Opinion Mining by Transformation-Based Domain Adaptation

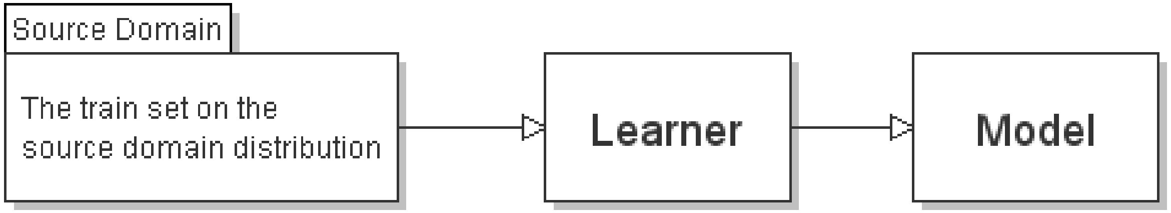
Róber Ormándi, István Hegedűs, Richárd Farkas

TSD-2010

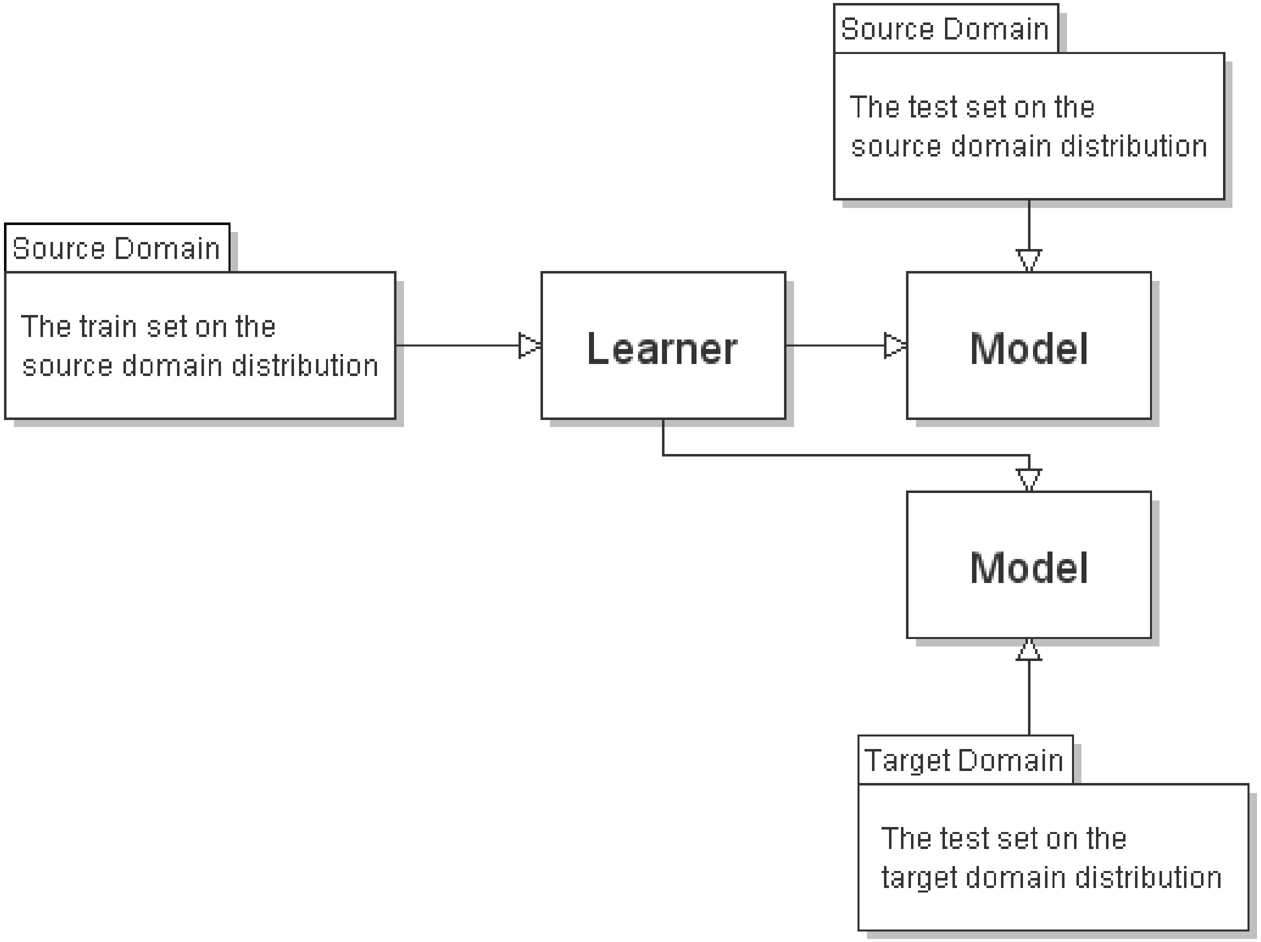
The Domain Adaptation Task

- Given two datasets from different domains, namely the source (S) and the target (T) domains
- The source has a huge, the target often has just a few amount of labeled samples (|S|>>|T|)
- The task is to set the labels of the target domain as precisely as possible, using the labeled source samples

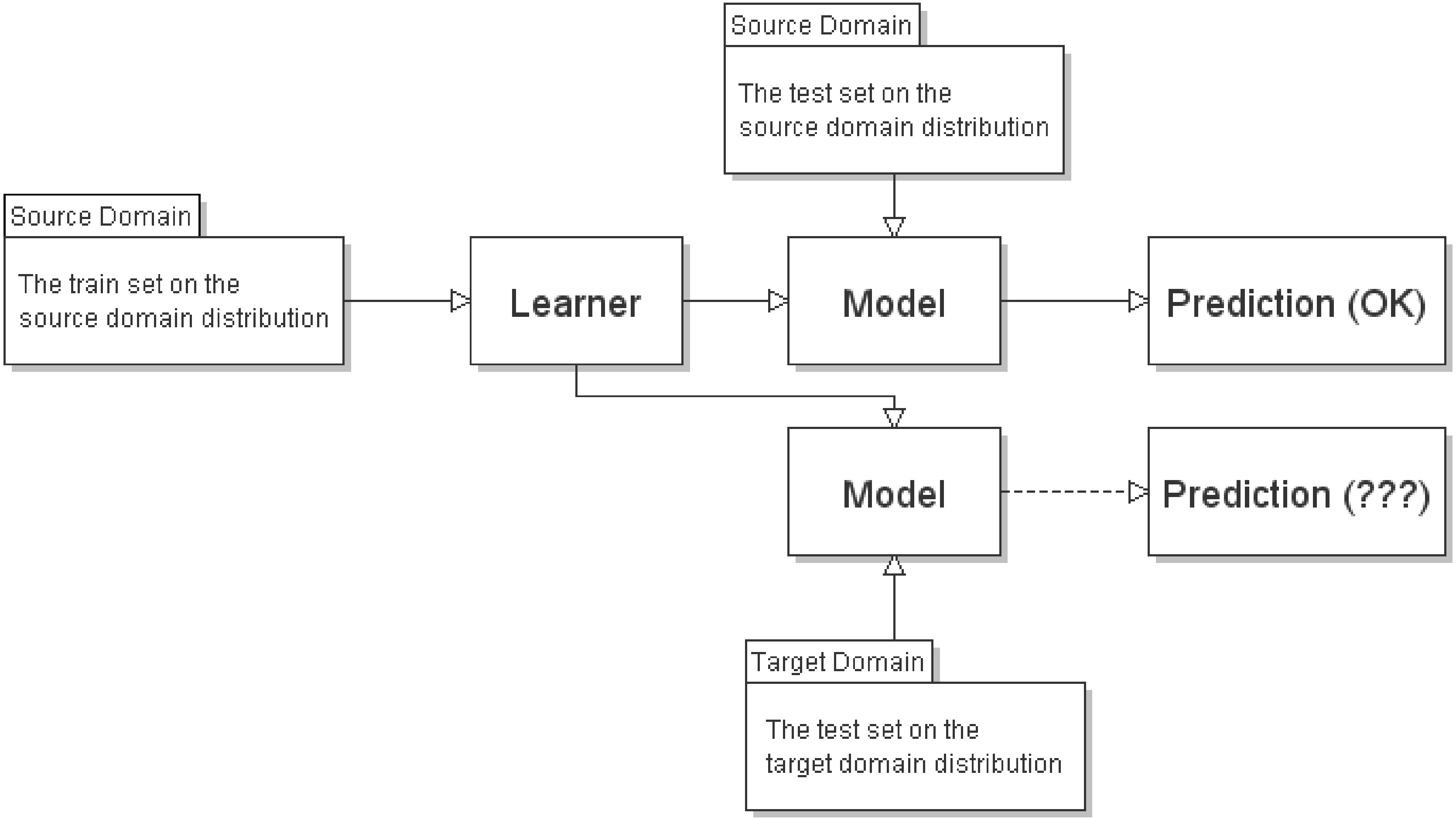
Basic Machine Learning



Basic Machine Learning


Source Domain
The test set on the source domain distribution
Source Domain
The train set on the source domain distribution
Learner
Model
Model
Target Domain
The test set on the target domain distribution

Basic Machine Learning



Source Domain

The test set on the source domain distribution

Source Domain

The train set on the source domain distribution

Learner

Model

Prediction (OK)

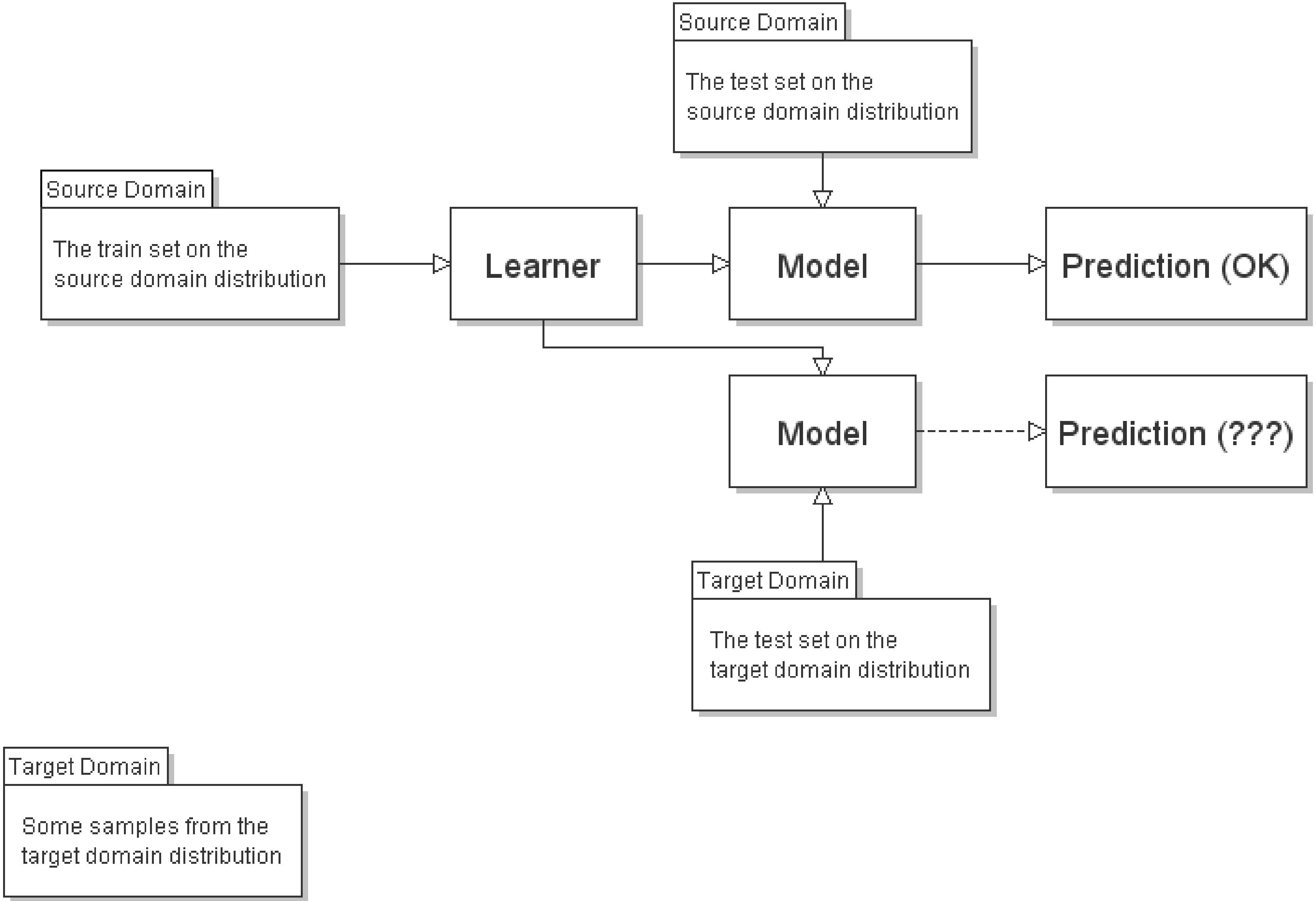
Model

Prediction (???)

