Evaluation

- Problems
  - Accuracy: for imbalanced data (skewed class distribution)
  - Cost of errors (misclassification)
- Visualization of performance
  - ROC curves: false positive rate vs. true positive rate
  - Cost curves: Expected cost vs. misclassification cost * class distribution
  - …..

Contingency matrix

$$TPR = \frac{\#TP}{\#P} = \frac{\#TP}{\#TP + \#FN}$$

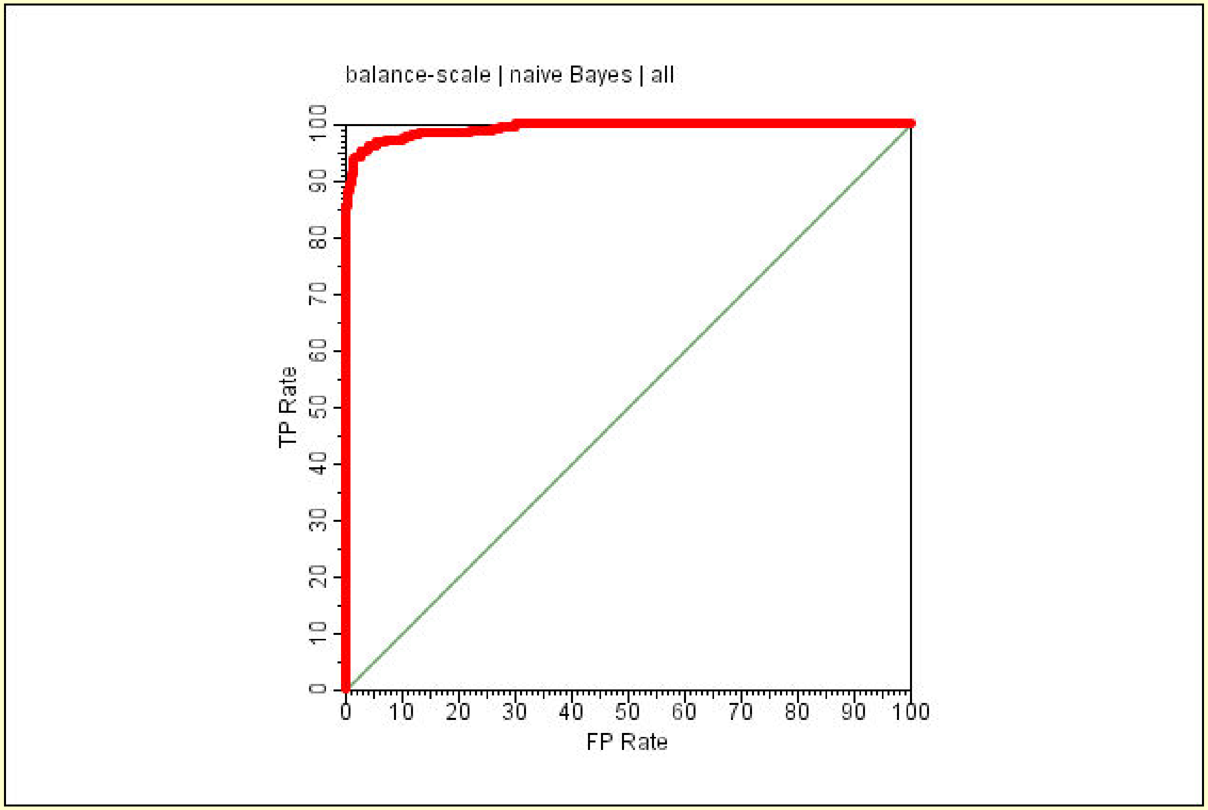
$$FPR = \frac{\#FP}{\#N} = \frac{\#FP}{\#FP + \#TN}$$

		Predicted	
		Positive	Negative
True	Positive	#TP	#FN
	Negative	#FP	#TN

ROC (Receiver Operating Characteristics) curves

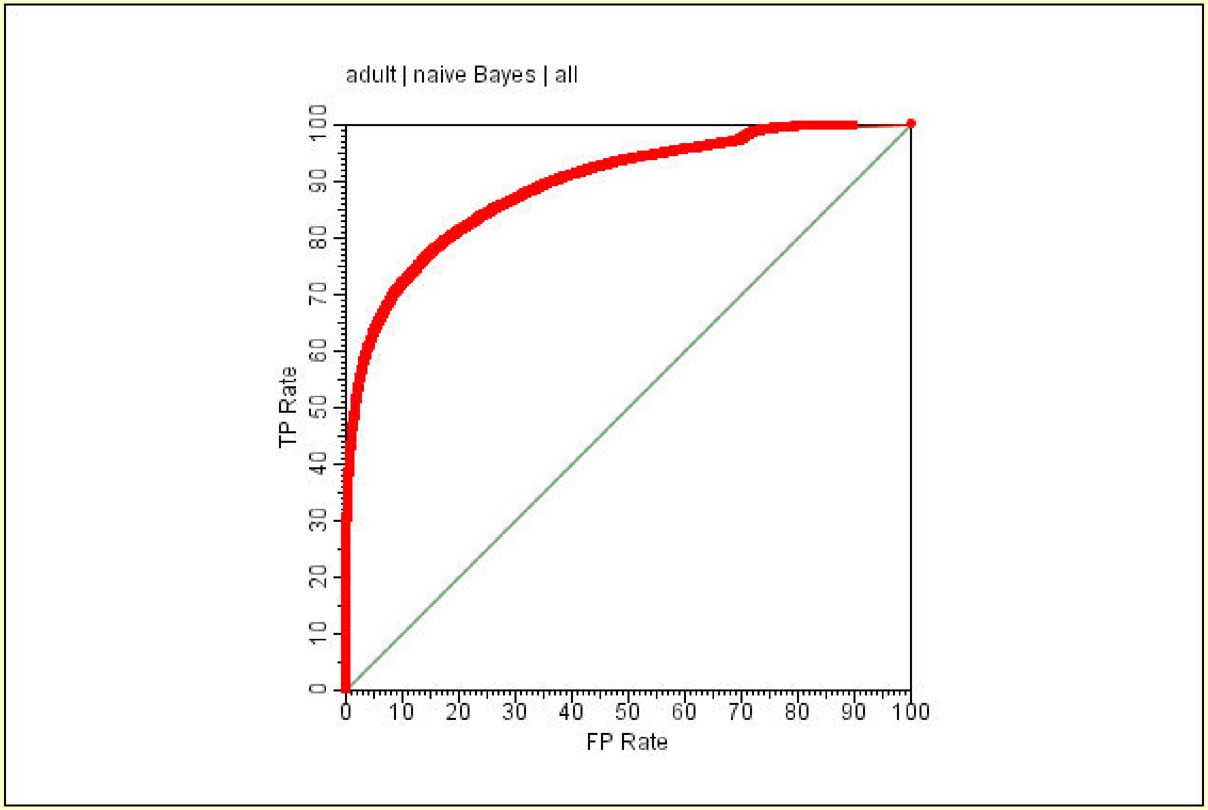
- TPR against FPR
- Classifier = point on the ROC graph
- Distribution-independent
- Ideal = <0, 1> (dominance: North-West)
- Linear interpolation: performance of a point *between* two classifiers
- Convex hull: non-dominated classifiers and interpolation between them
- Trivial=<0,0>, or <1,1> or <x,x>

Example of ROC curves



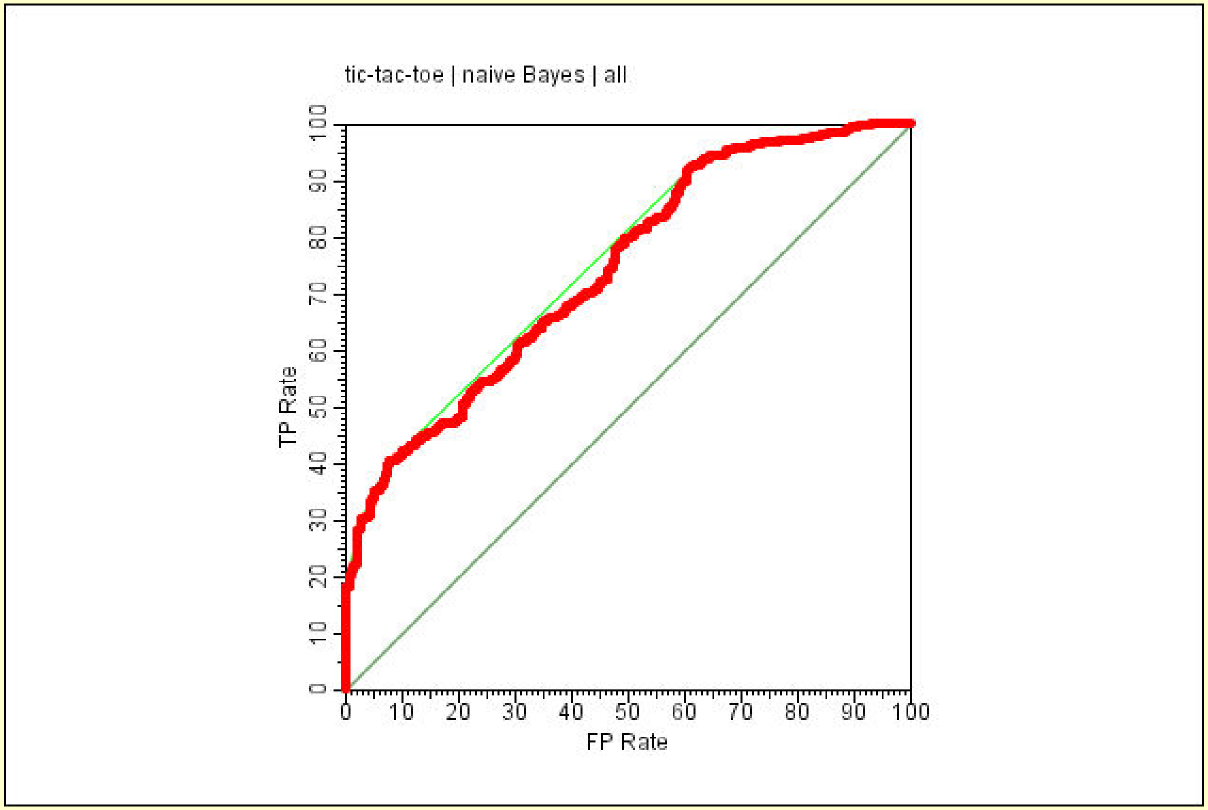
– Good separation between classes, convex curve

Example of ROC curves



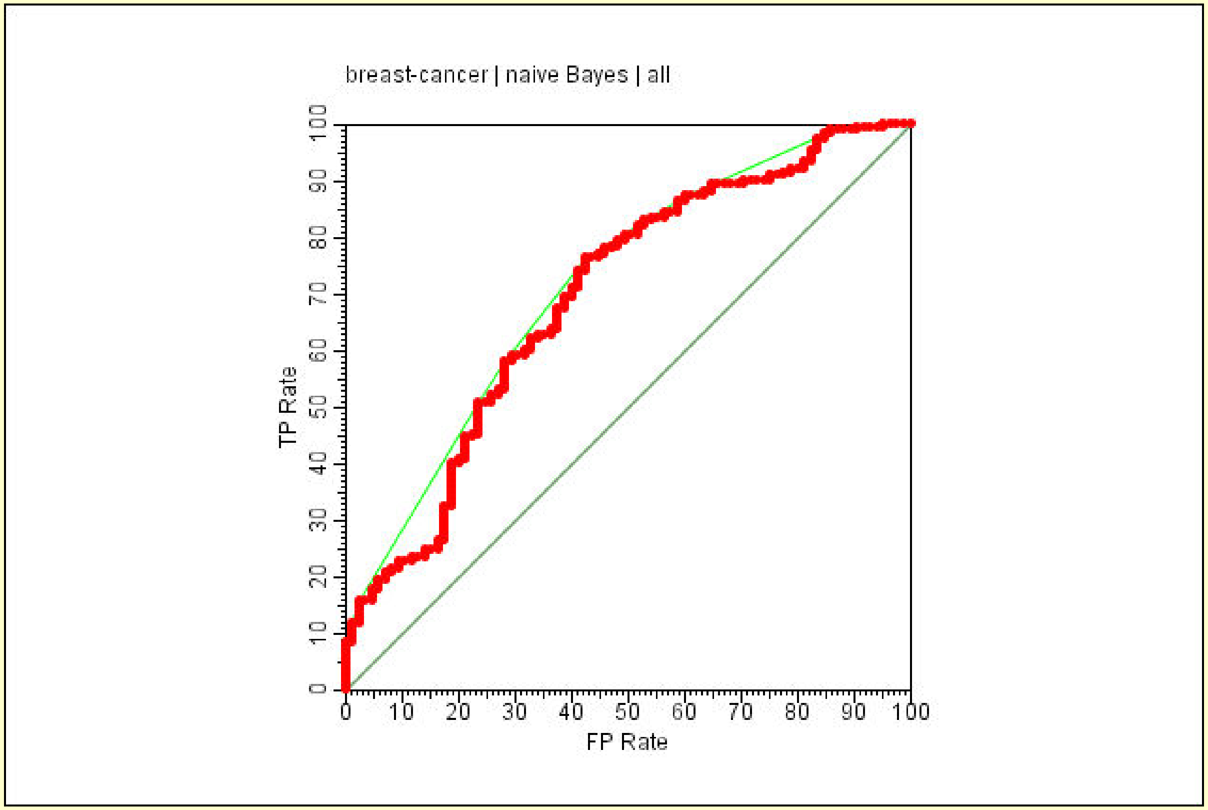
– Reasonable separation, mostly convex

Example of ROC curves



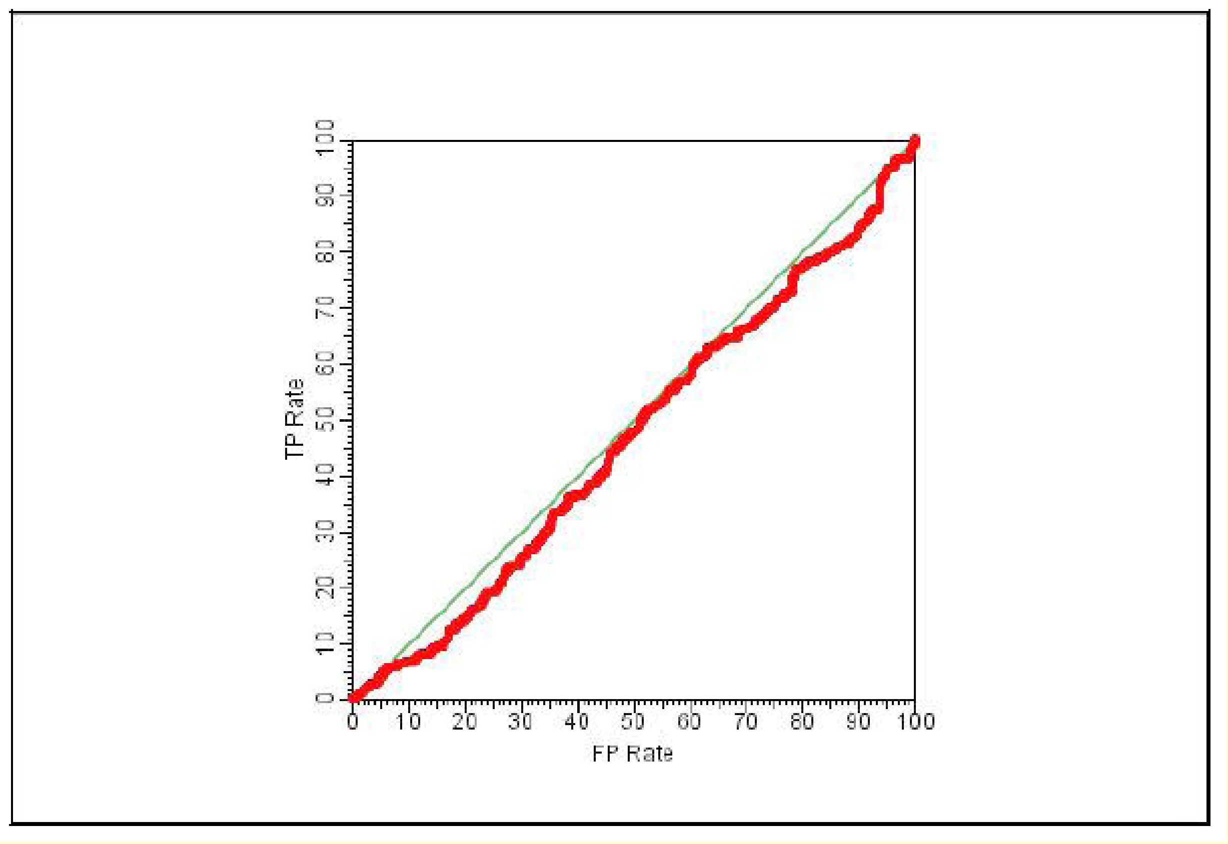
– Fairly poor separation, mostly convex

Example of ROC curves



– Poor separation, large and small concavities

Example of ROC curves



– Random performance

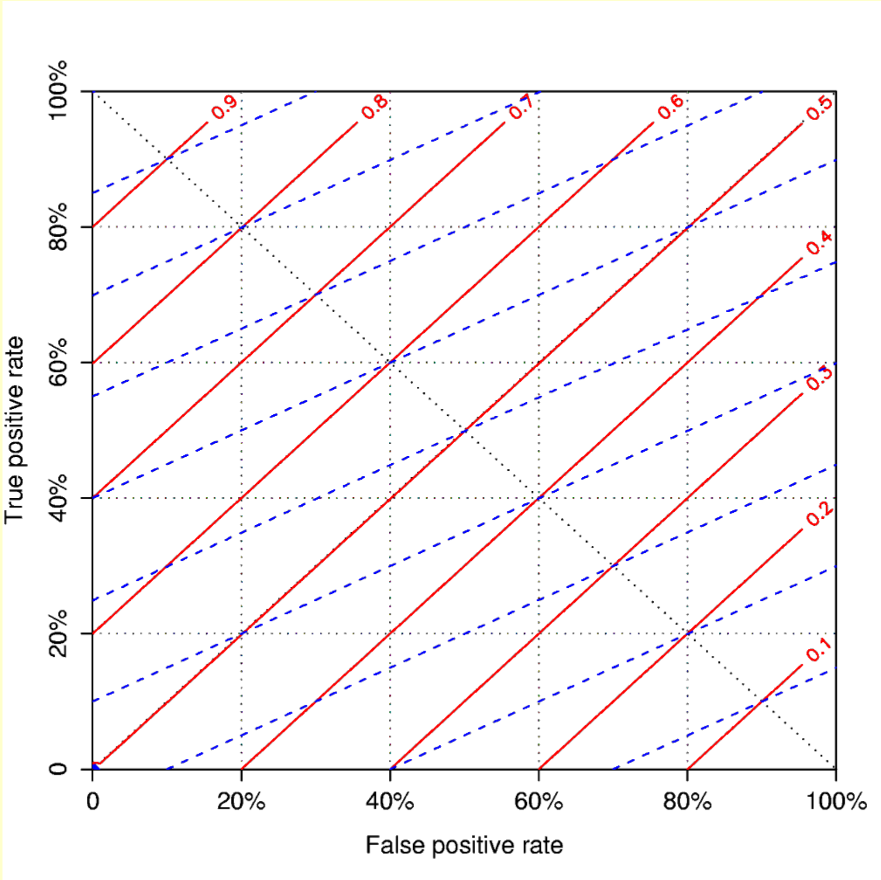
Producing ROC curves

- With a threshold:
  - prediction is numerical real-valued, decision is binary: positive $>\theta$
  - Bayesian classifier $P(\text{data}|+)/P(\text{data}|-)>\theta$
- Multiple classifiers from one algorithm
  - trained at different class ratios
  - using different misclassification costs
- The convex hull of different classifiers
  - trained on a single data set (fixed class ratio)

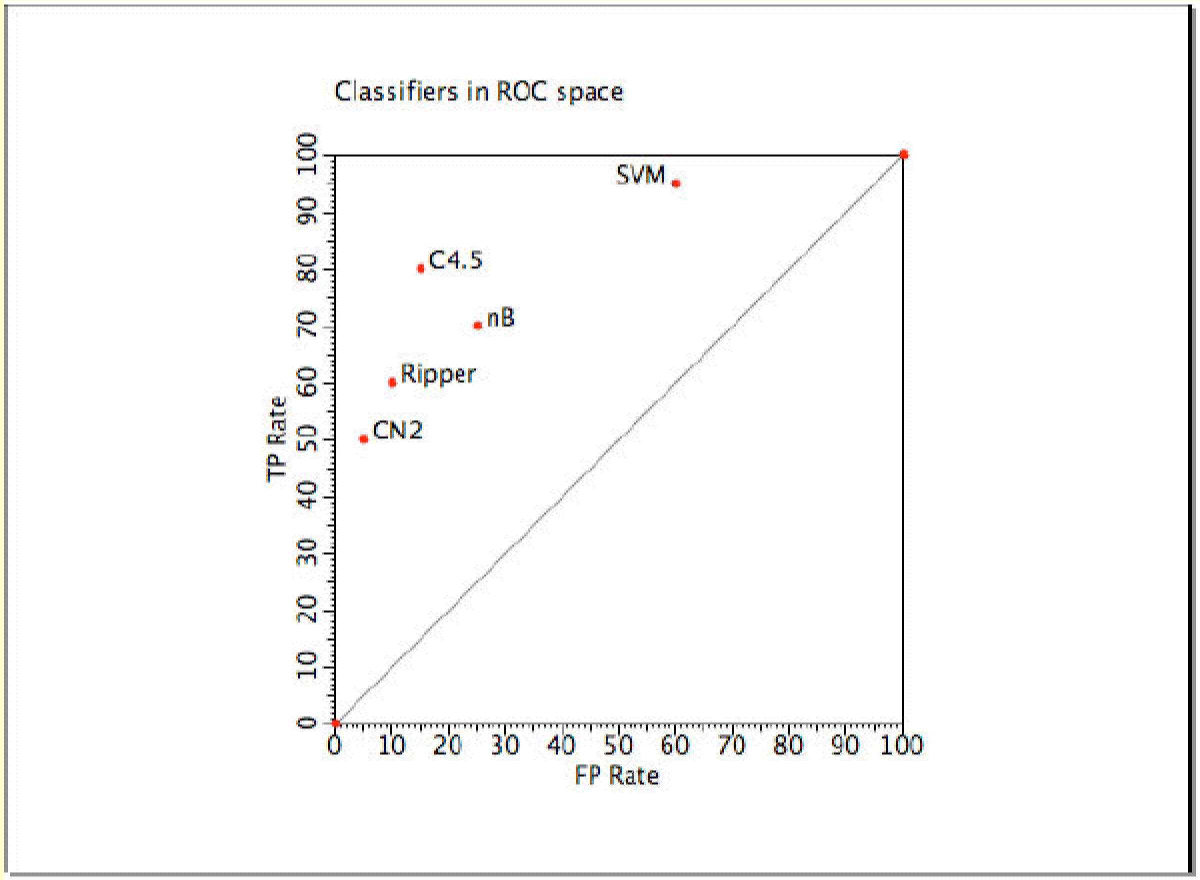
Iso-accuracy lines

- Red/Blue lines
  - Classifiers with the same accuracy
  - But at different distributions (pos/neg ratio)

- Intersection with diagonal *tpr =1- fpr*
- acc=(pos*tpr+neg*(1-fpr))/n
- acc=(pos*tpr+neg*tpr)/n
- acc/(pos+neg)/n = tpr
- acc = tpr

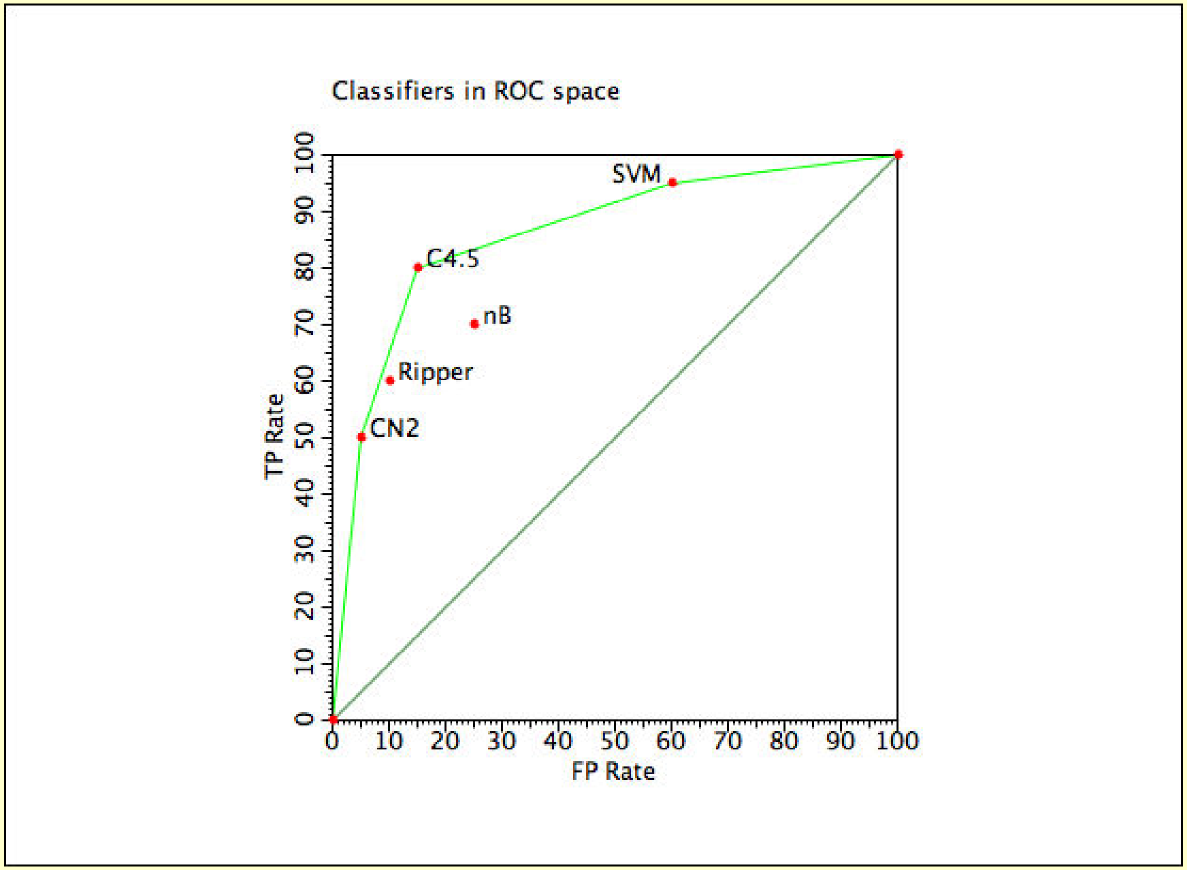


Comparing Learning Algorithms



Source: Peter Flach's tutorial on ROC curves, ICML 2004

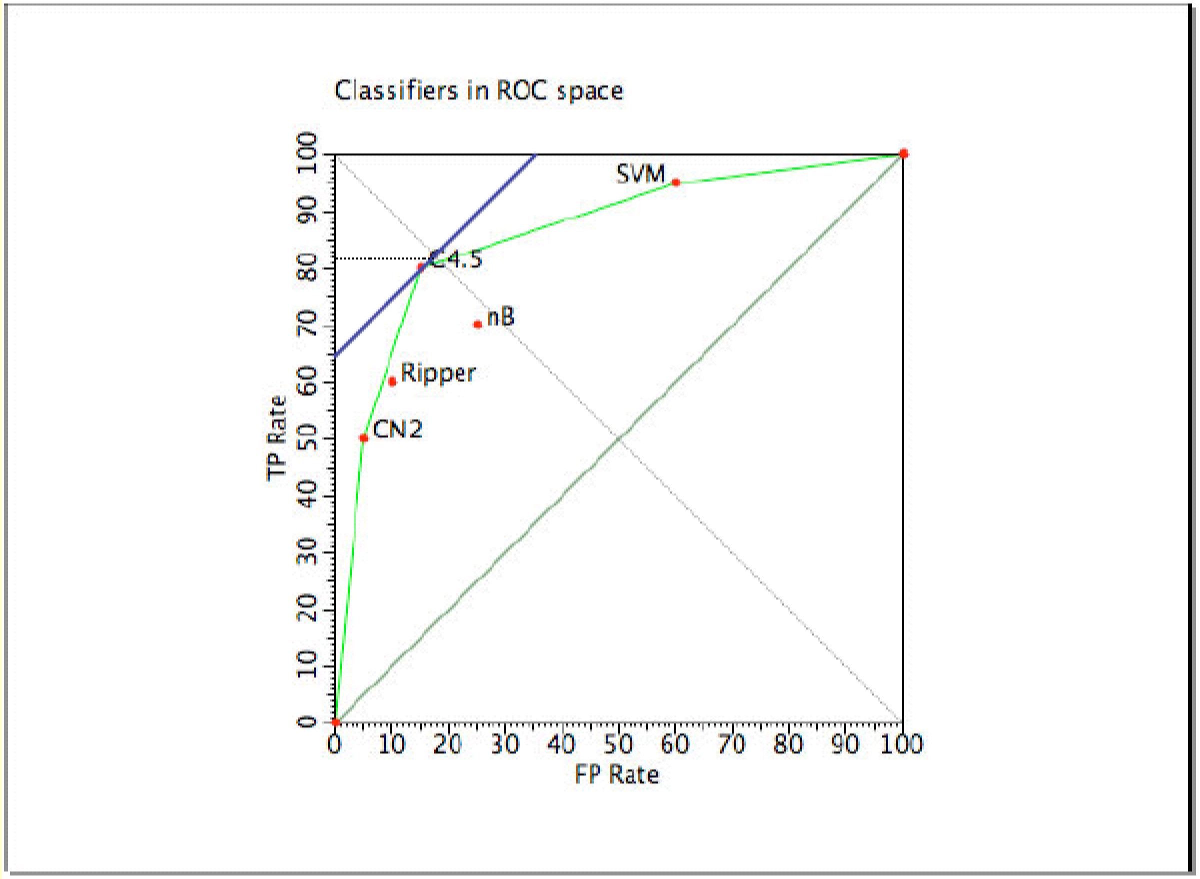
The Convex Hull



Classifiers on convex hull are optimal

Source: Peter Flach's tutorial on ROC curves, ICML 2004

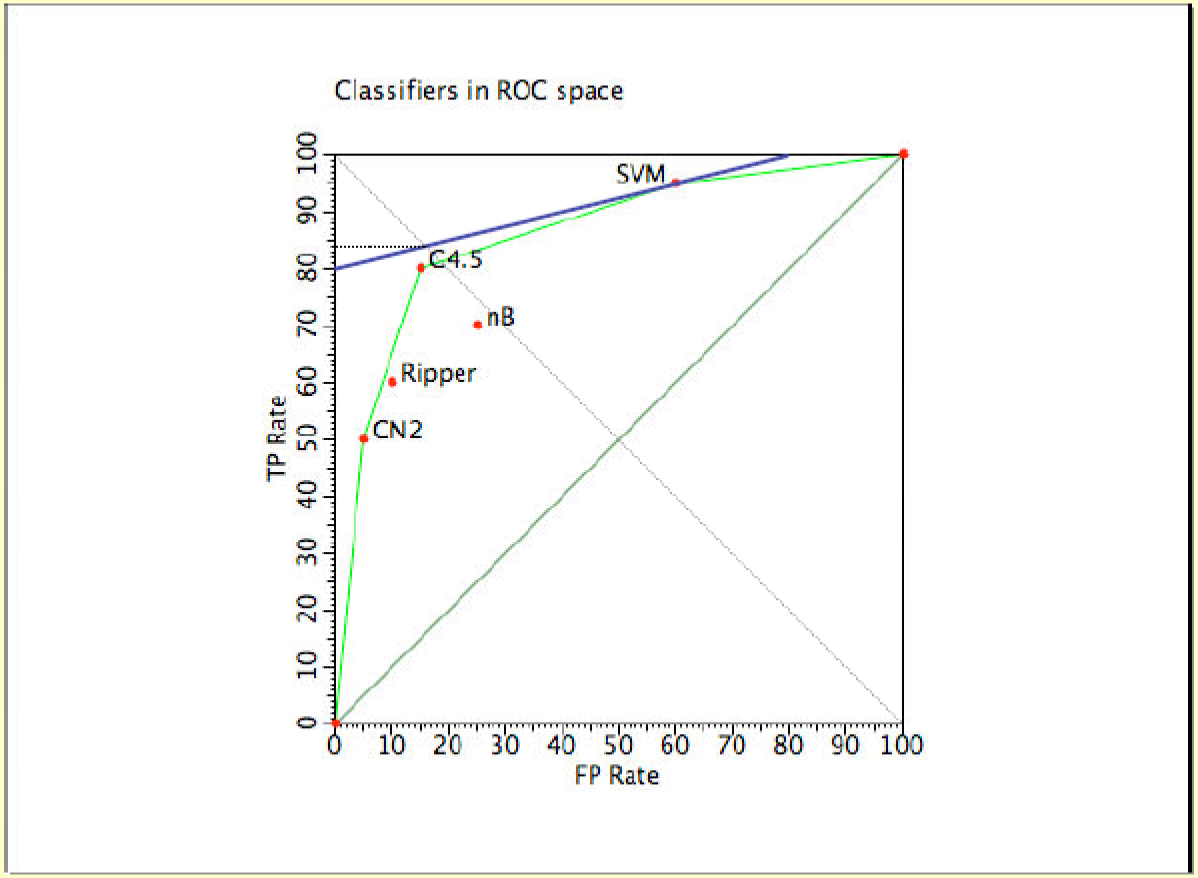
Choosing the Best



For uniform class distribution, C4.5 is optimal and achieves about 82% *accuracy*

Source: Peter Flach's tutorial on ROC curves, ICML 2004

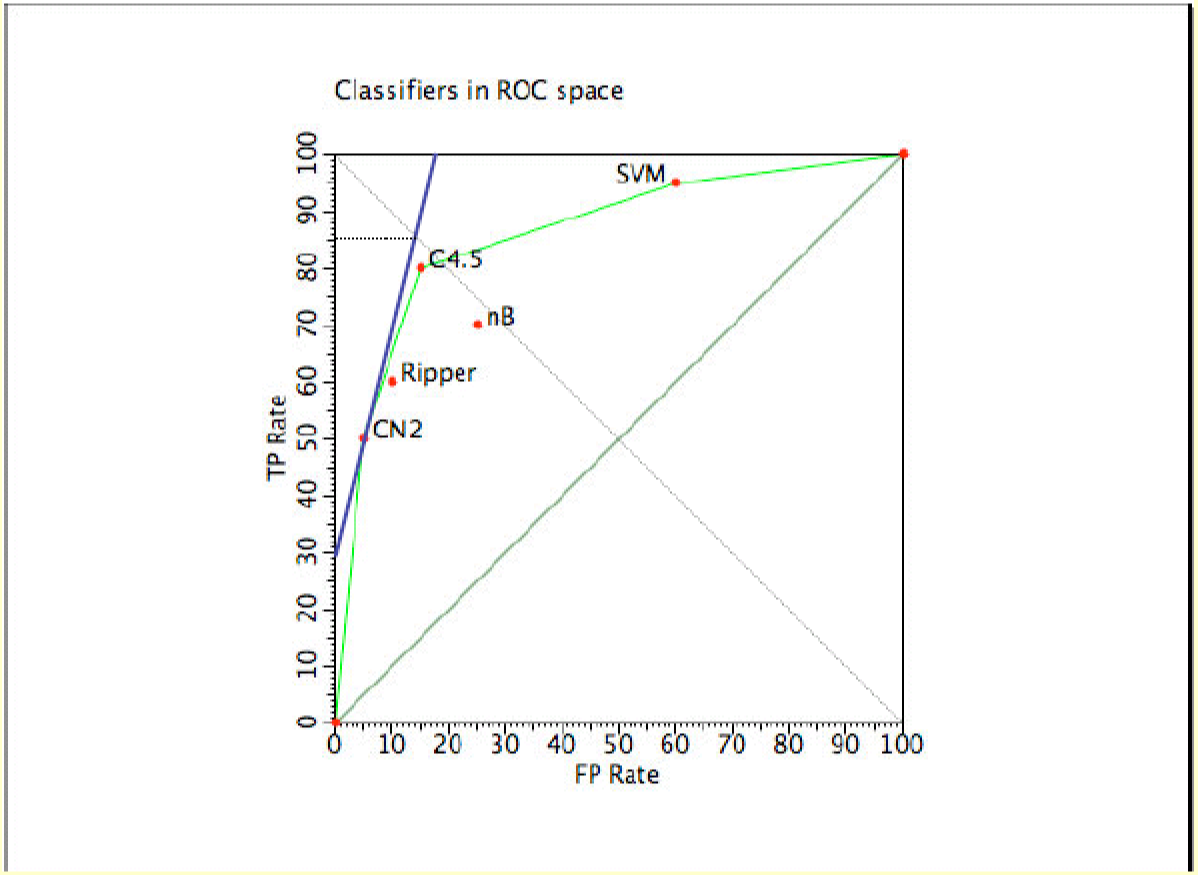
Choosing the Best



With four times as many +ves as –ves, SVM is optimal and achieves about 84% accuracy

Source: Peter Flach's tutorial on ROC curves, ICML 2004

Choosing the Best



With four times as many –ves as +ves, CN2 is optimal
 and achieves about 86% accuracy

Source: Peter Flach's tutorial on ROC curves, ICML 2004

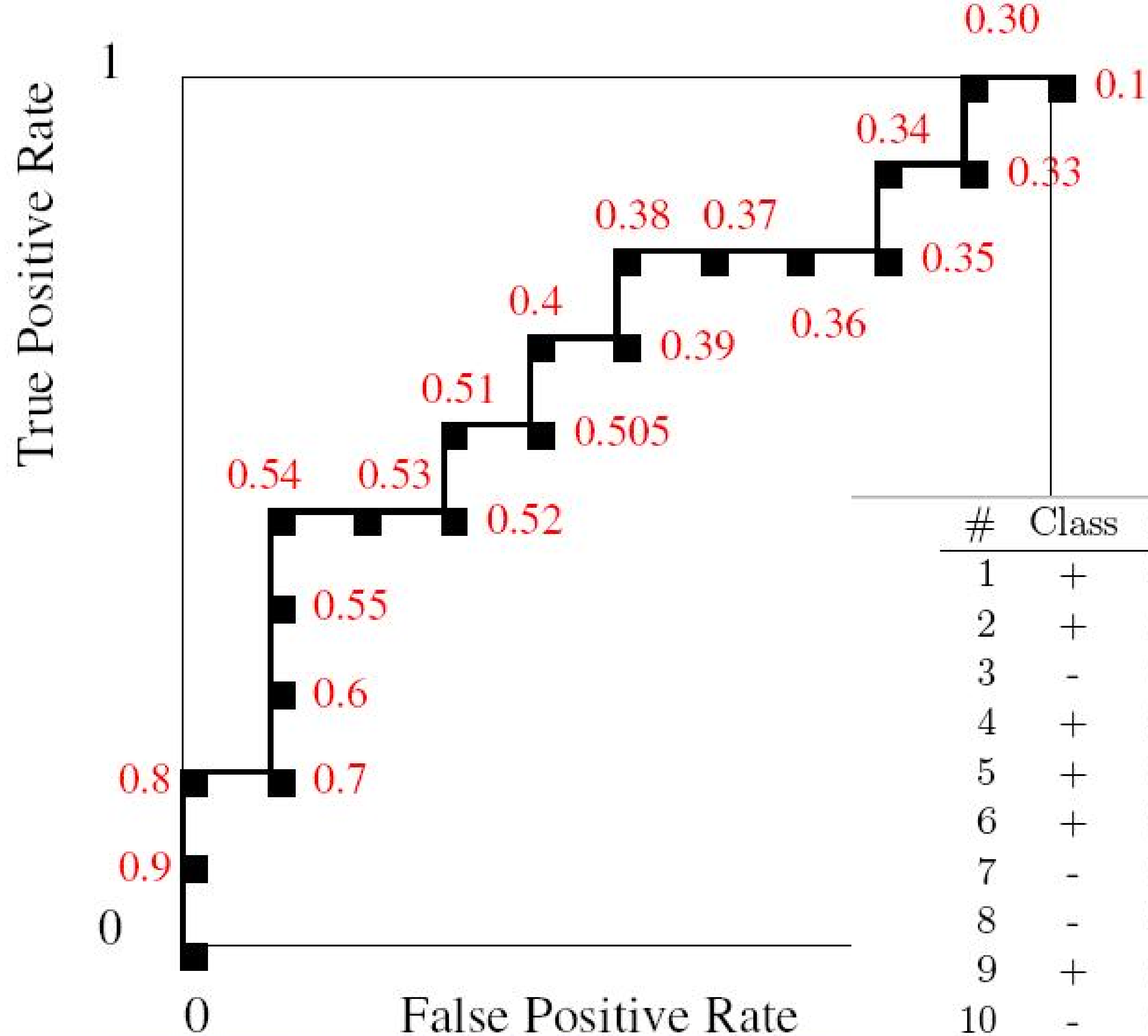
Rankers and classifiers

- A scoring classifier outputs scores f(x,+) and f(x,–) for each class
  - e.g. estimate class-conditional likelihoods P(x|+) and P(x|–)
  - scores don't need to be normalised
- f(x) = f(x,+)/f(x,–) can be used to rank instances from most to least likely positive
  - e.g. likelihood ratio P(x|+)/P(x|–)
- Rankers can be turned into classifiers by setting a threshold on f(x)

Drawing ROC curves for rankers

- Naïve method:
  - consider all possible thresholds
    - in fact, only k+1 for k instances
  - construct contingency table for each threshold
  - plot in ROC space
- Practical method:
  - rank test instances on decreasing score f(x)
  - starting in (0,0), if the next instance in the ranking is +ve move 1/Pos up, if it is –ve move 1/Neg to the right
    - make diagonal move in case of ties

ROC Curves



#	Class	Score	#	Class	Score
1	+	0.9	11	+	0.4
2	+	0.8	12	-	0.39
3	-	0.7	13	+	0.38
4	+	0.6	14	-	0.37
5	+	0.55	15	-	0.36
6	+	0.54	16	-	0.35
7	-	0.53	17	+	0.34
8	-	0.52	18	-	0.33
9	+	0.51	19	+	0.30
10	-	0.505	20	-	0.1

ROC curves for rankers

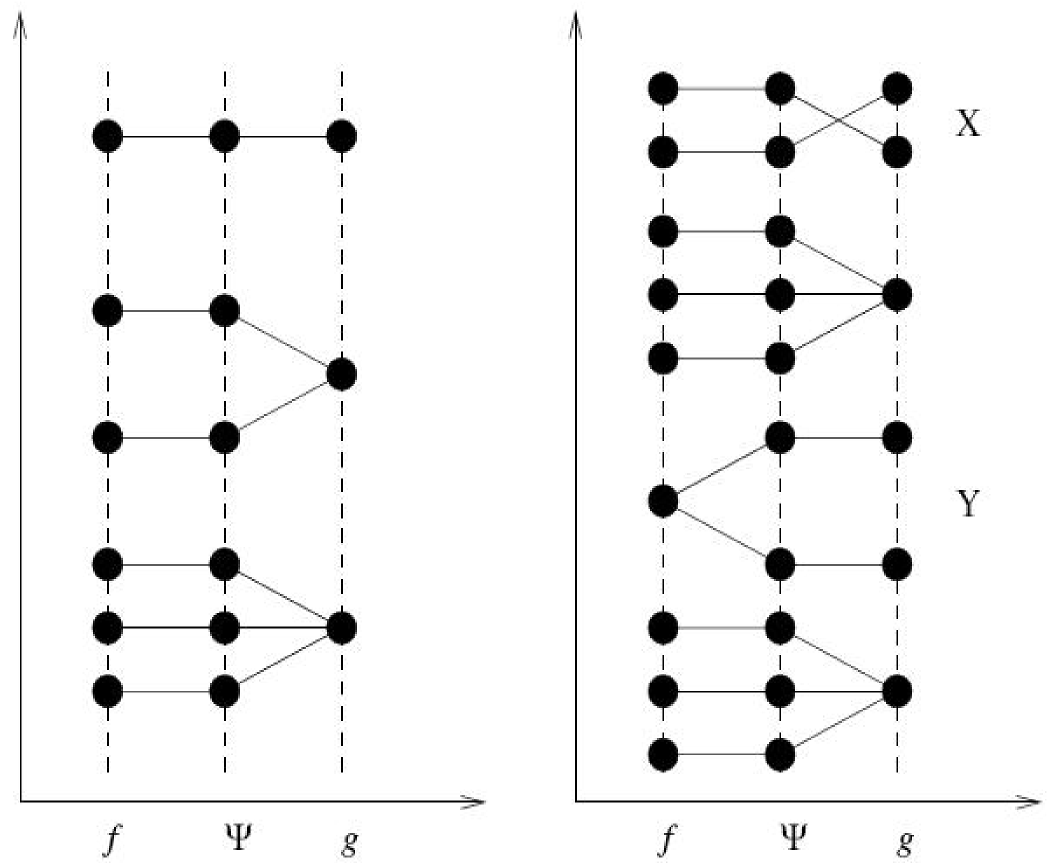
- Visualizes the quality of the ranker or probabilistic model on a test set,
  - without committing to a classification threshold
  - aggregates over all possible thresholds
- Curve slope indicates local class distribution
  - diagonal segment -> locally random behavior
- Concavities: locally worse than random behavior
  - convex hull corresponds to discretizing scores
  - can potentially do better: repairing concavities

The AUC metric

- The Area Under ROC Curve (AUC) assesses the ranking in terms of separation of the classes
  - all the +ves before the –ves: AUC=1
  - random ordering: AUC=0.5
  - all the –ves before the +ves: AUC=0
- AUC for comparing learning algorithms
  - a better measure than accuracy

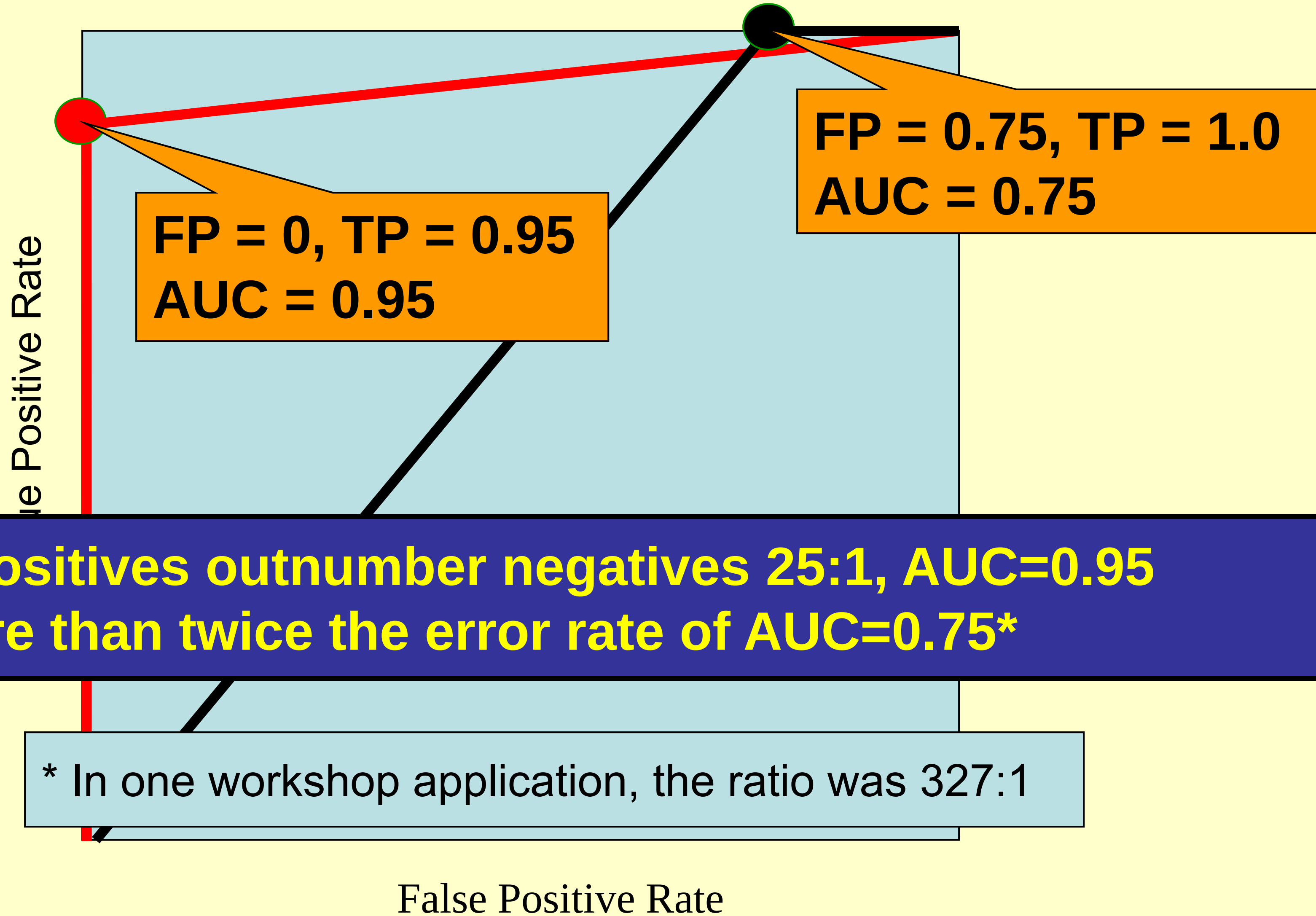
AUC – why it's a good measure

- It is more discriminant than accuracy and consistent with it
- Like ROC, it is not sensitive to imbalance



X is Consistency counter example
Y is Discriminancy counter example

Why sometimes it isn't



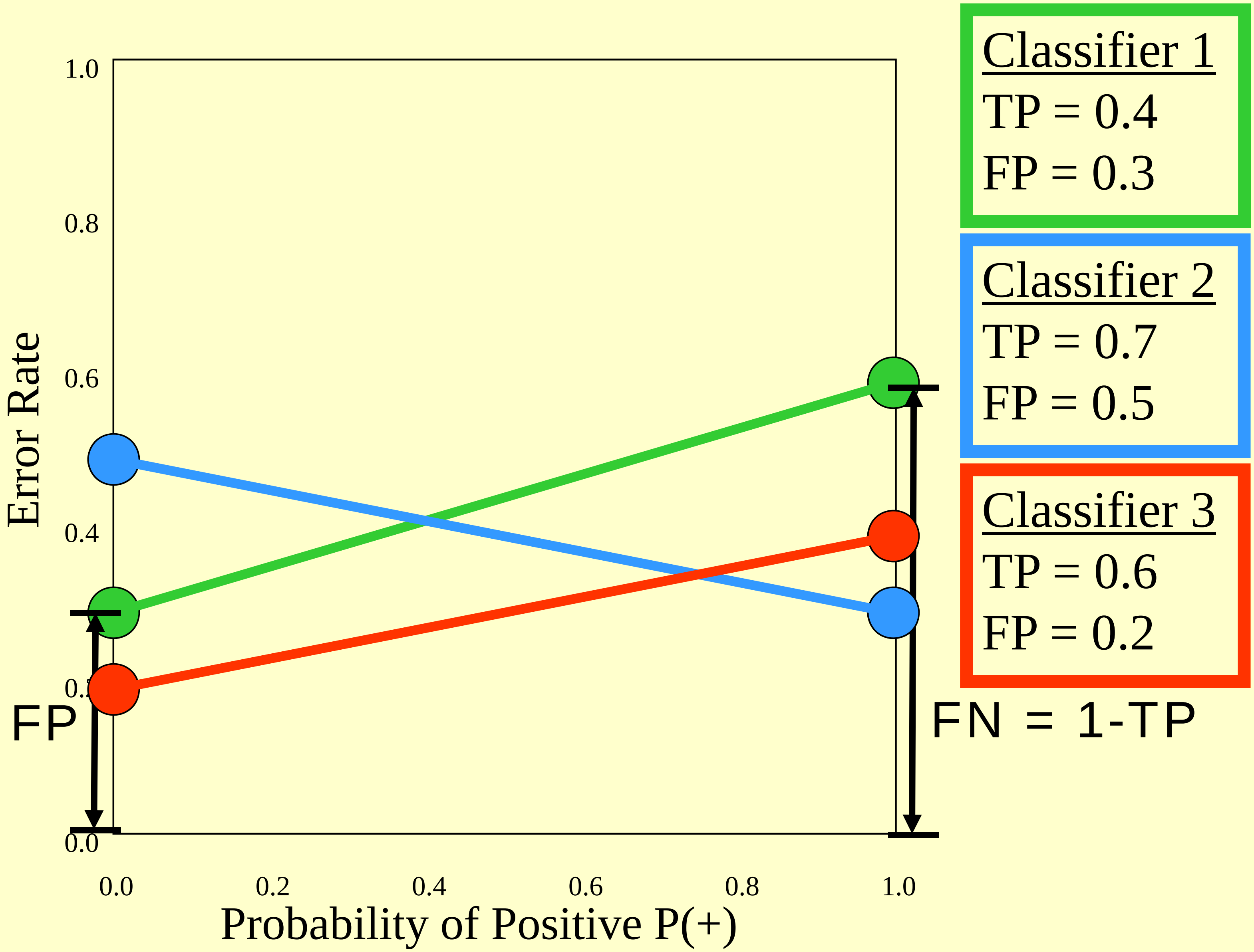
When positives outnumber negatives 25:1, AUC=0.95 has more than twice the error rate of AUC=0.75*

* In one workshop application, the ratio was 327:1

Single Values

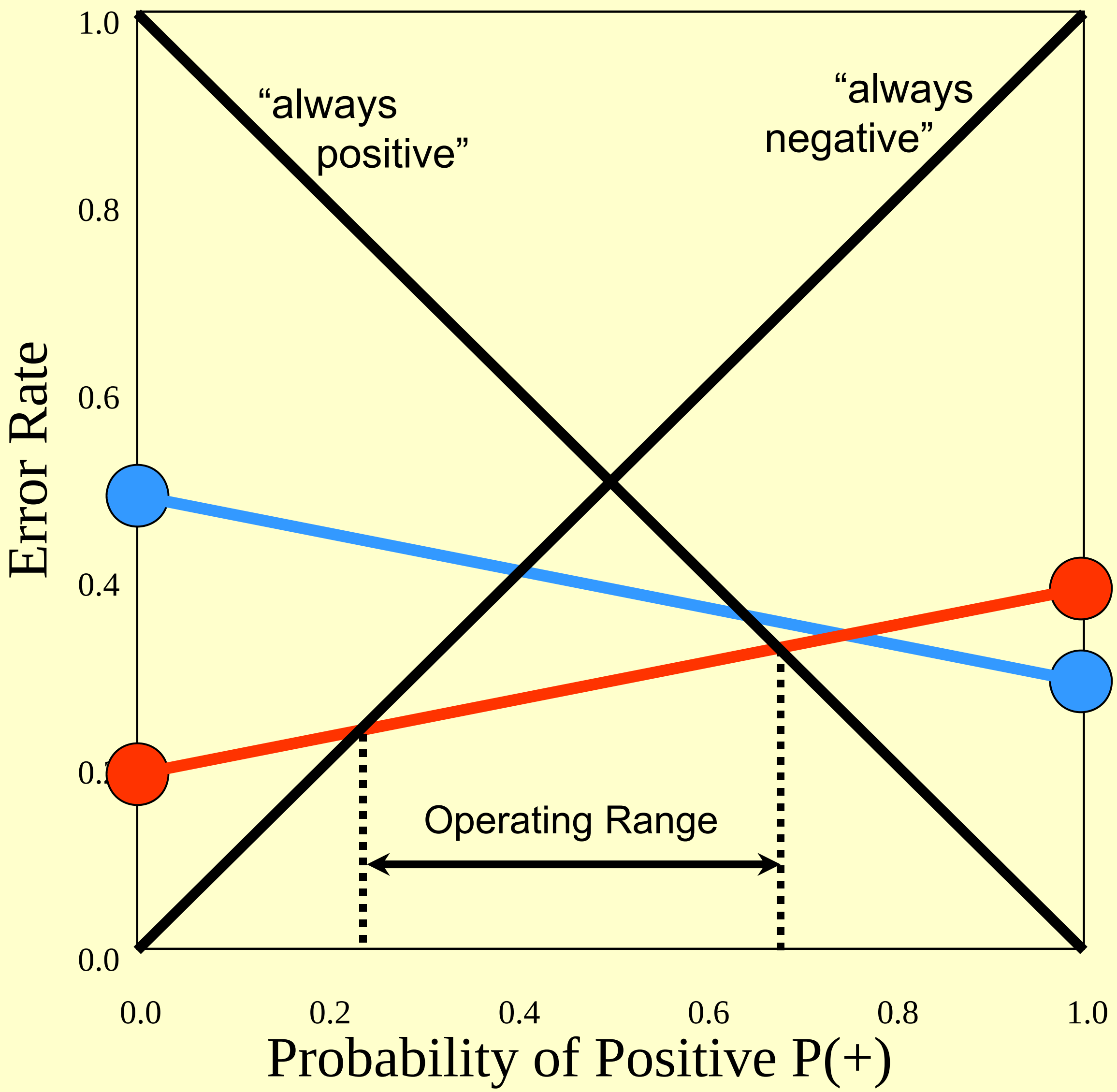
- Summarize performance
  - Easy to compare classifiers
- We know how to
  - average them,
  - compute confidence intervals,
  - test for significance, etc.
- But hide important differences

Cost Curves

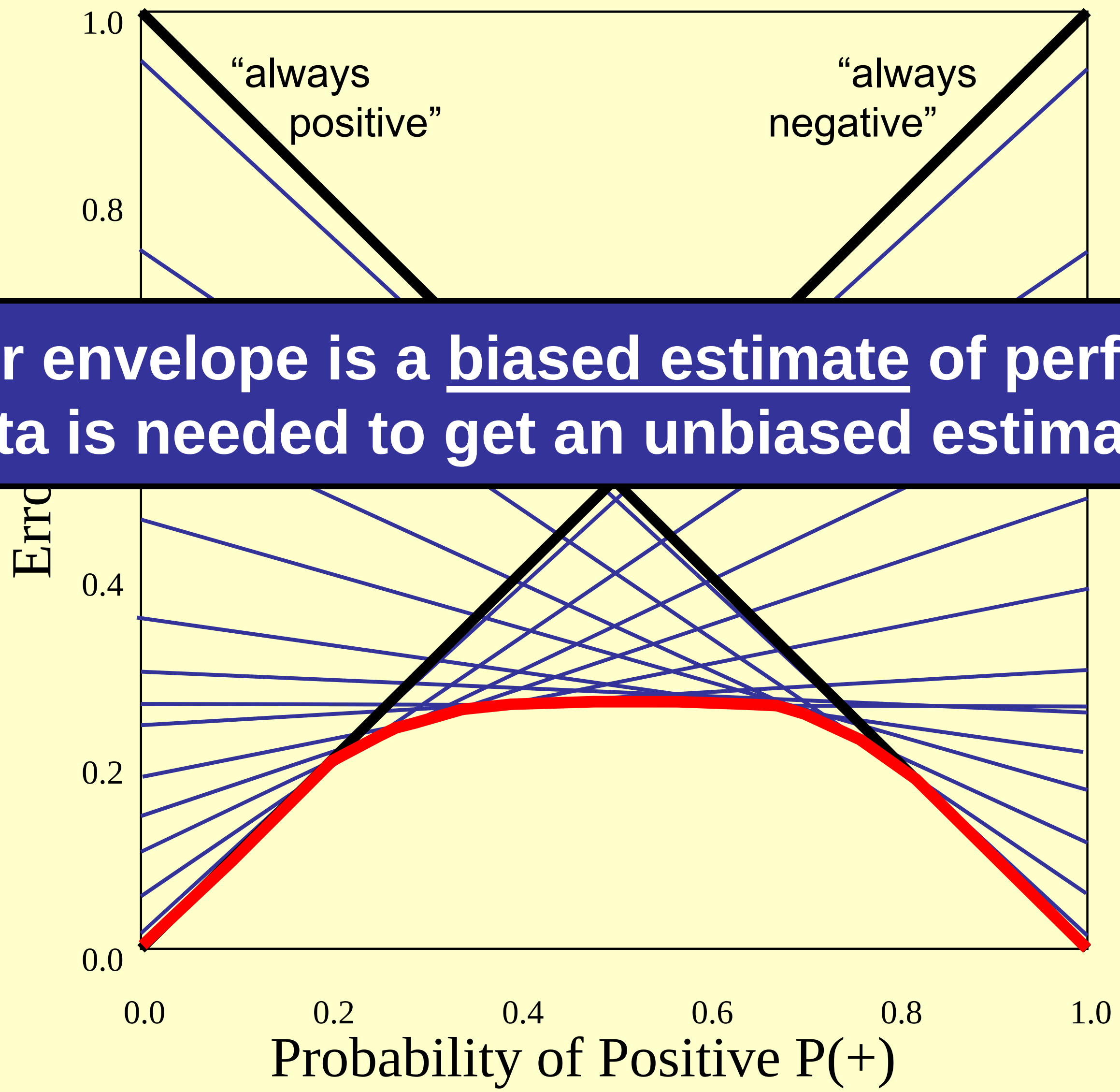


1.0
0.8
0.6
0.4
0.2
0.0
Error Rate
FP
FN = 1-TP
0.0
0.2
0.4
0.6
0.8
1.0
Probability of Positive P(+)
Classifier 1
TP = 0.4
FP = 0.3
Classifier 2
TP = 0.7
FP = 0.5
Classifier 3
TP = 0.6
FP = 0.2

