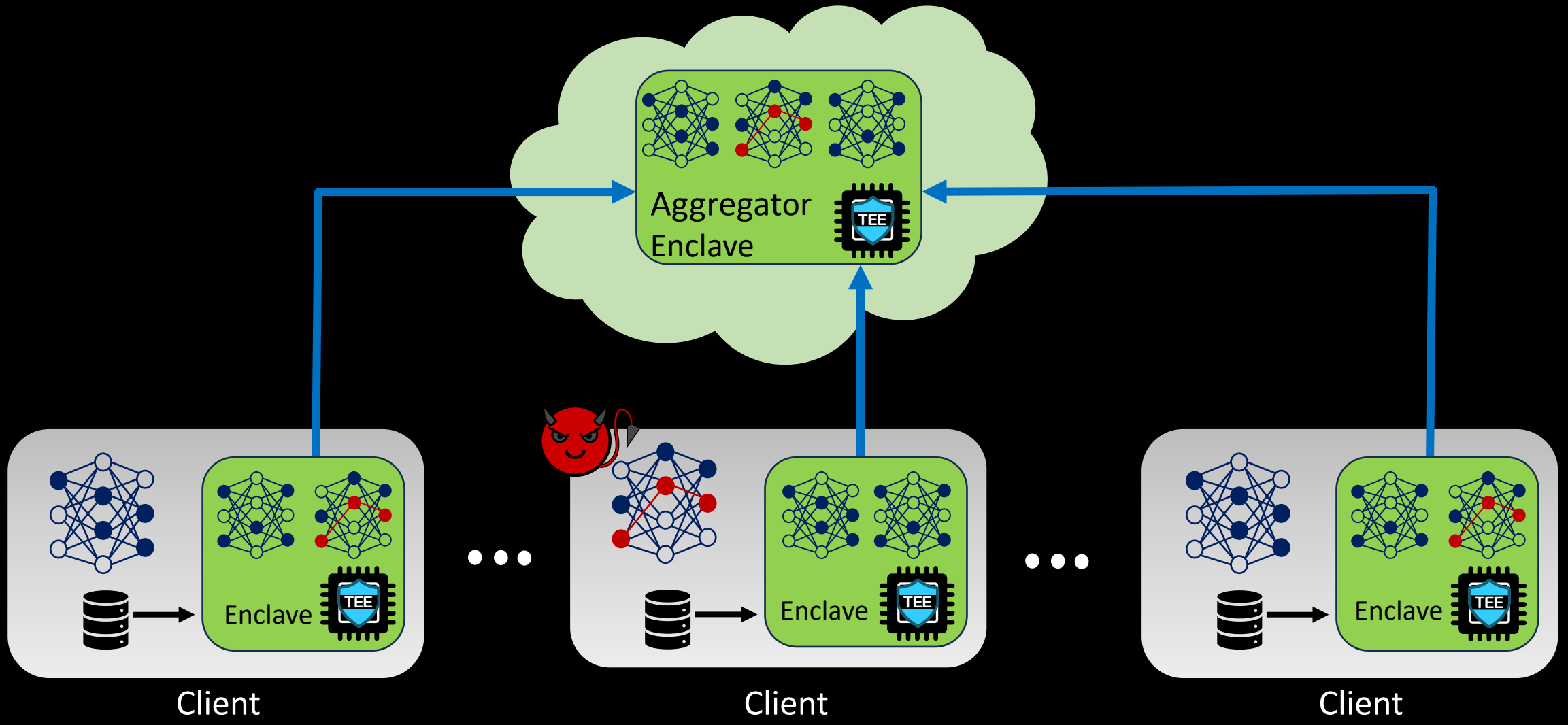
6) Clustering

7) Report suspicious models to server

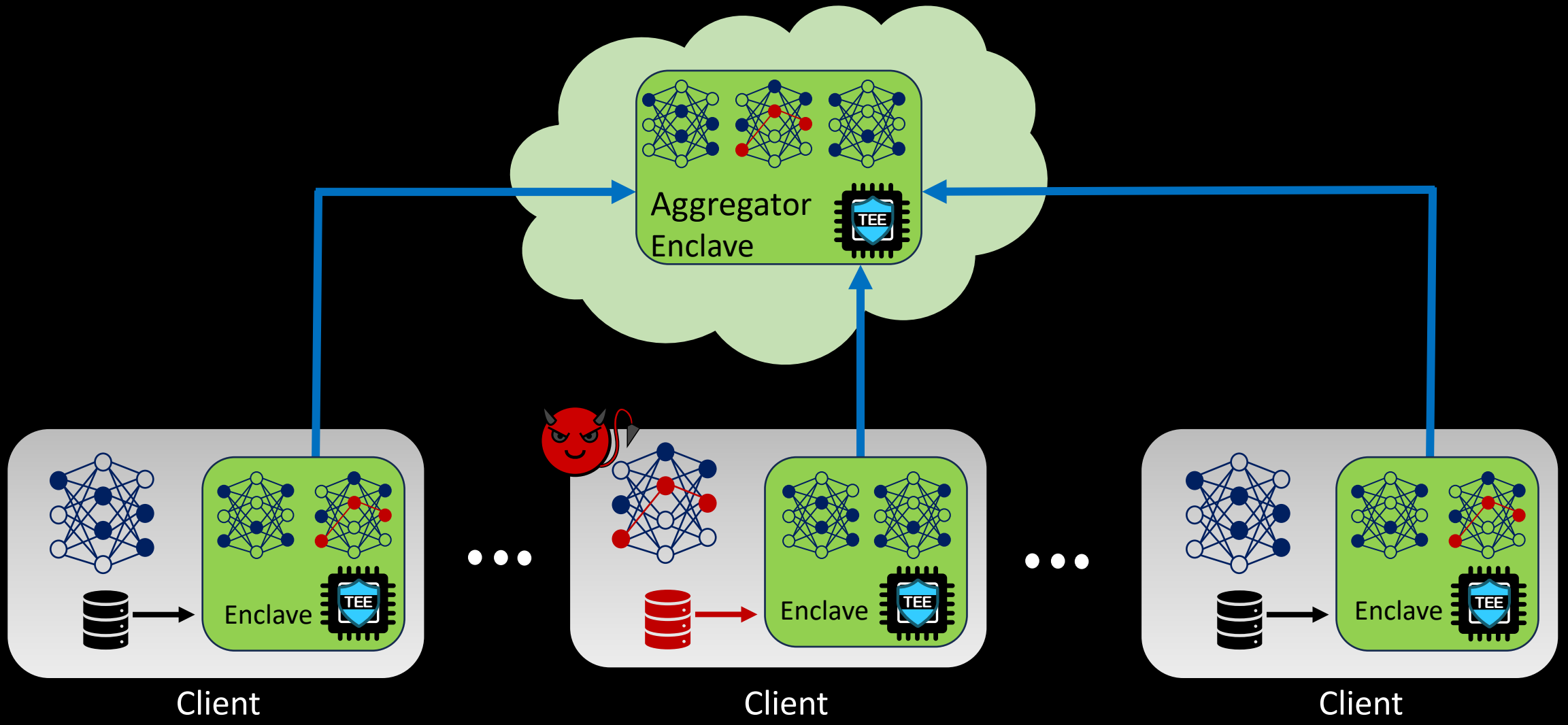
CrowdGuard – High Level Overview



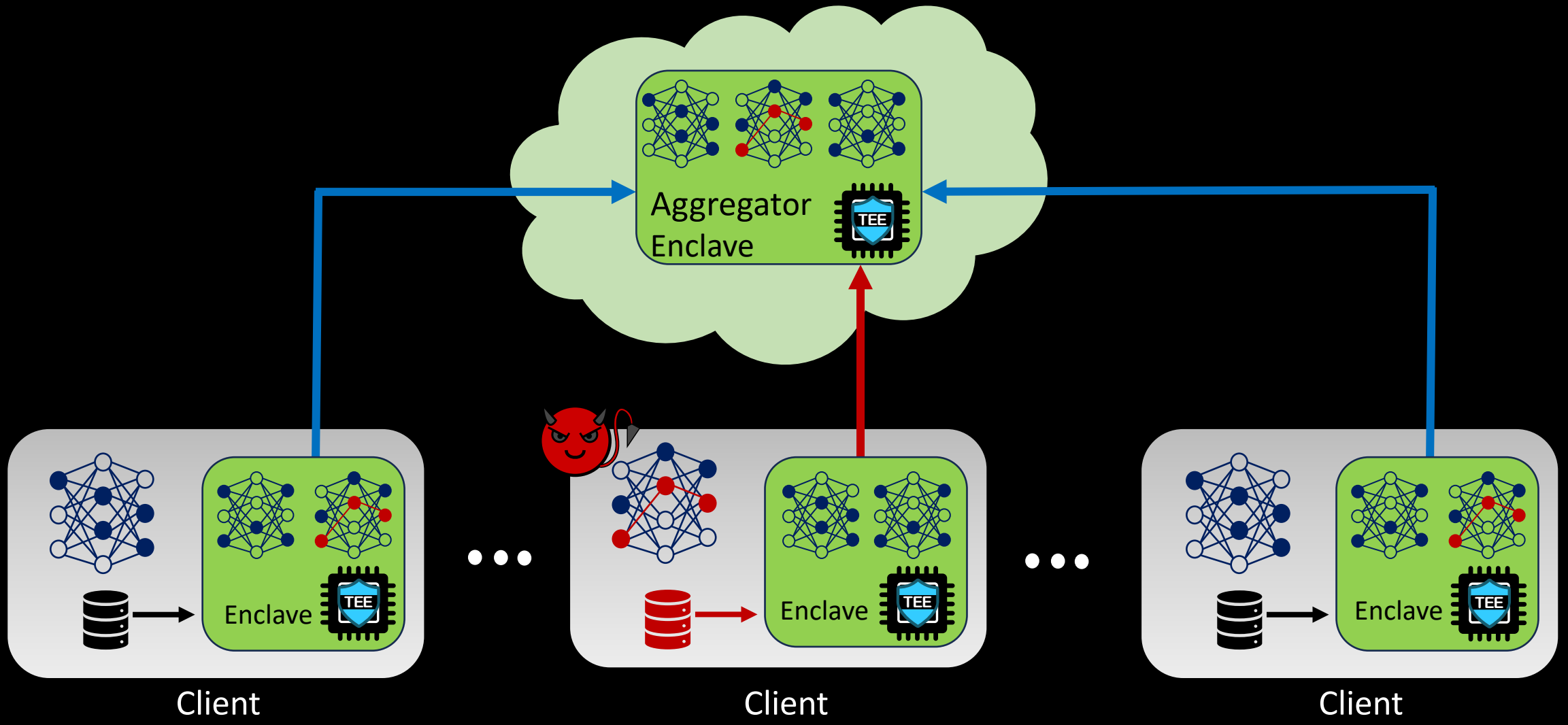
CrowdGuard – High Level Overview



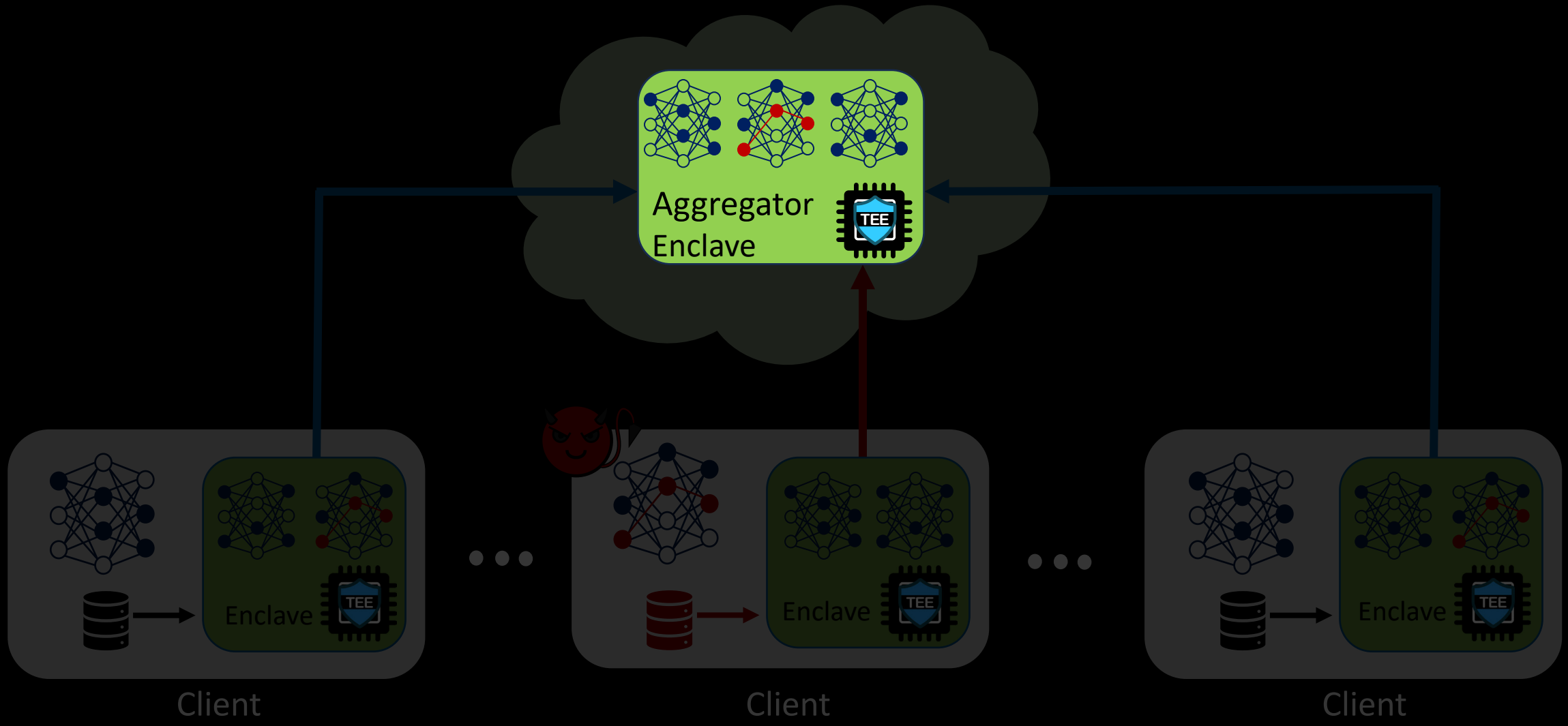
CrowdGuard – High Level Overview



CrowdGuard – High Level Overview



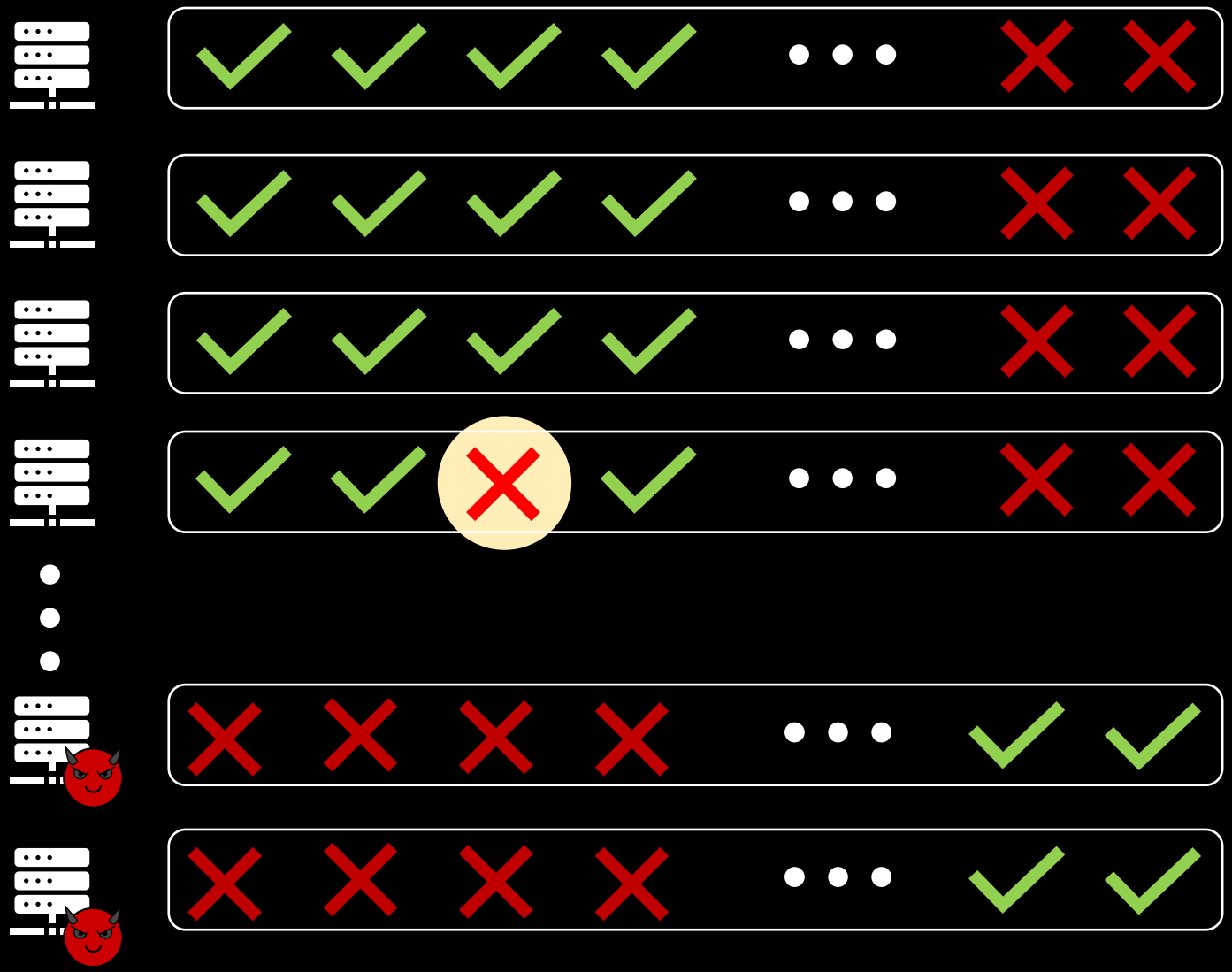
