# Prediction and Overfitting

CS4780/5780 – Machine Learning
Fall 2013

Thorsten Joachims
Cornell University

Reading: Mitchell Sections 3.6 – 3.7

# Example: Text Classification

- Task: Learn rule that classifies Reuters Business News
  - Class +: “Corporate Acquisitions”
  - Class -: Other articles
  - 2000 training instances
- Representation:
  - Boolean attributes, indicating presence of a keyword in article
  - 9947 such keywords (more accurately, word “stems”)

LAROCHE STARTS BID FOR NECO SHARES +

Investor David F. La Roche of North Kingstown, R.I., said he is offering to purchase 170,000 common shares of NECO Enterprises Inc at 26 dlrs each. He said the successful completion of the offer, plus shares he already owns, would give him 50.5 pct of NECO's 962,016 common shares. La Roche said he may buy more, and possible all NECO shares. He said the offer and withdrawal rights will expire at 1630 EST/2130 gmt, March 30, 1987.

SALANT CORP 1ST QTR FEB 28 NET -

Oper shr profit seven cts vs loss 12 cts. Oper net profit 216,000 vs loss 401,000. Sales 21.4 mln vs 24.9 mln. NOTE: Current year net excludes 142,000 dlr tax credit. Company operating in Chapter 11 bankruptcy.

# Decision Tree for “Corporate Acq.”

```
vs = 1: -
vs = 0:
|   export = 1:
…
|   export = 0:
|   |   rate = 1:
|   |   |   stake = 1: +
|   |   |   stake = 0:
|   |   |   |   debenture = 1: +
|   |   |   |   debenture = 0:
|   |   |   |   |   takeover = 1: +
|   |   |   |   |   takeover = 0:
|   |   |   |   |   |   file = 0: -
|   |   |   |   |   |   file = 1:
|   |   |   |   |   |   |   share = 1: +
|   |   |   |   |   |   |   share = 0: -
… and many more
```

**Learned tree of ID3:**

- has 299 nodes
- is consistent

**Accuracy of learned tree:**

- 11% error rate

Note: word stems expanded for improved readability.

# Learning as Prediction

**Definition:** *A particular* **instance of a learning problem** *is described by a probability distribution* $P(X, Y)$.

**Definition:** *A* **sample** $S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n))$ *is independently identically distributed (i.i.d.) according to* $P(X, Y)$ *if*

$$P(S = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n))) = \prod_{i=1}^{n} P(X = \vec{x}_i, Y = y_i)$$

# Sample Error and Generalization Error

**Definition:** *The* **error on sample** $S$ $Err_S(h)$ *of a hypothesis* $h$ *is* $Err_S(h) = \frac{1}{n}\sum_{i=1}^{n} \Delta(h(\vec{x}_i), y_i)$.

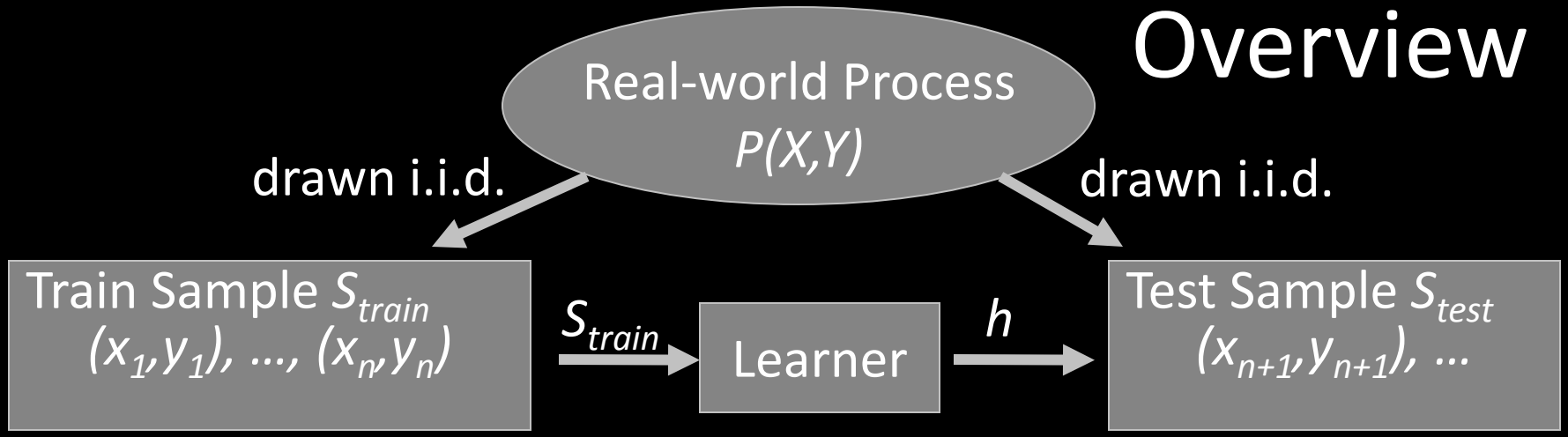
**Definition:** $\Delta(a, b)$ *is the 0/1-loss function*

