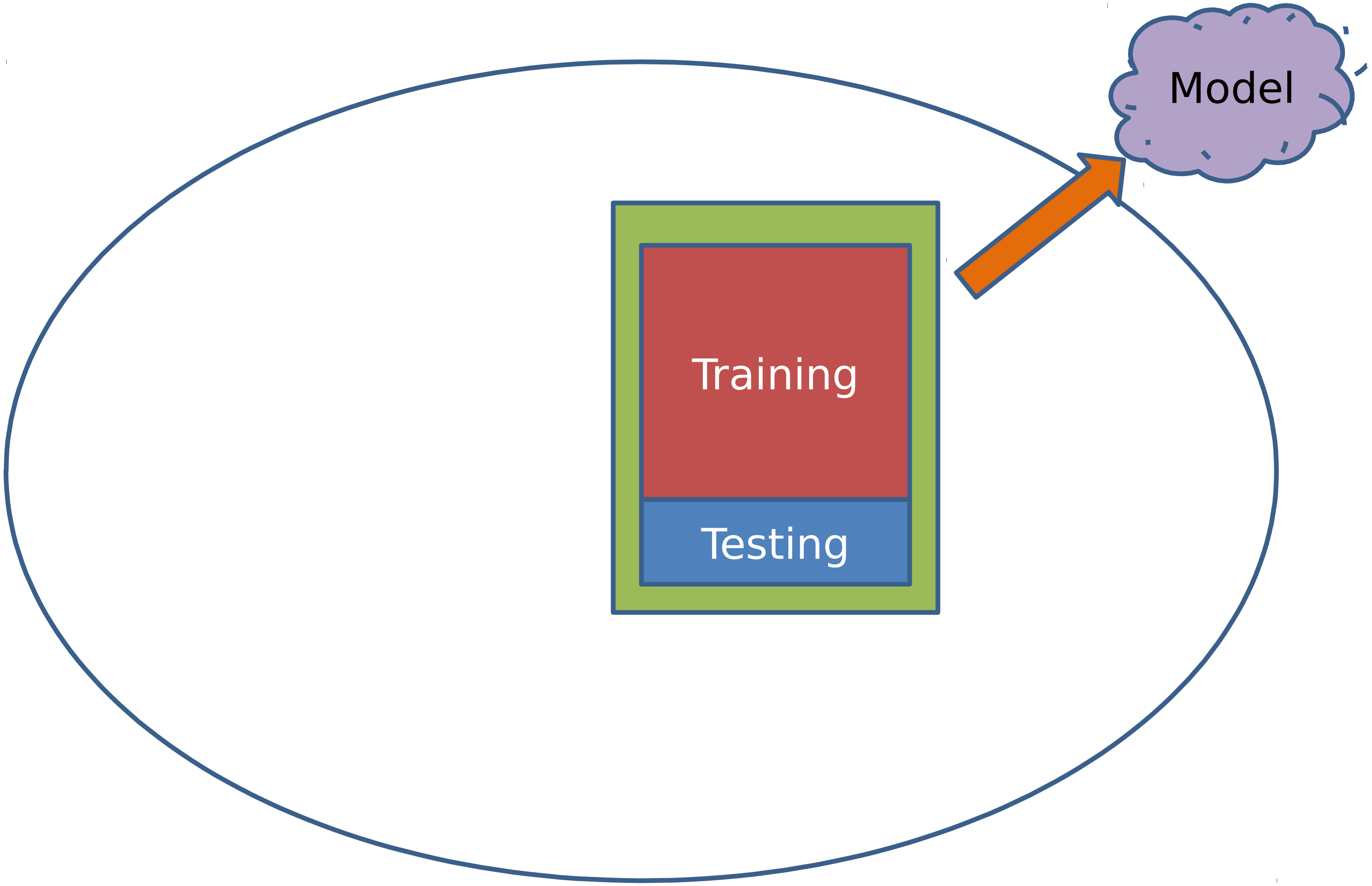
# Evaluating a Classifier

Jakramate Bootkrajang
CS, CMU

# Why testing a classifier

- You have built a classifier.
- You'd like to know how good your classifier is
- For example, you are building a classifier for classifying cancer patients from normal patients using microarray data
- You collect data from SuanDok Hospital
- You can only test the classifier on the (limited) data you have

# Types of data



# Basic Performance Measures (1/2)

- Confusion matrix

| | | **Predicted** | |
|---|---|---|---|
| | | Negative | Positive |
| Actual label | Negative | a<br>True Negative | b<br>False Positive |
| | Positive | c<br>False Negative | d<br>True Positive |

# Basic Performance Measures (2/2)

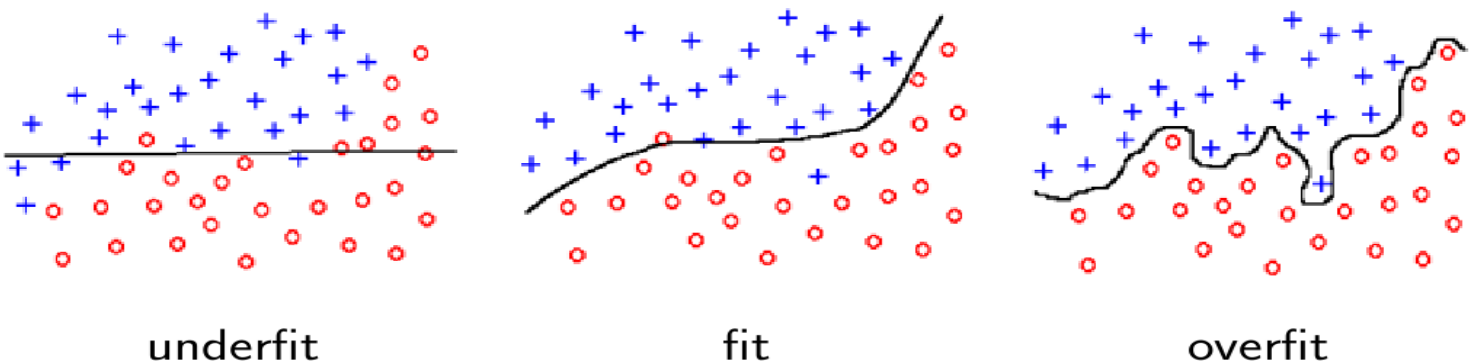
- **Accuracy** = (a + d) / (a + b + c + d) = (TN+TP)/total = 1-error
- **True positive rate** = d/(c + d), recall
- True negative rate = a/(a + b), specificity
- **False positive rate** = b/(a + b), false alarms
- False negative rate = c/(c + d), 1-specificity

# Performance measures based on types of data

- Performance on training data
  - Training error,  (accuracy)
- Performance on test data
  - Test error
- Performance on unseen data
  - Generalisation error
- Performance on validation data
  - Validation error (used for model selection)

