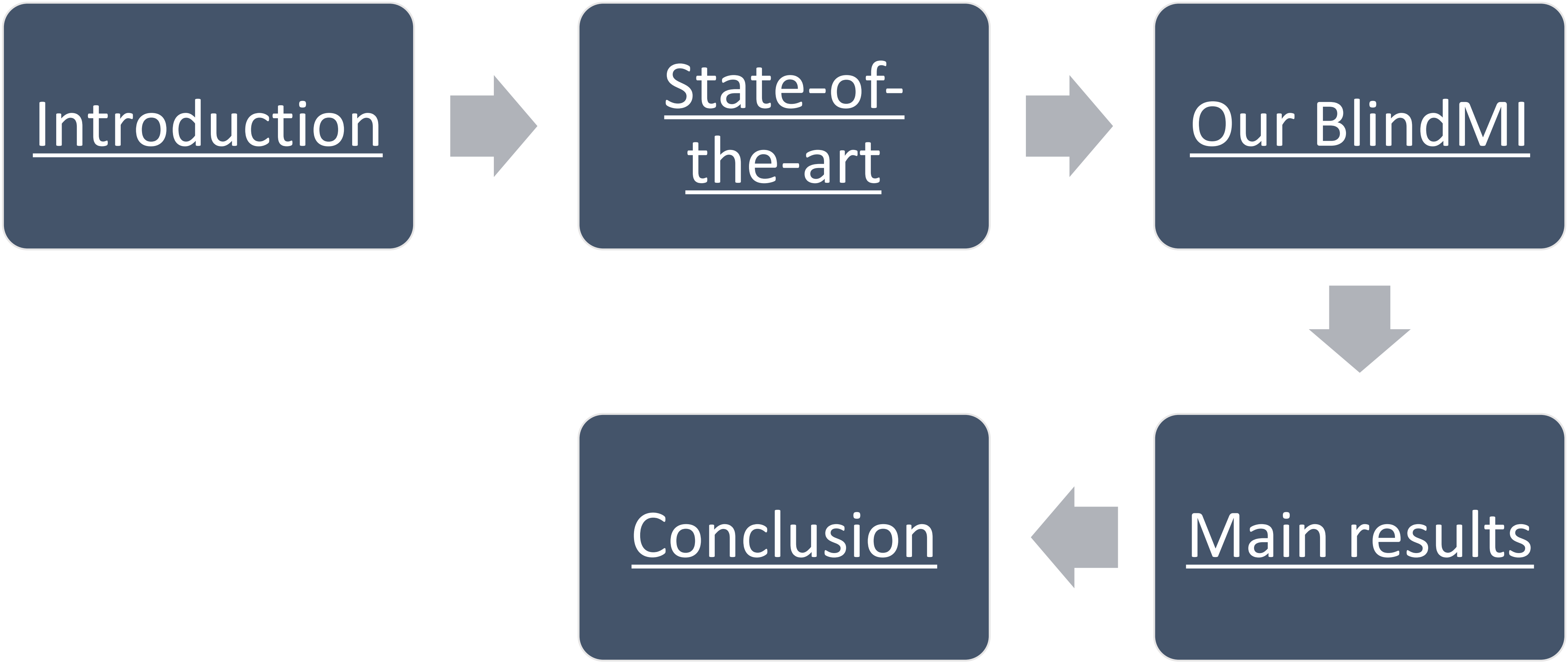
Practical Blind Membership Inference Attack via Differential Comparisons

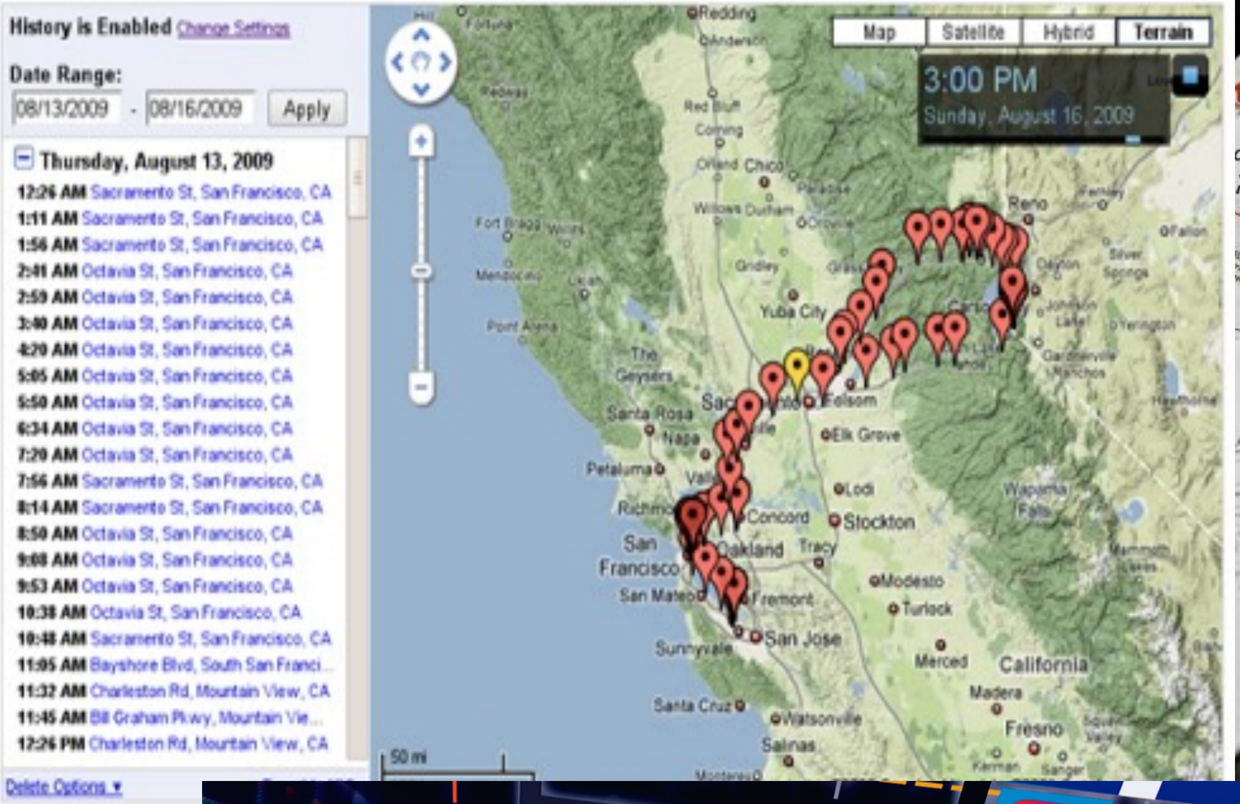
Bo Hui[1]*, Yuchen Yang[1]*, Haolin Yuan[1]*, Philippe Burlina[2], Neil Zhenqiang Gong[3], Yinzhi Cao[1]

