# Stolen Memories

## Leveraging Model Memorization for Calibrated White-Box Membership Inference

**Klas Leino*** & Matt Fredrikson | Carnegie Mellon University

# Membership Inference

- **Basic Idea:** adversary tries to predict whether a given point was part of a target model's training set

auxiliary data

even prior

training algorithm

- Why do we care about MI?
  - Membership may itself be sensitive/private
  - Serves as a practical measure of information leakage
  - Directly related to differential privacy

# Related Work

Shokri et al. 2016
*Membership Inference Attacks Against Machine Learning Models*

Yeom et al. 2017
*The Unintended Consequences of Overfitting: Training Data Inference Attacks*

Long et al. 2017
*Towards Measuring Membership Privacy*

Long et al. 2018
*Understanding Membership Inferences on Well-generalized Learning Models*

Salem et al. 2019
*Model and Data Independent Membership Inference Attacks on Machine Learning Models*

Black-box

Nasr et al. 2018
*Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning Under Passive and Active White-box Inference Attacks*

White-box

# Can We do Better with White-box Access?

Model parameters might leak significantly more information; *can we leverage them?*

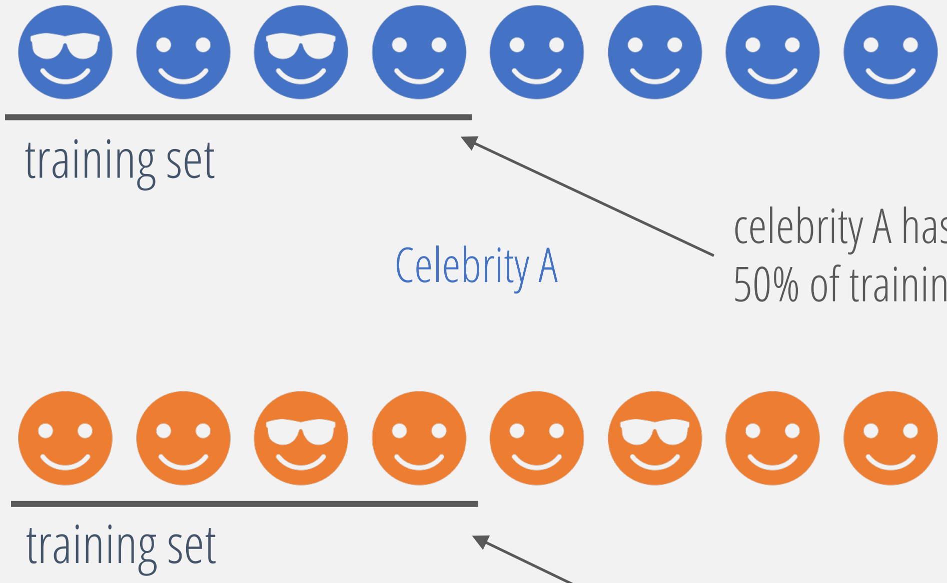
- Yes!
  - We show *how to explicitly identify memorization* in deep networks, and weaponize it for membership inference
  - We find our attacks can be calibrated for increased confidence in positive inferences
  - We evaluate the practicality of popular defenses against our attack

# Overview

- Privacy in Deep Learning
- Understanding Overfitting
- White-box Membership Inference
- Key Results

