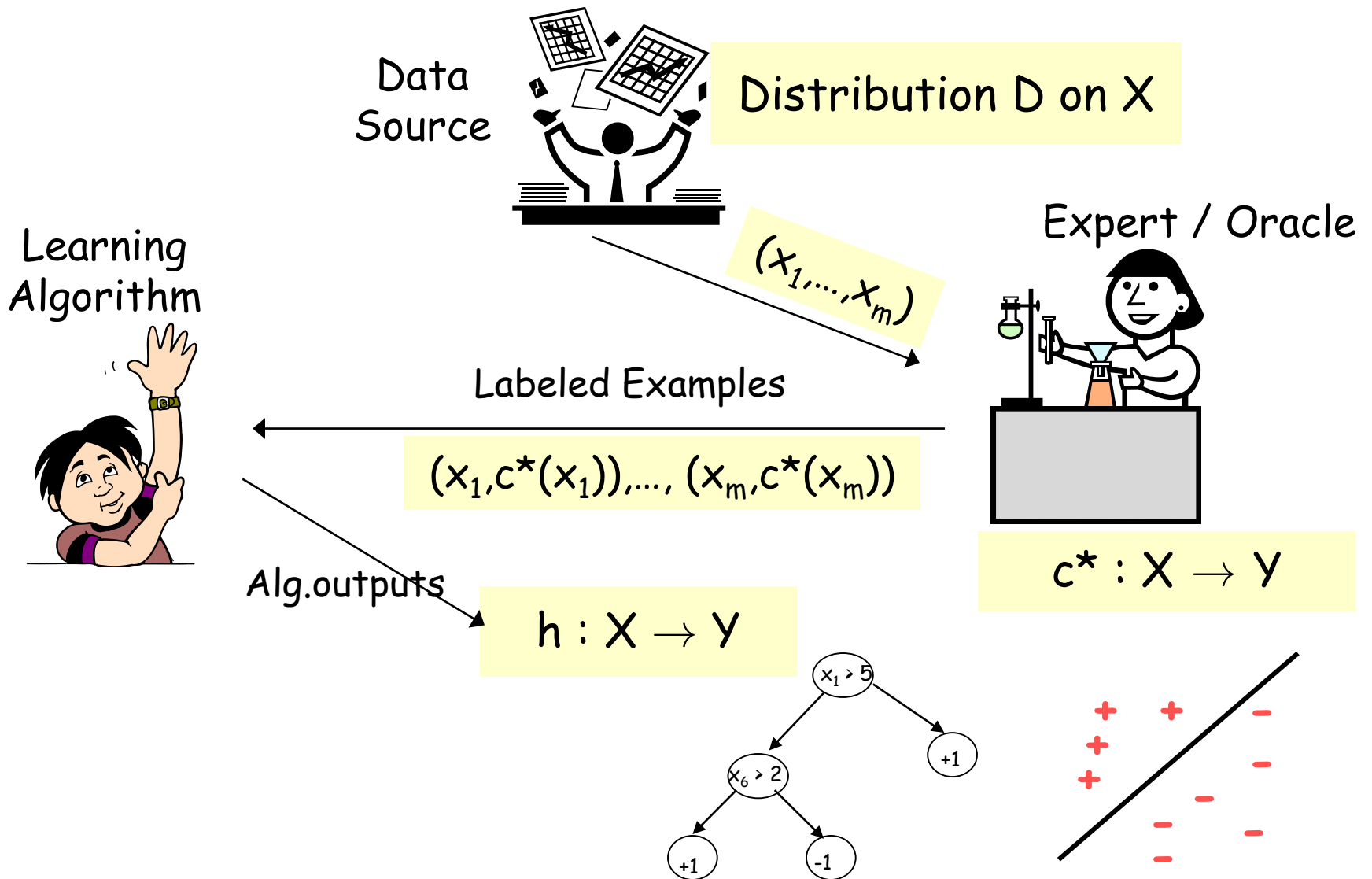


# Model Selection

Maria-Florina (Nina) Balcan

03/22/2019

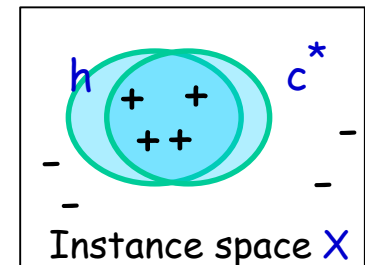
# PAC/SLT models for Supervised Learning



