



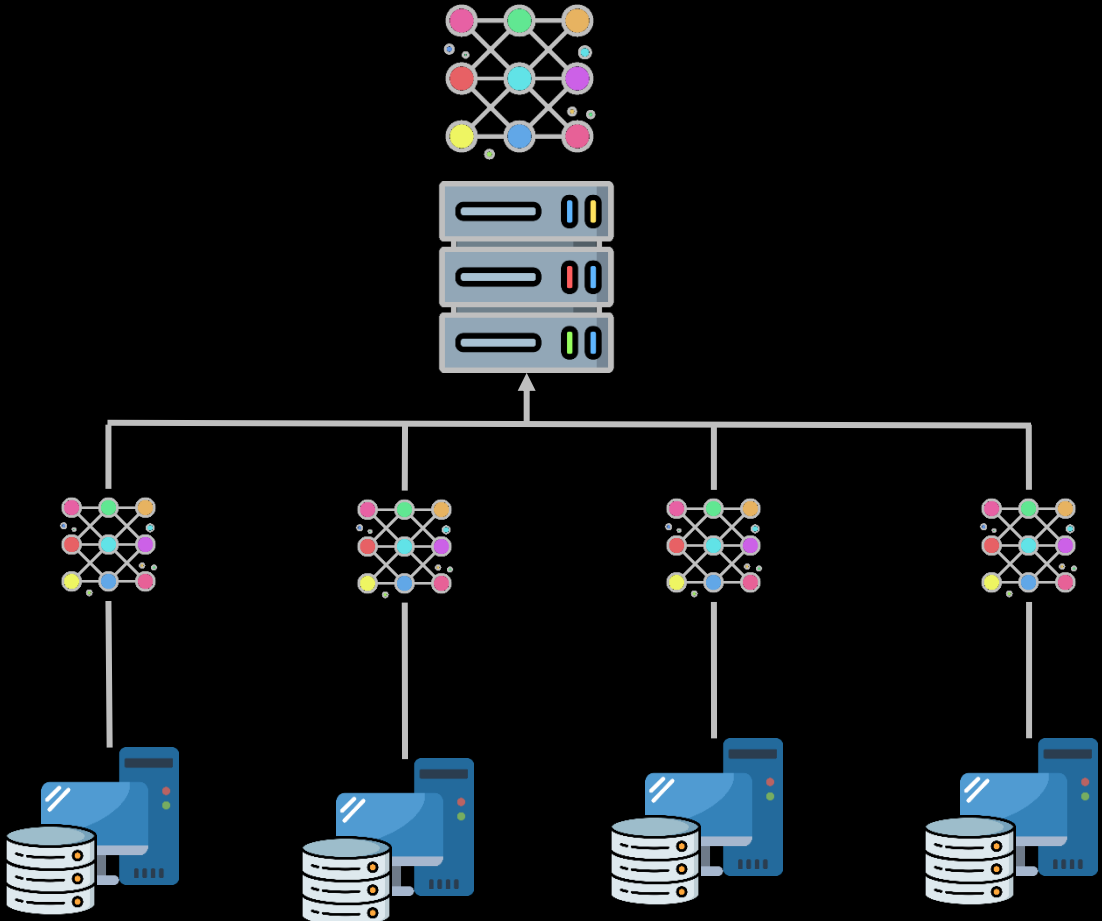
AutoAdapt

Automatic Adversarial Adaption for Stealthy Poisoning Attacks in Federated Learning

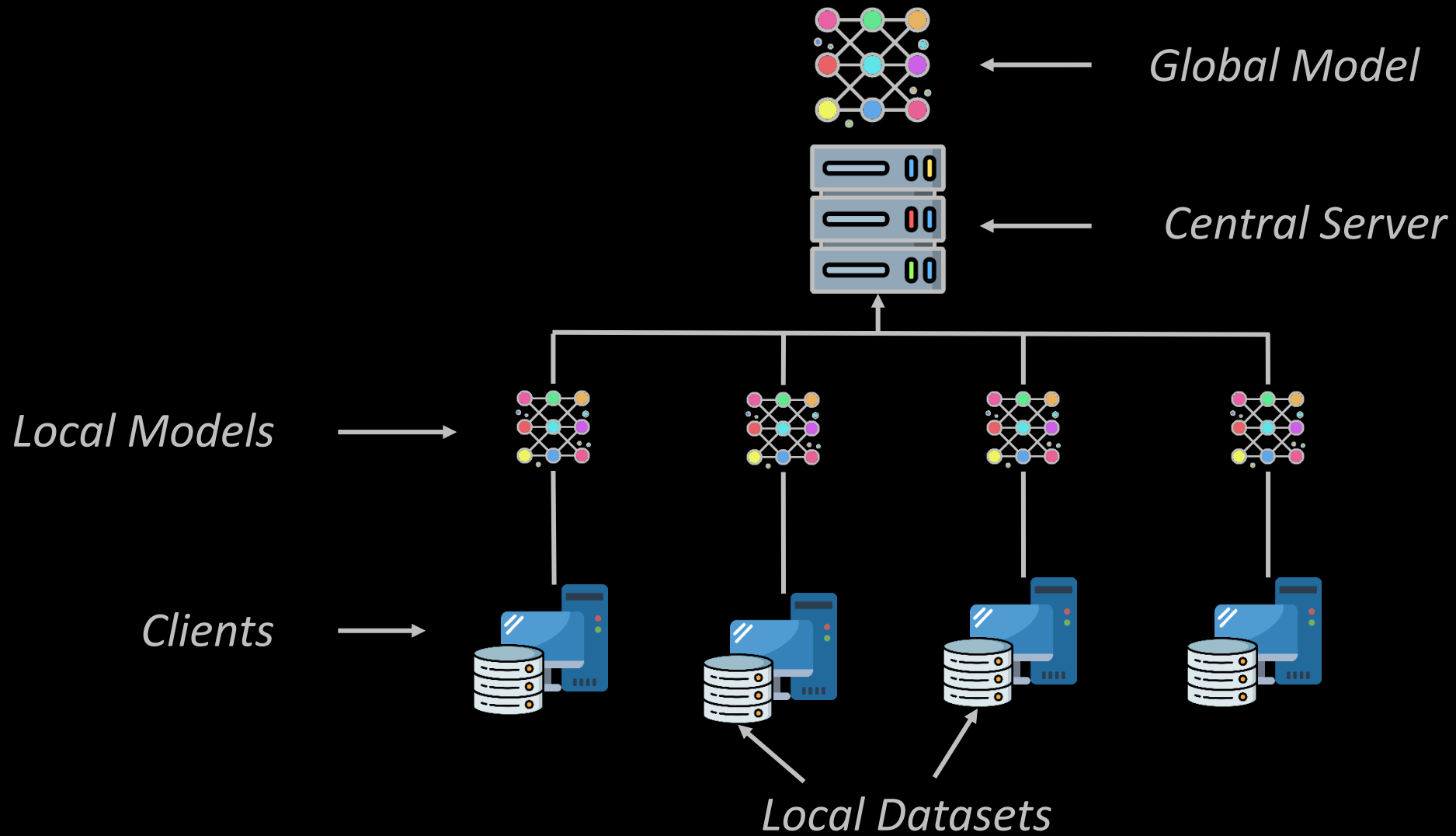
Torsten Krauss, Jan König, Alexandra Dmitrienko, and Christian Kanzow

