## Defining Data Science - Summary

### What is Data?

* Data is everywhere: text, phone numbers, time, etc.
* With computers and the Internet, data handling has become crucial for processing and communication.

### What is Data Science?

* Data Science: A field using scientific methods to extract knowledge from structured and unstructured data.
* Goals: Understand data, find hidden relationships, build models, and apply actionable insights.
* Involves statistics, probability, and expertise in specific domains (finance, medicine, etc.).
* Combines empirical, theoretical, computational, and data-driven paradigms of science.

### Related Fields

* **Databases**: How to store data efficiently.
* **Big Data**: Handling large volumes of data with simple structures.
* **Machine Learning**: Building predictive models from data.
* **Artificial Intelligence**: Creating complex models that mimic human thought.
* **Visualization**: Making data comprehensible through visual representations.

### Types of Data

* **Structured**: Well-organized (e.g., tables).
* **Unstructured**: No specific format (e.g., text files).
* **Semi-structured**: Some organization (e.g., JSON).

### Data Sources

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

### Data Journey Steps

1. **Data Acquisition**: Collecting data, sometimes using special techniques.
2. **Data Storage**: Deciding how to store data (relational databases, NoSQL, data lakes).
3. **Data Processing**: Converting data to a usable form, often using AI techniques.
4. **Visualization**: Creating visual representations to understand data.
5. **Training Predictive Models**: Using machine learning to make predictions from data.

### Digitalization and Digital Transformation

* **Digitalization**: Translating business processes into digital form.
* **Digital Transformation**: Using data science to guide business decisions and improve productivity.

### Introduction to Data Ethics

#### Overview

Data ethics involves the moral principles and values governing data generation, processing, sharing, and usage. It's increasingly vital as AI and data become integral to everyday applications, raising concerns about privacy, consent, bias, and fairness.

#### Key Trends

* **Data Marketplaces:** By 2022, 1-in-3 large organizations engage in data trading.
* **Data Growth:** By 2025, data creation and consumption are expected to exceed 180 zettabytes.
* **AI Pervasiveness:** Ethical considerations are crucial to prevent algorithm weaponization and ensure responsible AI deployment.

#### Basic Definitions

* **Ethics:** Moral principles guiding behavior.
* **Data Ethics:** Evaluates moral issues related to data and algorithms.
* **Applied Ethics:** Practical application of ethical principles.
* **Ethics Culture:** Organizational adoption of ethical practices.

#### Core Concepts

1. **Ethics Principles:**
   * **Accountability:** Responsibility for data and AI operations.
   * **Transparency:** Clear and understandable AI decisions.
   * **Fairness:** Elimination of biases in AI systems.
   * **Reliability & Safety:** Consistent and safe AI behavior.
   * **Privacy & Security:** Protection of personal data.
   * **Inclusiveness:** Design for diverse human needs.
2. **Ethics Challenges:**
   * **Data Ownership:** Rights over data usage.
   * **Informed Consent:** Users' understanding and agreement to data use.
   * **Intellectual Property:** Protection of economically valuable data.
   * **Data Privacy:** Safeguarding user data against unauthorized access.
   * **Right to Be Forgotten:** Users' right to data erasure.
   * **Dataset Bias:** Avoiding non-representative data selection.
   * **Data Quality:** Ensuring accurate and consistent data.
   * **Algorithm Fairness:** Preventing discrimination by AI.
   * **Misrepresentation:** Honest reporting and visualization of data.
   * **Free Choice:** Avoiding manipulative choice architectures.
3. **Case Studies:**
   * **Tuskegee Syphilis Study:** Informed consent issues.
   * **Netflix Prize:** Data privacy and de-anonymization.
   * **Street Bump App:** Collection bias and accessibility.
   * **Gender Shades Study:** Algorithmic fairness.
   * **Georgia COVID-19 Charts:** Data misrepresentation.
   * **ABCmouse:** Illusion of free choice.
   * **Facebook Data Breach:** Data privacy and user rights.

#### Applied Ethics

1. **Professional Codes:** Guidelines for ethical behavior in organizations (e.g., Data Science Association Code of Conduct).
2. **Ethics Checklists:** Practical tools for ethical compliance (e.g., Deon checklist).
3. **Ethics Regulations:** Laws and regulations governing data privacy (e.g., GDPR).
4. **Ethics Culture:** Building a culture that prioritizes ethical behavior and allows raising concerns early

## Defining Data

**Data**: Facts, information, observations, and measurements used for discoveries and informed decisions. A **data point** is a single unit within a dataset, which is a collection of data points.

### Raw Data

* **Raw Data**: Unprocessed data from its source. Needs organization for analysis.
* **Dataset Structure**: Can be structured, unstructured, or semi-structured.

### Types of Data

1. **Quantitative Data**: Numerical, measurable data (e.g., population, height, earnings). Used for mathematical analysis and trend discovery.
2. **Qualitative Data**: Categorical, subjective data (e.g., video comments, car models, favorite colors). Used for understanding preferences and patterns.

### Data Structures

1. **Structured Data**: Organized into rows and columns (e.g., spreadsheets, relational databases). Easier to relate, harder to modify.
2. **Unstructured Data**: No specific format (e.g., text files, video files). Easier to add new information, harder to analyze.
3. **Semi-structured Data**: Combination of structured and unstructured (e.g., JSON, HTML). Organized with metadata like tags and attributes.

### Sources of Data

* **Primary Data**: Generated by users directly (e.g., scientists collecting observations).
* **Secondary Data**: Collected for general use and shared with others.

### Common Data Sources

* **Databases**: Hosted and maintained with database management systems.
* **Files**: Audio, image, video files, spreadsheets.
* **Internet Sources**: Databases and files hosted online.
* **APIs**: Allow sharing data over the internet.
* **Web Scraping**: Extracts data from web pages.