ISLab

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Dados e Aprendizagem Automática

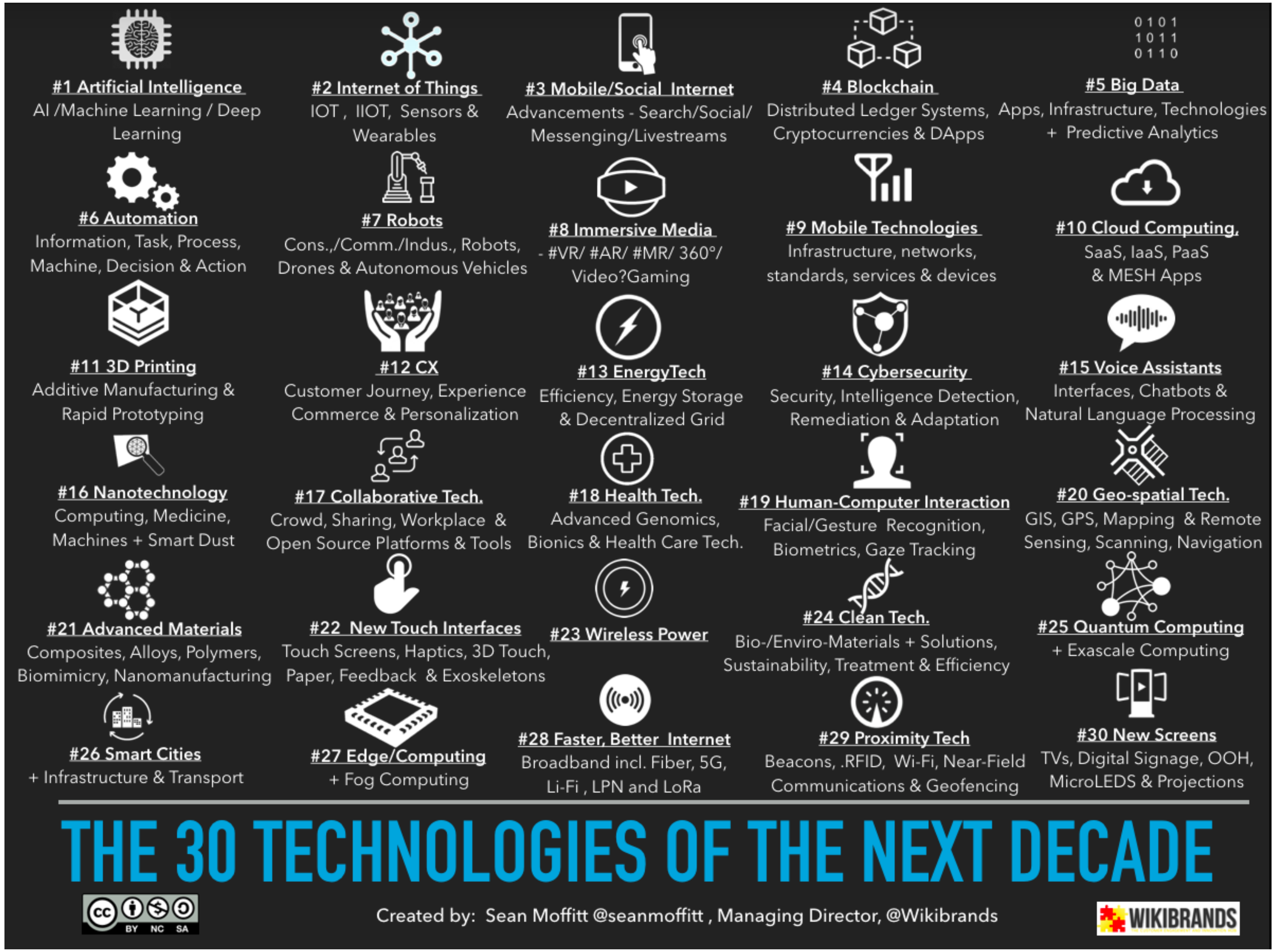
Data Science Pipeline

DAA @ MEI/1º ano – 1º Semestre
DAA @ MiEI/4º ano – 1º Semestre
Paulo Novais

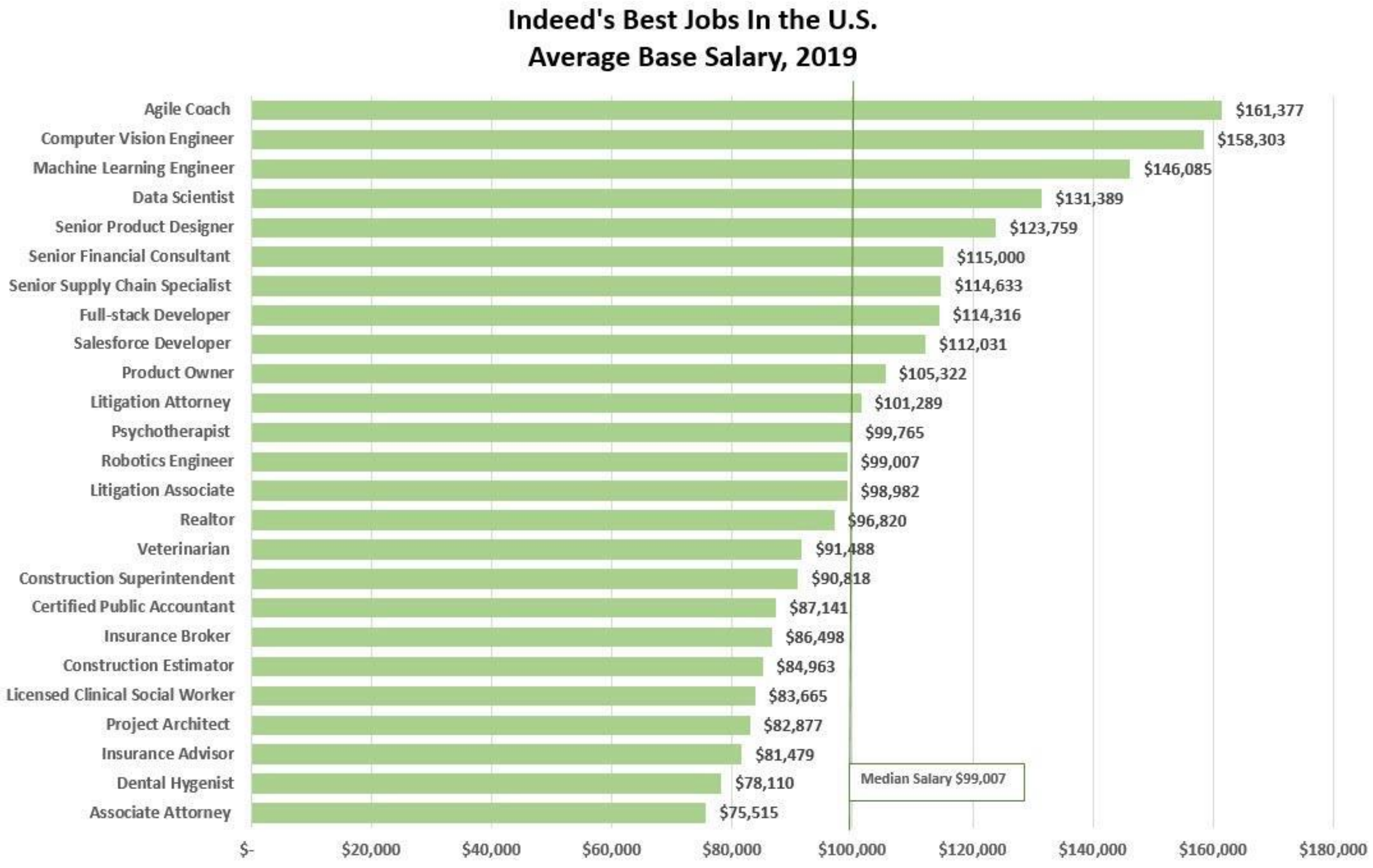
Contents

- Machine Learning vs Data Science;
- Terminology of AI
- Methodologies
    - CRISP-DM
    - SEMMA
    - PMML
- A ML Pipeline

