Information Visualization

Data sense making (1)

Lesson 5

Marilena Daquino
Assistant Professor

Department of Classical Philology and Italian Studies

marilena.daquino2@unibo.it

Table of contents

01

Stages of a project

From data to wisdom

01

SO WHAT?

Get insights

Data must provide new information.

Understand

Graphs must have a take-home message.

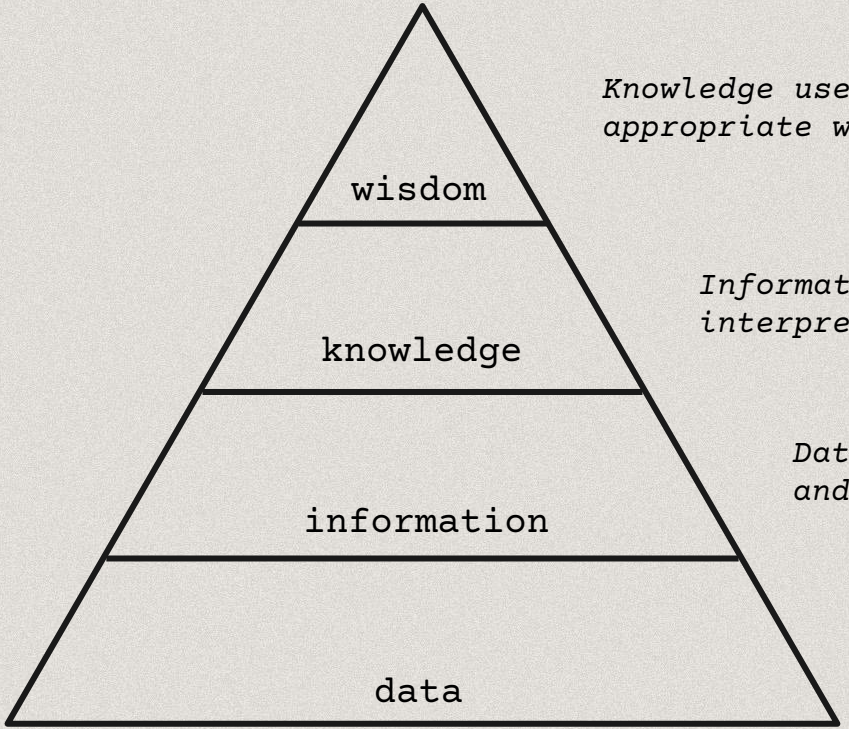
Make decisions

The interpretation of graphs must facilitate decision-making processes.

Data to wisdom pyramid

01

Stages of EDA



Knowledge used in an appropriate way

Information that has been interpreted

Data processed, visualised and contextualised

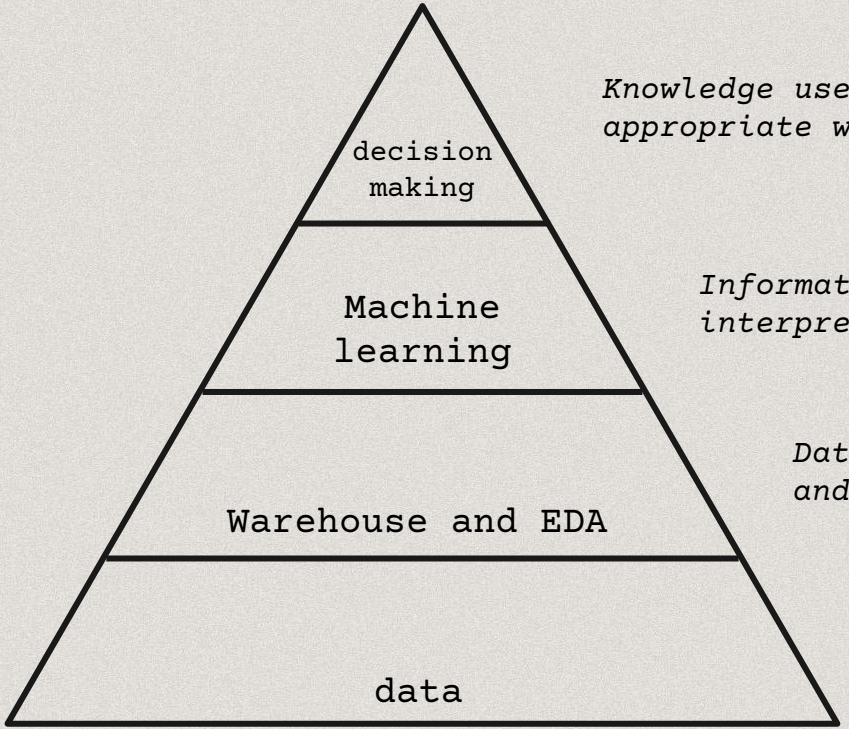
An abstraction of real world via measurement

