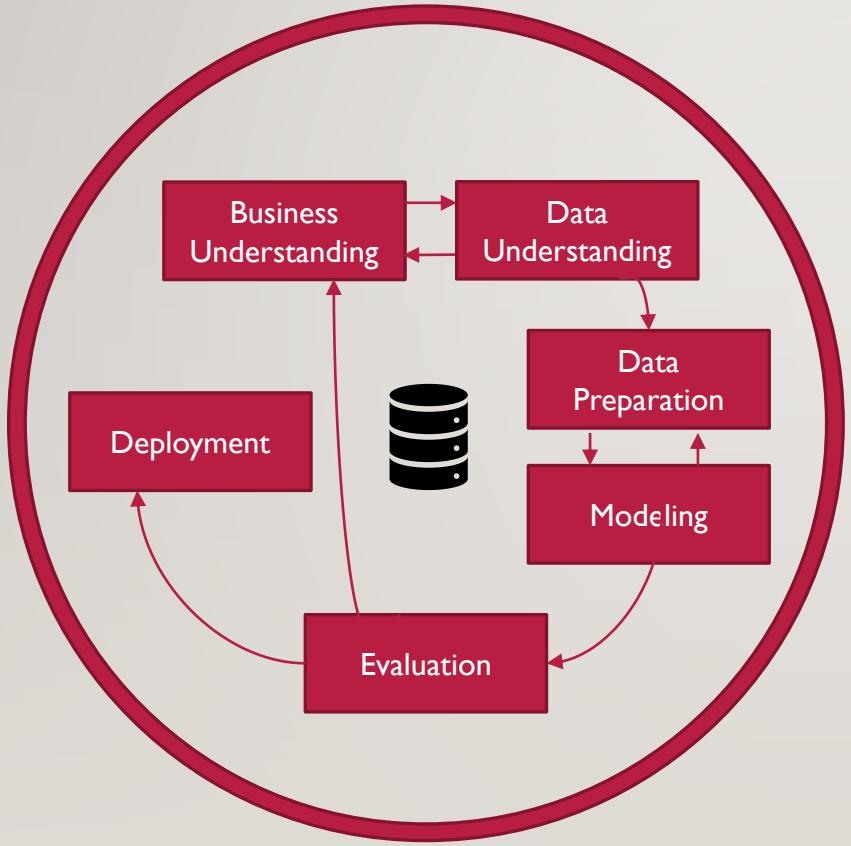
DATA SCIENCE METHODOLOGY

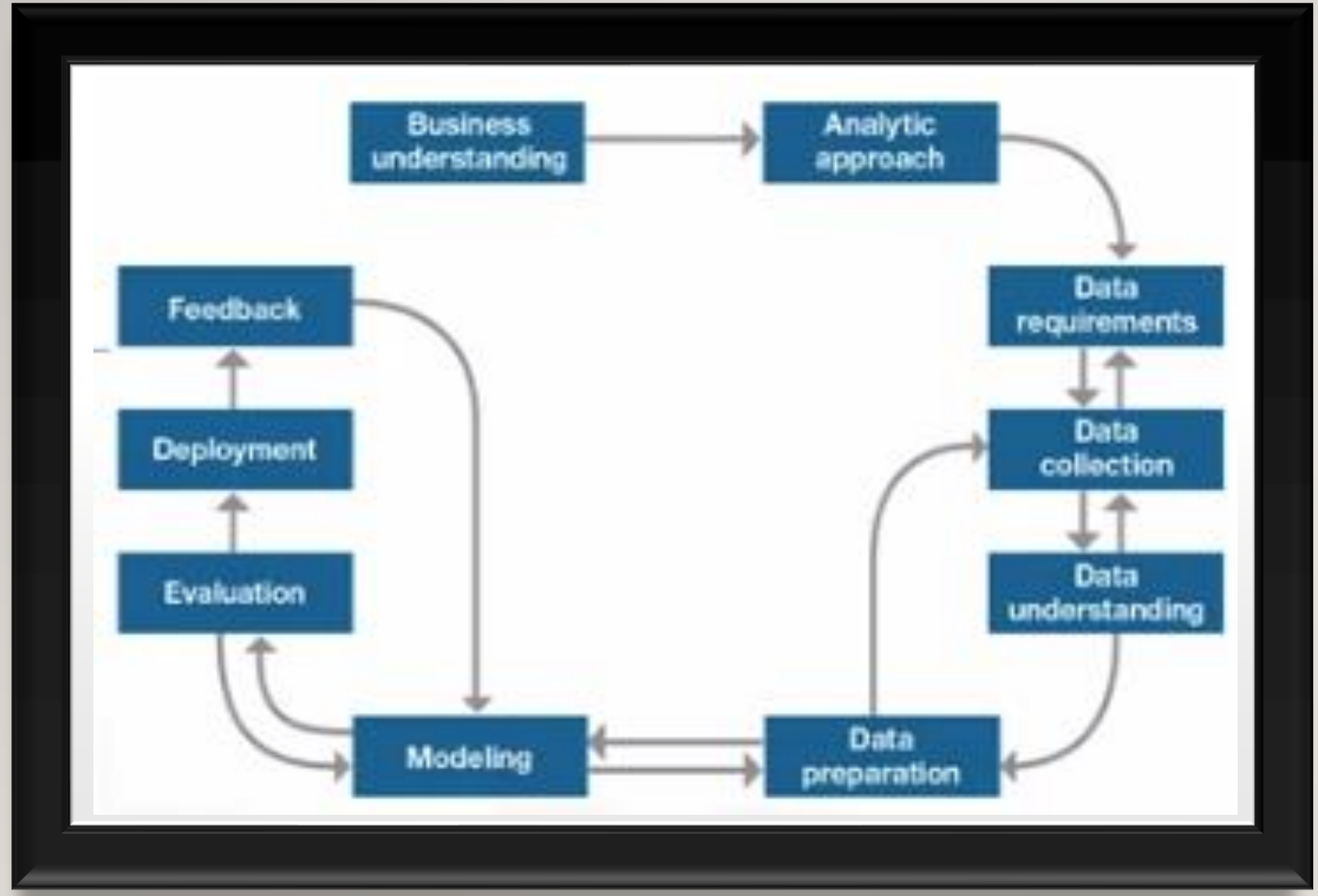
CRISP-DM

Fabio Lonardoni

CRISP-DM



Business Understanding: The intention of the project is outlined.
Data Understanding: Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected
Data Preparation: Data must be transformed into a useable subset unless it is determined that more data is needed. 70% of the time.
Modeling: the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining
Evaluation: The selected model must be tested with pre-selected test. Training mode.
Deployment: the model is used on new data outside of the scope of the dataset and by new stakeholders.



ITERATIVE PROCESS & NEVER ENDS

FROM PROBLEM TO APPROACH

BUSINESS UNDERSTANDING

- What is the problem that you are trying to solve?
  - Define the Goals
  - Define the Objectives
- It's important defining the best questions for the problem

ANALYTIC APPROACH

- How can you use data to answer the question?
- What is the best Machine Learning algorithm?

Descriptive	*What is the current status?*
Diagnostic (Statistical Analysis)	*What happened?* *Why is this happening?*
Predictive (Forecasting)	*What if these trends continue?* *What will happen next?*
Prescriptive	*How do we solve it?*

- It's important defining the best patterns needed to address the questions

FROM REQUIREMENTS TO COLLECTION

DATA REQUIREMENTS

- What are the data and what is the schema we need to automate the definition of the problem?
  - Identify the data
  - Identify the format
  - Identify the sources

DATA COLLECTION

- Collect all information we need
- Application of descriptive statistics and visualization techniques to assess the content, quality and initial insights about data

FROM UNDERSTANDING TO PREPARATION

DATA UNDERSTANDING

- Encompassing all activities related to constructing the dataset
- Is the data that you collected representative of the problem to be solved?

DATA PREPARATION

- Preparing data for modeling stage
  - Addressing missing values
  - Correcting invalid values
  - Addressing outliers
  - Removing duplicate data
  - Properly formatting the data
- Exploring the data further and making sure that it is in the right format for the machine learning algorithm that we selected in the analytic approach stage

FROM MODELING TO EVALUATION

MODELING

- Data Modelling focuses on developing models that are either descriptive or predictive.
- Preparing Training set and Test set

EVALUATION

- Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.
- Does the model used really answer the initial question or does it need to be adjusted?
- Ensuring that the model is working as intended
- Ensuring that the data are properly handled and interpreted
- Ensuring the model is designed as intended
- 2 Phases:
  - Diagnostic measures phase
  - Statistical significance testing

FROM DEPLOYMENT TO FEEDBACK

DEPLOYMENT

- Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test.

FEEDBACK

- Once in play, feedback from the users will help to refine the model and assess it for performance and impact. The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required.