Reshaping Datasets for Unsupervised Domain Adaptation

Boqing Gong

UCF Center for Research in Computer Vision

Joint work with Kristen Grauman and Fei Sha

Data-centric era

Experiments, observations, and simulations in science

Internet of things
Sensors everywhere

140 billion images, ***12M hourly***

300 hour new video every minute

200B tweets yearly, 500M daily

Great sources of discovery and knowledge

Google predicted flu outbreak two weeks before CDC, and now they collaborate.

twitter correctly predicted 2012 presidential election.

waze GPS provides real-time traffic information.

cytolon matches cancer patients to cord-blood donors in real-time.

Challenges

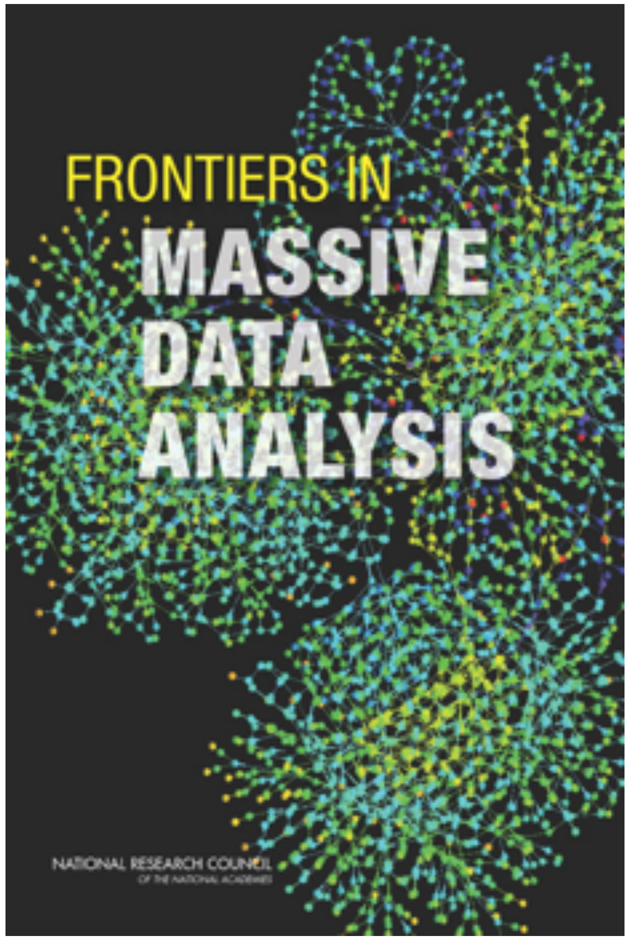
Dealing with highly distributed data

Coping with sampling biases and heterogeneity

Exploiting parallel and distributed architectures

Data visualization, integration, validation, security, sharing, etc.

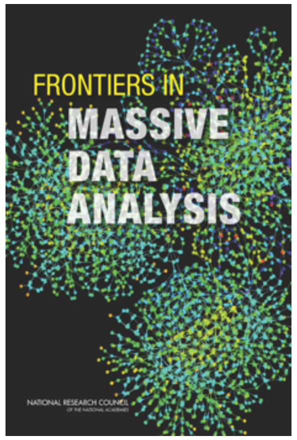
… …



National Academies Report

Sampling bias & heterogeneity

"(training) Data may have been collected according to a certain criterion …, but (testing) the inferences and decisions may refer to a different sampling criterion."



National Academies Report

Self-driving car: a case study

Self-driving car: a case study

Pedestrian detection and avoidance system

Sampling bias ➔ Performance significantly degrades [Dollár et al.'09]

The perils of mismatched domains

Cause: standard assumption in machine learning

Same underlying distribution for training and testing

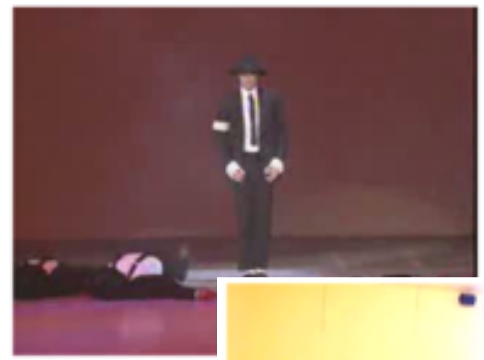
This is a realistic obstacle for autonomous systems

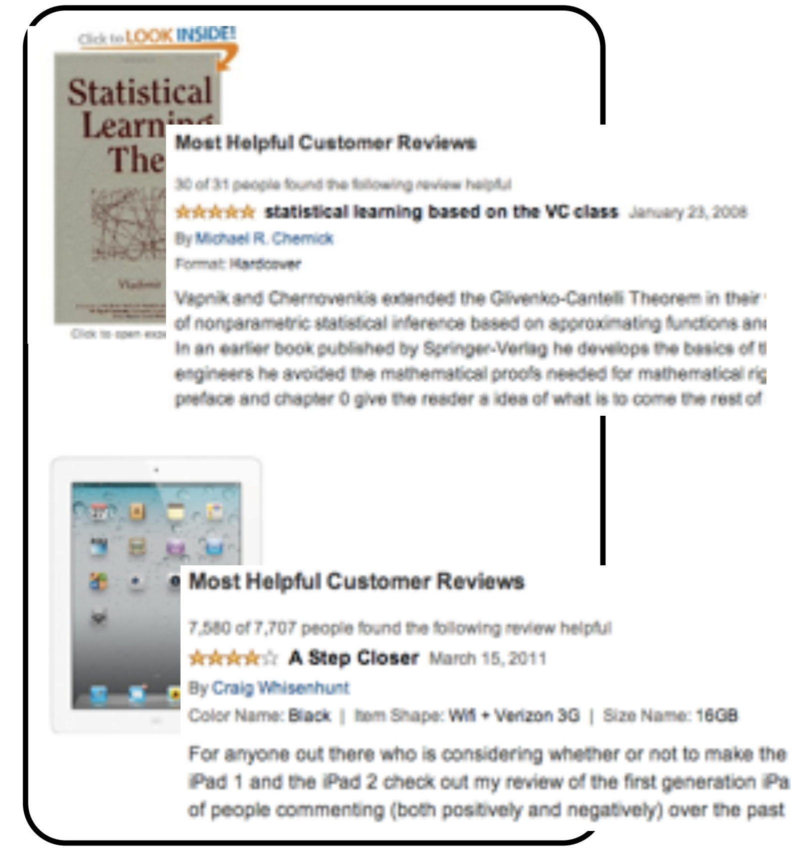
Systems often deployed to new environment, not lab reproducible

