Meta-Recognition, Machine Learning and the Open Set Problem

Walter J. Scheirer
Director of Research & Development, Securics, Inc.
Assistant Professor Adjoint at the University of Colorado at Colorado Springs

What is recognition in computer vision?

- Compare an object to a known set of classes, producing a similarity measure to each

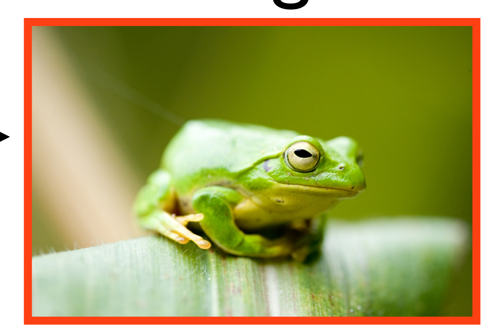
Quiet brown frog (cc) by Olivier Ffrench

What is this?

Teapot

Red teapot (cc) by fraise

Frog



Frog on corn leaf (cc) by Joi Ito

Girl

Lovely little girl:) (cc) by BirdCantFly

Why is recognition hard?

Eye (cc) by Michele Catania

The same object can cast an infinite number of different images onto the retina[1] (humans) or an innumerable number of images on a sensor (machine)

1. D. Cox, J. DiCarlo, and N. Pinto, MIT 6.963 Lecture, "A High-Throughput Approach to Discovering Good Forms of Visual Representation"

Image by Michele Catania "Eye" BY http://www.flickr.com/photos/cataniamichele/

Why is recognition hard?

fugu! ⓒ by svacher

fugu - top profile ⓒ by svacher

fugu - side profile ⓒ by svacher

Image by svacher “fugu!” BY http://www.flickr.com/photos/trufflepig/
Image by svacher “fugu – top profile” BY http://www.flickr.com/photos/trufflepig/
Image by svacher “fugu – side profile” BY http://www.flickr.com/photos/trufflepig/

Why is recognition hard?

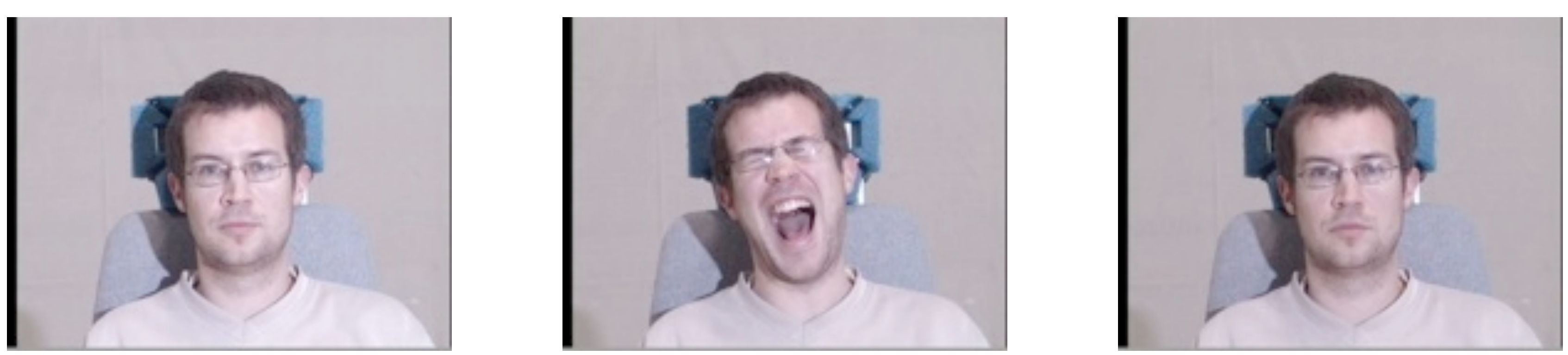


Image credit: CMU Multi-PIE Database, http://www.multipie.org/

What strategies do we have to approach this problem?

- Multiple-View Geometry
- 3D Modeling
- Invariant Feature Descriptors
- Data Fusion
- Machine Learning

Data Fusion

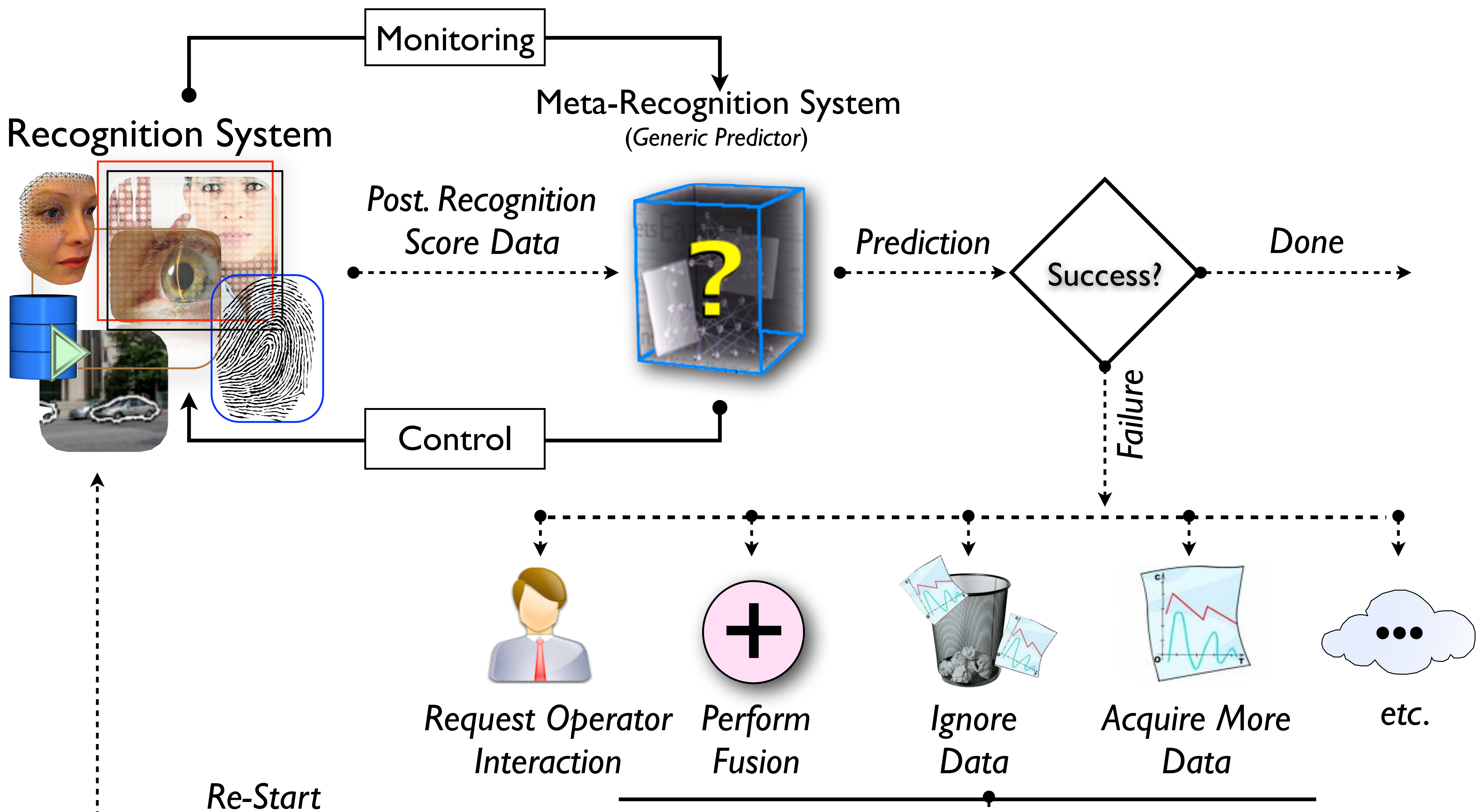
- A single algorithm is not a complete solution for a recognition task
- Combine information across algorithms and sensors[1]
  - Decision fusion
  - Score level normalization & fusion

Do this is a **robust** manner...

1. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer, 2006

Meta-Recognition

Goal: Predict if a recognition result is a success or failure



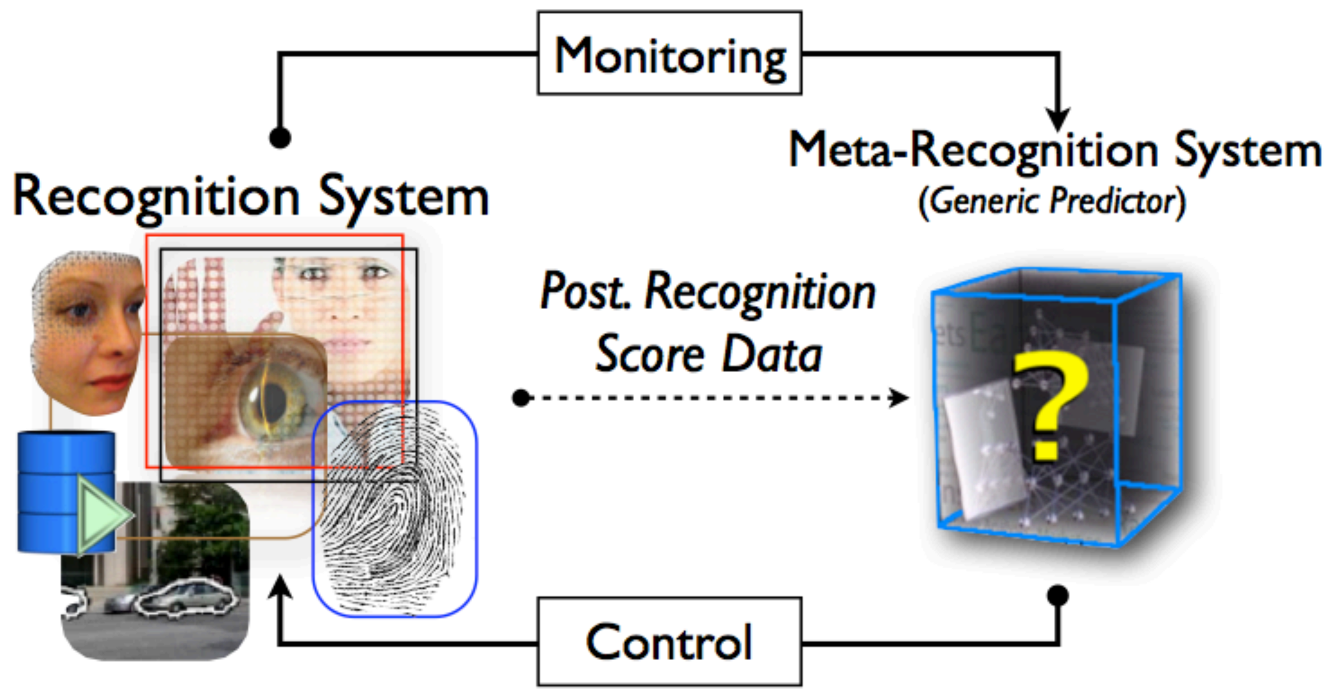
1. W. Scheirer et al., "Meta-Recognition: the Theory and Practice of Recognition Score Analysis," IEEE T-PAMI, August 2011

From Meta-Cognition to Recognition

- Inspiration: *Meta-Cognition* Study
  - "knowing about knowing[1]"
  - Example: If a student has more trouble learning history than math, she "knows" something about her learning ability and can take corrective action

1. J. Flavell and H. Wellman, "Metamemory," in *Perspectives on the Development of Memory and Cognition*, 1988, pp. 3-33

Meta-Recognition Defined



Let X be a recognition system. Y is a meta-recognition system when recognition state information flows from X to Y, control information flows from Y to X, and Y analyzes the recognition performance of X, adjusting the control information based on the observations.

Can’t we do this with say... image quality?

8 47

191 Gallery

Apparent quality is not always tied to rank.

- Quality is good as an “overall” predictor
  - Over a large series of data and time

- Quality does not work as a “per instance” predictor
  - One image analyzed at a time...

Challenges for Image Quality Assessment

- Interesting recent studies from the National Institute of Standards and Technology
  - Iris[1]: three different quality assessment algorithms lacked correlation
  - Face[2]: out of focus imagery was shown to produce better match scores

"Quality is not in the eye of the beholder; it is in the recognition performance figures!"   - Ross Beveridge

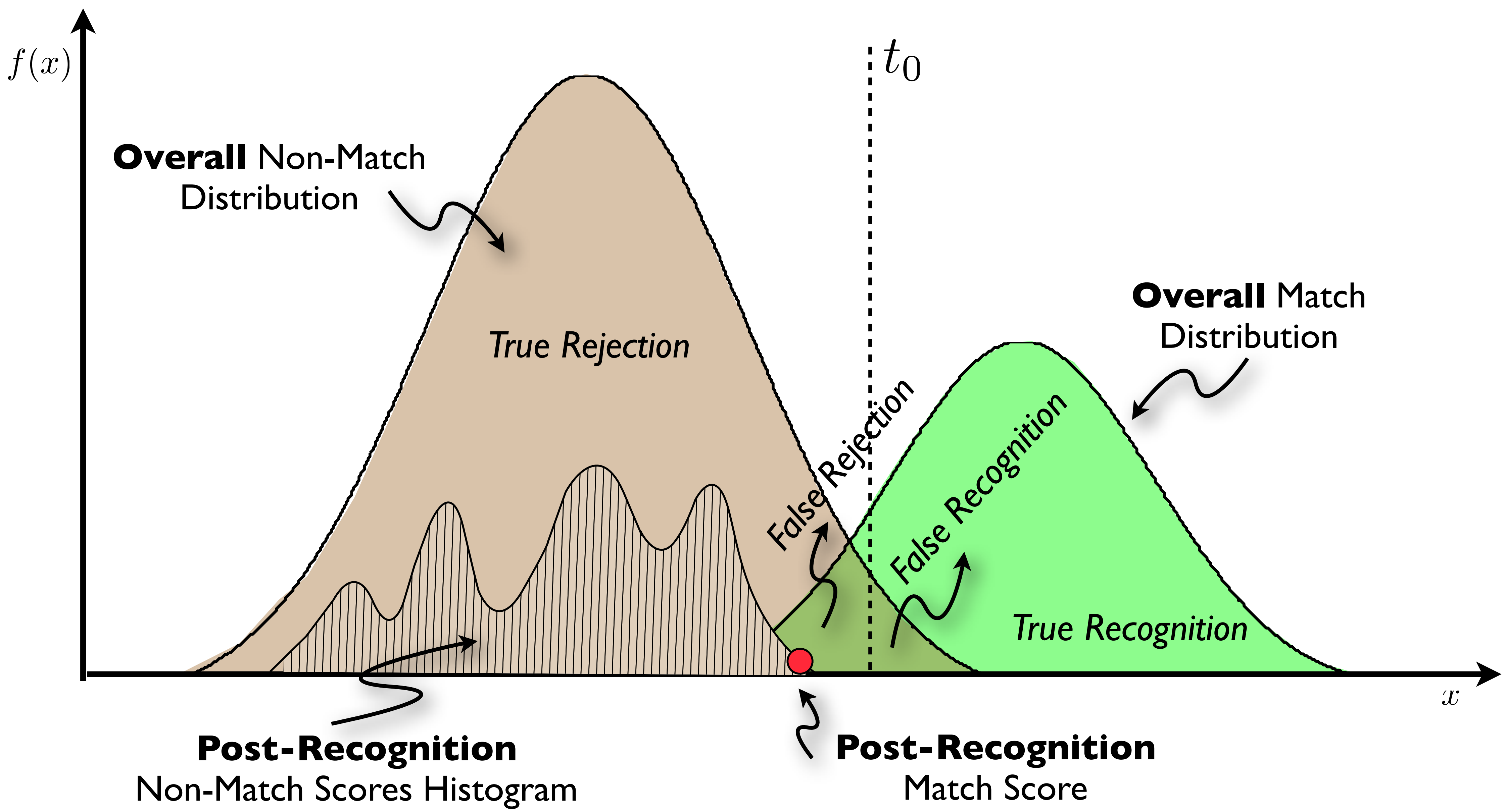
1. P. Flynn, "ICE Mining: Quality and Demographic Investigations of ICE 2006 Performance Results," MBGC Kick-off workshop, 2008
2. R. Beveridge, "Face Recognition Vendor Test 2006 Experiment 4 Covariate Study," MBGC Kick-off workshop, 2008

What about cohorts?

- A likely related phenomenon to Meta-Recognition
- Post-*verification* score analysis
- Model a distribution of scores from a pre-defined "cohort gallery" and then normalize data[1]
    - This estimate valid "score neighbors"
    - A claimed object should be followed by its cohorts with a high degree of probability
- Intuitive, but lacks a theoretical basis

1. S. Tulyakov et al., "Comparison of Combination Methods Utilizing t-normalization and Second Best Score Models," IEEE Workshop on Biometrics, 2008.

Recognition Systems



Formal definition of recognition

Find[1] the class label c^*, where p_k is an underlying probability rule and p_0 is the input distribution satisfying:

$$c^* = \underset{class\ c}{\operatorname{argmax}} \Pr(p_0 = p_c)$$

subject to $\Pr(p_0 = p_{c^*}) \geq 1 - \delta$, for a given confidence threshold δ. We can also conclude a lack of such class.

Probe: input image p_0 submitted to the system with corresponding class label c^*.

Gallery: all the classes c^* known by the recognition system.

1. G. Shakhnarovich, et al. "Face Recognition from Long-term Observations," ECCV, 2002.

Rank-1 Prediction as a Hypothesis Test

- Formalization of Meta-Recognition
  - Determine if the top K scores contain an outlier with respect to the current probe's match distribution
- Let $\mathcal{F}(p)$ be the non-match distribution, and $m(p)$ be the match score for that probe.
- Let $S(K) = s_1 \ldots s_k$ be the top K sorted scores

Hypothesis Test: H_0 (failure) : $\forall x \in S(K), x \in \mathcal{F}(p)$

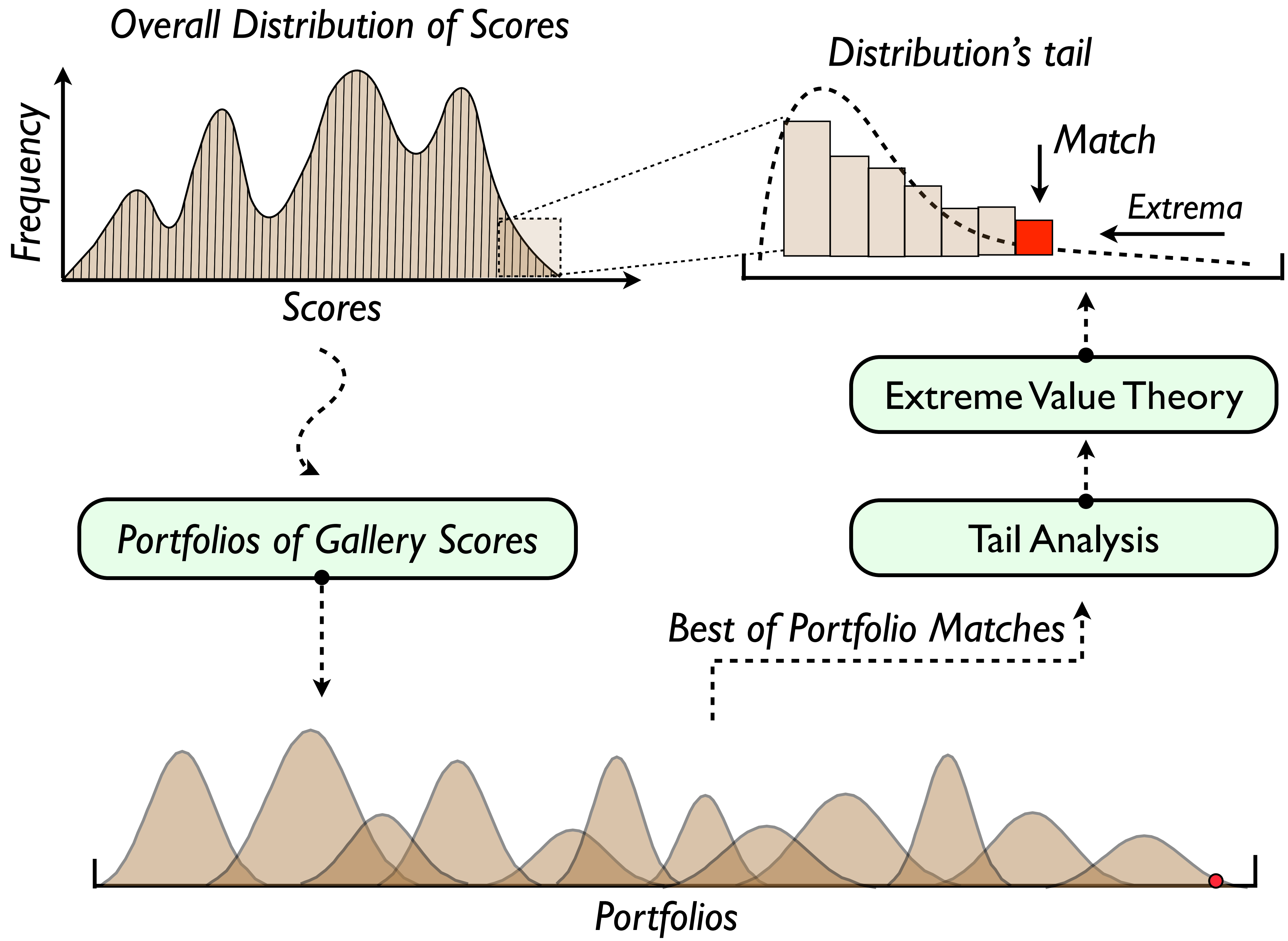
If we can reject H_0, then we predict success.

The Key Insight

We don’t have enough data to model the match distribution, but we have n samples of the non-match distribution - good enough for non-match modeling and outlier detection.

If the best score is a match, then it should be an outlier with respect to the non-match model.

A Portfolio Model of Recognition



The Extreme Value Theorem

Let $(s_1, s_2, \dots, s_n)$ be a sequence of i.i.d. samples. Let $M_n = \max\{s_1, \dots, s_n\}$. If a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and

$$\lim_{x\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$$

then if F is a non-degenerate distribution function, it belongs to one of three extreme value distributions[1].

The i.i.d. constraint can be relaxed to a weaker assumption of exchangeable random variables[2].

1. S. Kotz and S. Nadarajah, Extreme Value Distributions: Theory and Applications, 1st ed. World Scientific Publishing Co., 2001.

2. S. Berman, "Limiting Distribution of the Maximum Term in Sequences of Dependent Random Variables," Ann. Math. Stat., vol. 33, no. 3, pp. 894-908, 1962.

The Weibull Distribution

The sampling of the top-n scores always results in an EVT distribution, and is *Weibull* if the data are bounded[1].

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Choice of this distribution is not dependent on the model that best fits the entire non-match distribution.

1. *NIST/SEMATECH e-Handbook of Statistical Methods*, ser. 33. U.S. GPO, 2008

Rank-1 Statistical Meta-Recognition

Require: a collection of similarity scores S

1. **Sort** and retain the n largest scores, $s_1, \ldots, s_n \in S$;

2. **Fit** a Weibull distribution W_S to $s_2, \ldots, s_n$, skipping the hypothesized outlier;

3. **if** $Inv(W_S(s_1)) > \delta$ **do**

4. s_1 is an outlier and we reject the failure prediction (null) hypothesis H_0

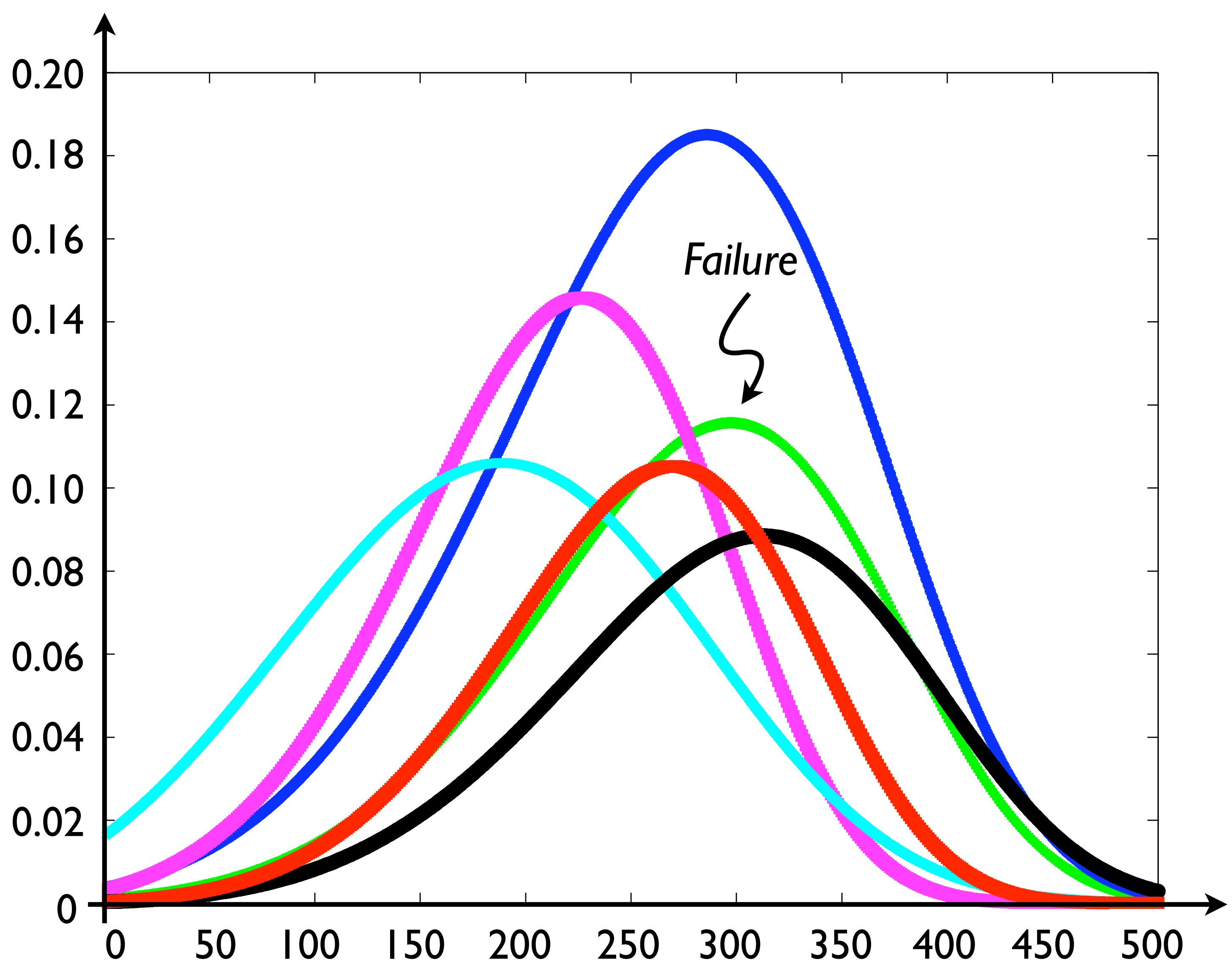
6. **end if**

δ is the hypothesis test "significance" level threshold

Good performance is often achieved using $\delta = 1 - 10^{-8}$

Can't we just look at the mean or shape of the distribution?

Per-instance success and failure distributions are not distinguishable by shape or position



The outlier test is necessary

Meta-Recognition Error Trade-off Curves

	Conventional Explanation	**Prediction**	**Ground Truth**
Case 1	False Accept	Success	O
Case 2	False Reject	Failure	O
Case 3	True Accept	Success	P
Case 4	True Reject	Failure	P

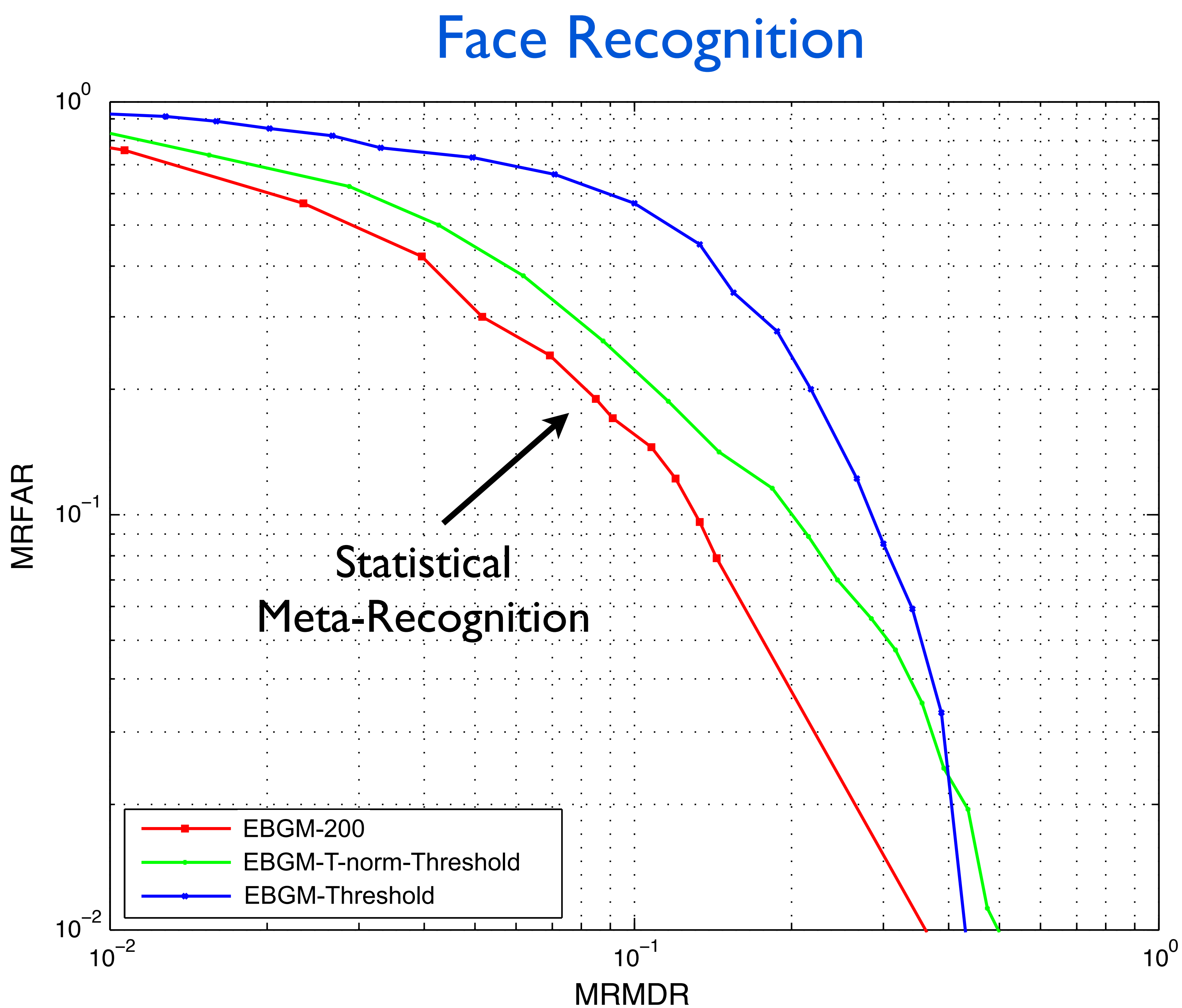
Meta-Recognition False Alarm Rate

$$\text{MRFAR} = \frac{|\text{Case 1}|}{|\text{Case 1}| + |\text{Case 4}|}$$

Meta-Recognition Miss Detection Rate

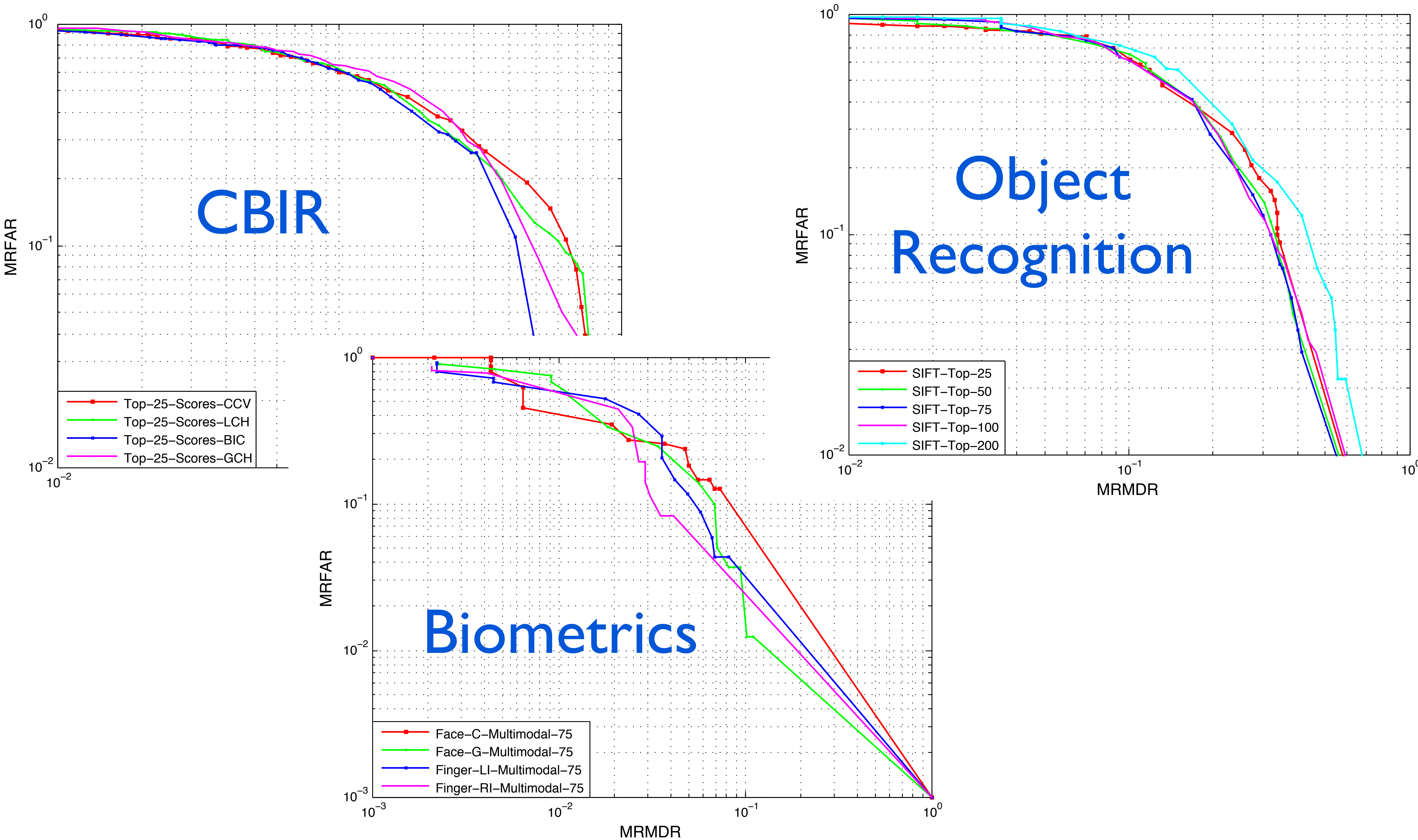
$$\text{MRFAR} = \frac{|\text{Case 2}|}{|\text{Case 2}| + |\text{Case 3}|}$$

Comparison with Basic Thresholding over Original and T-norm Scores



Points approaching the lower left corner minimize both errors

And meta-recognition works across all algorithms tested...

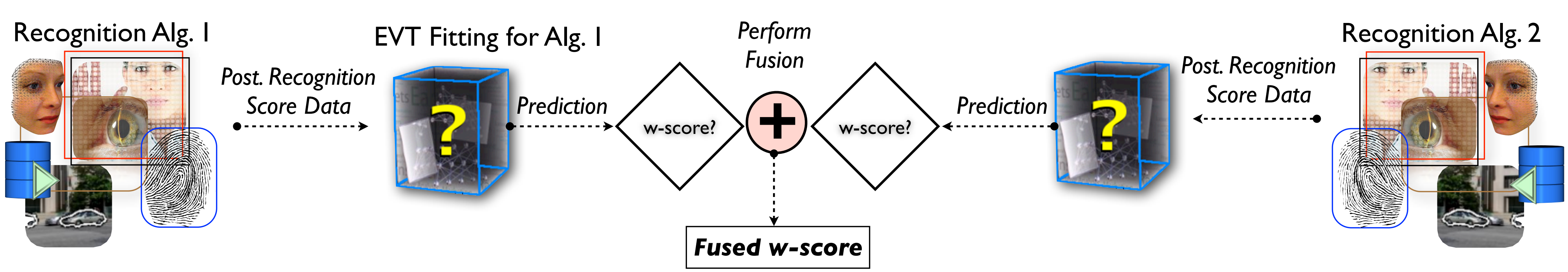


We can do score level fusion too...

Use the CDF of the Weibull model for score normalization:

$$\mathrm{CDF}(x) = 1 - e^{-(x/\lambda)^k}$$

We call this a *w-score*



w-score normalization

Require: a collection of scores S, of vector length m, from a single recognition algorithm j;

1. **Sort** and retain the n largest scores, $s_1, \ldots, s_n \in S$;

2. **Fit** a Weibull distribution W_S to $s_2, \ldots, s_n$, skipping the hypothesized outlier;

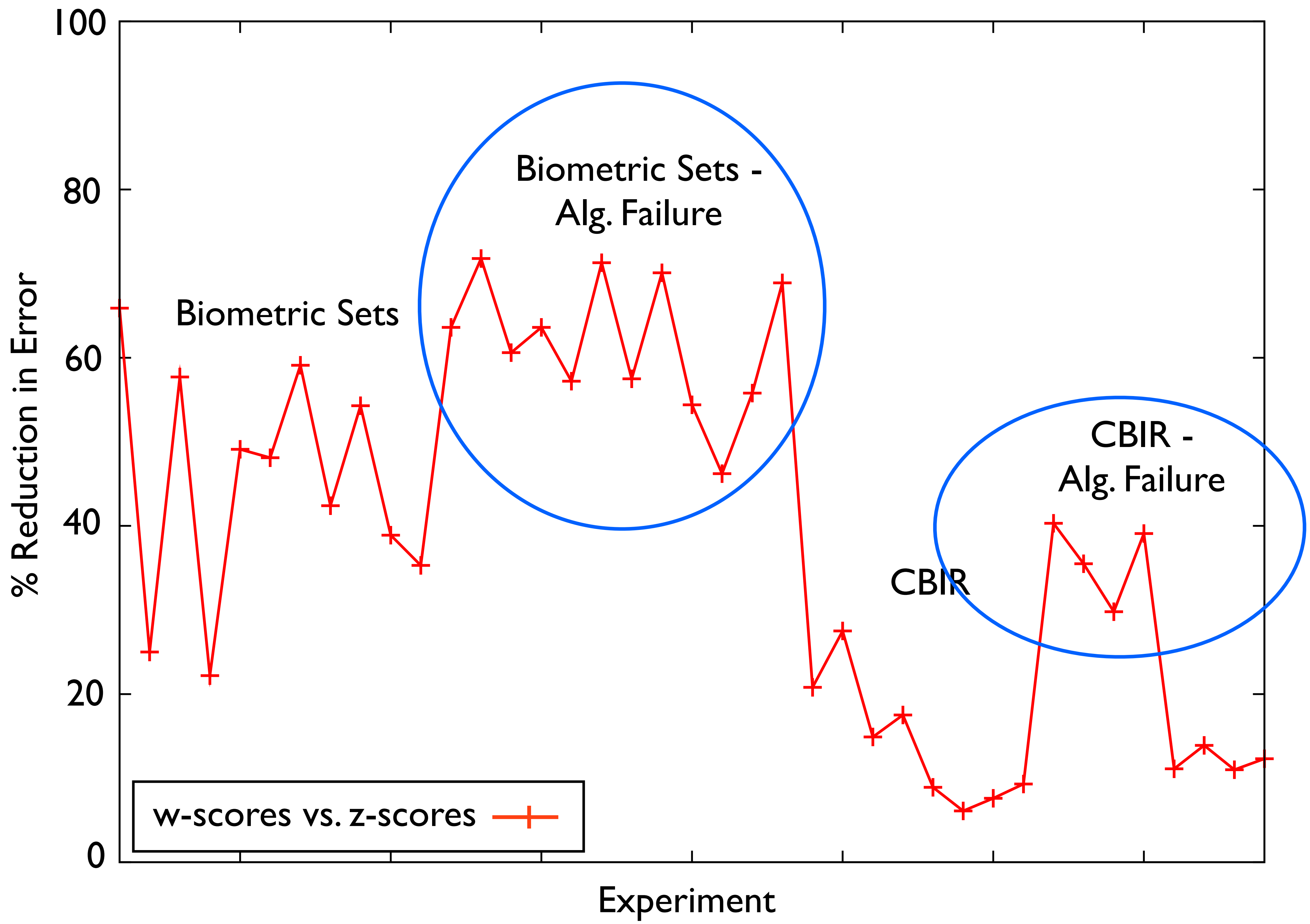
3. **While** $k < m$ **do**

4. $\quad s'_k = \mathrm{CDF}(s_k, W_S)$

5. $\quad k = k + 1$

6. **end while**

Error Reduction: Failing vs. Succeeding Algorithm



Let's take a step back and consider machine learning for recognition...

- Large-scale learning is a major recent innovation in computer vision
  - Feed lots of features to a learning algorithm, and let it find correlation
- How should we approach the multi-class problem[1] for general object recognition?
  - Is it a series of binary classifications?
  - Should it be performed by detection?
  - What if the classes are ill-sampled, not sampled at all, or undefined?

