Chapter 1: What is Data Science?

Learning Objectives

By the end of this chapter, you should be able to:

- Define data science and understand its multidisciplinary nature.

- Differentiate between data science, machine learning, statistics, and data analytics.

- Identify key components of a data science project lifecycle.

- Appreciate the impact of data science in real-world scenarios.

What is Data Science?

Data Science is a field that combines statistical techniques, computer science, and domain expertise to analyze and interpret complex data.

It involves collecting, processing, analyzing, visualizing, and interpreting large volumes of data to uncover patterns and support decision-making.

Core components of data science:

- **Data Collection**: Gathering raw data from various sources such as databases, APIs, or sensors.

- **Data Cleaning**: Handling missing values, outliers, and inconsistencies.

- **Exploratory Data Analysis (EDA)**: Understanding data distribution and relationships.

- **Modeling**: Applying machine learning or statistical models to solve problems.

- **Communication**: Presenting results using dashboards, reports, or visualizations.

Data Science vs Related Fields

Statistics: The backbone of data science, focusing on data analysis and inference.

Machine Learning: Subfield of AI that enables systems to learn from data without being explicitly programmed.

Data Analytics: Often focused on summarizing historical data to derive insights.

Artificial Intelligence (AI): A broader area involving creating systems that simulate human intelligence.

The Data Science Workflow (CRISP-DM)

A standard methodology used in data science projects:

1. **Business Understanding** - Define the problem and project objectives.
2. **Data Understanding** - Collect initial data and explore to identify data quality issues.
3. **Data Preparation** - Clean and format the data for modeling.
4. **Modeling** - Choose modeling techniques and build models.
5. **Evaluation** - Assess if the model meets business objectives.
6. **Deployment** - Deliver the final product or insights to stakeholders.

Real-World Applications

- **Healthcare**: Predicting disease outbreaks, patient diagnosis, personalized medicine.
- **Retail**: Inventory optimization, customer behavior analysis, recommendation systems.
- **Finance**: Credit scoring, algorithmic trading, fraud detection.
- **Education**: Student performance analysis, adaptive learning systems.
- **Transportation**: Route optimization, traffic prediction, autonomous vehicles.

Quick Terms to Know

- **Dataset**: A structured set of data, usually in tabular format.
- **Algorithm**: A series of steps or rules used to solve a problem.
- **Model**: A mathematical representation trained on data to make predictions or classifications.
- **Feature**: A variable or input used in modeling.
- **Label (Target)**: The outcome or variable we want to predict.
- **Training Data**: The data used to teach a model.

- **Test Data**: New data used to evaluate a model's performance.

Practice Questions

1. Define data science and explain why it's considered multidisciplinary.

2. What is the difference between data science and data analytics?

3. Describe the steps involved in the CRISP-DM process.

4. List three industries that benefit from data science and how.

Mini Task

Find a recent news article or research paper about data science or AI.

Summarize the problem it addresses, the data used, and the outcome.

Bonus: Identify which step(s) of the data science process it represents.