Statistical Methods for Open Set Recognition

Walter J. Scheirer and Terrance E. Boult

Companion Website

All material for this tutorial is available at:

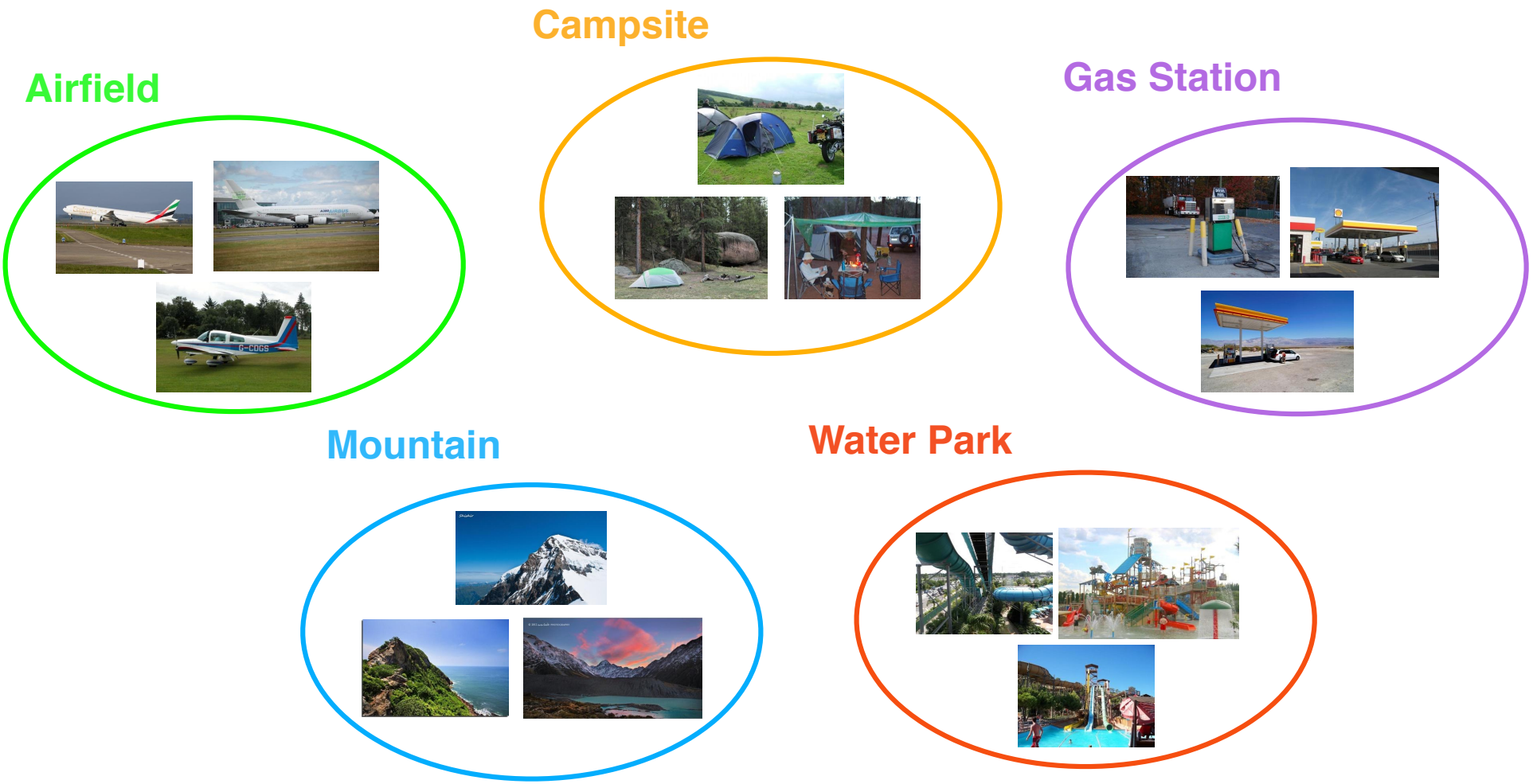
http://www.wjscheirer.com/misc/openset/

(Also linked to from the CVPR 2016 Website)

Part 1: An Introduction to the Open Set Recognition Problem

Benchmarks in computer vision

Assume we have examples from all classes:



Places2 Data Set (part of ILSVRC 2016)

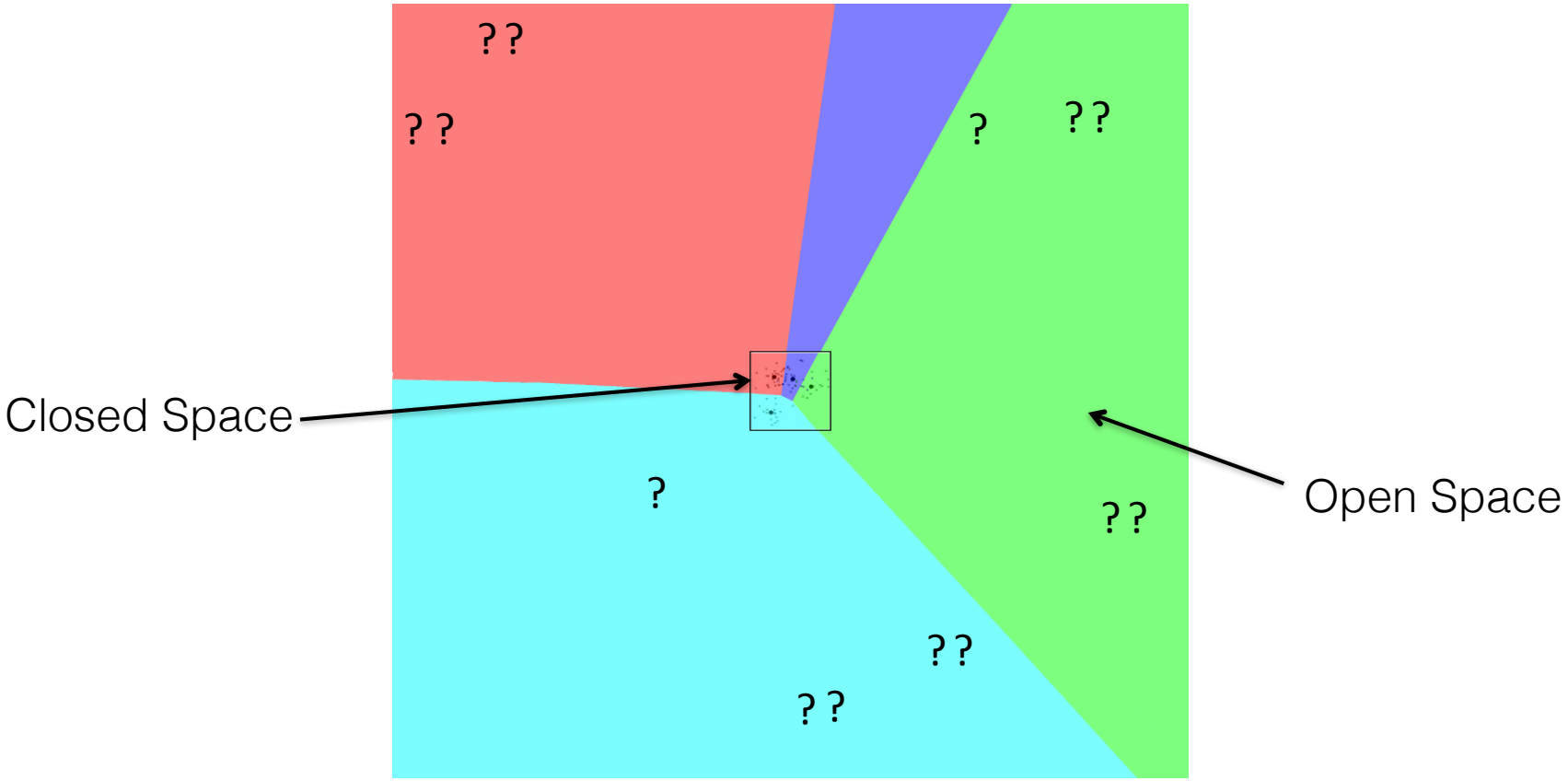
Out in the real world…

Be on the lookout for blue Ford sedans

while rejecting the trees, signs, telephone poles…

M. Milford, W.J. Scheirer, E. Vig, A. Glover, O. Baumann, J. Mattingley, and D.D. Cox, “Condition Invariant Top-Down Visual Place Recognition,” ICRA 2014.

Open Space in Classification



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?

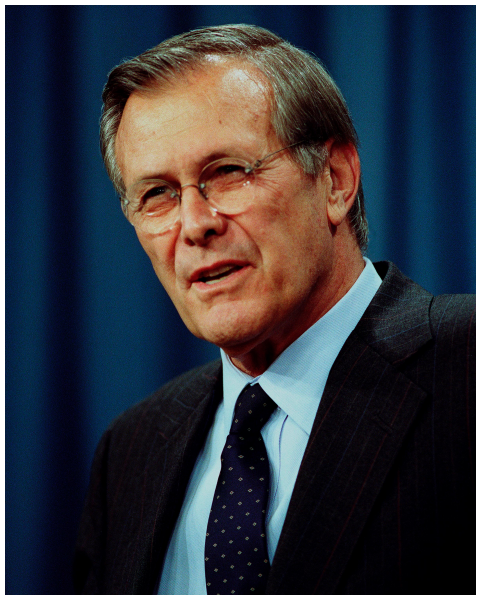
R. P. Duin and E. Pekalska, “Open Issues in Pattern Recognition,” in *Computer Recognition Systems*, M. Kurzynski, E. Puchala, M. Wozniak, and A. Zolnierek, Eds. Springer, 2005, pp. 27–42.

“There are known knowns…”

known classes: the classes with distinctly labeled positive training examples (also serving as negative examples for other known classes)

known unknown classes: labeled negative examples, not necessarily grouped into meaningful categories

unknown unknown classes: classes unseen in training



Definitions

Closed Set Recognition: all testing classes are known at training time

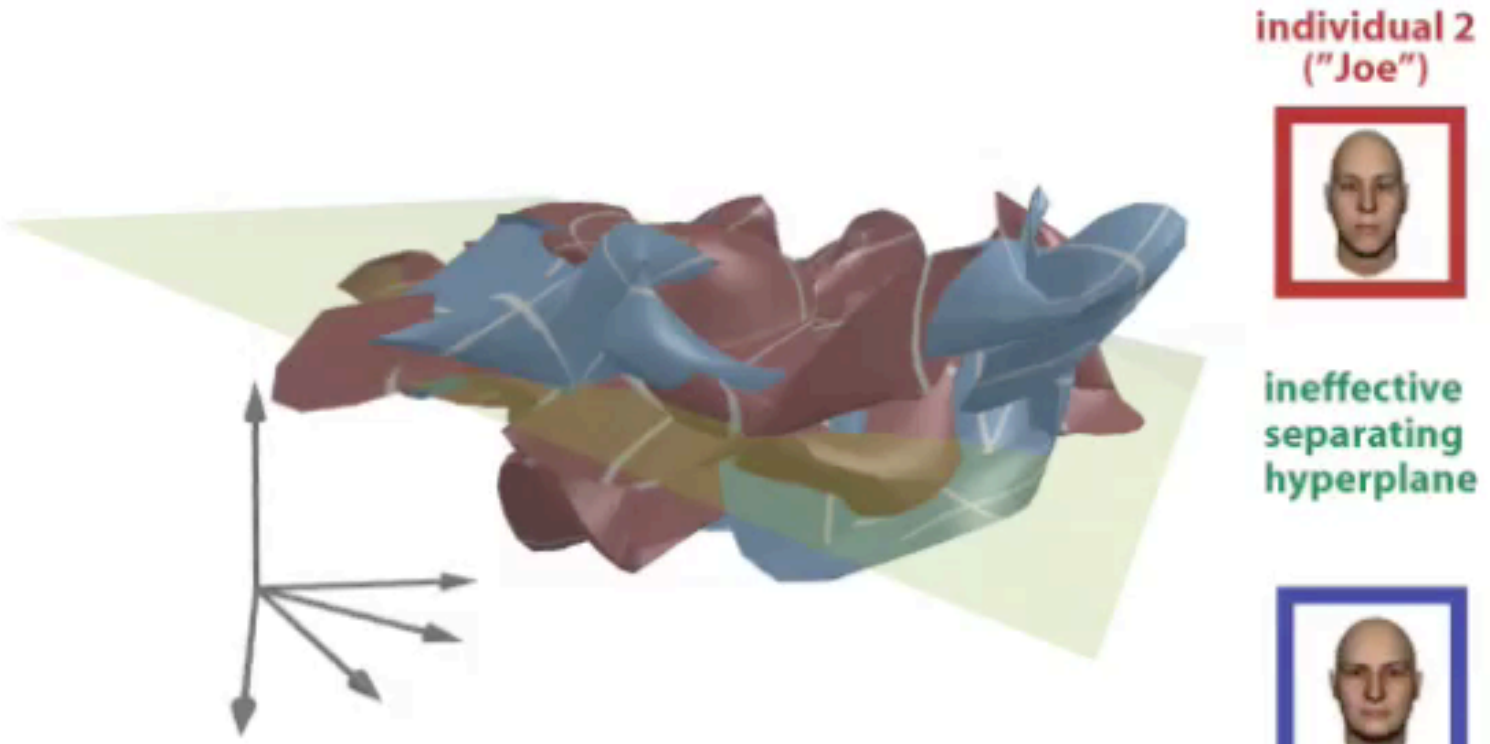
Open Set Recognition: incomplete knowledge of the world is present at training time, and unknown classes can be submitted to an algorithm during testing

The burden for the visual recognition community

- Results look better than they really are, which is misleads practitioners
- “Off-the-shelf” classifiers are not sufficient to solve the problem
- Open set problems are found in nearly every case where recognition algorithms are present

