Data Engineering Techniques (AID30010) — Unit 1

Introduction to Data Engineering and Preprocessing

Table of Contents

---

1. Data Engineering — Definition

1. Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information.
2. This information supports downstream use cases such as analysis and machine learning.
3. Data engineering is the intersection of data management, data architecture, and software engineering.
4. Data engineers manage the data engineering lifecycle — starting from ingestion and ending with serving data for use cases like analysis or machine learning.
5. The field has become increasingly important as organizations generate massive volumes of data and need reliable pipelines to make that data useful.

---

2. Key Concepts

Data Governance (DG)

1. Data Governance is the process of managing data availability, usability, integrity, and

security in an organization.
2. It defines rules, policies, and standards for handling and managing data properly across the organization.
3. It ensures that data is accurate, secure, and accessible only to authorized users.
4. Good data governance helps organizations comply with regulations and builds trust in the data being used.
5. It also defines who owns each dataset and who is responsible for maintaining its quality and accuracy.

Data Wrangling

1. Data Wrangling is the process of cleaning and transforming raw data into a usable format.
2. It involves removing errors, handling missing values, and correcting inconsistencies in the dataset.
3. It combines and organizes multiple complex datasets to make them easier to analyze together.
4. Data wrangling is often the most time-consuming part of the data process, taking up a large portion of a data engineer's or analyst's time.
5. The output of data wrangling is clean, structured data that is ready to be used in analysis or machine learning models.

---

3. Roles in the Data Engineering Ecosystem

Data Architect

1. Designs the overall data infrastructure and architecture of an organization.
2. Defines how data will be collected, stored, and managed across the organization.
3. Ensures the system is scalable, secure, and efficient for current and future needs.
4. Creates and maintains the blueprint that guides how all data systems connect and work together.
5. Works closely with data engineers and business teams to align technical design with business goals.

Data Engineer

1. Collects and processes raw data from different sources such as databases, APIs, and files.
2. Builds data pipelines and ETL processes to clean and transform data into usable formats.
3. Prepares reliable, high-quality data for analytics and machine learning use cases.
4. Deals with raw data that may contain human, machine, or instrument errors and fixes them.

5. Ensures data is consistently available, accessible, and in the right format for downstream users.

QA Engineer

1. Tests data systems and pipelines to ensure quality and accuracy at every stage.
2. Identifies bugs, errors, and inconsistencies in the system before they reach end users.
3. Verifies that the system meets required standards and performance benchmarks.
4. Designs test cases and validation checks to catch incorrect or incomplete data early.
5. Works alongside engineers to ensure the output data is trustworthy and production-ready.

Product Managers (PMs)

1. Define the product vision, strategy, and roadmap for the data product or platform.
2. Connect business goals, user needs, and technology teams to keep everyone aligned.
3. Manage the overall lifecycle of the product from idea to launch to improvement.
4. Act as the central hub between stakeholders, engineers, and analysts to prioritize work.
5. Ensure the product delivers real value to users and meets the goals of the organization.

Data Analysts

1. Analyze past and present data to understand trends and patterns in the business.
2. Create reports, dashboards, and visualizations to communicate insights clearly.
3. Help organizations understand what happened and why certain outcomes occurred.
4. Work with cleaned and processed data delivered by data engineers.
5. Answer specific business questions using queries, statistics, and data exploration techniques.

Data Scientists

1. Use cleaned data to build predictive models that forecast future outcomes.
2. Apply machine learning and statistical techniques to solve complex business problems.
3. Predict future trends and patterns from data using advanced algorithms.
4. Receive data that has already been cleaned and prepared by data engineers.
5. Communicate findings and model results to business teams to support data-driven decisions.

Stakeholders

1. Include owners, managers, employees, customers, shareholders, and suppliers.

2. Define business goals and requirements that guide what data needs to be collected and analyzed.
3. Use insights from data to make better business decisions and set future strategy.
4. May not be technical but rely heavily on dashboards and reports produced by data teams.
5. Their feedback and needs drive what the data engineering and analytics teams prioritize.

---

4. Data Engineering Lifecycle

1. Data Generation

1. Data is created from different sources such as applications, websites, sensors, databases, and user activities.
2. This is the starting point of the data engineering lifecycle — no pipeline exists without raw data first.
3. The volume and variety of data generated keeps growing with the rise of IoT, social media, and digital systems.

2. Data Ingestion

1. The generated data is collected and imported into the system using pipelines or tools like APIs, streaming systems, or batch processing.
2. Ingestion can be done in real time (streaming) for fast-moving data, or in batches for large periodic data loads.
3. This step is critical because poor ingestion leads to missing or delayed data in the rest of the pipeline.

3. Data Storage

1. The collected data is stored in data warehouses, data lakes, or databases so it can be accessed later for processing.
2. The choice of storage depends on the type of data — structured data goes to warehouses, raw data goes to lakes.
3. Storage must be reliable, scalable, and cost-effective to handle growing data volumes over time.

4. Data Transformation

1. Raw data is cleaned, filtered, and converted into a useful format using ETL/ELT processes so it becomes suitable for analysis.

