

---

# DATA SENSE MAKING

## TABLE OF CONTENTS

01

FROM DATA TO WISDOM

02

STAGES OF INFORMATION  
VISUALISATION



03

TYPES OF DATA EXPLORATION

04

HANDS-ON

---

# FROM DATA TO WISDOM

## REMINDER OF DATA SCIENCE OBJECTIVES



Get insights from  
data



Improve understanding  
of data

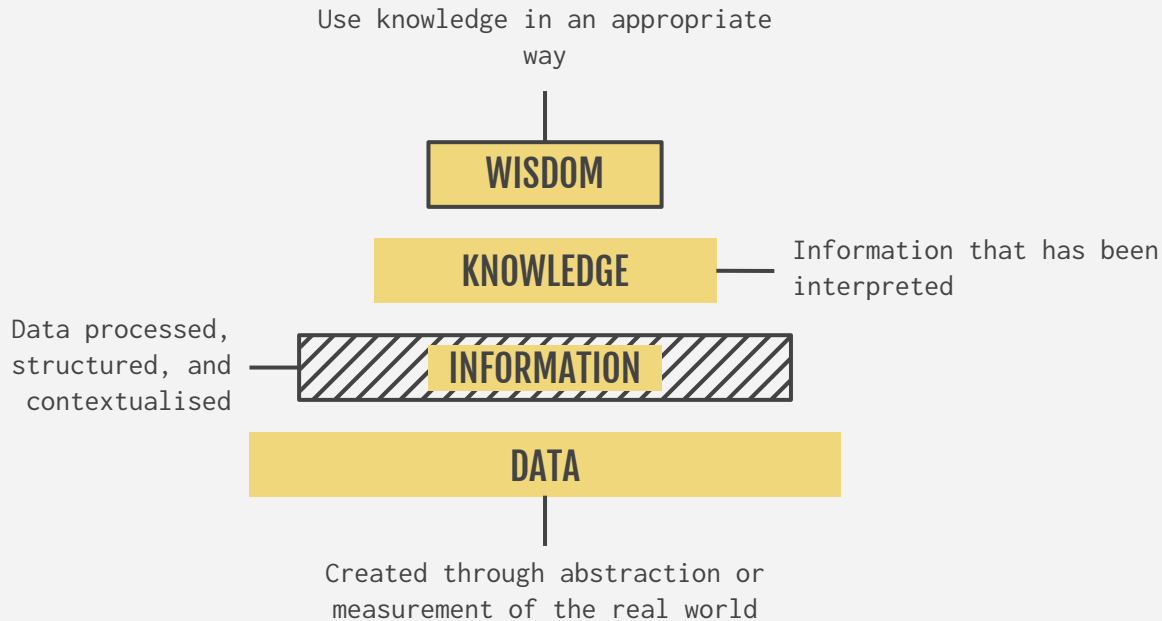


Make informed  
decisions

# DIKW PYRAMID

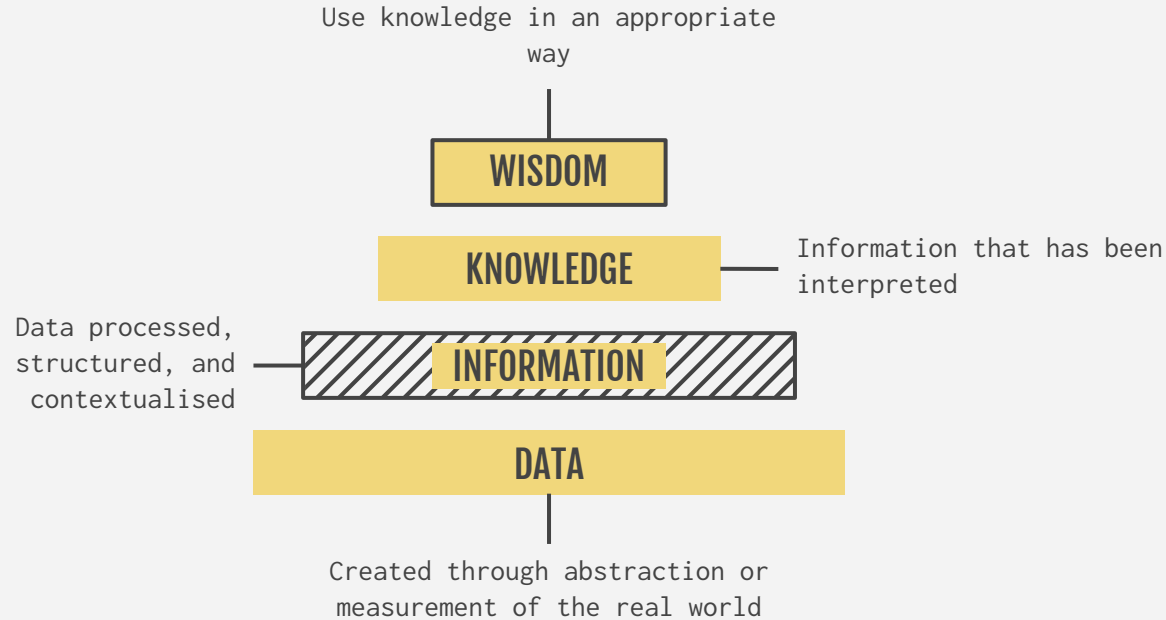
