The Data Mining Process :

a (not so) short introduction

Vincent Lemaire, Marc Boullé, Fabrice Clérot, Nicolas Voisine, Carine Hue, Françoise Fessant, Romain Trinquart

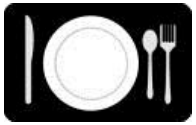
February 2017

Starting with an example

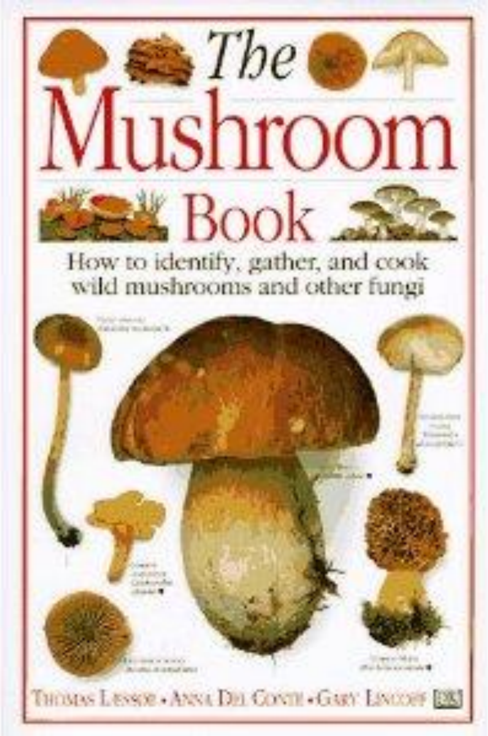
Pope Clement VII died in 1534 after eating poisoned mushrooms
(or so went the rumors at the time of his death)

Specify an objective

- Eat mushrooms, and avoid dying !
  - Thus : learn how to discern
    - Edible mushrooms



    - Poisoned mushrooms

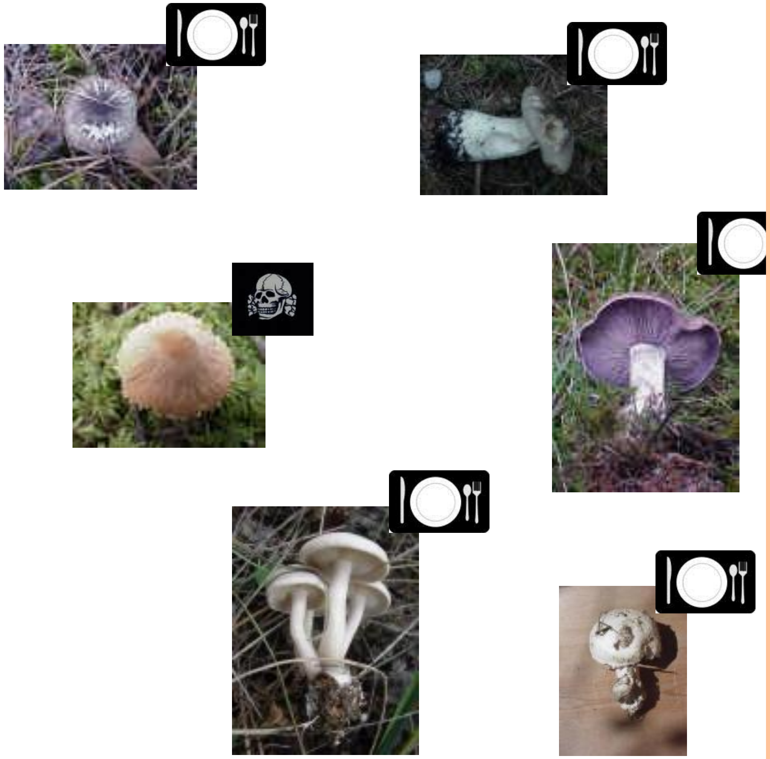

The Mushroom Book
How to identify, gather, and cook wild mushrooms and other fungi
Thomas Læssøe • Anna Del Conte • Gary Lincoff
DK

Data collection

- Picking of mushrooms
- Labeling by a pharmacist

Data collection

- Pick mushrooms
- Have them labeled by a pharmac



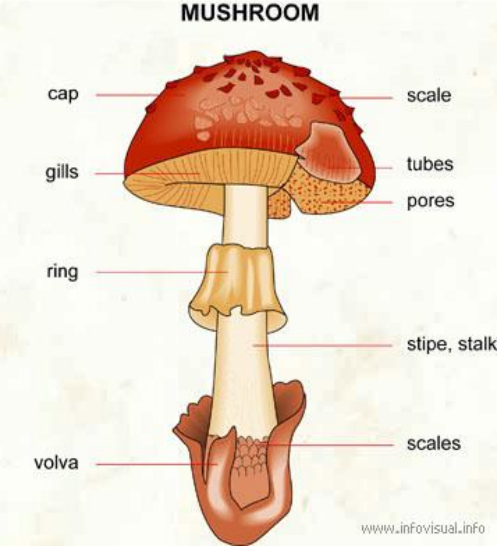
sounds obvious ?

but it’s not !

- need someone who knows where to pick mushrooms in the forest
- need a mushroom expert who knows the mushrooms properties
- need someone to specify how to pick mushrooms (do not pick the same one many times, for instance …)

Data Preparation

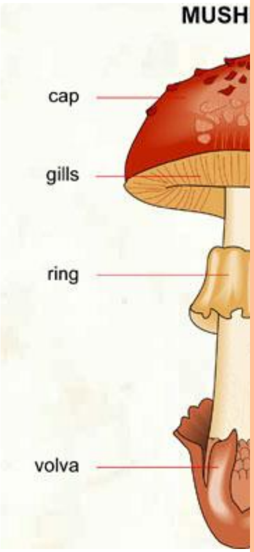
- Define a series of indicators
- Calculate the indicators for each picked mushroom
  - Cap Shape, Cap Surface , CapColor
  - Bruise
  - Odor
  - RingType
  - …



unrestricted

Data Preparation

- Define a series of indicators
- Calculate the indicators for each picked mushroom
  - Cap Shape, Cap Surface , CapColor
  - Bruise
  - Odor
  - RingType
  - …



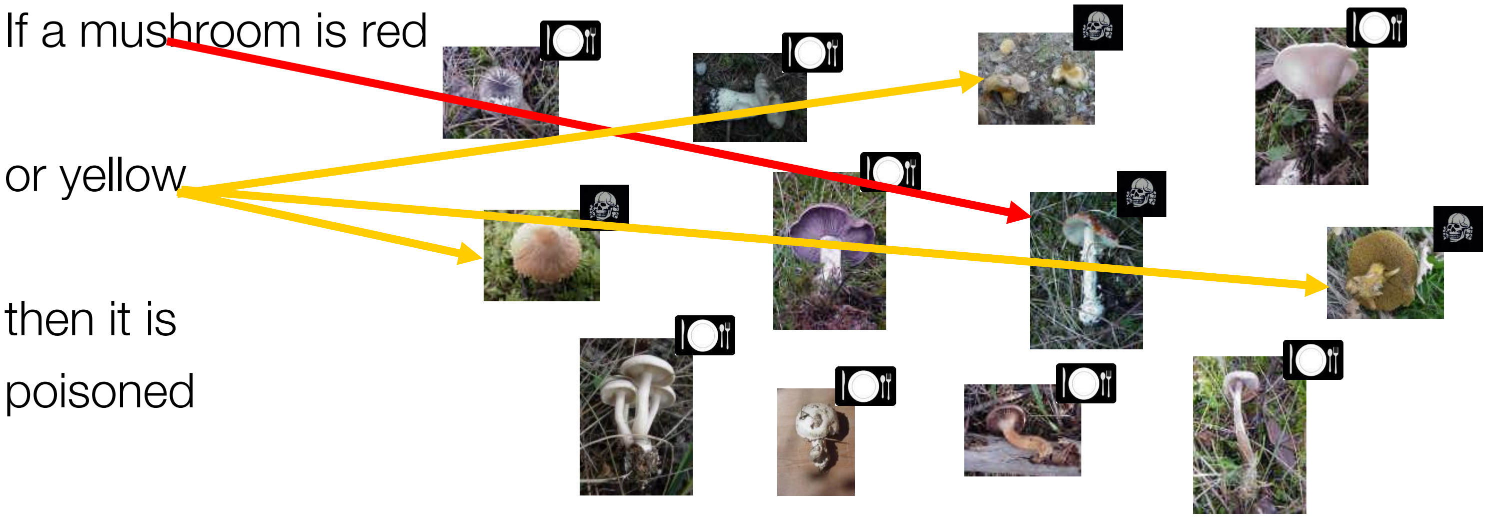
sounds obvious ?

but it’s not !

- still need the mushroom expert who knows the mushrooms properties
- need someone to prepare the data (don’t expect the mushroom expert to do that !)