[1]The Johns Hopkins University [2]The Johns Hopkins University Applied Physics Laboratory [3]Duke University

*Equal contribution

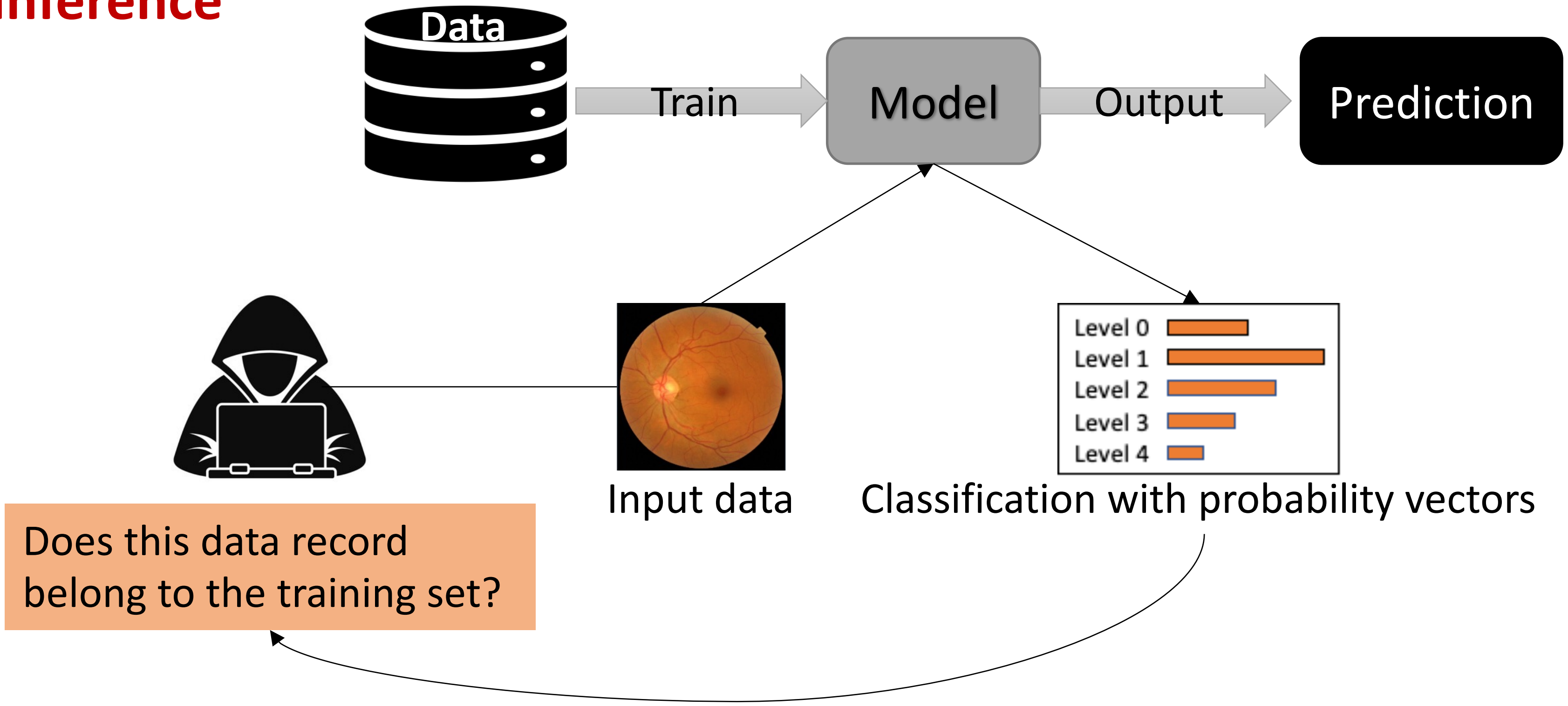

Introduction
State-of-the-art
Our BlindMI
Main results
Conclusion

Introduction



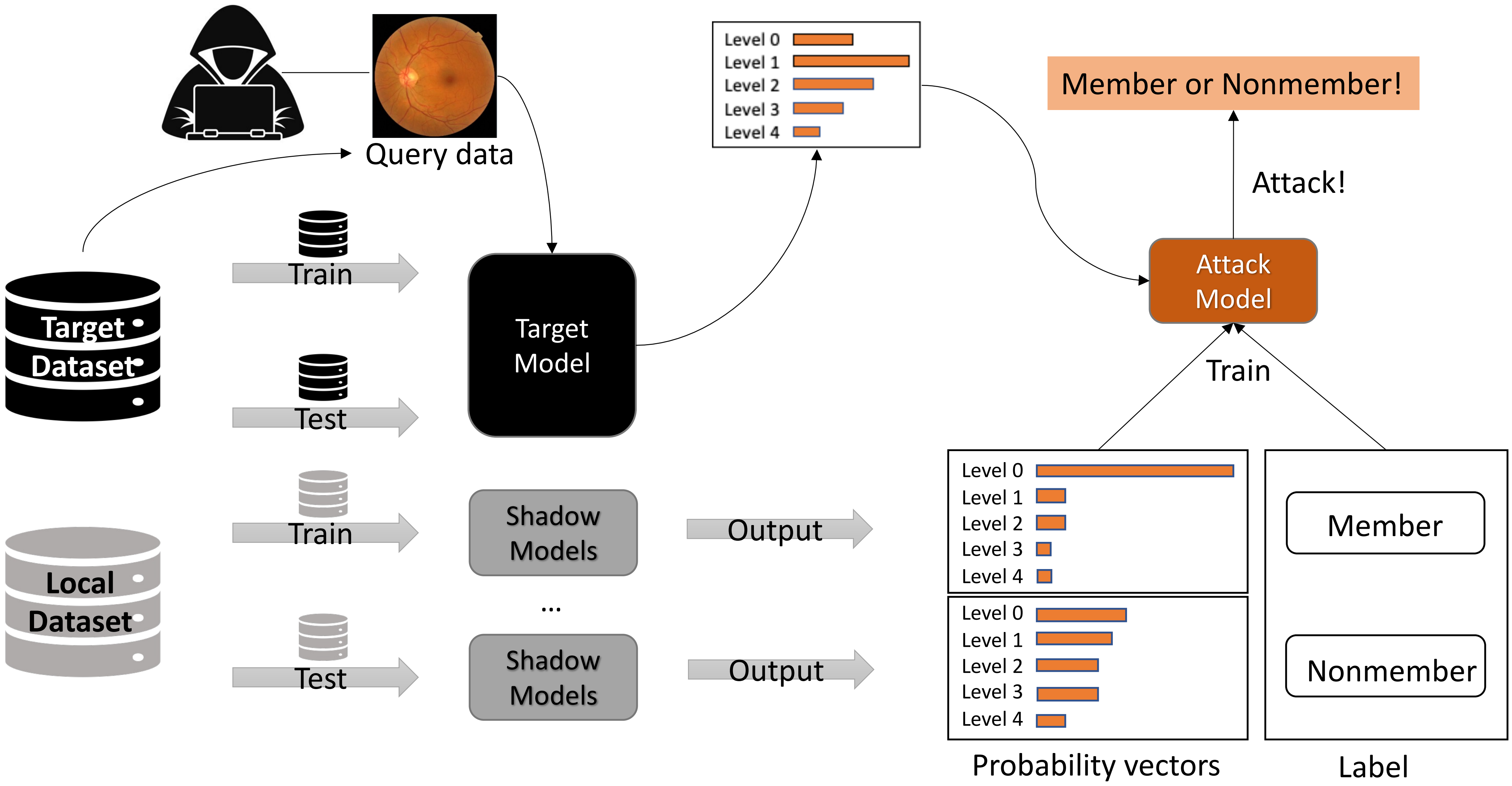
