CHAMELEON

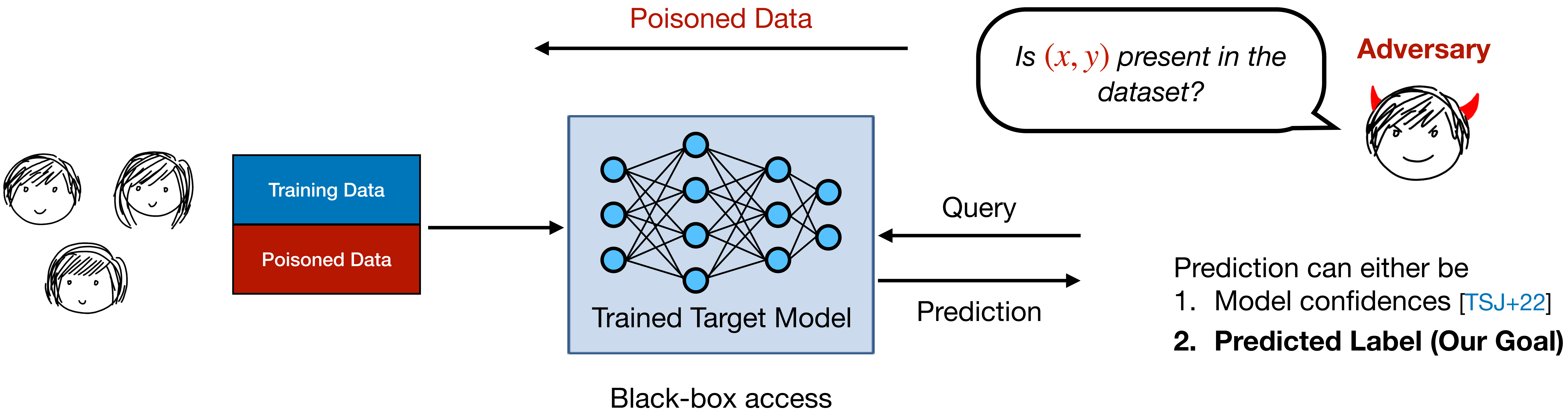
Increasing Label-Only Membership Leakage with Adaptive Poisoning

Harsh Chaudhari[*], Giorgio Severi[*], Alina Oprea[*], Jonathan Ullman[*]

[*]Northeastern University

Threat Model: Membership Inference

Infer if a challenge point (x, y) is present in the training set by querying the ML model.

