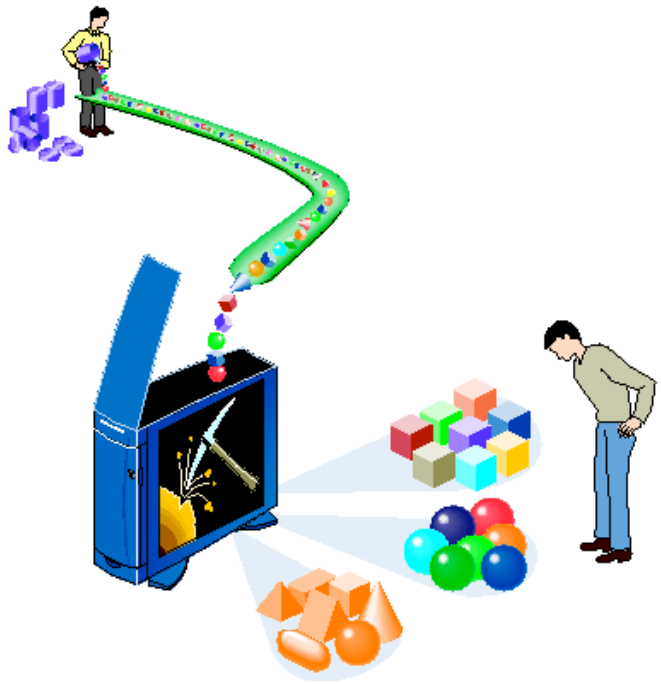
Supervised and Unsupervised Learning

Ciro Donalek

Ay/Bi 199 – April 2011

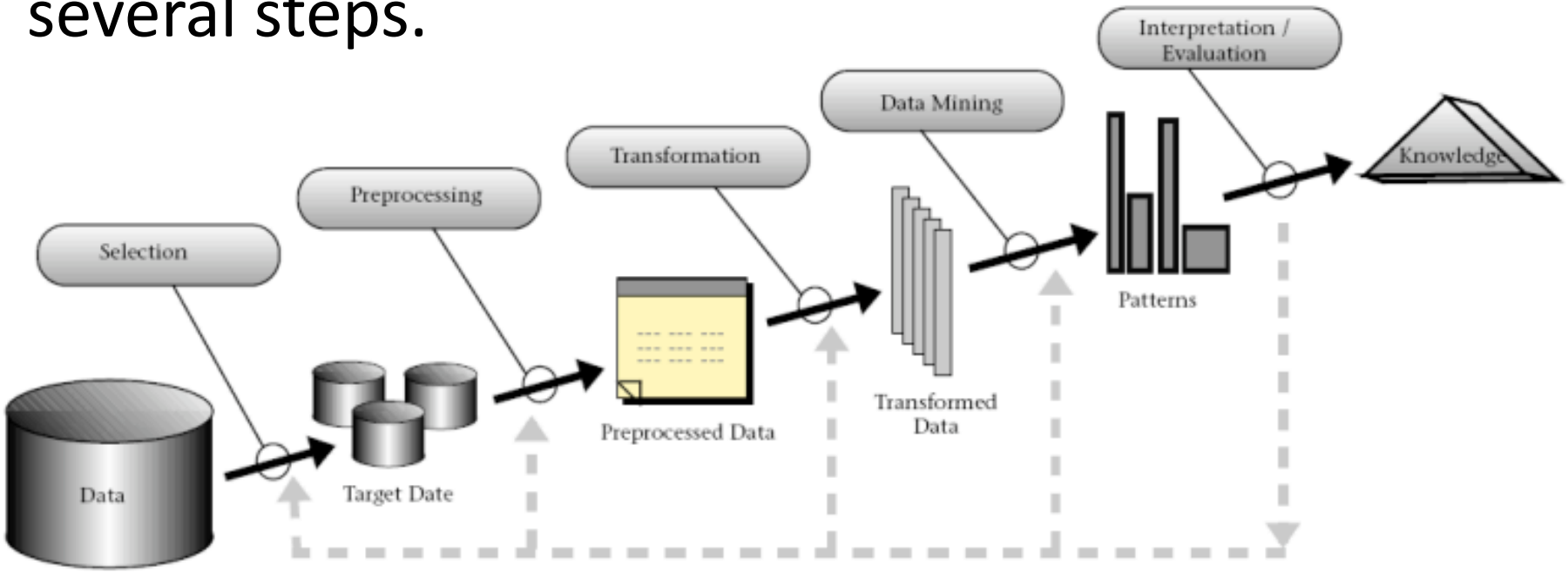
Summary

- KDD and Data Mining Tasks
- Finding the optimal approach
- Supervised Models
  - Neural Networks
  - Multi Layer Perceptron
  - Decision Trees
- Unsupervised Models
  - Kmeans
  - Self Organizing Maps
- Ensembles
- Links and References