# How Does Overfitting Manifest Itself?

- Hypothesis: idiosyncratic feature use provides *evidence* of membership



training set

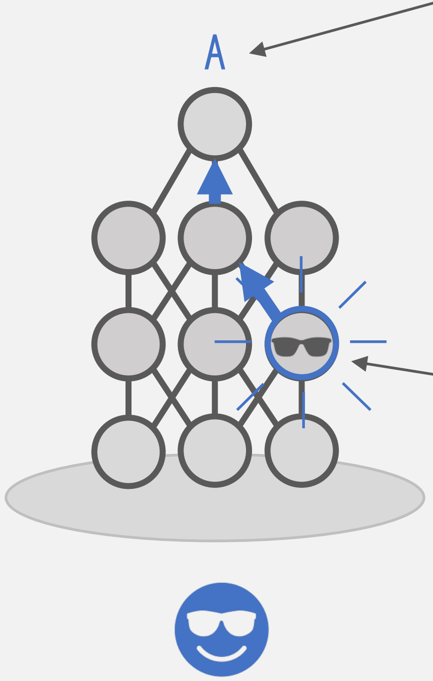
Celebrity A

celebrity A has sunglasses in 50% of training instances

training set

Celebrity B

celebrity B has sunglasses in 25% of training instances

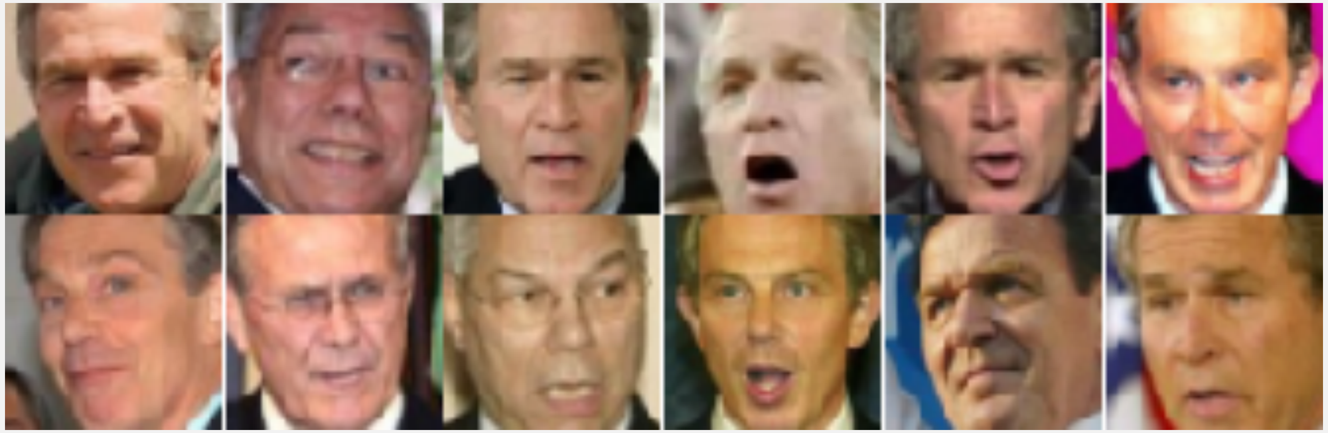


because sunglasses influence prediction of celebrity A, we infer sunglasses were more common in the training set than average for celebrity A

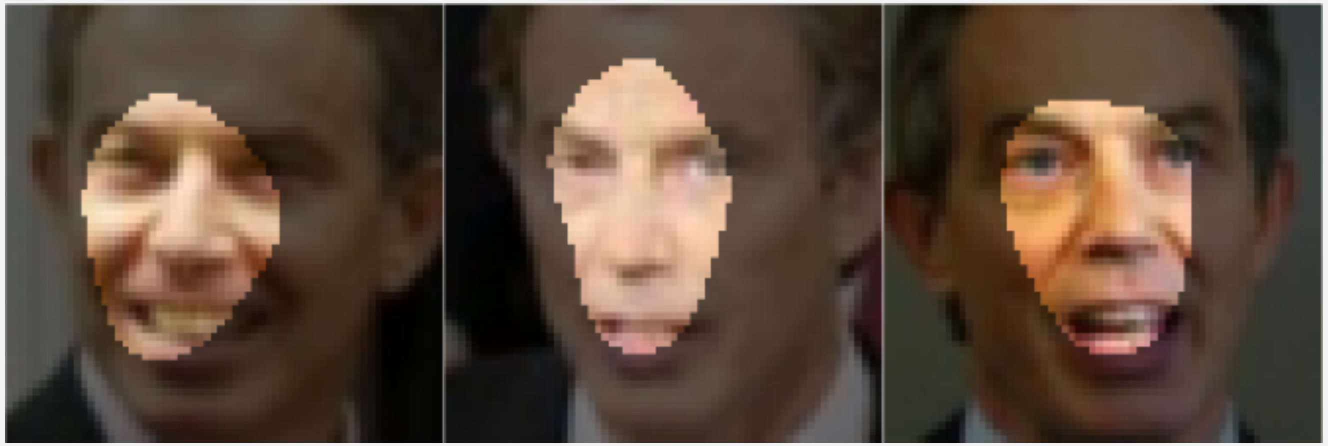
sunglasses are predictive in training set, so the model may learn to encode them