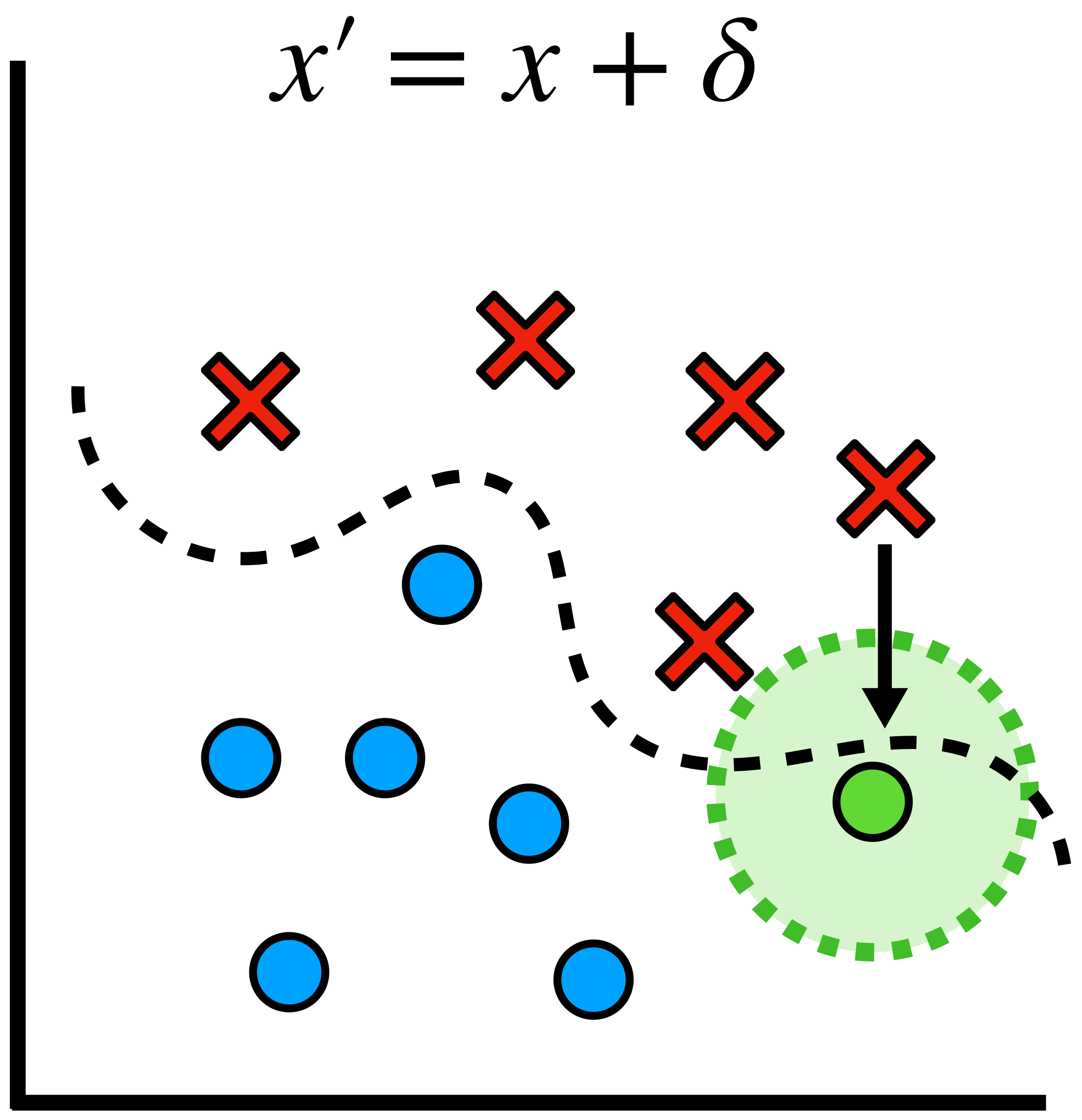


The success of the adversary is measured by achieving a **high TPR in a low FPR** regime.

[TSJ+22]: Tramèr et al. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. ACM CCS 2022.

Existing Label-Only Membership Inference Attacks

- **Gap Attack** [*YGF18*]**:**
    - Predicts misclassified point as a Non-Member.
    - Requires only one query to the target model for each challenge point.
- **Decision-Boundary Attack** [*CTC21, LZ21*]**:**
    - Uses a sample's distance from the Decision-Boundary (DB) to determine its membership status. Distance is measured using adversarial examples [*BRB18, CJW20*].
    - Works under the assumption that **non-members** lie closer to the Decision Boundary compared to **members**.
    - **Computationally expensive** approach, requires $\approx \mathbf{2000}$ queries to the target model for each point.

