Data Mining

INTRODUCTION

Matteo Golfarelli

Teaching methods and material

Classroom and laboratory lessons using the Weka open source software

The course includes two modules

- Data mining (36 ore): Introduces the basic concepts, describes the mining techniques to apply to structured data
- Text mining (18 ore): Describes how mining techniques need to be specialized to work effectively on textual data

Teaching methods and material

The exam consists of an elaborate and an oral exam on all the subjects of the course.

During the oral examination you may be required to use the Weka software

The choice of the elaborate must be agreed with the teacher

- Determines an additional score for [0..4] points
- Study of an algorithm among those in the literature
- Analysis of a data set with mining techniques

Teaching methods and material

All course topics are covered by downloadable slides from the teacher's site

The **textbook** for the **data mining** module is:

Pang-Ning Tan, Michael Steinbach, Vipin Kumar *Introduction to Data Mining*. Pearson International, 2006.

The **textbook** for the **text mining** module is:

Christopher Manning, Hinrich Schutze, Prabhakar Raghavan. Introduction to Information Retrieval. Cambridge University Press, 2008. (disponibile on line)

More details on the Weka software can be found in:

Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques, IInd Ed. Morgan Kaufmann, 2005.

The Data Mining Module

The goal of the DM module is twofold:

- Acquire practical skills for data analysis
  - Overview on Weka
  - Case studies and Exercises
- Review and broaden the toolbox of the techniques acquired during the Machine Learning course
  - Most of the techniques studied in Machine Learning are well-suited for real-valued attributes, but in data mining large part of the features are categorical (i.e. nominal or ordinal)

The Data Mining Module

The goal of the DM module is twofold:

- Acquire practical skills for data analysis
  - Overview on Weka
  - Case studies and Exercises
- Review and broaden the toolbox of the techniques acquired during the Machine Learning course

DATA MINING

- Classification
  - Naive-Bayes
  - Decision Tree
  - Rule-based
- Clustering
  - Hierarchical
  - DBScan
- Association Rules
- Outlier Detection

MACHINE LEARNING

- Classification
  - Bayes
  - K-NN
  - SVM
- Clustering
  - K-means
  - Expectation Maximization
- Dimensionality Reduction
- Neural Network
- Deep Learning

Data & Knowledge Engineering Profile

The DKE profile studies the modeling and algorithms needed to build and exploit knowledge for advanced business and scientific applications.

The reference applications are:

- Business Intelligence
- Semantic Web
- Internet-of-Things

The reference professional figures are:

- Data Scientists
- Designers and Consultants in Business Intelligence and Analytics
- Semantic web and IoT systems experts
- Technology experts in Big Data
- Project managers of high-tech projects

More Courses in the DKE Profile

- Big Data (Prof. Gallinucci)
- Business Intelligence (Prof. Rizzi)
- Project Management (Prof. Boschetti)
- Decision Support Systems (Prof. Maniezzo)
- Semantic Web (Prof. Carbonaro)

There is a specific Erasmus agreement with the Universidad Politecnica de Catalunya (Barcelona) including a master specialized on the DKE subjects

Why Mining Data?

The amount of data stored on computer is constantly increasing

- IoT data
- Social data
- Data on purchases / tax receipts
- Banking and credit card transactions

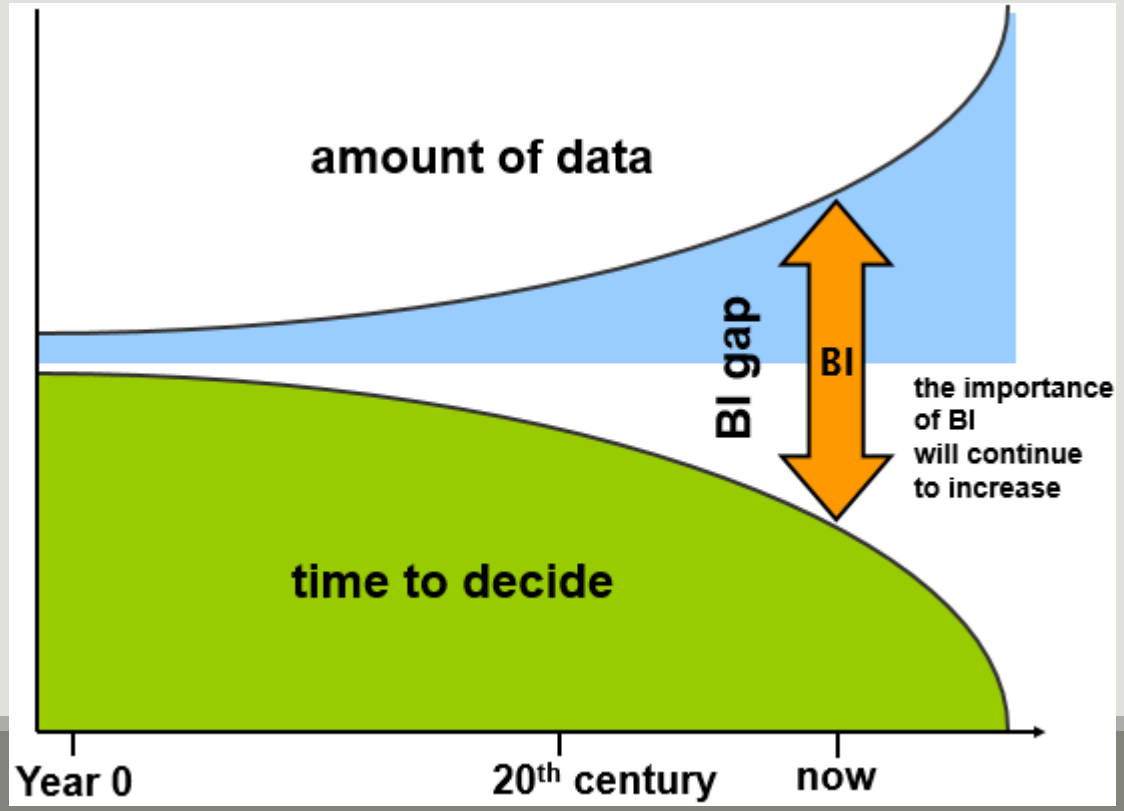
Hardware becomes more powerful and cheaper each day

Competitive pressure is constantly growing

- The information resource is a precious asset to overcoming competitors

Why Mining Data?

Most of the information on the data is not directly apparent
Men's guided analysis can take weeks to find useful information
Most of the data has never been analyzed

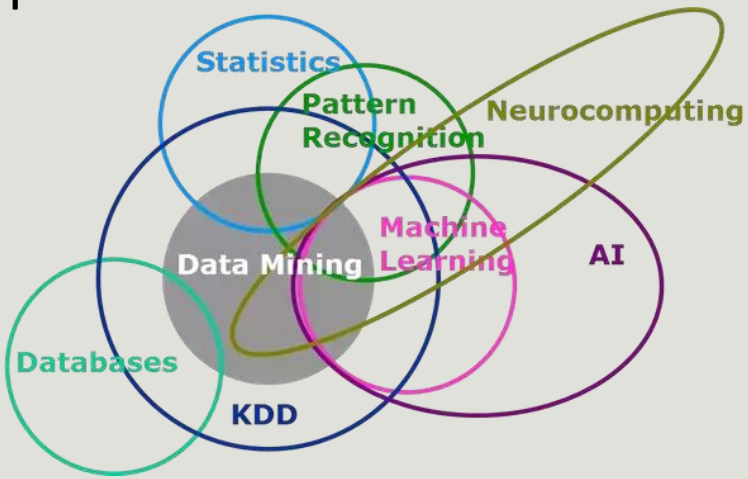


AI, Machine Learning & Data Mining

Although strongly interrelated, the term machine learning is formally distinct from the term Data Mining which indicates the computational process of pattern discovery in large datasets using machine learning methods, artificial intelligence, statistics and databases.