*Where is the Life we have lost in living?  
Where is the wisdom we have lost in knowledge?  
Where is the knowledge we have lost in information?*

T.S. Eliot, The rock, 1934.



# DIKW PYRAMID

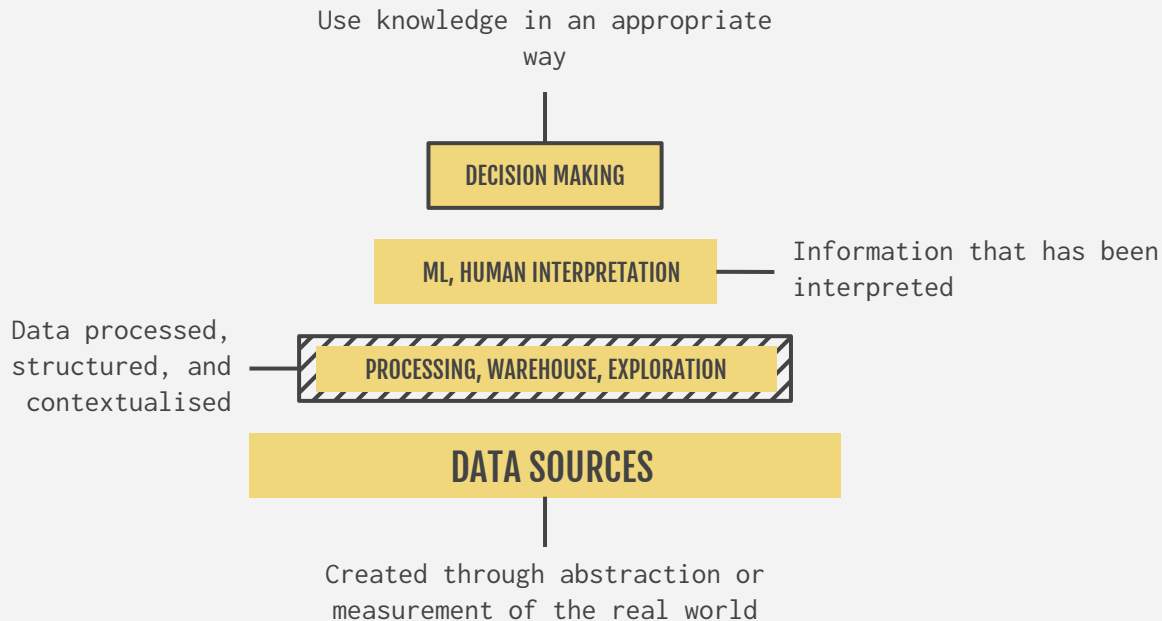
The length is **directly proportional** to the amount of data processed and **inverse proportional** to the informative results.



