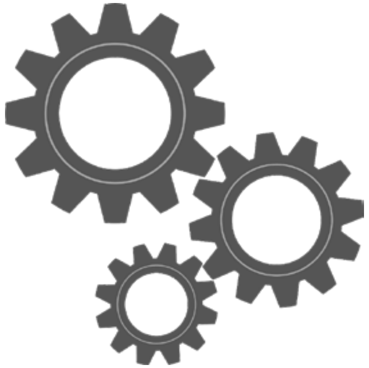
On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models

Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, Peter Dickinson

IEEE EuroS&P – 8 September 2021

benjamin.zhao@unsw.edu.au

Attacks on Machine Learning



ML Model

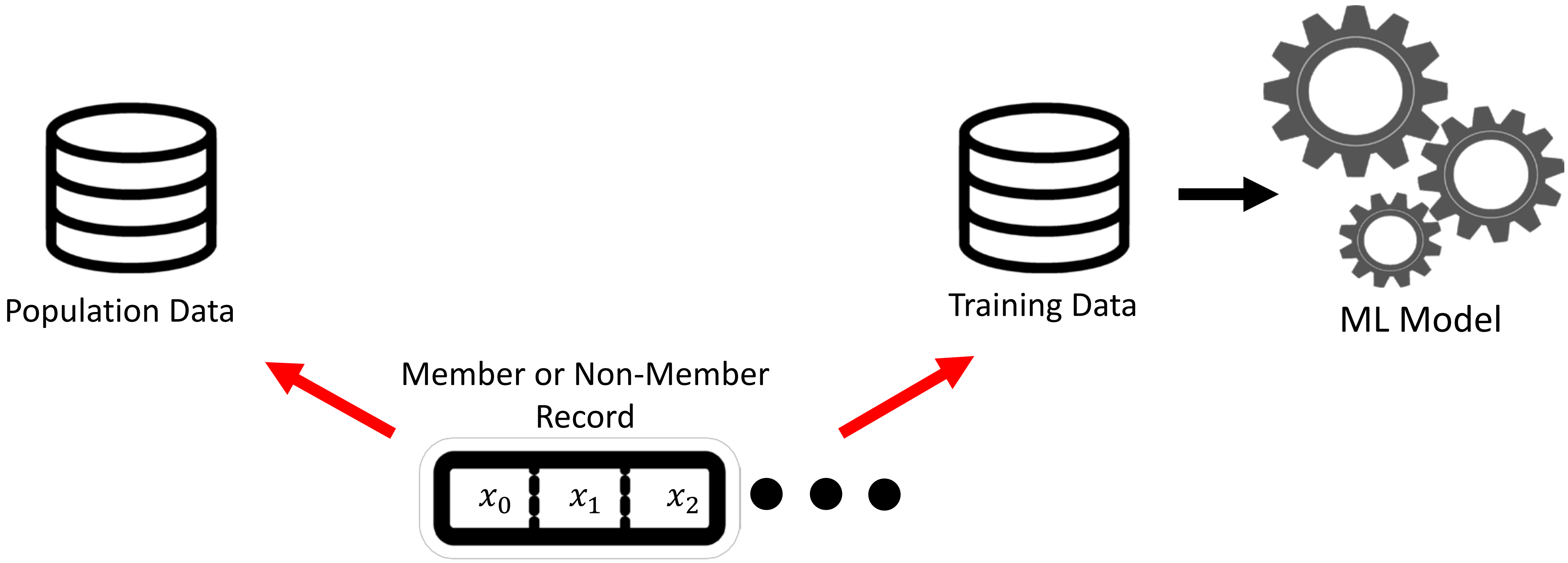
Unexpected Behaviours

- Adversarial Examples
- Poisoning Attacks
- Backdoor Attacks

Information Leakage

- Model Extraction
- **Membership Inference**
- **Attribute Inference**

Membership Inference



- Infer if any given record is from the training data.

Membership Inference Works

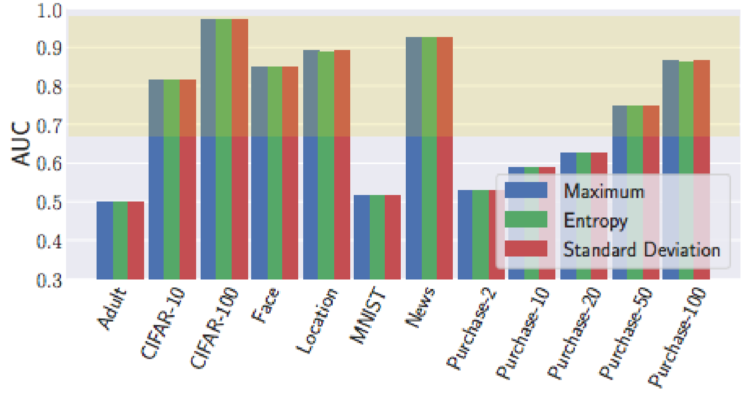
Membership Inference Attacks Against Machine Learning Models