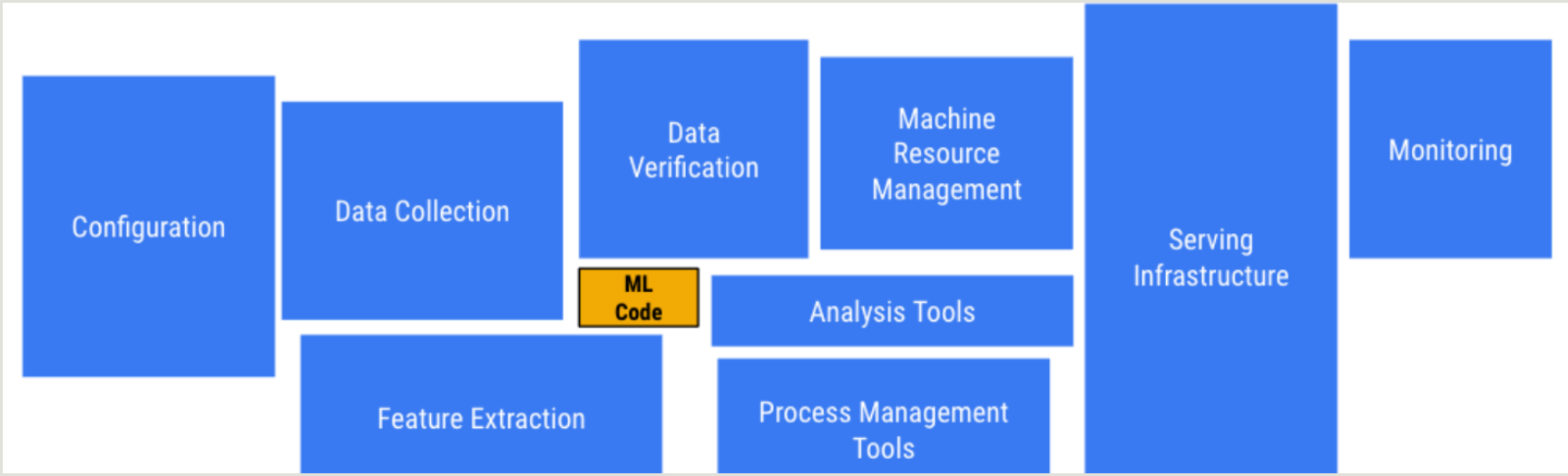
Apart from the actual analysis phase, data mining covers aspects of:

- Data management and pre-processing
- Modeling
- Identification of metrics of interest
- Visualization



AI, Machine Learning & Data Mining

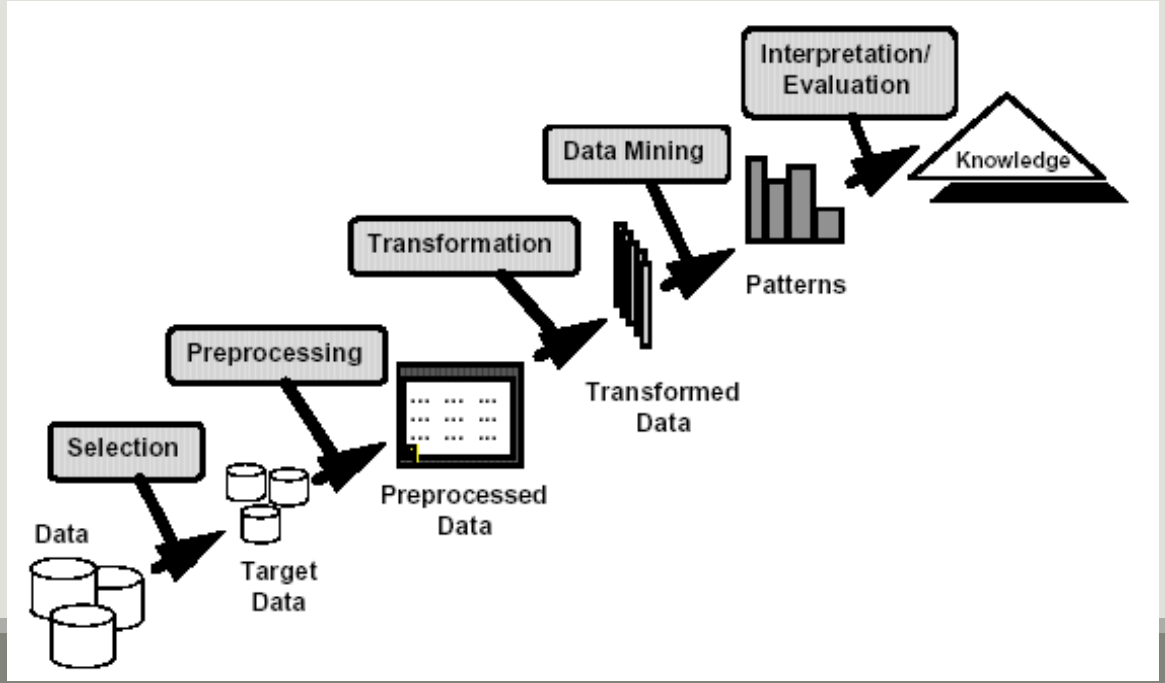
The role of Machine Learning in a real project is shown by the picture that lists all the typical activities. The larger the size, the longer the time taken by ana activity



Data Mining Definitions

Complex extraction of implicit, previously unknown and potentially useful data from the information.

Exploration and analysis, using automated and semi-automatic systems, of large amounts of data in order to find significant patterns



Analytics

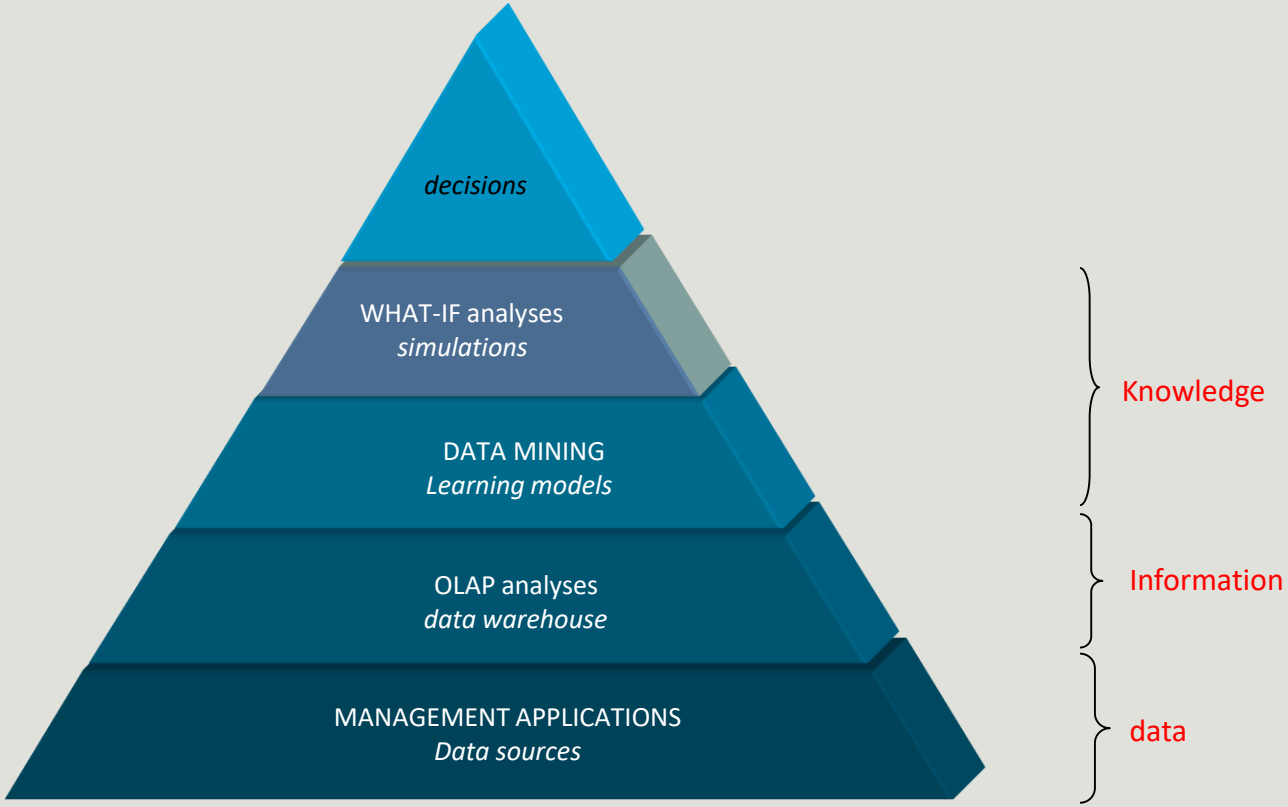
Analytics refers to software used to discovery, understanding and sharing of relevant patternin data. Analytics are based on the concurrent use of statistics, machine learning and operational research techniques. Analytics often exploit advanced visualization techniques