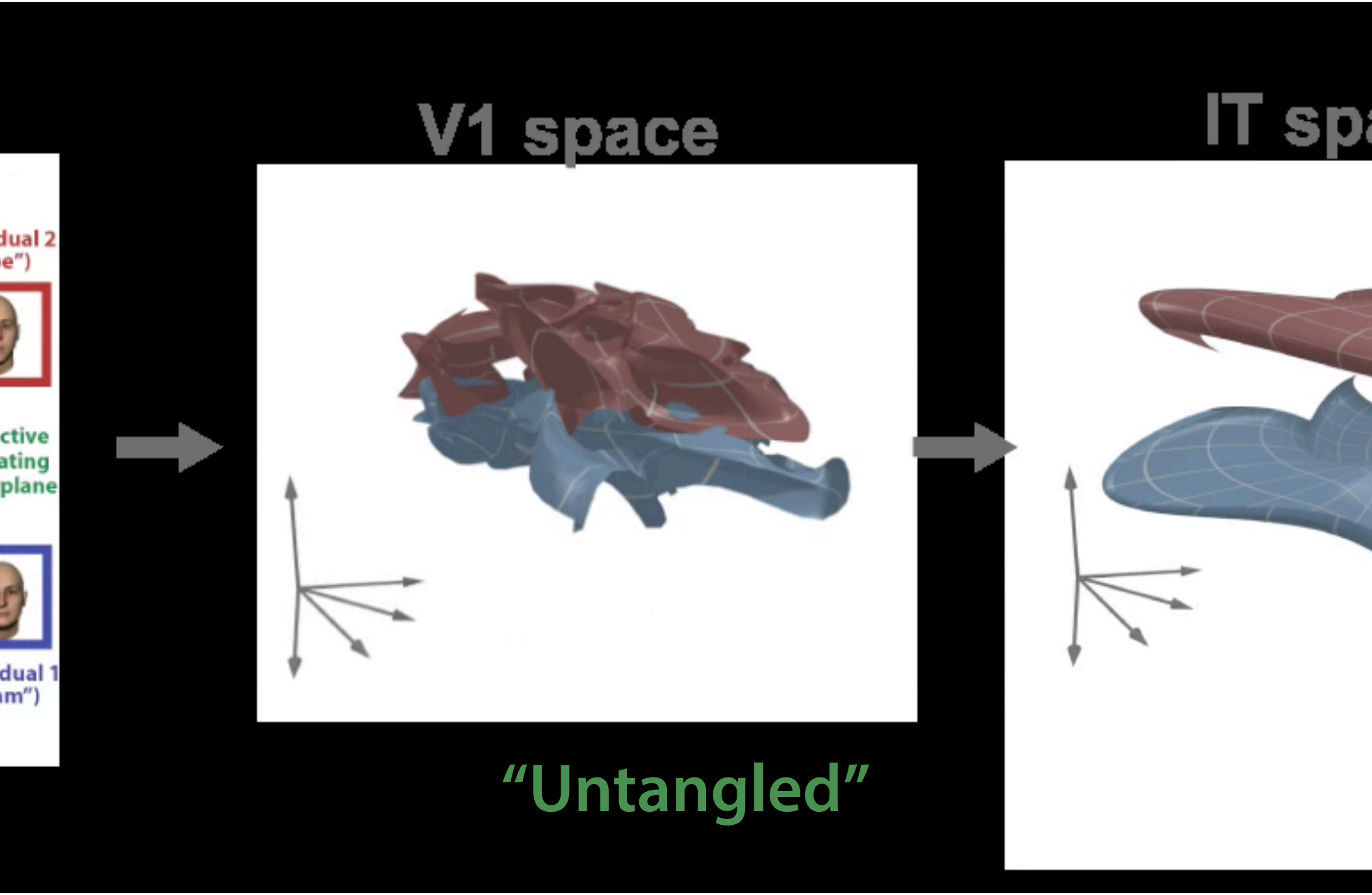
A surprising finding…

pixel space

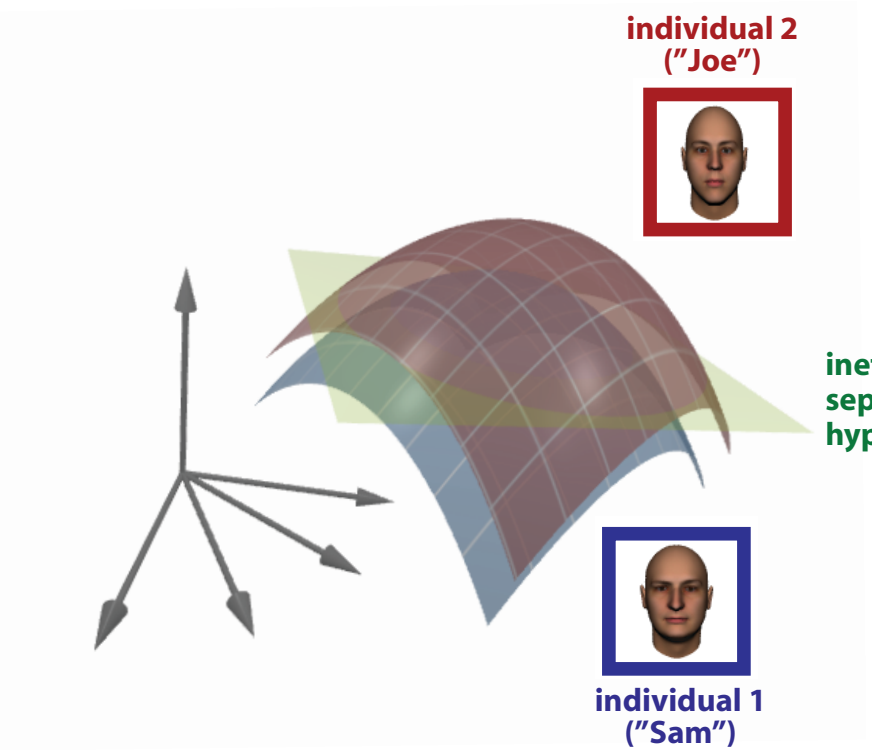


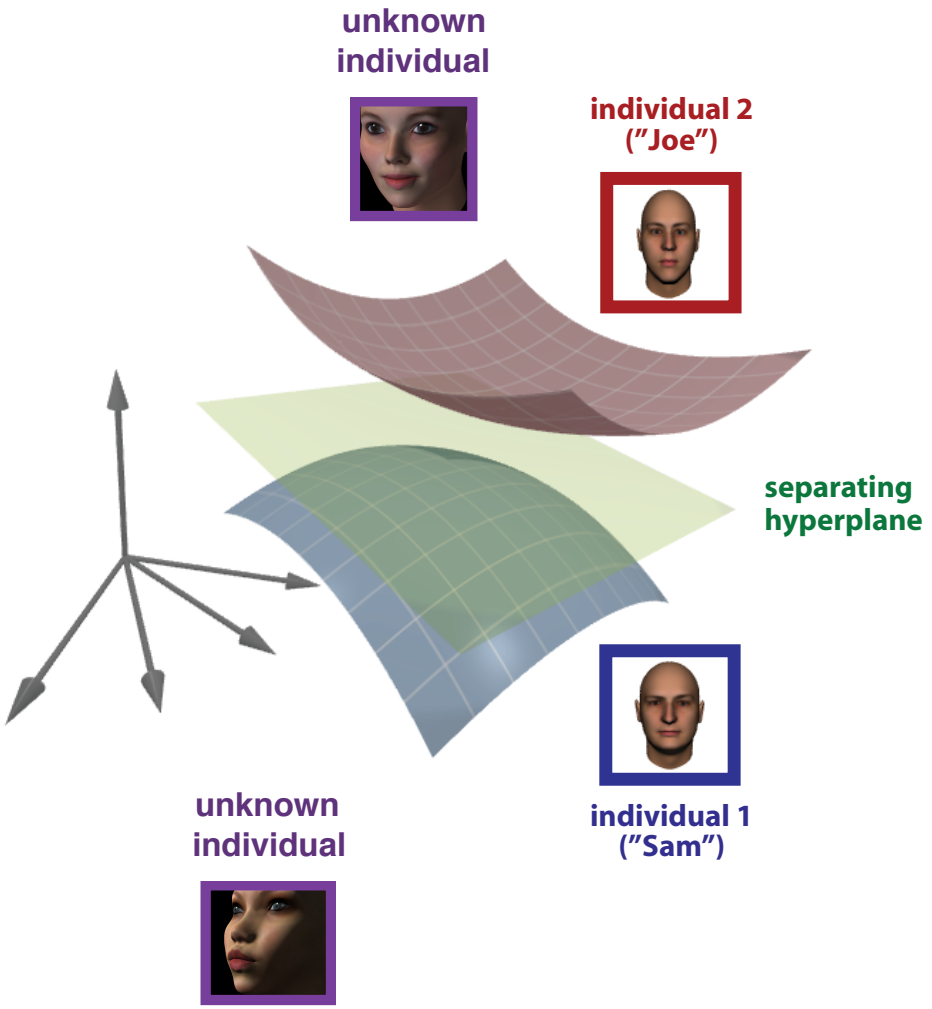
"Tangled"

Adapted from an image by D. D. Cox

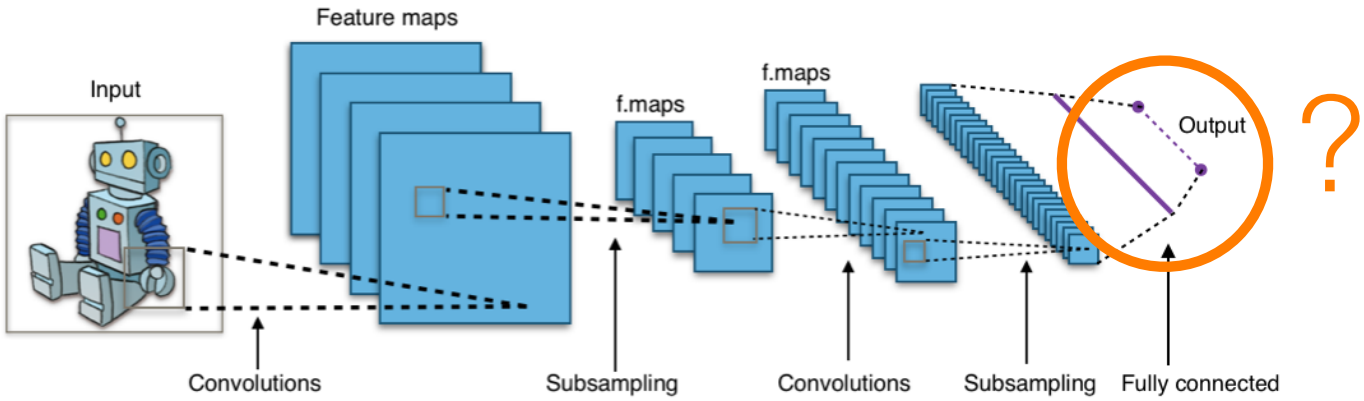

V1 space
IT spa
dual 2
e")
ctive
ating
plane
dual 1
am")
"Untangled"

Linear separation of CNN feature representations





Read-out layer



Softmax

$$P(y = j|\mathbf{x}) = \frac{e^{\mathbf{v_j}(\mathbf{x})}}{\sum_{i=1}^{N} e^{\mathbf{v_i}(\mathbf{x})}}$$

Sum over all of the classes

Linear SVM

$$min \frac{1}{2}||w||^2$$

subject to

$$y_i(w * x_i + b) \geq 1, \forall_i$$

Known positive or negative sample

Cosine Similarity

$$\frac{\mathrm{A} \cdot \mathrm{B}}{||\mathrm{A}||\ ||\mathrm{B}||} < \delta$$

Threshold determined empirically via known pairs

Evolving images to match CNN classes

A. Nguyen, J. Yosinski, and J. Clune, “Deep Neural Networks are Easily Fooled,” CVPR 2015.

But you don’t have to use tricky manipulations

GoogleNet Output

Label: Hammerhead Shark

Label: Blow Dryer

Label: Mosque

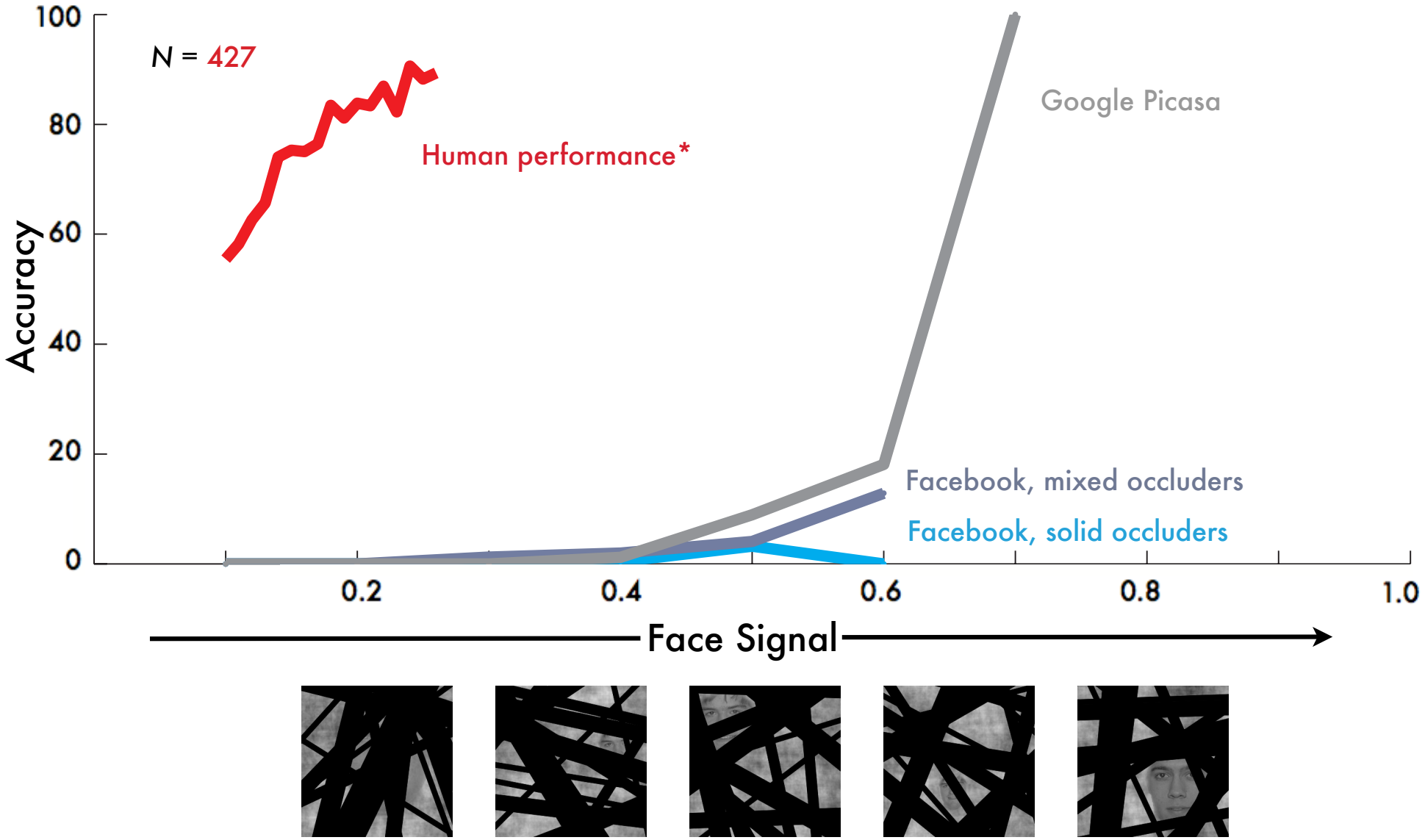
Label: Syringe

Label: Trimaran

Label: Missile

Are performance measures misleading us?

