Label Inference Attacks Against Vertical Federated Learning

Chong Fu Xuhong Zhang Shouling Ji Jinyin Chen Jingzheng Wu
Shanqing Guo Jun Zhou Alex X. Liu Ting Wang

2021

Big Data Era

- Locations
- Health records
- View histories
- …

Private user data

Data collection

Google

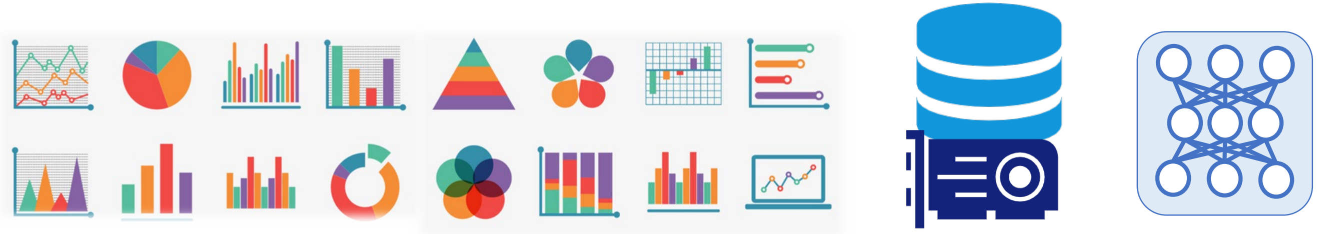
facebook

Tencent

IT companies'

- Apps
- Websites
- …

Data analysis (includes machine learning)

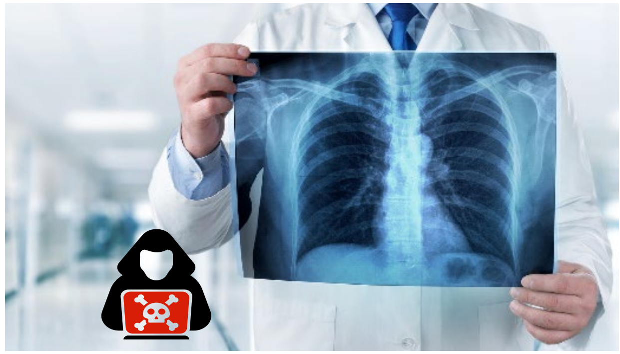


Data Leakage

2018, Facebook
exposed 87 million user data

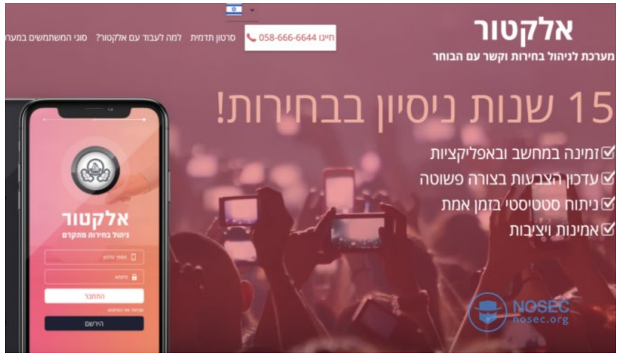
2019, Capital One Bank
leaked 106 million user data

2020, Marriott Hotel
breached 5.2 million user data



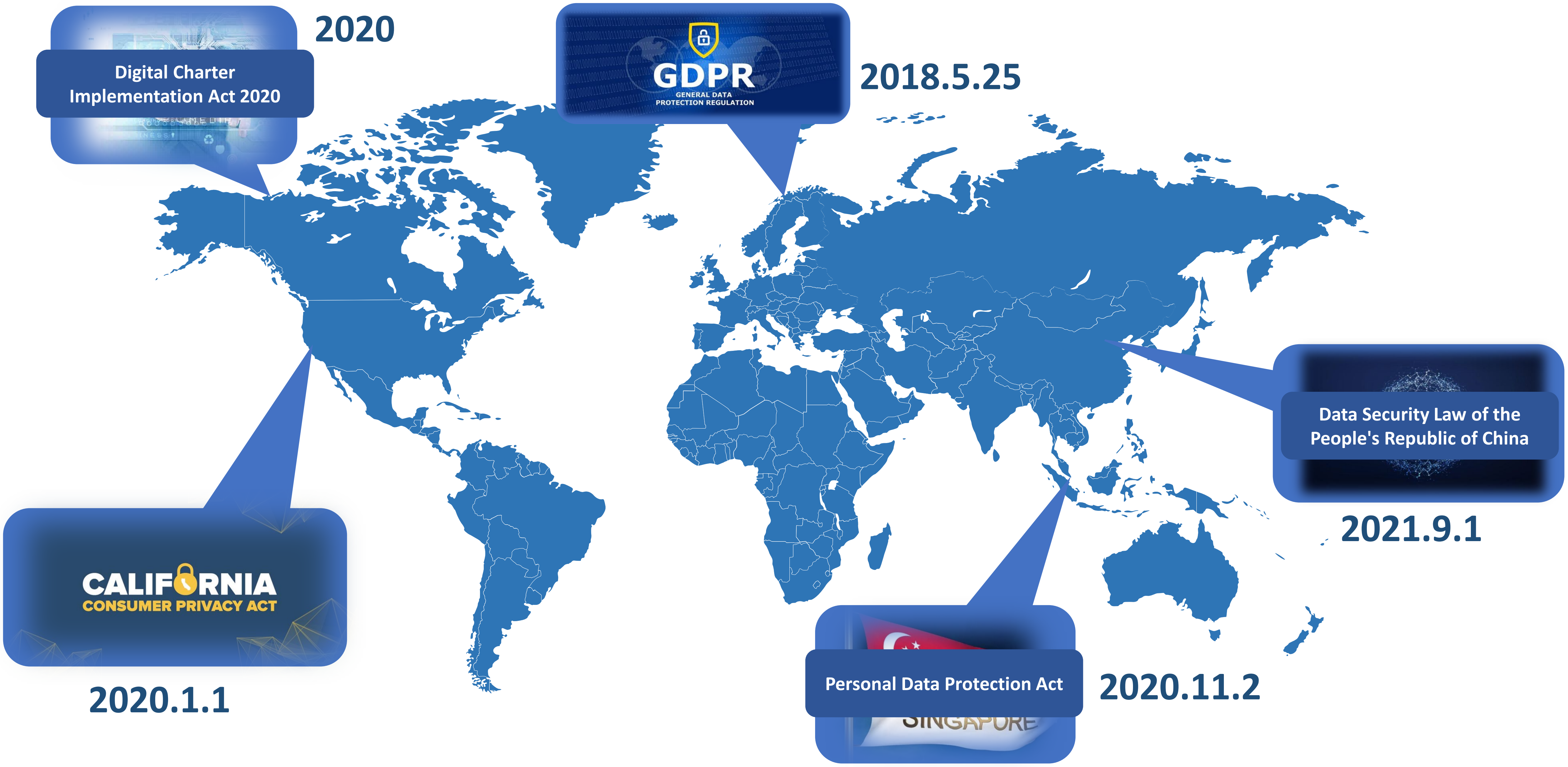
2020, Brazilian ministry of health
leaked 0.24 billion records

2020, Microsoft
exposed 250 million records



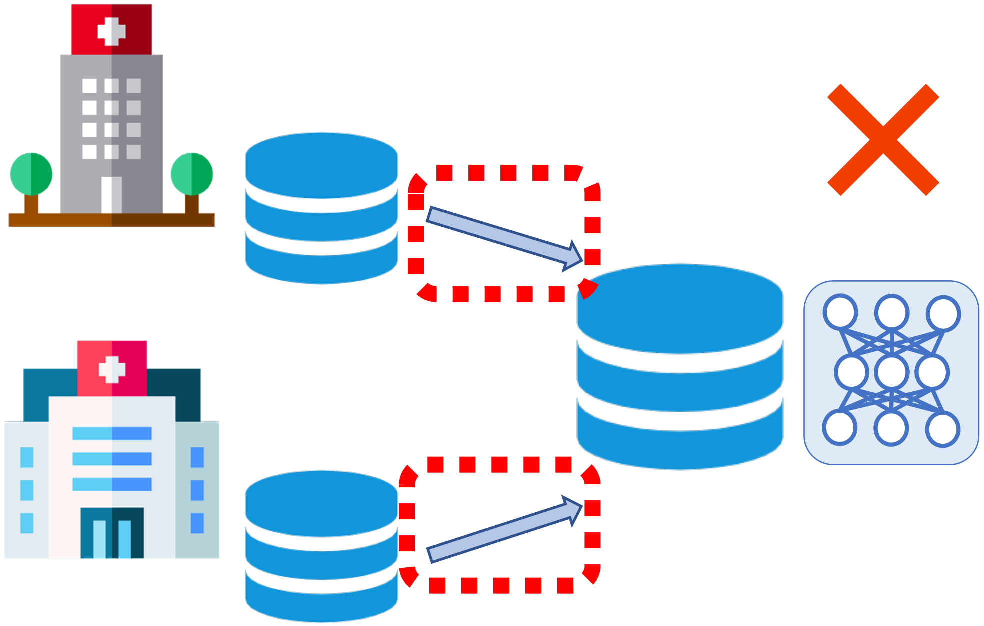
2020, 6.4 million voters'
data in Israel were leaked

User Data Protection Laws



The Dilemma of “Isolated Data”

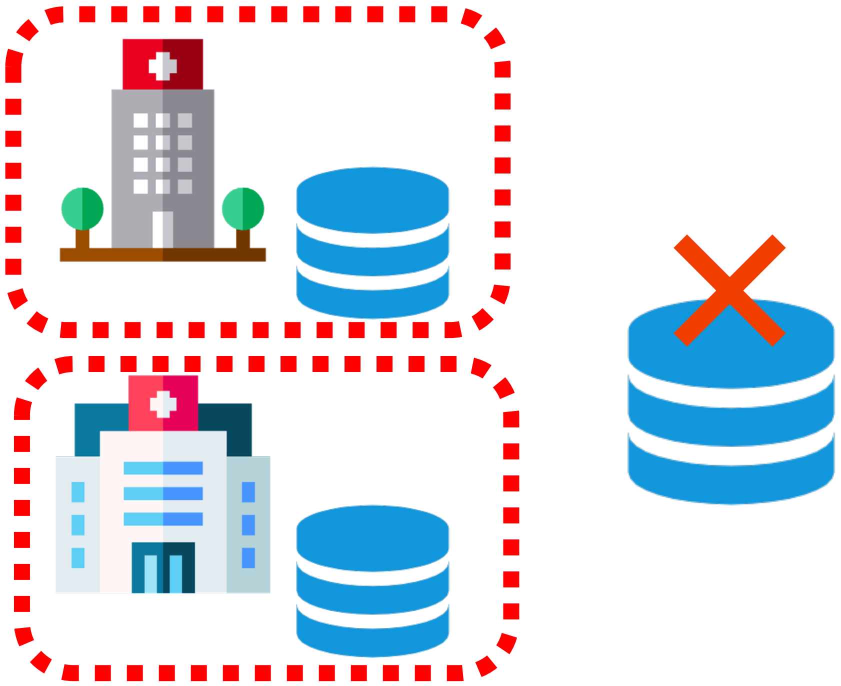
- **The dilemma of “isolated data”**



Traditional **centralized** machine learning breaks laws of user data protection.

The Dilemma of “Isolated Data”

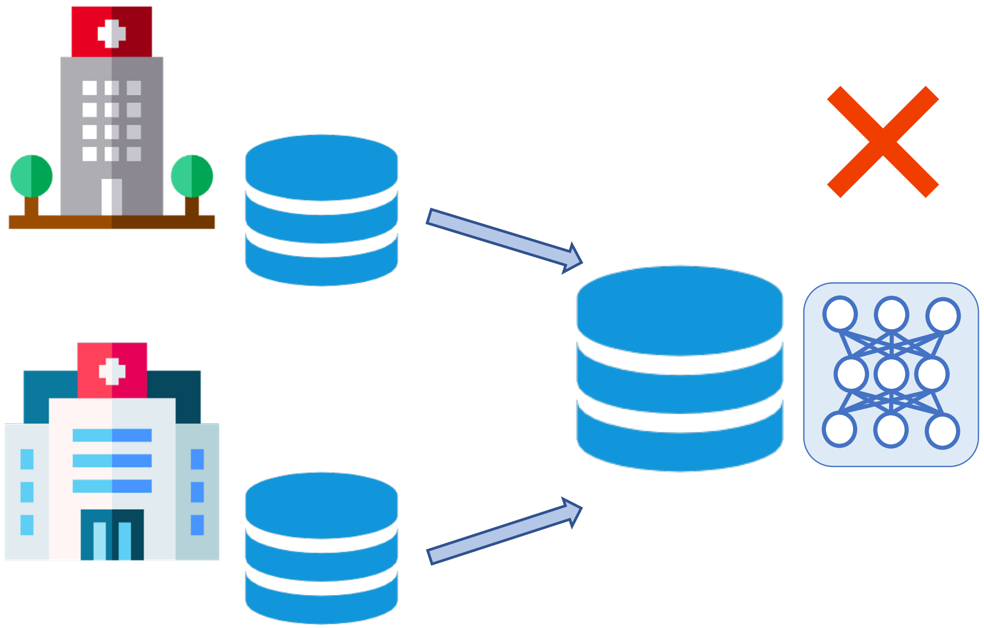
➢ **The dilemma of “isolated data”**



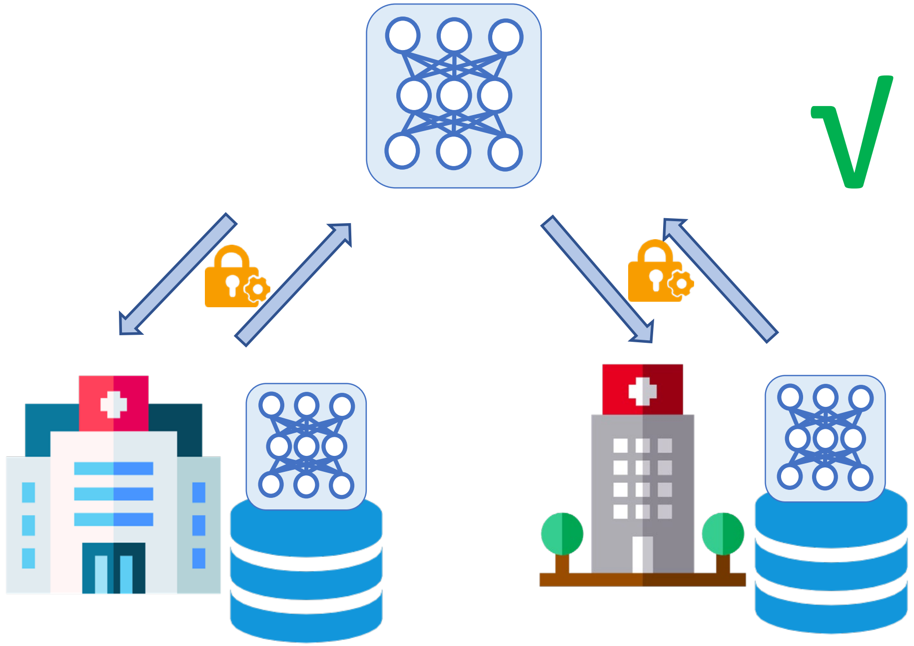
User data is **isolated** in different companies or organizations.

