

# Outline of this section

- 1 Overview of Transfer Learning
- 2 Categorization
- 3 Applications
- 4 Conclusion

# What is Transfer Learning? [Pan and Yang, 2010]

## Naive View (Transfer Learning)

*Transfer Learning (i.e. Knowledge Transfer, Domain Adaption) aims at applying knowledge learned **previously** to solve **new** problems faster or with better solutions.*

# What is Transfer Learning? [Pan and Yang, 2010]

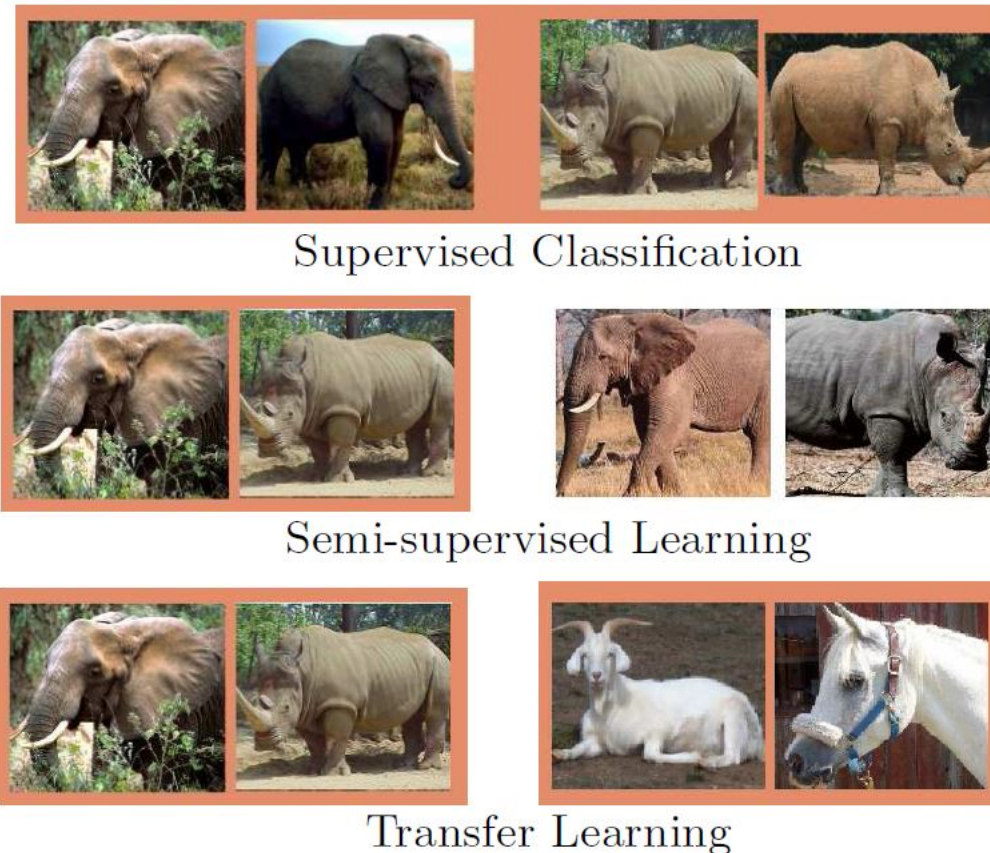
## Naive View (Transfer Learning)

*Transfer Learning (i.e. Knowledge Transfer, Domain Adaption) aims at applying knowledge learned **previously** to solve **new** problems faster or with better solutions.*

# Transfer Learning

- What to “Transfer” ?
- How to “Transfer” ?
- When to “Transfer” ?
- Machine Learning Scheme.
- Relationship with other ML tech?

# What is Transfer Learning? (Cont.)



**Figure :** Supervised classification uses labeled examples of elephants and rhinos; semi-supervised learning uses additional unlabeled examples of elephants and rhinos; transfer learning uses additional labeled datasets[Raina et al., 2007].

# Motivation

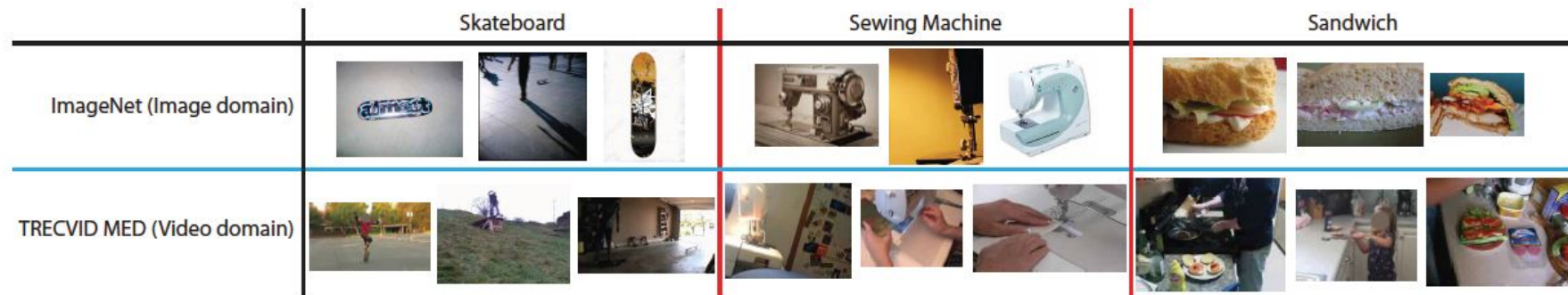
Why we need Transfer Learning[Tang et al., 2012]?

- **Labeled data are expensive and limited.**
- **Related data are cheap and sufficient.**

# Motivation

Why we need Transfer Learning[Tang et al., 2012]?

- **Labeled data are expensive and limited.**
- **Related data are cheap and sufficient.**



