Transfer Learning with Applications

Sinno Jialin Pan[1], Qiang Yang[2,3] and Wei Fan[3]

[1] Institute for Infocomm Research, Singapore

[2] Hong Kong University of Science and Technology

[3] Huawei Noah's Ark Research Lab, Hong Kong

Outline

- **Part I:** An overview of transfer learning – (Sinno J. Pan)
- **Part II: T**ransfer learning applications (Prof. Qiang Yang)
- **Part III:** Advanced research topics: heterogeneous transfer learning (Wei Fan)

Transfer Learning Overview

Sinno Jialin Pan (Ph.D.)

Lab Head, Text Analytics,
Data Analytics Department,
Institute for Infocomm Research (I2R), Singapore

Agency for Science, Technology and Research
SINGAPORE

1991-2011
Creating Growth
Enhancing Lives

A*STAR I2R

Transfer of Learning

A psychological point of view

- The study of dependency of human conduct, learning or performance on prior experience.
  - [Thorndike and Woodworth, 1901] explored how individuals would transfer in one context to another context that share similar characteristics.

- C++ → Java
- Maths/Physics → Computer Science/Economics

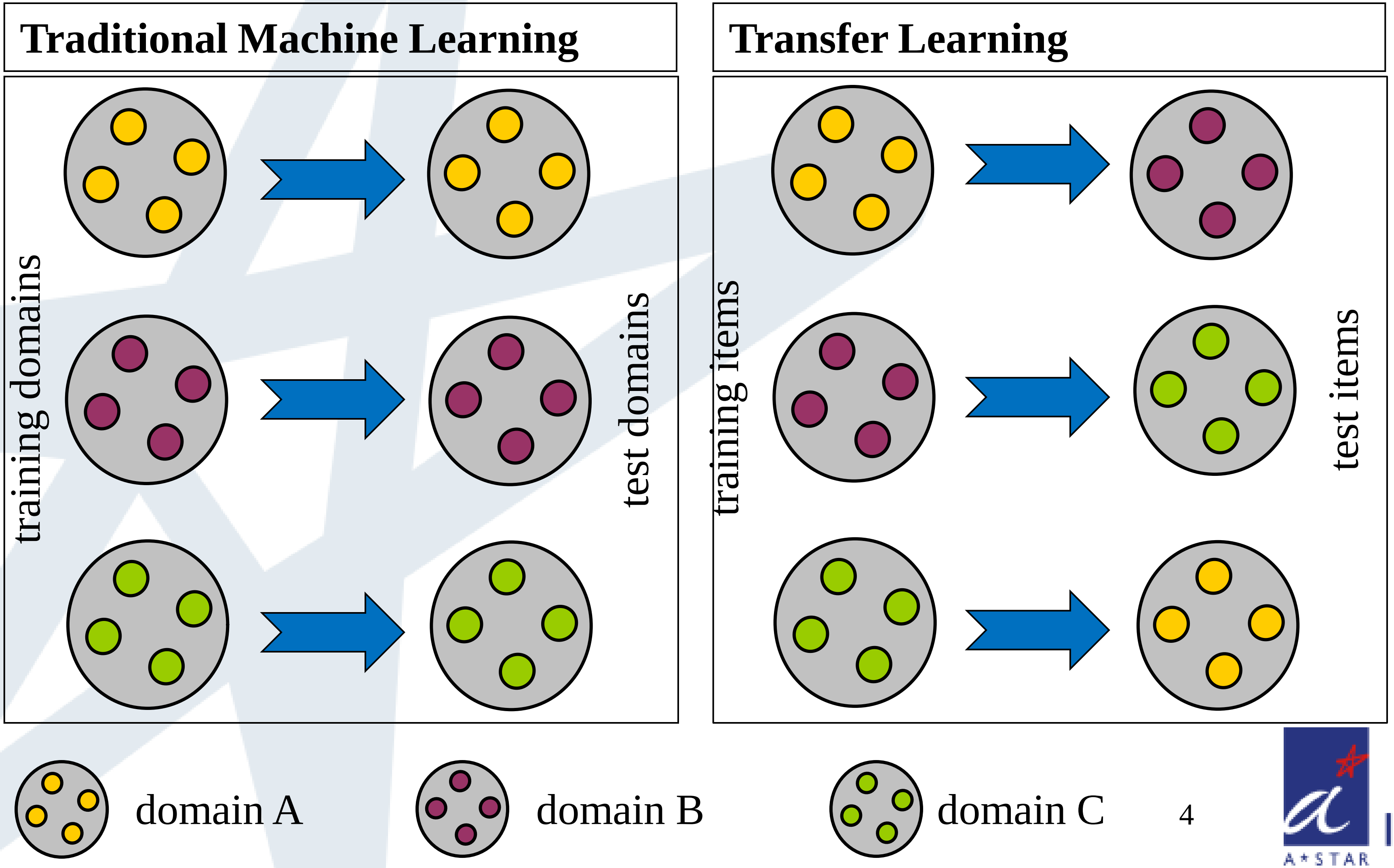
Transfer Learning

In the machine learning community

- The ability of a system to recognize and apply knowledge and skills learned in previous domains/tasks to novel tasks/domains, which share some commonality.

- Given a target domain/task, how to identify the commonality between the domain/task and previous domains/tasks, and transfer knowledge from the previous domains/tasks to the target one?

Transfer Learning



Transfer Learning

Different fields

- Transfer learning for reinforcement learning.

[Taylor and Stone, Transfer Learning for Reinforcement Learning Domains: A Survey, JMLR 2009]

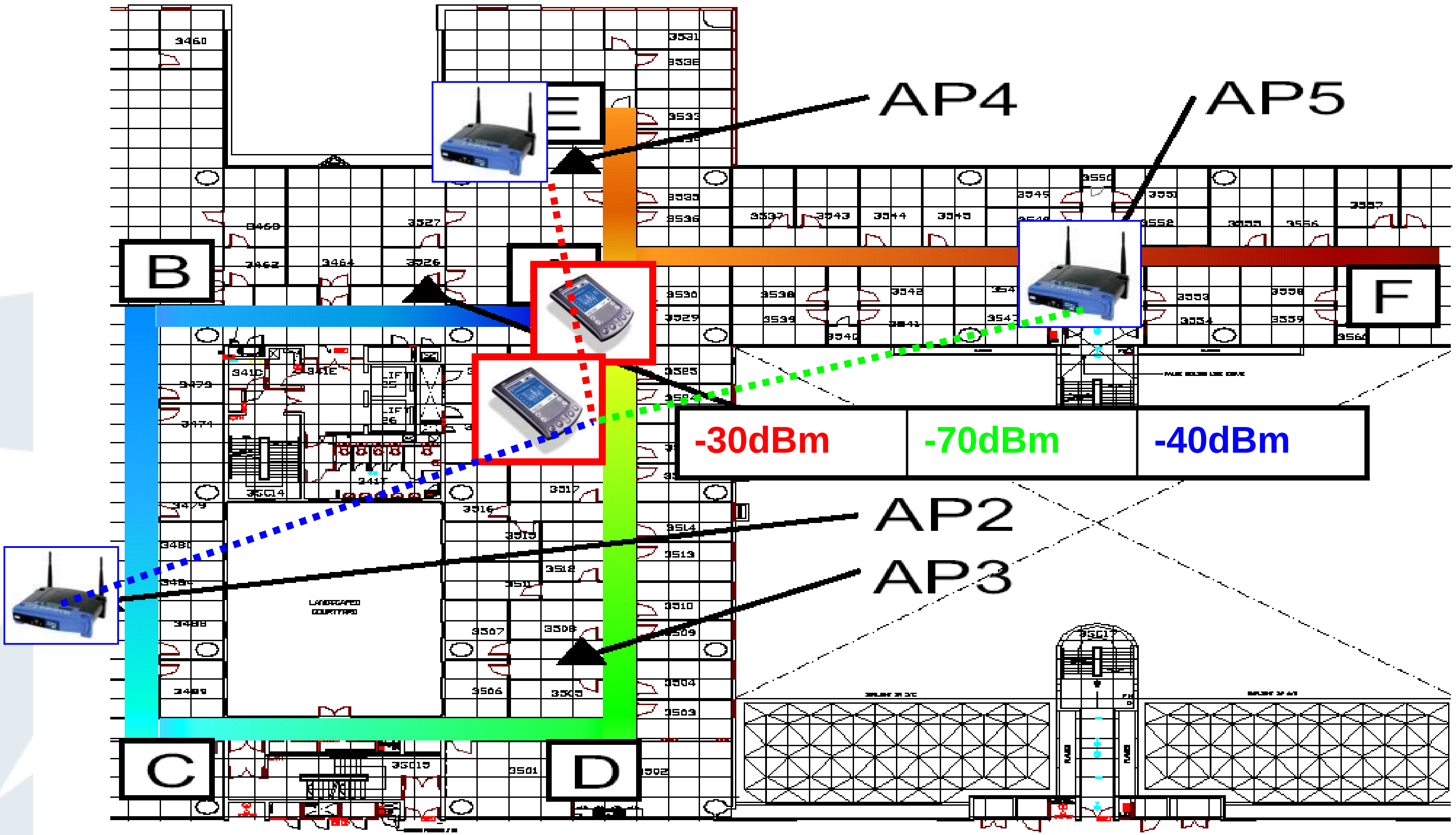
- Transfer learning for classification, and regression problems.

Focus!

[Pan and Yang, A Survey on Transfer Learning, IEEE TKDE 2010]

Motivating Example I:

Indoor WiFi localization



I²R
A*STAR

Indoor WiFi Localization (cont.)

Average Error Distance

Training

S=(-37dbm, .., -77dbm), L=(1, 3)
S=(-41dbm, .., -83dbm), L=(1, 4)
…
S=(-49dbm, .., -34dbm), L=(9, 10)
S=(-61dbm, .., -28dbm), L=(15,22)

Localization model

Test

S=(-37dbm, .., -77dbm)
S=(-41dbm, .., -83dbm)
…
S=(-49dbm, .., -34dbm)
S=(-61dbm, .., -28dbm)

~ 1.5 meters

Device A

Device A

Drop!

Training

S=(-33dbm, .., -82dbm), L=(1, 3)
…
S=(-57dbm, .., -63dbm), L=(10, 23)

Localization model

Test

S=(-37dbm, .., -77dbm)
S=(-41dbm, .., -83dbm)
…
S=(-49dbm, .., -34dbm)
S=(-61dbm, .., -28dbm)

~10 meters

Device B

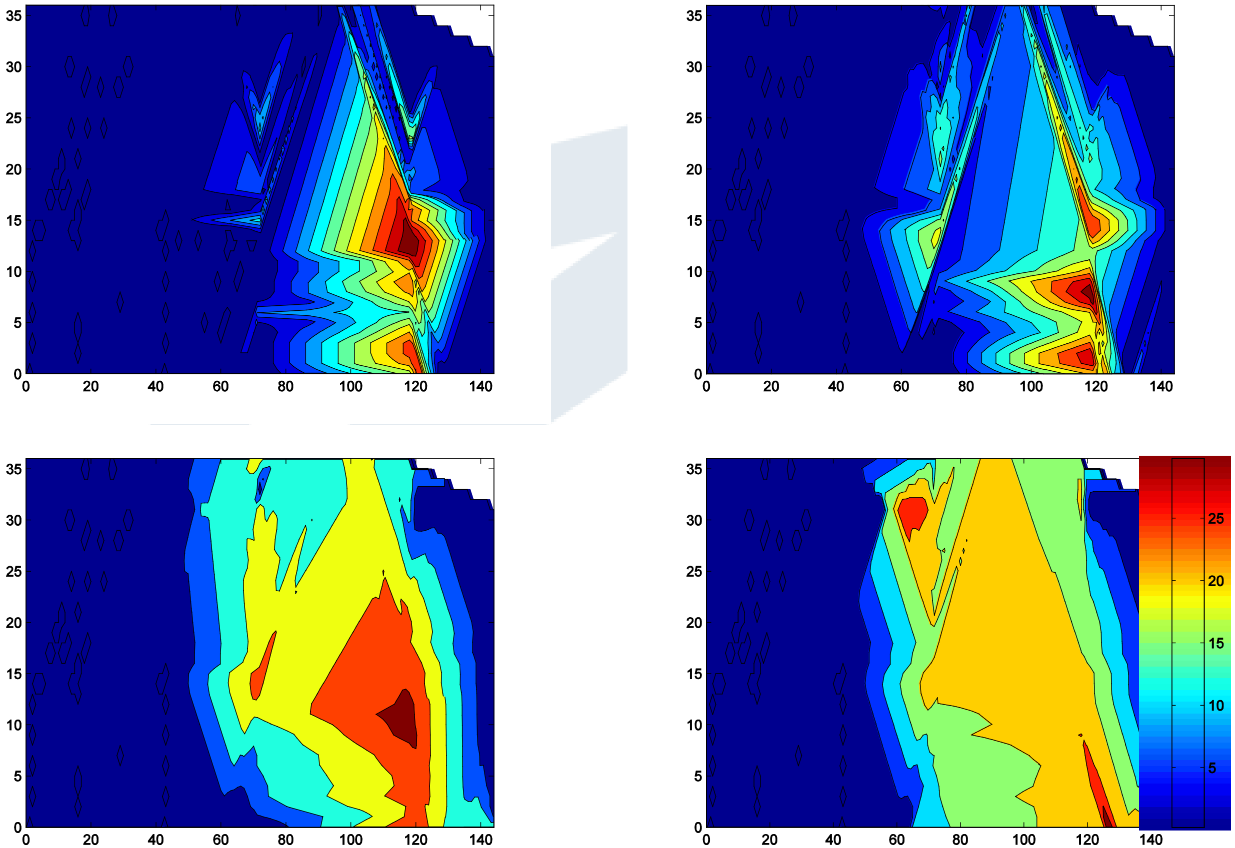
Device A

Difference between Domains

Time Period A | Time Period B

Device A

Device B



Motivating Example II:

Sentiment classification

10 hours ago
Edward Priz★ replied:

You know, this isn't the first time that "States Rights" has been used as a cover for racist policies. In fact, the whole "States Rights" thing has become a sort of code for heavy-handed racist policies, hasn't it? And it does provide a sort of contextual

10 hours ago
RICH HIRTH★ replied:

The issue here is probable cause. A police officer can question if he has probable cause, and he can document it. This law can be abused if being Latino is probable cause. That is license to harass for the police. As long as the law is applied fairly there

2 hours ago
Julia Gomez replied:

The Arizona law is so clearly unconstitutional that I do not think it will ever reach the point of being enforced. The article did not say so, but the Republican governor is afraid of a GOP primary electorate that is even more reactionary than usual. That is why she signed the bill, not because she thinks it is legally defensible.

Sentiment Classification (cont.)

Classification Accuracy

Training

10 hours ago
Edward Priz replied:
You know, this isn't the first time that "States Rights" has been used as a cover for racist policies. In fact, the whole "States Rights" thing has become a sort of code for heavy-handed racist policies, hasn't it? And it does provide a sort of contextual link with those heroic days when evil was confronted in places like Selma and Little Rock, doesn't it? Thanks for making that link explicit.

Electronics

Sentiment Classifier

Test

10 hours ago
Edward Priz replied:
You know, this isn't the first time that "States Rights" has been used as a cover for racist policies. In fact, the whole "States Rights" thing has become a sort of code for heavy-handed racist policies, hasn't it? And it does provide a sort of contextual link with those heroic days when evil was confronted in places like Selma and Little Rock, doesn't it? Thanks for making that link explicit.

Electronics

~ 84.6%

Drop!

Training

10 hours ago
RICH HIRTH replied:
The issue here is probable cause. A police officer can question if he has probable cause, and he can document it. This law can be abused if being Latino is probable cause. That is license to harass for the police. As long as the law is applied fairly there should not be a problem. As far as documentation, Most states have laws that citizens must carry valid state ID, and no one cares. There is no reason the Executive branch needed to get involved in what the Court should decide.

DVD

Sentiment Classifier

Test

10 hours ago
Edward Priz replied:
You know, this isn't the first time that "States Rights" has been used as a cover for racist policies. In fact, the whole "States Rights" thing has become a sort of code for heavy-handed racist policies, hasn't it? And it does provide a sort of contextual link with those heroic days when evil was confronted in places like Selma and Little Rock, doesn't it? Thanks for making that link explicit.

Electronics

~72.65%

I²R
A*STAR

Difference between Domains

	Electronics	Video Games
	(1) **Compact**; easy to operate; very good picture quality; looks **sharp**!	(2) A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.
	(3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and **sharp**.	(4) Very **realistic** shooting action and good plots. We played this and were **hooked**.
	(5) It is also quite **blurry** in very dark settings. I will never buy HP again.	(6) The game is so **boring**. I am extremely unhappy and will probably never buy UbiSoft again.

A Major Assumption in Traditional Machine Learning

- Training and future (test) data come from the same domain, which implies
  - Represented in the same feature spaces.
  - Follow the same data distribution.

In Real-world Applications

- Training and testing data may come from different domains, which have:
  - ❑Different marginal distributions, or different feature spaces:

    $$\mathcal{X}_S \neq \mathcal{X}_T, \text{ or } P_S(x) \neq P_T(x)$$

  - ❑Different predictive distributions, or different label spaces:

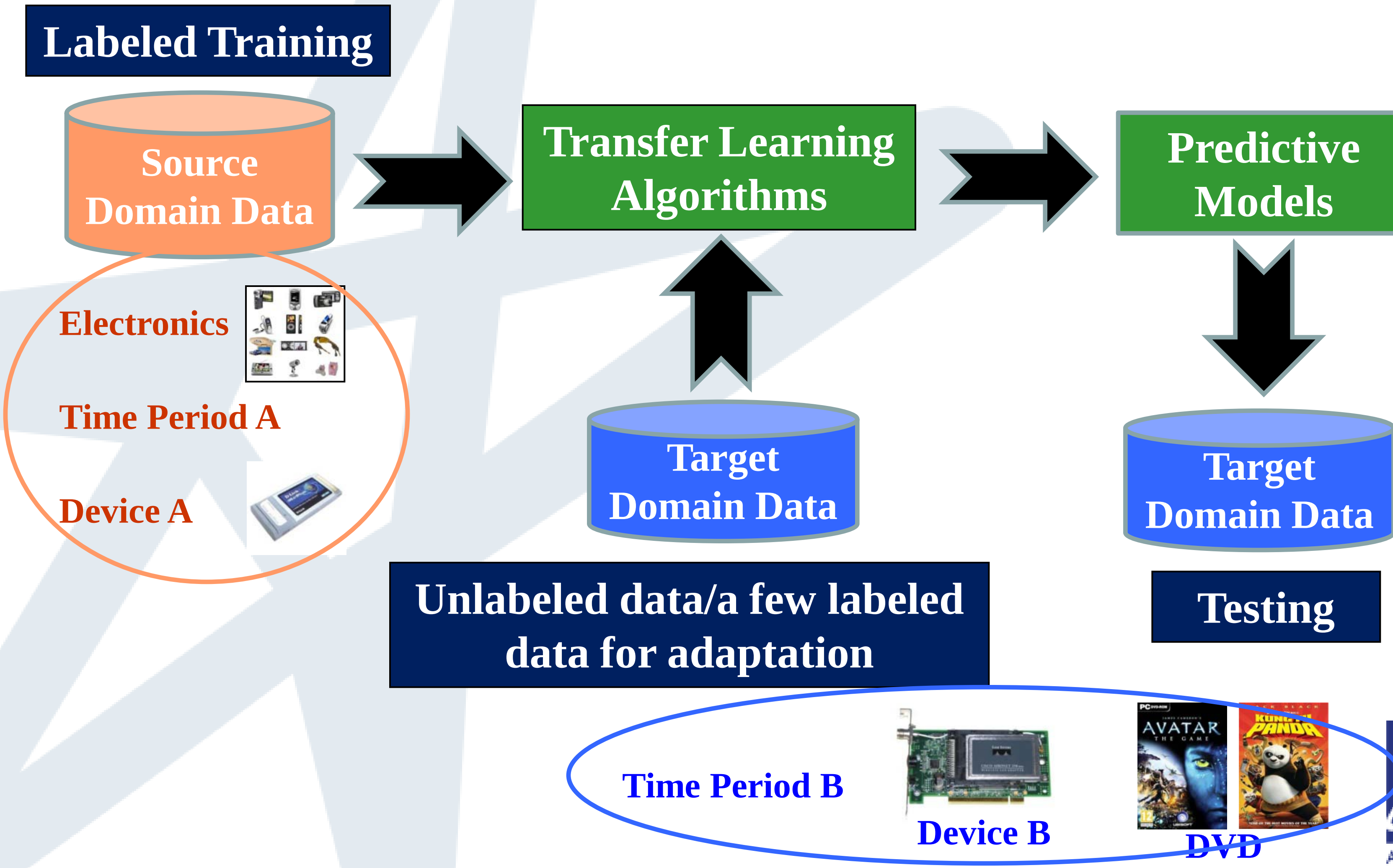
    $$\mathcal{Y}_S \neq \mathcal{Y}_T, \text{ or } f_S \neq f_T \ (P_S(y|x) \neq P_T(y|x))$$

How to Build Systems on Each Domain of Interest

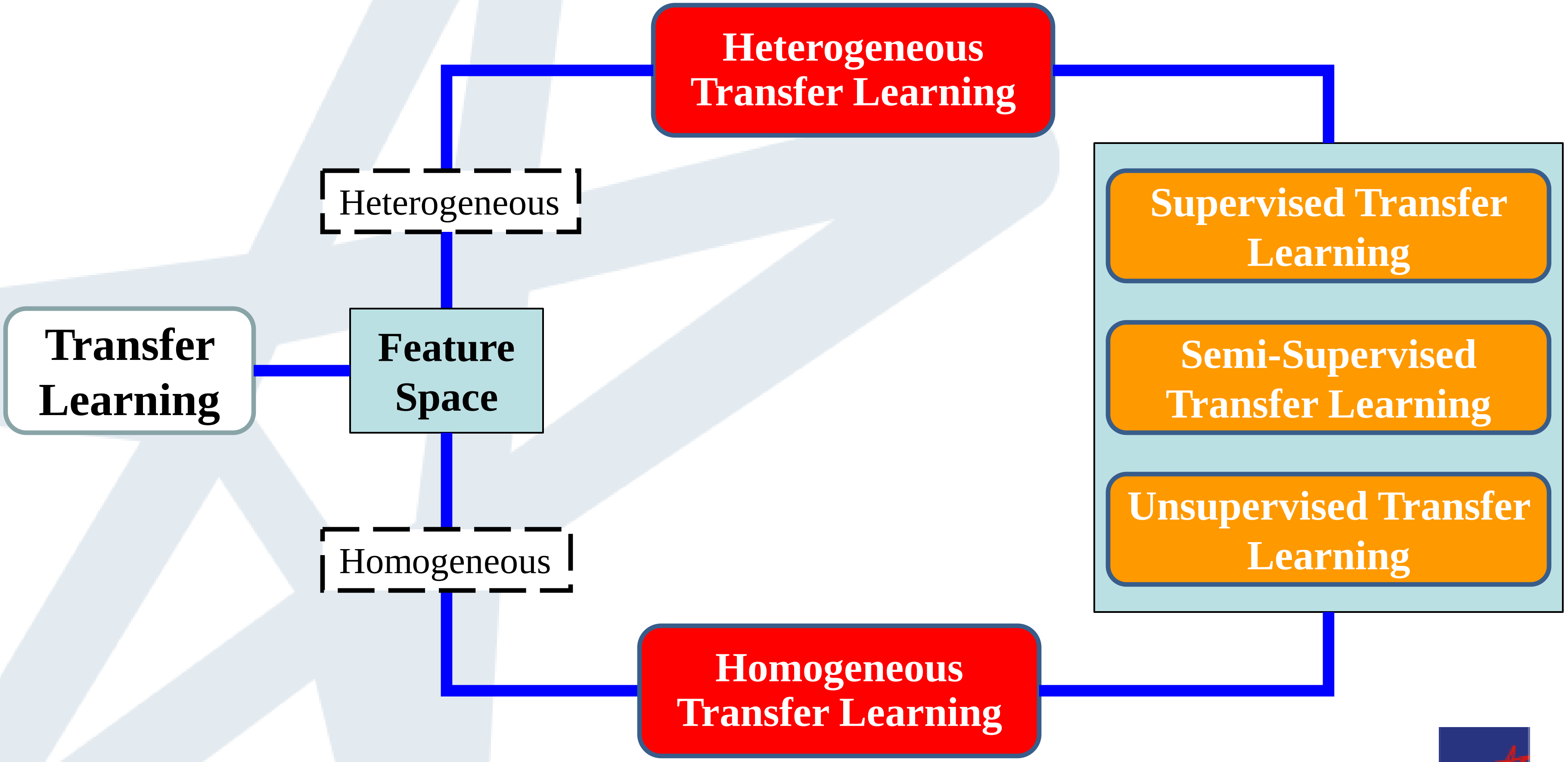
- Build every system from scratch?
  - Time consuming and expensive!

- Reuse common knowledge extracted from existing systems?
  - More practical!

The Goal of Transfer Learning



Transfer Learning Settings



Transfer Learning Approaches

Instance-based Approaches

Feature-based Approaches

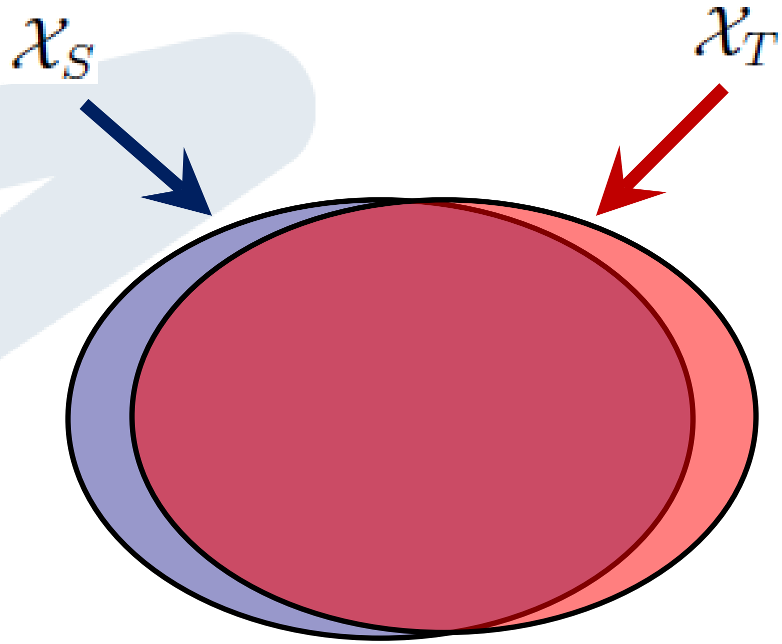
Parameter-based Approaches

Relational Approaches

Instance-based Transfer Learning Approaches

General Assumption

Source and target domains have a lot of overlapping features (domains share the same/similar support)



Instance-based Transfer Learning Approaches

Case I

Problem Setting

Given $\mathrm{D}_S = \{x_{S_i}, y_{S_i}\}_{i=1}^{n_S}$, $\mathrm{D}_T = \{x_{T_i}\}_{i=1}^{n_T}$,

Learn f_T, s.t. $\sum_i \epsilon(f_T(x_{T_i}), y_{T_i})$ is small,

where y_{T_i} is unknown.

Assumption

- $\mathcal{Y}_S = \mathcal{Y}_T$, and $P(Y_S|X_S) = P(Y_T|X_T)$,
- $\mathcal{X}_S \approx \mathcal{X}_T$,
- $P(X_S) \neq P(X_T)$.

Case II

Problem Setting

Given $\mathrm{D}_S = \{x_{S_i}, y_{S_i}\}_{i=1}^{n_S}$,

$\mathrm{D}_T = \{x_{T_i}, y_{T_i}\}_{i=1}^{n_T}$, $n_T \ll n_S$,

Learn f_T, s.t. $\epsilon(f_T(x_{T_i}), y_{T_i})$ is small, and

f_T has good generalization on unseen x_T^*.

Assumption

- $\mathcal{Y}_S = \mathcal{Y}_T$,

  but $f_S \neq f_T$ $(P_S(y|x) \neq P_T(y|x))$.

Instance-based Approaches

Case I

Given a target task,

$$
\begin{aligned}
\theta^* &= \arg\min \mathbb{E}_{(x,y)\sim P_T}\left[l(x,y,\theta)\right] \\
&= \arg\min \mathbb{E}_{(x,y)\sim P_T}\left[\frac{P_S(x,y)}{P_S(x,y)} l(x,y,\theta)\right] \\
&= \arg\min \int_y \int_x P_T(x,y)\left(\frac{P_S(x,y)}{P_S(x,y)} l(x,y,\theta)\right) dx dy \\
&= \arg\min \int_y \int_x P_S(x,y)\left(\frac{P_T(x,y)}{P_S(x,y)} l(x,y,\theta)\right) dx dy \\
&= \arg\min \mathbb{E}_{(x,y)\sim P_S}\left[\frac{P_T(x,y)}{P_S(x,y)} l(x,y,\theta)\right]
\end{aligned}
$$

Instance-based Approaches

Case I (cont.)

If $P_S(x, y) = P_T(x, y)$

$$\theta^* = \arg\min \mathbb{E}_{(x_T, y_T) \sim P_T}[l(x_T, y_T, \theta)]$$

$$\Updownarrow$$

$$\theta^* = \arg\min \mathbb{E}_{(x_S, y_S) \sim P_S}[l(x_S, y_S, \theta)]$$

$$\Downarrow$$

$$\theta^* = \arg\min \sum_{i=1}^{n_S} l(x_{S_i}, y_{S_i}, \theta) + \lambda \Omega(\theta)$$

Instance-based Approaches

Case I (cont.)

Assumption: $\{P_S(x) \neq P_T(x),\ P_S(y|x) = P_T(y|x)\} \Rightarrow P_S(x,y) \neq P_T(x,y)$

$$
\begin{aligned}
\theta^* &= \arg\min \mathbb{E}_{(x,y)\sim P_S}\left[\frac{P_T(x,y)}{P_S(x,y)} l(x,y,\theta)\right] \\
&= \arg\min \mathbb{E}_{(x,y)\sim P_S}\left[\frac{P_T(x)P_T(y|x)}{P_S(x)P_S(y|x)} l(x,y,\theta)\right] \\
&= \arg\min \mathbb{E}_{(x,y)\sim P_S}\left[\frac{P_T(x)}{P_S(x)} l(x,y,\theta)\right]
\end{aligned}
$$

$$
\text{Denote } \beta(x) = \frac{P_T(x)}{P_S(x)},
$$

$$
\theta^* = \arg\min \sum_{i=1}^{n_S} \beta(x_{S_i}) l(x_{S_i}, y_{S_i}, \theta) + \lambda\Omega(\theta)
$$

Instance-based Approaches

Case I (cont.)

How to estimate $\beta(x) = \dfrac{P_T(x)}{P_S(x)}$?

A simple solution is to first estimate $P_T(x)$, $P_S(x)$, respectively, and calculate $\dfrac{P_T(x)}{P_S(x)}$.

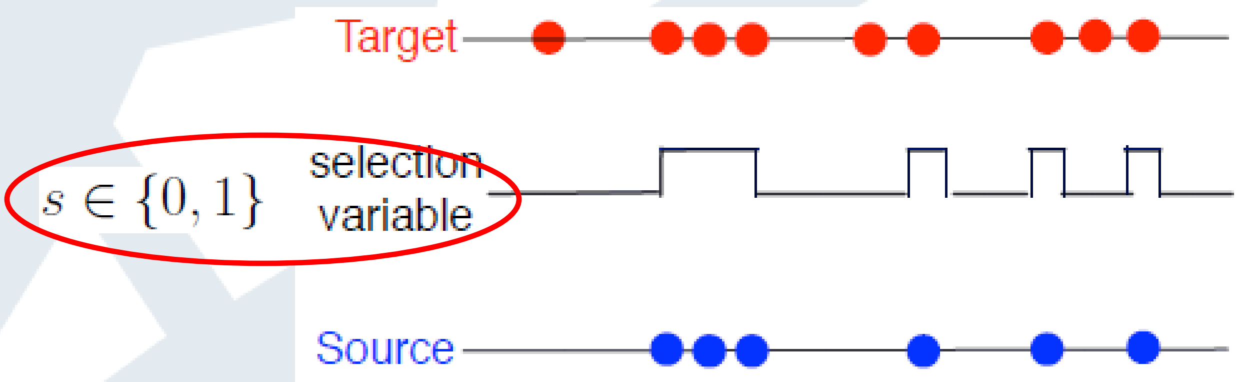
An alterative solution is to estimate $\dfrac{P_T(x)}{P_S(x)}$ directly.

Correcting Sample Selection Bias / Covariate Shift
[Quionero-Candela, *etal*, Data Shift in Machine Learning, MIT Press 2009]

Instance-based Approaches

Correcting sample selection bias

- Imagine a ***rejection*** sampling process, and view the source domain as samples from the target domain



Assumption: sample selection bias is caused by the data generation process

Instance-based Approaches

Correcting sample selection bias (cont.)

- The distribution of the selector variable maps the target onto the source distribution

$$P_S(x) \propto P_T(x) P(s=1|x)$$

$$\Downarrow$$

$$\beta(x) = \frac{P_S(x)}{P_T(x)} \propto \frac{1}{P(s=1|x)}$$

[Zadrozny, ICML-04]

- ➢ Label instances from the source domain with label 1
- ➢ Label instances from the target domain with label 0
- ➢ Train a binary classifier

Instance-based Approaches

Kernel mean matching (KMM)

Maximum Mean Discrepancy (MMD)

Given $\mathbf{X}_S = \{x_{S_i}\}_{i=1}^{n_S}$, $\mathbf{X}_T = \{x_{T_i}\}_{i=1}^{n_T}$, drown from $P_S(x)$ and $P_T(x)$, respectively,

$$\mathrm{Dist}(P(X_S), P(X_T)) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \Phi(x_{S_i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \Phi(x_{T_j}) \right\|_{\mathcal{H}}$$

[Alex Smola, Arthur Gretton and Kenji Kukumizu, ICML-08 tutorial]

Instance-based Approaches

Kernel mean matching (KMM) (cont.)

