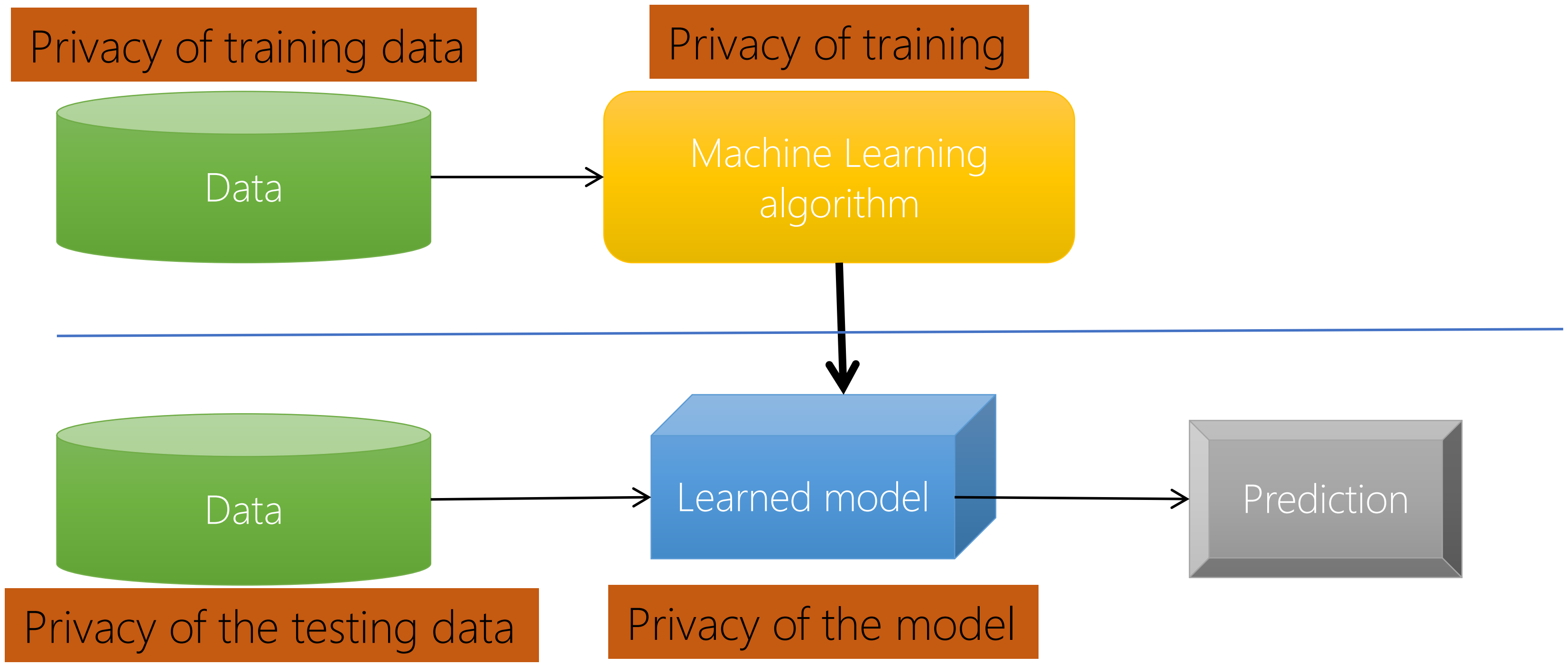
UNIVERSITY of VIRGINIA

AAAI-23 Bridge: AI & Law

Tianhao Wang
University of Virginia

Privacy Issues in Machine Learning



Privacy of Testing Data: Attribute Inference

Attacker Goal: Extract private inputs by leveraging the outputs and ML model.