both celebrities A and B have sunglasses in 25% of instances

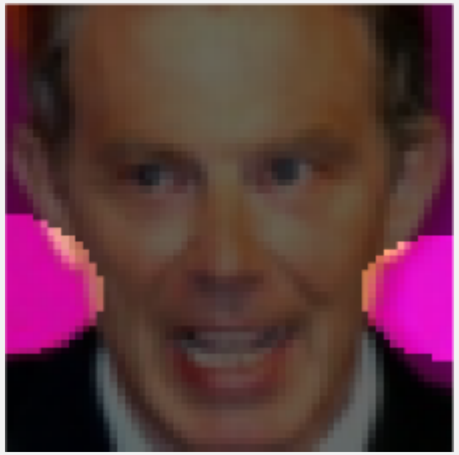
# Example



Sample of LFW training instances



Typical explanations on test instances of Tony Blair



Explanation on training instance of Tony Blair with distinctive pink background. The model uses the background to classify the instance as Tony Blair.

# Key Idea

Membership information is leaked through the target model's idiosyncratic use of features. Features that are distributed differently in the training data from how they are distributed in the general population provide evidence for or against membership.

Next we would like to formalize this intuition…

# Overview

- Privacy in Deep Learning
- Understanding Overfitting
- White-box Membership Inference
    - Bayes-optimal Membership Inference
    - Extending to Deep Models
- Key Results

# Formal Intuition

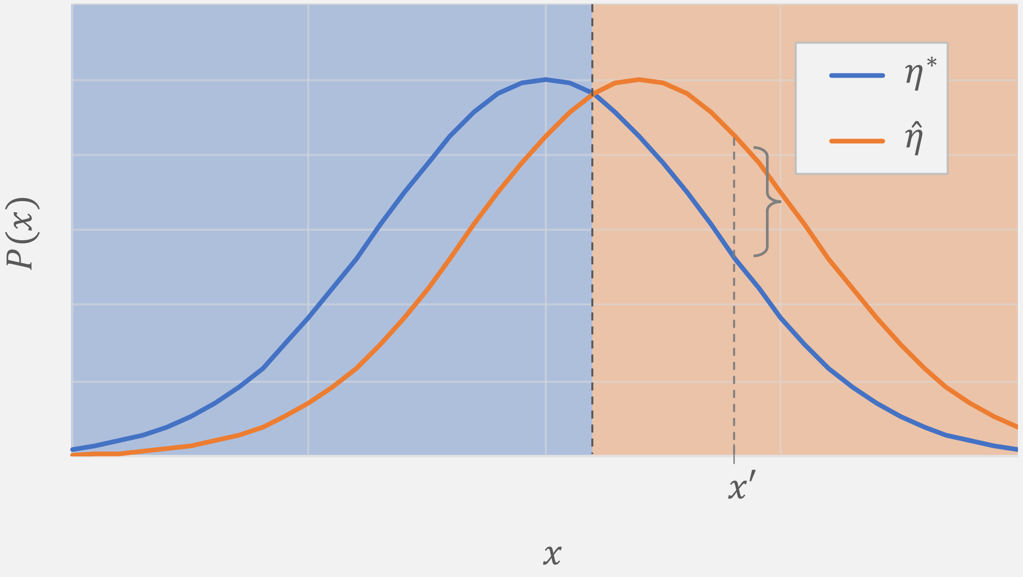
- Consider two normal distributions, $\eta^*$ and $\hat{\eta}$
- $x'$ is more likely to have been drawn from $\hat{\eta}$ than from $\eta^*$ if

$$\Pr_{\eta^*}[x'] < \Pr_{\hat{\eta}}[x']$$

- We can make a classifier that predicts whether $x'$ was drawn from $\eta^*$ or $\hat{\eta}$