1. R. P. Duin and E. Pekalska, "Open Issues in Pattern Recognition," in *Computer Recognition Systemsn,* M. Kurzynski, E. Puchala, M. Wozniak, and A. Zolnierek, Eds. Springer, 2005

Closed Set Recognition

- How well are we really doing on recognition tasks?
- The problem we'd like to solve: scene understanding given an image never seen before
- The problem data sets solve: given a set of known classes, and corresponding '+' and '-' labels, distinguish between these classes
  - Caltech 101 & 256
  - LabelMe
  - ImageNet
- Training and Testing on the same data[1]

1. A. Torralba and A. A. Efros, "Unbiased Look at Dataset Bias," IEEE CVPR, 2011

Closed Set Recognition

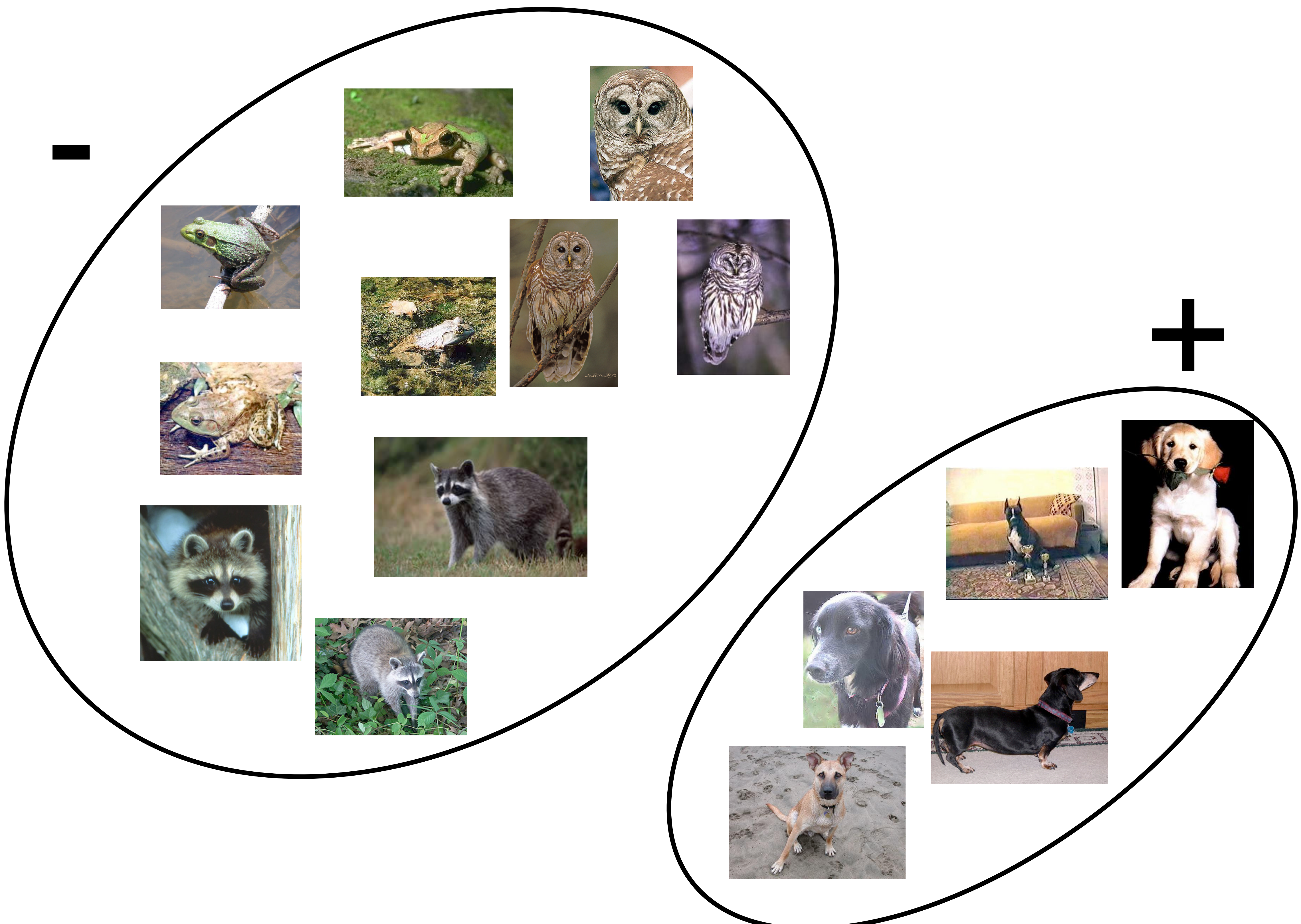


Image credit: Caltech-256, http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Open Set Recognition

- There are classes not seen in training that occur in testing
- Suppose the "other" classes are known
  - we generally cannot have enough positive samples to balance the negative samples

"All positive examples are alike; each negative example is negative in its own way[1]"

1. X. Zhou and T. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap," IEEE CVPR, 2001

Open Set Recognition

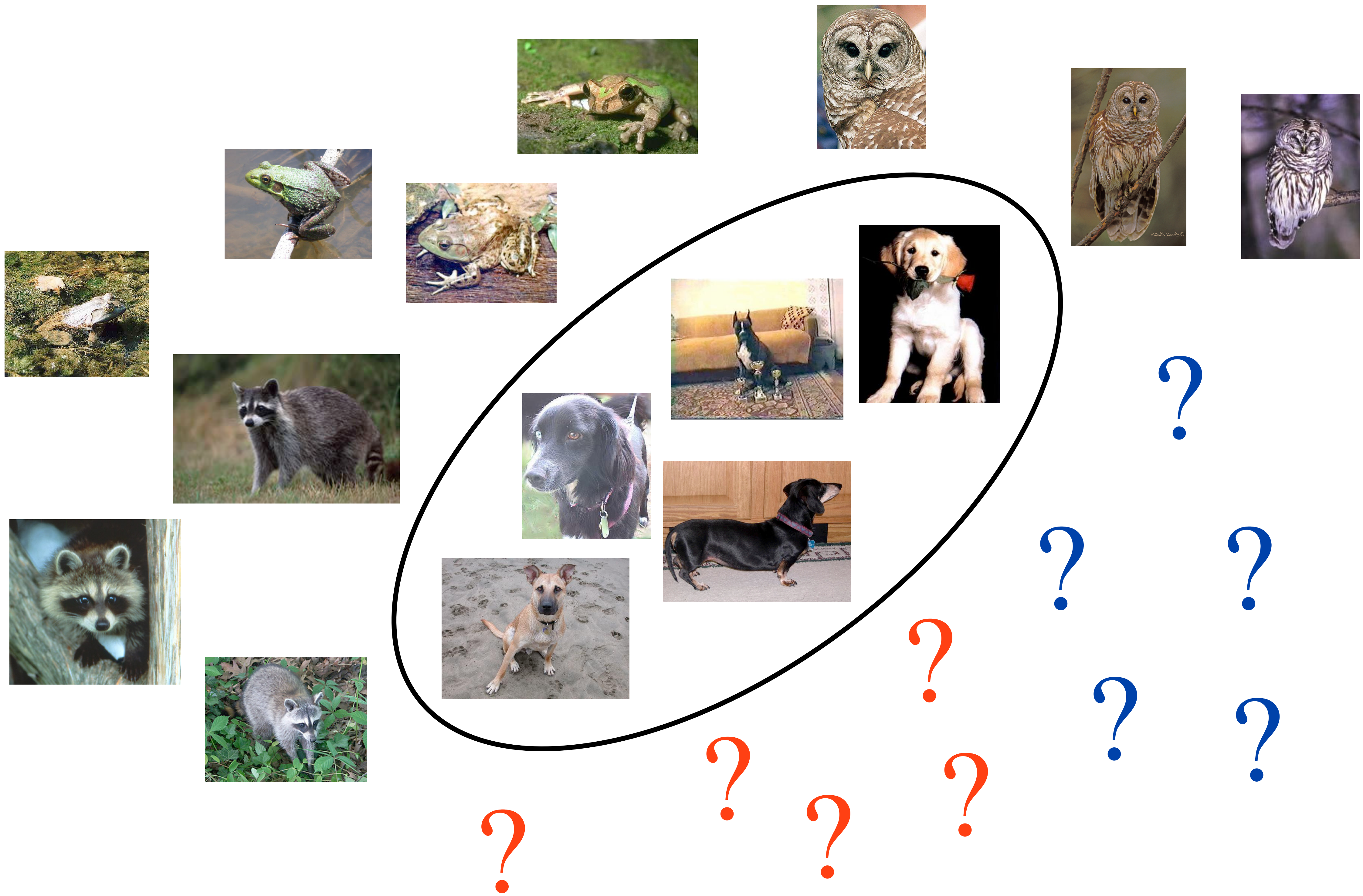


Image credit: Caltech-256, http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Formalization of Open Set Recognition Problem

- A class is a distribution $\mathcal{P}$
- A sample V is labeled $L = +1$ if it belongs to the class to be recognized and $L = -1$ for any other class
- Training samples from $\mathcal{P}$: $\hat{V} = \{v_1, \dots, v_m\}$
- Training samples from other known classes $\mathcal{K}$: $\hat{K} = \{k_1, \dots, k_n\}$
- The larger universe of unknown negative classes $\mathcal{U}$
- Test data: $\{t_1, \dots, t_z\}$, $t_i \in \mathcal{P} \cup \mathcal{K} \cup \mathcal{U}$
- A measurable recognition function f for a class $\mathcal{P}$

Recognition Risk: $R(f) = \mathbb{E}(sign(f(V)) \neq L)$

Formalization of Open Set Recognition Problem

- A few notes on Risk
  - Ensure that the risk of a false positive (over generalization) is proportional to the volume of space which is labeled positive
  - Ensure that over specialization occurs if we define the region too narrowly around the training data
  - Good solutions to the open set recognition problem require minimizing the volume of space representing the learned recognition function f
    - *Outside the support of positive samples*

Formalization of Open Set Recognition Problem

- We also need to optimize a data error measure:

$$\mathcal{D}(f(v_i); f(k_j)); \ (v_i \in \hat{V}, k_j \in \hat{K})$$

$\mathcal{D}$ could be: inverse F-measure over the training data, inverse training precision for a fixed training recall, inverse training recall for a fixed training precision...