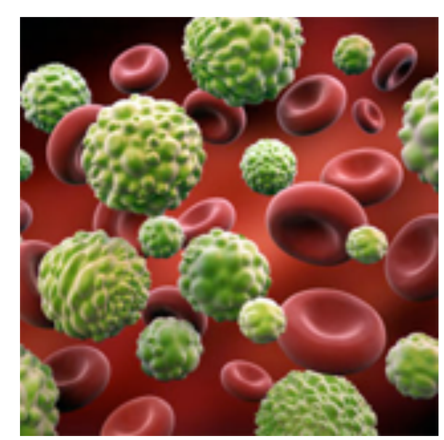
Expensive to collect training data from each type of target environment

Systems naturally degrade; environment dynamically evolves

Mismatches are common to many areas







Biology: different subjects

Abstract form: *unsupervised* domain adaptation (DA)

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)

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

Different distributions

Objective

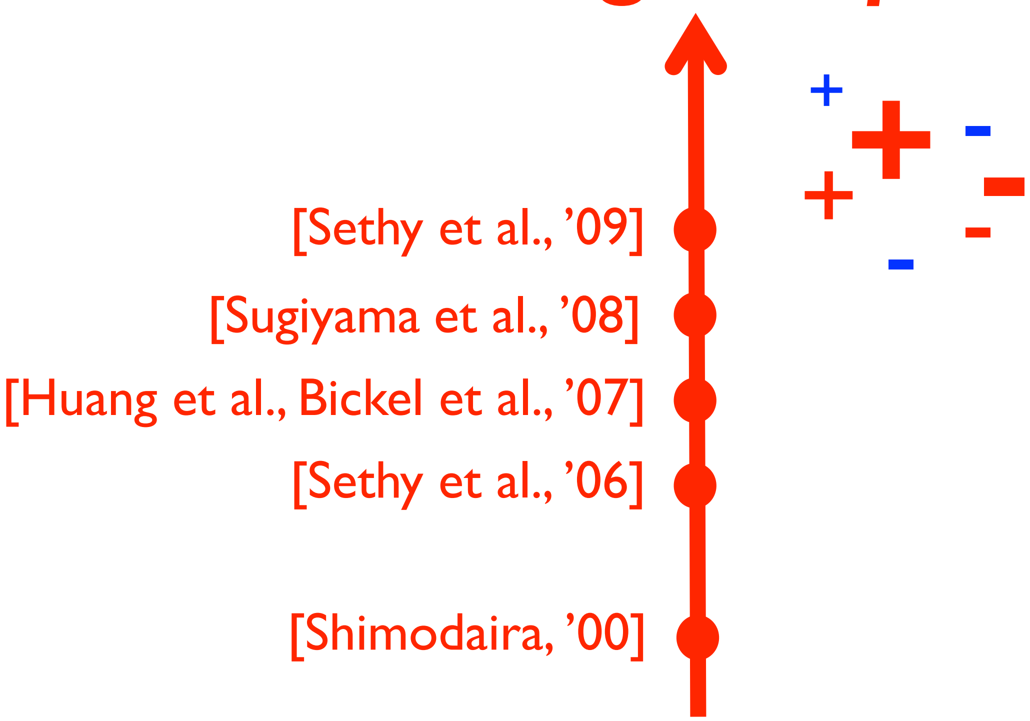
Learn models to work well on target

Background on DA

2009, 10	Computer vision: classification
2000s	Machine learning, NLP: DA, covariate shift, sampling bias
1990s	Speech: speaker adaptation
1970s	Statistics & econometrics: **sampling bias**

