Connecting the Dots with **Landmarks**:

Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation

Boqing Gong
University of Southern California

Joint work with Kristen Grauman and Fei Sha

The perils of mismatched domains

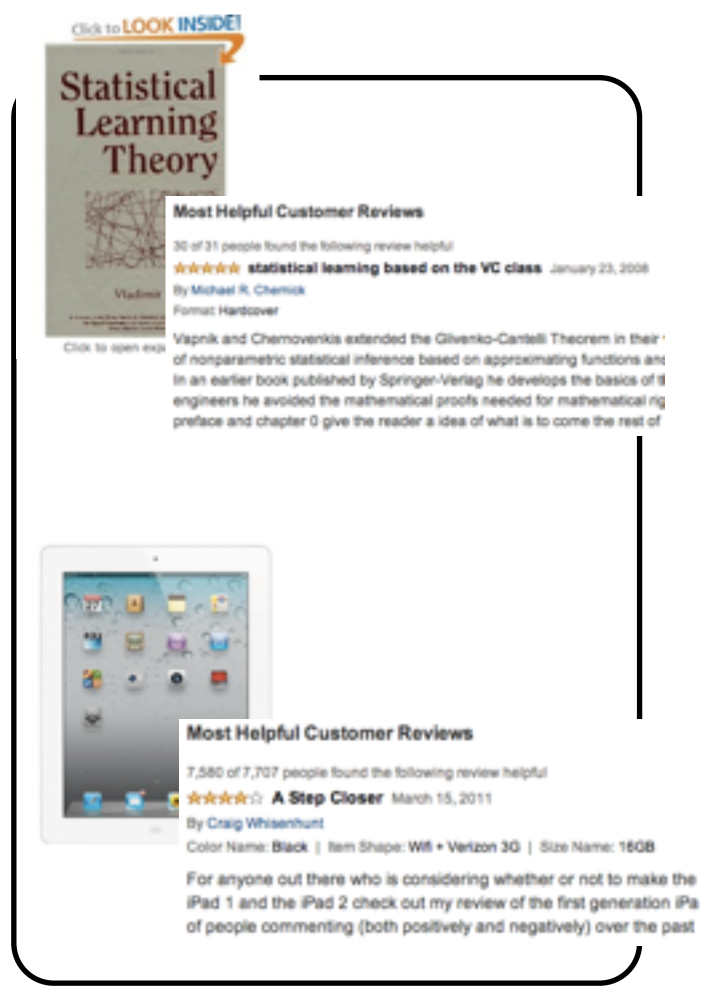
TRAIN → **TEST**

Poor cross-domain generalization

- Different underlying distributions
- Overfit to datasets' ***idiosyncrasies***

Images from [Saenko et al.'10].

Common to many areas



Computer vision

Text processing

Speech recognition

Language modeling

etc.

Unsupervised domain adaptation

Setup

Source domain (with labeled data)

$$D_{\mathcal{S}} = \{(x_m, y_m)\}_{m=1}^{\mathsf{M}} \sim P_{\mathcal{S}}(X, Y)$$

Target domain (no labels for training)

$$D_{\mathcal{T}} = \{(x_n, y_n)\}_{n=1}^{\mathsf{N}} \sim P_{\mathcal{T}}(X, Y)$$

Unsupervised domain adaptation

Setup

Source domain (with labeled data)

$$D_{\mathcal{S}} = \{(x_m, y_m)\}_{m=1}^{\mathsf{M}} \sim \boxed{P_{\mathcal{S}}(X,Y)}$$

Target domain (no labels for training)

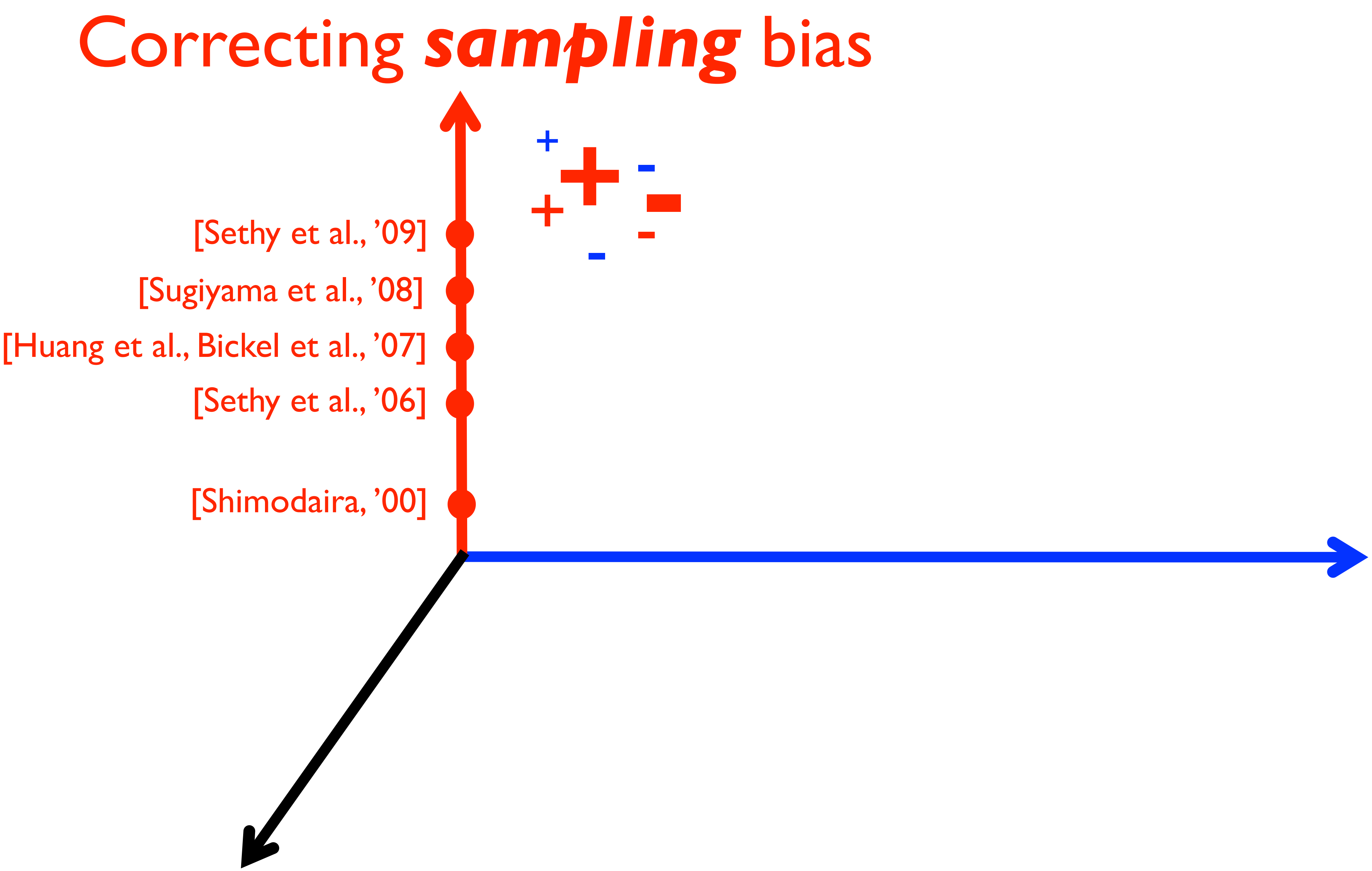
$$D_{\mathcal{T}} = \{(x_n, y_n)\}_{n=1}^{\mathsf{N}} \sim \boxed{P_{\mathcal{T}}(X,Y)}$$

