CRISP-DM

An Introduction to Effective Data Mining

Ricardo Fitas
21/04/2023

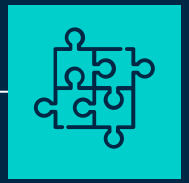
Ricardo Fitas

MSc Mechanical Engineering (FEUP) – Machine Design

Work Experience (Data & Optimization):

- Bosch Termotechik (2019)
- RWTH Institut für Textiltechnik (2020/21)
- RWTH Werkzeugmaschinenlabor (2021)
- INEGI CETRIB (2021/22)
- Research Assistant – Technical University of Munich (2022/23)
- Freelancer, Consultant, PhD Candidate

Table of Contents



01

INTRODUCTION TO DATA MINING AND CRISP-DM

What is data mining?
Benefits of CRISP-DM

02

CRISP-DM FRAMEWORK

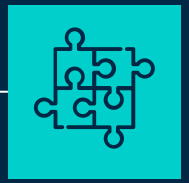
What are the different phases of CRISP-DM?

03

PYTHON LIBRARIES AND CASE STUDIES

Some practical applications of CRISP-DM

Table of Contents



01

INTRODUCTION TO DATA MINING AND CRISP-DM

What is data mining?
Benefits of CRISP-DM

02

CRISP-DM FRAMEWORK

What are the different phases of CRISP-DM?

03

PYTHON LIBRARIES AND CASE STUDIES

Some practical applications of CRISP-DM

Data Mining

"Finding patterns in historical data and then leveraging those patterns on current data to make future predictions."

—Keith McCormick

Data Mining

- Typically uses statistical methods and machine learning algorithms;
- Has become na essential tool for businesses, organizations, and governments to make data-driven decisions and improve their operations.

Data Mining vs Data Analysis

- The difference lies in their focus on patterns:
  - Data Mining is focused on discovering hidden patterns (by using statistical and machine learning algorithms, and usually in the perspective of future deployment);
  - Data Analysis is focused on understanding underlying patterns (usually involve exploring and understanding the data related to the question we want to answer).

Data Mining vs Data Analysis

- Examples of typical questions related to data analysis:
  - What is the average purchase amount by customer segment?
  - What is the distribution of customer satisfaction ratings by region? Are there regional differences in the factors contributing to customer satisfaction?
- Examples of typical questions related to data mining:
  - Which customers are most likely to churn and stop using our product, and what factors contribute to their likelihood of churn?
  - What are key factors that influence the success of marketing campaigns, and how can we use this information to improve future campaigns?