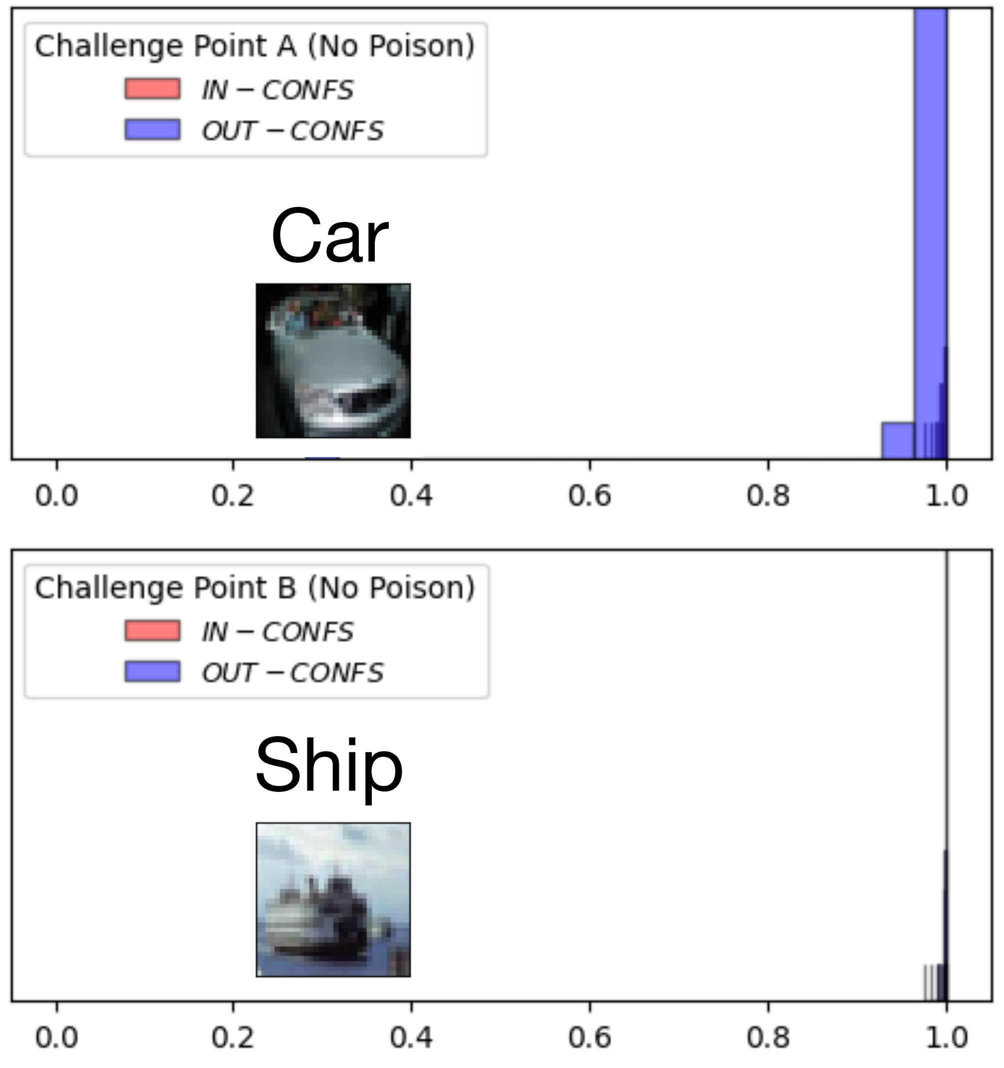


[YGF+18]: Yeom et al. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. IEEE CSF 2018.
[CTC+21]: C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot. Label-only membership inference attacks. ICML 2021.
[LZ21]: Z. Li and Y. Zhang. Membership leakage in label-only exposures. ACM CCS 2021.
[BRB18]: W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. ICLR 2018.
[CJW20]: J. Chen, M. I. Jordan, and M. J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. IEEE S&P 2020.