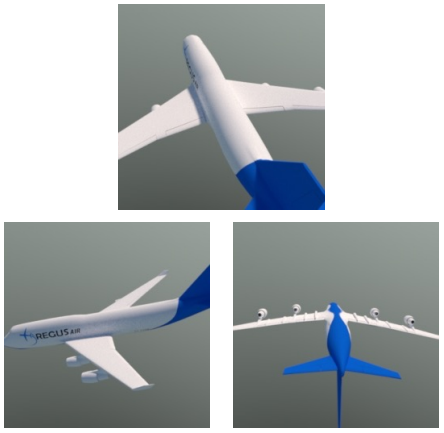
Psychophysics on the Model



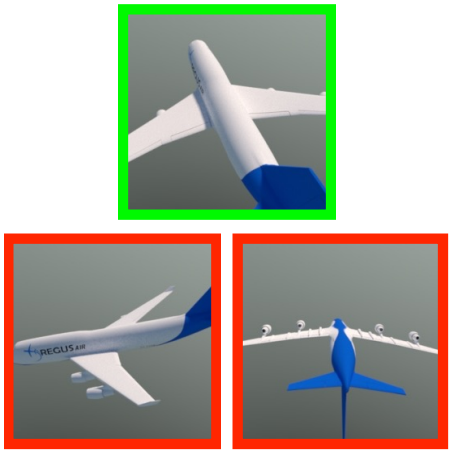
W.J. Scheirer, S. Anthony, K. Nakayama, and D. D. Cox, “Perceptual Annotation: Measuring Human Vision to Improve Computer Vision,” IEEE T-PAMI, 36(8) August 2014.

Psychophysics pipeline

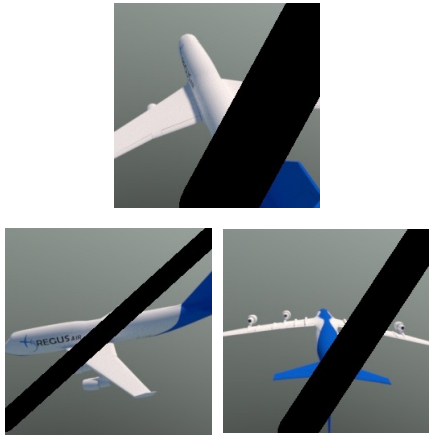
1. Render Class Canonical View (CCV) Candidates



2. CCV Classifier



3. Manipulate Chosen Variable



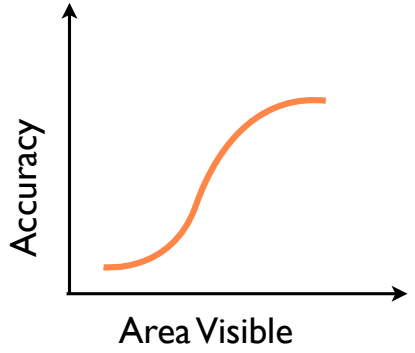
4. Classify Images

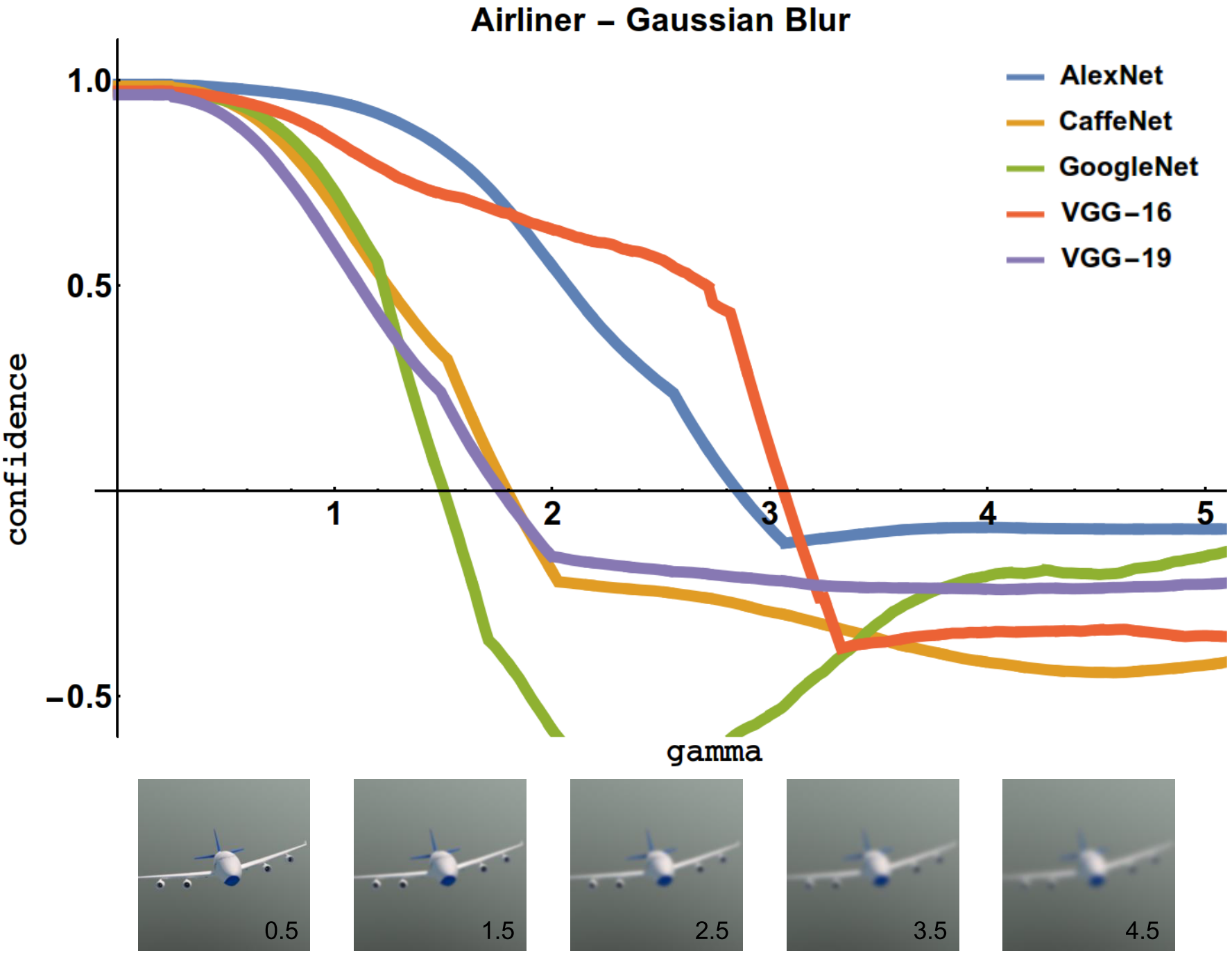
Fish

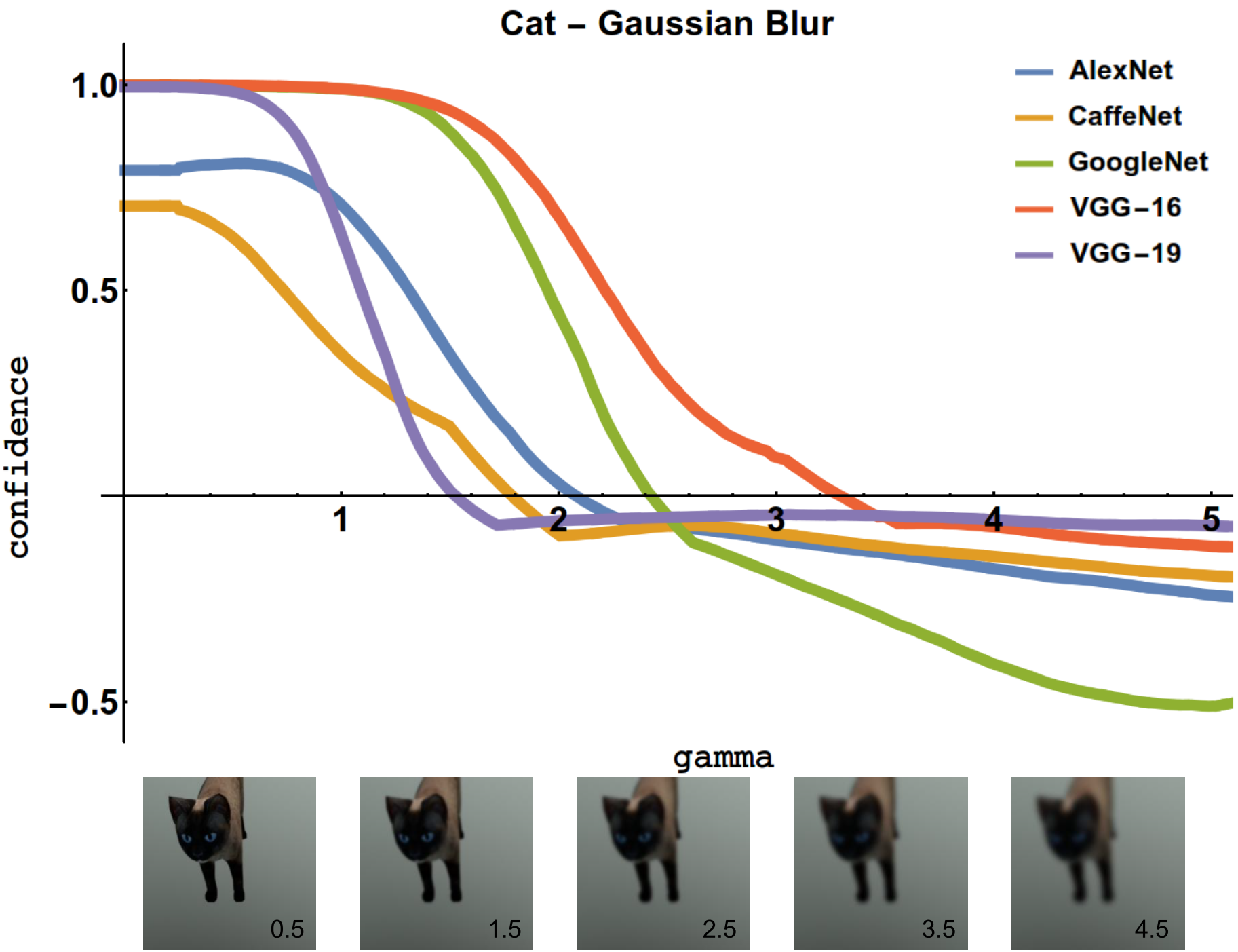
Plane

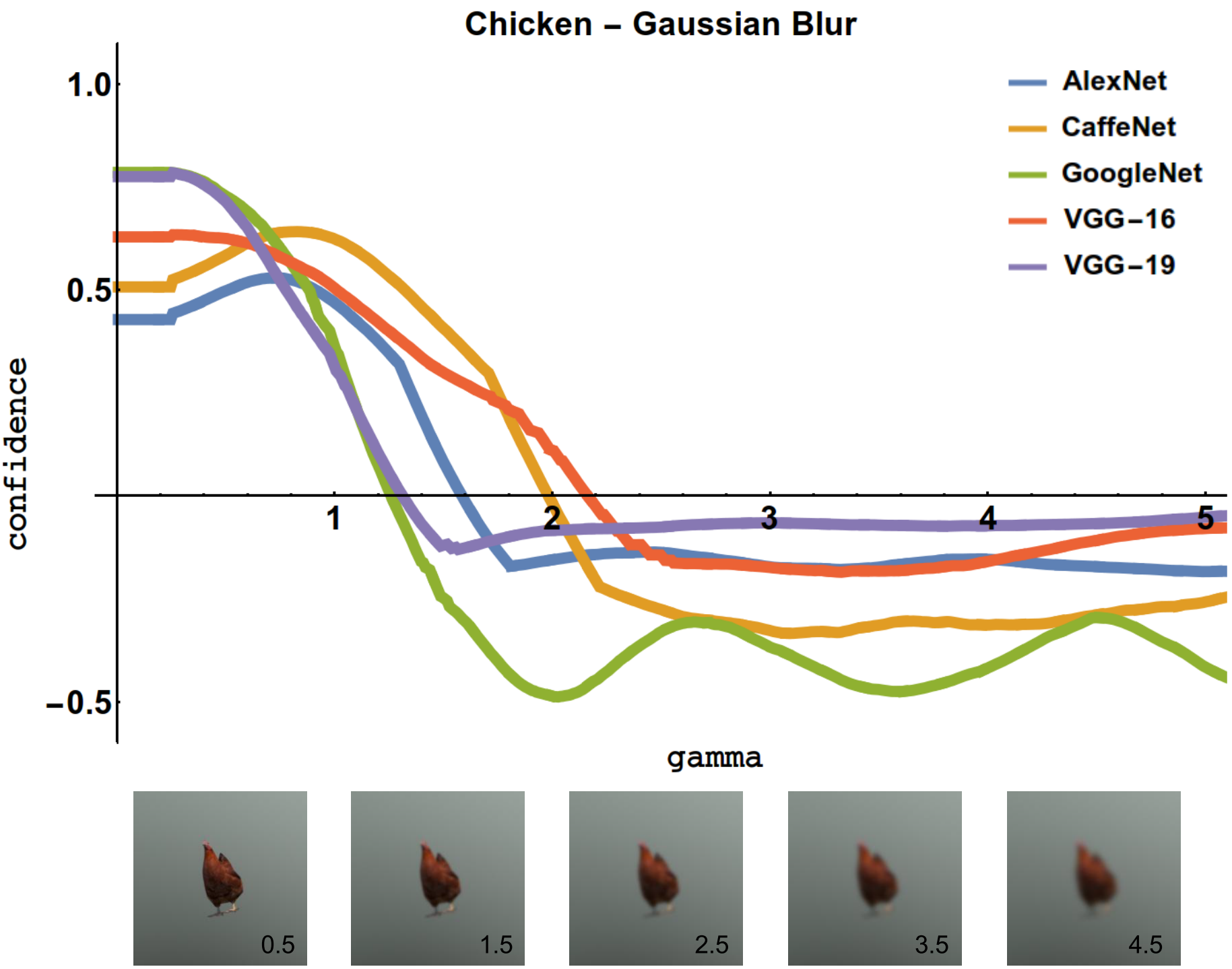
Skyscraper

5. Generate Psychometric Curve



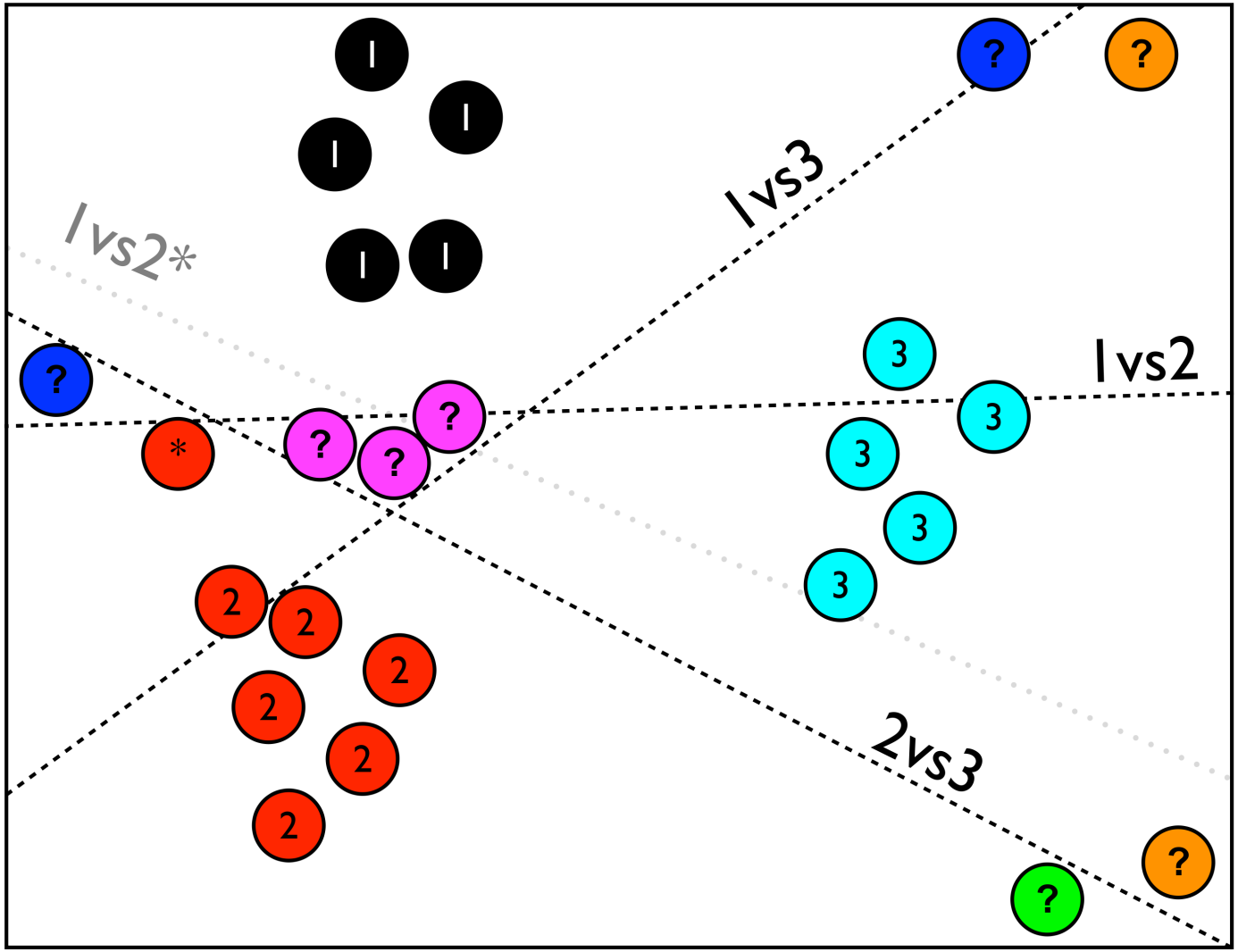

Airliner – Gaussian Blur
AlexNet
CaffeNet
GoogleNet
VGG–16
VGG–19
1.0
0.5
–0.5
confidence
1
2
3
4
5
gamma
0.5
1.5
2.5
3.5
4.5


Cat – Gaussian Blur
AlexNet
CaffeNet
GoogleNet
VGG–16
VGG–19
1.0
0.5
–0.5
confidence
1
2
3
4
5
gamma
0.5
1.5
2.5
3.5
4.5

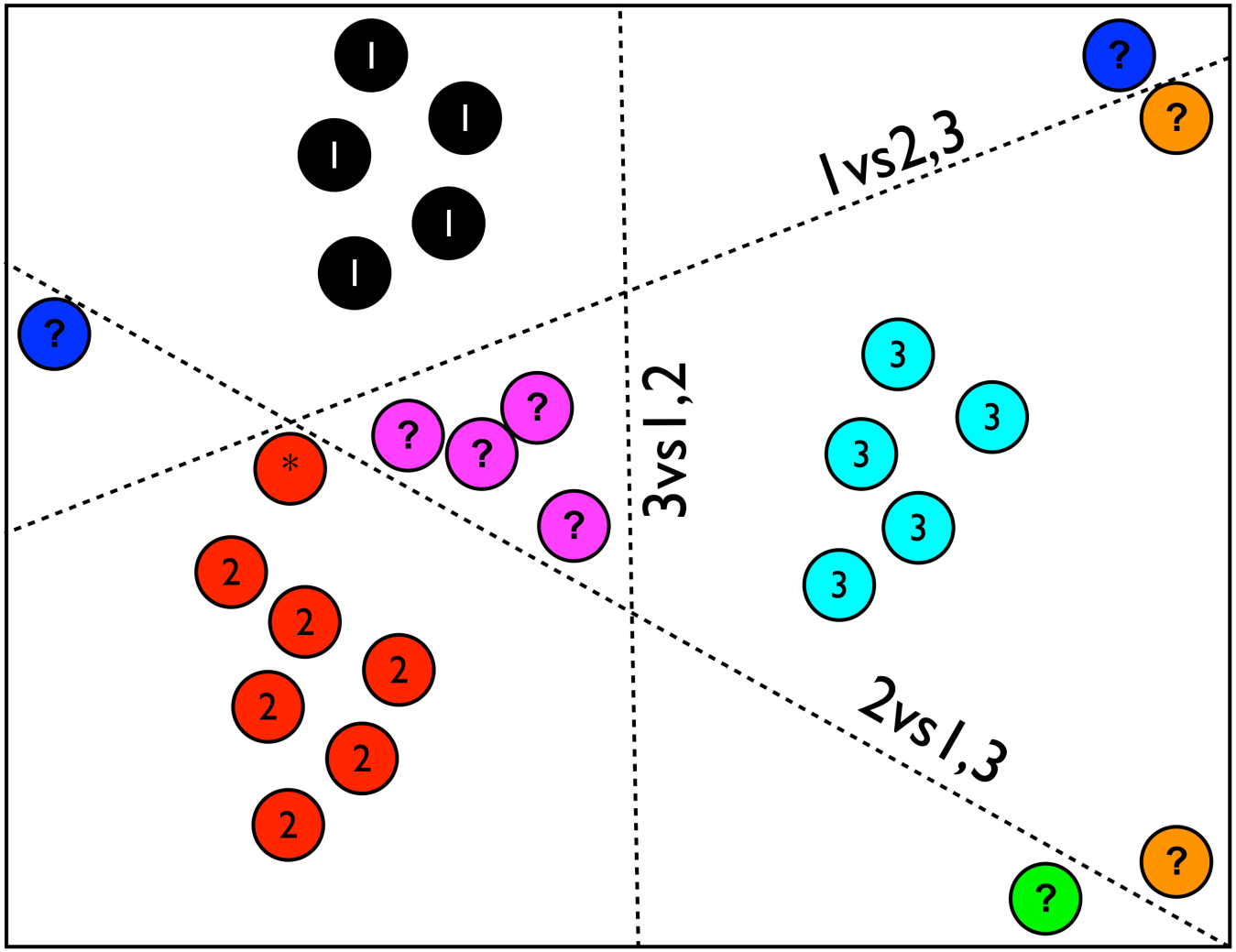

Chicken – Gaussian Blur
AlexNet
CaffeNet
GoogleNet
VGG–16
VGG–19
confidence
1.0
0.5
–0.5
1
2
3
4
5
gamma
0.5
1.5
2.5
3.5
4.5

What standard options do we have to solve this problem?

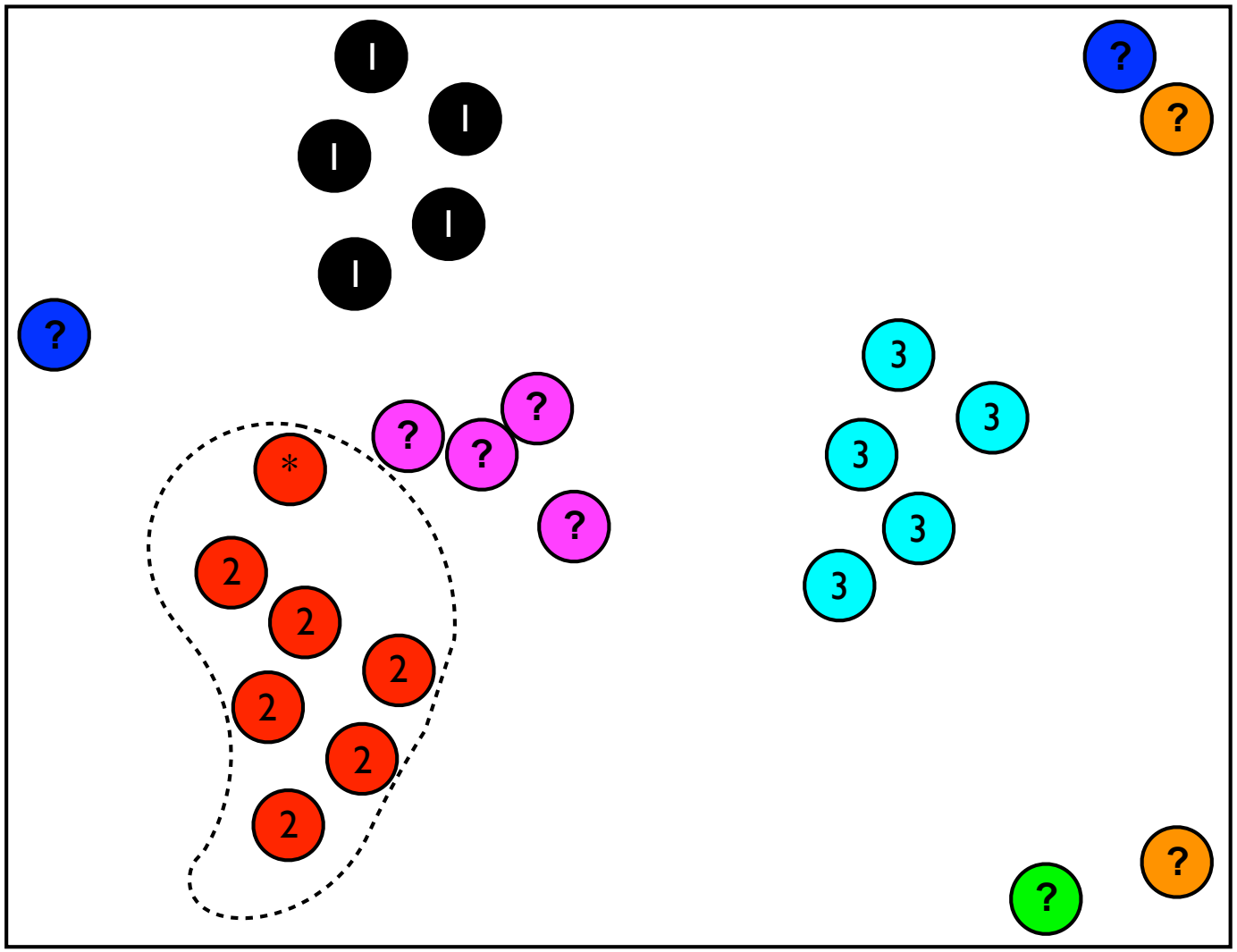
Binary Classification



Multi-class 1-vs-All Classification



1-class Classification



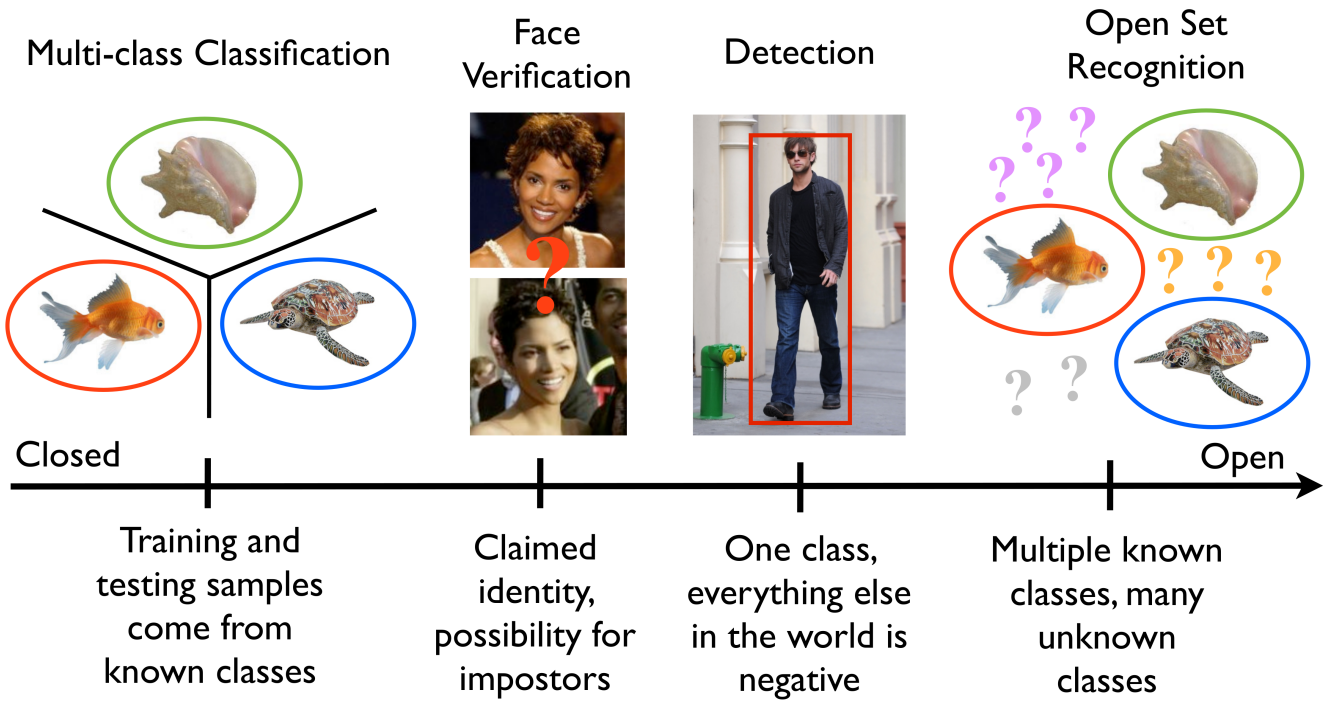
B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and “R. Williamson. Estimating the Support of a High-dimensional Distribution,” Technical report, Microsoft Research, 1999.

“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.

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	0.38
Typical Detection	1	100,000	1,000,000	0.55
Object Recognition	88	12	88	0.63
Object Recognition	88	6	88	0.74
Object Recognition	212	6	212	0.83

Fundamental multi-class recognition problem

$$\operatorname*{argmin}_f \left\{ R_{\mathcal{I}}(f) := \int_{\mathbb{R}^d \times \mathbb{N}} L(x, y, f(x)) P(x, y) \right\}$$

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

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

open space

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's missing from our definition of open space risk?

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 definition doesn't tell us how to define $\mathcal{O}$

Incorporating open space risk into a model

- Discriminative models?

  Don't address unknown unknowns in open space

- Generative models?

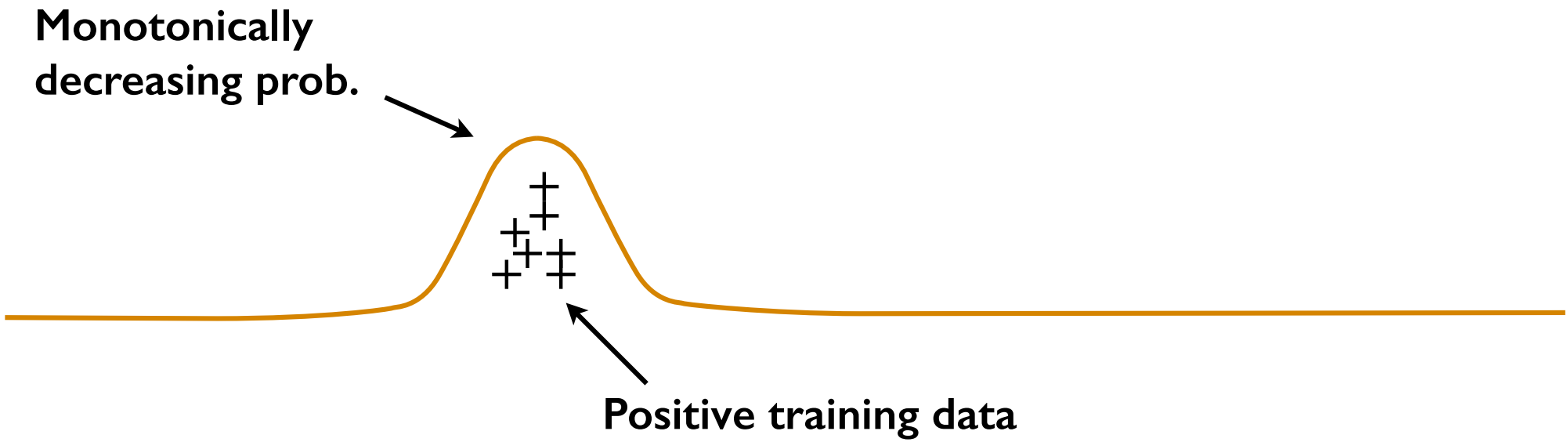
  Don't address unknown unknowns in open space

- Hard negative mining (Felzenszwalb et al. 2010)?

Not possible to mine examples from unknown classes

Abating Process

- Model enforced decay of probability away from supporting evidence



W. Scheirer, L. Jain, and T. Boult, “Probability Models for Open Set Recognition,” IEEE T-PAMI 2014

The Compact Abating Probability Model

Conceptual example: if we are labeling location data using training data only from Campinas, Brazil, it would be risky it would be risky to apply that model to South Bend, Indiana

Distance

Far

Idea: ensure that the recognition function is decreasing away from the training data, so that thresholding it limits the labeled region.

Definition of Open Space

$$\mathcal{O} = S_o - \bigcup_{i \in N} B_r(x_i)$$

closed ball of radius r centered around training sample x_i

Treat r as a problem specific parameter

Abating Bound

$$\forall x, x_i : \quad 0 < K(x, x_i) \leq A(\|x - x_i\|)$$

Positive Definite Kernel (e.g., RBF)

Non-negative finite square integrable continuous decreasing function

When $\forall x, \exists x^* \mid f(x) \leq A(\|x - x^*\|)$,
f is abating because the spatial influence decreases with distance from x^*

Abating Probabilistic Point Model

Fusion Operator
(e.g., sum or product)

$$M(x) = p_f(F(K(x, x_1) \ldots K(x, x_m)); y)$$

Model

Probability of points associating becomes less intense as the spatial separation of any two points increases.

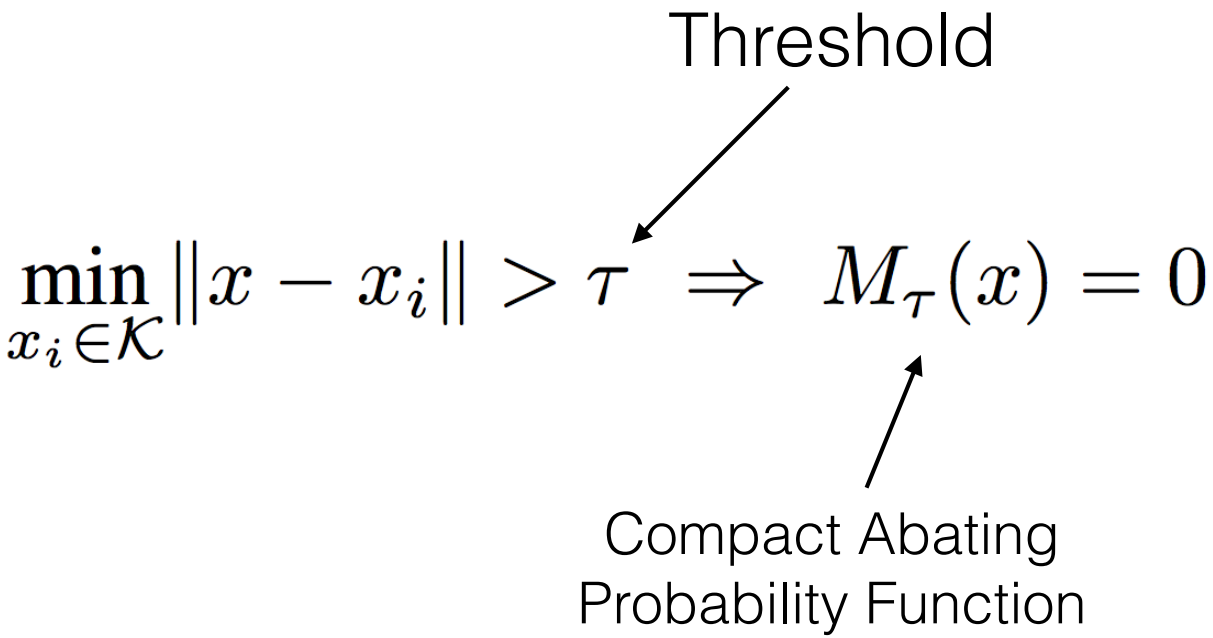
Fused Abating Property

After fusion there is an abating bound function centered at x_0 such that the fused value F is bounded from above by that abating function.

$$F(K(x, x_1) \dots K(x, x_m)) \leq A_{x'}(\|x' - x\|)$$

Abating Bound Function

Compact Abating Probability (CAP) Model



Features beyond a given thresholded τ from the closest training point have zero probability

Theorem 1

Let $M_{\tau,y}(x)$ *be a probabilistic recognition function that uses a CAP model over a known training set for class* y, where $\exists x_i \in \mathcal{K} \mid M_{\tau,y}(x_i) > 0$. *Let open space risk be* $R_{\mathcal{O}}(f)$ *and open space be* $\mathcal{O}$. *If* r *satisfies* $r > \tau$, *then* $R_{\mathcal{O}}(M_{\tau,y}) = 0$,

What does this mean?

When the CAP distance threshold is smaller than the open space radius, the CAP model has zero open space risk.

Proof of Theorem 1

Let x *be any point in* $\mathcal{O}$. *Since* $x \in \mathcal{O}$ *implies* $x \notin \bigcup_{i \in N} B_r(x_i)$, *we have* $\forall x_i \in \mathcal{K}, \|x - x_i\| > r > \tau$. *Therefore, by the compact abating property* $M_{\tau,y}(x) = 0$. *Placing this into the numerator of* $R_{\mathcal{O}}(f)$ *yields* $\int_{\mathcal{O}} M_{\tau,y}(x)dx = 0$ *and zero open space risk.* □

Corollary 1

Thresholding CAP model probability manages Open Space Risk

For any CAP model, considering only points with sufficiently high probability will reduce open space risk. In particular, consider a canonical sum kernel-based CAP model with a probability threshold $0 \leq \delta_\tau \leq 1$ *such that for the set of points* $x_i \in \mathcal{K}$ *and coefficients* $\vartheta_i > 0$ $p_f(\sum_i \vartheta_i K(x, x_i); y) \geq \delta_\tau$. *Increasing* δ_τ *decreases open space risk, and there exists a* δ_τ^* *such that any greater threshold produces zero open space risk*

How do we get from Corollary 1 to an algorithm?

- No guarantee that the model assigns positive labels within the compact support region
  - CAP ensures that there is a zero probability of doing so outside the region
- Quality of the CAP model depends on how well the probabilities model the actual underlying positive region
- **1-class SVM + Non-linear (RBF) kernel**

Theorem 2

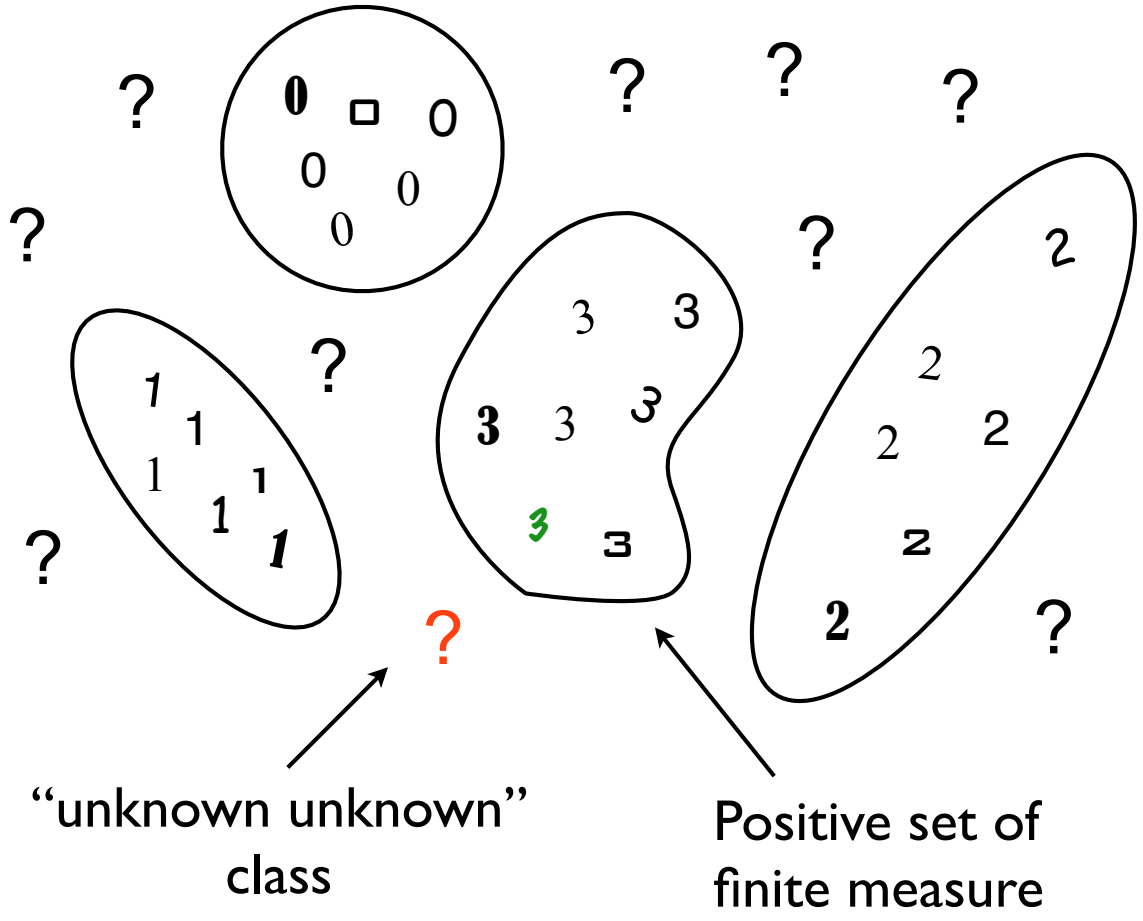
RBF One-Class SVM yields CAP model

Let $x_i \in \mathcal{K}, i = 1 \ldots m$ be the training data for class y. Let O-SVM be a 1-class SVM with a square integrable monotonically decreasing RBF kernel K defined over the training data, with associated Lagrangian multipliers $\alpha_i > 0$, then $\Sigma_i \alpha_i y_i K(x, x_i)$ yields a CAP model. □

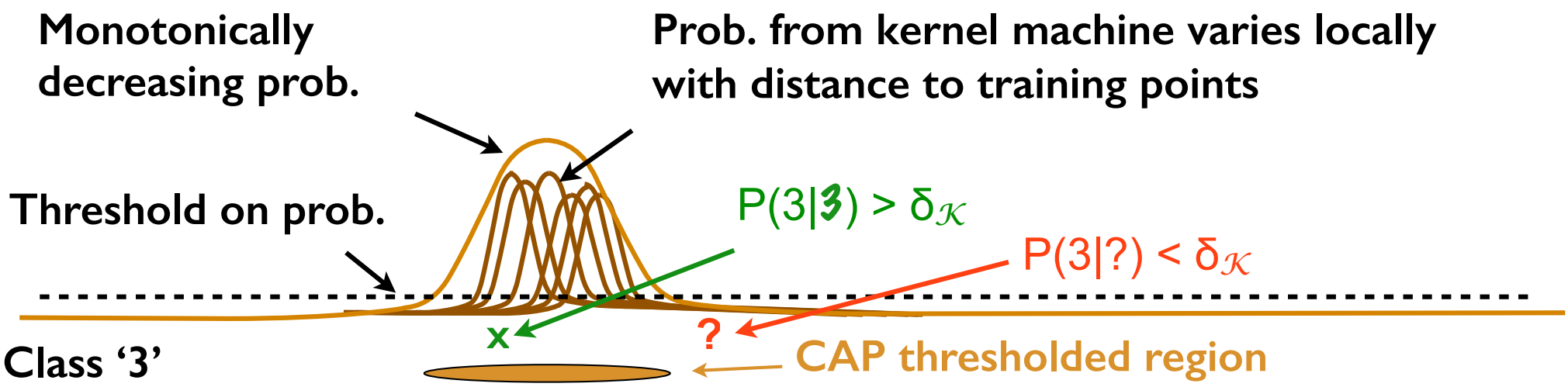
Proof of Theorem 2

Since 1-Class SVM has only positive data, we can view this function as a canonical sum over positive definite kernels. Let $g = \sum_i \vartheta_i = \sum_i \alpha_i y_i$. *Let* $i^* = \operatorname{argmin}_i \|x' - x_i\|$, *then it is sufficient to let* $A_{x'} = gK(x, x_{i^*})$, *which by the theorem's kernel assumption is monotonically decreasing and in the space of square integrable functions. Hence* $gK(x, x_{i^*})$ *is an abating bound function for the sum, yielding a CAP model.*

Goal: Multi-class Open Set Recognition



Model: Compact Abating Probability



Do any of the well known approaches from the literature apply?

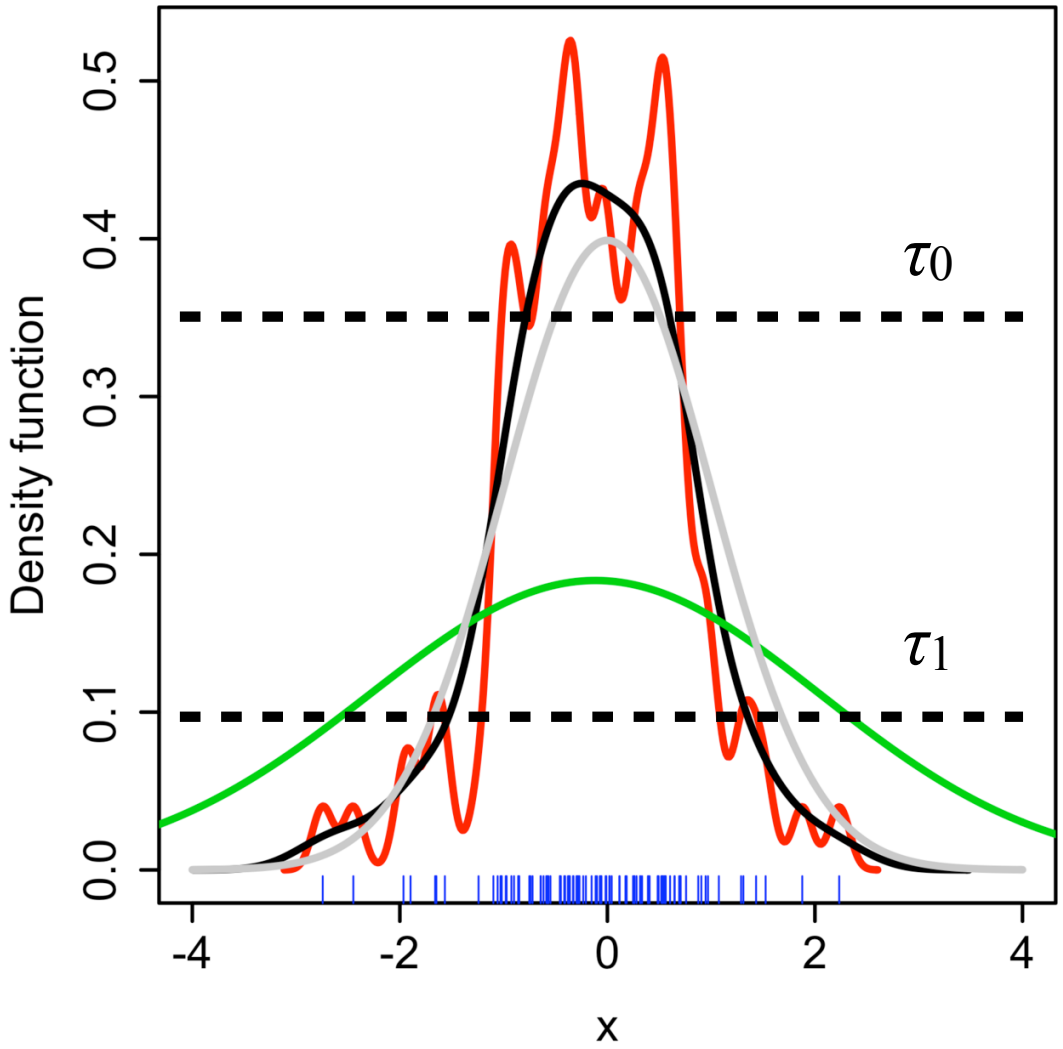
Kernel Density Estimation (KDE)

D.M.J. Tax, Ph.D. Dissertation “One-class classification: Concept learning in the absence of counter-examples” 2001

1. Fit a Gaussian distribution to the positive training data for a class

2. Empirically estimate a threshold τ over the resulting density

Kernel Density Estimation



Comparison of 1D bandwidth selectors

KDE Pitfalls

- Nearly always results in overfitting for visual recognition problems
- Choice of Gaussian distribution questionable in many circumstances
- How do we estimate a good τ ?

Support Vector Data Description (SVDD)

D.M.J. Tax and R.P.W. Duin: Support vector data description. Machine Learning 54, 45–66

- Hypersphere with the minimum radius is estimated around the positive class data that encompasses almost all training points.

Support Vector Data Description (SVDD)

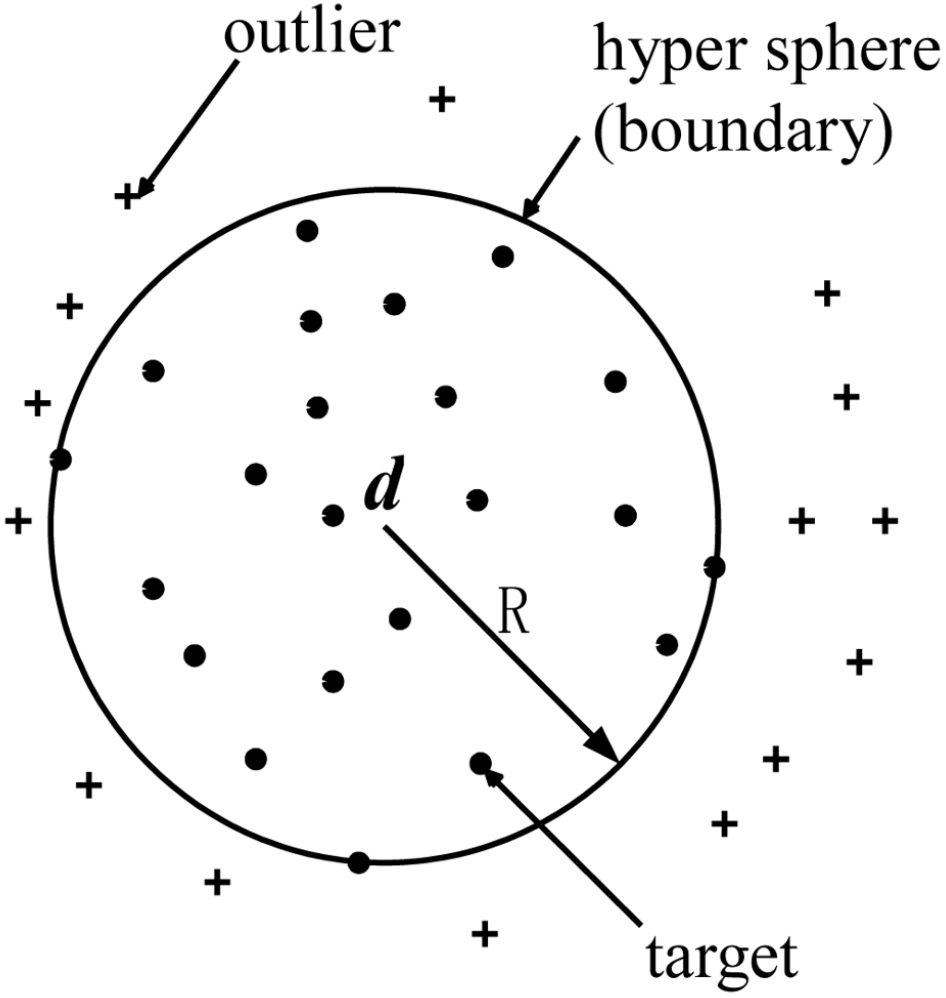


Image credit: Shen et al. *Sensors* 2012

Support Vector Data Description (SVDD)

- Sensitive to feature scaling (Tax and Duin ASCI 2002)
- Difficult to solve using good numerical optimization (Chang et al. NTU Tech. Report 2013)
- Far less effective than binary classifiers when some sampling of negatives is available
  - Overfits the training data

1-Class SVM

- Only positive data at training time
- “Origin” defined by the kernel serves as the only member of a “second class”
- Training object yields a binary classifier f
- When used, usually for outlier or anomaly detection

1-Class SVM

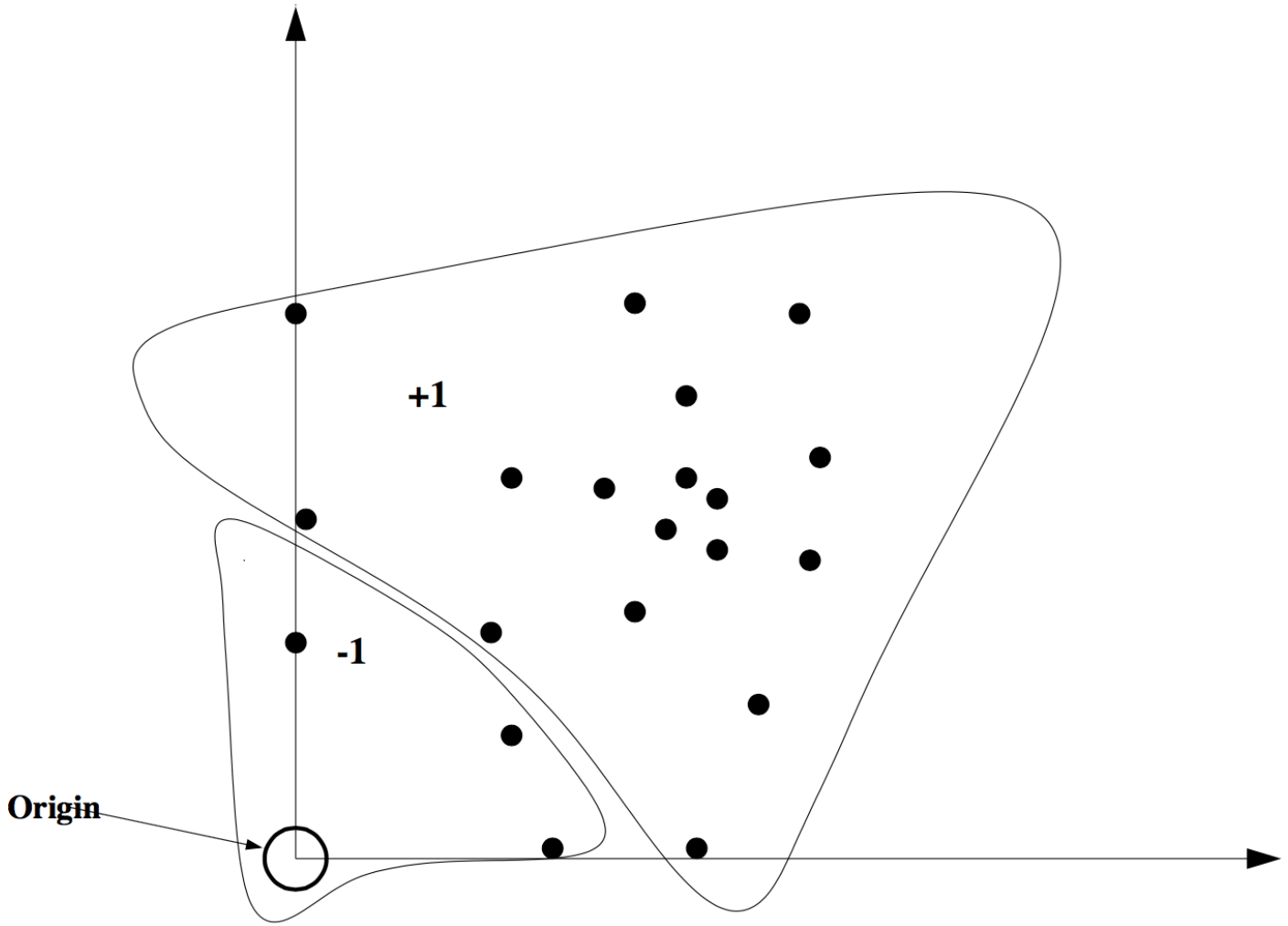
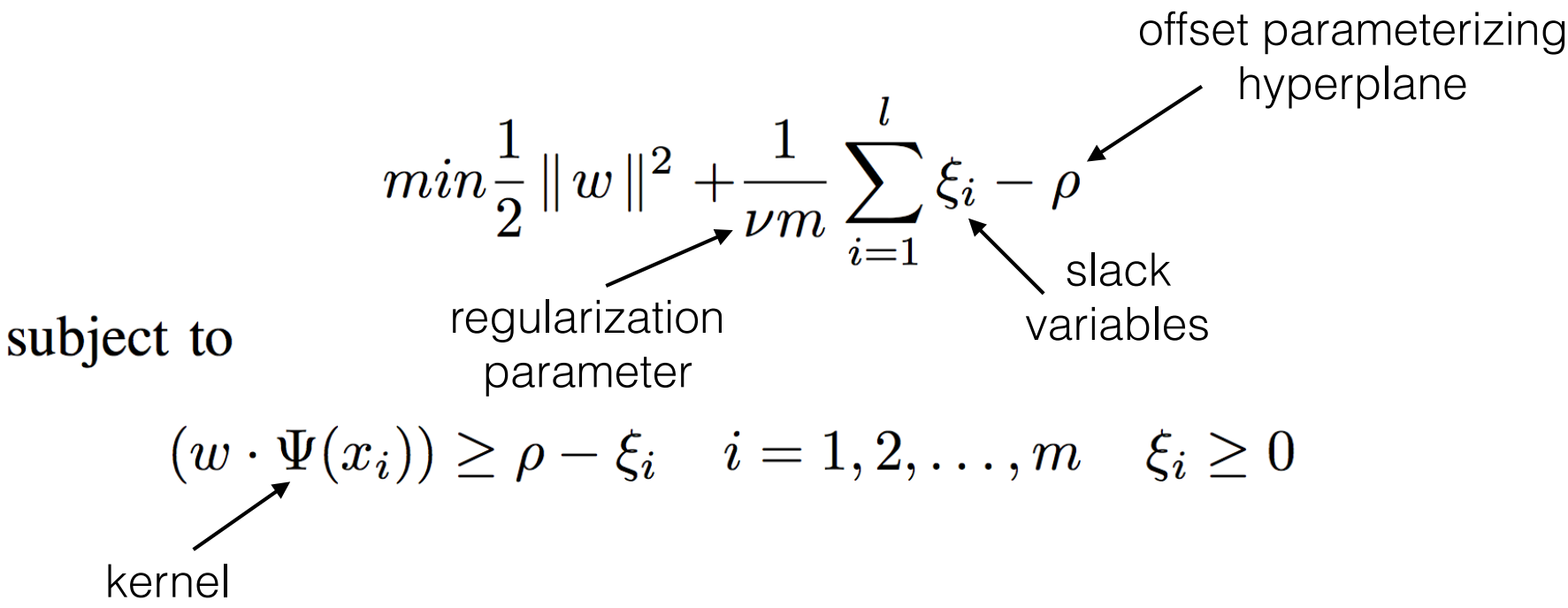


Image credit: L. Manevitz and M. Yousef, “One-Class SVMs for Document Classification” JMLR 2001

1-Class SVM Objective



ν controls the upper bound on training error

1-Class SVM Implementation

LIBSVM (linear and RBF)

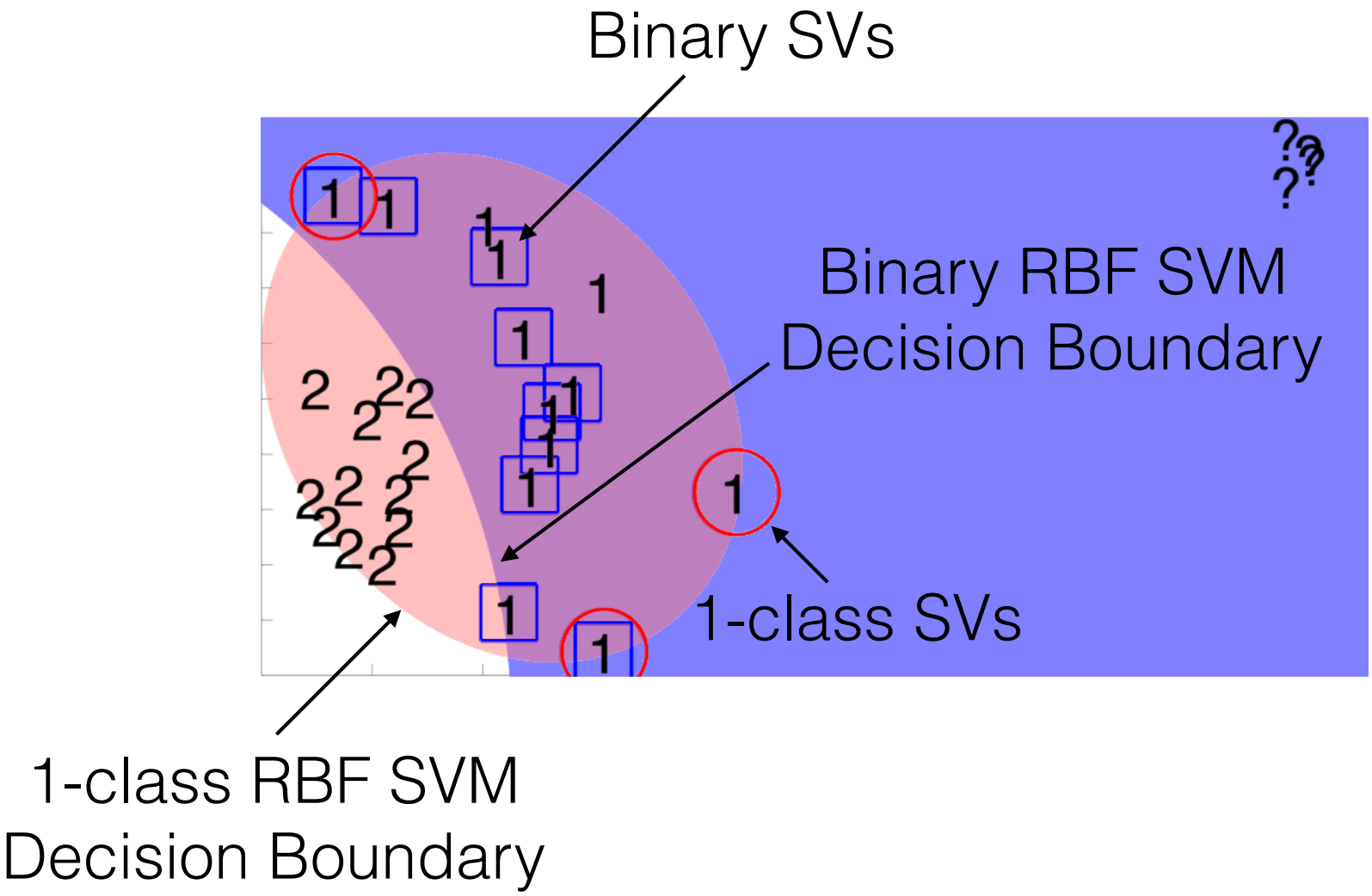
Usage: svm-train [options] training_set_file [model_file]
options:
-s svm_type : set type of SVM (default 0)
0 -- C-SVC (multi-class classification)
1 -- nu-SVC (multi-class classification)
2 -- one-class SVM
3 -- epsilon-SVR (regression)
4 -- nu-SVR (regression)

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

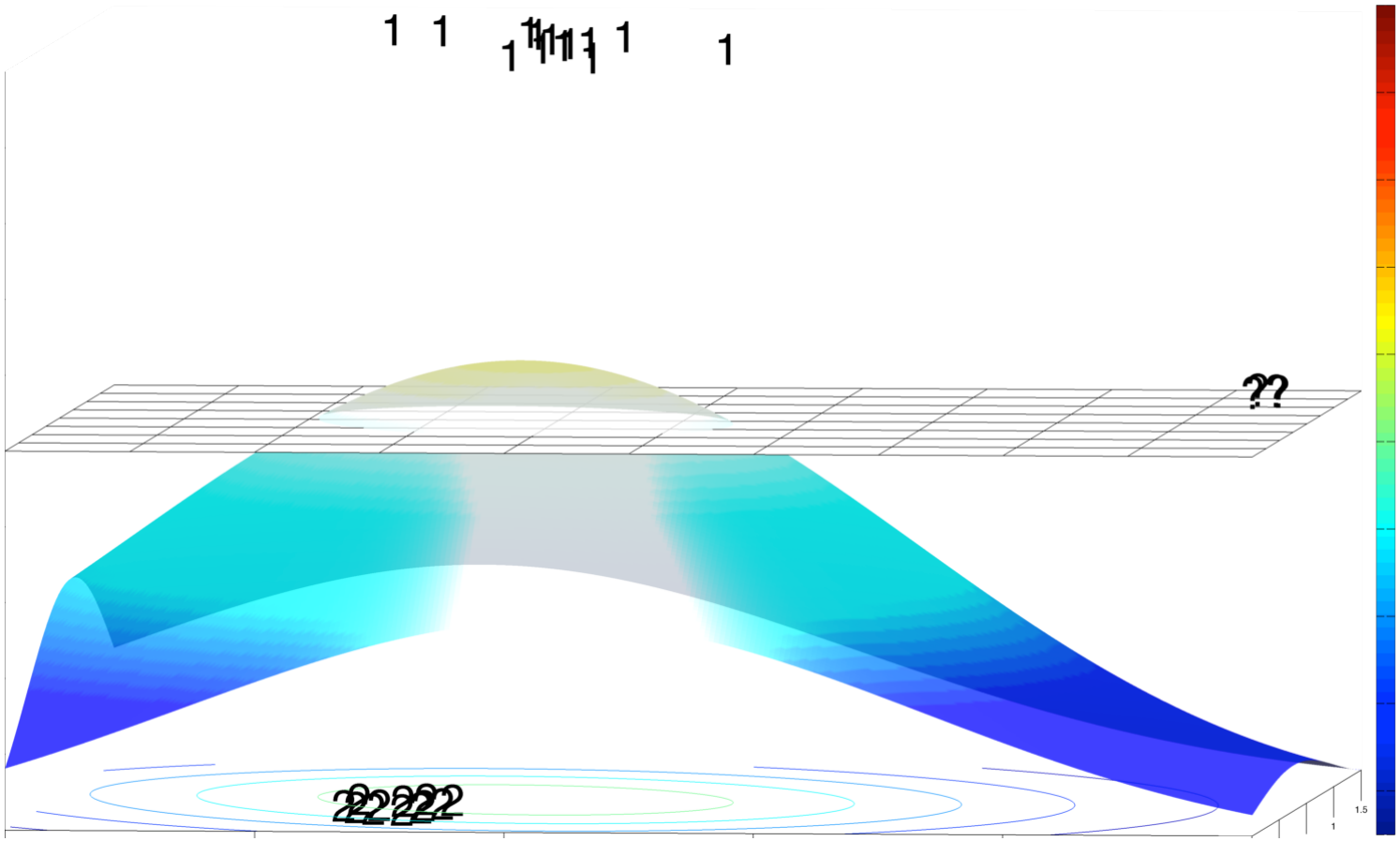
- Zhou and Huang *Multimedia Systems* 2003
  - Kernel and parameter selection
    - Gaussian kernels lead to over-fitting
    - Parameters chosen in *ad hoc* fashion
    - An issue in other domains too!

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

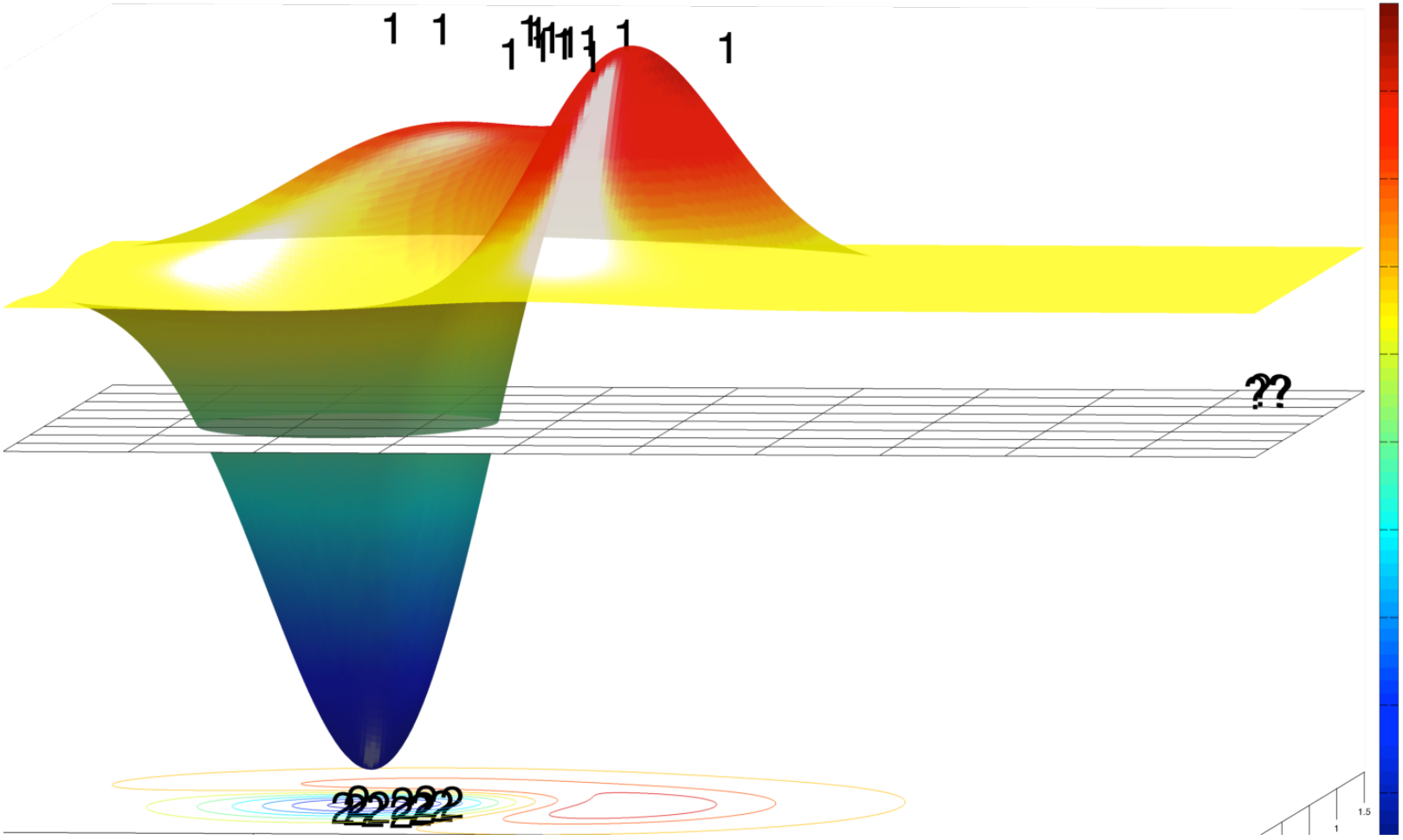
Problems with Existing Models for Binary Problems



Normalized decision scores for 1-Class SVM



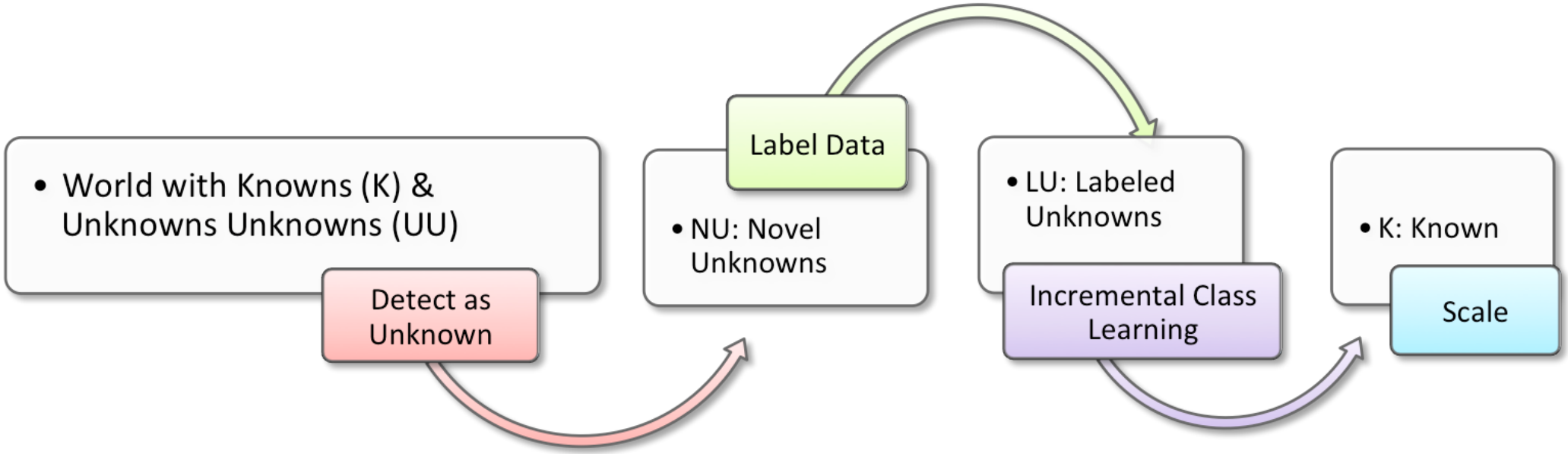
Normalized decision scores for Binary SVM



Other machine learning approaches

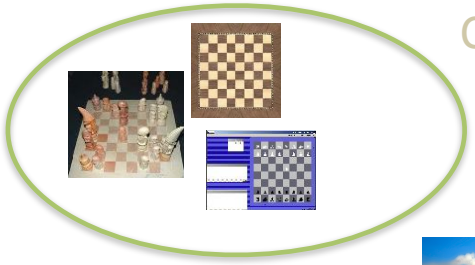
- M. Rohrbach, M. Stark, and B. Schiele, “Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting,” in IEEE CVPR, 2011.
- C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer,” in IEEE CVPR, 2009.
- E. Bart and 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.

Open World Recognition



A. Bendale and T. E. Boult. "Towards open world recognition," CVPR 2015

Out in the Real-World



Chess board

airplanes

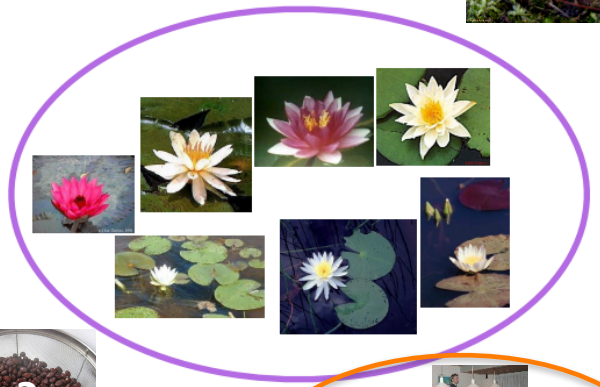
elephant

Boxing glove

soccer ball

Detect New Category

water lilly



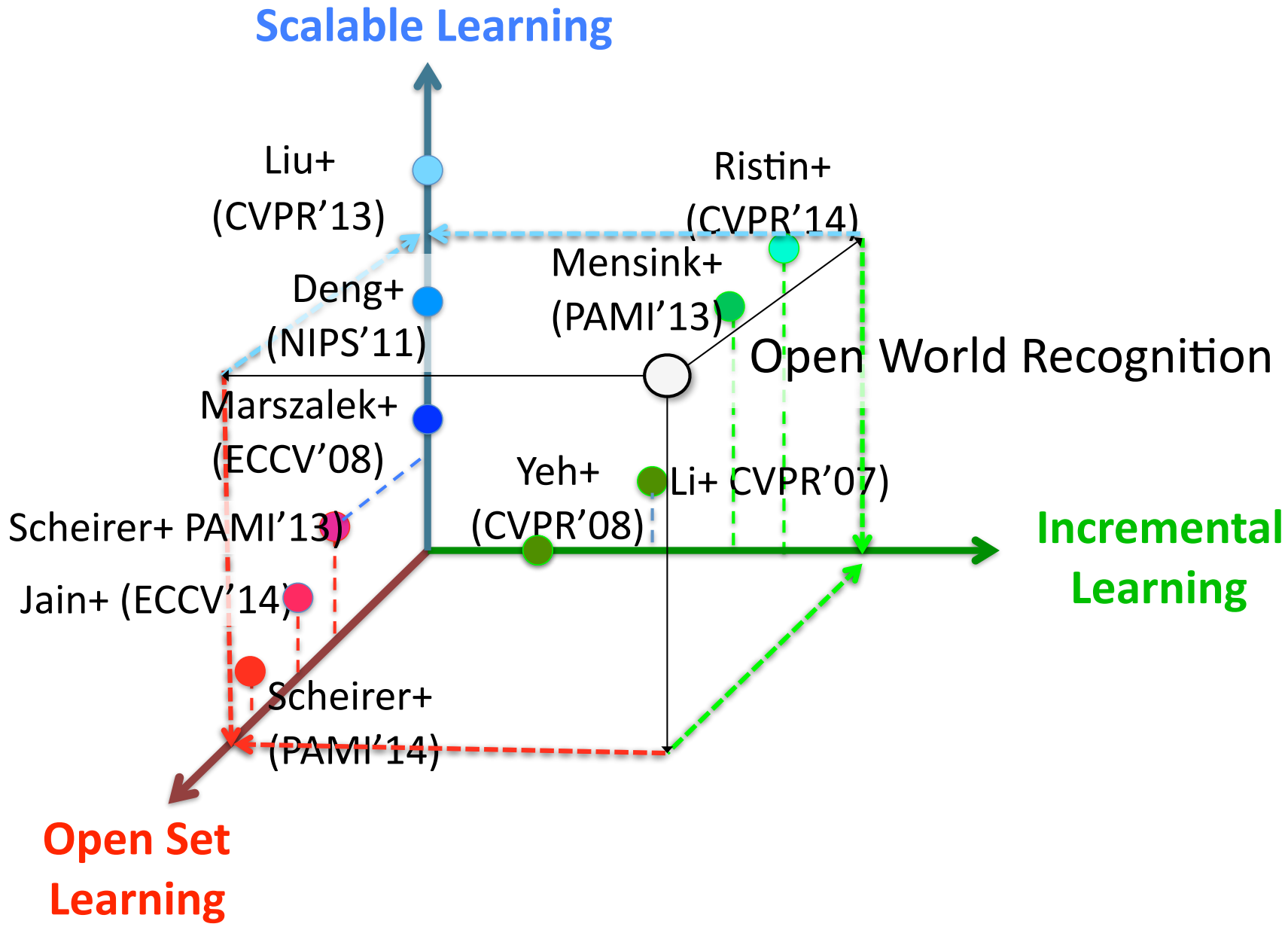
Pool Table

car

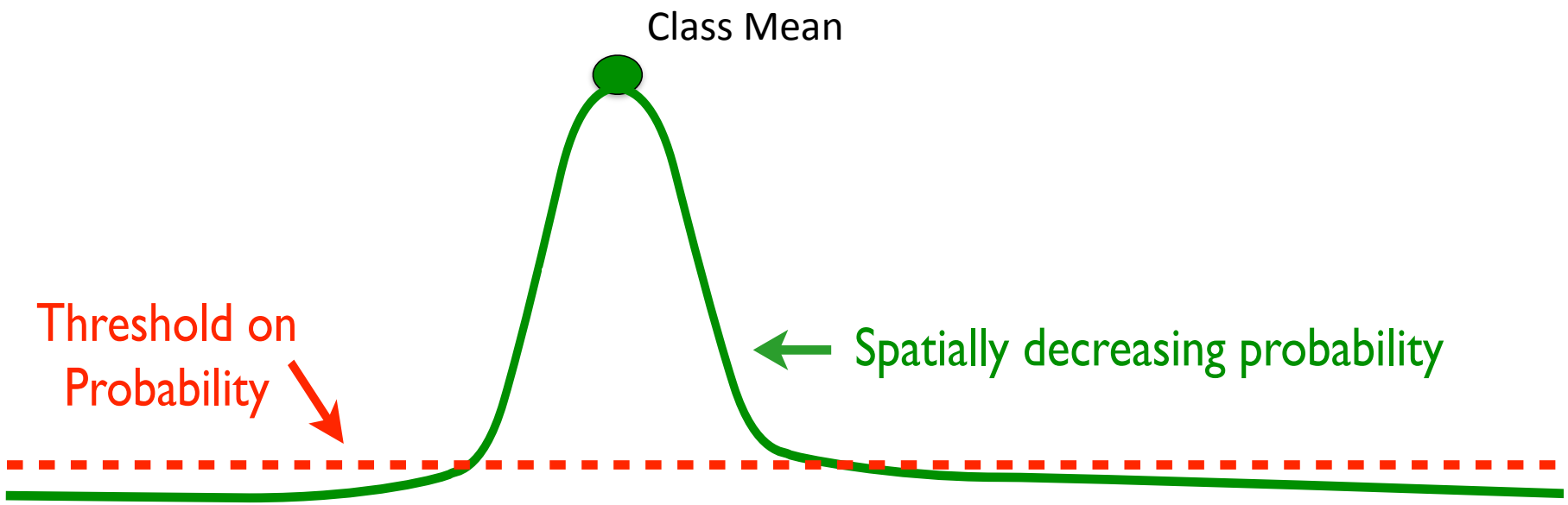
Calculator

Bowling Pin

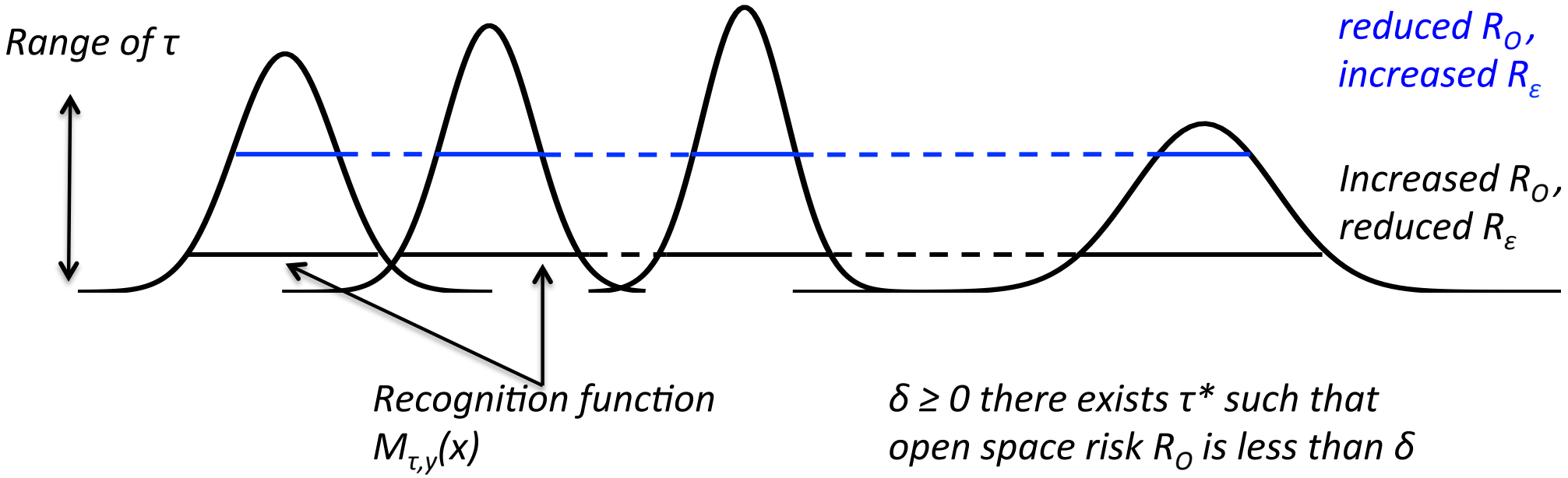
Related Work



Recall the CAP Model:



Theorem on Open Space Risk for Model Combination



Theorem: Open Space Risk for Transformed Spaces