Machine Learning
&
Data analytics

Workshop during
NewTech4Steel project meeting
Buttrio, 13.-14.11.2018

Machine learning – Definition

•Machine Learning (ML) is part of the field of Data Analytics

•ML is a field of artificial intelligence

ML uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed

See Wikipedia

Machine learning – Classification of learning tasks

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback.

Semi-supervised learning: The computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.

Active learning: The computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Reinforcement learning: Data (in form of rewards and punishments) are given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

Machine learning – Categorisation by desired output

Classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
In **regression**, also a supervised problem, the outputs are continuous rather than discrete.
In **clustering,** a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
Density estimation finds the distribution of inputs in some space.
Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics.

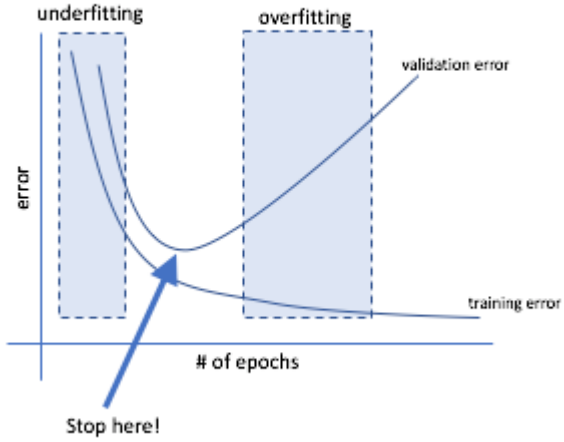
Supervised learning

Learning a function that maps an input to an output based on example input-output pairs

Regression	Classification
•Statistical regression methods •Artificial Neural Networks •Support Vector Machines	•Statistical classification methods (e.g. Discriminant Analysis, Bayes) •Artificial Neural Networks •Support Vector Machines •Decision Tree algorithms •Distance/density based classifiers (e.g. k-Nearest Neighbor, c-means)

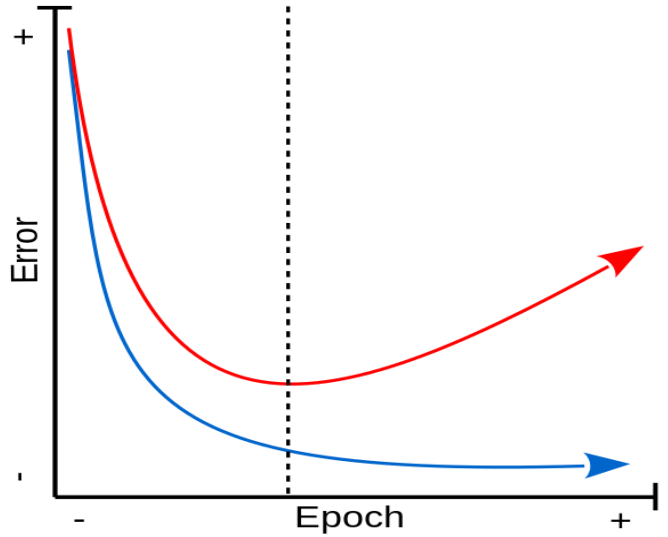
How to efficiently exploit available data?

- Build the best possible model
- Avoid overfitting
- Common point to most data driven modelling appr



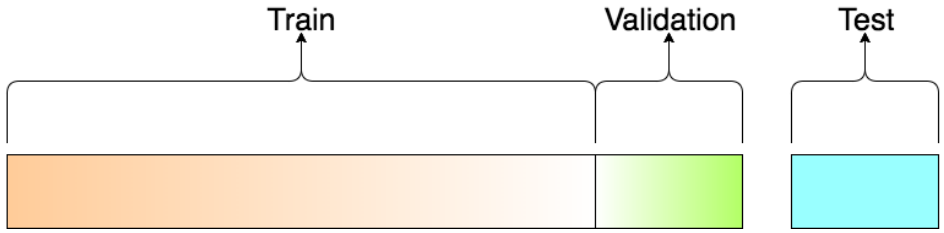
Overfitting

- It happens when a statistical (*data driven*) model adapts too much to experimental data
- Lacks in **generalization**
- Due to higher number of model parameters with respect to data samples
- Limits <u>in practice</u> the usefulness of a model

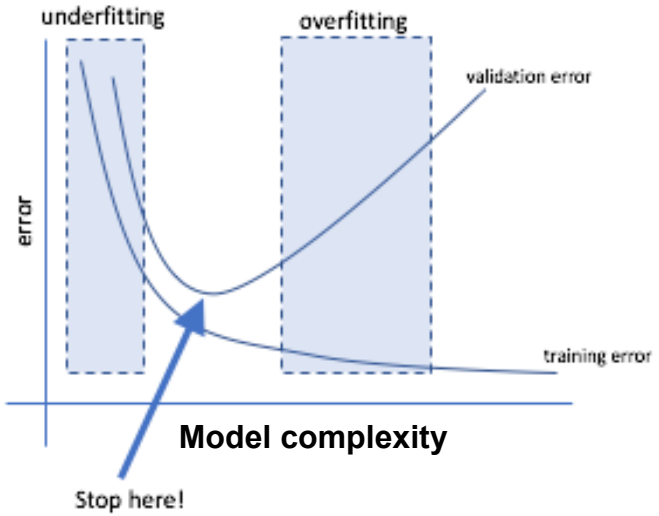


An efficient split of available data

The idea is to monitor the model performance, **during the training**, on data **not used for parameters tuning** and to use this information to maximize model performance.



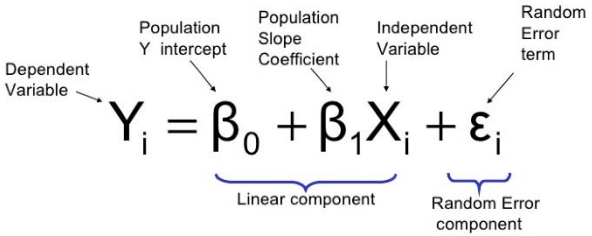
- Training data are used to tune models parameters all through the training
- Validation data might be used to define model hyperparameters or decide when to stop the training
- Test data are used for assessing real model performance

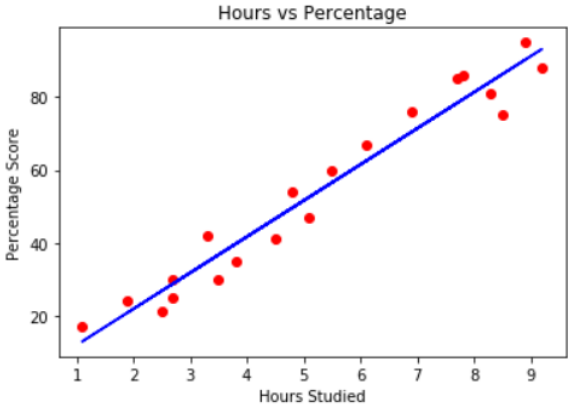


Regression analysis aims at making explicit the relation among one (target) dependent variable and its (independent) predictors.

- Basic method yet useful and used

Univariate linear regression





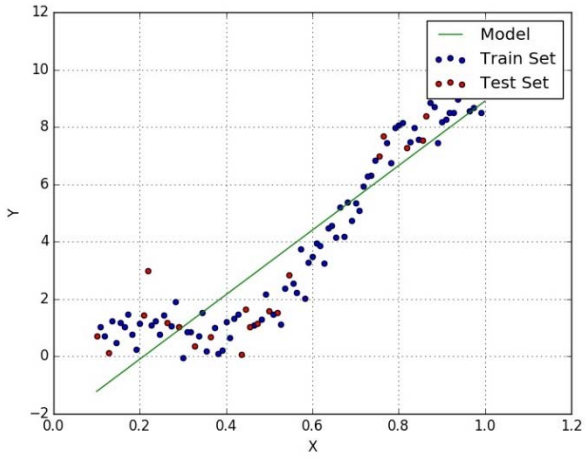
Determination of regression coefficients

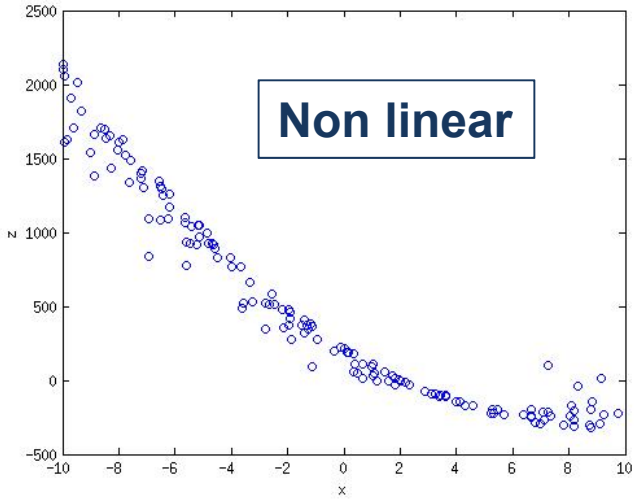
- Depends on the number of parameters (coefficents) (see the multivariate case)
  - More coefficients than observation → nope
  - Same number →linear system, unique solution
  - More observations that coefficients → optimization: **Least Square**

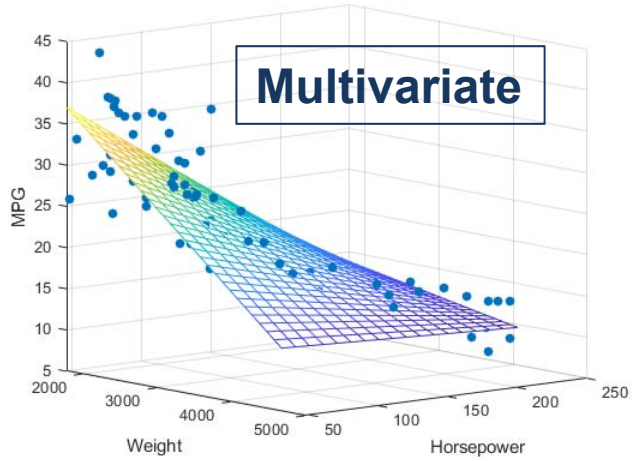
Not only linear! Go complex!

