# IDS-Unit-1

15 June 2025 07:30

Data Science

• Data Science is defined as the study of data, where it comes from, what it represents and the way by which it can be transformed into valuable inputs and resources to create business and IT strategies.

(or)

• Data Science is the study of data, it involves developing methods of recording, storing and analysing data to efficiently extract useful information.

• The goal of Data Science is to gain insights and knowledge from any type of data both structured or unstructured.

• Data Science is a field of study that combines domain expertise, programming skills and knowledge of mathematics and statistics to extract meaningful insights from data.

(or)

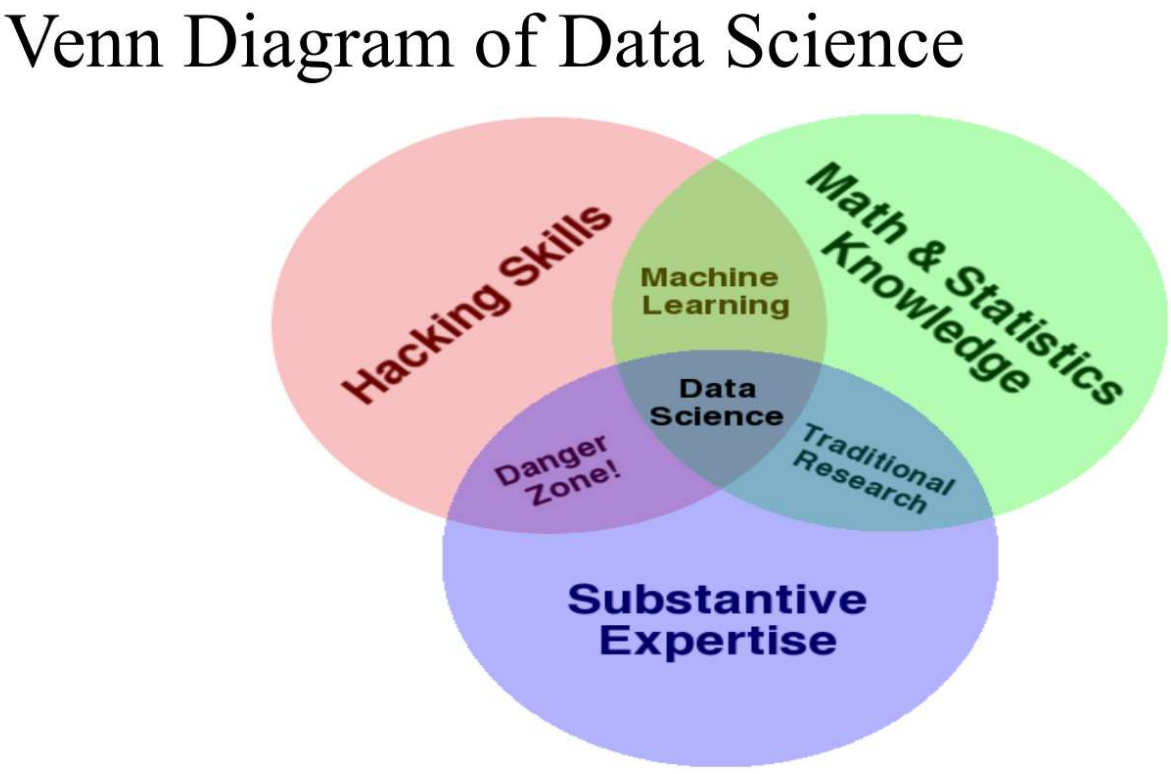
• Data Science is a multidisciplinary field of scientific data, Algorithms, Systems, Processes in order to extract insights from heterogeneous huge amount of data.

