Membership Inference Attacks against Classifiers

Ninghui Li

Department of Computer Science

Purdue University

Privacy Concerns regarding Machine Learning Models

- Secrecy of the model
  - Model parameters, hyperparameters in training, etc.
- Privacy of the training data
  - Membership inference attacks
    - Given a target model and a target instance, output whether the instance has been used to train the target model.
  - Attribute inference / instance reconstruction attacks
    - Knowing some attributes of an instance in training, infer unknown attributes
  - Representative instance reconstruction
    - Reconstruct representative instance of a class

Outline

- Why Study Membership Inference
- Differential Privacy
- Membership privacy and differential privacy
- Formulations of Membership Inference Attacks
- What We Currently Know about MI Attacks

Why Care about Membership Inference?

- Privacy incidents demonstrate
  - Privacy violation = positive membership disclosure

- No membership disclosure means no re-identification disclosure

Defining Privacy is Hard

- Lots of data privacy notions
  - E.g., k anonymity, l diversity, t closeness, and many others
- Why defining privacy is hard?
  - Difficult to agree on adversary goal.
  - Difficult to agree on adversary prior knowledge.
  - Too strong , then not achievable.
  - Too weak, then not enough.
  - Foremost, privacy is a complex social, legal, and moral concept

Privacy Incident 1: Netflix Movie Rating Data

- In 2006, Netflix released anonymized movie rating data for its Netflix prize challenge
  - “Anonymized data” includes date and value of movie ratings
- Knowing 6-8 approximate movie ratings and dates is able to uniquely identify a record with over 90% probability
  - Correlating with a set of 50 users from imdb.com yields two records
- Netflix cancels second phase of the challenge, was sued for privacy violation and settled the lawsuit

Arvind Narayanan, Vitaly Shmatikov: ***Robust De-anonymization of Large Sparse Datasets.*** *IEEE Symposium on Security and Privacy 2008: 111-125*

Re-identification occurs! Re-identification implies membership disclosure.

Privacy Incidence 2: Genome-Wide Association Study (GWAS)

- A typical study examines thousands of singe-nucleotide polymorphism locations (SNPs) in a given population of patients for statistical links to a disease.
- From aggregated statistics, one individual's genome, and knowledge of SNP frequency in background population, one can infer participation in the study.
  - The frequency of every SNP gives a very noisy signal of participation;
  - Combining thousands of such signals gives high-confidence prediction

N. Homer, et al. ***Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SND genotyping microarrays***. *PLoS Genet*, 4(8):e1000167+, 2008.

GWAS Privacy Issue

Published Data

	Disease Group Avg	Control Group Avg
SNP1=A	43%	…
SNP2=A	11%	…
SNP3=A	58%	…
SNP4=A	23%	…
…		

Adv. Info & Inference

Population Avg	Target individual Info	Target in Disease Group
42%	yes	+
10%	no	-
59%	no	+
24%	yes	-

Membership disclosure occurs!

Outline

- Why Study Membership Inference?
- Differential Privacy
- Membership privacy and differential privacy
- Formulations of Membership Inference Attacks
- What We Currently Know about MI Attacks

What is Privacy?

It is complicated!

Some concepts from the book “Understanding Privacy” by Daniel J. Solove:

1. the right to be let alone
2. limited access to the self
3. secrecy—the concealment of certain matters from others;
4. control over others' use of information about oneself
5. personhood—the protection of one’s personality, individuality, and dignity;
6. intimacy—control over, or limited access to, one’s intimate relationships or aspects of life.

Formulation of "Privacy as Secrecy"

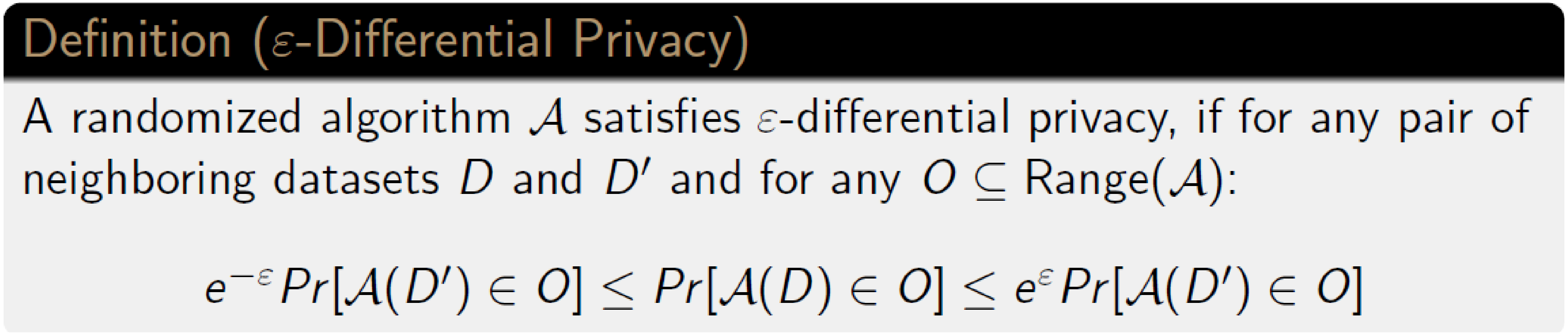
- Dalenius [in 1977] proposes this as privacy notion: "*Access to a statistical database should not enable one to learn anything about an individual that could not be learned without access*."
  - Similar to the notion of semantic security for encryption
  - Not possible if one wants utility!