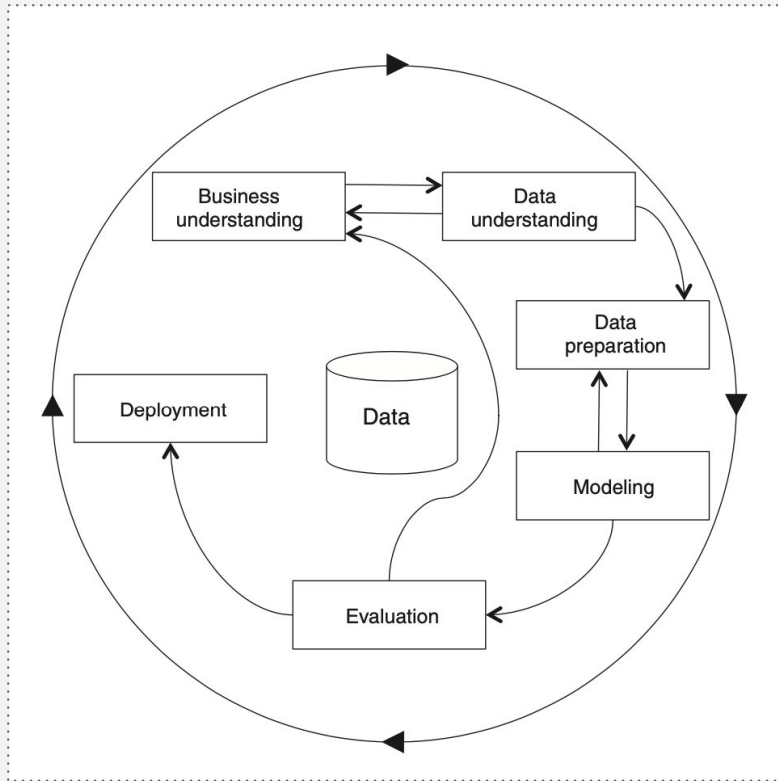
# DATA SCIENCE PYRAMID

The DIKW pyramid corresponds to **data science activities**.

Developers usually spend most of the time in the first two stages, and less in the top two stages.



## CRISP-DM PROCESS



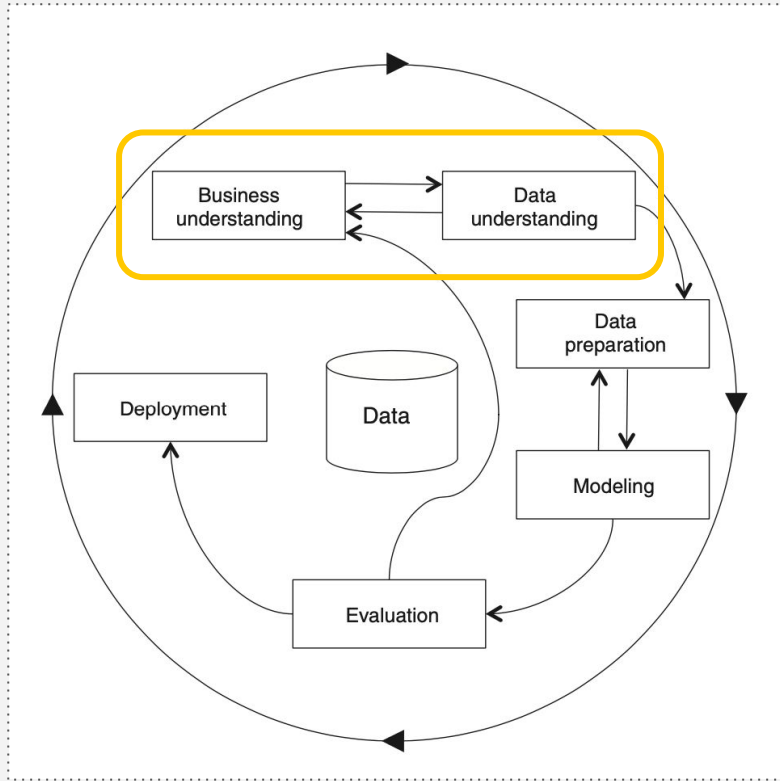
Data science activities are part of an iterative life-cycle.

