Membership Inference Attacks against Machine Learning Models

Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov
(Cornell Tech)

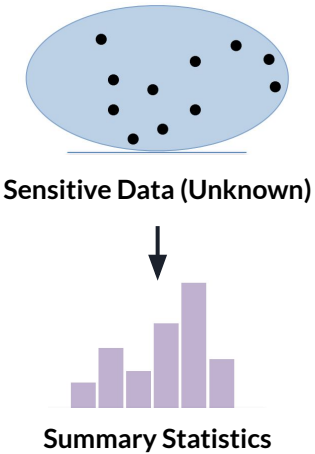
Paper appears in S&P '2017
Slide deck for discussion in UIUC CS 562 by Muhammad Adil Inam

Background - Membership Inference

- Has been previously used in the context of **statistics**
- Given the summary statistics (e.g Mean) on each attribute, can be identify if the target data record was in the sensitive data

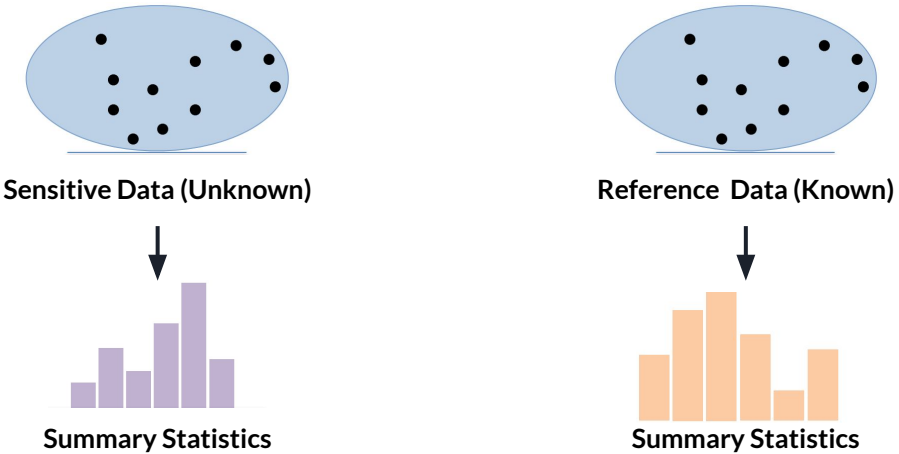
Background - Membership Inference

- Has been previously used in the context of **statistics**
- Given the summary statistics (e.g Mean) on each attribute, can be identify if the target data record was in the sensitive data



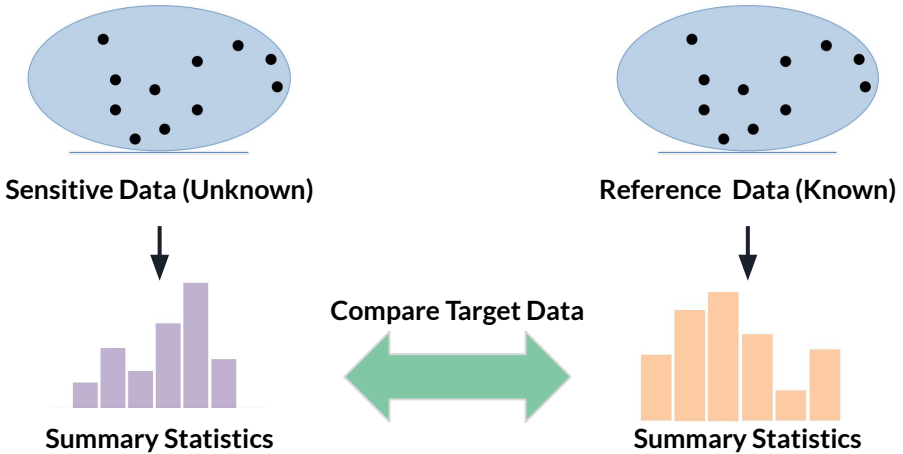
Background - Membership Inference

- Has been previously used in the context of **statistics**
- Given the summary statistics (e.g Mean) on each attribute, can be identify if the target data record was in the sensitive data