Federated Learning

- The dilemma of “isolated data”
- Federated learning (FL)

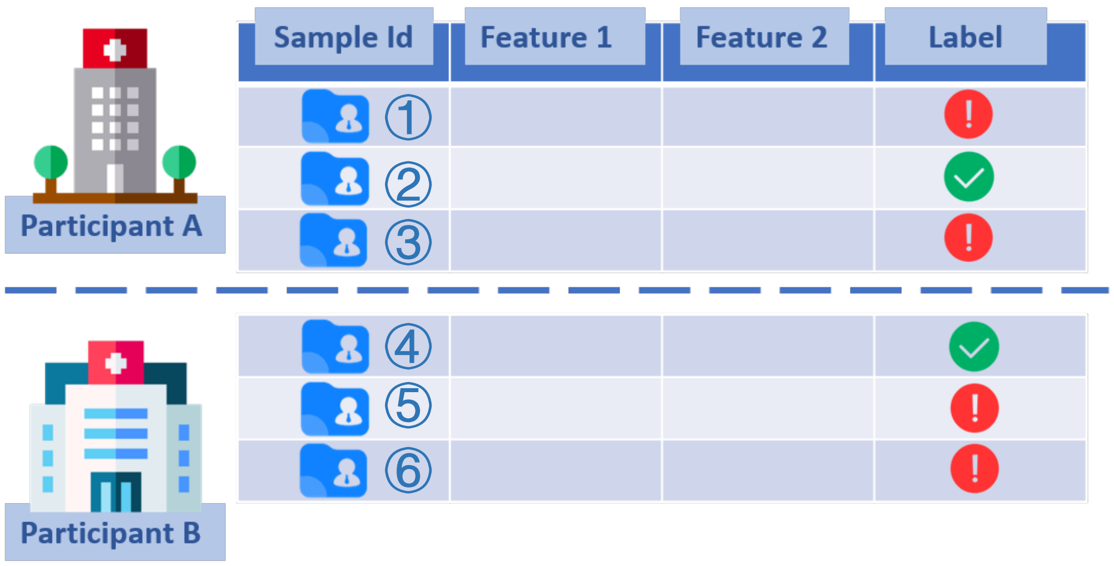


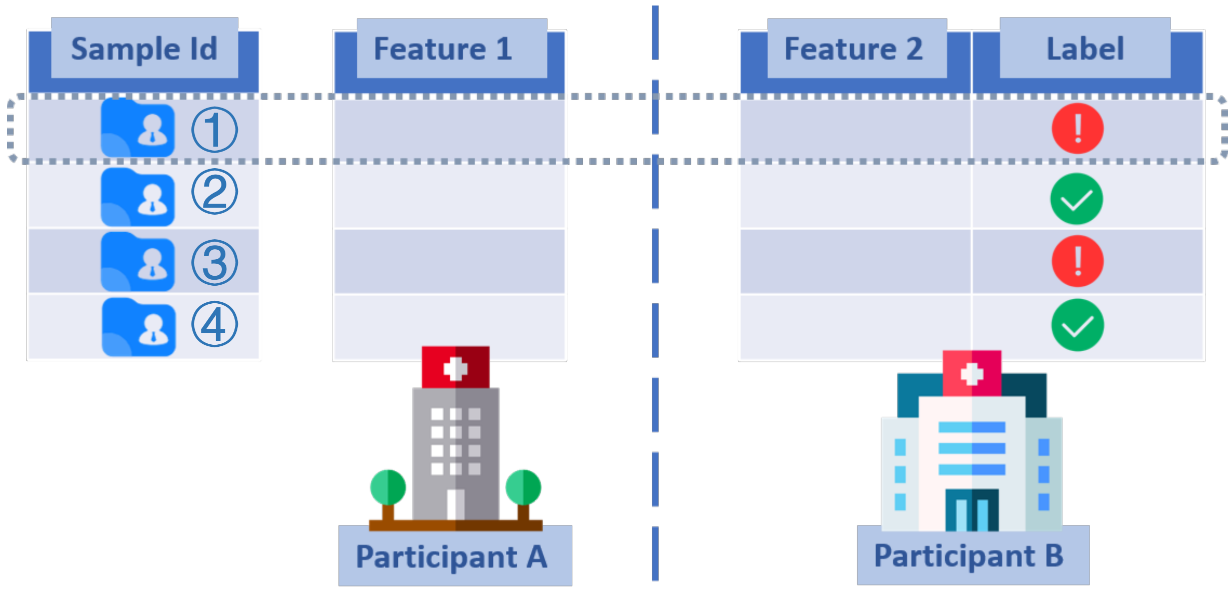
Traditional **centralized** machine learning breaks laws of user data protection.



FL allows multiple participants to collaboratively train a machine learning model **without revealing their local data**.

Horizontal Federated Learning & Vertical Federated Learning





Horizontal federated learning (HFL):
Datasets share the **same feature space** but **differ in the sample space**.

Vertical federated learning (VFL):
Datasets share the **same sample space** but **differ in the feature space**.

Federated Learning Is Widely Used

➢ FL is being widely used in industry. Worldwide IT companies put much effort into developing FL systems.

TensorFlow Federated from Google

PySyft from OpenMined

Federated AI Technology Enabler (FATE) from Tencent

Fedlearner from ByteDance

PaddleFL from Baidu

Federated Learning Has Vulnerabilities

An adversarial participant in federated learning may:

Infer private information of other participants

- Infer membership [Oakland' 19]
- Infer class representatives [CCS' 17]
- Infer sample properties [Oakland' 19]
- Reconstructing training samples [NeurIPS' 19]
- …

Attack the federated model

- Inject backdoor to the federated model [ICLR' 20]
- Poison the federated model [USENIX Security' 20]
- …

- *Above studies have thoroughly analyzed the privacy and security risks of HFL.* ***However, the privacy risks of VFL remain unexplored.***
- *We reveal and shed lights on the vulnerability of VFL to the* ***label inference*** *attacks.*

Label Inference Attacks

Illustration of Label Inference Attacks Against VFL with Model Splitting

- Several participants collaboratively train a VFL model.

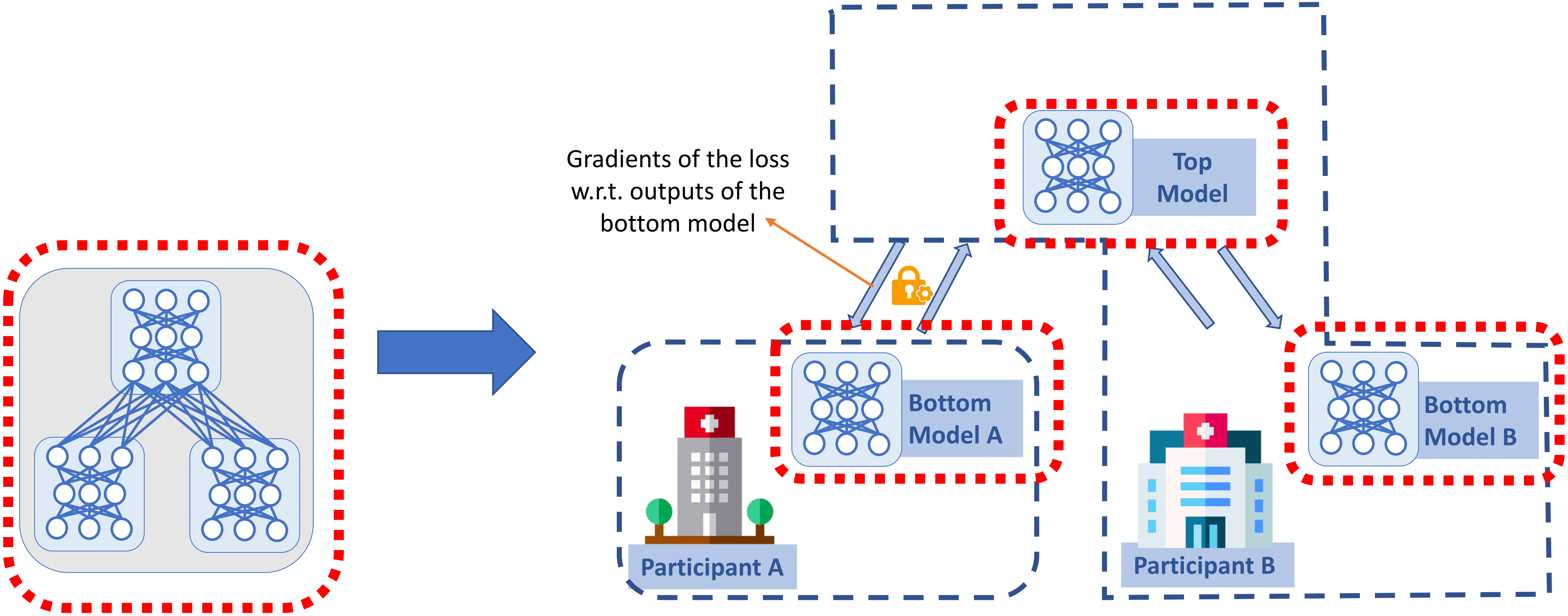
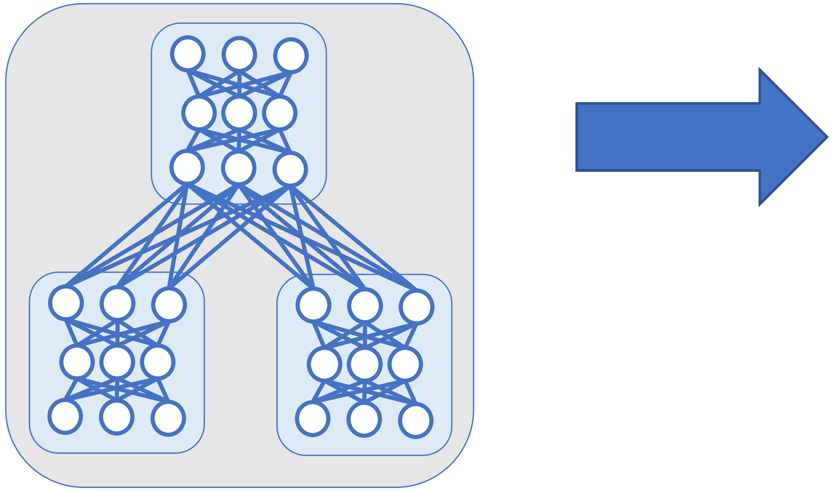


Illustration of Label Inference Attacks Against VFL with Model Splitting

- Several participants collaboratively train a VFL model.
- Every participant in VFL holds **partial features**.



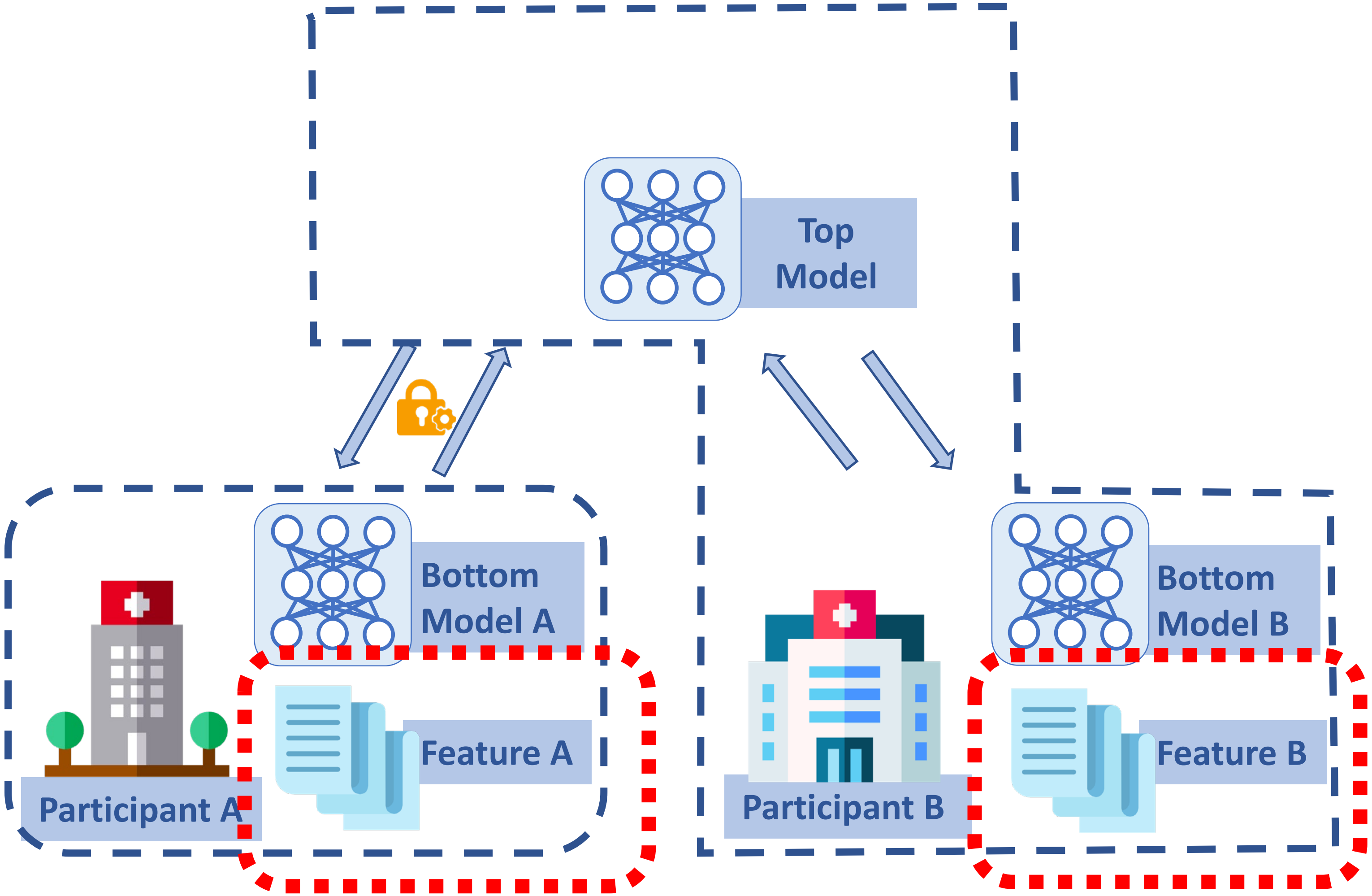
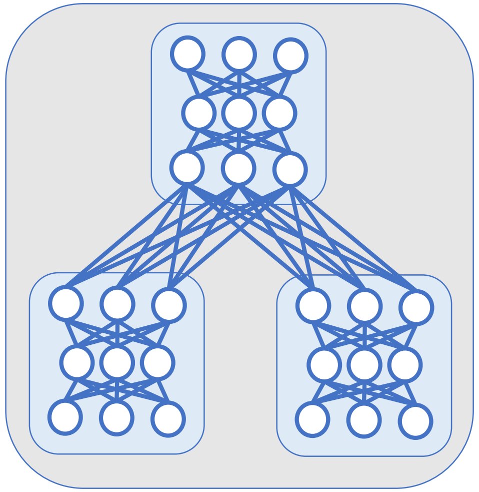


Illustration of Label Inference Attacks Against VFL with Model Splitting

- Several participants collaboratively train a VFL model.
- Every participant in VFL holds **partial features**.
- The **labels** are privately owned by one participant. This participant also controls the server running the **top model**.



