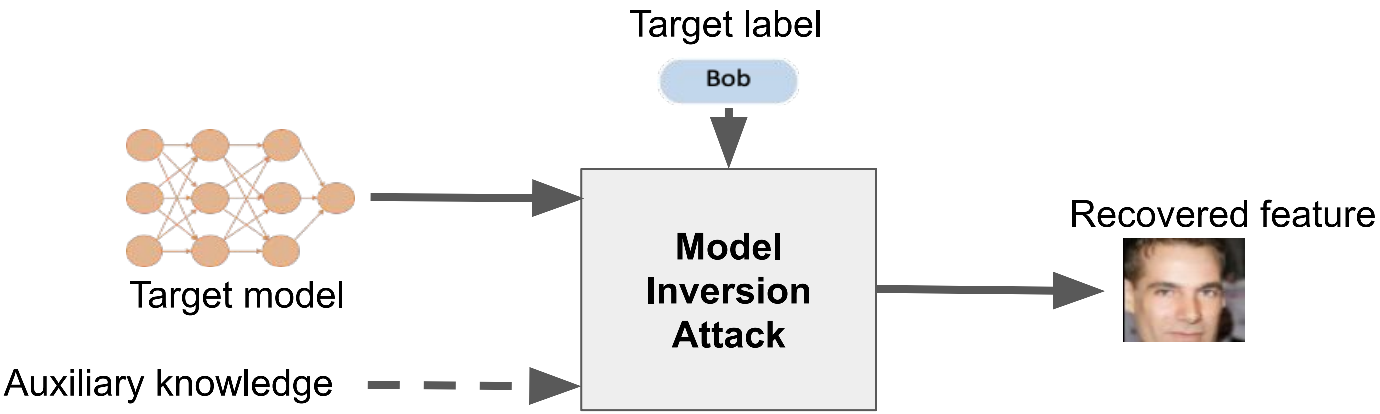
# Improving Robustness to Model Inversion Attacks via Mutual Information Regularization

**Tianhao Wang**, Yuheng Zhang, Ruoxi Jia

HARVARD
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
VT

# Background: Model Inversion (MI) Attack

- Goal: Given the access to a model, recover private training data associated with some target label
  - Blackbox: the attacker can only query the model
  - Whitebox: the attacker has the access to the model parameters



# Background: Attack Algorithms

- Attacks on different models: Linear regression [FLJLPR14], decision tree, and neural networks [FJR15, YCL19, SBBFZ20, ZJPWLS20]
- Common algorithm: Output the feature that is mostly likely to produce the target label under the target network, i.e. computing MLE $\max_x p(y|x)$
- Able to recover sensitive attributes for not only training data but also test data drawn independently from domain distribution.

# Prior Work on Defending MI Attacks

- Differential Privacy (DP)
  - Observed through empirical studies that DP cannot provide protection against MI attacks with reasonable model utility [FLJLPR14, ZJPWLS20].
  - Our paper presents a theoretical analysis that explains the inefficacy of DP.
- Model Specific Defenses
  - Decision tree: place sensitive features at a particular depth [FJR15].
  - DNN (black-box): injecting uniform noise to confidence scores [SBBFZ20], reducing their precision [FJR15] or dispersion [YCL19].

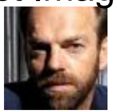
*Our defense is model agnostic and effective for **both** blackbox and whitebox settings.*

# Our Defense Goal

- Both the recovery of *training images* and *test images* would incur privacy loss to the target identity.