*The **length** is directly proportional to the amount of data processed and inverse proportional to the informative results.*

Data science pyramid

01

Stages of EDA



Knowledge used in an appropriate way

Information that has been interpreted

Data processed, visualised and contextualised

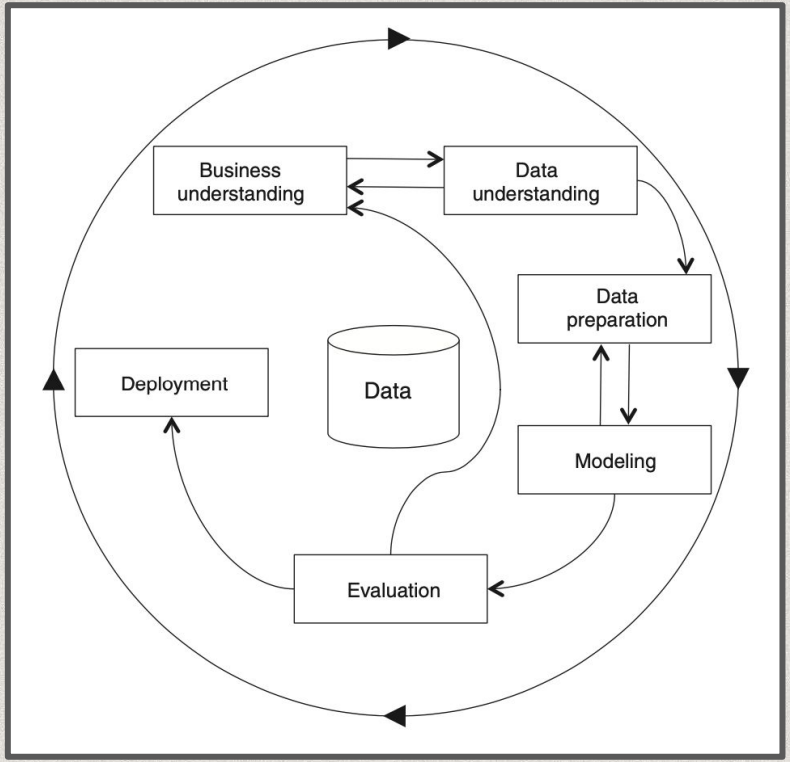
An abstraction of real world via measurement

*The **length** is directly proportional to the amount of data processed and inverse proportional to the informative results.*

CRISP-DM process

01

Stages



Data science activities are part of an iterative life-cycle.

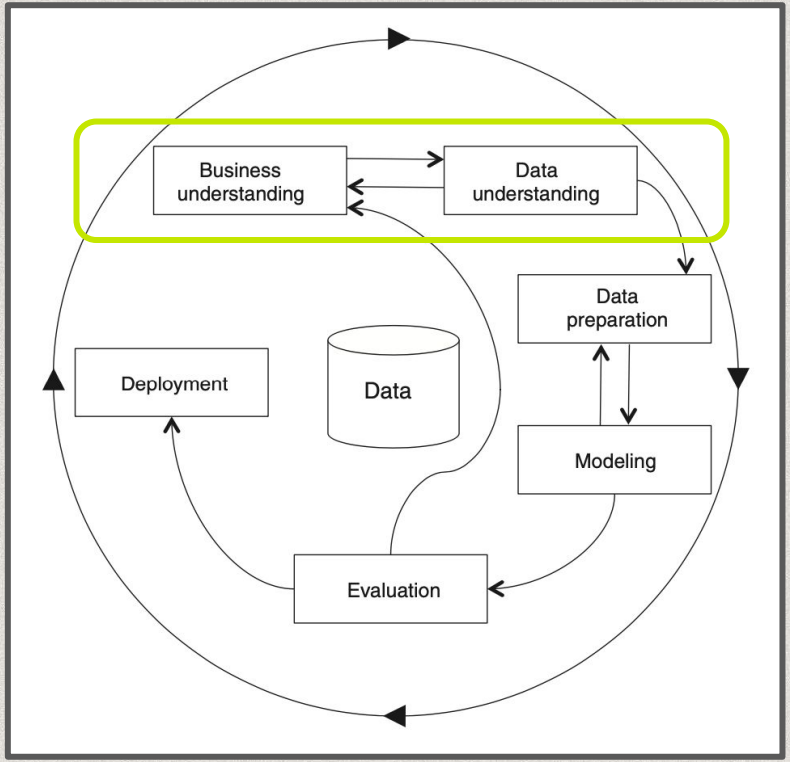
One of the most used models for describing the data mining process is called **Cross Industry Standard Process for Data Mining**.

