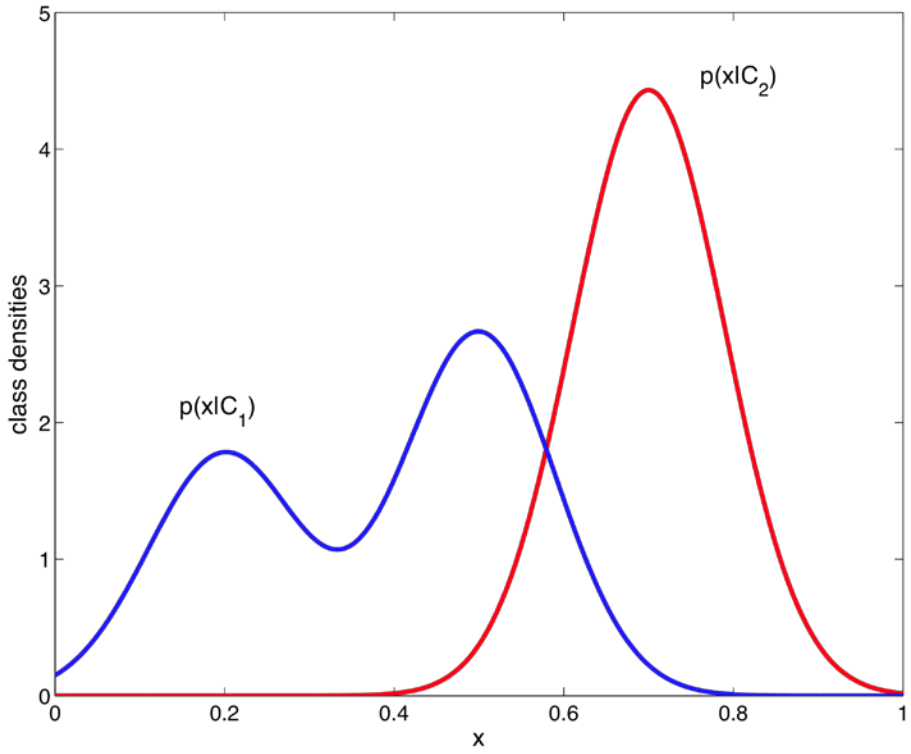
Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011
Prof. Erik Sudderth

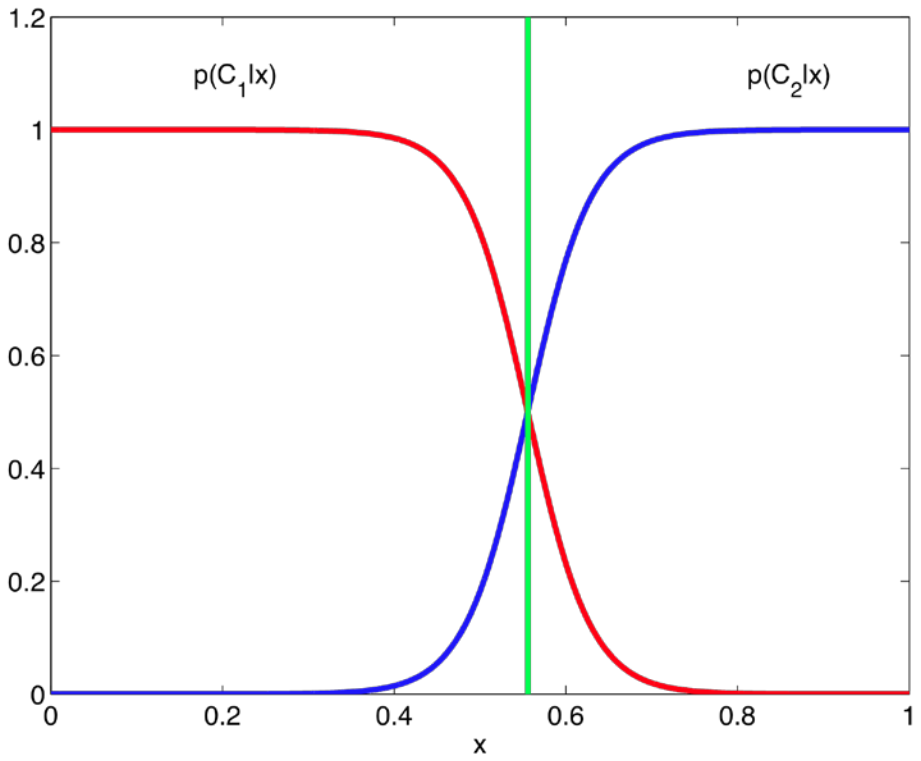
Lecture 6: Decision Theory, Model Selection & Validation