Reza Shokri, Cornell Tech, shokri@cornell.edu
Marco Stronati[*], INRIA, marco@stronati.org
Congzheng Song, Cornell, cs2296@cornell.edu
Vitaly Shmatikov, Cornell Tech, shmat@cs.cornell.edu

Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models

Ahmed Salem[*], Yang Zhang[*§], Mathias Humbert[†], Pascal Berrang[*], Mario Fritz[*], Michael Backes[*]
[*]CISPA Helmholtz Center for Information Security,
{ahmed.salem, yang.zhang, pascal.berrang, fritz, backes}@cispa.saarland
[†]Swiss Data Science Center, ETH Zurich and EPFL, mathias.humbert@epfl.ch



Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting

Samuel Yeom[*] Irene Giacomelli[†] Matt Fredrikson[*] Somesh Jha[†]
[*]Carnegie Mellon University, [†]University of Wisconsin–Madison

	Our work	Shokri et al. [7]
Attack complexity	Makes only one query to the model	Must train hundreds of shadow models
Required knowledge	Average training loss L_S	Ability to train shadow models, e.g., input distribution and type of model
Precision	0.505 (MNIST) 0.694 (CIFAR-10) 0.874 (CIFAR-100)	0.517 (MNIST) 0.72-0.74 (CIFAR-10) > 0.99 (CIFAR-100)
Recall	> 0.99	> 0.99

Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning

Milad Nasr, University of Massachusetts Amherst, milad@cs.umass.edu
Reza Shokri, National University of Singapore, reza@comp.nus.edu.sg
Amir Houmansadr, University of Massachusetts Amherst, amir@cs.umass.edu

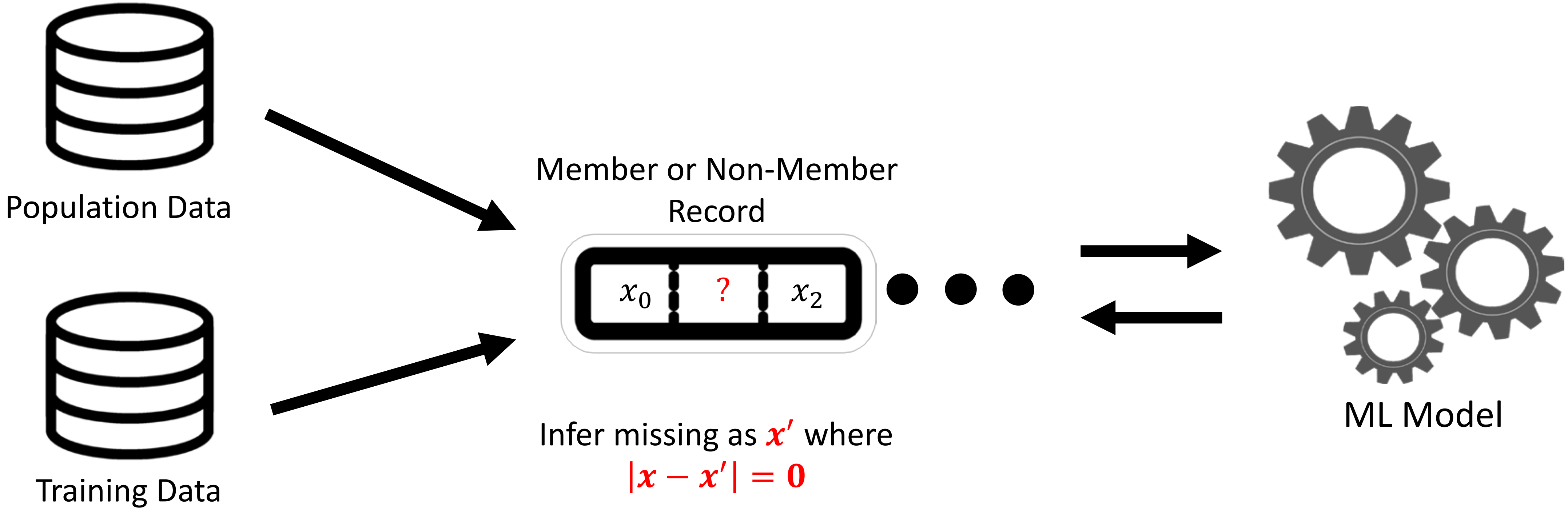
White-box vs Black-box: Bayes Optimal Strategies for Membership Inference