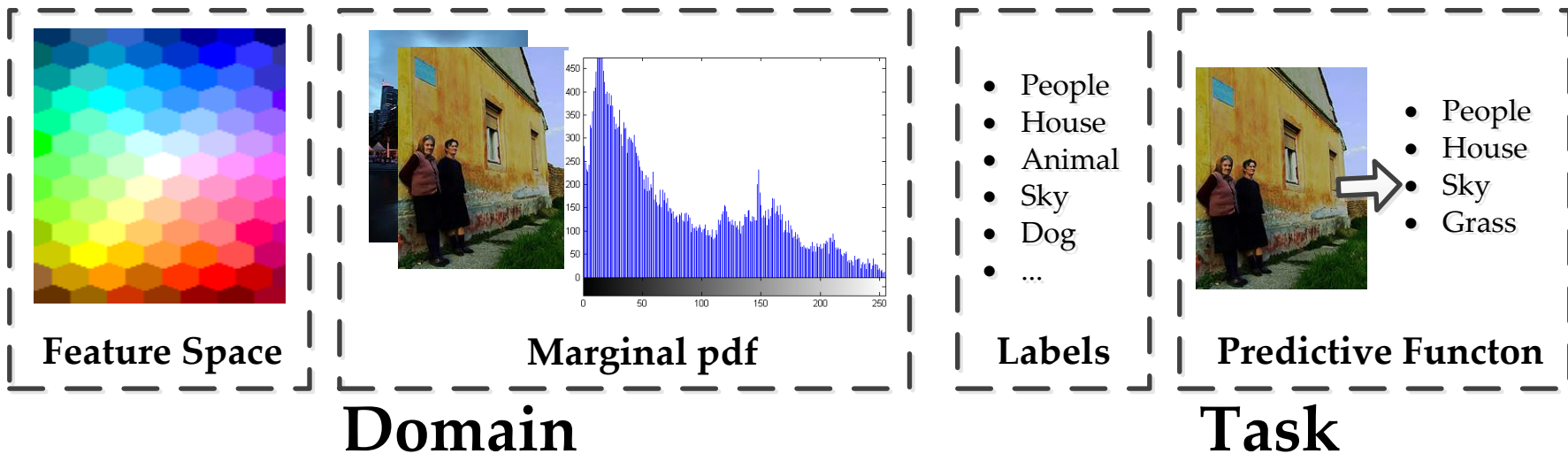
**Figure :** Object detector for static image is easy to obtain. However, the labeled data for video task are limited and expensive.

# Terminologies

- **Domain:** A domain  $\mathcal{D} = \{\mathcal{X}, P(X)\}$  consists of two components: a feature space  $\mathcal{X}$  and a marginal prob distribution  $P(X), X \in \mathcal{X}$ .
- **Task:** Given a specific domain, a task  $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$  also consists of two components: a label space  $\mathcal{Y}$  and the predictive function  $f(\cdot) = P(y|x)$ . *The predictive is unknown for us but can be learned from training data, which consists of data pair  $(x_i, y_i)$ .*

# Terminologies

- **Domain:** A domain  $\mathcal{D} = \{\mathcal{X}, P(X)\}$  consists of two components: a feature space  $\mathcal{X}$  and a marginal prob distribution  $P(X), X \in \mathcal{X}$ .
- **Task:** Given a specific domain, a task  $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$  also consists of two components: a label space  $\mathcal{Y}$  and the predictive function  $f(\cdot) = P(y|x)$ . *The predictive is unknown for us but can be learned from training data, which consists of data pair  $(x_i, y_i)$ .*





## Terminologies (Cont.)

- **Source/Target Domain Data:** a set of labeled data  $D_S$  and  $D_T$

$$D_S = \{(x_{S_1}, y_{S_1}) \dots (x_{S_{n_S}}, y_{S_{n_S}})\}, D_T = \{(x_{T_1}, y_{T_1}) \dots (x_{T_{n_T}}, y_{T_{n_T}})\}$$

In most cases,  $0 \leq n_T \ll n_S$ .

## Terminologies (Cont.)

- **Source/Target Domain Data:** a set of labeled data  $D_S$  and  $D_T$

$$D_S = \{(x_{S_1}, y_{S_1}) \dots (x_{S_{n_S}}, y_{S_{n_S}})\}, D_T = \{(x_{T_1}, y_{T_1}) \dots (x_{T_{n_T}}, y_{T_{n_T}})\}$$

In most cases,  $0 \leq n_T \ll n_S$ .

### Definition (Transfer Learning)

*Given a source domain  $\mathcal{D}_S$  and learning task  $\mathcal{T}_S$ , a target domain  $\mathcal{D}_T$  and learning task  $\mathcal{T}_T$ , **transfer learning** aims to help improve the learning of the target predictive function  $f_T(\cdot)$  in  $\mathcal{D}_T$  using the knowledge in  $\mathcal{D}_S$  and  $\mathcal{T}_S$ , where  $\mathcal{D}_S \neq \mathcal{D}_T$  or  $\mathcal{T}_S \neq \mathcal{T}_T$ .*

# Transfer of Learning

A psychological point of view

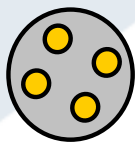
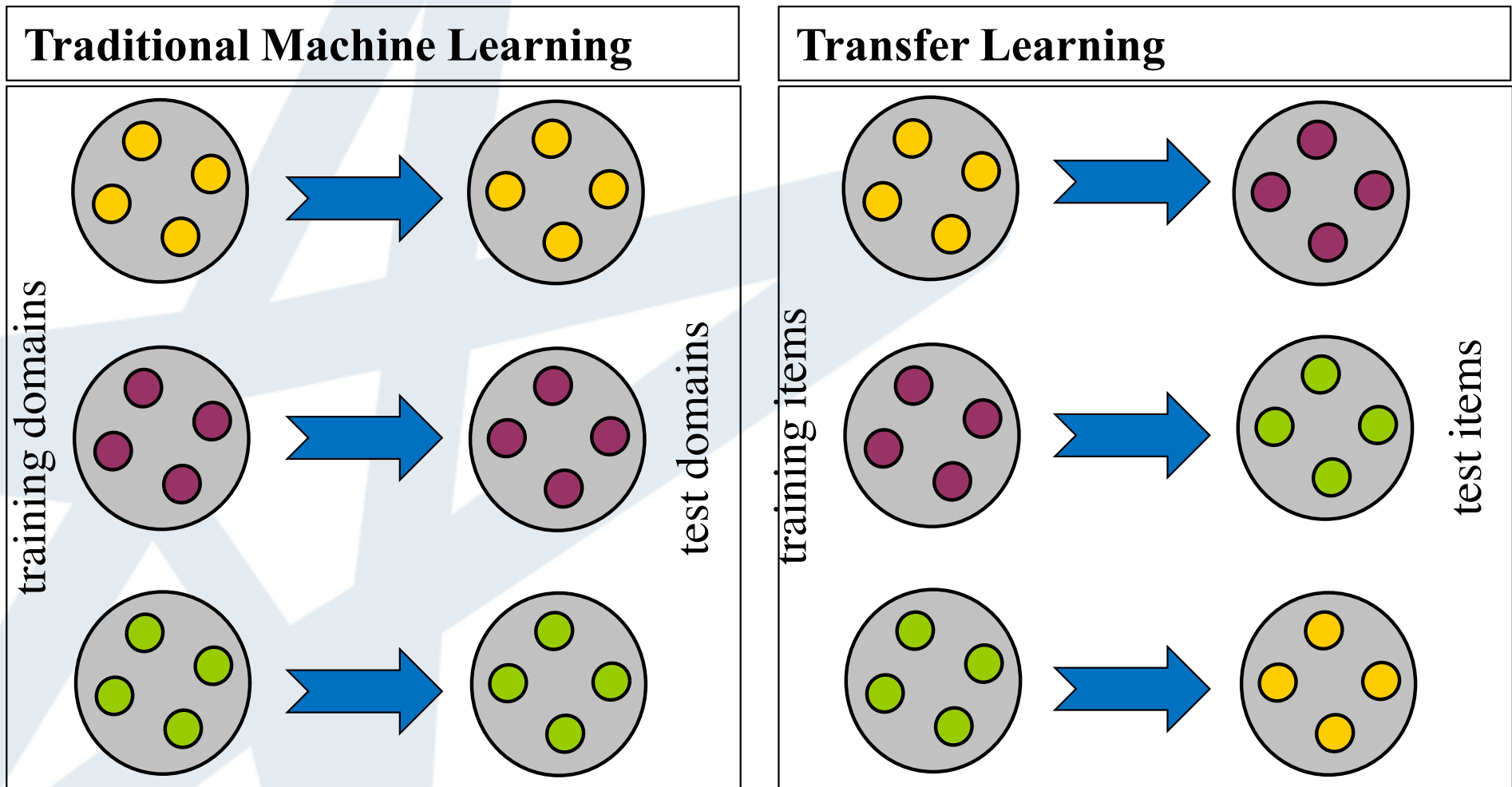
- The study of dependency of human conduct, learning or performance on prior experience.
  - [Thorndike and Woodworth, 1901] explored how individuals would transfer in one context to another context that share similar characteristics.
  - C++ → Java
  - Maths/Physics → Computer Science/Economics

