Adaptation for Objects and Attributes

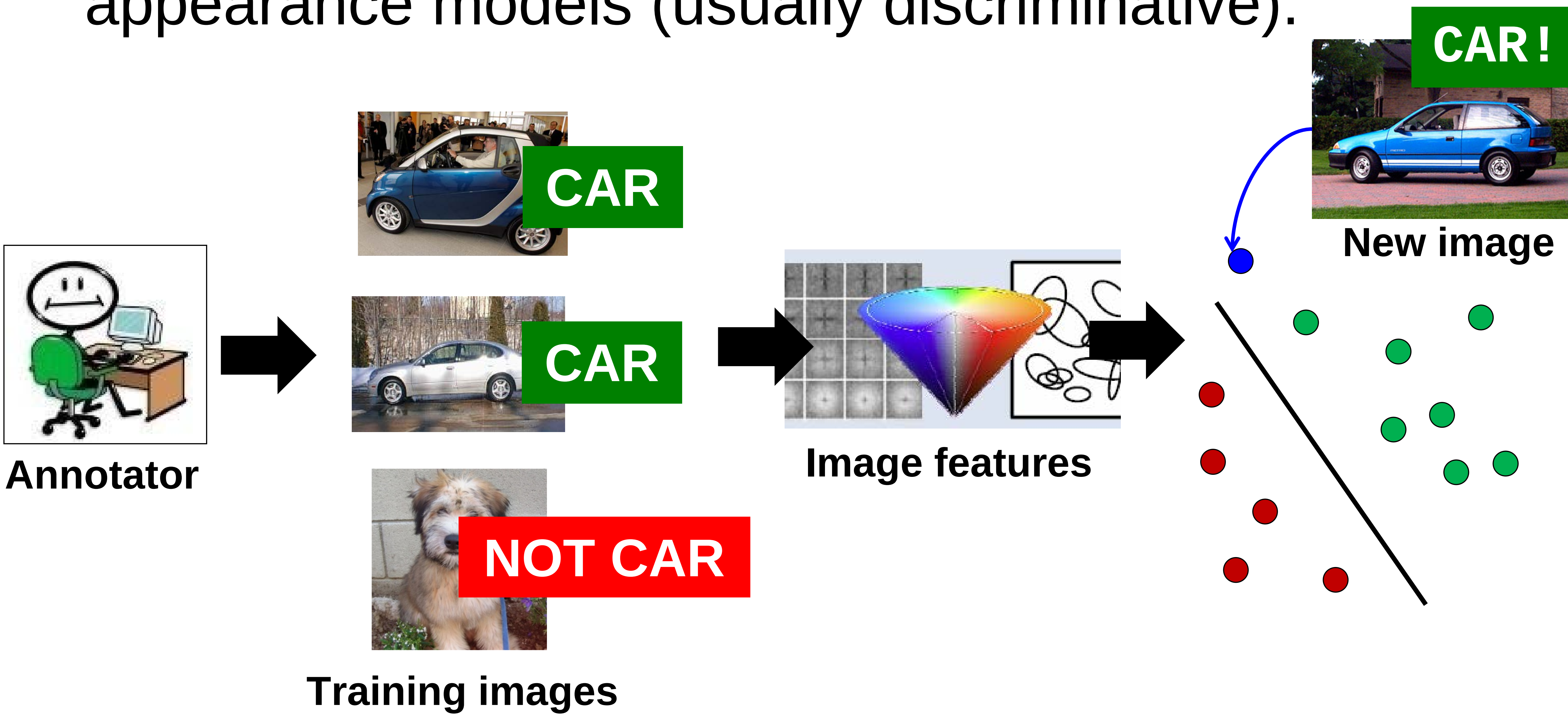
Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Adriana Kovashka (UT Austin),
Boqing Gong (USC), and Fei Sha (USC)

THE UNIVERSITY OF TEXAS AT AUSTIN

Learning-based visual recognition

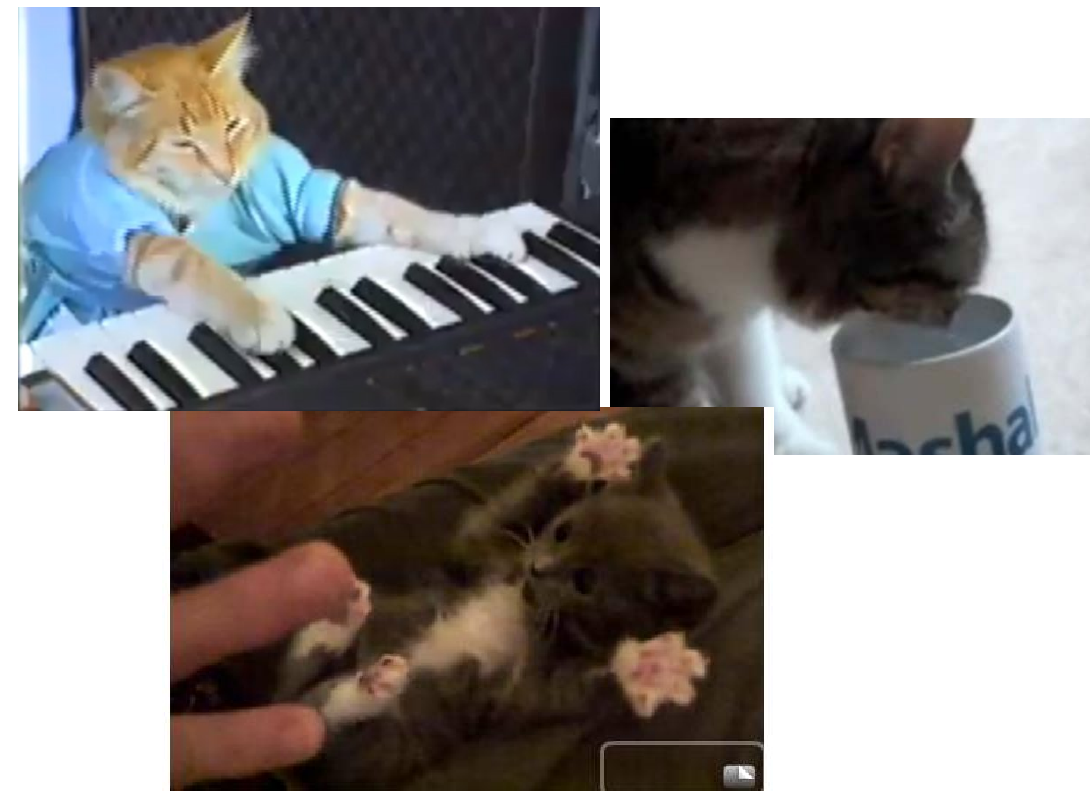
Last 10+ years: impressive strides by *learning* appearance models (usually discriminative).



Typical assumptions

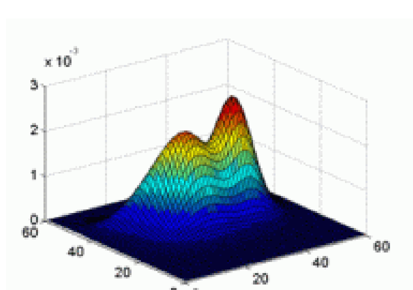
1. Test set will look like the training set.
2. Human labelers "see" the same thing.

Mismatched domains

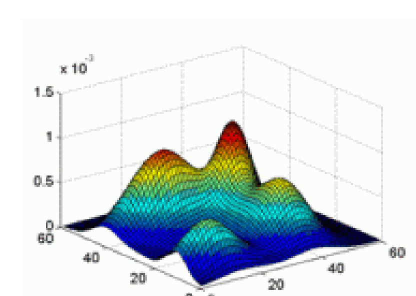


TRAIN → **TEST**

Flickr



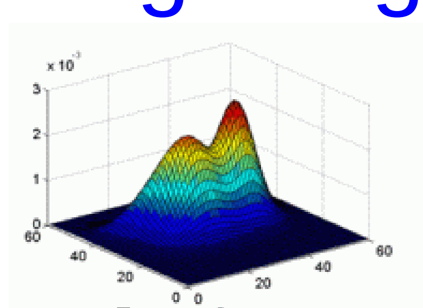
YouTube



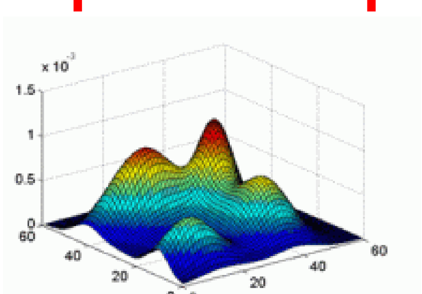
Mismatched domains

TRAIN → **TEST**

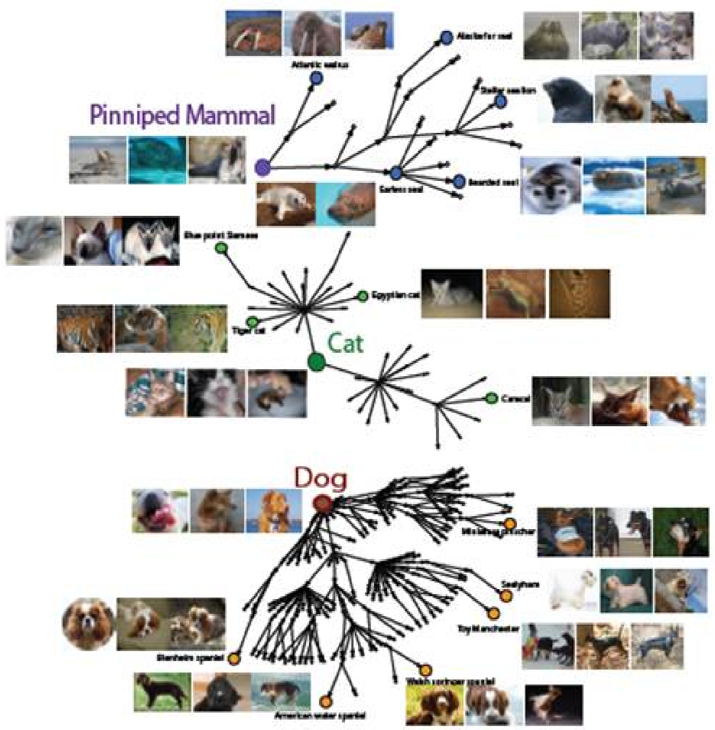
Catalog images



Mobile phone photos

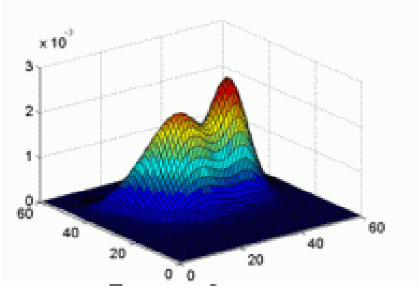


Mismatched domains

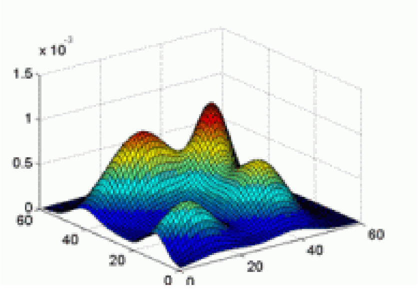


TRAIN → **TEST**

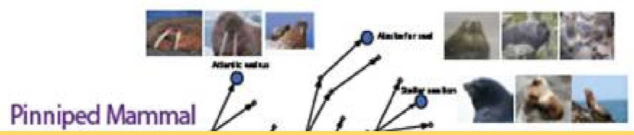
ImageNet



PASCAL VOC



Mismatched domains

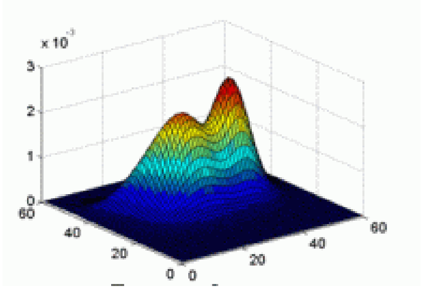


"It is worthwile to note that, even with 140K training ImageNet images, we do not perform as well as with 5K PASCAL VOC training images."

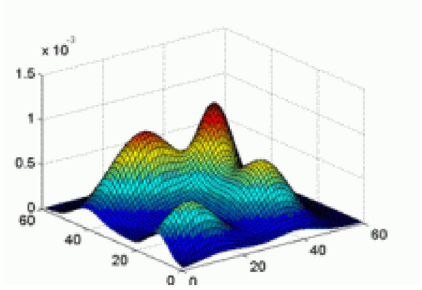
– Perronnin et al. CVPR 2010

TRAIN → **TEST**

ImageNet



PASCAL VOC



Mismatched domains

Problem: Poor cross-domain generalization

- Different underlying distributions
- Overfit to datasets' idiosyncrasies

Possible solution:

Unsupervised domain adaptation

Unsupervised domain adaptation

Setup

Source domain (with labeled data)

$$D_S = \{(x_m, y_m)\}_{m=1}^{M} \sim P_S(X, Y)$$

Target domain (no labels for training)

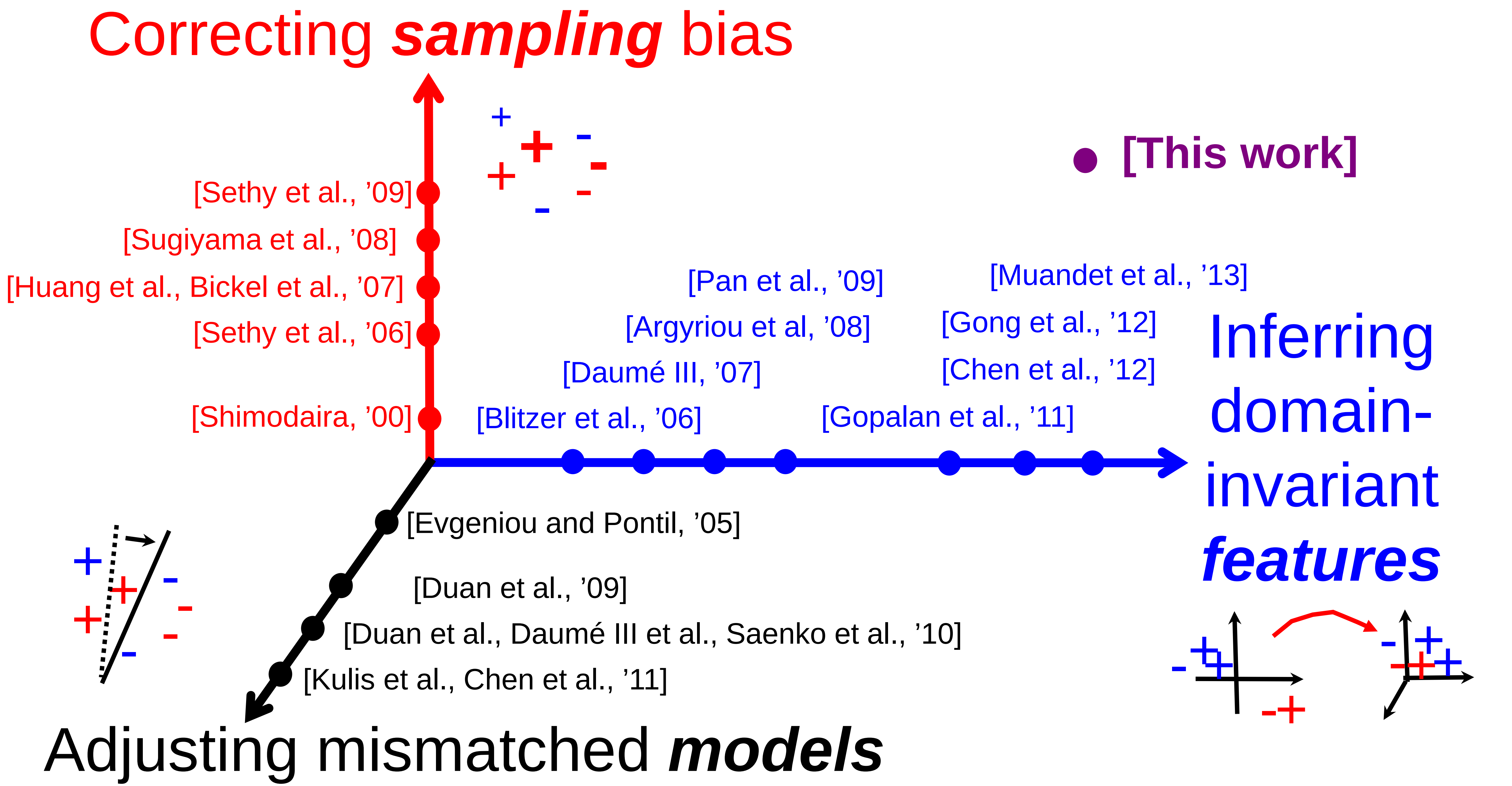
$$D_T = \{(x_n, y_n)\}_{n=1}^{N} \sim P_T(X, Y)$$

