EncoderMI: Membership Inference against Pre-trained Encoders in Contrastive Learning

Hongbin Liu*, Jinyuan Jia*, Wenjie Qu †, Neil Zhenqiang Gong*

Duke University*, Huazhong University of Science and Technology †

Contrastive Learning

- **Learn the general features of a dataset without labels**

Teaching the model which data points are similar or different.

Unlabeled Data

Encoder

Downstream Classifier

Contrastive Learning

- **Similar Sample Stay Close to Each Other and Dissimilar Ones are Far Apart.**

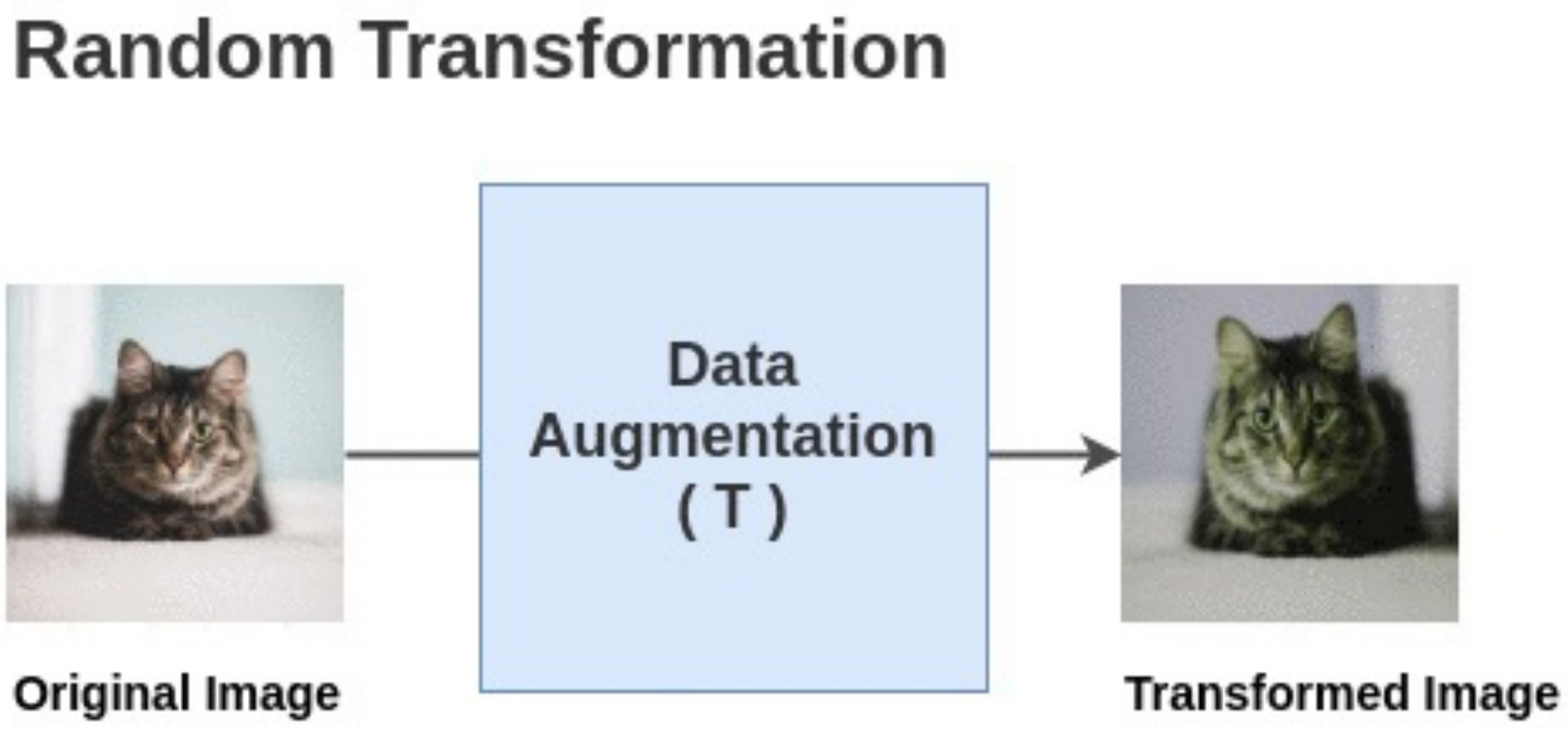







Contrastive Learning

- **Data Augmentation**

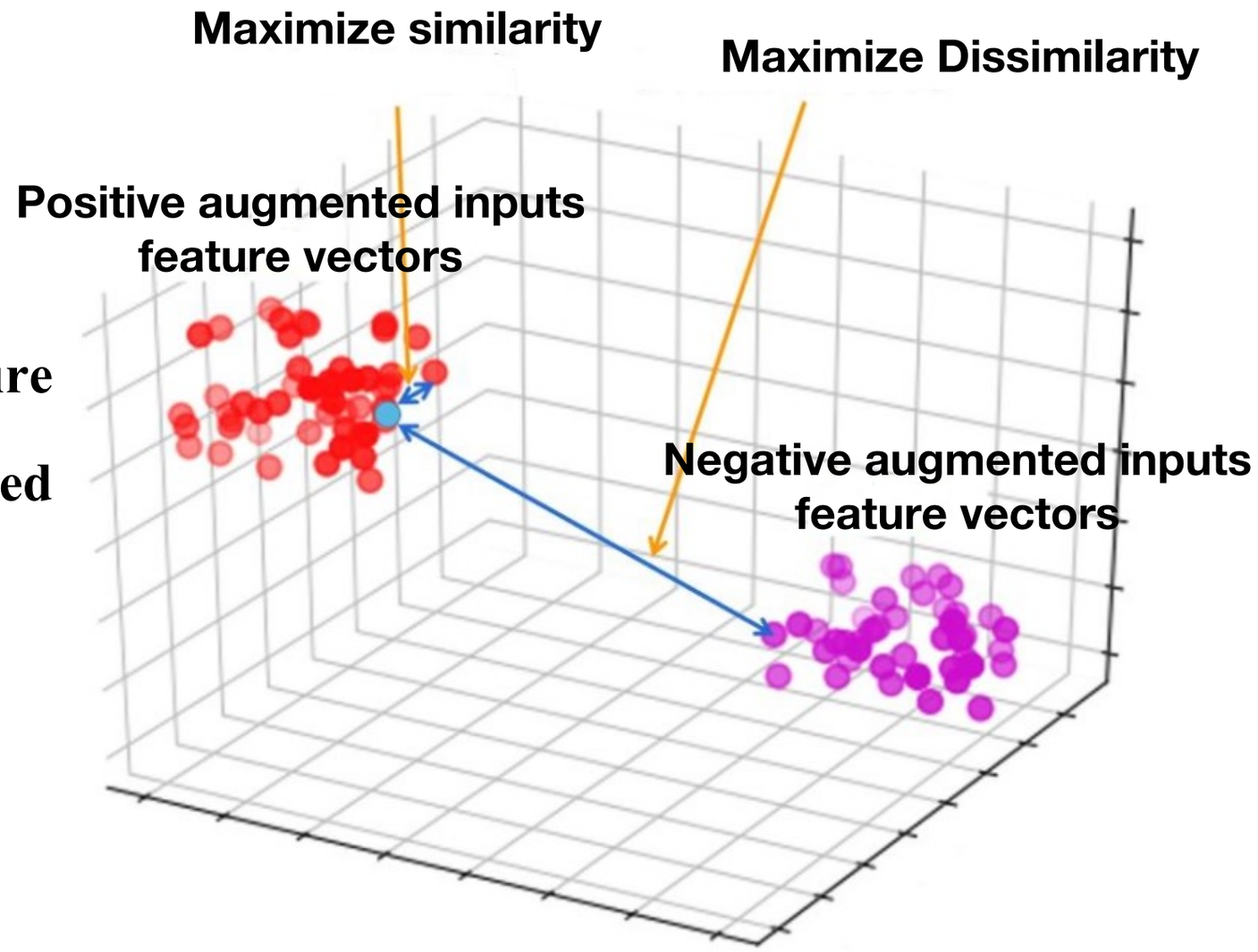


Create random input (called augmented input) by a sequence of random operations

Contrastive Learning

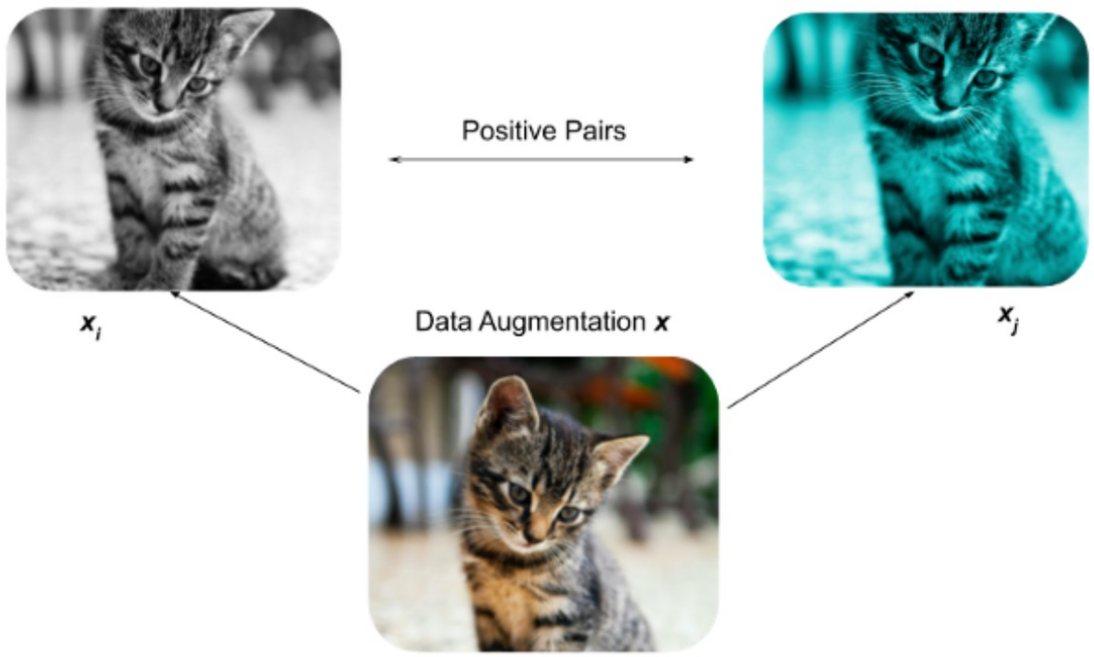
- **Learning Goal**

Outputs similar (or dissimilar) feature vectors for augmented inputs created from the same (or different) input