Example: 538 Steak Survey on BigML.com

The model $f(\boxed{x_1, \ldots,} \boxed{x_n}) = \boxed{y}$

Household income
Whether person gambles
Whether cheated on significant other

Prediction of how person likes steak prepared:
- rare
- medium-rare
- medium
- medium-well
- well-done

Plus confidence value
Normalized vector of class confidences each in [0,1]

Attribute Inference Attack

Evaluate f with x_n=0 and x_n=1

Return x_n that gives y

Also consider likelihood

The model $f(x_1, \ldots, x_n) = y$

Prediction of how person likes steak prepared:
- rare
- medium-rare
- medium
- medium-well
- well-done

Plus confidence value

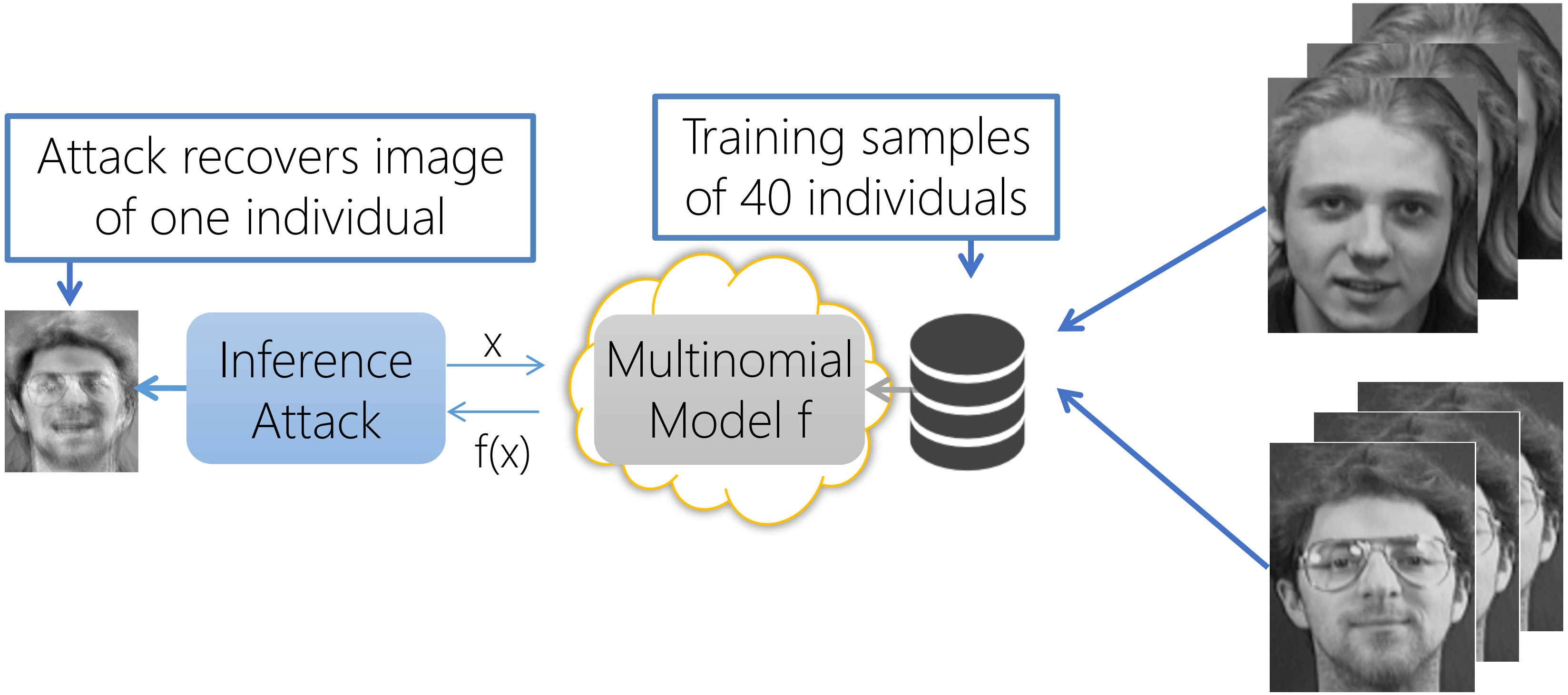
Household income
Whether person gambles
Whether cheated on significant other

[Fredrikson, Lantz, Jha, Lin, Page, Ristenpart 2014]

Extraction Attack

$f(x_1, \ldots, x_n) = [p_{Bob}, \ldots, p_{Jake}]$

Given y, infer $x_1, \ldots, x_n$ assuming they are all unknowns



Search for x that maximizes p_{Bob} using gradient descent

[Fredrikson, Jha, Ristenpart 2015]

Training Data: Membership Inference

Goal: Infer whether x is used to train the model.

