Dataset Condensation for Data Privacy

Undergraduate Research Showcase ‘22

Eshan Gujarathi
eshan.rg@iitgn.ac.in

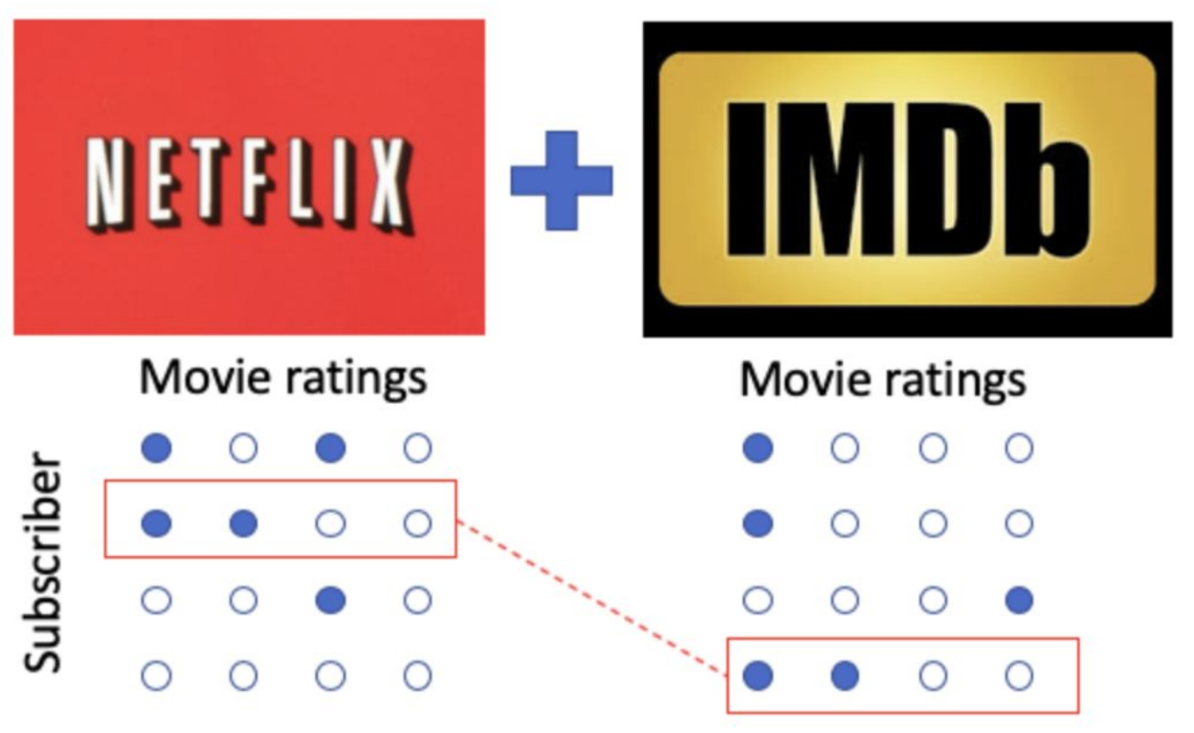
Hitarth Gandhi
hitarth.g@iitgn.ac.in

Advisor: Prof. Anirban Dasgupta

Motivation

With the advent of Machine Learning as a service, it has become increasingly simple for people to use a service provider like Google or Amazon to train their own ML models. Some models are then available as a black-box for open use. But even these black-box models can leak information about the data its trained on.