Target Domain

The test set on the target domain distribution

Domain Adaptation



Source Domain

The train set on the source domain distribution

Learner

Model

Prediction (OK)

Source Domain

The test set on the source domain distribution

Model

Prediction (???)

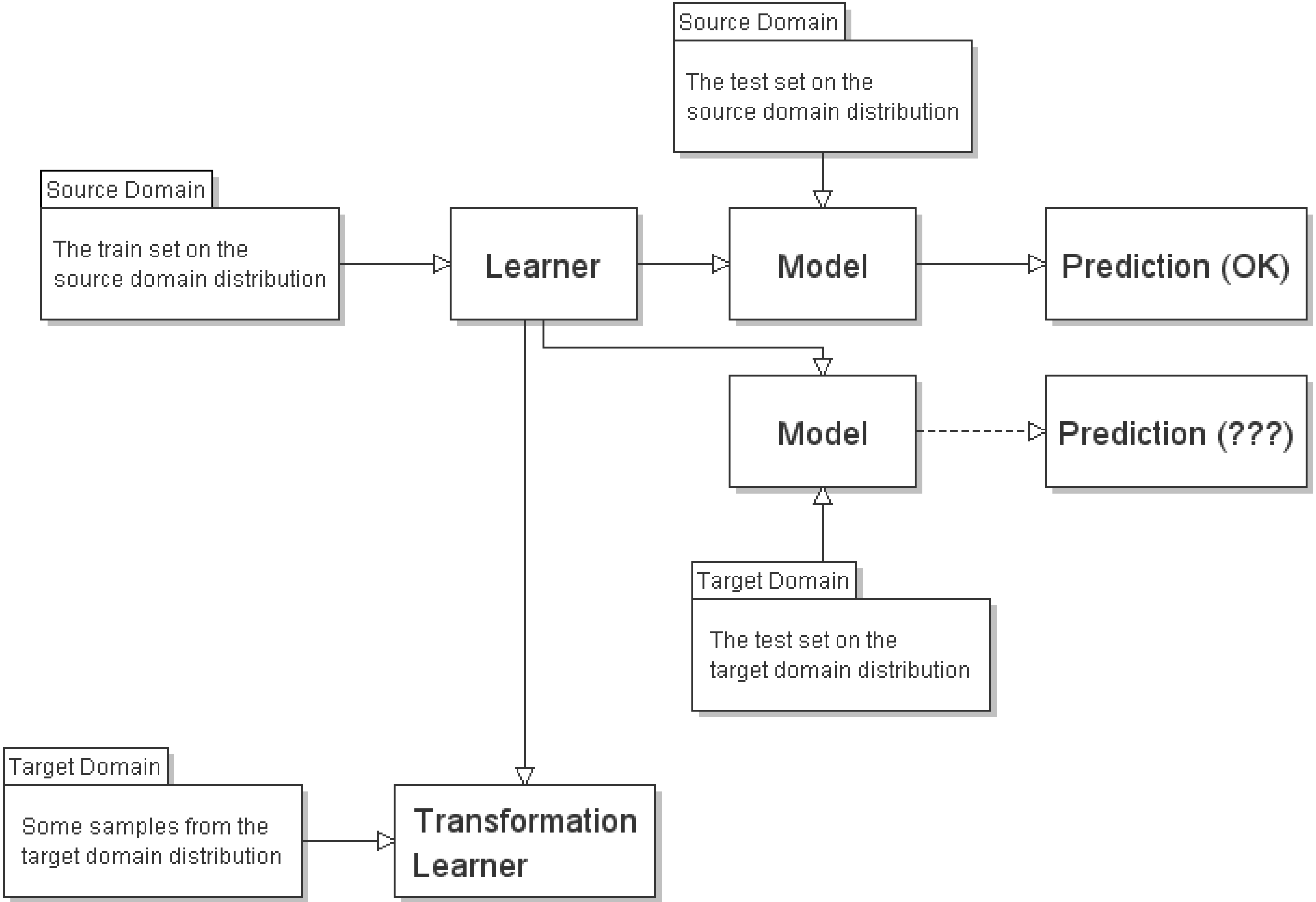
Target Domain

The test set on the target domain distribution

Target Domain

Some samples from the target domain distribution

Our Approach



Source Domain

The test set on the source domain distribution

Source Domain

The train set on the source domain distribution

Learner

Model

Prediction (OK)

Model

Prediction (???)

Target Domain

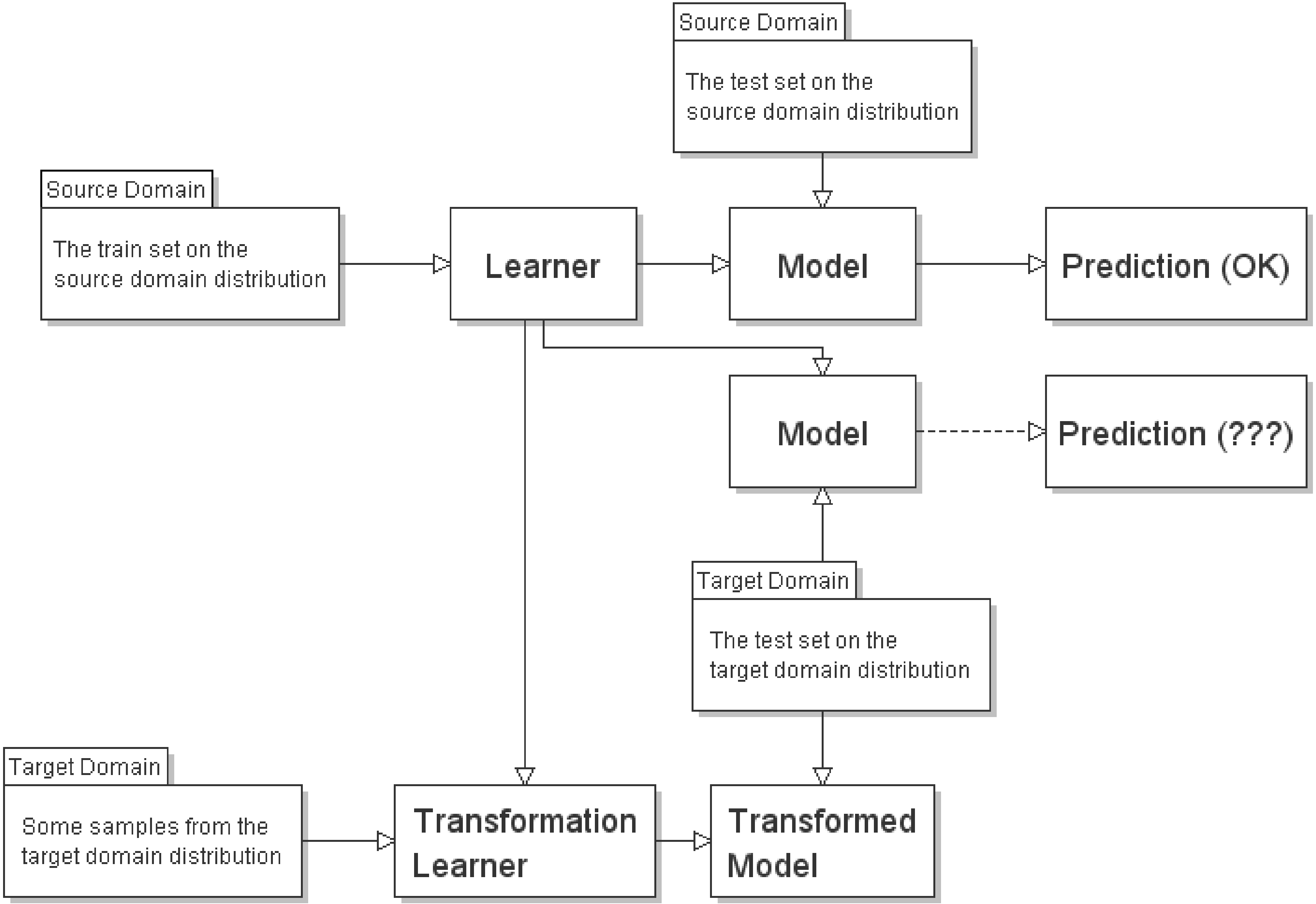
The test set on the target domain distribution

Target Domain

Some samples from the target domain distribution

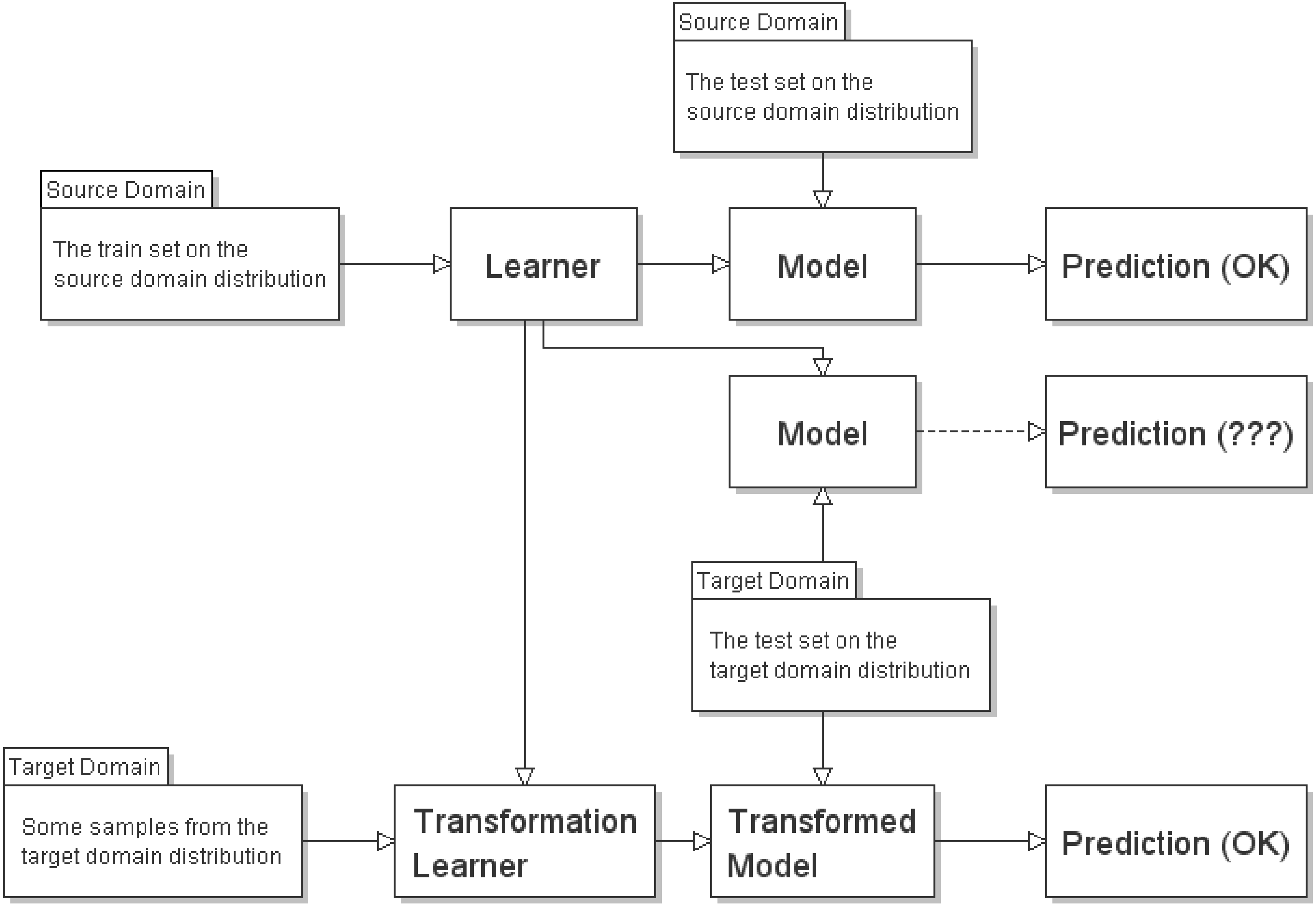
Transformation Learner

Our Approach



Source Domain
The test set on the source domain distribution
Source Domain
The train set on the source domain distribution
Learner
Model
Prediction (OK)
Model
Prediction (???)
Target Domain
The test set on the target domain distribution
Target Domain
Some samples from the target domain distribution
Transformation Learner
Transformed Model

Our Approach



Source Domain

The test set on the source domain distribution

Source Domain

The train set on the source domain distribution

Learner

Model

Prediction (OK)

Model

Prediction (???)