[Huang *etal*., NIPS-06]

$$\arg\min_{\beta} \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \beta(x_{S_i}) \Phi(x_{S_i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \Phi(x_{T_j}) \right\|$$

$$s.t \quad \beta(x_{S_i}) \in [0,\ B] \text{ and } \left| \frac{1}{n_S} \sum_{i=1}^{n_S} \beta(x_{S_i}) - 1 \right| \leq \epsilon.$$

Instance-based Approaches

Direct density ratio estimation

[Sugiyama *etal*., NIPS-07, Kanamori *etal*., JMLR-09]

Recall $\beta(x) = \dfrac{P_T(x)}{P_S(x)}$

Let $\widetilde{\beta}(x) = \sum_{\ell=1}^{b} \alpha_\ell \psi_\ell(x)$, and denote $\widetilde{P}_T(x) = \widetilde{\beta}(x) P_S(x)$

KL divergence loss

Least squared loss

$$\arg\min_{\{\alpha_\ell\}_{\ell=1}^{b}} \mathrm{KL}[P_T(x) || \widetilde{P}_T(x)]$$

[Sugiyama *etal*., NIPS-07]

$$\arg\min_{\{\alpha_\ell\}_{\ell=1}^{b}} \int_{X_S \cup X_T} \left(\widetilde{\beta}(x) - \beta(x)\right)^2 P_S(x) dx$$

[Kanamori *etal*., JMLR-09]

Instance-based Approaches

Case II

- $\mathcal{Y}_S = \mathcal{Y}_T$,
  but $f_S \neq f_T$ $(P_S(y|x) \neq P_T(y|x))$.
- Intuition: Part of the labeled data in the source domain can be reused in the target domain after re-weighting

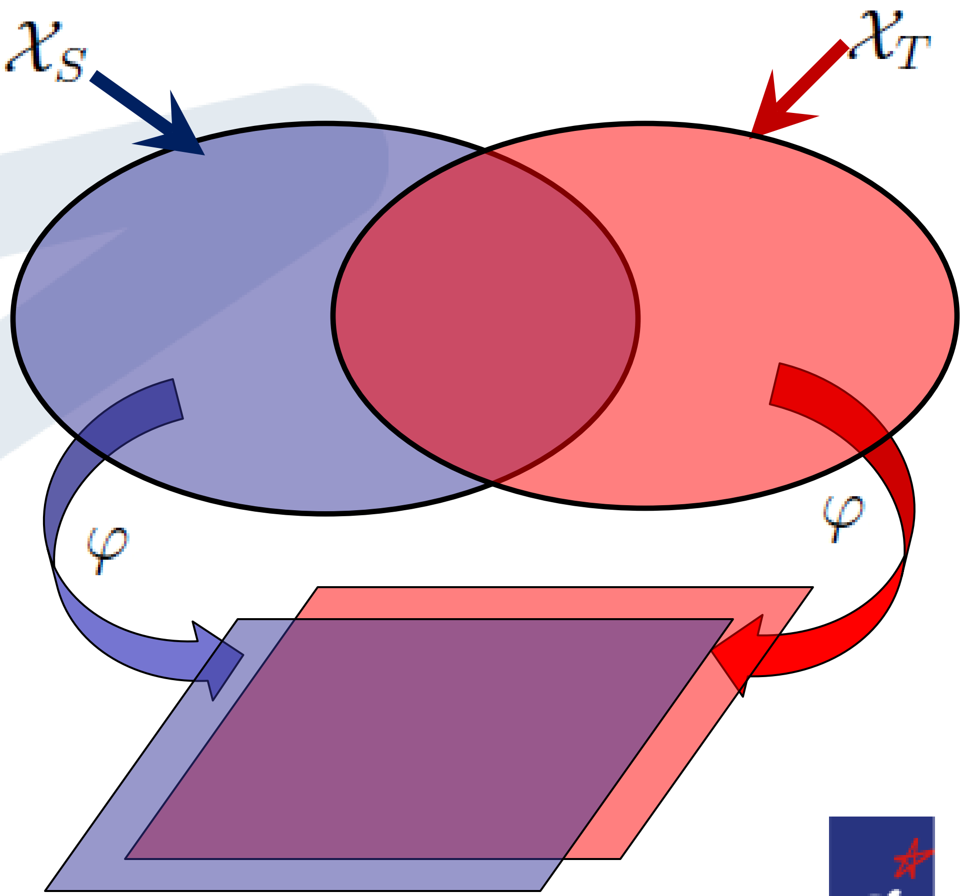
Instance-based Approaches

Case II (cont.)

- **TrAdaBoost** [Dai *etal* ICML-07]
  - For each boosting iteration,
    - Use the same strategy as AdaBoost to update the weights of target domain data.
    - Use a new mechanism to decrease the weights of misclassified source domain data.

Feature-based Transfer Learning Approaches

When source and target domains only have some overlapping features. (lots of features only have support in either the source or the target domain)



Feature-based Transfer Learning Approaches (cont.)

How to learn φ ?

- Solution 1: Encode application-specific knowledge to learn the transformation.

- Solution 2: General approaches to learning the transformation.

Feature-based Approaches

Encode application-specific knowledge

Electronics	Video Games
(1) **Compact**; easy to operate; very good picture quality; looks **sharp**!	(2) A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.
(3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and **sharp**.	(4) Very **realistic** shooting action and good plots. We played this and were **hooked**.
(5) It is also quite **blurry** in very dark settings. I will never_buy HP again.	(6) The game is so **boring**. I am extremely unhappy and will probably never_buy UbiSoft again.

Feature-based Approaches

Encode application-specific knowledge (cont.)

Electronics

	compact	sharp	blurry	hooked	realistic	boring
	1	1	0	0	0	0
	0	1	0	0	0	0
	0	0	1	0	0	0

Training

$$y = f(x) = \mathrm{sgn}(w \cdot x^T), \qquad w = [1, 1, -1, 0, 0, 0]$$

Prediction

Video Game

	compact	sharp	blurry	hooked	realistic	boring
	0	0	0	1	0	0
	0	0	0	1	1	0
	0	0	0	0	0	1

Feature-based Approaches

Encode application-specific knowledge (cont.)

Electronics	Video Games
(1) **Compact**; easy to operate; very ***good*** picture quality; looks **sharp**!	(2) A very ***good*** game! It is action packed and full of ***excitement***. I am very much **hooked** on this game.
(3) I purchased this unit from Circuit City and I was very ***excited*** about the quality of the picture. It is really ***nice*** and **sharp**.	(4) Very **realistic** shooting action and ***good*** plots. We played this and were **hooked**.
(5) It is also quite **blurry** in very dark settings. I will ***never_buy*** HP again.	(6) The game is so **boring**. I am extremely ***unhappy*** and will probably ***never_buy*** UbiSoft again.

Feature-based Approaches

Encode application-specific knowledge (cont.)

- Three different types of features
  - Source domain (***Electronics***) specific features, e.g.,
    compact, sharp, blurry
  - Target domain (***Video Game***) specific features, e.g.,
    hooked, ***realistic***, ***boring***
  - Domain independent features (pivot features), e.g.,
    good, excited, ***nice***, ***never_buy***

Feature-based Approaches

Encode application-specific knowledge (cont.)

- How to identify ***pivot*** features?
  - Term frequency on both domains
  - Mutual information between features and labels (source domain)
  - Mutual information on between features and domains

- How to utilize pivots to ***align*** features across domains?
  - Structural Correspondence Learning (SCL) [Biltzer *etal.* EMNLP-06]
  - Spectral Feature Alignment (SFA) [Pan *etal*. WWW-10]

Feature-based Approaches

Structural Correspondence Learning (SCL)

- **Intuition**
  - Use ***pivot*** features to construct ***pseudo*** tasks that related to target classification task
  - Model correlations between ***pivot*** features and other features using multi-task learning techniques
  - Discover new shared features by exploiting the feature correlations

Structural Correspondence Learning

Algorithm

- Identify *P* ***pivot*** features
- Build *P* classifiers to predict the pivot features from remaining features
- Discover ***shared*** feature subspace
  - Compute top *K* ***eigenvectors***
  - Project original features into eigenvectors to derive new shared features
- Train classifiers on the source using ***augmented*** features (original features + new features)

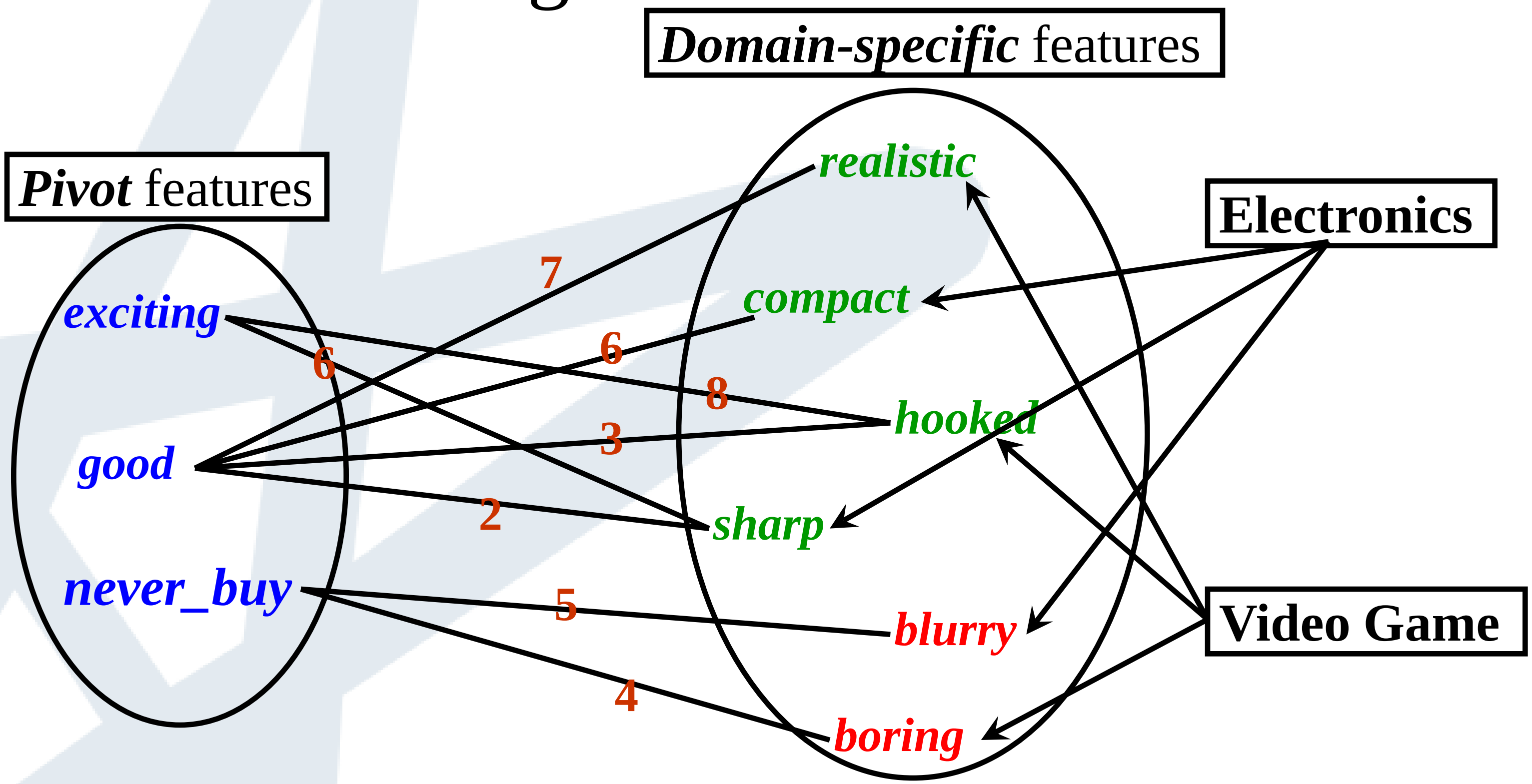
Feature-based Approaches

Spectral Feature Alignment (SFA)

- **Intuition**
  - Use a ***bipartite*** graph to model the correlations between ***pivot*** features and other features
  - Discover new shared features by applying ***spectral clustering*** techniques on the graph

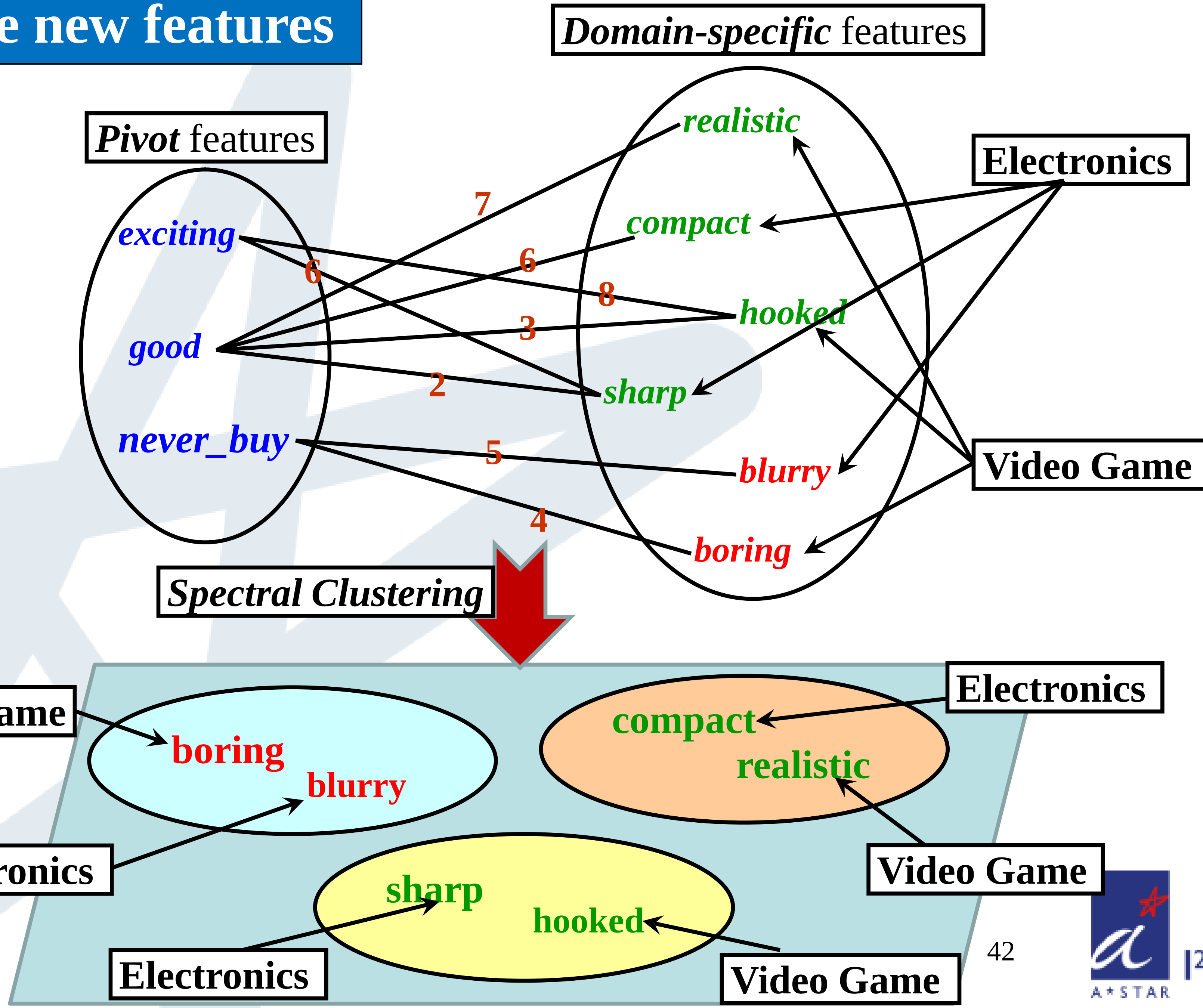
Spectral Feature Alignment (SFA)

High level idea



- ➢ If two ***domain-specific*** words have connections to more common ***pivot*** words in the graph, they tend to be aligned or clustered together with a higher probability.
- ➢ If two ***pivot*** words have connections to more common ***domain-specific*** words in the graph, they tend to be aligned together with a higher probability.

Derive new features



Spectral Feature Alignment (SFA)

Derive new features (cont.)

Electronics

	sharp/hooked	compact/realistic	blurry/boring
	1	1	0
	1	0	0
	0	0	1

Training

$$y = f(x) = \mathrm{sgn}(w \cdot x^T), \qquad w = [1, 1, -1]$$

Prediction

Video Game

	sharp/hooked	compact/realistic	blurry/boring
	1	0	0
	1	1	0
	0	0	1

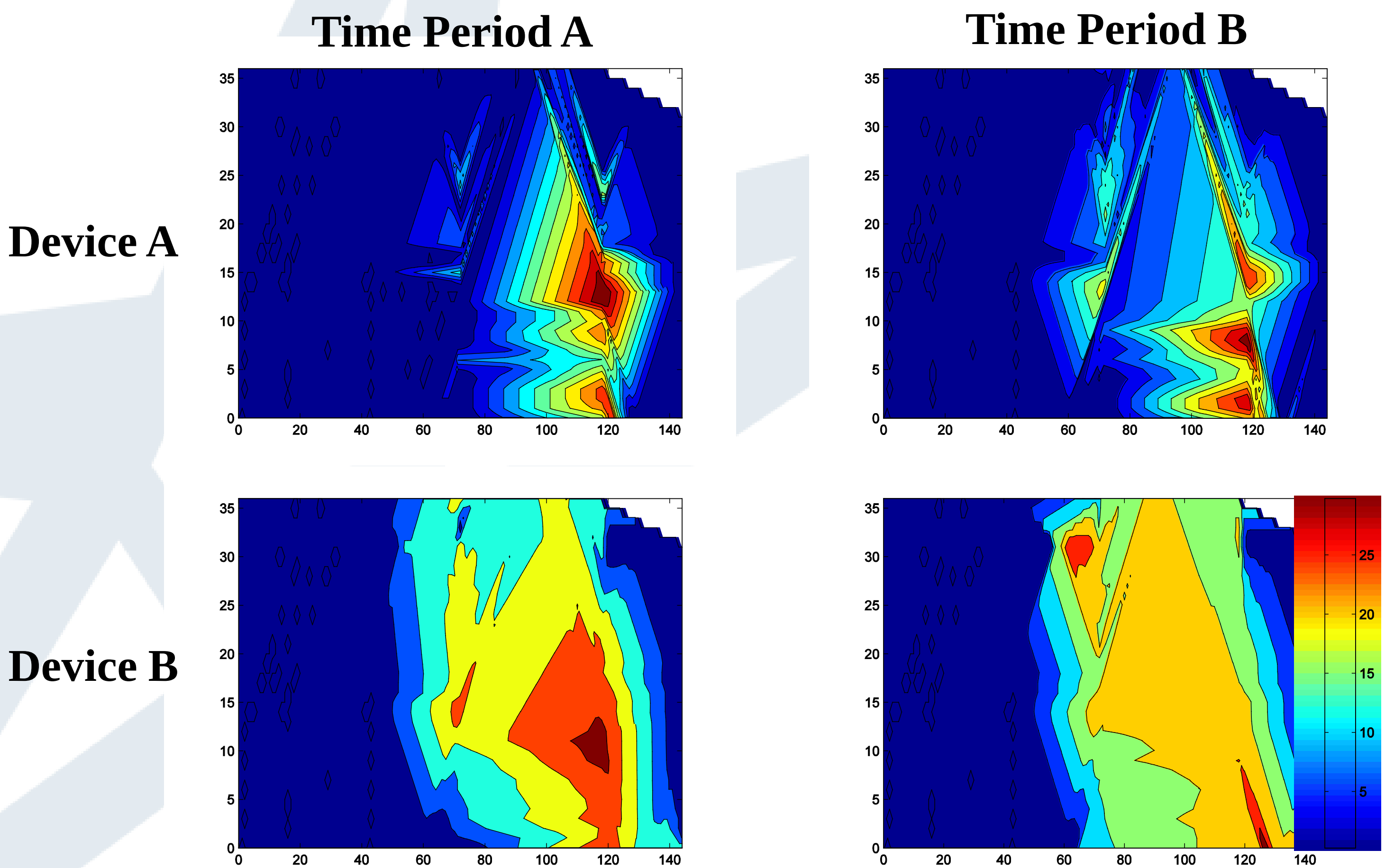
Spectral Feature Alignment (SFA)

Algorithm

- Identify P ***pivot*** features
- Construct a ***bipartite*** graph between the pivot and remaining features.
- Apply ***spectral clustering*** on the graph to derive new features
- Train classifiers on the source using ***augmented*** features (original features + new features)

Feature-based Approaches

Develop general approaches



Feature-based Approaches

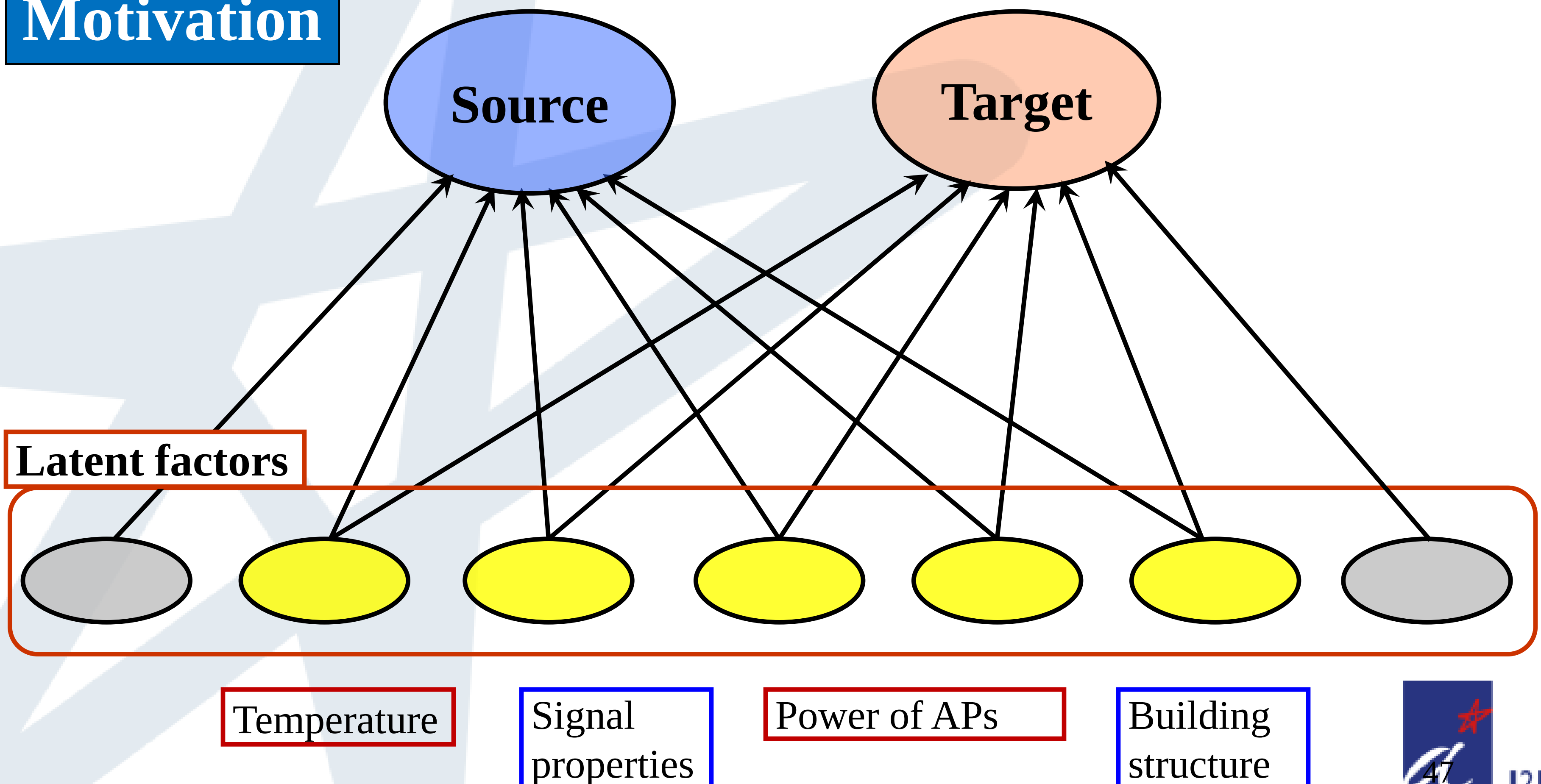
General approaches

- Learning features by minimizing distance between distributions
- Learning features inspired by multi-task learning
- Learning features inspired by self-taught learning

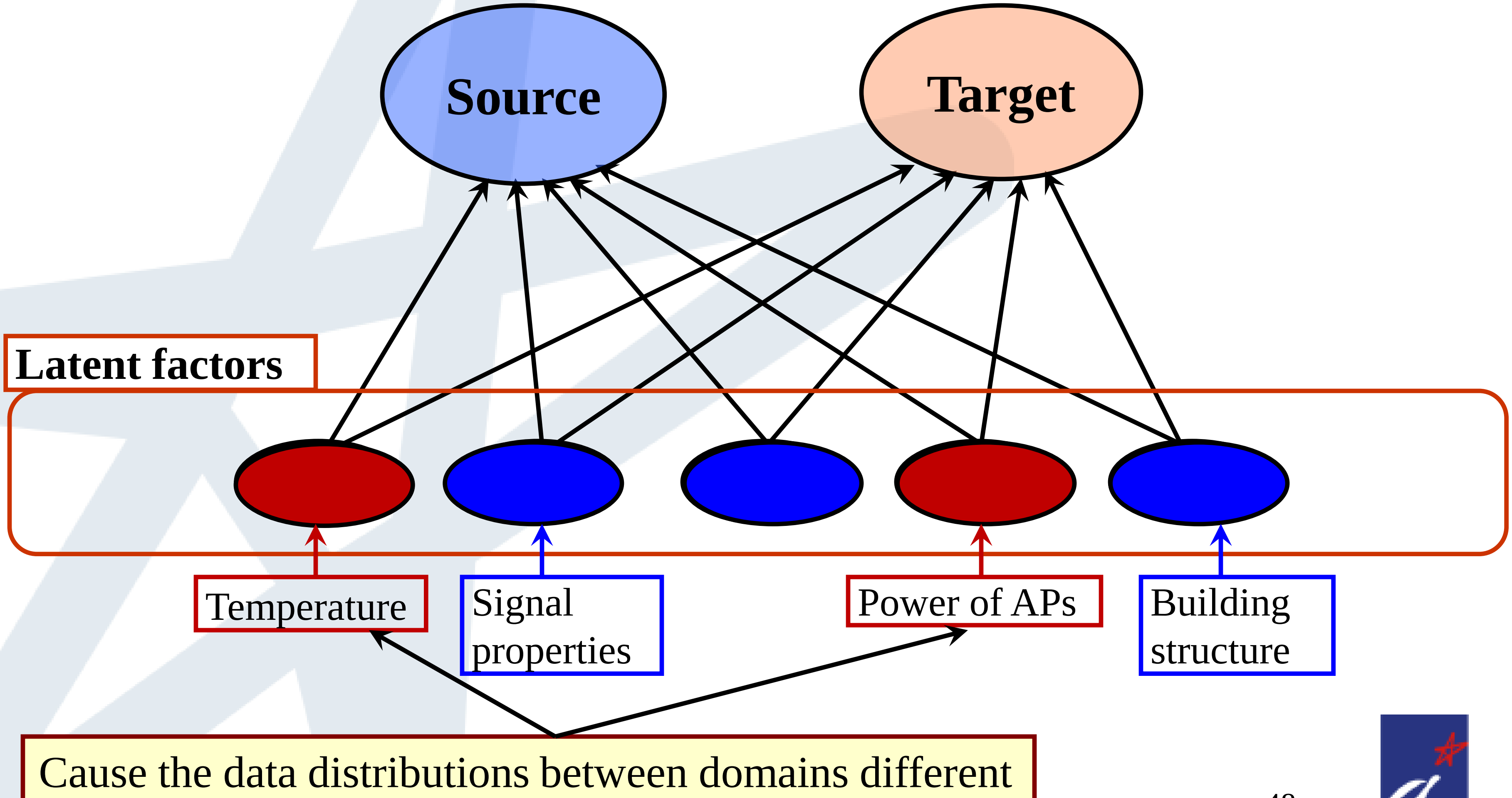
Feature-based Approaches

Transfer Component Analysis [Pan *etal*., IJCAI-09, TNN-11]

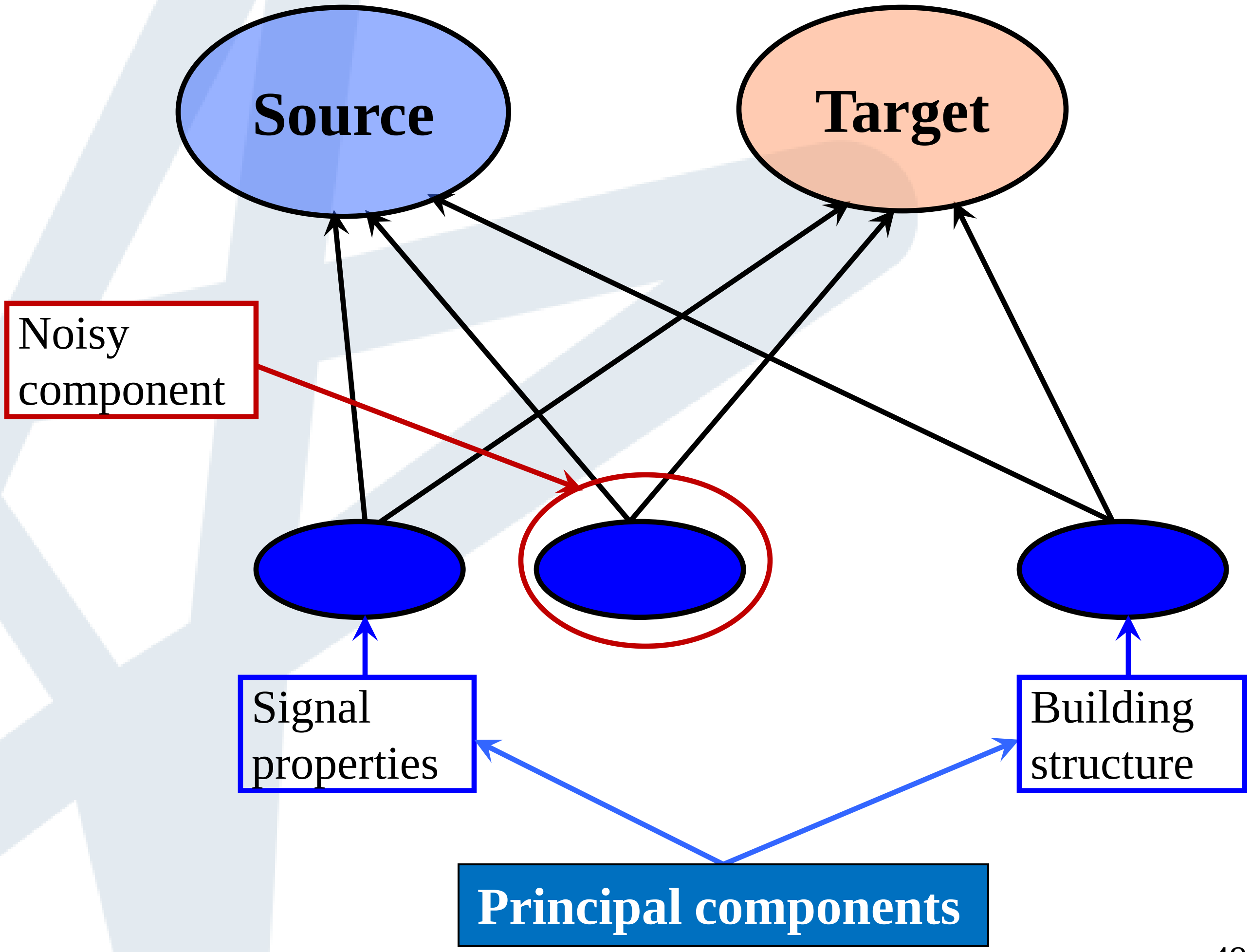
Motivation



Transfer Component Analysis (cont.)

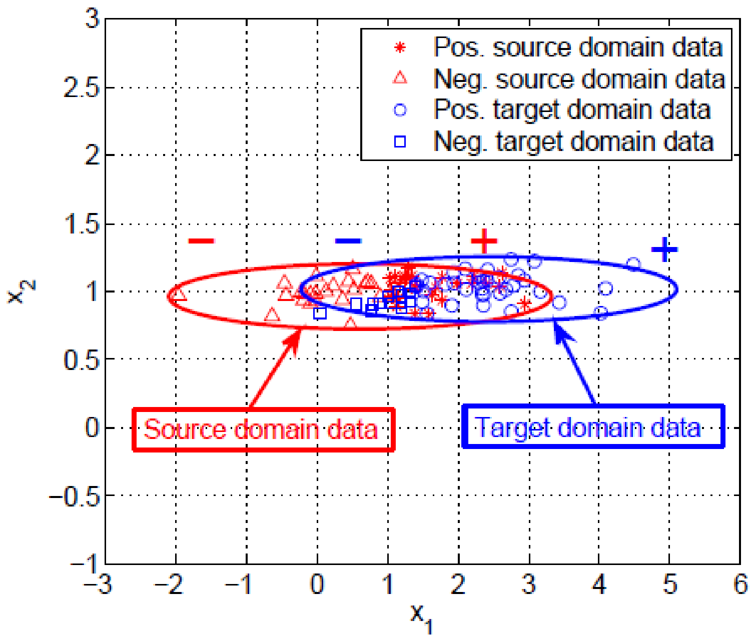


Transfer Component Analysis (cont.)



Transfer Component Analysis (cont.)

Learning φ by only minimizing distance between distributions may map the data onto noisy factors.



Transfer Component Analysis (cont.)

Main idea: the learned φ should map the source and target domain data to the latent space spanned by the factors which can reduce domain difference and preserve original data structure.

High level optimization problem

$$\min_{\varphi} \quad \mathrm{Dist}(\varphi(\mathbf{X}_S), \varphi(\mathbf{X}_T)) + \lambda\Omega(\varphi)$$

$$\text{s.t.} \quad \text{constraints on } \varphi(\mathbf{X}_S) \text{ and } \varphi(\mathbf{X}_T)$$

Transfer Component Analysis (cont.)

Recall: Maximum Mean Discrepancy (MMD)

Given $\mathbf{X}_S = \{x_{S_i}\}_{i=1}^{n_S}$, $\mathbf{X}_T = \{x_{T_i}\}_{i=1}^{n_T}$, drown from $P_S(x)$ and $P_T(x)$, respectively,

$$\mathrm{Dist}(P(X_S), P(X_T)) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \Phi(x_{S_i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \Phi(x_{T_j}) \right\|_{\mathcal{H}}$$

Transfer Component Analysis (cont.)

$$\begin{aligned}\mathrm{Dist}(\varphi(\mathbf{X}_S), \varphi(\mathbf{X}_T)) &= \left\| \mathbb{E}_{x\sim P_T(x)}[\Phi(\varphi(x))] - \mathbb{E}_{x\sim P_S(x)}[\Phi(\varphi(x))] \right\| \\ &\approx \left\| \frac{1}{n_S}\sum_{i=1}^{n_S}\Phi(\varphi(x_{S_i})) - \frac{1}{n_T}\sum_{i=1}^{n_T}\Phi(\varphi(x_{T_i})) \right\|\end{aligned}$$

Assume $\Psi = \Phi \circ \varphi$ a RKHS, with kernel $k(x_i, x_j) = \Psi(x_i)^\top \Psi(x_j)$

$$\mathrm{Dist}(\varphi(\mathbf{X}_S), \varphi(\mathbf{X}_T)) = \mathrm{tr}(KL)$$

$$K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix} \in \mathbb{R}^{(n_S+n_T)\times(n_S+n_T)}, L_{ij} = \begin{cases} \frac{1}{n_S^2} & x_i, x_j \in X_S, \\ \frac{1}{n_T^2} & x_i, x_j \in X_T, \\ -\frac{1}{n_S n_T} & \text{otherwise.} \end{cases}$$

Transfer Component Analysis (cont.)

$$\min_{\varphi} \quad \mathrm{Dist}(\varphi(\mathbf{X}_S), \varphi(\mathbf{X}_T)) + \lambda\Omega(\varphi)$$

$$\text{s.t.} \quad \text{constraints on } \varphi(\mathbf{X}_S) \text{ and } \varphi(\mathbf{X}_T)$$

$$\Downarrow$$

$$\min_{\varphi} \quad \mathrm{tr}(KL) + \lambda\Omega(\varphi)$$

$$\text{s.t.} \quad \text{constraints on } \varphi(\mathbf{X}_S) \text{ and } \varphi(\mathbf{X}_T)$$

- The kernel function can be a highly nonlinear function of φ
- A direct optimization of minimizing the quantity w.r.t. φ can get stuck in poor local minima

Transfer Component Analysis (cont.)

