Classifier Verification

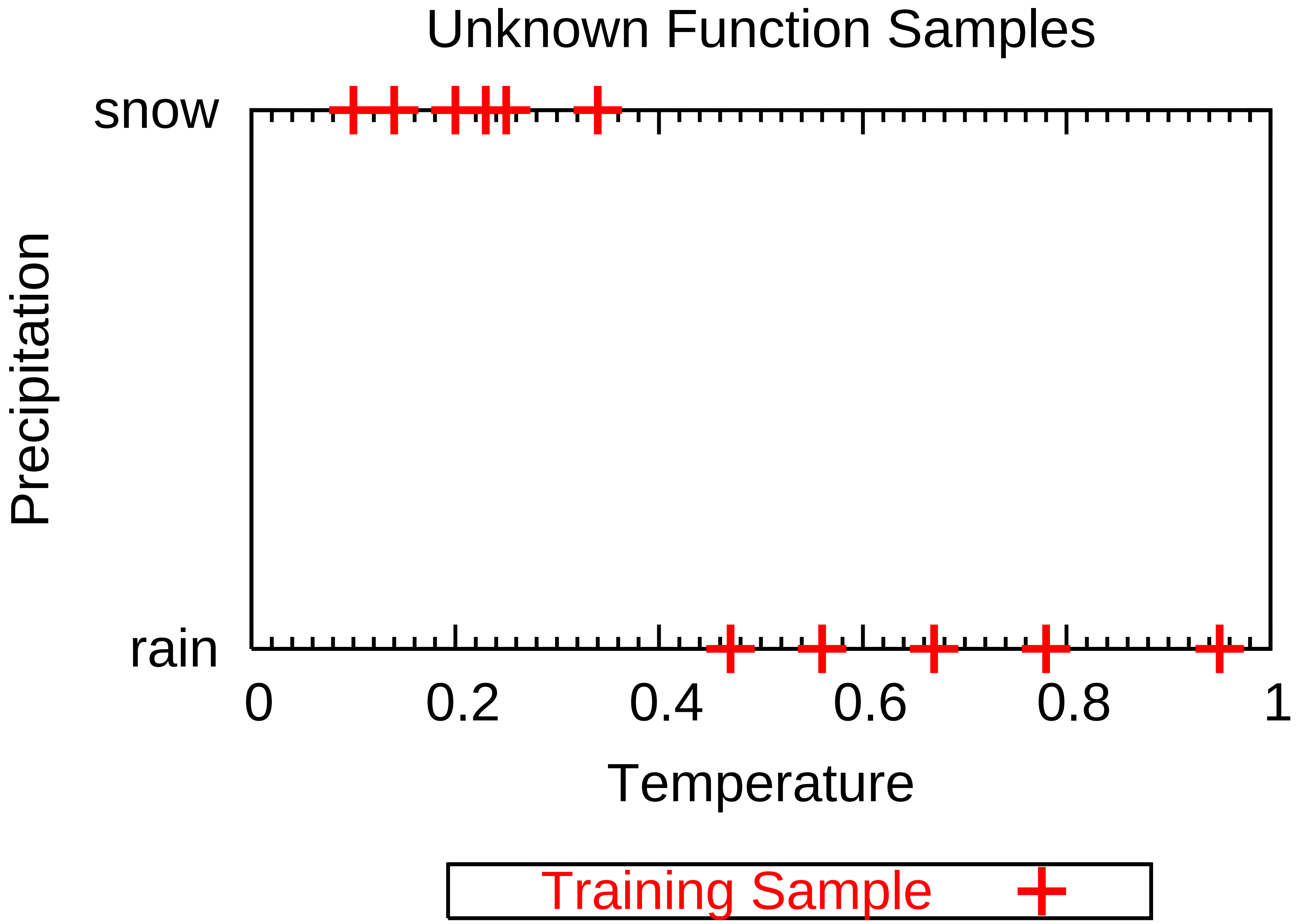
John Langford

The Classification Problem:

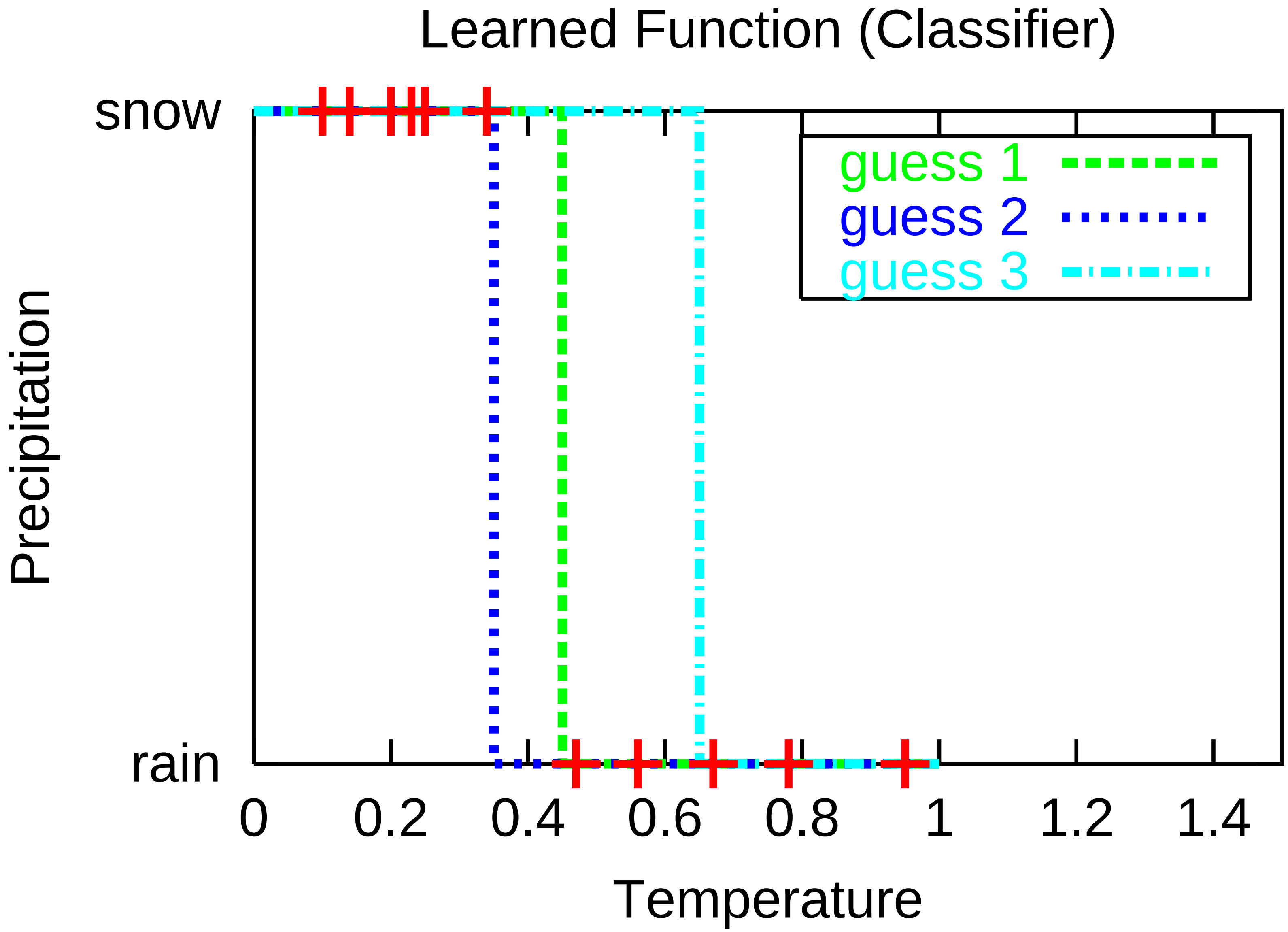
1. Get (input, class label) pairs from unknown function