Privacy In Machine Learning

- Model
- Data
  - **Membership Inference**



Membership Inference Attack (State-of-the-art)

Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.



What if the shadow model is not like the target model?

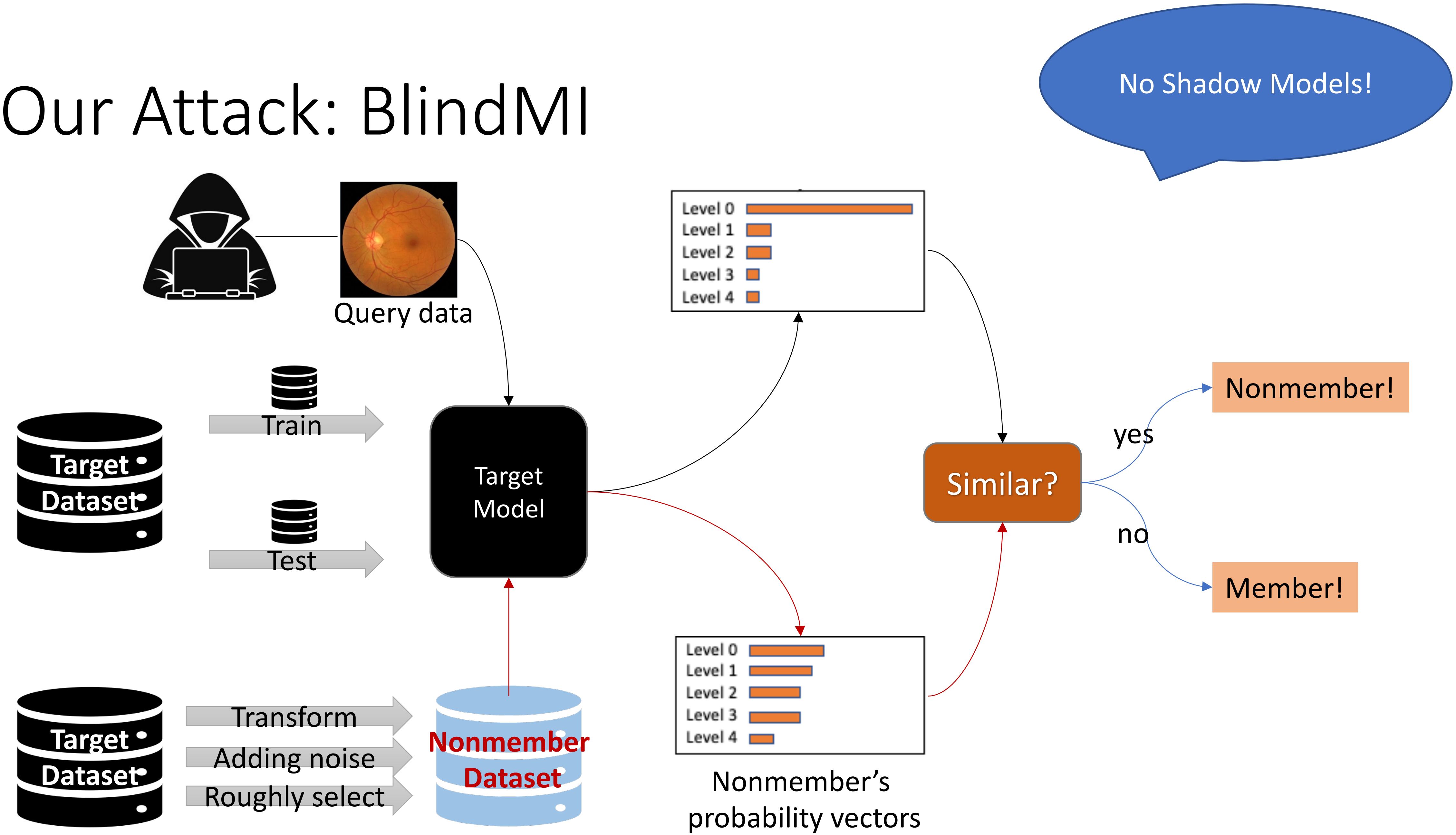
The attack F-1 score decreases.