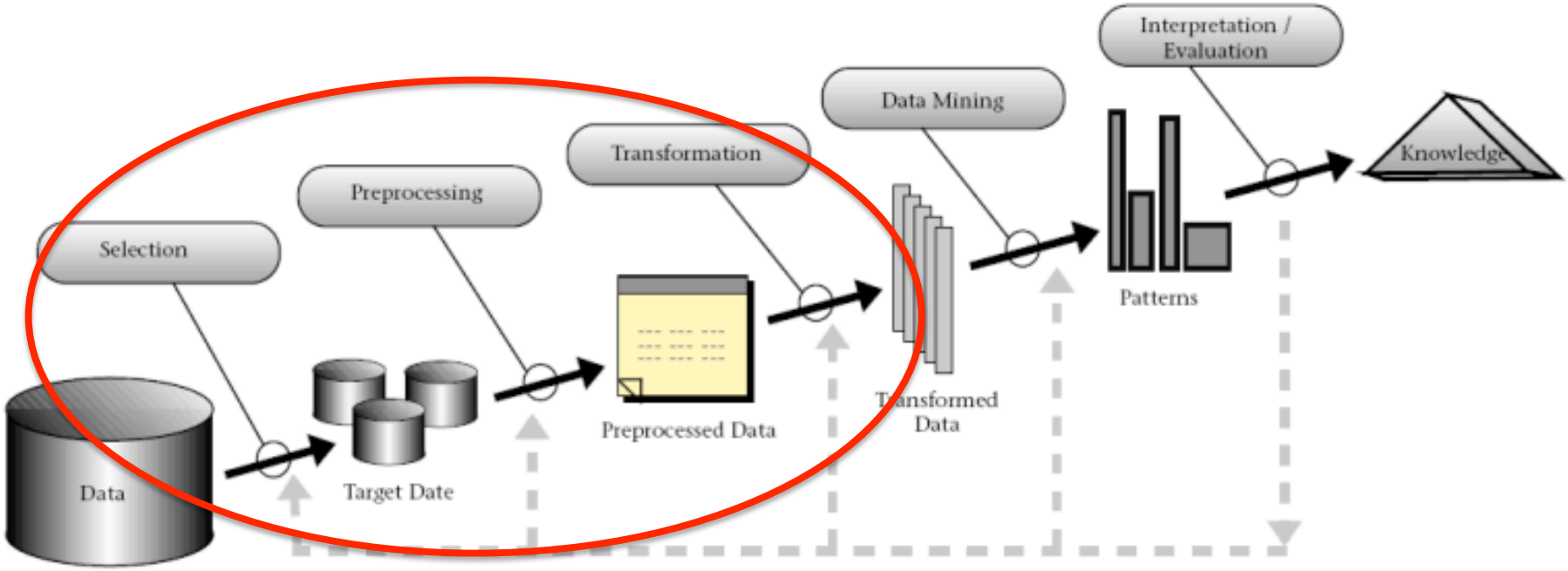
Knowledge Discovery in Databases

- KDD may be defined as: "*The non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*".
- KDD is an interactive and iterative process involving several steps.



You got your data: what's next?

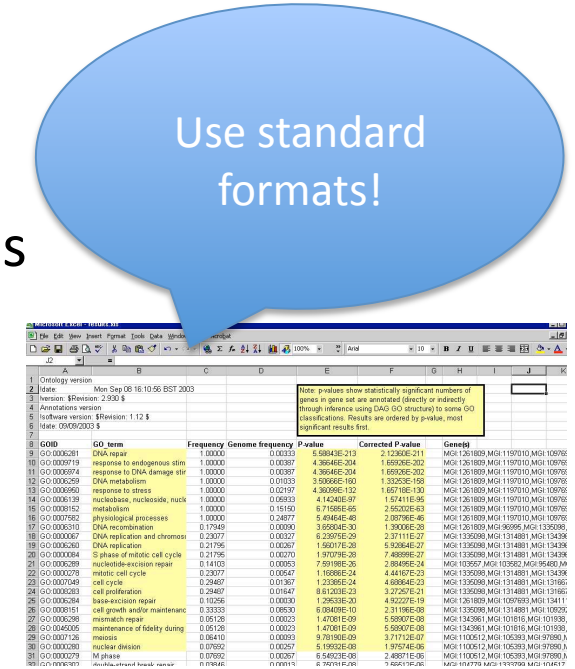
What kind of analysis do you need? Which model is more appropriate for it? …


Interpretation / Evaluation
Data Mining
Transformation
Preprocessing
Selection
Knowledge
Patterns
Transformed Data
Preprocessed Data
Target Date
Data

Clean your data!

- Data preprocessing transforms the raw data into a format that will be more easily and effectively processed for the purpose of the user.

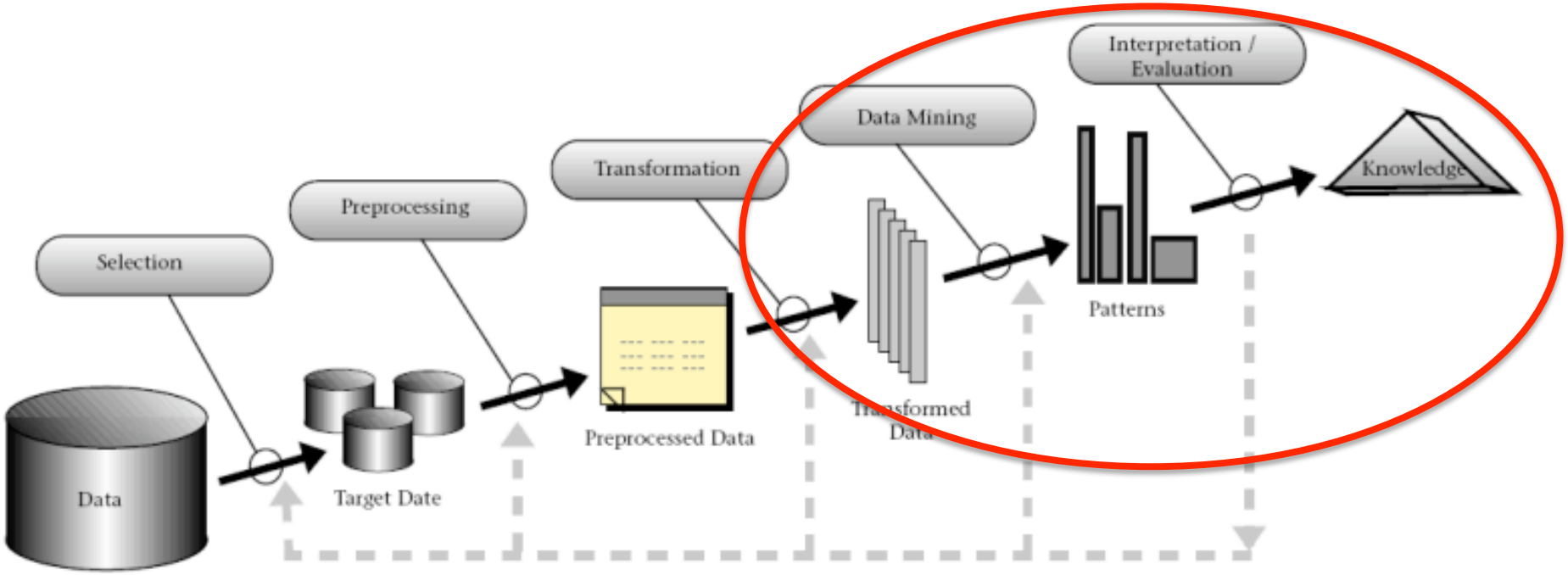
- Some tasks
  - *sampling*: selects a representative subset from a large population of data;
  - *Noise treatment*
  - strategies to handle ***missing data****:* sometimes your raws will be incomplete, not all parameters are measured for all samples.
  - *normalization*
  - *feature extraction*: pulls out specified data that is significant in some particular context.