Different distributions

Objective

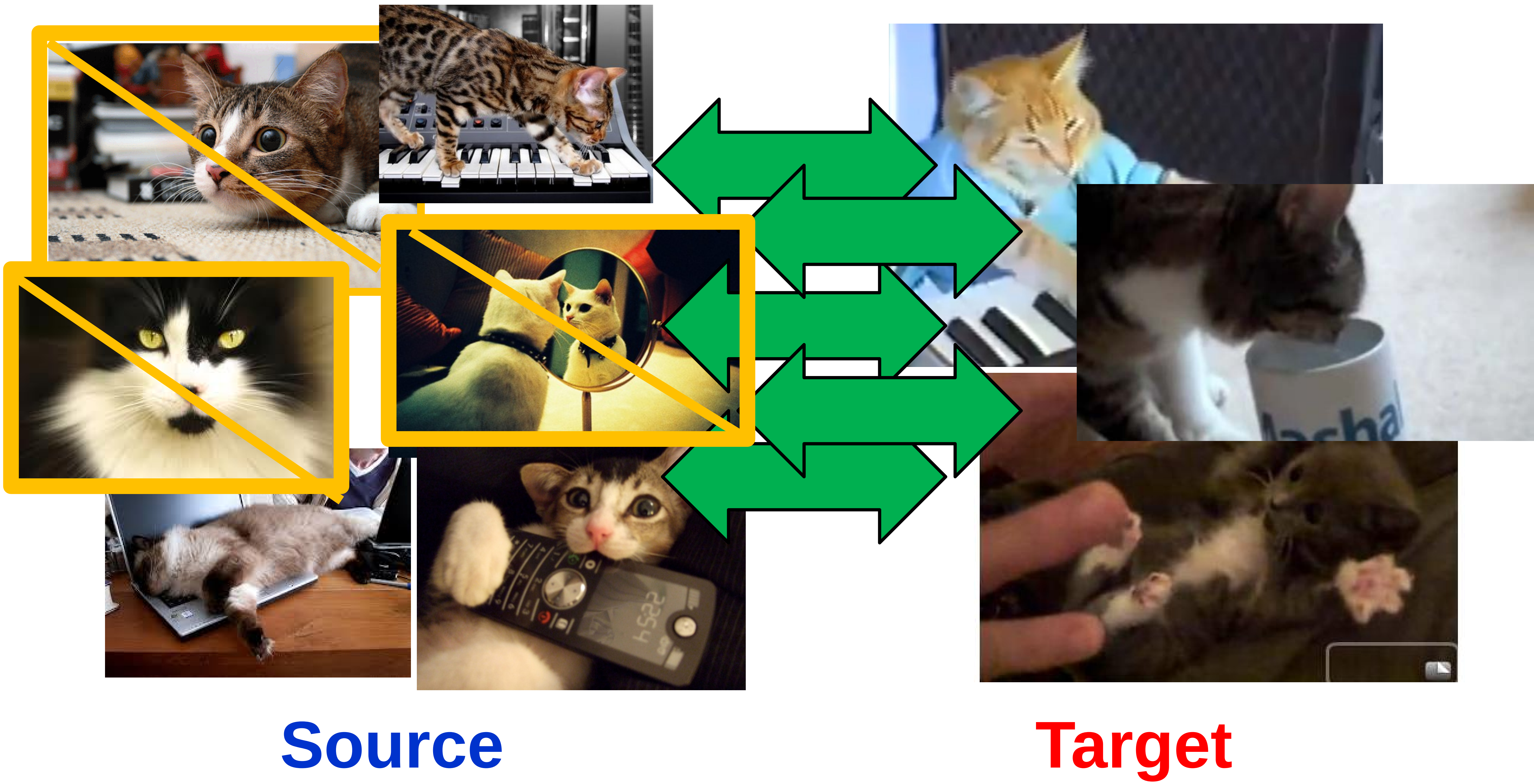
Learn classifier to work well on the target

Much recent research



Problem

Existing methods attempt to adapt *all* source data points, including “hard” ones.



Problem

Existing methods attempt to adapt *all* source data points, including “hard” ones.

Our idea

Automatically identify the “most adaptable” instances

Use them to create series of easier auxiliary domain adaptation tasks

[Gong et al., ICML 2013]

Landmarks

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

Source

Target

[Gong et al., ICML 2013]

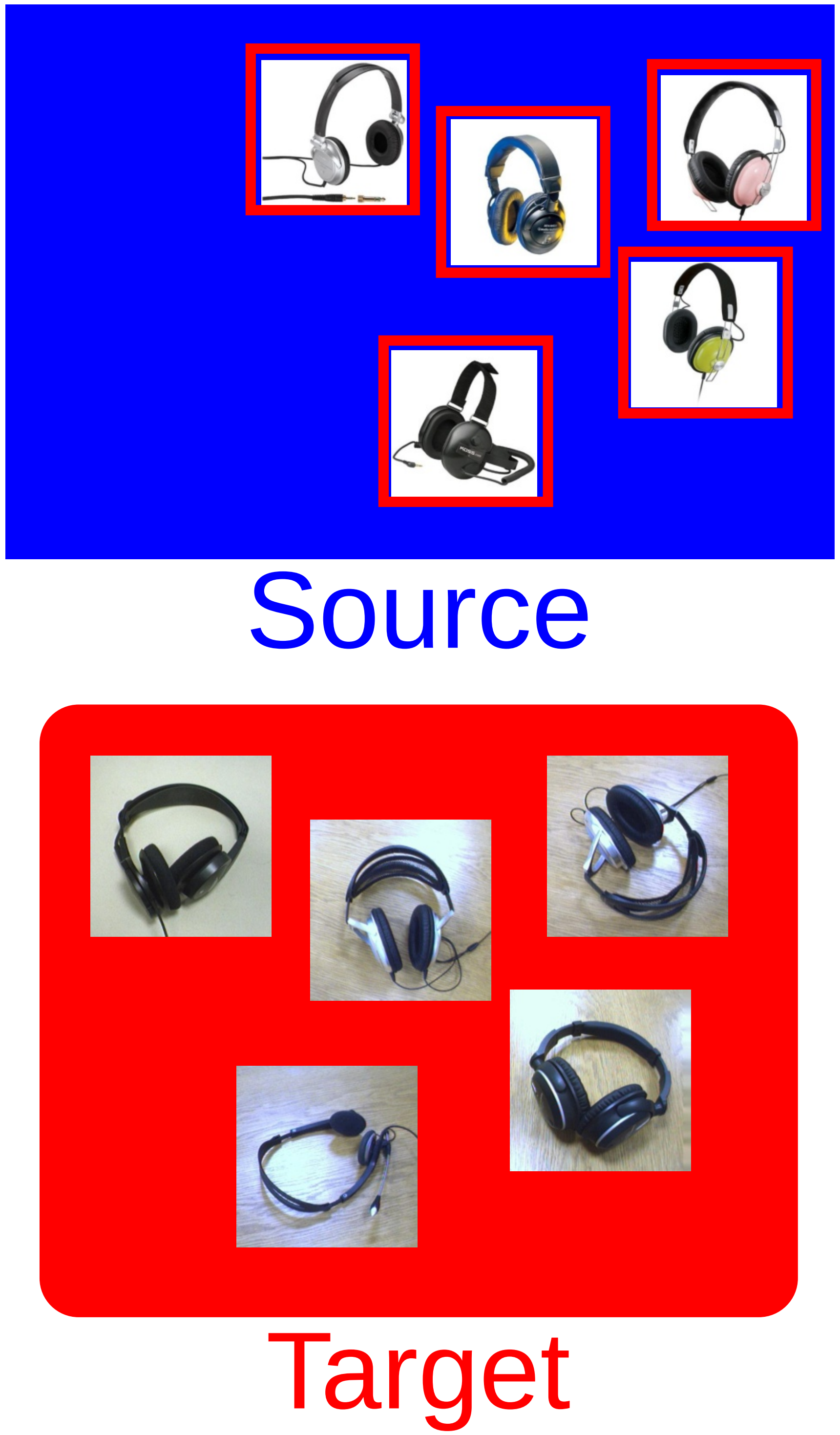
Landmarks

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

Roles:

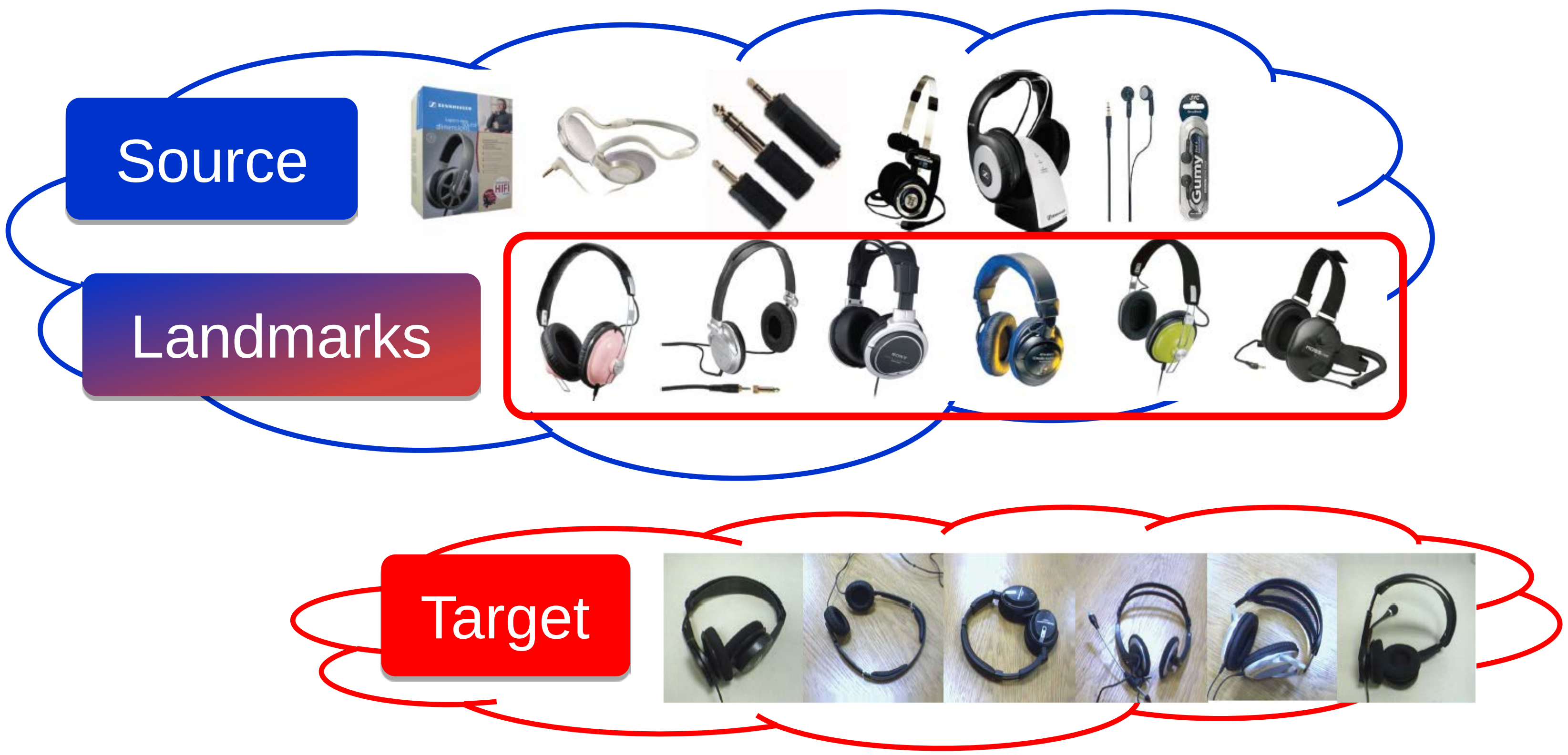
Ease adaptation difficulty

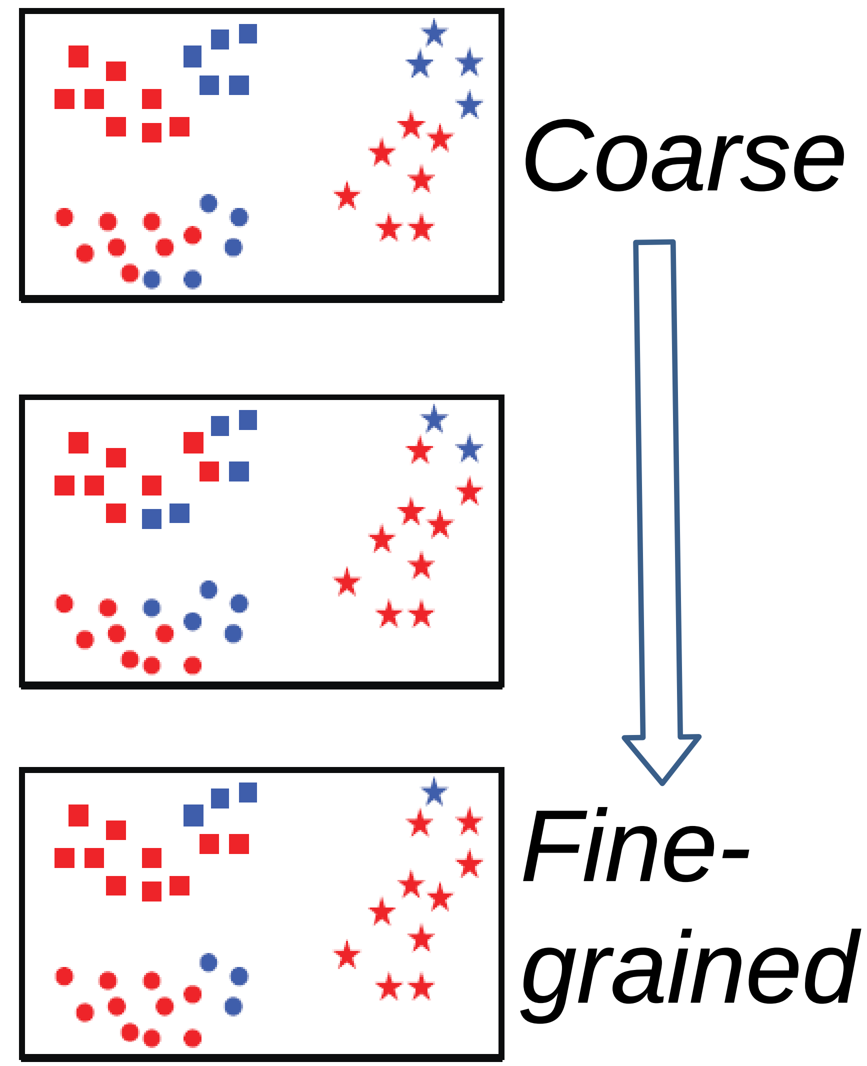
Provide discrimination (biased to target)



[Gong et al., ICML 2013]

Key steps



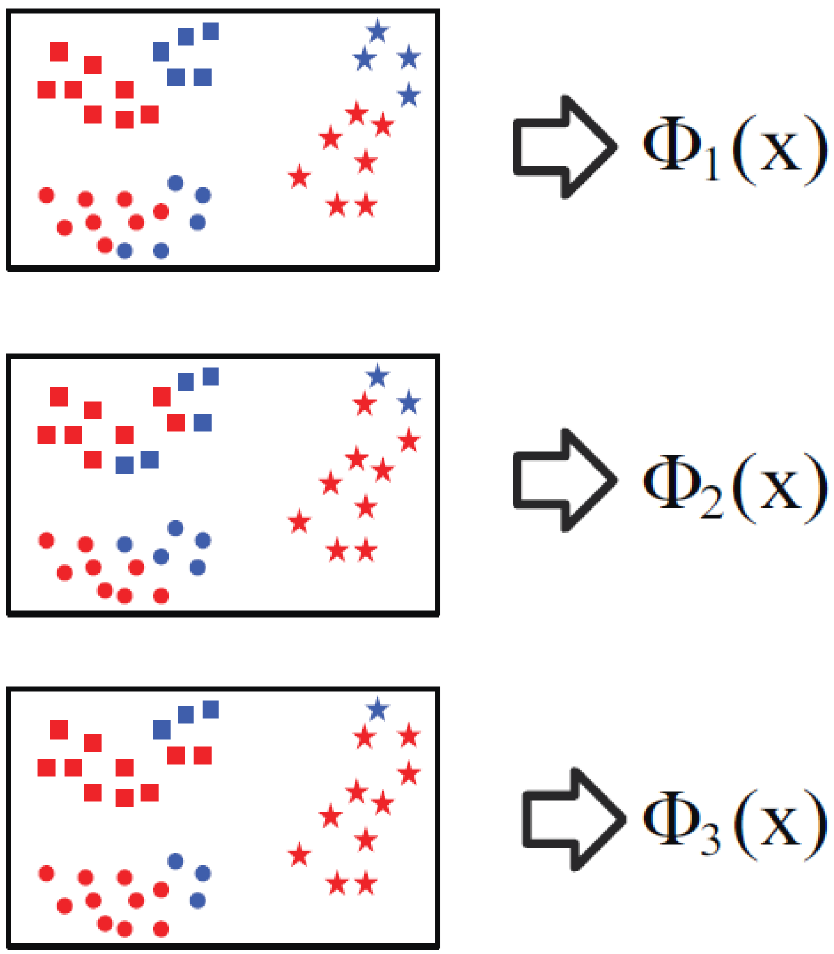


1 Identify landmarks

at multiple scales.

[Gong et al., ICML 2013]

Key steps

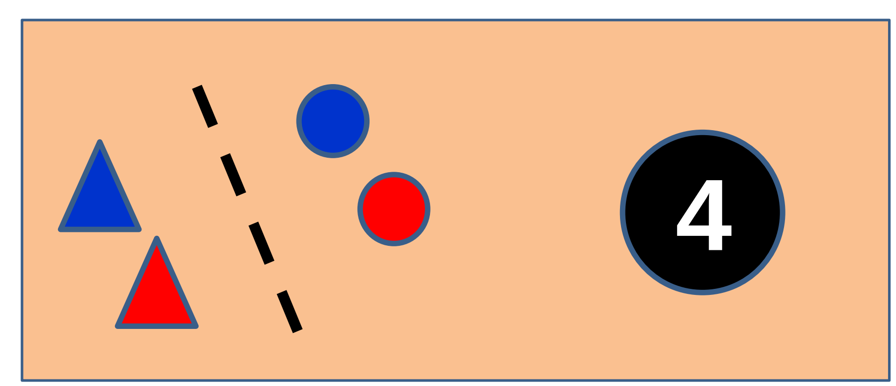


2 Construct auxiliary domain adaptation tasks

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

3

Obtain domain-invariant features



Predict target labels

[Gong et al., ICML 2013]

Identifying landmarks

Objective

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

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

[Gong et al., ICML 2013]

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)

[Gong et al., ICML 2013]

Method for identifying landmarks

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

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

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

[Gong et al., ICML 2013]

Method for identifying landmarks

Convex relaxation

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

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

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

[Gong et al., ICML 2013]

Scale for landmark similarity?

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

Gaussian kernels

How to choose the bandwidth?

Our solution:

Examine distributions at <u>multiple</u> granularities

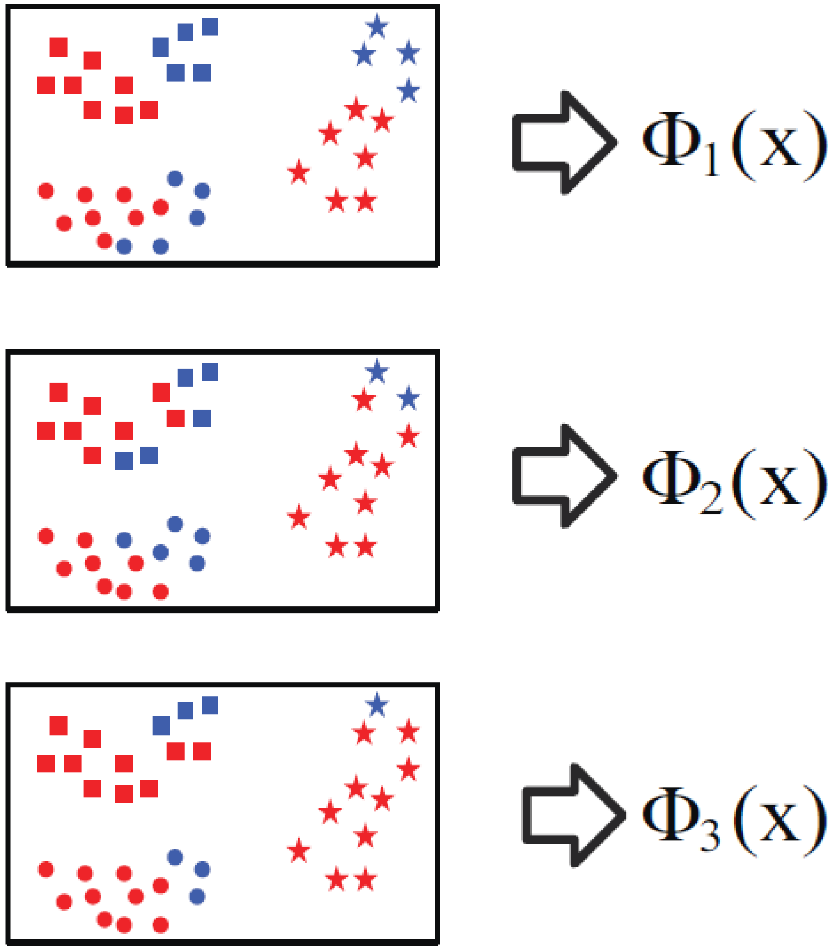
Multiple bandwidths→multiple sets of landmarks

[Gong et al., ICML 2013]

Landmarks at multiple scales

[Gong et al., ICML 2013]

Key steps



2 Construct auxiliary domain adaptation tasks

[Gong et al., ICML 2013]

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)

[Gong et al., ICML 2013]

Constructing easier auxiliary tasks

At each scale σ

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

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

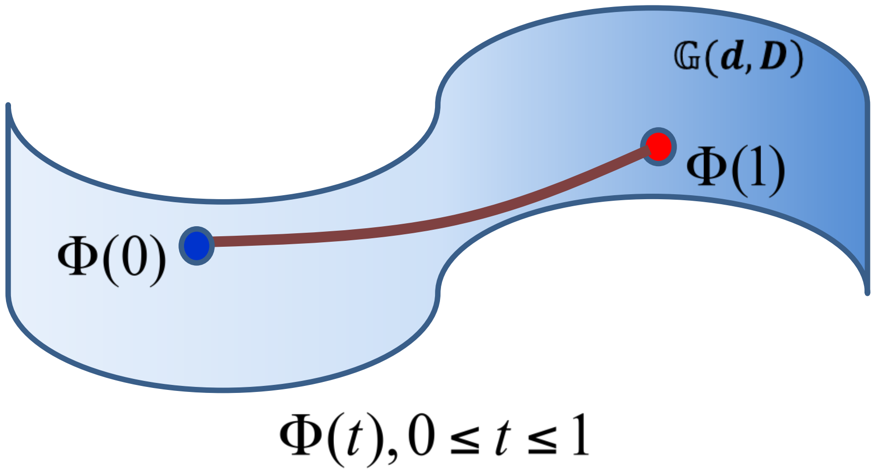
Intuition: distributions are closer (cf. Theorem 1)

[Gong et al., ICML 2013]

Constructing easier auxiliary tasks

Each task provides new basis of features via geodesic flow kernel (GFK):

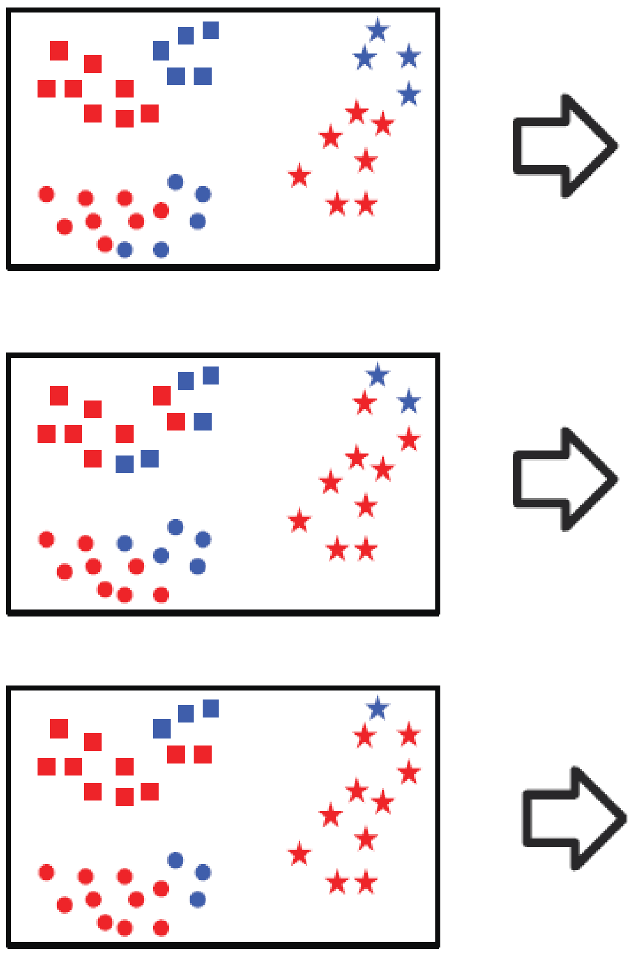
$$K_\sigma(x_i, x_j) = \int_0^1 (\Phi_\sigma(t)'x_i)'(\Phi_\sigma(t)'x_j)\mathrm{d}t = x_i G_\sigma x_j$$



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

[Gong et al., CVPR 2012]

Key steps



$\Phi_1(x)$

$\Phi_2(x)$

$\Phi_3(x)$

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 MKL

Obtain domain-invariant features

Combining features discriminatively

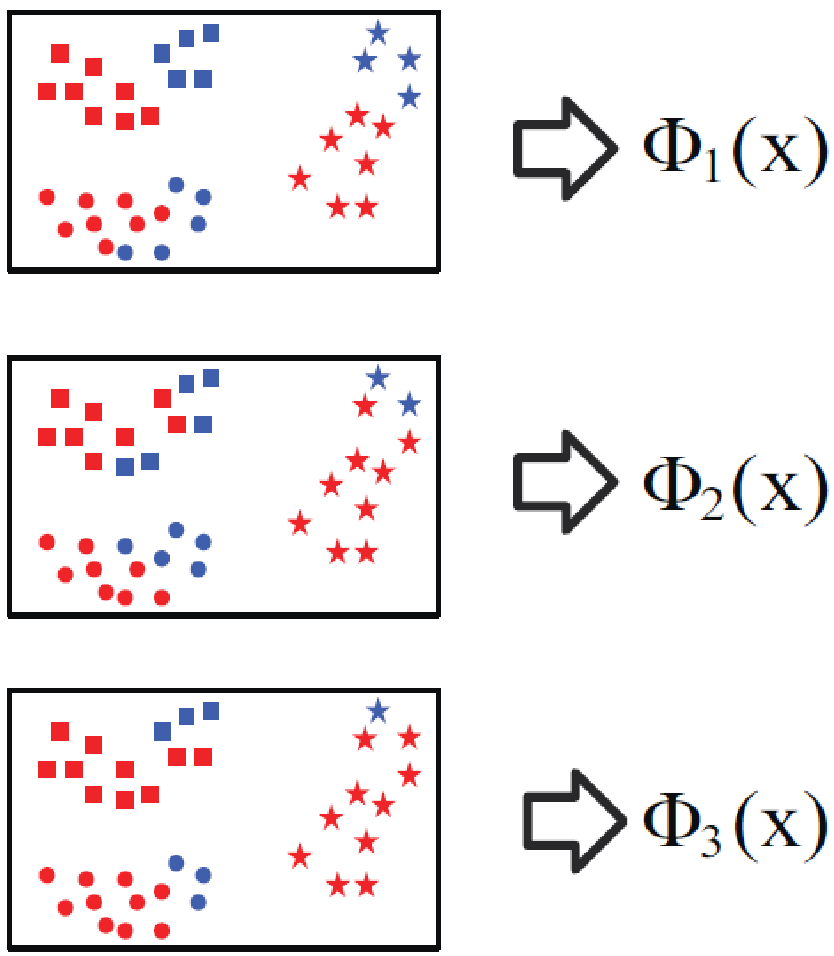
Multiple kernel learning on the labeled landmarks

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

Arriving at domain-invariant feature space

Discriminative loss biased to the target

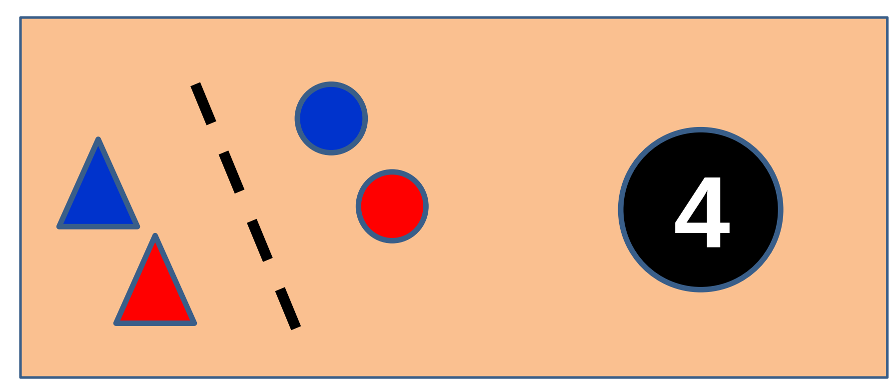
Key steps



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

Experiments

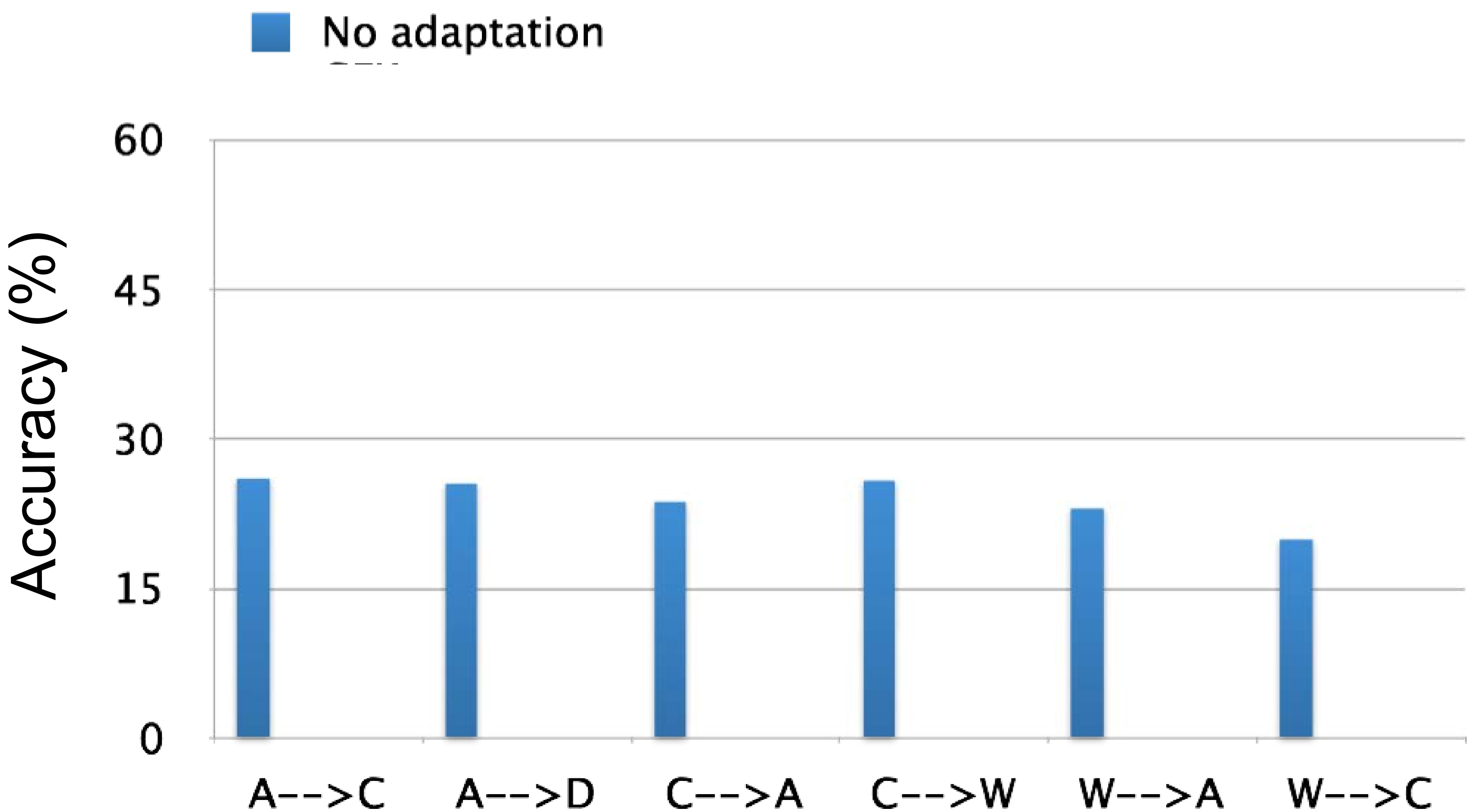
Four vision datasets/domains on visual object recognition

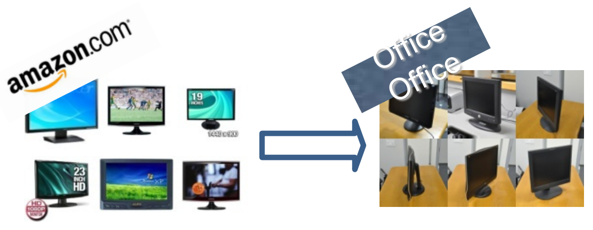
[Griffin et al. ’07, Saenko et al. 10’]

Four types of product reviews on sentiment analysis

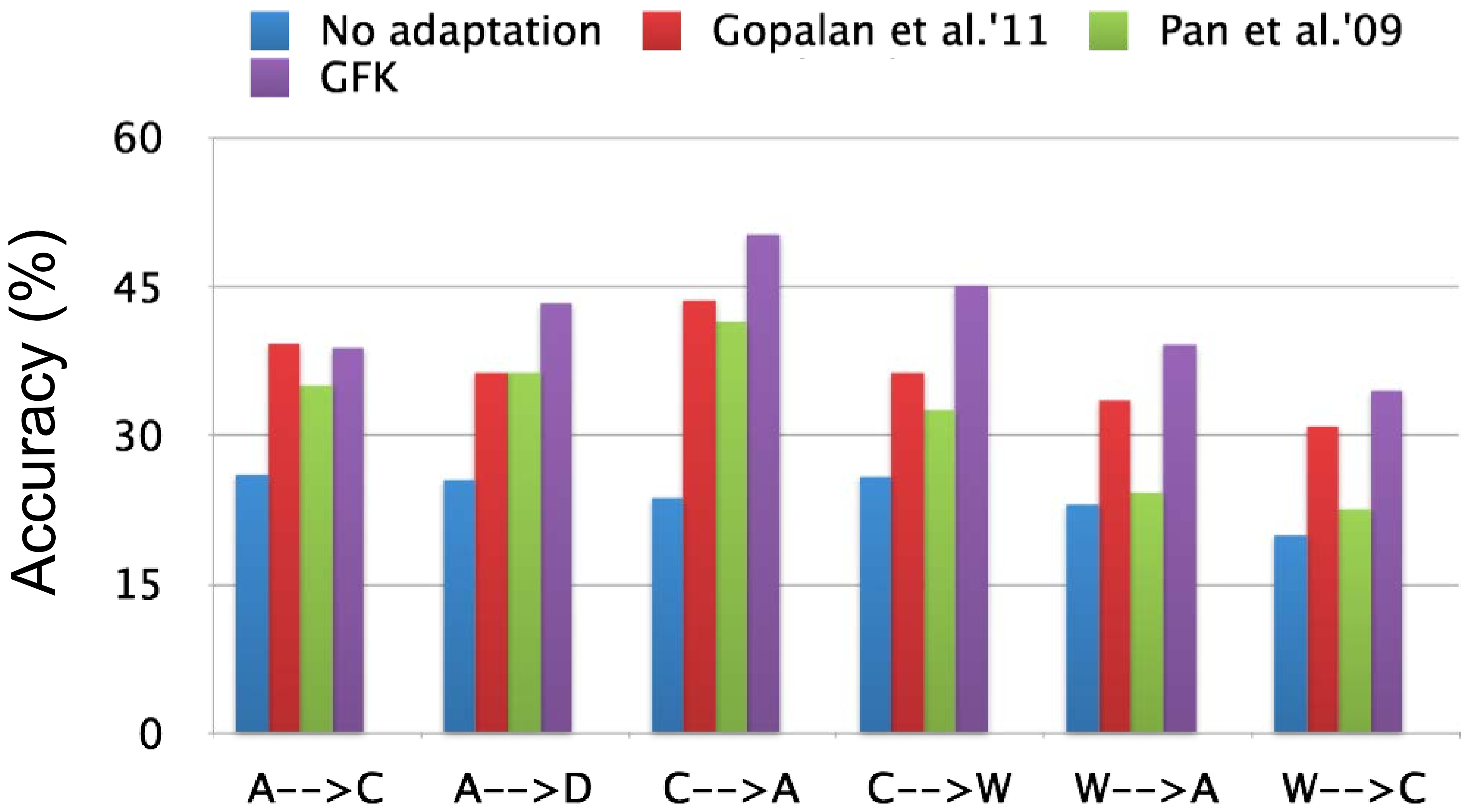
Books, DVD, electronics, kitchen appliances *[Biltzer et al. ’07]*

Cross-dataset object recognition

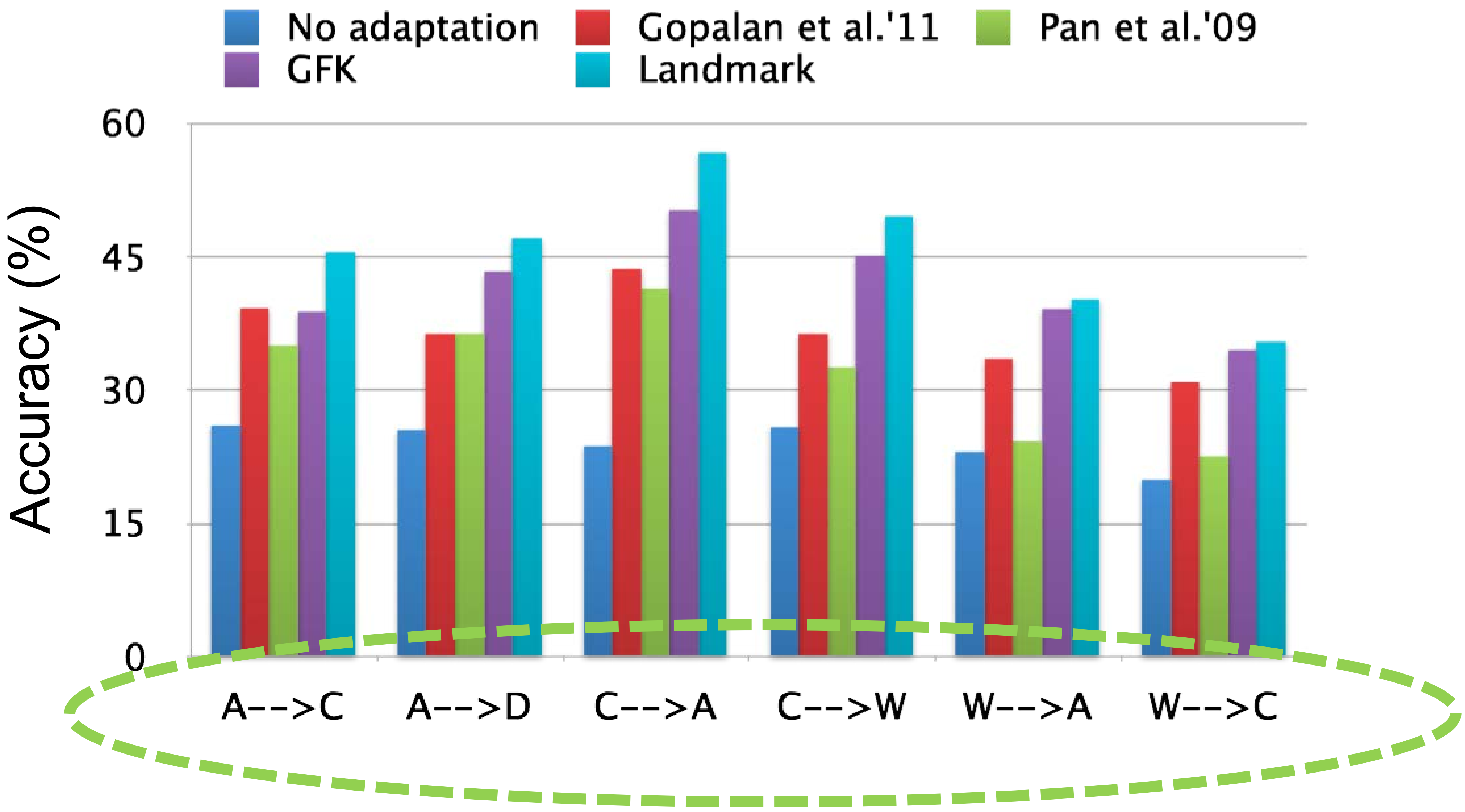




Cross-dataset object recognition



Cross-dataset object recognition



Datasets as domains?