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Binary MAP Estimation



Likelihood Functions



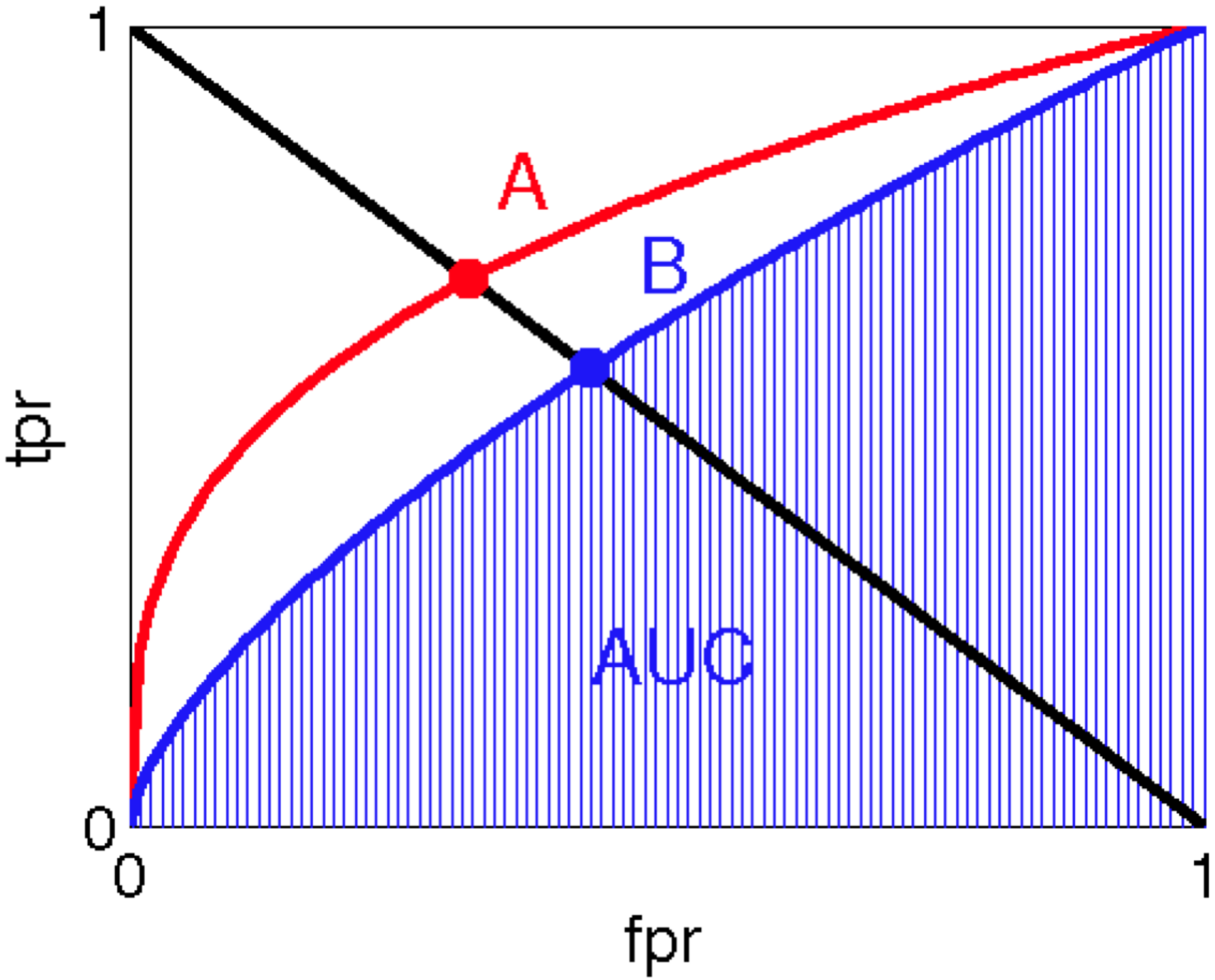
Posterior Probabilities

False Positives vs. False Negatives

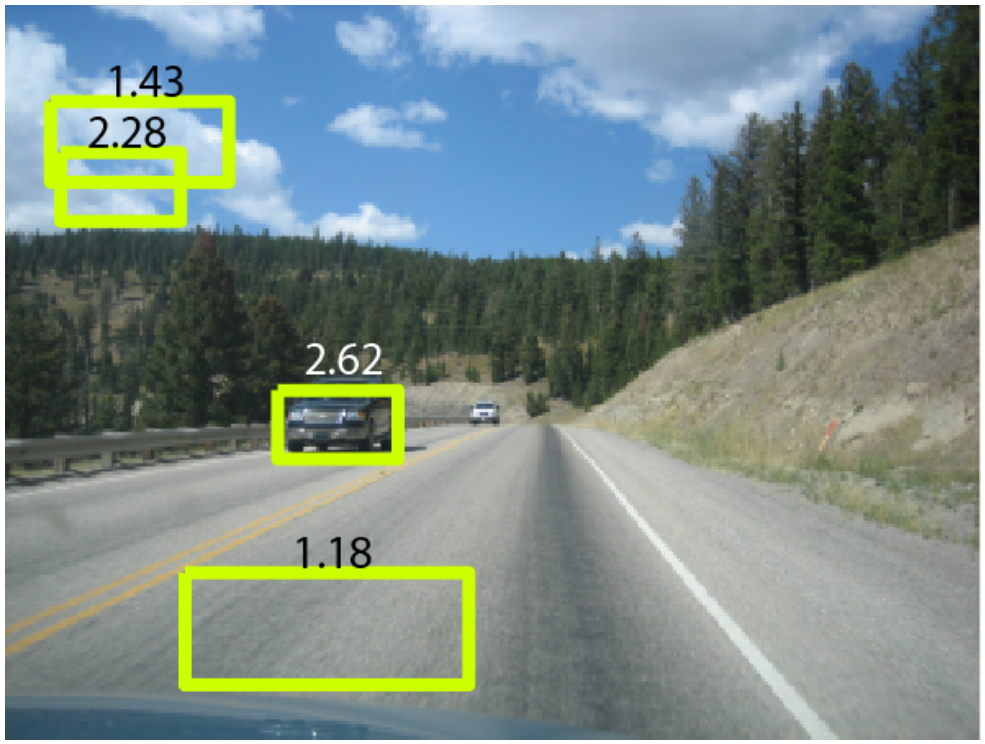
		Truth		
		1	0	Σ
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = FN + TN$
	Σ	$N_+ = TP + FN$	$N_- = FP + TN$	$N = TP + FP + FN + TN$

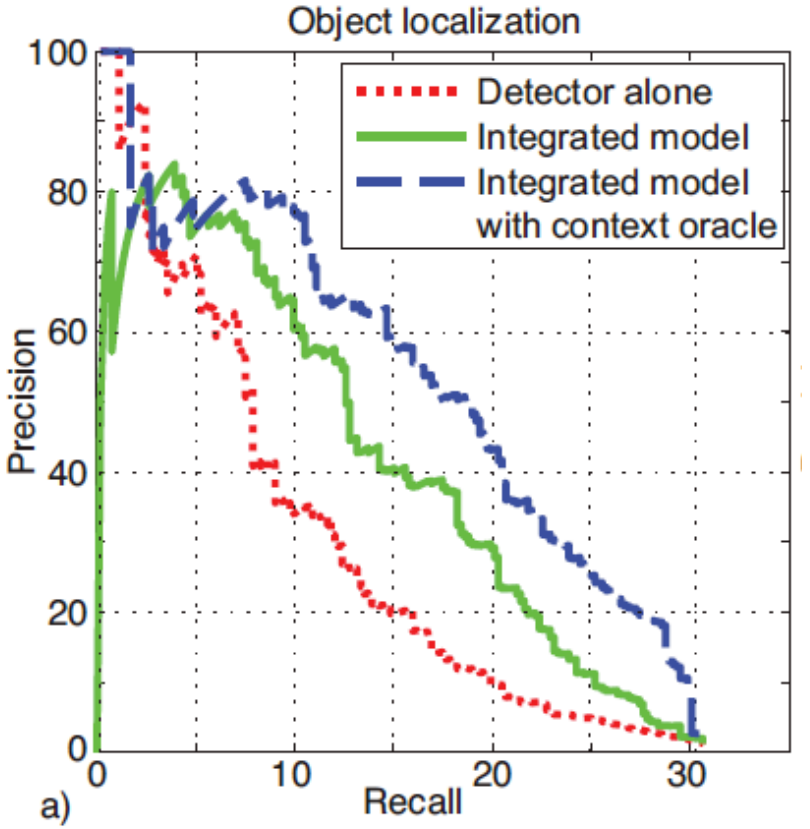
	$y = 1$	$y = 0$	$y = 1$	$y = 0$
$\hat{y} = 1$	$TP/\hat{N}_+$=precision	$FP/\hat{N}_+$=FDP	TP/N_+=TPR	FP/N_-=FPR
$\hat{y} = 0$	$FN/\hat{N}_-$	$TN/\hat{N}_-$=NPV	FN/N_+=FNR	TN/N_-=TNR

ROC Curves



Example: Object Detection



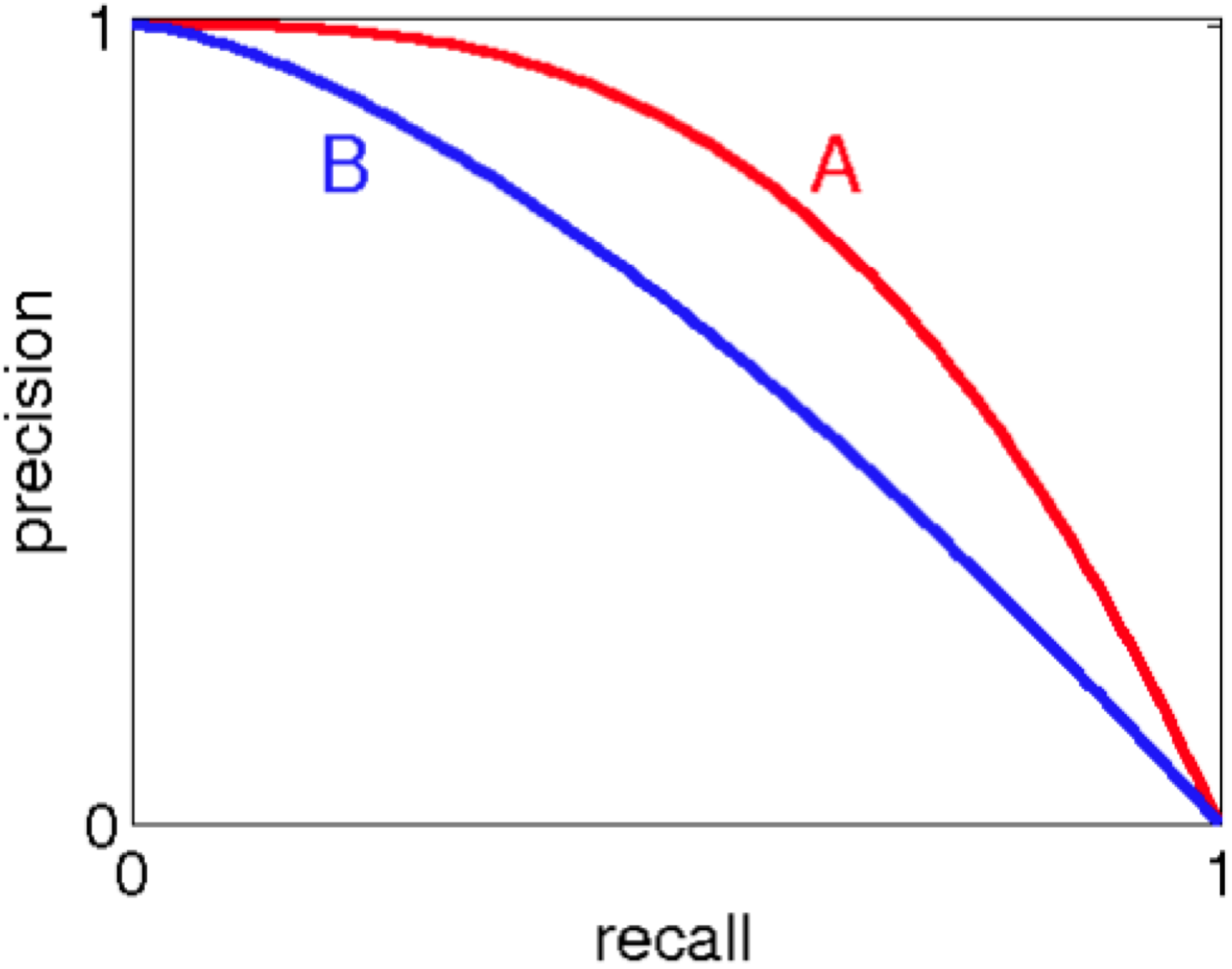


Fei-Fei, Fergus, Torralba, ICCV 2009

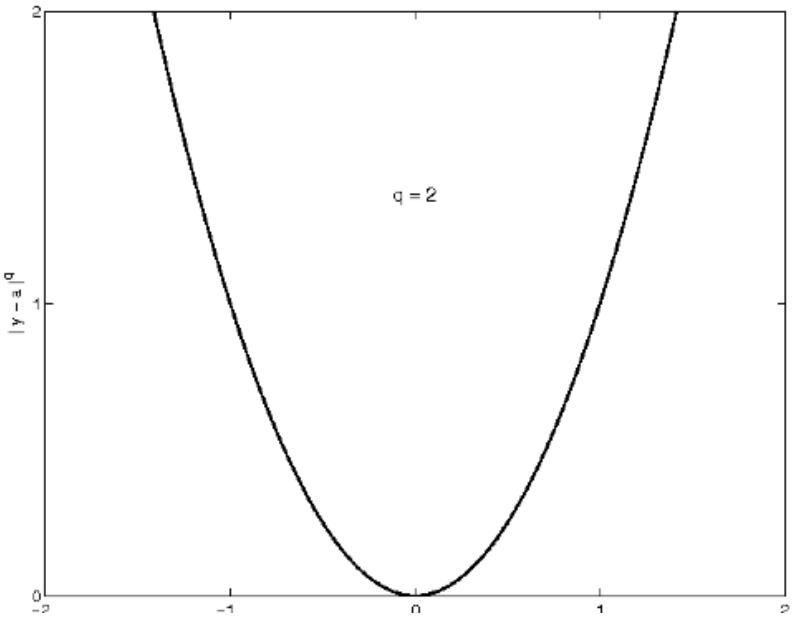
The number of *negative* examples may not be well defined:

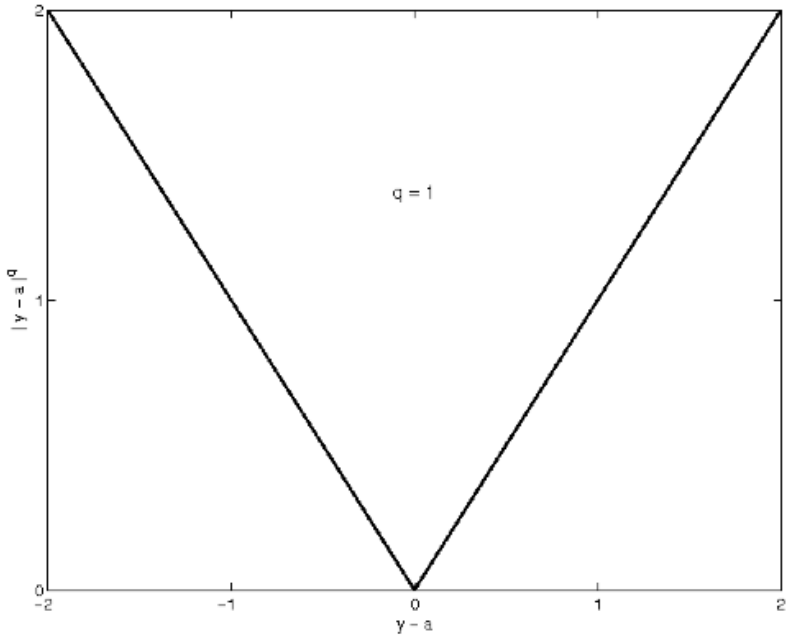
- How many windows not containing a car are there in an image?
- How many documents not about cars exist in the world?

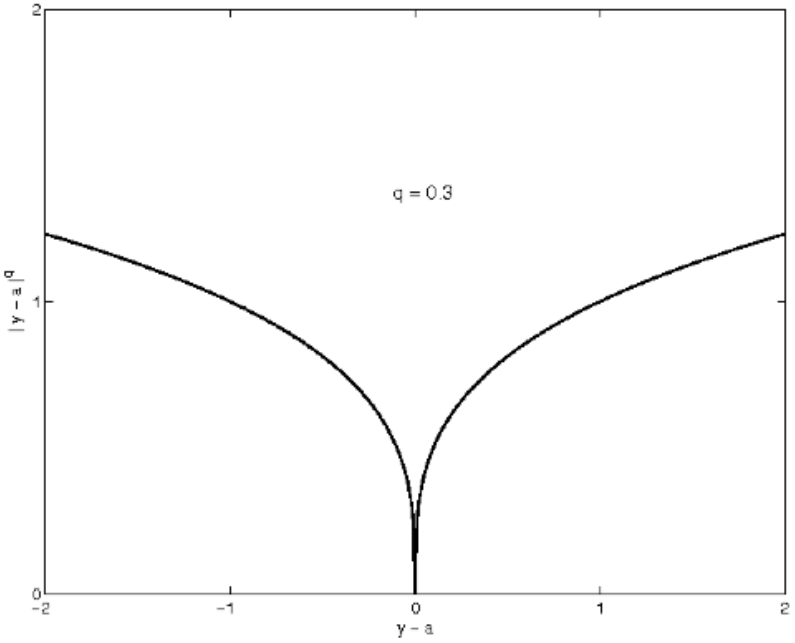
Idealized Precision-Recall Curve

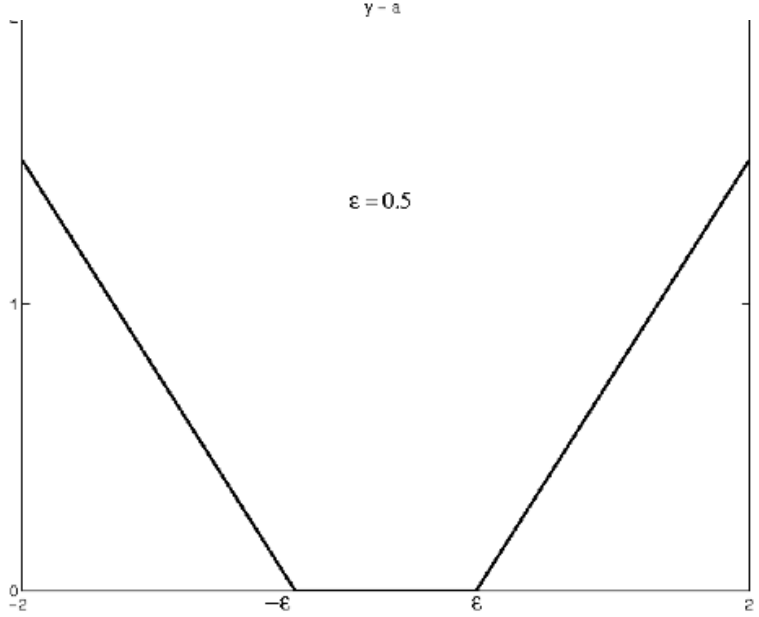


Continuous Loss Functions









What are Good Loss Functions?

Bayesian color constancy

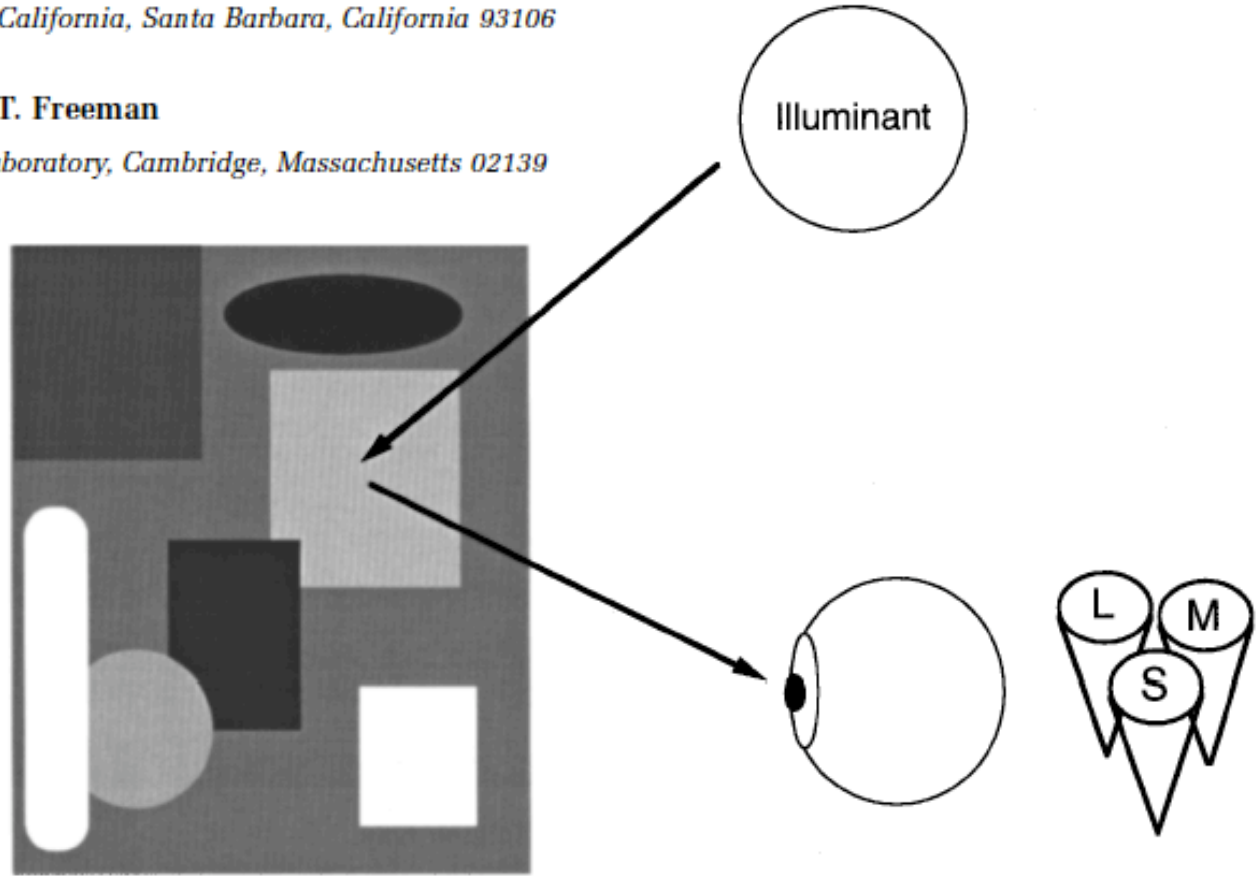
Journal of the Optical Society of America A, July 1997

