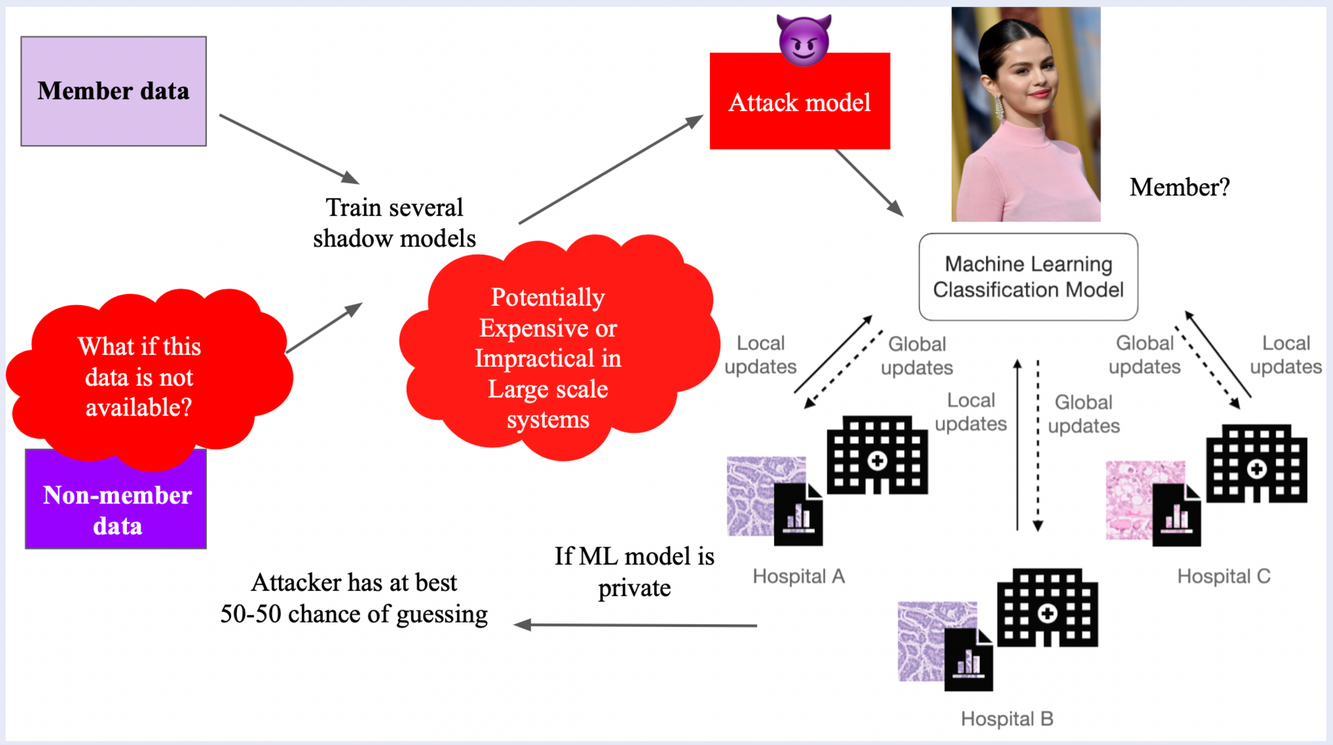
BLACK-BOX PRIVACY AUDITING OF MACHINE LEARNING PIPELINES

Mishaal Kazmi, Alireza Akbari, Hadrien Laurette, Sébastien Gambs, Mathias Lécuyer

DOES PRIVACY AUDITING OF MACHINE LEARNING (ML) MODELS HAVE TO BE SO COSTLY?

THE PROBLEM

- Machine Learning models are susceptible to **training data memorization** and may leak this data, therefore, causing a **breach in privacy.**
- Existing solutions require alterations in the training pipeline of ML models, leading to **expensive** re-training of these models especially in **large-scale industry settings.**



RQ: CAN WE ESTIMATE THE PRIVACY GUARANTEES OF AN ML MODEL WITHOUT THE NEED FOR RETRAINING/RESAMPLING?