Background - brief review

Correcting *sampling* bias

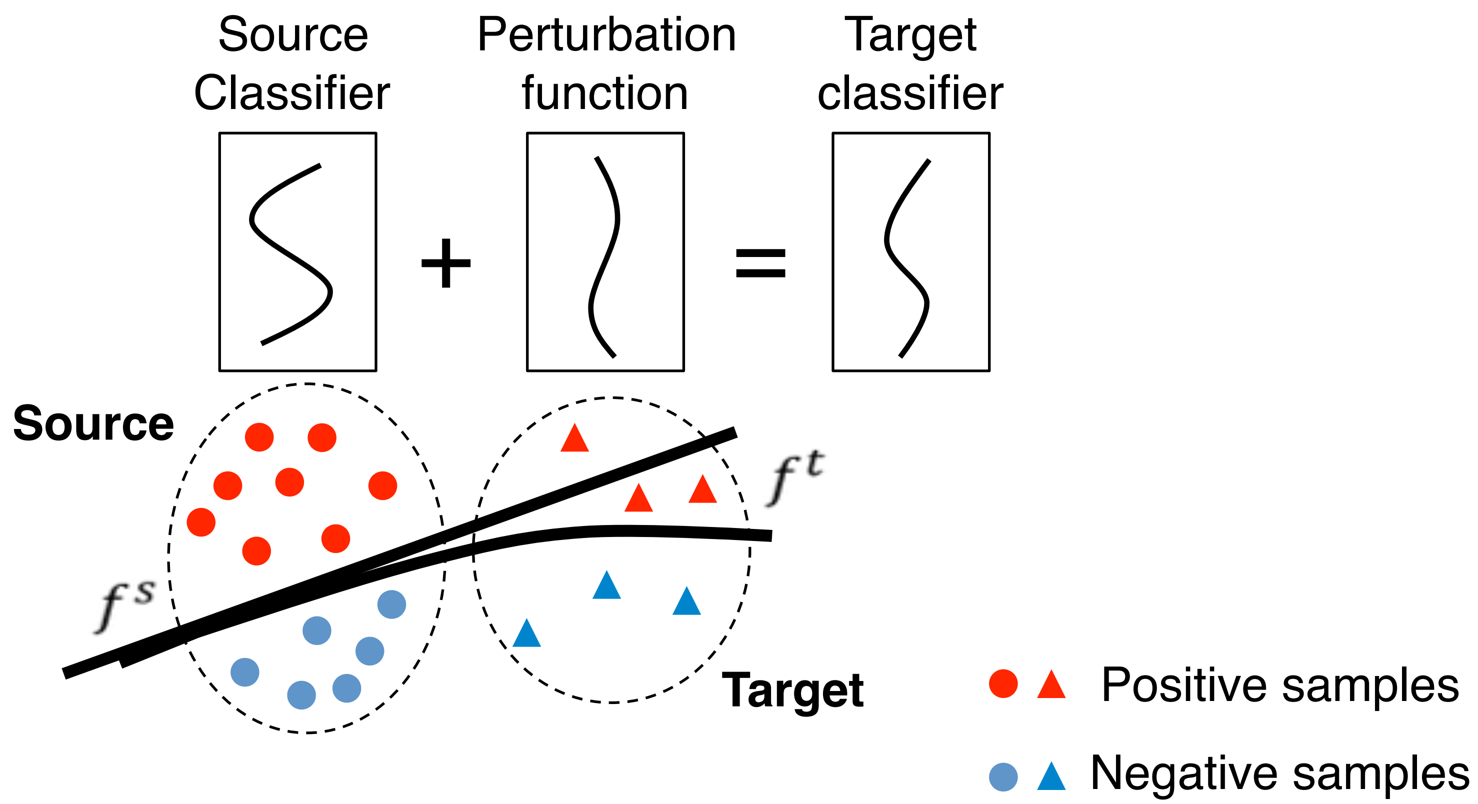


Re-weight source instances

$$\mathbb{E}_{\mathcal{T}}[h(\mathbf{x}) \neq y] \approx \mathbb{E}_{\mathcal{S}}\, \omega(\mathbf{x})\ [h(\mathbf{x}) \neq y]$$

$$\omega(\mathbf{x}) : \text{instance weight}$$

Background - brief review



[Evgeniou and Pontil, '05]

[Duan et al., '09]

[Duan et al., Daumé III et al., Saenko et al., '10]

[Kulis et al., Chen et al., '11]

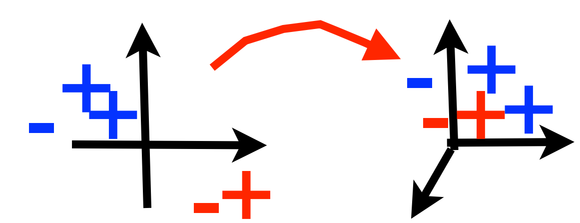
Adjusting mismatched ***models***

Background - brief review

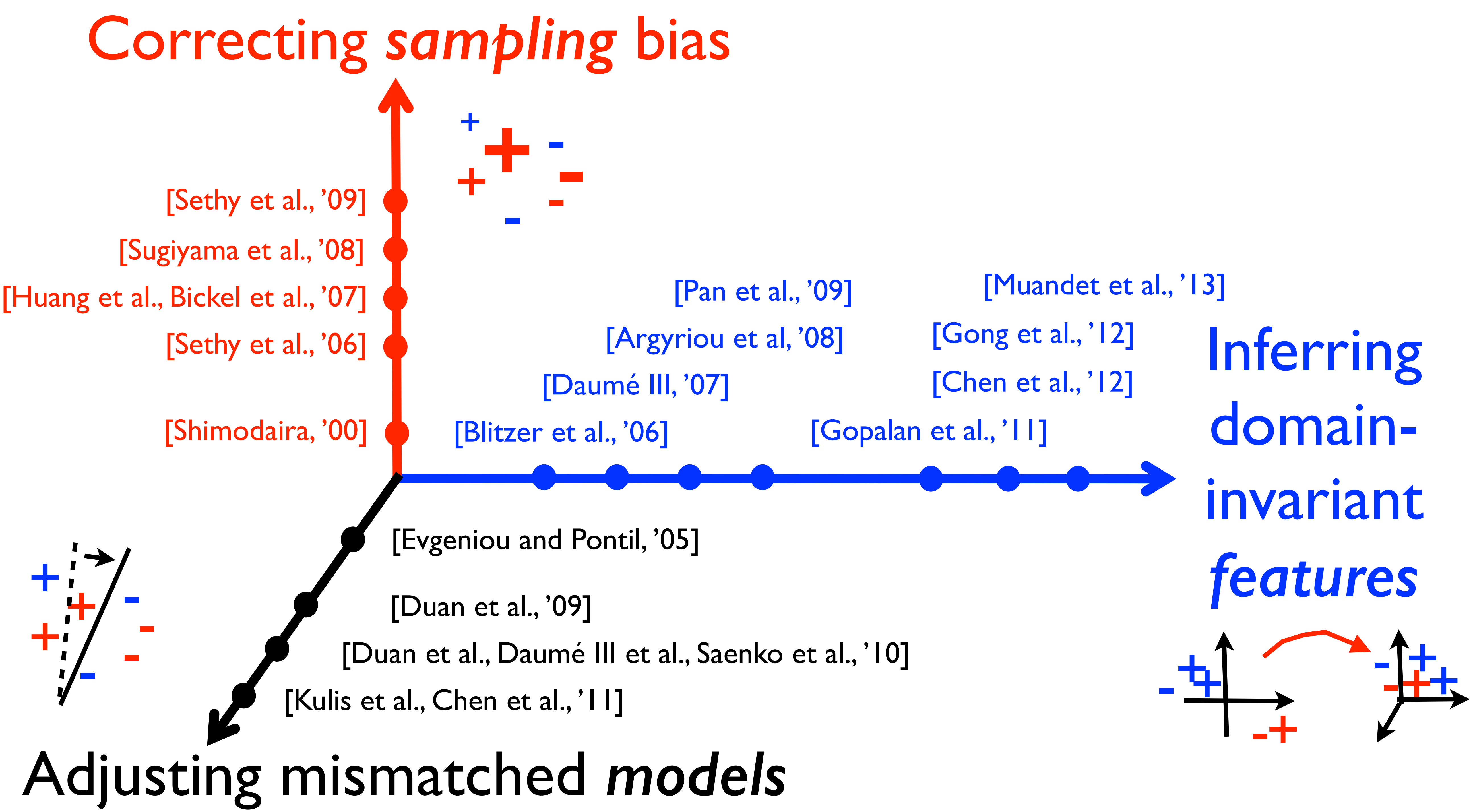
$$\mathbf{x} \mapsto \mathbf{z}, \quad \text{s.t.}$$

