

An aerial photograph of a city waterfront at sunset. The sun is low on the horizon, casting a warm, golden glow over the scene. The water is dark blue with many small sailboats scattered across it. The city buildings are visible on the left side, with some trees in the foreground. The overall atmosphere is peaceful and scenic.

# Evaluating Machine Learning Methods: Part 2

CS 760@UW-Madison



# Goals for the last lecture



you should understand the following concepts

- bias of an estimator
- learning curves
- stratified sampling
- cross validation
- confusion matrices
- TP, FP, TN, FN
- ROC curves

# Goals for the lecture



you should understand the following concepts

- PR curves
- confidence intervals for error
- pairwise  $t$ -tests for comparing learning systems
- scatter plots for comparing learning systems
- lesion studies

# Recall: ROC



		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{true positive rate (recall)} = \frac{\text{TP}}{\text{actual pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{false positive rate} = \frac{\text{FP}}{\text{actual neg}} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$



# ROC curves

Does a low false-positive rate indicate that most positive predictions (i.e. predictions with confidence  $>$  some threshold) are correct?

suppose our TPR is 0.9, and FPR is 0.01

fraction of instances that are positive	fraction of positive predictions that are correct
0.5	0.989
0.1	0.909
0.01	0.476
0.001	0.083



# Other accuracy metrics

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

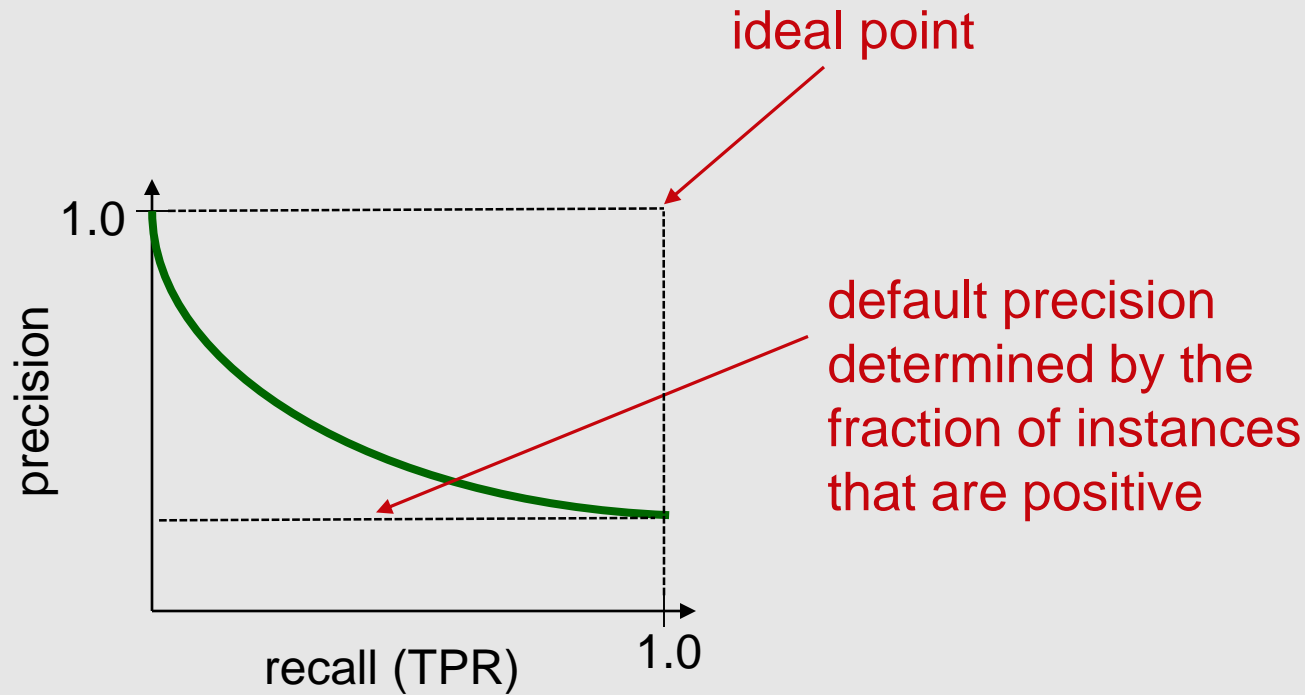
$$\text{recall (TP rate)} = \frac{\text{TP}}{\text{actual pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{precision (positive predictive value)} = \frac{\text{TP}}{\text{predicted pos}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