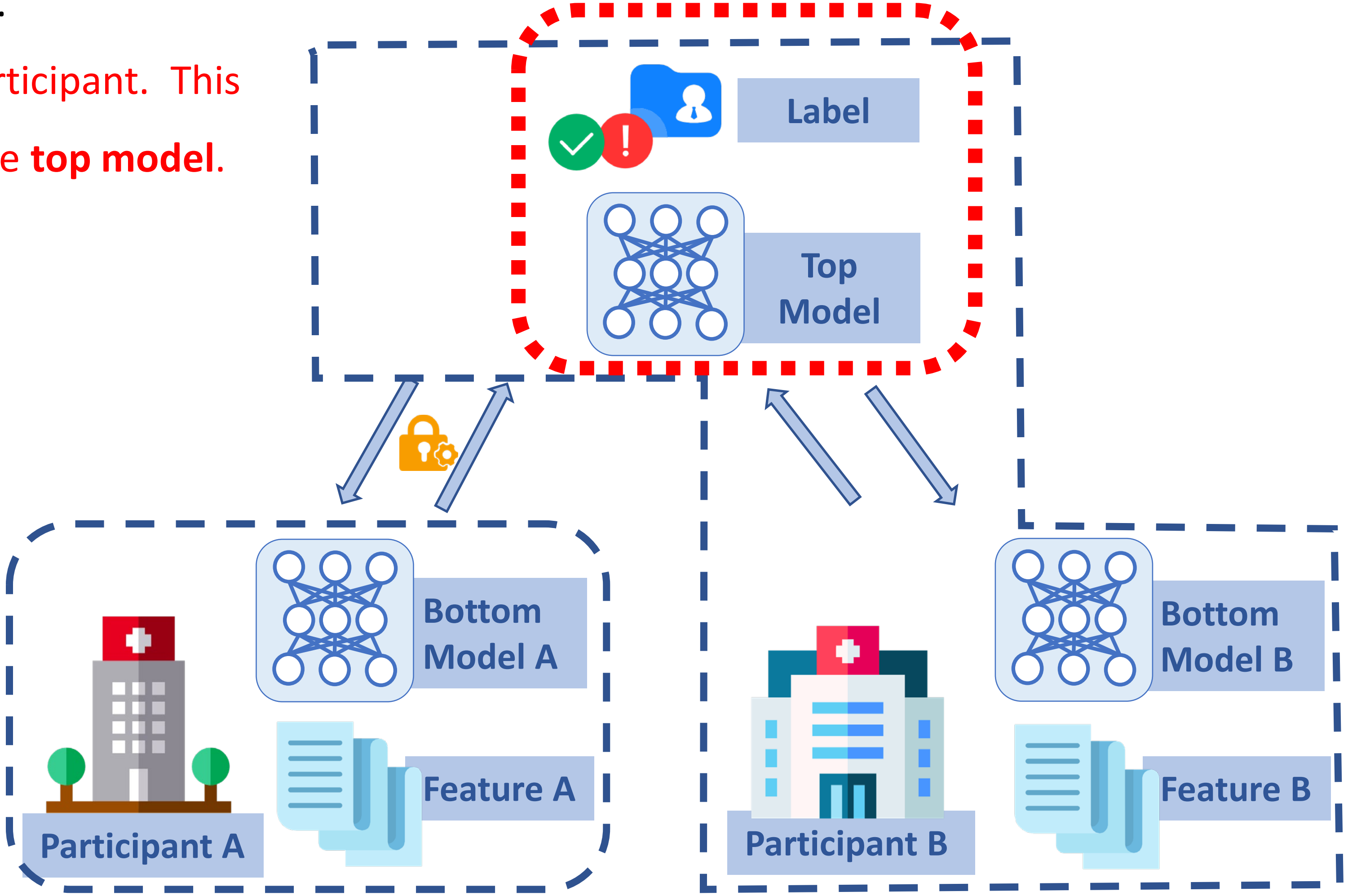
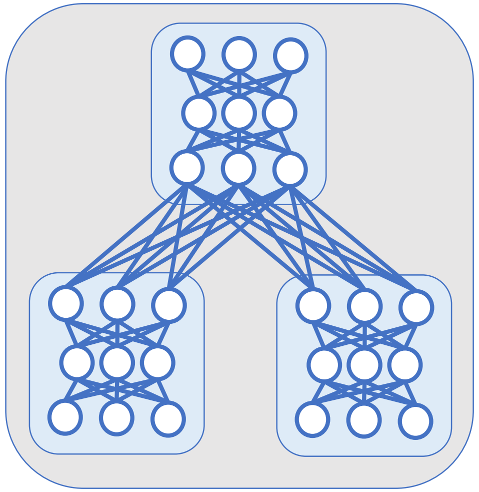
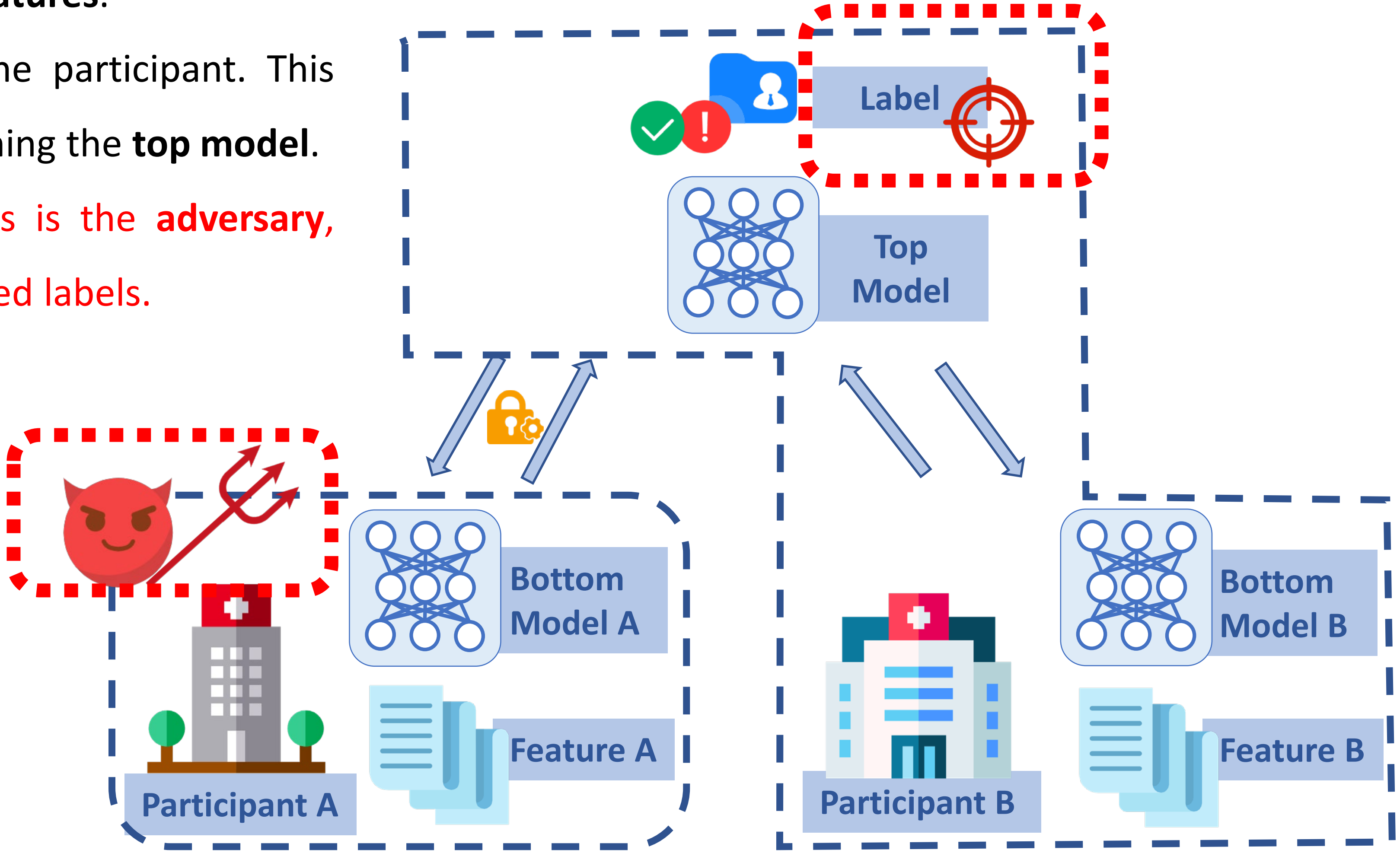


Illustration of Label Inference Attacks Against VFL with Model Splitting

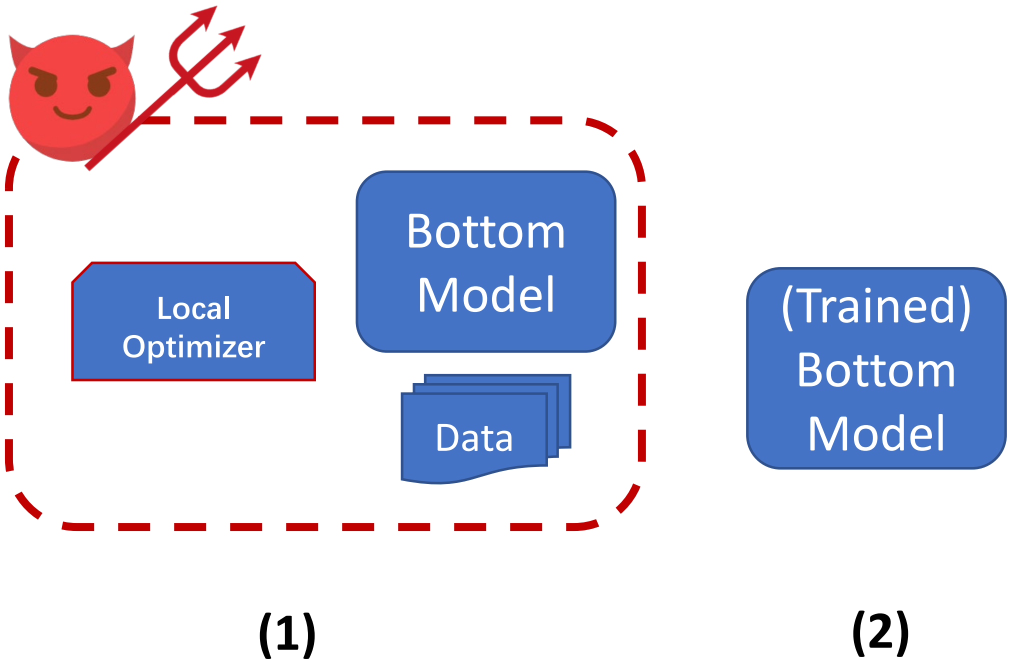
- Several participants collaboratively train a VFL model.
- Every participant in VFL holds **partial features**.
- The **labels** are privately owned by one participant. This participant also controls the server running the **top model**.
- One of the participants without labels is the **adversary**, whose goal is to infer the privately owned labels.





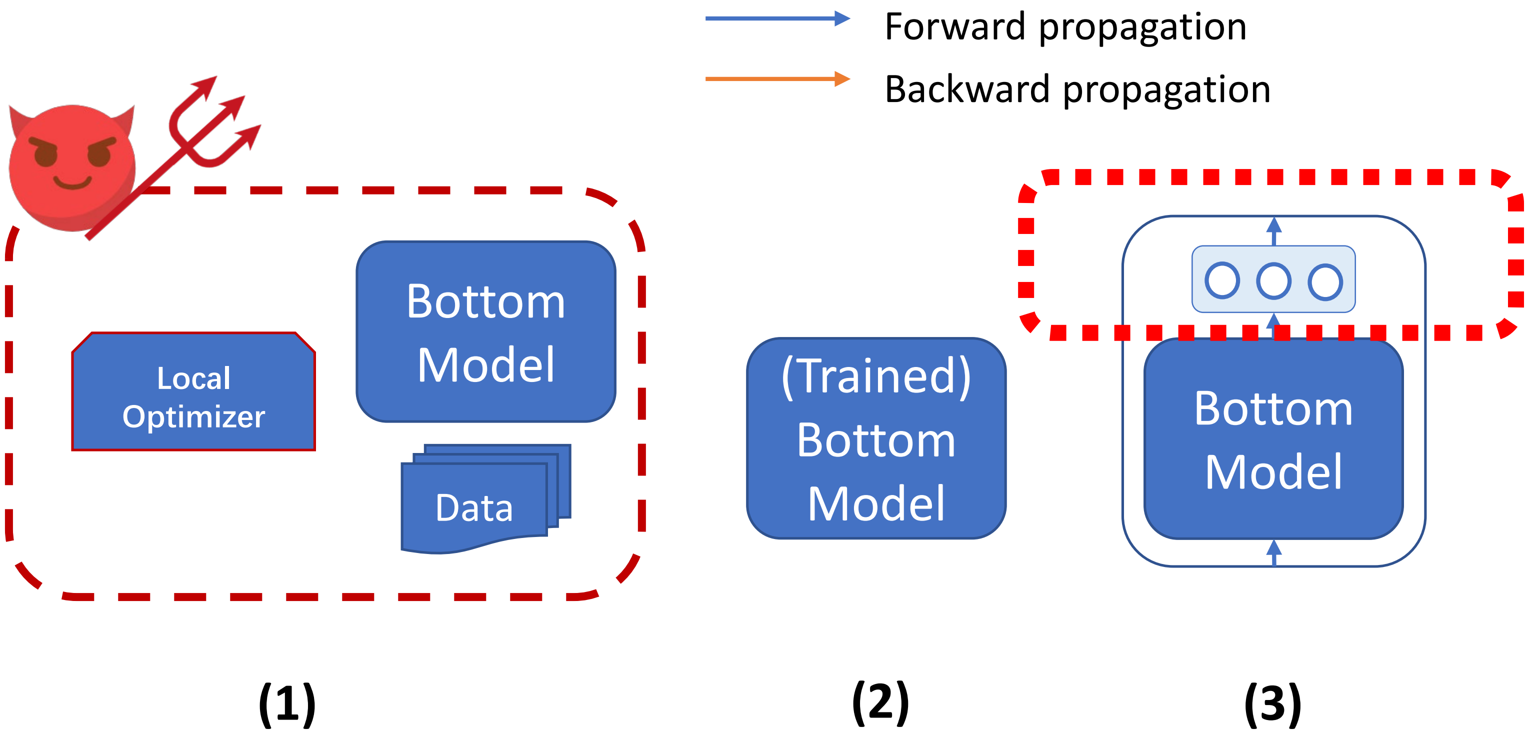
Attack 1: Passive Label Inference Attack

- Exploit the locally **owned bottom model**.



