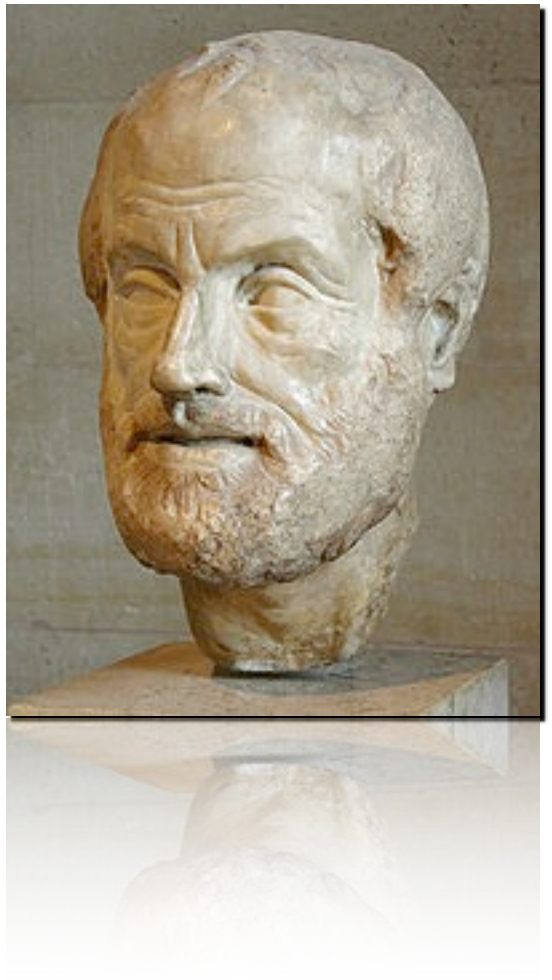
# Classification

# Classification

- Aristotle was one of the first to use classification in 300 B.C.
- His organism classification was based on color of the organism's blood.
- Later he organized them by physical characteristics.



# Classification

Statistical classification is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters, etc) and based on a training set of previously labeled items.

# Classification

Models of data with a categorical response are called classifiers. A classifier is built from ***training data***, for which classifications are known. The classifier assigns new test data to one of the categorical levels of the response.

Parametric methods, like Discriminant Analysis, fit a parametric model to the training data and interpolate to classify test data. Nonparametric methods, like Classification Trees, use other means to determine classifications. In this sense, classification methods are analogous to Nonlinear Regression.

# Classification

Statistical classification algorithms are typically used in ***pattern recognition*** systems.

Sometimes the term “classification” is synonymous with what is commonly known in machine learning as ***clustering***.

# Examples of classification algorithms include:

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

# Variance

The variance of a sample is one measure of statistical ***dispersion***, averaging the squared distance of its possible values from the expected value (mean). Whereas the mean is a way to describe the location of a distribution, the variance is a way to capture its scale or ***degree of being spread out***. The unit of variance is the square of the unit of the original variable. The positive square root of the variance, called the standard deviation, has the same units as the original variable and can be easier to interpret for this reason.

If random variable X has expected value (mean) μ = E(X), then the variance Var(X) of X is given by:

$$Var(x) = E[(X - \mu)^2]$$

# Covariance Matrix

The covariance matrix is a matrix of covariances between elements of a vector. It is the natural generalization to higher dimensions of the concept of the variance of a scalar-valued variable.

If entries in the column vector $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$

are variables, each with finite variance, then the covariance matrix Σ is the matrix whose *(i, j)* entry is the covariance

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

# Measures of Similarity

- It is useful to be able to construct 'similarity' measures. These include:
  - **The Euclidean distance**, the "ordinary" distance between two points that one would measure with a ruler (can calculate using the Pythagorean theorem).
  - **The Mahalanobis distance**, a distance measure introduced by P. C. Mahalanobis in 1936. It is based on correlations between variables by which different patterns can be identified and analyzed.

# The Euclidean Distance

The Euclidean distance between points *P* $(p_1, p_2, \dots p_n)$ and *Q* $(q_1, q_2, \dots q_n)$ is:

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

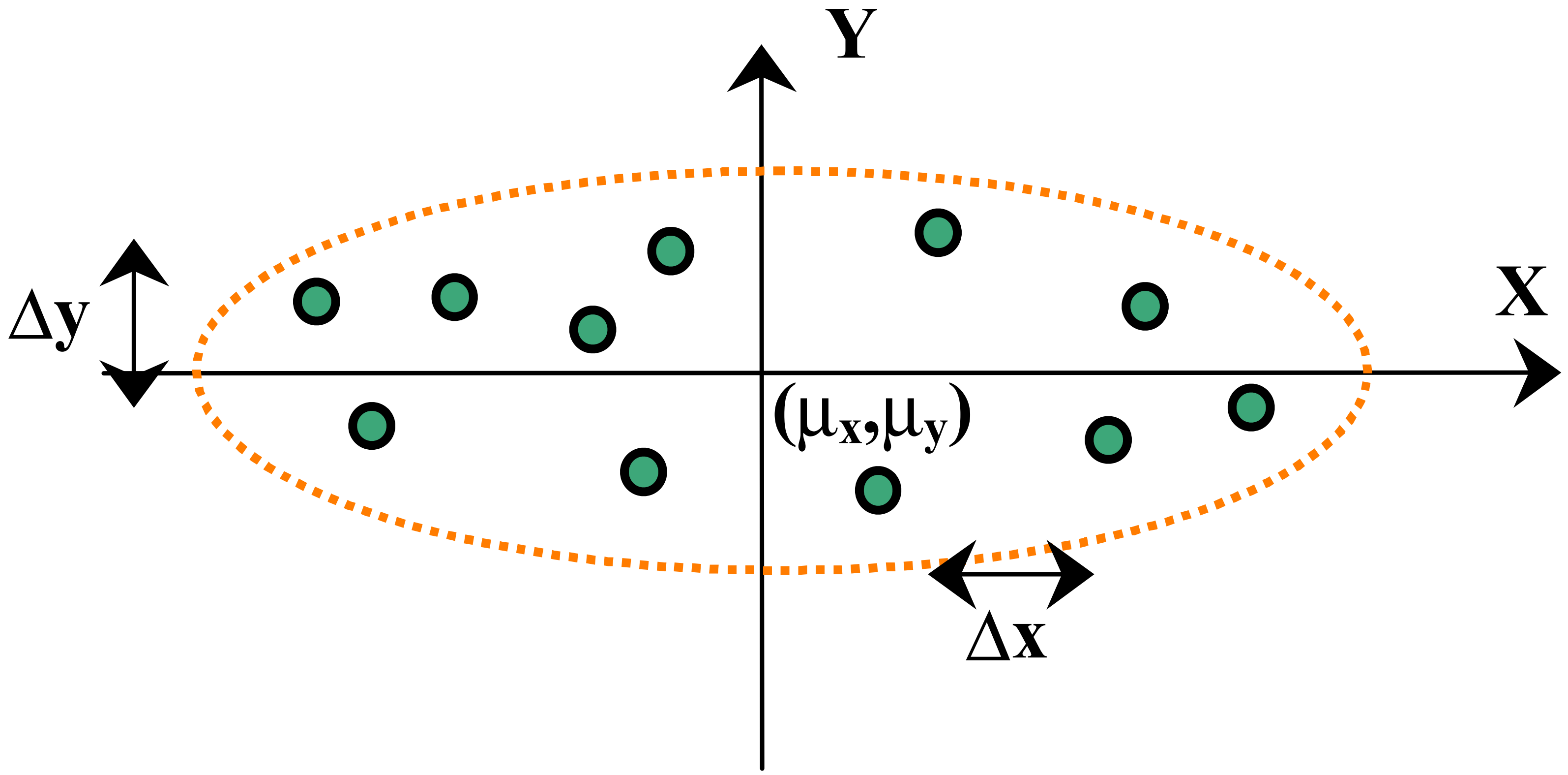
# The Mahalanobis distance

Formally, the Mahalanobis distance from a group of values with mean **µ** and covariance matrix **S** for a multivariate vector ***x*** is defined as:

$$D_m(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

# Mahalanobis Distance

Euclidian distance weights all dimensions (variables) equally, however, statistically they may not be the same:



Euclidian distance $\Delta x = \Delta y$

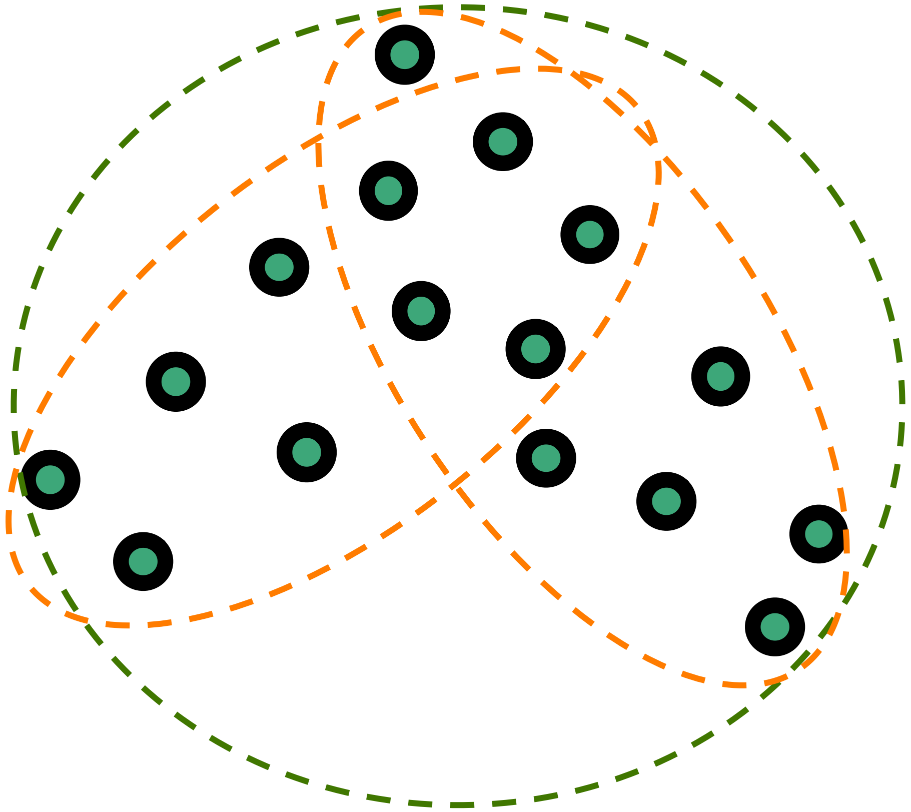
However statistically $\Delta x < \Delta y$

# Over fitting

If our classifier is trained on a small data set it is inadvisable to build too complex a classification boundary