Goal: balance the risk with the data error measure, all while being subject to hard constraints from the positive training data and/or negative training data

Formalization of Open Set Recognition Problem

The Open Set Recognition Problem

Using training data with positive samples, and other known class samples, and a data error measure, find a measurable recognition function f, where f(x) > 0 implies positive recognition, and f is defined by:

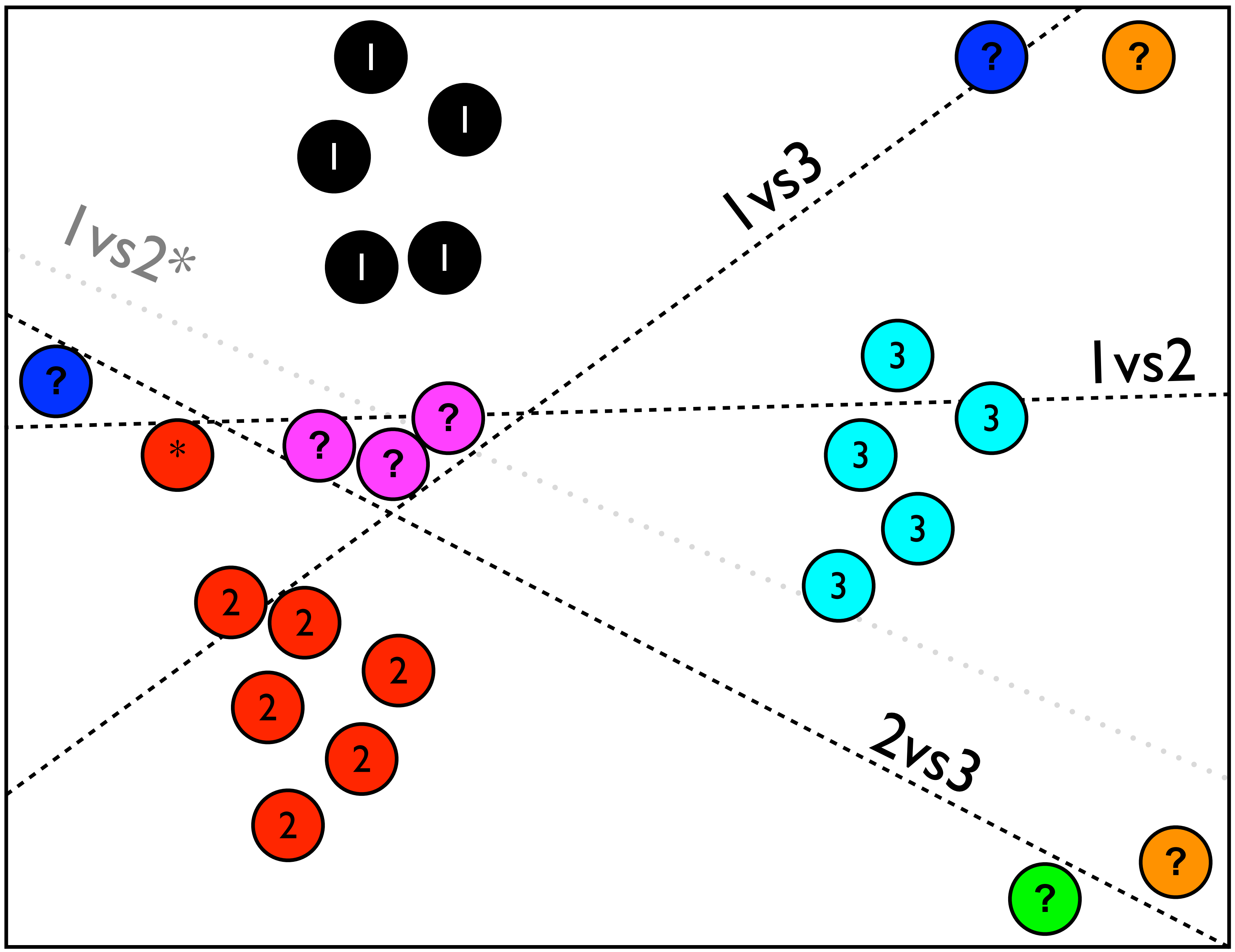
$$\operatorname*{argmin}_{f} \{R(f) + \lambda_r \mathcal{D}(f(v_i); f(k_j))\}$$

subject to

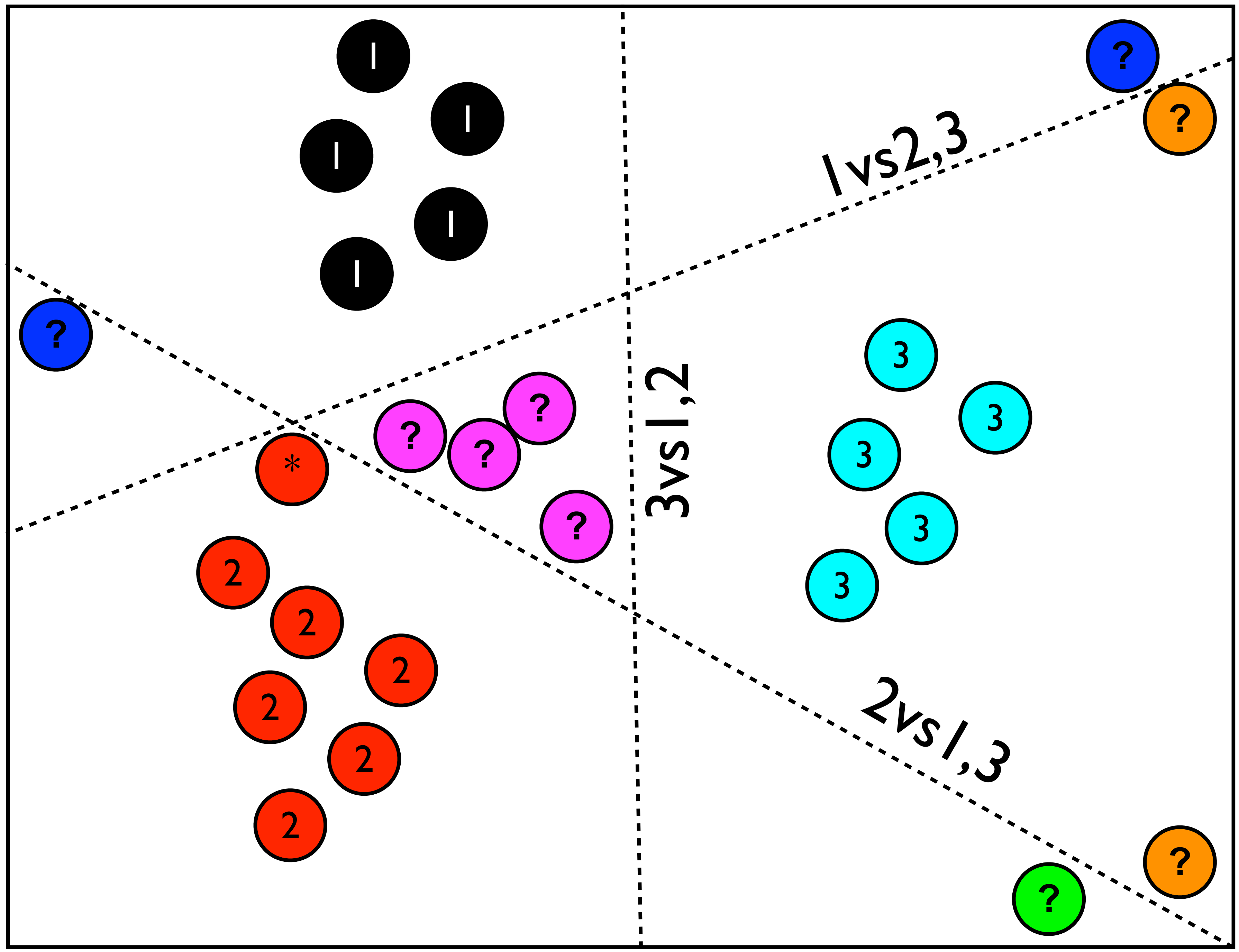
$$m\alpha \ \leq \sum_{i=1}^{m} \phi(f(v_i)) \quad \text{and} \quad n\beta \geq \sum_{j=1}^{n} \phi(f(k_j))$$

where λ_r specifies the regularization tradeoff between risk and data, where $\alpha \geq 0$ and *$\beta \geq 0$ allow a prescribed limit on true positive and/or false positive rates, and Φ is a given loss function.*

The trouble with binary classification



The trouble with 1-vs-All classification



One Solution: 1-class SVM

- Formulation by Schölkopf et al.[1]
  - Origin defined by the kernel function serves as the only member of a "second class"
  - Find the best margin with respect to the origin
  - The resulting function f takes the values
    - +1 in a region capturing most of the training data points
    - -1 elsewhere

1. B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the Support of a High-dimensional Distribution," Microsoft Research, Tech. Rep. MSR-TR-99-87, 1999

One Solution: 1-class SVM

To separate the training data from the origin, the algorithm solves the following quadratic programming problem for w and ρ to learn f:

$$min \frac{1}{2} \| w \|^2 + \frac{1}{\nu m} \sum_{i=1}^{l} \xi_i - \rho$$

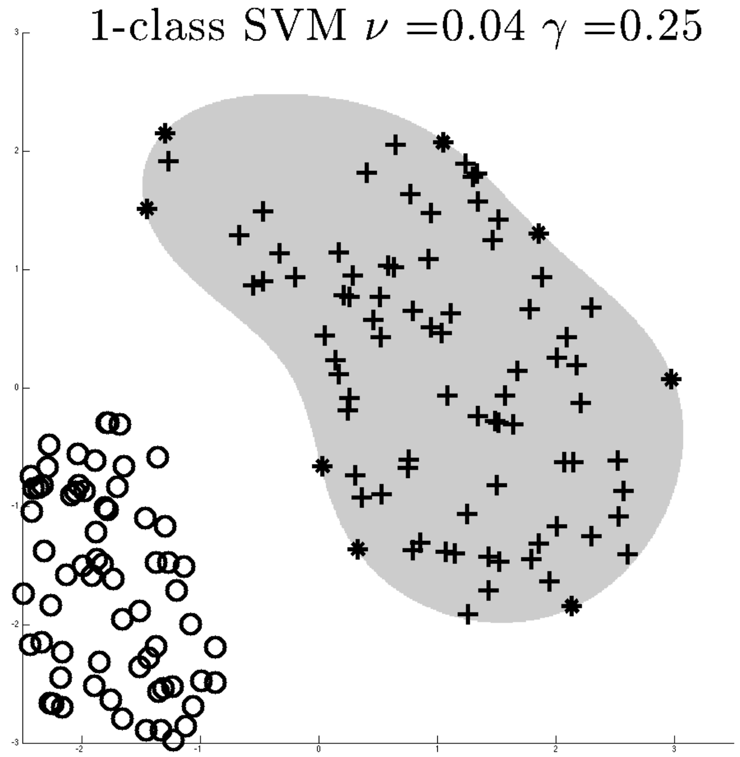
subject to

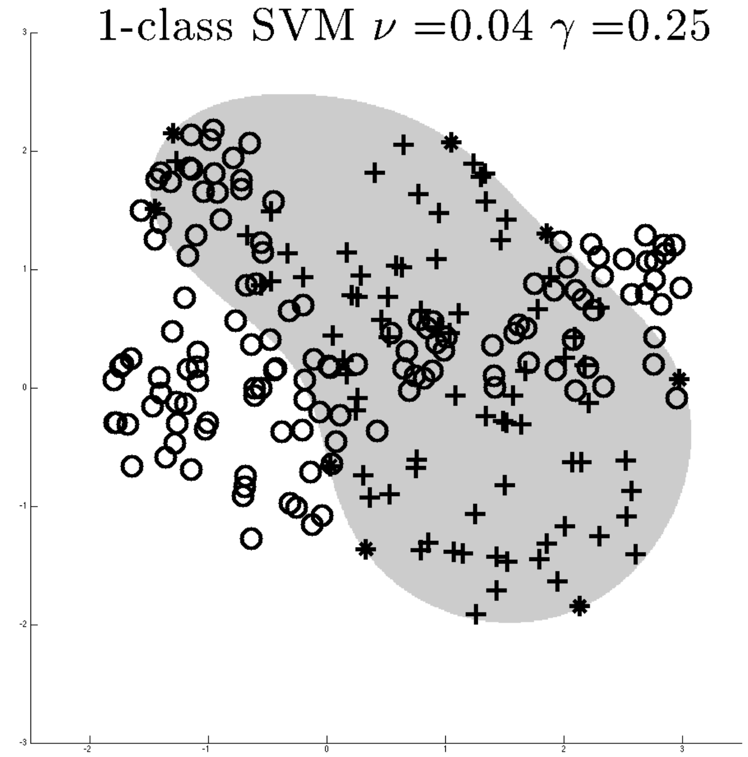
$$(w \cdot \Psi(x_i)) \geq \rho - \xi_i \quad i = 1, 2, \ldots, m \quad \xi_i \geq 0$$

The kernel function Ψ impacts density estimation and smoothness. The regularization parameter $v \in (0, 1]$ controls the trade-off between training classification accuracy and and the smoothness term $\| w \|$, and also impacts the number of support vectors.