Understanding data science begins with three basic areas:

• Math/statistics: use of equations and formulas to perform analysis

• Computer programming: ability to use code to create outcomes on the computer

• Domain knowledge: This refers to understanding the problem domain (medicine, finance, social science, and so on)




Screen clipping taken: 15-06-2025 07:38

1. Identify and describe at least three real-world scenarios where Data Science is applied effectively.

• The great thing about data science is that it is not limited to one facet of society, one domain, or one department of a university; it is virtually everywhere.
• There are number of applications where data science can be used:
1. It helps in getting ideas of what customers would love to Purchase or eat according to their previous browsing history/purchase history.
2. Data Science also helps in making future predictions.
3. Data Science also helps in getting recommendations

1. Finance
2. Public policy
3. Politics
4. Healthcare
5. Urban planning
6. Education
7. Libraries

Finance
• Through capturing and analyzing new sources of data, building predictive
models and running real-time simulations of market events, they help the finance
industry obtain the information necessary to make accurate predictions.
• Data scientists in the financial sector may also partake in fraud detection and risk reduction.
• Banks and other loan sanctioning institutions collect a lot of data about the borrower in the initial "paperwork" process.
• Data science practices can minimize the chance of loan defaults via information such as customer profiling, past expenditures, and other essential variables that can be used to analyze the probabilities of risk and default.
• Data Science is used in financial institutions to identify the creditworthiness of potential customers

Public Policy:
---
• public policy is the application of policies, regulations, and laws to the problems of society through the actions of government and agencies for the good of a citizenry.
• Many branches of social sciences (economics, political science, sociology, etc.) are foundational to the creation of public policy.
• Data science helps governments and agencies gain insights into citizen behaviors that affect the quality of public life, including traffic, public transportation, social welfare, community wellbeing, etc.
• This information, or data, can be used to develop plans that address the betterment of these areas

Politics
• Politics is a broad term for the process of electing officials who exercise the policies that govern a state.
• Data scientists analyzed former US President Obama's 2008 presidential campaign success with Internet-based campaign efforts.
• Data scientists have been quite successful in constructing the most accurate voter targeting models and increasing voter participation.
• In 2016, the campaign to elect Donald Trump was a brilliant example of the use of data science in social media to tailor individual messages to individual people.
• The data analytics firm obtained data on approximately 87 million Facebook users from an academic researcher in order to target political ads during the 2016 US presidential campaign.

Healthcare:
---
• Healthcare is another area in which data scientists keep changing their research approach and practices.
• Medical Imaging
• Genemoics
• Drug Discovery
• Predictive Analysis
• Monitoring patients health :wearable devices
• Tracking and preventing diseases
• Providing virtual assistance

Urban Planning:
---
• Many scientists and engineers have come to believe that the field of urban planning is ripe for a significant – and possibly disruptive –change in approach as a result of the new methods of data science

Education:
---
• Schools, Colleges, Universities have large amount of student data such as academic records, grades, results, personal interests, cultural interests etc to handle.
The analysis of this data can help them in finding advanced methods for improving/enhancing the student learning.
• Improve adaptive learning
• Better parent involvement
• Better Assessment of Teachers
• Improve student performance
• Better Organization
• Regular updates in curriculum
• Student recruitment

Libraries:
---
• Data science is also frequently applied to libraries.

• Jeffrey M. Stanton has discussed the overlap between the task of a data science professional and that of a librarian. In his article, he concludes, “In the near future, the ability to fulfill the roles of citizenship will require finding, joining, examining, analyzing, and understanding diverse sources of data […] Who but a librarian will stand ready to give the assistance needed, to make the resources accessible, and to provide a venue for knowledge creation when the community advocate arrives seeking answers?”

==========

2. How does Data Science differ from Information Science? Highlight key distinctions with examples

Data Science and Information Science are closely related fields, but they differ in focus, methods, and applications. Here's a breakdown of their key distinctions with examples:

| **Aspect** | Data Science | Information Science |
|---|---|---|
| Primary Focus | Extracting insights and knowledge from data using statistical, computational, and machine learning techniques. | Organizing, managing, and retrieving information effectively for human use. |
| Goal | Predictive modeling, pattern recognition, decision-making. | Information organization, accessibility, and usability. |

## 2. Tools and Techniques

| **Aspect** | Data Science | Information Science |
|---|---|---|
| Techniques | Machine learning, data mining, statistical analysis, data visualization. | Information retrieval, classification, metadata design, human-computer interaction. |
| Tools | Python, R, SQL, TensorFlow, Pandas. | MARC, Dublin Core, XML, library systems, UX tools. |

Example:

- *Data Science:* Using Python and scikit-learn to build a recommendation engine.
- *Information Science:* Creating metadata schemas for digital archives.