	Target Model	Shadow Model	Attack F1-Score
CIFAR-100	ResNet50	ResNet50	0.9384
		VGG16	0.7217
		CNN	0.8861
CUB	ResNet101	ResNet101	0.9675
		VGG19	0.8486
		DensNet121	0.6389

How we deal with this problem?

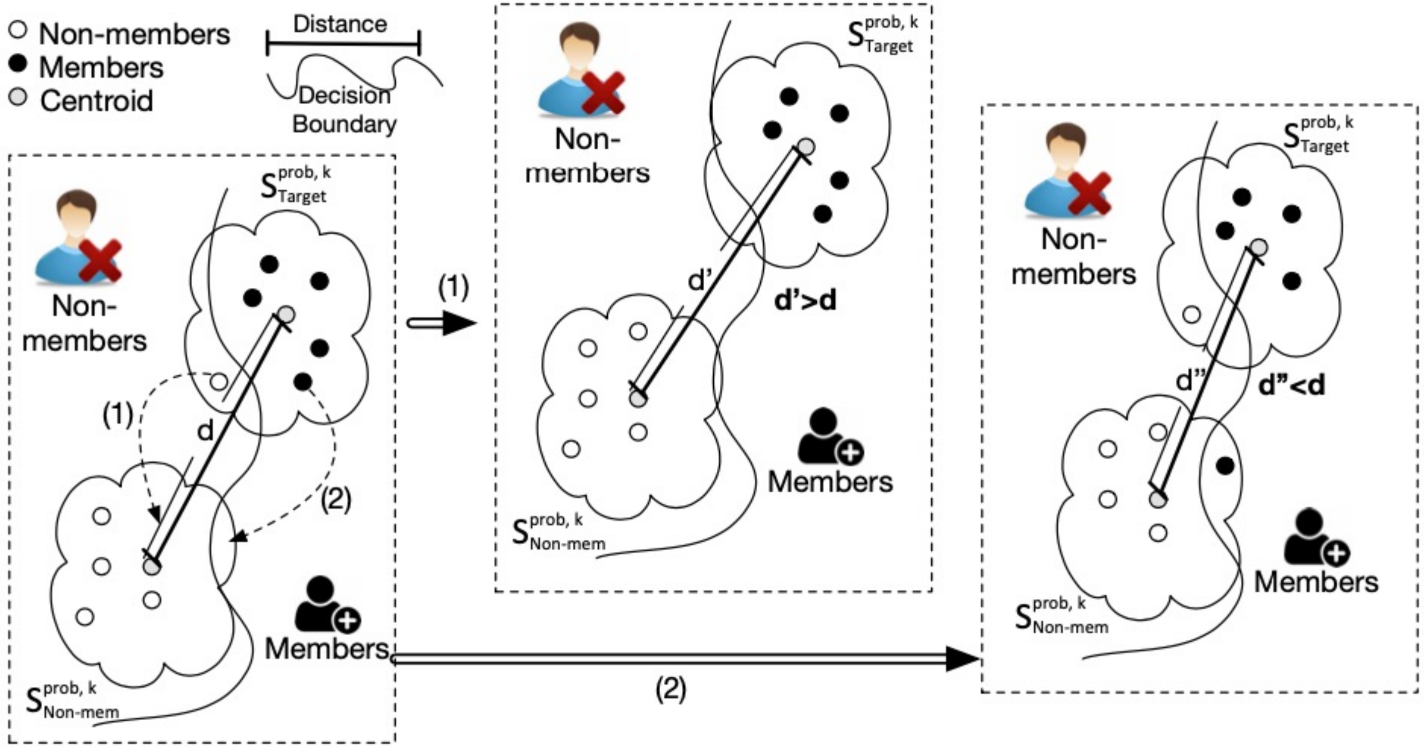
Give up the shadow models!

Our Attack: BlindMI



Variations

- BlindMI-1Calss:
  - Train a one-class SVM model on the nonmember set
- BlindMI-Diff:
  - A novel approach: differential comparison



Main results

Dataset description

Dataset	# of classes	Description	Resolution	Training set size
Adult	2	census income records	N/A	16,280
EyePACS	5	retina images with diabetic retinopathy	150×150	10,000
CH-MNIST	8	histological images of colorectal cancer	64×64	2,500
Location	30	mobile users' location check-in records	N/A	2,505
Purchase-50	50	shoppers' purchase histories	N/A	10,000
Texas	100	inpatients stays in health facilities	N/A	10,000
CIFAR-100	100	object recognition dataset	32×32	10,000
Birds-200	200	photos of birds species	150×150	5,894

