Practical Blind Membership Inference Attack via Differential Comparisons

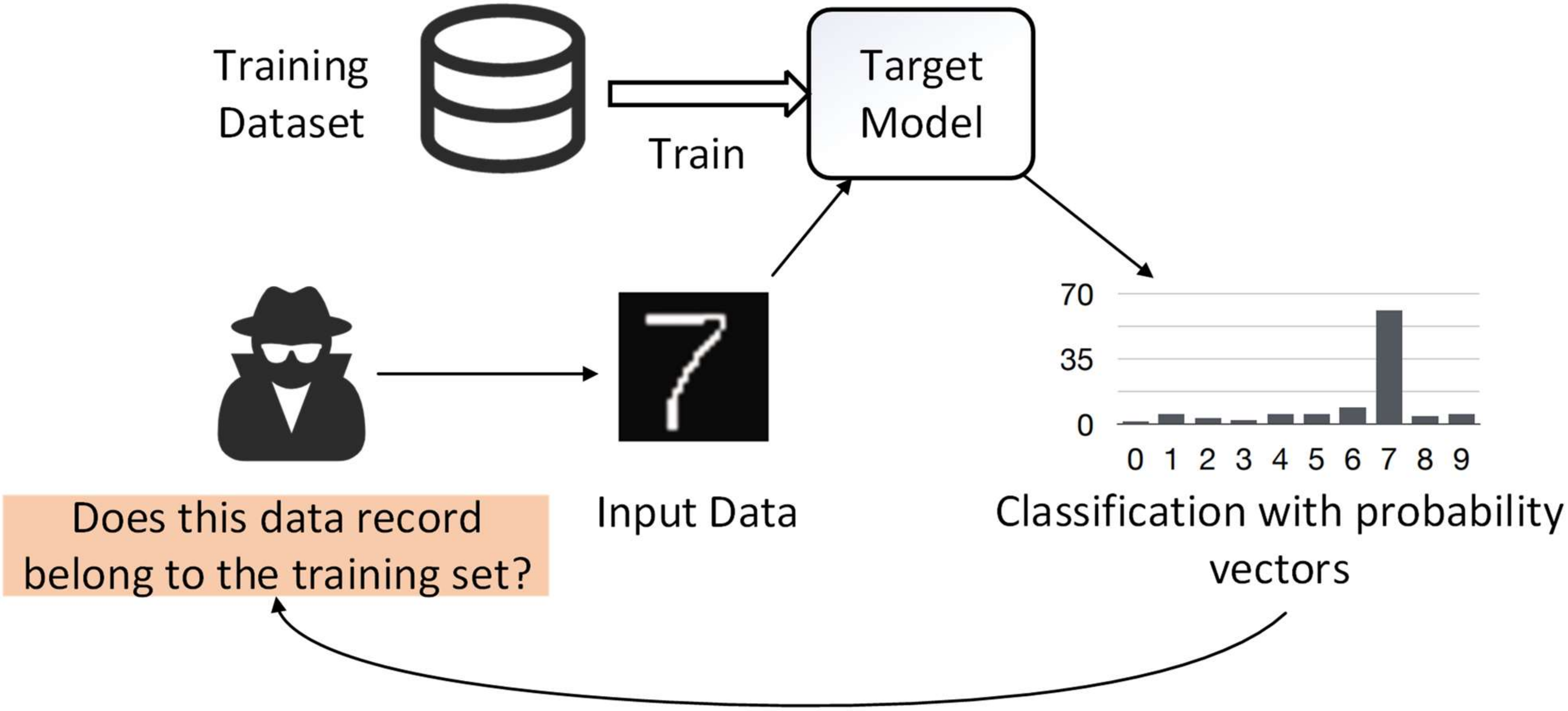
Bo Hui†∗, Yuchen Yang†∗, Haolin Yuan†∗, Philippe Burlina ‡, Neil Zhenqiang Gong§, and Yinzhi Cao†

†The Johns Hopkins University ‡The Johns Hopkins University Applied Physics Laboratory §Duke University

