# Introduction to Data Science

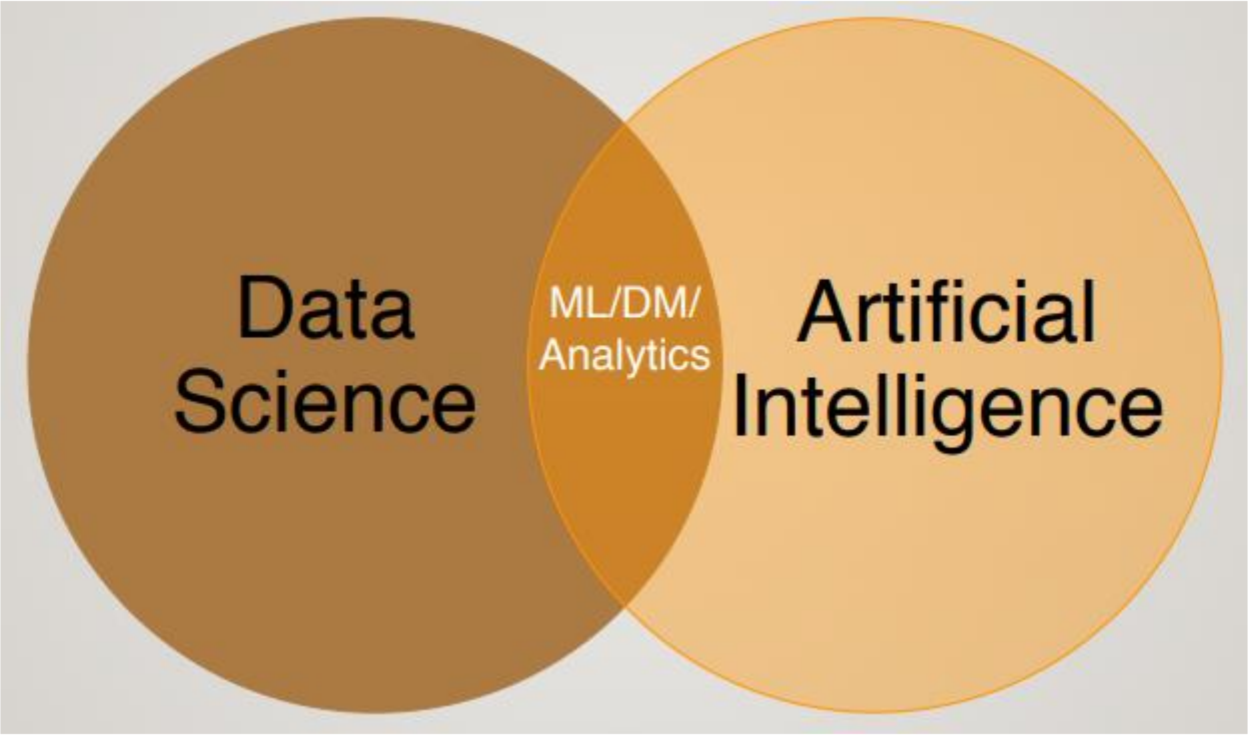
# Define

- “Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured.”
- It combines aspects of statistics, machine learning, and domain expertise to analyze data and make informed decisions.

# DATA SCIENCE AND BIG DATA

- They are not the “same thing”
- **Big data = crude oil**
- Big data is about extracting “crude oil”, transporting it in “mega tankers”, siphoning it through “pipelines”, and storing it in “massive silos”
- Data science is about refining the “crude oil”

# DATA SCIENCE AND ARTIFICIAL INTELLIGENCE



# Key Concepts

- **Data:** The raw material of data science. It can be structured (organized in tables or databases), semi-structured (like emails or social media posts), or unstructured (images, videos, audio).
- **Analysis:** The process of examining data to identify patterns, trends, and relationships. This can involve statistical methods, machine learning algorithms, and data visualization techniques.
- **Insights:** The valuable information extracted from data analysis. Insights can be used to make better decisions, improve products and services, and gain a competitive advantage.

# Key Concepts

The field of data science typically involves three key areas:

- **Data collection and processing:** This process, also referred to as data preparation, involves gathering data from various sources and cleaning it to ensure accuracy and reliability. The collected data may come from databases, spreadsheets, online sources, and other types of data storage systems.
- **Data analysis:** Data scientists use statistical methods and machine learning algorithms to explore and analyze the data. This step helps to identify patterns, correlations, trends, and other insights hidden in the data.
- **Data interpretation and communication:** Once the analysis is complete, data scientists interpret the results and communicate them to stakeholders in a way that’s easy to understand. This often involves creating visualizations, reports, and presentations.

# Importance of data science

In Various Industries

From healthcare to finance, data science is playing a  crucial role in various industries. By analyzing large  amounts of data, businesses can reduce costs, increase  efficiency, and improve customer satisfaction. In this  section, we will explore the benefits of data science in  different sectors.

