

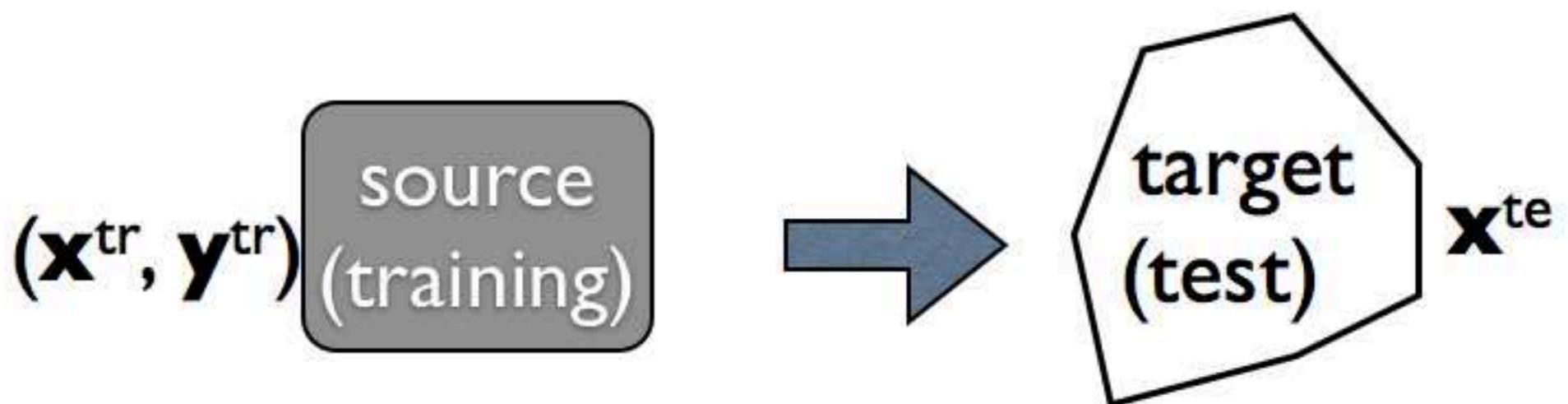
# Domain adaptation under **target and conditional shift**

Presenter: Kun Zhang

joint work with  
Bernhard Schölkopf, Krikamol Muandet, Zhikun Wang

Dept. Empirical Inference  
Max Planck Institute for Intelligent Systems  
Tübingen, Germany

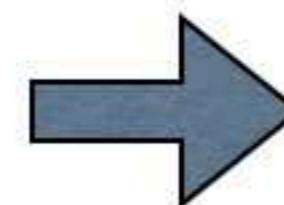
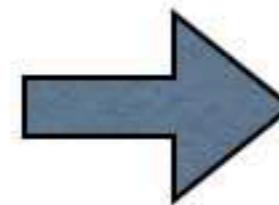
# Why domain adaptation



- Traditional supervised learning:

$$P_{XY}^{te} = P_{XY}^{tr}$$

- might not be the case in practice:



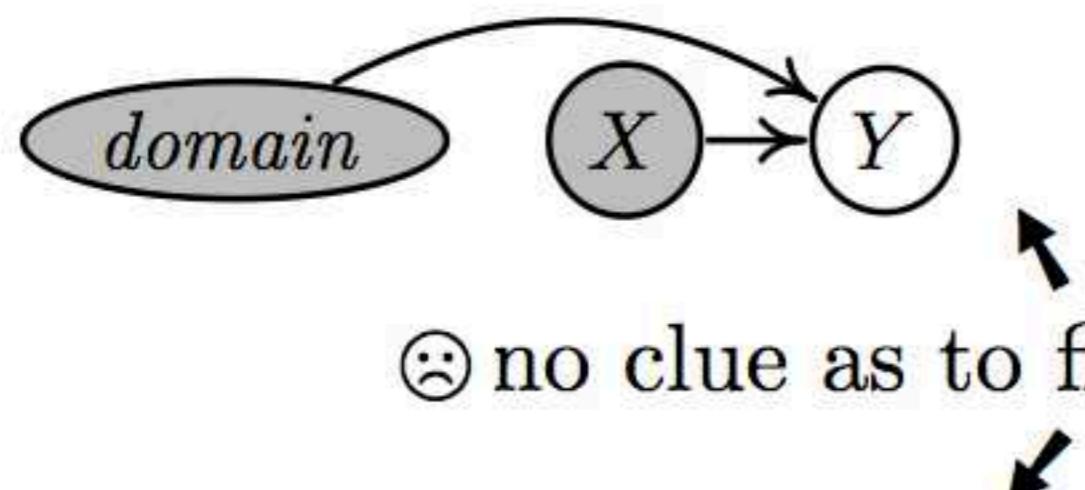
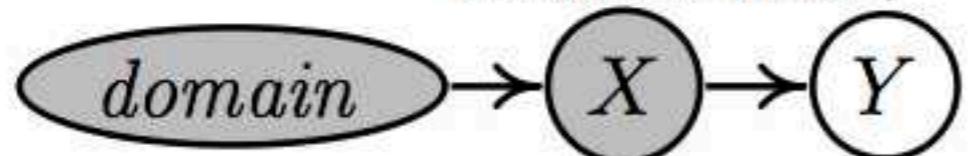
# Outline

- **Problem:** domain adaptation from  $(\mathbf{x}^{\text{tr}}, \mathbf{y}^{\text{tr}})$  to  $\mathbf{x}^{\text{te}}$
- **Possible causal models** for domain adaptation & solutions
  - Target shift
  - Location-scale conditional shift
  - Location-scale generalized target shift (target + conditional shift)

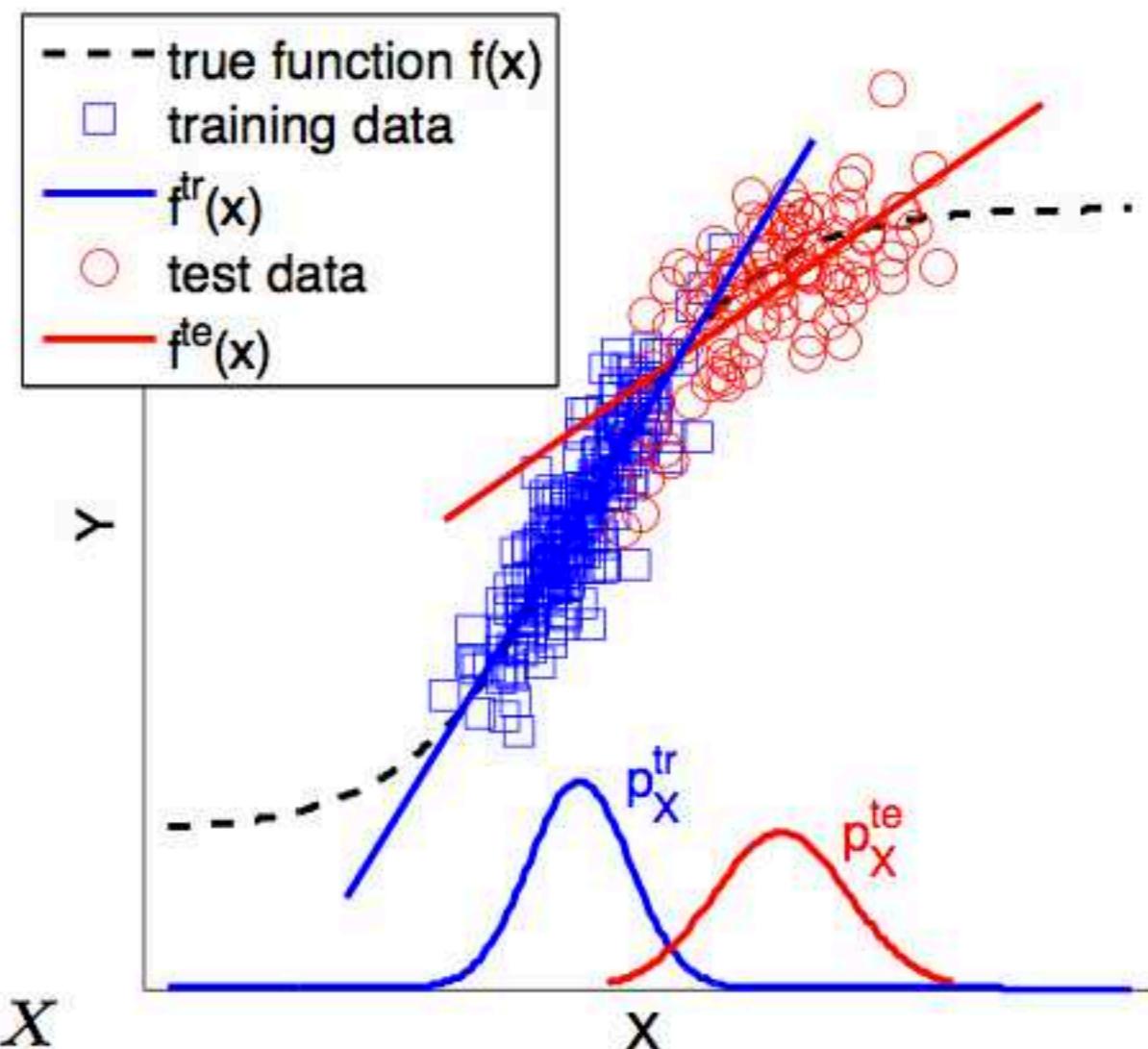
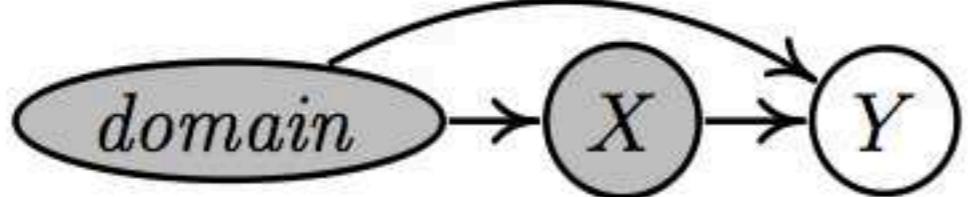
# Possible situations for domain adaptation: When $\mathbf{X} \rightarrow \mathbf{Y}$

## covariate shift

(Shimodaira00; Sugiyama et al.08; Huang et al.07,  
Gretton et al.08...)