# Transfer Learning

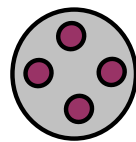
In the machine learning community

- 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?

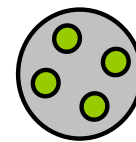
# Transfer Learning



domain A



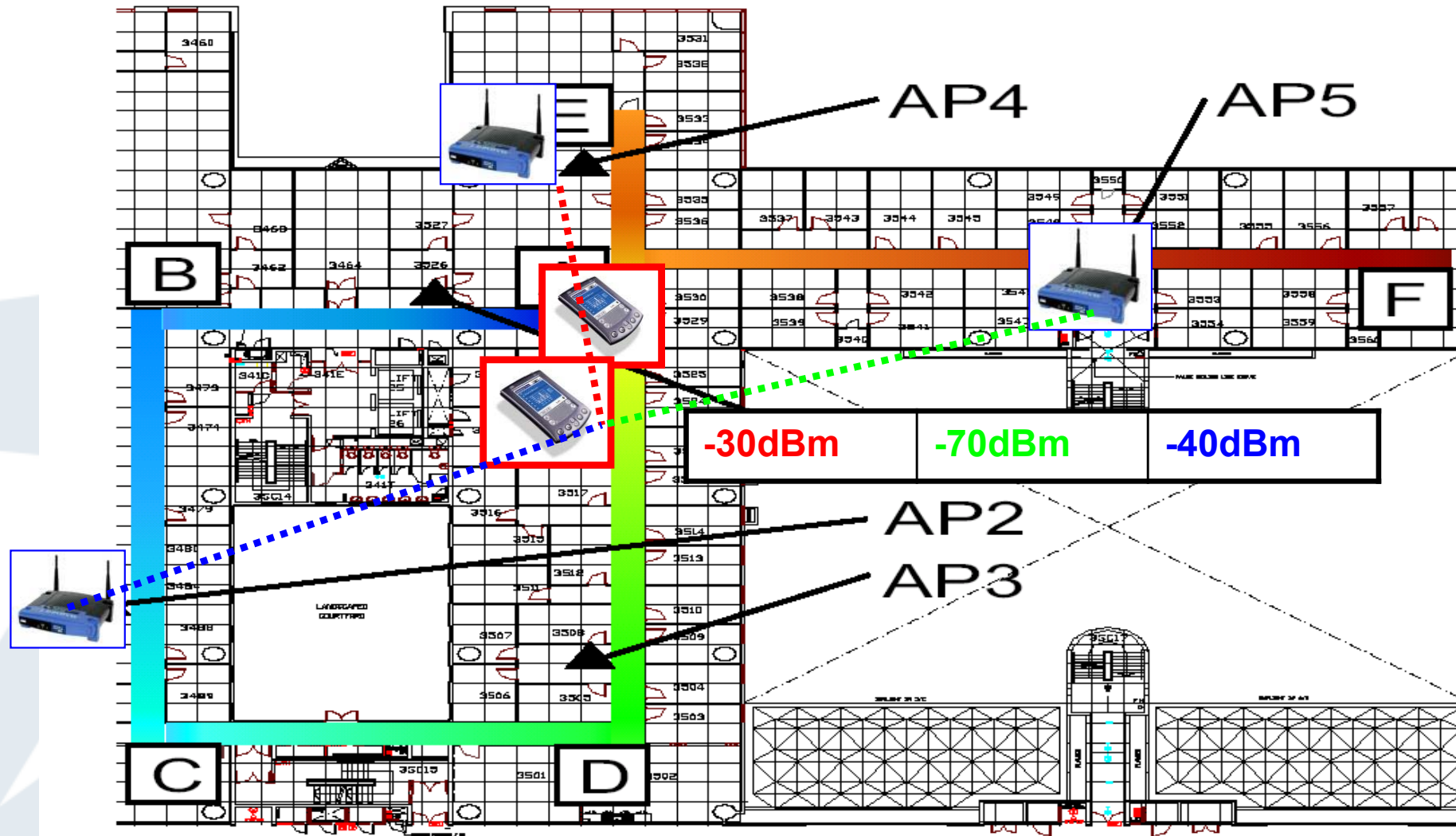
domain B



domain C

# Motivating Example I:

## Indoor WiFi localization



# Indoor WiFi Localization (cont.)

**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)



Device A

**Localization model**

**Test**

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



Device A

Average Error Distance

~ 1.5 meters



**Training**

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



Device B

**Localization model**

**Test**

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




Device A


~10 meters

# 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.)

**Training**

10 hours ago  
Edward Prtiz 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 Prtiz 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