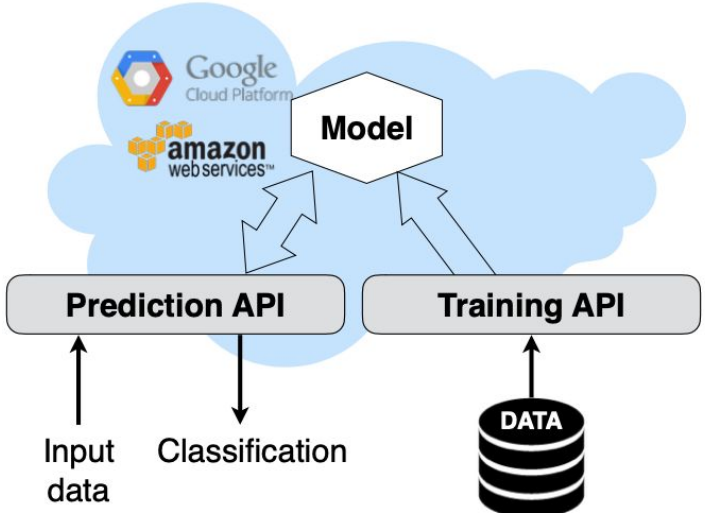
Background - Membership Inference

- Has been previously used in the context of **statistics**
- Given the summary statistics (e.g Mean) on each attribute, can be identify if the target data record was in the sensitive data



Background - ML as a Service

- Internet giants such as Google and Amazon are offering “machine learning as a service.”
- The service then makes the model available to the customer, typically as a black-box API



Membership Inference in ML models

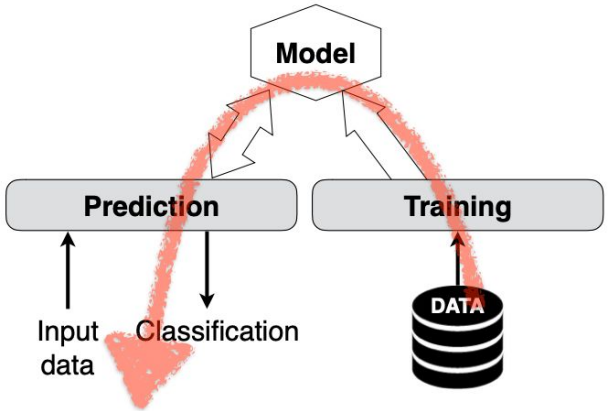
- *Given a machine learning model and a record, determine whether this record was used as part of the model's training dataset or not*

- Do ML predictions leak information about training data?

Membership Inference in ML models

- *Given a machine learning model and a record, determine whether this record was used as part of the model's training dataset or not*

- Do ML predictions leak information about training data?



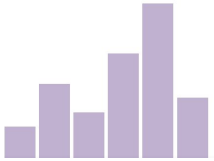
Privacy Implications of Data Leakage

- In most cases, when the underlying training dataset is not sensitive, data leakage has little to no implications
- However, in certain cases with sensitive data, it can directly lead to a privacy breach.
- **Scenario:** knowing that a certain patient's clinical record was used to train a model associated with a disease (e.g, to determine the appropriate medicine dosage or to discover the genetic basis of the disease) can reveal that the patient has this disease.

Main Intuition and Insight

- ML models overfit to their training data due to lack of generalization
- The **classification behaviour** varies if the input in “from the training set” or “not from the training set.
- *Can we recognize the difference between these behaviors?*