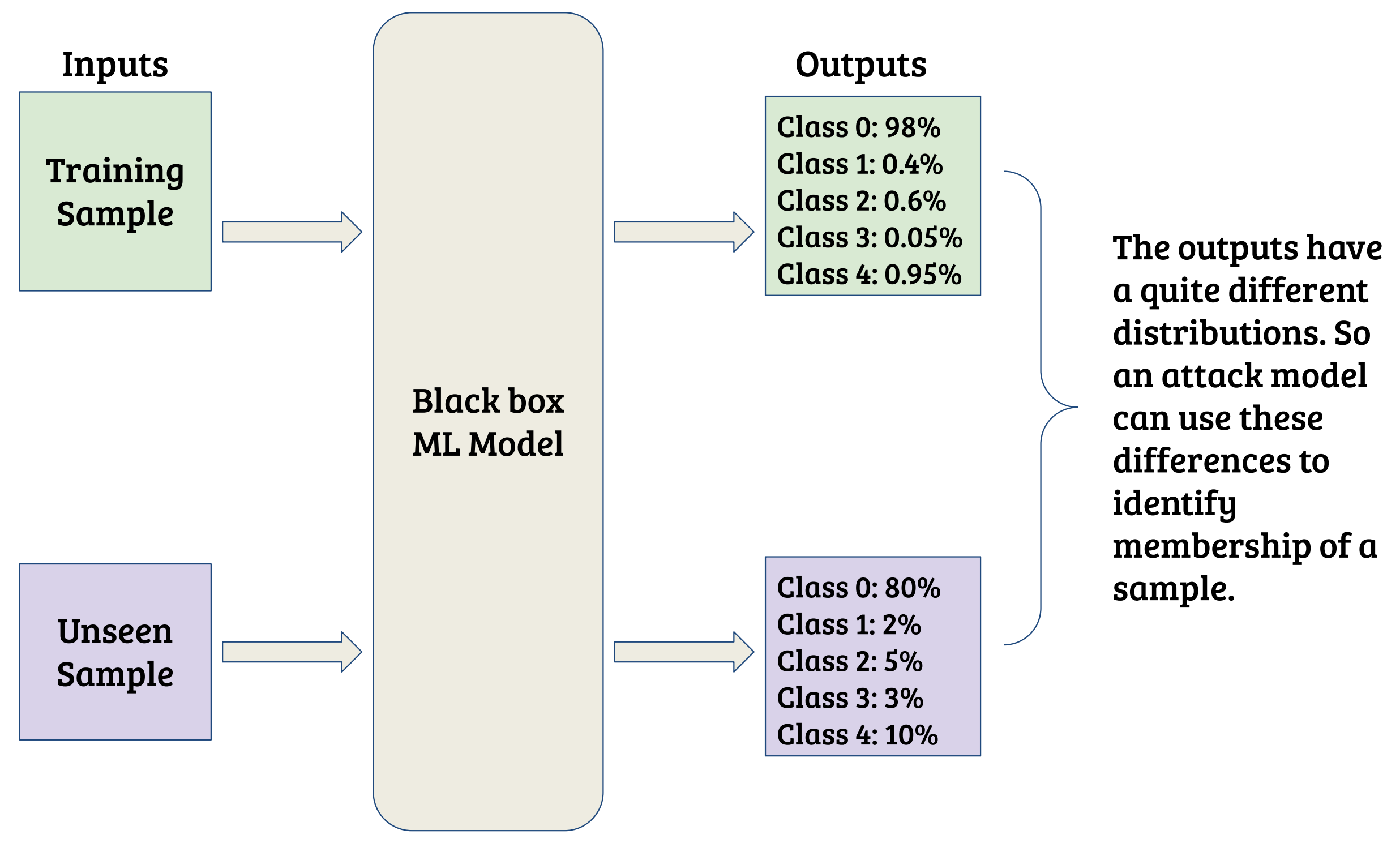
Many people can attack the model using a Membership Inference Attack and extract data from the model. Some methods try to mitigate this problem by anonymizing the data hence restricting the information gain of the attacker. But in past years researchers have figured out a way to de-anonymize the datasets using some other dataset. Thus we need more robust methods for data privacy.