# Precision/recall curves



A *precision/recall curve* plots the precision vs. recall (TP-rate) as a threshold on the confidence of an instance being positive is varied



# Precision/recall curve example



predicting patient risk for VTE

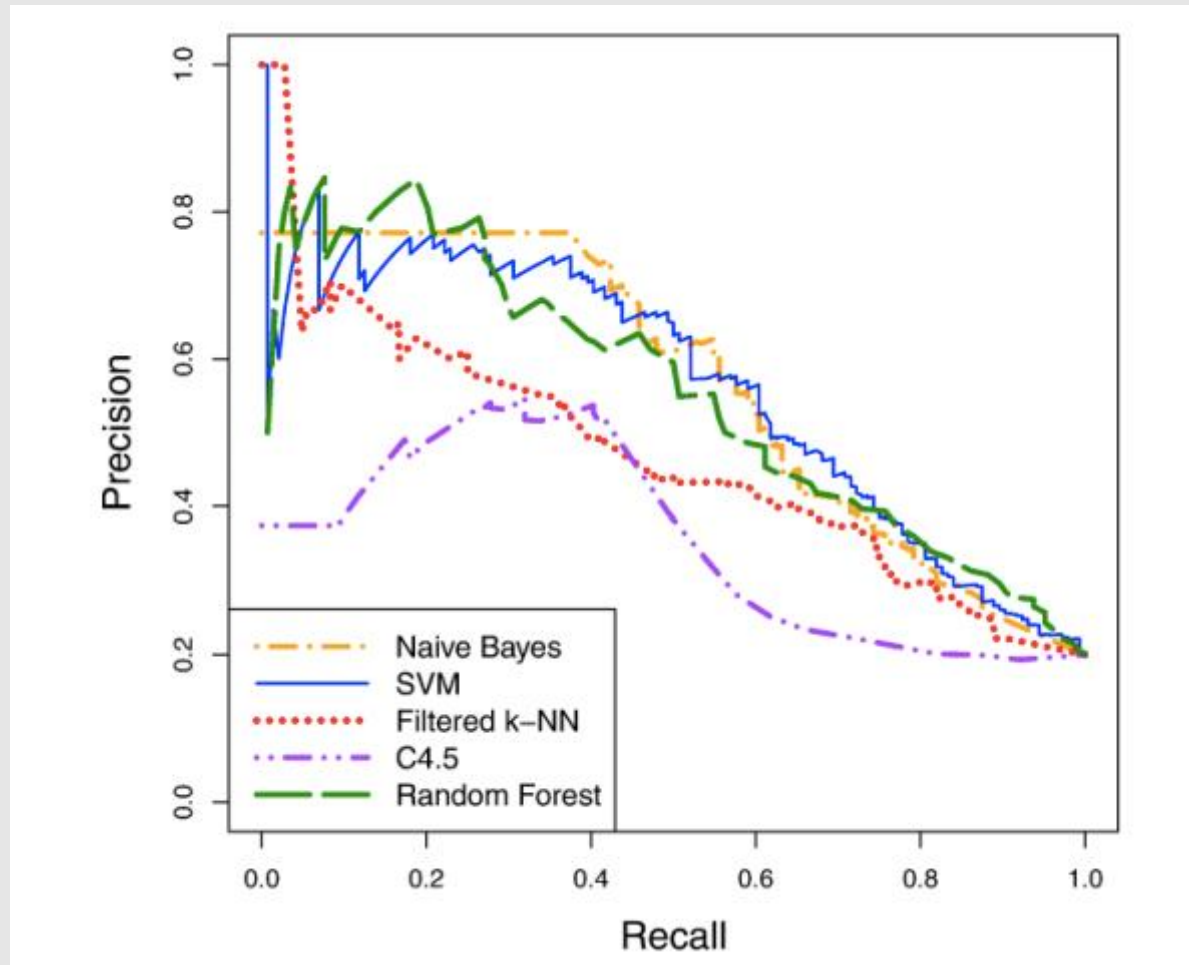


figure from Kawaler et al., *Proc. of AMIA Annual Symposium*, 2012





# How do we get one ROC/PR curve when we do cross validation?

## Approach 1

- make assumption that confidence values are comparable across folds
- pool predictions from all test sets
- plot the curve from the pooled predictions