Missing Data

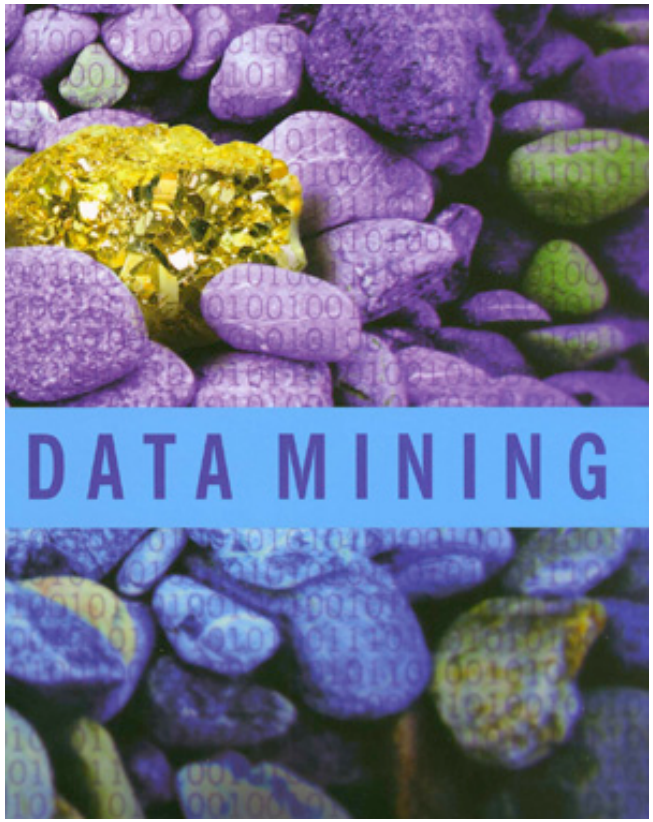
- Missing data are a part of almost all research, and we all have to decide how to deal with it.
- Complete Case Analysis: use only raws with all the values
- Available Case Analysis
- Substitution
  - Mean Value: replace the missing value with the mean value for that particular attribute
  - Regression Substitution: we can replace the missing value with historical value from similar cases
  - Matching Imputation: for each unit with a missing y, find a unit with similar values of x in the observed data and take its y value
  - Maximum Likelihood, EM, etc

- Some DM models can deal with missing data better than others.


Interpretation / Evaluation
Data Mining
Transformation
Preprocessing
Selection
Knowledge
Patterns
Transformed Data
Preprocessed Data
Target Date
Data

Data Mining

- Data Mining is about automating the process of searching for patterns in the data.
- More in details, the most relevant DM tasks are:
  - association
  - sequence or path analysis
  - **clustering**
  - **classification**
  - **regression**
  - visualization



Finding Solution via Purposes

- What kind of analysis do you need?

- Regression
  - predict new values based on the past, inference
  - compute the new values for a dependent variable based on the values of one or more measured attributes
- Classification:
  - divide samples in classes
  - use a trained set of previously labeled data
- Clustering
  - partitioning of a data set into subsets (clusters) so that data in each subset ideally share some common characteristics

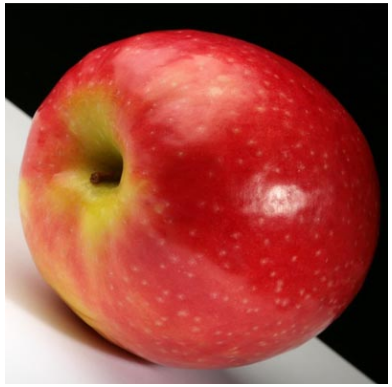
- Classification is in a some way similar to the clustering, but requires that the analyst know ahead of time how classes are defined.

Cluster Analysis

starshadow78/Flickr

How many clusters do you expect?

starshadow78/Flickr



starshadow78/Flickr

starshadow78/Flickr

Search for Outliers

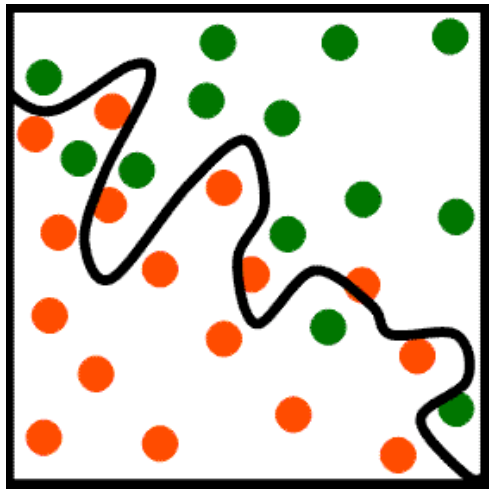
starshadow78/Flickr

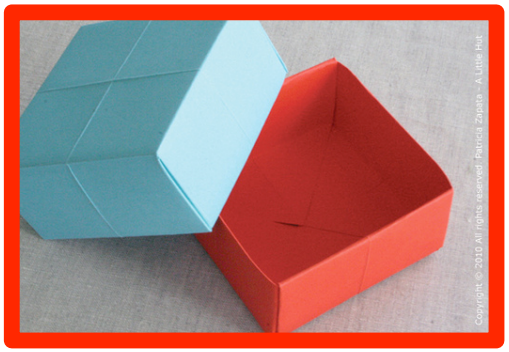
Classification

- Data mining technique used to predict group membership for data instances. There are two ways to assign a new value to a given class.

- **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
  - 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)

