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,
† Swiss Data Science Center, ETH Zurich and EPFL

Network and Distributed Systems Security (NDSS) Symposium 2019

By Lingxiao Kong

2019/7/30

Content

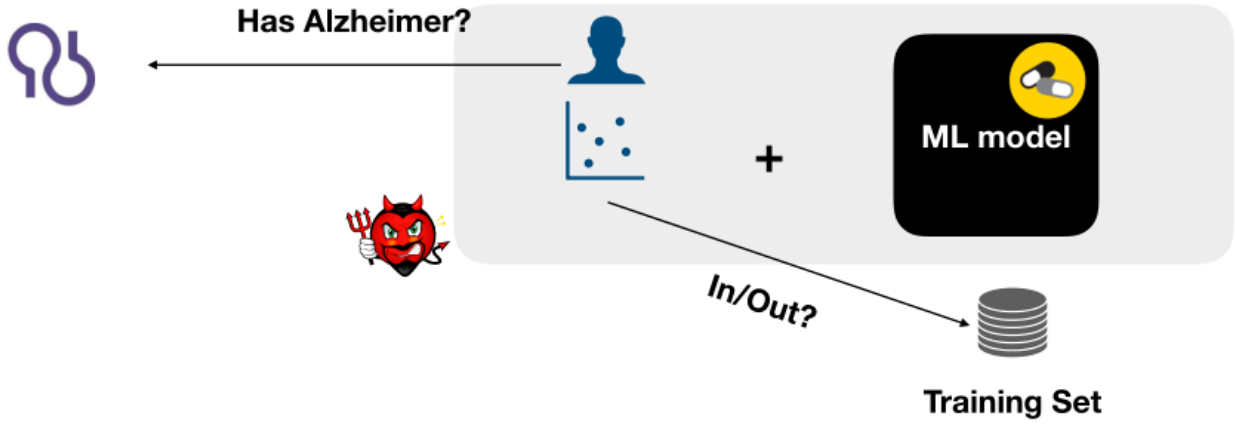
- Background
- Contribution
- Attacks
- Defence mechanisms

Background

- ML models are trained on sensitive data
  - Financial data
  - Location and activity data
  - Biomedical data
- ML models are vulnerable to various security and privacy attacks
  - Model Inversion Attacks
  - Adversarial sample attacks

Background

- Membership inference attack
  - An adversary aims to determine whether a data item (also referred to as a data point) was used to train an ML model or not.
  - Successful membership inference attacks can cause severe consequences



State Of The Art

- Shokri et al. present the first membership inference attack against machine learning models[1].
  - The target model is a black-box API
  - Construct multiple shadow models to mimic the target model's behavior and derive the data necessary
  - Train attack models
- Two main assumptions
  - Attacker needs to establish multiple shadow models with each one sharing the same structure as the target model.
  - Attacker needs to establish multiple shadow models with each one sharing the same structure as the target model.

Contribution

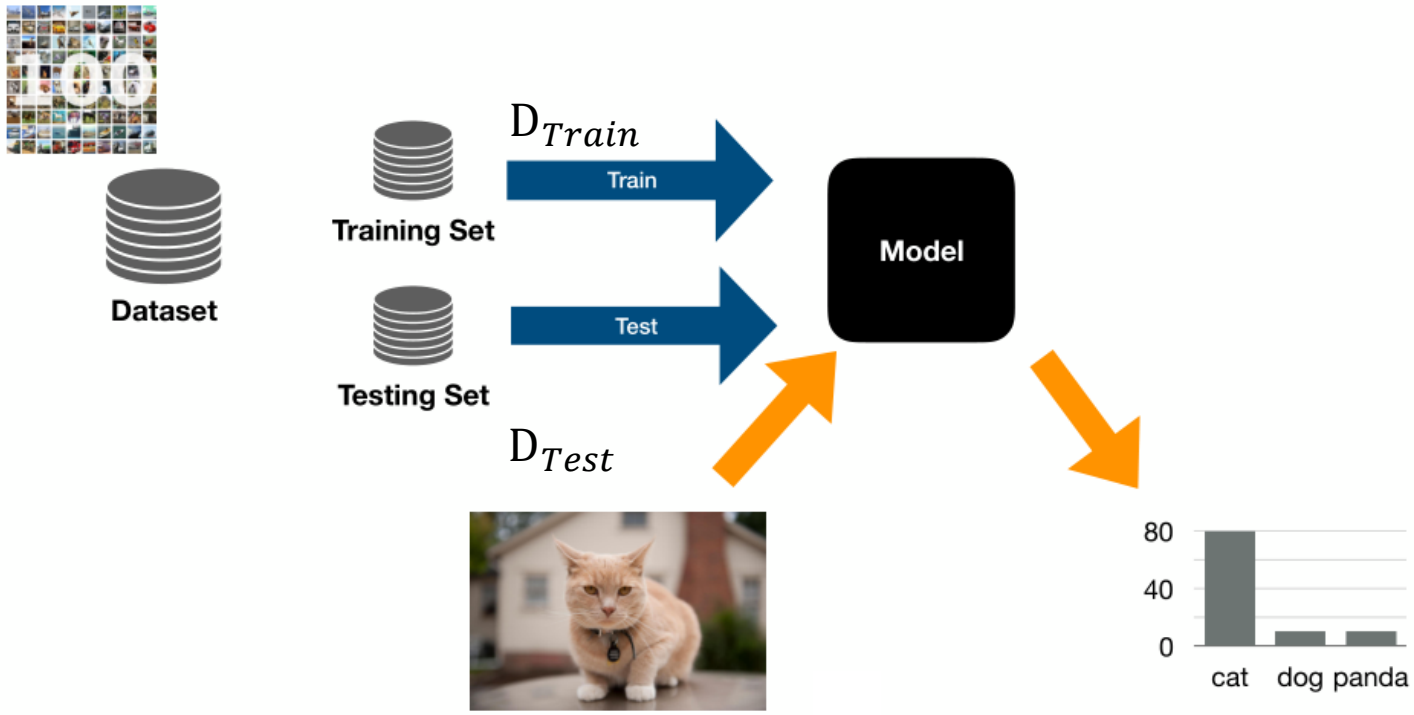
- This paper study three different types of adversaries based on the design and training data of shadow models.

Adversary type	Shadow model design		Target model's
	No. shadow models	Target model structure	training data distribution
Shokri et al. [38]	multiple	✓	✓
Our adversary 1	1	-	✓
Our adversary 2	1	-	-
Our adversary 3	-	-	-

- Propose two defense mechanisms, namely dropout and model stacking, and demonstrate their effectiveness experimentally.

Attacks

- An ML classifier is essentially a function *M* that maps a data point X (a multidimensional feature vector) to an output vector *M*(X) = Y.



Membership inference attack

- Given a target data point $\mathbf{X}_{Train}$, a trained machine learning model *M*, and external knowledge of an adversary, denoted by *K*, a membership inference attack *(attack model)* can be defined as the following function.

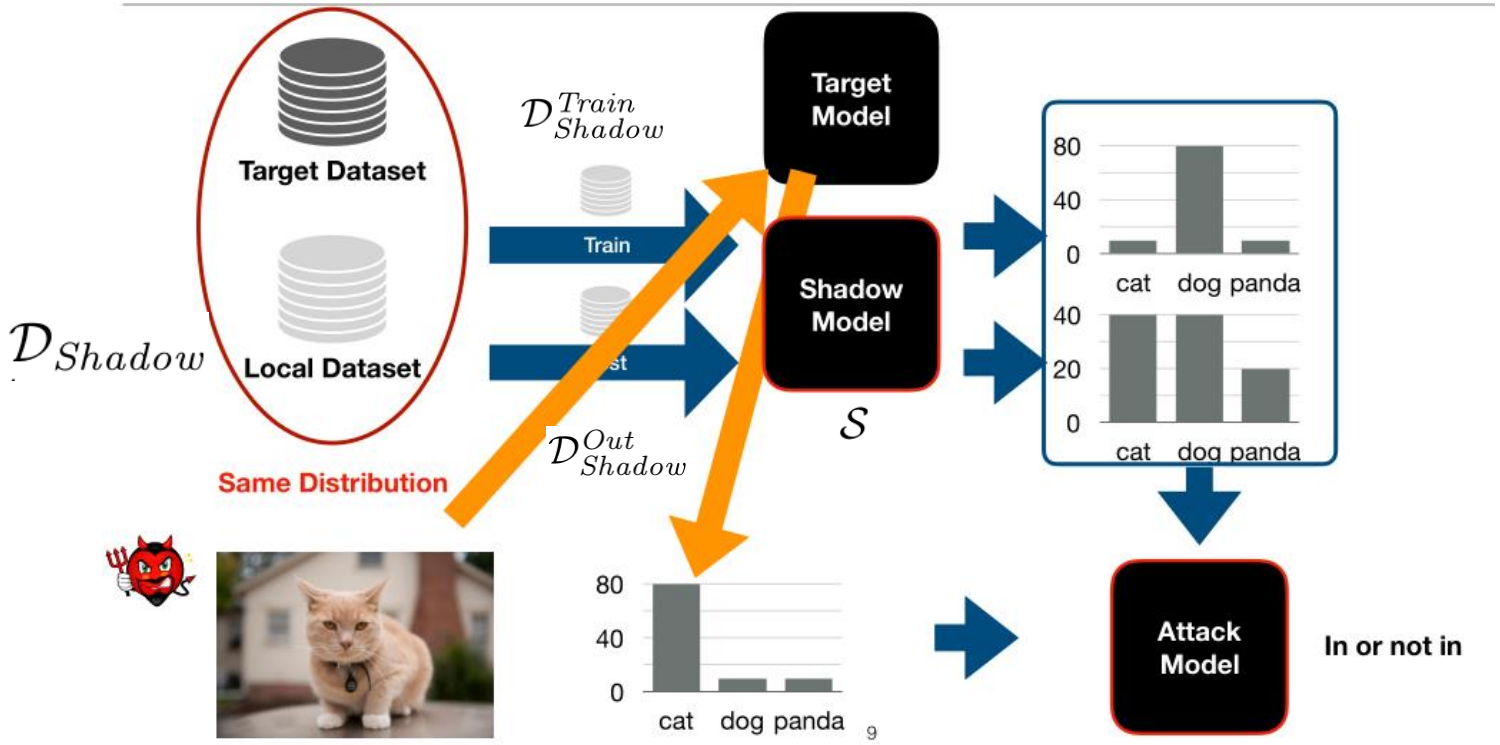
$$\mathcal{A} : \mathbf{x}_{Target}, \mathcal{M}, \mathcal{K} \rightarrow \{0, 1\}$$

  0 means $\mathbf{X}_{Train}$ is not a member of *M*'s training dataset $\mathbf{D}_{Train}$ and 1 otherwise.

- Assume the adversary only has black-box access to the target model, the adversary can submit a data point to *M* and then obtain the probabilistic output, i.e., *M($\mathbf{X}_{Train}$)*.

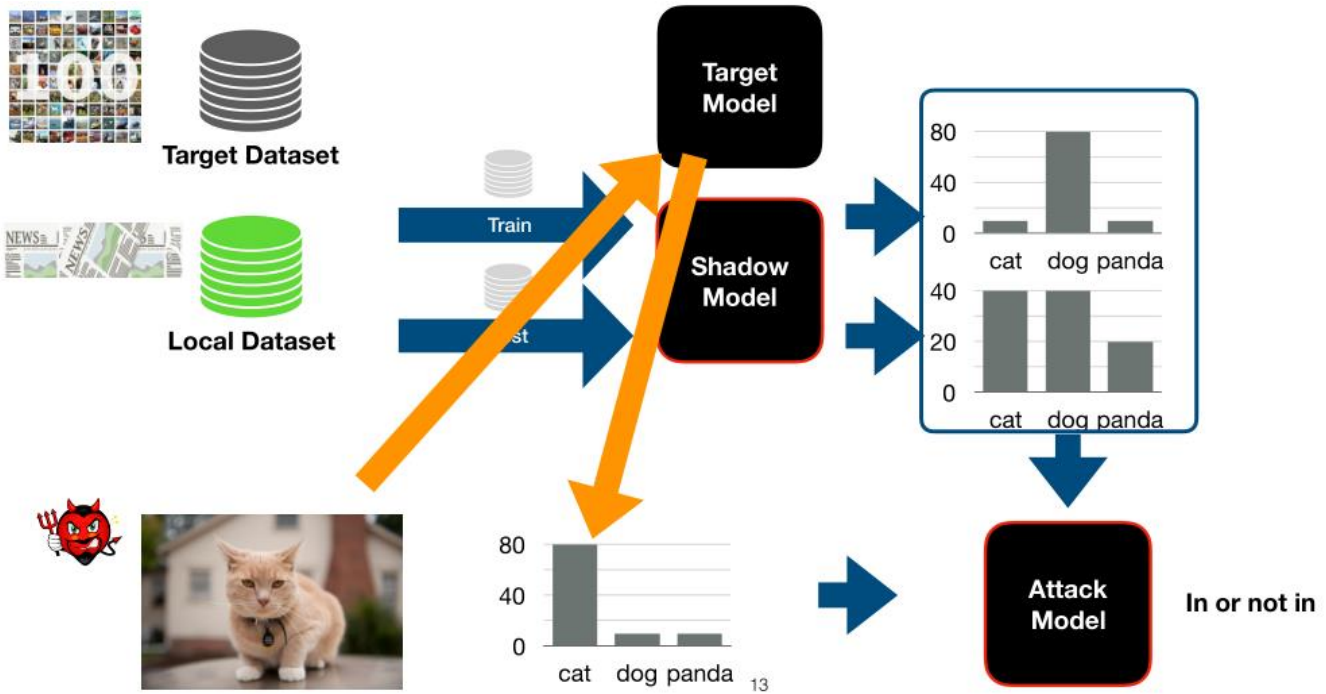
First Attack(Adversary 1)

- Assume a dataset that comes from the same distribution as the target model's training data.
- Start by using only one instead of multiple shadow models to mimic the target model's behavior.
- The adversary knows the target model's algorithm and hyperparameters and implements her shadow model in the same way.



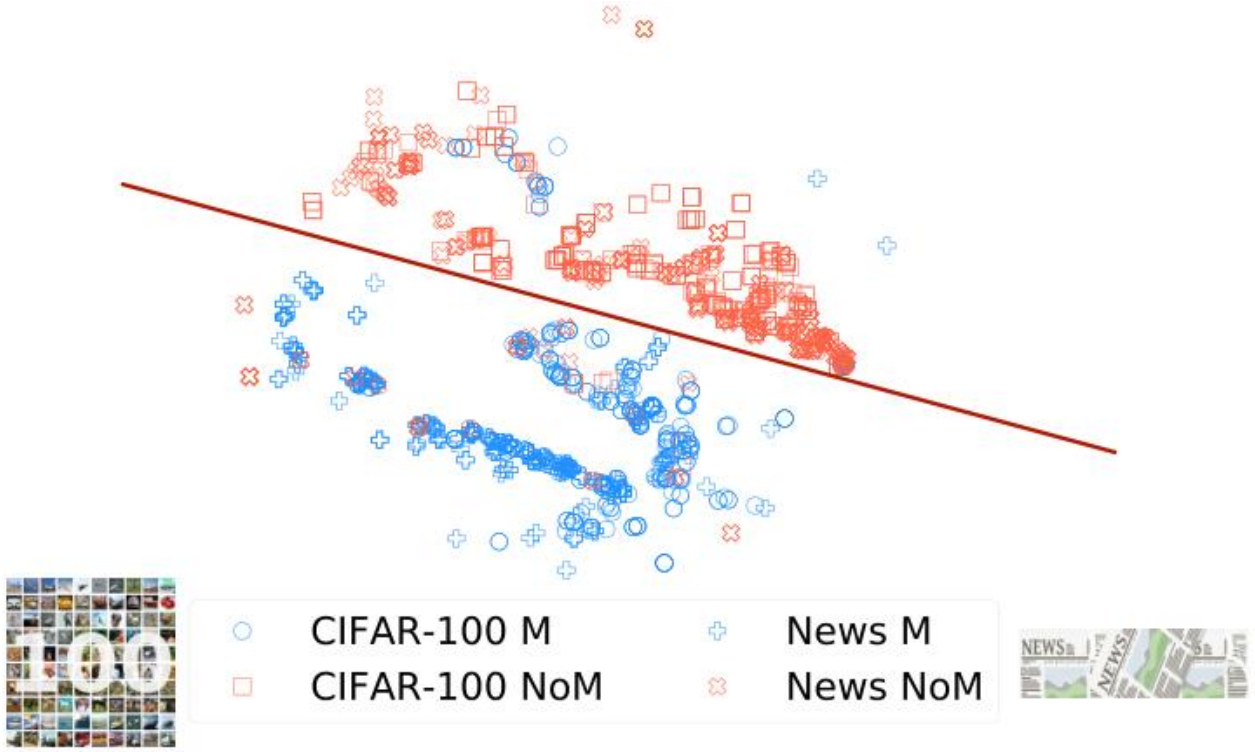
Data Transferring Attack (Adversary 2)

- Use an existing dataset that comes from a different distribution than the target model's training data to train her shadow model.
- The shadow model summarize the membership status of a data point in the training set of a machine learning model.



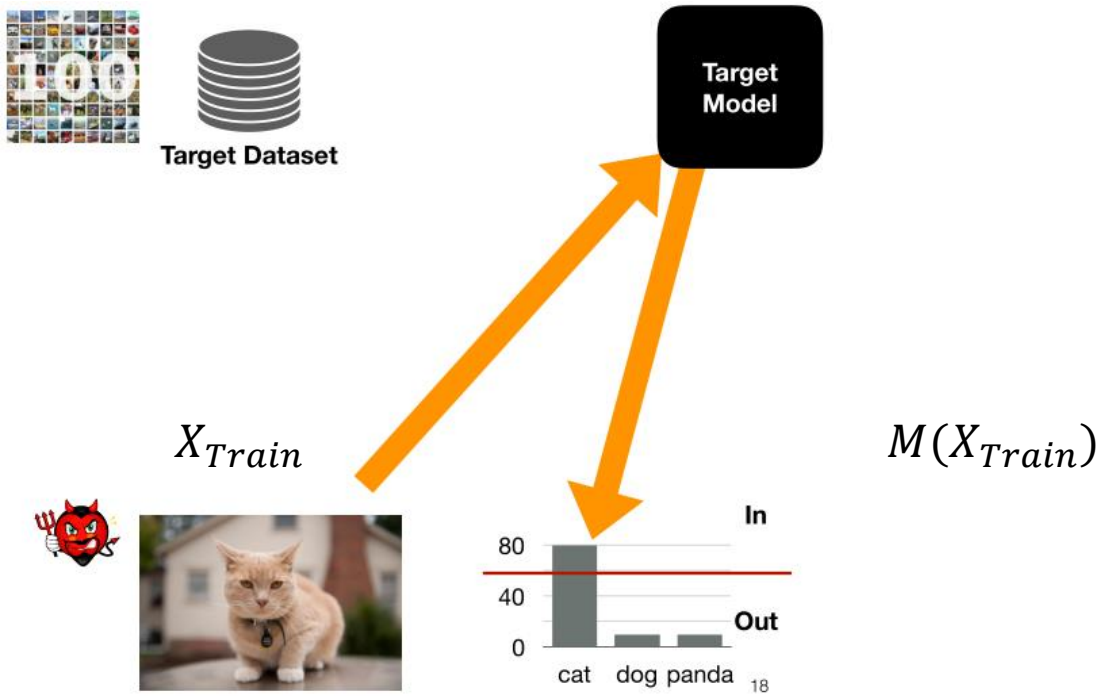
Reason

- The member and non-member points of these datasets are tightly clustered together and follow a common decision boundary.
- The attack model trained on one dataset can effectively infer the membership status of points in the other dataset.

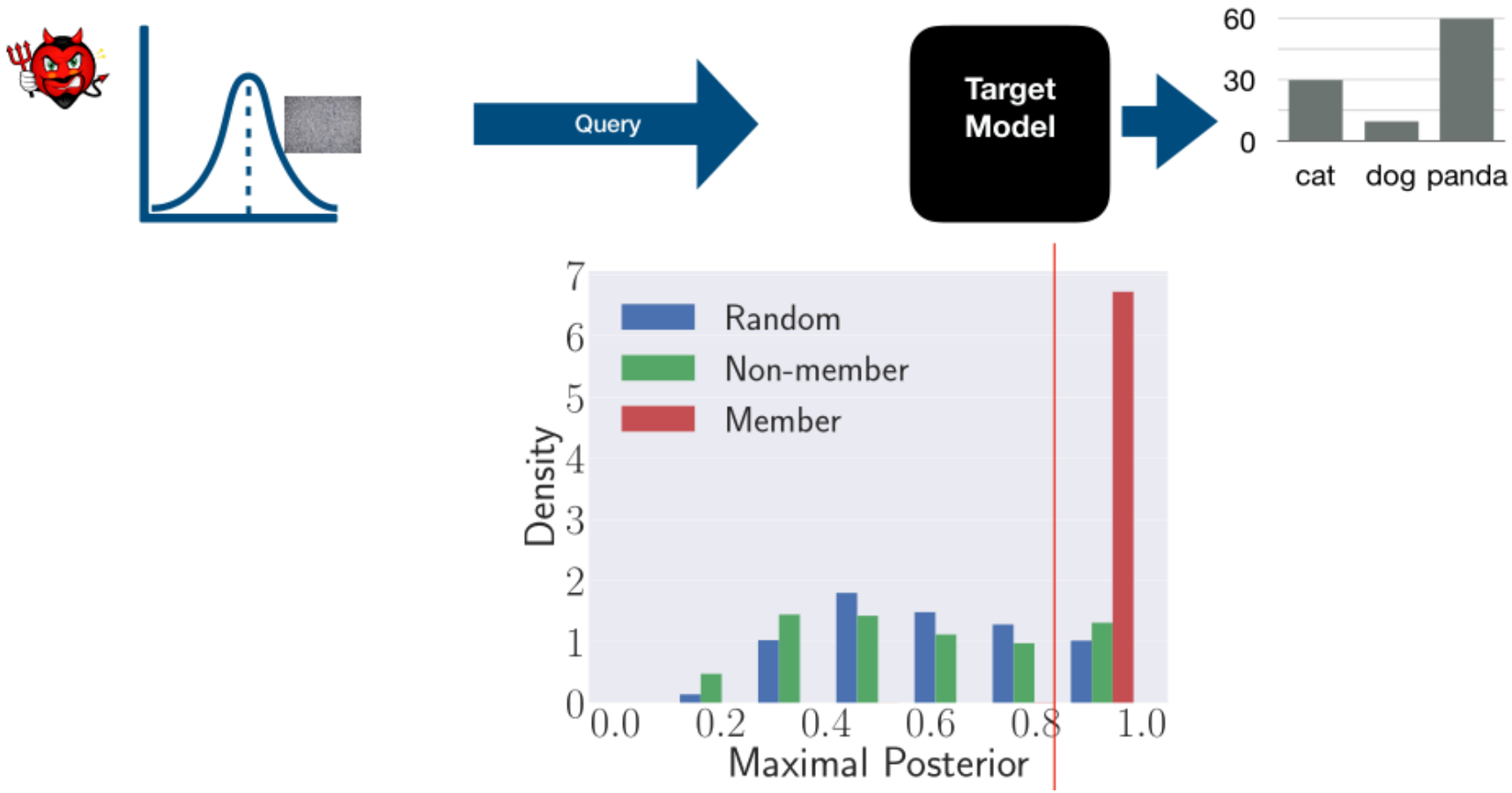


Third Attack (Adversary 3)

- All the third attack could rely on is the target model's output posteriors after querying her target data point.

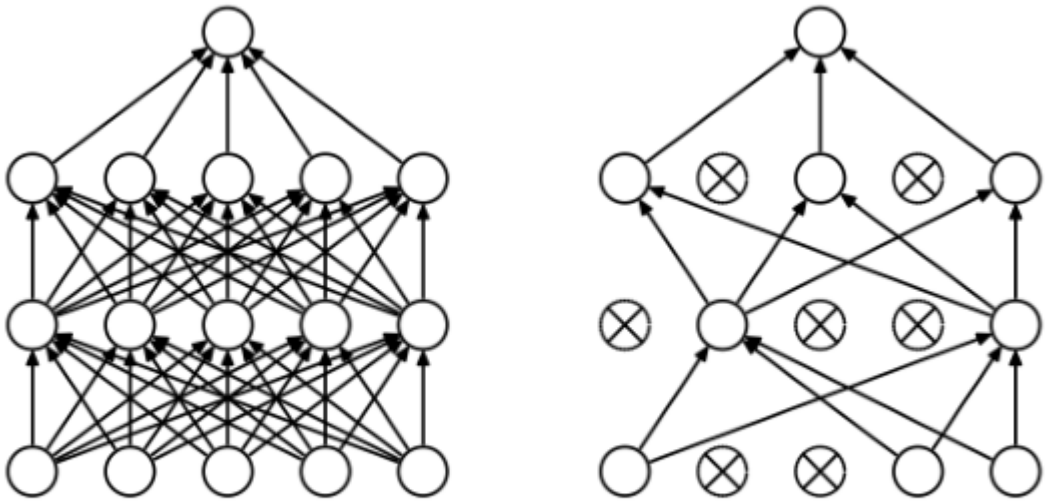


Threshold Choosing



Defense

- **Dropout** is a very effective method to reduce overfitting based on empirical evidences. It is executed by randomly deleting in each training iteration a fixed proportion (dropout ratio) of edges in a fully connected neural network model.

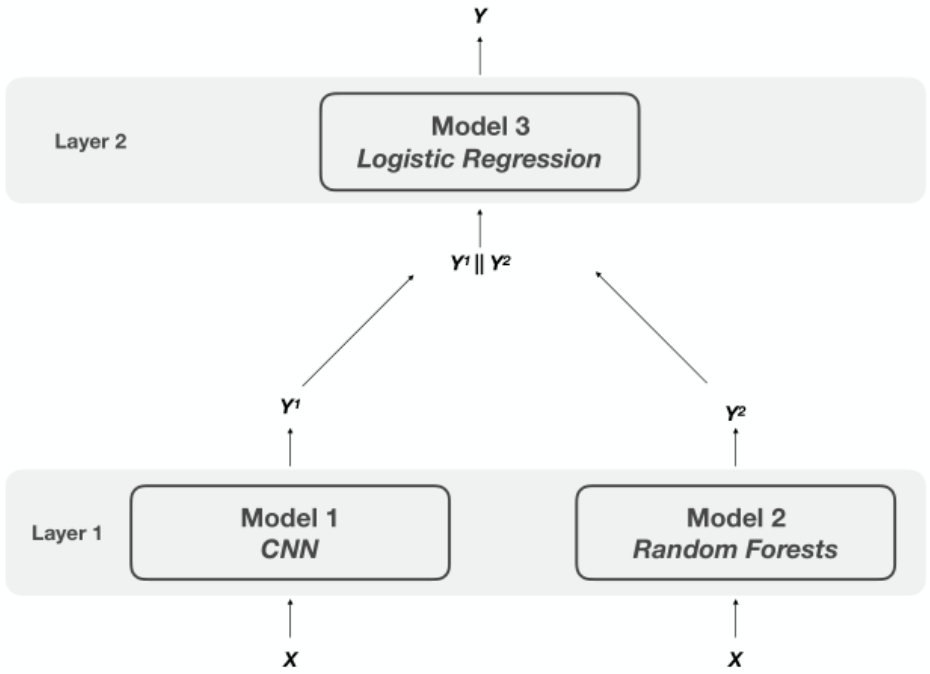


(a) Standard Neural Net (b) After applying dropout.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting", JMLR 2014

Model Stacking

- The intuition behind this defense is that if different parts of the target model are trained with different subsets of data, then the complete model should be less prone to overfitting.This can be achieved by using ensemble learning. neural network model.



THANKS