THINK BEFORE YOU CLICK!

FRAUDULENT EMAIL DETECTION

Advaith Shyamsunder Rao
Falgun Malhotra
Hsiao-Chun (Jeanie) Hung
Vanshita Gupta

ENRON SCANDAL

Emails played a significant role in uncovering the Enron scandal and providing evidence of the fraudulent activities within the company

CORPORATE FRAUD & ACCOUNTING MISCONDUCT

Deceptive financial practices to inflate profits and conceal substantial debts; manipulating financial statements, and exploiting accounting loophole

FINANCIAL INCENTIVES

Inflating the stock price allowed executives to profit personally as well as meet market expectations

THE EMAILS

Over 500K emails from approximately 150 Enron employees

DATASET

Enron Email Dataset lacks the availability of labeled frauds in the dataset.

Therefore, we use the combination of:

1. Enron email dataset
2. Phishing Dataset
3. Social Engineering Dataset

```
Message-ID: <1695208.1075857585551.JavaMail.evans@thyme>
Date: Tue, 14 Nov 2000 12:18:00 -0800 (PST)
From: mariamarcelle@hotmail.com
To: jarnold@enron.com
Subject: bank wire wsex
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: "maria marcelle" <mariamarcelle@hotmail.com>
X-To: jarnold@enron.com
X-cc:
X-bcc:
X-Folder: \John_Arnold_Dec2000\Notes Folders\Notes inbox
X-Origin: Arnold-J
X-FileName: Jarnold.nsf

Dear Mr.. Lavorato,

The $1500 that you sent to us in October, has not been credited to our
account.If those funds were sent through AM TRADE INTERNATIONAL, you need to
have your bank send an amendment message stating that the respective funds
are intended for final credit to World Sports Exchange/ ACCT # 12307915.

Most likely, those funds are sitting at the Antigua Overseas Bank.
```

2 GOLD TEST SETS

To perform high-quality testing, the project uses

GOLD FRAUD SET

Contains 1000 curated fraud emails from the phishing and social engineering dataset.

In this fraud dataset, we assess how many fraudulent emails our model misses.

SANITY SET

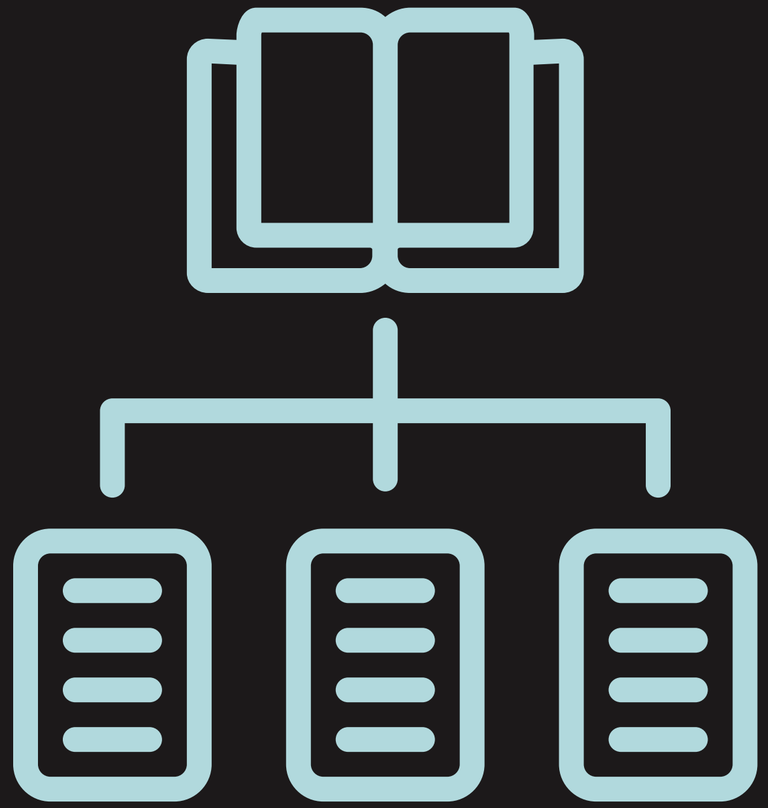
Contains 250K curated internal email communication emails between employees at Enron.

In this dataset, we conduct a precision test on our model to determine if it incorrectly flags any non-fraudulent emails as fraud.

LABEL ANNOTATION

To label the Enron email dataset, two heuristics are used to filter suspicious emails and label them into fraud and non-fraud classes.

1. Email Signals
2. Automated ML labeling



EMAIL SIGNALS

Email Signal based heuristics are used to specifically filter and target suspicious emails for fraud labeling. The signals used are:

1. Person Of Interest
2. Suspicious Folders
3. Sender Type
4. Low Communication
5. Contains Replies and Forwards

count	20131
mean	12.3
std	104.9
min	1
25%	1
50%	1
75%	4
max	5486

AUTOMATED ML LABELING

1. Phishing Annotation

   High precision model trained on the Phishing dataset.

2. Social Engineering Annotation

   High precision model trained on the Social Engineering dataset.

The two ML Annotator models use TFIDF to embed the input text and make use of SVM models with Gaussian Kernel.

DATA BREAKDOWN

Dataset Breakdown

	Fraud	Non-Fraud
Enron Dataset	2327	445090
Phishing Dataset	4976	12515
Social Engineering Dataset	4160	6475

Train Breakdown

Label	Distribution
0	214080
1	10463

DATA PREPROCESSING

01 LINK REMOVAL

We remove several types of links from the email body such as URLs and href tags and replace them with a common <URL> token.

02 HTML REPLACEMENT

We replace HTML in the email body with the text inside the HTML content. Usually, HTML tags are present in the email body when images are shared.

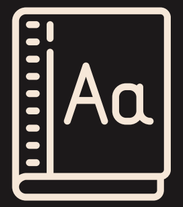
03 NEW LINES AND UNICODE REMOVAL

We remove new lines, emojis, and other special characters from the email text.

04 REMOVE SPECIFIC PATTERNS

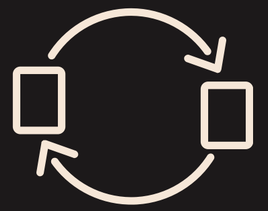
We remove specific patterns from our text such as Boilerplate, Signatures, and Greetings.

DATA AUGMENTATION



Synonym Replacement

Stopword Removal



Swapping Noun Phrases

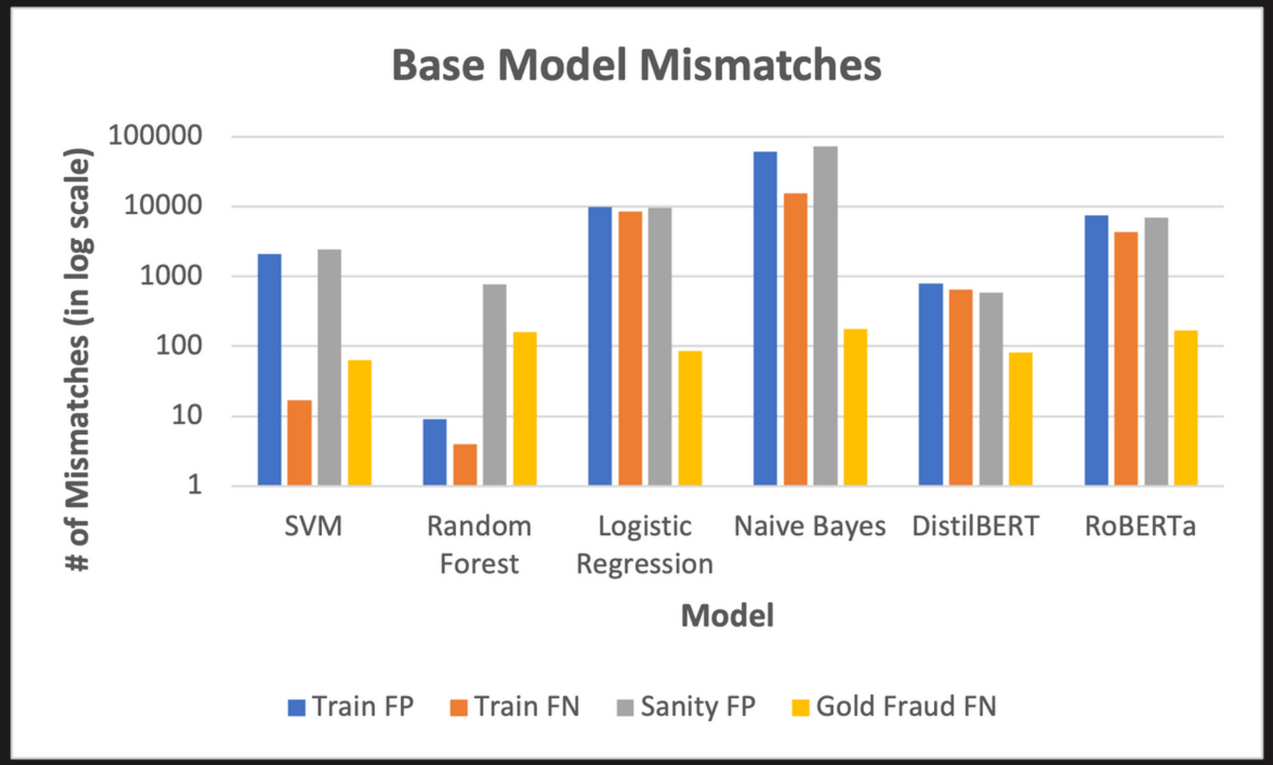
MODELING

6 different classification models

1. (Baseline) SVM
2. Naive Bayes
3. Logistic Regression
4. Random Forest
5. Pretrained DistilBERT Base
6. Pretrained RoBERTa Large

We also performed hyperparameter optimization to pick the best version of the models on our dataset, ensuring low False Positive rate in the Sanity set and low False Negative rate in the Gold Fraud set.

RESULTS



Base Models

1. SVM
2. Random Forest
3. Logistic Regression
4. Naive Bayes
5. DistilBERT
6. RoBERTa

DIFFERENTIAL PRIVACY

4 differentially private models

1. Random Forest
2. Logistic Regression
3. Naive Bayes
4. BERT

DIFFERENTIAL PRIVACY - BERT

1. Adding noise to the BERT parameters
2. Noise is introduced into the gradients of deep neural networks via DP-SGD
3. The goal of this strategy is to reduce the model's ability to memorize specific input instances
4. Gaussian noise to our parameters
5. Gradient Clipping

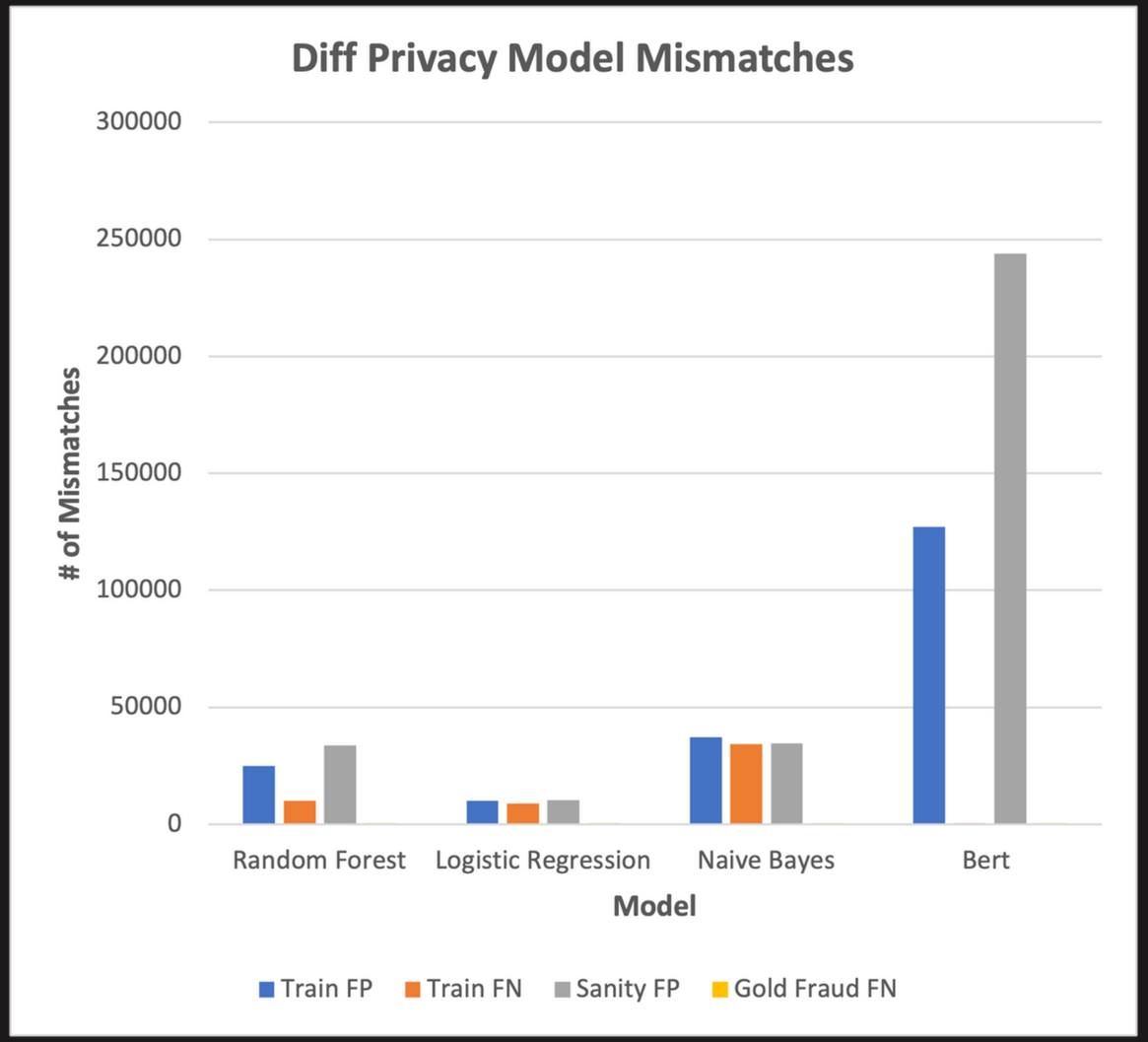
Sigma (Noise Multiplier)	0.37
C (Max Grad Norm)	0.1
Epsilon	1e-08
Target Epsilon	7.5
Delta	1/len(training_data) = 3.4e-05

DIFFERENTIAL PRIVACY - BERT

1. Adding noise to the BERT parameters
2. Noise is introduced into the gradients of deep neural networks via DP-SGD
3. The goal of this strategy is to reduce the model's ability to memorize specific input instances
4. Gaussian noise to our parameters
5. Gradient Clipping

Sigma (Noise Multiplier)	0.37
C (Max Grad Norm)	0.1
Epsilon	1e-08
Target Epsilon	7.5
Delta	1/len(training_data) = 3.4e-05

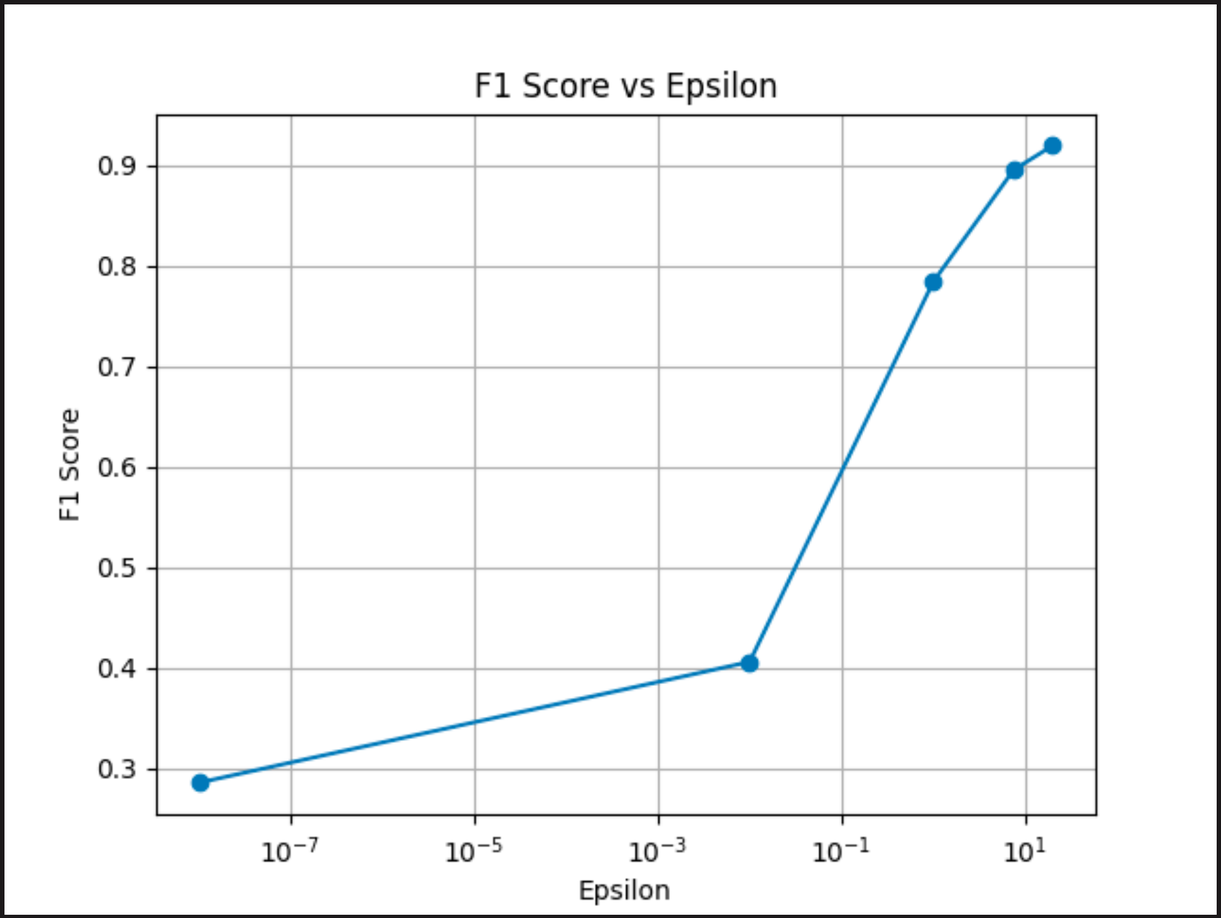
RESULTS



Diff Privacy Models

1. Random Forest
2. Logistic Regression
3. Naive Bayes
4. BERT

PRIVACY-ACCURACY TRADE-OFF

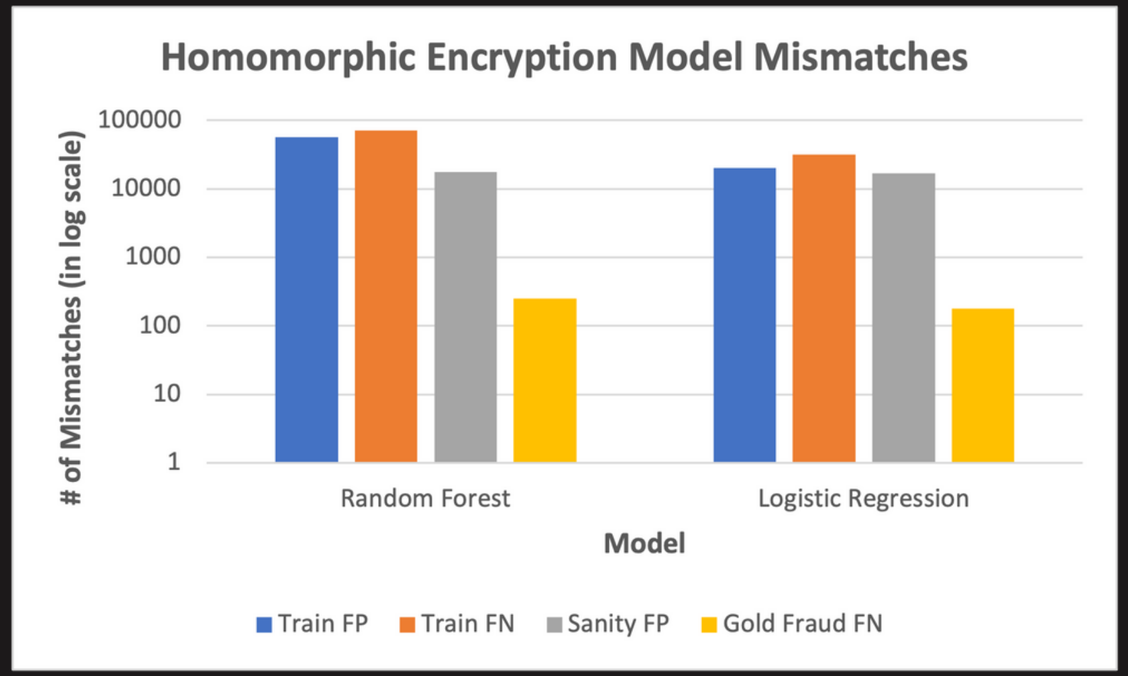


- Visualizing Privacy budget alongside F1-Score for Logistic Regression model.
- Smaller values of Epsilon provides stronger privacy guarantee but may reduce the model accuracy.

Best Eval F1 Score = 0.919
For Epsilon = 20

HOMOMORPHIC ENCRYPTION

1. Fully Homomorphic Encryption(FHE)
2. Training is performed on non-encrypted data
3. Quantize the model using Concrete-ML
4. Convert/Compile the model to its FHE Equivalent
5. Inference done through server-side encryption
6. Fraud Prediction decrypted at client-side



HOMOMORPHIC ENCRYPTION

The benefits:

1. Enables secure sharing and analysis of private data without privacy compromise
2. Allows mathematical operations on encrypted data without revealing the actual data
3. Random noise is added before encryption for enhanced security
4. Noise grows with each operation, potentially overflowing on the actual data
5. Categorized based on mathematical computations on the ciphertext's type and frequency

The Types:

	Actions	Number of Operations
Partially Homomorphic Encryption (PHE)	One (Addition or multiplication)	Unlimited
Somewhat Homomorphic Encryption (SHE)	Two (Addition and multiplication)	Limited
Fully Homomorphic Encryption (FHE)	Two (Addition and multiplication)	Unlimited

CONCRETE-ML FRAMEWORK

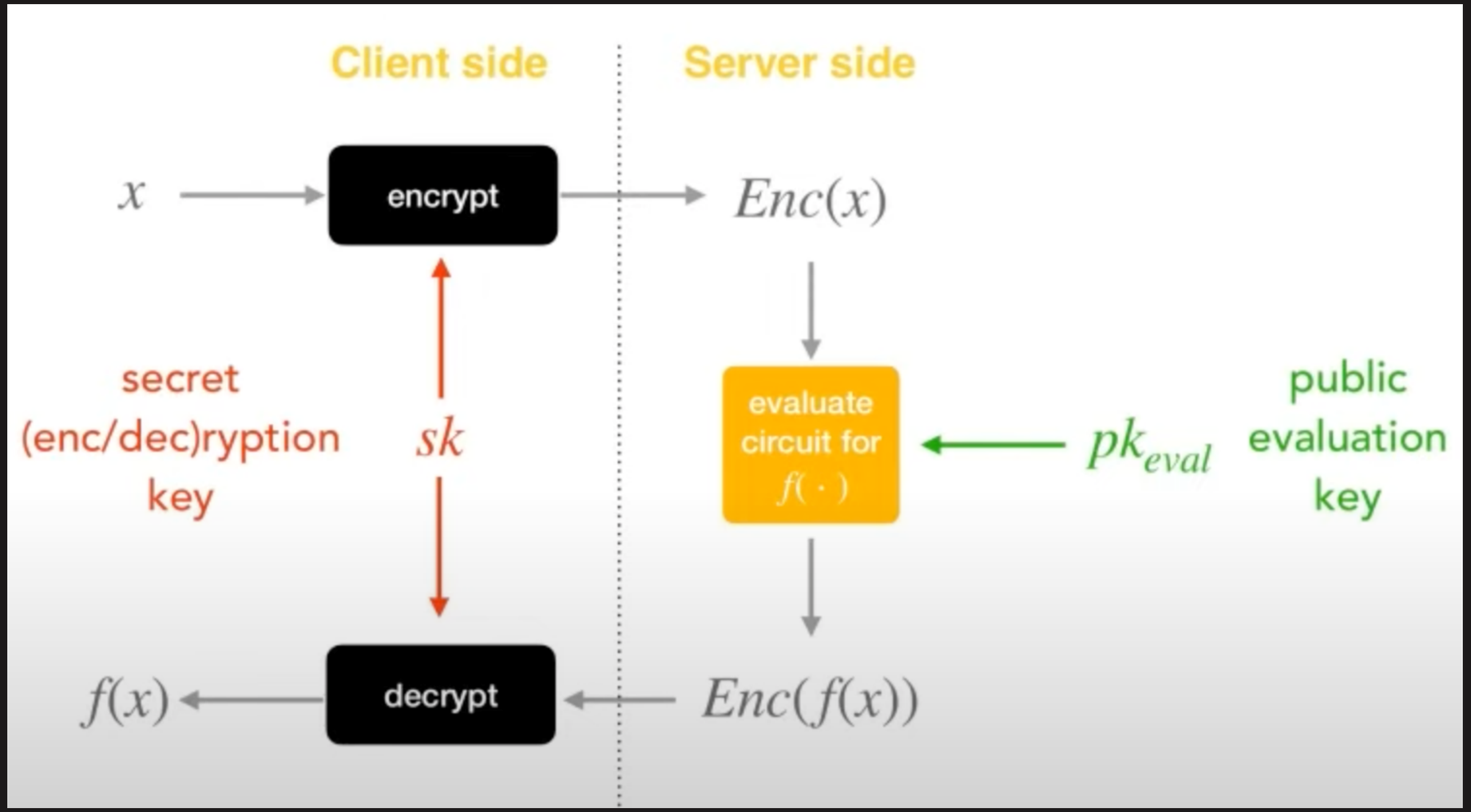
Concrete ML is a Privacy-Preserving Machine Learning (PPML) open-source set of tools built on top of Concrete by Zama.

It aims to simplify the use of fully homomorphic encryption (FHE).

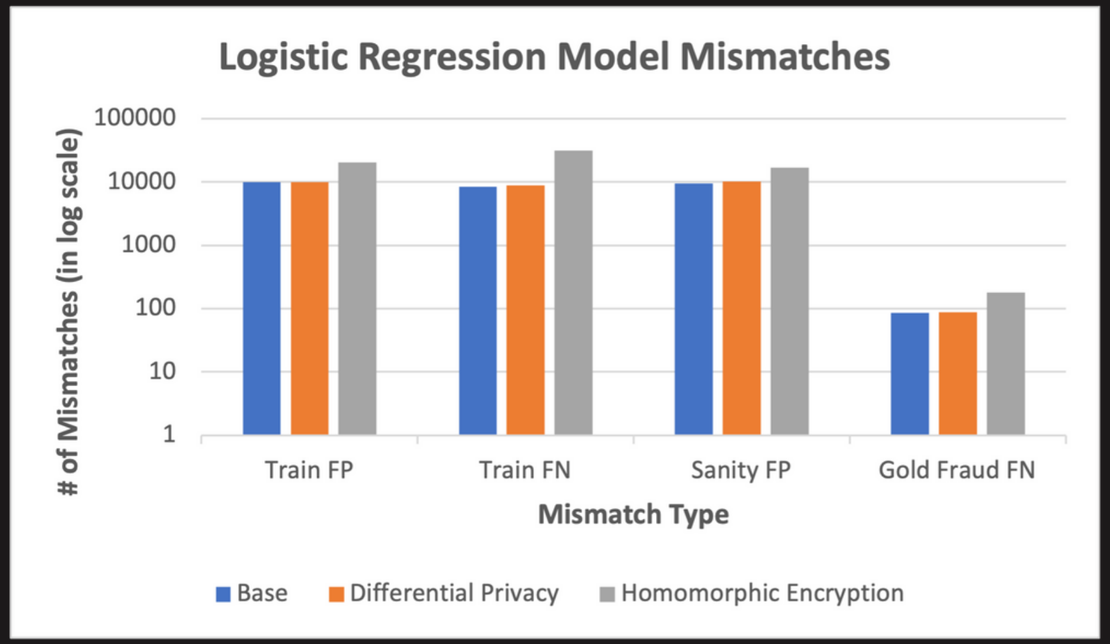
It features a variant of TFHE supporting leveled and fast bootstrapped operations

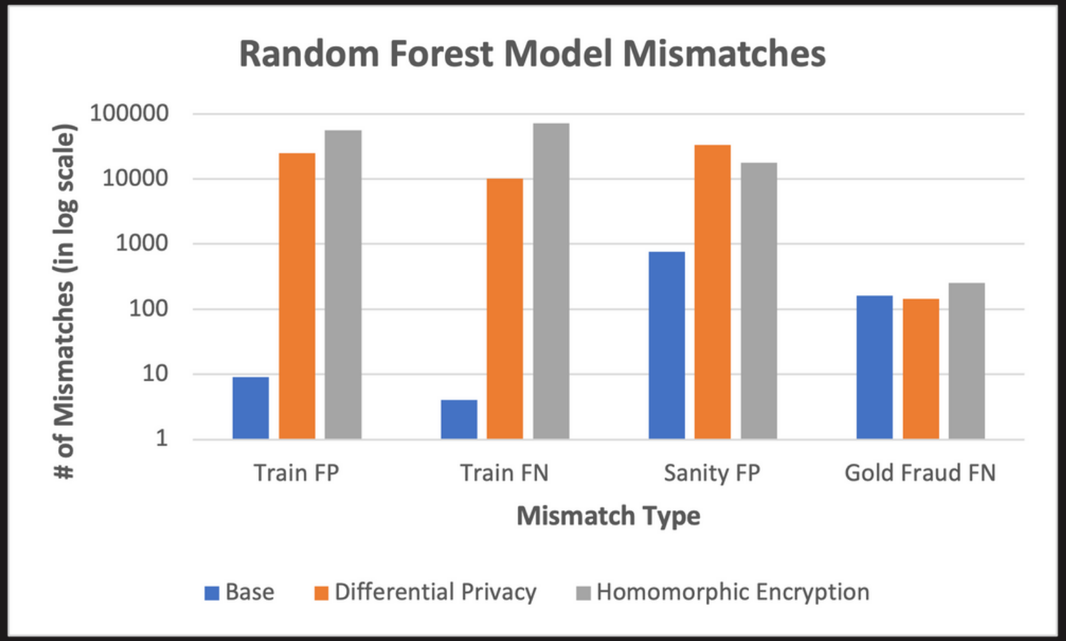
	BGV	CKKS	TFHE-LIB	TFHE-CONCRETE
Operations	Leveled	Leveled	Bootstrapped	Leveled + Bootstrapped
Non-linear functions	Approximate	Approximate	Exact	Exact or Approximate
Data Types	Integers	Reals	Boolean	Boolean + Integers

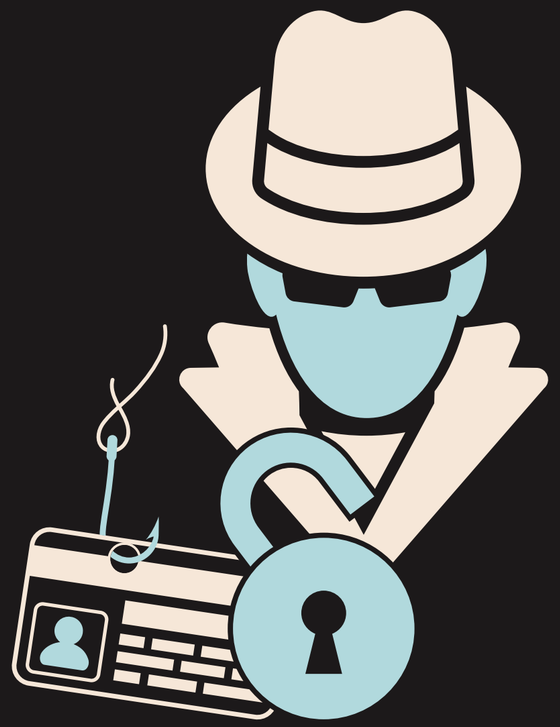
FHE


Client side
Server side
x
encrypt
$Enc(x)$
secret
(enc/dec)ryption
key
sk
evaluate
circuit for
$f(\cdot)$
pk_{eval}
public
evaluation
key
$f(x)$
decrypt
$Enc(f(x))$

RESULTS


Logistic Regression Model Mismatches
# of Mismatches (in log scale)
100000
10000
1000
100
10
1
Train FP
Train FN
Sanity FP
Gold Fraud FN
Mismatch Type
Base
Differential Privacy
Homomorphic Encryption


Random Forest Model Mismatches
# of Mismatches (in log scale)
100000
10000
1000
100
10
1
Train FP
Train FN
Sanity FP
Gold Fraud FN
Mismatch Type
Base
Differential Privacy
Homomorphic Encryption



FUTURE SCOPE

1. SMPC, Federated learning
2. More training for Bert models
3. Better quality labels

THINK BEFORE YOU CLICK!

PROTECT YOURSELF FROM FRAUD EMAILS

Don't share your personal information online!