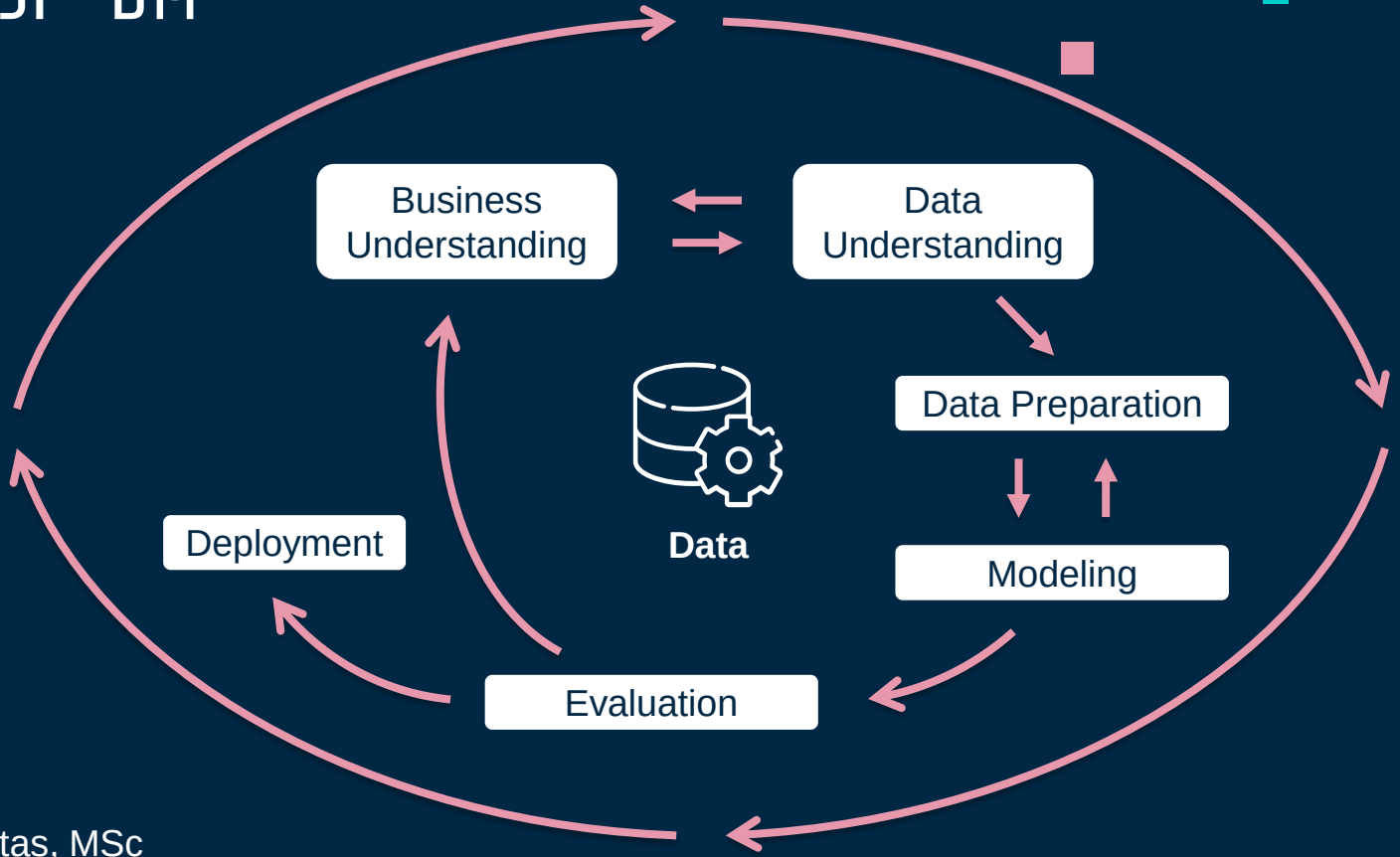
CRISP-DM

- Cross-Industry Standard Process for Data Mining (CRISP-DM);
- It is a well-established methodology for data mining projects;
- It provides a structured approach to guide organizations through the data mining process, from understanding the business problem to deploying the solution;
- The framework consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Why is CRISP-DM important?

- Minimization of the risk of failure
- Maximizing the changes of success

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Alternatives to CRISP-DM

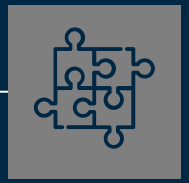
- Six Sigma: focus on improving the quality of processes;
- Lean: continuous improvement (value stream mapping, kanban, 5s);
- Agile: focus on delivery and adaptability of customer's requirements;
- Knowledge Discovery in Databases (KDD): focus on discovering patterns in order to make further decisions.

Data Mining Techniques

- Classification: Categorizing data into predefined classes or categories based on known attributes or features.
- Clustering: Grouping data points based on similarities in their attributes or features.
- Association rule mining: Finding patterns or relationships between different items in a dataset, such as in market basket analysis. E.g., Apriori
- Regression: Predicting a continuous output variable based on input variables.
- Anomaly detection: Identifying unusual or unexpected observations in a dataset that may indicate a problem. E.g., clustering-based anomaly detection
- Text mining: Extracting useful information from unstructured text data, such as social media posts or customer reviews. E.g., sentimental analysis



Table of Contents



01

INTRODUCTION TO DATA MINING AND CRISP-DM

What is data mining?
Benefits of CRISP-DM

02

CRISP-DM FRAMEWORK

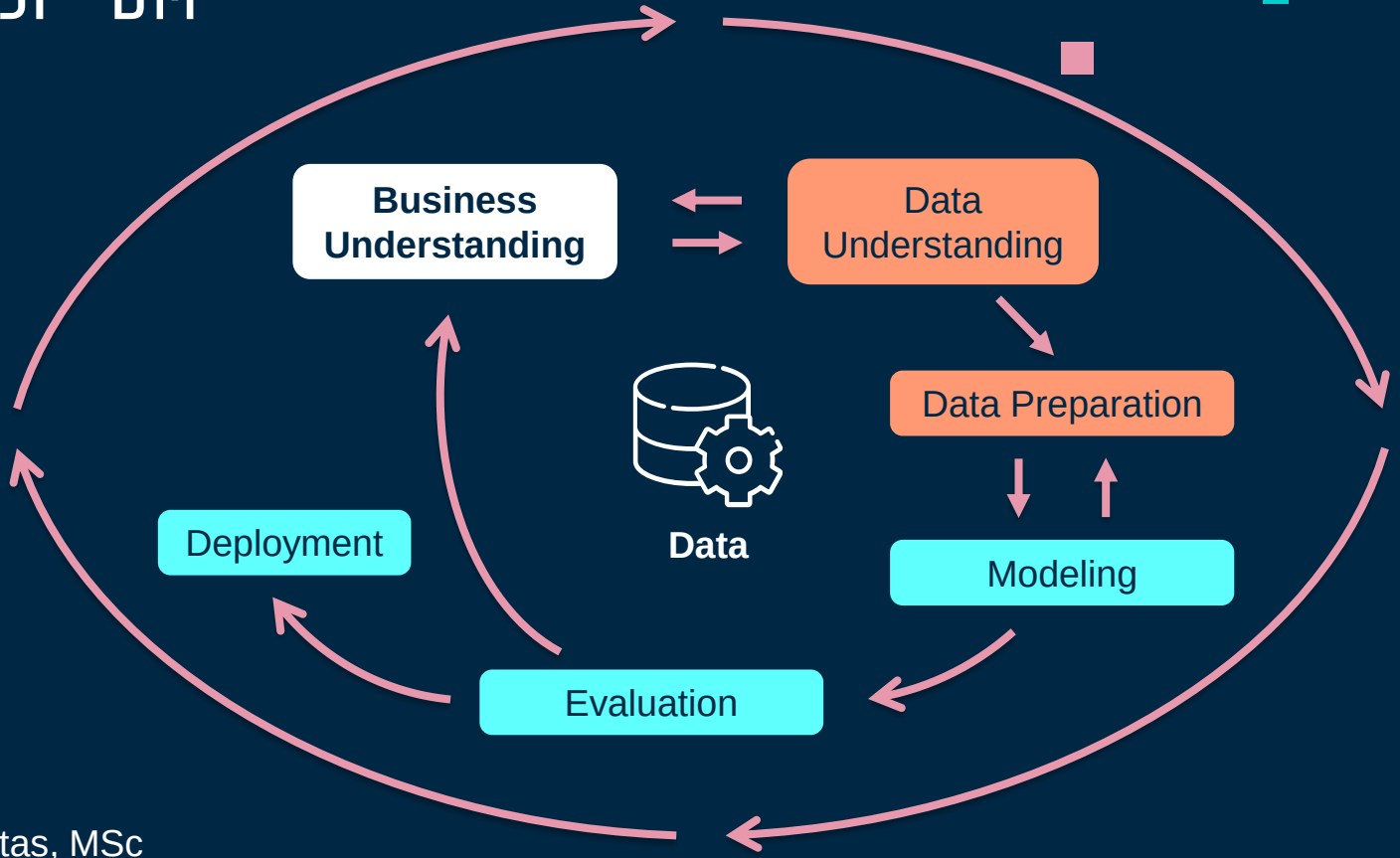
What are the different phases of CRISP-DM?