Analytics in BI 2.0 play the same role data mining played in BI 1.0

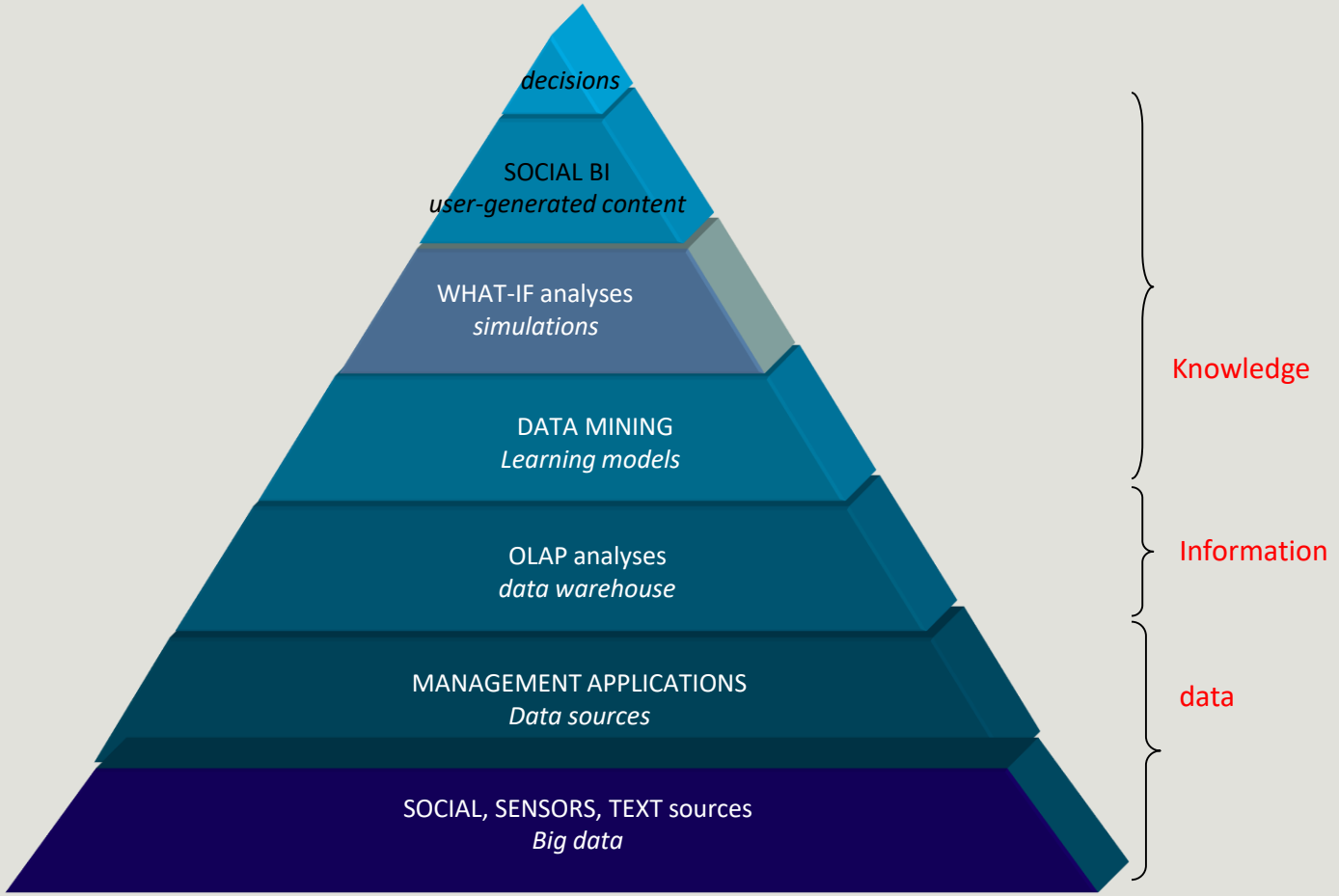
Data Mining solutions have spread much less than DW ones due to the:

- Complexity and costs
- Needs of an expert for results understanding
- lack of certainty of meeting the project goals