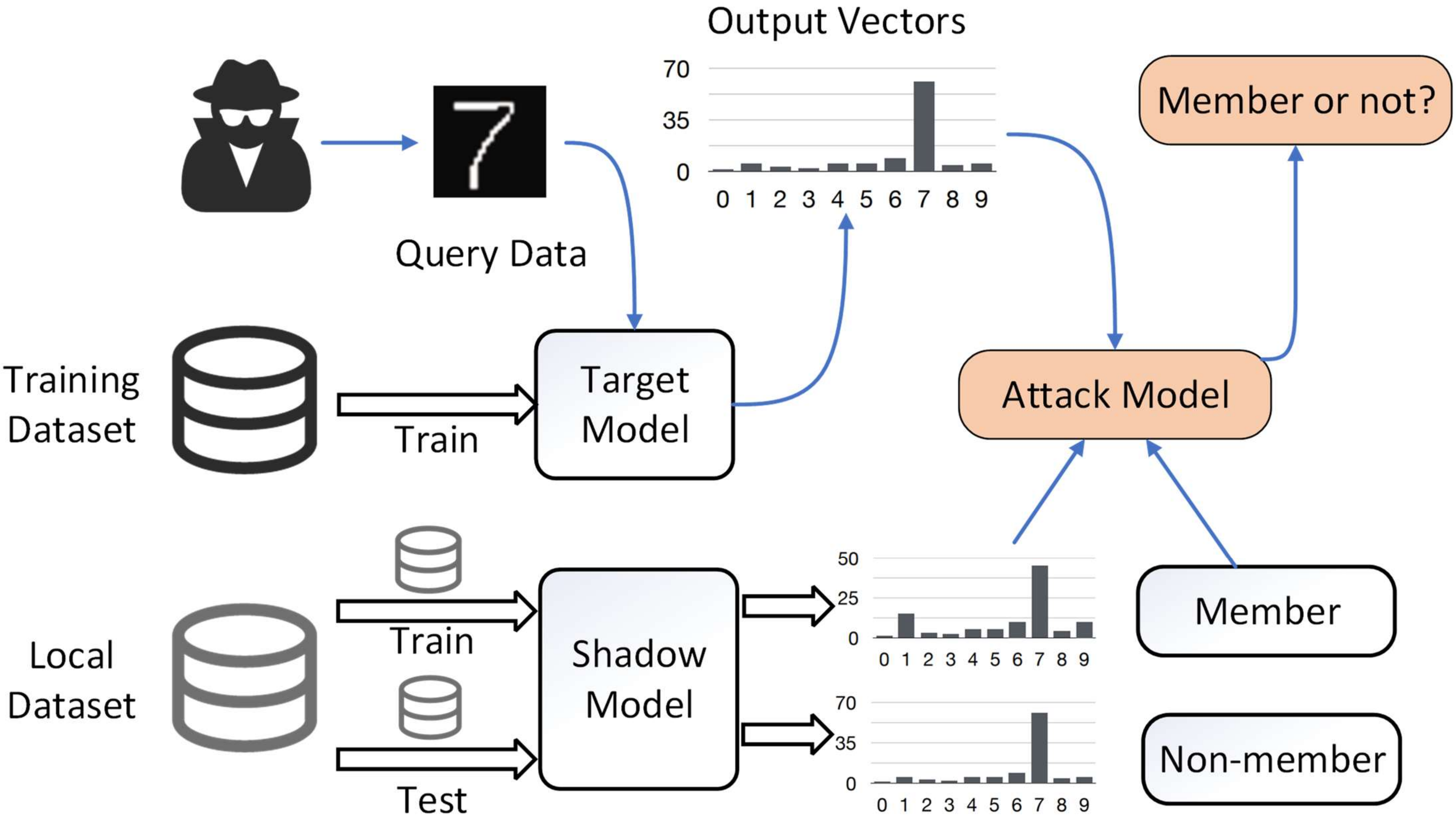
Privacy Problem of ML

- **Model**
- **Data**

Membership Inference


Training Dataset
Train
Target Model
70
35
0
0 1 2 3 4 5 6 7 8 9
Does this data record belong to the training set?
Input Data
Classification with probability vectors

State-of-the-art Attack



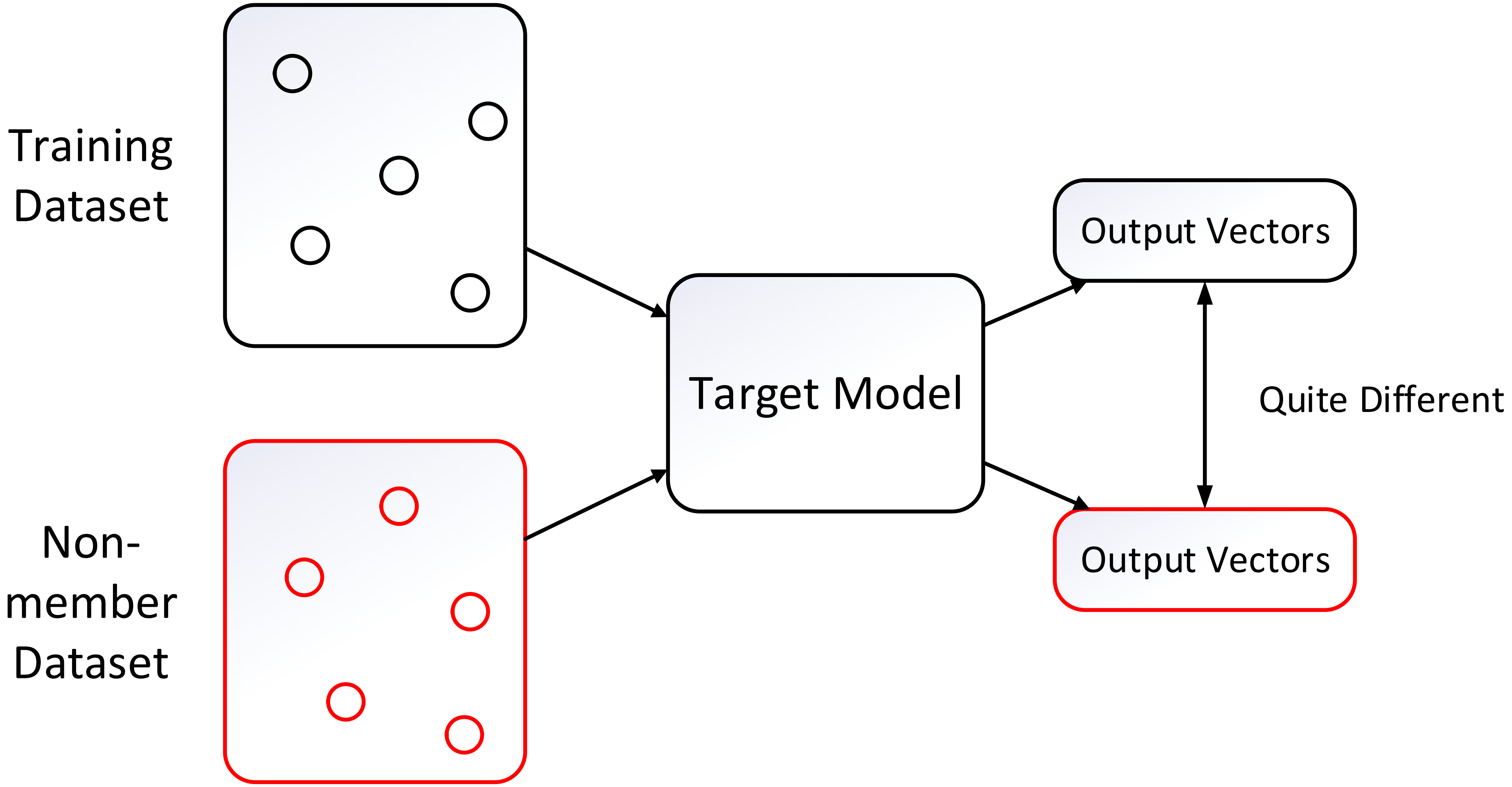
State-of-the-art Attack

- **The Attack Performance**

	Target Model	Shadow Model	Attack F1-Scores
CIFAR-100	ResNet50	ResNet50	0.9384
		VGG16	0.7217
		CNN	0.8861
CUB	ResNet101	RestNet101	0.9675
		VGG19	0.8486
		DensNet121	0.6389

