Recognition by Association

ask not "What is this?"
but "What is it ***like***?"

Tomasz Malisiewicz
joint work with Alyosha Efros

October 27, 2008
Learning Lunch

Carnegie Mellon
THE ROBOTICS INSTITUTE

Goal and Approach

- **Goal**: Recognize many different types of objects inside an image

- **Observation**: Recognition becomes easier once we have the correct segmentation

- **Approach**: Use a segment-centric object representation and an exemplar-based non-parametric recognition model

Tomasz Malisiewicz, Alexei A. Efros. Recognition by Association via Learning Per-exemplar Distances. In CVPR, June 2008.

Understanding an Image

slide by Fei Fei, Fergus & Torralba

Object naming

Object naming / Object categorization

Object naming / Object categorization

sky

building

flag

face

banner

wall

street lamp

bus

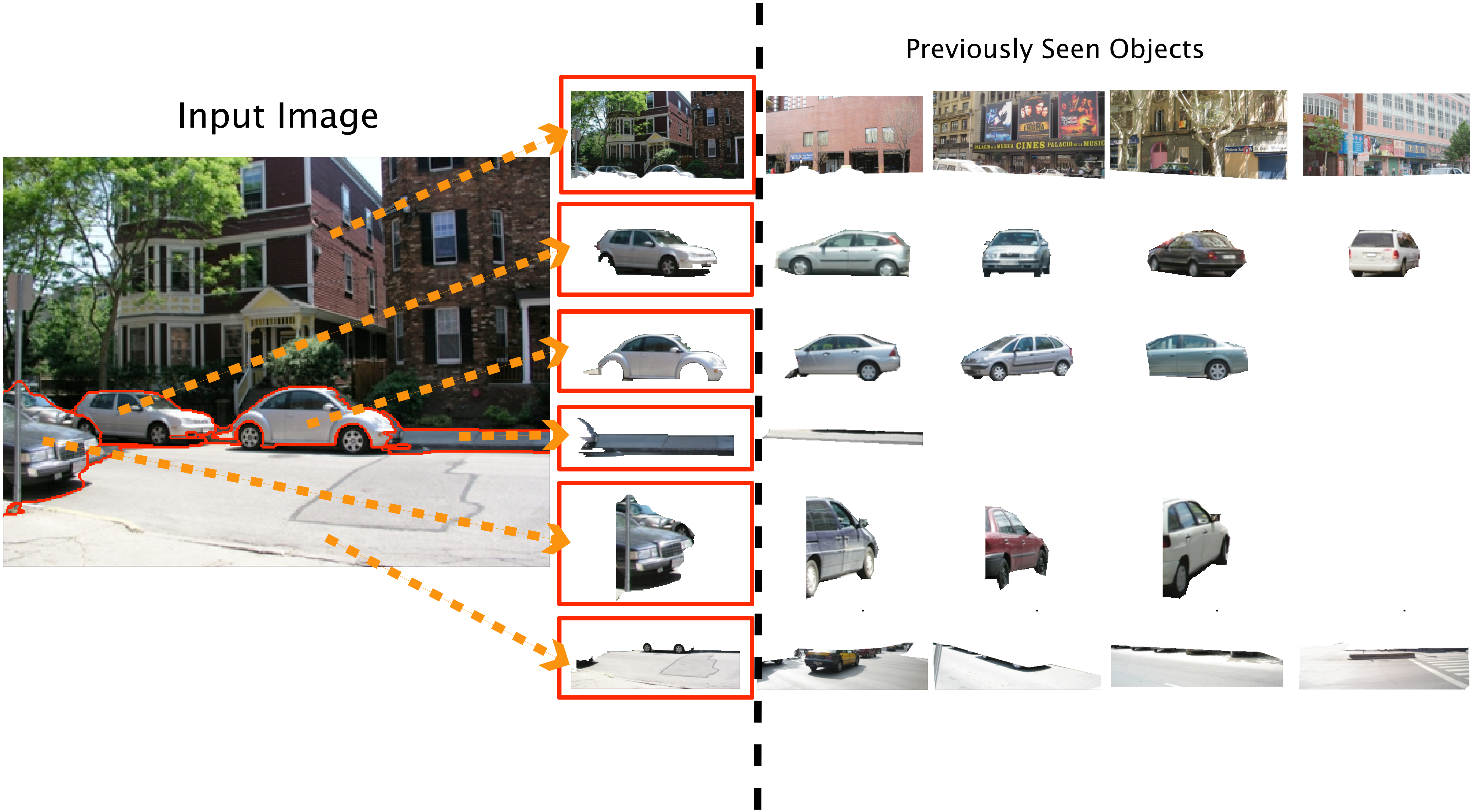
bus

cars

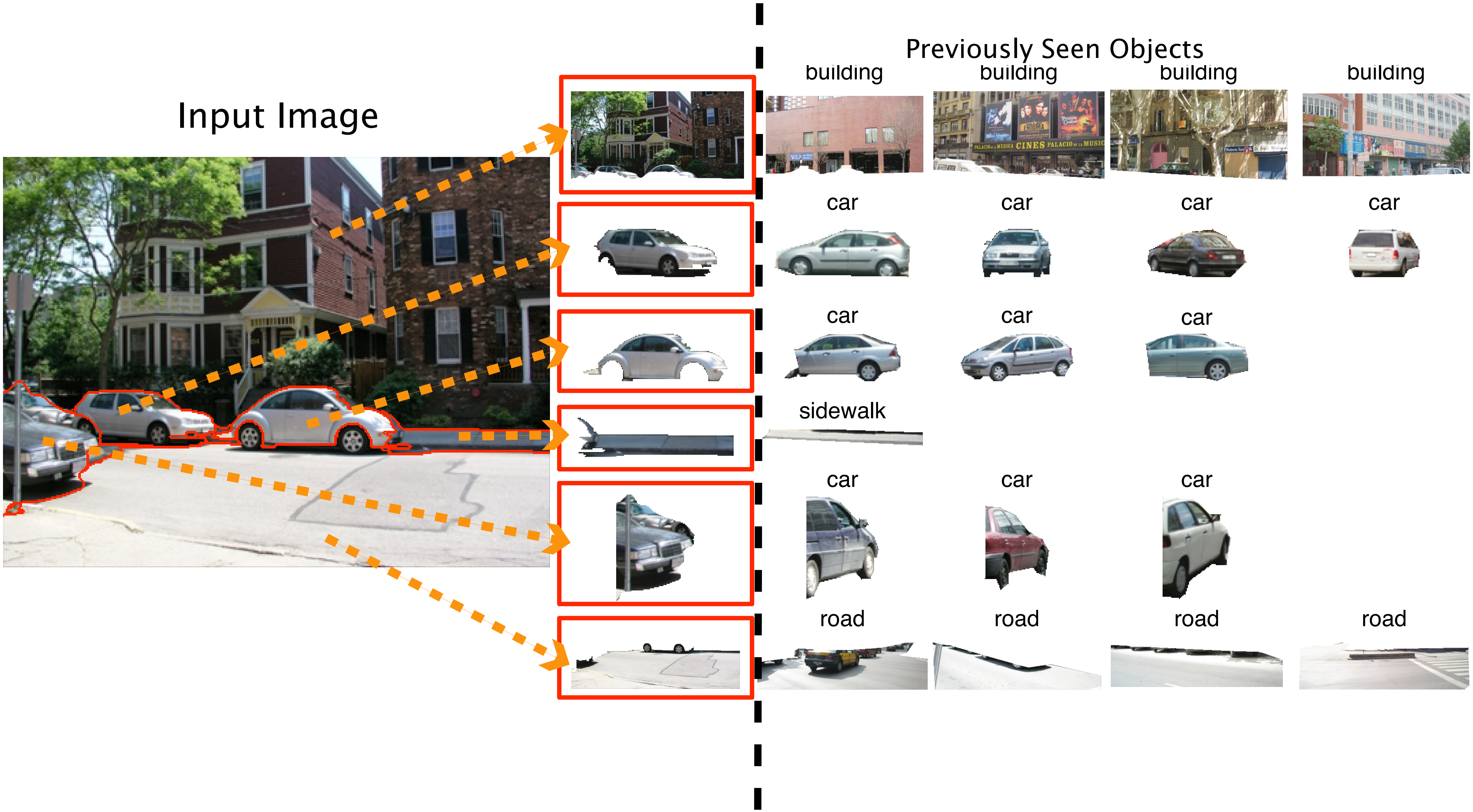
Different way of looking at recognition

Input Image

Different way of looking at recognition



Different way of looking at recognition



Our Contributions

- Posing Recognition as Association
  - Use large number of object exemplars

Our Contributions

- Posing Recognition as Association
  - Use large number of object exemplars

- Learning Object Similarity
  - Different distance function per exemplar

Our Contributions

- Posing Recognition as Association
  - –Use large number of object exemplars

- Learning Object Similarity
  - –Different distance function per exemplar

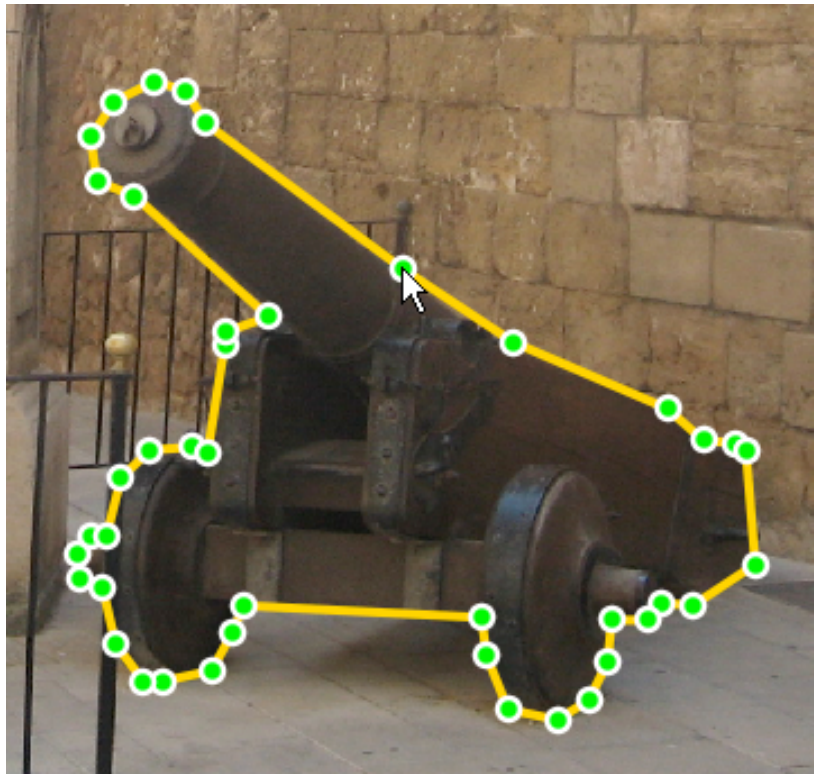
- Recognition-Based Object Segmentation
  - –Use multiple segmentation approach

Object Exemplars

- Extract objects from LabelMe with labels such as road, car, sky, tree, building, person
- Use the segmentation masks and labels provided by LabelMe annotators

LabelMe **Dataset**

12,905 Object Exemplars
171 unique 'labels'



B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision, May, 2008.