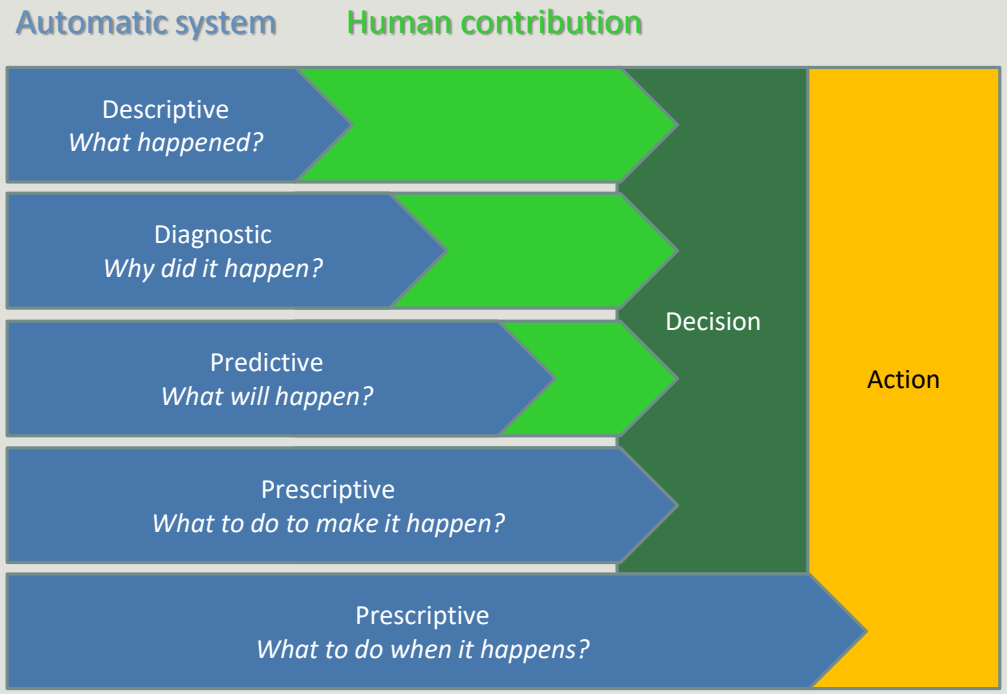
The BI 1.0 pyramid



The BI 2.0 pyramid

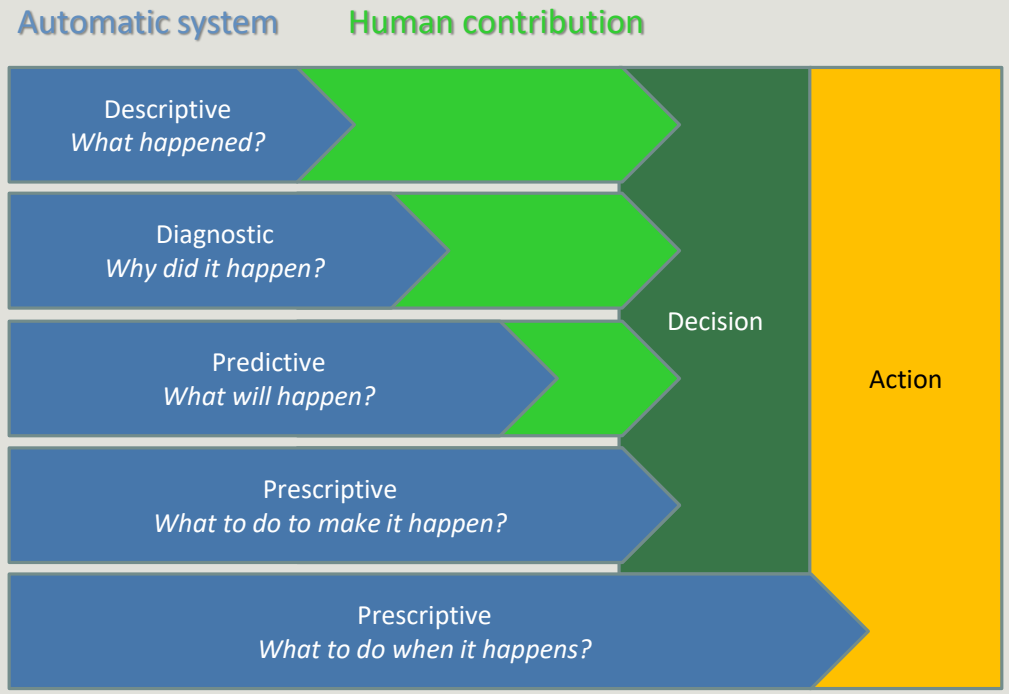


The BI 2.0 pyramid



As a function of the level of automation of decision

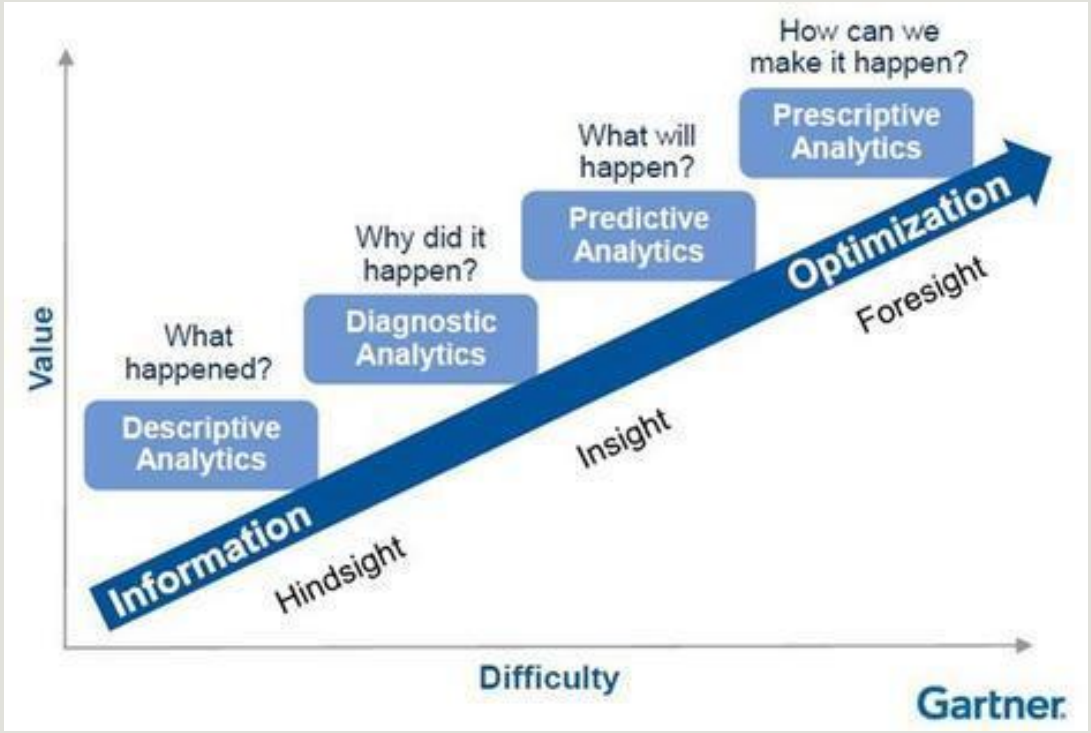
The BI 2.0 pyramid



As a function of the level of automation of decision

Which is the best solution to be adopted?

The BI 2.0 pyramid



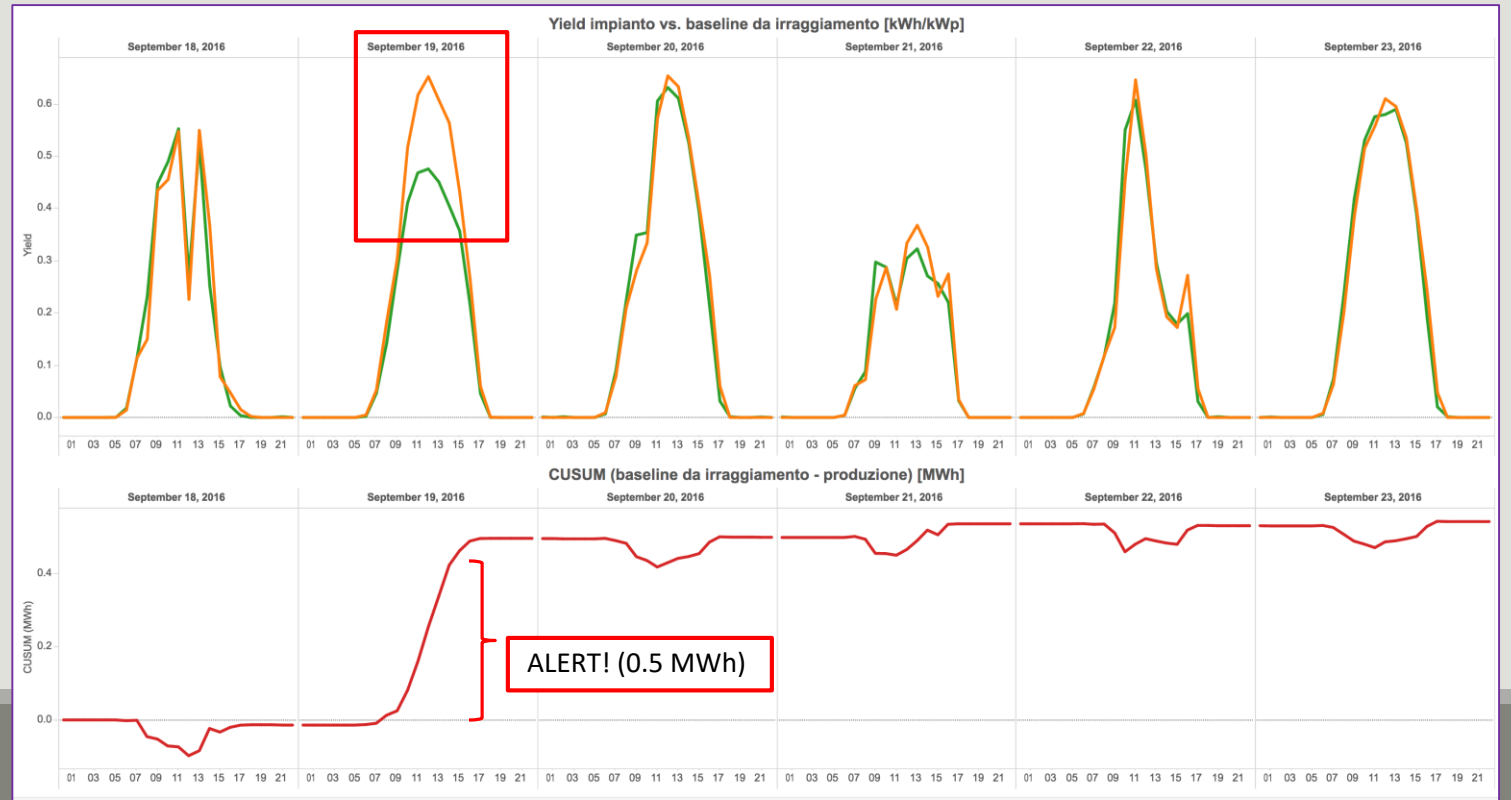
As a function of the level of automation of decision