Effectiveness: the distance *does* increase

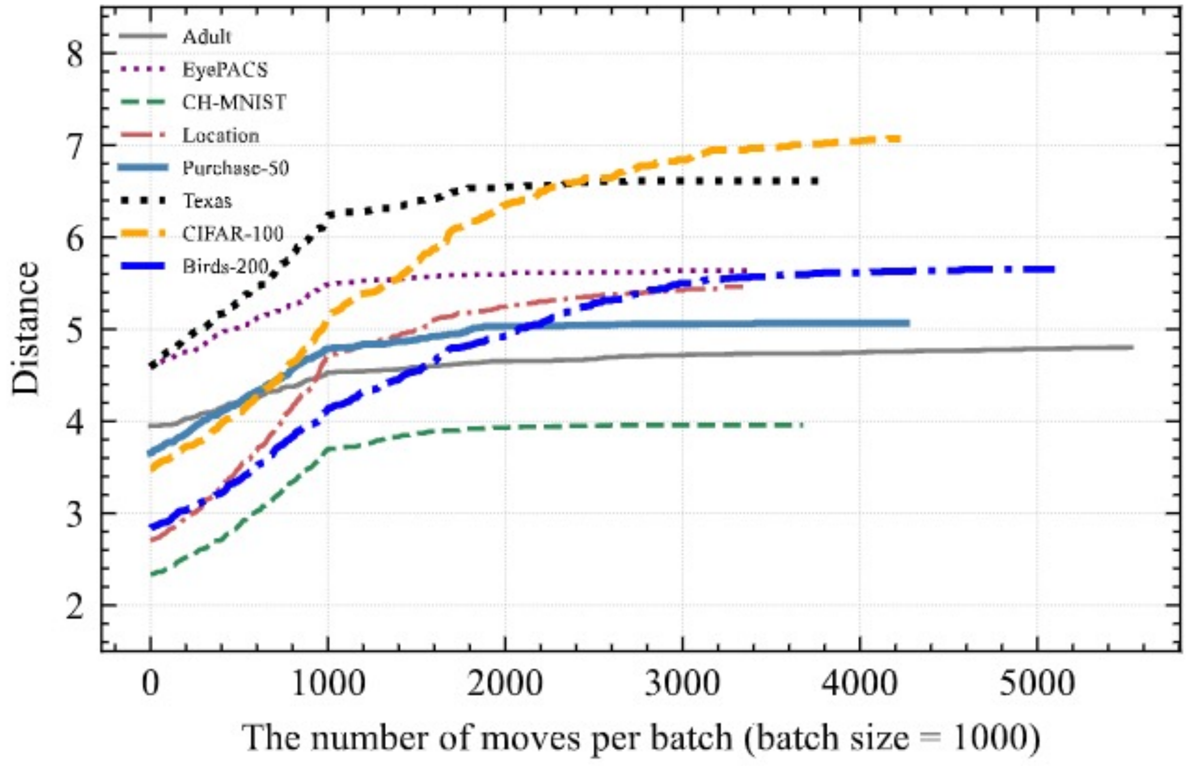


Fig. 8. Distance vs. # of iterations per batch for BLINDMI-DIFF-w/o.

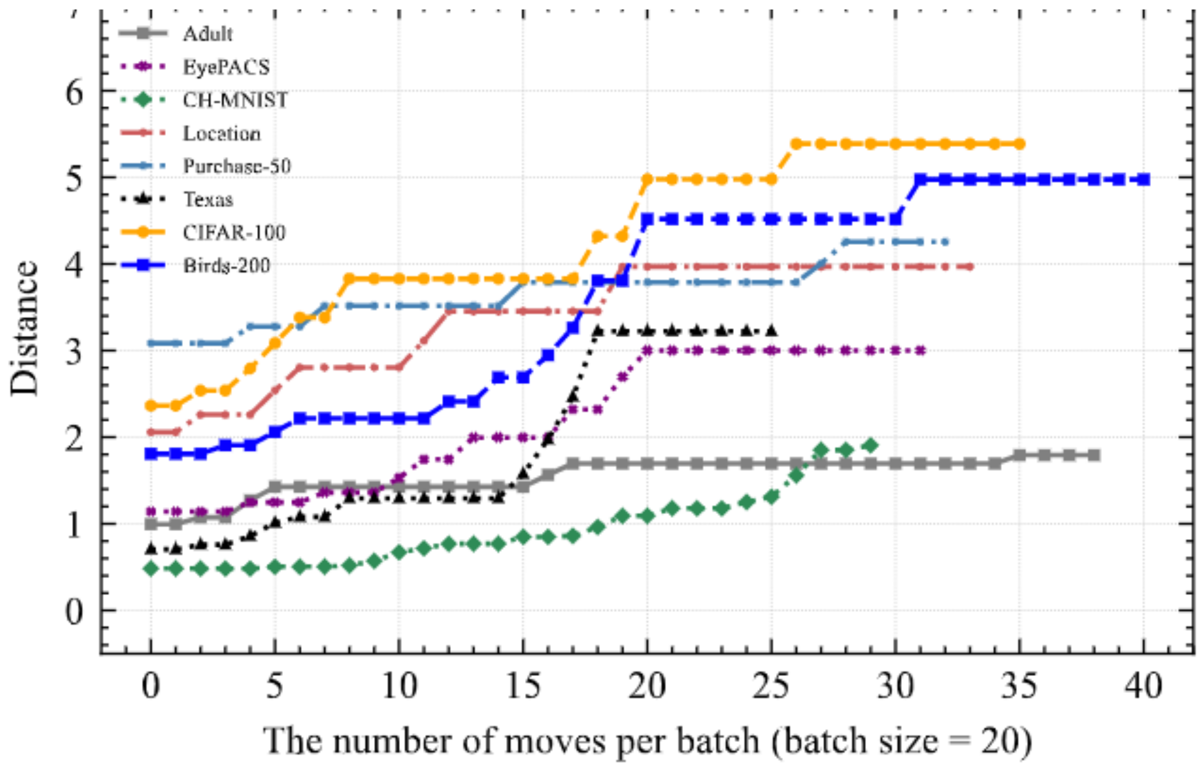


Fig. 7. Distance vs. # of iterations per batch for BLINDMI-DIFF-w/.

State-of-the-art attacks description

- ***NN:*** train a NN model from all features. [1]
- ***Top3-NN:*** train a NN model from top three features. [3]
- ***Top1-Threshold:*** compare the top feature with a threshold. [3]
- ***Loss-Threshold:*** compute a cross-entropy loss and compare. [2]
- ***Label Only:*** classify a sample as a member if the predicted class is correct. [2]
- ***Top2+True:*** our improved version of Top3-NN with the ground-truth label.

[1] Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP).

[2] S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting" 2018 IEEE 31st Computer Security Foundations Symposium (CSF)

[3] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, “Ml-leaks: Model and data independent membership inference attacks and defense son machine learning models.” 2019 Network and Distributed Systems Security Symposium (NDSS).