Domain 1

Domain 2

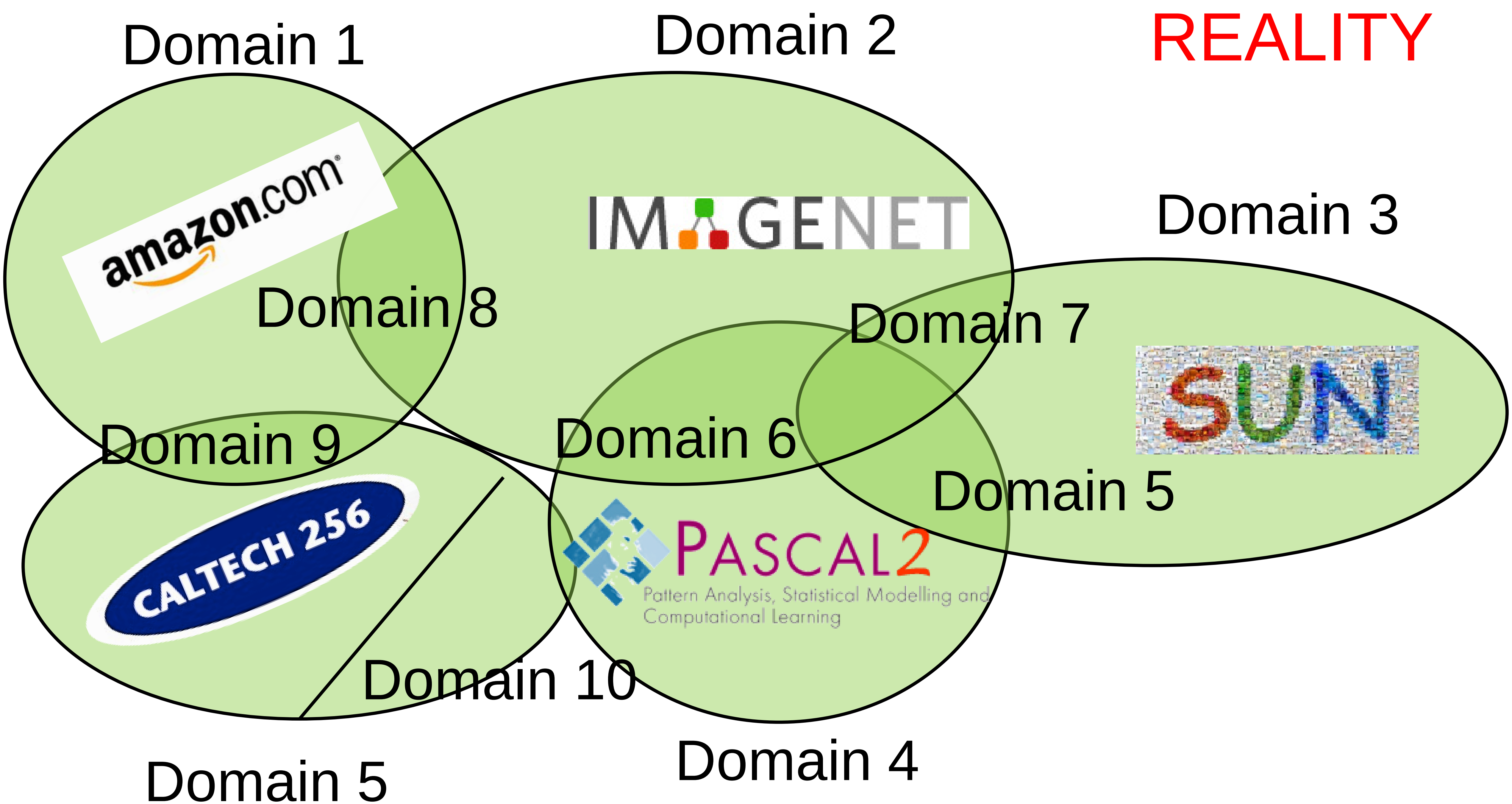
ASSUMED

Domain 3

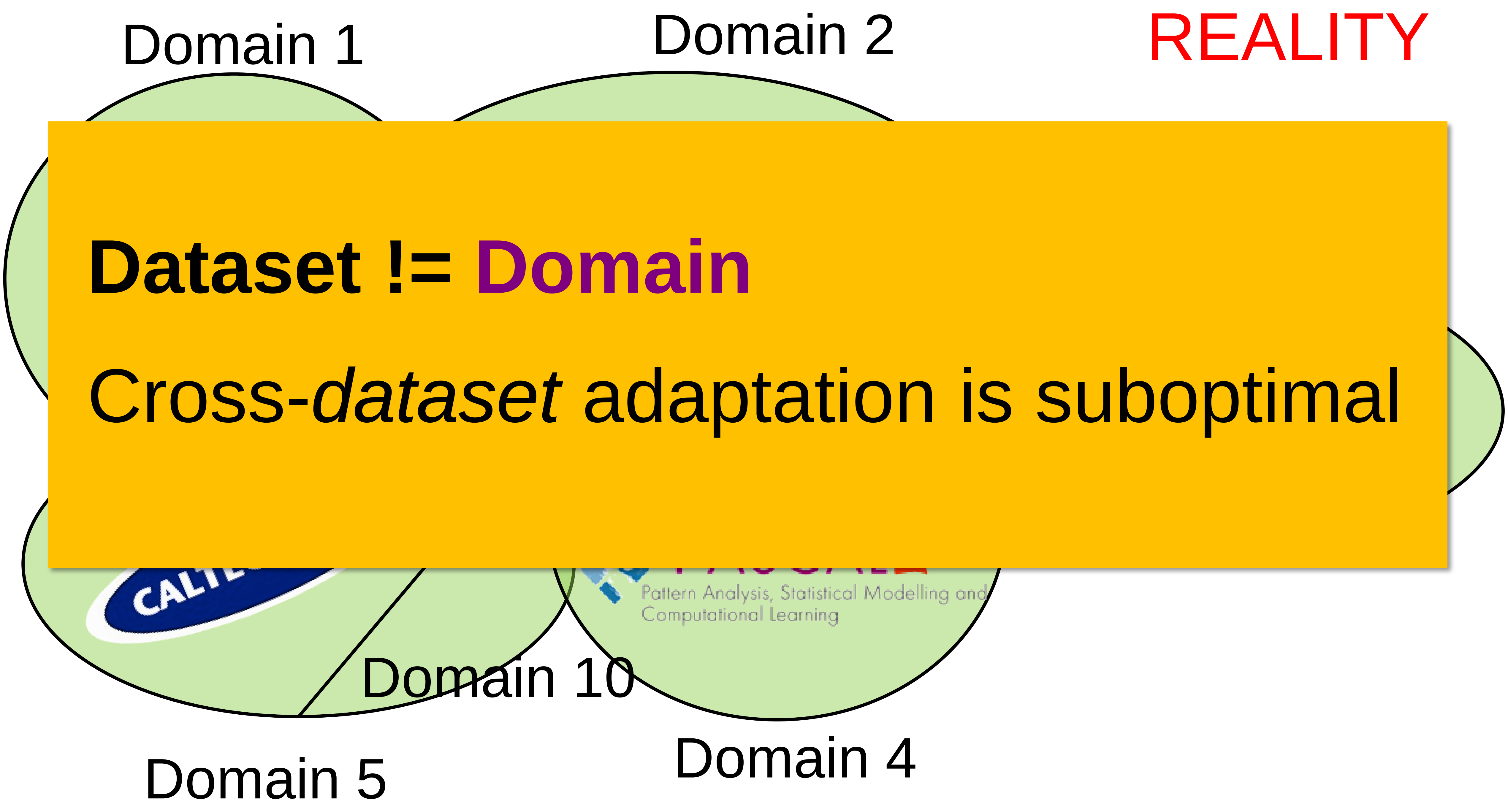
Domain 5

Domain 4

Datasets as domains?



Datasets as domains?



Dataset != Domain

Cross-*dataset* adaptation is suboptimal

How to define a domain?

NLP: *Language*-specific domains ✓

Speech: *Speaker*-specific domains ✓

Vision: ??

pose-specific? *illumination*-specific?
occlusion? *image resolution*? *background*?

Challenges:

Many continuous factors ***vs.*** few discrete
Factors **overlap** and interact

Discovering latent visual domains

We propose to discover domains – “reshaping” them to cross dataset boundaries

Maximum distinctiveness

$$\max_{\{z_{mk}\}} \sum_{k \neq k'} d(P(k), P(k')) \longrightarrow \text{MMD}$$

where $z_{mk} = \begin{cases} 1 & \text{if } \mathbf{x}_m \text{ belongs to domain } k \\ 0 & \text{else} \end{cases}$

Maximum learnability

Determine *K* with domain-wise cross-validation

[Gong et al., NIPS 2013]

Results: discovering domains

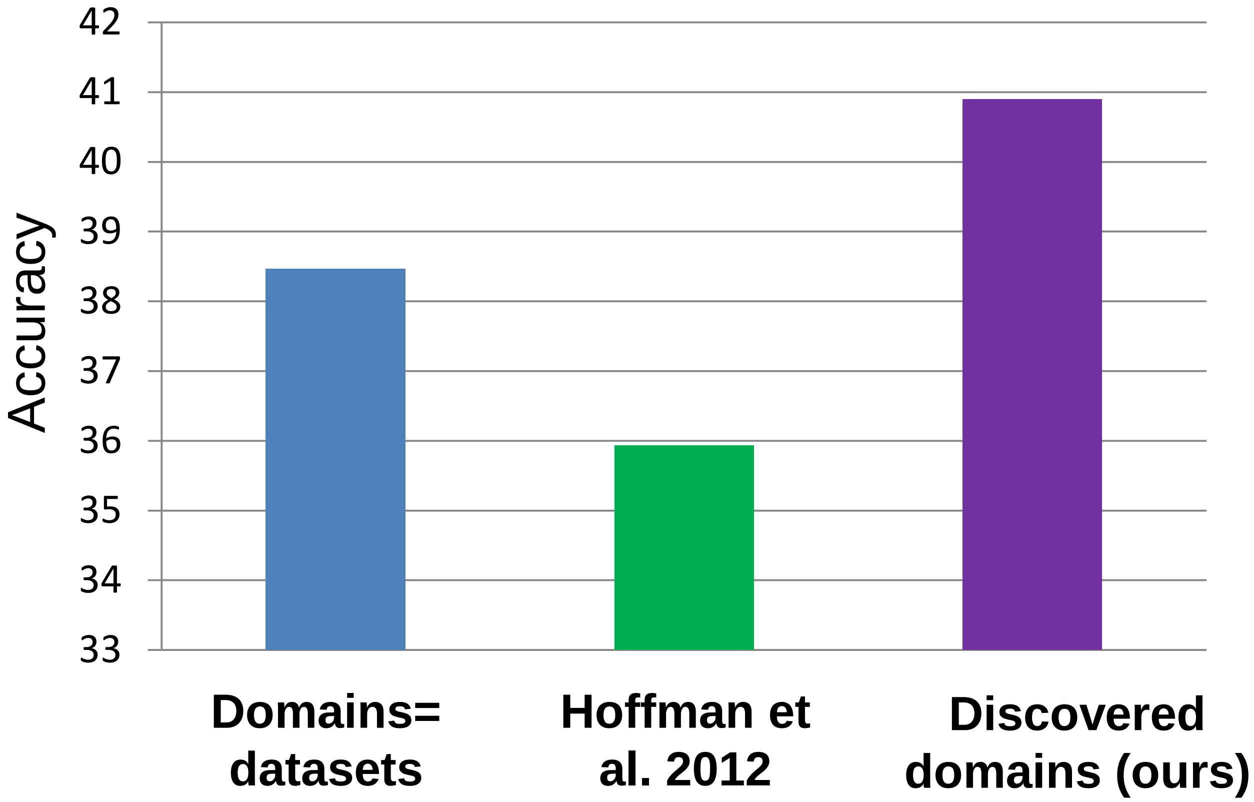
Discovered domain I

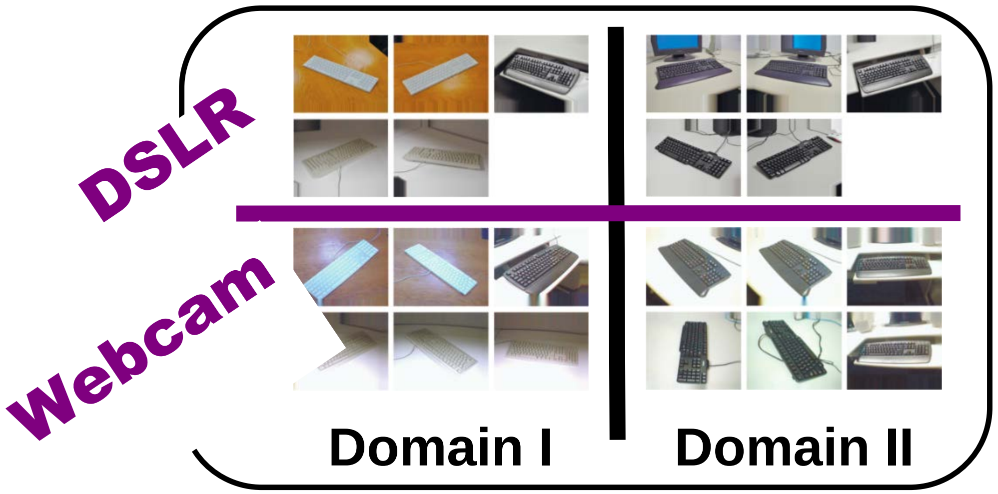
Discovered domain II

[Gong et al., NIPS 2013]