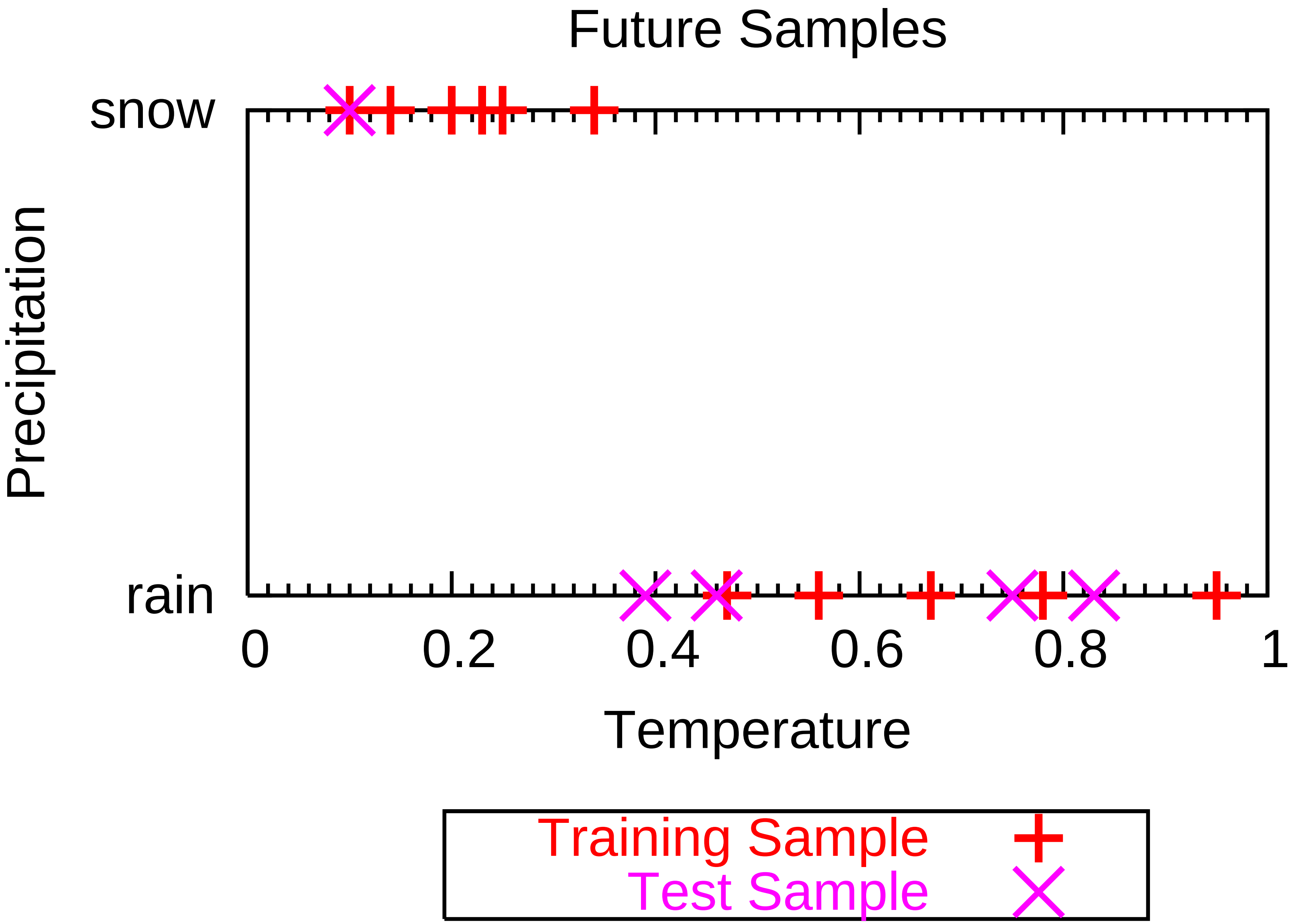
2. Guess the unknown function

3. ... to minimize errors on future inputs



Unknown Function Samples
snow
rain
Precipitation
0
0.2
0.4
0.6
0.8
1
Temperature
Training Sample


Learned Function (Classifier)
snow
rain
Precipitation
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Temperature
guess 1
guess 2
guess 3



Future Samples
snow
rain
Precipitation
0
0.2
0.4
0.6
0.8
1
Temperature
Training Sample
Test Sample

Learning = Prediction ability

- We can't expect any prediction ability, in general.

- We can expect prediction ability, if examples come independently, sometimes.

Definitions

X = input space

$Y = \{0, 1\}$ = output space

$c : X \to Y$ = classifier

$D(x, y)$ = distribution over x and y

$S = (x, y)^m \sim D^m$ = sample set

$e_D(c) \equiv \Pr_{x,y \sim D}(c(x) \neq y)$ = "true error"

$\hat{e}_S(c) \equiv \Pr_{x,y \sim U(S)}(c(x) \neq y)$ = "train error"

Interlude: Flip some coins to see how randomness behaves.

Outline

1. Estimation techniques for comparing classifiers

    - Training error?

    - Test error?

    - Cross validation error?

    - Bayes optimal?

2. Worst-case techniques for comparing classifiers

Training error

Compare classifiers based on training error.

Absurd, no one uses.

Counter example: memorize training set and predict randomly off training set

... or pick the smallest number of heads.

Test error

Compare classifiers based on error on a test set.

- $+$ unbiased

- - confidence unclear

  test error $= 0$

Test error

Compare classifiers based on error on a test set.

- $+$ unbiased

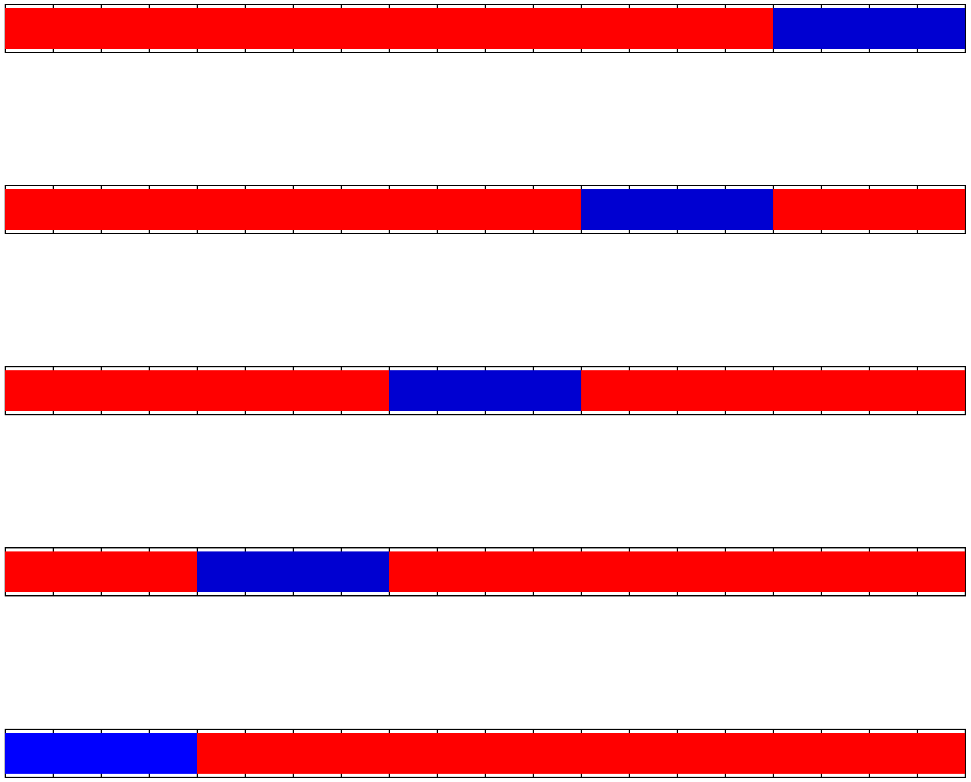
- - confidence unclear

  test error $= 0$

  On one example

Cross validation error

Train and Test many times, then average errors



Cross validation, continued

+ unbiased (for appropriate notion)

- confidence unclear

Example:

1. label drawn uniform from $\{0, 1\}$.
2. Classifier $= 1$ (always) or 0 (always).
3. Learning algorithm $= 1$ if odd 0 if even.

Bayes Optimal

Start with a prior over classifiers. Use Bayes law to update based upon training set evidence and assume the result is optimal.

- + optimal

- - Prior can never be correctly specified in practice

  Bad prior $\Rightarrow$ overfitting.

  Example: Assume uniform prior on n classifiers with zero error rate.

  Real data is totally noisy.

  Acquire $\log(n)$ examples, and overfit.

Outline

1. Estimation techniques for comparing classifiers

   - Training error: useless, untrustworthy.
   - Test error: Useful, if you have a large test set.
   - Cross validation error: Probably the best here
   - Bayes optimal: not a form of validation

2. Worst-case techniques for comparing classifiers

Outline

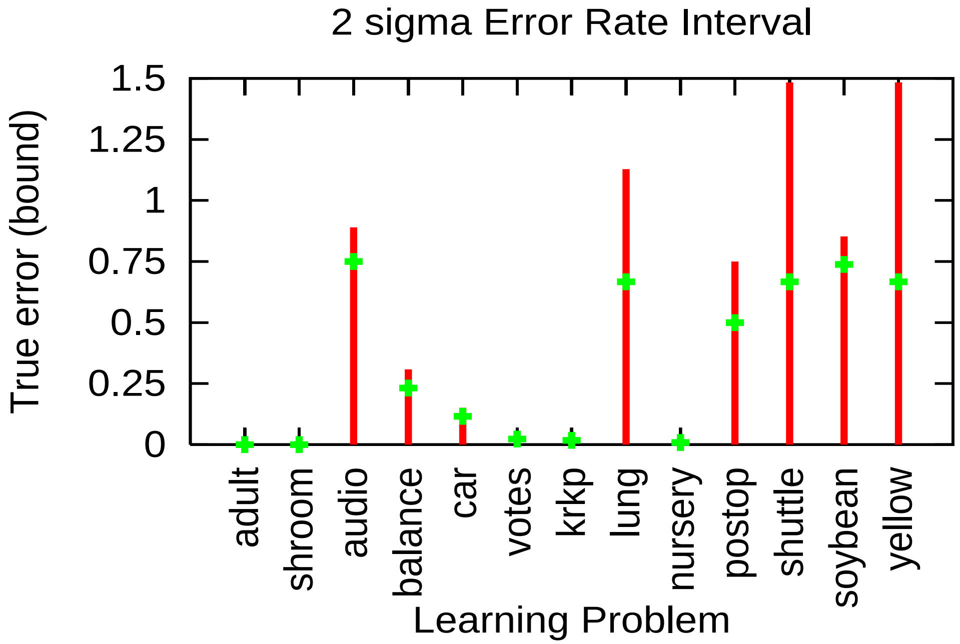
1. Estimation techniques for comparing classifiers

2. Worst-case techniques for comparing classifiers
   - Empirical confidence interval?
   - AIC/BIC/MDL?
   - Real confidence intervals?