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

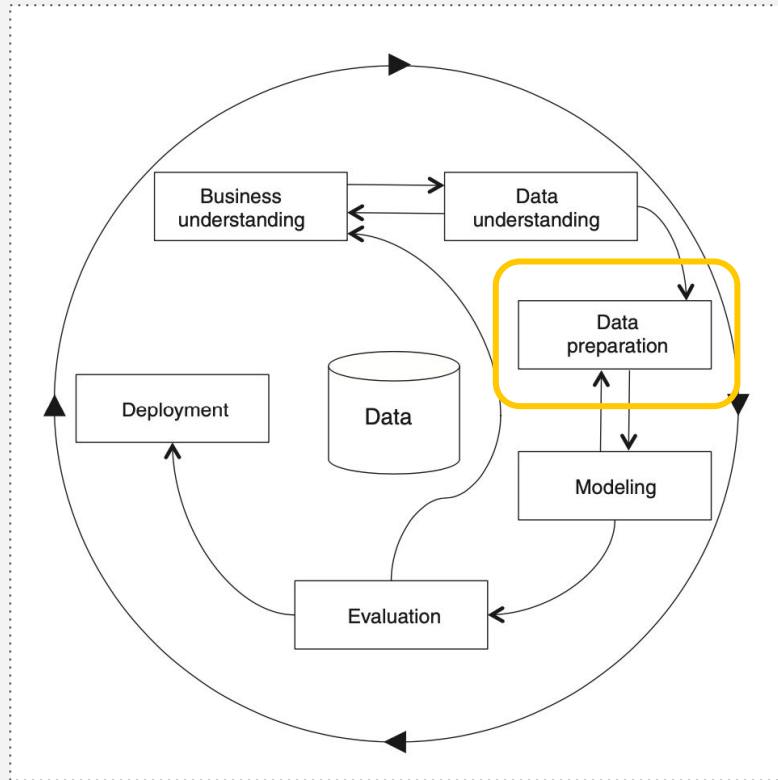


In **Business understanding and data understanding** developers define the goals of the project according to the needs of the commissioner.

Include **identification of a problem**  
and **data exploration** (to see if  
adequate data are available).

If there are data, the process proceeds. If there are no data available developers choose another problem to tackle.

## CRISP-DM PROCESS

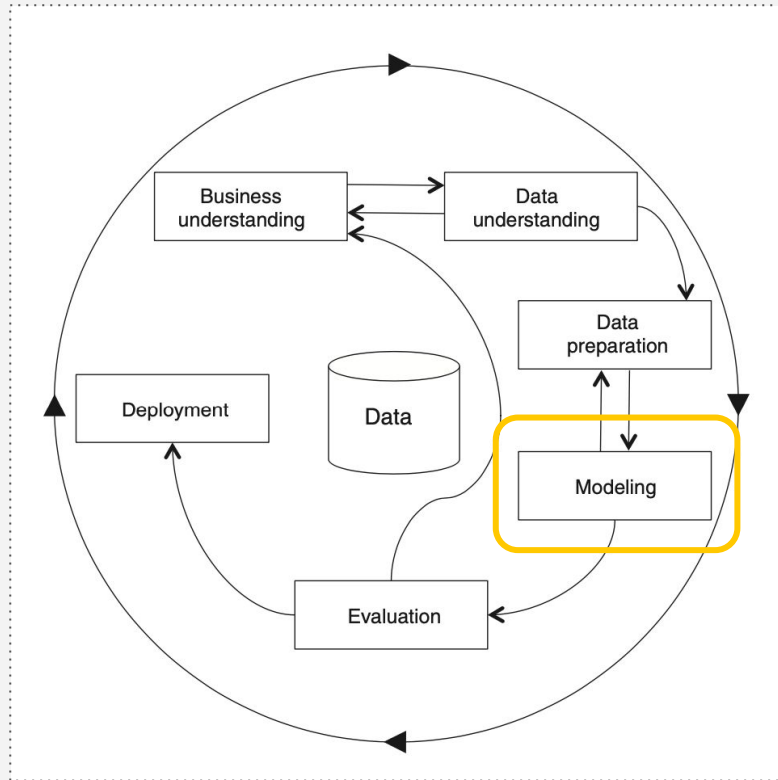


In **Data preparation**, developers **create the dataset** for the analysis.

It may require to **integrate several** data sources, where inconsistencies must be resolved. Data are mapped, merged, and moved to a dataset for data analysis purposes. This process is called **ETL** (extraction, transformation and load)

Secondly, **data-quality** checks are performed.

# CRISP-DM PROCESS



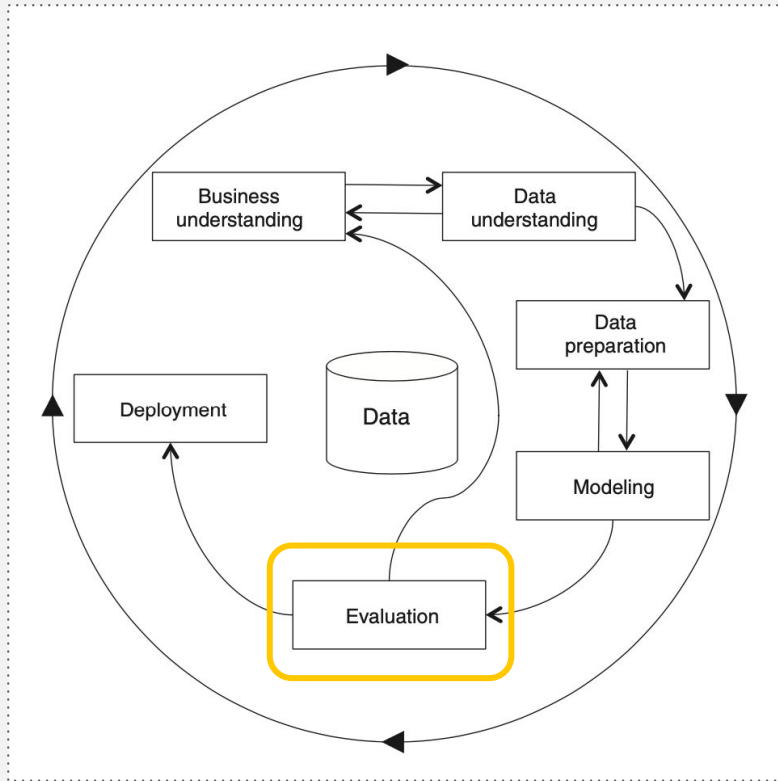
In **modeling**, automatic algorithms are applied to extract patterns of interest and to create a model that encodes such patterns.

Usually, **Machine Learning** methods are here applied to understand which algorithm better fits the data and helps to extract the patterns.

A model can also be a **decision tree**.