$$P_{\mathcal{S}}(z, y) \approx P_{\mathcal{T}}(z, y)$$

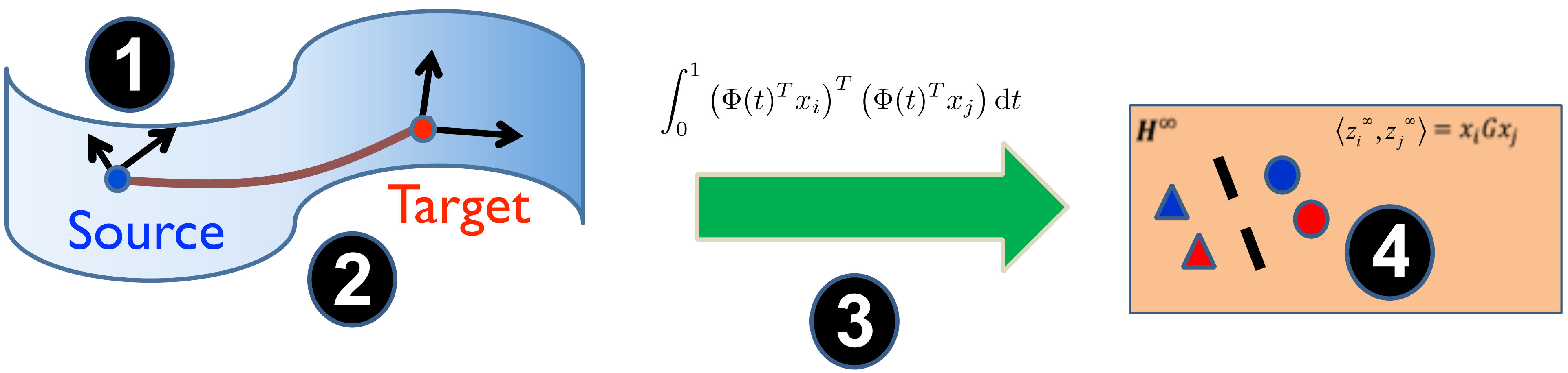
Inferring domain-invariant *features*



Background - quick review



GFK: inferring a domain-invariant feature space



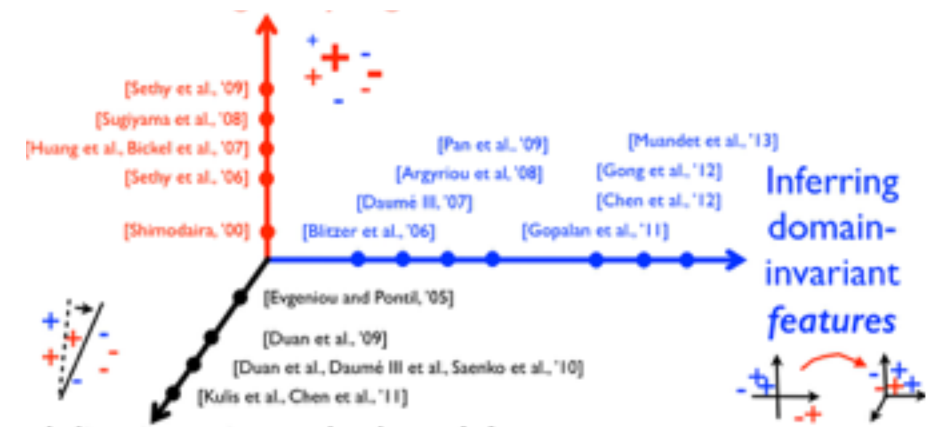
1. Exploit subspace structure in data
2. Model domain shift with geodesic flow
3. **Derive a domain-invariant kernel**
4. Classify target data in the kernel space

[Gong *et al.*, CVPR'12]

Key to domain adaptation

"*to reduce source-target discrepancy*"

Snags in previous methods



Forced adaptation

Attempting to adapt **all source instances**, including "hard" ones

Implicit discrimination

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

Key to domain adaptation

"***to reduce source-target* domain *discrepancy***"

What is a source domain?

Is it always fixed?

Can we ***reshape*** it?

What constitutes a domain?

In speech and NLP:

Speakers

Languages

Article topics

…other **factors**

In computer vision:

Factors?

Pose

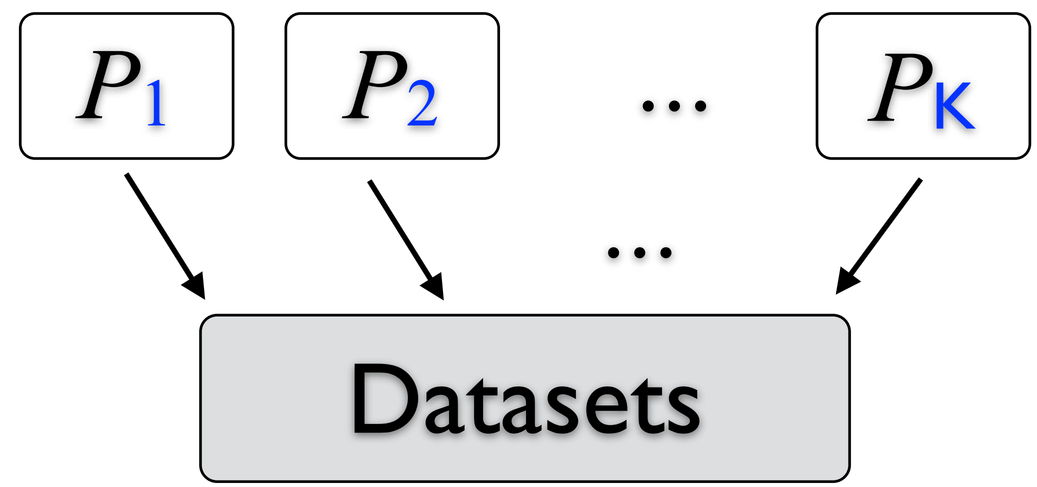
Lighting

Fore/ Background

Occlusion

Many factors overlap & interact

Some questions revolving around "domain"



Adapt-abilities
of different domains
[Gong et al., IJCV'14, CVPR'12]

What is a domain?
Reshaping data according to domains from which they come?
[Gong et al., NIPS'13]

Our key insights

Forced adaptation from a prefixed source domain

➔ Select the best instances for adaptation

Implicit discrimination

➔ Approximate discriminative loss on target

Selecting most adaptable source instances

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

[Gong *et al.*, ICML'13]

Selecting most adaptable source instances

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

Source

Identifying landmarks:

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

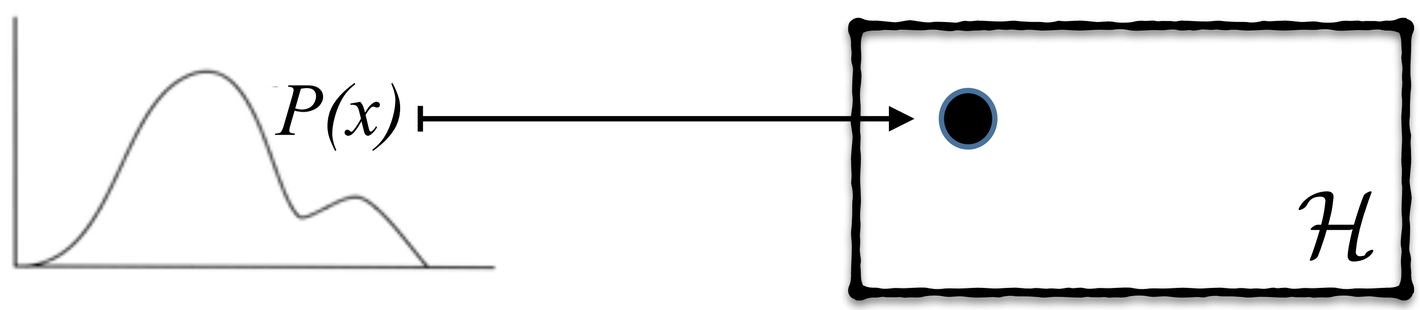
$$\min_{\text{landmarks}} \quad d(P_{\mathcal{L}}, P_{\mathcal{T}})$$

Target

[Gong *et al.*, ICML'13]

Kernel embedding of distributions

$$\mu[P] \triangleq \mathbb{E}_x[\phi(x)]$$



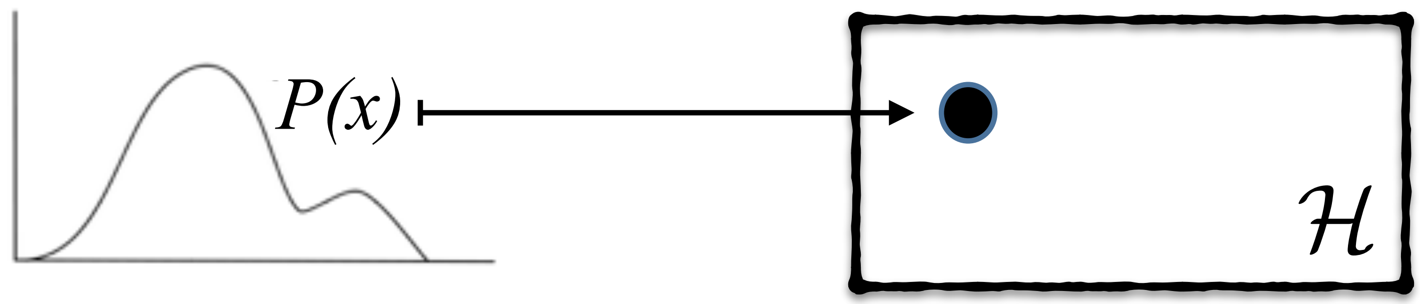
μ maps distribution P to Reproducing Kernel Hilbert Space

μ is injective if $\phi(\cdot)$ is characteristic

[Müller'97, Gretton et al.'07, Sriperumbudur et al.'10]

Kernel embedding of distributions

$$\mu[P] \triangleq \mathbb{E}_x[\phi(x)]$$



Empirical kernel embedding:

$$\hat{\mu}[P] = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \quad x_i \sim P$$

Identifying landmarks by matching kernel embeddings

Integer programming

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

where

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

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

Solving by relaxation

Convex relaxation

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

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

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

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 [Gong et al., ICML'13, IJCV'14] for details)

What do landmarks look like?

Landmark based domain adaptation

Identifying landmarks → New source-target adaptation task → Adaptation by GFK

Experimental study

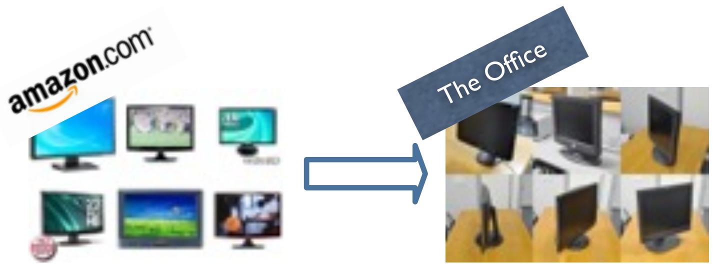
Four vision datasets/domains on visual object recognition

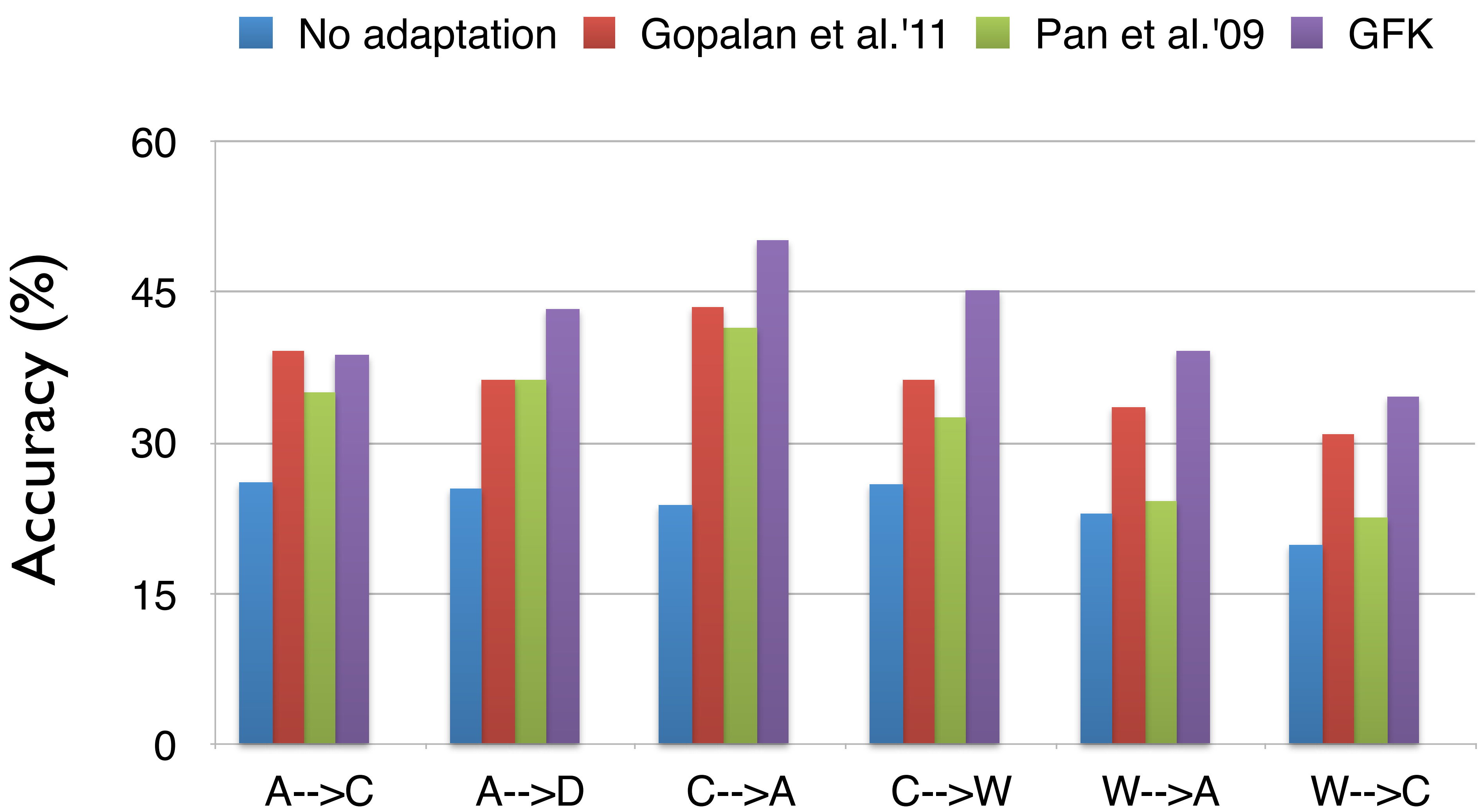
[Griffin et al. '07, Saenko et al. 10']

Four types of product reviews on sentiment analysis

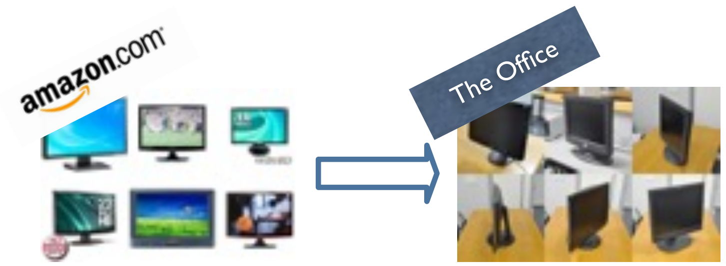
Books, DVD, electronics, kitchen appliances [Biltzer et al. '07]