2. This step handles missing values, removes duplicates, and standardizes formats across different data sources.
3. Transformed data is consistent, accurate, and ready for analytics, reporting, or machine learning models.

5. Data Serving / Consumption

1. The processed data is delivered to end users for analytics, reporting, dashboards, or machine learning models.
2. This is the final stage where the value of all the previous steps becomes visible to the business.
3. Different users consume data in different ways — analysts use dashboards, scientists use model pipelines, and executives use reports.

---

5. What is Data Mining?

Definition

1. Data Mining is the process of extracting useful patterns and knowledge from large amounts of data.
2. It is also known as Knowledge Discovery in Databases (KDD).
3. Data mining analyzes data from sources such as databases, text, web, images, and transactions.
4. The patterns discovered must be interesting — meaning they must be valid, novel, useful, and understandable.
5. Simple search or query processing is not considered data mining — the patterns must be non-trivial and previously unknown.

Characteristics of Patterns

1. The discovered patterns must be valid and accurate based on the data.
2. They should be novel or previously unknown to the user or the system.
3. They must be useful and help in making better business or research decisions.
4. They must be understandable so that humans can interpret and act on them.
5. Patterns that are trivial or already obvious do not count as data mining results.

Why Data Mining is Important

1. Businesses generate huge volumes of data due to rapid computerization — from terabytes to petabytes to zettabytes.

2. Data mining helps discover patterns, trends, and relationships hidden inside this large data.
3. It supports better business decisions and enables automated prediction of future behaviors.
4. It allows automated discovery of previously unknown patterns without manually inspecting every record.
5. Without data mining, the massive amounts of data collected would remain unused and provide no value.

Applications of Data Mining

1. Market analysis and business intelligence — finding customer buying patterns and preferences.
2. Trend prediction and behavior analysis — predicting which customers are likely to leave or buy next.
3. Healthcare — discovering patterns in patient data to improve diagnosis and treatment.
4. E-commerce and finance — detecting fraud, recommending products, and forecasting sales.
5. Other specialized forms include: text mining, web mining, audio/video mining, and social network mining.

Goal of Data Mining

1. The main goal is Knowledge Discovery from databases (KDD) — turning raw data into actionable knowledge.
2. It provides efficient tools and techniques to analyze very large datasets automatically.
3. It converts raw data into useful knowledge for decision making in business and research.
4. It enables automated analysis that would be impossible to do manually at large scale.
5. Example: discovering that 80% of customers who buy cheese and milk also buy bread — written as: Cheese, Milk → Bread.

---

6. KDD (Knowledge Discovery in Databases) Process

KDD Process — Overview

1. KDD is the process of extracting useful knowledge from large datasets.
2. It converts raw data into meaningful patterns and information through a series of steps.
3. Data mining is a core step inside the KDD process, not the entire process itself.
4. The KDD process starts from understanding the problem and ends with presenting discovered knowledge.

5. Each step feeds into the next — the quality of earlier steps directly affects the quality of final results.

Step 1 — Understanding the Domain

1. Understand the application domain, relevant background knowledge, and the goals of the end user.
2. Identify the goals of the end users or define the specific business problem to be solved.
3. Select the target dataset or choose a subset of variables relevant to the analysis.
4. This step ensures the entire process is focused on answering the right question from the start.
5. Without a clear understanding of the domain, the patterns discovered may be irrelevant or misleading.

Step 2 — Data Cleaning and Preprocessing

1. Remove noise, errors, and outliers from the data to improve its quality.
2. Handle missing or inconsistent values using appropriate methods like mean imputation or deletion.
3. Prepare the data to make it accurate and suitable for analysis.
4. This step also accounts for time sequence information and known changes in the data over time.
5. Poorly cleaned data leads to incorrect patterns — "No quality data, no quality mining results."

Step 3 — Data Reduction and Projection

1. Select important features or attributes from the dataset that are relevant to the goal.
2. Reduce the number of variables using dimensionality reduction techniques to simplify the data.
3. Simplify the data while keeping all the important information and relationships intact.
4. This step helps in reducing computation time and making the data mining step more efficient.
5. Techniques include attribute selection, principal component analysis, and concept hierarchy generation.

Step 4 — Choosing the Data Mining Task

1. Decide the type of analysis needed based on the goal of the problem.
2. Tasks may include classification, clustering, regression, association rule mining, or anomaly detection.

3. The task depends entirely on the goal — whether you want to predict, group, find patterns, or detect outliers.
4. Choosing the wrong task leads to results that do not answer the original business question.
5. This step bridges understanding the problem and technically solving it.

Step 5 — Choosing the Algorithm

1. Select the appropriate data mining algorithm based on the task chosen in the previous step.
2. Decide the model and parameters to be used — different algorithms suit different data types and sizes.
3. Match the algorithm with the overall KDD process criteria and objectives.
4. Multiple algorithms may be tried and compared to find the one that gives the best results.
5. The choice of algorithm affects the accuracy, speed, and interpretability of the results.