:( no clue as to find  $P_{Y|X}^{te}$



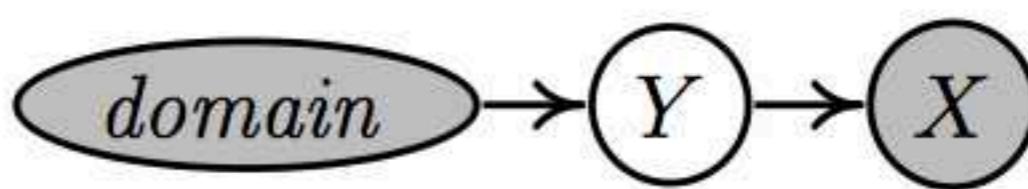
# Possible situations for domain adaptation: When $Y \rightarrow X$

- $Y$  is usually the cause of  $X$  (especially for classification)

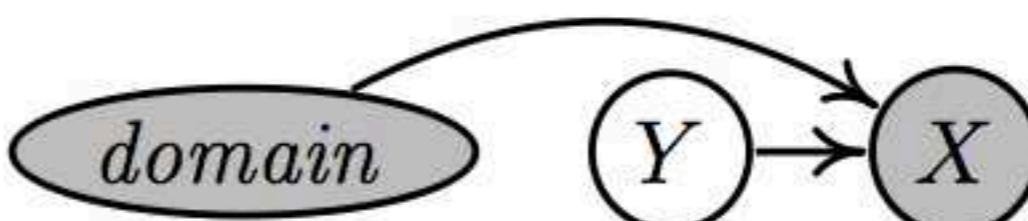
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9



- Target shift (TarS)

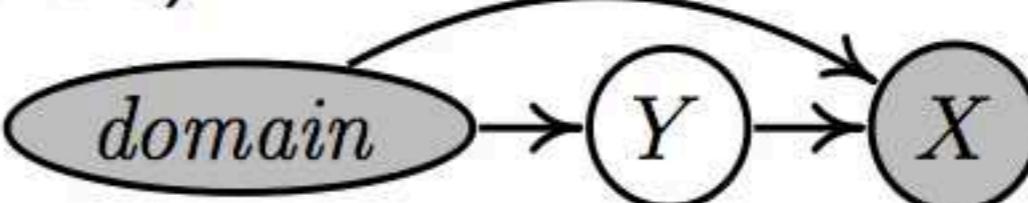


- Conditional shift (ConS)

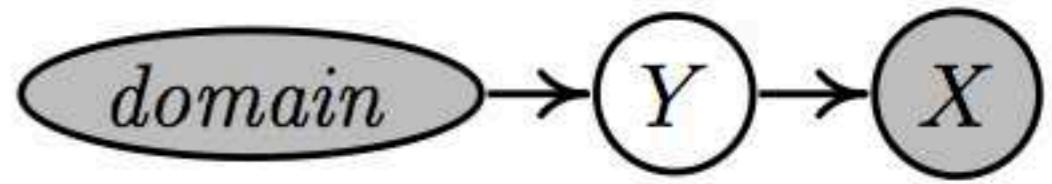


$P_{X}^{te}$   
helps  
find  
 $P_{Y|X}^{te}$

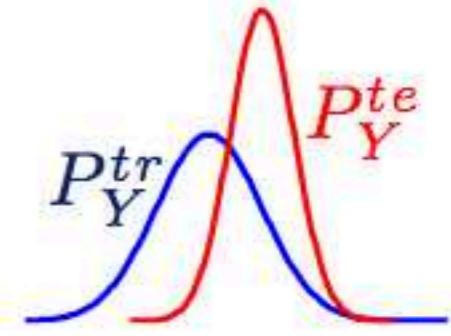
- Generalized target shift (GeTarS)



# Target shift



- $P_Y^{te} \neq P_Y^{tr}$ , but  $P_{X|Y}^{te} = P_{X|Y}^{tr}$ , and furthermore
  - **richness:** the support of  $P_Y^{tr}$  is richer
  - **invertibility:** only one  $P_Y \xrightarrow[P_{X|Y}^{tr}]{} P_X^{te}$
- find the learning machine on test domain by importance reweighting



$$\begin{aligned}
 R[P^{te}, \theta, l(x, y, \theta)] &= \mathbb{E}_{(X, Y) \sim P^{te}} [l(x, y, \theta)] = \int P_{XY}^{tr} \cdot \frac{P_{XY}^{te}}{P_{XY}^{tr}} \cdot l(x, y, \theta) dx dy \\
 &= \mathbb{E}_{(X, Y) \sim P^{tr}} \cdot \underbrace{\frac{P_Y^{te}}{P_Y^{tr}}}_{\triangleq \beta^*(y)} \cdot \underbrace{\frac{P_{X|Y}^{te}}{P_{X|Y}^{tr}}}_{\triangleq \gamma^*(x, y)} \cdot l(x, y, \theta) dx dy,
 \end{aligned}$$

$\triangleq \beta^*(y)$     $\triangleq \gamma^*(x, y) \equiv 1 !$

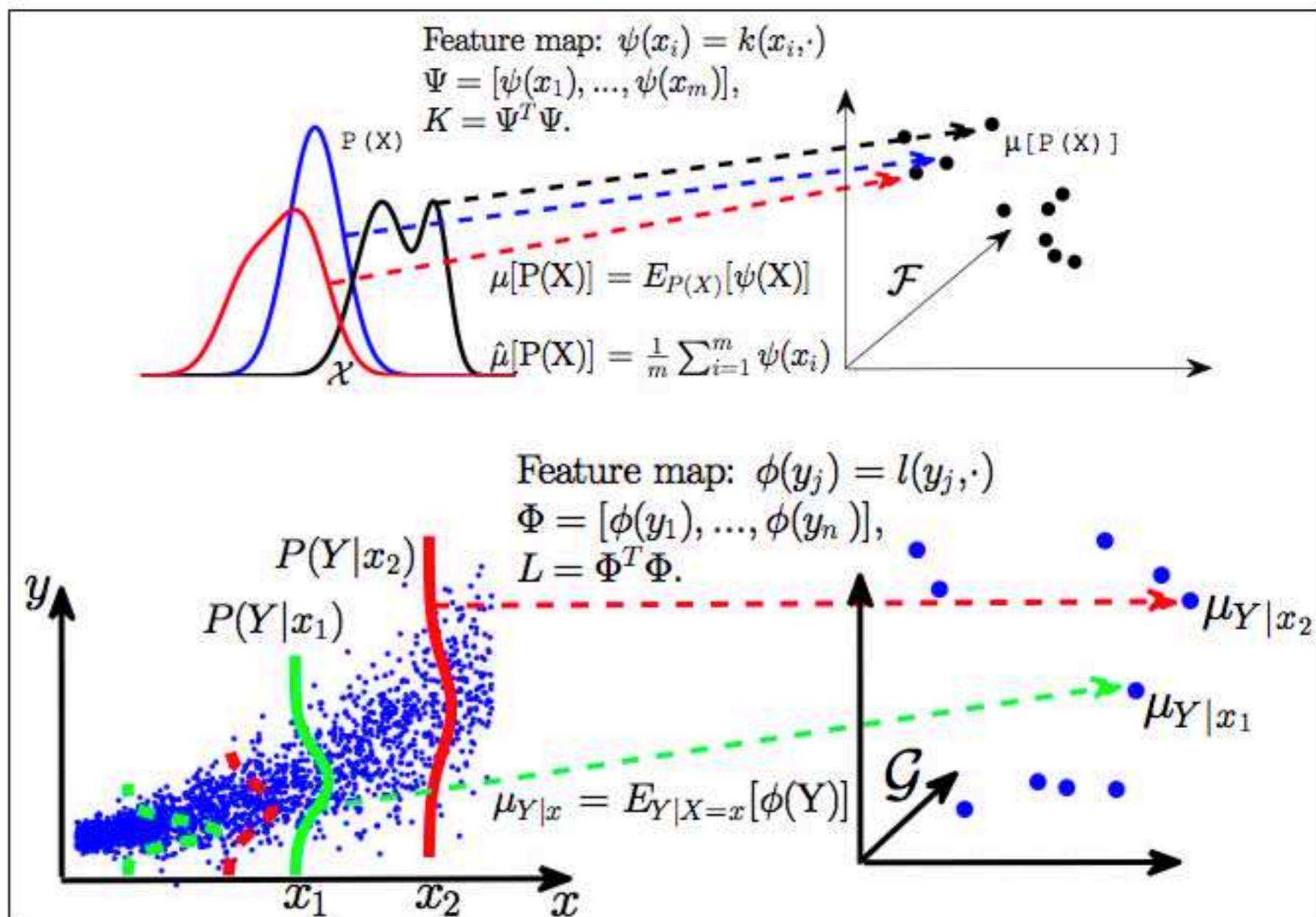
$$\widehat{R} = \frac{1}{m} \sum_{i=1}^m \beta^*(y_i^{tr}) \gamma^*(x_i^{tr}, y_i^{tr}) l(x_i^{tr}, y_i^{tr}; \theta) = \frac{1}{m} \sum_{i=1}^m \beta^*(y_i^{tr}) \cdot l(x_i^{tr}, y_i^{tr}; \theta)$$

- ratio  $\beta^*(y)$  can be estimated by  $\min. \mathcal{D}\left(P_X^{te}, \int P_Y^{tr} \beta(y) P_{X|Y}^{tr} dy\right)$ : difficult !