[Pan *etal*., AAAI-08]

Learning φ $\Rightarrow$ (1) learning K

(2) low-dimensional reconstructions of $\mathbf{X}_S$ and $\mathbf{X}_T$ based on K

To minimize the distance between domains

To maximize the data variance

$$\text{Learning } K \Rightarrow \min_{K \succeq 0} \operatorname{tr}(KL) - \lambda \operatorname{tr}(K)$$

To preserve the local geometric structure

$$\text{s.t.} \quad K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2, \ \forall (i,j) \in \mathcal{N},$$

$$K\mathbf{1} = \mathbf{0}, \ K \succeq 0.$$

Low-dimensional constructions of $\mathbf{X}_S, \ \mathbf{X}_T \Rightarrow$ PCA on K

- It is a SDP problem, expensive!
- It is transductive, cannot generalize on unseen instances!
- PCA is post-processed on the learned kernel matrix, which may potentially discard useful information.

Transfer Component Analysis (cont.)

$K = \widetilde{K} W W^\top \widetilde{K}$ where $W \in \mathbb{R}^{(n_S+n_T)\times m}$ and $m \ll n_S + n_T$.

Parametric kernel

Learning K $\Rightarrow$ learning a low-rank matrix W

Regularization term

Minimize distance between domains

$$\min_W \ \mathrm{tr}(W^\top \widetilde{K} L \widetilde{K} W) + \lambda \mathrm{tr}(W^\top W)$$

$$\text{s.t.} \quad W^\top \widetilde{K} H \widetilde{K} W = I$$

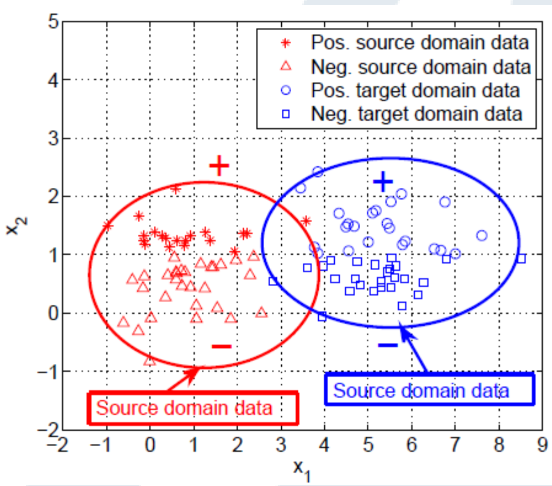
Maximize data variance

$W^* \Leftrightarrow$ m leading eigenvectors of $(\widetilde{K} L \widetilde{K} + \lambda I)^{-1} \widetilde{K} H \widetilde{K}$

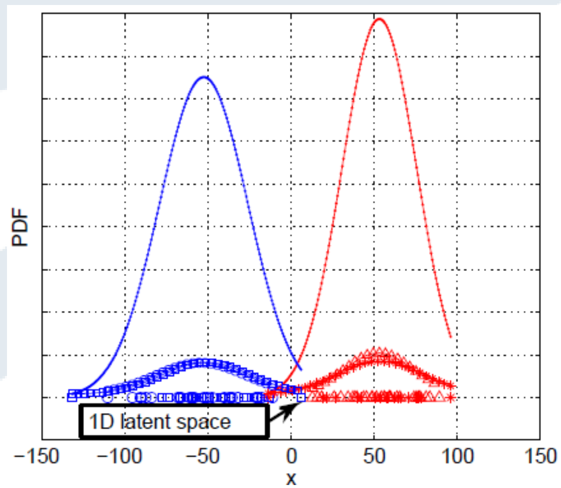
Transfer Component Analysis (cont.)

An illustrative example

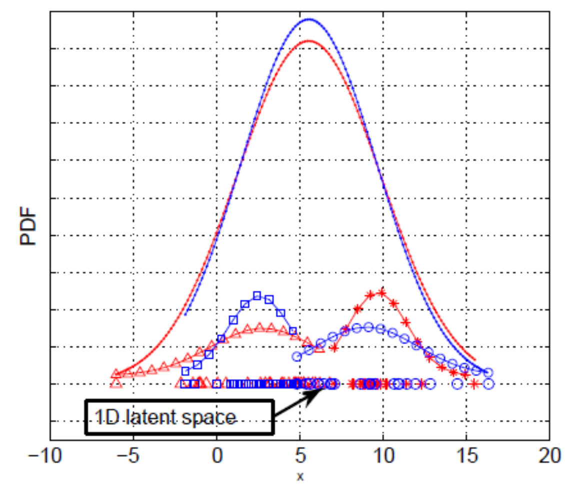
Latent features learned by **PCA** *and* **TCA**



Original feature space



PCA



TCA

Feature-based Approaches

Multi-task Feature Learning

General Multi-task Learning Setting

Given $\mathrm{D}_S = \{x_{S_i}, y_{S_i}\}_{i=1}^{n_S}$, $\mathrm{D}_T = \{x_{T_i}, y_{T_i}\}_{i=1}^{n_T}$,

where n_S and n_T are small,

Learn f_S, f_T, s.t. $\sum_{t \in \{S,T\}} \sum_i \epsilon(f_t(x_{t_i}), y_{t_i})$ is small.

- **Assumption:** If tasks are related, they should share some **good** common features.
- **Goal:** Learn a low-dimensional representation shared across related tasks.

Feature-based Approaches

Multi-task Feature Learning (cont.)

Assume $f(x) = \langle \theta, (U^\top x) \rangle = \theta^\top (U^\top x)$, where $\theta \in \mathbb{R}^k, x \in \mathbb{R}^m, U \in \mathbb{R}^{m \times k}$

$$\{\Theta^*, U^*\} = \arg\min \sum_{t \in \{S,T\}} \sum_{i=1}^{n_t} l(U^\top x_{t_i}, y_{t_i}, \theta_t) + \lambda_1 \Omega(\Theta)$$

s.t. constraints on U.

$\Theta = [\theta_S,\ \theta_T] \in \mathbb{R}^{k \times 2}$

- U is full rank ($U \in \mathbb{R}^{m \times k}, k = m$), Θ is sparse. [Argyriou *etal.*, NIPS-07]
- U is low rank ($U \in \mathbb{R}^{m \times k}, k \ll m$). [Ando and Zhang, JMLR-05] [Ji *etal*, KDD-08]

Feature-based Approaches

Self-taught Feature Learning

- **Intuition:** There exist some higher-level features that can help the target learning task even only a few labeled data are given.
- **Steps:**

1) Learn higher-level features from a lot of unlabeled data.
2) Use the learned higher-level features to represent the data of the target task.
3) Training models from the new representations of the target task with corresponding labels.

Feature-based Approaches

Self-taught Feature Learning

- **How to learn higher-level features**
  - Sparse Coding [Raina etal., 2007]
  - Deep learning [Glorot *etal*., 2011]

Parameter-based Transfer Learning Approaches

Assume $f(x) = \langle \theta, x \rangle = \theta^\top x = \sum_{i=1}^{m} \theta_i x_i,$ where $\theta, x \in \mathbb{R}^m$.

$$\begin{aligned} \theta_S^* &= \arg\min \sum_{i=1}^{n_S} l(x_{S_i}, y_{S_i}, \theta_S) + \lambda\Omega(\theta_S), \\ \theta_T^* &= \arg\min \sum_{i=1}^{n_T} l(x_{T_i}, y_{T_i}, \theta_T) + \lambda\Omega(\theta_T) \end{aligned}$$

Tasks are learned independently

Motivation: A well-trained model θ_S^* has learned a lot of structure. If two tasks are related, this structure can be transferred to learn θ_T^* .

Parameter-based Approaches

Multi-task Parameter Learning

Assumption:

If tasks are related, they may share similar parameter vectors.

For example, [Evgeniou and Pontil, KDD-04]

Common part

Specific part for individual task

$$\theta_S = \theta_0 + v_S$$

$$\theta_T = \theta_0 + v_T$$

$$\{\theta_S^*,\ \theta_T^*\} = \arg\min \sum_{t \in \{S,T\}} \sum_{i=1}^{n_t} l(x_{t_i}, y_{t_i}, \theta_t) + \lambda \Omega(\theta_0, v_S, v_T)$$

Parameter-based Approaches

Multi-task Parameter Learning (cont.)

A general framework:

Denote $\Theta = [\theta_S,\ \theta_T]$,

$$f(\Theta) = \sum_{t \in \{S,T\}} \left\| \theta_t - \frac{1}{2} \sum_{s \in \{S,T\}} \theta_s \right\|^2$$

$$\Theta^* = \arg\min \sum_{t \in \{S,T\}} \sum_{i=1}^{n_t} l(x_{t_i}, y_{t_i}, \theta_t) + \lambda_1 \mathrm{tr}(\Theta^\top \Theta) + \lambda_2 f(\Theta)$$

$$\sum_{t \in \{S,T\}} \|\theta_t\|^2$$

[Zhang and Yeung, UAI-10]

$$f(\Theta) = \mathrm{tr}(\Theta^\top \Sigma^{-1} \Theta)$$

$$\text{s.t.} \quad \Sigma \succeq 0 \text{ and } \mathrm{tr}(\Sigma) = 1.$$

[Agarwal *etal*, NIPS-10]

$$f(\Theta) = \sum_{t \in \{S,T\}} \left\| \theta_t - \widetilde{\theta}_t^{\mathcal{M}} \right\|^2$$

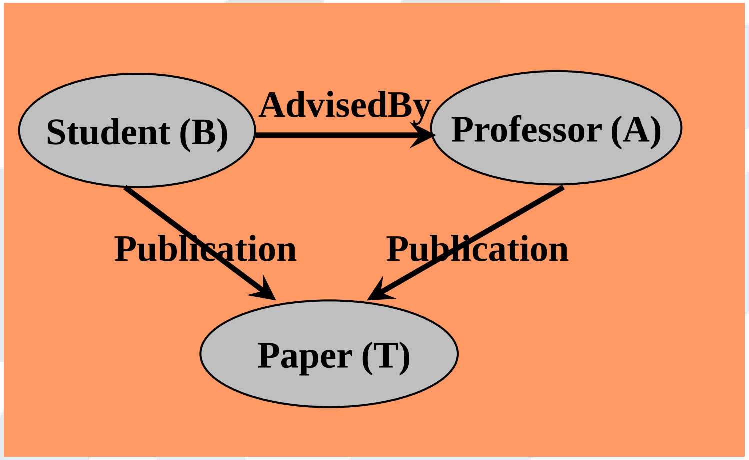
Relational Transfer Learning Approaches

- **Motivation:** If two relational domains (data is non-i.i.d) are related, they may share some similar relations among objects. These relations can be used for knowledge transfer across domains.

Relational Transfer Learning Approaches (cont.)

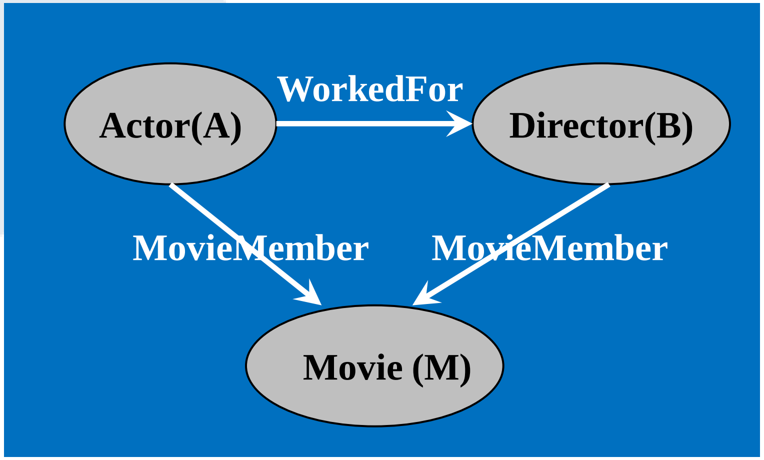
[Mihalkova *etal*., AAAI-07, Davis and Domingos, ICML-09]

Academic domain (source)



AdvisedBy (B, A) ∧ Publication (B, T) => Publication (A, T)

Movie domain (target)



WorkedFor (A, B) ∧ MovieMember (A, M) => MovieMember (B, M)

P1(x, y) ∧ P2 (x, z) => P2 (y, z)

Relational Approaches

Relational Adaptive bootstraPping [Li *etal*., ACL-12]

Task: sentiment summarization

- What is the opinion expressed on?
  - To construct lexicon of ***topic*** or ***target*** words
- How is the opinion expressed?
  - To construct lexicon of ***sentiment*** words

Sentiment lexicon (camera)

great, amazing, light recommend, excellent, etc.
artifacts, noise, never but, boring, etc.

Topic lexicon (camera)

camera, product, screen, photo, size, weight, quality, price, memory, etc.

Relational Approaches

Relational Adaptive bootstraPping (RAP) (cont.)

Reviews on cameras

The **camera** is **great**.
It is a very **amazing product**.
I highly **recommend** this **camera**.
Photos had some **artifacts** and **noise**.

Reviews on movies

This **movie** has **good script**, **great casting**, **excellent acting**.
This **movie** is so **boring**.
The **Godfather** was the most **amazing movie**.
The **movie** is **excellent**.

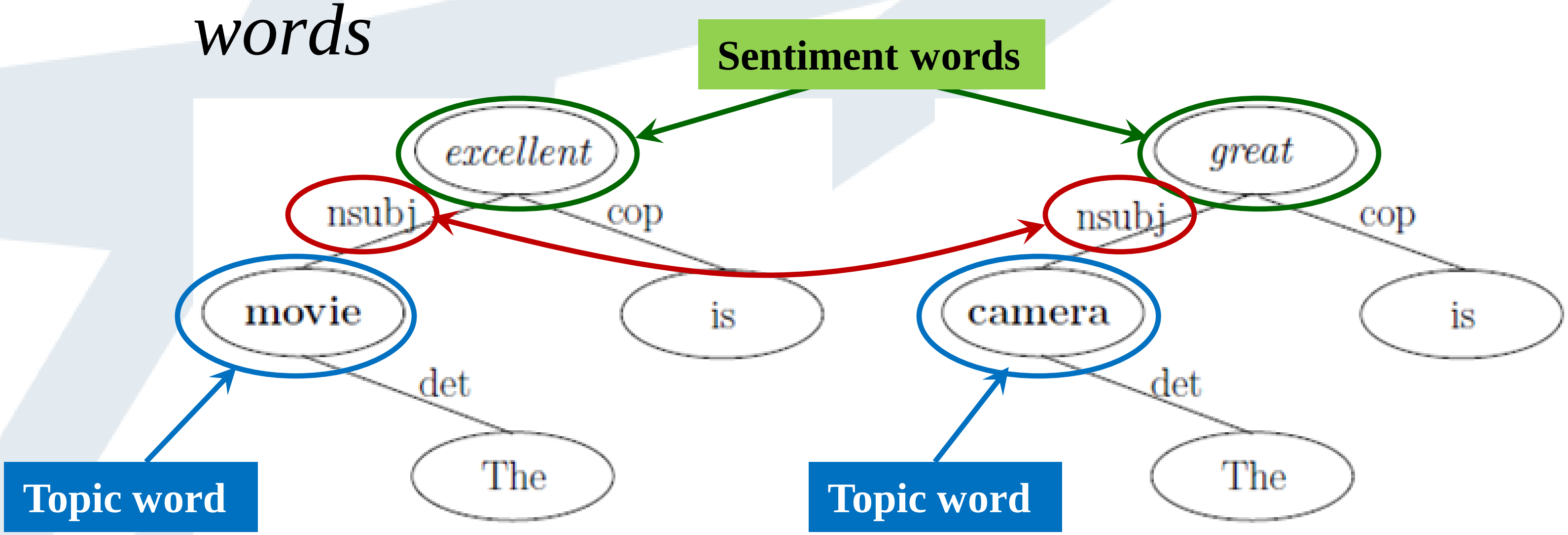
Relational Approaches
RAP (cont.)

- ➢Bridge between cross-domain sentiment words
  - – *Domain independent (general) sentiment words*

- ➢ Bridge between cross-domain topic words

Relational Approaches
RAP (cont.)

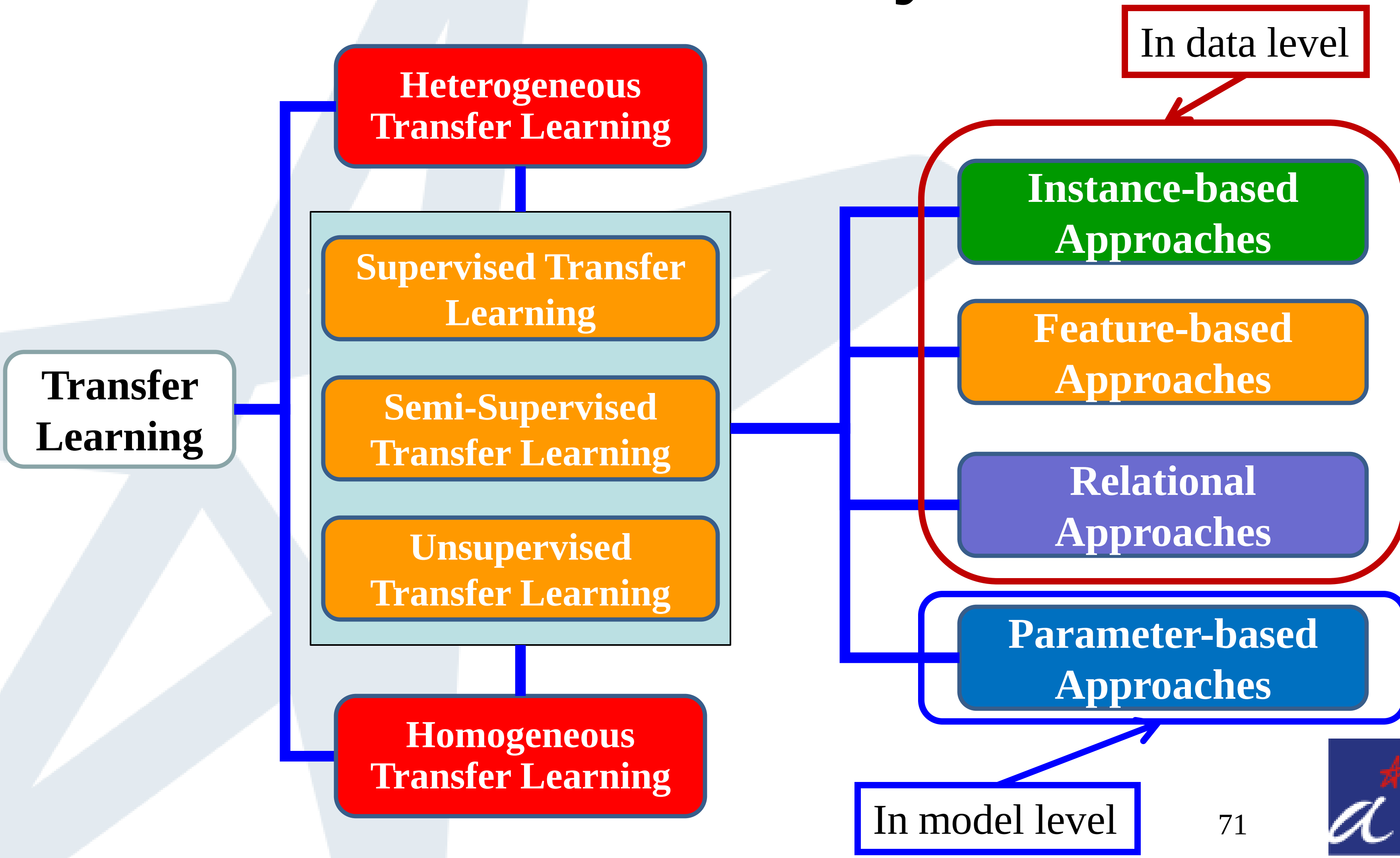
- Bridge between cross-domain topic words
  - *Syntactic structure between topic and sentiment words*



(b) Movie domain. (a) Camera domain.

Common syntactic pattern: *"topic word"* – ***nsubj*** – "sentiment word"

Summary


In data level
Heterogeneous Transfer Learning
Transfer Learning
Supervised Transfer Learning
Semi-Supervised Transfer Learning
Unsupervised Transfer Learning
Homogeneous Transfer Learning
Instance-based Approaches
Feature-based Approaches
Relational Approaches
Parameter-based Approaches
In model level

Some Advanced Research Issues in Transfer Learning

- How to transfer knowledge across heterogeneous feature spaces
- Active learning meets transfer learning
- Transfer learning from multiple sources

Reference

- [Thorndike and Woodworth, The Influence of Improvement in one mental function upon the efficiency of the other functions, 1901]
- [Taylor and Stone, Transfer Learning for Reinforcement Learning Domains: A Survey, JMLR 2009]
- [Pan and Yang, A Survey on Transfer Learning, IEEE TKDE 2009]
- [Quionero-Candela, *etal*, Data Shift in Machine Learning, MIT Press 2009]
- [Biltzer *etal*.. Domain Adaptation with Structural Correspondence Learning, *EMNLP* 2006]
- [Pan *etal*., Cross-Domain Sentiment Classification via Spectral Feature Alignment, WWW 2010]
- [Pan *etal*., Transfer Learning via Dimensionality Reduction, AAAI 2008]

Reference (cont.)

- [Pan *etal.*, Domain Adaptation via Transfer Component Analysis, IJCAI 2009]
- [Evgeniou and Pontil, Regularized Multi-Task Learning, KDD 2004]
- [Zhang and Yeung, A Convex Formulation for Learning Task Relationships in Multi-Task Learning, UAI 2010]
- [Agarwal *etal*, Learning Multiple Tasks using Manifold Regularization, NIPS 2010]
- [Argyriou *etal.*, Multi-Task Feature Learning, NIPS 2007]
- [Ando and Zhang, A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data, JMLR 2005]
- [Ji *etal*, Extracting Shared Subspace for Multi-label Classification, KDD 2008]

Reference (cont.)

- [Raina *etal.*, Self-taught Learning: Transfer Learning from Unlabeled Data, ICML 2007]
- [Dai *etal.*, Boosting for Transfer Learning, ICML 2007]
- [Glorot *etal.*, Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach, ICML 2011]
- [Davis and Domingos, Deep Transfer vis Second-order Markov Logic, ICML 2009]
- [Mihalkova *etal.*, Mapping and Revising Markov Logic Networks for Transfer Learning, AAAI 2007]
- [Li *etal.*, Cross-Domain Co-Extraction of Sentiment and Topic Lexicons, ACL 2012]

Reference (cont.)

- [Sugiyama *etal*., Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation, NIPS 2007]
- [Kanamori *etal*., A Least-squares Approach to Direct Importance Estimation, JMLR 2009]
- [Cristianini *etal*., On Kernel Target Alignment, NIPS 2002]
- [Huang *etal*., Correcting Sample Selection Bias by Unlabeled Data, NIPS 2006]
- [Zadrozny, Learning and Evaluating Classifiers under Sample Selection Bias, ICML 2004]

Thank You

Selected Applications of Transfer Learning

Qiang Yang and Sinno J. Pan

2013 PAKDD Tutorial

Brisbane, Australia

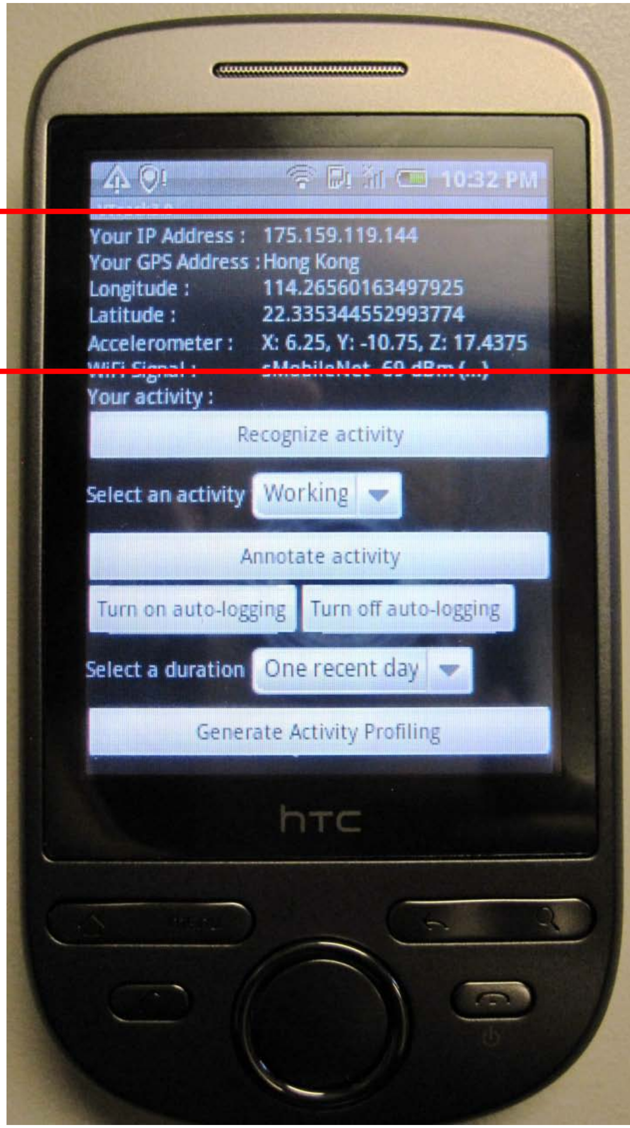
Part I. Cross Domain Transfer Learning for Activity Recognition

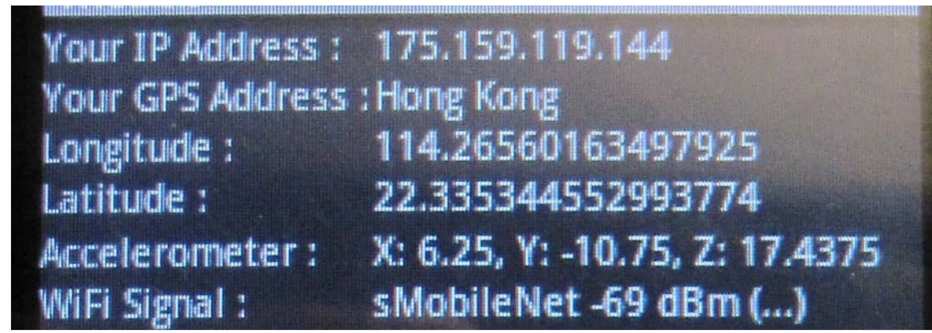
- Vincent W. Zheng, Derek H. Hu and Qiang Yang. Cross-Domain Activity Recognition. In *Proceedings of the 11th International Conference on Ubiquitous Computing* (**Ubicomp-09**), Orlando, Florida, USA, Sept.30-Oct.3, 2009.
- Derek Hao Hu, Qiang Yang. Transfer Learning for Activity Recognition via Sensor Mapping. *In Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (IJCAI-11), Barcelona, Spain, July 2011

Demo

- [Annotation]()

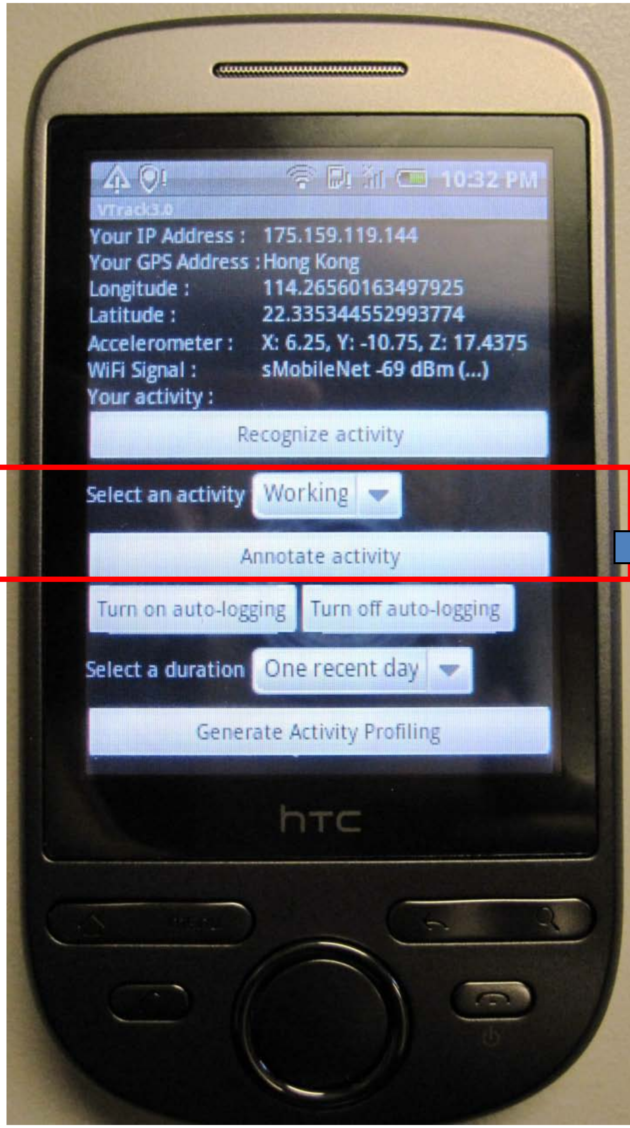
eHealth Demo

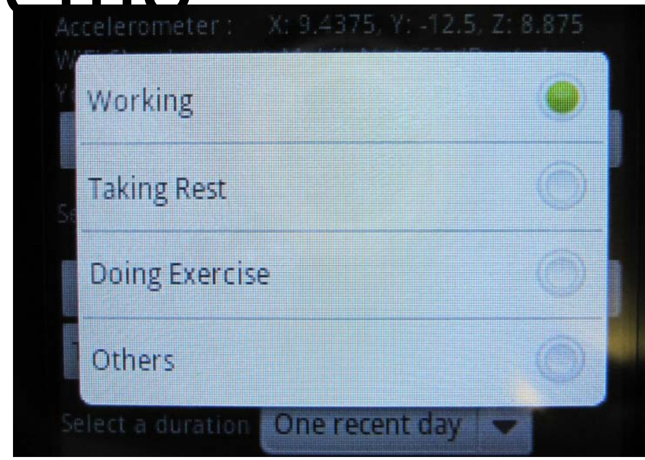


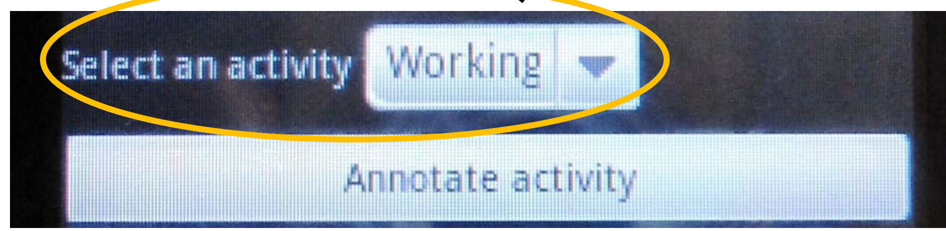


Sensor data

eHealth demo

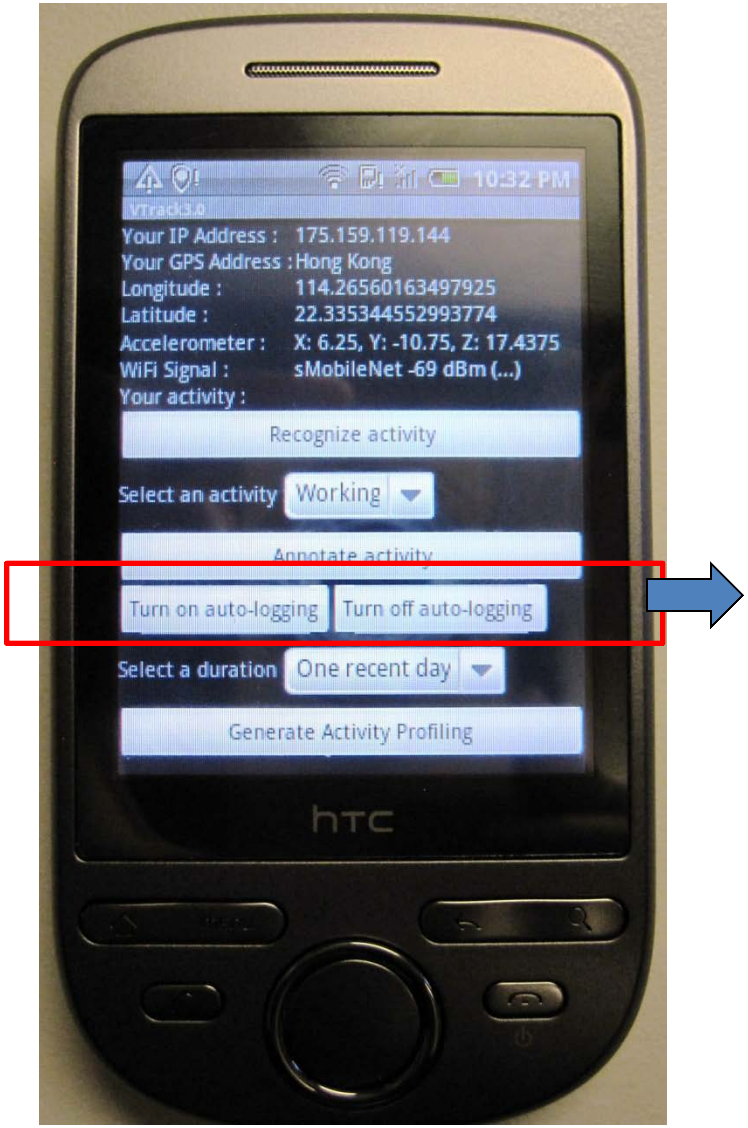


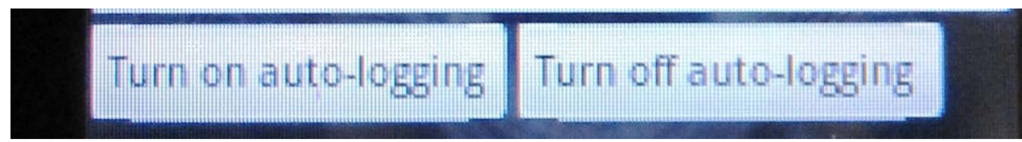




Activity annotation

eHealth demo



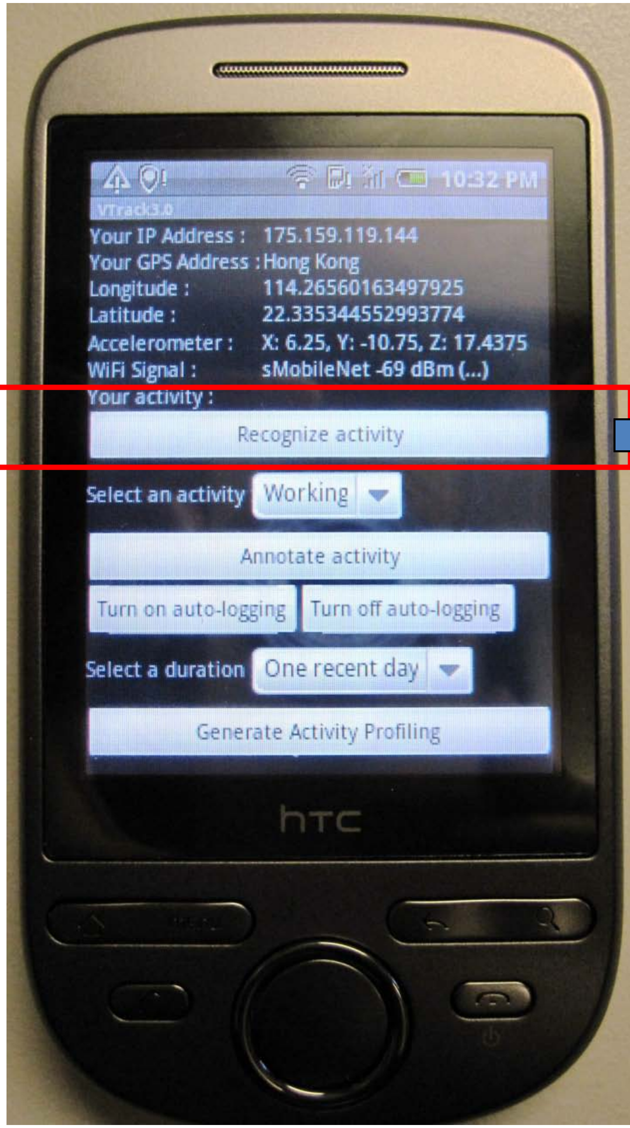


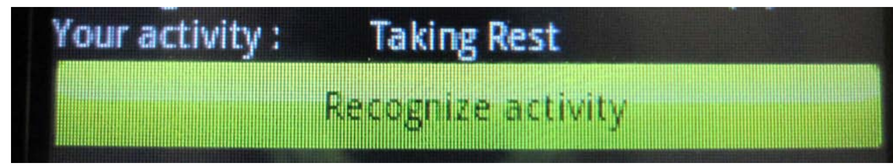
Auto logging / activity recognition
(service in background)