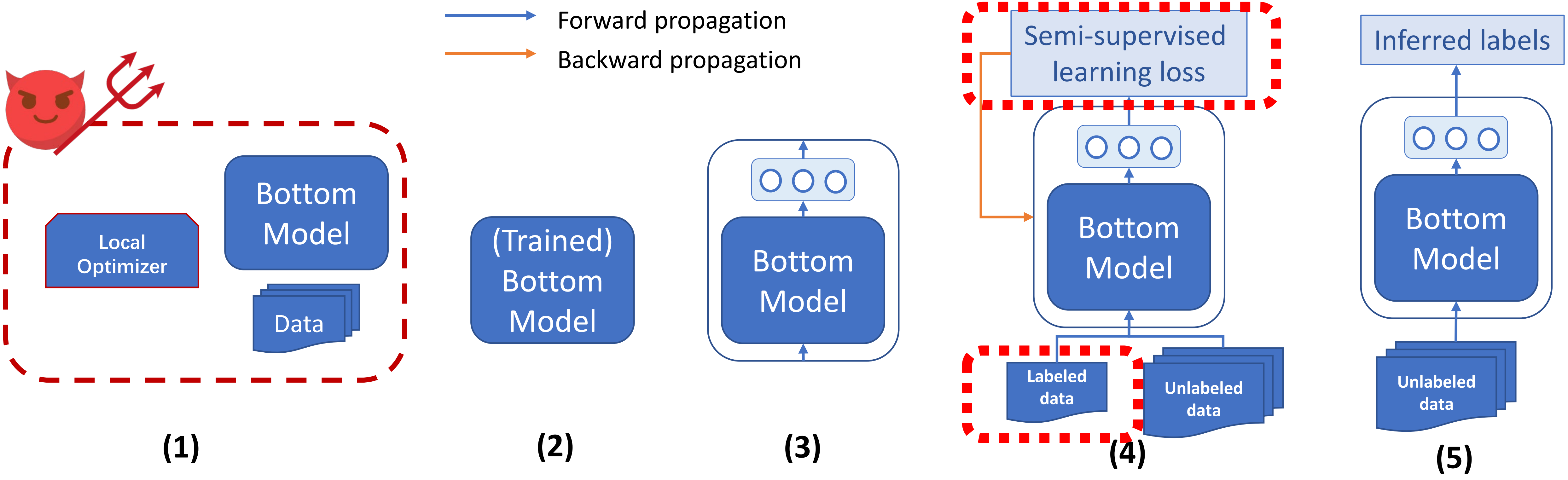
Attack 1: Passive Label Inference Attack

- Exploit the locally **owned bottom model**.
- Fine-tune the bottom model with an **additional classification layer**.



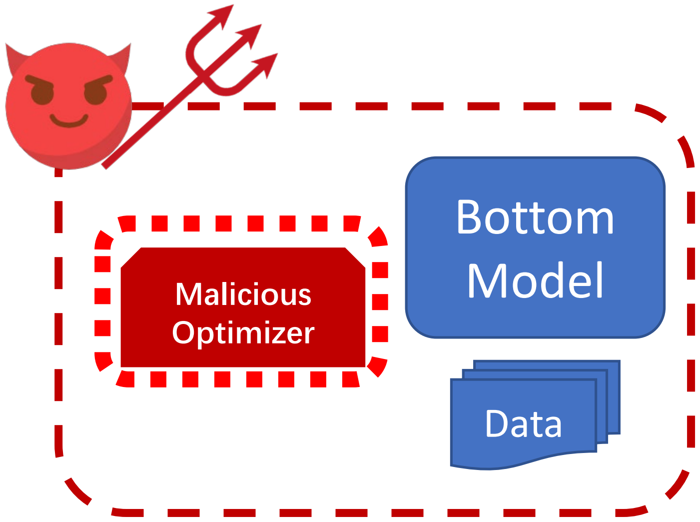
Attack 1: Passive Label Inference Attack

- ➢ Exploit the locally **owned bottom model**.
- ➢ Complete the bottom model with an **additional classification layer**.
- ➢ Use a small amount of auxiliary labeled data to **fine-tune** the bottom model in a **semi-supervised** manner.



Attack 2: Active Label Inference Attack

- **Accelerate** the local model's learning during training



(1)

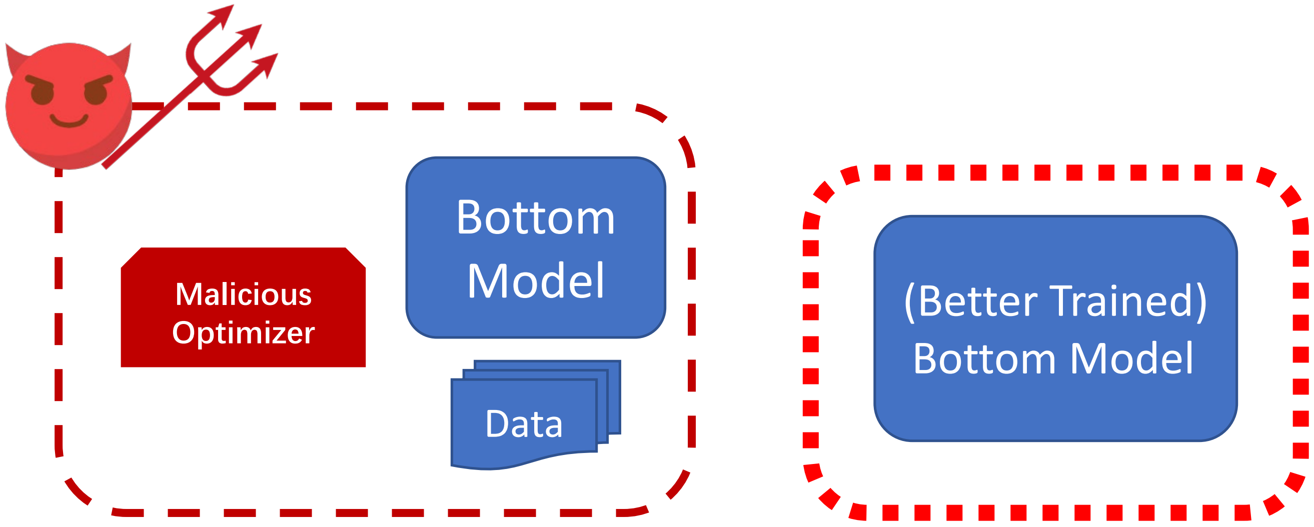
Algorithm 1 Local malicious optimization of the adversary's bottom model

Require: Momentum parameter β, the gradient scaling factor's resetting parameter γ, maximum gradient scaling factor r_{max}, minimum gradient scaling factor r_{min}, learning rate η, initial bottom model parameters Θ, initial gradient velocity v.

```
while stopping criterion not met do
    Receive G_output from the server
    G ← Backward(G_output)
    for each parameter θ in Θ and its gradient g_θ in G do
        v_θ ← β · v_θ + (1 − β) · g_θ
        if is not the first criterion then
            r_θ ← 1.0 + γ · (v_θ ÷ v_last)
            r_θ ← Max(r_θ, r_min)
            r_θ ← Min(r_θ, r_max)
            v_θ ← r_θ · v_last
        end if
        v_last ← v_θ
        θ ← θ − η · v_θ
    end for
end while
```

Attack 2: Active Label Inference Attack

- **Accelerate** the local model's learning during training
- Better expressiveness of the bottom model
- The VFL model is tricked to **rely more on the adversary's bottom model**



(1) (2)

Algorithm 1 Local malicious optimization of the adversary's bottom model

Require: Momentum parameter β, the gradient scaling factor's resetting parameter γ, maximum gradient scaling factor r_{max}, minimum gradient scaling factor r_{min}, learning rate η, initial bottom model parameters Θ, initial gradient velocity v.

while stopping criterion not met **do**
    Receive G_{output} from the server
    $G \leftarrow Backward(G_{output})$
    for each parameter θ in Θ and its gradient g_θ in G **do**
        $v_\theta \leftarrow \beta \cdot v_\theta + (1-\beta) \cdot g_\theta$
        if is not the first criterion **then**
            $r_\theta \leftarrow 1.0 + \gamma \cdot (v_\theta \div v_{last})$
            $r_\theta \leftarrow Max(r_\theta, r_{min})$
            $r_\theta \leftarrow Min(r_\theta, r_{max})$
            $v_\theta \leftarrow r_\theta \cdot v_{last}$
        end if
        $v_{last} \leftarrow v_\theta$
        $\theta \leftarrow \theta - \eta \cdot v_\theta$
    end for
end while

Illustration of Label Inference Attack Against VFL without Model Splitting

- Several participants collaboratively train a VFL model.
- During the forward propagation, this participant **sums up outputs of all the bottom models** to get the final output.

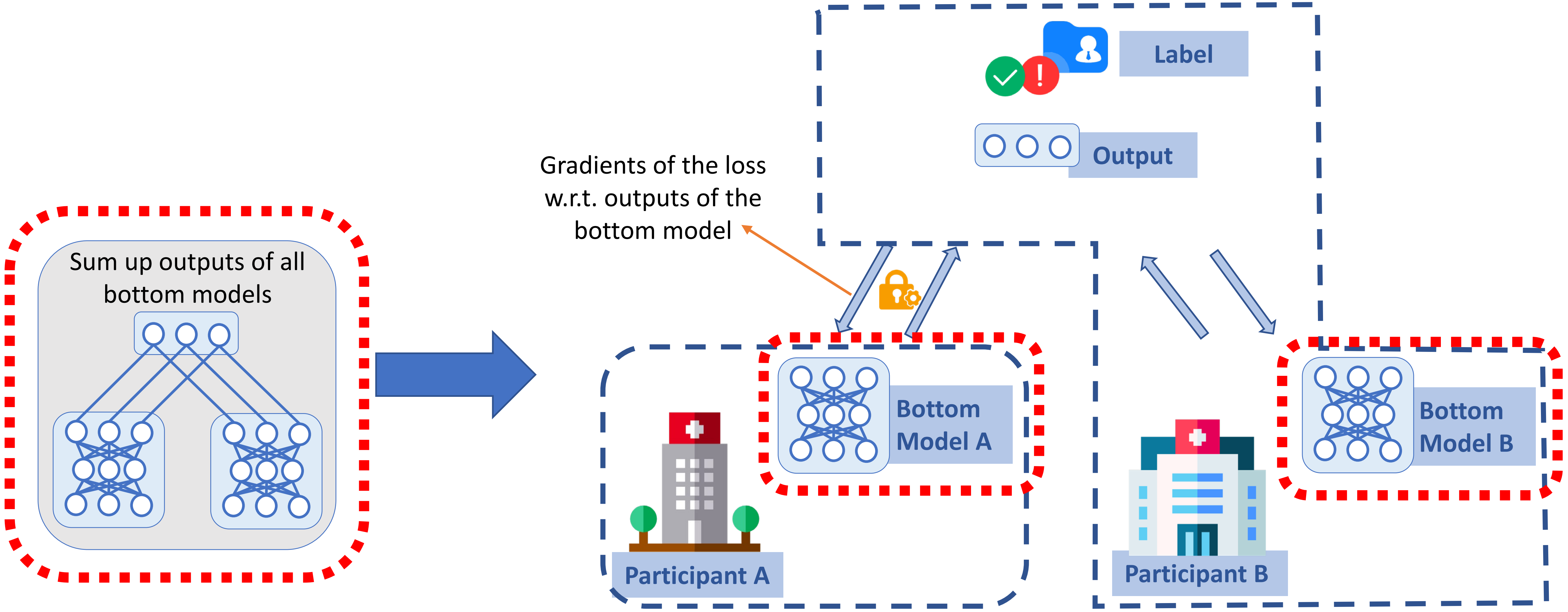
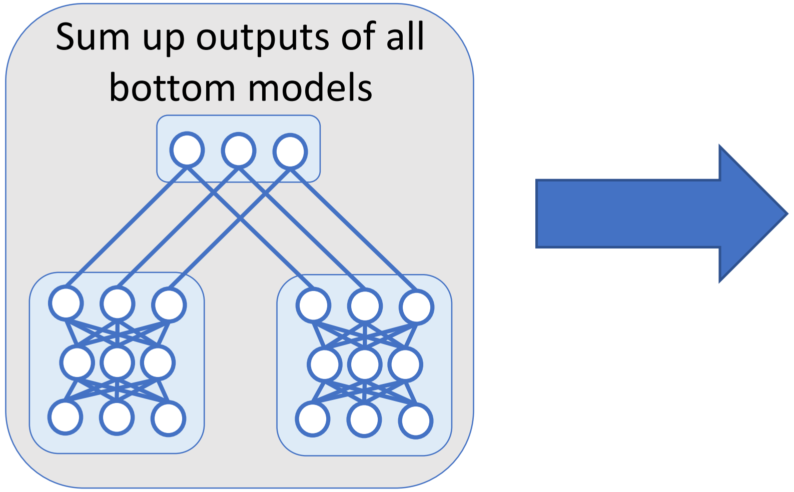


Illustration of Label Inference Attack Against VFL without Model Splitting

- Several participants collaboratively train a VFL model.
- During the forward propagation, this participant **sums up outputs of all the bottom models** to get the final output.
- Every participant holds **partial features**.
- The **labels** are privately owned by one participant.



