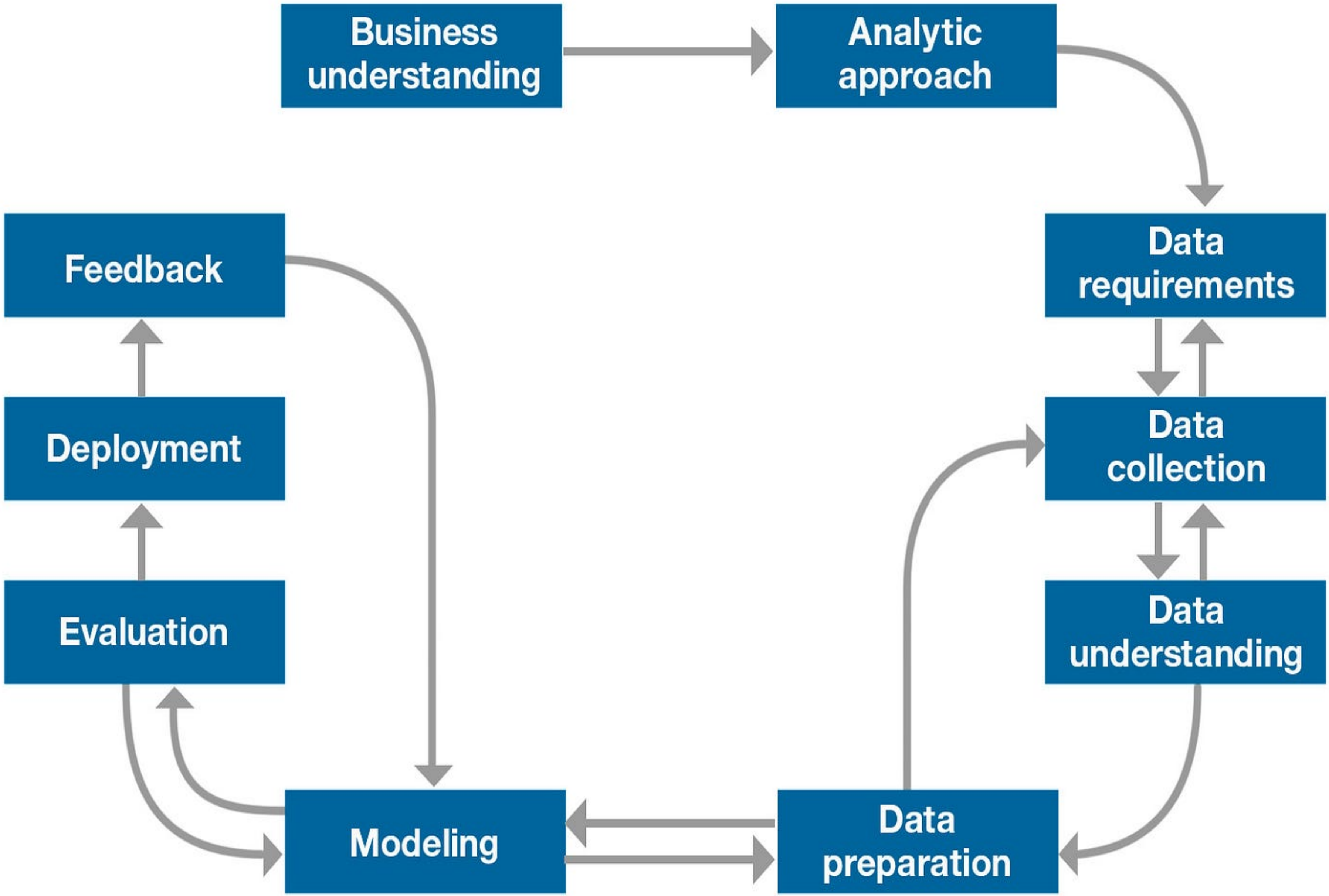
# K. BHASKAR

TOPIC : DATA SCIENCE METHODOLOGY

PIN . NO : 21AK5A0305

# What is Data Science Methodology ?

- A Data Science methodology is a **structured approach to problem-solving using data.**
- It involves understanding the business problem, collecting and preparing the data, modeling and evaluating the data, deploying the model, and monitoring and maintaining its performance.
- A data science life cycle (also known as a data science methodology) describes the step-by-step approach you take to deliver a project.