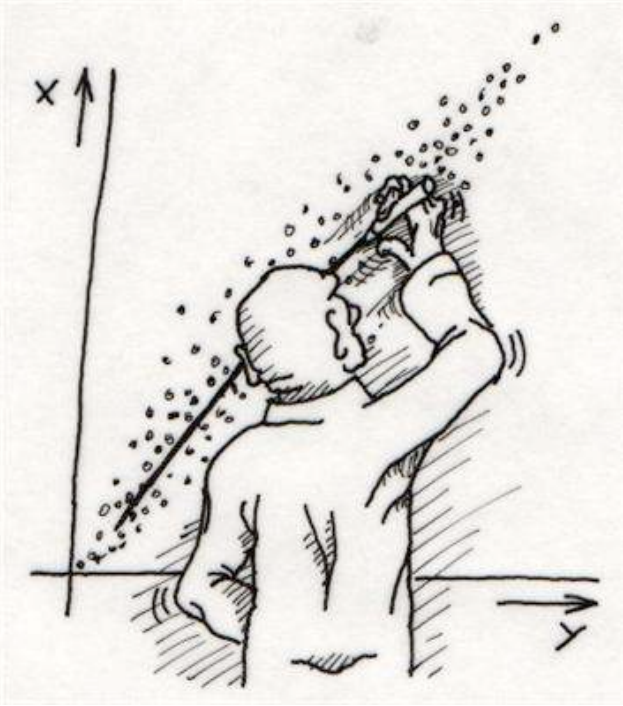




starshadow78/Flickr

Regression / Forecasting

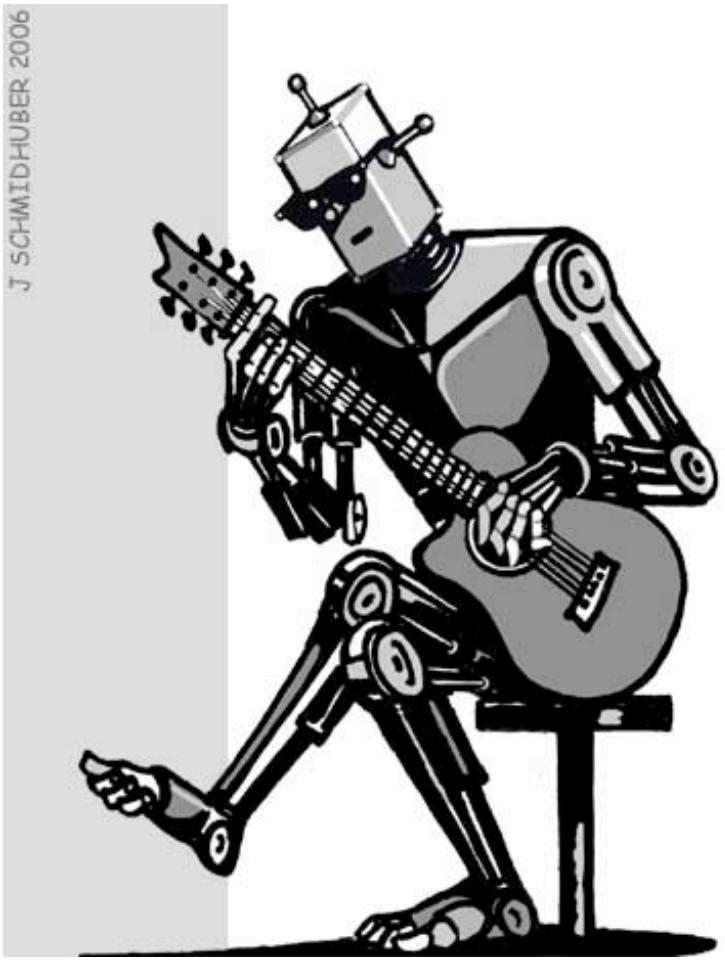
- Data table statistical correlation
  - mapping without any prior assumption on the functional form of the data distribution;
  - machine learning algorithms well suited for this.
- Curve fitting
  - find a well defined and known function underlying your data;
  - theory / expertise can help.



Machine Learning

- To learn: *to get knowledge of by study, experience, or being taught*.

- Types of Learning
  - Supervised
  - Unsupervised



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.
- Cluster significance and labeling.
- The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes.

Supervised Learning

- Training data includes both the input and the desired results.
- For some examples the correct results (targets) are known and are given in input to the model during the learning process.
- The construction of a proper training, validation and test set (Bok) is crucial.
- These methods are usually fast and accurate.
- Have to be able to **generalize**: give the correct results when new data are given in input without knowing a priori the target.

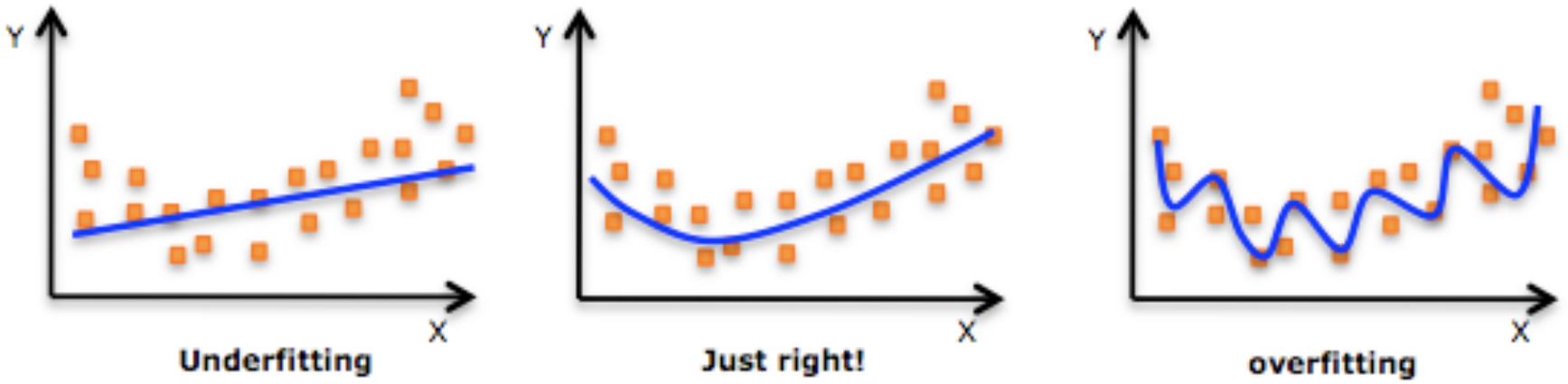
Generalization

- Refers to the ability to produce reasonable outputs for inputs not encountered during the training.

In other words: NO PANIC when "never seen before" data are given in input!

A common problem: OVERFITTING

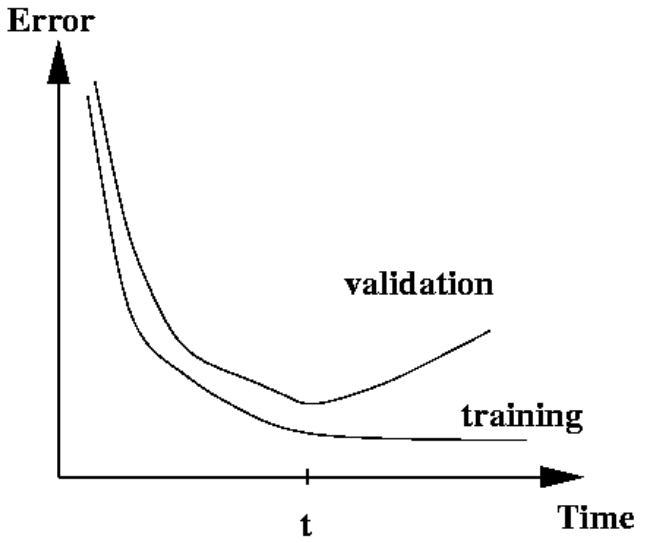
- Learn the “data” and not the underlying function
- Performs well on the data used during the training and poorly with new data.



Use proper training sets, early stopping.