Measuring Similarity

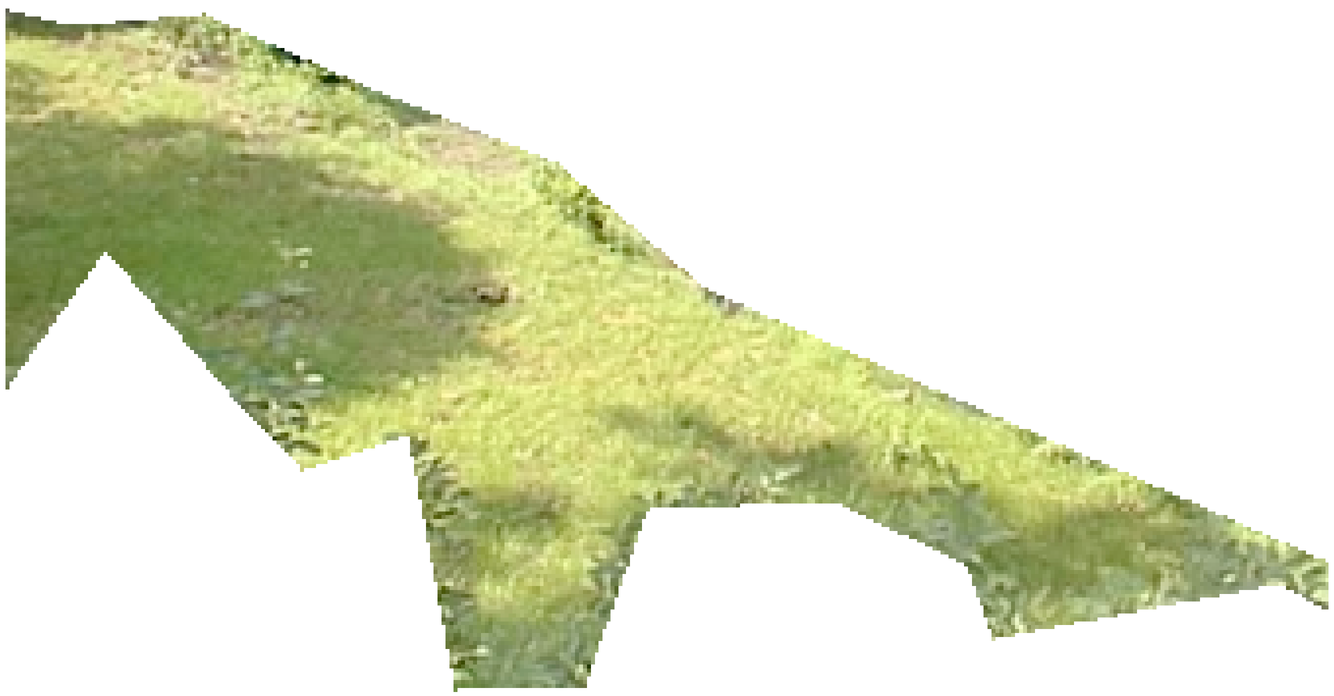
- How are objects similar?

Measuring Similarity

- How are objects similar?

Measuring Similarity

- How are objects similar?



Measuring Similarity

- How are objects similar?

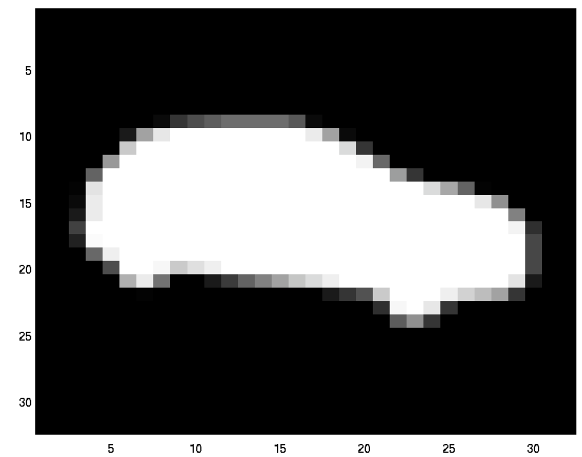
Measuring Similarity

- How are objects similar?

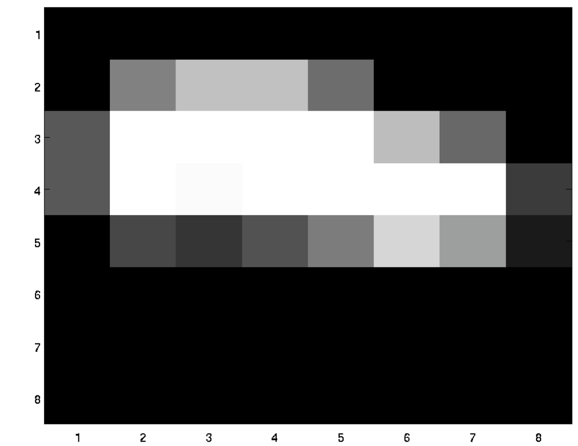
Exemplar Representation

Type	Name	Dimension
Shape	Centered Mask	32x32=1024
	BB Extent	2
	Pixel Area	1
Texture	Right Boundary Tex-Hist	100
	Top Boundary Tex-Hist	100
	Left Boundary Tex-Hist	100
	Bottom Boundary Tex-Hist	100
	Interior Tex-Hist	100
Color	Mean Color	3
	Color std	3
	Color Histogram	33
Location	Absolute Mask	8x8=64
	Top Height	1
	Bot Height	1

Centered Mask



Absolute Position Mask

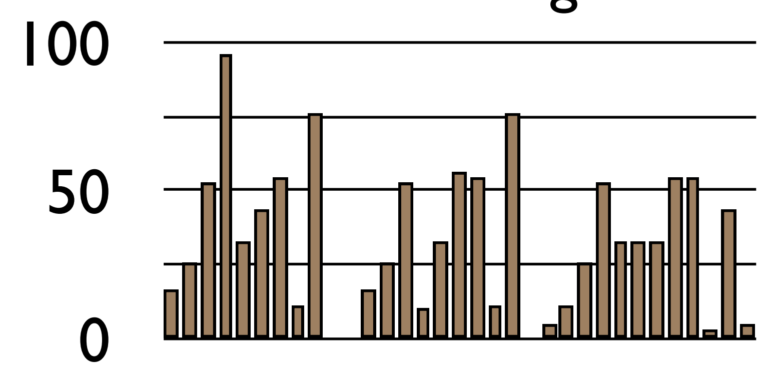


Texton Histogram

Top & Bottom Height

Boundary Texton Hist

Color Histogram



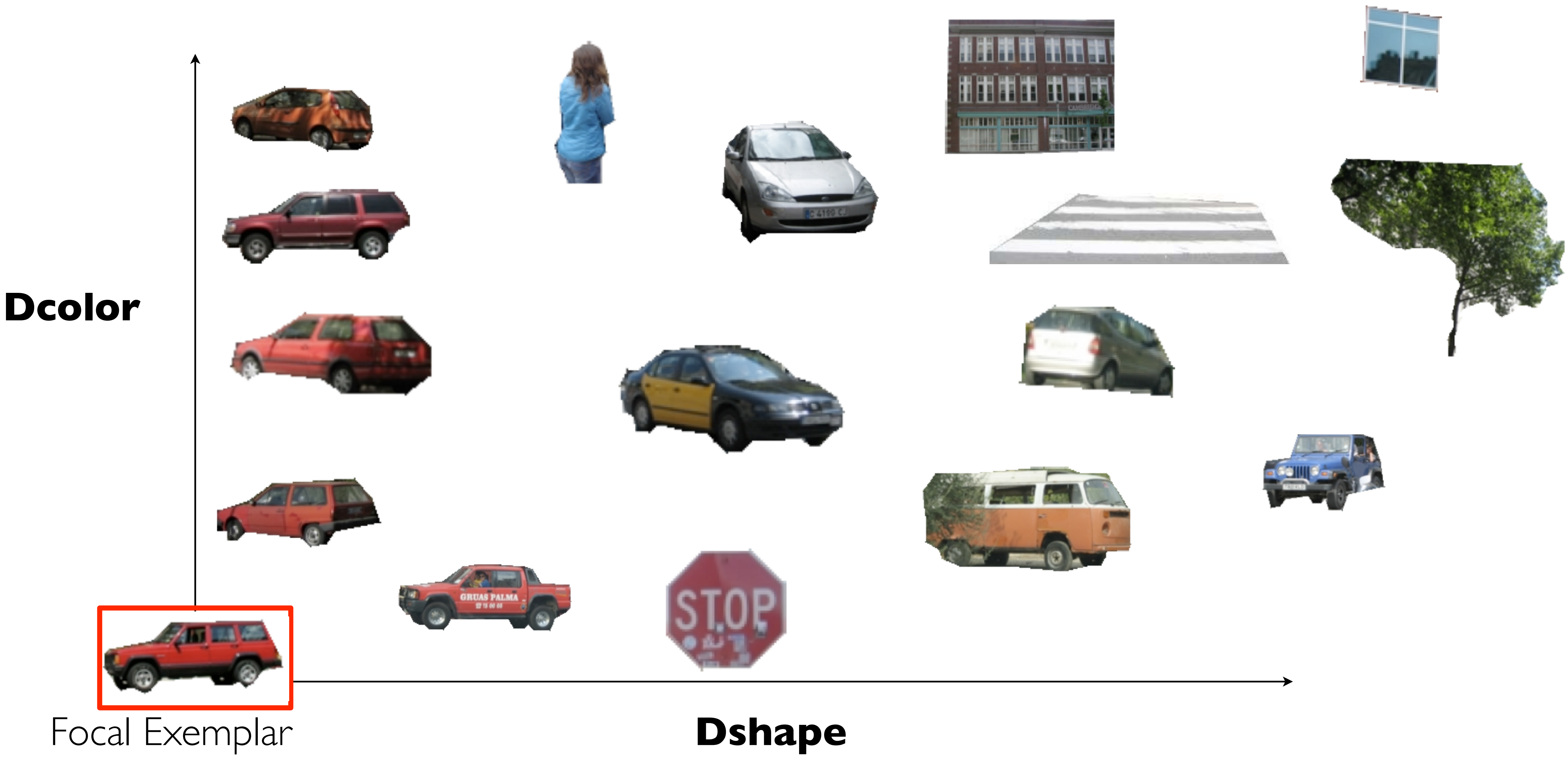
Learning a Per-Exemplar Similarity Measure

- We create a scalar distance between two objects by weighing the elementary distances differently
- A different set of weights -- a distance function -- is learned per exemplar

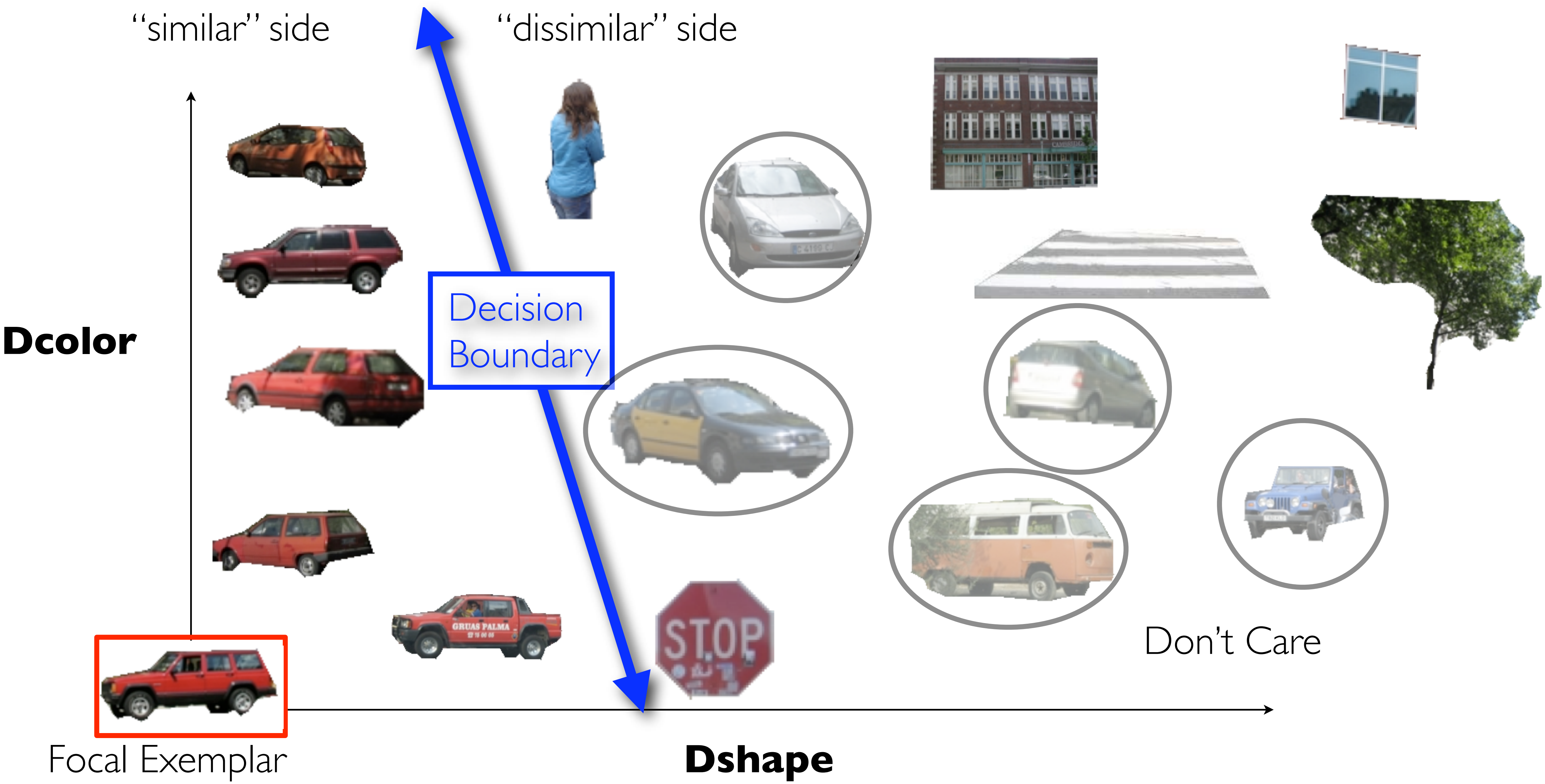
[1] Andrea Frome, Yoram Singer, Jitendra Malik. "Image Retrieval and Recognition Using Local Distance Functions." In NIPS, 2006.

[2] Andrea Frome, Yoram Singer, Fei Sha, Jitendra Malik. "Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification." In ICCV, 2007.

Learning Distance Functions

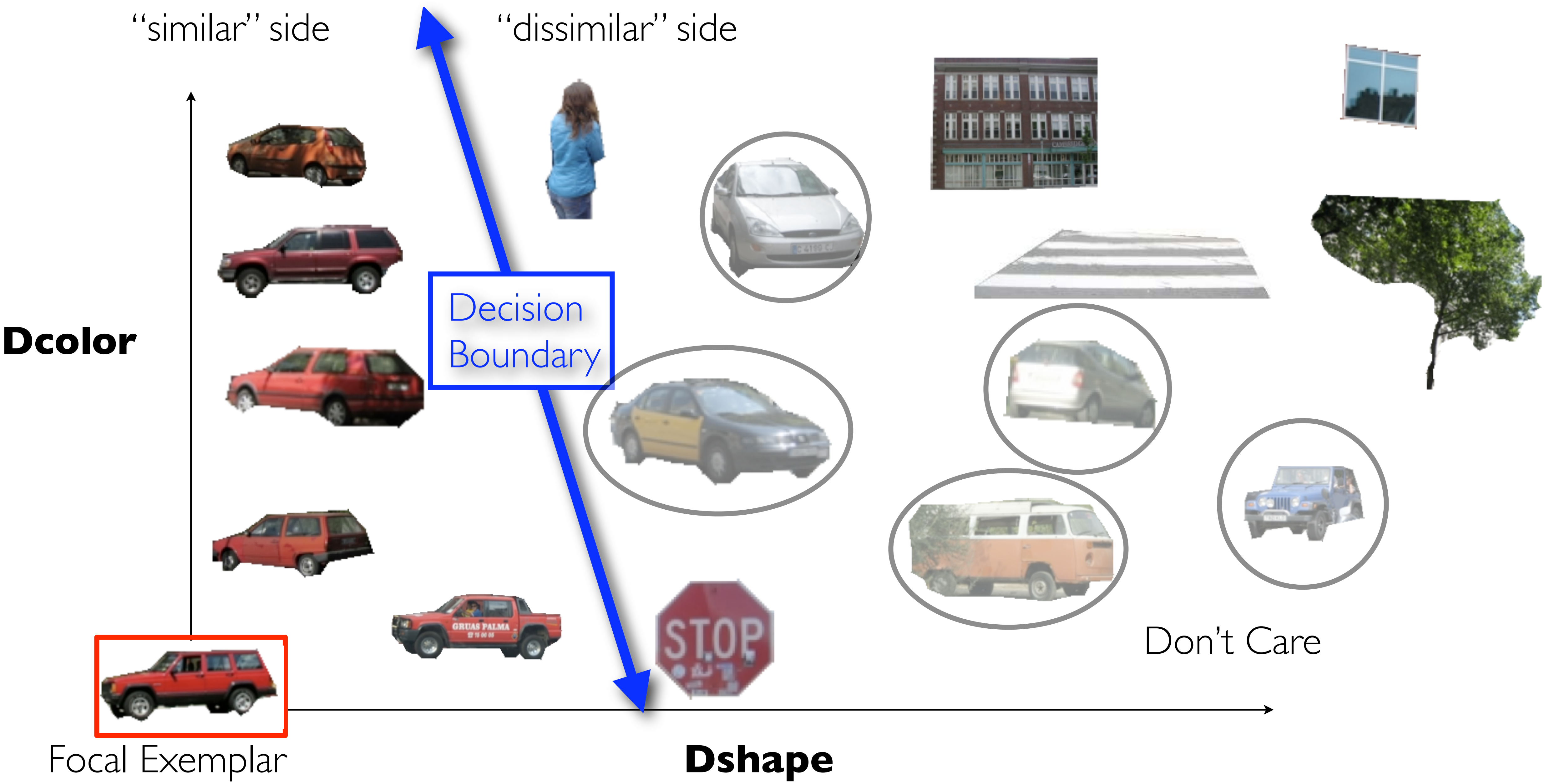


Learning Distance Functions



Learning Distance Functions

$$f(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{i \in C} \alpha_i L(-\mathbf{w} \cdot \mathbf{d}_i) + \sum_{i \notin C} L(\mathbf{w} \cdot \mathbf{d}_i)$$



w: positive weight vector

binary vector encodes which K exemplars are forced to be similar.

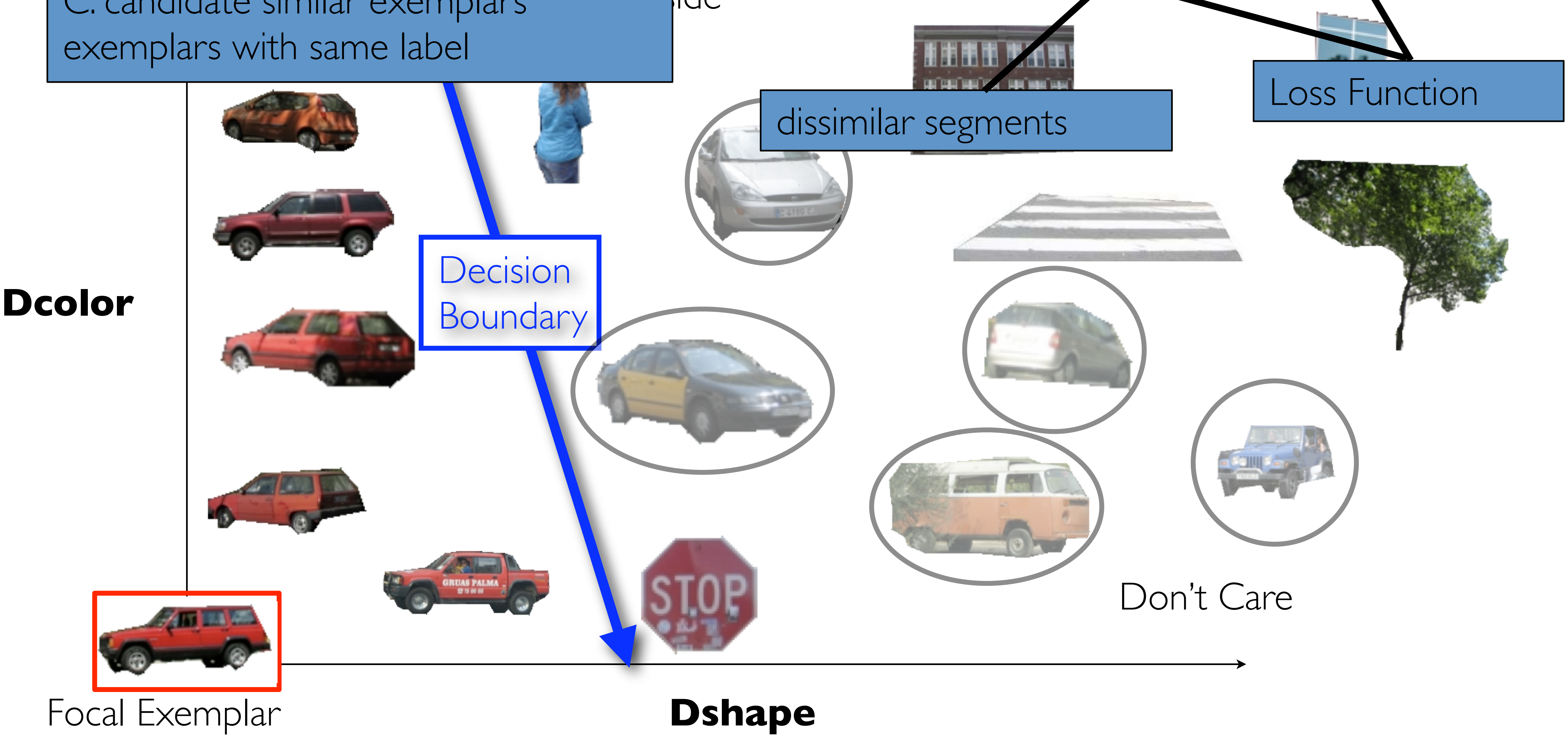
Learnin tions

$$f(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{i \in C} \alpha_i L(-\mathbf{w} \cdot \mathbf{d}_i) + \sum_{i \notin C} L(\mathbf{w} \cdot \mathbf{d}_i)$$

