HSLU Hochschule Luzern

CRISP-DM Introduction Data Science Projects Structue and Phases

Einführung

Umberto Michelucci
23. Februar 2023

FH Zentralschweiz

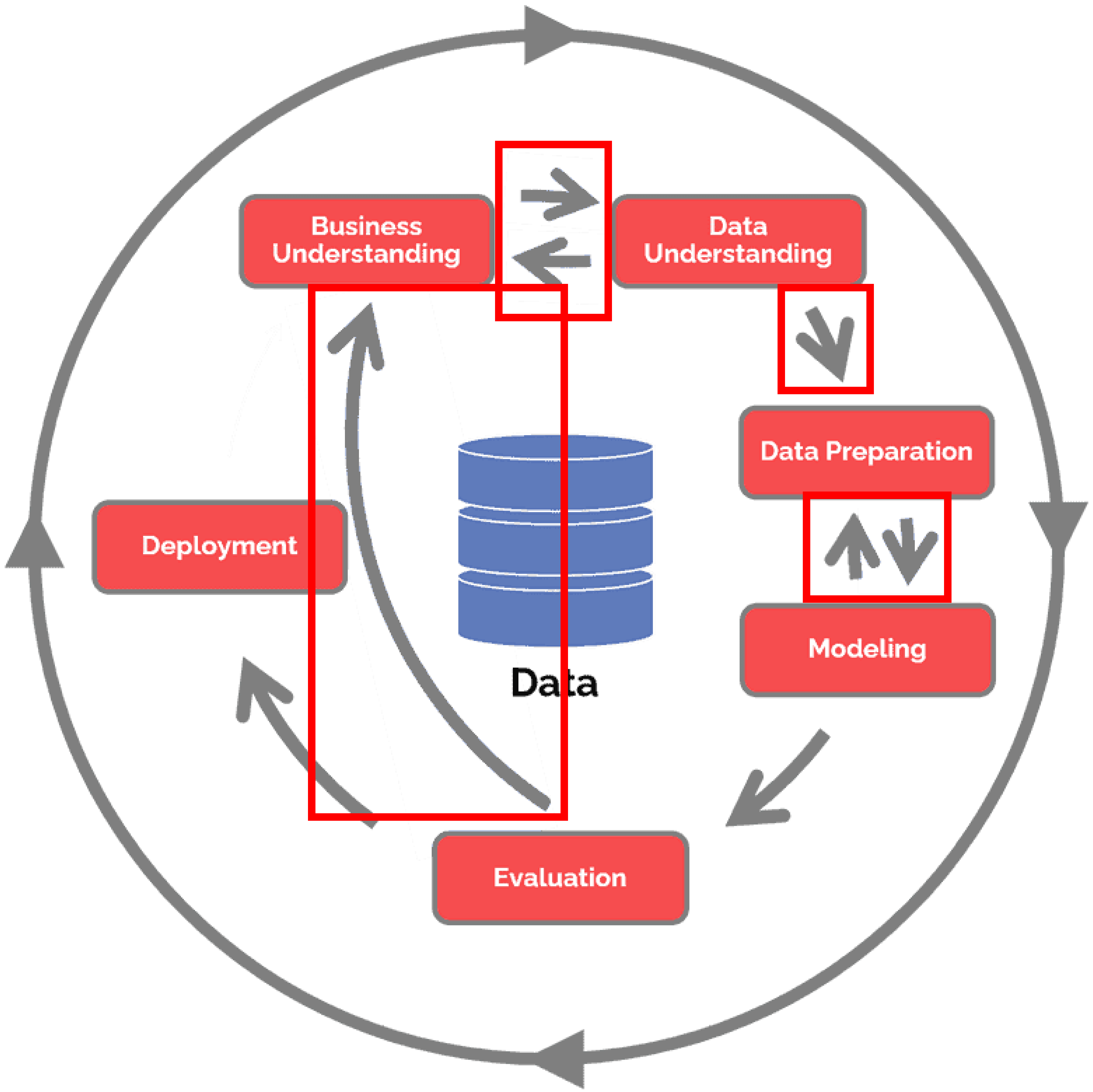
What is CRISP-DM?

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model that serves as the base for a data science process. It has six phases:

1. **Business understanding** – What does the business need?
2. **Data understanding** – What data do we have / need? Is it clean?
3. **Data preparation** – How do we organize the data for modeling?
4. **Modeling** – What modeling techniques should we apply?
5. **Evaluation** – Which model best meets the business objectives?
6. **Deployment** – How do stakeholders access the results?

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

CRISP-DM - Phasen



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

Phase 1 - Business Understanding

Any good project starts with a deep understanding of the customer’s needs!

1. **Determine business objectives:** You should first “thoroughly understand, from a business perspective, what the customer really wants to accomplish.”* and then define business success criteria.
2. **Assess situation:** Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
3. **Determine data mining goals:** In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.
4. **Produce project plan:** Select technologies and tools and define detailed plans for each project phase.

* https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf (Last accessed 23rd Feb. 2023)

Phase 2 - Data Understanding

1. **Collect initial data: Acquire** the necessary data and (if necessary) load it into your analysis tool (for example Python).
2. **Describe data:** Examine the data and **document** its surface properties like data format, number of records, or field identities.
3. **Explore data:** Dig deeper into the data. Query it, **visualize** it, and identify relationships among the data.
4. **Verify data quality:** How clean/dirty is the data? **Document** any quality issues.

ACQUIRE – DOCUMENT - VISUALISE

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

Phase 3 - Data Preparation

A common rule of thumb is that 80% of the project is data preparation.

1. **Select data**: Determine which data sets will be used and **document** reasons for inclusion/exclusion.
2. **Clean data**: Often this is the lengthiest task. Without it, you'll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.
3. **Construct data**: Derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields.
4. **Integrate data**: Create new data sets by combining data from multiple sources.
5. **Format data**: Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.
6. **Save the final dataset!**

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

Phase 4 - Modeling

What is widely regarded as data science's **most exciting** work is also often the **shortest phase** of the project.

1. **Select modeling techniques**: Determine which algorithms to try (e.g. regression, neural net).
2. **Generate test design**: Pending your modeling approach, you might need to split the data into training, test, and validation sets.
3. **Build model**: As glamorous as this might sound, this might just be executing a few lines of code like "reg = LinearRegression().fit(X, y)".
4. **Assess model**: Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

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

Phase 5 - Evaluation

the *Evaluation* phase looks more broadly at which model best meets the business and what to do next.

1. **Evaluate results**: Do the models meet the business success criteria? Which one(s) should we approve for the business?
2. **Review process**: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.
3. **Determine next steps**: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

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

Phase 6 - Deployment

"Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise."

–CRISP-DM Guide

A model is not particularly useful unless the customer can access its results.

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

Phase 6 - Deployment

1. **Plan deployment**: Develop and document a plan for deploying the model.
2. **Plan monitoring and maintenance**: Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.
3. **Produce final report**: The project team documents a summary of the project which might include a final presentation of data mining results.
4. **Review project**: Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

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

Is CRISP-DM Agile**?

CRISP-DM indirectly advocates agile principles and practices by stating: “The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next.”*

* https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf (Last accessed 23rd Feb. 2023)
** https://www.atlassian.com/agile/scrum (Last accessed 23rd Feb. 2023)