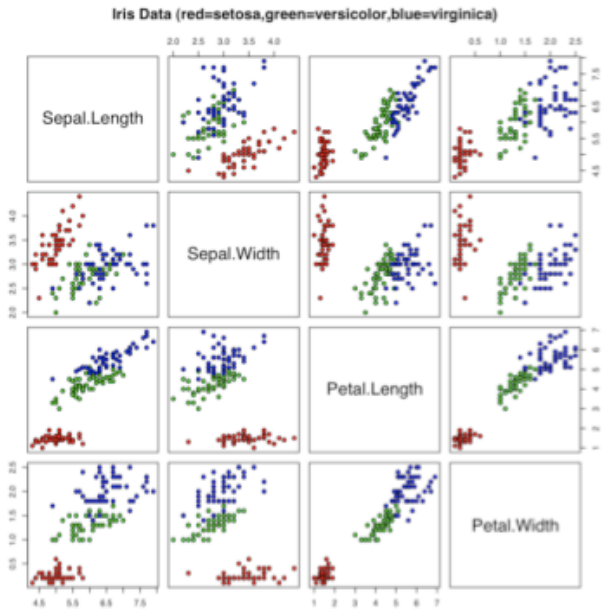
Datasets

- **Training set**: a set of examples used for learning, where the target value is known.
- **Validation set**: a set of examples used to tune the architecture of a classifier and estimate the error.
- **Test set**: used only to assess the performances of a 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.



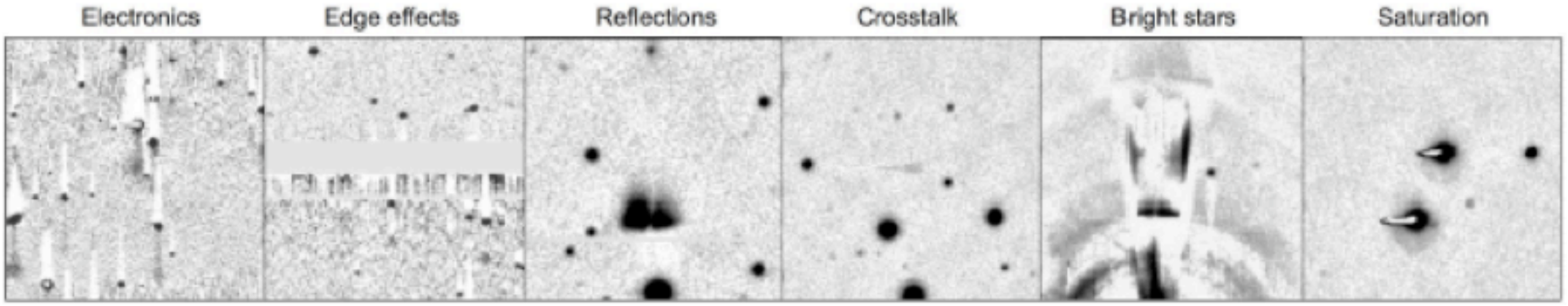
IRIS dataset

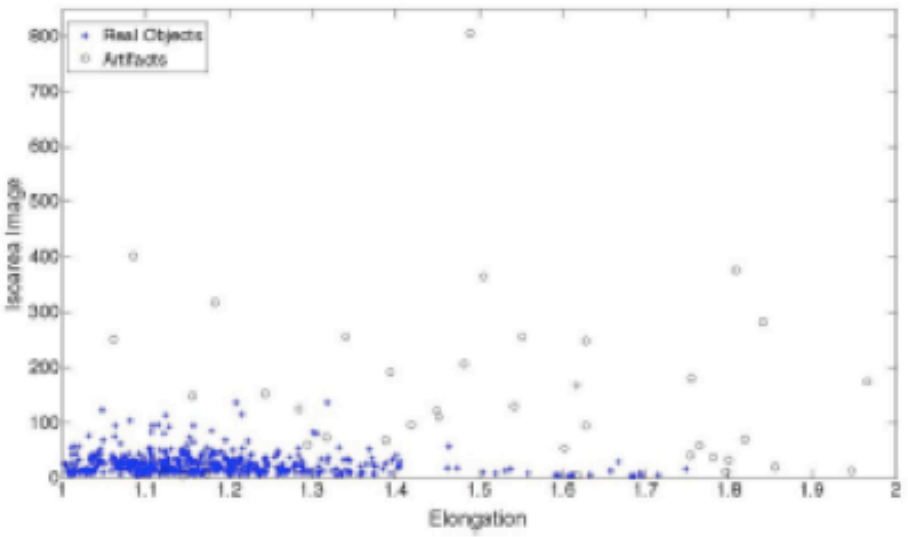
- IRIS
  - consists of 3 classes, 50 instances each
  - 4 numerical attributes (sepal and petal length and width in cm)
  - each class refers to a type of Iris plant (Setosa, Versicolor, Verginica)
  - the first class is linearly separable from the other two while the 2nd and the 3rd are not linearly separable

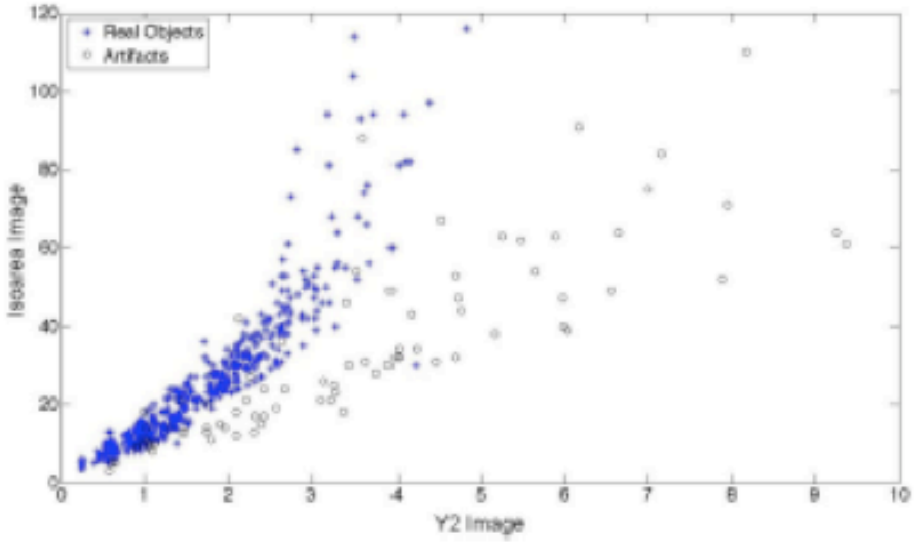


Artifacts Dataset

- PQ Artifacts
  - 2 main classes and 4 numerical attributes
  - classes are: true objects, artifacts







