Introduction to Data Analytics

Data Science Methodology

Learning Objectives

By the end of this lesson, you will be able to:

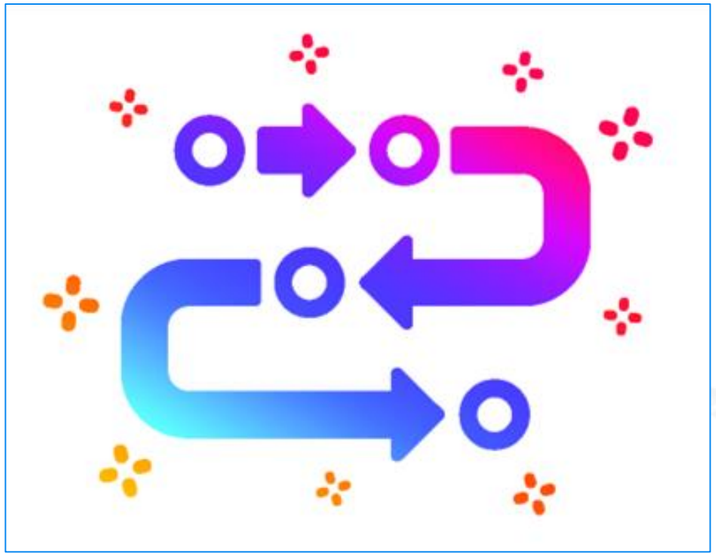
- Explain data science methodology
- Describe the various stages of data science methodology

DATA AND
ARTIFICIAL INTELLIGENCE

Data Science Methodology: Overview

Data Science Methodology

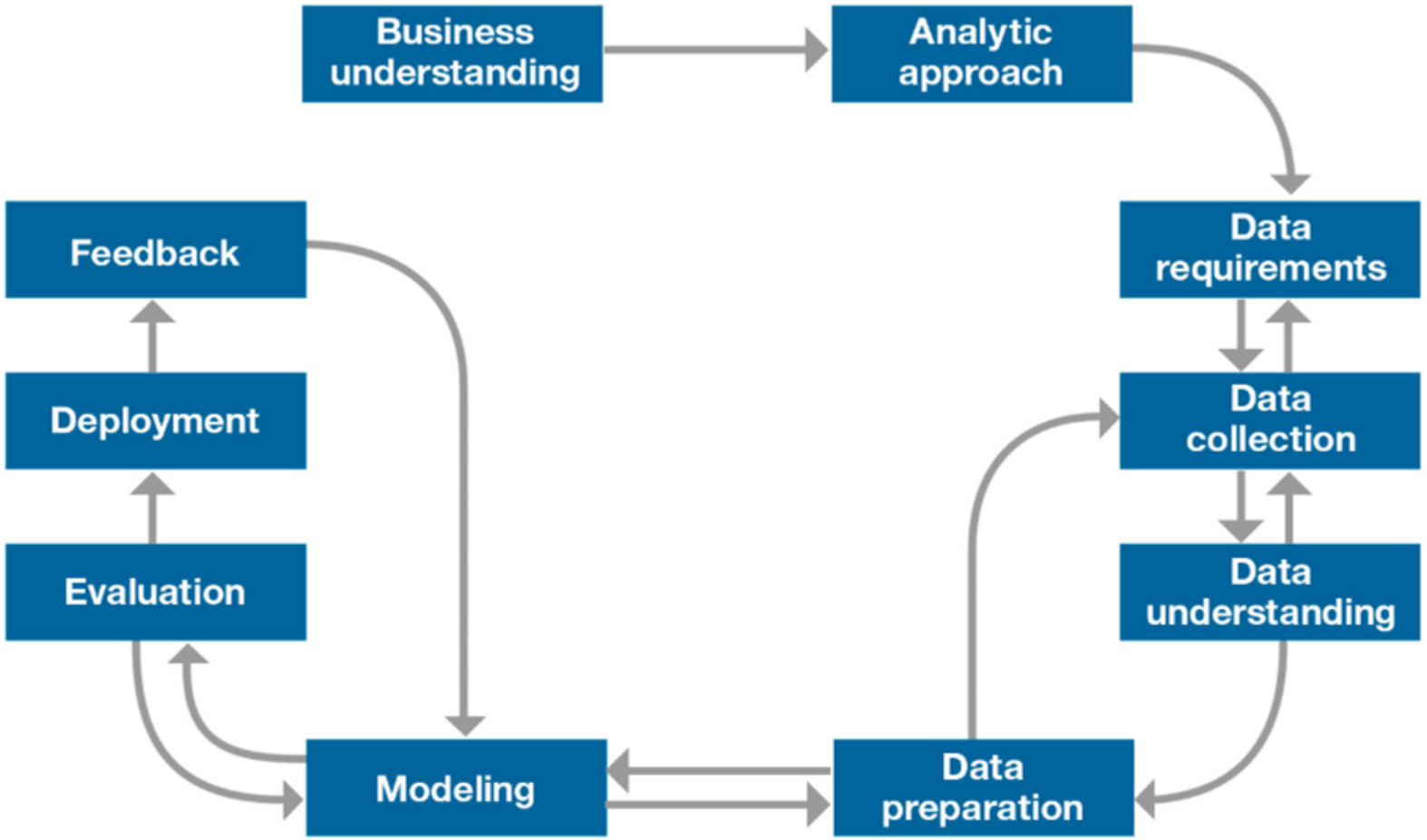
- A methodology is a process with a defined input to achieve a defined output.
- It drives activities within a given domain and does not depend on technologies or tools.
- Data science methodology is an iterative methodology leveraged to produce repeatable and successful results.