Contrastive Learning Inference

- **Observation**



Similar Feature Vectors

X_1 $\{n_1, \ldots \ldots\}$
X_2 $\{n_2, \ldots \ldots\}$

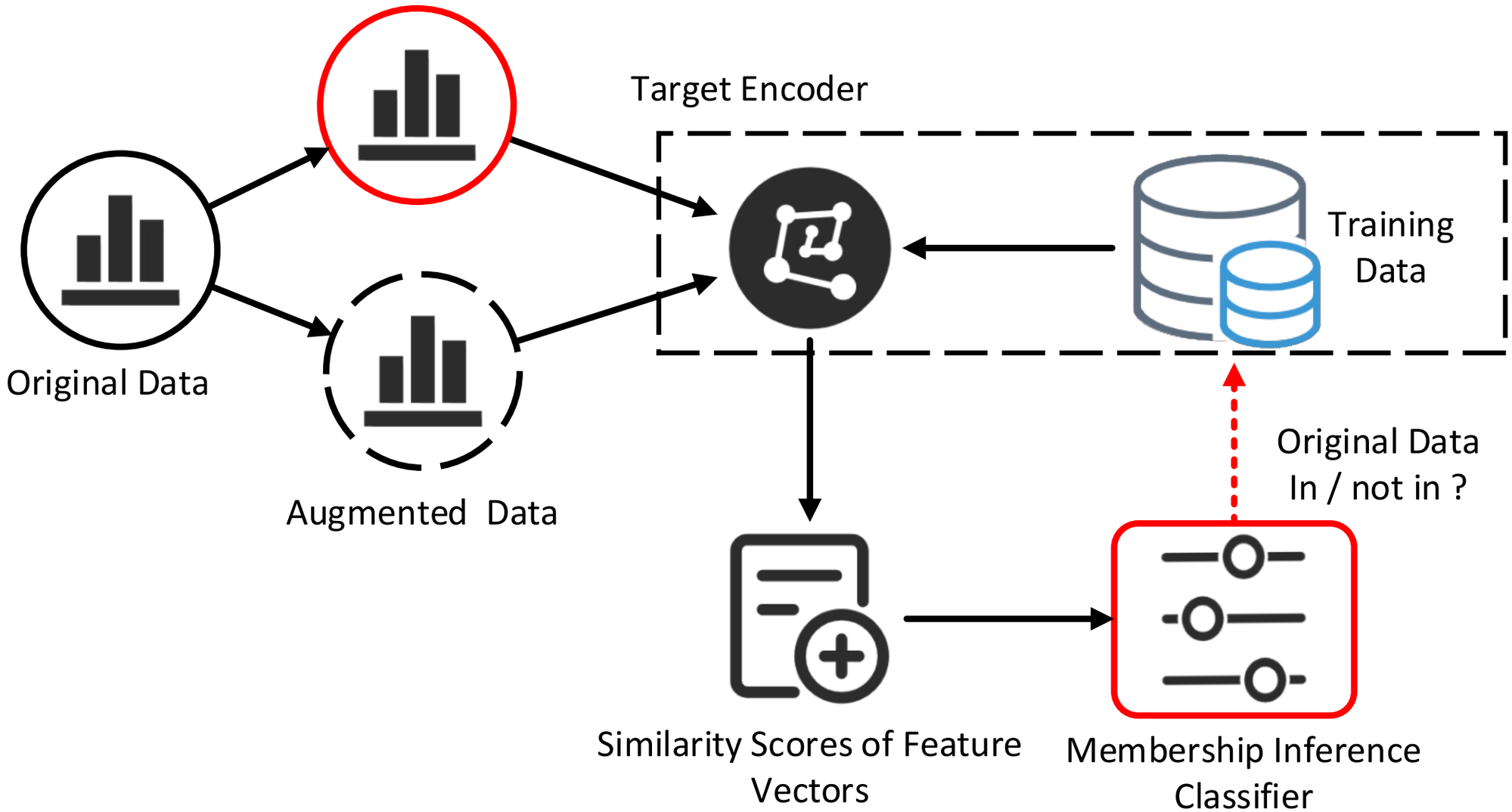
Overfitted

output more similar feature vectors for augmented inputs created from an input

Membership Inference

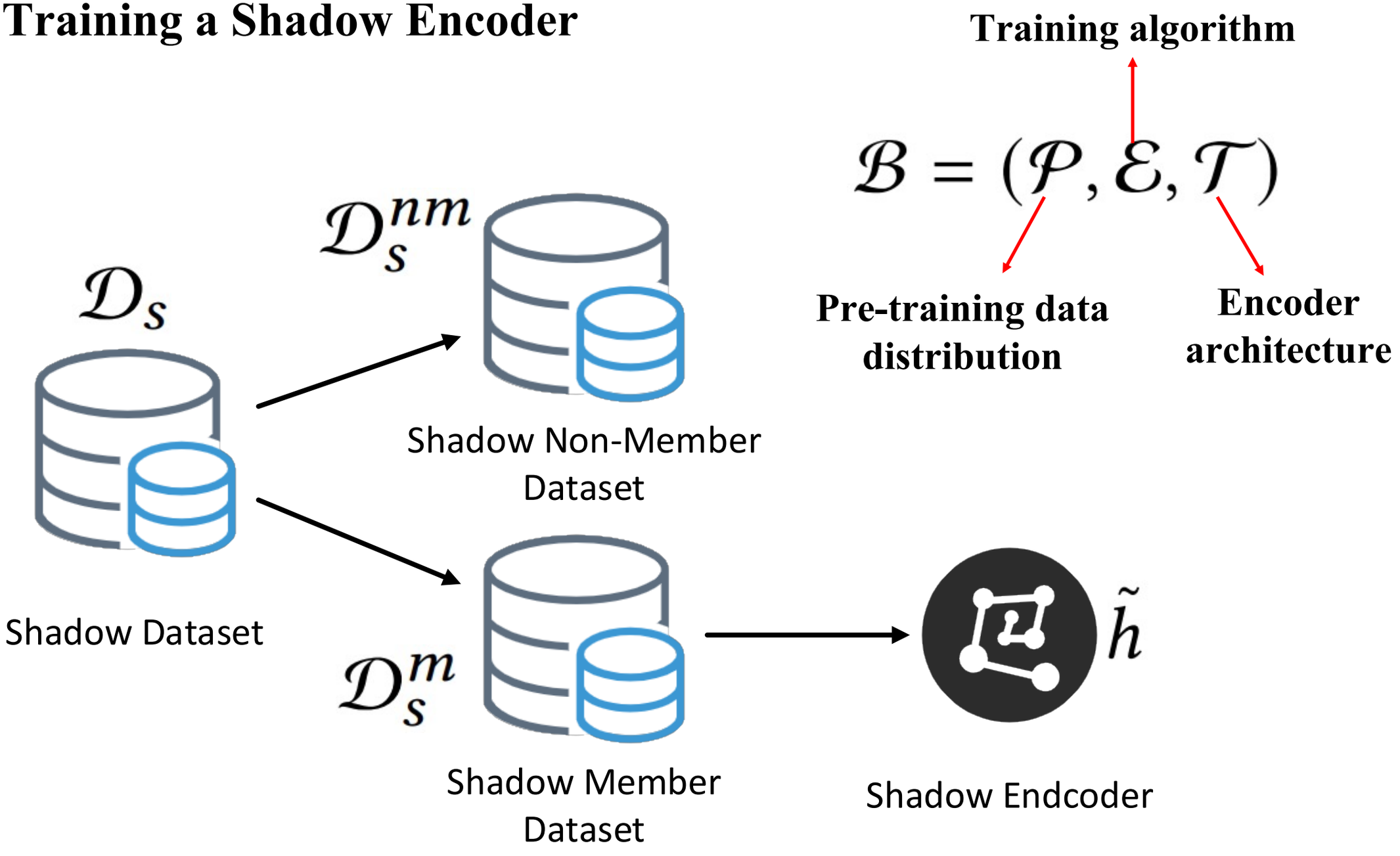
Inference Classifiers

