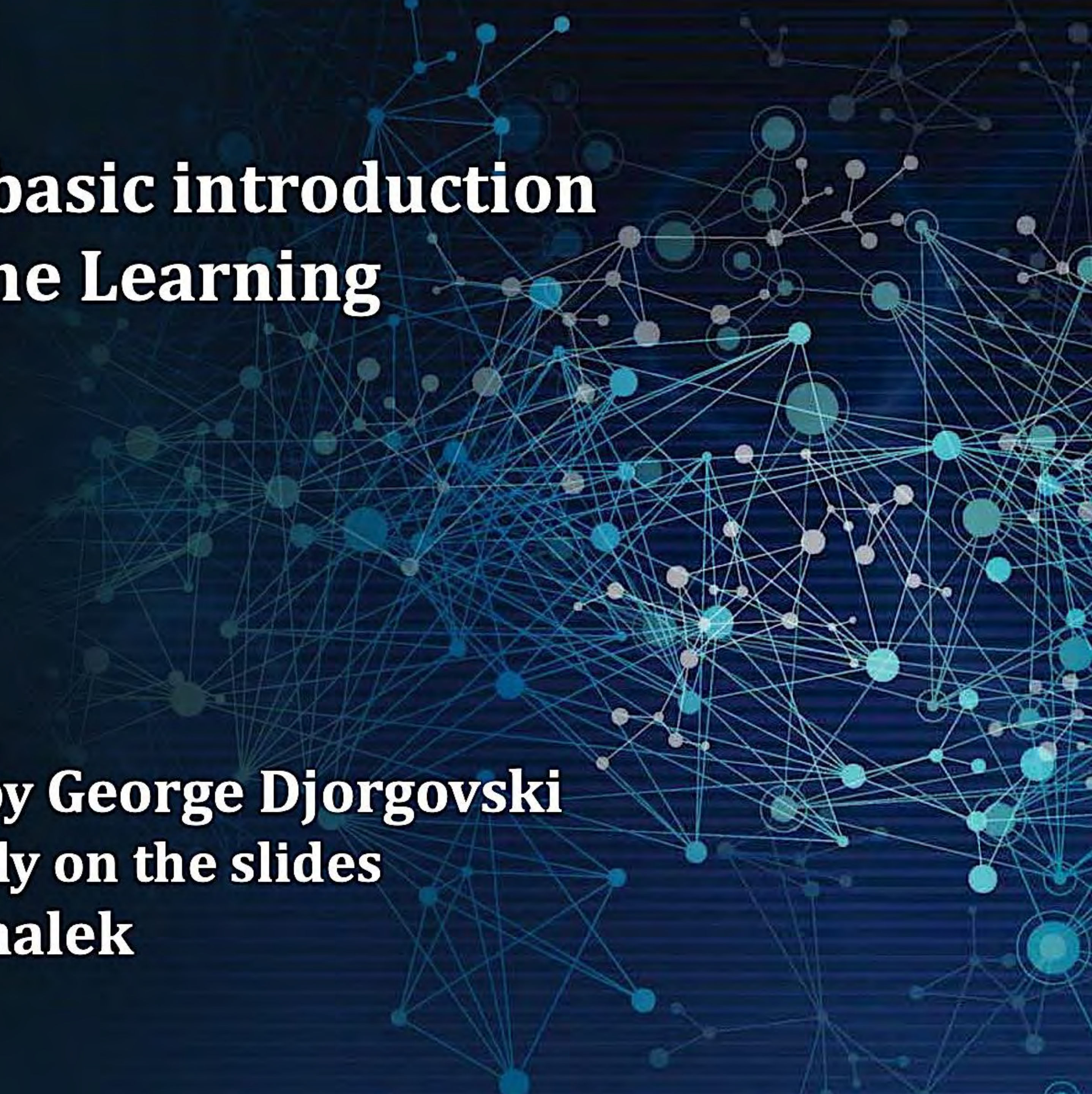
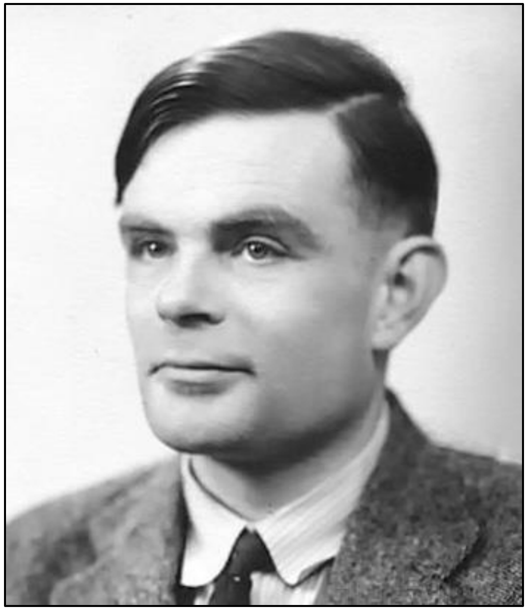
A (very) basic introduction to Machine Learning



Presented by George Djorgovski
based largely on the slides
by Ciro Donalek

How did AI Start

Alan Turing (1950),
"Computing Machinery
and Intelligence"



J. C. R. Licklider (1956),
" Man-Computer Symbiosis"

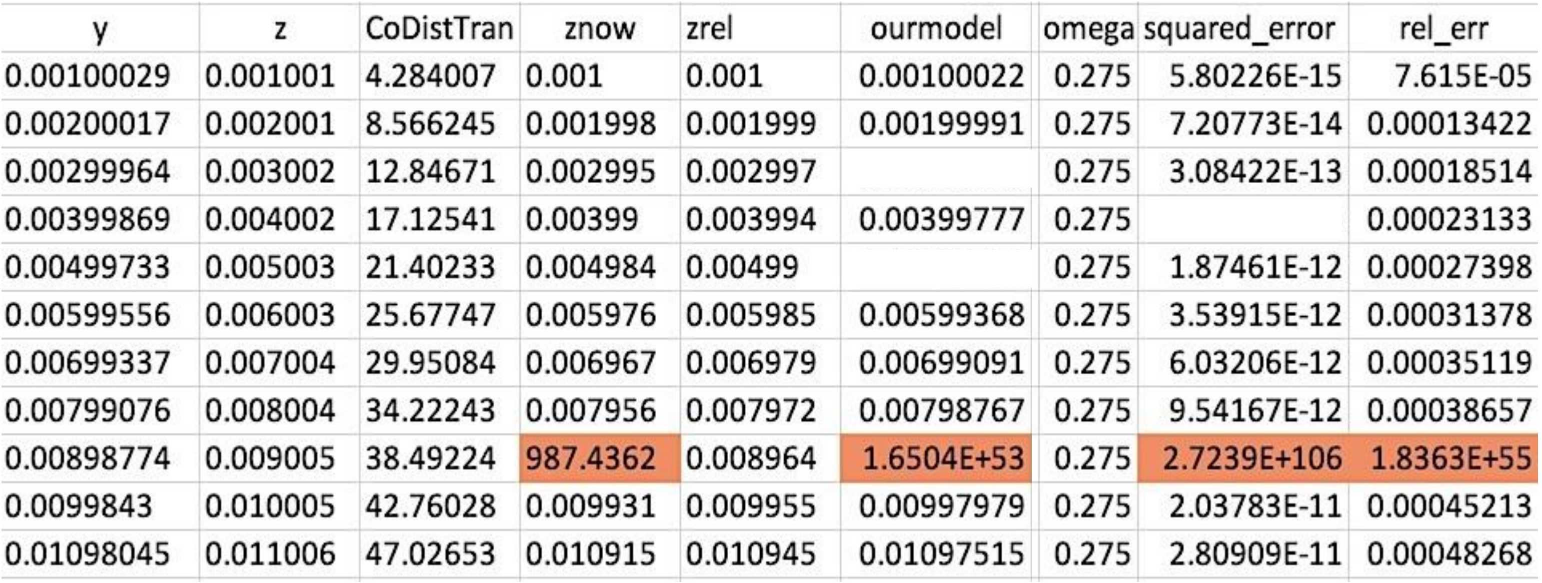
Dartmouth AI Workshop (1956)
Including Marvin Minsky, John Holland,
Claude Shannon, Herbert Simon, at al.

ML: historic definition

- Machine Learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed (Arthur Samuel – 1959).

Data Representation in Machine Learning

Imagine data represented as a large spreadsheet:



y	z	CoDistTran	znow	zrel	ourmodel	omega	squared_error	rel_err
0.00100029	0.001001	4.284007	0.001	0.001	0.00100022	0.275	5.80226E-15	7.615E-05
0.00200017	0.002001	8.566245	0.001998	0.001999	0.00199991	0.275	7.20773E-14	0.00013422
0.00299964	0.003002	12.84671	0.002995	0.002997		0.275	3.08422E-13	0.00018514
0.00399869	0.004002	17.12541	0.00399	0.003994	0.00399777	0.275		0.00023133
0.00499733	0.005003	21.40233	0.004984	0.00499		0.275	1.87461E-12	0.00027398
0.00599556	0.006003	25.67747	0.005976	0.005985	0.00599368	0.275	3.53915E-12	0.00031378
0.00699337	0.007004	29.95084	0.006967	0.006979	0.00699091	0.275	6.03206E-12	0.00035119
0.00799076	0.008004	34.22243	0.007956	0.007972	0.00798767	0.275	9.54167E-12	0.00038657
0.00898774	0.009005	38.49224	987.4362	0.008964	1.6504E+53	0.275	2.7239E+106	1.8363E+55
0.0099843	0.010005	42.76028	0.009931	0.009955	0.00997979	0.275	2.03783E-11	0.00045213
0.01098045	0.011006	47.02653	0.010915	0.010945	0.01097515	0.275	2.80909E-11	0.00048268

Some data may be *missing*

Some may be **wrong**

Some columns may contain no useful information ↑

Each row (***data or feature vector***) is one instance of what you are analyzing (galaxies, genes …). There could be tens … thousands … millions … billions of rows (***N***)

Each column is one of the quantities you are measuring – that is the ***data dimensionality***. There could be a few … tens … thousands … millions … billions of dimensions (***D***)

Computational Cost and Complexity

In general, computational cost of training and implementing NN and other ML models depends on the size of data sets and their dimensionality:

Computational cost = [data size] × [data dimensionality]

Scales with the number of input data vectors N, often as $N \log N$, which is fine

Scales with the **number of data dimensions *D***, usually as a power law, D^n, where $n >$ 2 or higher, or an exponential, e^D, which is *expensive*

For example, if $n = 2$, $D = 5$, cost ~ 25, if $D = 1000$, cost ~ a million

This is the "curse of *dimensionality*"

This is why ***dimensionality reduction*** is ***critical***

There are many methods, depending on the use case

Quick definitions

- **Supervised Learning**: for some of the samples, the desired output is known and it is used during the training process.
- **Unsupervised Learning**: the model is not provided with the correct results during the training; can be used to cluster the input data in classes on the basis of their statistical properties only.
- **Semi-Supervised Learning**: combines both labeled and unlabeled examples to generate an appropriate function or classifier.

Algorithms

- There are many good tools out there, but you need to choose the right ones for your needs.
- No "one size fits all" solution.

Supervised Algorithms

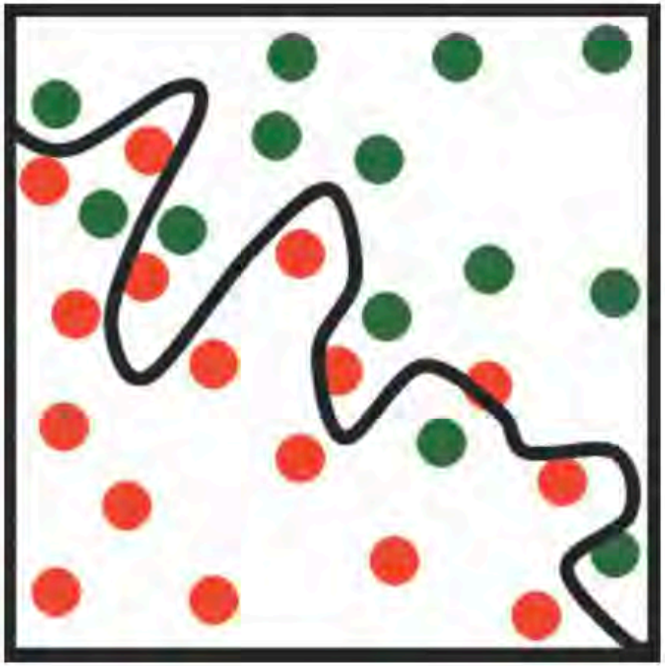
Neural Networks (MLP)
Boltzmann Machines
RBM
Decision Trees
Nearest Neighbor
Naive Bayes Classifiers
Bayesian Networks
Gaussian Processes
Regression
...

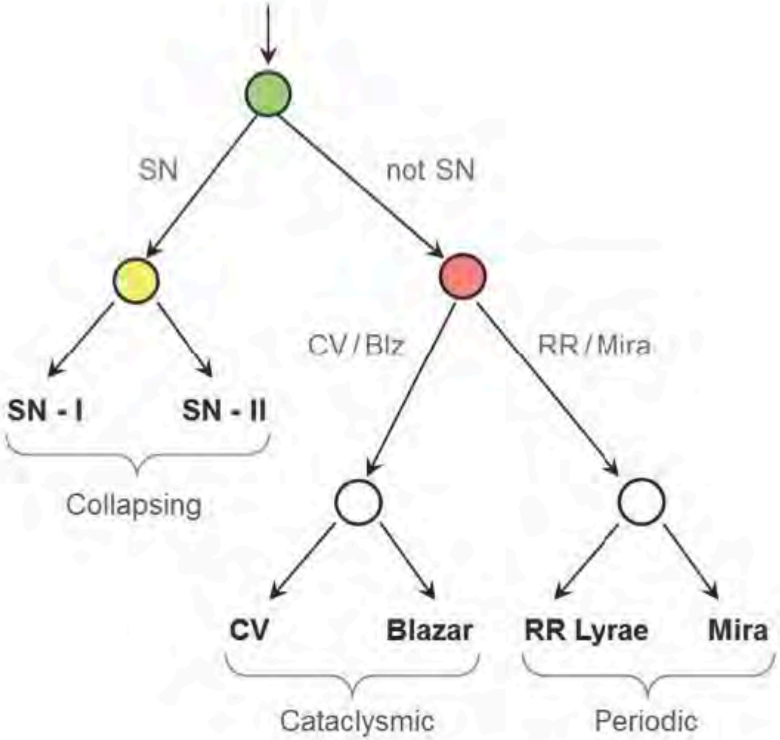
Unsupervised Algorithms

K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal Surfaces
...

DM tasks: Classification

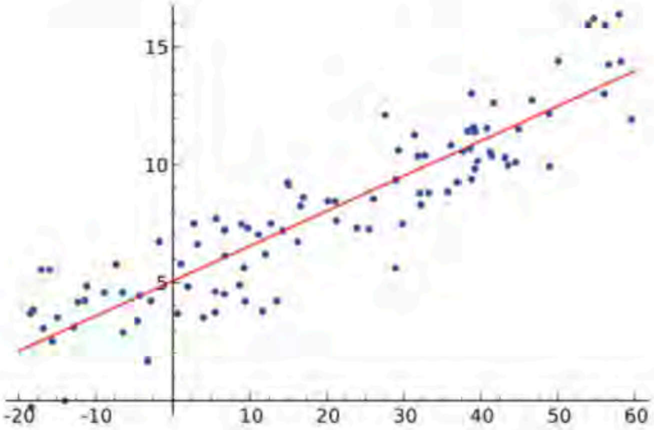
- Assign samples into categories (classes) based on a predictable attribute.
- The goal of classification is to accurately predict the target class for each case in the data set.