*T. Dalenius, **Towards a methodology for statistical disclosure control**. Statistik Tidskrift 15, pp. 429–444, 1977.*

Impossibility of “Privacy as Secrecy”: The Smoker Example

- Assume that smoking causes lung cancer is not yet public knowledge, and an organization conducted a study that demonstrates this connection and now wants to publish the results.
- A smoker Carl was not involved in the study, but complains that publishing the result of this study affects his privacy, because others would know that he has a higher chance of getting lung cancer, and as a result he may suffer damages, e.g., his health insurance premium may increase.
- Can Carl legitimately complain about his privacy being violated by pubishing results of the study?

Differential Privacy



Definition (ε-Differential Privacy)

A randomized algorithm $\mathcal{A}$ satisfies ε-differential privacy, if for any pair of neighboring datasets D and D' and for any $O \subseteq \text{Range}(\mathcal{A})$:

$$e^{-\varepsilon} Pr[\mathcal{A}(D') \in O] \leq Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} Pr[\mathcal{A}(D') \in O]$$

Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam D. Smith: ***Calibrating Noise to Sensitivity in Private Data Analysis.*** *TCC 2006: 265-284*

Genius of Idea Behind DP

- Identify an ideal world of privacy for each individual
  - the world where the individual’s data is removed
- Require a mechanism to simulate the ideal world for each individual
- DP does not need to deal with data correlation
- DP simulates the definition that privacy is “control over others' use of information about oneself”

The Personal Data Principle

- Data privacy means giving an individual control over his or her personal data. An individual's privacy is not violated if no personal data about the individual is used.
- Privacy does not mean that no information about the individual is learned, or no harm is done to an individual; enforcing that is infeasible and unreasonable.

Some Caveats of Applying DP

- How neighboring datasets is defined.
- What constitutes an individual’s data: One individual’s data or personal data under one individual’s control
- Group privacy
- Moral challenge
- Choosing epsilon value
- Learning models and applying to individuals

What Constitutes An Individual’s Personal Data?

- Is the genome of my parents, children, sibling, cousins “my personal information”?

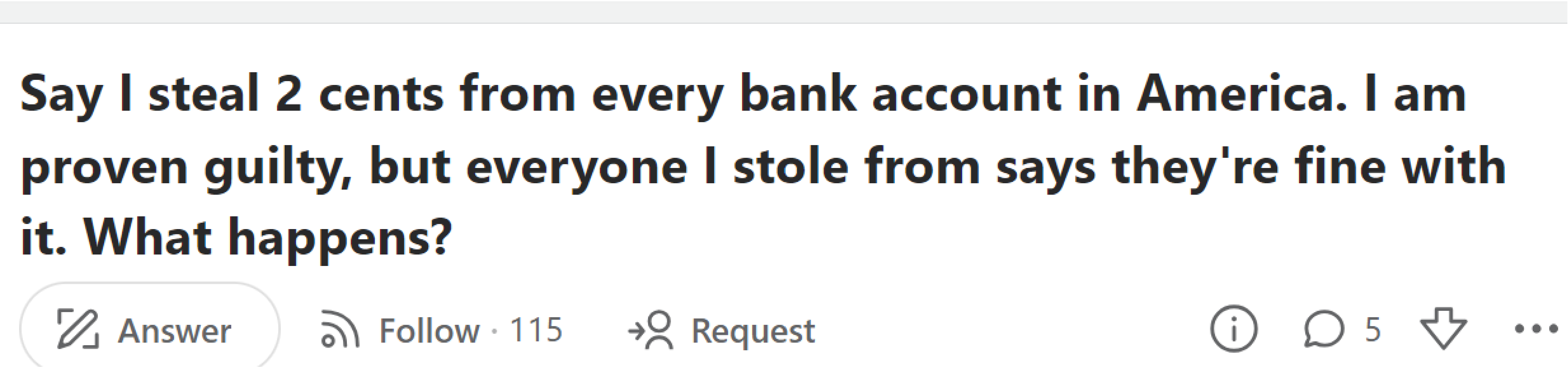
- Example: DeCode Genetics, based in Reykjavík, says it has collected full DNA sequences on 10,000 individuals. And because people on the island are closely related, DeCode says it can now also extrapolate to accurately guess the DNA makeup of nearly all other 320,000 citizens of that country, including those who never participated in its studies.

Such legal and ethical questions still need to be resolved

- Evidences suggest that such privacy concerns will be recognized.
- In 2003, the supreme court of Iceland ruled that a daughter has the right to prohibit the transfer of her deceased father's health information to a Health Sector Database, not because her right acting as a substitute of her deceased father, but in the recognition that she might, on the basis of her right to protection of privacy, have an interest in preventing the transfer of health data concerning her father into the database, as information could be inferred from such data relating to the hereditary characteristics of her father which might also apply to herself.

https://epic.org/privacy/genetic/iceland_decision.pdf

A Moral Challenge to DP



Say I steal 2 cents from every bank account in America. I am proven guilty, but everyone I stole from says they're fine with it. What happens?

- If one makes profit from applying DP to a dataset of many individuals, isn’t this morally equivalent to the above?

