Lecture 11

Classification analysis

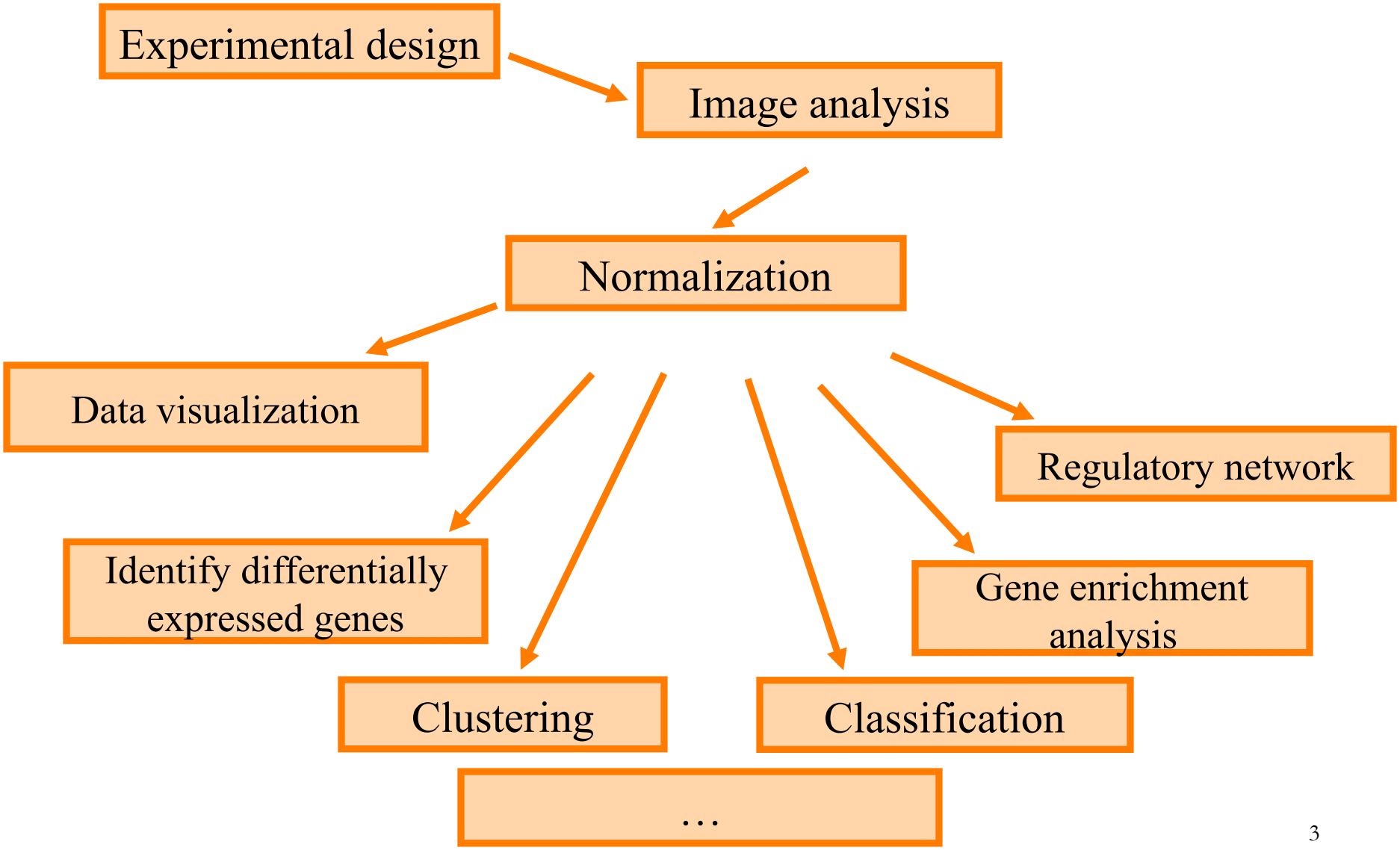
MCB 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

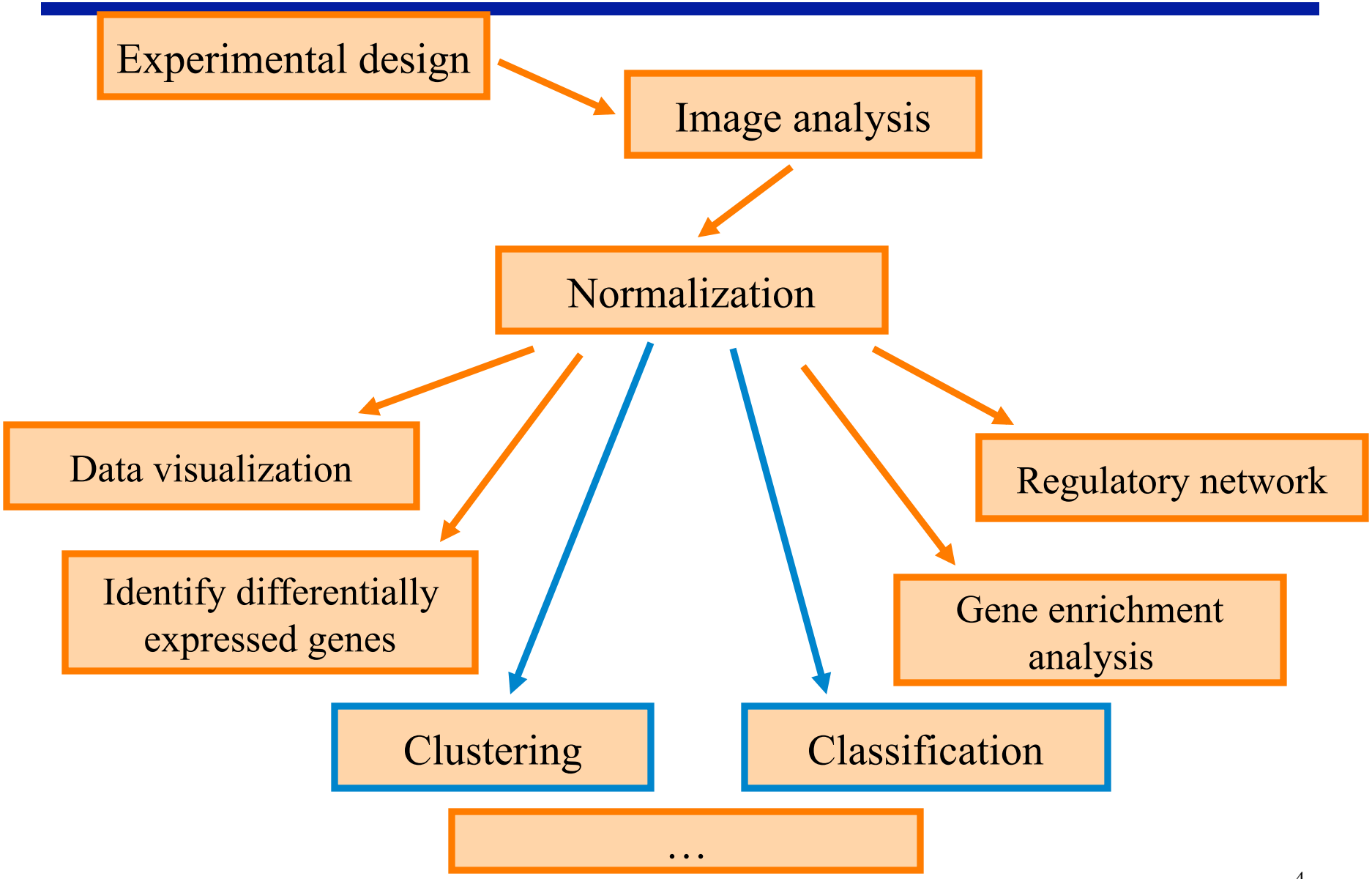
Outline

- Introduction to classification
- Why gene select
- Performance assessment
- Case study