# PAC/SLT models for Supervised Learning

- $X$  - feature/instance space; distribution  $D$  over  $X$   
e.g.,  $X = \mathbb{R}^d$  or  $X = \{0,1\}^d$
- Algo sees training sample  $S: (x_1, c^*(x_1)), \dots, (x_m, c^*(x_m))$ ,  $x_i$  i.i.d. from  $D$ 
  - labeled examples - drawn i.i.d. from  $D$  and labeled by target  $c^*$
  - labels  $\in \{-1,1\}$  - binary classification
- Algo does optimization over  $S$ , find hypothesis  $h$ .
- Goal:  $h$  has small error over  $D$ .

$$err_D(h) = \Pr_{x \sim D}(h(x) \neq c^*(x))$$



**Bias:** fix hypothesis space  $H$  [whose complexity is not too large]

- Realizable:  $c^* \in H$ .
- Agnostic:  $c^*$  "close to"  $H$ .

# Sample Complexity: Finite Hypothesis Spaces

## Realizable Case

### Theorem

$$m \geq \frac{1}{\varepsilon} \left[ \ln(|H|) + \ln\left(\frac{1}{\delta}\right) \right]$$

labeled examples are sufficient so that with prob.  $1 - \delta$ , all  $h \in H$  with  $err_D(h) \geq \varepsilon$  have  $err_S(h) > 0$ .

So, if  $c^* \in H$  and can find consistent fns, then only need this many examples to get generalization error  $\leq \varepsilon$  with prob.  $\geq 1 - \delta$

## Agnostic Case

What if there is no perfect  $h$ ?

**Theorem** After  $m$  examples, with probab.  $\geq 1 - \delta$ , all  $h \in H$  have  $|err_D(h) - err_S(h)| < \varepsilon$ , for

$$m \geq \frac{1}{2\varepsilon^2} \left[ \ln(|H|) + \ln\left(\frac{2}{\delta}\right) \right]$$

# Sample Complexity: Infinite Hypothesis Spaces

## Realizable Case

### Theorem

$$m = O\left(\frac{1}{\varepsilon} \left[ VCdim(H) \log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right) \right]\right)$$

labeled examples are sufficient so that with probab.  $1 - \delta$ , all  $h \in H$  with  $err_D(h) \geq \varepsilon$  have  $err_S(h) > 0$ .

# Sample Complexity: Infinite Hypothesis Spaces

## Theorem (agnostic case)

$$m = O\left(\frac{1}{\epsilon^2} \left(VCdim(H) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

labeled examples are sufficient s.t. with probability at least  $1 - \delta$   
for all  $h$  in  $H$   $|err_D(h) - err_S(h)| \leq \epsilon$

## Statistical Learning Theory Style

With prob at least  $1 - \delta$  for all  $h$  in  $H$

$$err_D(h) \leq err_S(h) + \sqrt{\frac{1}{2m} \left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right)}.$$

