



^{sa}
Pepite

From Data to Performance

Introduction to data analysis





Why data mining ?

◆ Understand the past

- ◆ Explain the behavior of key performance indicators
- ◆ Transform ad-hoc know how into formal operation rules
- ◆ Identify conditions where operation is better
- ◆ Identify weaknesses and failure root causes

◆ Manage current situations

- ◆ Take better decisions for operation
- ◆ Detect drift of performance
- ◆ Check efficiencies of improvement actions

◆ Forecast the future

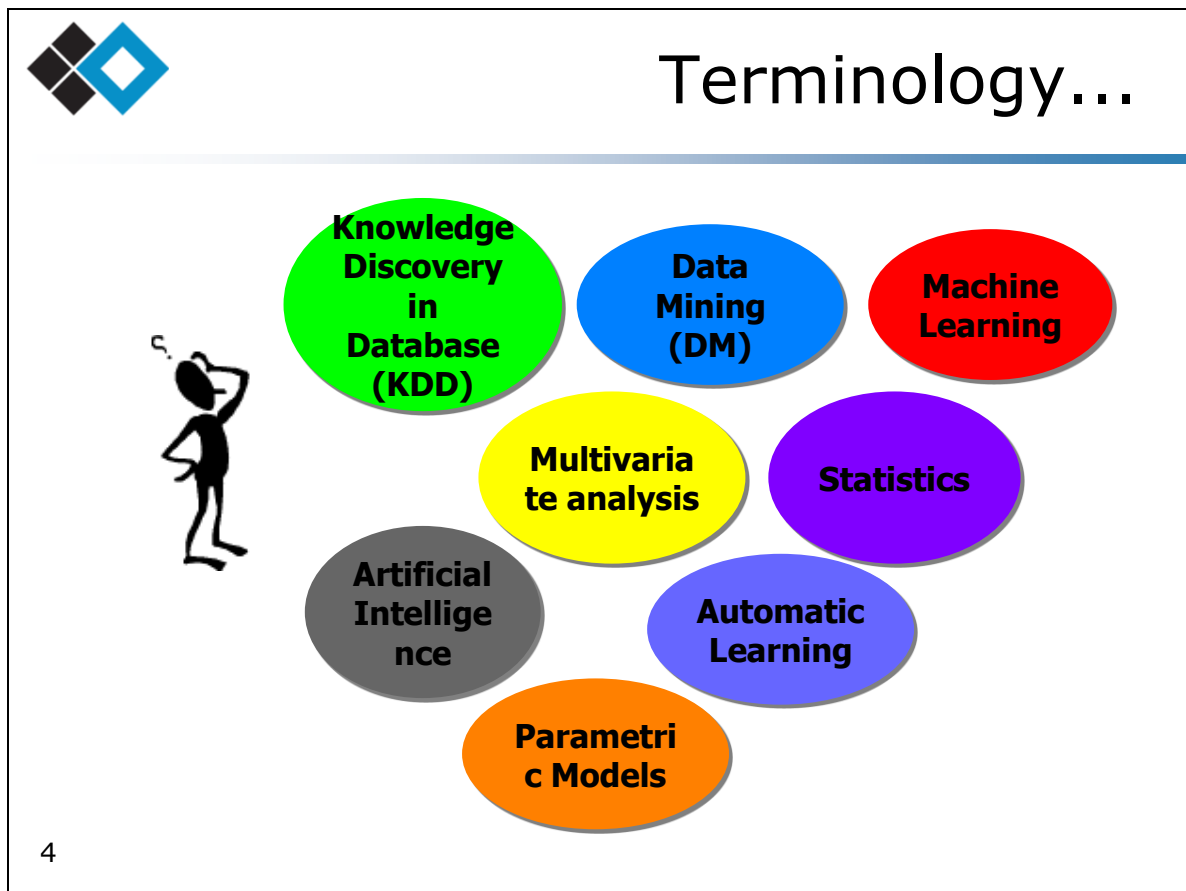
- ◆ Forecast future behaviors and performance
- ◆ Schedule improvement and actions and countermeasures

2





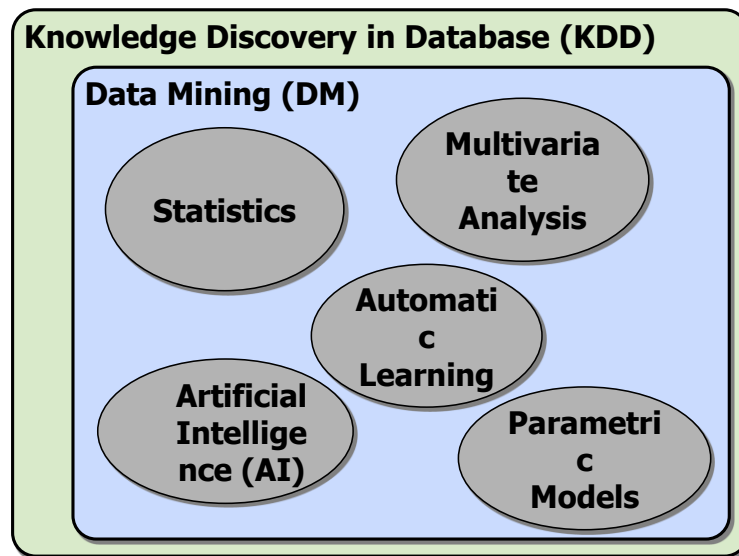
-
1. Some key definition
 2. Data analysis tools
 3. Methodology
 4. The phases, step by step
 5. Key success factors



When talking about data mining it is not always clear what this really means: are we talking about an algorithm, a methodology, a process? in the next slides we will try to clarify a little bit these concepts....



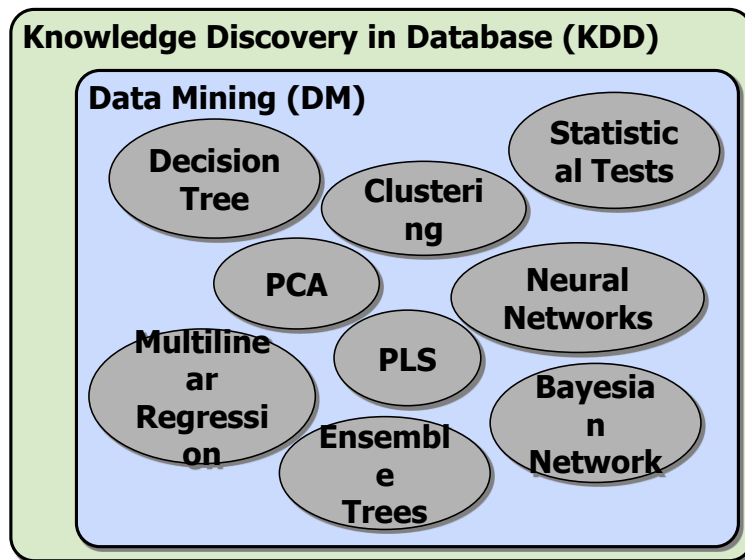
Terminology - The "schools"



First of all it is very important to understand that data mining is a part of a broader concept called Knowledge Discovery in Database. Data Mining is composed of several tools that can be clustered in families (statistics, automatic learning,...).



Terminology - The technologies



6

There are plenty of algorithms available for Data Mining. An algorithm can also have plenty of variants minor or major. Considering the previous slide, we can really associate one algorithm with one family. For example automatic learning tools clearly use statistics...



Data Mining - Definitions

- ◆ Knowledge Discovery in Databases (KDD)

Complex **process** leading to the identification of new information, valid, understandable and actionable starting from database.

- ◆ Data Mining

Set of tools : data visualization, descriptive statistics, pattern recognition, artificial intelligence used to describe, classify, forecast and cluster data.

KDD is more a process; it is defined as a set of activities (data collection, data cleaning, validation, data mining, models deployment, etc...)

Data mining is an activity part of the Knowledge Discovery Process.



Definitions


◇ Database :

- Collection of objects (observations, instances, states)
 - Described by attributes (variables, measurements, parameters, factors) with numerical or discrete values
 - Organized in the form of a table:

| Object Nb | Timestamp | T° | Flow | Pressure | ... | Quality |
|-----------|-----------|----|------|----------|-----|---------|
| o-1 | 00:00:00 | 35 | 14.3 | 2.51 | ... | High |
| o-2 | 00:15:00 | 30 | 14.1 | 2.89 | ... | Medium |
| o-3 | 00:30:00 | 30 | 13.2 | 1.67 | ... | High |
| o-4 | 00:45:00 | 31 | 15.6 | 2.09 | ... | Low |
| o-5 | 01:00:00 | 36 | 14.0 | 2.56 | ... | |
| o-6 | 01:15:00 | 39 | 17.9 | 4.71 | ... | High |
| o-7 | 01:30:00 | 34 | 14.4 | 3.98 | ... | Medium |

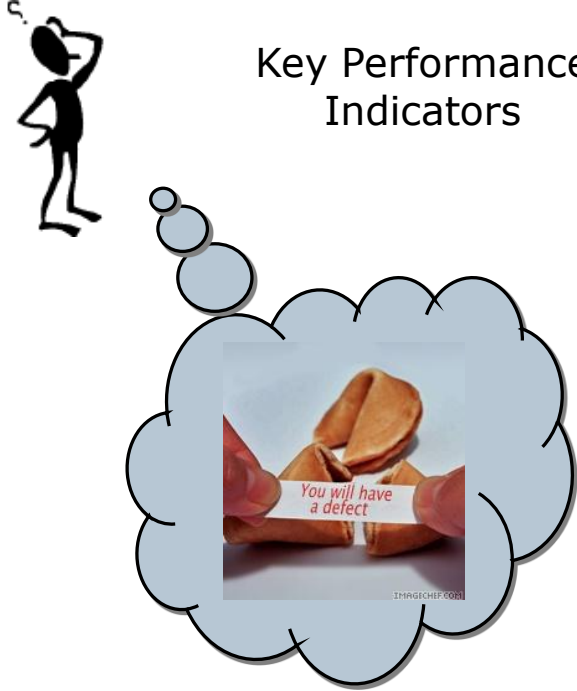
8

A database from the data mining point of view is a table containing data usually organized in rows for the objects (aka observation, individuals,...) and in columns for the attributes (aka parameters, variables, inputs, factors,...). The purpose of data mining is to discover trends, patterns, correlations, etc. among these data. In the modeling phase, data mining will learn a model that links some attributes (inputs) with other attributes (outputs).



What is a model ?