Technologies of the next decade

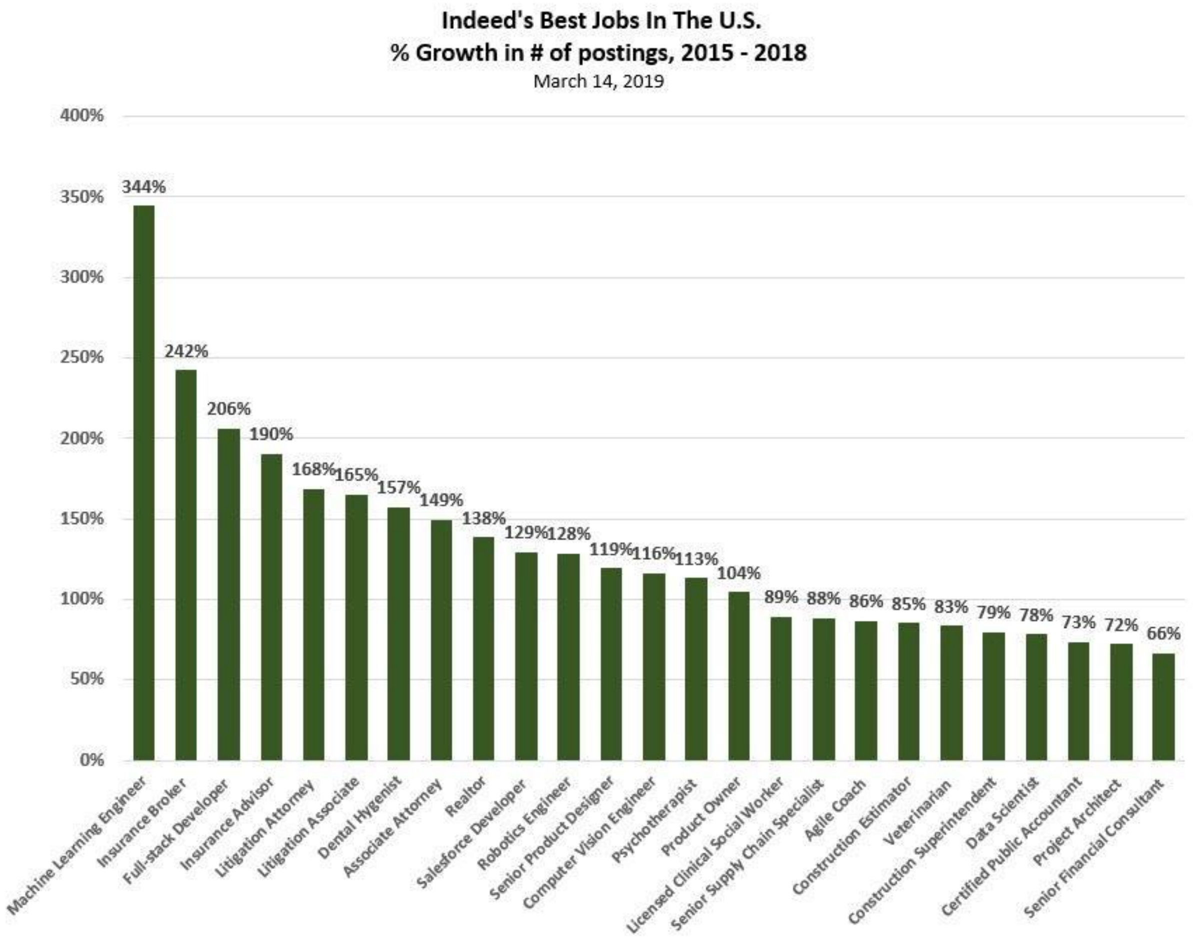


Motivation



(https://www.forbes.com/sites/louiscolumbus/2019/03/17/machine-learning-engineer-is-the-best-job-in-the-u-s-according-to-indeed/#2d134a177bb0)

Motivation



(https://www.forbes.com/sites/louiscolumbus/2019/03/17/machine-learning-engineer-is-the-best-job-in-the-u-s-according-to-indeed/#2d134a177bb0)

Terminology of AI

- Artificial Intelligence
- Machine Learning
- Deep Learning
- Data Science
- …

Terminology of AI

Machine Learning vs Data Science

Machine Learning

- A -> B system
- pt-pt = Aprendizagem Automática (?)
- *“Field of study that gives computer the ability to learn without being explicitly programmed.”*

Arthur Samuel

Usually results in a software artefact

Data Science

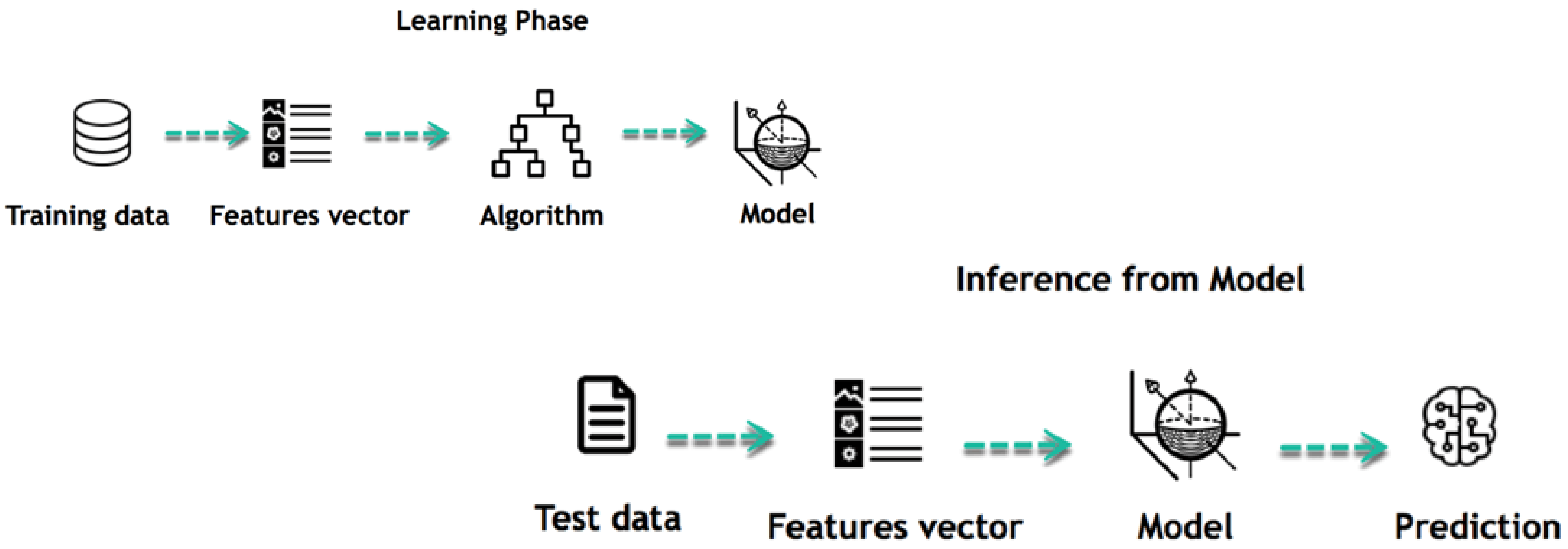
- Analyse sets of data (datasets)
- pt-pt = Ciência dos Dados (?)
- Science of extracting knowledge and insights directly from data

Usually results in slides and reports

There is no universal adhrence!!!

Terminology of AI

How does it work?



Source: Machine Learning Tutorial for Beginners
https://www.guru99.com/machine-learning-tutorial.html

Learning

- How does it happen? A simple example …

Important Concepts

- **Training set**
  - Dataset describing a particular problem
- **Test Set**
  - Data set against which the model will be tested
- **Input Variables**
  - Set of variables that characterize each instance of the problem
- **Output Variables**
  - Set of variables that answer the problem