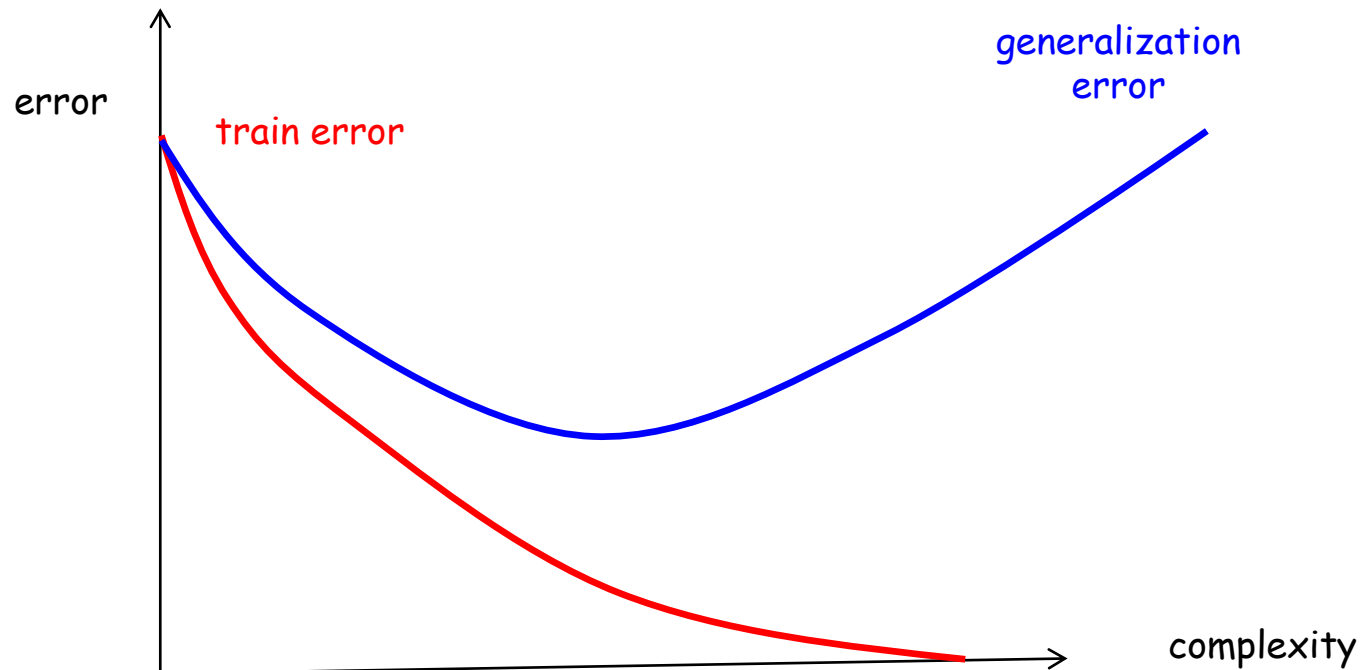
Can we use our bounds for  
model selection?