Operating Range

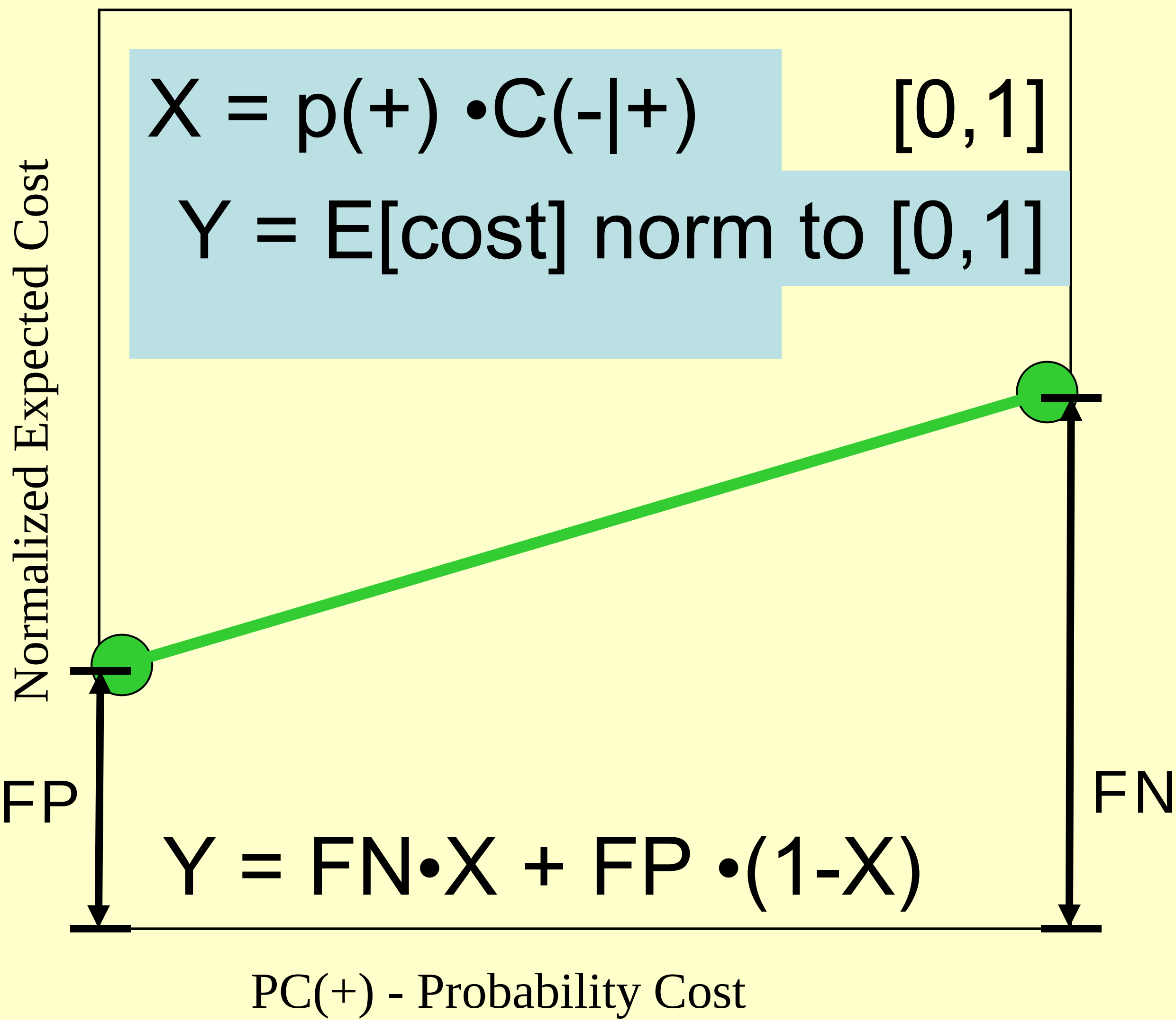


Lower Envelope

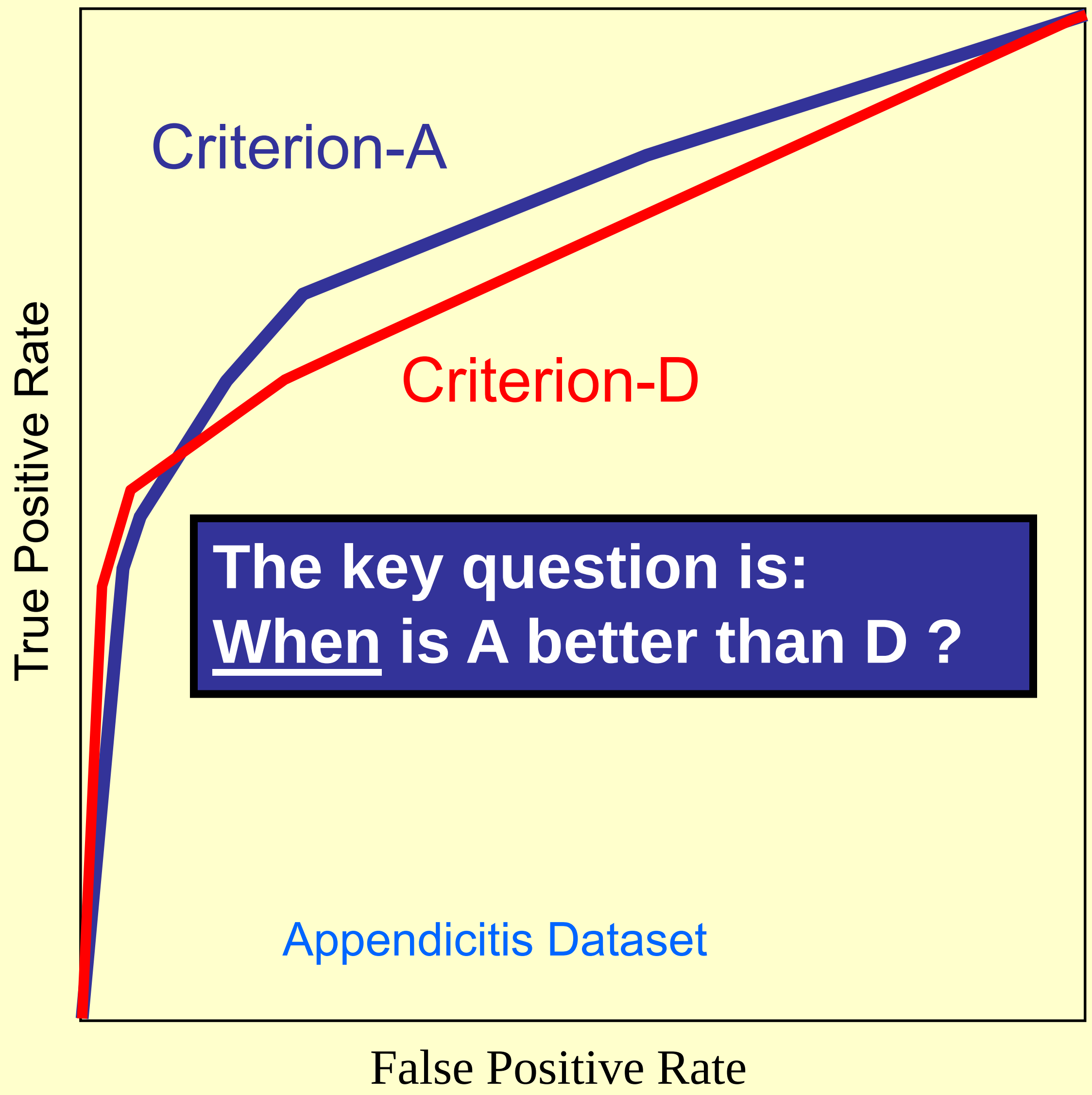


The lower envelope is a <u>biased estimate</u> of performance. Fresh data is needed to get an unbiased estimate.

