# The Complexity of Learning Verification

John Langford

# Ground Rules

1. I assume you know the basics
   (a) skip proofs
   (b) stay at a high level
2. No citations (for clarity).
3. Don't hesitate to ask questions

What is the complexity of verifying successful learning?

Answer: Relates to description length

1. Focus on Classification

2. Assume I.I.D. samples from unknown $D$ on $X \times \{0, 1\}$

3. Description length = Probability measure on a set of objects

# Outline

1. Basics: one classifier

2. Description length of classifier

3. Description length of labels given unlabeled data

4. Bit Questions

5. Relationship with other complexities

## Model: Definitions

$X$ = input space

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

$c : X \to Y$ = classifier

All samples are drawn independently from some unknown distribution $D(x, y)$.

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

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

$\hat{c}_S \equiv \sum_{x,y \in S} I\left[c(x) \neq y\right]$ = number of errors on $S$

# Model: basic quantities

Theorem: (Test Set Bound) For all classifiers $c$, for all $D$, for all $\delta \in (0,1]$ with probability $1-\delta$:

$$\text{(Exact) } c_D \leq \max\left\{ p : \sum_{i=0}^{\hat{c}_S} \binom{m}{i} p^i (1-p)^{m-i} \geq \delta \right\} \equiv \text{Test}(m, \hat{c}_S, \delta)$$

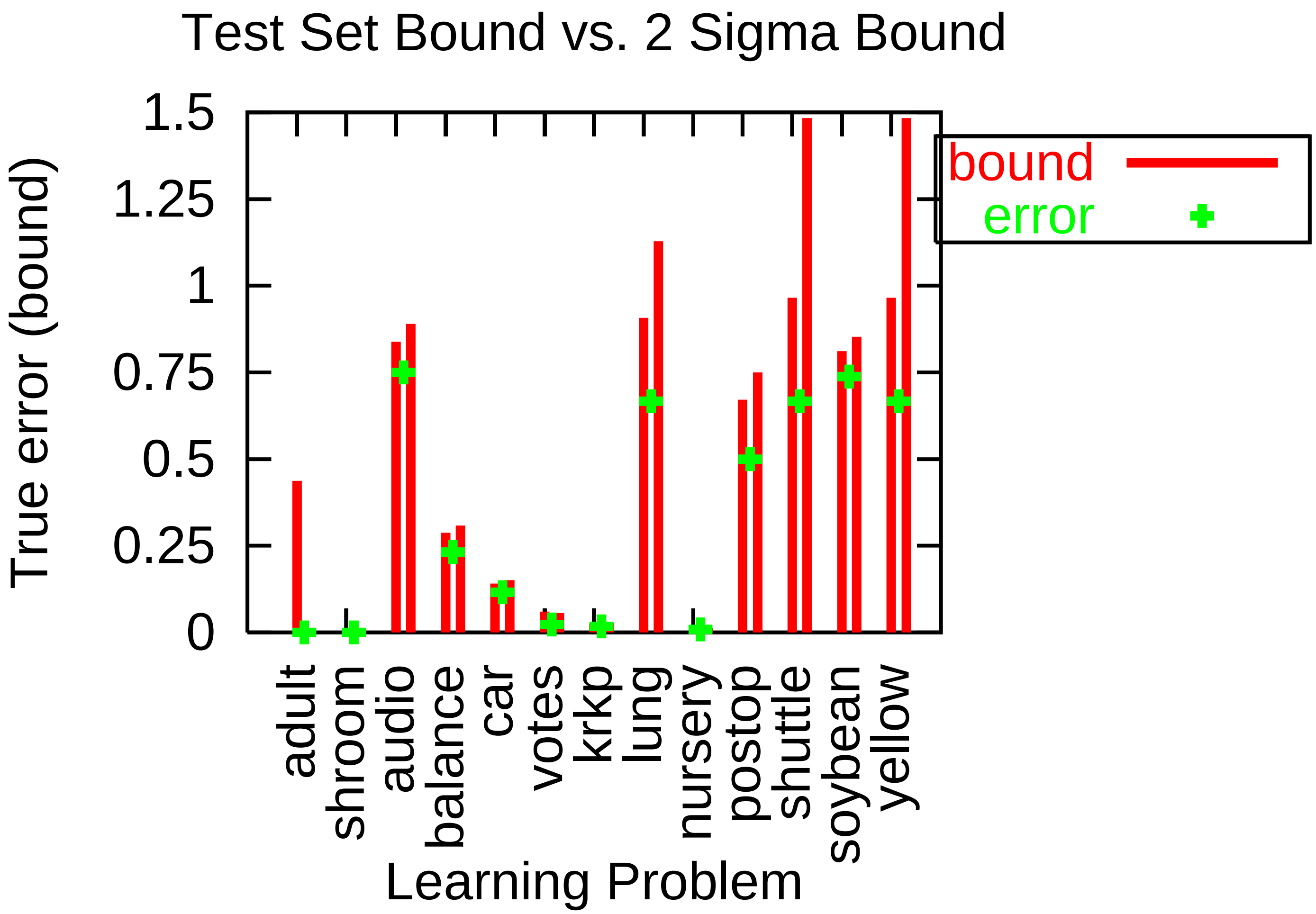
$$\text{(Close Approximation) } \Rightarrow \text{KL}\left(\frac{\hat{c}_S}{m} || c_D\right) \leq \frac{\ln \frac{1}{\delta}}{m}$$