Modeling

- After the analysis of the explanatory variables, the following rule is kept:



If a mushroom is red

or yellow

then it is

poisoned

Test of the results

- Test the rule on a new collect

sounds obvious ?

but it’s not !

- still need someone who knows where to pick mushrooms in the forest
- still need someone to specify how to pick mushrooms (do not pick the same ones as before, for instance …)

Test of the results

- Test the rule on a new collect

Test of the results

- Test the rule on a new collect

sounds obvious ?

but it’s not !

- need someone to cook the mushrooms

Test of the results

- Test the rule on a new collect

Test of the results

- Test the rule on a new collect

+

sounds obvious ?

but it’s not !

- you might have avoided killing a new pope by having the mushroom expert evaluate your modelling results !

Deploy the solution

- Following the evaluation deploy the solution …
- … or not !

Deploy the solution

- Following the evaluation deploy the solution …
- … or not !

sounds obvious ?

but it’s not !

- putting a model in production is <u>always</u> a risky business :
- « **better safe than sorry** » should be the data-miner’s motto

enough with popes and mushrooms !

let's proceed to our

data-mining …

some tasks and some definitions …

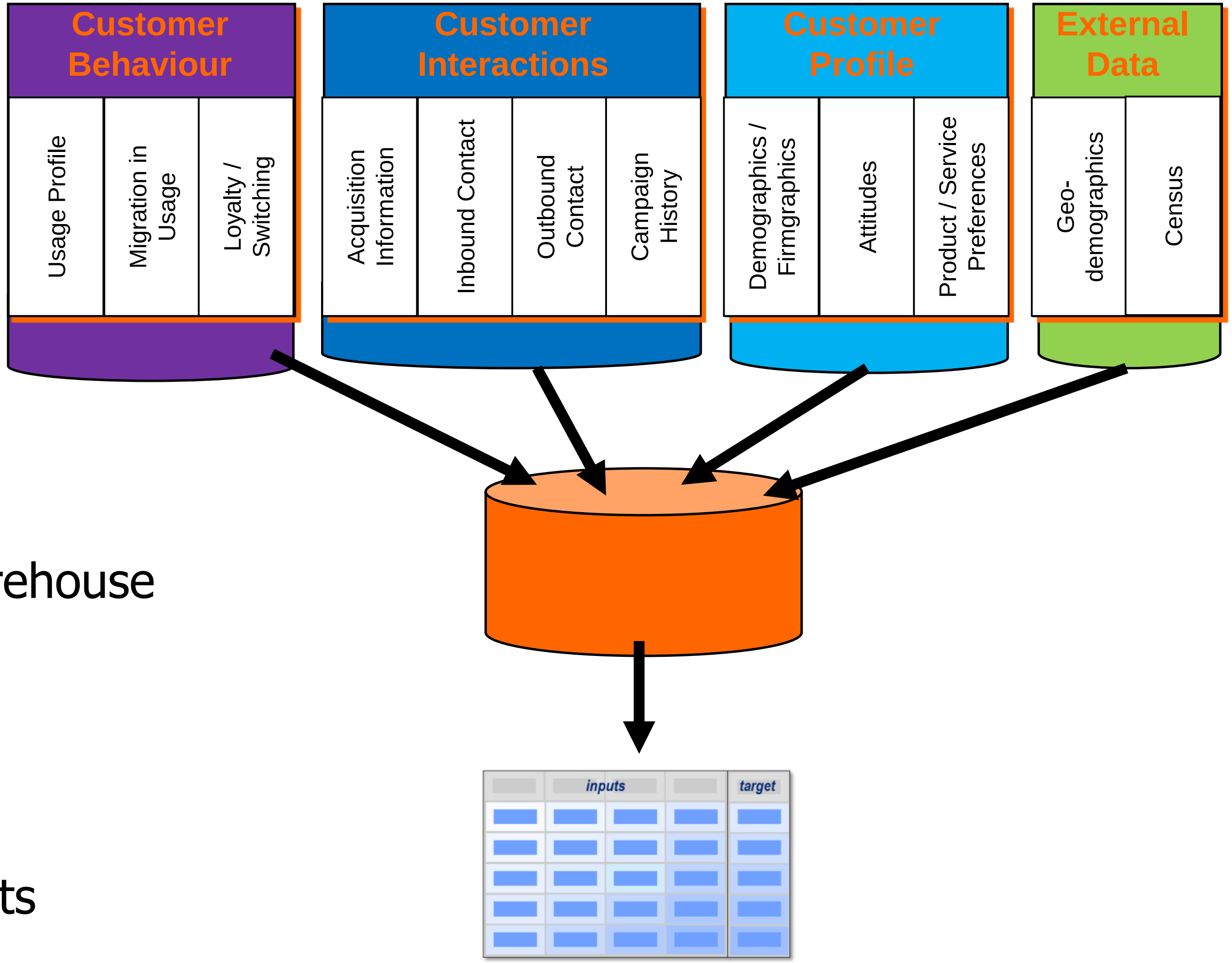
unrestricted

where do the data come from ?

unrestricted

From Data Sources to DataMarts

Data Sources



Customer Behaviour
- Usage Profile
- Migration in Usage
- Loyalty / Switching

Customer Interactions
- Acquisition Information
- Inbound Contact
- Outbound Contact
- Campaign History

Customer Profile
- Demographics / Firmgraphics
- Attitudes
- Product / Service Preferences

External Data
- Geo-demographics
- Census

More

History
Normalized
Detailed

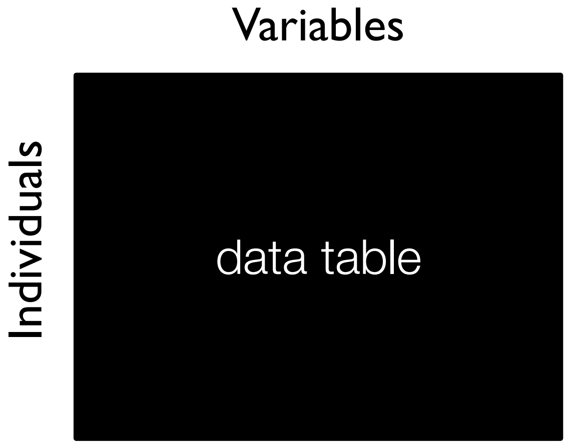
Data Warehouse

Data Marts

Less

data table

- row :
  - individual (or case)
- column :
  - variable (or feature, or agregate)
    - variable have « types »
      - categorical (red / green / blue)
      - numerical (temperature)
    - a more accurate classification exists
      - nominal, ordinal, interval, ratio
      - khiops reduces these four types to the two above



what do we do with a data table ?

Что дѣлать?

Наболѣвшіе вопросы нашего движенія

Н. ЛЕНИНА.

. . . „Партійная борьба придаетъ партіи силу и жизненность, величайшимъ доказательствомъ слабости партіи является ея расплывчатость и притупленіе рѣзко обозначенныхъ границъ, партія укрѣпляется тѣмъ, что очищаетъ себя“ . . . (Изъ письма Лассаля къ Марксу отъ 24 іюня 1852 г.)

Цѣна 1 руб.
Preis 2 Mark — 2.50 Francs.

STUTTGART
Verlag von J. H. W. Dietz Nachf. (G. m. b. H.)
1902

What's to be done
Lenin, 1902

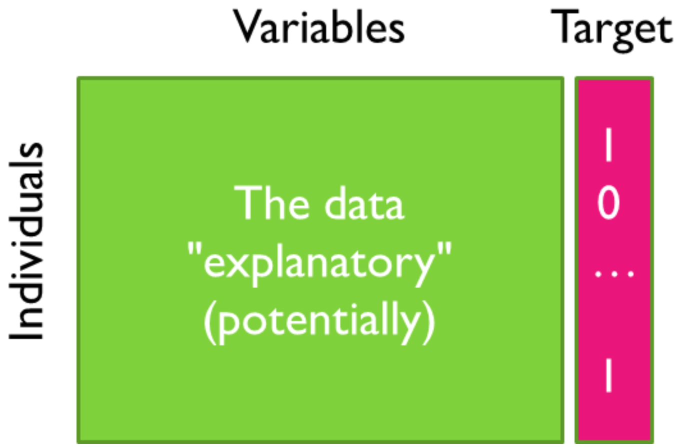
descriptive analysis overview

descriptive analysis

- get aware of the data set on a per aggregate basis
  - missing values, consistency
  - univariate statistics
- get aware of the relationships between few aggregates
  - mostly pairs of aggregates
    - scatterplots, correlation coefficients

Supervised analysis overview

Supervised analysis



Variables

Target

Individuals

The data "explanatory" (potentially)