Statistical Issues in Microarray Analysis



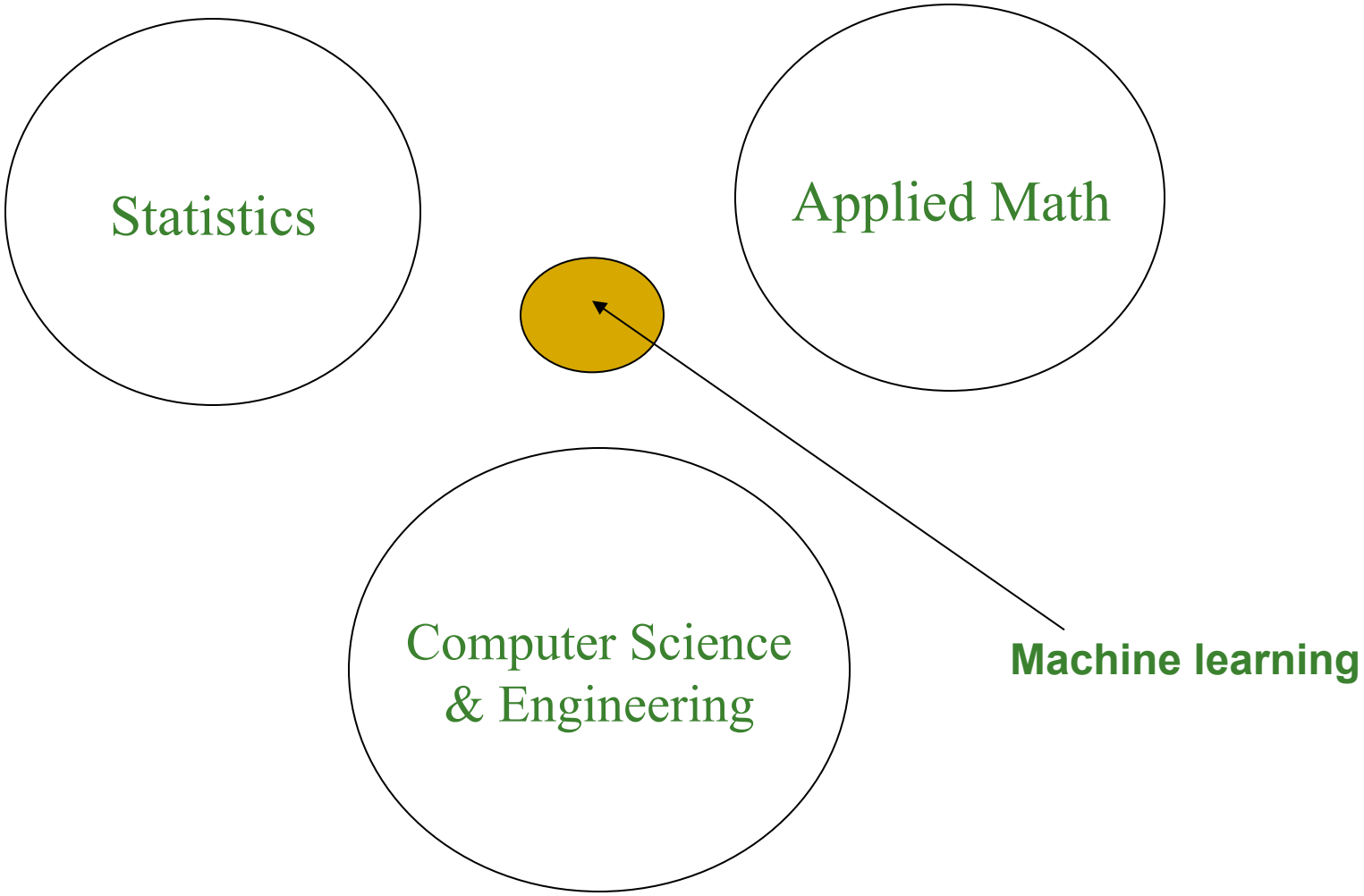
Statistical Issues in Microarray Analysis



Cluster analysis vs. Classification

- Alternative terminology
  - —— computer science: **unsupervised** and **supervised learning.**
  - —— biological literature: **class discovery** and **class prediction.**

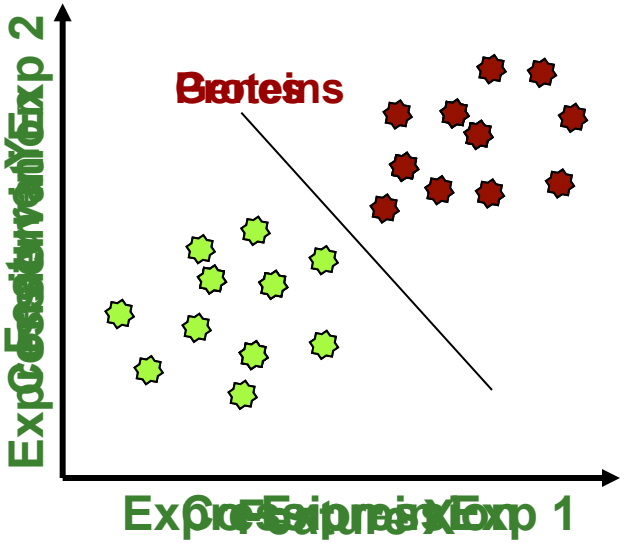
Machine learning (classification + clustering) =(supervised analysis +unsupervised analysis)

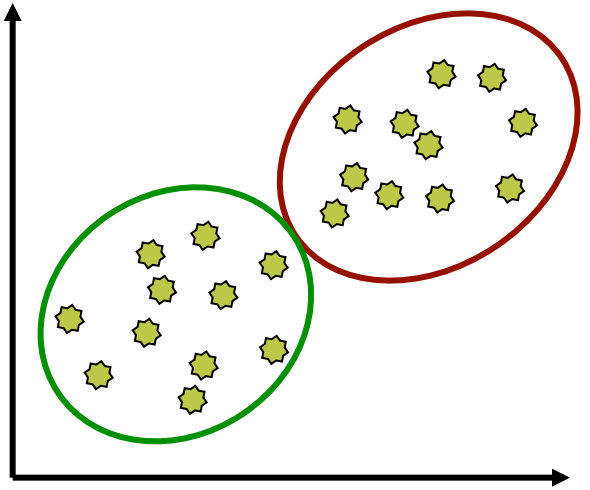


A very interdisciplinary field with long history

The Basic Idea – classification & clustering

- Objects characterized by one or more features
- Classification
  - Have labels for some points
  - Want a "rule" that will accurately assign labels to new points
  - Supervised learning
- Clustering
  - No labels
  - Group points into clusters based on how "near" they are to one another
  - Identify structure in data
  - Unsupervised learning



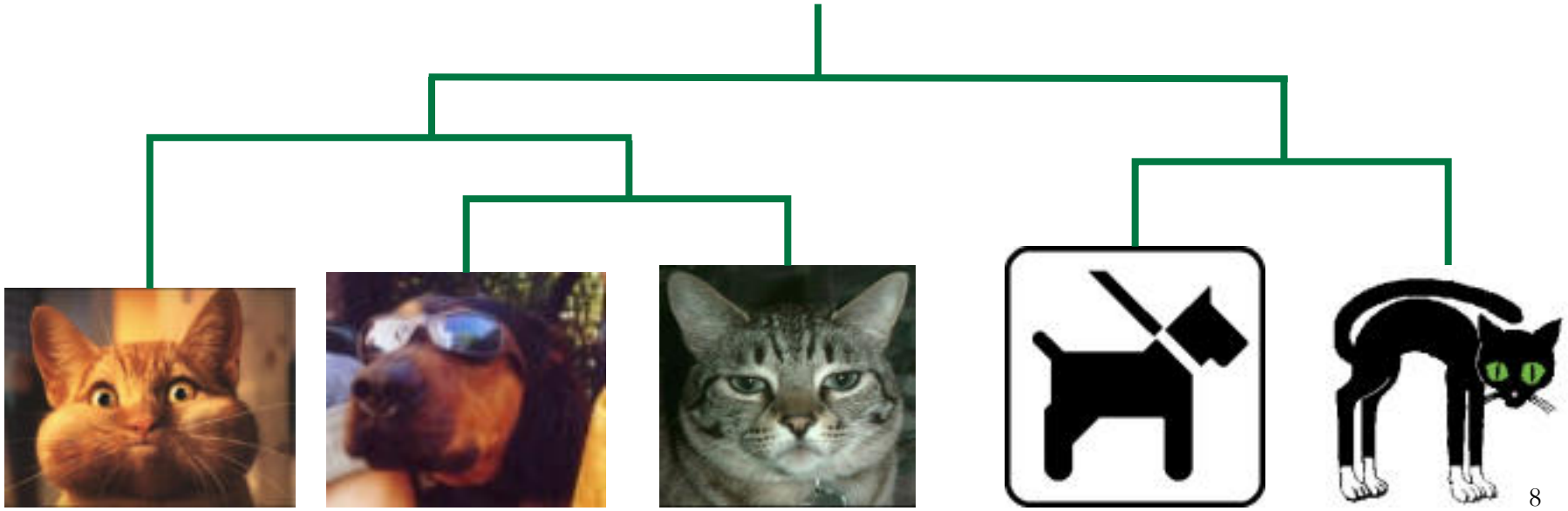


Un-supervised analysis

Calvin, I still don't know the difference between cats and dogs …

I don't know it either.