**DISCLAIMER** In this course we will use less sophisticated methods for the analysis, but the process still applies.

# CRISP-DM PROCESS

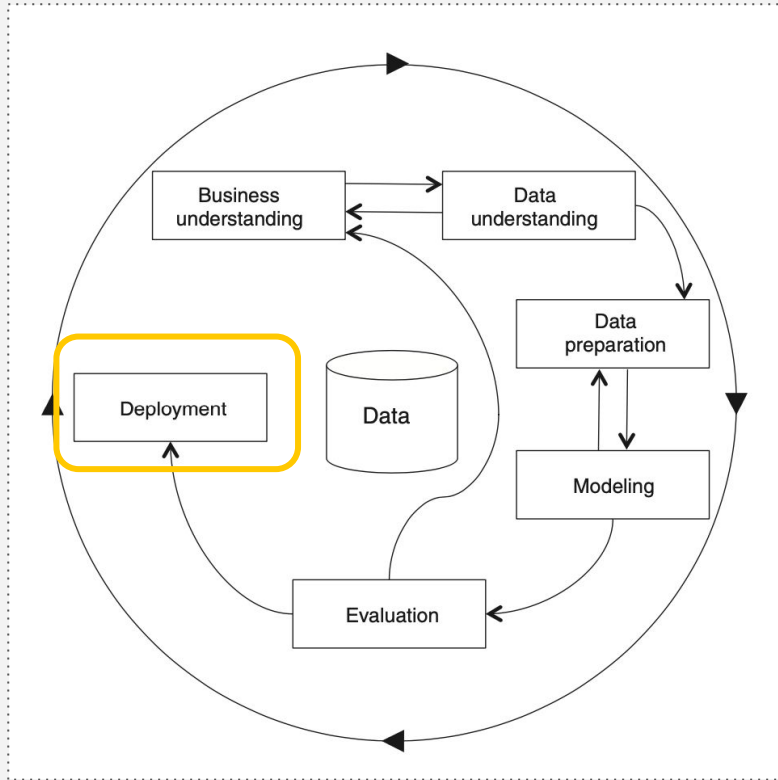


In the **evaluation**, developers test their model with respect to the initial goals.

Are the objectives achieved? What is missing? What can be done better?

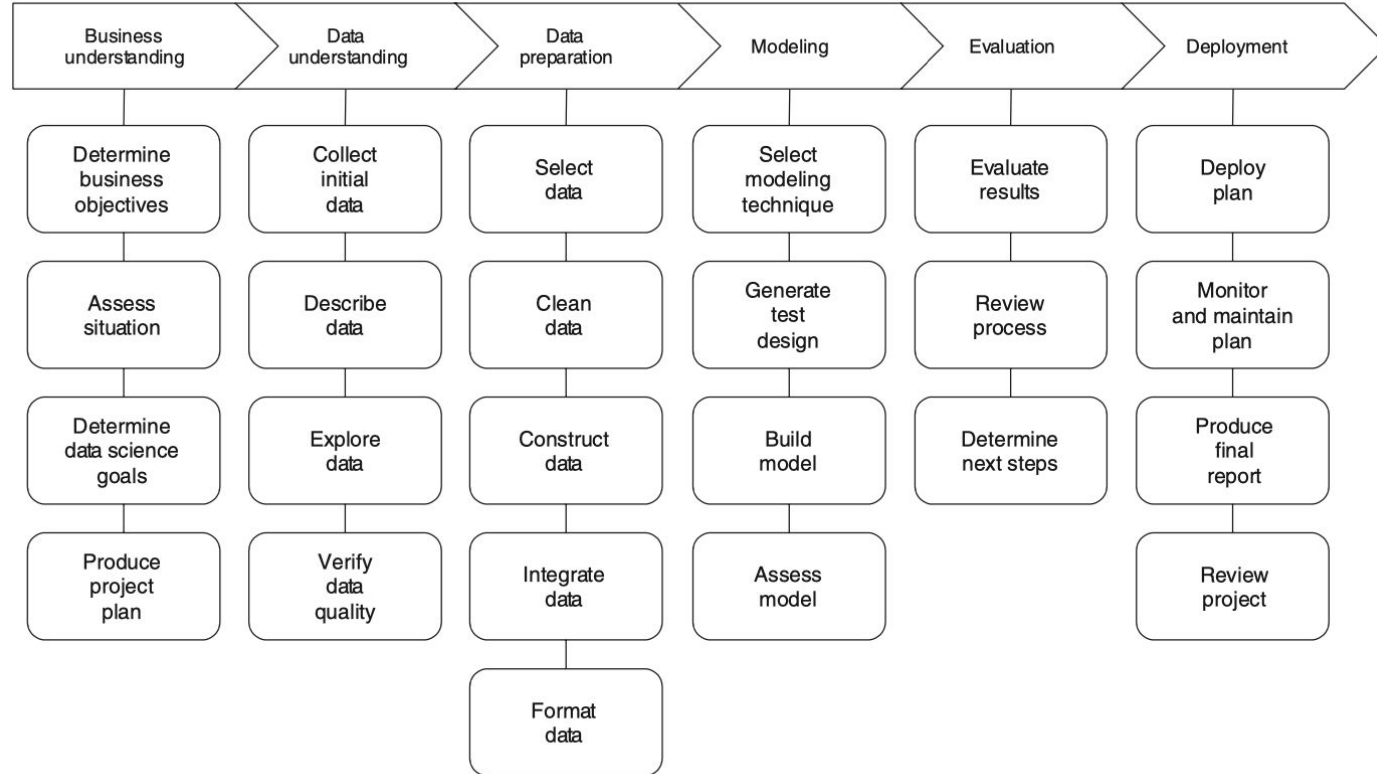
# CRISP-DM PROCESS

In the **deployment**, developers study how to integrate their results in the original infrastructure of the commissioner.