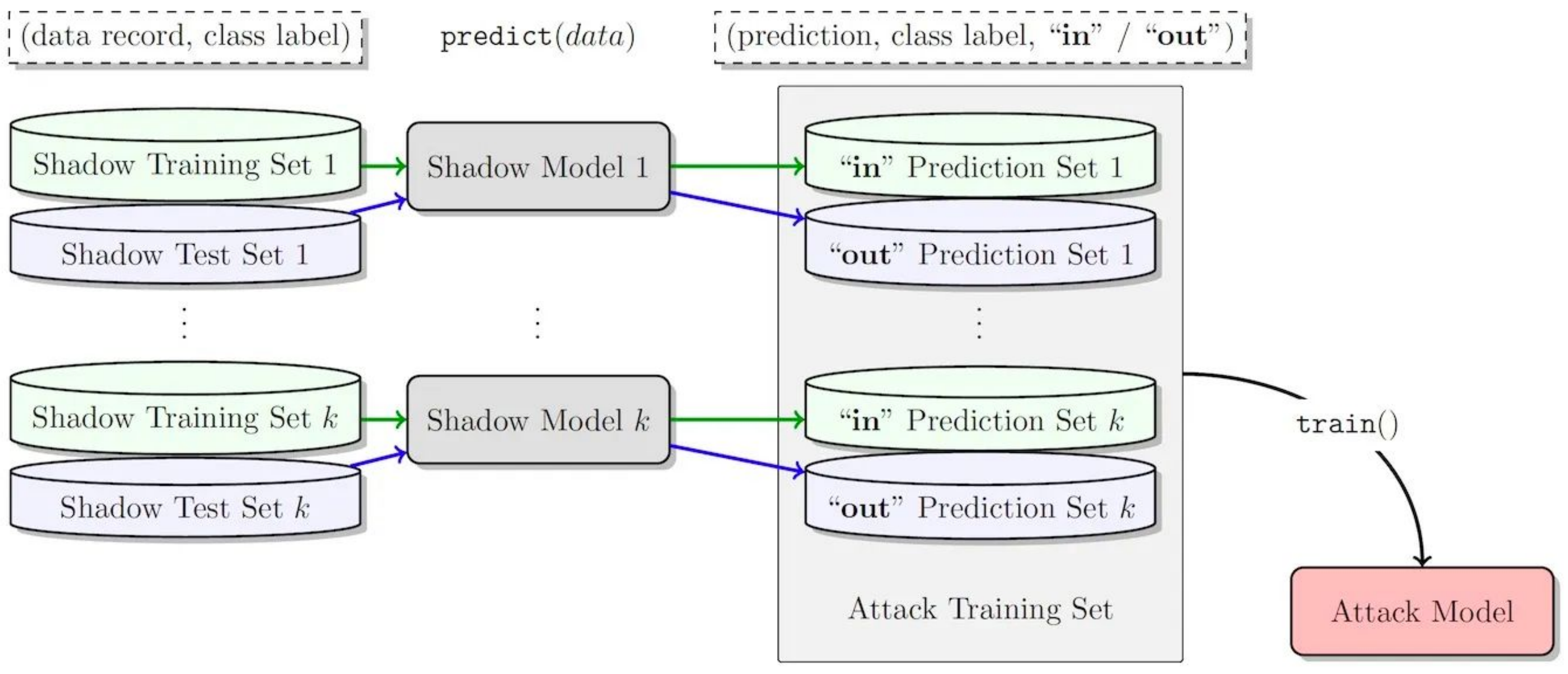
Introduction

Membership Inference Attack (MIA)

The attacker has access to a black box trained target model that outputs the prediction vector given an input image.

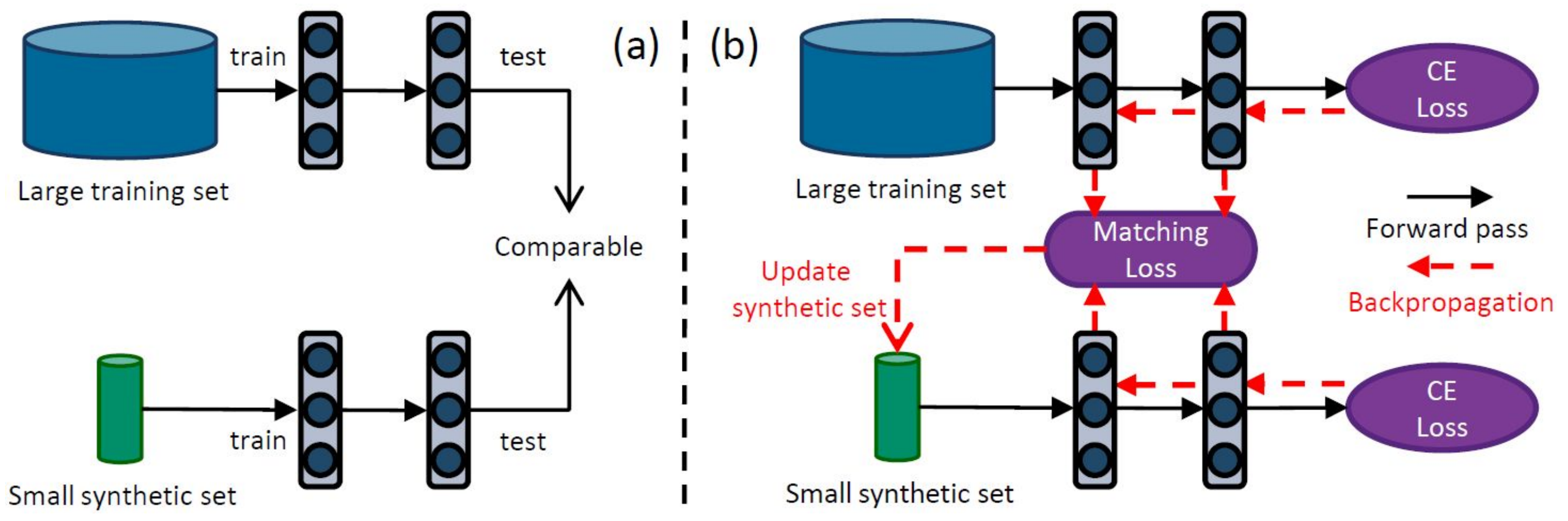


In this attack, we train *k* shadow models and get their prediction vectors for the training set and the testing set. Then we label the points in the training set as ***“in(y=1)”*** and the points in test set as ***“out(y=0)”***. Thus we now get a dataset which has inputs as the softmax outputs of shadow models and outputs as 0 or 1 depending its membership. So we use this dataset to train a simple attack classifier.



