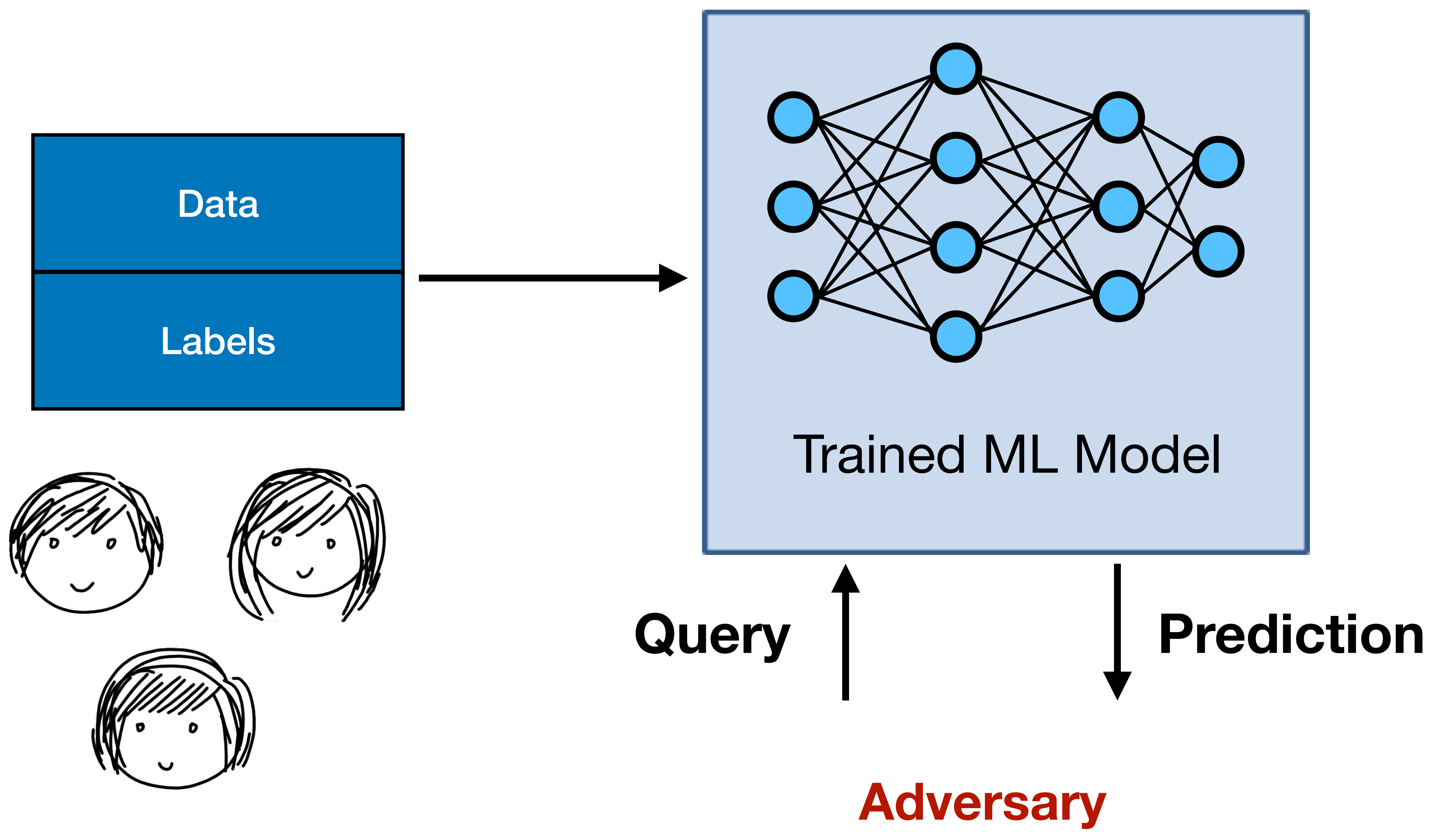
SNAP: Efficient Extraction of Private Properties with Poisoning

Harsh Chaudhari[*], John Abascal[*], Alina Oprea[*],
Matthew Jagielski[†], Florian Tramèr[‡], Jonathan Ullman[*]

[*]Northeastern University, [†]Google Research, [‡]ETH Zurich

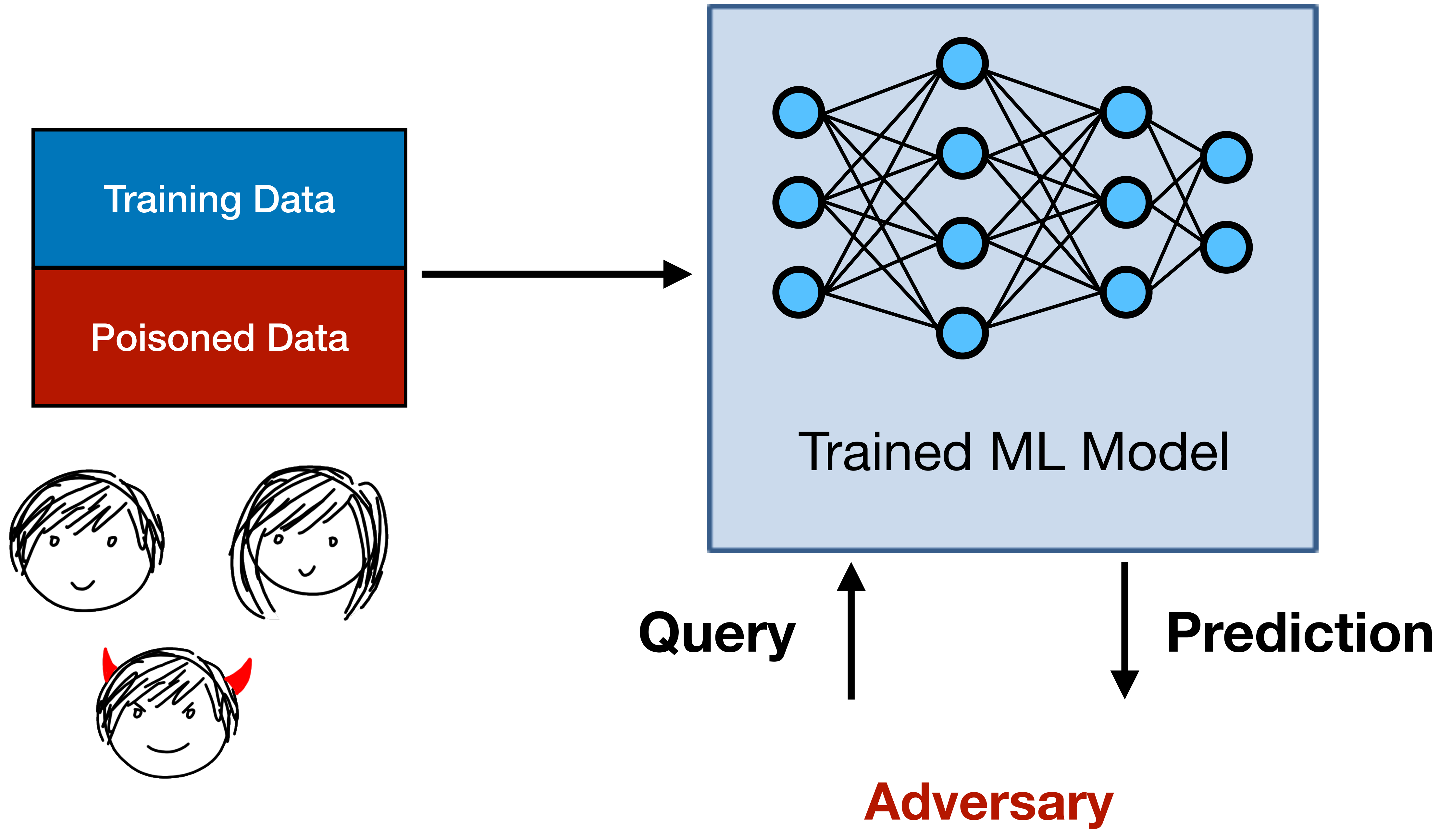
Privacy Attacks in Machine Learning



- **Membership Inference:** Determine if a data sample was present in the training set of the ML model [*SSS17, YGF18, CCN21*].
- **Attribute Inference:** Extract the missing attribute of a training record [*JE22, MDK22*].
- **Property Inference:** Learn properties of a group of individuals about the dataset [*GWY18, ZTO21, SE22*].