Outline

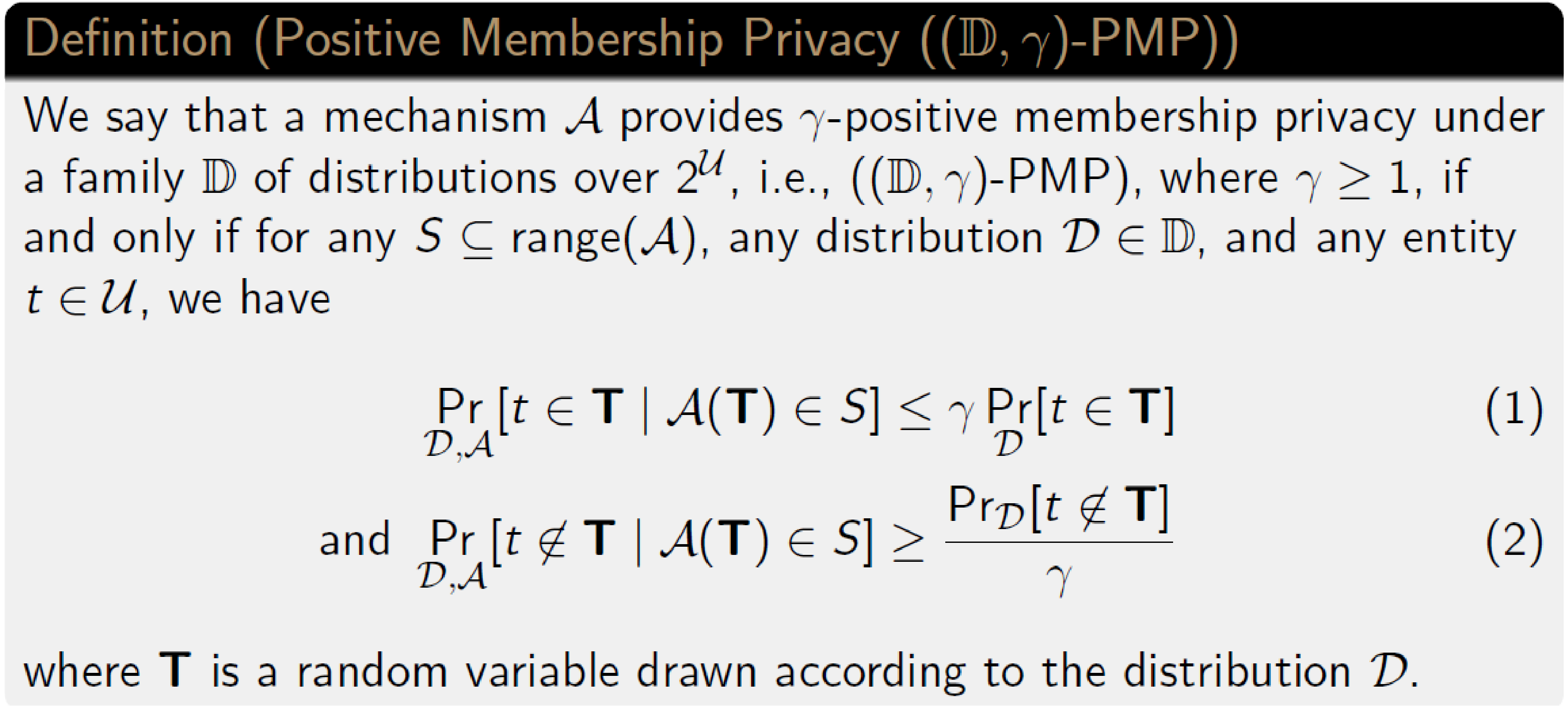
- Why Study Membership Inference?
- Differential Privacy
- Membership privacy and differential privacy
- Formulations of Membership Inference Attacks
- What We Currently Know about MI Attacks

A Formal Membership Privacy Framework

- Adversary has some prior belief about the input dataset (modeled by a prob. dist. over all possible datasets)
  - Gives the prior probability of any t’s membership
- Adversary updates belief after observing output of the algorithm, via Bayes rule
  - Obtains posterior probability of t’s membership
- For any t, posterior belief should not change too much from prior
- Membership privacy is parameterized by the family of prior distributions the adversary is allowed to have

Ninghui Li, Wahbeh H. Qardaji, Dong Su, Yi Wu, Weining Yang: ***Membership privacy: a unifying framework for privacy definitions.*** *CCS 2013: 889-900*

Positive Membership Privacy



Definition (Positive Membership Privacy ($(\mathbb{D},\gamma)$-PMP))

We say that a mechanism $\mathcal{A}$ provides γ-positive membership privacy under a family $\mathbb{D}$ of distributions over $2^{\mathcal{U}}$, i.e., ($(\mathbb{D},\gamma)$-PMP), where $\gamma \geq 1$, if and only if for any $S \subseteq \text{range}(\mathcal{A})$, any distribution $\mathcal{D} \in \mathbb{D}$, and any entity $t \in \mathcal{U}$, we have

$$\Pr_{\mathcal{D},\mathcal{A}}[t \in \mathbf{T} \mid \mathcal{A}(\mathbf{T}) \in S] \leq \gamma \Pr_{\mathcal{D}}[t \in \mathbf{T}] \qquad (1)$$

$$\text{and} \quad \Pr_{\mathcal{D},\mathcal{A}}[t \notin \mathbf{T} \mid \mathcal{A}(\mathbf{T}) \in S] \geq \frac{\Pr_{\mathcal{D}}[t \notin \mathbf{T}]}{\gamma} \qquad (2)$$

where $\mathbf{T}$ is a random variable drawn according to the distribution $\mathcal{D}$.