Key Performance Indicators



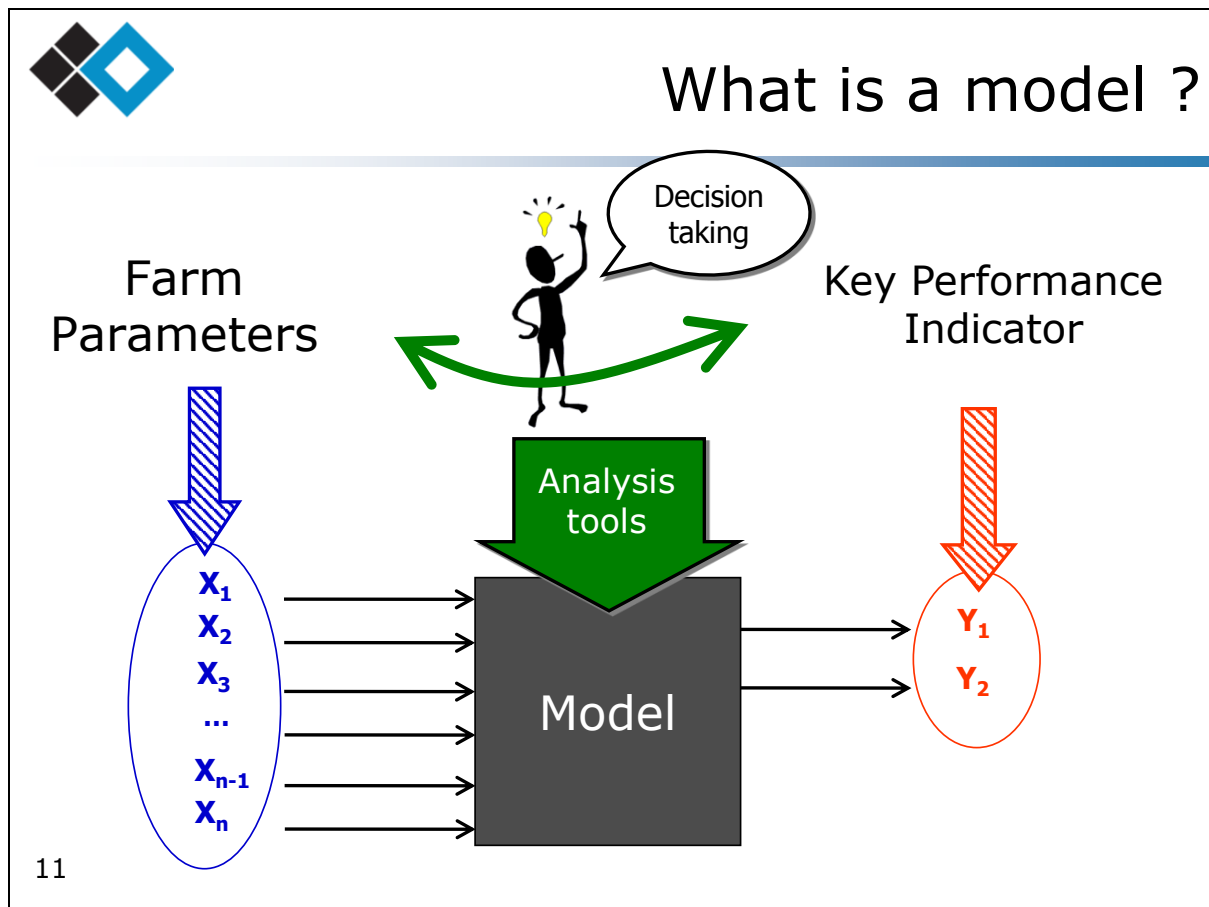
9

The slide features a logo in the top left corner, a title 'What is a model ?' in the top right, and the text 'Key Performance Indicators' in the center. A stick figure is shown thinking, with a thought bubble containing an image of a fortune teller's slip that says 'You will have a defect'. The number '9' is in the bottom left corner.

Key performance indicators help decisions makers to have a high-level insight to monitor the performance. Typically KPI are: yield, energy efficiency, malformation rate, etc. However it is not always very easy to understand the trends of the indicators and diagnose it.

The diagram is titled "What is model ?" in the top right corner. On the left, the text "Farm Parameters" is positioned. On the right, the text "Key Performance Indicator" is positioned. A stick figure stands in the center, with a question mark above its head. Two green arrows originate from the stick figure, one pointing left towards "Farm Parameters" and one pointing right towards "Key Performance Indicator". Below the stick figure is a thought bubble containing an image of a fortune teller's slip. The slip is held by two hands and has the text "You will have a defect" written on it in red. The number "10" is located in the bottom left corner of the diagram's frame.


On the other hand, it is more and more easy and cost effective to monitor a set of parameters and archive past values of these parameters in a database. Even if experts on the field are aware that these monitored parameters have an impact on the KPI, it is not an easy task to connect these parameters with KPIs.



In this context data mining can bring an effective solution to connect farm parameters with KPIs like malformation rate. Indeed, with data analysis we can automate the historical monitoring of parameters values and trends of the KPI. We call this process automatic learning of KPI models.



-
1. Some key definition
 - 2. Data analysis tools**
 3. Methodology
 4. The phases, step by step
 5. Key success factors



Type of (data mining) problem

- ◆ **Dependency** between parameters
 - Example : detect similar behavior in historical data, identify parameters having similar/independant behaviors
- ◆ **Regression**
 - Example : forecast the temperature of a tank depending on weather parameters
- ◆ **Classification**
 - Example : find root causes of malformation type
- ◆ **Clustering** (detect groups of similar objects or attributes)
 - Example : identify similar batches behaviors

13

Depending of the objective, there are several different types of problem we can tackle with data mining.

If we want to check if parameters are dependant and/or correlated we can use dependency tools like a correlation matrix, statistical tests, etc... This can be useful if we want to automatically select parameters that have an impact on a KPI.

Often we want to predict a continuous value (output or KPI) with other attributes (parameters or inputs). A typical example would be a temperature in a tank that we cannot measure because it would be too expensive to monitor it continuously. We can expect to predict this temperature with a regression model applied on indirect measurements. If the output we want to predict is not continuous (black, red, small, low, high,...) we will apply classification methods. For example, from historical data we could obtain a malformation risk model that could predict the malformation level of a batch (low, medium, high).

Another type of problem is the detection of a regime or of similar individuals. There are also techniques that will allow to detect automatically individuals with similar behavior. This could help for example fish farmers to detect production batches with a similar life history.