1

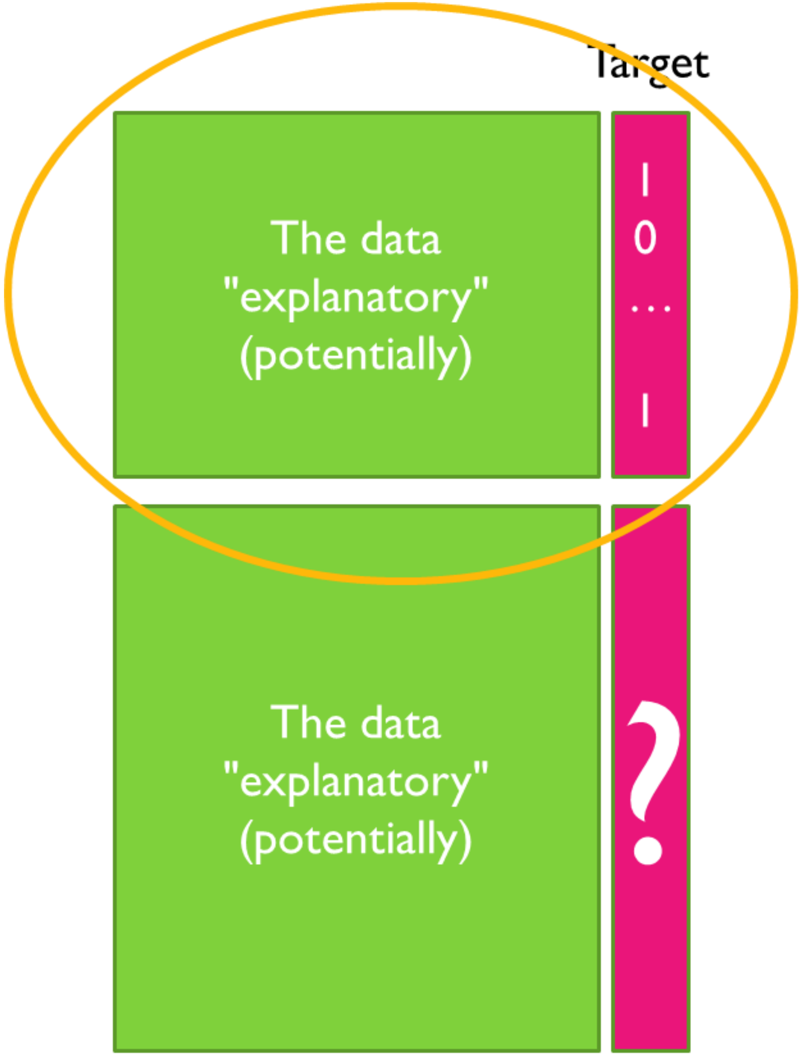
0

...

1

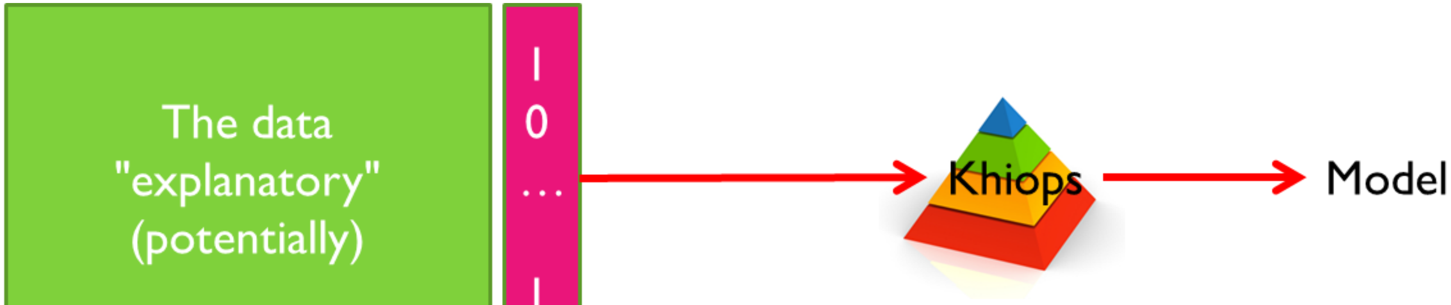
Symbolic target: "classification"
Continue target: "regression"

Supervised analysis


Target
The data "explanatory" (potentially)
1
0
...
1
The data "explanatory" (potentially)
?

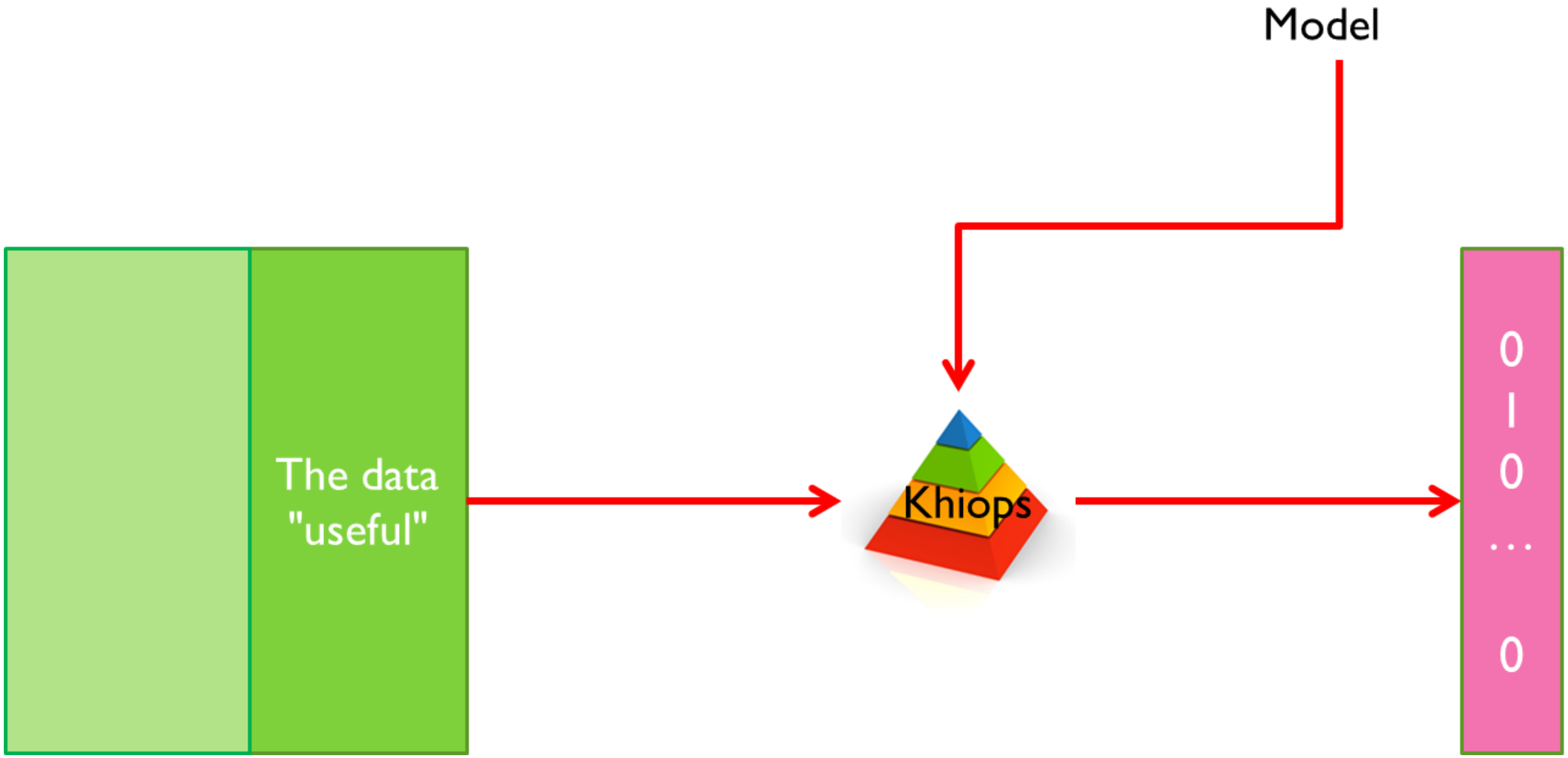
Supervised analysis
=
A target is identified
+
A representative sample of examples is available
+
We want to estimate the target on the rest of the base

Supervised Analysis: Modeling



Model:
Target_Estimated = $\mathbb{F}$(useful explanatory variables)

Supervised Analysis: deployment



unrestricted

exploratory data analysis overview

context and goal

- a data set
  - available as a data table
  - line = individual of the analysis
  - column= agregate describing the individuals
- « make sense » of the data set
  - fuzzy goal ?
  - can be made more accurate in different ways
    - can the data be described in a « simpler » space ?
    - can the individuals be grouped according to different « behaviours » ?
    - is this individual an « outlier » and why ?