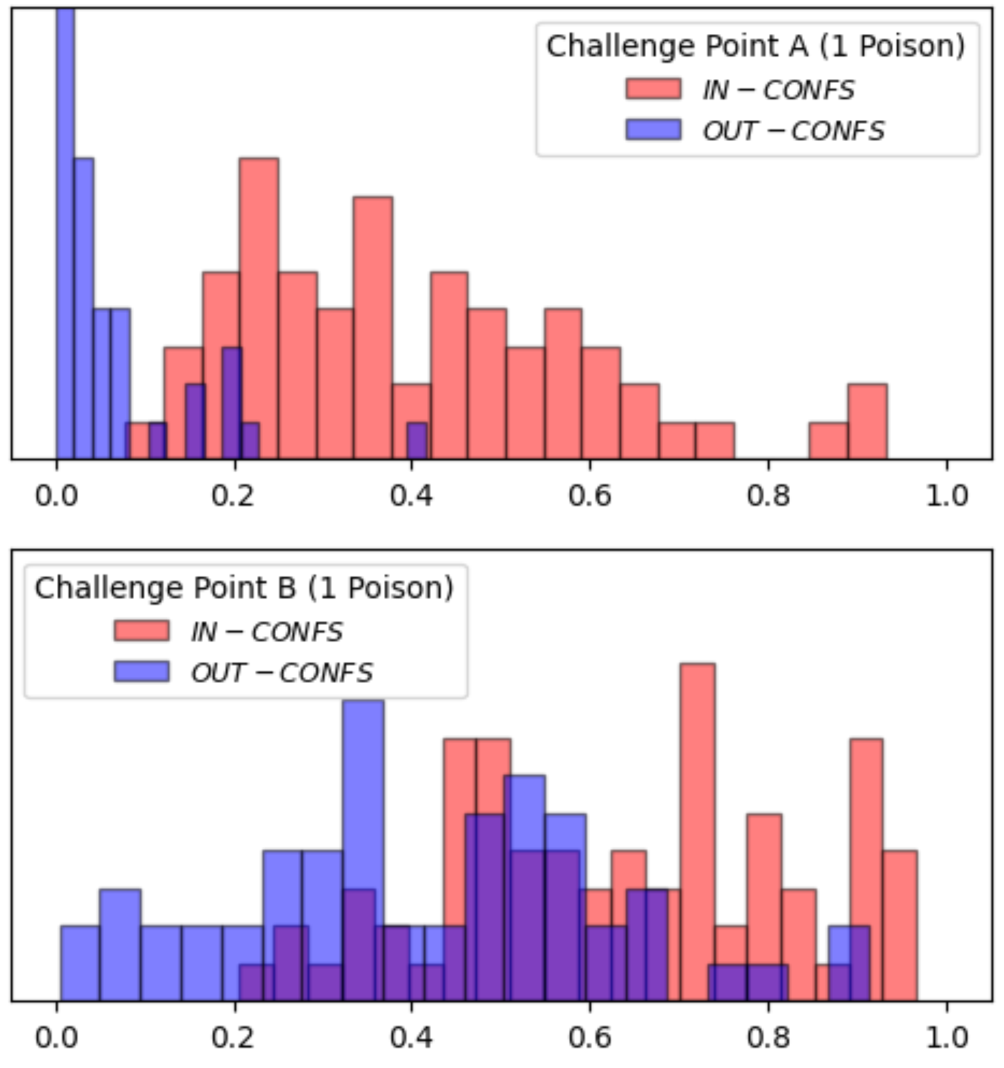
Our Contributions

- **Existing** Label-Only Membership Inference attacks **Fail** in the low False Positive Rate regime.
- New Label-Only MI attack **CHAMELEON** that uses **Adaptive Poisoning** and **Membership Neighborhood** strategies to succeed in the low FPR regime.
- Advantages: **17.5x** higher TPR at 1%FPR than prior work [*CTC21*,*LZ21*], while requiring **39x** fewer queries.
- Provide a **Theoretical Analysis** to understand Impact of poisoning on our MI attack.
- Comprehensive Evaluation: Tested on **4 Datasets** over **6 Model Architectures.**