03

PYTHON LIBRARIES AND CASE STUDIES

Some practical applications of CRISP-DM

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

1st Phase: Business Understanding

- It is crucial to understand the business objectives and goals of the project to ensure that the data mining results and relevant and actionable for the organization;
- It consists of four tasks:
  - Determine Business Objectives
  - Assess Situation
  - Determine Data Mining Goals
  - Produce Project Plan

1st Phase: Business Understanding

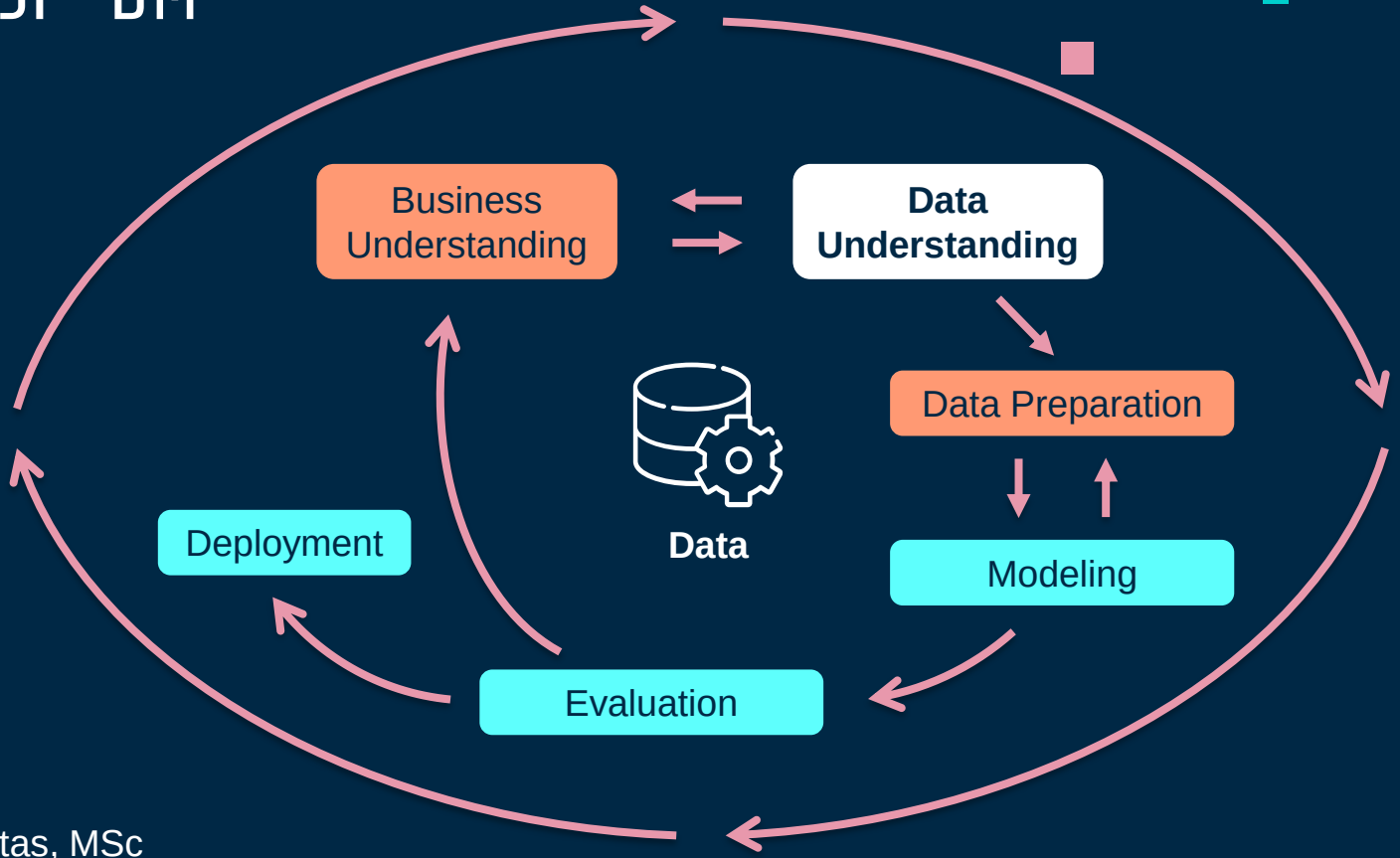
Why is Business Understanding a phase of CRISP-DM?

