HSLU Hochschule Luzern

File Structure in a Data Science Project
Industry Best Practices

Einführung

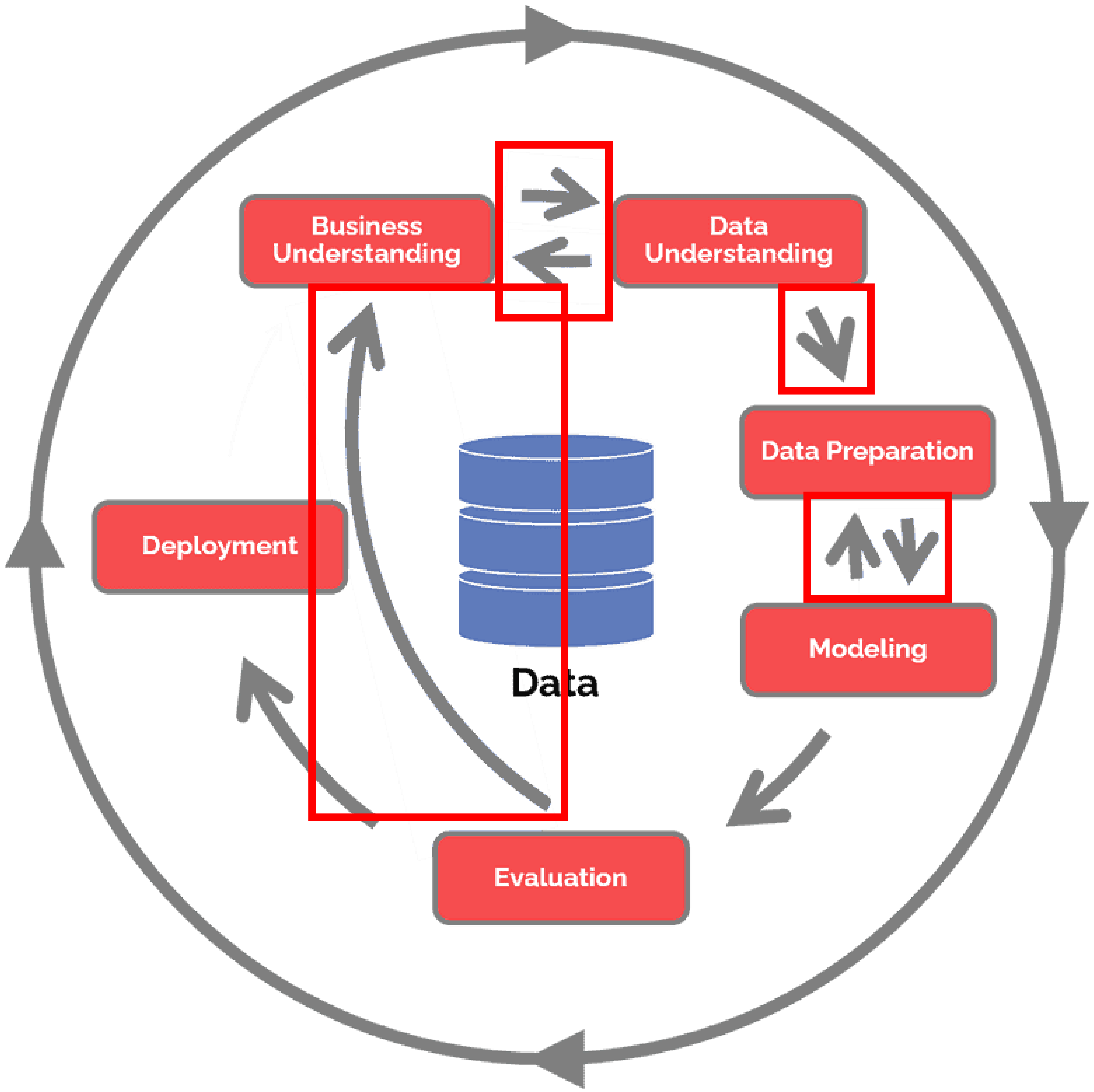
Dr. Umberto Michelucci

3. Februar 2025

FH Zentralschweiz