# Danger of Overfitting

- Overfitting = model fits to the data so well
- Overfitting is the opposite of to generalise
- Training error is small but test/generalisation error is large.
- Causes: too small dataset,  train the model for too long



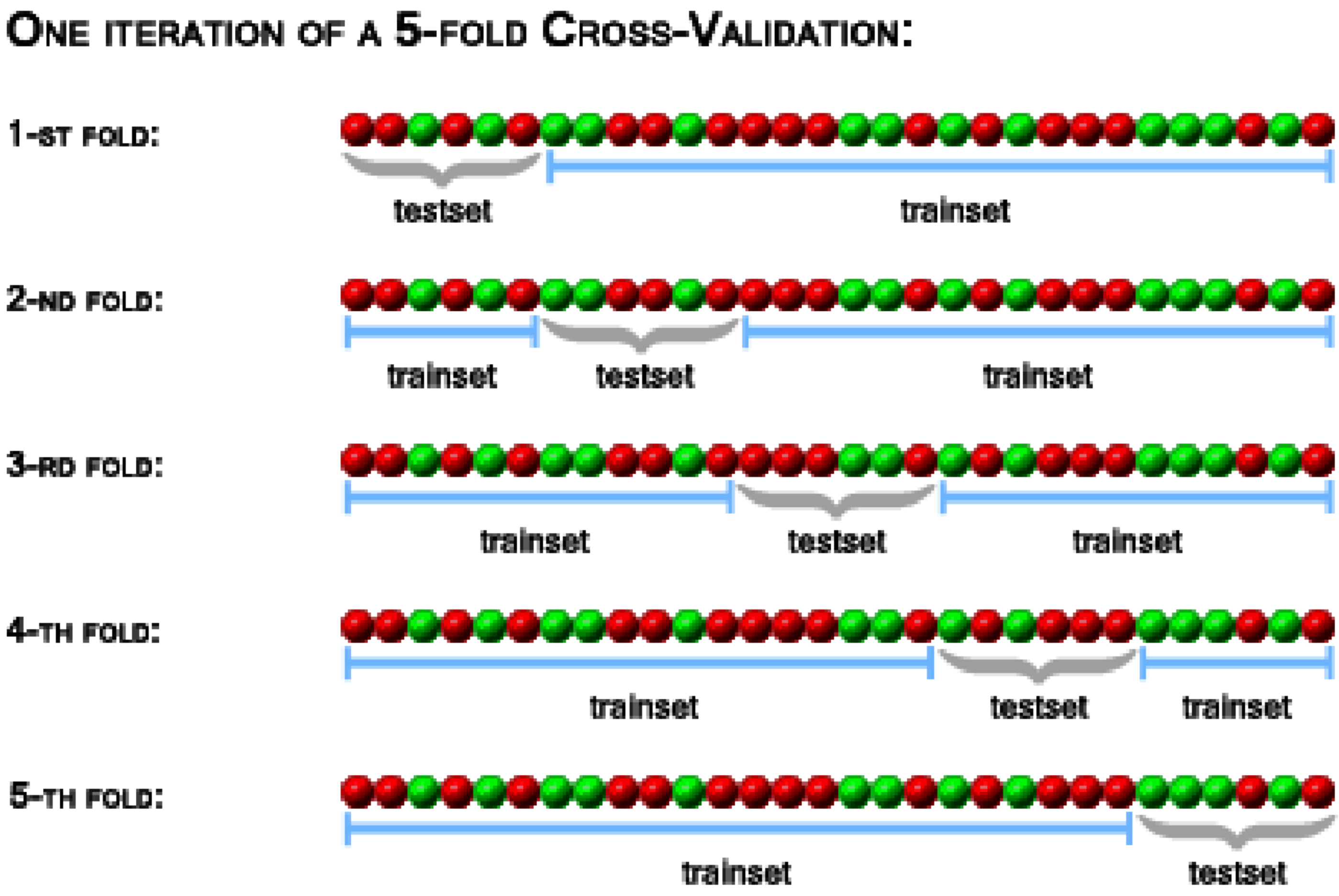
# Theoretically expand your data

- Motivation
  - More training data gives better generalisation
  - More test data gives better classification error probability
- Partitioning
  - Holdout
  - Cross validation
  - Bootstrap

# Hold out method

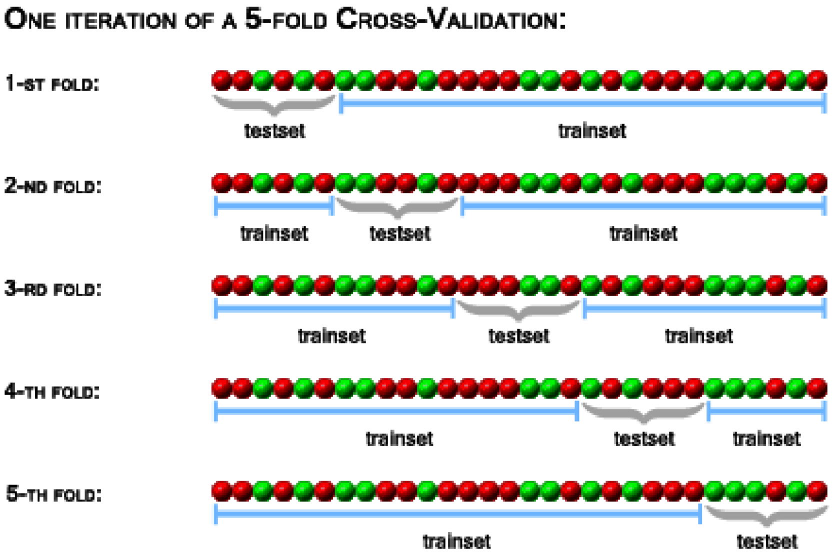
- Dataset is randomly partitioned into two independent sets.
  - Training set (e.g., 2/3 of data) for the model construction.
  - Test set (e.g., 1/3 of data) is ***hold out*** for accuracy estimation of the classifier.
- Repeat the hold out *k* times,
- Final accuracy is the average of the accuracies obtained

# K-fold Cross validation



# K-fold Cross validation

- The dataset is randomly divided into K disjoint sets of equal size.
- Train the classifier K times, each with different held out test set.
- Estimated error is the mean of K errors



# Leave-one-out Cross validation

- A special case of K-fold CV with K=n
- Where n is the number of samples
- n experiments are preformed using n-1 examples for training and the remaining example for testing
- Computationally expensive

