Data Science

Introduction

Lecture No. 1

Instructor: Abdul Rehman

Books & Resources

- Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

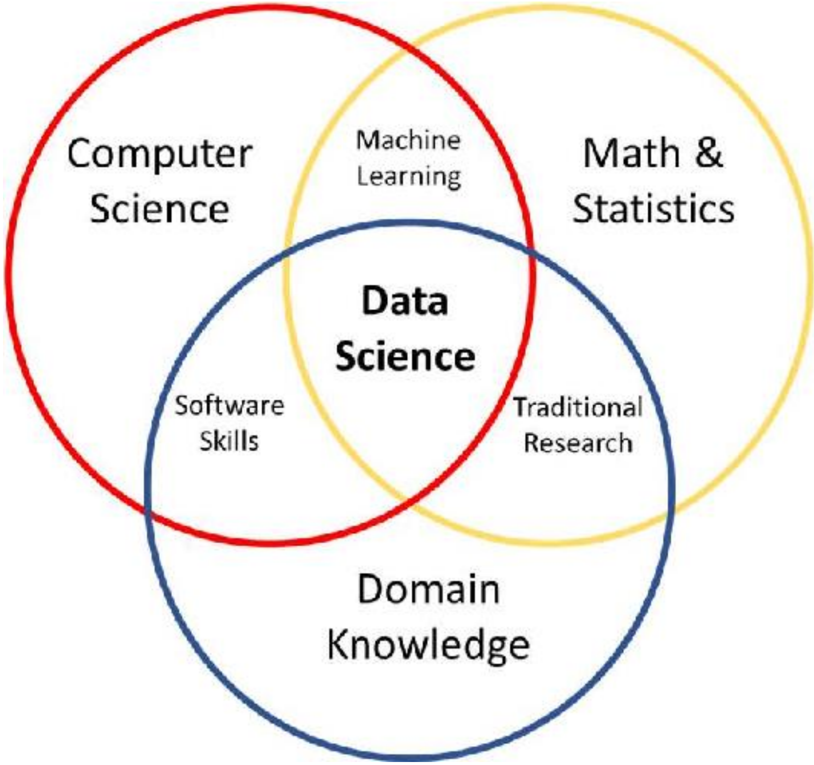
Lecture Outlines

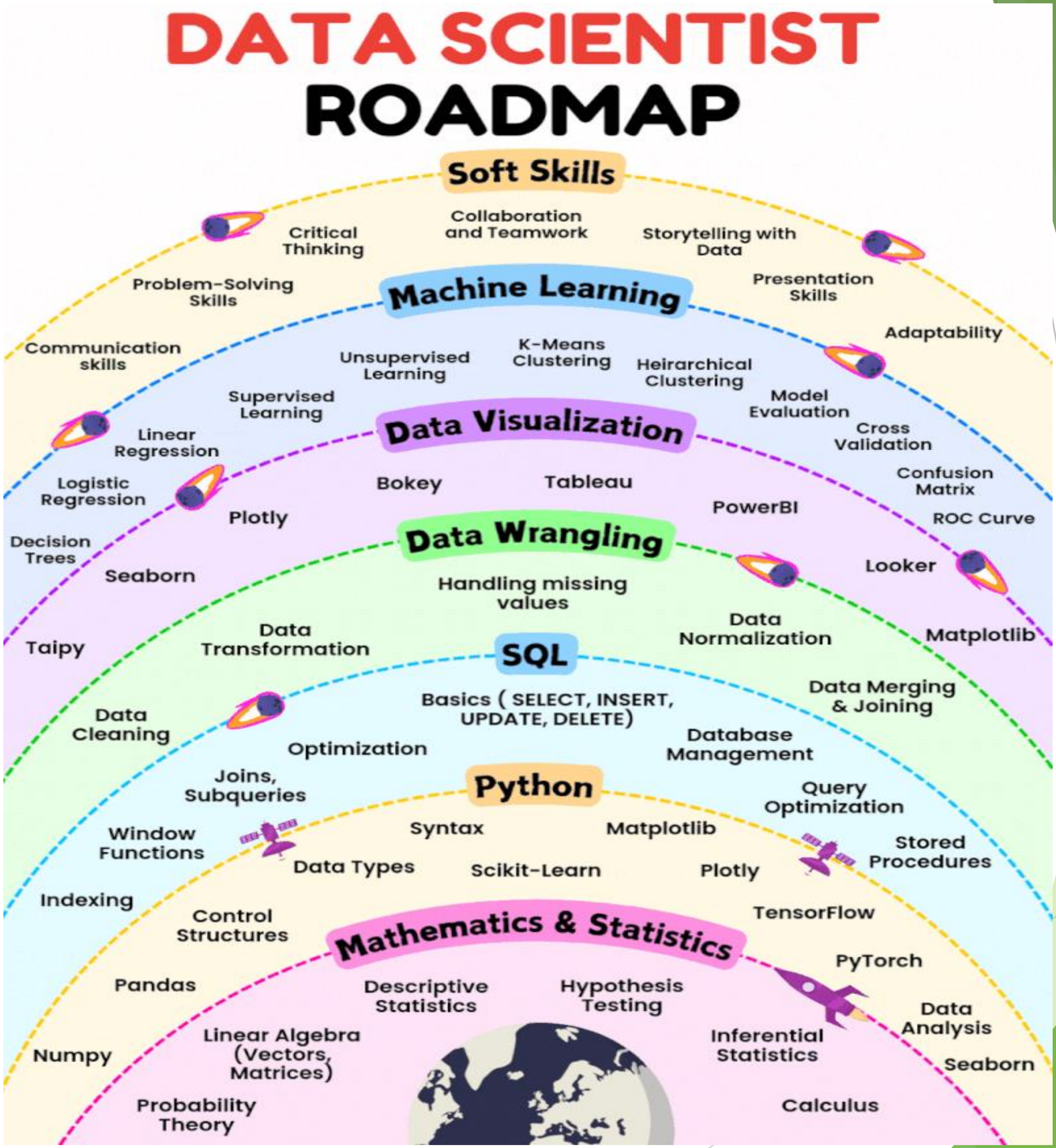
- ❑ What is Data Science?
- ❑ Key Components of Data Science
- ❑ Key Characteristics of Big Data (5 Vs)
- ❑ What is Machine Learning?
- ❑ Difference between AI and ML
- ❑ Applications of ML
- ❑ Dataset and their types (how to get datasets?)
- ❑ Introduction to Data and their types
- ❑ What are features and labels in dataset
- ❑ Skills Required for a Data Scientist

What is Data Science?

- **Data Science** is the process of extracting knowledge and insights from structured and unstructured data using techniques like statistics, mathematics, machine learning, data visualization, and domain knowledge.
- **Data science** is a multidisciplinary field.
- It is also used to solve business problem

Data Science is Multidisciplinary Field





DATA SCIENTIST ROADMAP

Soft Skills

Critical Thinking

Collaboration and Teamwork

Storytelling with Data

Problem-Solving Skills

Presentation Skills

Communication skills

Adaptability

Machine Learning

Unsupervised Learning

K-Means Clustering

Heirarchical Clustering

Supervised Learning

Model Evaluation

Linear Regression

Cross Validation

Logistic Regression

Confusion Matrix

ROC Curve

Decision Trees

Data Visualization

Bokey

Tableau

PowerBI

Plotly

Seaborn

Looker

Taipy

Matplotlib

Data Wrangling

Handling missing values

Data Normalization

Data Transformation

Data Merging & Joining

Data Cleaning

SQL

Basics (SELECT, INSERT, UPDATE, DELETE)

Optimization

Database Management

Joins, Subqueries

Query Optimization

Window Functions

Stored Procedures

Indexing

Python

Syntax

Matplotlib

Data Types

Scikit-Learn

Plotly

TensorFlow

Control Structures

PyTorch

Pandas

Data Analysis

Numpy

Seaborn

Mathematics & Statistics

Descriptive Statistics

Hypothesis Testing

Linear Algebra (Vectors, Matrices)