where $\text{KL}(q||p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ for $q < p$

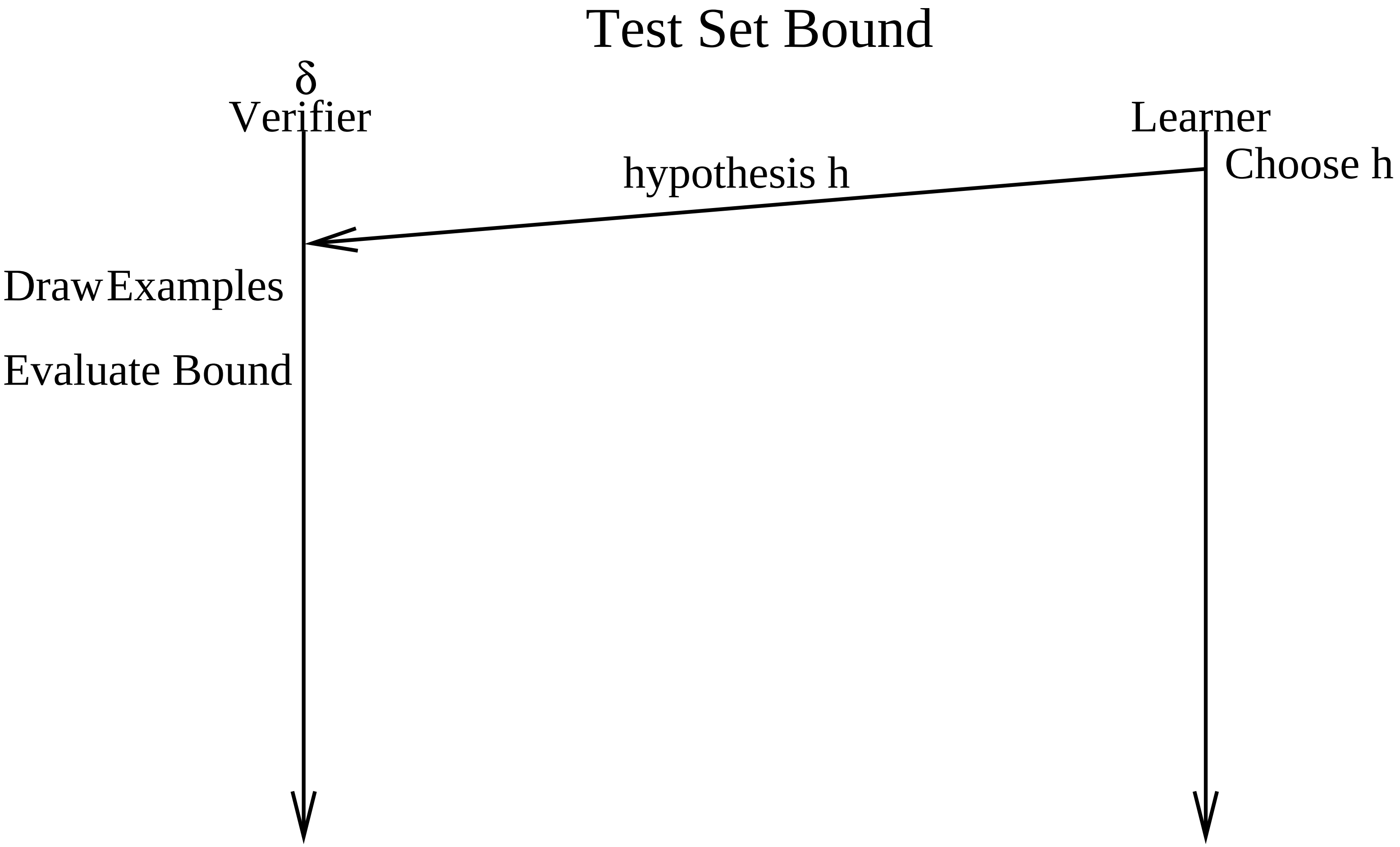
$$\text{(Loose Approximation) } \Rightarrow c_D \leq \frac{\hat{c}_S}{m} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