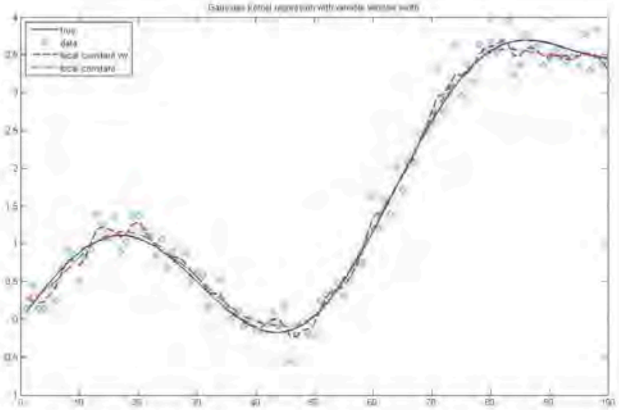




DM tasks: Regression

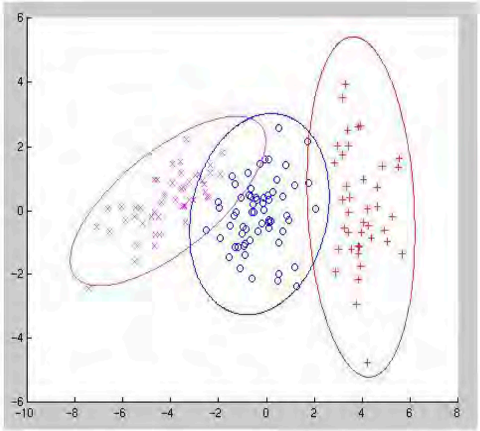
- Compute the new values for a dependent variable based on the values of one or more measured attributes.
- Examples:
  - predict wind velocities based on temperature, air pressure and humidity;
  - predict coupon redemption rate based on the face value, distribution method and distribution volume



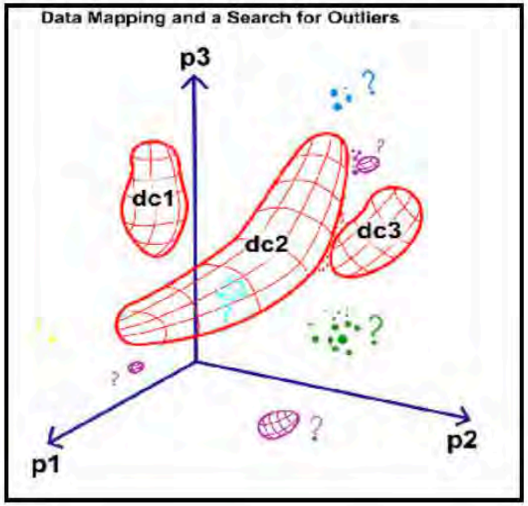


DM tasks: clustering

- Clustering
  - partitioning of a data set into subsets (clusters) so that data in each subset ideally share some common characteristics.

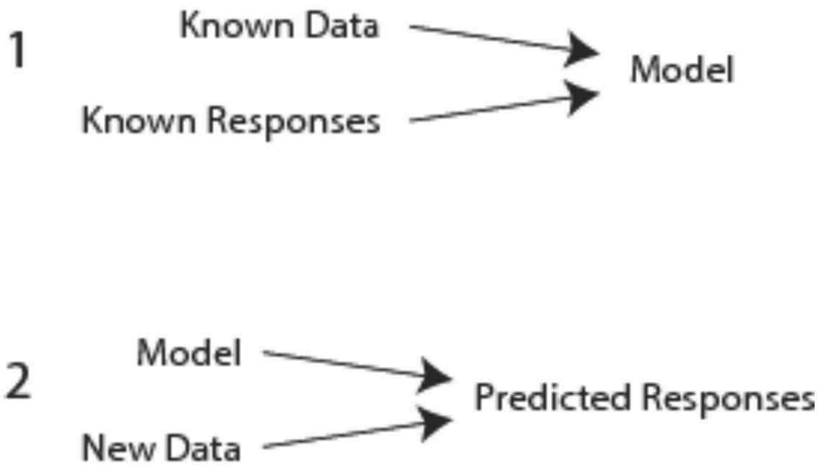


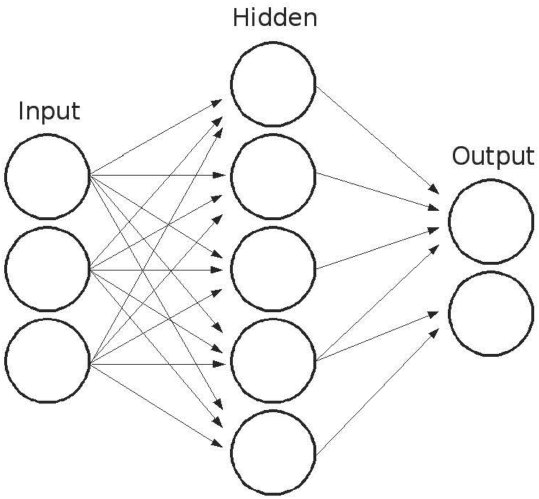
- Deviation Analysis (search for outliers)
  - anomalies;
  - peculiar objects.



Supervised Learning

- For some examples the correct results (targets) are known and are given in input to the model during the learning process.
- Generalization: ability of a learning machine to perform accurately on new, unseen examples.

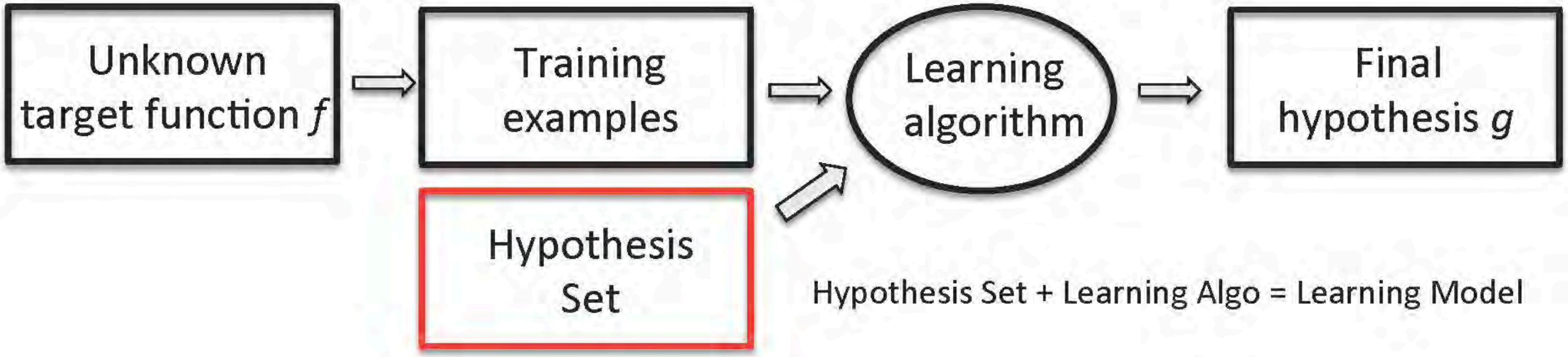




Images credits: Mathworks

Supervised Learning

- Training data consists of a set of training examples
- Ideal target function *f*
- Hypothesis *g* that best approximate *f*
- Learning algorithm
  - – connect target function and hypothesis
- Hypothesis Set
- Predict new inputs: $y_{new} = g(\mathbf{x}_{new})$



Hypothesis Set + Learning Algo = Learning Model

Supervised Learning in a nutshell

① Define the data to be used as a learning set

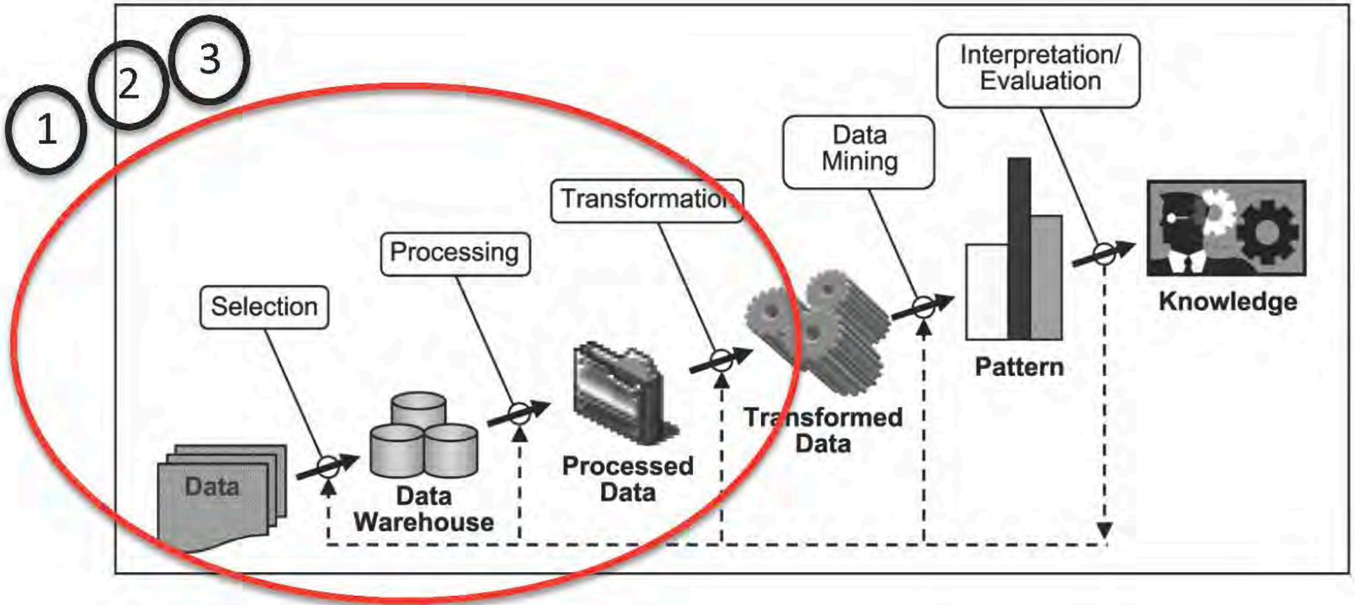
- eg, handwriting analysis: single character or entire words?

② Prepare the training set

- eg, create training, validation and test sets

③ Transform the input data in feature vectors (X,Y)

- eg, extract/select features to avoid the curse of dimensionality



Supervised Learning in a nutshell

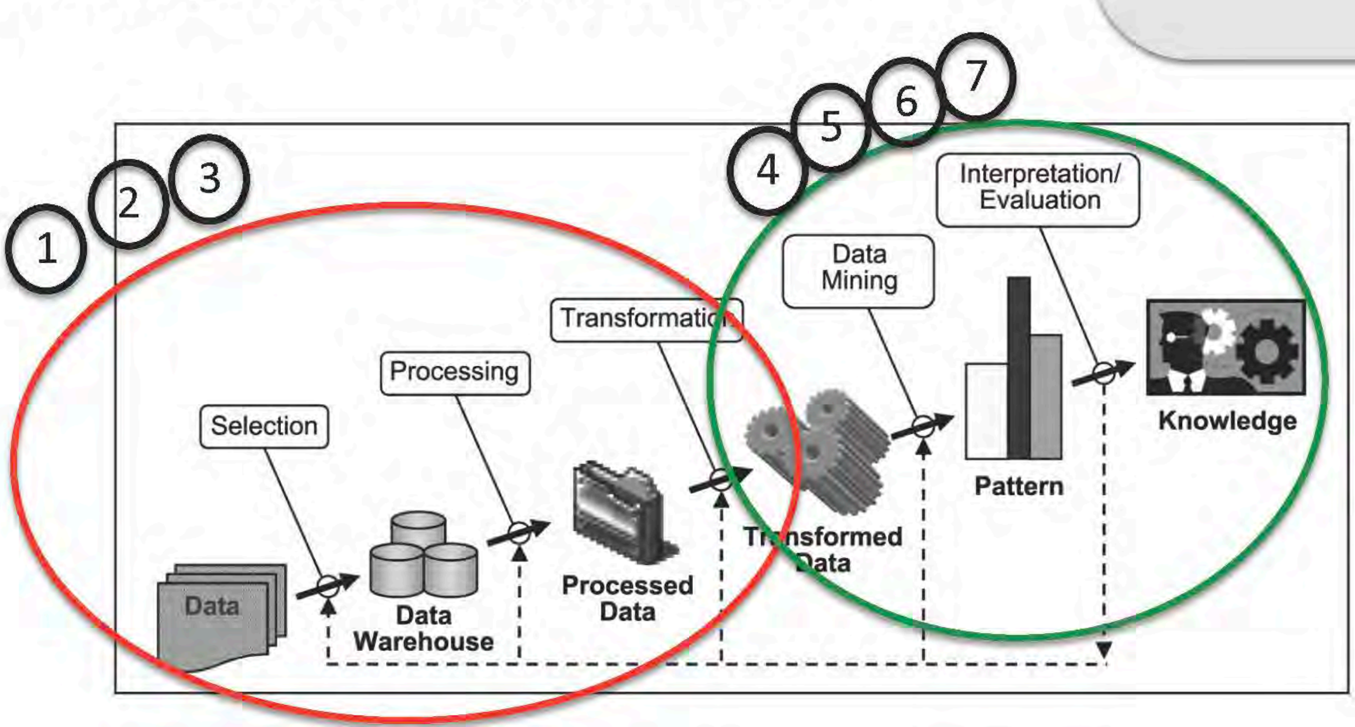
④ Choose the learning model
  – eg, Neural Network and Back propagation

⑤ Choose a validation model
  – eg, cross validation, random splits, etc

⑥ Run the algorithm, compute the accuracy and update until satisfied
  – eg, minimize the loss, minimize the MSE, etc

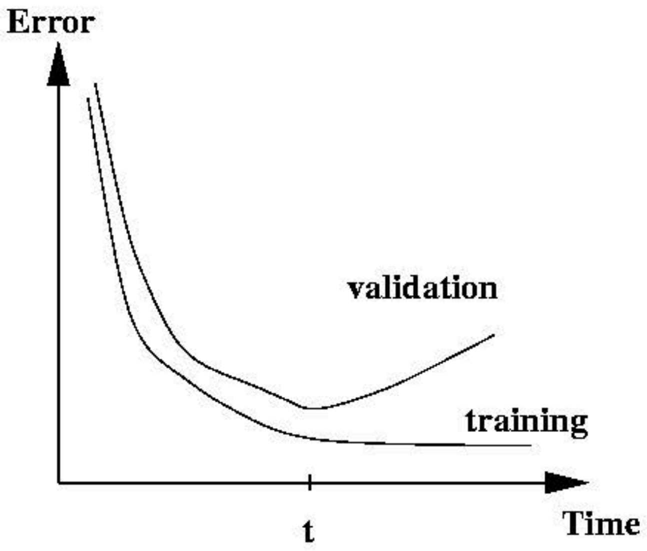
⑦ Use final model to make predictions

Supervised Algorithms
Neural Networks (MLP)
Boltzmann Machines
RBM
Decision Trees
Nearest Neighbor
Naive Bayes Classifiers
Bayesian Networks
GPR
...



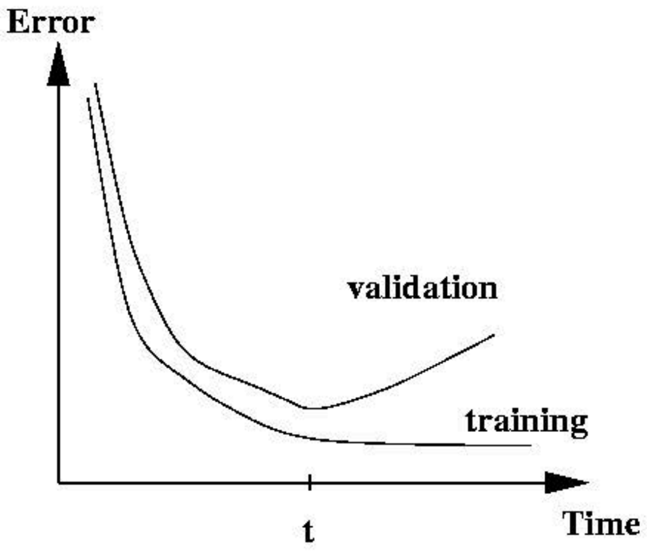
Training Set

- **Training set**: a set of examples used for learning, where the target value is known.
- The goal of the learning algorithm is to build a model which makes accurate predictions on the training set.
- Training set accuracy does not give a good indication about the generalization power of the model.
- Add a Validation set.