Different distributions

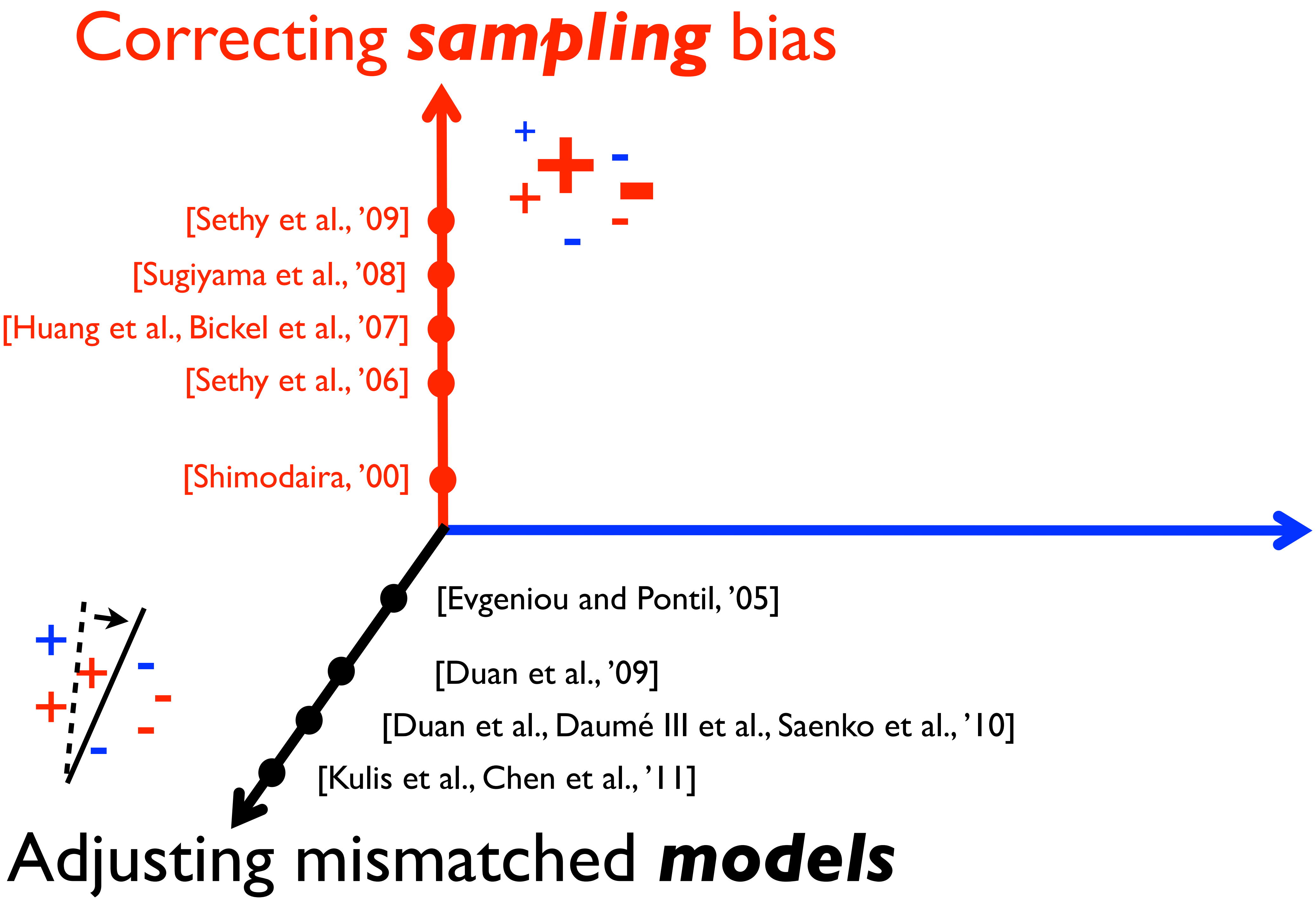
Objective

Learn classifier to work well on the target

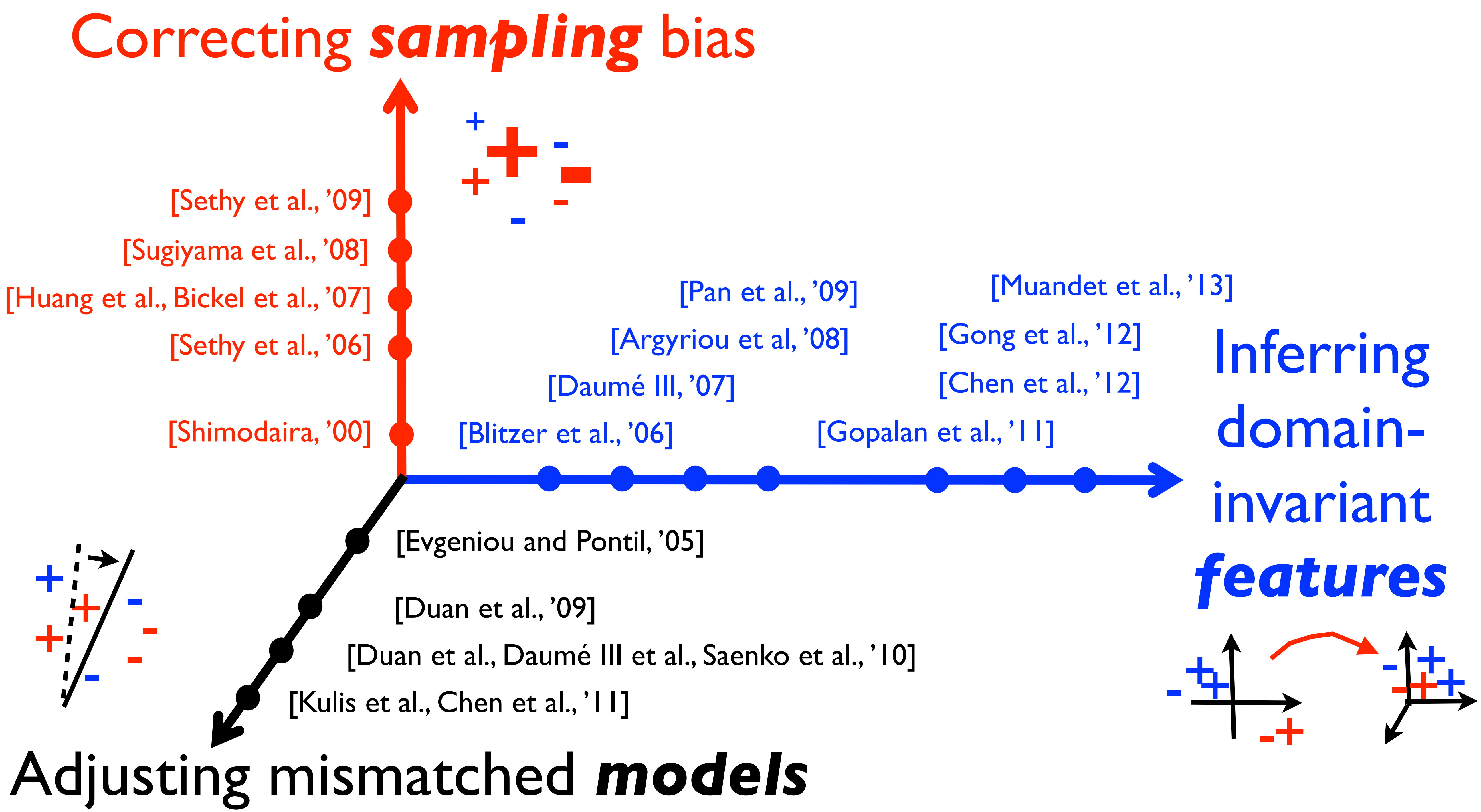
Many existing works



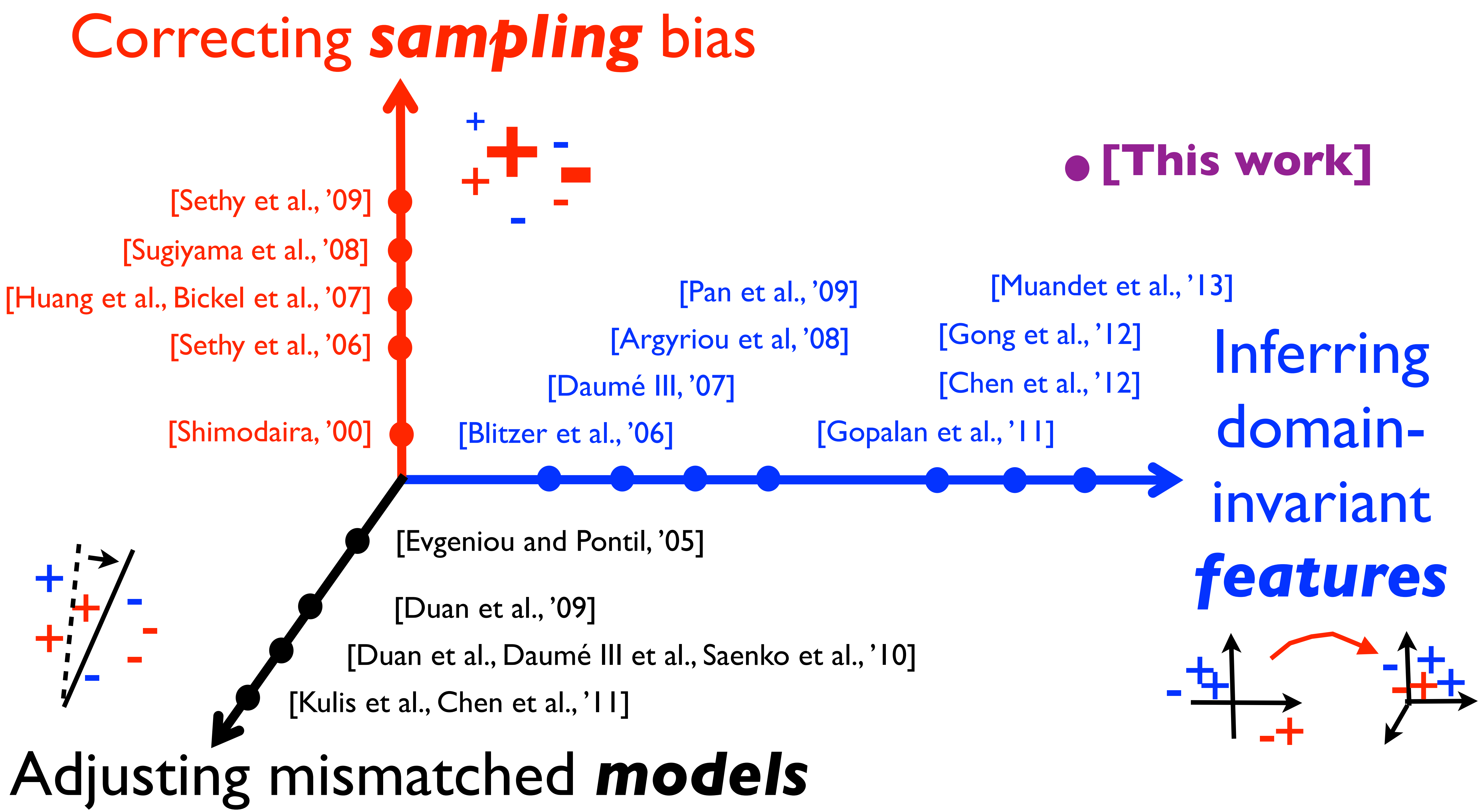
Many existing works



Many existing works



Many existing works



Snags

Forced adaptation

Attempting to adapt *all* source data points, including "hard" ones

Implicit discrimination

Learning discrimination biased to source, rather than optimized w.r.t. target

Our key insights

Forced adaptation

➔ Select the best instances for adaptation

Implicit discriminations

➔ Approximate discriminative loss on target

Landmarks

Landmarks are labeled source instances distributed similarly to the target domain.

Landmarks

Landmarks are labeled source instances distributed similarly to the target domain.

Landmarks

Landmarks are labeled source instances distributed similarly to the target domain.

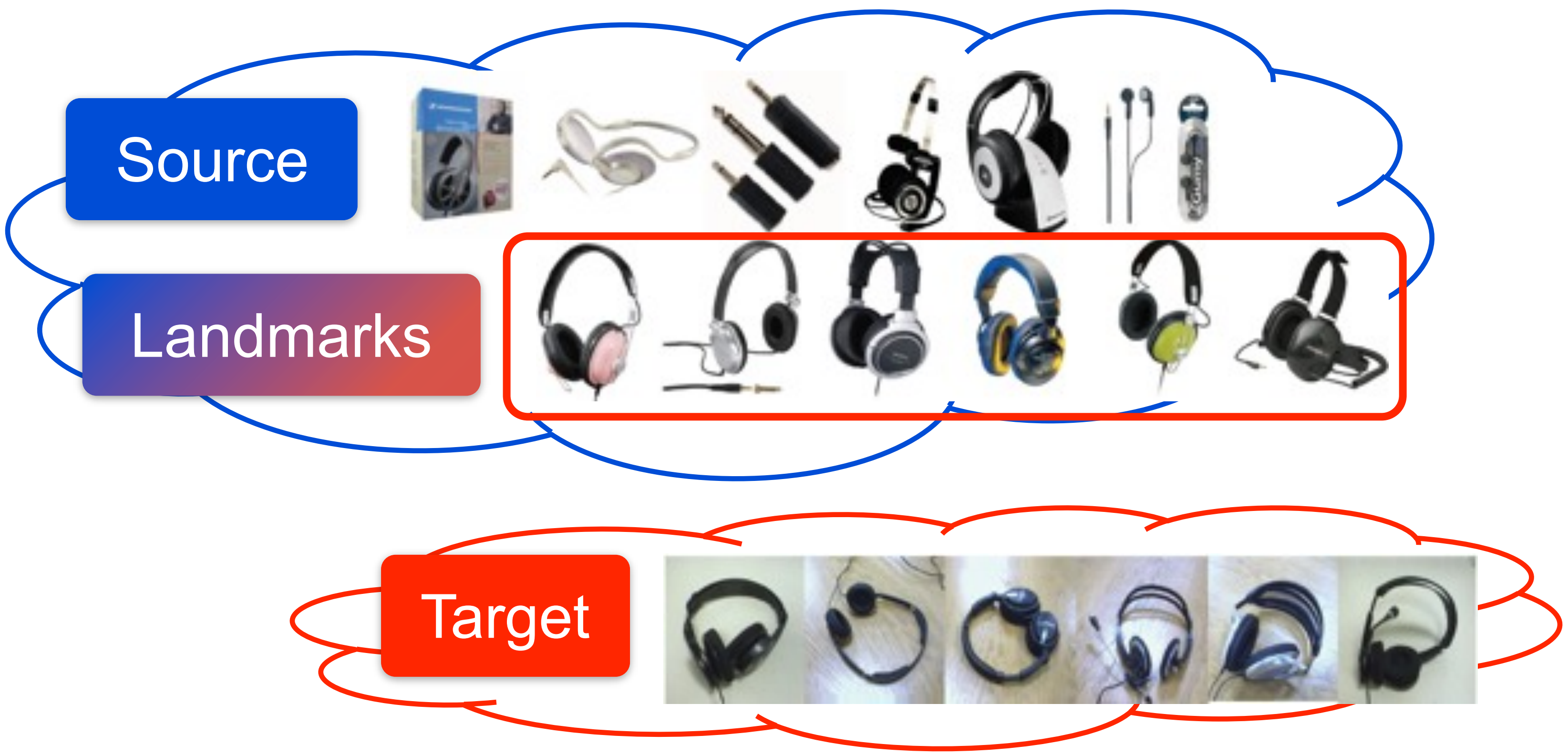
Landmarks

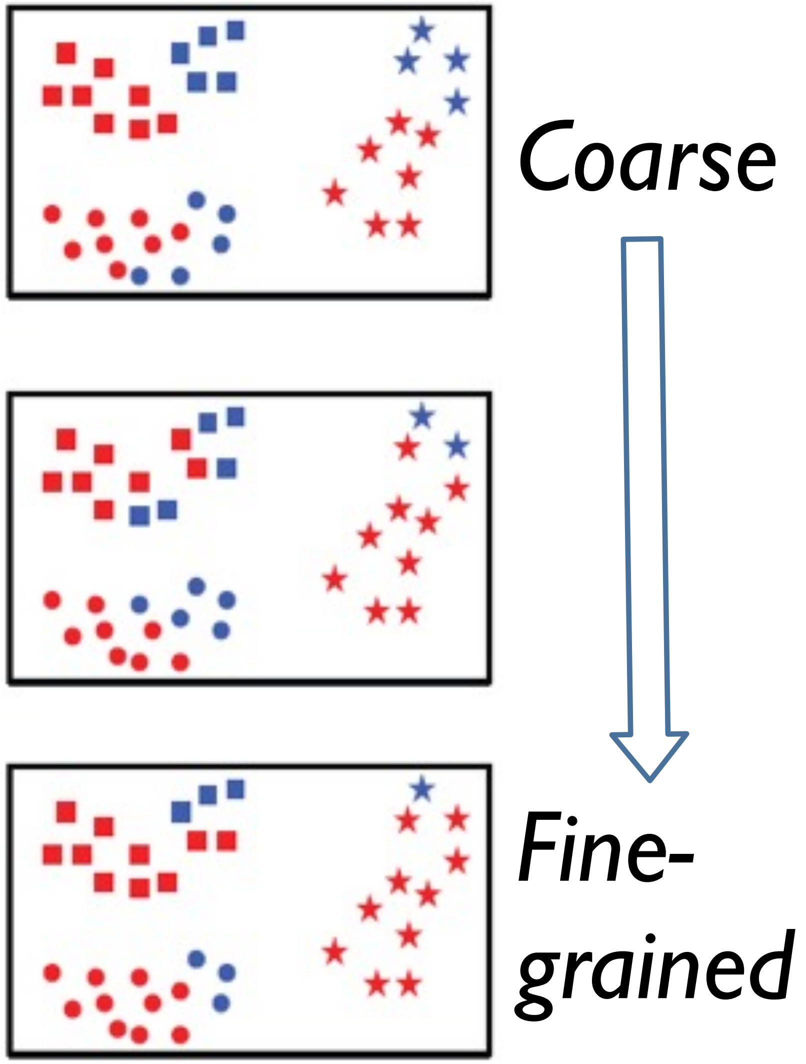
Landmarks are labeled source instances distributed similarly to the target domain.

Roles

- Ease adaptation difficulty
- Provide discrimination (biased to target)

