**What is Data?**

Data is the information we interact with daily, such as text, phone numbers, and time. Computers need data to perform computations. With the rise of the Internet, computers handle data for processing and communication more than for computations.

**What is Data Science?**

Data science is the field that uses scientific methods to extract knowledge and insights from data, both structured and unstructured, and apply these insights to various domains. Key aspects include:

1. Extracting knowledge from data. 2. Using scientific methods like statistics. 3. Applying insights to practical situations. 4. Handling both structured and unstructured data.

5. Domain expertise in areas like finance, medicine, and marketing.

**Related Fields in Data Science**

Databases: Storing data efficiently for faster processing.

Big Data: Handling large volumes of data with simple structures.

Machine Learning: Building models to predict outcomes.

Artificial Intelligence: Creating complex models that mimic human thought processes.

Visualization: Using visual tools to understand and interpret data.

**Types of Data**

Structured Data: Organized in tables, like phone lists and temperature logs.

Semi-structured Data: Partially organized, like JSON-formatted scientific papers.

-Unstructured Data: Unorganized, like text documents and raw video feeds.

**Data Sources**

Structured Data: IoT sensors, surveys, behavior analysis.

Unstructured Data: Texts, images, videos, web server logs.

Semi-structured Data: Social network graphs, photo collections.

**Steps in Data Science**

1. Data Acquisition: Collecting data, sometimes requiring special techniques.

2. Data Storage: Storing data in relational databases, NoSQL databases, or data lakes.

3. Data Processing: Converting data into usable forms, often using AI techniques.

4. Visualization: Visualizing data to understand and find relationships.

5. Training a Predictive Model: Using machine learning to make predictions from data.

**Digitalization and Digital Transformation**

Businesses digitize processes (digitalization) and apply data science to improve productivity (digital transformation). For example, an online course can be improved by analyzing data on student performance and time spent on modules, leading to better content and age recommendations

**Data Ethics Overview**

* **Definition**: Data Ethics studies and evaluates moral problems related to data, algorithms, and practices.
* **Importance**: As data usage grows, ethical guidelines are necessary to prevent harm and ensure the responsible use of data and AI.

**Key Concepts**

1. **Ethics Principles**-Shared values guiding acceptable behaviors in data and AI projects.
2. **Ethics Challenges**- Issues like data ownership, privacy, informed consent, and algorithmic fairness.
3. **Case Studies**- Real-world examples highlighting the impact of ethical lapses, such as the Netflix data breach and MIT's Gender Shades study.

**Applying Data Ethics**

1. **Professional Codes**-Guidelines for ethical behavior in professional settings.
2. **Ethics Checklists**-Practical tools to ensure ethical practices in projects.
3. **Ethics Regulations**-Laws and regulations governing data privacy and protection.
4. **Ethics Culture**-Creating an organizational culture that prioritizes ethical behavior and accountability.

. **Data Definition and Characteristics**

* **Data:** Facts, information, observations, and measurements used for discoveries and informed decisions.
* **Data Point:** A single unit of data within a dataset.
* **Dataset:** A collection of data points, which can come in various formats and structures.

**Types of Data**

* **Raw Data-** Unanalyzed data in its initial state from the source.
* **Quantitative Data-** Numerical data that can be measured and analyzed mathematically
* **Qualitative Data-** Non-numerical, subjective data capturing the quality of something

**Data Structures**

**Structured Data-** Organized into rows and columns, with specific rules for values (e.g., spreadsheets, relational databases)

**Unstructured Data-** Lacks a predefined format, making it flexible but harder to analyze (e.g., text files, video files).

**Semi-structured Data-** Combines elements of structured and unstructured data, following a flexible format with metadata (e.g., HTML, JSON).

**Sources of Data**

Data sources include primary data generated directly by users (e.g., scientists' observations) and secondary data collected for general use. Managed by database systems, databases store structured data for querying. Files encompass various formats like audio, images, and spreadsheets. Online sources host databases and files, accessible via APIs and web scraping for data extraction and sharing across digital platforms.

**Introduction to Statistics and Probability**

**Statistics and Probability** are crucial areas of mathematics in data science. While you can work with data without deep math knowledge, understanding basic concepts is beneficial.

**Probability and Random Variables**

* **Probability**: A number between 0 and 1 that shows how likely an event is. Example: Rolling a dice and getting an even number (3 out of 6) is 0.5.
* **Random Variables**: Values representing outcomes of random events, like dice rolls (1-6). These are discrete if countable and continuous if they cover a range (like bus arrival times).

**Probability Distribution**

* Discrete Variables: Described by a function P(X), summing to 1 over all possible values.
* Continuous Variables: Described by a probability density function p(x). Probability for intervals, not exact values.

**Mean, Variance, and Standard Deviation**

* Mean: Average of values.
* Variance: Spread of values around the mean.
* Standard Deviation: Square root of variance, showing spread in the same units as the data.

**Mode, Median, and Quartiles**

Mode: Most frequent value.

Median: Middle value when data is ordered.

Quartiles: Divide data into quarters, used to make box plots showing data distribution.

**Real-World Data**

Real-world data, like baseball player weights, often follows a probability distribution. We use histograms to visualize this distribution.

**Box Plot**: Visual tool to show mean, median, and quartiles of data.

**Normal Distribution**

Common in real-world data, characterized by mean and variance.

Used to generate random samples fitting real-world data distributions.

**Confidence Intervals**

Estimation range for the true mean of a population based on a sample.

Wider intervals mean higher confidence.

**Hypothesis Testing**

**Hypothesis**: Statement to be tested, like comparing heights of different baseball player roles.

Use confidence intervals and tests (e.g., t-tests) to support or reject hypotheses.

**Law of Large Numbers and Central Limit Theorem**

**Central Limit Theorem**: Mean of large samples from any distribution tends to be normally distributed.

**Law of Large Numbers**: Sample mean approaches true mean as sample size increases.

**Covariance and Correlation**

**Covariance**: Measures how two variables change together.

**Correlation**: Normalized covariance, indicating the strength and direction of the relationship between variables. Values range from -1 (negative correlation) to 1 (positive correlation).