## 3. Data vs. Information

| **Aspect** | Data Science | Information Science |
|---|---|---|
| Data Handling | Works with raw, structured, semi-structured, and unstructured data. | Focuses on processed data (information) and its organization. |
| Output | Models, predictions, dashboards. | Taxonomies, ontologies, information systems. |

Example:

- *Data Science:* Analyzing social media data to detect sentiment trends.
- *Information Science:* Structuring a knowledge base for a corporate intranet.

## 🧠 4. Interdisciplinary Roots

| **Aspect** | Data Science | Information Science |
|---|---|---|
| Origins | Statistics, computer science, mathematics. | Library science, cognitive science, information theory. |

Example:

- *Data Science:* A data scientist with a background in applied mathematics.
- *Information Science:* An information scientist trained in library and archival studies.

## 🚀 5. Application Domains

| **Aspect** | Data Science | Information Science |
|---|---|---|
| Industries | Finance, healthcare, e-commerce, tech. | Libraries, museums, academic institutions, UX design. |

Example:

- *Data Science:* Fraud detection in banking.
- *Information Science:* Designing user-friendly interfaces for digital libraries.

==========

3. Discuss how Data Science intersects with disciplines such as Statistics, Computer Science, and Business Analytics. Provide relevant examples.

Data Science is inherently interdisciplinary and intersects deeply with Statistics, Computer Science, and Business Analytics. Here's how each discipline contributes to and complements Data Science, along with relevant examples:

## 1. Statistics: The Foundation of Data Science

Intersection:

- Statistics provides the theoretical backbone for data analysis, hypothesis testing, and model validation.
- It helps in understanding data distributions, correlations, and drawing inferences.

Key Concepts Used:

- Probability theory
- Regression analysis
- Hypothesis testing
- Sampling methods

Example:

- A data scientist uses logistic regression to predict whether a customer will churn based on historical behavior.
- A/B testing is applied to evaluate the effectiveness of a new website layout.

### 2. Computer Science: The Engine Behind Data Science

Intersection:

- Computer Science enables the implementation of algorithms, data structures, and software systems for data processing.
- It supports scalable data handling, machine learning, and automation.

Key Concepts Used:

- Algorithms and data structures
- Programming (Python, R, SQL)
- Databases and data engineering
- Machine learning and AI

Example:

- Building a real-time recommendation system using collaborative filtering and deploying it via cloud infrastructure.
- Writing Python scripts to clean and transform large datasets for analysis.

### 3. Business Analytics: The Application Layer of Data Science

Intersection:

- Business Analytics focuses on using data to drive strategic decisions and improve business outcomes.
- It translates technical insights into actionable business strategies.

Key Concepts Used:

- KPI tracking
- Dashboard creation
- Predictive analytics
- ROI and cost-benefit analysis

Example:

- Using predictive modeling to forecast sales and optimize inventory.
- Creating interactive dashboards in Power BI or Tableau to monitor customer engagement metrics.

### Summary of Intersections

| Discipline | Contribution to Data Science | Example |
|---|---|---|
| Statistics | Data understanding & modeling | Predicting churn using regression |
| Computer Science | Data processing & algorithm design | Building a fraud detection system |
| Business Analytics | Strategic decision-making | Optimizing marketing campaigns |

==========

4. Classify the various types of data, and differentiate between structured and unstructured data using appropriate examples.

Data can be classified in several ways based on its nature, format, and how it's used. Here's a breakdown of the types of data and a comparison between structured and unstructured data with examples:

## Types of Data

### 1. Based on Measurement Scale

| Type | Description | Example |
|---|---|---|
| Nominal | Categorical data without order | Gender, colors, country names |
| Ordinal | Categorical data with a meaningful order | Customer satisfaction (low, medium, high) |

| | | |
|---|---|---|
| Interval | Numeric data with equal intervals, no true zero | Temperature in Celsius |
| Ratio | Numeric data with a true zero | Height, weight, age, income |

### 2. Based on Structure

| Type | Description | Example |
|---|---|---|
| Structured Data | Organized in rows and columns; easily searchable | Excel spreadsheets, SQL databases |
| Unstructured Data | No predefined format; harder to analyze | Emails, social media posts, videos |
| Semi-structured Data | Partially organized with tags or markers | JSON, XML, HTML files |

## VS Structured vs. Unstructured Data

| Feature | Structured Data | Unstructured Data |
|---|---|---|
| Format | Tabular (rows and columns) | Free-form (text, images, audio) |
| Storage | Relational databases (e.g., MySQL) | Data lakes, NoSQL databases |
| Ease of Analysis | Easy to query and analyze | Requires preprocessing and advanced tools |
| Examples | Sales records, employee databases | Tweets, YouTube videos, PDFs |
| Tools Used | SQL, Excel, BI tools | NLP, computer vision, Hadoop, Spark |

### Examples

- Structured:
  A table of customer data:
  | ID | Name | Age | Purchase Amount |
  |---|---|---|---|
  | 1 | Alice | 30 | $250 |
  | 2 | Bob | 45 | $180 |
- Unstructured:
  A customer review:
  *"I loved the product! The delivery was fast and the packaging was great."*

==========

5. What are the major challenges in working with unstructured data, and what techniques are commonly used to manage and process it?

Working with unstructured data presents several challenges due to its complexity, variability, and lack of predefined format. Here's a breakdown of the major challenges and the techniques commonly used to manage and process it:

## Major Challenges in Working with Unstructured Data

### 1. Lack of Structure

- No predefined schema makes it difficult to store and query.
- Example: Free-text reviews or social media posts.

### 2. Volume and Variety

- Unstructured data is often massive and comes in diverse formats (text, images, audio, video).
- Example: Petabytes of video content on YouTube.

### 3. Data Quality and Noise

- Contains irrelevant or redundant information.
- Example: Spam in email datasets or typos in user comments.

### 4. Complex Processing Requirements

- Requires advanced techniques like NLP, image recognition, or speech-to-text.
- Example: Extracting sentiment from customer feedback.

### 5. Storage and Scalability

- Traditional relational databases are not suitable.
- Needs scalable storage solutions like data lakes or NoSQL databases.

### 6. Security and Privacy

- Sensitive information may be embedded in unstructured formats.
- Example: Personal data in scanned documents or emails.

## 🛠️ Techniques to Manage and Process Unstructured Data

### 1. Natural Language Processing (NLP)

- Used for text analysis, sentiment detection, topic modeling.
- Tools: spaCy, NLTK, BERT, GPT models.

### 2. Text Mining and Information Extraction

- Extracts structured information from text.
- Example: Named Entity Recognition (NER) to identify names, dates, locations.

### 3. Computer Vision

- Analyzes image and video data.
- Tools: OpenCV, TensorFlow, YOLO, CNNs.

### 4. Speech and Audio Processing

- Converts audio to text and analyzes it.
- Tools: Google Speech-to-Text, Whisper, Kaldi.

### 5. Data Lakes and NoSQL Databases

- Store unstructured data efficiently.
- Tools: Hadoop, Amazon S3, MongoDB, Elasticsearch.

### 6. Machine Learning and Deep Learning

- Automates pattern recognition and classification.
- Example: Classifying emails as spam or not using supervised learning.
- ==========

6. Define data cleaning and elaborate on the key methods employed to clean raw datasets.

## What Is Data Cleaning?

Data cleaning (also known as data cleansing or data scrubbing) is the process of detecting and correcting (or removing) errors and inconsistencies in data to improve its quality and usability. It's a crucial step in data preprocessing before analysis, modeling, or visualization.

## 🔧 Key Methods of Data Cleaning

Here are the most commonly used techniques to clean raw datasets:

### 1. Handling Missing Data

- Methods:
- Remove rows or columns with too many missing values.
- Impute missing values using mean, median, mode, or predictive models.
- Example: Filling missing age values with the median age of the dataset.

### 2. Removing Duplicates

- Methods:
- Identify and drop duplicate rows based on key columns.
- Example: Removing repeated customer entries in a CRM database.

### 3. Correcting Data Types

- Methods:
- Convert columns to appropriate types (e.g., dates, integers, floats).
- Example: Changing a column from string to datetime format for time series analysis.

### 4. Standardizing Data

- Methods:
- Normalize formats (e.g., date formats, phone numbers).
- Convert text to lowercase or remove special characters.
- Example: Standardizing “NY”, “New York”, and “new york” to a single format.

### 5. Filtering Outliers

- Methods:
- Use statistical methods (e.g., z-score, IQR) to detect and remove or cap outliers.
- Example: Removing income values that are unrealistically high or low.

### 6. Validating Data Accuracy

- Methods:
- Cross-check values against known standards or rules.
- Example: Ensuring email addresses contain “@” and a domain.

### 7. Parsing and Tokenizing

- Methods:
- Break down complex fields into meaningful components.
- Example: Splitting full names into first and last names.

### 8. Handling Inconsistent Labels

- Methods:
- Unify categorical labels (e.g., “Male”, “male”, “M” → “Male”).
- Example: Standardizing product categories across different datasets.

==========

7 . Explain the concepts of Data Transformation and Data Reduction. How do these processes contribute to enhancing the performance of data models?

Both Data Transformation and Data Reduction are essential preprocessing steps in data science and machine learning. They help improve the efficiency, accuracy, and interpretability of models.

### 🔧 1. Data Transformation

Definition:

Data transformation involves converting data into a suitable format or structure for analysis or modeling.

#### ☑ Common Techniques:

- Normalization/Standardization: Scaling features to a common range (e.g., 0–1 or z-scores).
- Encoding Categorical Variables: Converting categories into numerical values (e.g., one-hot encoding).
- Log Transformation: Reducing skewness in data.
- Aggregation: Summarizing data (e.g., daily to monthly sales).
- Feature Engineering: Creating new features from existing ones.

#### 📈 Contribution to Model Performance:

- Improves model convergence and training speed.
- Reduces bias from dominant features.
- Makes data compatible with algorithms (e.g., tree-based models vs. linear models).

Example:

Transforming income data using a log scale to reduce the impact of extreme values.

### 📉 2. Data Reduction

Definition:

Data reduction aims to reduce the volume or dimensionality of data while preserving its essential characteristics.

### Common Techniques:

- Dimensionality Reduction:
- PCA (Principal Component Analysis): Projects data into fewer dimensions.
- t-SNE, UMAP: For visualization and clustering.
- Feature Selection:
- Removing irrelevant or redundant features.
- Techniques: correlation analysis, recursive feature elimination.
- Sampling:
- Using a representative subset of data.
- Aggregation:
- Grouping data to reduce granularity.

### Contribution to Model Performance:

- Reduces overfitting by eliminating noise.
- Speeds up training and inference.
- Enhances model interpretability.

Example:

Using PCA to reduce 100 image features to 10 principal components for faster classification.

### Summary Table

| **Aspect** | Data Transformation | Data Reduction |
|---|---|---|
| Goal | Make data suitable for modeling | Reduce data size or complexity |
| Focus | Format, scale, structure | Volume, dimensionality |
| Techniques | Normalization, encoding, log transform | PCA, feature selection, sampling |
| Benefits | Better model accuracy and compatibility | Faster training, less overfitting |

From <https://copilot.cloud.microsoft/?fromcode=cmc&redirectid=523035E14A4C4BF0AF19CB6097DA8817&auth=2>

What is Data?

• Webster's defines data as a plural form of datum as "something given or admitted especially as a basis for reasoning or inference.

There is also often a debate about what is the difference between data and information. In fact, it is common to use one to define the other (e.g., "data is a piece of information").

Measurement of Data

• Bit(Binary Digit) : A bit is a value of either a 1 or 0 (on or off).

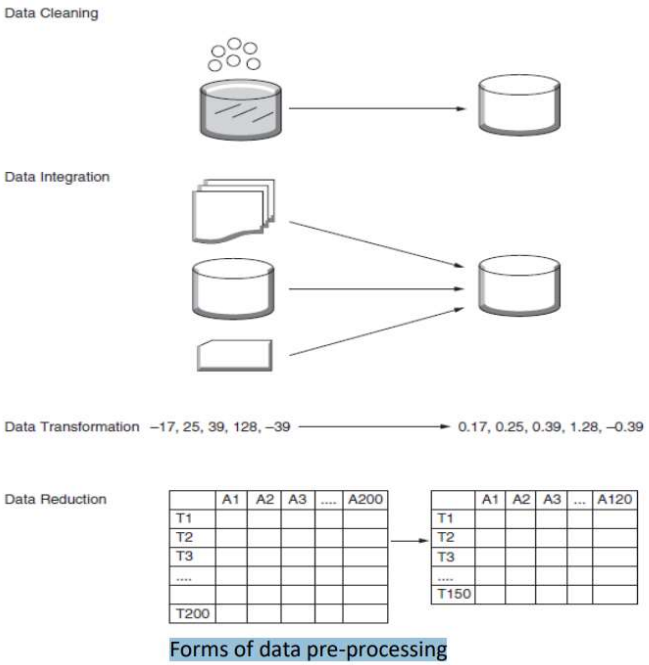
• Nibble : A Nibble is 4 bits.
• Byte : a Byte is 8 bits. 1 character, e.g. "a", is one byte.
• Kilobyte (KB) : 1024 bytes
• Megabyte (MB) : 1024 KB
• Gigabyte (GB) : 1024 MB(2^30 bytes)
• Terabyte (TB) : 1024 GB (2^40 bytes)
• Petabyte (PB) :1024 TB (2^50 bytes)
• Exabyte (EB) : 1024 PB (2^60 bytes)
• Zettabyte (ZB) : 1024 EB (2^70 bytes)
• Yottabyte (YB) : 1024 ZB (2^80 bytes)

5 Stages of Data
1. Capture : Data Acquisition, Extraction
2. Maintain: Data Warehousing, Data Cleaning, Data staging, Data Processing, Data Architecture
3. Process: Data Mining, Classification, modelling
4. Analyse: Predictive analysis, Regression techniques, Qualitative analysis
5. Communicate: Report generation, Visualization, Business Intelligence, Decision making

==========

Forms of data pre-processing

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction



Forms of data pre-processing

Screen clipping taken: 15-06-2025 17:58

==========

7. Explain the concepts of Data Transformation and Data Reduction.
How do these processes contribute to enhancing the performance of data models?

Data Transformation
• Data must be transformed so it is consistent and readable (by a system). The following five processes may be used for data transformation.
1. Smoothing: Remove noise from data.
2. Aggregation: Summarization, data cube construction.
3. Generalization: Concept hierarchy climbing.
4. Normalization: Scaled to fall within a small, specified range and aggregation. Some of the techniques that are used for accomplishing normalization (but we will not be covering them here) are:

a. Min–max normalization.
b. Z-score normalization.
c. Normalization by decimal scaling.
5. Attribute or feature construction.
a. New attributes constructed from the given ones.

Data Reduction
• Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained.
• One example of a large dataset that could warrant reduction is a data cube.
• Data cubes are multidimensional sets of data that can be stored in a spreadsheet.
• A data cube could be in two, three, or a higher dimension. The two of the most common techniques used for data reduction.
1. Data Cube Aggregation
2. Dimensionality Reduction

Data Discretization
• We are often dealing with data that are collected from processes that are continuous, such as temperature, ambient light, and a company's stock price.
• But sometimes we need to convert these continuous values into more manageable parts. This mapping is called discretization.
• In undertaking discretization, we are also essentially reducing data.
• There are three types of attributes involved in discretization:
a. Nominal: Values from an unordered set
b. Ordinal: Values from an ordered set
c. Continuous: Real numbers

==========

Forms of data pre-processing

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction