1. Introduction to Data Science

> **Definition:** Data Science is an interdisciplinary field that combines statistics, mathematics, programming, machine learning and domain knowledge to extract meaningful insights and knowledge from structured and unstructured data.

1.1 Need for Data Science

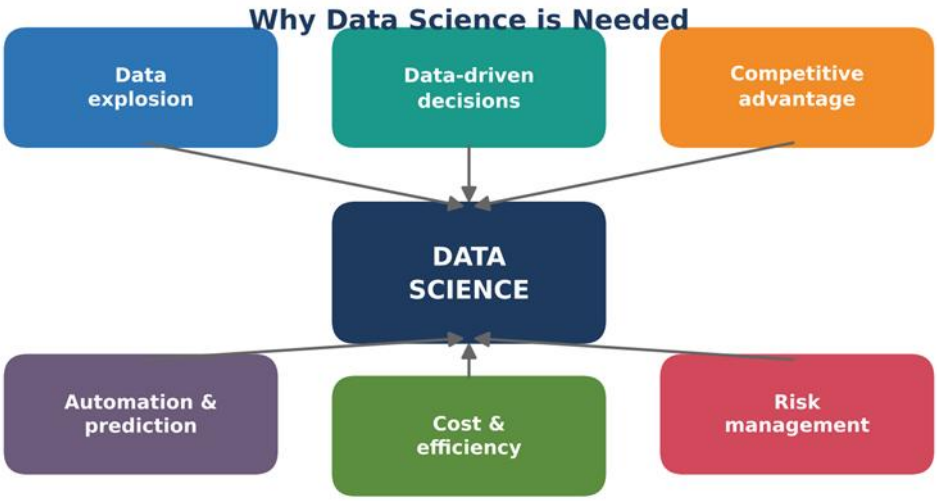

Fig 1.1: Reasons why Data Science is needed

- **Explosion of data:** social media, IoT and e-commerce generate zettabytes of data that need systematic analysis.
- **Data-driven decision making:** organizations replace intuition with evidence.
- **Competitive advantage:** personalization and demand forecasting (Amazon, Netflix).
- **Automation and prediction:** fraud detection, predictive maintenance, chatbots.
- **Cost reduction and efficiency:** optimized supply chains and resource usage.

- **Handling unstructured data:** text, images and video need advanced techniques.
- **Risk management:** credit scoring, insurance and security risk modelling.
- **Innovation:** enables new products such as recommendation engines and self-driving cars.