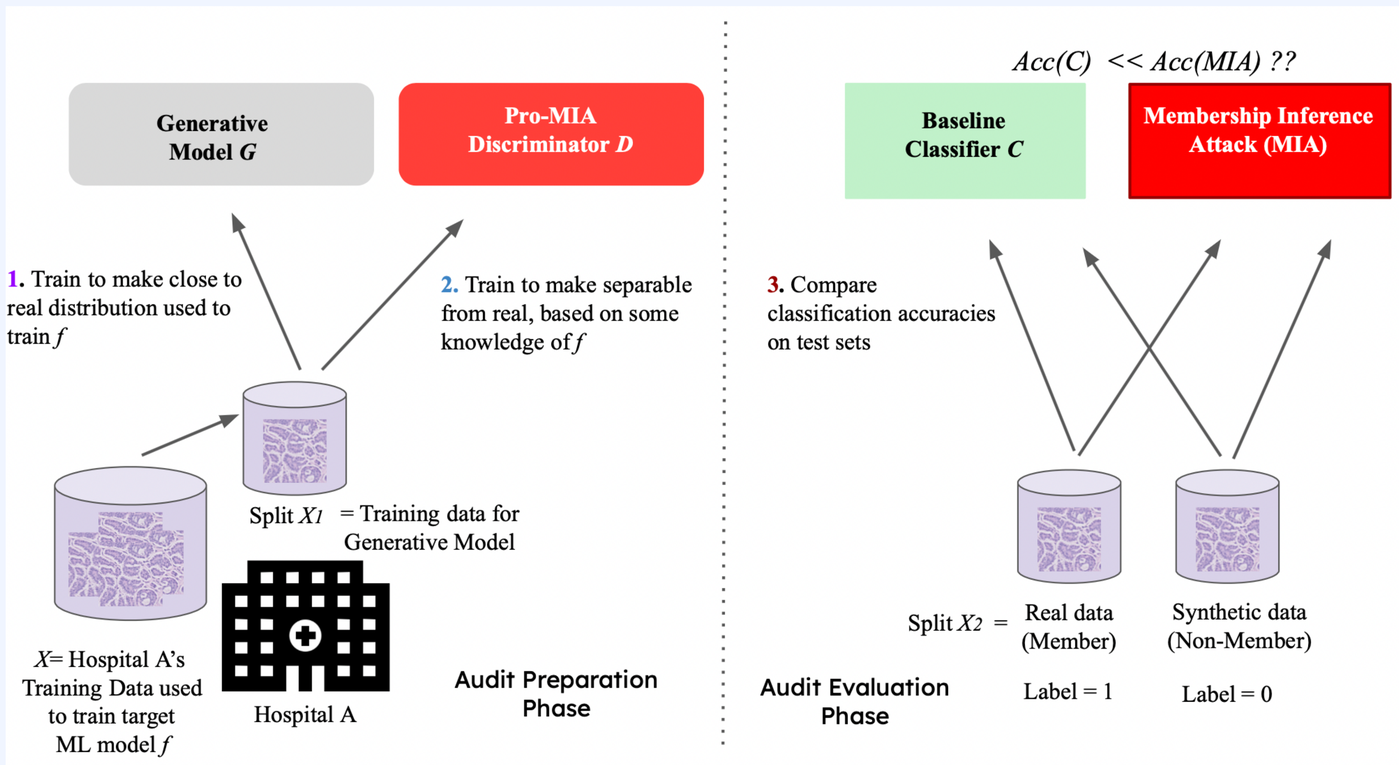
OUR CONTRIBUTIONS

A **black-box training data privacy auditing mechanism** that investigates the privacy guarantees of a machine learning model **without the need** for:

1. **retraining** the machine learning model,
2. **poisoning** the training data, thereby auditing the ML model not the algorithm
3. **knowing the whole dataset** the ML model has previously trained on

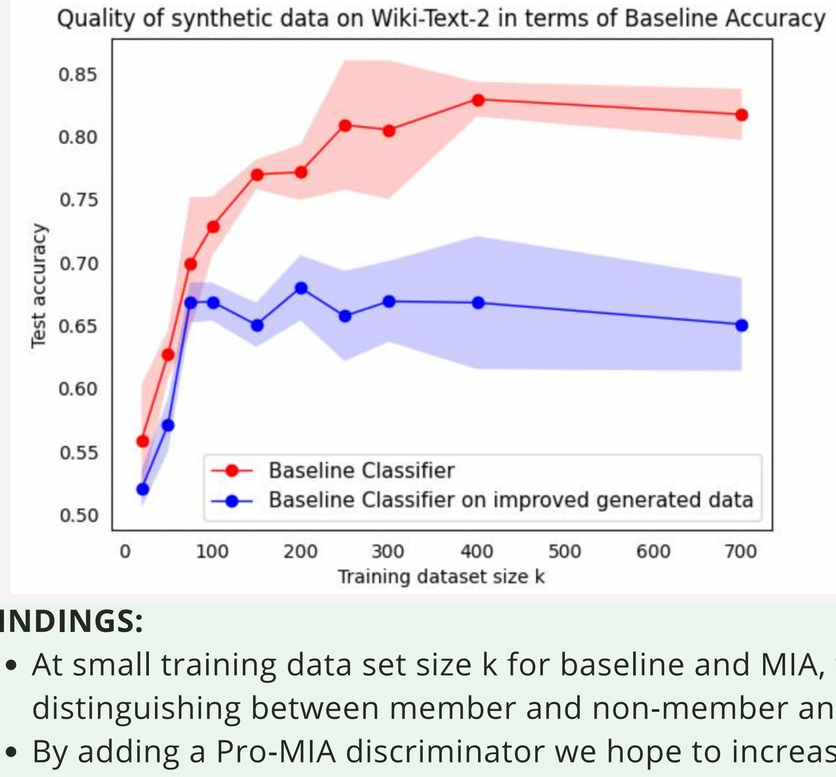
via the membership inference attack (MIA) on generated/synthetic (out) data

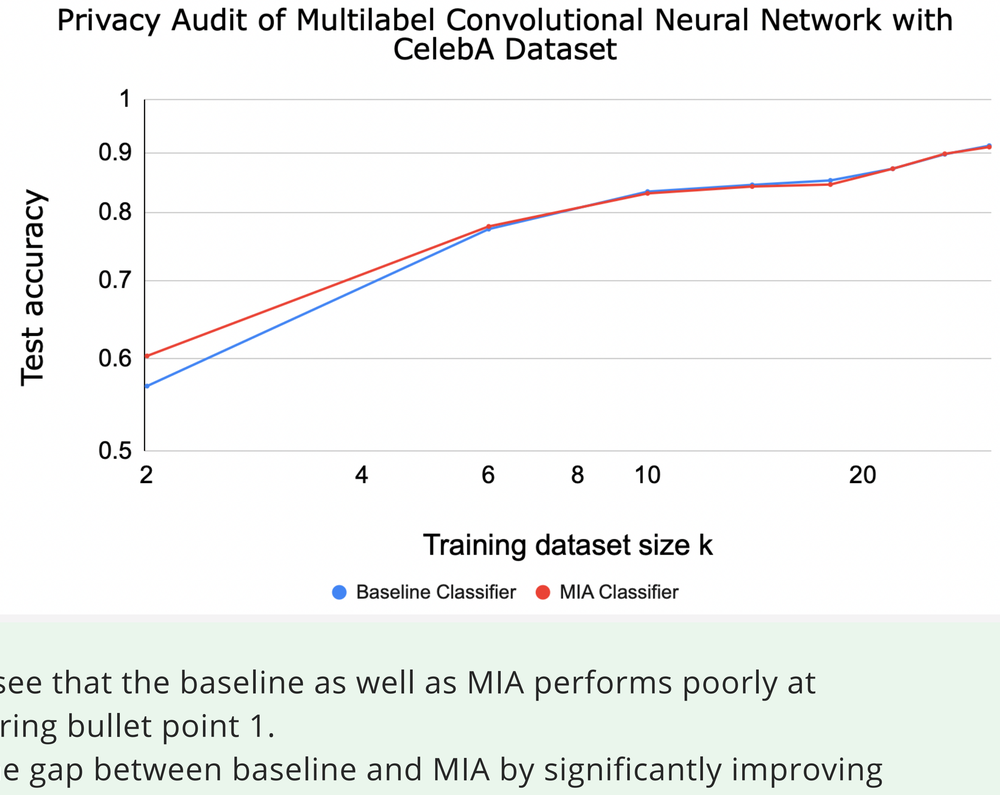
SYSTEM DESIGN



EVALUATION AND FINDINGS

- Can we synthesize non-member data that is good enough to fool a baseline classifier at distinguishing between member (real) vs non-member (fake) data?
- Can a membership classifier do much better than a baseline classifier having knowledge of the logits of the target model "f"?
- Can the accuracy of baseline vs MIA give a proxy of privacy leakage for a target ML model "f"?





FINDINGS:

- At small training data set size k for baseline and MIA, we see that the baseline as well as MIA performs poorly at distinguishing between member and non-member answering bullet point 1.
- By adding a Pro-MIA discriminator we hope to increase the gap between baseline and MIA by significantly improving the performance of the MIA while maintaining the low performance of the baseline. By this we hope to estimate the privacy leak attributed to the model f to help answer bullets 2 and 3.

SYSTOPIA UQÀM UBC

mkazmi@cs.ubc.ca