CrowdGuard – High Level Overview



Server Side Stacked Clustering

1) Receive Votes

Server Side Stacked Clustering



1) Receive Votes

Server Side Stacked Clustering

1) Receive Votes

2) First Clustering

Server Side Stacked Clustering

1) Receive Votes

2) First Clustering

3) Final Clustering

Server Side Stacked Clustering

1) Receive Votes

2) First Clustering

3) Final Clustering

4) Aggregate Accepted Models

Evaluation Overview

Data Distribution	TPR	TNR
CIFAR-10 – 1-class non-IID rates	100.0%	100.0%
CIFAR-10 – 1-class non-IID rates	100.0%	100.0%
CIFAR-10 – Dirichlet Distribution	100.0%	100.0%
CIFAR-10 – Normal	100.0%	100.0%
MNIST – 1-class non-IID rates	100.0%	100.0%

TPR: True-Positive-Rate
TNR: True-Negative-Rate

Evaluation Overview

Data Distribution	TPR	TNR
CIFAR-10 – 1-class non-IID rates	100.0%	100.0%
CIFAR-10 – 1-class non-IID rates	100.0%	100.0%
CIFAR-10 – Dirichlet Distribution	100.0%	100.0%
CIFAR-10 – Normal	100.0%	100.0%
MNIST – 1-class non-IID rates	100.0%	100.0%

Varied Attack Parameter	TPR	TNR
PMR $\in \{0.05, 0.1, \ldots, 0.45\}$	100.0%	100.0%
$\alpha \in \{0.1, \ldots, 0.9\}$	100.0%	100.0%
PDR $\in \{0.1, \ldots, 0.9\}$	100.0%	100.0%
LR $\in \{0.01, 0.001\}$	100.0%	100.0%

TPR: True-Positive-Rate
TNR: True-Negative-Rate

Evaluation Overview

Data Distribution	TPR	TNR
CIFAR-10 – 1-class non-IID rates	100.0%	100.0%
CIFAR-10 – 1-class non-IID rates	100.0%	100.0%
CIFAR-10 – Dirichlet Distribution	100.0%	100.0%
CIFAR-10 – Normal	100.0%	100.0%
MNIST – 1-class non-IID rates	100.0%	100.0%