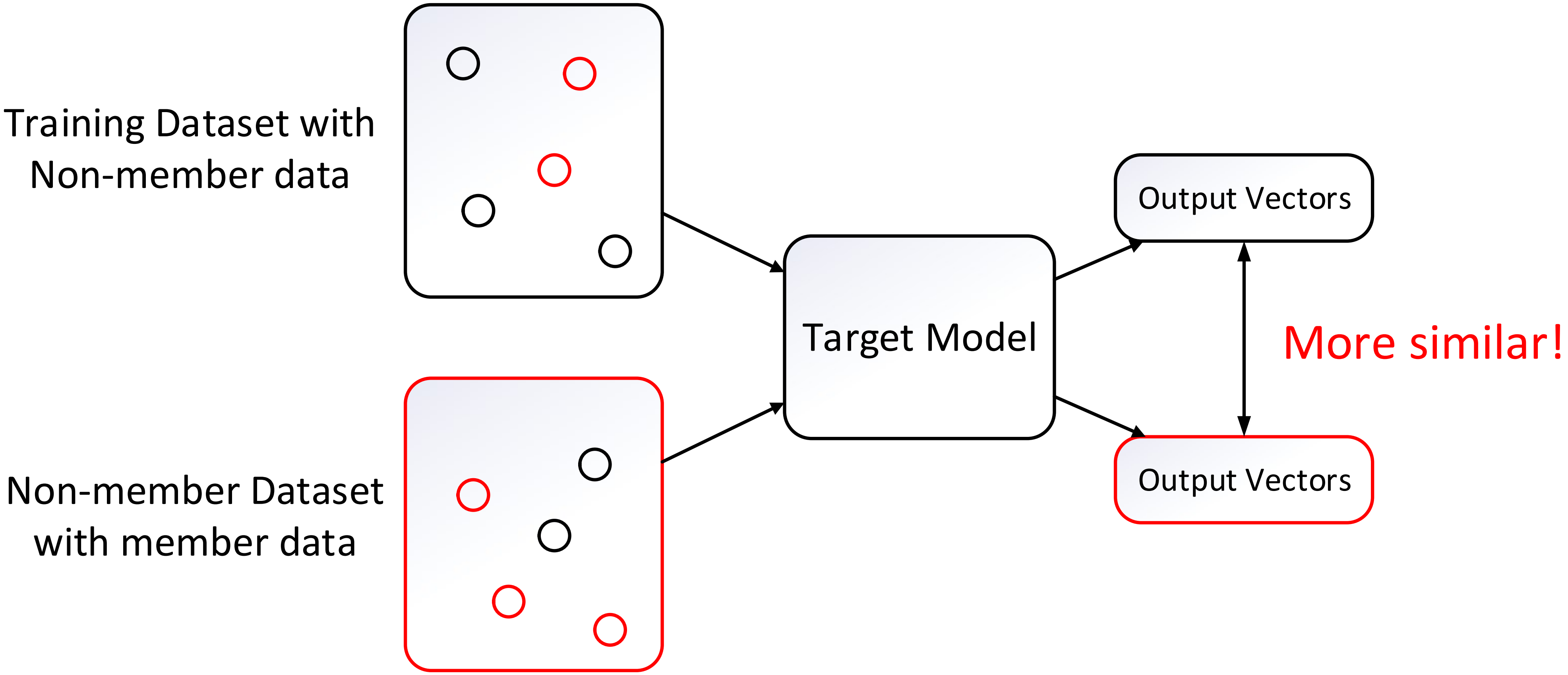
Motivation

- **Differential Comparison**



Motivation

- **Differential Comparison**



Training Dataset with Non-member data

Non-member Dataset with member data

Target Model

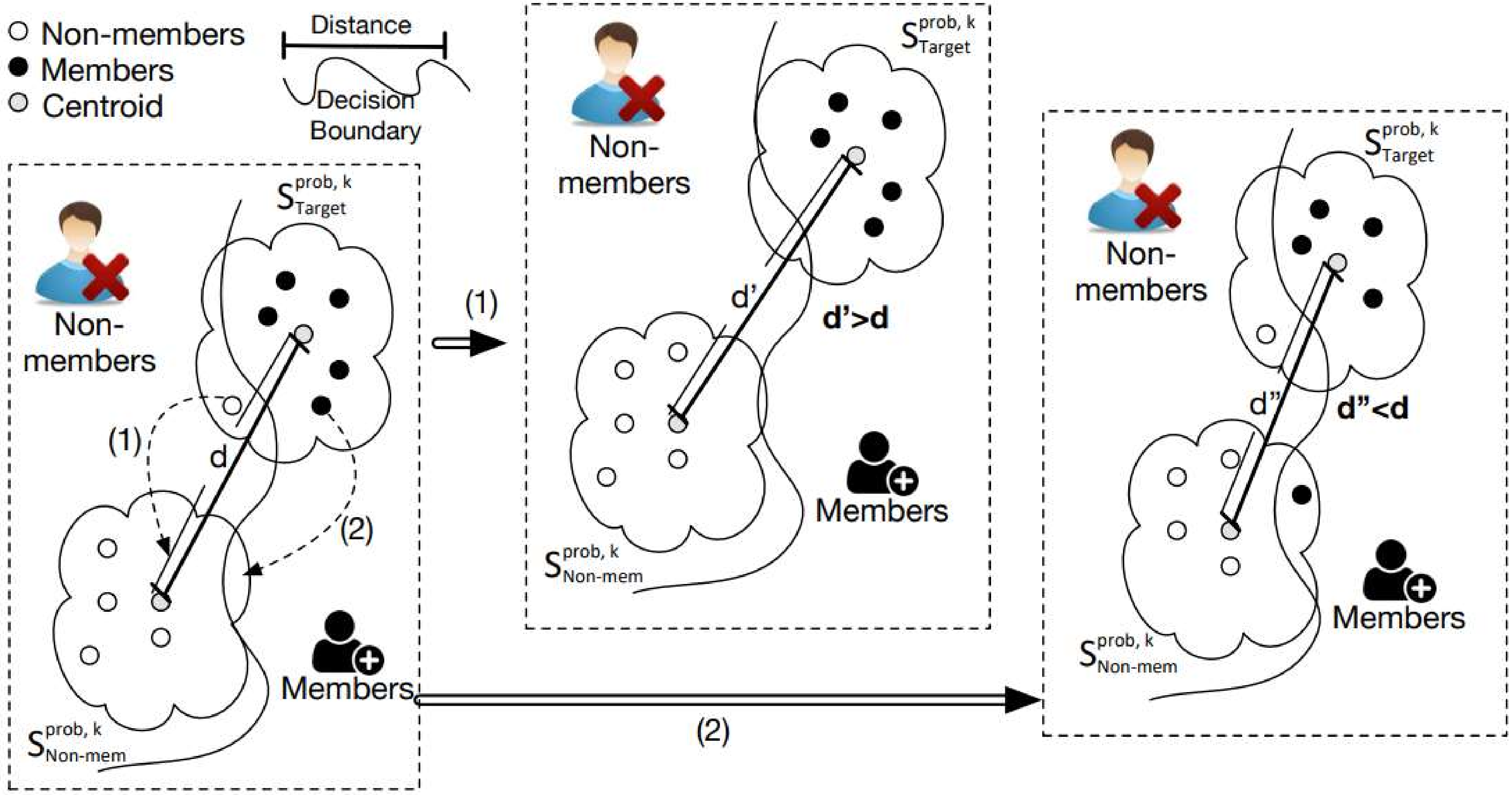
Output Vectors

Output Vectors

More similar!

BlindMI Attack

- **Differential Comparison**

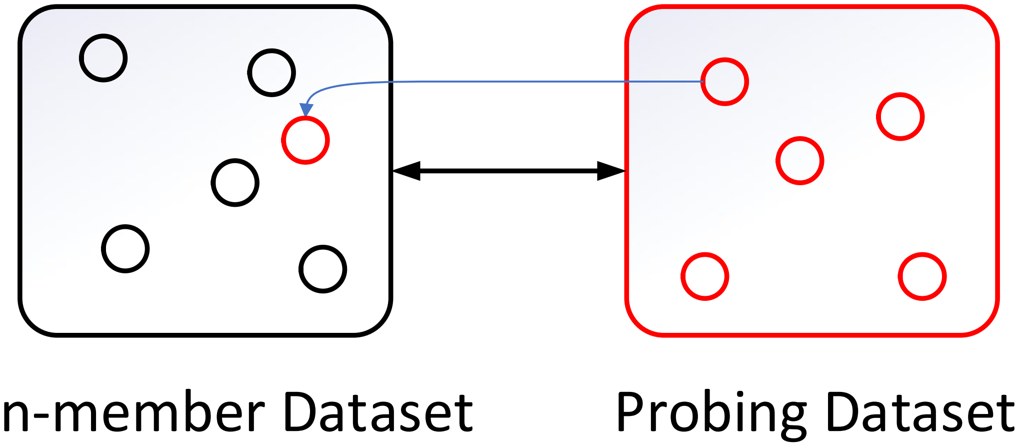


Some Details