E.g., γ=1.25, when Pr[t∈**T**]=0.8, Pr[t∈**T** | *A*(**T**) ∈*S*] ≤ min(0.8*1.25, 1-0.2/1.25)
= min(1,1-0.16) = 0.84

when Pr[t∈**T**]=0.2, Pr[t∈**T** | *A*(**T**) ∈*S*] ≤ min(0.2*1.25, 1-0.8/1.25)
= min(0.25, 1-0.64) = 0.25

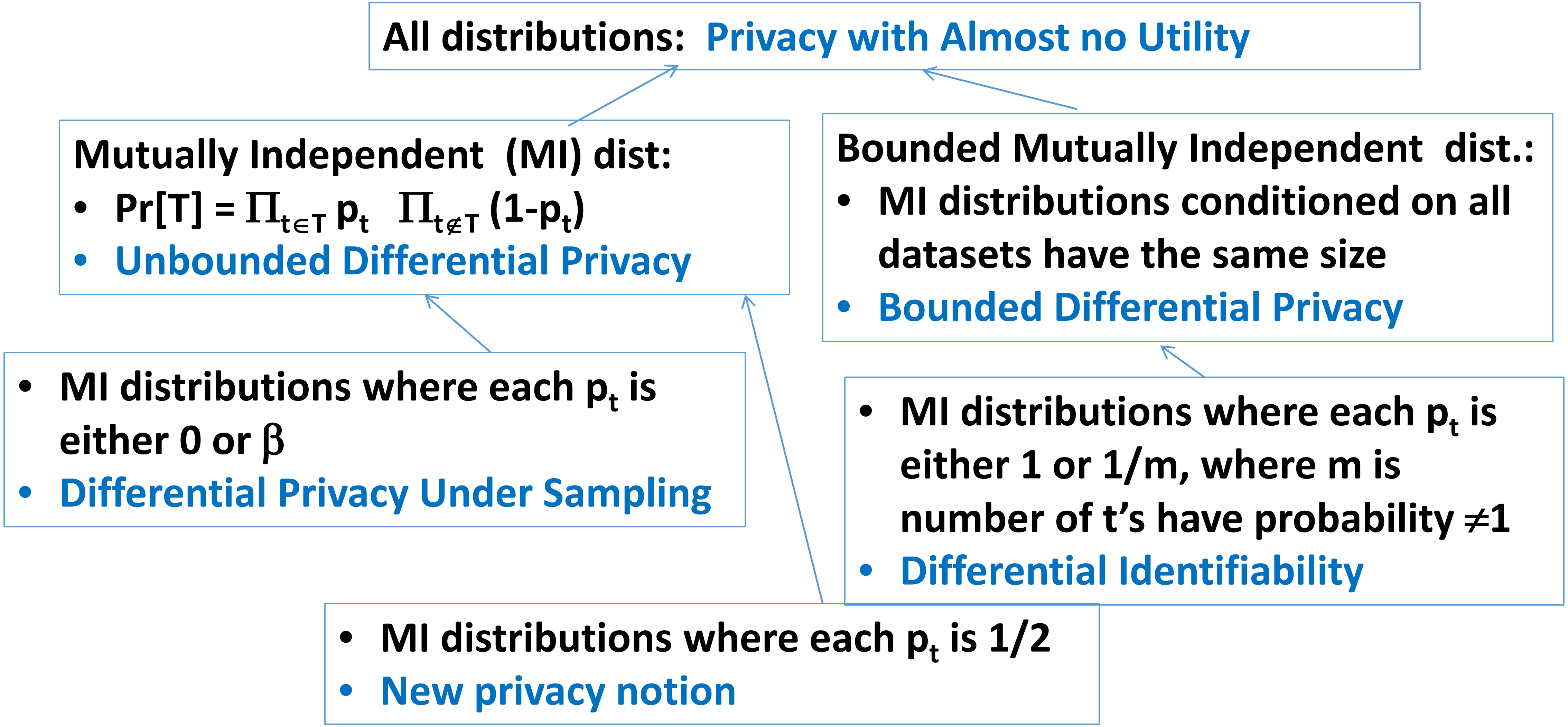
Impossibility of “Privacy as Secrecy” in the Membership Privacy Framework

- Membership privacy for the family of all possible distributions is infeasible
  - Requires publishing similar output distributions for two completely different datasets
  - Output has (almost) no utility
- Moral: One has to make some assumptions about the adversary’s prior belief
  - Assumptions need to be clearly specified and reasonable

Differential Privacy as Membership Privacy

- DP is equivalent to (positive + negative) MP under the family of all Mutually Independent (MI) distributions
    - Each MI distribution can be written as
      $\Pr[T] = \Pi_{t \in T} \, p_t \quad \Pi_{t \notin T} \, (1\text{-}p_t)$ where there is p_t for each t
- Differential privacy insufficient for membership privacy without independence assumption