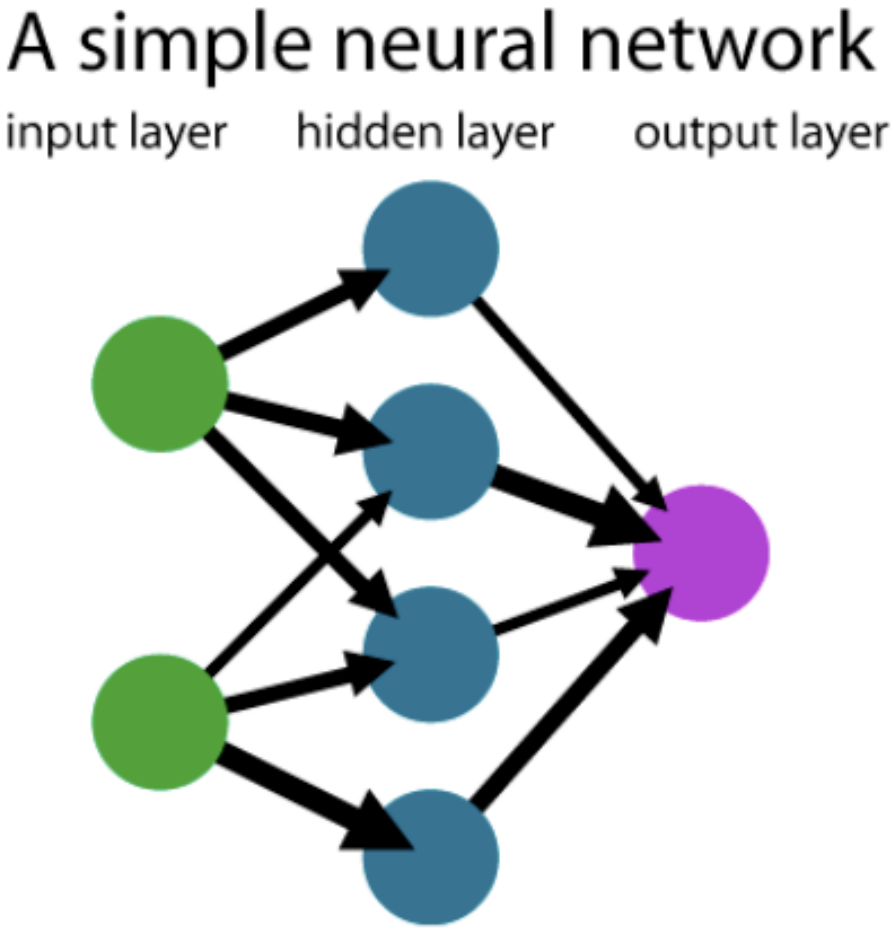
Data Selection

- "**Garbage in, garbage out** ": training, validation and test data must be representative of the underlying model
- All eventualities must be covered
- Unbalanced datasets
  - since the network minimizes the overall error, the proportion of types of data in the set is critical;
  - inclusion of a loss matrix (Bishop,1995);
  - often, the best approach is to ensure even representation of different cases, then to interpret the network's decisions accordingly.

Artificial Neural Network

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

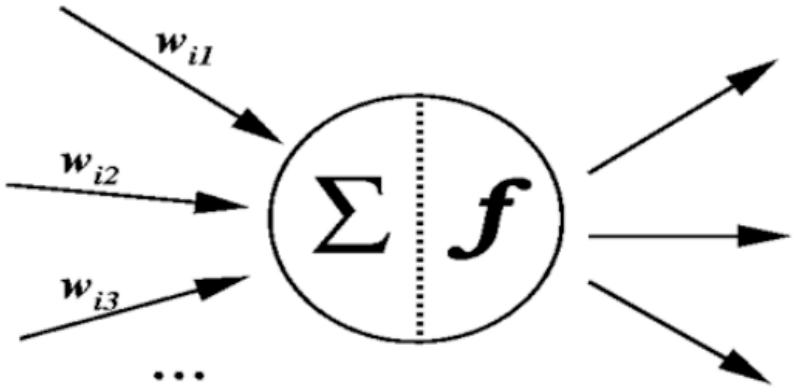
"a large number of highly interconnected simple processing elements (neurons) working together to solve specific problems"



A simple artificial neuron

- The basic computational element is often called a node or unit. It receives input 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:

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



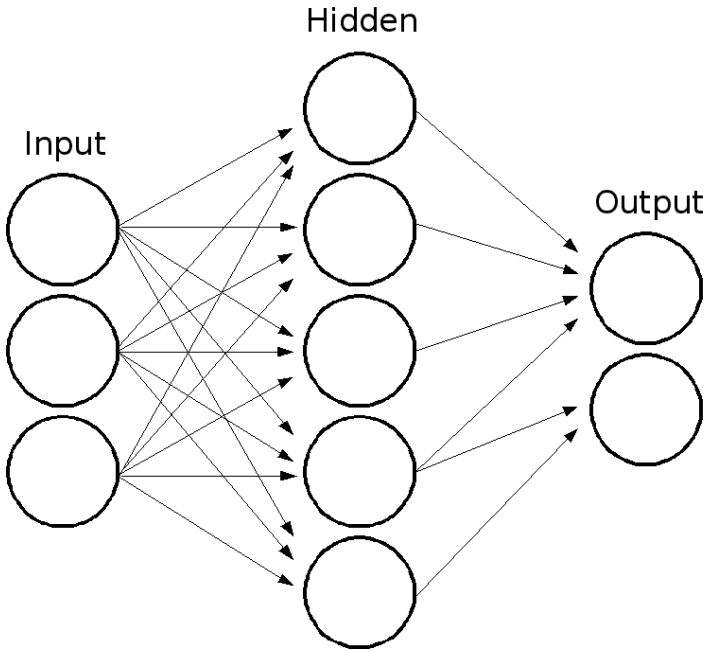
$$y_i = f(net_i)$$

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.
The values of the functions associated with the connections are called "weights".

The whole game of using NNs is in the fact that, in order for the network to yield appropriate outputs for given inputs, the weight must be set to suitable values.

The way this is obtained allows a further distinction among modes of operations.



Neural Networks: types

Feedforward: Single Layer Perceptron, MLP, ADALINE (Adaptive Linear Neuron), RBF

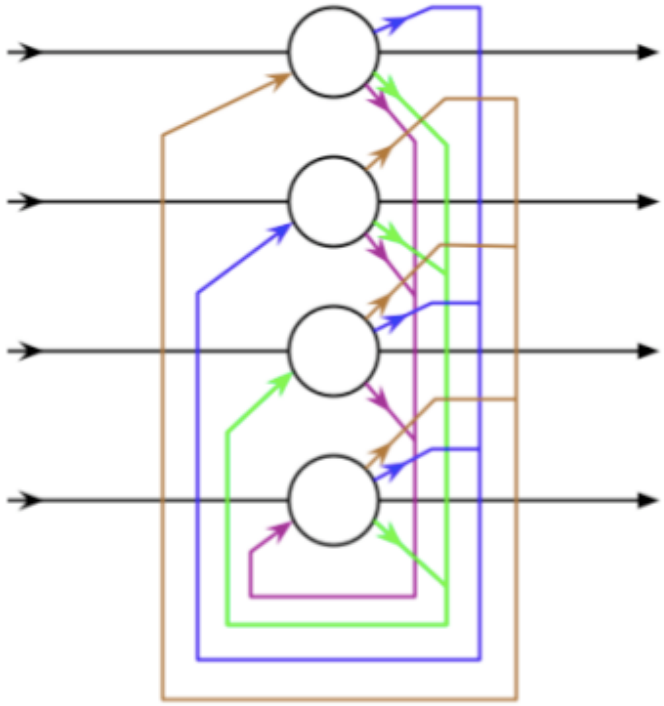
Self-Organized: SOM (Kohonen Maps)

Recurrent: Simple Recurrent Network, Hopfield Network.