Network and Distributed Systems Security Symposium (NDSS), 2024

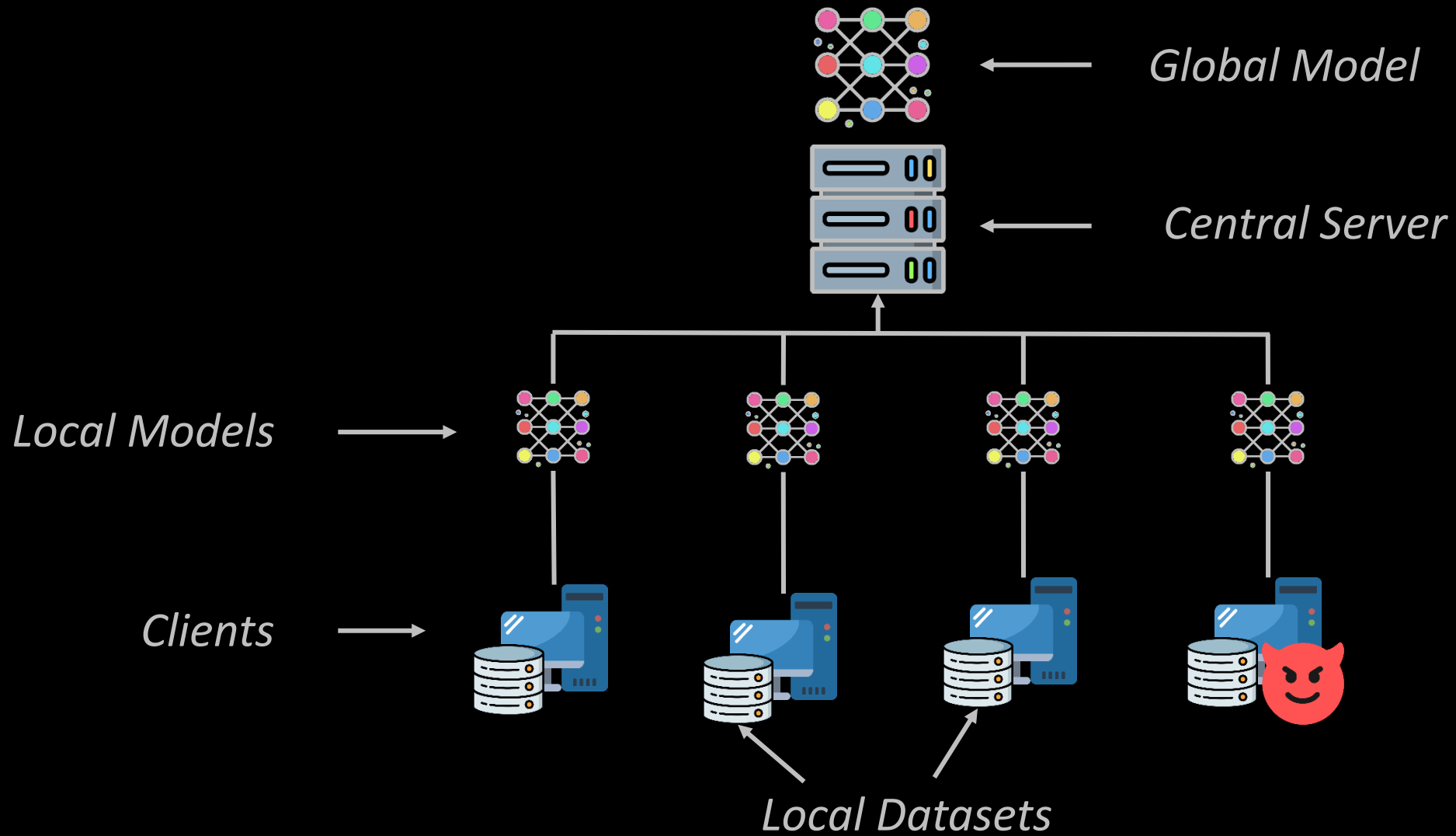
Scenario



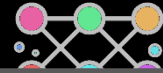
Scenario



Scenario



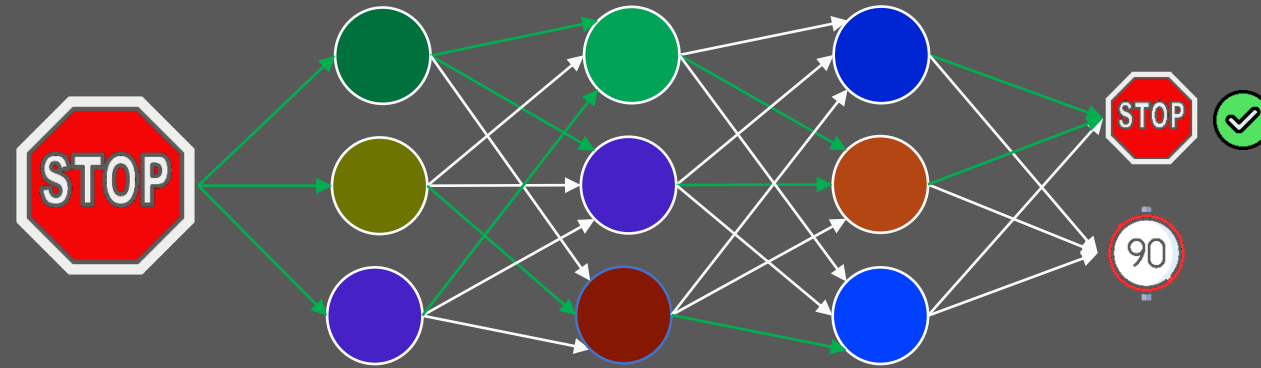
Scenario



Global Model

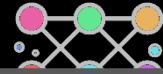
Targeted Poisoning Attack / Backdoor

Local



Local Datasets

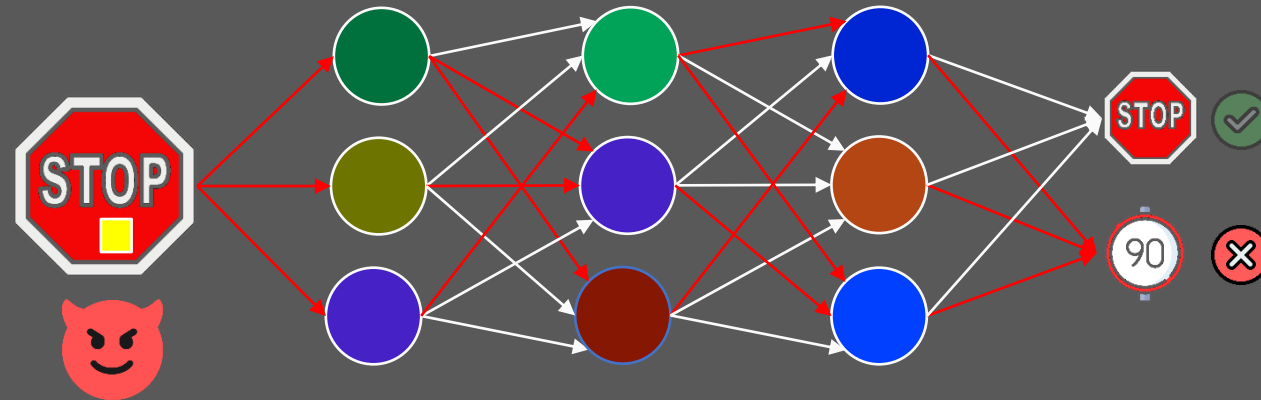
Scenario



Global Model

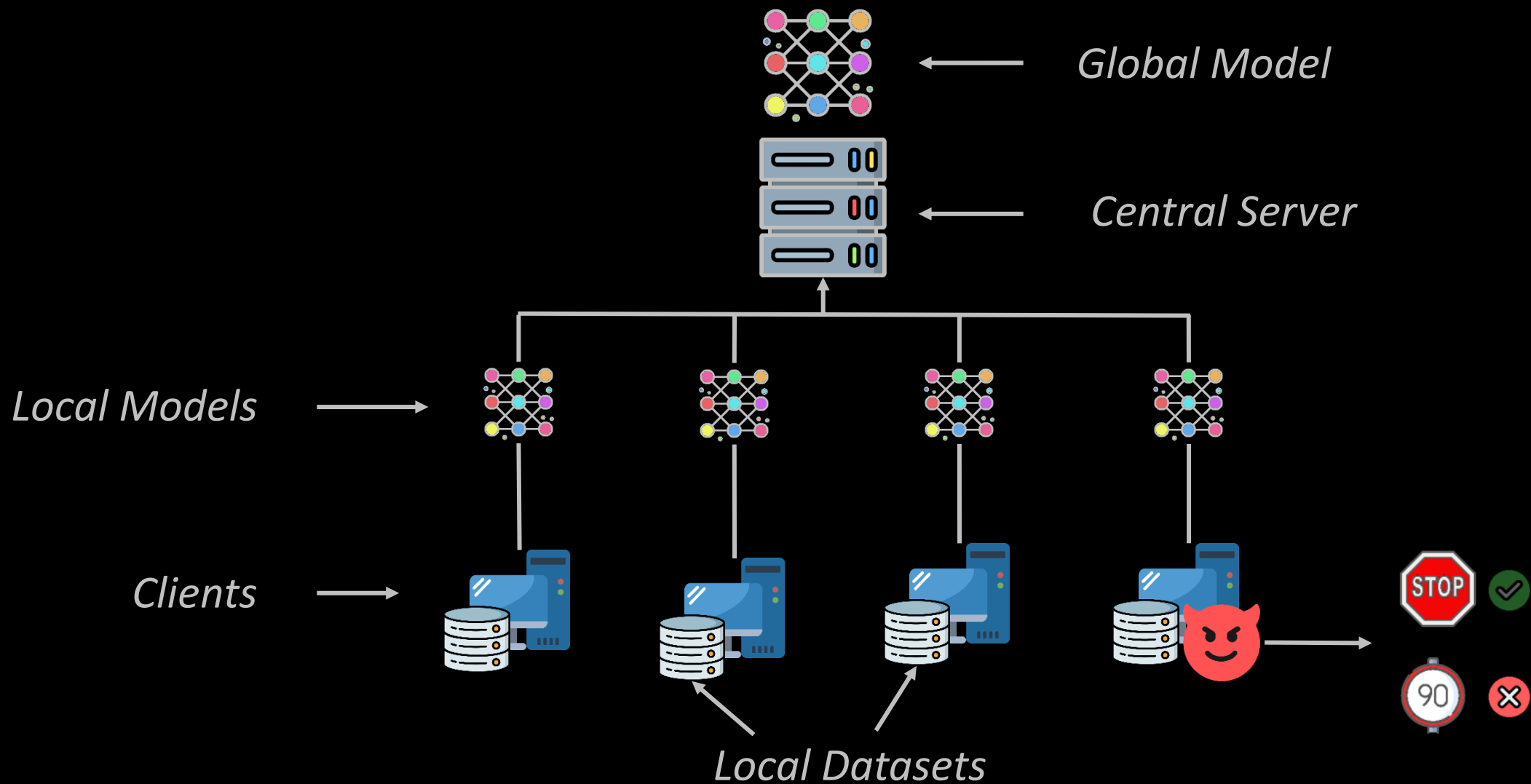
Targeted Poisoning Attack / Backdoor

Local

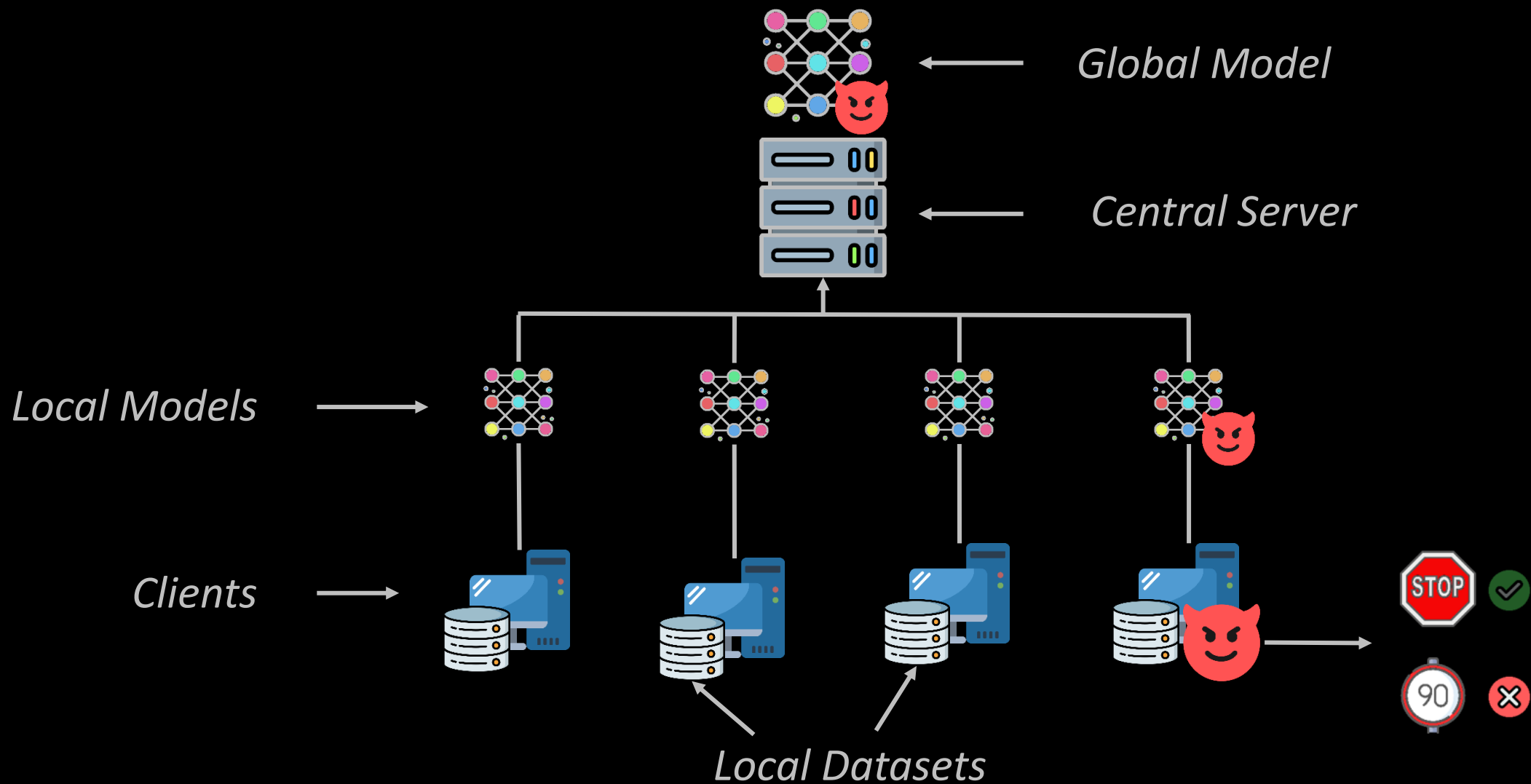


Local Datasets

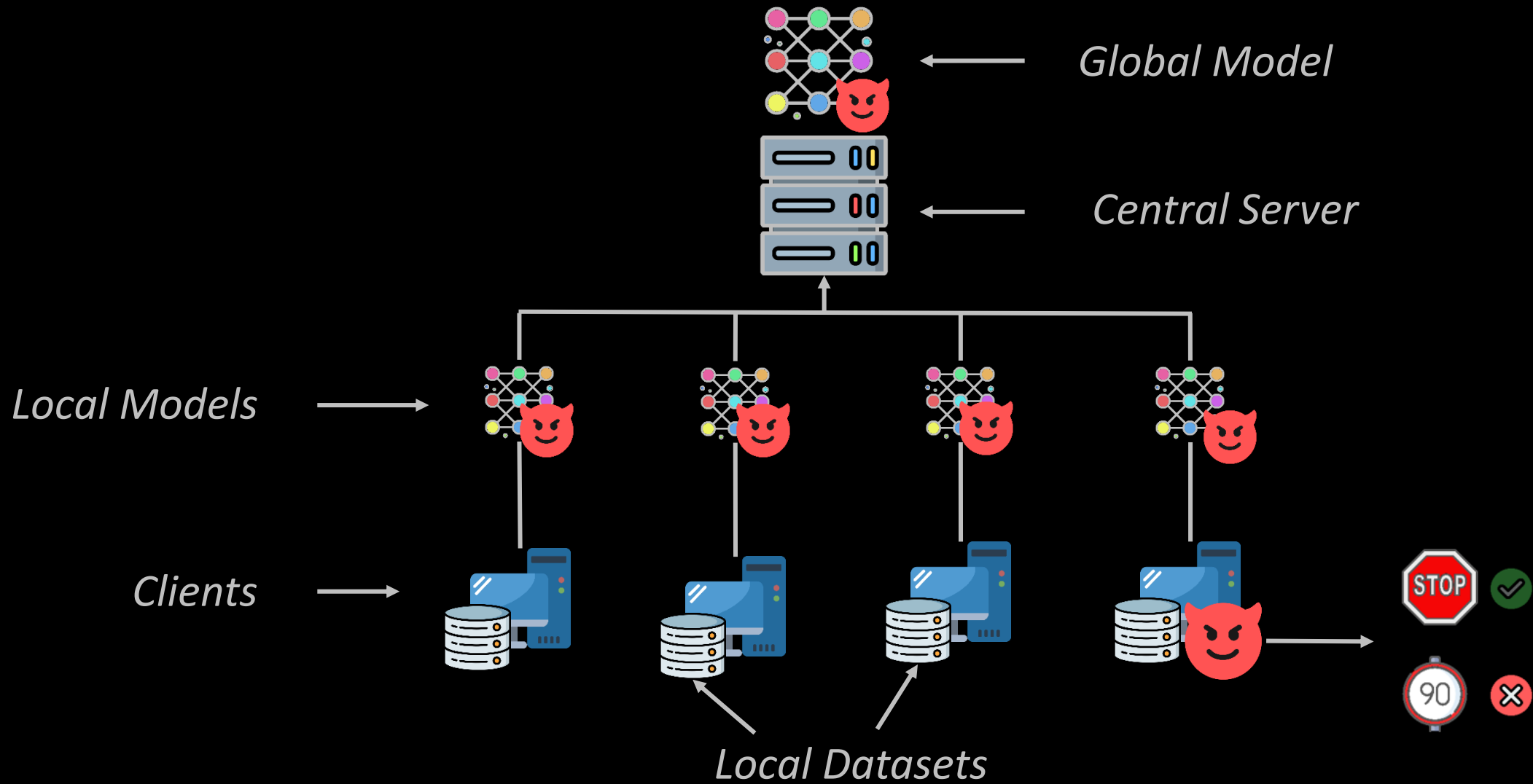
Scenario



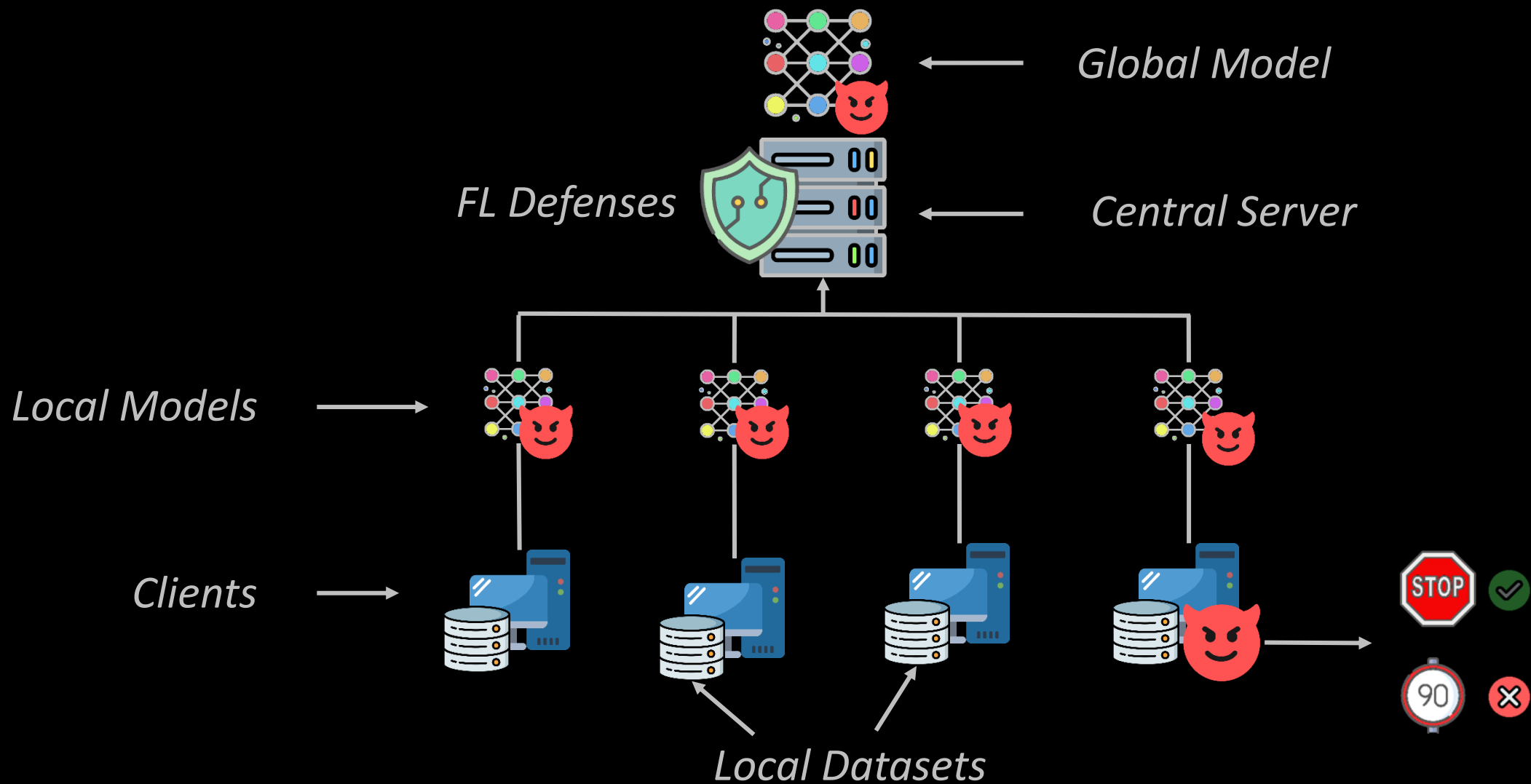
Scenario



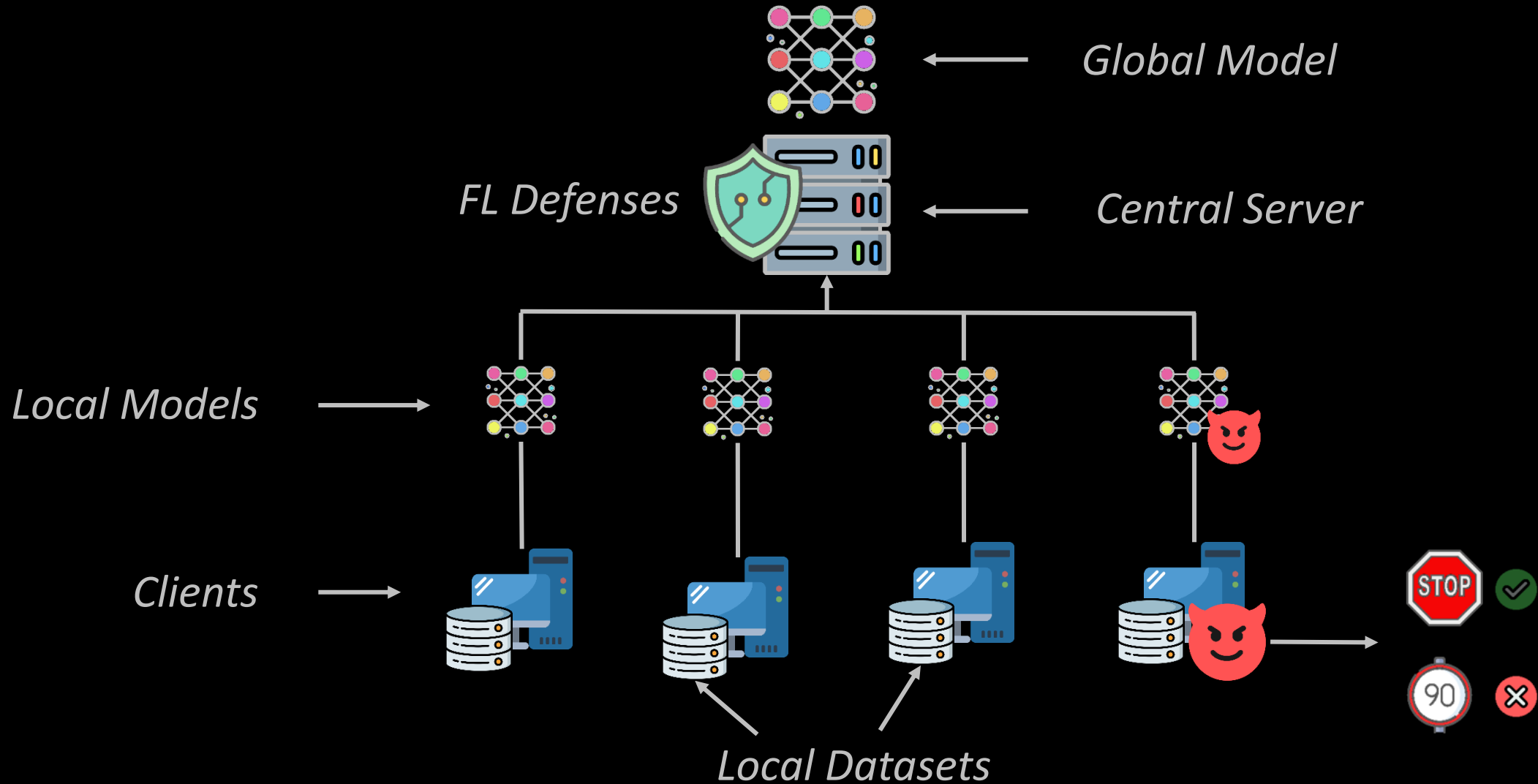
Scenario



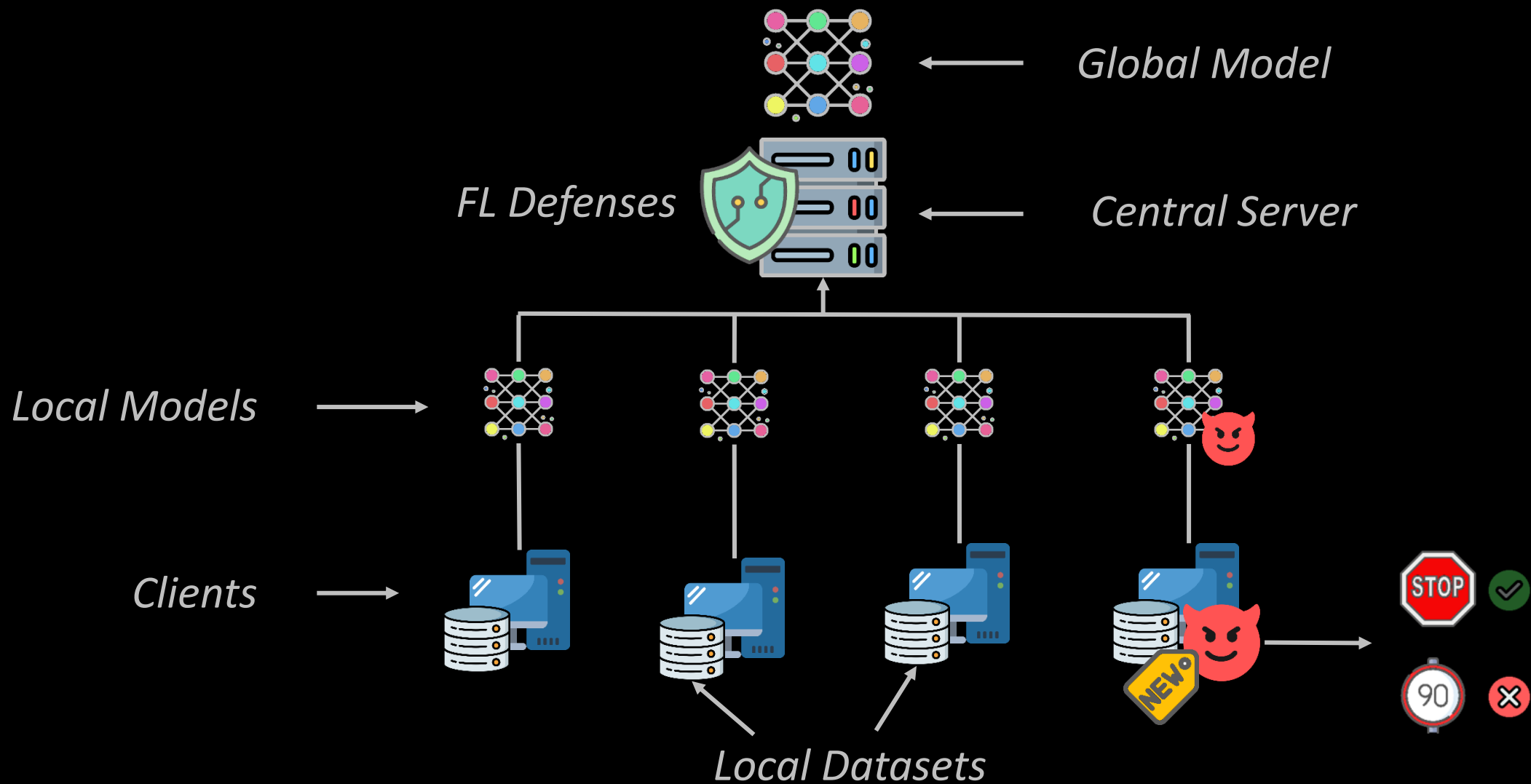
Scenario



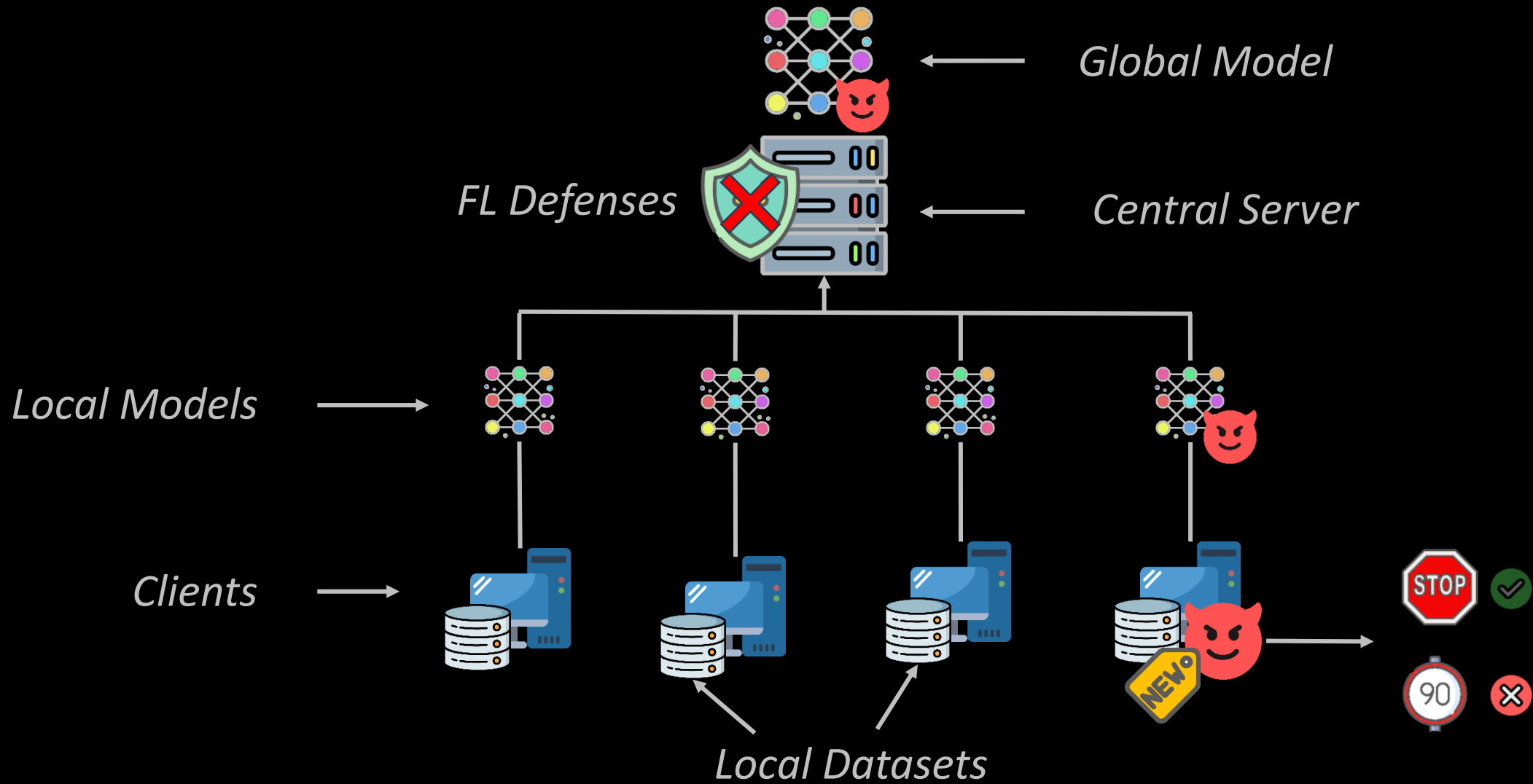
Scenario



Scenario



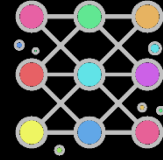
Scenario



Defenses against Poisoning in FL



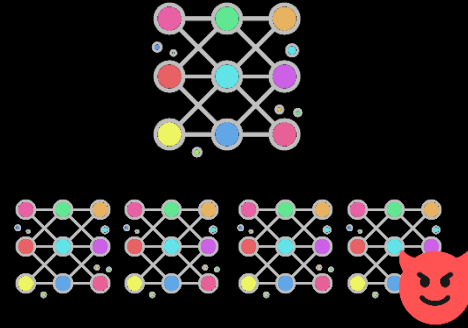