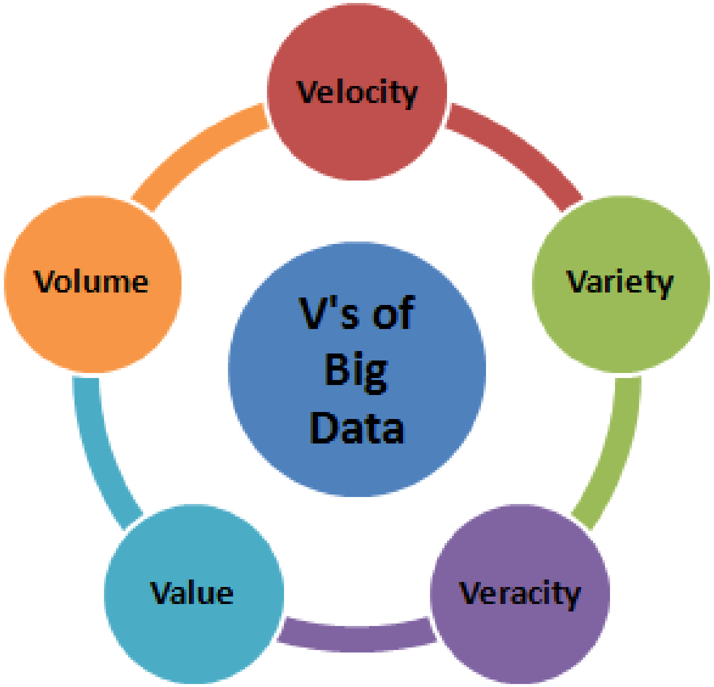
Inferential Statistics

Probability Theory

Calculus

Key Characteristics of Big Data (5 Vs)

- **Volume**: Massive amounts of data (terabytes to petabytes).
- **Velocity**: High-speed data generation (real-time streaming data).
- **Variety**: Different types of data (structured, unstructured, semi-structured).
- **Veracity**: Ensuring data quality and accuracy.
- **Value**: Extracting useful insights for decision-making.


V's of Big Data
Velocity
Variety
Veracity
Value
Volume

Key Components of Data Science

- **Data Collection** – Gathering structured and unstructured data from various sources (databases, APIs, web scraping).
- **Data Preprocessing** – Cleaning and transforming raw data into a usable format.
- **Exploratory Data Analysis (EDA)** – Identifying patterns, trends, and outliers using visualization tools.

Key Components of Data Science

- **Machine Learning** – Building predictive models using algorithms like regression, classification, and clustering.
- **Data Visualization** – Presenting findings through graphs, charts, and dashboards (using tools like Matplotlib, Power BI, and Tableau).
- **Decision Making** – Using insights for business intelligence, automation, and strategy.

Applications of Data Science

- **Healthcare**
  - Disease prediction and diagnosis
  - Drug discovery and personalized medicine
  - Medical image analysis (X-rays, MRIs, CTSCAN, etc.)
- **Finance**
  - Fraud detection and prevention
  - Algorithmic trading and investment analysis
- **Marketing**
  - Customer segmentation and targeting
  - Sentiment analysis on social media
  - Marketing campaign optimization

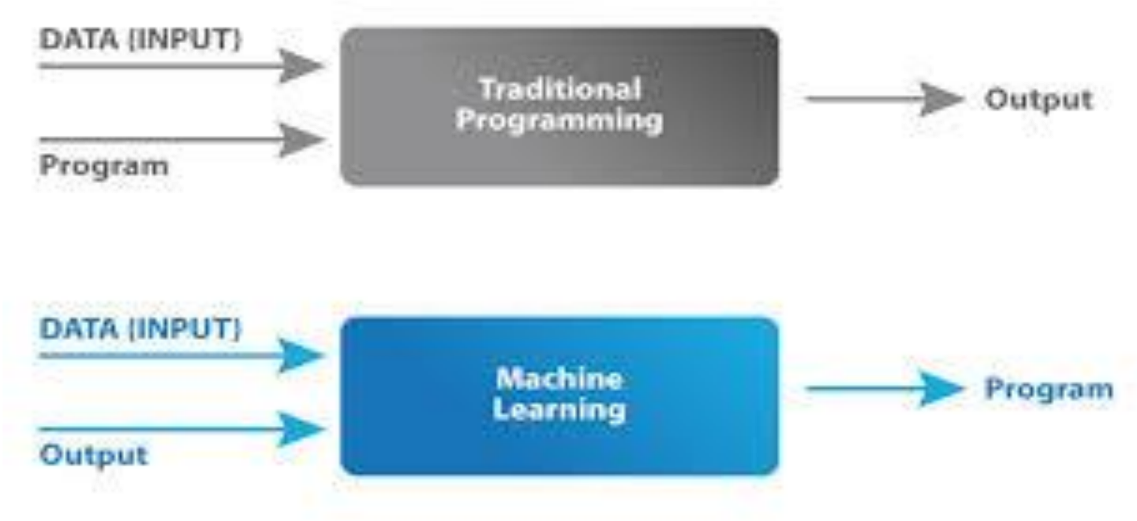
Applications of Data Science

- **E-Commerce**
  - Product recommendation systems
  - Customer behavior analysis
- **Transportation**
  - Route optimization (like Google Maps)
  - Autonomous vehicle navigation
- **Education**
  - Student performance prediction
- **Sports Analytics**
  - Player performance tracking
  - Game strategy optimization

Machine Learning & Artificial Intelligence

- Machine Learning is the field of study that give computers the capability to learn without being explicitly programed.
- Machine can learn itself from past data and automatically improve with experience.
- Artificial Intelligence comprises two words “Artificial” and “Intelligence”. Artificial refers to something which is made by humans or a non-natural thing and Intelligence means the ability to understand or think.

Traditional Programming and Machine Learning



- **Traditional Programming:** Rules + Data → Output.
- **Machine Learning:** Data + Output → Rules (model).

Examples of AI, ML, and DL

Example Scenario:

Let's say you want to develop a **self-driving car**:

- **AI:** The car should be able to drive itself intelligently like a human.
- **ML:** The car learns from driving data — when to brake, accelerate, or turn based on previous experiences.
- **DL:** The car uses deep learning models (like Convolutional Neural Networks) to recognize traffic signs, pedestrians, and lane markings from camera images.

Uses cases of ML & AI

- **For example:**

1. Fraud Detection (ML)
2. Email Filtering (ML)
3. Medical Diagnosis (ML)
4. Pattern Recognition (ML)
5. Face recognition, (ML)
6. Social media, (ML)
7. Movie recommendation systems, (ML)
8. Virtual assistants, (ML)

- **For example:**

1. Autonomous vehicles, (AI) [Tesla, Cruise, Waymo, Drones]
2. Chabot, (AI) [Gimini, Tesla, ChatGPT, Siri, Xiaowei, Jeeney AI, etc.]
3. NLP, (AI) [Google Translator]
4. Robotics (AI & ML)
5. Self-driving Car (AI & ML)

What is a Dataset

- A dataset is a collection of data organized in a structured format, often used for analysis, research, or training machine learning models.
- Datasets can contain various types of data, such as numerical values, text, images, or even videos, and are typically arranged in rows and columns (similar to a table).
- Each row represents
  - an instance/record), and
- Each column represents
  - a feature or attribute of the data

Types of Dataset

- **Structured datasets**: Organized in a tabular form (like a spreadsheet or database).
- **Unstructured datasets:** Data without a predefined format, like images, videos, or text documents.
- **Semi-structured datasets**: Have some organization but aren't as strictly structured, such as JSON or XML files.

Data Sets Sources

Datasets:

1. Kaggle Datasets
2. Amazon Datasets
3. UCI Machine Learning Repository
4. Google Data Search Engine
5. Microsoft Datasets
6. Awesome Public Dataset Collection
7. Scikit Learn Datasets
8. Computer Vision Datasets
9. Government Datasets (such as: https://www.data.gov/us)