- Data miners can develop a more effective solution that addresses the specific needs of the business;
- To ensure that the project is successful and that the insights generated from the data mining process are relevant and actionable.

Questions related to:

- Business goals: "Increase catalog sales to existing customers"
- Data mining goals: "Predict how many widgets a customer will buy, given their last purchases, demographic information, and prices"

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

2nd Phase: Data Understanding

- It is focused on getting to know the data that will be used in the project;
- Here, a context is given to the data in order to accurately develop a solution.
- It consists of four tasks:
  - Collect Initial Data
  - Describe Data
  - Explore Data
  - Verify Data Quality

2nd Phase: Data Understanding

Why is Business Understanding a phase of CRISP-DM?

- Data miners can identify issues and limitations with the data, that may impact the quality of the results.

Questions related to data quality verification:

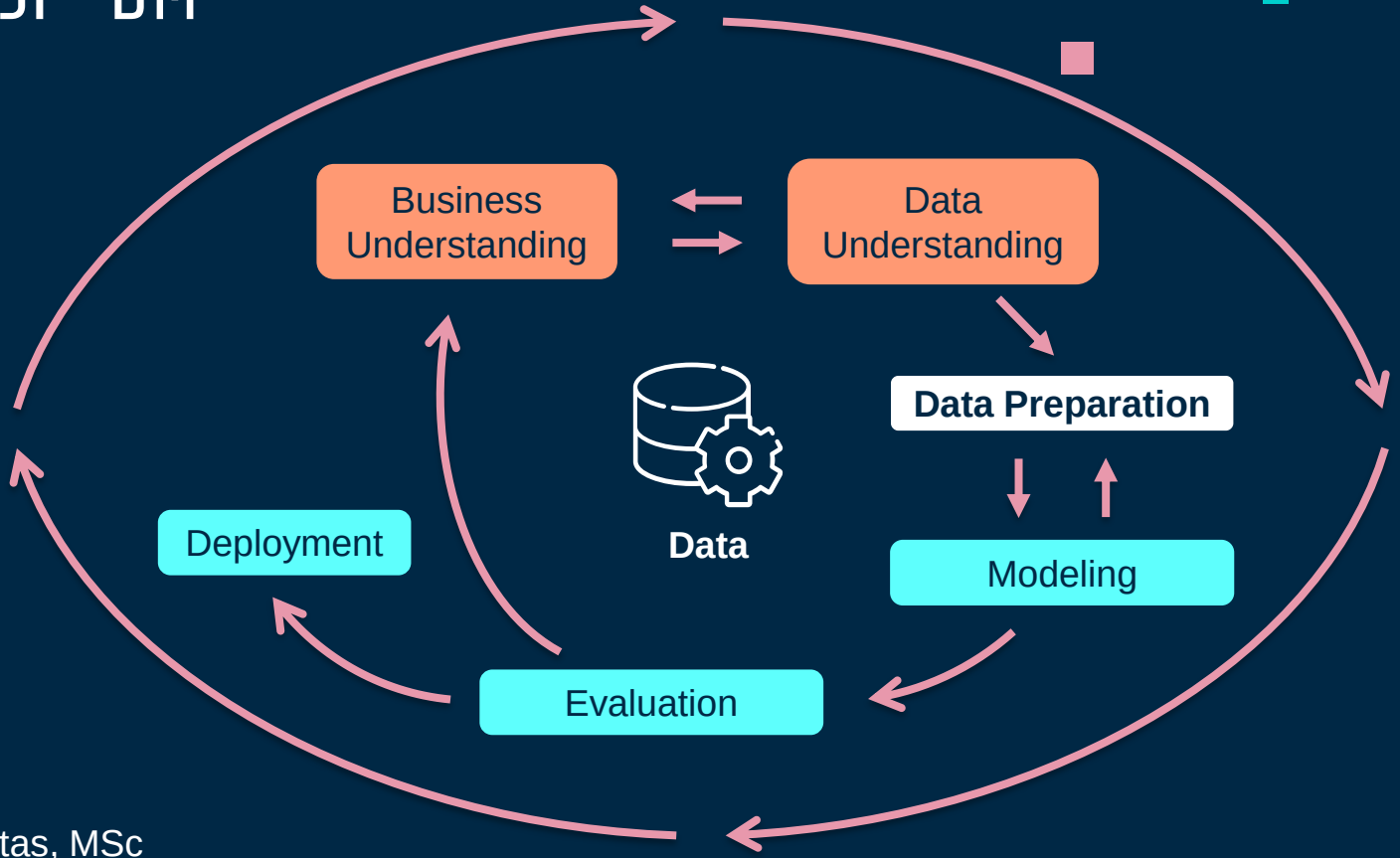
- "Is the data complete"?
- "Is it correct or does it contain errors"?
- "Are there missing values"?