FL Defenses



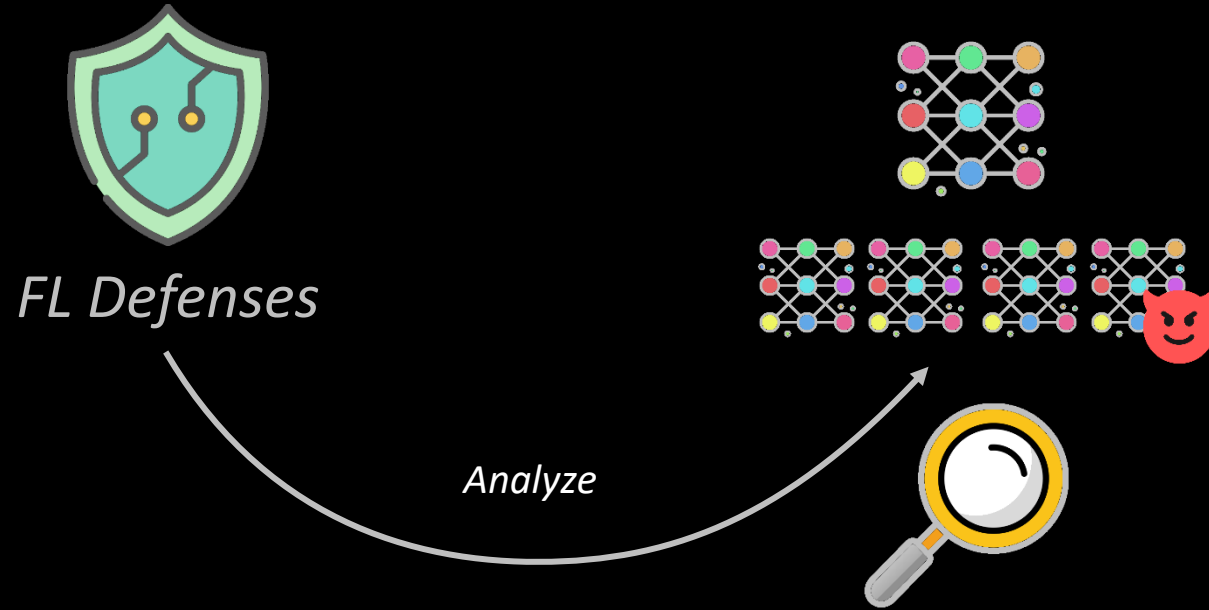
Defenses against Poisoning in FL



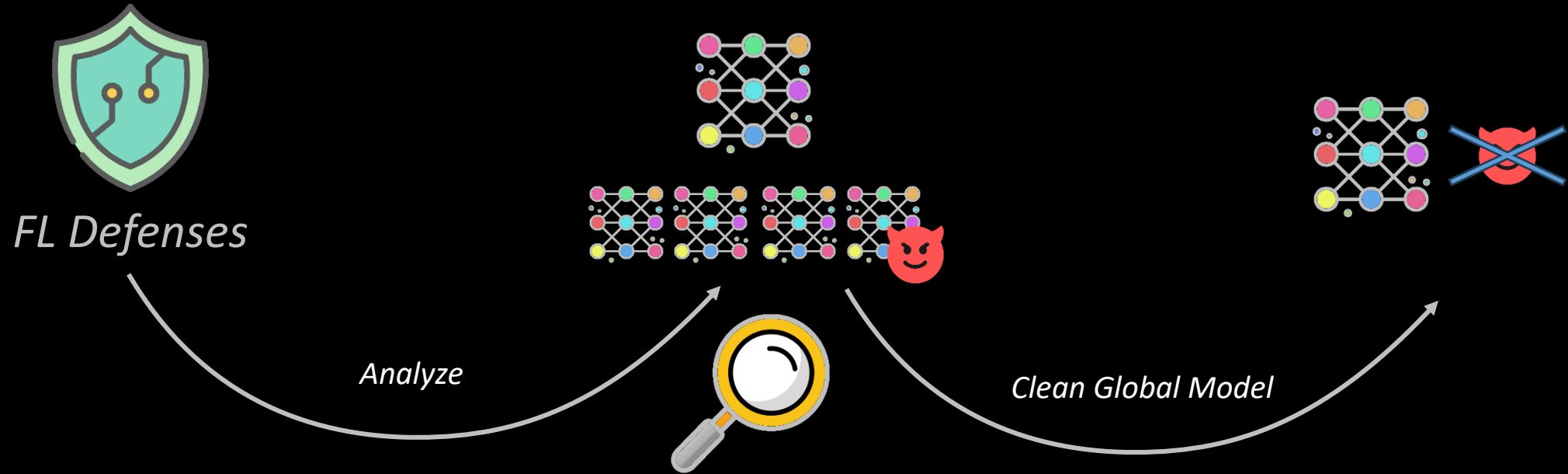
FL Defenses



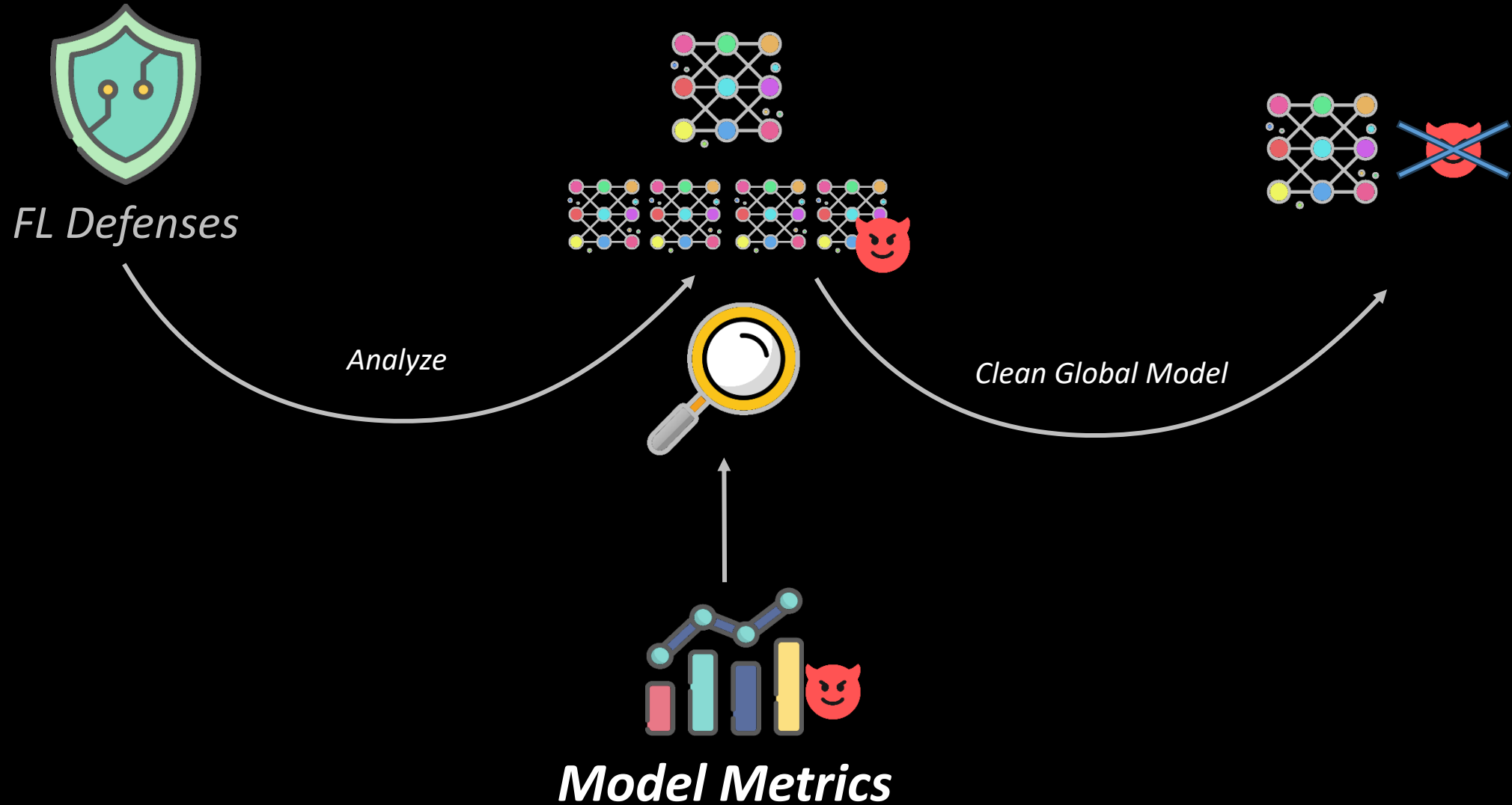
Defenses against Poisoning in FL



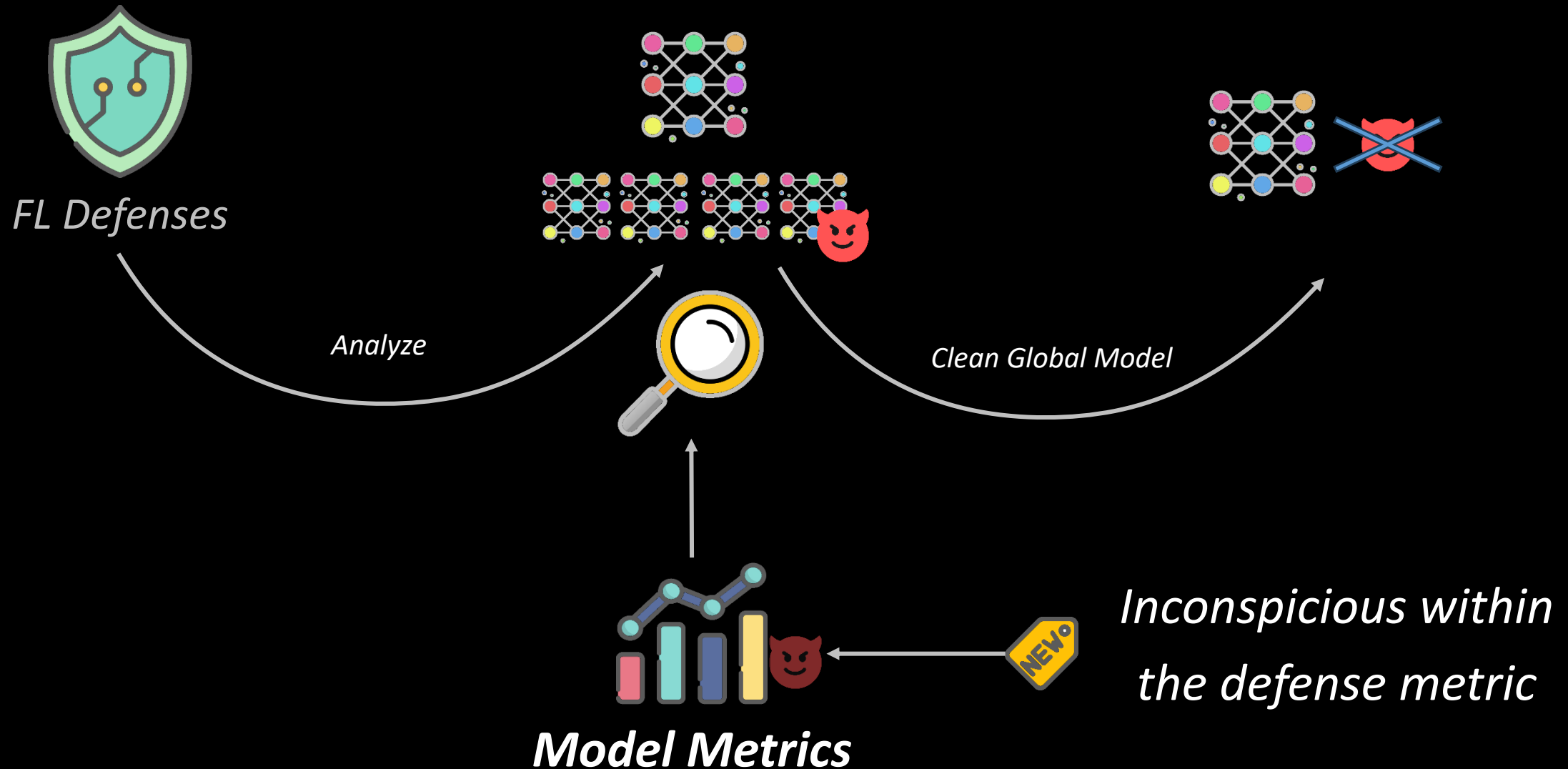
Defenses against Poisoning in FL



Defenses against Poisoning in FL



Defenses against Poisoning in FL

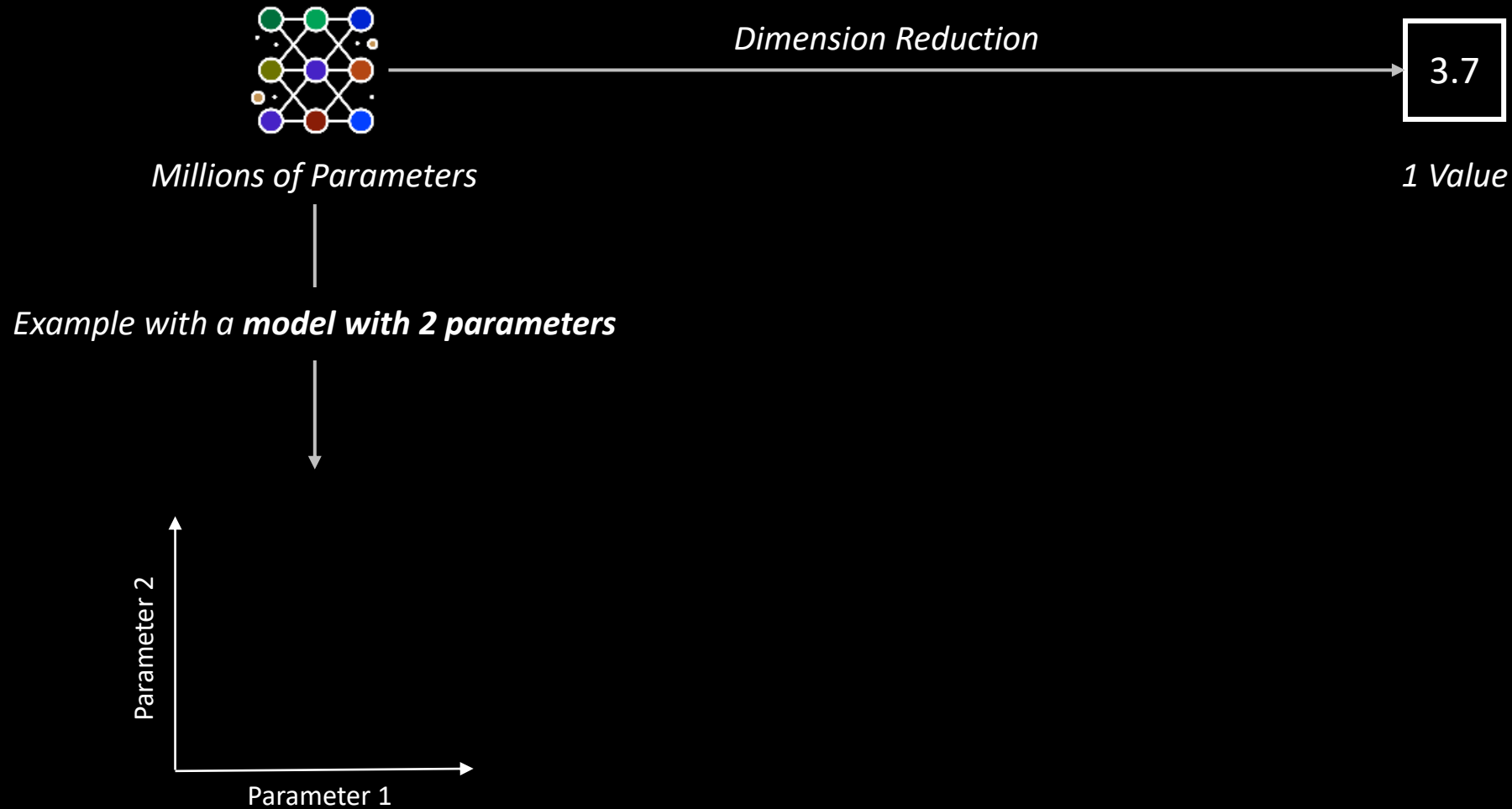


Model Metrics

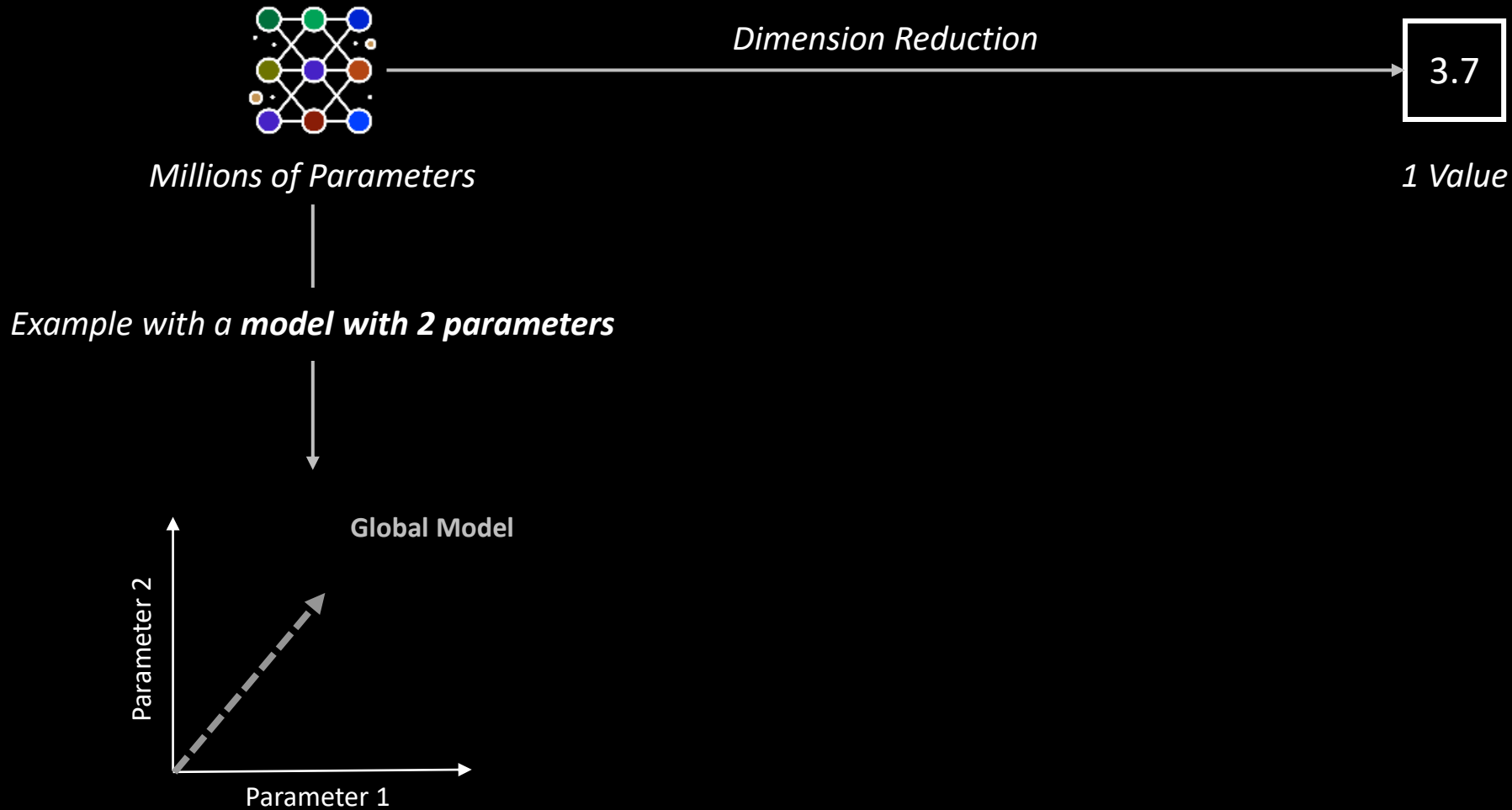
Model Metrics



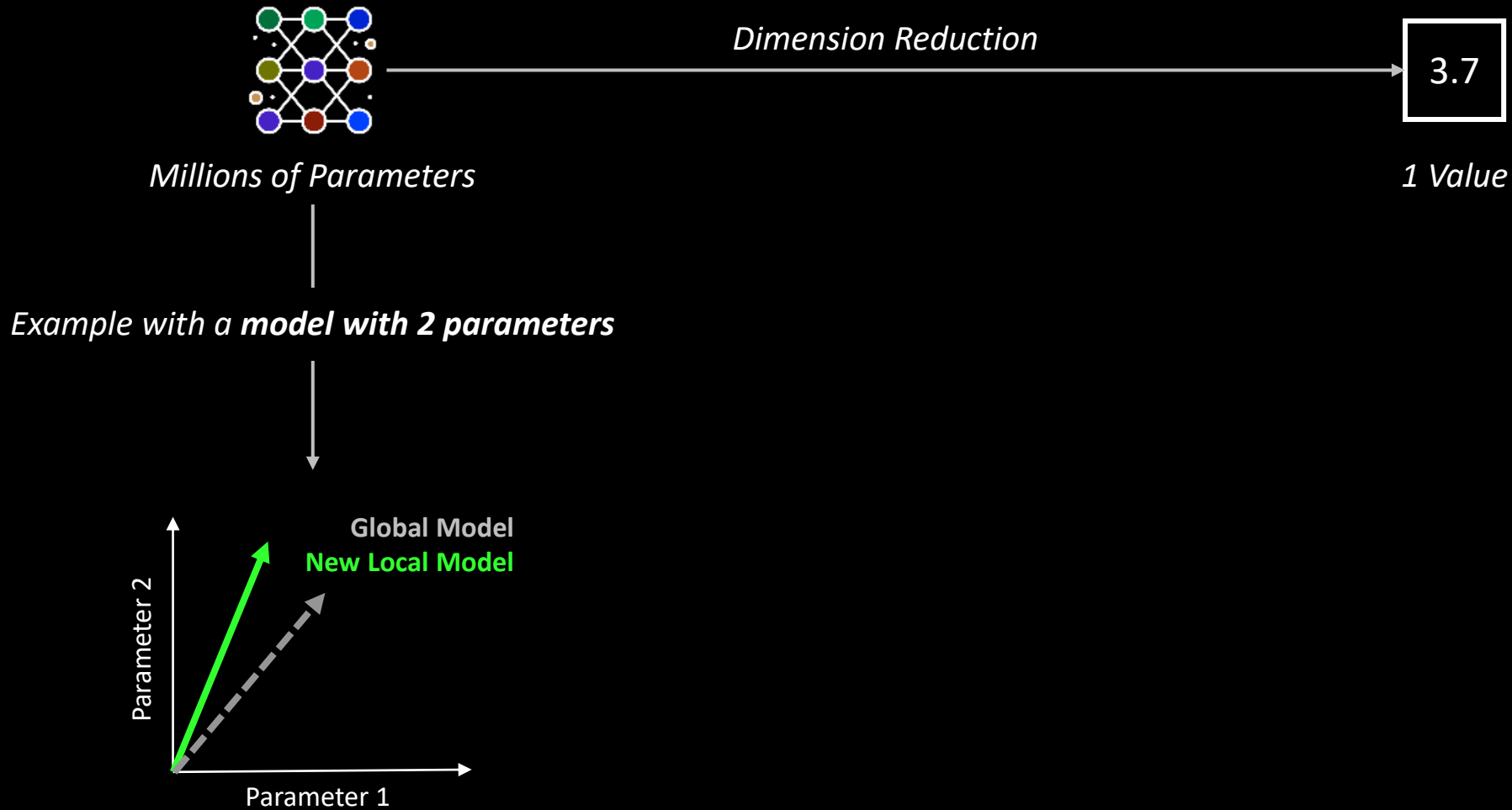
Model Metrics



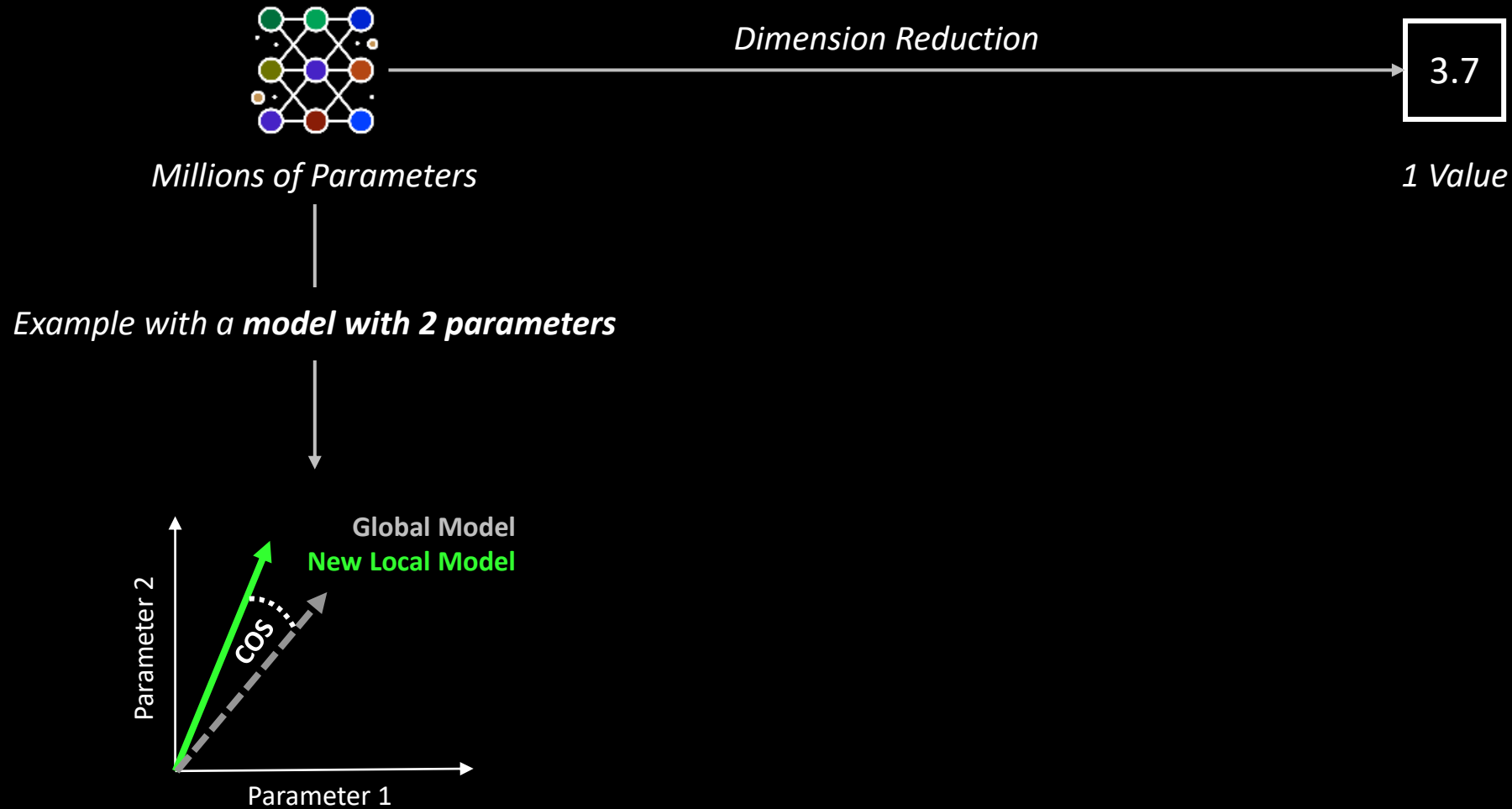
Model Metrics



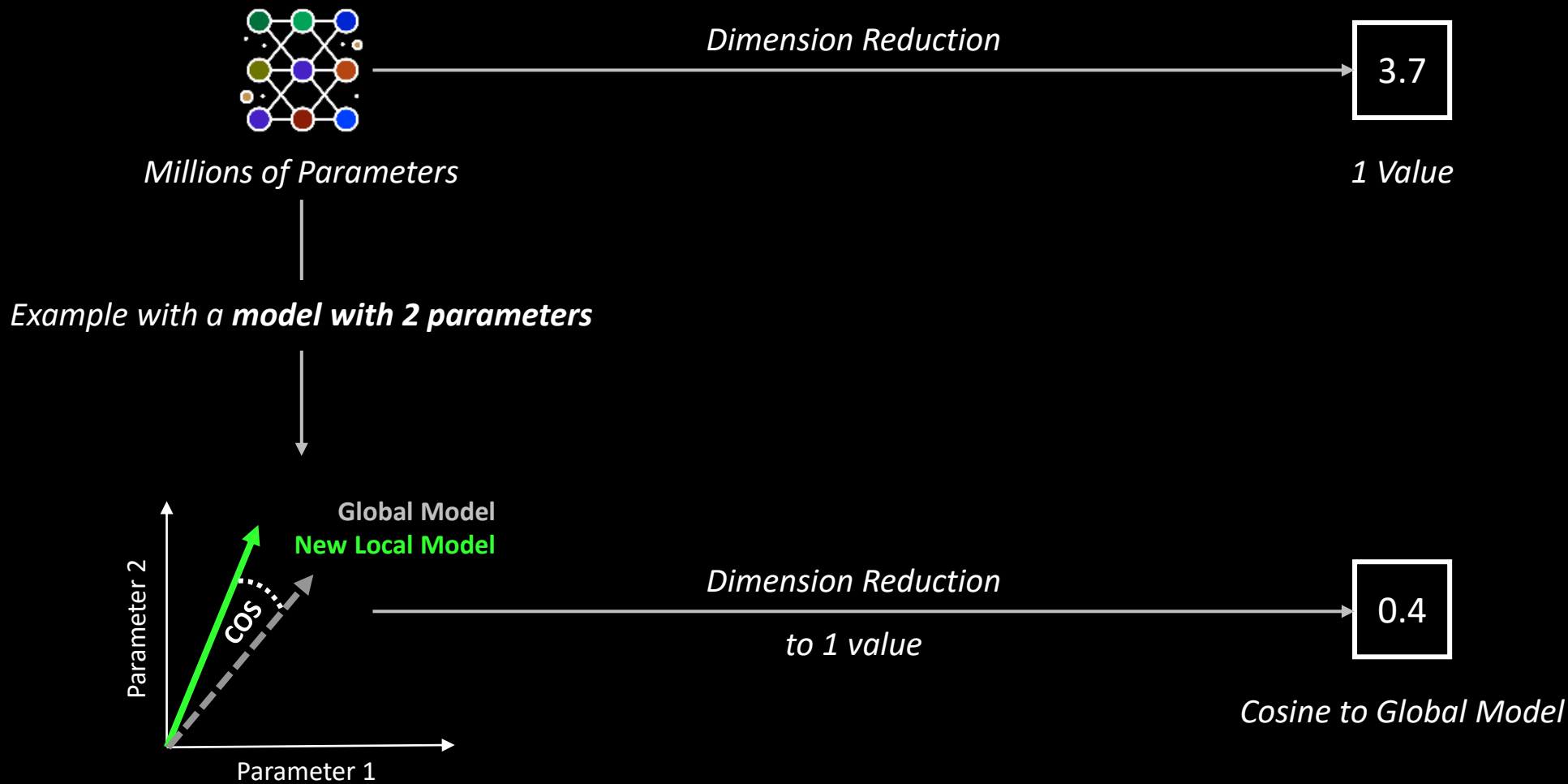
Model Metrics



Model Metrics



Model Metrics



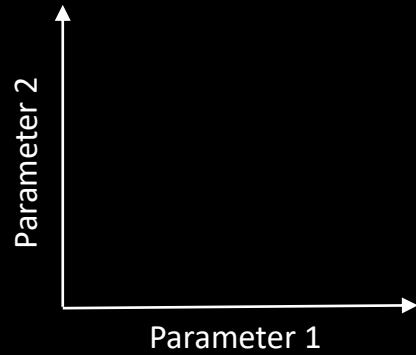
Attacking FL Systems



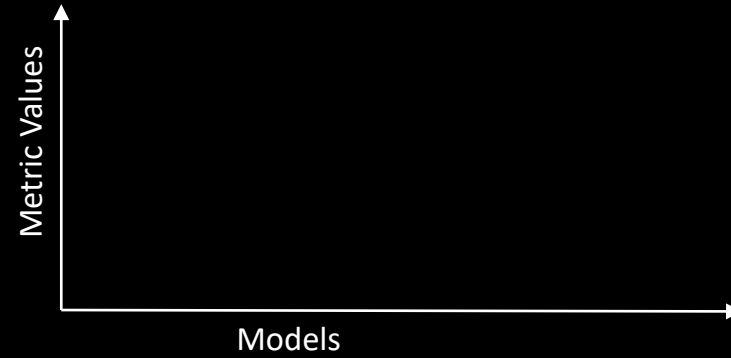
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



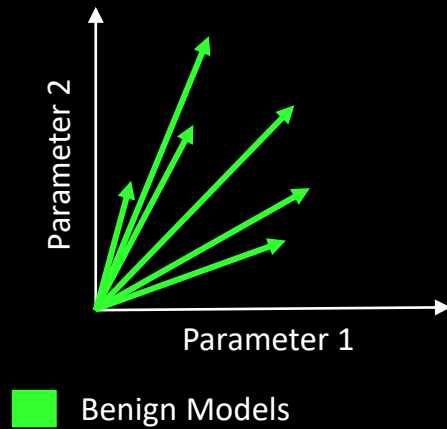
Dimension reduction of a model to one metric value