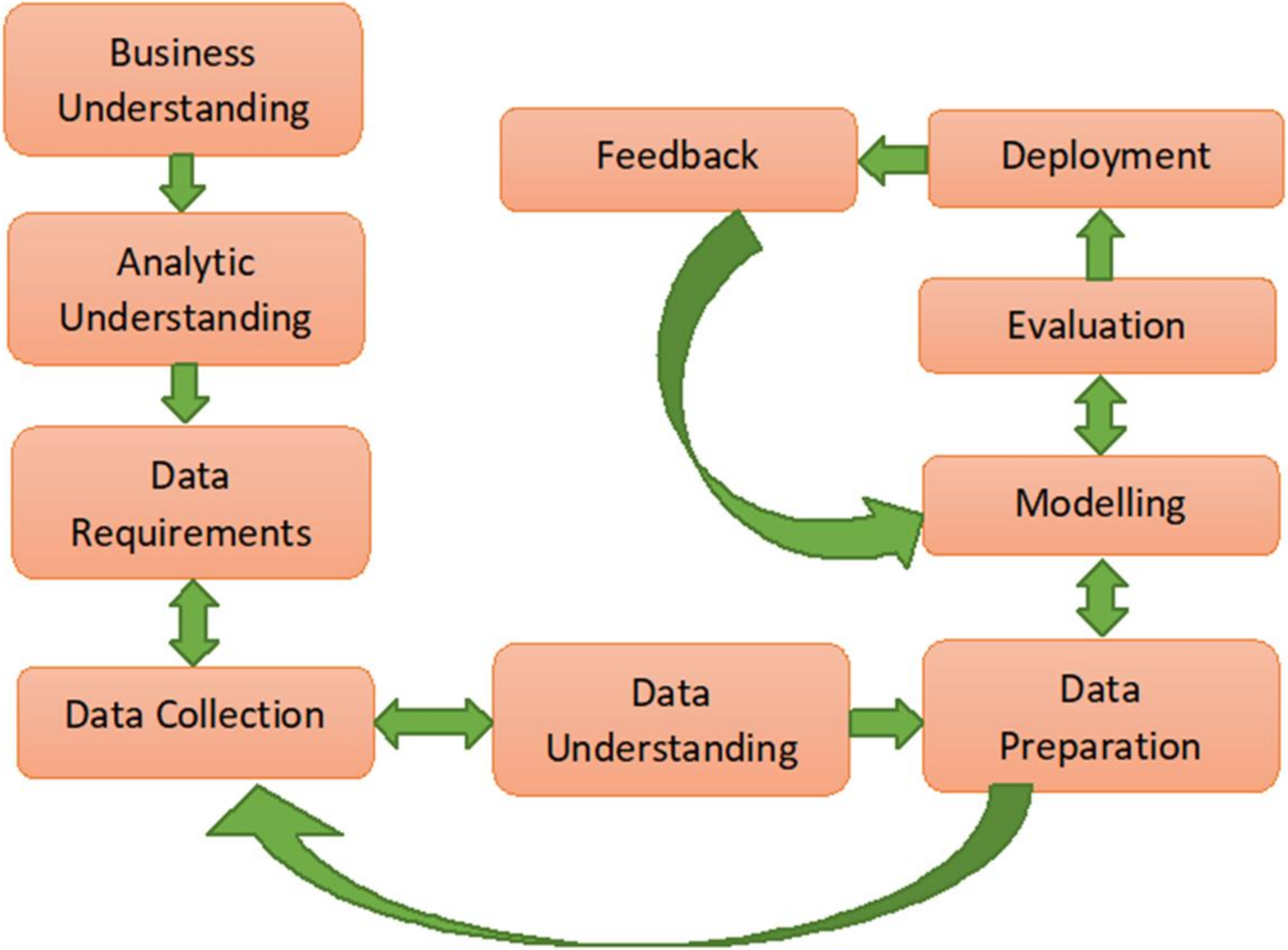
Business understanding
Analytic approach
Data requirements
Data collection
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Feedback

- Data scientists intuitively understand these steps, but documenting them can help increase repeatability and prevent forgetting a step.
- Data Science Methodology indicates the routine for finding solutions to a specific problem.
- This is a cyclic process that undergoes a critic behaviour guiding business analysts and data scientists to act accordingly.
- Data Science methodology is a structured approach to solving complex problems using data.