Analysis of dependence

- ◆ A dependence model describes the dependency between variables.
- ◆ There are 2 types of dependence models:
 - quantitativ
 - structural
- ◆ Nevertheless, just a few methods exist enabling to deduct a structure from the raw data, and these are limited to a small amount of data.
- ◆ Examples: correlation matrix analysis, dendrogramme analysis, principal components analysis (PCA), bayesian network

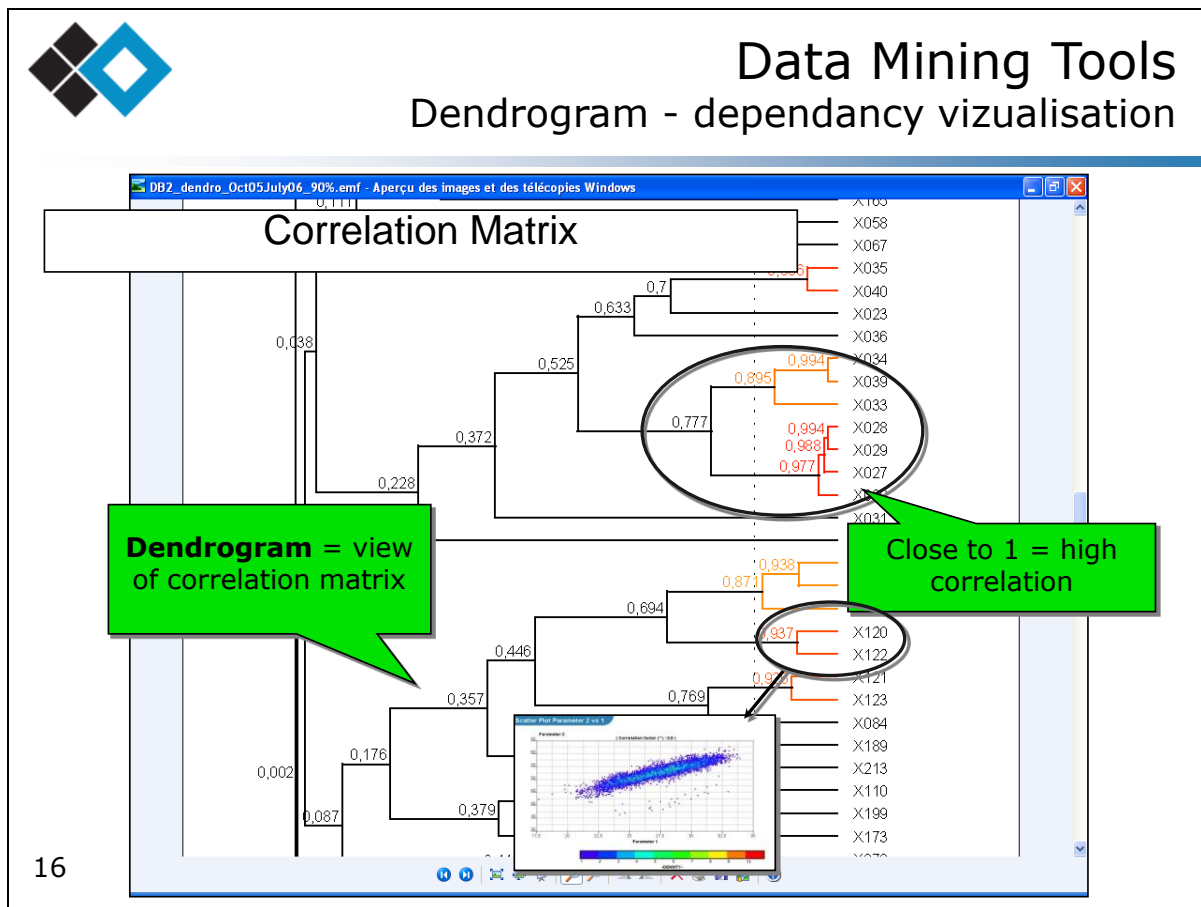


Tools...



15

In this part we will describe a few tools that are very useful for data mining tasks. Of course this is only a sample of the whole set of tools available in data mining.



A dendrogramme is a visualization tool for dependency problems. In the case shown here, the values shown at the nodes of the tree are the linear correlation coefficients (aka “Pearson”).

If two parameters are linearly correlated, the correlation coefficient is close to 1 (when the linear correlation is perfect, the correlation coefficient is 1). If the value at a node is 0.77, it means that all parameters at the right of the node are at least correlated with a correlation coefficient of 0.77.



Data Mining Tools Trees...



One of the most powerful tools in data mining are trees.



Data Mining Tools Trees...

- ◇ Induction trees are very popular because they can be easily interpreted by domain experts
- ◇ Approach : the data set is iteratively divided in subsets with a more less heterogeneous or variability of the output.

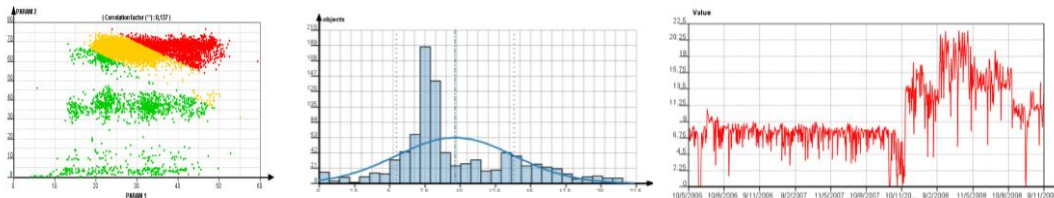


Trees are very useful models because they can extract automatically a set of rules from raw data. In the simple case shown here we can build a sex classifier model from a database of people with several parameters recorded for each individual (such height, weight, hair length,...). Decision tree learning algorithm will crunch all these data to identify what are the key parameters and test on this parameters that can help me to decide if a person is a man or a woman. Obviously the hair length is a trivial parameter in my sample.