Other kinds of regression

- Multivariate
- Non-linear
- Overcome some of the limitation of linear regression
- Similar solving approach (*provided a model)*



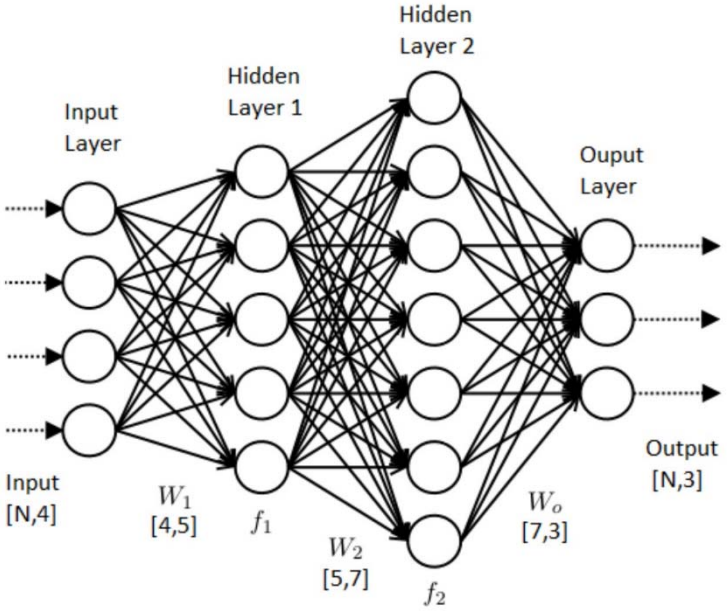




What is a neural network?

Mathematical model formed by a set of connected articifical neurons, organized in layers, inspired by the human brain

- Universal function approximator
- Robust wrt noise, outliers
- Generalization capabilities



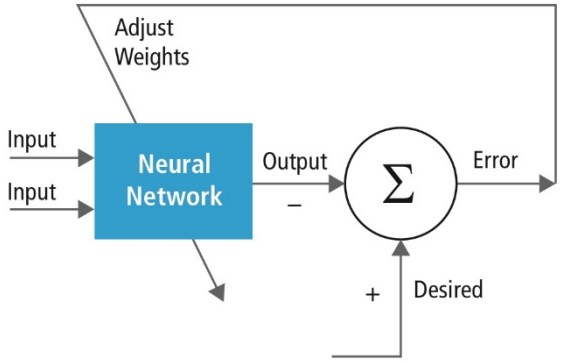
Signals flow through connected neurons

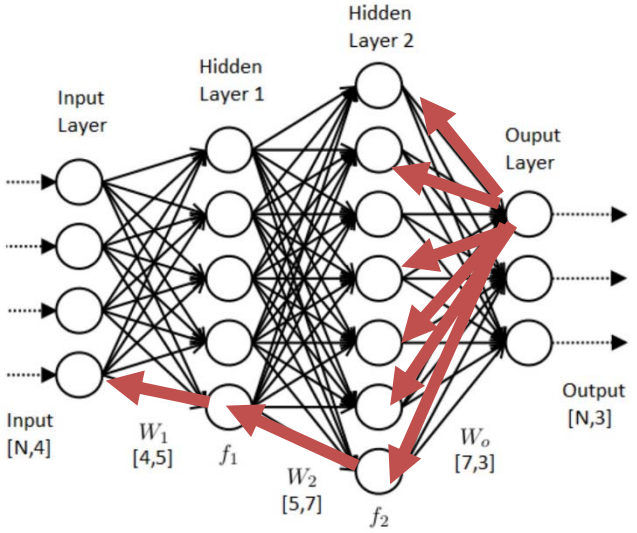
- Incoming signals are weighted and summed up
- Outgoing signal via a activation function (sigmoids…)

$$f(x) = k\left(\sum_i w_i g_i(x)\right)$$

Training a neural network

- Gradient descent based approach
- Back propagation: error is propagated backward in the network during the training epochs
- Until desired error is achieved, stall situations, maximum epochs reached or overfitting occurs



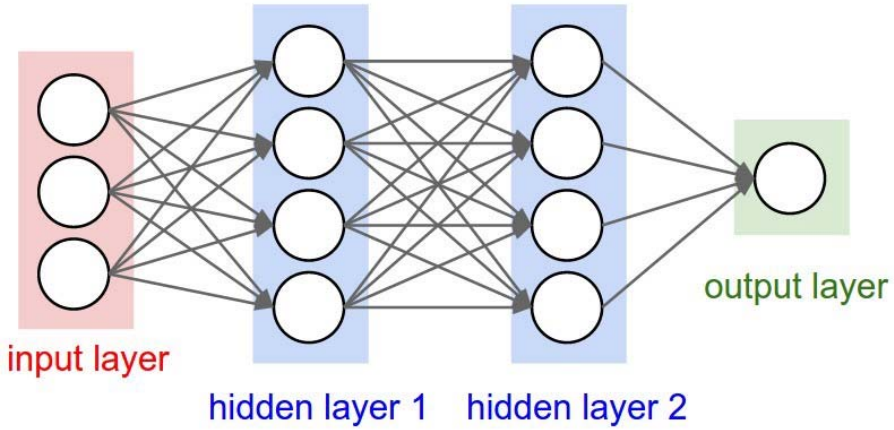


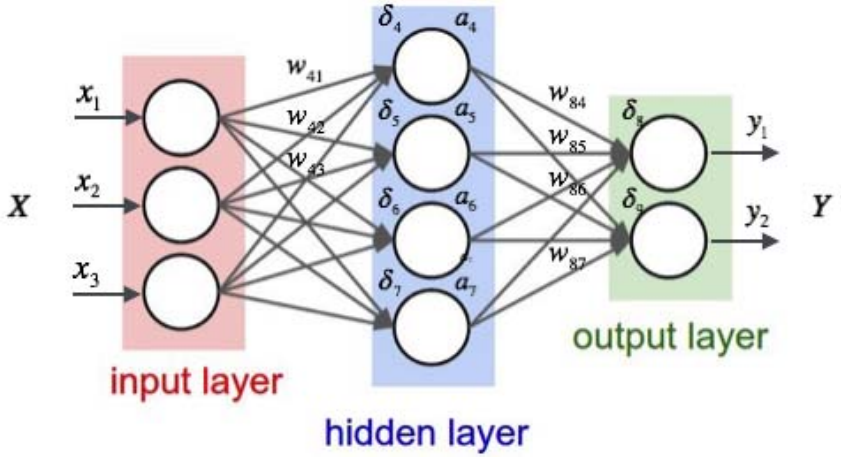
- Different paradigms of Back-Propagation exist: aiming at speed, generalization capabilities, robustness..

Approaches for the use of ANN in classification tasks:

- Single ANN output
  - Coded class
  - Ordinal relation among classes
- Multiple output neurons
  - One per class
  - A flavour of membership probability
  - Similarity among classes/samples

CONS: it's a black box



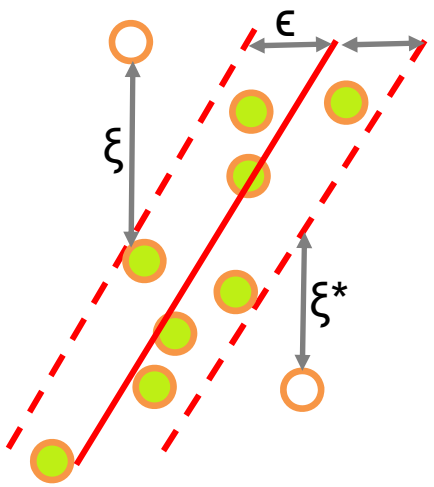


Some ANN based models in steel-making applications

- MLP model to predict the pig iron temperature at a blast furnace:
  A lot of process data of the BF are used to predict the temperature of the next tapping
  The model is re-trained with data of the last 3 months, if results are better => taking the new model, otherwise use the old model.

- Support Vector Machines

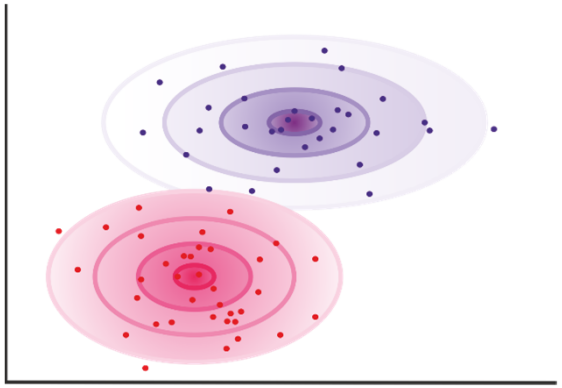
The regression function tries to minimize the sum of ϵ + ξ (simplified).



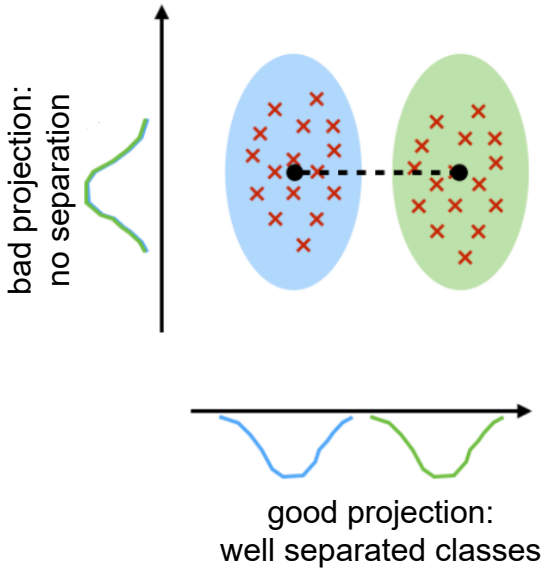
Non-linear regression is done by mapping the input data into a m-dimensional feature space where a linear model can be constructed.
(see Supervised Learning – Classification)

Machine learning – Supervised learning - Classification

- Statistical classification methods
  - (Naïve) Bayes classifier
    Basing on maximum likelihood theory, Bayes classifier offers a simple and good solution when members of the classes belong to Gauss distributions.

