

PrivacyRaven: Comprehensive Privacy Testing for Deep Learning

Suha S. Hussain

#### whoami



- Suha S. Hussain (@suhackerr)
  - CS Undergraduate at Georgia Tech
    - Threads: Theory & People
  - Security Engineering Intern at Trail of Bits
    - Cryptography Team

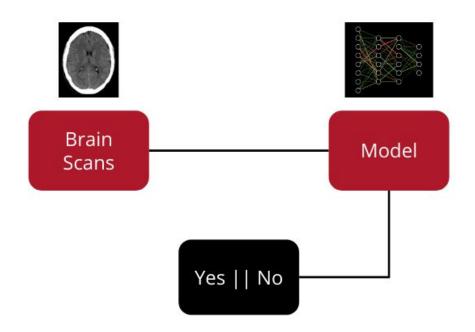


#### **Auditing Deep Learning**



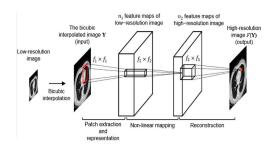
#### How can this system be attacked?

- Purpose: Detect a brain bleed from images of a scan
  - Black-box
  - Binary result
- Use PrivacyRaven to simulate privacy attacks



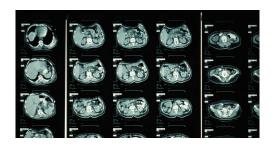
#### **Privacy Violations**





#### **Intellectual Property**

A substitute model was created from a **model extraction** attack.



#### **Data Reconstruction**

The adversary launched a **model inversion** attack.

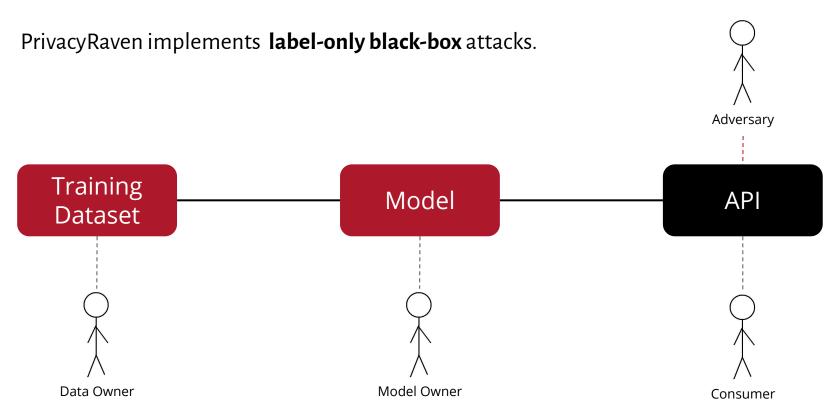


#### Re-identification

A **membership inference** attack was executed.

#### **Threat Model**





#### Affordances



- Determine the susceptibility of a model to different privacy attacks
- Evaluate privacy preserving machine learning techniques
- Develop novel privacy metrics and attacks
- Repurpose attacks for data provenance auditing and other use cases

## **Model Extraction**

TRAJL BITS

#### **Attack Objectives**



#### Model with High Accuracy

This attack is typically **financially motivated**.

Avoid paying for the target model in the future or profit off of extracted model.

#### Model with High Fidelity

This attack is typically reconnaissance-motivated.

Learn more about the original model and launch other classes of attacks.

#### A Framework for Model Extraction



Model extraction attacks can be partitioned into **multiple phases**.



#### Extract an MNIST model



#### Launch an attack in under 15 lines of code

```
model = train_mnist_victim()
def query_mnist(input_data):
    return get_target(model, input_data)
emnist_train, emnist_test = get_emnist_data()
attack = ModelExtractionAttack(query_mnist, 100,
    (1, 28, 28, 1),
    10,
    (1, 3, 28, 28),
    "copycat",
    ImagenetTransferLearning,
    1000,
    emnist_train,
    emnist_test,
```

#### **Extraction Results**



- Target Model Statistics
- Synthetic Dataset Details
- Substitute Model Statistics
- Accuracy Metrics
- Fidelity Metrics

### Membership Inference

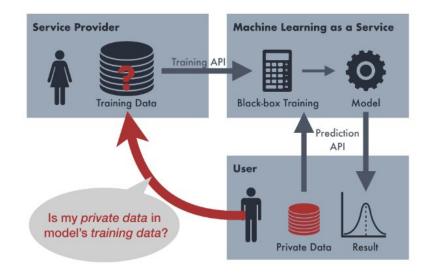
TRAJL BITS

#### An Overview of Membership Inference



#### Objective: Re-identification

- Integrates the extraction API
- Unique threat model



#### A Framework for Membership Inference



Membership inference attacks can also be partitioned into **multiple phases**.

#### Label-Only Membership Inference Attacks

Christopher A. Choquette-Choo\*†, Florian Tramèr‡, Nicholas Carlini‡, Nicolas Papernot\* University of Toronto\*, Vector Institute†, Stanford University†, Google<sup>§</sup>

1 INTRODUCTION

nettai corrections [1, 4], or financial information [1, 4]. Problem in infantry to Popule are continued score, Popular and Comparison of the Comparison of t information permissing experiturally to according of the model's solution operation permission generalized by the consideration of the configuration of the

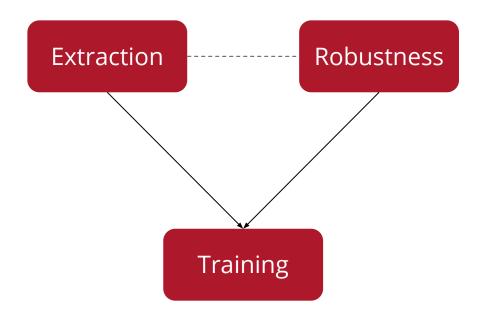
prediction confidence that models exhibit on their training data [7, 10, 11, 12, 13, 14]. This difference in prediction the training data [7, 10, 11, 12, 13, 14]. This difference in prediction the training defenses either implicitly or explicitly rely on a strategy

Advance—Mandroodly inference situate, or one of the singlar flower of privacy behalps for mandred to moving models great
and the privacy behalps for mandred to moving models great
truth to model. For the single models of the model of a stage of dark
truth to model for the single models of the model of a stage of dark
truth to model. For the single models of the model of a stage of dark
truth to model or mixed models or the model or models or the model or mixed models or the models or mixed models or the model or mixed models or the models or mixed models or mixed models or models or mixed models and mixed models or mixed models and mixed mo

labels when querying the trained model, without any prediction Machine learning algorithms are often trained on sensitive confidences. This threat model is more realistic in practiceor private user information, such as medical records [1, 3], as many machine learning models deployed in user-facing textual conversations [1, 4], of financial information [6, 4], or financial information [6, 4], or financial information [6, 4], products are utilizely to expose raw confidence scans.

set reveals that the victim indeed has cancer.

Existing membership inference attacks exploit the higher



## **Model Inversion**

TRAJL

#### An Overview of Model Inversion



#### Objective: Obtain memorized data

- More nebulous area of work
- Integrates the extraction API
- Trains an "inverse" network





#### **Upcoming Features**



- New interface for metrics visualizations
- Automated hyperparameter optimization
- Certifiable differential privacy verification
- Privacy thresholds and metric calculations
- Side channel and property inference attacks
- Federated learning and generative model attacks
- Built-in victim models implementing PPML techniques

## Thank you for your time!

Are there any questions?



#### Repository:

github.com/trailofbits/PrivacyRaven

#### **Contact:**

suha.hussain@trailofbits.com james.miller@trailofbits.com

# TRAIL OF BITS