- **Dataset Preparation for Differential Comparison**

  1) Generation of Non-members:

  - **Sample Transformation.**
  - **Random Perpetuation.**
  - **Cross-domain Samples.**

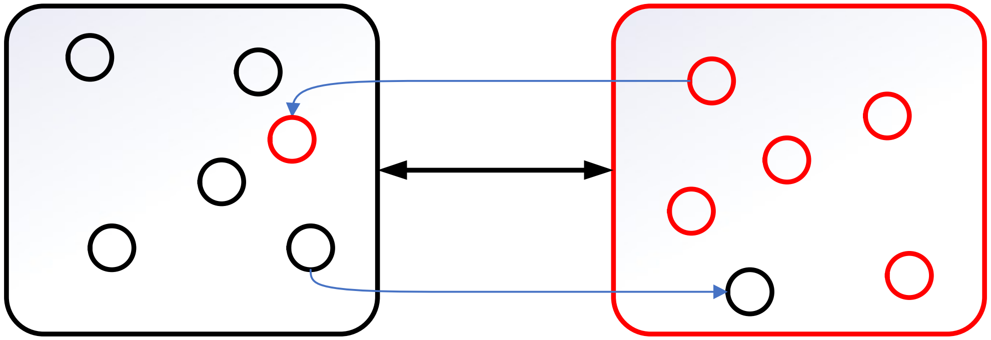


Some Details

- **Dataset Preparation for Differential Comparison**

  2) Rough Sample Separation:

  - **Clustering algorithm like k-means.**
  - **Separation based on the highest probability score.**