The simplest bringing value to the company

Descriptive Analytics

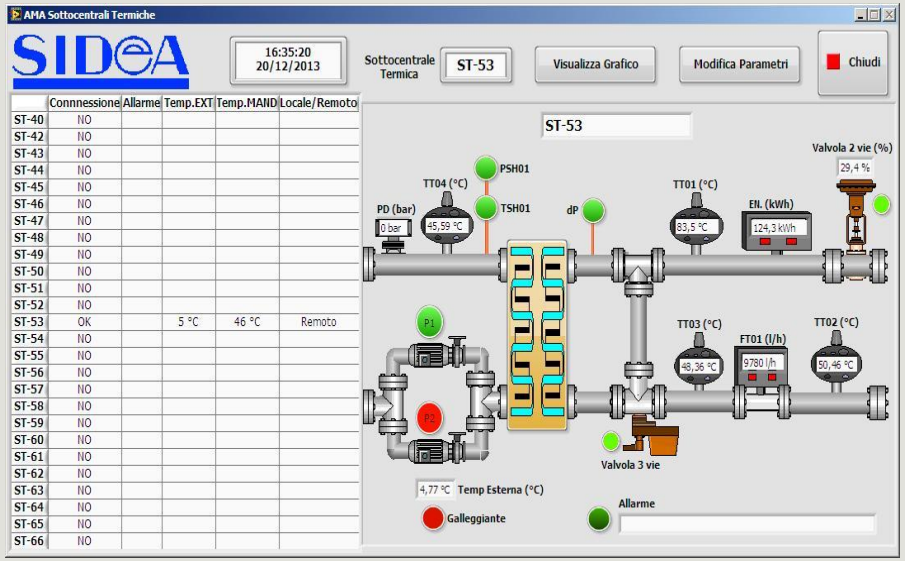
- Historical/past data is used to describe the system
- Dashboards and OLAP are the main visualization types

Energy baselines



Diagnostic Analytics

- Use the data to understand the causes
- Fault diagnosis
- Preventive alerting systems



	Connnessione	Allarme	Temp.EXT	Temp.MAND	Locale/Remoto
ST-40	NO				
ST-42	NO				
ST-43	NO				
ST-44	NO				
ST-45	NO				
ST-46	NO				
ST-47	NO				
ST-48	NO				
ST-49	NO				
ST-50	NO				
ST-51	NO				
ST-52	NO				
ST-53	OK		5 °C	46 °C	Remoto
ST-54	NO				
ST-55	NO				
ST-56	NO				
ST-57	NO				
ST-58	NO				
ST-59	NO				
ST-60	NO				
ST-61	NO				
ST-62	NO				
ST-63	NO				
ST-64	NO				
ST-65	NO				
ST-66	NO				

Plant data (from sensors) + fault data: Identify (i.e. classify) time series that lead to a fault

- Identify (i.e. classify) time series that lead to a fault by fault type
- Identify the sensors and engineer the features that bring more information (i.e. are strongly correlated with fault)

Predictive Analytics

Use the data to predict future values

- Simulation systems
- Time series prediction
- Failure prediction
- Sales predictions

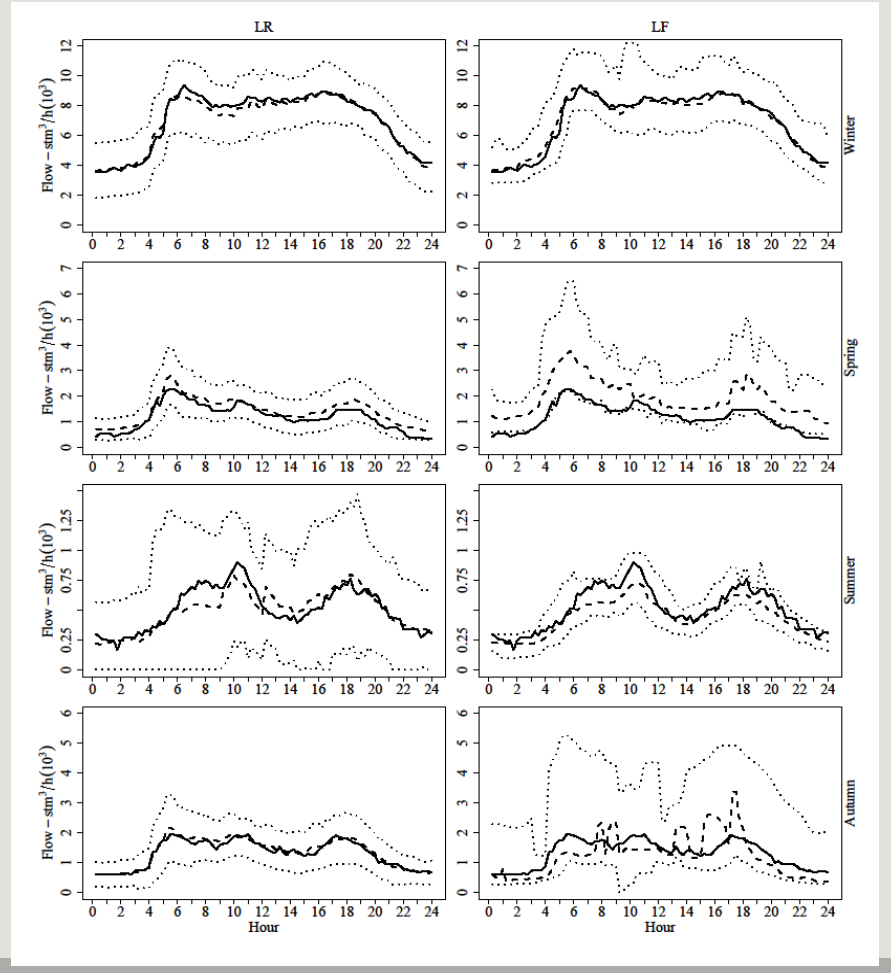
Gas consumption forecast for HERA

Input

- 1 year consumption time series
- 1 day meteo forecast
- Day of the week

Output

- 24 hours consumption forecast



Prescriptive Analytics

- Mono and multi goals optimization systems
- Tests alternative scenarios
- Decisional systems

- Optimize energy acquisition mix depending on needs, market and meteo
  - ✓ Photovoltaic is enough?
  - ✓ Buy from the network?
  - ✓ Biomass?
- Decide watering given an optimal terrain humidity profile
  - ✓ Watering now even if it will rain tomorrow?
  - ✓ Watering many times for short periods, or watering abbundatly few times?