# Bootstrap (bagging)

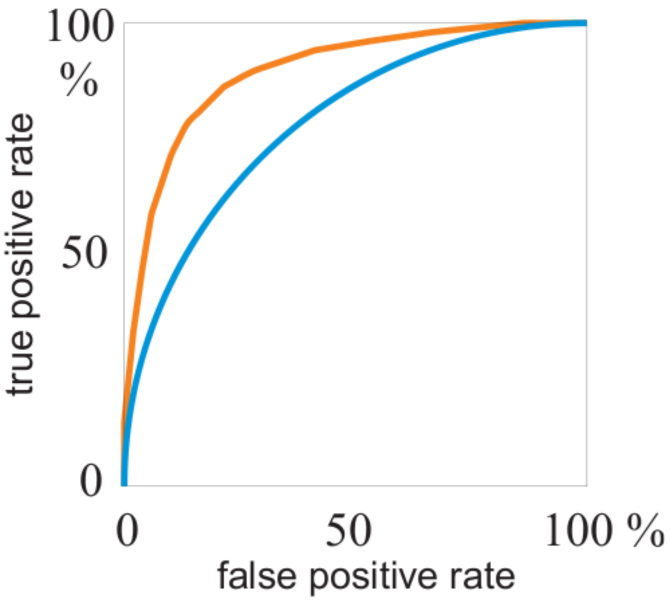
- The bootstrap uses sampling with replacement to form the training set.
- Given: the training set $T$ consisting of $n$ entries.
- Bootstrap generates $m$ new datasets $T_i$ each of size $n' < n$ by sampling $T$ uniformly with replacement. The consequence is that some entries can be repeated in $T_i$.
- In a special case (called 632 boosting) when $n' = n$, for large $n$, $T_i$ is expected to have $1 - \frac{1}{e} \approx 63.2\%$ of unique samples. The rest are duplicates.
- The $m$ statistical models (e.g., classifiers, regressors) are learned using the above $m$ bootstrap samples.
- The statistical models are combined, e.g. by averaging the output (for regression) or by voting (for classification).

# Recommended protocol

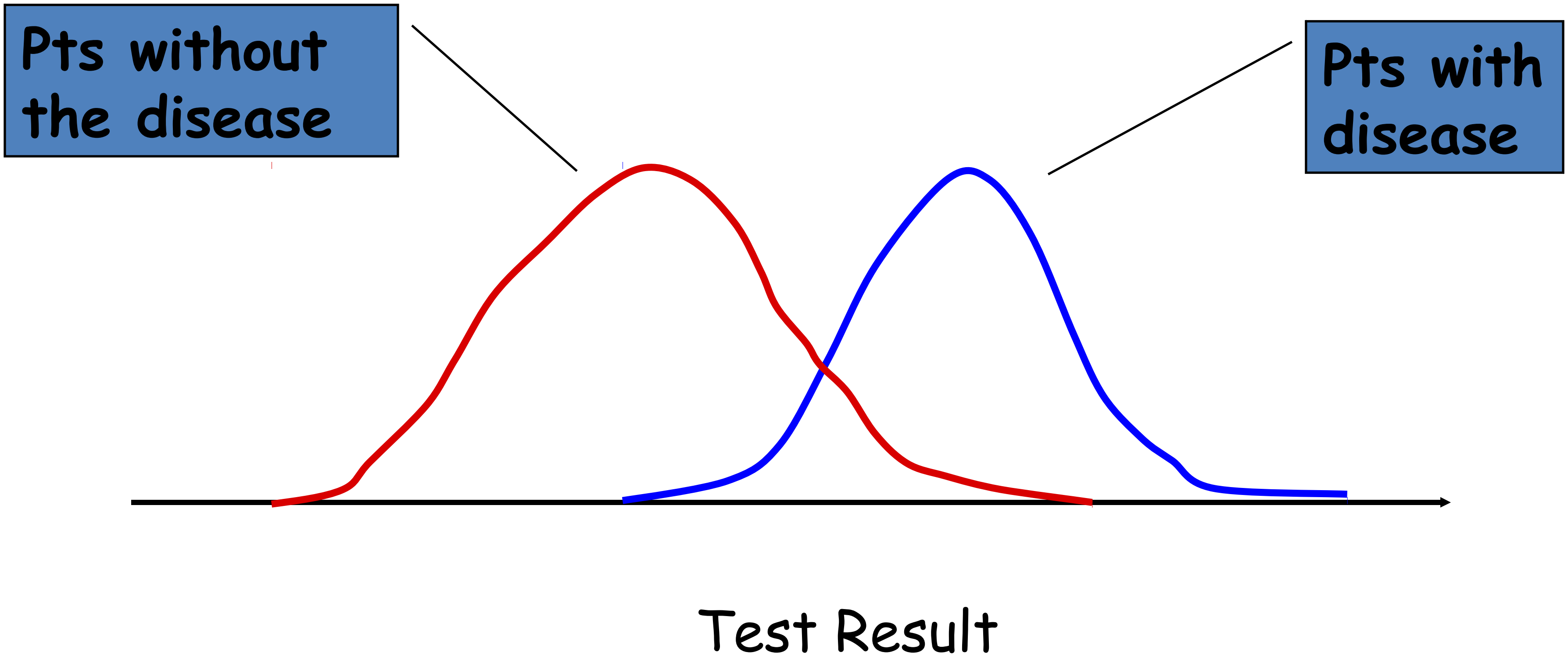
- Use $K$-fold cross-validation $(K = 5$ or $K = 10)$ for estimating performance estimates (accuracy, etc.).
- Compute the mean value of performance estimate, and standard deviation and confidence intervals.
- Report mean values of performance estimates and their standard deviations or 95% confidence intervals around the mean.