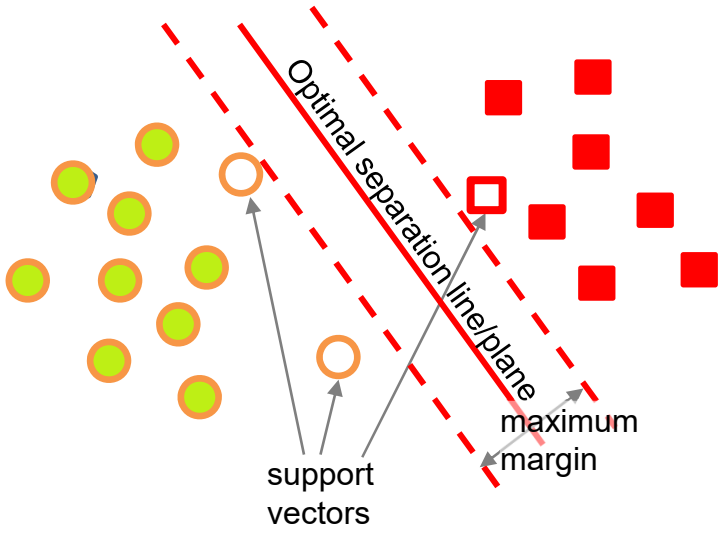


  - Discriminant (function) Analysis
    DA is used to determine which variables discriminate between two or more groups.



Since their calculation speed both regression methods are suitable to be used as fitting functions for Genetic algorithms !

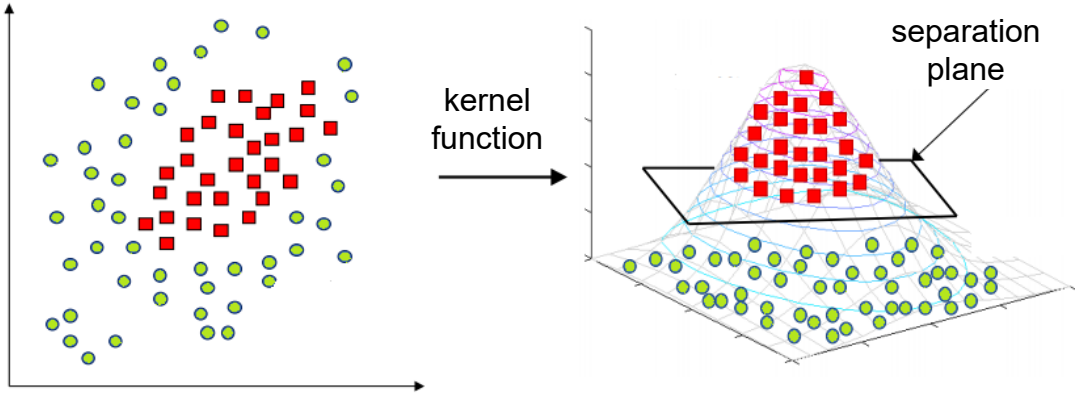
- Support Vector Machines



The separation plane is constructed using only the support vectors.

The optimization target is to find a plane which maximises the margin between the different classes.

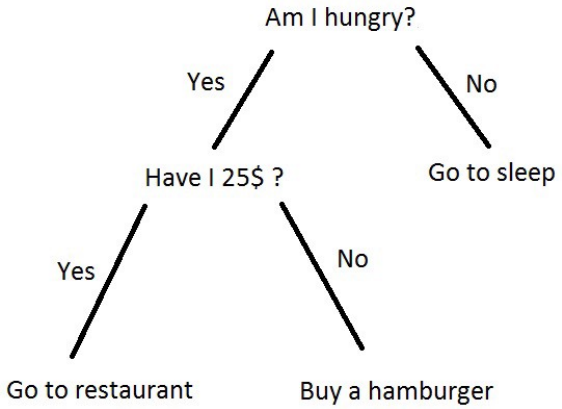
The problem of non-linear separation is solved by transforming input data into higher dimensions where linear separation is possible.

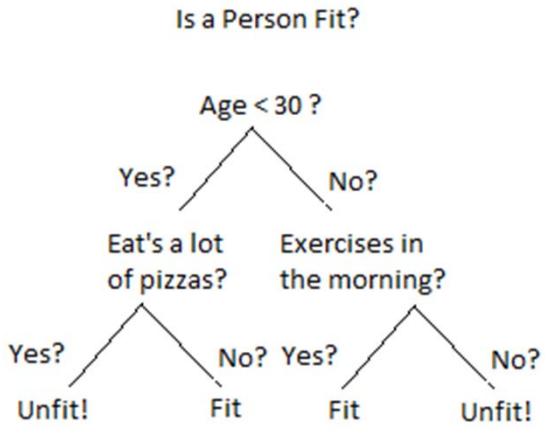


*often called **kernel trick***

Decision Trees

is a directed graph with a root and a defined path to each node or leaf





Predictive model, used for classification and decision making where:

- Each internal node represents a variable
- An outgoing arch a possible value for that variable
- Each leaf the prediction for the target variable given previous variables values

The path from the root to the leaf represents the decision making process.

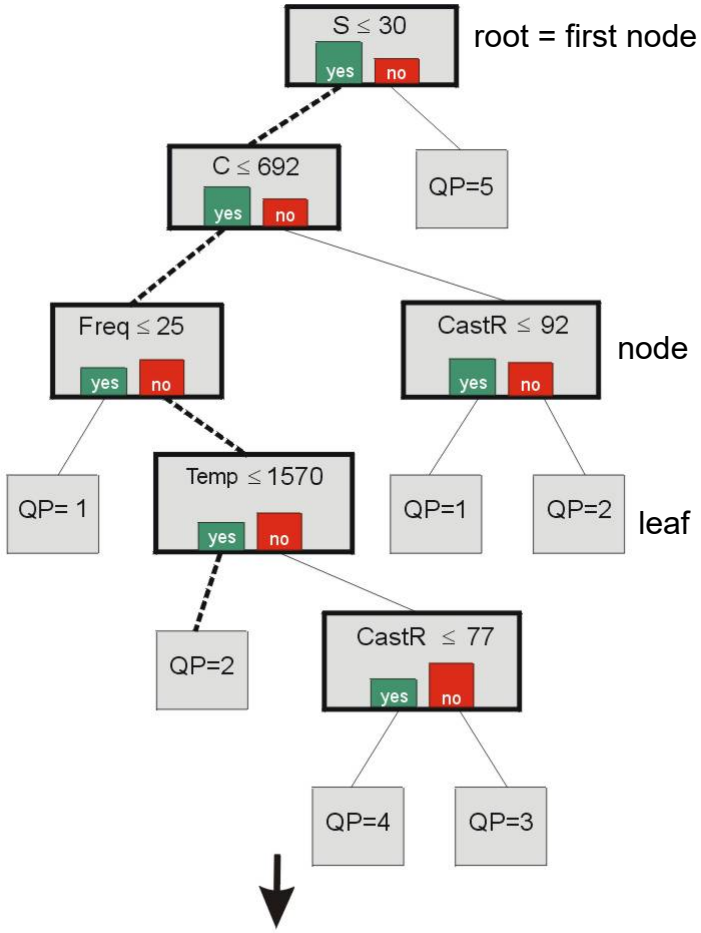
PROs:

- interpretable

CONs:

- Not so robust
- Problems may arise when coping with numeric data

- When using numeric data the tree becomes often a lot of single leafs, so interpretation and generalization becomes bad.
- But sometimes the tree offers new insights into the dependencies between independent variables and the target parameters.



IF S<30 AND C<692 AND Freq>25 AND TEMP<1570 THEN QP=2

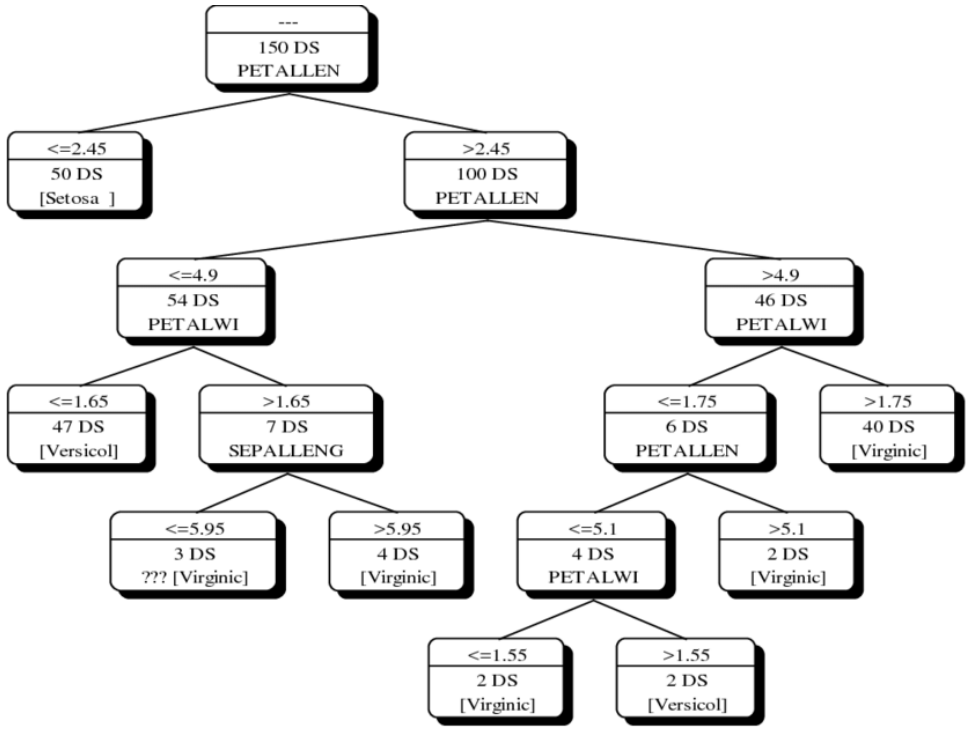
Decision Trees

Build the tree

- Exploit labelled data
- At each step the most «splitting» variable is selected according to target class
- .. until no split remains and a leaf is reached

Advances

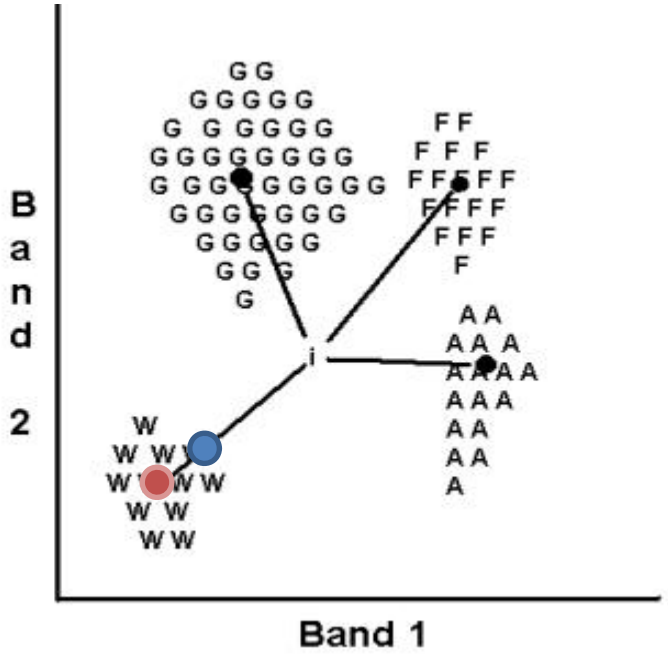
- Conflicts (in the leaves) are possible: more robustness and generalization
- Pruning techniques



Distance/density based classifiers

The idea: place items in the class to which they are closest

- Requires a measure of the distance among samples and *classes*
- Classes represented by
  - Centroids
  - Minimum, mean,percentile distance
  - Medoid (a representative object)
  - A set of individual points

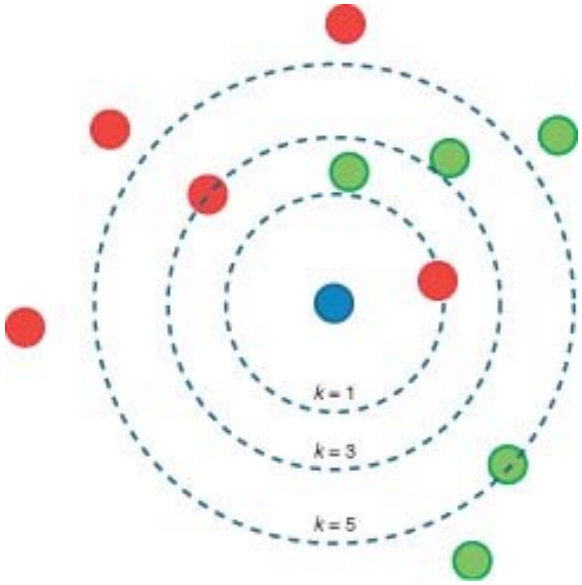


Distance based classifiers

The idea: place items in the class to which they are closest

K-Nearest-Neighbours algorithm

- Exploit labelled data including samples features and a target variable (class)
- No explicit training phase
- When a new sample is presented it is labelled as the majority of its K nearest other samples



Drawbacks

- Calculating the distances may be time consuming
- Some heuristics exist in order to limit the number of distances to be calculated

Unsupervised learning

"Learning from data that has not been labeled, classified or categorized"

- Clustering
- Anomaly detection
- Neural Networks
- Density estimation
- Dimensionality reduction

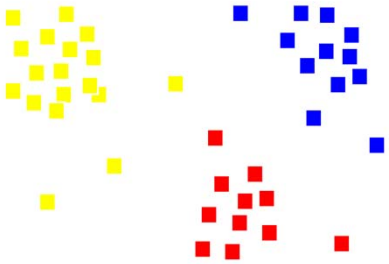
•Clustering (cluster analysis)
"Grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)."

•Hierarchical clustering

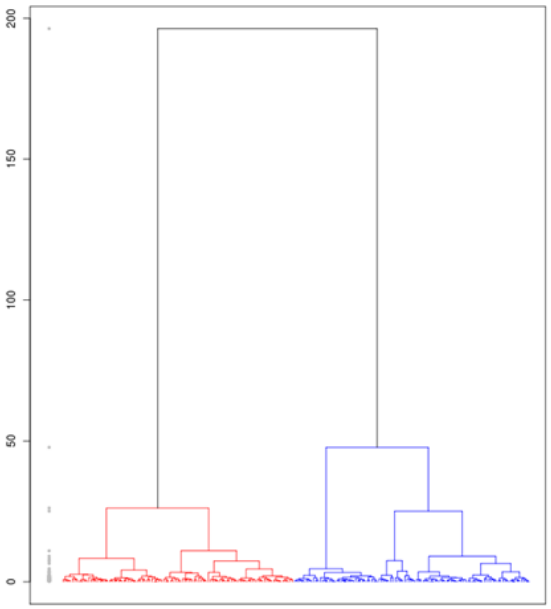
•Centroid clustering

•Distribution models

•Density models

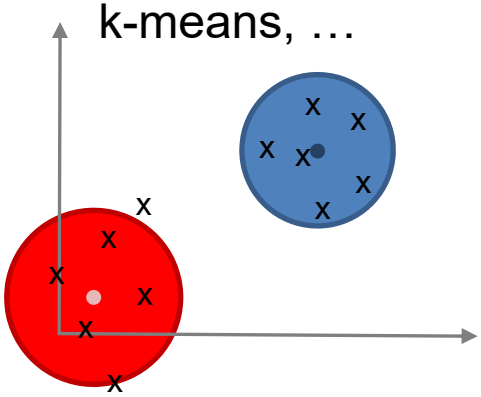


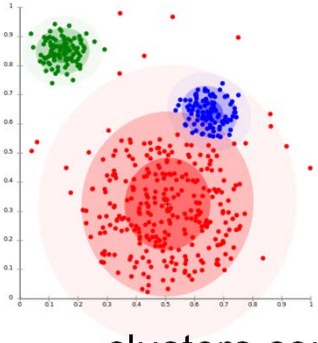
Machine learning – Unsupervised learning



Hierarchical clustering
applying dendrograms

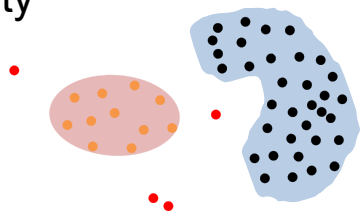
Centroid clustering,
e.g. nearest neighbor,
k-means, …





Density models

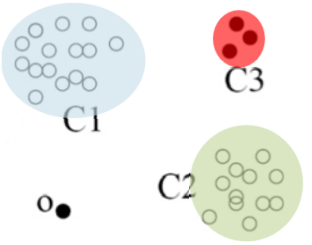
- clusters have a high density of objects
- clusters are separated by regions of low object density

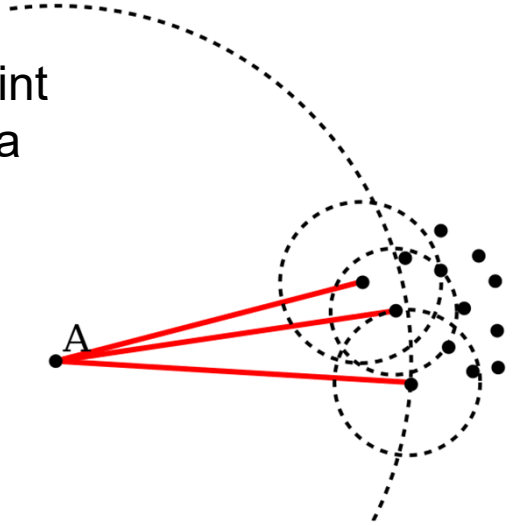


Clustering by
distribution models

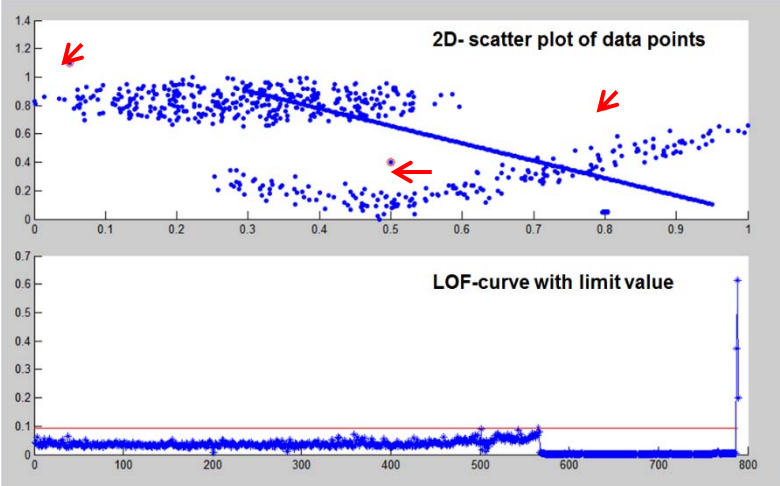
- clusters consists of objects belonging to the same distribution

- Anomaly detection
  - One-class SVMs
  - Cluster based outlier detection
    First cluster all data, then look at small clusters and calculate a factor depending on the cluster size and the similarity to the next cluster
  - Local Outlier Factor (LOF)
    Basic idea: comparing the local density of a point with the densities of its neighbors. Point A has a much lower density than its neighbors.

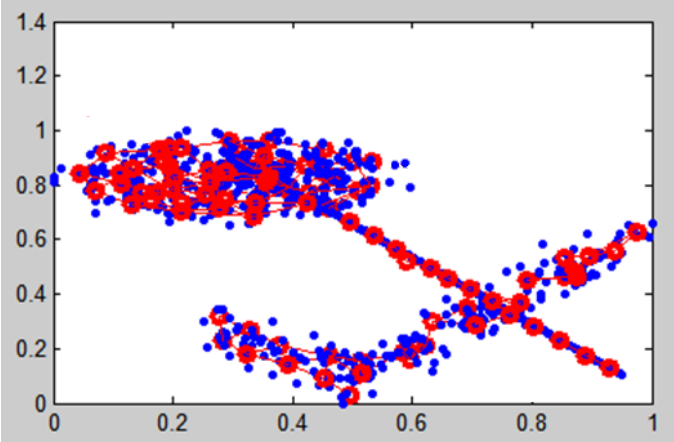




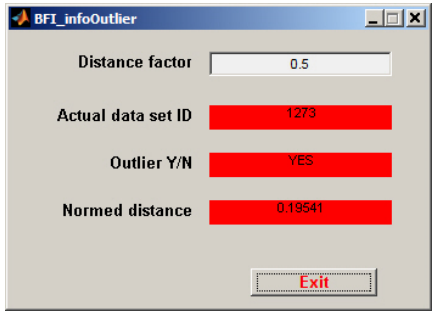
- Example: Outlier test in higher-dimensional data



Application of LOF method to 3-dimensional test data to identify automatically outliers in historical data.