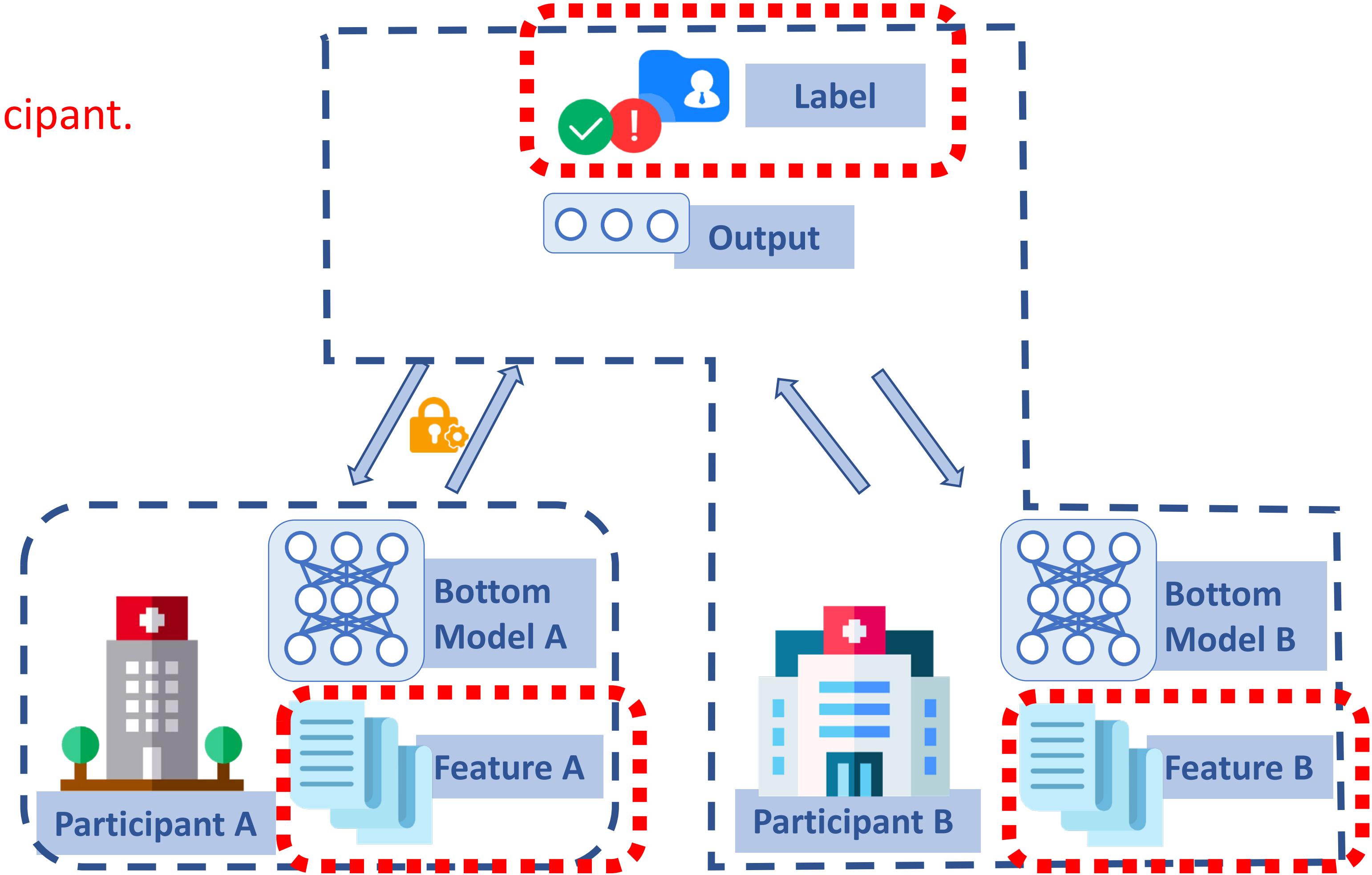
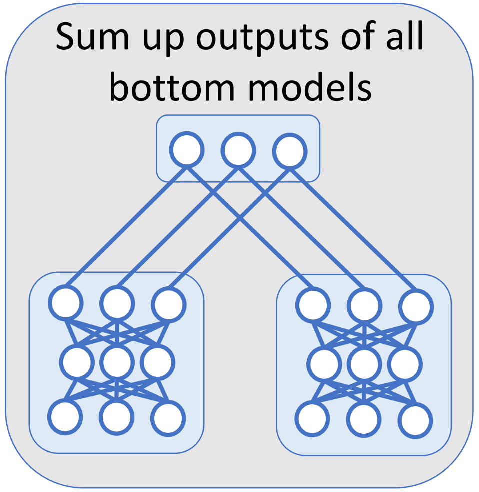
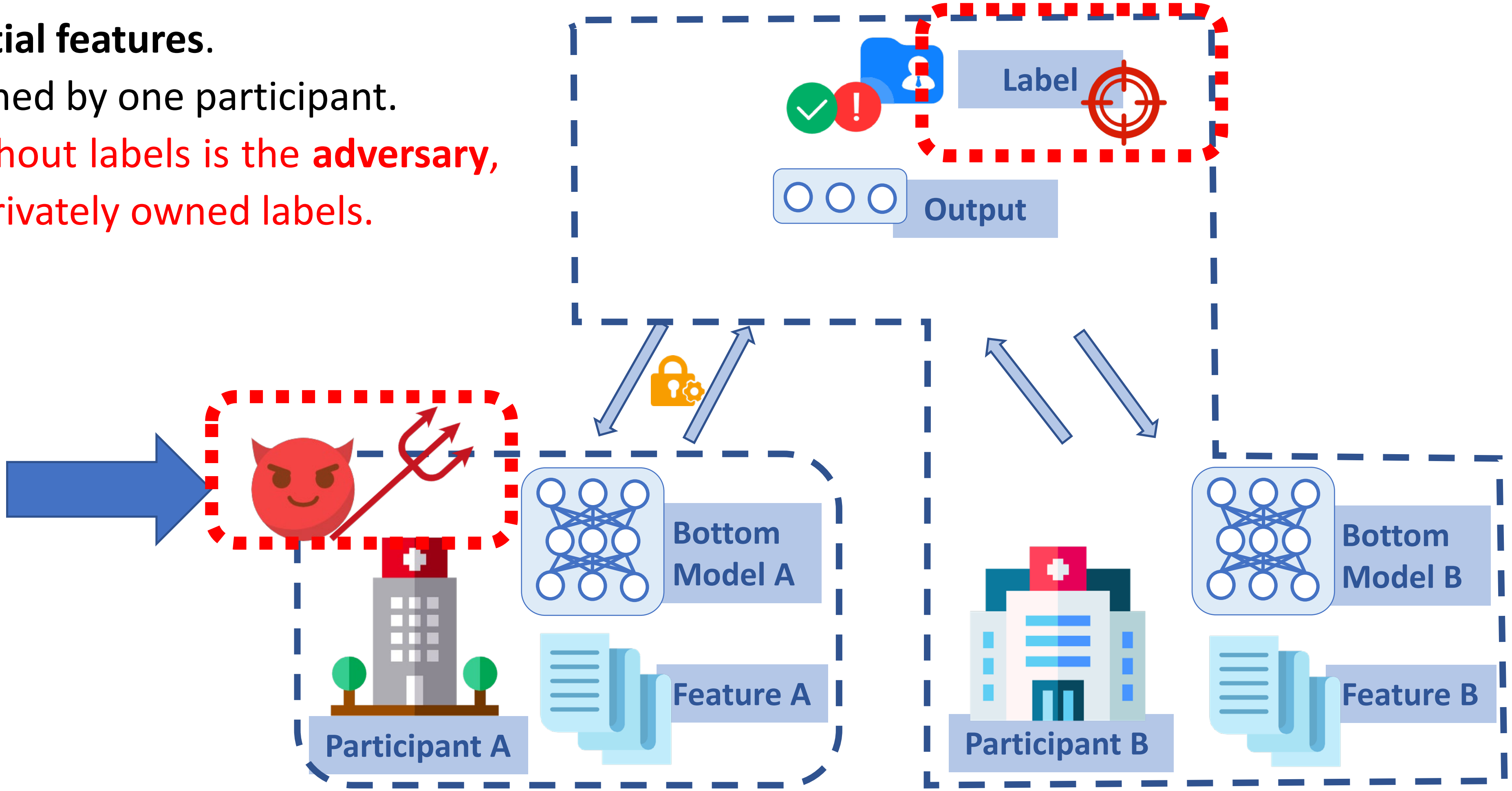


Illustration of Label Inference Attack Against VFL without Model Splitting

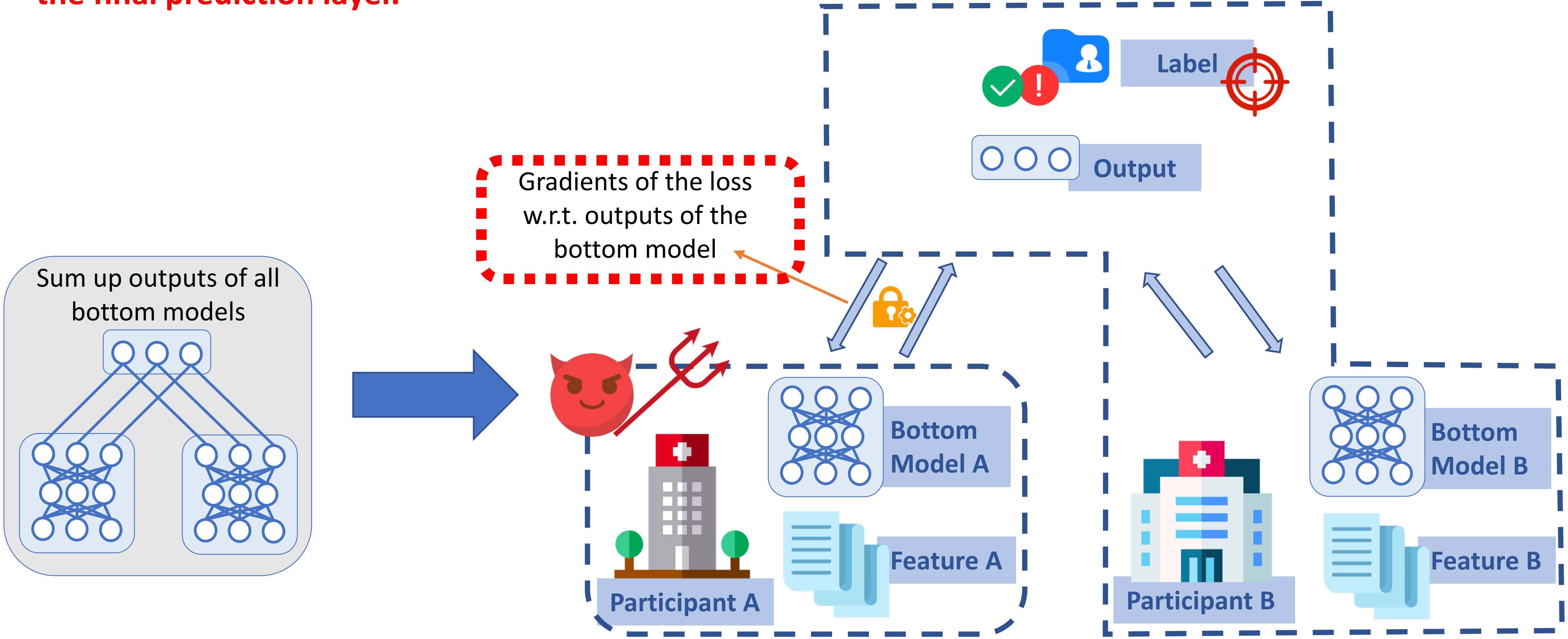
- Several participants collaboratively train a VFL model.
- During the forward propagation, this participant **sums up outputs of all the bottom models** to get the final output.
- Every participant holds **partial features**.
- The **labels** are privately owned by one participant.
- One of the participants without labels is the **adversary**, whose goal is to infer the privately owned labels.





Attack 3: Direct Label Inference Attack

- For VFL without model splitting, the adversary is able to receive the **gradients of the final prediction layer.**

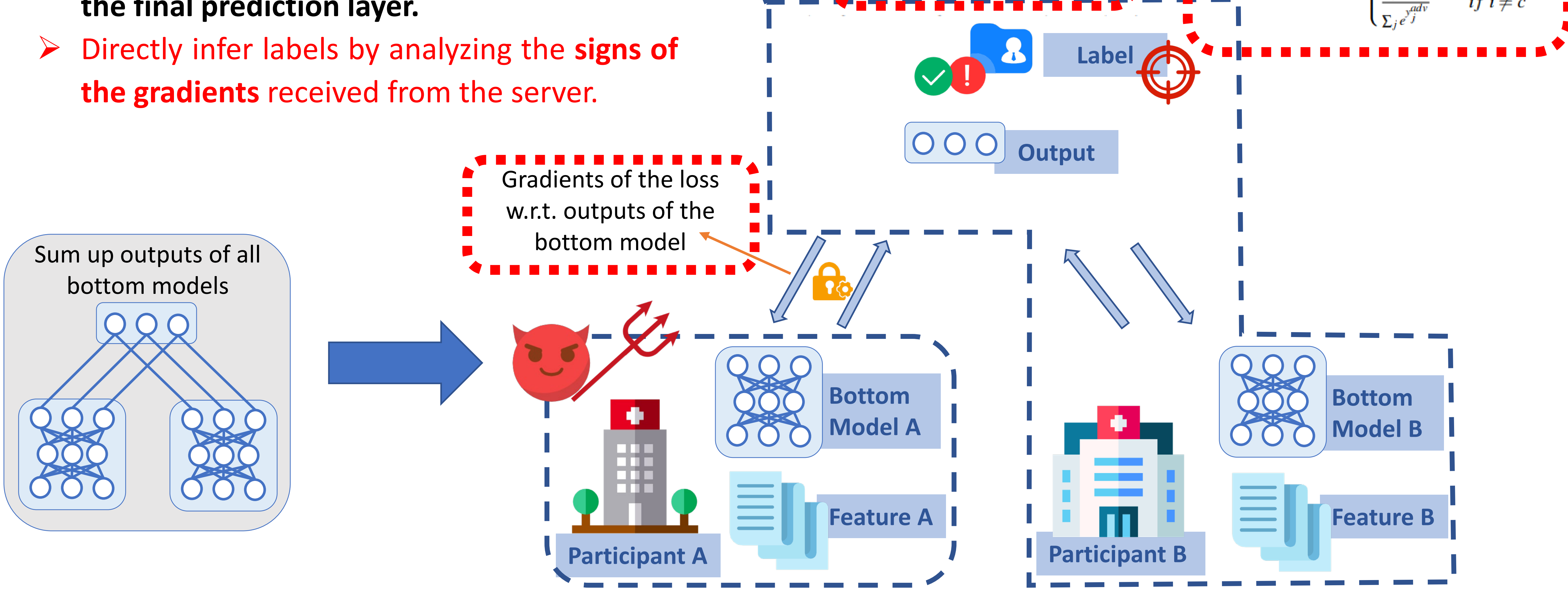


Attack 3: Direct Label Inference Attack

- For VFL without model splitting, the adversary is able to receive the **gradients of the final prediction layer.**
- Directly infer labels by analyzing the **signs of the gradients** received from the server.

$$loss(x,c) = -log\frac{e^{\sum_k y_c^k}}{\sum_j e^{\sum_k y_j^k}}$$

$$g_i^{adv} = \frac{\partial loss(x,c)}{\partial y_i^{adv}} = -\frac{\partial log e^{y_c^{adv}} - \partial log \sum_j e^{y_j^{adv}}}{\partial y_i^{adv}} = \begin{cases} -1 + \frac{e^{y_i^{adv}}}{\sum_j e^{y_j^{adv}}} & if\ i = c \\ \frac{e^{y_i^{adv}}}{\sum_j e^{y_j^{adv}}} & if\ i \neq c \end{cases}$$



Attack Evaluation

Experimental Setup

Datasets and model architectures

- **Various data types**

  Image, text, numerical feature and categorical feature.