Comparison with State-of-the-art Attacks

No more shadows
Add more stability

	Attack	Adult	EyePACS	CH-MNIST	Location	Purchase-50	Texas	CIFAR-100	Birds-200
Blind	NN	40.6 ± 7.32	69.1 ± 0.02	71.7 ± 3.53	78.4 ± 3.23	59.4 ± 11.9	76.7 ± 2.20	83.1 ± 3.53	58.3 ± 27.4
	Top3-NN	26.7 ± 7.25	69.5 ± 1.04	70.9 ± 4.03	78.1 ± 3.39	59.6 ± 12.1	76.8 ± 2.07	81.7 ± 6.66	68.6 ± 21.3
	Top1-Threshold	1.01 ± 0.44	71.1 ± 0.42	52.8 ± 17.6	22.7 ± 3.87	53.5 ± 7.26	0.67 ± 0.38	92.8 ± 1.72	71.4 ± 0.65
	BlindMI	**64.2 ± 1.59**	**77.7 ± 0.80**	**75.1 ± 1.49**	**86.2 ± 0.90**	**78.0 ± 0.31**	**85.5 ± 0.80**	**93.9 ± 0.63**	**96.8 ± 0.09**
Blackbox	Top2+True	52.1 ± 6.27	73.4 ± 0.41	75.4 ± 1.84	83.3 ± 2.24	62.9 ± 10.7	83.4 ± 1.29	80.9 ± 7.85	69.5 ± 25.6
	Loss-Threshold	56.2 ± 0.77	73.8 ± 0.57	71.8 ± 4.01	47.7 ± 19.7	48.1 ± 18.6	69.6 ± 9.60	85.6 ± 5.09	71.2 ± 13.7
	Label-Only	56.2 ± 5.28	72.8 ± 0.09	70.9 ± 1.54	75.3 ± 0.12	72.1 ± 0.07	79.7 ± 0.50	85.5 ± 0.47	86.4 ± 0.81
	BlindMI	**66.0 ± 0.28**	**80.6 ± 1.90**	**77.2 ± 1.83**	**87.3 ± 0.70**	**79.9 ± 0.57**	**86.7 ± 0.37**	**94.8 ± 0.14**	**97.2 ± 0.03**
Gray-Blind	NN	54.3 ± 5.50	72.3 ± 0.08	73.5 ± 1.99	85.6 ± 0.71	77.0 ± 0.36	83.4 ± 0.83	93.2 ± 0.46	96.8 ± 0.28
	Top3-NN	56.4 ± 9.27	74.8 ± 0.37	73.6 ± 1.80	85.7 ± 0.69	77.2 ± 0.34	83.4 ± 0.90	93.2 ± 0.80	93.2 ± 0.03
	Top1-Threshold	1.01 ± 0.44	71.1 ± 0.42	52.8 ± 17.6	22.7 ± 3.87	53.5 ± 7.26	0.67 ± 0.38	92.8 ± 1.72	71.4 ± 0.65
	BlindMI	**64.2 ± 1.59**	**77.7 ± 0.80**	**75.1 ± 1.49**	**86.2 ± 0.90**	**78.0 ± 0.31**	**85.5 ± 0.80**	**93.9 ± 0.63**	**96.8 ± 0.09**
Graybox	Top2+True	66.0 ± 0.50	77.3 ± 0.69	75.1 ± 2.03	86.0 ± 0.55	78.4 ± 0.25	85.7 ± 0.18	93.8 ± 0.53	96.9 ± 0.18
	Loss-Threshold	57.0 ± 0.84	76.8 ± 0.68	73.0 ± 2.90	75.9 ± 4.96	71.8 ± 2.70	76.5 ± 4.81	87.1 ± 3.39	85.3 ± 0.89
	Label-Only	56.2 ± 5.28	72.8 ± 0.09	70.9 ± 1.54	75.3 ± 0.12	72.1 ± 0.07	79.7 ± 0.50	85.5 ± 0.47	86.4 ± 0.81
	BlindMI	**66.0 ± 0.30**	**80.6 ± 1.90**	**77.2 ± 1.83**	**87.3 ± 0.70**	**79.9 ± 0.57**	**86.7 ± 0.37**	**94.8 ± 0.14**	**97.2 ± 0.03**

$\triangle$ 0 $\triangle$ 28.2 $\triangle$ 17.6 $\triangle$ 38.5

Different nonmember generations:

- Transformation is the best.

TABLE XI. MMD STATISTICAL TESTS OF BLINDMI-DIFF WITH NONMEMBER DATASETS GENERATED VIA DIFFERENT METHODS (EACH VALUE IS THE MMD WITH STANDARD ERROR OF THE MEAN BETWEEN CORRESPONDING SAMPLES AND REAL-WORLD NON-MEMBERS IN THE TEST DATASET.)

Sample trans	Random perp	Random generation	Cross domain	Training set
0.194 ± 0.009	0.438 ± 0.039	3.024 ± 1.024	0.225 ± 0.015	1.864 ± 0.022

TABLE XII. F1-SCORE (%) WITH STANDARD ERROR OF MEAN FOR DIFFERENT KERNEL FUNCTIONS OF BLINDMI-DIFF

		Gaussian (default)	Laplacian	Linear	Sigmoid	Polynomial
DIFF-w/	Adult	**64.2±1.59**	60.3±0.38	40.7±0.20	51.1±0.41	58.4±1.02
	EyePACS	**77.7±0.80**	67.3±0.31	71.8±0.93	72.8±0.87	73.9±0.88
	CH-MNIST	**75.1±1.49**	73.1±0.92	72.4±0.53	71.3±0.71	72.7±1.20
	Location	**86.2±0.90**	85.1±2.42	83.4±0.98	79.8±1.52	76.7±0.17
	Purchase-50	**78.0±0.31**	68.9±0.50	75.8±0.61	71.1±1.05	66.0±0.99
	Texas	**85.5±0.80**	83.6±0.47	81.2±0.29	80.9±0.49	81.9±1.72
	CIFAR-100	**93.9±0.63**	93.3±0.79	87.9±1.09	86.9±1.02	90.1±0.83
	Birds-200	**96.8±0.09**	91.9±1.32	95.7±1.06	94.4±1.31	93.9±0.96

Different kernel functions:

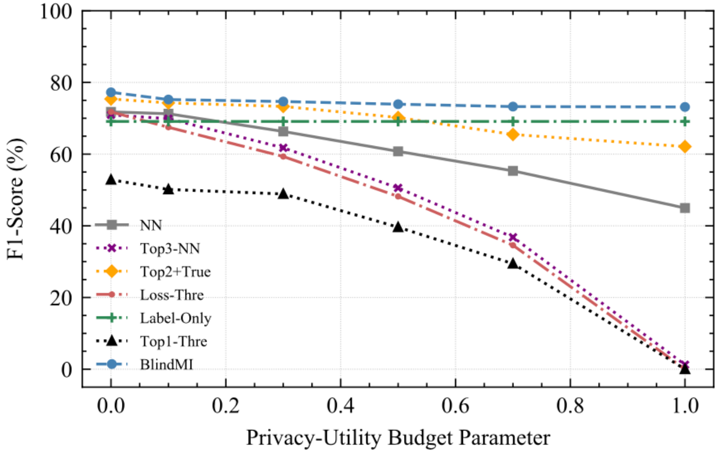
- Gaussian is the best.

Evaluation against State-of-the-art Defenses

MemGuard:

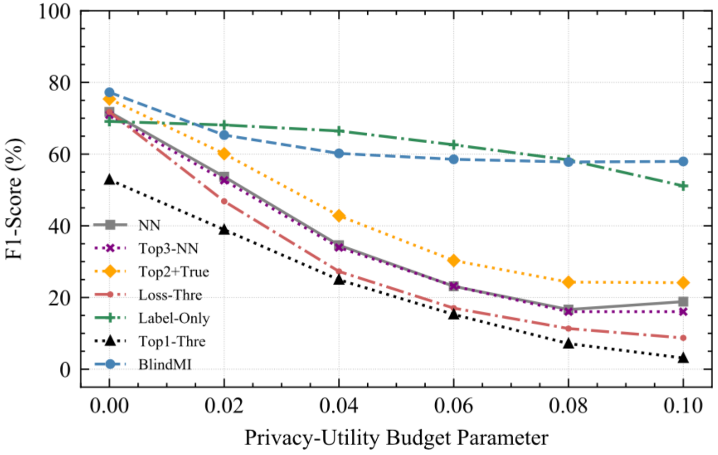
Add carefully crafted perturbation to the target model's output and turns it into an adversarial example to fool the attacker's classifier.

Outperform 5% to 75%



(a) MemGuard on CH-MNIST

Outperform 8% to 59%



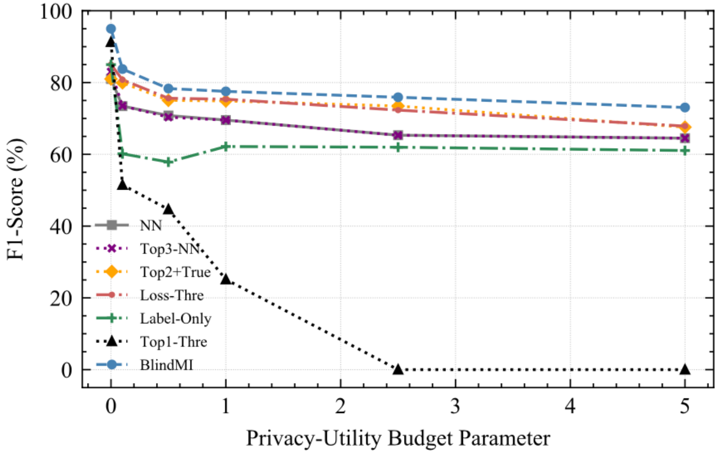
(b) DP-Adam on CH-MNIST

DP-Adam:

Add perturbations to the training process such that no single training sample has a significant impact on the learned target model.

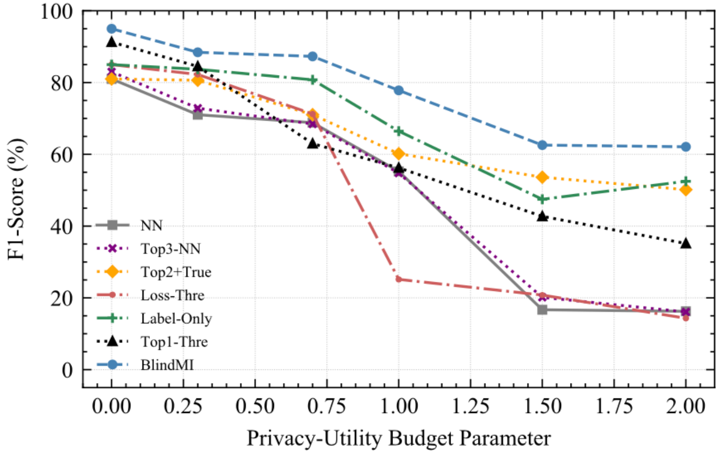
MMD+Mixup:

Adopt Maximum Mean Discrepancy to reduce the gap between the softmax distributions of the training and validation sets during training.



(c) MMD+Mix-up on CIFAR-100

Outperform 5% to 75%



(d) Adversarial Regularization on CIFAR-100

Outperform 10% to 60%

Adversarial Regularization:

Model MI attacks as a regularization term to be used in regularizing the training of the target model.

