What is Data Science?

Data Science is the process of **collecting, cleaning, analyzing, and interpreting data** to gain insights and make informed decisions.

It combines techniques from **statistics**, **computer science**, and **domain knowledge** to solve real-world problems.

Key Steps in Data Science

Step	Description
1. Data Collection	Gathering raw data from sources like websites, databases, APIs, IoT devices, and sensors.
2. Data Cleaning	Fixing errors, handling missing or inconsistent data — this takes up about **80% of a data scientist's time**.
3. Data Analysis	Using **statistical and visualization methods** to find trends, correlations, and patterns.

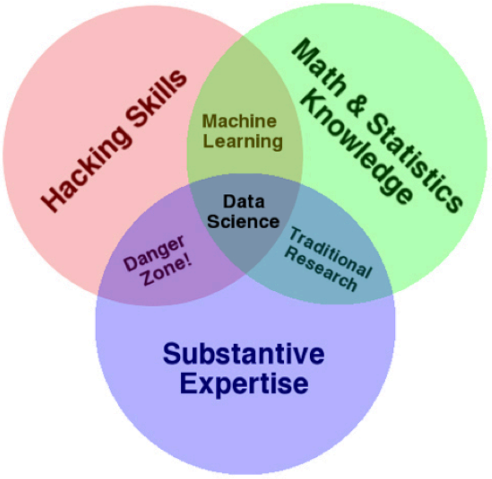
4. Model Building	Applying **machine learning algorithms** to predict outcomes or classify data.
5. Interpretation & Communication	Explaining insights clearly using **visuals, reports, and dashboards** for better decision-making.

Data Science Venn Diagram

The **Venn Diagram of Data Science** shows the intersection of three major skill areas:

Skill Area	Description
Mathematics & Statistics	Understanding data distributions, probability, and hypothesis testing.
Computer Science	Programming, automation, databases, and machine learning implementation.

Domain Expertise	Knowledge of the specific field (finance, healthcare, marketing, etc.) to interpret results meaningfully.



Data Science lies at the intersection of **Mathematics**, **Programming**, and **Domain Knowledge**.

The Data Science Lifecycle

The **Data Science Lifecycle** is a structured process to extract actionable insights from raw data.

It ensures each step — from problem understanding to deployment — is systematic and repeatable.

---

1. Problem Definition

- Clearly define the **goal** or **business question**.
- Understand **what needs to be predicted, optimized, or explained**.

---

2. Data Collection

- Gather **relevant and sufficient** data from reliable sources.

- Can include structured (databases) and unstructured (text, images) data.

---

3. Data Cleaning (Data Preprocessing)

- Handle missing values, remove duplicates, and correct inconsistencies.
- Transform data into a suitable format for analysis.
- Most **time-consuming** phase (around 70–80% of total effort).

---

4. Data Exploration (EDA – Exploratory Data Analysis)

- Visualize data to find **patterns, trends, and relationships**.
- Identify outliers, variable correlations, and potential biases.
- Tools: Matplotlib, Seaborn, Pandas Profiling.

---

5. Model Building

- Choose appropriate **machine learning algorithms** (e.g., regression, classification, clustering).
- Train models using historical data.
- Optimize parameters to improve accuracy.

---

6. Model Evaluation

- Assess performance using **metrics** like accuracy, precision, recall, RMSE, or F1-score.
- Use techniques like **cross-validation** to avoid overfitting.

---

7. Deployment

- Integrate the trained model into **real-world systems** or **applications**.
- Can involve APIs, dashboards, or cloud-based deployment.

8. Communication & Reporting

- Present insights using **visualizations, reports, or dashboards**.
- Translate technical results into **actionable business insights**.
- Communication is key — decision-makers rely on clear reporting.

---

9. Maintenance & Iteration

- Continuously **monitor and update** the model as new data comes in.
- Re-train models to maintain accuracy and reliability over time.

Summary

Stage	Purpose
Problem Definition	Understand what needs to be solved
Data Collection	Gather data from various sources
Data Cleaning	Prepare clean, usable data
EDA (Exploration)	Identify trends and patterns
Model Building	Train predictive models
Evaluation	Check accuracy and reliability
Deployment	Integrate model into use
Communication	Present insights clearly
Maintenance	Keep model updated and accurate

Cheat Sheet

Concept	Key Notes
Data Science	Extracts knowledge and insights from data
Core Skills	Statistics, Programming, Domain Knowledge
Popular Tools	Python, R, SQL, Pandas, NumPy,

	Matplotlib, Scikit-learn
Main Goal	Turn data into actionable insights
Lifecycle Phases	Define → Collect → Clean → Explore → Build → Evaluate → Deploy → Communicate → Maintain
EDA	Explore data visually to find hidden patterns
Machine Learning	Automates prediction and decision-making
Communication	Essential for conveying results effectively