Demo

- Recognition

eHealth demo



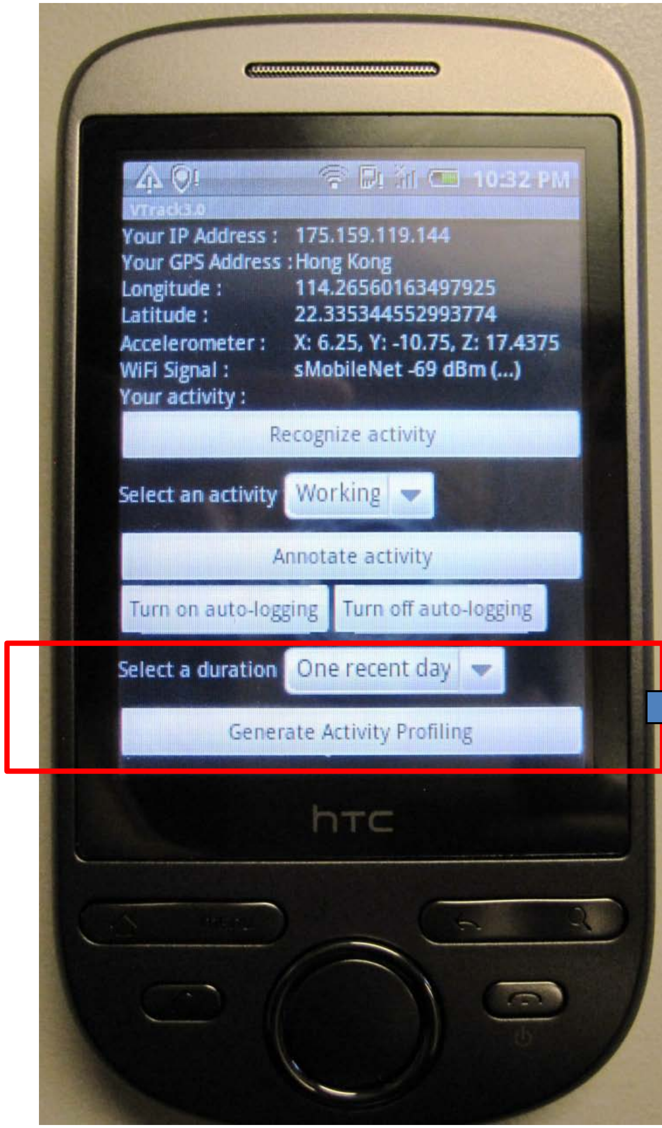


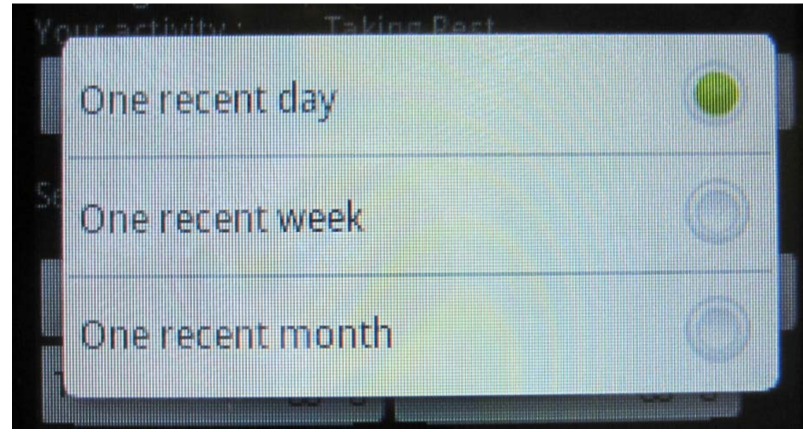
Real-time activity recognition

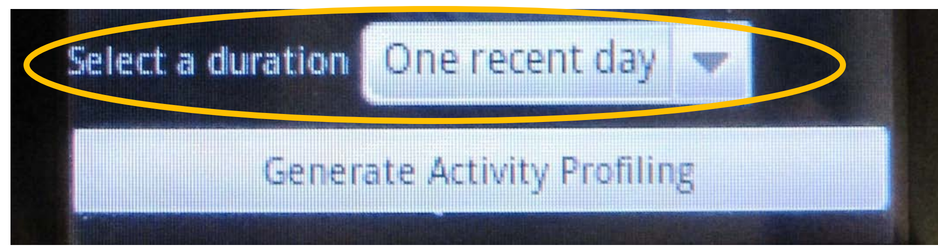
Demo

- Profiling

eHealth demo

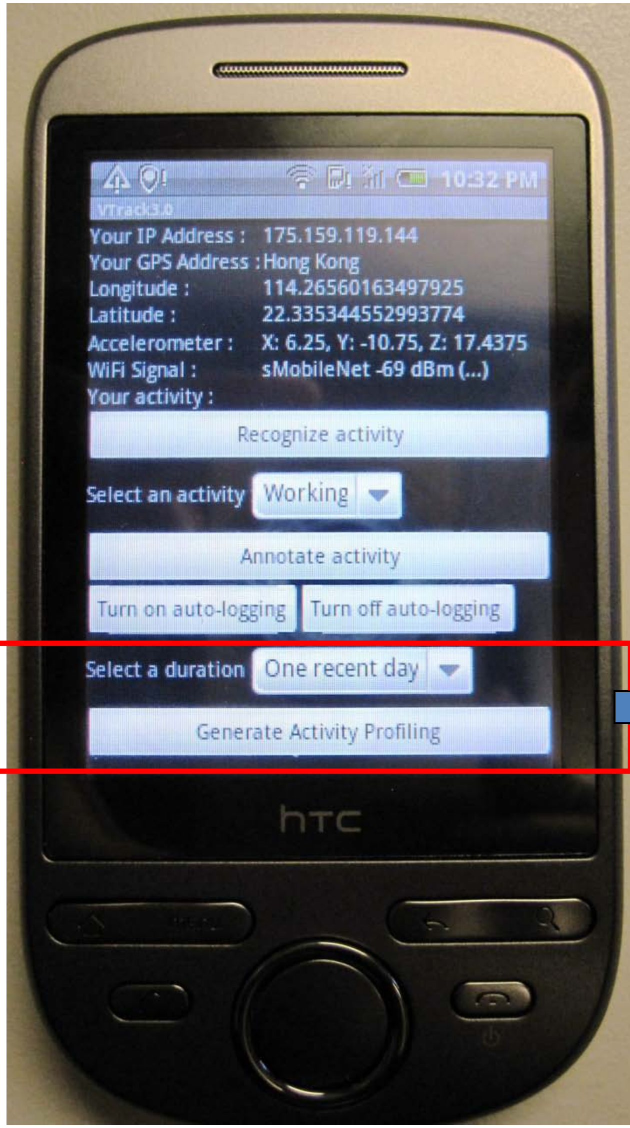


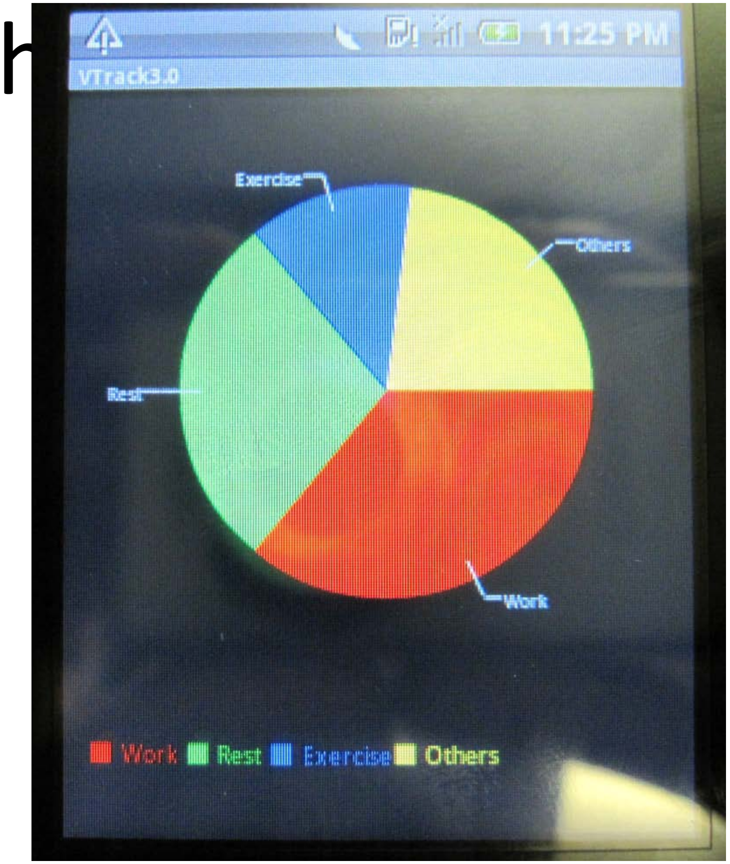


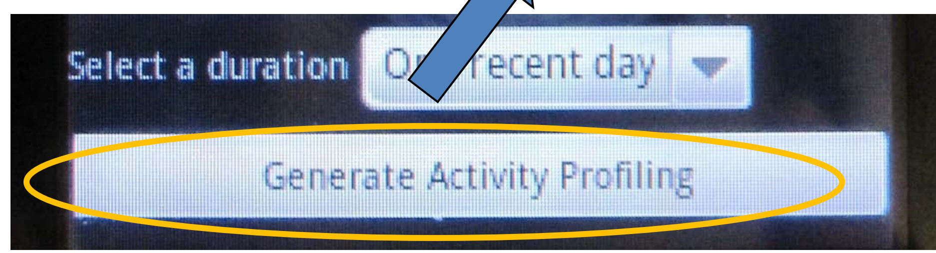


Activity profiling

eHealth

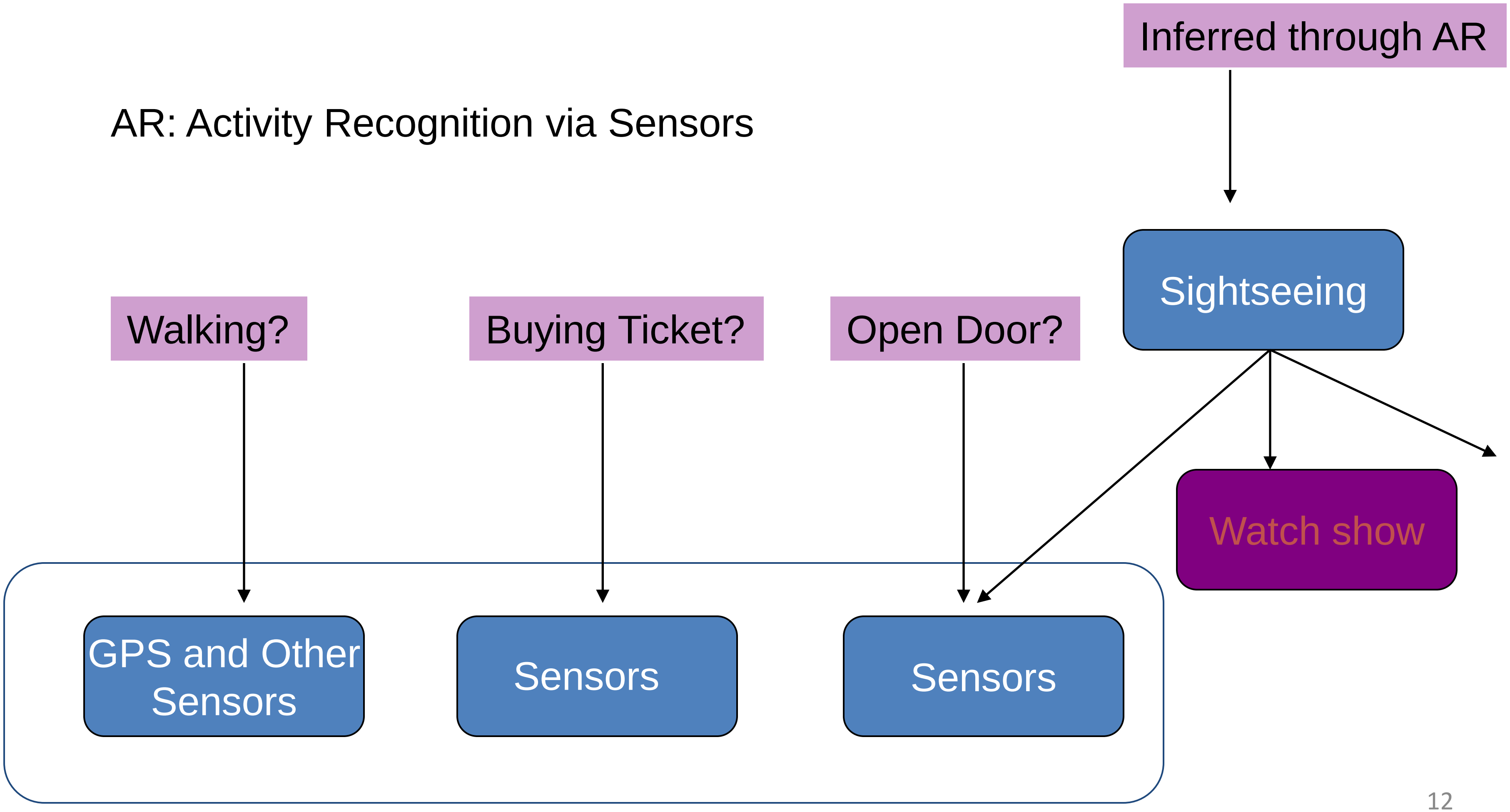






Activity profiling for health management

Key Problem: Recognizing Actions and Context (Locations)



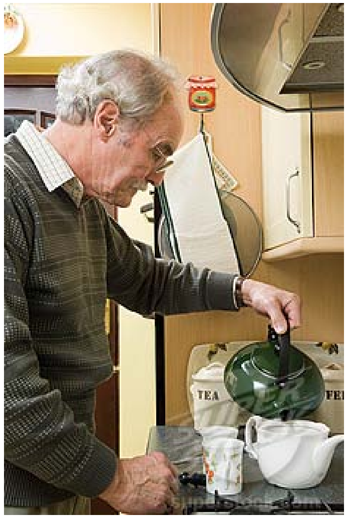
1. Cross-Domain Activity Recognition

[Zheng, Hu, Yang: UbiComp-2009, PCM-2011]

- Challenge:
  - – Some activities without data (partially labeled)
- Cross-domain activity recognition
  - – Use other activities with available labeled data

Making coffee

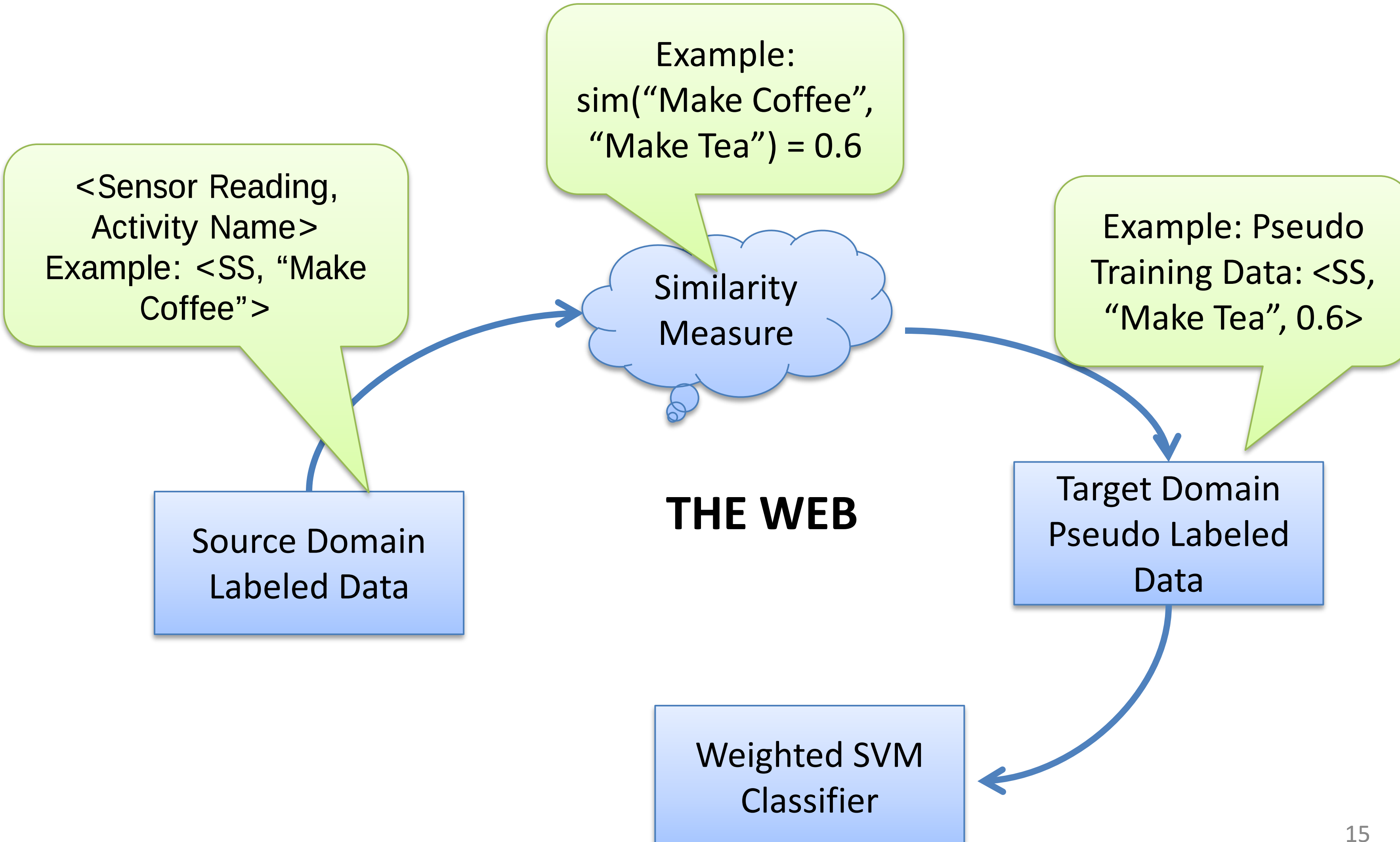
- Happen in kitchen
- Use cup, pot
- …



Making tea

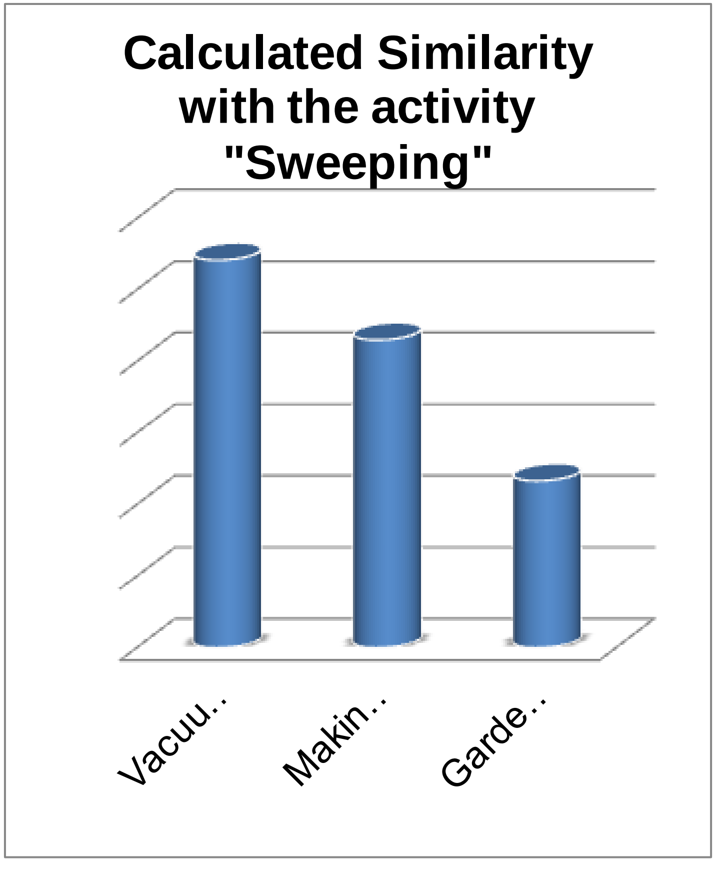

Cleaning Indoor
Source Domain
Sweeping
Swiffering
Mopping
Vacuuming
Dusting
Sweeping
Making-the-bed
Putting-things-away
Organizing
Disposing-Garbage
Taking-out-trash
Dealing-with-Garbage
Cleaning-a-surface
Scrubbing
Cleaning-a-surface
Cleaning miscellaneous
Cleaning-miscellaneous
Cleaning-background
Cleaning-background
Cleaning Indoor
Activity Transfer
Gardening
Mowing lawn
Gardening
Yardwork
Target Domain 1
Yardwork-miscellaneous
Yardwork-miscellaneous
Laundry
Washing-laundry
Drying-laundry
Washing/Drying-laundry
Washing-laundry-background
Drying-laundry-background
Washing/Drying-laundry-background
Folding-laundry
Putting-away-laundry
Ironing
Dealing-with-clothes
Laundry
Laundry-miscellaneous
Laundry-miscellaneous
Hand-washing-dishes
Drying-dishes
Putting-away-dishes
Dealing-with-dishes
Target Domain 2
Loading-dishwasher
Unloading-dishwasher
Loading/unloading-dishwasher
Dishwashing
Dishwashing-miscellaneous
Dishwashing-miscellaneous
Dishwashing

System Workflow



Calculating Activity Similarities

- How similar are two activities?
  - Use Web search results
  - TFIDF: Traditional IR similarity metrics (cosine similarity)
  - Example
    - Mined similarity between the activity “sweeping” and “vacuuming”, “making the bed”, “gardening”



Datasets: MIT PlaceLab

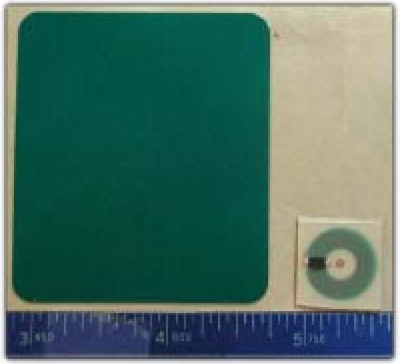
http://architecture.mit.edu/house_n/placelab.html

- MIT PlaceLab Dataset (PLIA2) [Intille et al. Pervasive 2005]
- Activities: Common household activities

Datasets: Intel Research Lab

- Intel Research Lab [Patterson, Fox, Kautz, Philipose, ISWC2005]
  - Activities Performed: 11 activities
  - Sensors
    - RFID Readers & Tags
  - Length:
    - 10 mornings

1	Using the bathroom
2	Making oatmeal
3	Making soft-boiled eggs
4	Preparing orange juice
5	Making coffee
6	Making tea
7	Making or answering a phone call
8	Taking out the trash
9	Setting the table
10	Eating breakfast
11	Clearing the table



Picture excerpted from [Patterson, Fox, Kautz, Philipose, ISWC2005].

Cross-Domain AR: Performance

	Accuracy with Cross Domain Transfer	# Activities (Source Domain)	# Activities (Target Domain)	Baseline (Random Guess)	Supervised (Upper bound)
Intel Research Lab Dataset	63.2%	5	6	16.7%	78.3%
Amsterdam Dataset	65.8%	4	3	33.3%	72.3%
MIT Dataset (Cleaning to Laundry)	58.9%	13	8	12.5%	-
MIT Dataset (Cleaning to Dishwashing)	53.2%	13	7	14.3%	-

- **Activities in the source domain and the target domain are generated from ten random trials, mean accuracies are reported.**

Derek Hao Hu and Qiang Yang, IJCAI 2011

Transferring Across Feature Space

+

Transferring Across Label Space

→

Transfer from Source Domain to Target Domain

$$p(y_t \mid x_t) = \sum_{c^{(i)} \in \mathbf{L}_s} p(c \mid x_t) \cdot p(y_t \mid c)$$

Proposed Approach

- Final goal: Estimate $p(\mathbf{y}_t \mid \mathbf{x}_t)$
  - We ha $p(\mathbf{y_t}|\mathbf{x_t}) = \sum_{\mathbf{c}^{(i)} \in \mathcal{L}_s} p(\mathbf{c}|\mathbf{x_t}) \cdot p(\mathbf{y_t}|\mathbf{c})$
  - $p(\mathbf{y_t}|\mathbf{x_t}) \approx p(\hat{\mathbf{c}}|\mathbf{x_t}) \cdot p(\mathbf{y_t}|\hat{\mathbf{c}}) \quad (\hat{\mathbf{c}} = \arg\max_{\mathbf{c} \in \mathcal{L}_s} p(\mathbf{c}|\mathbf{x_t}))$ e:

Feature Transfer

Label Transfer

Experiments

- Datasets
  - UvA dataset [van Kasteren et al. Ubicomp 2008]
  - MIT Placelab (PLIA1) dataset [Intille et al. Ubicomp 2006]
  - Intel Research Lab dataset [Patterson et al. ISWC 2005]
- Baseline
  - Unsupervised Activity Recognition Algorithm [Wyatt et al. 2005]
- Different sensors for different datasets

State-based sensors for UvA dataset

A series of different wired sensors for MIT dataset

RFID sensor for Intel Research Lab Dataset

Experiments: Different Feature & Label Spaces

K	MIT → UvA Acc(Var)
K = 5	**59.8% (4.2%)**
K = 10	57.5% (4.1%)
K = 15	51.0% (4.8%)
K = 20	41.0% (4.1%)
Unsupervised	47.3%(4.1%)

Table 3: Algorithm performance of transferring knowledge from MIT PLIA1 to UvA dataset

K	MIT → Intel Acc(Var)
K = 5	60.5% (4.2%)
K = 10	**61.2% (3.8%)**
K = 15	53.2% (4.1%)
K = 20	42.0% (2.5%)
Unsupervised	42.8%(3.8%)

Table 4: Algorithm performance of transferring knowledge from MIT PLIA1 to Intel dataset

- Source: MIT PLIA1 dataset
  Target: UvA (Intel) datasets

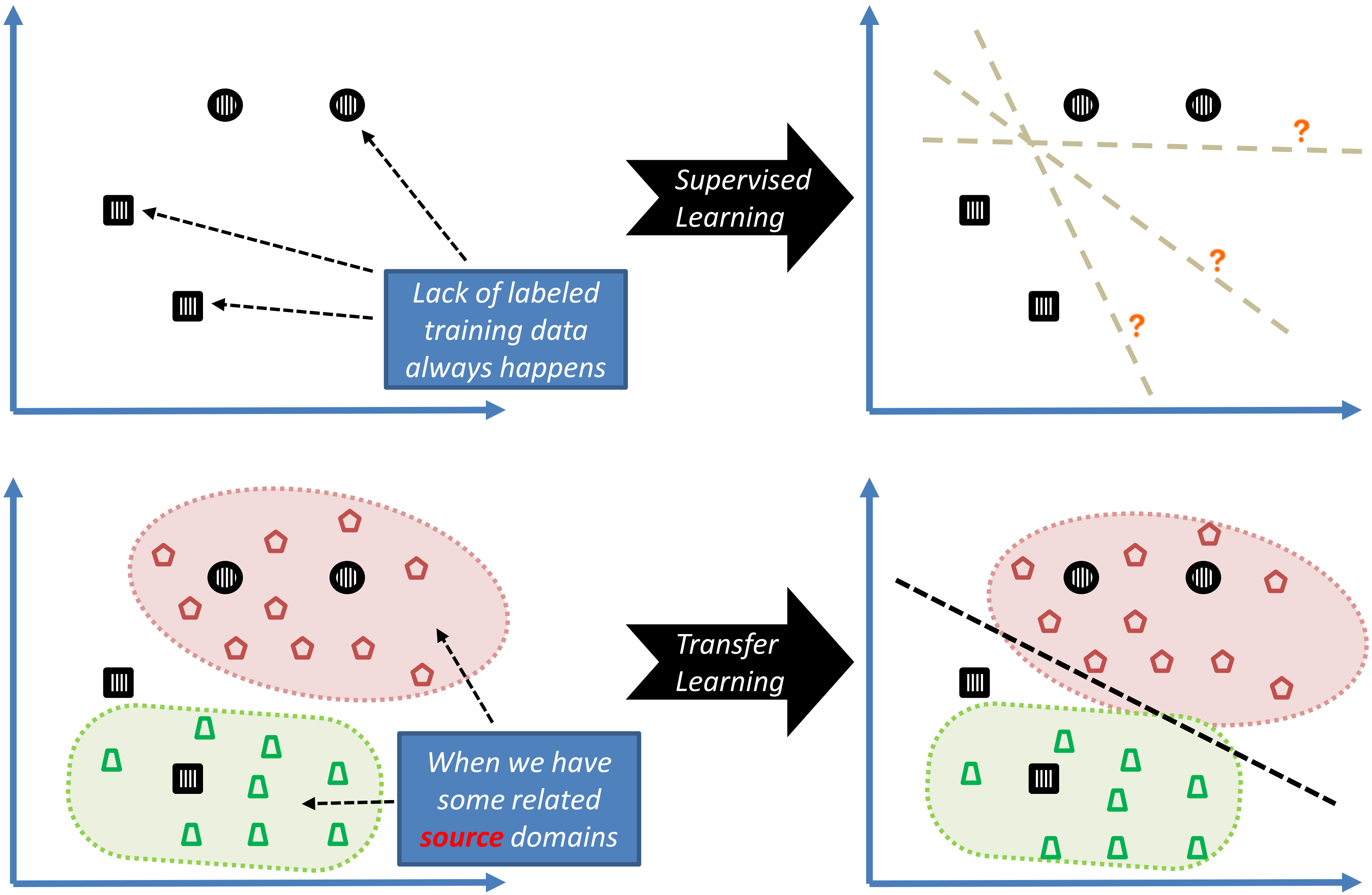
Part II

- Source Free Transfer Learning
- Evan Wei Xiang, Sinno Jialin Pan, Weike Pan, Jian Su and Qiang Yang. Source-Selection-Free Transfer Learning. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11), Barcelona, Spain, July 2011.

Source-Selection *free* Transfer Learning

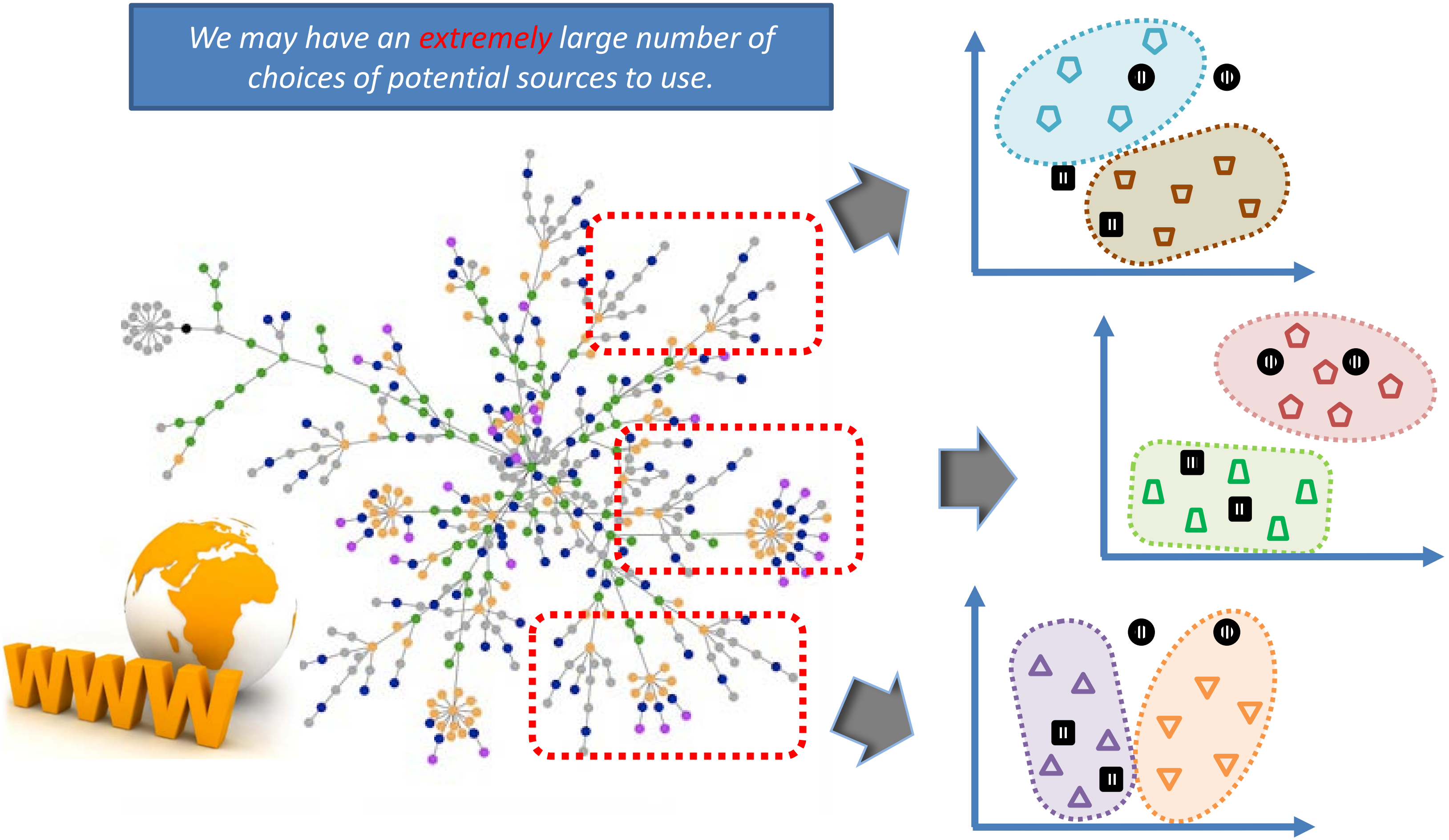
Evan Xiang, Sinno Pan, Weike Pan,
Jian Su, Qiang Yang

Transfer Learning



Where are the "right" source data?