DATA AND
ARTIFICIAL INTELLIGENCE

Stages of Data Science Methodology

Stages of Data Science Methodology



Source: https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

Stages of Data Science Methodology

Business understanding is the first stage of the data science methodology and lays the foundation for a successful end result.

- This stage identifies key business sponsors, steering committee, and internal sponsors.
- It helps understand business and customer needs and identify who needs the analytical solution.
- It includes defining the problem, project objectives, and solution requirements from a business perspective.

Stages of Data Science Methodology

- The analytic approach determines business requirements as well as data requirements.
- It identifies the analytic methods, hardware and software, data content, formats, and representations to be used.

Analytic Approach

Stages of Data Science Methodology

- The requirement stage is specific to identifying necessary data with its initial source and appropriate format.

- This stage has multiple sub-stages including data acquisition, data wrangling, data analysis and data modeling.

Data Requirements

Stages of Data Science Methodology

- In collection stage, data scientists identify and gather the available relevant data as a good quality input data is required for a great output.

- Data scientists evaluate the volume and properties of the data and understand the distribution of each attribute.

- High-performance platforms and in-database analytic functionality enable data scientists to use large data sets.

Data Collection

Stages of Data Science Methodology

Data scientists use descriptive statistics and visualization techniques to:

- Understand data content
- Assess data quality
- Discover initial insights about the data

Data Understanding

Stages of Data Science Methodology

- The data preparation stage includes activities to construct a data set for data modeling.

- This stage includes cleaning of data, eliminating duplicates, formatting data from multiple sources, and transforming data into more useful variables.

- Data scientists are capable of creating explanatory variables through a combination of domain knowledge and existing structured variables.

Stages of Data Science Methodology

- The modeling stage applies predictive model on historical data to obtain the outcome.
- This stage helps organizations gain intermediate insights and future trends, leading to strategic improvements.
- Using exploratory data analytics, data scientists attempt multiple algorithms to find the best model for the available data set.

Stages of Data Science Methodology

- Once the model is developed, data scientists evaluate the model to understand its quality and ensure that it addresses the business problem.
- In model evaluation, diagnostic measures are computed and outputs such as tables and graphs are evaluated.
- During the evaluation phase, data mining result is evaluated for novelty and usefulness.

Evaluation

Stages of Data Science Methodology

Review the whole evaluation process with the following steps:

- Summarize activities that are missed
- Ensure that the model is correctly built
- Identify failures and misleading steps
- Determine the plan of action based on findings
- Analyze and estimate the potential for improvement

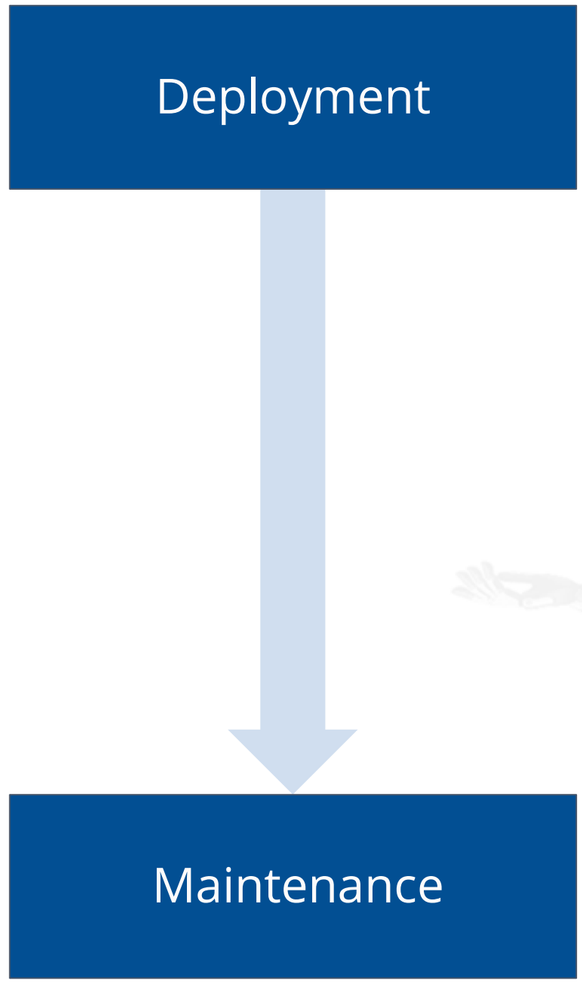
Stages of Data Science Methodology

- In the deployment stage, a satisfactory model should be deployed into the production environment.
- It involves multiple groups, skills, and technologies.
- It requires planning on how knowledge can be propagated to users.

Stages of Data Science Methodology

In this maintenance phase, identify:

- What could change in the environment?
- How will the accuracy be monitored?
- When should the data mining model not be used?
- Will business objectives change over time?
- What kind of report is required?
- Were initial data mining goals met?
- Who will be target groups for reports?



Stages of Data Science Methodology

In this last stage of feedback, review the whole framework by:

- Interviewing people involved in the project
- Interviewing end users and identifying improvement areas
- Summarizing the feedback and documenting the experience

Key Takeaways

- Data science methodology is the process that drives activities within a given domain.

- The different stages of methodology: *Business understanding, analytical approach, data requirements, data collection, data understanding, data preparation, modeling, evaluation, deployment,* and *feedback*.