Example

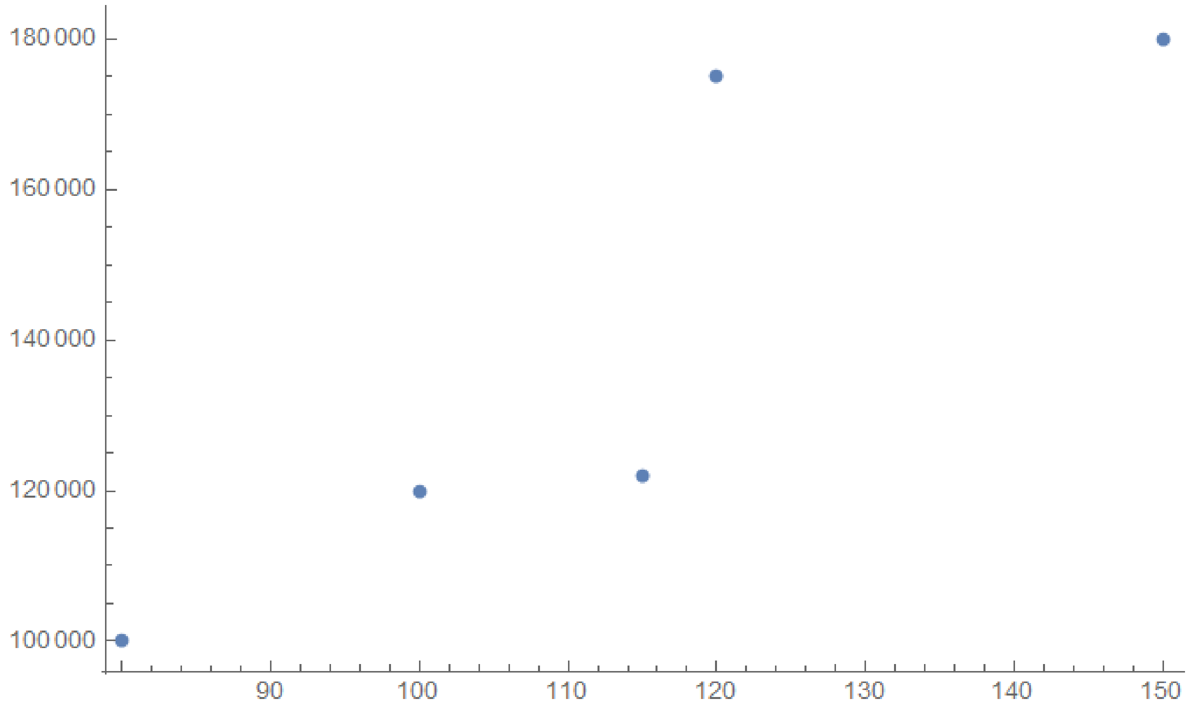
House prices by area

Area	Price
120	175000
150	180000
80	100000
100	120000
115	122000

$m = 5$

$x^{(1)} = 120$

$y^{(3)} = 100000$



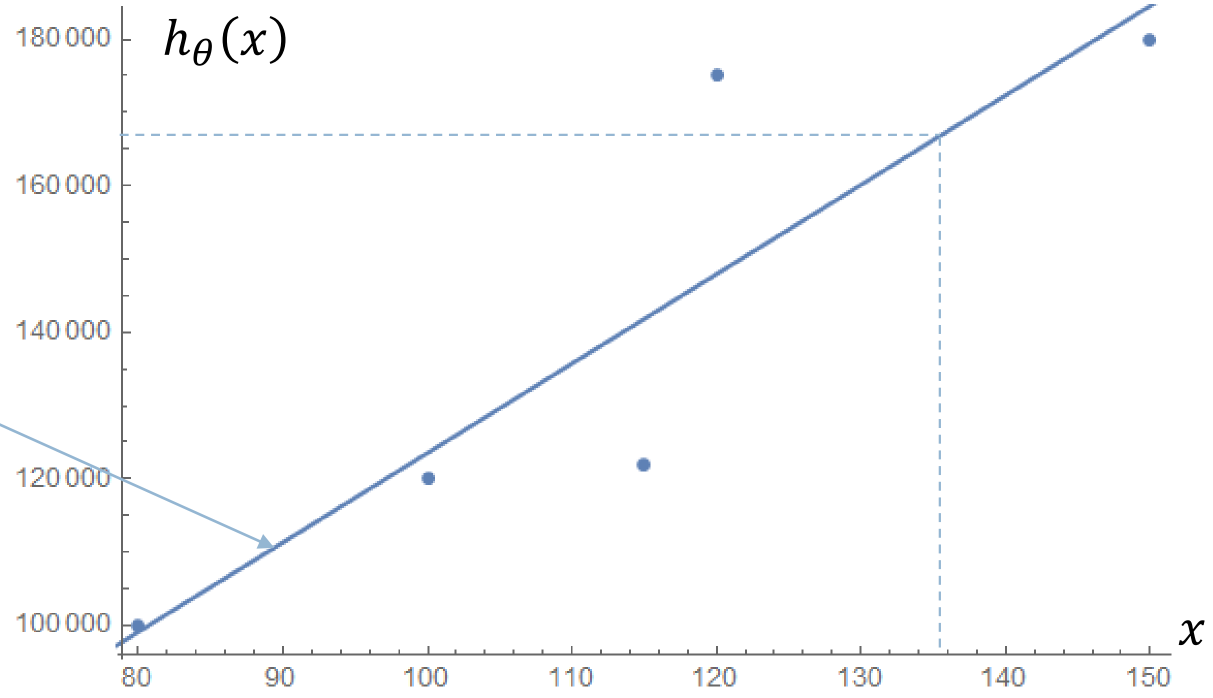
Example

- House prices by area

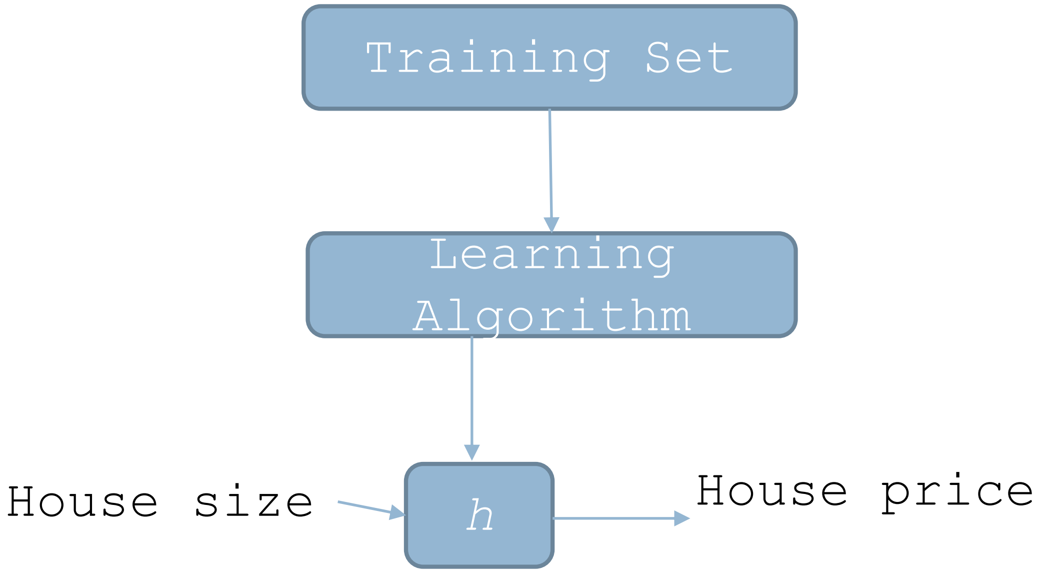
Area	Price
120	175000
150	180000
80	100000
100	120000
115	122000

Function that models the problem h

$$h_\theta(x) = \theta_0 + \theta_1 x$$



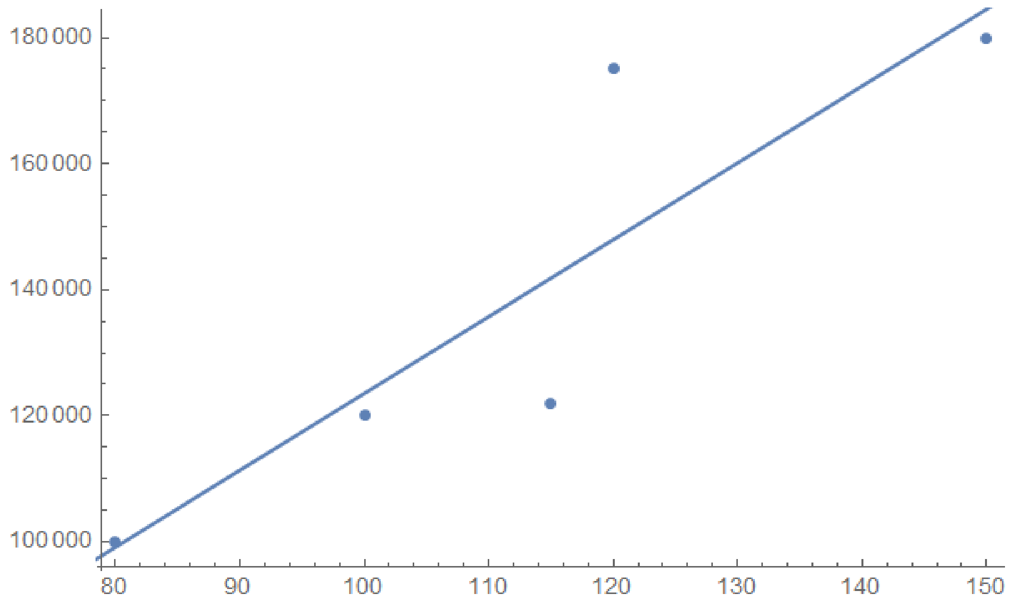
Operation



Models

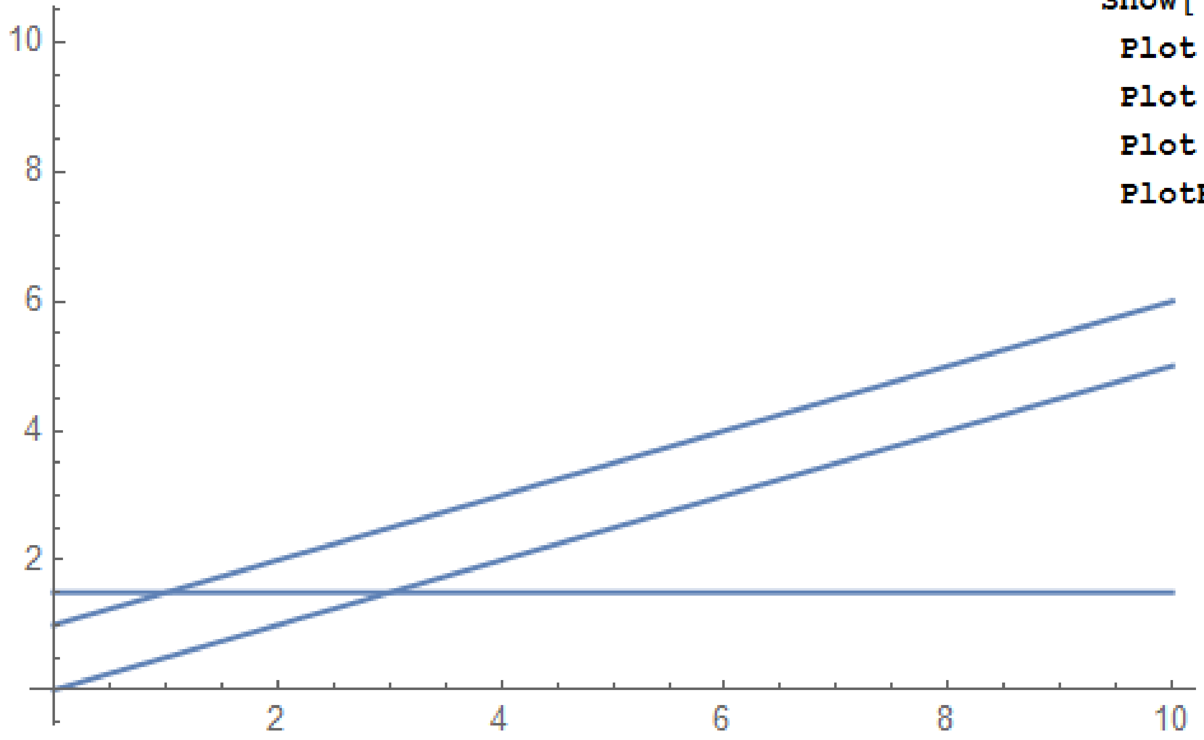
In this example, a model was built based on a linear regression with a variable

There are many different models, with different degrees of complexity



Models

For the same problem we can create different models. How to choose the best?



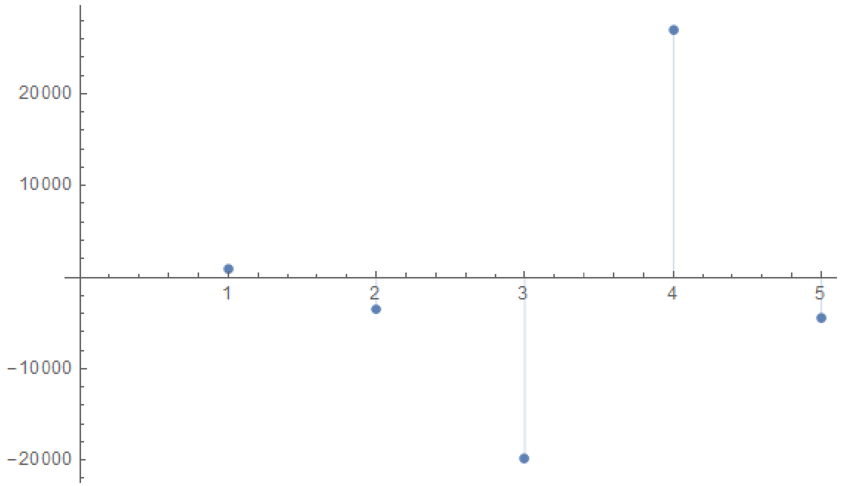

```
Show[
 Plot[1.5 + 0 * x, {x, 0, 10}],
 Plot[0 + 0.5 * x, {x, 0, 10}],
 Plot[1 + 0.5 * x, {x, 0, 10}],
 PlotRange → {{0, 10}, {0, 10}}, AxesOrigin → {0, 0}]
```

