**Data** refers to any information we interact with, from text to numbers, and is essential for computations, especially with computers. The Internet has amplified the role of computers in data handling, emphasizing data processing and communication.

**Data Science** is a field that uses scientific methods to extract knowledge and insights from both structured and unstructured data and applies these insights to various domains. It encompasses understanding data, using probability and statistics, and applying practical insights to real-world situations.

Key aspects of data science include:

* Extracting knowledge from data to find hidden relationships and build models.
* Using scientific methods, initially viewed as a subset of statistics but now recognized as broader.
* Applying insights to real business situations.
* Handling both structured and unstructured data.
* Requiring domain expertise in areas like finance, medicine, and marketing.

Data science intersects with related fields:

* **Databases:** Structuring data for efficient processing.
* **Big Data:** Managing and processing large data sets.
* **Machine Learning:** Developing models to predict outcomes.
* **Artificial Intelligence:** Creating complex models that mimic human thought processes.
* **Visualization:** Making data comprehensible through visual representation.

**Types of Data**:

* **Structured:** Well-organized data, like tables.
* **Semi-structured:** Data with some organizational structure, like JSON files.
* **Unstructured:** Raw data, like text or video.

**Sources of Data**:

* **Structured:** IoT sensors, surveys, user behavior analysis.
* **Unstructured:** Texts, images, videos, web server logs.
* **Semi-structured:** Social network graphs, group dynamics data.

**Data Science Processes**:

1. **Data Acquisition:** Collecting data from various sources.
2. **Data Storage:** Deciding on storage methods based on future queries; options include relational databases, NoSQL databases, and data lakes.
3. **Data Processing:** Converting raw data into usable forms, often using AI techniques.
4. **Visualization/Human Insights:** Visualizing data to understand and find relationships.
5. **Training a Predictive Model:** Using machine learning to create models for decision-making based on data.

In essence, data science involves collecting, storing, processing, visualizing, and modeling data to derive actionable insights and make informed decisions.

**Ethics** is derived from the Greek "ethikos," meaning character or moral nature. It governs behavior based on shared values and principles rather than laws, influencing corporate governance and regulations.

**Data Ethics** is a branch of ethics focused on moral issues related to data, algorithms, and practices, including responsible innovation and ethical programming.

**Applied Ethics** involves the practical application of moral considerations to real-world actions, ensuring they align with ethical values.

**Ethics Culture** operationalizes applied ethics across organizations by defining principles, incentivizing compliance, and encouraging ethical behaviors consistently.

**Ethics Concepts**

1. **Ethics Principles**:
   * **Accountability**: Responsibility for ethical operations.
   * **Transparency**: Clarity in data and AI actions.
   * **Fairness**: Addressing biases to ensure equitable AI.
   * **Reliability & Safety**: Consistent and safe AI behavior.
   * **Privacy & Security**: Protecting data and user privacy.
   * **Inclusiveness**: Designing AI for diverse needs.
2. **Ethics Challenges**:
   * **Data Ownership**: Control and rights over personal data.
   * **Informed Consent**: Users' understanding and agreement.
   * **Intellectual Property**: Rights over intangible creations.
   * **Data Privacy**: Protecting user identity and privacy.
   * **Right to Be Forgotten**: Deleting personal data on request.
   * **Dataset Bias**: Avoiding non-representative data.
   * **Data Quality**: Ensuring accurate and consistent data.
   * **Algorithm Fairness**: Preventing discrimination by AI.
   * **Misrepresentation**: Honest data communication.
   * **Free Choice**: Avoiding manipulative choice architectures.

**Case Studies Highlighting Ethics Violations**

1. **Informed Consent**: Tuskegee Syphilis Study.
2. **Data Privacy**: Netflix data prize.
3. **Collection Bias**: Boston's Street Bump app.
4. **Algorithmic Fairness**: MIT Gender Shades Study.
5. **Data Misrepresentation**: Georgia COVID-19 charts.
6. **Illusion of Free Choice**: ABCmouse subscription issues.
7. **Data Privacy & User Rights**: Facebook Data Breach.

**Applied Ethics**

1. **Professional Codes**: Guidelines for ethical behavior.
2. **Ethics Checklists**: Tools like Deon, Privacy Audit, and AI Fairness Checklists for practical application.
3. **Ethics Regulations**: Laws like GDPR, CCPA, HIPAA to ensure data protection.
4. **Ethics Culture**: Building shared values and emotional connections within organizations to address systemic ethical issues.

Ethics in data science involves defining and operationalizing principles, addressing challenges, learning from case studies, applying professional codes and checklists, complying with regulations, and fostering a robust ethics culture.

**Raw Data**: Data in its initial, unanalyzed state from its source. To be useful, it needs to be organized into structured, unstructured, or semi-structured formats.

**Quantitative Data**: Numerical observations that can be measured and analyzed mathematically, like population figures, heights, and earnings.

**Qualitative Data**: Categorical data that captures subjective qualities and is often non-numerical or non-mathematical, like comments, car models, and favorite colors.

**Structured Data**: Organized into rows and columns with defined rules, making it easily related to other structured data but challenging to alter. Examples include spreadsheets and relational databases.

**Unstructured Data**: Lacks a defined format or rules, making it flexible but harder to analyze. Examples include text files, messages, and videos.

**Semi-structured Data**: Combines elements of both structured and unstructured data, organized but not in strict rows and columns. Examples include HTML, CSV files, and JSON.

**Sources of Data**:

* **Primary Data**: Generated by its original user or collector.
* **Secondary Data**: Collected for general use and shared with others.
* **Common Sources**: Databases, files (audio, image, video, spreadsheets), internet sources, APIs, and web scraping.