Data Mining Tools Visualization

- ◆ Mandatory for these tasks:
 - Explore data
 - Particular regime
 - Abnormalities, drift
 - Obvious correlation
 - Interpret models
 - Understand the model
 - Performance of the model
 - Distribution and understanding of model errors
 - COMMUNICATE, COMMUNICATE, COMMUNICATE



19

Visualization is an essential tool for the preliminary analysis of numbers and raw data. Visualization is used to explore the data and check their rough properties. It gives already a first idea of the data quality.

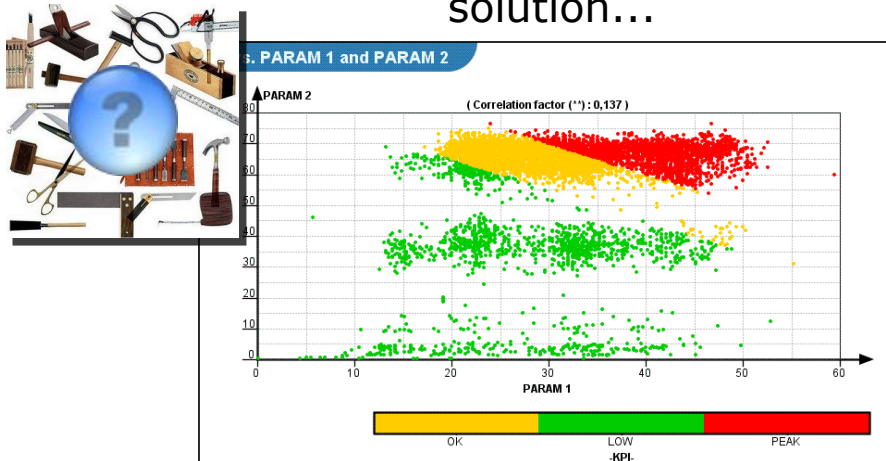
Visualization is also very useful to have an idea of the models' performance.

Typical tools : histogram, trends, scatter plots, etc... It is also important that these tools are dynamic so that we are able to select easily and zoom into data.



What is the best tool ?

Complex tools are not the bullet proof solution...



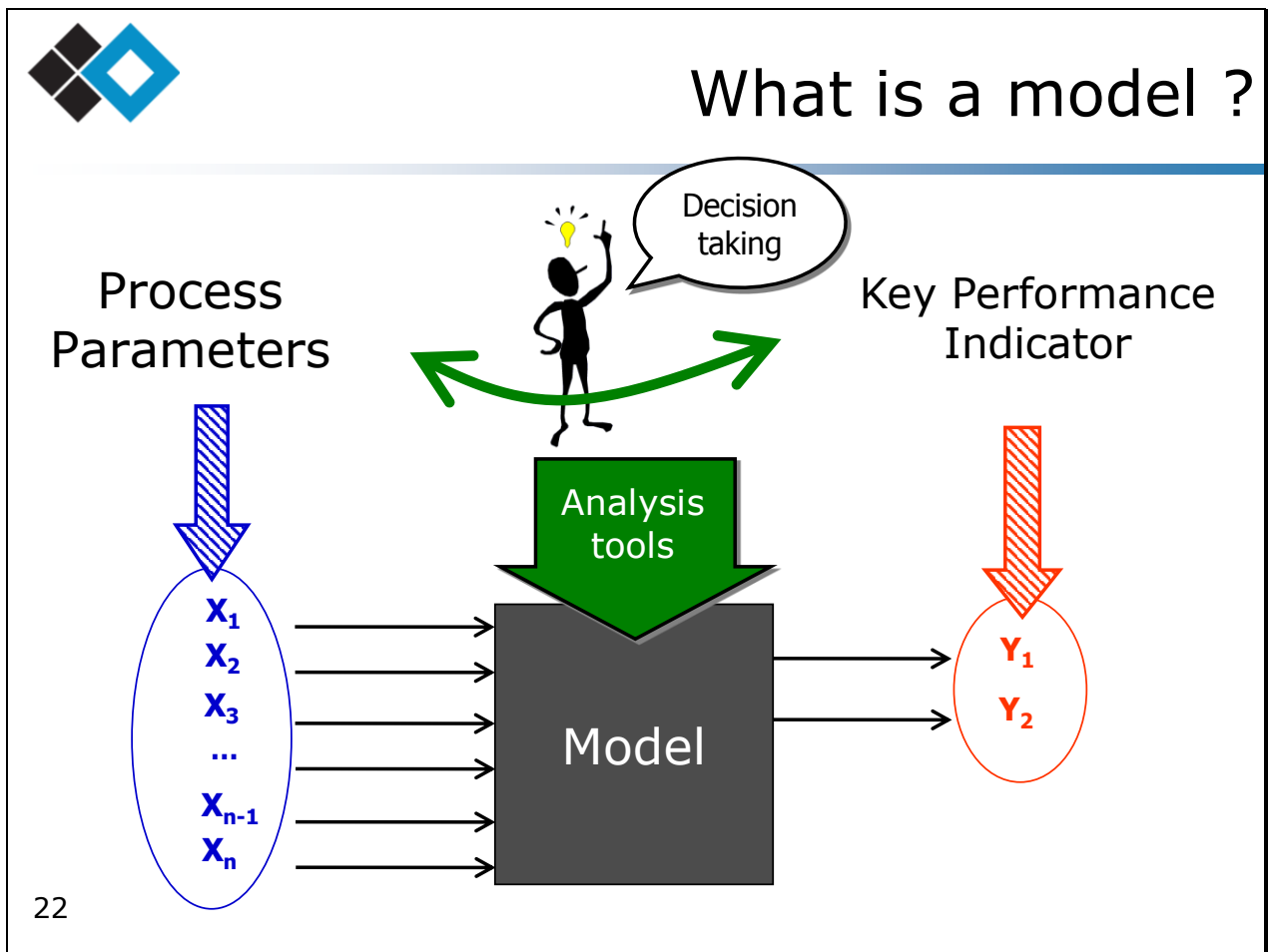
... but smart combination of tools will always bring most of the value

From our experience, we know that effective solutions are given not only by one method but by a combination of several tools. Sometimes simple tools can give valuable results on complex problems...