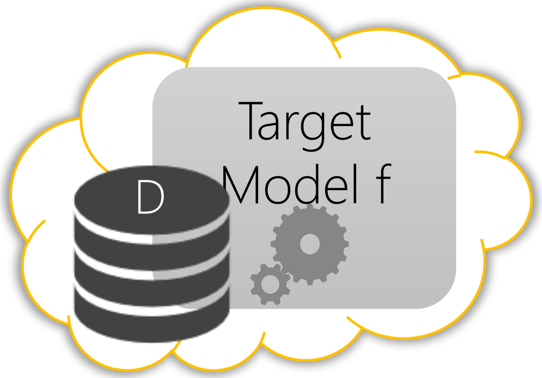
Why: what is the privacy concern?

Assume f can predict cancer-related health outcomes.

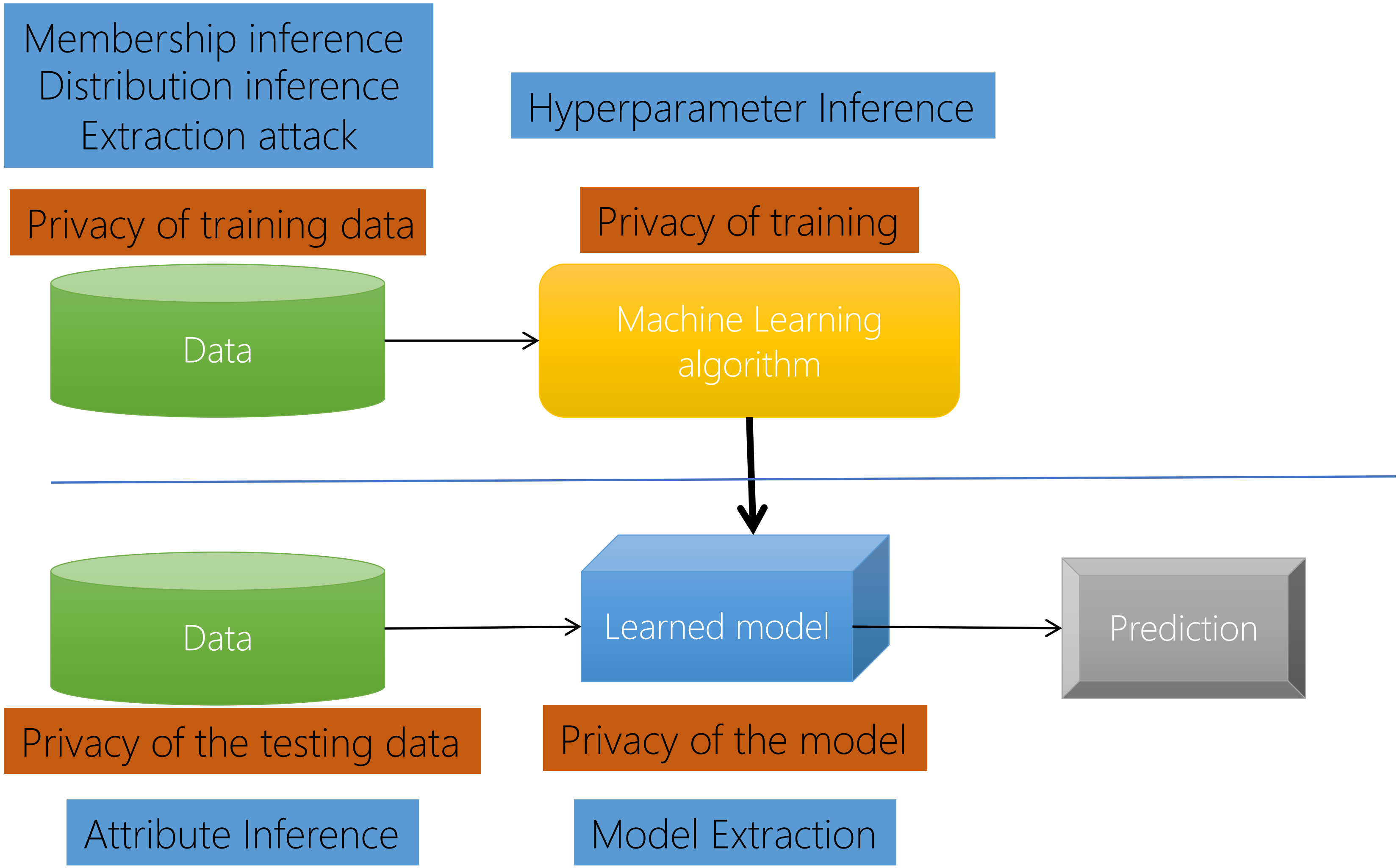
If x is used to train f, x may have health issues.