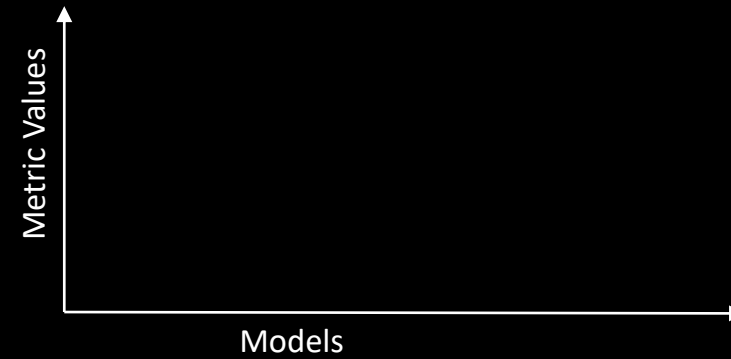
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



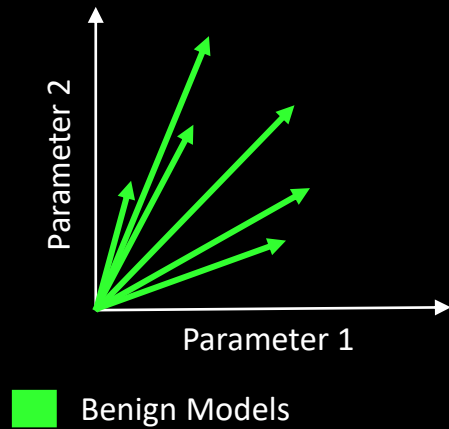
Dimension reduction of a model to one metric value



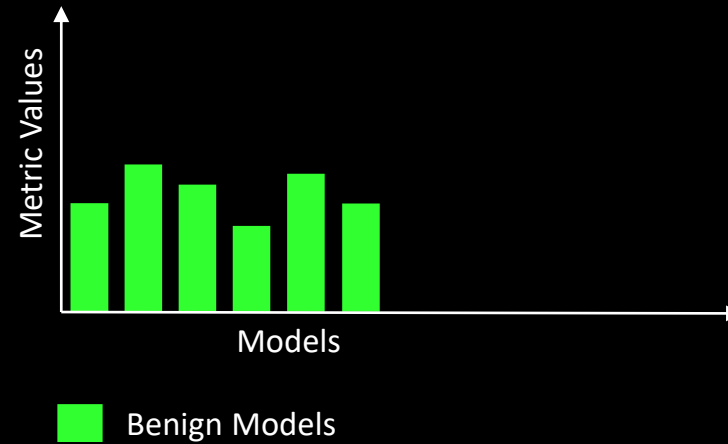
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



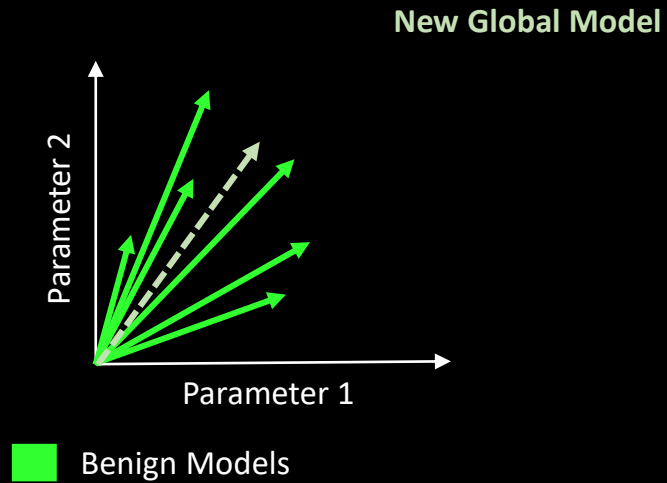
Dimension reduction of a model to one metric value



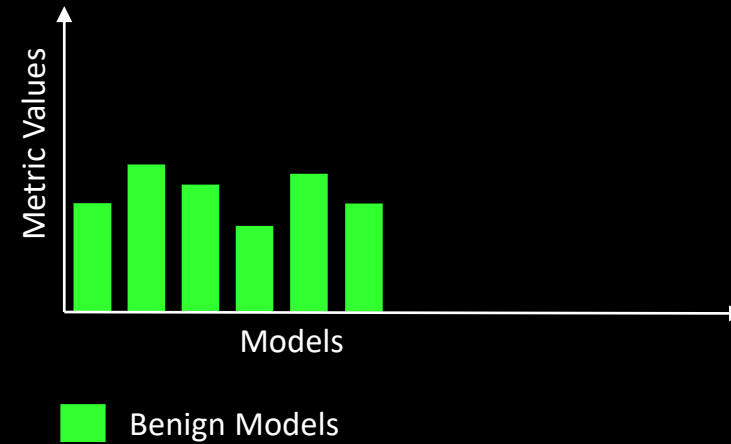
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



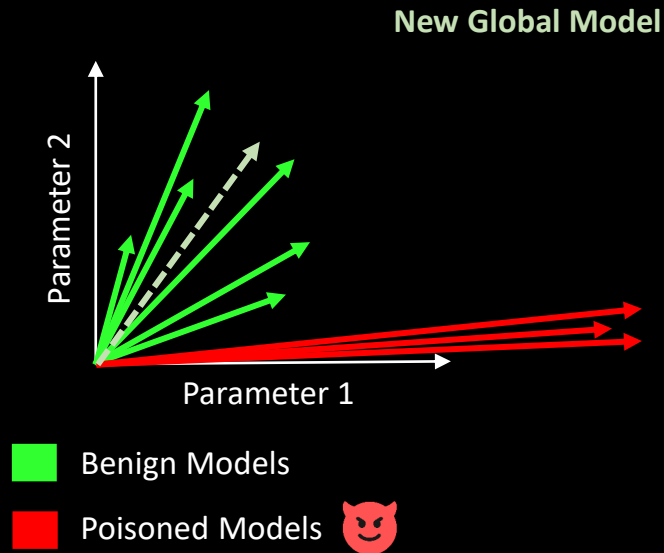
Dimension reduction of a model to one metric value



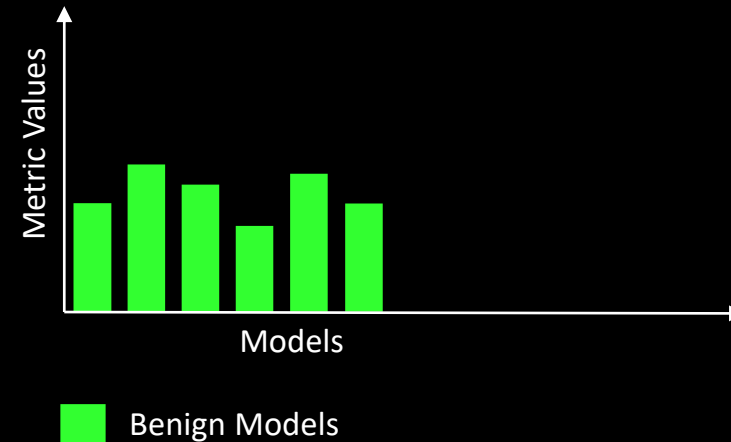
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



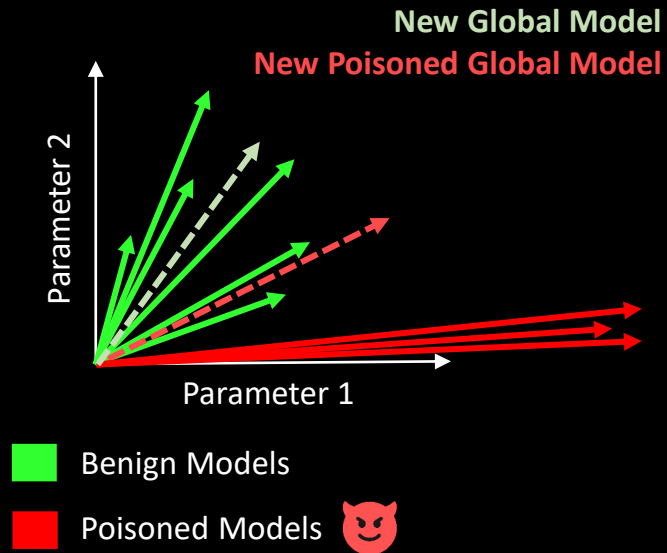
Dimension reduction of a model to one metric value



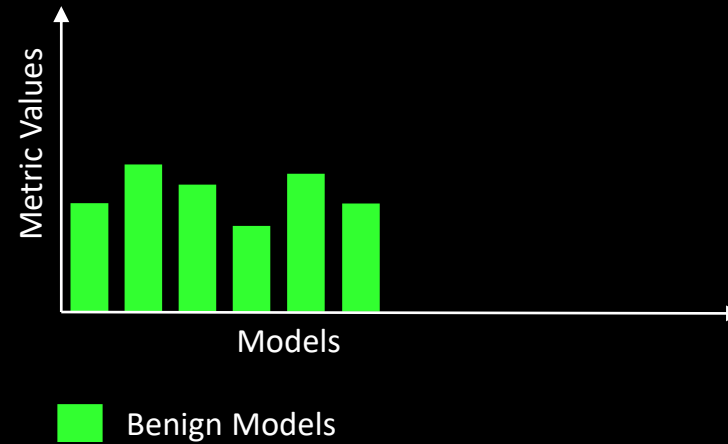
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



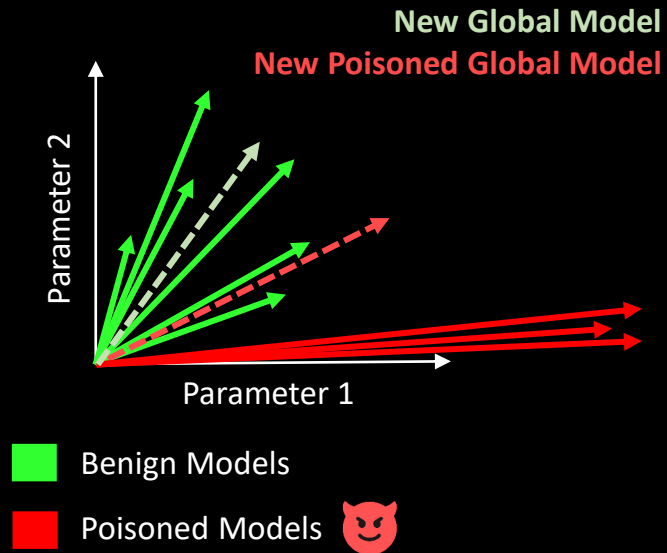
Dimension reduction of a model to one metric value