C: candidate similar exemplars
exemplars with same label

dissimilar segments

Loss Function



Learning Distance Functions

$$f(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{i \in C} \alpha_i L(-\mathbf{w} \cdot \mathbf{d}_i) + \sum_{i \notin C} L(\mathbf{w} \cdot \mathbf{d}_i)$$

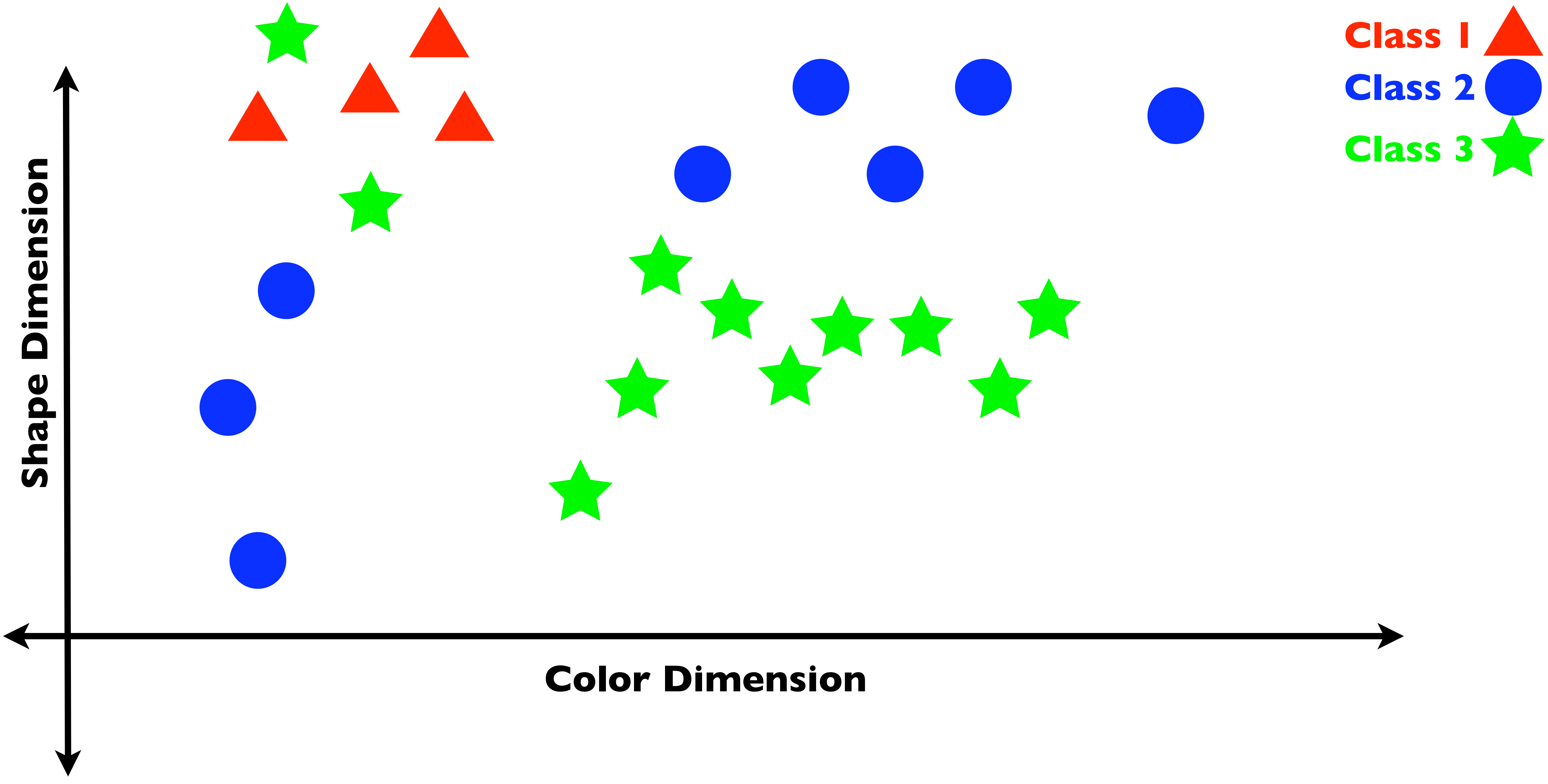
Iterative Optimization

$$\boldsymbol{\alpha}^k = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i \in C} \alpha_i L(-\mathbf{w}^{\mathbf{k}} \cdot \mathbf{d}_{\mathbf{i}})$$

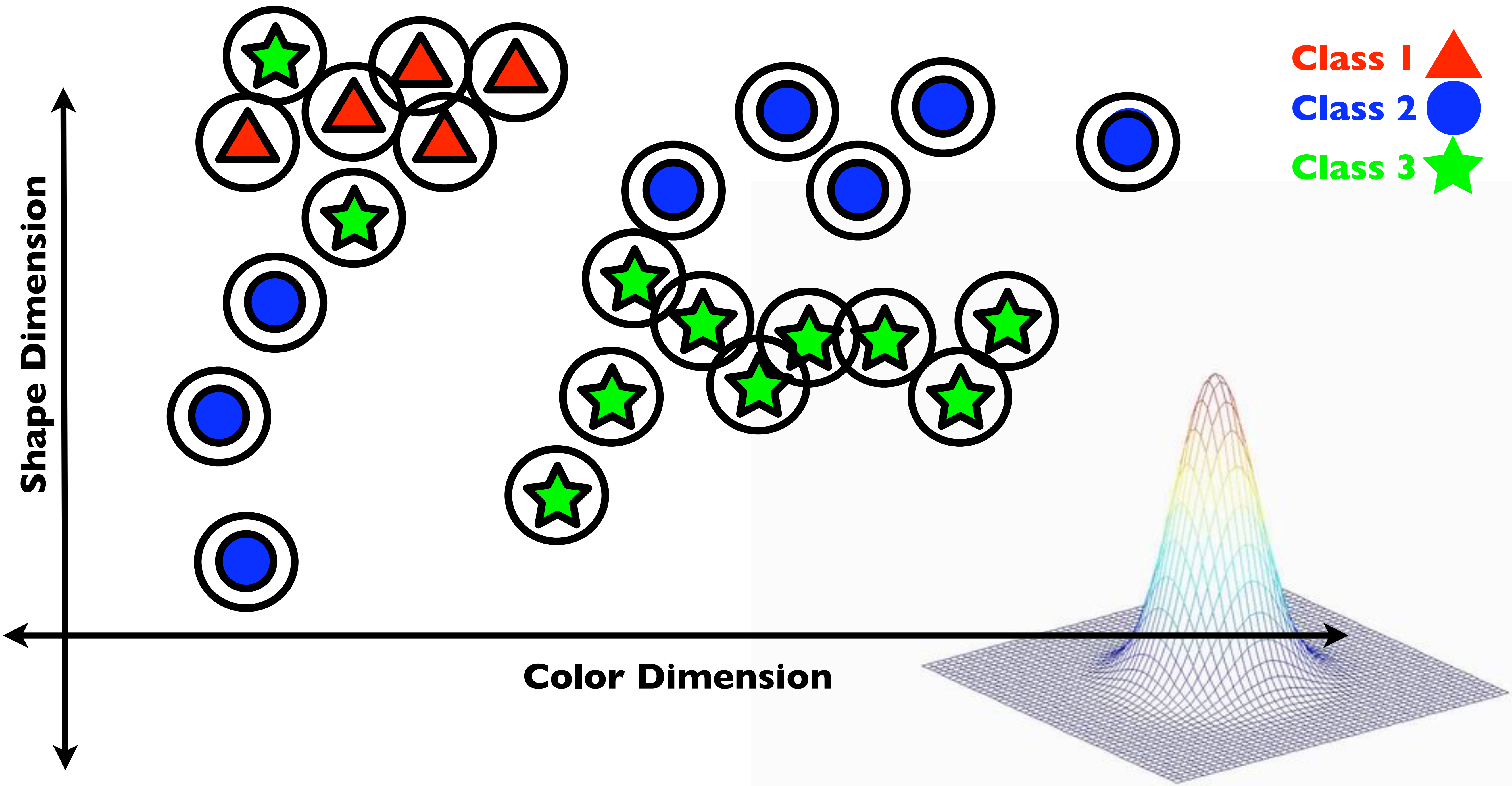
$$\mathbf{w}^{k+1} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i:\alpha_i^k = 1} L(-\mathbf{w} \cdot \mathbf{d}_i) + \sum_{i \notin C} L(\mathbf{w} \cdot \mathbf{d}_i)$$

alpha sums to K=10 (forced number of similar exemplars)
L: squared hinge-loss function (SVM optimization)
initialize with texton histogram distance (works well for a wide array of objects!)