# ROC- Receiver Operating Characteristic

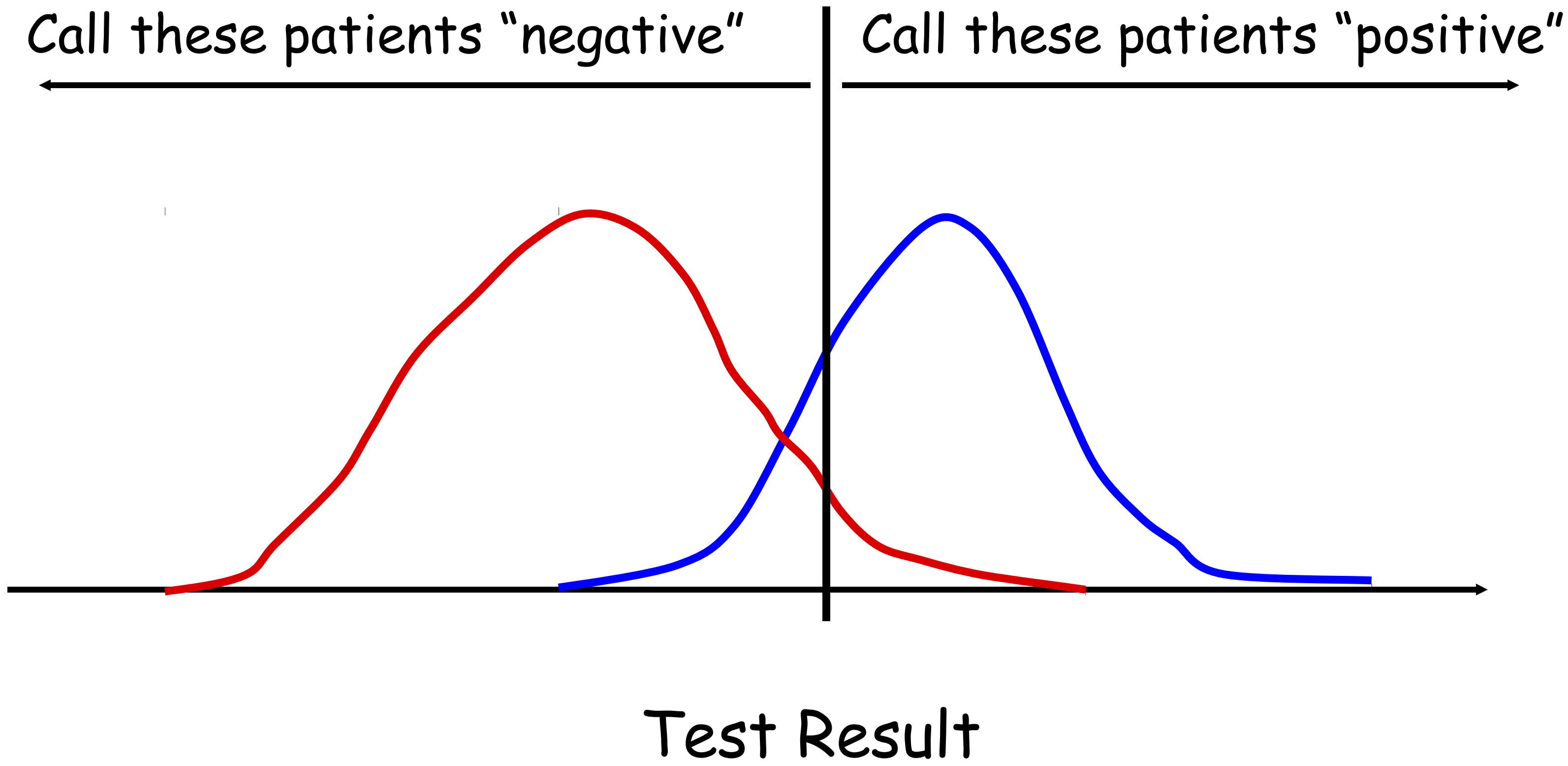
- Called also often ROC curve.
- Originates in WWII processing of radar signals.
- Useful for the evaluation of dichotomic classifiers performance.
- Characterizes degree of overlap of classes for a single feature.
- Decision is based on a single threshold $\Theta$ (called also operating point).
- Generally, false alarms go up with attempts to detect higher percentages of true objects.
- A graphical plot showing (hit rate, false alarm rate) pairs.
- Different ROC curves correspond to different classifiers. The single curve is the result of changing threshold $\Theta$.