# *The typical steps involved in a Data Science methodology:*

- Data Preparation: In this stage, the data science team cleans, transforms, and prepares the data for analysis. This step is critical to ensure that the data is in the right format and quality for analysis.
- Data Understanding: In this stage, the data science team identifies and collects the data required for the analysis. They explore the data to understand its structure, quality, and completeness.
- Business Understanding:In this stage, the business problem is defined, and the objective of the analysis is identified. The data science team should work closely with the business stakeholders to understand the problem and define the goals.
- Data Modeling: In this stage, the data science team selects the appropriate modeling techniques to analyze the data and build predictive models. This stage involves selecting the right algorithms, tuning the model parameters, and validating the model.

- Evaluation: In this stage, the data science team evaluates the model’s performance and its ability to solve the business problem. They use various evaluation metrics to determine the effectiveness of the model and make improvements if necessary.
- Deployment: In this stage, the data science team deploys the model in the production environment, integrating it into the business processes, and ensuring that it is working correctly.
- Monitoring and Maintenance: In this stage, the data science team monitors the model’s performance in the production environment, making necessary changes and improvements to ensure that it continues to work effectively.


Business Understanding
Analytic Understanding
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment
Feedback

- **Business Understanding:**
  Before solving any problem in the Business domain it needs to be understood properly. Business understanding forms a concrete base, which further leads to easy resolution of queries. We should have the clarity of what is the exact problem we are going to solve.
- **Analytic Understanding:**
  Based on the above business understanding one should decide the analytical approach to follow. The approaches can be of 4 types: Descriptive approach (current status and information provided), Diagnostic approach(a.k.a statistical analysis, what is happening and why it is happening), Predictive approach(it forecasts on the trends or future events probability) and Prescriptive approach( how the problem should be solved actually).
- **Data Requirements:**
  The above chosen analytical method indicates the necessary data content, formats and sources to be gathered. During the process of data requirements, one should find the answers for questions like ‘what’, ‘where’, ‘when’, ‘why’, ‘how’ & ‘who’.

- **Data Collection:**
  Data collected can be obtained in any random format. So, according to the approach chosen and the output to be obtained, the data collected should be validated. Thus, if required one can gather more data or discard the irrelevant data.
- **Data Understanding:**
  Data understanding answers the question "Is the data collected representative of the problem to be solved?". Descriptive statistics calculates the measures applied over data to access the content and quality of matter. This step may lead to reverting the back to the previous step for correction.
- **Data Preparation:**
  Let's understand this by connecting this concept with two analogies. One is to wash freshly picked vegetables and second is only taking the wanted items to eat in the plate during the buffet. Washing of vegetables indicates the removal of dirt i.e. unwanted materials from the data.

- **Modelling:**

  Modelling decides whether the data prepared for processing is appropriate or requires more finishing and seasoning. This phase focuses on the building of predictive/descriptive models.

- **Evaluation:**

  Model evaluation is done during model development. It checks for the quality of the model to be assessed and also if it meets the business requirements. It undergoes diagnostic measure phase (the model works as intended and where are modifications required) and statistical significance testing phase (ensures about proper data handling and interpretation).

- **Deployment:**

  As the model is effectively evaluated it is made ready for deployment in the business market. Deployment phase checks how much the model can withstand in the external environment and perform superiorly as compared to others.

# *Feed Back :*

Feedback is the necessary purpose which helps in refining the model and accessing its performance and impact. Steps involved in feedback define the review process, track the record, measure effectiveness and review with refining .

# *SUMMARY :*

After successful abatement of these 10 steps, the model should not be left untreated, rather based on the feedbacks and deployment appropriate update should be made. As new technologies emerge, new trends should be reviewed so that the model continually provides value to solutions.

*THANK YOU*