# Benefits of data science for businesses

Better Decision Making

Data science helps businesses make better decisions based on insights gained from analyzing large amounts of data.

Increased Efficiency

Data science allows businesses to automate processes, reduce  costs, improve efficiency, and streamline operations.

Improved Customer Experience

By analyzing customer data, businesses can personalize their offerings, improve customer satisfaction, and drive customer retention.

# Applications of Data Science

- **Business:** Customer segmentation, fraud detection, personalized recommendations, market research, risk assessment.
- **Healthcare:** Disease diagnosis, drug discovery, personalized medicine, medical imaging analysis, patient monitoring.
- **Finance:** Algorithmic trading, risk management, credit scoring, fraud detection, financial forecasting.
- **Social Media:** Sentiment analysis, recommendation systems, targeted advertising, network analysis.
- **Government:** Predictive policing, disaster response, urban planning, public health surveillance.

# Data Science Challenges

1. **Data Quality Issues:**
   - **Inaccurate Data:** Errors, inconsistencies, and outdated information can lead to flawed analyses and misleading conclusions.
   - **Incomplete Data:** Missing values can hinder analysis and reduce the effectiveness of models.
2. **Data Volume and Velocity:**
   - **Big Data:** The sheer volume of data generated today can be overwhelming to process and analyze efficiently.
   - **Real-time Data:** The need to analyze streaming data in real-time presents challenges for data processing and model deployment.

# Data Science Challenges

3. **Data Privacy and Security:**
   - **Data Breaches:** Sensitive data is vulnerable to cyberattacks, leading to potential harm to individuals and organizations.
   - **Regulations:** Compliance with data privacy regulations (e.g., GDPR, CCPA) adds complexity and constraints to data usage.
4. **Lack of Skilled Professionals:**
   - **Talent Gap:** There is a significant shortage of skilled data scientists with the necessary expertise in areas like machine learning, statistics, and programming.
   - **Interdisciplinary Skills:** Finding professionals with both technical skills and domain expertise is challenging.

# Data Science Challenges

5. **Explainability and Interpretability:**
   - **Black Box Models:** Many advanced machine learning models are complex and difficult to understand, making it hard to explain their decisions and build trust.
   - **Bias and Fairness:** Models can inadvertently reflect biases present in the training data, leading to unfair or discriminatory outcomes.
6. **Communication and Collaboration:**
   - **Bridging the Gap:** Effectively communicating complex technical concepts to non-technical stakeholders is crucial but often challenging.
7. **Ethical Considerations:**
   - **Transparency and Accountability:** Ensuring transparency and accountability in the use of data and the development of AI systems.

# Introduction to Data Science

# Software Engineering for Data Science

Bridges the gap between data science and software development.

It involves **applying software engineering principles & practices to the data science lifecycle**, ensuring that data-driven solutions are not only effective but also efficient, scalable, and maintainable.

Software Engineering for Data Science

## Key Concepts

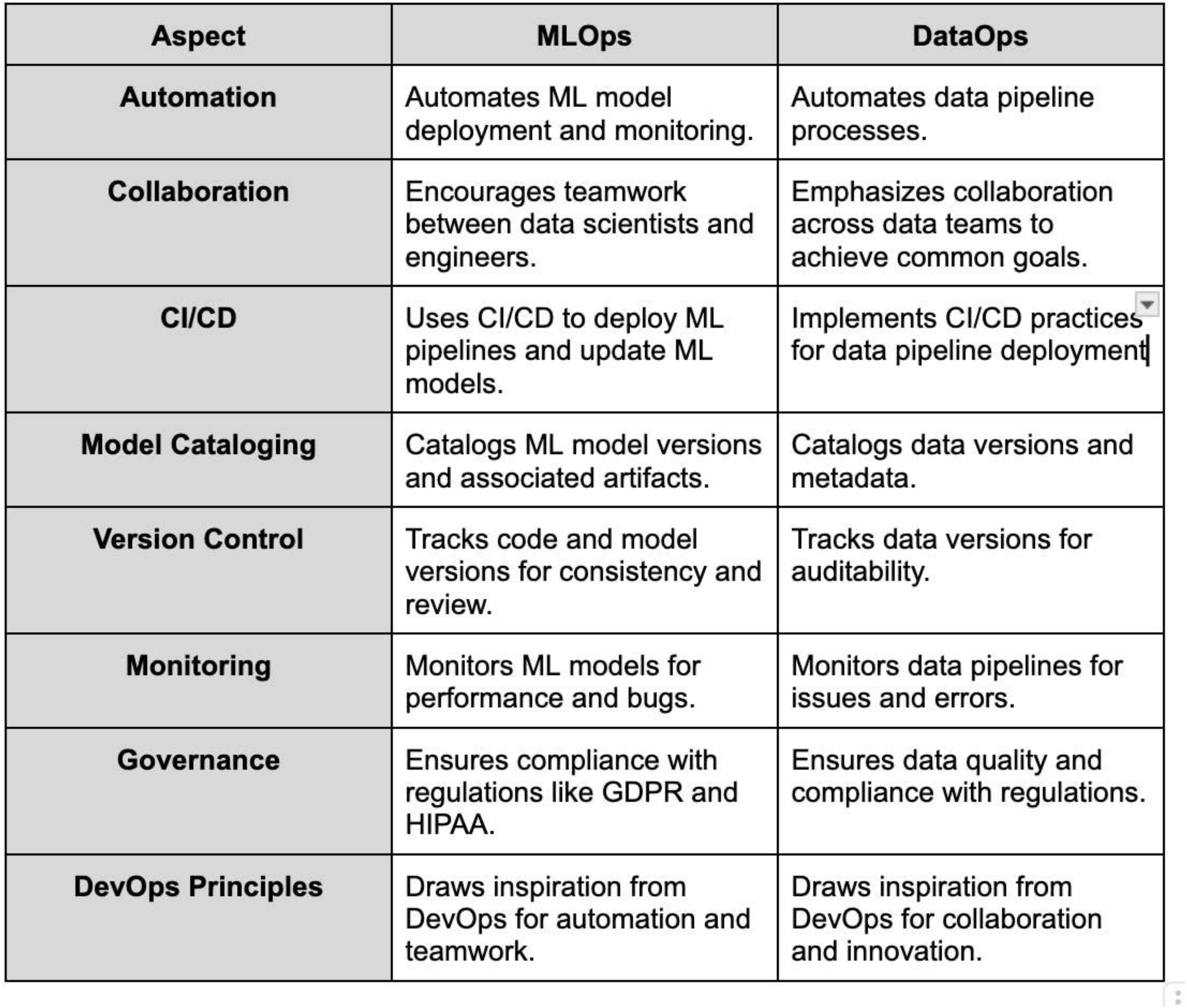
- **DataOps:** A set of practices that aim to shorten the system development lifecycle while delivering features, fixes, and updates frequently and reliably.
  - It focuses on the entire data pipeline, from data ingestion and transformation to analysis and visualization.
- **MLOps:** A set of practices that automate and streamline the machine learning (ML) lifecycle, enabling faster and more reliable development and deployment of ML models.
  - It covers the entire ML workflow, from data preparation and model training to deployment, monitoring, and retraining.

# Software Engineering for Data Science

## Core Principles

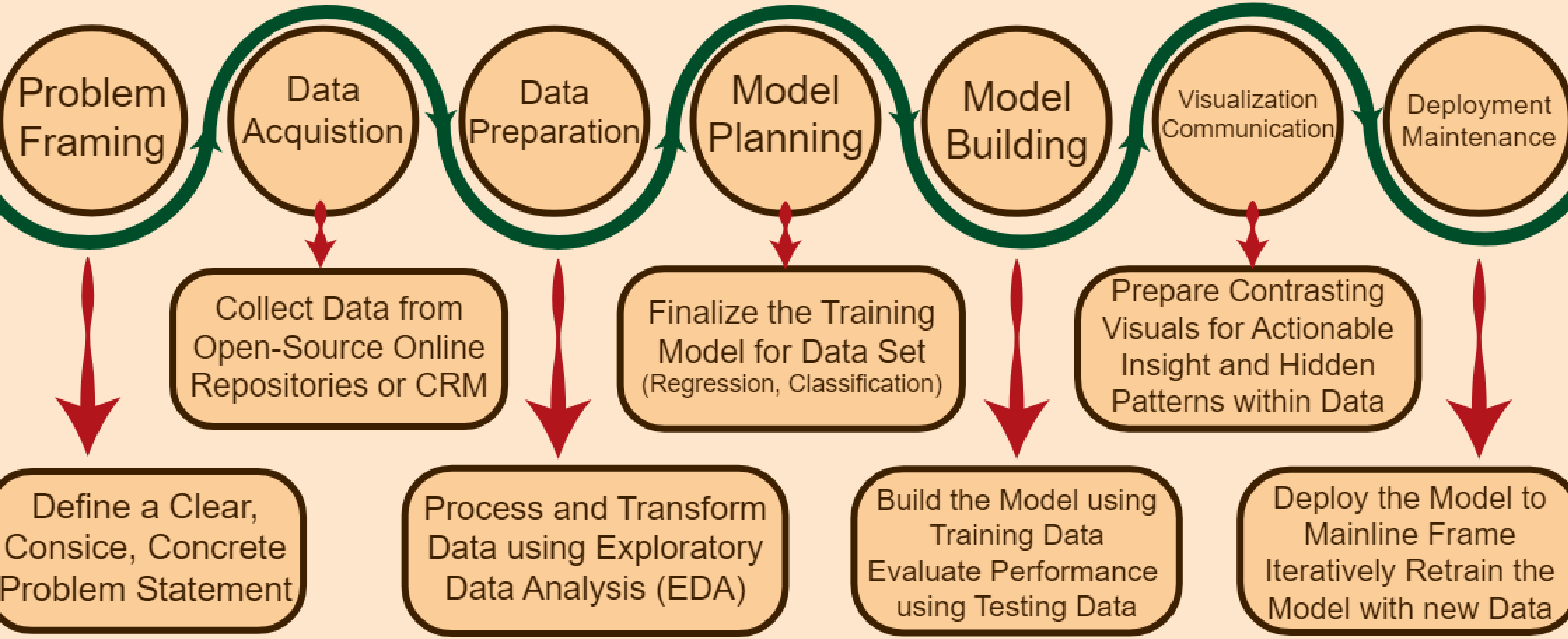
- **Version Control:** Using tools like Git to track changes in code, data, and models.
- **Testing and Quality Assurance:** Implementing unit tests, integration tests, and end-to-end tests to ensure the accuracy and reliability of data pipelines and ML models.
- **Continuous Integration and Continuous Delivery (CI/CD):** Automating the build, test, and deployment processes to accelerate the development cycle.
- **Infrastructure as Code:** Defining and managing infrastructure (e.g., servers, databases, cloud resources) using code, making it easier to reproduce and scale.
- **Scalability & Performance:** Designing systems that can handle large volumes of data and high-throughput workloads.
- **Reproducibility:** Ensuring that experiments and models can be easily reproduced to validate results and maintain consistency.
- **Collaboration:** Fostering collaboration between data scientists, software engineers, and other stakeholders.

# Software Engineering for Data Science



| Aspect | MLOps | DataOps |
|---|---|---|
| **Automation** | Automates ML model deployment and monitoring. | Automates data pipeline processes. |
| **Collaboration** | Encourages teamwork between data scientists and engineers. | Emphasizes collaboration across data teams to achieve common goals. |
| **CI/CD** | Uses CI/CD to deploy ML pipelines and update ML models. | Implements CI/CD practices for data pipeline deployment |
| **Model Cataloging** | Catalogs ML model versions and associated artifacts. | Catalogs data versions and metadata. |
| **Version Control** | Tracks code and model versions for consistency and review. | Tracks data versions for auditability. |
| **Monitoring** | Monitors ML models for performance and bugs. | Monitors data pipelines for issues and errors. |
| **Governance** | Ensures compliance with regulations like GDPR and HIPAA. | Ensures data quality and compliance with regulations. |
| **DevOps Principles** | Draws inspiration from DevOps for automation and teamwork. | Draws inspiration from DevOps for collaboration and innovation. |

# DATA SCIENCE PROCESS



Problem Framing

Define a Clear, Consice, Concrete Problem Statement

Data Acquistion

Collect Data from Open-Source Online Repositories or CRM

Data Preparation

Process and Transform Data using Exploratory Data Analysis (EDA)

Model Planning

Finalize the Training Model for Data Set (Regression, Classification)

Model Building

Build the Model using Training Data Evaluate Performance using Testing Data

Visualization Communication

Prepare Contrasting Visuals for Actionable Insight and Hidden Patterns within Data

Deployment Maintenance

Deploy the Model to Mainline Frame Iteratively Retrain the Model with new Data

# Data Science Process Roles

Data Scientist: The core role responsible for analyzing data, building and evaluating models, and extracting meaningful insights. They possess strong statistical, mathematical, and programming skills.

Data Engineer: Focuses on building and maintaining the data infrastructure, including data pipelines, data warehouses, and data lakes. They ensure data quality, availability, and accessibility for analysis.

Data Analyst: Gathers, cleans, and prepares data for analysis. They perform exploratory data analysis (EDA) and generate reports and visualizations to communicate insights to stakeholders.

# Data Science Process Roles

**Machine Learning Engineer:** They ensure model performance, scalability, and reliability.

**Business Analyst:** Understands business needs and translates them into data science problems. They act as a bridge between the business and the data science team.

**Data Architect:** Designs and implements data solutions, including data models, databases, and data integration strategies

# Types of Data Analytics

DESCRIPTIVE (BUSINESS INTELLIGENCE AND DATA MINING)

PREDICTIVE (FORECASTING)

PRESCRIPTIVE (OPTIMIZATION AND SIMULATION)

DIAGNOSTIC ANALYTICS

# Descriptive Analytics

- It looks at data and analyses past event for insight as to how to approach future events.
- Descriptive analytics looks at past performance and understands the performance by mining historical data to understand the cause of success or failure in the past.
- Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.
- Examples of Descriptive analytics are company reports that provide historic reviews like: Data Queries, Reports, Descriptive Statistics, Data dashboard

# Predictive Analytics

- Predictive analytics turn the data into valuable, actionable information.
- It uses data to determine the probable outcome of an event or a likelihood of a situation occurring.
- Predictive analytics holds a variety of statistical techniques from modeling, machine, learning, data mining, and game theory that analyze current and historical facts to make predictions about a future event.
- Techniques that are used for predictive analytics are:
  - Linear Regression
  - Time series analysis and forecasting
  - Data Mining

# Prescriptive Analytics

- Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefits from the predictions and showing the decision maker the implication of each decision option.
- For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.
- This type of analytics talks about an analysis that is based on rules and recommendations, to prescribe a certain analytical path for an enterprise.
- At the next level, prescriptive analytics will automate decisions and actions—how can we make that happen?

# Diagnostic Analytics

- In diagnostic analytics, most enterprises start to apply data analytics to answer diagnostic questions such as how and why something happened.
- Some may also call this behavioural analytics.
- Diagnostic analytics is about looking into the past and determining why a certain thing happened. This type of analytics usually revolves around working on a dashboard.
- Use historical data over other data to answer any question or for the solution of any problem. We try to find any dependency and pattern in the historical data of a particular problem.

# 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

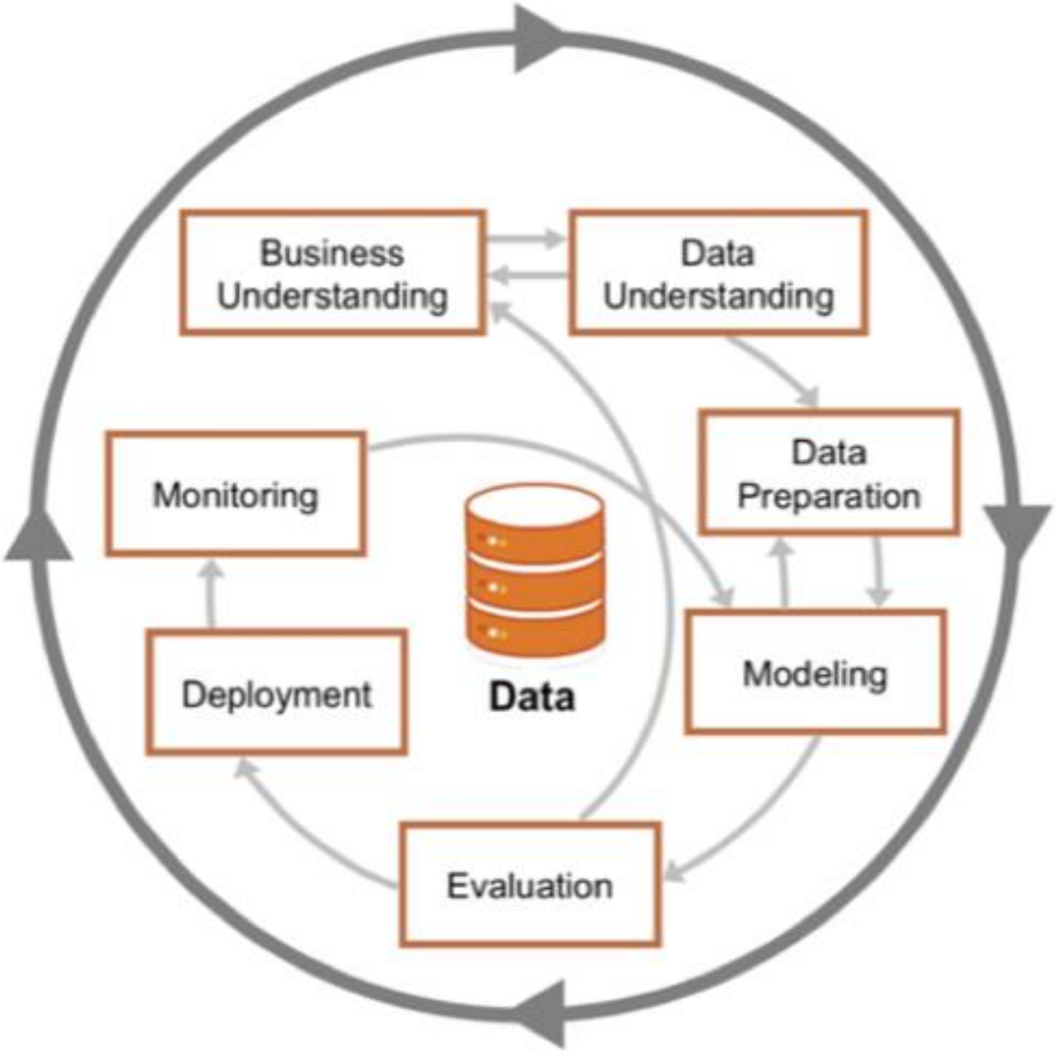
CRISP-DM stands for Cross-Industry Standard Process for Data Mining.

- 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.