Generation of clusters in the cleaned data set



Detection of outliers by calculating a density based distance between actual data set and derived clusters.

GNG is a derivative of SOM and produces clusters only at places were data points exist. So complex shapes in n-dimensional input space can be represented by cluster centers

The problem

- Currently a large amount of data is collected (especially in industrial plants)
- When building a data driven predictive model not all the variables are related to the target
- **Selecting related variables, why?**
  - Using non-related variables may degrade model performance
  - Gain knowledge from data
  - Faster training, less complex models
  - Less overfitting

- Variable selection reduces the dimension of data **without transforming** data
- particularly important when the number of potential input variables is considerable with respect to the number of available measurements.

Variables selection approaches

- **Filter:** variables are ranked according to a goodness measure (i.e. correlation,…)
  - Fast
  - Not related to the used model
  - Flexible

T-test
Wilconox test
Correlation
Fisher criterion
Single variable performance
Mutual information

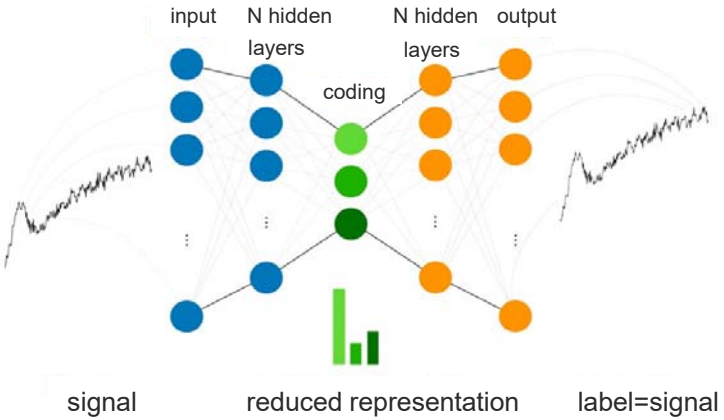
- **Wrappers:** variables subset performance evaluated on the basis of the employed model
  - Efficient in performance
  - May be time consuming
  - Unstable

Exhaustive selection
Sequential selection
GIVE-A-GAP

- **Embedded:** the selection is part of the learning machine
  - Too specific, not always possible

- Neural Networks
  - Autoencoder networks (formerly known as bottle-neck)
  - *The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction*

- Autoencoder learns abstraction of
- input data by **reducing network topology**
- Reduction of **dimension** = **compaction**
- Deep Learning often selects several
- Autoencoders one after the other
- as a preliminary step (**stacking**)



- BFI uses special activation functions, which are motivated by **physical processes**
  - Example: **Karhunen-Loevé Eigenspace Analysis** of process variables
  - to establish a non-linear anomaly detector

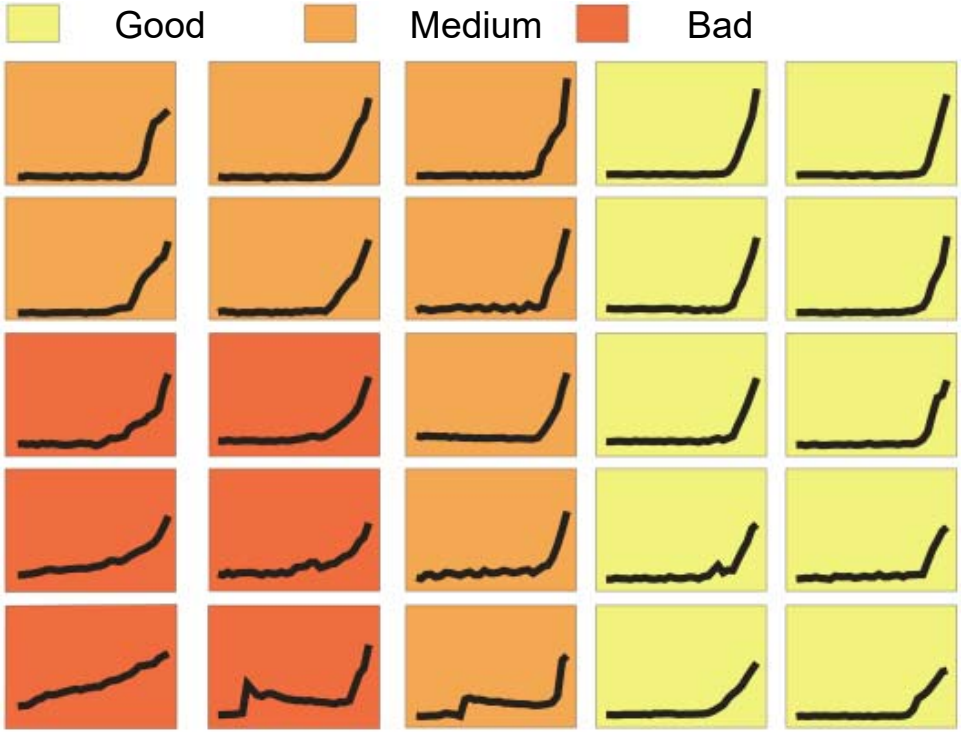
- Neural Networks
  - Self Organising Map (Kohonen Feature Map)
  - *SOM produces a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction.*

Application of SOM

-to cluster signal curves

-then label clusters

-use as classifier



Big Data Analytics = Deep Learning ???

*"**Big Data analytics** is the process of collecting, organizing and analyzing large sets of data (called Big Data) to discover patterns and other useful information."*

Deep Learning is part of a broader family of machine learning methods based on learning data representations.
Methods of Deep Learning are

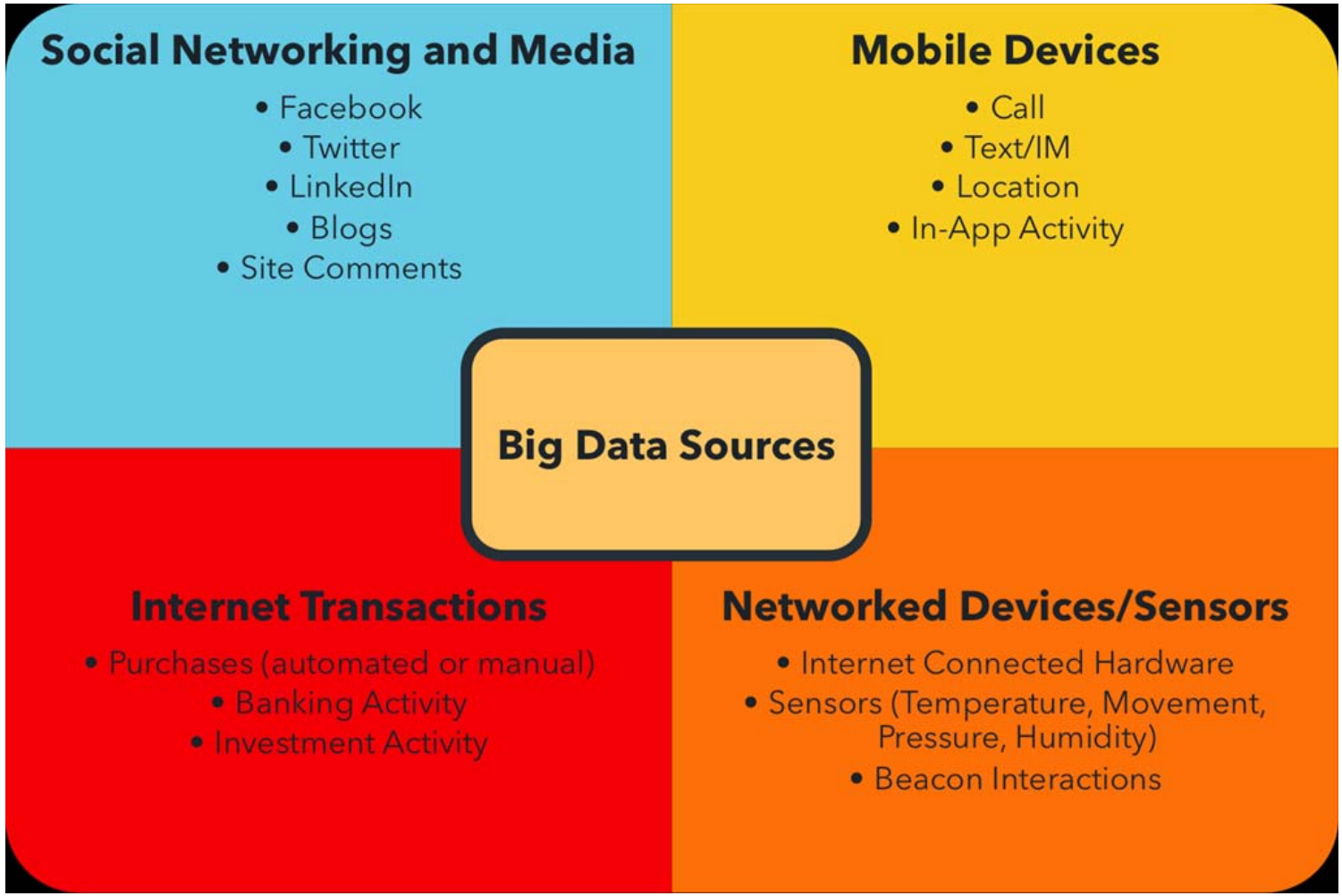
- Deep neural networks (supervised)
- Recurrent networks (supervised)
- Convolutional networks (supervised)
- Deep belief networks (unsupervised)

*"**Big Data analytics** is the process of collecting, organizing and analyzing large sets of data (called Big Data) to discover patterns and other useful information."*