Key steps

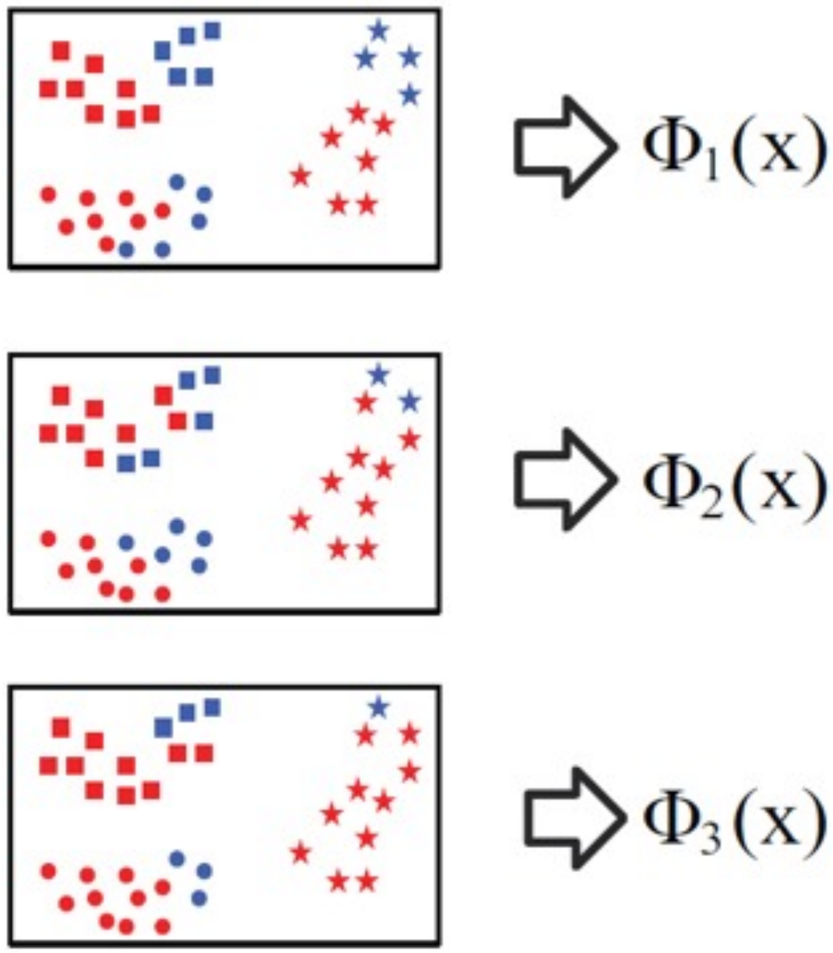




❶ Identify landmarks

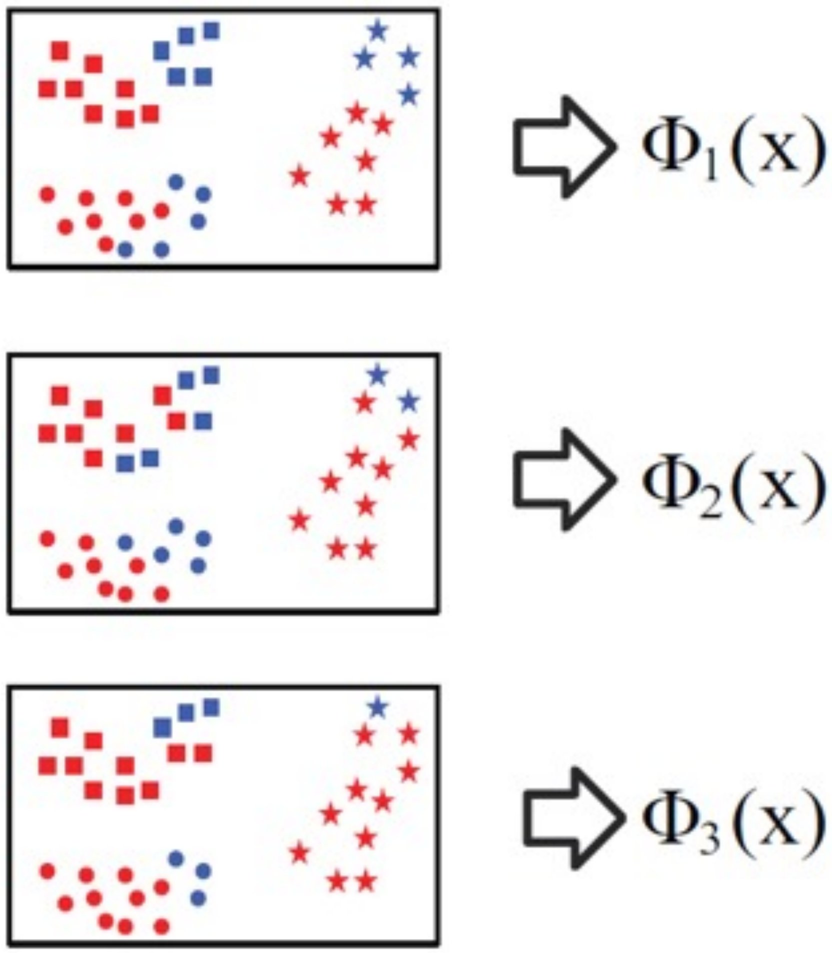
at multiple scales.

Key steps



2 Construct auxiliary domain adaptation tasks

Key steps

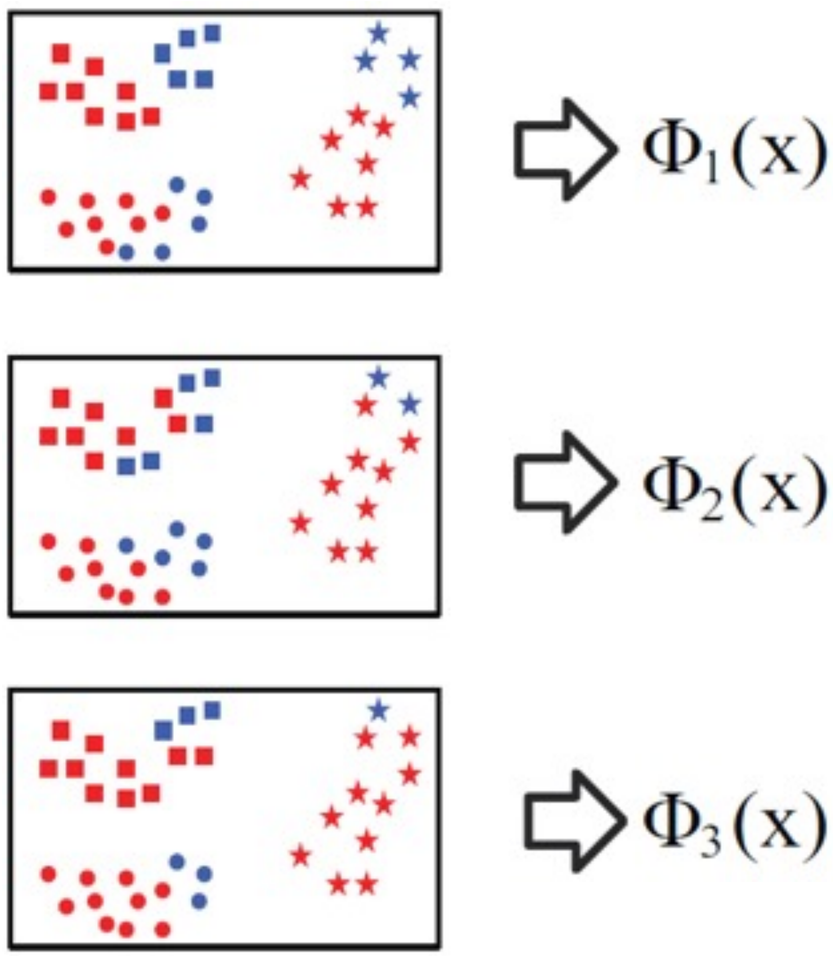


2 Construct auxiliary domain adaptation tasks

$$\Phi(x)=\begin{bmatrix} \Phi_1(x)\cdot w_1 \\ \Phi_2(x)\cdot w_2 \\ \Phi_3(x)\cdot w_3 \end{bmatrix}$$

3 Obtain domain-invariant features

Key steps

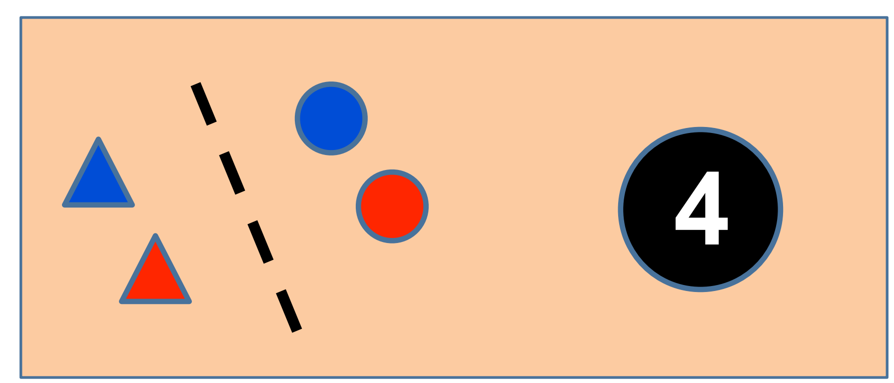


2 Construct auxiliary domain adaptation tasks

$$\Phi(\mathrm{x}) = \begin{bmatrix} \Phi_1(\mathrm{x}) \cdot w_1 \\ \Phi_2(\mathrm{x}) \cdot w_2 \\ \Phi_3(\mathrm{x}) \cdot w_3 \end{bmatrix}$$

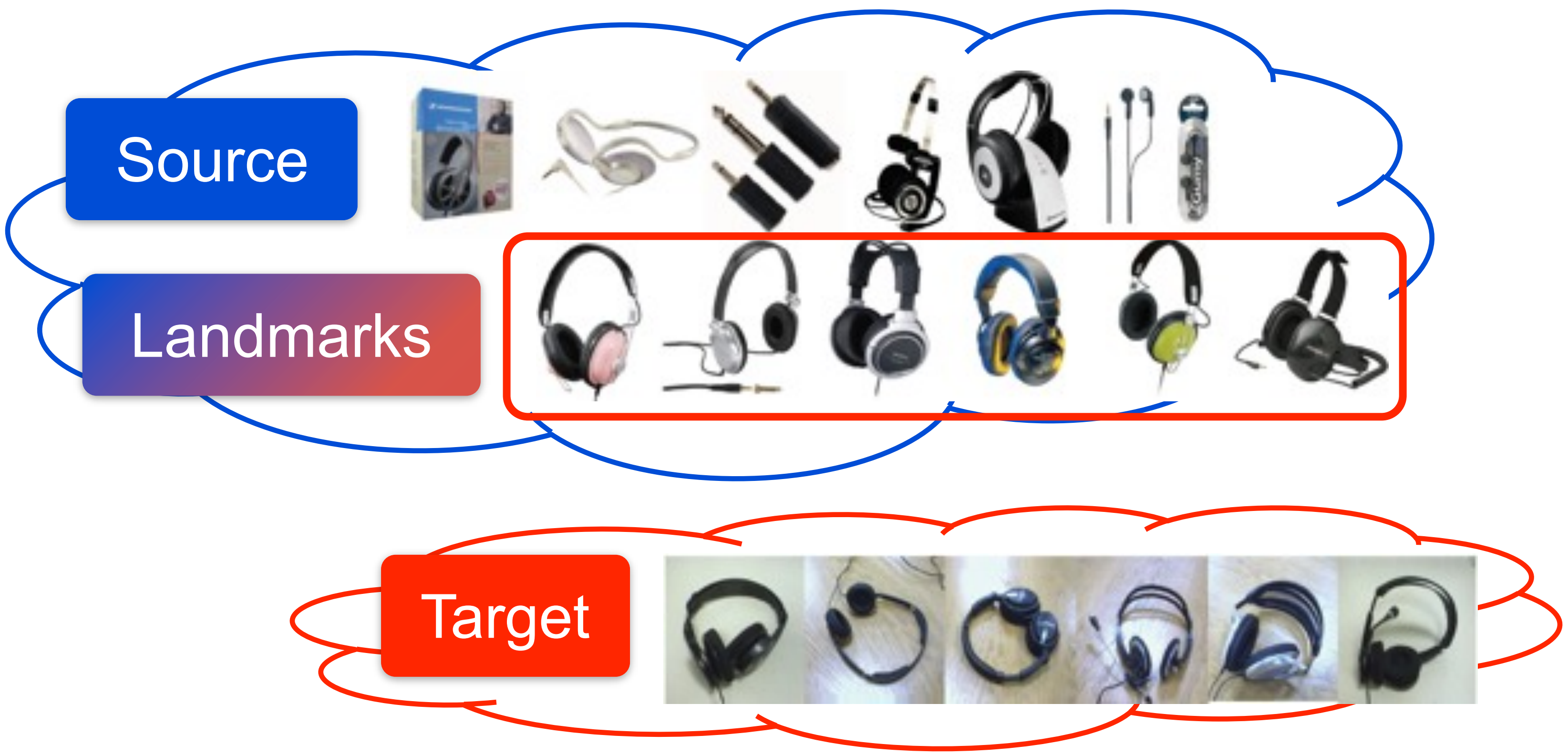
3

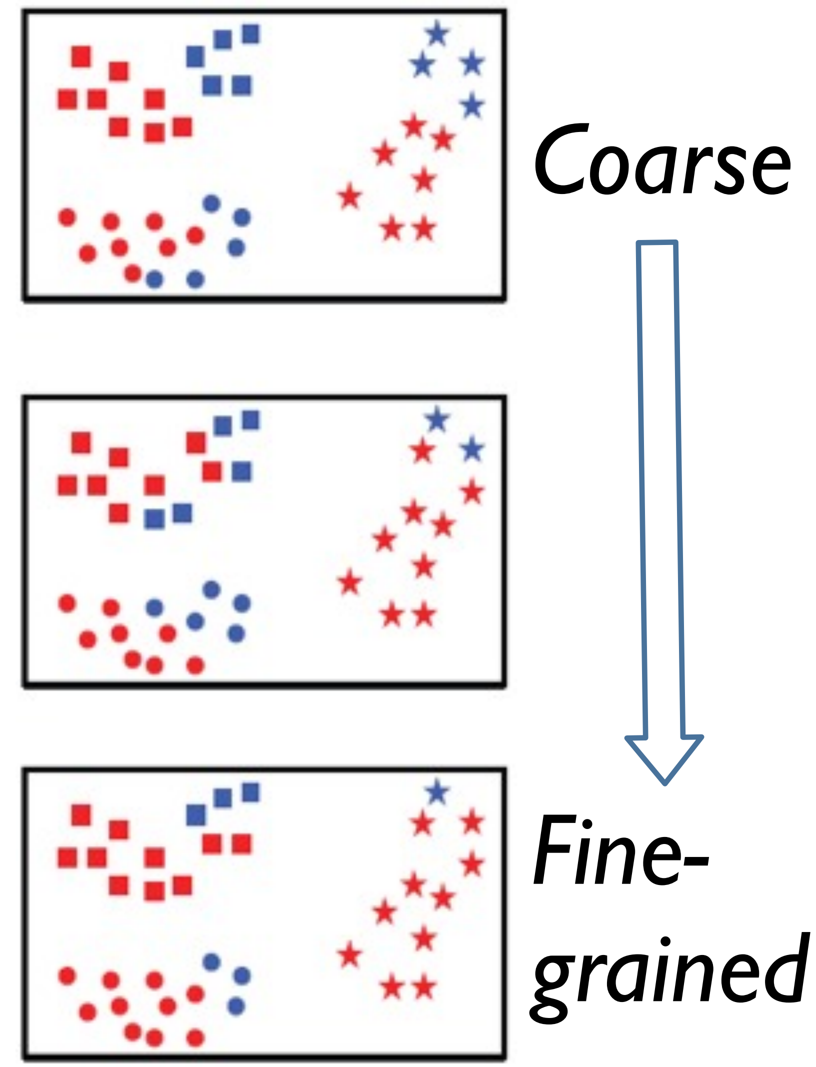
Obtain domain-invariant features



Predict target labels

Key steps





1 Identify landmarks

at multiple scales.