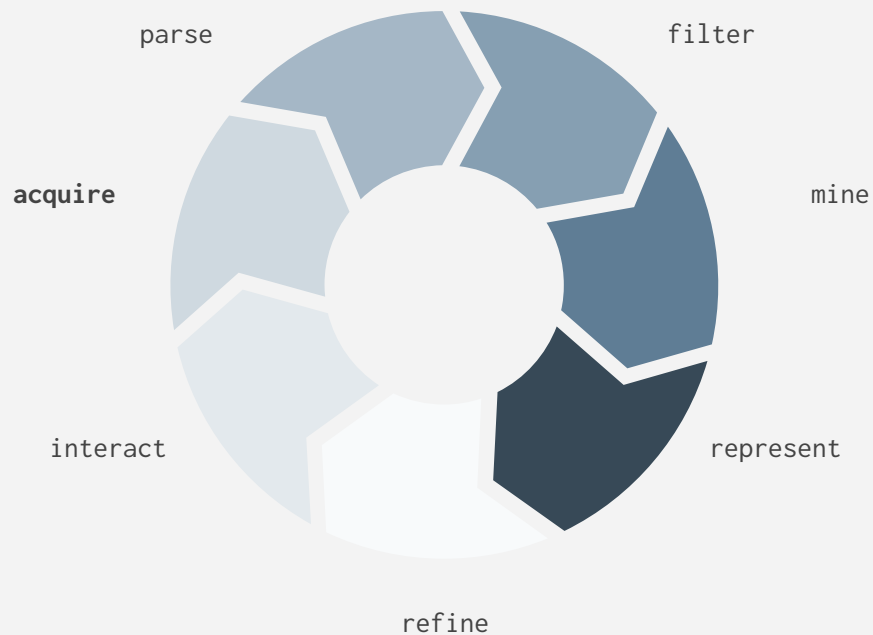
# CRIS



---

# STAGES OF INFORMATION VISUALISATION

# THE SEVEN STAGES OF INFORMATION VISUALISATION





# ACQUIRE DATA

## ARTchives

We start from the RDF dump of the ARTchives project, which includes information about art historians, their collections, research topics, and related cultural institutions.

Always start with **existing data** and discover what is the **added value** of the dataset (e.g. find relevant information that characterise it, find patterns in data).



# ACQUIRE DATA PRELIMINARIES

## ARTchives

The dataset is meant  
to facilitate  
historiographical  
research.

It's rather small  
(25-30 records).

Includes Wikidata and  
local terms.

Ask some preliminary **questions**:

- Who created the dataset and why?
  - How big is it?
  - What do fields mean?

**Get familiar** with a few records.



## ACQUIRE DATA DEFINE ATTRIBUTES FOR THE ANALYSIS

### ARTchives

*Example Univariate:*  
the distribution of  
the property  
**birthplace**. How many  
historians are  
annotated with that  
property?

Select the number of (dependent or independent) data attributes that you want to work on.

**Univariate.** A single variable studied against other independent variables.

**Bivariate.** Two dependent variables studied against other independent variables.

**Trivariate.** Three dependent variables studied against other independent variables.

**Multivariate.** Multiple dependent variables studied against other independent variables.



## PARSE DATA

### ARTchives

The data are parsed via python library RDFLib, which allows us to manipulate graph data.

After you acquire the data, these need to be parsed—changed into a format that tags each part of the data with its intended use.



# FILTER AND MINE DATA

## ARTchives

To answer specific questions we may need to **filter** some data out, e.g. data about historians' birthplaces.

In order to use python libraries for data analysis and visualisation we need to **convert** filtered data into other formats, e.g. a table.

After you transform the raw data into a more suitable format, you can perform operations such as **filtering**, **sorting**, **re-organising** so that patterns can be easily identified.

This step involves math, statistics, and data mining.



# REPRESENT DATA

## ARTchives

The use some python libraries for plotting information and get some new insight from the data we have.

This step determines the basic form that a set of data will take. Some data sets are shown as lists, others are structured like trees, and so forth.

How you choose to represent the data can influence the very first step (what data you **acquire**) and the third step (what particular pieces you **extract**).



## REFINE DATA

### ARTchives

After interpreting the visualizations we will tweak them to highlight most meaningful insights.

Graphic design methods are used to clarify the representation by calling more attention to particular data.



# INTERACT WITH DATA

Interaction means **letting the user control or explore the data**. Interaction might cover things like selecting a subset of the data or changing the viewpoint.





---

# TYPES OF DATA EXPLORATION

## LEVELS OF ANALYSIS

STATISTICAL

TEMPORAL

GEOSPATIAL

**Profiling** (at  
micro-meso-macro level)

**WHEN:** evolution of  
variables over time  
variable

**WHERE:** trajectories and  
space dimension of  
variables


## LEVELS OF ANALYSIS

**TOPICAL**

**NETWORK**

**WHAT:** analysis of  
categorical variables

**WITH WHOM:** relations  
between data points



ELECTRONIC PUBLISHING  
AND DIGITAL STORYTELLING  
Lesson 5

---

# HANDS-ON

# WHAT'S THE PLAN? SET UP YOUR PROJECT!

Install Jupyter notebook locally and create the notebook for your python code

**CREATE A JUPYTER  
NOTEBOOK**

First steps to answer a research question via data visualisation

**DATA ACQUISITION /  
PARSING / FILTERING**

Install and use some python libraries for **exploring** data

**DATA REPRESENTATION**

# INSTALL AND LAUNCH JUPYTER NOTEBOOK

Instructions: <https://jupyter.org/install>

In the shell run:

```
pip install notebook
```

Then, in the shell, move to the folder  
where your code is  
(e.g. `cd Desktop/dhdk_epds/tutorials/`)  
and run:

```
jupyter notebook
```

# CREATE A JUPYTER NOTEBOOK

When the browser opens let's explore together what is there and **let's create your first Python file.**

Top-right menu: New > Python 3

- Rename the file
- Have a look at the editor menu
- Create cells and define the type of content (markdown or code)
  - Basics of markdown
  - Example of python code
    - Run

# INSTALL PYTHON LIBRARIES

Pandas

```
pip install pandas
```

```
pip install pandas_profiling
```

Seaborn

```
pip install seaborn
```



## MOVE TO THE TUTORIAL

Open the course repository on the browser  
<https://github.com/marilenadaquino/epds>  
Go to the folder tutorials/ and open the file  
**dataviz\_tutorial.ipynb**

**You can**

- Either browse it in the browser
- Download the file, include it in a folder  
and open it with jupyter notebook

## WANT TO USE COLAB INSTEAD?

Open the course repository on the browser  
<https://github.com/marilenadaquino/epds>  
Go to the folder tutorials/ and open the file  
**dataviz\_tutorial.ipynb**

Replace “<https://github.com/>” with  
“<https://github.com/tocolab.com/>” in the URL of  
the ipynb

The first time you run the cell including  
imports, include for each package the following  
instruction (replace placeholder <PACKAGENAME>)

```
!pip install <PACKAGENAME>
```

# HOMEWORK

Create your first Jupyter notebook and submit it in a week!

<https://forms.gle/EeyyG5cStdNpUfAp9>

In this Jupyter notebook you'll have to:

- Acquire / Parse / Filter ARTchives data in order to answer the following question:

**What are the most referenced people in ARTchives archival collections?**

- Represent the data in a bar chart



---

# THANKS

Does anyone have any questions?

marilena.daquino2@unibo.it

[github](#)