Alexandre Sablayrolles[1 2] Matthijs Douze[2] Yann Ollivier[2] Cordelia Schmid[1] Hervé Jégou[2]

LOGAN: Membership Inference Attacks Against Generative Models[*]

Jamie Hayes[†], Luca Melis[†], George Danezis, and Emiliano De Cristofaro
University College London
{j.hayes, l.melis, g.danezis, e.decristofaro}@cs.ucl.ac.uk

Attribute Inference



- Infer information on missing attribute(s) with access to the ML Model.
- Is there any *advantage* to inferring attributes when in or out of the training data. (Learning from the Distribution versus Learning from inclusion)

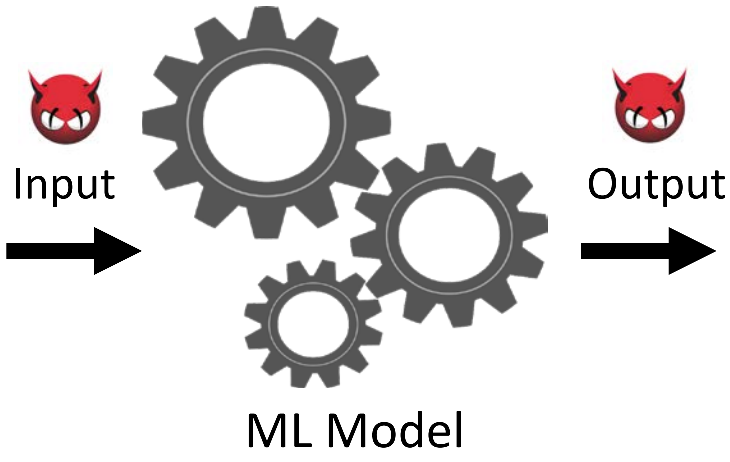
Evaluating Attribute Inference

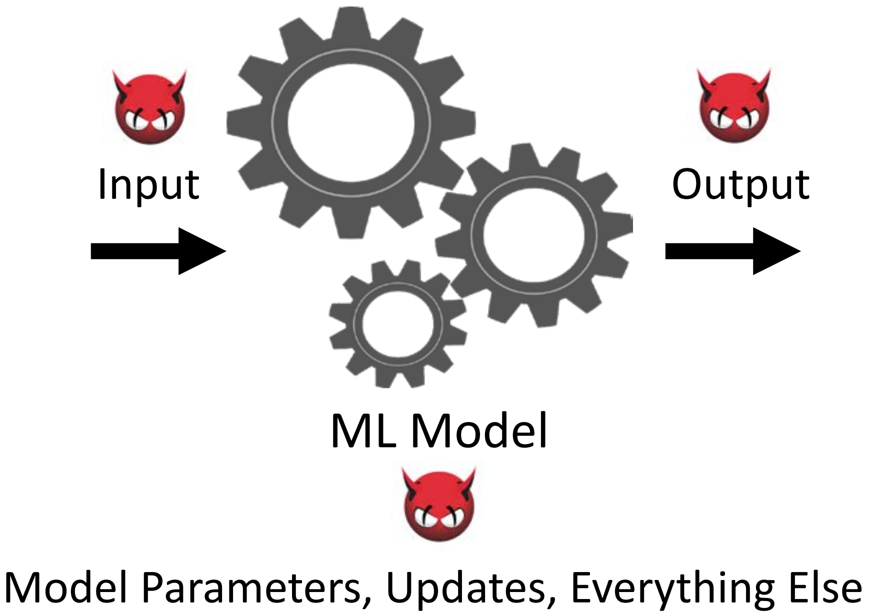
Infer 15 (Most Important) Missing Features

AI	Loc-30	Pur-2	Pur-10	Pur-20	Pur-50	Pur-100
Conf	7.78E-4	1.38E-5	-3.69E-4	2.16E-4	2.00E-3	1.65E-3
Loss	7.76E-4	-9.79E-5	5.57E-3	6.69E-3	4.59E-3	5.09E-3
Shadow	8.00E-4	-2.00E-4	2.17E-3	2.63E-3	4.10E-3	4.20E-3

The models above are vulnerable to ***Membership*** Inference, however there is negligible advantage when performing ***Attribute*** Inference