Identifying landmarks

Objective

$$P_{\mathcal{L}}(\text{landmarks}) \approx P_{\mathcal{T}}(\text{target})$$

Identifying landmarks

Objective

$$P_{\mathcal{L}}(\text{landmarks}) \approx P_{\mathcal{T}}(\text{target})$$

$$\min_{\text{landmarks}} \quad d(P_{\mathcal{L}}, P_{\mathcal{T}})$$

Identifying landmarks

Objective

$$P_{\mathcal{L}}(\text{landmarks}) \approx P_{\mathcal{T}}(\text{target})$$

$$\min_{\text{landmarks}} \quad d(P_{\mathcal{L}}, P_{\mathcal{T}})\,?$$

Maximum mean discrepancy (MMD)

Empirical estimate [Gretton et al. '06]

$$d(P_{\mathcal{L}}, P_{\mathcal{T}}) = \left\| \frac{1}{\mathsf{L}} \sum_{l=1}^{\mathsf{L}} \phi(x_l) - \frac{1}{\mathsf{N}} \sum_{n=1}^{\mathsf{N}} \phi(x_n) \right\|_{\mathcal{H}}$$

$\mathcal{H}$ a universal RKHS

$\phi(\cdot)$ kernel function induced by $\mathcal{H}$

x_l the l-th **landmark** (from the source domain)

Method for identifying landmarks

Integer programming

$$\min_{\{\alpha_m\}} \quad \left\| \frac{1}{\sum_i \alpha_i} \sum_{m=1}^{\mathsf{M}} \alpha_m \phi(x_m) - \frac{1}{\mathsf{N}} \sum_{n=1}^{\mathsf{N}} \phi(x_n) \right\|_{\mathcal{H}}^2$$

where

$$\alpha_m = \begin{cases} 1 & \text{if } x_m \text{ is a landmark for the target} \\ 0 & \text{else} \end{cases}$$

$$m = 1, 2, \cdots, \mathsf{M}$$

Method for identifying landmarks

Convex relaxation

$$\min_{\{\alpha_m\}} \quad \left\| \frac{1}{\sum_i \alpha_i} \sum_{m=1}^{\mathsf{M}} \alpha_m \phi(x_m) - \frac{1}{\mathsf{N}} \sum_{n=1}^{\mathsf{N}} \phi(x_n) \right\|_{\mathcal{H}}^2$$

Method for identifying landmarks

Convex relaxation

$$\min_{\{\alpha_m\}} \quad \left\| \boxed{\frac{1}{\sum_i \alpha_i}} \sum_{m=1}^{\mathsf{M}} \alpha_m \phi(x_m) - \frac{1}{\mathsf{N}} \sum_{n=1}^{\mathsf{N}} \phi(x_n) \right\|_{\mathcal{H}}^2$$

Method for identifying landmarks

Convex relaxation

$$\min_{\{\alpha_m\}} \quad \left\| \boxed{\frac{1}{\sum_i \alpha_i}} \sum_{m=1}^{\mathsf{M}} \alpha_m \phi(x_m) - \frac{1}{\mathsf{N}} \sum_{n=1}^{\mathsf{N}} \phi(x_n) \right\|_{\mathcal{H}}^2$$

$$\boxed{\beta_m = \frac{\alpha_m}{\sum_i \alpha_i} \rightarrow \text{Quadratic programming}}$$

$$\min_{\beta} \quad \beta^T K^s \beta - \frac{2}{\mathsf{N}} \beta^T K^{st} \mathbf{1}$$

How to choose the kernel functions?

$$\min_{\beta} \quad \beta^T K^s \beta - \frac{2}{\mathrm{N}} \beta^T K^{st} \mathbf{1}$$

Gaussian kernels

Plus: universal (characteristic)
Minus: how to choose the bandwidth?

How to choose the kernel functions?

$$\min_{\beta} \quad \beta^T K^s \beta - \frac{2}{\mathsf{N}} \beta^T K^{st} \mathbf{1}$$

Gaussian kernels

Plus: universal (characteristic)
Minus: how to choose the bandwidth?

Our solution: bandwidth---granularity

Examining distributions at multiple granularities
Multiple bandwidths, multiple sets of landmarks

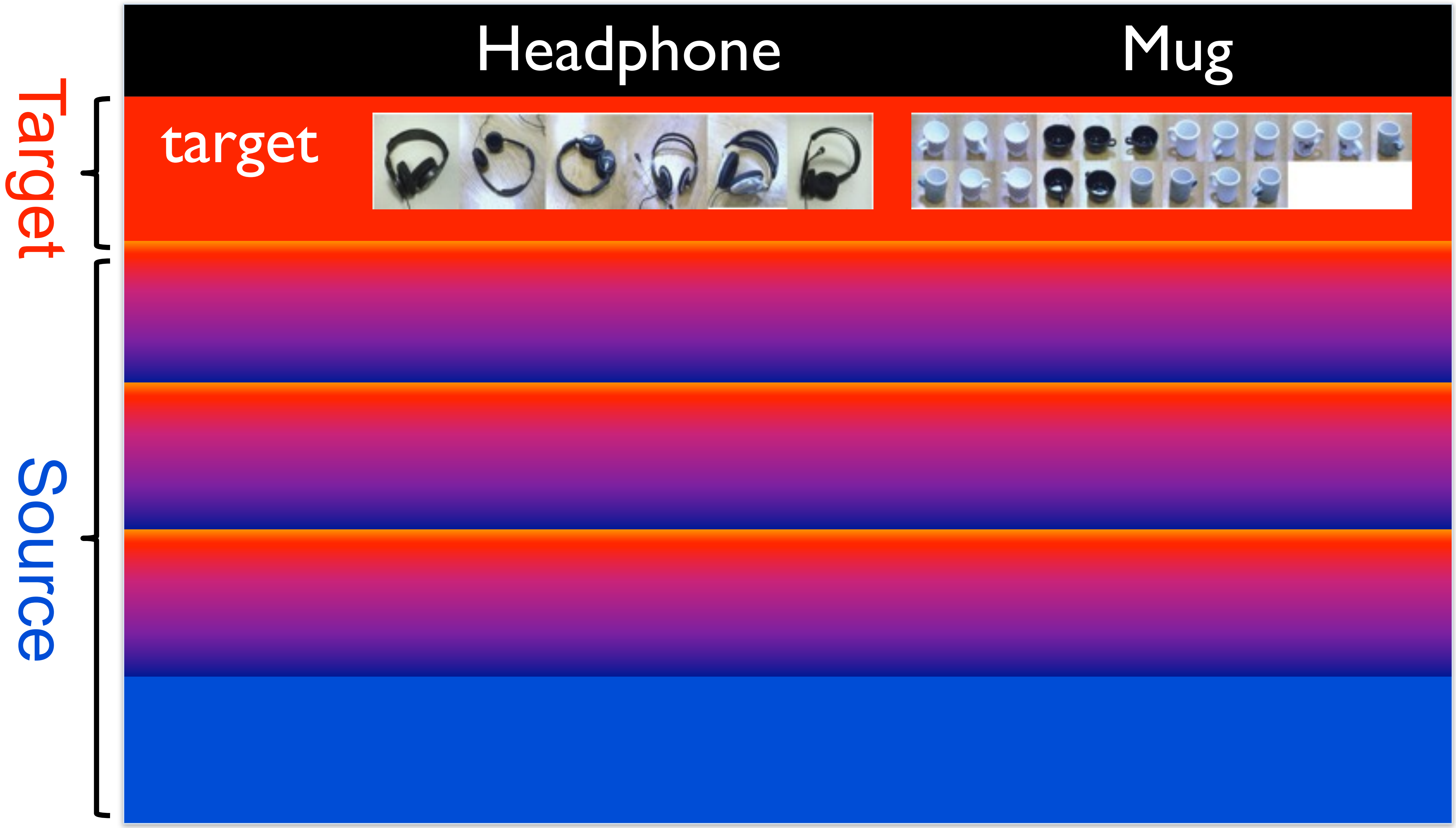
Other details

Class balance constraint

Recovering $\alpha_m^\star$ from $\beta_m^\star (= \frac{\alpha_m}{\sum_i \alpha_i})$

(See paper for details)

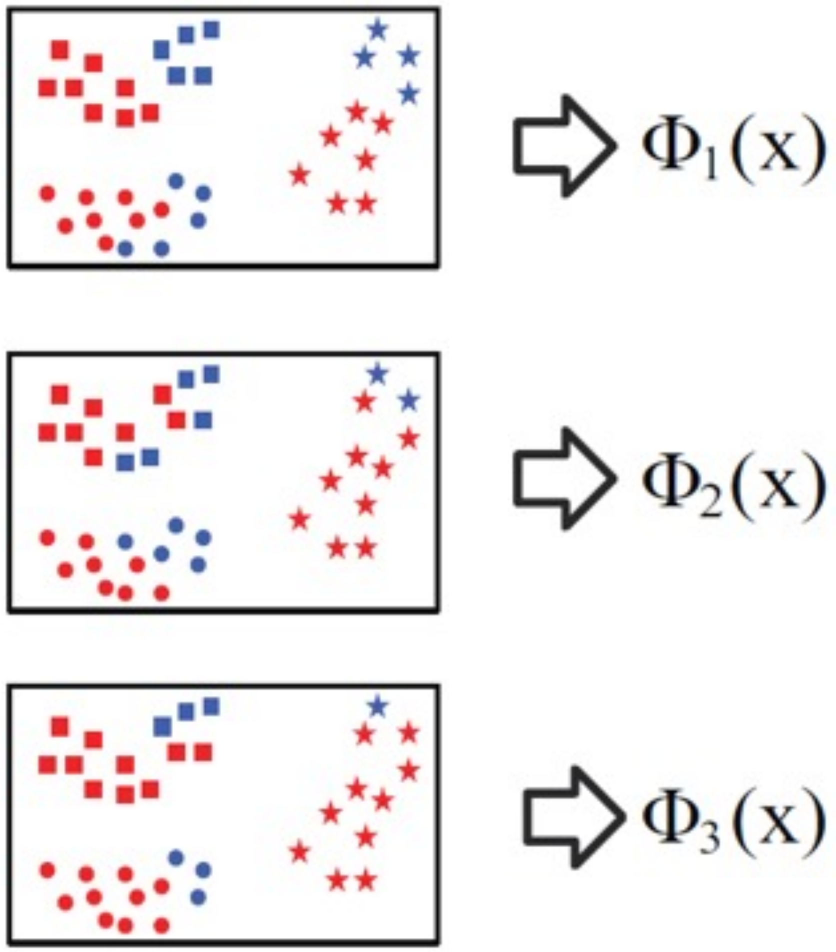
What do landmarks look like?



What do landmarks look like?

What do landmarks look like?

Key steps



2 Construct auxiliary domain adaptation tasks

Constructing easier auxiliary tasks

At each scale σ

$$\textcolor{blue}{\text{New source}} = \textcolor{blue}{\text{Source}} \setminus \text{Landmarks}$$

$$\textcolor{red}{\text{New target}} = \textcolor{red}{\text{Target}} \cup \text{Landmarks}$$

Intuition: distributions are closer (cf. Theorem 1)

Constructing easier auxiliary tasks

At each scale σ

$$\text{New source} = \text{Source} \setminus \text{Landmarks}$$

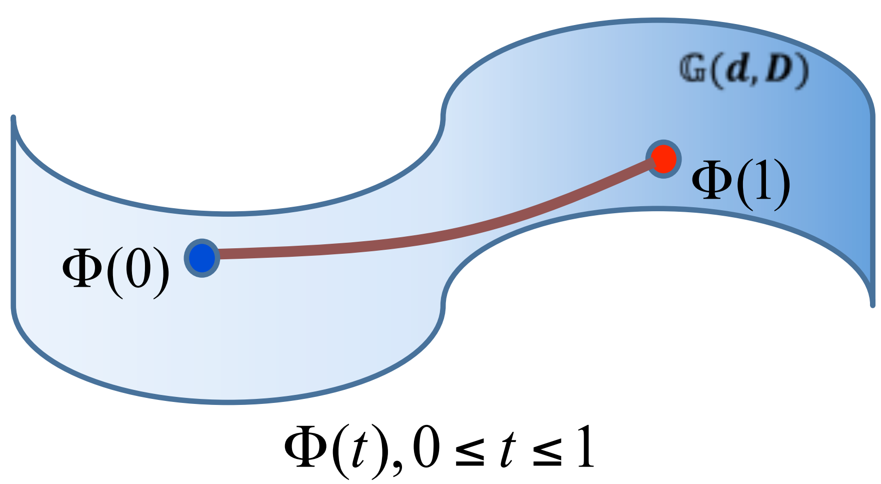
$$\text{New target} = \text{Target} \cup \text{Landmarks}$$

Intuition: distributions are closer (cf. Theorem 1)

Auxiliary tasks → new basis of features

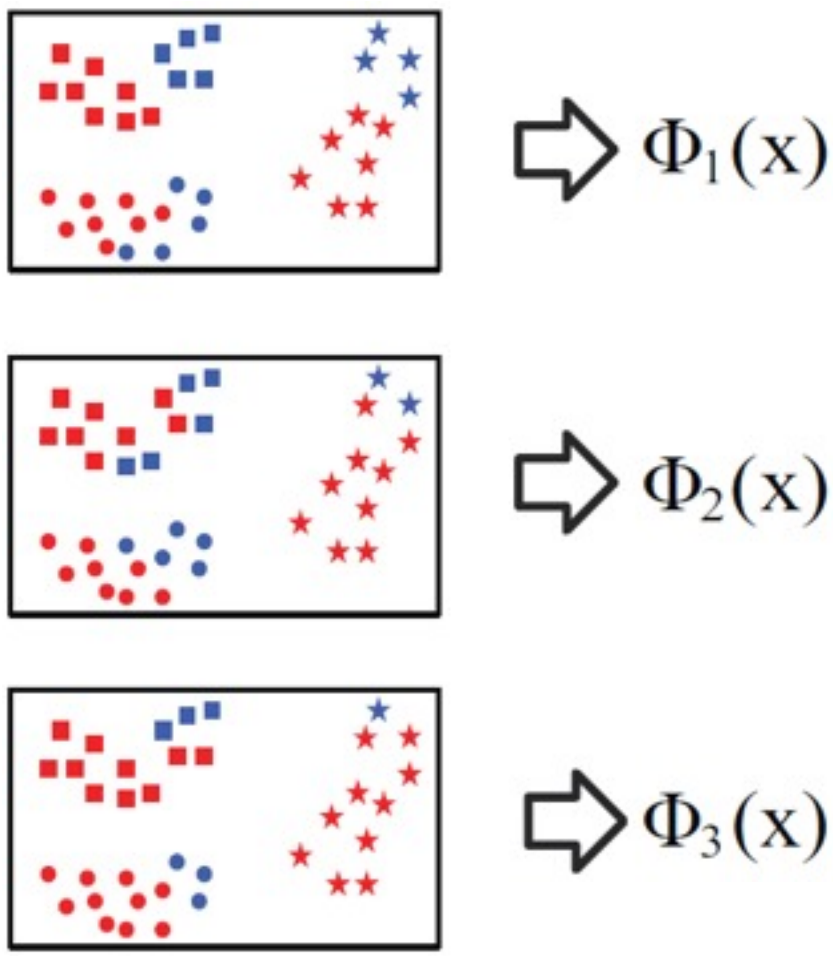
by a geodesic flow kernel (GFK) based method

$$K_\sigma(x_i, x_j) = \int_0^1 (\Phi_\sigma(t)' x_i)' (\Phi_\sigma(t)' x_j) \mathrm{d}t = x_i G_\sigma x_j$$