# Specific Example

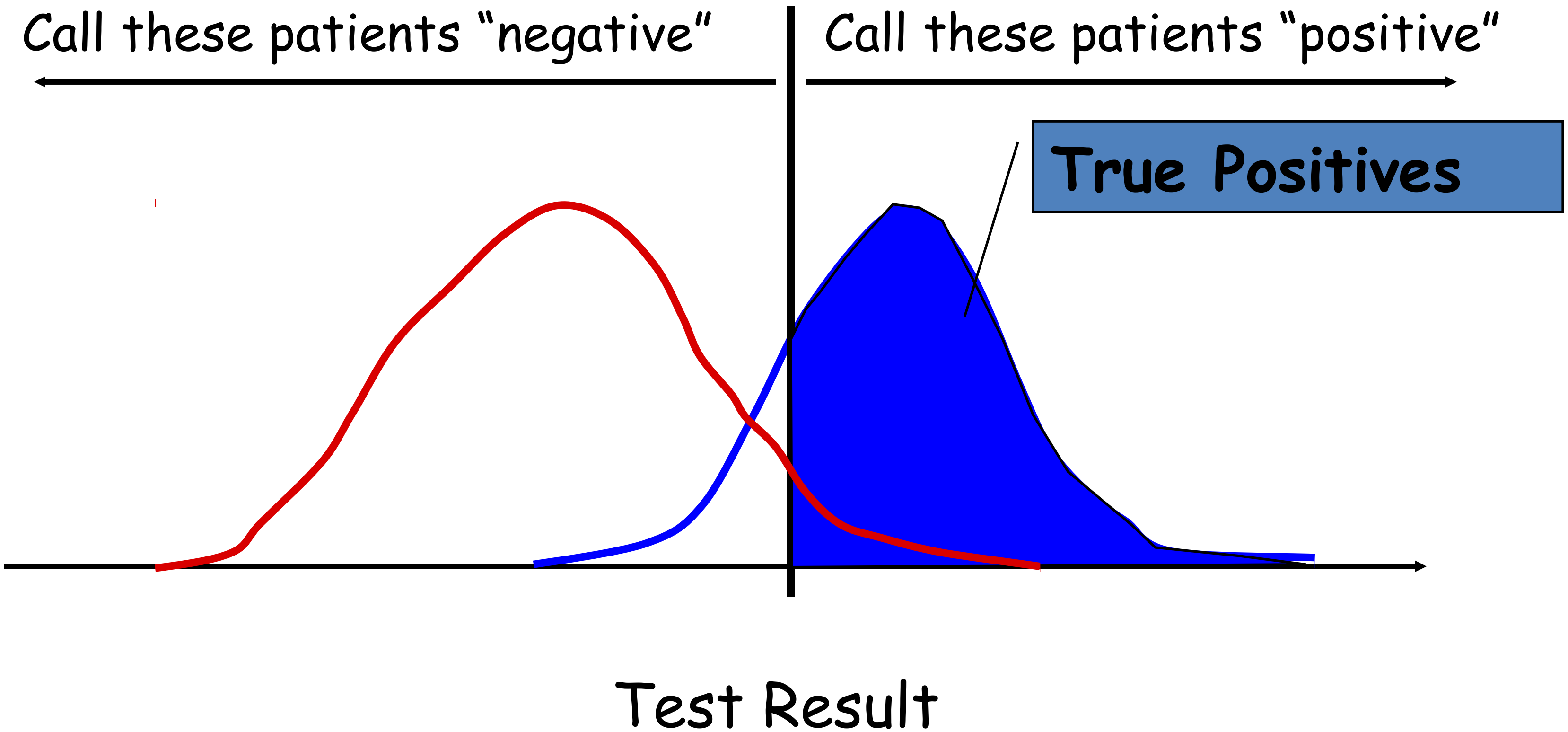


# Threshold

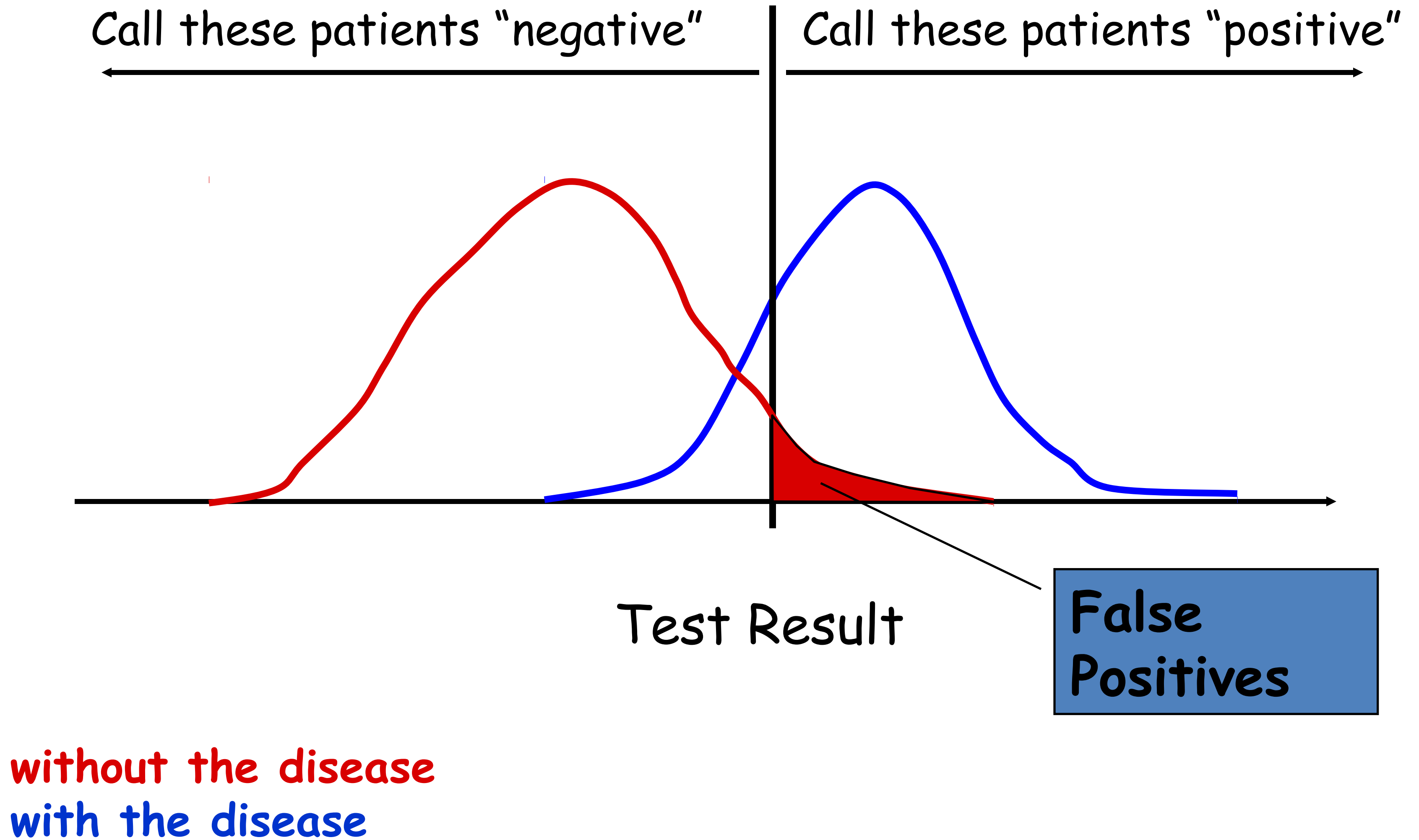


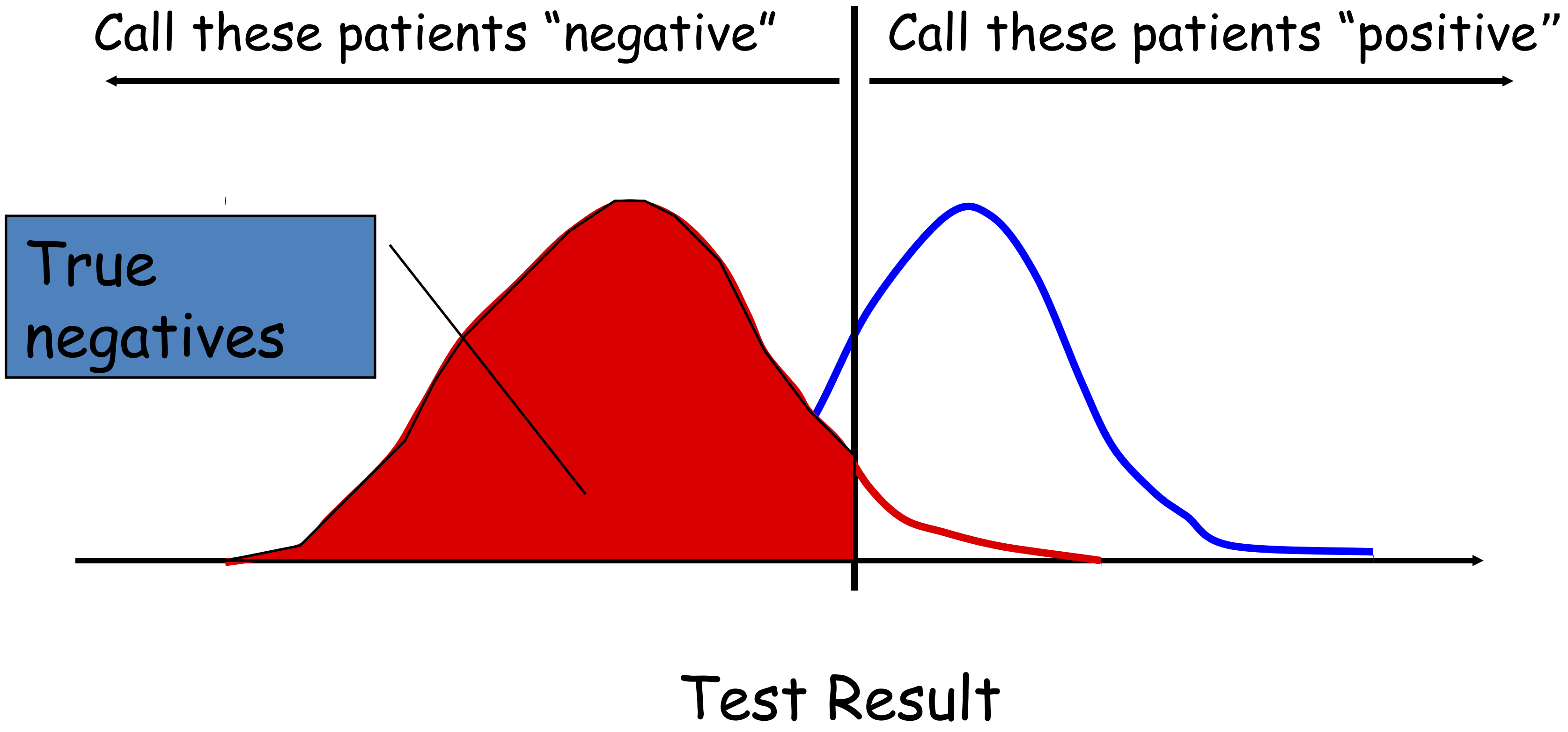
Call these patients "negative"
Call these patients "positive"
Test Result