It is independent from any software or data analysis technique.

CRISP-DM process

01

Stages

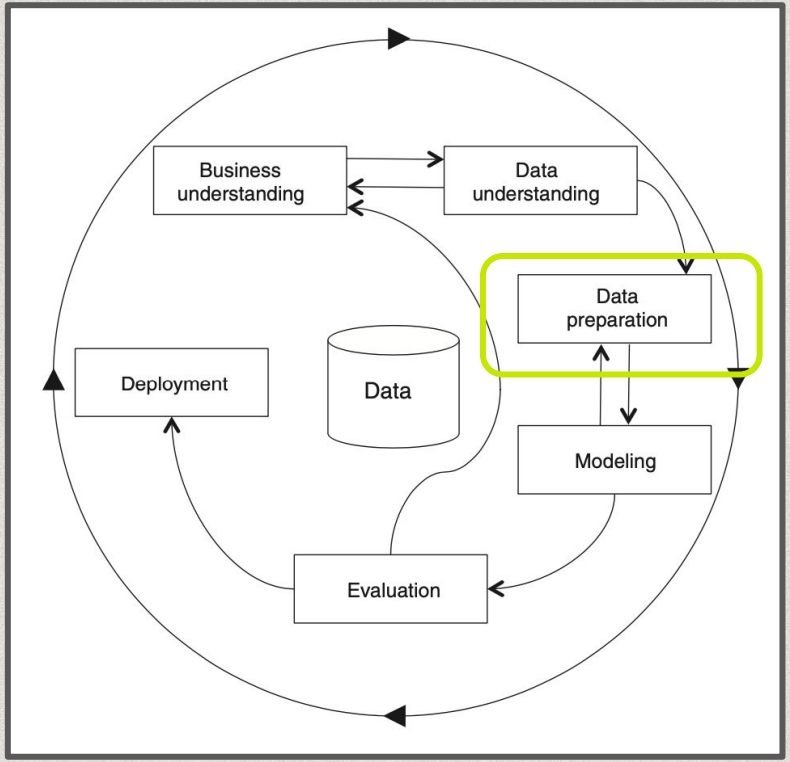


- Set the goal according to needs of a commissioner
  (e.g. support art historians in understanding historiographic trends, write grant proposals)
- Identify the problem
  (e.g. scholars do not know what are the most/least studied topics)
- Identify adequate data sources