1.2 Scope of Data Science

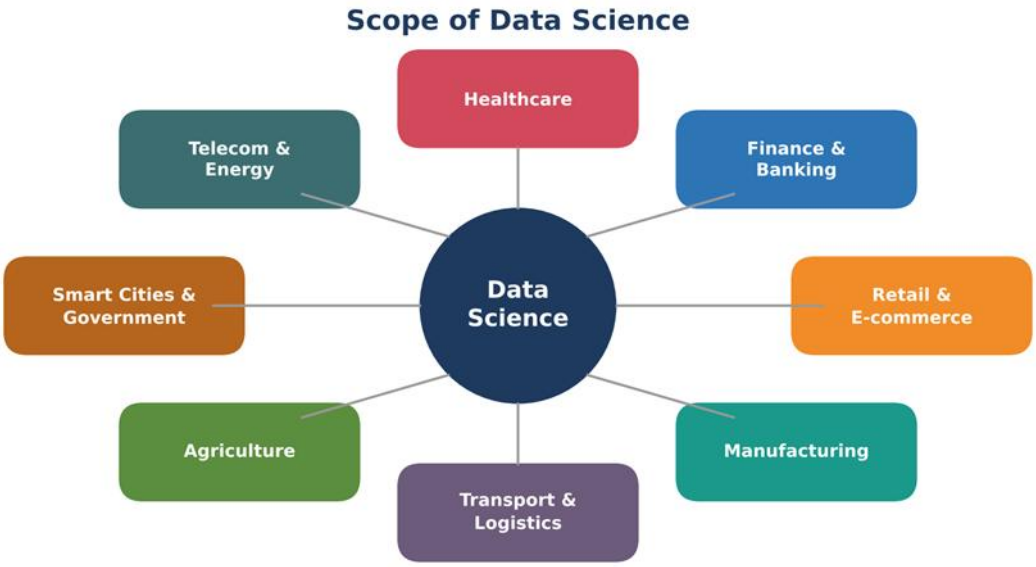

Fig 1.2: Application areas (scope) of Data Science

Domain	Applications
Healthcare	Disease prediction, medical imaging, drug discovery
Finance / Banking	Fraud detection, algorithmic trading, credit risk
Retail / E-commerce	Recommendation systems, customer segmentation
Manufacturing	Predictive maintenance, quality control
Transportation	Route optimization, traffic prediction, autonomous vehicles

Domain	Applications
Agriculture	Crop yield prediction, precision farming
Smart cities / Government	Policy analytics, crime prediction, utility management
Telecom / Energy	Churn prediction, smart grids

Career roles: Data Scientist, Data Analyst, Machine Learning Engineer, Data Engineer, BI Developer.

1.3 Data Science Lifecycle

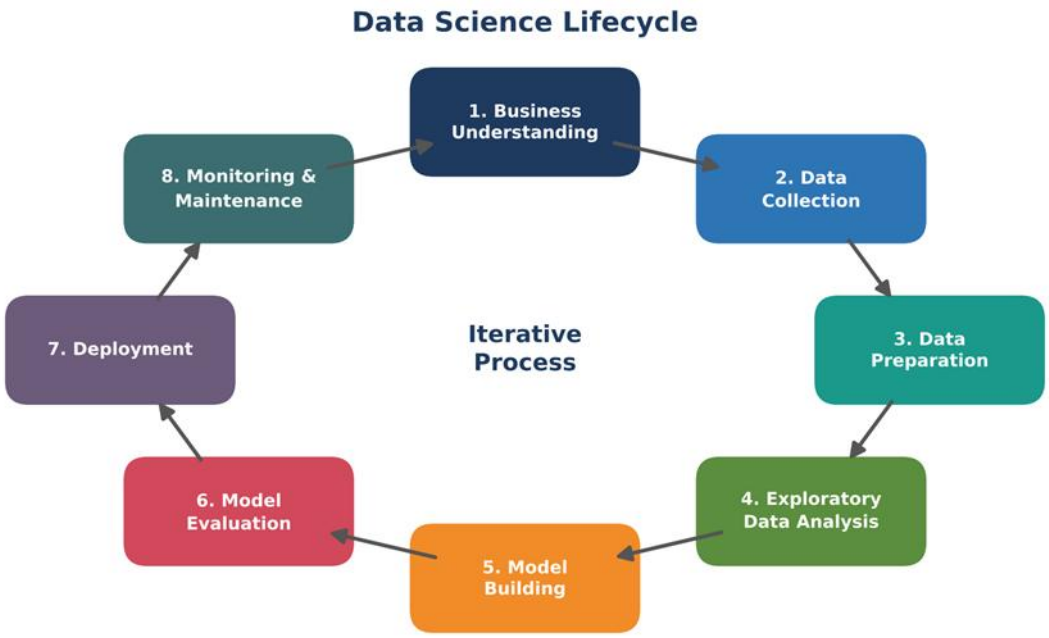

Fig 1.3: Data Science lifecycle (iterative)

Phase	Description
1. Business Understanding	Define the problem, objectives and success metrics (KPIs).

Phase	Description
2. Data Collection	Gather data from databases, APIs, sensors and web scraping.
3. Data Preparation	Handle missing values, duplicates and outliers; transform, normalize and encode. Takes about 60-80% of total effort.
4. Exploratory Data Analysis	Summary statistics, visualization, correlations and pattern discovery.
5. Model Building	Feature engineering; select and train algorithms (regression, classification, clustering).
6. Model Evaluation	Accuracy, precision, recall, F1-score, RMSE, cross-validation.
7. Deployment	Integrate into an application, API or dashboard.
8. Monitoring & Maintenance	Track model drift and retrain with new data.

Conclusion: The lifecycle is iterative; results from later stages often send the team back to earlier ones. Data Science turns raw data into decisions across every industry.