Target Domain

The test set on the target domain distribution

Target Domain

Some samples from the target domain distribution

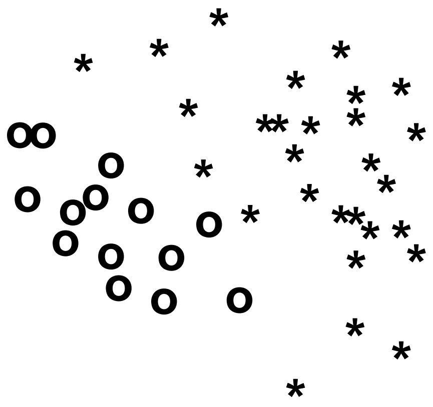
Transformation Learner

Transformed Model

Prediction (OK)

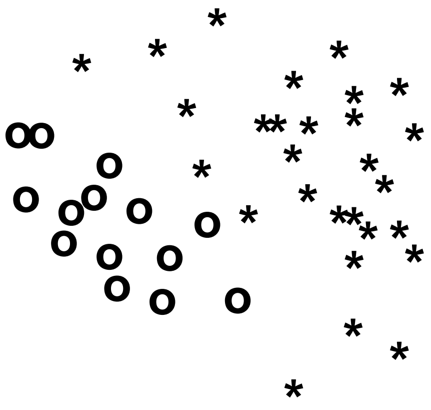
Transformation Based Domain Adaptation

Source Domain



Transformation Based Domain Adaptation

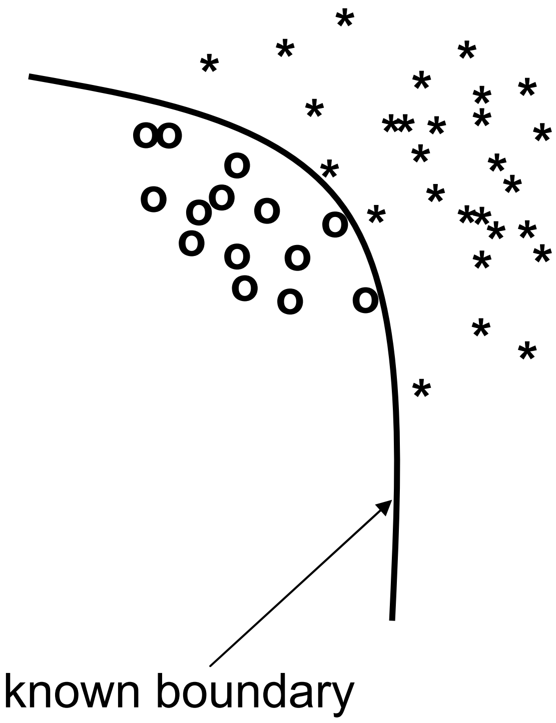
Source Domain



use a machine learning method

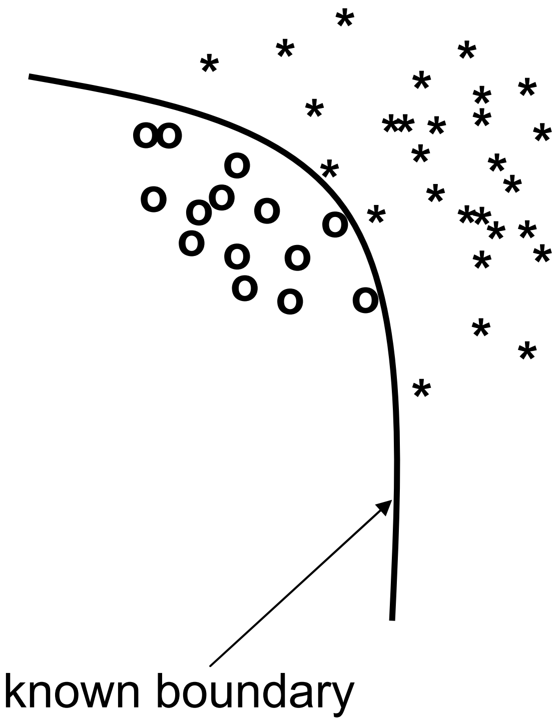
Transformation Based Domain Adaptation

Source Domain

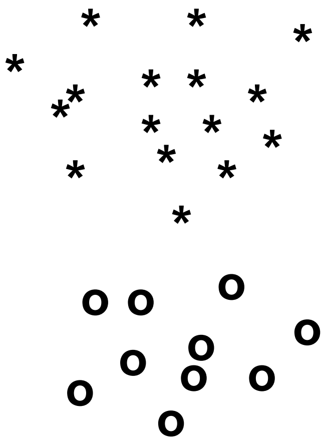


Transformation Based Domain Adaptation

Source Domain

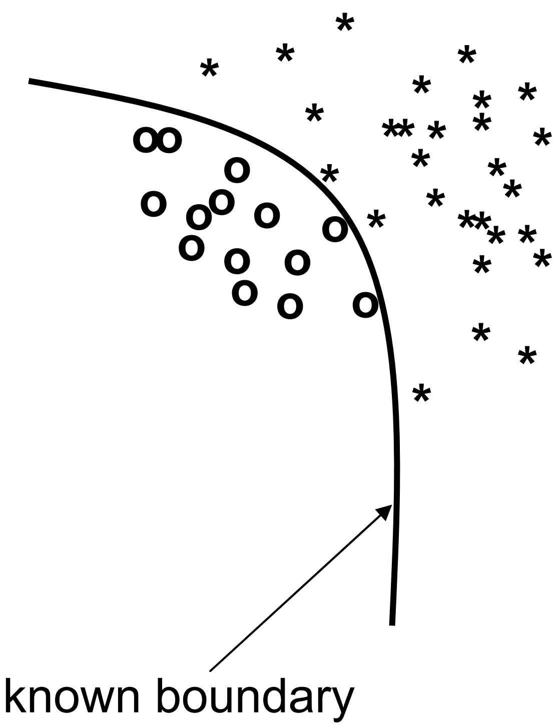


Target Domain



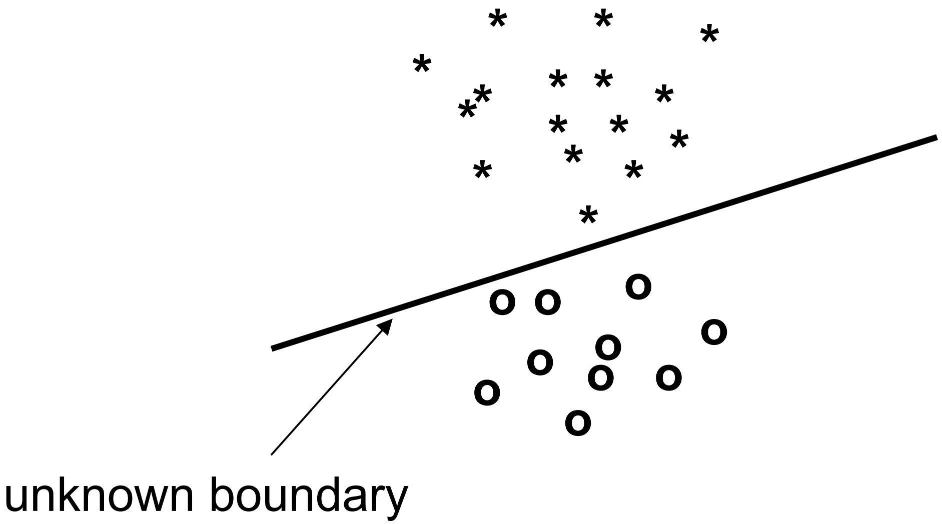
Transformation Based Domain Adaptation

<u>Source Domain</u>



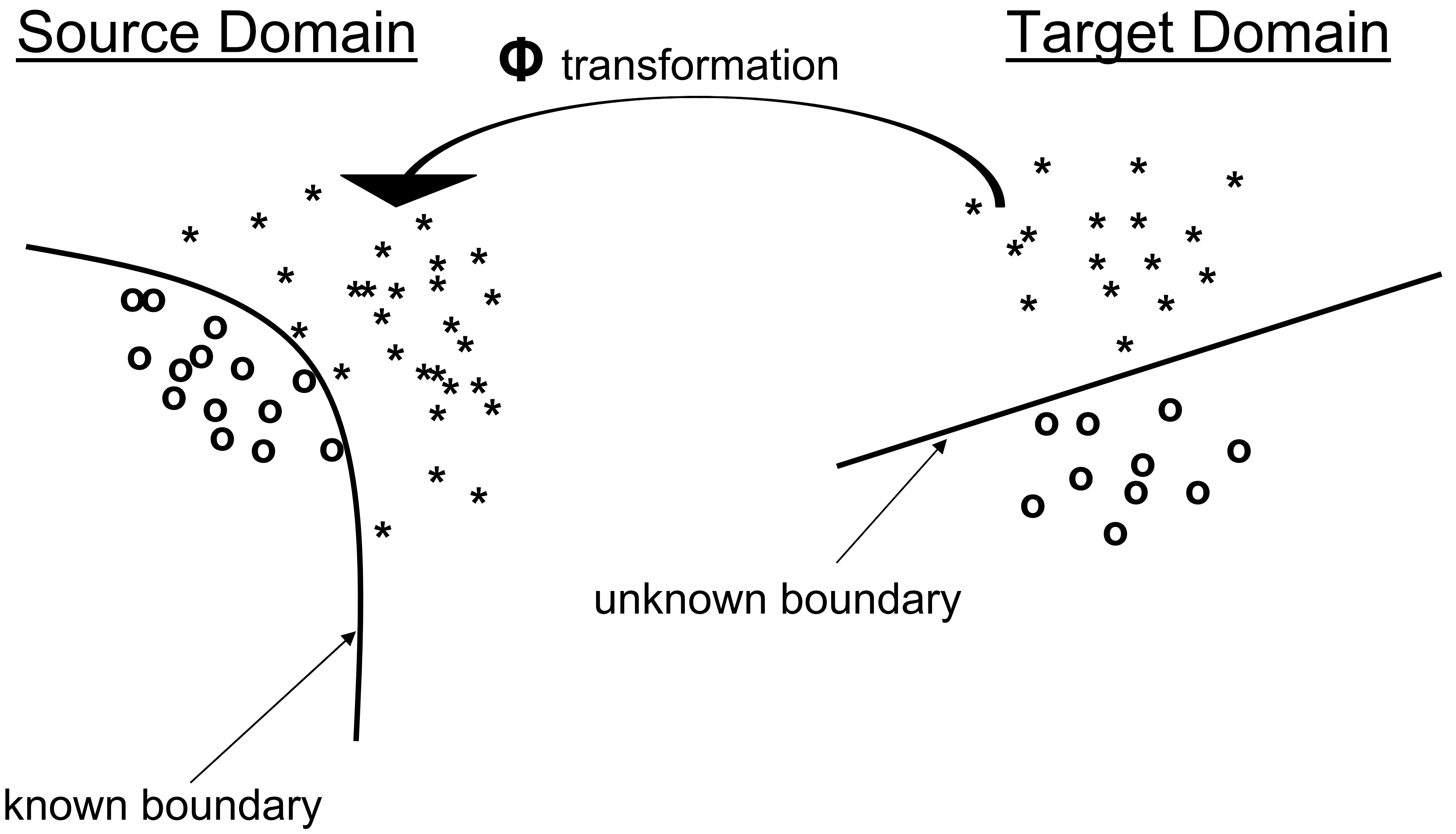
known boundary

<u>Target Domain</u>