How?

By observing the behavior of f. *Statistical-based* or *shadow models* + *meta-classifier*.



Privacy Issues in Machine Learning



Defenses against Privacy Attacks

Defense strategies against privacy attacks in ML can be broadly classified into:

- Heuristic-based solutions:
  - ML-specific techniques
  - Anonymization
  - Distributed learning
- Privacy-as-control:
  - Encryption techniques
  - Unlearning
- Quantify disclosure:
  - Differential privacy

ML-Specific Defenses

Overfitting is one of the reasons for information leakage

- Dropout
- Early stopping
- Removing outliers
- Weight smoothing

Anonymization

Removing identifying information in the data

The remaining information in the data can be used for identifying the individual data instances: 87% of all Americans can be uniquely identified using 3 bits of information: ZIP code, birth date, and gender

Information for the Governor of Massachusetts identified from information released by an insurance group

User ID	Name	Address	Account Type	Subscription Date
001	Alice	123 A St	Pro	01/02/20
002	Bob	234 B St	Free	02/03/21
003	Charlie	456 C St	Pro	03/04/18

User ID	Name	Address	Account Type	Subscription Date
001			Pro	01/02/20
002			Free	02/03/21
003			Pro	03/04/18

Homomorphic Encryption (HE)

HE allows computations on encrypted data (without decrypting it)

- Encrypted data can be analyzed and manipulated without revealing the original data

HE encrypts the data, and applies an algebraic system (e.g., additions and multiplications) to allow computations while the data is still encrypted

- Only the person who has a matching key can access the decrypted results

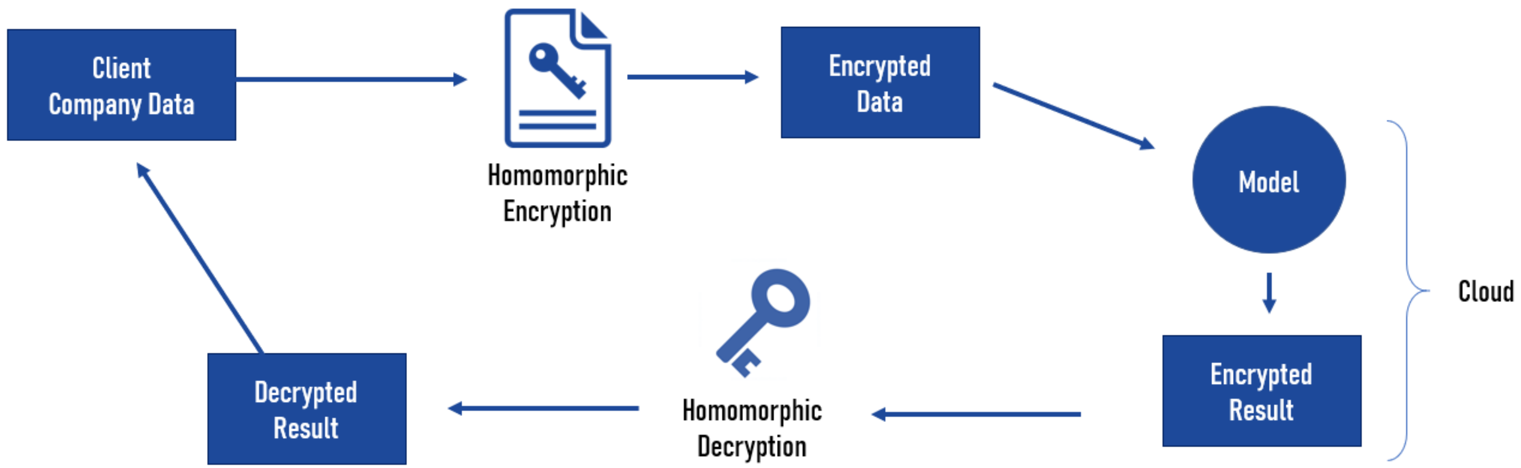


Figure form: [Homomorphic Encryption & Machine Learning: New Business Models]