We may have an extremely large number of choices of potential sources to use.



Outline of Source-Selection-Free Transfer Learning (SSFTL)

- ***Stage 1: Building base models***

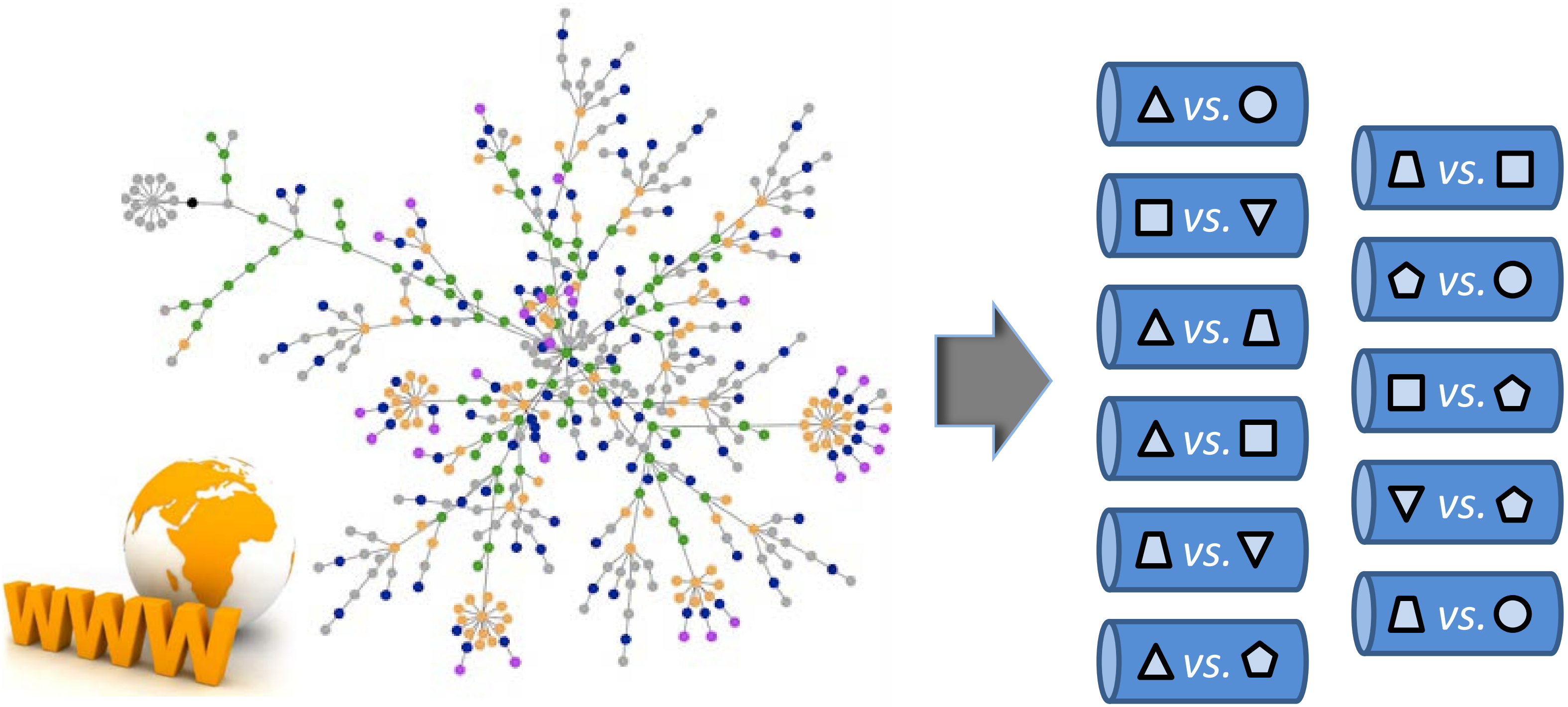
- ***Stage 2: Label Bridging via Laplacian Graph Embedding***

- ***Stage 3: Mapping the target instance using the base classifiers & the projection matrix***

- ***Stage 4: Learning a matrix W to directly project the target instance to the latent space***

- ***Stage 5: Making predictions for the incoming test data using W***

SSFTL – Building base models

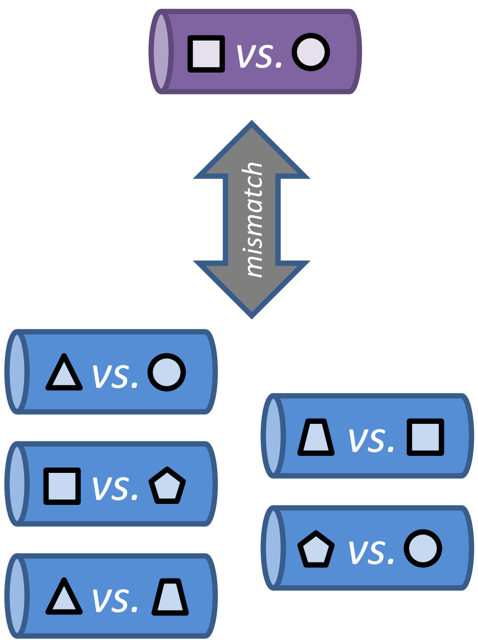


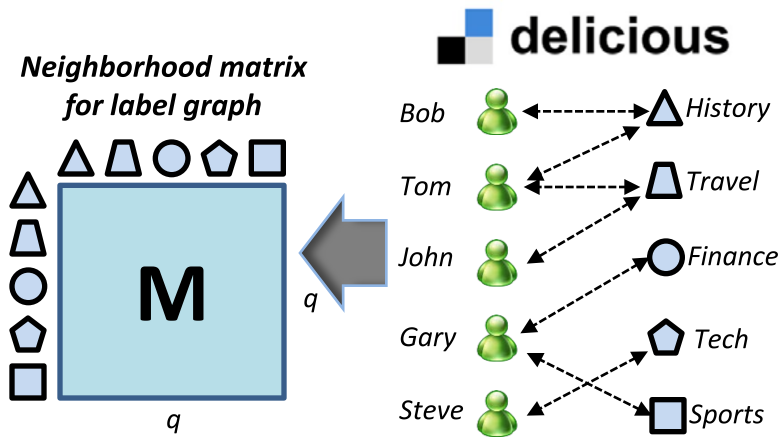
*From the taxonomy of the online information source, we can "**Compile**" a lot of base classification models*

SSFTL – Label Bridging via Laplacian Graph Embedding

Problem

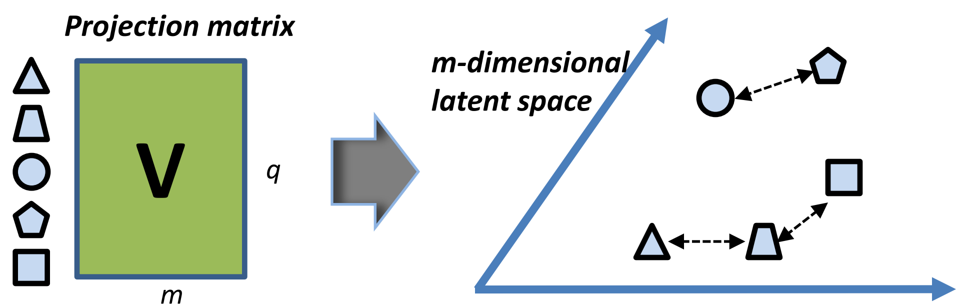
However, the ***label spaces*** *of the based classification models and the target task can be* ***different***





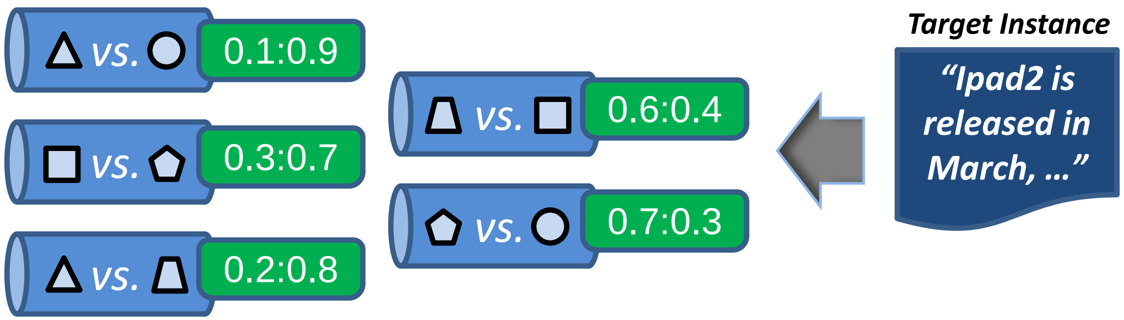
Laplacian Eigenmap
[Belkin & Niyogi,2003]

Since the label names are usually short and sparse, , in order to uncover the intrinsic relationships between the target and source labels, we turn to some ***social media*** *such as Delicious, which can help to bridge different label sets together.*



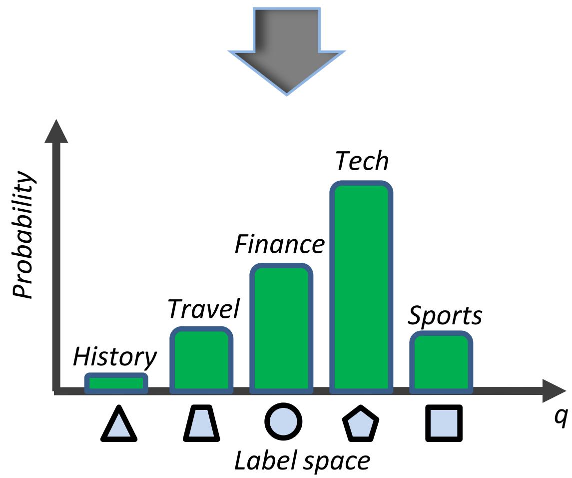
The ***relationships*** *between labels, e.g., similar or dissimilar, can be represented by the* ***distance*** *between their corresponding prototypes in the latent space, e.g., close to or far away from each other.*

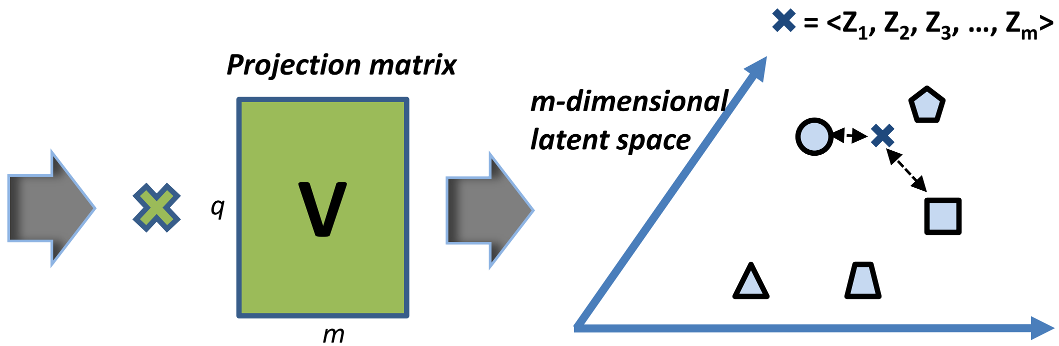
SSFTL – Mapping the target instance using the base classifiers & the projection matrix V



*For each target instance, we can obtain a **combined result on the label space** via aggregating the predictions from all the base classifiers*

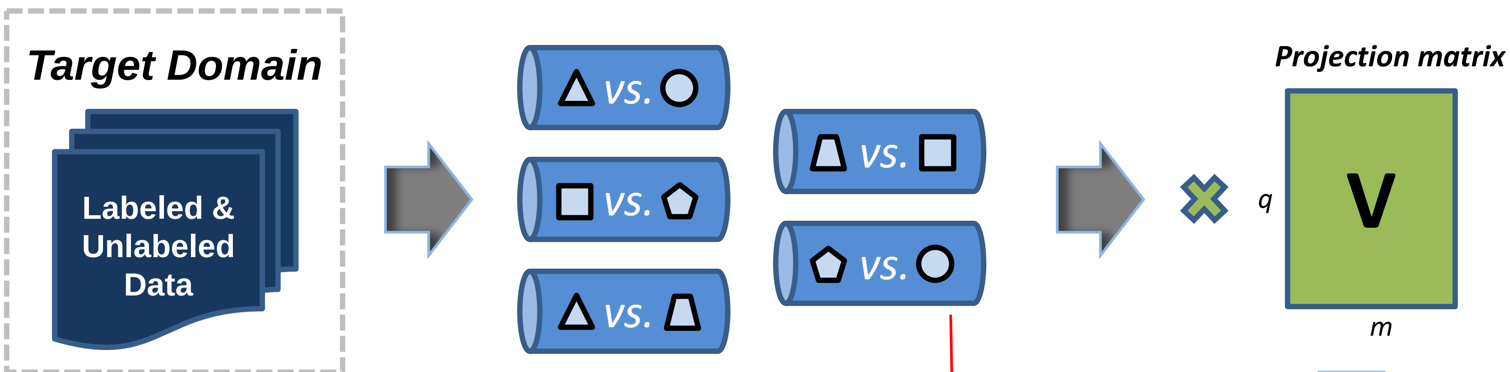
*Then we can use the **projection matrix V** to transform such combined results from the **label** space to a **latent** space*





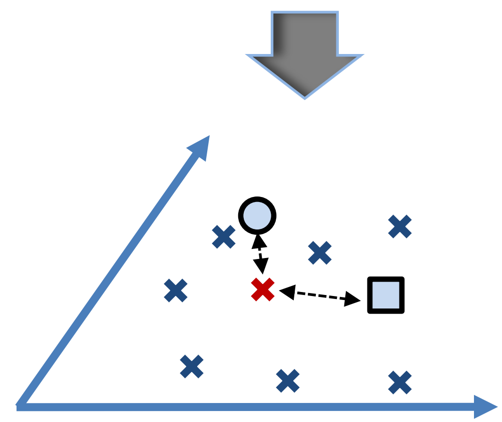
*However, do we need to recall the base classifiers during the **prediction** phase?*
*The answer is **No**!*

SSFTL – Learning a matrix W to directly project the target instance to the latent space



For each target instance, we first aggregate its prediction on the base label space, and then project it onto the latent space

$$\mathbf{V}'\mathbf{F}^u_S = \mathbf{V}'\sum_{i=1}^{k} \varepsilon_i \mathbf{F}^u_{S_i}$$



Loss on unlabeled data

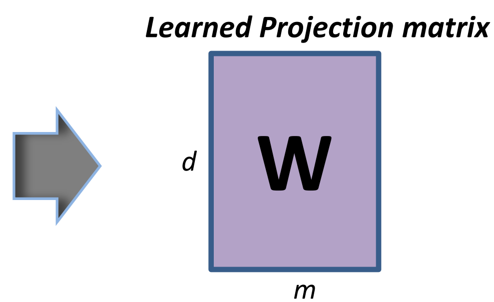
$$\Omega_{\mathbf{D}^u_T}(\mathbf{W}) = \frac{1}{n-\ell}\|\mathbf{W}'\mathbf{X}^u - \mathbf{V}'\mathbf{F}^u_S\|_F^2$$

Loss on labeled data

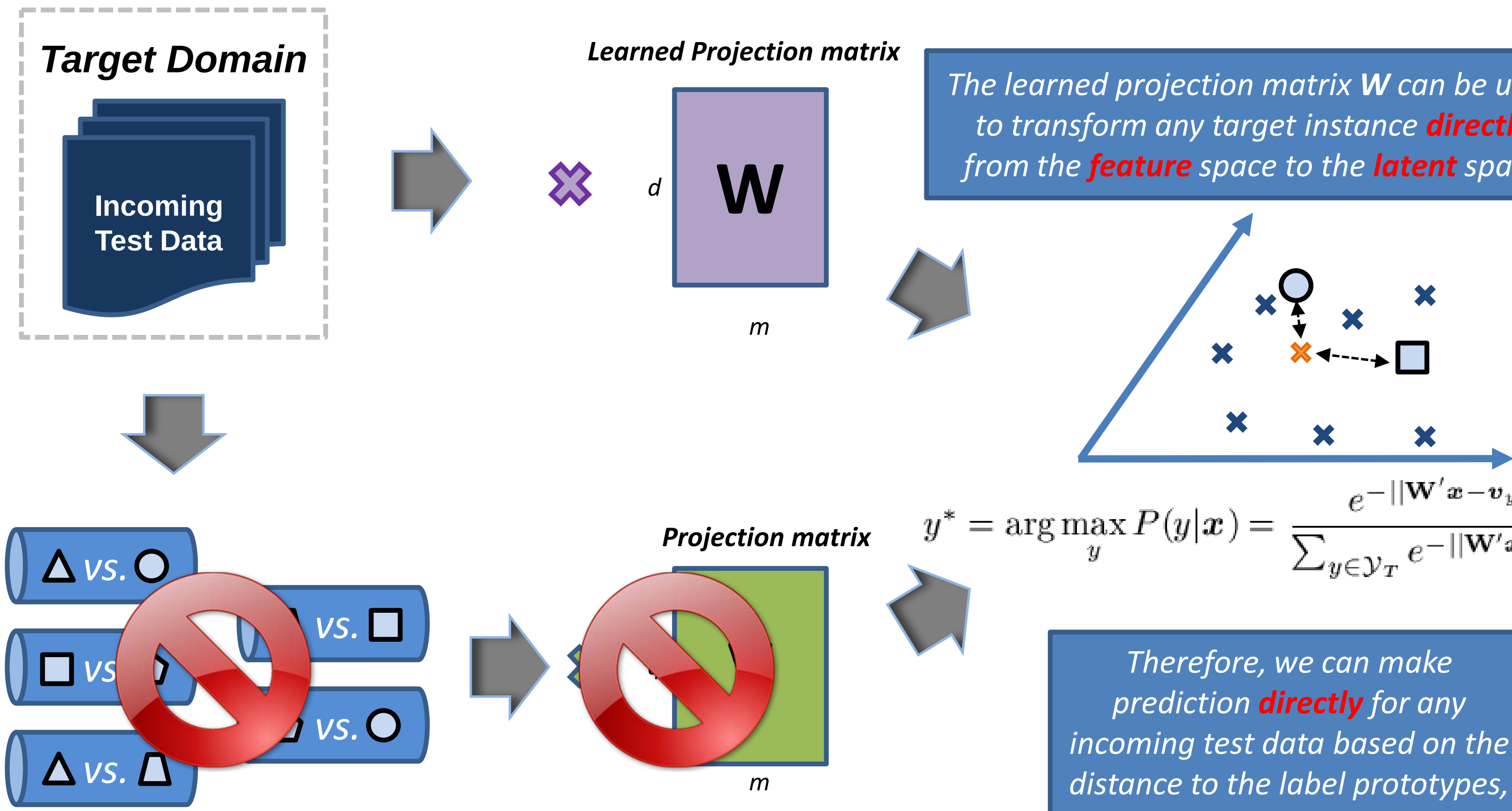
$$\Omega_{\mathcal{D}^\ell_T} = \frac{1}{\ell}\left\|\mathbf{W}'\mathbf{X}^\ell - \mathbf{V}'_T\phi(\mathbf{Y}^\ell))\right\|_F^2$$

Our regression model

$$\min_{\mathbf{W}} \Omega_{\mathcal{D}^\ell_T}(\mathbf{W}) + \lambda_1\|\mathbf{W}\|_F^2 + \lambda_2\Omega_{\mathcal{D}^u_T}(\mathbf{W})$$



SSFTL – Making predictions for the incoming test data



The learned projection matrix ***W*** can be used to transform any target instance ***directly*** from the ***feature*** space to the ***latent*** space

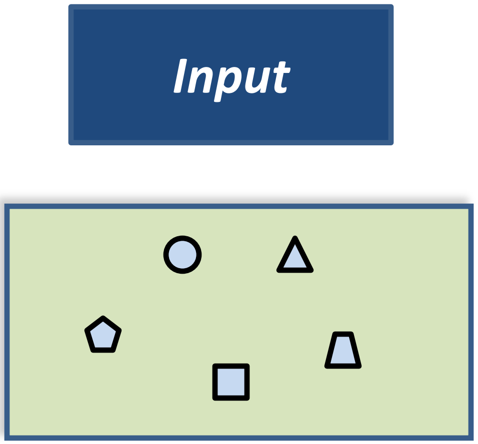
$$y^* = \arg\max_y P(y|\boldsymbol{x}) = \frac{e^{-\|\mathbf{W}'\boldsymbol{x} - \boldsymbol{v}_y\|_2^2}}{\sum_{y \in \mathcal{Y}_T} e^{-\|\mathbf{W}'\boldsymbol{x} - \boldsymbol{v}_y\|_2^2}}$$

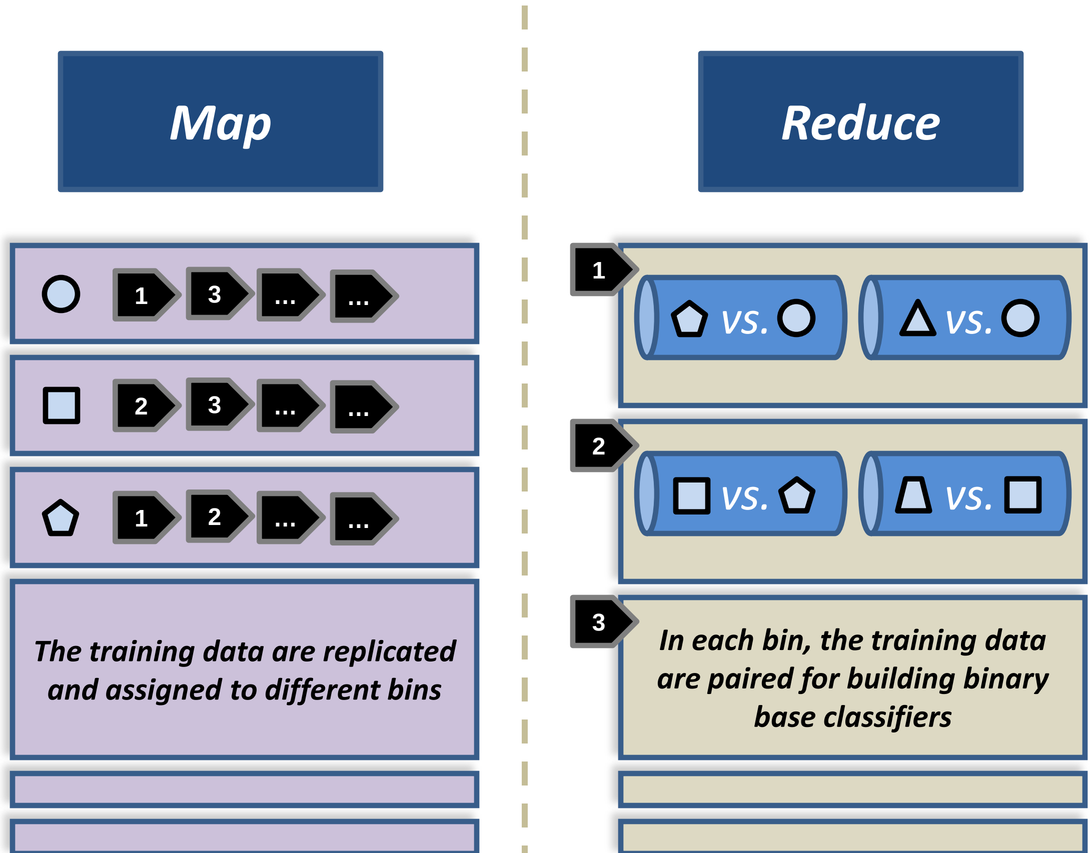
Therefore, we can make prediction ***directly*** for any incoming test data based on the distance to the label prototypes, ***without calling the base classification models***

Experiments - Datasets

- ***Building Source Classifiers with Wikipedia***
  - 3M articles, 500K categories (mirror of Aug 2009)
  - 50, 000 pairs of categories are sampled for source models
- ***Building Label Graph with Delicious***
  - 800-day historical tagging log (Jan 2005 ~ March 2007)
  - 50M tagging logs of 200K tags on 5M Web pages
- ***Benchmark Target Tasks***
  - 20 Newsgroups (190 tasks)
  - Google Snippets (28 tasks)
  - AOL Web queries (126 tasks)
  - AG Reuters corpus (10 tasks)

SSFTL - Building base classifiers Parallelly using MapReduce





If we need to build 50,000 base classifiers, it would take about ***two days*** *if we run the training process on a* ***single server.***
Therefore, we distributed the training process to a cluster with ***30 cores*** *using MapReduce, and finished the training within* ***two hours****.*

These pre-trained source base classifiers are ***stored*** *and* ***reused*** *for different incoming target tasks.*

Experiments - Results

Table 1: Comparison results under varying numbers of labeled data in the target task (accuracy in %).

Dataset	0		5			10			20		
	RG	SSFTL	SVM	TSVM	SSFTL	SVM	TSVM	SSFTL	SVM	TSVM	SSFTL
20NG	50.0	**80.3**	69.8	75.7	**80.6**	72.5	81.0	**81.6**	79.1	83.7	**84.5**
Google	50.0	**72.5**	62.1	69.7	**73.4**	64.5	73.2	**75.7**	67.3	73.8	**80.3**
AOL	50.0	**71.0**	72.1	74.1	**74.3**	73.7	76.8	**77.7**	79.2	77.8	**80.7**
Reuters	50.0	**72.7**	69.7	63.3	**74.3**	75.9	63.7	**76.9**	79.5	66.7	**80.1**

Unsupervised SSFTL

Semi-supervised SSFTL

Our regression model

$$\min_{\mathbf{W}} \boxed{\Omega_{\mathcal{D}_T^{\ell}}(\mathbf{W})} + \lambda_1 ||\mathbf{W}||_F^2 + \lambda_2 \Omega_{\mathcal{D}_T^{u}}(\mathbf{W})$$

-Parameter setttings-
Source models: *5,000*
Unlabeled target data: *100%*
lambda_2: *0.01*

Experiments - Results

Table 2: Comparison results on varying numbers of source classifiers (accuracy in %).

Dataset	Number of source classifiers for SSFTL						
	250	500	1K	2K	5K	10K	20K
20NG	76.3	78.2	80.3	82.5	84.5	85.1	**85.6**
Google	70.6	73.1	76.6	78.5	80.3	**80.4**	80.2
AOL	67.6	76.6	78.0	78.8	80.7	**81.2**	79.1
Reuters	72.2	74.0	76.7	78.0	**80.1**	79.6	78.1

For each target instance, we first aggregate its prediction on the base label space, and then project it onto the latent space

$$\mathbf{V}'\mathbf{F}^u_S = \mathbf{V}'\boxed{\sum_{i=1}^{k}}\varepsilon_i \mathbf{F}^u_{S_i}$$

Loss on unlabeled data

$$\Omega_{\mathbf{D}^u_T}(\mathbf{W}) = \frac{1}{n-\ell}\|\mathbf{W}'\mathbf{X}^u - \mathbf{V}'\mathbf{F}^u_S\|^2_F$$

Our regression model

$$\min_{\mathbf{W}} \Omega_{\mathcal{D}^\ell_T}(\mathbf{W}) + \lambda_1\|\mathbf{W}\|^2_F + \lambda_2\,\Omega_{\mathcal{D}^u_T}(\mathbf{W})$$

-Parameter setttings-
Mode: *Semi-supervised*
Labeled target data: *20*
Unlabeled target data: *100%*
lambda_2: *0.01*

Experiments - Results

Table 3: Comparison results on varying size of unlabeled data in the target task (accuracy in %).

Dataset	Unlabeled data involved in SSFTL				
	20%	40%	60%	80%	100%
20NG	80.5	80.9	81.8	84.0	**84.5**
Google	74.5	74.9	76.4	77.9	**80.3**
AOL	73.4	75.7	77.1	78.2	**80.7**
Reuters	75.5	77.7	77.8	78.7	**80.1**

Our regression model

$$\min_{\mathbf{W}} \Omega_{\mathcal{D}_T^\ell}(\mathbf{W}) + \lambda_1 ||\mathbf{W}||_F^2 + \lambda_2 \boxed{\Omega_{\mathcal{D}_T^u}(\mathbf{W})}$$

-Parameter setttings-
Mode: *Semi-supervised*
Labeled target data: *20*
Source models: *5,000*
lambda_2: *0.01*

Experiments - Results

Table 4: Overall performance of SSFTL under varying values of λ_2 (accuracy in %).

Dataset	λ_2 of SSFTL						
	0	0.001	0.01	0.1	1	10	100
20NG	83.2	84.1	84.5	**85.3**	84.8	83.3	79.3
Google	76.6	79.1	**80.3**	78.7	78.2	77.4	74.3
AOL	78.3	79.5	**80.7**	79.1	78.8	76.3	73.4
Reuters	75.5	78.2	**80.1**	78.5	76.0	72.1	68.5

Supervised SSFTL

Semi-supervised SSFTL

Our regression model

$$\min_{\mathbf{W}} \Omega_{\mathcal{D}_T^\ell}(\mathbf{W}) + \lambda_1 ||\mathbf{W}||_F^2 + \boxed{\lambda_2} \Omega_{\mathcal{D}_T^u}(\mathbf{W})$$

-Parameter setttings-
Labeled target data: *20*
Unlabeled target data: *100%*
Source models: *5,000*

Experiments - Results

Table 5: Analysis on weighted and uniform SSFTL under varying number of labeled data (accuracy in %).

Dataset	Uniform SSFTL				Weighted SSFTL			
	5	10	20	30	5	10	20	30
20NG	72.8	80.7	81.2	81.9	80.6	81.6	84.5	85.9
Google	64.1	67.0	70.8	77.2	73.4	75.7	80.3	81.1
AOL	69.8	71.7	72.1	74.8	74.3	77.7	80.7	82.5
Reuters	69.7	70.3	75.5	78.8	74.3	76.9	80.1	82.6

For each target instance, we first aggregate its prediction on the base label space, and then project it onto the latent space

$$\mathbf{V}'\mathbf{F}_S^u = \mathbf{V}'\sum_{i=1}^{k} \boxed{\varepsilon_i} \mathbf{F}_{S_i}^u$$

Loss on unlabeled data

$$\Omega_{\mathbf{D}_T^u}(\mathbf{W}) = \frac{1}{n-\ell}\|\mathbf{W}'\mathbf{X}^u - \mathbf{V}'\mathbf{F}_S^u\|_F^2$$

Our regression model

$$\min_{\mathbf{W}} \Omega_{\mathcal{D}_T^\ell}(\mathbf{W}) + \lambda_1\|\mathbf{W}\|_F^2 + \lambda_2 \Omega_{\mathcal{D}_T^u}(\mathbf{W})$$

-Parameter setttings-
Mode: *Semi-supervised*
Labeled target data: *20*
Source models: *5,000*
Unlabeled target data: *100%*
lambda_2: *0.01*

Related Works

Table 6: Summary of some related transfer learning works.

Transfer learning methods	*Scalability*	*Diff. label*
RSP [Shi *et al.*, 2009]	×	√
EigenTransfer [Dai *et al.*, 2009]	×	√
MTL-MI [Quadrianto *et al.*, 2010]	×	√
DAM [Duan *et al.*, 2009]	√	×
LWE [Gao *et al.*, 2008]	√	×
SSFTL	√	√

Conclusion

- ***Source-Selection-Free Transfer Learning***
  - *When the potential auxiliary data is embedded in very **large online** information sources*

- ***No need for task-specific source-domain data***
  - *We compile the label sets into a **graph Laplacian** for automatic label bridging*

- ***SSFTL is highly scalable***
  - *Processing of the online information source can be done **offline** and **reused** for different tasks.*

Q & A

Advance Research Topics in Transfer Learning

Wei Fan

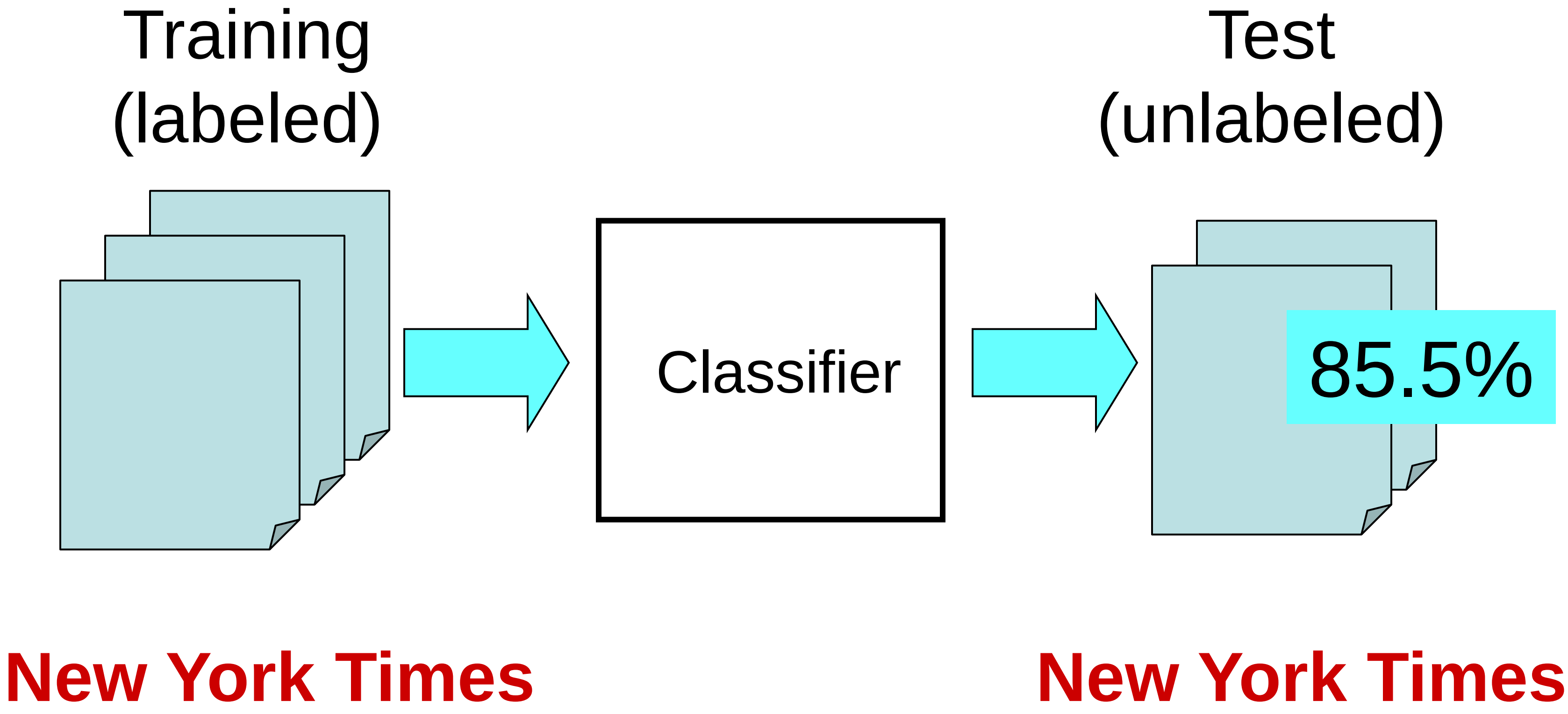
Huawei Noah's Ark Research Lab, Hong Kong

Predictive Modeling with Heterogeneous Sources

Xiaoxiao Shi   Qi Liu   Wei Fan
Qiang Yang   Philip S. Yu

Why learning
with heterogeneous sources?