[SSS+17]: Shokri et al. Membership Inference against Machine Learning Models. IEEE S&P 2017.
[YGF+18]: Yeom et al. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. IEEE CSF 2018.
[CCN+21]: Carlini et al. Membership Inference Attacks from First Principles. IEEE S&P 2021.
[JE22]: Jayaraman et al. Are Attribute Inference Attacks Just Imputation? ACM CCS 2022.
[MDK+22]: Mehnaz et al. Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models. USENIX 2022.
[GWY+18]: Ganju et al. Property inference attacks on fully connected neural networks using permutation invariant representations. ACM CCS 2018.
[ZTO21]: Zhang et al. Leakage of dataset properties in Multi-Party machine learning. USENIX 2021.
[SE22]: Suri et al. Formalizing and estimating distribution inference risks. PETS 2022.

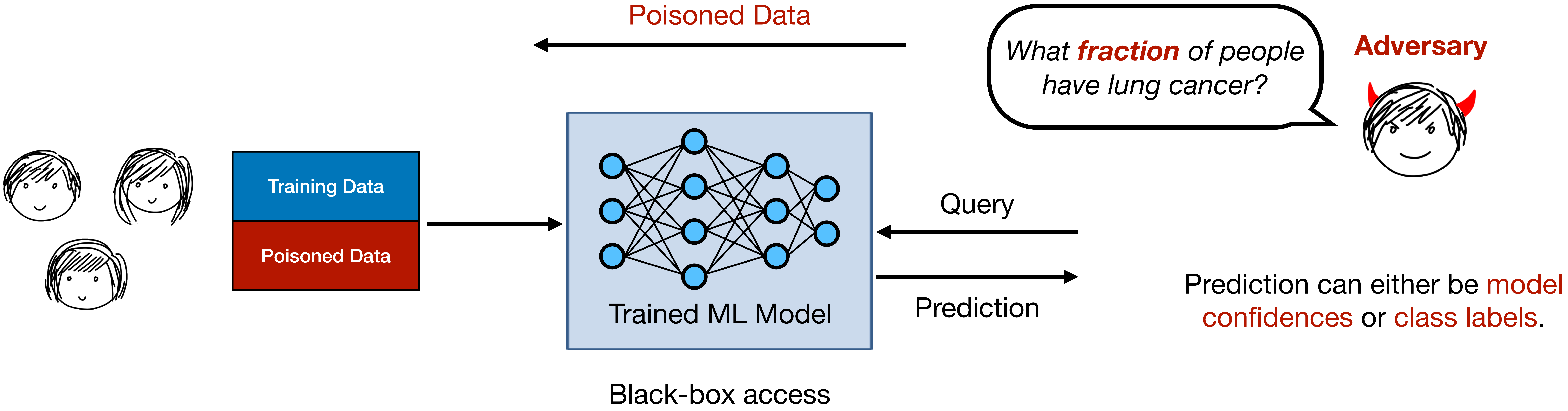
Amplifying Privacy Leakage with Poisoning



- **Membership Inference:** [*TSJ22, CSS22*] showed 8x better attack success than [*CCN21*].
- **Attribute Inference:** [*TSJ22*] showed 30x better attack success than [*MDK22*].
- **Property Inference:** [*MGC22*] showed 2x better attack success than [*GWY18*].

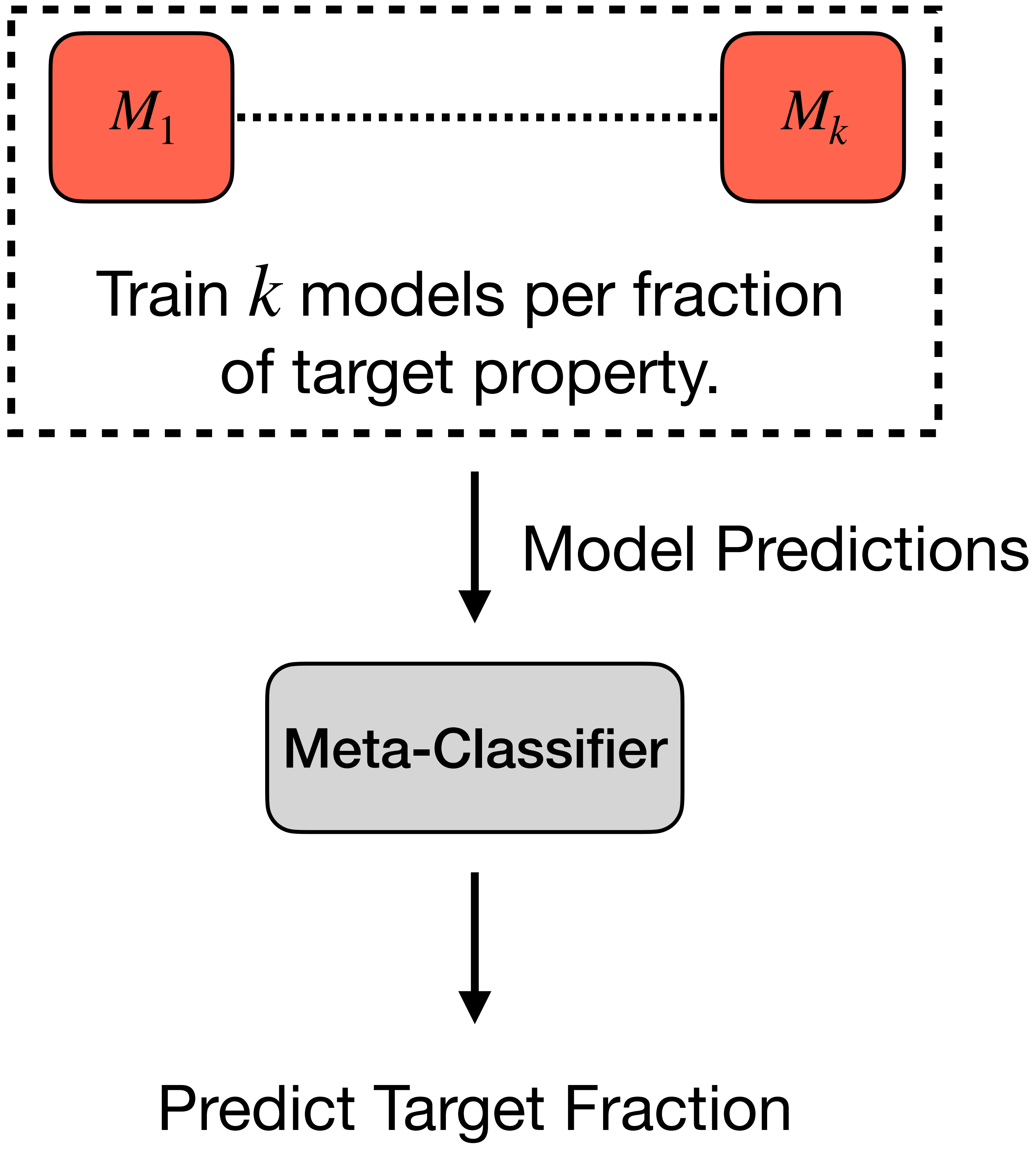
[TSJ+22]: Tramèr et al. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. ACM CCS 2022.
[CSS+22]: Chen et al. Amplifying Membership Exposure via Data Poisoning. NeurIPS 2022.
[MGC22]: Mahloujifar et al. Property inference from poisoning. IEEE S&P 2022.

Threat Model: Property Inference



The success of the adversary is measured by distinguishing between two fractions of the target property.