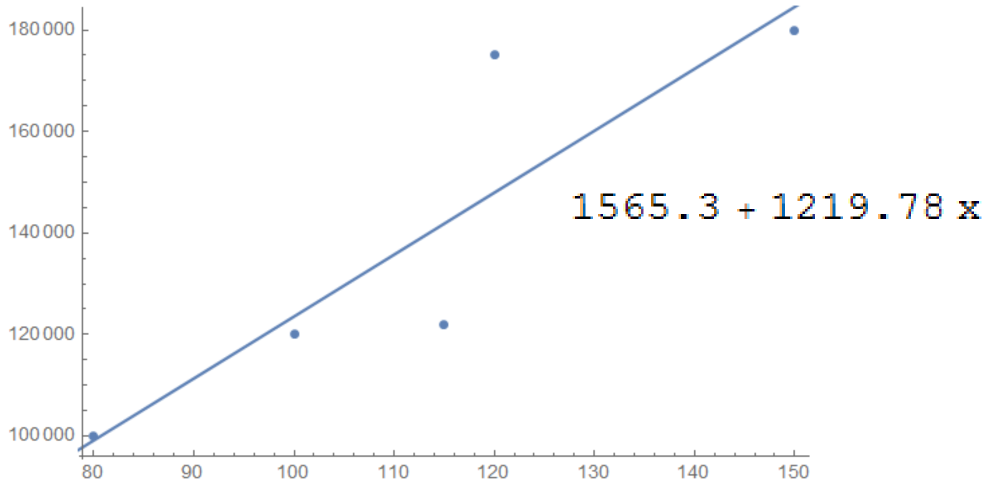
Key idea

Choose values of θ_0 and θ_1 that minimize the distance between h_θ (x) and y for each pair of values (x, y) in the training model

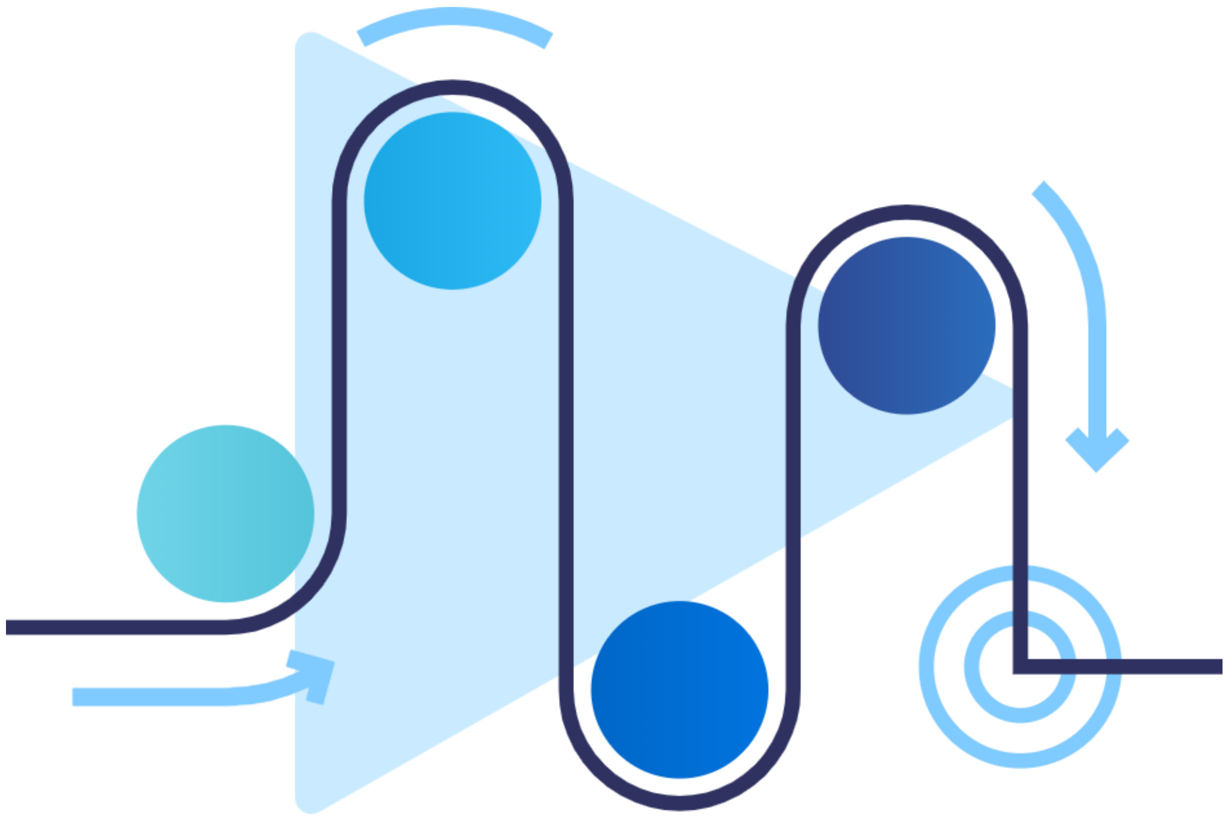
$$minimize \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

```
Total[Abs /@ lm["FitResiduals"]]
55828.4
```





Methodologies



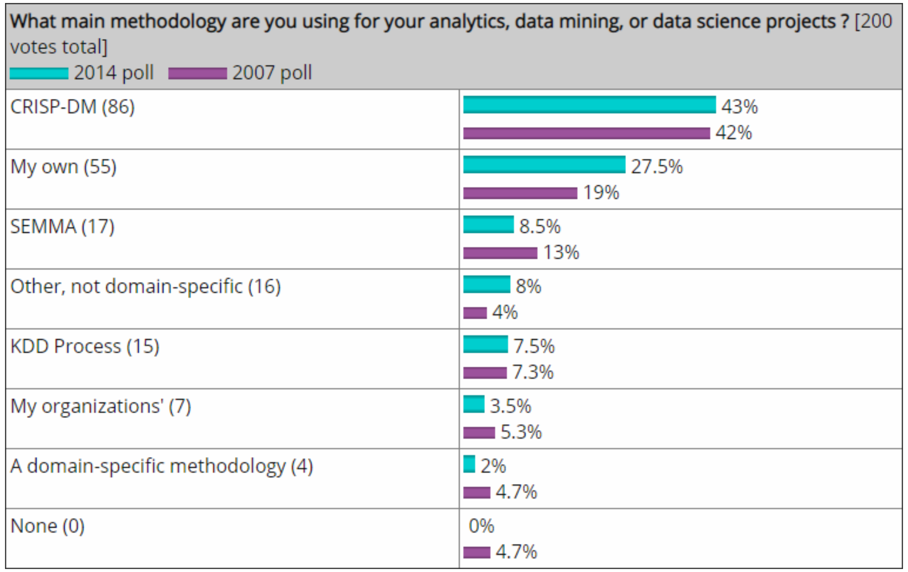
Methodologies for Knowledge Extraction

- A Methodology for Knowledge Extraction (Data Mining) describes and creates a set of steps that the development of a Knowledge Extraction Project must go through to solve problems.
- Framing a KE/DM process under a methodology:
  - Ensures greater robustness;
  - Facilitates its understanding, implementation, and development;
  - Allows process replication;
  - Assists in project planning and management;
  - It gives "maturity" to the KE/DM process;
  - Encourage adoption of best practices.

Methodologies

Why standard methodologies?

- Allows projects to be replicated
- Aid project planning and management
- Encourage best practices and help to obtain better results



(https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html)

Methodologies for Knowledge Extraction

- CRISP-DM
  - **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
    (Daimler Chrysler, SPSS, NCR)

- SEMMA
  - **S**ample, **E**xplore, **M**odify, **M**odel and **A**ssess
    (SAS Institute Inc.)

- PMML
  - **P**redictive **M**odel **M**arkup **L**anguage
    (Angoss Software, Magnify, Univ. Illinois, NCR, SPSS

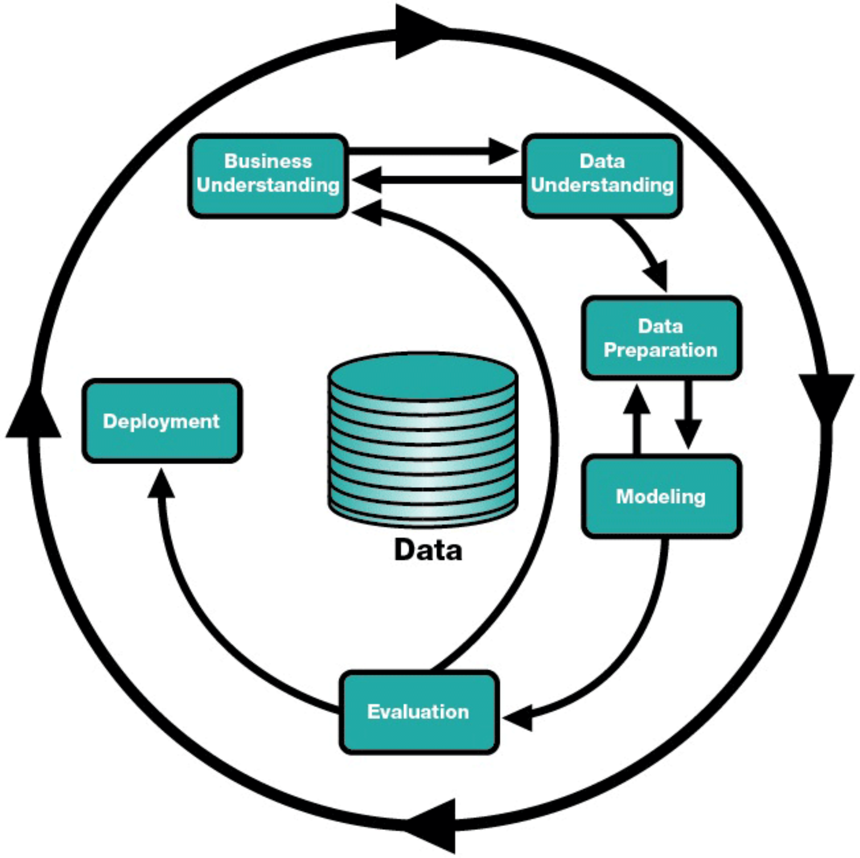
CRoss Industry Standard Process for Data Mining

- **CRISP-DM**
  (Daimler Chrysler, SPSS, NCR)

- Objectives:
  - Define a KE process for the industry;
  - Build and provide support tools;
  - Ensure the quality of CE projects;
  - Reduce specific CE knowledge needed to conduct a CE process.

CRISP-DM

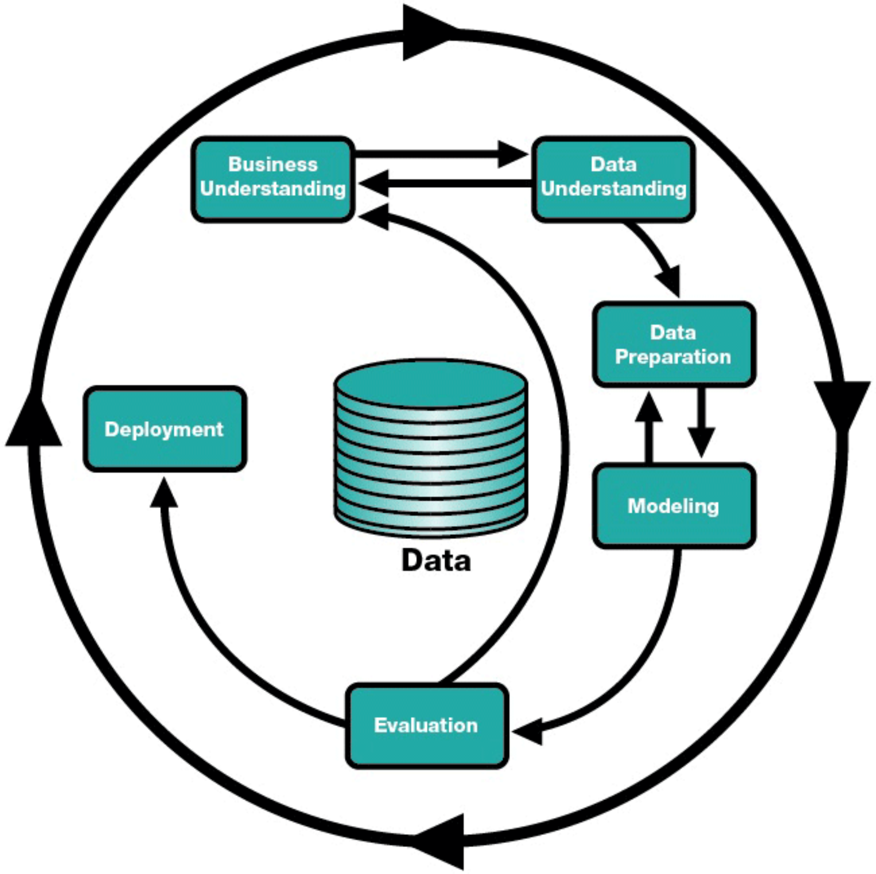
- Process model with a view to defining a "script" for the development of KE projects, which takes place in 6 stages:
  - Business Understanding;
  - Data Understanding;
  - Data preparation;
  - Modeling;
  - Evaluation;
  - Deployment.



(http://crisp-dm.eu/reference-model/)

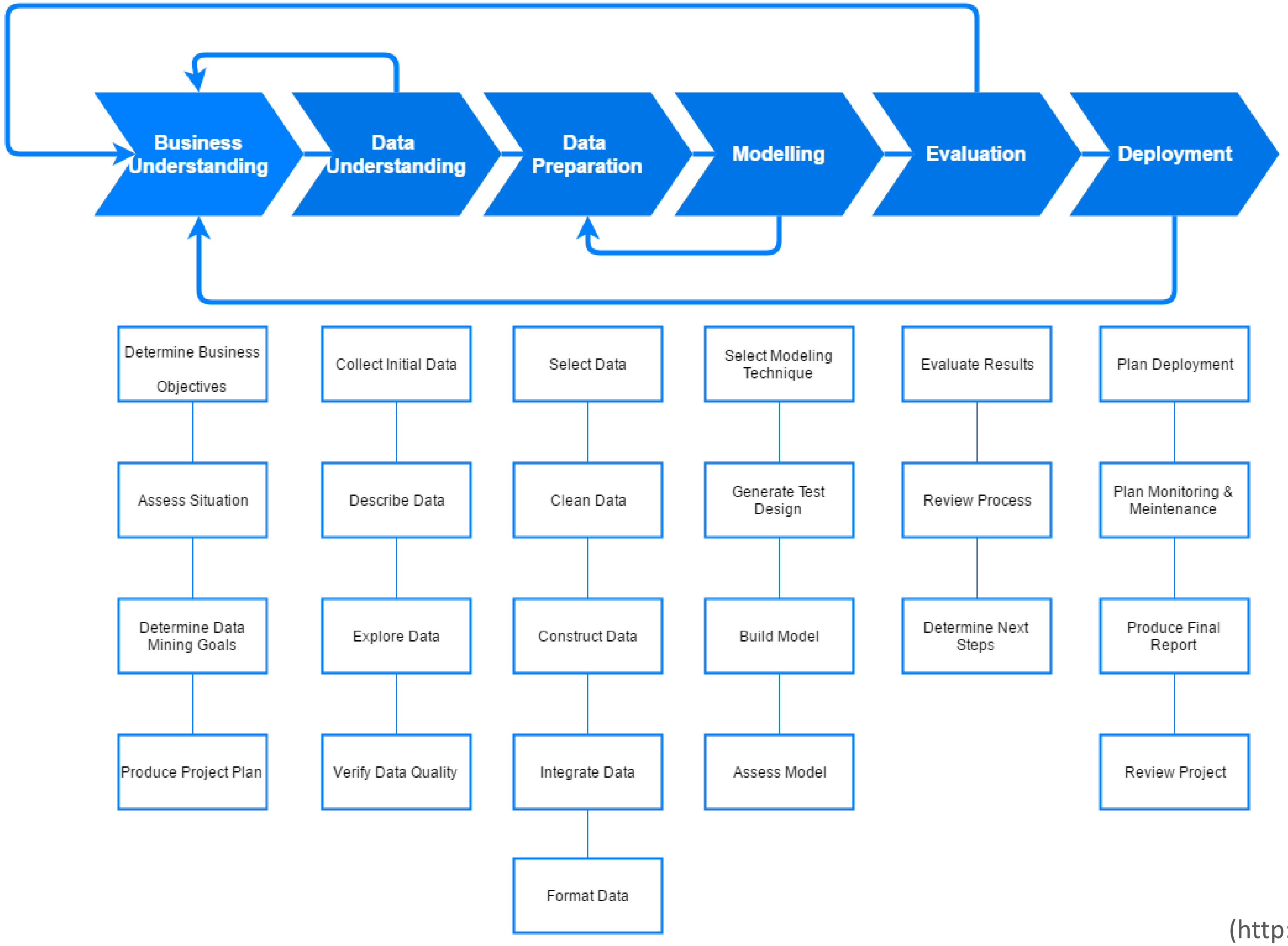
CRISP-DM

- Business Understanding
  - Understanding the project objectives and defining the KE problem;
- Data Understanding
  - Obtain data and identify data quality;
- Data Preparation
  - Selection of attributes and data cleaning;
- Modeling
  - Experimentation with KE tools;
- Evaluation
  - Comparison of results with business objectives;
- Deployment
  - Putting the model into production.



(http://crisp-dm.eu/reference-model/)

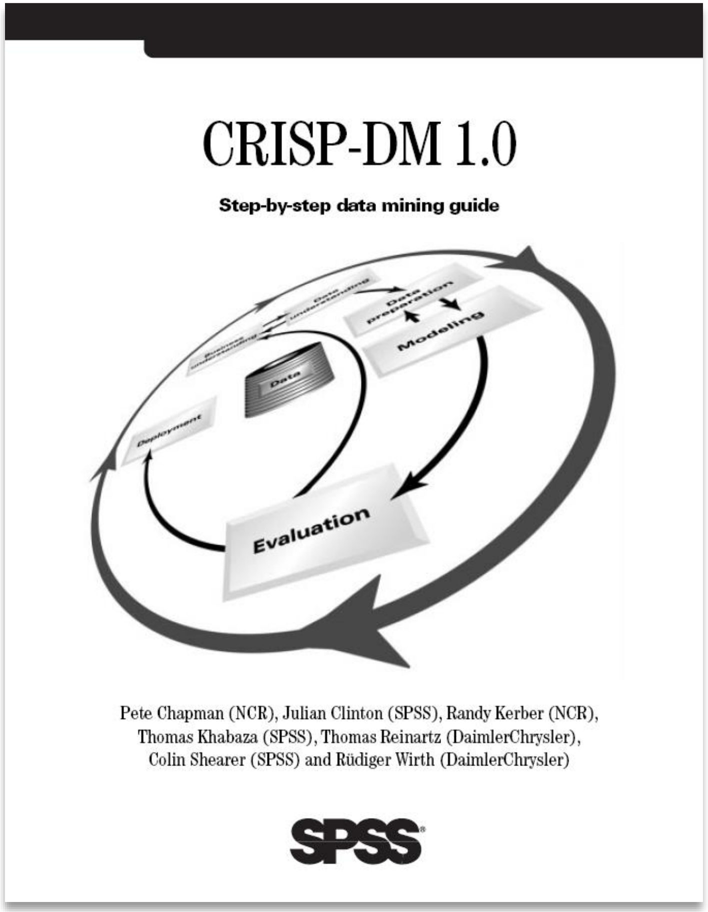
CRISP-DM



(http://crisp-dm.eu/business-understanding/)

CRISP-DM

- “CRISP-DM 1.0: Step-by-step data mining guide”,
  Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth



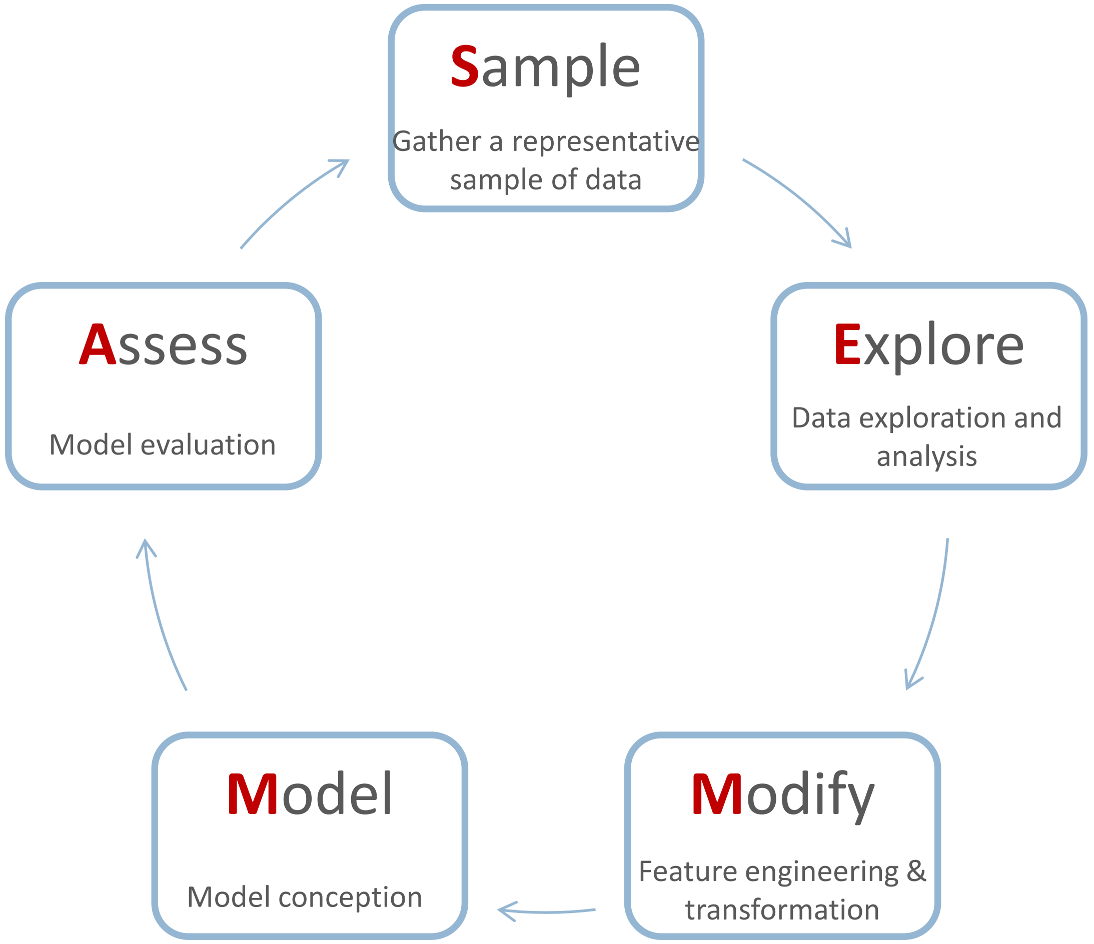
Sample, Explore, Modify, Model and Assess;

- SEMMA
- Data Mining product developed by SAS Institute Inc.;
- SAS definition:
  - "Data Mining is the process of extracting knowledge and complex relationships from large volumes of data."
- Motivation:
  - need to define, standardize and integrate Data Mining systems or processes in production cycles.

SEMMA

Divide the Data Mining process into 5 steps:

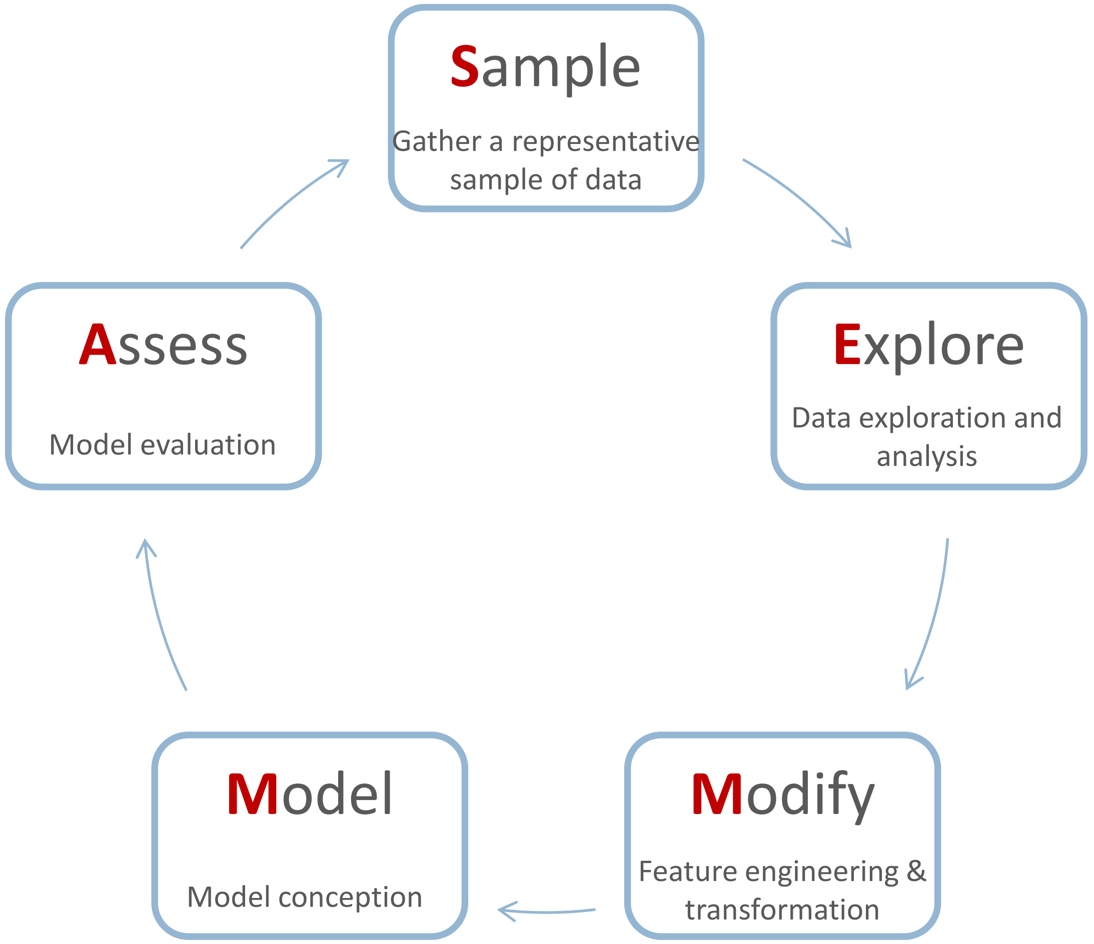
- Sample:
  - Data extraction from the problem universe;
  - It bases the Data Mining process on the concept of “sample” of the problem;
  - Small and significant sample;
  - Provides flexibility and speed in the data processing.
- Explore;
- Modify;
- Model;
- Assess.



SEMMA

Divide the Data Mining process into 5 steps:

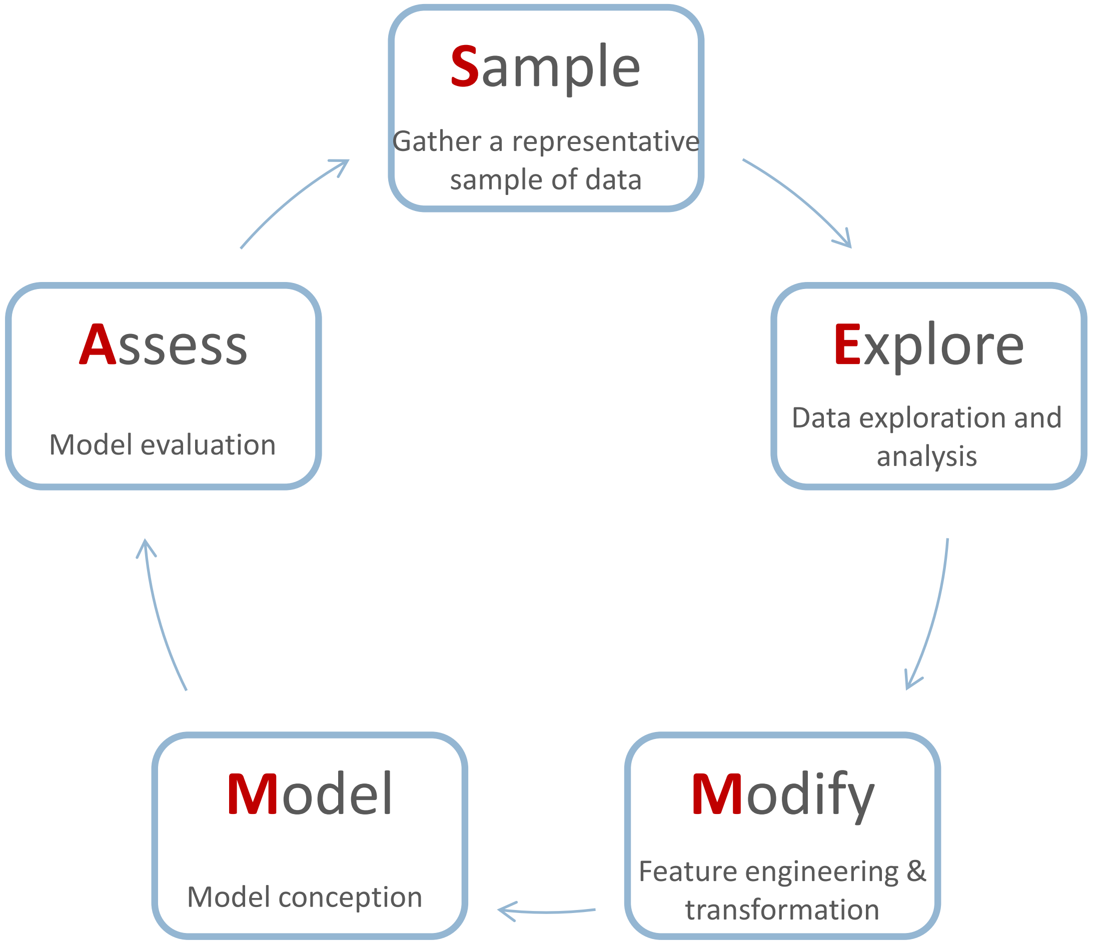
- Sample;
- Explore:
  - Visual and/or numerical exploration of trends;
  - Refinement of the discovery process (mining);
  - Statistical techniques: linear regression, least squares, Poisson distribution, etc.;
  - Search for unforeseen trends in data;
- Modify;
- Model;
- Assess.



