ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models

Ahmed Salem*, Yang Zhang*§, Mathias Humbert†, Pascal Berrang*,
Mario Fritz*, Michael Backes*
*CISPA Helmholtz Center for Information Security,
{ahmed.salem, yang.zhang, pascal.berrang, fritz, backes}@cispa.saarland
†Swiss Data Science Center, ETH Zurich and EPFL, mathias.humbert@epfl.ch

Xinyu Wang

19-03-28

NSEC Lab

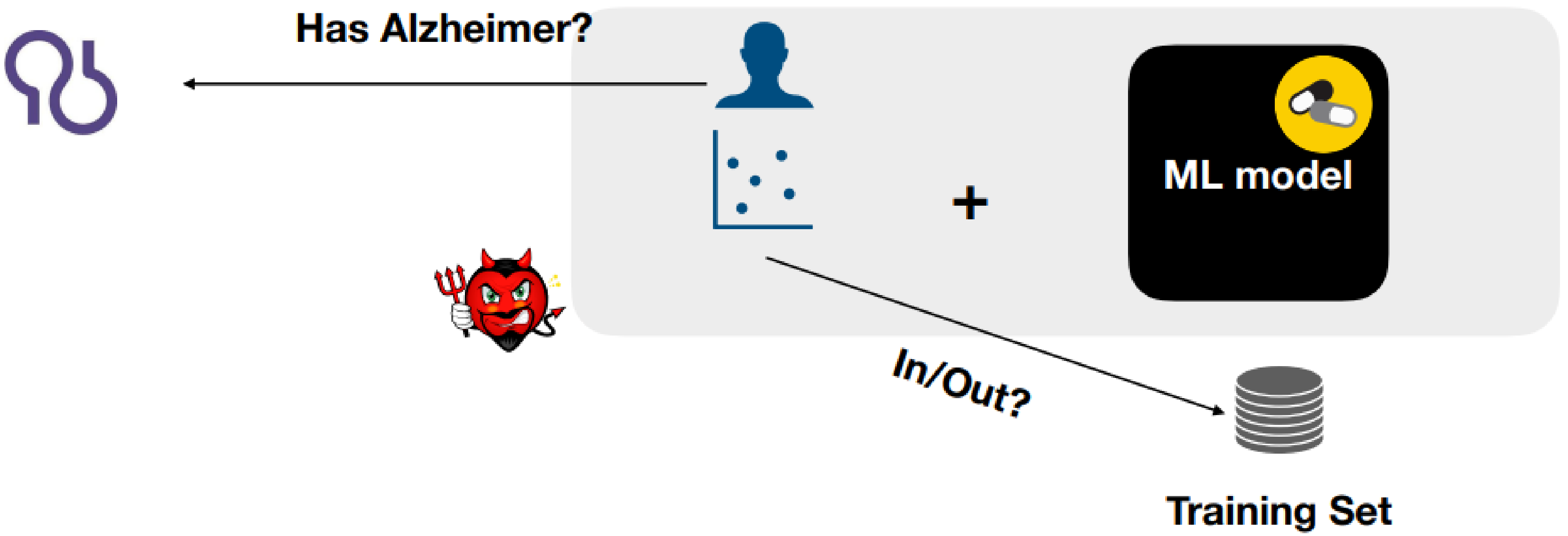
OUTLINE

- Background About Membership Inference Attack
- Commentary on Previous Work
- Proposed Attacks
- Proposed Defenses
- Conclusion

BACKGROUND

Training data can be sensitive:

- Financial data
- Location and activity data
- Biomedical data
- Etc.