Attacking FL Systems



Exemplary visualization of a model with 2 parameters



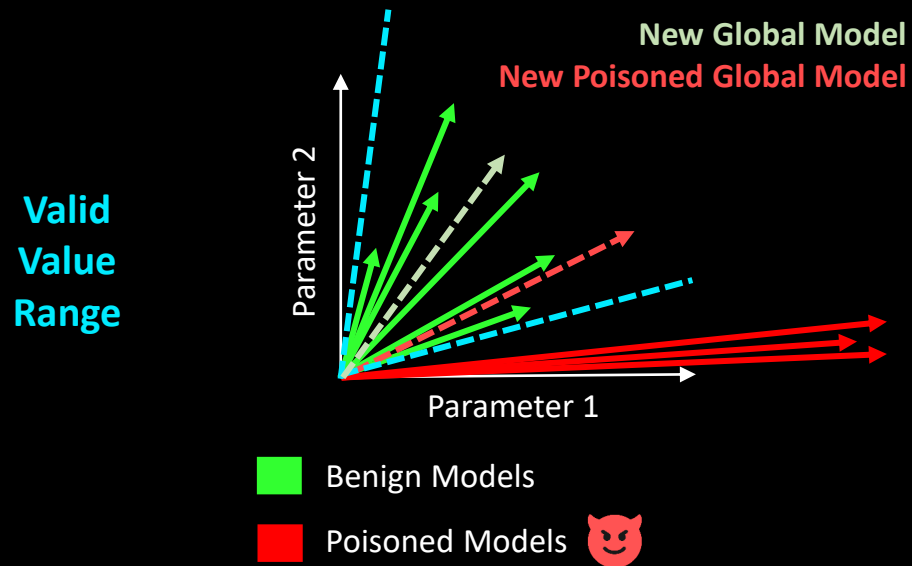
Dimension reduction of a model to one metric value



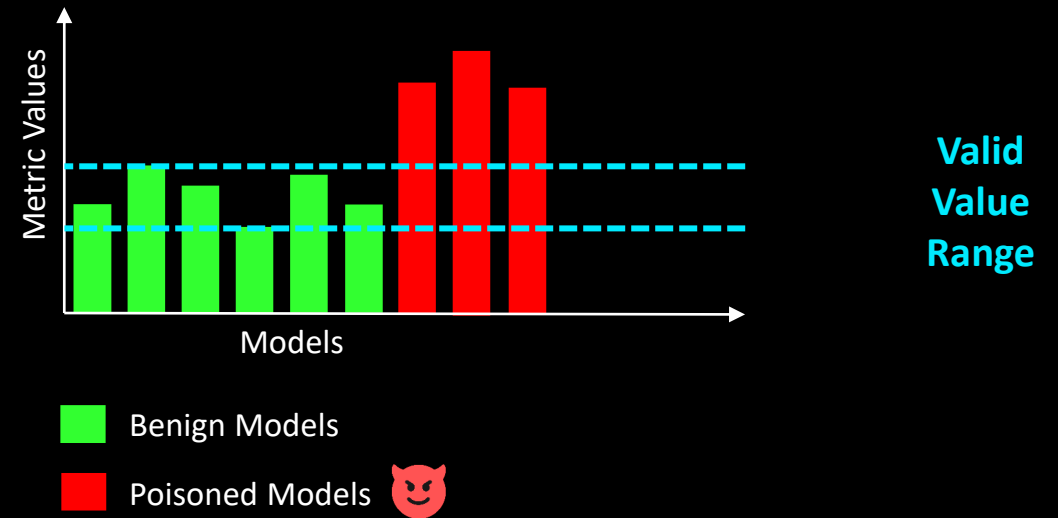
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



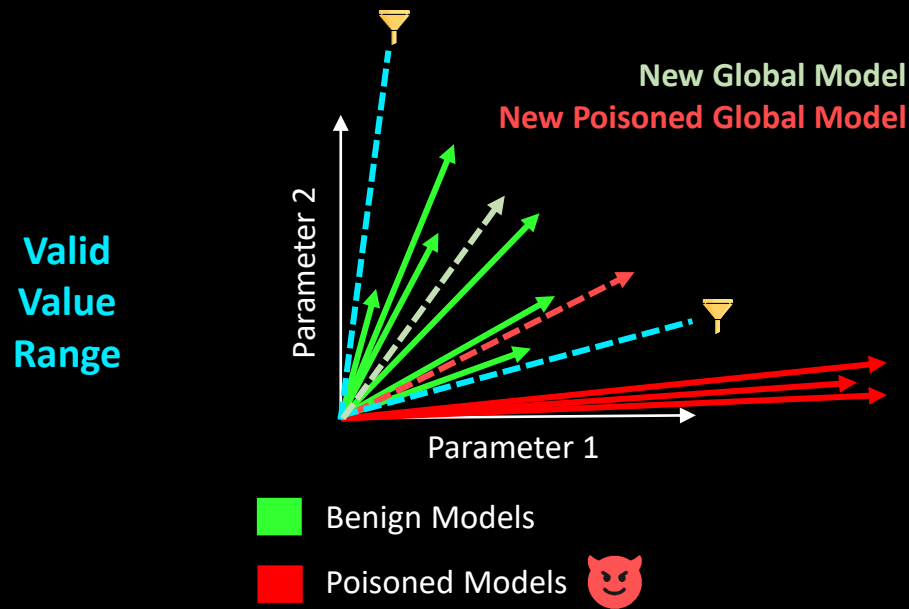
Dimension reduction of a model to one metric value



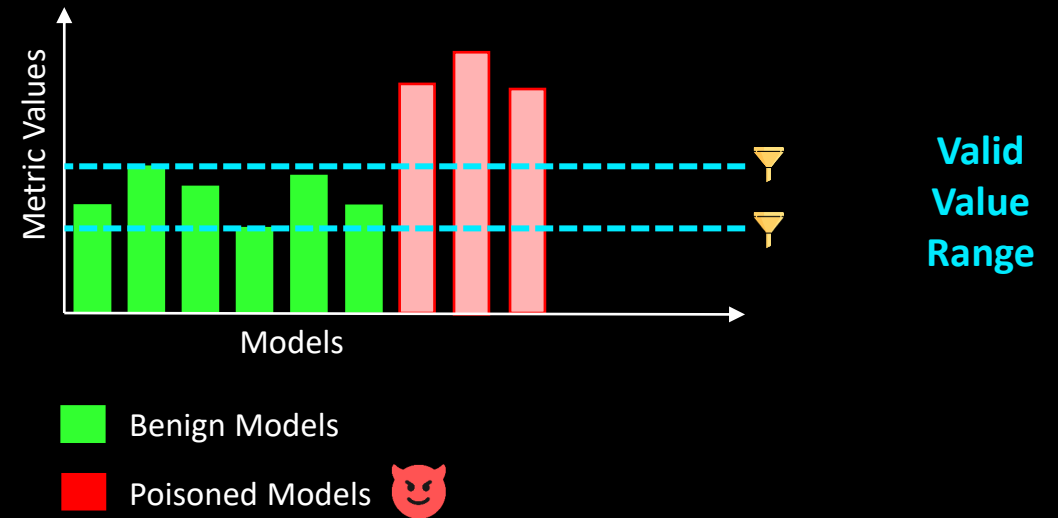
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



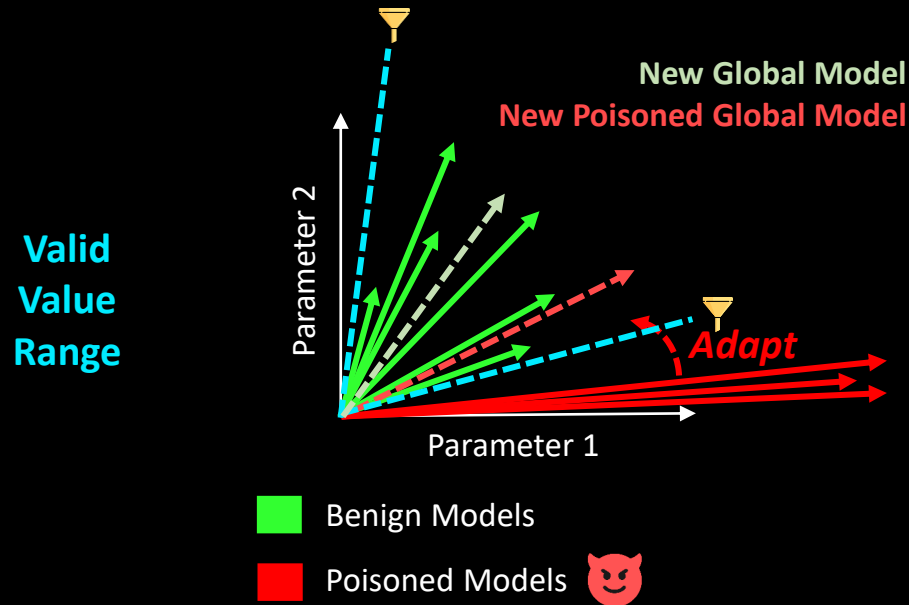
Dimension reduction of a model to one metric value



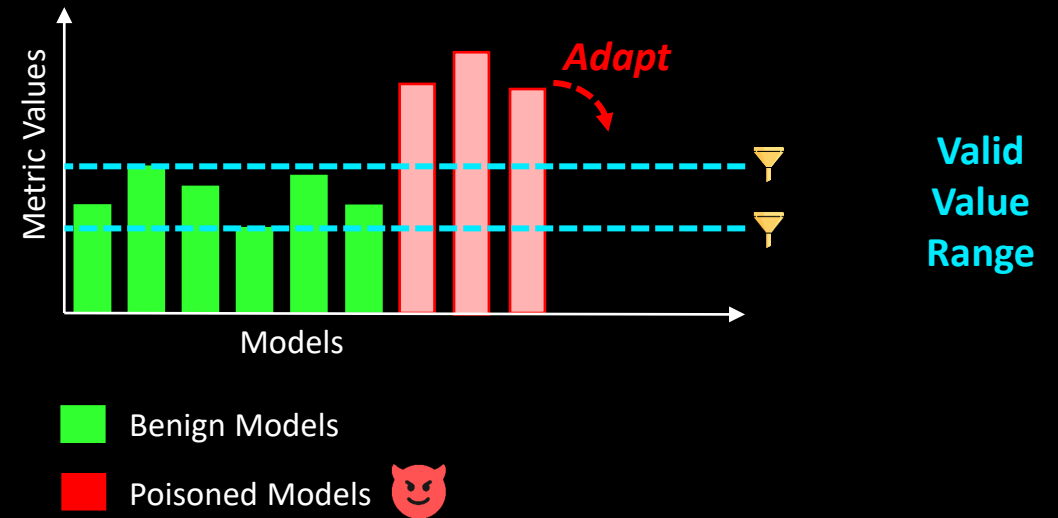
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



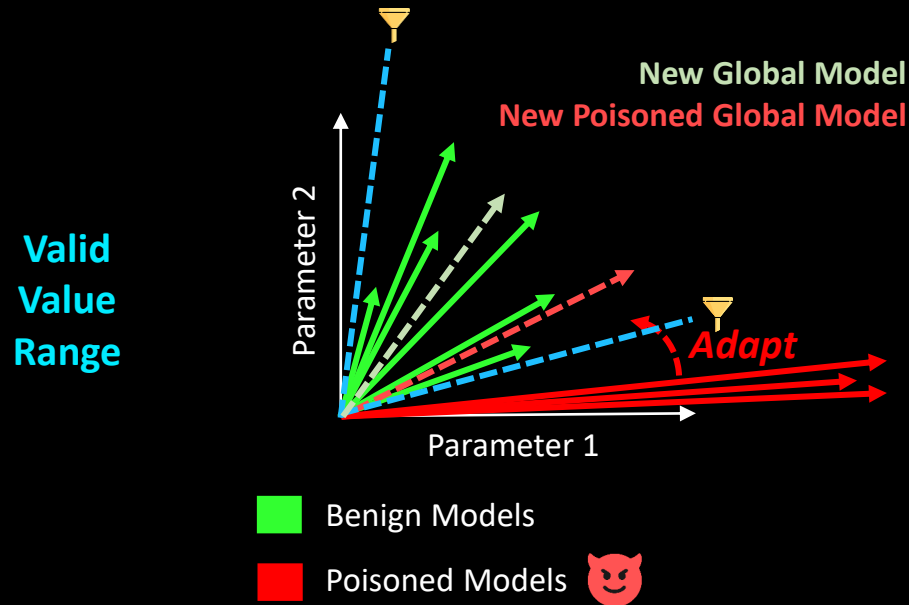
Dimension reduction of a model to one metric value



Attacking FL Systems



Exemplary visualization of a model with 2 parameters



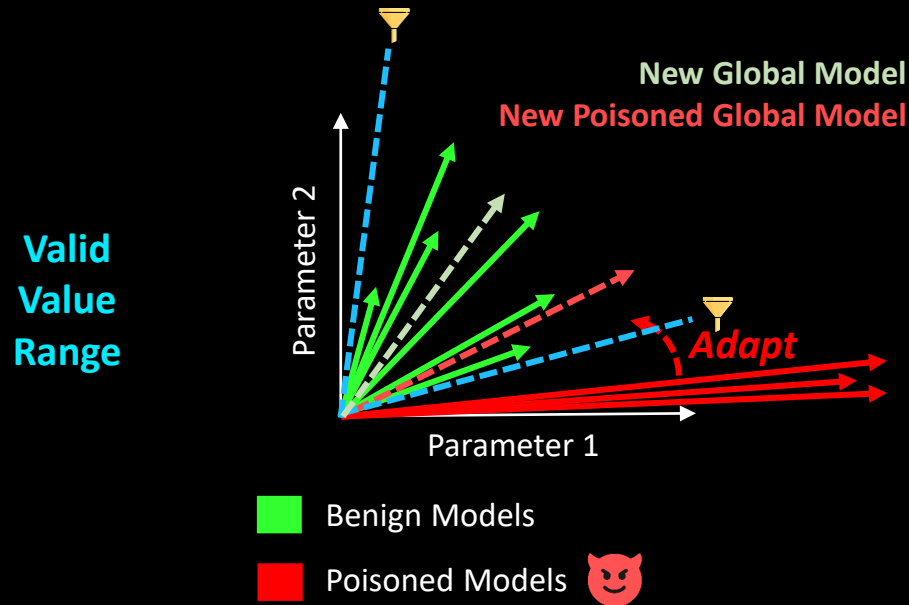
Dimension reduction of a model to one metric value



Attacking FL Systems



Exemplary visualization of a model with 2 parameters



Dimension reduction of a model to one metric value

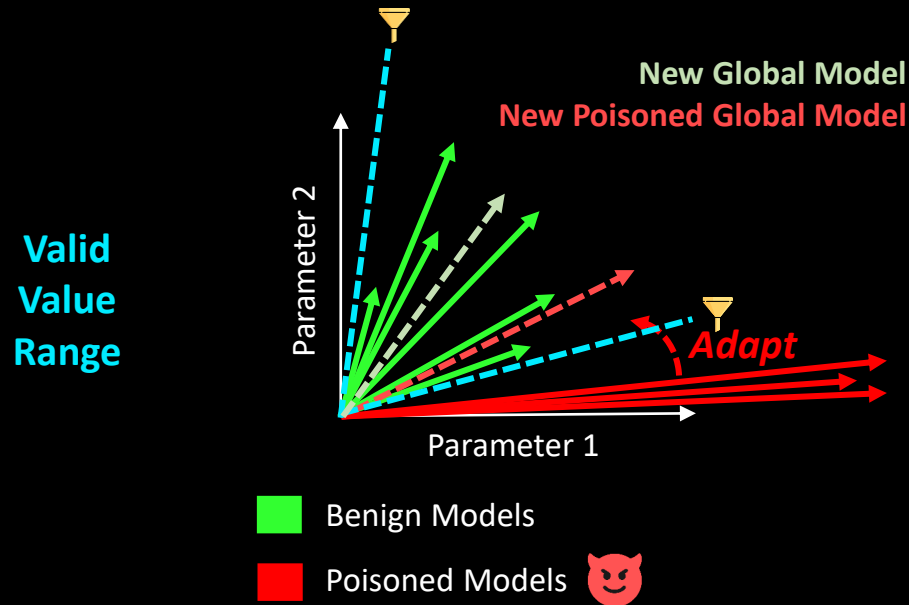


$$Loss = Loss_{data} + Loss_{adaption}$$

Attacking FL Systems



Exemplary visualization of a model with 2 parameters



Dimension reduction of a model to one metric value



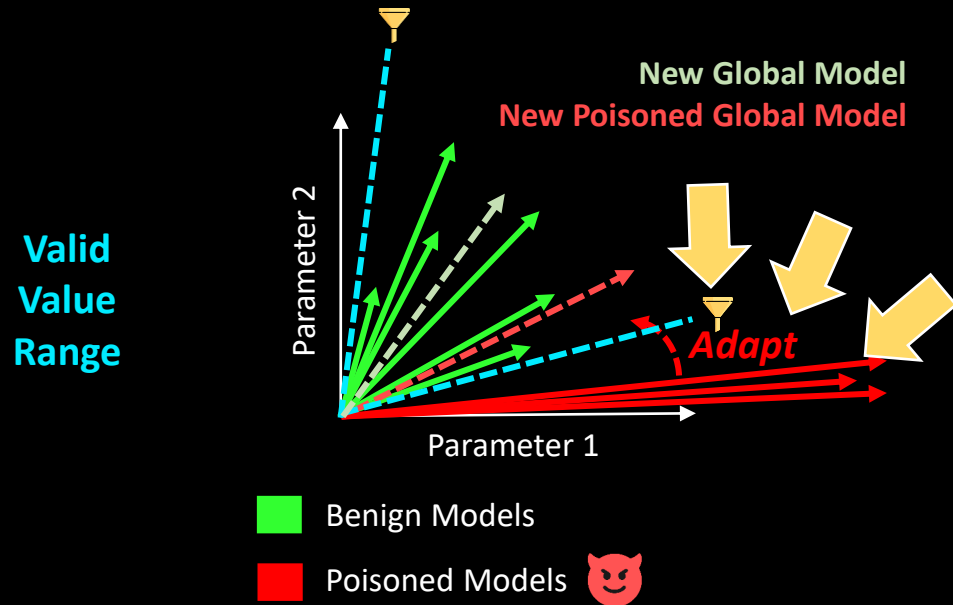
$$Loss = Loss_{data} + Loss_{adaption}$$

→ Our work improves the current method for adversarial adaption

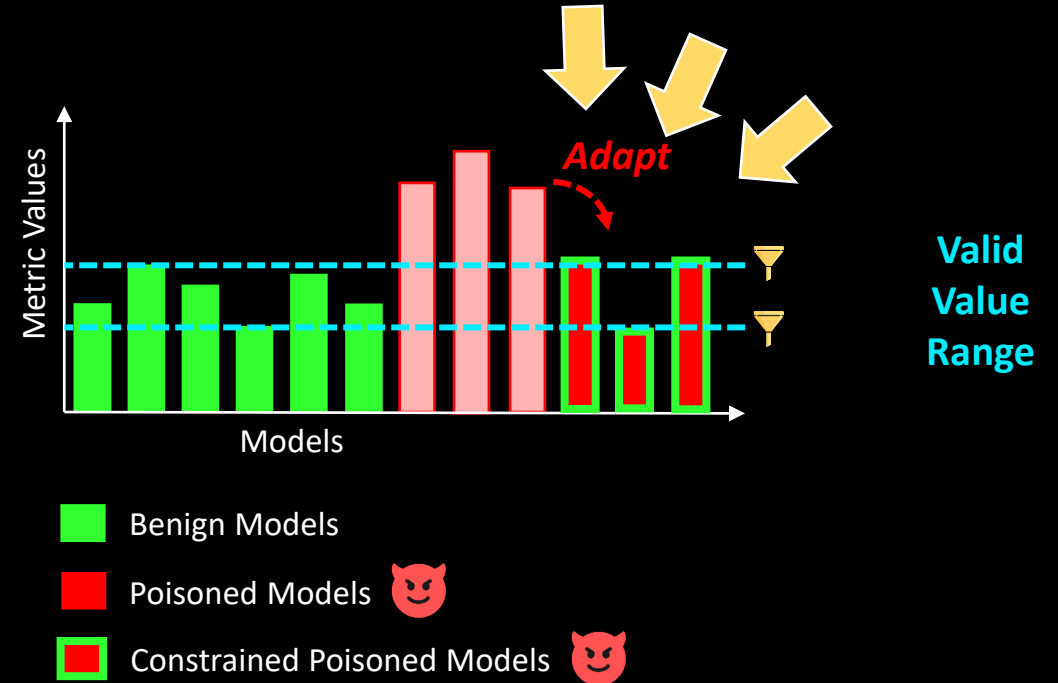
Attacking FL Systems



Exemplary visualization of a model with 2 parameters



Dimension reduction of a model to one metric value

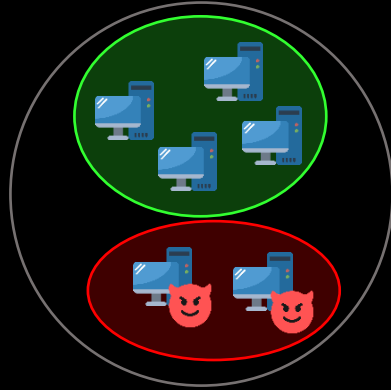


$$Loss = Loss_{data} + Loss_{adaption}$$

→ Our work improves the current method for adversarial adaption

Classic Adaption Workflow

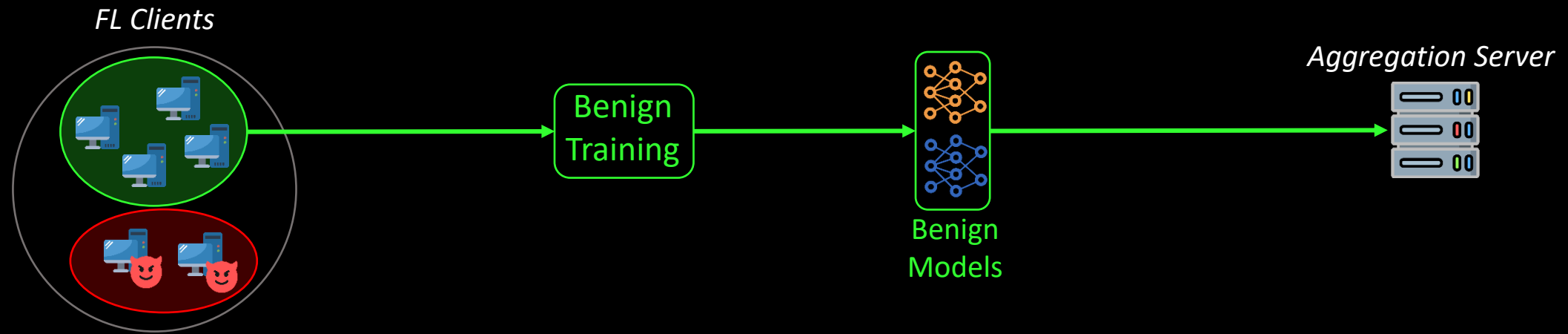
FL Clients



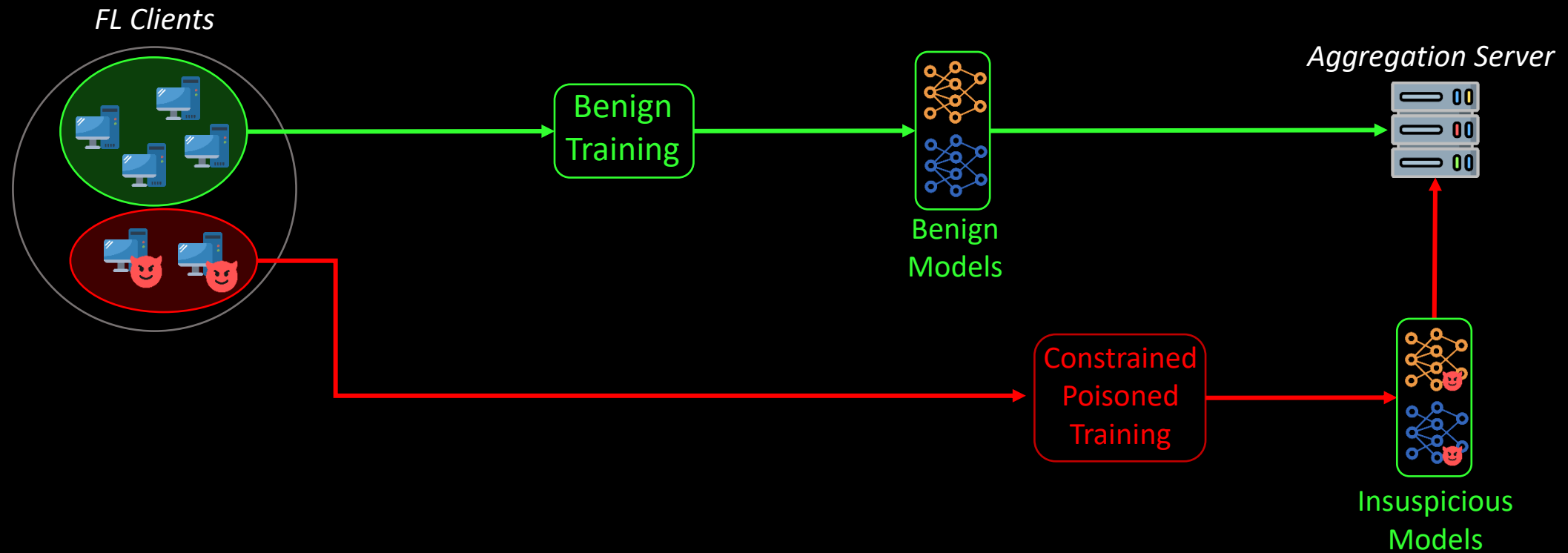
Aggregation Server



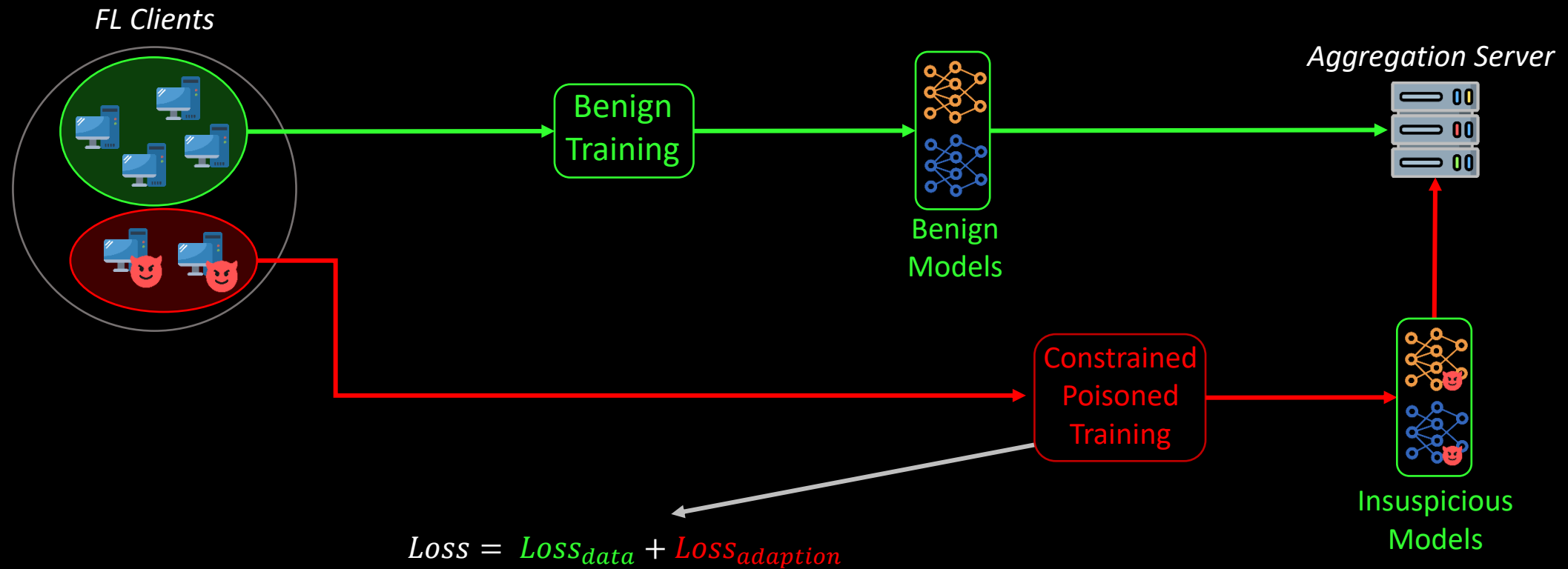
Classic Adaption Workflow



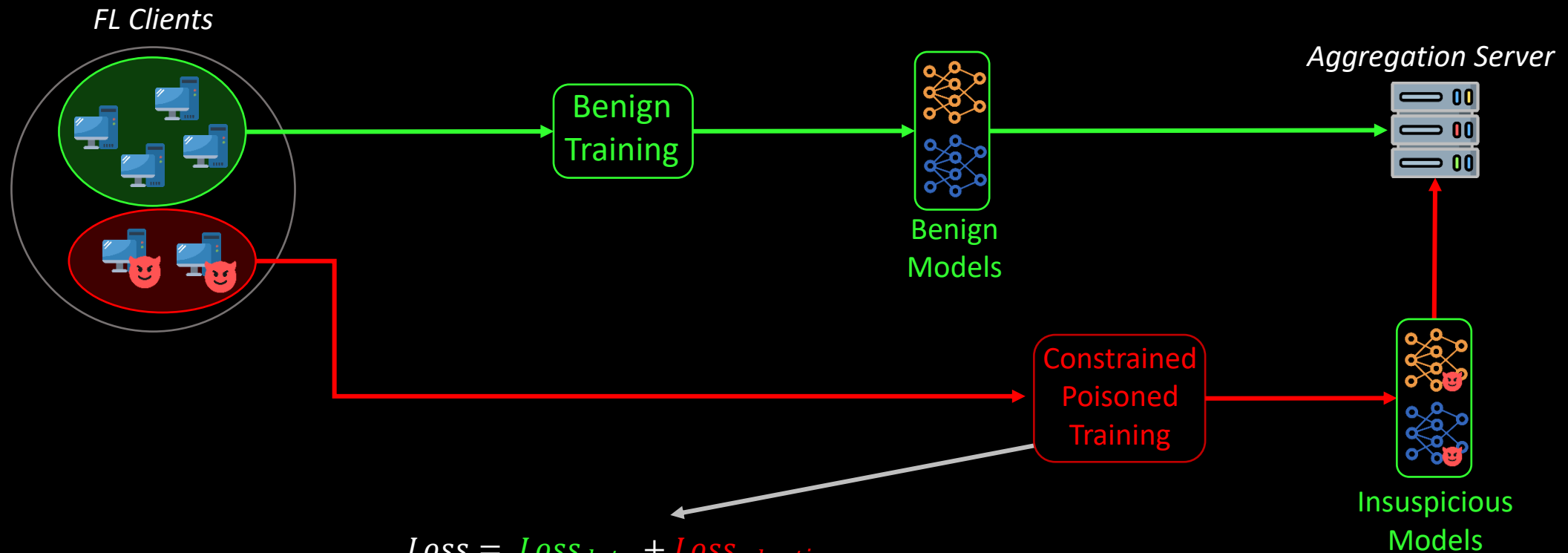
Classic Adaption Workflow



Classic Adaption Workflow



Classic Adaption Workflow



$$Loss = LOSS_{data} + LOSS_{adaption}$$

Trade-Off [1]

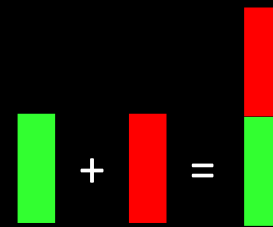
$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

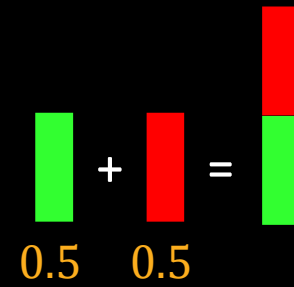
Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



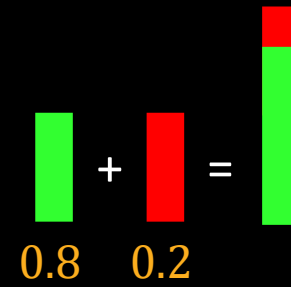
Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error $\alpha = [0.1, \dots, 0.9]$

Challenges

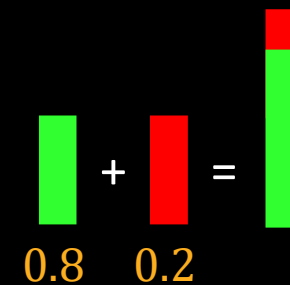
$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

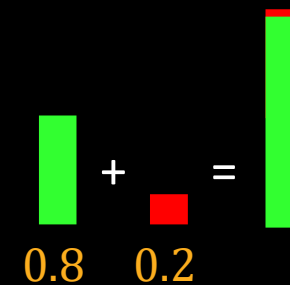
- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$



Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$



Challenges

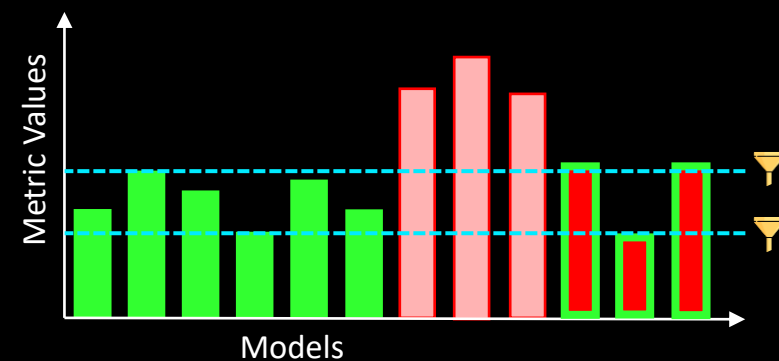
$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $LOSS_{data}$ and $LOSS_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

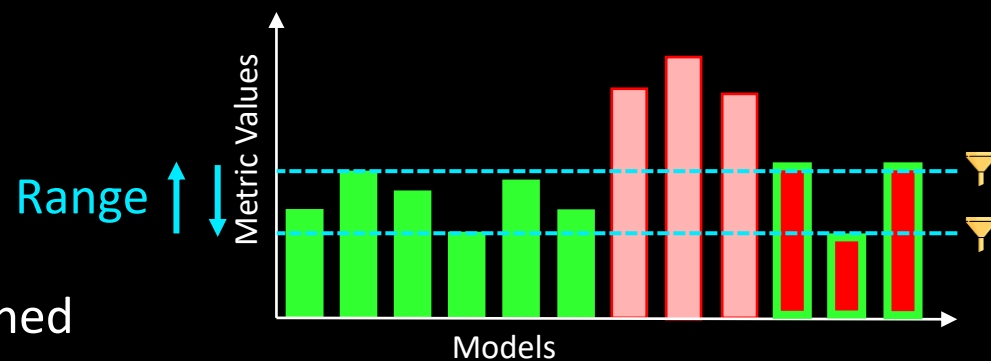
- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined



Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

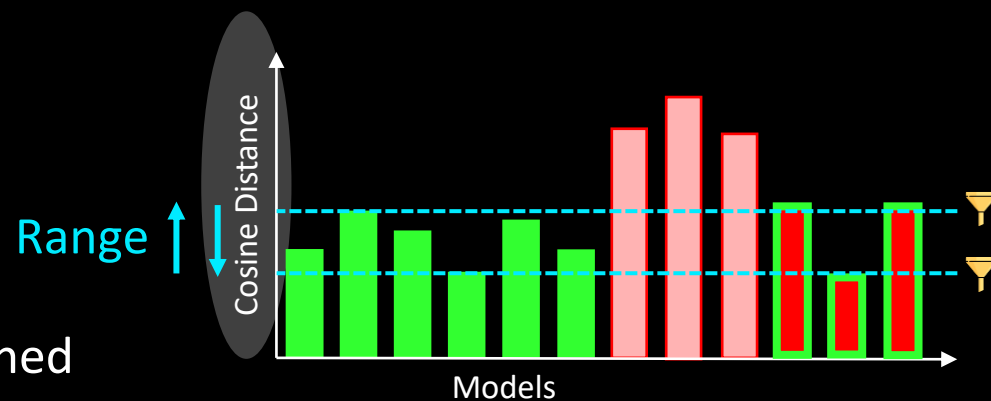
- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined



Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$

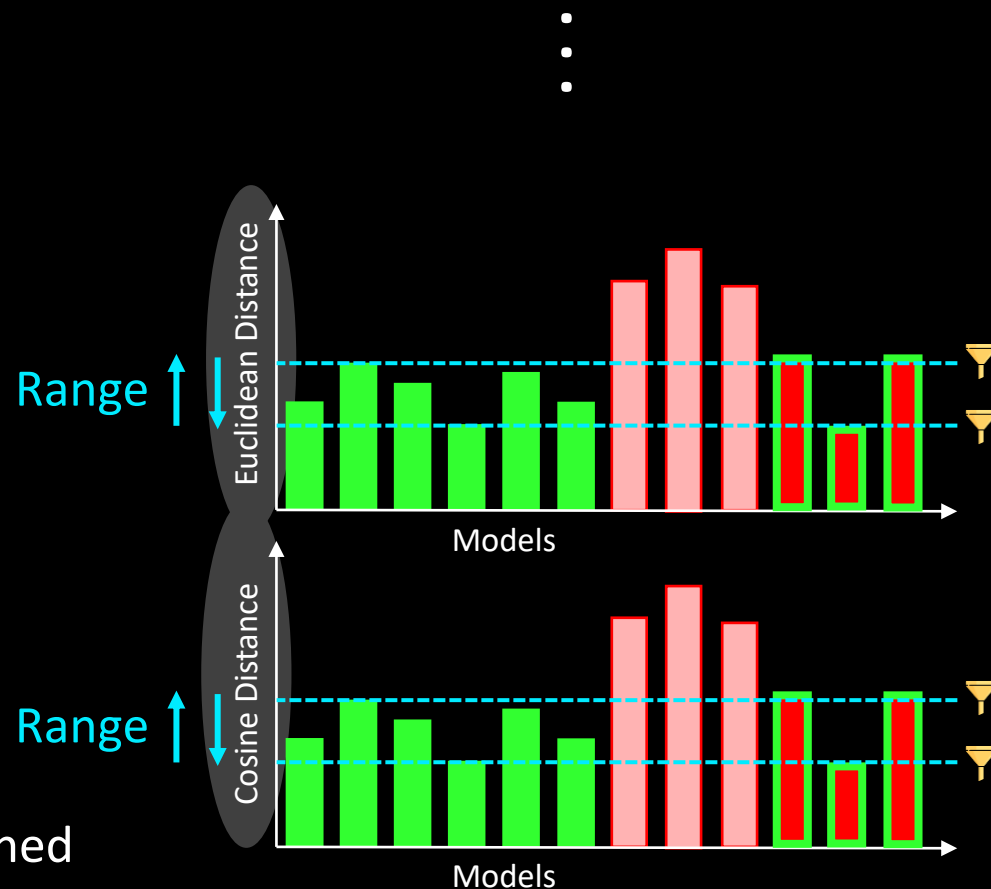
- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined



Challenges

$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $LOSS_{data}$ and $LOSS_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined



Challenges

$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $LOSS_{data}$ and $LOSS_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined

Challenges



$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption} \longrightarrow$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $LOSS_{data}$ and $LOSS_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

Augmented Lagrangian Method
for constraint optimization problems
&
adapted for **multiple range constraints**

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined

Challenges

AutoAdapt

$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption}$$



$$Loss = LOSS_{data} + LOSS_{AutoAdapt}$$

$$LOSS_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta LOSS_j)|^2 - \gamma_j^2)$$

Iterative update of γ_j for each $LOSS_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta LOSS_j, & \text{if } LOSS_j \geq 0 \\ 0, & \text{if } LOSS_j < 0 \end{cases}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $LOSS_{data}$ and $LOSS_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined

Challenges

AutoAdapt

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

- ① Fixed α during training
- ② Manual adjustment of α via trial and error
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$
- ④ Handling of **multiple inequality (range)** constraints undefined

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

- 1 Fixed α during training

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

① Fixed α during training



γ_j individual for each $Loss_j$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

① Fixed α during training



γ_j individual for each $Loss_j$

γ_j automatically adjusted during training

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

- ① Fixed α during training



γ_j individual for each $Loss_j$

γ_j automatically adjusted during training



- ② Manual adjustment of α via trial and error



Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

- ① Fixed α during training



γ_j individual for each $Loss_j$

γ_j automatically adjusted during training



- ② Manual adjustment of α via trial and error ✓
- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$ ✓

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

AutoAdapt

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

- ① Fixed α during training



γ_j individual for each $Loss_j$

γ_j automatically adjusted during training



- ② Manual adjustment of α via trial and error 



β is automatically initialized

- ③ Ill-Conditioning of $Loss_{data}$ and $Loss_{adaption}$ 

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

4

Handling of **multiple inequality**
(range) constraints undefined

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

4

Handling of **multiple inequality**
(range) constraints undefined

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality**

(range) constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7 \longrightarrow$ Unsatisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

4

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta Loss_j)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta \cdot -5)|^2 - \gamma_j^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta \cdot Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + \beta \cdot -5)|^2 - \gamma_j^2)$$

- ④ Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta \cdot Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, 6 + 1 - 5)|^2 - 6^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Challenges

AutoAdapt

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + 1 - 5)|^2 - 6^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|\max(0, \gamma_j + 1 - 5)|^2 - 0^2)$$

- ④ Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = \frac{1}{2\beta} \sum_{j=1}^m (|0|^2 - 0^2)$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



AutoAdapt

$$Loss = Loss_{data} + Loss_{AutoAdapt}$$

$$Loss_{AutoAdapt} = 0$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Challenges