exploratory data analysis versus supervised analysis

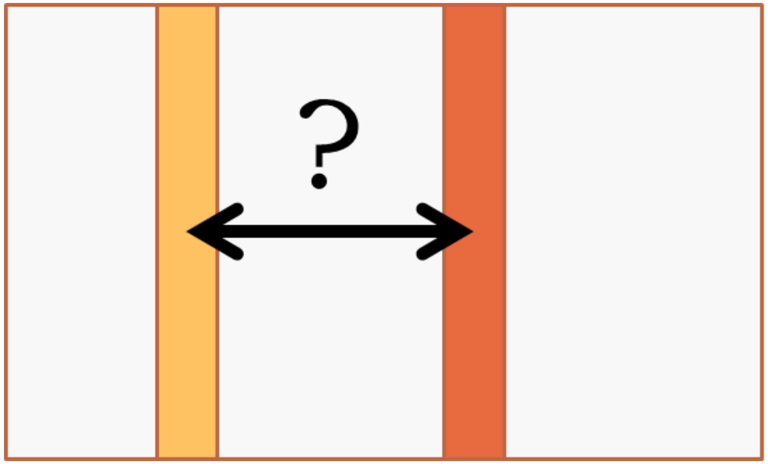
- supervised analysis
  - there is a particular aggregate (« target ») to be explained from the others (« explicative variables »)
  - the result is a model
  - the model performance can be objectively measured
    - as the reconstruction of the target
- exploratory data analysis = unsupervised analysis
  - all aggregates have the same role
  - there is no « objective » measure of performance
    - the quality of an exploratory data analysis lies in the knowledge the analyst gains on the data

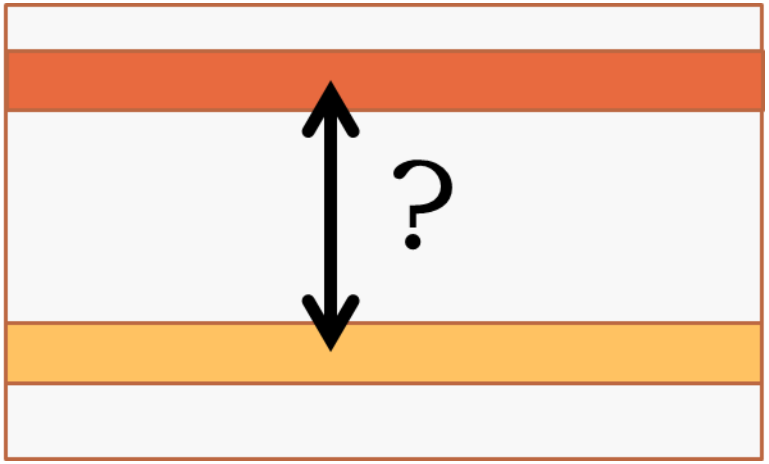
exploratory data analysis

- specifying the purpose of the analysis is crucial
    - there are as many analysis as there are possible purposes
        - at least as many evaluations as there are purposes

exploratory data analysis

- two main categories
  - build a « simpler » representation space
    - focuses on the relationships between aggregates
  - build a simpler description of the individuals (a few « typical behaviours »)
    - focuses on the links between individuals

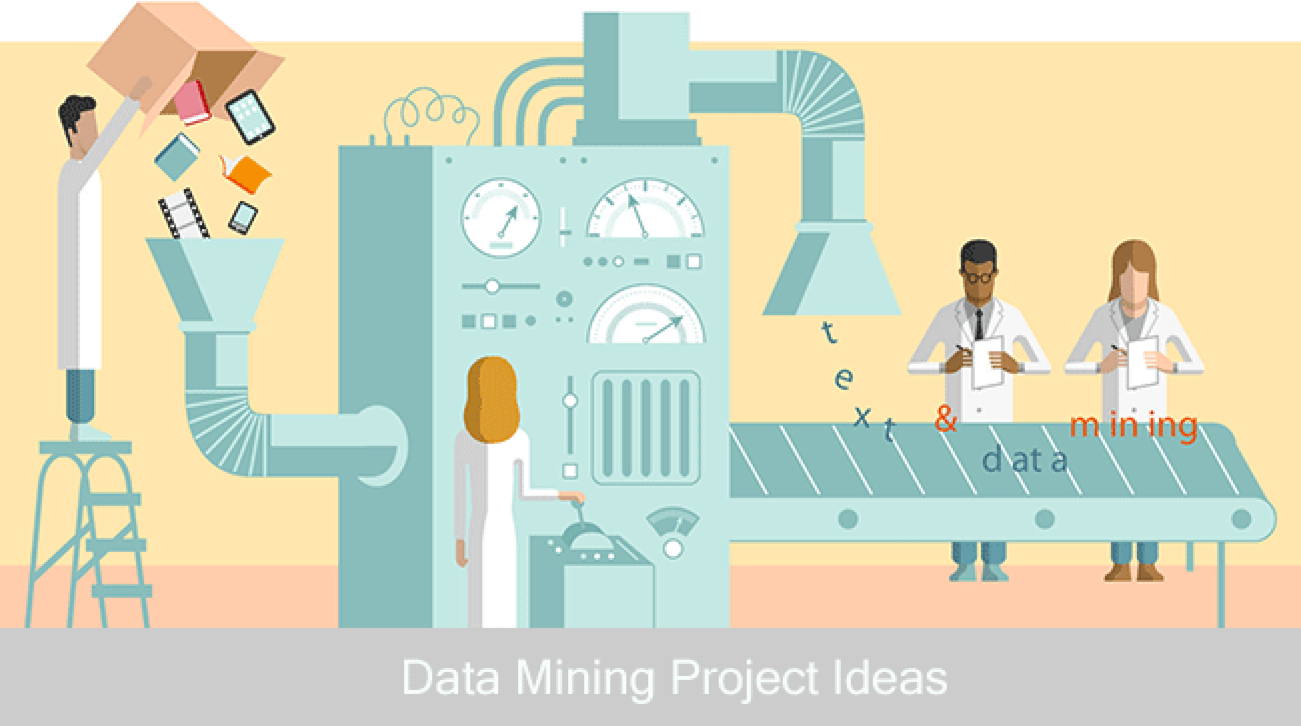




exploratory data analysis

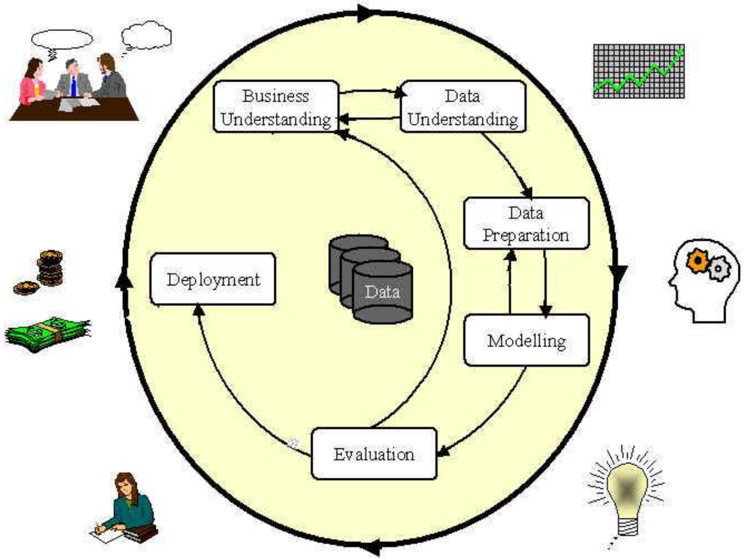
- the two approaches are often chained
  - simplify the representation space, then build a few typical behaviours in this space
  - build a few typical behaviours, then define a space allowing such behaviours to be described simply

