

Introduction to Data Science

Introduction to Data Science

Quick reference guide – concise, organized, and easy to skim.

1. What Is Data Science?

Aspect	Description
Definition	A multidisciplinary field that extracts knowledge and insights from structured and unstructured data using statistical, computational, and domain-specific techniques.
Core Goal	Turn raw data into actionable information for decision-making, product development, and strategic planning.
Key Disciplines	• Statistics & probability • Computer science & software engineering • Domain expertise (e.g., finance, health, marketing) • Machine learning & AI • Data engineering & visualization

2. The Data Science Workflow

1. Problem Formulation

- Define business or research question.
- Identify success metrics (accuracy, ROI, etc.).

2. Data Acquisition

- Sources: databases, APIs, web scraping, sensors, public datasets.
- Consider data licensing & privacy regulations (GDPR, CCPA).

3. Data Exploration & Cleaning

- Inspect data (summary stats, visual checks).
- Handle missing values, outliers, duplicate records.

Introduction to Data Science

- Perform type conversion, normalization, encoding.

4. Feature Engineering

- Create informative variables (aggregations, ratios, time-based features).
- Dimensionality reduction (PCA, t-SNE) when needed.

5. Modeling

- Choose algorithm family (regression, classification, clustering, deep learning).
- Split data (train/validation/test) ? avoid leakage.
- Hyperparameter tuning (grid search, Bayesian optimization).

6. Evaluation

- Metrics: accuracy, precision/recall, F1, ROC-AUC, RMSE, MAE, R², etc.
- Cross-validation, confusion matrix, residual analysis.

7. Deployment & Monitoring

- Package model (Docker, REST API, cloud service).
- Set up logging, drift detection, performance alerts.

8. Communication & Visualization

- Storytelling with charts, dashboards, and narrative.
- Tailor visualizations to audience (executives vs. technical peers).

3. Essential Tools & Languages

Category	Popular Options
Programming	Python (pandas, NumPy, scikit-learn, TensorFlow, PyTorch), R (tidyverse, caret, mlr)
Data Storage	SQL (PostgreSQL, MySQL), NoSQL (MongoDB, Cassandra), Data lakes (AWS S3, Azure Blob)
Big Data	Apache Spark, Hadoop, Dask
Visualization	Matplotlib, Seaborn, Plotly, ggplot2, Tableau, Power BI

Introduction to Data Science

Category	Popular Options
Version Control	Git + GitHub/GitLab/Bitbucket
Workflow & Automation	Jupyter Notebook/Lab, RStudio, Apache Airflow, Prefect, MLflow
Cloud Platforms	AWS (SageMaker, Redshift), GCP (BigQuery, AI Platform), Azure (ML Studio, Synapse)

4. Fundamental Concepts

4.1 Statistics & Probability

- **Descriptive stats:** mean, median, variance, percentiles.
- **Probability distributions:** Normal, Binomial, Poisson, Exponential.
- **Inference:** hypothesis testing, confidence intervals, p-values.
- **Bayesian thinking:** priors, posteriors, MAP estimation.

4.