Introduction to Data Science

Overview of Data Science

What is Data Science?

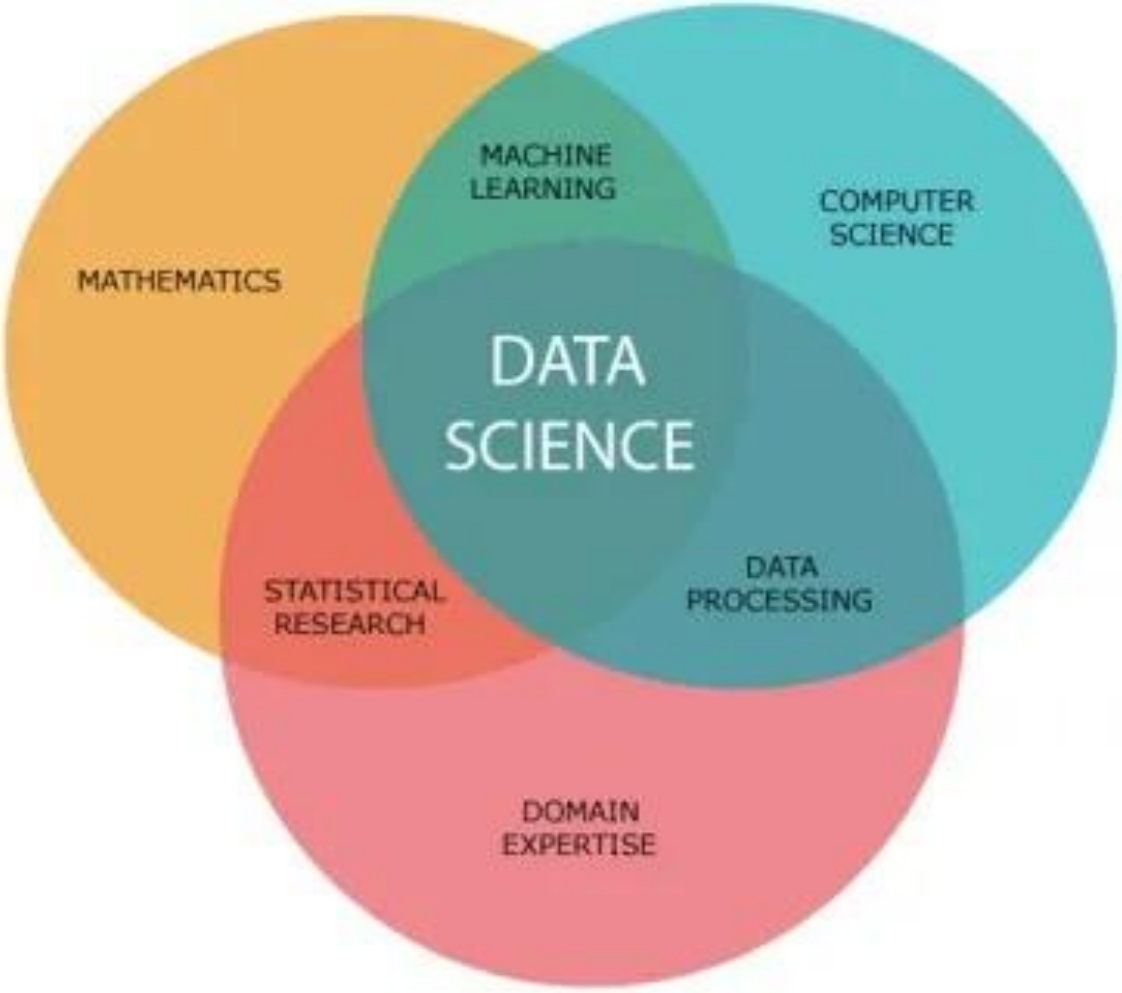
An interdisciplinary field focused on extracting insights from data using techniques from mathematics, statistics, and computer science."

Key Components

- Data Collection
- Data Processing
- Analysis & Modeling
- Visualization & Communication

Importance

Drives innovation in AI, business, and research.


MACHINE
LEARNING
COMPUTER
SCIENCE
MATHEMATICS
DATA
SCIENCE
STATISTICAL
RESEARCH
DATA
PROCESSING
DOMAIN
EXPERTISE

Jargons of Data Science

Key Terms in Data Science

- **Data Wrangling:** Cleaning and transforming raw data.
- **Feature Engineering:** Selecting and transforming variables for models.
- **Machine Learning:** Algorithms that learn patterns from data.
- **Big Data:** Extremely large datasets processed with advanced tools.
- **ETL:** Extract, Transform, Load (data pipeline process).
- **NLP:** Natural Language Processing for text analysis

Modern Data Ecosystem

Definition

The interconnected environment of tools, technologies, and processes for managing data.

Key Components

- **Data Sources:** IoT devices, databases, APIs, etc.
- **Data Storage:** Data lakes, cloud platforms, relational databases.
- **Data Processing:** Batch vs. real-time processing.
- **Data Analysis:** Advanced analytics and ML platforms.
- **Data Visualization:** Dashboards, reports, BI tools.

Data Science Lifecycle


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co

Phases

1. Problem Definition
2. Data Collection
3. Data Cleaning
4. Data Exploration
5. Model Building
6. Evaluation
7. Deployment

Visual: A circular or linear flow diagram with arrows connecting each phase.

Trends, Markets, and Applications of Data Science

Trends

- AI and Deep Learning.
- AutoML (Automated Machine Learning).
- Explainable AI (XAI).
- Data privacy and security (GDPR, CCPA).

Markets

- Health
- Finance
- Social Media
- E-commerce

Applications

- Predictive analysis
- Fraud detection
- NLP
- Recommendation systems

Tools and Technologies in Data Science

- **Programming Languages:** Python, R, SQL.
- **Libraries:** TensorFlow, PyTorch, Scikit-learn.
- **Data Visualization:** Tableau, Power BI, Matplotlib.
- **Big Data:** Hadoop, Spark.
- **Cloud Platforms:** AWS, Google Cloud, Azure.

Data Scientists and Their Roles

Role Description:

- Professionals who analyze and interpret complex data to solve problems.

Key Responsibilities:

- Collect and process data.
- Build predictive models.
- Communicate insights to stakeholders.

Skills Required:

- Statistical knowledge, programming, machine learning, domain expertise.

Data Scientist job Reality