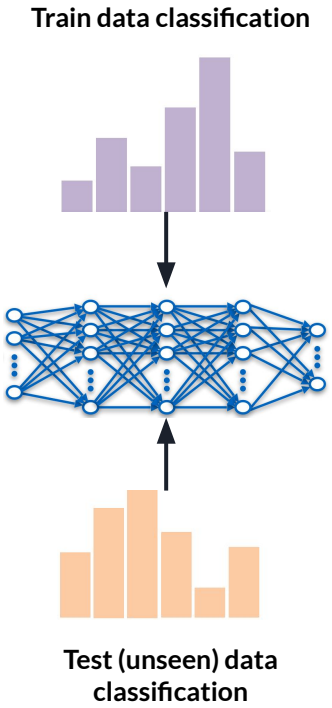
Train data classification



Test (unseen) data classification

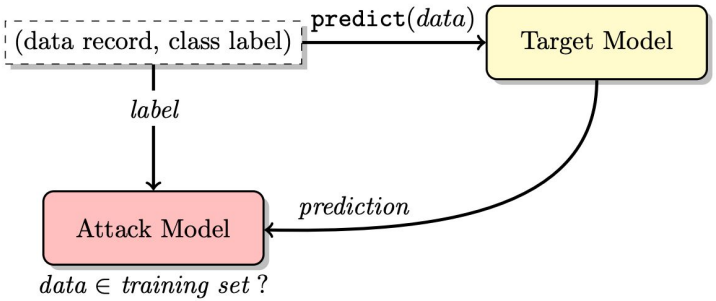
Main Intuition and Insight

- ML models overfit to their training data due to lack of generalization
- The **classification behaviour** varies if the input in “from the training set” or “not from the training set.
- *Can we recognize the difference between these behaviors?*
  - **Train a ML model** to recognize the difference! (Attack Model)



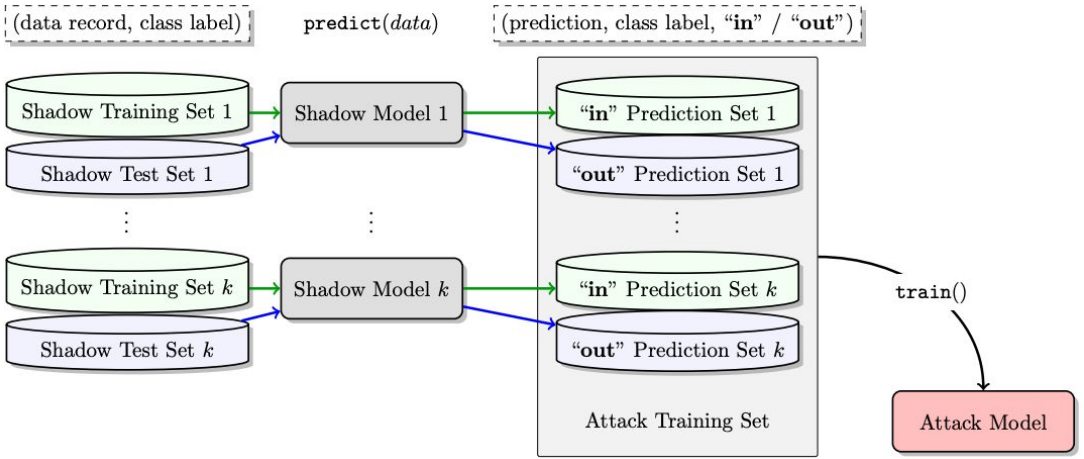
Constraints on the Attack Model

- No prior knowledge about the training algorithm, model type or models' parameters of the target (Black Box setting)
- No access to internal computations of the target model
- No prior knowledge about the underlying distribution of trained data



Shadow Models

- Create multiple shadow models that **imitate** the behavior of the target model
- Train the attack model on the labeled inputs and outputs of the shadow models



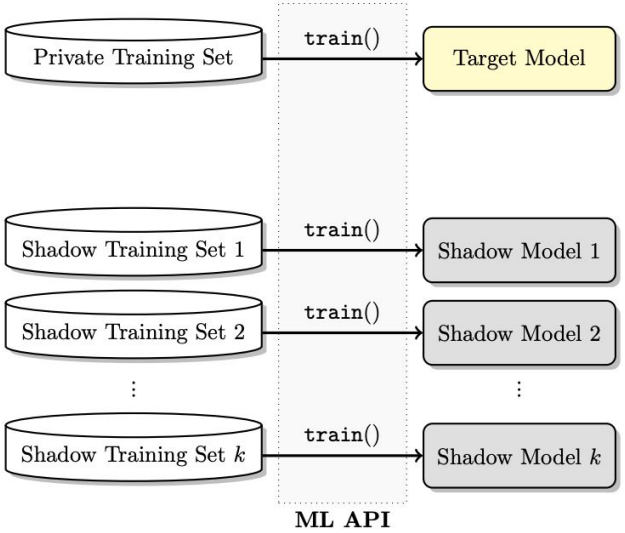
1. How to define the type and architecture for shadow models?

2. How to obtain data for training shadow models?

Model Type and Architecture

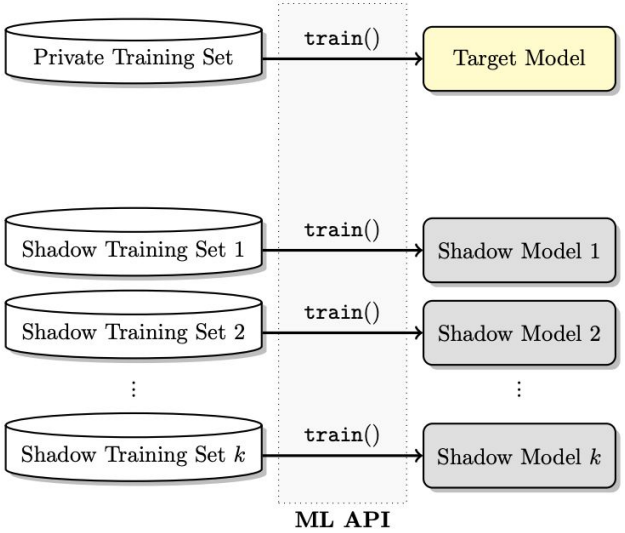
- An attacker can use exactly the same service (e.g., Google Prediction API) to train the shadow model as was used to train the target model

- **Assumption:** For the same classification task, cloud services would make use of similar models



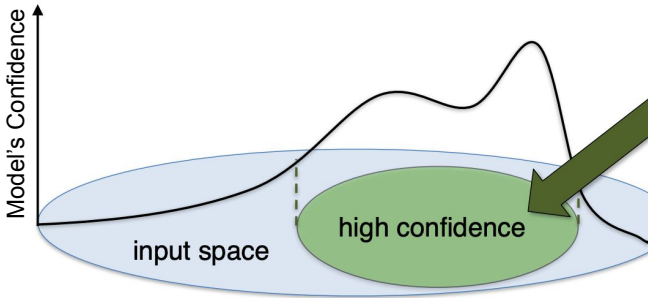
Model Type and Architecture

- An attacker can use exactly the same service (e.g., Google Prediction API) to train the shadow model as was used to train the target model
- **Assumption:** For the same classification task, cloud services would make use of similar models
- ***Discussion Question?***
  - How realistic is this given assumption?



Data : Model-based Synthesis

- Generate **synthetic** training data for the shadow models using the target model itself
- Records that are classified by the target model with high confidence should be statistically similar to the target's training dataset
- Using a hill-climbing algorithm, **search** the space of possible data records to find inputs that are classified by the target model with high confidence

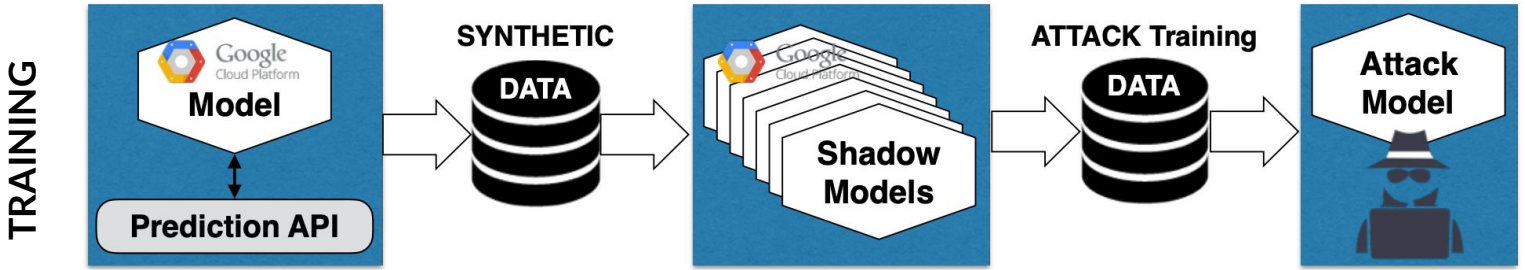


Data : Other Methods

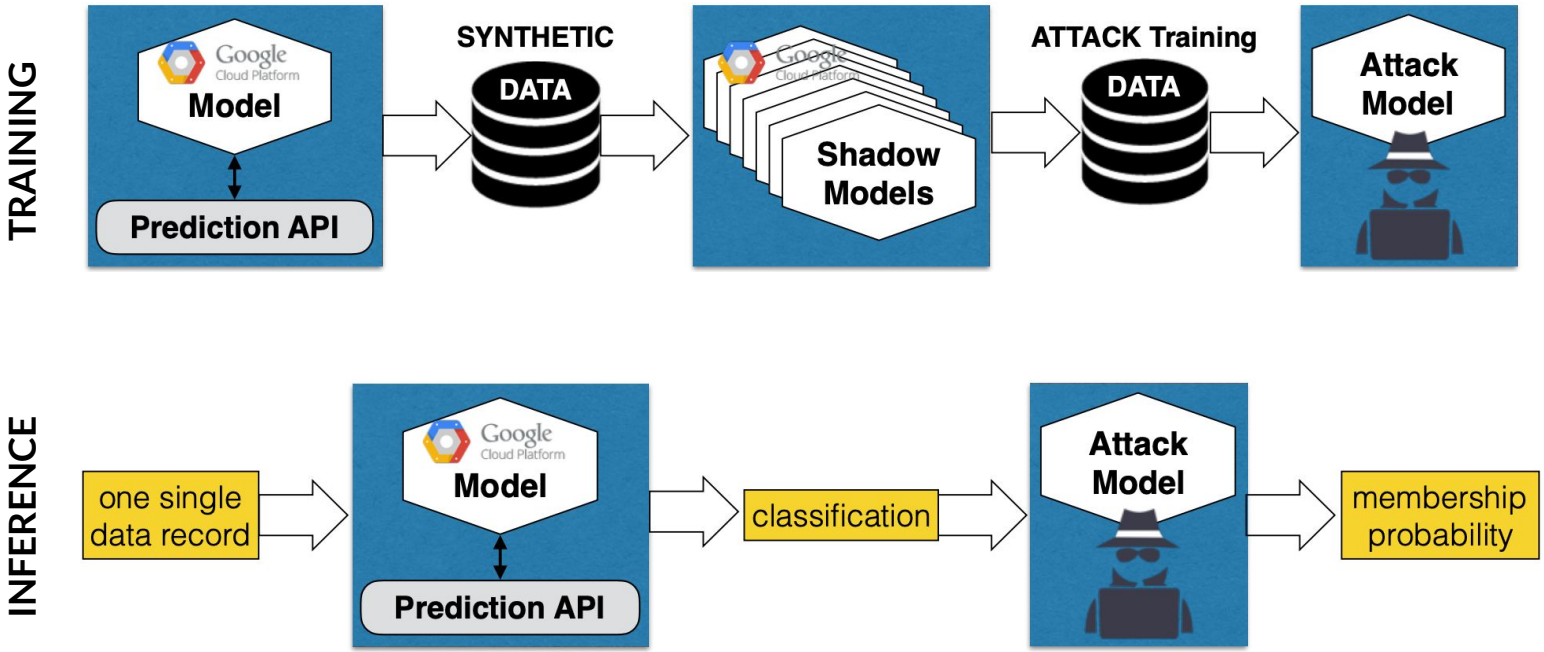
- **Statistics-based synthesis -** if the attacker have some statistical information about the population of the target model's training data, synthetic data can be generated by independently sampling the value of each feature from its own marginal distribution

- **Noisy real data -** an attacker may have access to some data that is similar to the target model's training data and can be considered as a "noisy" version. In the experiments, they simulate this by flipping the (binary) values of 10% or 20% randomly selected features

Complete Workflow



Complete Workflow

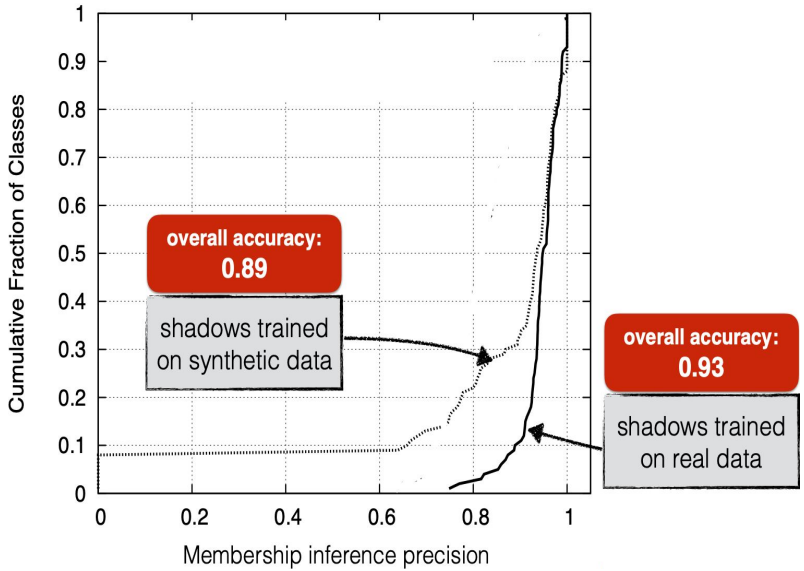


Discussion Questions

- In practice, would it cause trouble because you would need to send the API interface a lot of requests?
- What is the effect of the parameter k i.e the number of shadow models?
- Is synthetic data generation feasible for complex datasets?

Evaluation - Attack Model

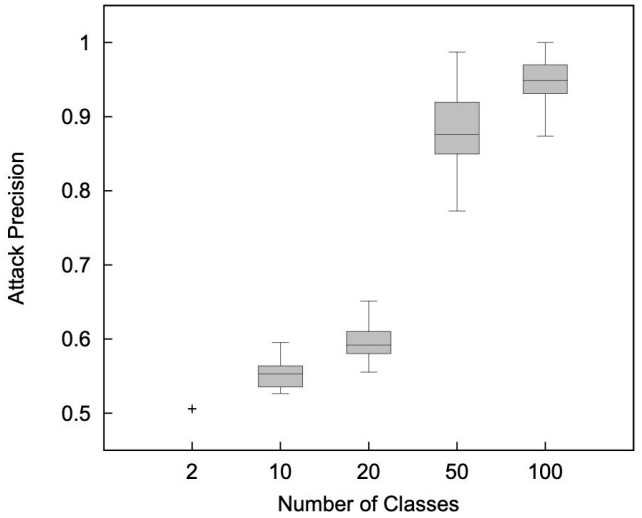
- For the majority of the target model's classes, their attack achieves high precision
- Demonstrates an attacker can efficiently generate high confidence inputs that a membership inference attack can be trained on, with only black-box access to the target model



Purchase Dataset, Google, Membership Inference Attack (100 classes)

Evaluation - Number of Classes

- The more classes, the more signals about the internal state of the model are available to the attacker
- As the number of classes increases, the model needs to extract more distinctive features from the data i.e need to remember more about their training data, thus they leak more information



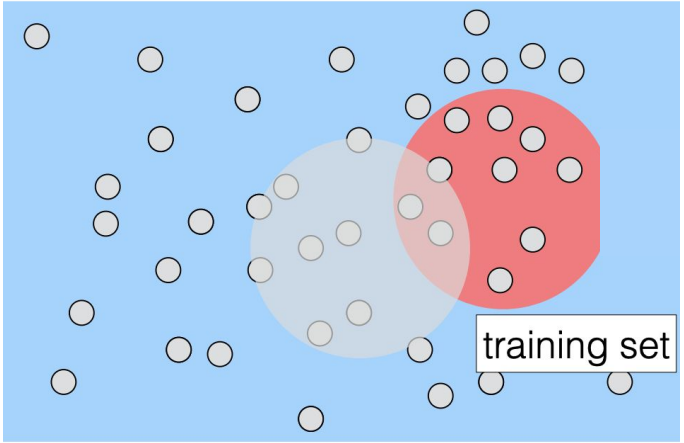
Purchase Dataset, Google, Membership Inference Attack

Mitigation Strategies

- Restrict the prediction vector to top k classes
- Coarsen precision of the prediction vector - round the classification probabilities in the prediction vector down to d floating point digits - smaller d leads to less leakage
- Increase entropy of the prediction vector - modify (or add) the softmax layer and increase its normalizing temperature
- Make use of regularization - Ridge and Lasso etc

Utility vs Privacy

- Utility: Does the model generalize to data outside the training set?
- Privacy: Does the model leak information about data in the training set?
- Overfitting is the common enemy of the two - hence the goals of the two are aligned



Discussion Questions

Discussion Questions

- Does differential privacy defeat membership inference attacks?
- Will the attack still be successful if the attacker only gets access to the prediction value of top “k” labels (say k=1) from the target model?
- Does this work only with neural networks or also with older ML techniques?