2nd Phase: Data Understanding

Business Understanding ⇄ **Data Understanding**

- The data understanding phase can also lead to a deeper understanding of the business and its processes;
- By exploring the data, analysts may uncover previously unknown relationships or patterns that can inform and guide business decisions;
- E.g.: data analysis reveal that certain products are selling better in certain geographic regions or during certain times of the year, which can help a business to refine its marketing and sales strategies.

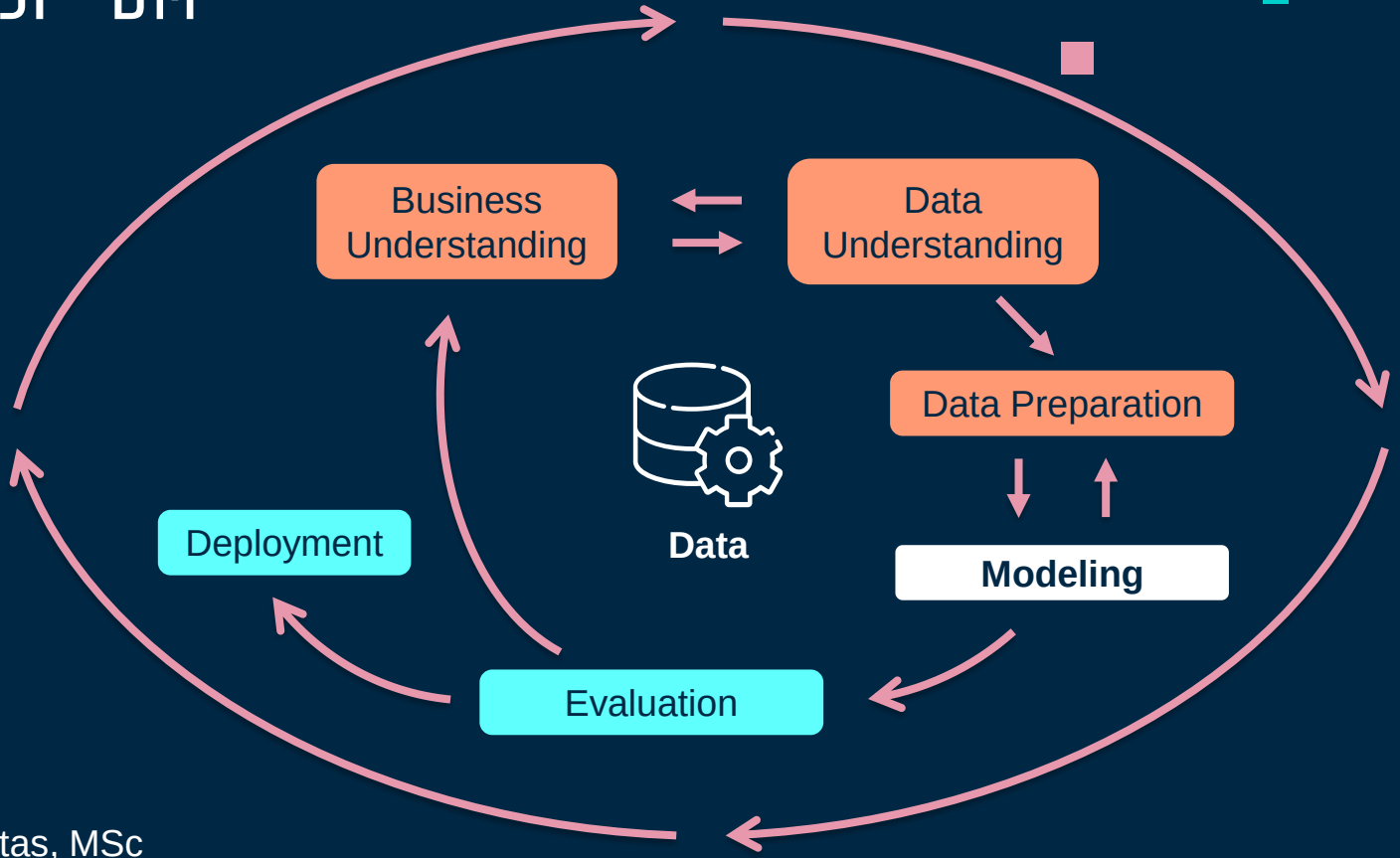
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

3rd Phase: Data Preparation

- Primary goal: to ensure that the data is in a suitable format and quality for effective analysis;
- Data preparation allows data analysts to identify and correct errors, inconsistencies, and missing values in the dataset and can ensure that the resulting models are accurate, reliable, and effective.
- It consists of five tasks:
  - Select Data
  - Clean Data
  - Construct Data
  - Integrate Data
  - Format Data

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

4th Phase: Modeling

- Focused on developing and building predictive models;
- Is the heart of the data mining process;
- Patterns and insights are discovered.
- It consists of four tasks:
  - Select Modeling Technique
  - Generate Test Design
  - Build Model
  - Assess Model