the data-mining project

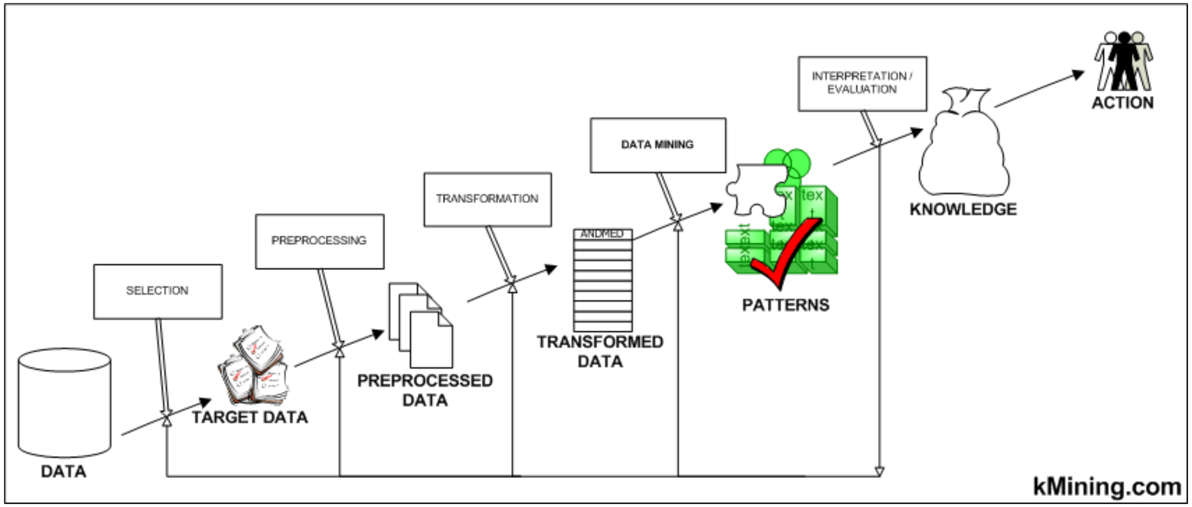


unrestricted

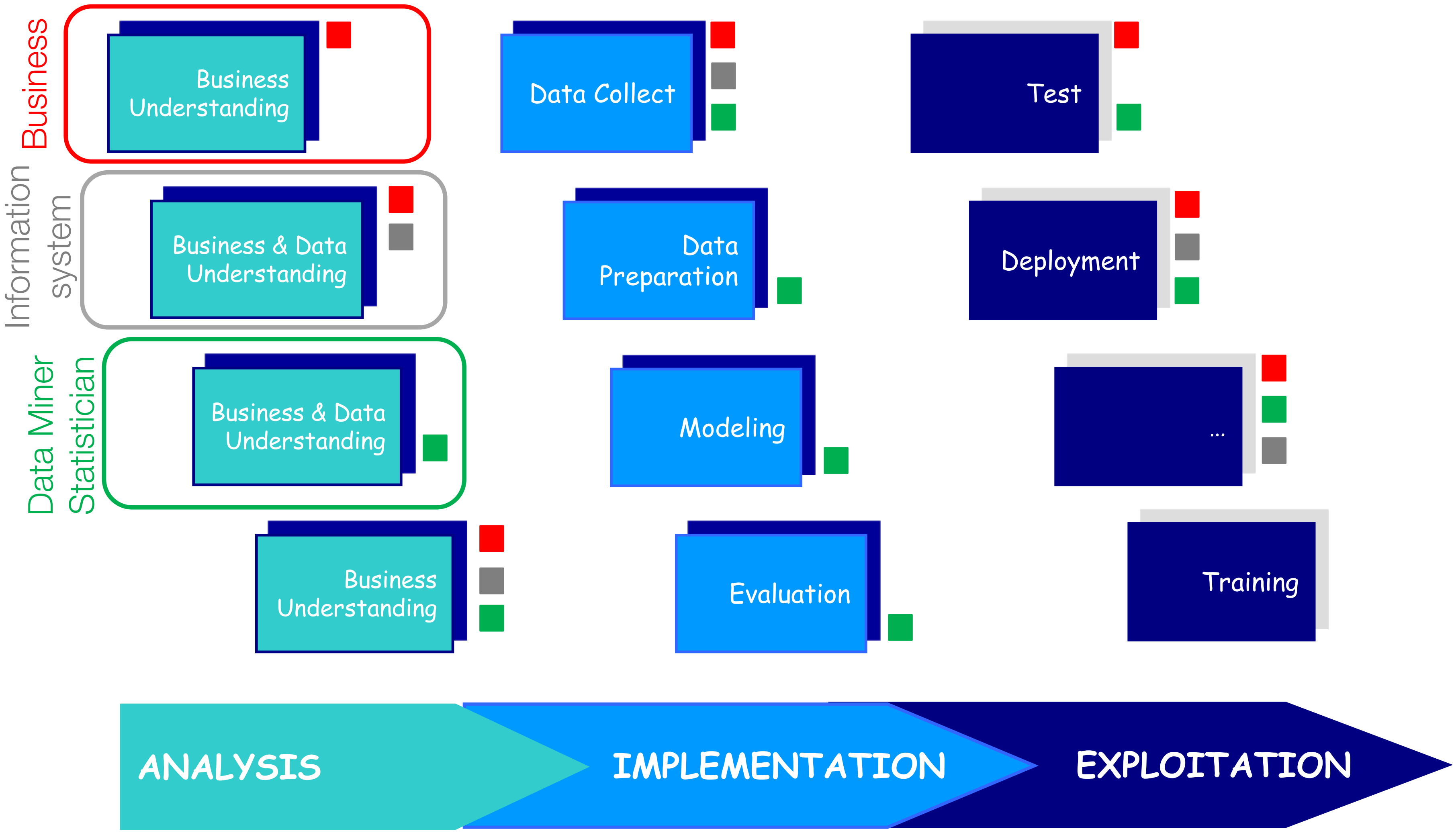
The Data Mining Project



Two complementary views

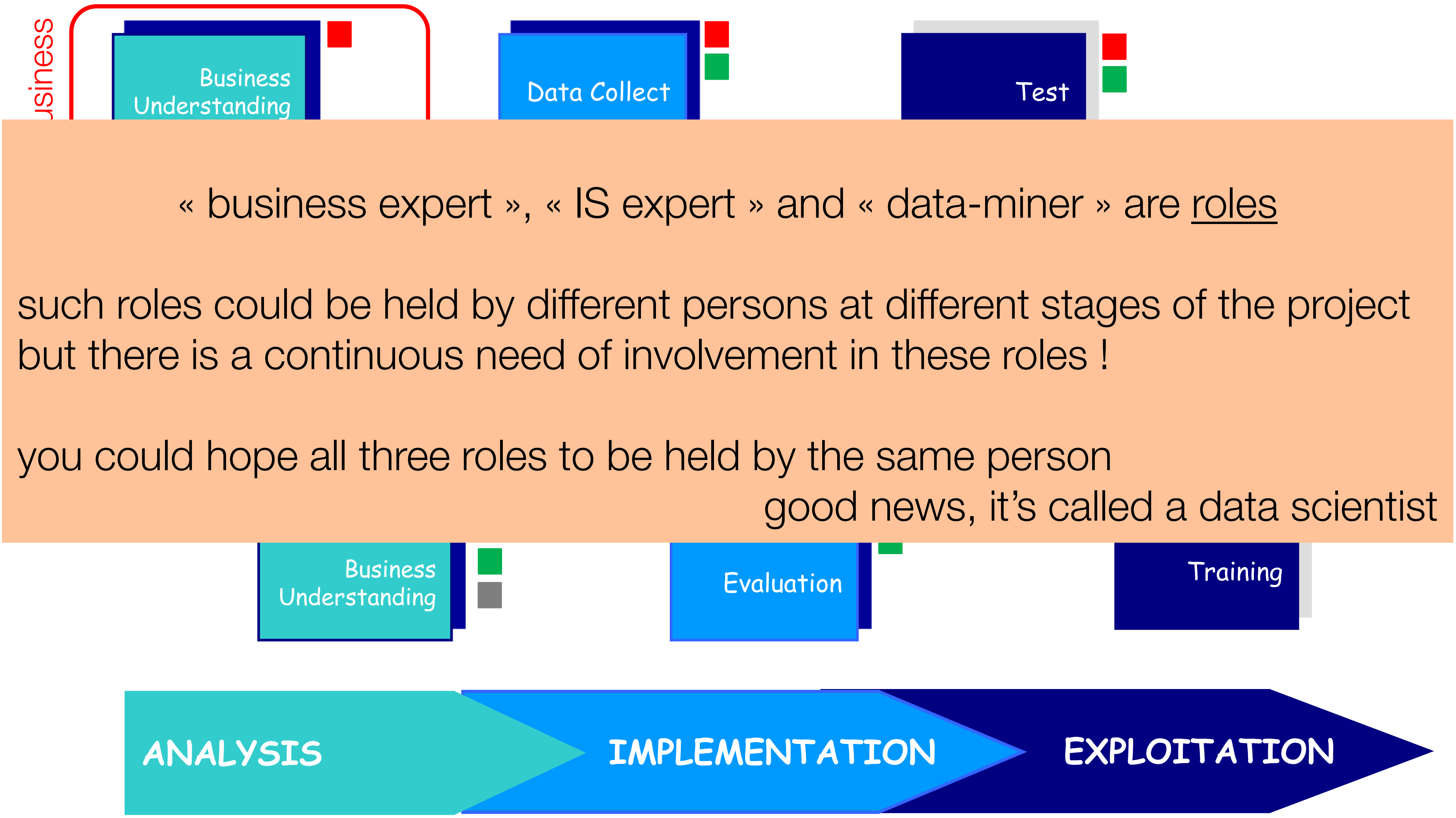


"The" Data Mining Project



unrestricted

“The” Data Mining Project



« business expert », « IS expert » and « data-miner » are <u>roles</u>

such roles could be held by different persons at different stages of the project but there is a continuous need of involvement in these roles !