Classification Accuracy

~ 84.6%



**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 Prtiz 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%

# Difference between Domains

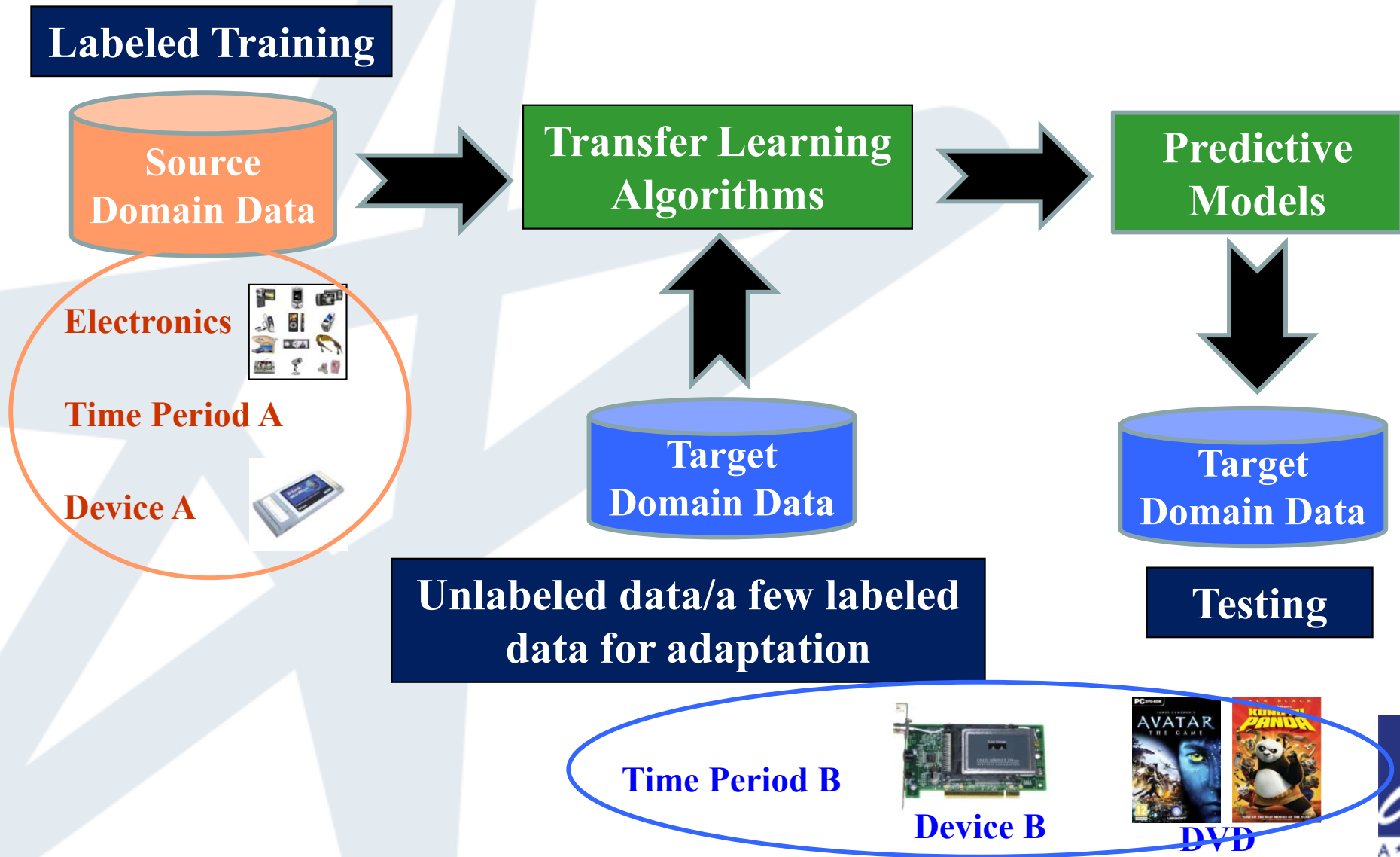


Electronics	Video Games
(1) <b>Compact</b> ; easy to operate; very good picture quality; looks <b>sharp</b> !	(2) A very good game! It is action packed and full of excitement. I am very much <b>hooked</b> on this game.
(3) I purchased this unit from Circuit City and I was very excited about the quality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

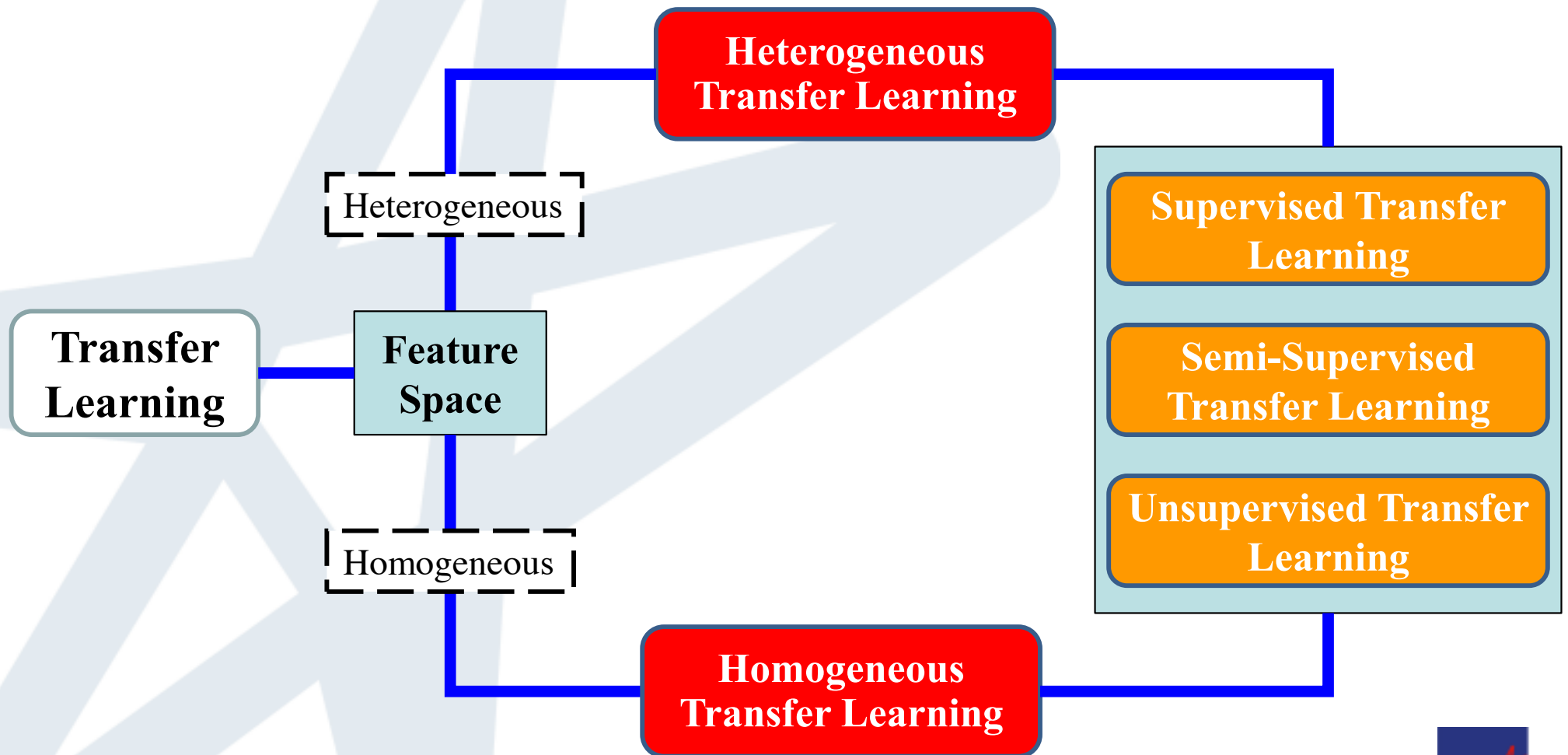
# How to Build Systems on Each Domain of Interest