Let's try to figure it out together …



Supervised analysis

Calvin, I still don't know the difference between cats and dogs …
Oh, now I get it!!

Don't worry! I'll show you once more:

Class 1: cats



Class 2: dogs


Supervised analysis

= learning from examples, classification

— Assume we have already seen groups of healthy and sick people. Now let's diagnose the next person walking into the hospital.

— We know that a group of genes have function X (and these others don't). Let's find more genes with function X.

— We know many gene-pairs that are functionally related (and many more that are not). Let's extend the number of known related gene pairs.

Known structure in the data needs to be generalized to new data.

Un-supervised analysis

= clustering

— Are there groups of genes that behave similarly in all conditions?

— Disease X is very heterogeneous. Can we identify more specific sub-classes for more targeted treatment?

No structure is known. We first need to find it. Exploratory analysis.

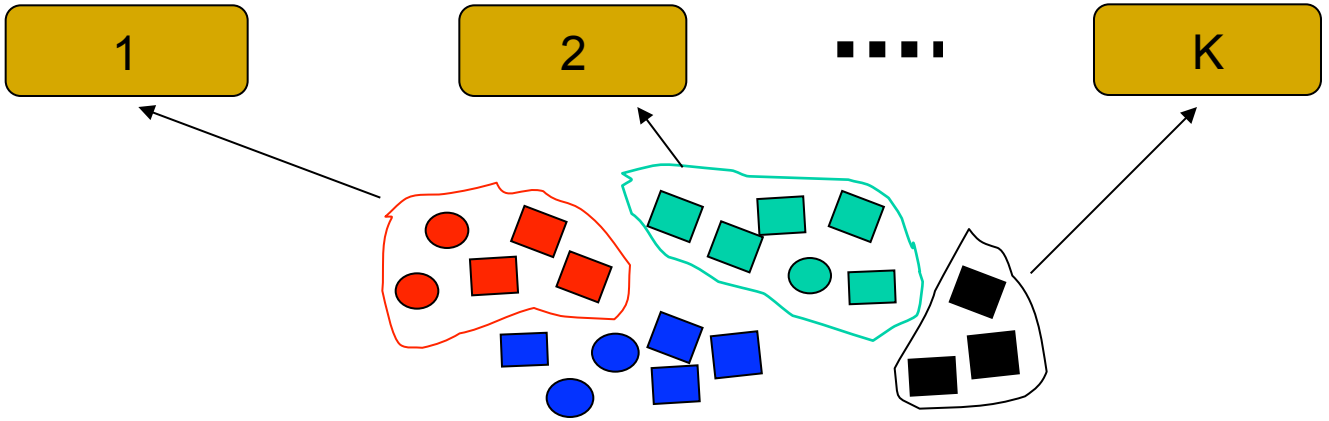
Supervised analysis: setup

- **Training set**
  - Data: microarrays
  - Labels: for each one we know if it falls into our class of interest or not (binary classification)
- **New data (test data)**
  - Data for which we don't have labels.
  - Eg. Genes without known function
- **Goal: Generalization ability**
  - Build a classifier from the training data that is good at predicting the right class for the new data.

Basic principles of classification/discrimination

Each object associated with a class label (or response) $Y \in \{1, 2, \ldots, K\}$ and a feature vector (vector of predictor variables) of G measurements: $X = (X_1, \ldots, X_G)$

Aim: predict Y from X.