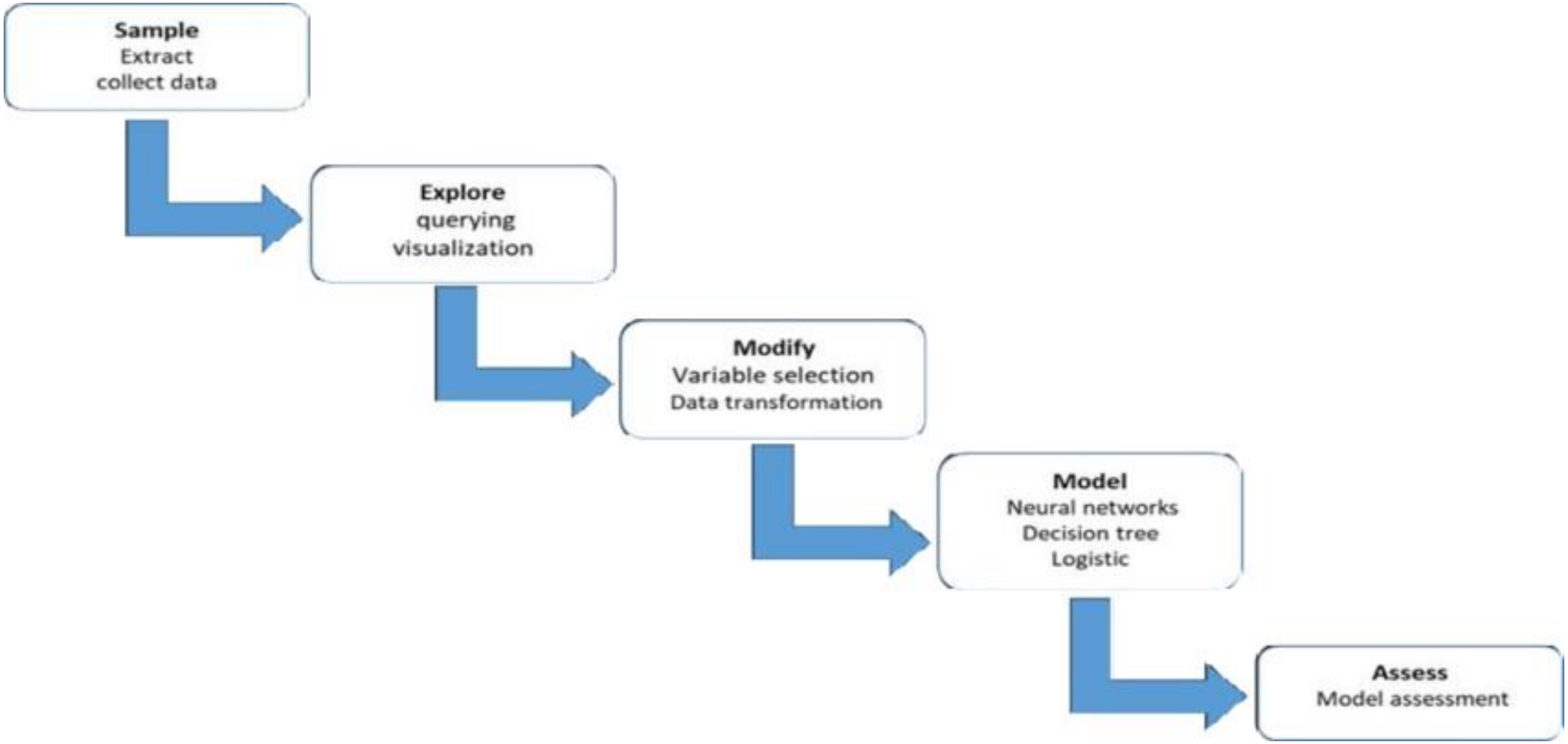
**4. Model:**

- **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:**

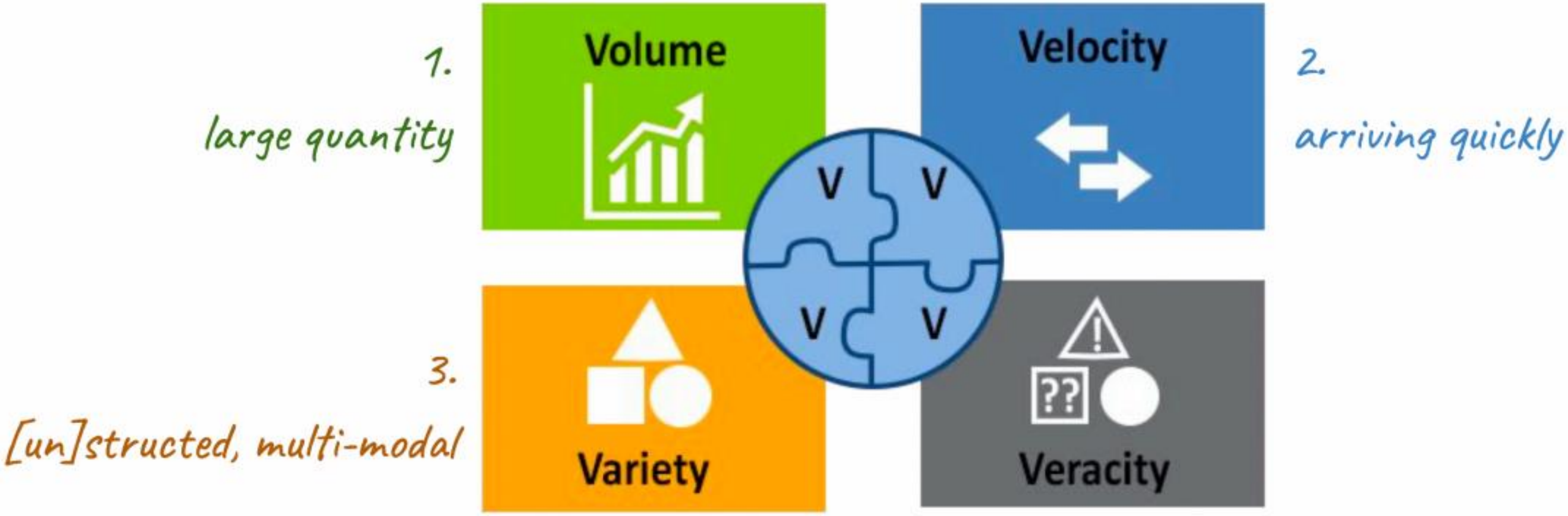
- Evaluate the findings and their implications.

**4. Maintain:**

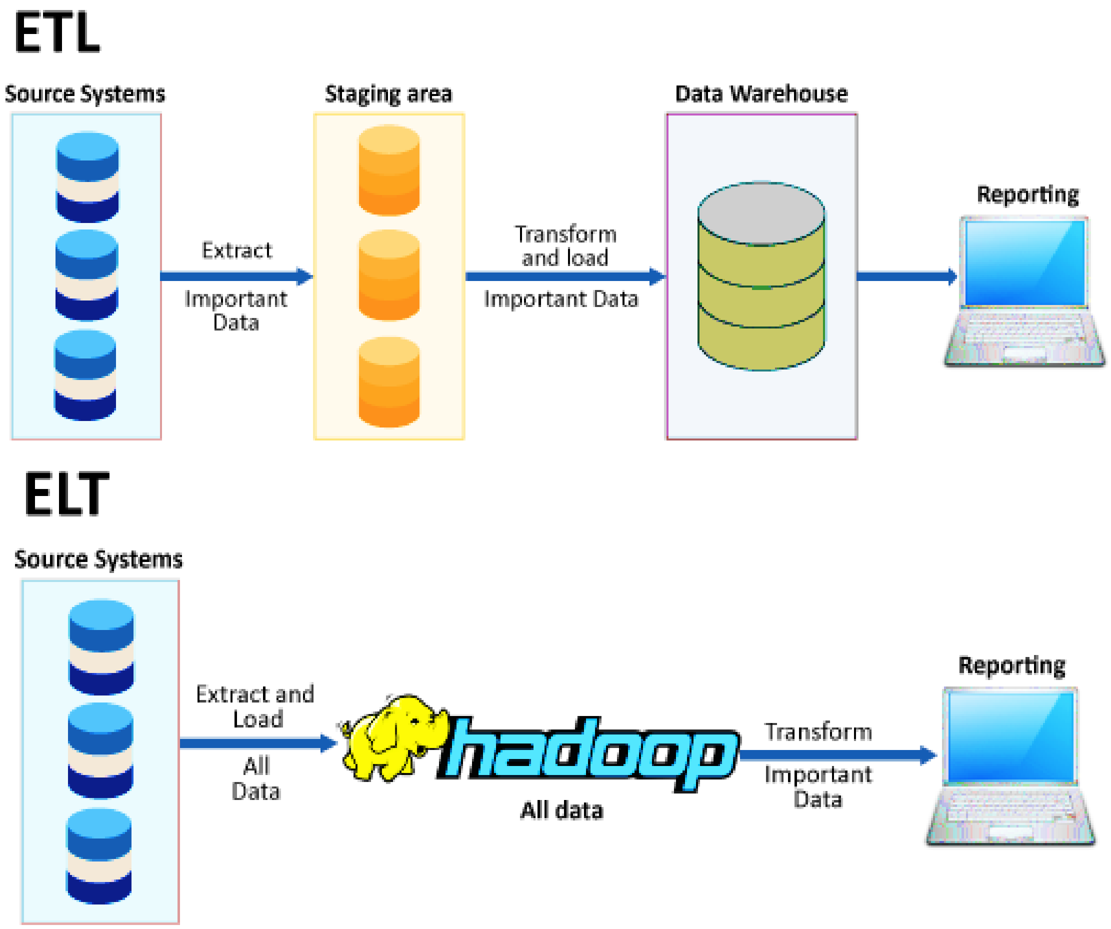
- **Update and refine the analysis as new data becomes available.**
- This may involve:
  - **Retraining models** with new data to improve their performance.
  - **Updating data sources** and re-running the analysis.
  - **Incorporating new insights** and adjusting the analysis accordingly.