Stochastic: Boltzmann machines, RBM.

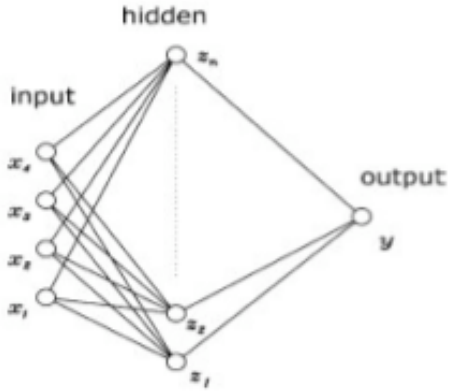
Modular: Committee of Machines, ASNN (Associative Neural Networks), Ensembles.

Others: Instantaneously Trained, Spiking (SNN), Dynamic, Cascades, NeuroFuzzy, PPS, GTM.

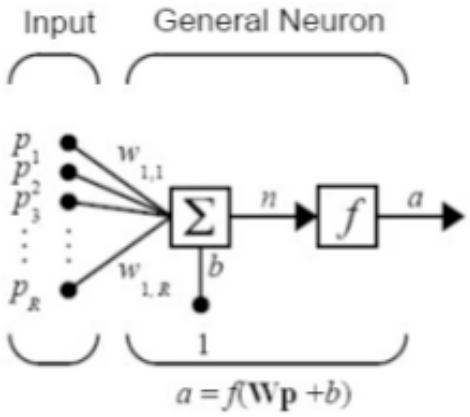


Multi Layer Perceptron

- 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.



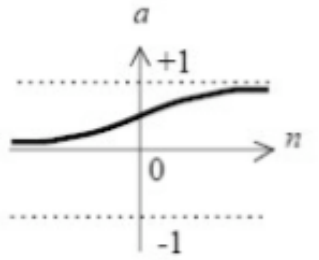
The architecture of a two layer MLP.



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

Where

R = number of elements in input vector

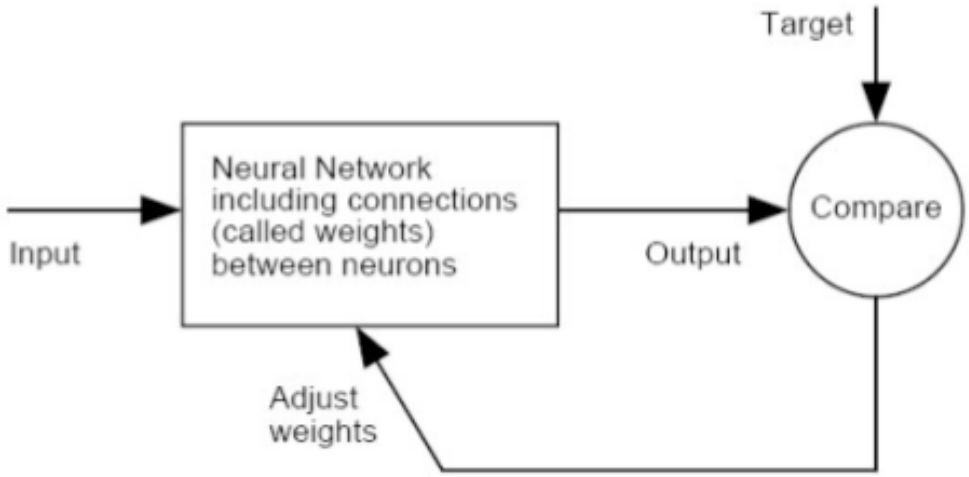


$a = logsig(n)$

Learning Process

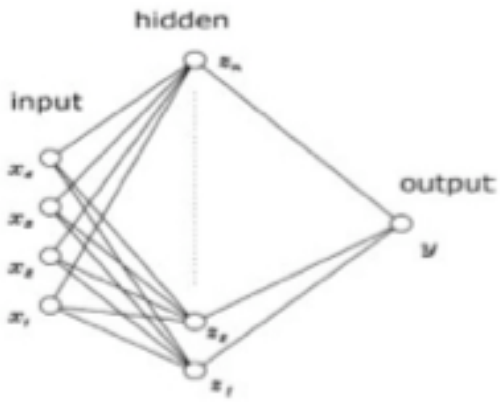
- Back Propagation
  - the output values are compared with the target to compute the value of some predefined error function
  - the error is then fedback through the network
  - using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function

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



Hidden Units

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

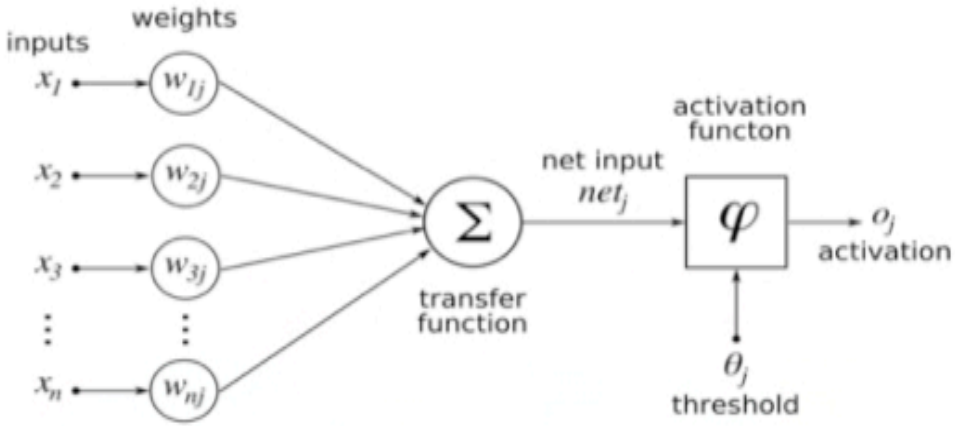


The architecture of a two layer MLP.

- 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.