- 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**

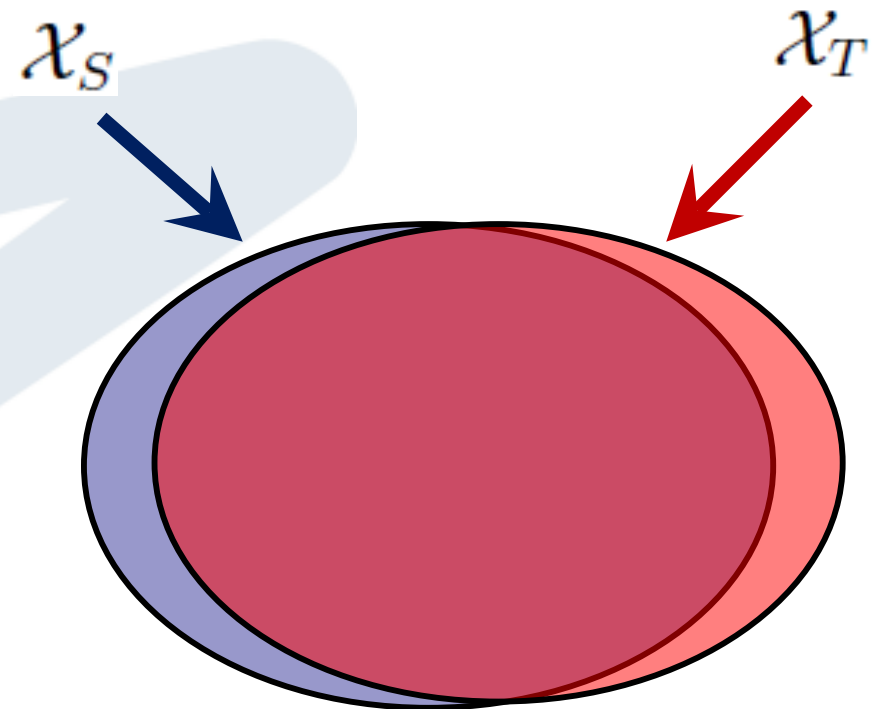
**Parameter-based  
Approaches**

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

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

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

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

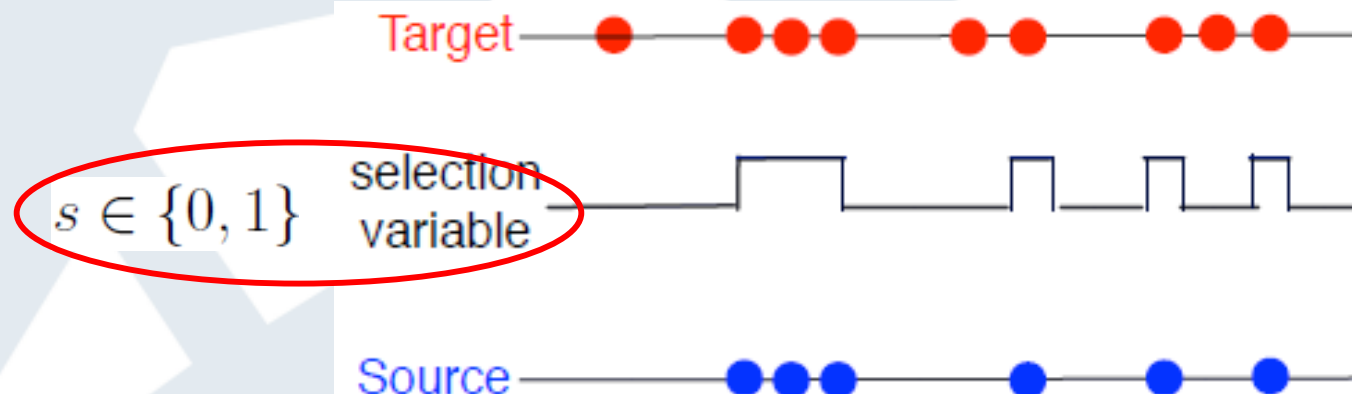
- $\mathcal{Y}_S = \mathcal{Y}_T$ ,  
but  $f_S \neq f_T$  ( $P_S(y|x) \neq P_T(y|x)$ ).



# Instance-based Approaches

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

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



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

[Zadrozny, ICML-04]

- Labeled instances from the source domain with label 1
- Unlabeled instances from the target domain with label 0
- Train a binary classifier

# Instance-based Approaches

correcting sample selection bias(cont.)

- Use  $\beta(x)$  to train a binary classifier. The classifier is regarded as the selection variable mentioned above.
- Put the labeled data in source domain and the unlabeled data in target domain together. Use the classifier to resample the data.