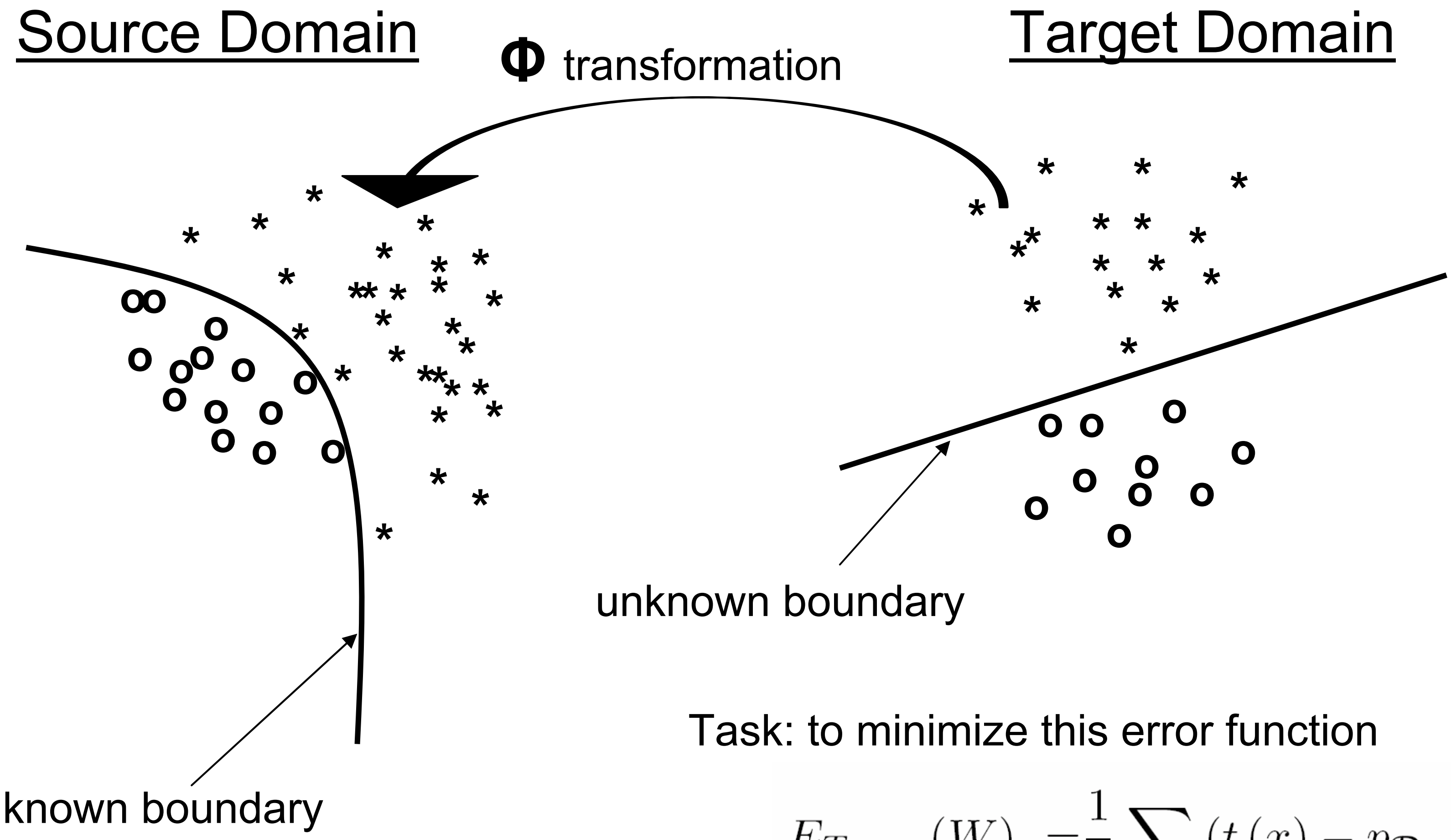


unknown boundary

Transformation Based Domain Adaptation



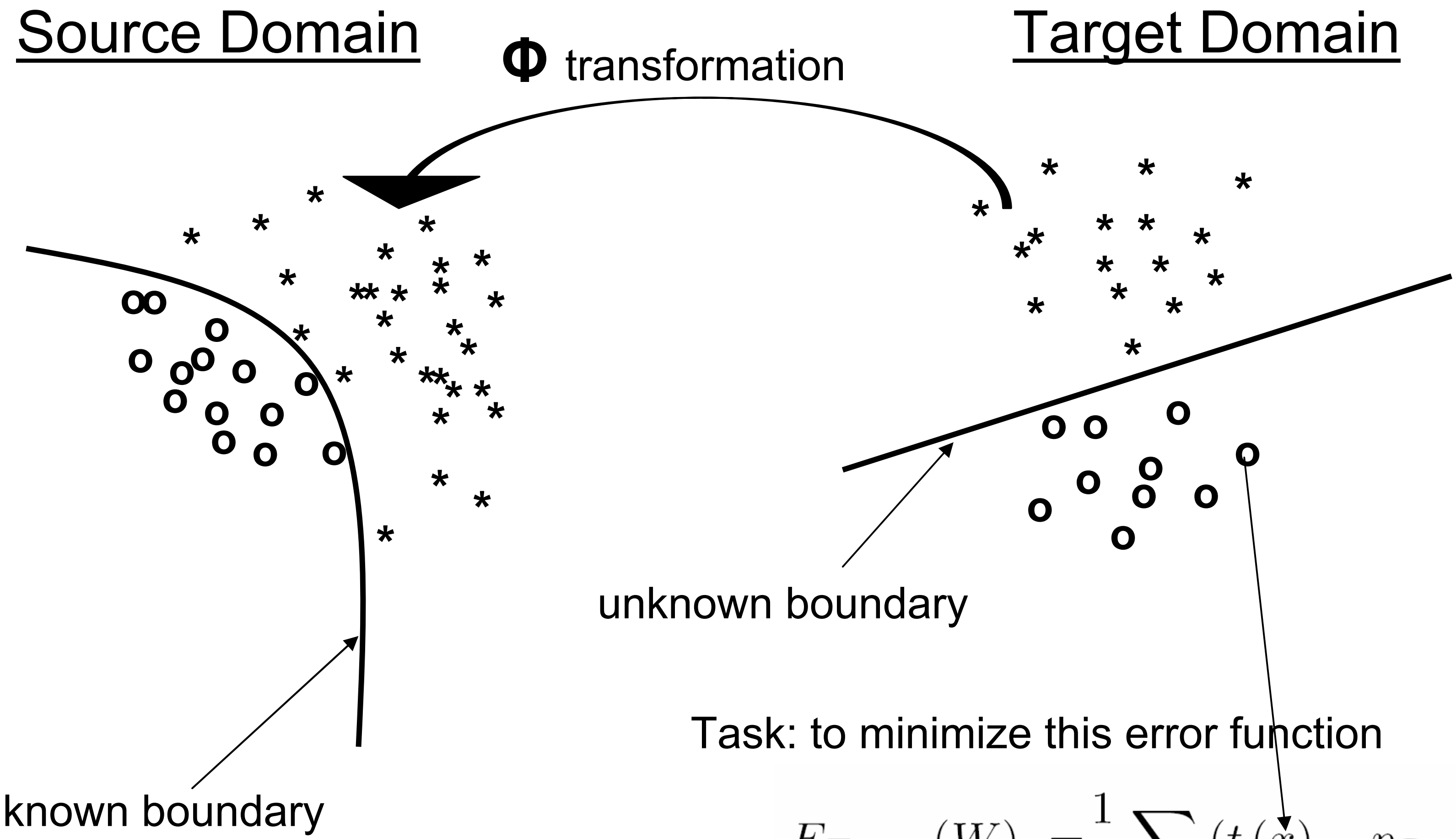
Transformation Based Domain Adaptation



Task: to minimize this error function

$$E_{T,p_{\mathcal{D}_S}}(W) \; = \frac{1}{2} \sum_{x \in T} \left(t(x) - p_{\mathcal{D}_S}(Wx) \right)^2$$

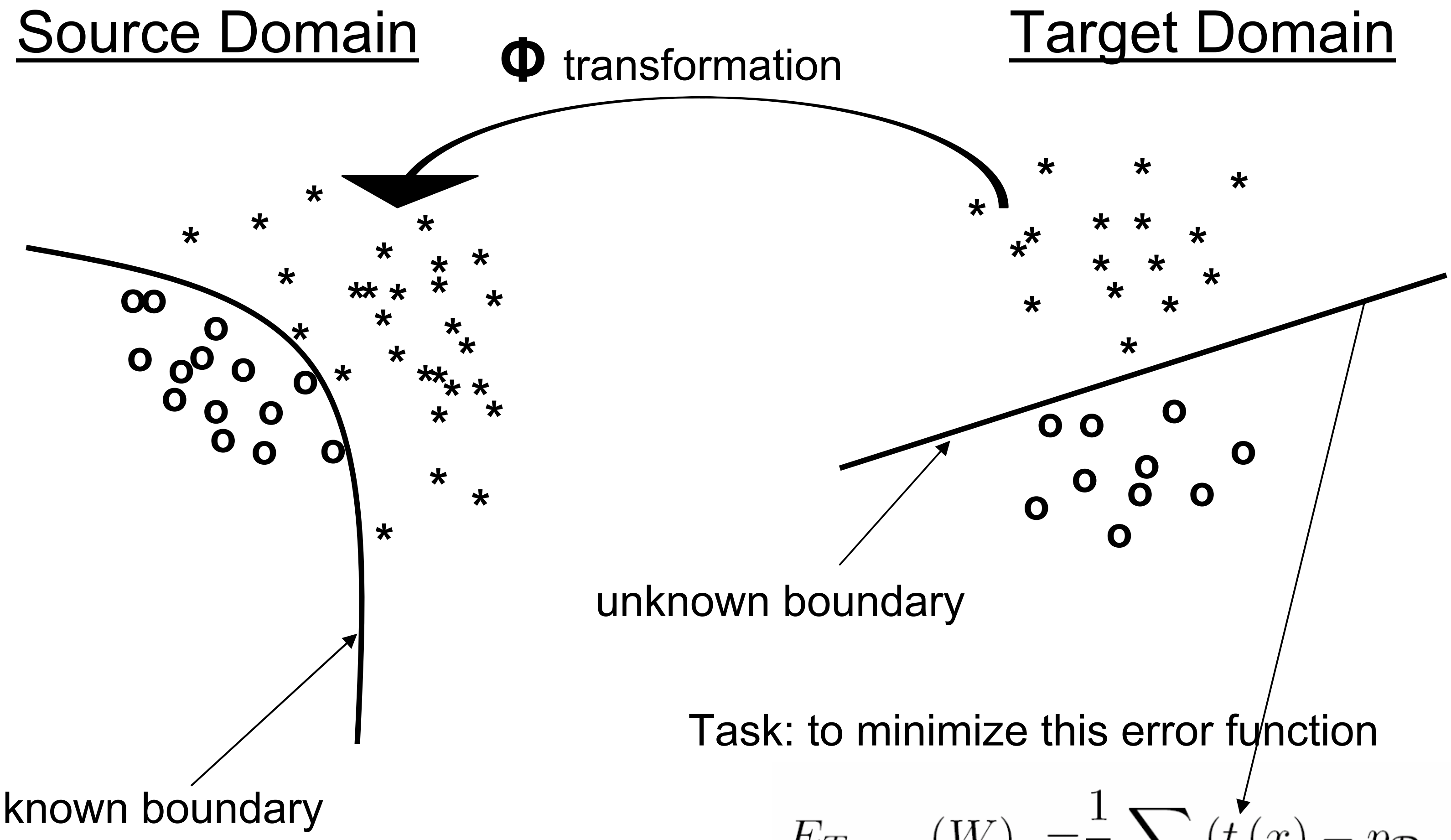
Transformation Based Domain Adaptation



Task: to minimize this error function

$$E_{T,p_{\mathcal{D}_S}}(W) \ =\frac{1}{2}\sum_{x\in T}\left(t\left(x\right)-p_{\mathcal{D}_S}\left(Wx\right)\right)^2$$

Transformation Based Domain Adaptation



Source Domain

Target Domain

Φ transformation

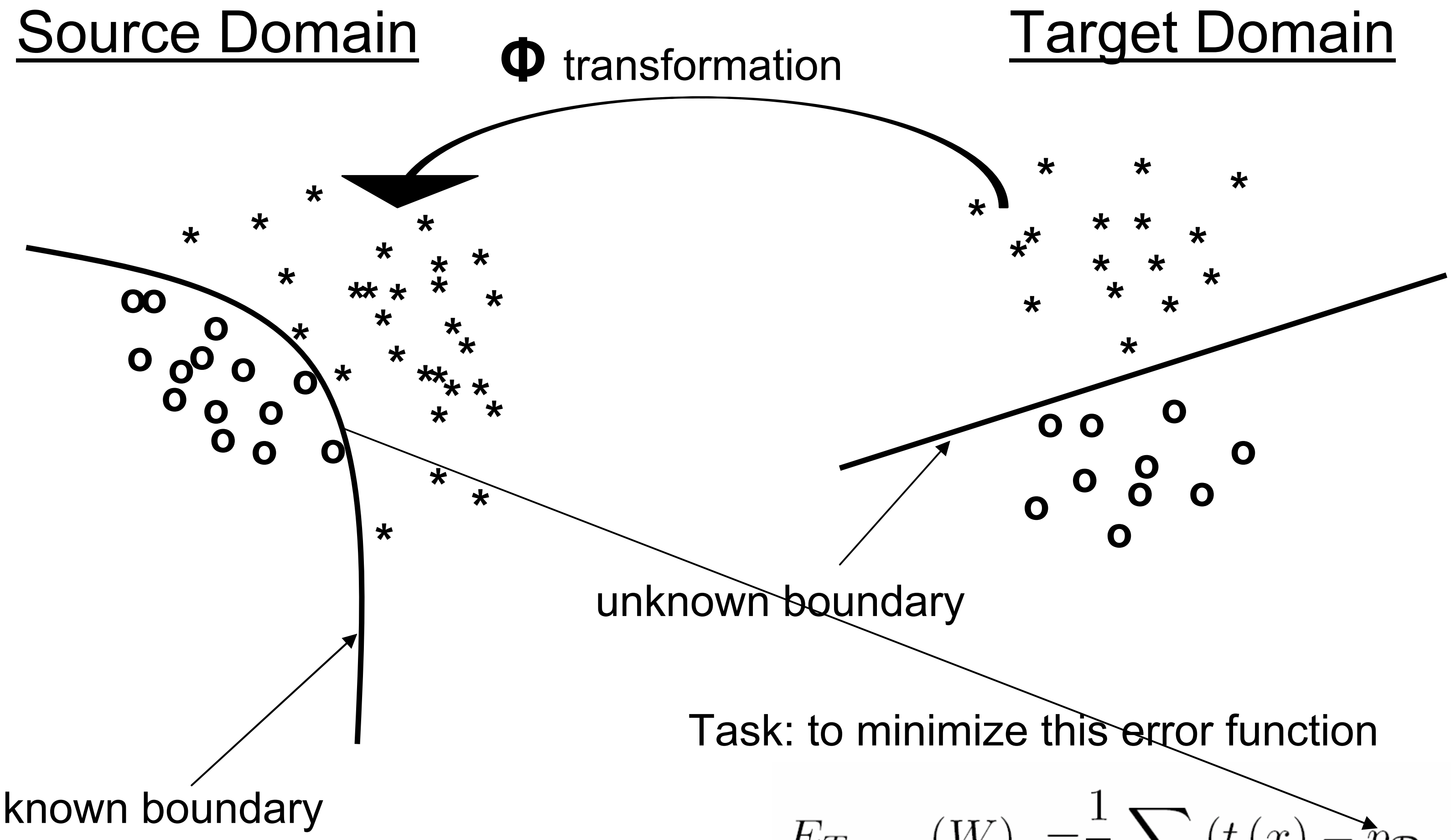
unknown boundary

known boundary

Task: to minimize this error function

$$E_{T,p_{\mathcal{D}_S}}(W) = \frac{1}{2} \sum_{x \in T} \left(t(x) - p_{\mathcal{D}_S}(Wx) \right)^2$$