Attacks Threat Model

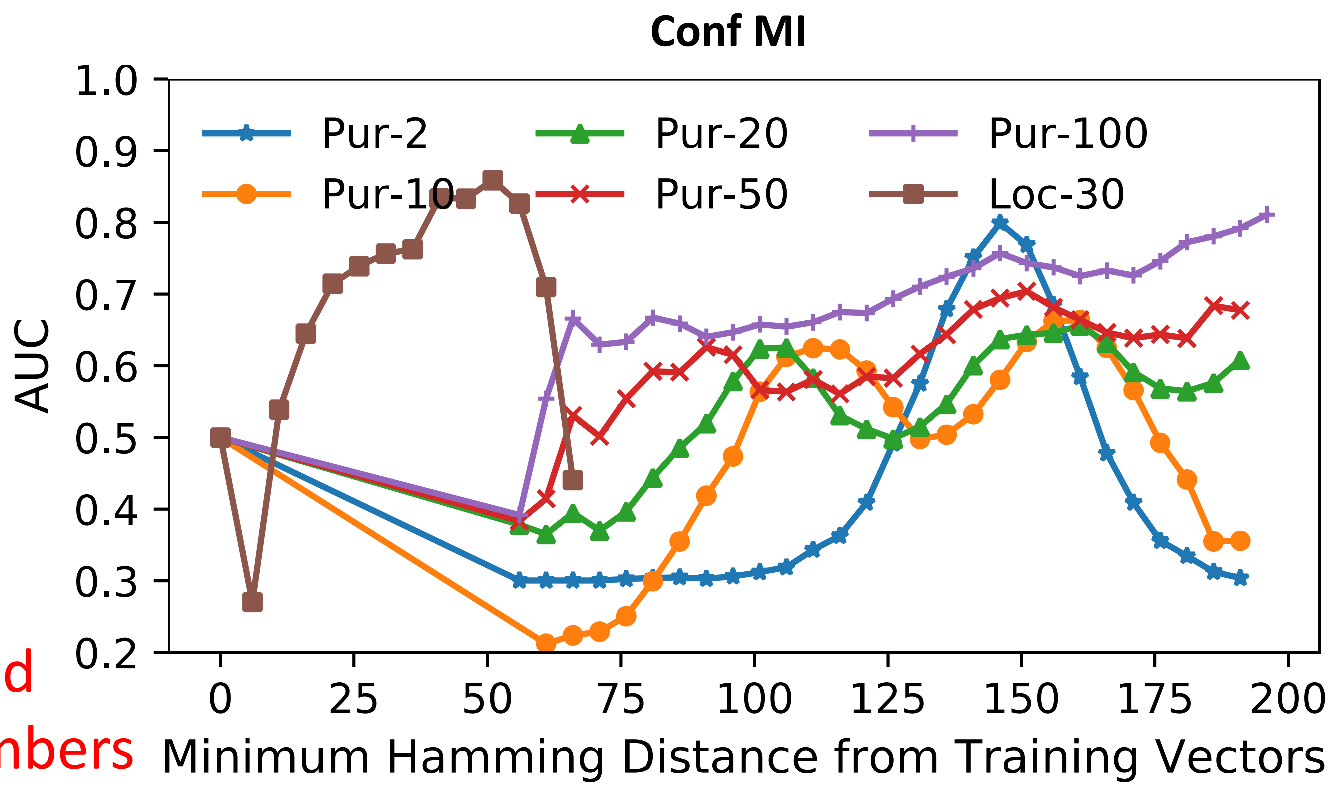




- 3 **Black Box** attacks
  - Shadow MI (Shokri et al.)
  - Loss MI (Yeom et al.)
  - Confidence MI (Salem et al.)

- 2 **White Box** attacks
  - Local MI (Nasr et al.)
  - Global MI (Nasr et al.)