Membership Privacy Notions



All distributions: Privacy with Almost no Utility

Mutually Independent (MI) dist:

- Pr[T] = $\Pi_{t \in T}\, p_t \quad \Pi_{t \notin T}\,(1-p_t)$
- Unbounded Differential Privacy

Bounded Mutually Independent dist.:

- MI distributions conditioned on all datasets have the same size
- Bounded Differential Privacy

- MI distributions where each p_t is either 0 or β
- Differential Privacy Under Sampling

- MI distributions where each p_t is either 1 or 1/m, where m is number of t's have probability $\neq$1
- Differential Identifiability

- MI distributions where each p_t is 1/2
- New privacy notion

Outline

- Why Study Membership Inference?
- Differential Privacy
- Membership privacy and differential privacy
- Formulations of Membership Inference Attacks
- What We Currently Know about MI Attacks

MI Attack Effectiveness as an Empirical Measure of Privacy

- Intuitively, training an ML model offers some level of privacy, since the model uses aggregated information
- But one cannot prove privacy of classifiers
- If we understood what are the best MI attacks, then level of resistance to MI attack indicates level of privacy
    - Similar to how we understand security of cryptographic primitives
- Privacy of ML models can then be empirically measured

Adversary Models for MI Attacks against Classifiers (Model Access)

- Knowledge of and access to the target model
  - Black-box: Can query the target model
  - White-box: Exact parameters of the target model
  - Federated: Model parameters during training (e.g., Federated Learning)

Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov: ***Membership Inference Attacks Against Machine Learning Models****. IEEE Symposium on Security and Privacy 2017: 3-18*

Milad Nasr, Reza Shoki, Amir Houmansadr: **Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning.** IEEE Symposium on Security and Privacy 2019: 739-753

Adversary Models for MI Attacks (2)

- Auxiliary Information
  - Distribution of training data
  - Model architecture and training recipe for target model
    - Enable training of **shadow models**
  - Knowledge of some non-members
  - Knowledge of some members

Some Example MI Attacks

- Use prediction label
  - Baseline attack: If prediction is correct, then conclude member
- Use predicted confidence for the correct label (equivalent cross-entropy loss)
  - Global: loss below some global threshold indicates member
  - Class: loss below some class-specific threshold
  - Instance: loss below some instance-specific threshold

More Example MI Attacks

- White-box
  - Use activation map of neurons
  - Use gradients of the target instance in the target model
    - Members should be smaller gradient norm

Metrics of Membership Inference

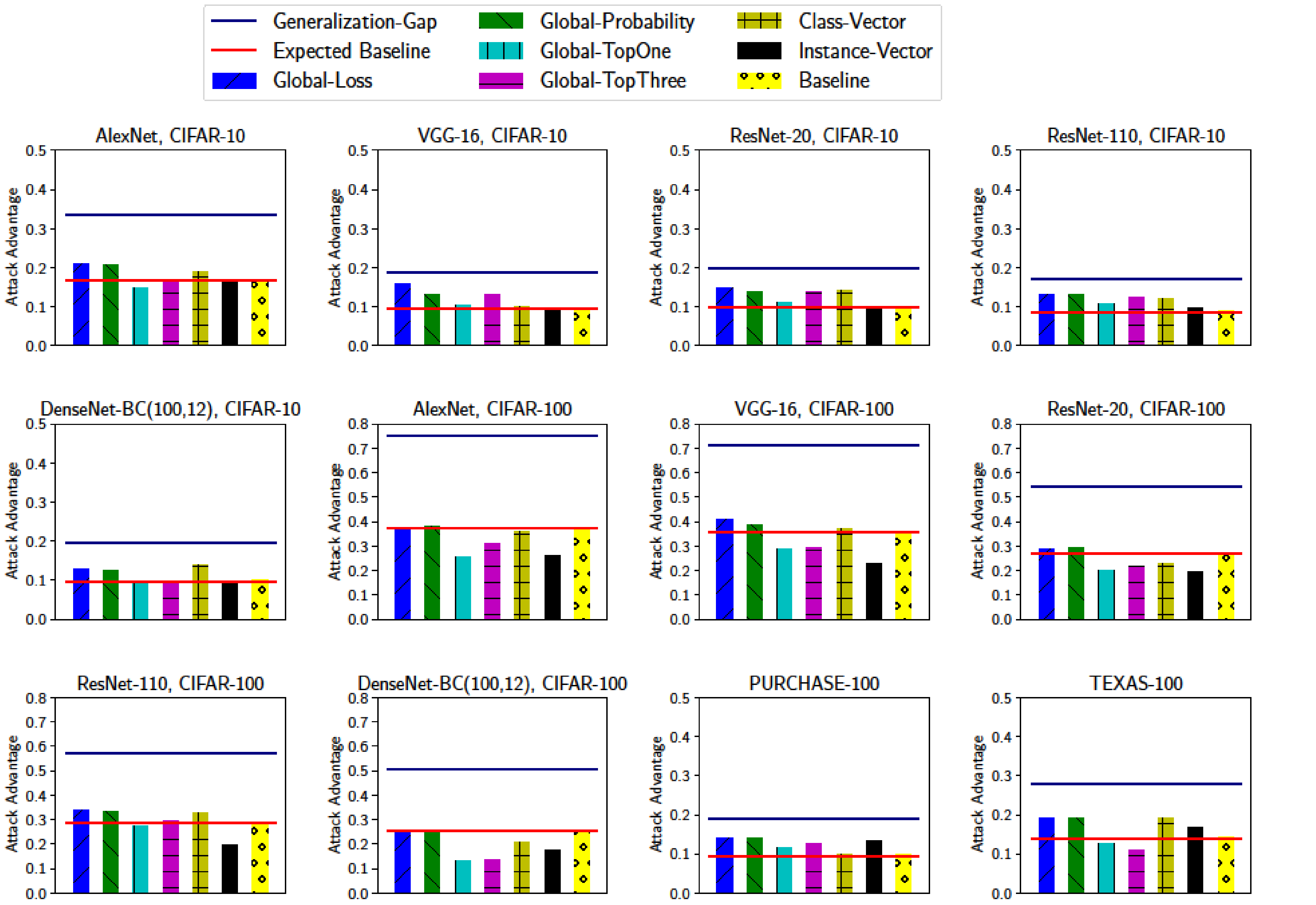
- Challenge: What distribution of members/non-members for evaluation?
  - Often assumed: 50-50
- Standard classification metrics
  - Accuracy, AUC
- Metrics focusing on high-confidence (the most vulnerable) instances: TPR at low FPR
  - Consider worst-case effect of MI attacks