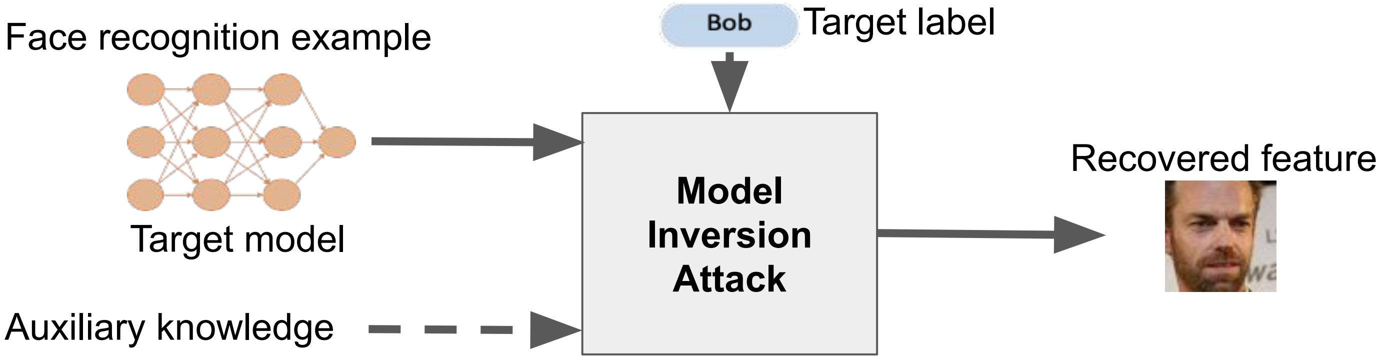
Training Image

Test Image



→ Both reveals the guy's face

- Design an algorithm to protect the training data distribution, instead of just training data set.



# MID: Mutual Information Regularization based Defense

- Intuition: if the output distribution $\hat{Y} = f(X)$ is independent from X, the attacker cannot learn anything about X's distribution.
- Method: Regularize the loss function by the *mutual information* between model's input and output distribution.
  - The mutual information is a measure of the mutual dependence between the two variables.

$$\min_{f \in \mathcal{H}} E_{(x,y) \sim p_{X,Y}(x,y)}[\mathcal{L}(y, f(x))] + \lambda \mathcal{I}(X, \hat{Y})$$

Original Loss Function

Regularizer Coefficient

mutual information between input and prediction

$$\mathcal{I}(X, \hat{Y}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)}\right) dy dx$$

- Challenge: mutual information is computationally expensive.

# Instantiation of MID

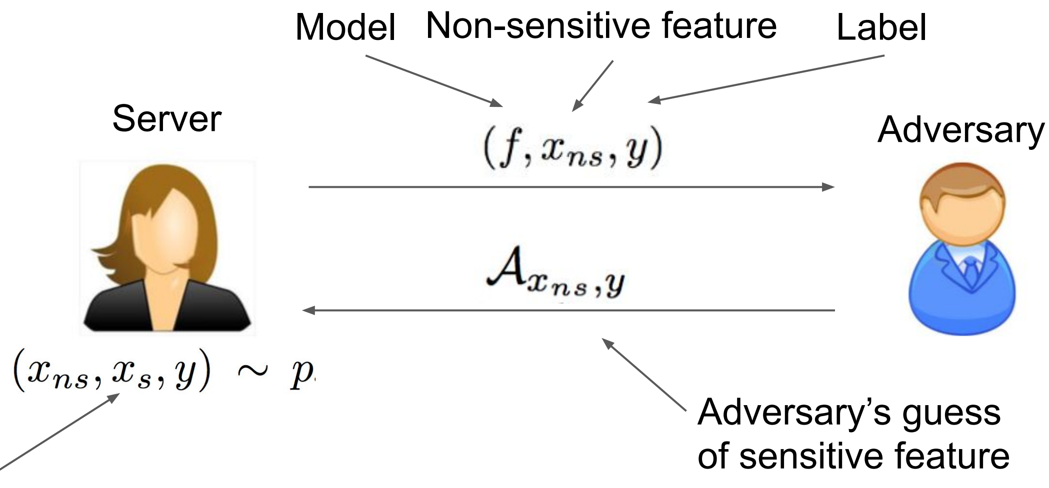
- Linear regression: Taylor-expansion based approximation
- Decision tree: modify ID3
- Deep Neural Networks: information bottleneck technique [AFDM16, ST17]
  - Regard the neural network as a Markov chain $Y - X - Z - \hat{Y}$

label feature stochastic encoding of intermediate layer prediction

By data processing inequality, we have $\mathcal{I}(X, \hat{Y}) \leq \mathcal{I}(X, Z)$ → new training loss → $\min_{\theta} -\mathcal{I}(Z; Y) + \lambda \mathcal{I}(Z, X)$

# Formalization of MI attack

- We formalize the MI attacks and quantify its distributional privacy loss.
- First attempt of modeling the privacy loss of members in the population.