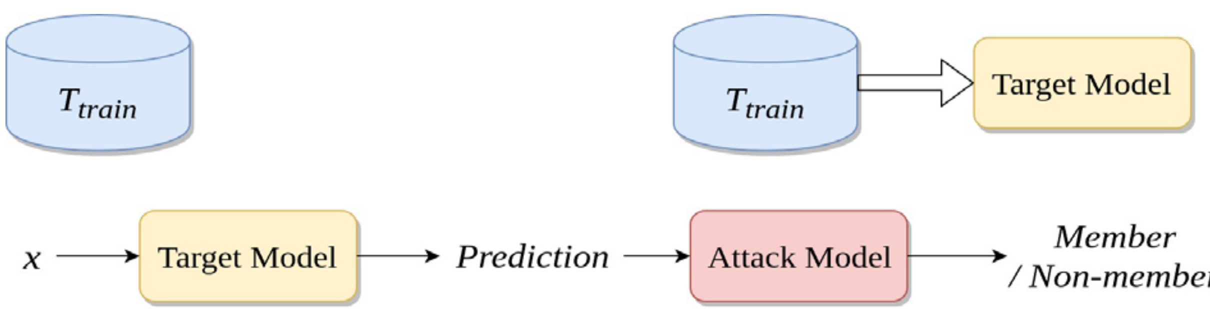
BACKGROUND

- Shokri et al. ,Oakland 2017

Membership Inference Attacks Against Machine Learning Models

Reza Shokri, Cornell Tech, shokri@cornell.edu
Marco Stronati*, INRIA, marco@stronati.org
Congzheng Song, Cornell, cs2296@cornell.edu
Vitaly Shmatikov, Cornell Tech, shmat@cs.cornell.edu

- **Membership Inference**: Given a machine learning model (target model) and a record (*x*), determine whether this record was used as part (member) of the model's training dataset or not.

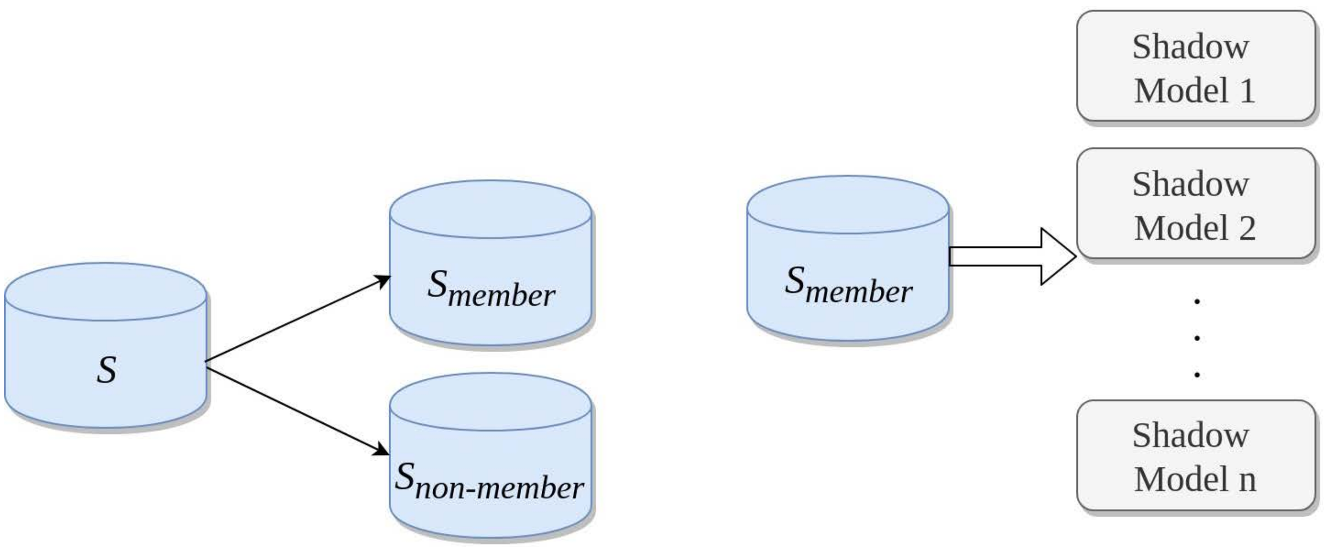


BACKGROUND

Shokri et al. proposed a three-step approach:

1. Shadow model training

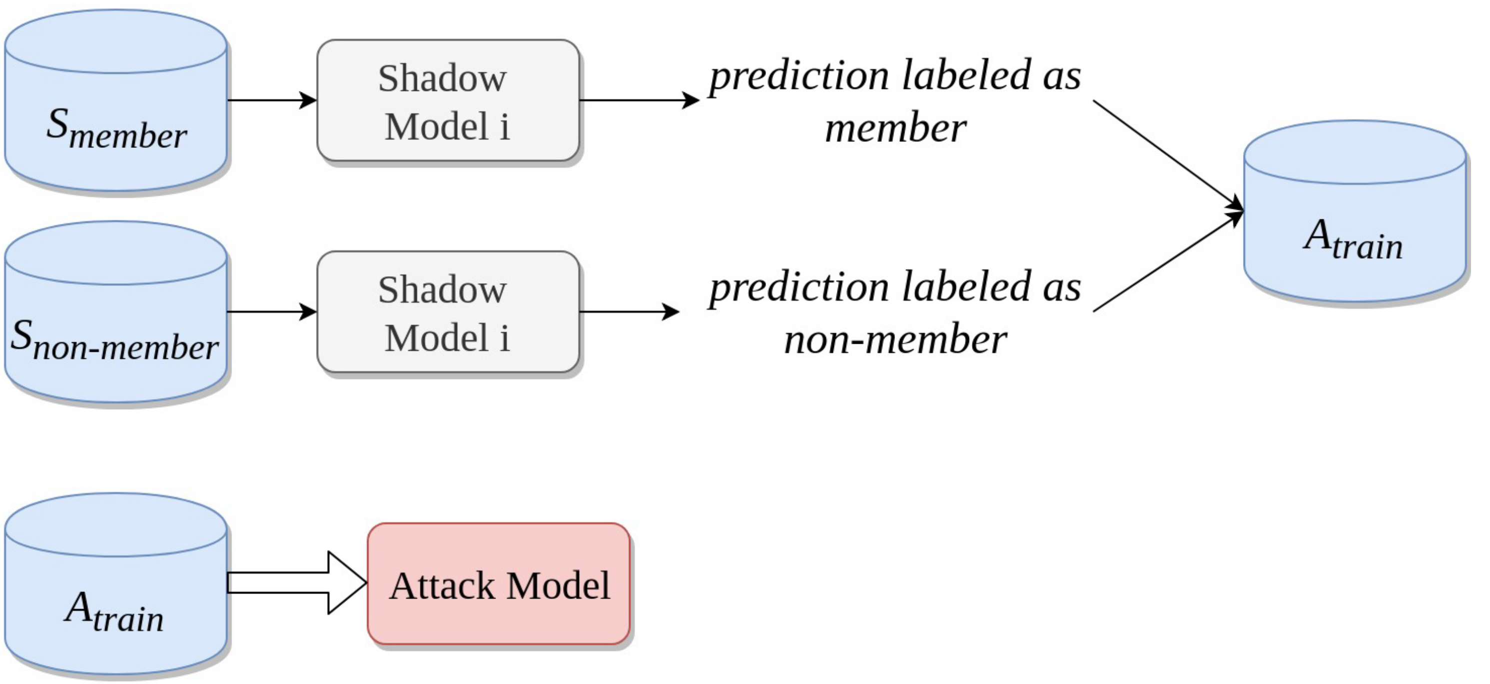
Assume the attacker can get a shadow training set S, which shares the same distribution with T_{train}.



BACKGROUND

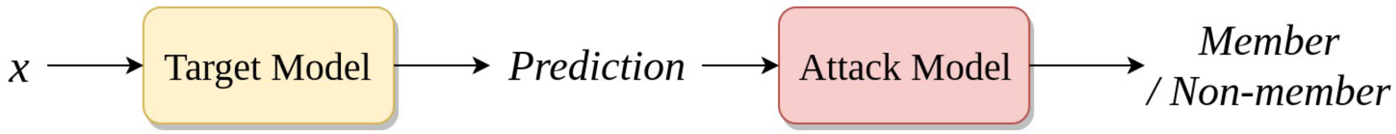
2. Attack model training

Get the attack training set A_{train} from shadow training set (S_{member} and $S_{non\text{-}member}$) and shadow models.



BACKGROUND

3. Membership inference



In the “attack model training” step we have modeled the relationship between prediction and membership

Therefore, with the prediction of data record *x*, we can predict the membership of *x*.

BACKGROUND

Three strong assumptions

- **Multiple shadow models**: The attacker has to train multiple shadow models
  - to obtain a large training dataset for the attack model
- **Model dependent**: The attacker knows the structure of the target model
  - training algorithm, and
  - hyperparameters
- **Data dependent**: The attacker can get a shadow training dataset S
  - S shares the same distribution with T_{train} (training dataset of the target model)

COMMENTARY

Three strong assumptions

- **Multiple shadow models**
- **Model dependent**
- **Data dependent**

These strong assumptions limit the scenario of the membership inference attack.

Therefore, this paper tries to relax these assumptions step-by-step.