Validation Set

- Set of examples used to tune the architecture of a classifier and estimate the error.
- Used for model selection.
- The validation data has to be representative of the range of inputs the classifier is likely to encounter.
- How to create it?
  - gather new data;
  - random split:
    - 80-20
    - cross-validation

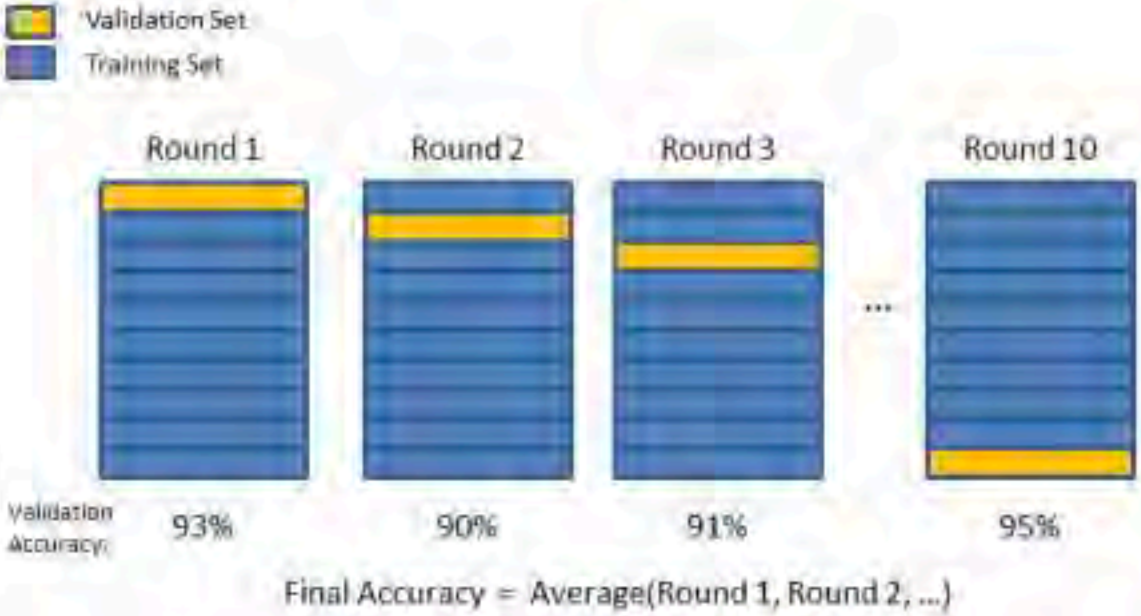


Cross-Validation

- C-V techniques are used for assessing how the results of a statistical analysis will generalize to an independent data set.
- Exhaustive Cross-Validation
  - leave one out cross validation (LOOCV)
  - leave p-out cross validation
- Non-exhaustive Cross-Validation
  - *k*-fold cross validation
  - repeated random sub-sampling validation
- Choose also according to your model/task.

K-fold cross-validation

- How it works:
  - randomly partition the original into *k* subsamples;
  - of the *k* subsamples, one is retained as the validation data for testing the model, and the remaining *k* – 1 are used as training data;
  - the process is then repeated *k times*, with each of the *k* subsamples used exactly once as the validation data;
  - the *k* results can be averaged (or otherwise combined) to produce a single estimation.



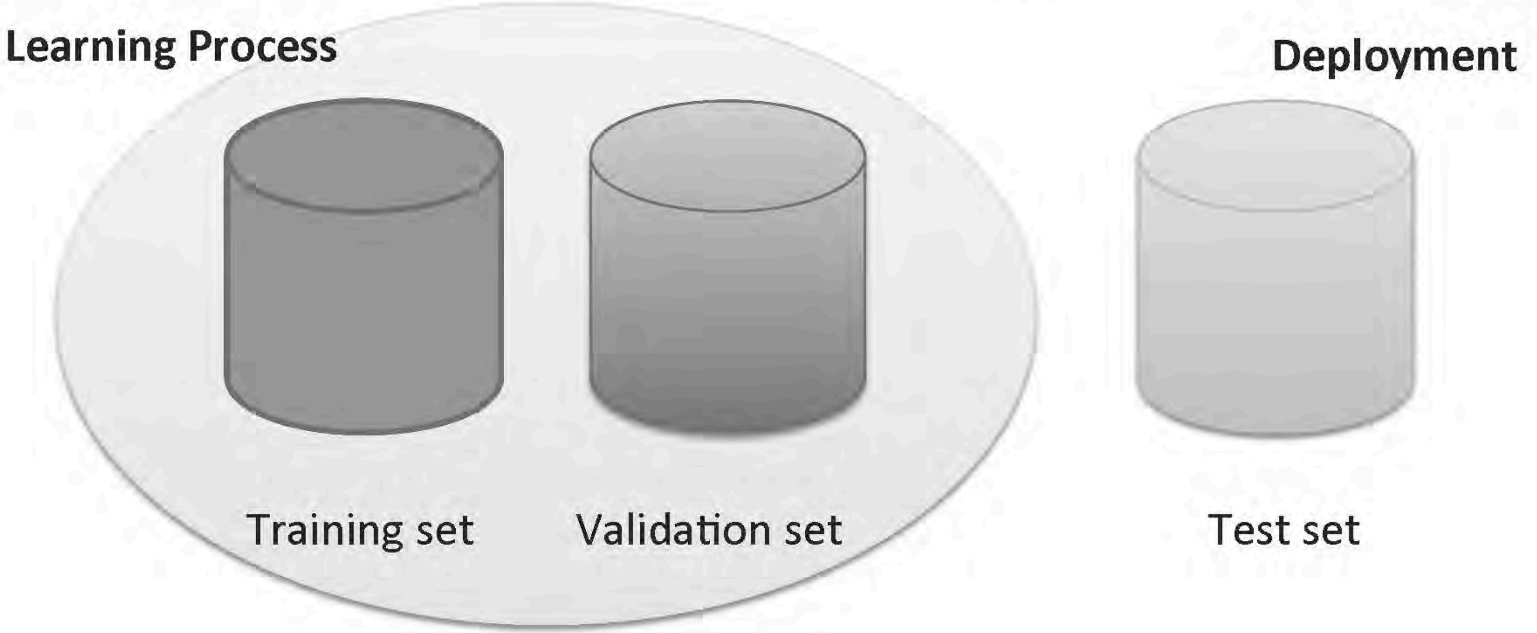
Picture credit: chrisjmccormick

Repeated random sub-sampling

- Repeated Random Sub-Sampling
  - at each step: randomly split the dataset into two subsets: training and validation;
  - compute the validation errors.
  - average the results over the splits.
- Advantage: proportion of the sets not dependent on the number of folds.
- Disadvantage: some samples may never be selected for validation, some may be selected more than once.

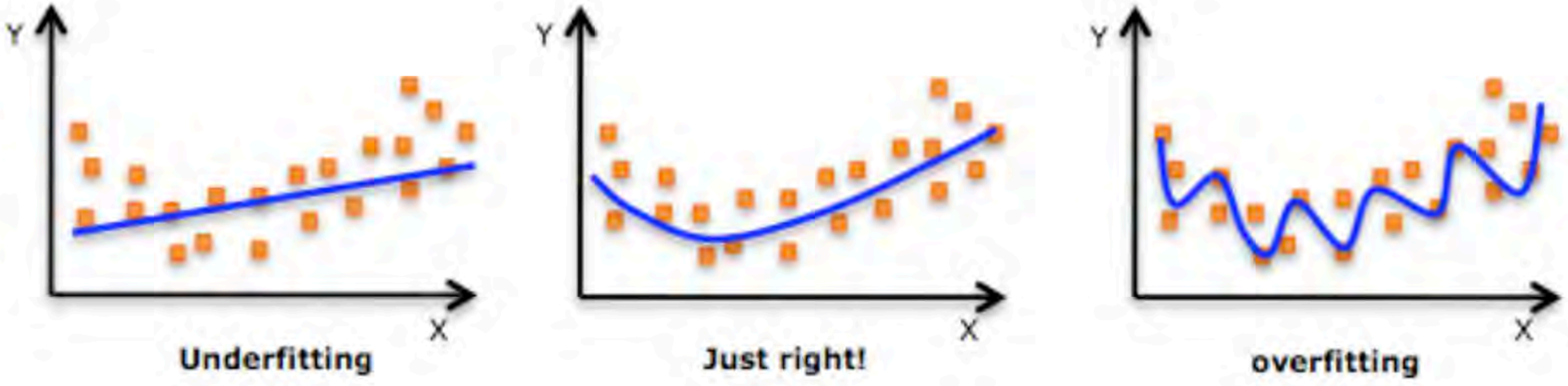
Test Set

- **Test set**: used only to assess the performances of a fully trained classifier.
- It is **never used** during the training process so that the error on the test set provides an unbiased estimate of the generalization error.



A common problem: OVERFITTING

- Model is not be able to generalize.
- Learn the "data" and not the underlying function.
- Performs well on the data used during the training and poorly with new data.
- How to avoid: cross-validation, early stopping, regularization, Bayesian priors, model comparison.



Overfitting in supervised learning

- Example: overfitting in supervised learning.
- Blu is the training error, red the validation error, over time.
- If the validation error increase while the training error decrease it is a warning sign for overfitting.

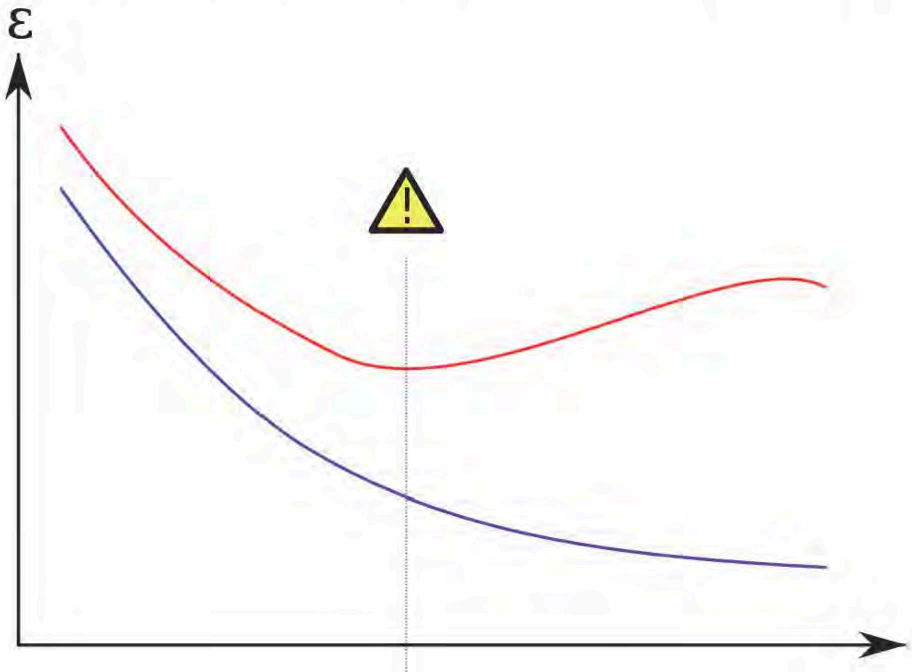
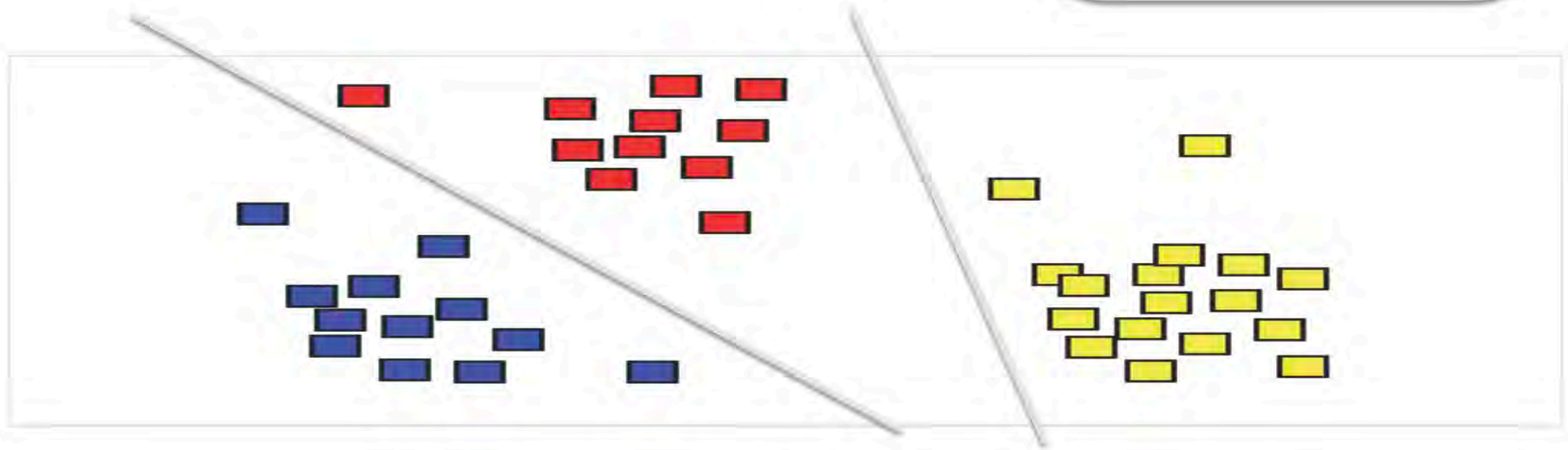


Image credit: Wikipedia.

Unsupervised Learning

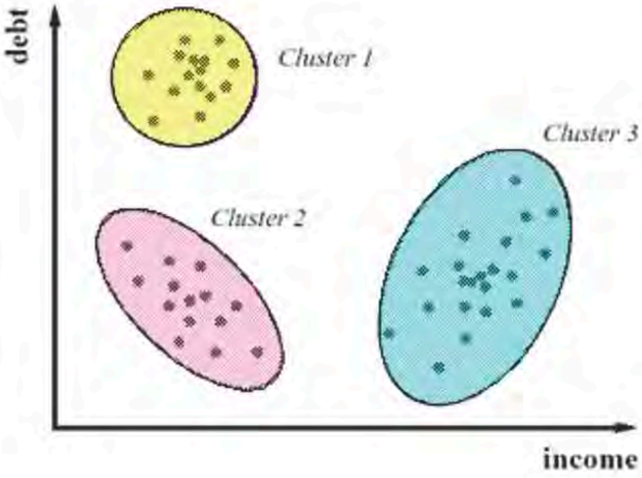
- Supervised: (input, correct output)
  Unsupervised Learning: (input)
- The model is **not** provided with the correct results during the training.
- Can be used to cluster the input data in classes **on the basis of their statistical properties only**.

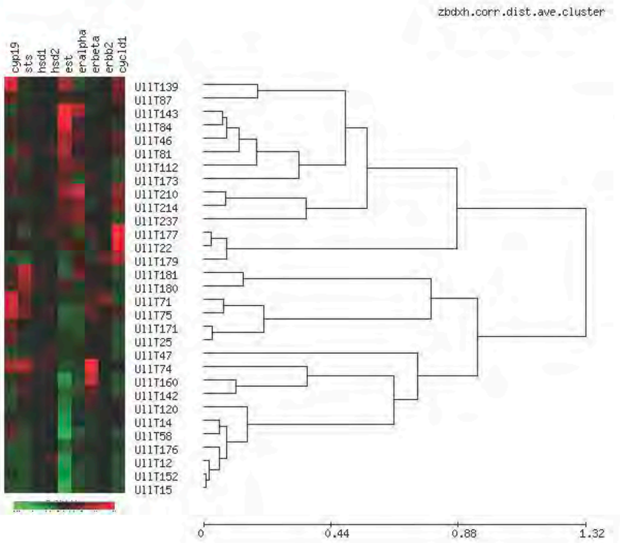
Unsupervised Algorithms
K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal
Surfaces
...

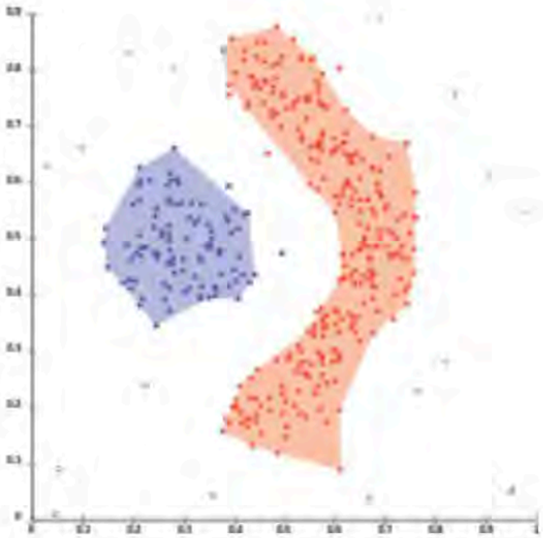


Types of Clustering

- PARTITIONING: construct various partitions and then evaluate them based on some criterion.
- HIERARCHICAL: finds successive clusters using previously established clusters (can be agglomerative or divisive).
- DENSITY-BASED: based on connectivity and density functions.

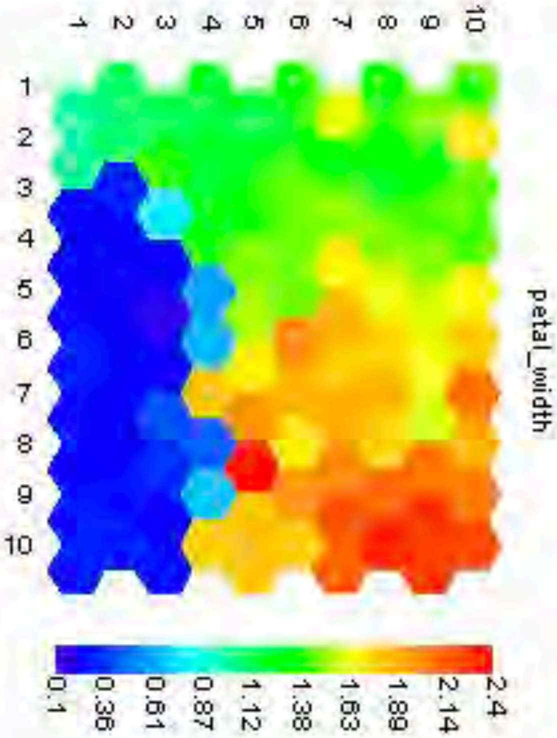






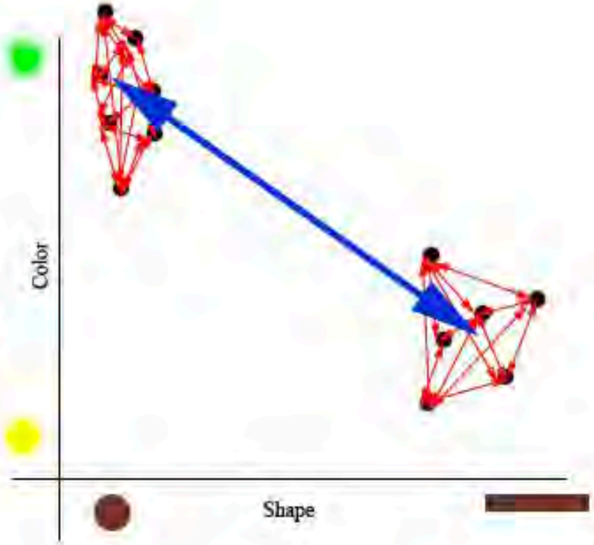
Labeling

- Assign a known label to an object.
- The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes.
- Helps to identify clusters.



Good Clustering?

- Clusters can be evaluated with "internal" as well as "external" measures:
  - internal measures are related to the inter/intra cluster distance;
  - external measures are related to how representative are the current clusters to "true" classes.
- A good clustering is one where:
  - the intra-cluster distance is minimized: defined as the sum of distances between objects in the same clusters;
  - the inter-cluster distance is maximized: defined as the distances between different clusters.
- Define "distance".



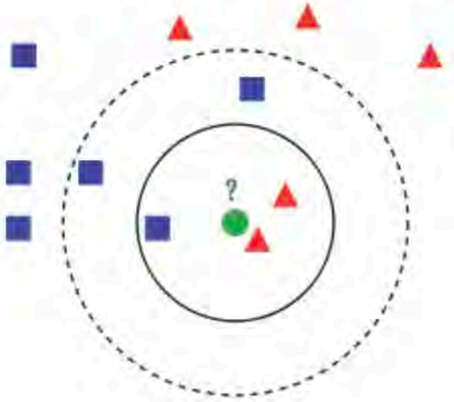
Distance between clusters

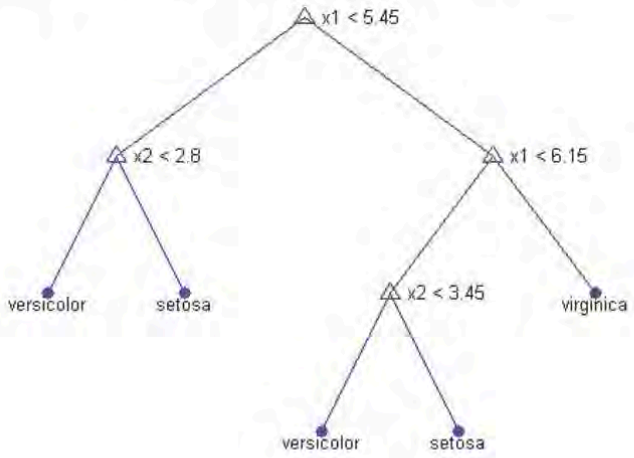
- Single link
  - smallest distance between an element in one cluster and an element in the other
    $\text{dis}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- Complete link
  - largest distance between an element in one cluster and an element in the other
    $\text{dis}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- Average
  - average distance between an element in one cluster and an element in the other
  - i.e., $\text{dis}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- Centroid
  - distance between the centroids of two clusters
    $\text{dis}(K_i, K_j) = \text{dis}(C_i, C_j)$
- Medoid
  - distance between the medoids of two clusters
    $\text{dis}(K_i, K_j) = \text{dis}(M_i, M_j)$

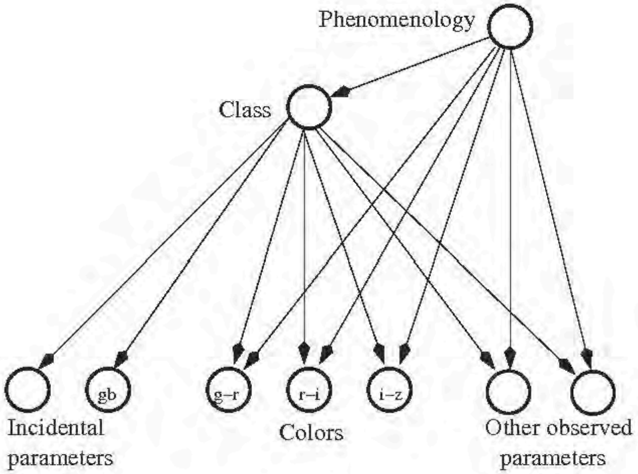
Credit: Data Mining: Concepts and Techniques, J. Han, M. Kamber

A two step process: model usage

- Model usage: classifying future or unknown objects
  - estimate accuracy;
    - use an independent test set;
    - eg, accuracy rate: percentage of test samples that are correctly classified;
    - if the accuracy is acceptable, use the model to classify data whose labels are not known.
  - Output: crispy or probabilistic.





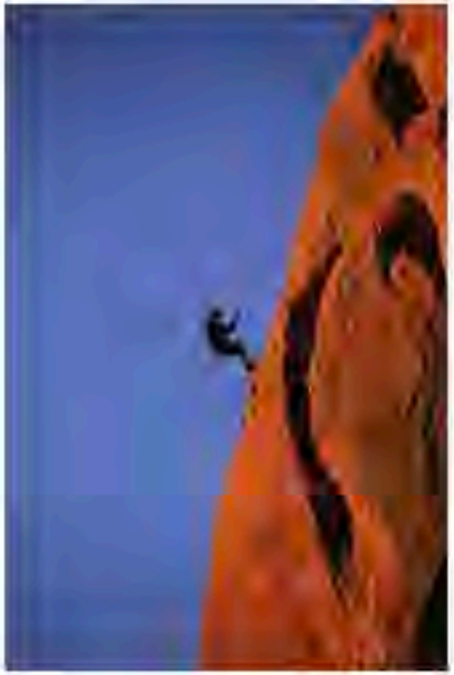


Crispy vs Probabilistic

- **Crispy classification**
  - given an input, the classifier returns its label
- **Probabilistic classification**
  - given an input, the classifier returns its probabilities to belong to each class;
  - useful when some mistakes can be more costly than others;
  - allow thresholds (e.g., give me only data >90%)
  - winner take all and other rules
    - assign the object to the class with the highest probability (WTA)
    - …but only if its probability is greater than 40% (WTA with thresholds)

Classifiers evaluation

- Accuracy
  - ability of correctly predicti class labels.
- Speed
  - training time, classification time.
- Scalability
  - classifying data sets with millions of examples and hundreds of attributes with reasonable speed.
- Robustness
  - ability of handling missing data, noise, etc.
- Interpretability



Preparing the data

- Relevance analysis
  - remove redundant and irrelevant attributes;
  - correlation analysis can be used to examine whether two variables changes together in a consistent manner.
    - Pearson coefficient for linear correlation;
  - feature selection.

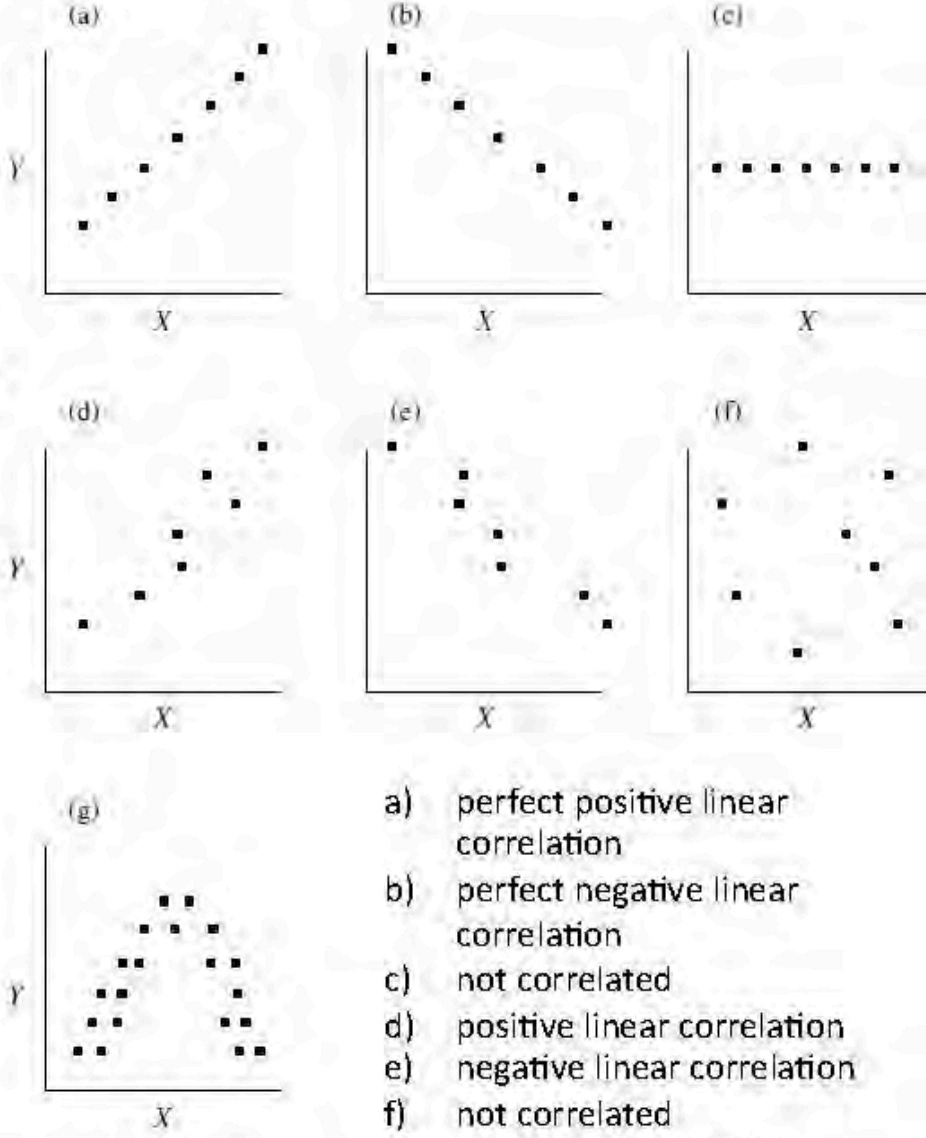
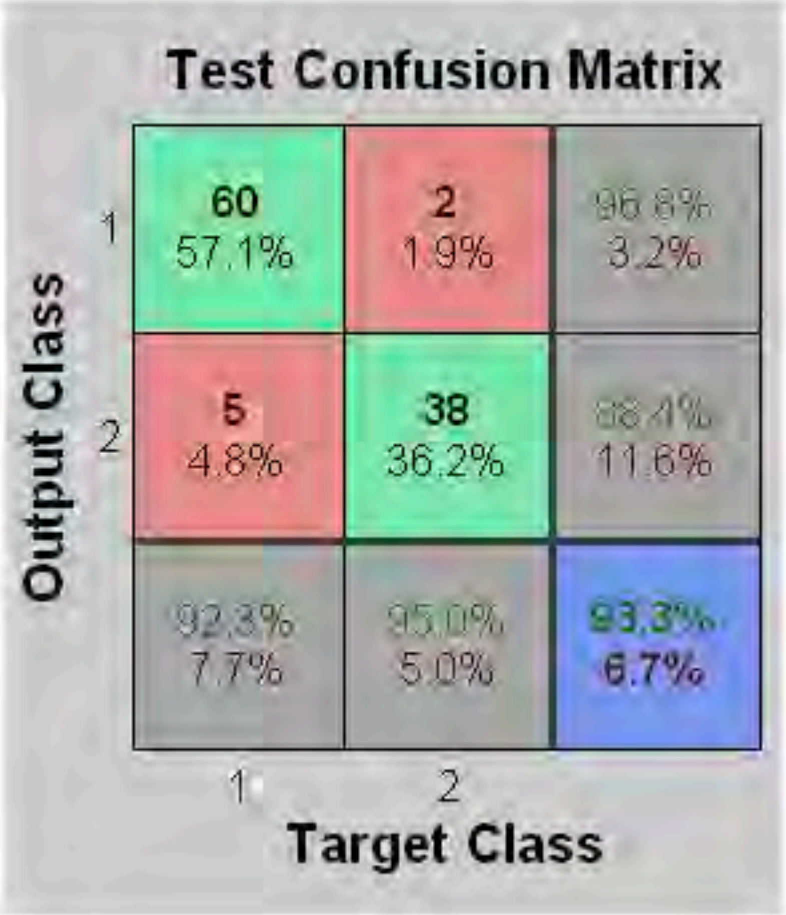


Image credits: onfolio

a) perfect positive linear correlation
b) perfect negative linear correlation
c) not correlated
d) positive linear correlation
e) negative linear correlation
f) not correlated
g) non-linear correlation