Demo

## Test set bounds superior to common practice



## Interpretation: Interactive Proof of Learning



# Outline

1. Basics: One Classifier

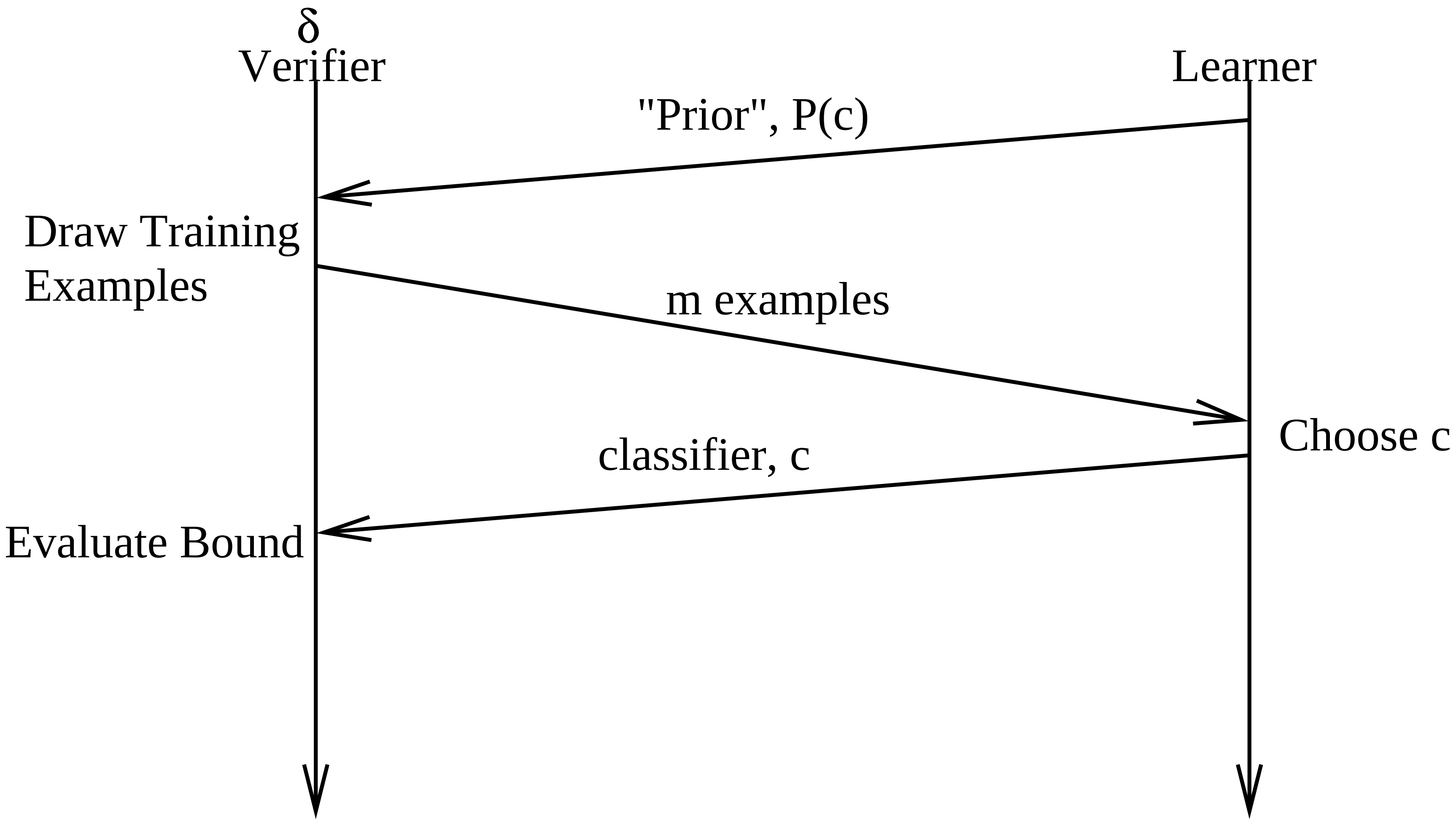
2. Description length of Classifier

3. Description length of labels given unlabeled data

4. Bit Questions

5. Relationship with other Complexities

# Occam's Razor Bound Protocol



## Occam's Razor Bound

Theorem: (Occam's Razor Bound) For all "priors" $P(c)$ over the classifiers $c$, for all $D$, for all $\delta \in (0,1]$ with probability at least $1-\delta$:

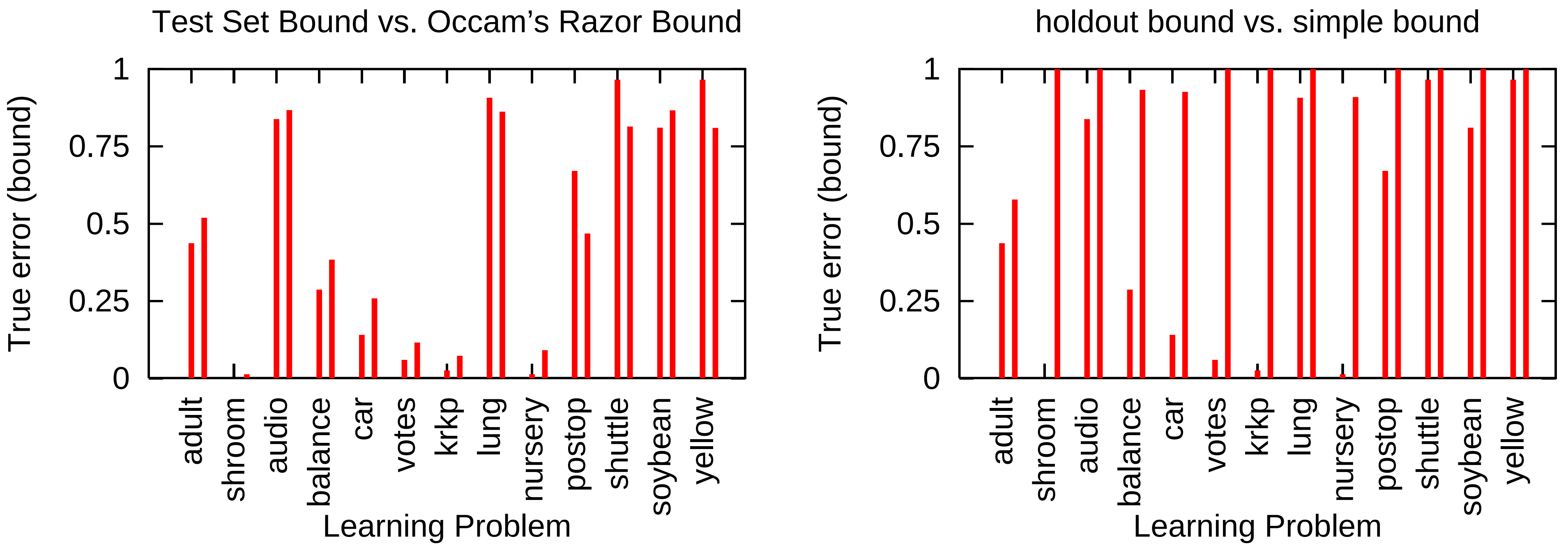
$$\forall c: \; c_D \leq \text{Test}(m, \hat{c}_S, \delta P(c))$$

$$\Rightarrow \text{KL}\left(\frac{\hat{c}_S}{m} || c_D\right) \leq \frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{m}$$

$$\Rightarrow c_D \leq \frac{\hat{c}_S}{m} + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{2m}}$$