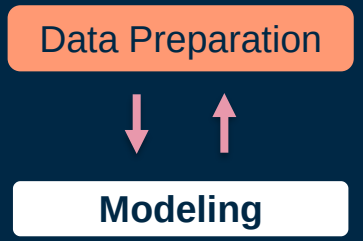
4th Phase: Modeling

Incorrect data:

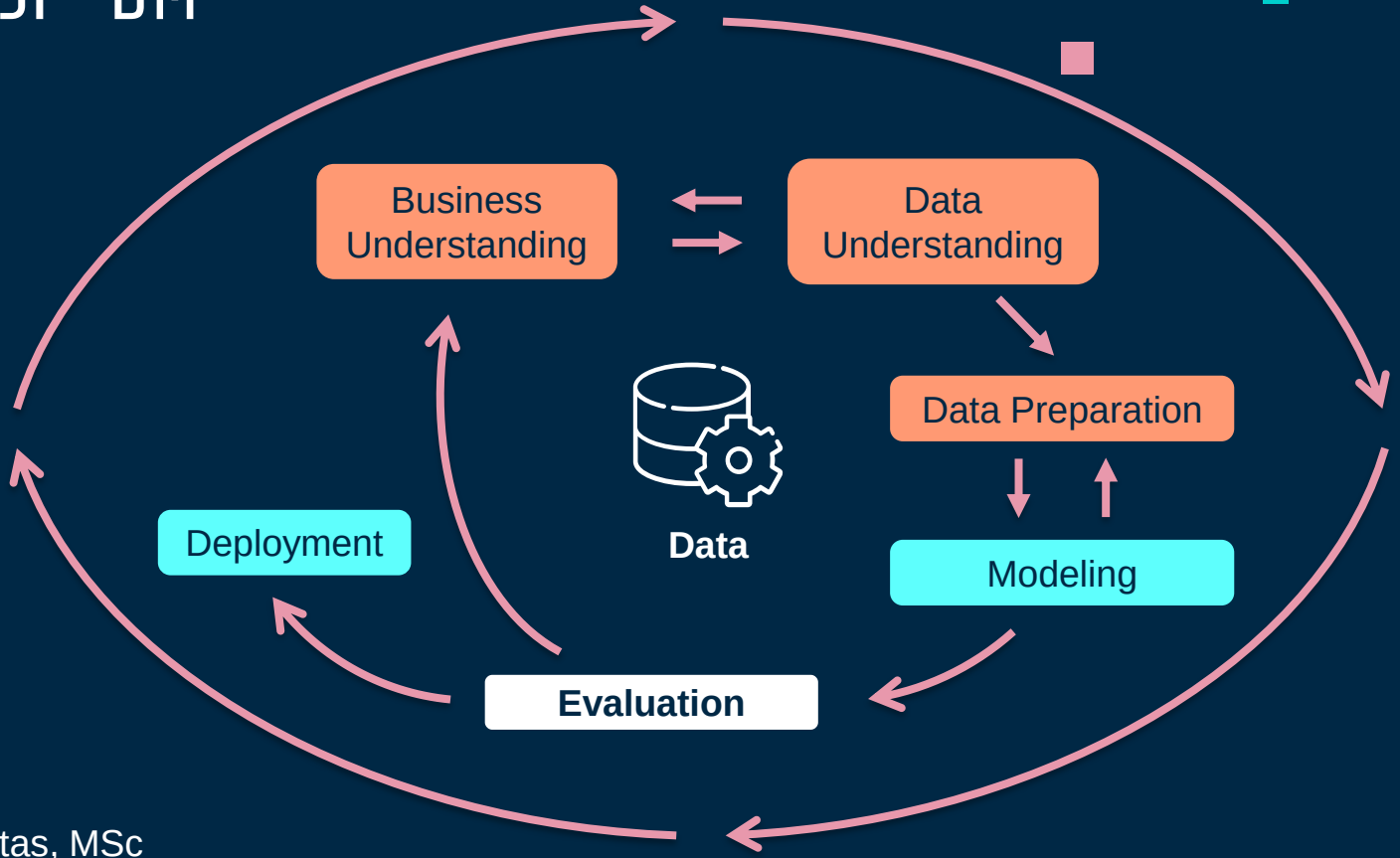
- It is discovered that the dataset contains incomplete or incorrect data that was missed during the data preparation phase;
- It may be necessary to go back to the data preparation phase to correct the errors or fill in the missing data;

Performance issues:

- If the performance is not satisfactory or does not meet the desired accuracy, it may be necessary to go back to the data preparation phase to identify the issues and prepare the data more effectively for modeling.