Dataset Condensation

Dataset Condensation aims to generate a small set of synthetic images that can match the performance of a network trained on a large image dataset. The method realizes this goal by learning a synthetic set such that a deep network trained on it and the large set produces similar gradients w.r.t. the parameters.

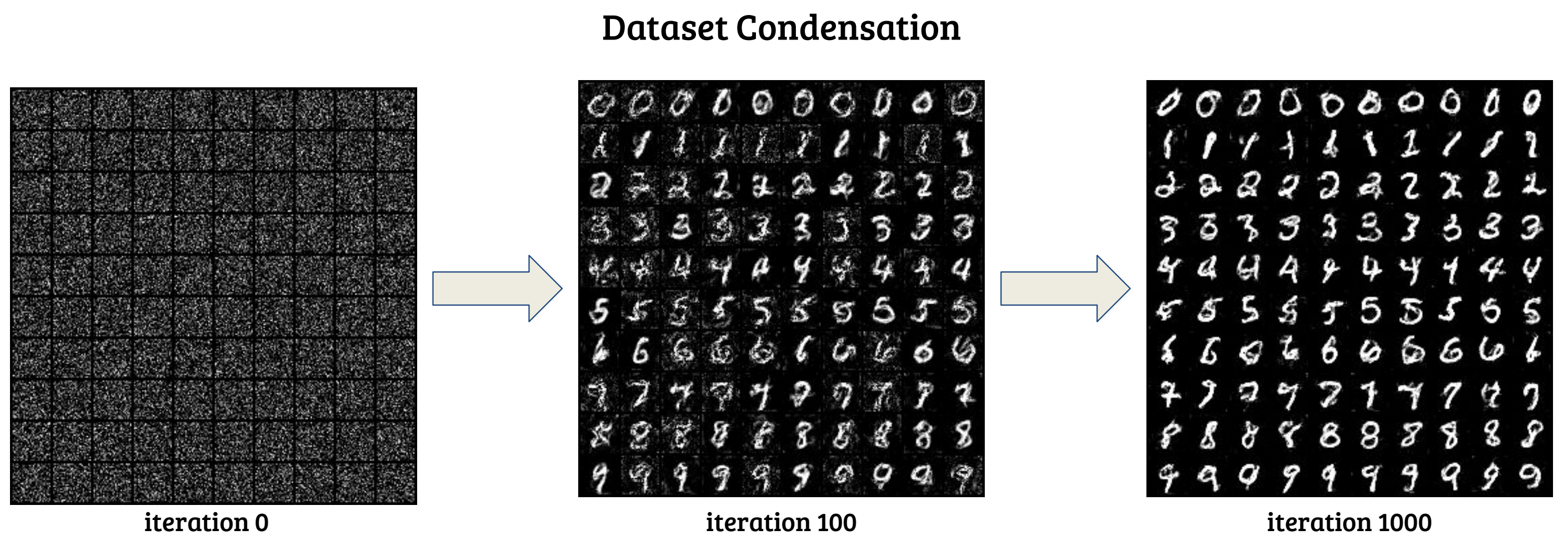


Why Dataset Condensation?