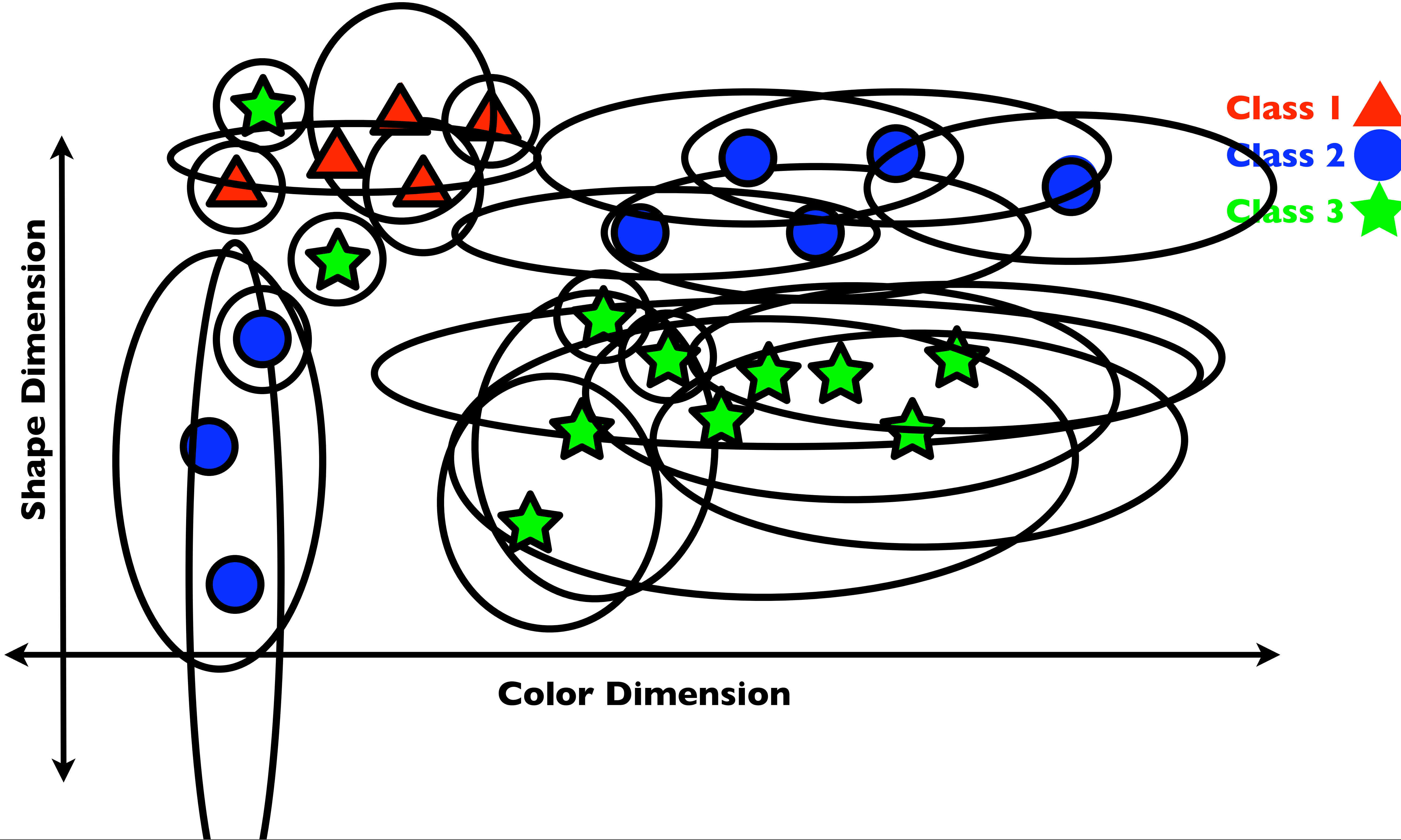
Non-parametric density estimation



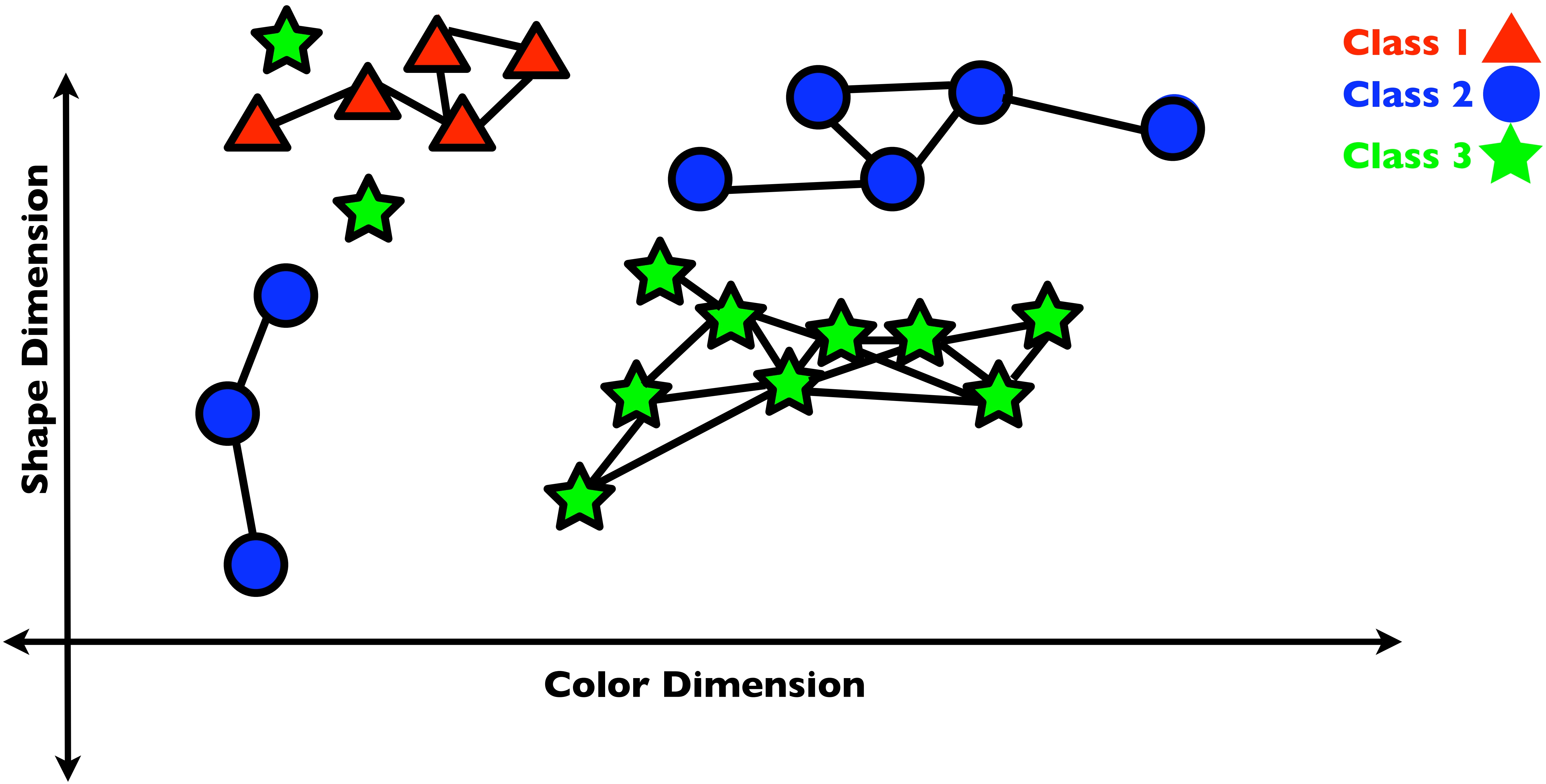
Non-parametric density estimation



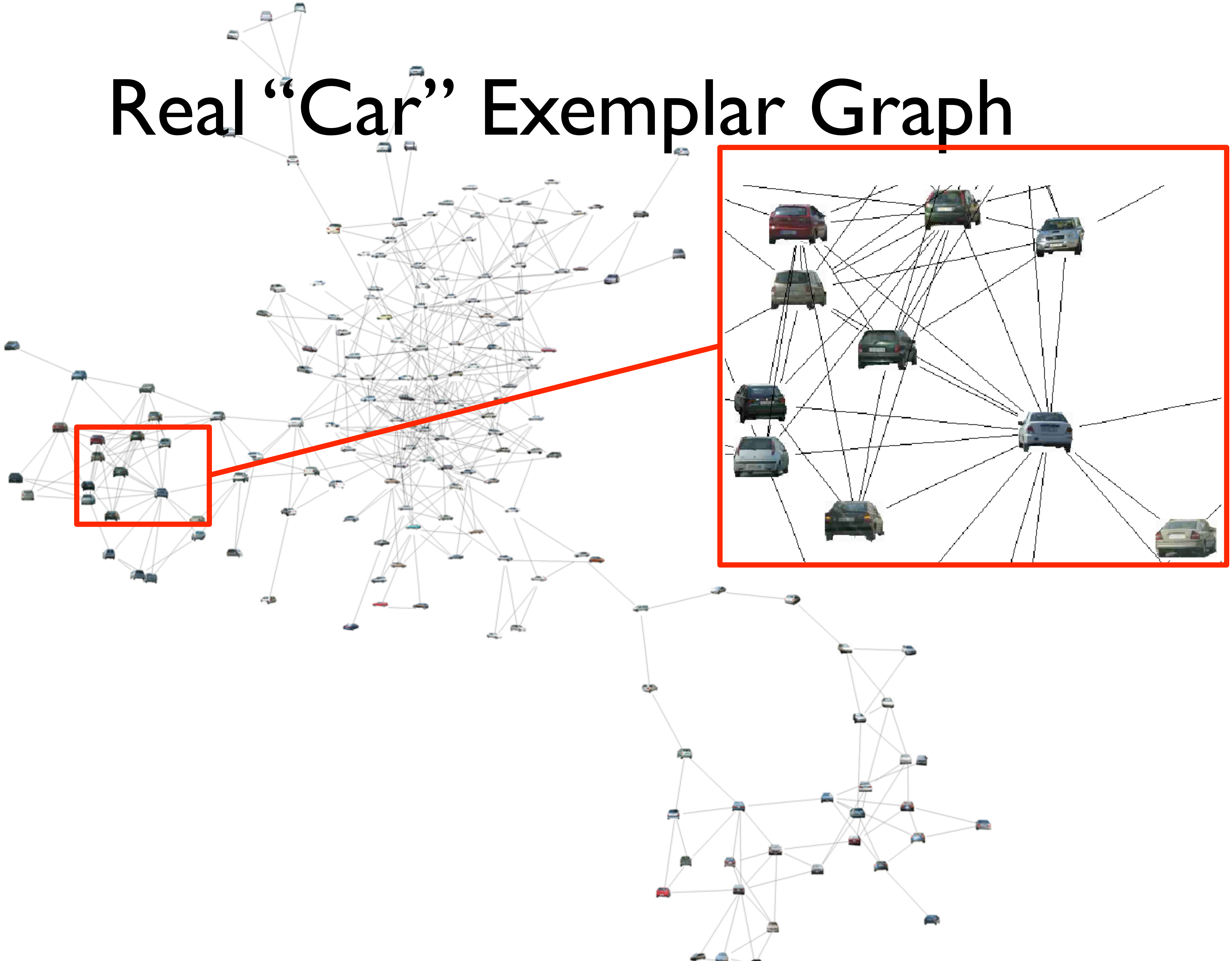
Non-parametric density estimation



Exemplar Graph



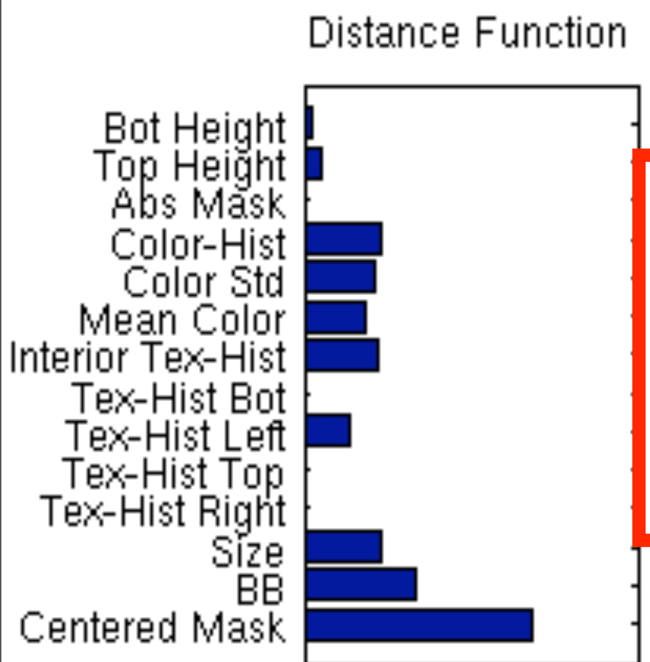
Real "Car" Exemplar Graph



Visualizing Distance Functions (Training Set)