$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$



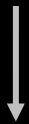
AutoAdapt

$$Loss = Loss_{data} + 0$$

$$Loss_{AutoAdapt} = 0$$

④

Handling of **multiple inequality (range)** constraints undefined



Modeling \leq constraints: $Loss_j \leq 0$

Iterative update of γ_j for each $Loss_j$ before optimizer step:

$$\gamma_j = \begin{cases} \gamma_j + \beta Loss_j, & \text{if } Loss_j \geq 0 \\ 0, & \text{if } Loss_j < 0 \end{cases}$$

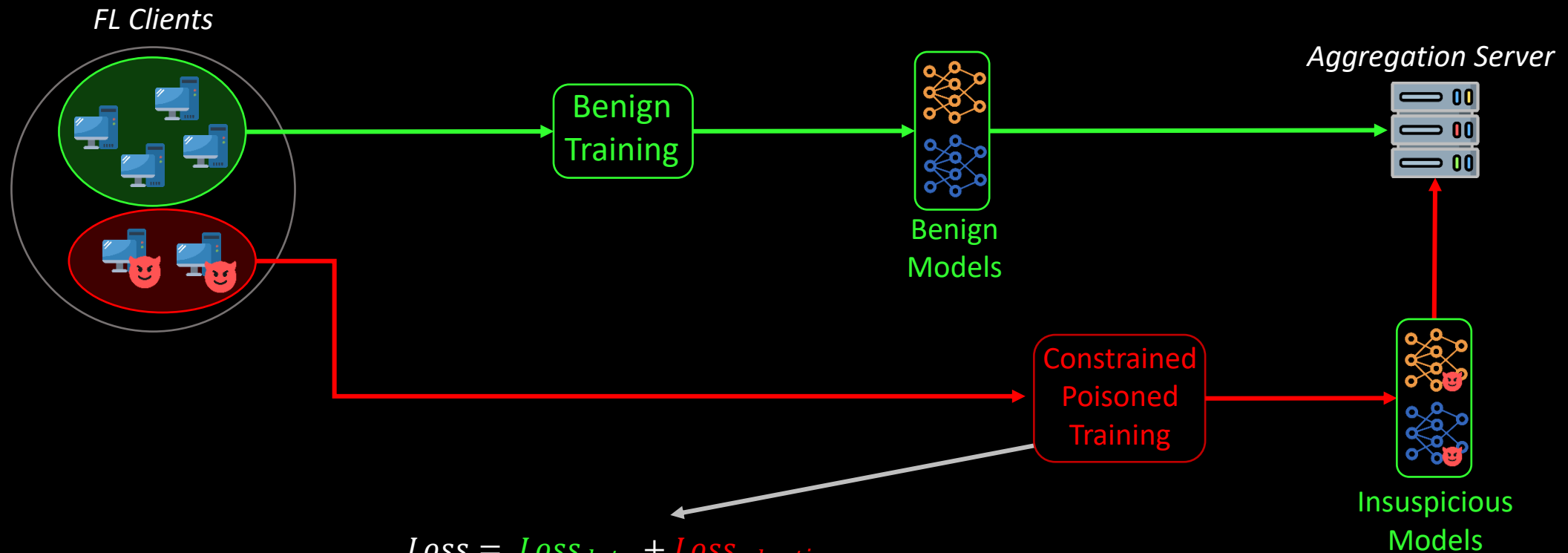
$Loss_j = 7$ \longrightarrow Unsatisfied Constraint