# Kernel mean embedding

(Smola et al. 07; Gretton et al. 07; Song et al. 09)

- $P_X$  has a unique embedding  $\mu[P_X] = \mathbb{E}_{X \sim P_X}[\psi(X)]$  for characteristic kernels.
- Conditional embedding of  $P_{Y|X}$  is an operator from  $\mathcal{F}_X$  to  $\mathcal{G}_Y$ :  $\mathcal{U}_{Y|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ ;  $\mathcal{C}_{YX}$  and  $\mathcal{C}_{XX}$  are *uncentered* cross- and auto-covariance operators.
- $\mu[P_X] = \mathcal{U}_{X|Y} \cdot \mu[P_Y]$ .
- $\hat{\mathcal{U}}_{X|Y} = \Psi(L + \lambda I)^{-1}\Phi^\top$ .



# Correcting TarS by reweighting target to match covariate with KMM

how to find  
 $\beta^*(y) = \frac{P_Y^{te}}{P_Y^{tr}}$ ?

$$P_Y^{new} = \beta(y) P_Y^{tr}$$
$$\downarrow \cdots \cdots \cdots \downarrow$$
$$P_{X|Y}^{tr}$$
$$\downarrow \cdots \cdots \cdots \downarrow$$
$$P_X^{new} \approx P_X^{te}$$

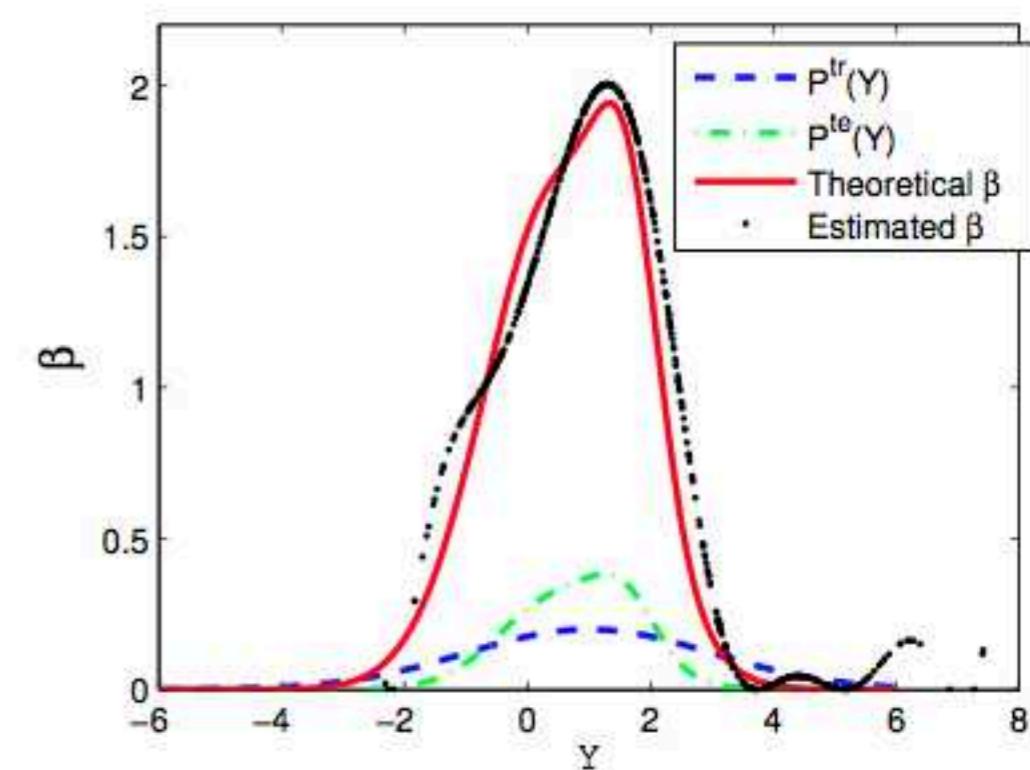
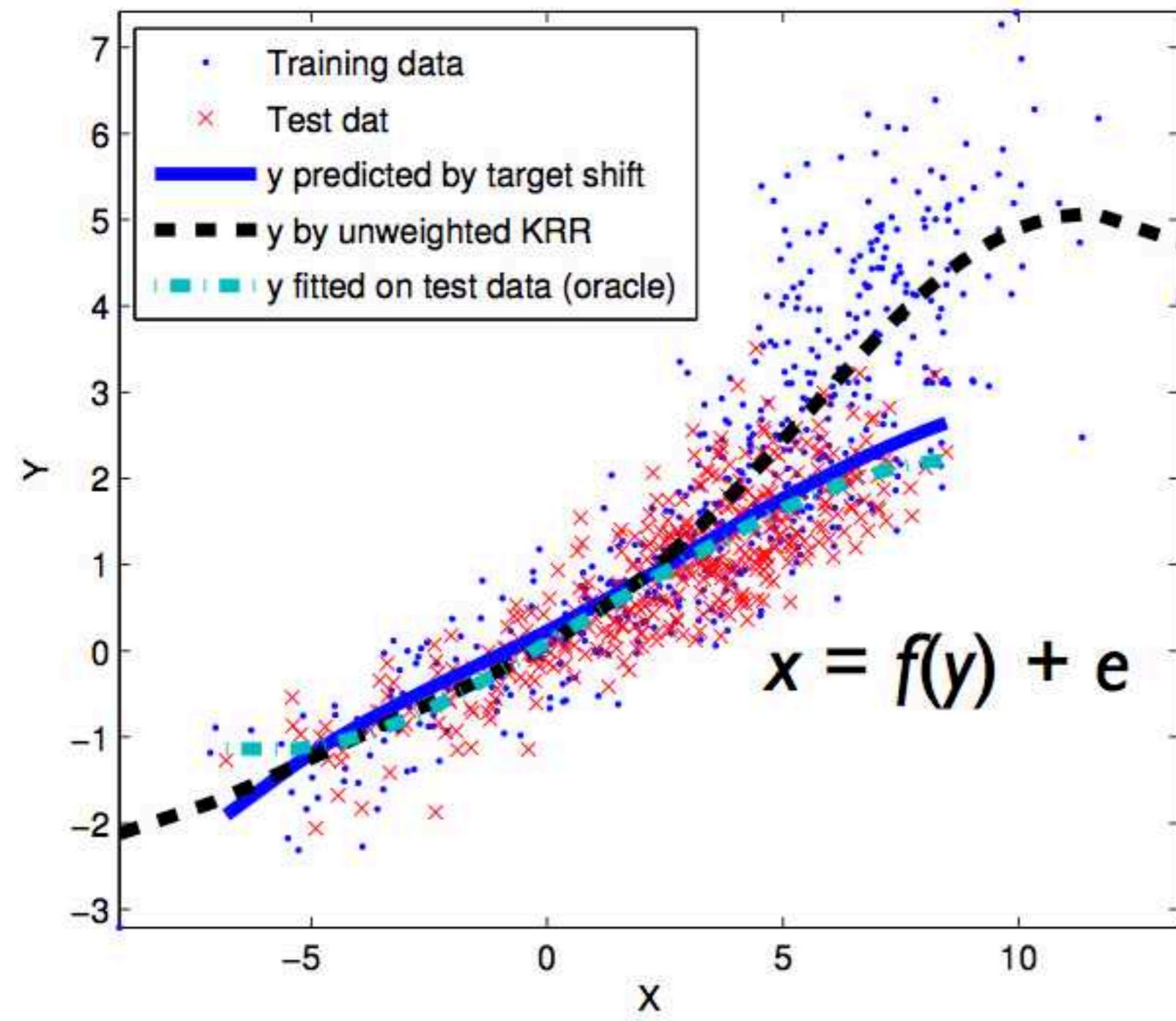
$\beta(y)$  can be estimated by matching  $P_X^{new}$  with  $P_X^{te}$  :-)