Empirical confidence interval

k= number of test errors, m = number of examples

$\mu = \frac{k}{m}$, $\sigma^2 = \frac{2k(m-k)}{m(m-1)}$, pick bound $= \frac{k}{m} + 2\sigma$



AIC/BIC/MDL

Empirically, very “large” classifiers can be useful so this often provides the wrong intuition or is not easy to apply. Some “large” classifiers:

- Bayes optimal (full posterior)
- neural networks
- boosted decision trees.

Analysis is almost always asymptotic $\Rightarrow$ it may not apply to your finite dataset.

1. Estimation techniques for comparing classifiers

2. Worst-case techniques for comparing classifiers

   - Empirical confidence interval: falls down on the job

   - AIC/BIC/MDL: difficult to apply

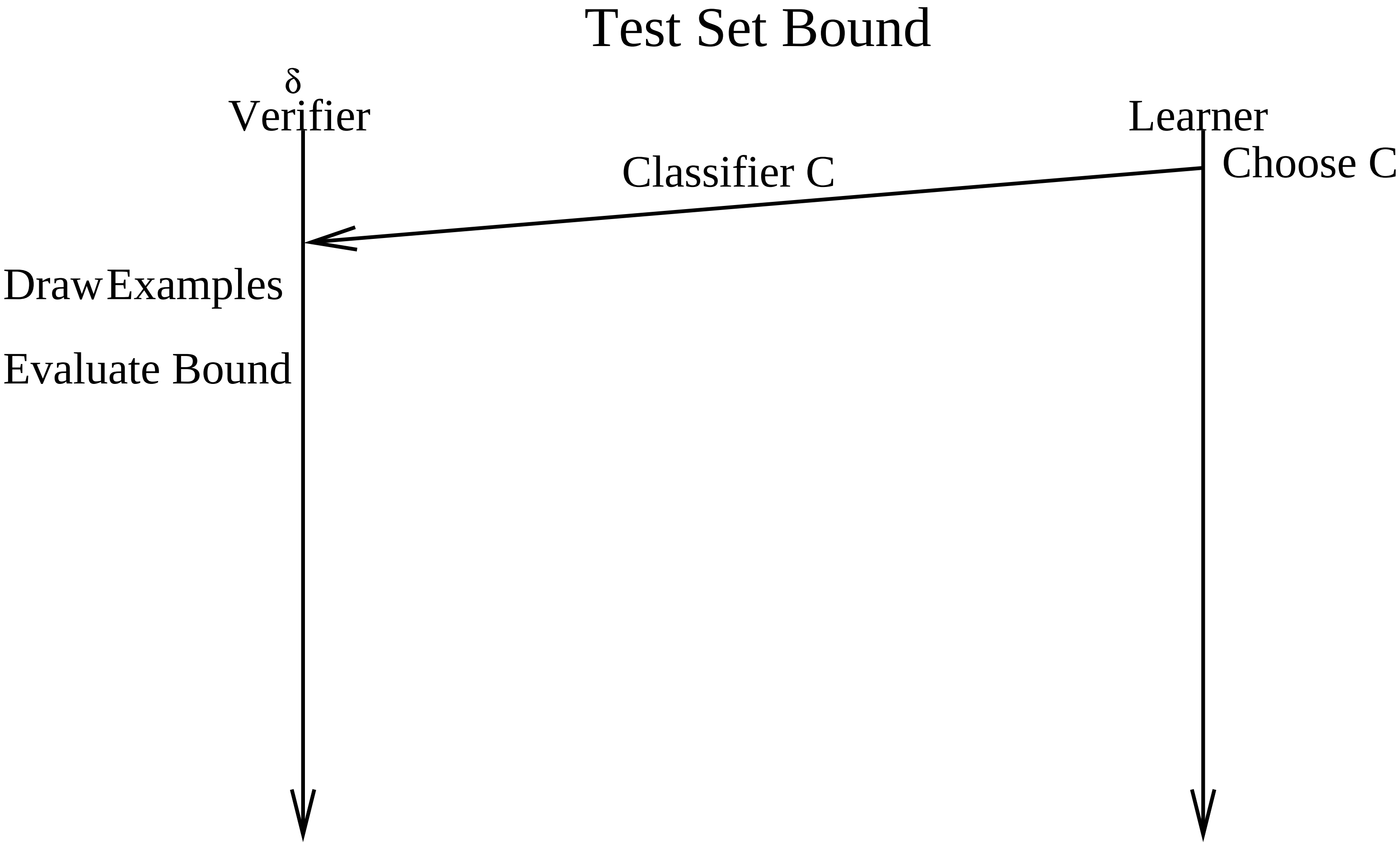
   - Real confidence intervals

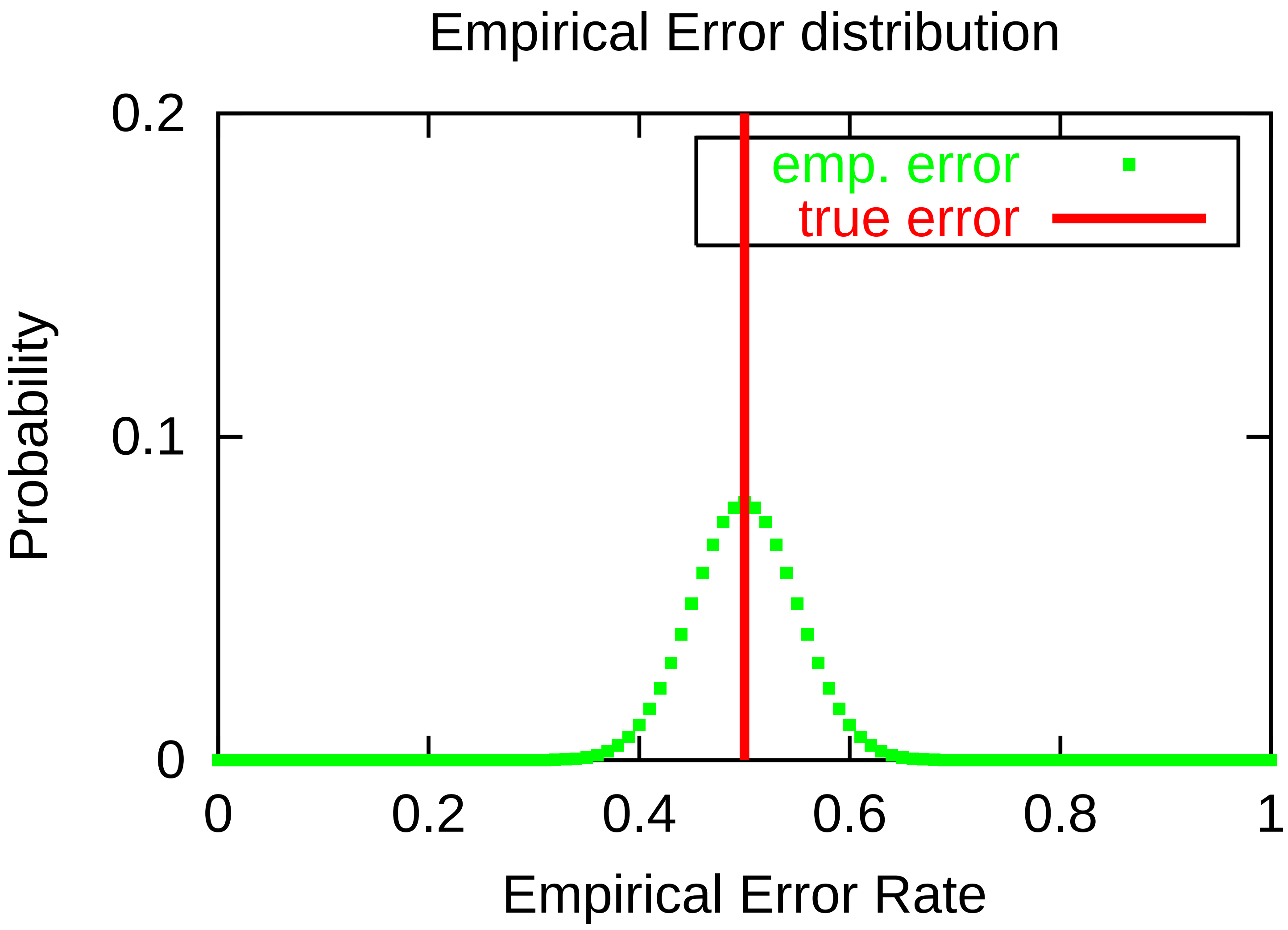
     - Test set bounds

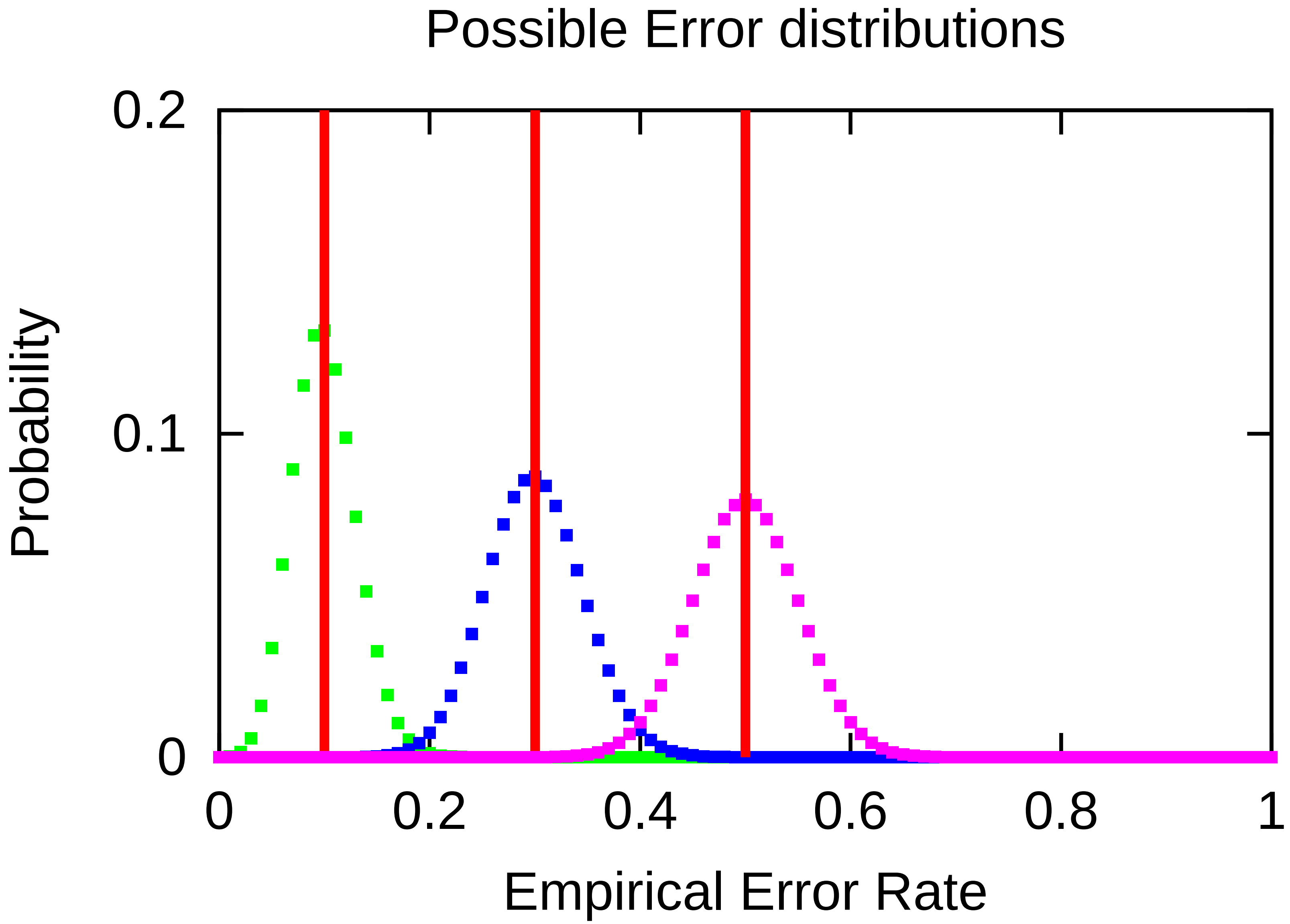
     - Train set bounds

     - Results

Interpretation: Interactive Proof of Learning




Empirical Error distribution
0.2
0.1
0
Probability
emp. error
true error
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

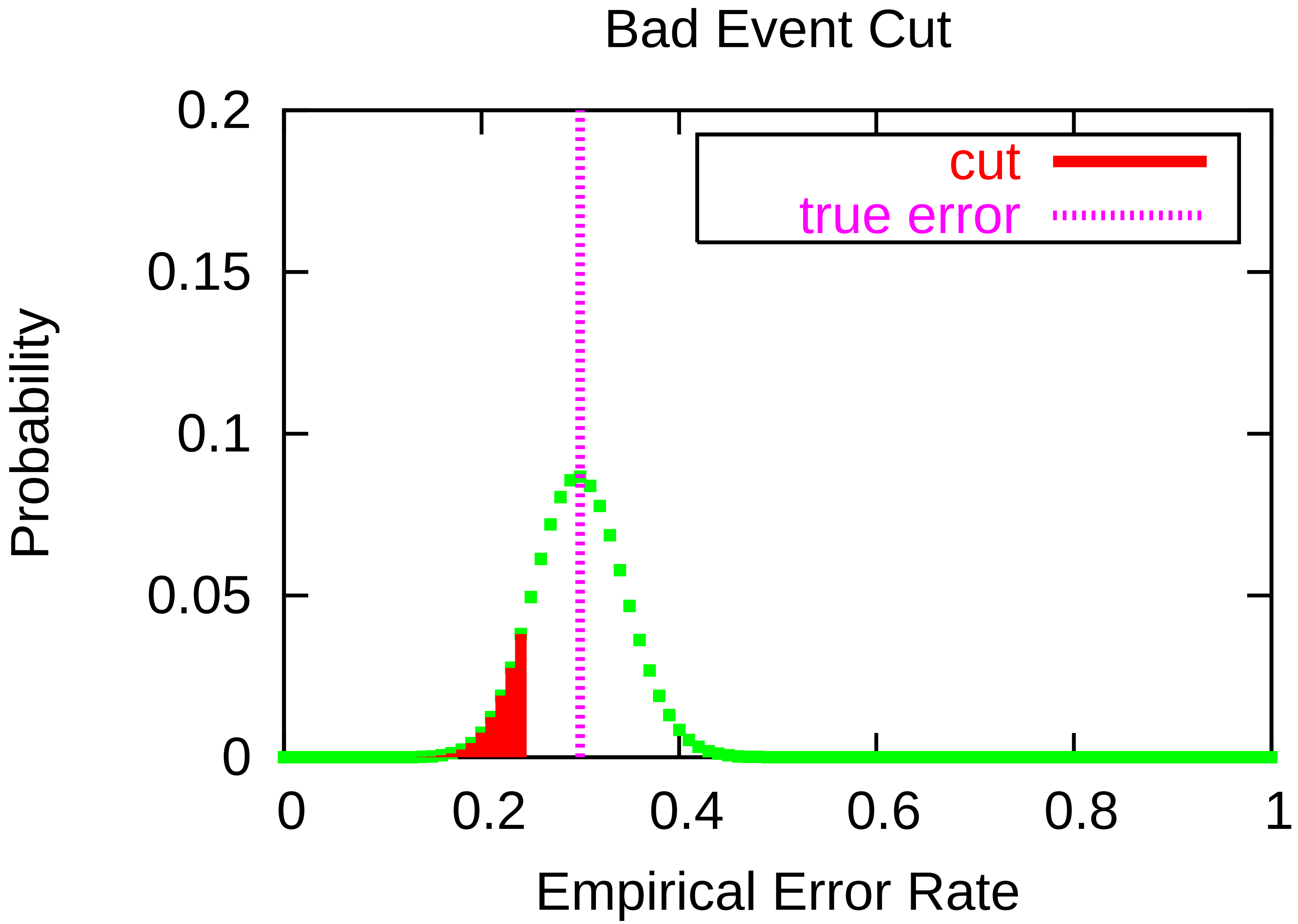

Possible Error distributions
Probability
Empirical Error Rate
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1

How can you make a Bound?