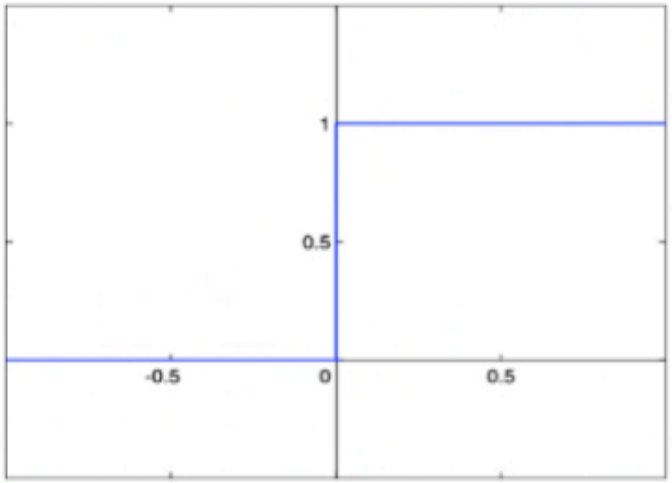
Activation and Error Functions

- **Error Functions**
  - measure of the discrepancy between the network output values and the target;
  - sum of the squared errors (SSE), cross entropy (CE), etc.
- **Activation Functions**
  - used by most units to transform their inputs;
  - needed to introduce non linearity into the network
  - linear, logistic, tanh, softmax...



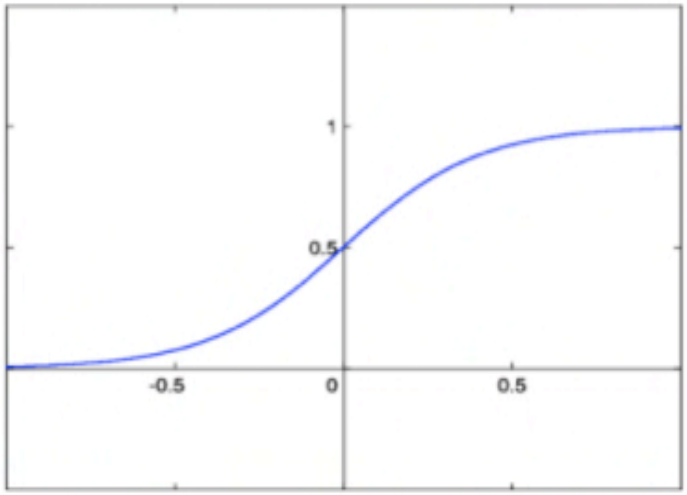
Using a Multilayer Perceptron with a softmax activation function and cross-entropy error, the network outputs can be interpreted as the conditional probabilities $p(C_1|\mathbf{x})$ and $p(C_2|\mathbf{x})$ where $\mathbf{x}$ is the input vector, C_1 the first class, C_2 the second class.

Activation Functions



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.

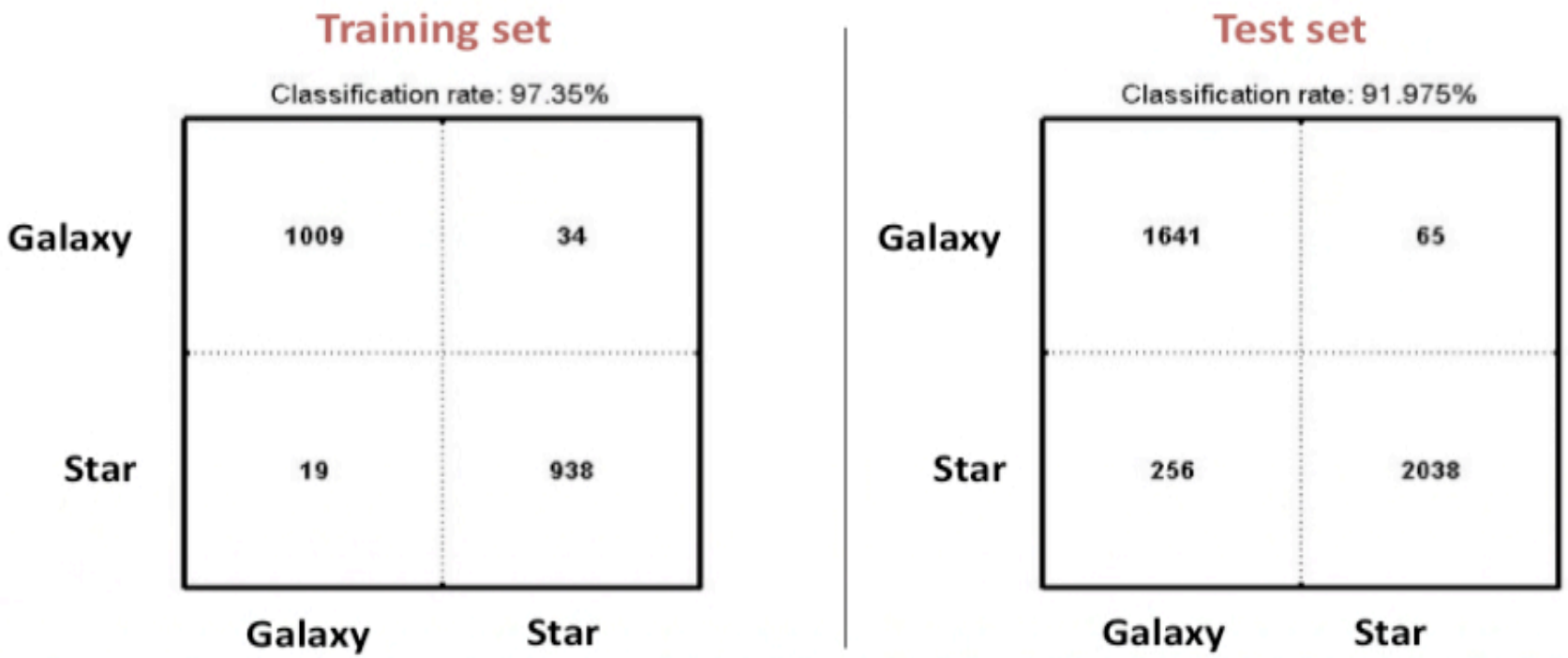


Sigmoid function
Has the property of being similar to the step function, but with the addition of a region of uncertainty.

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

Results: confusion matrix

In the confusion matrix the network prediction Y are compared with the target T: the rows represent the true classes and the columns the predicted classes.



Results: completeness and contamination

The performances of the classifiers are rated based on the following three criteria. Supposing we have 2 classes A and B:

✓ **completeness**: the percentage of objects of class A correctly classified as such;

✓ **contamination**: the percentage of objects of class A incorrectly classified as objects belonging to the class B;

✓ **classification rate**: the overall percentage of objects correctly classified.

Exercise: compute completeness and contamination for the previous confusion matrix (test set)

Decision Trees

- Is another classification method.
- A decision tree is a set of simple rules, such as "if the sepal length is less than 5.45, classify the specimen as setosa."
- Decision trees are also nonparametric because they do not require any assumptions about the distribution of the variables in each class.