Probing Dataset 1 Probing Dataset 2

Some Details

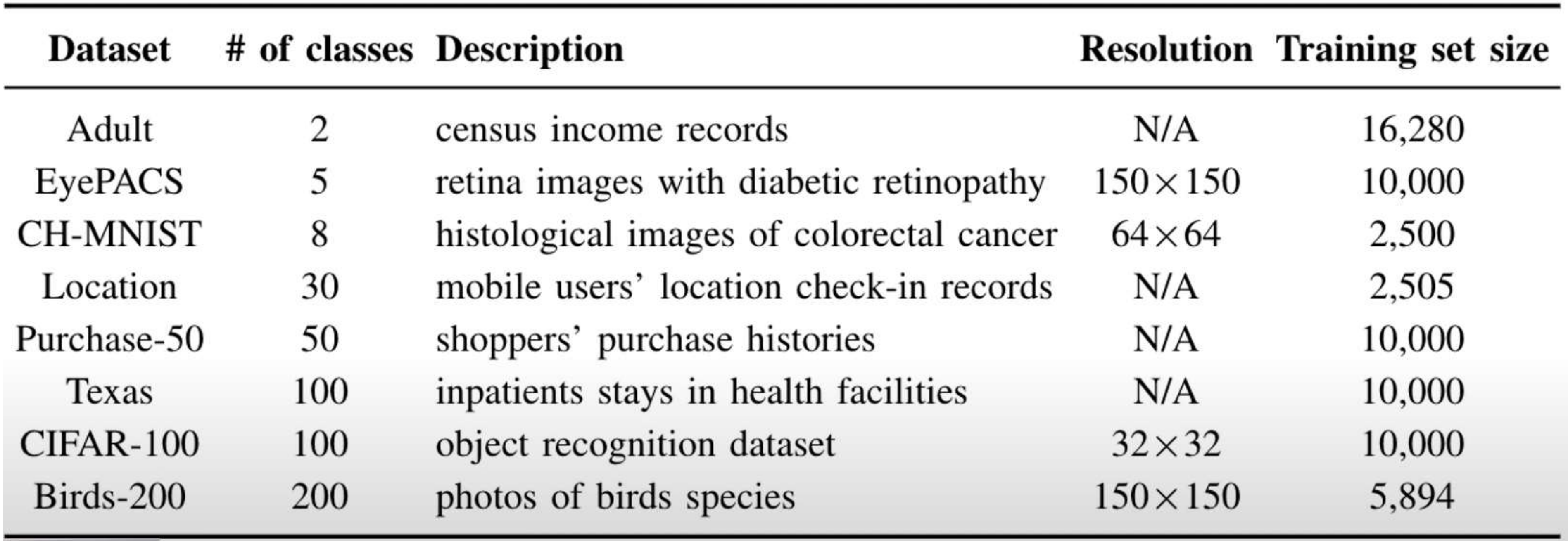
- **Distance Calculations**

$$D(S_{target}^{prob,k}, S_{nonmem}^{prob,k}) = \left\| \frac{1}{n_t} \sum_{i=1}^{n_t} \phi(y_i) - \frac{1}{n_n} \sum_{j=1}^{n_n} \phi(y'_j) \right\|_{\nu}$$

