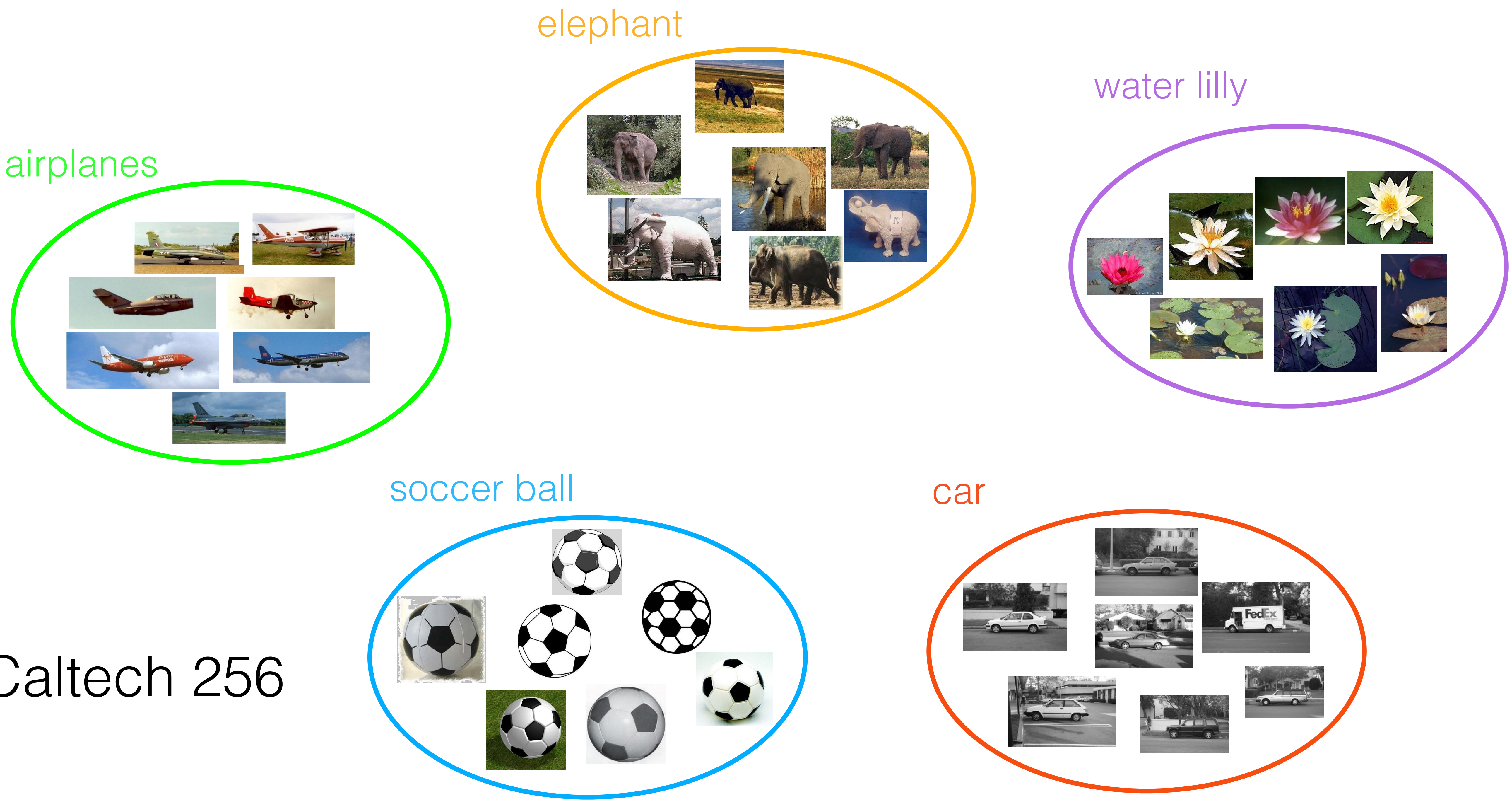
Emerging Work in Open Set Recognition for Vision and Language

Walter J. Scheirer
School of Engineering and Applied Sciences,
Department of Molecular and Cellular Biology, and
Center for Brain Science, Harvard University

Part 1: The Open Set Recognition Problem

Benchmarks in computer vision

Assume we have examples from all classes:



Caltech 256

Out in the real world…

Detect the cars in this image

while rejecting the trees, signs, telephone poles…

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

Zhao and Huang (with some help from Tolstoy)
CVPR 2001

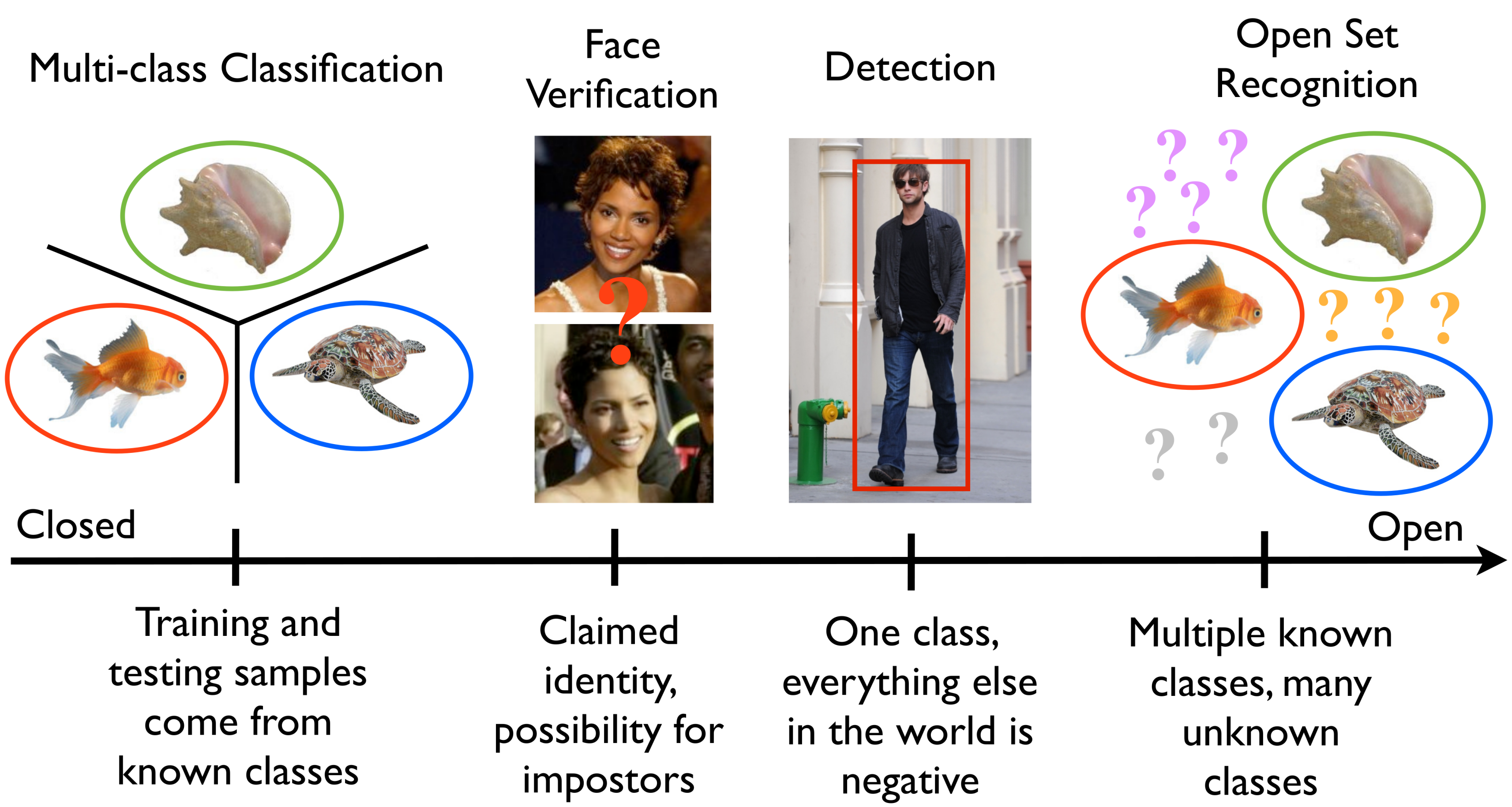
X. Zhou and T. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap," in IEEE CVPR, 2001.

What is the general recognition problem?

- Duin and Pekalska*: how one should approach multi-class recognition is still an open issue
  - Is it a series of binary classifications?
  - Is it a search performed for each possible class?
  - What happens when some classes are ill-sampled, not sampled at all or undefined?

R. P. Duin and E. Pekalska, “Open Issues in Pattern Recognition,” in *Computer Recognition Systems*, M. Kurzynski, E. Puchala, M. Wozniak, and A. Zolnierek, Eds. Springer, 2005, pp. 27–42.

Vision problems in order of "openness"



W.J. Scheirer, A. Rocha, A. Sapkota, and T. Boult, "Towards Open Set Recognition," IEEE T-PAMI, 35(7) July 2013.

Let's formalize openness

$$\text{openness} = 1 - \sqrt{\frac{2 \times |\text{training classes}|}{|\text{testing classes}| + |\text{target classes}|}}$$

Examples of openness values

	Targets	Training	Testing	Openness
Typical Multi-class	x	x	x	0%
Face Verification	12	12	50	38%
Typical Detection	1	100,000	1,000,000	55%
Object Recognition	88	12	88	63%
Object Recognition	88	6	88	74%
Object Recognition	212	6	212	83%

Fundamental multi-class recognition problem

$$\underset{f}{\operatorname{argmin}} \left\{ \underbrace{R_{\mathcal{I}}(f)}_{\text{Ideal Risk}} := \int_{\mathbb{R}^d \times \mathbb{N}} \underbrace{L(x, y, f(x))}_{\text{Loss Function}} \underbrace{P(x, y)}_{\text{Joint Distribution}} \right\}$$

Ideal Risk Loss Function Joint Distribution

Undefined for
open set recognition!