general population

training set

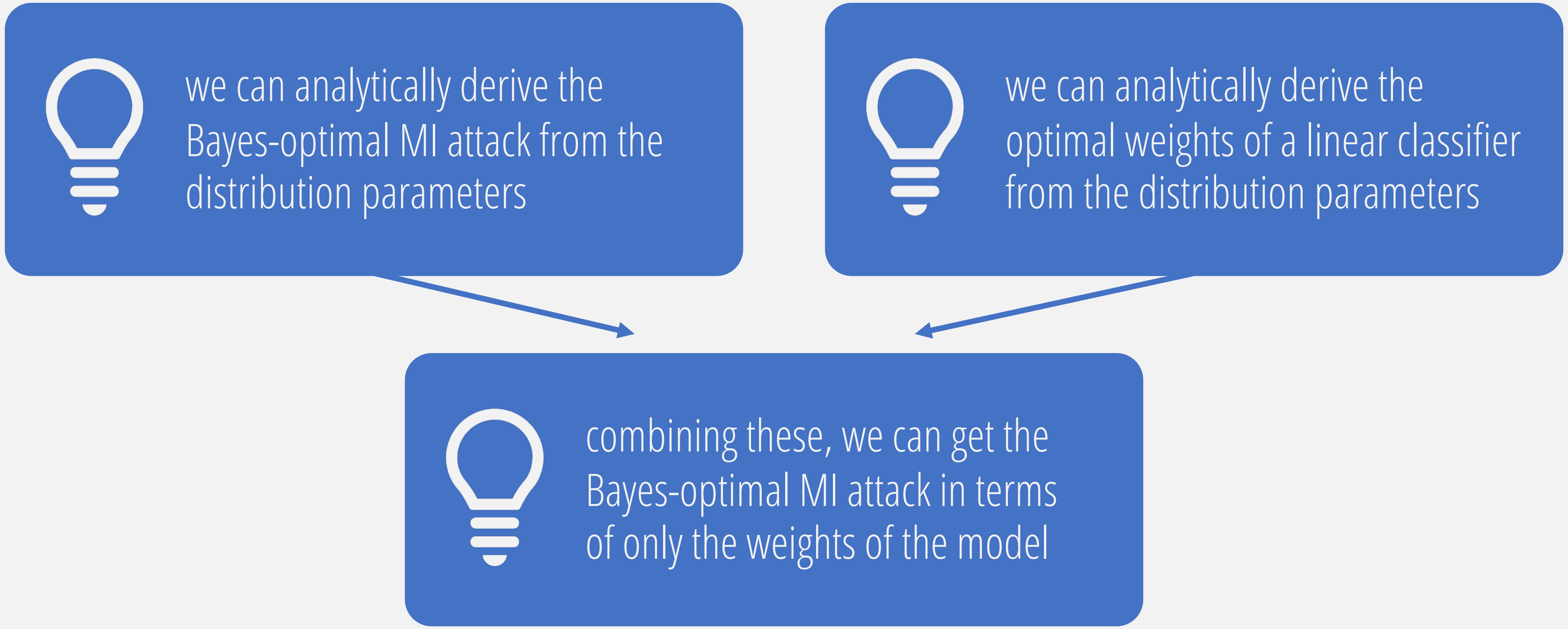


An attacker won't be able to know $\eta^*$ or $\hat{\eta}$

weights from target model convey relevant information about training distribution

# Bayes-optimal Membership Inference

- Assume the data is Gaussian and satisfies the naïve Bayes assumption
- Model the training set as a Gaussian distribution using its empirical mean and covariance
- Assume the target model is a linear model



we can analytically derive the Bayes-optimal MI attack from the distribution parameters

we can analytically derive the optimal weights of a linear classifier from the distribution parameters

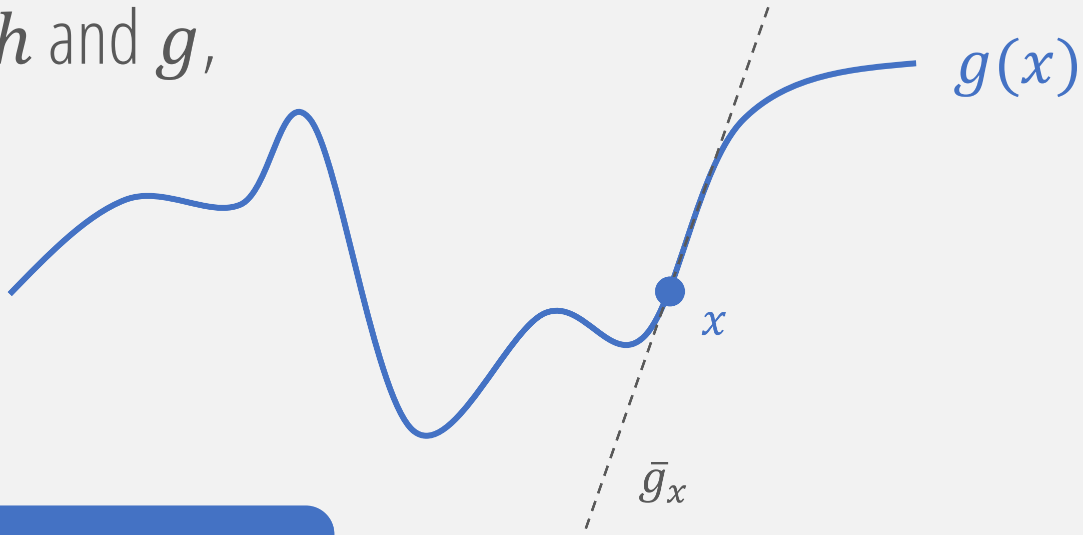
combining these, we can get the Bayes-optimal MI attack in terms of only the weights of the model