David H. Brainard

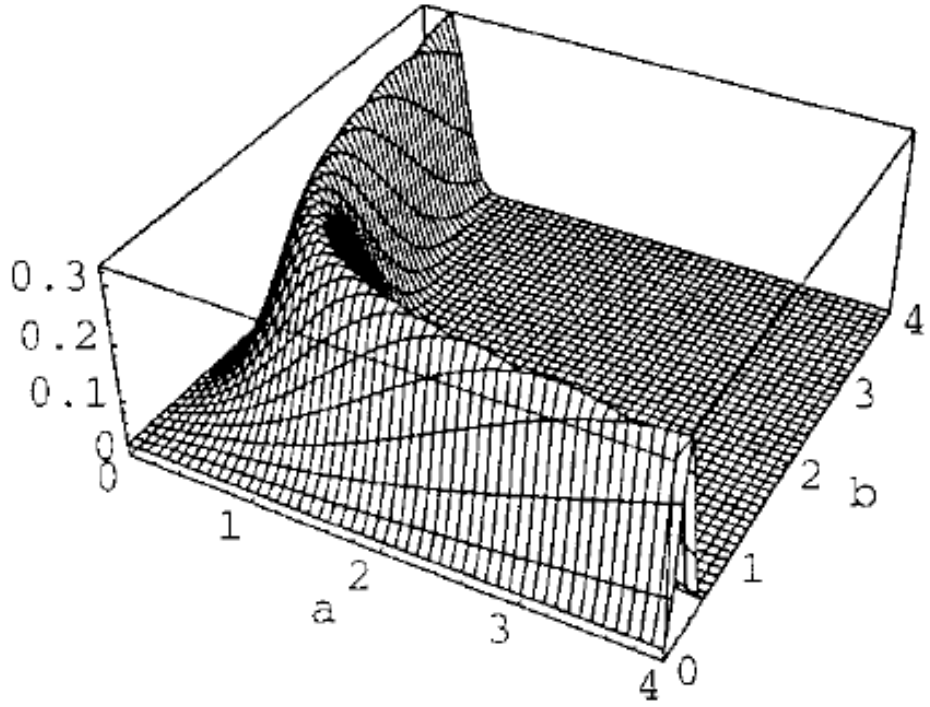
Department of Psychology, University of California, Santa Barbara, California 93106

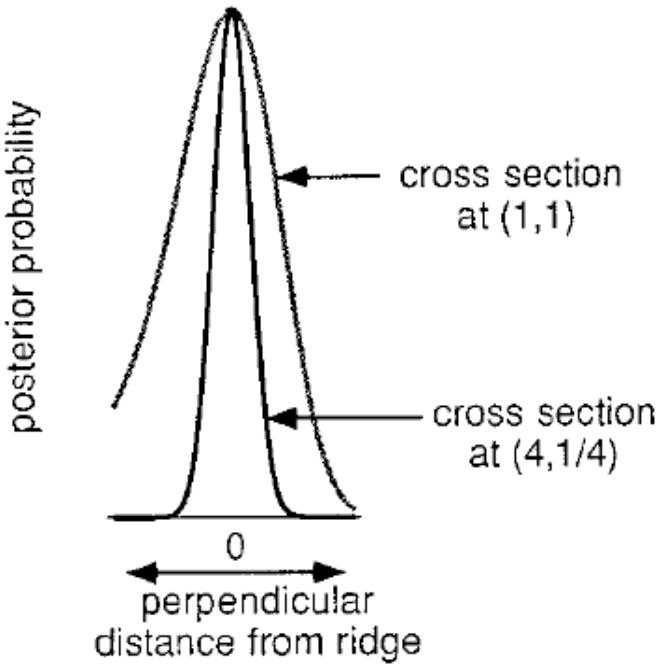
William T. Freeman

MERL, a Mitsubishi Electric Research Laboratory, Cambridge, Massachusetts 02139

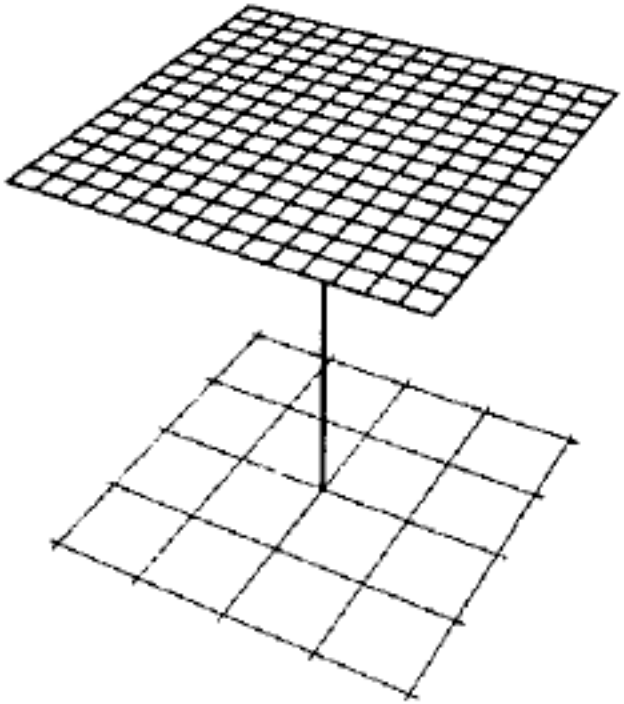


Toy Example

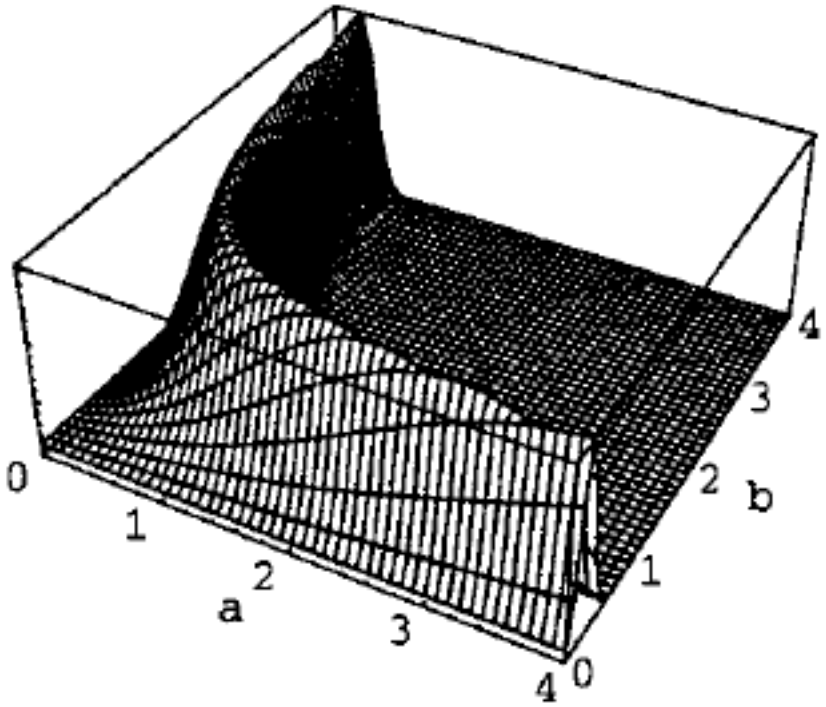

0.3
0.2
0.1
0
0
1
2
3
4
a
0
1
2
3
4
b


posterior probability
cross section at (1,1)
cross section at (4,1/4)
0
perpendicular distance from ridge