A. Smola, "Learning with Kernels," Ph.D. dissertation, Technische Universität Berlin, Berlin, Germany, November 1998.

Open Space

Open Space

- Open space is the space far from known data
- We need to address the infinite half-space problem of linear classifiers
- Principle of Indifference*
  - If there is no known reason to assign probability, alternatives should be given equal probability
  - One problem: we need the distribution to integrate to 1!

J.M. Keynes, A Treatise on Probability. Macmillan & Company, Limited, 1921.

Open Space Risk

Open Space Risk: the relative measure of open space to the full space

open space

$$R_{\mathcal{O}}(f) = \frac{\int_{\mathcal{O}} f(x)dx}{\int_{S_o} f(x)dx}$$

Open space + positive training examples

The open set recognition problem

Preliminaries

Space of positive class data: $\mathcal{P}$

Space of other known class data: $\mathcal{K}$

Positive training data: $\hat{V} = \{v_1, \ldots, v_m\}$ from $\mathcal{P}$

Negative training data: $\hat{K} = \{k_1, \ldots, k_n\}$ from $\mathcal{K}$

Unknown negatives appearing in testing: $\mathcal{U}$

Testing data: $\mathcal{T} = \{t_1, \ldots, t_z\}, t_i \in \mathcal{P} \cup \mathcal{K} \cup \mathcal{U}$

Assume the problem openness is > 0

The open set recognition problem

Minimize open set risk:

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ R_{\mathcal{O}}(f) + \lambda_r R_{\mathcal{E}}(f(\hat{V} \cup \hat{K})) \right\}$$

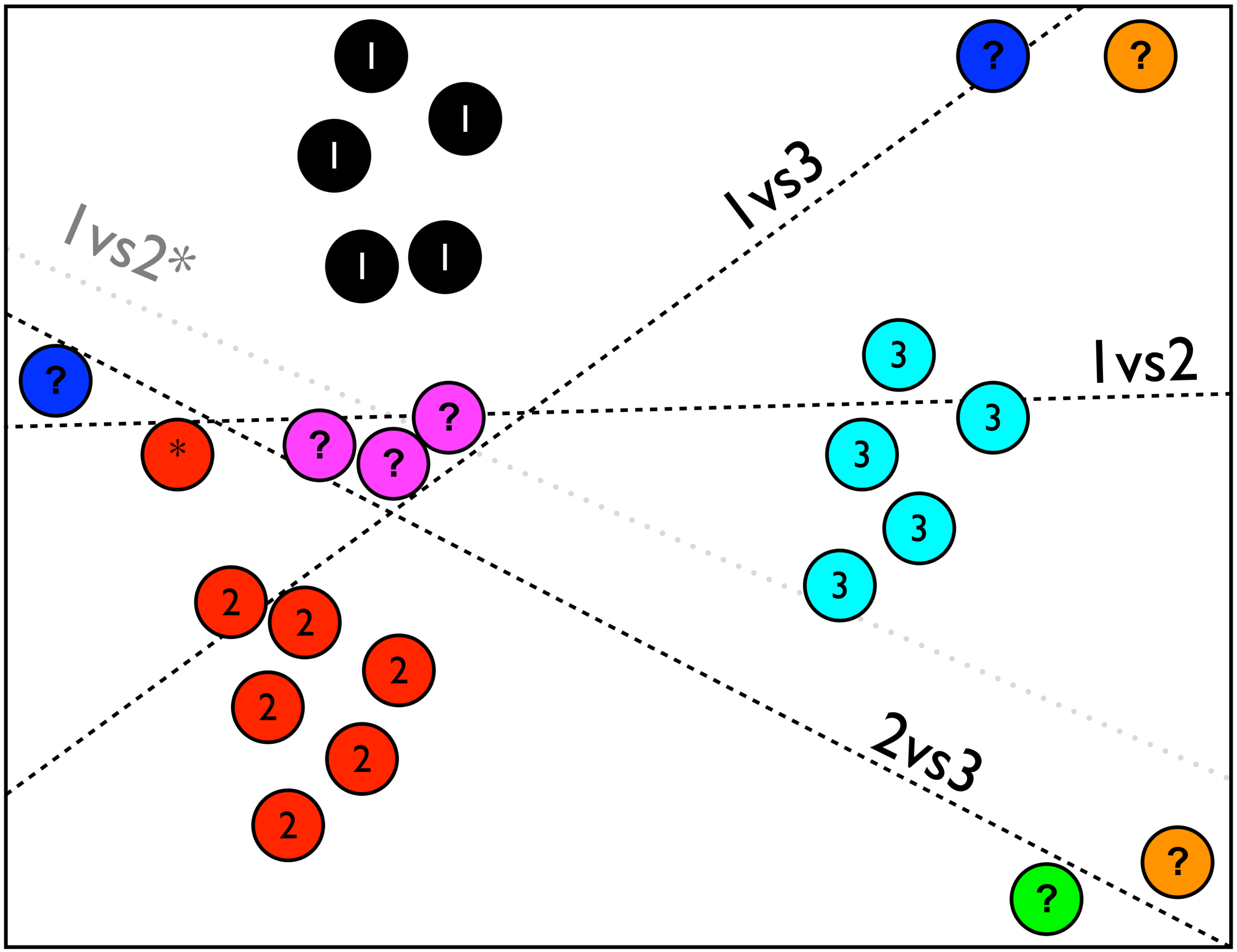
Open Space Risk Associated with $\mathcal{U}$

Regularization Constant

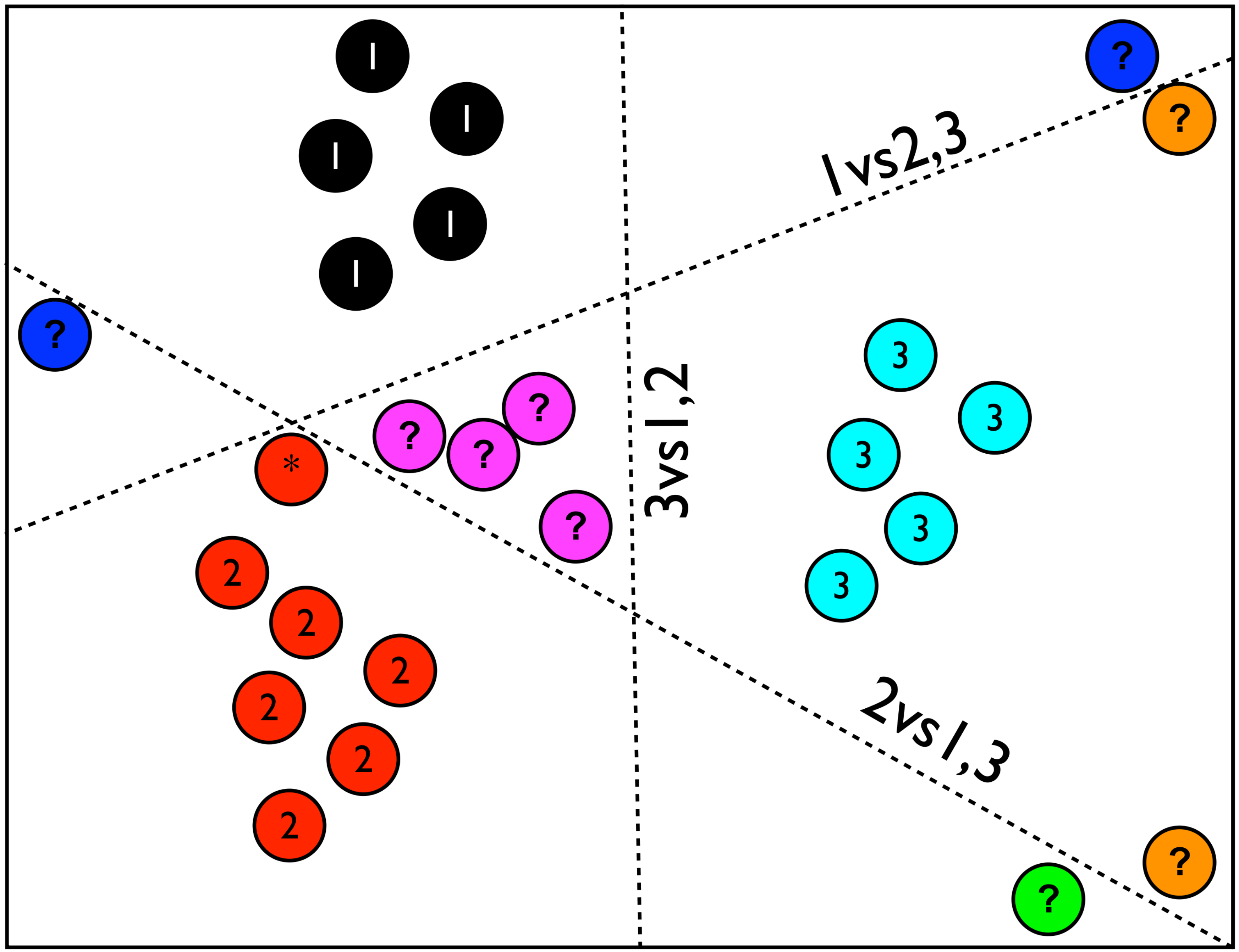
Empirical Risk Function

What options do we have to solve this problem?

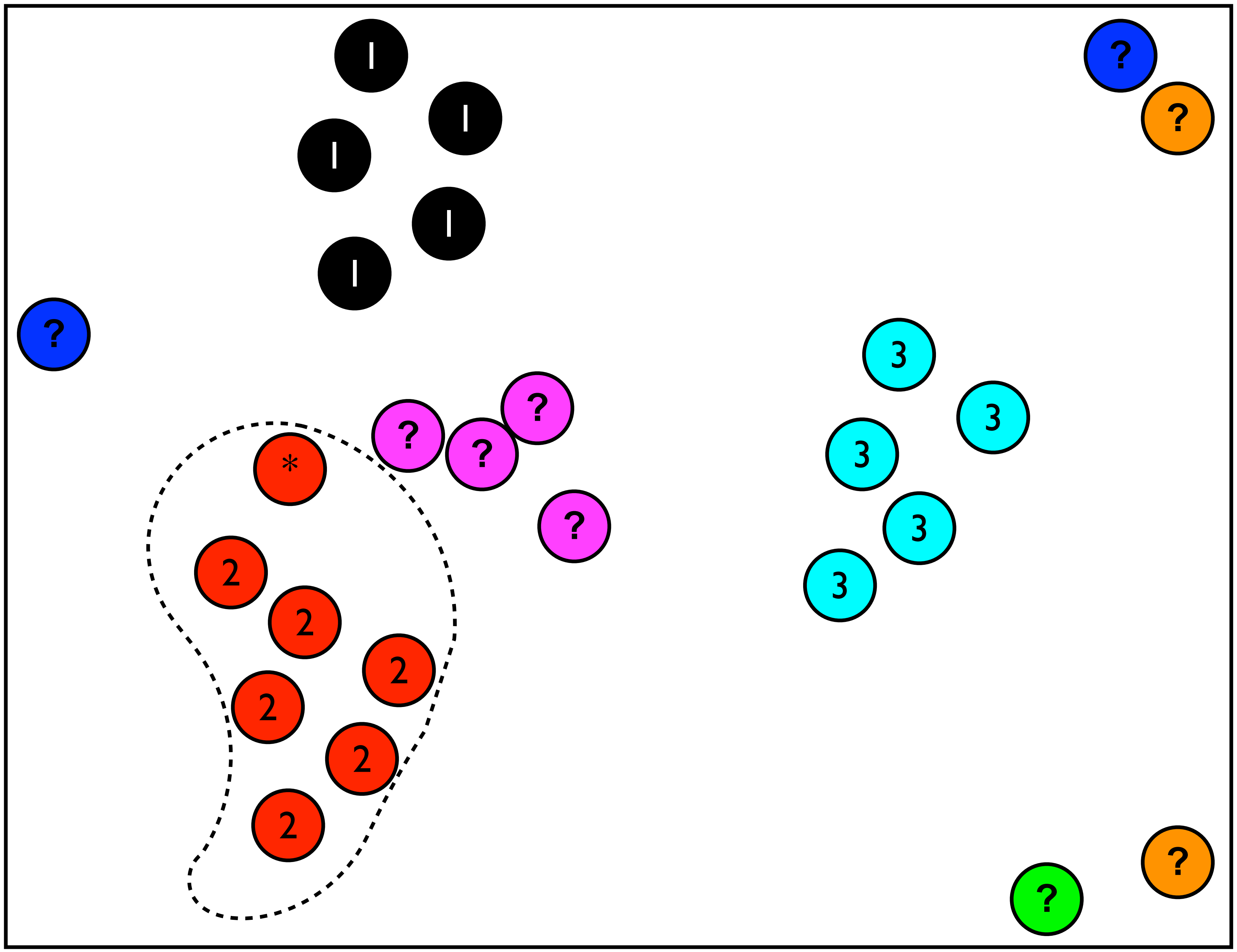
Binary Classification



Multi-class 1-vs-All Classification



1-class Classification



B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and “R. Williamson. Estimating the Support of a High-dimensional Distribution,” Technical report, Microsoft Research, 1999.

Why didn't the 1-class SVM catch on?

- Zhou and Huang *Multimedia Systems* 2003
  - Kernel and parameter selection
    - Gaussian kernels lead to over-fitting
    - Parameters chosen in *ad hoc* fashion
    - An issue in other domains too!

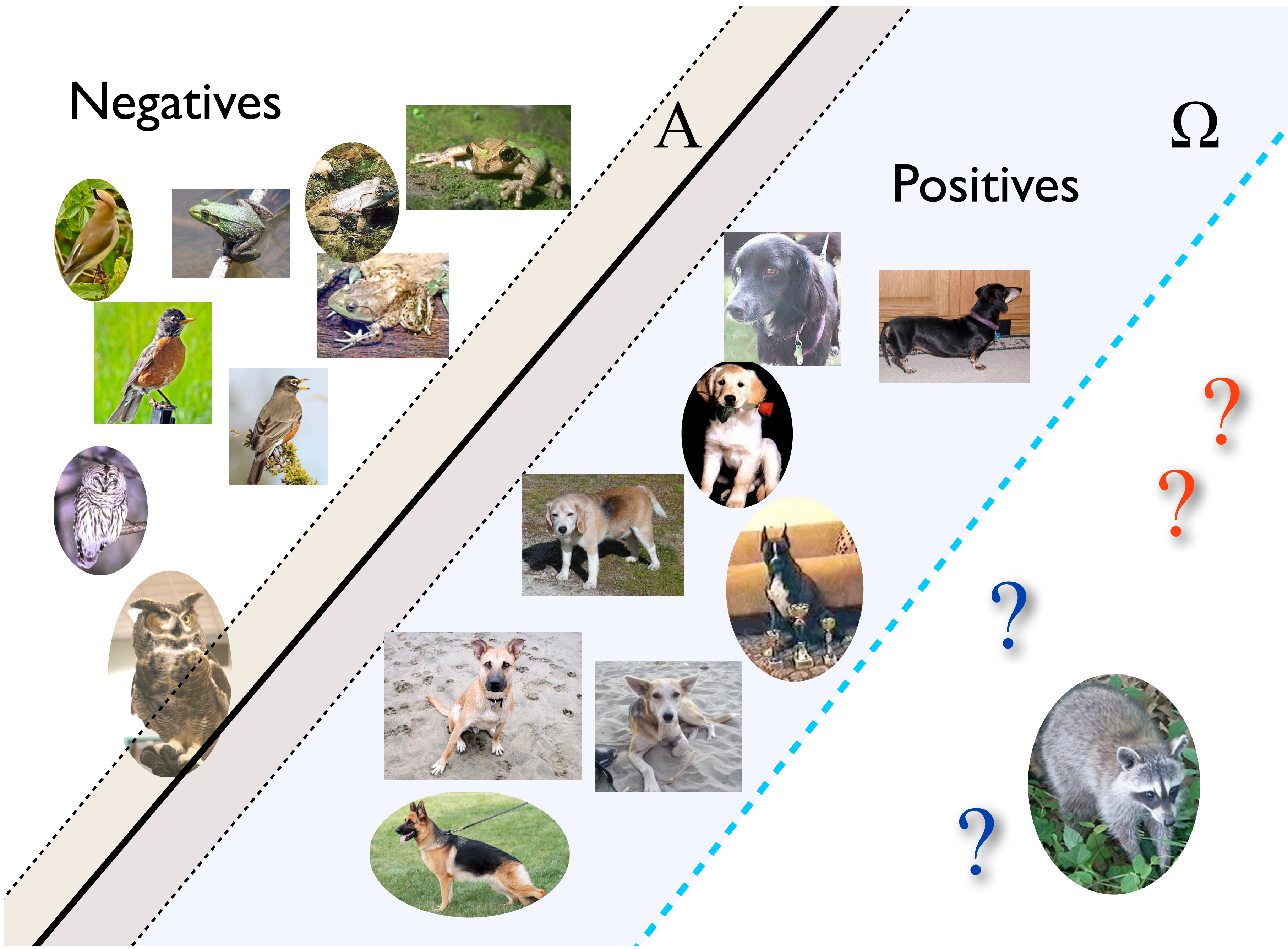
X. Zhou and T. Huang, "Relevance Feedback in Image Retrieval: A Comprehensive Review," *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

Other approaches

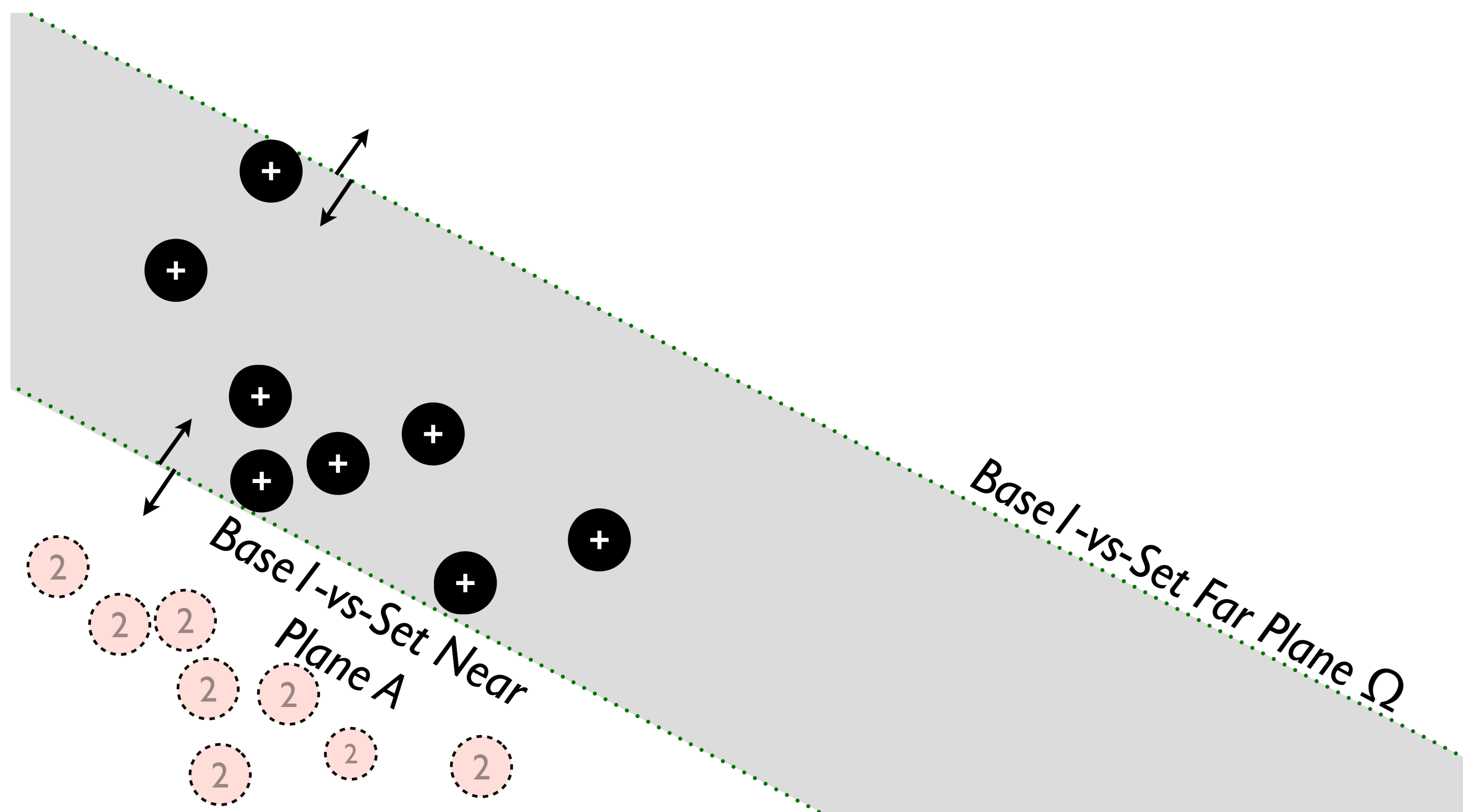
- M. Rohrbach, M. Stark, and B. Schiele, “Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting,” in IEEE CVPR, 2011.
- C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer,” in IEEE CVPR, 2009.
- E. Bart and S. Ullman, “Single-example Learning of Novel Classes Using Representation by Similarity,” BMVC, 2005.
- M. Palatucci, D. Pomerleau, G. Hinton, and T.M. Mitchell, “Zero-shot Learning with Semantic Output Codes,” NIPS, 2009.
- L. Wolf, T. Hassner, and Y. Taigman, “The One-shot Similarity Kernel,” ICCV 2009.
- G. Heidemann, “Unsupervised Image Categorization,” Image and Vision Computing, vol. 23, no. 10, pp. 861–876, October 2004.

Let’s include open space risk in our optimization problem

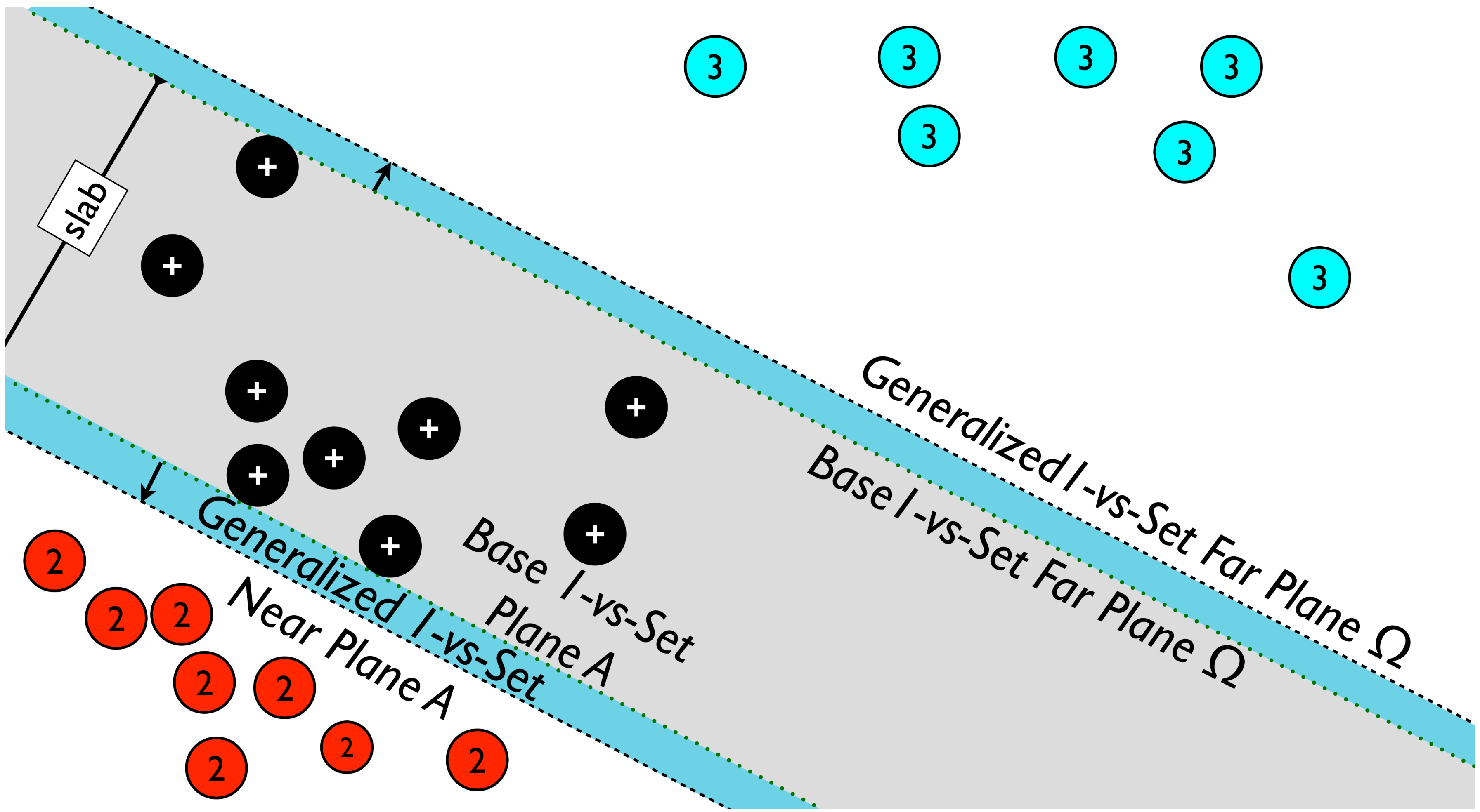
Slab Model



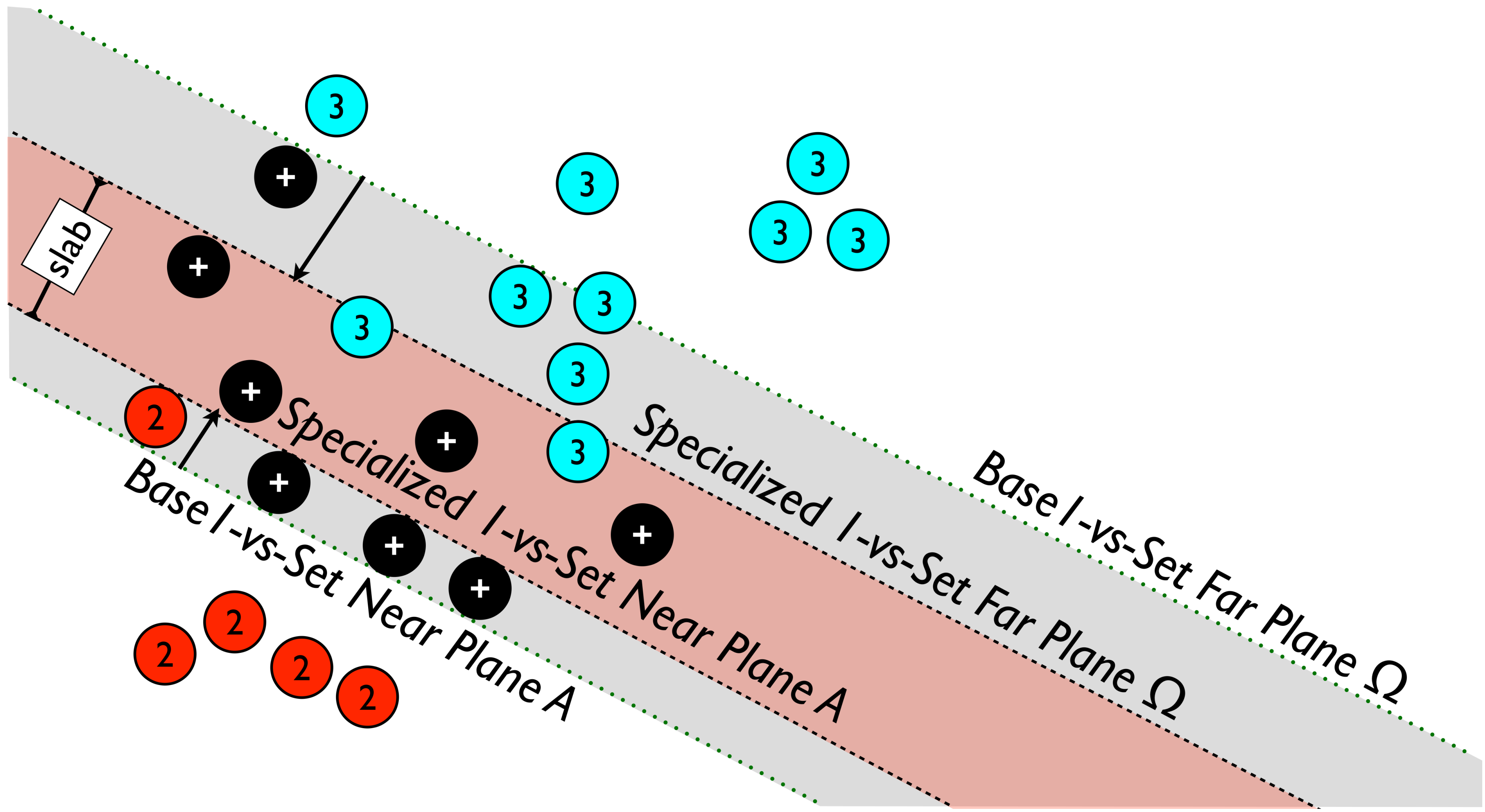
Base Linear 1-vs-Set Machine



Generalization



Specialization



Open space risk for linear slab model

δ_A Marginal distance of near plane

δ_Ω Marginal distance of far plane

δ^+ Separation needed to account for all positive data

$$\frac{\delta_\Omega - \delta_A}{\delta^+}$$

Overgeneralization risk

$$\frac{\delta^+}{\delta_\Omega - \delta_A}$$

Overspecialization risk

Open space risk for linear slab model

Two additional terms

Importance of open space around A

Importance of open space around Ω

$$R_\varsigma = \frac{\delta_\Omega - \delta_A}{\delta^+} + \frac{\delta^+}{\delta_\Omega - \delta_A} + p_A\omega_A + p_\Omega\omega_\Omega$$

Margin around A

Margin around Ω

Sketch of the 1-vs-Set Machine training algorithm

1. Train a linear SVM f using $\hat{V}$ and $\hat{K}$

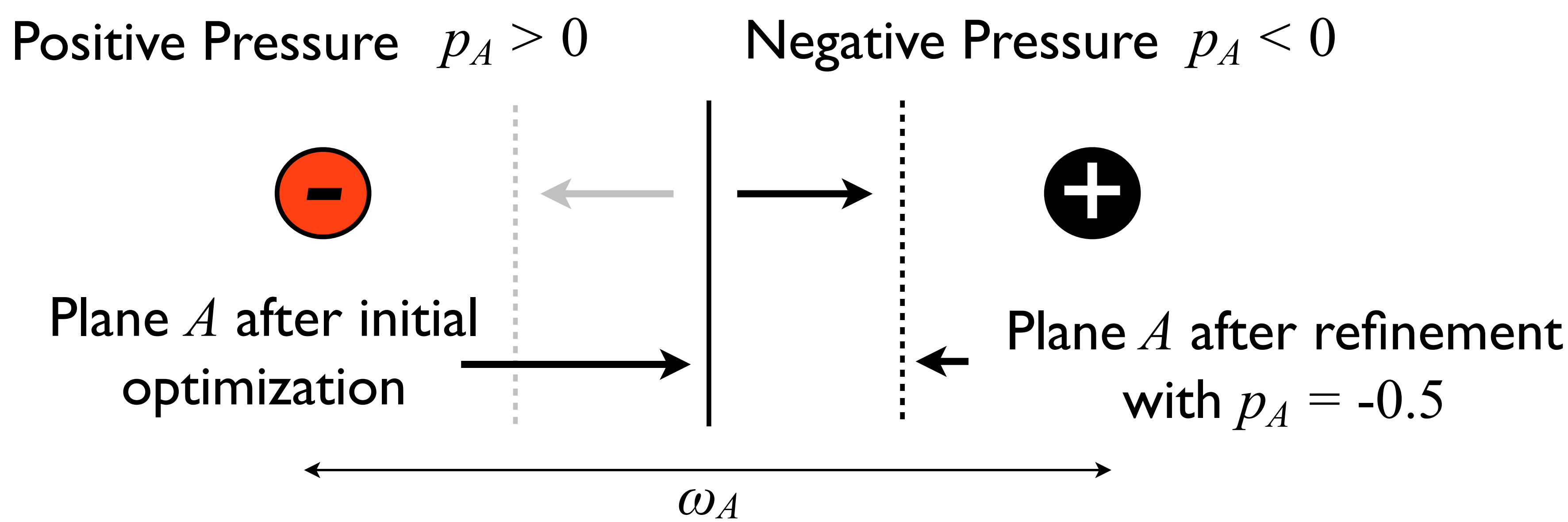
2. Generate decision scores for each training point in $\hat{V}$ and $\hat{K}$

3. Sort decision scores, where s_k is the minimum and s_j is the maximum

4. Initialize A to margin plane of f, and Ω to s_j

5. Iteratively move A to s_{k+1} or s_{k-1}, Ω to s_{j-1} or s_{j+1} to minimize $R_\varsigma(f) + \lambda_r R_\mathcal{E}$

1-vs-Set Machine Plane Refinement



1-vs-Set Machine Prediction

function PREDICT(t_x, f, A, Ω)
  if ($A \leq f(t_x)$ **and** $f(t_x) \leq \Omega$) **then Return** $+1$
  else Return -1
  end if
end function

How can we evaluate open set recognition in a controlled manner?

Open Set Object Recognition

Cross-data set methodology*

Training: Caltech 256

known classes

Testing: Caltech 256 + ImageNet

known classes +

unknown classes

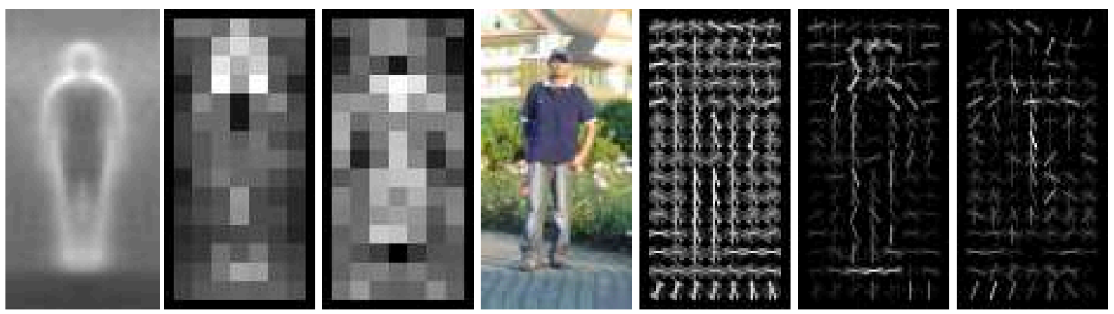
Open Universe of 88 classes: 1 positive class, n training classes, 87 negative testing classes (532,400 images)

Open Universe of 212 classes: 1 positive class, n training classes, 211 negative testing classes (13,610,400 images)

A. Torralba and A. A. Efros, “Unbiased Look at Dataset Bias,” in IEEE CVPR 2011.

Features

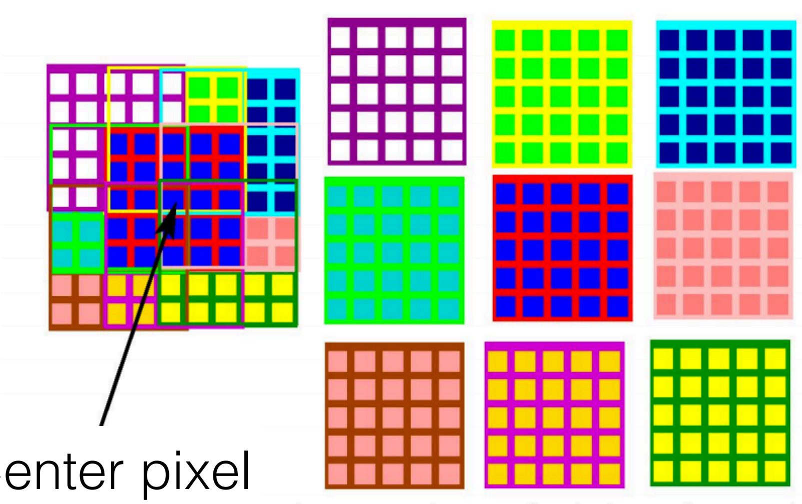
Histogram of Oriented Gradients



(Dalal and Triggs 2005) © 2005 IEEE

N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in IEEE CVPR, 2005

LBP-like descriptor



A. Sapkota, B. Parks, W.J. Scheirer, and T. Boult, “FACE-GRAB: Face Recognition with General Region Assigned to Binary Operator

1-vs-Set Machine vs. Typical SVMs

2-tailed paired t-test	binary 1-vs-Set	binary linear	binary RBF	1-class 1-vs-Set	1-class linear	1-class RBF
binary 1-vs-Set (HOG 88)		**	**	**	**	**
binary linear (HOG 88)	—		—	++	++	++
binary RBF (HOG 88)	—	++		++	++	++
1-class 1-vs-Set (HOG 88)	—	—	—		**	—
1-class linear (HOG 88)	—	—	—	—		—
1-class RBF (HOG 88)	—	—	—	—	++	
binary 1-vs-Set (HOG 212)		**	*	**	**	**
1-class 1-vs-Set (HOG 212)	—	—	—		—	*
binary 1-vs-Set (LBP-like 88)		**	**	**	**	**
1-class 1-vs-Set (LBP-like 88)	—	—	—		**	—
binary 1-vs-Set (LBP-like 212)		*	—	**	**	**
1-class 1-vs-Set (LBP-like 212)	—	—	—		**	—

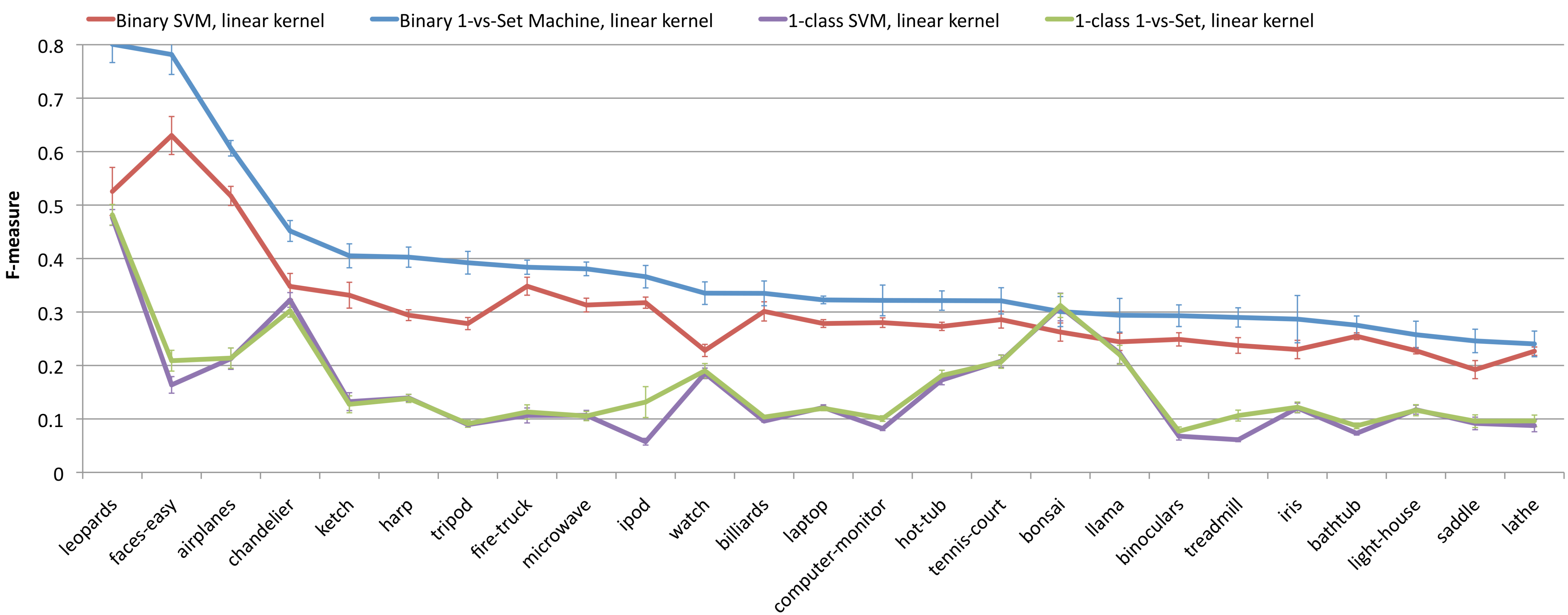
** 1-vs-Set Machine is statistically significant at $p < 0.01$

* 1-vs-Set Machine is statistically significant at $p < 0.05$

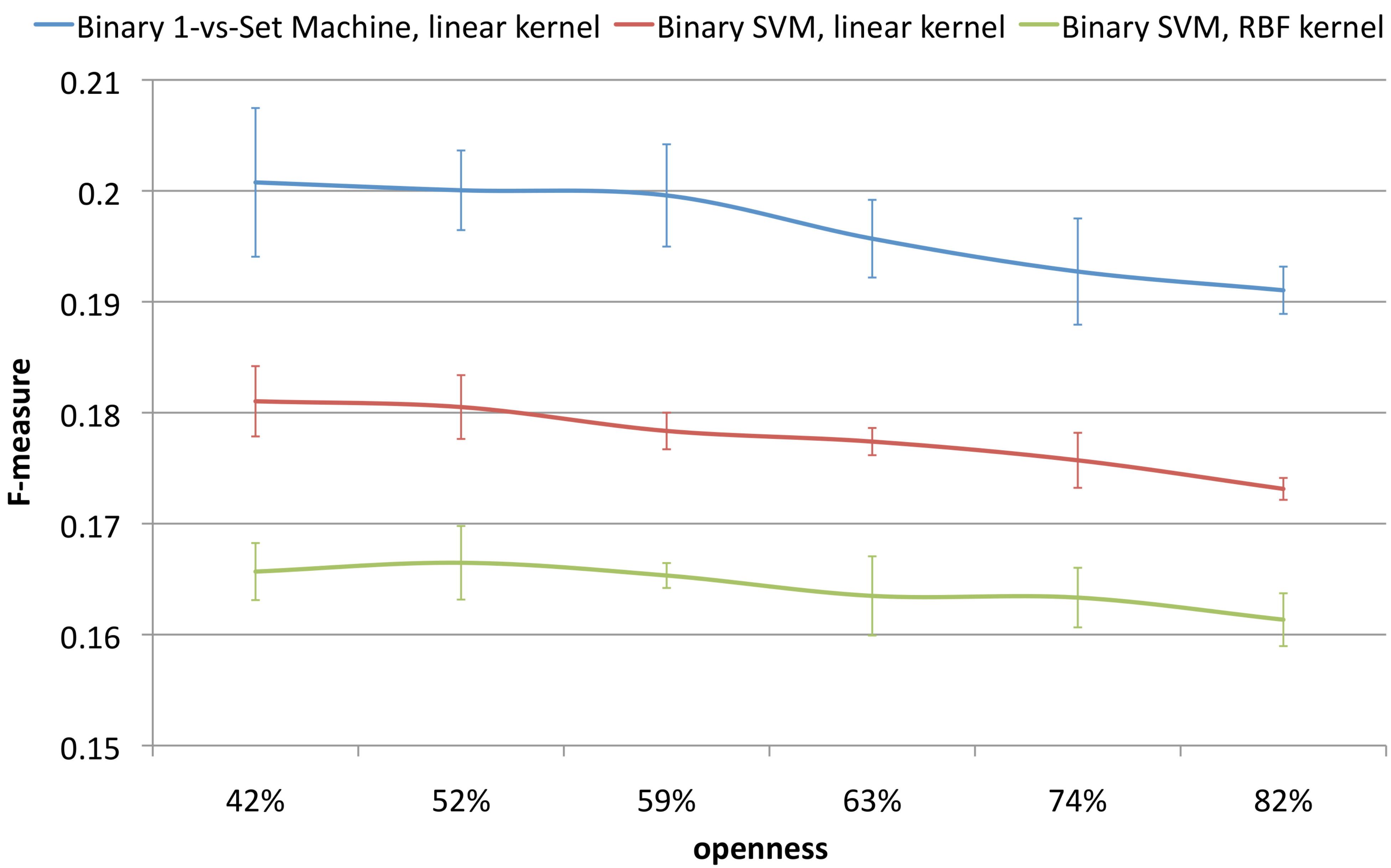
++ Baseline Machine is statistically significant at $p < 0.01$

— No statistical significance

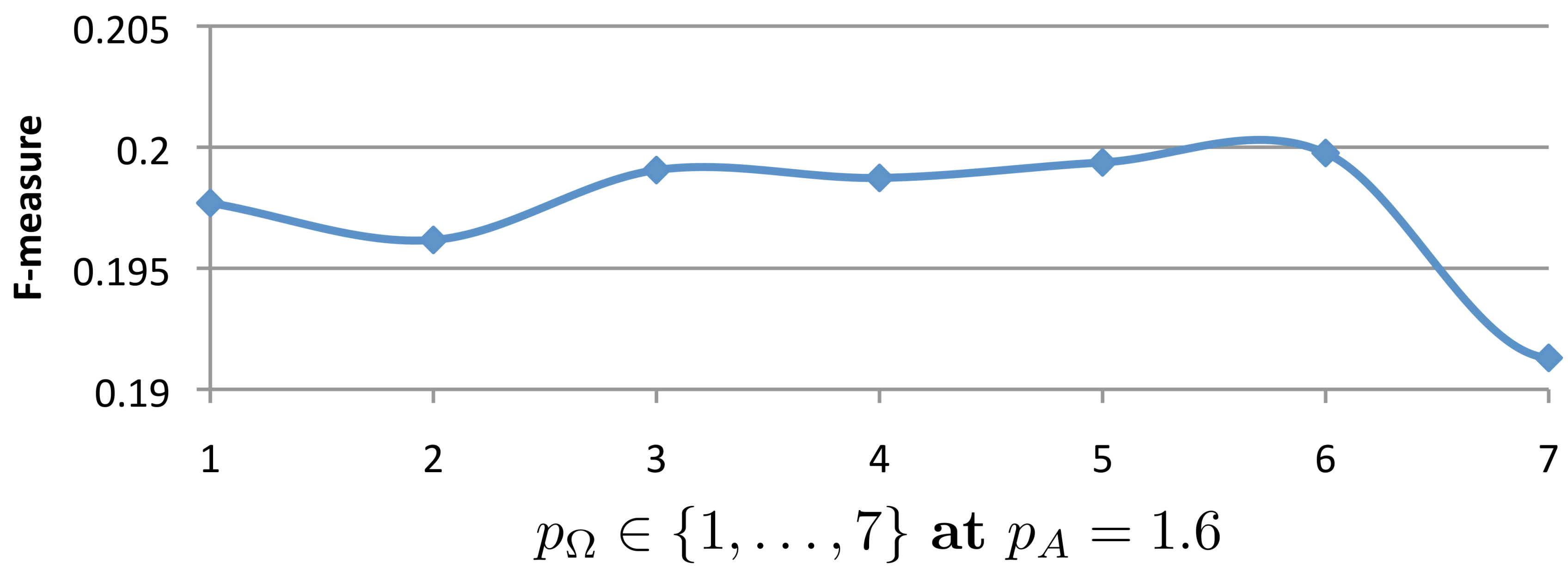
Top 25 classes for the open universe of 88 classes

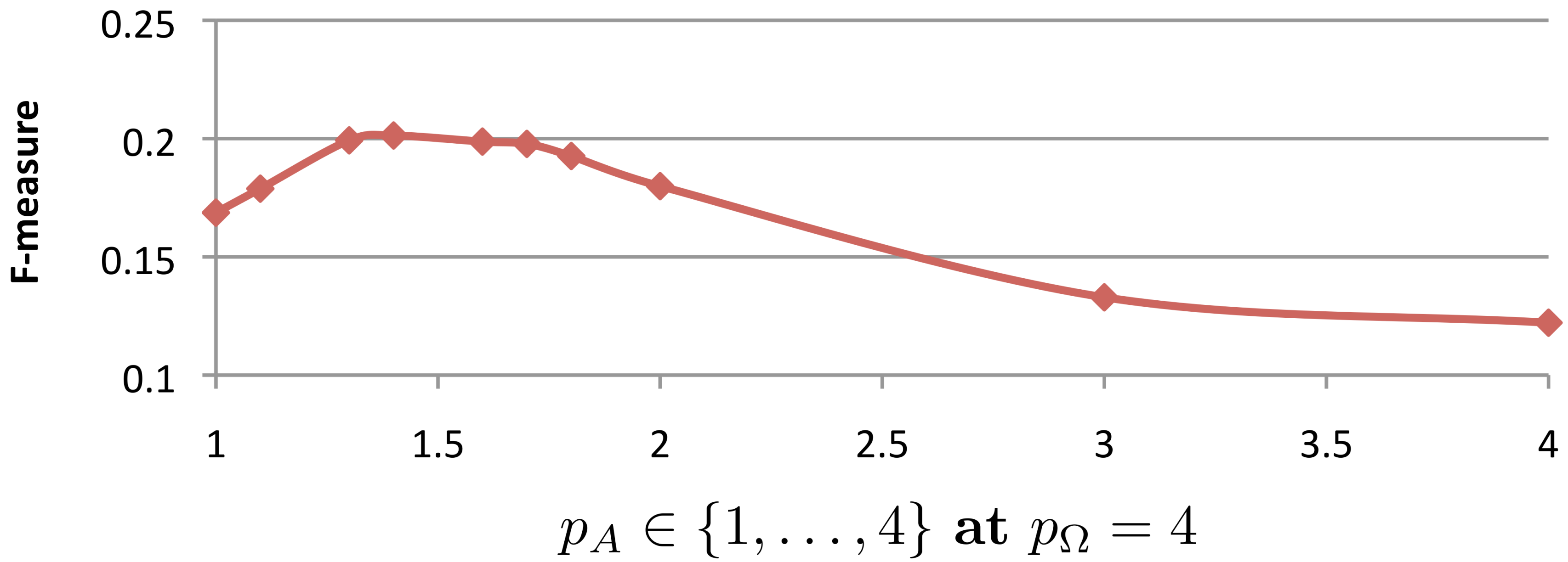


F-measure as a function of openness



Near and far plane pressures for open universe of 88 classes

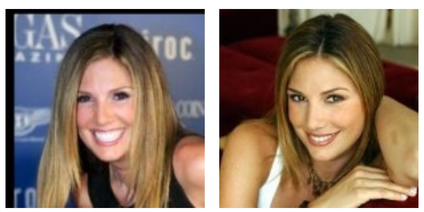




The second plane has an impact on recognition performance

Open Set Face Verification

Labeled Faces in the Wild



Genuine Pair

Impostor Pair

Impostor Pair

Impostor Pair

Gallery classes: 12 people with at least 50 images

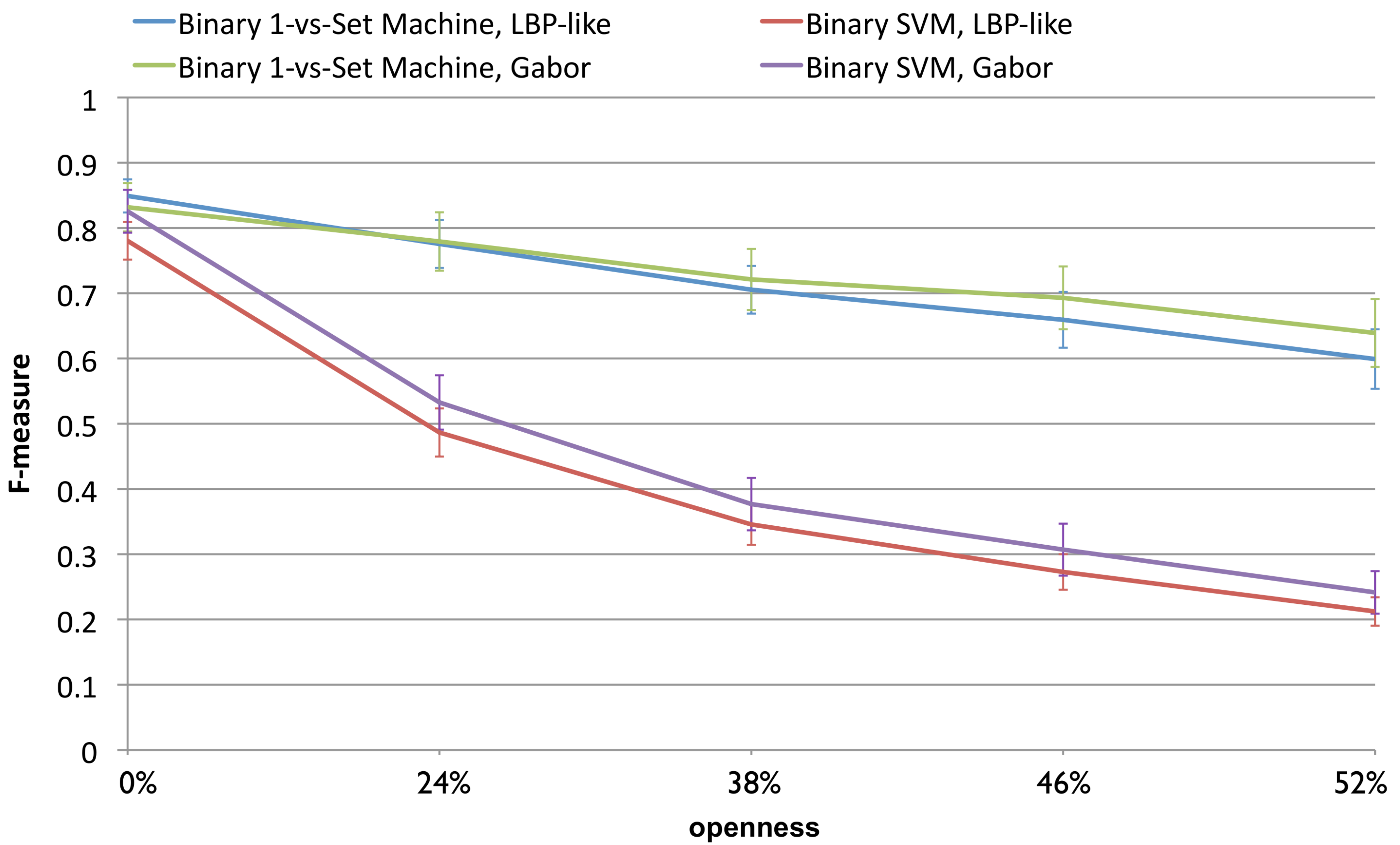
Impostor classes: 82 other people in LFW

1,316 test images across all classes

Features: LBP-like and Gabor*

N. Pinto, J. J. DiCarlo, and D. D. Cox, “How Far Can You Get with a Modern Face Recognition Test Set Using Only Simple Features?” in IEEE CVPR, 2009.

Open set face verification



Conclusions

Classification assumptions need to be questioned for open set recognition.

Open Space Risk helps us quantify the risk associated with the unknown.

Reducing open space leads to more stable solutions when lots of unknown data appears during testing.

Part 2: Open Set Influence Recognition

We find open set problems in NLP too

An emerging area: **Stylometry**

Machine learning is now able to detect quantifiable style markers in written language.

Common applications: authorship attribution, genre tagging, **textual reuse**

No text exists in isolation: we cannot make *a priori* assumptions on characteristics of style.

Intertextuality in Literature

Kristeva: “Any text is constructed as a mosaic of quotations; any text is the absorption and transformation of another.”

The nature of textual reuse is widely varied:

- Direct quotations
- Loose lexical correspondance
- Idea reuse
- Sound reuse

J. Kristeva, “Word, Dialog and Novel,” ed. Toril Moi, The Kristeva Reader, New York: Columbia University Press, 1986, p. 34-61.

Quantitative Intertextuality

Since the problem is one of pattern recognition, it is a good candidate for automated assistance by computers.

Quantitative Intertextuality is the algorithmic study of information reuse in any semiotic system.

Applications:

- Scholarly work (*e.g.* digital humanities)
- Practical applications (*e.g.* digital forensics)

Case Study: Paul the Deacon’s *Angustae Vitae*

Paul the Deacon: 8th century monk, intellectual, and court poet of the Lombards.

Catullus: influential Latin poet of the late Roman Republic.

Angustae Vitae is an epistolary poem that juxtaposes poetry in the classical world with poetry in a Christian monastic context.