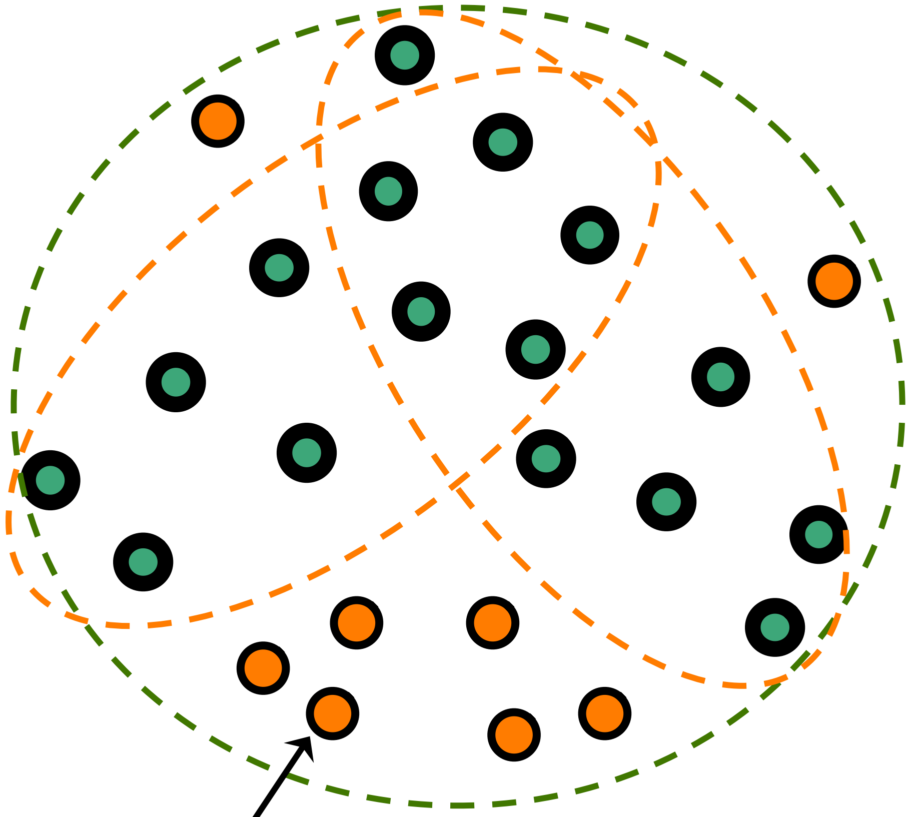
# Over fitting

If our classifier is trained on a small data set it is inadvisable to build too complex a classification boundary



# Over fitting

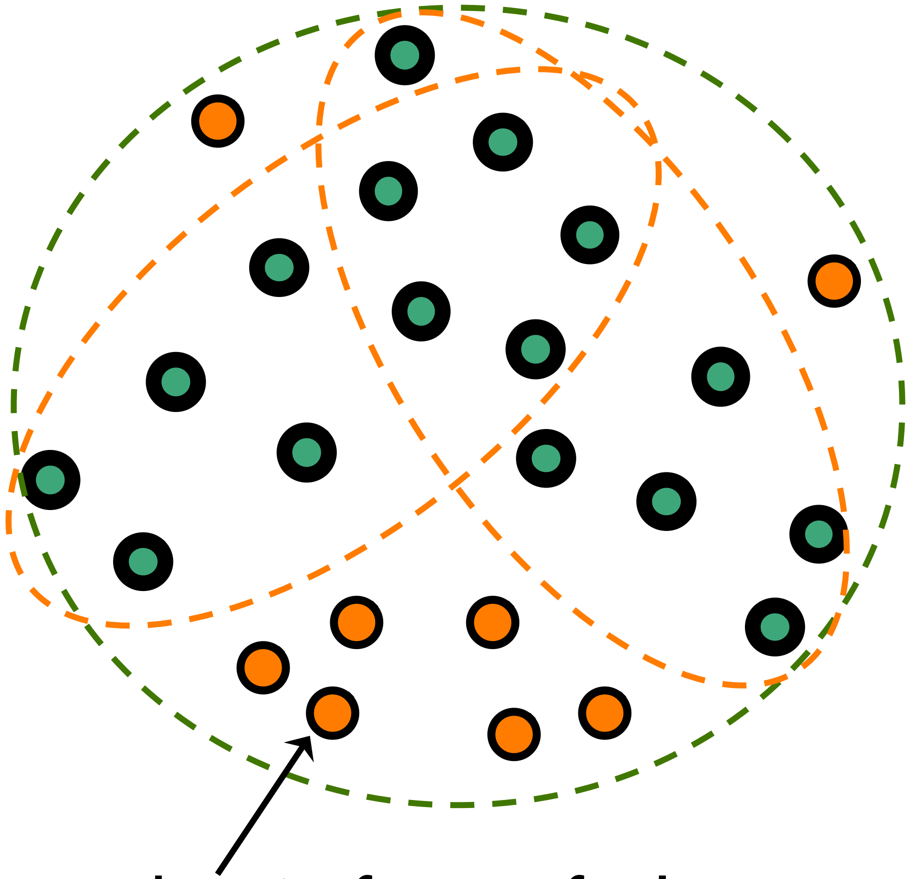
If our classifier is trained on a small data set it is inadvisable to build too complex a classification boundary



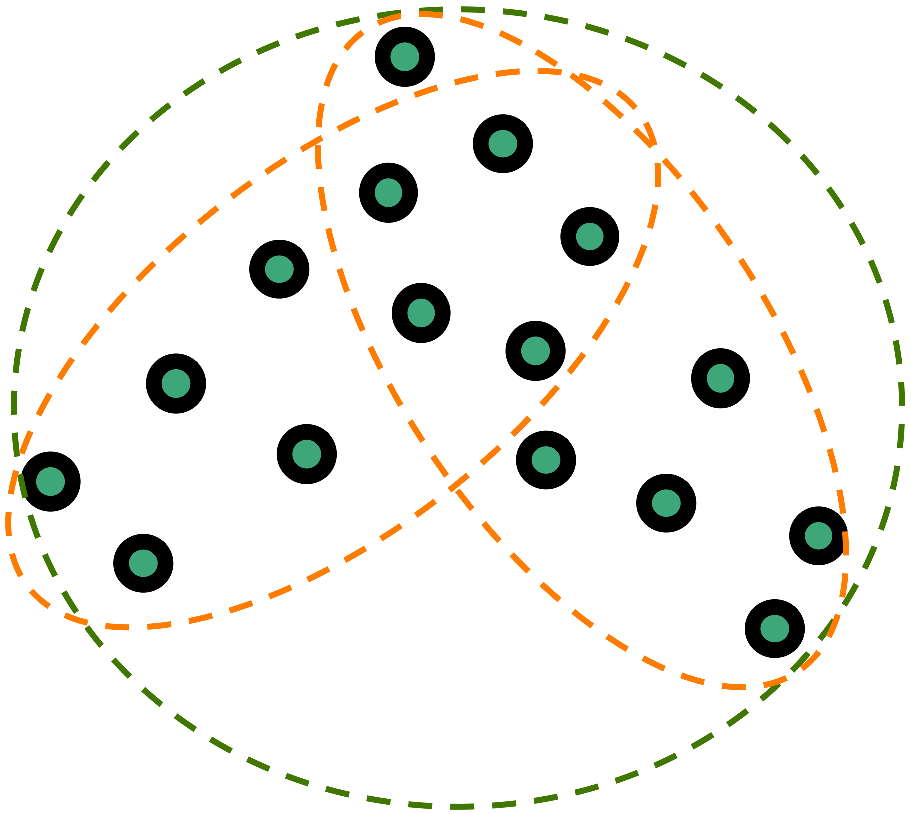
More data in favor of a loose boundary implies overfitting