Query

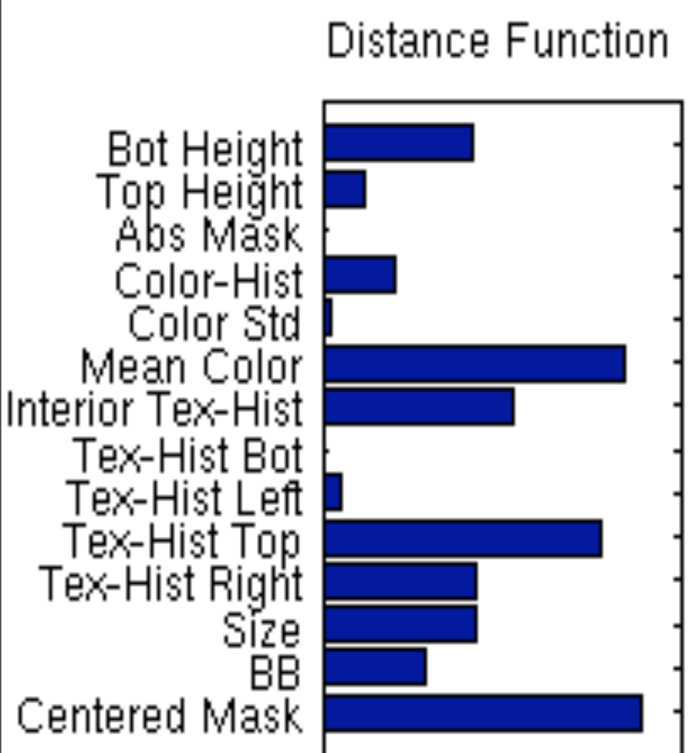
Top Neighbors with Tex-Hist Dist



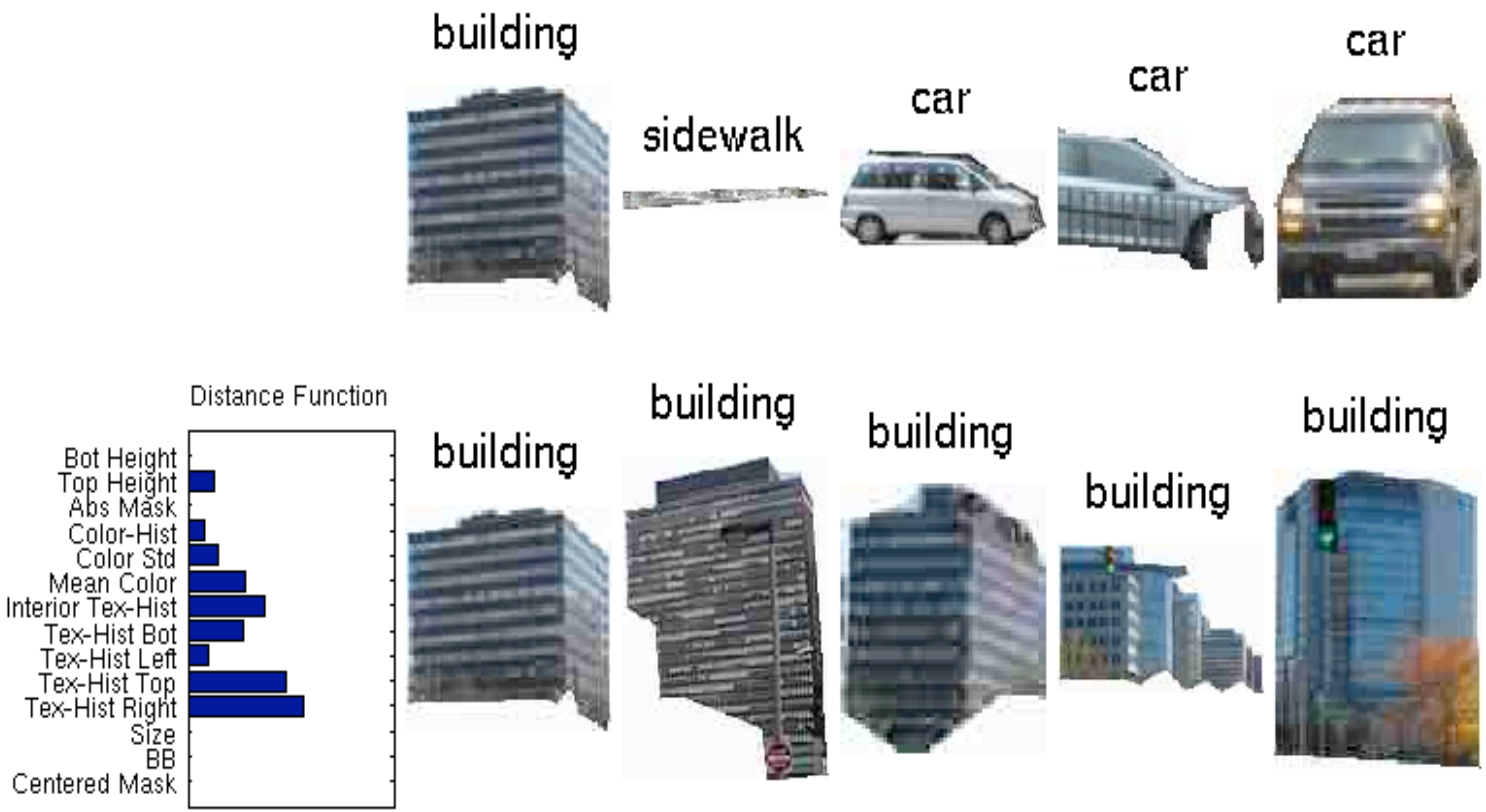
Query

Top Neighbors with Learned Dist

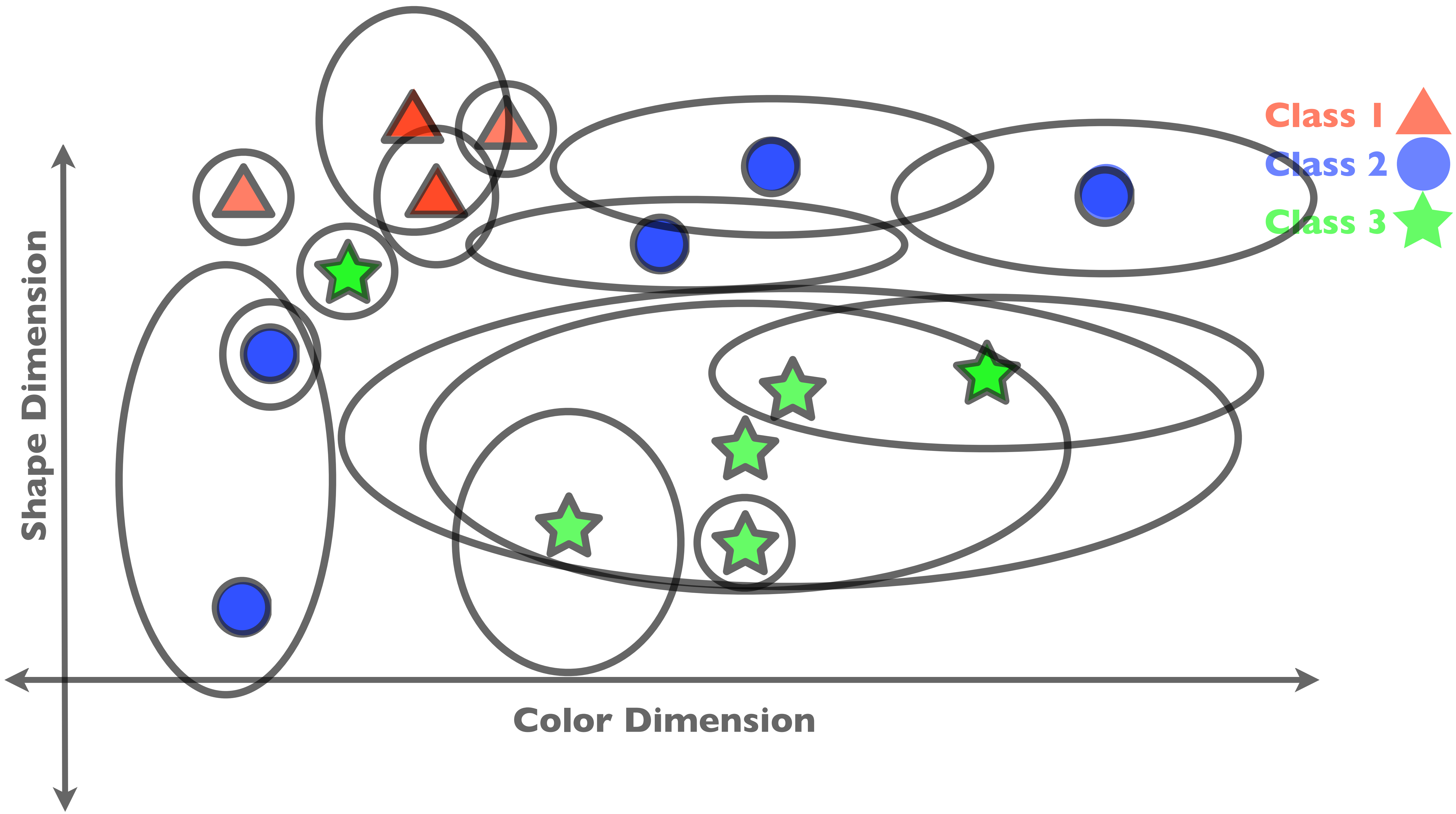
Visualizing Distance Functions (Training Set)