# Some definitions ...

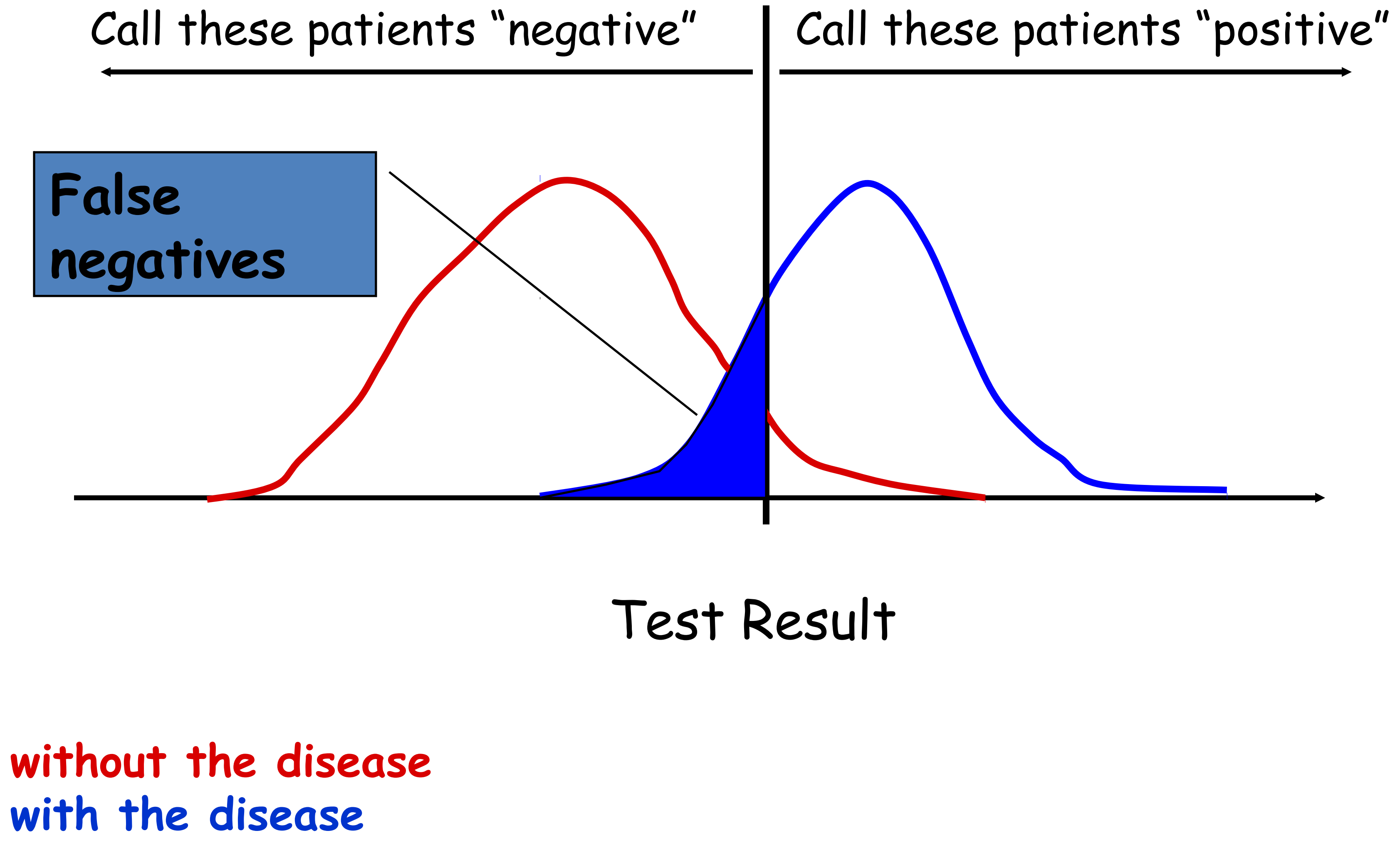



without the disease
with the disease

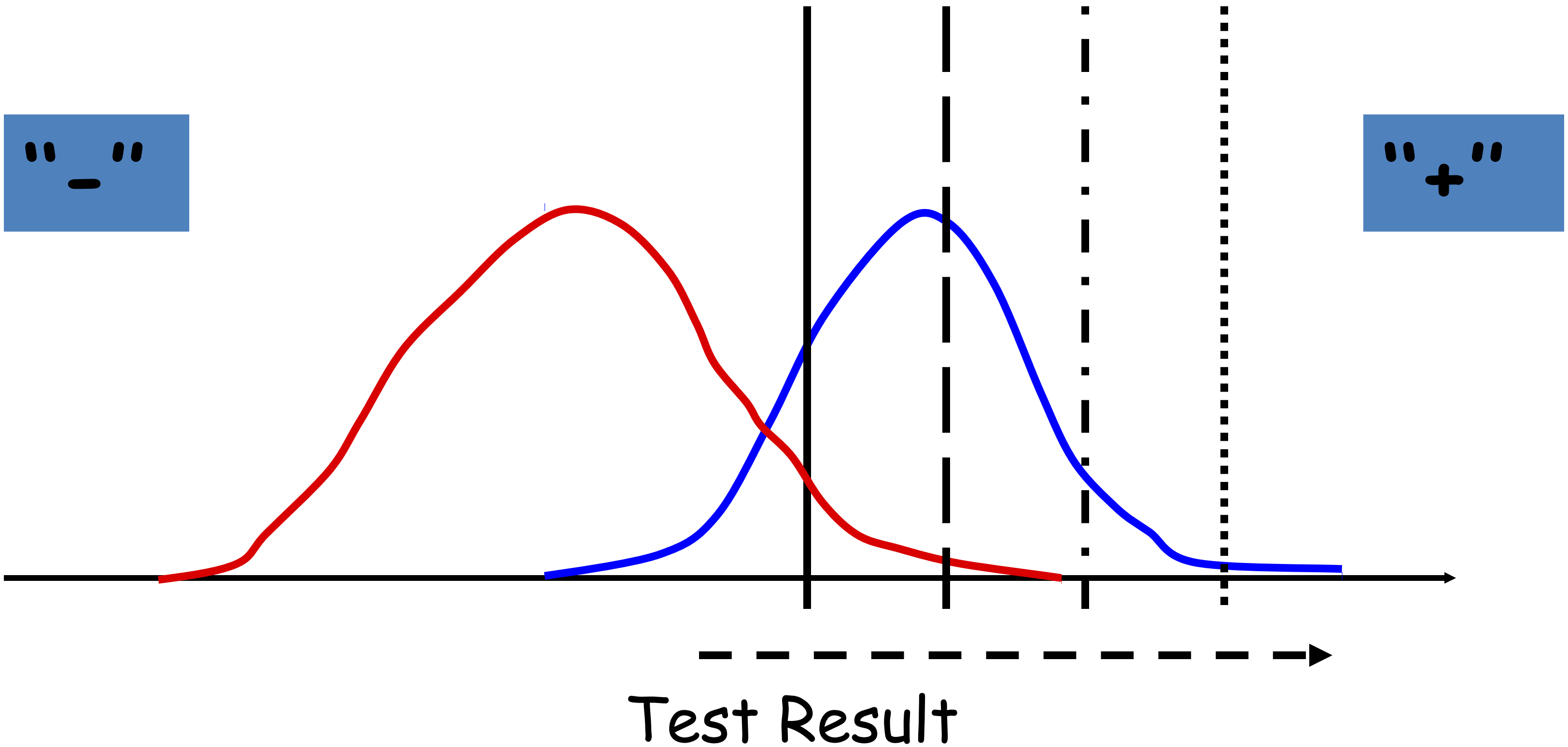

Call these patients "negative"
Call these patients "positive"
Test Result
False Positives
without the disease
with the disease




without the disease
with the disease

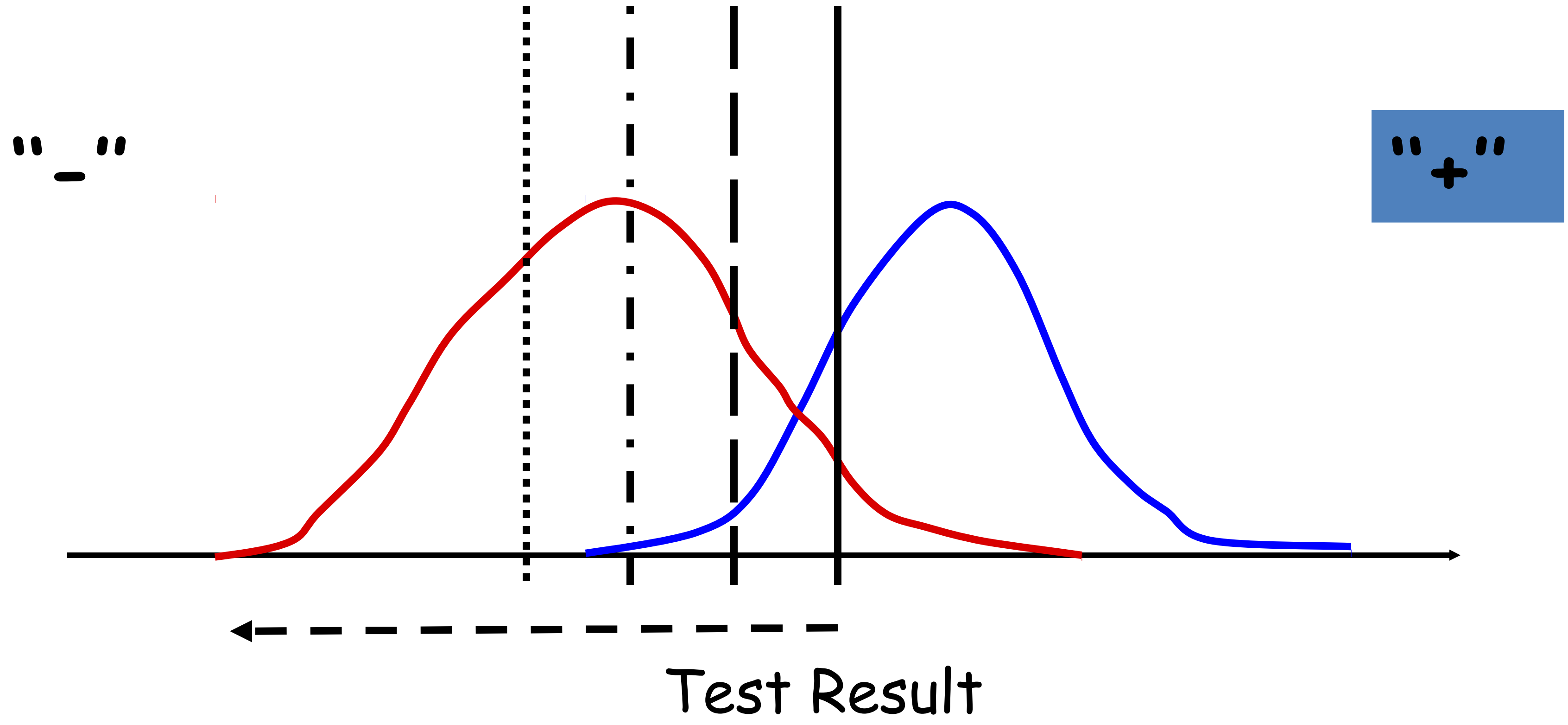

Call these patients "negative"
Call these patients "positive"
False negatives
Test Result
without the disease