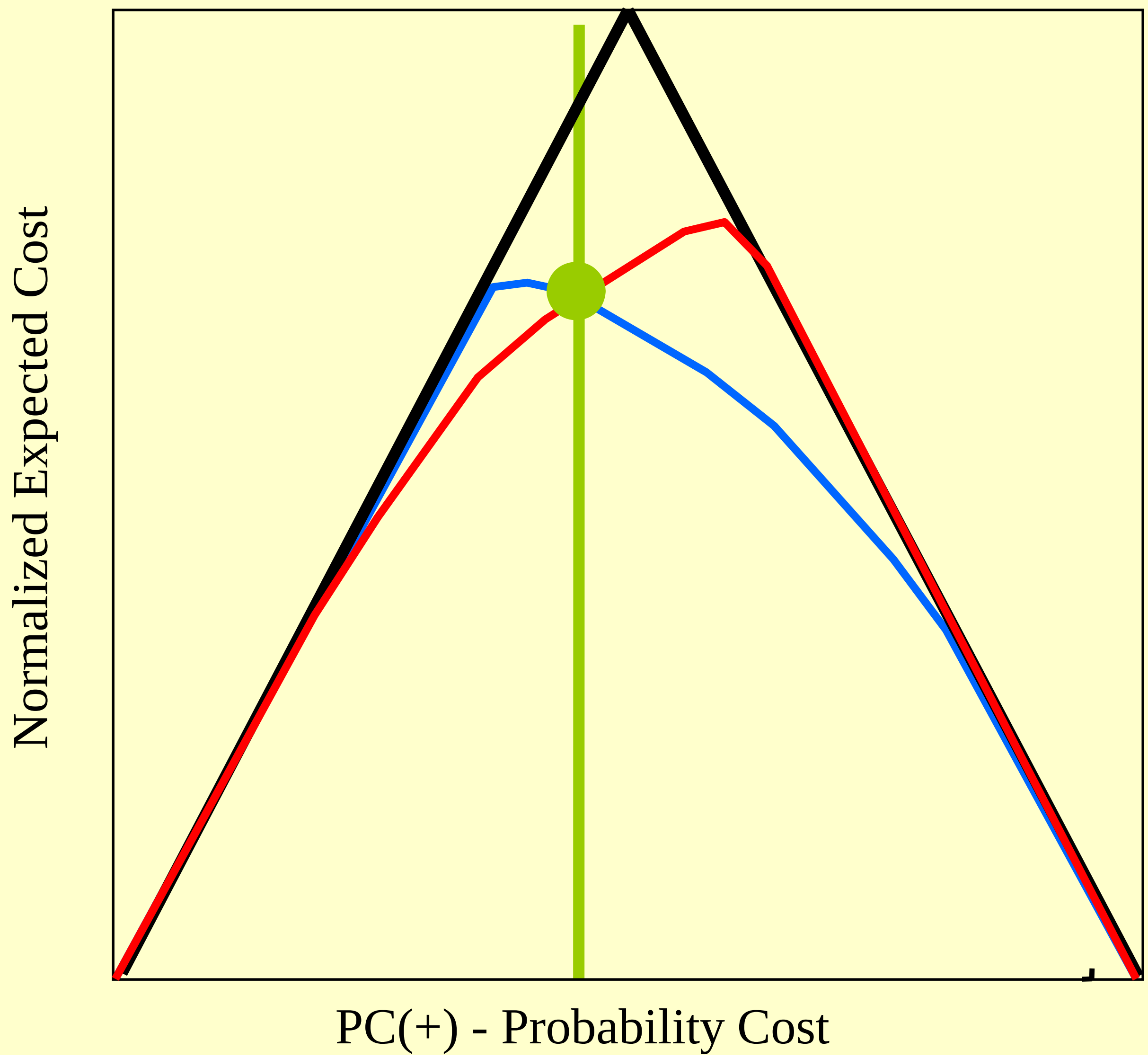
Taking Costs Into Account



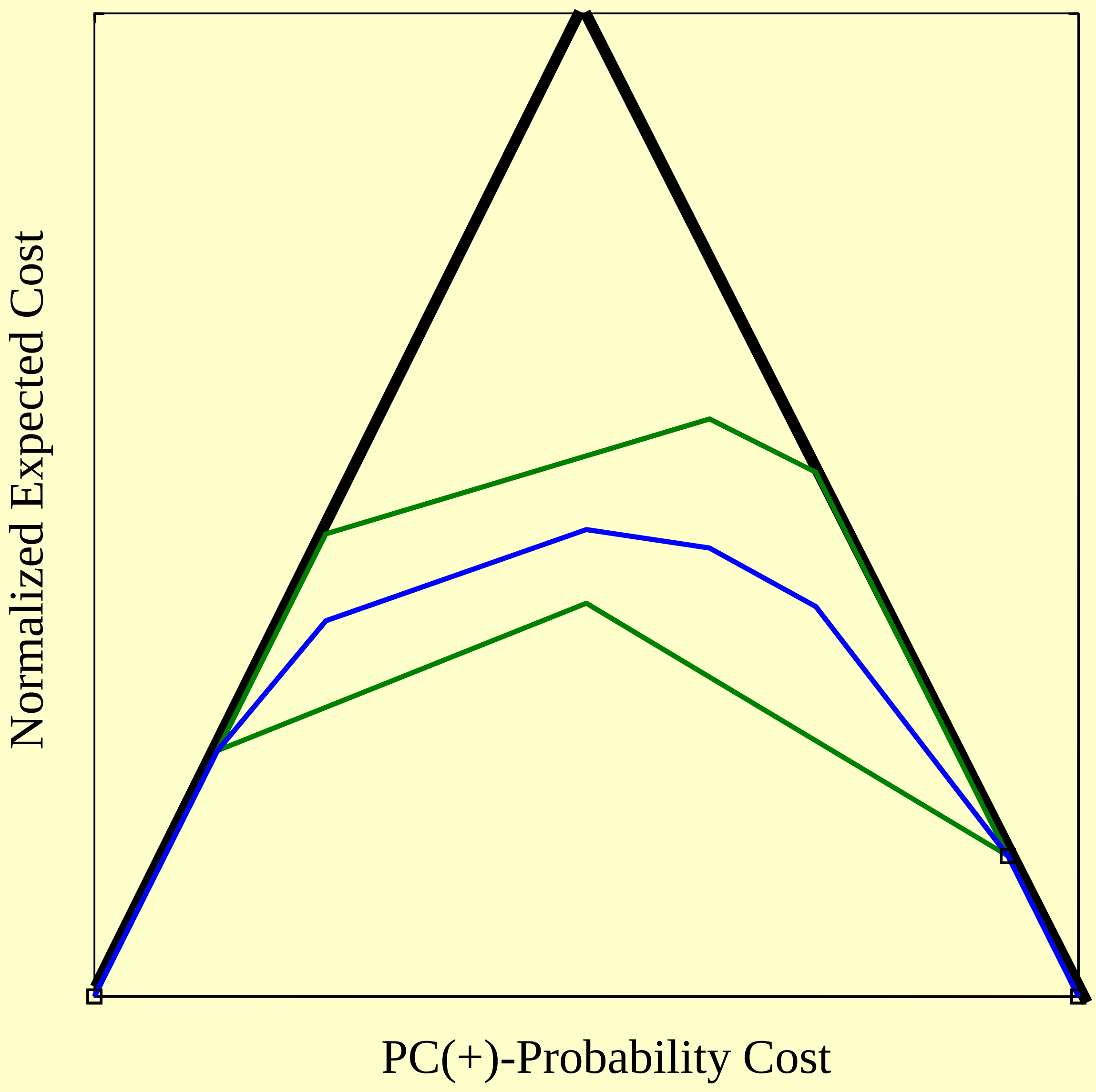
2 Splitting Criteria for C4.5



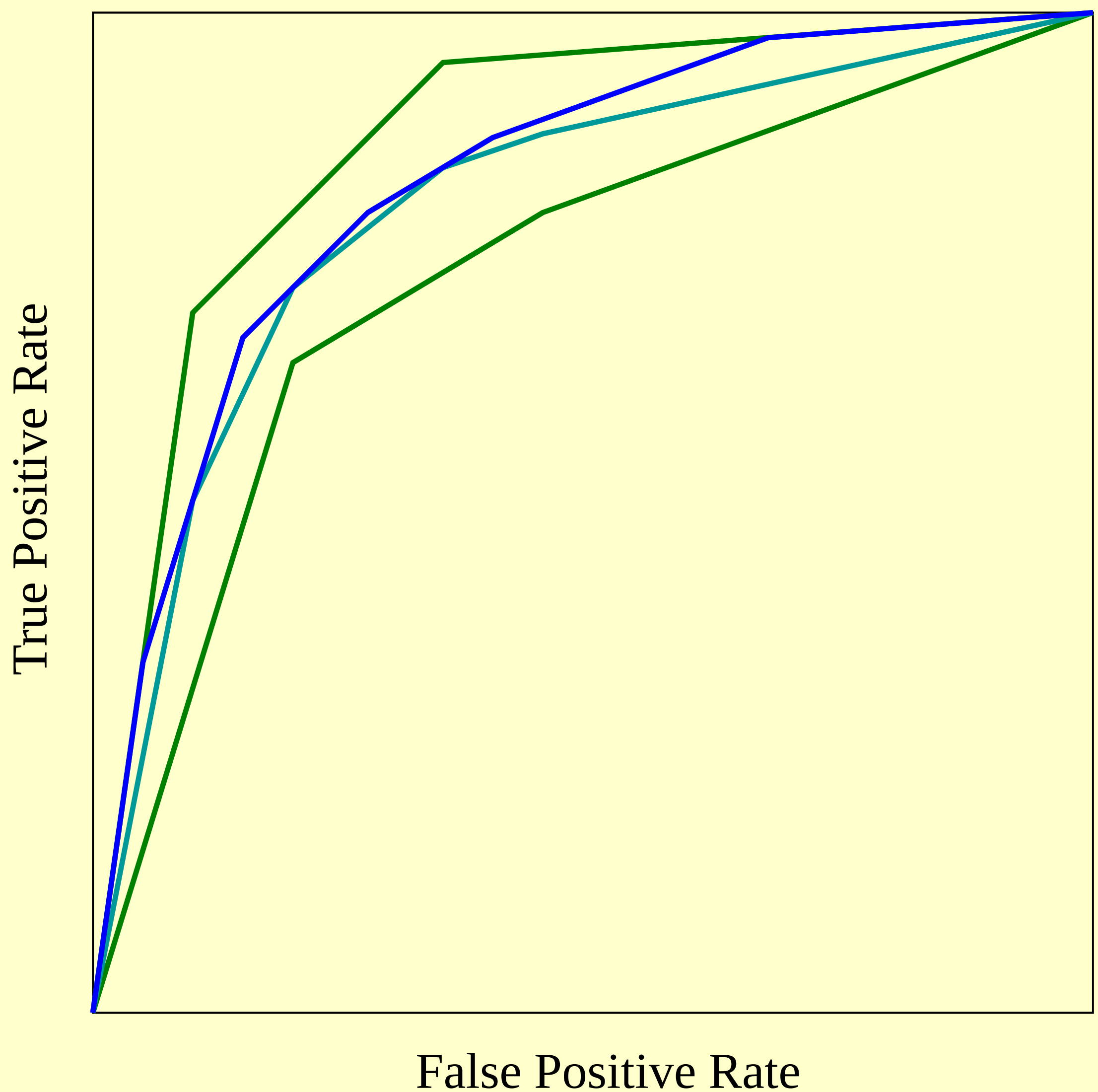
Comparing Cost Curves



Averaging Cost Curves



Averaging ROC Curves



Confidence Intervals

True	Predicted	
	pos	neg
pos	78	22
neg	40	60

True	Predicted	
	pos	neg
pos	75	25
neg	45	55

True	Predicted	
	pos	neg
pos	83	17
neg	38	62

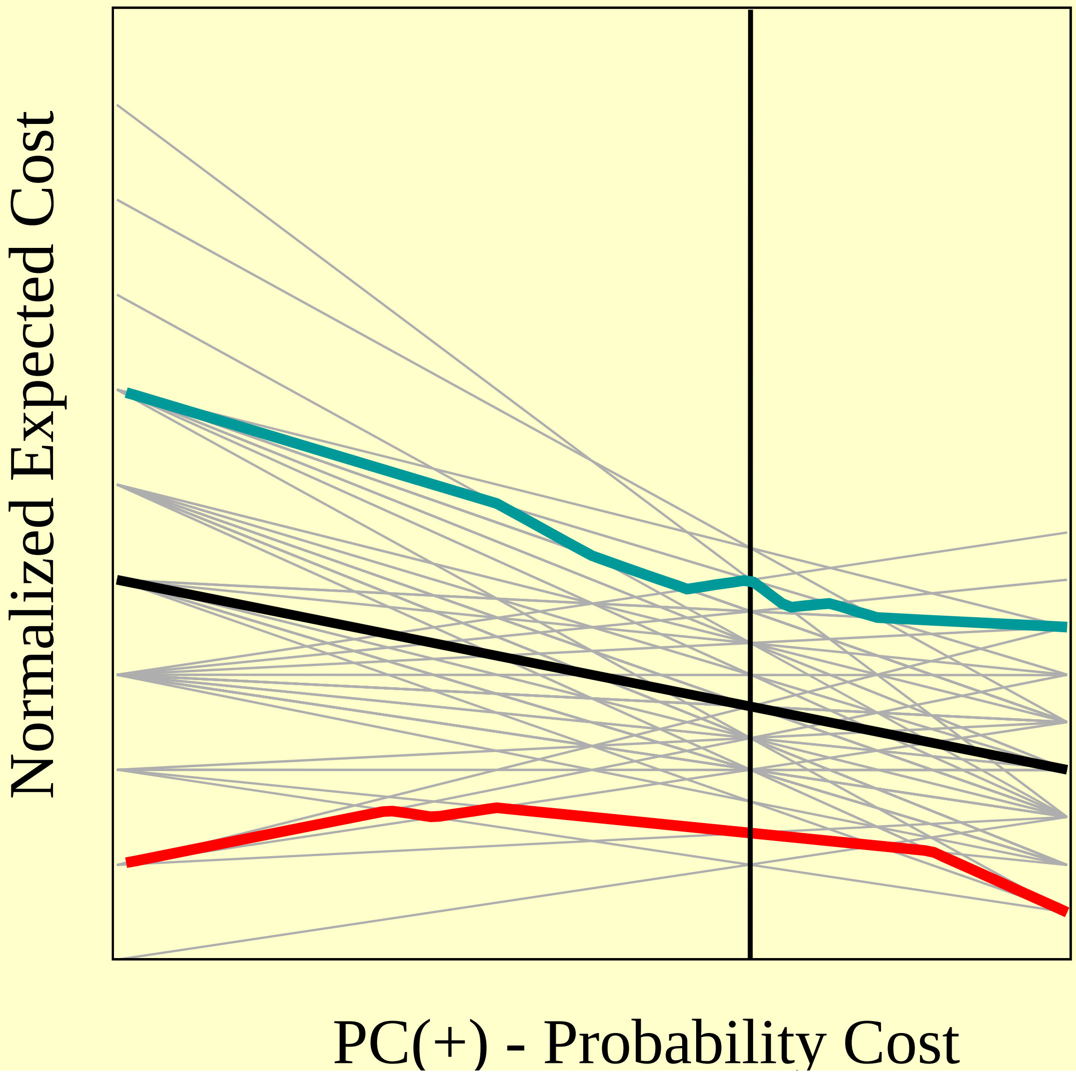
<u>Original</u>
TP = 0.78
FP = 0.4

<u>Resample #1</u>
TP = 0.75
FP = 0.45

<u>Resample #2</u>
TP = 0.83
FP = 0.38

Resample confusion matrix 10000 times and take 95% envelope

Confidence Interval Example



Paired Resampling to Test Statistical Significance

For the 100 test examples in the negative class:

Predicted by Classifier1	**Predicted by Classifier2**	
	pos	**neg**
pos	**30**	**10**
neg	**0**	**60**

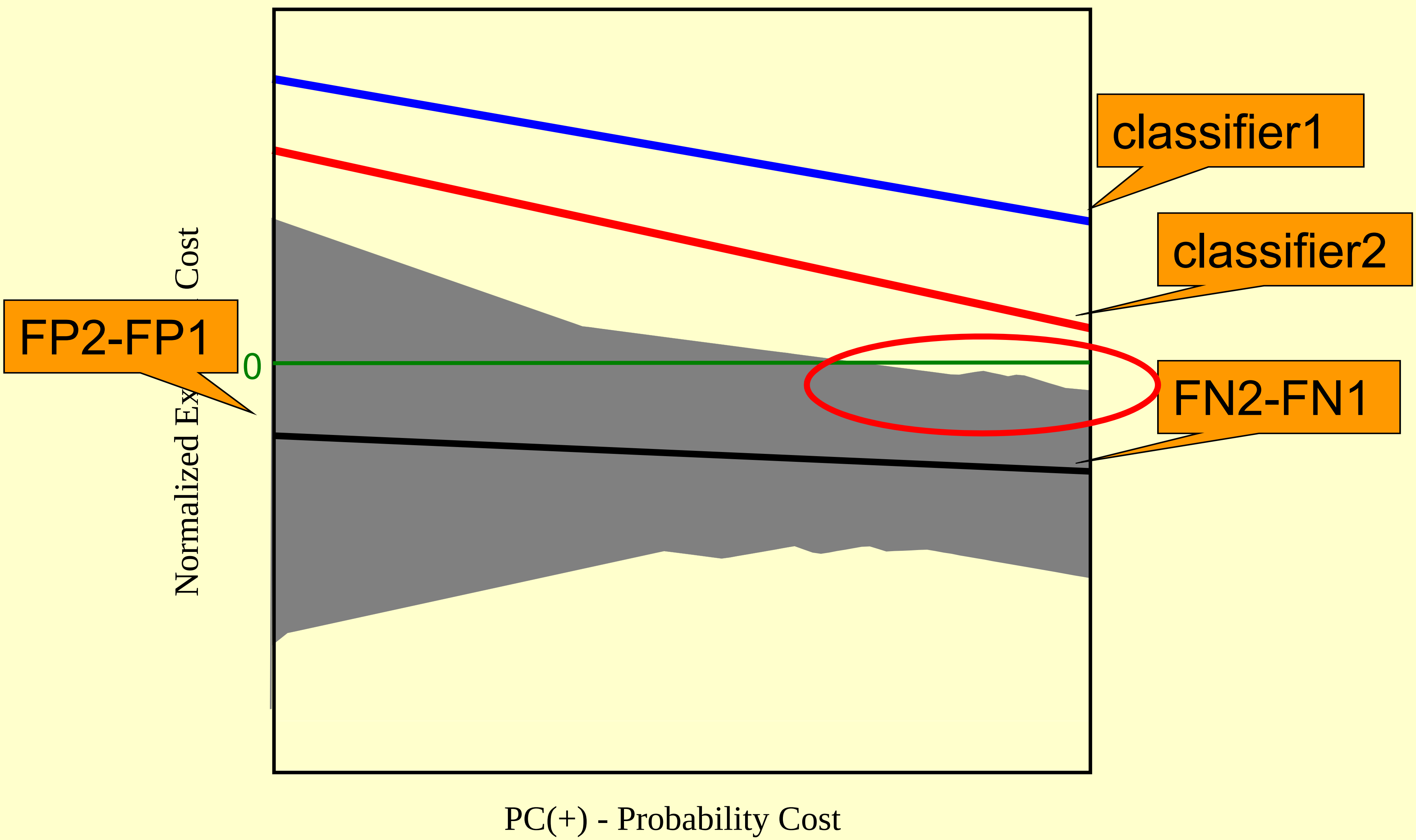
FP for classifier1: (30+10)/100 = 0.40

FP for classifier2: (30+0)/100 = 0.30

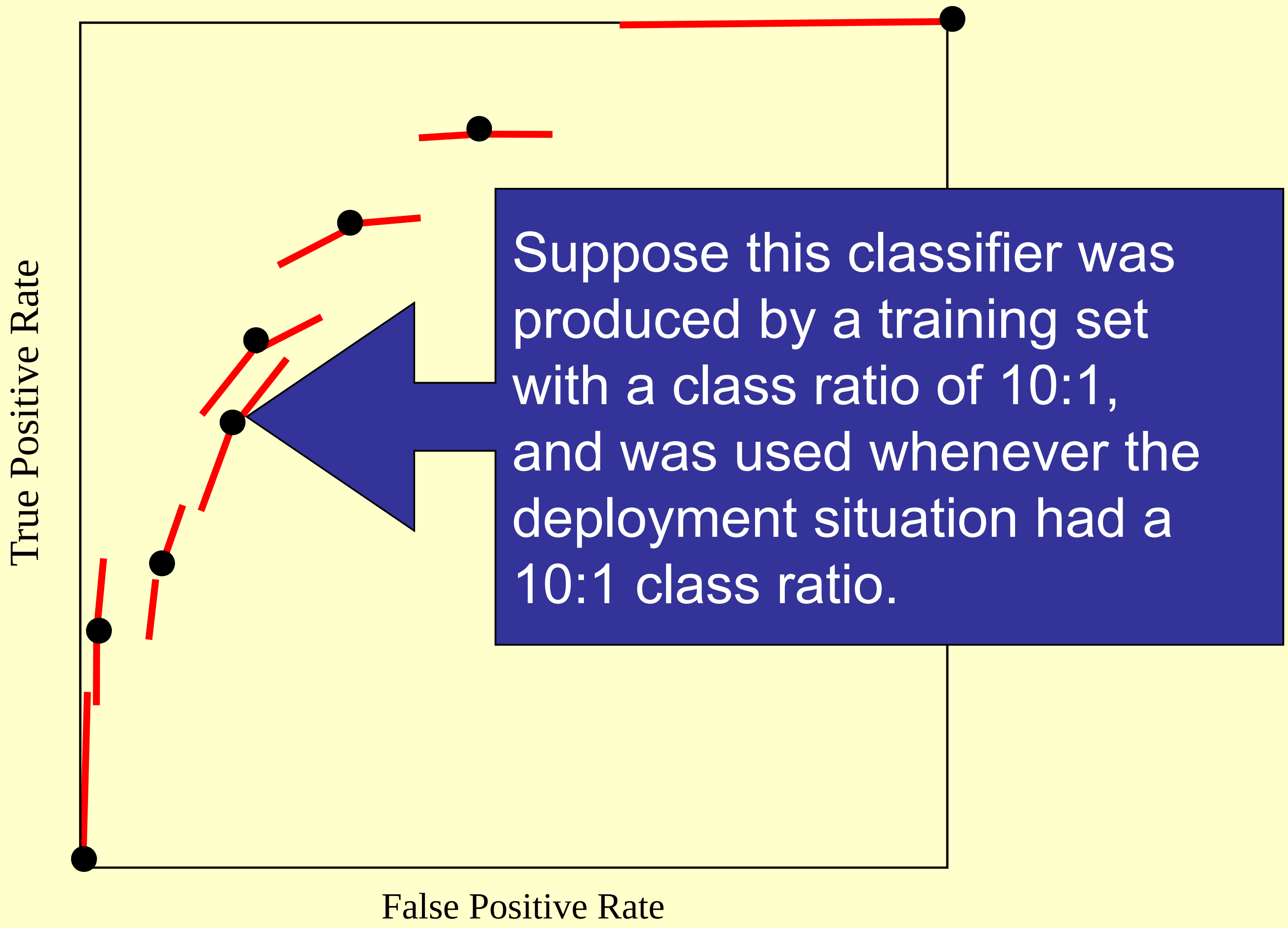
FP2 – FP1 = -0.10

Resample this matrix 10000 times to get (FP2-FP1) values.
Do the same for the matrix based on positive test examples.
Plot and take 95% envelope as before.

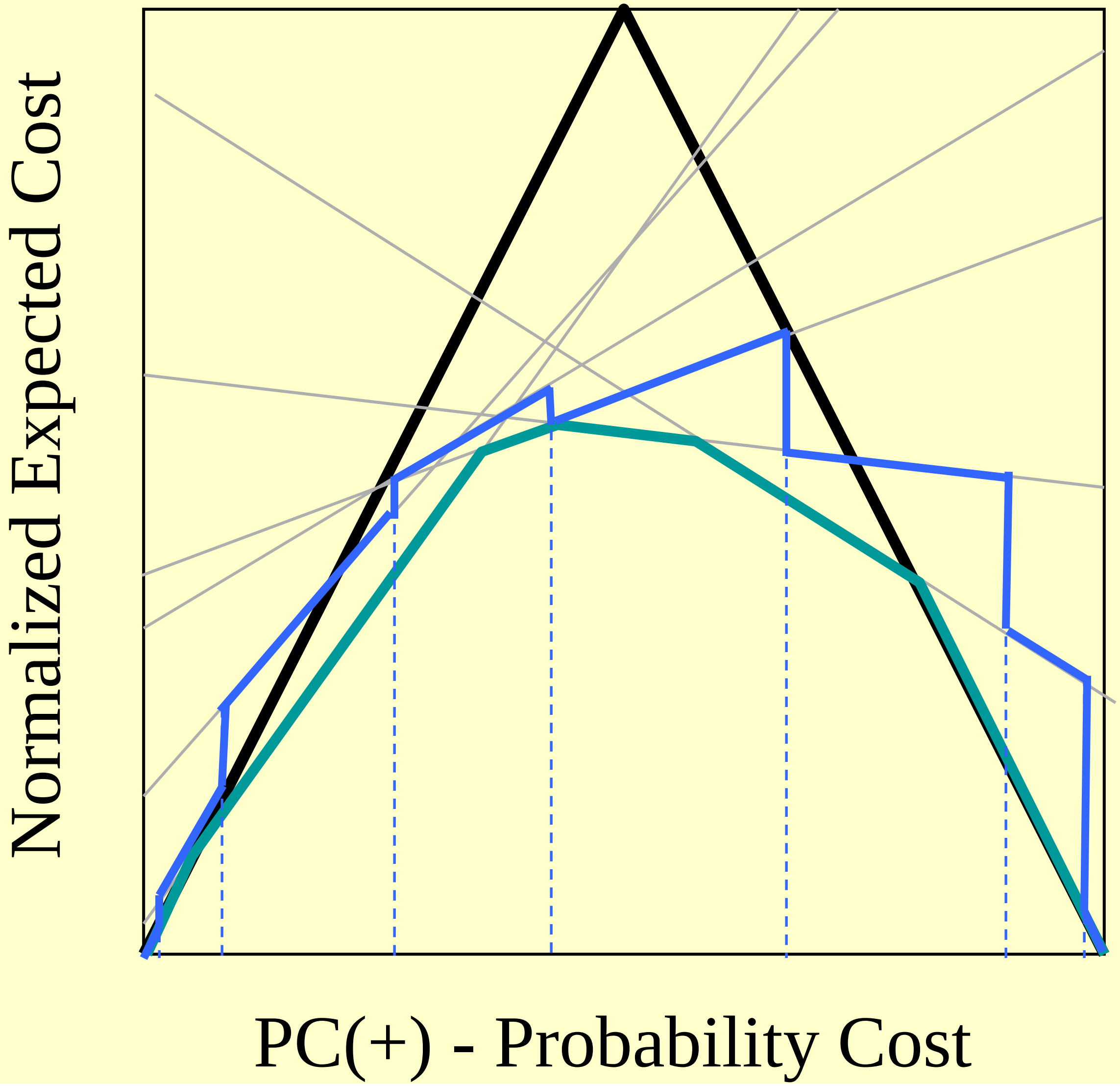
Low correlation = Low significance



ROC, Selection procedure



Cost Curves, Selection Procedure



A Personal Opinion: You Decide

- ROC curves
  - Show the inherent trade-off between TPR/FPR
- AUC
  - Is better than accuracy
  - But does not show <u>when</u> one classifier is better than another.
- Cost curves enable easy visualization of
  - Average performance (expected cost)
  - Operating range
  - Confidence intervals on performance
  - Difference in performance and its significance.