$$\Delta(a,b) = \begin{cases} 0 & if(a == b) \\ 1 & else \end{cases}$$

**Definition:** *The* **prediction/generalization/true error** $Err_P(h)$ *of a hypothesis* $h$ *for a learning task* $P(X, Y)$ *is*

$$Err_P(h) = \sum_{\vec{x} \in X, y \in Y} \Delta(h(\vec{x}), y) P(X = \vec{x}, Y = y).$$

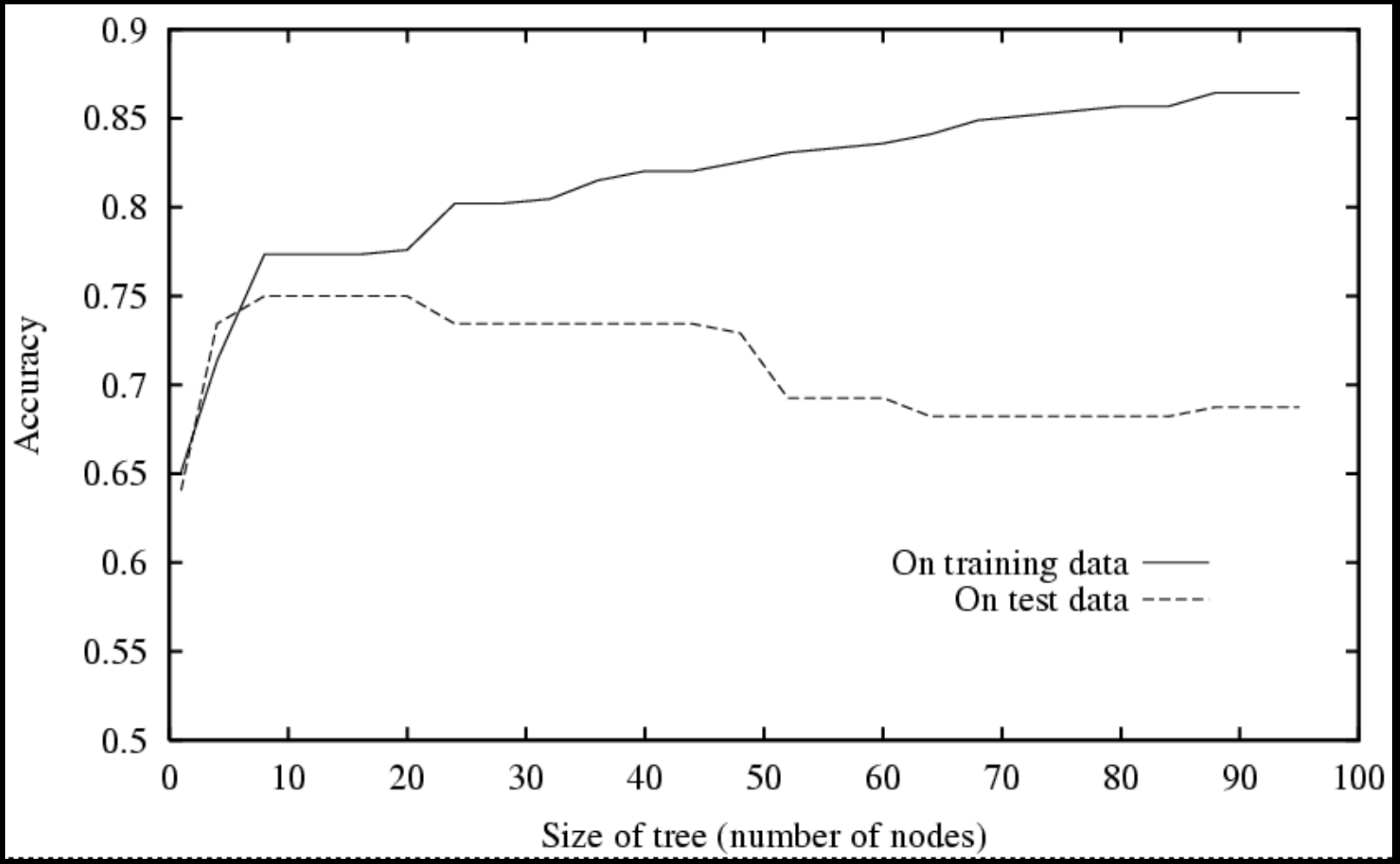
# Learning as Prediction Overview




- Goal: Find $h$ with small prediction error $Err_P(h)$ over $P(X,Y)$.
- Strategy: Find (any?) $h$ with small error $Err_{S_{train}}(h)$ on training sample $S_{train}$.

- Training Error: Error $Err_{S_{train}}(h)$ on training sample.
- Test Error: Error $Err_{S_{test}}(h)$ on test sample is an estimate of $Err_P(h)$ .