## Approach 2 (for ROC curves)

- plot individual curves for all test sets
- view each curve as a function
- plot the average curve for this set of functions

# Comments on ROC and PR curves



## both

- allow predictive performance to be assessed at various levels of confidence
- assume binary classification tasks
- sometimes summarized by calculating *area under the curve*

## ROC curves

- insensitive to changes in class distribution (ROC curve does not change if the proportion of positive and negative instances in the test set are varied)
- can identify optimal classification thresholds for tasks with differential misclassification costs

## precision/recall curves

- show the fraction of predictions that are false positives
- well suited for tasks with lots of negative instances



# Confidence intervals on error

Given the observed error (accuracy) of a model over a limited sample of data, how well does this error characterize its accuracy over additional instances?

Suppose we have

- a learned model  $h$
- a test set  $S$  containing  $n$  instances drawn independently of one another and independent of  $h$
- $n \geq 30$
- $h$  makes  $r$  errors over the  $n$  instances

our best estimate of the error of  $h$  is

$$\text{error}_S(h) = \frac{r}{n}$$



# Confidence intervals on error

With approximately  $C\%$  probability, the true error lies in the interval

$$error_S(h) \pm z_C \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

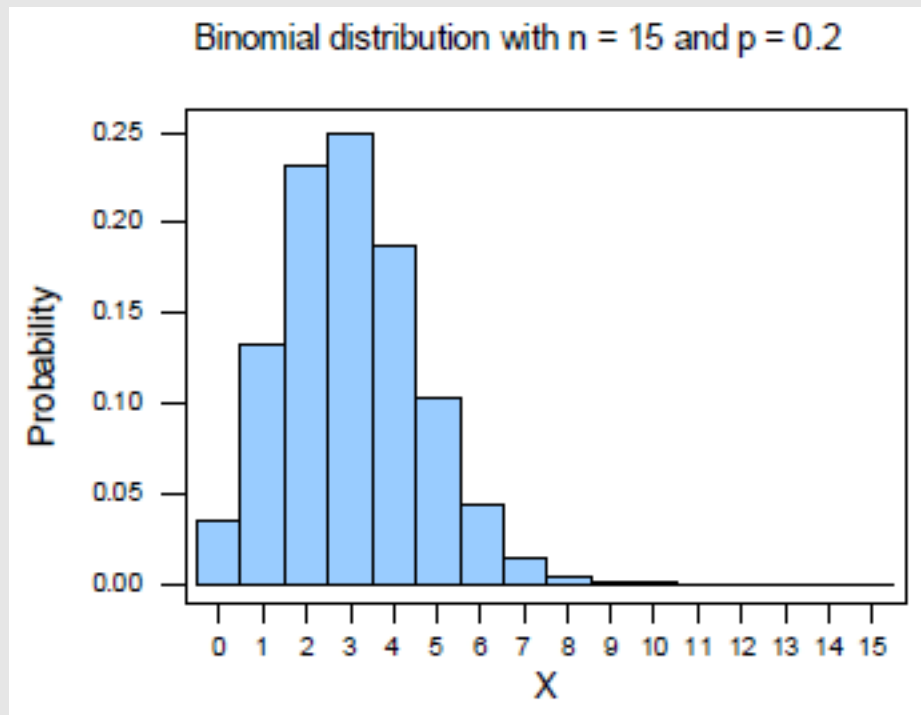
where  $z_C$  is a constant that depends on  $C$  (e.g. for 95% confidence,  $z_C = 1.96$ )



# Confidence intervals on error

How did we get this?