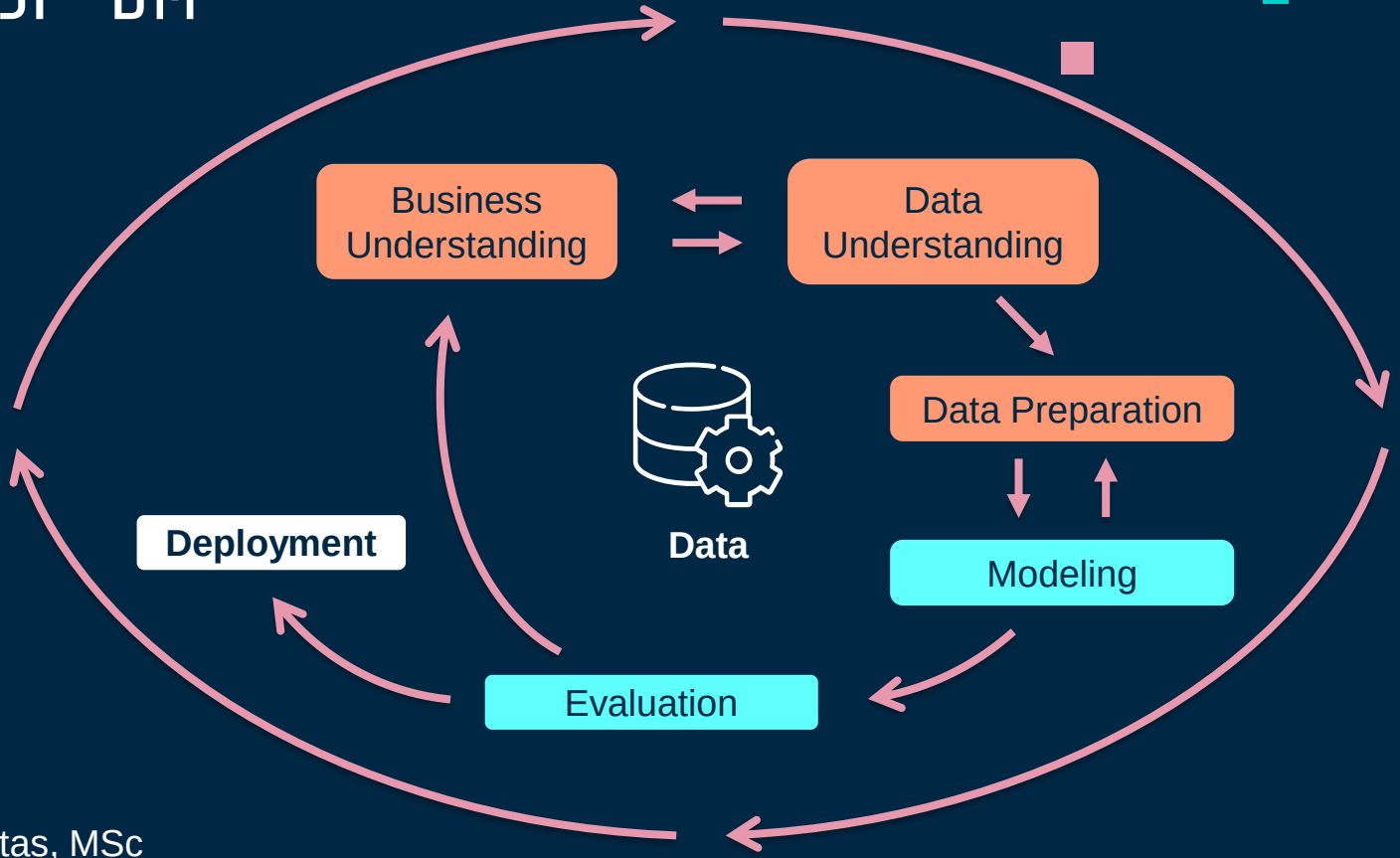
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

5th Phase: Evaluation

- After the predictive model being built, it is crucial to assess if the solution found is useful for the organization;
- It consists of three tasks:
  - Evaluate Results
  - Review Process
  - Determine Next Steps
- If the model meets the business objectives and performs well on the testing dataset, it may be ready for deployment;
- This is a Business evaluation, not a Model evaluation!

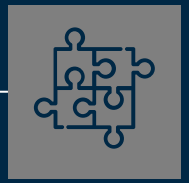
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

6th Phase: Deployment

- Is the final phase in the data mining process, where the predictive models are put into use and the organization can start reaping the benefits of the data mining project;
- It consists of four tasks:
  - Plan Deployment
  - Plan Monitoring & Maintenance
  - Produce Final Report
  - Review Project

Table of Contents



01

INTRODUCTION TO DATA MINING AND CRISP-DM

What is data mining?
Benefits of CRISP-DM

02

CRISP-DM FRAMEWORK

What are the different phases of CRISP-DM?

03

PYTHON LIBRARIES AND CASE STUDIES

Some practical applications of CRISP-DM

Python Libraries

- Business understanding: no python libraries, but jupyter notebook, excel and power-point are mainly used for the discussion of the requirements.
- Data understanding: pandas, seaborn/matplotlib, numpy, pandas-profiling, missingno, scipy;
- Data preparation: pandas, scikit-learn, numpy, category_encoders, imbalanced-learn, deature-engine, pandasql;
- Modeling: scikit-learn, keras/tensorflow, xgboost/lightgbm/catboos, statsmodels, yellowbrick, shap, eli5, seaborn;
- Evaluation: numpy, matplotlib, pandas, scipy;
- Deployment: Django, AWS, dash.