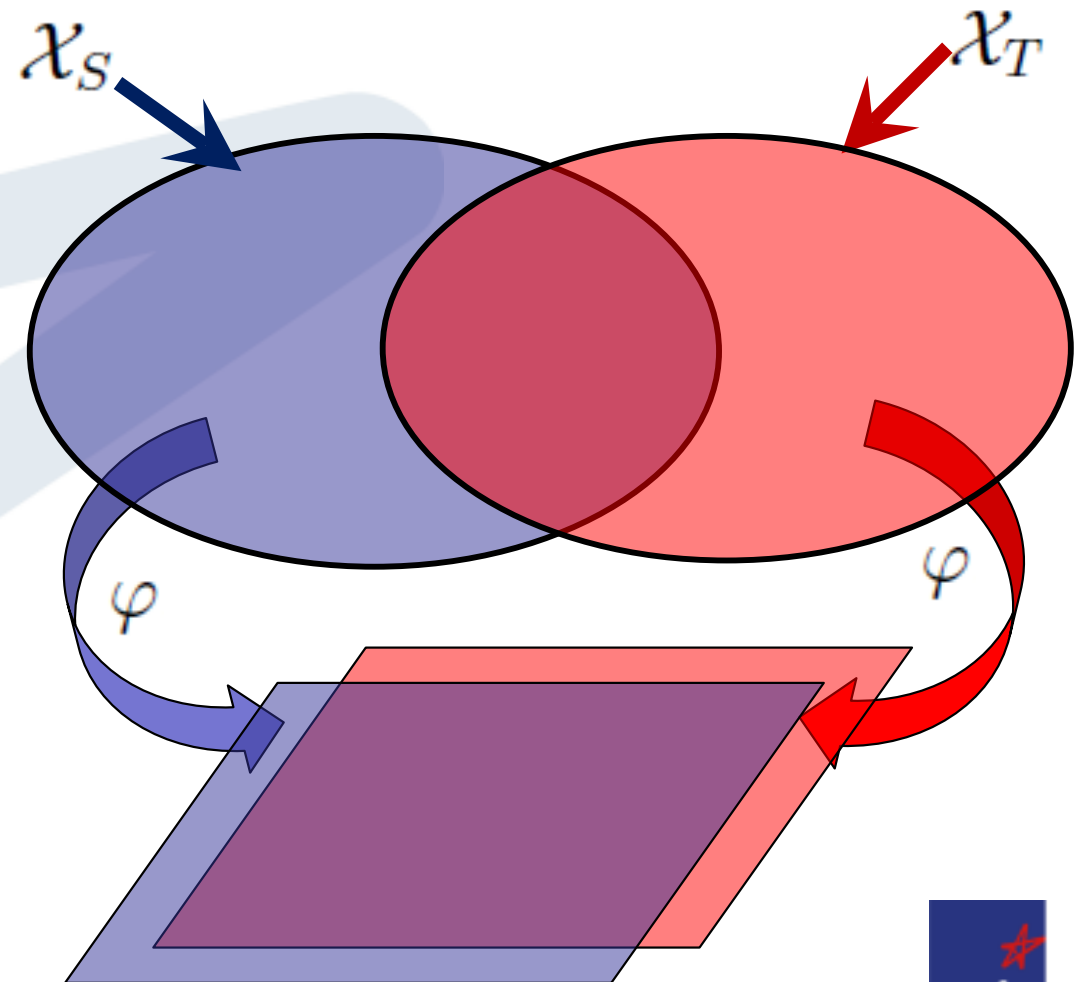
# Instance-based Approaches

correcting sample selection bias(cont.)

- Now we get a new sample subset of source domain data. The sample selection bias has been corrected in the new subset. Data in the sample subset and the target domain follow the same distribution.
- We can train our model on the sample subset of source domain data. Then directly use the model on target 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 Approaches

Encode application-specific knowledge



(1) **Compact**; easy to operate; very good picture quality; looks **sharp**!



(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**.



(5) It is also quite **blurry** in very dark settings. I will never\_buy HP again.

## Video Games

(2) A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.




(4) Very **realistic** shooting action and good plots. We played this and were **hooked**.

(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

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

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

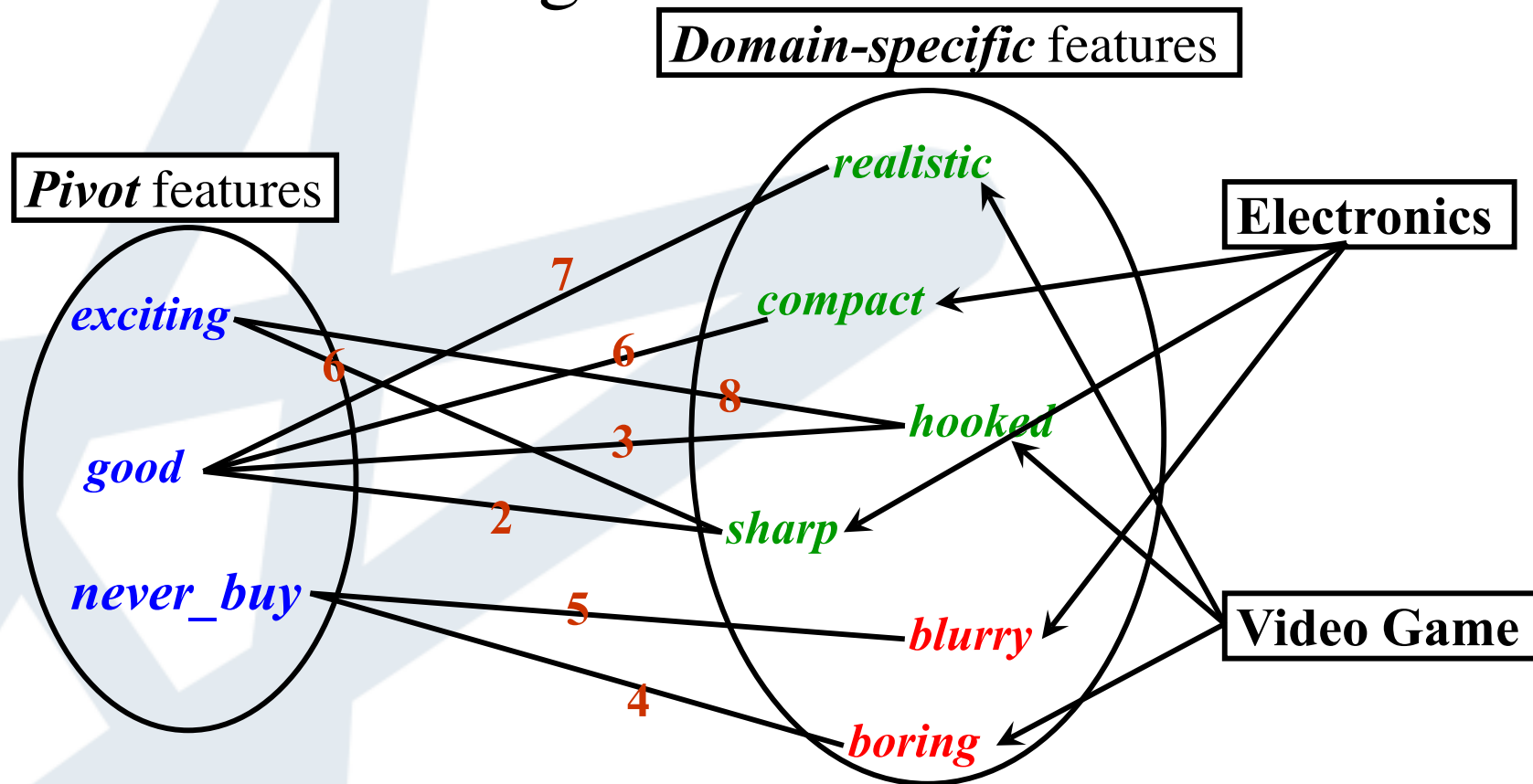
## Spectral Feature Alignment (SFA)