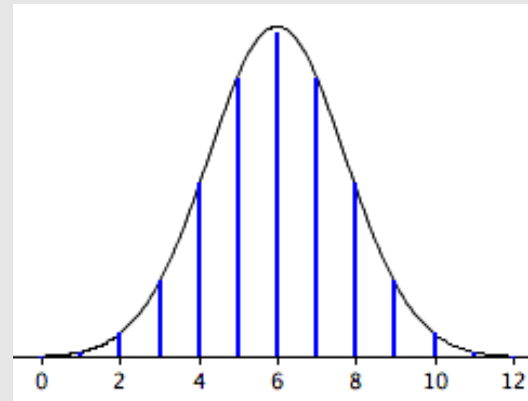
1. Our estimate of the error follows a binomial distribution given by  $n$  and  $p$  (the true error rate over the data distribution)



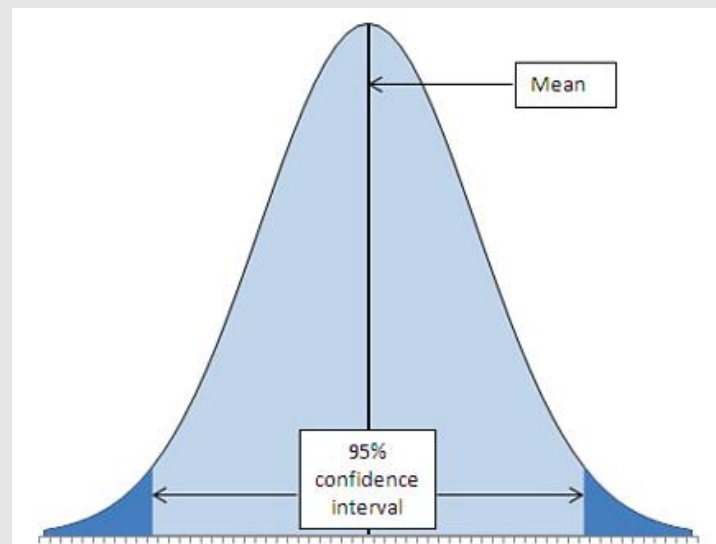
2. Most common way to determine a binomial confidence interval is to use the *normal approximation* (although can calculate exact intervals if  $n$  is not too large)

# Confidence intervals on error

2. When  $n \geq 30$ , and  $p$  is not too extreme, the normal distribution is a good approximation to the binomial



3. We can determine the  $C\%$  confidence interval by determining what bounds contain  $C\%$  of the probability mass under the normal



# Comparing learning systems



How can we determine if one learning system provides better performance than another

- for a particular task?
- across a set of tasks / data sets?



# Motivating example

	<u>Accuracies on test sets</u>				
System A:	80%	50	75	...	99
System B:	79	49	74	...	98
$\delta$ :	+1	+1	+1	...	+1

- Mean accuracy for System A is better, but the standard deviations for the two clearly overlap
- Notice that System A is always better than System B



# Comparing systems using a paired $t$ test



- consider  $\delta$ 's as observed values of a set of i.i.d. random variables
- *null hypothesis*: the 2 learning systems have the same accuracy
- *alternative hypothesis*: one of the systems is more accurate than the other
- hypothesis test:
  - use paired  $t$ -test to determine probability  $p$  that mean of  $\delta$ 's would arise from null hypothesis
  - if  $p$  is sufficiently small (typically  $< 0.05$ ) then reject the null hypothesis



# Comparing systems using a paired $t$ test

1. calculate the sample mean

$$\bar{\delta} = \frac{1}{n} \sum_{i=1}^n \delta_i$$

2. calculate the  $t$  statistic

$$t = \frac{\bar{\delta}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (\delta_i - \bar{\delta})^2}}$$

3. determine the corresponding  $p$ -value, by looking up  $t$  in a table of values for the Student's  $t$ -distribution with  $n-1$  degrees of freedom