you could hope all three roles to be held by the same person

good news, it’s called a data scientist

“The” Data Mining Project

« business expert », « IS expert » and « data-miner » are <u>roles</u>

such roles could be held by different persons at different stages of the project but there is a continuous need of involvement in these roles !

you could hope all three roles to be held by the same person

good news, it’s called a data scientist

bad news, data scientists do not exist

ANA

MPLEMENTATION

EXPLOITATION

antique artistic view of a data scientist
the Chimera of Arrezzo (V century BC)

“The” Data Mining Project

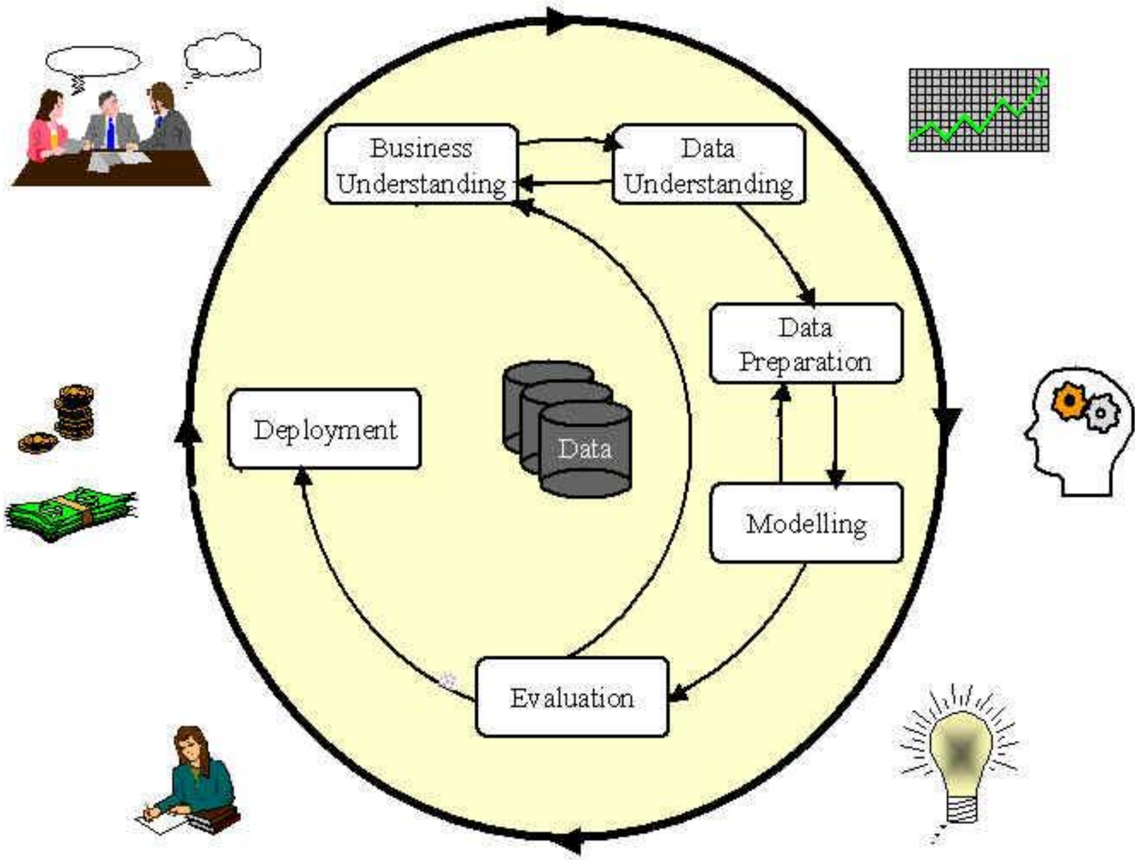
Business
Business Understanding
Data Collect
Test

warning : « IS expert » hides indeed <u>two roles</u>

- expertise on the <u>data sources</u>
  - obviously mandatory for the beginning of the process
- expertise on the <u>deployment constraints</u>
  - no need to build a memory-hungry model if there are strong memory limitations at the deployment stage !

- note : expertise on the data sources is also necessary at the deployment stage
  - no need to deploy a model if the data cannot be fetched at deployment time !

Step by Step



unrestricted

Business Understanding:

“Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.” (Oracle Definition)

Business Understanding:

“Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.” (Oracle Definition)

- Definition of a business perspective
  - For example: prediction of the customers who are going to churn
- Inventory
  - Success criteria (improvement of at least 50% of the targeting effectiveness)
  - Available resources (IS, expertise ...)
  - Constraints (legal issues, intelligibility results ...)
  - Time span for the first full iteration
- Formulation into a Data Mining project
  - Supervised classification : categorical target (i.e churn or not)
  - Regression : continuous target (age of customer)
  - Segmentation : No target

Business Understanding:

“Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.” (Oracle Definition)

- Definition of a business perspective
  - For example: prediction of the customers who are going to churn
- Inventory
  - Success criteria (improvement of at least 50% of the targeting effectiveness)
  - Available resources (IS, expertise ...)
  - Constraints (legal issues, intelligibility results ...)
- Formulation into a Data Mining project
  - Supervised classification : categorical target (i.e churn or not)
  - Regression : continuous target (age of customer)
  - Segmentation : No target

it might be much more complex …

in general, no business question is reducible to a single data-mining operation

therefore a data mining project consists in many data-mining operations pipelined in (hopefully) clever way

Business Understanding:

“Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.” (Oracle Definition)

- Definition of a business perspective
  - For example: prediction of the customers who are going to churn
- Inventory
  - Success criteria (improvement of at least 50% of the targeting effectiveness)
  - Available resources (IS, expertise ...)
  - Constraints (legal issues, intelligibility results ...)
- Formulation into a Data Mining project
  - Supervised classification : categorical target (i.e churn or not)
  - Regression : continuous target (age of customer)
  - Segmentation : No target

- Actions:
  - Evaluation
    - agree on one main kpi for success (and one only)
      - could be ROI if it is measurable
      - many other secondary kpis can also be defined
  - Definition of the work plan and milestones
    - identification of the key resources holding the three roles
  - Prior identification of data repositories
    - no data, no project
- Results:
  - Specifications
- Recommendations:
  - Ensure good involvement on the BU side
  - Clearly secure the key resources <u>for the time of the project</u>

Business Understanding:

"Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives." (Oracle Definition)

- Definition of a business perspective
  - For example: prediction of the custom who are going to churn
- Inventory
  - Success criteria (improvement of at lea 50% of the targeting effectiveness)
  - Available resources (IS, expertise ...)
  - Constraints (legal issues, intelligibility results ...)
- Formulation into a Data Mining project
  - Supervised classification : categorical targ (i.e churn or not)
  - Regression : continuous target (age of customer)
  - Segmentation : No target

- Actions:

sounds obvious ?

but it's not !

- this phase of the project is often overlooked or poorly treated
- « we have a business problem, let's rush to a data scientist and expect miracles ! » … voodoo analytics
- things are really improving, however
- but be careful anyway

Data Understanding:

“Start by collecting data, then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information.” (Oracle Definition)

Data Understanding:

“Start by collecting data, then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information.” (Oracle Definition)

- Identify the available data sources
- Describe macroscopic characteristics of the data
  - volume, nature, ...
- Explore the data (descriptive statistics)
- Check data quality (missing values ...)
- Check the data representativeness

- Implementation:
  - Meetings, investment of the BU side and the Information System
  - Working expertise of a statistician

Data Understanding:

"Start by collecting data, then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information." (Oracle Definition)

- Identify the available data sources
- Describe macroscopic characteristics of the data
  - volume, nature, ...
- Explore the data (descriptive statistics)
- Check data quality (missing values ...)
- Check the data representativeness

- Implementation:
  - Meetings, investment of the BU side and the Information System
  - Working expertise of a statistician

- Define the scope of the study : which individuals should be retained
  - keep all them or draw a sample (N)
- Create a dictionary to document "metadata" :
  - which native variable to use (L1)
  - which new variables to create (L2)
    - Construction of variables taking into account the business knowledge
- Build the data mart* :
  - a flat data table of (N) lines and (L1+L2) columns
  - draw a part of the data mart to create a 'modeling data table'

*A data mart contains often different domains of data such as customer, billing, uses, contacts

Modeling:

"Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed."

Modeling:

"Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed."

- Khiops aims at making this phase as automatic as possible
  - proof later in the tutorial !