- i.e., minimizing

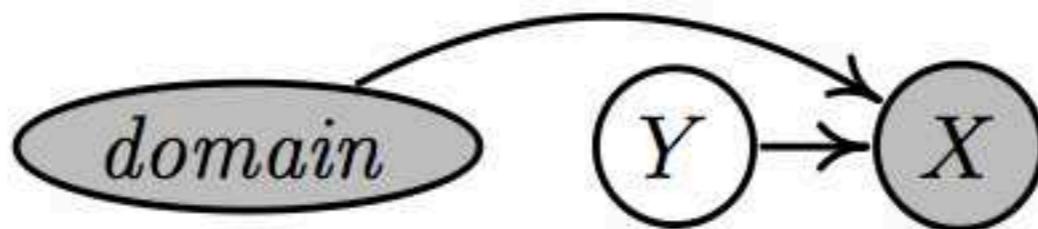
$$\begin{aligned} \|\hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}]\|^2 &= \left\| \hat{u}_{X|Y} \cdot \frac{1}{m} \sum_{i=1}^m \beta_i \phi(y_i^{tr}) - \frac{1}{n} \sum_{i=1}^n \psi(x_i^{te}) \right\|^2 \\ &= \underbrace{\frac{1}{m^2} \beta^\top L (L + \lambda_m I)^{-1} K (L + \lambda_m I)^{-1} L \beta}_\triangleq A - \underbrace{\frac{2}{mn} \mathbf{1}_n^\top K^c (L + \lambda_m I)^{-1} L \beta}_\triangleq M + \text{const} \end{aligned}$$

- QP problem: unique solution to  $\beta$ !
- reparameterization such that  $\beta$  is a function of & smooth in  $y$ : still a QP problem

# Correction for TarS: An illustration

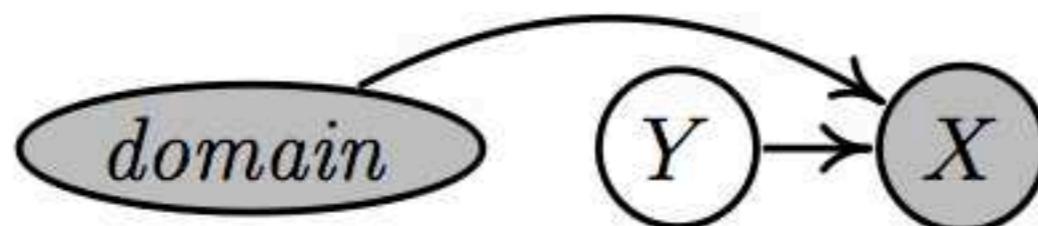


# Conditional shift



- If  $P_{X|Y}^{te} \neq P_{X|Y}^{tr}$ , possible to determine  $P_{Y|X}^{te}$ ?
- In general, not possible: marginal  $P_X^{te}$  do not contain enough information to determine  $P_{X|Y}^{te}$  (or  $P_{Y|X}^{te}$ )
- Change in  $P_{X|Y}$  must be constrained

# Location-scale conditional shift

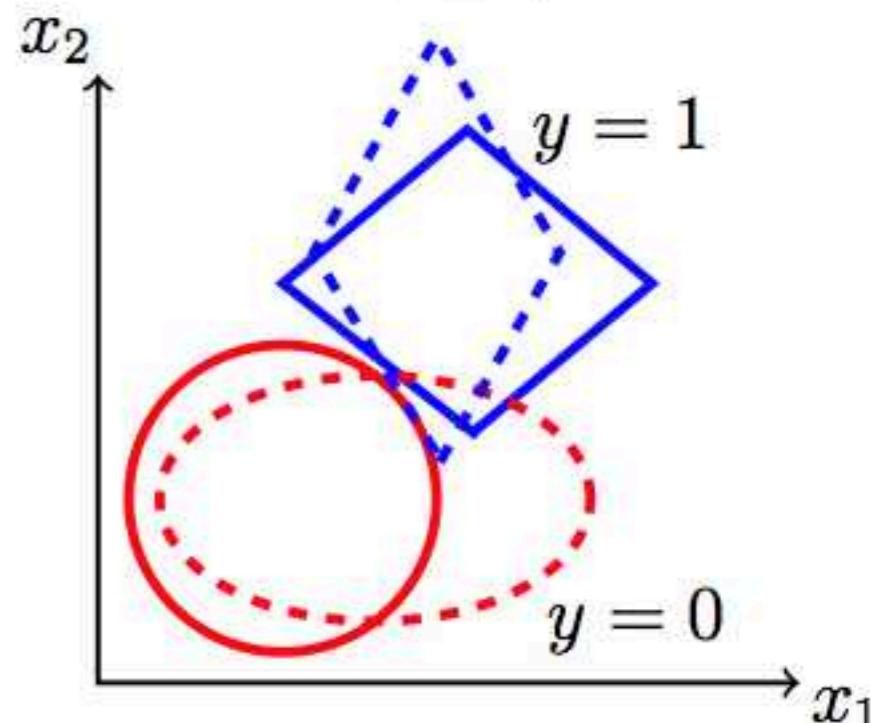


contours of  $P_{X|Y}$  (solid: training domain, dashed: test domain)

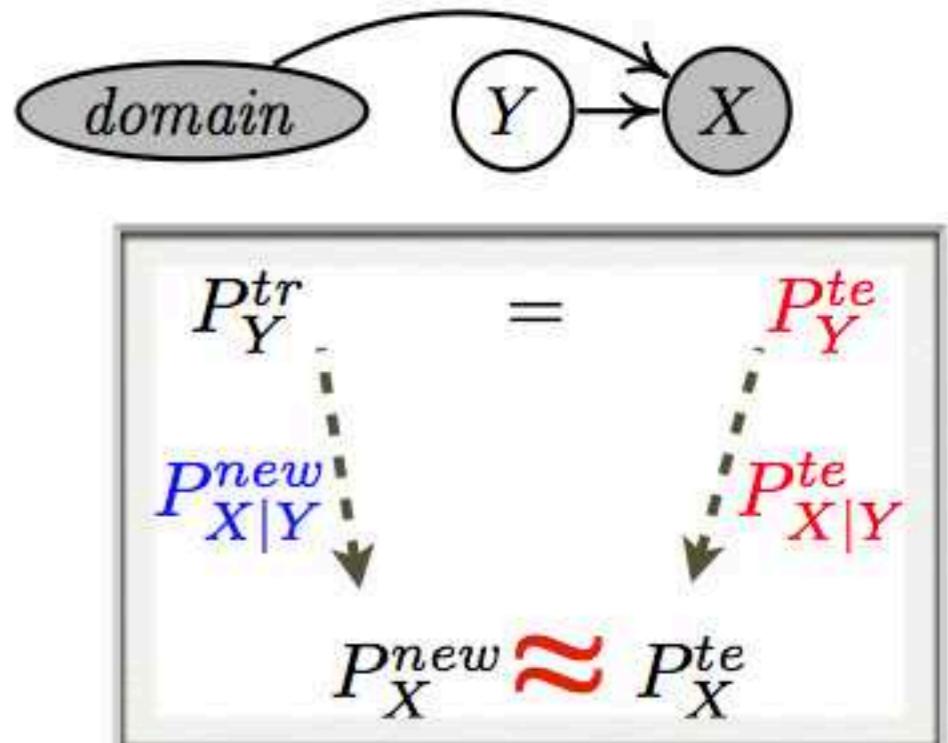
- For each  $y$ , *domain* could change the scale and mean of each feature
- does not change the dependence structure between features
- Assumption A<sup>ConS:</sup>

$\exists \mathbf{w}$  and  $\mathbf{b}$ , such that  $P_{X|Y}^{new} = P_{X|Y}^{te}$ ,  
where  $X^{new} = \mathbf{w}(Y^{tr}) \circ X^{tr} + \mathbf{b}(Y^{tr})$

- Can we find  $\mathbf{w}$  and  $\mathbf{b}$  ?



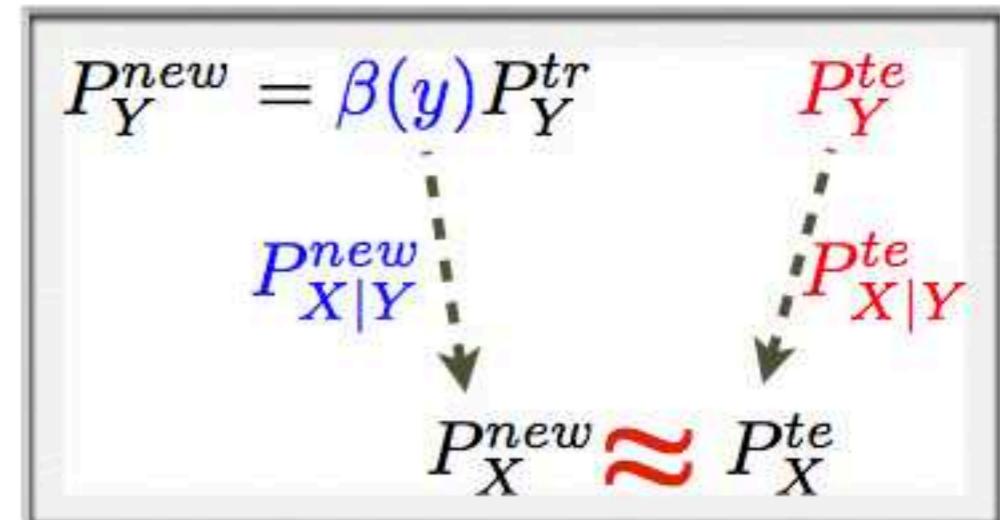
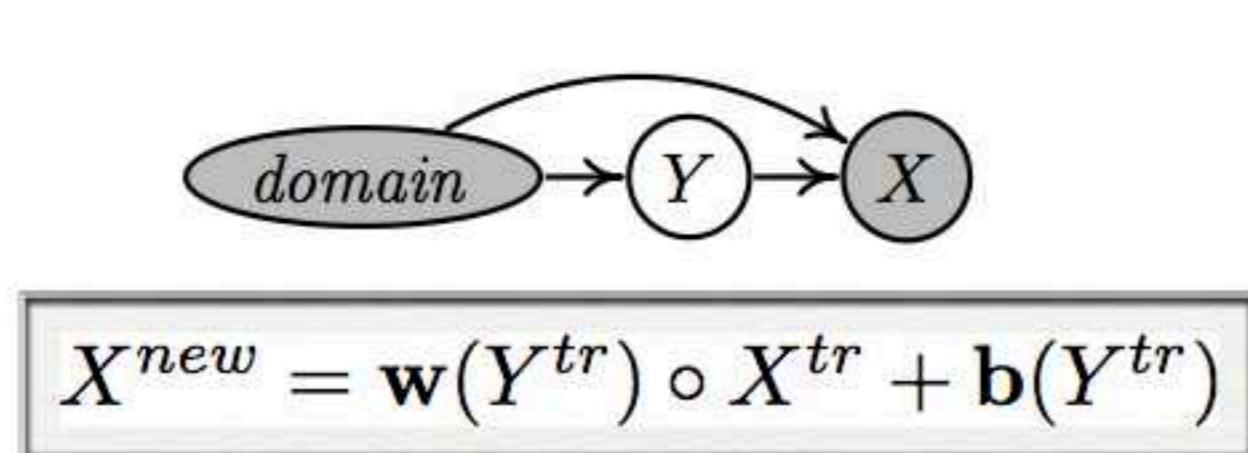
# Identifiability & solution



- $P_{X|Y}^{\text{new}}$  is theoretically identifiable under mild conditions (c.f. Theo 2)
  - Necessary condition:  $P_{X|Y}^{\text{tr}}(x|y_i)$ ,  $i = 1, \dots, C$ , are linearly independent after any LS transformation)
- Estimation: SCG for optimization  
by minimizing  $\left\| \hat{\mu}[P_X^{\text{new}}] - \hat{\mu}[P_X^{\text{te}}] \right\|^2 = \left\| \hat{\mathcal{U}}[P_{X|Y}^{\text{new}}] \hat{\mu}[P_Y^{\text{tr}}] - \hat{\mu}[P_Y^{\text{te}}] \right\|^2$ ,  
such that  $P_{X|Y}^{\text{new}} = P_{X|Y}^{\text{te}}$  with  $X^{\text{new}} = w(Y^{\text{tr}}) \cdot X^{\text{tr}} + b(Y^{\text{tr}})$
- regularization to prefer little change in the conditional

$$J^{\text{reg}} = \frac{\lambda_{LS}}{m} \cdot \|\mathbf{W} - \mathbf{1}_m \mathbf{1}_d^\top\|_F^2 + \frac{\lambda_{LS}}{m} \cdot \|\mathbf{B}\|_F^2$$

# LS generalized target shift



- LS-GeTarS: **target-shift** + **location-scale conditional shift**
- $P_Y^{\text{te}}$  and  $P_{X|Y}^{\text{te}}$  can be uniquely recovered under mild conditions (c.f.Theo 3)
- Solution:  $\min. \left\| \hat{\mu}[P_X^{\text{new}}] - \hat{\mu}[P_X^{\text{te}}] \right\|^2 = \left\| \hat{\mathcal{U}}[P_{X|Y}^{\text{new}}] \hat{\mu}[P_X^{\text{new}}] - \hat{\mu}[P_X^{\text{te}}] \right\|^2$  $= \frac{1}{m^2} \underline{\beta^\top \Omega \tilde{K} \Omega^\top \beta} - \frac{2}{mn} \underline{\mathbf{1}_n^\top \tilde{K}^c \beta}$ 
  - alternate between optimizing  $\beta$  (for TarS) and optimizing  $\mathbf{w}$  and  $\mathbf{b}$  (for ConS)

# To find the learning machine under LS- GeTarS

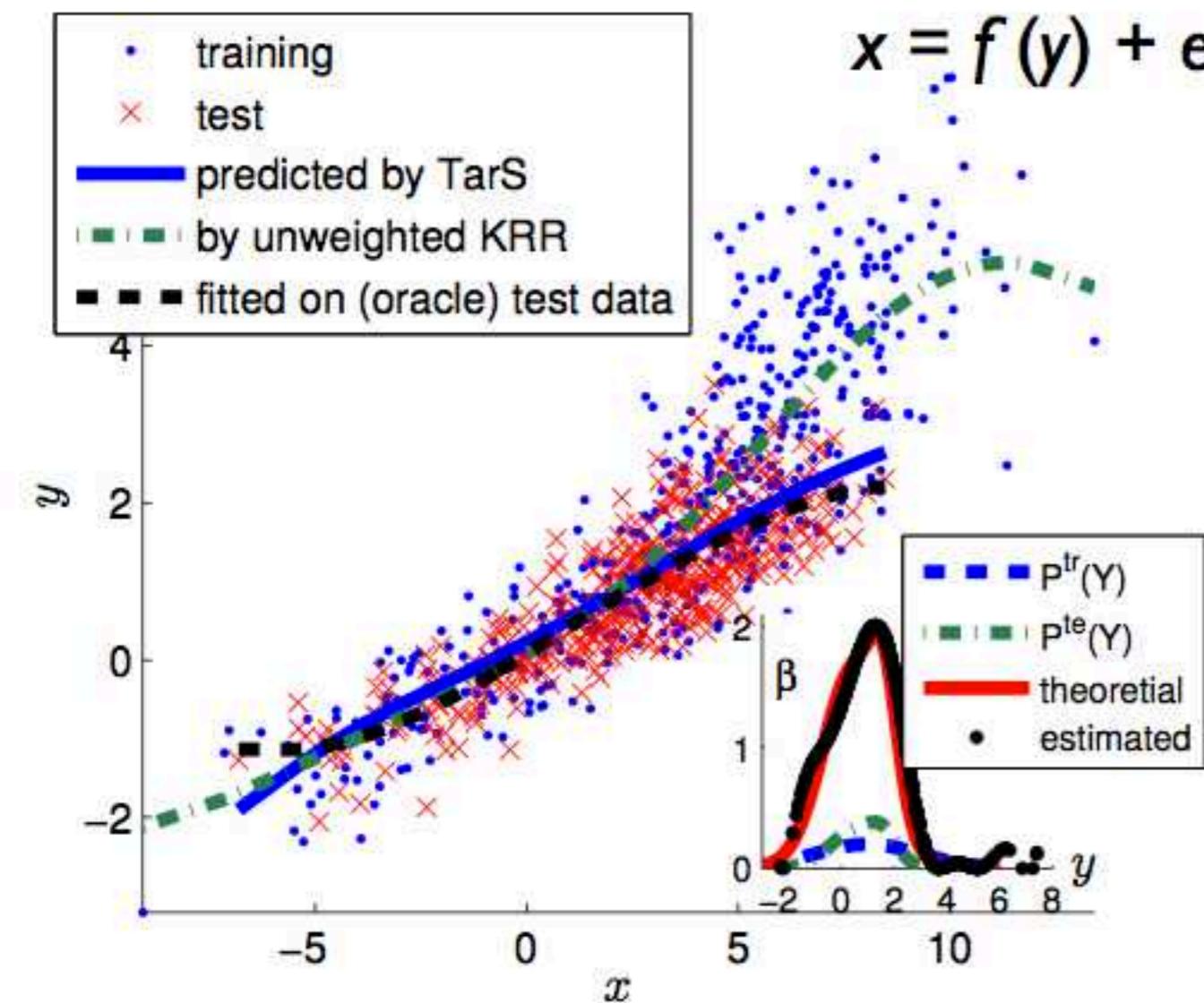
- With sample transformation + importance reweighting

$$X^{new} = \mathcal{T}(X^{tr}, Y^{tr}) \text{ satisfies } P_{X|Y}^{new} = P_{X|Y}^{te}$$