Transformation Based Domain Adaptation



Source Domain

Target Domain

Φ transformation

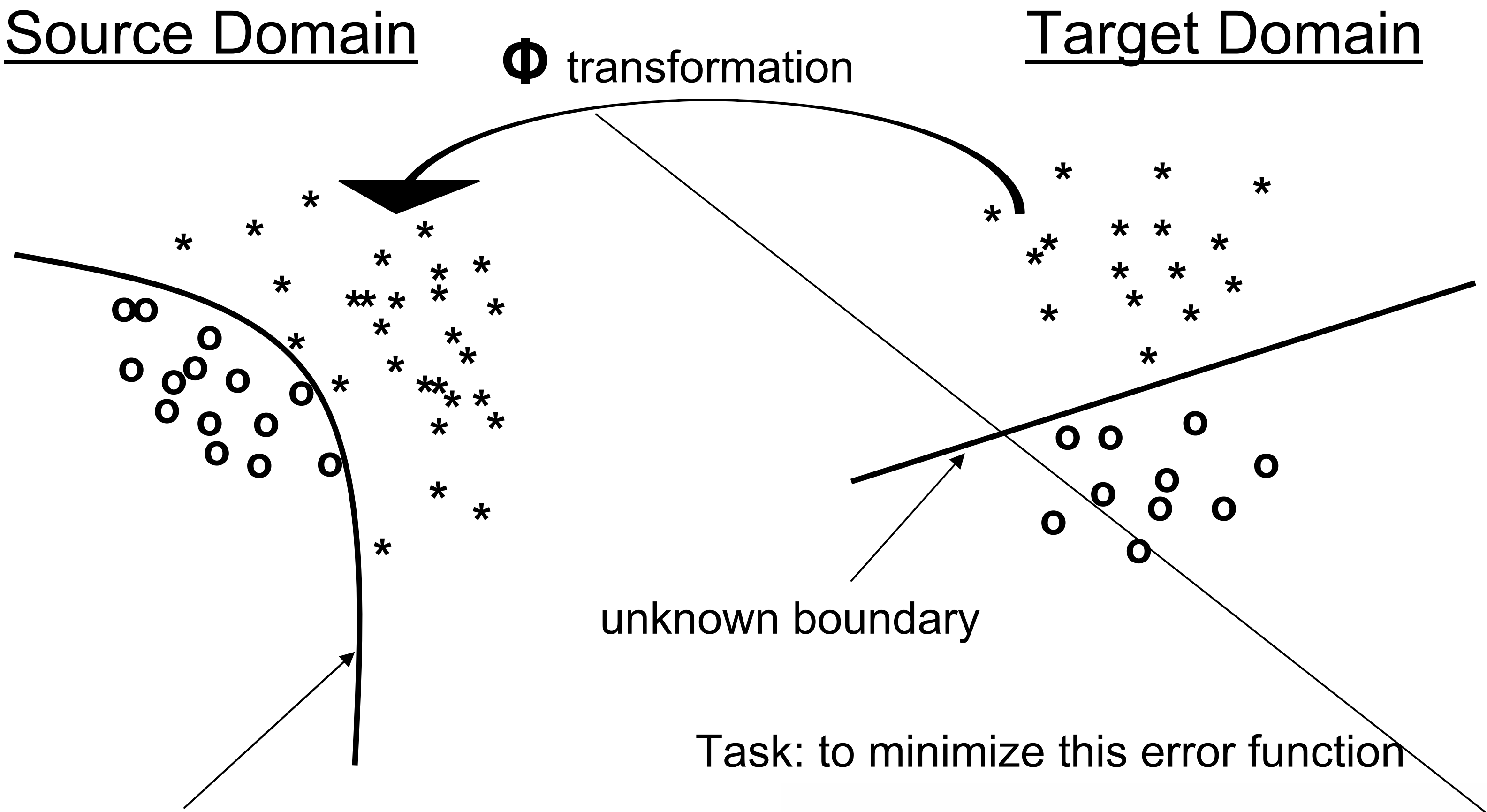
unknown boundary

known boundary

Task: to minimize this error function

$$E_{T,p_{\mathcal{D}_S}}(W) = \frac{1}{2}\sum_{x\in T}\left(t(x) - p_{\mathcal{D}_S}(Wx)\right)^2$$

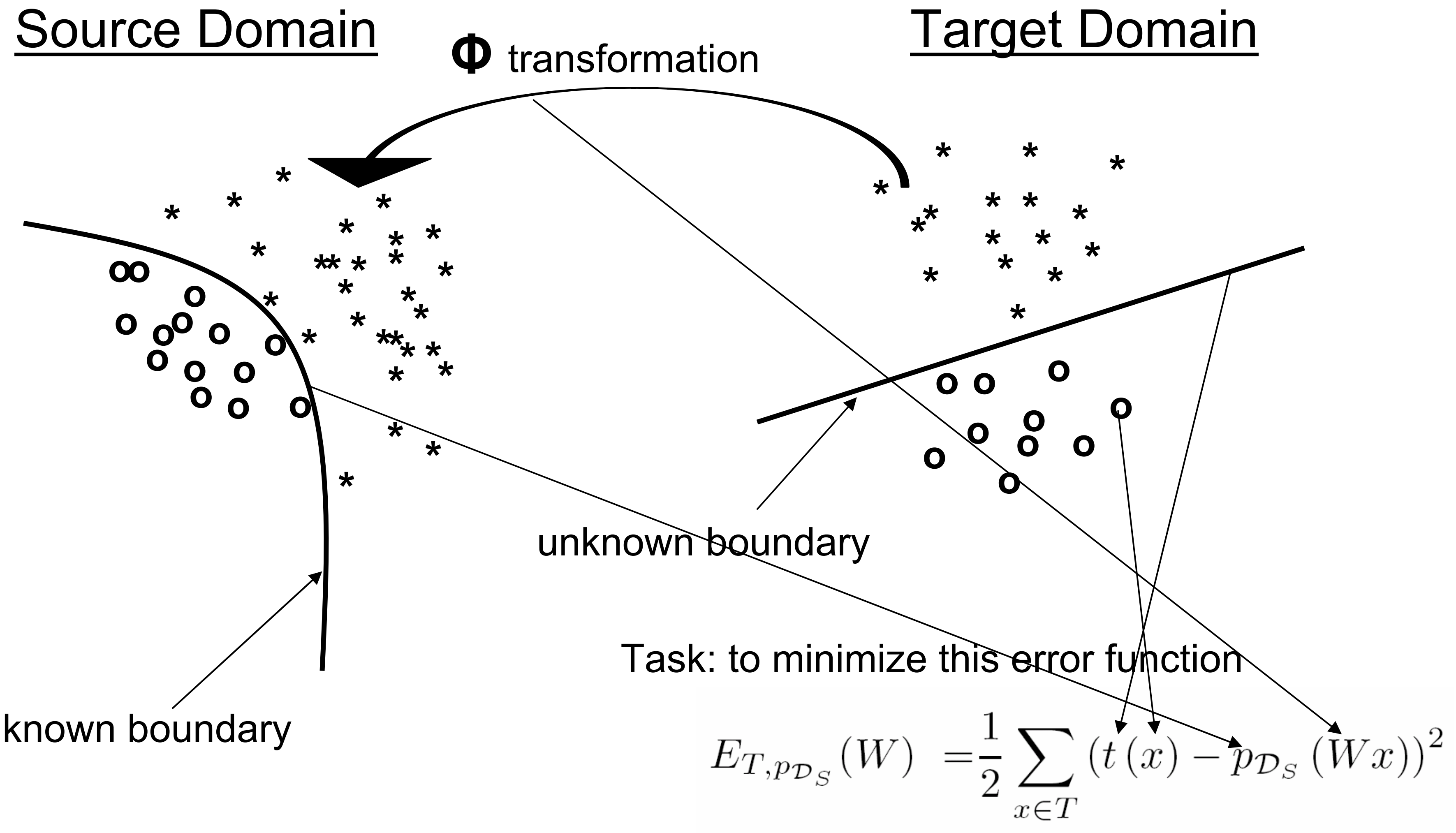
Transformation Based Domain Adaptation



Task: to minimize this error function

$$E_{T,p_{\mathcal{D}_S}}(W) \ =\frac{1}{2}\sum_{x\in T}\left(t\left(x\right)-p_{\mathcal{D}_S}\left(Wx\right)\right)^2$$