Completeness and Contamination

- **Completeness**: the percentage of objects of a given class correctly classified as such. (ex, class 1: 96.8% compl.);
- **Contamination**: for each class, the percentage of objects of other classes incorrectly classified as objects belonging to that class (ex, class 1: 7.7% cont.)
- **Precision**: 1-Contamination



Test Confusion Matrix

Output Class	1	2	
1	60 57.1%	2 1.9%	96.8% 3.2%
2	5 4.8%	38 36.2%	88.4% 11.6%
	92.3% 7.7%	95.0% 5.0%	93.3% 6.7%

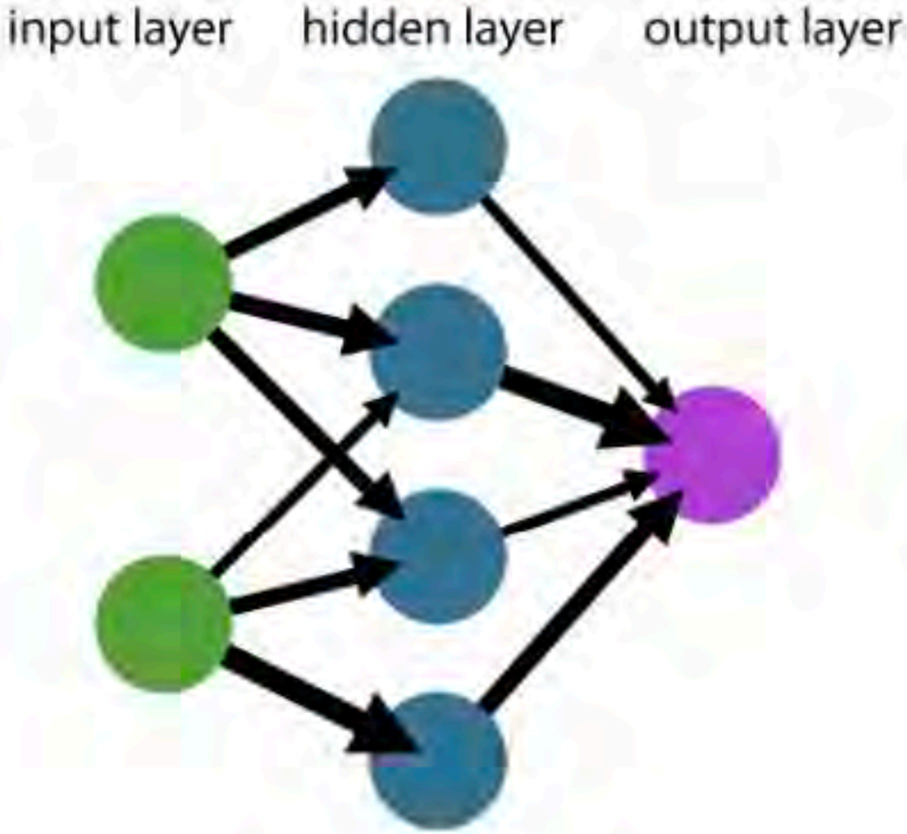
Target Class

Artificial Neural Networks

An Artificial Neural Network is an information processing paradigm that is inspired by the way biological nervous systems process information:

> **"a large number of highly interconnected simple processing elements (neurons) working together to solve specific problems"**

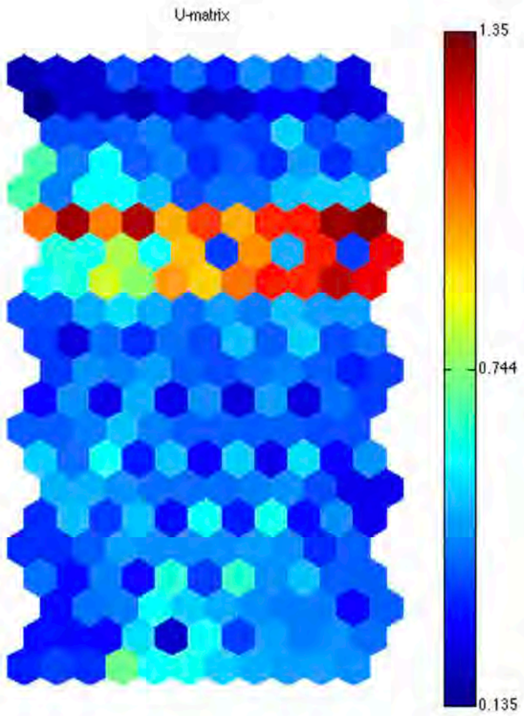
A simple neural network

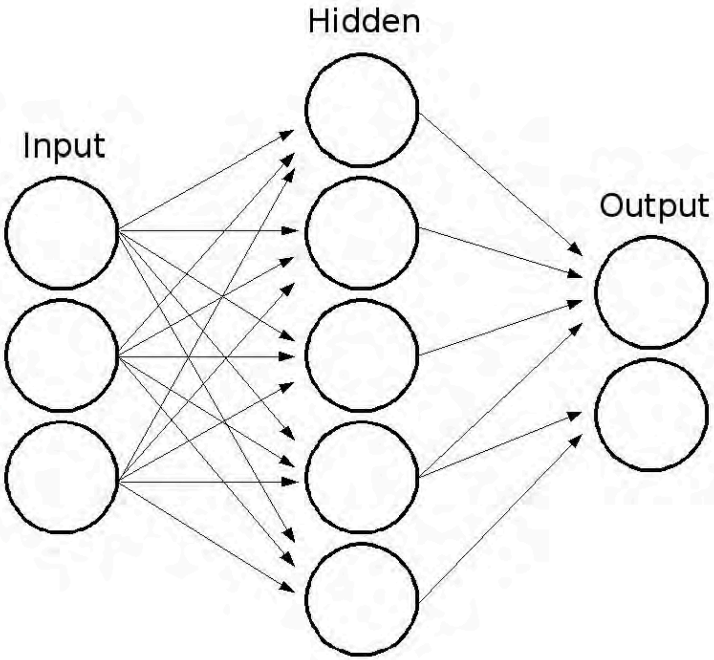


Neural Networks

A Neural Network is usually structured into an input layer of neurons, one or more hidden layers and one output layer.

Neurons belonging to adjacent layers are usually fully connected and the various types and architectures are identified both by the different topologies adopted for the connections as well by the choice of the activation function.





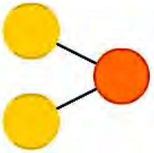
A mostly complete chart of

Neural Networks

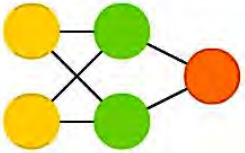
©2016 Fjodor van Veen - asimovinstitute.org

- 🟡◯ Backfed Input Cell
- 🟡 Input Cell
- 🟡△ Noisy Input Cell
- 🟢 Hidden Cell
- 🟢◯ Probablistic Hidden Cell
- 🟢△ Spiking Hidden Cell
- 🟠 Output Cell
- 🟠◯ Match Input Output Cell
- 🔵 Recurrent Cell
- 🔵◯ Memory Cell
- 🔵△ Different Memory Cell
- 🟣 Kernel
- 🟣◯ Convolution or Pool

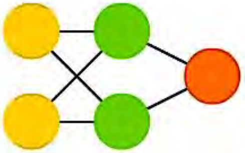
Perceptron (P)



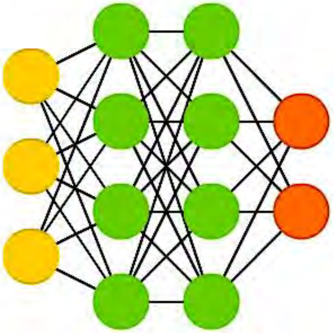
Feed Forward (FF)



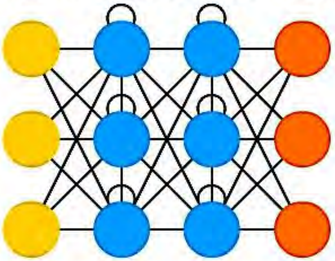
Radial Basis Network (RBF)



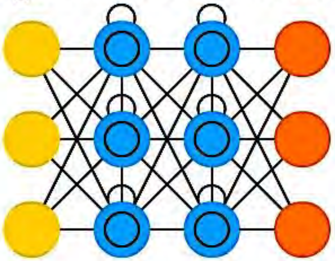
Deep Feed Forward (DFF)



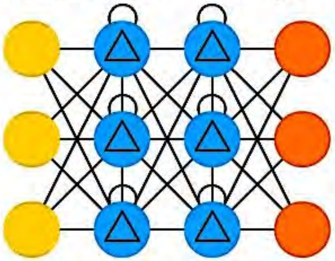
Recurrent Neural Network (RNN)



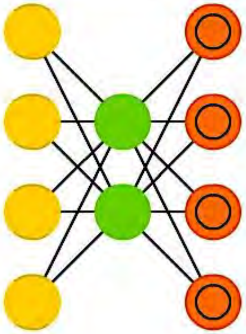
Long / Short Term Memory (LSTM)



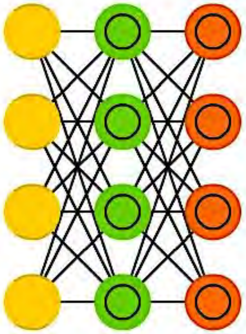
Gated Recurrent Unit (GRU)



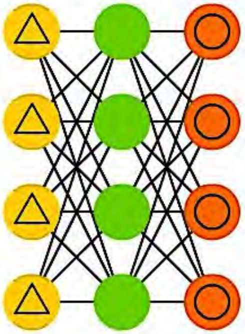
Auto Encoder (AE)



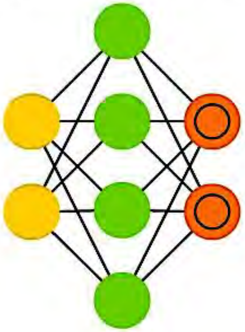
Variational AE (VAE)

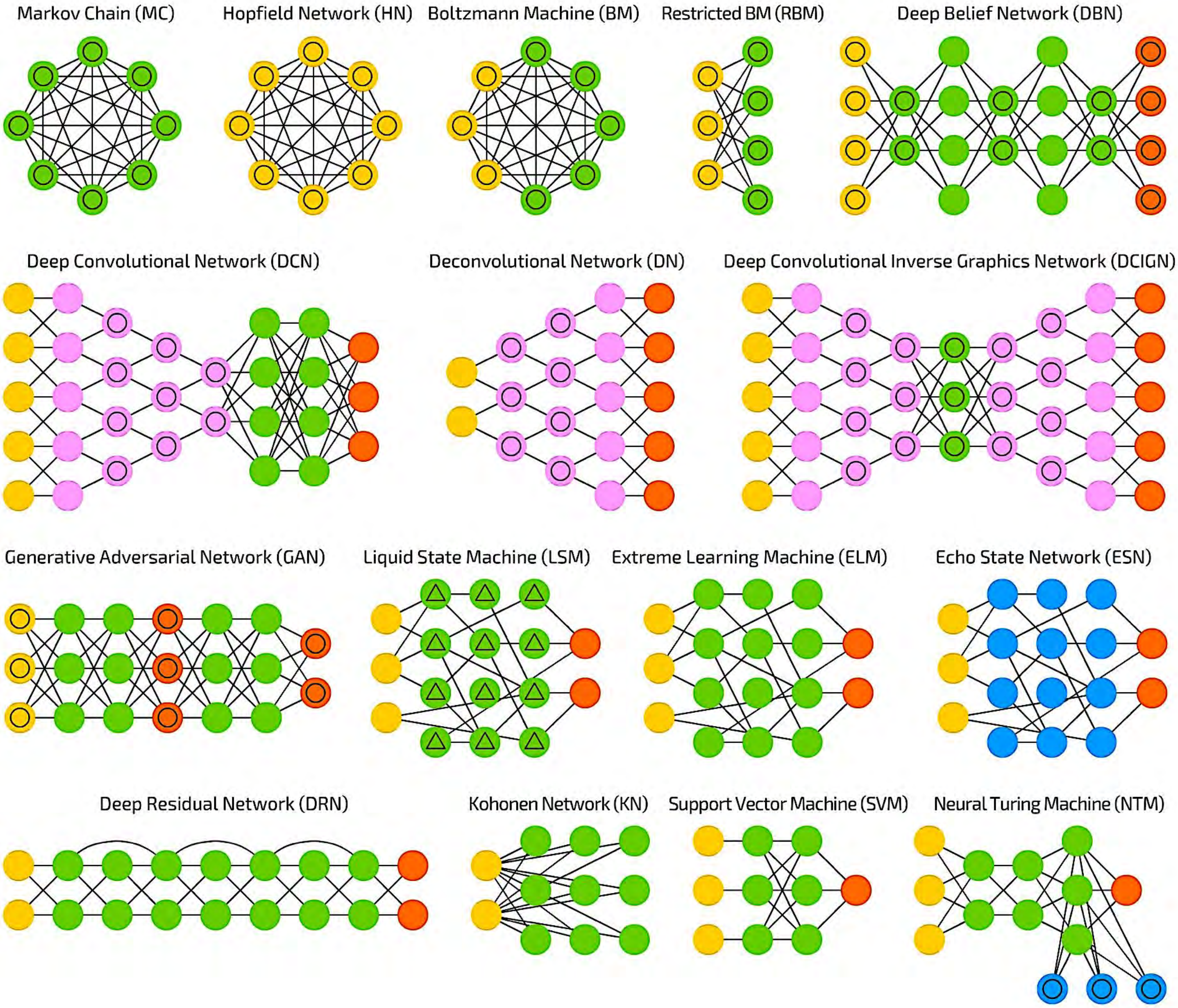


Denoising AE (DAE)



Sparse AE (SAE)



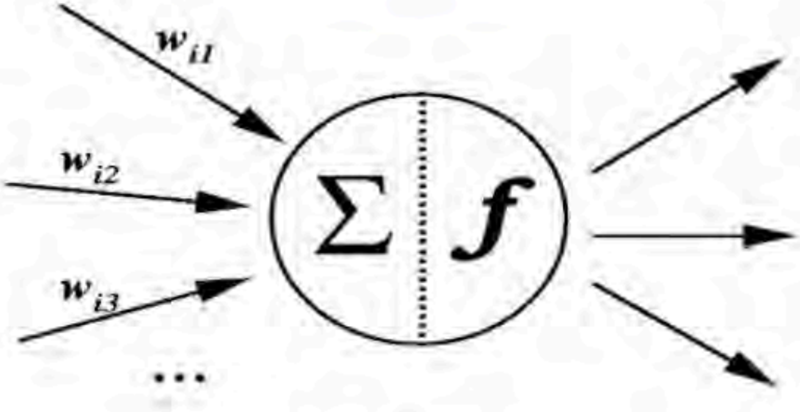

Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Kohonen Network (KN)
Support Vector Machine (SVM)
Neural Turing Machine (NTM)

And more...