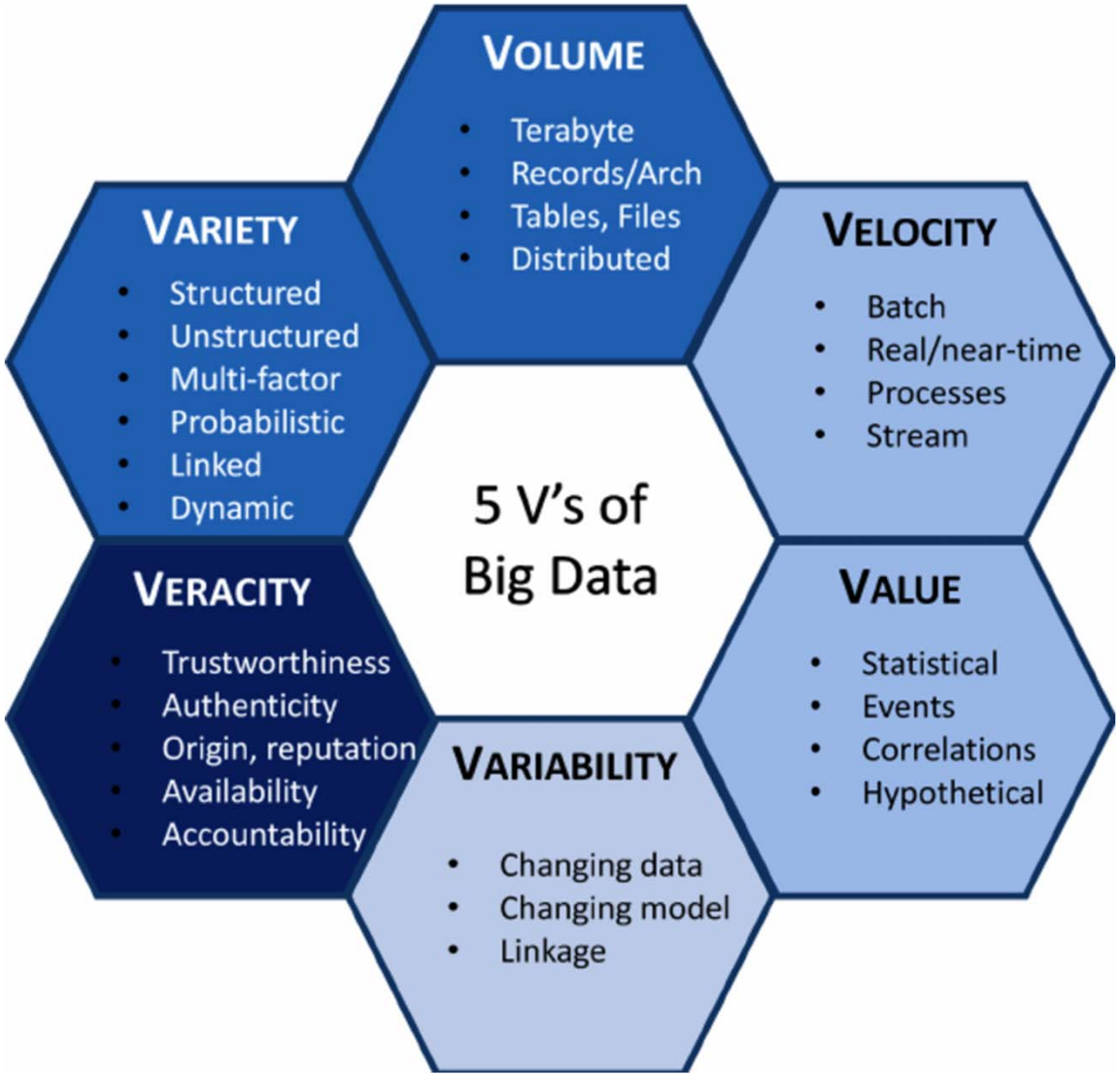
Motivation:

- Nowadays data are continuously generated
- Multiple sources, multiple kind of data
  - Images
  - Sounds
  - Transactions
  - Purchases
  - Positions
  - Text
  - Social networks
  - … and sensors in the industrial field
- **I**nternet **O**f **T**hings
- These data are a **resource**, but their efficient exploitation is complicate

They are not only LARGE: they are BIG


Social Networking and Media
• Facebook
• Twitter
• LinkedIn
• Blogs
• Site Comments
Mobile Devices
• Call
• Text/IM
• Location
• In-App Activity
Big Data Sources
Internet Transactions
• Purchases (automated or manual)
• Banking Activity
• Investment Activity
Networked Devices/Sensors
• Internet Connected Hardware
• Sensors (Temperature, Movement, Pressure, Humidity)
• Beacon Interactions

Big Data Analytics

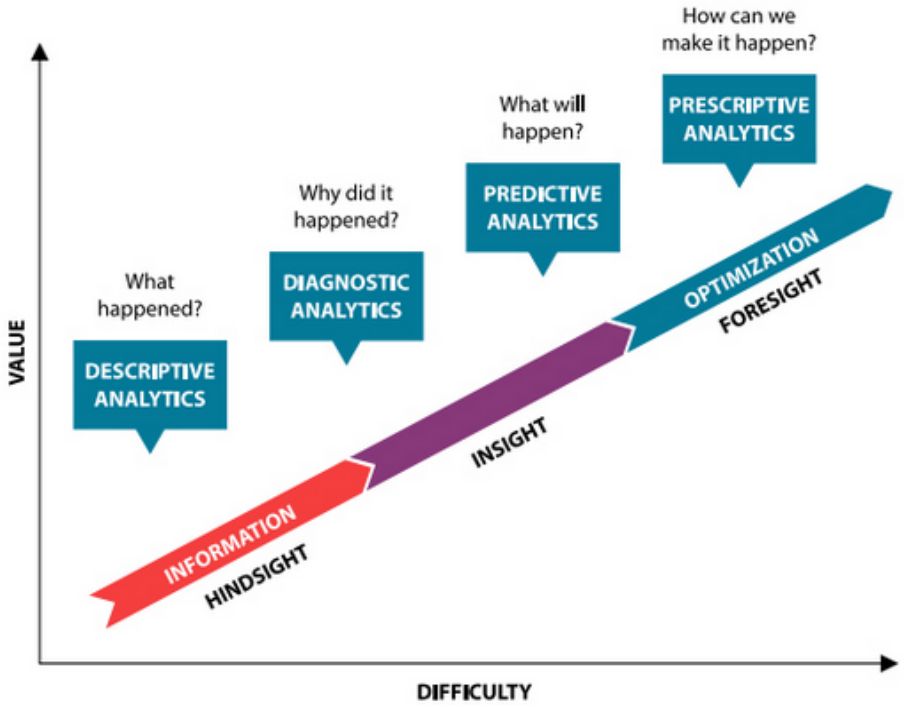


Main points:

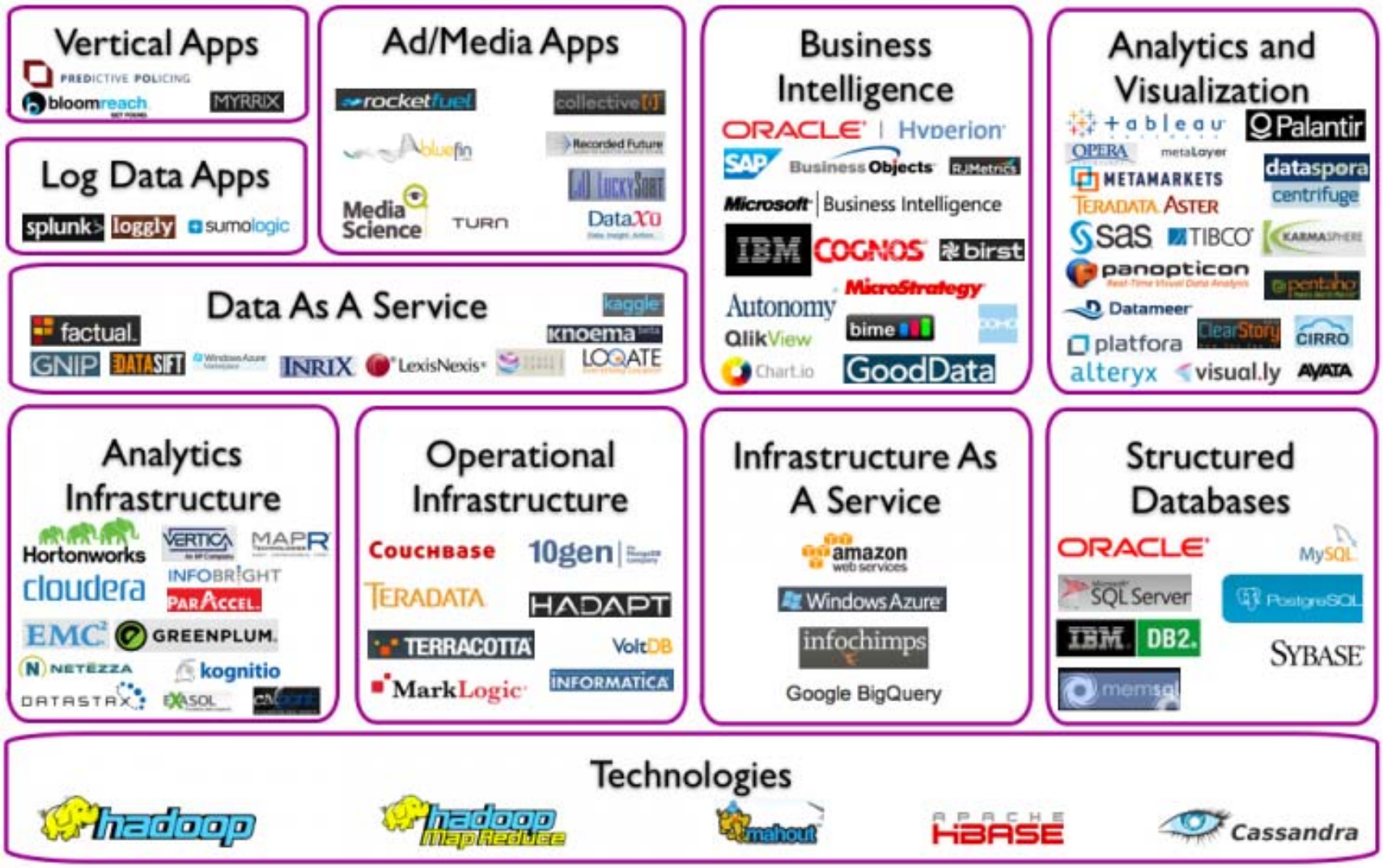
- Focus on the volume and on heretogeneity of data
- Structured and NON structured data