Standard Supervised Learning



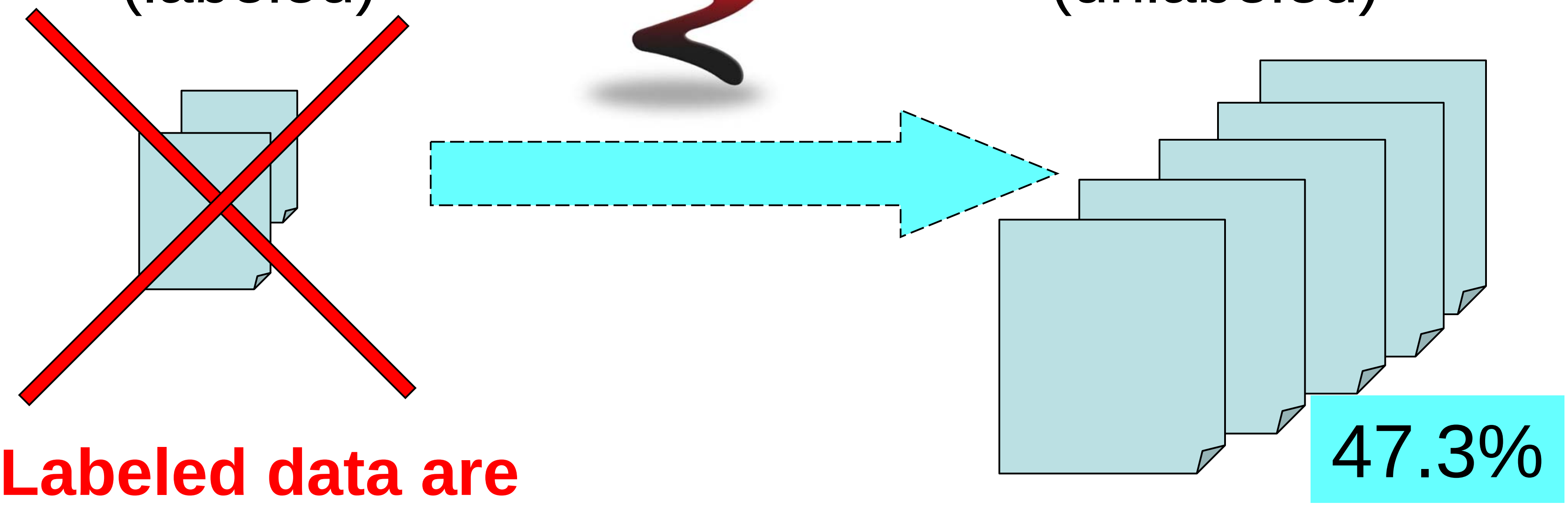
Why heterogeneous sources?

In Reality…

How to improve the performance?

Training (labeled)

Test (unlabeled)



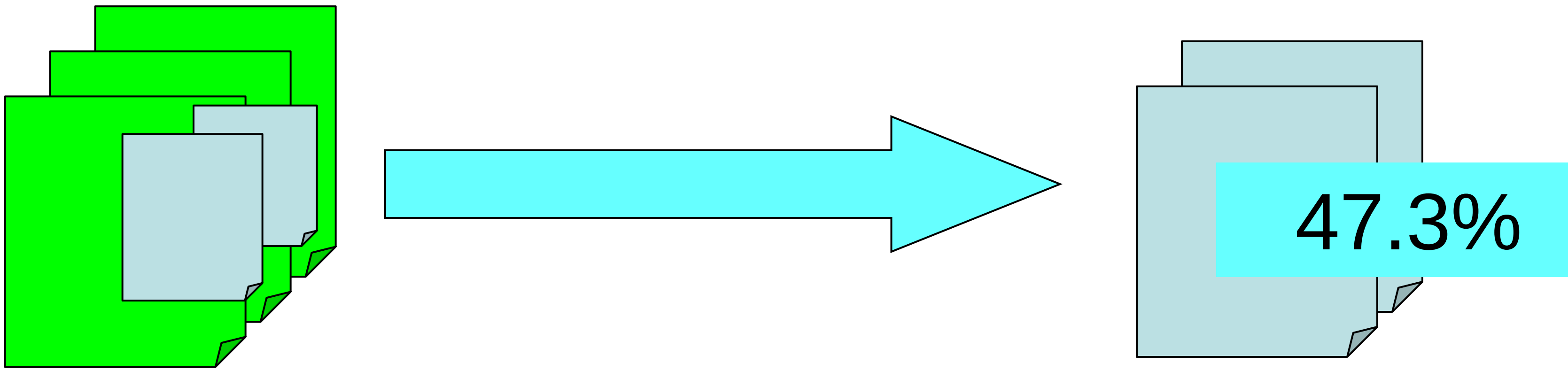
Labeled data are insufficient!

New York Times

Why heterogeneous sources?

Labeled data from other sources

Target domain test (unlabeled)



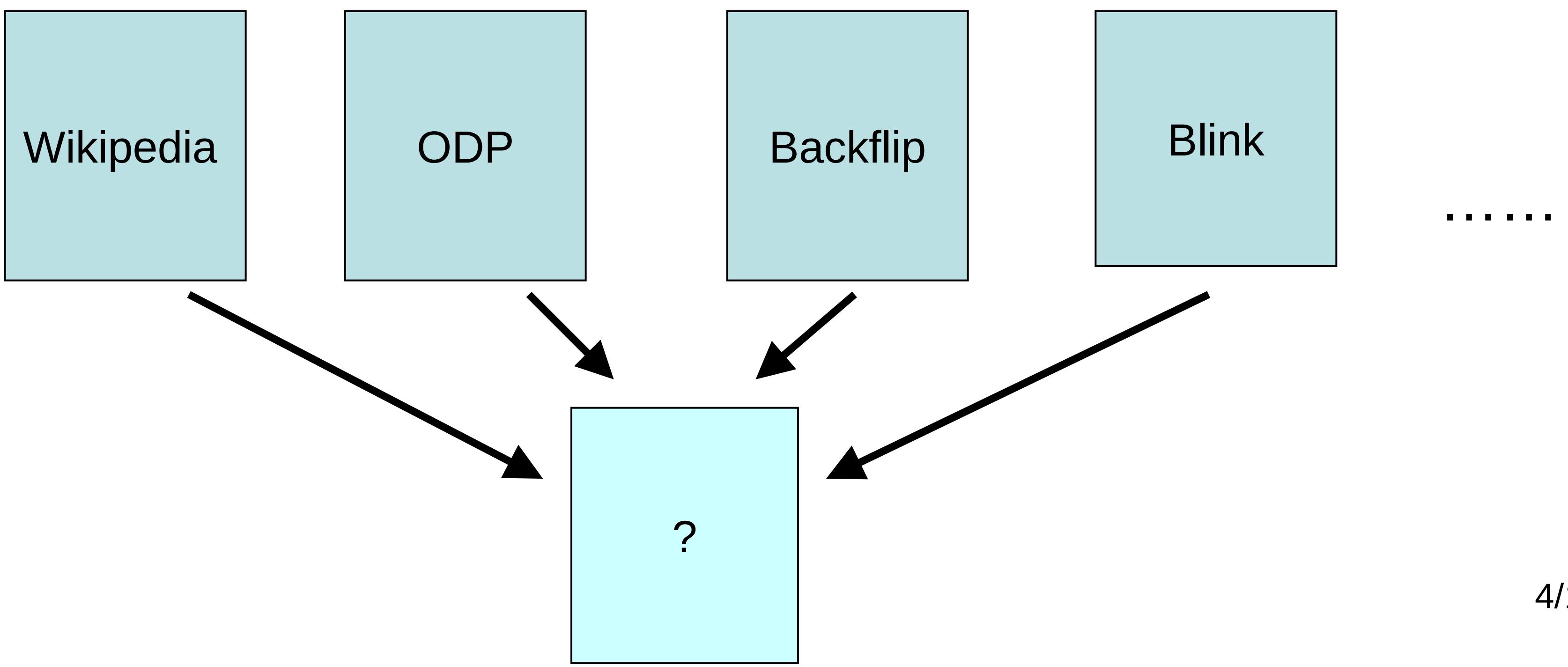
Reuters

New York Times

1. Different distributions
2. Different outputs
3. Different feature spaces

Real world examples

- Social Network:
  - Can various bookmarking systems help predict social tags for a new system given that their outputs (social tags) and data (documents) are different?



Real world examples

- Applied Sociology:
  - Can the suburban housing price census data help predict the downtown housing prices?

?

#rooms	#bathrooms	#windows	price
5	2	12	XXX
6	3	11	XXX

#rooms	#bathrooms	#windows	price
2	1	4	XXXXX
4	2	5	XXXXX

Other examples

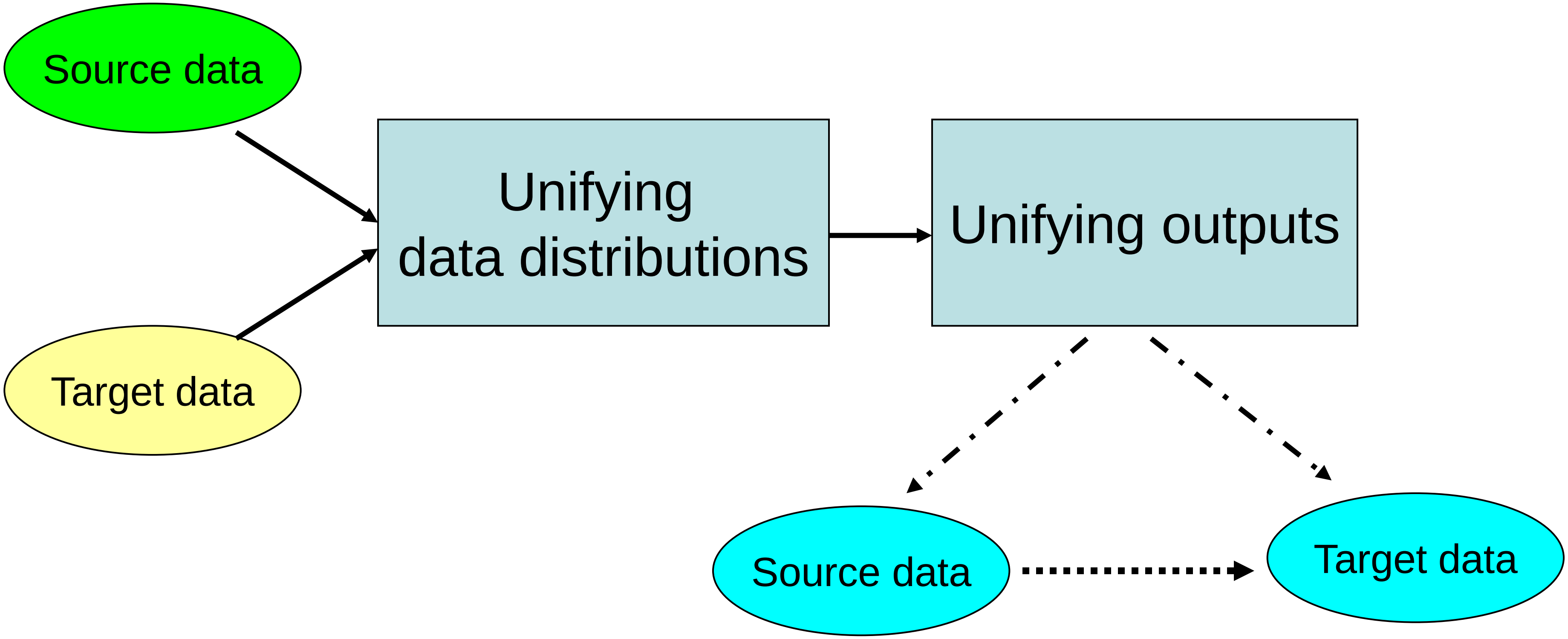
- Bioinformatics
  - Previous years' flu data → new swine flu
  - Drug efficacy data against breast cancer → drug data against lung cancer
  - ……
- Intrusion detection
  - Existing types of intrusions → unknown types of intrusions
- Sentiment analysis
  - Review from SDM→ Review from KDD

Learning with Heterogeneous Sources

- The paper mainly attacks two sub-problems:
  - Heterogeneous data distributions
    - Clustering based KL divergence and a corresponding sampling technique
  - Heterogeneous outputs (to regression problem)
    - Unifying outputs via preserving similarity.

Learning with Heterogeneous Sources

- General Framework



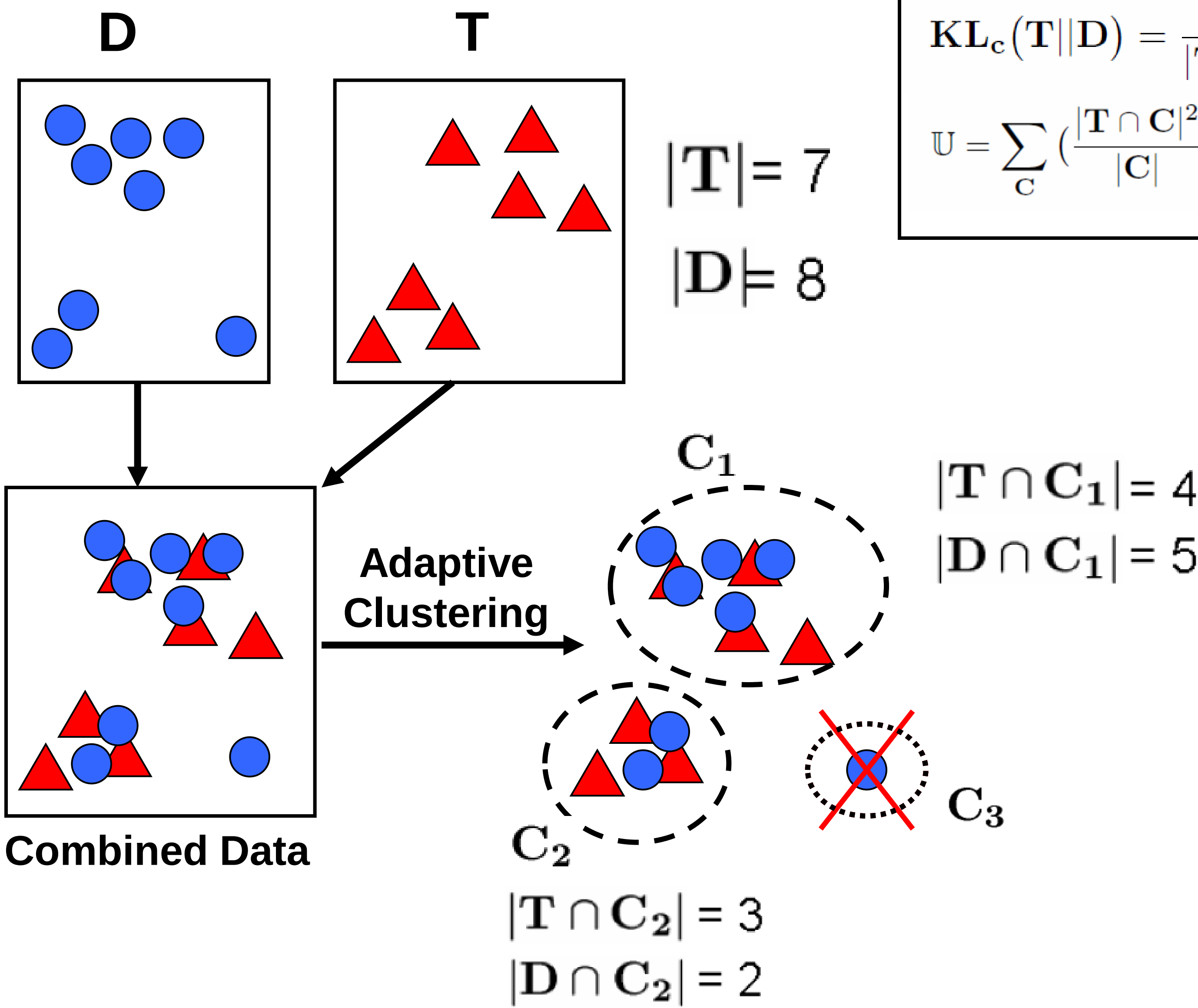
Unifying Data Distributions

- Basic idea:
  - Combine the source and target data and perform clustering.
  - Select the clusters in which the target and source data are similarly distributed, evaluated by KL divergence.

An Example

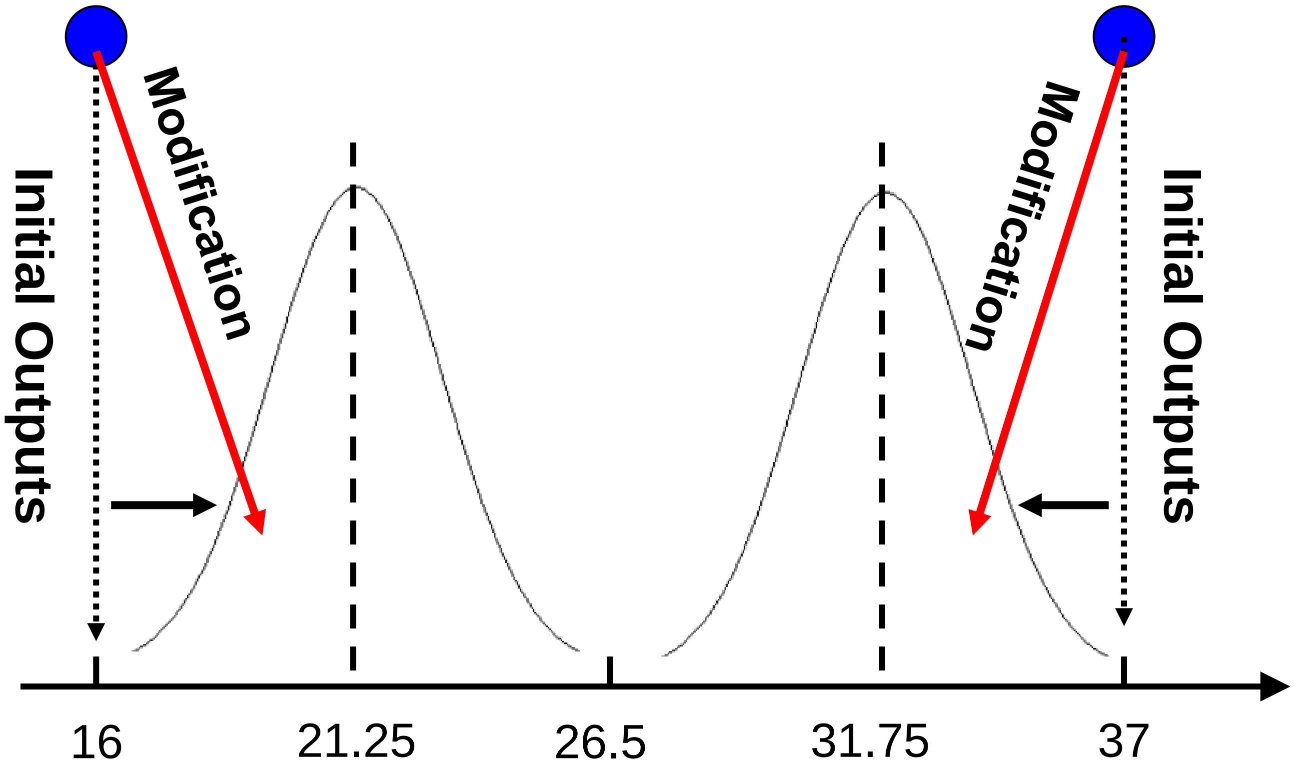
$$\mathbf{KL_c}(\mathbf{T}||\mathbf{D}) = \frac{2}{|\mathbf{T}|}\mathbb{U} + \log\frac{|\mathbf{D}|}{|\mathbf{T}|}$$

$$\mathbb{U} = \sum_{\mathbf{C}}\left(\frac{|\mathbf{T}\cap\mathbf{C}|^2}{|\mathbf{C}|}\log\frac{|\mathbf{T}\cap\mathbf{C}|}{|\mathbf{D}\cap\mathbf{C}|}\right)$$



Unifying Outputs

- Basic idea:
  - Generate initial outputs according to the regression model
  - For the instances similar in the original output space, make their new outputs closer.


Initial Outputs
Modification
Modification
Initial Outputs
16
21.25
26.5
31.75
37

Experiment

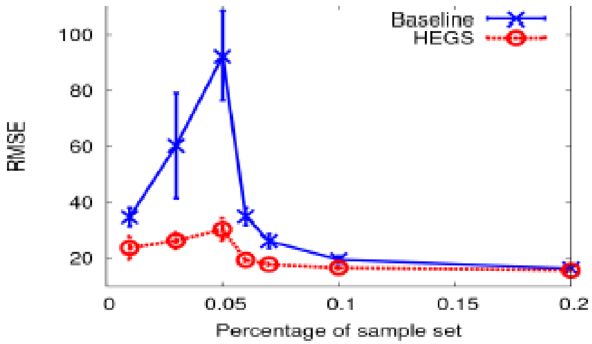
- Bioinformatics data set:

Table 1: Description of the data sets (#Feature =161)

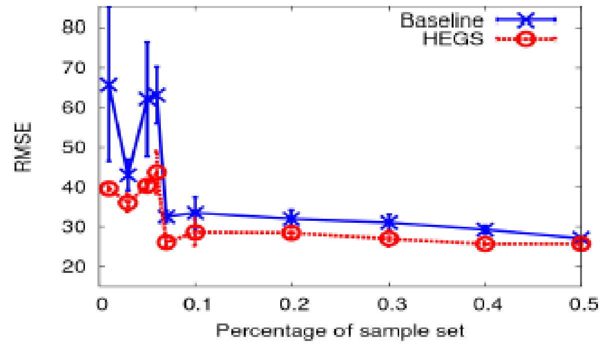
Order	Type	Size	Scale	References
1	Regression	2431	0∼99.99	[8]
2	Regression	561	1∼127.8	[8]
3	Regression	601	0∼100	[8]
4	Regression	290	2.1∼98	[15]
5	Regression	344	0.2∼98.5	[15]
6	Classification	7443	4 classes	[10]
7	Classification	196	2 classes	[16]

Note: Some references, such as [8], refer to several data sets from different research groups

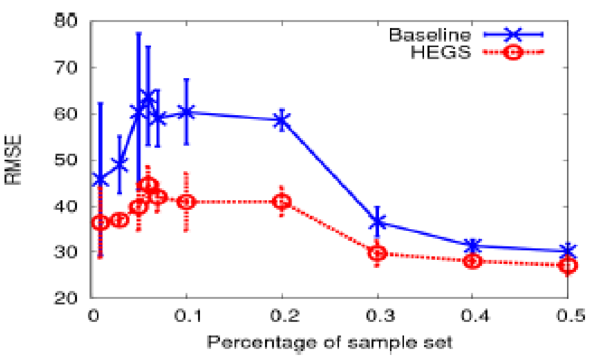
Experiment



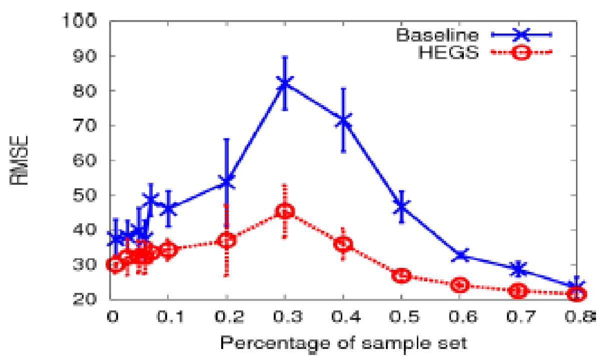
(a) Data set 1



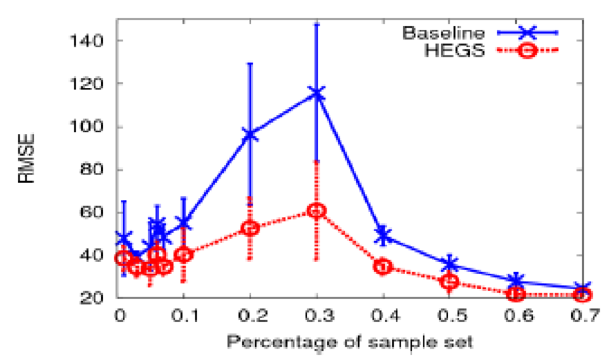
(b) Data set 2



(c) Data set 3



(d) Data set 4



(e) Data set 5

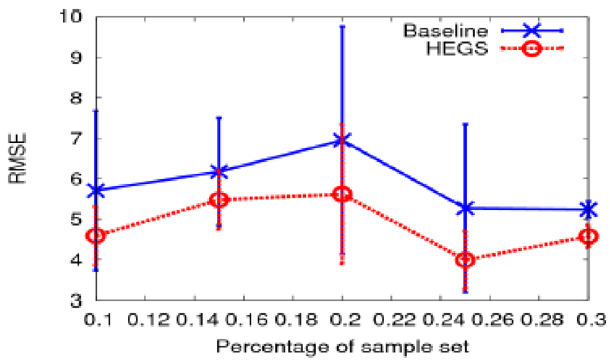
Experiment

- Applied sociology data set:

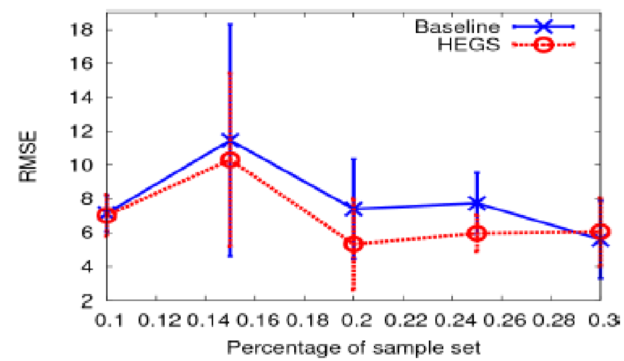
Table 2: Description of the data sets (#Feature =18)

Name	Size	Scale
Newton	18	2.47~21.46
Boston Roxbury	19	12.03~36.98
Lynn	22	6.58~27.71
Boston Savin Hill	23	15.17~34.02
Cambridge	30	1.73~29.53
Somerville	15	11.12~34.41
South Boston	10	3.53~18.46
Brookline	11	7.67~18.66
East Boston	11	10.29~19.01
Quincy	11	9.38~29.55

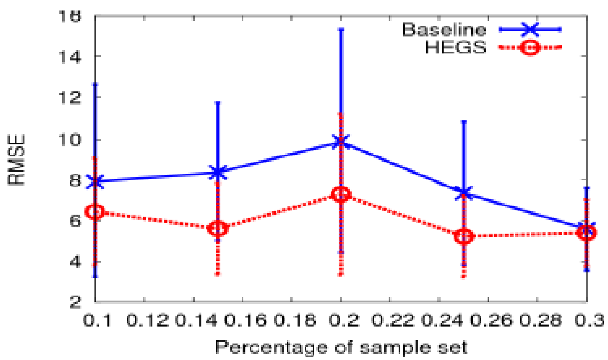
Experiment



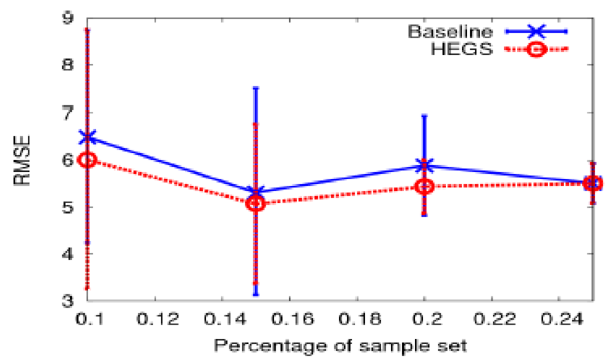
(a) Newton



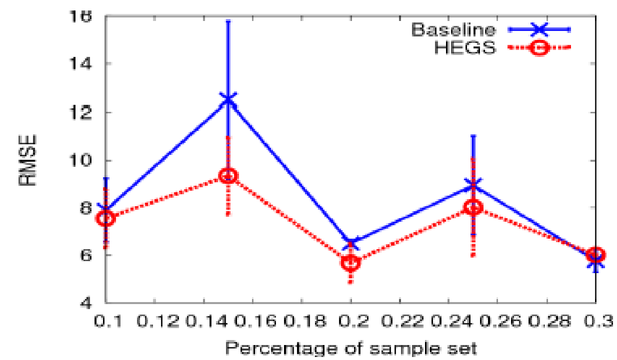
(b) Boston Roxbury



(c) Lynn



(d) Boston Savin Hill



(e) Cambridge

Conclusions

- Problem: Learning with Heterogeneous Sources:
  - Heterogeneous data distributions
  - Heterogeneous outputs
- Solution:
  - Clustering based KL divergence help perform sampling
  - Similarity preserving output generation help unify outputs

Transfer Learning on Heterogeneous Feature Spaces via Spectral Transformatio

Xiaoxiao Shi, Qi Liu, Wei Fan,
Philip S. Yu, and Ruixin Zhu

Motivation

Standard Supervised Learning

Training documents (labeled)

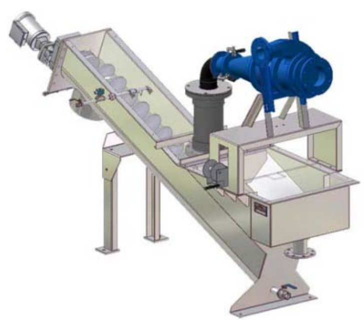

German Alert Followed Namibian Airport Scare

Namibian authorities halted and searched a flight bound for Munich on Wednesday morning after luggage screeners found an untagged laptop bag containing batteries wired to a fuse and clock, a discovery made just hours before Germany issued a rare security alert.

On Thursday, Air Berlin said no explosives were found in the bag.

A statement by the Namibia Airports Company said that "a suspicious parcel" had been found in a luggage screening area in the airport at Windhoek, Namibia's capital, at 8:50 a.m. local time Wednesday, and that as a result Air

Classifier



Test documents (unlabeled)


Italy's Political Houdini May Not Escape This Time

ROME — Over the years, Prime Minister Silvio Berlusconi has marveled Italians with his Houdini-like powers to escape the trickiest political traps and bounce back when all odds were against him.

But this time around, as a political crisis deepens and looks poised to bring down the government within weeks, something is notably different in Italy.

The New York Times

The New York Times

In Reality…

How to improve the performance?

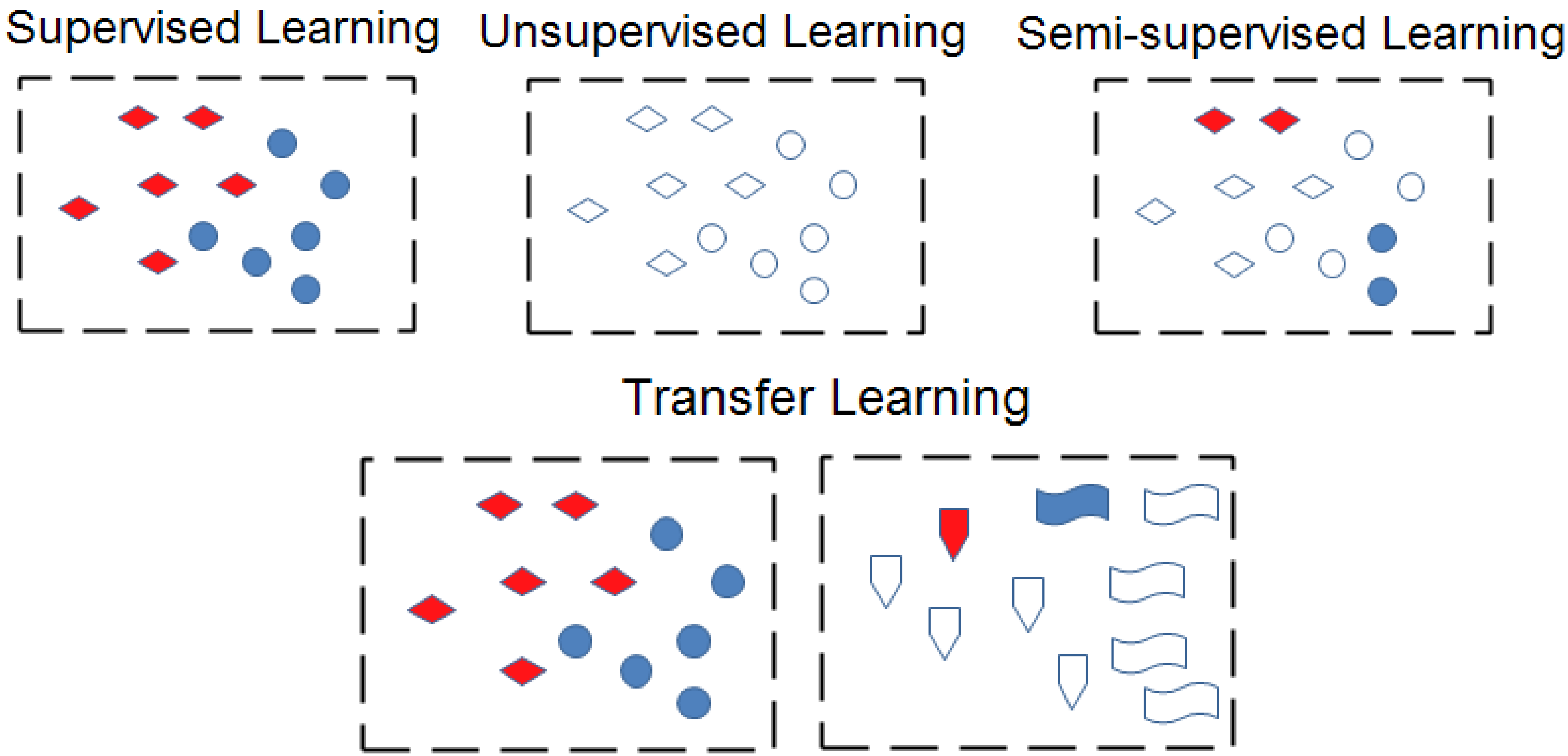
Training (labeled)

Labeled data are insufficient!