Limitations of [MGC 22]



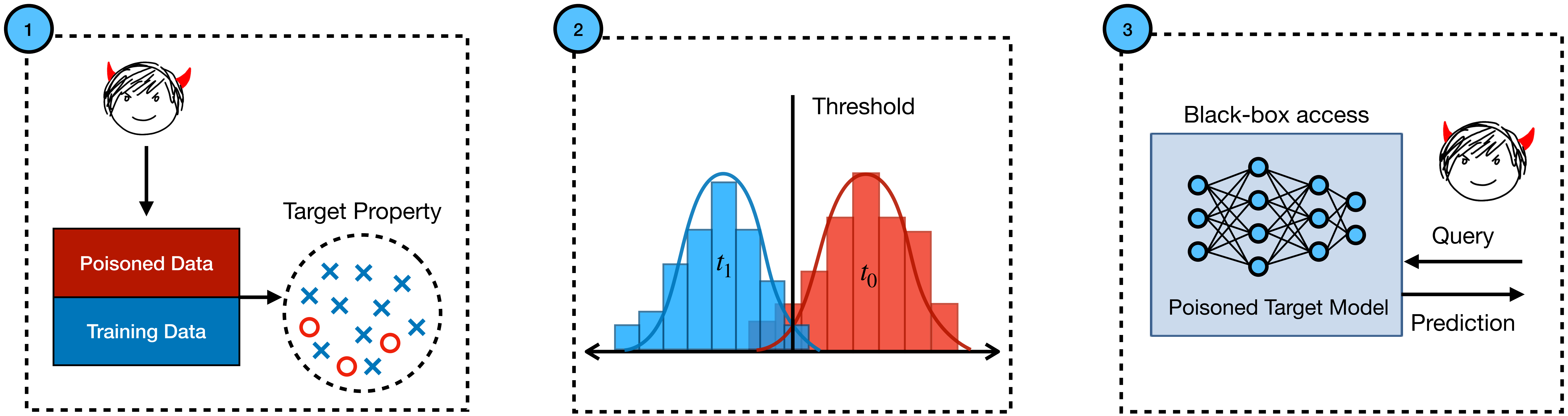
- **Drawbacks:**
  - **Computationally expensive** approach, requires training $\mathbf{k \approx 500}$ shadow models per fraction.
  - Requires a **large poisoning rate** for high attack accuracy.

[MGC22]: Mahloujifar et al. Property inference from poisoning. IEEE S&P 2022.

Our Contributions

- New property inference attack with poisoning: **SNAP**.
- Advantages: **34% higher attack accuracy**, **56x faster**, and **4-6x less poisoning** than prior work [*MGC22*].
- Backed by a **Theoretical Framework** for Model Confidence Learning under poisoning.
- Extensions: **Label-only**, **Property Existence**, and **Size Estimation**.
- Evaluation: Tested over **18 properties** with **attack accuracy** $\geq 90\,\%$ at **low** poisoning.

SNAP Attack Overview

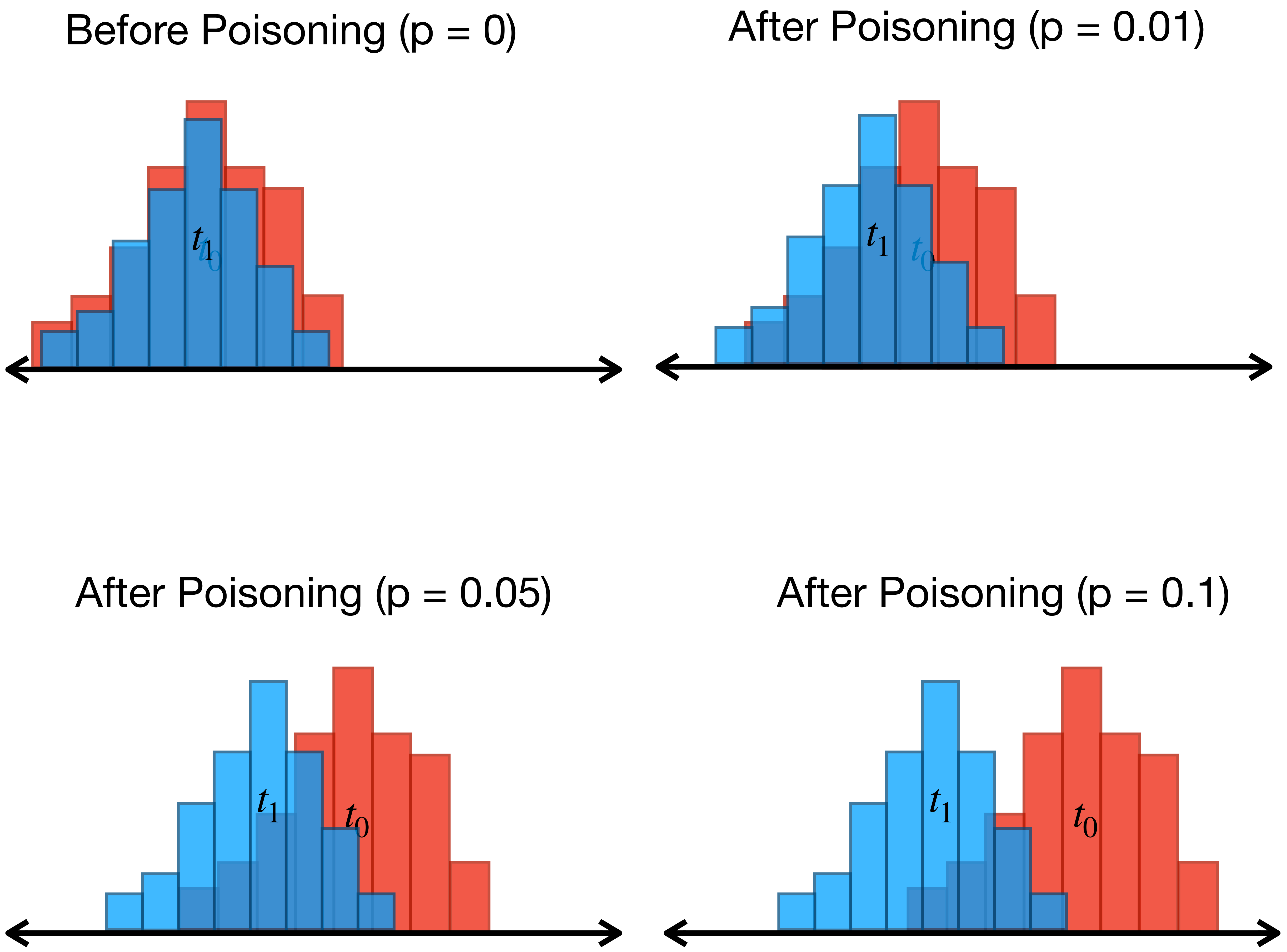


Data Poisoning

Model Confidence Learning

Distinguishing Test

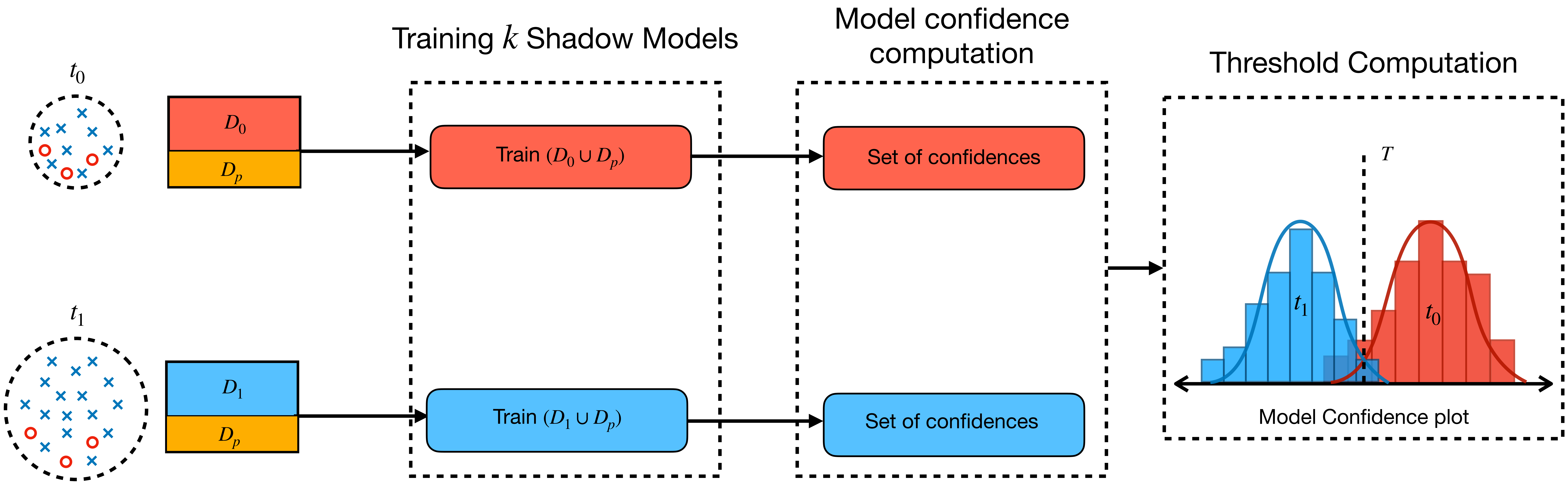
SNAP Attack: Insights



- Consider two fractions $t_0 < t_1$ of the target property.
- Poisoning disparately impacts distribution of confidences for the two fractions.
- Poisoning causes higher misclassification rate for fraction t_0.
- Confidence separation can be used as a tool for distinguishing test.
- Mount a subpopulation poisoning attack [JSH21].
- Theoretical analysis explaining the separation in confidences.

