Summary:

What Is Data:

The omnipresence and importance of data in our lives, especially with the advent of computers and the internet. It highlights how computers store, process, transmit, and manipulate data, enabling tasks from emailing to online searches. Understanding these processes is crucial for managing data effectively and ensuring privacy and security in the digital age.

What Is Data Science:

It is a scientific field that applies methods to extract knowledge from structured and unstructured data. Its primary goal is to uncover hidden relationships within data and build models for practical insights across various domains like finance, medicine, and marketing. The field utilizes scientific methods such as statistics, focusing on generating actionable insights that can be applied in real-world business scenarios.

Types Of Data:

Structured Data: Structured data refers to organized and formatted data that is typically stored in databases or other data repositories.

Semi-Structured: Semi-structured data doesn't fit neatly into traditional relational databases but has some organizational properties.

Unstructured Data: Unstructured data refers to information that lacks a predefined format or organization, such as text documents, emails, videos, and social media posts.

What To Do With Data:

- **Data Acquisition**: Collecting data, which can be straightforward (e.g., from a web application) or complex (e.g., buffering IoT sensor data via IoT Hub).

- **Data Storage:** Choosing the right storage method based on future query needs:

  Relational Databases: Use SQL to query structured data in tables and schemas.

NoSQL Databases: Store complex data (e.g., JSON, graphs) without schemas but with limited querying capabilities.

Data Lake Storage: Store large, unstructured datasets for big data processing, often using formats like Parquet.

- **Data Processing:** Transforming data into a usable format for visualization or model training.

- **Visualizations:** Visualizing data to uncover relationships and insights, often using statistical techniques to test hypotheses or correlations.

- **Training a Predictive Model:** Using Machine Learning to build models that make predictions based on new data with similar structures.

DATA ETHICS:

A third of large organizations are expected to trade data through online marketplaces, making it easier for app developers to integrate data-driven insights and automation. However, the rise of AI and data usage also brings potential harms, such as the weaponization of algorithms and ethical concerns about data privacy and user influence. With the prediction that data creation and consumption will reach 180 zettabytes by 2025, data scientists will have unprecedented access to personal data, raising ethical issues around behavioral profiling and decision-making manipulation.

Data ethics, a branch of ethics focusing on data, algorithms, and practices, is crucial to addressing these challenges. It involves the study and evaluation of moral problems related to data handling and algorithm usage. Applied ethics is the practical application of these moral considerations to ensure real-world actions and processes align with ethical values. An ethics culture operationalizes these principles across organizations, defining ethical standards, incentivizing compliance, and promoting desired behaviors to ensure consistent and scalable ethical practices.

How Data Is Define:

Raw Data: Raw data is unprocessed data in its original state. It needs to be organized into a structured format for human and technological analysis. Data structures are classified into three types: structured (e.g., databases), unstructured (e.g., text files, images), and semi-structured (e.g., JSON, XML).

Quantitative Data: Quantitative data consists of numerical observations that can be measured and analyzed mathematically, such as population size, height, or earnings. This data can be further analyzed to uncover trends, like seasonal AQI patterns or rush hour traffic probabilities.

Qualitative Data: Qualitative data, or categorical data, captures subjective qualities that can't be measured numerically, such as video comments, car models, or favorite colors.

Structured Data: Structured data is organized into rows and columns with specific rules for each column to ensure data accuracy, such as a customer spreadsheet with non-empty, numeric phone numbers. It allows easy relation to other structured data but is challenging to restructure. Examples include spreadsheets, relational databases, and bank statements.

Unstructured Data: Unstructured data lacks a defined format and is not organized into rows or columns, making it easier to add new information but harder to analyze. For example, sensor data recording both barometric pressure and temperature without altering existing data. Examples include text files, messages, and video files.

Semi-Structured Data: Semi-structured data combines aspects of structured and unstructured data, it is organized. Examples include HTML, CSV files, and JSON.