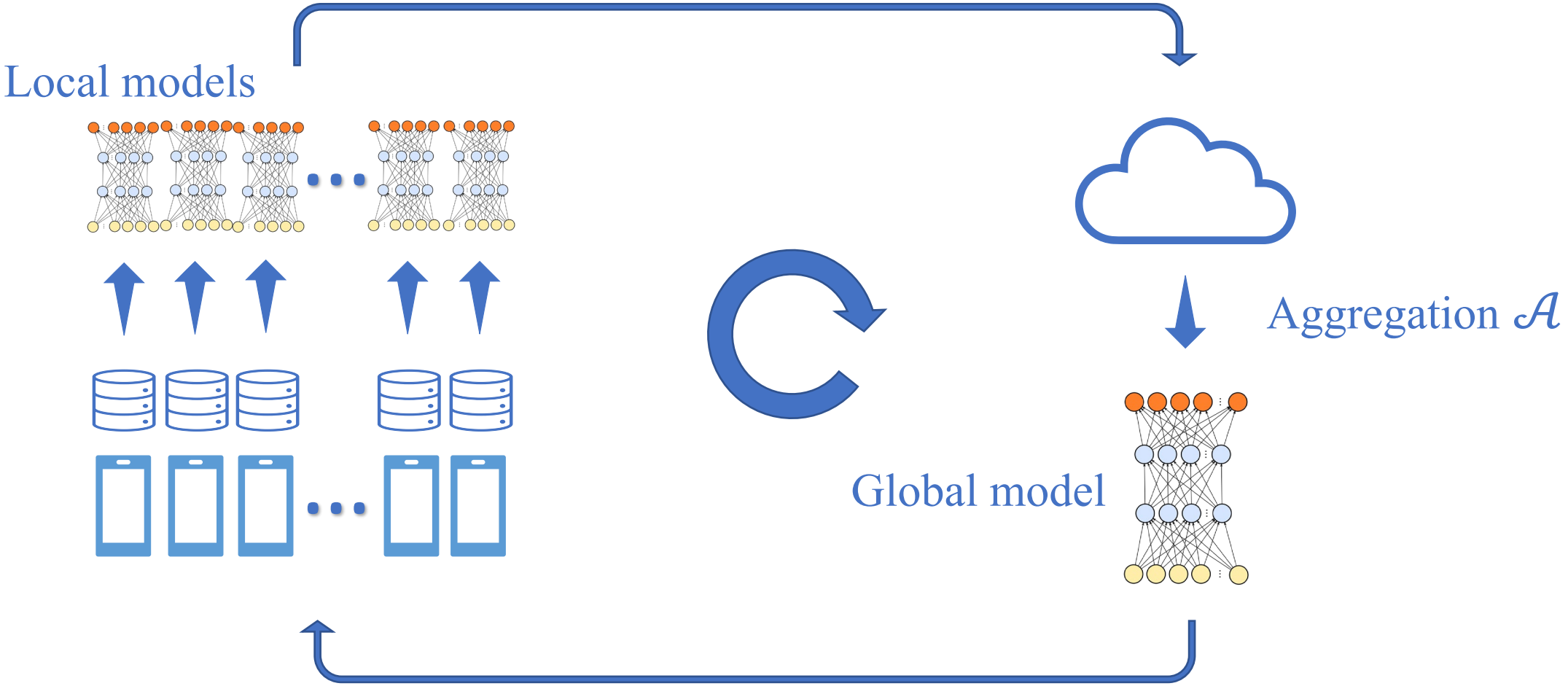
Local Model Poisoning Attacks to Byzantine-Robust Federated Learning

Minghong Fang*, **Xiaoyu Cao*** , Jinyuan Jia, and Neil Gong

Duke UNIVERSITY

IOWA STATE UNIVERSITY

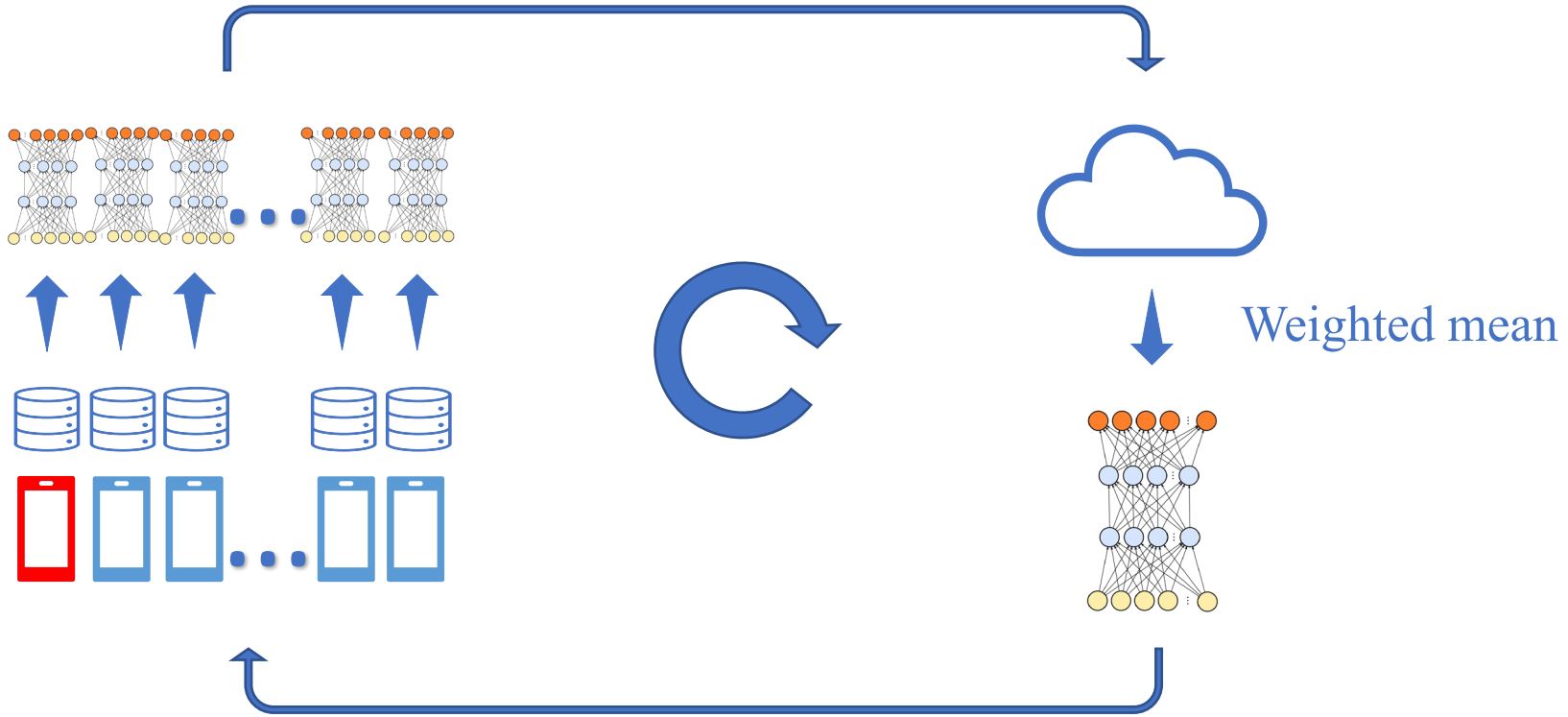
Federated Learning: Collaborative Learning with Decentralized Data



Real-world Applications

Federated Learning is Vulnerable to Attacks

A single malicious worker can arbitrarily manipulate the global model.



Byzantine-robust Federated Learning

- Byzantine-robust aggregation rules as a defense.
  - Learn a good global model given bounded Byzantine workers.
- Main idea:
  - Filter out statistical outliers.
- Existing methods:
  - Krum, Trimmed-mean, coordinate-wise median, Bulyan, …

Limitations

- Asymptotic bounds
  - Order-optimal bounds on parameters.
  - No guarantee on classification results.

- Strong assumptions
  - IID data
  - Strongly-convex

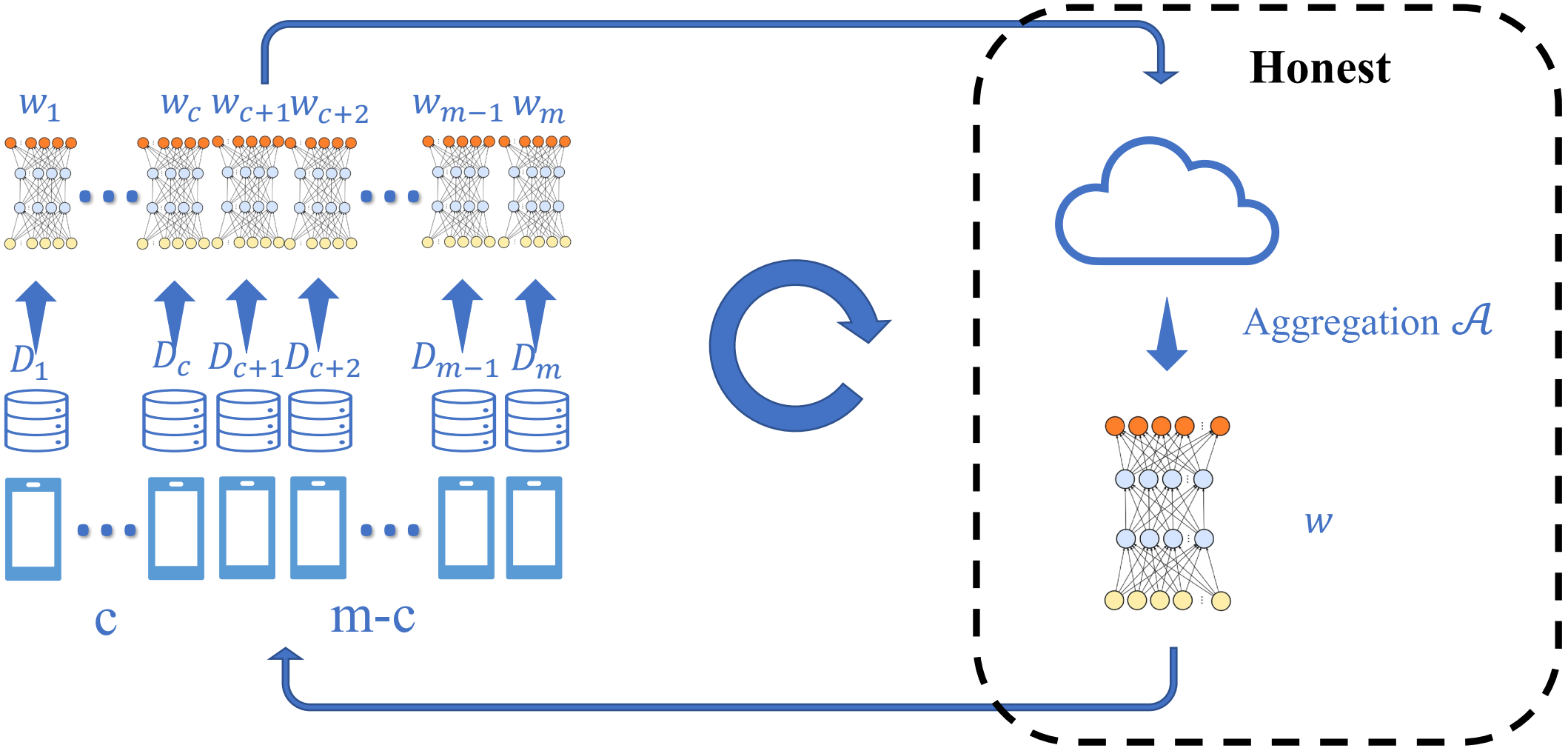
Our work

Increase classification error rates of global models by sending manipulated local models.

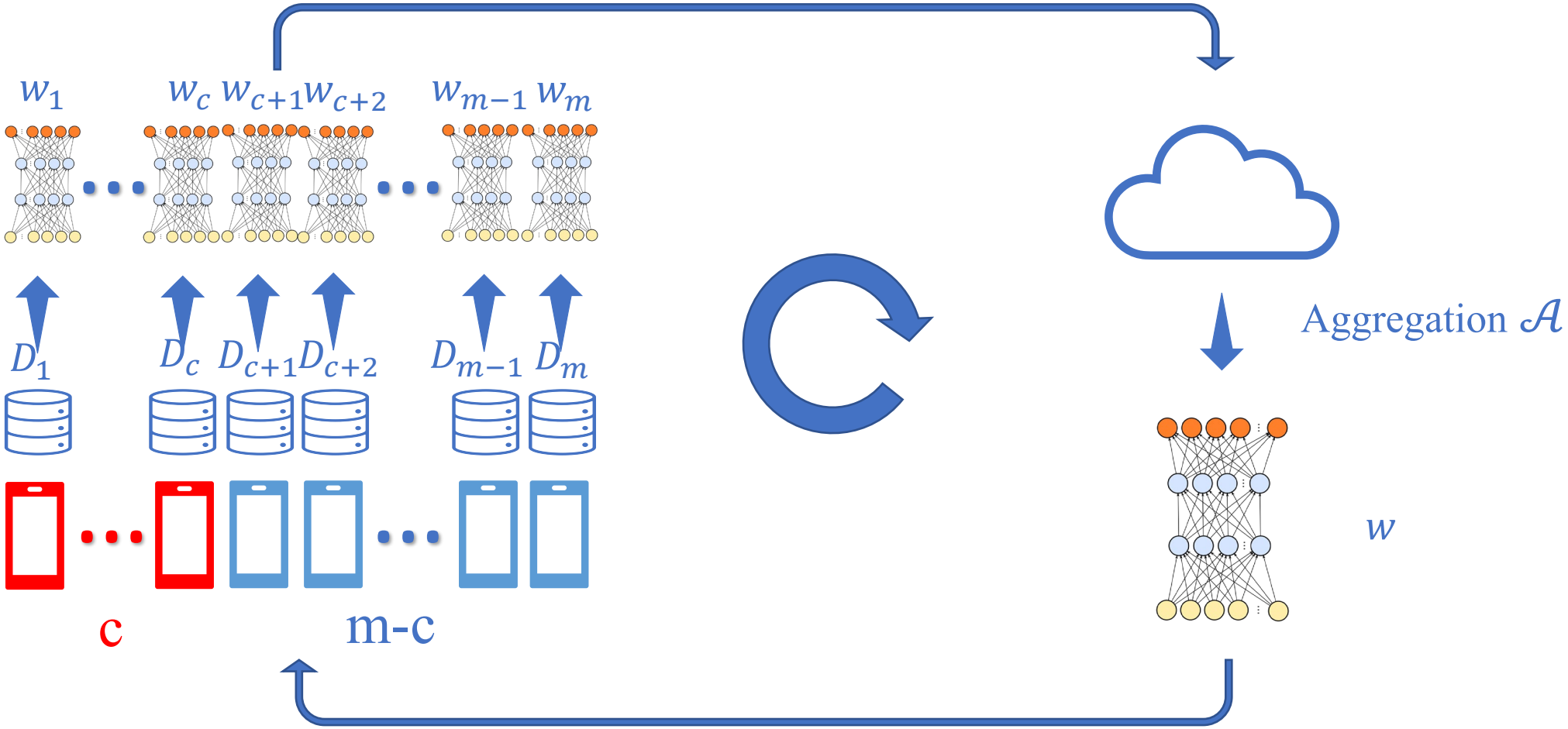
Threat Model

- Attacker's goal:
  - High classification error rate.
- Attacker's capability:
  - Controls some workers.
  - Sends arbitrary local models.
- Attacker's knowledge:
  - Compromised workers: everything.
  - Server: know aggregation or not.
  - Benign workers: know model/data or not.

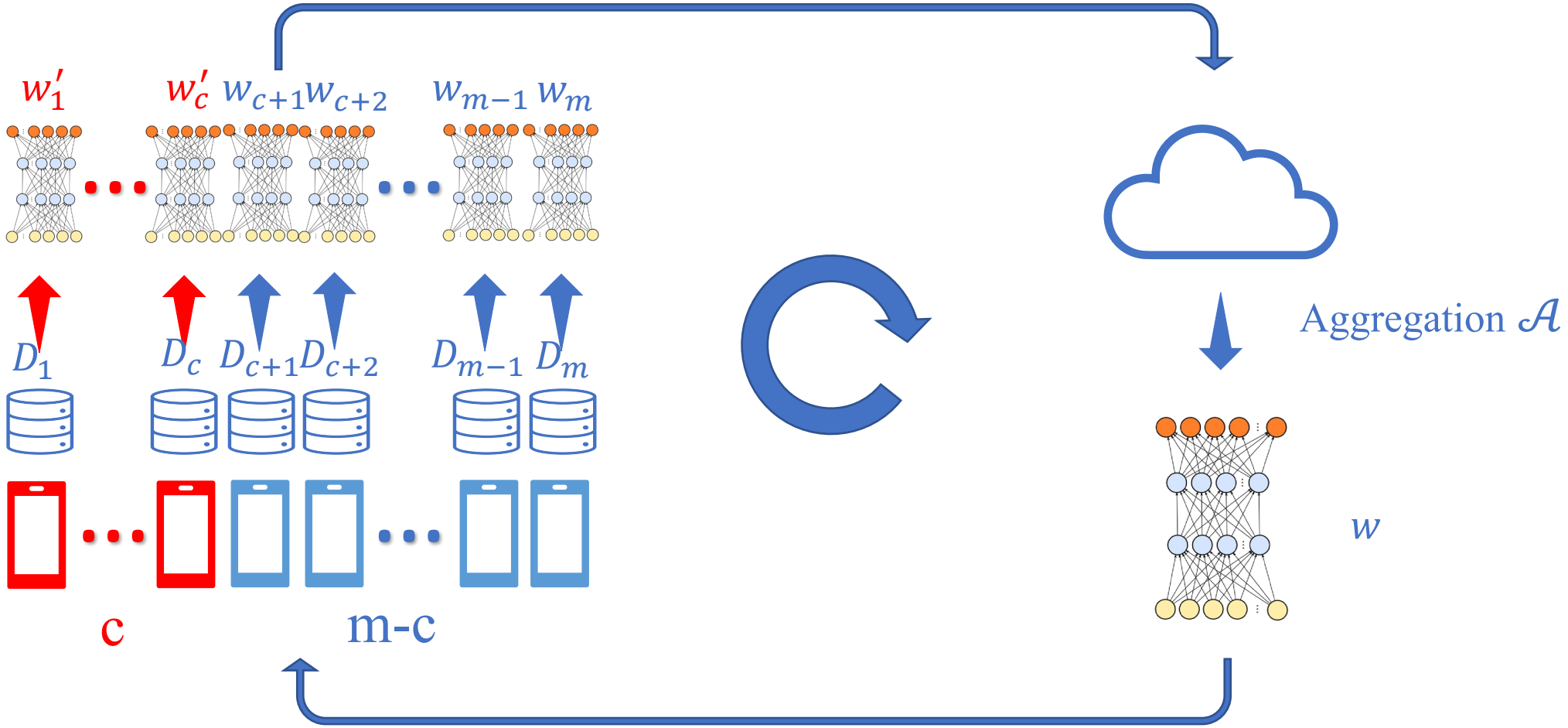
Threat Model



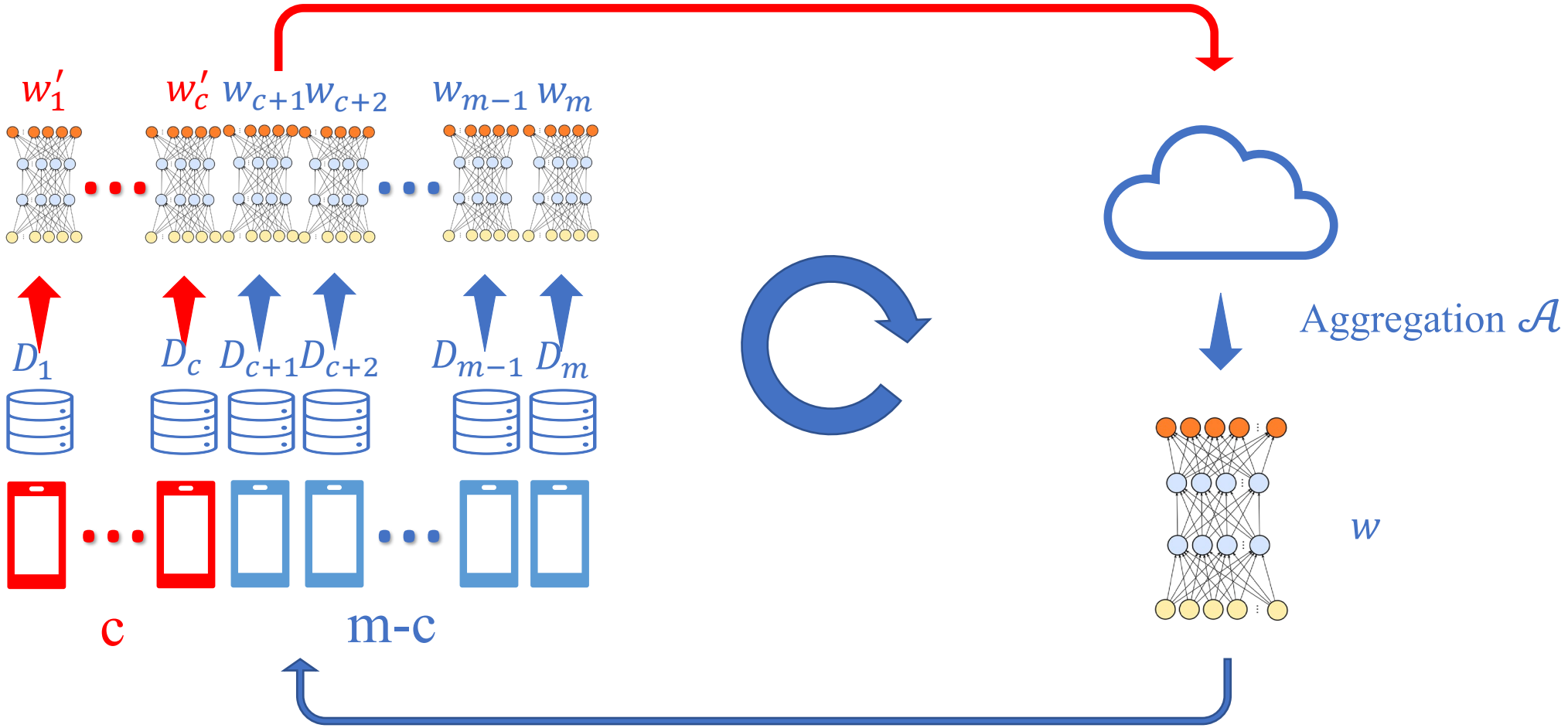
Threat Model



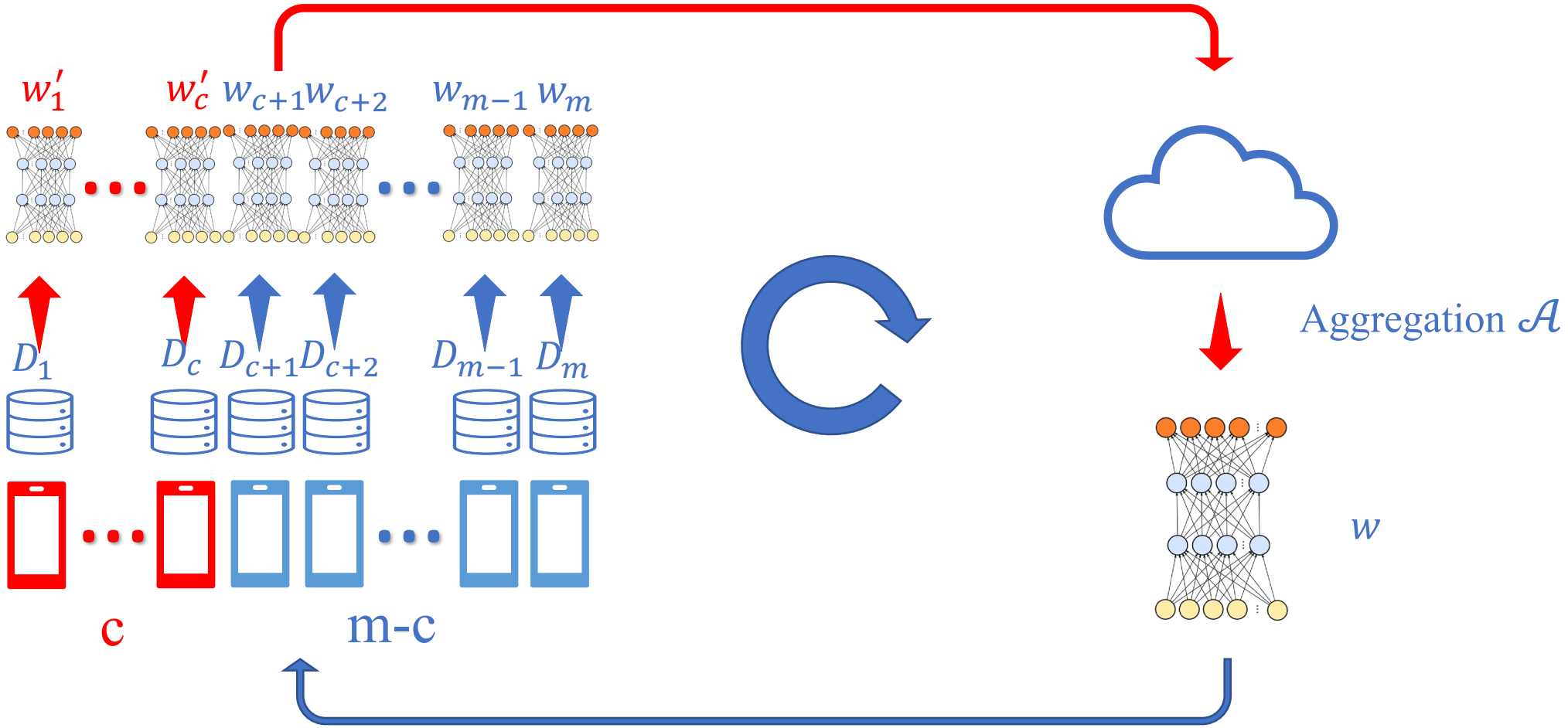
Threat Model



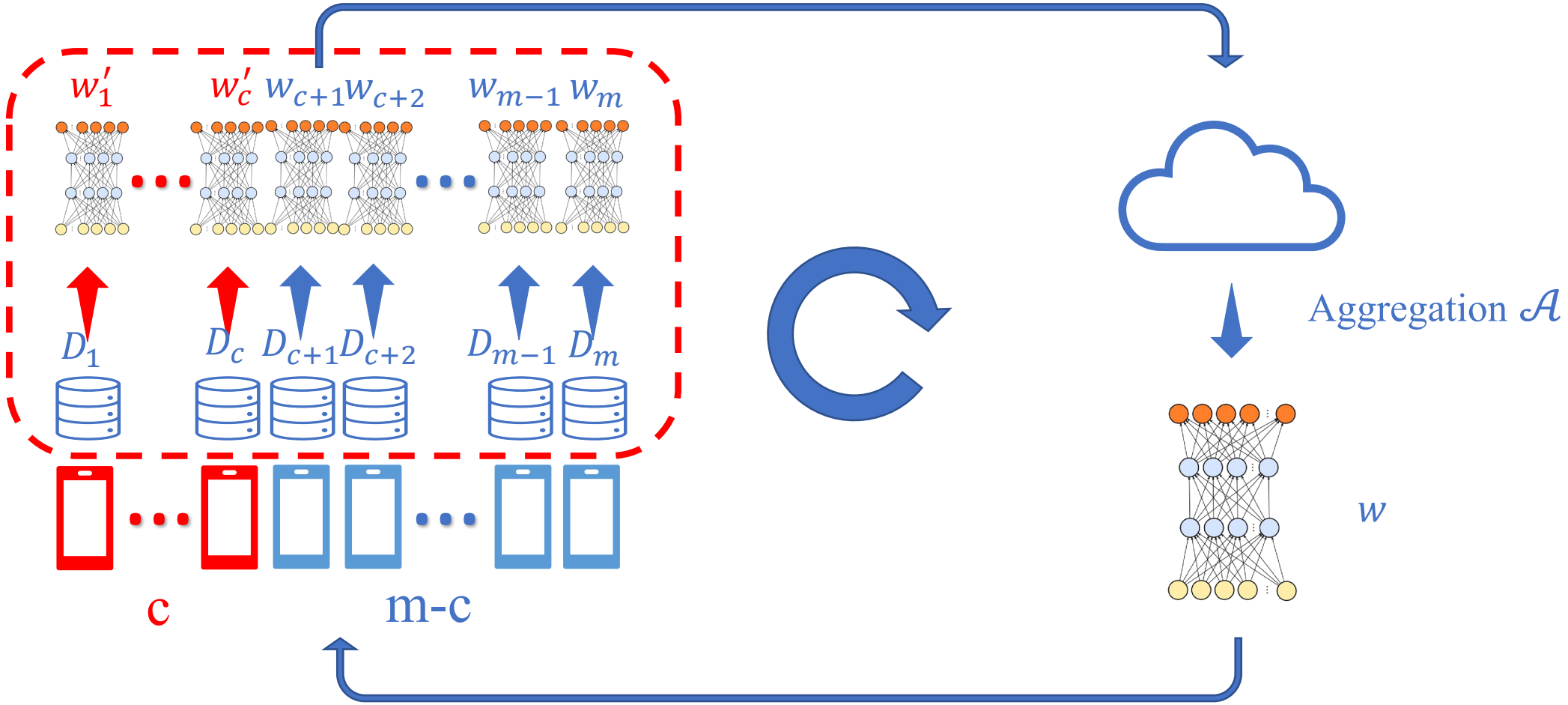
Threat Model



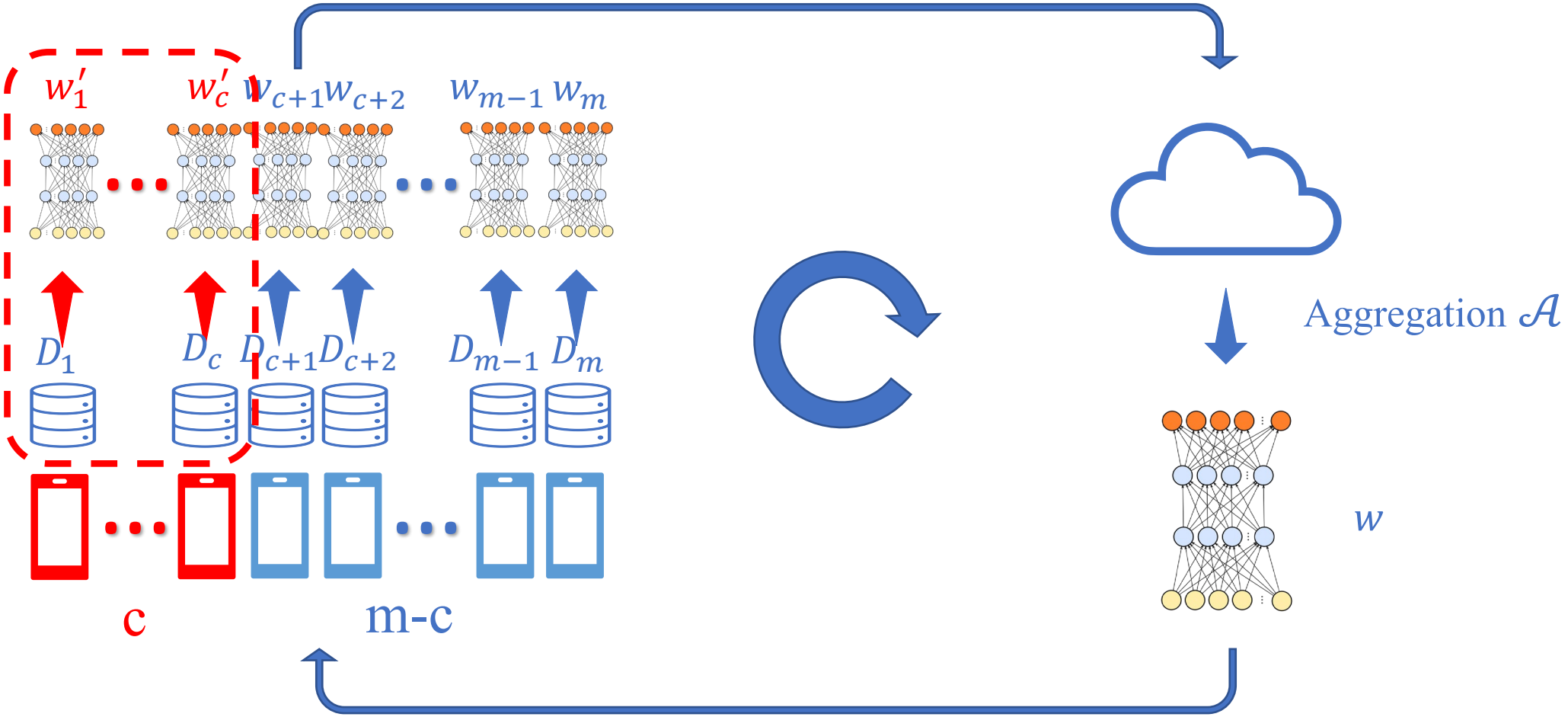
Threat Model



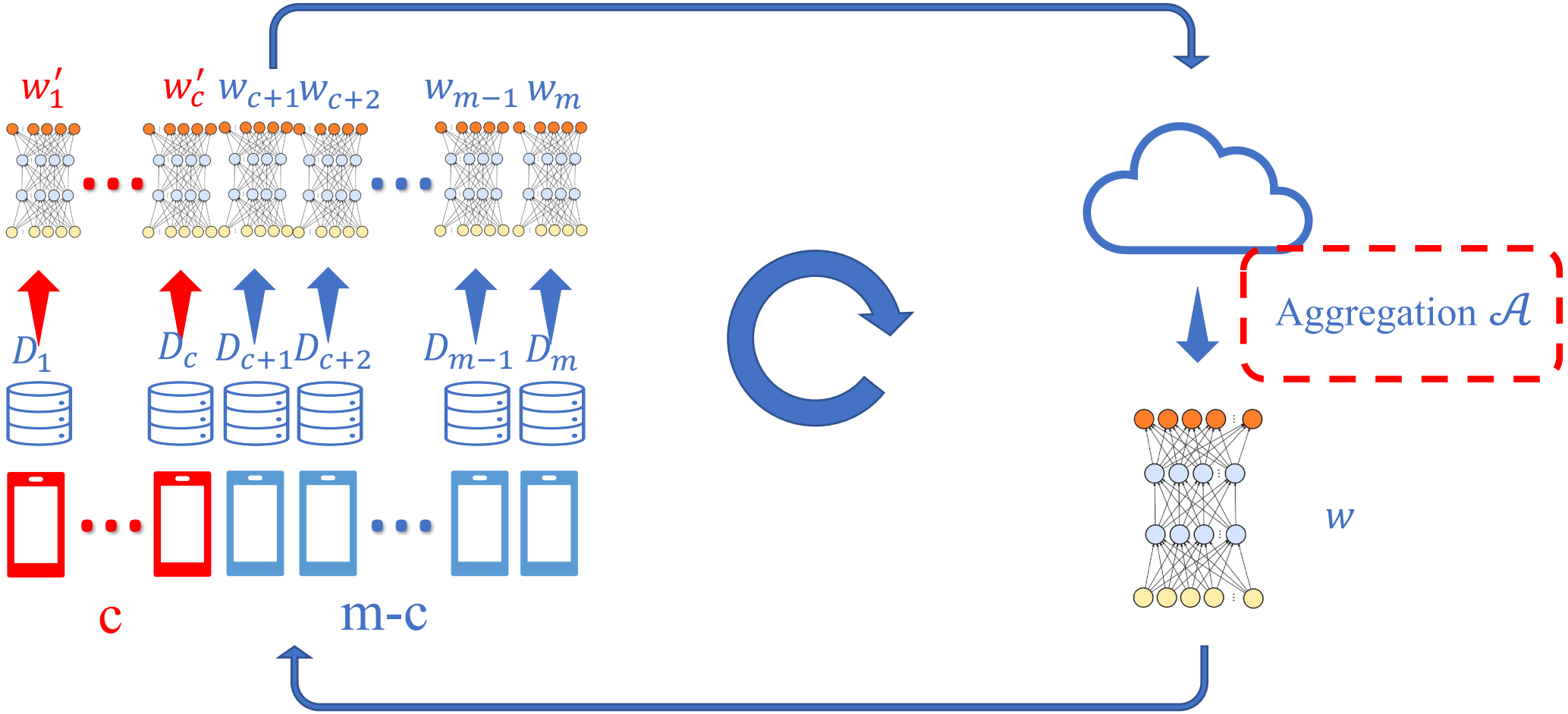
Full-knowledge Attack



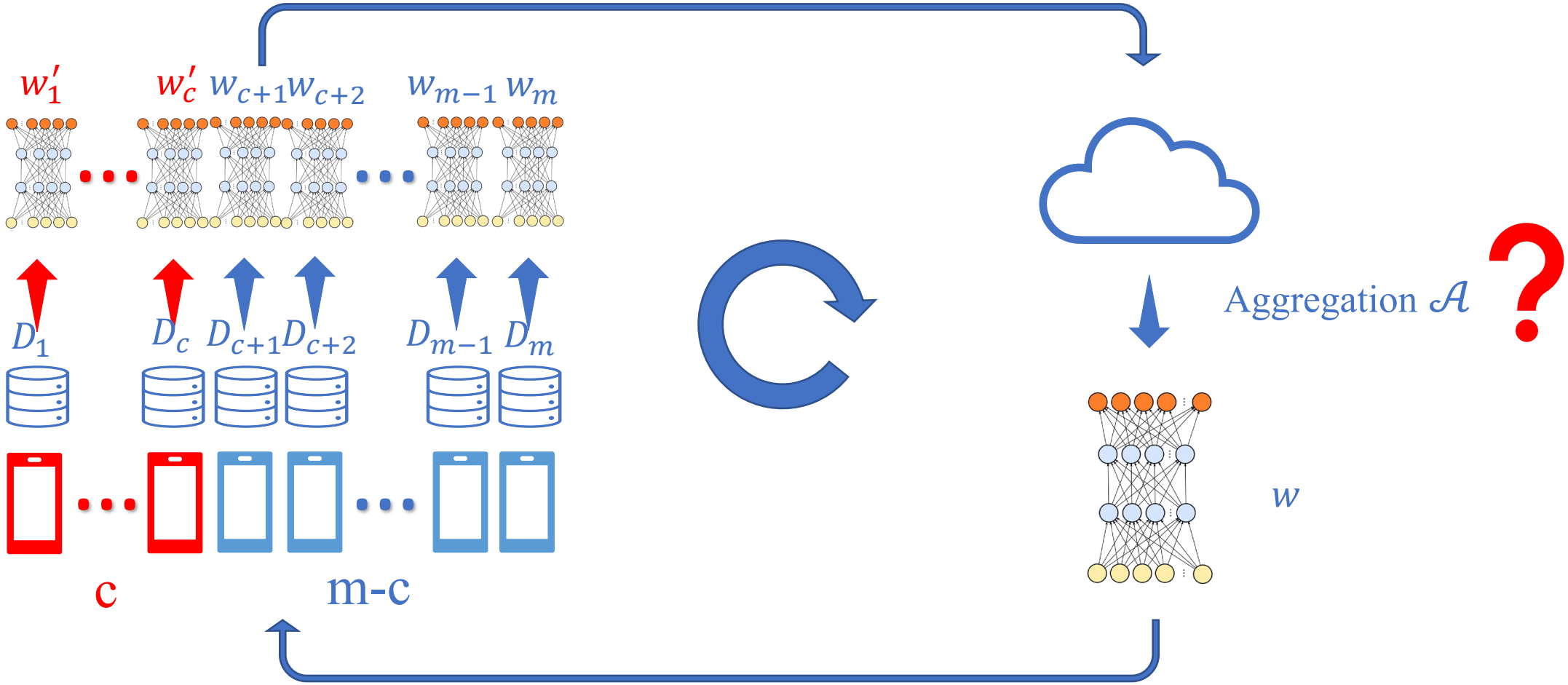
Partial-knowledge Attack



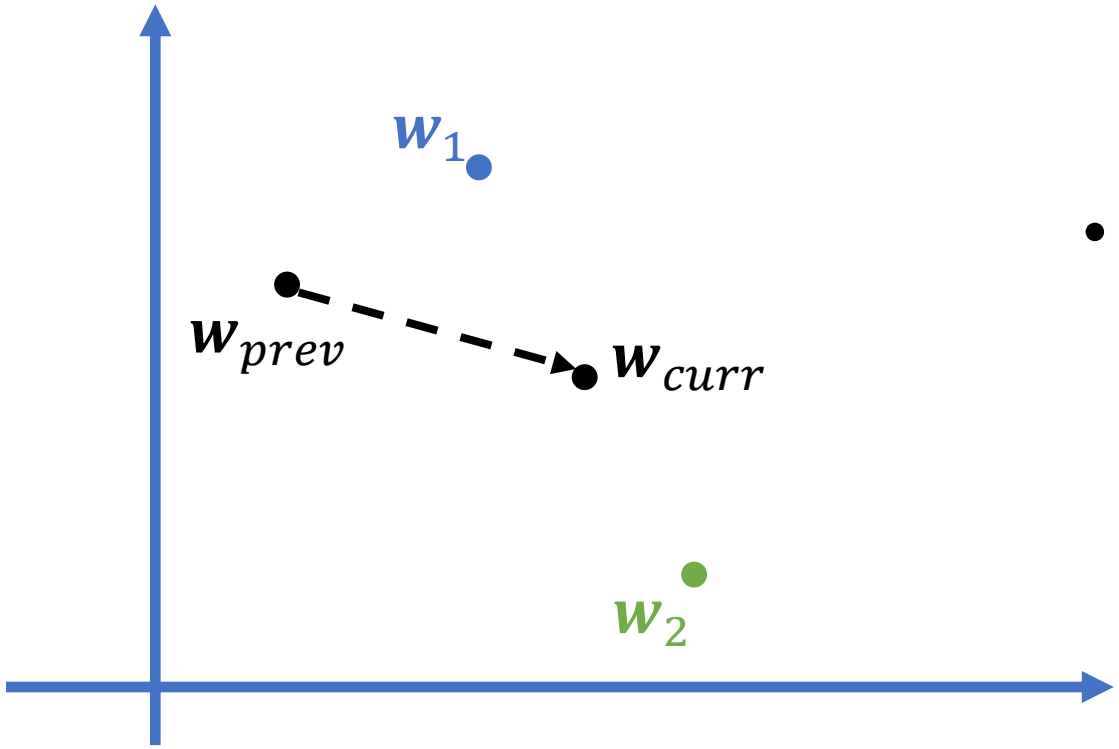
Aggregation Rule is Known



Aggregation Rule is Unknown

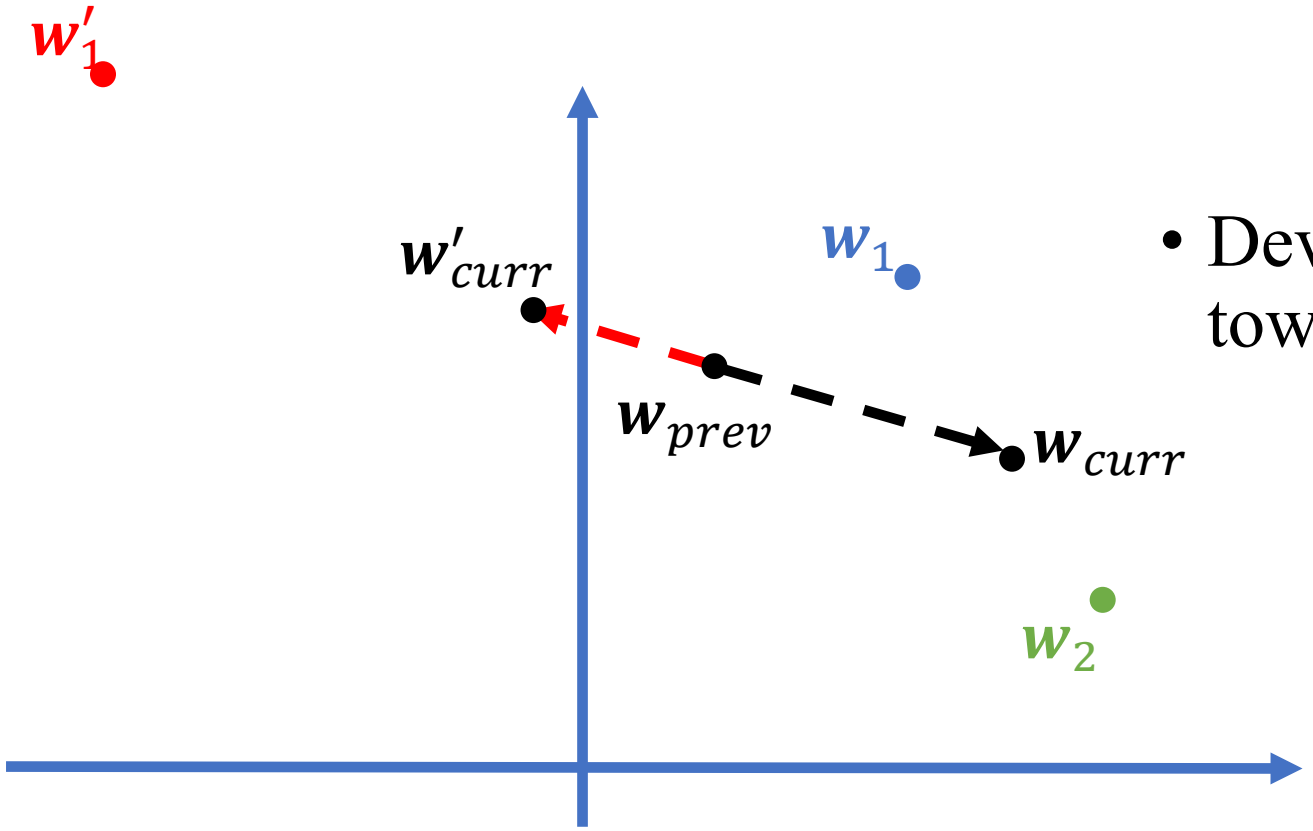


Our Idea



- No attack: global model changes along some direction.

Our Idea



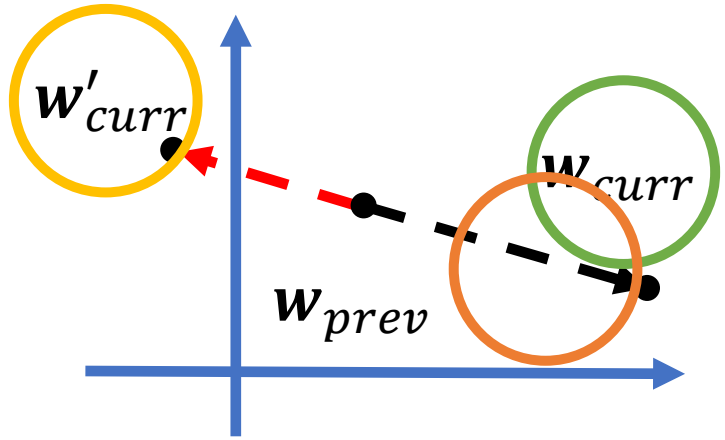
- Deviate global model the most towards inverse of the direction.

Formulate Optimization Problem

$$\max_{\boldsymbol{w}'_1,\ldots,\boldsymbol{w}'_c} \boldsymbol{s}^T(\boldsymbol{w} - \boldsymbol{w}')$$

Subject to $\boldsymbol{w} = \mathcal{A}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c, \boldsymbol{w}_{c+1}, \ldots, \boldsymbol{w}_m)$

$\boldsymbol{w}' = \mathcal{A}(\boldsymbol{w}'_1, \ldots, \boldsymbol{w}'_c, \boldsymbol{w}_{c+1}, \ldots, \boldsymbol{w}_m)$



where $\boldsymbol{s}$ is a column vector of the changing directions of $\boldsymbol{w}$.

Our formula is general to **any** aggregation rule $\mathcal{A}$.

Solving the Optimization Problem

- Full knowledge:
  - $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c, \boldsymbol{w}_{c+1}, \ldots, \boldsymbol{w}_m$ are known.
  - Directly solve the optimization problem.

- Partial knowledge:
  - Only $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c$ are known.
  - Estimate $\boldsymbol{w}$ with $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c$.

- Aggregation unknown:
  - Take a guess on $\mathcal{A}$.

Experimental Settings

- 100 workers
  - 20 compromised by default
- Non-IID data distribution
- Datasets:
  - MNIST (default)
  - Fashion-MNIST
  - CH-MNIST
  - Breast Cancer Wisconsin (Diagnostic)

Our Attack is Effective

	NoAttack	Gaussian	LabelFlip	Partial	Full
Krum	0.11	0.10	0.10	**0.75**	**0.77**
Trimmed Mean	0.06	0.07	0.07	**0.14**	**0.23**
Median	0.06	0.06	0.16	**0.28**	**0.32**

Our attacks can effectively increase error rates

Our Attack is Effective

There is no attack

	NoAttack	Gaussian	LabelFlip	Partial	Full
Krum	0.11	0.10	0.10	**0.75**	**0.77**
Trimmed Mean	0.06	0.07	0.07	**0.14**	**0.23**
Median	0.06	0.06	0.16	**0.28**	**0.32**

Our attacks can effectively increase error rates

Our Attack is Effective

Adding gaussian random noise to local models

	NoAttack	Gaussian	LabelFlip	Partial	Full
Krum	0.11	0.10	0.10	**0.75**	**0.77**
Trimmed Mean	0.06	0.07	0.07	**0.14**	**0.23**
Median	0.06	0.06	0.16	**0.28**	**0.32**

Our attacks can effectively increase error rates

Our Attack is Effective

Flip labels of local training data

	NoAttack	Gaussian	LabelFlip	Partial	Full
Krum	0.11	0.10	0.10	**0.75**	**0.77**
Trimmed Mean	0.06	0.07	0.07	**0.14**	**0.23**
Median	0.06	0.06	0.16	**0.28**	**0.32**

Our attacks can effectively increase error rates

Our Attack is Effective

Our attack, partial knowledge

	NoAttack	Gaussian	LabelFlip	Partial	Full
Krum	0.11	0.10	0.10	**0.75**	**0.77**
Trimmed Mean	0.06	0.07	0.07	**0.14**	**0.23**
Median	0.06	0.06	0.16	**0.28**	**0.32**

Our attacks can effectively increase error rates

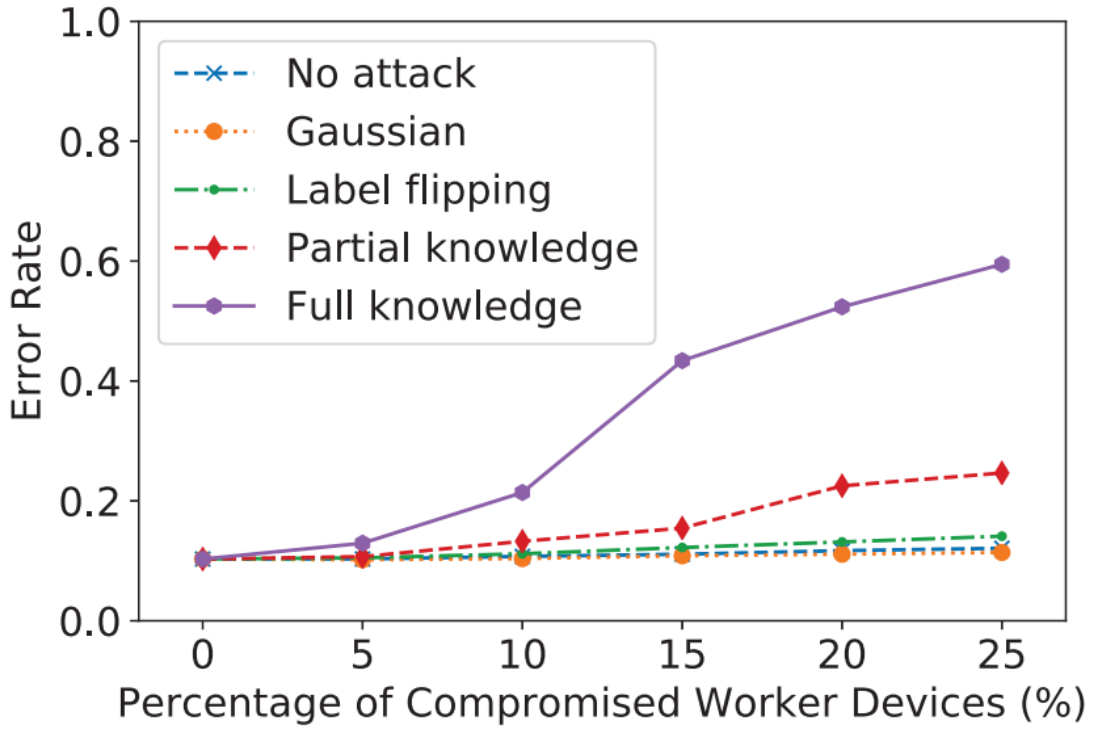
Our Attack is Effective

Our attack, full knowledge

	NoAttack	Gaussian	LabelFlip	Partial	Full
Krum	0.11	0.10	0.10	**0.75**	**0.77**
Trimmed Mean	0.06	0.07	0.07	**0.14**	**0.23**
Median	0.06	0.06	0.16	**0.28**	**0.32**

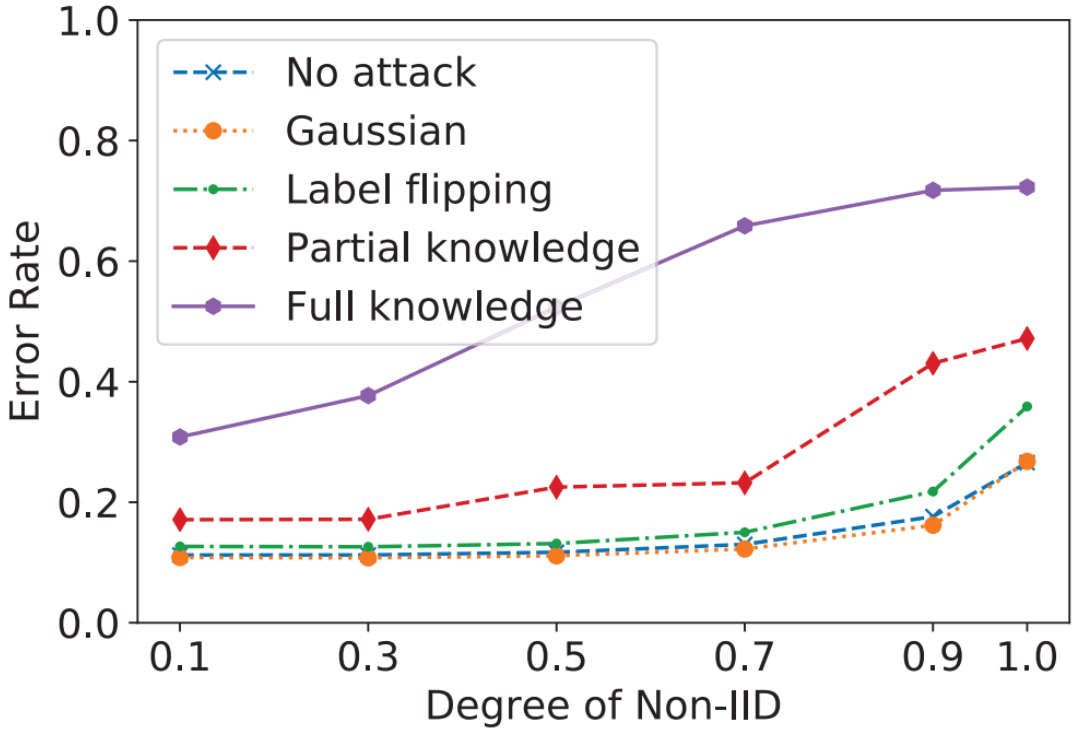
Our attacks can effectively increase error rates

More Compromised, Better Attack



Our attacks are more effective with more compromised workers

More Non-IID, Better Attack



Our attacks are more effective when data are more Non-IID among workers

Our Attacks Transfer between Aggregation Rules

	Krum	Trimmed mean	Median
No attack	0.14	0.12	0.13
Krum attack	0.70	0.15	0.18
Trimmed mean attack	0.14	0.25	0.20

Comparing with Data Poisoning Attacks

	NoAttack	DataPoisoning	Partial	Full
Mean	0.10	0.11	**0.54**	**0.69**
Krum	0.23	0.24	**0.85**	**0.89**
Trimmed Mean	0.12	0.12	**0.27**	**0.32**
Median	0.13	0.13	**0.19**	**0.21**

Our attacks are much more effective than data poisoning attacks

Possible Defenses

- Error Rate based Rejection (ERR):
  - Remove local models based on validation error rates.
- Loss Function based Rejection (LFR):
  - Remove local models based on validation loss.
- Union

Defense Results

	No attack	Krum	Trimmed mean
Krum	0.14	0.72	0.13
Krum + ERR	0.14	0.62	0.13
Krum + LFR	0.14	0.58	0.14
Krum + Union	0.14	0.48	0.14
Trimmed mean	0.12	0.15	0.23
Trimmed mean + ERR	0.12	0.17	0.21
Trimmed mean + LFR	0.12	0.18	0.12
Trimmed mean + Union	0.12	0.18	0.12
Median	0.13	0.17	0.19
Median + ERR	0.13	0.21	0.25
Median + LFR	0.13	0.20	0.13
Median + Union	0.13	0.19	0.14

- The defenses are effective in some cases while not in others
- We need more advanced defenses against our attacks

Conclusion

- We propose a general local model poisoning attack for any Byzantine-robust federated learning.
- Our attack can increase error rates of global models in Byzantine-robust federated learning.
- New defenses are needed to defend against our attacks.

Thank you!

- For any questions, please email
  - xiaoyu.cao@duke.edu
  - myfang@iastate.edu