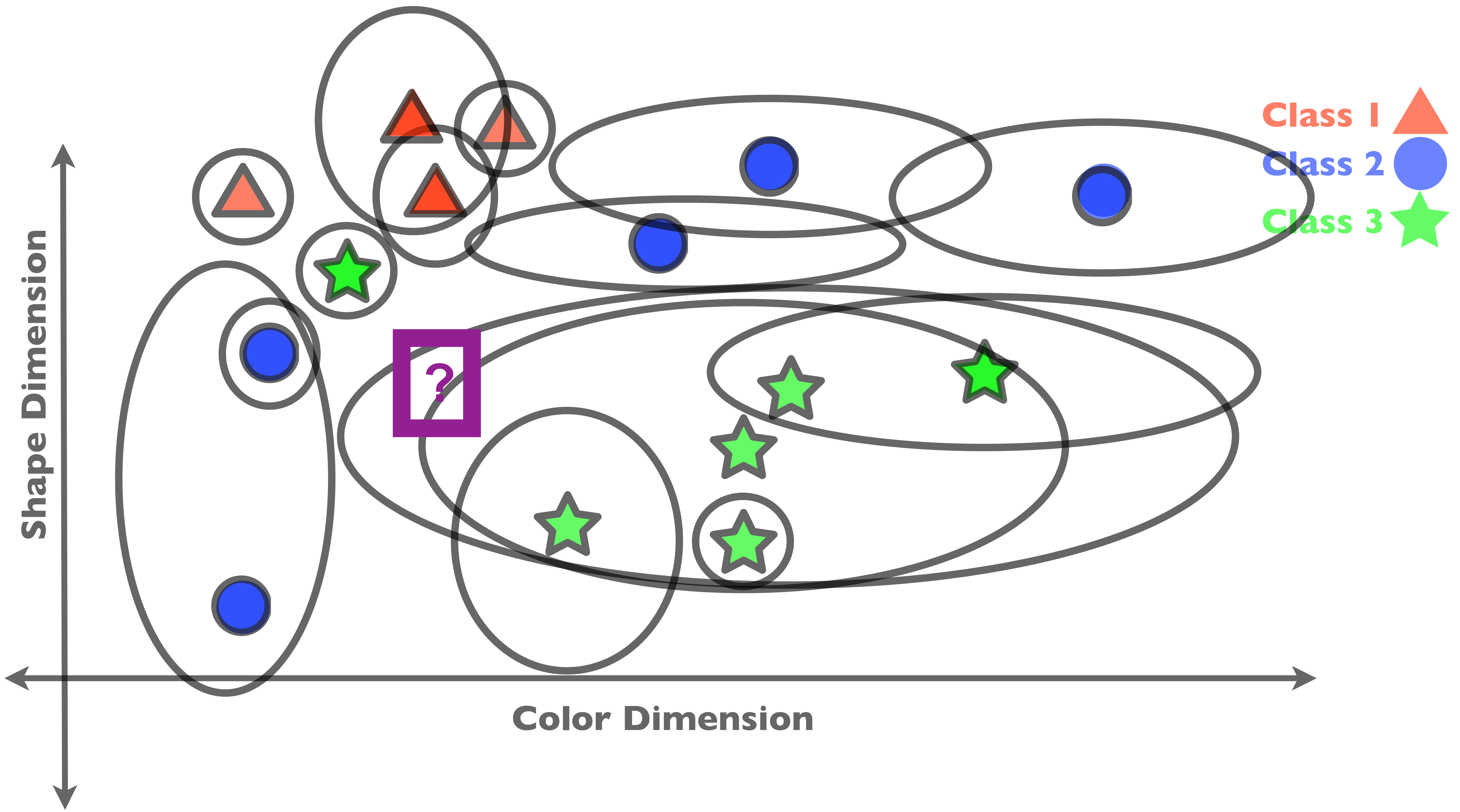
Visualizing Distance Functions (Training Set)