Evaluating Existing Membership Inference



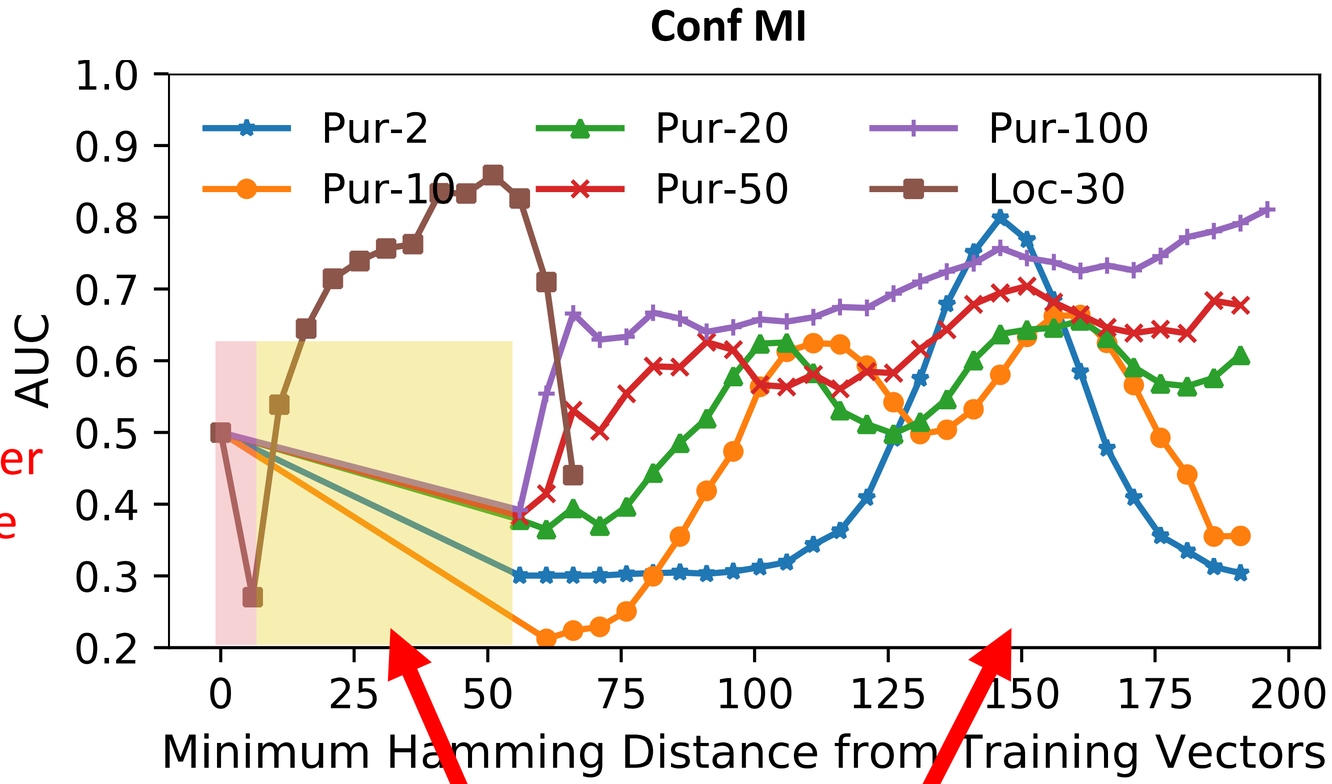
CIFAR Dataset in the paper

AUC computed between members and groupings of non-members at various distances.

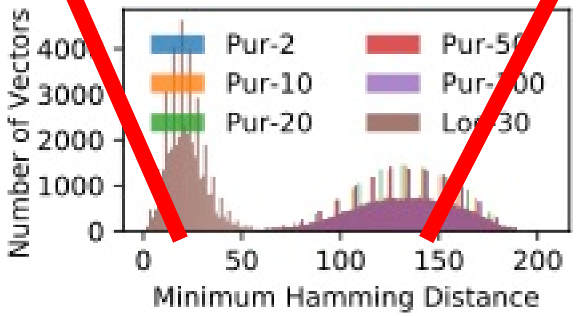
Hamming distance considered as Purchase and Location datasets are *Binary*

Evaluating Existing Membership Inference

No non-member vectors at close distances

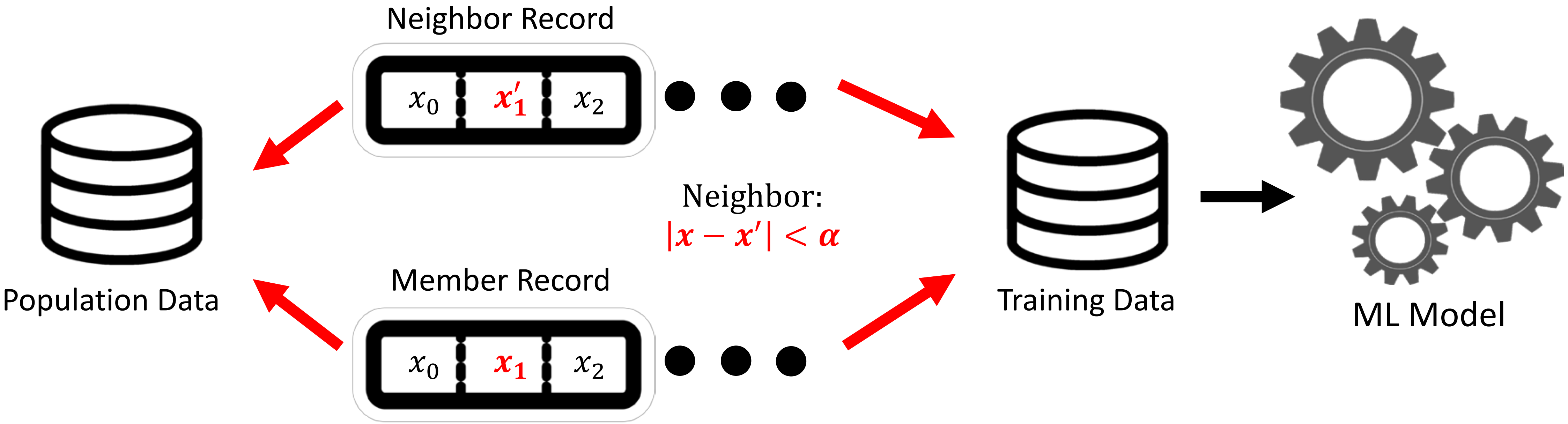


Vectors further away can be inferred more easily



The distribution of non-members from the population do not show the full picture.