$$y_i \in S_{target}^{prob,k},\ y'_j \in S_{nonmem}^{prob,k}$$

Evaluations

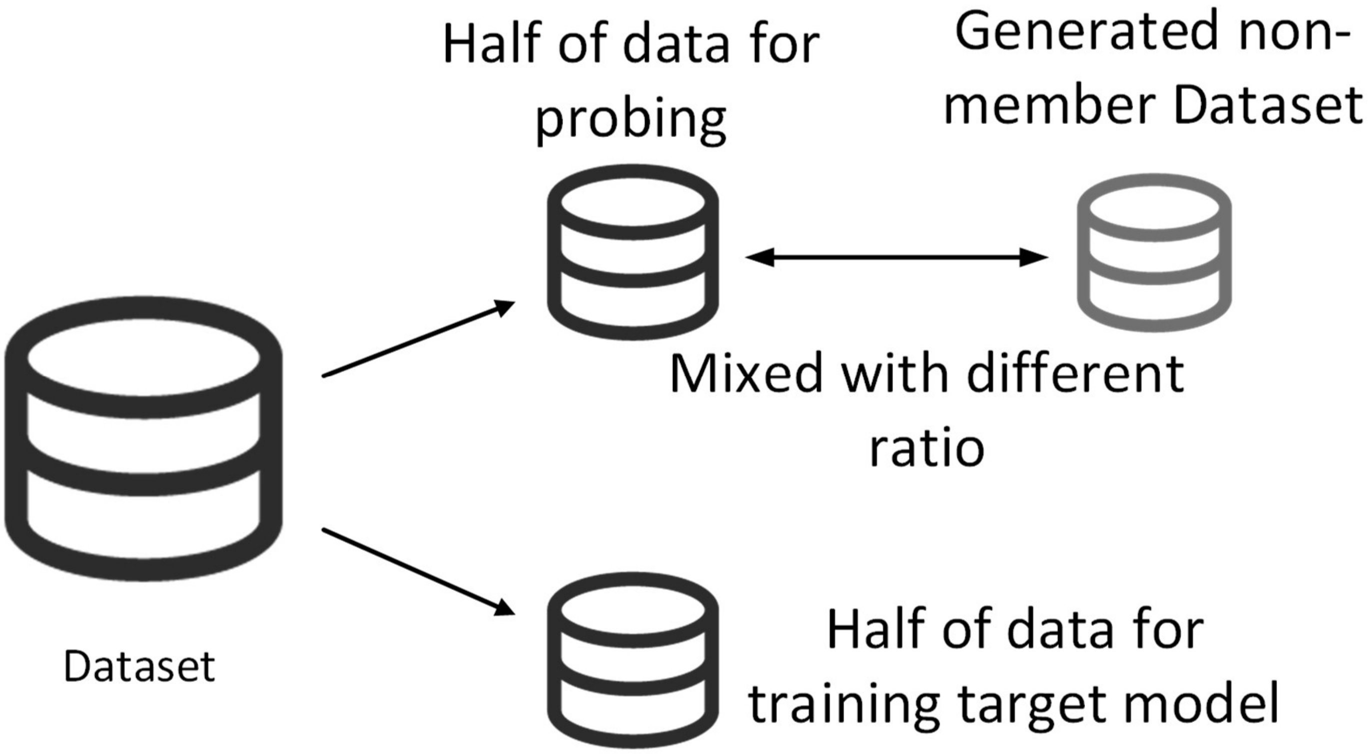
- **Eight Datasets**



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

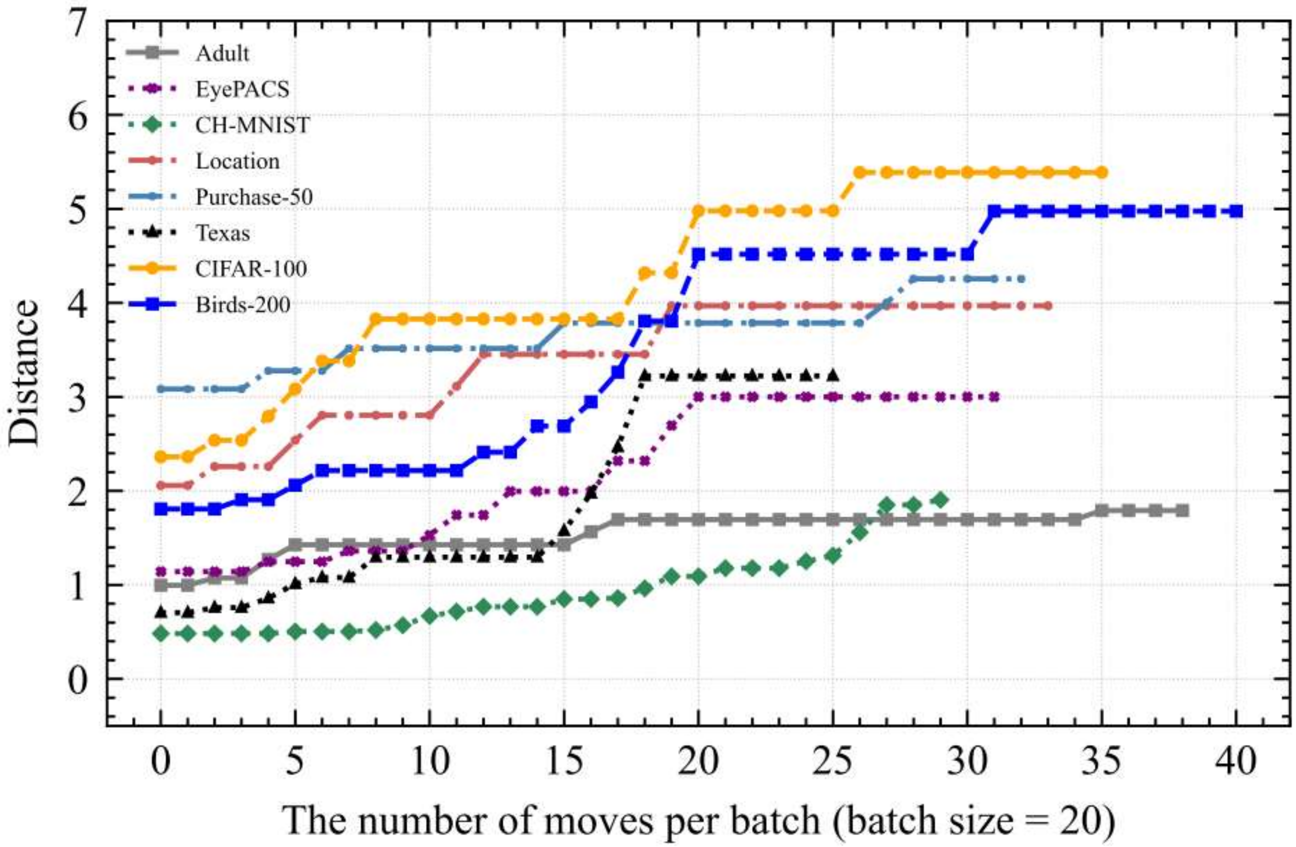
Evaluations

- **Training and Probing Datasets**


Half of data for probing
Generated non-member Dataset
Mixed with different ratio
Dataset
Half of data for training target model