The BI adoption path

The adoption of BI solutions is incremental and rarely allows steps to be skipped
This is because it is ***risky***, ***costly*** and ***useless*** to adopt advanced solutions before completely exploiting simple ones

- Managers are not ready
  - ✓ Not in the right mindset
- Data are not ready
  - ✓ Not of enough quality
- Company processes are not ready
  - ✓ Not defined to rely on and to be reactive to data

Beware of consultants and software vendors who offer advanced analytics if you barely exploit the corporate data warehouse

Turning your company in a data-driven one

The term ***data-driven company*** refers to companies where decisions and processes are supported by data

- Decisions are based on quantitative rather than qualitative knowledge
- Processes & Knowledge are an asset of the company and are not lost if managers change

The gap between a data-driven decision and a *good* decision is a *good* manager

Adopting a data-driven mindset goes far beyond adopting a business intelligence solution and entails:

✓ Create a data culture

✓ Change the mindset of managers

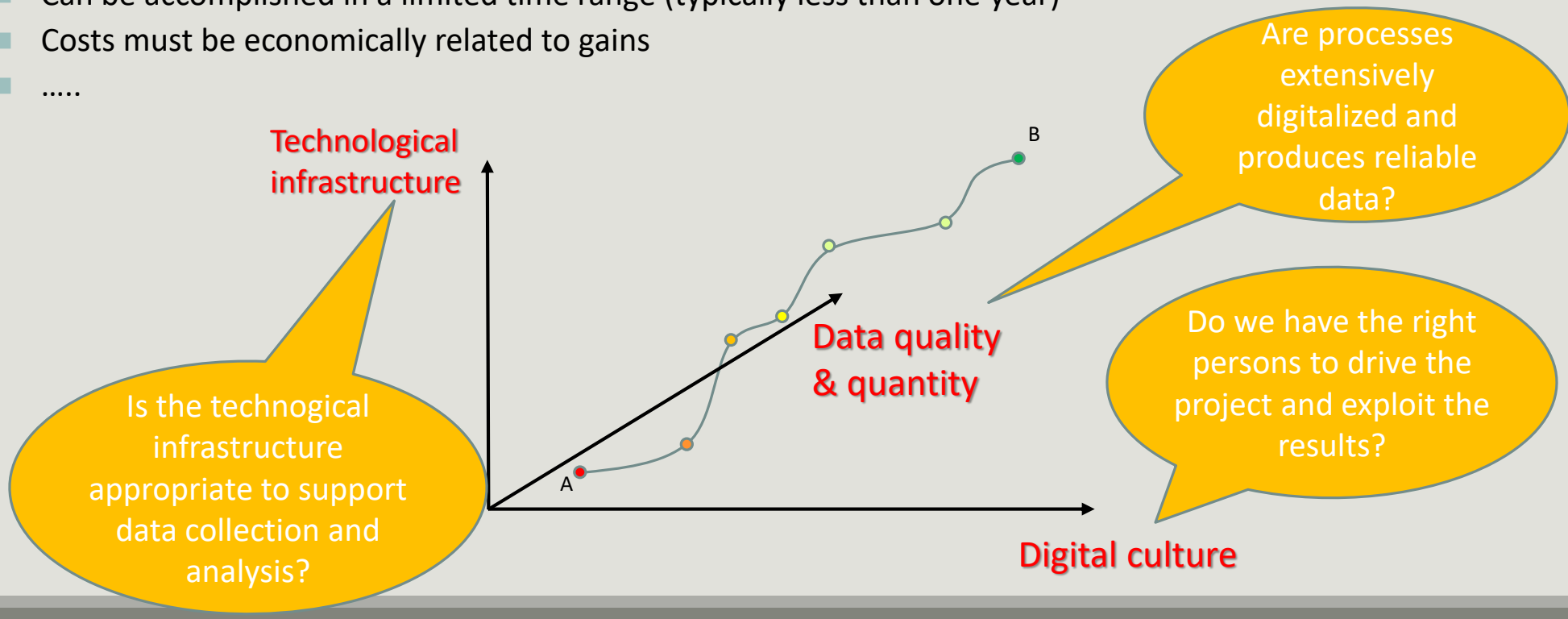
✓ Change processes

✓ Improve the quality of all the data

Turning your company in a data-driven one

Digitalization is a journey that involves three main dimensions. Moving from A to B is a multi-year process made of intermediate goals each of wich must be feasible

- Solves a company pain and brings value
- Can be accomplished in a limited time range (typically less than one year)
- Costs must be economically related to gains
- …..



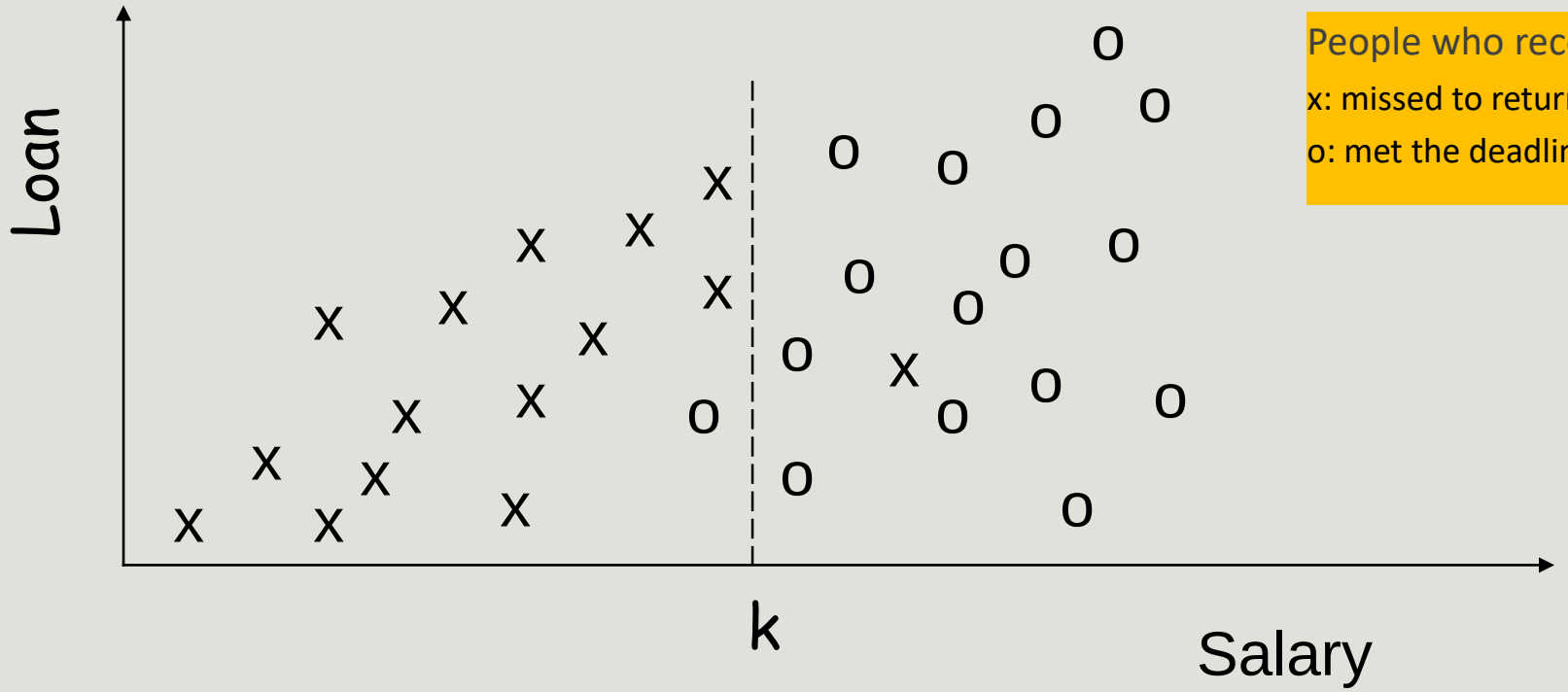
Pattern

A pattern is a synthetic representation rich in semantics of a set of data; usually expresses a recurring pattern in data, but can also express an exceptional pattern