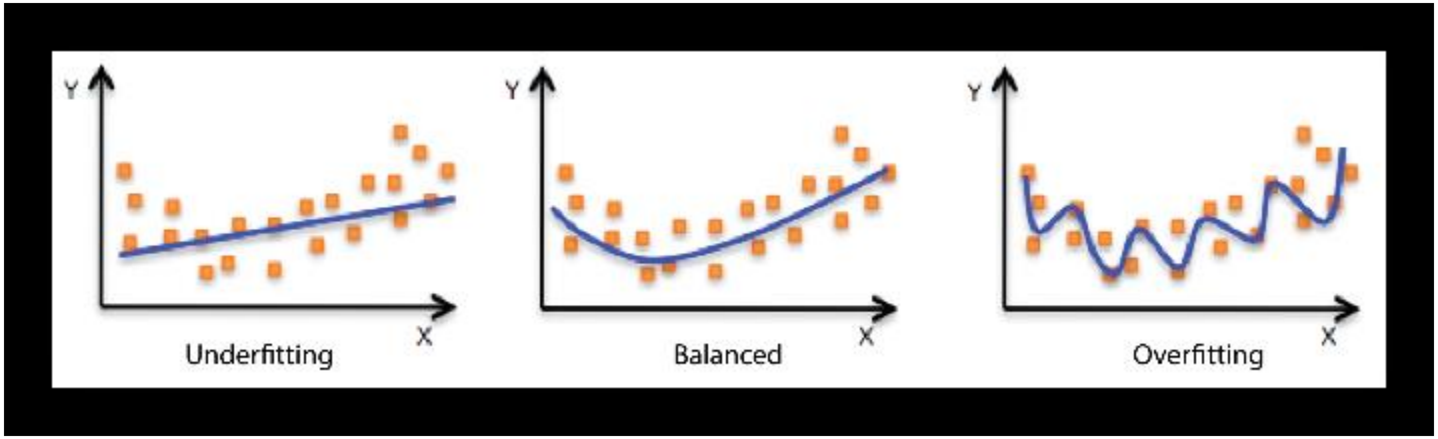
- there are many more tools and techniques than Khiops in a data-miner's toolbox
  - but this is not an introductory machine learning or data-mining course
  - we are concerned with the data-mining process …
  - … of which modelling is just a phase among others

- we shall not visit the zoo of data-mining tools today !

Modeling:

"Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed."

- the most common pitfall in the modeling phase
  - overfitting !



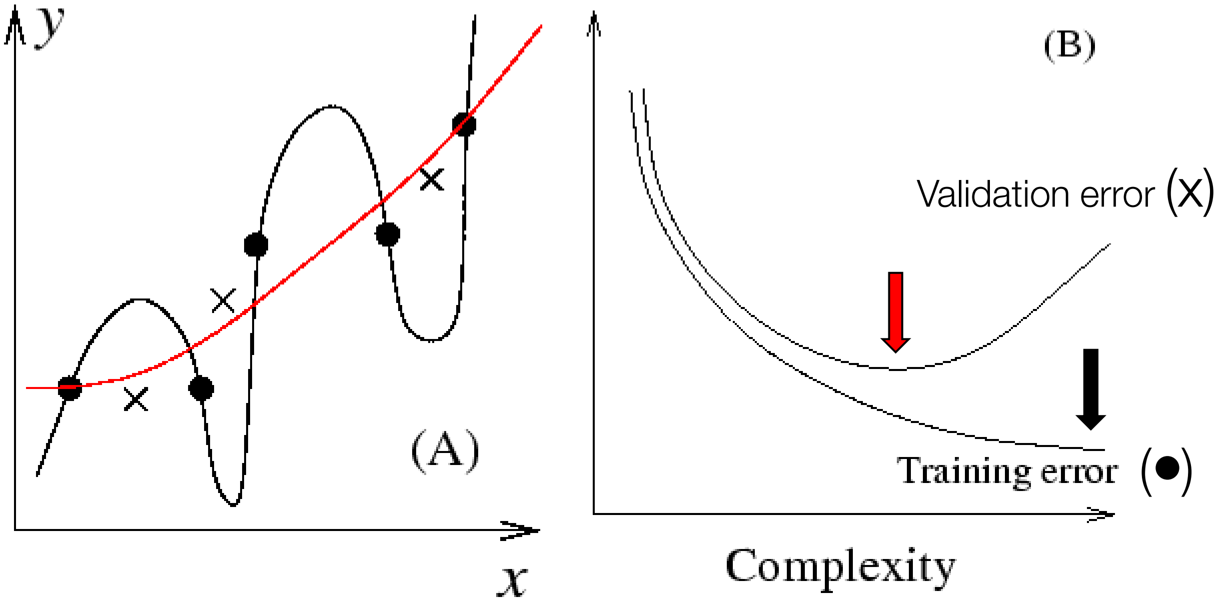
« Let well alone »
« Le mieux est l'ennemi du bien »

unrestricted

Modeling:

“Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.”

- control of the overfitting
  - split the data set into train, validation and test sets
  - train models of increasing complexity on the training set and control the performance on the validation set
  - keep the complexity level that performs best on the validation set



Modeling:

"Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed." (Oracle Definition)"

Train dataset

The data used to construct or discover a predictive relationship are called the training data set.

Validation dataset

setting model parameters for approaches that search through training data for empirical relationships and which tend to overfit the data (meaning that they can identify apparent relationships in the training data that do not hold in general).

Test dataset

A test set is a set of data that is independent of the training data, but that follows the same probability distribution as the training data.

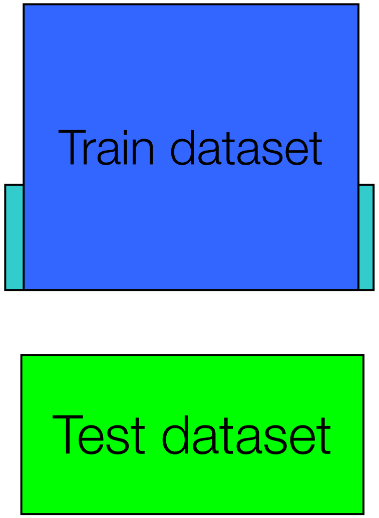
see http://en.wikipedia.org/wiki/Test_set

If a model fit to the training set also fits the test set well, minimal over fitting has taken place. If the model fits the training set much better than it fits the test set, over fitting is likely the cause.

Modeling:

“Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.” (Oracle Definition)”

Khiops uses a powerful regularization technique



you can use all your data !

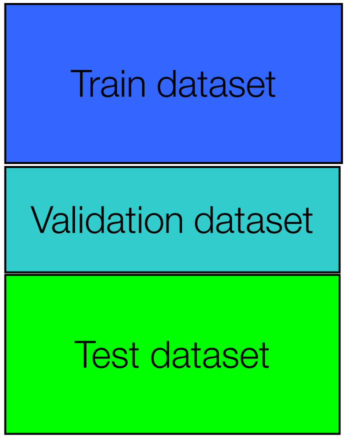
.

you still need a test set to double-check …

Modeling:

"Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed."

- What should be done?
  - Identify your modeling problem
  - Select a modeling technique
  - Specify an evaluation protocol
    - performance evaluation criterion
    - split the data in train / validation / test
  - Elaborate the model
    - optimize the criterion on the training set
    - control the performance on the validation set
  - Evaluate the model
    - compute the performance o
  - Produce a modeling report
    - with variable influence on the
  - Interpret the results
  - Look for possible improvements (
  - Implement these improvements a


Train dataset
Validation dataset
Test dataset

sounds obvious ?

but it's not !

- seems this could go on forever …
- … but data-mining is <u>always</u> a time-constrained task

Evaluation

"Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. Determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results is reached." (Oracle Definition)

Evaluation

“Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. Determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results is reached.” (Oracle Definition)

- Evaluation of the impact of the modeling
  - Evaluation of the objectives of the project
  - Collection of all the elements required to deploy the model

- Implementation
  - Use of the model on ‘new’ data to have a ‘new’ temporal behavior assessment (stationarity?)
  - Meeting ‘GoNoGo’ before the deployment

- Results
  - Evaluation Report
  - Decision to deploy
    - or not
      - new process to try to improve the results ?
      - or stop

Deployment

“Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.”

Deployment

“Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.”

- Objectives of this step:
  - Deployment type: deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.”
  - …
- Implementation
  - Meeting between domain expert, members of the information system and data mining expert…
- Results:
  - Deployment Specifications
  - First Implementation with the assistance of the data mining
- Report of the end of the project … for the data miner…

Deployment

"Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process."