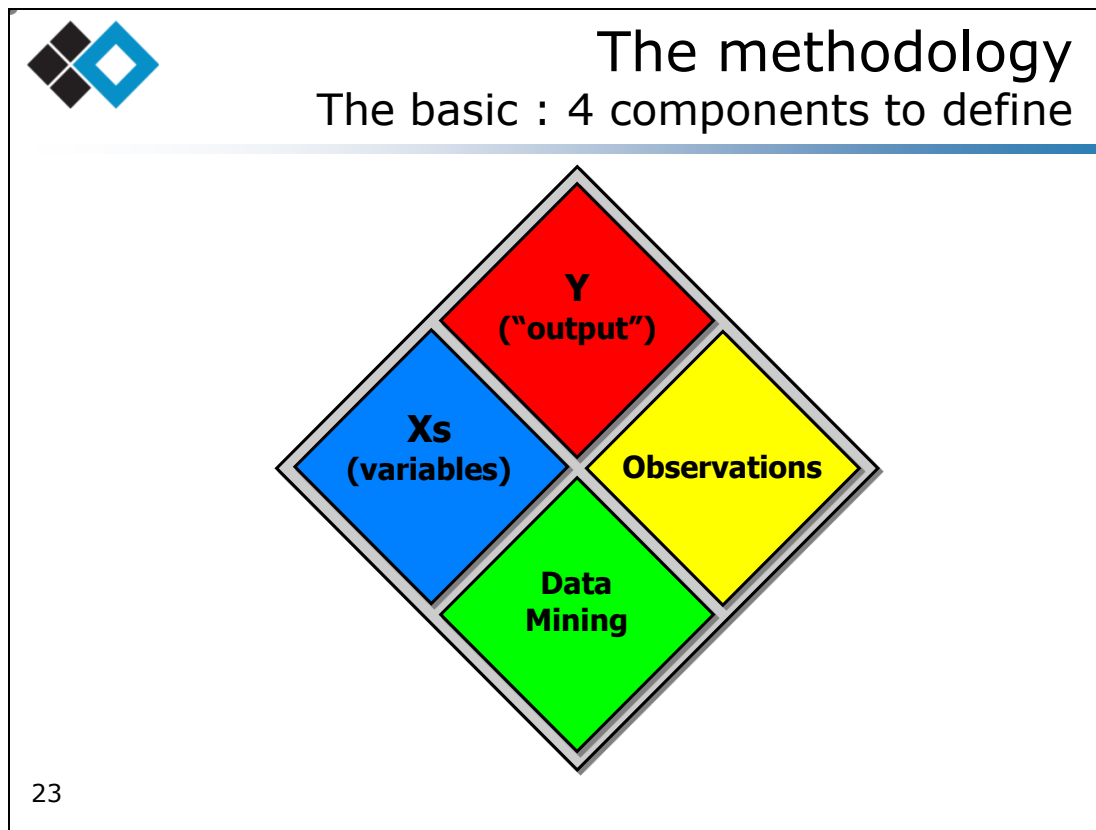


1. Some key definition
2. Data analysis tools
- 3. Methodology**
4. The phases, step by step
5. Key success factors

Having tools without methodology will probably fail or at best make you lose a lot of time.




Just as a reminder, what we want to achieve is to build a model that will link KPIs and input parameters.





Before building a model, we need to have these four components:

1. The input parameters/variables; some data mining tools will automatically discard a subset of these inputs because they have no impact on the outputs.
2. The outputs (KPIs for example)
3. The observations (with inputs and outputs recorded) that will be used to train and test the models
4. The data mining tools



Build model with data mining Where is the magic ?





No miracle
keystroke...

24

Having data and tools is not a enough... you cannot expect to transform data into gold only by pushing on a model keystroke! Data mining is not magic...



Data mining : where is the magic ?

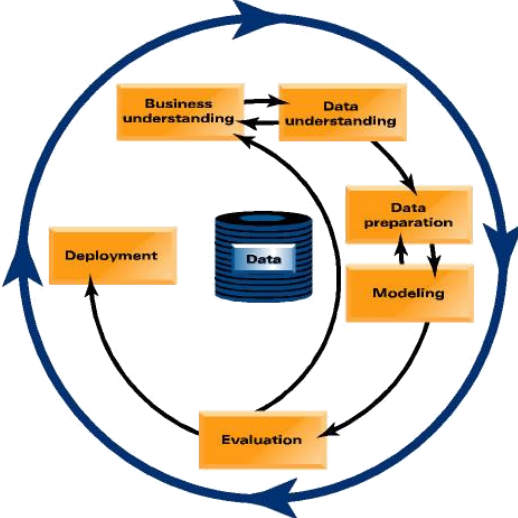


“Garbage in...

...garbage out.”

25

One of the main reasons of this failure of magic is that you don't have the control on data quality... whatever the tools you are using you won't be able to increase the information level contained in your hat. At best you can get rid of the noise but if there is only garbage in your data... you will end up with nothing than transformed garbage...