- **Various model architectures**

  ResNet, BERT and fully connected neural networks.

- The VFL models get **good performance on the original tasks**.
  - Top-1 accuracy:
    - CIFAR-10: 82.80%
    - CINIC-10: 73.69%
    - Yahoo Answers: 71.67%
    - Criteo: 71.32%
  - Top-5 accuracy on CIFAR-100: 75.11%
  - F1 score on BHI: 83.40%

Dataset	Bottom Model Architecture	Top Model Architecture
CIFAR-10	ResNet-18	FCNN-4
CIFAR-100	ResNet-18	FCNN-4
CINIC-10	ResNet-18	FCNN-4
Yahoo Answers	BERT	FCNN-4
Criteo	FCNN-3	FCNN-3
BHI	ResNet-18	FCNN-4

Experimental Setup

Datasets and model architectures

- **Various data types**

  Image, text, numerical feature and categorical feature.

- **Various model architectures**

  ResNet, BERT and fully connected neural networks.

- The VFL models get **good performance on the original tasks**.
  - Top-1 accuracy:
    - CIFAR-10: 82.80%
    - CINIC-10: 73.69%
    - Yahoo Answers: 71.67%
    - Criteo: 71.32%
  - Top-5 accuracy on CIFAR-100: 75.11%
  - F1 score on BHI: 83.40%

Dataset	**Bottom Model Architecture**	**Top Model Architecture**
CIFAR-10	ResNet-18	FCNN-4
CIFAR-100	ResNet-18	FCNN-4
CINIC-10	ResNet-18	FCNN-4
Yahoo Answers	BERT	FCNN-4
Criteo	FCNN-3	FCNN-3
BHI	ResNet-18	FCNN-4

Experimental Setup

Datasets and model architectures

- **Various data types**

  Image, text, numerical feature and categorical feature.

- **Various model architectures**

  ResNet, BERT and fully connected neural networks.

- The VFL models get **good performance on the original tasks**.
  - Top-1 accuracy:
    - CIFAR-10: 82.80%
    - CINIC-10: 73.69%
    - Yahoo Answers: 71.67%
    - Criteo: 71.32%
  - Top-5 accuracy on CIFAR-100: 75.11%
  - F1 score on BHI: 83.40%

Dataset	Bottom Model Architecture	Top Model Architecture
CIFAR-10	ResNet-18	FCNN-4
CIFAR-100	ResNet-18	FCNN-4
CINIC-10	ResNet-18	FCNN-4
Yahoo Answers	BERT	FCNN-4
Criteo	FCNN-3	FCNN-3
BHI	ResNet-18	FCNN-4

Performance of the Passive/Active/Direct Label Inference Attack

- Good attack performance.
- The active label inference attack outperforms the passive label inference attack.

Dataset	Train Set Size	Test Set Size	Number of Classes	Known Label Quantity Per Class	Metric	Attack Performance			
						Train Set		Test Set	
						Passive	Active	Passive	Active
CIFAR-10	50,000	10,000	10	4	Top-1 Acc	0.8024	**0.8484**	0.6299	**0.6342**
CIFAR-100	50,000	10,000	100	4	Top-5 Acc	0.6267	**0.6732**	0.4319	**0.4700**
CINIC-10	180,000	90,000	10	4	Top-1 Acc	0.7206	**0.7818**	0.5440	**0.5995**
Yahoo Answers	50,000	20,000	10	10	Top-1 Acc	0.6335	**0.6424**	0.6370	**0.6419**
Criteo	80,000	20,000	2	50	Top-1 Acc	0.6828	**0.6879**	0.6785	**0.6830**
BHI	69,181	17,296	2	35	F1 Score	0.7614	**0.7824**	0.7519	**0.7673**

- The direct label inference attack can infer all labels in the training dataset (100% top-1 accuracy).

Performance of the Passive/Active/Direct Label Inference Attack

- ➢ Good attack performance.
- ➢ The active label inference attack outperforms the passive label inference attack.

Dataset	Train Set Size	Test Set Size	Number of Classes	Known Label Quantity Per Class	Metric	Attack Performance			
						Train Set		Test Set	
						Passive	Active	Passive	Active
CIFAR-10	50,000	10,000	10	4	Top-1 Acc	0.8024	**0.8484**	0.6299	**0.6342**
CIFAR-100	50,000	10,000	100	4	Top-5 Acc	0.6267	**0.6732**	0.4319	**0.4700**
CINIC-10	180,000	90,000	10	4	Top-1 Acc	0.7206	**0.7818**	0.5440	**0.5995**
Yahoo Answers	50,000	20,000	10	10	Top-1 Acc	0.6335	**0.6424**	0.6370	**0.6419**
Criteo	80,000	20,000	2	50	Top-1 Acc	0.6828	**0.6879**	0.6785	**0.6830**
BHI	69,181	17,296	2	35	F1 Score	0.7614	**0.7824**	0.7519	**0.7673**

- ➢ The direct label inference attack can infer all labels in the training dataset (100% top-1 accuracy).

Performance of the Passive/Active/Direct Label Inference Attack

- Good attack performance.
- The active label inference attack outperforms the passive label inference attack.

Dataset	Train Set Size	Test Set Size	Number of Classes	Known Label Quantity Per Class	Metric	Attack Performance			
						Train Set		Test Set	
						Passive	Active	Passive	Active
CIFAR-10	50,000	10,000	10	4	Top-1 Acc	0.8024	**0.8484**	0.6299	**0.6342**
CIFAR-100	50,000	10,000	100	4	Top-5 Acc	0.6267	**0.6732**	0.4319	**0.4700**
CINIC-10	180,000	90,000	10	4	Top-1 Acc	0.7206	**0.7818**	0.5440	**0.5995**
Yahoo Answers	50,000	20,000	10	10	Top-1 Acc	0.6335	**0.6424**	0.6370	**0.6419**
Criteo	80,000	20,000	2	50	Top-1 Acc	0.6828	**0.6879**	0.6785	**0.6830**
BHI	69,181	17,296	2	35	F1 Score	0.7614	**0.7824**	0.7519	**0.7673**

- The direct label inference attack can infer all labels in the training dataset (100% top-1 accuracy).

Impact of the Amount of Auxiliary Labeled Data & Comparison with Direct Semi-supervised Learning

- More auxiliary labeled samples indeed increases the attack accuracy.
- However, as the number of auxiliary labeled samples grows, the attack accuracy increases more and more slowly.
- The trained bottom model contains much information for label inference.

Known Label Quantity	**Passive Label Inference**		**Direct Semi**	
	Training Dataset	Test Dataset	Training Dataset	Test Dataset
10	**0.6554**	**0.5235**	0.1157	0.1138
20	**0.7080**	**0.5542**	0.1187	0.1166
40	**0.8024**	**0.6299**	0.1698	0.1683
120	**0.8406**	**0.6305**	0.1866	0.1846
320	**0.8544**	**0.6392**	0.3286	0.3218

Experiment on CIFAR-10. Attack performance is measured by top-1 accuracy.

Impact of the Amount of Auxiliary Labeled Data & Comparison with Direct Semi-supervised Learning

- More auxiliary labeled samples indeed increases the attack accuracy.
- However, as the number of auxiliary labeled samples grows, the attack accuracy increases more and more slowly.
- The trained bottom model contains much information for label inference.

Known Label Quantity	**Passive Label Inference**		**Direct Semi**	
	Training Dataset	Test Dataset	Training Dataset	Test Dataset
10	**0.6554**	**0.5235**	0.1157	0.1138
20	**0.7080**	**0.5542**	0.1187	0.1166
40	**0.8024**	**0.6299**	0.1698	0.1683
120	**0.8406**	**0.6305**	0.1866	0.1846
320	**0.8544**	**0.6392**	0.3286	0.3218

Experiment on CIFAR-10. Attack performance is measured by top-1 accuracy.

Impact of the Amount of Auxiliary Labeled Data & Comparison with Direct Semi-supervised Learning

- More auxiliary labeled samples indeed increases the attack accuracy.
- However, as the number of auxiliary labeled samples grows, the attack accuracy increases more and more slowly.
- The trained bottom model contains much information for label inference.

Known Label Quantity	Passive Label Inference		Direct Semi	
	Training Dataset	Test Dataset	Training Dataset	Test Dataset
10	**0.6554**	**0.5235**	0.1157	0.1138
20	**0.7080**	**0.5542**	0.1187	0.1166
40	**0.8024**	**0.6299**	0.1698	0.1683
120	**0.8406**	**0.6305**	0.1866	0.1846
320	**0.8544**	**0.6392**	0.3286	0.3218

Experiment on CIFAR-10. Attack performance is measured by top-1 accuracy.

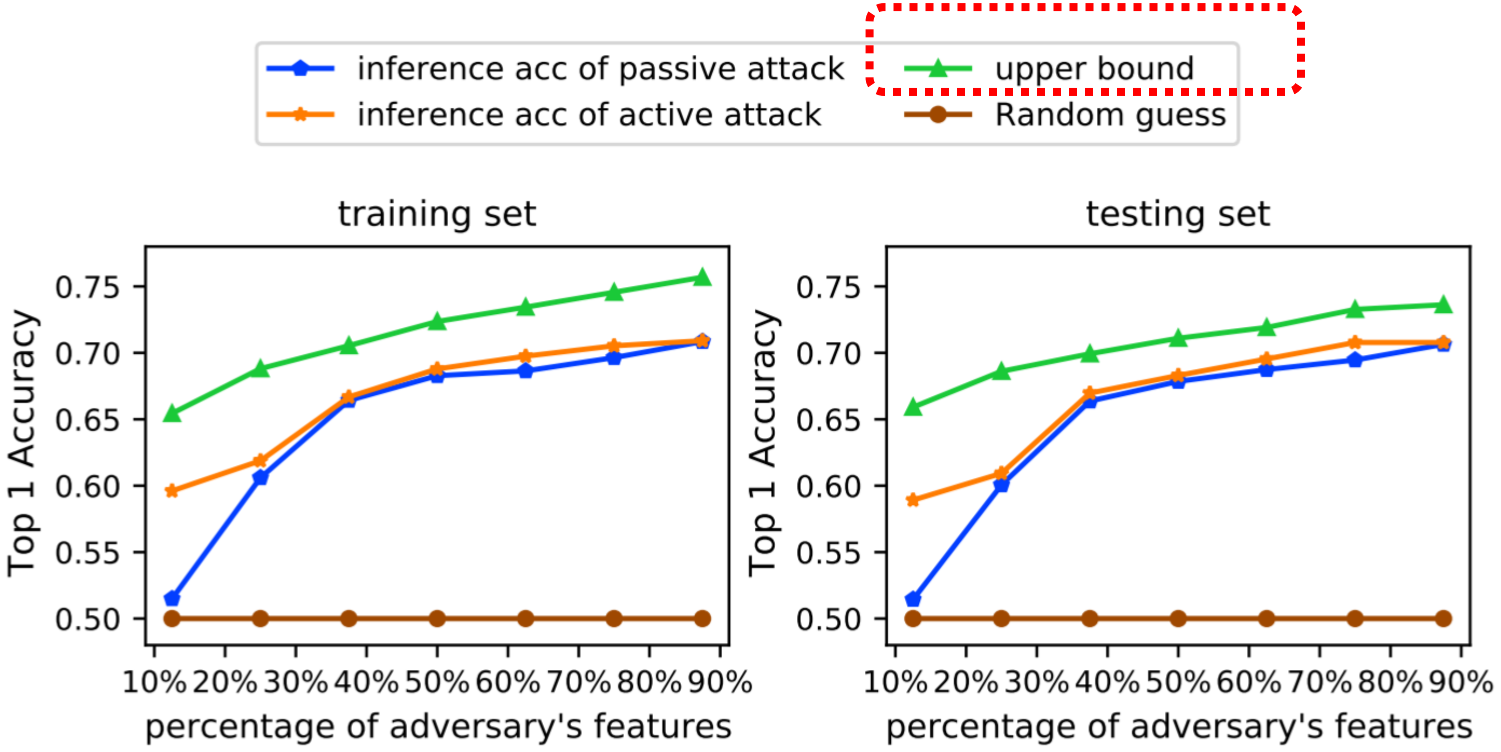
The Active Label Inference Attack's Influence on the Federated Model's Performance on the Original Task

- The active label inference attack has a **very small impact** on the federated model's performance on **the original task.**

Dataset	Metric	Model Performance under:	
		No Attack	Active Attack
CIFAR-10	Top-1 Acc	**0.8280**	0.8139
CIFAR-100	Top-5 Acc	**0.7511**	0.7500
CINIC-10	Top-1 Acc	0.7369	**0.7400**
Yahoo Answers	Top-1 Acc	**0.7167**	0.7120
Criteo	Top-1 Acc	**0.7132**	0.7128
BHI	F1 Score	0.8340	**0.8504**