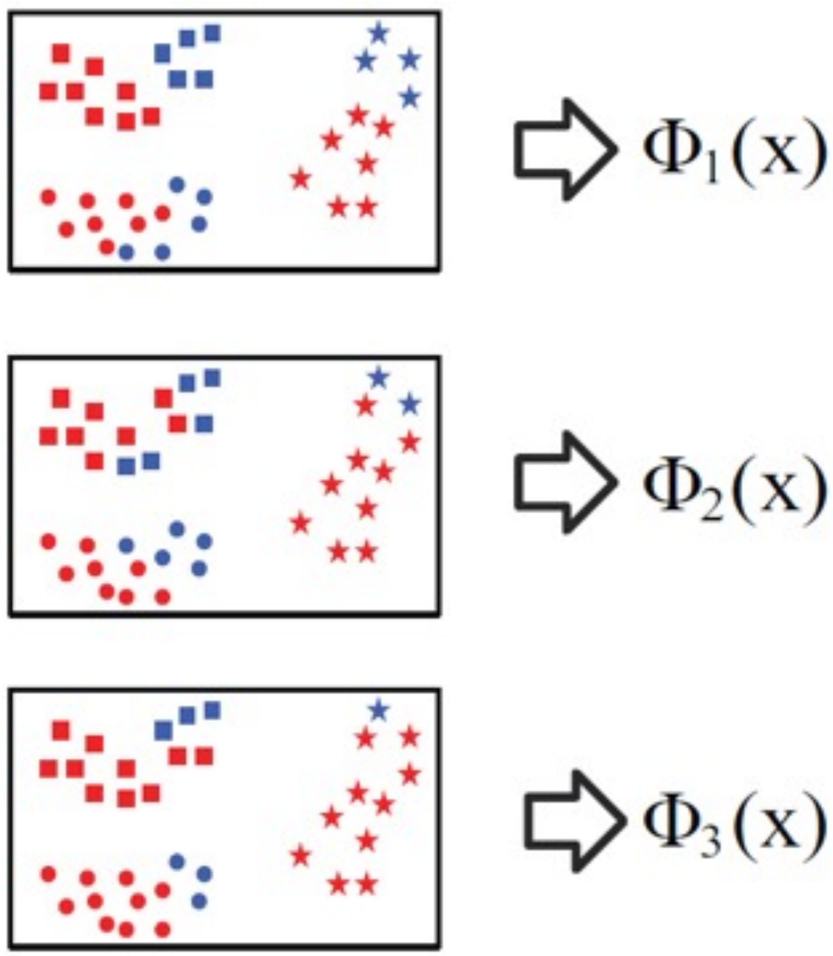
- Integrate out domain changes
- Obtain domain-invariant representation [Gong, et al. '12]

Key steps



2 Construct auxiliary domain adaptation tasks

Key steps



2 Construct auxiliary domain adaptation tasks

$$\Phi(x)=\begin{bmatrix}\Phi_1(x)\cdot w_1\\ \Phi_2(x)\cdot w_2\\ \Phi_3(x)\cdot w_3\end{bmatrix}$$

3 Obtain domain-invariant features

Combining features discriminatively

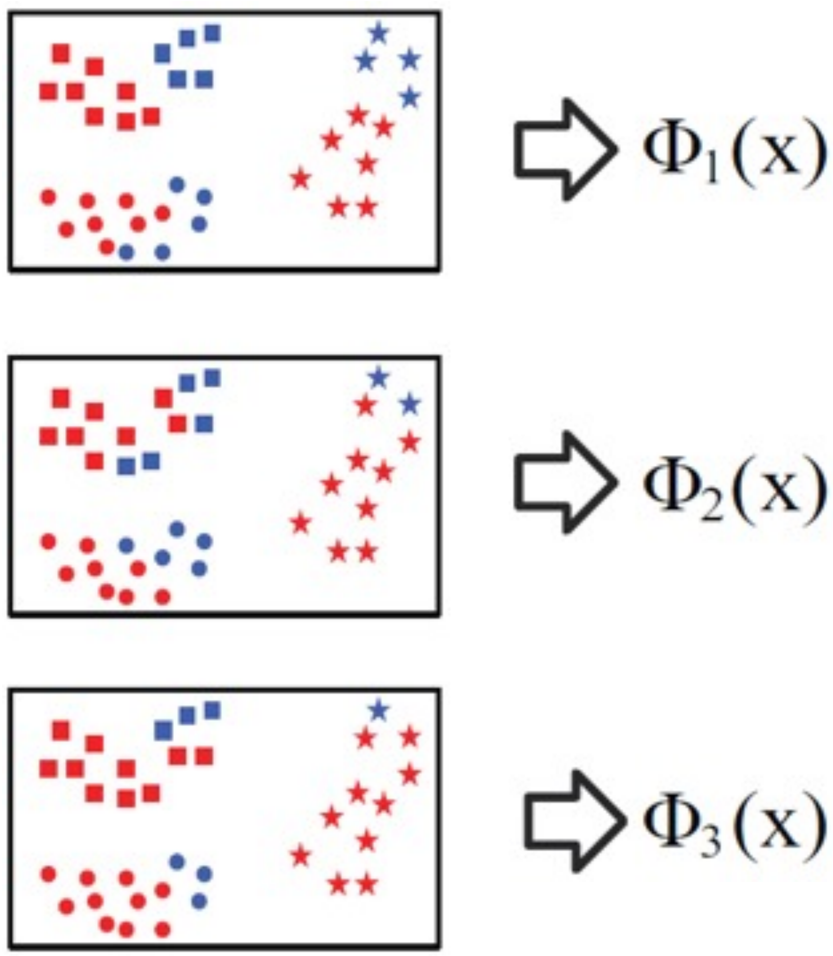
Multiple kernel learning on the labeled landmarks

$$F = \sum_{\sigma} w_{\sigma} G_{\sigma}, \quad \text{s.t.} \quad w_{\sigma} \geq 0, \sum_{\sigma} w_{\sigma} = 1$$

Arriving at domain-invariant feature space

Discriminative loss biased to the target

Key steps

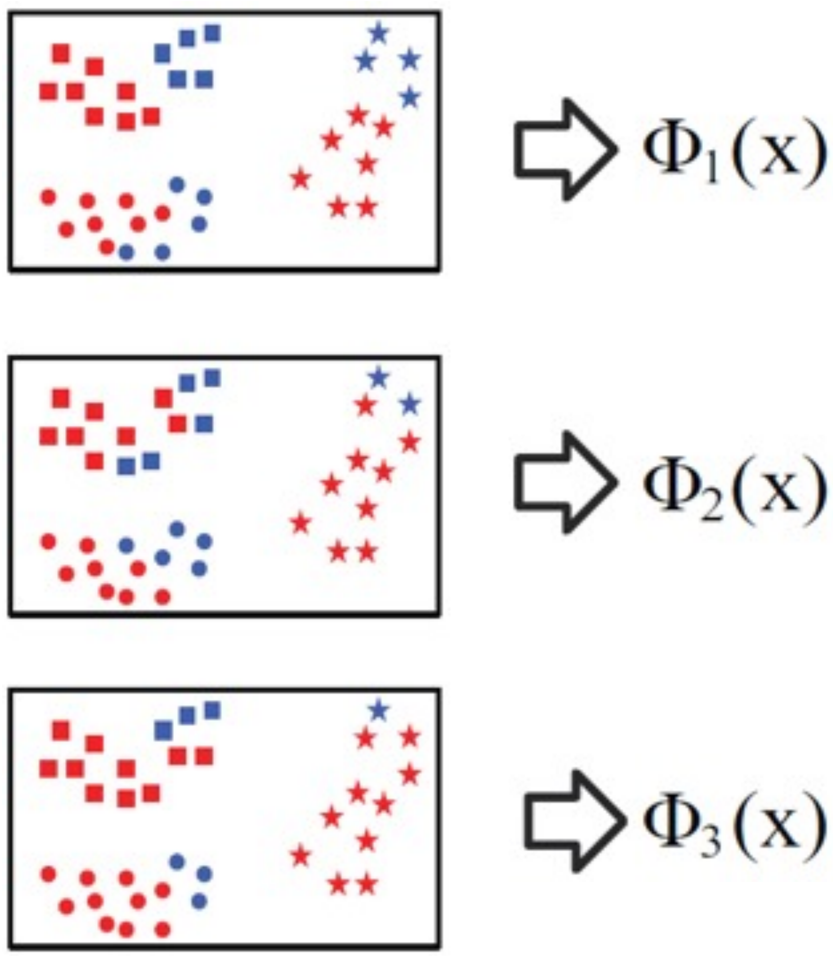


$$\Phi(x) = \begin{bmatrix} \Phi_1(x) \cdot w_1 \\ \Phi_2(x) \cdot w_2 \\ \Phi_3(x) \cdot w_3 \end{bmatrix}$$

3 Obtain domain-invariant features

2 Construct auxiliary domain adaptation tasks

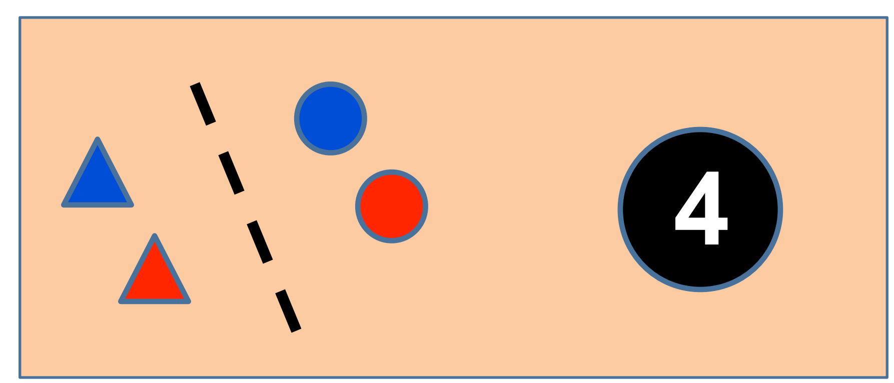
Key steps



2 Construct auxiliary domain adaptation tasks

$$\Phi(x) = \begin{bmatrix} \Phi_1(x) \cdot w_1 \\ \Phi_2(x) \cdot w_2 \\ \Phi_3(x) \cdot w_3 \end{bmatrix}$$

3 Obtain domain-invariant features



Predict target labels

Experimental study

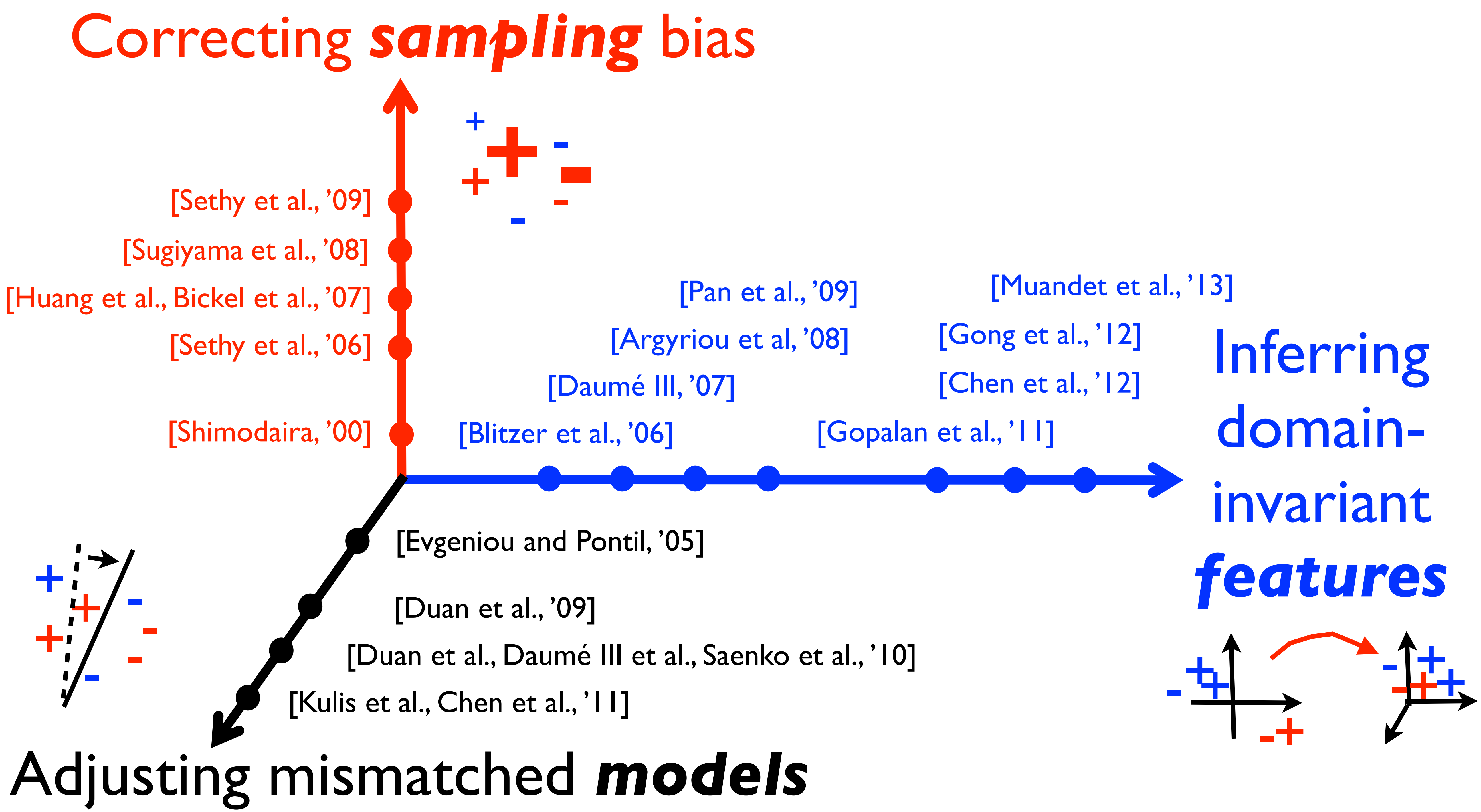
Four vision datasets/domains on visual object recognition