# True Error, Training Error, Overfitting

**Model selection:** trade-off between decreasing training error and keeping  $H$  simple.

$$\text{err}_D(h) \leq \text{err}_S(h) + \sqrt{\frac{VCdim(H)}{m}} + \dots$$

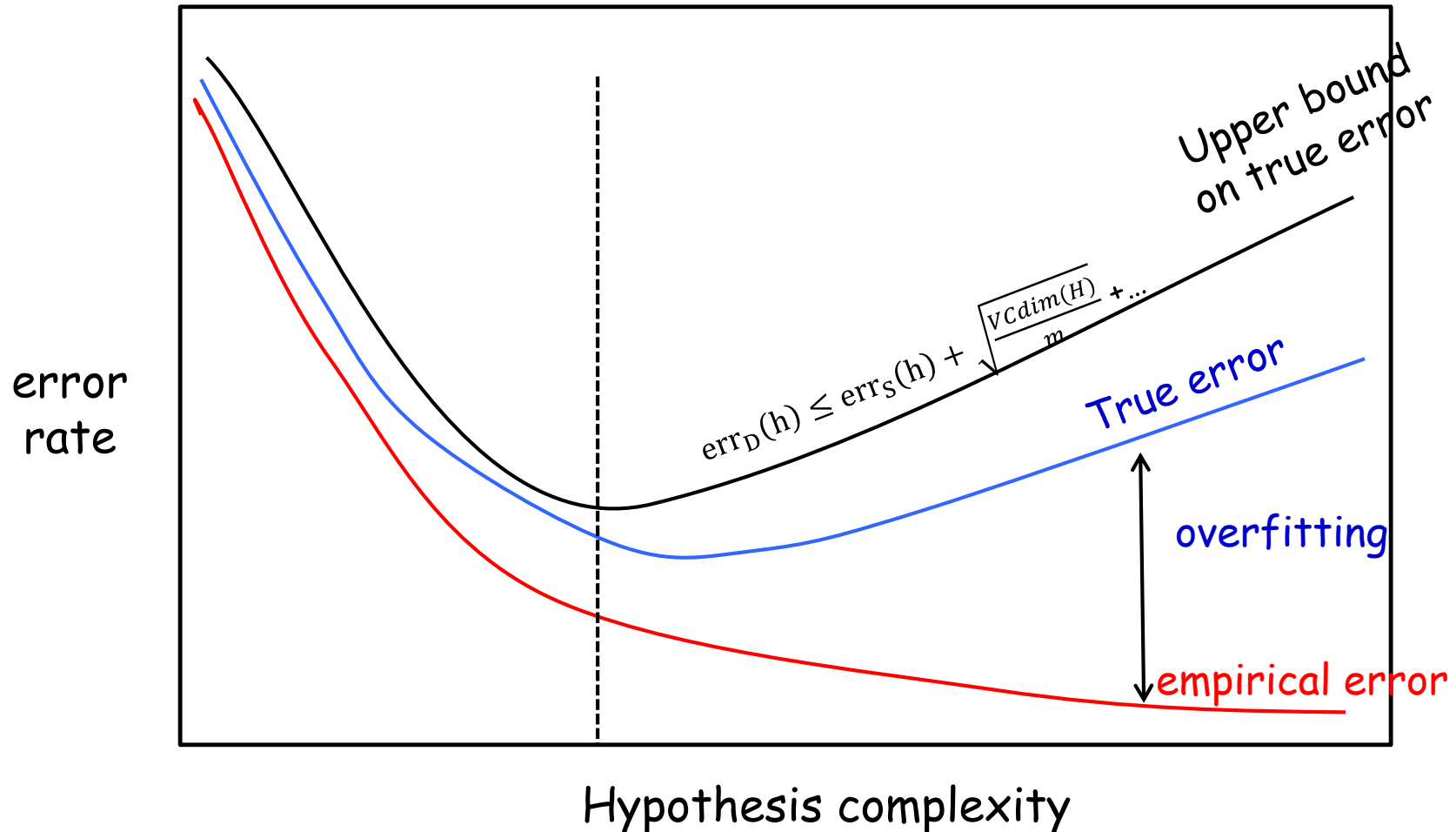




# Structural Risk Minimization (SRM)

$$H_1 \subseteq H_2 \subseteq H_3 \subseteq \dots \subseteq H_i \subseteq \dots$$

(E.g.,  $H_i$  = decision trees of depth  $i$ )