MAP Loss Function

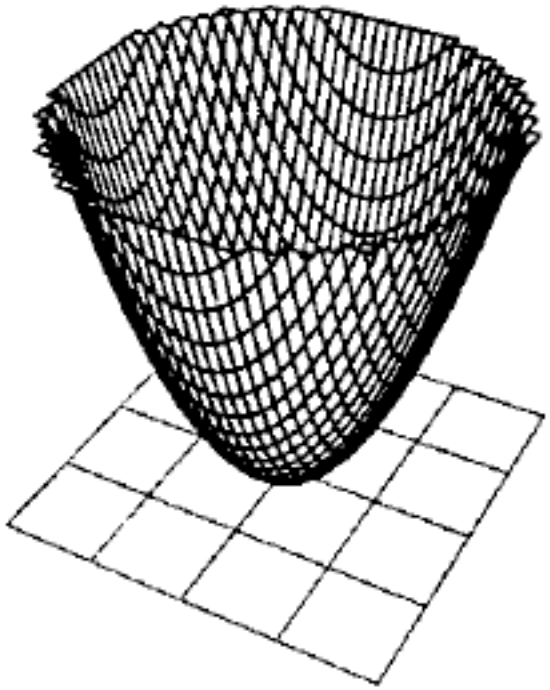


(a) MAP loss function

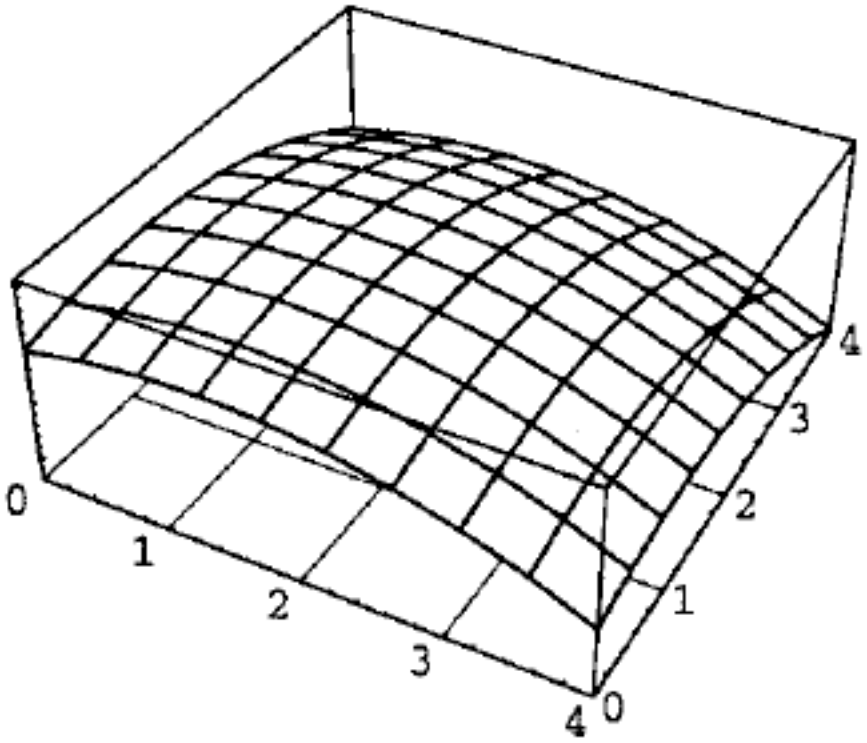


(d) (minus) MAP expected loss

Quadratic Loss Function

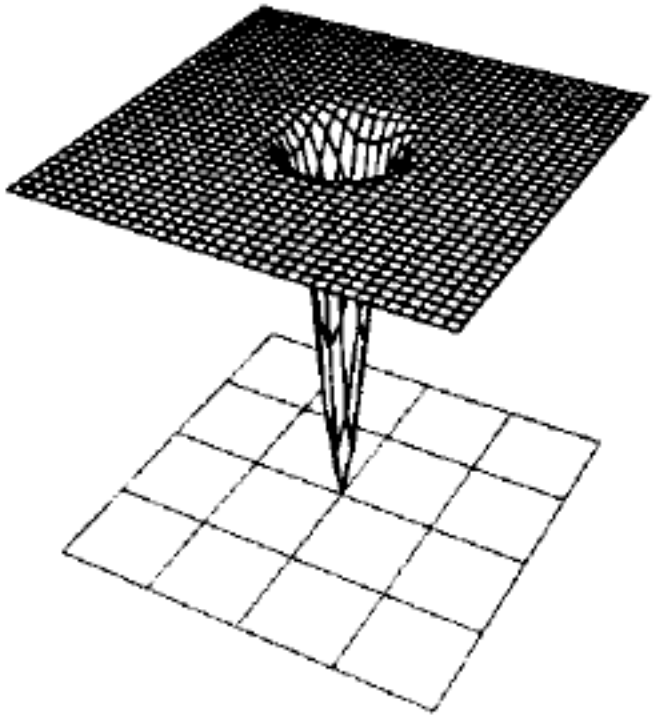


(b) MMSE loss function

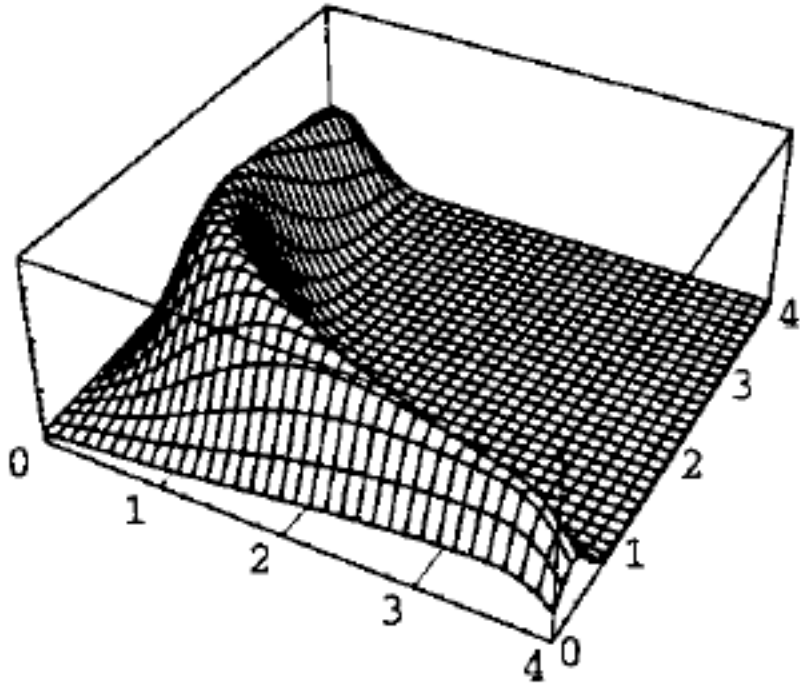


(e) (minus) MMSE expected loss

Local Mass Loss Function

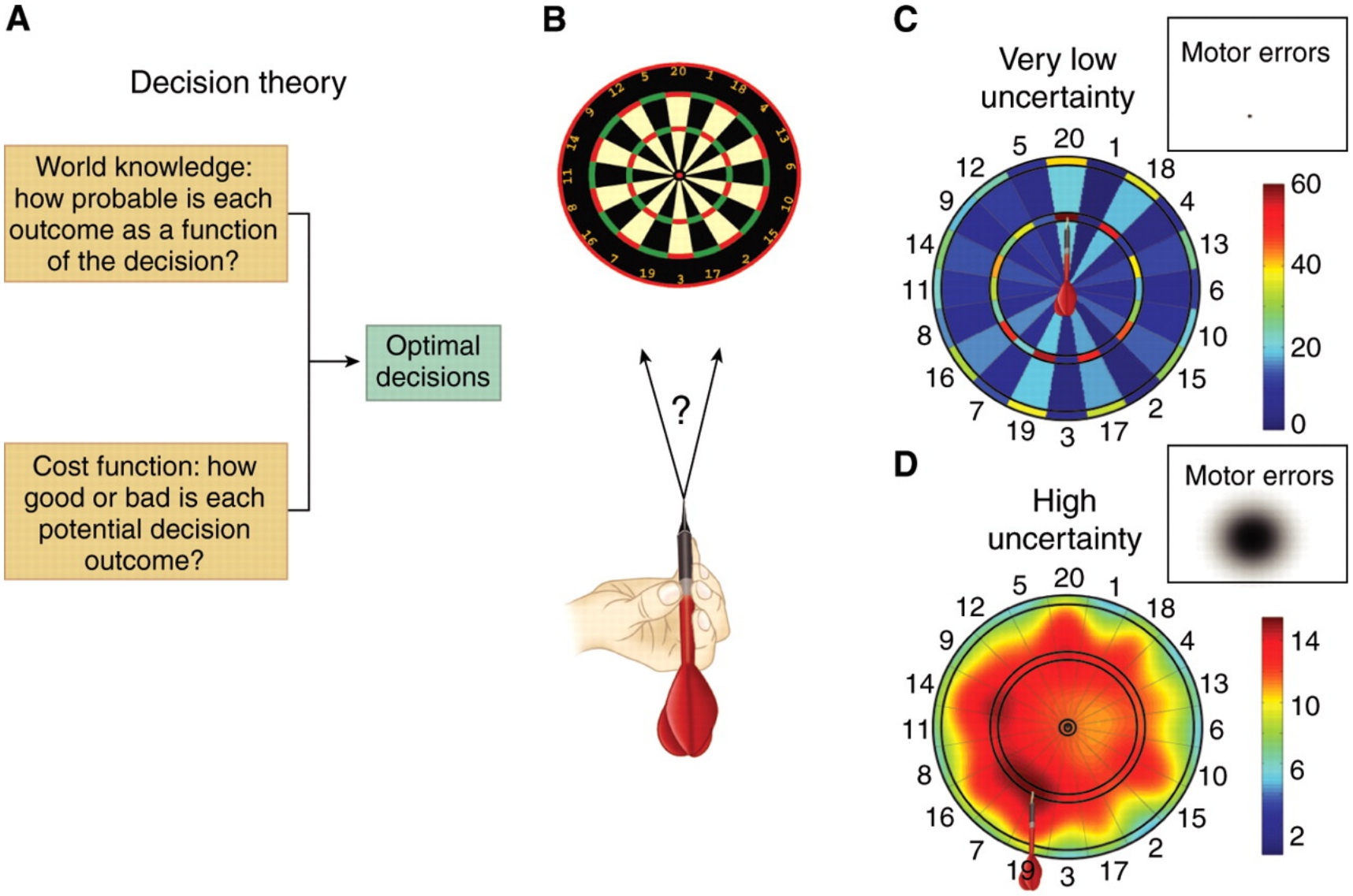


(c) MLM loss function



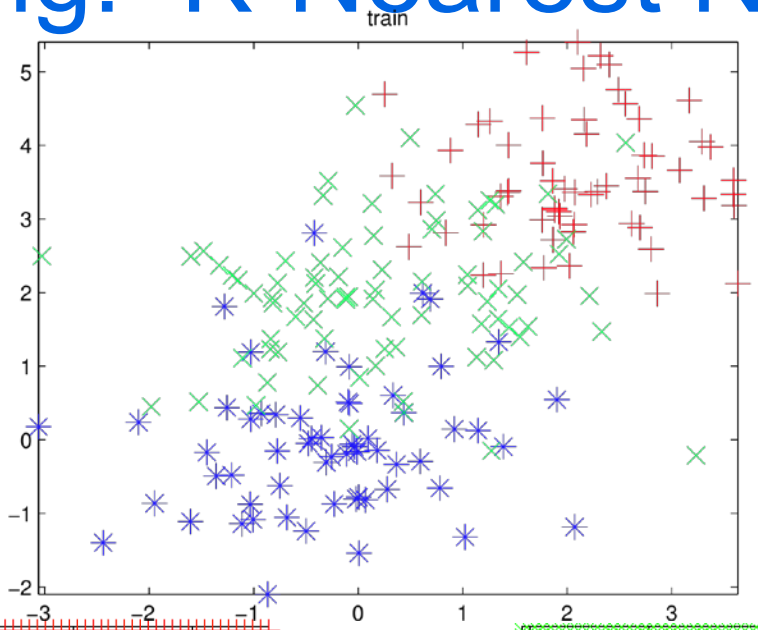
(f) (minus) MLM expected loss

Modeling Human Decisions

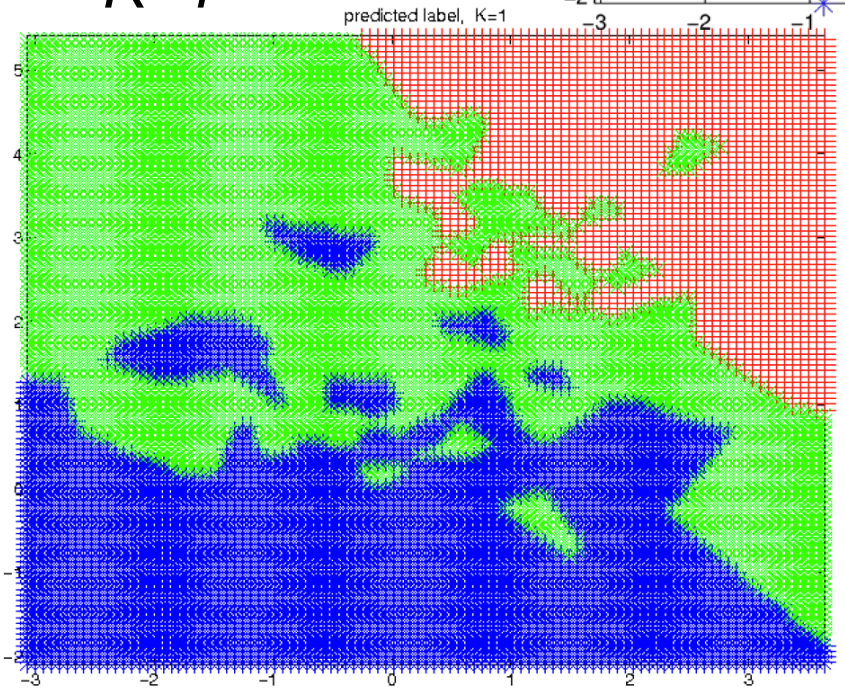


Koerding, Science Magazine, Oct. 2007

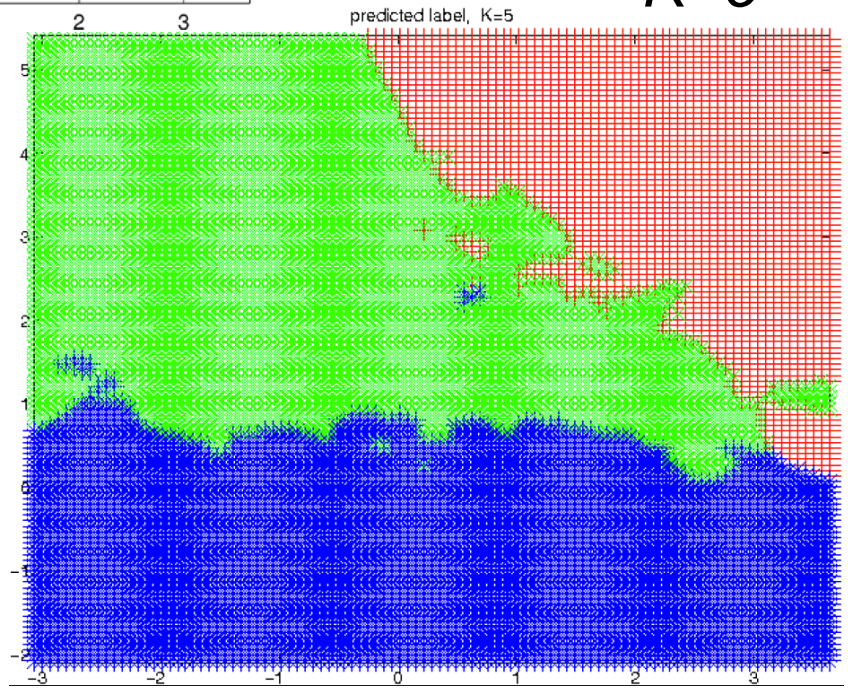
Overfitting: K-Nearest Neighbors



K=1



K=5



Training and Test Data

- Several candidate learning algorithms or models, each of which can be fit to data and used for prediction
- How can we decide which is best?

Approach 1: Split into train and test data

- Learn parameters of each model from training data