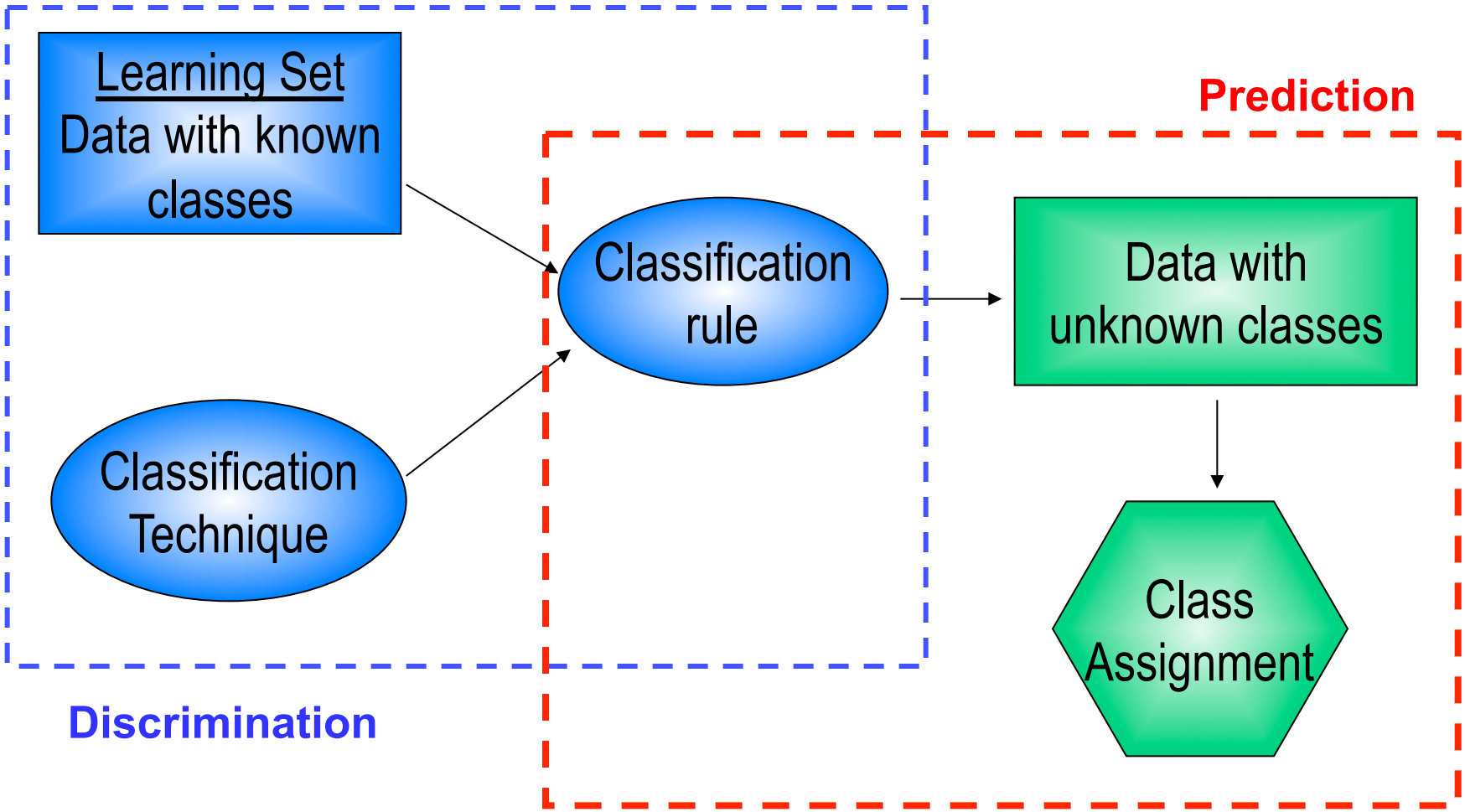
Predefined Class {1,2,…K}

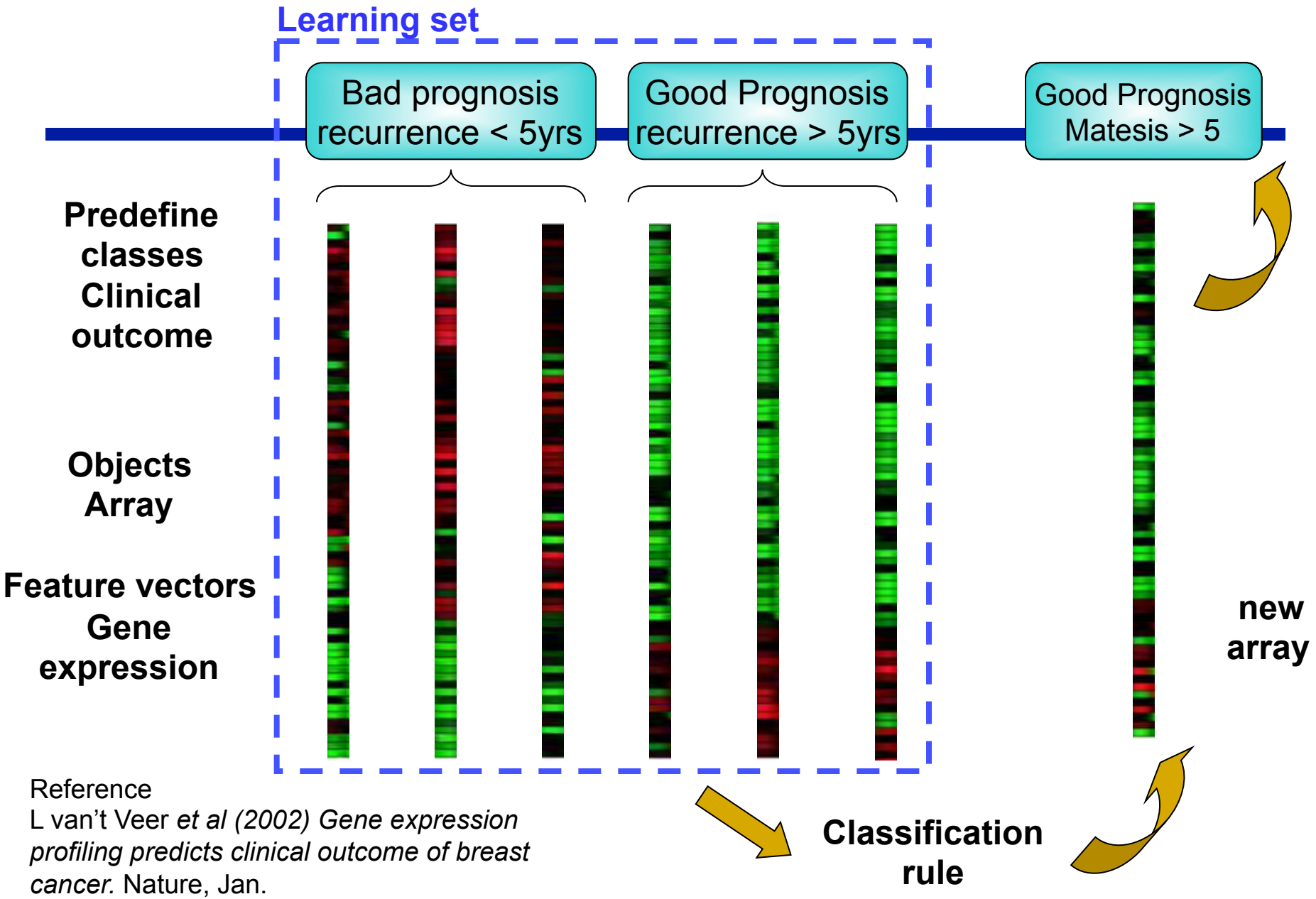
X = Feature vector {colour, shape}

Classification rule ?

X = {black, round}
Y = ?