A pattern must be:

- Valid on data with a certain degree of confidence
- It can be understood from the syntax and semantic point of view, so that the user can interpret it
- Previously unknown and potentially useful, so that the user can take actions accordingly

Example



People who received a loan

x: missed to return a rate

o: met the deadlines

- Pattern:
  - ✓ `IF` salary < k `THEN` missed rate

Pattern Types

Association rules

- Let you determine the logical implications of the dataset, and then identify the groups of affinity between objects

Classifiers

- Allow you to derive a model for classifying data according to a set of a priori assigned classes

Decision trees

- Are special type of classifiers that allow to identify, in order of importance, the causes that lead to an event occurring

Clustering

- Groups elements in a set, depending on their characteristics, of apriori unknow classes

Time series

- They allow the detection of recurring or atypical patterns in complex data sequences

What is NOT Data Mining

What is not Data Mining	What Data Mining is
– Look for a number in the phone book	– Discover that some surnames are more common in certain regions (eg Casadei, Casadio, ... in Romagna)
– Query a search engine to search for information on "Amazon“	–Group documents returned by a search engine based on context information (eg "Amazon rainforest", "Amazon.com")

Where Data Mining Comes from?

This discipline stands in the middle between several areas

- Machine learning / artificial intelligence
- Pattern recognition
- Statistics
- Databases

Traditional analytical techniques are unsuitable for many reasons

- Quantity of data
- High dimensionality of data
- Heterogeneity of data

Data Mining Applications

Predictive systems

- Exploit some features to predict the unknown values of other features
  - Classification
  - Regression
  - Outlier detection

Descriptive systems

- Find user-readable patterns that can be understood by human users
  - Clustering
  - Association rules
  - Sequential pattern

A Definition for Classification

Given a record set (*training set*)

- Each record is composed by a set of *attributes*, where one ot them represents the *class* of the record.

Find a *model* for the class attribute expressing the attribute value as a function of the remaining attributes/features

Goal: unclassified record must be assigned to a class in the most accurate way

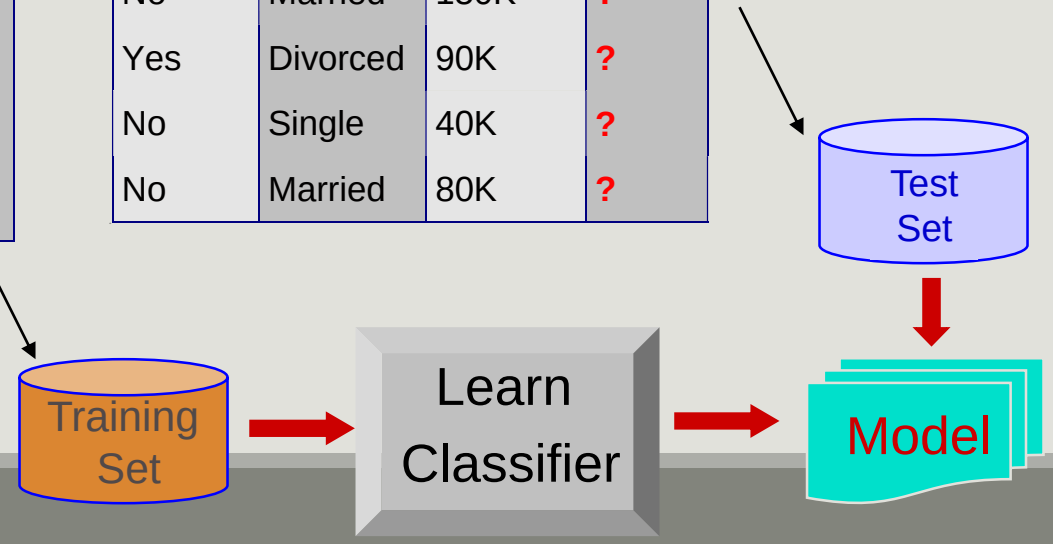
- A *test set* is used to determine the model accuracy. Typically, the data set is split in training set and test set. The first one is used to build the model, the second one to validate it.

An Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

Direct Marketing

- Goal: Reduce the cost of email marketing by ***defining*** the set of customers that, with the highest probability, will buy a new product
- Technique:
  - Exploit the data collected during the launch of similar products
  - We know which customers bought and which one did not buy
  - Such information *{buy, not buy}* becomes the *class attribute*
  - Collect all the available information about each customers: demographics, life style, previous contact with the company
    - Job, Income, age, gender, etc.
    - Use such information as an input to as input attributes to train the model

Classification: Application 2

Fraud detection

- Goal: predict the fraudulent use of credit cards
- Approach:
  - Use past transactions and information about their owners as attributes
    - When a user buy, what does she buy, does she pay late, etc.
  - Label past transactions as *fraudulent* or *legitimate*
  - This information is the classification attribute
  - Build a model for the two classes of transactions
  - Use the model to detect fraudolent behaviors of the next transactions for a specific credit card

Classification: Application 2

Fraud detection

- Goal: predict the fraudulent use of credit cards
- Approach:
  - Use past transactions and information about their owners as attributes
    - When a user buy, what does she buy, does she pay late, etc.
  - Label past transactions as *fraudulent* or *legitimate*
  - This information is the classification attribute
  - Build a model for the two classes of transactions
  - Use the model to detect fraudolent behaviors of the next transactions for a specific credit card

Classification: Application 3

Churn detection

- Goal: Predict customers who are willing to go to a competitor.
- Approach:
  - Use the purchasing data of individual users (present and past) to find the relevant attributes
  - How often does the user contact the company, where he calls, at what times of day he calls more frequently, what is his financial situation, is married, etc.
  - Label users as *loyal* or *not loyal*
  - Find a pattern that defines loyalty

A Definition for Clustering

Given a set of points, each featuring a set of attributes, and having a similarity measure between points, find subset (i.e. cluster) of points such that:

Points belonging to a cluster are more similar to each other than those belonging to other clusters

Similarity measures

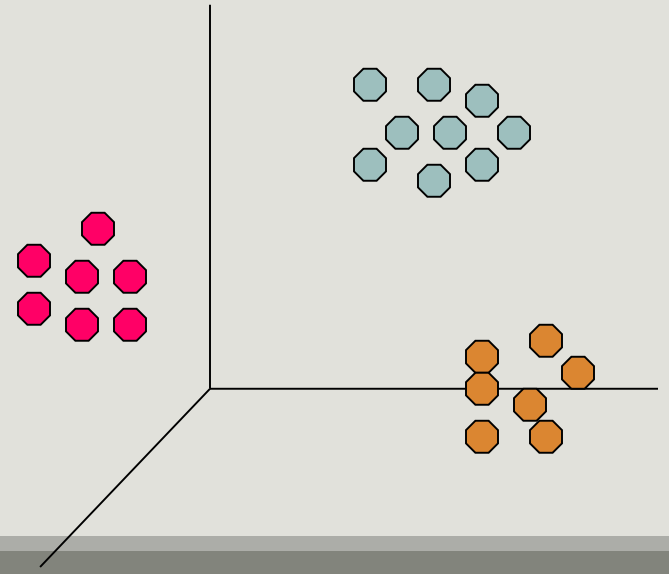
- Euclidean distance is applicable if point attributes assume continuous values
- Many other measures are available or can be defined for each specific domain