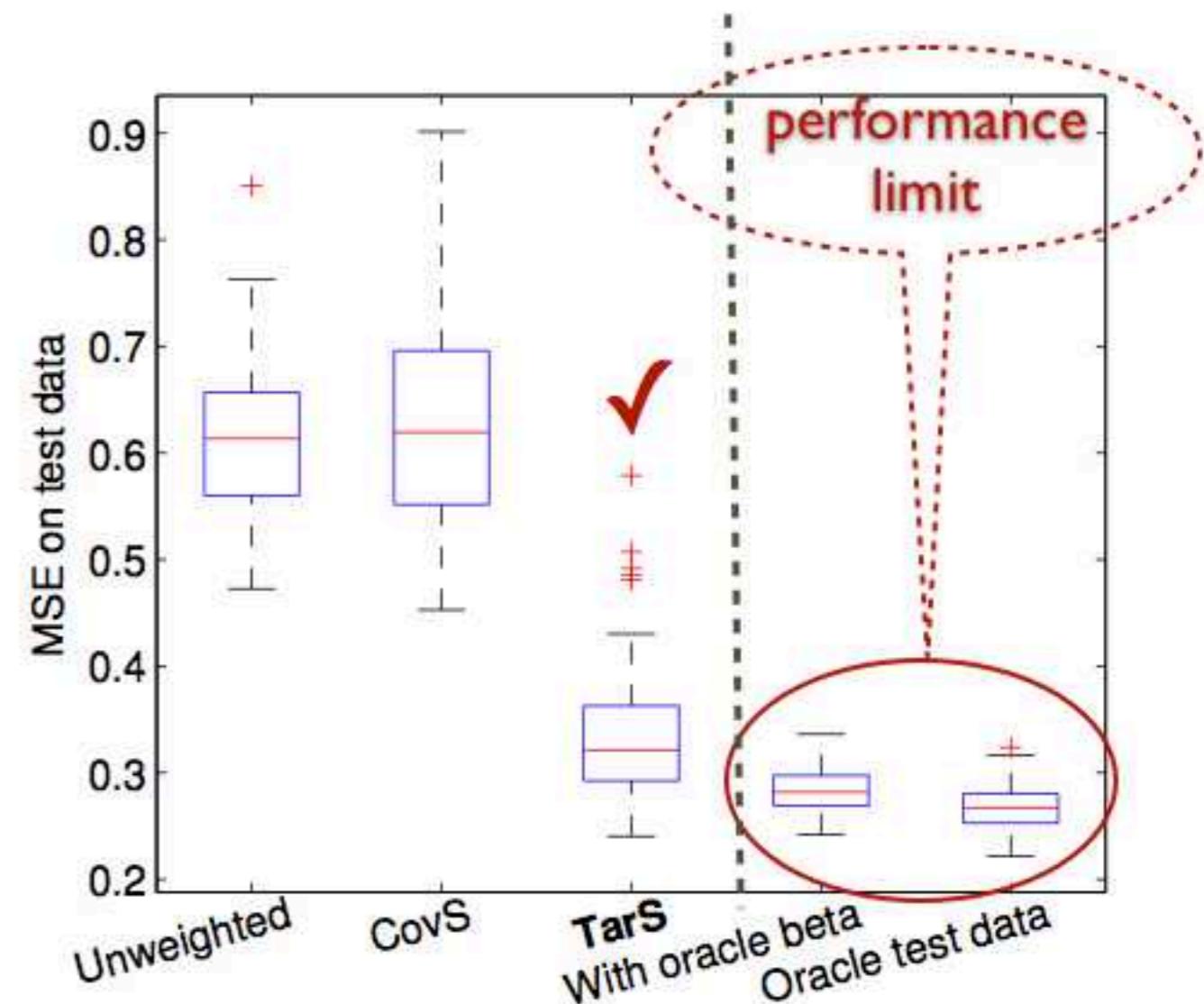
$$\begin{aligned} R[P^{te}, \theta, l(x, y, \theta)] &= \mathbb{E}_{(X, Y) \sim P^{te}} [l(x, y, \theta)] = \int P_Y^{tr} \cdot \beta^*(y) \cdot P_{X|Y}^{te} \cdot l(x, y, \theta) dx dy \\ &= \int P_Y^{tr} \cdot \beta^*(y) \cdot P_{X|Y}^{new} \cdot l(x, y, \theta) dx dy, = \underline{\mathbb{E}_{(X, Y) \sim P_{XY}^{new}} \beta^*(y) \cdot l(x, y, \theta)}, \end{aligned}$$

$$R_{emp}[P^{te}, \theta, l(x, y, \theta)] = \frac{1}{m} \sum_{i=1}^m \underline{\beta^*(y_i^{tr})} \underline{l(x_i^{new}, y_i^{tr}, \theta)}.$$

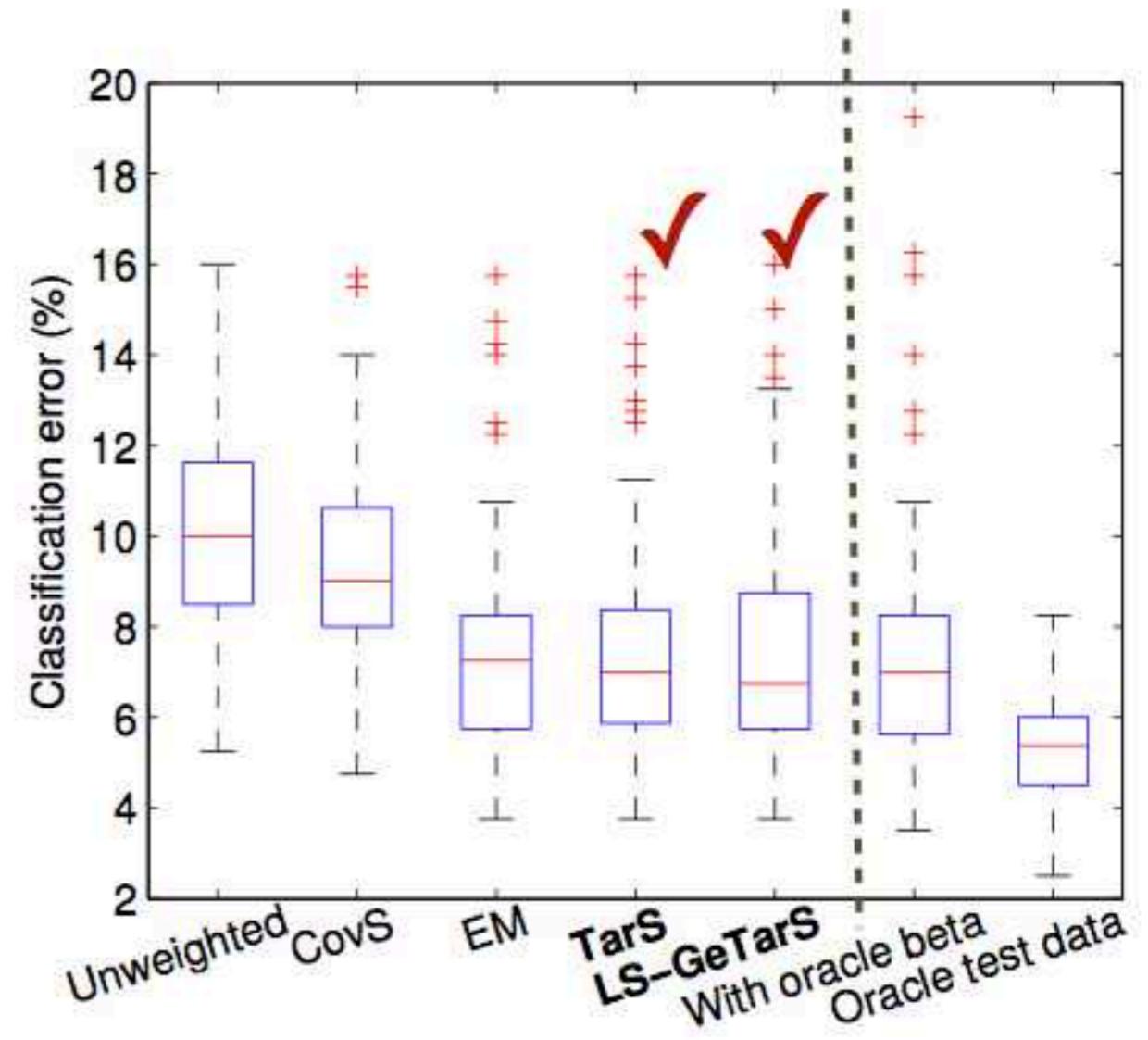
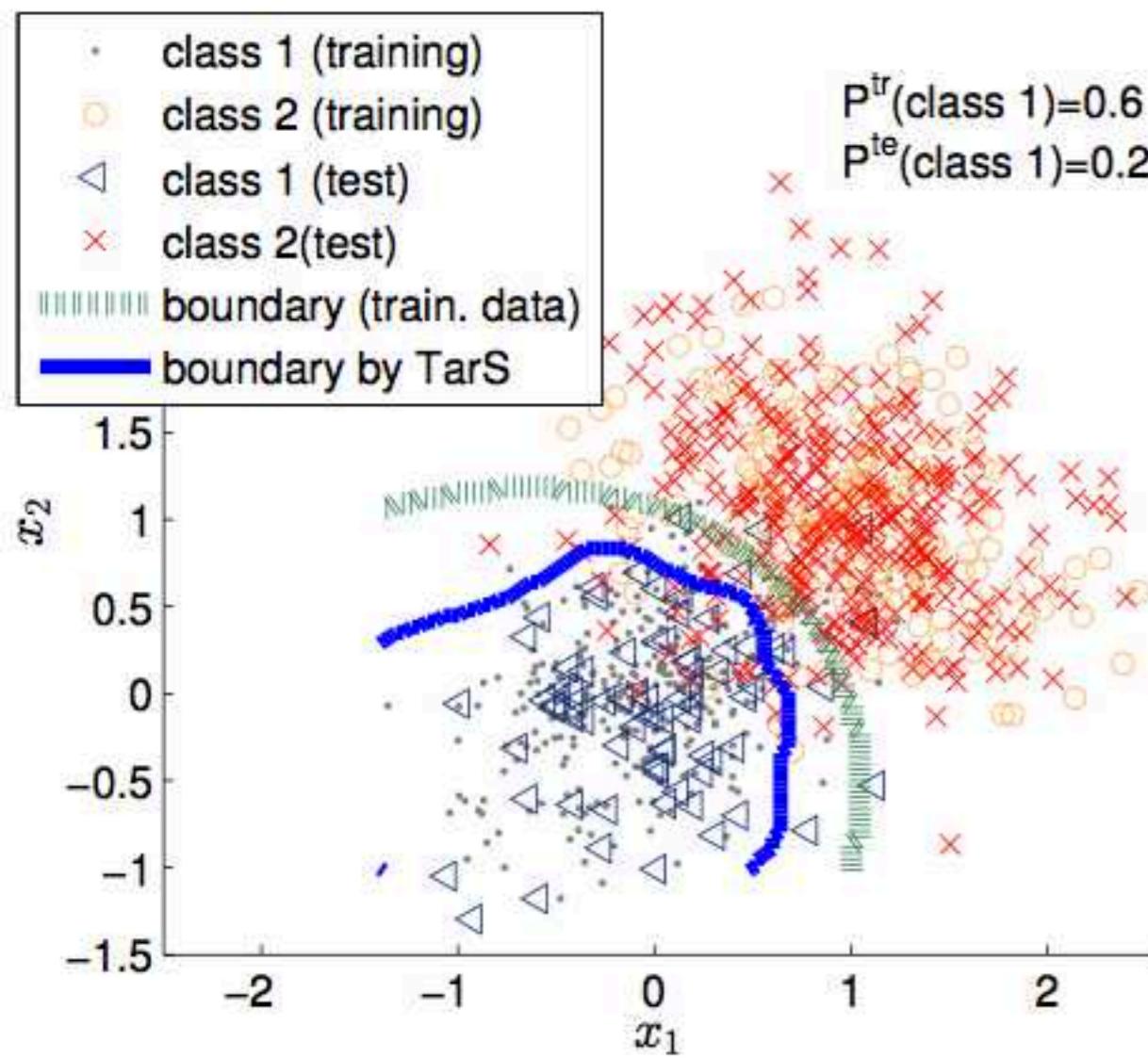
# Simulation I: Regression under TarS