A simple Artificial Neuron

- The basic computational element is called “node” or “unit”.
- Receives inputs from some other units, or from an external source .
- Each input has an associated weight **w**, which can be modified so as to model synaptic learning.
- The unit computes some function of the weighted sum of its inputs:

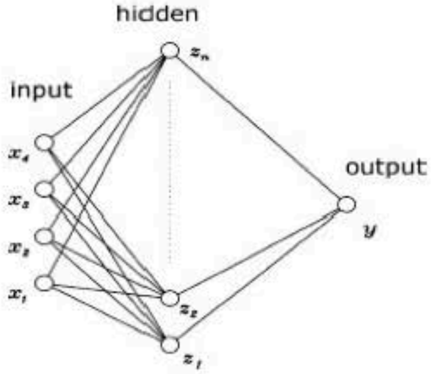
$$y_i = f(\sum_j w_{ij} y_j)$$



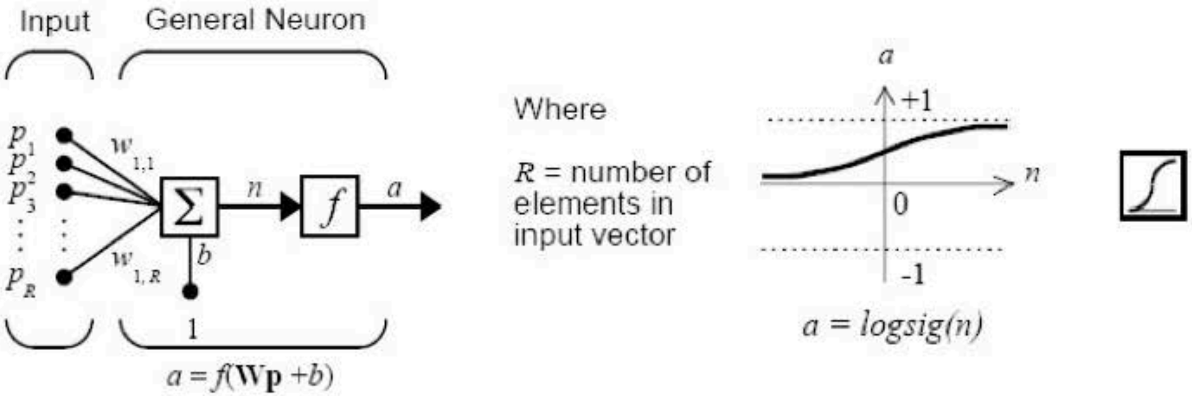
$y_i = f(net_i)$

Multi-Layer Perceptron

- The MLP is one of the most used supervised model: it consists of multiple layers of computational units, usually interconnected in a feed-forward way.
- Each neuron in one layer has direct connections to all the neurons of the subsequent layer.



The architecture of a two layer MLP.



$a = f(\mathbf{W}\mathbf{p} + b)$

Where

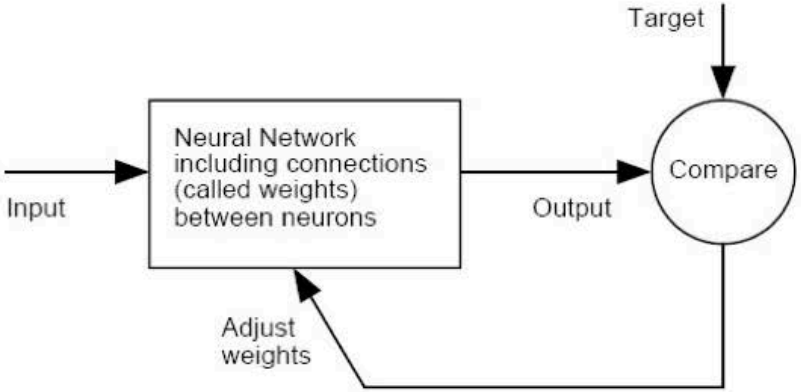
R = number of elements in input vector

$a = logsig(n)$

Learning Process

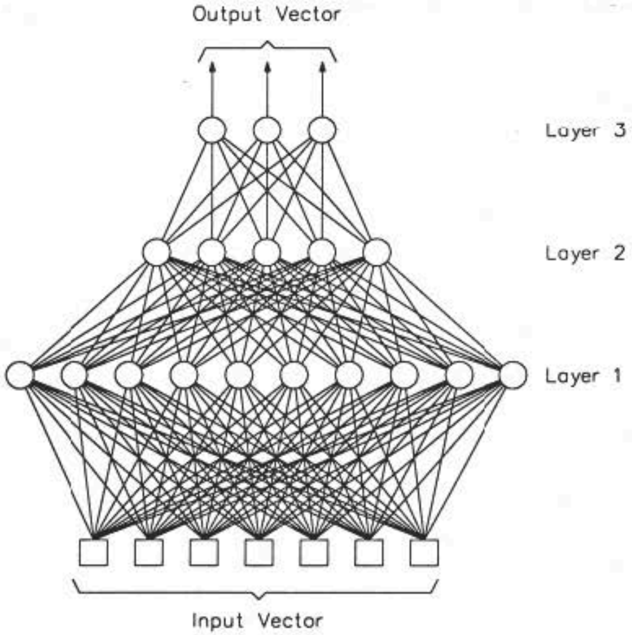
- Back Propagation
  - the output values are compared with the target to compute the value of some predefined error function;
  - the error is then feedback through the network;
  - using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function.

After repeating this process for a sufficiently large number of training cycles, the network will usually converge.


Target
Neural Network including connections (called weights) between neurons
Input
Output
Compare
Adjust weights

Hidden Units

- The best number of hidden units depend on:
  - number of inputs and outputs;
  - number of training cases;
  - the amount of noise in the targets;
  - the complexity of the function to be learned;
  - the activation function, etc.

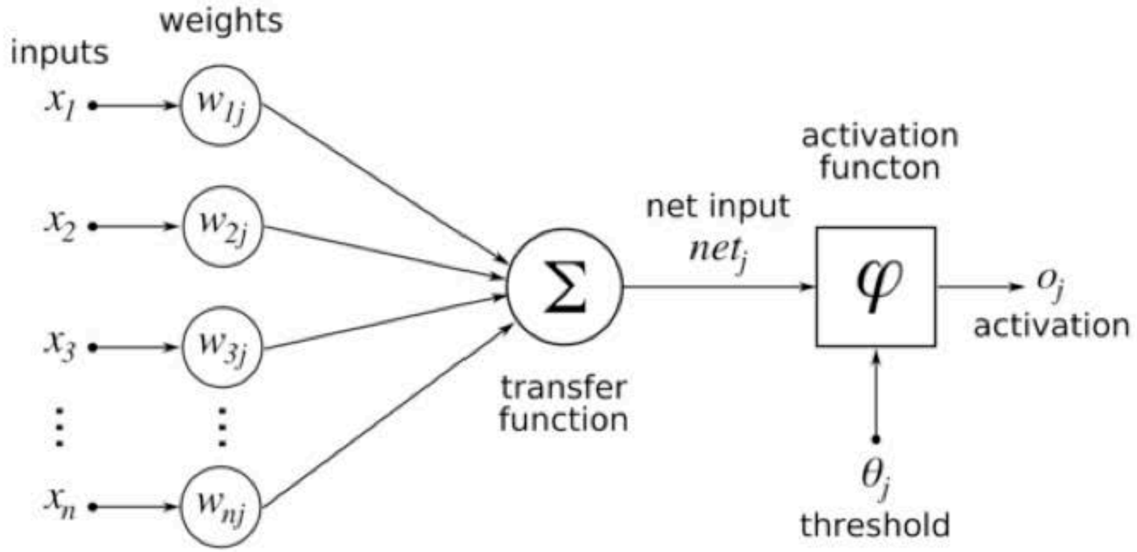


- Too few hidden units => high training and generalization error, due to underfitting and high statistical bias.
- Too many hidden units => low training error but high generalization error, due to overfitting and high variance.
- Rules of thumb don't usually work.

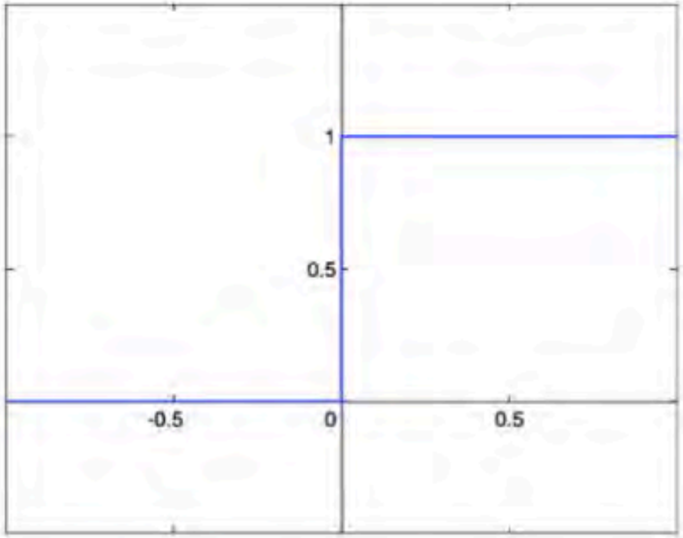
Image Credits: statistics4u

Activation Functions

- Activation Functions
  - used by most units to transform their inputs;
  - needed to introduce non-linearity into the network;
  - linear, logistic, tanh, softmax...

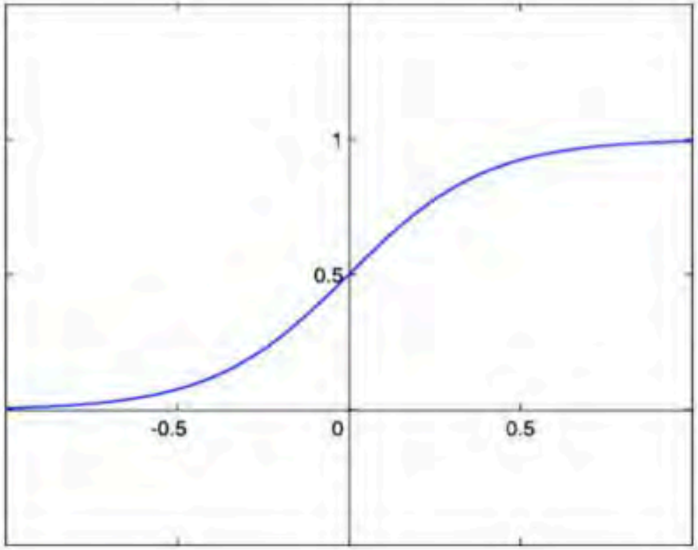


Activation Functions: examples



Step function
The output is a certain value A1, if the input sum is above a certain threshold and A0 if the input sum is below a certain threshold.

When we want to classify an input pattern into one of two groups, we can use a binary classifier with a step activation function.



Sigmoid function
Has the property of being similar to the step function, but with the addition of a region of uncertainty.

Sigmoid functions in this respect are very similar to the input-output relationships of biological neurons.

$$\sigma(t) = \frac{1}{1 + e^{-\beta t}}$$

Activation and Error Functions

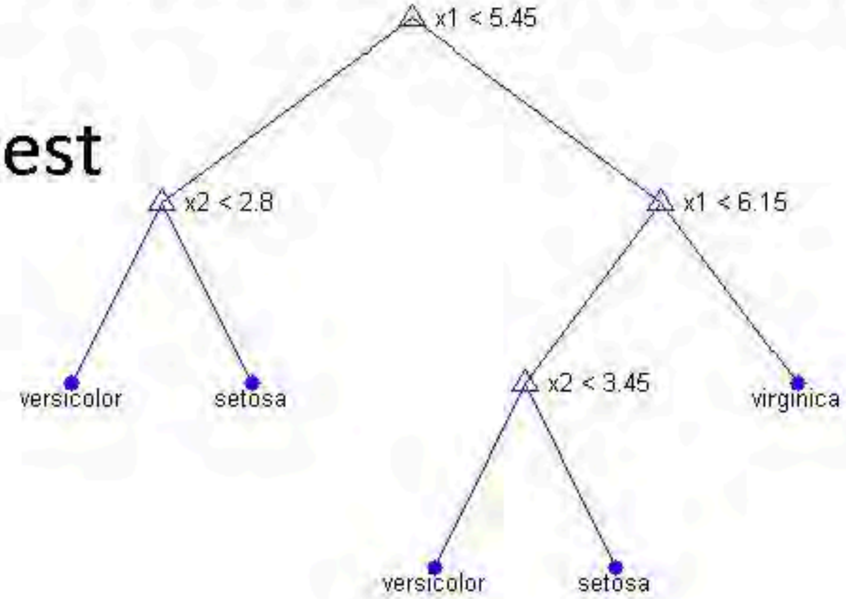
- Error Functions
  - most methods for supervised learning require a measure of the discrepancy between the netwrok output values and the target;
  - sum of the squared errors (SSE), cross entropy (CE), etc.

$$E_p = \frac{1}{2} \sum_j \left(t_j^p - y_j^p \right)^2$$

Using a Multilayer Perceptron with a softmax activation function and cross-entropy error, the network outputs can be interpreted as the conditional probabilities $p(C_1|\mathbf{x})$ and $p(C_2|\mathbf{x})$ where $\mathbf{x}$ is the input vector, C_1 the first class, C_2 the second class.

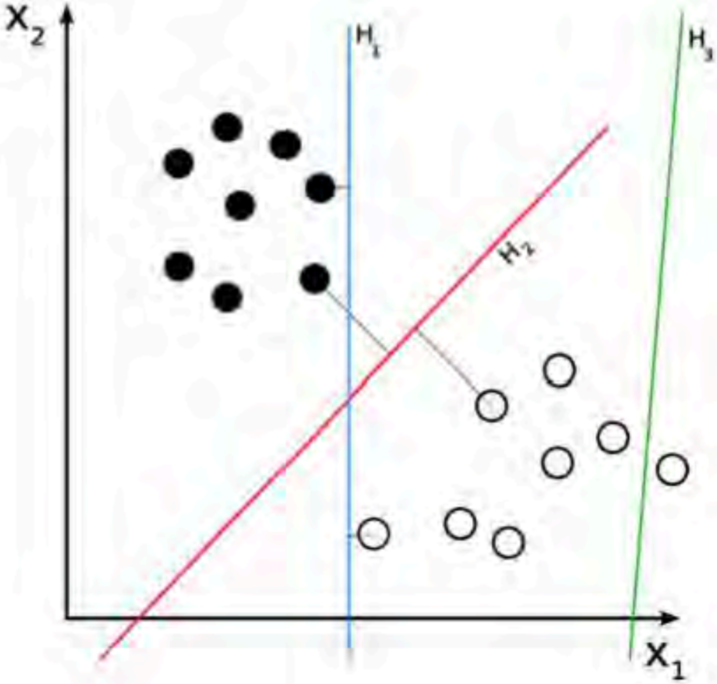
Decision Trees

- Supervised classification method.
- A decision tree is a set of simple rules, such as “if the feature 1 is less than x and feature 2 is greater than y, classify the specimen as AGN”.
- Non-parametric: they do not require any assumptions about the distribution of the variables in each class.
- Internal nodes denotes test on the attributes.
- Leaves represent the class labels.



Support Vector Machines

- Support Vector Machines (SVM) are a group of supervised learning methods that can be applied to classification or regression.
- For any particular set of two-class objects, an SVM finds the unique hyperplane having the maximum margin.



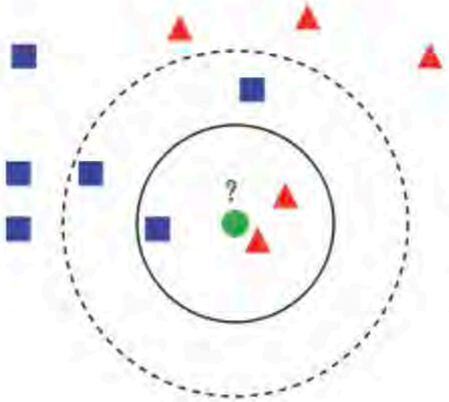
H3 (green) doesn't separate the 2 classes.

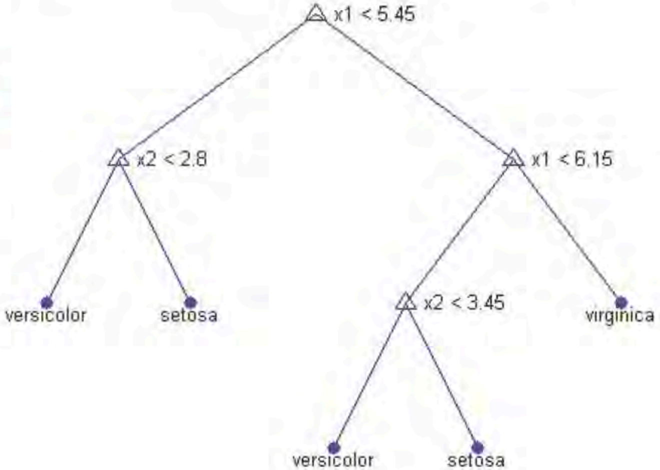
H1 (blue) does, with a small margin.