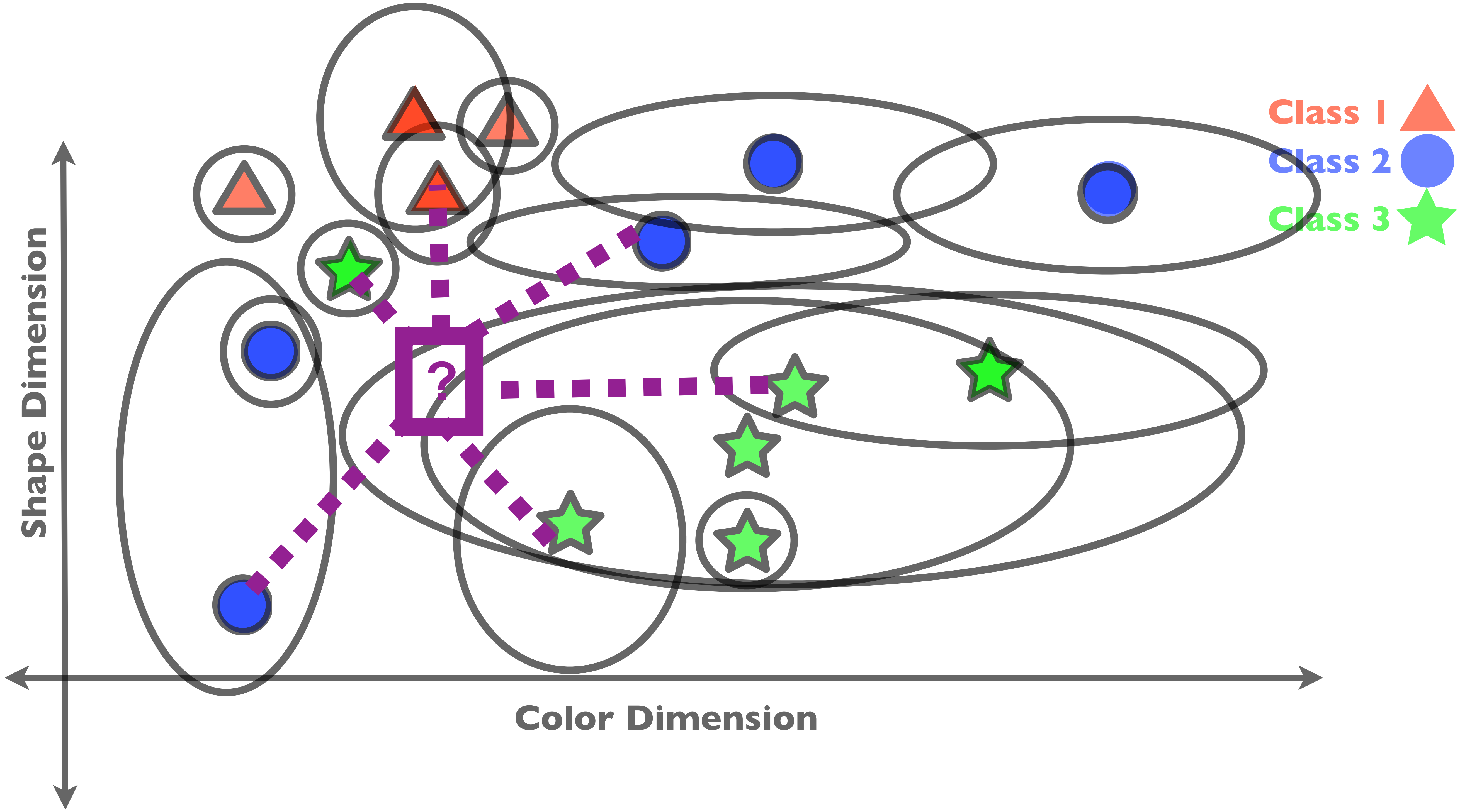
Recognition Time



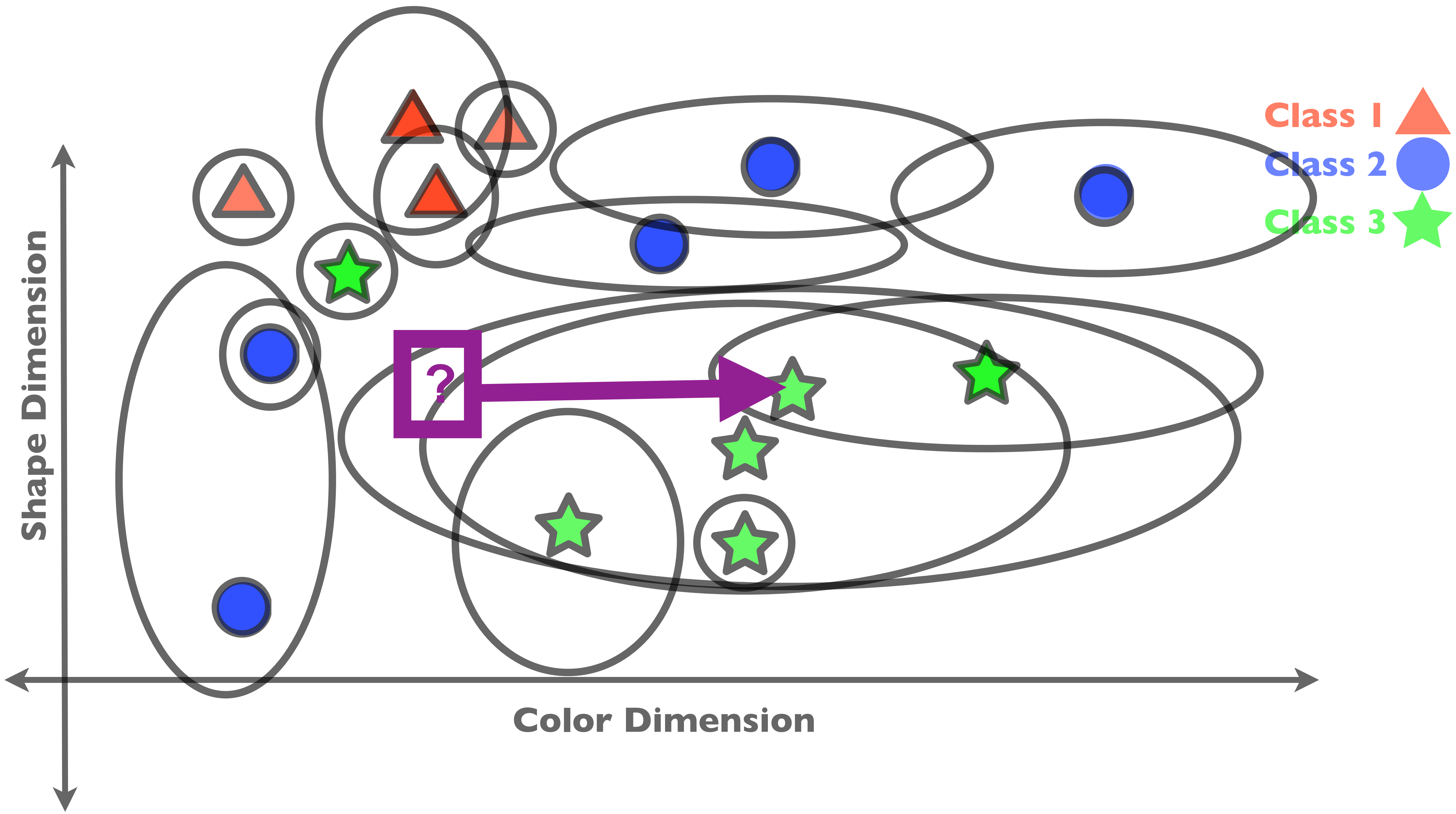
Recognition Time



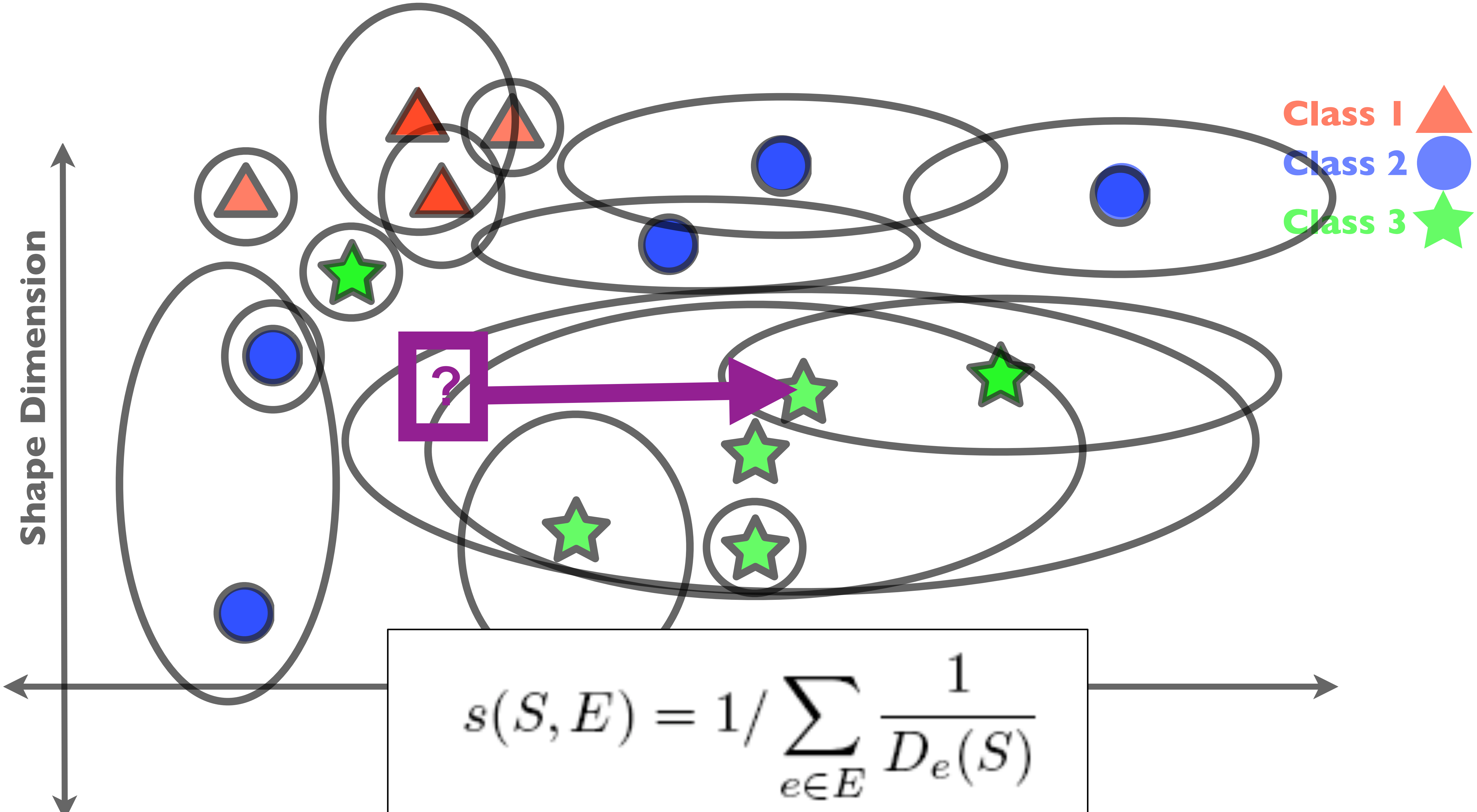
Recognition Time



Recognition Time



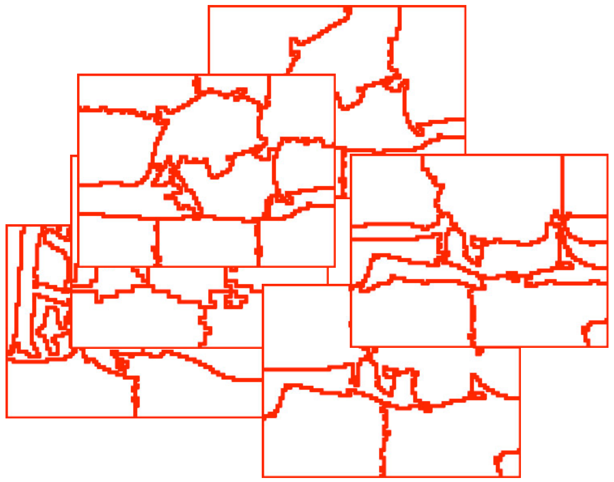
Recognition Time



$$s(S, E) = 1 / \sum_{e \in E} \frac{1}{D_e(S)}$$

Object Segmentation via Recognition

- Generate Multiple Segmentations (Hoiem 2005, Russell 2006, Malisiewicz 2007*)
  - Mean-Shift and Normalized Cuts
  - Use pairs and triplets of adjacent segments
  - Generate about 10,000 segments per image



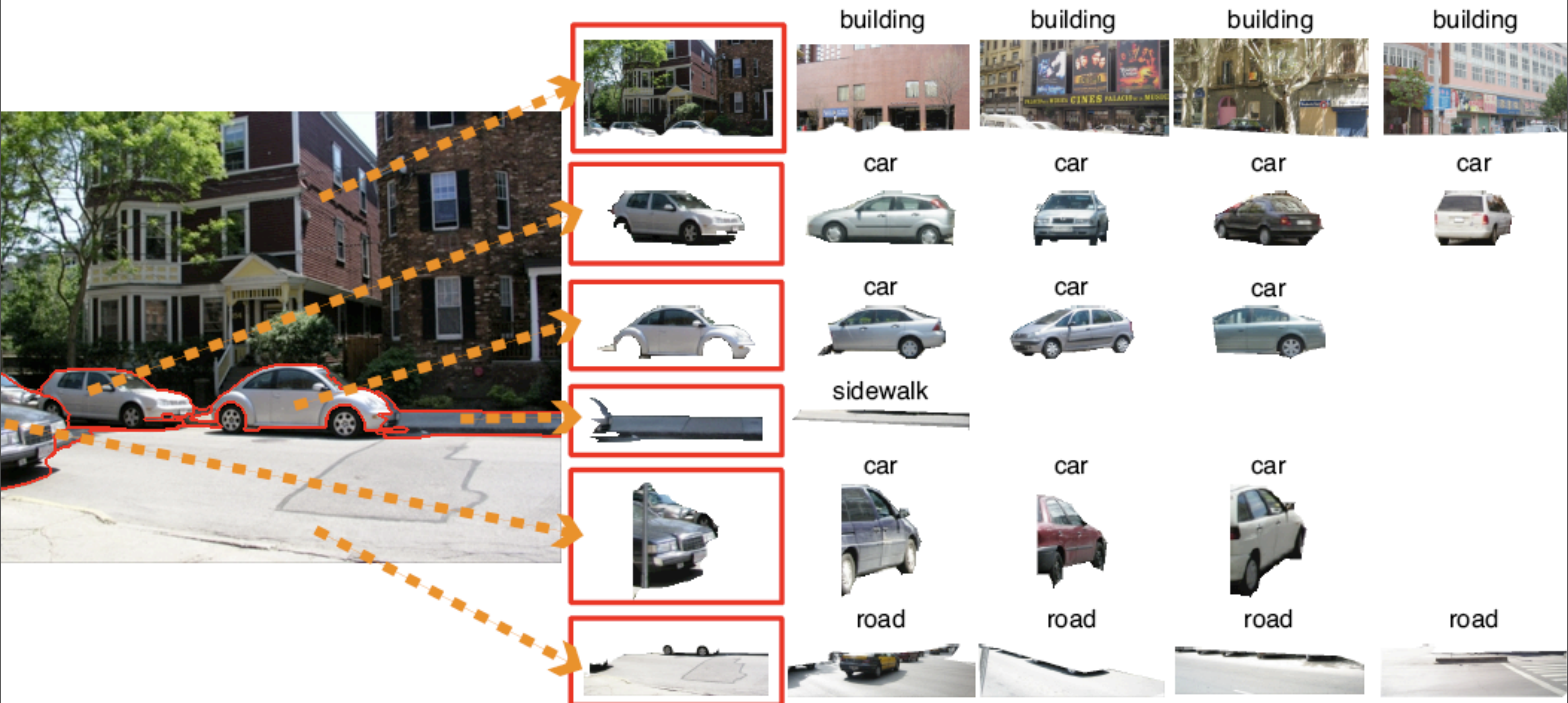
- Enhance training with bad segments
- Apply learned distance functions to bottom-up segments

Tomasz Malisiewicz, Alexei A. Efros. Improving Spatial Support for Objects via Multiple Segmentations, In BMVC 2007.

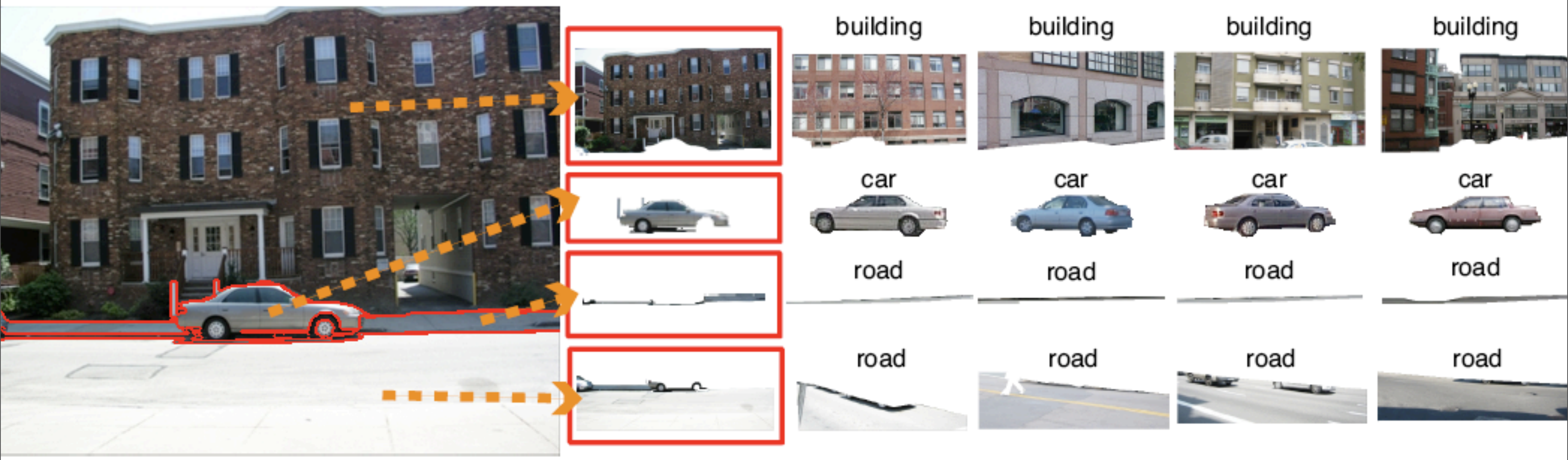
Top Object Hypotheses in Test Set

Toward Image Parsing

Toward Image Parsing



Toward Image Parsing



Observations + Conclusions

- Exemplar model and segment-centric features work well for both free-form stuff like grass and fixed-extent things like cars
- Distance Functions are good at localizing objects for which we have observed many instances
- Success relies on having ground truth segmentations during learning
- Need a clever way to integrate object hypotheses to parse images