with the disease

# Moving the Threshold: right




without the disease
with the disease

# Moving the Threshold: left

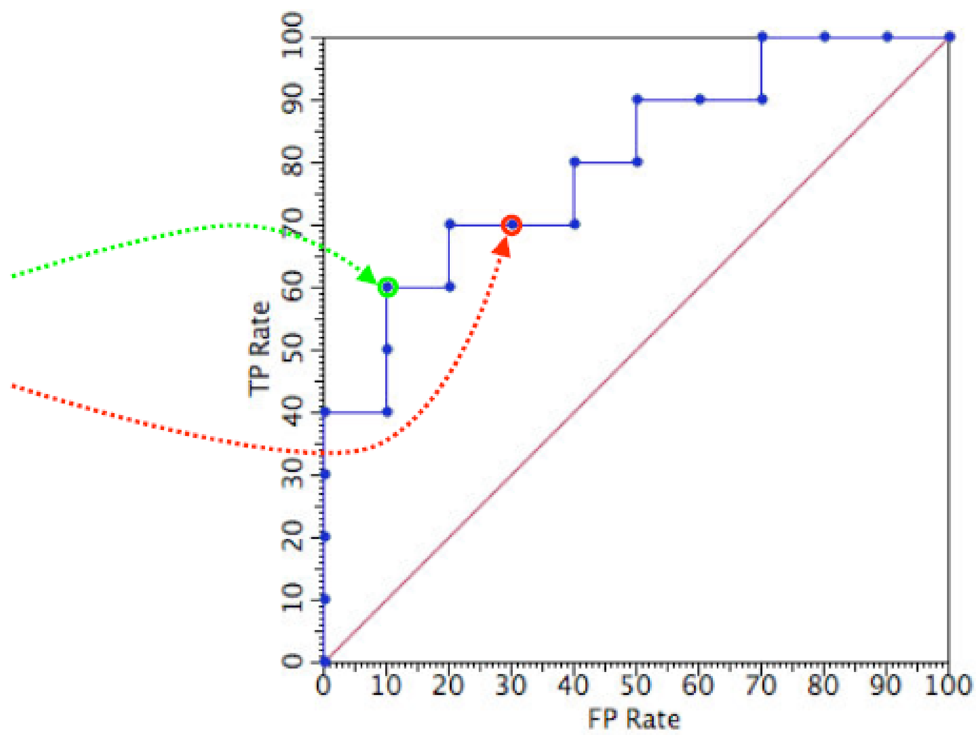



without the disease
with the disease

# How to plot

- Sort the predictions based on confidences
- Start from the most sure prediction
  - Set a threshold equals to the confidences (scores)
  - If we see + we move up, else we move left
- Continue until we reach the least sure prediction

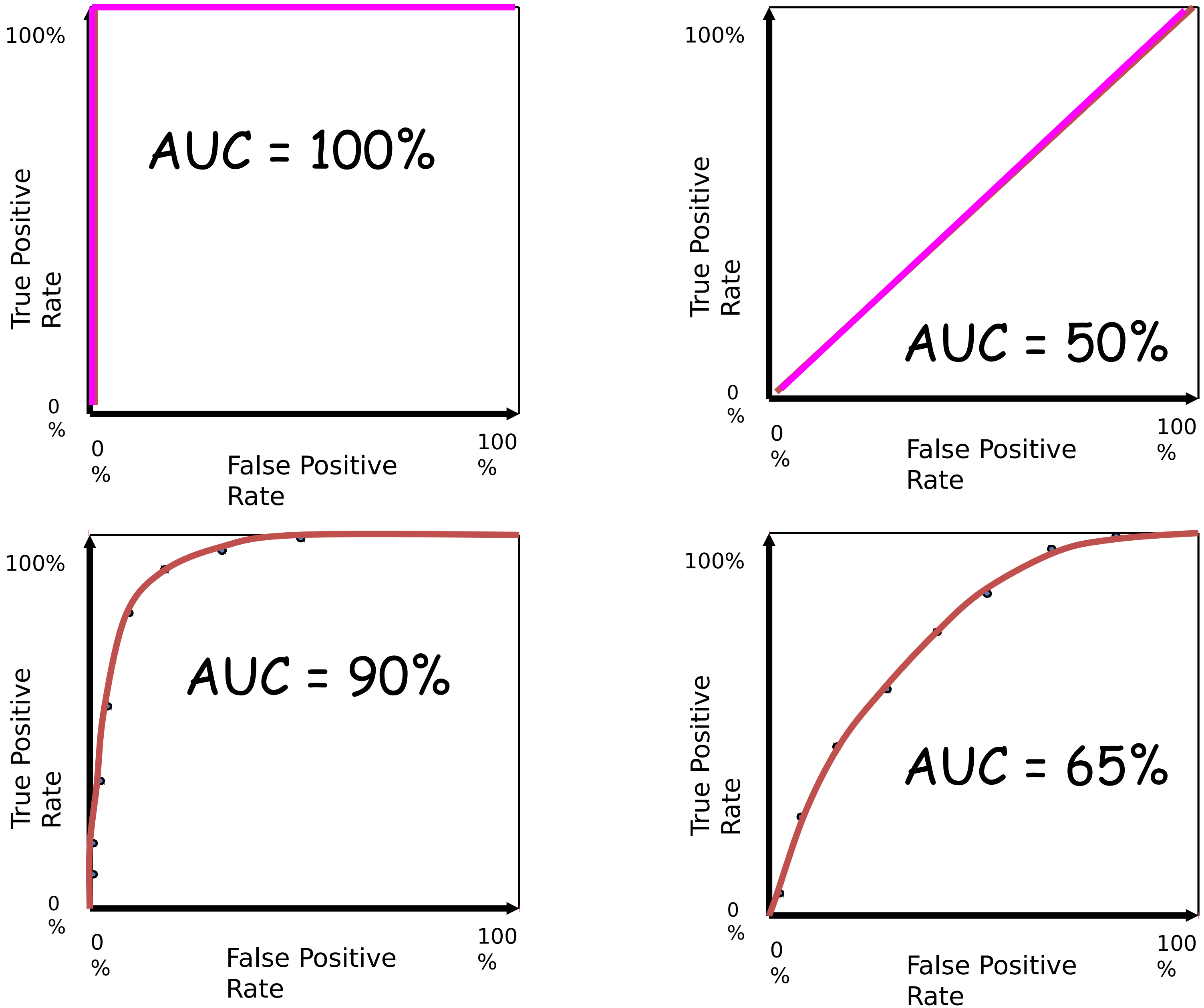
| Class | Score |
|---|---|
| + | 0.98 |
| + | 0.93 |
| + | 0.87 |
| + | 0.84 |
| – | 0.79 |
| + | 0.73 |
| + | 0.67 |
| – | 0.62 |
| + | 0.57 |
| – | 0.54 |
| – | 0.48 |
| + | 0.43 |
| – | 0.37 |
| + | 0.34 |
| – | 0.28 |
| – | 0.24 |
| + | 0.18 |
| – | 0.12 |
| – | 0.09 |
| – | 0.03 |



# Area under ROC curve (AUC)

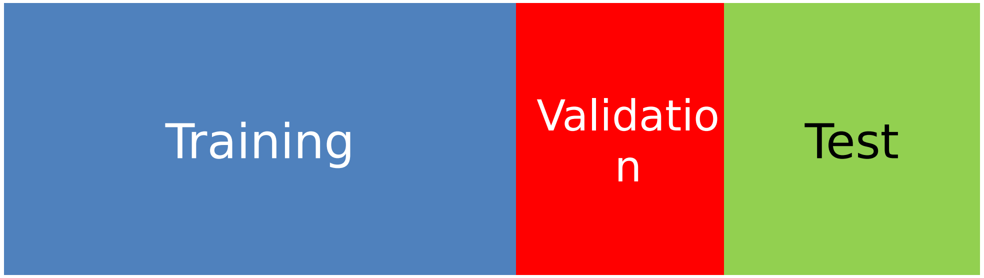
- *Overall measure* of test performance
- Equals to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one

# AUC for ROC curves



# Parameter tuning

- Most classifiers have parameters
- Their values need to be chosen
- For example, the value of k for an k-NN
- Use validation set for this purpose



# No free lunch theorem

- All algorithms that search for an extremum of a cost function perform exactly the same when averaged over all possible cost functions.
- So, for any search/optimization algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class.

# Summary

- Ways to evaluate a classifier
  - Based on information of the confusion matrix
- Ways to increase the size of the dataset
- Receiver Operating Curve
- Area Under ROC Curve
- Parameter tuning
  - Using another hold out validation set
- No free lunch theorem
  - No best classifier.