[Griffin et al. '07, Saenko et al. 10']

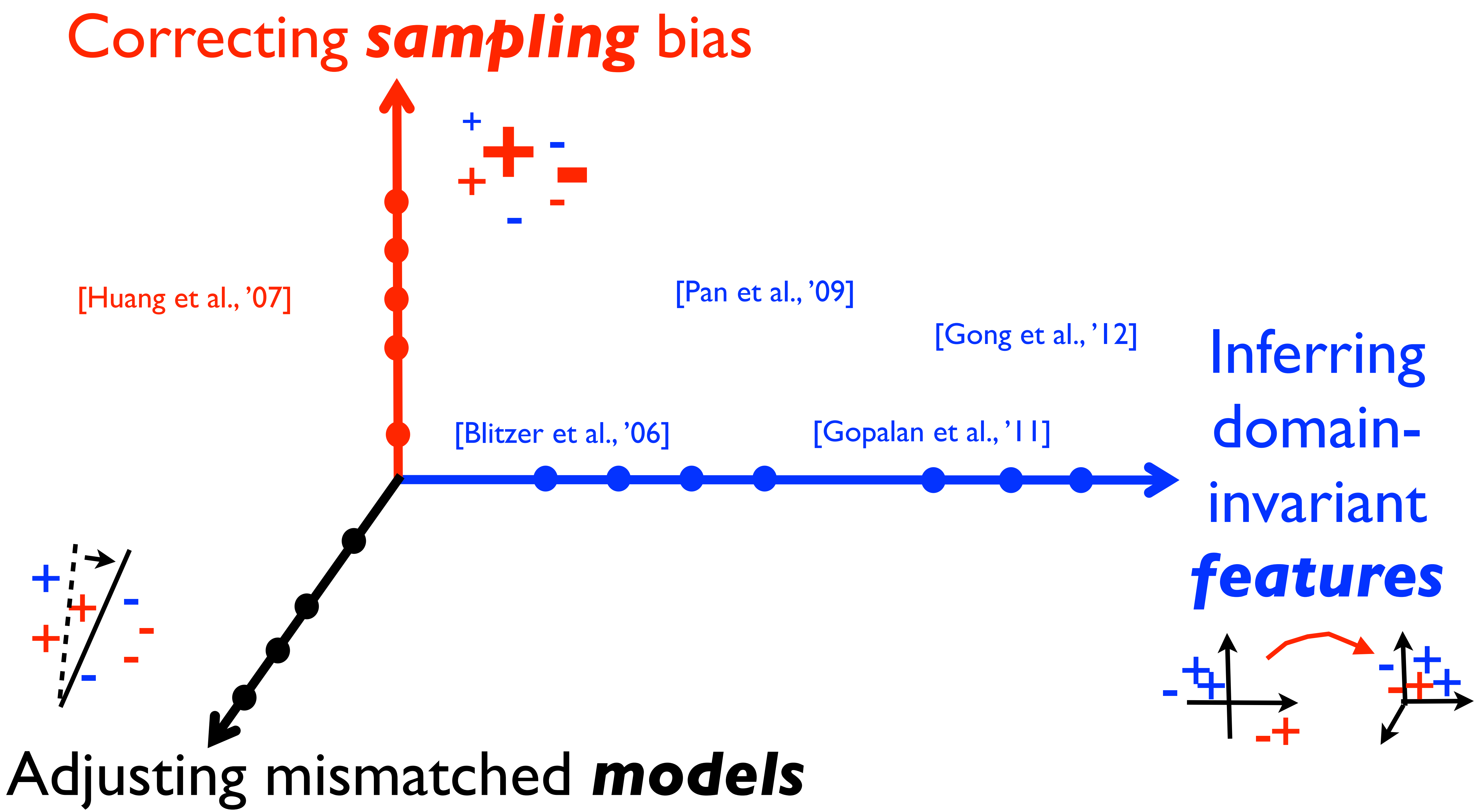
Four types of product reviews on sentiment analysis

Books, DVD, electronics, kitchen appliances [Biltzer et al. '07]