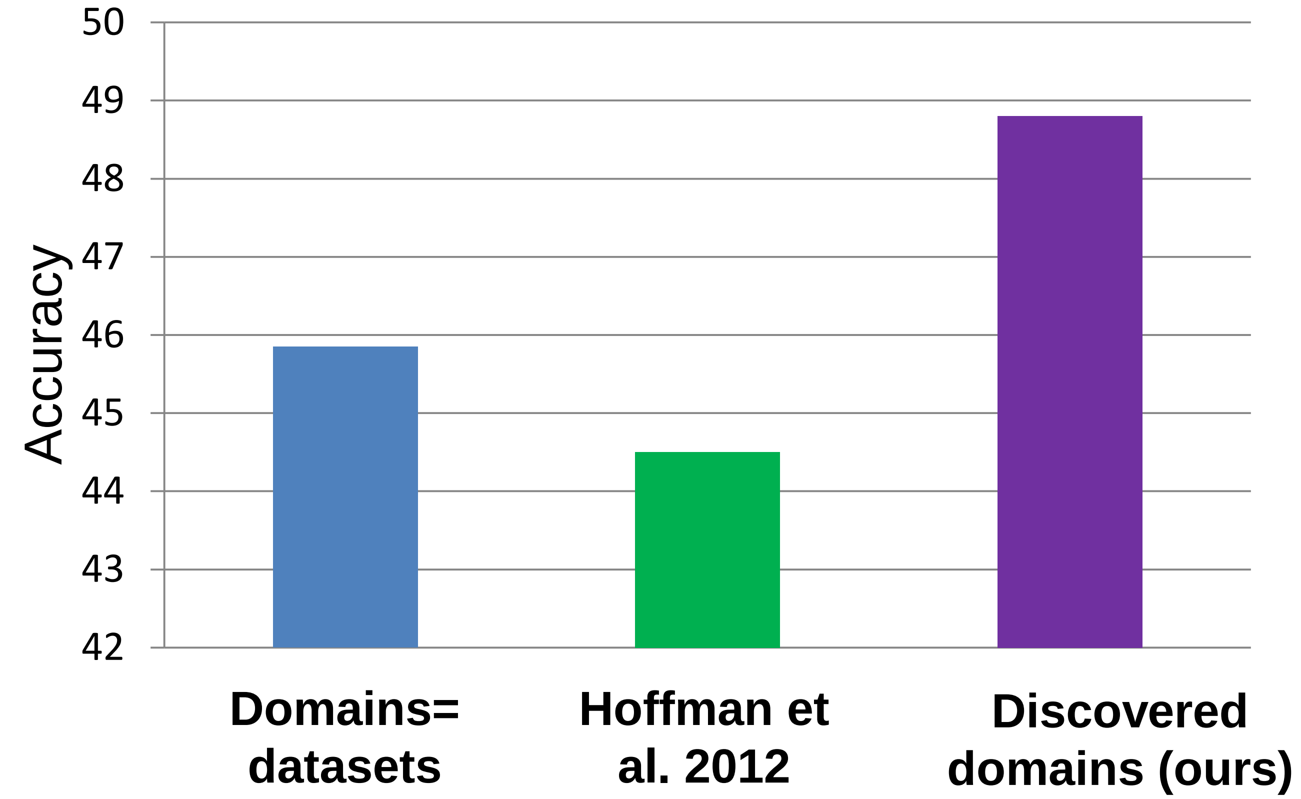
Results: discovering domains

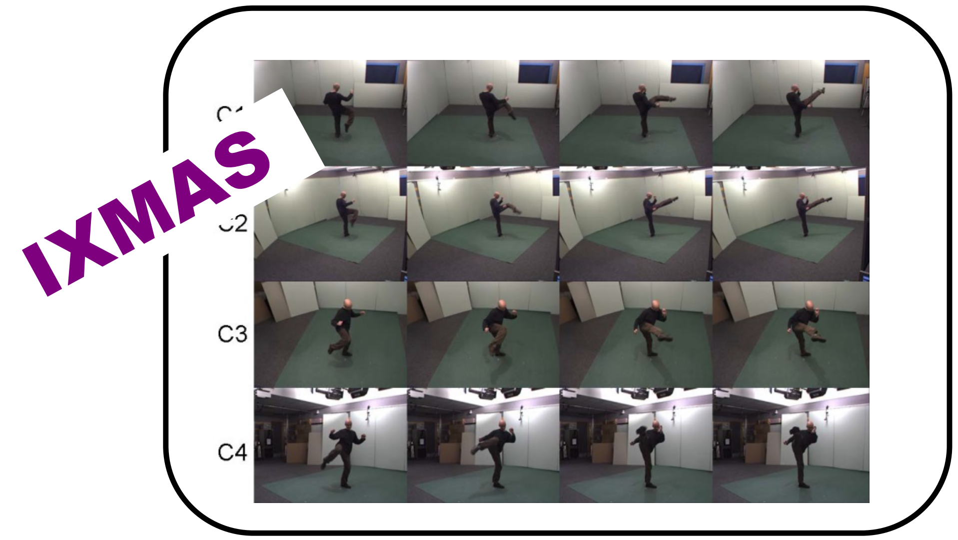
Cross-dataset object recognition





Cross-viewpoint action recognition





Summary so far

landmarks

labeled source instances
distributed similarly to the target

auxiliary tasks provably easier to solve

discriminative loss despite unlabeled target

reshaping datasets to latent domains

discover cross-dataset domains

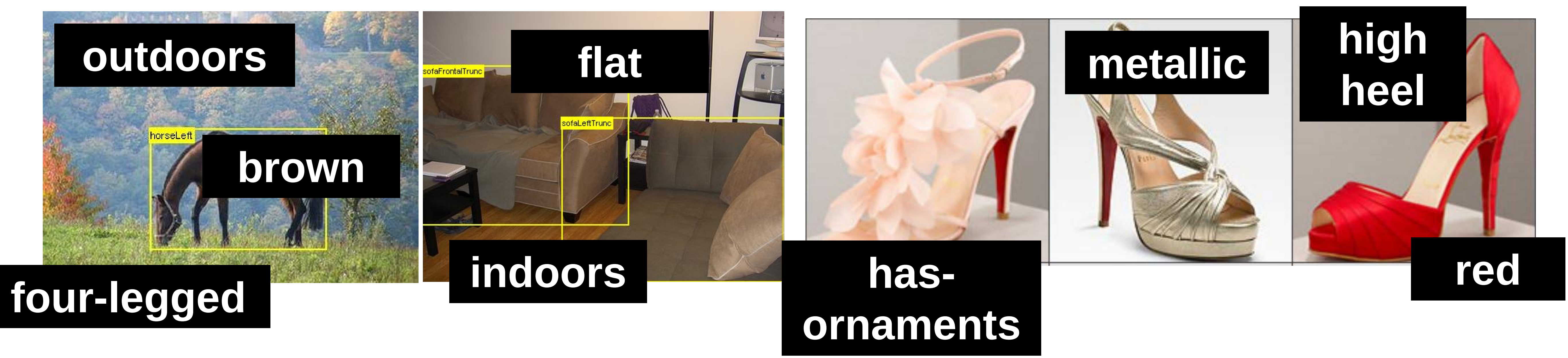
maximally distinct & learnable

Typical assumptions

1. Test set will look like the training set.
2. Human labelers “see” the same thing.

Visual attributes

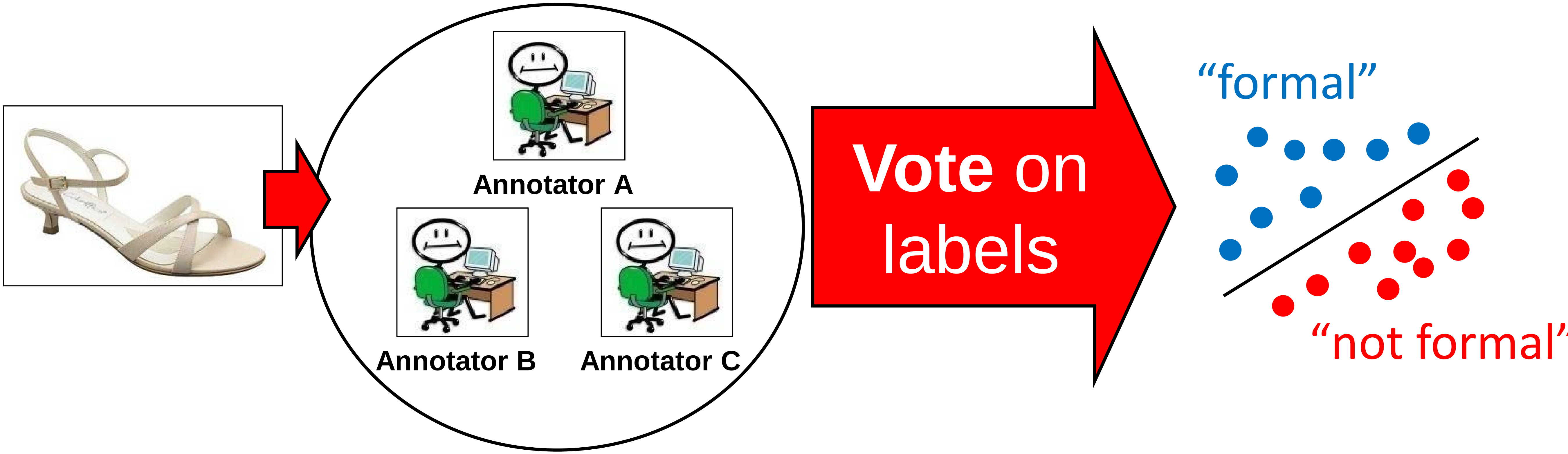
- High-level semantic properties shared by objects
- Human-understandable and machine-detectable



[Oliva et al. 2001, Ferrari & Zisserman 2007, Kumar et al. 2008, Farhadi et al. 2009, Lampert et al. 2009, Endres et al. 2010, Wang & Mori 2010, Berg et al. 2010, Branson et al. 2010, Parikh & Grauman 2011, …]

Standard approach

Learn one monolithic model per attribute



Problem

There may be valid perceptual differences within an attribute.

Formal?

User labels:
50% “yes”
50% “no”

Binary attribute

More ornamented?

or

User labels:
50% “first”
20% “second”
30% “equally”

Relative attribute

Imprecision of attributes

Fine-grained meaning

Overweight?

or just

Chubby?

Imprecision of attributes

Context

Is  *formal?*

= *formal* wear for a conference? OR

= *formal* wear for a wedding?

Imprecision of attributes

Cultural

Is  *blue* or *green*?

English: “blue”

Russian: “neither”
(“голубой” vs. “синий”)

Japanese: “both”
(“青” = blue *and* green)

But do we need to be that precise?

Yes. Applications like image search require that user's perception matches system's predictions.

"white high heels"

"less formal than these"

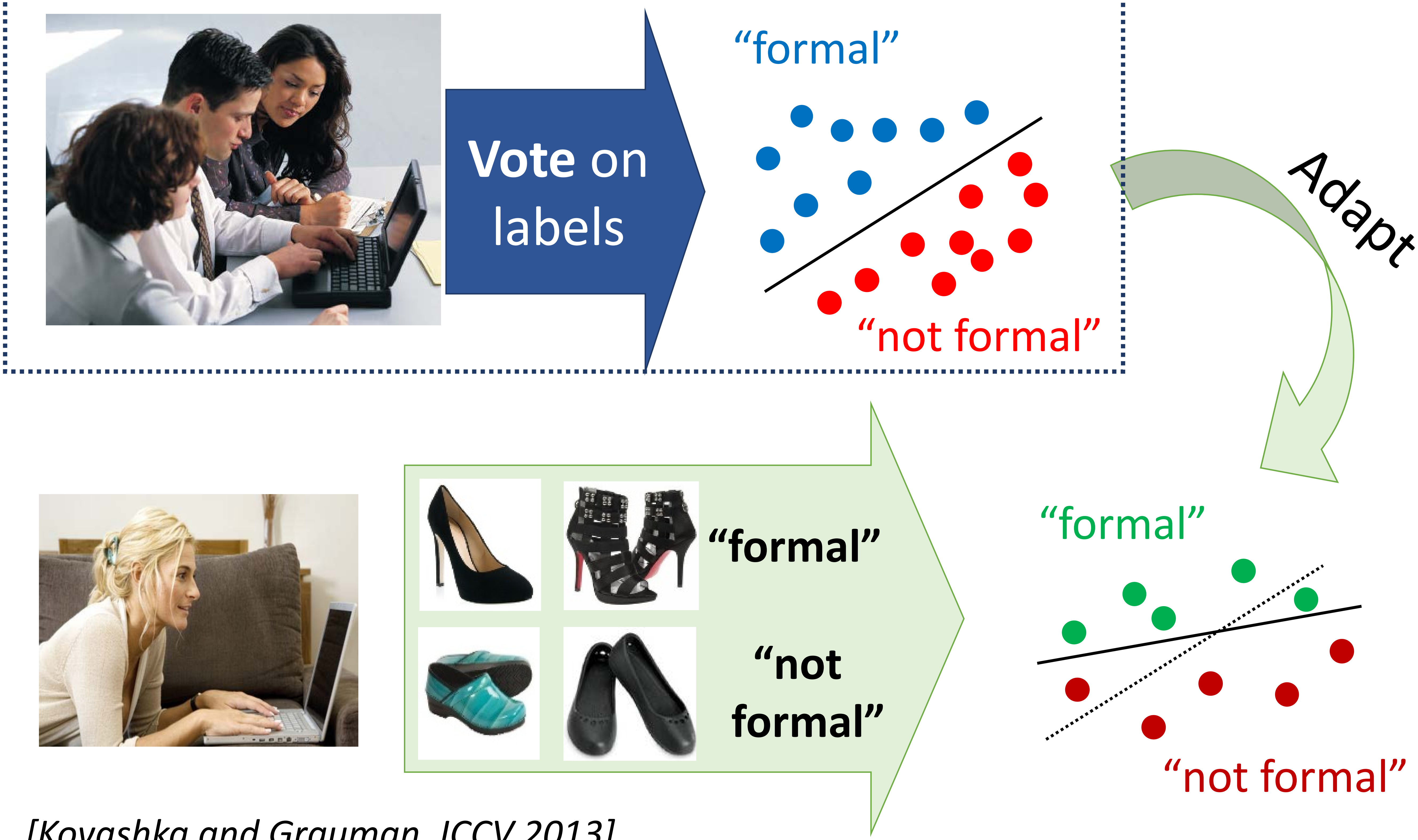
[WhittleSearch, Kovashka et al. CVPR 2012]

Our idea

- Treat learning perceived attributes as an adaptation problem.
- Adapt generic attribute model with minimal user-specific labeled examples.
- Obtain implicit user-specific labels from user's search history

[Kovashka and Grauman, ICCV 2013]

Our idea



[Kovashka and Grauman, ICCV 2013]

Learning adapted attributes

- Adapting binary attribute classifiers:

Given user-labeled data $D_b = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$

and generic model $\boldsymbol{w}'_b$,

$$\min_{\boldsymbol{w}_b} \frac{1}{2} \boxed{\|\boldsymbol{w}_b - \boldsymbol{w}'_b\|^2} + C \sum_{i=1}^{N} \xi_i,$$

$$\text{subject to} \quad y_i \boldsymbol{x}_i^T \boldsymbol{w}_b \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

J. Yang et al. ICDM 2007.

Learning adapted attributes

- Adapting relative attribute rankers:

Given user-labeled data $D_r = \{(x_{i_1} \succ x_{j_1})\}_{i=1}^{N}$

and generic model $\boldsymbol{w}'_r$,

$$\min_{\boldsymbol{w}_r} \frac{1-\delta}{2}\|\boldsymbol{w}_r\|^2 + \frac{\delta}{2}\|\boldsymbol{w}_r - \boldsymbol{w}'_r\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad \boldsymbol{w}_r^T\boldsymbol{x}_{i_1} - \boldsymbol{w}_r^T\boldsymbol{x}_{i_2} \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i,$$

B. Geng, et al. TKDE 2010.

Collecting user-specific labels

- **Explicitly** from actively requested labels

  Seek labels on uncertain and diverse images

- **Implicitly** from search history

  o Transitivity

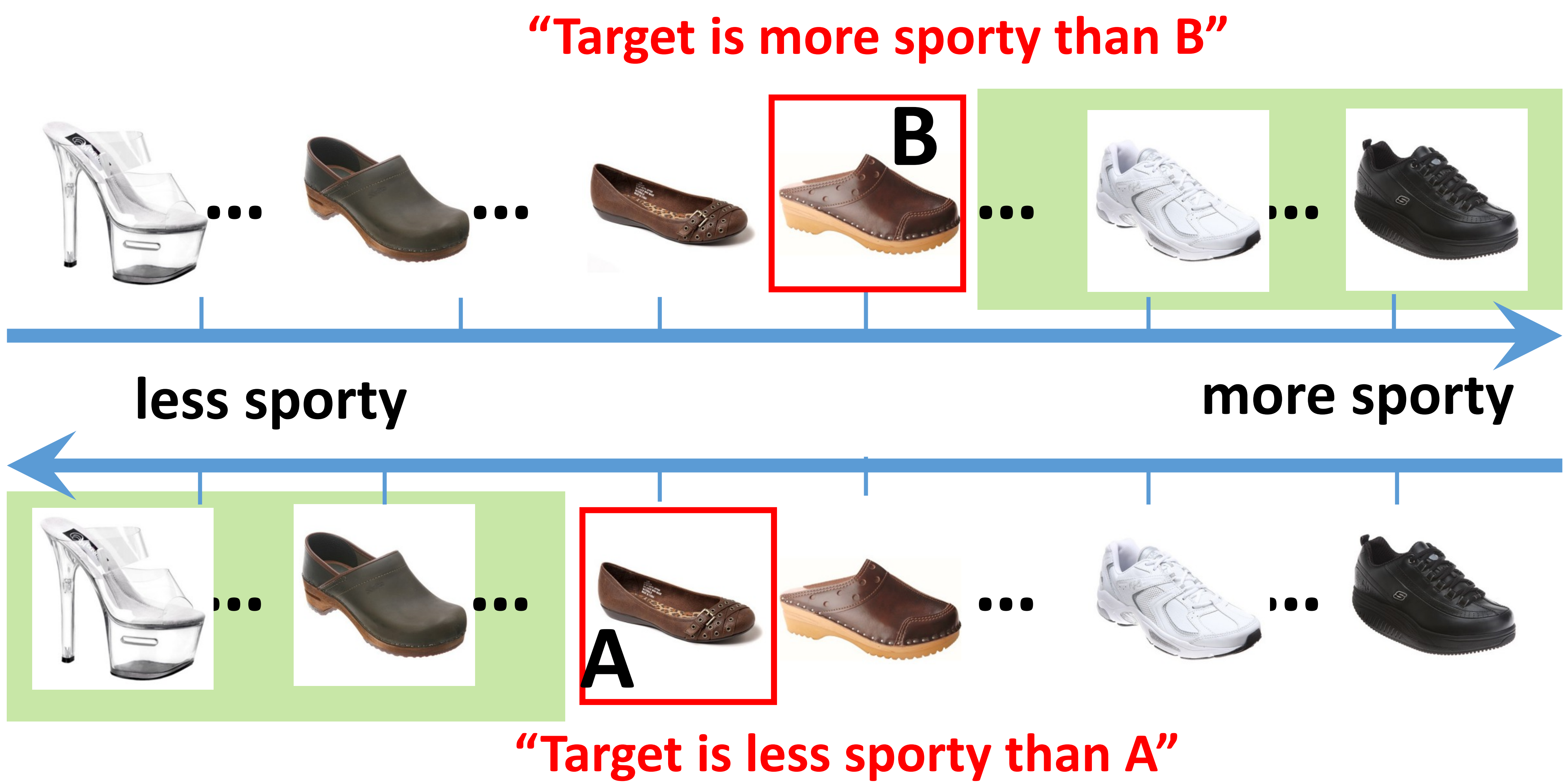
  “My target is…

  less formal than 

  more formal than  “

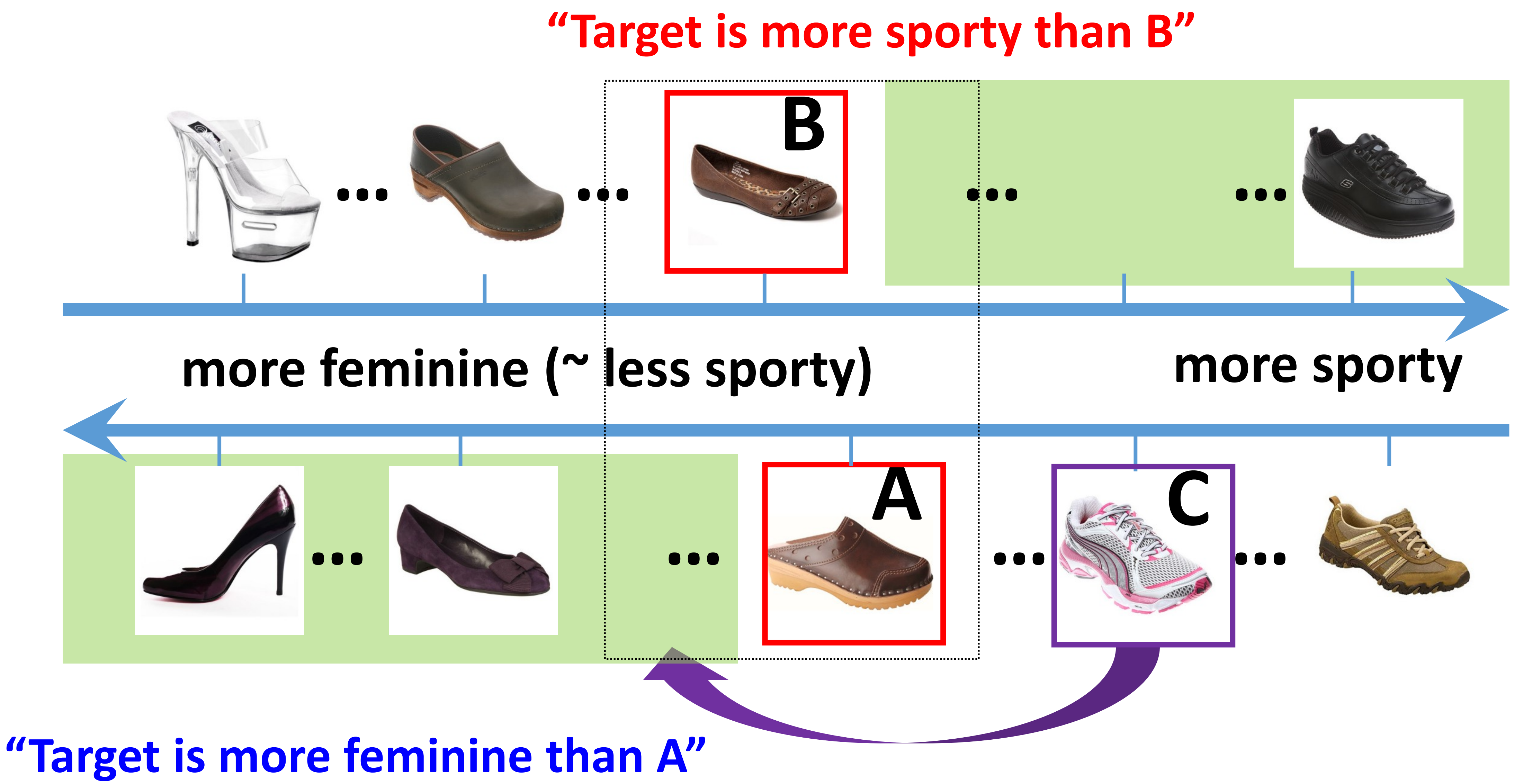
  o Contradictions

Inferring implicit labels



User's feedback history can reveal mismatch in perceived and predicted attributes