Cluster representation

A 3D clustering found using the Euclidean distance

Intra-cluster distances are minimized

Inter-cluster distances are maximized



Clustering: Application 1

Market Segmentation:

- Goal: Split customers into distinct subsets to target specific marketing activities
- Approach:
  - Gather information about customer lifestyle and geographic location
  - Find clusters of similar customer
  - Measure cluster quality by verifying whether the purchasing patterns of customers belonging to the same cluster are more similar to those of distinct clusters

Clustering: Application 2

Document Clustering

- Goal: Find clusters of documents that are similar on the basis of the most relevant terms that they contain
- Approach:
  - Identify the terms that occur most frequently in the different documents.
  - Define a frequency-based similarity measure and use it to create clusters.

Clustering: Application 2

Points to be clustered: 3204 Los Angeles Times articles

Similarity measure: number of common words between two documents (excluding some common words).

Category	*# articles*	*#properly classsified*	*% properly classsified*
Finance	555	364	66%
Foreign policy	341	260	76%
National Chronicle	273	36	13%
Local chronicle	943	746	79%
Sport	738	573	78%
Entertainment	354	278	79%

A Definition for Association Rules

Given a set of records each consisting of multiple elements belonging to a given collection

It produces rules of dependence that predict the occurrence of one of the elements in the presence of others.

TID	*Record*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Beer, Diapers, Milk

Rule:

{Milk} --> {Coke}

{Diapers, Milk} --> {Beer}

Association Rules: Application 1

Marketing and sales promotion:

- Suppose you have discovered the association rule

{Bagels, … } --> {Potato Chips}

- **Potato Chips as a consequent:** the information can be used to understand what actions to take to increase its sales
- **Bagels as an antecedent:** the information can be used to understand which products might be affected if the store interrupts the sale of Bagels

Association Rules: Application 2

Arrangement of the goods.

- Goal: Identify products purchased together from a sufficiently large number of customers.
- Approach: uses data from tax receipts to find dependencies between products.
- A famous association rule
  - If a customer buys diapers and milk then they will most likely buy bee
  - So do not be surprised if you find the beers beside the diapers!

Association Rules: Application 3

Inventory management:

- Goal: A household repair company wants to study the relationship between reported malfunctions and spare parts required to properly equip their vehicles and reduce visits to their homes.
- Approach: Processes the data about the spare parts used in the previous assistances to look for co-occurrence patterns.

A Definition for Regression

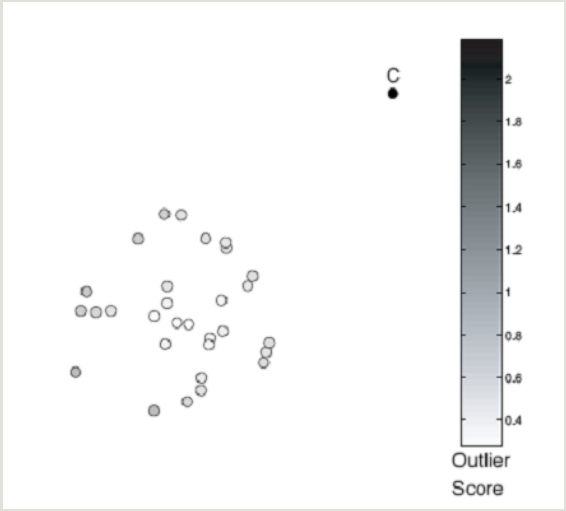
Predict the value of a continuous variable based on values of other variables assuming a linear / nonlinear dependency pattern.

Examples:

- Predict the sales revenue of a new product based on advertising investments.
  Predict the wind speed depending on temperature, humidity, atmospheric pressure
- Prediction of the stock market trend.

A Definition for Outlier Detection

Identify deviations from normal behavior



Applications:

- Identification of fraud in the use of credit cards
- Identification of network intrusions

Data Mining Bets

- Scalability
- Multidimensionality of the data set
- Complexity and heterogeneity of the data
- Data quality
- Data Properties
- Privacy Keeping
- Processing in real-time

A Methodology for Data Mining: CRISP-DM

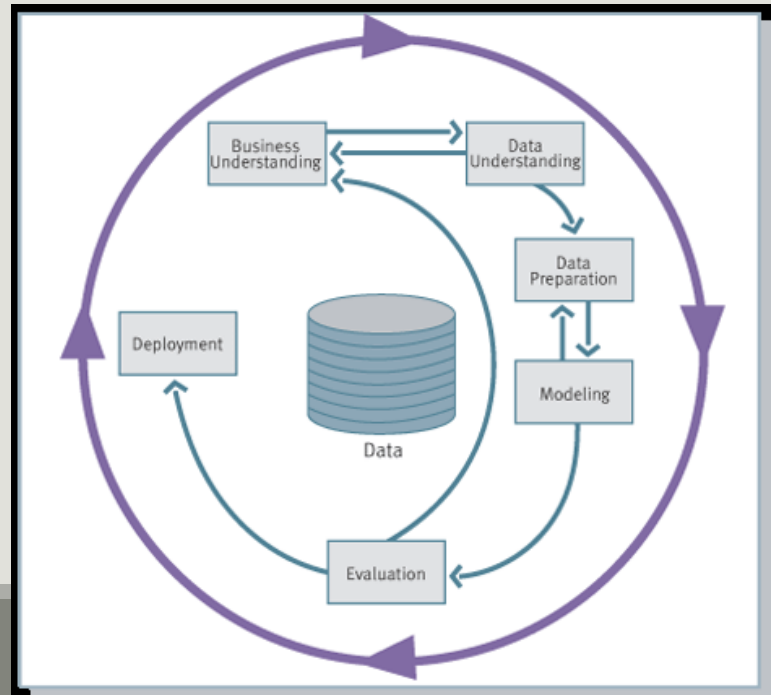
A Data mining project requires a structured approach; choosing the best algorithm is only one of the success factors

The **CRISP-DM** methodology is one of the most structured proposals to define the fundamental steps of a Data Mining project

The six stages of the life cycle are not strictly sequential.

Turning back on activities already
done it is often necessary

http://www.crisp-dm.org/



CRISP-DM steps

1) **Understanding the Application Domain:** understanding project goals from the user's point of view, translate the user's problem into a data mining problem, and define a project plan
2) **Understanding the data:** preliminary data collection aimed at identifying quality problems and conducting preliminary analyzes to identify the salient characteristics
3) **Data Preparation:** includes all the tasks needed to create the final dataset: selecting attributes and records, transforming and cleaning data

CRISP-DM steps

4) **Model Creation:** Several data mining techniques are applied to the dataset also with different parameters in order to identify what makes the model more accurate
5) **Evaluation of Model and Results:** The model(s) obtained from the previous phase are analyzed to verify that they are sufficiently precise and robust to respond adequately to the user's objectives
6) **Deployment:** The built-in model and acquired knowledge must be made available to users. This phase can therefore simply lead to the creation of a report or may require implementation of a user-controlled controllable data mining system