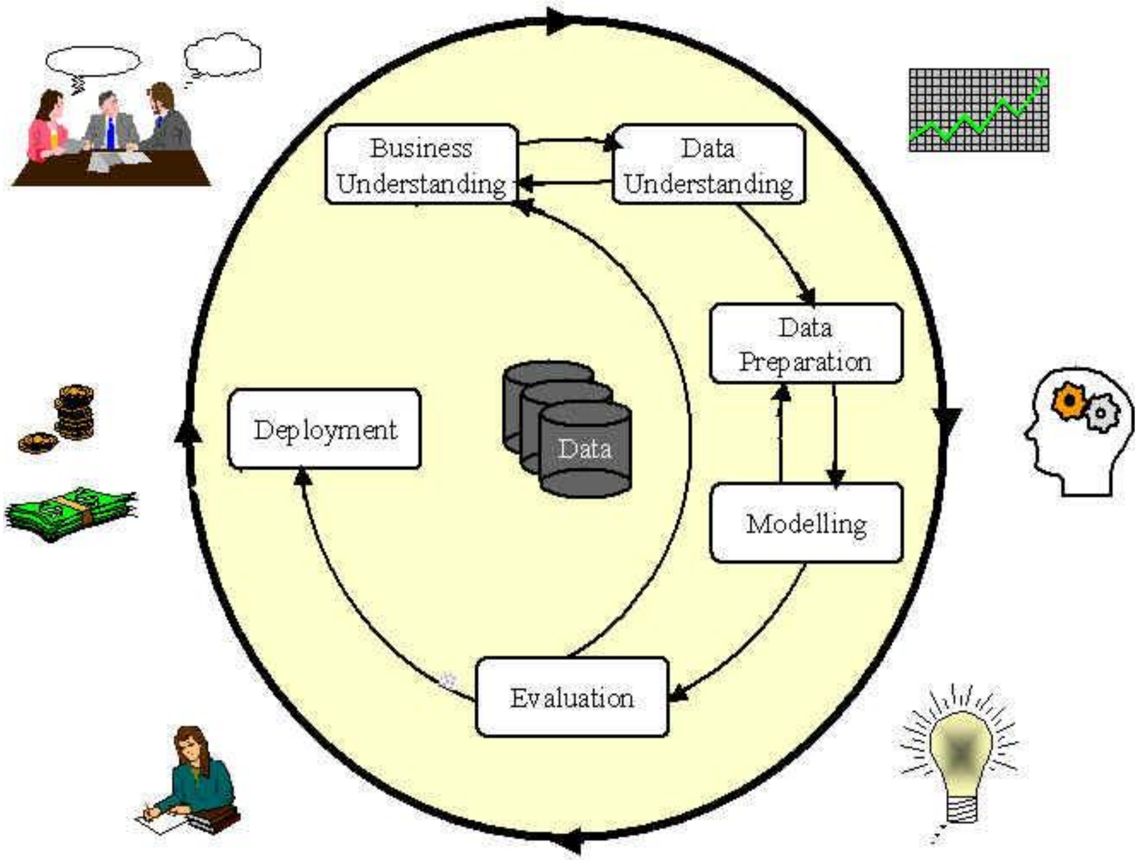
- Objectives of this step:
  - Deployment type: deployment can be as simple as generating a report or as complex as implementing a
  - …
- Implementation
  - Meeting between domain expert, men mining expert…
- Results:
  - Deployment Specifications
  - First Implementation with the ass
- Report of the end of the project

sounds obvious ?

but it's not !

- it would be a very poor practice to discover at this stage that deployment brings technical constraints (cpu or memory usage limitation, …) which cannot be met by the chosen modeling technique
- deployment technical constraints must be known at the beginning of the project !

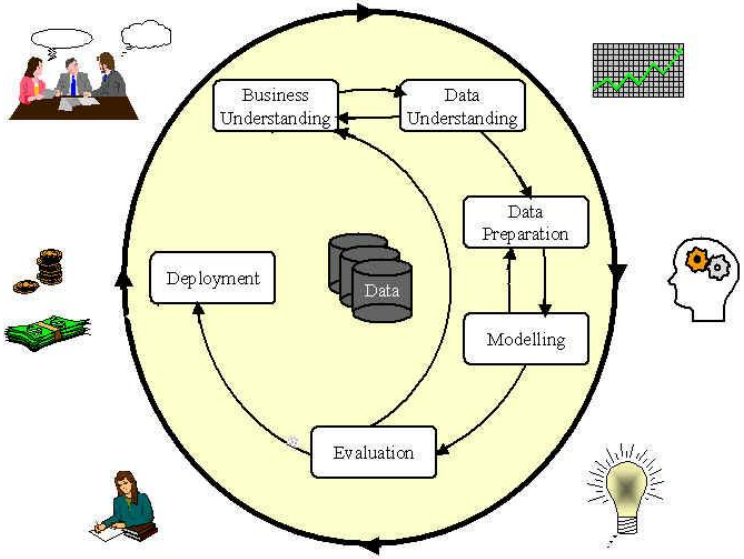
Step by Step – again ?



CRISP-DM++ Methodology

- CRISP-DM: Cross-Industry Standard Process for Data Mining

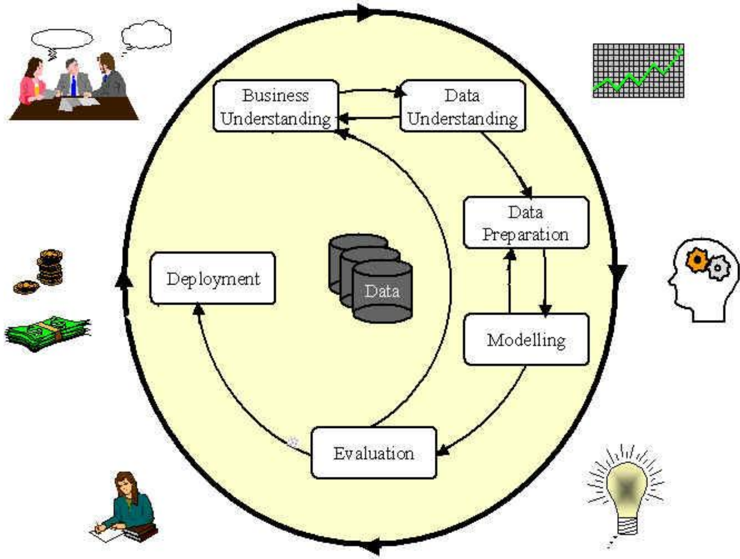
1. Collect of the needs
2. Sources Analysis
3. Explanatory Analysis of the data
4. Objective Formalization
5. Database constitution
6. Data Preparation
7. Modeling
8. Evaluation, Interpretation
9. Test
10. Deployment
11. Training



Khiops can help: 3
Khiops is dedicated to: 6, 7, 8, 9, 10
Add-On Khiops interpretation is dedicated to: 8

CRISP-DM$_{++}$ a little note on « agility »

- the main "fail" point is at the evaluation step
- if you want to "fail fast" as a data-miner, you need to
  - have productive tool for preparation, modeling, evaluation and interpretation
    - endlessly tweaking model parameters is not a data-miner's job !
  - be confident enough in your tools, results and insights to ask for more data (and therefore more work from others !) to improve the performance



CRISP-DM$_{++}$ Methodology

- CRISP-DM: Cross-Industry Standard Process for Data Mining

1. Collect of the needs
2. Sources Analysis
3. Explanatory Analysis of the data
4. Objective Formalization
5. Database constitution
6. Data Preparation
7. Modeling
8. Evaluation, Interpretation
9. Test
10. Deployment
11. Training

WANTED !

Business people
Information system
Data Miner

what ? not a word about “big data” ?

a word about "big data" !

- Data-mining is often « secondary data analysis »
  - analysis on data which have not been specifically collected for the task but have been gathered for other tasks, or with no specific task in mind
    - the data is « there », why not use it ?
    - sure … but who can tell us about the collection bias ?
  - you never have « all » the data, you only have the data that has been chosen, collected
    - however large, the data does not « speak for itself » …

- In this respect data-mining is different from classical statistics where gathering the data « from the field » is part of the job
  - careful experimental planning and design allows to reduce the size of the data collected while maintaining the power of the analysis

a word about “big data” !

- Data-mining is often « secondary data analysis »
- In this respect data-mining is different from classical statistics where gathering the data « from the field » is part of the job

- Big data infrastructures are changing the picture as they allow to gather all data « from the field » (in principle)
  - the good news is that data-miners can rely on a lot of statistical techniques (tests, experimental designs …) to build better data sets
  - the bad news is that they have to become familiar with these techniques

thank you

En moulinant n’importe quel volume de données, on trouve toujours une combinaison qui répond à une question que l’on ne se posait pas. Cela revient un peu à tirer une flèche dans un mur avec un arc, puis d’aller peindre une cible autour de la flèche.

(http://www.francetvinfo.fr/sante/maladie/maladie-de-lyme-la-fable-de-la-pommade-antibiotique-miracle_1979767.html)