Compare with test set bound: $\delta \to \delta P(c)$

# Results for Decision Trees, 80/20 train/test split




Left bar = test set bound, right bar = Occam's Razor or Uniform Bound

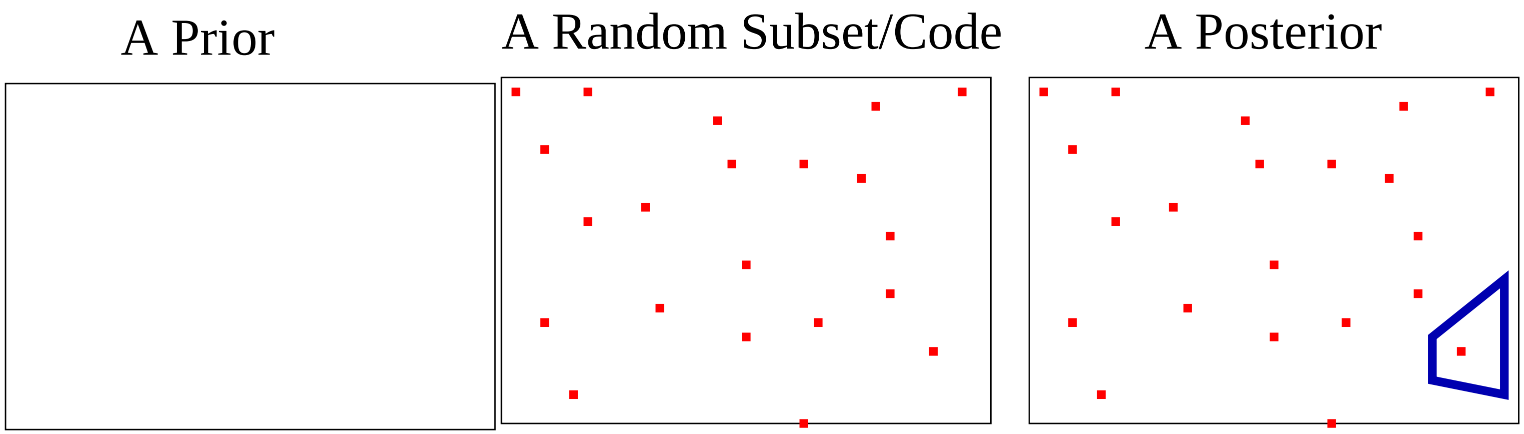
Prior is *necessary* for sane results

# Variations

1. PAC-Bayes Bound: $c \rightarrow$random process generating $c$. $P(c) \rightarrow P(c|\text{process})$

2. Sample Compression Bound: $P(c)$ specified implicitly by algorithm, $m \rightarrow m_c$

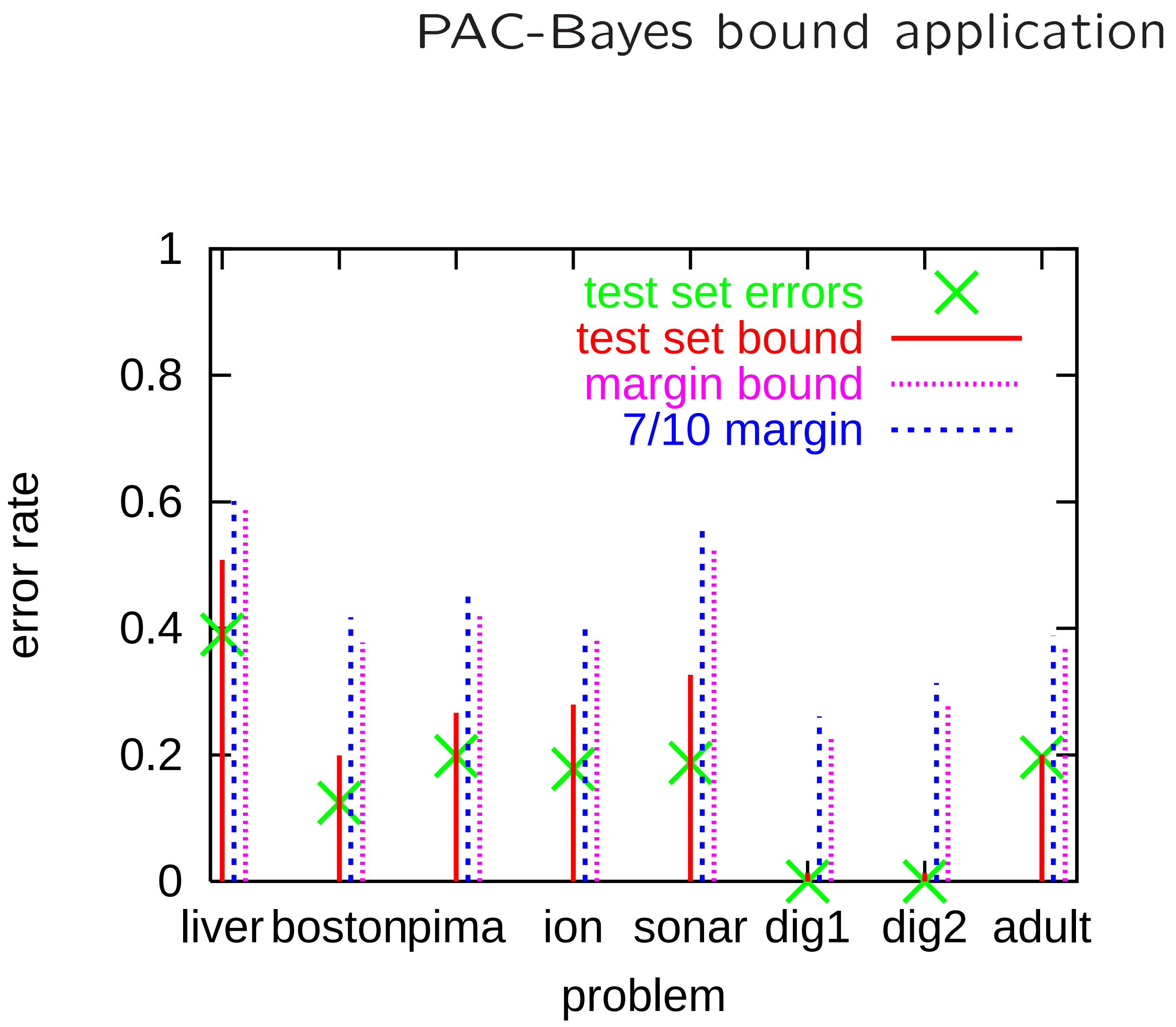
# The PAC-Bayes Bound




For all $Q(c)$, description length of $c \sim Q|P \leq \mathrm{KL}(Q||P) + \ln(m+1)$

Theorem: (PAC-Bayes Bound) For all "priors" $P(c)$ over the classifiers $c$, for all $D$, for all $\delta \in (0,1]$ with probability $\geq 1-\delta$:

$$\forall Q(c): \ \mathrm{KL}\left(E_{c\sim Q}\frac{\hat{c}_S}{m} || E_{c\sim Q} c_D\right) \leq \frac{\mathrm{KL}(Q||P) + \ln\frac{m+1}{\delta}}{m}$$



PAC-Bayes bound application
test set errors
test set bound
margin bound
7/10 margin
error rate
1
0.8
0.6
0.4
0.2
0
liver
boston
pima
ion
sonar
dig1
dig2
adult
problem