- Need for specific tools for satisfying 5v
  - Algorithms
    - Suitable analytics for heterogeneous data
    - Parallelization
    - Distributed computing
  - Databases
    - Fast retrieving and storage
    - Efficient organization
    - From data-base to data-lake
  - Hardware
    - Toward faster machines

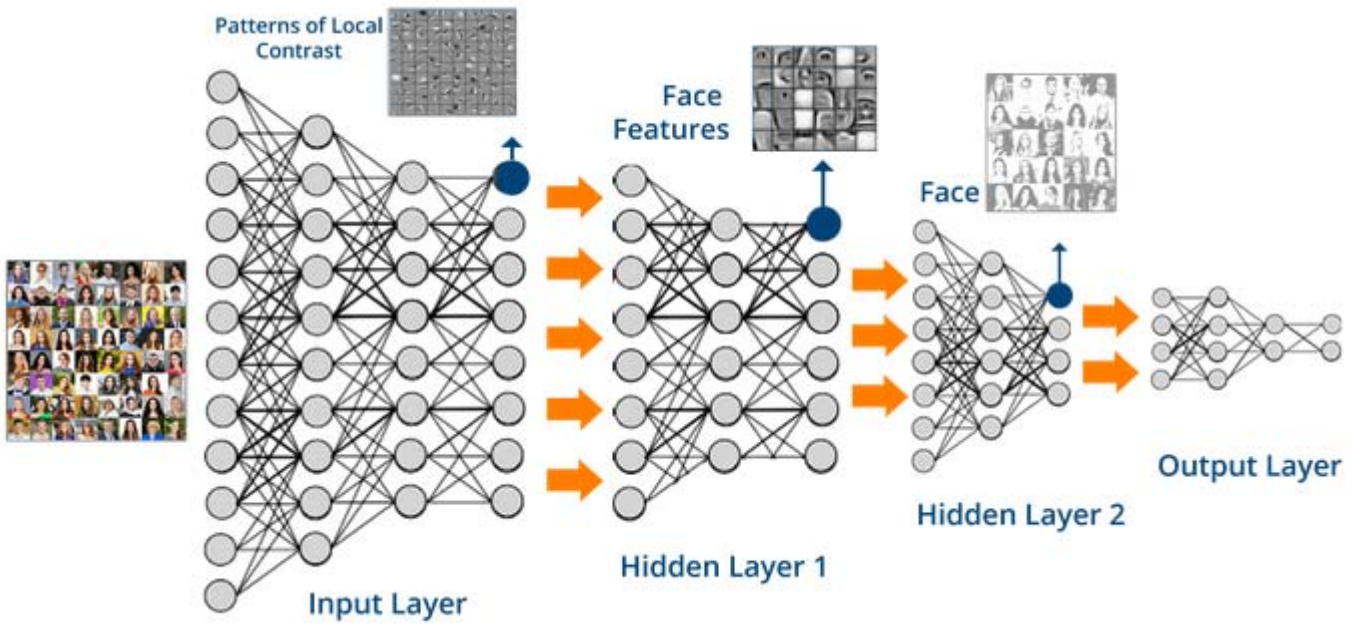
Big Data analytics steps



Existing tools



Deep Learning



What is deep learning

Deep Learning is part of a broader family of machine learning methods based on learning data representations.

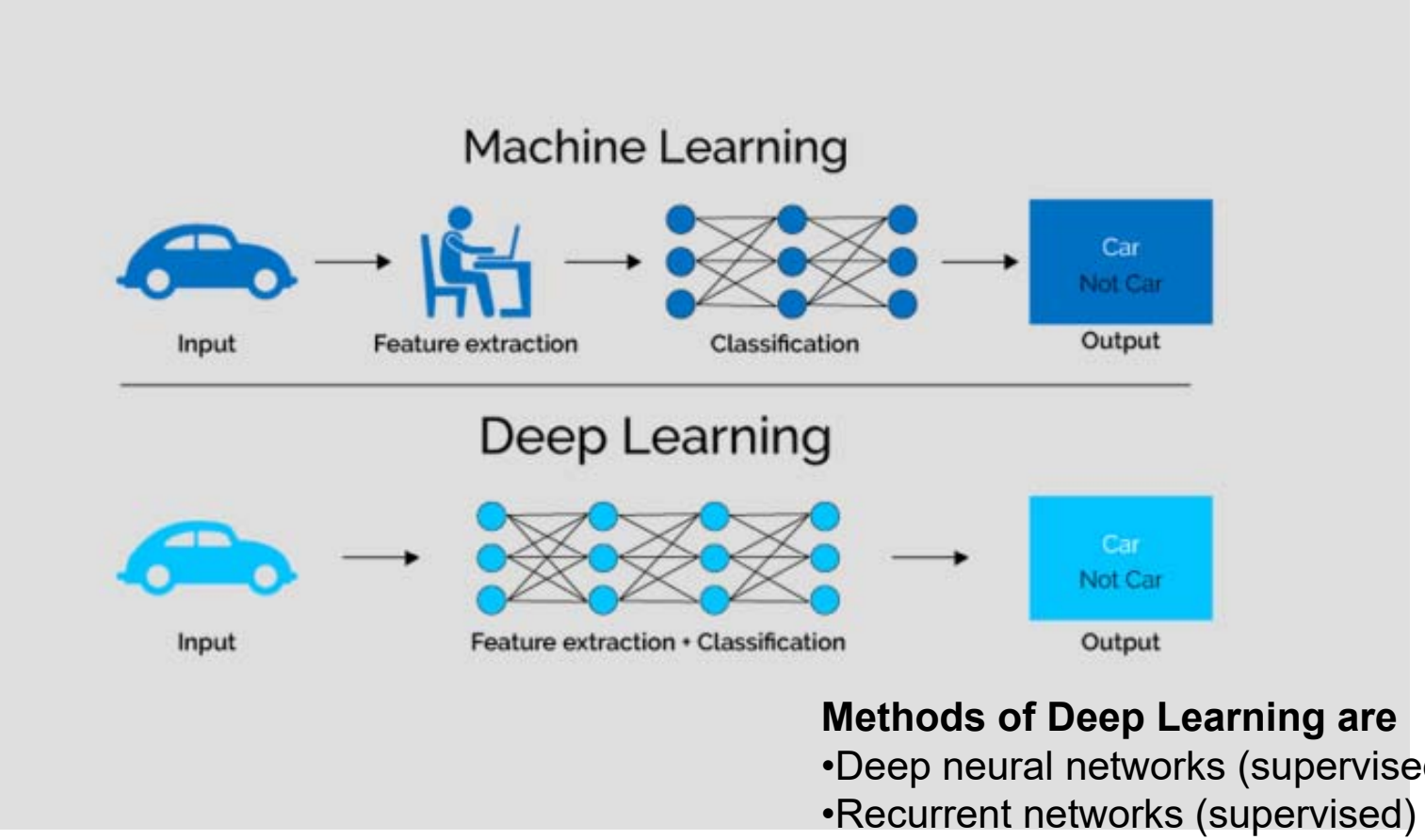
Each level of representation represents a set of features or concepts deriving from previous levels and used at next levels.

Why today?

Tools formanaging such data structures are available

- Hardware (CPUs, GPUs,..)
- Algorithms

Machine learning & Deep learning

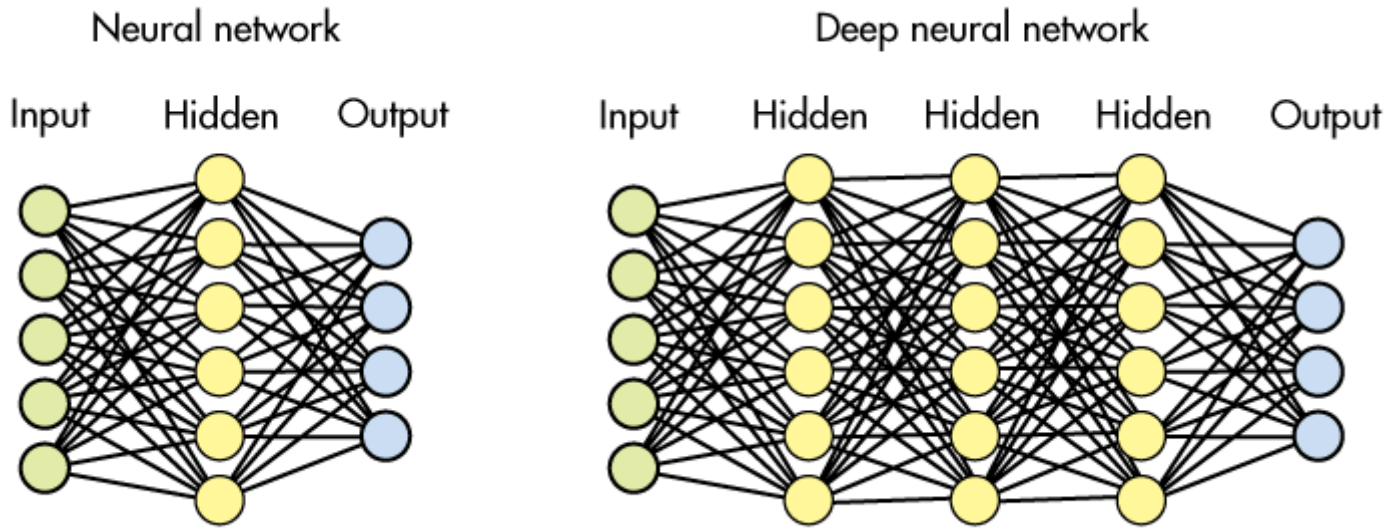


Methods of Deep Learning are

- Deep neural networks (supervised)
- Recurrent networks (supervised)
- Convolutional networks (supervised)
- Deep belief networks (unsupervised)