Backdoor Type	TPR	TNR
Pixel-Trigger	100.0%	100.0%
Label Swap Backdoor	100.0%	100.0%
Semantic Trigger	100.0%	100.0%
Multi-Backdoor Attack	100.0%	100.0%

Varied Attack Parameter	TPR	TNR
PMR $\in \{0.05, 0.1, \ldots, 0.45\}$	100.0%	100.0%
$\alpha \in \{0.1, \ldots, 0.9\}$	100.0%	100.0%
PDR $\in \{0.1, \ldots, 0.9\}$	100.0%	100.0%
LR $\in \{0.01, 0.001\}$	100.0%	100.0%

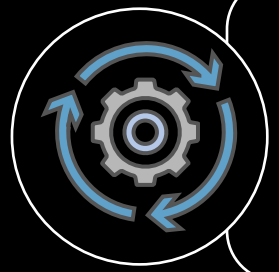
TPR: True-Positive-Rate
TNR: True-Negative-Rate

Conclusion

- Federated Learning allows joint DNN training without sharing data
- Distributed setting allows malicious clients injecting backdoors
- Existing defenses make strong assumptions on data scenario or adversaries

Conclusion

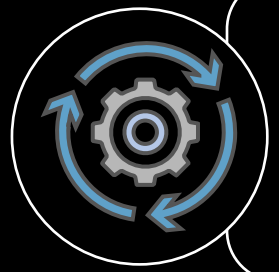
- Federated Learning allows joint DNN training without sharing data
- Distributed setting allows malicious clients injecting backdoors
- Existing defenses make strong assumptions on data scenario or adversaries



- Client-Side analysis inside TEEs using local data
- HLBIM metric measures changes of updates to detect poisoned models
- Server-Side algorithm for robust vote aggregation

Conclusion

- Federated Learning allows joint DNN training without sharing data
- Distributed setting allows malicious clients injecting backdoors
- Existing defenses make strong assumptions on data scenario or adversaries



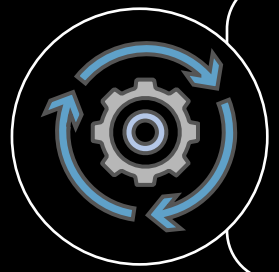
- Client-Side analysis inside TEEs using local data
- HLBIM metric measures changes of updates to detect poisoned models
- Server-Side algorithm for robust vote aggregation

- Effectively mitigates backdoor attacks, even in non-IID scenarios
- Behavior analysis prevents filtering of benign models
- Trusted hardware prevents privacy attacks

Conclusion

Integrated in Intel's Framework OpenFL

- Federated Learning allows joint DNN training without sharing data
- Distributed setting allows malicious clients injecting backdoors
- Existing defenses make strong assumptions on data scenario or adversaries



- Client-Side analysis inside TEEs using local data
- HLBIM metric measures changes of updates to detect poisoned models
- Server-Side algorithm for robust vote aggregation

- Effectively mitigates backdoor attacks, even in non-IID scenarios
- Behavior analysis prevents filtering of benign models
- Trusted hardware prevents privacy attacks

Additional information

Evaluation Results – Comparison Against SotA

Approach	BA	MA	TPR	TNR	PRC
No Attack	0.0%	62.0%	-	-	-
No Defense	80.0%	61.5%	-	-	-
Differential Privacy	80.0%	50.6%	-	-	-
Zhao et al.	100.0%	61.2%	-	-	-
Median	0.0%	10.0%	-	-	-
FoolsGold	0.0%	10.0%	100.0%	9.0%	47.4%
Krum	100.0%	63.8%	88.9%	0.0%	42.1%
Auror	80.0%	68.4%	0.0%	100.0%	-
CrowdGuard	0.0%	62.0%	100.0%	100.0%	100.0%

BA: Backdoor Accuracy
MA: Main Task Accuracy
TPR: True-Positive-Rate
TNR: True-Negative-Rate
PRC: Precision