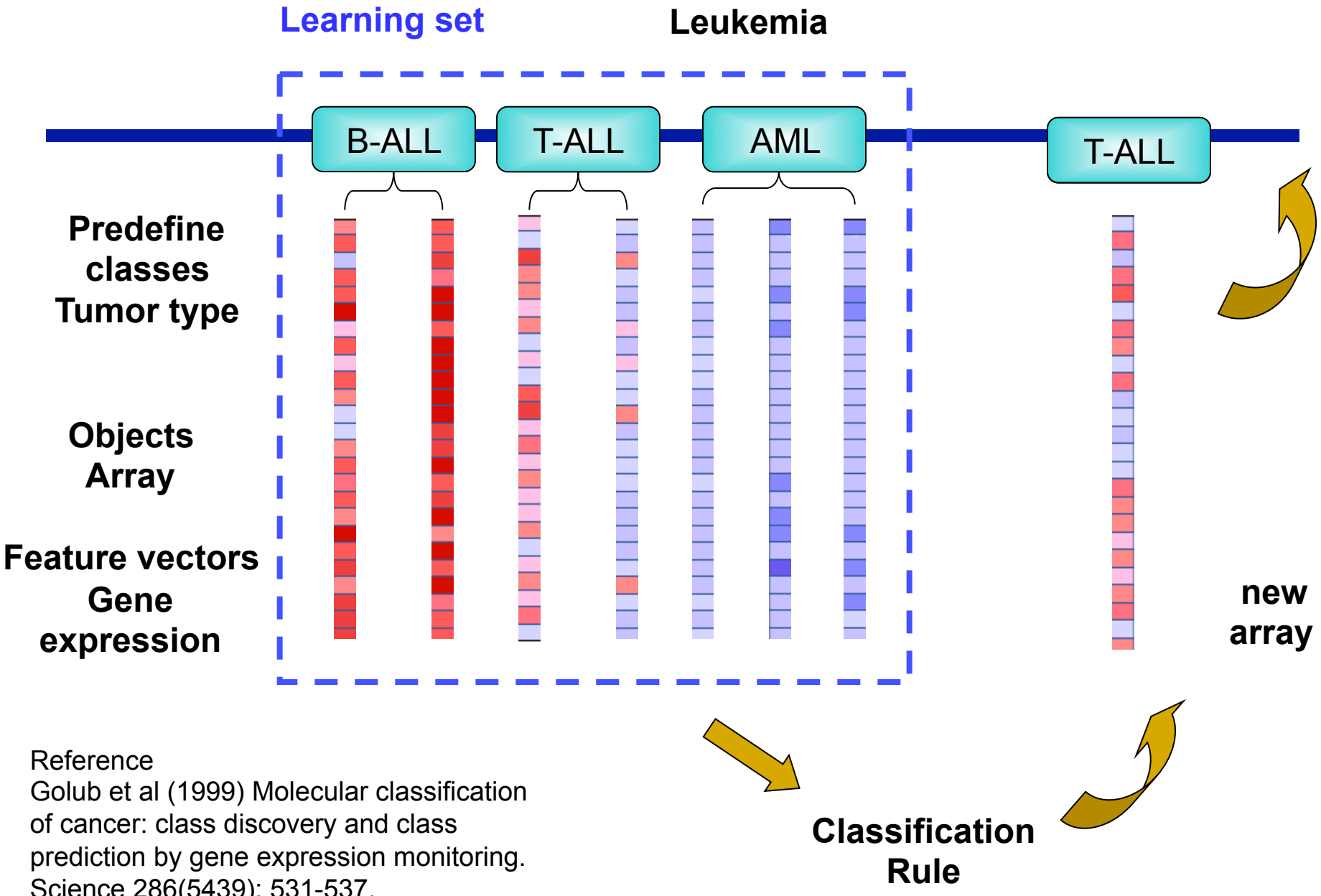
Discrimination and Allocation





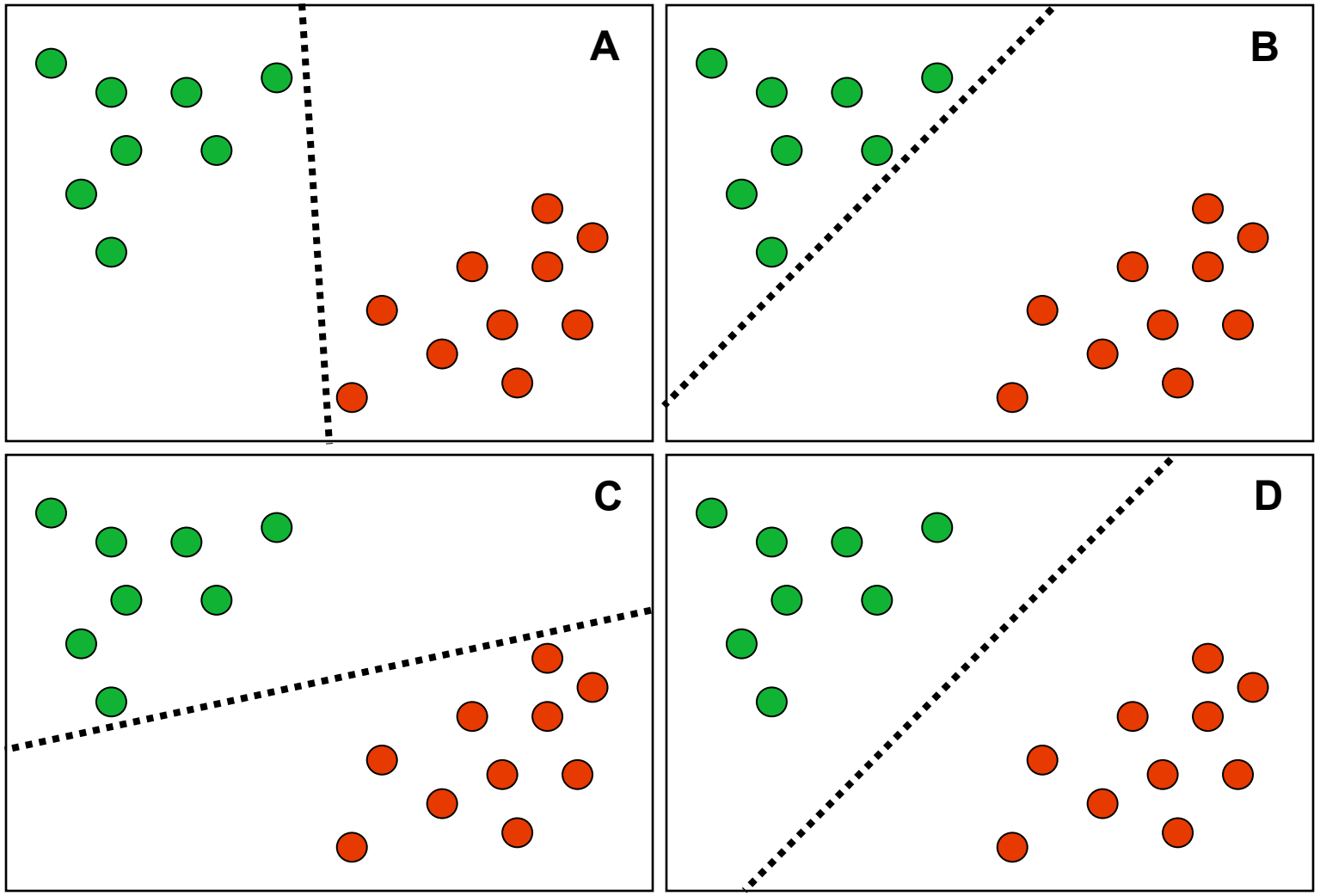
Reference
L van't Veer *et al (2002) Gene expression profiling predicts clinical outcome of breast cancer.* Nature, Jan.

.

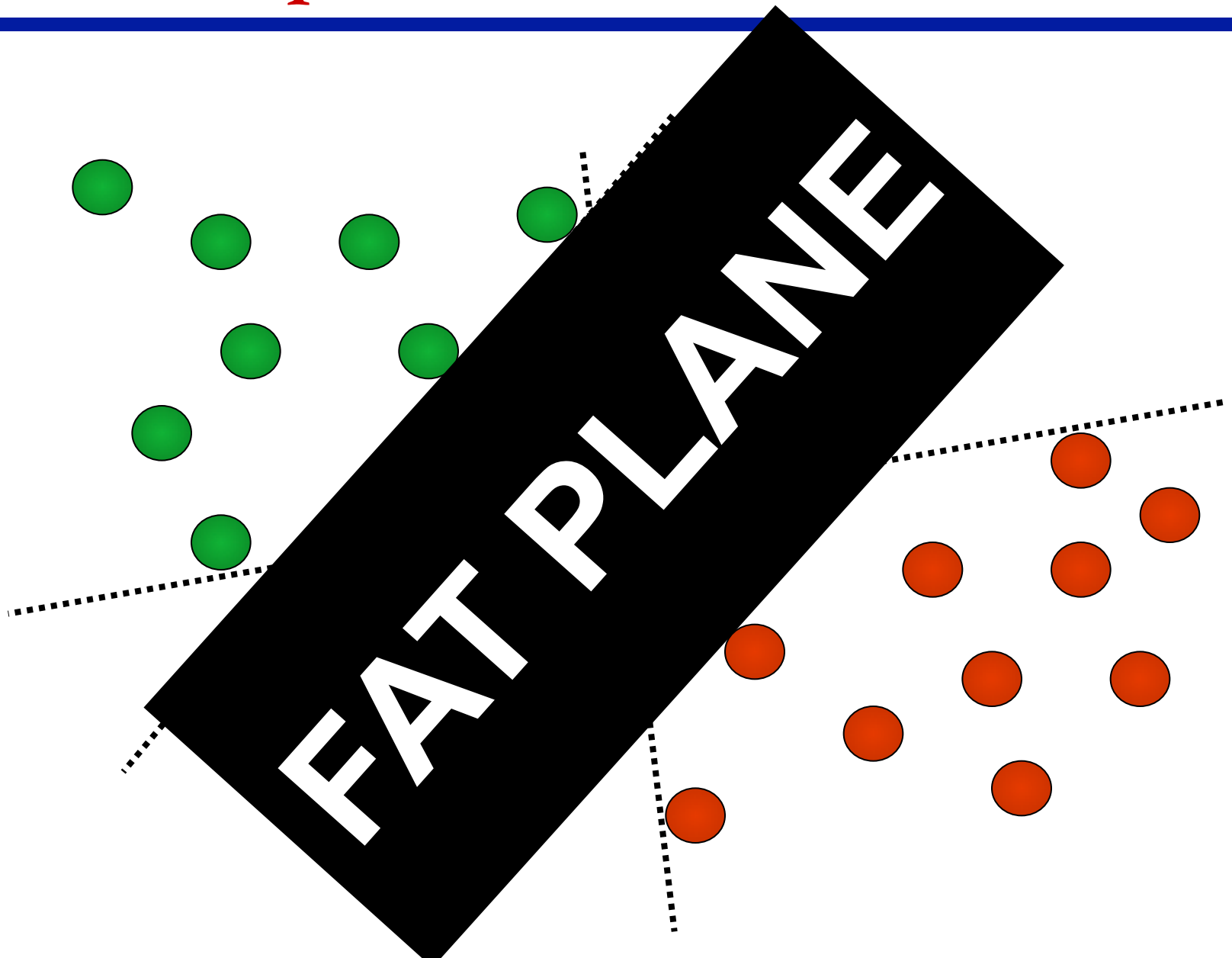


Reference
Golub et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531-537.