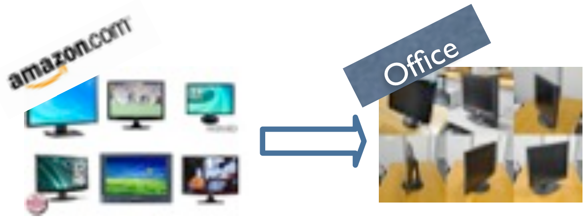
Comparing with

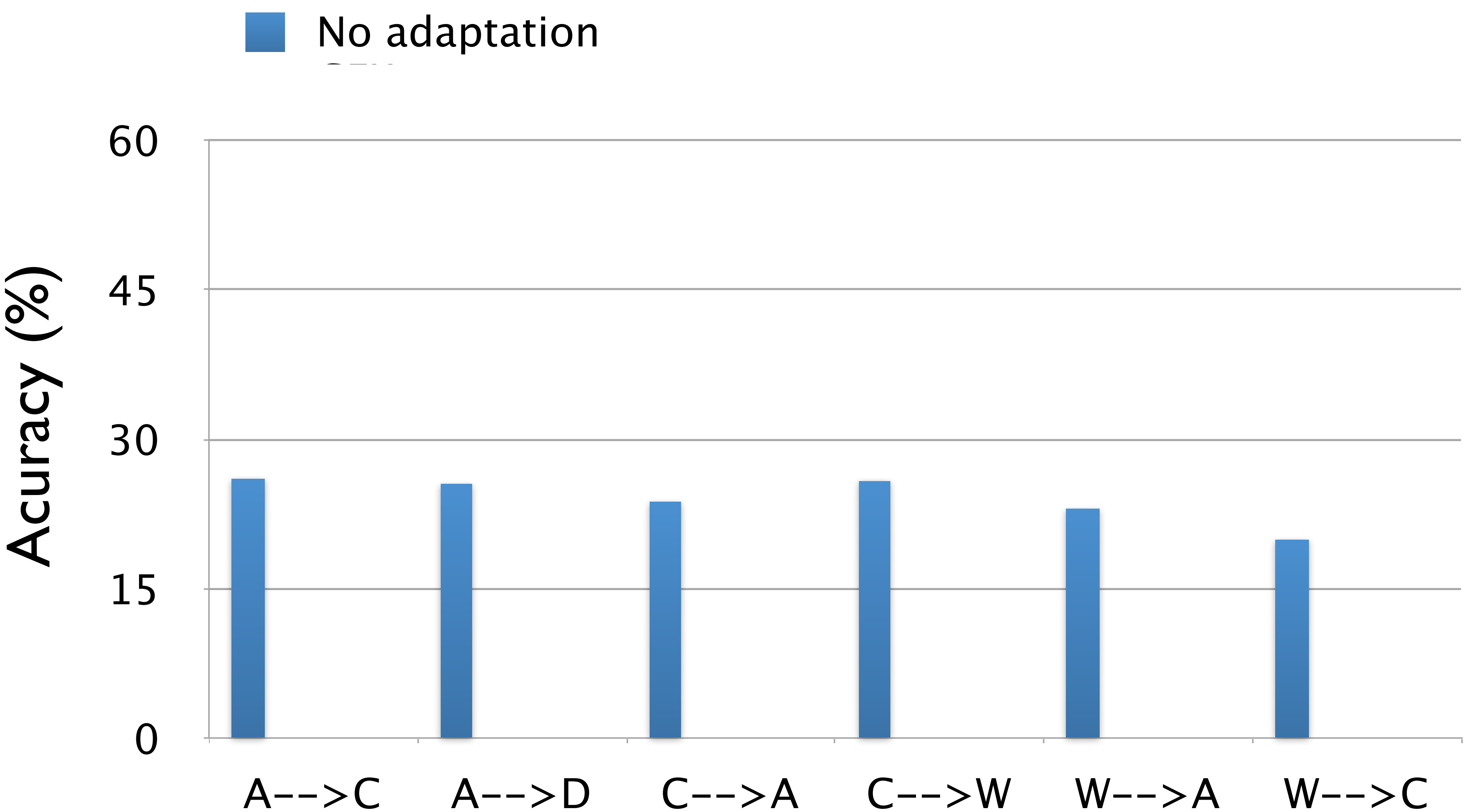


Comparing with

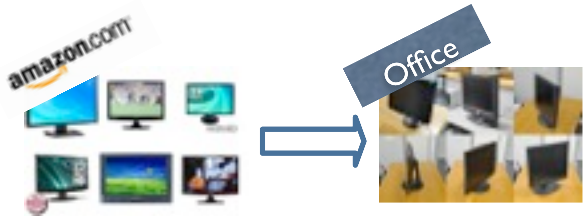


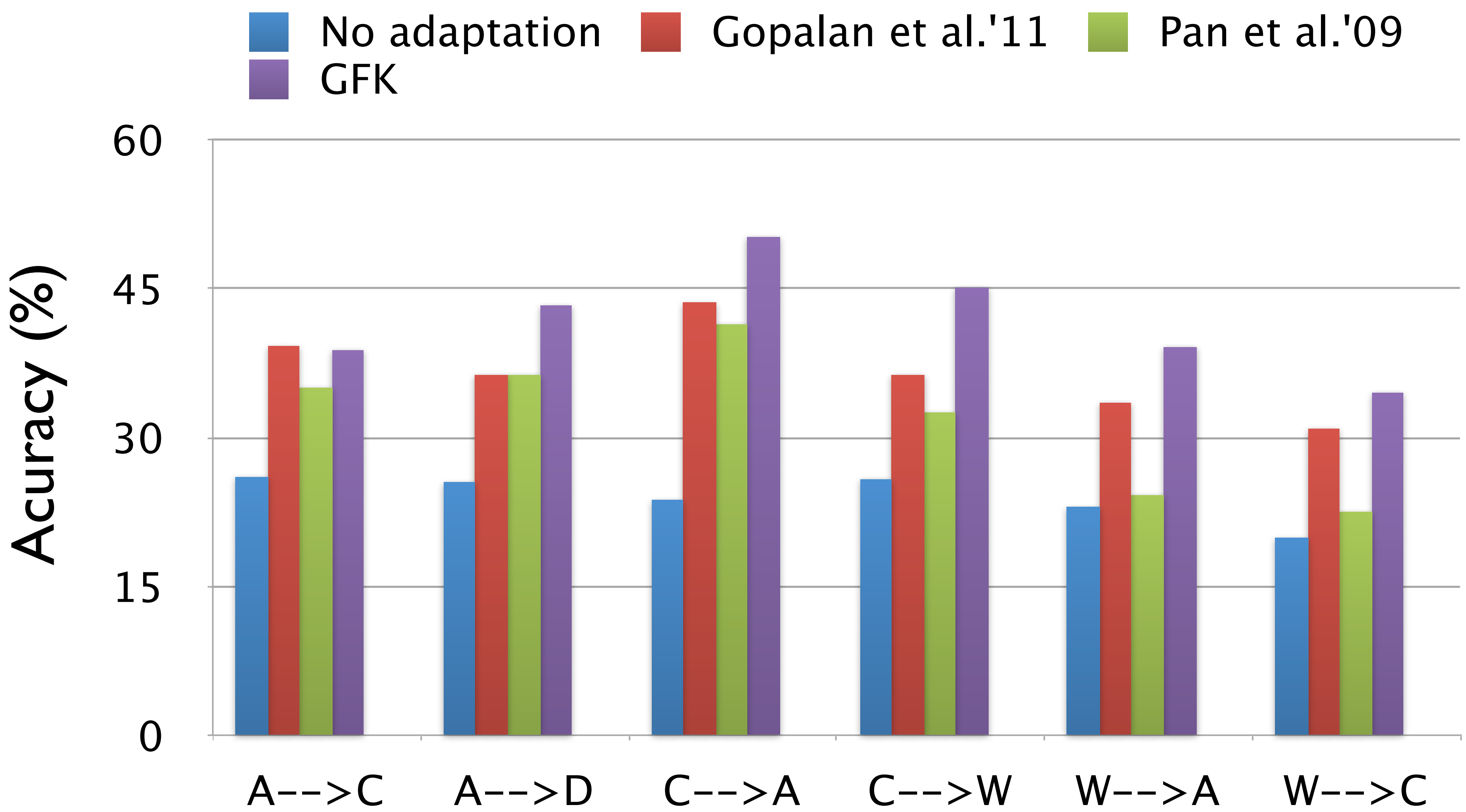
Object recognition



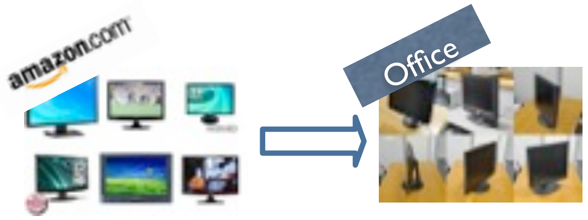


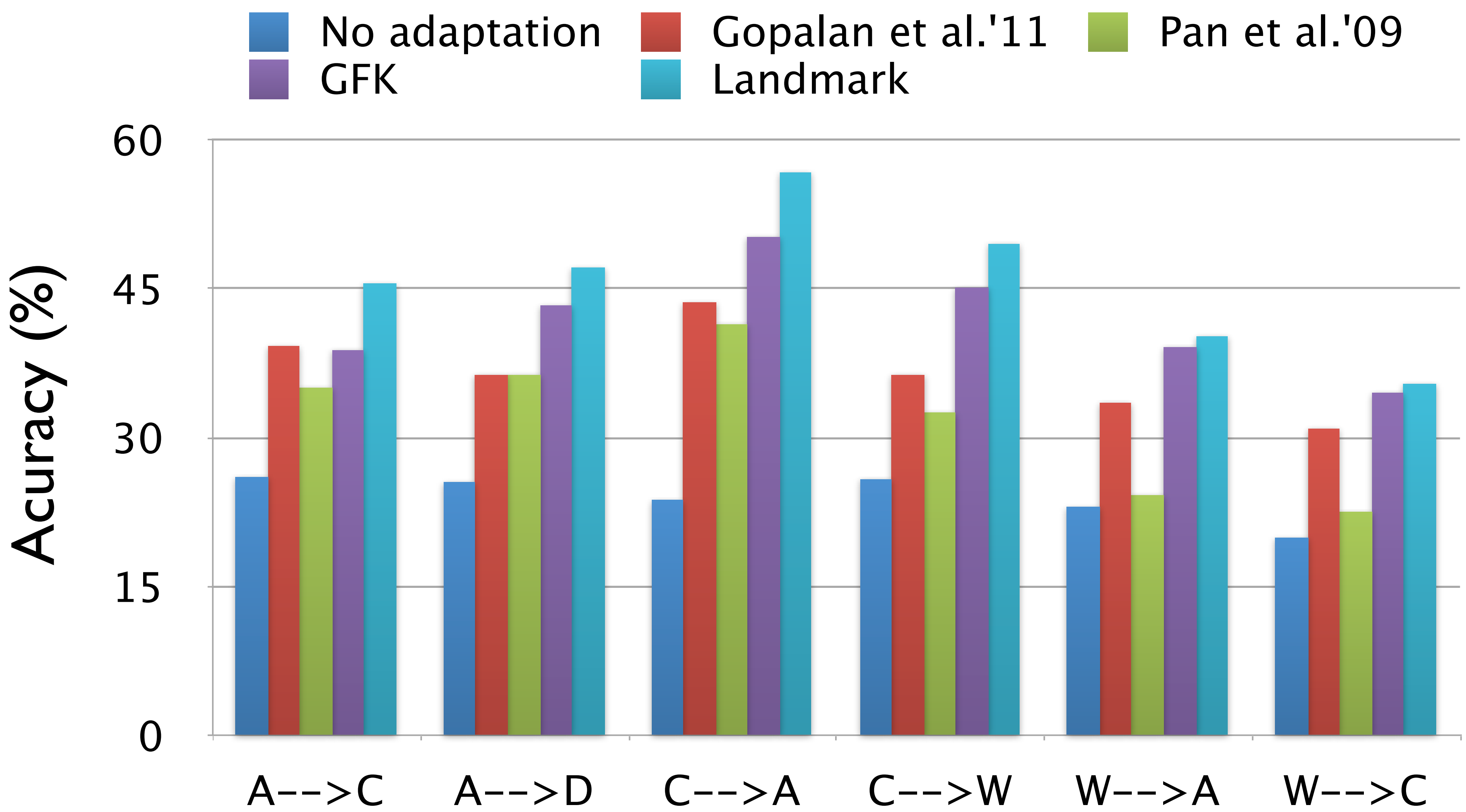
Object recognition



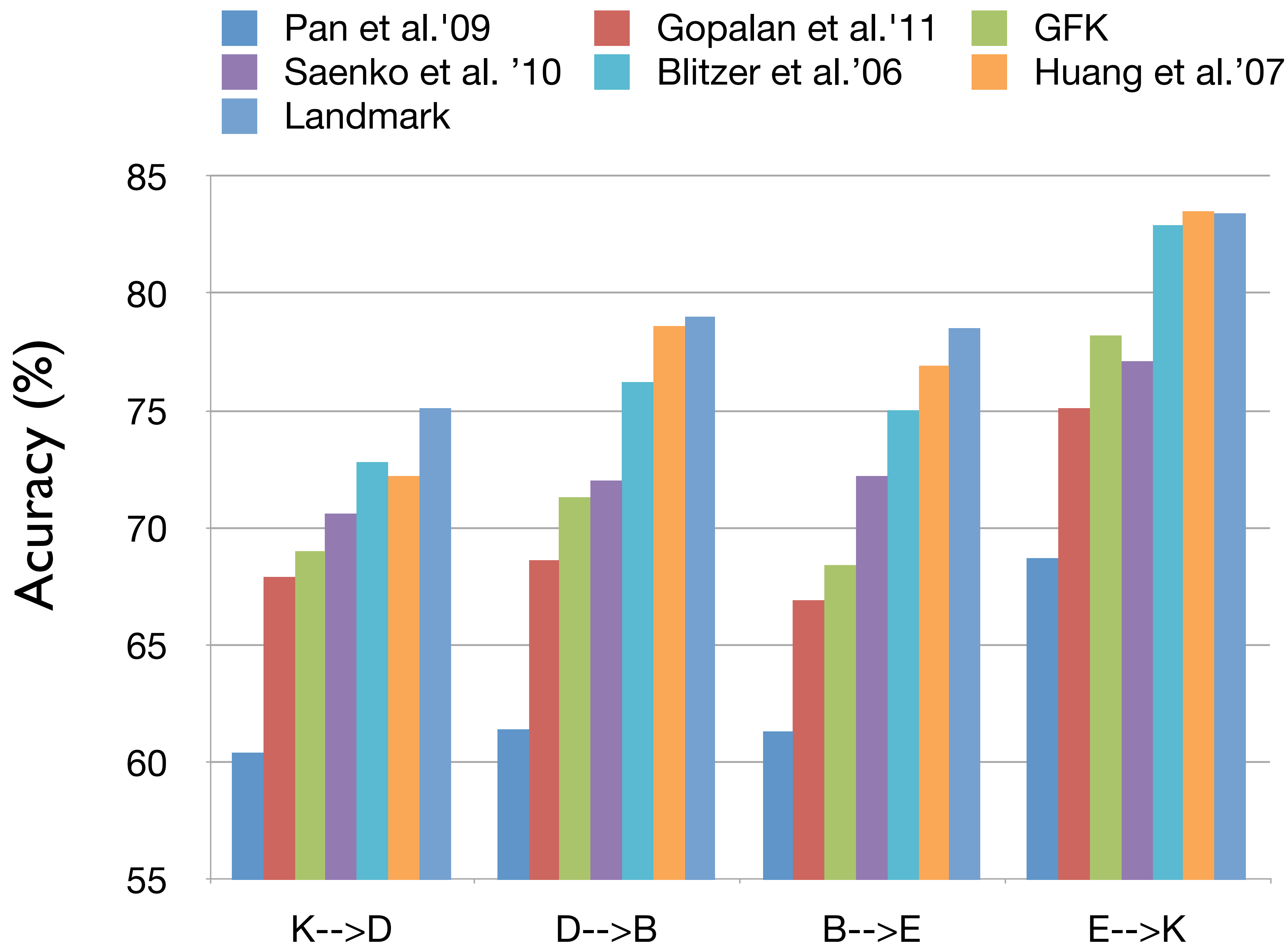


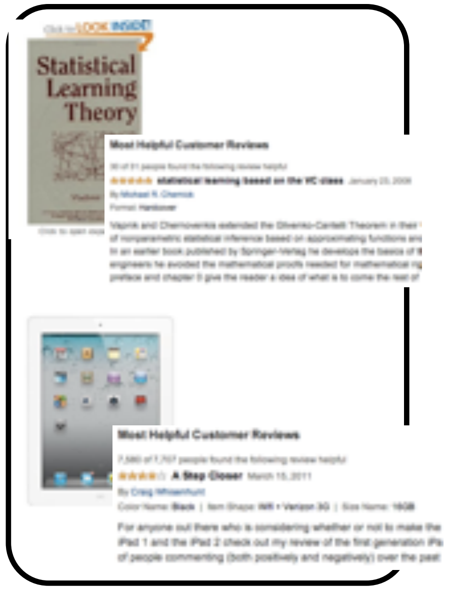
Object recognition





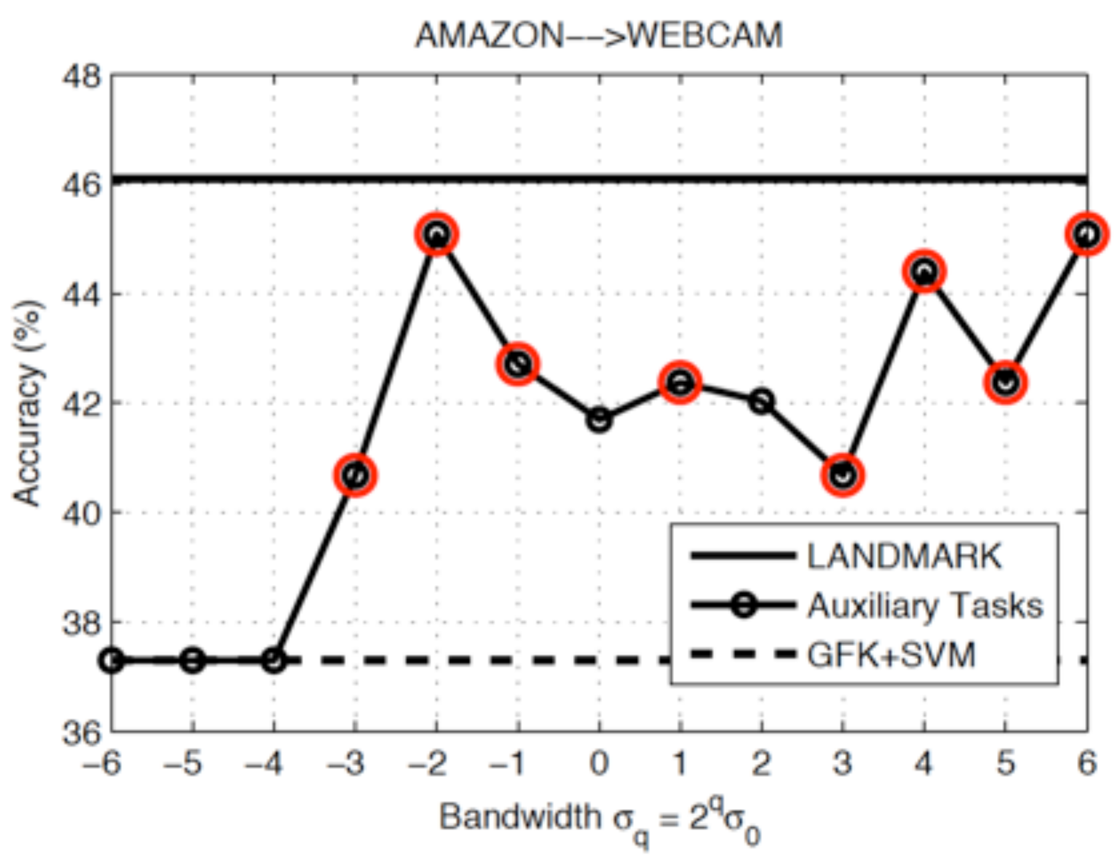
Sentiment analysis

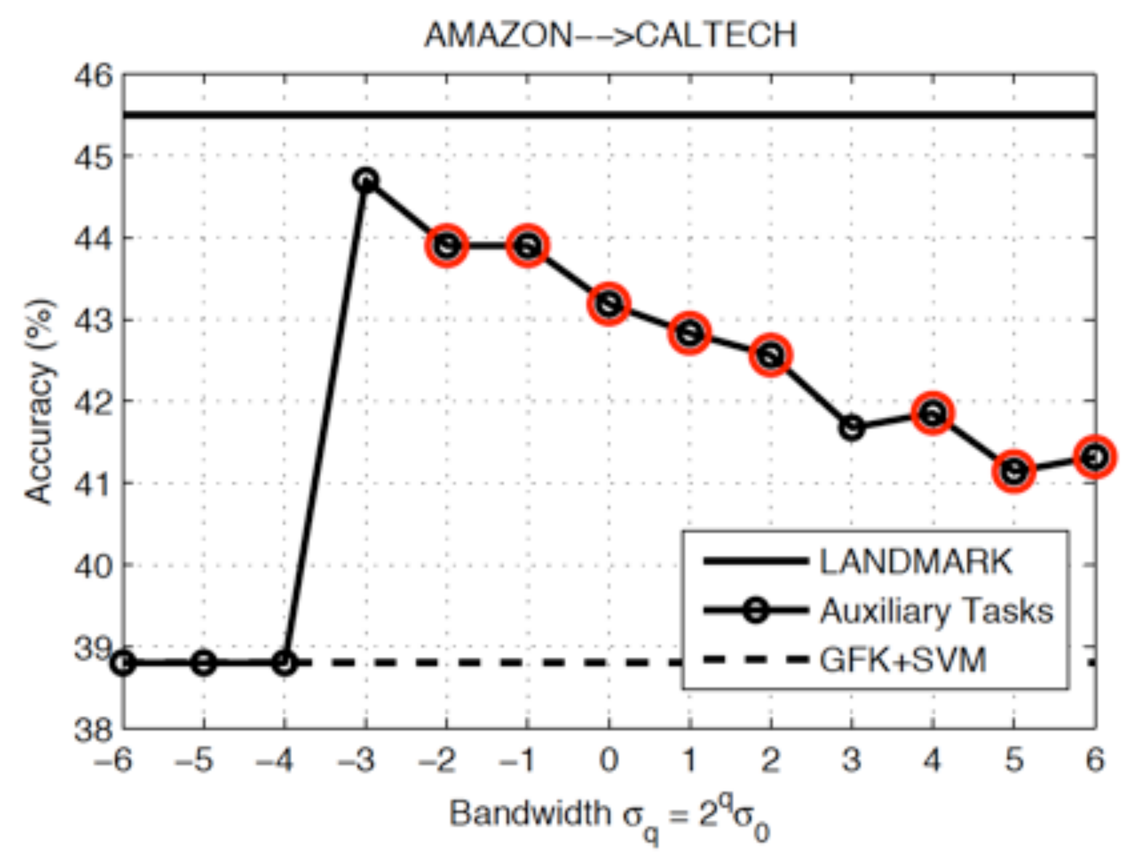




Auxiliary tasks easier to solve

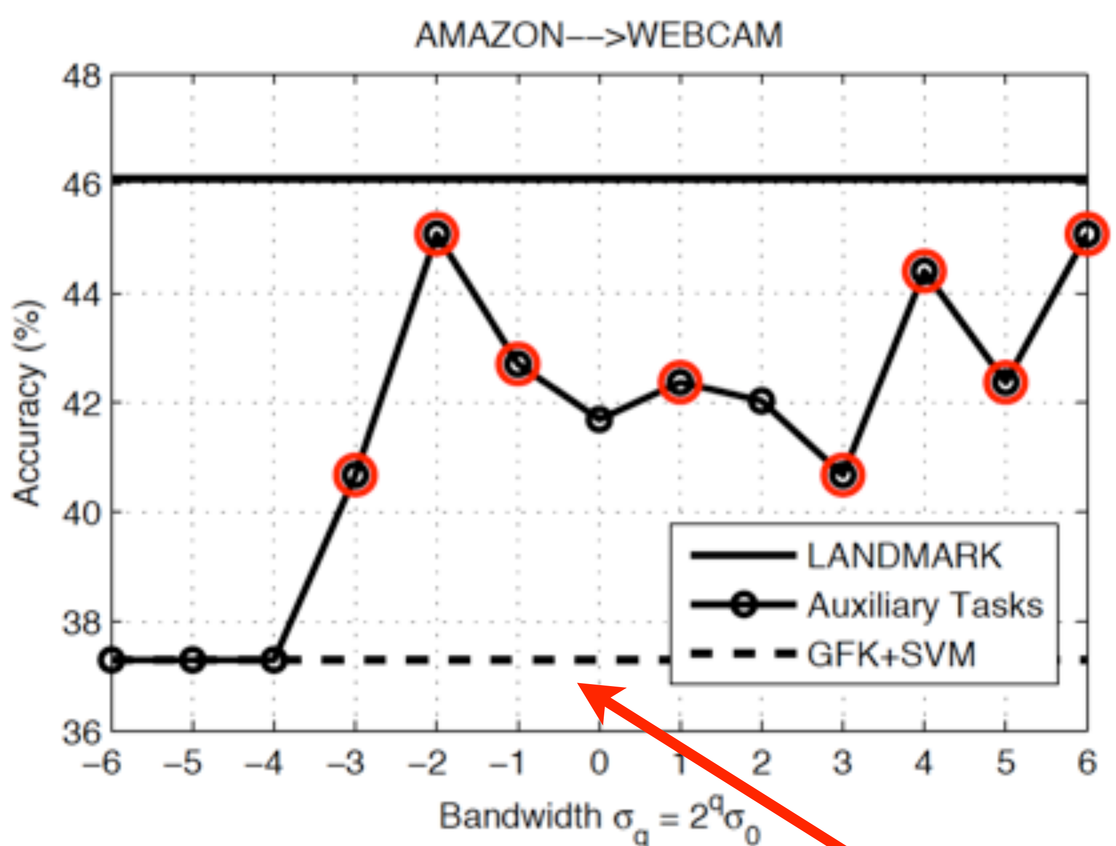
Empirical results on visual object recognition

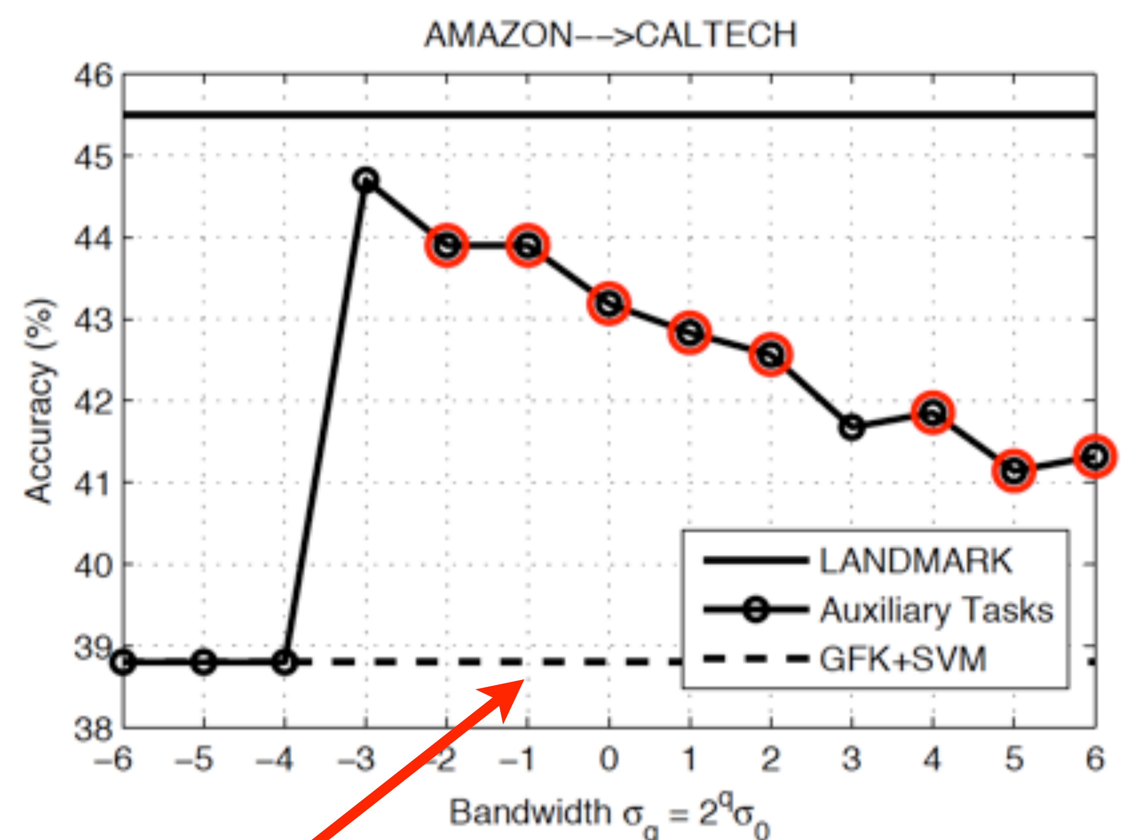