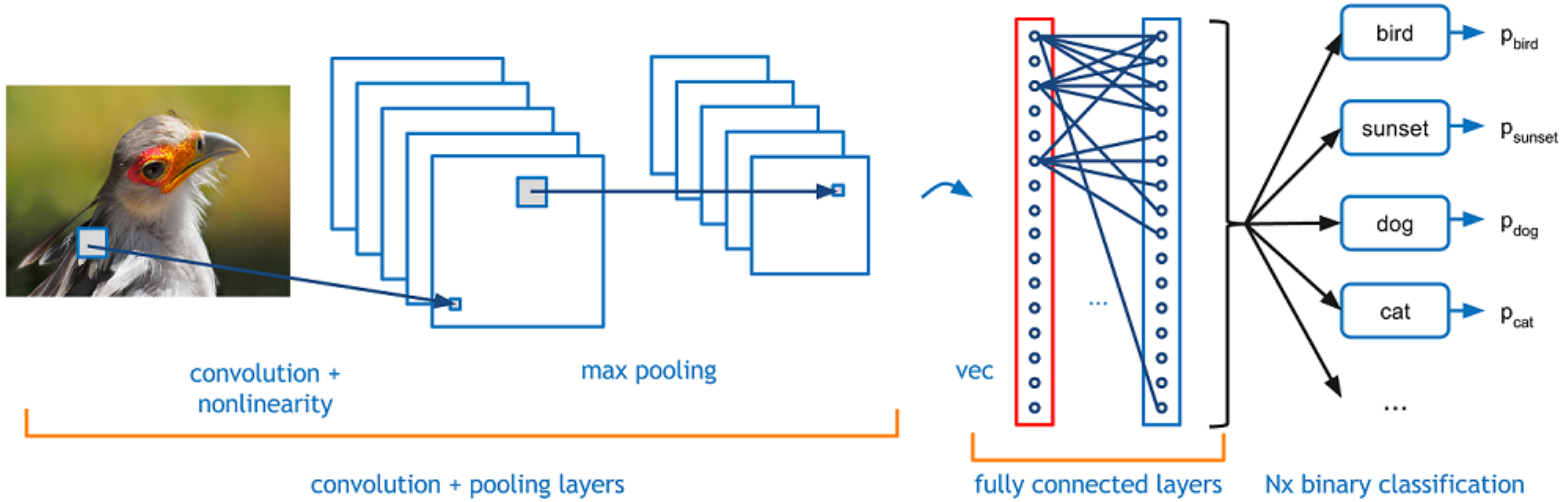
Deep learning popular architectures

Deep neural networks



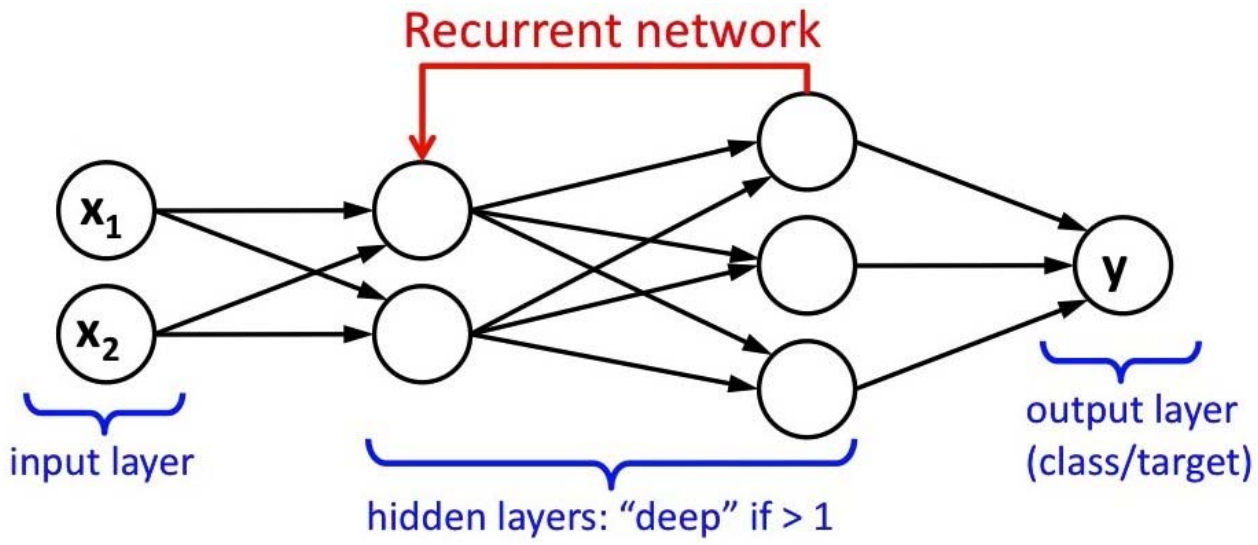
- More than 1 hidden layer.. Goes deep!
- Can be trained by means of standard ANN training algorithms
- intermediate layers build up multiple layers of **abstraction** [if we're doing visual pattern recognition, then the neurons in the first layer might learn to recognize edges, the neurons in the second layer could learn to recognize more complex shapes]
- **Problems** may start if we go very deep
  - Different learning speed among layers
  - The vanishing gradient problem

Convolutional neural networks



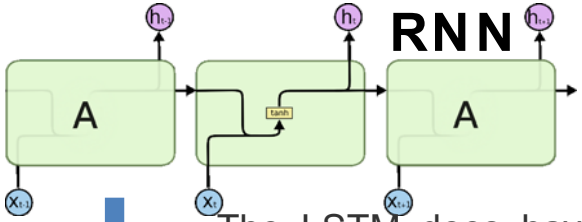
- Popular in image classification (but can be extended to any task)
- Image partitioned in small parts
  - Bio inspired (animal brain cortex organization)
- Each slide is filtered (convolved) and (after some maths) the first fully connected layer is created
  - Each input represent one or more abstract characteristics
- Then training is similar to standard FFNN

Recurrent networks



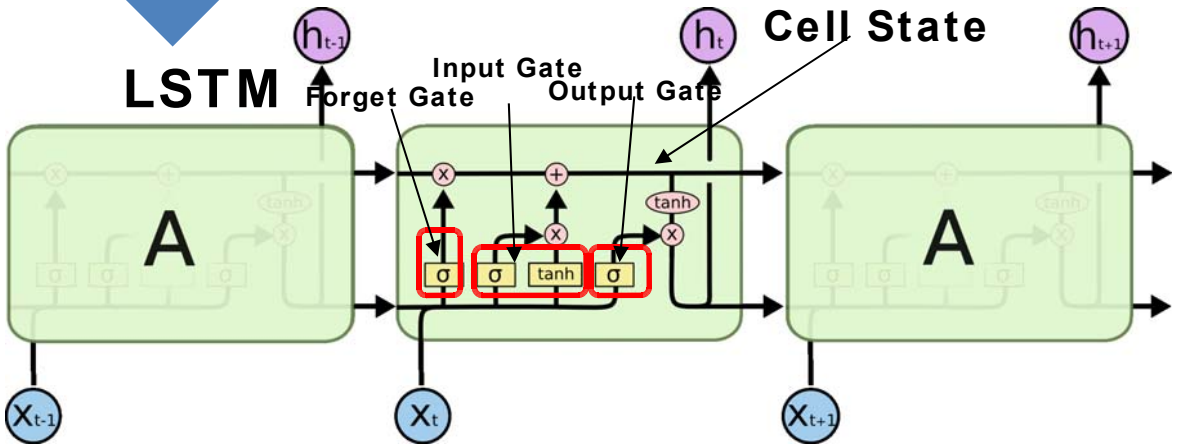
- Allows backward connection among neurons
- This gives the ANN a memory
  - Layers extend the memory
- Suitable for
  - Time series forecasting
  - Speech recognition
- Possible to train with standard algorithms. Special architectures need special algorithms

Long Short Term Memory



The key to LSTMs is the cell state, the horizontal line running throug the top of the diagram.
The cell state is kind of like a conveyor belt. It runs straight down th entire chain, with only some minor linear interactions. It's very easy fc information to just flow along it unchanged.

The LSTM does have the ability to remove or add information to the cell state, carefull regulated by structures called gates.
Gates are a way to optionally let information through. They are composed out of a sigmoi neural net layer and a pointwise multiplication operation.

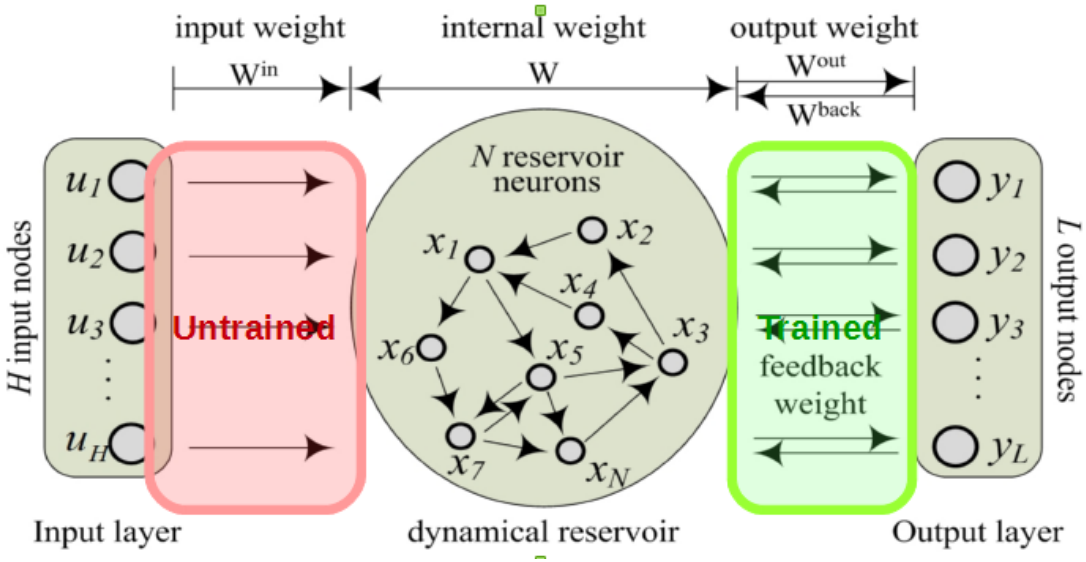


The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through. A value of zero means "let nothing through," while a value of one means "let everything through!"
An LSTM has three of these gates, to protect and control the cell state.

Echo State Networks

Echo State Networks combine fast training and performance for time series prediction.

Reservoir layer models non-linear dynamics (Markovian lavoured) component and does not need a *traditional* training.



Main ANN traning takes place on the output layer as a linear feed-forward component.

Main criticality concerns the selection of ESN parameters

Tools


Caffe
Caffe2
Chainer
Cognitive Toolkit
MATLAB
mxnet
PaddlePaddle
PYTORCH
TensorFlow
...
NVIDIA DEEP LEARNING SDK and CUDA