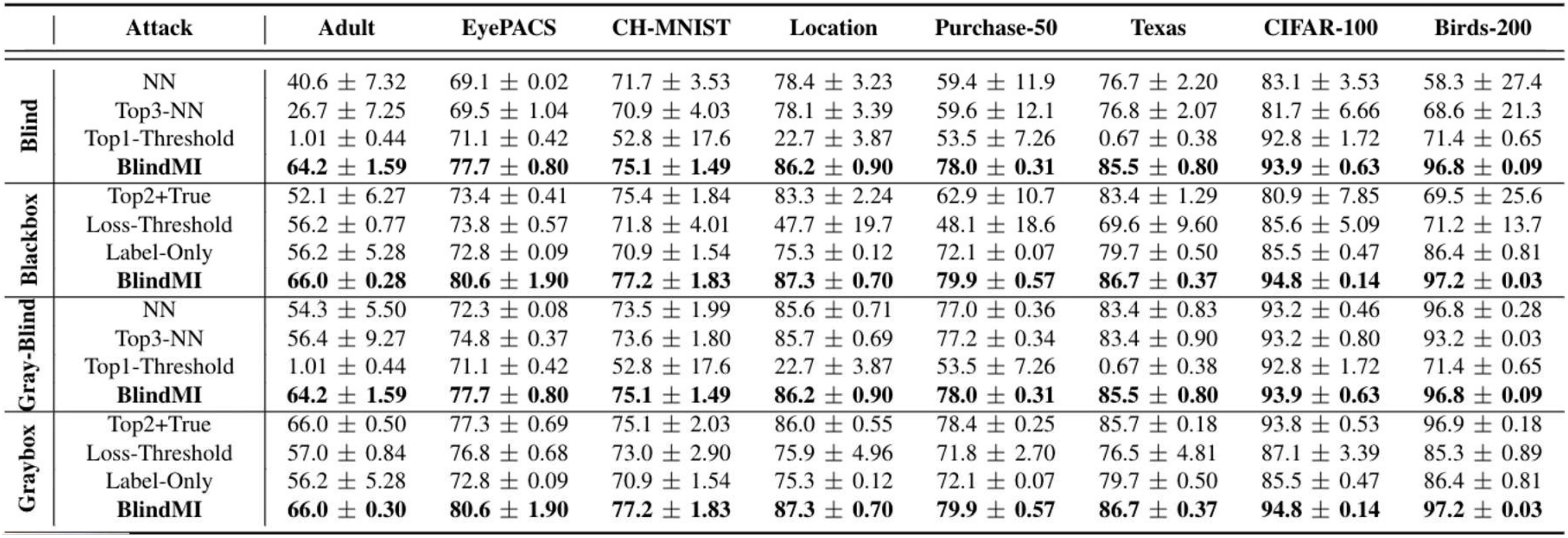
Evaluations

- **Distance Variations**



Evaluations

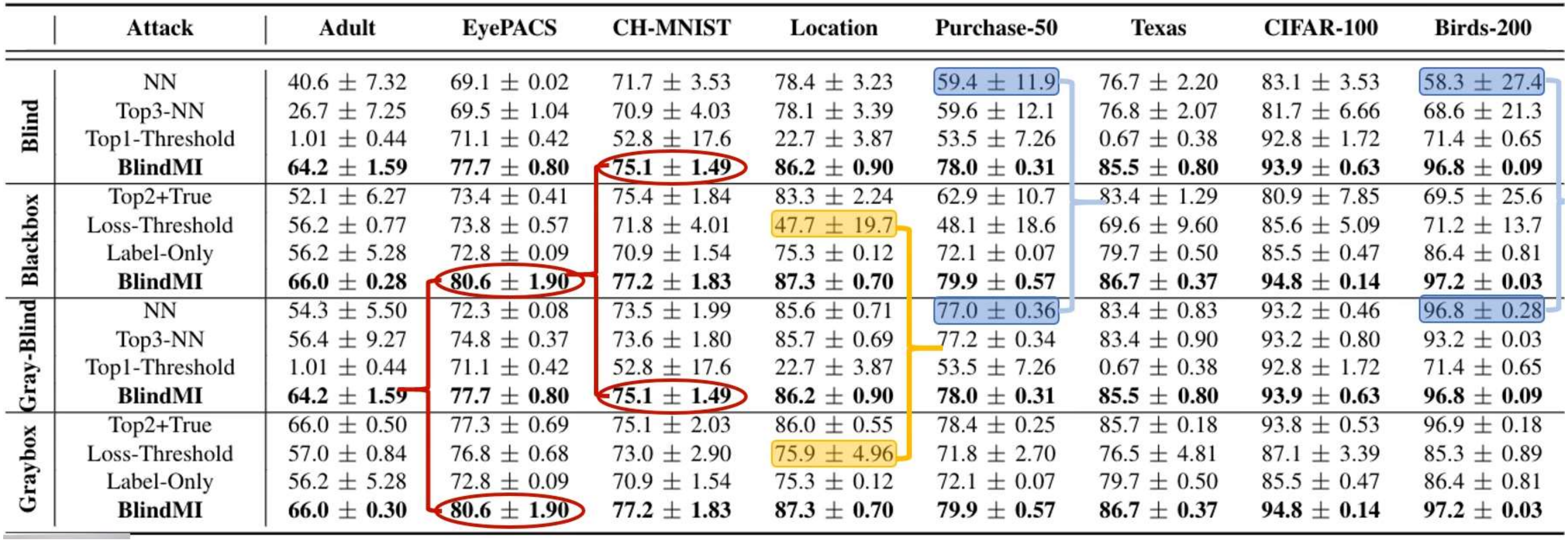
- **Comparison of State-of-the-art attacks**