Transformation Based Domain Adaptation



Source Domain

Target Domain

Φ transformation

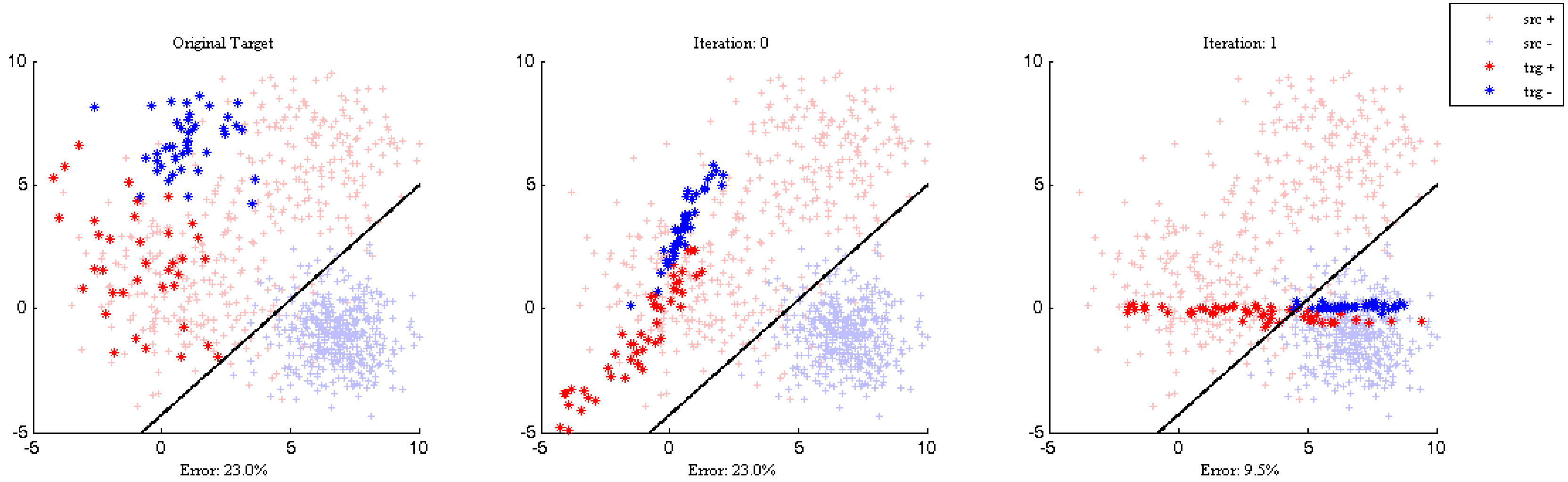
unknown boundary

known boundary

Task: to minimize this error function

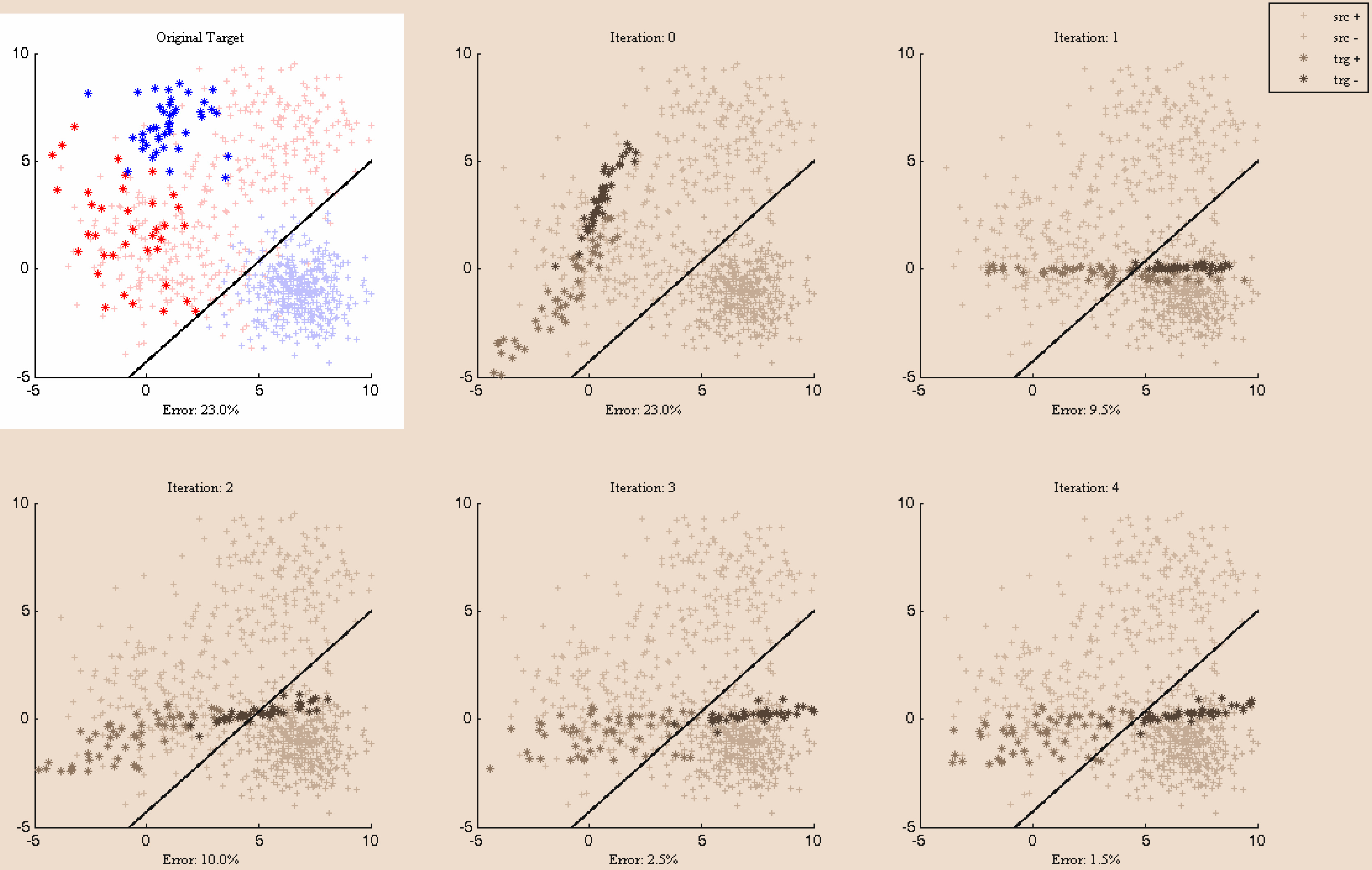
$$E_{T,p_{\mathcal{D}_S}}(W) \; = \frac{1}{2}\sum_{x\in T}\left(t\left(x\right) - p_{\mathcal{D}_S}\left(Wx\right)\right)^2$$

Synthetic Database



- 2D points and 2 classes
- Source train/test 800/200 samples
- Target distribution is the „rotated” source distribution (by 90°)