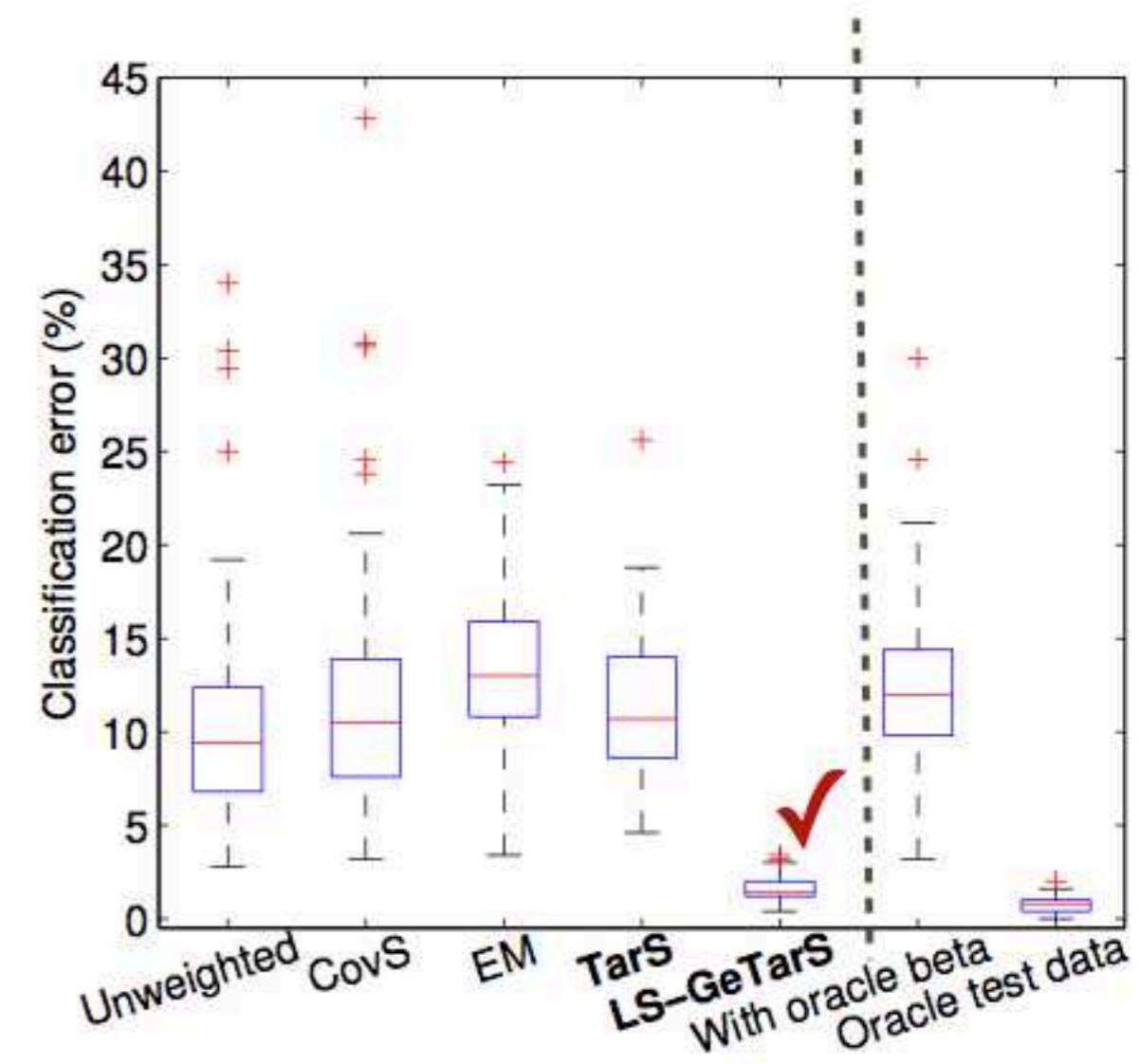
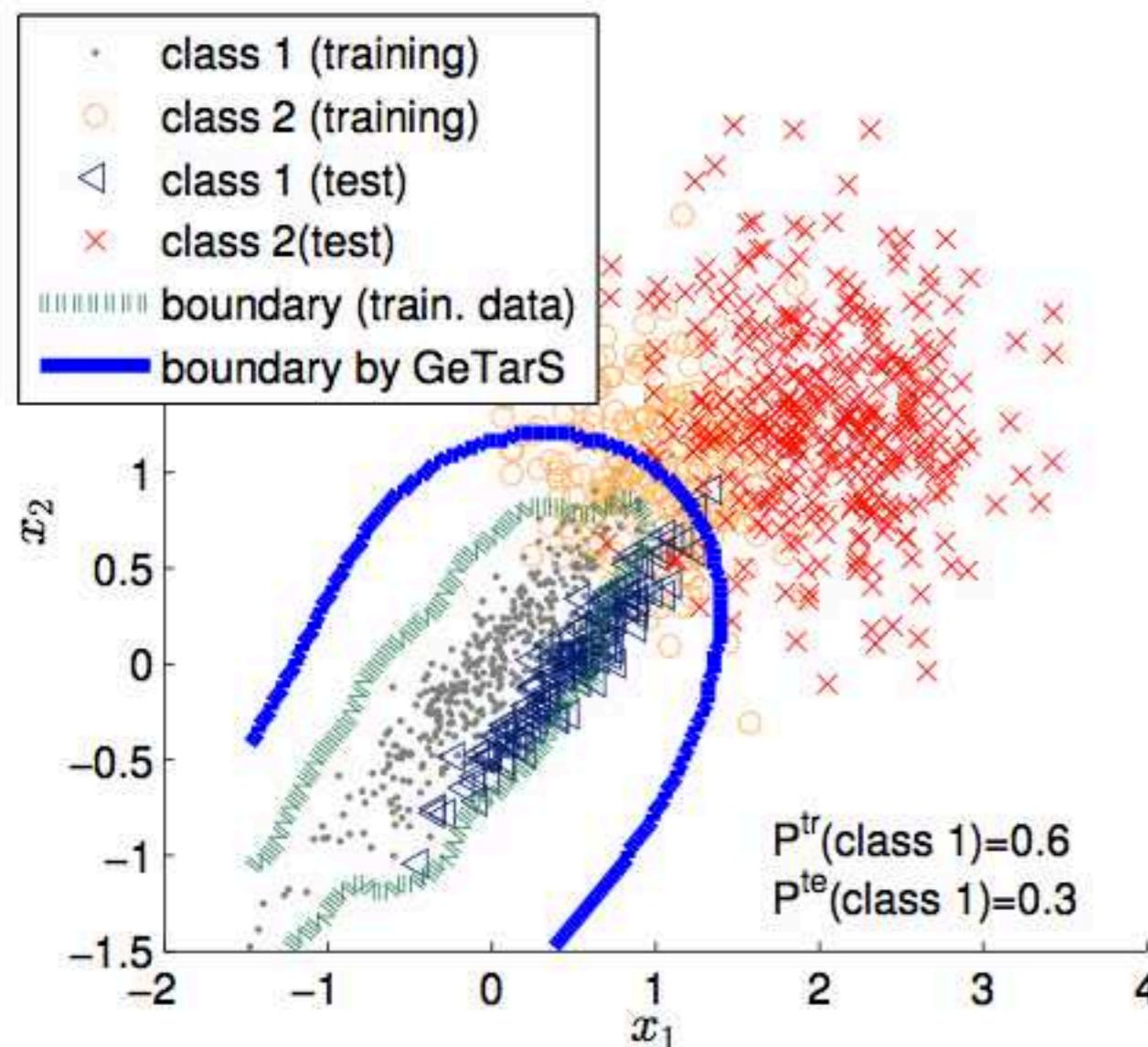
$$x = f(y) + e$$



# Simulation 2: Classification under TarS

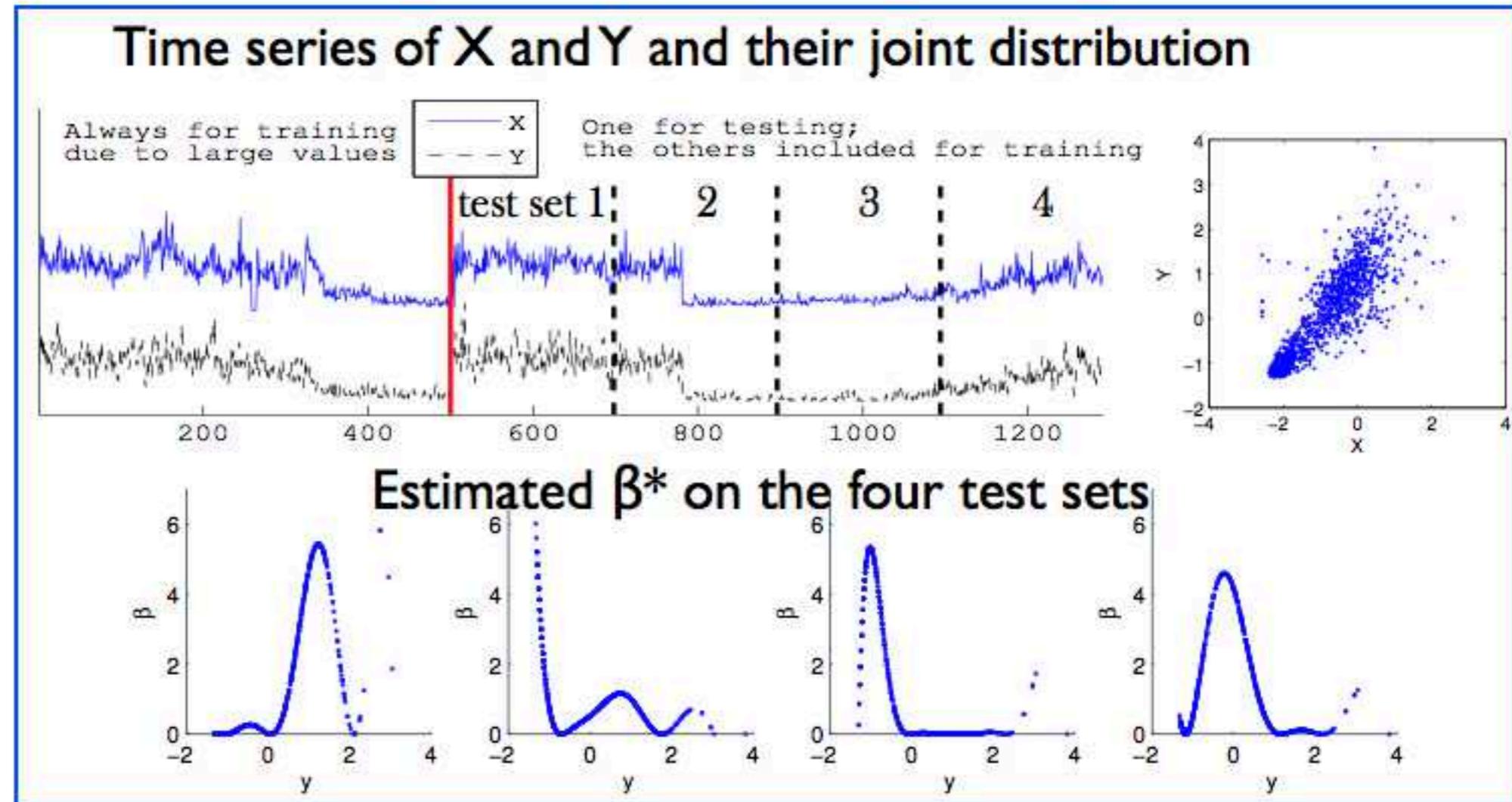


# Simulation 3: Classification under LS-GeTarS



# Regression under TarS: Real data

- Cause-effect pair 68
- X: # bytes sent by a computer and  
Y: # open http connections at the same time
- $Y \rightarrow X$
- TarS greatly improves the prediction performance for Y



Prediction performance (MSE) on test data

Test set	Unweight.	CovS	CovS ( $q = 0.5$ )	TarS	TarS ( $q=0.5$ )	
1	0.3789	0.3844	0.3802	0.3310	<b>0.3229</b>	✓
2	0.0969	0.1126	0.1071	0.0937	<b>0.0887</b>	
3	0.0578	0.0673	0.0659	<b>0.0466</b>	0.0489	
4	0.2054	0.2126	0.2136	0.2008	<b>0.1630</b>	✓

# Remote sensing image classification

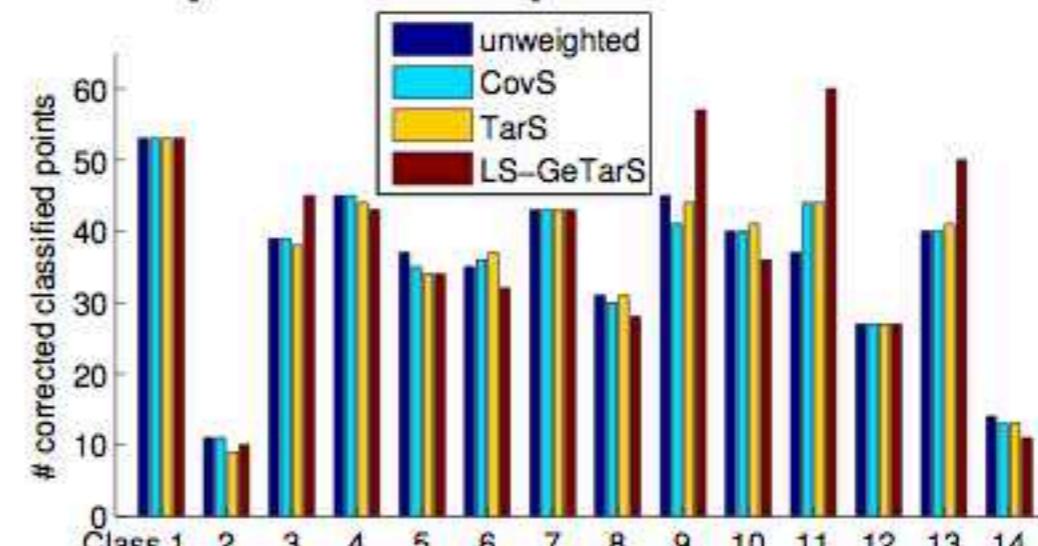
- two domains (area 1 & area 2)
- 14 classes

Class	Number of patterns			
	Area 1		Area 2	
	$TR_1$	$TS_1$	$TR_2$	$TS_2$
Water	69	57	213	57
Hippo grass	81	81	83	18
Floodplain grasses1	83	75	199	52
Floodplain grasses2	74	91	169	46
Reeds1	80	88	219	50
Riparian	102	109	221	48
Firescar2	93	83	215	44
Island interior	77	77	166	37
Acacia woodlands	84	67	253	61
Acacia shrublands	101	89	202	46
Acacia grasslands	184	174	243	62
Short mopane	68	85	154	27
Mixed mopane	105	128	203	65
Exposed soil	41	48	81	14
Total	1242	1252	2621	627

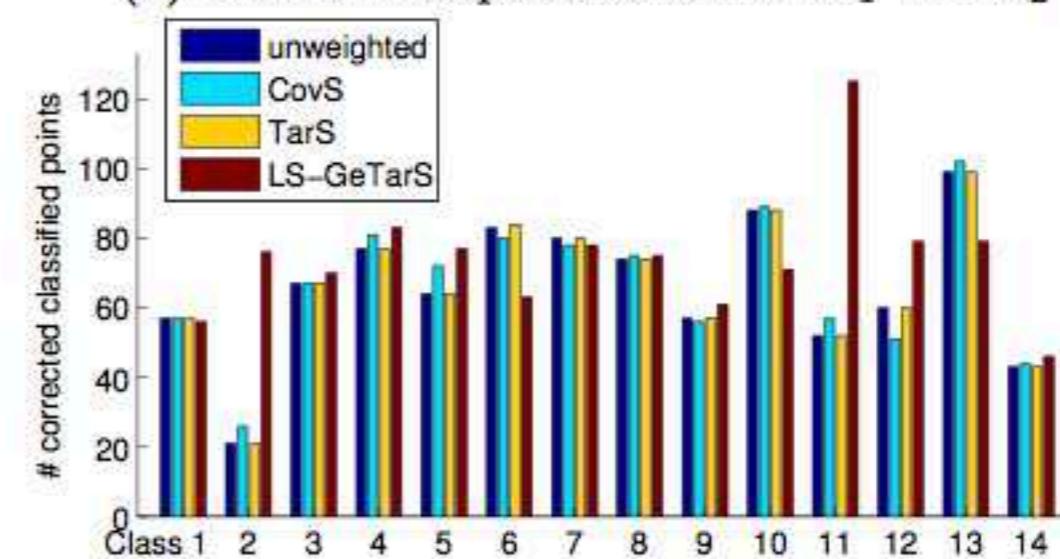
Misclassification rates by different methods

Problem	Unweight	CovS	TarS	LS-GeTarS
$TR_1 \rightarrow TS_2$	20.73%	20.73%	20.41%	11.96%
$TR_2 \rightarrow TS_1$	26.36%	25.32%	26.28%	13.56%

# correctly classified points for each class



(a) Domain adaptation from  $TR_1$  to  $TS_2$



(a) Domain adaptation from  $TR_2$  to  $TS_1$

# Summary

- Different causal models underlying covariate shift, target shift, conditional shift, and generalized target shift
- efficiently solved with kernel mean embedding
- Nothing comes from nothing: What to transfer ?
  - Background causal info. helps