The diagram illustrates the CRISP-DM methodology as a continuous cycle. It features six orange rectangular boxes arranged in a circle, connected by a large blue circular arrow pointing clockwise. The boxes are labeled: 'Business understanding' (top-left), 'Data understanding' (top-right), 'Data preparation' (middle-right), 'Modeling' (bottom-right), 'Evaluation' (bottom), and 'Deployment' (middle-left). In the center of the cycle is a blue cylinder icon labeled 'Data'. Bidirectional arrows connect 'Business understanding' and 'Data understanding'. Bidirectional arrows also connect 'Data preparation' and 'Modeling'. A large blue arrow at the top of the cycle points from 'Business understanding' to 'Data understanding'. A large blue arrow at the bottom of the cycle points from 'Evaluation' to 'Deployment'. A large blue arrow on the right side of the cycle points from 'Modeling' to 'Data preparation'. A large blue arrow on the left side of the cycle points from 'Deployment' to 'Business understanding'. The text 'CRISP-DM = A standard Data Mining Process' and 'CROSS Industry Standard Process for Data Mining' is positioned above the diagram. The number '26' is in the bottom-left corner, and the URL 'www.crisp-dm.org' is in the bottom-right corner.

The methodology

Data Mining is sustained by a process

CRISP-DM = A standard Data Mining Process
CROSS Industry Standard Process for Data Mining

26

www.crisp-dm.org

We can see here that data mining is not an obvious task. This is why the industry has defined standard processes to support data mining tasks. CRISP_DM is the most used process for data mining; it describes very accurately the various tasks and phases of a data mining project as well as the interactions between these tasks.

"All models are wrong,
some are usefull"

George Box

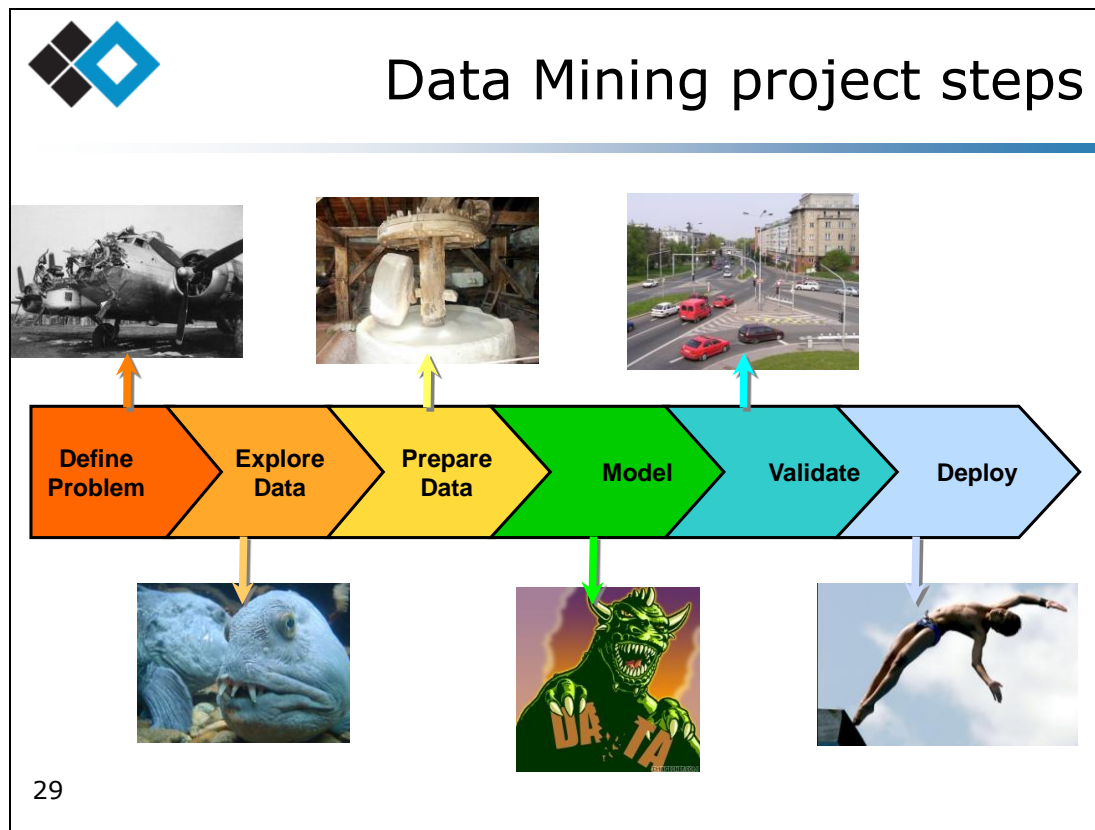


A little bit of humility....

George Edward Pelham Box (18 October 1919 –) is one of the most influential statisticians of the 20th century and a pioneer in the areas of quality control, time series analysis, design of experiments and Bayesian inference.



-
1. Some key definition
 2. Data analysis tools
 3. Methodology
 - 4. The phases, step by step**
 5. Key success factors



Following the CRISP-DM we can divide a data mining project into different phases.

Problem definition : this is where you transform a business objective/problem in a type of data mining problem. You will address here these typical activities: assess the data available (the "X" and "Y") and how these data can be used, collect the data to build a database, etc.

Explore data : at this step we use a lot data visualization to check the quality of the data and understand obvious correlation. Some data might be discarded (bad measurements, too many missing values, etc...).

Prepare the data : some available data needs to be prepared so that they can be fed to the data mining tools; typical preparation is: replace missing values, filtering, discretization, fast fourier transformation, etc. Typically at this phase, we also select the data subsets used for training the model and for testing it.

Model : at this phase we set up the learning algorithm and we feed it with the training subsets. Normally this phase is very quick.