PROPOSED ATTACKS

Strong assumptions:

1. **Multiple shadow models**
2. **Model dependent**
3. **Data dependent**

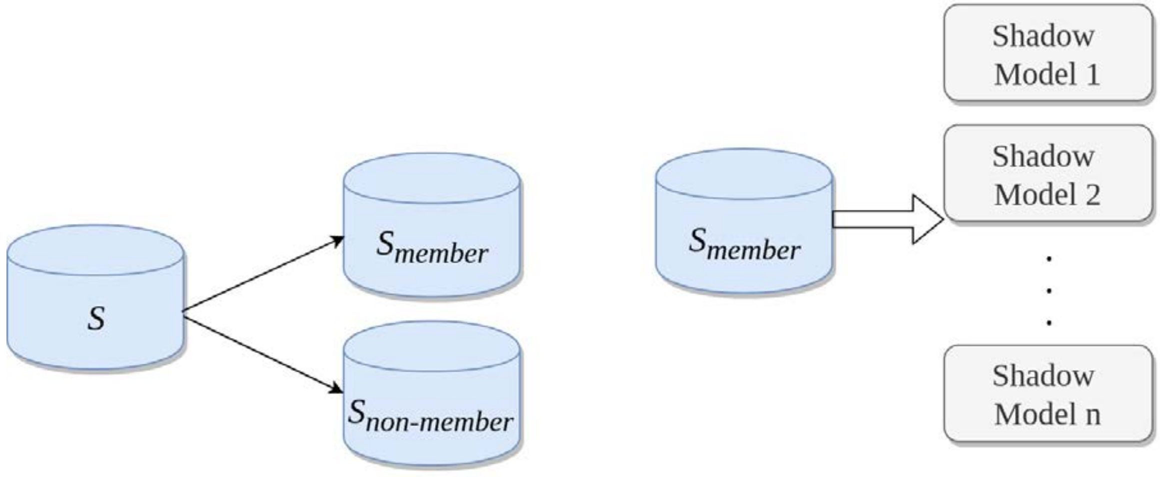
Relax strong assumptions step-by-step:

1. Relax assumption 1: using only one shadow model
2. Relax assumption 2: independence of model structure
3. Relax assumption 3: independence of data distribution

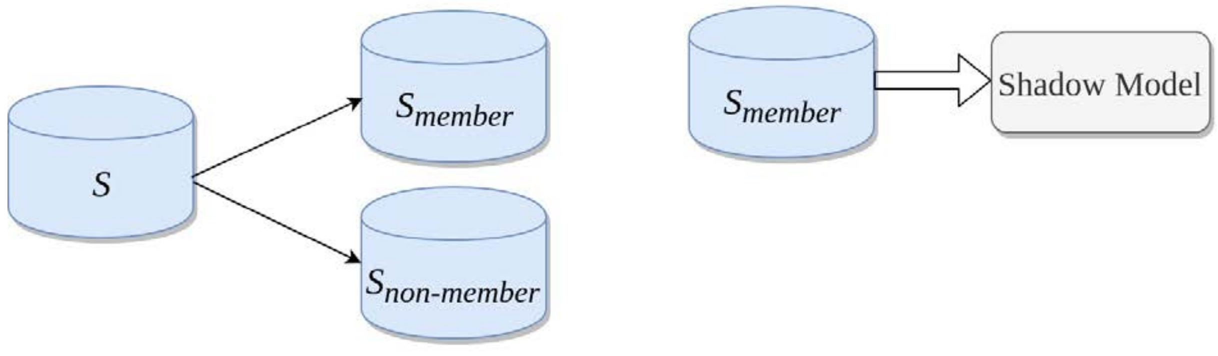
PROPOSED ATTACKS

Step 1: using only one shadow model

Shokri:



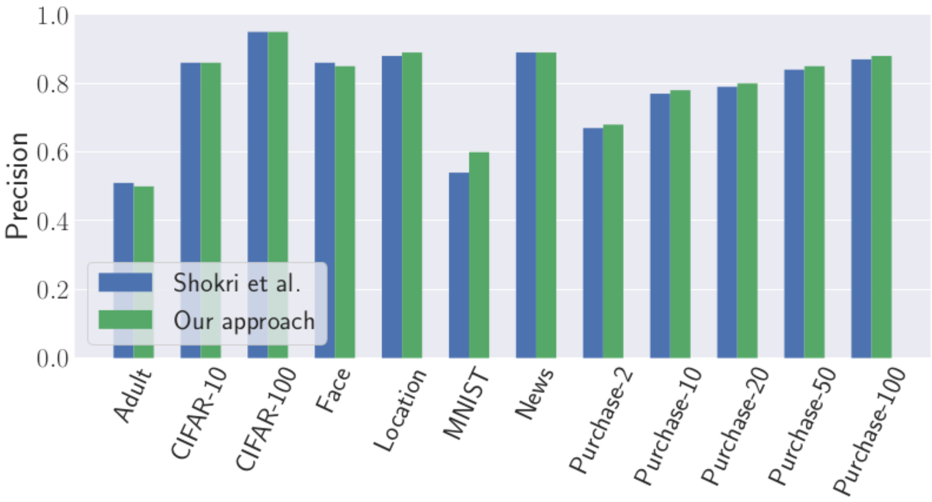
One shadow model:



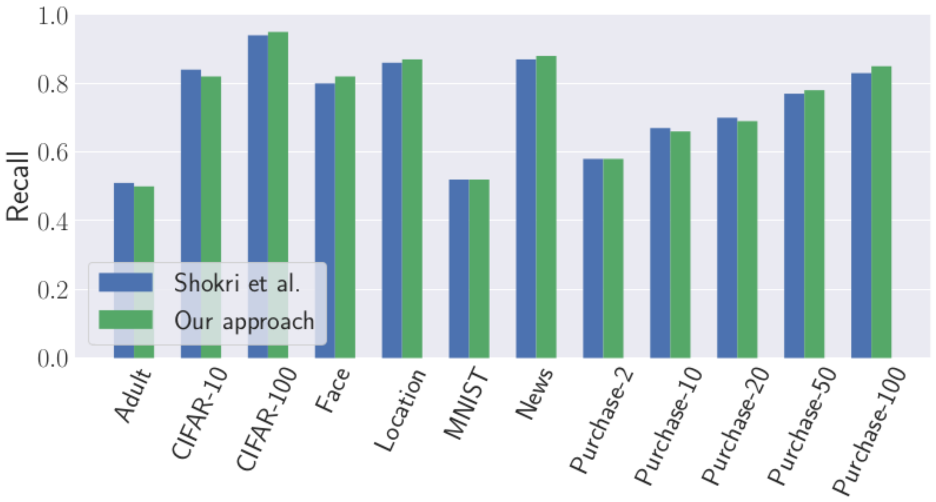
PROPOSED ATTACKS

Step 1: using only one shadow model

Results: Performance is similar to Shokri attack.



(a) Precision.



(b) Recall.

Fig. 1: Comparison of the first adversary's performance with Shokri et al.'s using all datasets. (a) precision, (b) recall.

PROPOSED ATTACKS

Step 2: independence of model structure

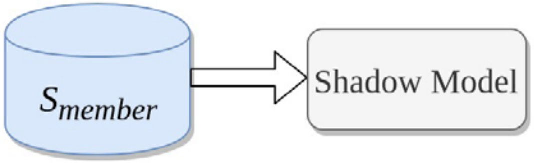
Experiments show:

- Changing hyperparameters have no significant effect on the performance
- Simply changing training algorithm of the shadow model leads to bad performance
  - Therefore this paper proposes a technique called *combining attack*