- Clear Neoteric influence, and surprisingly reminiscent of Catullus

C. Forstall, S. Jacobson, and W. Scheirer, “Evidence of Intertextuality: Investigating Paul the Deacon’s *Angustae Vitae,*” Literary and Linguistic Computing, vol. 26, no. 3, 2011.

The Clues

Angustae Vitae is peppered with classical intertexts, but it remains an open question as to whether Paul the Deacon had actually read Catullus.

But we know…

- The diction and thematic models recall the Neoterics
- Paul the Deacon was well versed in the poetry of Horace, Ovid and Vergil
- Content suggests familiarity with Catullus 1, 2, 50 and 68

The Clues

Catullus 2

PASSER, deliciae meae puellae,
quicum ludere, quem in sinu tenere,
cui primum digitum dare appetenti
et acris solet incitare morsus
cum desiderio meo nitenti
carum nescio quid lubet iocari,
credo ut, cum gravis acquiescet ardor,
sit solaciolum sui dolaris,
tecum ludere sicut ipsa possem
et tristis animi levare curas!

Angustae Vitae, lines 1-4:

Angustae vitae fugiunt consortia Musae,
Claustrorum septis nec habitare volunt,
Per rosulenta magis cupiunt sed ludere prata,
Pauperiem fugiunt, deliciasque colunt:

Feature: the functional n-gram

A functional bi-gram is an n-gram-based feature that describes frequently appearing information.

Functional n-grams for **sound** are:

Character-level features

Stand-ins for phonemes

Similar to function words

Elements of most of the lexicon

$$P(e_n|e_{n-N+1}^{n-1}) = \frac{C(e_{n-N+1}^{n-1}e_n)}{C(e_{n-N+1}^{n-1})} \iff freq(e_{n-N+1}^{n-1}e_n) > \phi$$

Forstall, C.W. and W.J. Scheirer. 2009. "Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound," Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1.2.1-23.

The functional n-gram process

Select x of the most frequently occurring n-grams in a sample:

804	er	560	ti
778	qu	555	us
726	is	513	at
723	en	512	nt
709	re	503	ae
685	te	501	ta
651	es	470	tu
615	um	468	ri
604	in	454	or
574	it	452	am

$\longrightarrow$

Compute the probability features:

$$\frac{\text{Count}(\text{“er”})}{\text{Count}(\text{“e”})} = 0.179785$$

$$\vdots$$

$$\frac{\text{Count}(\text{“re”})}{\text{Count}(\text{“r”})} = 0.275447$$

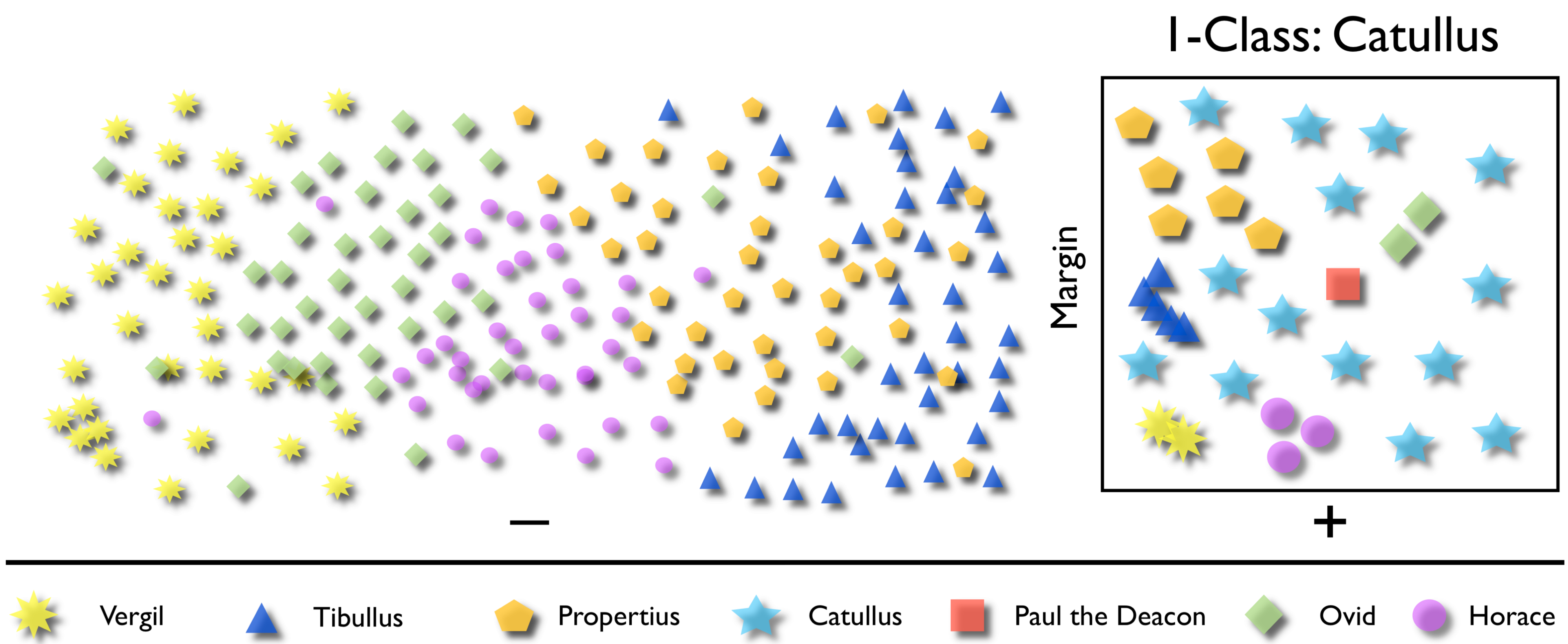
Open Set Influence Analysis

We want to test the stylistic similarity of any poet to Catullus.

Methodology: train a 1-class SVM on representative samples from Catullus.

- Recall that 1-class SVMs tend to overfit the training data.
- We don't need to generalize - we are only interested in samples that fall within the support of the training data.

Open Set Influence Analysis



Low-Probability Analysis

Words that occur infrequently are often intertexts.
Incorporate this into the influence analysis process:

Fix a desired probability range for words that occur infrequently.

Scan for n-gram sequences composed of only those words in a particular passage (ignoring all others).

Use as an additional feature to augment our vector of functional n-grams.

$$(P_{low} < \Pr(\text{word}_1) < P_{high}) \ \ldots \ (P_{low} < \Pr(\text{word}_2) < P_{high}) \ \ldots \ (P_{low} < \Pr(\text{word}_n) < P_{high})$$

Low-Probability Analysis

Key n-gram sequences common to *Angustae Vitae* and Catullus:

delic(ias|iae) ludure

flagra(t|ns|bat) amor_

redde miser poema

Evidence of Catullan Influence in *Angustae Vitae*

Identifying a functional n-gram of interest:

Looked at most frequently occurring character-level bigrams in Catullus 1 - 64.

Rank	bi-gram	Probability		Rank	bi-gram	Probability
1	er	0.180	← Most Frequent	6	te	0.251
2	qu	1.000		7	es	0.146
3	is	0.169		8	um	0.164
4	en	0.162		9	in	0.140
5	**re**	**0.275**	← Most Probable	10	it	0.133

Evidence of Catullan Influence in *Angustae Vitae*

Sample 're' training data for the poems of Catullus (features most similar to poems 1 & 2):

Feature	Poems	Feature	Poems
0.458	1 and 2	0.435	38 and 39
0.412	2b and 3	0.480	44 and 45
0.455	4 and 5	0.480	50 and 51
0.524	7 and 8	0.444	62
0.500	13	0.463	64
0.406	17 and 21	0.464	64

Evidence of Catullan Influence in *Angustae Vitae*

Samples classified positively with Catullus out of all samples:

More Like Catullus		**Less Like Catullus**	
Text	**Positive Class.**	**Text**	**Positive Class.**
Angustae Vitae	**1/1**	Ovid *Amores*	2/40
Propertius *Elegies*	6/40	Horace *Epistles*	3/40
Tibullus *Elegies*	5/40	Virgil *Aeneid*	2/35

Evidence of Catullan Influence in *Angustae Vitae*

Samples classified positively with Catullus out of all samples after refinement using low-probability features:

More Like Catullus		Less Like Catullus	
Text	**Positive Class.**	**Text**	**Positive Class.**
Angustae Vitae	**1/1**	Ovid *Amores*	1/40
Propertius *Elegies*	5/40	Horace *Epistles*	1/40
Tibullus *Elegies*	5/40	Virgil *Aeneid*	1/35

Conclusions

Quantitative Intertextuality has the potential to draw out new aspects of stylistic influence and reference.

Functional n-gram feature provided a tool to quantify a notable stylistic similarity in sound between Catullus 1 & 2 and *Angustae Vitae*

Low probability analysis was able to refine results.

Wrapping Up

Further Reading

- W.J. Scheirer, A. Rocha, A. Sapkota, and T. Boult, “Towards Open Set Recognition,” IEEE T-PAMI, 35(7) July 2013.
- C. Forstall, S. Jacobson, and W. Scheirer, “Evidence of Intertextuality: Investigating Paul the Deacon’s *Angustae Vitae,*” Literary and Linguistic Computing, vol. 26, no. 3, 2011.
- W.J. Scheirer, L. Jain, and T. Boult, “Probability Models for Open Set Recognition,” To Appear in IEEE T-PAMI, 2014.
- L. Jain, W.J. Scheirer, and T. Boult, “Multi-Class Open Set Recognition Using Probability of Inclusion,” ECCV, 2014.

Code

1-vs-Set Machine on GitHub:
https://github.com/tboult/libSVM-onevset

Data sets:
http://www.metarecognition.com/openset/