### ➤ 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

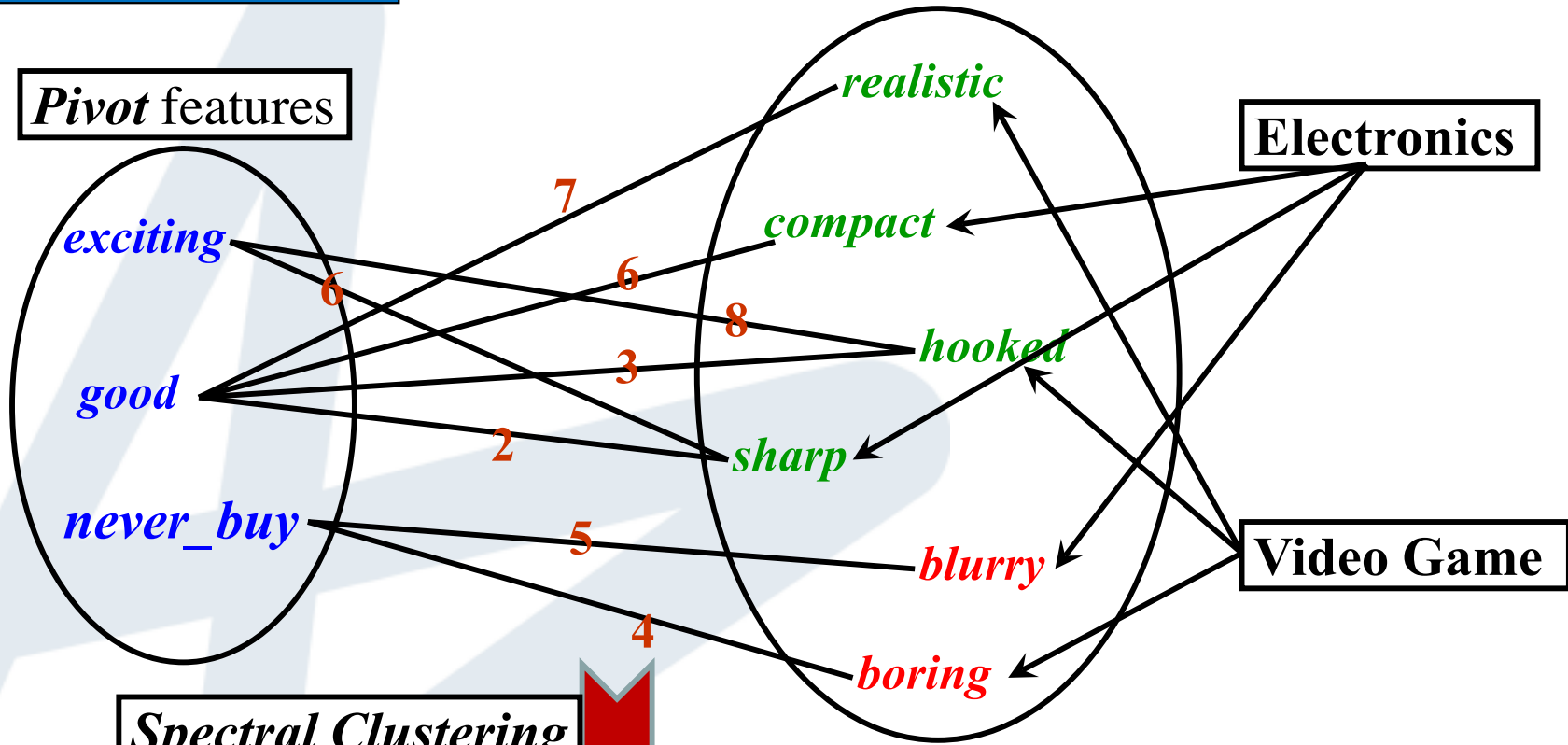


- If two *domain-specific* words have connections to more common *pivot* words in the graph, they tend to be aligned or clustered together with a higher probability.
- If two *pivot* words have connections to more common *domain-specific* words in the graph, they tend to be aligned together with a higher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

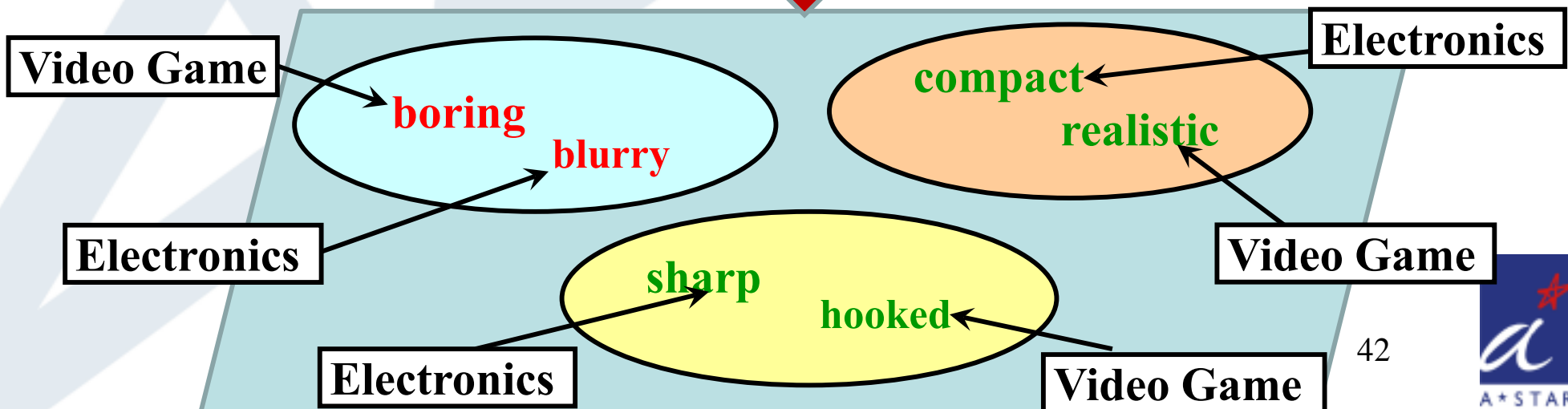
# Derive new features

## Domain-specific features

### Pivot features






### Spectral Clustering



# 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

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

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)

# Parameter-based Transfer Learning Approaches

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

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

$$\theta_T^* = \arg \min \sum_{i=1}^{n_T} l(x_{T_i}, y_{T_i}, \theta_T) + \lambda \Omega(\theta_T)$$

Tasks are learned independently

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

# 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

$$\begin{aligned}\theta_S &= \theta_0 + v_S \\ \theta_T &= \theta_0 + v_T\end{aligned}$$