# Over fitting

If our classifier is trained on a small data set it is inadvisable to build too complex a classification boundary

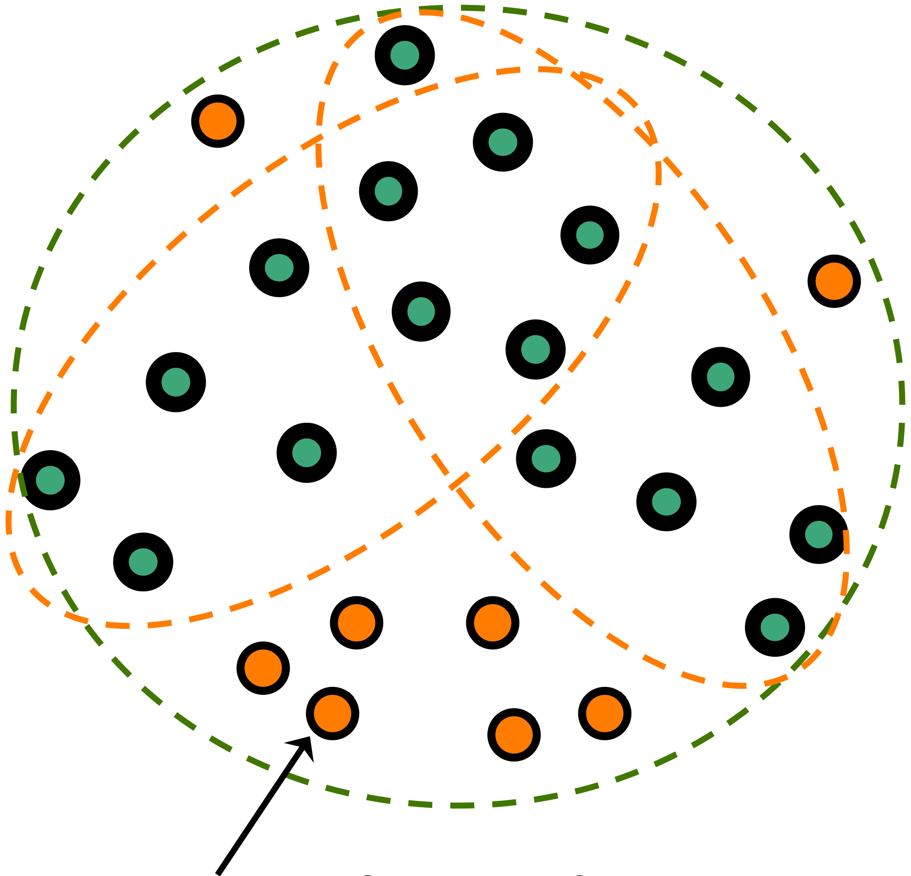


More data in favor of a loose boundary implies overfitting

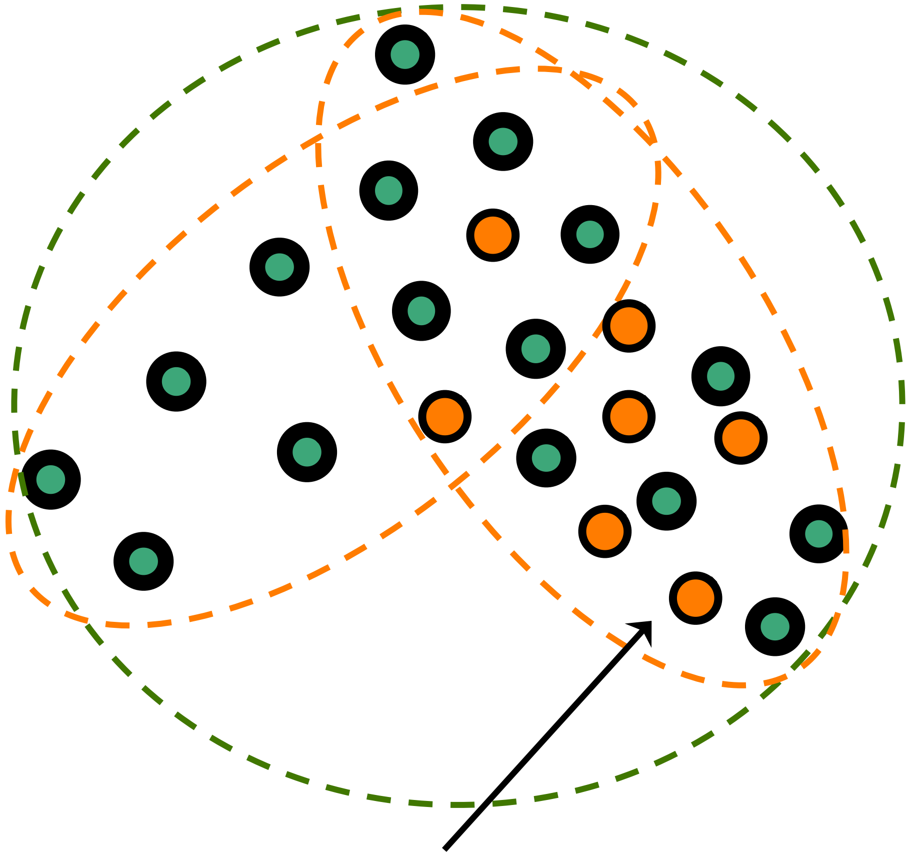


# Over fitting

If our classifier is trained on a small data set it is inadvisable to build too complex a classification boundary