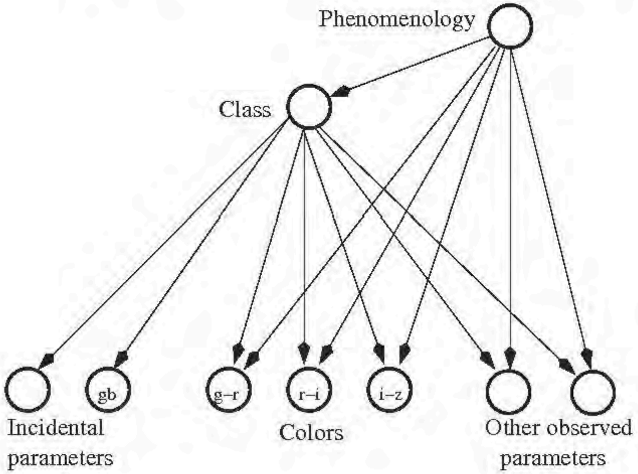
H2 (red) does with the maximum margin.

Combining Models

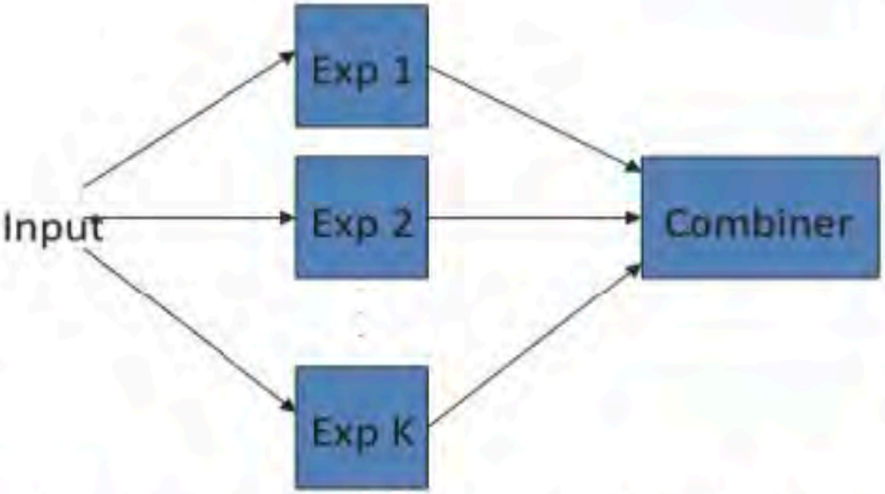
- It is often found that improved performance can be obtained combining model together.
  - individual classifiers may be optimized and trained differently;
  - some classifiers could work better than others in recognizing some classes when certain input attributes are present;
  - some can deal with missing data while some others not.



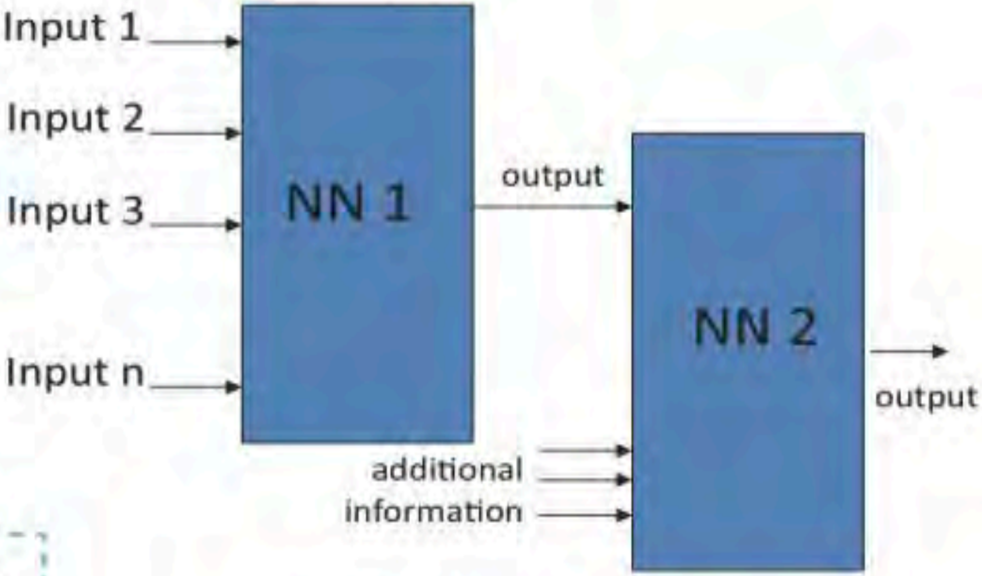




Committee of Machines



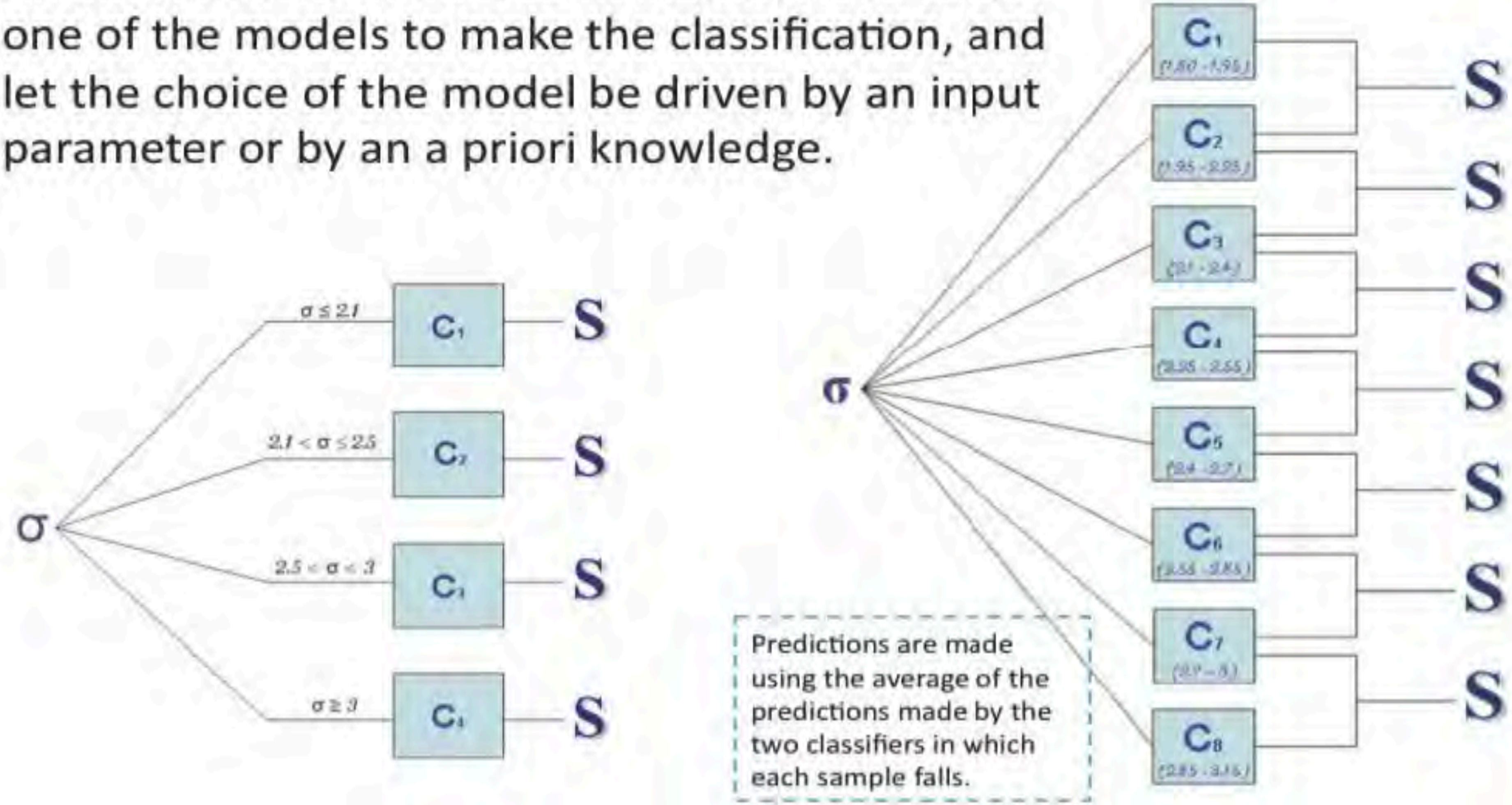
Committee Machines: combination of experts that "vote" together on a given example.



Two-level Network.

A priori Knowledge

An alternative of model combinations is to select one of the models to make the classification, and let the choice of the model be driven by an input parameter or by an a priori knowledge.



Predictions are made using the average of the predictions made by the two classifiers in which each sample falls.

What is Clustering?

- Unsupervised Learning
- **Cluster hypothesis**: *objects in the same cluster behave similarly with respect to relevance to information needs*.
  - points in the same cluster are likely to be of the same type.
  - finding natural groupings among objects.

Unsupervised Algorithms

K-Means

Self-Organizing Maps

RDF

Fuzzy Clustering

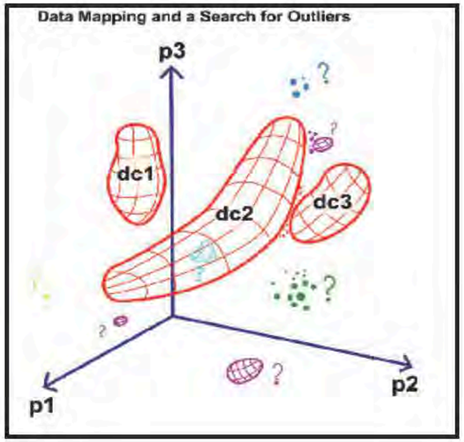
CURE

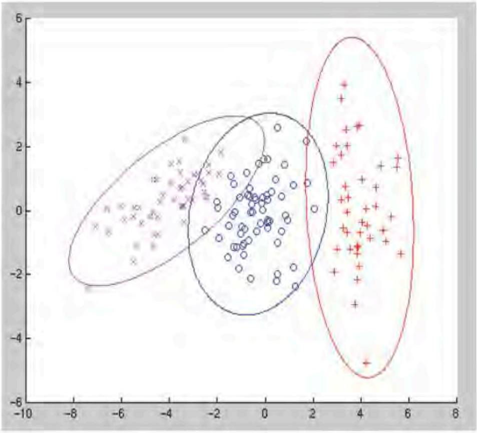
ROCK

Vector Quantization

Probabilistic Principal Surfaces

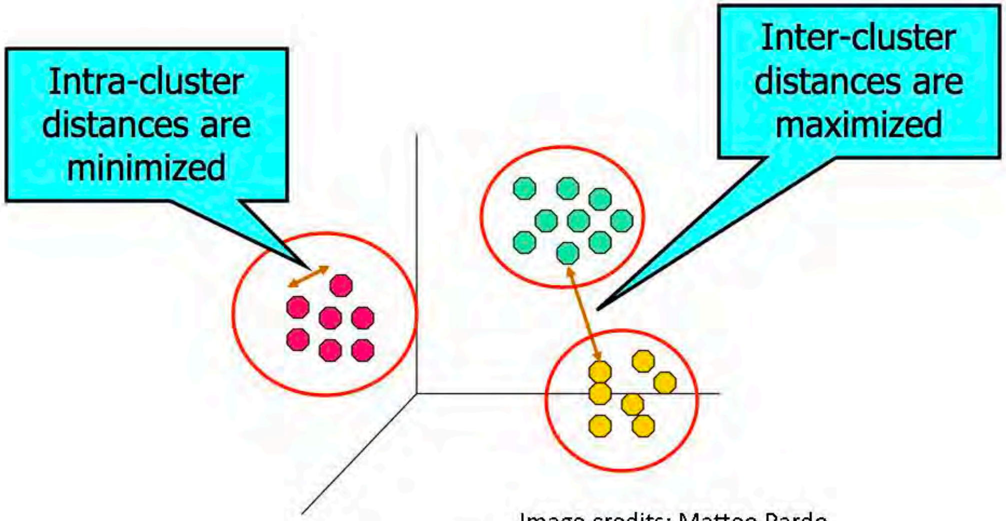
...





Internal Measures

- A good clustering is one where:
  - the intra-cluster distance is minimized: defined as the sum of distances between objects in the same clusters;
  - the inter-cluster distance is maximized: defined as the distances between different clusters.



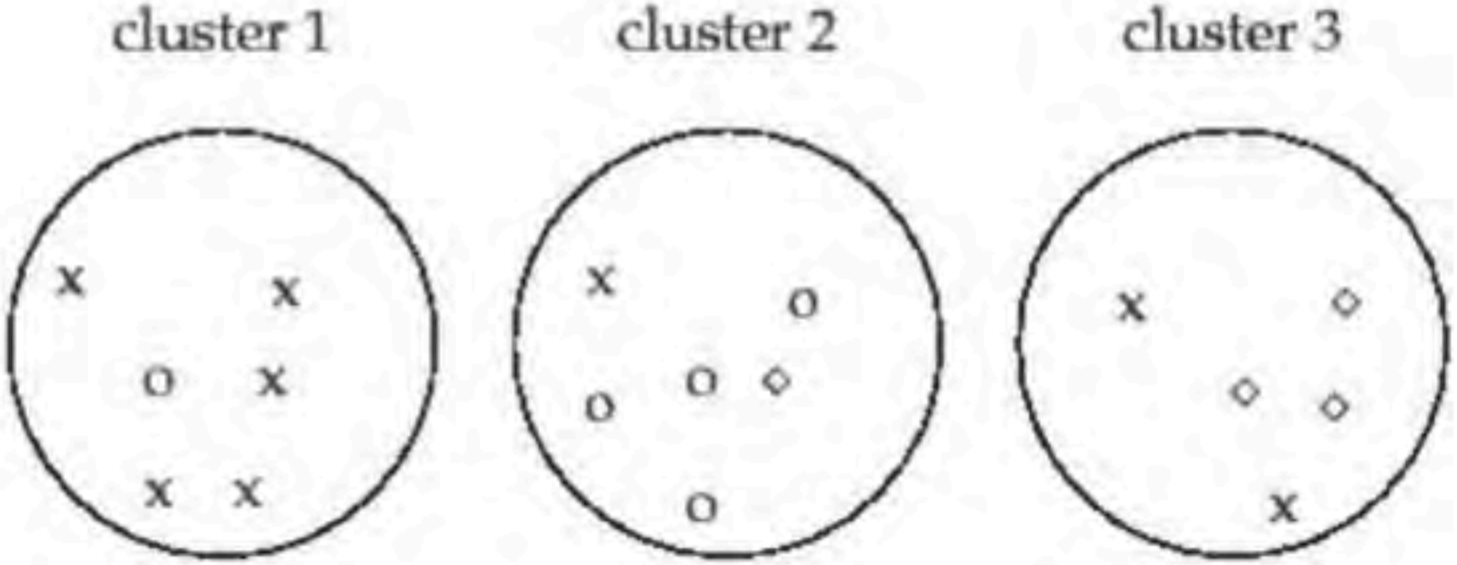
Intra-cluster distances are minimized

Inter-cluster distances are maximized

Image credits: Matteo Pardo

Purity

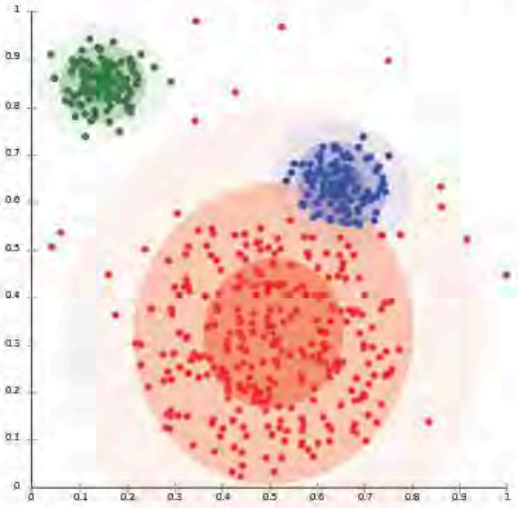
- ***Purity***:
  - assign each cluster to the class which is most frequent;
  - measure the accuracy by counting the number of correctly assigned samples per class.
  - problem: purity=1 if each cluster contain just one sample!