SEMMA

Divide the Data Mining process into 5 steps:

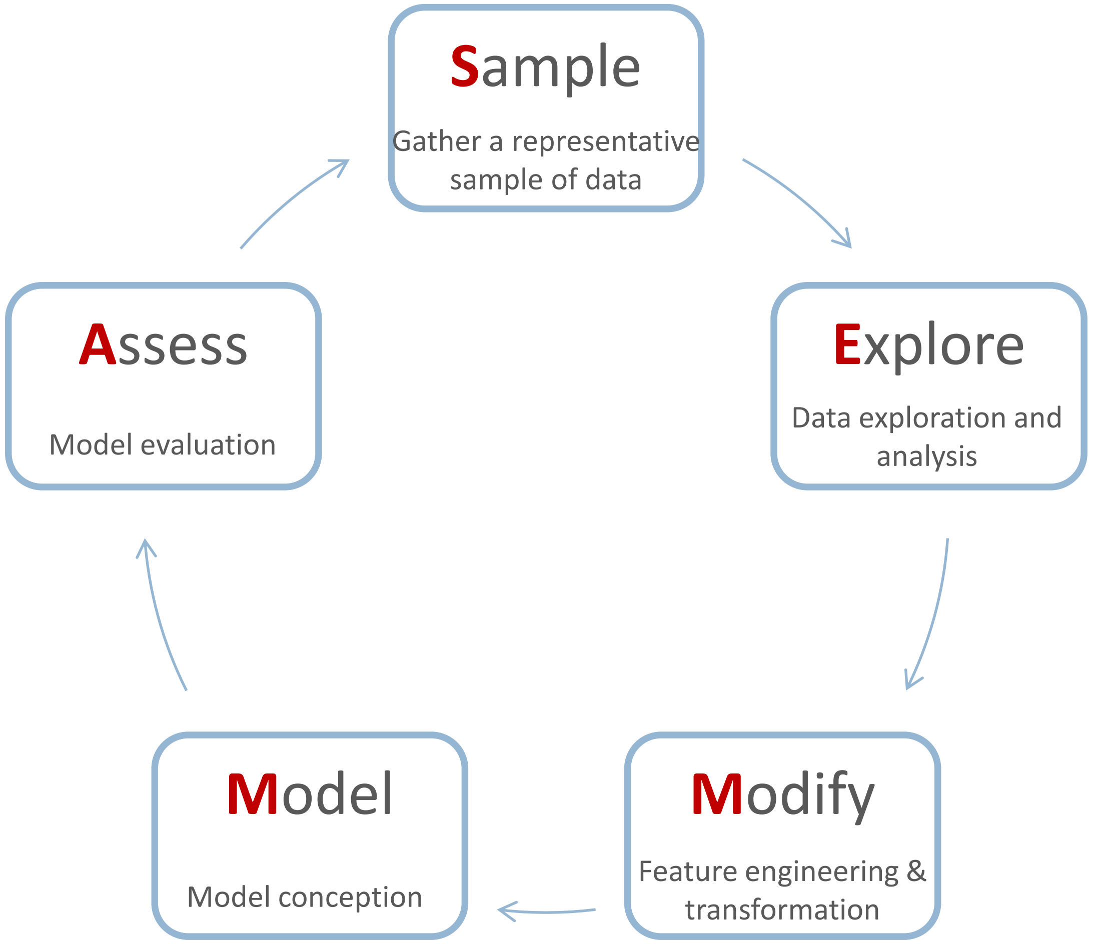
- Sample;
- Explore;
- Modify:
  - The concentration of all necessary modifications;
  - Inclusion of information;
  - Selection or introduction of new variables;
  - Objective: create, select and adapt variables for the next step;
- Model;
- Assess.



SEMMA

Divide the Data Mining process into 5 steps:

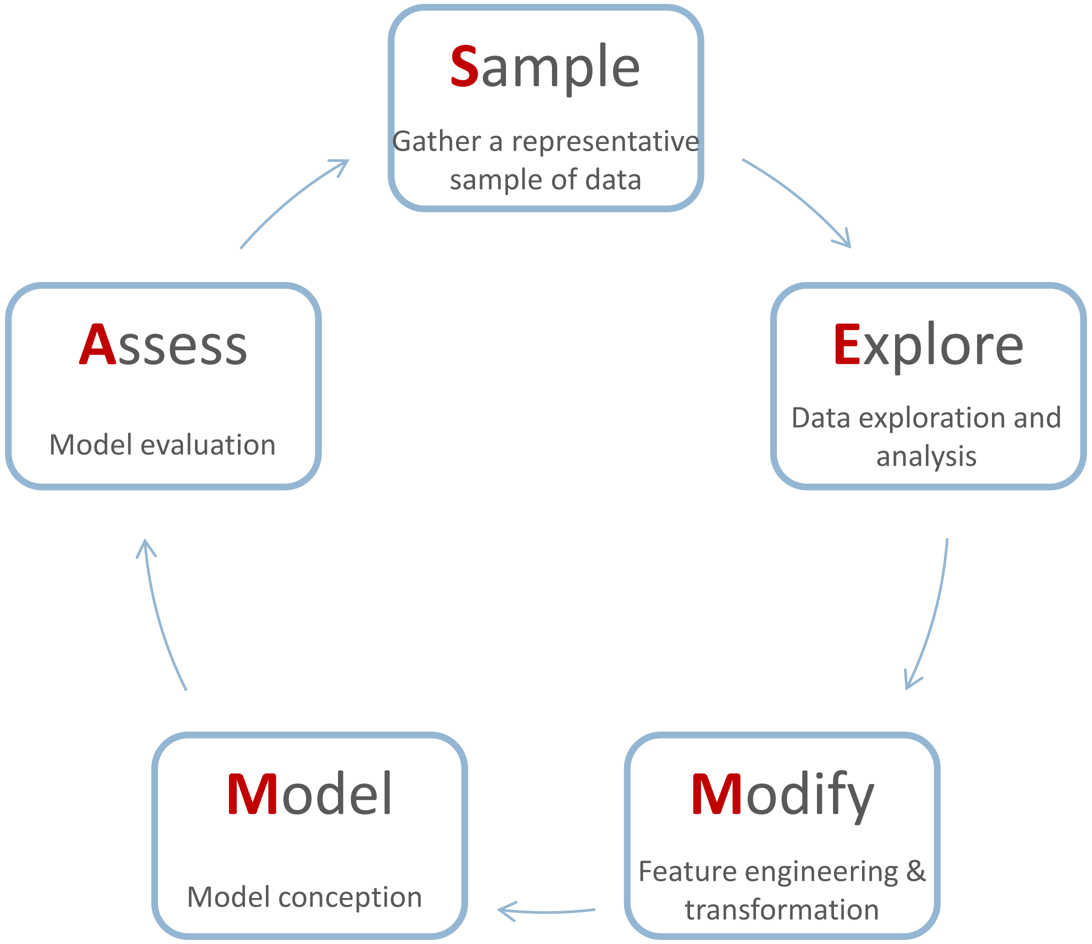
- Sample;
- Explore;
- Modify;
- Model:
  - Definition of data mining model construction techniques: artificial neural networks, decision trees, linear regression, etc.;
  - Dependent on the type of data present in each model (eg, ANN are more suitable in problems where the data has complex relationships);
- Assess.



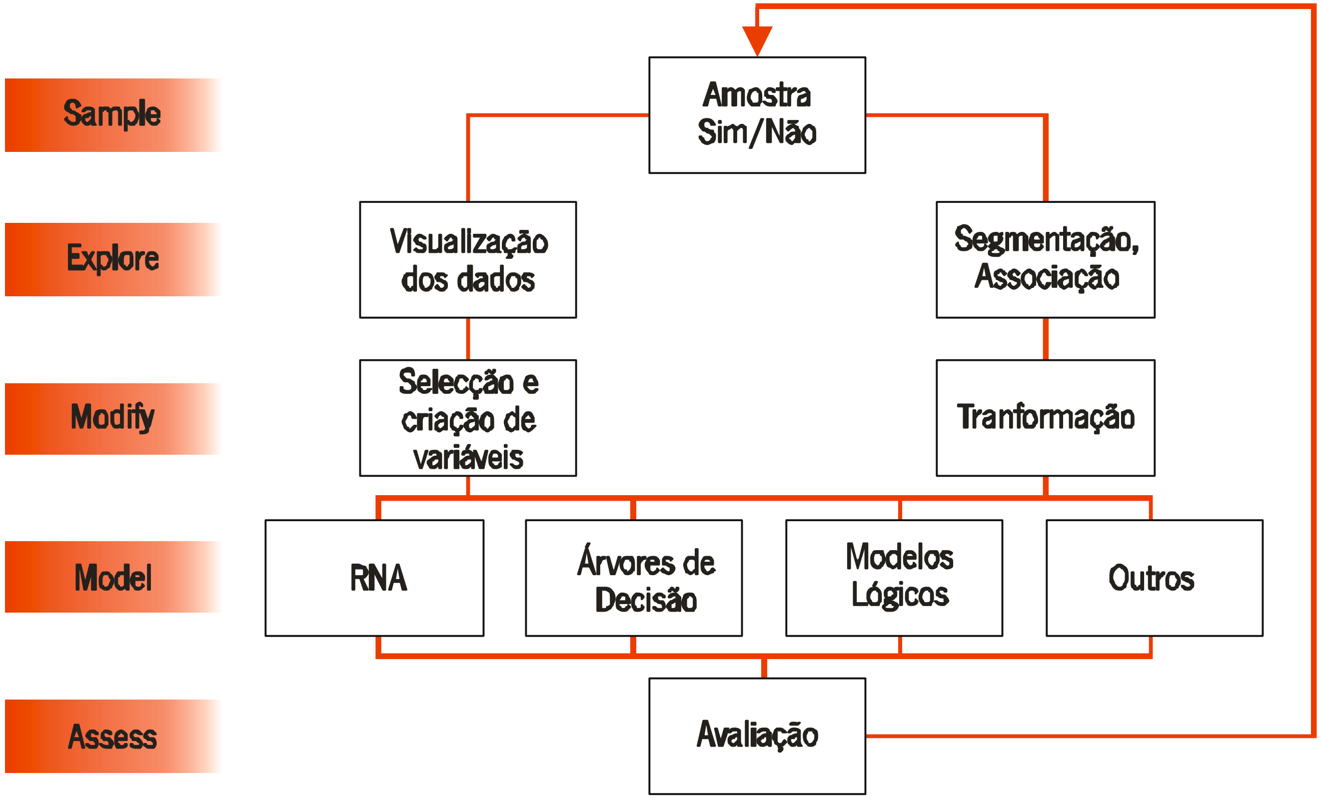
SEMMA

Divide the Data Mining process into 5 steps:

- Sample;
- Explore;
- Modify;
- Model;
- Assess:
  - Performance measurement of the model built for Data Mining;
  - Applying the model to a sample of test data;
  - Model adjustment procedure.