Outline

- Why Study Membership Inference?
- Differential Privacy
- Membership privacy and differential privacy
- Formulations of Membership Inference Attacks
- What We Currently Know about MI Attacks

What Do We Know about MIA Under Average-Case Metrics

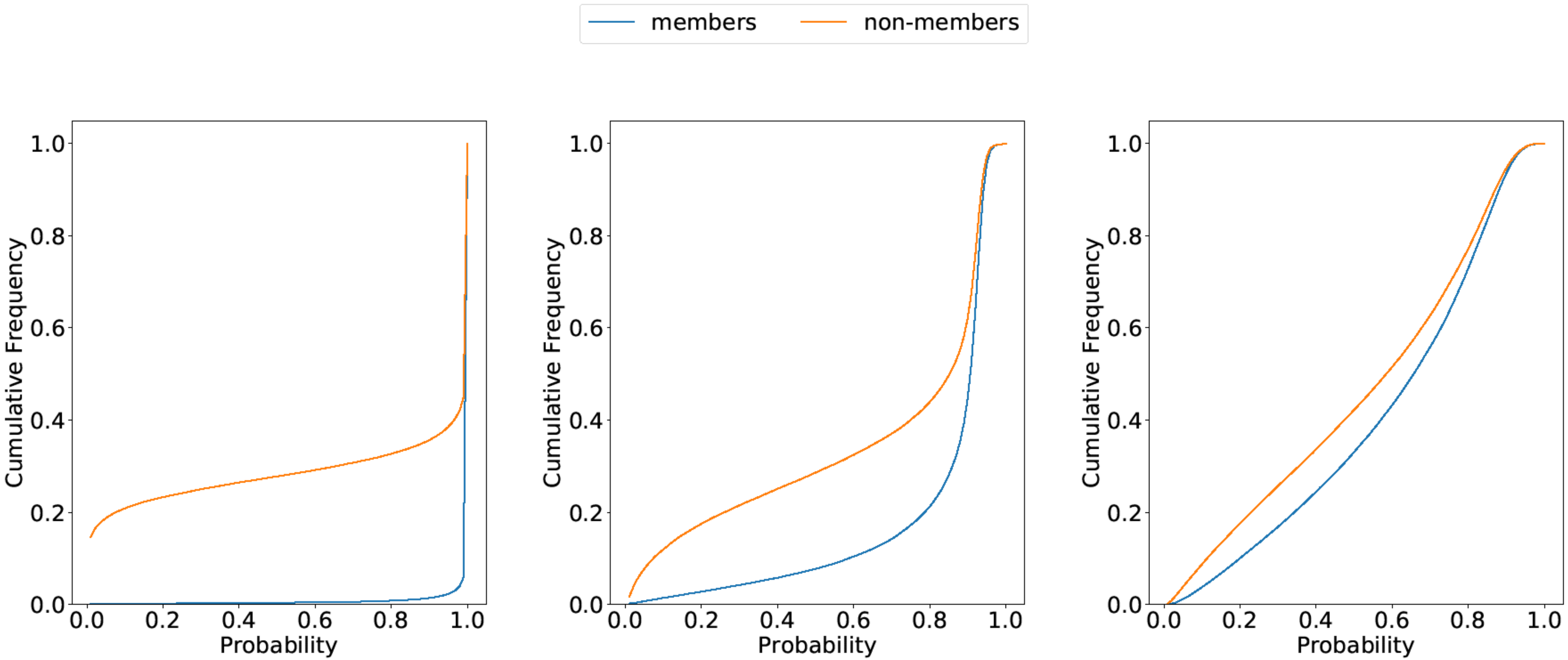
- Best attack exploits overfitting of the model
- Generalization gap (training accuracy – testing accuracy) plays an important role in MI advantage
- Baseline attack hard to beat



Jiacheng Li, Ninghui Li, Bruno Ribeiro: ***Membership Inference Attacks and Defenses in Classification Models.*** *CODAS PY 2021: 5-16*

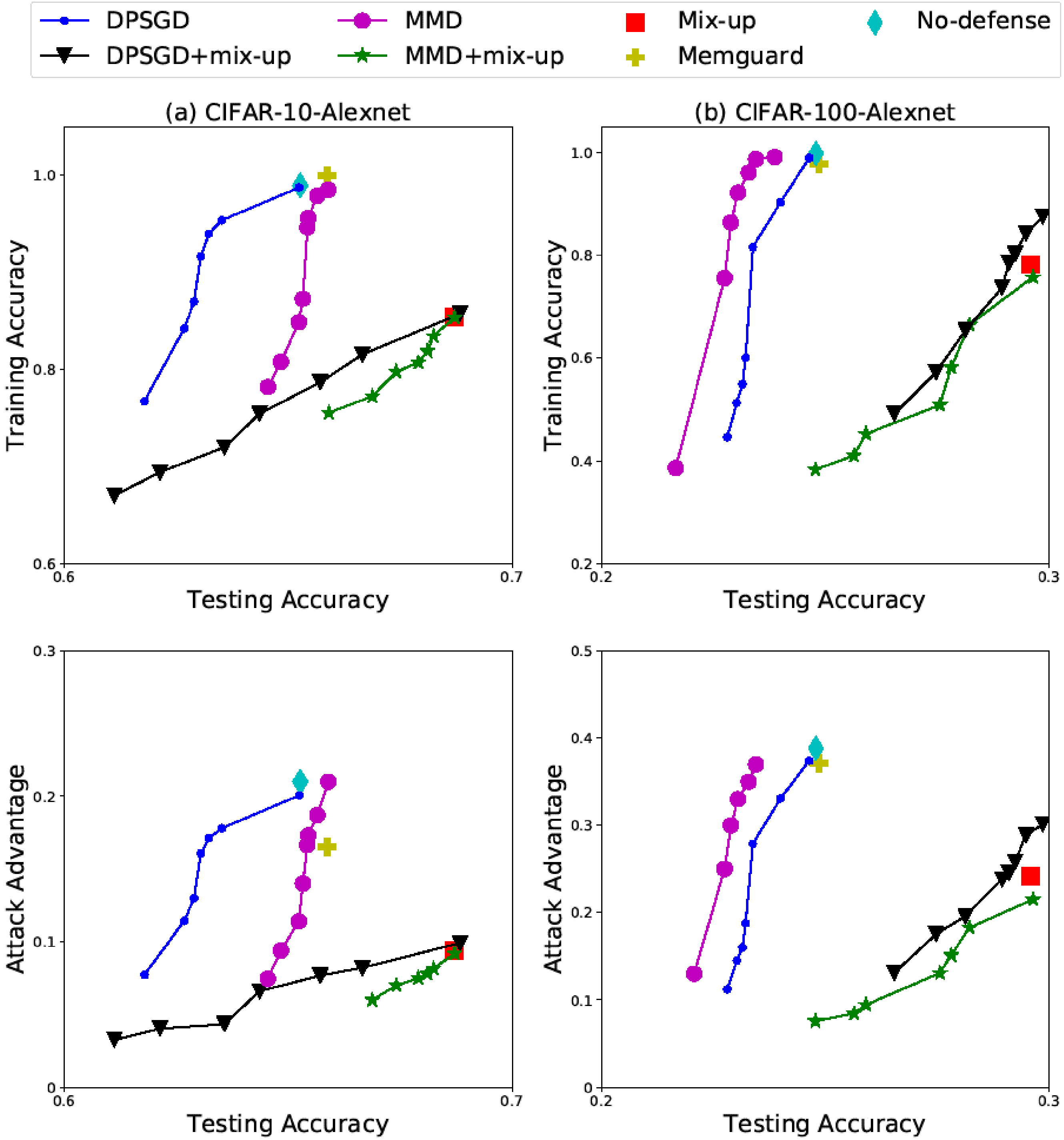
Proposed Defense Methods

- Intuition:
  - Reduce the generalization gap by intentionally reducing the training accuracy
  - Match the probability output distribution of training set with non-members
- Methods:
  - Mix-up training augmentation
  - Mean maximum discrepancy based regularization
  - DP-SGD: Add noises during training to satisfy DP (use noises sufficiently only for very large epsilon)



(a)No Defense (b)Mix-up only (c)Mix-up & MMD loss

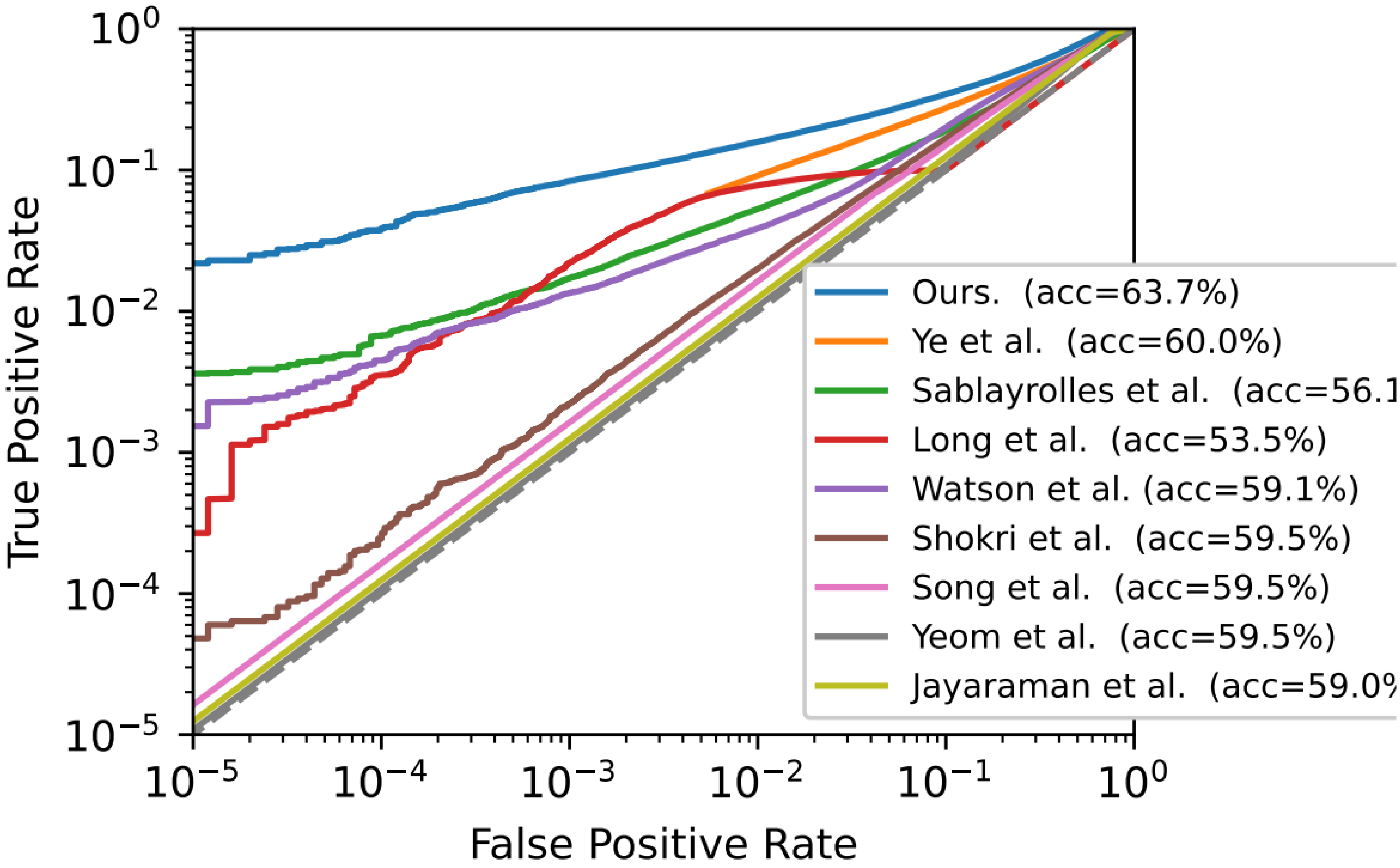
ResNet-20, CIFAR-100


DPSGD
MMD
Mix-up
No-defense
DPSGD+mix-up
MMD+mix-up
Memguard
(a) CIFAR-10-Alexnet
(b) CIFAR-100-Alexnet
Training Accuracy
Testing Accuracy
Attack Advantage
1.0
0.8
0.6
0.7
0.2
0.4
0.3
0.5
0.1
0

What Do We Know about MIA Considering Only High-Confidence Instances

- Consider TPR at very low FPR (e.g., 0.1% or 0.001%)
- Instance-specific hypothesis testing attacks
    - Train many shadow models, each using a subset of target instances
    - For each instance x, there are many models trained using x, and many without
    - Learn two Gaussian distributions (members vs non-members) of the loss
    - Output bayes prediction for the observed loss.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, Florian Tramèr: ***Membership Inference Attacks From First Principles.*** *IEEE Symposium on Security and Privacy 2022: 1897-1914*



Picture From

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, Florian Tramèr: ***Membership Inference Attacks From First Principles.*** *IEEE Symposium on Security and Privacy 2022: 1897-1914*

What do we Know about MIA in Federated Learning Setting?

Basic federated learning setting

- The central server initialize the central model weights
- For each communication round:
    - clients download the current central model weights from central server
    - clients perform local model update (e.g. SGD for a few batches)
    - clients send the updated model (or the parameter difference) to the central server
    - central server aggregates the updated model and obtain a new centralmodel
- perform the above training steps until convergence

Intuition of Attacks

- Gradient vectors of different training instances are orthogonal at later gradient update rounds
  - an overparameterized model has significantly more parameters than training instances
  - two high-dimensional Gaussian random vectors with zero mean and diagonal covariance matrix (isotropic) are nearly orthogonal
- The adversary observes model weights and updates
  - If the target instance not used by the client, then we expect the update is orthogonal to the gradient of that instance; otherwise, they are not

Jiacheng Li, Ninghui Li, Bruno Ribeiro: Effective passive membership inference attacks in federated learning against overparameterized models. ICLR 2023

Open Questions

- How effective are different defense mechanisms when considering high-confidence instances?
- How far are we from sufficiently understanding MI attacks in white-box classifiers to use it as measure of privacy?

Thank You!

- Questions?