Chameleon: Insights



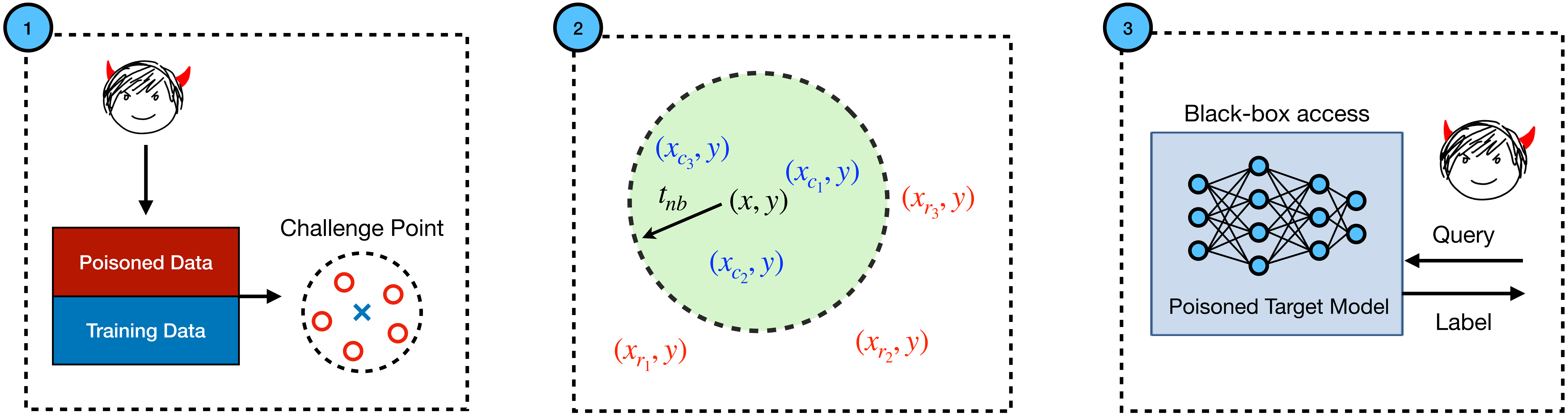
No Poisoning



Under Poisoning
(1 poisoned sample)

- Non poisoned models, whether (x, y) was in the training set (IN) or not (OUT), will likely correctly classify (x, y).
- If over-poisoning, both IN and OUT models will likely missclassify the challenge point.
- Add enough poisoned points, such that IN models correctly classify while OUT models misclassify the challenge point.
- Each challenge point requires different amount of poisoned points.
- Find neighbors of the challenge point with similar poisoning-induced behavior to enhance attack success.

Chameleon Attack: Building Blocks



Adaptive Poisoning **Membership Neighborhood** **Distinguishing Test**

Comparison with Prior Work on CIFAR-100

Label-Only Attack	*TPR@0.1%FPR*	*TPR@1%FPR*	*TPR@5%FPR*	*AUC*	*MI Accuracy*
Gap [*YGC18*]	0%	0%	0%	73.8%	73.8%
Decision-Boundary (DB) [*CTC21, LZ21*]	0.02%	3.6%	23.0%	84.9%	81.1%
Chameleon (Ours)	**29.6%**	**52.5%**	**70.9%**	**92.6%**	**85.2%**

- Achieves **370x** and **17.5x** higher TPR than DB attack at 0.1% and 1% FPRs respectively.
- Also **improves** upon the (average case) **AUC and MI Accuracy** metrics.
- **39x** more **query efficient** than DB when mounting the attack.

Conclusion

- We show that prior Label-Only MI attacks [*CTC21, LZ21*] **fail** in the low FPR regime.
- We propose a novel Label-Only MI attack that uses **adaptive poisoning** and **membership neighborhood** strategies to achieve **High TPR**.
- We also provide a **theoretical analysis** explaining the impact of data poisoning on Label-Only MI.
- Differential Privacy can be used an **effective defense** against our Chameleon attack, but comes at the **expense** of model utility.

Thank You

chaudhari.ha@northeastern.edu