The Impact of the Quantity of the Adversary’s Features

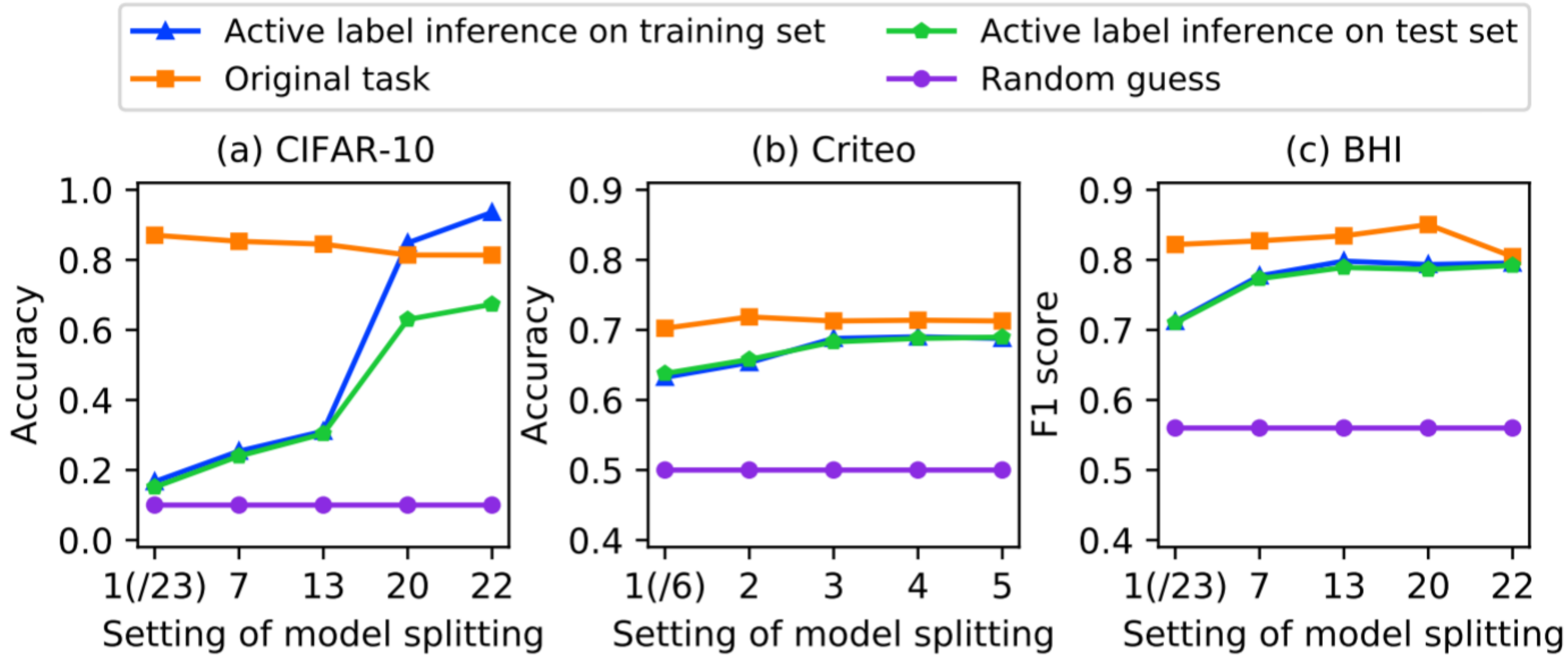
- The quantity of the adversary’s local features determines **the upper bound** of the attack performance.
- The active label inference attack can only boost the attack performance **within this upper bound**.



The impact of the quantity of the adversary’s features on Criteo. The *upper bound* is obtained using all the labels to directly train an inference model with the adversary’s features.

The Impact of the Complexity of the Bottom Model

- The more layers the adversary's bottom model has, the better the active label inference attack performs.
- VFL models with simpler tasks face a greater risk of label leakage.



Performance of the Active Attack in Multi-party Setting

- Attack performance degrades as the number of participants increases.
- Label inference attacks threats multi-party VFL even when there are 8 participants.

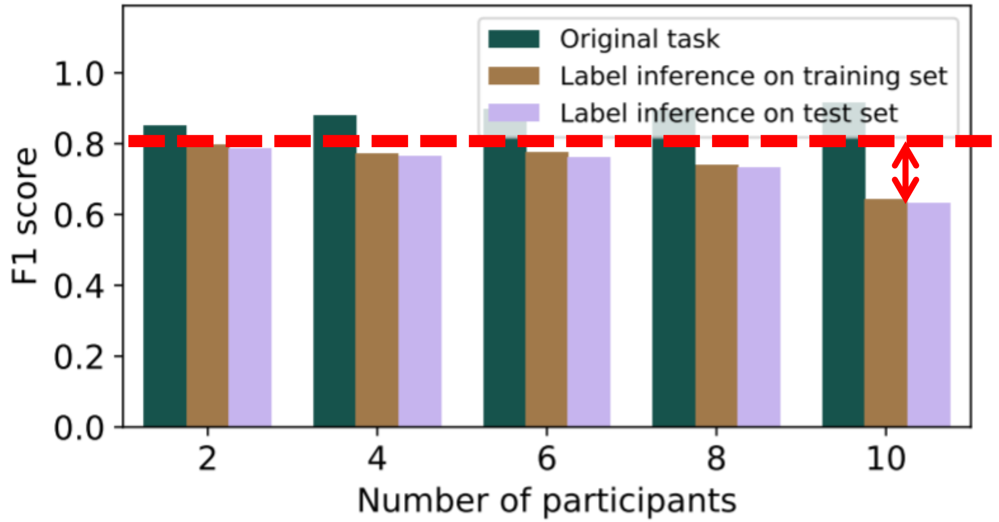


Figure 3: Performance of the active attack in multi-party setting on BHI.

Performance of the Active Attack in Multi-party Setting

- Attack performance degrades as the number of participants increases.
- Label inference attacks threats multi-party VFL even when there are 8 participants.

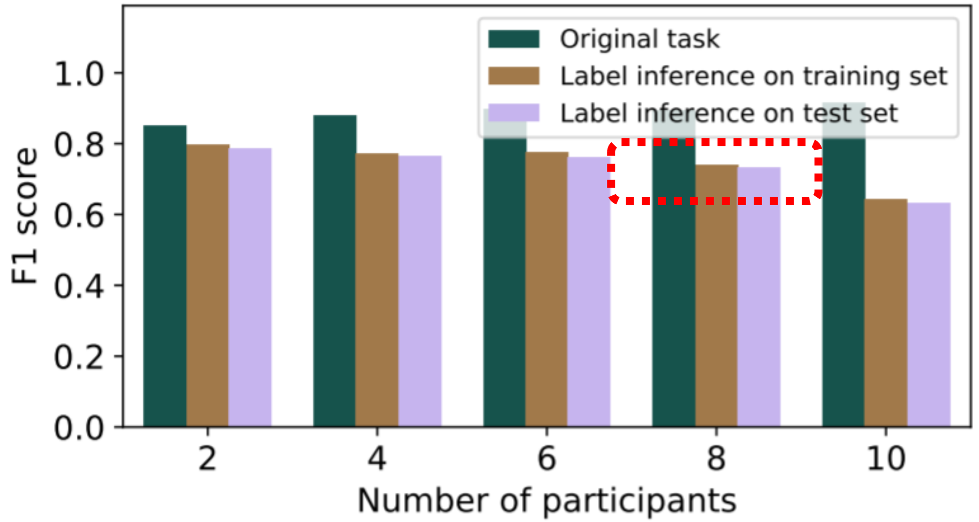
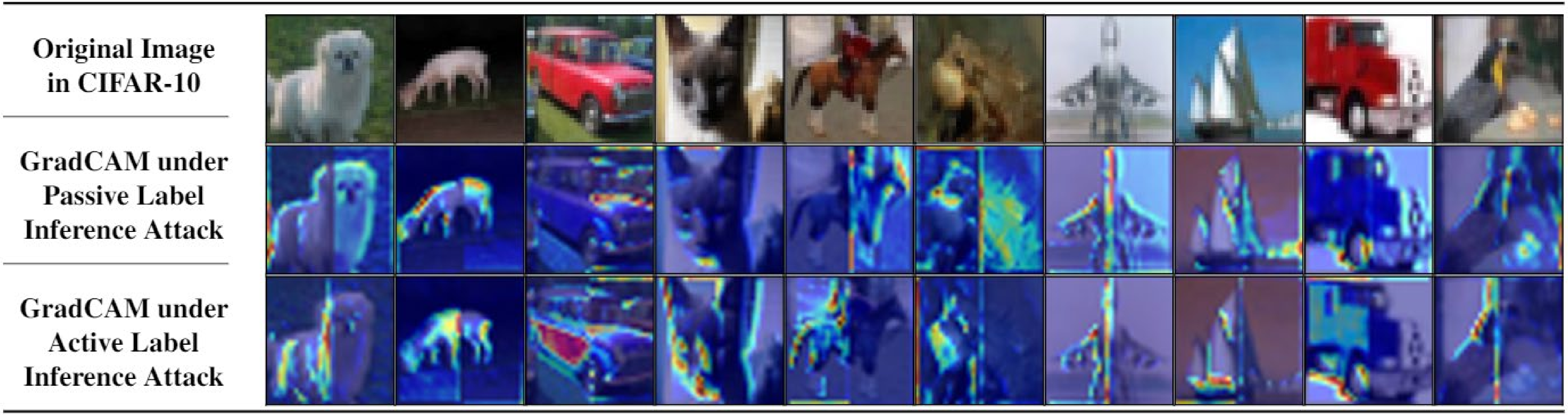


Figure 3: Performance of the active attack in multi-party setting on BHI.

Analysis

Why the Active Label Inference Attack Works (1)

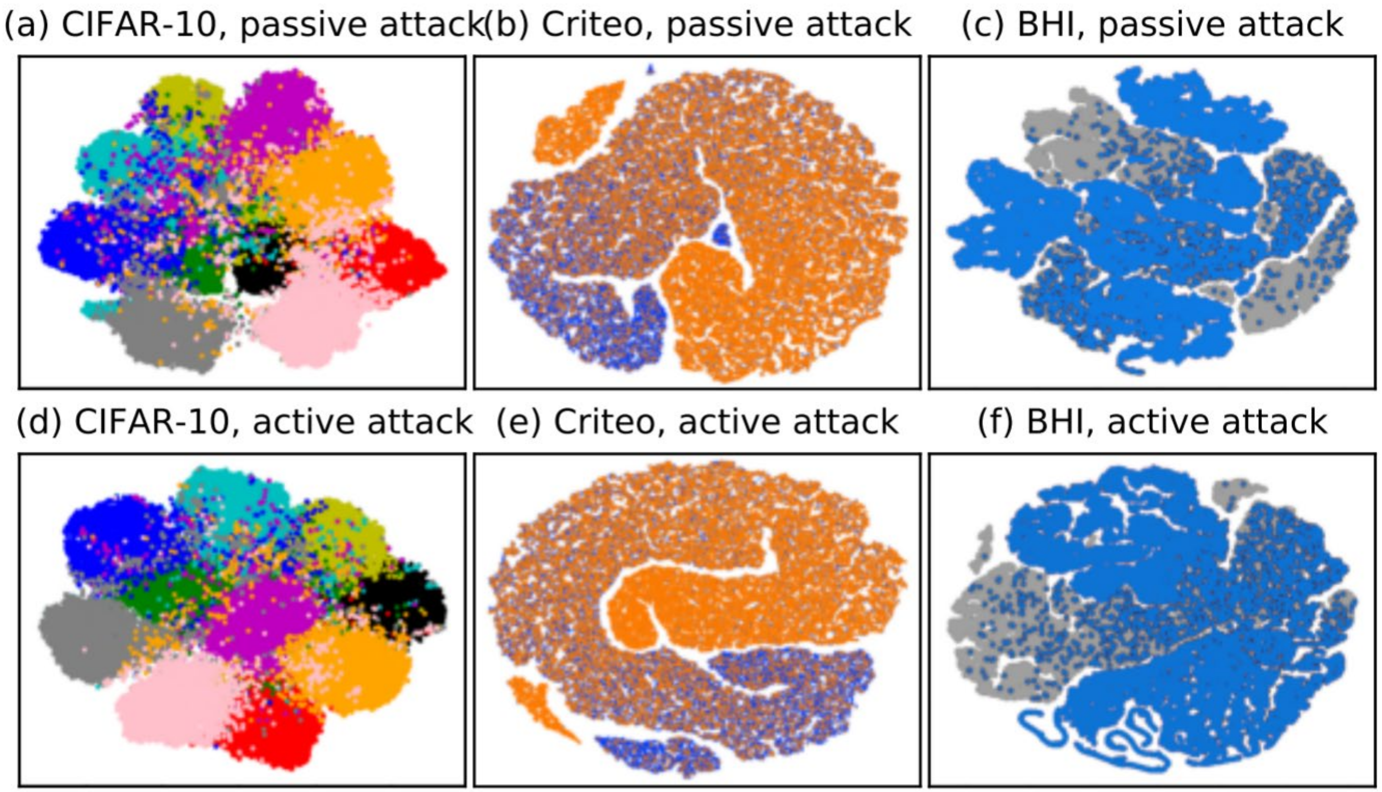
- More attention is drawn to the adversary's datum under the active attack.



GradCAM visualization of some training samples under the passive or active label inference attacks on CIFAR-10. The left half of the image is the datum of the adversary.

Why the Active Label Inference Attack Works (2)

- The adversary's bottom model learns **better representations of raw local data** under the active attack.