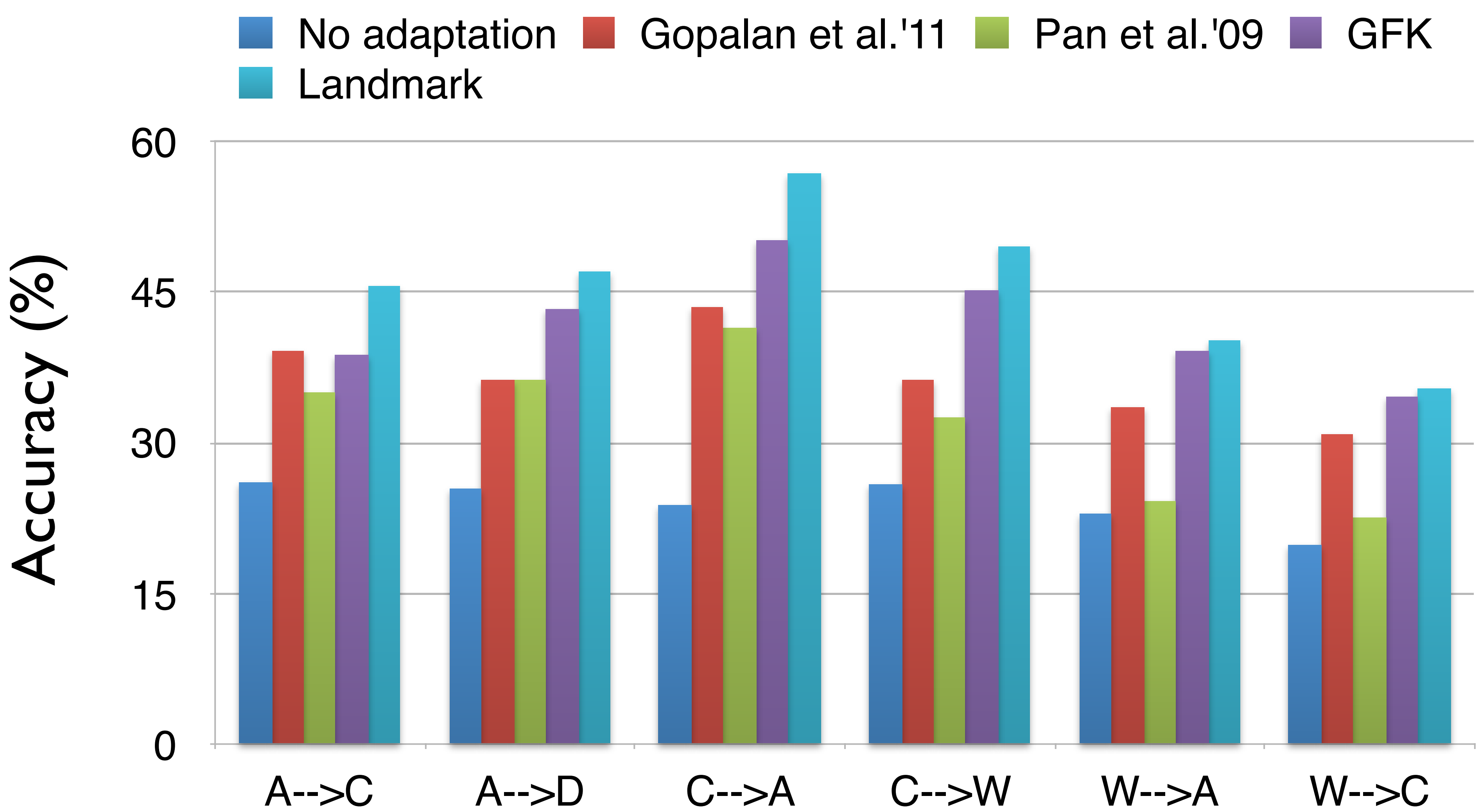
Comparison results: object recognition



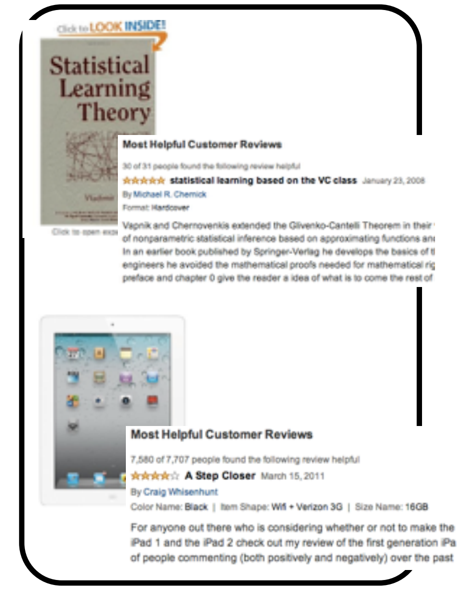


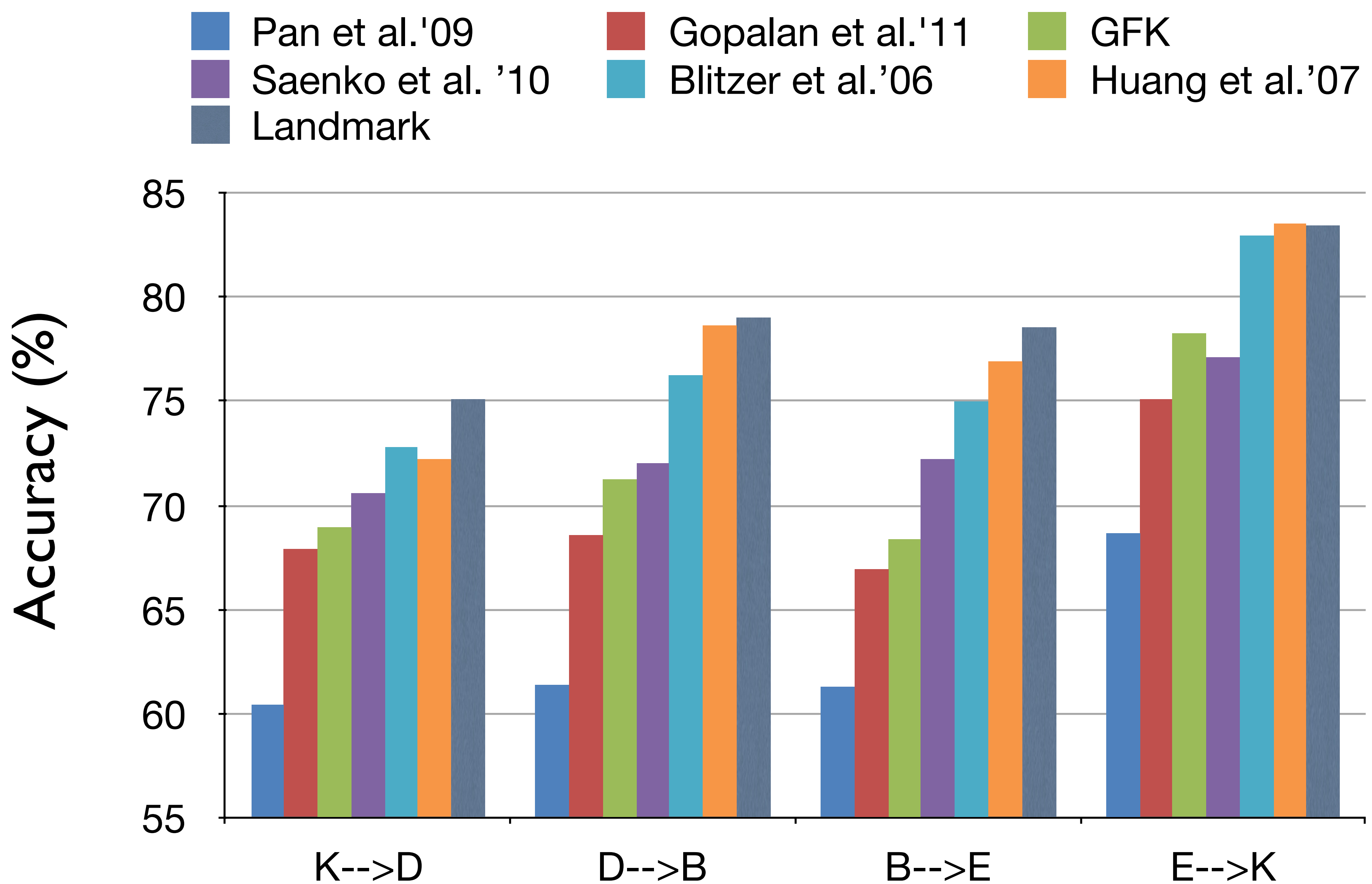
Comparison results:
object recognition



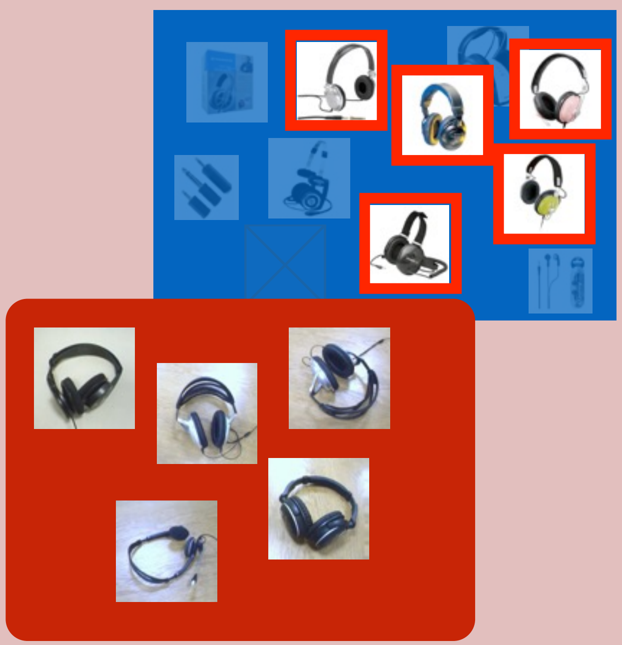


Comparison results: sentiment analysis





Summary - Landmarks



Landmarks

[Gong *et al.*, ICML'13]

- *Labeled source instances, distributed similarly to target*
- *Better approximation of discriminative loss of target*
- *Automatically identifying landmarks*
- *Benefiting other adaptation methods*

Key to domain adaptation

“*to reduce* *source*-*target* domain *discrepancy*”

What is a source domain?

Landmarks: reshaped target-oriented source

What if no a priori knowledge about target?

What constitutes a domain?

Amazon images from [Saenko et al.'10].

What constitutes a domain?

Domain I

Domain II

Two axiomatic properties for latent domains

I. Maximum **distinctiveness**:

Identifying distinct domains maximally different **in distribution** from each other

II. Maximum **learnability**

Being able to derive strong discriminative models from the identified domains

[Gong *et al.*, NIPS'13]

I. Maximum **distinctiveness**

Domains maximally different in distribution from each other

$$\max_{\{z_{mk}\}} \quad \sum_{k \neq k'} \hat{d}(P_k, P_{k'}; \{z_{mk}\})$$