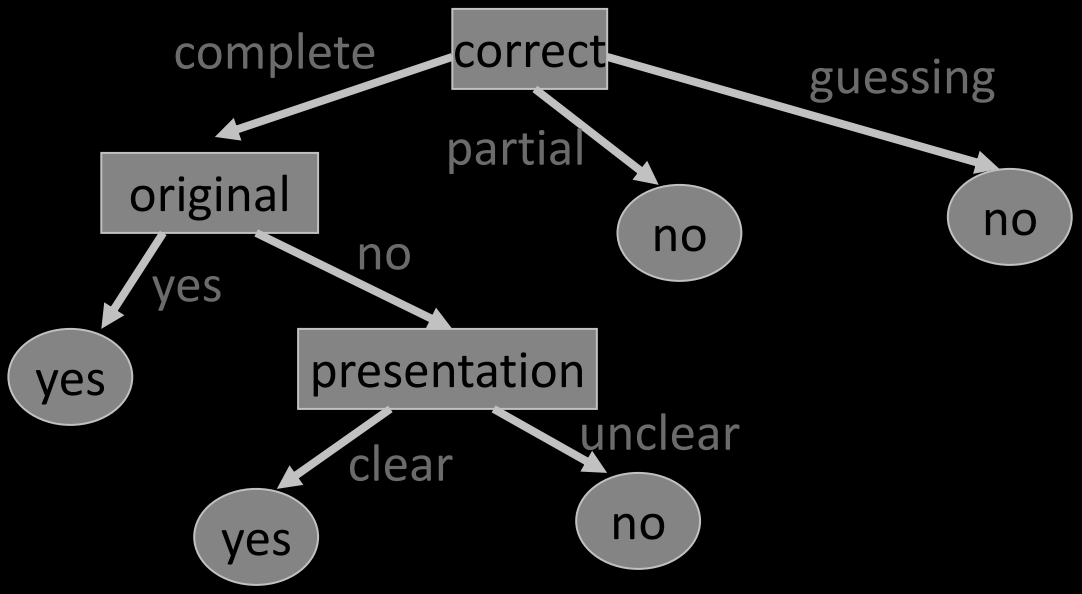
# Overfitting




- Note: Accuracy = 1.0-Error

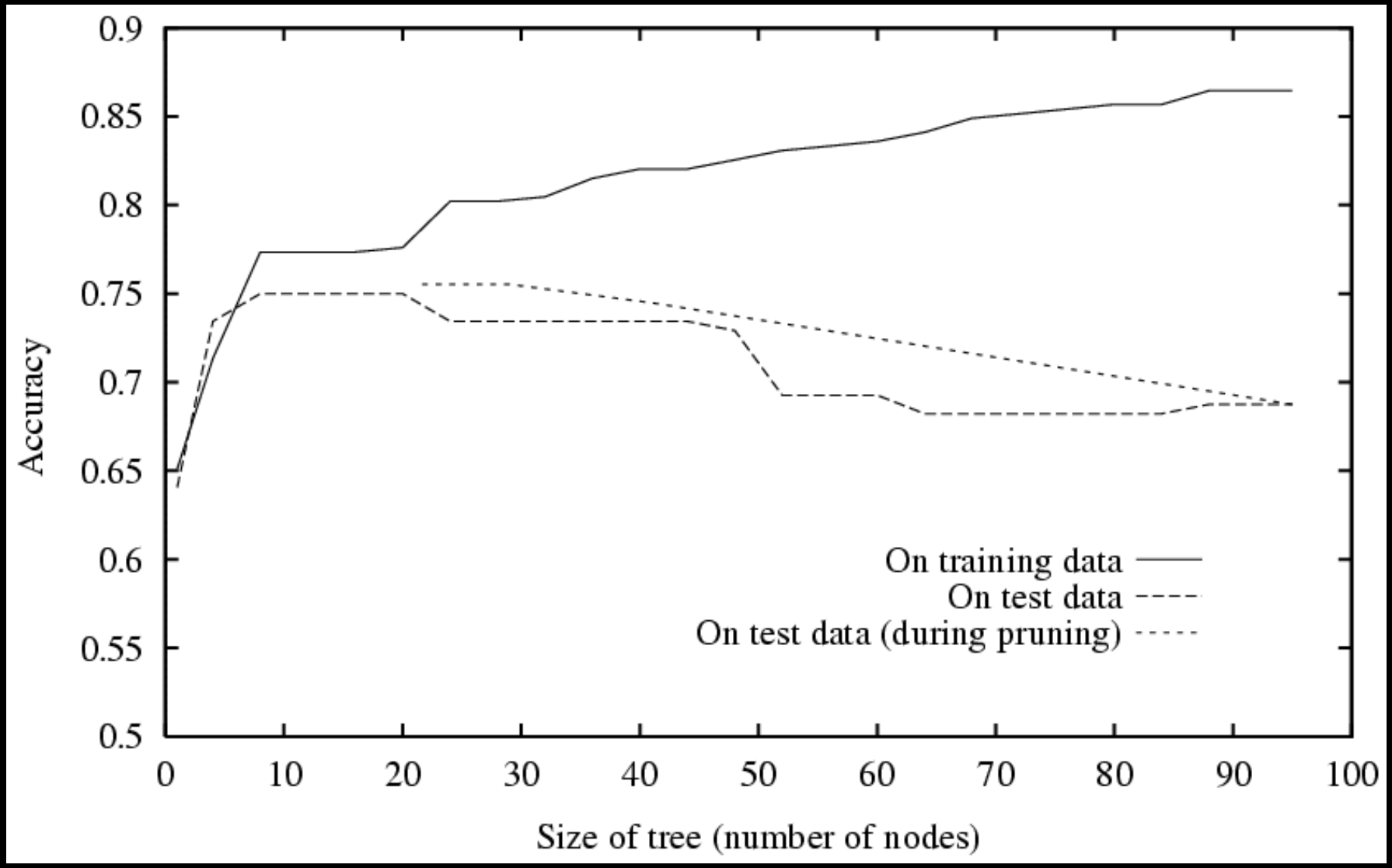
[Mitchell]

# Decision Tree Example: revisited




| | C O P | $A^+$ |
|---|---|---|
| $\vec{x}_1$ | ( c, y, c ) | $y_1 = +1$ |
| $\vec{x}_2$ | ( c, n, u ) | $y_2 = -1$ |
| $\vec{x}_3$ | ( c, y, u ) | $y_3 = +1$ |
| $\vec{x}_4$ | ( c, n, c ) | $y_4 = +1$ |
| $\vec{x}_5$ | ( p, y, c ) | $y_5 = -1$ |
| $\vec{x}_6$ | ( g, y, c ) | $y_6 = -1$ |
| $\vec{x}_7$ | ( c, y, c ) | $y_7 = +1$ |
| $\vec{x}_8$ | ( c, y, u ) | $y_8 = +1$ |
| $\vec{x}_9$ | ( p, y, c ) | $y_9 = -1$ |
| $\vec{x}_{10}$ | ( c, y, c ) | $y_{10} = +1$ |

# Reduced-Error Pruning



# Text Classification Example Results

- Unpruned Tree (ID3 Algorithm):
  - Size: 437 nodes Training Error: 0.0% Test Error: 11.0%
- Early Stopping Tree (ID3 Algorithm):
  - Size: 299 nodes Training Error: 2.6% Test Error: 9.8%
- Reduced-Error Pruning (C4.5 Algorithm):
  - Size: 167 nodes Training Error: 4.0% Test Error: 10.8%
- Rule Post-Pruning (C4.5 Algorithm):
  - Size: 164 tests Training Error: 3.1% Test Error: 10.3%
  - Examples of rules
    - IF vs = 1 THEN - [99.4%]
    - IF vs = 0 & export = 0 & takeover = 1 THEN + [93.6%]