# Outline

1. Basics: One Classifier

2. Description length of Classifier

3. Description length of labels given unlabeled data

4. Bit Questions

5. Relationship with other Complexities

## Transductive PAC-MDL Bound

$S = (x,y)^m \sim D^m$ = train set, $S' = (x,y)^n \sim D^n$ = test set

$\sigma : X^{|S|+|S'|} \to Y^{|S|+|S'|}$ = transductive classifier

$\hat{\sigma}_S \equiv \sum_{x,y \in S} I(\sigma(S,S')(x) \neq y)$ = train error

$\hat{\sigma}_{S'} \equiv \sum_{x,y \in S'} I(\sigma(S,S')(x) \neq y)$ = test error

Theorem: (PAC-MDL Bound) For all measures $P(\sigma)$, with probability $1-\delta$ over draws of $S, S' \sim D^{m+n}$:

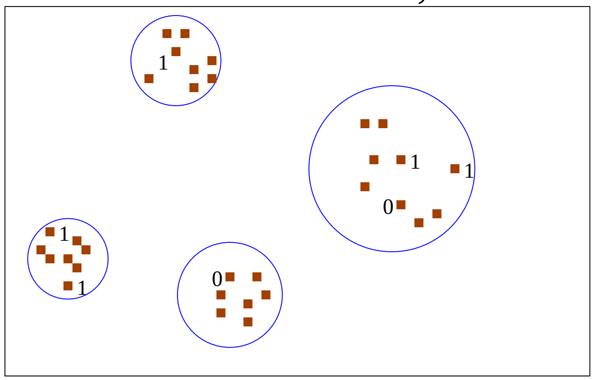
$$\forall \sigma : \ \hat{\sigma}_{S'} \leq \max \left\{ b : \sum_{t=b}^{\hat{\sigma}_S + b} \frac{\binom{m}{\hat{\sigma}_S + b - t}\binom{n}{t}}{\binom{m+n}{\hat{\sigma}_S + b}} \geq P(\sigma)\delta \right\}$$