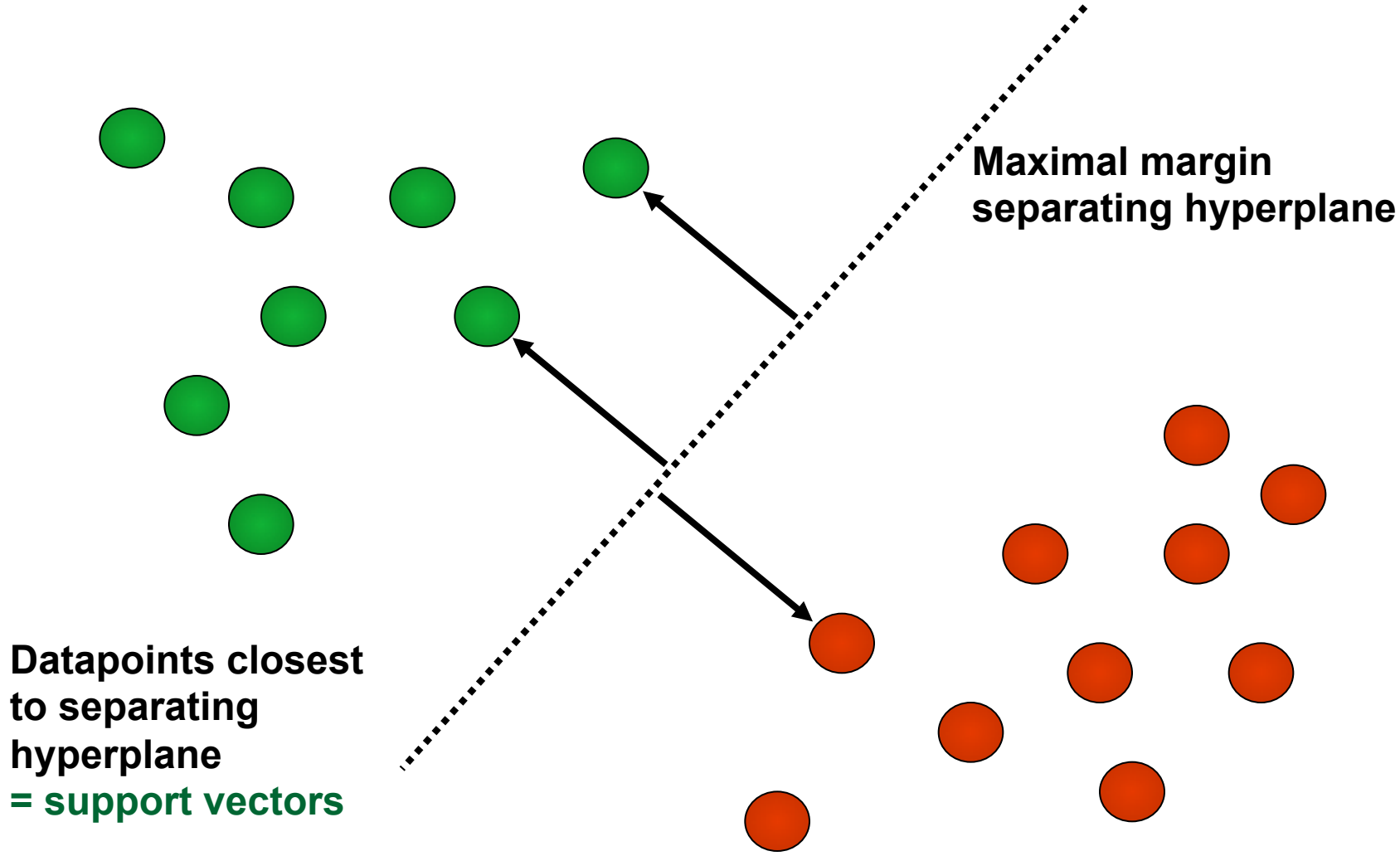
Classification: Which line separates best?



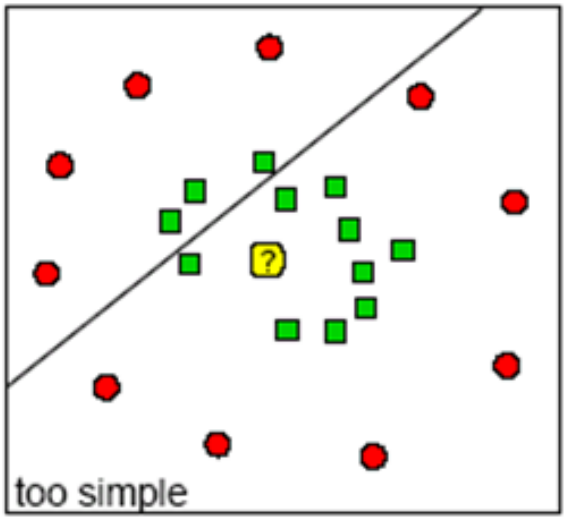
No sharp knife, but a …

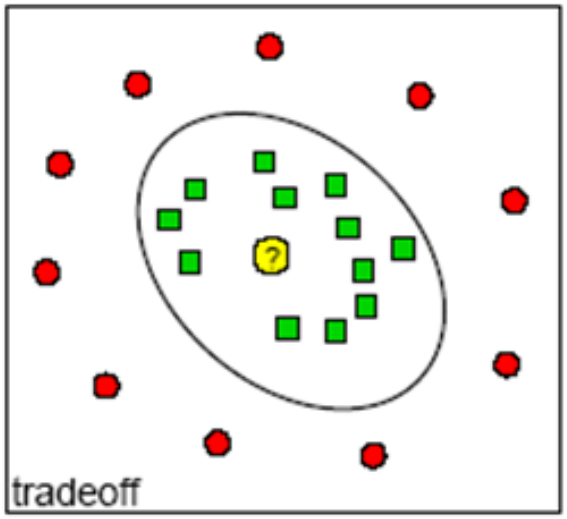


Support Vector Machines



Underfitting and Overfitting





negative example
positive example
new patient

Examples of classification algorithms

- Linear classifiers
  - Fisher's linear discriminant
  - Logistic regression
  - Naive Bayes classifier
  - Support vector machines
- Quadratic classifiers
- k-nearest neighbor
- Boosting
- Decision trees
- Neural networks
- Bayesian networks
- Hidden Markov models

Gene select

In tumor “classification” using gene expression data:

- Identification of new/unknown tumor classes using gene expression profiles (unsupervised learning - clustering)
- Classification of malignancies into known classes (supervised learning - discrimination)

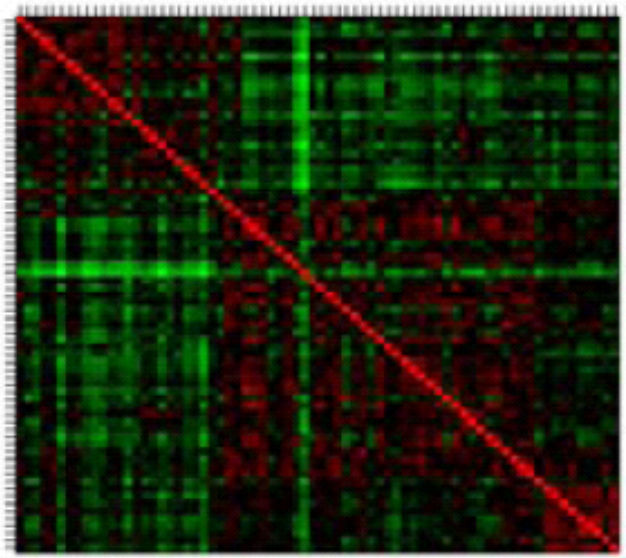
- Identify the “marker” genes that characterize the different tumor classes (feature or variable selection)

Why gene selection?

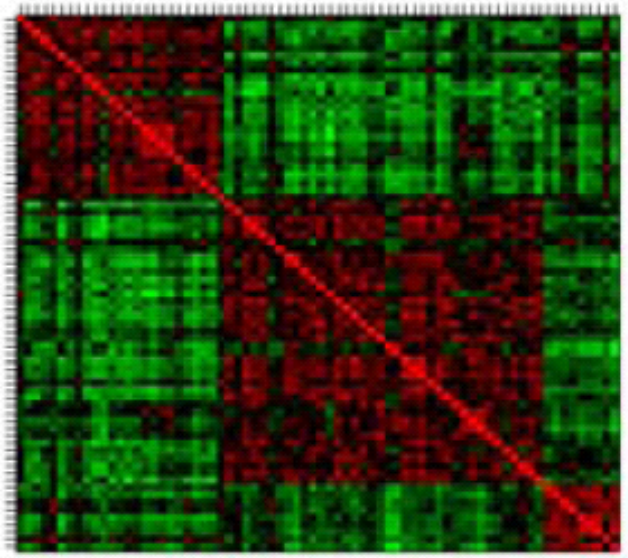
- Identify marker genes that characterize different tumor status.
- Many genes are redundant and will introduce noise that lower performance.
- Lead to better classification performance by removing variables that are noise with respect to the outcome
- Can eventually lead to a diagnosis chip. (“breast cancer chip”, “liver cancer chip”)

Gene selection

Why select features?