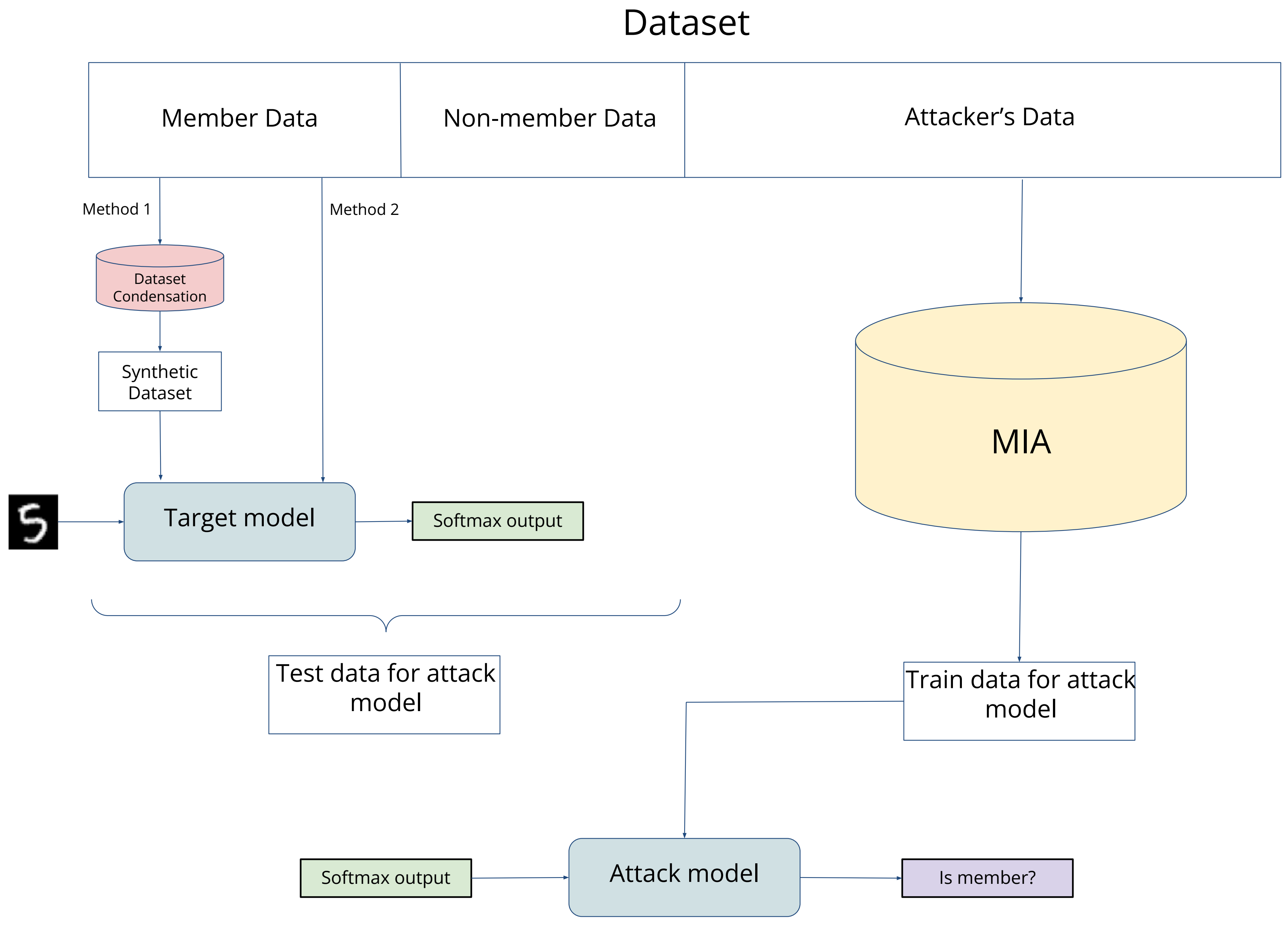
- Efficiency: Creates a smaller synthetic dataset from the the original dataset capturing features of the entire data. The synthetic data can later be used to train a network from scratch in a fraction of the original computational load with very less decrease in accuracy.
- Data Privacy for free: Creates new training data that is not part of the original dataset. It is difficult to backtrack from the synthetic dataset to the original dataset. Also, the attacker does not have any access to the synthetic dataset since it does not exist. Thus the model trains in the same way but provides privacy guarantees against attackers.

Experiments

We perform dataset condensation to make a synthetic dataset with only 10 images per class. We can see the dataset in making below.



Once we have the synthetic data, we can use it to train the target model. We propose the following pipeline for performing MIA experiments.



Results

Classification accuracy on simple CNN to show efficiency of data condensation.

Dataset	Accuracy
Original	~99%
Original (Subset)	~87%
Condensed	~90%

Preliminary results for attacker accuracy (for entire mnist data as member and attacker data)

Dataset	Accuracy
Original	~62%
Condensed	~41%

Future Work:

1. Implement the above pipeline and compare attack accuracy for original and synthetic dataset using dataset condensation.
2. Experiment with different dataset condensation techniques to give maximum privacy.

Acknowledgements

We are extremely grateful to Prof. Anirban Dasgupta for his guidance and support throughout the project. We would also like to thank the Undergraduate Research Showcase team for giving us the opportunity to present our work. Lastly, we are thankful to IIT Gandhinagar for providing us with the resources needed to complete the project.

The codebase and report for the project can be found here