

Introduction to Data Science

Introduction to Data Science

Key concepts, workflow, tools, and career pathways

1. What is Data Science?

Aspect	Description
Definition	An interdisciplinary field that extracts knowledge and insights from structured and unstructured data using statistical, mathematical, and computational techniques.
Core Disciplines	• Statistics & probability • Computer science & programming • Domain expertise (e.g., finance, healthcare, marketing)
Goal	Turn raw data into actionable information that supports decision-making, product development, or scientific discovery.

2. The Data Science Process (CRISP-DM Inspired)

1. Business / Problem Understanding

- Define objectives, success metrics, and constraints.
- Identify stakeholders and required deliverables.

2. Data Acquisition

- Sources: databases, APIs, web scraping, sensors, public datasets.
- Consider data licensing, privacy, and security.

3. Data Exploration & Cleaning

- Inspect data quality (missing values, outliers, inconsistencies).
- Perform **EDA** (Exploratory Data Analysis) with visualizations and summary statistics.
- Apply transformations: imputation, encoding, scaling, feature engineering.

4. Modeling / Analysis

Introduction to Data Science

- Choose appropriate algorithms (regression, classification, clustering, time series, deep learning).
- Split data (train/validation/test) and tune hyperparameters.
- Evaluate using metrics aligned with business goals (e.g., accuracy, RMSE, AUC, F1 score).

5. Interpretation & Communication

- Translate model outputs into insights.
- Use visual storytelling (charts, dashboards, reports).
- Explain model behavior (feature importance, SHAP values, LIME).

6. Deployment & Monitoring

- Package models as APIs, batch jobs, or embedded services.
- Set up monitoring for data drift, performance decay, and resource usage.
- Plan for maintenance, versioning, and reproducibility.

7. Feedback Loop

- Collect new data, user feedback, and business results.
- Iterate on earlier steps to improve the solution.

3. Fundamental Techniques

Category	Typical Methods	When to Use
Descriptive Statistics	Mean, median, variance, quantiles, correlation	Summarizing data, spotting trends
Inferential Statistics	Hypothesis testing, confidence intervals, ANOVA	Drawing conclusions about populations
Supervised Learning	Linear/Logistic regression, decision trees, random forests, gradient boosting, SVM, neural nets	Predicting a known target variable
Unsupervised Learning	K-means, hierarchical clustering, DBSCAN, PCA, t-SNE	Finding hidden structure, dimensionality reduction

Introduction to Data Science

Category	Typical Methods	When to Use
Time Series Analysis	ARIMA, Prophet, LSTM, seasonal decomposition	Forecasting with temporal data
Natural Language Processing	Tokenization, TF-IDF, word embeddings, transformers	Working with text data
Recommender Systems	Collaborative filtering, matrix factorization, content-based methods	Personalizing suggestions

4. Core Tools & Libraries

Domain	Popular Tools	Typical Use
Programming	Python, R	General data manipulation and modeling
Data Manipulation	pandas, dplyr, data.table	Cleaning, reshaping, aggregating data
Visualization	Matplotlib, Seaborn	