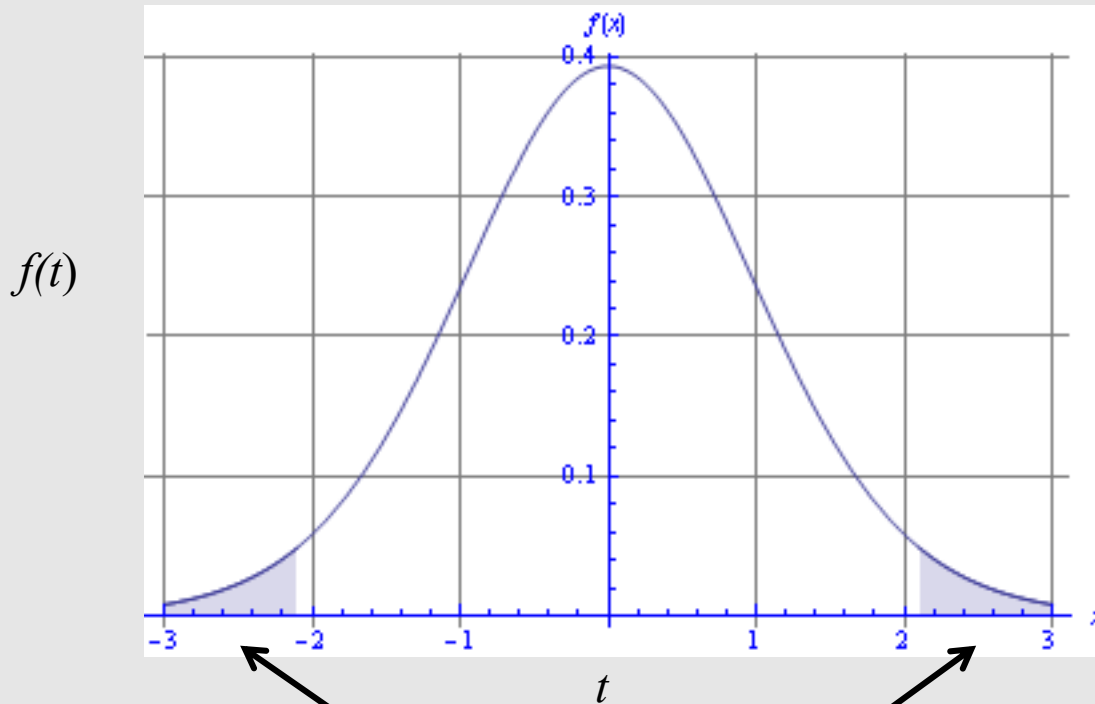
TABLE B.2 THE  $t$  DISTRIBUTION

Table entries are values of  $t$  corresponding to proportions in one tail or in two tails combined.

df	PROPORTION IN ONE TAIL				
	0.25	0.10	0.05	0.025	0.01
1	1.000	3.078	6.314	12.706	31.821
2	0.816	1.886	2.920	4.303	6.965
3	0.765	1.638	2.353	3.182	4.541
4	0.741	1.533	2.132	2.776	4.047
5	0.727	1.476	2.015	2.571	3.965
6	0.718	1.440	1.963	2.447	3.747
7	0.711	1.415	1.945	2.365	3.689
8	0.706	1.397	1.930	2.306	3.646
9	0.701	1.381	1.915	2.262	3.619
10	0.700	1.372	1.912	2.228	3.599
11	0.697	1.363	1.906	2.201	3.581
12	0.695	1.356	1.902	2.179	3.565
13	0.694	1.350	1.898	2.160	3.551
14	0.692	1.345	1.895	2.145	3.539
15	0.691	1.341	1.893	2.131	3.529
16	0.690	1.337	1.891	2.120	3.521
17	0.689	1.334	1.889	2.110	3.515
18	0.688	1.330	1.888	2.101	3.510
19	0.688	1.328	1.888	2.093	3.506
20	0.687	1.325	1.887	2.086	3.502
21	0.686	1.323	1.887	2.080	3.500
22	0.686	1.321	1.887	2.074	3.498
23	0.685	1.319	1.887	2.069	3.497
24	0.685	1.318	1.887	2.064	3.496
25	0.684	1.316	1.887	2.060	3.495
26	0.684	1.315	1.887	2.056	3.494
27	0.684	1.314	1.887	2.052	3.493
28	0.683	1.313	1.887	2.048	3.492
29	0.683	1.311	1.887	2.045	3.492
30	0.683	1.310	1.887	2.042	3.491
40	0.681	1.303	1.881	2.021	3.473
60	0.679	1.296	1.871	2.000	3.460
80	0.677	1.290	1.863	1.980	3.451
100	0.674	1.282	1.853	1.960	3.438

By H. B. Fisher and J. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 9th ed. London: Longman Group Ltd., 1963. Reprinted by permission of John Wiley & Sons, Inc. Adapted and revised with permission of the Addison-Wesley Longman Publishing Co.

# Comparing systems using a paired $t$ test



The null distribution of our  $t$  statistic looks like this

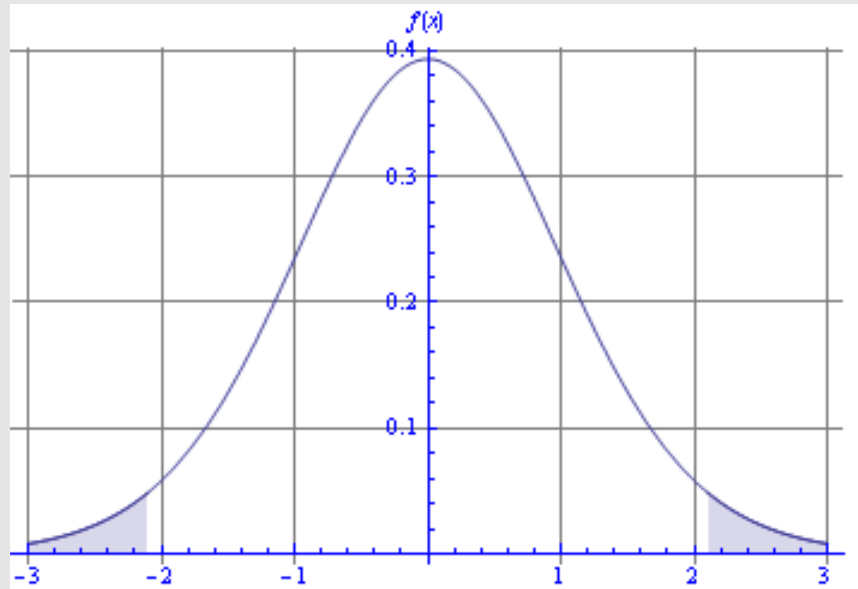
The  $p$ -value indicates how far out in a tail our  $t$  statistic is

If the  $p$ -value is sufficiently small, we reject the null hypothesis, since it is unlikely we'd get such a  $t$  by chance

for a two-tailed test, the  $p$ -value represents the probability mass in these two regions



# Why do we use a two-tailed test?



- a two-tailed test asks the question: is the accuracy of the two systems different
- a one-tailed test asks the question: is system A better than system B
- a priori, we don't know which learning system will be more accurate (if there is a difference) – we want to allow that either one might be



# Comments on hypothesis testing to compare learning systems

- the paired  $t$ -test can be used to compare two learning systems
- other tests (e.g. McNemar's  $\chi^2$  test) can be used to compare two learned models
- a statistically significant difference is not necessarily a large-magnitude difference



# Scatter plots for pairwise method comparison

We can compare the performance of two methods *A* and *B* by plotting (*A performance*, *B performance*) across numerous data sets

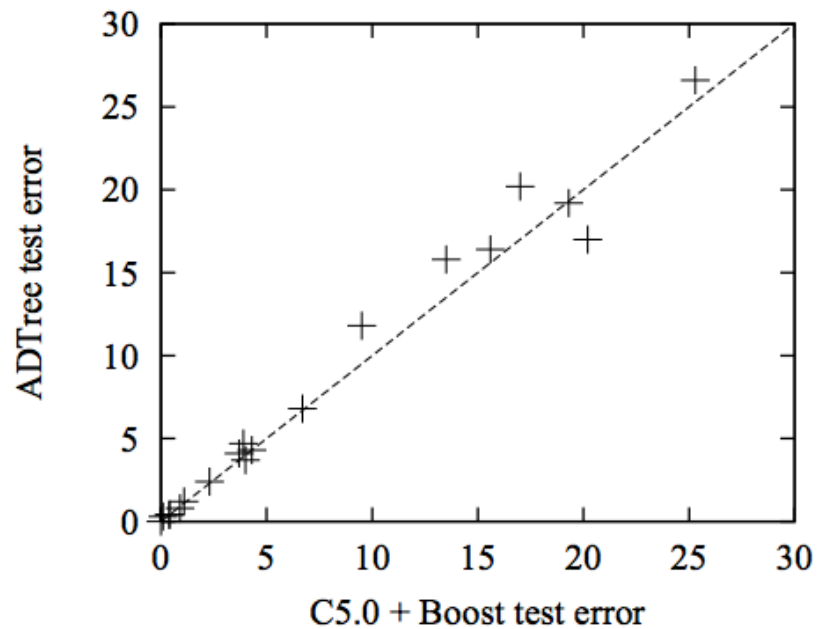


figure from Freund & Mason, *ICML* 1999

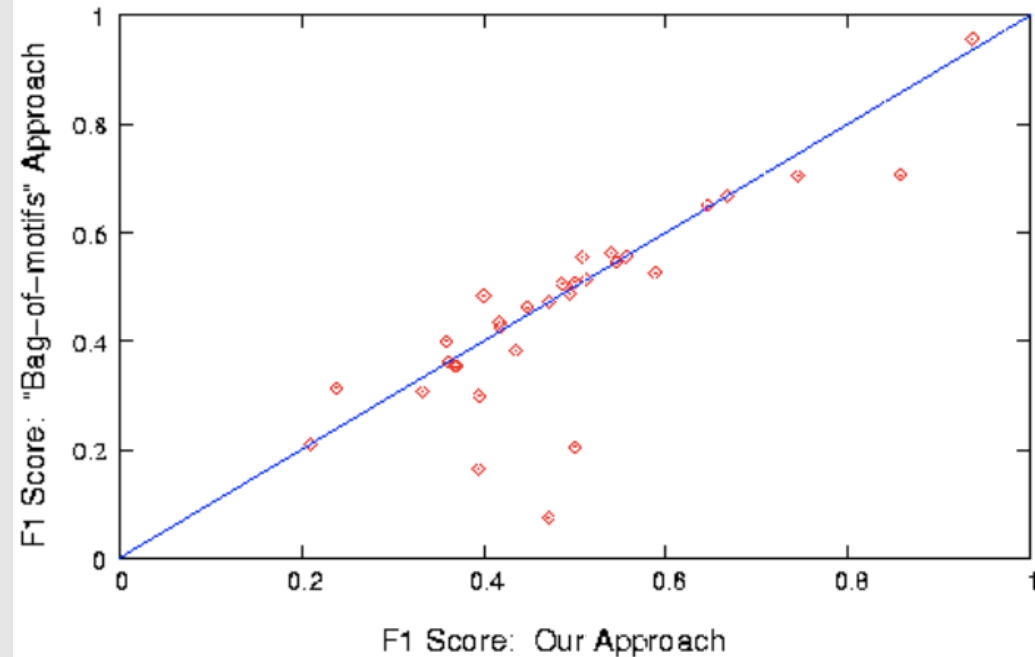


figure from Noto & Craven, *BMC Bioinformatics* 2006

# Lesion studies



We can gain insight into what contributes to a learning system's performance by removing (lesioning) components of it

The ROC curves here show how performance is affected when various feature types are removed from the learning representation

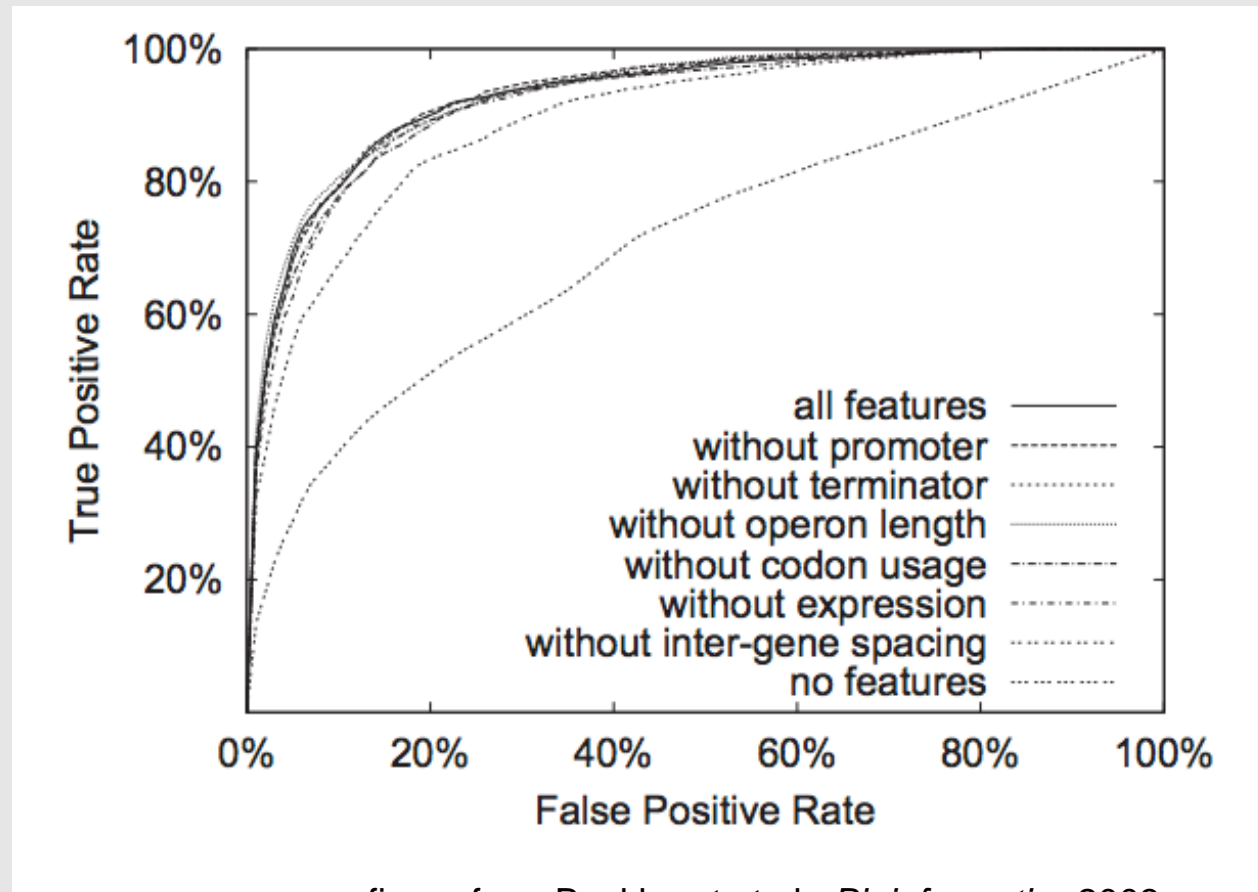


figure from Bockhorst et al., *Bioinformatics* 2003

# To avoid pitfalls, ask



1. Is my held-aside test data really representative of going out to collect new data?
  - Even if your methodology is fine, someone may have collected features for positive examples differently than for negatives – should be randomized
  - Example: samples from cancer processed by different people or on different days than samples for normal controls



# To avoid pitfalls, ask



2. Did I repeat my entire data processing procedure on every fold of cross-validation, using only the training data for that fold?
  - On each fold of cross-validation, did I ever access in any way the label of a test instance?
  - Any preprocessing done over entire data set (feature selection, parameter tuning, threshold selection) must not use labels

# To avoid pitfalls, ask



3. Have I modified my algorithm so many times, or tried so many approaches, on this same data set that I (the human) am overfitting it?
  - Have I continually modified my preprocessing or learning algorithm until I got some improvement on this data set?
  - If so, I really need to get some additional data now to at least test on



# THANK YOU

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, and Pedro Domingos.