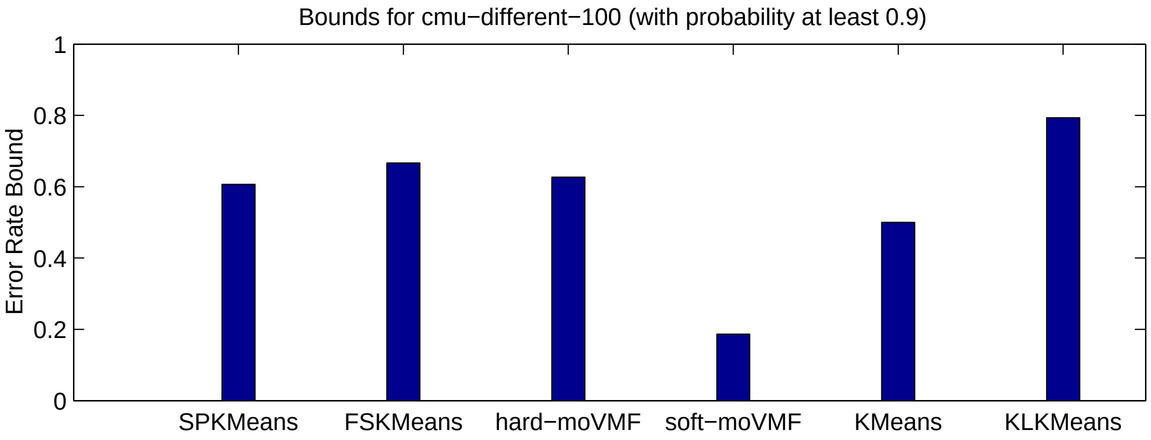
# Clustering application

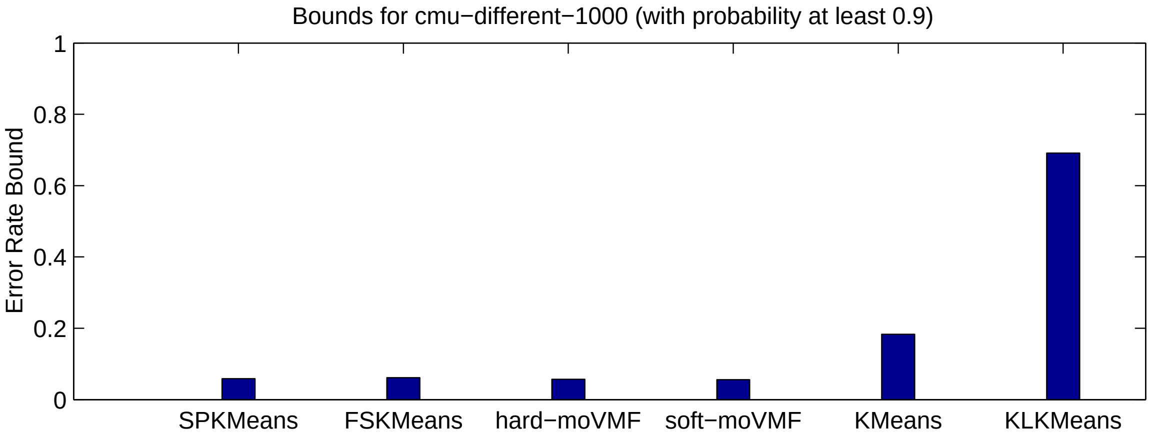
Suppose $z^t = (x,y)^m, x^n$

1. Cluster on $x^{m+n}$
2. Choose best constant label $y$ for each cluster using $y^m$.
3. For each $x$, predict the label of it's cluster.

A Set of Clustered Points, With Labels






$$P(\sigma) = \frac{2^{-k}}{k(k+1)r}$$

$k$ = number of clusters, $r$ = number of random initializations

## Implications (Transduction $\rightarrow$ Induction Reductions)

1. Occam's Razor Bound: (Limit $n \rightarrow \infty$)

2. VC bound (Finite $n$ reduction)

3. Rademacher Complexity (?)

## Outline

1. Basics: One Classifier
2. Description length of Classifier
3. Description length of labels given unlabeled data
4. Bit Questions
5. Relationship with other Complexities

## Is the bit description length the final answer?

No. Extra info $\Rightarrow$ tighter bounds.

Forms of extra information:

1. Error rates of other classifiers $\Rightarrow$ Shell bounds

2. Partial knowledge of the correct classifier $\Rightarrow$ Bayesian analysis

## Why the bit?

1. Easy to compute. (*Always* computable on any computer!) Contrast: VC dimension/entropy, Covering Number, Rademacher

2. Flexible. (Algorithm doesn't need to be guessed in advance. Practitioner can just adapt it.) Contrast: ERM analysis

3. Tight. Matching "lower upper" bounds up to $\delta \to \delta - \delta^2$.

4. Nicely quantifiable notion of complexity. Contrast: most computational complexities

# Outline

1. Basics: One Classifier
2. Description length of Classifier
3. Description length of labels given unlabeled data
4. Bit Questions
5. Relationship with other Complexities

## The Complexity of Learning

1. (Valiant-style) PAC analysis: Requires additional assumption "The target concept is a monomial." Mixes Information theoretic & computational complexity.

2. (Vapnik-style) ERM&SRM analysis: Bit bounds don't fix algorithm $\Rightarrow$ applies to many more algorithms.

3. Bayesian analysis: Requires additional assumption: correct prior.

4. Online learning/Competitive analysis: More fundamental (no i.i.d. assumption). Is competition learning?

# MDL/MAP

Tight "lower upper" bounds $\Rightarrow$ can't do better than bound minimization in general.

MDL/MAP is more agressive than bound$\Rightarrow$ MDL/MAP can overfit.

(See Peter Grunwald's talk)

Nonclassification problems

There exists reduction from $\{$Reinforcement Learning, Regression, Cost Sensitive Classification, etc...$\}$ to Classification

$\Rightarrow$

Verification of Classification $\Rightarrow$ Verification of $\{$Reinforcement Learning, Regression, Cost Sensitive Classification, etc...$\}$

# Conclusion

Description length = complexity of learning verification

1. Flexible. Easy to adapt to many learning algorithms.

2. Powerful. The tightest empirical and theoretical results.

3. Easy. Many people can use and tweak the bit bounds.

For many details, code & papers, see:

http://hunch.net/~jl/projects/prediction_bounds/prediction_bounds.html