- **Structure of Inference**



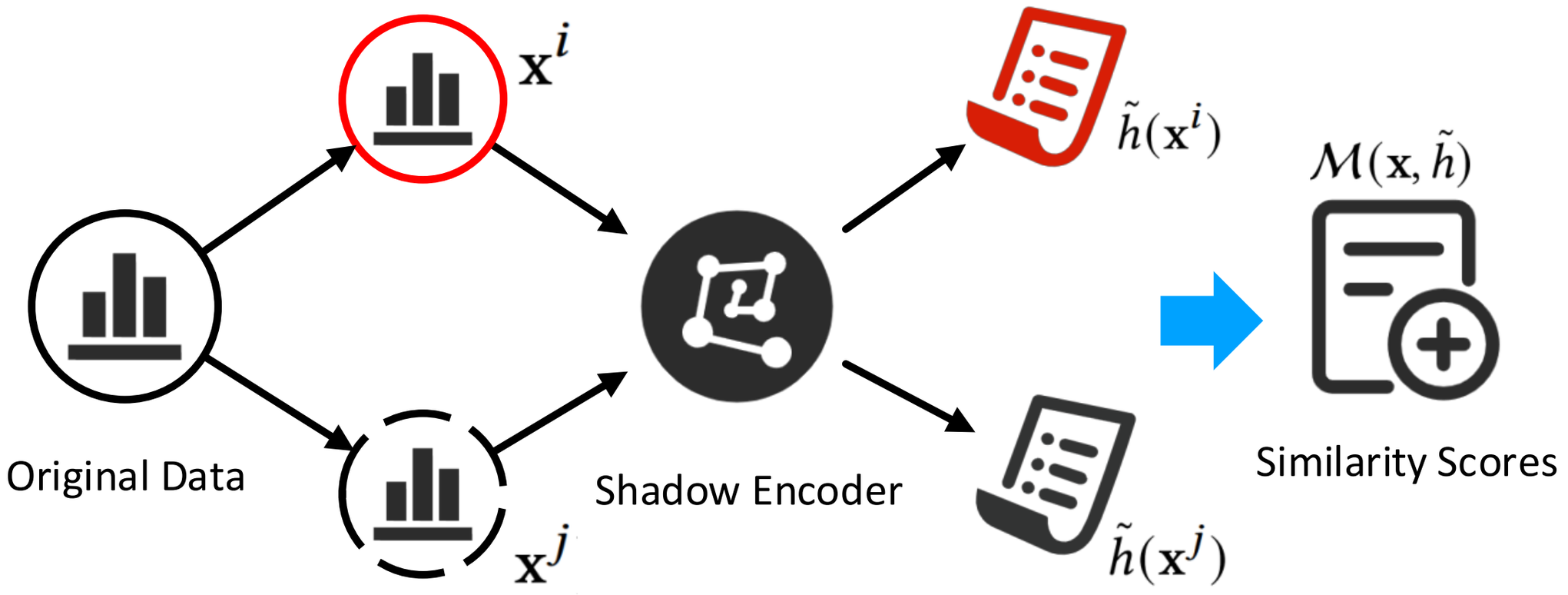
Inference Classifiers

- **Training a Shadow Encoder**



Inference Classifiers

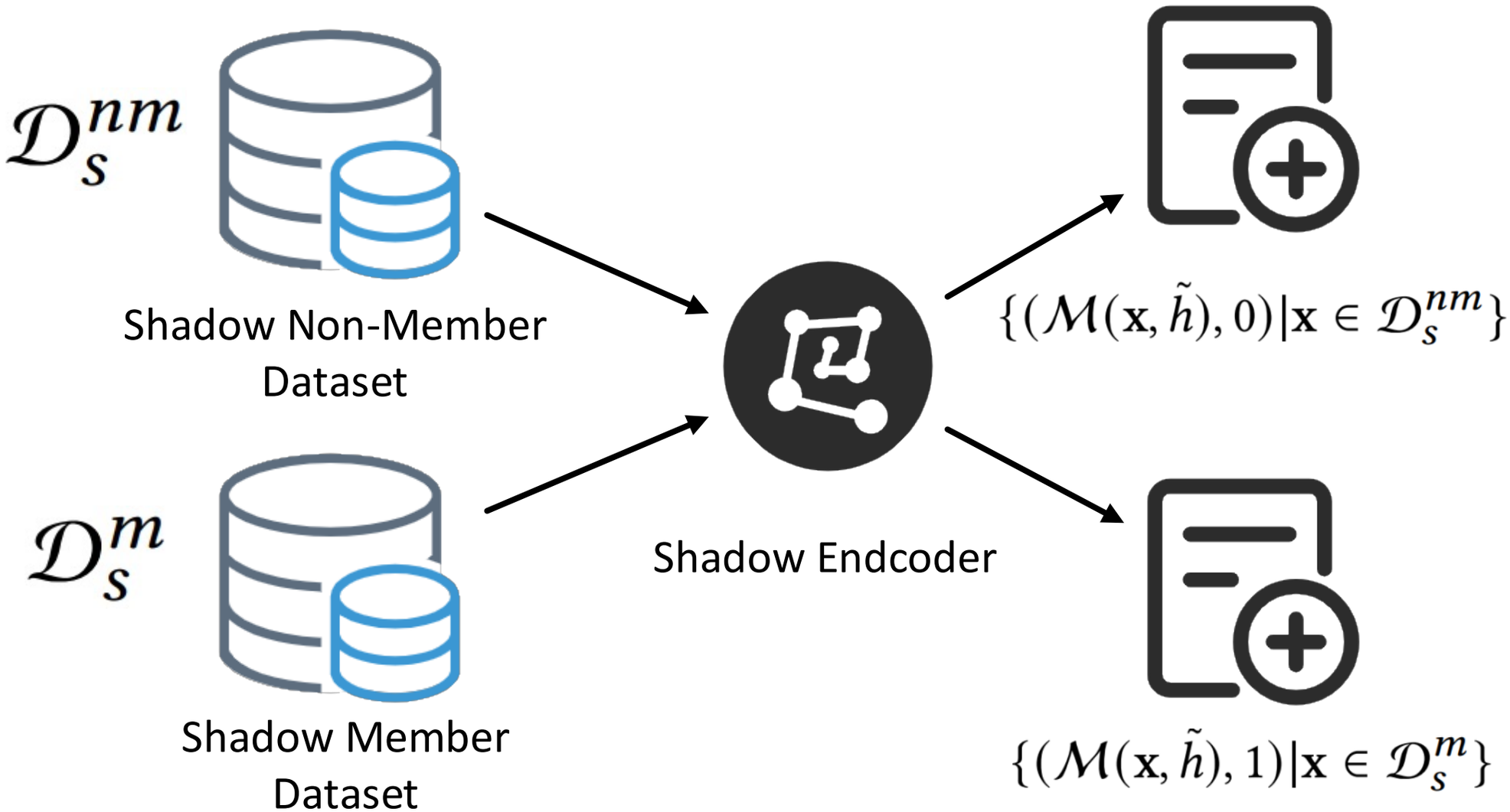
- **Extracting Membership Features**



$$\mathcal{M}(\mathbf{x}, \tilde{h}) = \{S(\tilde{h}(\mathbf{x}^i), \tilde{h}(\mathbf{x}^j)) | i \in [1, n], j \in [1, n], j > i\}$$

