# Introduction to Data Science

# Data Science Process

Different methodologies used in data science

- CRISP-DM Methodology,
- SEMMA,
- BIG DATA LIFE CYCLE,
- SMAM.

# CRISP-DM: Methodology

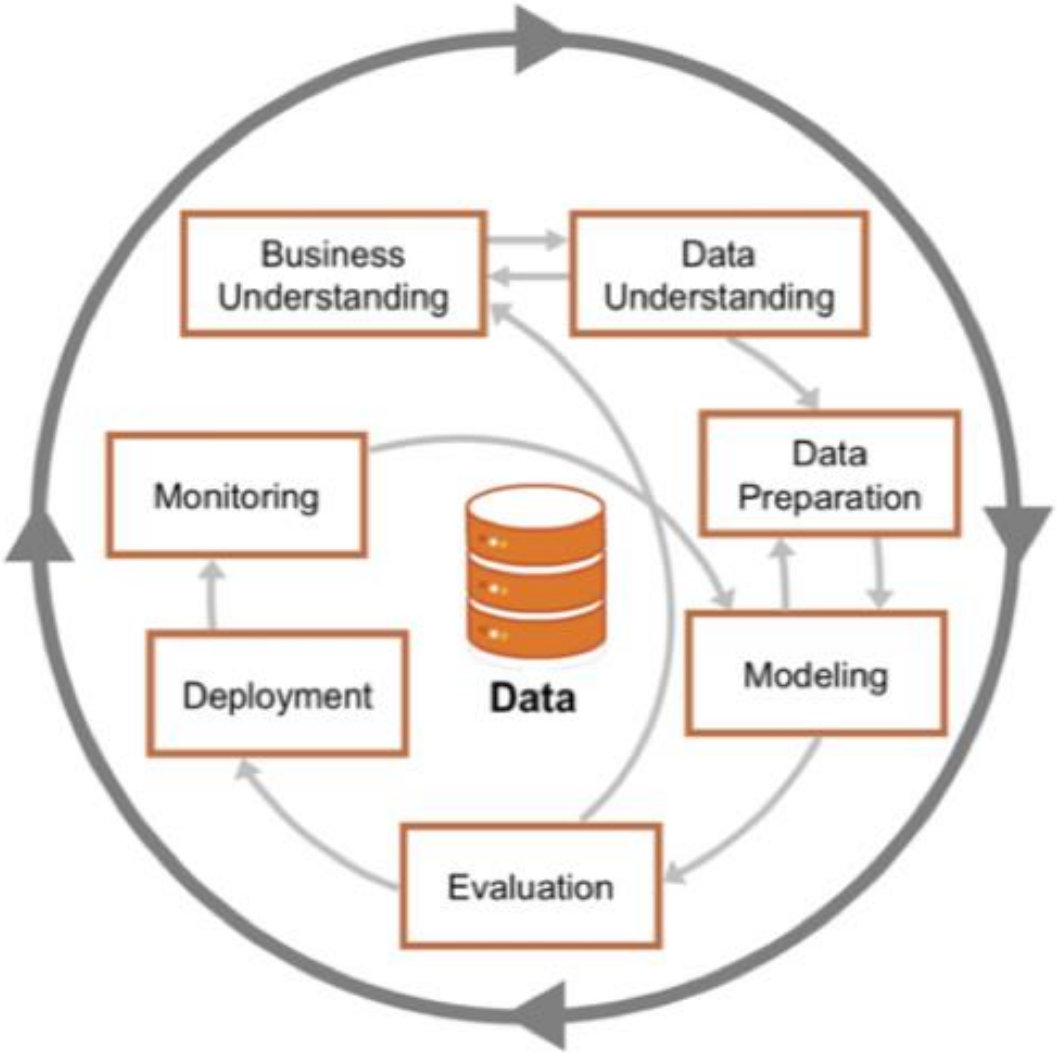
CRISP-DM stands for Cross-Industry Standard Process for Data Mining.

- Widely adopted methodology
- Provides a structured approach for planning & executing DM projects.
- Designed to be adaptable across various industries and applications.
- Key Characteristics of CRISP-DM
  - **Iterative:** The process is not strictly linear. You may need to revisit previous phases as you progress.
  - **Flexible:** It can be adapted to various project sizes and complexities.
  - **Industry-Neutral:** Applicable across different domains and sectors.
  - **Focus on Business Value:** Emphasizes understanding business needs and aligning data mining efforts accordingly.

# CRISP-DM: Data Mining Operations

1. Business Understanding:
   1. Determine business objectives and requirements.
   2. Assess situation and resources.
   3. Determine data mining goals.
2. Data Understanding:
   1. Collect initial data.
   2. Describe data.
   3. Explore data.
   4. Verify data quality.
3. Data Preparation:
   1. Select and Clean data.
   2. Construct data.
   3. Integrate data.
   4. Format data.
4. Data Modeling:
   1. Select modeling techniques.
   2. Generate test design.
   3. Build and Assess models.
5. Evaluation:
   1. Evaluate results.
   2. Review process.
   3. Determine next steps.
6. Deployment:
   1. Plan deployment.
   2. Plan monitoring and maintenance.
   3. Produce final report.
   4. Review project.

# CRISP-DM: Methedology



# SEMMA

**SEMMA** is a data mining methodology developed by the SAS Institute. It outlines a five-step process for extracting meaningful insights from data:

**1. Sample:**

- **Select a representative subset of the data** for analysis. Necessary to manage the computational complexity of working with large datasets.
- Sampling techniques can include random sampling, stratified sampling, and cluster sampling.

**2. Explore:**

- **Conduct exploratory data analysis (EDA)** to understand the characteristics of the data.
- This involves visualizing the data, identifying patterns, and detecting anomalies.
- Common EDA techniques include histograms, scatter plots, box plots…

# SEMMA

**3. Modify:**

- **Transform and prepare the data for modeling.**
  - **Data cleaning:** Handling missing values, outliers, and inconsistencies.
  - **Feature engineering:** Creating new variables or transforming existing ones.
  - **Data transformation:** Scaling or normalizing data to improve model accuracy.

