# Combining Test and Train Set Bounds

John Langford

Test set bound (applied) = bound on future error rate using holdout set

Train set bound (COLT) = bound on future error rate using training set

Can we use *both* sources of information?

## The simple method

$\delta$ = probability of bound failure.

$\mathrm{bound}_{\mathrm{test}}(\delta)$

$\mathrm{bound}_{\mathrm{train}}(\delta)$

Use $\min\left(\mathrm{bound}_{\mathrm{test}}\left(\frac{\delta}{2}\right), \mathrm{bound}_{\mathrm{train}}\left(\frac{\delta}{2}\right)\right)$

Can we do better?

For 2 test sets:

$\min\left(\mathrm{bound}_{\mathrm{test}\ 1}\left(\frac{\delta}{2}\right), \mathrm{bound}_{\mathrm{test}\ 2}\left(\frac{\delta}{2}\right)\right)$ is inefficient.

We want:

$$\mathrm{bound}_{\mathrm{test}}(\delta) \to \frac{\mathrm{bound}_{\mathrm{test}}(\delta)}{\sqrt{2}}$$

(or even better)

1. Test & Train set bounds

2. Combinations

3. A combination theorem

4. Technique and Method

## The Model

$X$ = (arbitrary) input space

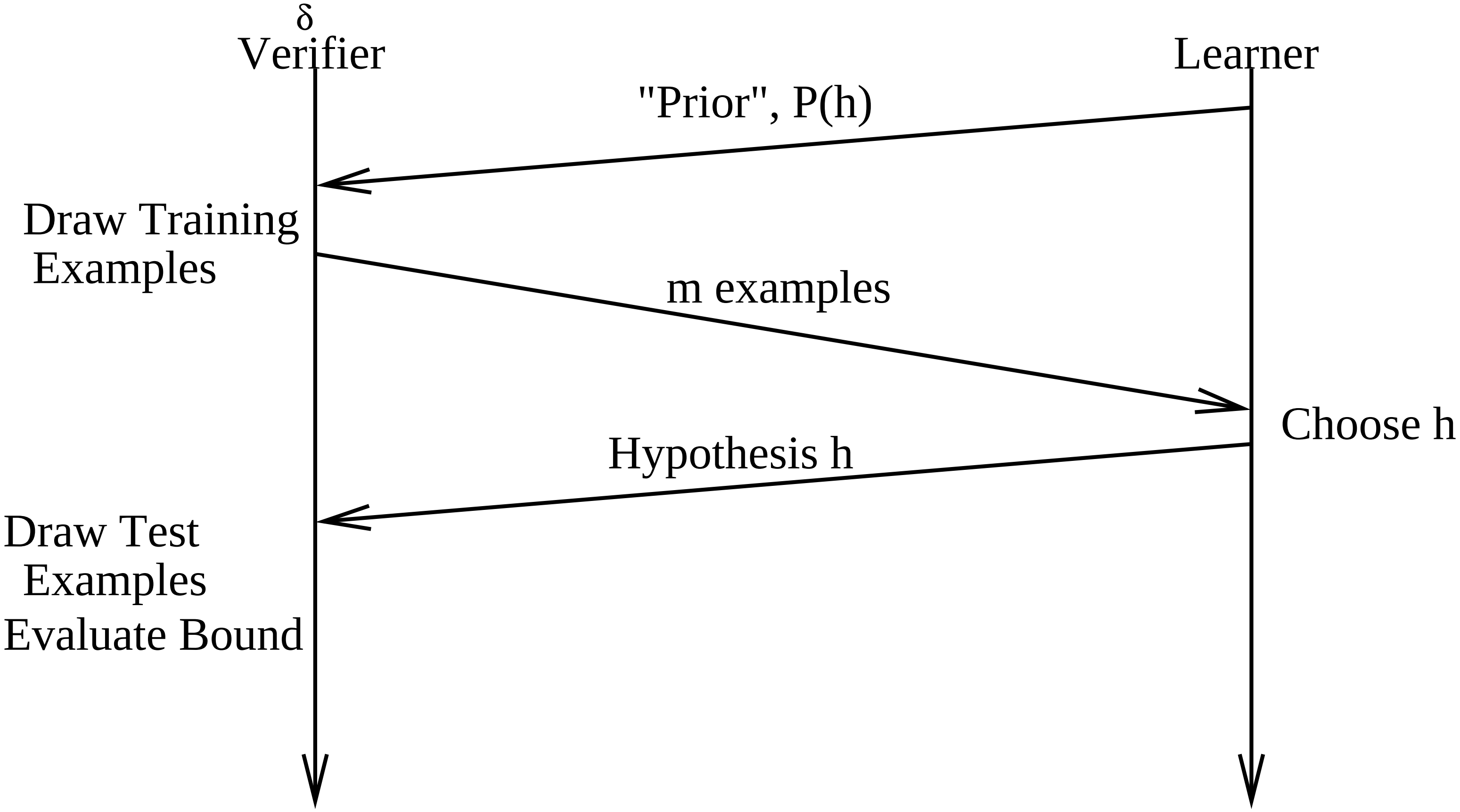
$Y \in \{0, 1\}$ = prediction space

Assume: Examples $(X, Y)$ are drawn independently from distribution $D$

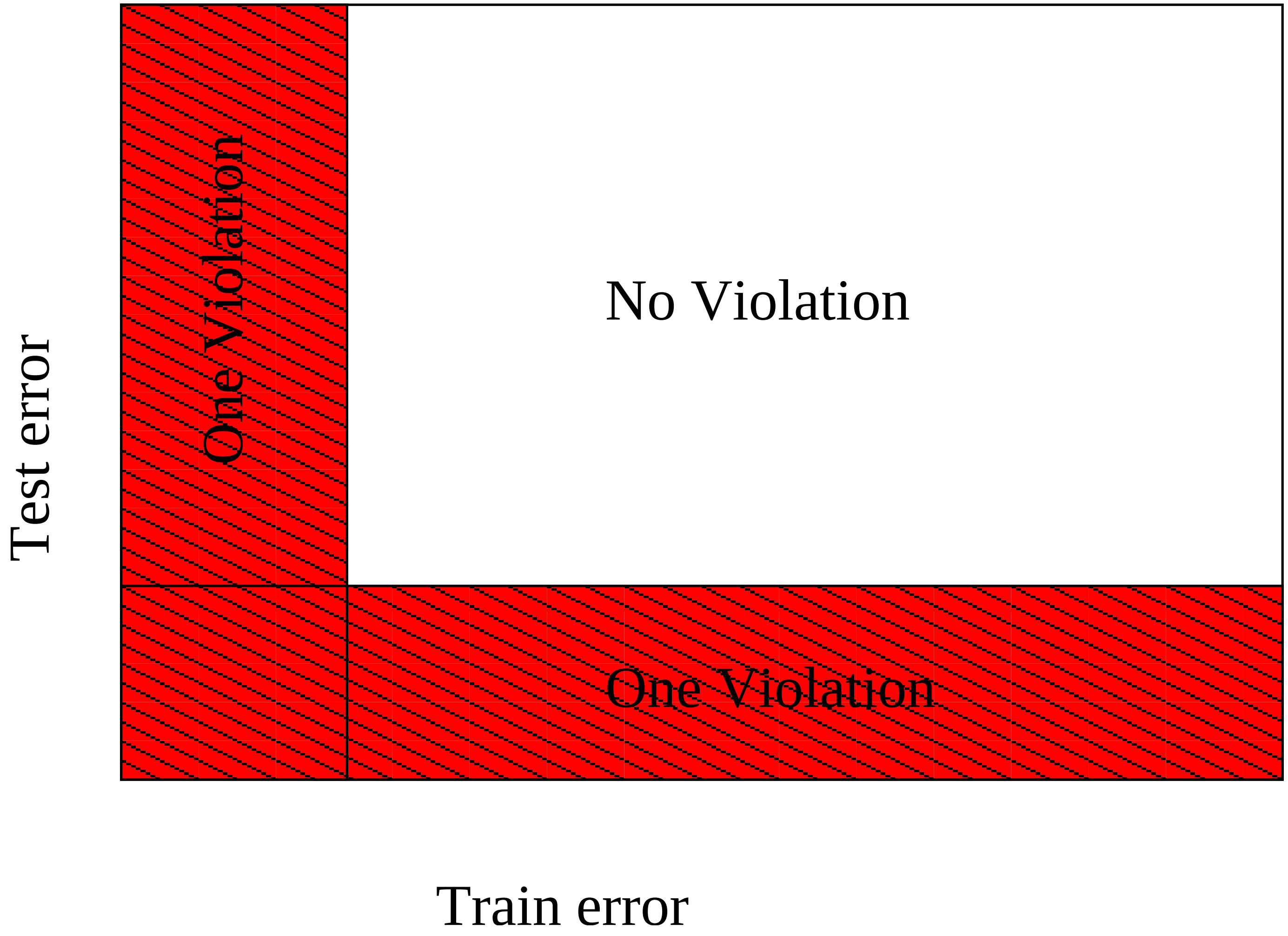
# Terminology

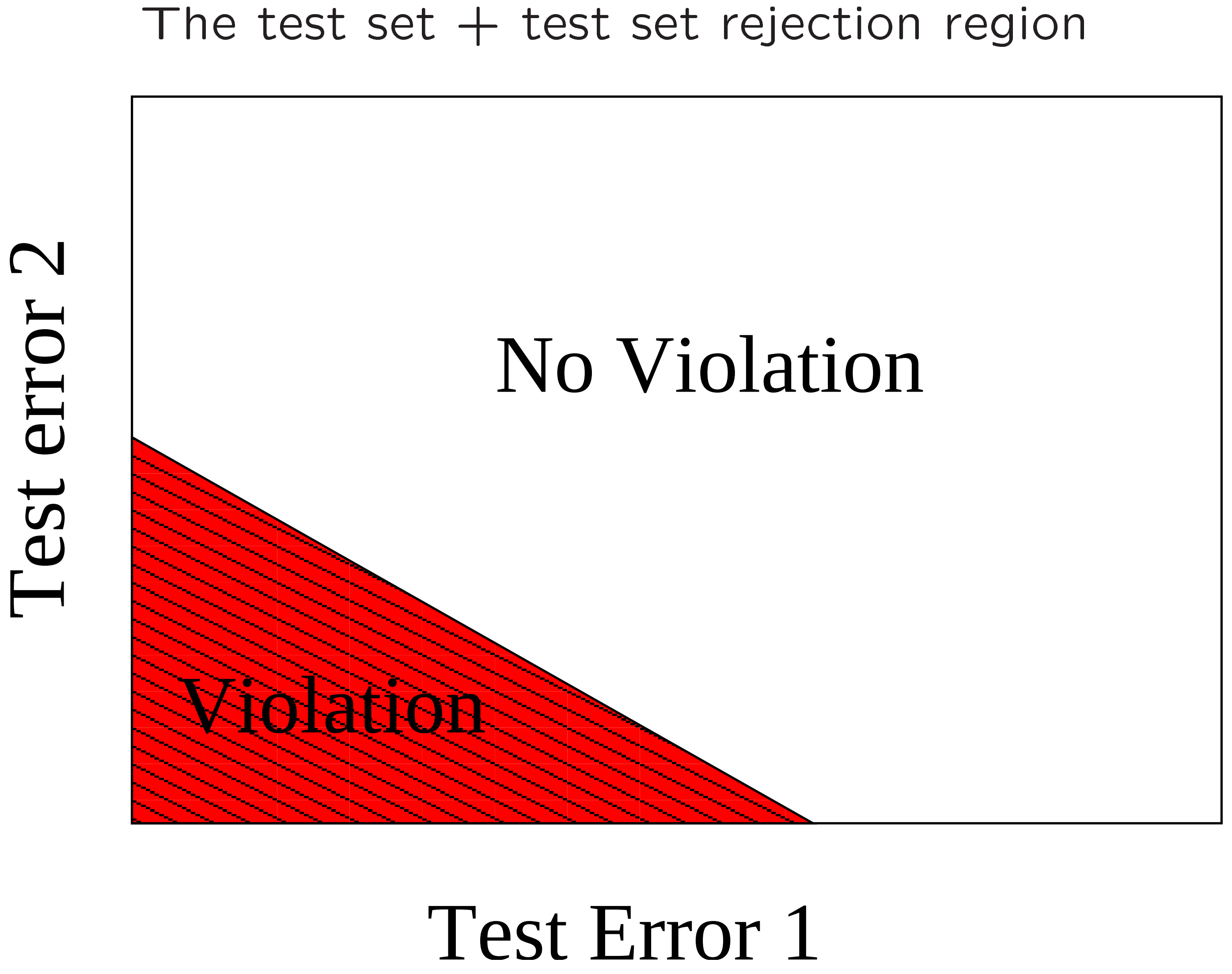
- Samples $= S$
- Learning algorithm $A: \; S \rightarrow h$
- True Error rate $e_D(h) = \mathrm{Pr}_D(h(x) \neq y)$
- Empirical Error rate $\hat{e}_S(h) = \mathrm{Pr}_S(h(x) \neq y)$