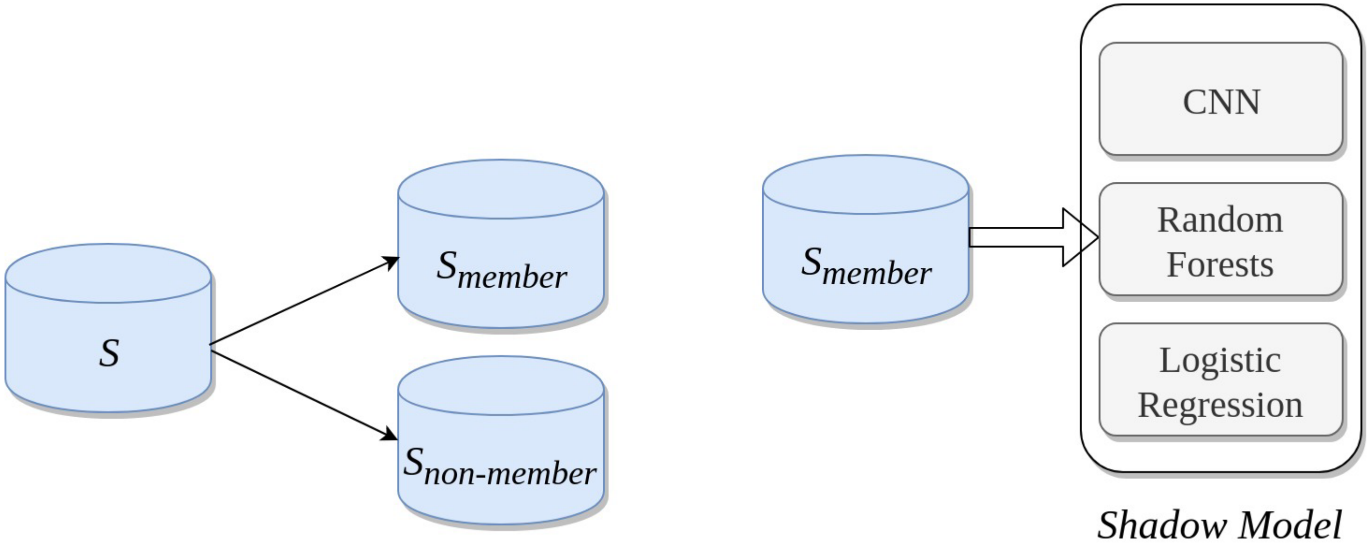
PROPOSED ATTACKS

Step 2: independence of model structure

One shadow model:



Combining attack: train sub-shadow models using a variety of different training algorithms and combine them



PROPOSED ATTACKS

Step 2: independence of model structure

Results: similar performance or even better

Classifier	With target model structure		Combining attack	
	Precision	Recall	Precision	Recall
Multilayer perceptron	0.86	0.86	0.88	0.85
Logistic regression	0.90	0.88	0.90	0.88
Random forests	1.0	1.0	0.94	0.93

PROPOSED ATTACKS

Step 3: independence of data distribution

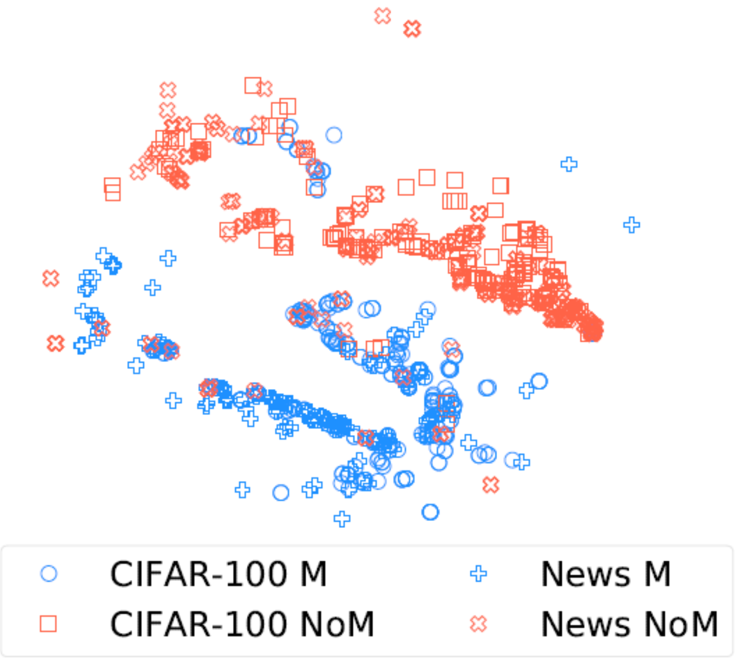
Data transferring attack: use dataset from a different distribution to train the shadow model

Target model:

Shadow model:

PROPOSED ATTACKS

Step 3: independence of data distribution



(a)

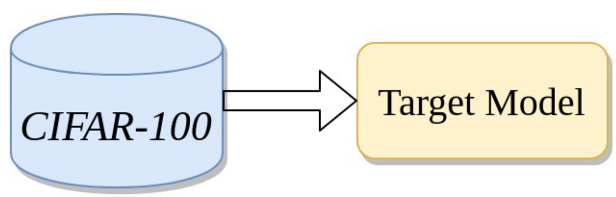
Intuition: different datasets share similar relations between prediction and membership