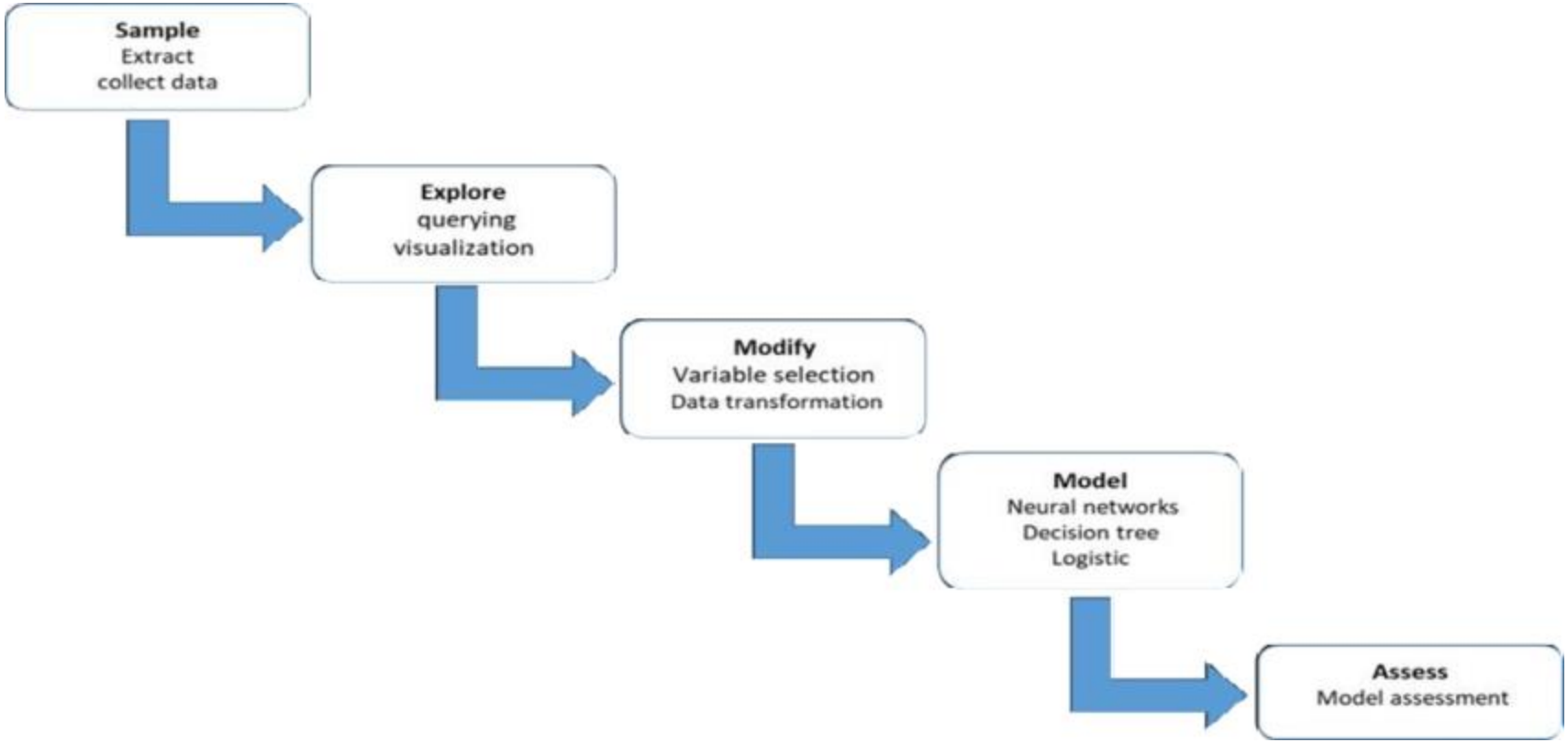
**4. Model:**

- **Build and train predictive models** using appropriate algorithms.
  - **Regression:** Predicting continuous values.
  - **Classification:** Predicting categorical values.
  - **Clustering:** Grouping similar data points together.

**5. Assess:**

- **Evaluate the performance of the models** using appropriate metrics.
- This helps to determine the accuracy, reliability, and generalizability of the models.
- Common evaluation metrics include accuracy, precision, recall, and F1-score.

# SEMMA


Sample
Extract
collect data
Explore
querying
visualization
Modify
Variable selection
Data transformation
Model
Neural networks
Decision tree
Logistic
Assess
Model assessment

# SMAM

- **SMAM** stands for **Sample, Mine, Assess, Maintain**. It's a simplified data science methodology, particularly useful for initial data exploration and analysis. Here's a breakdown:

**1. Sample:**

- Select a representative subset of the data.

**2. Mine:**

- Apply data mining techniques to discover patterns and relationships within the data.

**3. Assess:**

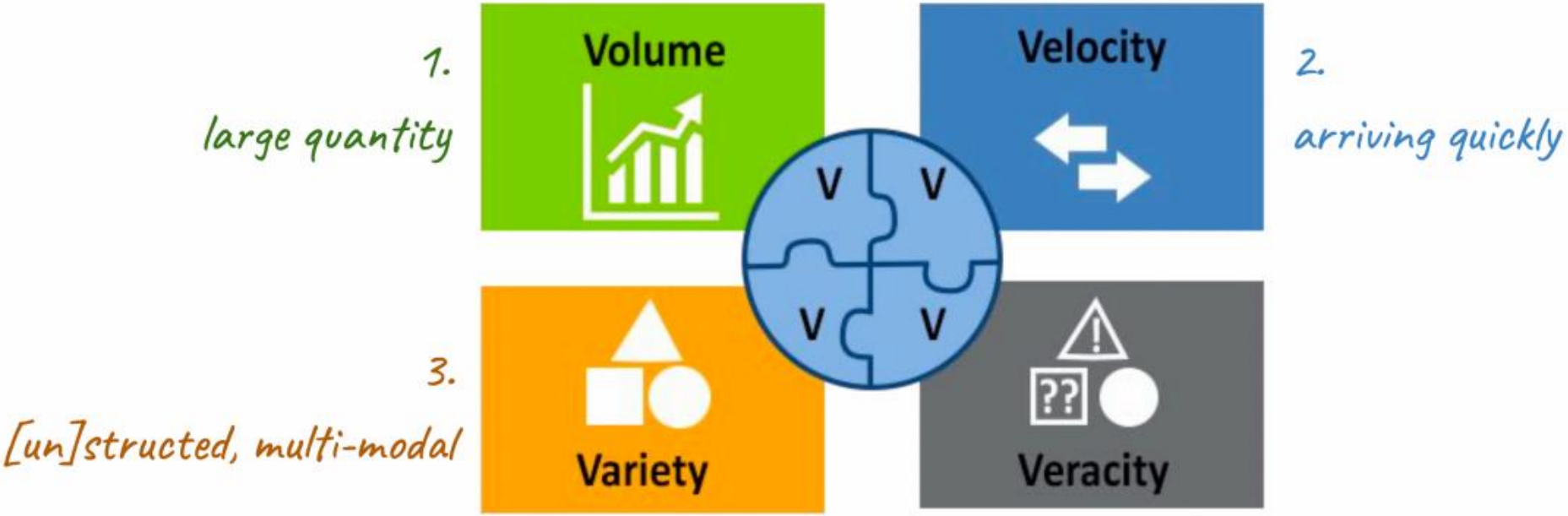
- Evaluate the findings and their implications.

**4. Maintain:**

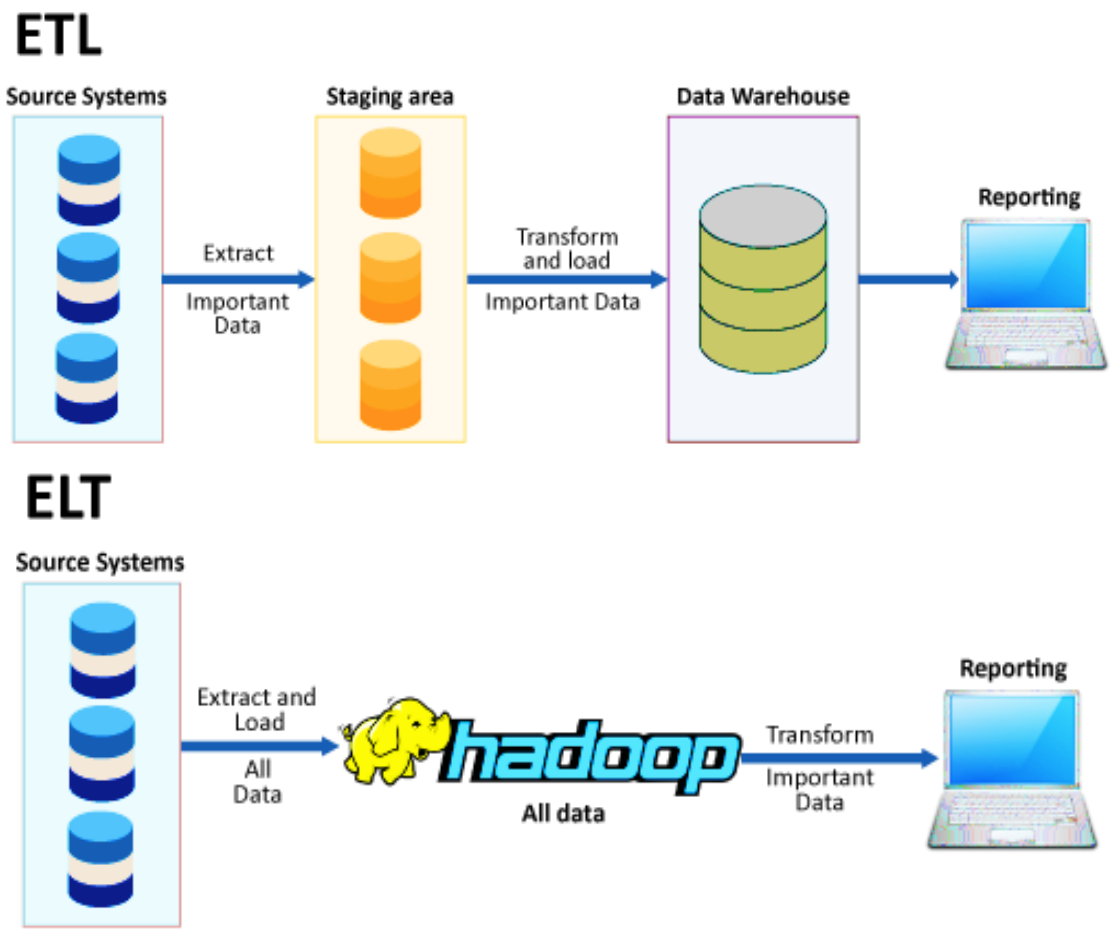
- **Update and refine the analysis as new data becomes available.**
- This may involve:
  - **Retraining models** with new data to improve their performance.
  - **Updating data sources** and re-running the analysis.
  - **Incorporating new insights** and adjusting the analysis accordingly.

# Big Data

ANALYSES WHICH CAN HANDLE THE 3 VS
AND DO IT WITH QUALITY (VERACITY)



# ETL vs ELT


ETL
Source Systems
Staging area
Data Warehouse
Reporting
Extract
Important
Data
Transform
and load
Important Data
ELT
Source Systems
Extract and
Load
All
Data
hadoop
All data
Transform
Important
Data
Reporting

# Big Data Life Cycle