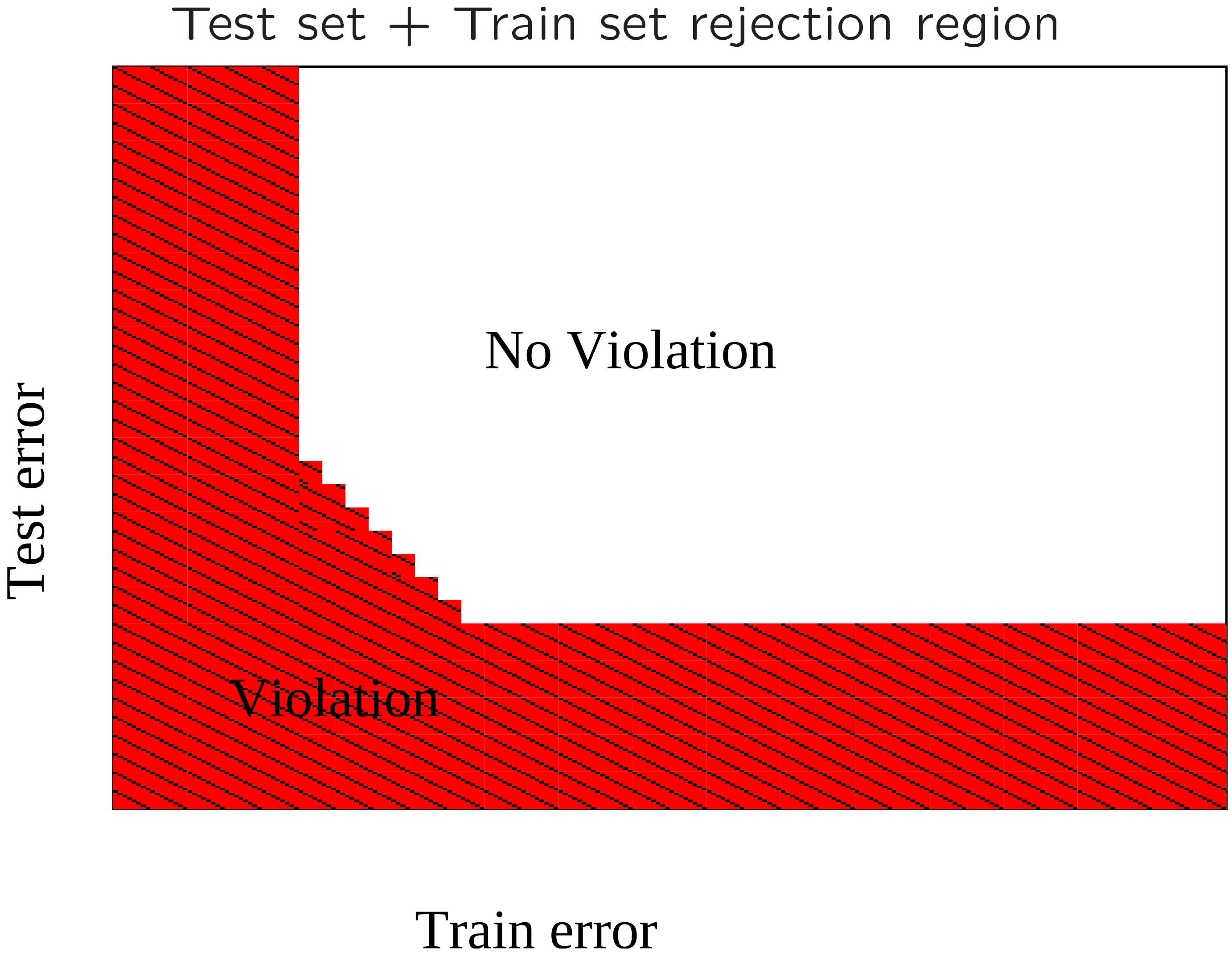
# Train and Test Bound



The "or" rejection region




The test set + test set rejection region
No Violation
Violation
Test error 2
Test Error 1



Test set + Train set rejection region
No Violation
Violation
Test error
Train error

1. Test & Train set bounds
2. Combinations
3. A combination theorem
4. Technique and Method

Bound = rejection set

Let $\phi_P(\delta)$ = rejection set dependent on $P$ and $\delta$.

$$\Pr_{x \sim P}(x \in \phi_P(\delta)) \leq \delta$$

Randomized bound = Randomized rejection set

$$E_{w \sim Q} \Pr_{x \sim P}(x \in \phi_{P,w}(\delta)) \leq \delta$$

$$\Rightarrow \Pr_{w \sim Q,\ x \sim P}(x \in \phi_{P,w}(\delta)) \leq \delta$$

Randomized bound = Randomized rejection set

$$E_{w \sim Q} \Pr_{x \sim P}(x \in \phi_{P,w}(\delta)) \leq \delta$$

$$\Rightarrow \Pr_{w \sim Q,\ x \sim P}(x \in \phi_{P,w}(\delta)) \leq \delta$$

$Q$ = test set

$P$ = train set

Problem: Approximate test set bound $\Rightarrow$ expectation not known

Solution:

1. Restrain $\phi_{P,w}(\delta)$ to be monotonic decreasing in $\hat{e}_{\mathsf{test}}$

2. if $\forall x \; F_1(x) \leq F_2(x)$ and $f(x) \geq 0$ then

$$\int f(x) dF_1(x) \leq \int f(x) dF_2(x)$$

1. Test & Train set bounds

2. Combinations

3. A combination theorem

4. Technique and Method

# The Combined Bound

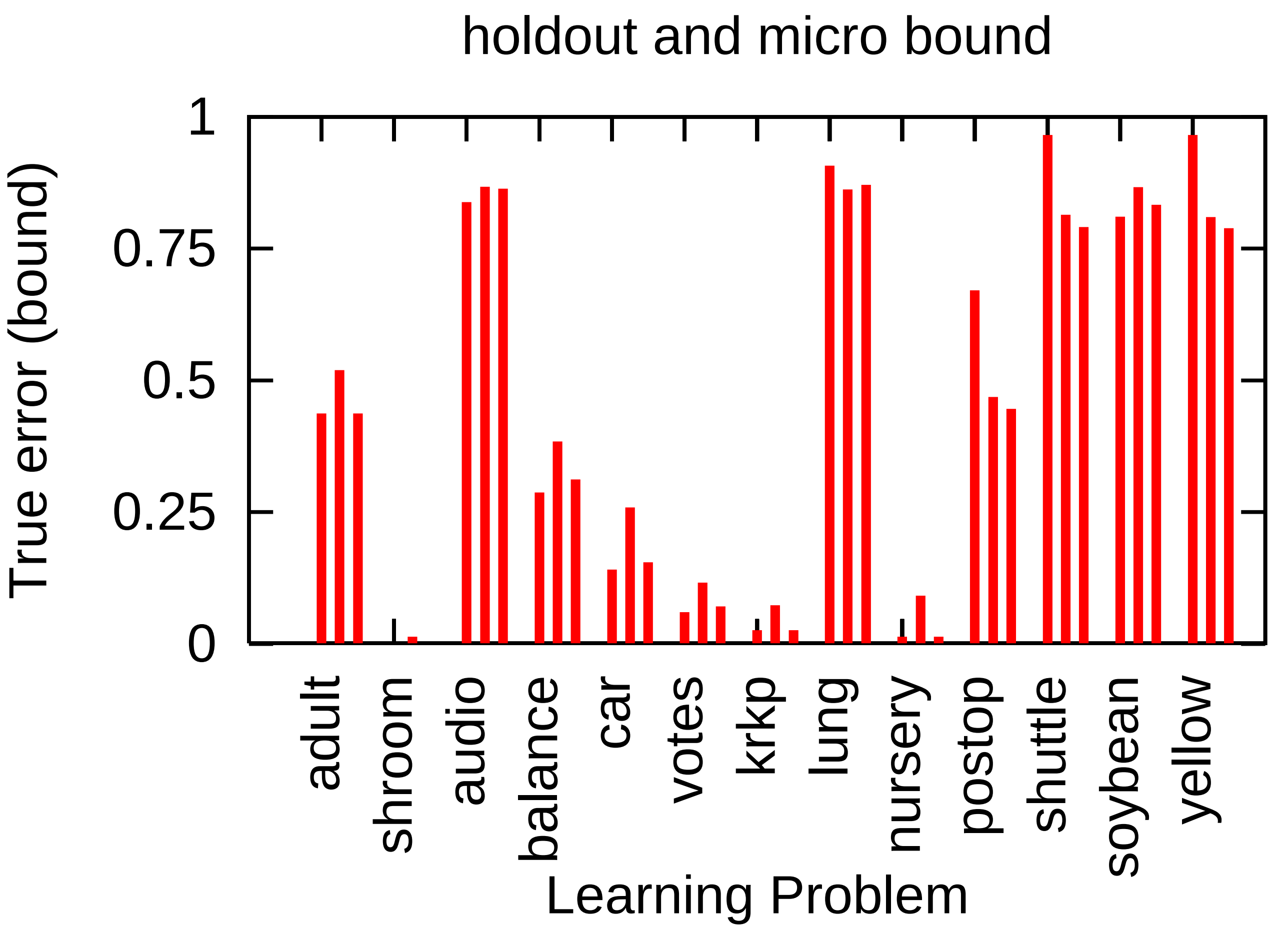
1. Fix bound.

2. Bound implies CDF, $F^{-1}_{\text{train}}(\delta)$, $F^{-1}_{\text{test}}(\delta)$.

3. Let rejection region $= F_{\text{train}}(\hat{e}_{\text{train}}(h))F_{\text{test}}(\hat{e}_{\text{test}}(h)) \leq t(\delta)$ where $t(\delta)$ satisfies:

$$\Pr_{x_{\text{test}}, x_{\text{train}}}(F_{\text{test}}(x_{\text{test}})F_{\text{train}}(x_{\text{train}}) \leq t(\delta)) \leq \delta$$

4. Binary search on true error rate $\Rightarrow$ worst consistent true error rate $=$ bound




left = test, middle = train, right = train and test

Demo

Conclusion: This approach works, and (empirically) never harms much while sometimes helping a lot.

1. Allows use of train set bounds "for free" in applied research.
2. Makes COLT-style bounds applicable even when not completely tight.

Open:

1. Can we speed up the combined bound calculation?
2. Is this the right combination?