PROPOSED ATTACKS

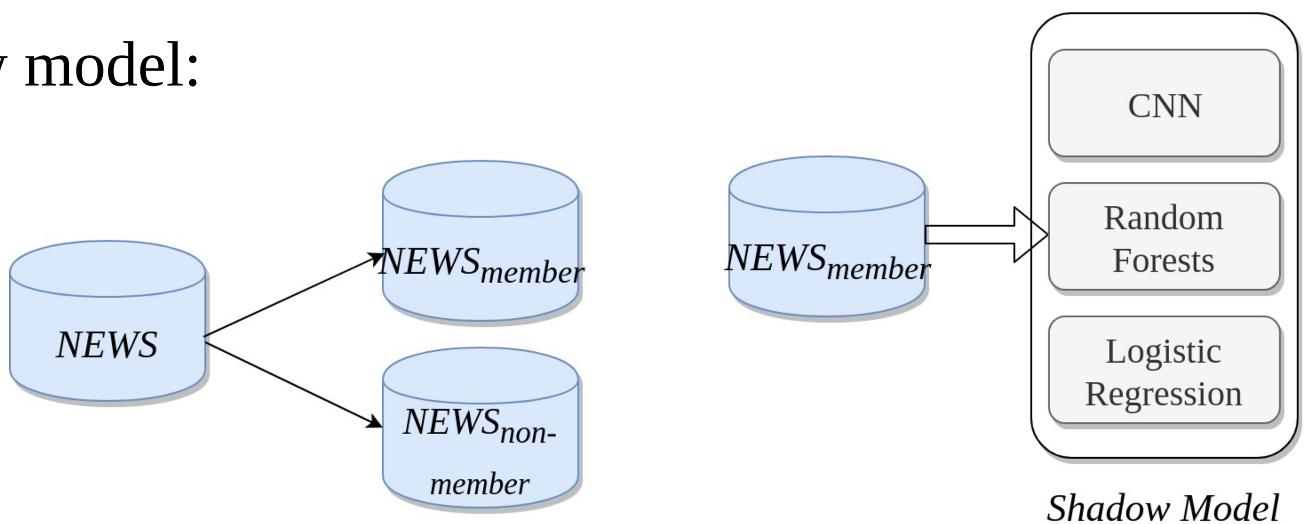
Step 3: independence of data distribution

Data transferring attack: use dataset from a different distribution to train the shadow model

Target model:



Shadow model:



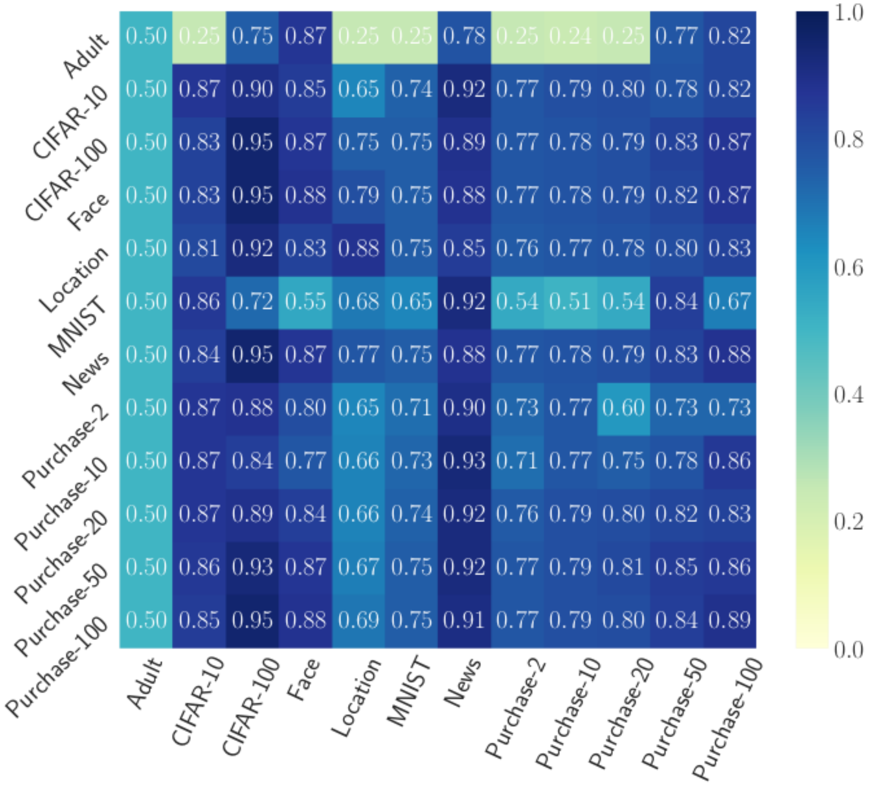
PROPOSED ATTACKS

Step 3: independence of data distribution

Results:

For instance,

- Use CIFAR-100 to attack Face: precision remains 0.95
- Use CIFAR-100 to attack News precision improves from 0.88 to 0.89