Auxiliary tasks easier to solve

Empirical results on visual object recognition

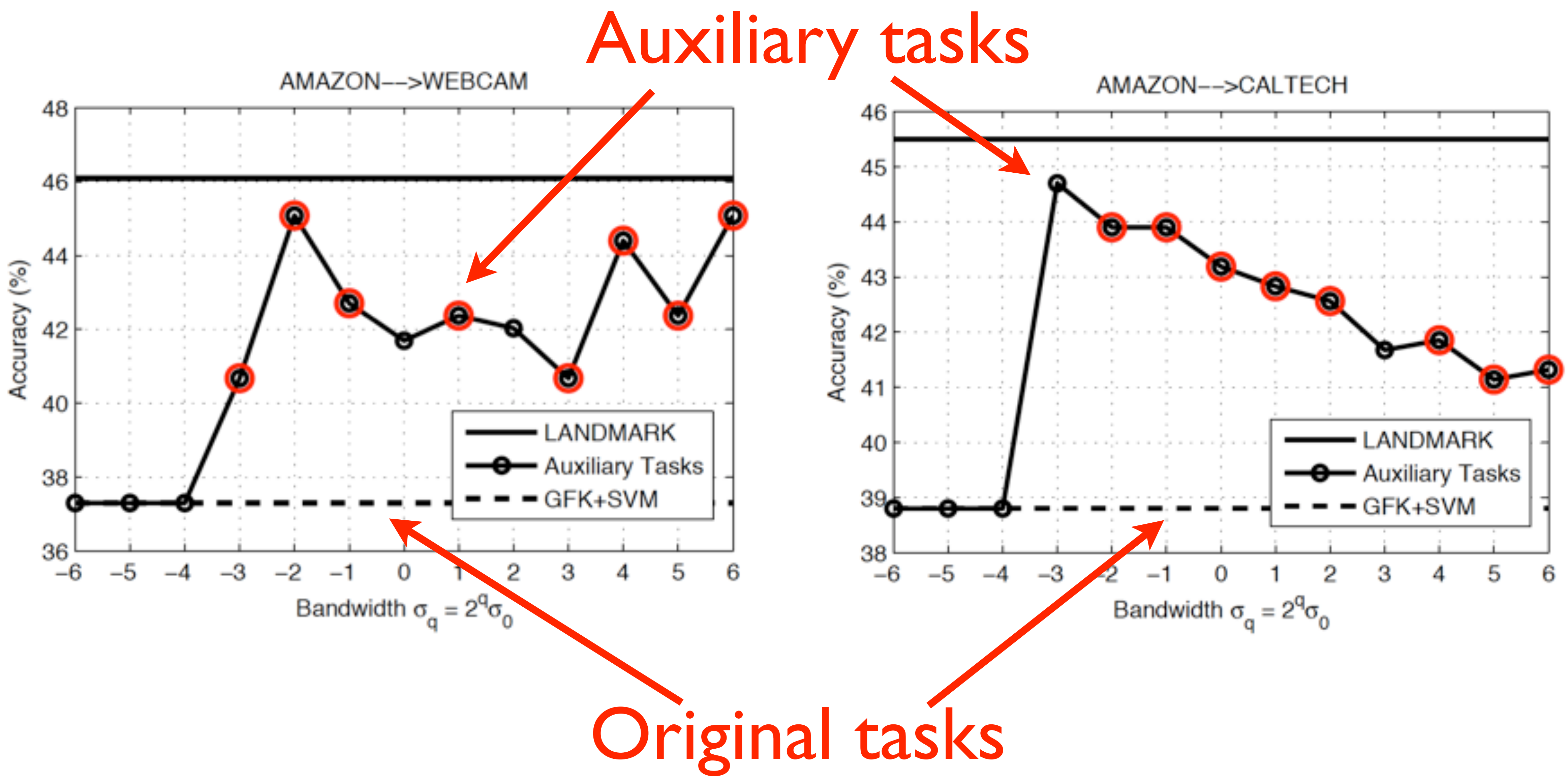




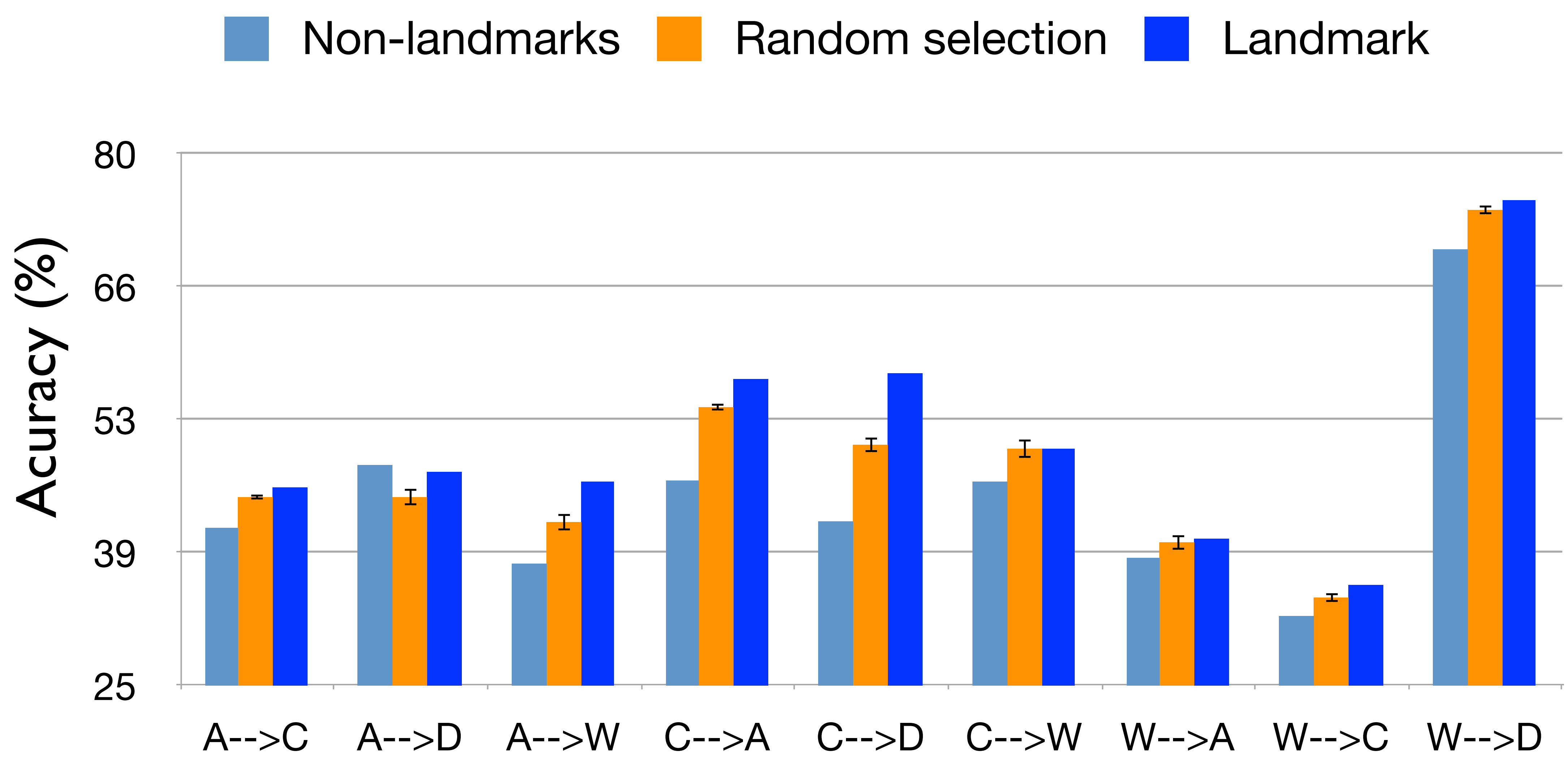
Original tasks

Auxiliary tasks easier to solve

Empirical results on visual object recognition



Landmarks good proxy to target *discrimination*



Summary

landmarks

an intrinsic structure, shared between domains

labeled source instances
distributed similarly to the target

auxiliary tasks provably easier to solve

discriminative loss despite unlabeled target

Outperformed the state-of-the-art

What do landmarks look like?

Dropping class balance constraint $P_{\mathcal{L}}(Y|X) = P_{\mathcal{S}}(Y|X)$?