Strong Membership Inference



- Infer if member vectors/neighbor vectors are in the training dataset
- Is there any *advantage* to inferring membership when in (member vectors) or out (neighbour vectors) of the training data.

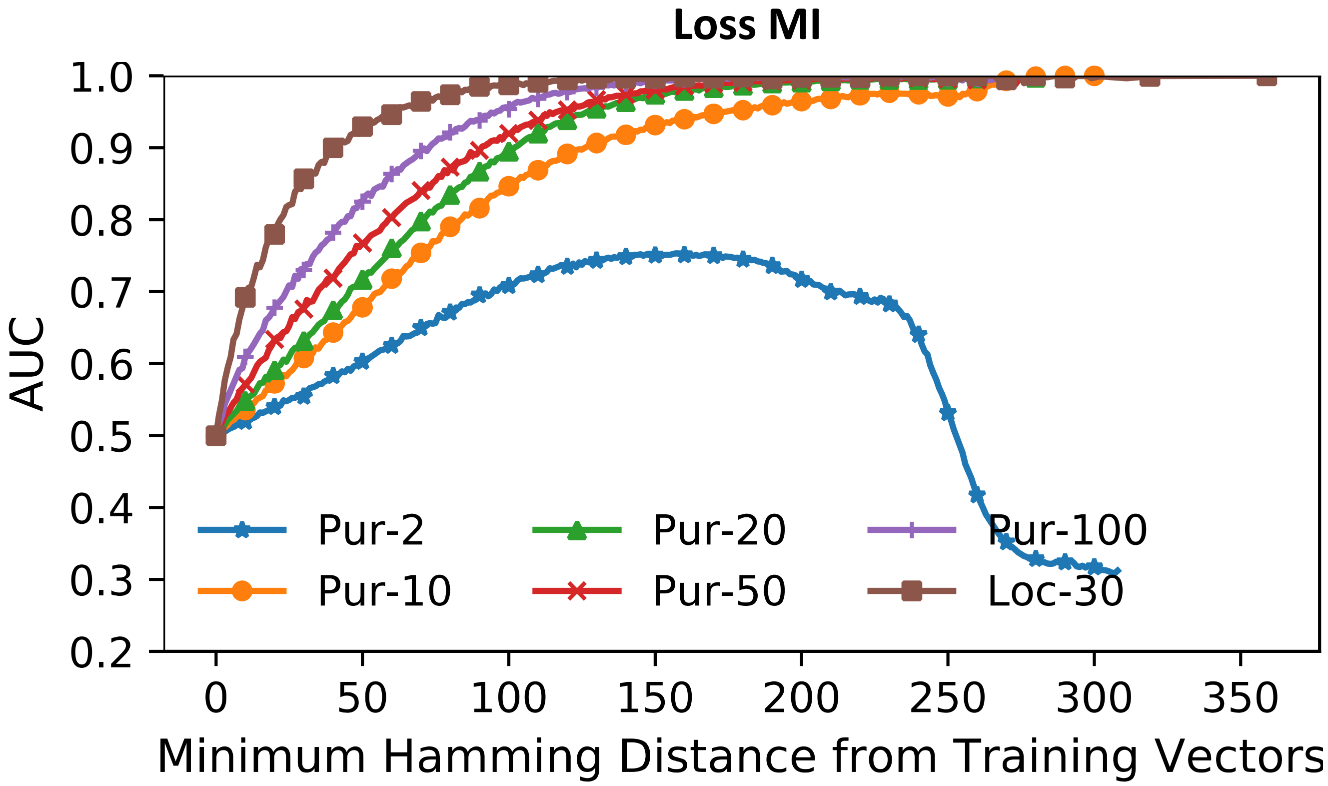
SMI Theoretical results

- A successful **Membership Inference** attack does not imply a successful **Strong Membership Inference** attack
  - (Theorem 1 in paper)

- **Strong Membership Inference ⇔ Attribute Inference**, assuming r-neighbour distinguishability holds
  - (Theorem 2 in paper)