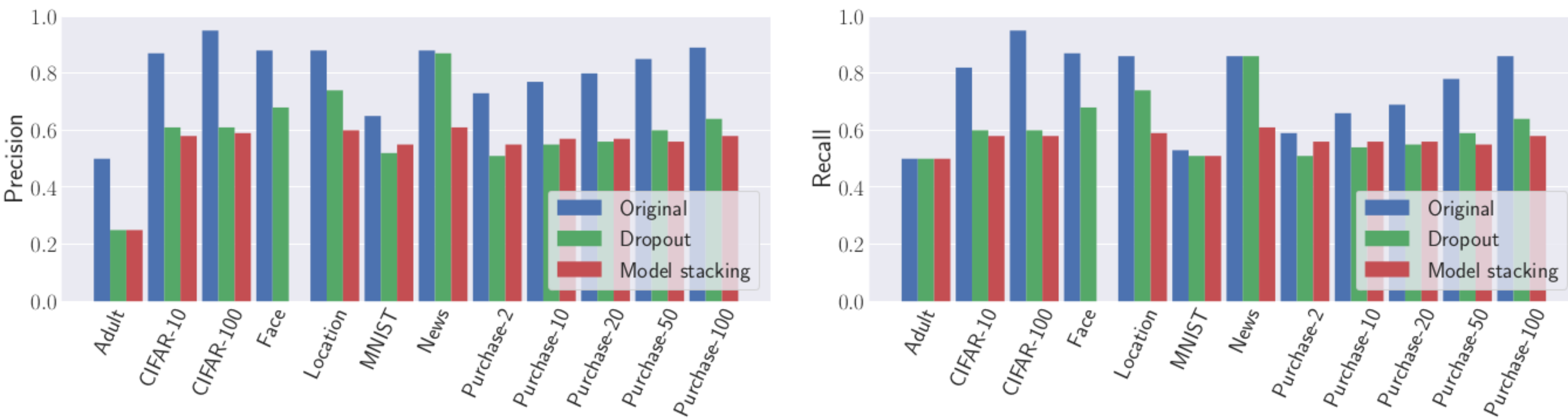


(a) Precision

PROPOSED DEFENSES

Principle: reduce overfitting

- Dropout
- Model Stacking



(a) (b)

Fig. 13: Comparison of the first adversary's performance under both of the defense mechanisms. (a) precision, (b) recall.

PROPOSED DEFENSES

Consider the effect on the target model's accuracy

- Dropout
- Model Stacking

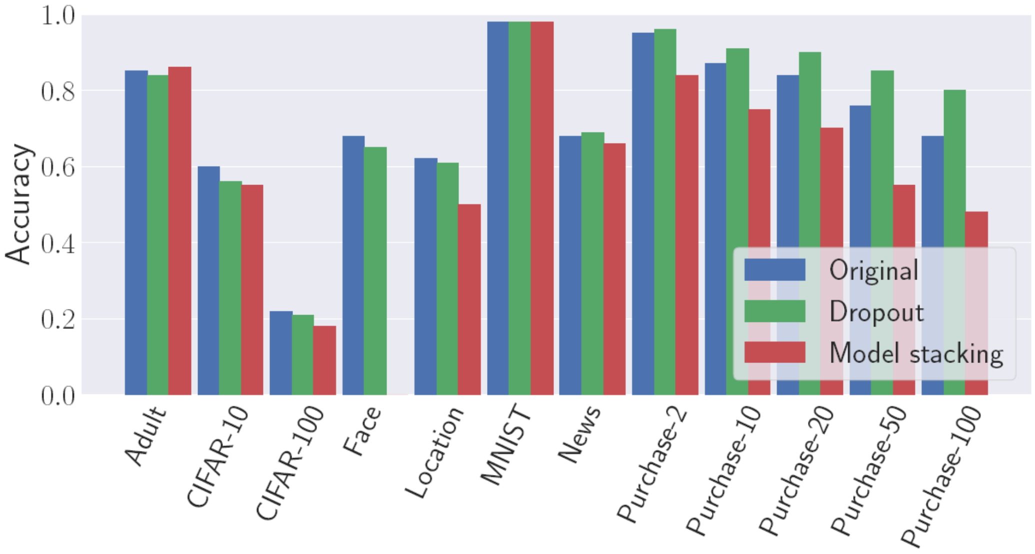


Fig. 15: Comparison of the target model's accuracy under both of the defense mechanisms.