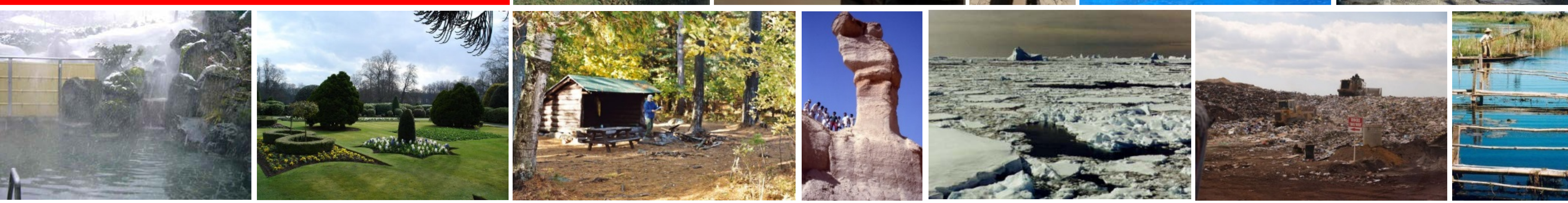
Inferring implicit labels



User's feedback history can reveal mismatch in perceived and predicted attributes

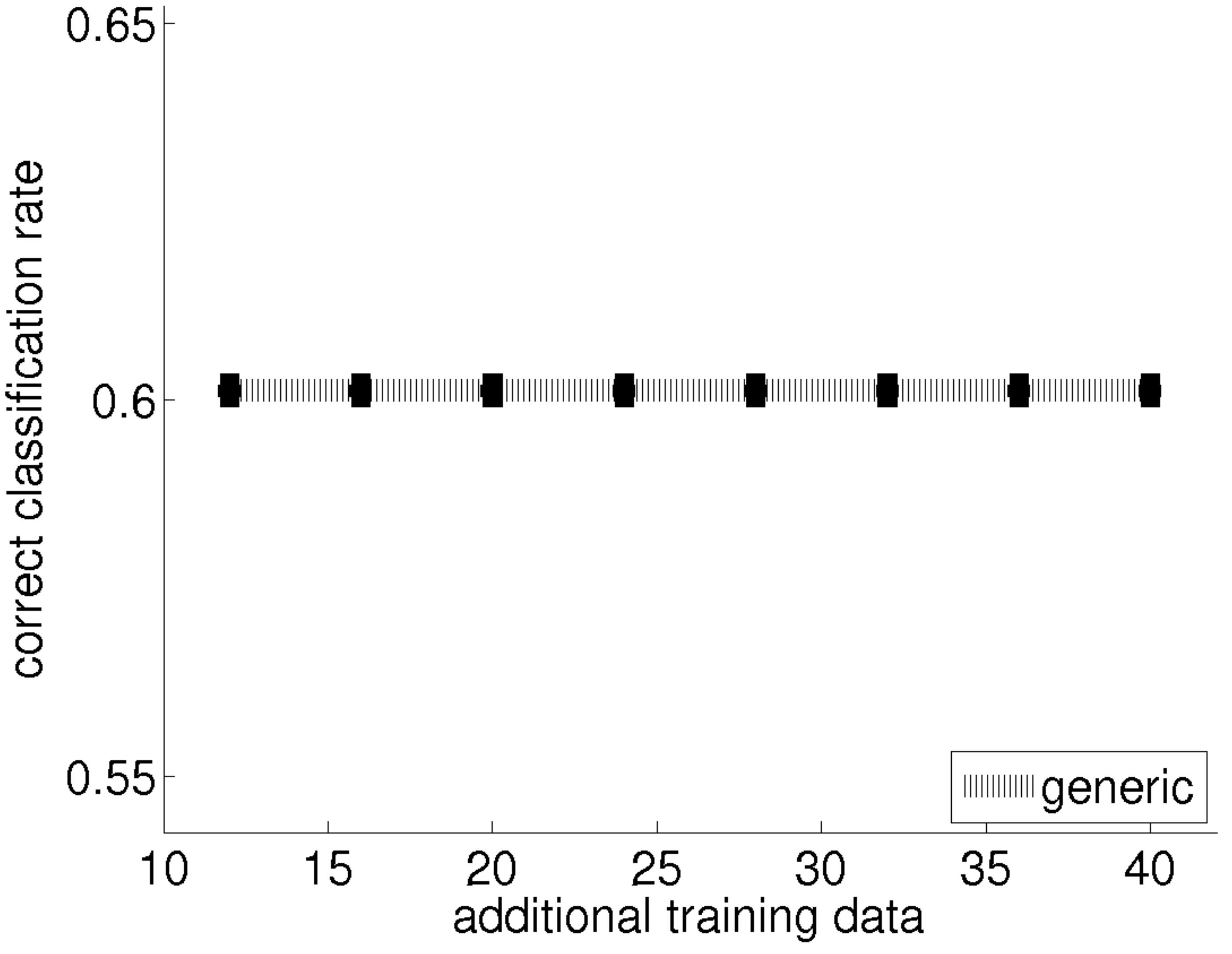
Datasets

SUN Attributes:
14,340 scene images
12 attributes:
“sailing”, “hiking”,
“vacationing”, “open area”,
“vegetation”, etc.



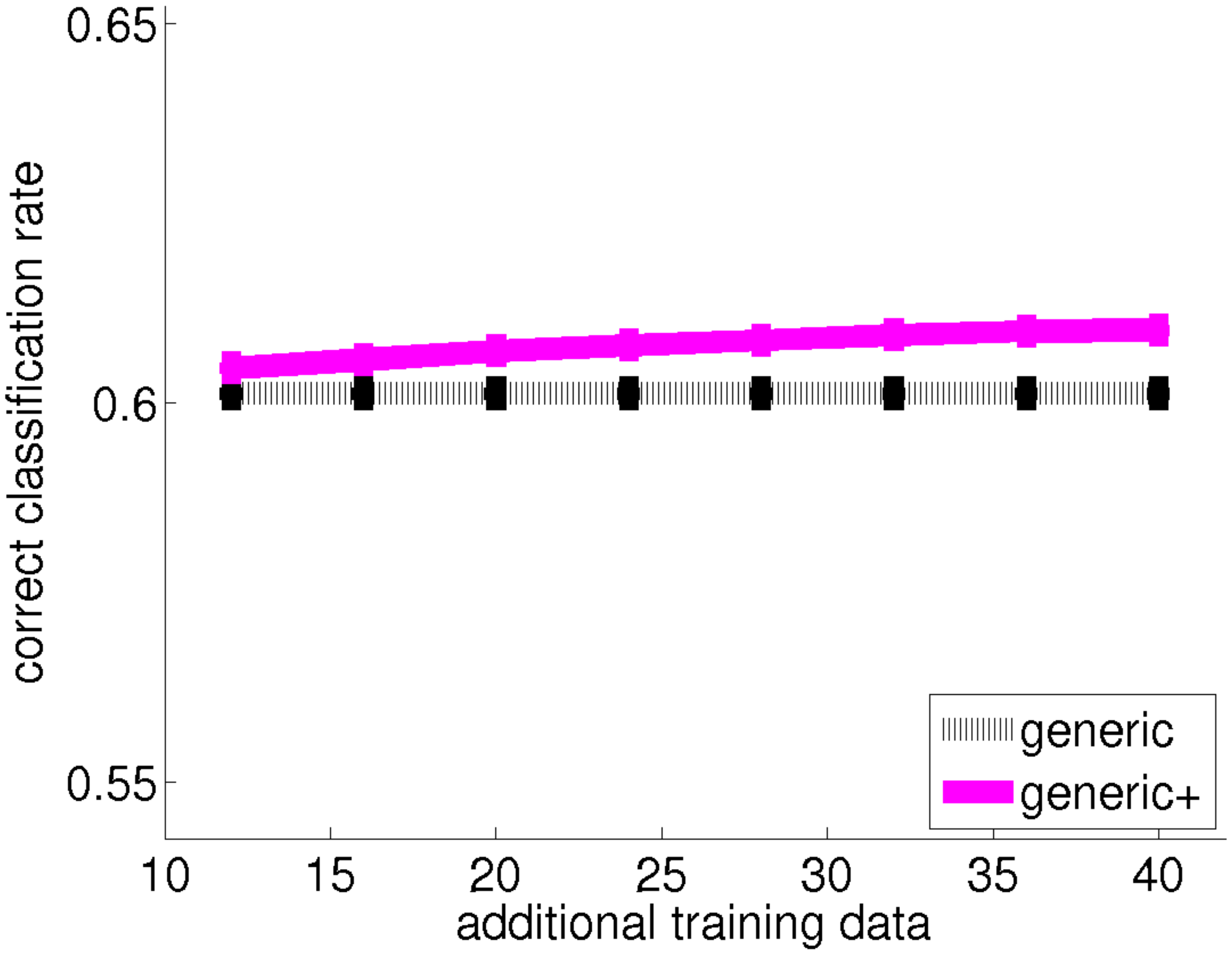
Shoes:
14,658 shoe images;
10 attributes:
“pointy”, “bright”, “high-heeled”, “feminine” etc.

Adapted attribute accuracy



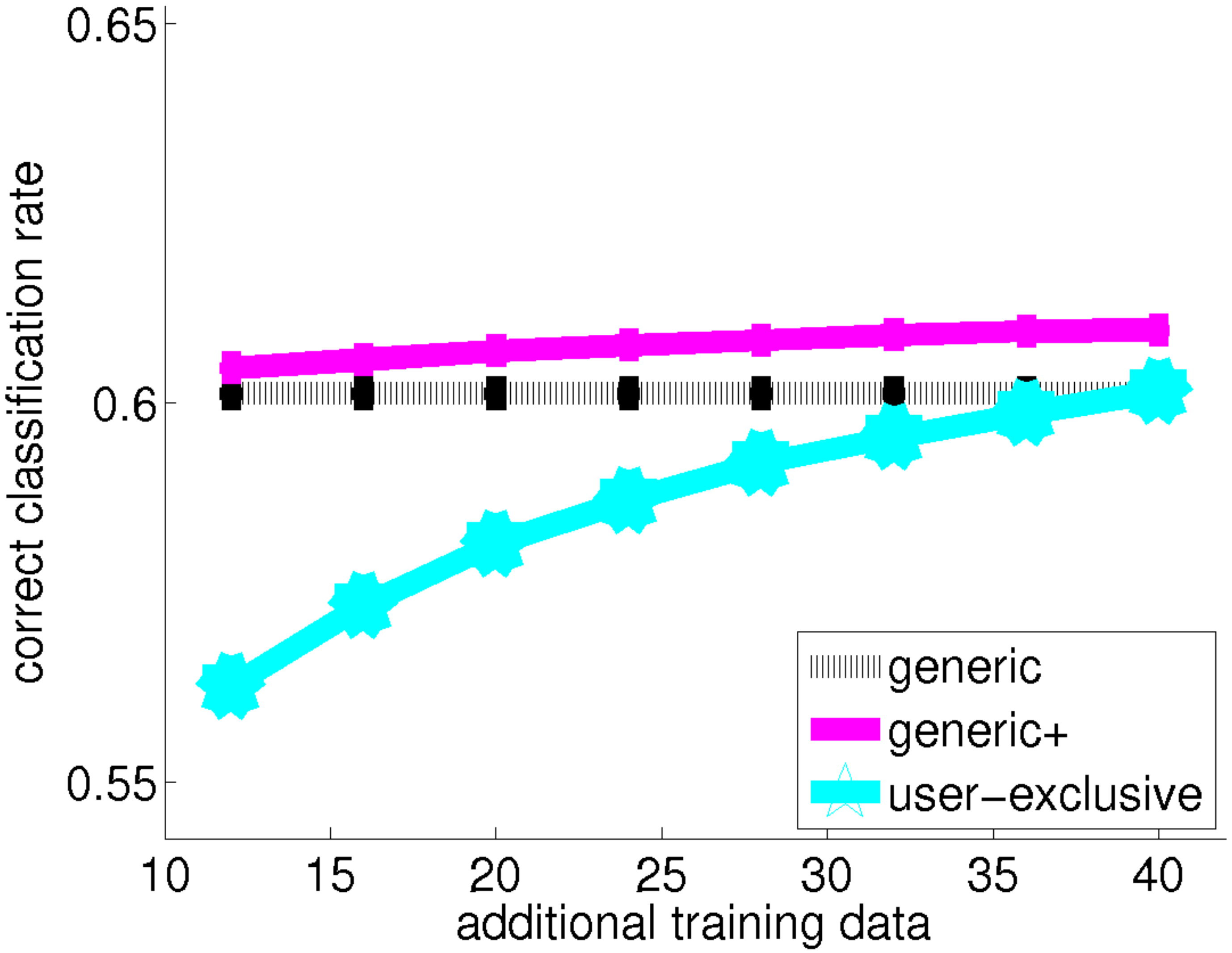
- 3 datasets
- 22 attributes
- 75 total users