Specific part for individual task

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



# Relational Transfer Learning

## Approaches

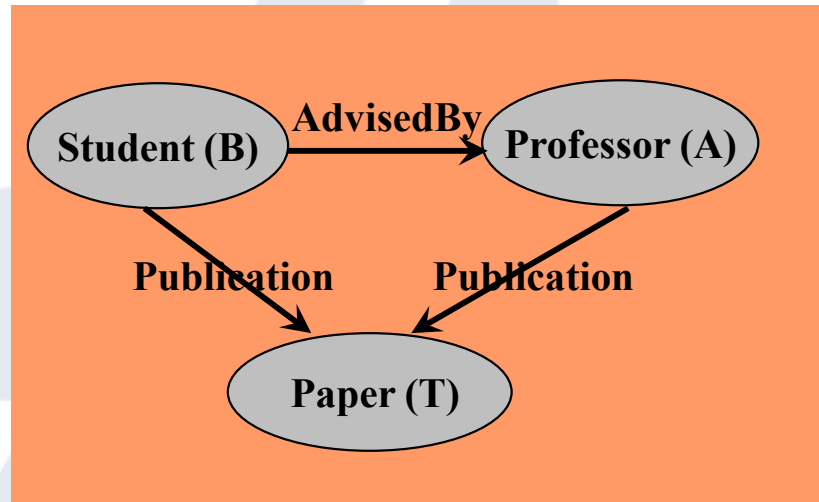
- **Motivation:** If two relational domains (data is non-i.i.d) are related, they may share some similar relations among objects. These relations can be used for knowledge transfer across domains.

# Relational Transfer Learning

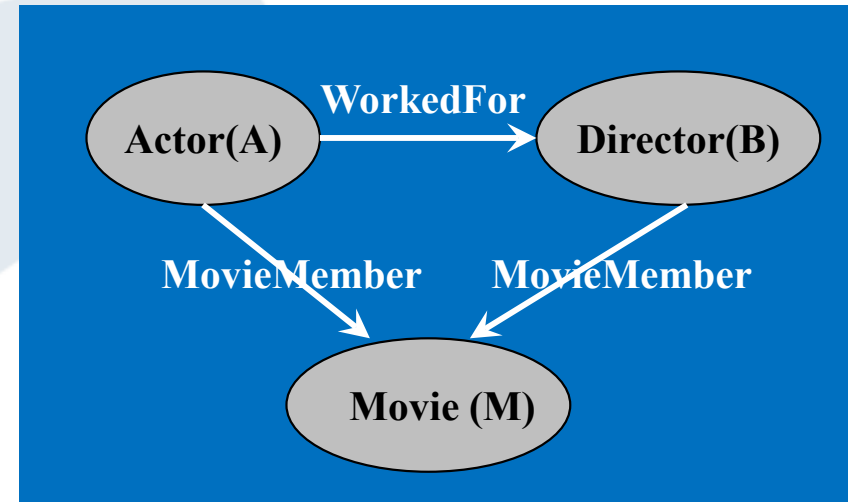
## Approaches (cont.)

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

Academic domain (source)



Movie domain (target)

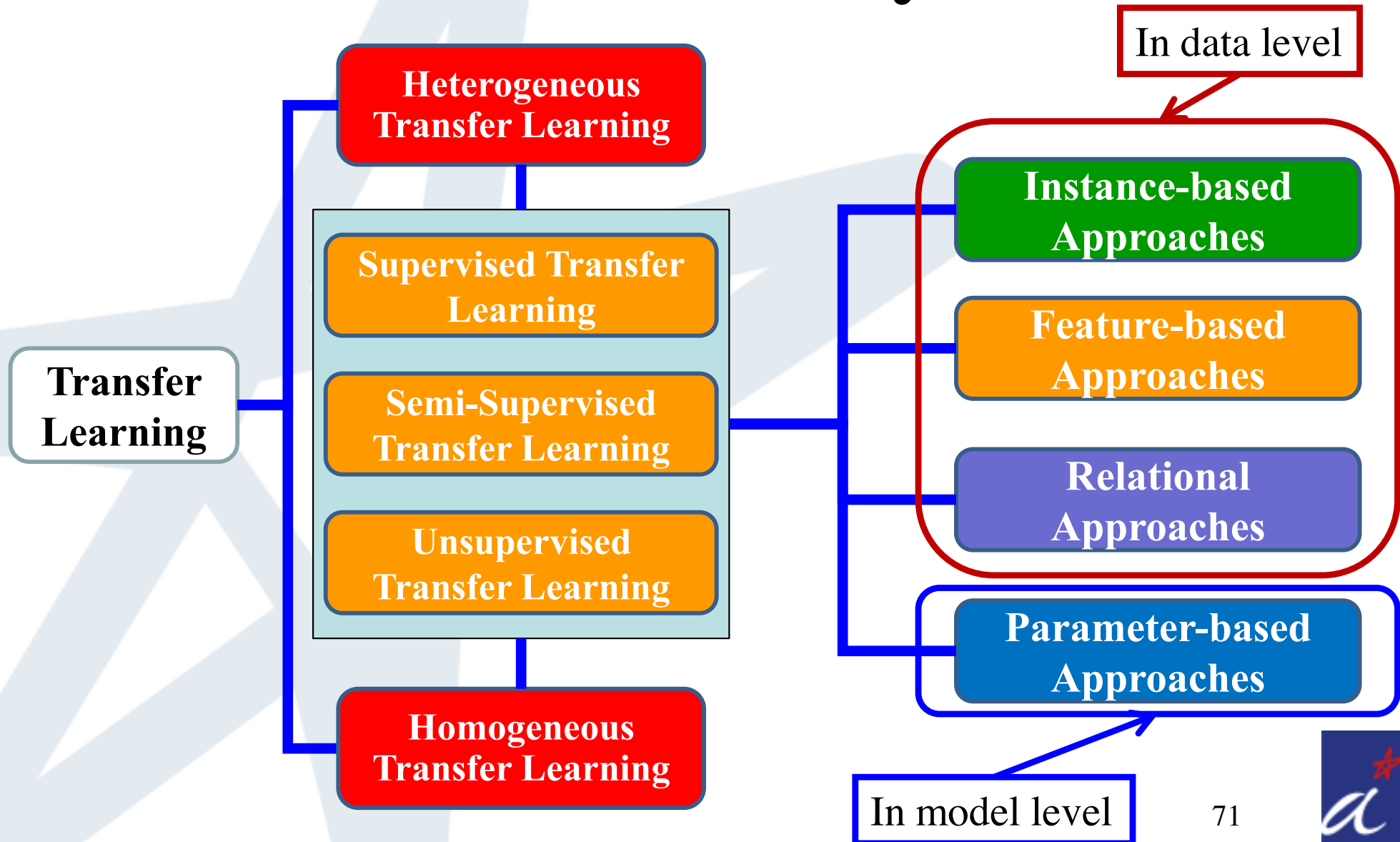


AdvisedBy (B, A)  $\wedge$  Publication (B, T)  
 $\Rightarrow$  Publication (A, T)

WorkedFor (A, B)  $\wedge$  MovieMember (A, M)  
 $\Rightarrow$  MovieMember (B, M)

$P1(x, y) \wedge P2(x, z) \Rightarrow P2(y, z)$

# Summary





**Thank You**