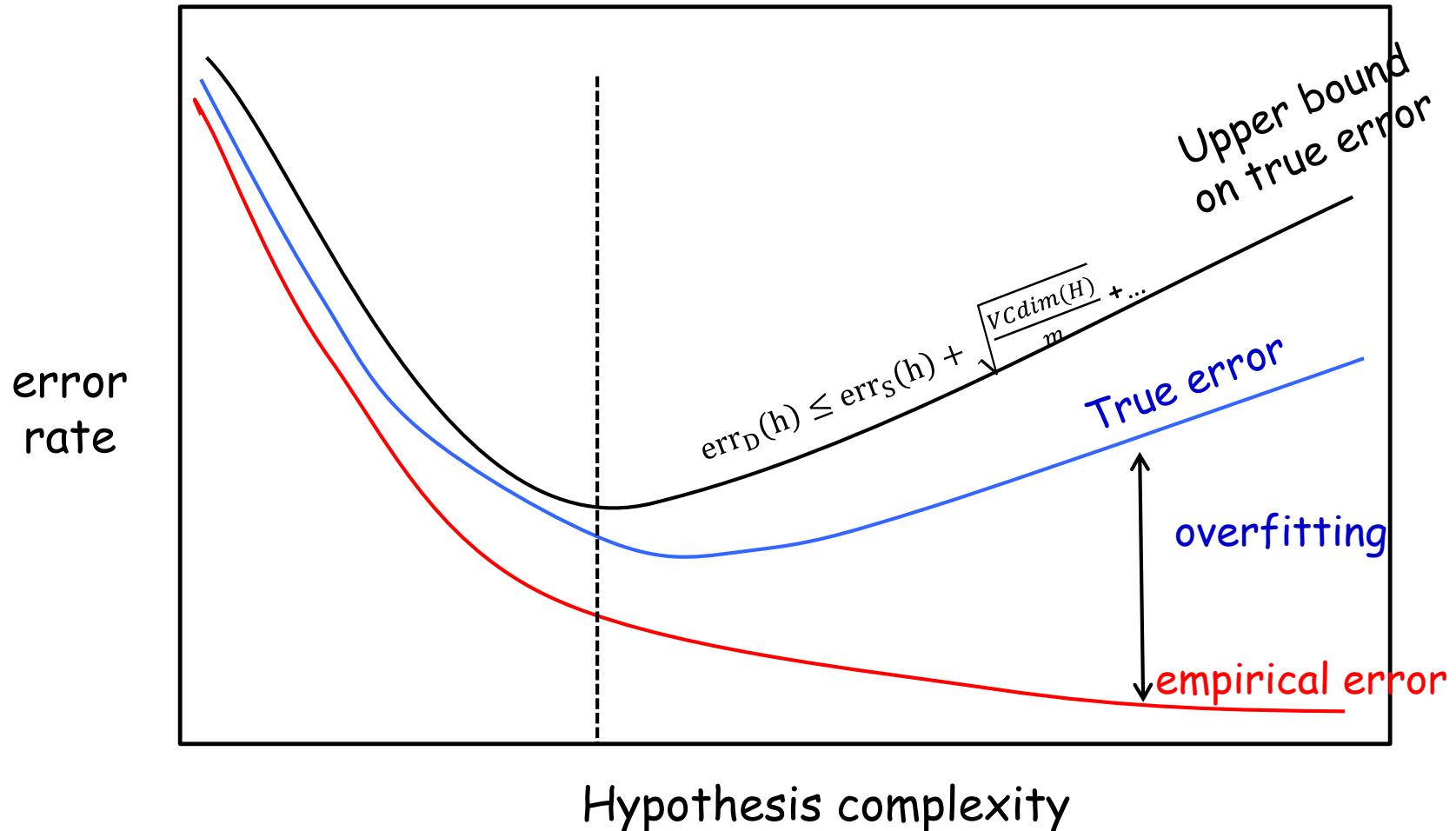


# What happens if we increase $m$ ?

Black curve will stay close to the red curve for longer, everything shift to the right...

# Structural Risk Minimization (SRM)

$$H_1 \subseteq H_2 \subseteq H_3 \subseteq \dots \subseteq H_i \subseteq \dots$$



# Structural Risk Minimization (SRM)

- $H_1 \subseteq H_2 \subseteq H_3 \subseteq \dots \subseteq H_i \subseteq \dots$

- $\hat{h}_k = \operatorname{argmin}_{h \in H_k} \{\operatorname{err}_S(h)\}$

As  $k$  increases,  $\operatorname{err}_S(\hat{h}_k)$  goes down but complex. term goes up.

- $\hat{k} = \operatorname{argmin}_{k \geq 1} \{\operatorname{err}_S(\hat{h}_k) + \operatorname{complexity}(H_k)\}$

Output  $\hat{h} = \hat{h}_{\hat{k}}$

**Claim:** W.h.p.,  $\operatorname{err}_D(\hat{h}) \leq \min_{k^*} \min_{h^* \in H_{k^*}} [\operatorname{err}_D(h^*) + 2\operatorname{complexity}(H_{k^*})]$

# Techniques to Handle Overfitting

- **Structural Risk Minimization (SRM).**  $H_1 \subseteq H_2 \subseteq \dots \subseteq H_i \subseteq \dots$   
Minimize gener. bound:  $\hat{h} = \operatorname{argmin}_{k \geq 1} \{ \operatorname{err}_S(\hat{h}_k) + \operatorname{complexity}(H_k) \}$ 
  - Often computationally hard....
  - Nice case where it is possible: M. Kearns, Y. Mansour, ICML'98, "A Fast, Bottom-Up Decision Tree Pruning Algorithm with Near-Optimal Generalization"
- **Regularization:** general family closely related to SRM
  - E.g., SVM, regularized logistic regression, etc.
  - minimizes expressions of the form:  $\operatorname{err}_S(h) + \lambda ||h||^2$
- **Cross Validation:**
  - Hold out part of the training data and use it as a proxy for the generalization error

# What you should know

- The importance of sample complexity in Machine Learning.
- Understand meaning of PAC bounds (what PAC stands for, meaning of parameters  $\epsilon$  and  $\delta$ ).
- Shattering, VC dimension as measure of complexity, form of the VC bounds.
- Model Selection, Structural Risk Minimization.