- $p =$ bias of a coin
- $\hat{p} =$ observed head rate in m flips

$$\Pr(p - \hat{p} > \epsilon) = \sum_{i=0}^{\lfloor m(p-\epsilon) \rfloor} \binom{m}{i} p^i (1-p)^{m-i}$$

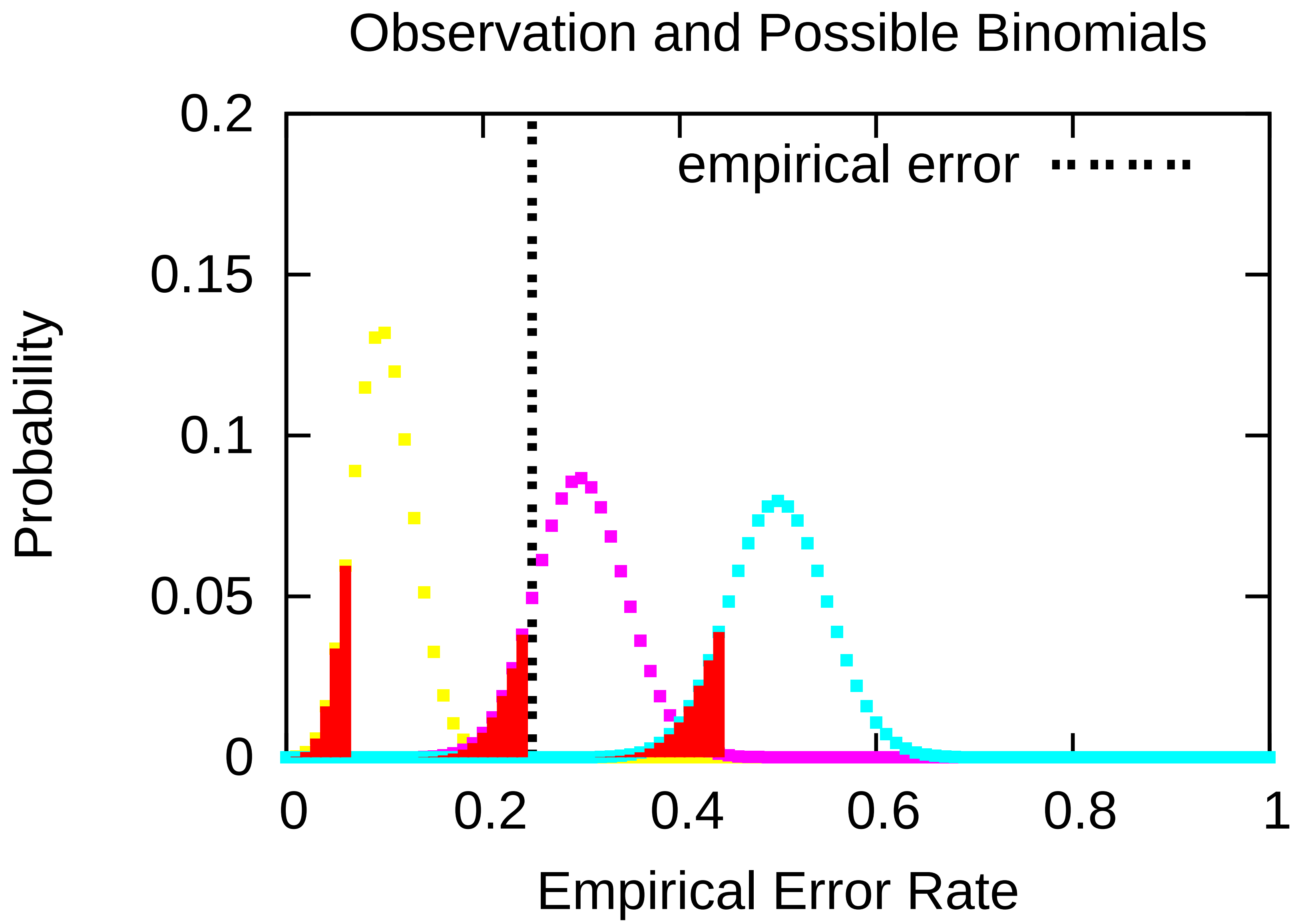
$$\leq e^{-2m\epsilon^2}$$

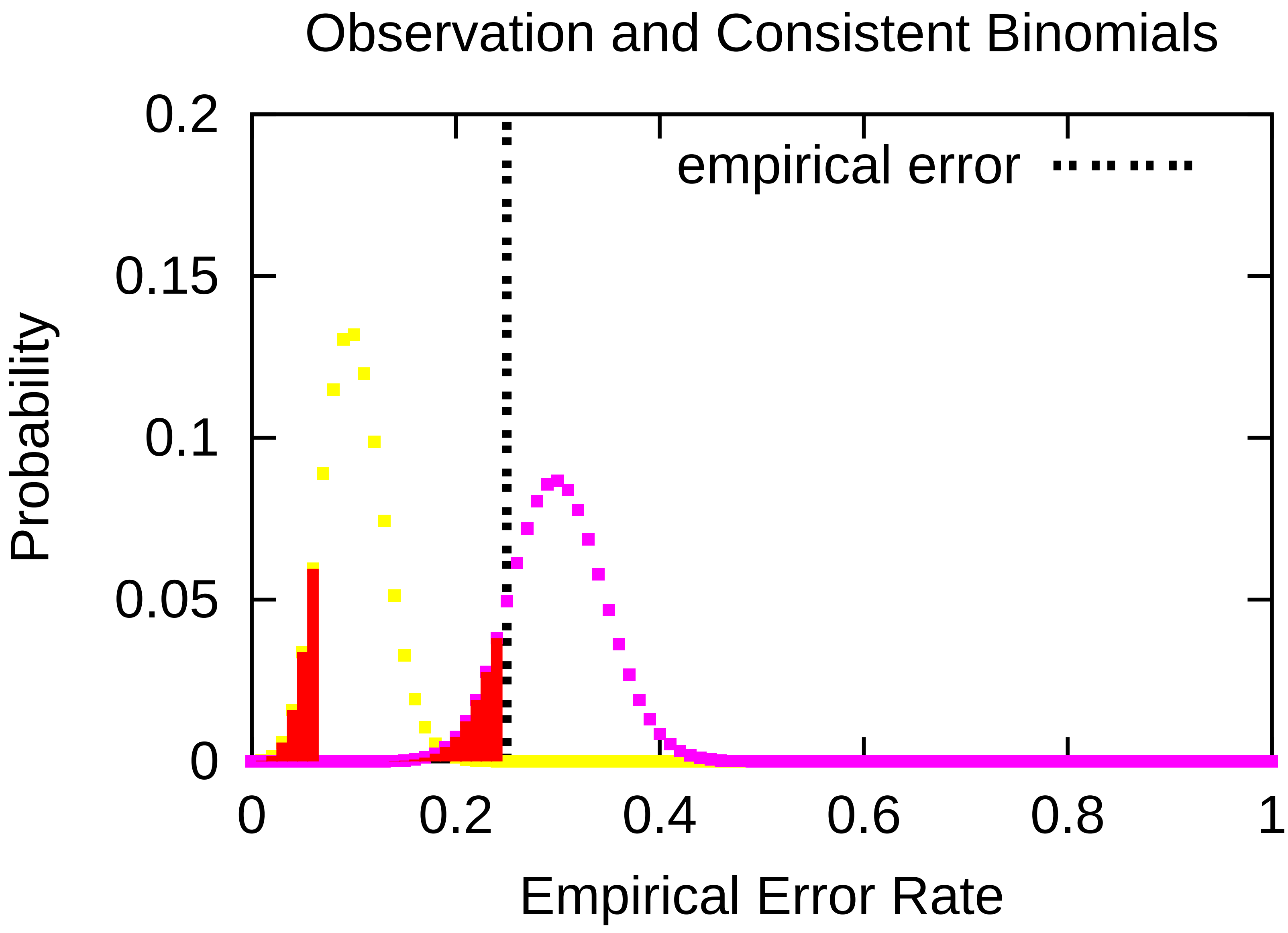


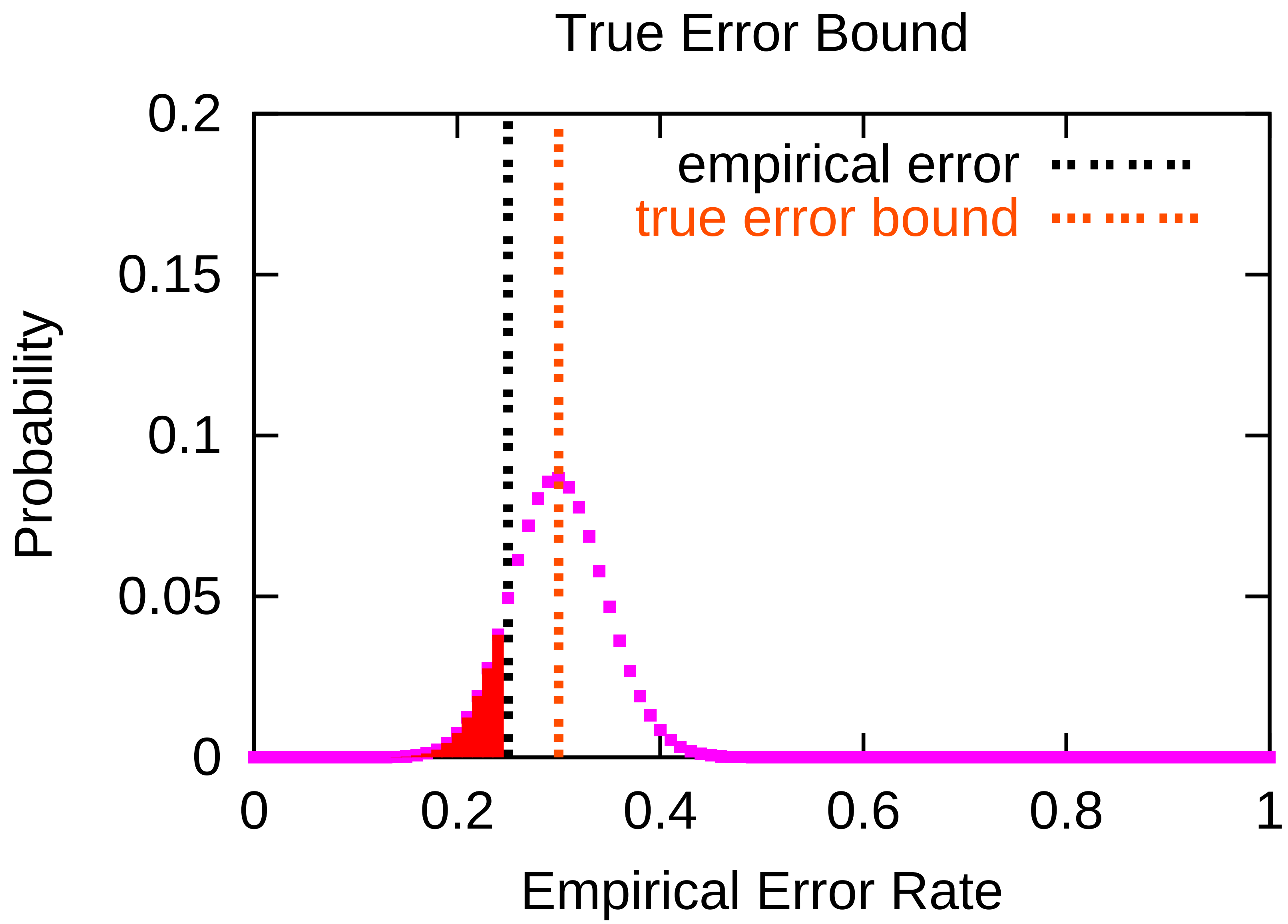
Bad Event Cut
cut
true error
Probability
Empirical Error Rate
0
0.05
0.1
0.15
0.2
0
0.2
0.4
0.6
0.8
1