SEMMA - process



in “Data Mining – Descoberta de Conhecimento em Bases de Dados”
Manuel Filipe Santos, Carla Azevedo

Predictive Model Markup Language

- PMML;
- Developed by Data Mining researchers and several companies (NCR, SPSS, etc.);
- The PMML specification is in the development and consolidation phase (version 4.2.1);
- Used by several applications (IBM DB2 Data Warehouse Edition v.10.5, SAS Enterprise Miner v.5.1, v.5.3, v.7.1, v.13.1, SPSS Statistics v.21); (http://www.dmg.org/products.html)
- Expand to make it a standard for the WWW;
- PMML is a language for describing Data Mining models;
- It uses XML to represent DM models.

PMML: examples


```
sepal_length,sepal_width,petal_length,petal_width,class
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
...
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
...
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
...
```

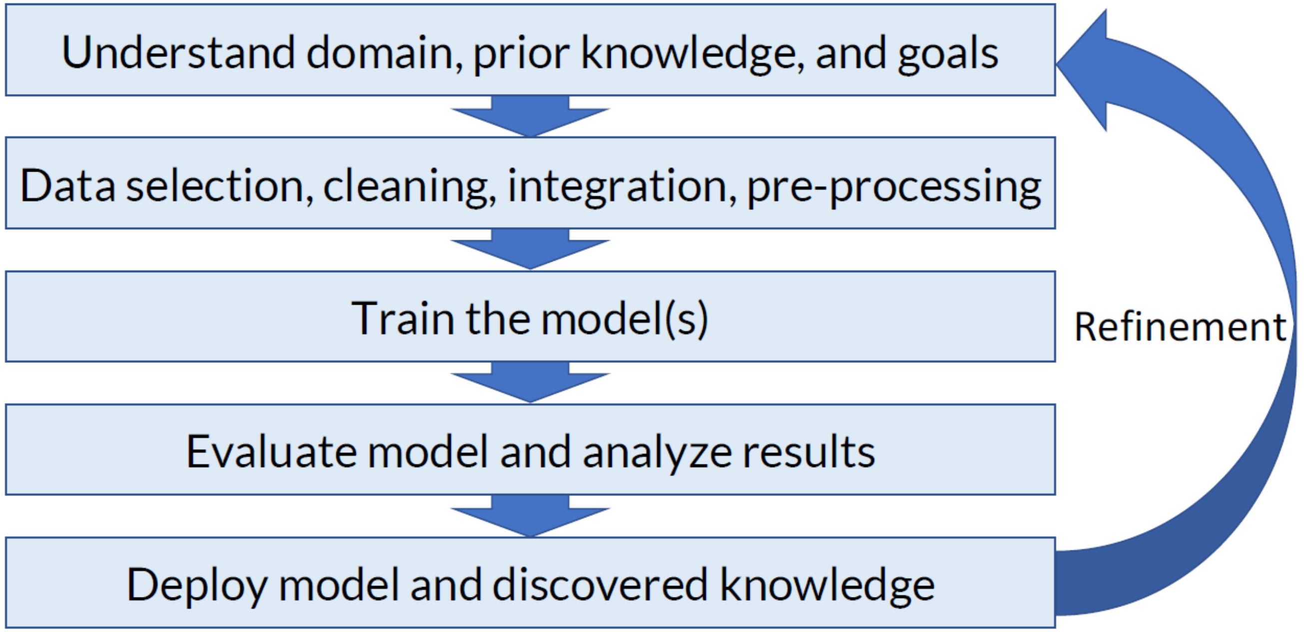

```
<PMML version="2.0">
–
<Header copyright="Copyright (c) 2001, Oracle Corporation.  All rights reserved.">
<Application name="Oracle 9i Data Mining" version="9.2.0"/>
</Header>
–
<DataDictionary numberOfFields="1">
<DataField name="item" optype="categorical"/>
</DataDictionary>
–
<TransformationDictionary>
–
<DerivedField name="PETAL_LENGTH">
–
<Discretize field="PETAL_LENGTH">
–
<DiscretizeBin binValue="1-1.59">
<Interval closure="closedOpen" leftMargin="1.0" rightMargin="1.59"/>
</DiscretizeBin>
–
<DiscretizeBin binValue="1.59-2.18">
<Interval closure="closedOpen" leftMargin="1.59" rightMargin="2.18"/>
</DiscretizeBin>
–
<DiscretizeBin binValue="2.18-2.77">
<Interval closure="closedOpen" leftMargin="2.18" rightMargin="2.77"/>
</DiscretizeBin>
```

PMML

- Allow applications to use multiple data sources without worrying about the differences between them;
- Allow the combined and/or cooperative use of Data Mining models;
- Allow the administration of DM models based on business areas.

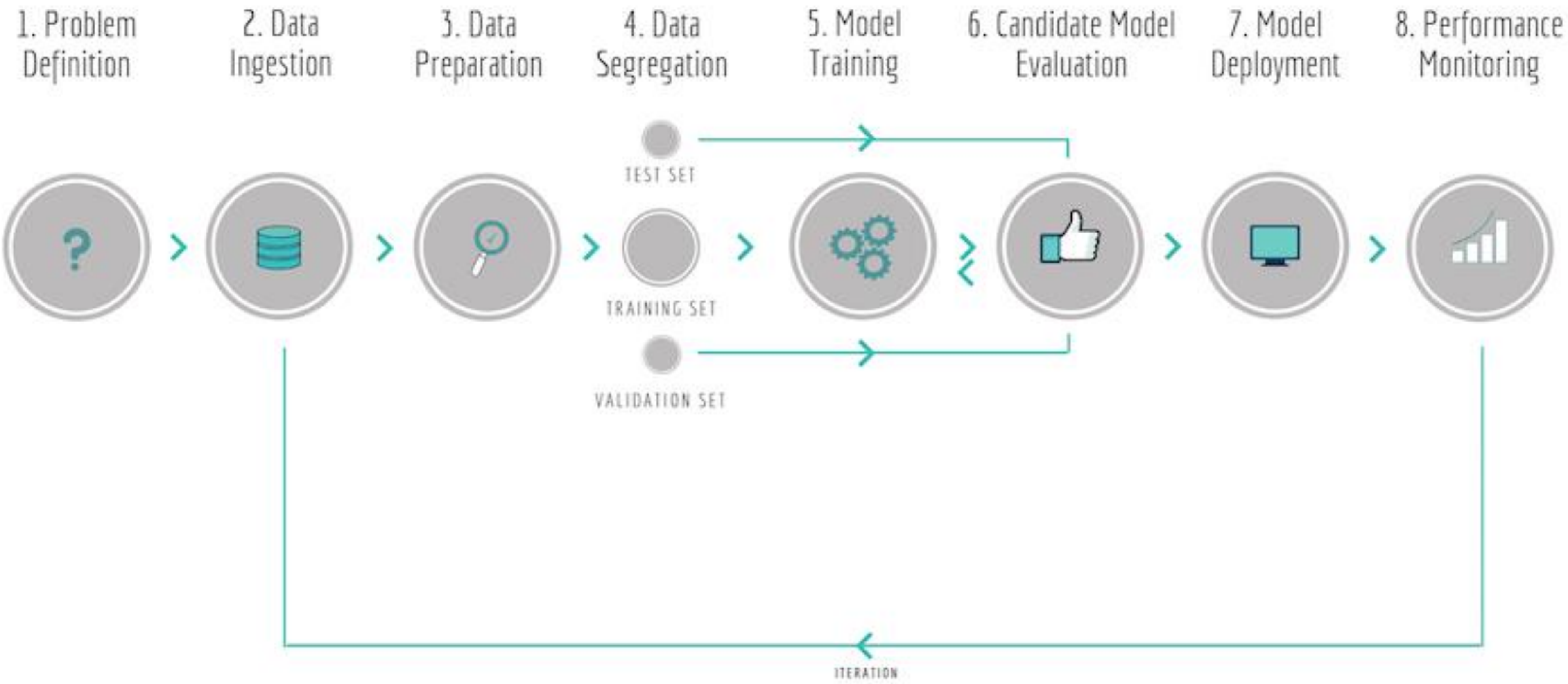
In practice

Learning is equivalent to searching (**optimization**) through the space of potential hypotheses (**representation**) to find one that best fits (**evaluation**) the training.



Source: CIS 419/519
Applied Machine Learning
Eric Eaton, University of Pennsylvania
www.seas.upenn.edu/~cis519

A Machine Learning Pipeline



(https://towardsdatascience.com/architecting-a-machine-learning-pipeline-a847f094d1c7)

References

- “CRISP-DM 1.0: Step-by-step data mining guide”, Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth.
- SAS Enterprise Miner: www.sas.com/technologies/analytics/datamining/miner/semma.html
- Data Mining Group (DMG): www.dmg.org www.dmg.org/faq.html

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Dados e Aprendizagem Automática

Data Science Pipeline

DAA @ MEI/1º ano – 1º Semestre
DAA @ MiEI/4º ano – 1º Semestre
Paulo Novais