Validate : we use the test subset to compute the statistical reliability of the model. If the model is not a black box model, we can use the expertise of the field knowledge to check that the model makes sense. If the validity phase concludes that the model is reliable enough we decide to deploy it “on-line”.


Deploy : at this stage, we can technically embed the models in the infrastructure used to monitor and control the operations. We can also deploy the model (that can a objective knowledge) through trainings.



1. Some key definition
2. Data analysis tools
3. Methodology
4. The phases, step by step
- 5. Key success factors**

We will describe here the key success factors that would minimize the risk of failure in a data mining project.

Basically there are two sources of risk : technical factors and organizational factors.



Technical success factors

- ◆ Data collection
 - ◆ Information system in place
 - ◆ Historian/datawarehouse
 - ◆ High availability of data
 - ◆ High quality of data

31

Technical success factors are important to assess because they can delay or even stop the data mining process.

The main issue regarding the technical factors is related to the data collection process. There are several factors that can impact the data collection process:

1st: the information system in place : how is the data organized and archived? Are datawarehouse/flat file/excel files used? How is the recorded data organized in the database? do we have an easy access to the data?

2nd: the data-warehouse and the historian : if a datawarehouse (relational database system) or historian is installed, data collection will be probably much easier.

3rd: the high availability of data is also very important
(some important parameters might not be archived even if they are monitored)

4th: quality of data : the “garbage in - garbage out” paradigm. If we do not have enough quality in the database (measurements error, failure in telecommunication,...) it will be difficult to extract valuable information from the data.



Human success factors

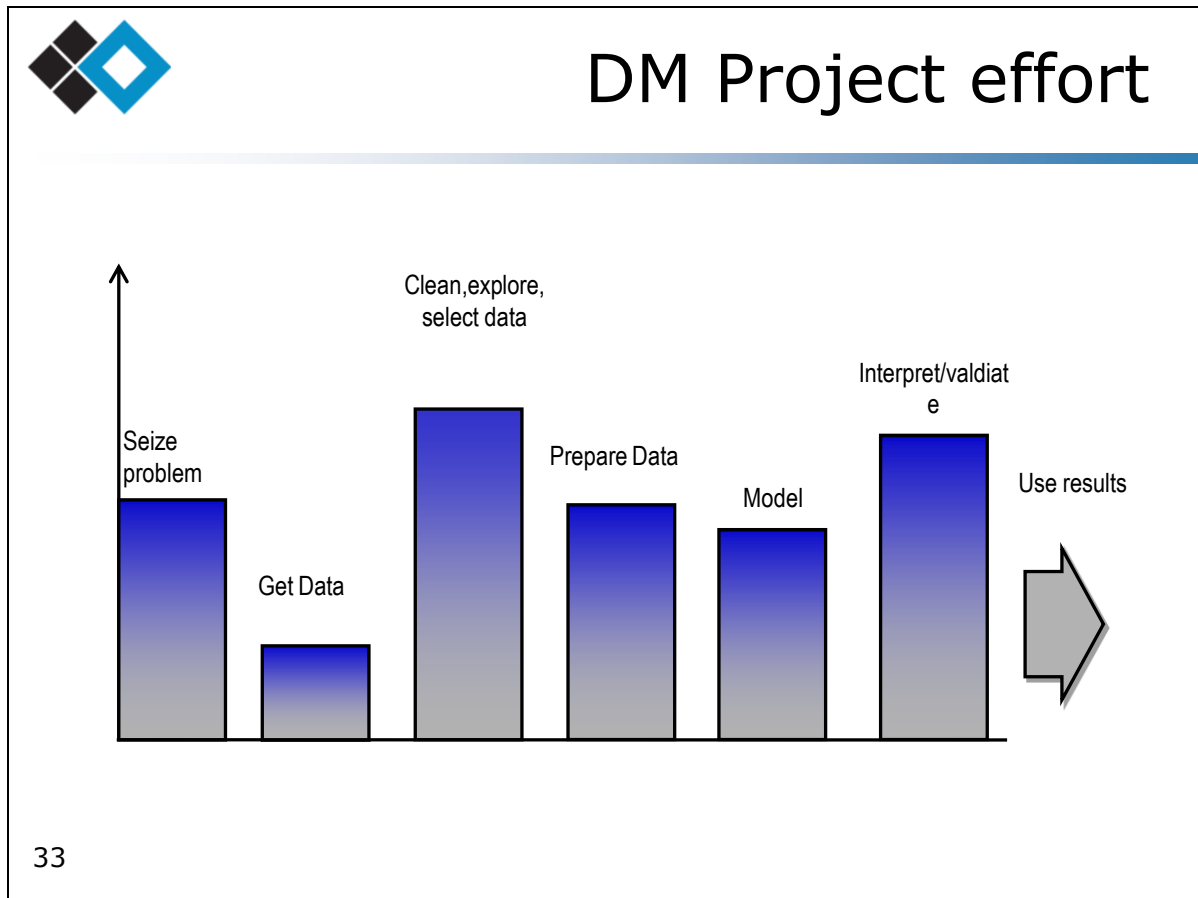
- ◆ Human resources
 - “Multi-hats” : technical, operation, IT
 - Continuous improvement culture
 - “Black belts” & “Green belts”
- ◆ Positive experience on applying data analysis

32

Human success factors are very important. Data mining requires the involvement of a team of experts (in IT, in data mining and in the business area where data mining is applied). Communication is key to minimize the risk and avoid any delay or too high expectations.

If Data Mining is in the hands of a team familiar with continuous improvement, there is much more chance to get faster and better results.

If there is already a success track of data mining it will also be much more easy to lead data manage projects to successful results.



This figure shows the expected effort required for the different phase of a project. IT is interesting to note that the real data-mining phase is not the most time-consuming.