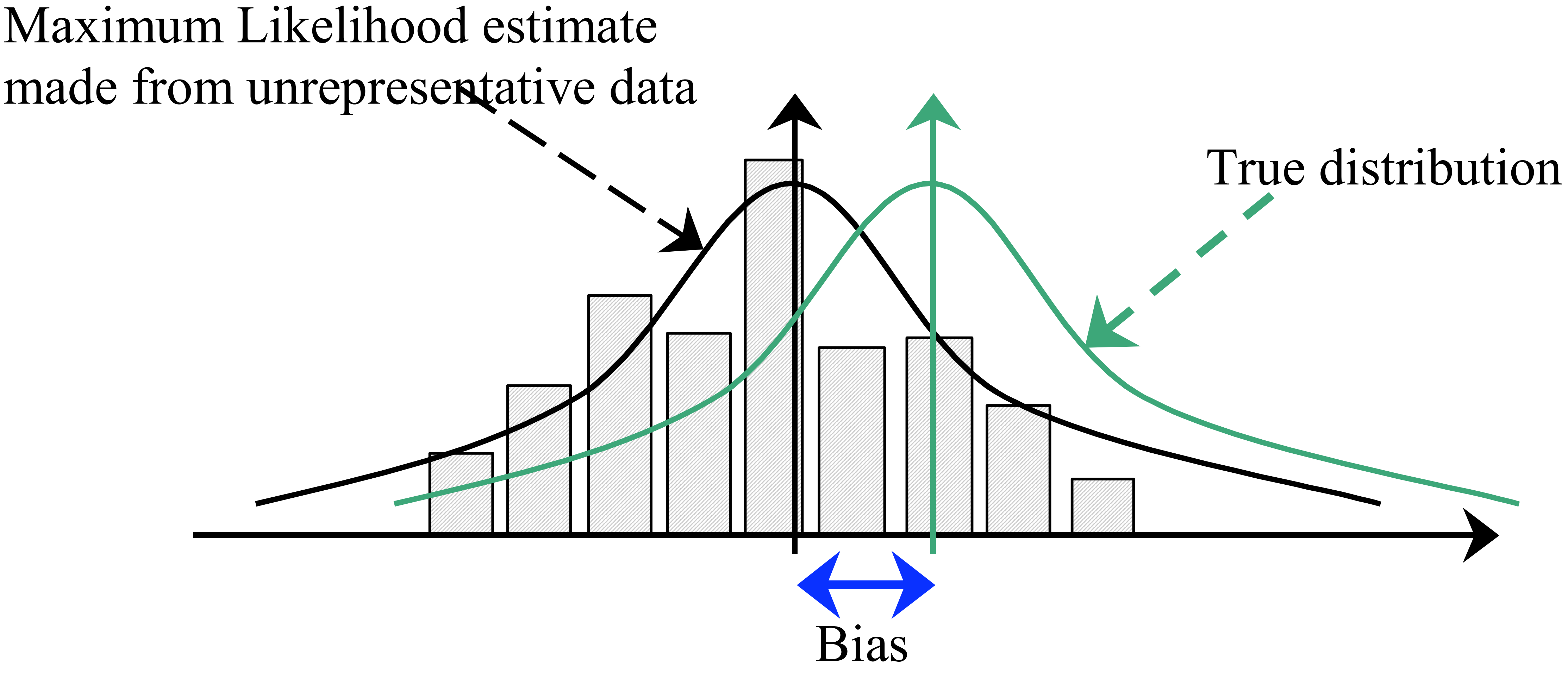
More data in favor of a loose boundary implies overfitting



More data could favor the mixture

# Bias

- Parameters of a distribution are normally computed from a data set.
- The difference between a true mean and an estimated mean is termed bias.



# Trade off between bias and variance

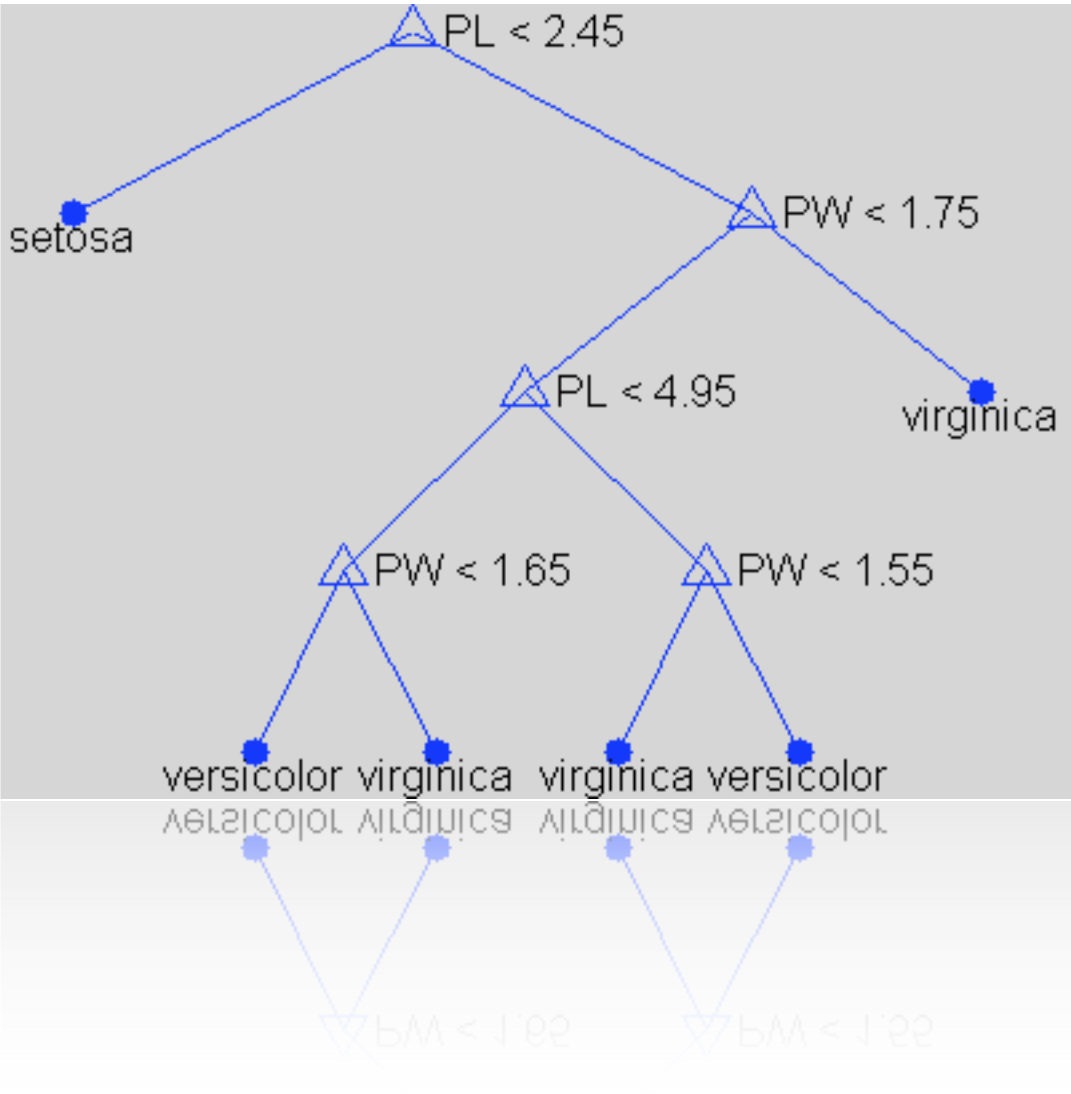
- Statistically it turns out that there is a trade off between bias and variance:
  - Low order models (eg linear) tend to be biased
  - High order models have greater variance.
- For large training sets variance tends to decrease for a fixed bias, hence higher order models may be more appropriate.

# Decision Trees

- Parametric models specify the form of the relationship between predictors and a response. In many cases, however, the form of the relationship is unknown, and a parametric model requires assumptions and simplifications. Regression Decision Trees offer a nonparametric alternative.
- Decision tree (also known as classification tree) methods are a good choice when the data-mining task is classification or prediction of outcomes and the goal is to ***generate rules that can be easily understood, explained, and translated into a natural query language***.

# Decision Trees

A decision tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. In creating a decision tree we determine a hierarchical set of rules that provides an efficient classification of the dataset.

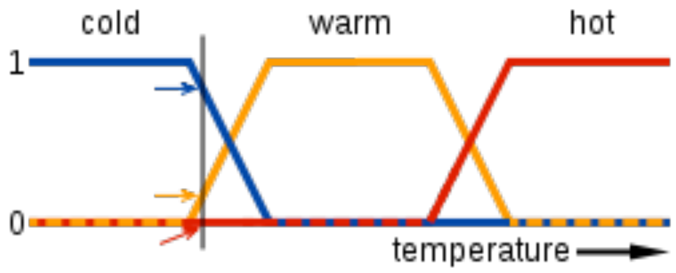


# Fuzzy Inference

The concept of Fuzzy Logic (FL) was conceived by Zadeh as a way of processing data by allowing partial set membership rather than crisp set membership or non-membership (Zadeh, 1965), hence the term "fuzzy".

The point of fuzzy logic is to ***map an input space to an output space***, and the primary mechanism for doing this is ***a list of if-then statements called rules***, as opposed to modeling a system mathematically.

***All rules are evaluated in parallel, and the order of the rules is unimportant.*** The rules themselves are useful because they refer to variables and the adjectives that describe those variables.



e.g. instead of saying the temperature is 20℃, we say it is warm

# Fuzzy Logic

- Zadeh reasoned that people do not require precise, numerical information input, and yet they are capable of highly adaptive control. Fuzzy Logic provides a simple way to arrive at a definite conclusion ***based upon vague, ambiguous, imprecise, noisy, or missing input information***.
- Fuzzy Logic is almost synonymous with the theory of fuzzy sets, a theory that relates to classes of objects with unsharp boundaries in which membership is ***a matter of degree***.

# Fuzzy Logic

- Fuzzy Logic was conceived as a better method for sorting and handling data but has proven to be an excellent choice for many control system applications since it mimics human control logic.
- It can be built into anything from small, hand-held products to large computerized process control systems. It uses an imprecise but very descriptive language to deal with input data more like a human operator. It is very robust and forgiving of operator and data input and often works when first implemented with little or no tuning.