Huge set of unlabeled documents

The New York

47.3%



Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Transfer Learning

Labeled data from other sources

REUTERS

Target domain test (unlabeled)

The New York Times

Heterogeneous datasets:

1. Different data distributions: $P(x_{train})$ and $P(x_{test})$ are different
2. Different outputs: y_{train} and y_{test} are different
3. Different feature spaces: x_{train} and x_{test} are different

- WiFi-based localization tracking [Pan et al'08]
- Collaborative Filtering [Pan et al'10]
- Activity Recognition [Zheng et al'09]
- Text Classification [Dai et al'07]
- Sentiment Classification [Blitzer et al ‘07]
- Image Categorization [Shi et al’10]
- … …

Issues

- Different data distributions: $P(x_{train})$ and $P(x_{test})$ are different

Chicago Tribune focuses more on Chicago local news

REUTERS focuses more on global news

WIKIPEDIA The Free Encyclopedia focuses more on scientific/objective documents

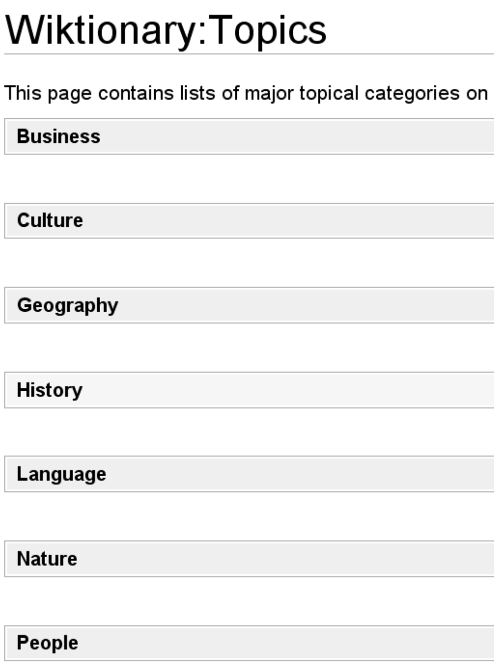
Issues

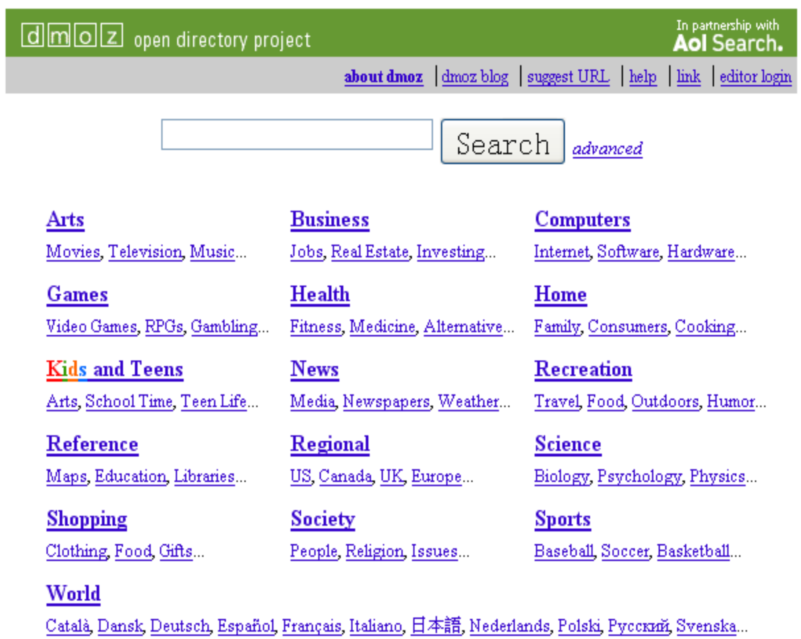
- Different outputs: y_{train} and y_{test} are different

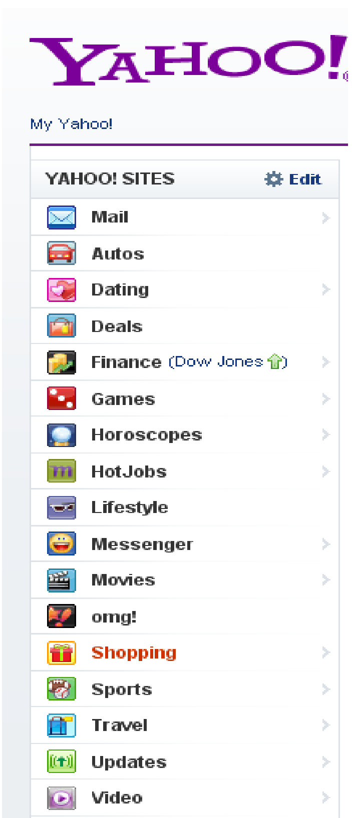
Wikipedia

ODP

Yahoo!

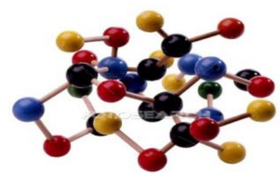






Issues

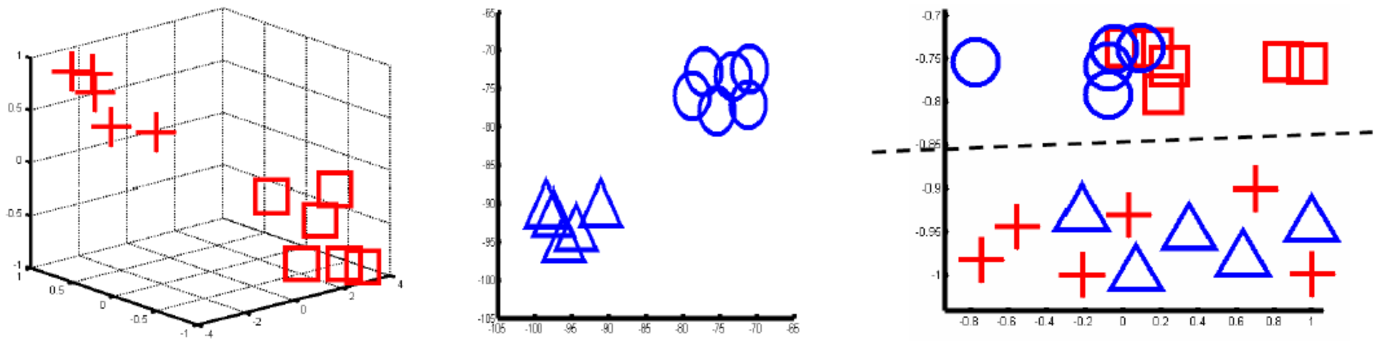
- Different feature spaces (the focus on the paper)
  - Drug efficacy tests:
    - Physical properties
    - To roperties



  - Image Classification
    - Wavelet features
    - Color histogram

Unify different feature spaces

- Different number of features; different meanings of the features, **no common feature, no overlap**.
- Projection-based approach **HeMap**
  - Find a projected space where (1) the source and target data are similar in distribution; (2) the original



(a) 3-D data (b) 2-D data (c) Projected space

Unify different feature spaces via HeMap

Optimization objective of HeMap:

$$\min_{\mathbf{B_T},\mathbf{B_S}} \ell(\mathbf{B_T},\mathbf{T}) + \ell(\mathbf{B_S},\mathbf{S}) + \beta \cdot \mathbf{D}(\mathbf{B_T},\mathbf{B_S}) \quad (1)$$

$\ell(\mathbf{B_T},\mathbf{T}) = \|\mathbf{B}$ $\ell(\mathbf{B_S},\mathbf{S}) = \|$ $\mathbf{D}(\mathbf{B_T},\mathbf{B_S}) = \frac{1}{2}(\ell(\mathbf{B_T},\mathbf{S}) + \ell(\mathbf{B_S},\mathbf{T}))$

The linear proj error

The linear pr error

The difference between the projected data

where $\mathbf{B_T} \in \mathbb{R}^{r\times k}, \mathbf{B_S} \in \mathbb{R}^{q\times k}$ are the projected matrices of $\mathbf{T}$ and $\mathbf{S}$ respectively.

Unify different feature spaces via HeMap

With some derivations, the objective can be reformulated as (more details can be found in the paper):

Theorem 1: The minimization problem in Eq. (4) is equivalent to the following maximization problem:

$$\min_{\mathbf{B}_\mathbf{T}^\top \mathbf{B}_\mathbf{T}=\mathbf{I},\, \mathbf{B}_\mathbf{S}^\top \mathbf{B}_\mathbf{S}=\mathbf{I}} G = \max_{\mathbf{B}^\top \mathbf{B}=\mathbf{I}} \mathbf{tr}(\mathbf{B}^\top \mathbf{A}\mathbf{B}) \qquad (6)$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_\mathbf{T} \\ \mathbf{B}_\mathbf{S} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix}. \qquad (7)$$

$$\mathbf{A}_1 = 2\mathbf{T}\mathbf{T}^\top + \frac{\beta^2}{2}\mathbf{S}\mathbf{S}^\top, \mathbf{A}_4 = \frac{\beta^2}{2}\mathbf{T}\mathbf{T}^\top + 2\mathbf{S}\mathbf{S}^\top$$

$$\mathbf{A}_2 = \mathbf{A}_3^\top = \beta(\mathbf{S}\mathbf{S}^\top + \mathbf{T}\mathbf{T}^\top)$$

Algorithm flow of HeMap

Construct matrix $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix}$

$\downarrow$

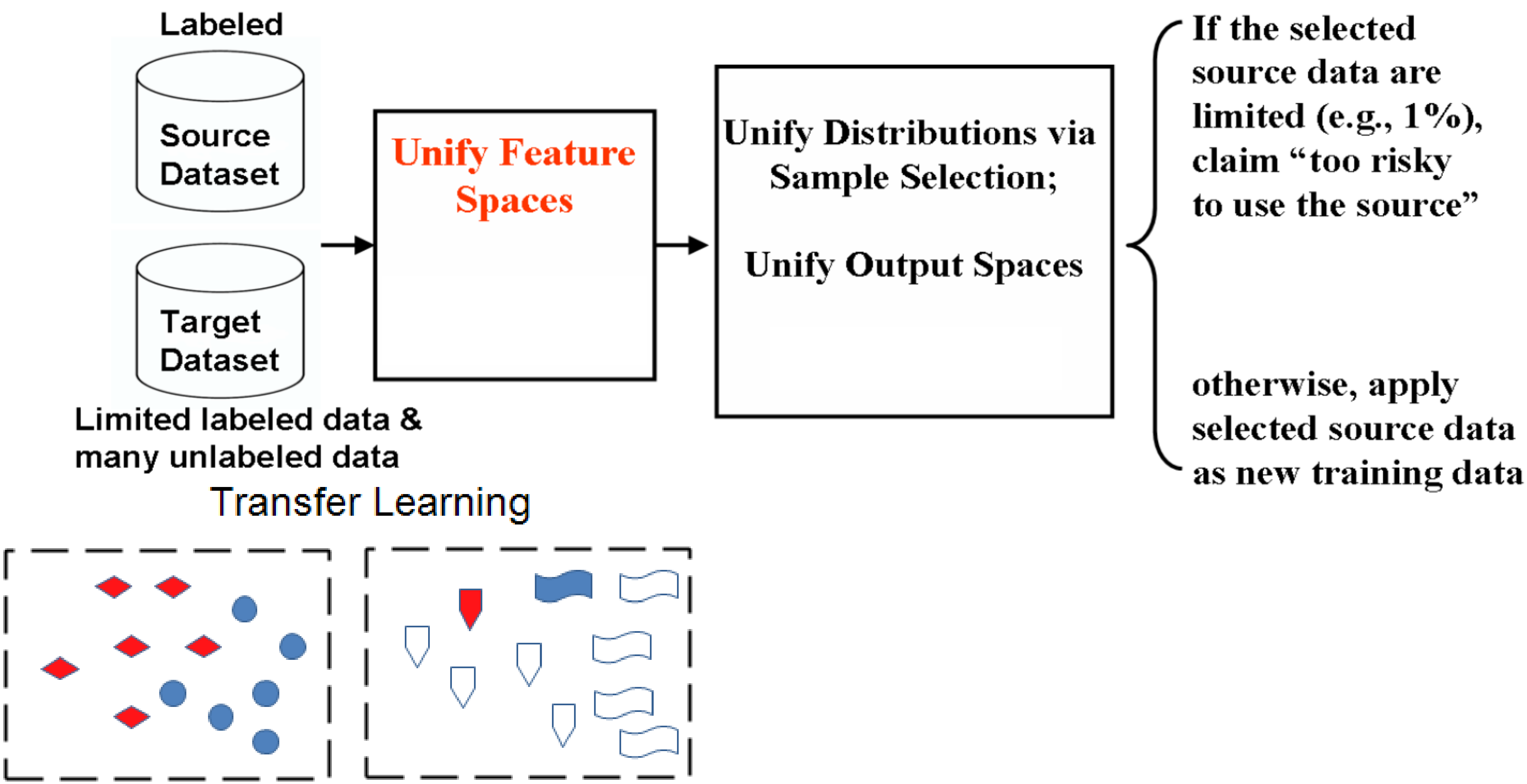
$\mathbf{A}_1 = 2\mathbf{T}\mathbf{T}^\top + \frac{\beta^2}{2}\mathbf{S}\mathbf{S}^\top, \mathbf{A}_4 = \frac{\beta^2}{2}\mathbf{T}\mathbf{T}^\top + 2\mathbf{S}\mathbf{S}^\top$

Calculate the top-k eigenvalues of $\mathbf{A}$, and their corresponding eigenvectors $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_k]$.

$\downarrow$

$\mathbf{B}_\mathbf{T}$ is the first half rows of $\mathbf{U}$; $\mathbf{B}_\mathbf{S}$ is the second half rows of $\mathbf{U}$.

Generalized HeMap to handle heterogeneous data (different distributions, outputs and feature spaces)



Unify different distributions and outputs

- Unify different distributions
  - Clustering based sample selection [Shi etc al,09]
- Unify different outputs
  - Bayesian like scheme

$$p(y|\mathbf{x}) = \sum_{v} (p(v|\mathbf{x}) p(y|v)) \qquad (11)$$

where $\mathbf{x}$ is the data to be predicted; y is the target label; and v denotes the output from the source task.

Generalization bound

Theorem 4: Let $\mathcal{H}$ be a a hypothesis space. Let $\mathbf{T}$ be unlabeled samples of size r. Let $\mathbf{S}$ be a labeled sample of size q generated by drawing ϑq points from target data and $(1-\vartheta)q$ points from source data. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of the error on $\mathbf{S}$ and $h^* = \min_{h \in \mathcal{H}} \epsilon(h)$ is the target risk minimizer, then with probability at least 1 - δ (over the choice of the samples),

$$\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{g(\hat{h})\log(2q) - \log\delta}{2q}}$$

$$+ 2(1-\alpha)\left(\frac{1}{2}\underline{\mathbb{d}(\mathbf{T},\mathbf{S})} + 4\sqrt{\frac{2g(\hat{h})\log r + \log\frac{4}{\delta}}{r}} + \underline{\xi}\right)$$

α and β are domain-specific parameters; $g(\hat{h})$ is model complexity

Principle I: minimize the difference between target and source datasets

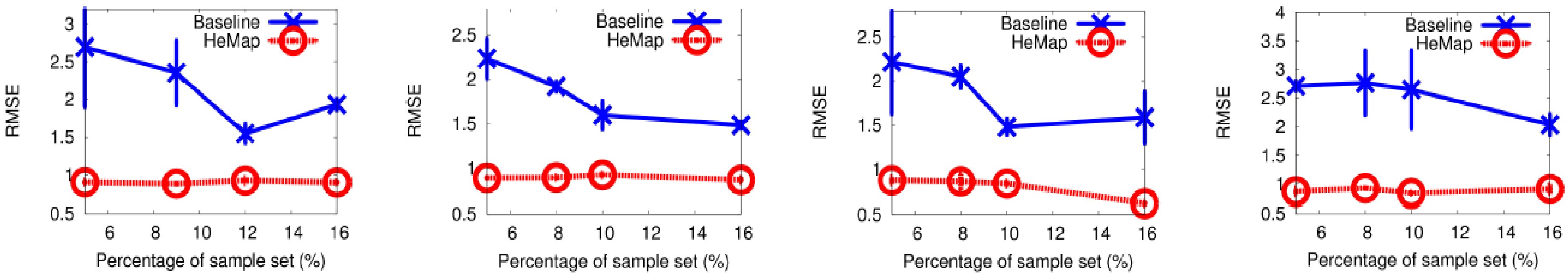
$\xi = \min_{h \in \mathcal{H}} \epsilon_{\mathbf{T}}(h) + \epsilon_{\mathbf{S}}(h)$

Principle II: minimize the combined expected error by maintaining the original structure (minimize projection error)

Experiments

- Drug efficacy prediction
  - The dataset is collected by the College of Life Science and Biotechnology of Tongji University, China. It is to predict the efficacy of drug compounds against certain cell lines.
  - The data are generated in two different feature spaces
    - general descriptors: refer to **physical** properties of compounds
    - drug-like index: refer to simple **topological** indices of compounds.

Experiments

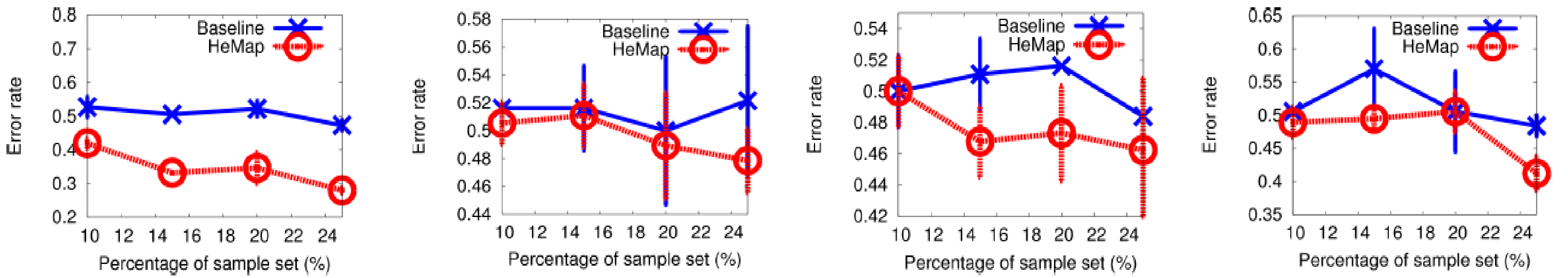


(a) Target is data set 1; source is data set 2

(b) Target is data set 2; source is data set 1

(c) Target is data set 3; source is data set 4

(d) Target is data set 4; source is data set 3



(e) Target is data set 5; source is data set 6

(f) Target is data set 6; source is data set 5

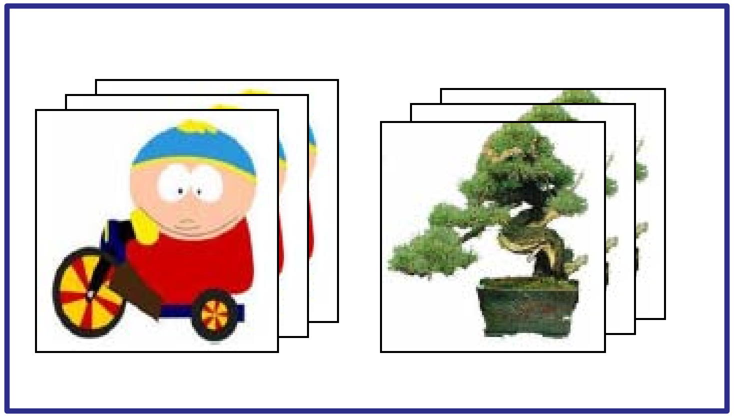
(g) Target is data set 7; source is data set 8

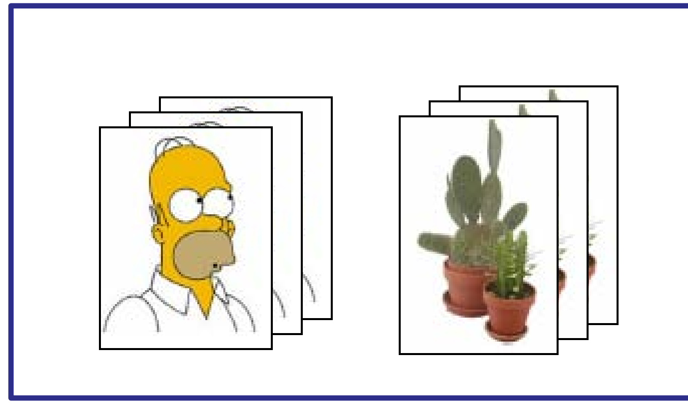
(h) Target is data set 8; source is data set 7

Experiments

- Image classification

Cartman & Bonsai



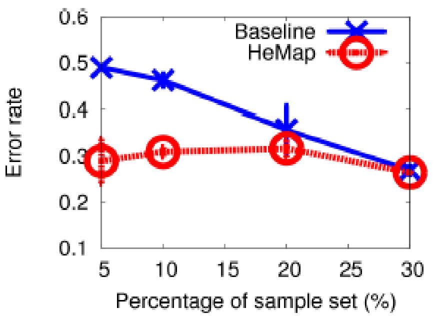


Homer Simpson & Cactus

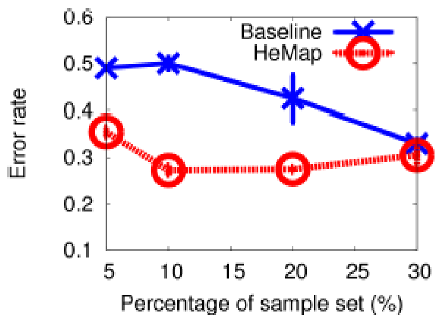
Homer Simpson & Coin

Superman & CD

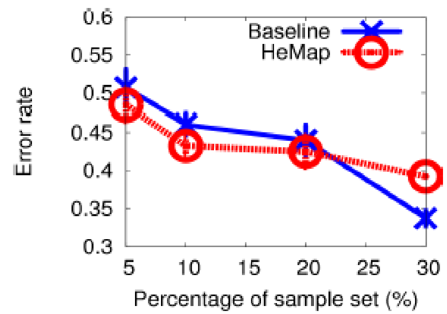
Experiments



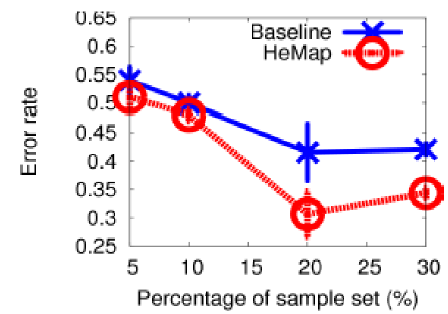
(a) Target is Cartman and Bonsai; source is Homer Simpson and Cactus



(b) Target is Homer Simpson and Cactus; source is Cartman and Bonsai



(c) Target is Homer Simpson and Coin; source is Superman and CD



(d) Target is Superman and CD; source is Homer Simpson and Coin

Conclusions

- Extends the applicability of supervised learning, semi-supervised learning and transfer learning by using heterogeneous data:
  - Different data distributions
  - Different outputs
  - **Different feature spaces**
- Unify different feature spaces via linear projection with two principles
  - Maintain the original structure of the data
  - Maximize the similarity of the two data in the projected space

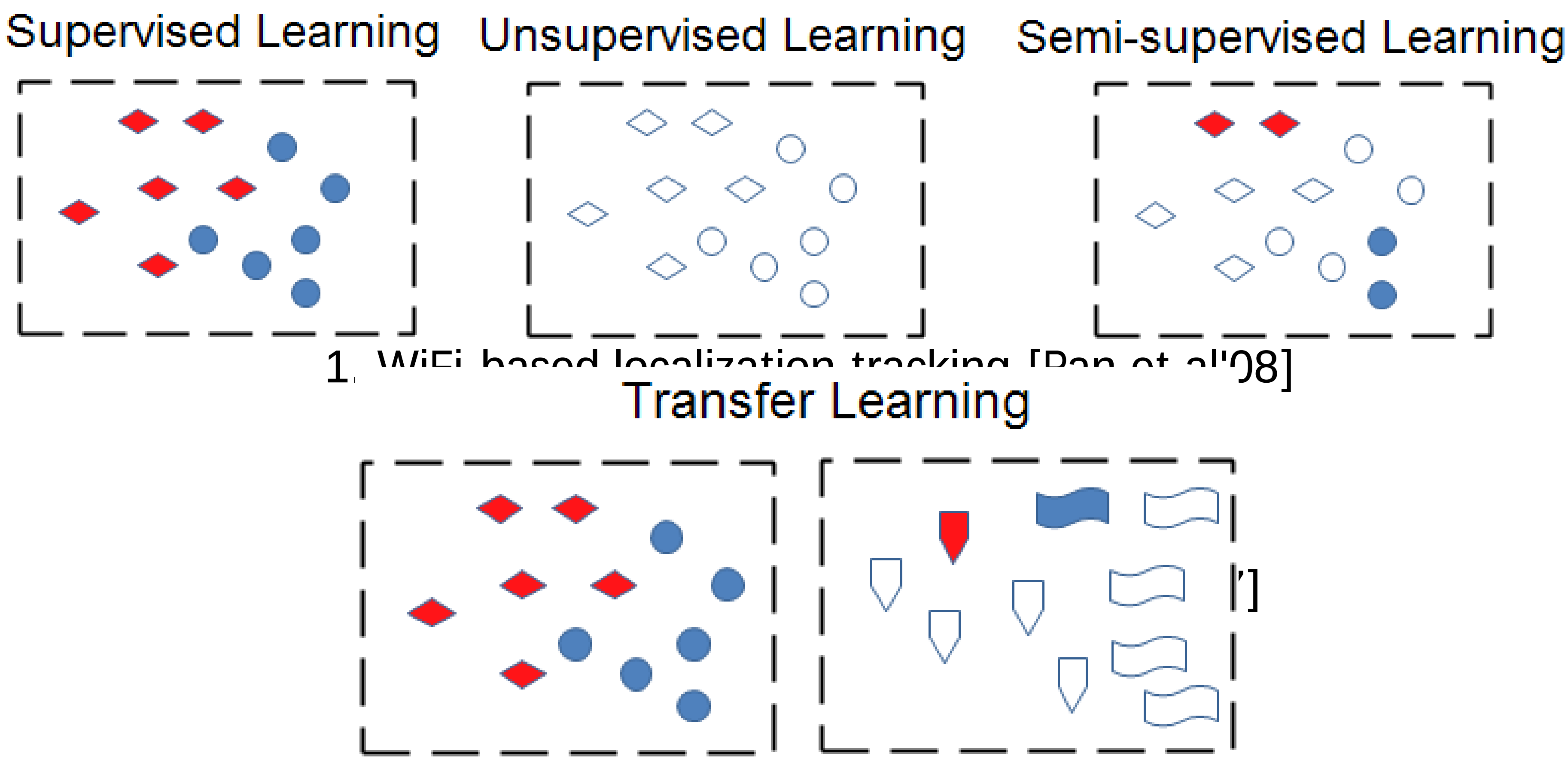
Cross Validation Framework to Choose Amongst Models and Datasets for Transfer Learning

Erheng Zhong[¶], Wei Fan[‡], Qiang Yang[¶],
Olivier Verscheure[‡], Jiangtao Ren[†]

Transfer Learning: What is it

Definition

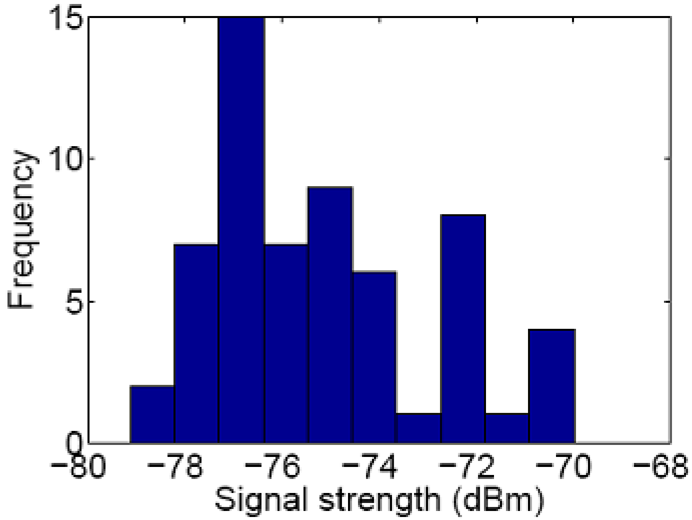
“source-domains” to improve “target-domain”: short of labeled information.



1. WiFi based localization tracking [Pan et al'08]

Application

Indoor WiFi localization tracking



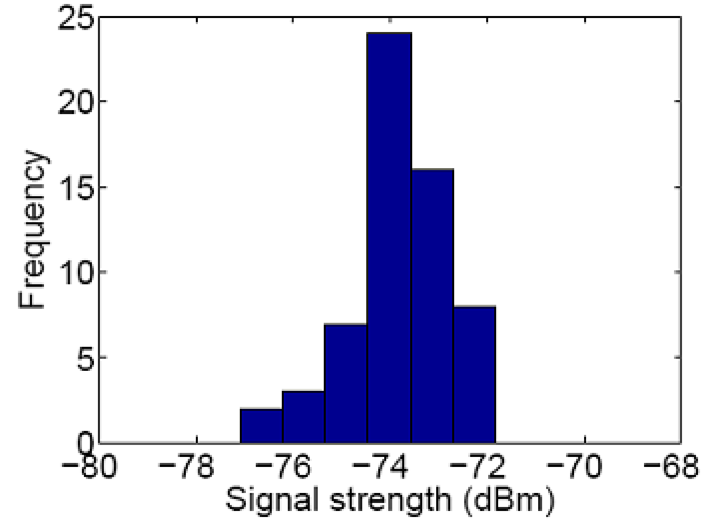
(a) WiFi signal at time period 1

Ly)

7) m

11) m

Transfer



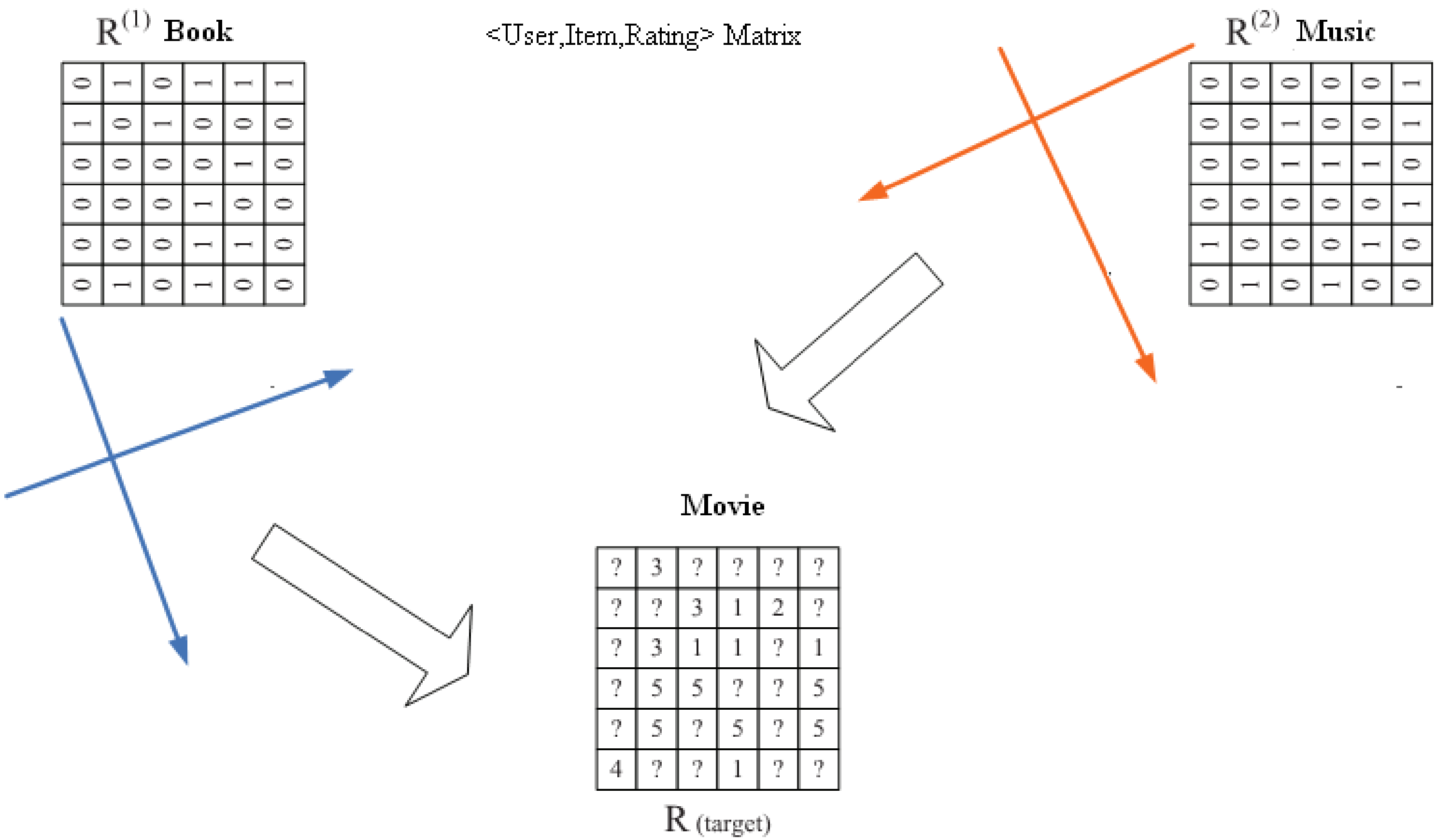
(b) WiFi signal at time period 2

ıe access poin

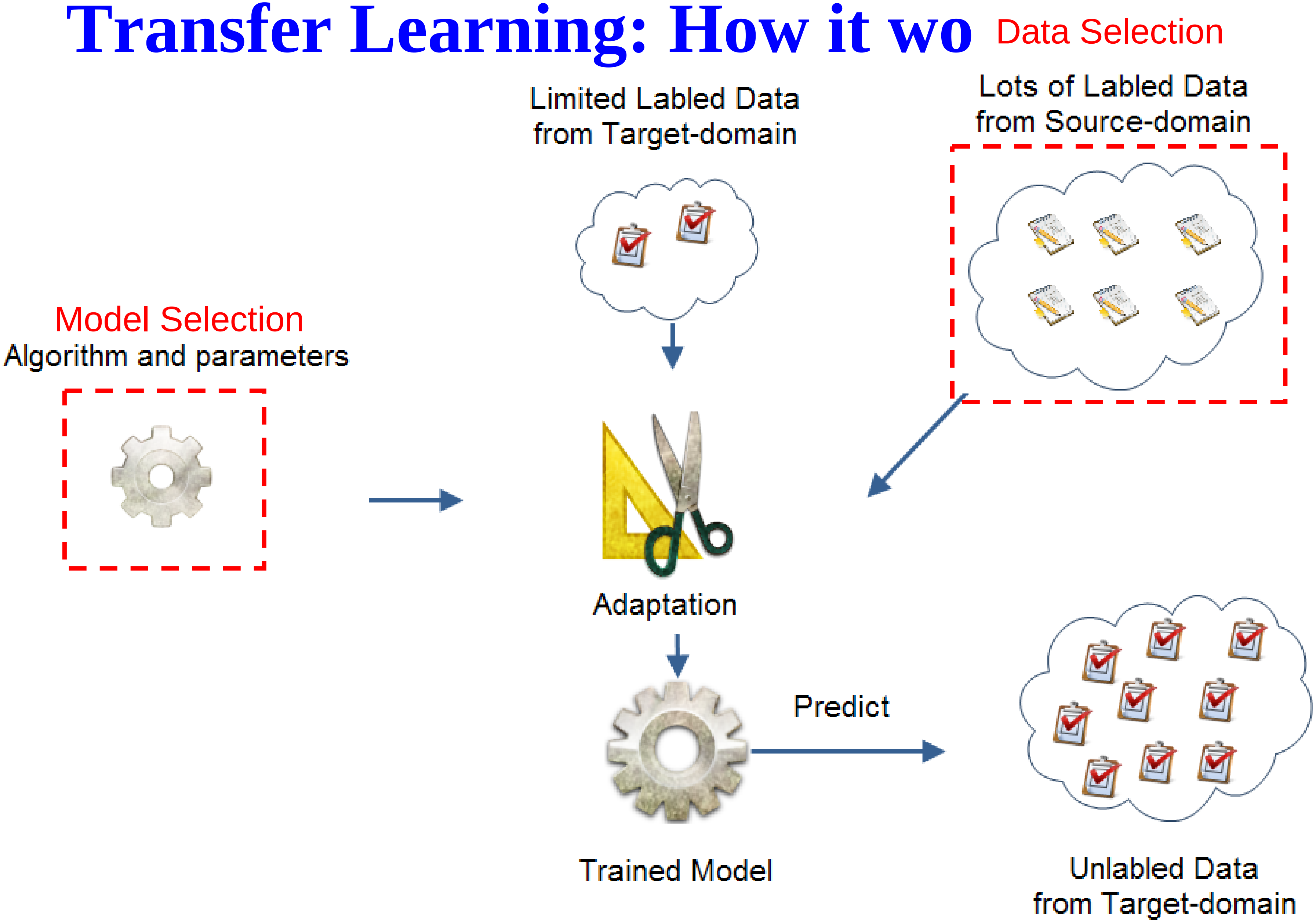
(Lx, Ly) is the coordinate of location.