Python - Example

```
# Import necessary libraries
import pandas as pd; import numpy as np;
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

# Load and clean the data
data = pd.read_csv("data.csv"); data = data.dropna()

# Explore the data
# ... (insert exploratory data analysis code here)
```

Python - Example

```
# Prepare the data
X = data[["feature_1", "feature_2", "feature_3"]]; y = data["target"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Model the data
model = LinearRegression(); model.fit(X_train, y_train)

# Evaluate the model
y_pred = model.predict(X_test); mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse); r2 = r2_score(y_test, y_pred)

# Visualize the results
plt.scatter(y_test, y_pred)
plt.xlabel("True Values"); plt.ylabel("Predictions");
plt.show()
```

Cases of success

Amazon: optimization of recommendations. They report that the algorithm's recommendations have led to a 29% increase in revenue. [1]

T-Mobile: development a predictive model to identify potential customer churn. They report that this model has helped them reduce churn by 50%. [2]

IBM: improvement of their lead generation process. They report that the new process led to a 400% increase in lead volume. [3]

Royal Bank of Scotland: development of a model to detect fraudulent transactions. They report that the model helped them reduce fraud by 60%. [4]

Ford: optimization of their supply chain. They report that this optimization has led to a 40% reduction in inventory levels. [5]

Case Study

A retail company wants to increase sales of its new product line of organic foods. The company has invested significant resources in developing it and wants to generate more revenue. The company has noticed that sales of organic food products have been lower than expected; they want to understand why.

- The company's revenue in the past year was $100 million.
- The organic food product line represents 15% of the company's total revenue.
- The organic food product line was launched six months ago.
- Sales of the organic food product line have been lower than expected.
- The company wants to understand why organic food product line sales have been lower than expected and how they can increase sales.



Case Study

Business Understanding:

- Goal: to increase sales of the new product line of organic foods; to generate more revenue from the organic food product line.
- Stakeholders: company's management, investors, and customers.
- Problem: "sales of the organic food products have been lower than expected"
- Factors that may be contributing to the problem: could be a lack of awareness about the product line, poor marketing, competition from other brands, or quality issues with the products.
- What is needed: to identify the key factors that are contributing to lower sales and develop a strategy to address these factors.



Case Study

Data Understanding:

- Gather data on sales: The company should gather data on sales of the organic food product line over the past six months. Ex. of variables: Date of sale, sales revenue and volume, customer demographics, product details, marketing campaigns. Data from other companies can be relevant to gather.
- Data description: The data is presented in a tabular format and includes 15 entries. The table contains six variables: date of sale, sales revenue, sales volume, customer demographics, product details, and marketing campaigns. The table also includes two issues: an incorrect revenue recorded on January 1st and an incorrect marketing campaign recorded on January 14th.

Case Study

Data Understanding:

- SKU901 (Organic Oranges) has the highest sales revenue and volume ($1,700 and 170 units, respectively), while SKU567 (Organic Carrots) has the lowest sales revenue and volume ($700 and 70 units, respectively).
- Influencer marketing generated the highest sales revenue and volume for two SKUs (SKU789 and SKU901) while social media ads generated the lowest sales revenue and volume for two SKUs (SKU456 and SKU567).
- Highest revenue: Female, San Francisco, 35-44 age group; Lowest revenue: Female, Miami, 35-44 age group

Case Study

Data Understanding (business understanding context):

- There are only 15 data entries, more data is needed;
- There is missing data from the rest of the days (total were six months);
- The table is incomplete since it does not include information on returns, discounts, etc.

Case Study

Data Preparation:

- Handling with not fully-correct entries: deletion, imputation
- Imputation methods: full-data average value, average value after filtering
- First sales revenue entry: changed to 1000$ due to correlation with sales volume
- Split customer demographics into gender, age range and location
- Split product details into SKU and Product Category
- Other transformations: categorization, deletion of “units” and “$"

Case Study

Modeling:

- Input labels: Gender, Age Range, Location, Product Category, Marketing Campaign
- Output labels: Sales revenue / sales volume (r = 1)
- Technique: regression analysis - Linear Regression
- Assessment criteria: R^2

Case Study

Evaluation:

- San Francisco is the region where it potentially leads to higher sales revenue;
- Potential buyers are 25 to 34 years old;
- Assuming that the predicted sales are an estimation for the average, for each combination, the average of sales revenue is equal to 1280$, whereas originally the average was 1150$;
- It would lead to an increase of 12% on sales of the production line, or an increase of at least 1% on total sales;
- With discounts: a regression model is more likely to be used; evaluation shall consider other variables such as the real profit.

Case Study

Deployment:

- The company shall focus their production on their top selling products: SKU345, 901, 890, and 678 – but continuing exploring the other products to collect more data or implementing the production of other products.

References

[1] Amazon: "Amazon.com Recommendations: Item-to-Item Collaborative Filtering", Greg Linden, Brent Smith, and Jeremy York (2003).

[2] T-Mobile: "Predictive analytics for customer churn from traditional to big data: A survey", Mohammad Alsheikh and Reda Alhajj (2016).

[3] IBM: "IBM's New Approach to Data-Driven Marketing: Agile Marketing Hub", IBM (2016).

[4] Royal Bank of Scotland: "CRISP-DM Application for Fraud Detection", F. De Luca et al. (2010).

[5] Ford: "Optimizing Supply Chain Inventory Management", Ford (2014).

Thanks!

Do you have any questions?

rfitas99@gmail.com
+351 925076673

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik