Introduction

Data science is a multidisciplinary field that combines **statistics, mathematics, and computer science** to extract meaningful insights and knowledge from large datasets. It involves collecting, cleaning, and analyzing data to solve complex problems, uncover hidden patterns, and drive data-driven decision-making.

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and making future predictions.

By using Data Science, companies can make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

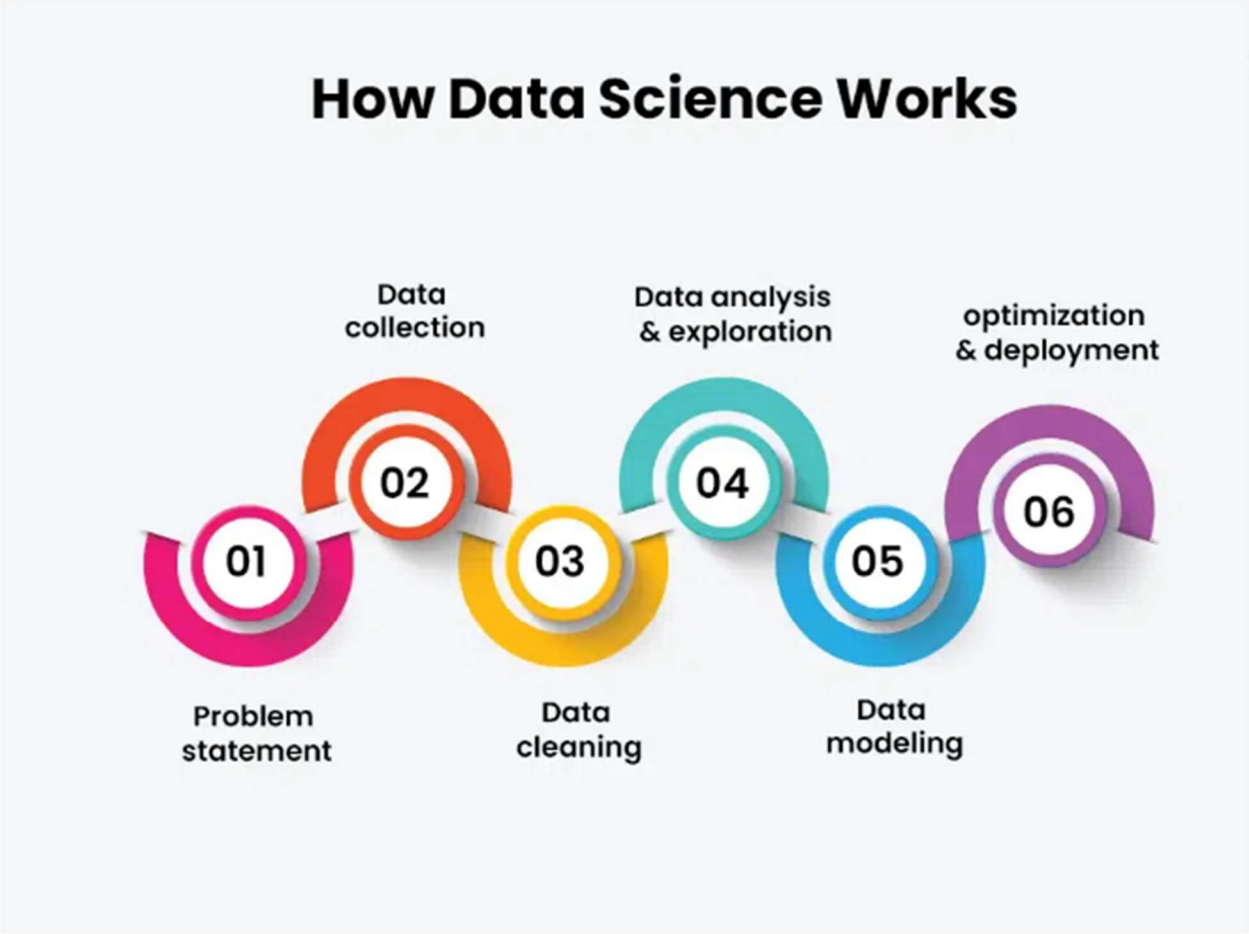
Applications

Key Industry Applications

- **Healthcare**:
    - **Medical Imaging**: Deep learning models analyze X-rays, MRIs, and CT scans to detect anomalies like tumors with high accuracy.
    - **Drug Discovery**: Algorithms simulate molecular interactions to predict drug efficacy, significantly shortening the development timeline.
    - **Predictive Diagnostics**: Models analyze patient history and genetic data to forecast disease progression and suggest personalized treatment plans.
- **Finance & Banking**:
    - **Fraud Detection**: Real-time monitoring systems identify suspicious transaction patterns, such as unusual spending locations or amounts, to prevent financial loss.
    - **Credit Scoring**: Lenders use behavioral data beyond traditional credit scores—like bill payment history and social media activity—to assess a borrower's creditworthiness.

    - **Algorithmic Trading**: Complex mathematical models execute high-frequency trades in microseconds by analyzing market data and news sentiment.

- **E-Commerce & Retail**:
    - **Recommendation Engines**: Platforms like Amazon and Netflix use collaborative filtering to suggest products or content based on a user's past behavior and similar profiles.
    - **Demand Forecasting**: Retailers analyze seasonal trends and historical sales to maintain ideal stock levels, reducing waste and overstocking costs.
    - **Dynamic Pricing**: Algorithms adjust product prices in real-time based on competitor pricing, market demand, and customer behavior.
- **Transportation & Logistics**:
    - **Route Optimization**: Google Maps and logistics companies like UPS use data to identify the most fuel-efficient and fastest routes by predicting traffic jams and delays.
    - **Autonomous Vehicles**: Self-driving cars process real-time sensor data from cameras and GPS to navigate safely and make independent driving decisions.
- **Entertainment & Social Media**:
    - **Content Personalization**: Social platforms like Instagram personalize feeds and advertisements by tracking user interactions like "likes" and "watch time".
    - **Sentiment Analysis**: Marketers use natural language processing (NLP) to gauge public opinion on social media to refine their brand strategies.



What is Big Data?

Big Data refers to vast and rapidly growing volumes of data that are too large and complex for traditional data processing tools to manage.

This data comes in many forms structured (e.g., tables), semi-structured (e.g., JSON, XML), and unstructured (e.g., text, images, video).

With the explosion of devices, sensors, online services, and digital platforms, data is now generated at an unprecedented rate. This growth makes it essential for organizations to adopt advanced tools and technologies to capture, store, analyze, and utilize this data effectively.

Big Data transforms raw information into actionable insights that help companies gain a competitive edge.

Traits of Big Data

Traits of Big data are primarily defined by the 5 V's: Volume, Velocity, Variety, Veracity, and Value. These characteristics differentiate big data from traditional datasets by its scale, speed of generation, and complexity of formats.

The 5 V's of Big Data

- **Volume**: Volume refers to a large size of data generated and stored every second using IoT devices, social media, videos, financial transactions, and customer logs. The data generated from the devices or different sources can range from terabytes to petabytes and beyond. To manage such large quantities of data requires robust storage solutions and advanced data processing techniques.
- **Velocity**: The speed with which data is generated, processed, and analyzed. With the development and usage of IoT devices and real-time data streams, the velocity of data has expanded tremendously, demanding systems that can process data instantly to derive meaningful insights.
- **Variety**: Data comes in multiple formats-text, audio, images, videos, logs, sensor data, etc. Handling all these types together is complex. This diversity requires advanced tools for data integration, storage, and analysis.
- **Veracity**: Refers to the trustworthiness and accuracy of the data. Inconsistent, duplicated, or noisy data can lead to wrong insights.
- **Value**: The ability to convert large volumes of data into useful insights. Big Data's ultimate goal is to extract meaningful and actionable insights that can lead to better decision-making, new products, enhanced consumer experiences, and competitive advantages.

Web Scraping

Web scraping is an automated method to extract large amounts of data from websites. This data, usually in HTML format, is converted into structured formats like spreadsheets or databases for further use.

It can be done through online tools, APIs, or custom code. While major websites like Google, Twitter, and Facebook offer APIs for structured data access, web scraping is often used for sites that lack such options or restrict data access.

Web scraping involves two main components:

Crawler: An AI algorithm that navigates the web and follows links to find the required data.

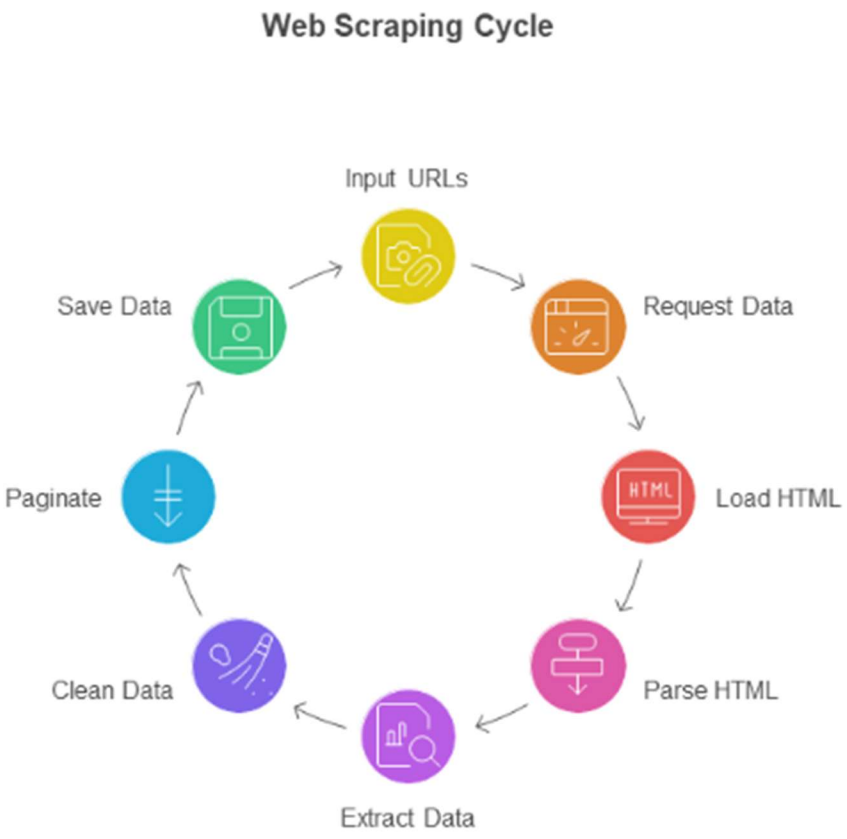
Scraper: A tool designed to extract the identified data from websites, with its design varying based on the project's complexity and scope.

How Web Scrapers Work?

Web Scrapers can extract all the data on particular sites or the specific data that a user wants. Ideally, it's best if you specify the data, you want so that the web scraper only extracts that data quickly. For example, you might want to scrape an Amazon page for the types of juicers available, but you might only want the data about the models of different juicers and not the customer reviews.

Web Scraping End-to-End Flow

- **Input**: Give URLs + specify what data you want (e.g., product name & price only).
- **Request**: Scraper visits each URL like a browser (sends HTTP GET).
- **Load**: Downloads HTML (runs JavaScript if needed for dynamic pages).
- **Parse**: Turns HTML into a navigable structure.
- **Extract**: Finds & pulls only the targeted data using selectors/XPath.
- **Clean**: Trims, converts, and organizes data into rows.
- **Paginate**: Follows "Next" links and repeats until done.
- **Save**: Exports clean data such as CSV, Excel, JSON, or database.



Types of Web Scrapers

Web Scrapers can be categorized based on different criteria such as development type, platform, and execution environment.

Based on Development Type

- **Self-built Web Scrapers**
    - Created from scratch using programming languages like Python or JavaScript.
    - Require advanced coding knowledge.
    - Offer full customization and flexibility.
    - More features demand deeper technical expertise.
- **Pre-built Web Scrapers**

- Already developed tools that can be easily downloaded and run.
- Offer user-friendly interfaces and advanced customization options.
- Suitable for users with little or no coding experience.

Based on Platform

- **Browser Extension Web Scrapers**
  - Installed directly as extensions in browsers like Chrome or Firefox.
  - Easy to use and quick to set up.
  - Limited by browser capabilities — cannot perform complex or large-scale scraping tasks.
- **Software Web Scrapers**
  - Standalone applications installed on your computer.
  - More advanced and feature-rich than browser-based scrapers.
  - Not limited by browser restrictions but require installation and system resources.

Based on Execution Environment

- **Cloud Web Scrapers**
  - Operate on cloud servers provided by scraper vendors.
  - Don’t use your computer’s CPU or RAM.
  - Allow multitasking since data scraping runs remotely.
- **Local Web Scrapers**
  - Run directly on your own computer.
  - Depending on local system resources (CPU, RAM).
  - May slow down your system during heavy scraping tasks.

What is Web Scraping Used for?

Web Scraping has multiple applications across various industries. Let's check out some of these now!

1. Price Monitoring

Web Scraping can be used by companies to scrap the product data for their products and competing products as well to see how it impacts their pricing strategies. Companies can use this data to fix the optimal pricing for their products so that they can obtain maximum revenue.

2. Market Research

Web scraping can be used for market research by companies. High-quality web scraped data obtained in large volumes can be very helpful for companies in analyzing consumer trends and understanding which direction the company should move in the future.

3. News Monitoring

Web scraping news sites can provide detailed reports on the current news to a company. This is even more essential for companies that are frequently in the news or that depend on daily news for their day-to-day functioning. After all, news reports can make or break a company in a single day!

4. Sentiment Analysis

If companies want to understand the general sentiment for their products among their consumers, then Sentiment Analysis is a must. Companies can use web scraping to collect data from social media websites such as Facebook and Twitter as to what the general sentiment about their products is. This will help them in creating products that people desire and moving ahead of their competition.

5. Email Marketing

Companies can also use Web scraping for email marketing. They can collect Email ID's from various sites using web scraping and then send bulk promotional and marketing Emails to all the people owning these Email ID's.

Analysis vs Reporting

In data science, **reporting** and **analysis** are often used interchangeably, but they represent distinct stages of the data lifecycle.

Reporting organizes and presents historical data in structured, static formats (e.g., dashboards, tables) to show "what happened," while analysis interprets data through modeling and investigation to understand "why" it happened and predict future outcomes. Reporting monitors performance, whereas analysis provides actionable insights for decision-making.

Reporting in Data Science

- **Focus:** Descriptive (What happened?).
- **Purpose:** Organizing, summarizing, and presenting raw data.
- **Nature:** Static, standardized, and periodic (e.g., monthly dashboards).
- **Output:** Canned reports, dashboards, alerts.
- **Example:** A monthly sales dashboard showing total revenue by region.

Analysis in Data Science

- **Focus:** Diagnostic, predictive, and prescriptive (Why? What next? What to do?).
- **Purpose:** Interpreting data to discover patterns and insights.
- **Nature:** Dynamic, flexible, and ad-hoc.
- **Output:** Actionable insights, recommendations, and forecasts.
- **Example:** Investigating why a specific region's sales dropped and recommending a price adjustment.

Key Differences

- **Role:** Reporting provides information; analysis provides insights.
- **Direction:** Reporting looks backward at historical data; analysis looks toward future actions.
- **Complexity:** Reporting is generally simpler and automated, while analysis requires deeper, human-driven investigation.

- **Goal:** To turn raw data into knowledge.