Results on Synthetic Dataset



Original Target

Error: 23.0%

Iteration: 0

Error: 23.0%

Iteration: 1

Error: 9.5%

Iteration: 2

Error: 10.0%

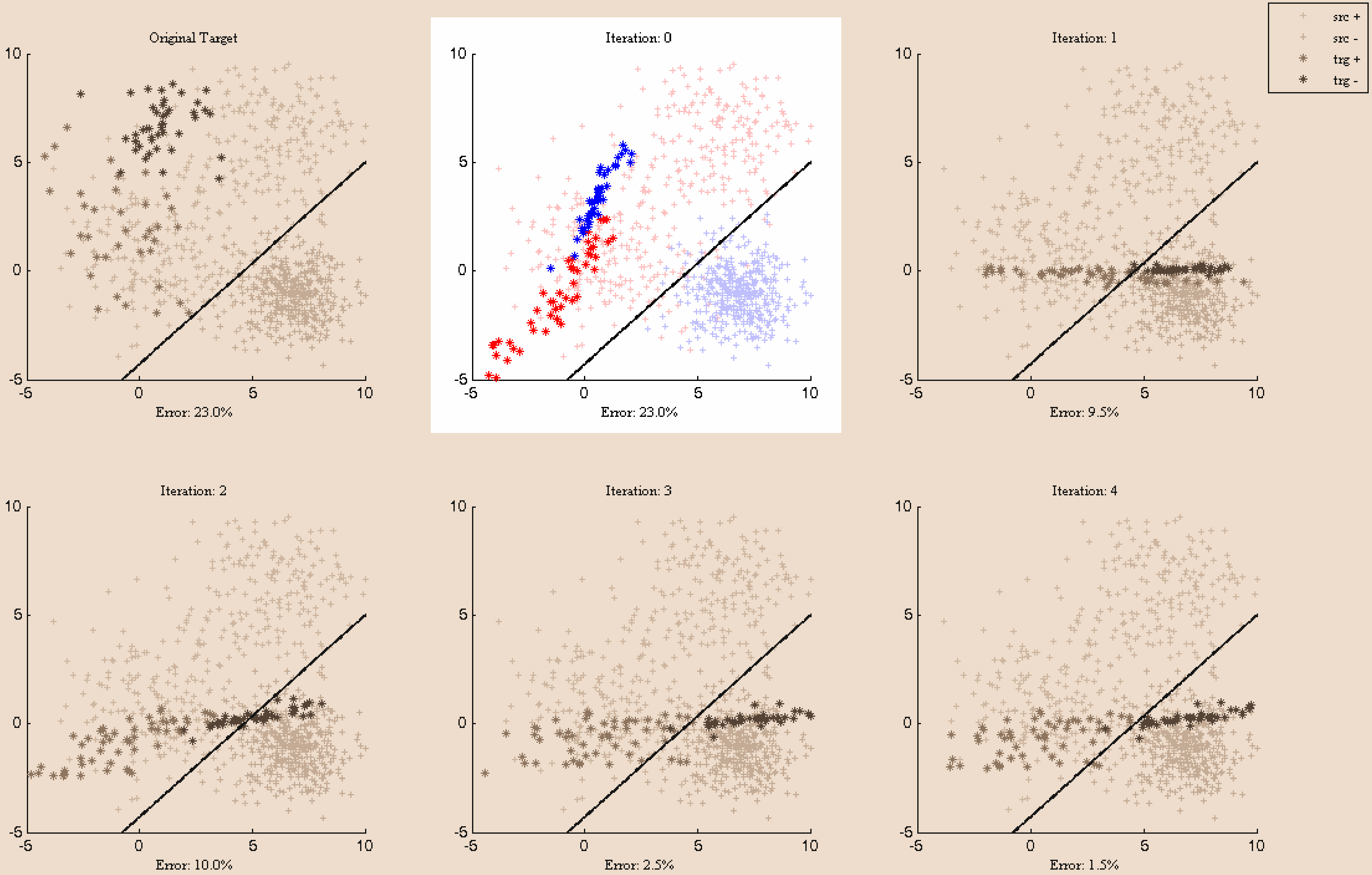
Iteration: 3

Error: 2.5%

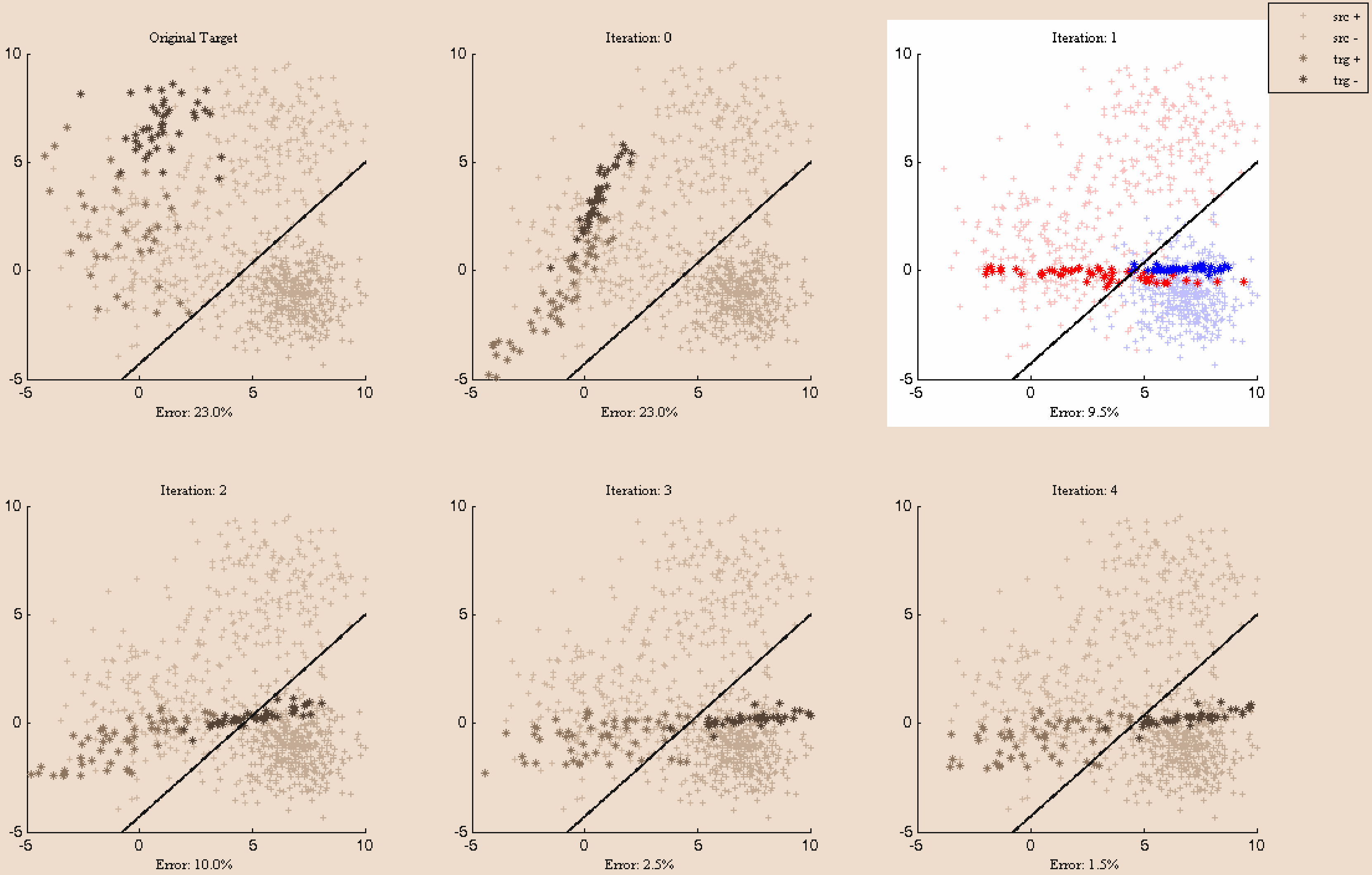
Iteration: 4

Error: 1.5%

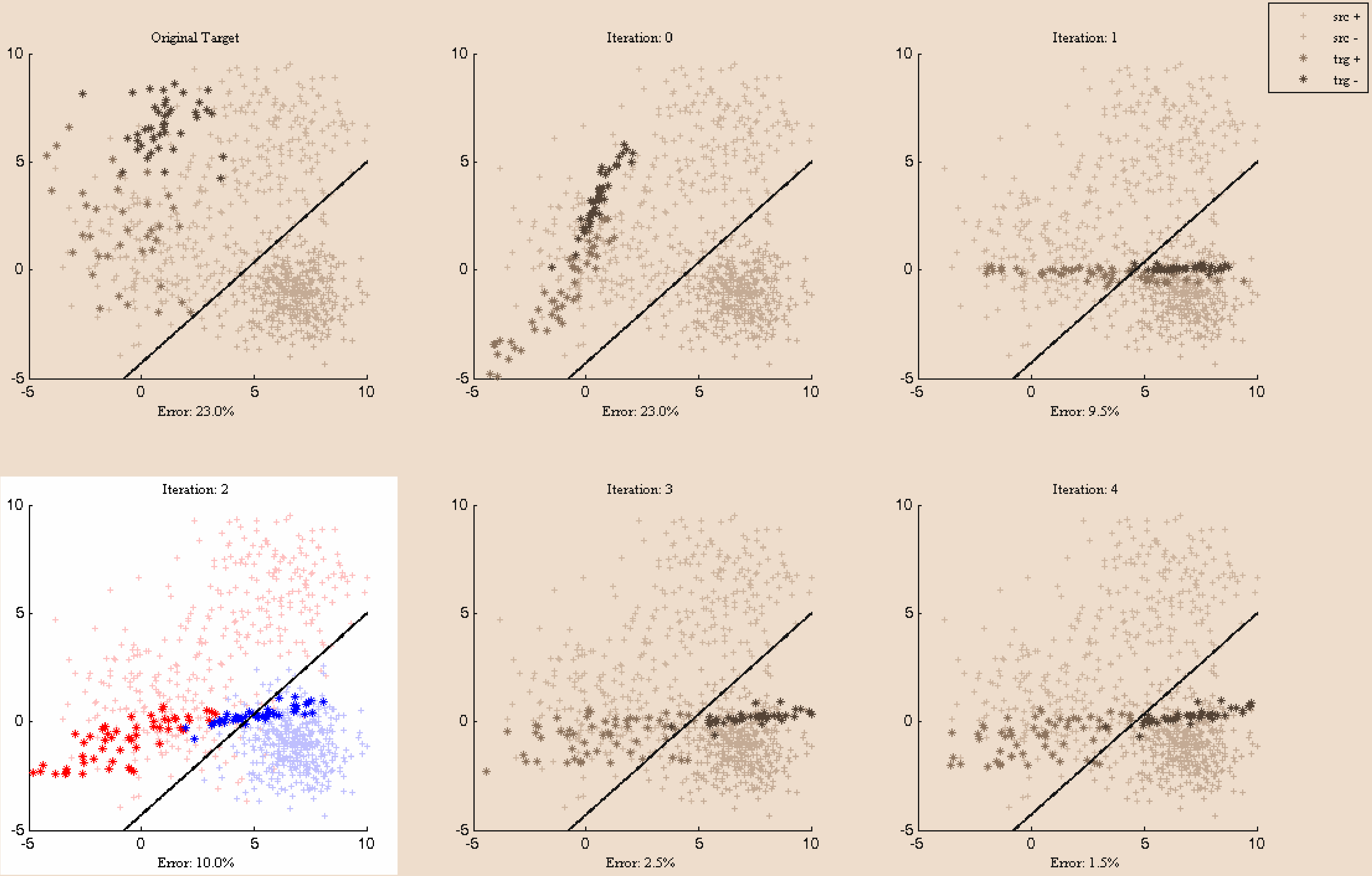
Results on Synthetic Dataset



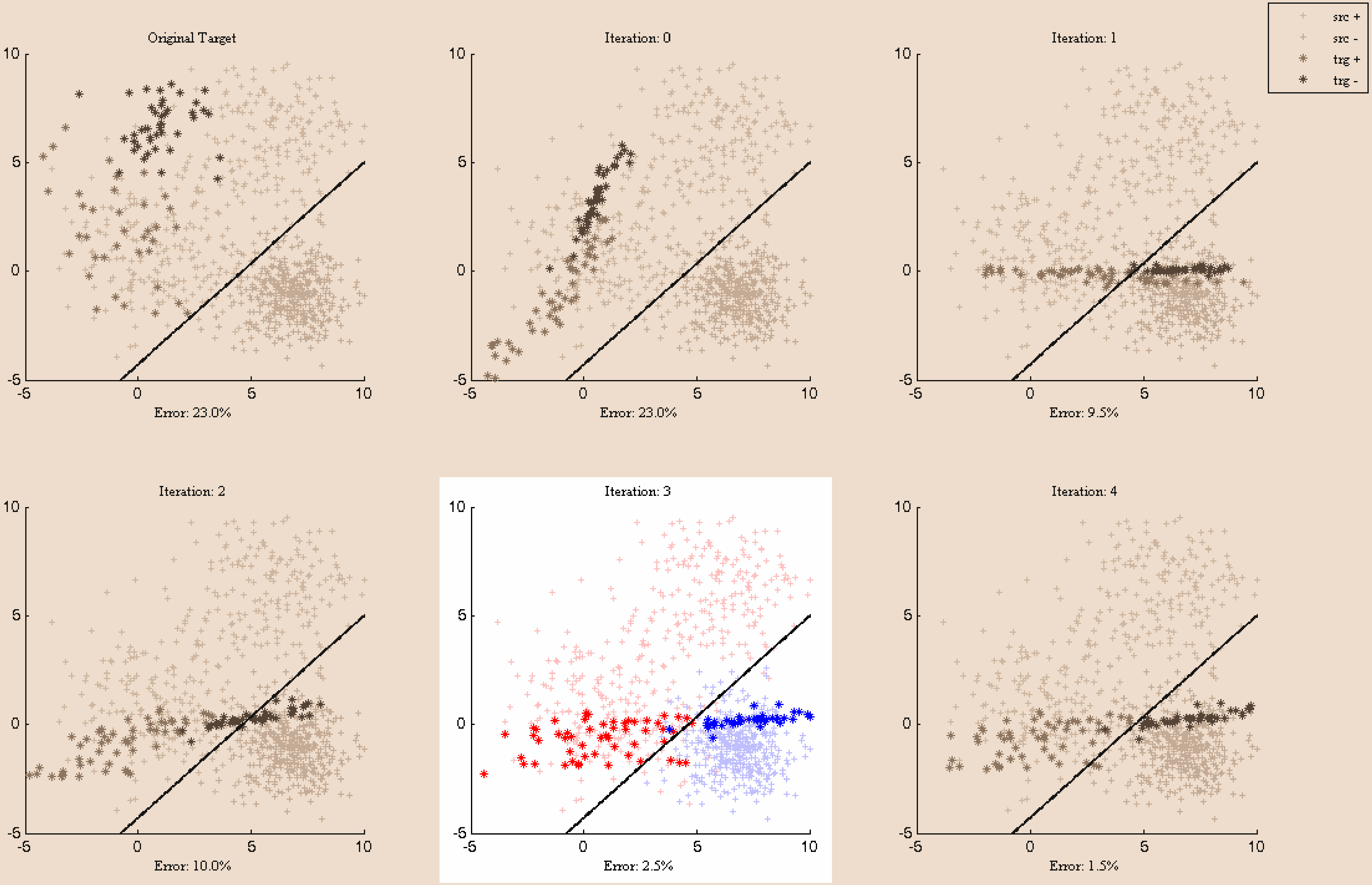
Results on Synthetic Dataset



Results on Synthetic Dataset



Results on Synthetic Dataset



Original Target

Error: 23.0%

Iteration: 0

Error: 23.0%

Iteration: 1

Error: 9.5%

Iteration: 2

Error: 10.0%

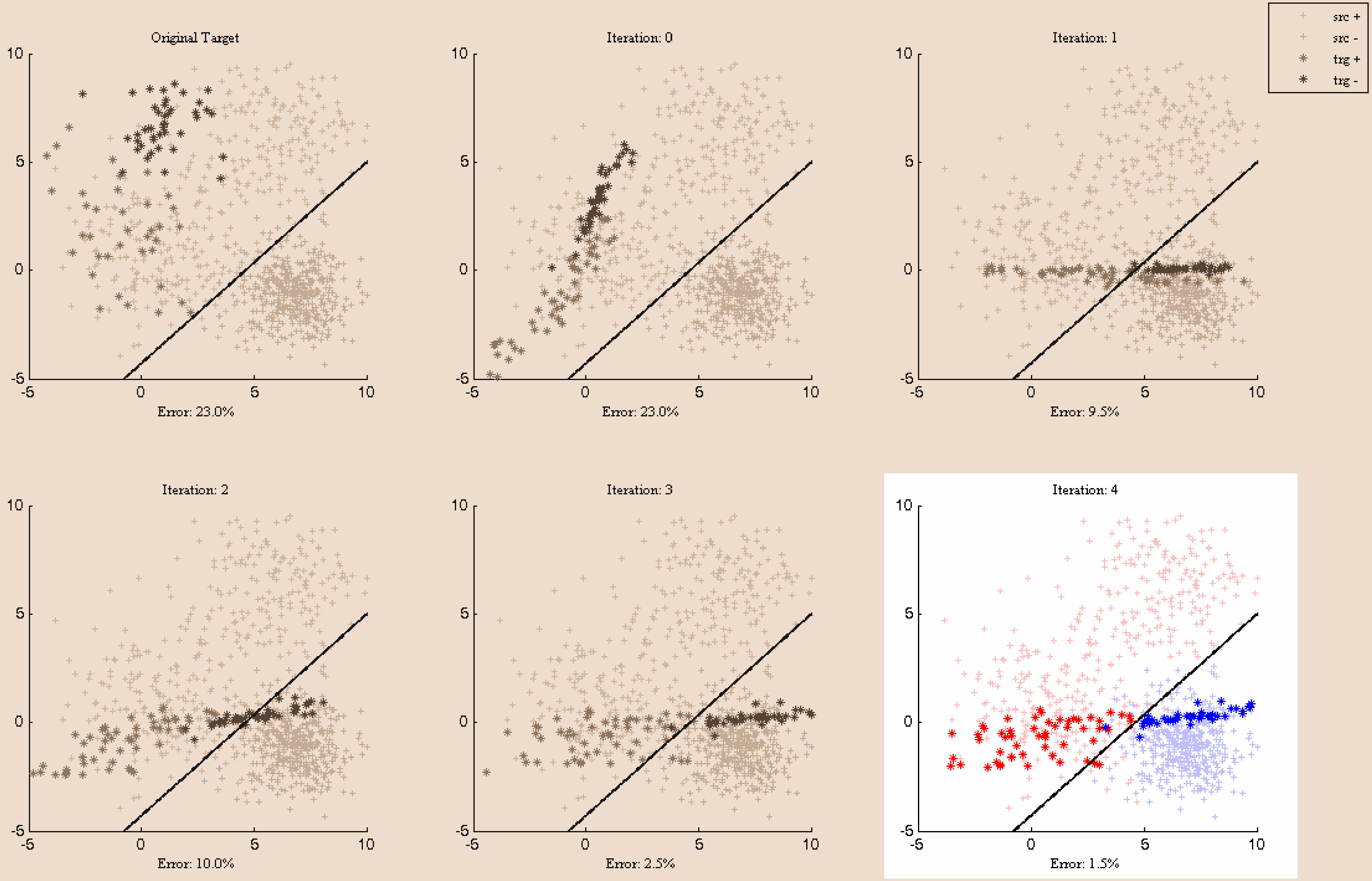
Iteration: 3

Error: 2.5%

Iteration: 4

Error: 1.5%

Results on Synthetic Dataset



Original Target

Error: 23.0%

Iteration: 0

Error: 23.0%

Iteration: 1

Error: 9.5%

Iteration: 2

Error: 10.0%

Iteration: 3

Error: 2.5%

Iteration: 4

Error: 1.5%

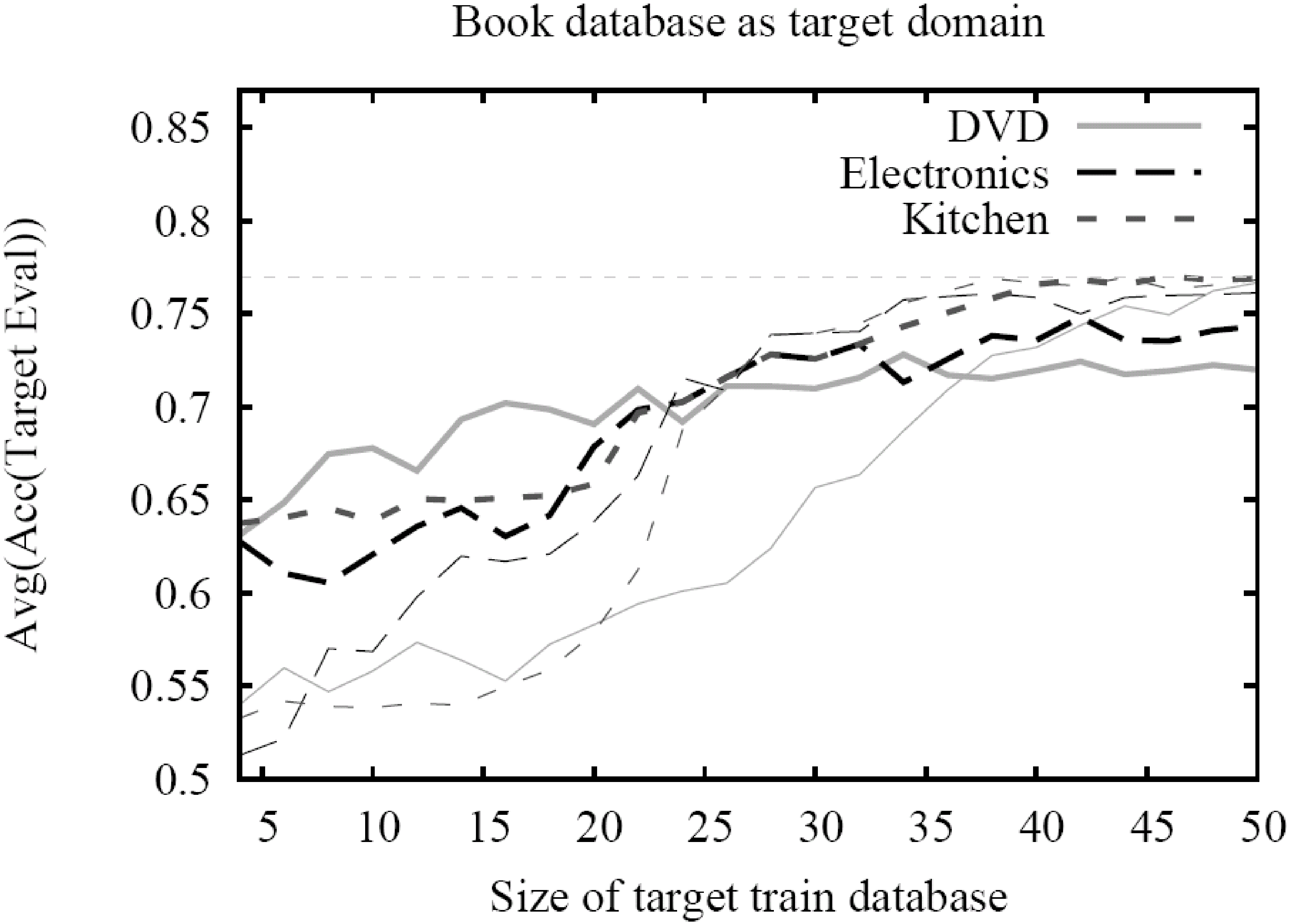
Multi-Domain Sentiment Database

- Reviews for 4 types of product/domain from the Amazon.com
  - Books
  - DVDs
  - Electronics
  - Kitchen appliances
- Balanced 1000-1000 positive and negative samples in ~ 5000 dimensions