# Big Data

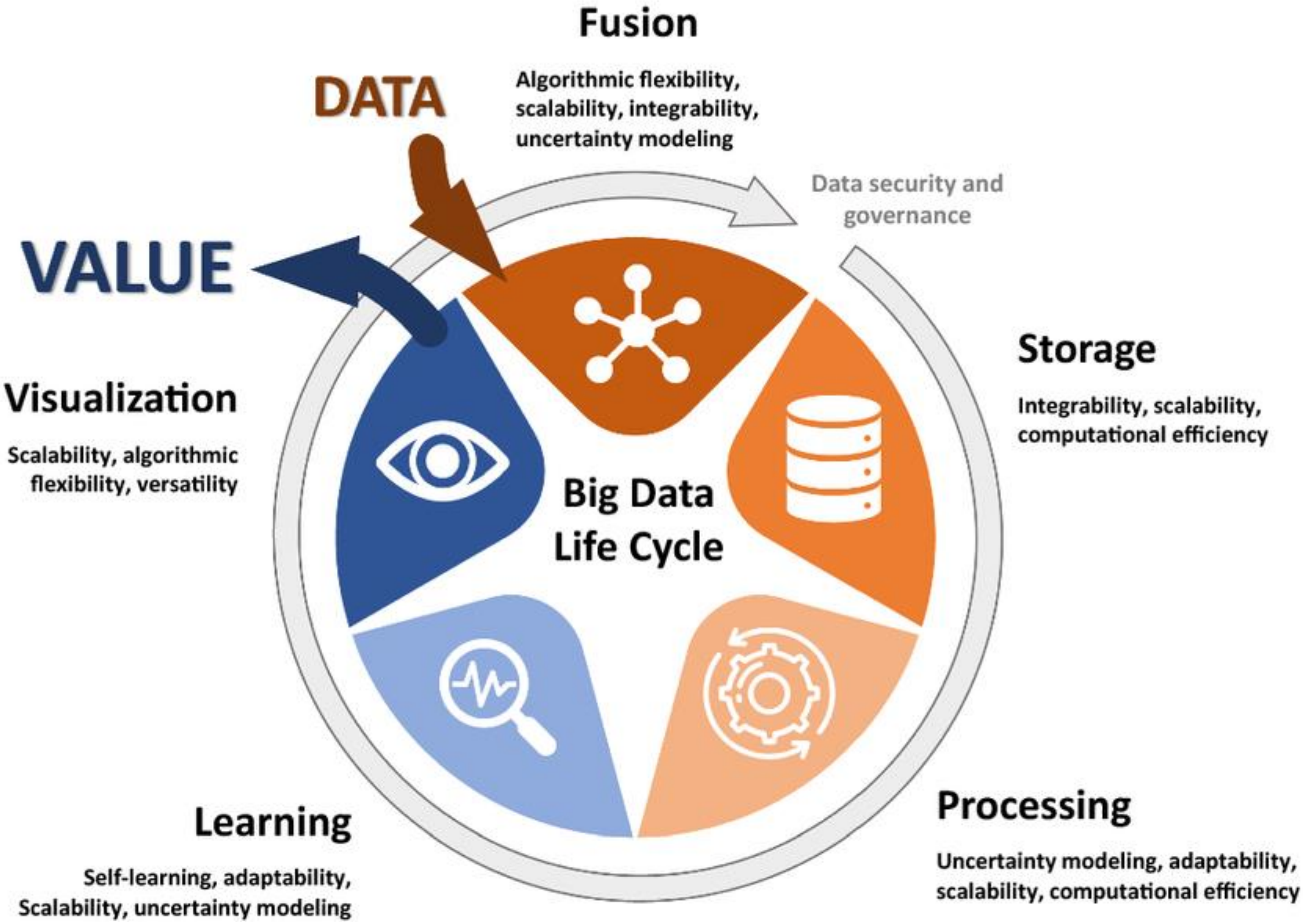
ANALYSES WHICH CAN HANDLE THE 3 VS
AND DO IT WITH QUALITY (VERACITY)



# ETL vs ELT


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

# Big Data Life Cycle


Fusion
Algorithmic flexibility, scalability, integrability, uncertainty modeling
DATA
Data security and governance
VALUE
Storage
Integrability, scalability, computational efficiency
Visualization
Scalability, algorithmic flexibility, versatility
Big Data Life Cycle
Processing
Uncertainty modeling, adaptability, scalability, computational efficiency
Learning
Self-learning, adaptability, Scalability, uncertainty modeling