Q: Given the tail is cut, what is our bound?

A: The largest true error that did not cut the observation.


Observation and Possible Binomials
empirical error
Probability
Empirical Error Rate
0
0.05
0.1
0.15
0.2
0
0.2
0.4
0.6
0.8
1


Observation and Consistent Binomials
empirical error
Probability
Empirical Error Rate
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1


True Error Bound
Probability
Empirical Error Rate
empirical error
true error bound
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

Train set bounds

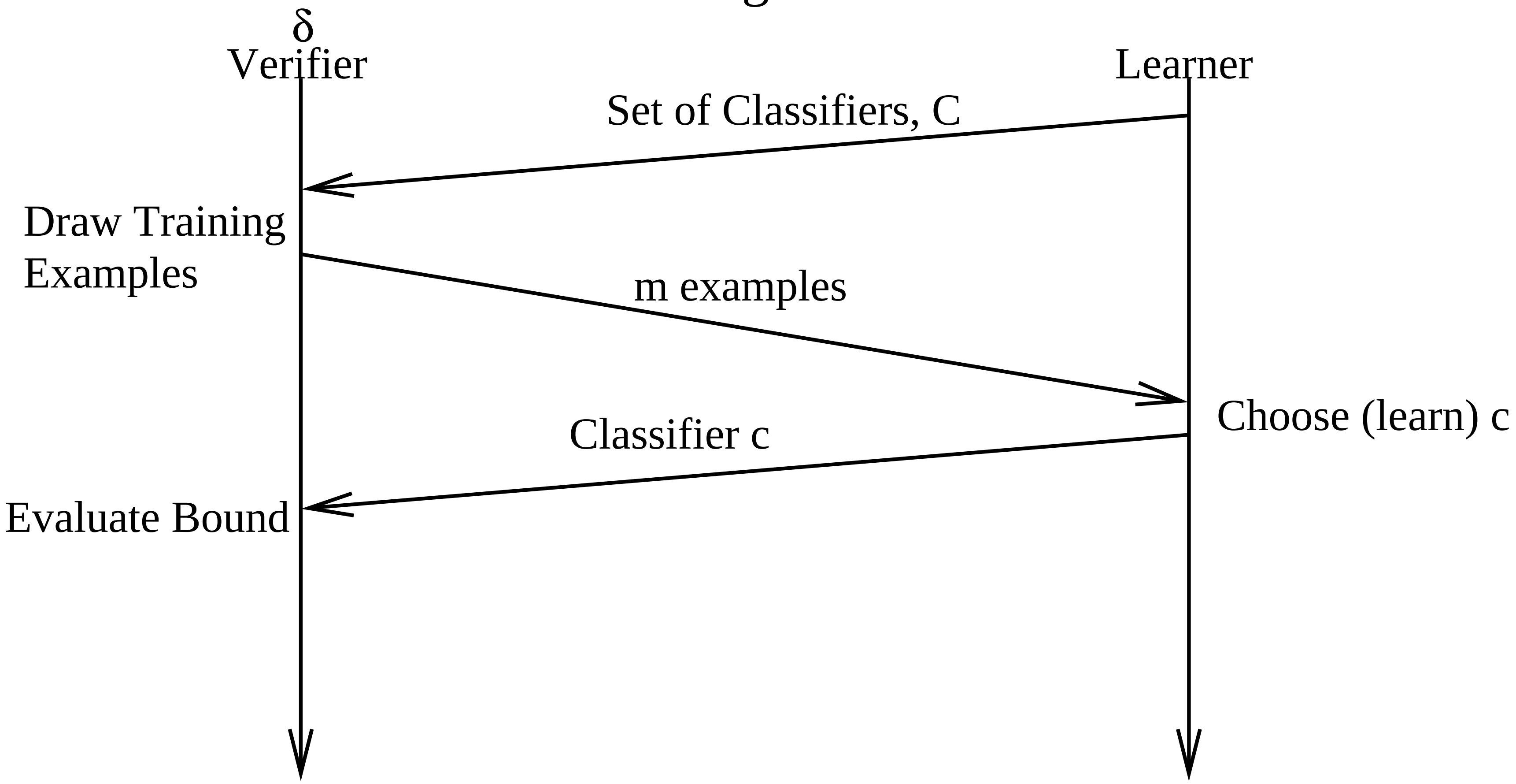
We can verify if we have learned.

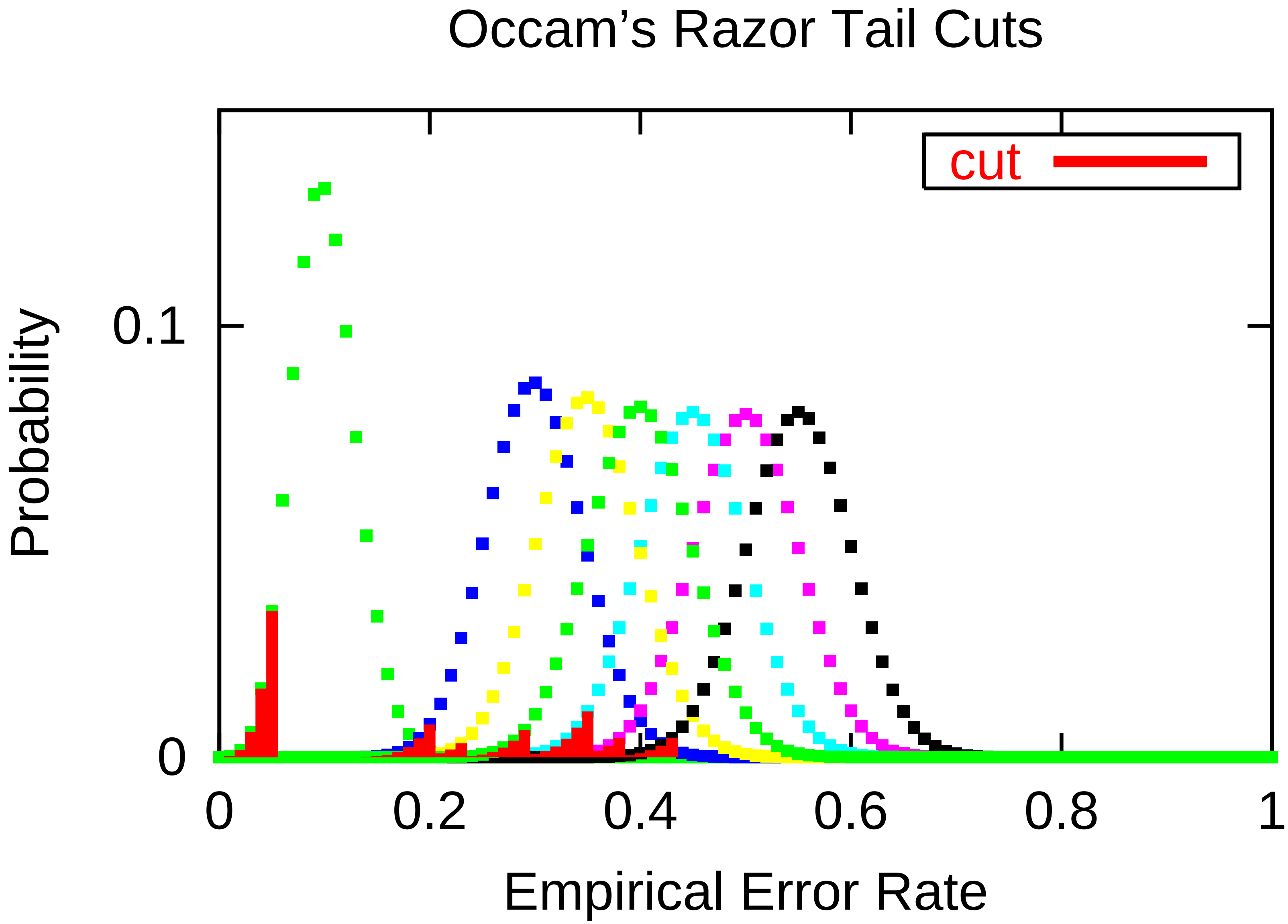
But: learning on one classifier is degenerate.

Q: What if there are multiple classifiers?

A: Cut all tails and add up the probability of error

Training Set Bound

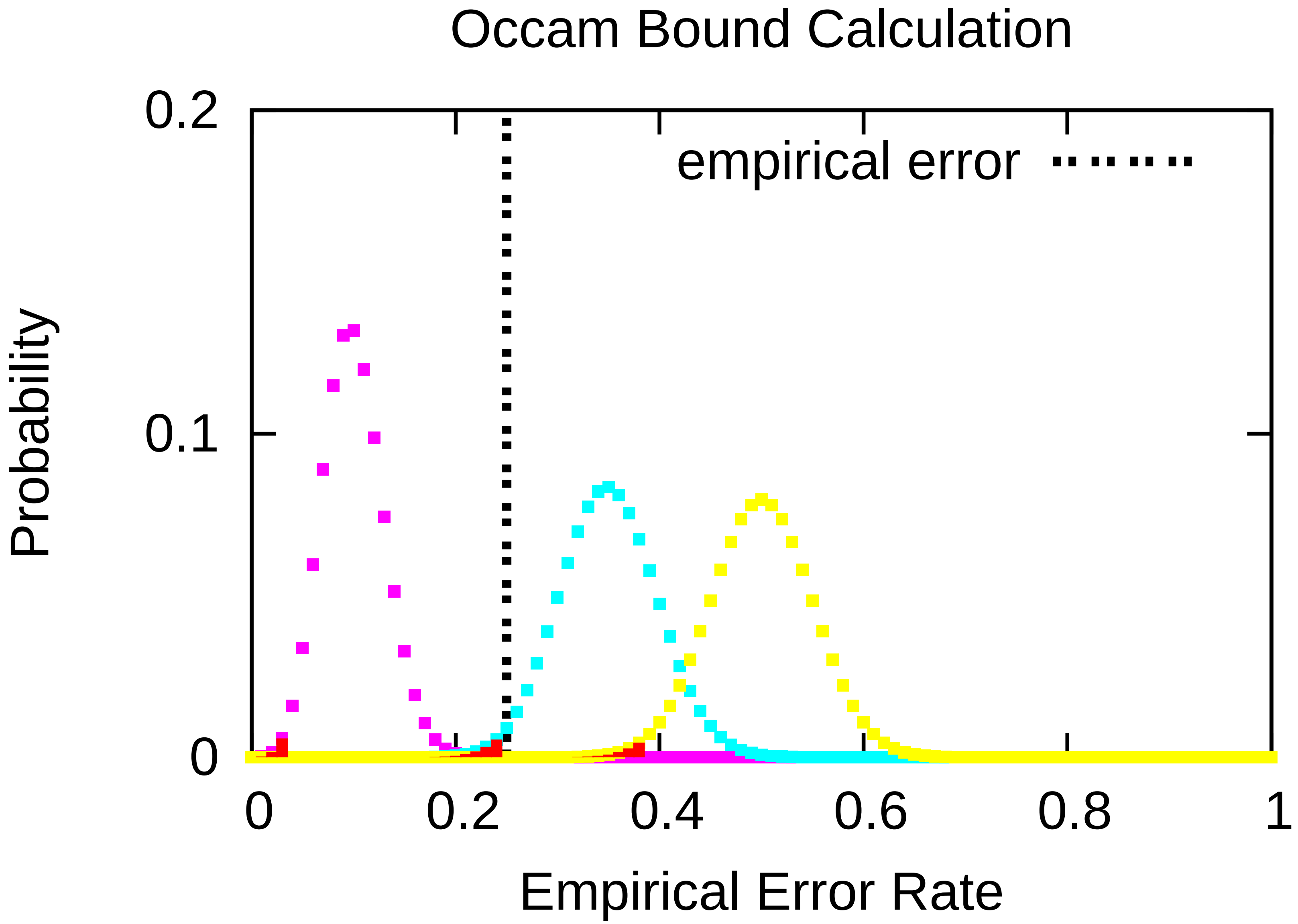




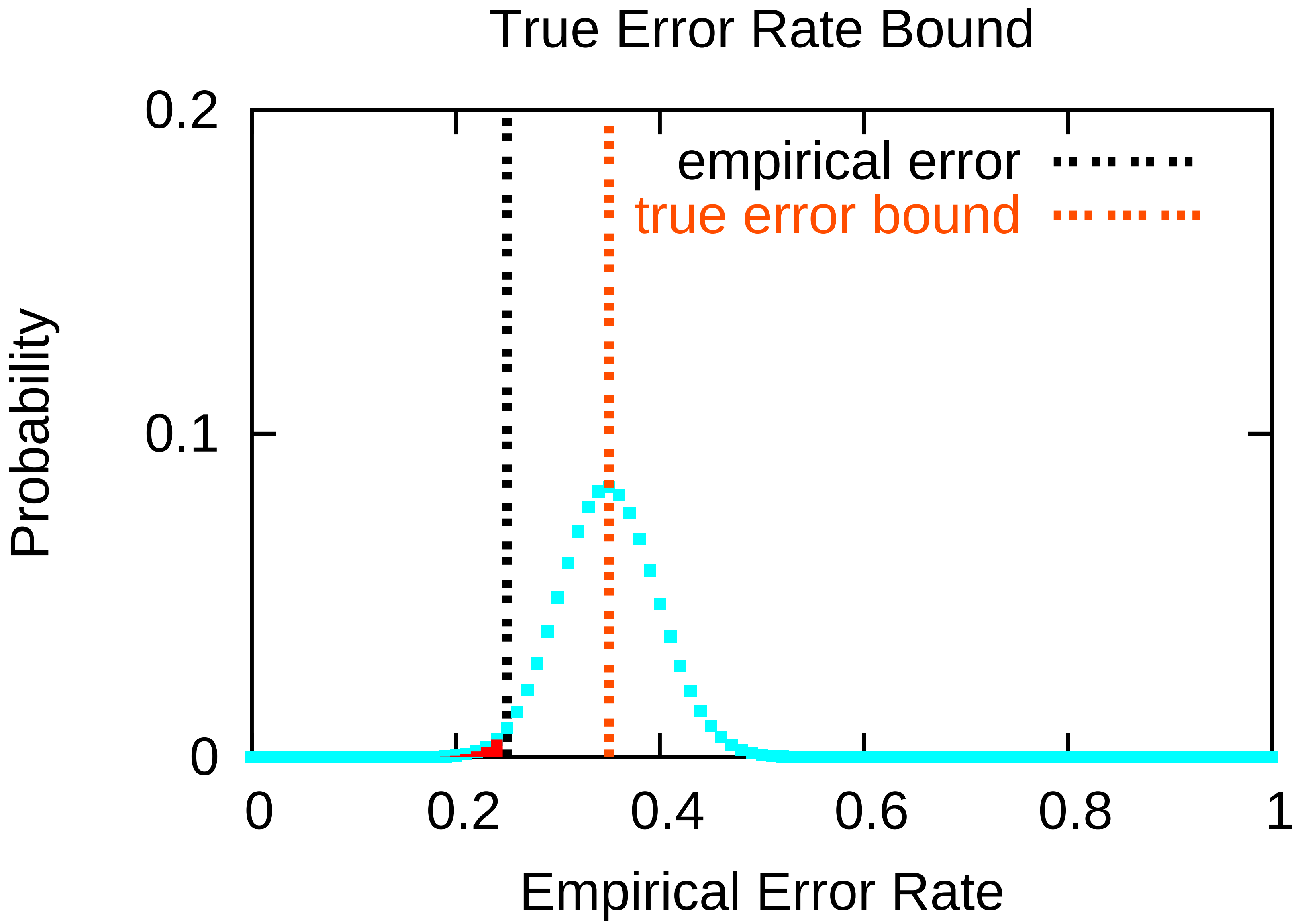
Occam's Razor Tail Cuts
Probability
0.1
0
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate
cut

Q: How do we calculate a bound?

A: Given c we know $p(c)$ so calculate the worst classifier consistent with a binomial tail cut at mass $\delta p(c)$.



