BASIC

Data science is an interdisciplinary field that uses statistics, programming, and AI/ML with domain expertise to extract actionable insights from data and support informed decision-making.

Difference between AI and ML and Data Science

Aspect	Artificial Intelligence (AI)	Machine Learning (ML)	Data Science
Primary Goal	Build systems that mimic human intelligence (reasoning, decision-making, perception)	Enable systems to learn patterns from data and improve automatically	Extract insights and knowledge from data to support decisions
Scope	Broadest field – includes ML, Deep Learning, NLP, CV, Robotics, Planning	Subset of AI focused only on learning algorithms	Interdisciplinary field combining statistics, ML, data engineering, and domain knowledge
Core Techniques	Rule-based systems, search algorithms, knowledge graphs, ML, DL, reinforcement learning	Supervised, unsupervised, reinforcement learning, neural networks	Statistics, EDA, data visualization, ML models, hypothesis testing
Output / Result	Intelligent behaviour (chatbots, autonomous agents, game-playing AI)	Predictive or adaptive models (recommendations, classification, forecasting)	Actionable insights, reports, dashboards, models
Data Dependency	May work with or without data (e.g., expert systems)	Strongly data-driven – performance improves with more data	Entirely data-centric – data collection, cleaning, and analysis are core

One-line Summary

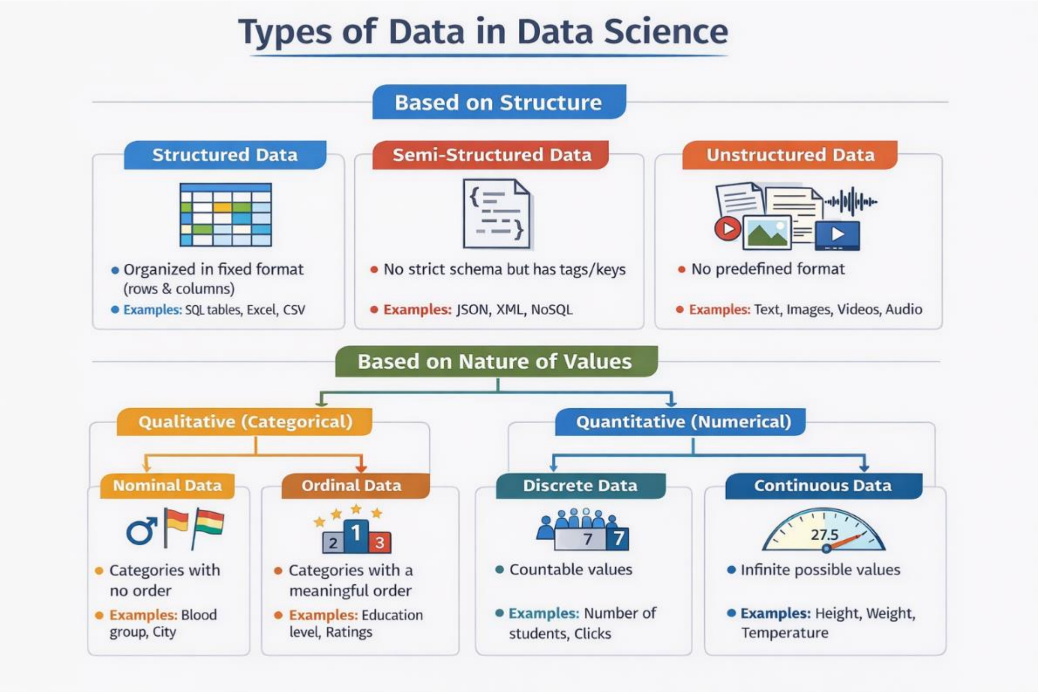
AI → “Make machines intelligent”

ML → “Let machines learn from data”

Data Science → “Turn data into decisions”

DATA

Data is a collection of raw facts or measurements that have little meaning on their own but become useful information when processed and analyzed.



1. Structured Data

Structured data is data that is highly organized and stored in a fixed schema (rows and columns).
It is easy to store, search, and analyze using traditional databases and SQL.

Examples: Excel sheets, CSV files, relational databases (MySQL, PostgreSQL)

---

2. Semi-Structured Data

Semi-structured data does not follow a strict table format but contains tags, keys, or markers that provide structure and hierarchy.

Examples: JSON files, XML files, HTML pages, NoSQL documents

---

3. Unstructured Data

Unstructured data has no predefined format or schema and is harder to analyze using traditional methods.
It often requires advanced techniques like NLP or computer vision.

Examples: Text documents, images, videos, audio recordings, social media posts

---

4. Qualitative (Categorical) Data

Qualitative data represents descriptive characteristics or labels rather than numerical values.
It is used to classify or group data into categories.

Examples: Gender, color, country, product type

---

5. Quantitative (Numerical) Data

Quantitative data consists of numerical values that represent counts or measurements and can be analyzed mathematically.

Examples: Age, income, temperature, sales amount

---

6. Discrete Data

Discrete data is a type of quantitative data that consists of countable, finite, or whole values.
It cannot take fractional values.

Examples: Number of students, number of purchases, defects count

---

7. Continuous Data

Continuous data can take any value within a given range, including decimals, and is typically measured rather than counted.

Examples: Height, weight, time, distance

---

8. Nominal Data

Nominal data is a type of categorical data where the categories have no natural order or ranking.

Examples: Blood group, city name, nationality

---

9. Ordinal Data

Ordinal data is categorical data where the categories have a meaningful order, but the differences between them are not numerically measurable.

Examples: Education level, customer satisfaction ratings (low–medium–high)

STEPS INVOLVED IN DATA SCIENCE PROCESS (DS lifecycle)

(learn each highlighted point and understand their explanation)

The data science process involves several steps, each with distinct objectives and methods for execution. Here's a breakdown of each step:

1. Business problem

Objective: Identify and understand the problem or business challenge.
Ways to Perform:

- Engage with stakeholders to define the problem clearly.
- Establish the objectives and key results (OKRs) to understand the desired outcome.
- Frame the problem in terms of data and analytics, such as classification, regression, or clustering.

2. Data Collection

Objective: Gather the data necessary to solve the problem.
Ways to Perform:

- **Structured Data:** Collect from databases, spreadsheets, or APIs.
- **Unstructured Data:** Collect from text, images, videos, or sensor data.
- **Internal Sources:** Extract data from internal company systems, logs, and databases.
- **External Sources:** Use publicly available datasets, third-party APIs, or purchase datasets from data vendors.
- Ensure data is relevant, up-to-date, and legally compliant (privacy laws, consent, etc.).

3. Exploratory Data Analysis (EDA)

Objective: Explore and analyze data to gain insights, identify patterns, and formulate hypotheses.
Ways to Perform:

- **Visualizations:** Use histograms, box plots, scatter plots, heatmaps, etc., to understand data distributions and relationships.
- **Summary Statistics:** Calculate basic statistics like mean, median, standard deviation, correlation coefficients, etc.
- **Check Data Quality:** Identify missing values, outliers, or anomalies in data that may affect analysis.

- **Understand Patterns:** Use clustering or dimensionality reduction techniques (e.g., PCA) to visualize high-dimensional data.

4. Data Cleaning and Preprocessing

Objective: Prepare the data for analysis by removing inconsistencies, correcting errors, and transforming it into a usable format.
Ways to Perform:

- **Handle Missing Data:** Remove, impute, or estimate missing values using methods like mean, median, or prediction models.
- **Remove Duplicates:** Identify and remove duplicate records.
- **Data Transformation:** Standardize or normalize values, convert data types (e.g., text to categories), and apply encoding techniques (e.g., One-Hot Encoding).
- **Feature Engineering:** Create new features that can improve model performance (e.g., date-time features, interaction terms).
- **Outlier Detection:** Detect and handle outliers that could skew analysis.

5. Feature Selection

Objective: Choose the most relevant features or create new ones that will contribute to the model's predictive power.
Ways to Perform:

- **Correlation Matrix:** Identify strongly correlated features and remove redundancies.
- **Feature Importance:** Use models (e.g., Random Forests, XGBoost) or statistical tests to rank features by importance.
- **Dimensionality Reduction:** Use PCA, t-SNE, or autoencoders to reduce the feature space.
- **Interaction Features:** Create new features based on combinations or transformations of existing features.

6. Model Selection and Training

Objective: Choose the appropriate machine learning model(s) and train them on the prepared data.
Ways to Perform:

- **Model Selection:** Choose algorithms based on the type of problem (e.g., linear regression for continuous targets, decision trees for classification, clustering algorithms for unsupervised tasks).
- **Cross-Validation:** Use techniques like k-fold cross-validation to avoid overfitting and assess model performance.
- **Hyperparameter Tuning:** Use grid search or random search to find the best hyperparameters.
- **Train/Test Split:** Divide data into training and testing sets to evaluate the model's generalizability.

7. Model Evaluation

Objective: Evaluate the trained model's performance using appropriate metrics.
Ways to Perform:

- **Accuracy Metrics:** Use metrics like accuracy, precision, recall, F1-score, and AUC-ROC for classification tasks.
- **Error Metrics:** For regression tasks, use RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), or R-squared.
- **Confusion Matrix:** For classification models, analyze true positives, false positives, true negatives, and false negatives.
- **Cross-Validation Scores:** Evaluate the stability and robustness of the model with cross-validation.
- **Performance Comparison:** Compare multiple models to identify the best-performing one.

8. Model Deployment

Objective: Implement the model in a real-world environment so it can generate predictions on new data.
Ways to Perform:

- **Deployment Platforms:** Deploy models on cloud services (e.g., AWS, GCP, Azure) or on-premises servers.
- **Integration with Applications:** Integrate models into business applications or products (e.g., recommendation systems, chatbots).
- **APIs and Microservices:** Expose models through APIs for real-time predictions.
- **Model Monitoring:** Continuously monitor model performance in production to detect data drift, concept drift, or performance degradation.

9. Model Maintenance and Monitoring

Objective: Ensure the model continues to perform well over time and adapt to changes in data.
Ways to Perform:

- **Performance Tracking:** Continuously monitor model performance and set thresholds for acceptable accuracy.
- **Retraining:** Periodically retrain models with fresh data to adapt to new trends or shifts in underlying patterns.
- **Feedback Loops:** Incorporate user feedback or new data sources to improve model accuracy.
- **Drift Detection:** Monitor for concept drift or data drift and trigger model retraining as needed.