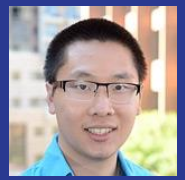
"SECRETS ARE LIES, SHARING IS CARING, PRIVACY IS THEFT."

Dave Eggers, “The Circle”

An Introduction to Machine Learning Privacy

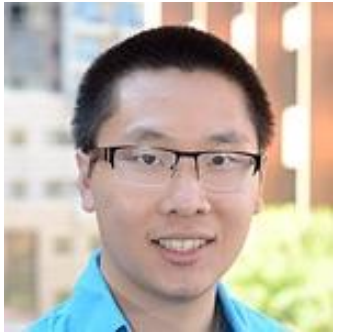



Who are we?



Nahid Farhady

Software Developer



Vincent Pham

Software Developer

Outline

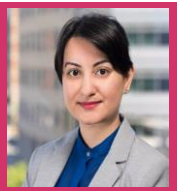
- Introduction
- Privacy attacks
- Defenses
- Demo





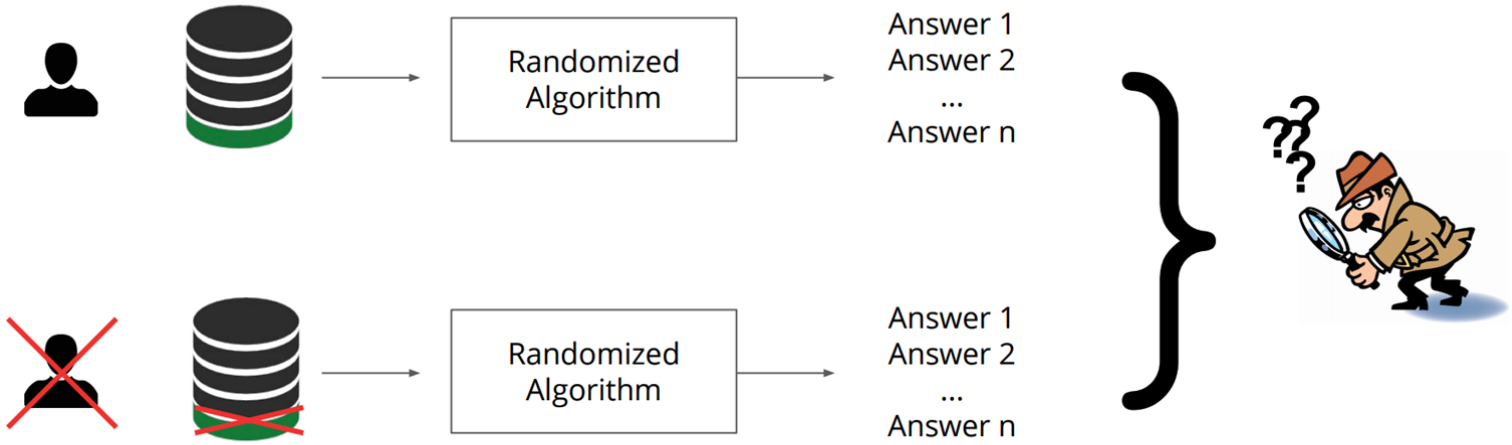
Motivation

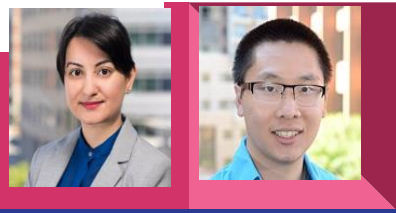
- Machine Learning needs data.
- What if data has sensitive information?
  - Medical data, financial data, etc
- Even if we trust the algorithms running the model and the model developers, sometimes output of the data and statistics of the output can reveal information.
- The threat model can be blackbox or whitebox:
  - Adversary can make a potentially unbounded number of queries
  - Adversary has access to model internals



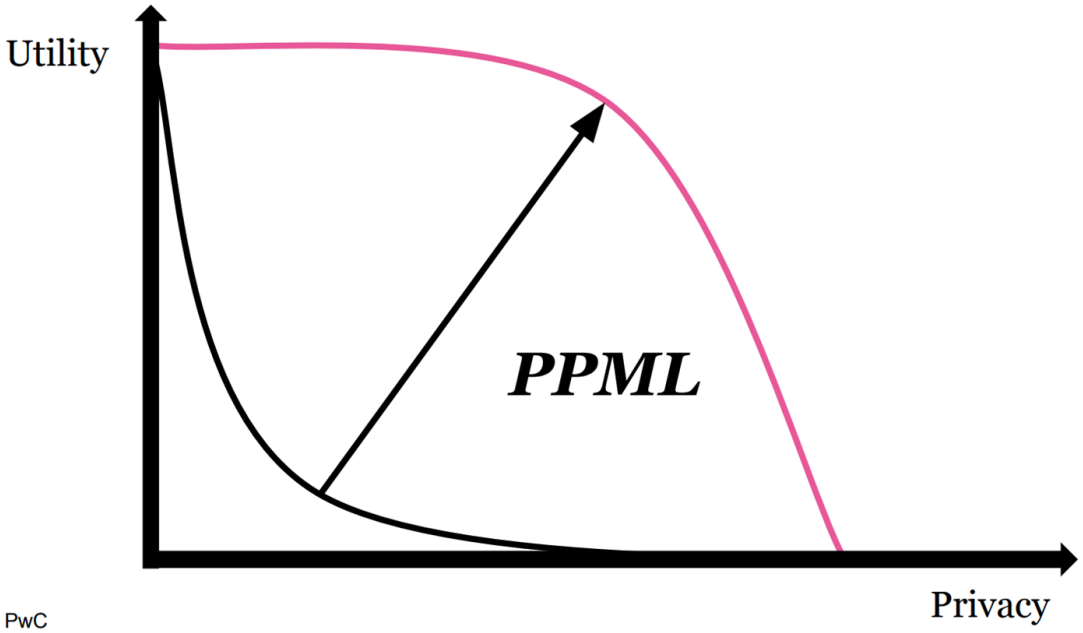


What is Privacy?

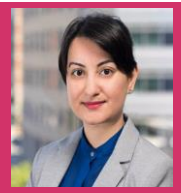




Privacy vs Utility Tradeoff



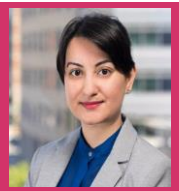
PwC





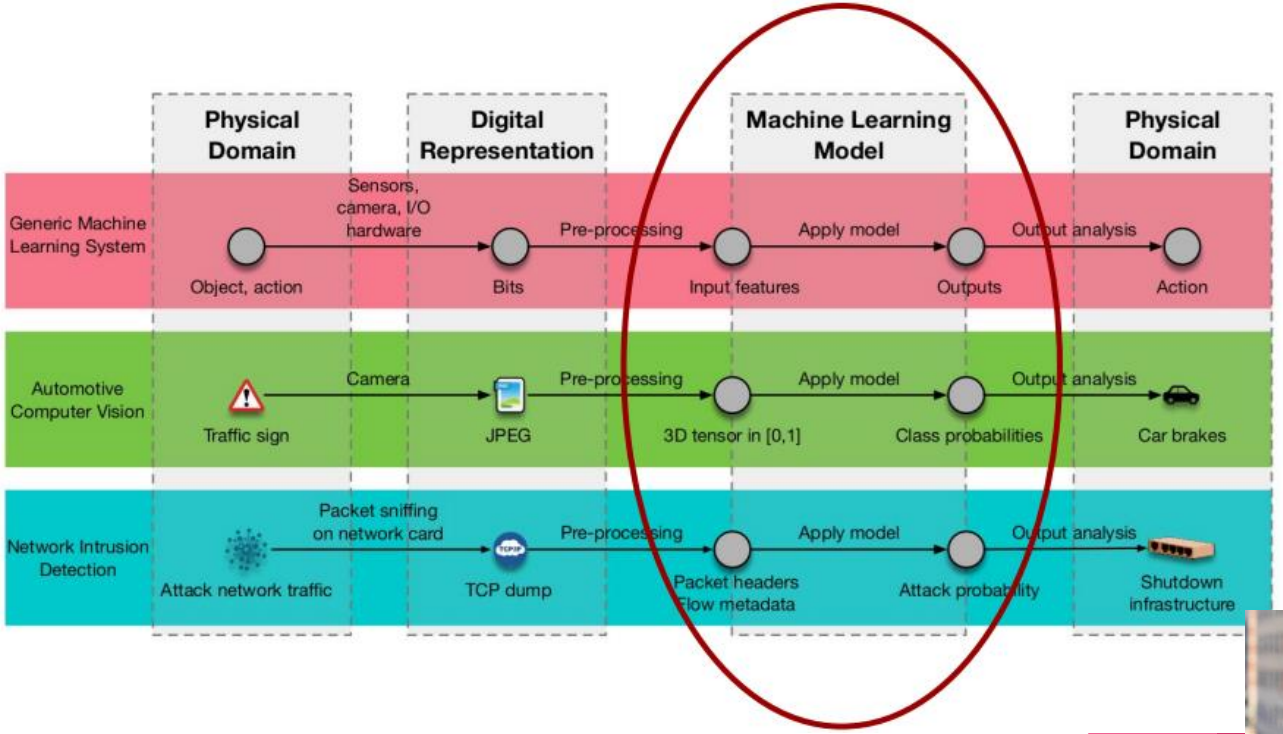
Privacy Attacks

- Classical
  - Private data in training
- ML enabled
  - Reconstruction attacks
  - Linkage
  - De-anonymization
  - Model inversion
  - Membership inference

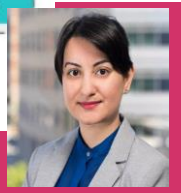




The Attack Surface



Reference: https://project.inria.fr/wifs2017/files/2017/12/WIFS_T2_Papernot.pdf

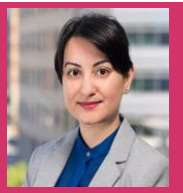




Linkage Attack

	HOSPITAL DATA SET (Found In Data Set)	VOTER DATA SET (Found in Data Set)	NON-VOTERS (In Population)
1	Not in Hospital Data	Male 1/1/1945 02138 — Can't Re-identify (No Match)	
2	Male 1/2/1945 02138	Not in Voter Data	Male 1/2/1945 02138 — Can't Re-identify (No Match)
3	Male 1/3/1945 02138; Male 1/3/1945 02138	Male 1/3/1945 02138 — Can't Re-identify (> 1 Match)	
4	Male 1/4/1945 02138 — Can't Re-identify (> 1 Match)	Male 1/4/1945 02138; Male 1/4/1945 02138	
5	Male 1/5/1945 02138 — *Presumed* Re-identification (Has Only 50% Chance of Being a Correct Match)	Male 1/5/1945 02138	Male 1/5/1945 02138 — Directly Protected From Re-identification
6	Male 1/6/1945 02138 — *Correct Re-identification*	Male 1/6/1945 02138	

Identification of the Governor of Massachusetts health records by combining with voter information





Linkage Attack

5-digit ZIP code, Birthdate, and Gender uniquely identify **87%** of the population in the U.S.

- Latanya Sweeney on 1990 Census





Linkage Attack

5-digit ZIP code, Birthdate, and Gender uniquely identify **64%** of the population in the U.S.

- Philippe Golle on the 2000 Census





Linkage Attack

NETFLIX

Date
Ratings: 1-5

IMDb

Name
Date
Ratings





De-anonymization Attack (Netflix Example)



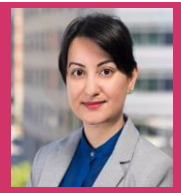
- “Power and Terror: Noam Chomsky in Our Times” and “Fahrenheit 9/11”



- “Jesus of Nazareth” and “Gospel of John”

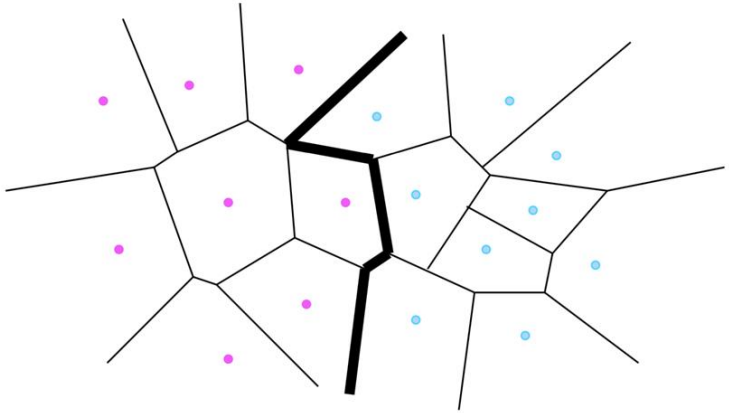


- “Super Size Me”

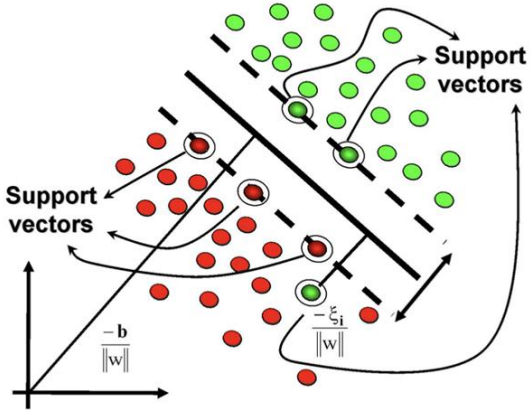




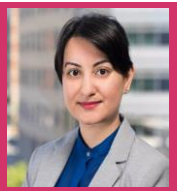
Reconstruction Attack



Nearest neighbor classifier

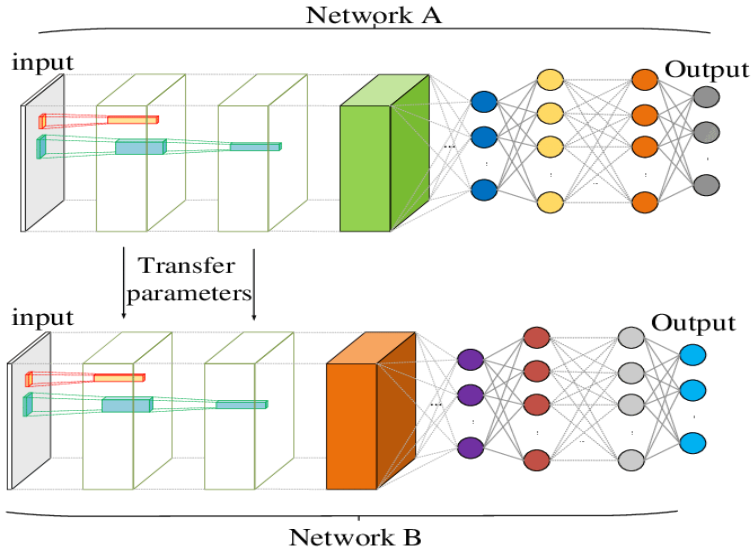


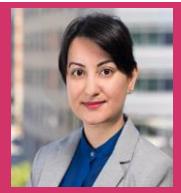
Kernel-based SVM





Reconstruction Attack







Model Inversion Attack

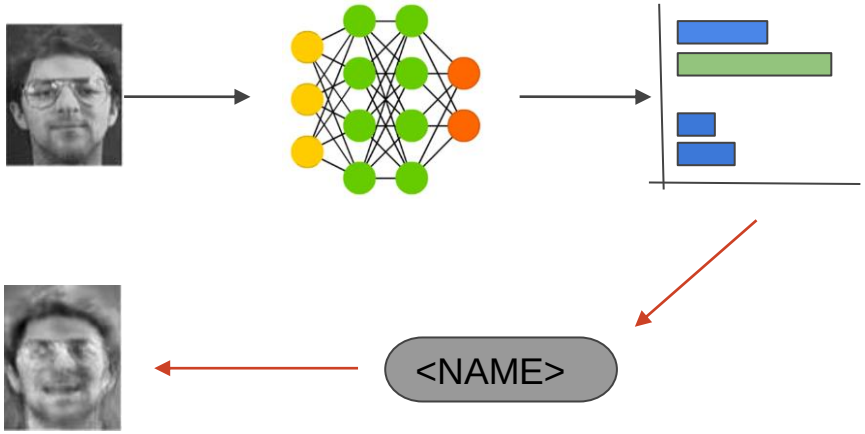
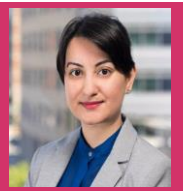


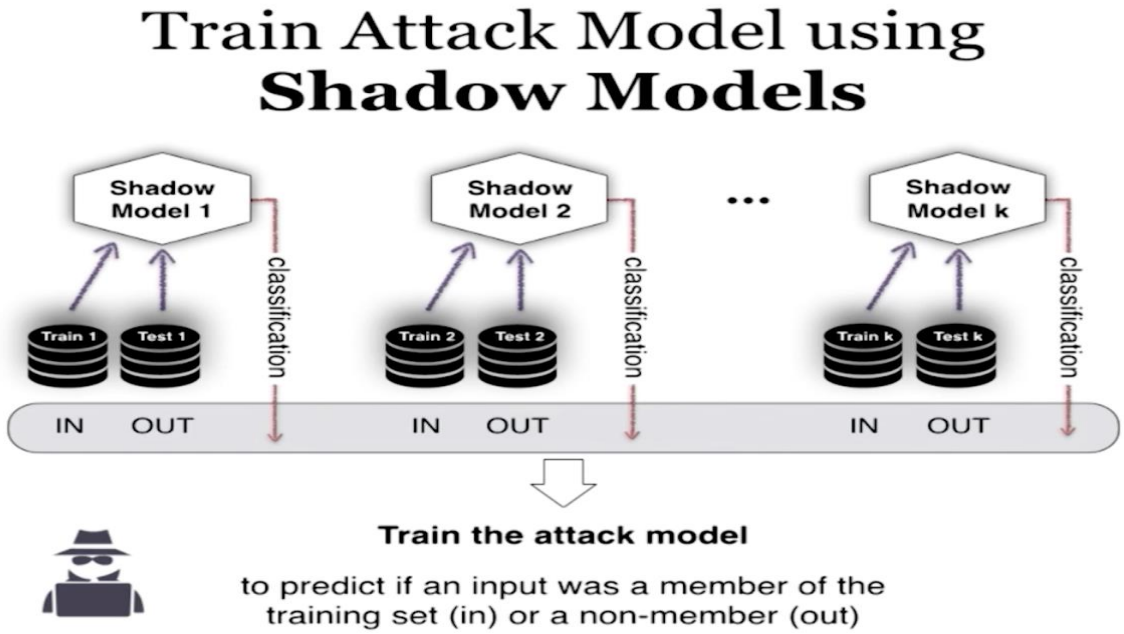


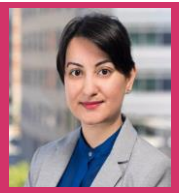
Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.





Membership Inference Attack







Query Attack

Google

Social security

social security
social security **card**
social security **disability**
social security **calculator**
social security **office richmond va**
social security **login**
social security **benefits**
social security **act**
social security **office richmond**
social security **for John Smith is 555-55-5555**

Google Search | I'm Feeling Lucky

Report inappropriate predictions



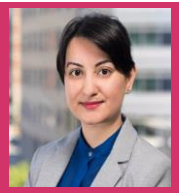


Privacy Attack Comparison

Approach	Main Attack Vector	Gist	Use Case Example
Linkage	Input	Combine two or more data source to determine sensitive information	A company's data is leaked with anonymized fields which attacker combines with an open dataset to infer identity.
De-anonymization	Input, Computation, Result	Reveal a hidden information from the user or even the identity of the user.	A person has preferences tied to a data on a site they believed was secured, but attackers finds a loophole to this info.
Reconstruction	Computation	Obtain raw features stored in the model.	Models such as KNN and SVM stores the raw features in the model used for prediction, thus an attacker with full access to the model can obtain the input features.
Model Inversion	Result	Use model's predictions to obtain the input features.	An API is publically available that returns probabilities, attackers can train a new model on the probability to obtain a best guess of the original features.
Membership Inference	Result	Generate multiple models to train an attack model that determines whether a record was used to train the production model.	A person with known cancer is identified as having been trained by the model which attacker's can infer additional attributes on the person.
Query	Computation	Takes advantage of over-memorized models to obtain sensitive information.	An attacker may brute force through autocomplete functions on sensitive keywords.

Privacy Preserving ML

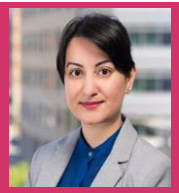
- Training Data Privacy
  - The guarantee that a malicious actor will not be able to reverse-engineer the training data.
- Input Privacy
  - The guarantee that a user's input data cannot be observed by other parties, including the model creator.
- Output Privacy
  - The guarantee that the output of a model is not visible by anyone except for the user whose data is being inferred upon.
- Model Privacy
  - The guarantee that the model cannot be stolen by a malicious party.





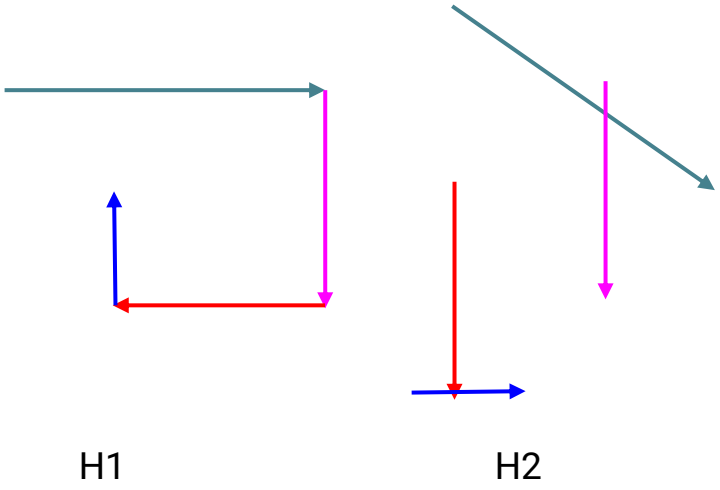
Privacy Aware Machine Learning

- Secure computation
  - Homomorphic encryption
  - Secure multi-party computation
- Privacy preservation
  - Federated machine learning
  - Differential privacy methods
  - PATE

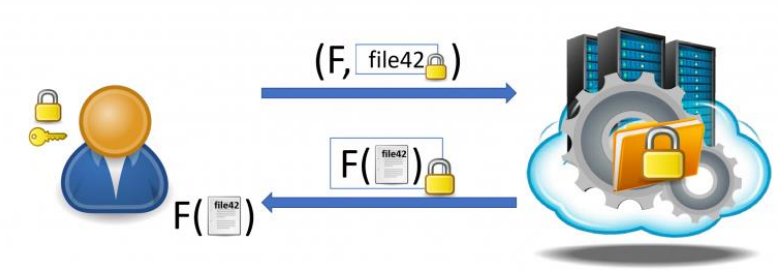




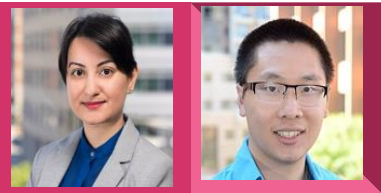
Homomorphic Encryption



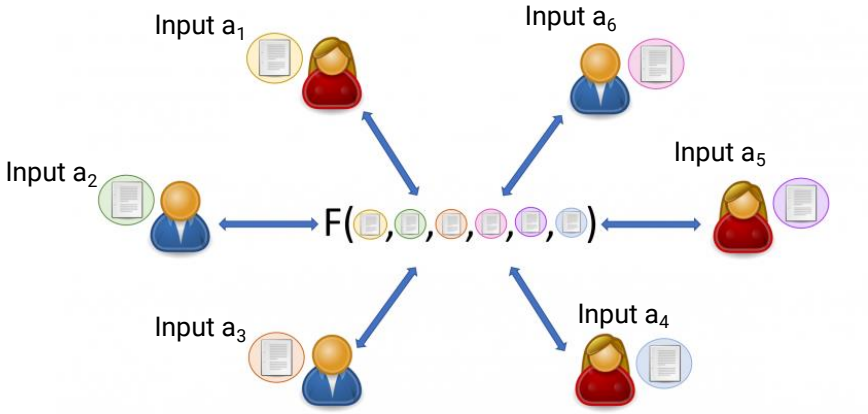
operation(plain)
==
decrypt(operation(encrypt(plain)))



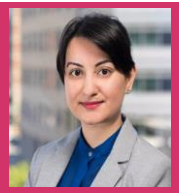
- H1 and H2 have a different shape
- They are not very comparable



Secure MultiParty Computation

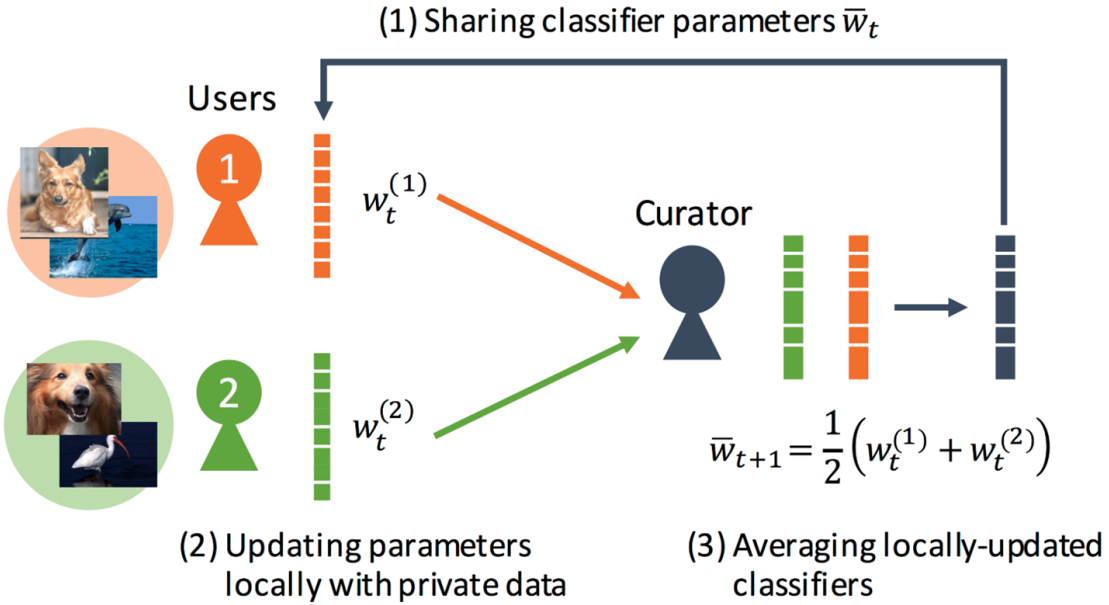


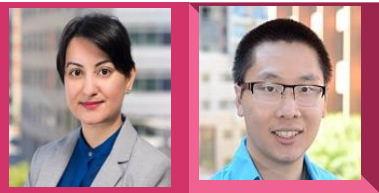
The purpose is to compute some value for $F(a_1,a_2,a_3,a_4,a_5,a_6)$





Federated Machine Learning

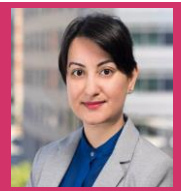




Differential Privacy

Data +  → Algorithm → Outcome

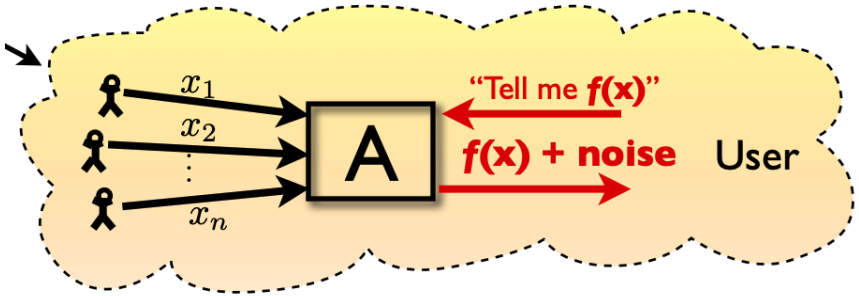
Data +  → Algorithm → Outcome

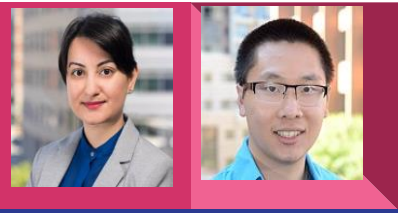




Differential Privacy

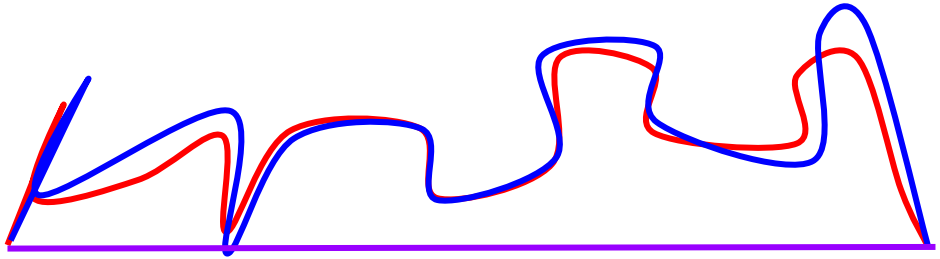
- May be interactive:
  - Non-interactive: release pre-defined summary stats + noise
  - Interactive: respond to user requests





Differential Privacy

What we want is a protocol that has a probability distribution over outputs:



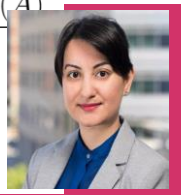
- such that if person i changed their input from X_i to any other allowed X'_i, the relative probabilities of any output do not change by much.
- This would effectively allow that person to pretend their input was any other value they wanted.

Bayes rule: $\frac{\Pr(x_i|output)}{\Pr(x_i'|output)} = \frac{\Pr(output|x_i)}{\Pr(output|x_i')} \cdot \frac{\Pr(x_i)}{\Pr(x_i')}$

Formula

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events
$P(A|B)$ = probability of A given B is true
$P(B|A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B





Differential Privacy

It's a property of a protocol A which you run on some dataset X producing some output A(X).

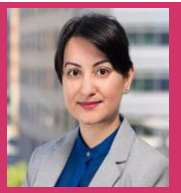
A is α-differentially private if for any two neighbor datasets S, S' (differ in just one element $x_i \rightarrow x_i'$),

For all outcomes V

$$e^{-\alpha} \leq Pr(A(S)=v)/Pr(A(S')=v) \leq e^{\alpha}$$

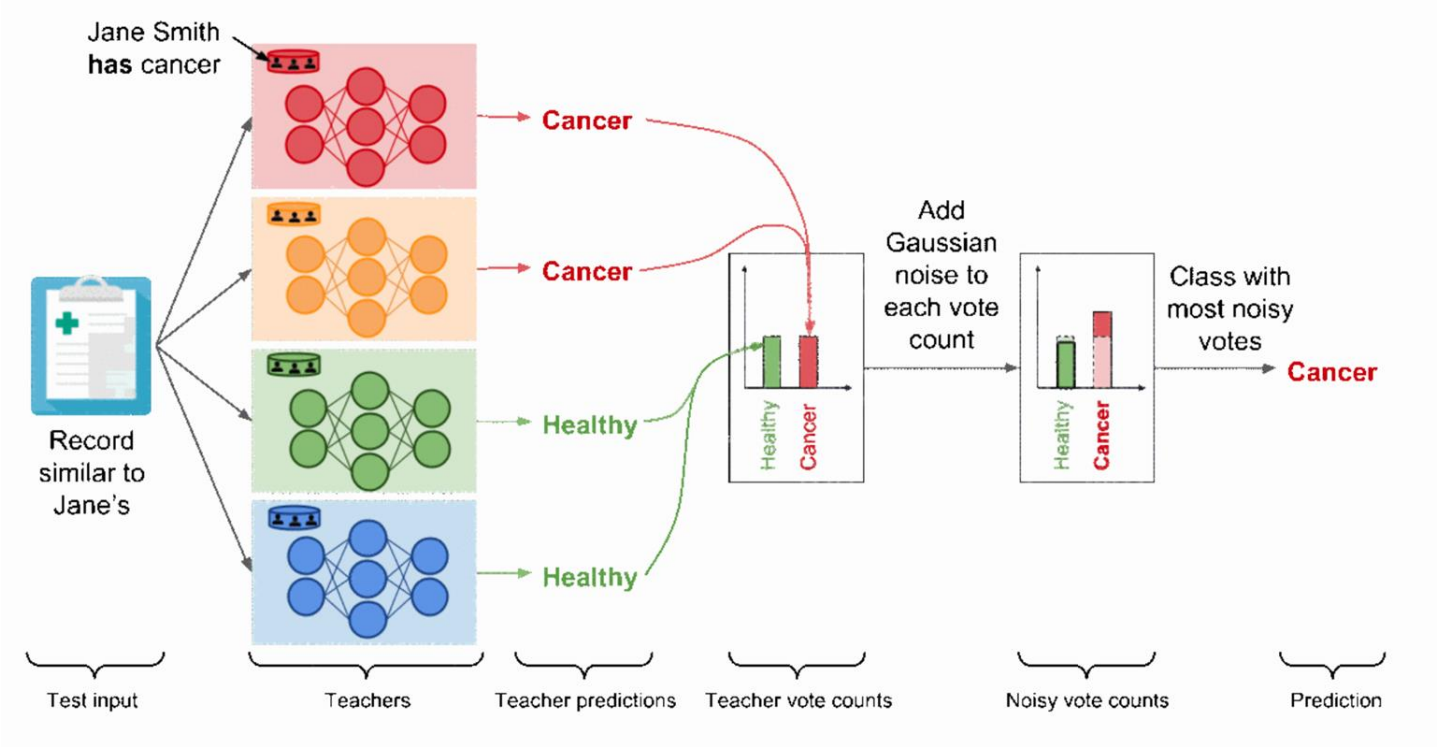
X_1
X_2
X_3
X_4
X_5

X_1
X_2
X'_3
X_4
X_5





Private Aggregation of Teacher Ensembels



Reference:http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html

Defense method comparison

Approach	Protection	Emphasis	Gist	Common use case	Practical use case
Federated Learning	Data owner	Privacy	Put model to customer	Centralized model on sensitive, distributed data	Android autocomplete
Differential Privacy	Data owner	Privacy	Inject noise	Proovable privacy-utility trade-off	Chrome usage statistics
Multi-Party Computation	Model owner	Security	Build up model without sharing data	Secure aggregation	Millionaire's problem
Homomorphic Encryption	Model owner	Security	Save mathematical properties while encrypting	Outsource model design	Numeraire
Blockchain	Attribution	Incentive	Secure compensation for contributions	Build very large model, attribute fairly	Self-driving cars

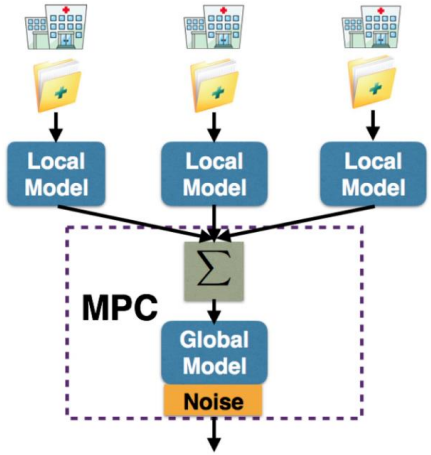
Reference: Privacy-Preserving Machine Learning - Silver Bullet for PwC?

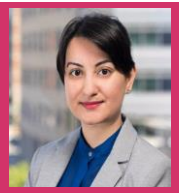




Combinatory Methods

- Homomorphic Encryption + Differential Privacy
- Secure Multiparty Computation + Differential Privacy
- Federated Learning + Differential Privacy + Secure Multiparty
- Homomorphic Encryption + PATE
- Secure Multiparty Computation + PATE
- Federated Learning + PATE + Homomorphic Encryption







Practical Methods for Privacy Preservation

- TensorFlow (TF-federated, TF-private, TF-encrypted)
  - https://github.com/tensorflow/privacy
- PyTorch (Openmined / PySyft)
  - https://github.com/OpenMined/PySyft
- IBM Privacy Package
  - https://github.com/IBM/differential-privacy-library

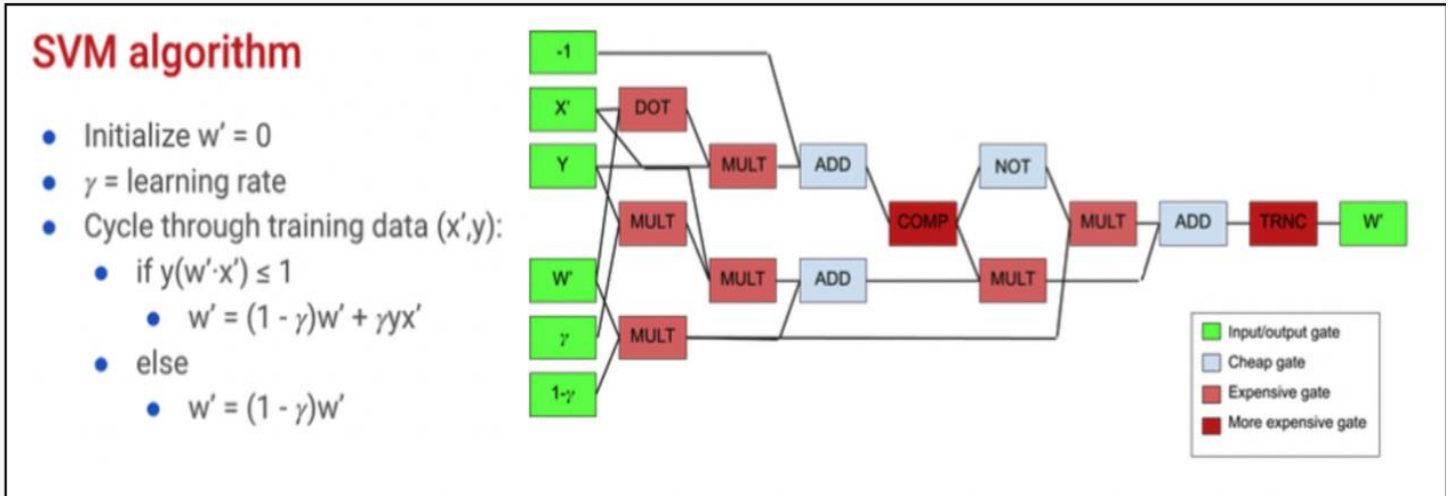


Questions?

Please reach out with questions

- Vincent.pham@gmail.com
- Nahid.farhadyghalaty@capitalone.com

MPC on SVM



Reference: https://blog.trailofbits.com/2019/10/04/multi-party-computation-on-machine-learning/

Mult-key Homomorphic Encryption (MKHE)