No feature selection



Top 100 feature selection
Selection based on variance

-1 +1

Correlation plot
Data: Leukemia, 3 class

Gene selection

Methods fall into three categories:

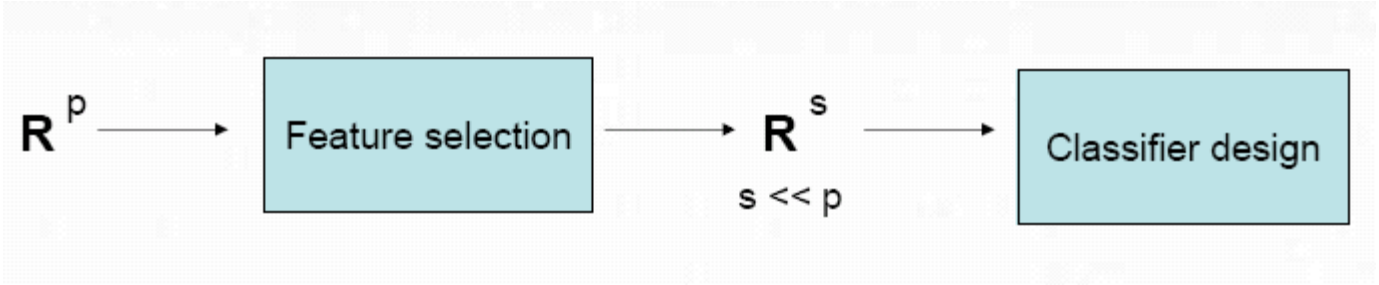
1. Filter methods
2. Wrapper methods
3. Embedded methods

Filter methods are simplest and most frequently used in the literature.

Features (genes) should be selected only from the training set used to build the model (and not the entire set)

Gene selection

Filter method:



- Features (genes) are scored according to the evidence of predictive power and then are ranked. Top *s* genes with high score are selected and used by the classifier.
- Scores: t-statistics, F-statistics, signal-noise ratio, …
- The # of features selected, *s*, is then determined by cross validation.

Advantage: Fast and easy to interpret.

How Good is the Classifier?

The Rule

We *must* test our classifier on a different set from the training set: the labeled test set

The Task

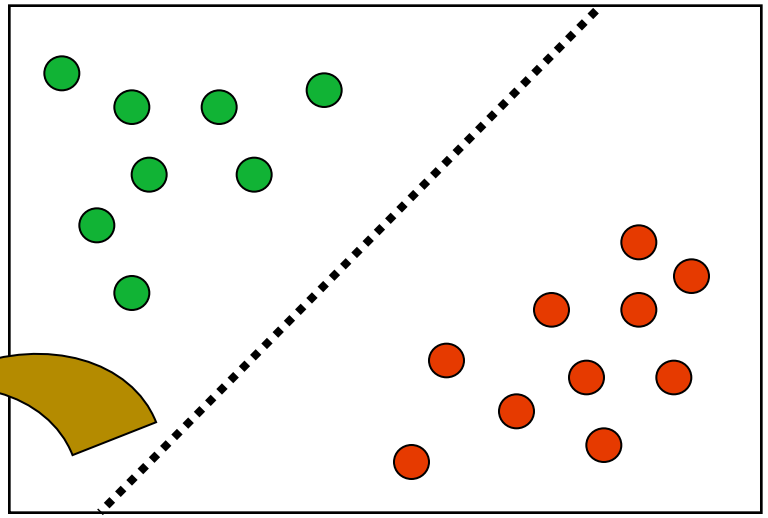
Calculate misclassification error rate on the test set: proportion of count of misclassified objects.

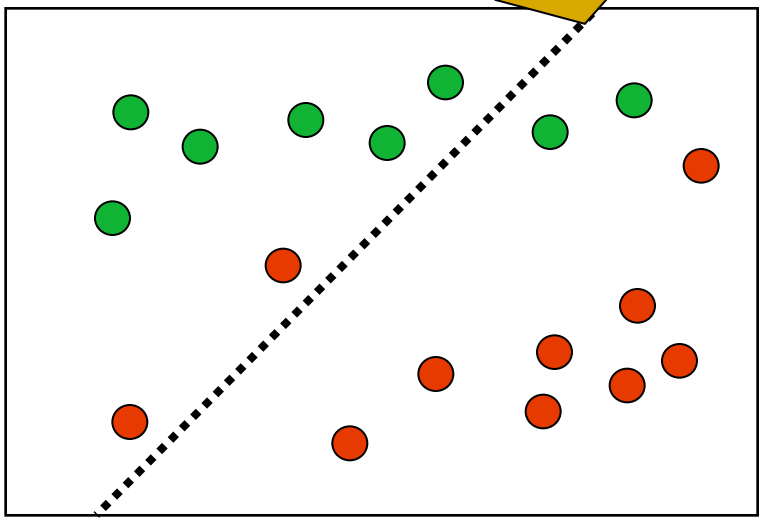
How Good is the Classifier?

Training error: how well do we do on the data we trained the classifier on?

But how well will we do in the future, on new data?

Test error: How well does the classifier generalize?





Same classifier (= line)

New data from same classes

The classifier will usually perform worse than before:

Test error > training error

- If it is binary classes (i.e., true/false, normal/ disease), we can even classify each object in the test set and count the number of each type of error. – then use sensitivity & specificity – ROC curve

Binary Classification Errors

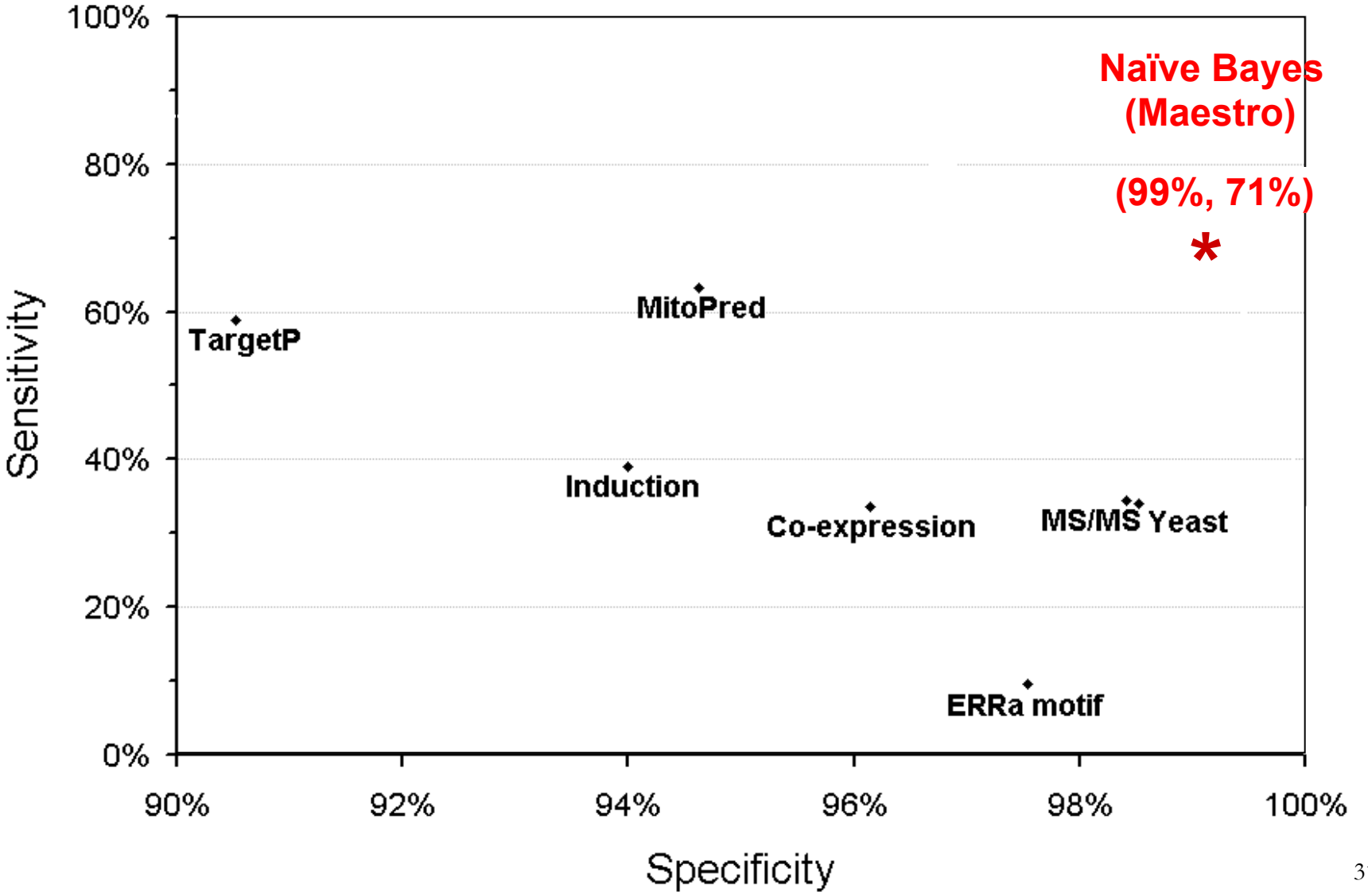
	True	False
Predicted True	TP	FP
Predicted False	FN	TN