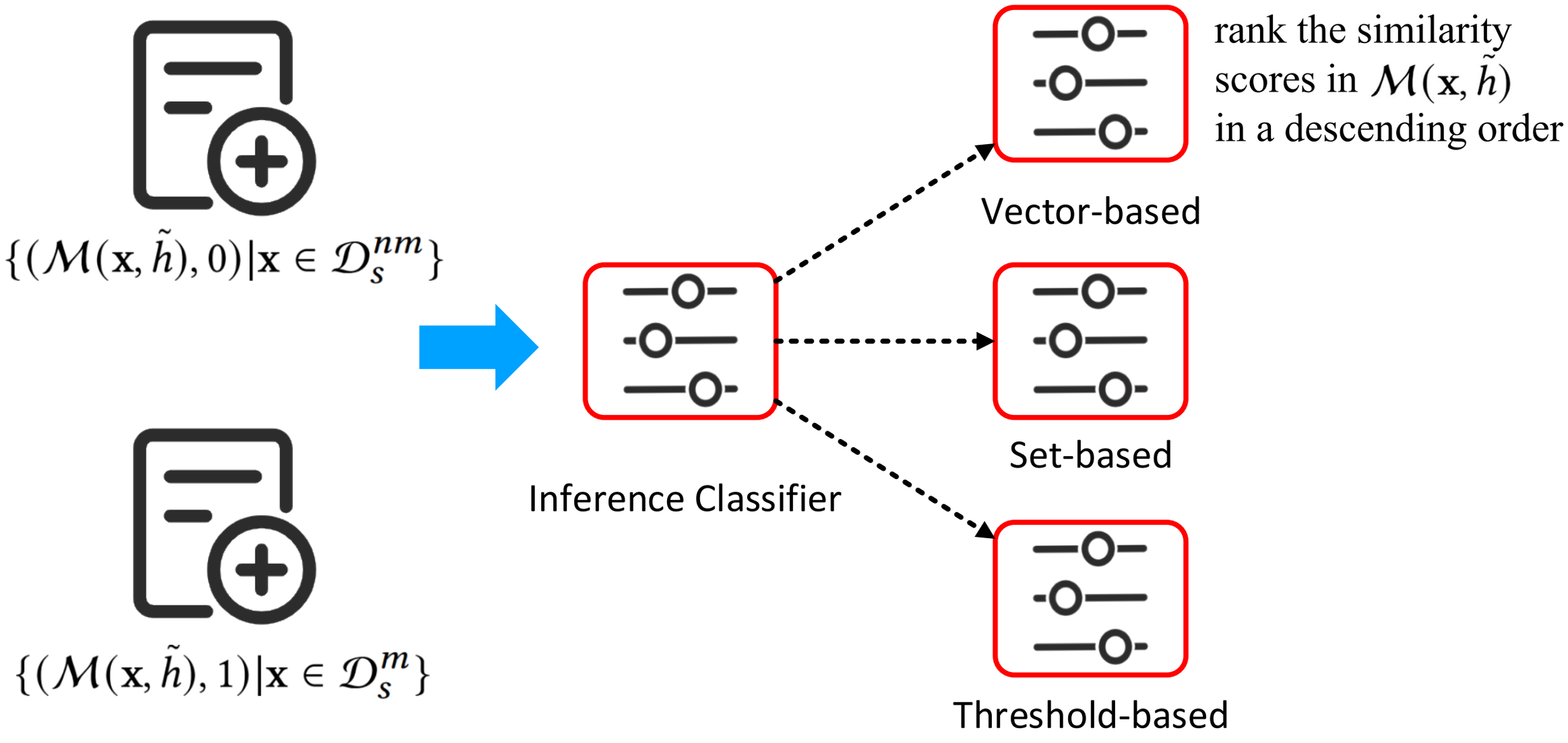
Inference Classifiers

- **Constructing an Inference Training Dataset**



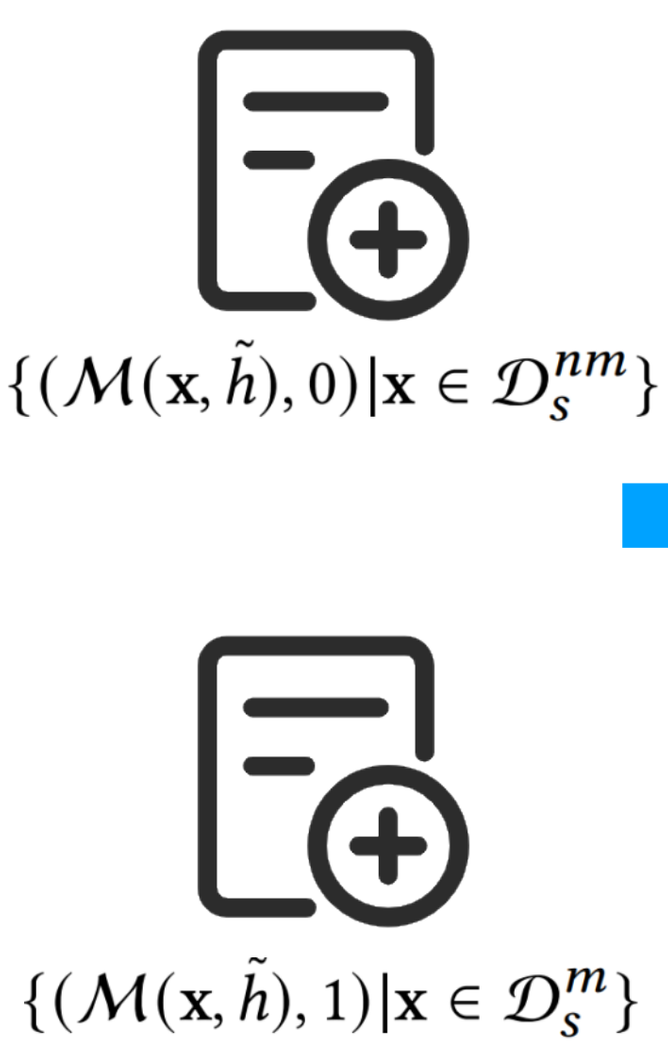
Inference Classifiers

- **Building Inference Classifiers**



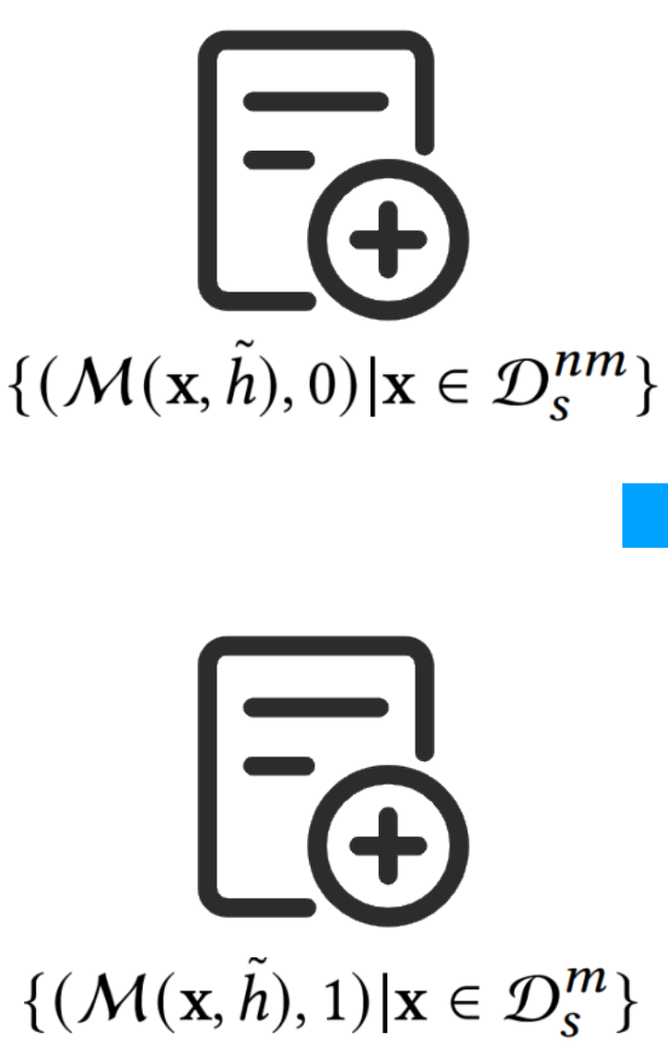
Inference Classifiers

- **Building Inference Classifiers**



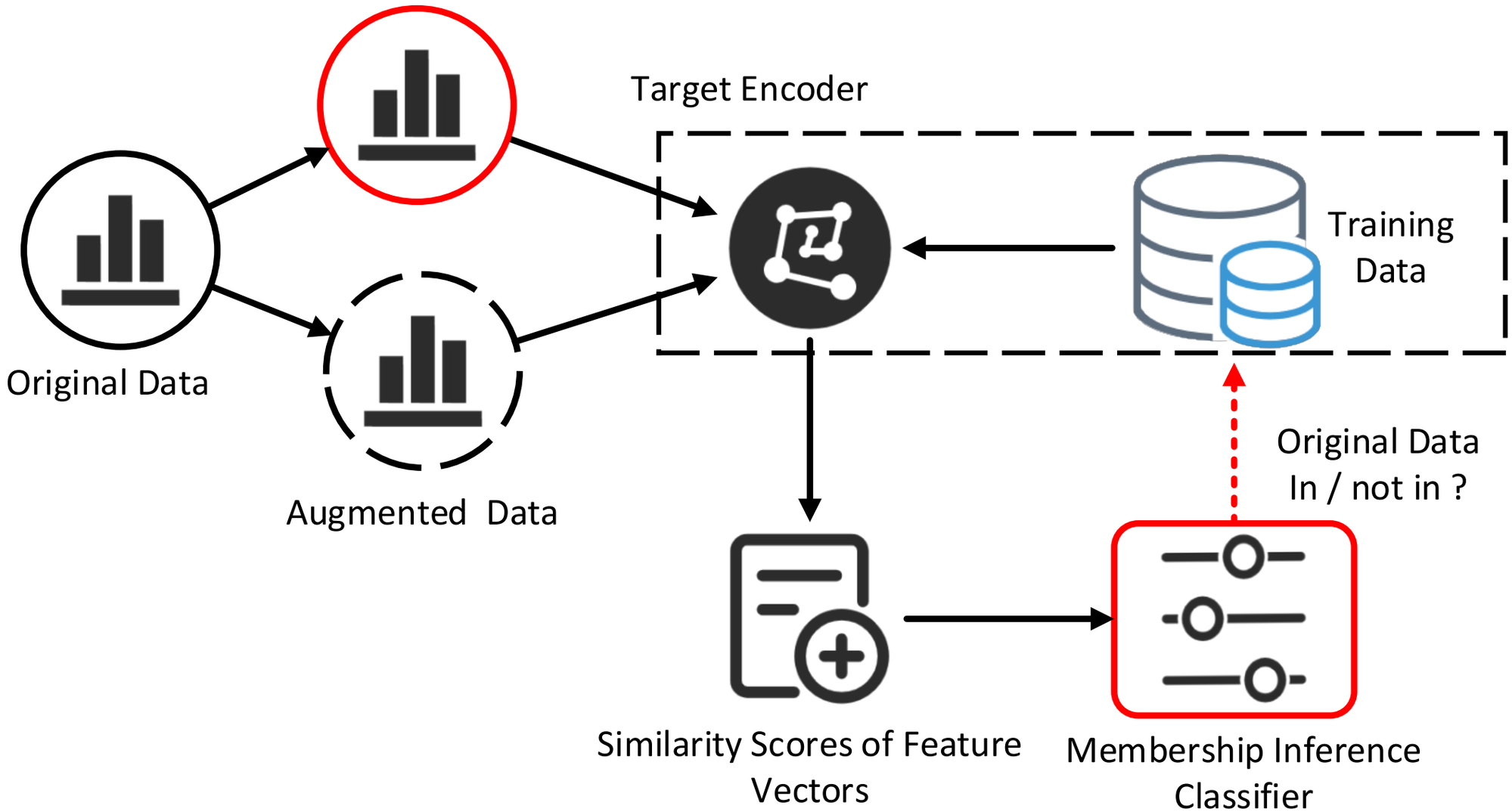
Inference Classifiers

- **Building Inference Classifiers**



Inference Classifiers

- **Structure of Inference**



Evaluation

- **Experimental Setup**

Datasets:

CIFAR10, STL10, and Tiny-ImageNet

Training target encoders:

ResNet18 Architecture, MoCo v_1 algorithm

Training shadow encoders

Data distribution: 20,000 images from same or different dataset.

Architecture: ResNet18 or VGG-11

Algorithm: MoCo v1 or SimCLR

Evaluation

- **Experimental Setup**

Building inference classifiers:

EncoderMI-V: a fully connected network with two hidden layers.