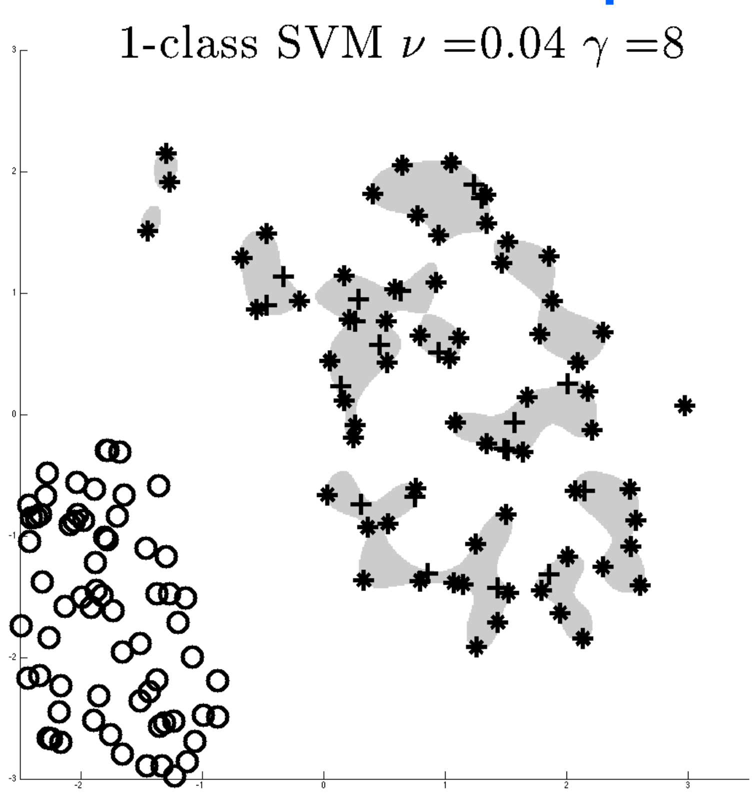
1-Class SVM

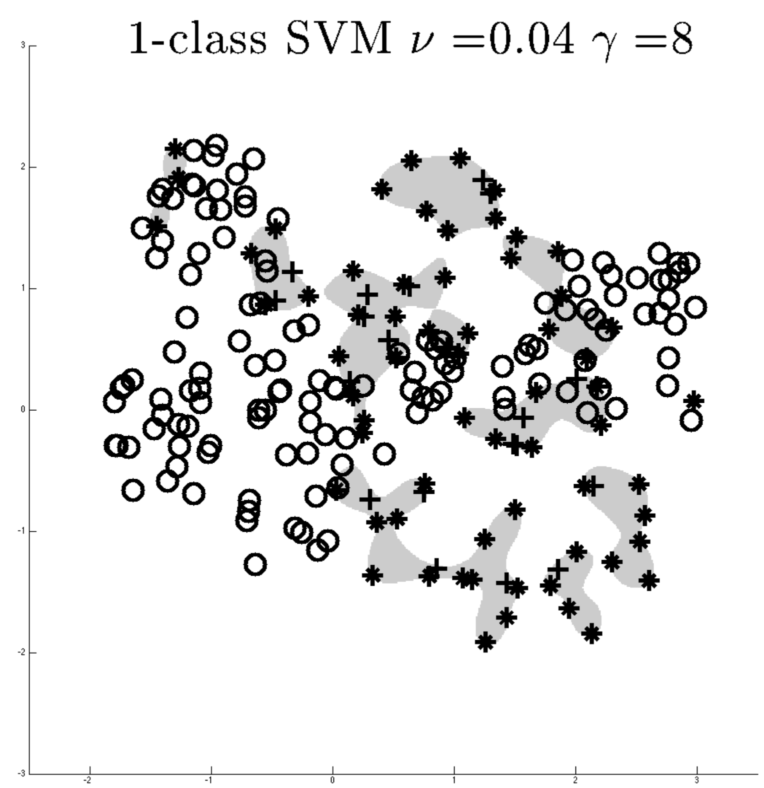
Generalization



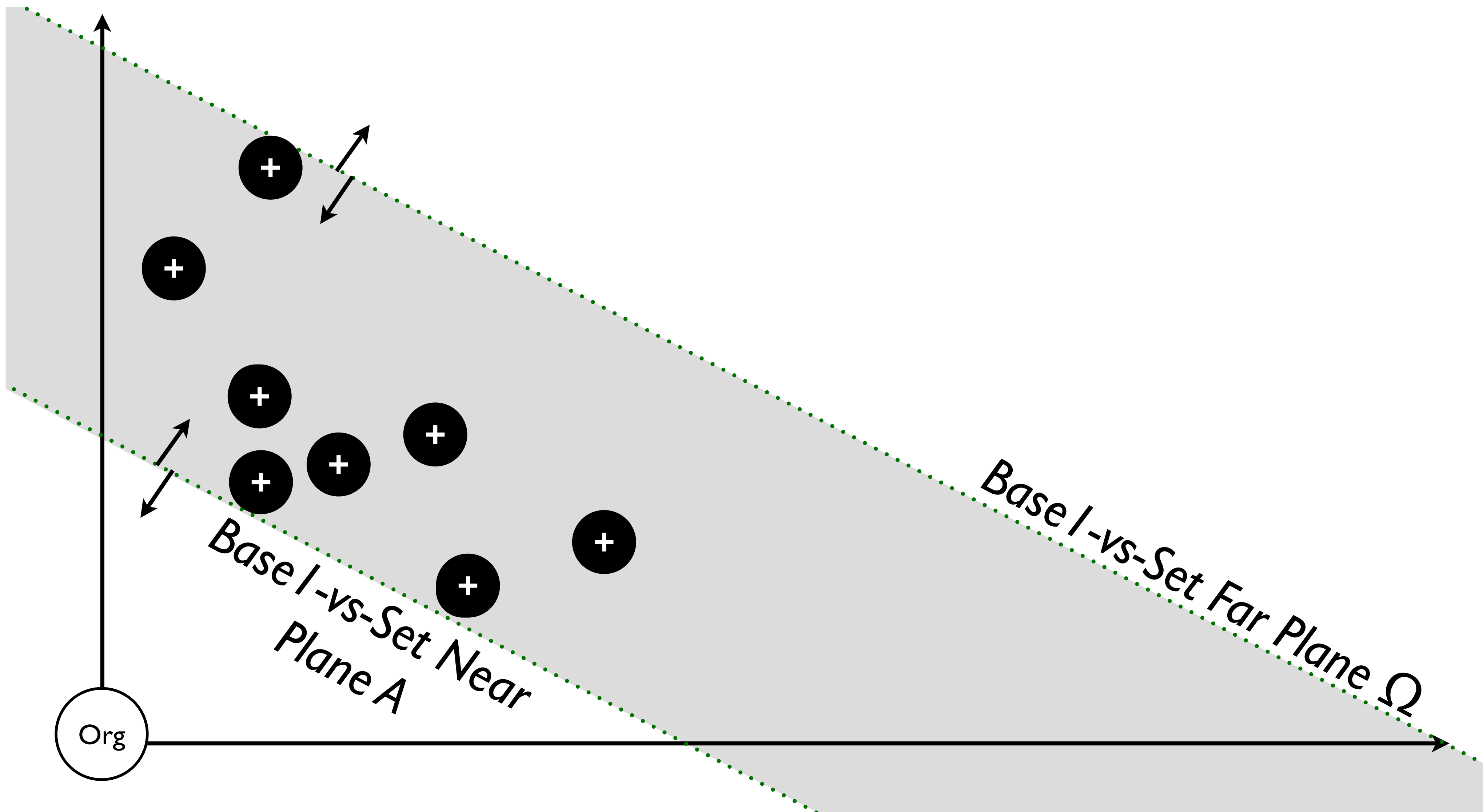


Specialization

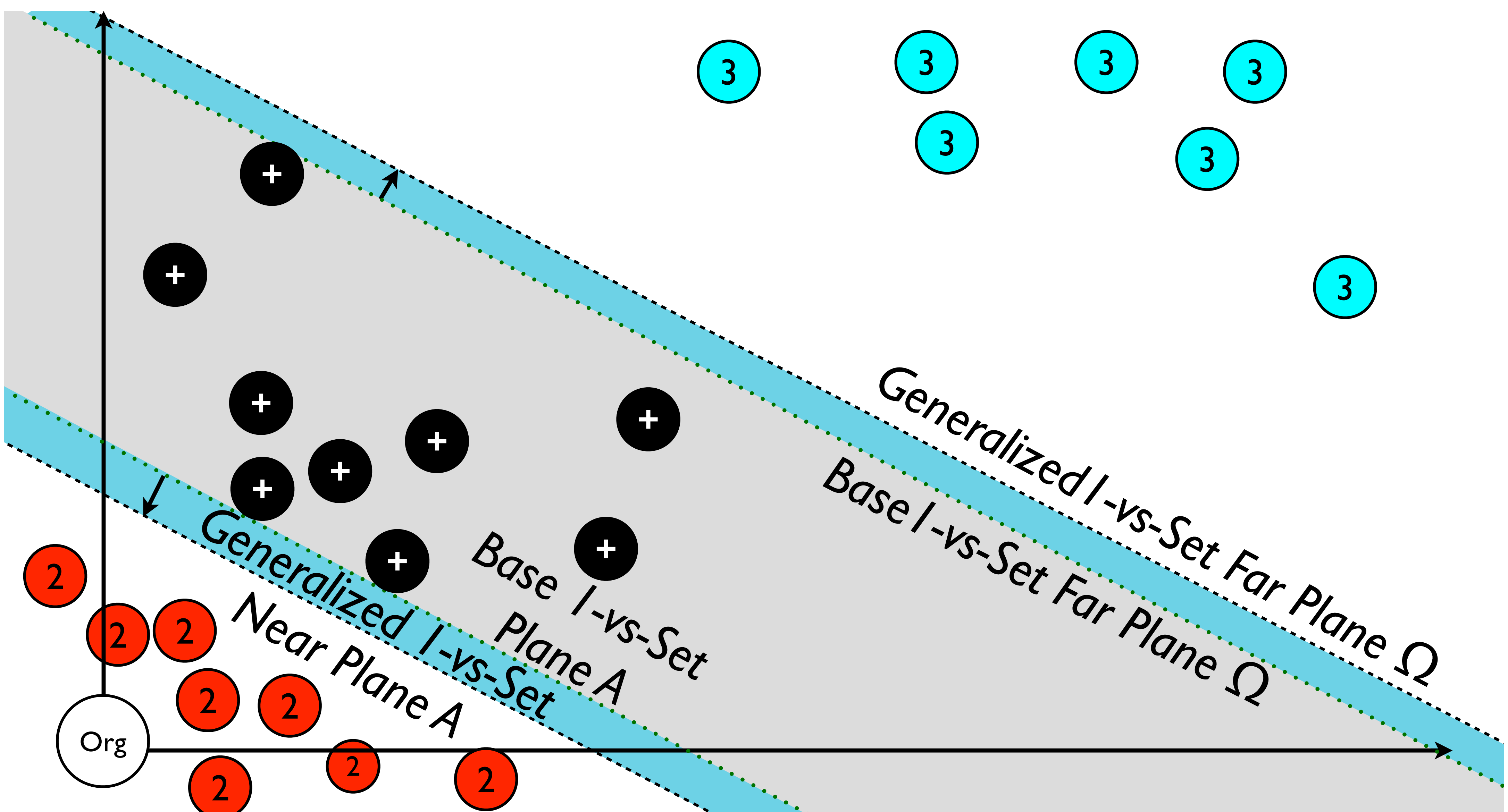




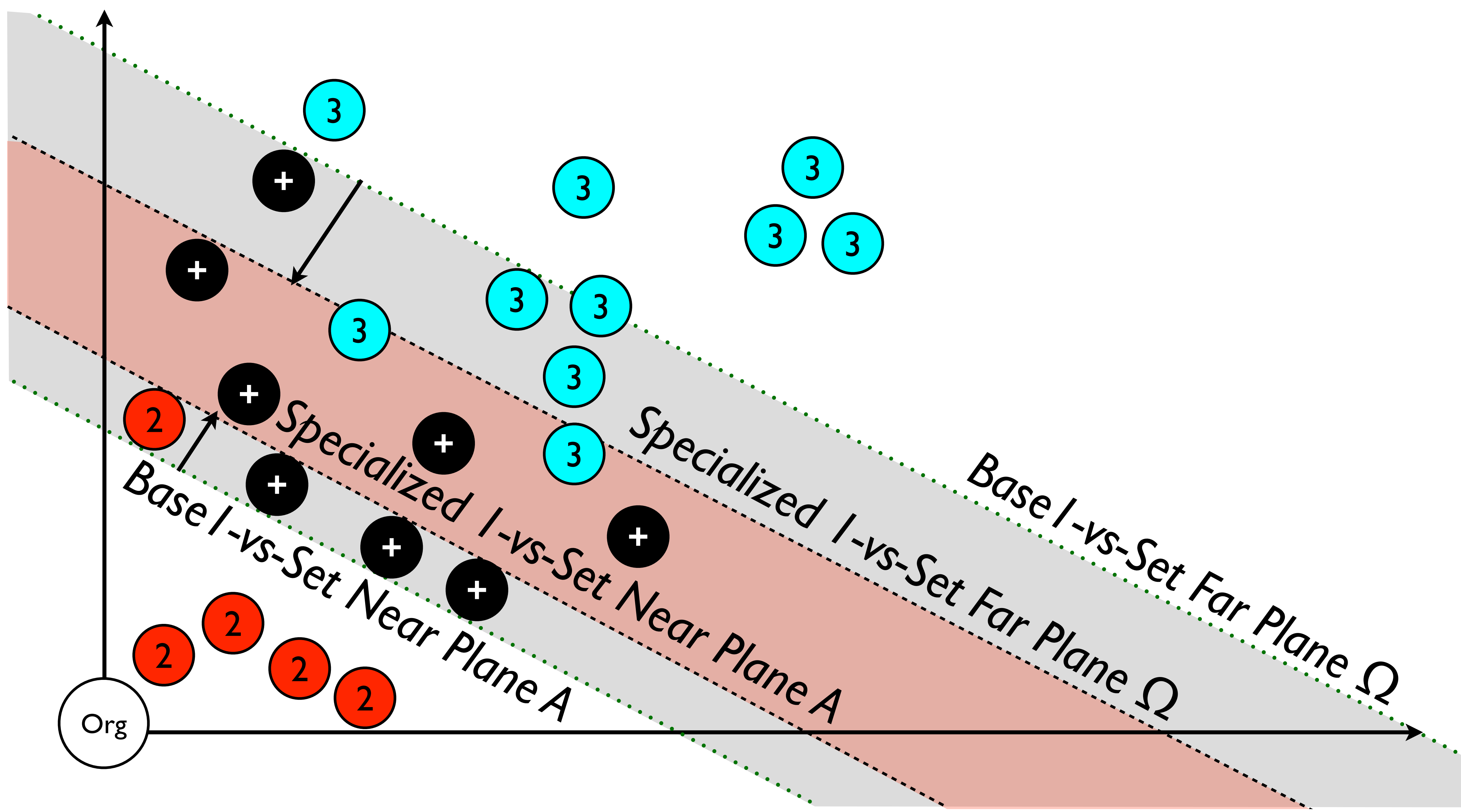
Start with a 1-class SVM



Generalization



Specialization



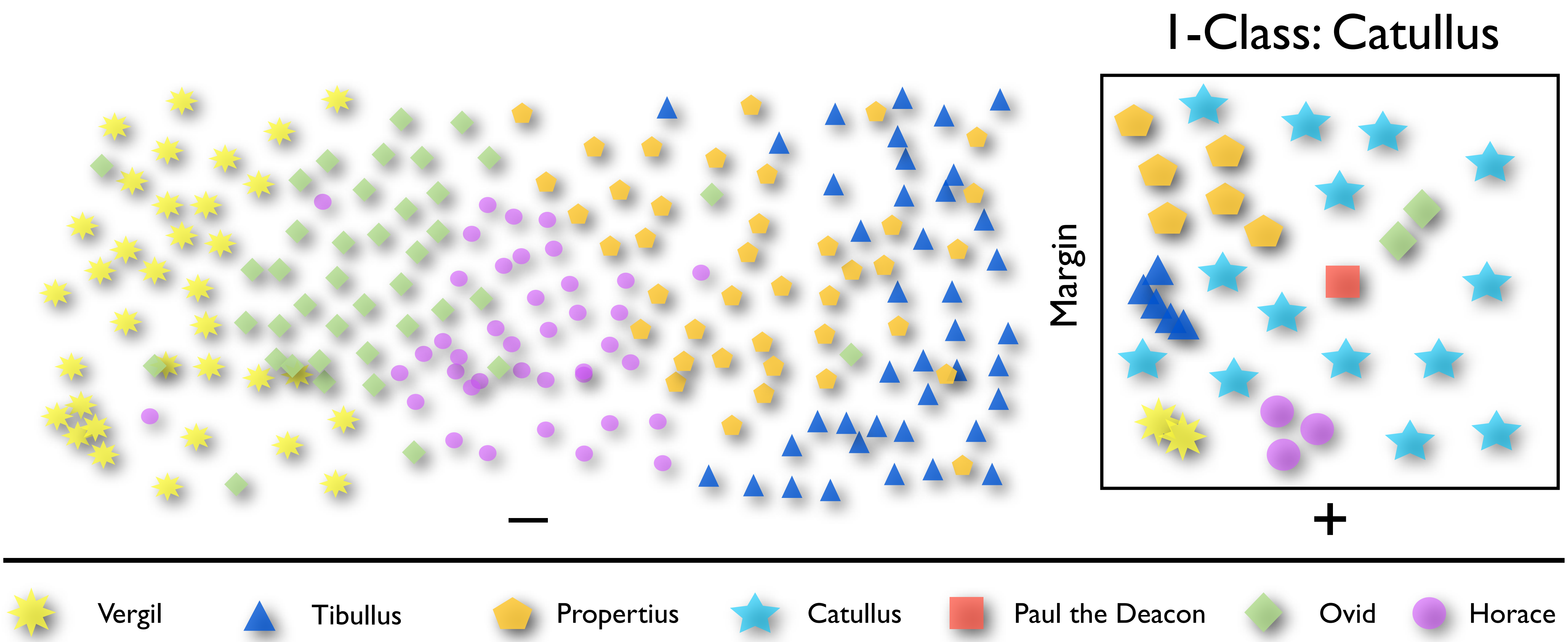
Where is this work heading?

- The 1-vs-Set Machine as an initial solution for open set recognition
- New classes of learning algorithms to specifically address the open set problem
- Application Area: Computational Linguistics
  - The recognition problem occurs here too

Taking literary theory into practice!

Open Set Intertextuality

Sometimes confusion is a good thing[1]...



1. C. Forstall, S. Jacobson and W. Scheirer, "Evidence of Intertextuality: Investigating Paul the Deacon's *Angustae Vitae,*" Literary & Linguistic Computing, 2011

Questions?