Sensitivity = TP/(TP+FN) Specificity = TN/(TN+FP)

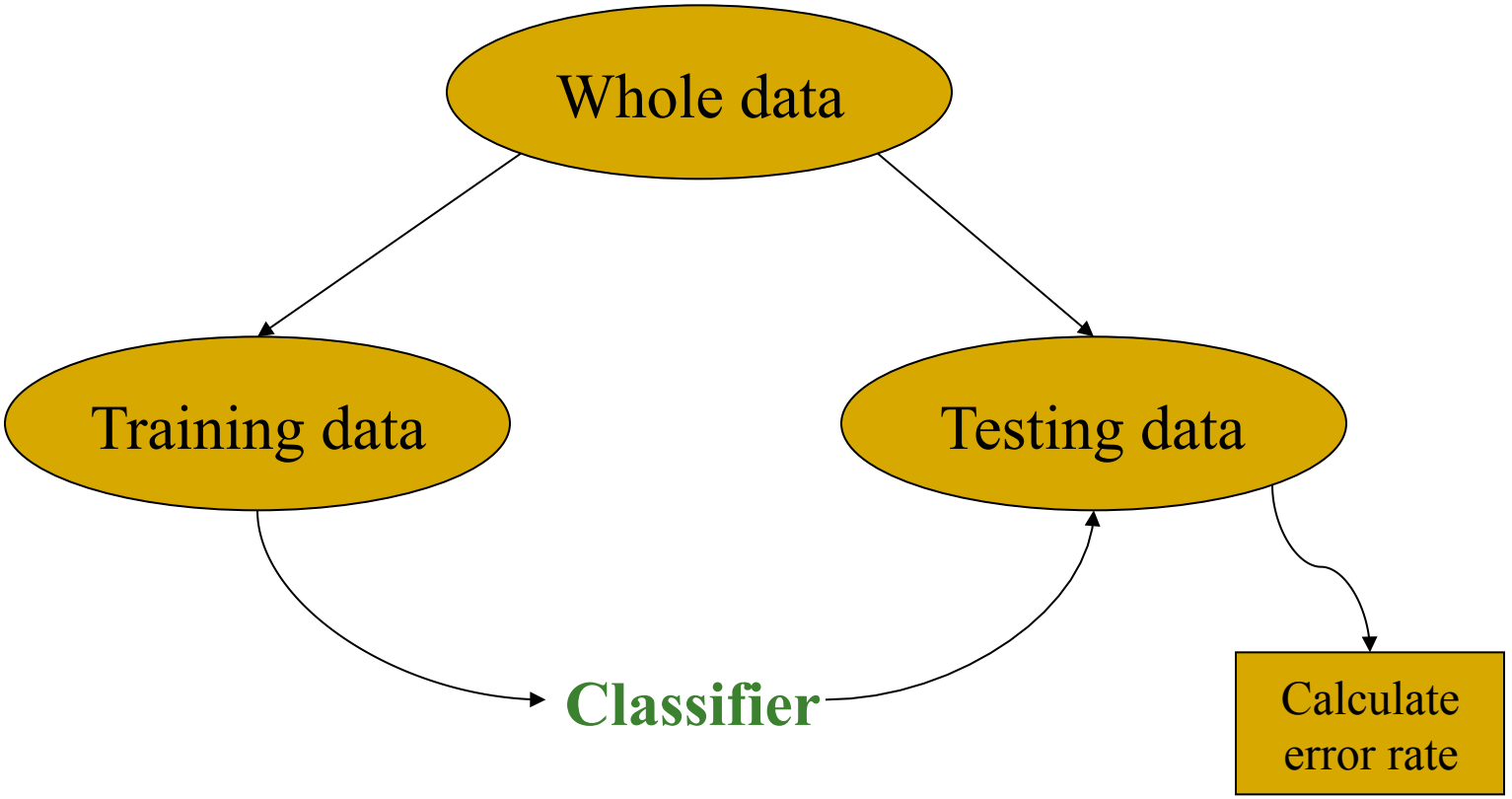
- Sensitivity
  - Fraction of all Class1 (True) that we correctly predicted at Class 1
  - *How good are we at finding what we are looking for*
- Specificity
  - Fraction of all Class 2 (False) called Class 2
  - *How many of the Class 2 do we filter out of our Class 1 predictions*

In both cases, the higher the better

Maestro Outperforms Existing Classifiers



If there is no test data available, use Cross Validation technique



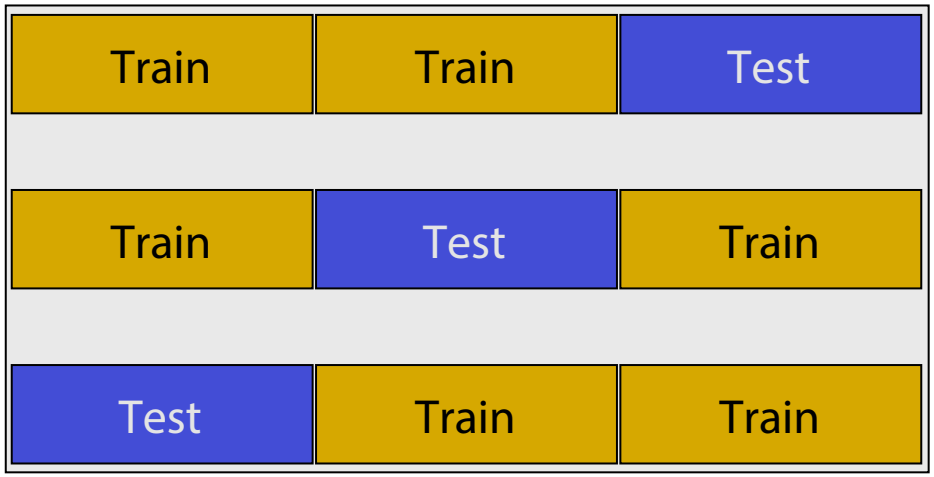
Cross-validation

Training error

Test error

V-fold Cross-validation

Here for V=3



Cross-validation

- V-fold cross validation:

  Cases in learning set randomly divided into V subsets of (nearly) equal size. Build classifiers by leaving one set out; compute test set error rates on the left out set and averaged.

  10-fold cross validation is popular in the literature.

- Leave-one-out cross validation

  Special case: V=n.

Packages needed for classification analysis

```
####  install the required packages for the first time ###
source("http://www.bioconductor.org/biocLite.R")
biocLite("ALL")
biocLite("genefilter")
biocLite("hgu95av2.db")
biocLite("MLInterfaces")

install.packages("gplots")
install.packages("e1071")
```

Then …

```
#####  load the packages ###
library("ALL")   ## or without quotes
library("genefilter")
library("hgu95av2.db")
library("MLInterfaces")
library("gplots")
library("e1071")
```