CRISP-DM process

01

Stages



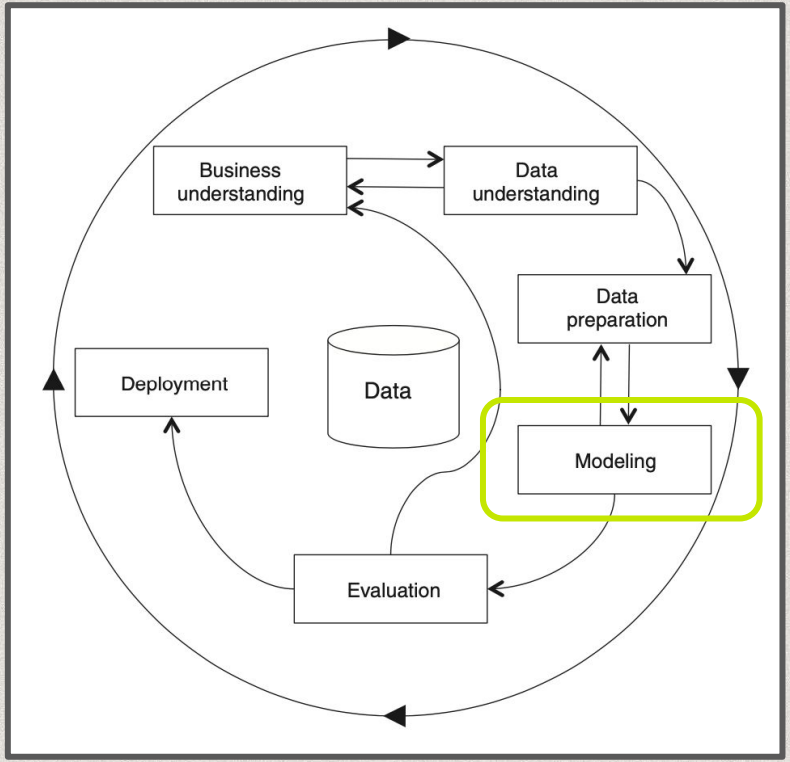
- Gather data for the analysis
  (may require data integration processes, such as mapping, disambiguation, reconciliation, inconsistency checks, merging or dumping).
- Ensure data quality
  (e.g. check data types, convert strings to numbers, extract information from natural language texts)

This process is called **ETL** (extraction, transformation and load)

CRISP-DM process

01

Stages



- Apply algorithms
  (automatic algorithms to understand patterns, e.g. correlation).

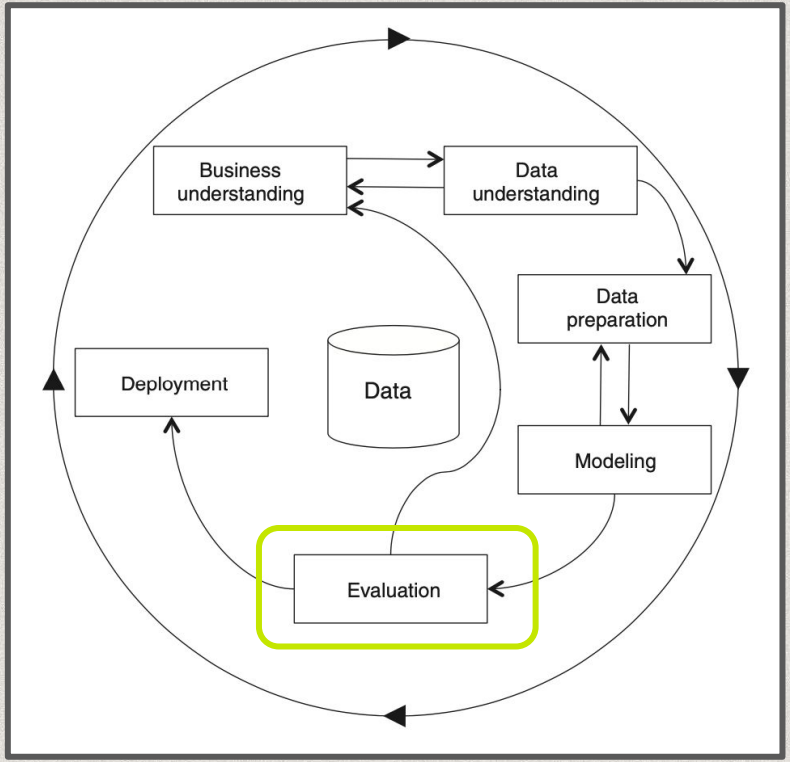
Machine learning algorithms can be applied to data to extract a model that can be reused in new contexts
(e.g. to classify artworks by style)

Less sophisticated methods can achieve similar results.

CRISP-DM process

01

Stages



- Evaluate algorithms and results
  (e.g. precision and recall or classification models, human interpretation of data visualizations).

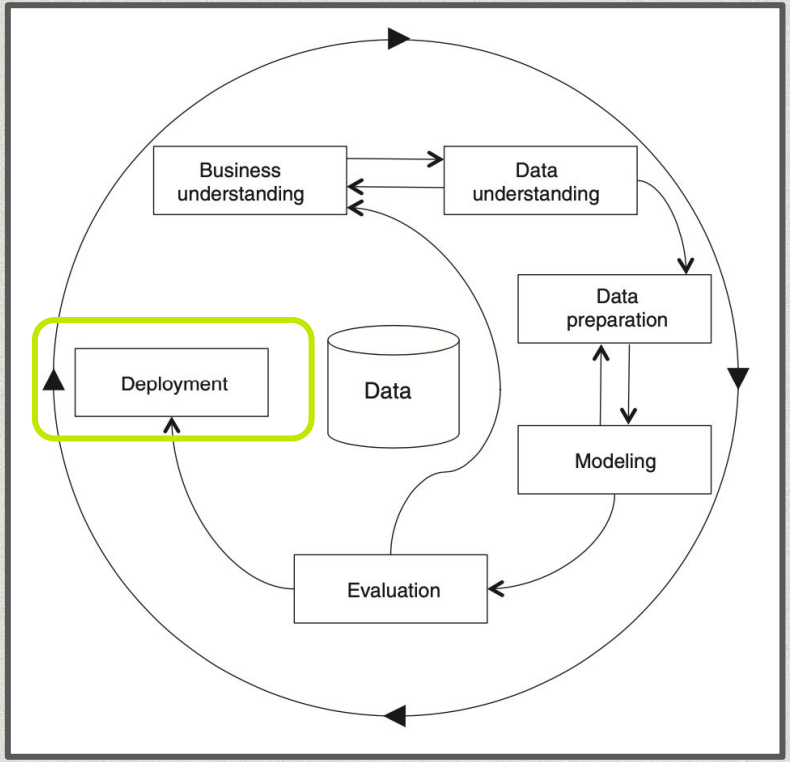
Here you figure your **insights** and **take-home message**, and you understand whether results are **useful** to any.

If results are not satisfying, maybe goals/sources/methods should be revised.

CRISP-DM process

01

Stages



- Integration with existing services or publication of new ones
  (e.g. include a recommender system in an existing catalogue or create a new interface to showcase results).

01

Stages of EDA

In an exploratory data analysis project (yours!), the goal is to provide visual evidence of peculiarities of a phenomenon describe in your data sources, starting from some research question.

01

Stages of EDA

In an exploratory data analysis project (yours!), the goal is to provide visual evidence of peculiarities of a phenomenon describe in your data sources, starting from some research question.

Spoiler alert!

This is how you should organise the jupyter notebook of your project

Seven stages of EDA

01

Stages of EDA

Acquire and parse

Get data from sources (RDF), parse them with a programming language (RDFlib python) in an IDE/GUI (jupyter)

Filter and mine

Get only the data you need, clean and transform them in appropriate data structures (tables)

Represent

Select the best charts and use data viz libraries to show data (seaborn, plotly)

Refine

Highlight the take home message (choose title, colors, describe patterns)

Interact

Provide an environment to present and interact with results (the website: HTML/CSS/JS)

02

Hands-on

Install Jupyter, clean data and visualize information.

Types of analysis

Statistics

Profiling (micro-meso-macro level)

Temporal

WHEN: evolution of variables over time

Geo-spatial

WHERE: trajectories and space dimension of variables

Types of analysis

Topical

WHAT: analysis of categorical variables

Network

WITH WHOM: relations and distance between data points

Hands-on

02

Get all the materials

Install jupyter...

In the terminal

```
pip install notebook
jupyter notebook
```

...or open colab

Login gmail, go to Colab
Select New notebook

Install packages

In the terminal/shell
(if IDE or Jupyter)

```
pip install pandas
pip install pandas_profiling
pip install seaborn
```

Tutorial

Open the tutorial:
in GitHub, Colab
or Jupyter (download)

Exercise

02

Assignment

Review

Review the tutorial

Exercise

Solve the problems (time to code!)

Fill in the form with your answers

TODO

Come prepared! Install these libraries

```
pip install sparql-dataframe
pip install pyproj
pip install mlxtend
pip install networkx
```

Thanks!

Do you have any questions?

marilena.daquino2@unibo.it

https://github.com/marilenadaquino/information_visualization