# Summary of Attack on Linear Models

- Train a *proxy model* on the auxiliary data
- Compare the weights of the proxy model to those of the target model to create an *attack model*
- Features that are used differently in the target model from in the proxy model are used to determine membership

# Layer-wise Attacks

- We can apply the same principle to a layer of a deep network
- We take a *slice* of the network, $\boldsymbol{f}$: two functions, $\boldsymbol{h}$ and $\boldsymbol{g}$, such that $\boldsymbol{f} = \boldsymbol{g} \circ \boldsymbol{h}$
  - $h$ computes features
  - $g$ classifies using these features



target model and proxy model must share the same fixed internal representation, i.e., $\hat{h} = \tilde{h}$

we can use a *local linear approximation* of $\hat{g}$ and $\tilde{g}$ to apply the linear attack described previously to the remainder of the network

# Combining Layers

We can train one such attack for each layer of the model and combine the results using a *meta attack model**

**see paper for mor details*

# Overview

- Privacy in Deep Learning
- Understanding Overfitting
- White-box Membership Inference
- Key Results
  - Performance
  - Differential Privacy

# Key Results

- Outperforms black-box attacks
- Ability to calibrate for precision
- Capable of high-precision inferences even on models that generalize well

| Dataset | Generalization Error | Our Precision |
|---|---|---|
| MNIST | 1% | 75% |

# Differential Privacy

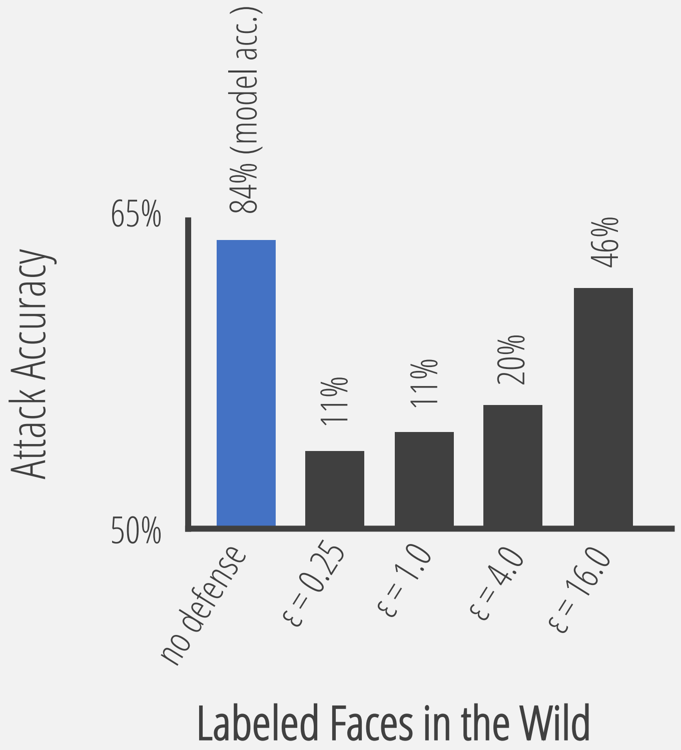
**Formal Guarantee***: ε-differential privacy guarantees that no adversary can achieve an accuracy greater than $\underline{e^{\varepsilon}/2}$

note that this implies DP gives *no meaningful guarantee* if $\varepsilon \geq \ln(2) \approx 0.69$

**result from Yeom et al. 2017*

# Is Differential Privacy a Good Defense?

- Small ε reduces attack accuracy, but is costly in terms of model utility
- Large ε may reduce attack accuracy, but not always



# Conclusion

- What else is in the paper?
  - Detail on analysis
  - More sophisticated variants of our attack
  - Additional experimental results
  - Evaluation of other defenses, e.g., regularization

### Key Takeaways

- Membership information is leaked through sensitive feature usage
- Using white-box information we can improve upon the state-of-the art MI attacks
- Large-$\varepsilon$-DP may not confer significant privacy protection

# Thank You!

contact kleino@cs.cmu.edu with questions