	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**

Evaluations

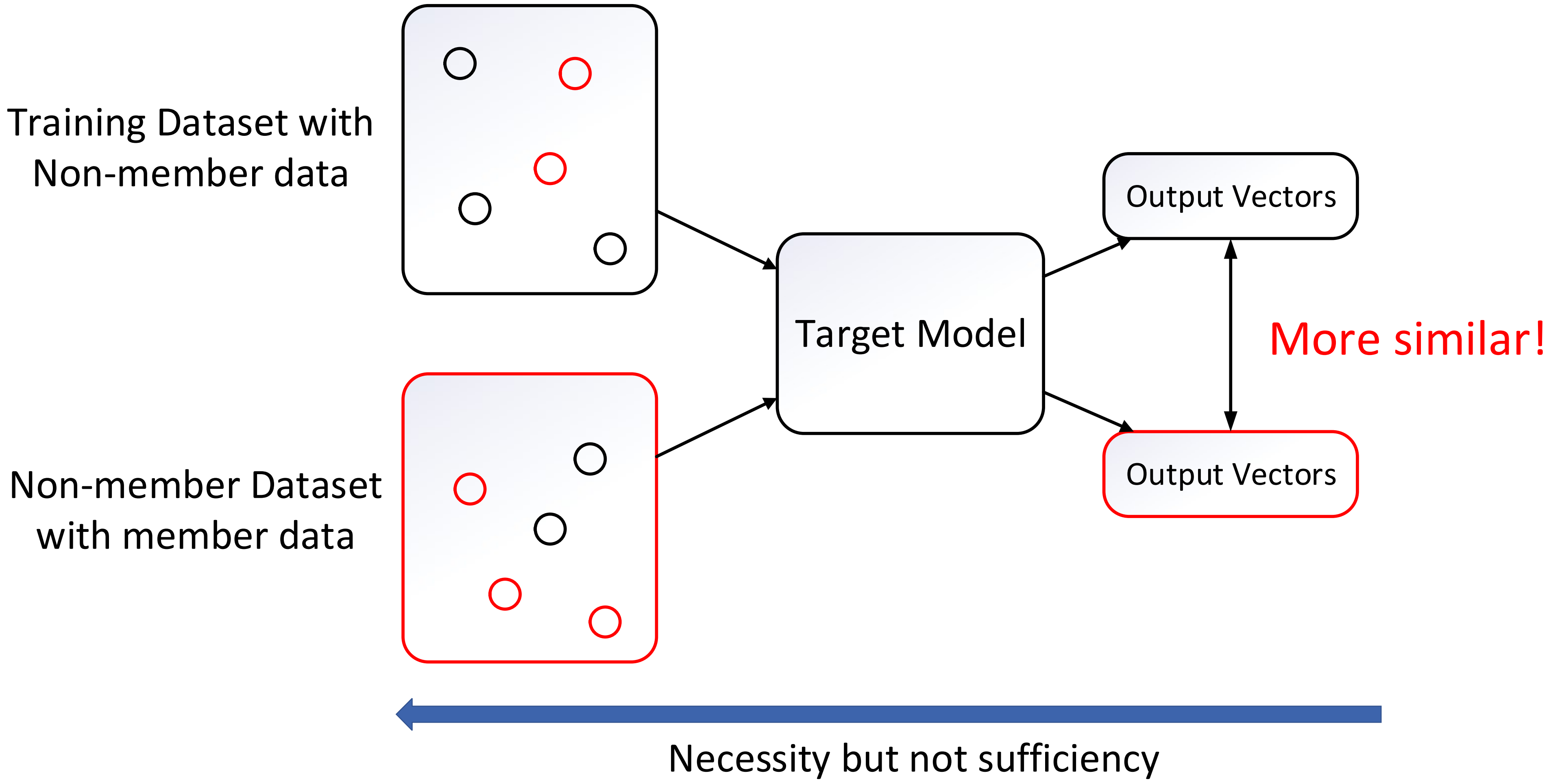
- **Comparison of State-of-the-art attacks**



	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**

Problems

- **Motivations**



Training Dataset with Non-member data

Non-member Dataset with member data

Target Model

Output Vectors

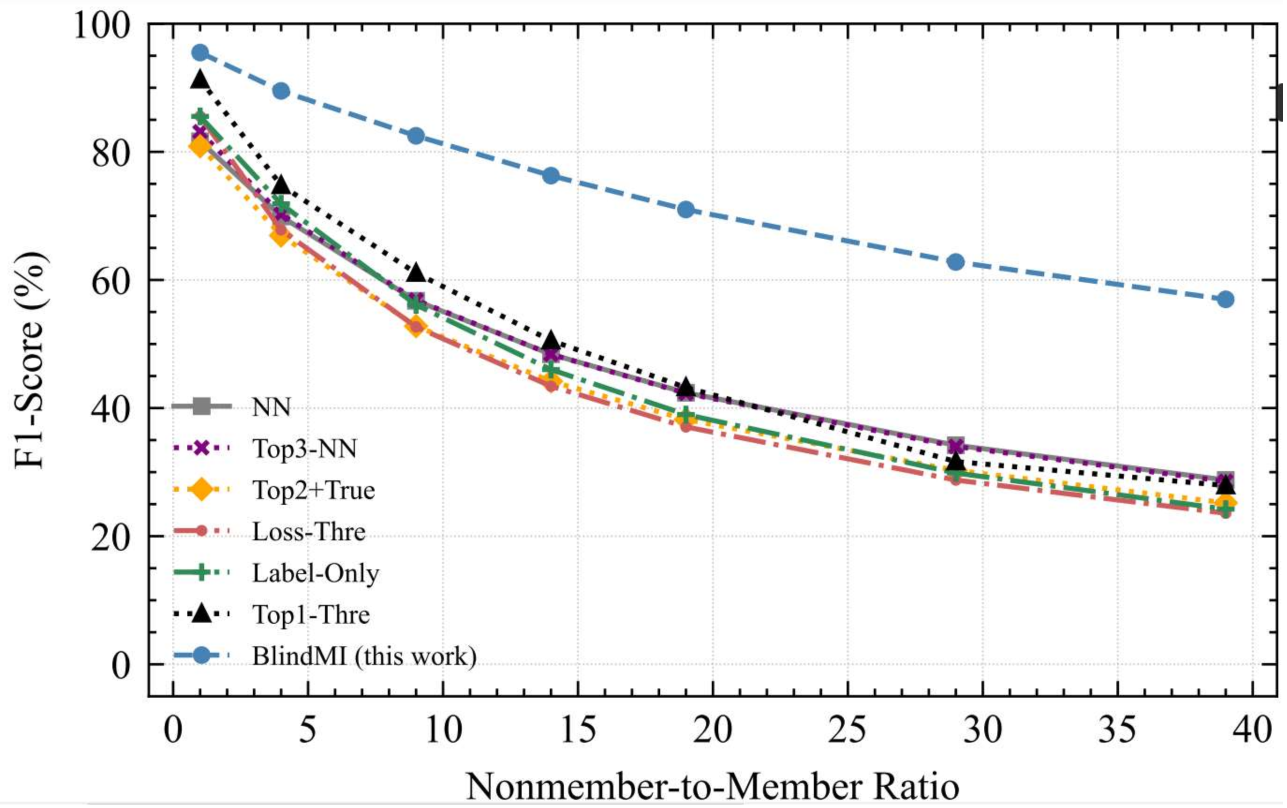
Output Vectors

More similar!

Necessity but not sufficiency

Problems

- **Prob Dataset**


F1-Score (%)
Nonmember-to-Member Ratio
NN
Top3-NN
Top2+True
Loss-Thre
Label-Only
Top1-Thre
BlindMI (this work)