F1-Score vs. Nonmember-to-Member Ratio

- Ratio↑ Attack↓
- BlindMI outperform 35%

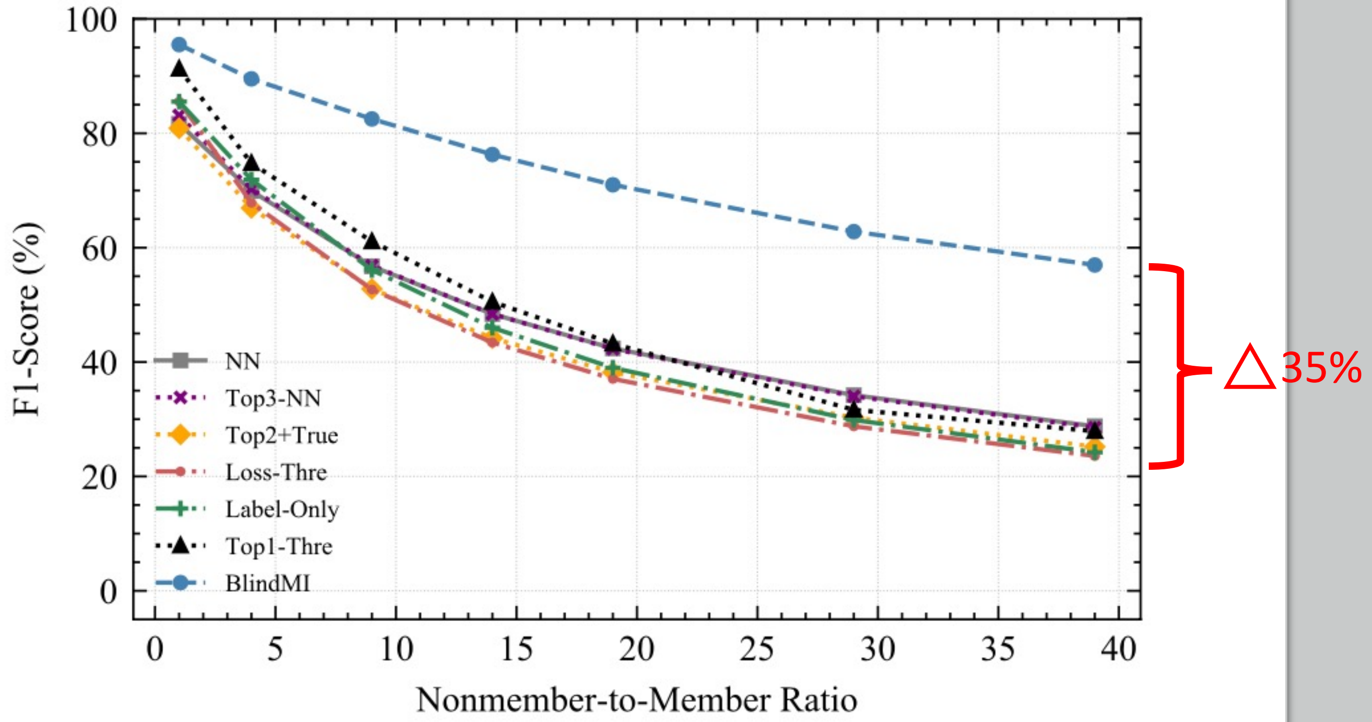


Fig. 4. F1-Score of Various Attacks vs. Nonmember-to-Member Ratio on CIFAR-100.

F1-score vs. # of Classes

- Class↑ Attack↑
- BlindMI outperform 5%-30%

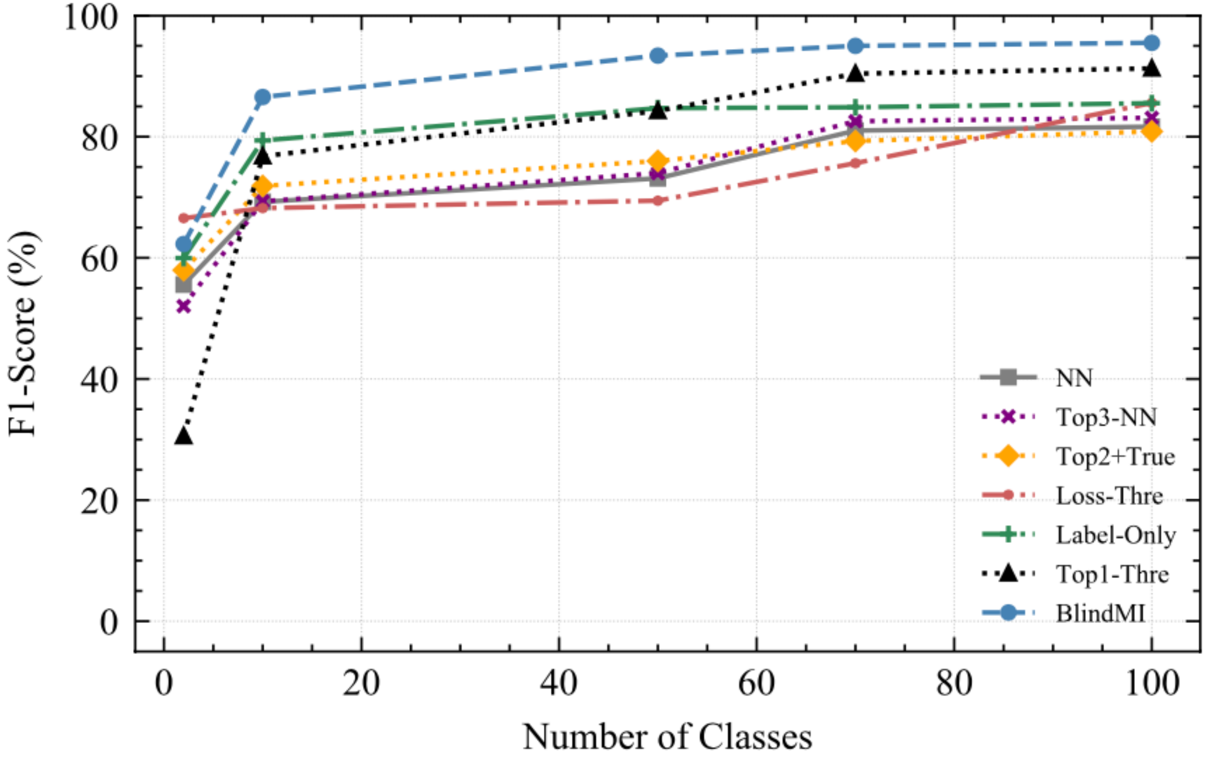


Fig. 5. F1-Score of Various Attacks vs. # of classes on CIFAR.

Conclusion

- We design a membership inference attack BlindMI using a novel technique, called differential comparison.
- Our evaluation shows that BlindMI outperforms state-of-the-art MI attacks under different settings.
- Our implementation is open-source at this repository:
- https://github.com/hyhmia/BlindMI