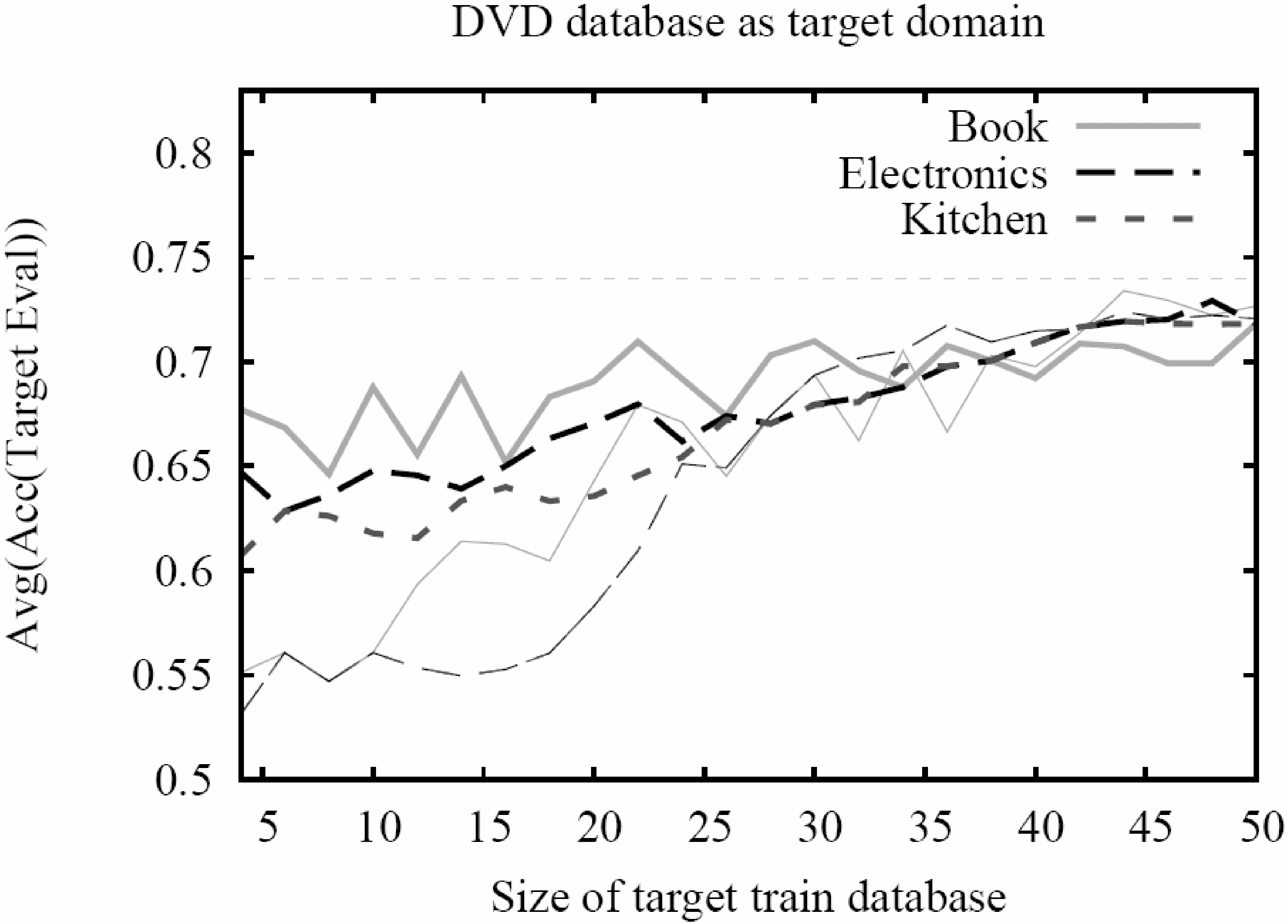
PreProcessing steps

- Train/test cut, 80/20% randomly
- Feature reduction by InfoGain score and PCA
  - we held the feature with > 0 score
  - compressed the dataset by PCA
- The results shows the average accuracy score of 10 runs

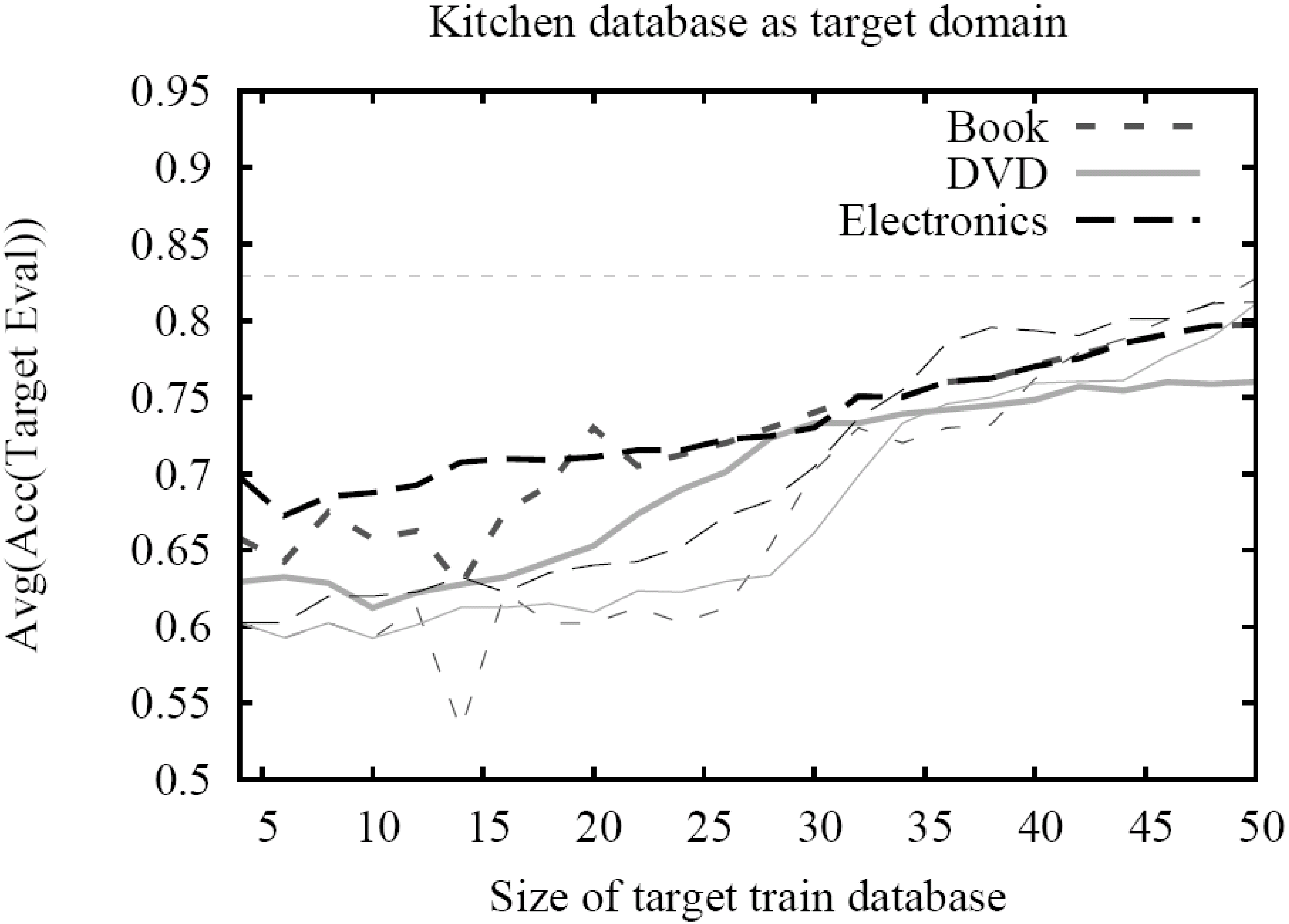
Results


Book database as target domain
DVD
Electronics
Kitchen
Avg(Acc(Target Eval))
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
5
10
15
20
25
30
35
40
45
50
Size of target train database

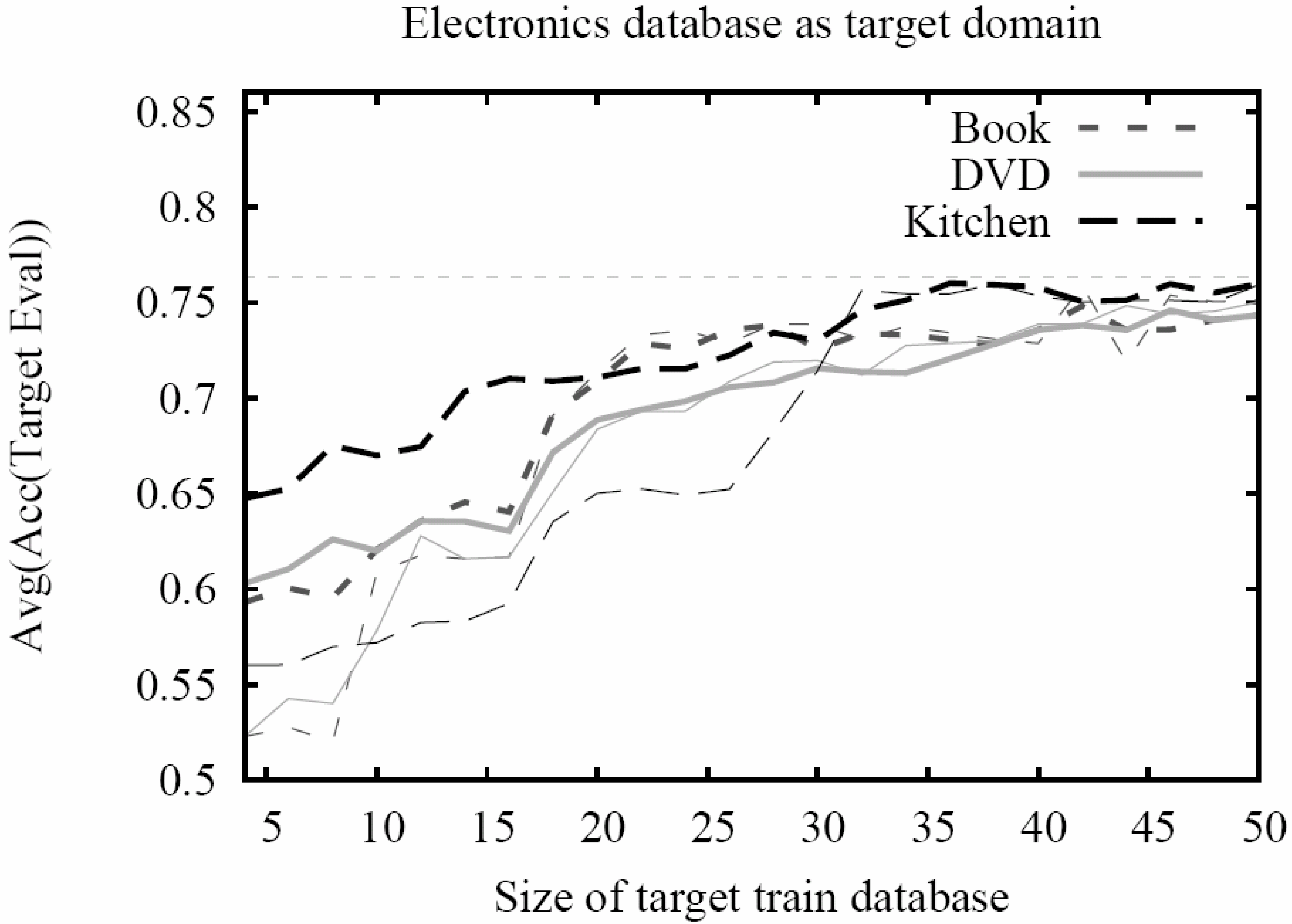
Results


DVD database as target domain
Book
Electronics
Kitchen
Avg(Acc(Target Eval))
0.8
0.75
0.7
0.65
0.6
0.55
0.5
5
10
15
20
25
30
35
40
45
50
Size of target train database

Results


Kitchen database as target domain
Book
DVD
Electronics
Avg(Acc(Target Eval))
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
5
10
15
20
25
30
35
40
45
50
Size of target train database

Results


Electronics database as target domain
Book
DVD
Kitchen
Avg(Acc(Target Eval))
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
5
10
15
20
25
30
35
40
45
50
Size of target train database