- Evaluate all models on test data, and pick best performer

Problem:

- Over-estimates test performance ("lucky" model)
- Learning algorithms can *never* have access to test data

Training, Test, and Validation Data

- Several candidate learning algorithms or models, each of which can be fit to data and used for prediction
- How can we decide which is best?

Approach 2: Reserve some data for validation

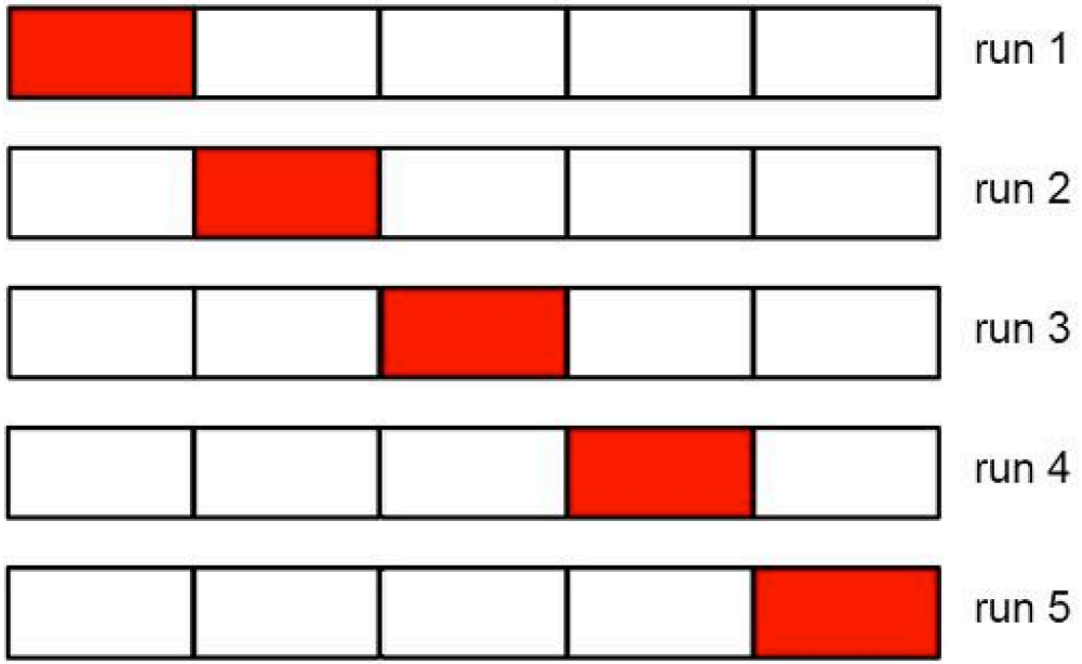
- Learn parameters of each model from training data
- Evaluate models on validation data, pick best performer

Problem:

- Wasteful of training data (learning can't use validation)
- May bias selection towards overly simple models

Cross-Validation

- Divide training data into K equal-sized *folds*
- Train on K-1 folds, evaluate on remainder
- Pick model with best average performance across K trials



How many folds?

- *Bias:* Too few, and effective training dataset much smaller
- *Variance:* Too many, and test performance estimates noisy
- *Cost:* Must run training algorithm once per fold, expensive
- *Practical rule of thumb:* 5-fold or 10-fold cross-validation
- *Theoretically troubled:* Leave-one-out cross-validation, K=N

Model Selection: Bayes' Factors

$$BF_{1,0} := \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)}$$

Bayes factor $BF(1,0)$	Interpretation
$B < \frac{1}{100}$	Decisive evidence for H_0
$B < \frac{1}{10}$	Strong evidence for H_0
$\frac{1}{10} < B < \frac{1}{3}$	Moderate evidence for H_0
$\frac{1}{3} < B < 1$	Weak evidence for H_0
$1 < B < 3$	Weak evidence for H_1
$3 < B < 10$	Moderate evidence for H_1
$B > 10$	Strong evidence for H_1
$B > 100$	Decisive evidence for H_1

As suggested by Jeffreys. Caveats: Can exhibit sensitivity to choice of priors for each model's parameters.
Most reliable when comparing pairs of "similar" models.

Bayesian Ockham's Razor

"Plurality must never be posited without necessity."

William of Ockham