EncoderMI-S: DeepSets

EncoderMI-T: A pre-determined threshold

Evaluation metrics:

Accuracy, precision, and recall

Compared methods:

5 methods aim to infer members of a classifier or embedding model.

Experimental Results

- **Existing Membership Inference Methods are Insufficient**

(a) Baseline-A

Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	55.1	53.4	73.1
STL10	54.3	53.7	62.2
Tiny-ImageNet	47.3	48.2	68.3

(b) Baseline-B

Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	54.6	63.1	58.2
STL10	–	–	–
Tiny-ImageNet	51.8	53.7	47.6

(c) Baseline-C

Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	52.8	54.1	43.1
STL10	50.5	50.1	57.9
Tiny-ImageNet	50.2	52.1	42.3

(d) Baseline-D

Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	50.7	50.6	51.0
STL10	50.1	49.9	50.3
Tiny-ImageNet	49.5	49.3	49.2

(e) Baseline-E

Pre-training dataset	Accuracy	Precision	Recall
CIFAR10	64.5	63.8	67.2
STL10	67.0	65.7	71.3
Tiny-ImageNet	68.6	67.8	70.8

Experimental Results

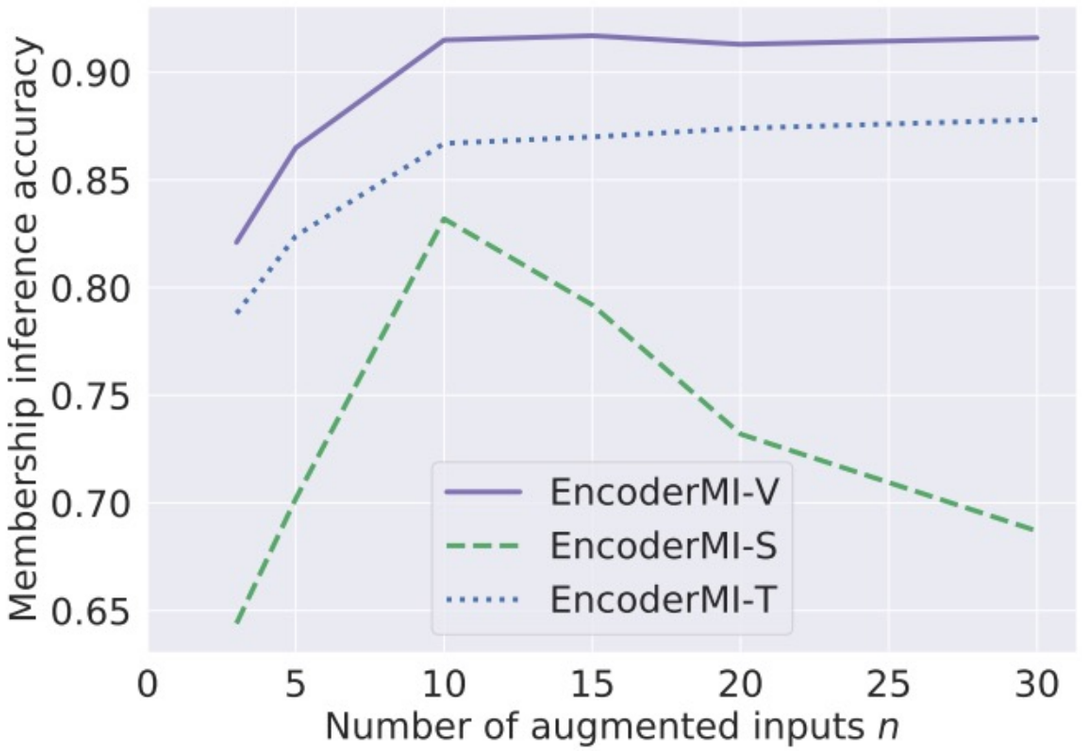
- **The Proposed Methods are Effective**

Pre-training data distribution	Encoder architecture	Training algorithm	Accuracy			Precision		
			Encod-erMI-V	Encod-erMI-S	Encod-erMI-T	Encod-erMI-V	Encod-erMI-S	Encod-erMI-T
×	×	×	88.7 (1.81)	84.9 (1.73)	85.3 (1.67)	86.0 (1.98)	81.5 (2.03)	81.8 (1.74)
√	×	×	93.0 (1.74)	88.2 (1.68)	90.0 (1.44)	90.1 (1.39)	85.4 (1.45)	86.8 (1.23)
×	√	×	89.1 (1.63)	86.4 (1.64)	85.7 (1.29)	83.3 (1.88)	84.0 (1.84)	80.1 (1.63)
×	×	√	94.1 (1.07)	91.3 (1.03)	94.1 (0.91)	90.7 (0.88)	90.3 (0.87)	93.5 (0.79)
√	√	×	94.4 (1.38)	90.4 (1.33)	91.5 (1.26)	97.4 (0.96)	94.1 (0.91)	93.8 (0.91)
√	×	√	96.1 (0.67)	91.6 (0.69)	94.1 (0.54)	93.8 (0.73)	90.4 (0.68)	94.2 (0.62)
×	√	√	94.5 (0.59)	91.8 (0.56)	92.0 (0.53)	92.3 (0.93)	94.4 (0.91)	94.1 (0.86)
√	√	√	96.5 (0.51)	92.0 (0.47)	94.3 (0.43)	96.6 (0.72)	92.9 (0.59)	94.9 (0.57)