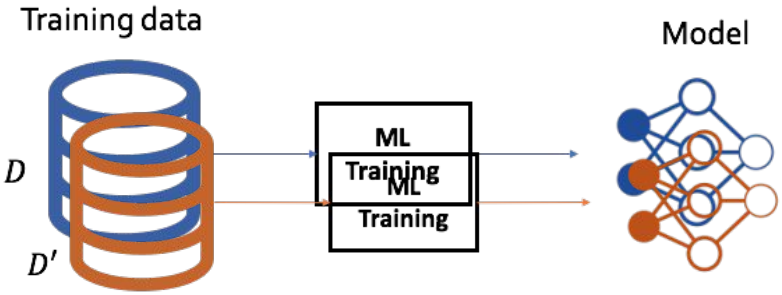



The adversary wins if the guess for $x_s$ is correct.
MI Privacy loss = privacy of adversary winning

# Characterizing MI privacy loss of DP models

**Definition 1** (Differential Privacy). *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ be a randomized mechanism. We say that $\mathcal{M}$ is $(\epsilon, \delta)$-differentially private if for every two adjacent datasets $S \sim S'$ and every subset $R \subseteq \mathcal{R}$,*

$$Pr[\mathcal{M}(S) \in R] \leq e^{\epsilon} Pr[\mathcal{M}(S') \in R] + \delta \qquad (9)$$




Several earlier empirical studies suggest that DP is not able to defend against model inversion attack with any reasonable model performance [FLJLPR14, ZJPWLS20]!

# Characterizing MI privacy loss of DP models

- Main result: when the learning algorithm is (ε, δ)-differentially private, the MI privacy loss is tightly upper bounded by $\frac{e^{n\epsilon}-1}{e^{n\epsilon}+1}+\frac{2(e^{n\epsilon}-1)}{(e^{n\epsilon}+1)(e^{\epsilon}-1)}\delta$ .
  - n: number of training data

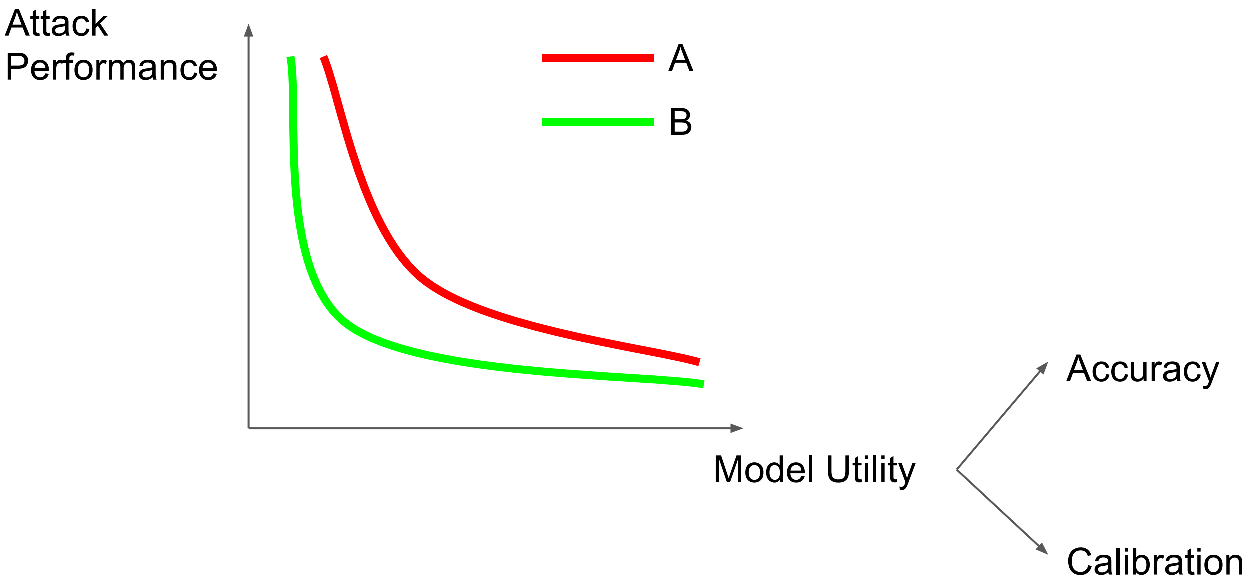
To make this bound small, the privacy budget ε needs to be set as o(1 / #training data)!

# Evaluation: Baselines

- Attack algorithms:
  - MAP [FLJLPR14, FJR15] // Black+white-box
  - Knowledge Alignment [YCL19] // Black-box
  - Update Leaks [SBBFZ20] // Black-box
  - GMI [ZJPWLS20] // White-box
- Defense baseline:
  - Differential Privacy
  - Set priority depth of sensitive attributes for decision tree
  - Noisy confidence scores for black-box DNNs

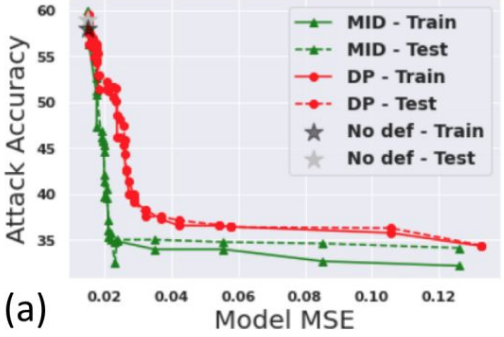
# Evaluation: Metrics

- Evaluate the performance of a defense mechanism in terms of the *privacy-utility tradeoff*.



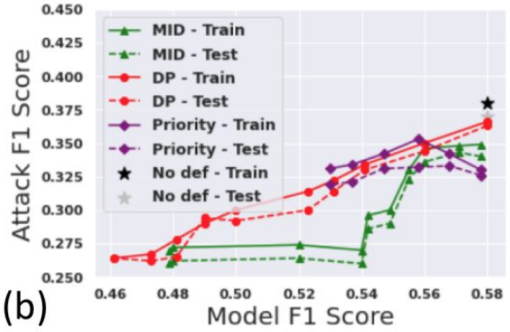
# Defense results for blackbox MI attacks

MAP on linear regression




(a)

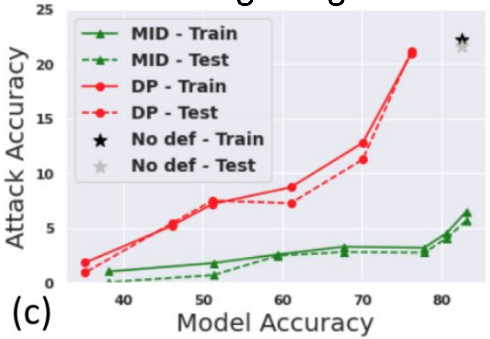
MAP on decision tree




(b)

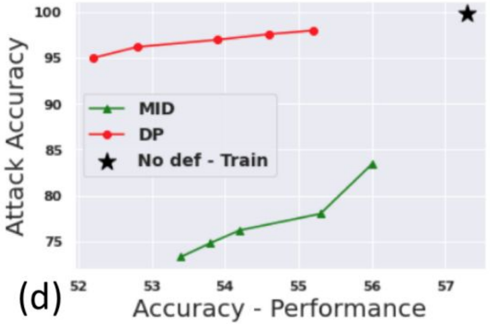
- The more predictive power the model has, the more vulnerable it is to the attacks.
- Our defense can significantly improve the model robustness for any fixed model performance.

Knowledge Alignment




(c)

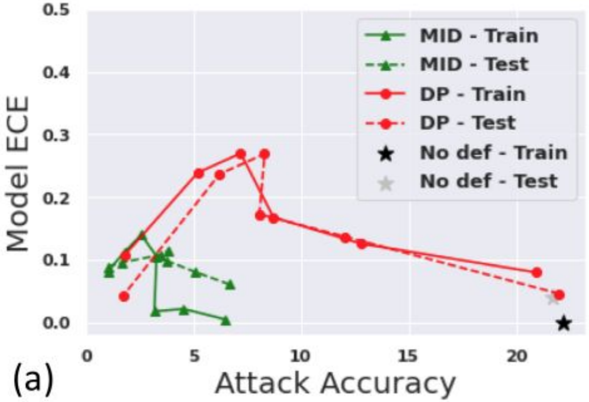
Update-Leaks



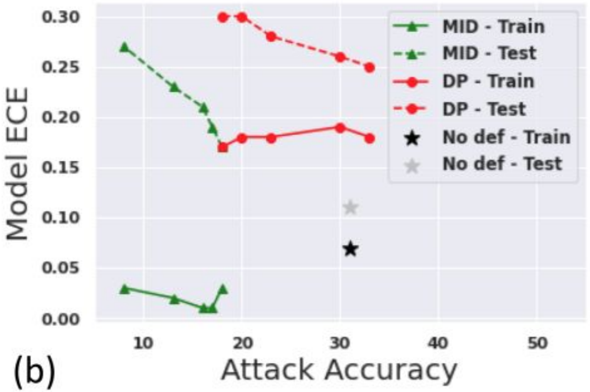

(d)

# Model Calibration

Knowledge Alignment



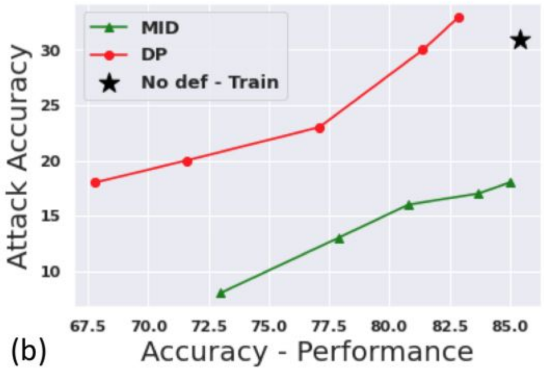

Update Leaks



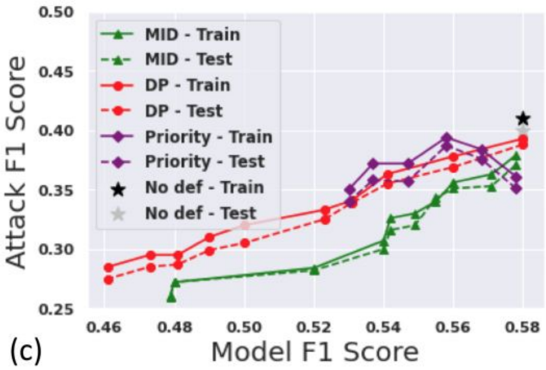

Expected Calibration Error (ECE) measures the mismatch between the model accuracy and confidence.

- Important for evaluating a risk model.

# Defense results for whitebox MI attacks


GMI




MAP with white-box counts

# Future Work

- Defending MI Attacks with computational security.

Thank you!

Contact: tianhaowang@fas.harvard.edu

# References


Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., & Ristenpart, T. (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)* (pp. 17-32).

Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 1322-1333).

Yang, Z., Chang, E. C., & Liang, Z. (2019). Adversarial neural network inversion via auxiliary knowledge alignment. *arXiv preprint arXiv:1902.08552*.

Salem, A., Bhattacharya, A., Backes, M., Fritz, M., & Zhang, Y. (2020). Updates-leak: Data set inference and reconstruction attacks in online learning. In *29th {USENIX} Security Symposium ({USENIX} Security 20)* (pp. 1291-1308).

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., & Song, D. (2020). The secret revealer: generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 253-261).

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.