Structure your files according to the phases of your project

CRISP-DM - Phasen

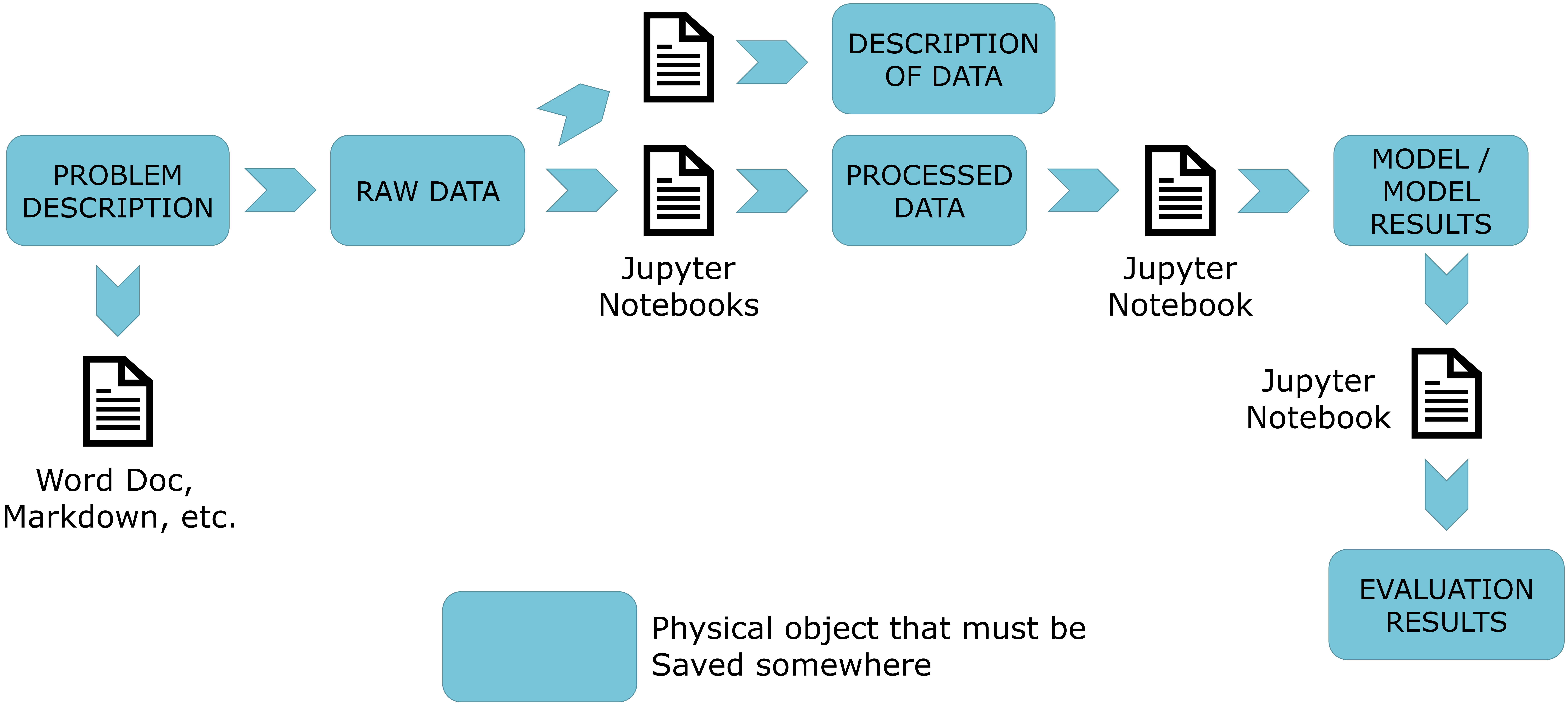


Source: https://www.datascience-pm.com/crisp-dm-2/ (last accessed 23rd Feb. 2023)