Recall		
Encod-erMI-V	Encod-erMI-S	Encod-erMI-T
90.1 (1.96)	95.3 (1.67)	95.9 (1.44)
97.8 (1.26)	93.2 (1.22)	97.1 (1.11)
96.3 (1.22)	91.1 (1.29)	96.1 (1.08)
97.4 (0.92)	91.3 (1.22)	95.6 (0.93)
90.8 (1.46)	87.1 (1.37)	89.4 (1.22)
97.6 (0.99)	92.3 (1.02)	95.7 (0.88)
96.7 (0.92)	90.6 (0.79)	92.7 (0.77)
97.0 (0.93)	92.4 (0.89)	93.2 (0.91)

Experimental Results

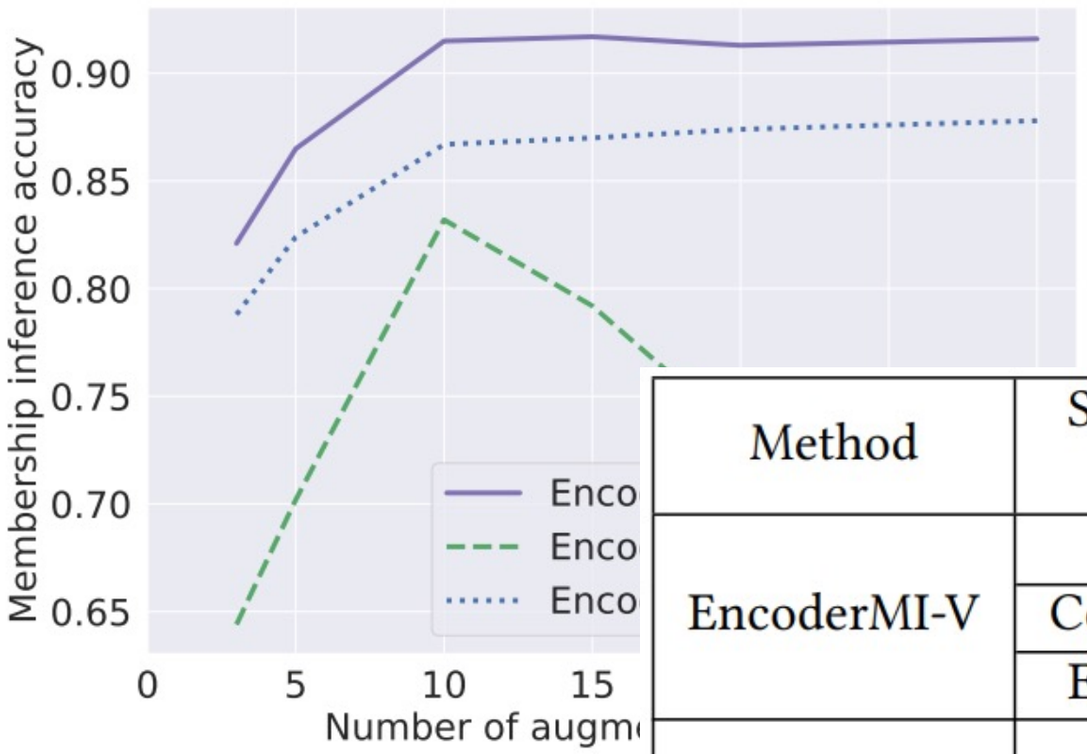
- **Impact of Augmentation and Similarity Metric**



	imilarity metric	Accuracy	Precision	Recall
	Cosine	91.5	90.0	93.5
	orrelation	89.3	87.9	92.2
	uclidean	88.9	85.3	94.5
EncoderMI-S	Cosine	83.2	80.5	87.9
	Correlation	76.3	75.5	78.5
	Euclidean	75.8	73.6	81.4
EncoderMI-T	Cosine	86.7	85.7	89.0
	Correlation	80.6	79.8	81.6
	Euclidean	80.7	80.1	82.5

Experimental Results

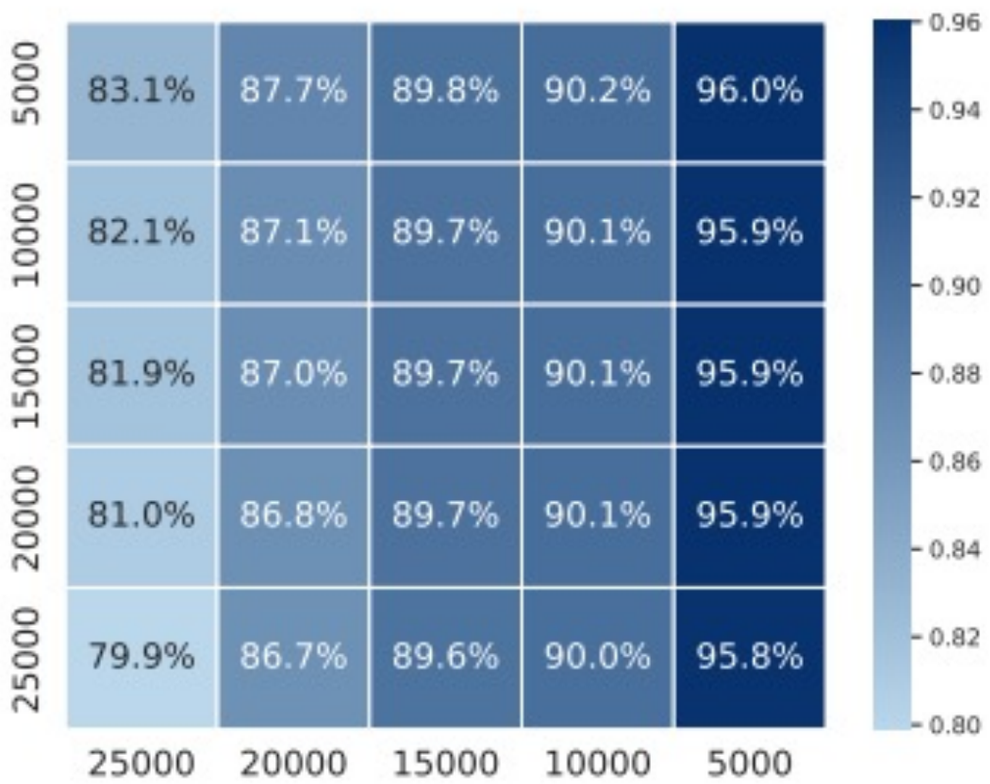
- **Impact of Augmentation and Similarity Metric**



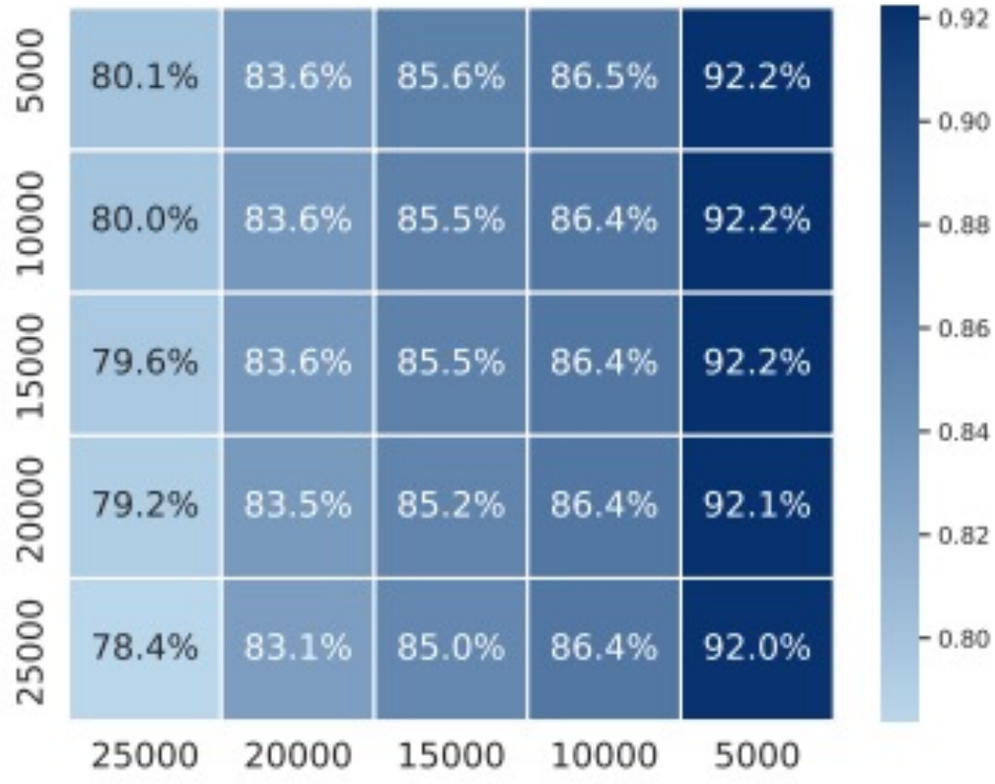
Method	Similarity metric	Accuracy	Precision	Recall
EncoderMI-V	Cosine	91.5	90.0	93.5
	Correlation	89.3	87.9	92.2
	Euclidean	88.9	85.3	94.5
EncoderMI-S	Cosine	83.2	80.5	87.9
	Correlation	76.3	75.5	78.5
	Euclidean	75.8	73.6	81.4
EncoderMI-T	Cosine	86.7	85.7	89.0
	Correlation	80.6	79.8	81.6
	Euclidean	80.7	80.1	82.5

Experimental Results

- **Impact of the size of the pre-training and shadow datasets**



(a) EncoderMI-V

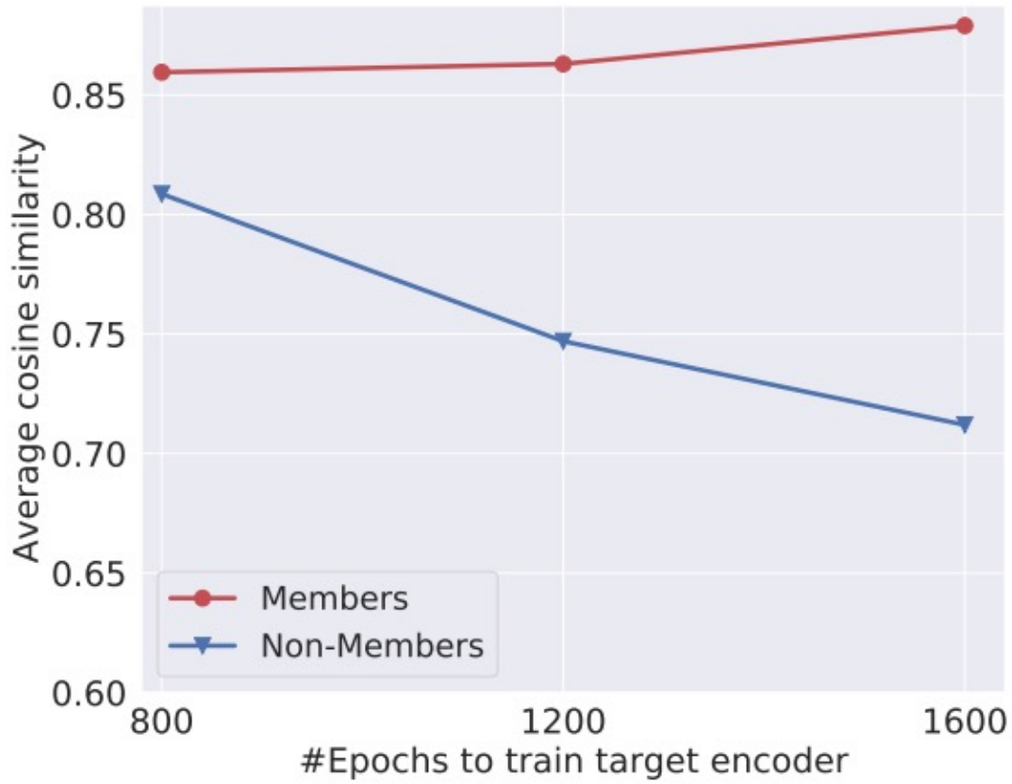


(b) EncoderMI-S

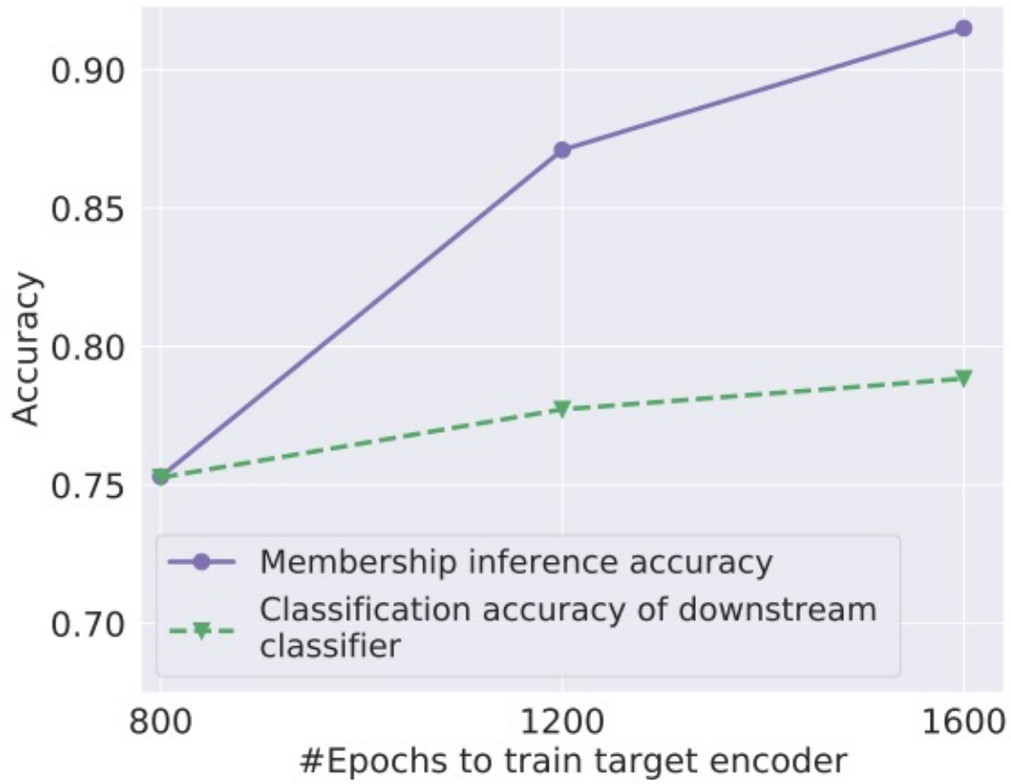
pre-training dataset (x-axis) and the shadow dataset (y-axis)

COUNTERMEASURES

- **Preventing Overfitting via Early Stopping**



(a) Overfitting of the target encoder



(b) Inferability-utility tradeoff