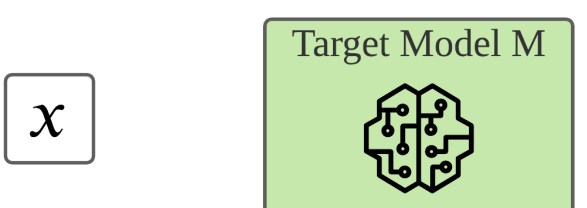
Membership Inference Attacks against Machine Learning Models

Lukas Gehrke

Membership Inference Attacks

Overview [Shokri et al. 2017]

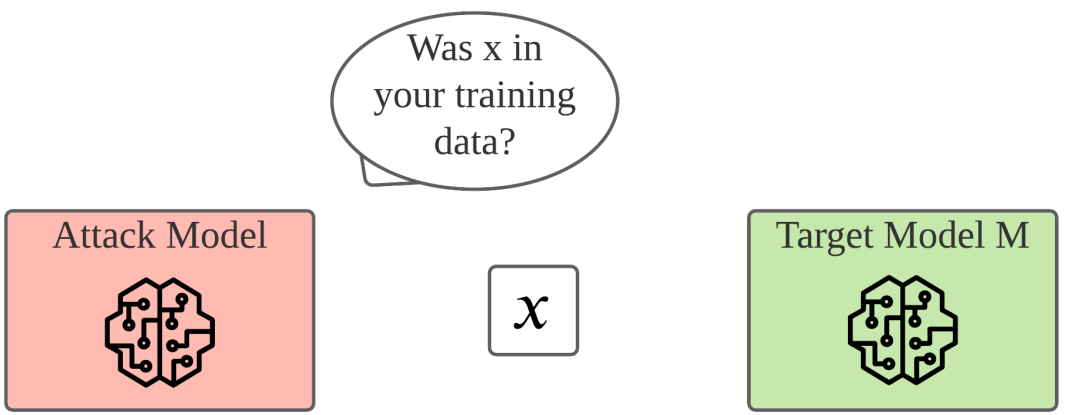
- Membership Inference Attack (MIA): Was data entry x in the training data of target model M?



Membership Inference Attacks

Overview [Shokri et al. 2017]

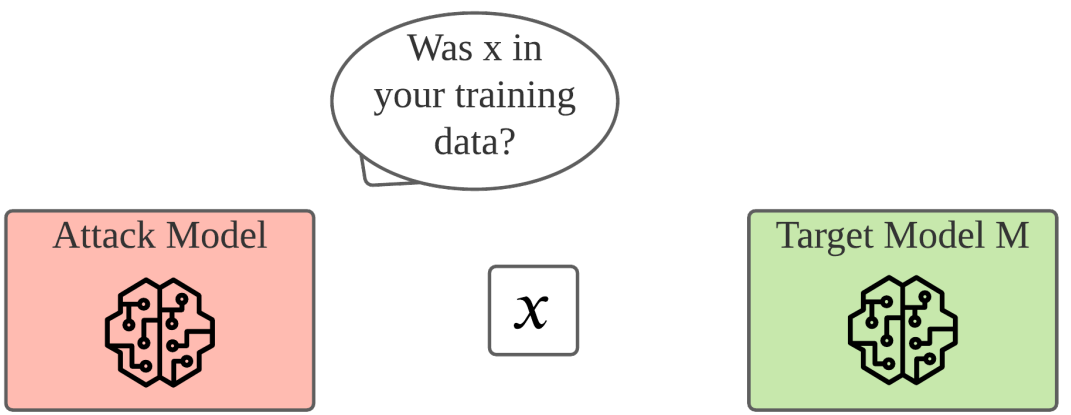
- Membership Inference Attack (MIA): Was data entry d in the training data of target model M?
- *"Turn Machine Learning against itself"* - Train machine learning models to perform Membership Inference Attacks against other models.



Membership Inference Attacks

Overview [Shokri et al. 2017]

- Membership Inference Attack (MIA): Was data entry d in the training data of target model M?
- *"Turn Machine Learning against itself"* - Train machine learning models to perform Membership Inference Attacks against other models.
- Focus on *supervised learning*: Benefit from models learning the relationship between data and labels as well as their tendency to overfit.



Membership Inference Attacks

"Quantify membership information leakage through the prediction outputs of machine learning models."

[Shokri et al. 2017]

Membership Inference Attacks

Background

- Wide-spread usage of Machine Learning as a Service (MLaaS) due to possibilities and advantages (easy setup, cheap infrastructure)[1]

[1] https://learn.g2.com/trends/machine-learning-service-mlaas; retrieved 2021-01-06

Membership Inference Attacks

Background

- Wide-spread usage of Machine Learning as a Service (MLaaS) due to possibilities and advantages (easy setup, cheap infrastructure)[1]
- Use cases for MLaaS:
  - NLP: Sentiment or topic analysis in emails, social media or reviews[2]
  - Image and Video Processing: Traffic analysis, shelve analysis in retail[2]

[1] https://learn.g2.com/trends/machine-learning-service-mlaas; retrieved 2021-01-06
[2] https://monkeylearn.com/blog/mlaas/; retrieved 2021-01-06

Membership Inference Attacks

Background

- Wide-spread usage of Machine Learning as a Service (MLaaS) due to possibilities and advantages (easy setup, cheap infrastructure)[1]
- Use cases for MLaaS:
  - NLP: Sentiment or topic analysis in emails, social media or reviews[2]
  - Image and Video Processing: Traffic analysis, shelve analysis in retail[2]
  - Forecasting and Prediction - Example [Shokri et al. 2017]: For an app, *analyze users behavior to train a model that learns when users are most likely to buy. Use the model inside the app to increase sales of in-app purchases*.

[1] https://learn.g2.com/trends/machine-learning-service-mlaas; retrieved 2021-01-06
[2] https://monkeylearn.com/blog/mlaas/; retrieved 2021-01-06

Membership Inference Attacks

Threats

- Entries of *privacy-sensitive* datasets could be exposed.
- Example [Truex et al. 2019]:
  - A MLaaS model used to determine ideal medical treatment for patients with a certain disease
  - The fact that a person A's clinical record was in the training data exposes that A has the disease.

Membership Inference Attacks

Threats

- Entries of *privacy-sensitive* datasets could be exposed.
- Example [Truex et al. 2019]:
  - A MLaaS model used to determine ideal medical treatment for patients with a certain disease
  - The fact that a person A's clinical record was in the training data exposes that A has the disease.
  - Consequences: A may be blackmailed with exposure, the owner of the model may be confronted with legal action; both may face economical disadvantages

Membership Inference Attacks

Attack Definition, Notations as in [Shokri et al. 2017]

Given

Membership Inference Attacks

Attack Definition, Notations as in [Shokri et al. 2017]

Given

- black-box query access to a model $f_{target}()$, called **target** model,
- D^{train}_{target}, the yet unknown set of training data of $f_{target}()$
- and a data record x,

determine if $x \in D^{train}_{target}$.

Membership Inference Attacks

Attack Definition, Notations as in [Shokri et al. 2017]

Given

- black-box query access to a model $f_{target}()$, called **target** model,
- D^{train}_{target}, the yet unknown set of training data of $f_{target}()$
- and a data record x,

determine if $x \in D^{train}_{target}$.

Attack Realization

- Train an **attack** model $f_{attack}()$ to learn differences in predictions of $f_{target}()$.

Membership Inference Attacks

Attack Definition, Notations as in [Shokri et al. 2017]

Given

- black-box query access to a model $f_{target}()$, called **target** model,
- D^{train}_{target}, the yet unknown set of training data of $f_{target}()$
- and a data record x,

determine if $x \in D^{train}_{target}$.

Attack Realization

- Train an **attack** model $f_{attack}()$ to learn differences in predictions of $f_{target}()$.
  - Assumption: $f_{target}()$ behaves differently on entries $\hat{x} \in D^{train}_{target}$ and $\tilde{x} \notin D^{train}_{target}$.
  - Let outputs of $f_{target}()$ be probability vectors of size c_{target} - the number of classes $f_{target}()$ assigns to data records - where each entry is in $[0, 1]$ and represents the probability that a record x is in class $y^{\{i\}}$ ($i \in 1 \ldots c$).

Membership Inference Attacks

Attack Definition, Notations as in [Shokri et al. 2017]

Given

- black-box query access to a model $f_{target}()$, called **target** model,
- D^{train}_{target}, the yet unknown set of training data of $f_{target}()$
- and a data record x,

determine if $x \in D^{train}_{target}$.

Attack Realization

- Train an **attack** model $f_{attack}()$ to learn differences in predictions of $f_{target}()$.
  - Assumption: $f_{target}()$ behaves differently on entries $\hat{x} \in D^{train}_{target}$ and $\tilde{x} \notin D^{train}_{target}$.
  - Let outputs of $f_{target}()$ be probability vectors of size c_{target} - the number of classes $f_{target}()$ assigns to data records - where each entry is in $[0, 1]$ and represents the probability that a record x is in class $y^{\{i\}}$ ($i \in 1 \dots c$).

  $\rightarrow$ For records $\hat{x} \in D^{train}_{target}$ the prediction outputs of $f_{target}()$ should reveal different patterns in distribution of probabilities than for records $\tilde{x} \notin D^{train}_{target}$.

Membership Inference Attacks

Attack Realization (continued)

- Create n **shadow** models $f_{shadow^1}() \ldots f_{shadow^n}()$ that imitate the behavior of $f_{target}()$. (e.g. by using the *same service* that has been used for $f_{target}()$)

Membership Inference Attacks

Attack Realization (continued)

- Create n **shadow** models $f_{shadow^1}() \ldots f_{shadow^n}()$ that imitate the behavior of $f_{target}()$. (e.g. by using the *same service* that has been used for $f_{target}()$)
  - Their training datasets $D^{train}_{shadow^1} \ldots D^{train}_{shadow^n}$ and testing datasets $D^{test}_{shadow^1} \ldots D^{test}_{shadow^n}$ are both known and have the same format as D^{train}_{target}.
  - Also the predictions of the shadow models have the same format as the predictions of $f_{target}()$.

Membership Inference Attacks

Attack Realization (continued)

- Create n **shadow** models $f_{shadow^1}() \ldots f_{shadow^n}()$ that imitate the behavior of $f_{target}()$. (e.g. by using the *same service* that has been used for $f_{target}()$)
  - Their training datasets $D^{train}_{shadow^1} \ldots D^{train}_{shadow^n}$ and testing datasets $D^{test}_{shadow^1} \ldots D^{test}_{shadow^n}$ are both known and have the same format as D^{train}_{target}.
  - Also the predictions of the shadow models have the same format as the predictions of $f_{target}()$.

  $\rightarrow$ The shadow models inputs and predictions serve as training data for $f_{attack}()$ to learn $f_{target}()$'s different reactions to known and unknown data records.

Membership Inference Attacks

Attack Realization (continued)

- Create n **shadow** models $f_{shadow^1}() \dots f_{shadow^n}()$ that imitate the behavior of $f_{target}()$. (e.g. by using the *same service* that has been used for $f_{target}()$)
  - Their training datasets $D^{train}_{shadow^1} \dots D^{train}_{shadow^n}$ and testing datasets $D^{test}_{shadow^1} \dots D^{test}_{shadow^n}$ are both known and have the same format as D^{train}_{target}.
  - Also the predictions of the shadow models have the same format as the predictions of $f_{target}()$.

  $\rightarrow$ The shadow models inputs and predictions serve as training data for $f_{attack}()$ to learn $f_{target}()$'s different reactions to known and unknown data records.
- Perform the attack by querying data records to $f_{target}()$ and classifying the outputs as $\in$ or $\notin D^{train}_{target}$ with $f_{attack}()$.

Membership Inference Attacks

Attack Realization (continued)

- Create n **shadow** models $f_{shadow^1}() \dots f_{shadow^n}()$ that imitate the behavior of $f_{target}()$. (e.g. by using the *same service* that has been used for $f_{target}()$)
  - Their training datasets $D^{train}_{shadow^1} \dots D^{train}_{shadow^n}$ and testing datasets $D^{test}_{shadow^1} \dots D^{test}_{shadow^n}$ are both known and have the same format as D^{train}_{target}.
  - Also the predictions of the shadow models have the same format as the predictions of $f_{target}()$.

  $\rightarrow$ The shadow models inputs and predictions serve as training data for $f_{attack}()$ to learn $f_{target}()$'s different reactions to known and unknown data records.

- Perform the attack by querying data records to $f_{target}()$ and classifying the outputs as $\in$ or $\notin D^{train}_{target}$ with $f_{attack}()$.

https://moviedatabase.fandom.com/wiki/Mia_Wallace

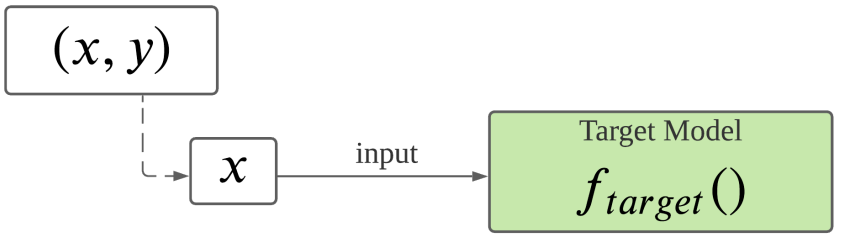
Membership Inference Attacks

Visualization

(x, y)

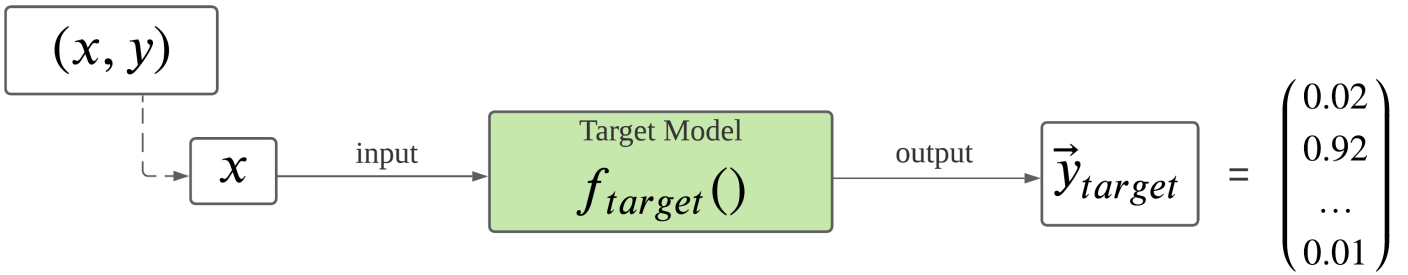
Membership Inference Attacks

Visualization



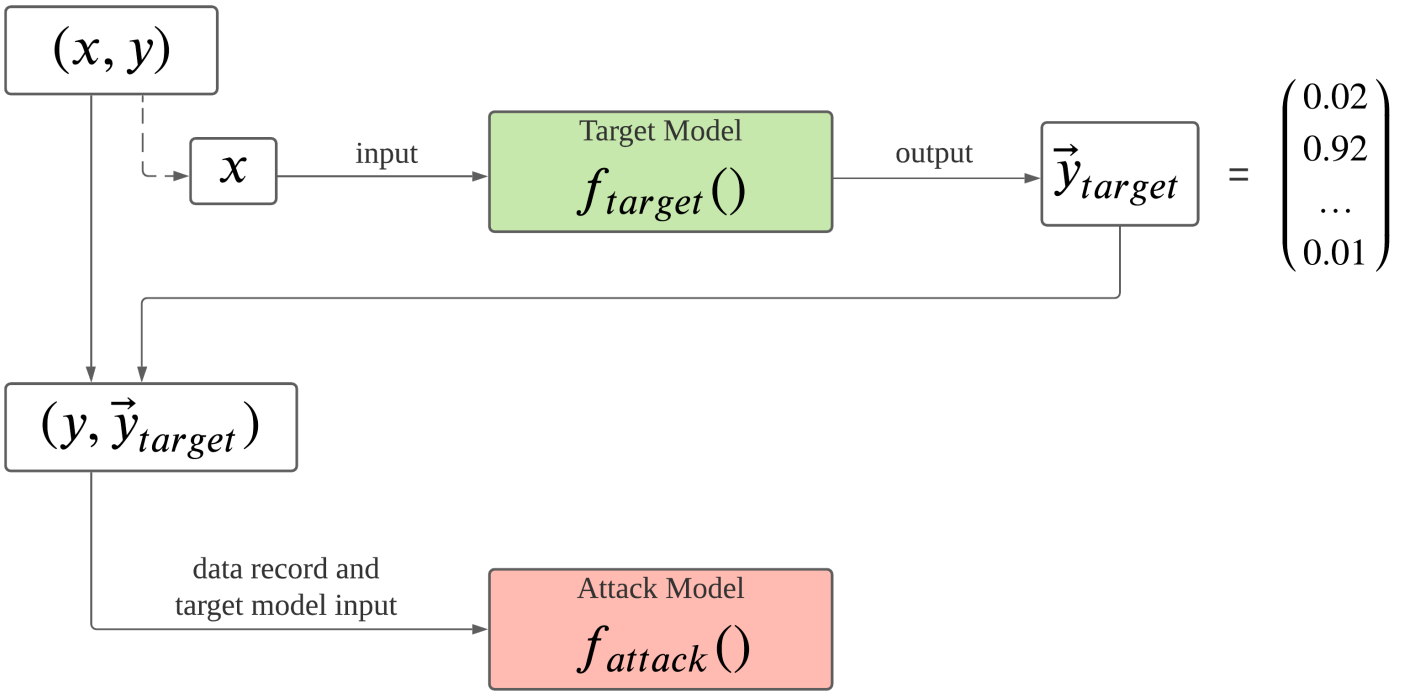
Membership Inference Attacks

Visualization



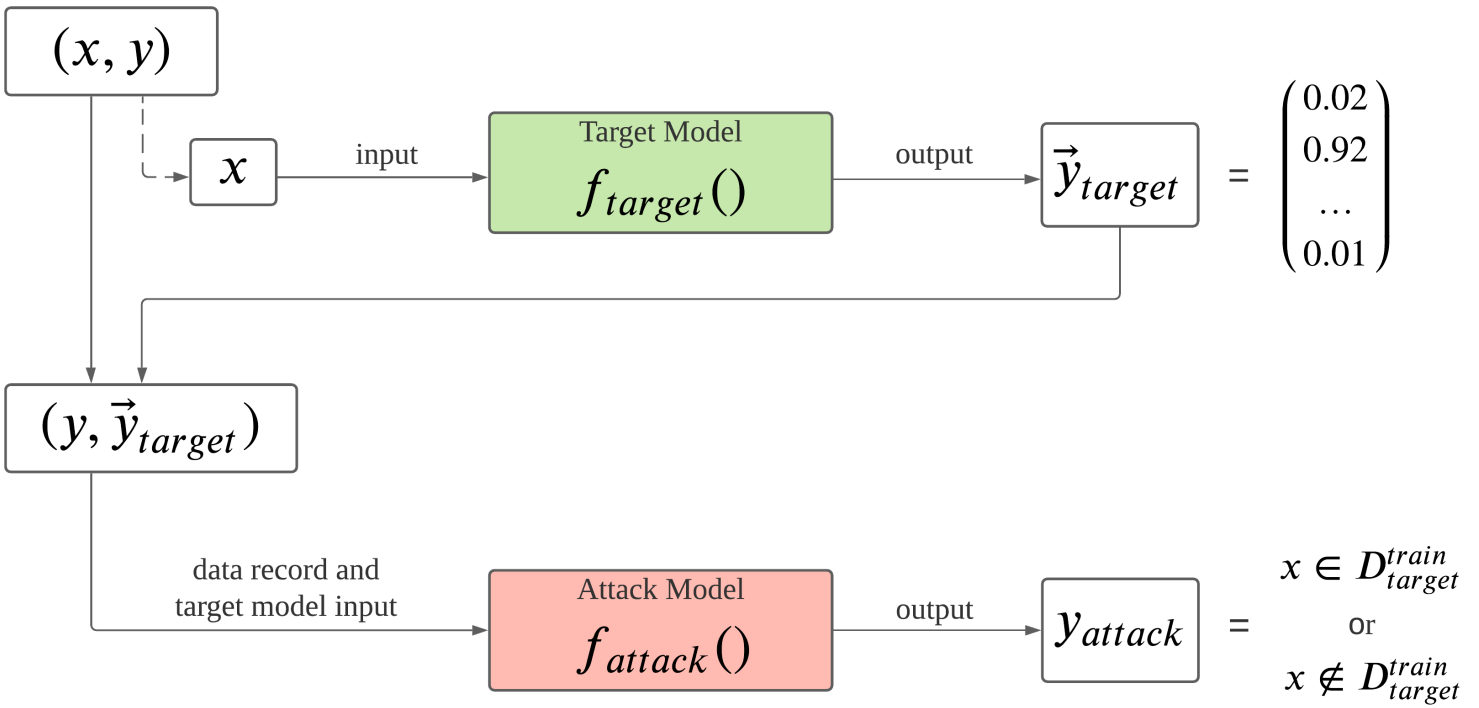
Membership Inference Attacks

Visualization



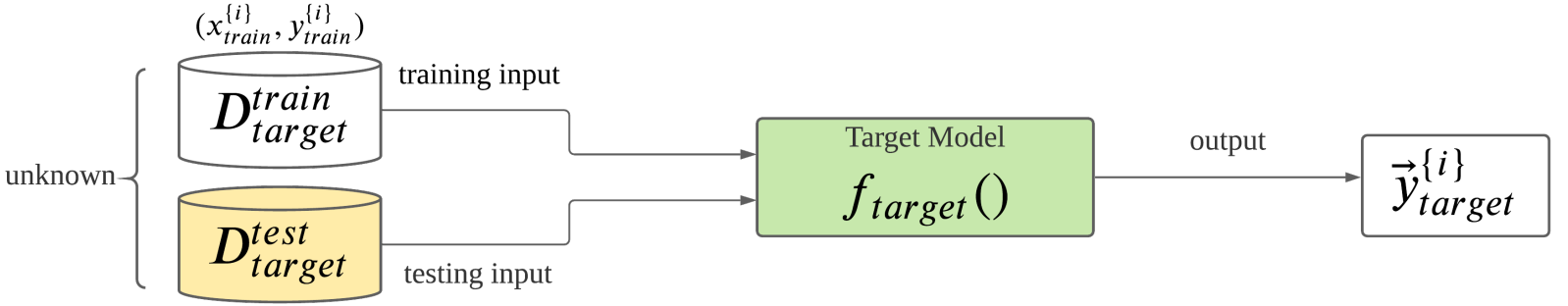
Membership Inference Attacks

Visualization



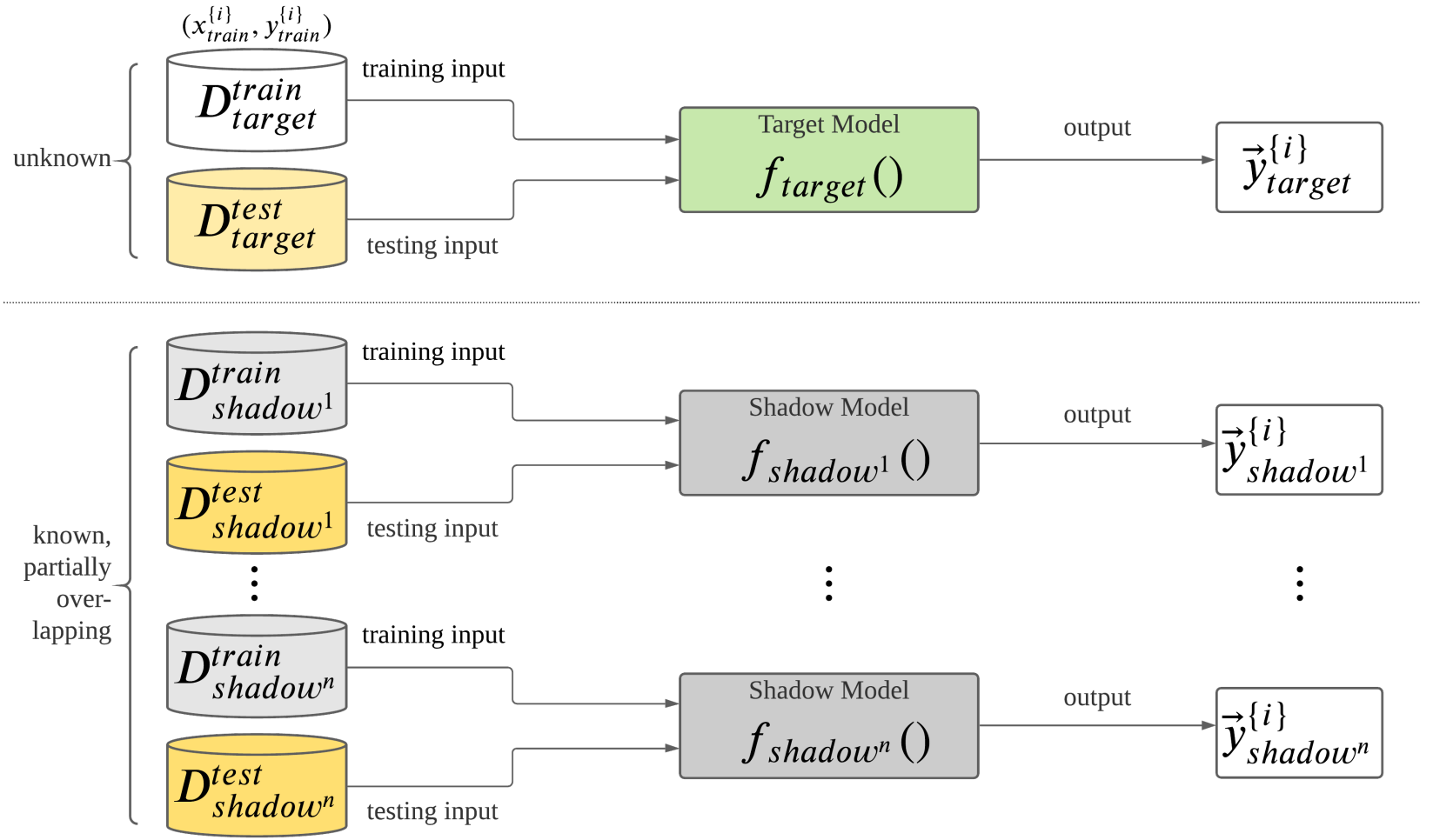
Membership Inference Attacks

Visualization



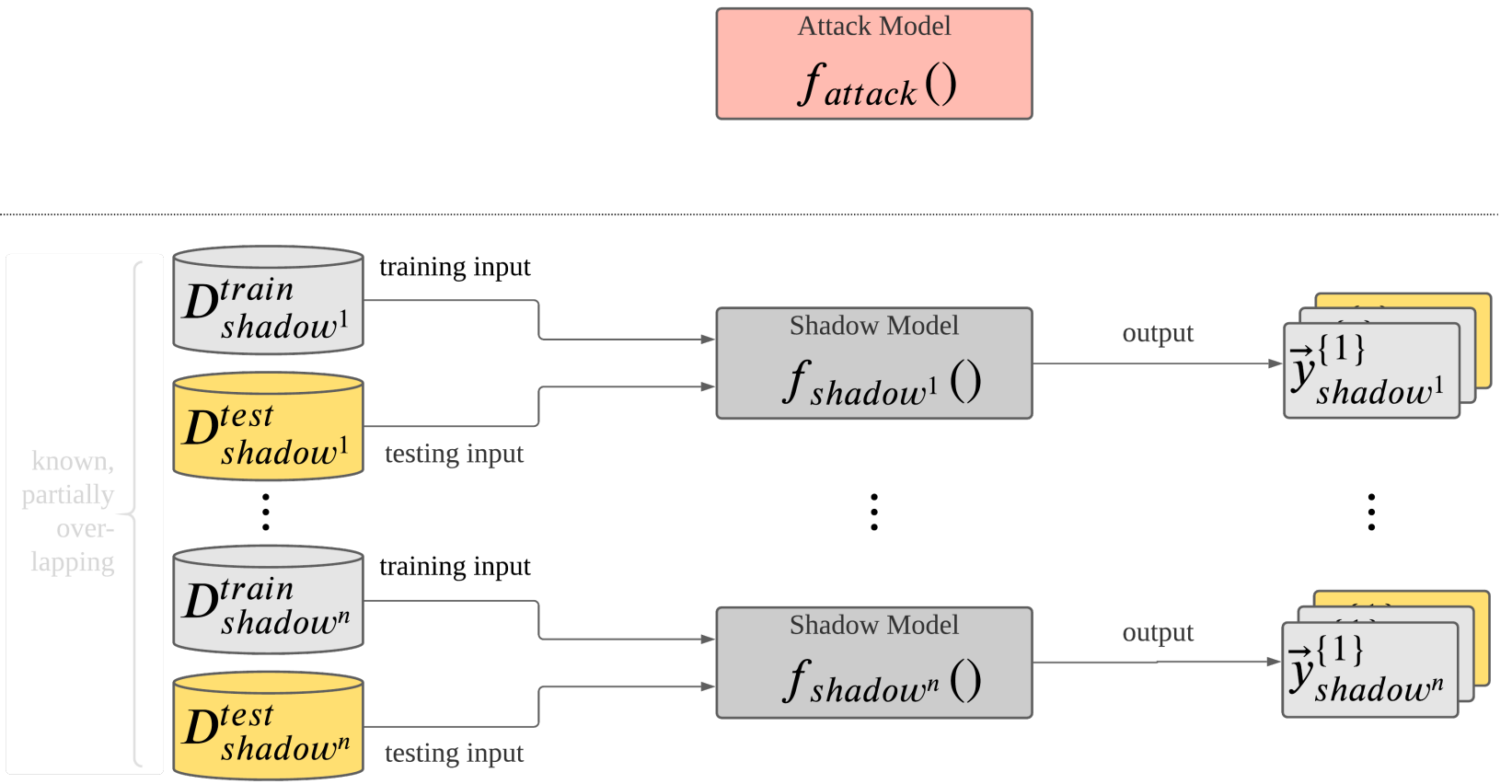
Membership Inference Attacks

Visualization



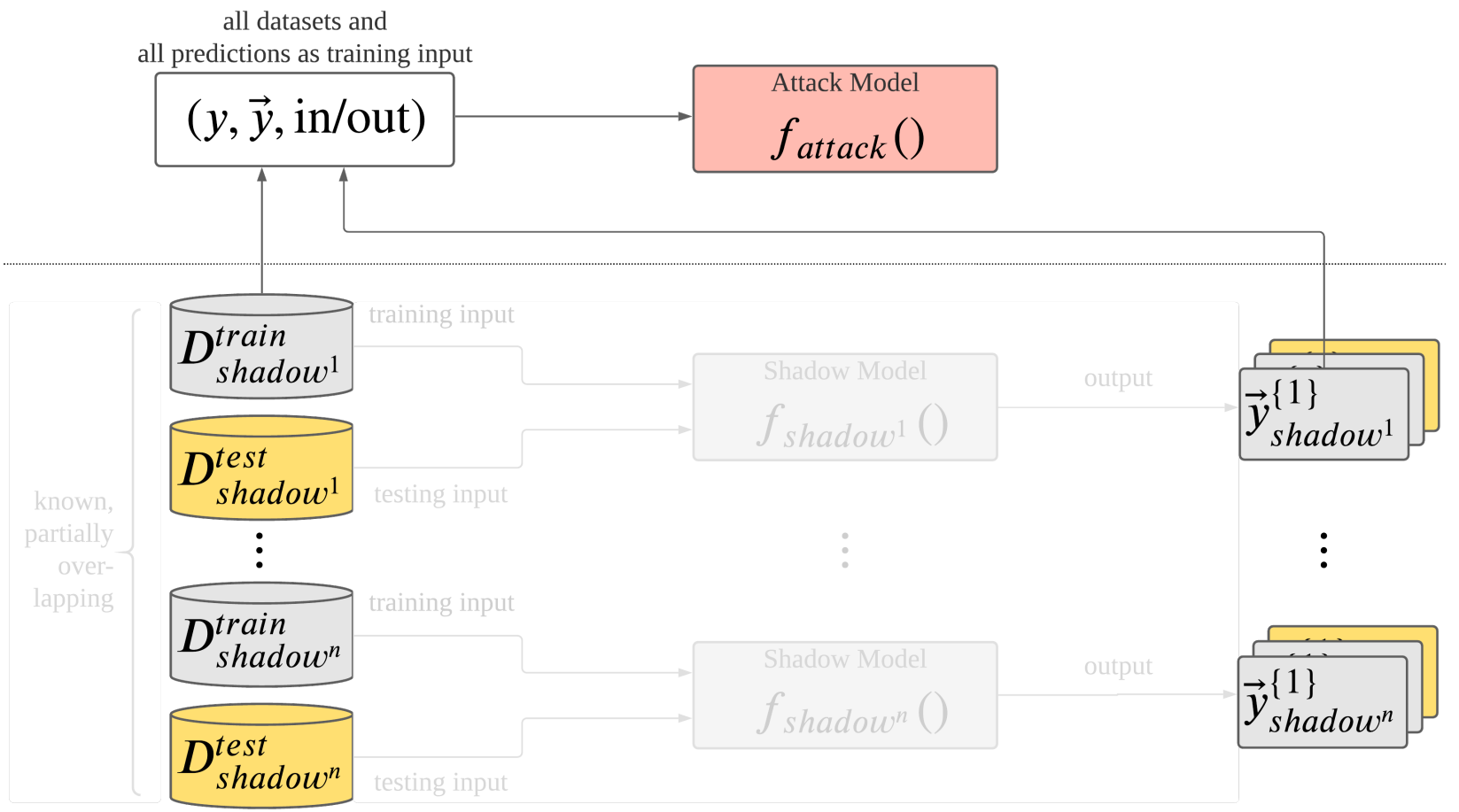
Membership Inference Attacks

Visualization



Membership Inference Attacks

Visualization



Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup, **Parameters**

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup, **Parameters**

- **Target models**: MLaaS models by Google Prediction API and Amazon ML as well as a self-implemented neural network

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup, **Parameters**

- **Target models**: MLaaS models by Google Prediction API and Amazon ML as well as a self-implemented neural network
- **Shadow model training data** (synthetic or noisy):
  - *Model-based synthesis*: Generating synthetic training data iteratively by using the target model itself
  - *Statistics-based synthesis*: Using statistical knowledge about the population underlying the target models data
  - *Noisy data*: Using entries of the real datasets and flipping of 10 or 20% of random features

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup, **Parameters**

- **Target models**: MLaaS models by Google Prediction API and Amazon ML as well as a self-implemented neural network
- **Shadow model training data** (synthetic or noisy):
  - *Model-based synthesis*: Generating synthetic training data iteratively by using the target model itself
  - *Statistics-based synthesis*: Using statistical knowledge about the population underlying the target models data
  - *Noisy data*: Using entries of the real datasets and flipping of 10 or 20% of random features
- **Datasets**:
  - UCI Adult, Purchases, CIFAR, MNIST, Locations, Hospital Stays
  - Tasks: Binary up to 100-class classification
  - Features: Images, binary (446 - 6,170), ordinal attributes (14)
  - Entities: 5,000 up to 70,000

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy
- Experiments: Comparison of . . .

. . . different **training set sizes**

. . . different **numbers of target classes**

. . . the six **datasets**

. . . the three **target models**

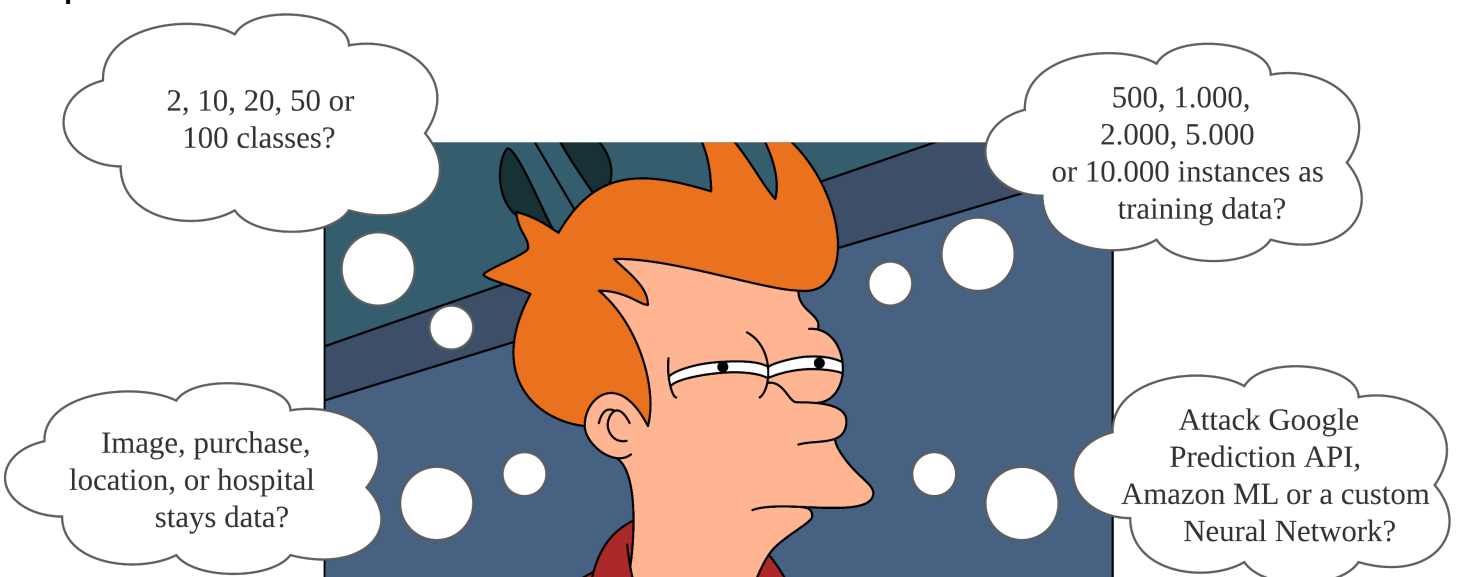
. . . the three **shadow model data generation techniques**

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy
- Experiments:



https://knowyourmeme.com/memes/futurama-fry-not-sure-if

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy
- Experiments: Comparison of . . .
  - . . . different **training set sizes**
  - . . . different **numbers of target classes**
  - . . . the six **datasets**
  - . . . the three **target models**
  - . . . the three **shadow model data generation techniques**

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy
- Experiments: Comparison of . . .
  - . . . different **training set sizes**
  - . . . different **numbers of target classes**
  - . . . the six **datasets**
  - . . . the three **target models**
  - . . . the three **shadow model data generation techniques**

Results

- The higher the **number of classes**, the better the attack works.

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy
- Experiments: Comparison of . . .

. . . different **training set sizes**

. . . different **numbers of target classes**

. . . the six **datasets**

. . . the three **target models**

. . . the three **shadow model data generation techniques**

Results

- The higher the number of classes, the better the attack works.
- The higher the **degree of overfitting**, the better the attack works.

Membership Inference Attacks

Experiments by [Shokri et al. 2017]

Setup (continued)

- Performance measures:
  - Target model: Training and testing accuracy
  - Attack model: Precision and recall, total accuracy
- Experiments: Comparison of . . .

. . . different **training set sizes**

. . . different **numbers of target classes**

. . . the six **datasets**

. . . the three **target models**

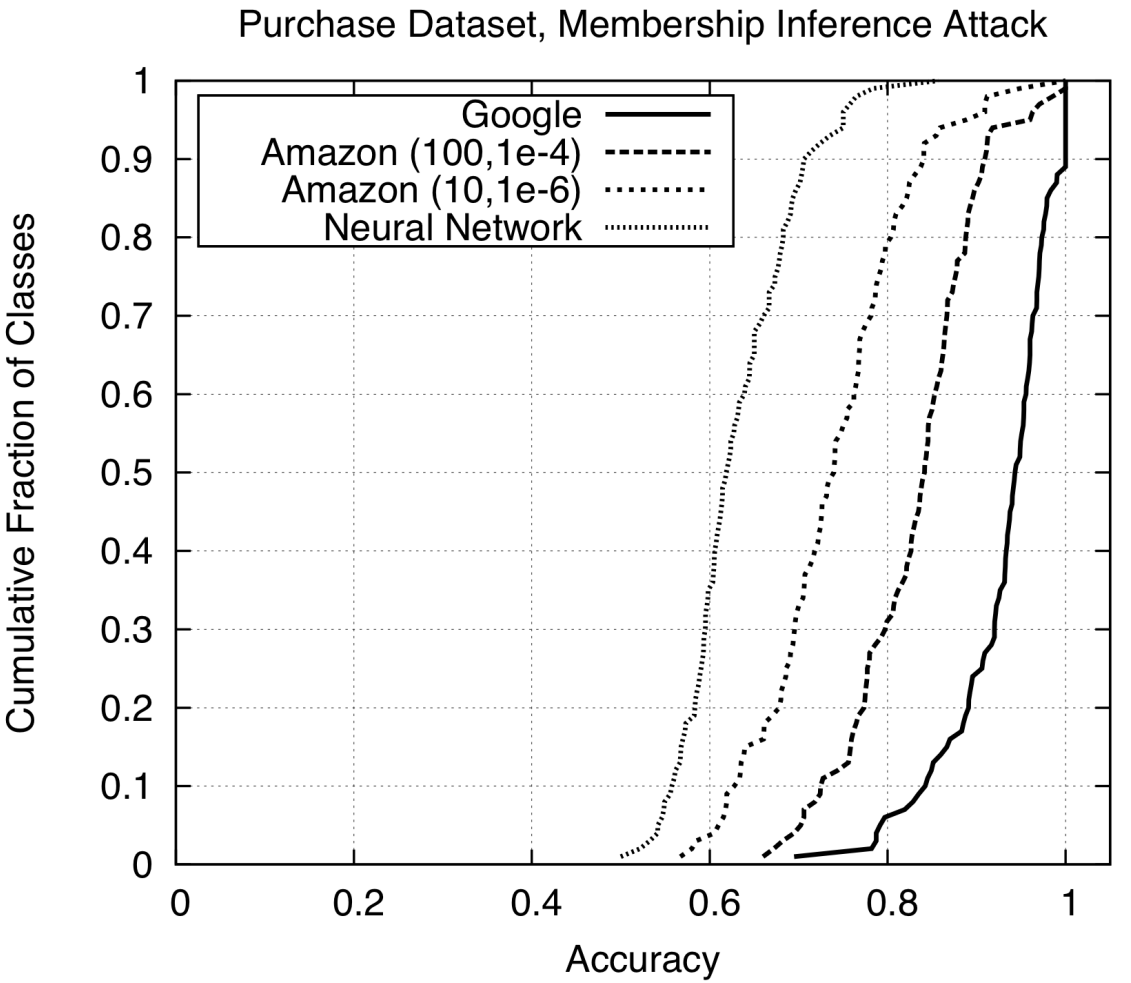
. . . the three **shadow model data generation techniques**

Results

- The higher the number of classes, the better the attack works.
- The higher the degree of overfitting, the better the attack works.
- For the shadow models, **even noisy or fully synthetic data work**.

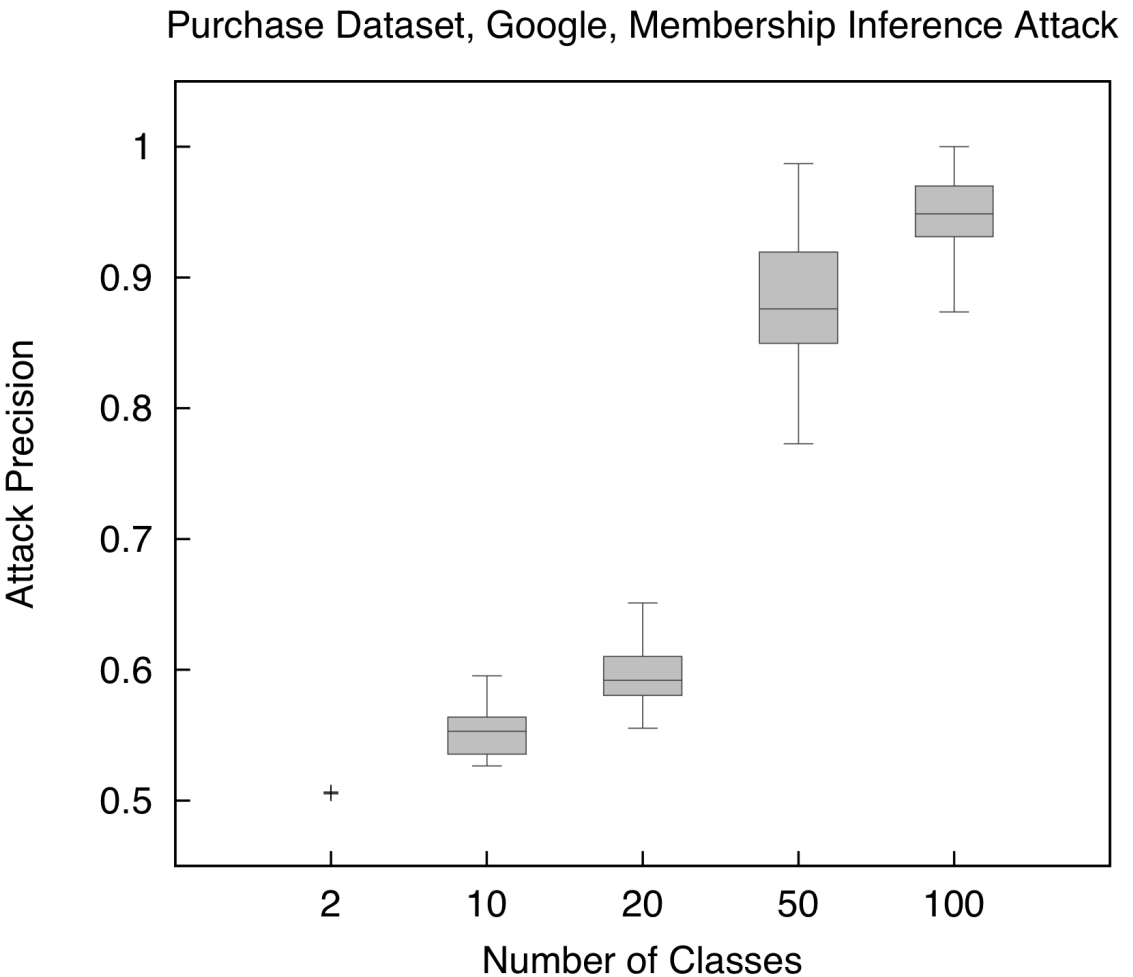
Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Comparison of Target Models



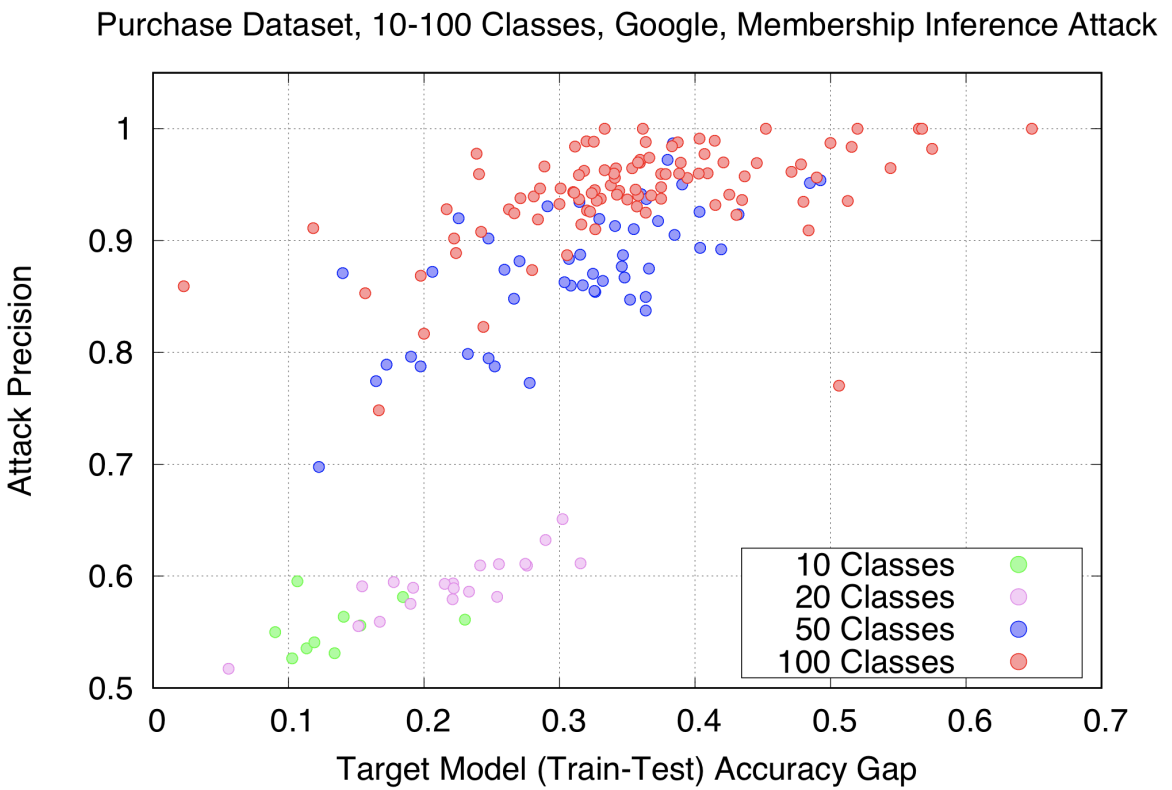
Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Effect of the Amount of Classes



Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Effect of Overfitting



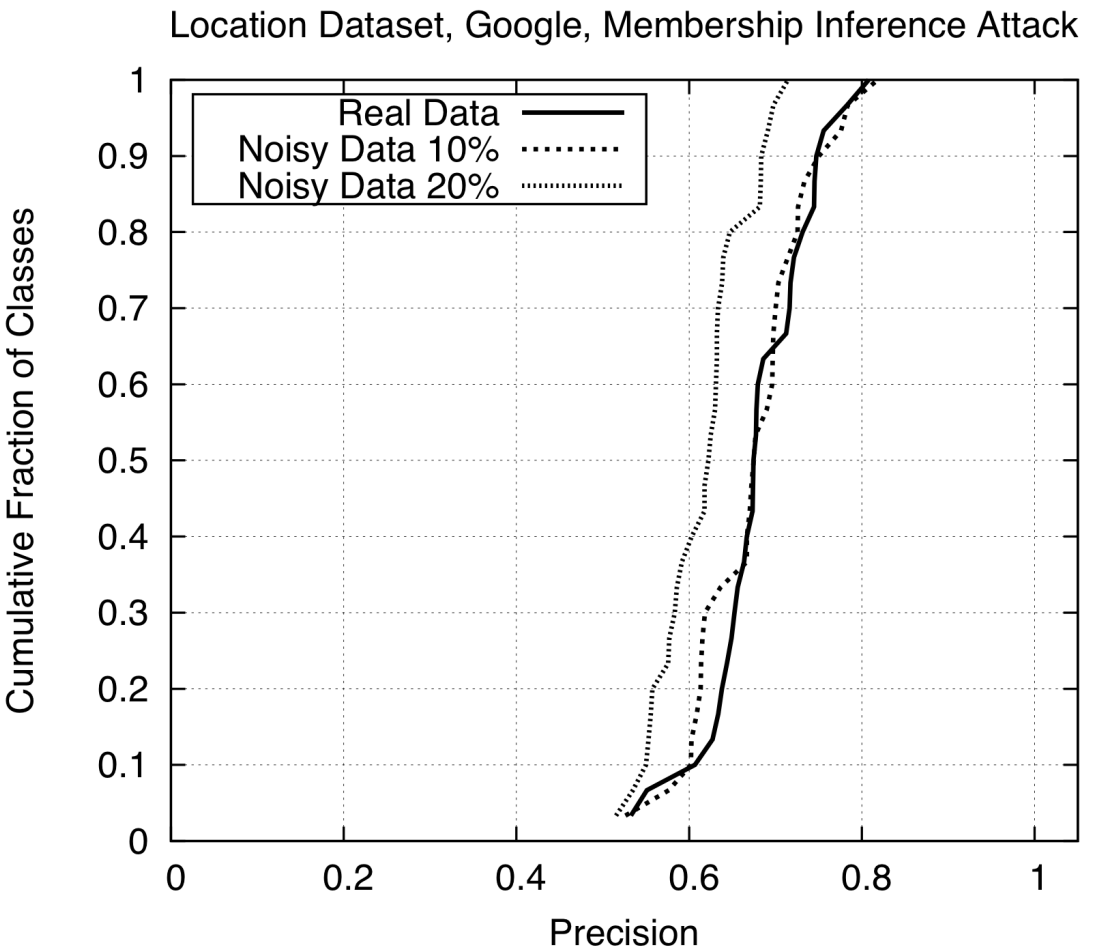
Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Effect of Overfitting Table

Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Effect of Noisy Data



Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Mitigation

Strategies

- Limit the number of classes in the prediction vector to the top k.
- Coarsen the exactness of entries in the prediction vector.
- Enlarge the entropy of the prediction vector.
- Apply regularization.

Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Mitigation

Strategies

- Limit the number of classes in the prediction vector to the top k.
- Coarsen the exactness of entries in the prediction vector.
- Enlarge the entropy of the prediction vector.
- Apply regularization.

Results

Hospital dataset	*Testing Accuracy*	*Attack Total Accuracy*	*Attack Precision*	*Attack Recall*
No Mitigation	0.55	0.83	0.77	0.95
Top $k = 3$	0.55	0.83	0.77	0.95
Top $k = 1$	0.55	0.82	0.76	0.95
Top $k = 1$ label	0.55	0.73	0.67	0.93
Rounding $d = 3$	0.55	0.83	0.77	0.95
Rounding $d = 1$	0.55	0.81	0.75	0.96
Temperature $t = 5$	0.55	0.79	0.77	0.83
Temperature $t = 20$	0.55	0.76	0.76	0.76
L2 $\lambda = 1e-4$	0.56	0.80	0.74	0.92
L2 $\lambda = 5e-4$	0.57	0.73	0.69	0.86
L2 $\lambda = 1e-3$	0.56	0.66	0.64	0.73
L2 $\lambda = 5e-3$	0.35	0.52	0.52	0.53

Membership Inference Attacks

Experiments by [Shokri et al. 2017] - Mitigation

Strategies

- Limit the number of classes in the prediction vector to the top k.
- Coarsen the exactness of entries in the prediction vector.
- Enlarge the entropy of the prediction vector.
- **Apply regularization.**

Results

Hospital dataset	*Testing Accuracy*	*Attack Total Accuracy*	*Attack Precision*	*Attack Recall*
No Mitigation	0.55	0.83	0.77	0.95
Top $k = 3$	0.55	0.83	0.77	0.95
Top $k = 1$	0.55	0.82	0.76	0.95
Top $k = 1$ label	0.55	0.73	0.67	0.93
Rounding $d = 3$	0.55	0.83	0.77	0.95
Rounding $d = 1$	0.55	0.81	0.75	0.96
Temperature $t = 5$	0.55	0.79	0.77	0.83
Temperature $t = 20$	0.55	0.76	0.76	0.76
L2 $\lambda = 1e-4$	0.56	0.80	0.74	0.92
L2 $\lambda = 5e-4$	0.57	0.73	0.69	0.86
L2 $\lambda = 1e-3$	0.56	0.66	0.64	0.73
L2 $\lambda = 5e-3$	0.35	0.52	0.52	0.53

Membership Inference Attacks

Further Research

- Model and Data Independent Membership Inference Attack [Humbert et al. 2019]
- MIA against adversarial defense methods [Shokri et al. 2019]
- Membership Inference Attacks against Generative Models (GANs) [Fritz et al. 2020]
- MIA against a label-only target model [Zhang et al. 2020]

Thank you.