Evaluating Strong Membership Inference

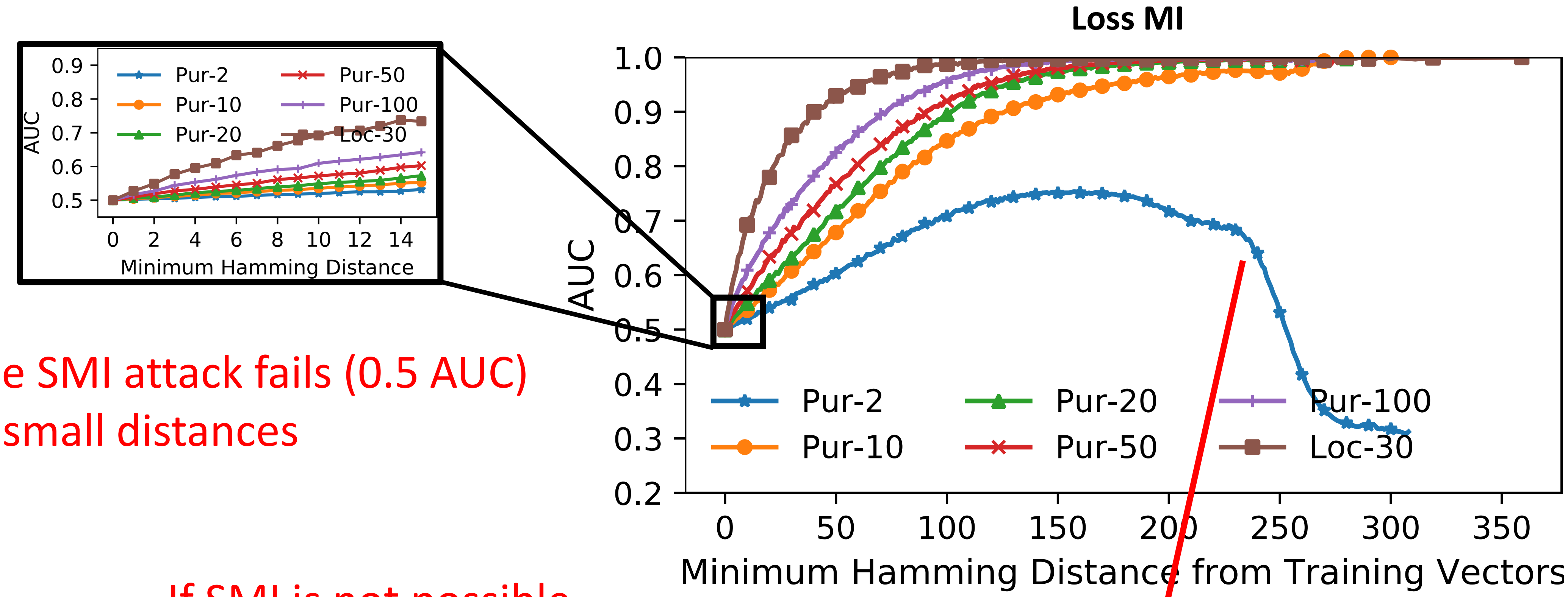
We perturb member vectors to deliberately produce off-distribution non-members.