Occam Bound Calculation
0.2
0.1
0
empirical error
Probability
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate


True Error Rate Bound
Probability
0.2
0.1
0
empirical error
true error bound
0
0.2
0.4
0.6
0.8
1
Empirical Error Rate

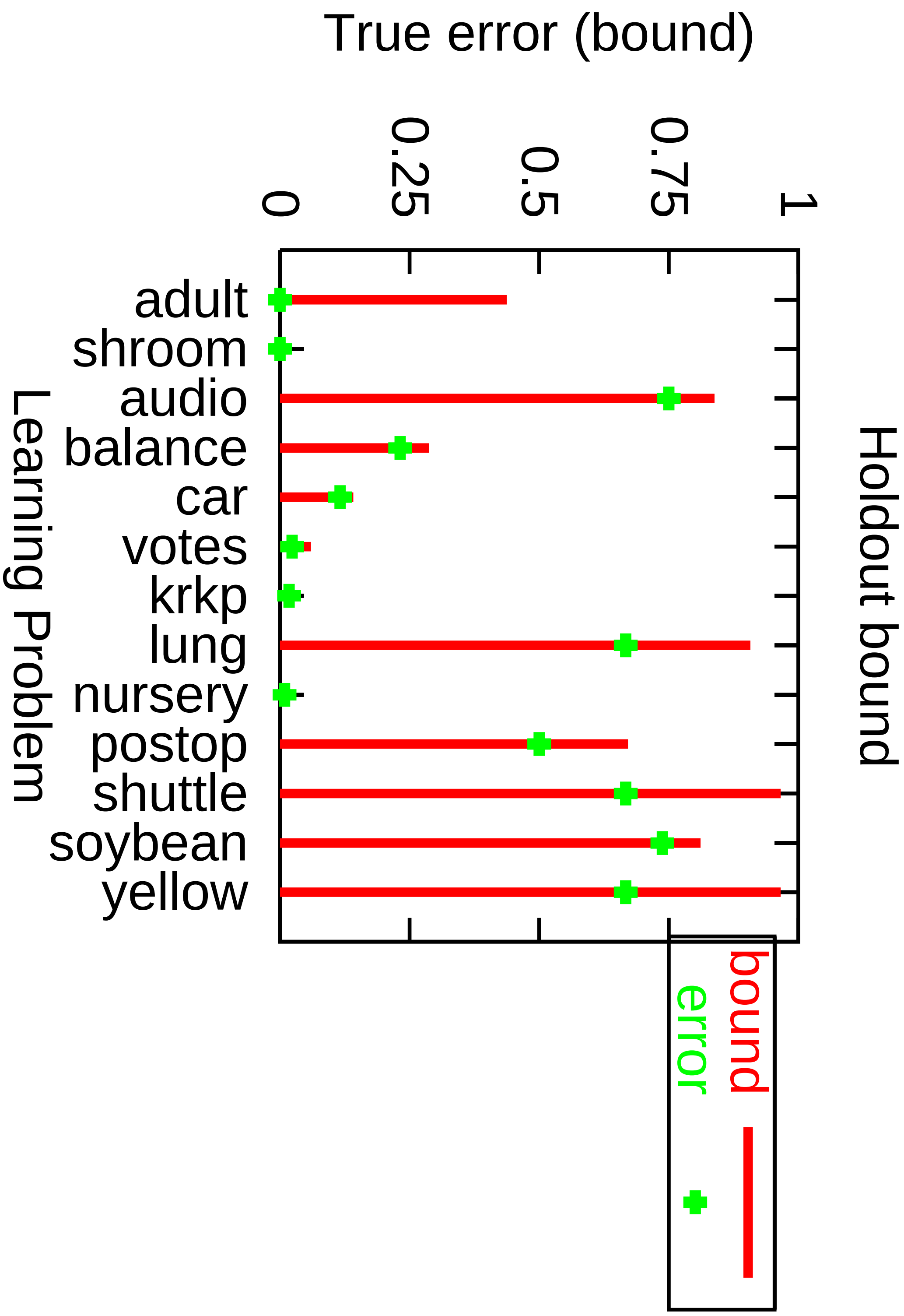
Side note

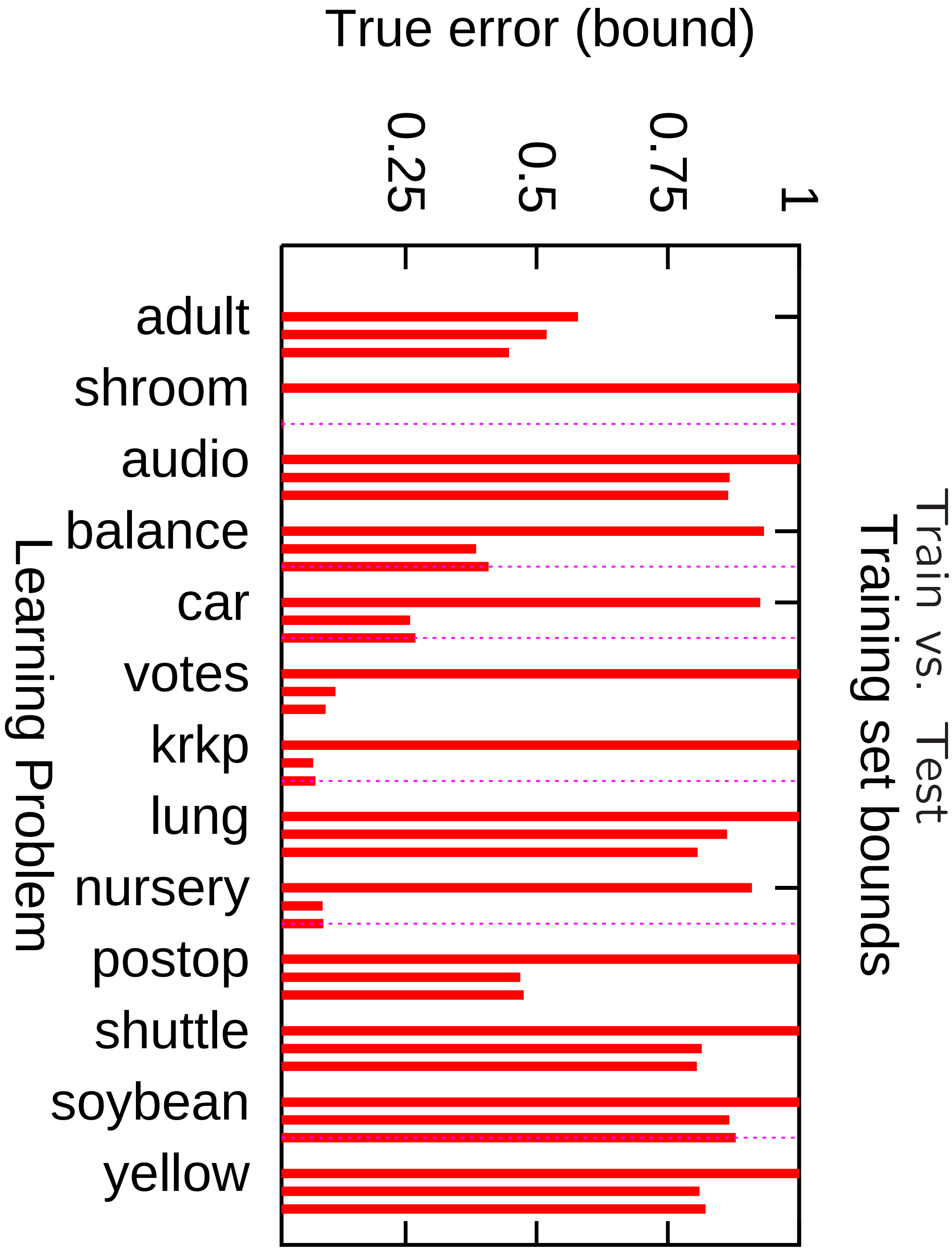
1. We have covered only the simplest and most basic of methods for train set based classifiers.

2. There are also other “interactive proof of learning” protocols. Can you guess how they work?

Results: Decision Trees

- ID3 decision tree + pruning
- probability of failure $= \delta = 0.1$
- Discrete problems from UCI database of Machine Learning problems.
- 100% of data used for training set bounds
- 80%/20% Train/Test split for test set bounds
- Minimal selection bias


True error (bound)
0
0.25
0.5
0.75
1
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
Learning Problem
Holdout bound
bound
error


True error (bound)
0.25
0.5
0.75
1
adult
shroom
audio
balance
car
votes
krkp
lung
nursery
postop
shuttle
soybean
yellow
Learning Problem
Train vs. Test
Training set bounds

1. Estimation techniques for comparing classifiers

2. Worst-case techniques for comparing classifiers

    - Empirical confidence interval: falls down on the job
    - AIC/BIC/MDL: difficult to apply
    - Real confidence intervals: the real winner
        - Test set bounds: Generally useful
        - Train set bounds: sometimes useful
        - Results

Conclusion

Use a confidence interval best for comparison.

Amongst estimation methods, test set or K-fold cross validation are reasonable.

Some code at:

http://www-2.cs.cmu.edu/~jcl/programs/bound/bound.html