$Loss_j = -5$ \longrightarrow Satisfied Constraint

$$\gamma_j = 6$$

$$\beta = 1$$

Classic Adaption Workflow

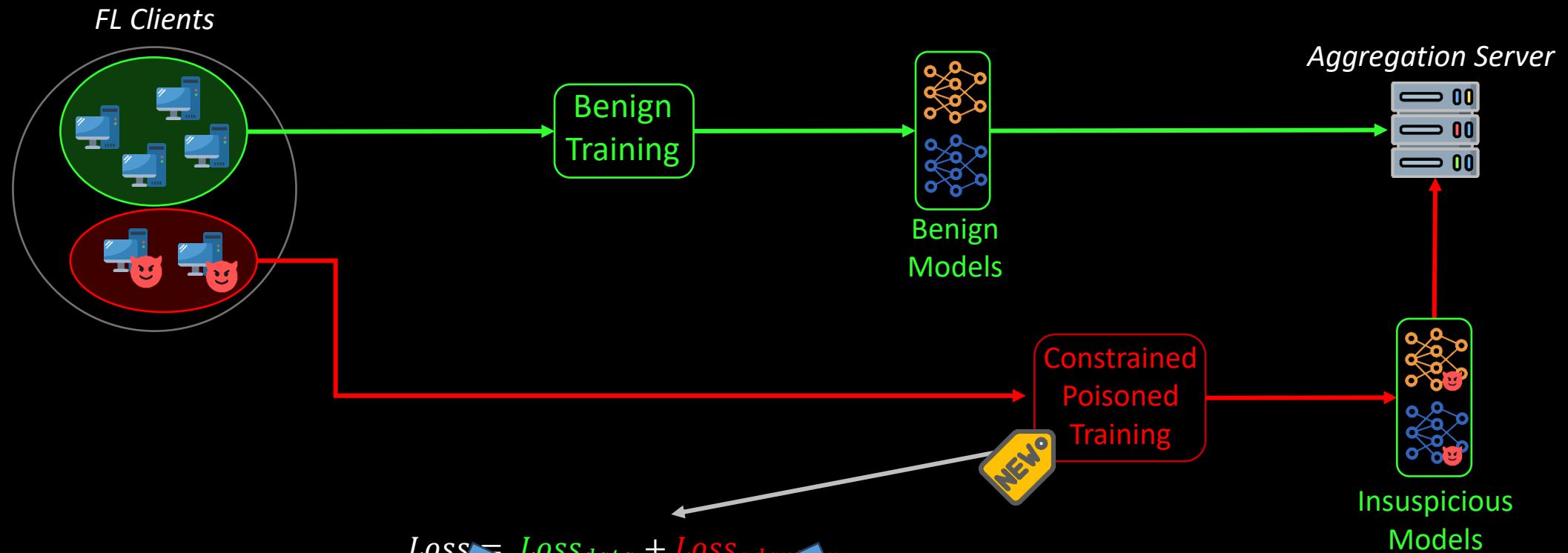


$$Loss = LOSS_{data} + LOSS_{adaption}$$

Trade-Off

$$Loss = \alpha \cdot LOSS_{data} + (1 - \alpha) \cdot LOSS_{adaption}$$

AutoAdapt Workflow

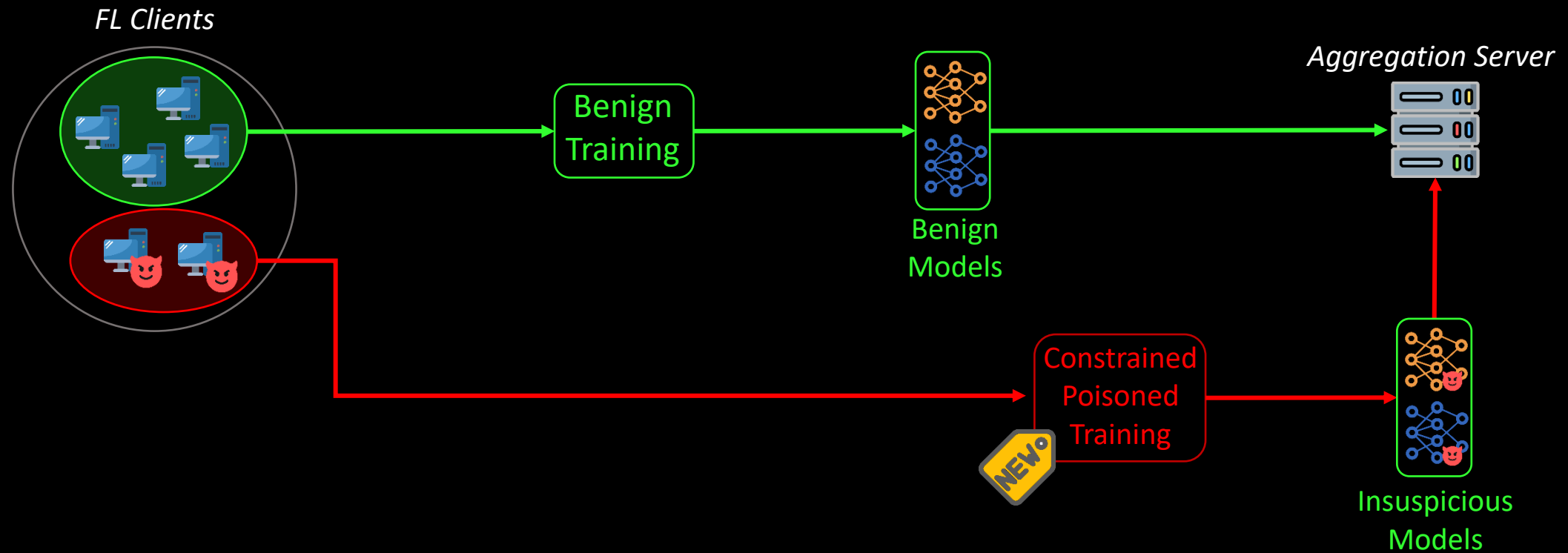


~~$$Loss = Loss_{data} + Loss_{adaptation}$$~~

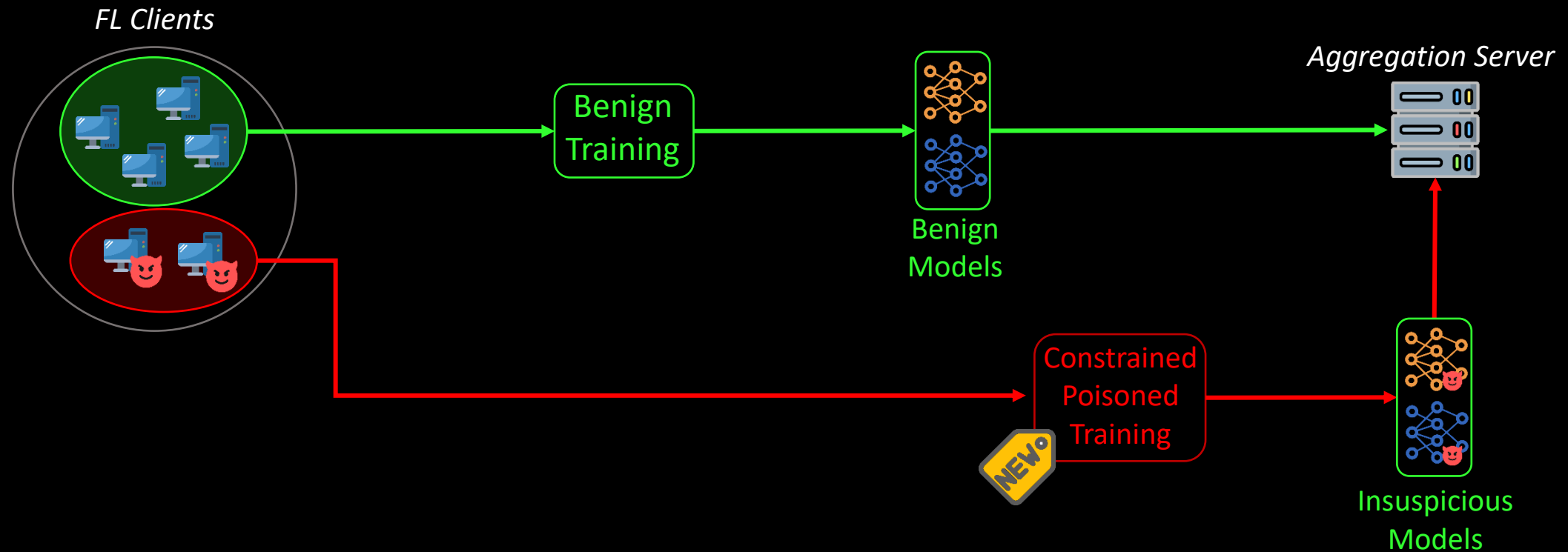
Trade-Off

~~$$Loss = \alpha \cdot Loss_{data} + (1 - \alpha) \cdot Loss_{adaption}$$~~

AutoAdapt Workflow



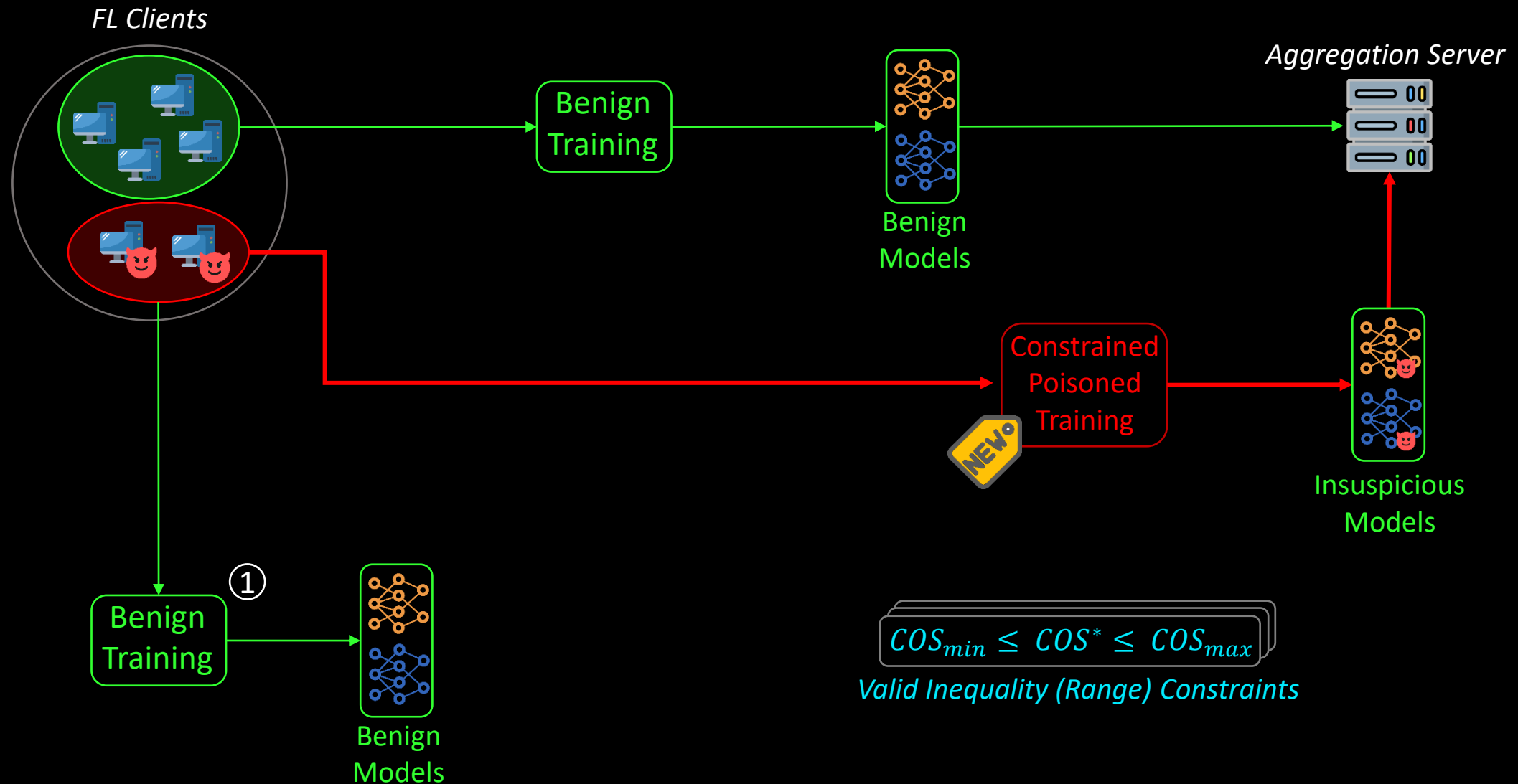
AutoAdapt Workflow



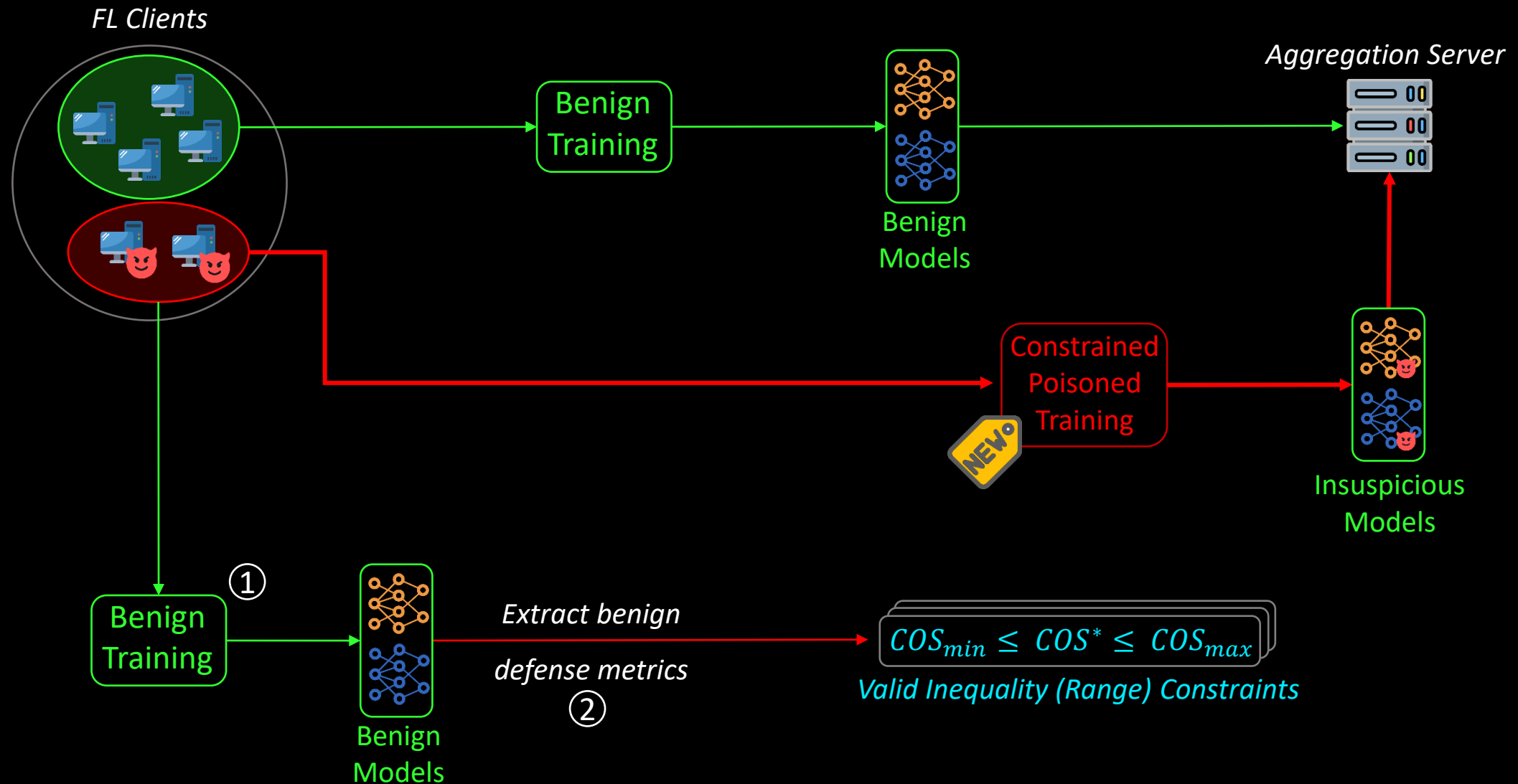
$$COS_{min} \leq COS^* \leq COS_{max}$$

Valid Inequality (Range) Constraints

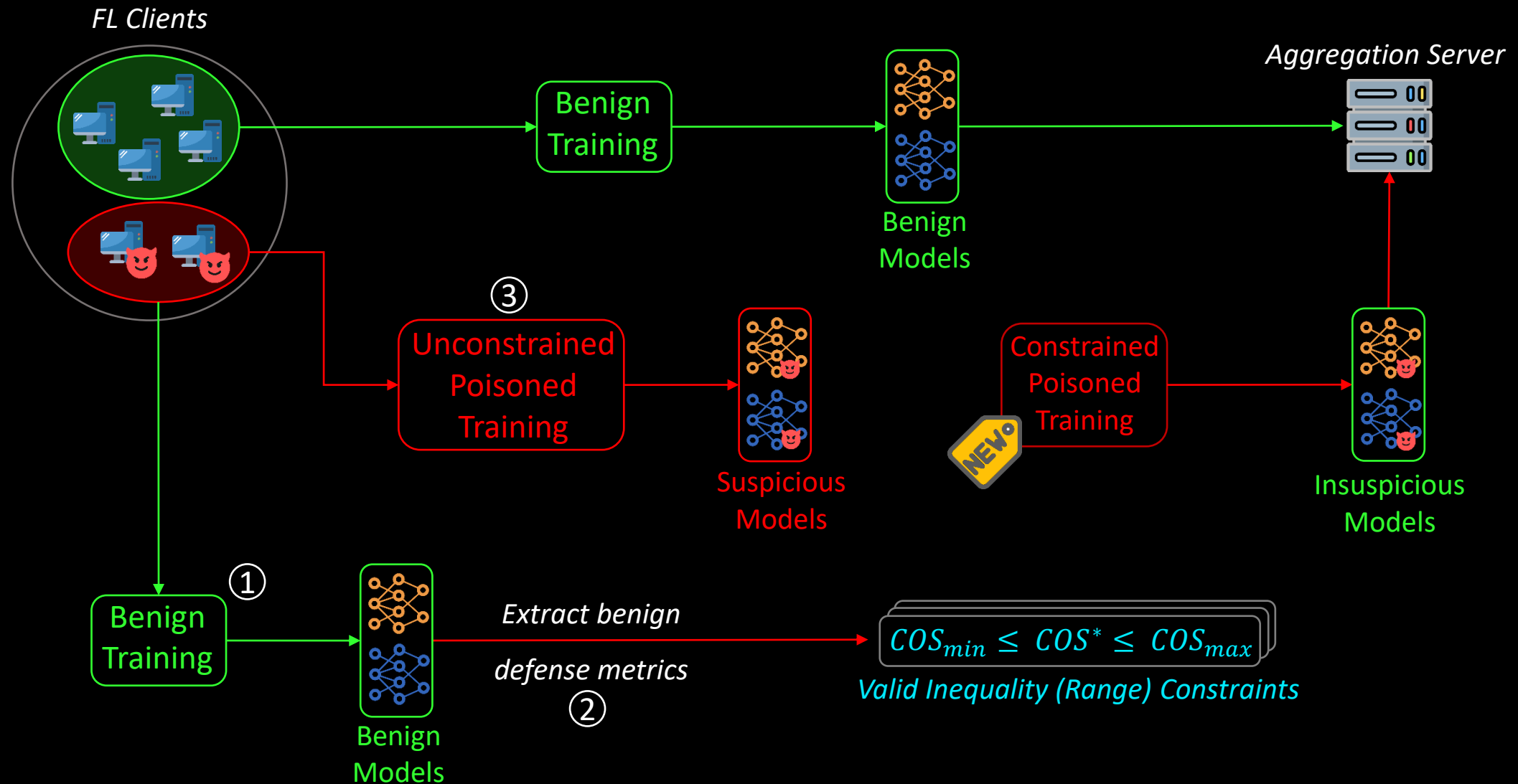
AutoAdapt Workflow



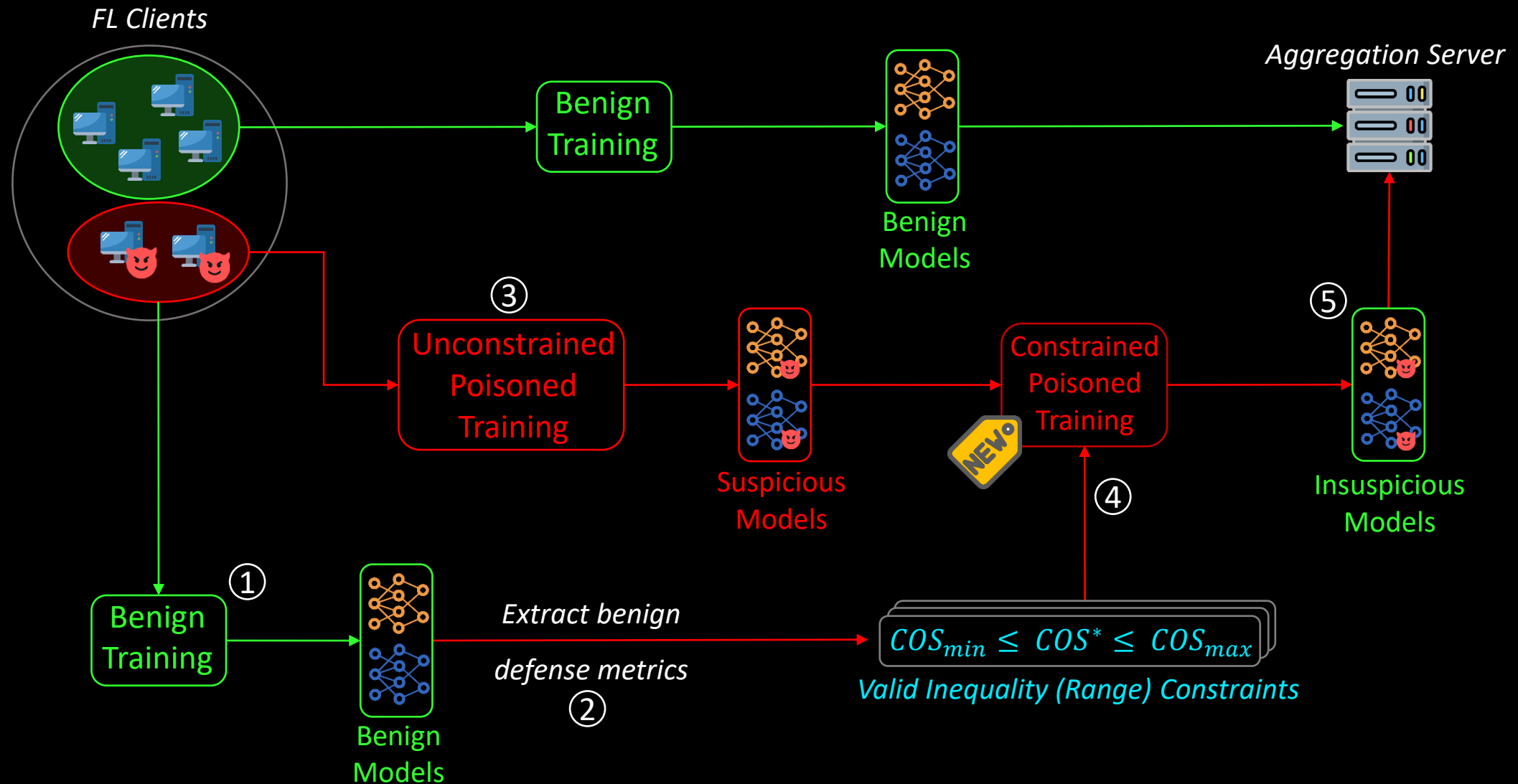
AutoAdapt Workflow



AutoAdapt Workflow



AutoAdapt Workflow



AutoAdapt – Experiments

Reported Setup:

- CIFAR-10
- ResNet-18
- 20 client
- 2560 Samples per client
- Poison Data Rate 0.1
- Poison Model Rate 0.45
- Semantic Backdoor

Conducted Experiments:

- 3 models
- 3 datasets
- 3 backdoors
- 4 defenses
- Different non-IID scenarios

AutoAdapt – Experiments

Model-Wise Metrics

- 1 Metric / Model

- 2 Constraints (1 Range)

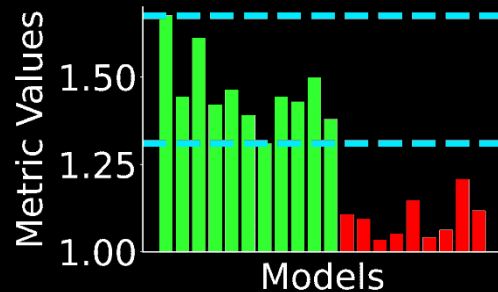


AutoAdapt – Experiments

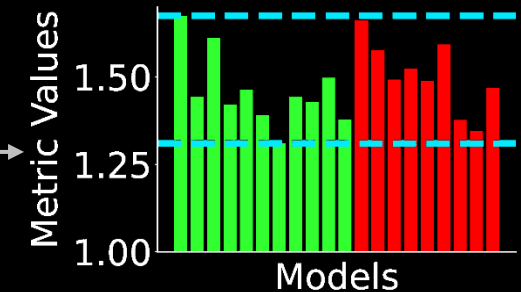
Model-Wise Metrics

- 1 Metric / Model

- 2 Constraints (1 Range)



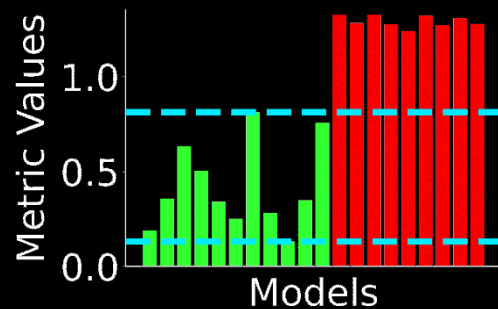
AutoAdapt



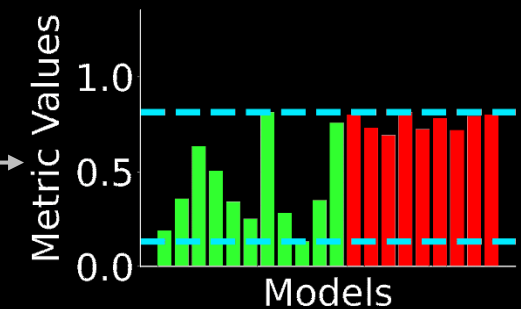
Layer-Wise Metrics (Last Layer)

- 14 Metrics / Model

- 28 Constraints (14 Ranges)



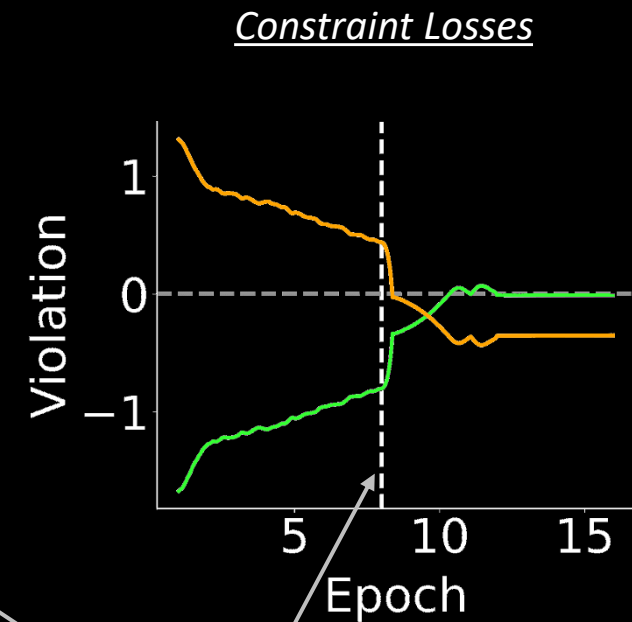
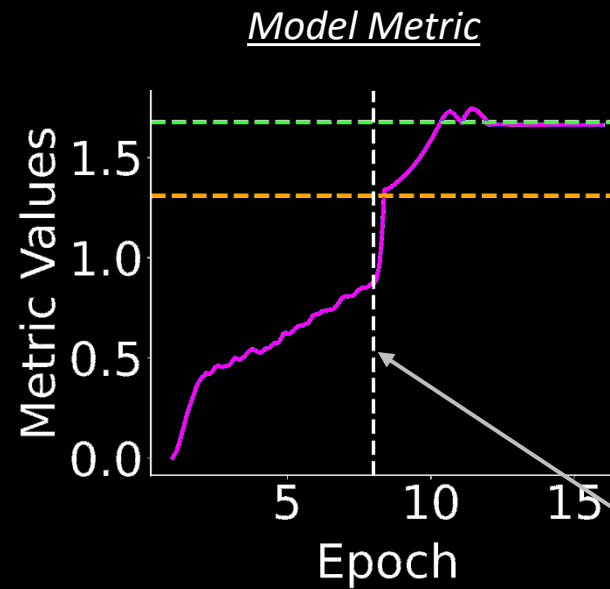
AutoAdapt



AutoAdapt – Experiments

Adaption Progression

- Adaption within 1-3 Epochs
- Alignment to one range threshold



Metric Value

----- Upper Range Constraint



----- Lower Range Constraint

Start of Constraint Training

AutoAdapt – Experiments

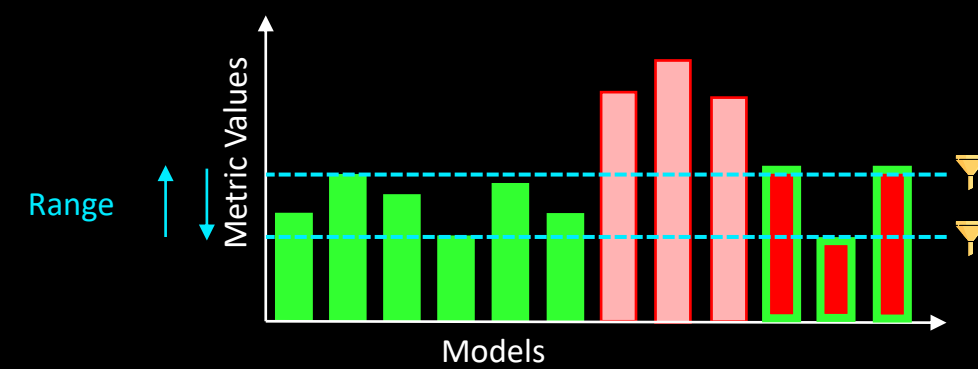
Runtime

- Faster FL Defense Testing
- Benefit for all Researchers

No Adaption	$\alpha = 0.9$	$\alpha = [0.1, \dots, 0.9]$	 AutoAdapt 3 epochs	 AutoAdapt 1 epoch
10.87s	25.46s	229.11s	20.50s	14.62s
+ 0 %	+ 134 %	+ 2007 %	+ 88 %	+ 34 %
Saved Time		+ 0 %	- 91 %	- 94 %
Speed-Up		x 0	11x faster	15x faster

Conclusion

- ① Replace fixed α with dynamic γ in Augmented Lagrangian
- ② Implicit handling of **multiple inequality (range)** constraints



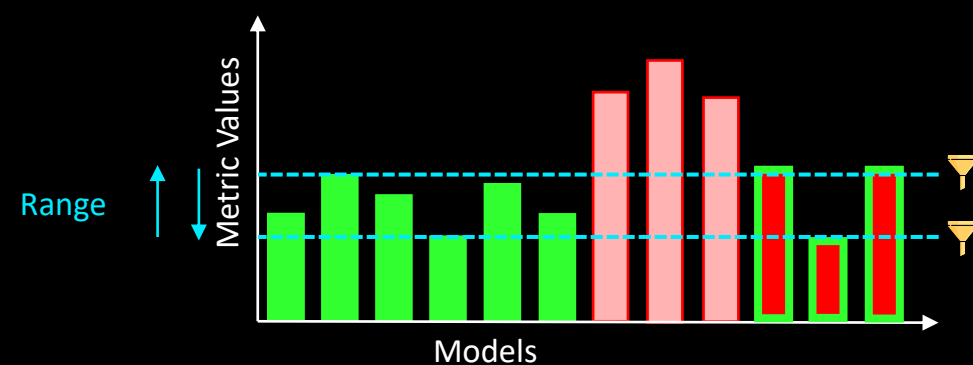
Conclusion

- ① Replace fixed α with dynamic γ in Augmented Lagrangian
- ② Implicit handling of **multiple inequality (range)** constraints

→ Metric adaption on a model-wise and layer-wise level

→ Successful adaption within 1-3 training epochs

→ 11-15x faster FL defense testing



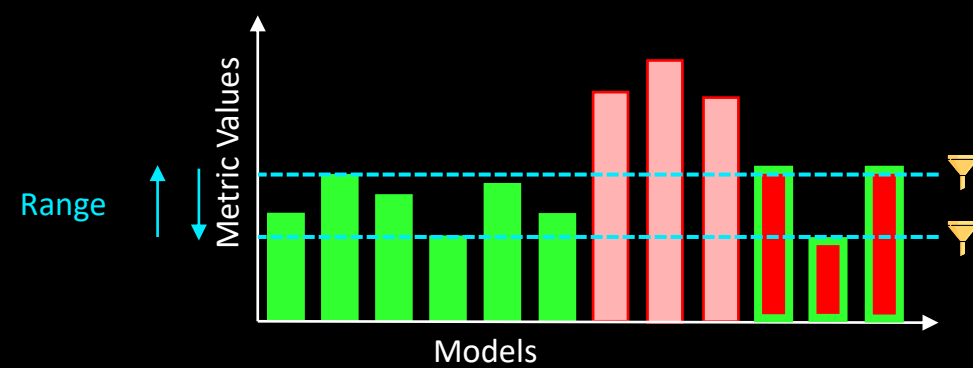
Conclusion

- ① Replace fixed α with dynamic γ in Augmented Lagrangian
- ② Implicit handling of **multiple inequality (range)** constraints

→ Metric adaption on a model-wise and layer-wise level

→ Successful adaption within 1-3 training epochs

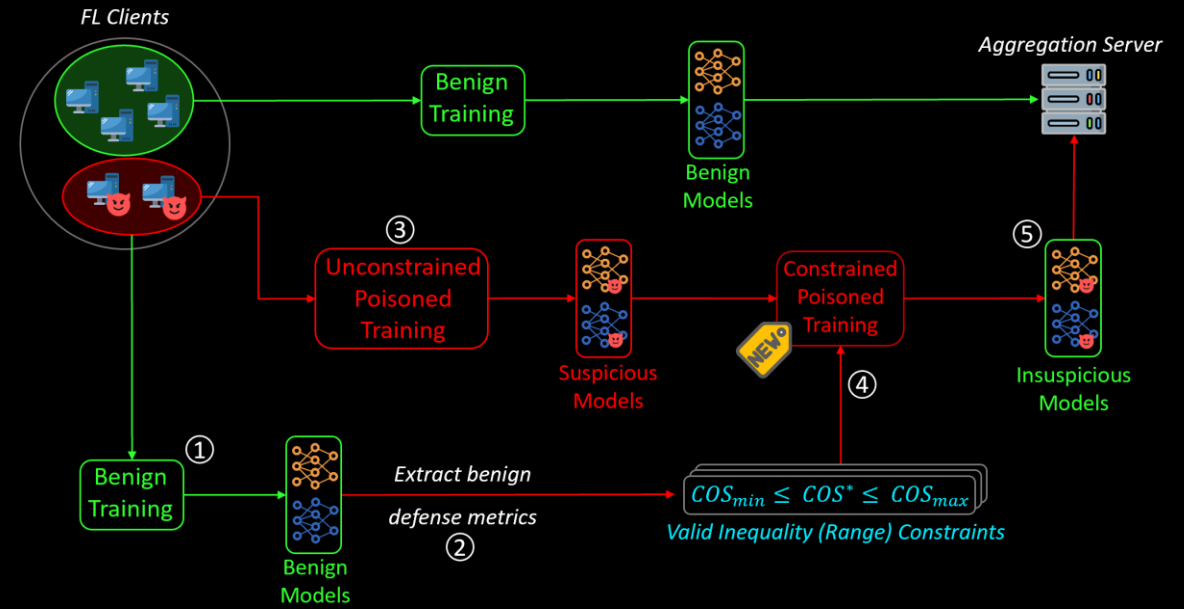
→ 11-15x faster FL defense testing



AutoAdapt: A useful tool to evaluate the robustness of FL poisoning defenses

Thank you!!11!!1

Any Questions?

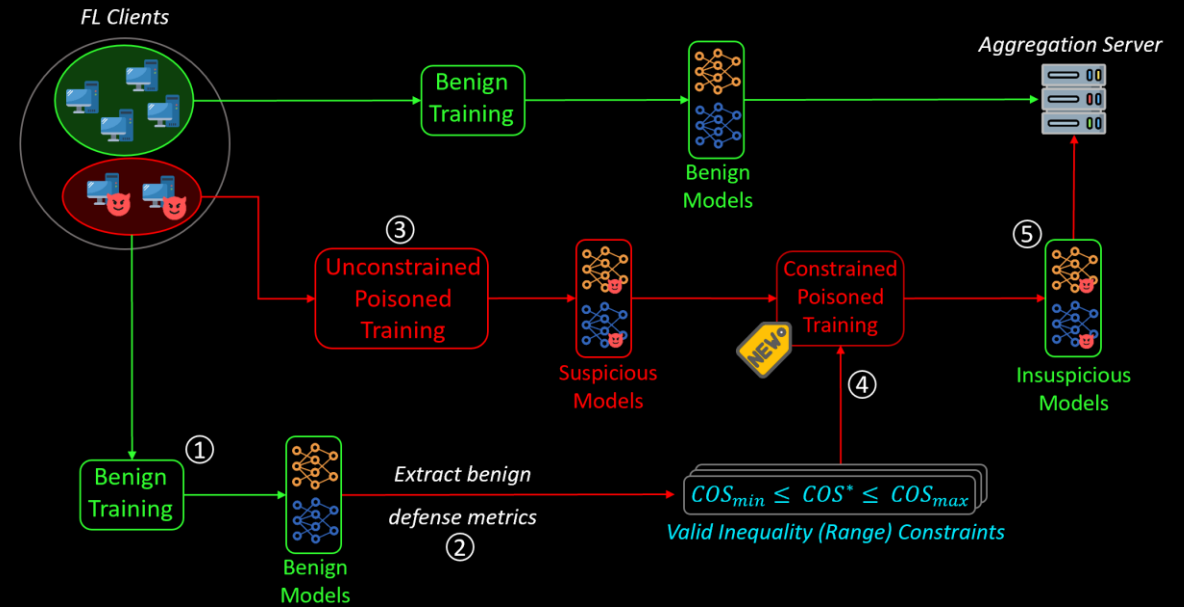


Torsten Krauß, Jan König, Alexandra Dmitrienko, Christian Kanzow

University of Würzburg

Thank you!!11!!1

Any Questions?



Torsten Krauß, Jan König, Alexandra Dmitrienko, Christian Kanzow

University of Würzburg