Files and phases

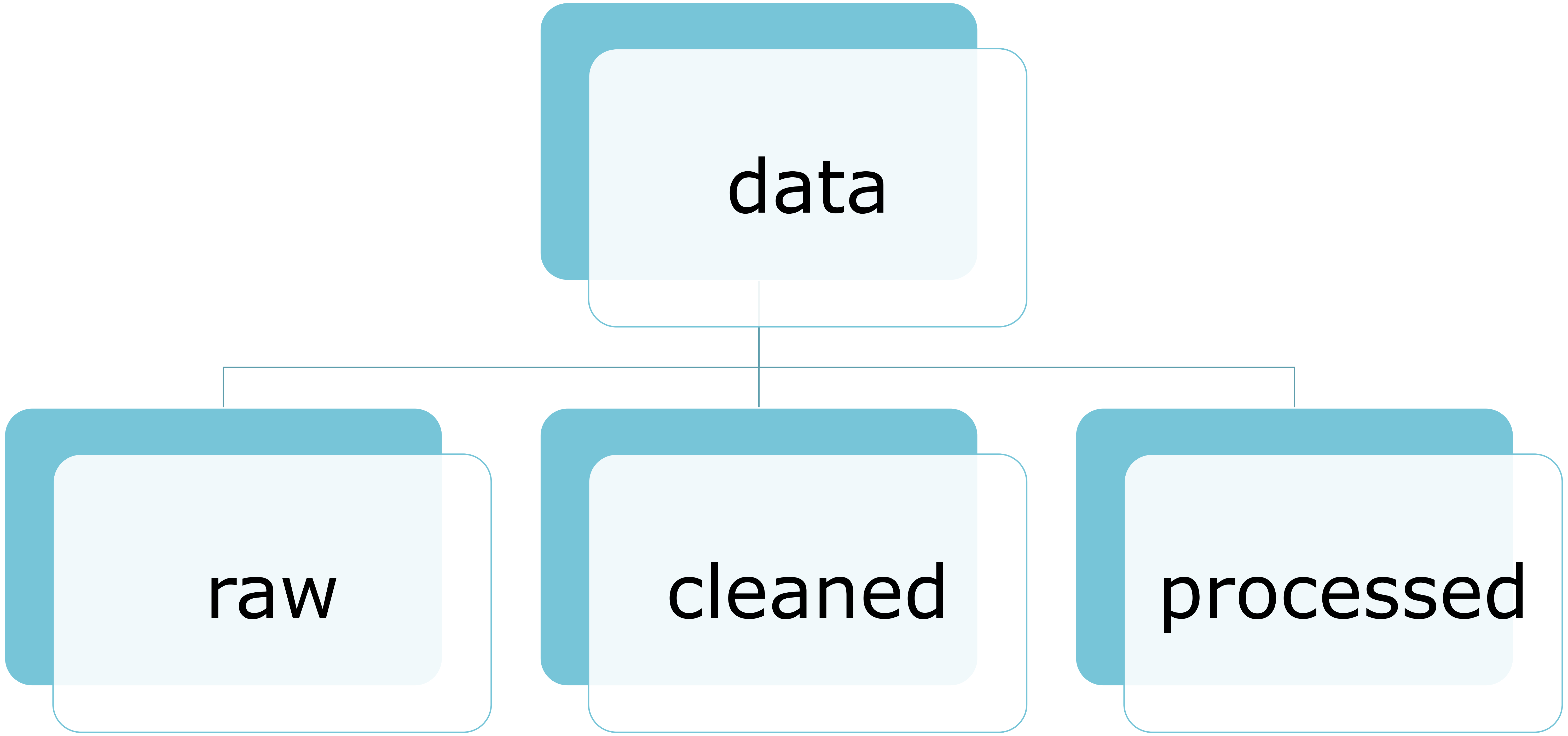
1. **Business understanding**: a short document (word, markdown, jupyter notebook) explaining in detail what is the problem you are trying to solve)
2. **Data understanding**: at least one (maybe multiple) jupyter notebooks in which you analyse the data you plan to use. Document it well and use visualisation as much as you can.
3. **Data preparation**: at least one jupyter notebook where you perform the data cleaning and preparation. At the end save the prepared dataset to use in the next phase.
4. **Modeling**: at least one juypter notebook (or script) where you train your model. Save the results and the models that you want to evaluate here (sometime training models takes many hours, you do not want to repeat that many times).
5. **Evaluation**: at least one jupyter notebook where you document the evaluation phase.
6. **Deployment**: often multiple documents, scripts, etc. depending on the situation and the project.

Files / Phases



Physical object that must be Saved somewhere

Example: data folder structure



Unmodified data with references, description, etc.

Cleaned data. For example without records with missing data.

The final dataset that will be used. Maybe multiple datasets merged, for example.

A couple of tips

1. Every jupyter notebook should run from top to end and always produce the same results.
2. Different experiments should be in different notebooks or save the results separately (if you want to try different dataset normalisation write separate pieces of code. DO NOT modify the code manually, save something and modify it again).
3. The entire project should be REPRODUCIBLE. You should be able to re-run all jupyter notebooks, from the first to the last and get the same exact results.