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

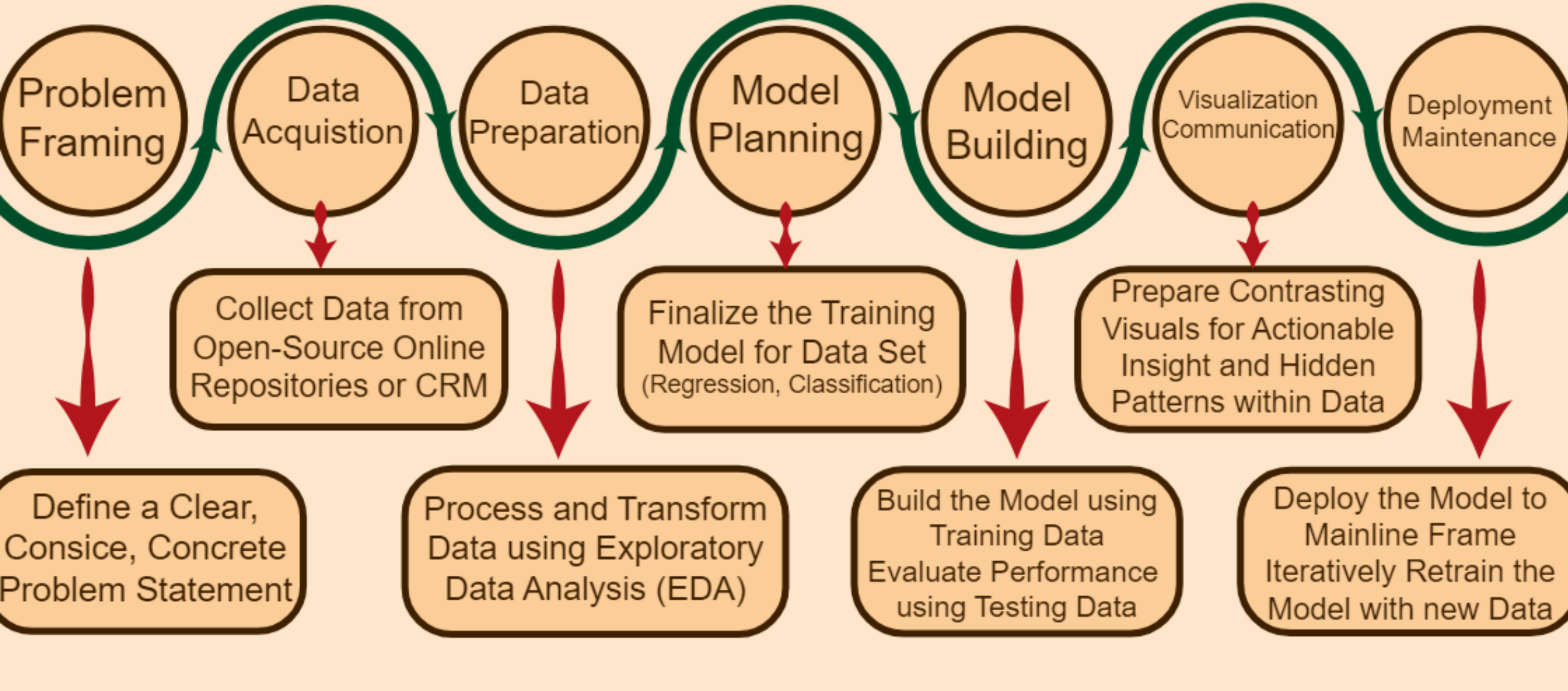
Core Principles

- **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



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.

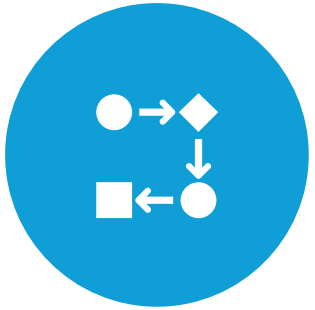
Business Analyst: Understands business needs and translates them into data science problems. They act as a bridge between the business and the data science team.

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.