[JSH+21]: Jagielski et al. Subpopulation data poisoning attacks. ACM CCS 2021.

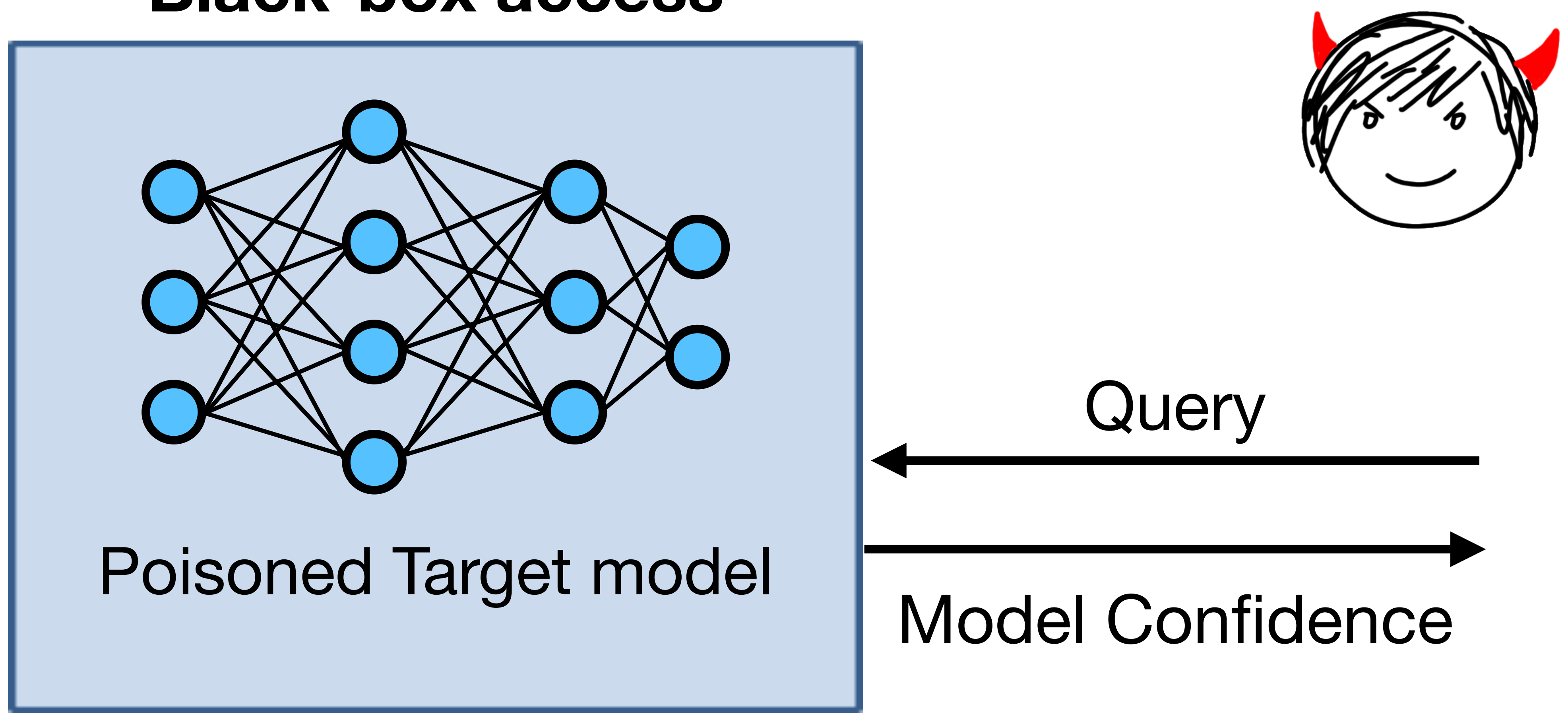
SNAP Attack: Model Confidence Learning



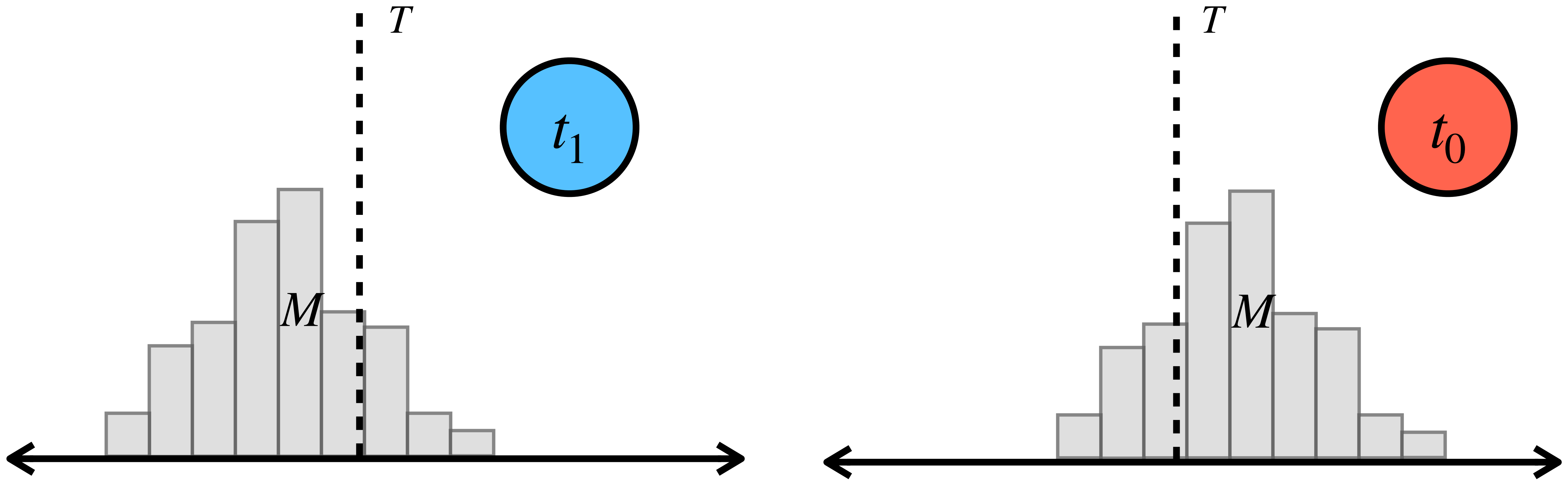
- [*MGC22*] requires $\mathbf{k \approx 1000}$ shadow models to train a meta-classifier for the distinguishing test.
- Our attack directly learns model confidences requiring $\mathbf{k \leq 8}$ for the distinguishing test.

SNAP Attack: Distinguishing Test

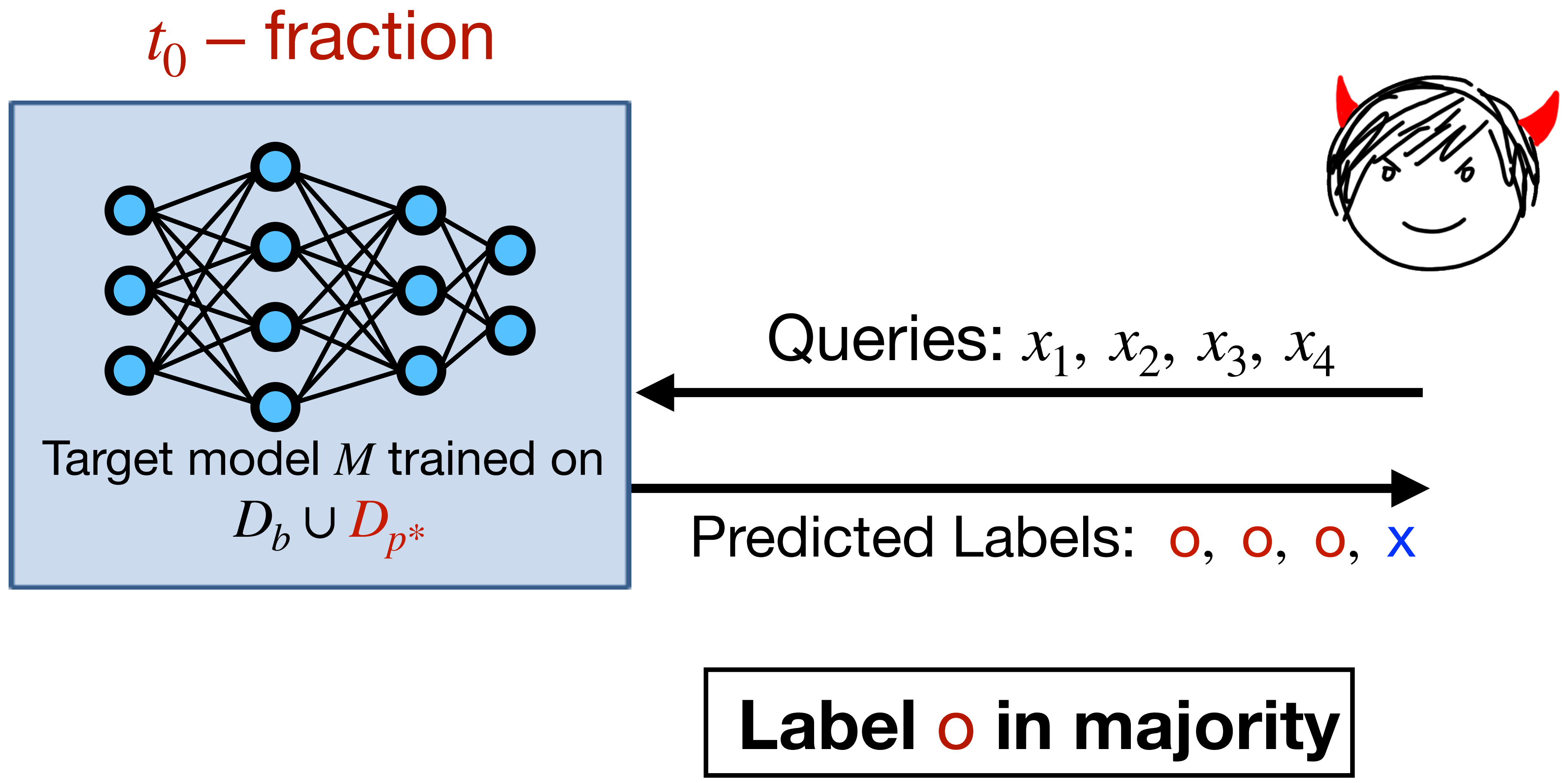
Black-box access

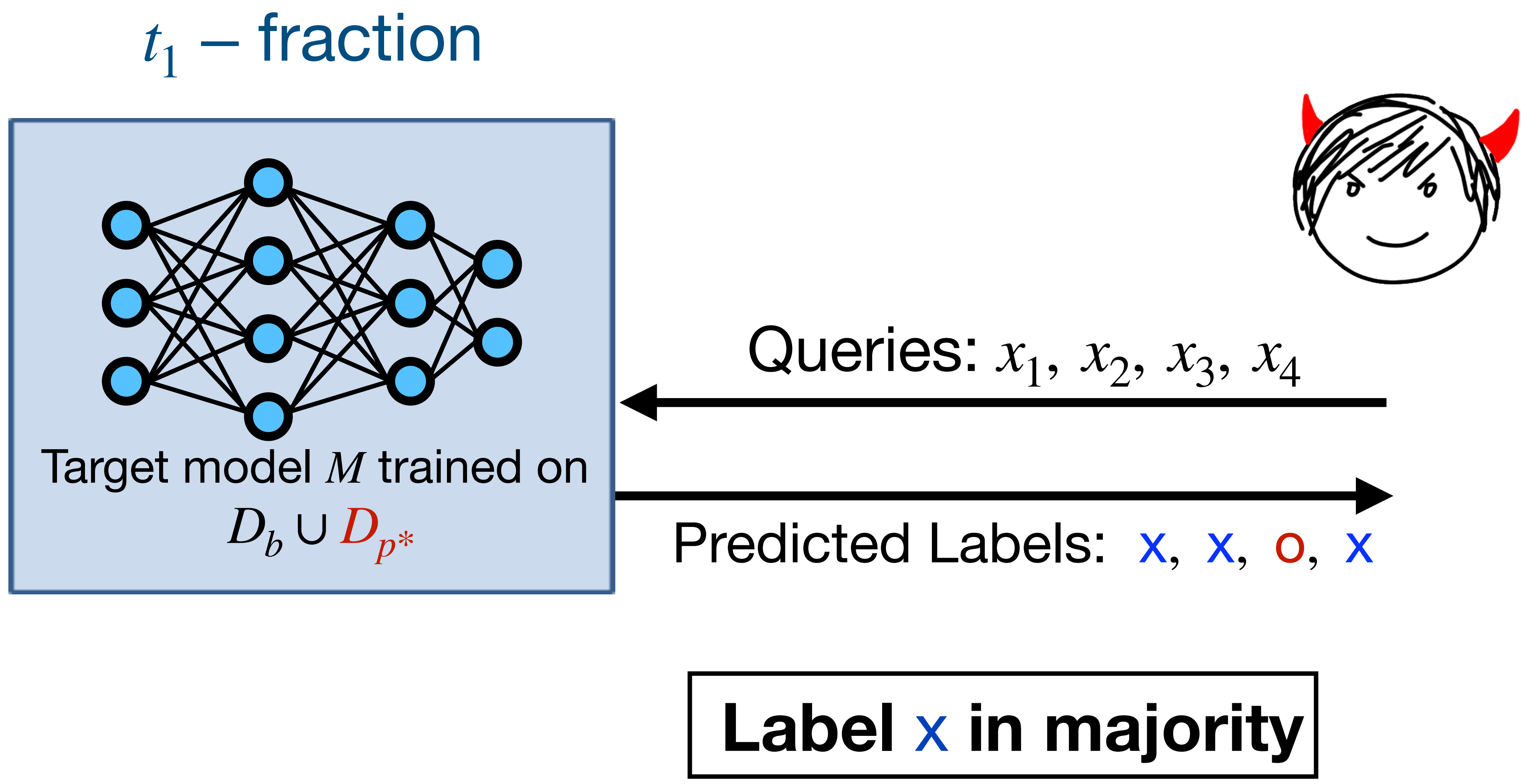


- Query the model on samples with the target property and obtain confidences.
- We provide **analysis on the total queries** attacker needs to succeed in the test.



SNAP Extension: Using Class Labels



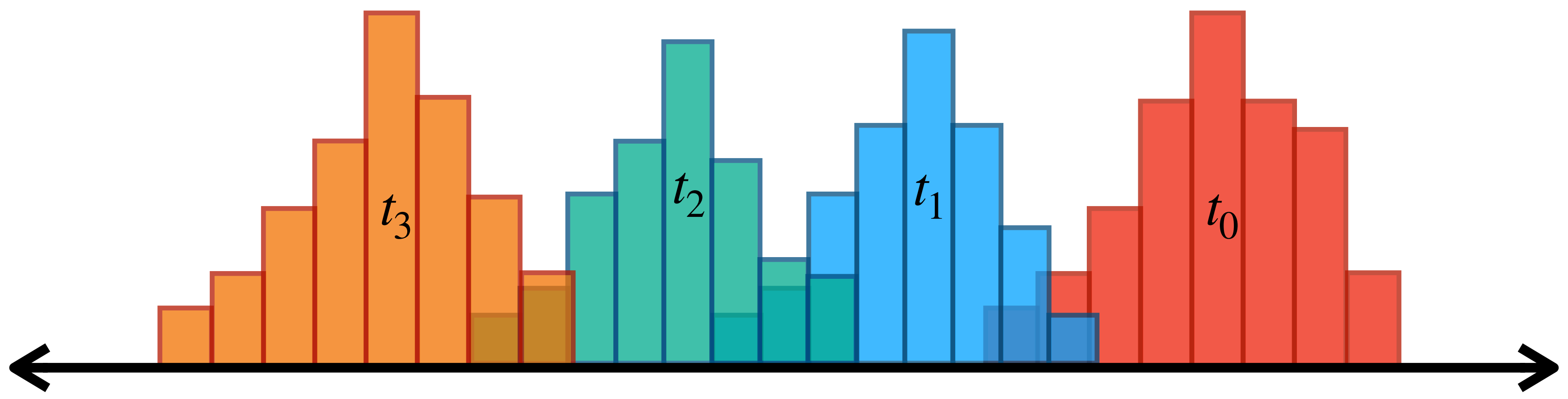


Key Insight:

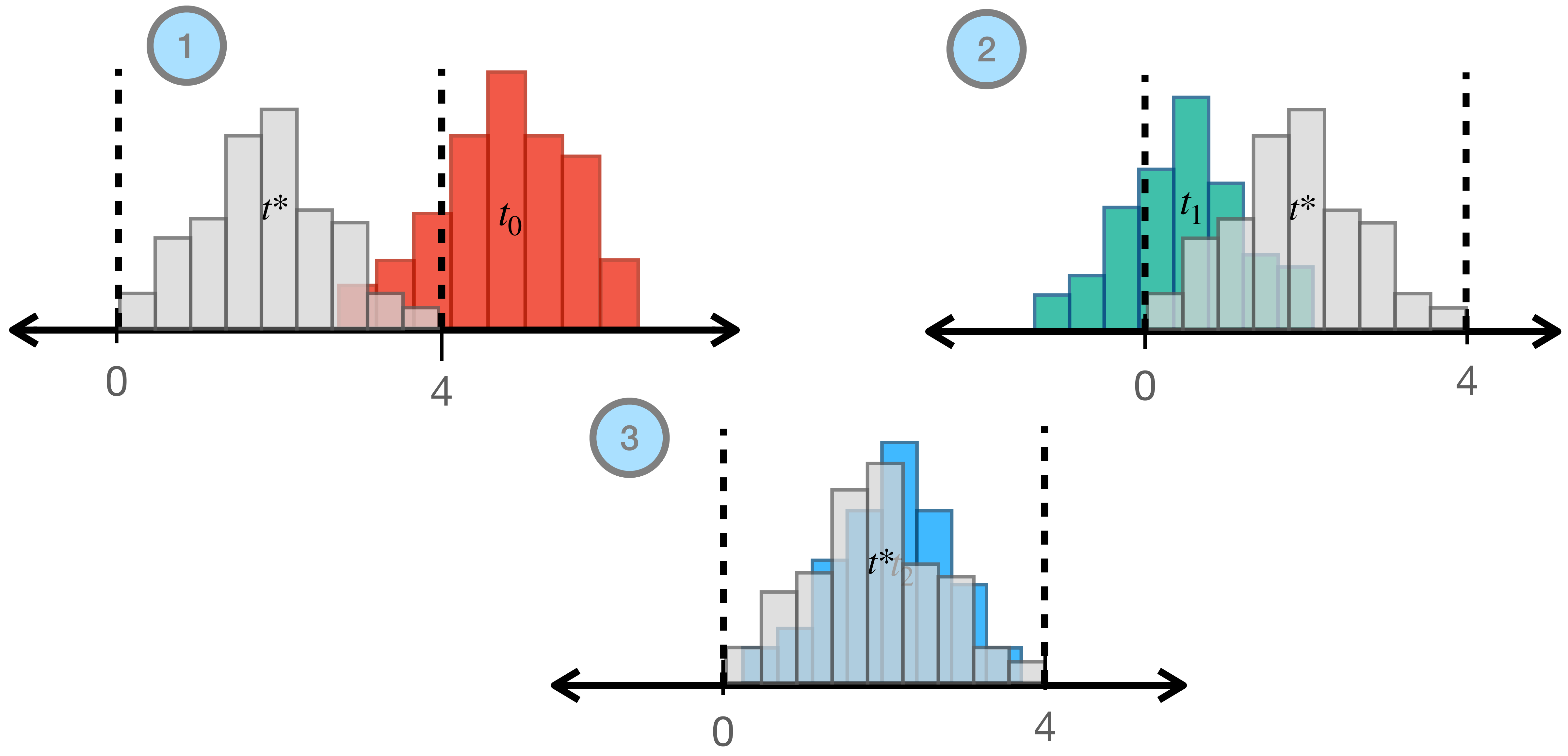
- Choose a poisoning rate p* such that
  - Labels for majority of the samples queried on a model trained on t_0 – fraction flips to the target label.
  - Labels for majority of the samples queried on a model trained t_1 – fraction stays the original label.
- p* is computed by analyzing the behavior of confidence distribution for the two fractions.

SNAP Extension: Size Estimation

Target Property Sizes: $0 \leq t_0 < t_1 < t_2 < t_3 \leq 1$



Size Estimation Algorithm



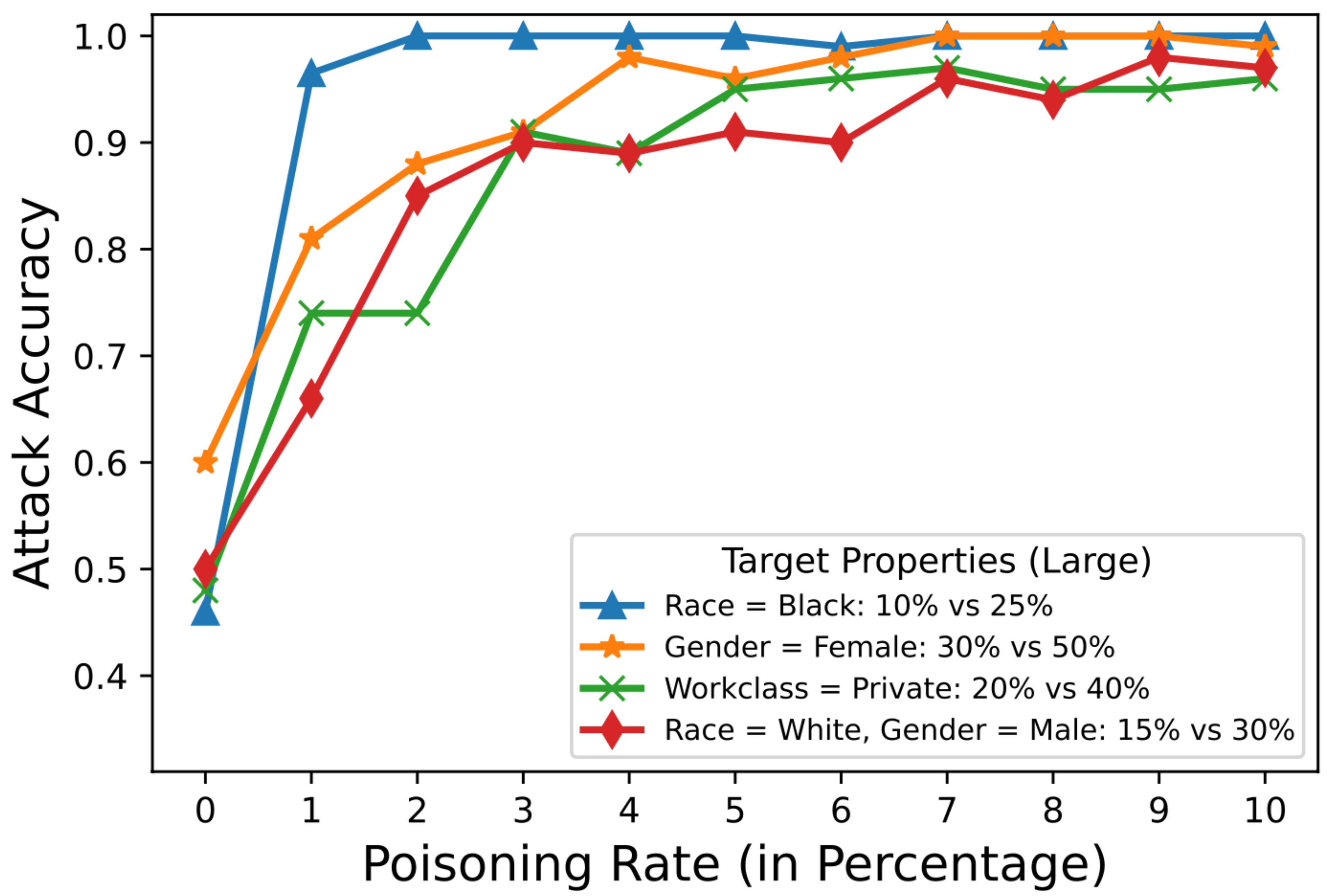
- **Insights:**
  - Very **realistic threat model**, adversary does not have knowledge of the target fraction t^*.
  - Given a fixed poison rate, the distributions follow a **strict size ordering**.
  - Given a set of ordered elements, we can exploit **binary search** to find the target fraction.
  - Previous approaches [*SE22, MGC22*] required $\mathbf{k \approx 20,000}$ shadow models to perform size estimation.
  - Our approach **exponentially drops** the number of shadow models to $\mathbf{k \leq 14}$.

Evaluation

- **Datasets:** Adult, Census, Bank Marketing and CelebA.
- **Target Properties:** We test on18 different target properties. Three broad categories:
    - **Large-sized:** Target property $> 10\,\%$ of the training set.
    - **Medium-sized:** $1\,\% \leq$ Target property $\leq 10\,\%$ of the training set.
    - **Small-sized:** Target property $< 1\,\%$ of the training set.
- **Model Architectures:** Feed-forward Neural Network and ResNet-18.
- **Evaluation Metric:**
    - **Attack Accuracy:** Accuracy of correctly distinguishing which fraction of the target property the model was trained on.
    - **Total Execution Time:** End-to-end running time to mount the attack.

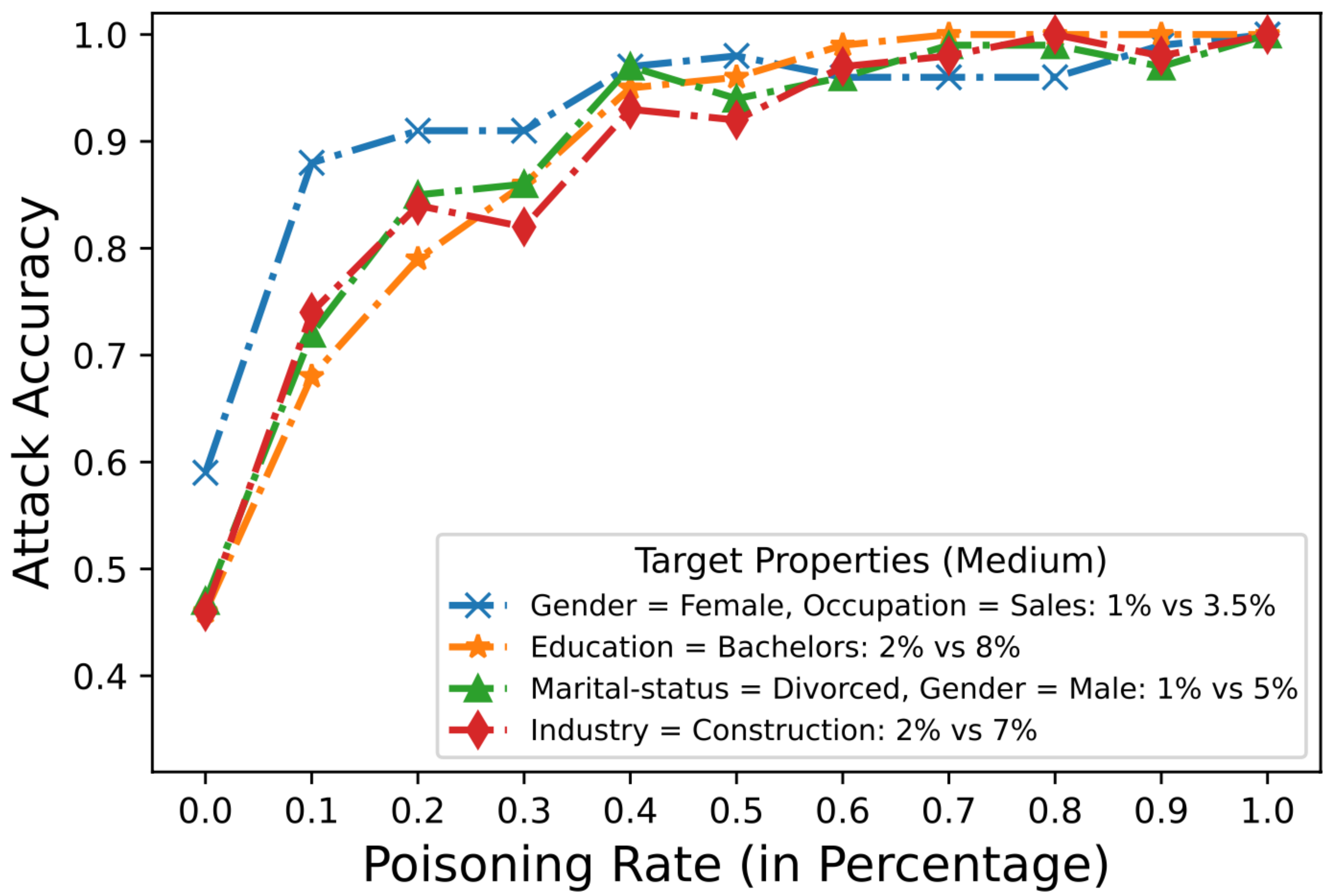
Evaluation of SNAP on Adult and Census

Large Properties $(\geq 10\%)$



Attack Accuracy $\geq 90\%$ with 5% poisoning.

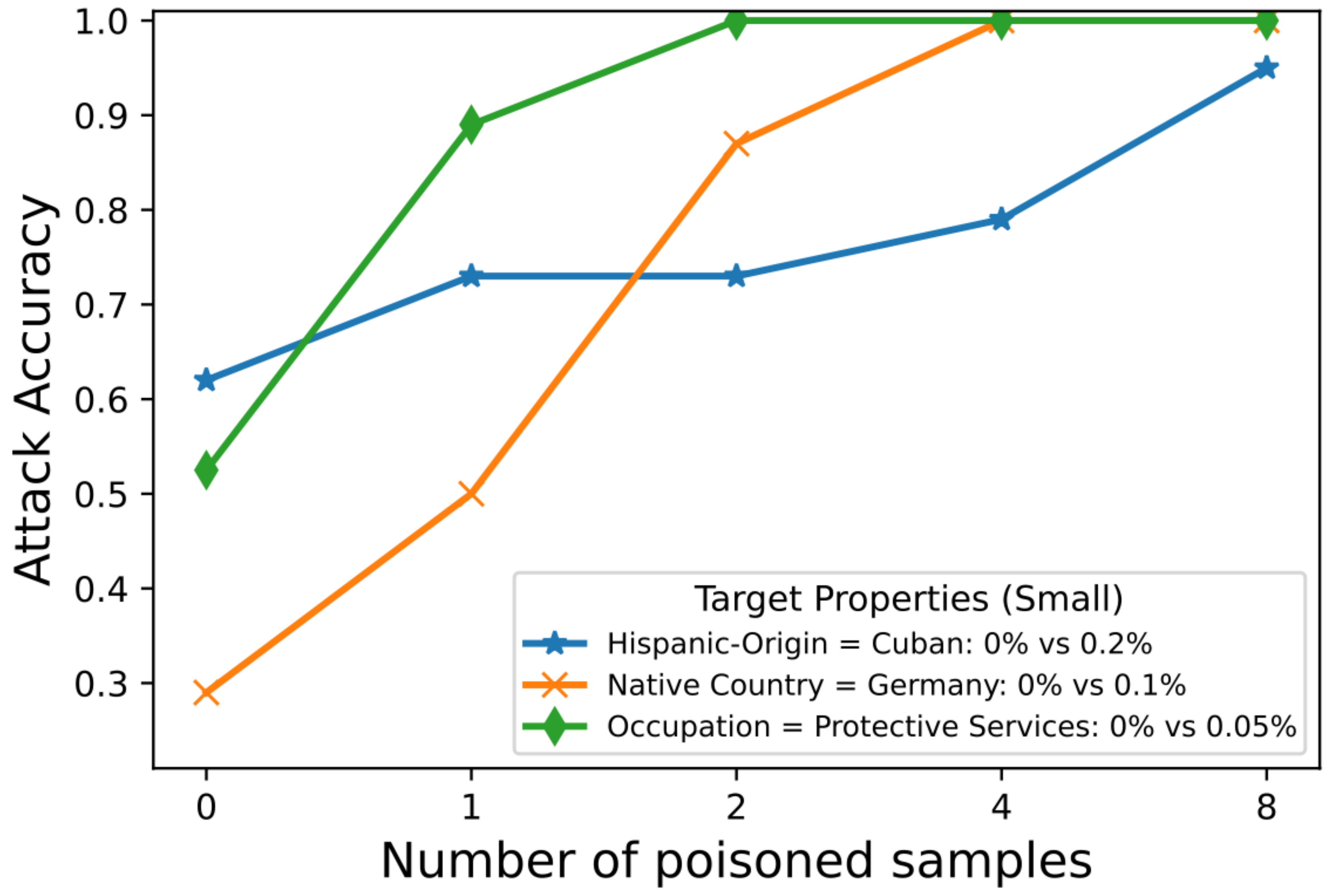
Medium Properties $(1\% - 10\%)$



Attack Accuracy $\geq 90\%$ with only 0.6% poisoning.

SNAP Extensions

Property Existence: Generalization of Membership Inference where $t_0 = 0$ and $t_1 > 0$



Attack Accuracy $\geq 90\%$ with **8** samples.

Property Size Estimation: Estimating the size of the target property.

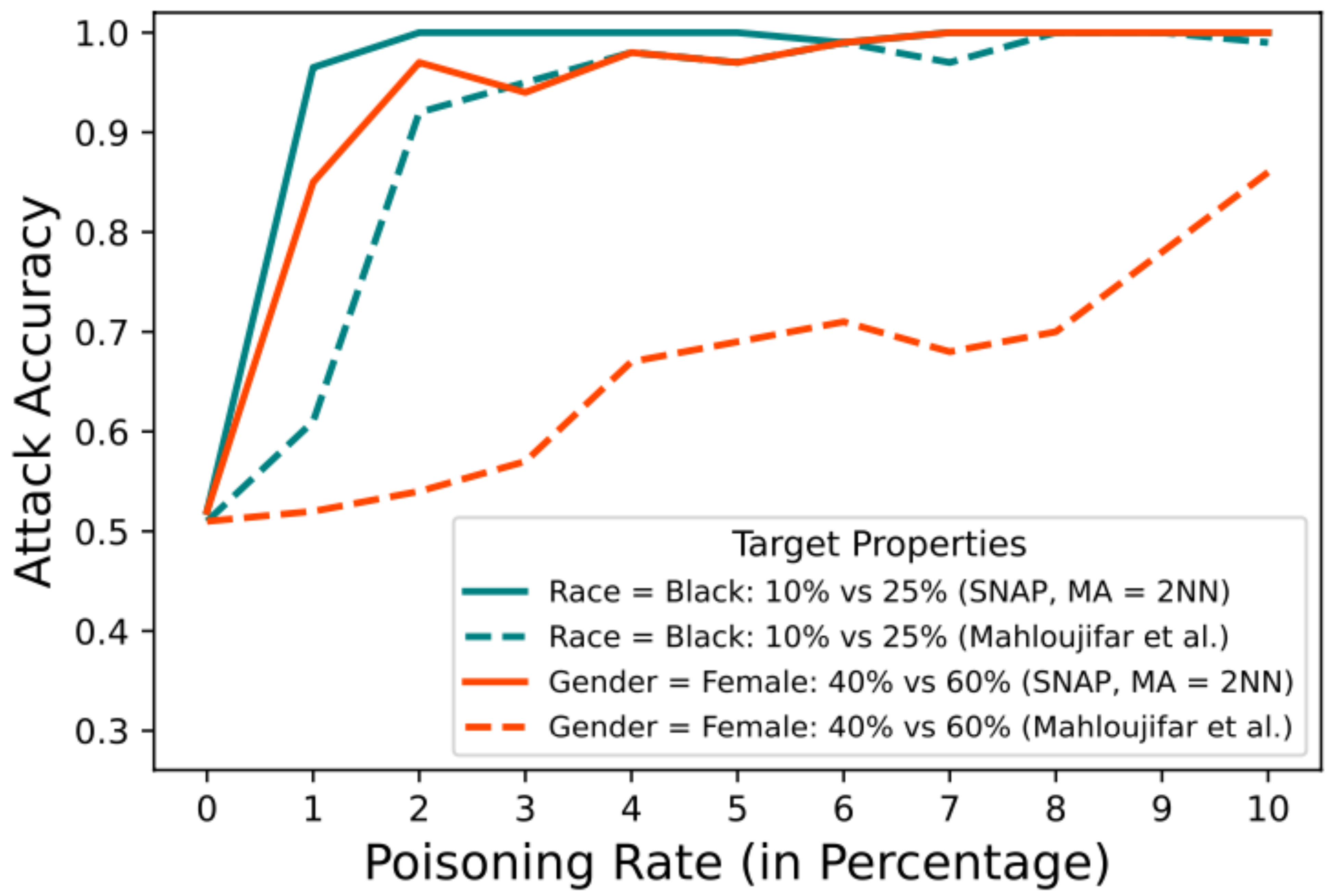
Medium sized Property	*Actual Size*	*Our Estimate*	
		0% poison	*1% poison*
Construction	**3%**	32.6%	**3.1%**
Female Sales	**3.9%**	24.9%	**4.3%**

Large sized Property	*Actual Size*	*Our Estimate*	
		0% poison	*5% poison*
White Male	**43%**	31.5%	**40%**
African-American	**10.2%**	17.5%	**9.3%**

Accurate estimation with low poisoning.
($\mathbf{1}\ \%$ for medium and $\mathbf{5}\ \%$ for large)

SNAP Comparison to [MGC 22]

SNAP using Model Confidences



- Achieves **34 %** higher attack accuracy than [MGC22].
- **56 ×** faster than [MGC22].
- Requires **4 – 6 ×** less poisoning than [MGC22].

SNAP using only Class Labels

Target Property	*Poisoning Rate*	*[MGC22]*	*SNAP*
White Male	5.7%	65%	**95%**
Private Sector	1.1%	56%	**94%**
Female	4.5%	70%	**98%**
African American	3.7%	97%	**100%**

Consistently **outperforms** [MGC22] and **same efficiency** benefits as our confidence attack.

DP-SGD as a Defense?

- Differential Privacy is traditionally designed to protect an individual sample’s privacy.
- DP is not intended to provide defense against property inference.
- Empirical confirmation on DP-SGD failing to prevent property inference attacks.

Target Property	*Attack Accuracy*			
	$\epsilon = 8$	$\epsilon = 4$	$\epsilon = 2$	$\epsilon = 1$
White Male	95%	98%	90%	75%
African-American	100%	100%	100%	98%

Conclusion

- We propose a novel property inference attack that is **more efficient**, requires **less poisoning** and has **higher attack accuracy** than previous work [*MGC22*].
- We provide a **theoretical framework** explaining the effectiveness and efficiency of our SNAP attack.
- We extend our attack to incorporate **label-only**, **property existence** and **property estimation** attacks.
- **Defending** against property inference attacks is still an **open problem**. We empirically evaluate and show that Differential Privacy is not enough to prevent property inference attacks.

Code

Thank You

chaudhari.ha@northeastern.edu