$$z_{mk} = \begin{cases} 1 & \text{if } x_m \in \text{the } k\text{-th domain} \\ 0 & \text{else} \end{cases}$$

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

II. Maximum **learnability**

Able to learn strong classifiers from domains

Within-domain cross-validation

$$\text{Accuracy}(\mathsf{K}) = \sum_{k=1}^{\mathsf{K}} \frac{\mathsf{M}_k}{\mathsf{M}} \text{Accuracy}_k$$

- Determining the number of domains K

Hard to manually define discrete domains

Amazon images from [Saenko et al.'10].

Our "reshaped" domains

Adapting from discovered domains > from datasets

Domain I **Domain II**

Summary - latent domains

Many factors overlap & interact

What constitutes domains?

[Gong *et al.*, NIPS'13]

- *Dataset ≠ domain*
- *Suboptimal to use DA methods for cross-dataset problem*
- *Discovering latent domains:*
  - *maximum distinctiveness*
  - *maximum learnability*

Key to domain adaptation

"*to reduce source-target discrepancy*"

What is a source domain?

- Landmarks: reshaped target-oriented source
- Discovering latent domains without target *a priori*

"to define domains / to reshape data well"

Thanks!

"*to reduce source-target discrepancy*"

What is a source domain?

Landmarks: reshaped target-oriented source

Discovering latent domains without target *a priori*

"to define domains / to reshape data well"