Application

Collaborative Filtering



Transfer Learning: How it wo

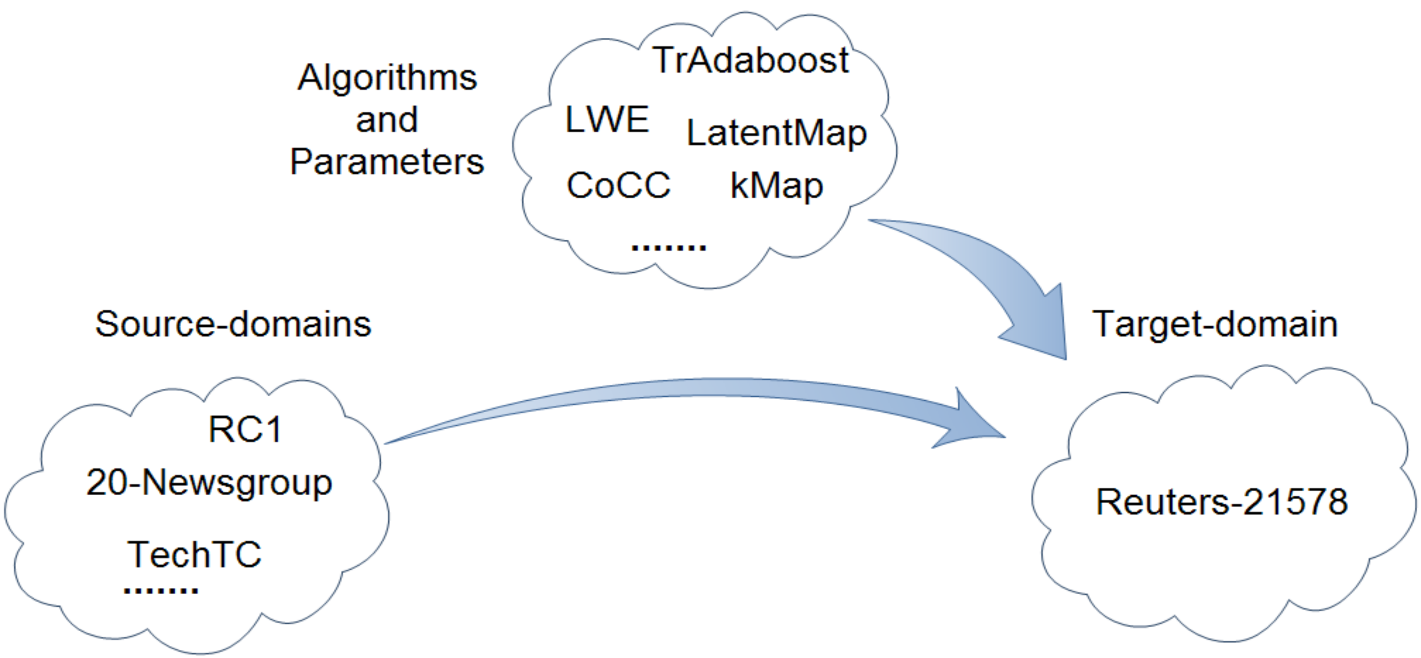


Re-cast: Model and Data Selection

(1) How to select the right transfer learning algorithms?

(2) How to tune the optimal parameters?

(3) How to choose the most helpful source-domain from a large pool of datasets?



Model & Data Selection Traditional Methods

1. Analytical techniques: AIC, BIC, SRM, etc.

$$\hat{f} = \arg\min_{f} \frac{1}{n} \sum_{\mathbf{x} \in X_s} \Big| P_s(y|\mathbf{x}) - P(y|\mathbf{x}, f) \Big| + \Theta_f$$

2. k-fold cross validation

$$\hat{f} = \arg\min_{f} \frac{1}{k} \sum_{j=1}^{k} \sum_{(\mathbf{x},y) \in S_j} \Big| P_s(y|\mathbf{x}) - P(y|\mathbf{x}, f_j) \Big|$$

Model & Data Selection Issuses

➡ $P_s(x) \neq P_t(x)$

The estimation is not consiste $\lim_{n \to \infty}(\hat{f}) \neq f^*$

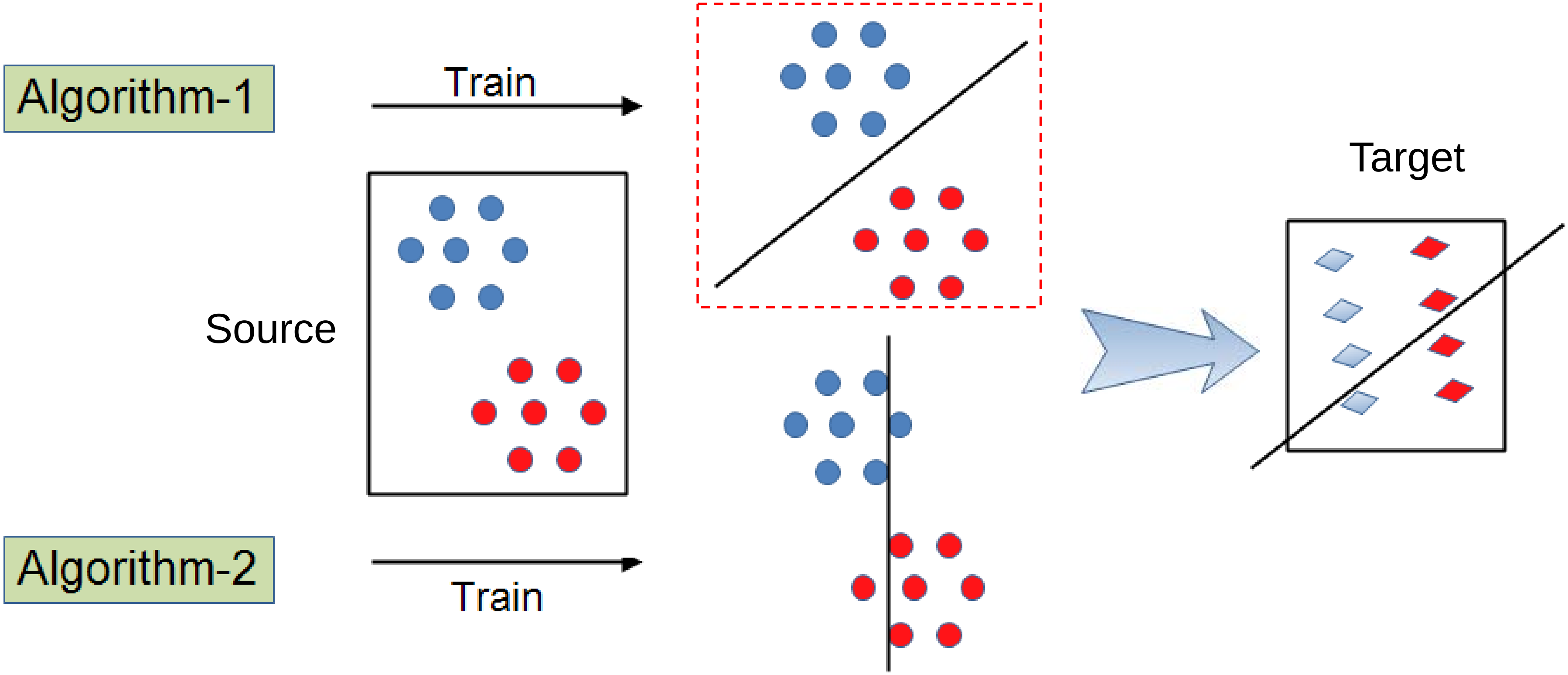
Ideal Hypothesis $f^* = \arg\min_{f} \mathbf{E}_{\mathbf{x} \sim P_t(\mathbf{x})} \left| P_t(y|\mathbf{x}) - P(y|\mathbf{x}, f) \right| + \Theta_f$

➡ $P_s(y \mid x) \neq P_t(y \mid x)$

A model approximating $P_s(y \mid x)$ is not necessarily close to $P_t(y \mid x)$

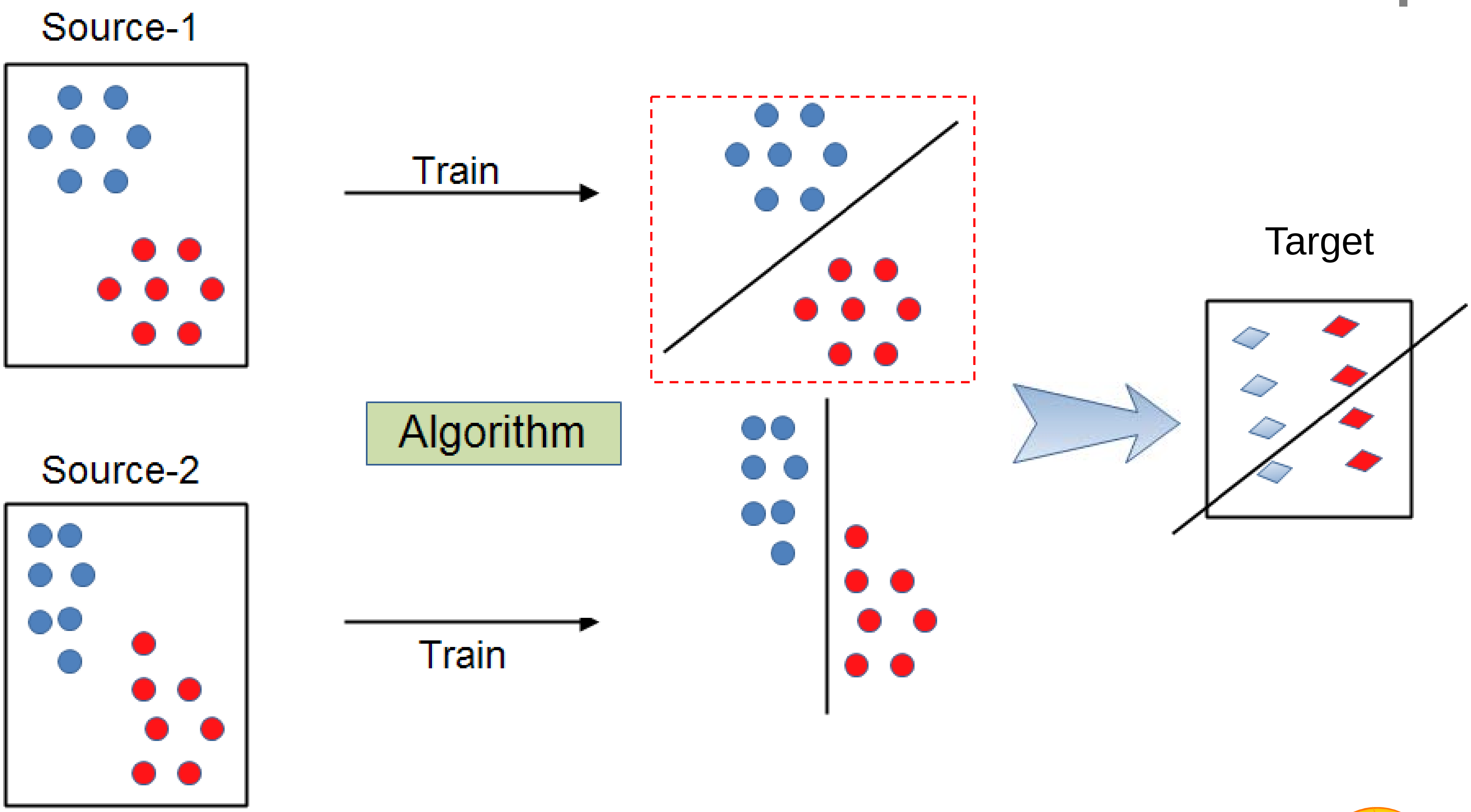
The number of labeled data in target domain is limited and thus the directly estimation of $P_t(y \mid x)$ is not reliable.

Model & Data Selection Model Selection Example



If we choose the wrong model....

Model & Data Selection Data Selection Example



If we choose the wrong source-domain....

Transfer Cross-Validation (TrCV)

Hard to calculate in practice

New criterion for transfer learning

$$\hat{f} = \arg\min_{f} \frac{1}{n} \sum_{\mathbf{x} \in X_s} \frac{P_t(\mathbf{x})}{P_s(\mathbf{x})} \left| P_t(y|\mathbf{x}) - P(y|\mathbf{x}, f) \right|$$

How to calculate this difference with limited labeled data?

Reverse Validation

...n between...

...en the conditional distribution

estima... and the true conditional distribution.

Practical method: Transfer Cross-Validation (TrCV)

$$\hat{f} = \arg\min_{f} \frac{1}{k} \sum_{j=1}^{k} \sum_{(\mathbf{x},y) \in S_j} \frac{P_t(\mathbf{x})}{P_s(\mathbf{x})} \left| P_t(y|\mathbf{x}) - P(y|\mathbf{x}, f) \right|$$

Density Ratio Weighting

Density Ratio Weighting

- The selected model is an unbiased estimator to the $\hat{f}$ ideal model f^*

Lemma 1. *$\ell_w(\hat{f}) + \Theta_{\hat{f}} = \ell^*(f^*) + \Theta_{f^*}$, when $n \to \infty$ and f^* and $\hat{f}$ belong to the same hypothesis class.*

$\ell^*(f^*)$ is the expected loss to approximate $P_t(y|\mathbf{x})$

$$\ell_w(\hat{f}) = \frac{1}{n} \sum_{\mathbf{x} \in X_s} \frac{P_t(\mathbf{x})}{P_s(\mathbf{x})} \left| P_t(y|\mathbf{x}) - P(y|\mathbf{x}, \hat{f}) \right|$$

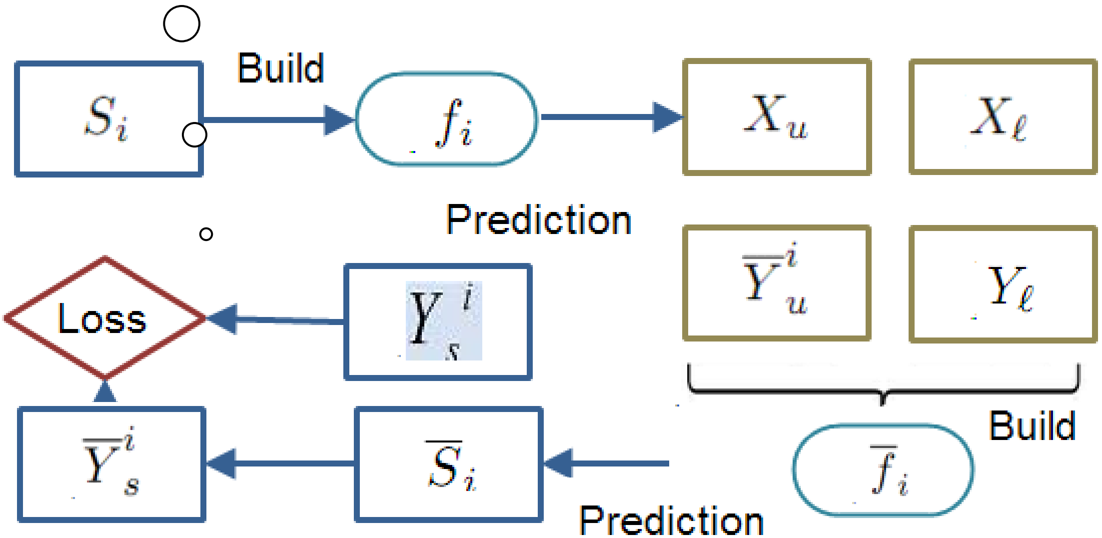
Θ_f is the model complexity

Important property to choose the right model even when P(x) and P(y|x) are different

- We adopt an existing method KMM (Huang et al'07) for density ratio weighting
- Reverse Validation to estimate $P_t(y|x) - P(y|x,f)$ (next slide)

$$|P_t(y|\mathbf{x}) - P(y|\mathbf{x}, f_i)|$$

Reverse Validation



S_i The source-domain data in i-th fold

$\overline{S}_i$ The remaining data

$\overline{Y}_u^i$ The predicted label of X_u in i-th fold

$\overline{Y}_s^i$ The predicted label of S_i in i-th fold

Y_s^i The true label of S_i in i-th fold

X_u X_ℓ The unlabeled and labeled target-domain data

Properties

- The selected model is an unbiased estimator to the ideal one. [Lemma 1]
- The model selected by the proposed method has a generalization bound over target-domain data. [Theorem 1]
- The value of reverse validati $r(\mathbf{x})$ is related to the difference between true conditional probability and mod $|P(y|x, f_i) - P_t(y|\mathbf{x})|$
- The confidence of TrCV has a bound.

$$Pr\left\{ -z < \frac{\varepsilon_u(f) - \varepsilon(f)}{\sqrt{\varepsilon(f) \cdot (1 - \varepsilon(f))/n}} < z \right\} \approx \lambda$$

$\varepsilon_u(f)$ the accuracy estimated by TrCV

$\varepsilon(f)$ the true accuracy of f

z $(1+\lambda)/2\text{-th}$ quantile point of the standard normal distribution

Experiment **Data Set**

- Wine Quality: two subsets related to red and white variants of the Portuguese “Vinho Verde” wine.

Data Set	$\lvert S \rvert$	$\lvert T \rvert$	Description
Red-White(RW)	1599	4998	physicochemical variables
White-Red(WR)	4998	1599	

For algorithm and parameters selection

Experiment Data Set

- Reuters-21578:the primary benchmark of text categorization formed by different news with a hierarchial structure.

Data Set	$\|S\|$	$\|T\|$	Description
orgs vs. people(ope)	1016	1046	Documents from different subcategories
orgs vs. places(opl)	1079	1080	
people vs. places(pp)	1239	1210	

For algorithm and parameters selection

Experiment Data Set

- SyskillWebert: the standard dataset used to test web page ratings, generated by the HTML source of web pages plus the user rating. we randomly reserve “Bands-recording artists” as source-domain and the three others as target-domain data.

Data Set	$\|S\|$	$\|T\|$	Description
Sheep(Sp)	61	65	Web pages with different contents
Biomedical(Bl)	61	131	
Goats(Gs)	61	70	

For algorithm and parameters selection

Experiment **Data Set**

- 20-Newsgroup: primary benchmark of text categorization similar to Reuters-21578

Data Set	S	T	$\|S\|$	$\|T\|$
comp	windows vs. motorcycles	graphics	1596	
vs.	pc.hardware vs. baseball	vs.	1969	1957
rec	mac.hardware vs. hockey	autos	1954	
sci	crypt vs. guns	electronics	1895	
vs.	med vs. misc	vs.	1761	1924
talk	space vs. religion	mideast	1612	

For source-domain selection

Experiment Baseline methods

- SCV: standard k-fold CV on source-domain
- TCV: standard k-fold CV on labeled data from target-domain
- STV: building a model on the source-domain data and validating it on labeled target-domain data
- WCV: using density ratio weighting to reduce the difference of marginal distribution between two domains, but ignoring the difference in conditional probability.

$$\hat{f} = \arg\min_{f} \frac{1}{k} \sum_{j=1}^{k} \sum_{(\mathbf{x},y)\in S_j} \frac{P_t(\mathbf{x})}{P_s(\mathbf{x})} \left| \boxed{P_s(y|\mathbf{x})} - P(y|\mathbf{x}, f_j) \right|$$

Experiment Other settings

- Algorithms:
  - Naive Bayes(NB), SVM, C4.5, K-NN and NNge(Ng)
  - TrAdaBoost(TA): instances weighting [Dai et al.'07]
  - LatentMap(LM): feature transform [Xie et al.'09]
  - LWE : model weighting ensemble [Gao et al.'08]

- Evaluation: if one criterion can select the better model in the comparison, it gains a higher measure value.

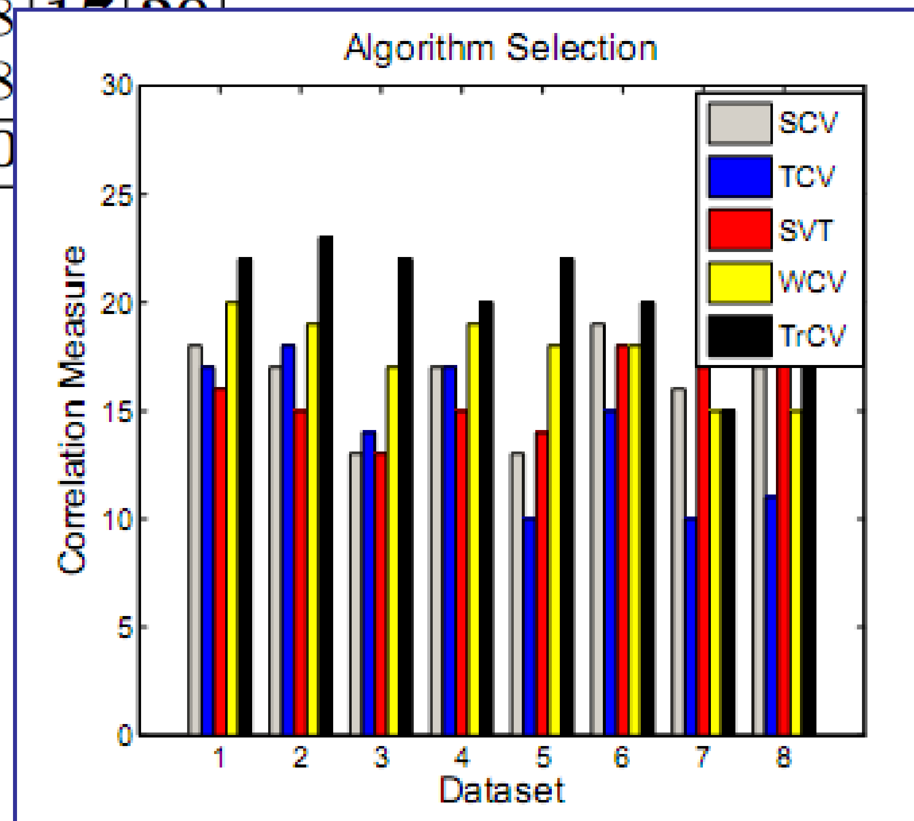
$$corr = C^2_{|\mathcal{H}|} - \sum_{f,g \in \mathcal{H}} \left[\Big(\varepsilon(f) - \varepsilon(g) \Big) \times \Big(v(f) - v(g) \Big) < 0 \right]$$

$\varepsilon(\cdot)$ and $v(\cdot)$ The accuracy and value of criteria (e.g TrCV, SCV, etc)

$C^2_{|\mathcal{H}|}$ The number of comparisions between models

Results Algorithm Selection

Method	RW	WR	ope	opl	pp	Sp	Bl	Gs
	Algorithm Selection							
SCV	18	17	13	17	13	19	16	17
TCV	17	18	14	17	10	15	10	11
STV	16	15	13	15	14	18	[illegible]	[illegible]
WCV	20	19	17	19	18	18	[illegible]	[illegible]
TrCV	**22**	**23**	**22**	**20**	**22**	**20**	[illegible]	[illegible]

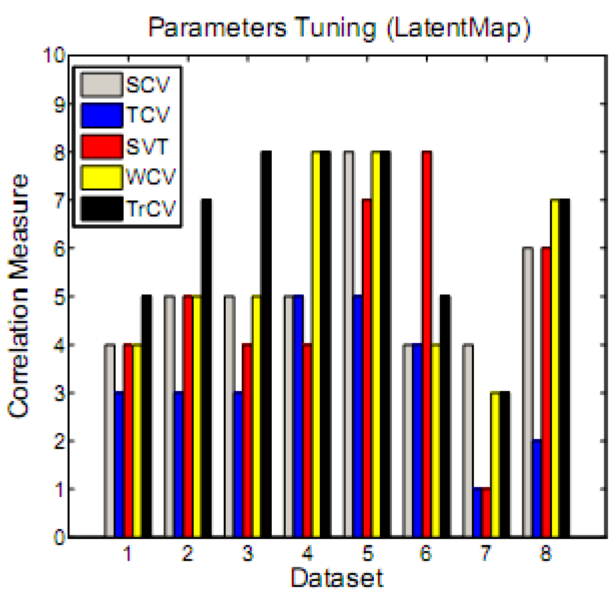


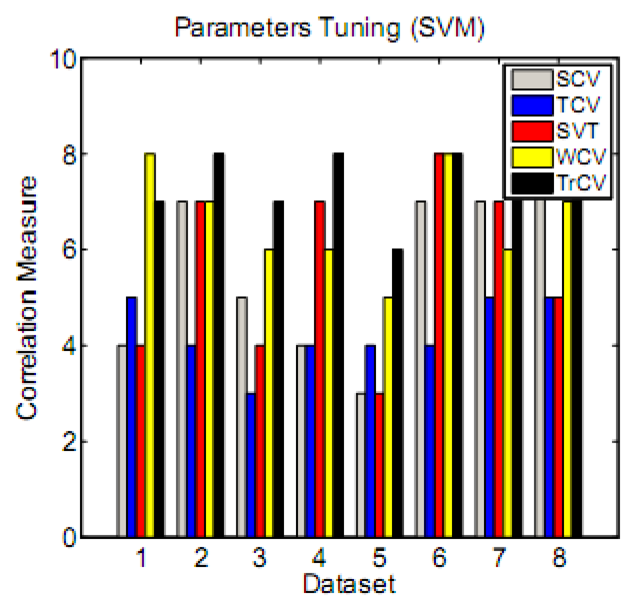
6 win and 2 lose!

Results Parameter Tuning

Method	RW	WR	ope	opl	pp	Sp	Bl	Gs	RW	WR	ope	opl	pp	Sp	Bl	Gs
	Parameter Tuning (LatentMap)								Parameter Tuning (SVM)							
SCV	4	5	5	5	**8**	4	**4**	6	4	7	5	4	3	7	7	**8**
TCV	3	3	3	5	5	4	1	2	5	4	3	4	4	4	5	5
STV	4	5	4	4	7	**8**	1	6	4	7	4	7	3	**8**	7	5
WCV	4	5	5	**8**	**8**	4	3	**7**	**8**	7	6	6	5	**8**	6	7
TrCV	**5**	**7**	**8**	**8**	**8**	5	3	**7**	7	**8**	**7**	**8**	**6**	**8**	**8**	**8**

13 win and 3 lose!

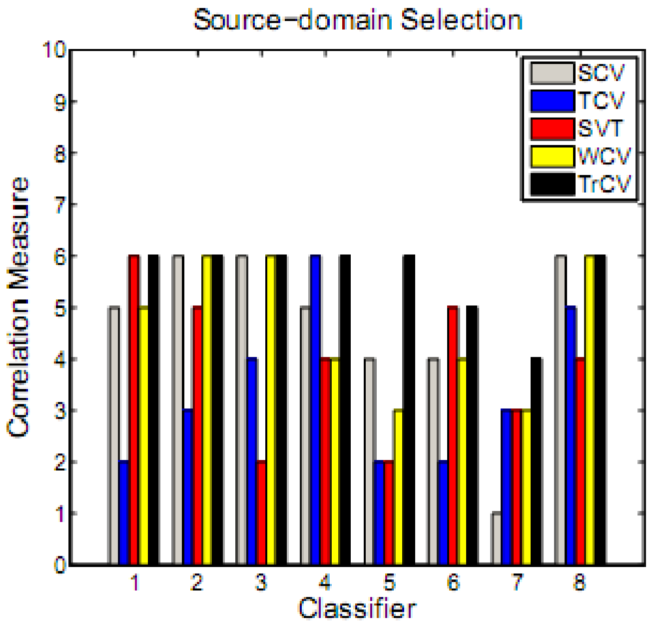




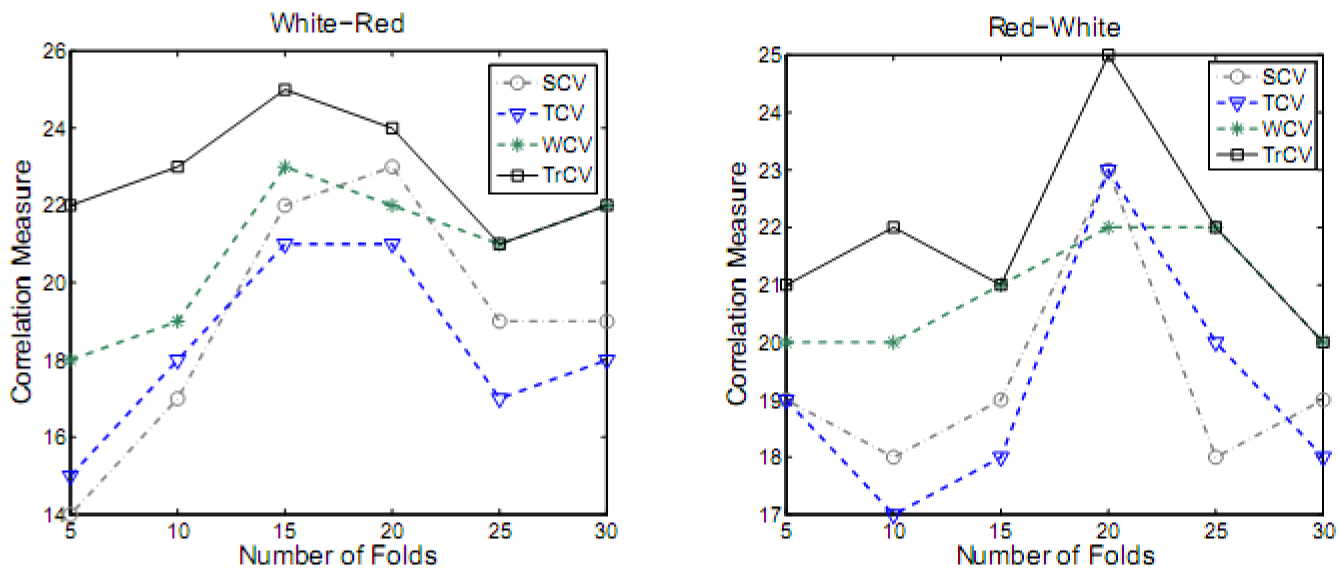
Results Source-domain Selection

Method	NB	SVM	C45	KNN	Ng	TA	LM	LWE	Pr
SCV	5	**6**	**6**	5	4	4	1	**6**	436
STV	2	3	4	**6**	2	2	3	5	371
TCV	**6**	5	2	4	2	**5**	3	4	399
WCV	5	**6**	**6**	4	3	4	3	**6**	442
TrCV	**6**	**6**	**6**	**6**	**6**	**5**	**4**	**6**	**512**

No lose!



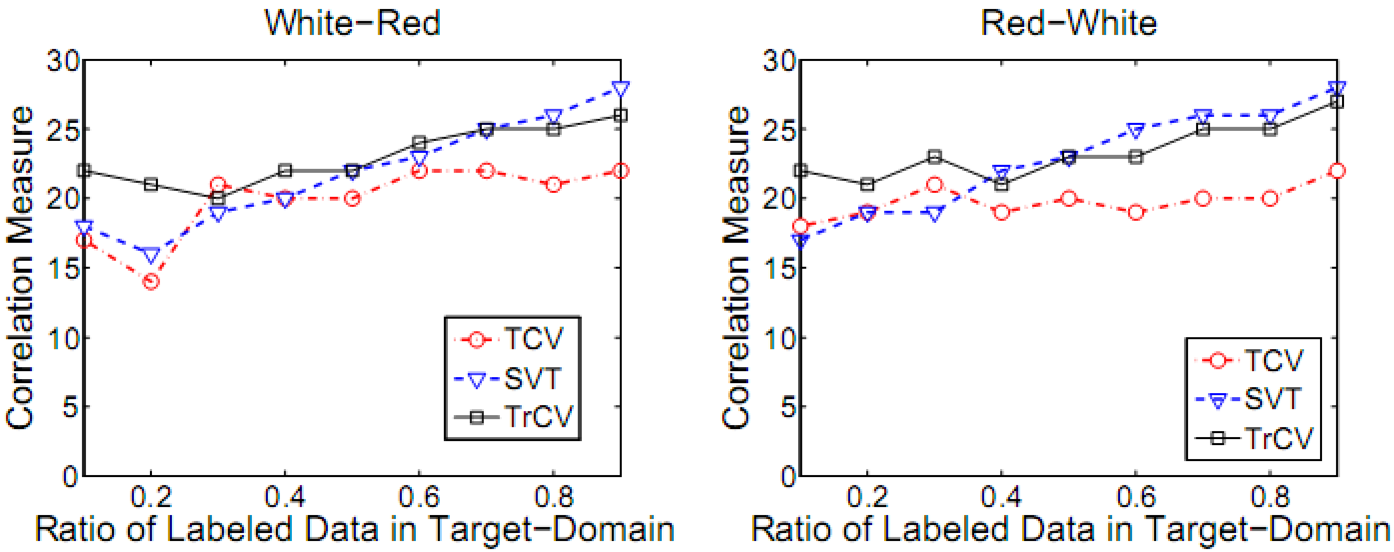
Results Parameter Analysis



(a) Different number of folds

TrCV achieves the highest correlation value under different number of folds from 5 to 30 with step size 5 .

Results Parameter Analysis



(b) Different number of labeled data in T

When only a few labeled data($< 0.4 \times |T|$) can be obtained in the target-domain, the performance of TrCV is much better than both SVT and TCV.

Conclusion

- Model and data selection when margin and conditional distributions are different between two domains.
- Key points
  - Point-1 Density weighting to reduce the difference between marginal distributions of two domains;
  - Point-2 Reverse validation to measure how well a model approximates the true conditional distribution of target-domain.
- Code and data available from the authors
  - www.weifan.info


Thank You!

Thanks!