Secure Multi-Party Computation

MPC allows two or more parties to jointly perform computation over their private data, without sharing the data

E.g., two banks want to know if they have both flagged the same individuals and learn about the activities by those individuals

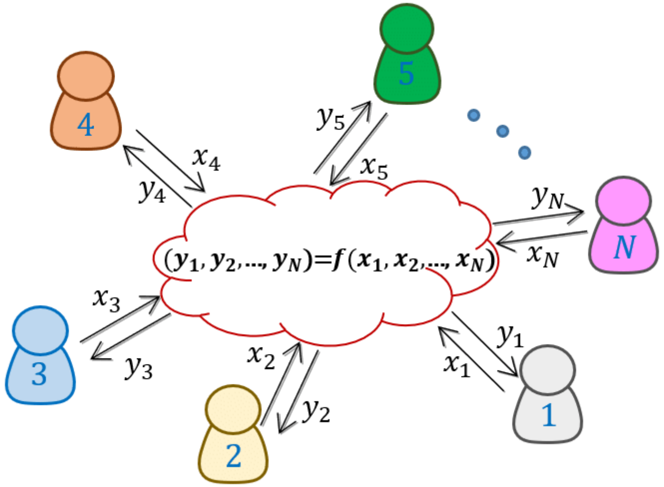


Figure form: Generation and Distribution of Quantum Oblivious Keys for Secure Multiparty Computation

Quantify Leakage

Allowing analysts to learn about *trends* in data, without revealing information specific to *individual data instances*

- Involves an intentional release of information, and attempt to control what can be learned from the released information

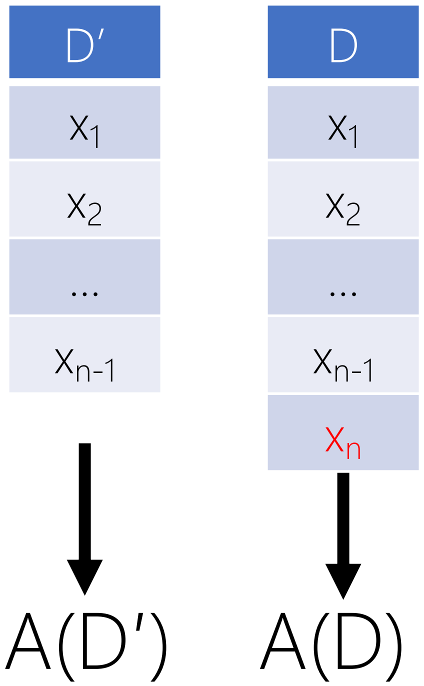
Related to data privacy is the *Fundamental Law of Information Recovery*

- *overly accurate estimates of too many statistics can completely destroy privacy*

There is an inevitable trade-off between privacy and accuracy/utility

Differential Privacy

Idea: Any output should be about as likely regardless of whether I am in the dataset or not. For each individual, the world after removing the individual's data is an ideal world of privacy for that individual.



Algorithm A satisfies ε-differential privacy if for any pair D and D′ that differ in one record, for possible output y,

$$e^{-\varepsilon} \leq \frac{\Pr[A(D)=y]}{\Pr[A(D')=y]} \leq e^{\varepsilon}$$

Parameter ε: strength of privacy protection, known as privacy budget. Goal is to simulate all these ideal worlds.

Smaller ε -> Stronger Privacy

Differential Privacy

- Obfuscation mechanisms for privacy protection
  - A randomization mechanism $\mathcal{M}(D)$ applies noise ξ to the outputs of a function $f(D)$ to protect the privacy of individual data instances, i.e., $\mathcal{M}(D) = f(D) + \xi$
    - Applies to SGD
    - Train a generative model
- Examples include:
  - 2014: Google's RAPPOR, for statistics on unwanted software hijacking users' settings
  - 2015: Google, for sharing historical traffic statistics
  - 2016: Apple, for improving its Intelligent personal assistant technology
  - 2017: Microsoft, for telemetry in Windows
  - 2020: LinkedIn, for advertiser queries
  - 2022: U.S. Census Bureau, for demographic data

Efforts Bridging Privacy and Law

User (experts and laypeople) perception/understanding of privacy risks and differential privacy

Discussion (among big companies and regulators) on how to set epsilon in differential privacy

Formalizing regulations (taken from e.g., GDPR) in computer science