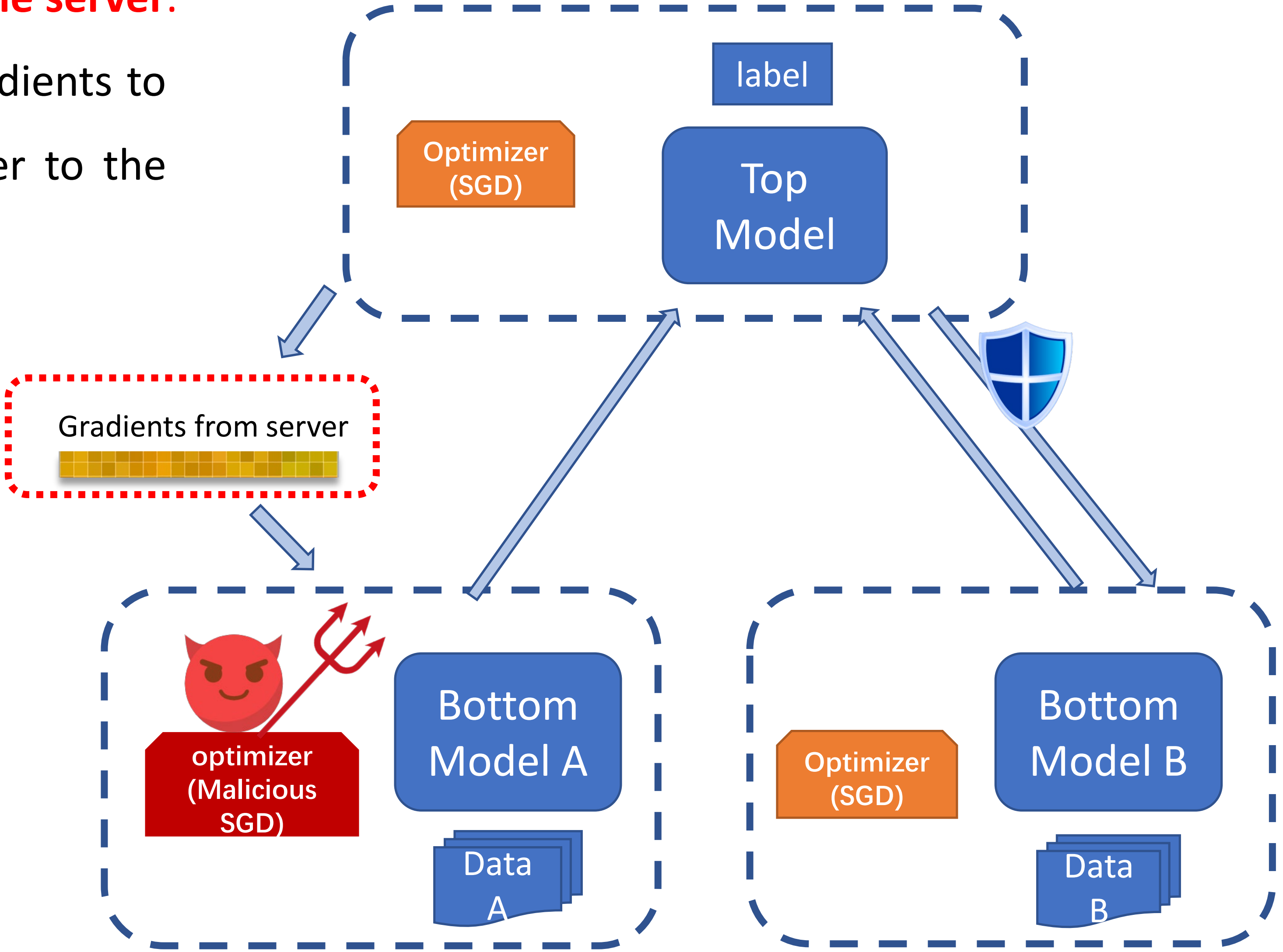


T-SNE projection of the outputs of the adversary's bottom model. Different color represents different labels.

Defense Evaluation

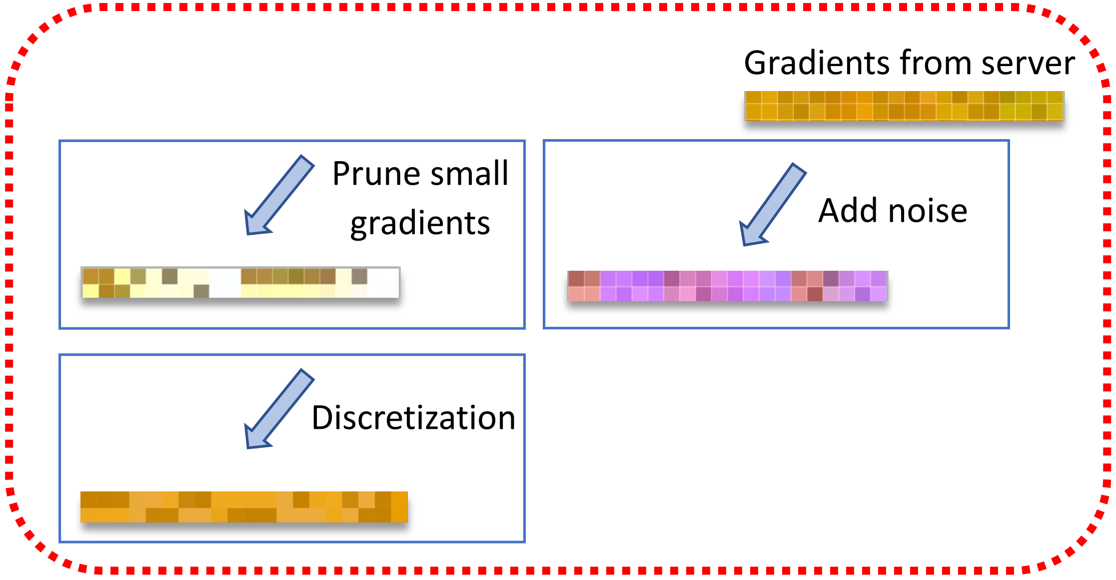
Possible Defense

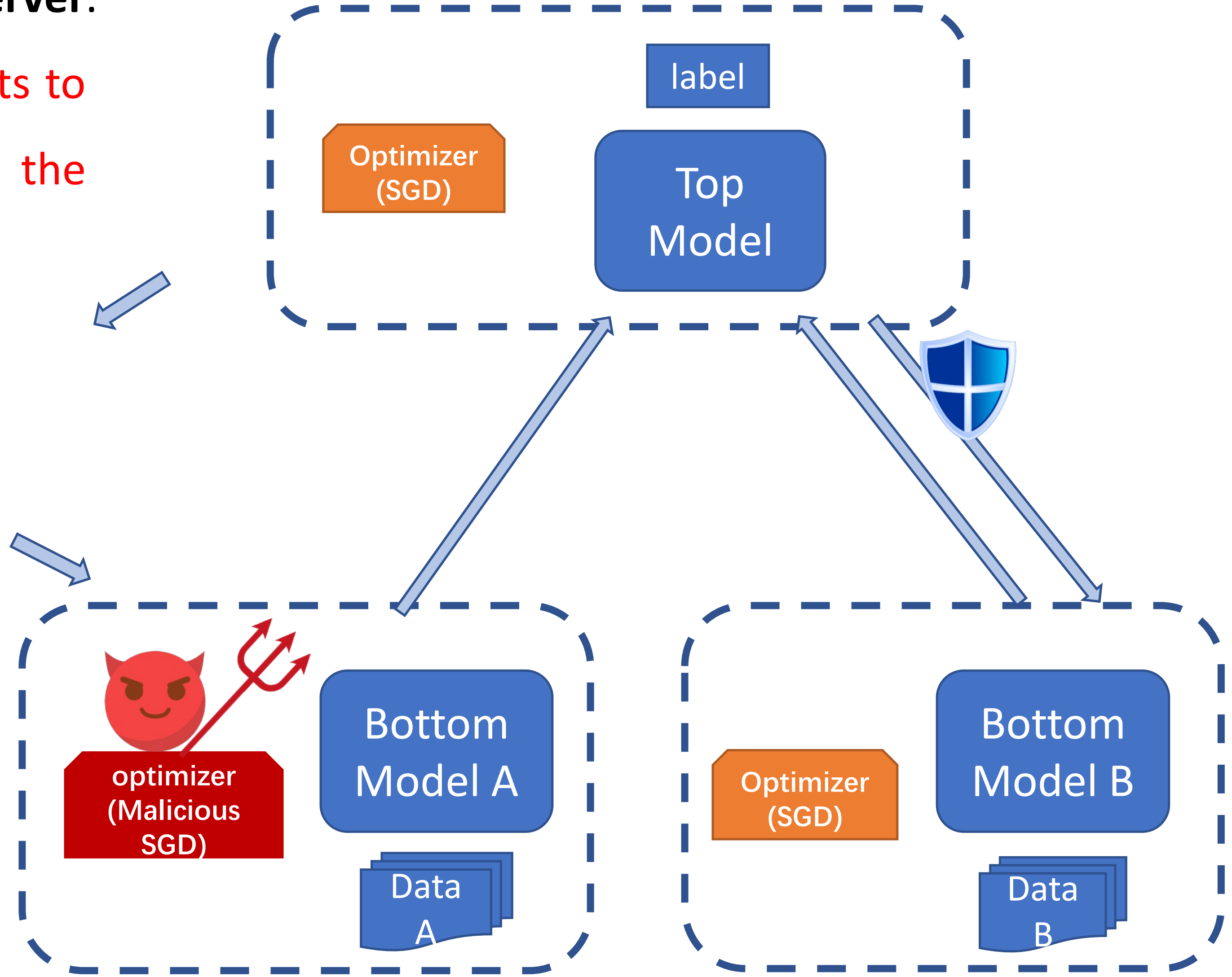
- In the training process of VFL, the **only** information sent to the adversary is the **gradients from the server**.
- Defense strategies can be applied to the gradients to prevent information leakage from the server to the adversary.



Possible Defense

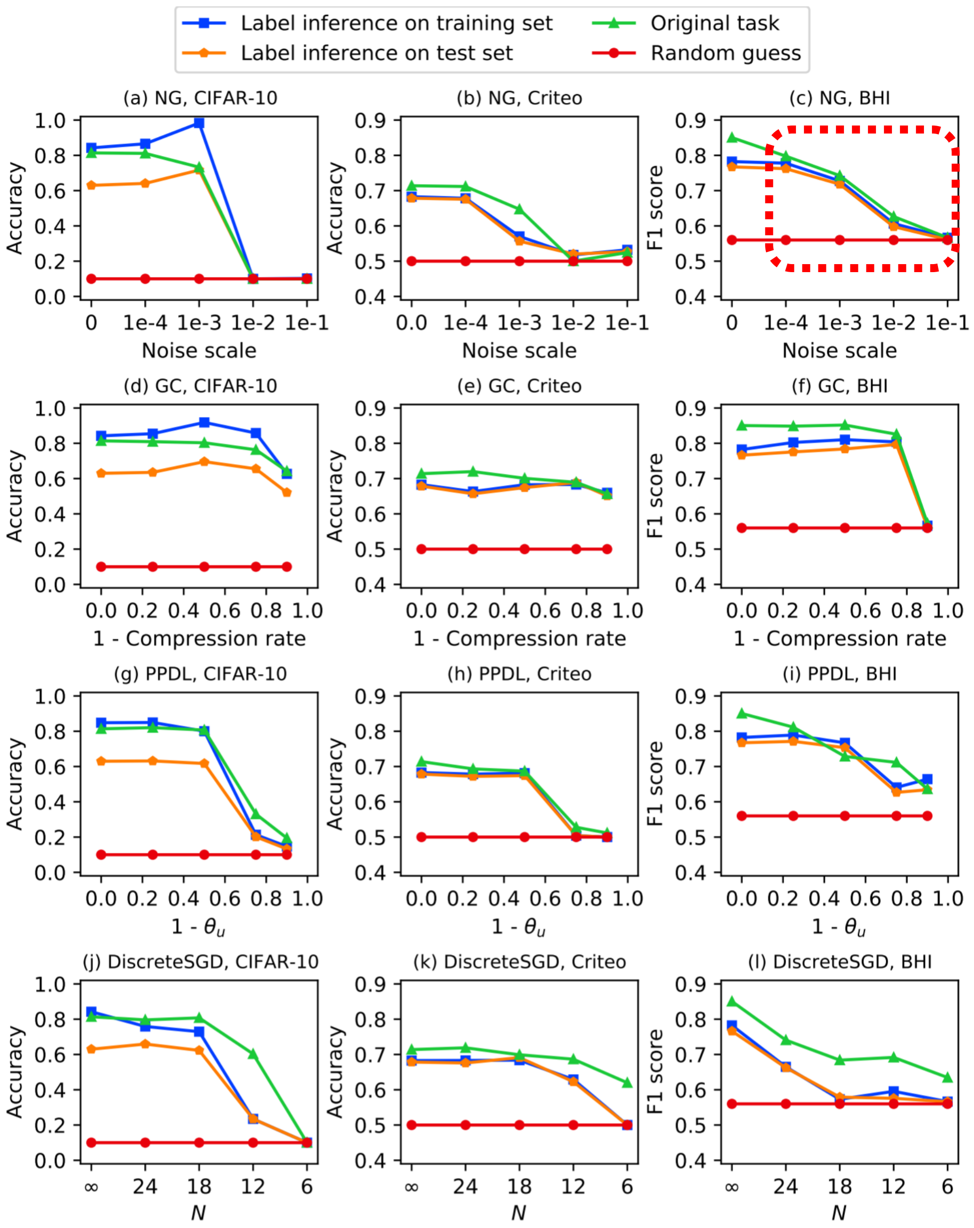
- In the training process of VFL, the **only** information sent to the adversary is the **gradients from the server**.
- Defense strategies can be applied to the gradients to prevent information leakage from the server to the adversary.





Defense Against the Active Label Inference Attack

- Three mainstream defense approaches: noisy gradients, gradient compression and privacy-preserving deep learning.
- DiscreteSGD, a customized version of the defense approach signSGD.
- These defense approaches are **not effective** against our active label inference attack.



Defense Against the Direct Label Inference Attack

- Two of the four evaluated defense approaches can successfully mitigate the direct label inference attack.

Table 6: Defenses against the direct label inference attack on CIFAR-10.

Defense Approach	Parameter	Parameter Set Value	Model Accuracy	Attack Accuracy
Noisy Gradients	Noise Scale	1e-4	0.8347	0.8063
		1e-3	0.8318	0.4906
		1e-2	0.7191	0.2452
		1e-1	0.1000	0.1265
Gradient Compression	Compression Rate	75%	0.8248	0.9997
		50%	0.8259	0.9931
		25%	0.8049	0.9245
		10%	0.1000	0.0058
Privacy-preserving Deep Learning	θ_u	0.75	0.8189	0.3904
		0.50	0.8216	0.3891
		0.25	0.1993	0.0972
		0.10	0.1000	0.0430
Discrete SGD	N	24	0.8145	0.9763
		18	0.7962	0.9330
		12	0.7471	0.9399
		6	0.6575	0.9087

Conclusion

Conclusion

- We reveal and shed lights on the new label leakage issue of VFL.
- We present three types of label inference attacks against VFL. We evaluate our attacks on various tasks under both two-participant and multi-participant settings and achieve good attack performance.
- We share insights about the underlying working mechanism of the active label inference attack, and present visualized proofs.
- We evaluate four possible defenses against our attacks and find that they are not effective against the passive/active attack, which motivates future work on better defenses.

Q&A

fuchong@zju.edu.cn