Adapted attribute accuracy



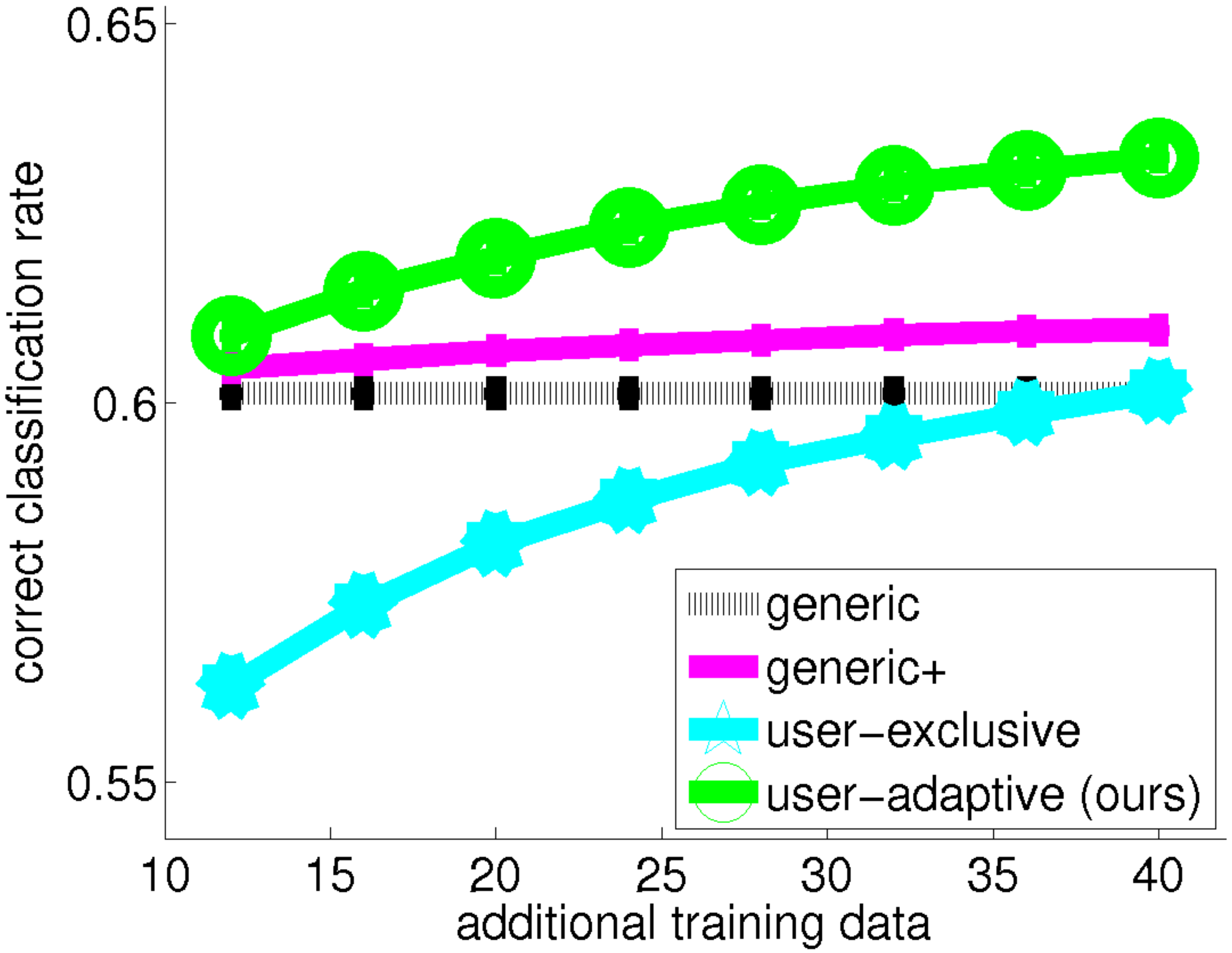
- 3 datasets
- 22 attributes
- 75 total users

Adapted attribute accuracy



- 3 datasets
- 22 attributes
- 75 total users

Adapted attribute accuracy



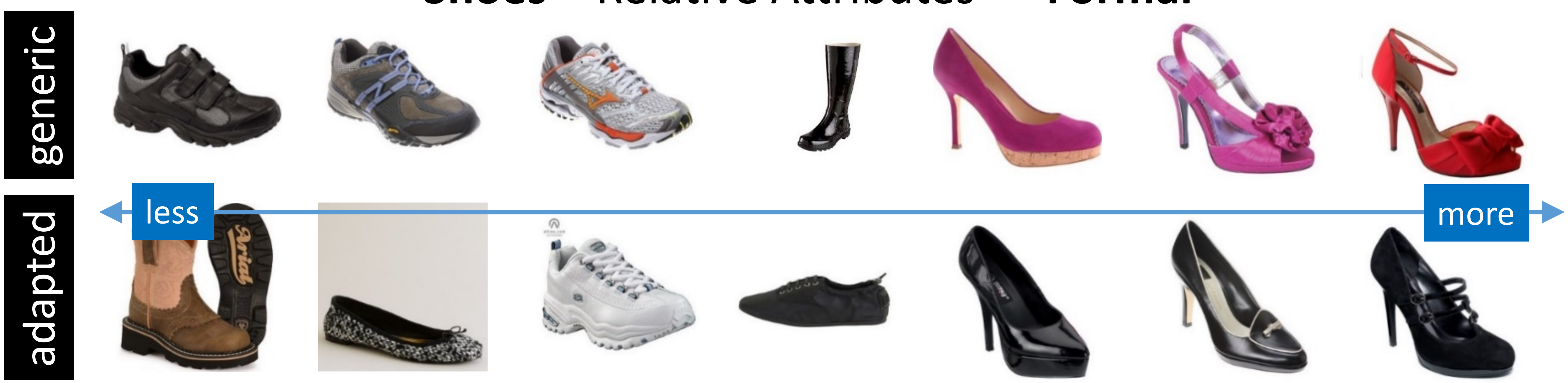
Adaptation approach most accurately captures perceived attributes

[Kovashka and Grauman, ICCV 2013]

Which images most influence adaptation?

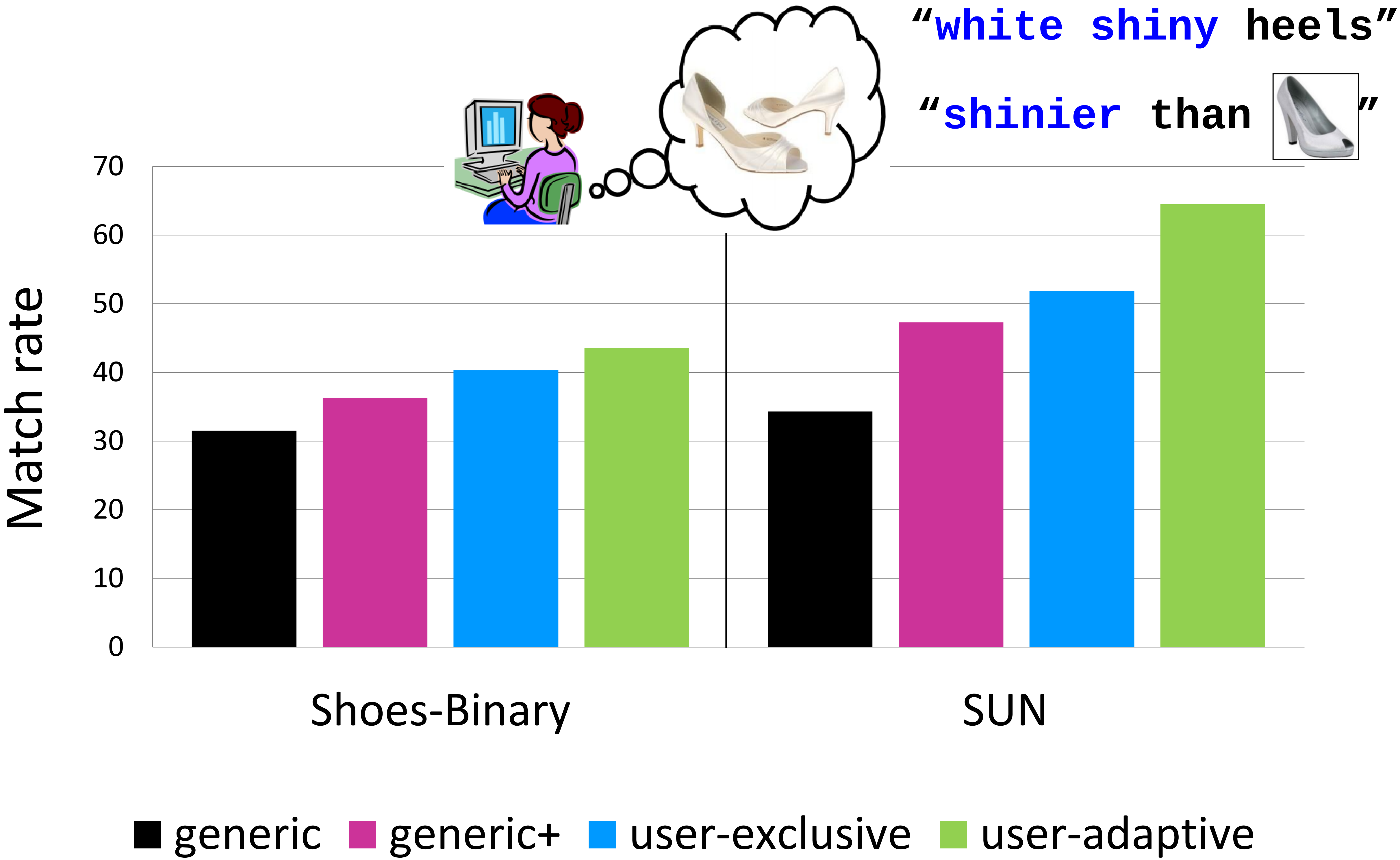
Visualizing adapted attributes

Shoes – Relative Attributes – **"Formal"**

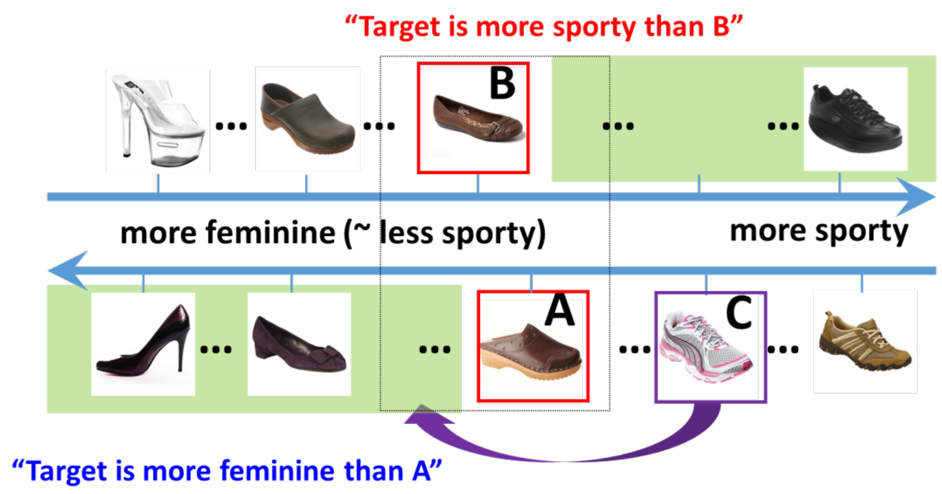


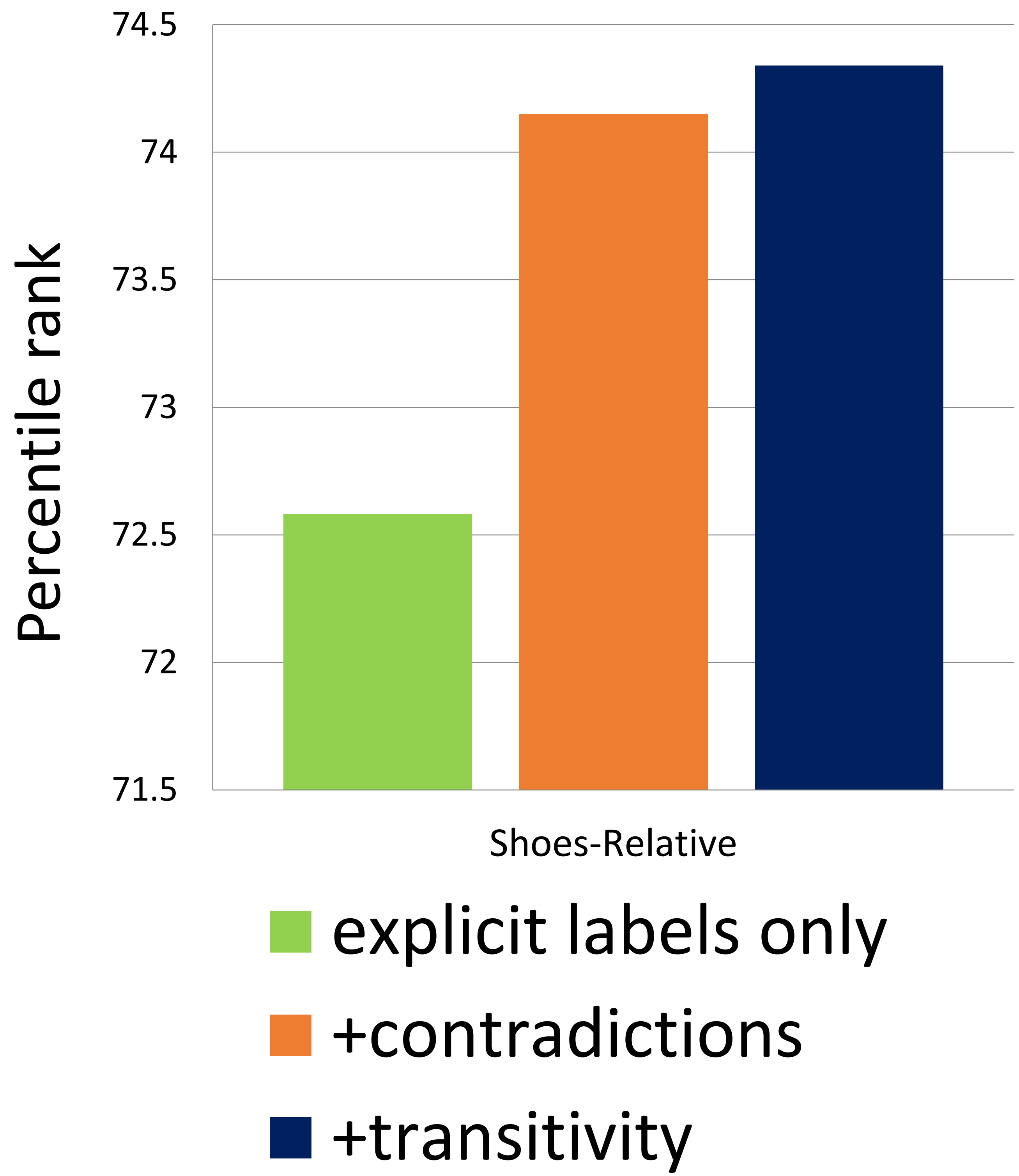
SUN – Binary Attributes – **"Vacationing"**

Personalizing image search with adapted attributes



Impact of implicit labels





Summary

- Practical concerns if learning visual categories:

  Test images can look different from training images!

  People do not perceive image labels universally!

- Domain adaptation methods help address them

  Landmark-based unsupervised adaptation

  Reshaping datasets into latent domains

  Adapt generic models to account for user-specific perception of attributes

References

- Attribute Adaptation for Personalized Image Search. A. Kovashka and K. Grauman. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, December 2013.
- Reshaping Visual Datasets for Domain Adaptation. B. Gong, K. Grauman, and F. Sha. In Proceedings of Advances in Neural Information Processing Systems (NIPS), Tahoe, Nevada, December 2013.
- Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation. B. Gong, K. Grauman, and F. Sha. In International Conference on Machine Learning (ICML), Atlanta, GA, June 2013.
- Geodesic Flow Kernel for Unsupervised Domain Adaptation. B. Gong, Y. Shi, F. Sha, and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 2012.