MI AUC increases as distance increases

More classes in a dataset is more vulnerable to MI

Evaluating Strong Membership Inference

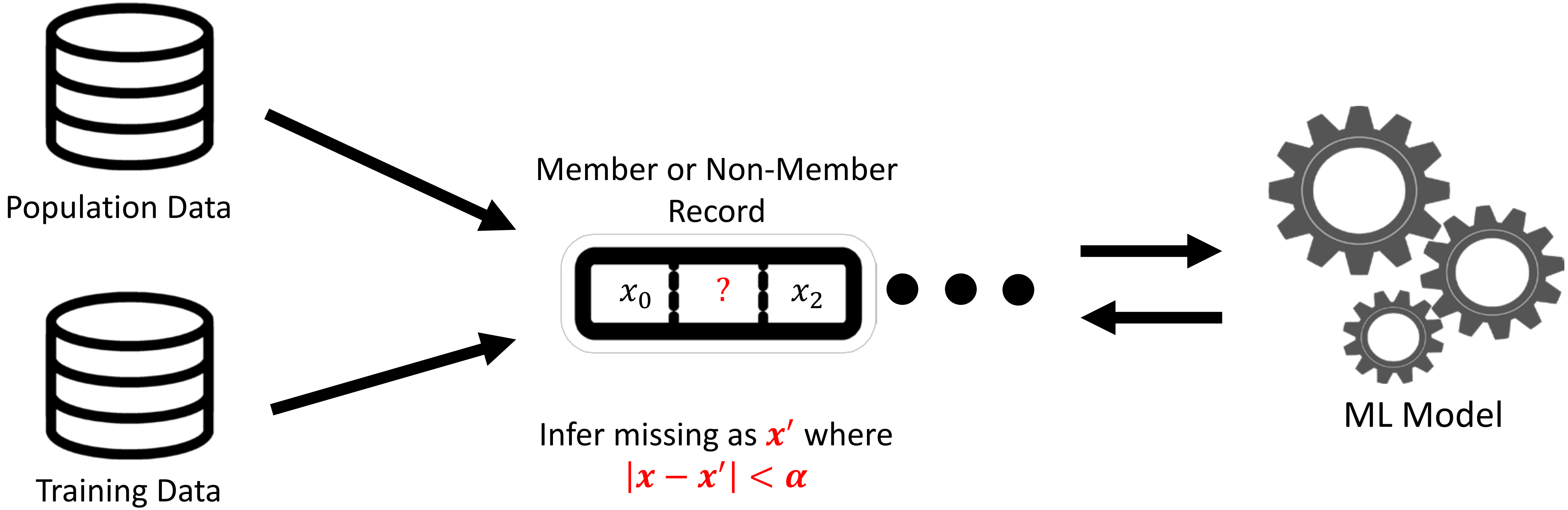


The SMI attack fails (0.5 AUC) at small distances

If SMI is not possible, then AI is not possible

Odd behaviour whereby the AUC decreases at extreme distances
- Dominant Class

Approximate Attribute Inference



- Infer approximate information on missing attribute(s) with access to the ML Model.
- Is there any *advantage* to inferring attributes when in or out of the training data.
  (Learning from the Distribution versus Learning from inclusion)

Evaluating Approximate Attribute Inference

AI	Loc-30	Pur-2	Pur-10	Pur-20	Pur-50	Pur-100
Conf	7.78E-4	1.38E-5	-3.69E-4	2.16E-4	2.00E-3	1.65E-3
Loss	7.76E-4	-9.79E-5	5.57E-3	6.69E-3	4.59E-3	5.09E-3
Shadow	8.00E-4	-2.00E-4	2.17E-3	2.63E-3	4.10E-3	4.20E-3

AAI	Loc-30	Pur-2	Pur-10	Pur-20	Pur-50	Pur-100
Conf	0.1609	0.0366	0.0516	0.0502	0.0958	0.1307
Loss	0.1030	0.0125	0.0516	0.0541	0.0789	0.1012
Shadow	0.0554	0.0054	0.0067	0.0149	0.0766	0.0964

Infer missing as x' where
$|x - x'| < \alpha$

We set α as 7.5, the distance of a random guess

It is possible to successfully infer approximate attributes significantly better than random guess when the target model is susceptible to membership inference.

Key Takeaways

1. It is difficult to infer exact attributes (AI), even if it is susceptible to MI.
2. Existing MI works consider datasets with vectors at large distances from each other.
3. The performance is close to a random guess (AUC = 0.5), for close non-members, problematic as SMI is needed for AI.
4. Dominating classes are less susceptible to MI and SMI attacks.
5. Observations of MI and SMI susceptibility is consistent across different ML architectures.
6. It is possible to approximately infer attributes (AAI), when susceptible to MI.
7. The more overfitted a target classification model, the more susceptible it is to AAI. AI remains difficult even with increased overfitting levels.

MACQUARIE University SYDNEY · AUSTRALIA

UNSW SYDNEY

CSIRO DATA 61

Australian Government
Department of Defence
Defence Science and Technology Group

Questions?

On the (In)Feasibility of Attribute Inference Attacks on Machine Learning Models

Benjamin Zi Hao Zhao[‡†], Aviral Agrawal[§*†], Catisha Coburn[¶], Hassan Jameel Asghar[*†], Raghav Bhaskar[†], Mohamed Ali Kaafar[*†], Darren Webb[¶], and Peter Dickinson[¶]
[*]Macquarie University, [‡]University of New South Wales, [†]Data61-CSIRO, [§]BITS Pilani K.K.Birla Goa campus, [¶]Cyber & Electronic Warfare Division, Defence Science and Technology Group, Australia

***Abstract*—With an increase in low-cost machine learning APIs, advanced machine learning models may be trained on private datasets and monetized by providing them as a service. However, privacy researchers have demonstrated that these models may leak information about records in the training dataset via membership inference attacks. In this paper, we take a closer look at another inference attack**

pact the attacks' likelihood and accuracy [27], [25], [34], [21], [31]. Our focus is on a related, and perhaps a more likely attack in practice, where the adversary with partial background knowledge of a target's record seeks to complete its knowledge of the missing attributes by observing the model's responses. This attack is called *model inversion* [5], [6], or in general *attribute inference* (AI) [34]. Yeom et al. [34] provide a formal definition of

Read more insights and details about our results.

https://arxiv.org/abs/2103.07101