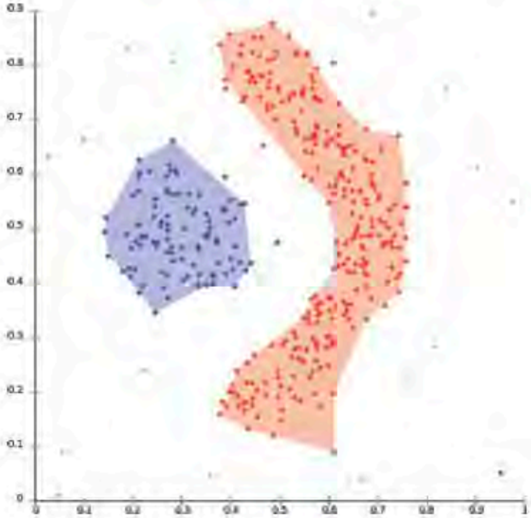


► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ◇, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

Types of Clustering

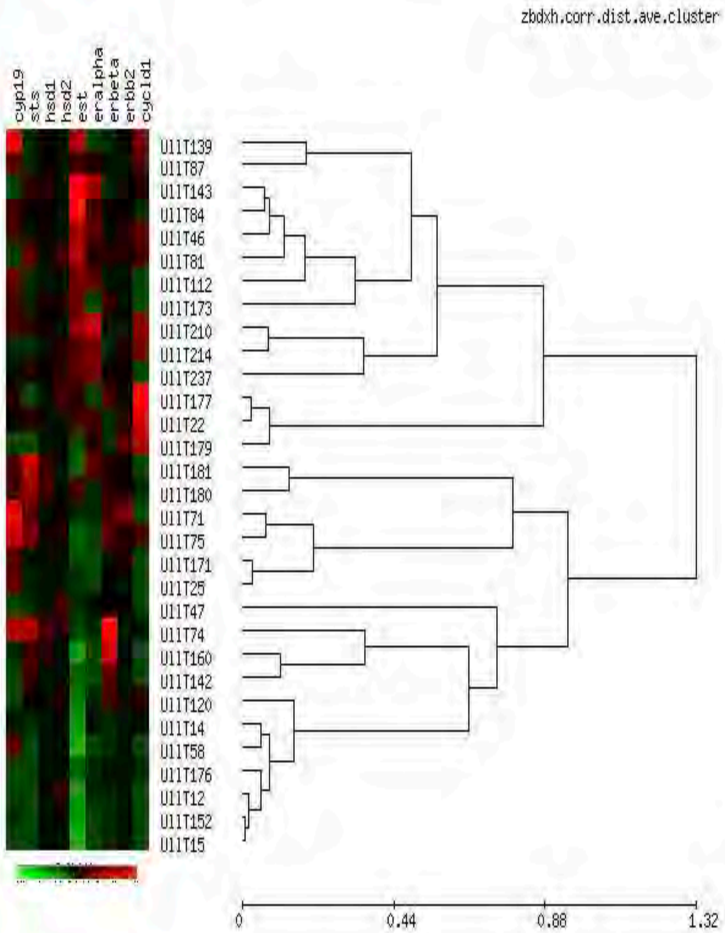
- MODEL BASED: assumes that the data were generated by a model and tries to recover the original model from the data (e.g., EM).
- DENSITY BASES: clusters are defined as areas of higher density. Objects in these sparse areas are usually considered to be noise and border points (e.g., DBSCAN).





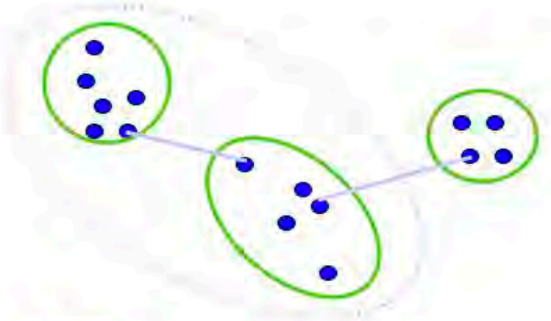
Hierarchical Clustering

- Find subsequent clusters using previously established ones.
- Cannot test all possible trees.
- Agglomerative (bottom-up): we start with each element in a separate cluster and merge them accordingly to a given property.
- Divisive (top-down): start with all the points in the same clusters and then divide them.

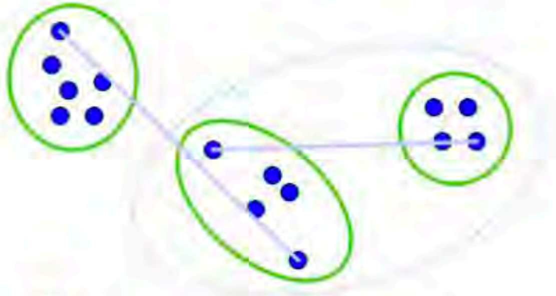


Distance between clusters

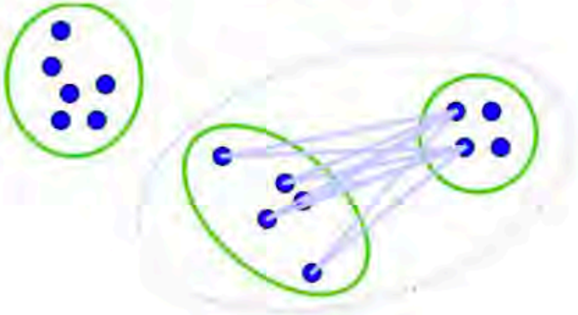
- Single link
  - smallest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = min(t_{ip}, t_{jq})$



- Complete link
  - largest distance between an element in one cluster and an element in the other
    $dis(K_i, K_j) = max(t_{ip}, t_{jq})$



- Average
  - average distance between an element in one cluster and an element in the other
  - i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$



- Centroid, Medoid

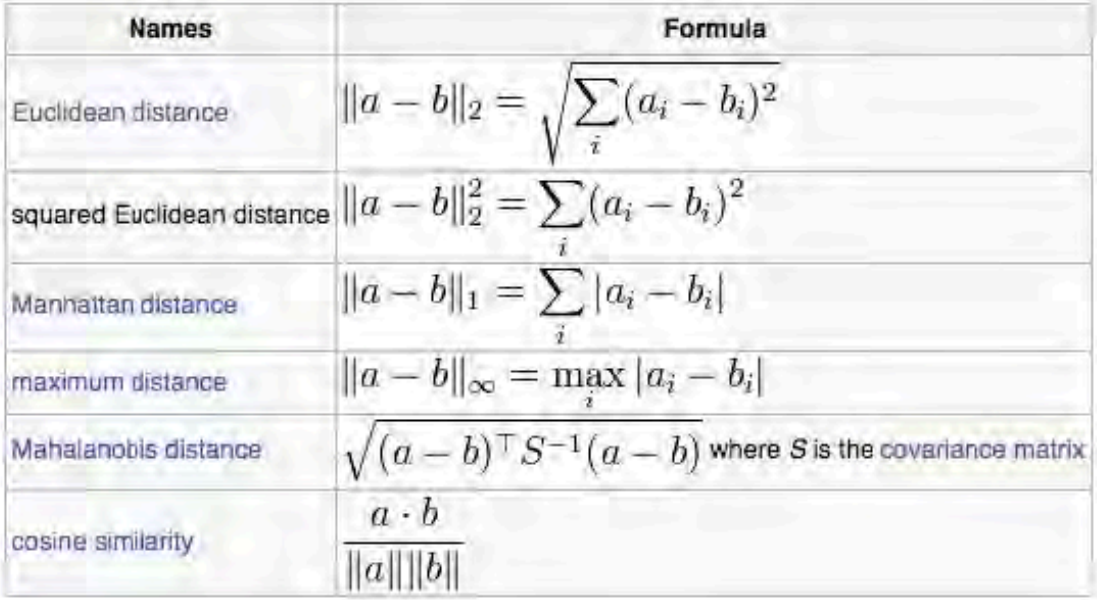
Ref: Data Mining: Concepts and Techniques, J. Han, M. Kamber
Image credit: Henry Lin

Similarity Measures

- Determine the similarity between two clusters and the shape of the clusters.
- Based on distance metrics.
- Distance metric: defines a distance between elements of a set:
  - non negativity: $dist(x,y) \geq 0$
  - symmetry: $dist(x,y) = dist(y,x)$
  - self-similarity: $dist(x,y)=0 \Leftrightarrow x=y$
  - triangular inequality: $dist(x,z) \leq dist(x,y) + dist(y,z)$

Common Distances

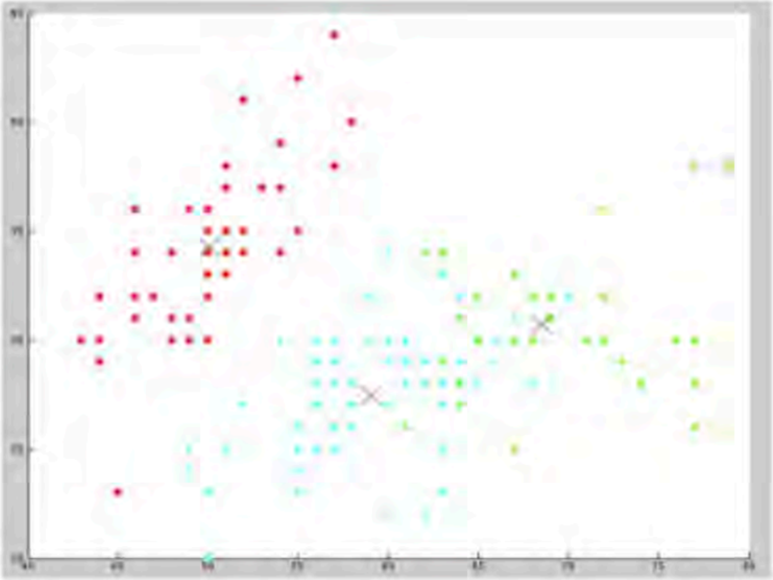
- Euclidian Distances
  - commonly used;
  - sphere shaped clusters;
  - Squared Euclidean Distance is not a metric as it does not satisfy the triangle inequality, however it is frequently used in optimization problems.



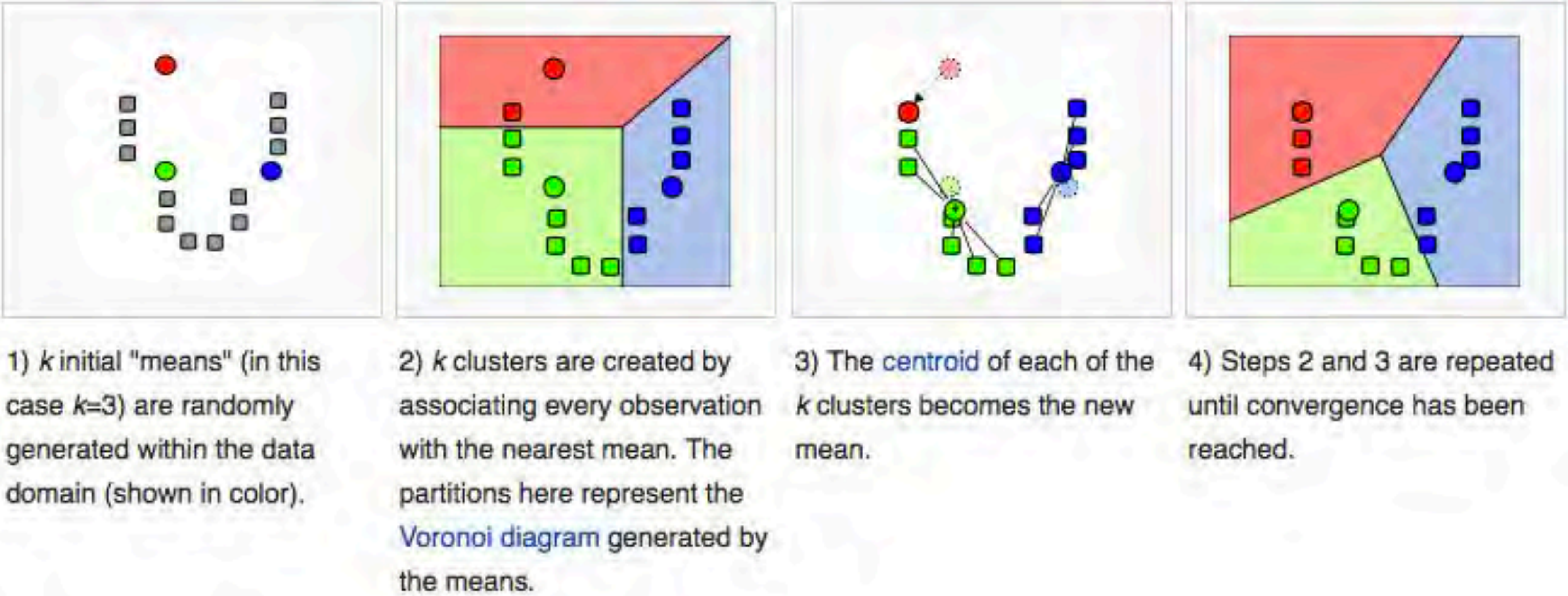
Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\|a\|\|b\|}$

K-Means

- Unsupervised clustering method
    - – partitioning algorithm
- Partition the data into k clusters, based on their features and distance from the centroids.
- Each cluster is defined by its centroid
- Goal: minimize the sum of the squared errors (SSE).



How it works



1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

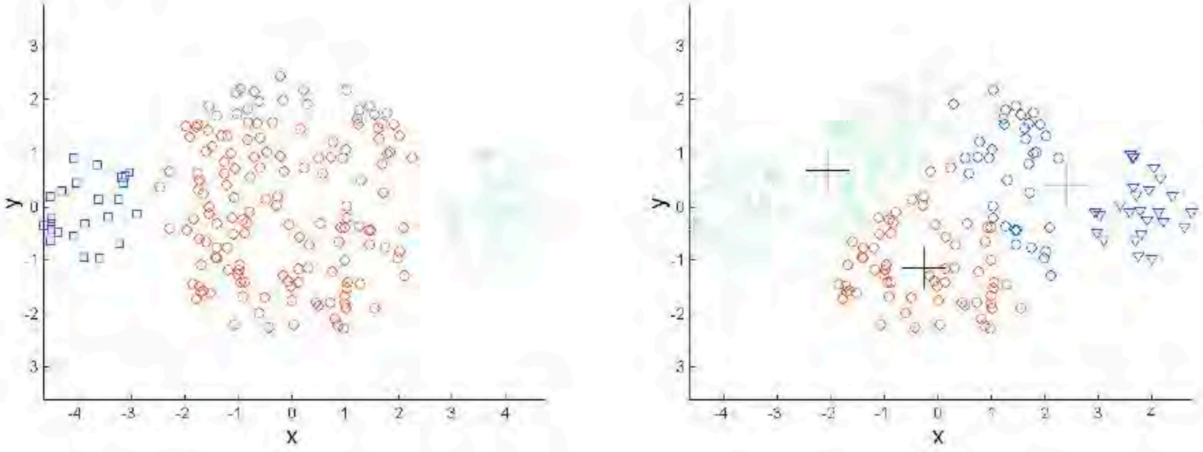
3) The centroid of each of the *k* clusters becomes the new mean.

4) Steps 2 and 3 are repeated until convergence has been reached.

Image credits: wikipedia

K-Means Pro and Cons

- Pro
  - Simple, use as benchmark
  - Relatively efficient: *O*(*t***k***n*), where *n* is the number of objects, *k* is the number of clusters, and *t* is the number of iterations.
  - easy to run multiple time with different k
- Cons
  - unable to handle noisy data
  - need to specify k
  - problems with clusters of different sizes, non-globular shapes, clusters differing in density



y
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
x
y
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
x