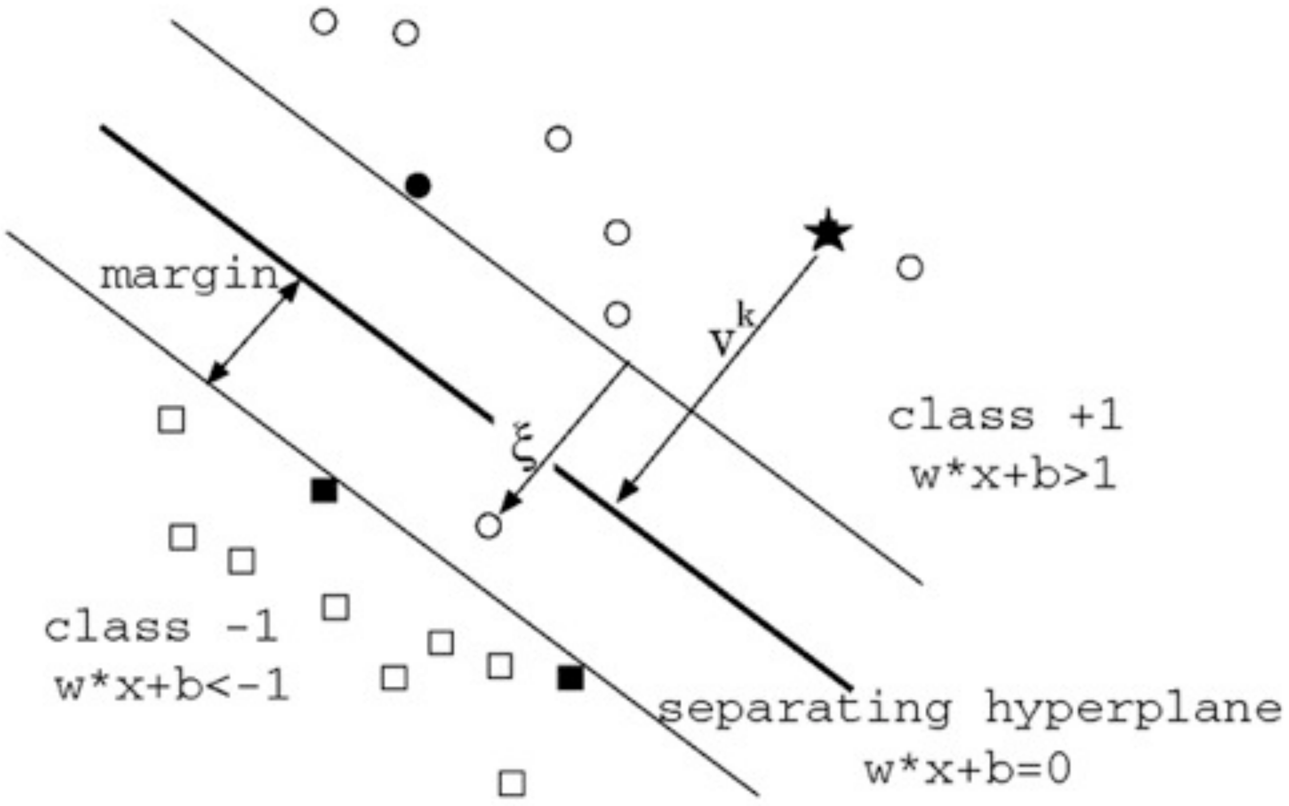
# Support Vector Machines

Support Vector Machines (SVM) are based on the concept of ***decision planes*** that define ***decision boundaries*** and were first introduced by Vapnik (Vapnik, 1995, 1998, 2000) and has subsequently been extended by others (Scholkopf et al., 2000, Smola and Scholkopf, 2004).

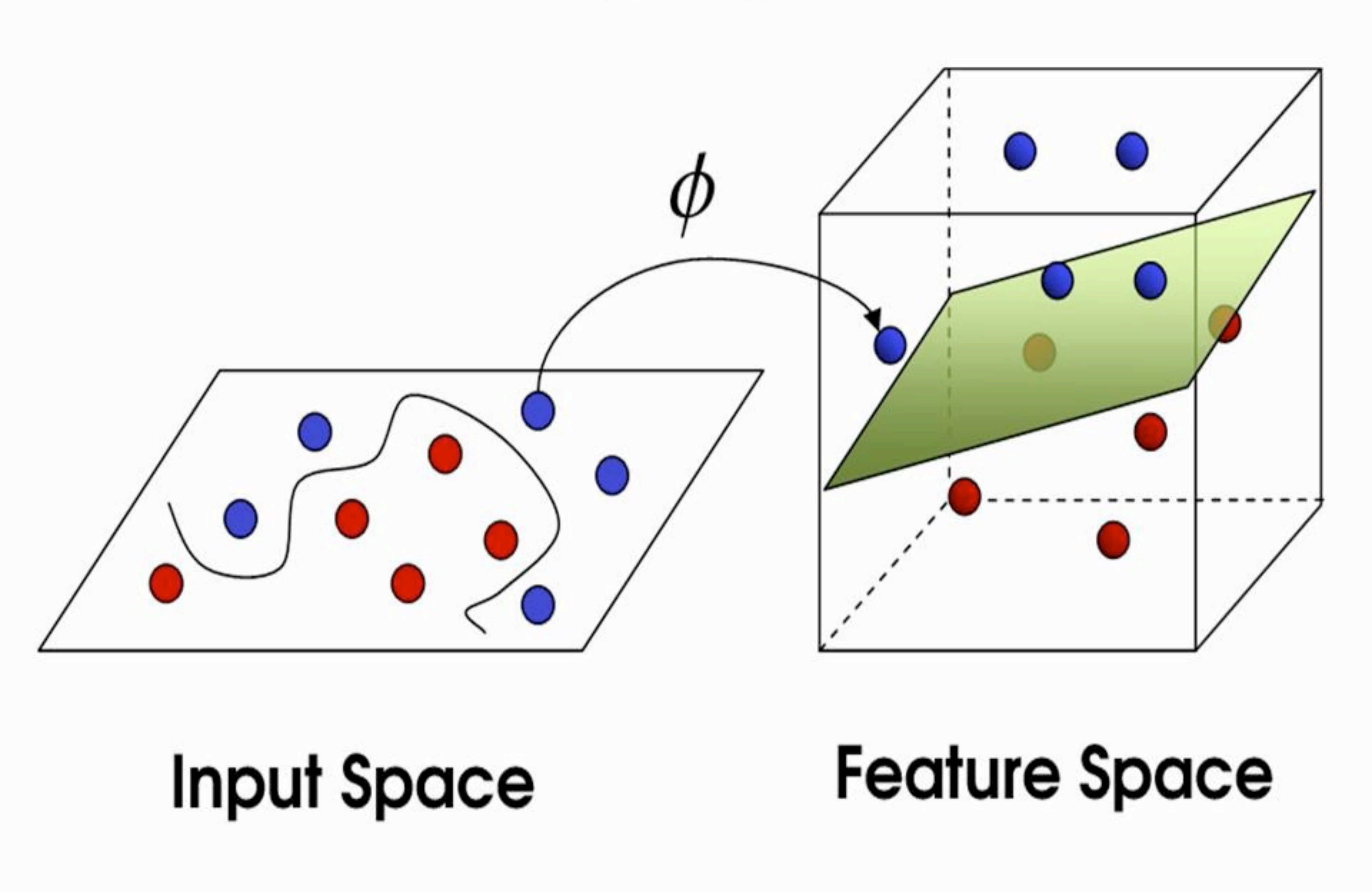
# Support Vector Machines

A decision plane is one that separates between a set of objects having different class memberships.

The simplest example is a linear classifier, i.e. a classifier that separates a set of objects into their respective groups with a line.



# Principle of Support Vector Machines (SVM)



# Support Vector Machines

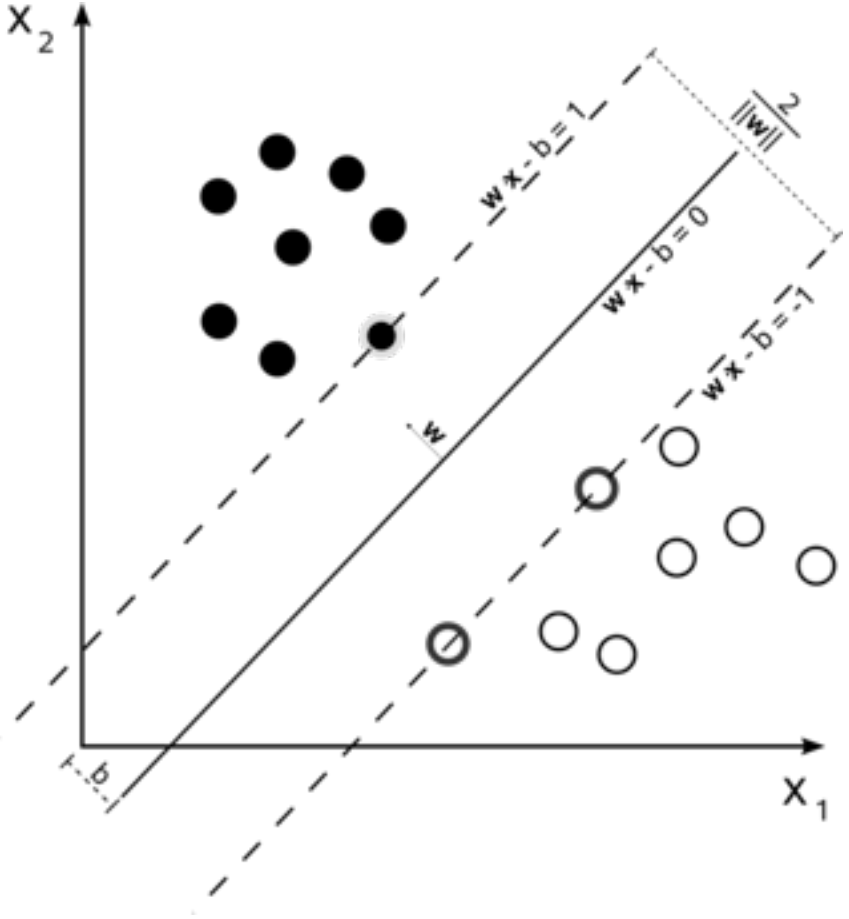
Support vector machines (SVMs) are a set of related ***supervised*** learning methods used for classification and regression.

Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a ***separating hyperplane*** in that space, one which ***maximizes the margin*** between the two data sets.

# Support Vector Machines

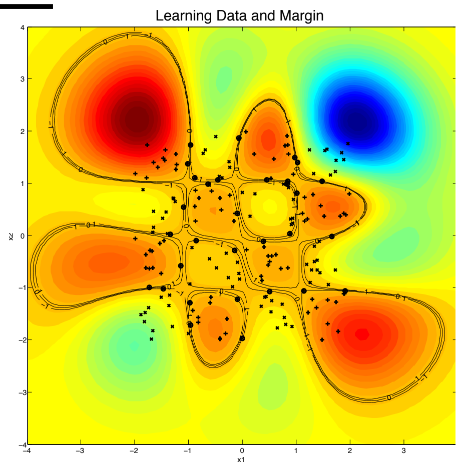
To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are “pushed up against” the two data sets.

Intuitively, a good separation is achieved by the hyperplane that has the ***largest distance to the neighboring data points*** of both classes, since in general ***the larger the margin the better the generalization error*** of the classifier.



# Examples

- In Matlab type and do the examples in:
  - `doc classify`
  - `doc treefit`
  - `doc mahal`
- Work through the examples in the Classification section of the statistics toolbox
- Get the source code and examples for SVM from http://asi.insa-rouen.fr/enseignants/~arakotom/toolbox/index.html. Run the example called exclass.m which illustrates a 2D SVM Classification.


Learning Data and Margin
x2
x1