Step 6 — Data Mining

1. Apply algorithms to discover patterns and relationships in data.
2. Examples include decision trees for classification, k-means for clustering, and regression for prediction.
3. This step extracts useful patterns such as rules, clusters, or trends from the dataset.
4. The patterns found here are in raw form and need to be evaluated and interpreted in the next step.
5. This is the core step of the entire KDD process — the actual knowledge extraction happens here.

Step 7 — Interpretation and Knowledge Presentation

1. Interpret the discovered patterns to understand their real-world meaning and relevance.
2. Evaluate whether the patterns are truly useful, meaningful, and not just statistical noise.
3. Present the knowledge using visualization tools, reports, or dashboards for easy understanding.
4. Consolidate the discovered knowledge and resolve any conflicts with previously known knowledge.
5. The final goal is to present clear, actionable insights that users and decision makers can act upon.

---

7. Types of Data — Structured, Semi-Structured, Unstructured

Structured Data

1. Structured data is data that is organized in a fixed format with rows and columns, similar to a table.
2. It follows a predefined schema, meaning every data entry must follow the same structure and format.
3. It is easily machine-readable and can be analyzed using databases and data analysis tools.
4. Structured data is commonly stored in relational database management systems (RDBMS).
5. Because of its organized nature, it requires minimal preprocessing before analysis.

Examples: Excel spreadsheets, CSV files, SQL database tables.

Semi-Structured Data

1. Semi-structured data has some level of organization but does not follow a strict table format.
2. It uses tags, markers, or metadata to create hierarchy and structure in the data.
3. The structure is flexible, meaning the order and size of data elements may vary.
4. Semi-structured data requires some preprocessing and transformation before analysis.
5. This type of data became very common with the growth of the World Wide Web and web applications.

Examples: JSON files, XML files, HTML documents.

Unstructured Data

1. Unstructured data does not have any predefined format or organizational structure.
2. It can exist in multiple formats such as text, images, audio, video, or documents.
3. It is not easily machine-readable and usually requires heavy preprocessing before analysis.
4. Most of the data generated today on the internet and digital platforms is unstructured.
5. Advanced technologies like data mining, machine learning, and AI are often used to analyze unstructured data.

Examples: Images (.jpeg, .png), videos (.mp4), audio files (.mp3, .wav), PDFs and text documents.

Comparison Table

Feature	Structured Data	Semi-Structured Data	Unstructured Data
Definition	Organized in a fixed tabular structure	Partial organization using tags or markers	No predefined structure
Data Organization	Highly organized with rows and columns	Moderately organized with flexible structure	No clear organization

Feature	Structured Data	Semi-Structured Data	Unstructured Data
Schema	Strict schema must be followed	Flexible schema	No schema
Ease of Analysis	Easy to analyze using database tools	Requires some preprocessing	Requires significant preprocessing
Examples	Excel, CSV, relational databases	JSON, XML, HTML	Images, videos, audio, text documents

8. Types of Data — Statistical / Analytical Data Types

Qualitative (Categorical) Data

1. Qualitative data refers to data that cannot be measured numerically but is categorized based on qualities or attributes.
2. It describes characteristics or categories such as gender, color, or type of product.
3. This type of data is usually expressed in the form of text, labels, symbols, images, or categories.
4. Qualitative data is often used in market research, surveys, and social studies.
5. Examples include gender (male/female), hair color, nationality, customer satisfaction level.

Nominal Data

1. Nominal data is a type of qualitative data where values represent names or labels of categories.
2. There is no meaningful order or ranking among the categories.
3. The numbers assigned to nominal data are only labels and have no mathematical meaning.
4. Nominal data is mainly used for classification or grouping of data.
5. Examples include hair color, blood group, gender, city names, product categories.

Ordinal Data

1. Ordinal data is categorical data that has a meaningful order or ranking between values.
2. Although the categories can be ranked, the exact difference between them is not known.
3. Ordinal data is often used in rating scales and surveys.
4. Arithmetic operations such as addition or subtraction cannot be meaningfully performed.
5. Examples include customer satisfaction levels (poor, average, good), class ranks, education level.

Quantitative Data

1. Quantitative data refers to data that can be measured or counted numerically.
2. It represents numerical values that can be used for statistical analysis.
3. Quantitative data answers questions such as how many, how much, or how often.
4. It can be analyzed using mathematical and statistical techniques.
5. Examples include height, weight, number of students, temperature, income.

Discrete Data

1. Discrete data consists of distinct and countable values.
2. The values are usually whole numbers or integers.
3. Discrete data cannot be divided into smaller fractional values.
4. It represents countable quantities.
5. Examples include number of students in a class, number of cars in a parking lot, number of books.

Continuous Data

1. Continuous data represents values that can take any value within a range.
2. These values are usually expressed as decimal or fractional numbers.
3. Continuous data represents measurable quantities such as height, weight, or temperature.
4. It can be divided into infinitely smaller parts.
5. Examples include height of a person, temperature, time, distance.

Comparison — Discrete vs. Continuous

Feature	Discrete Data	Continuous Data
Nature of data	Separate and distinct values	Measurable values within a range
Type of values	Whole numbers or integers	Decimal or fractional values
Measurement	Countable quantities	Measurable quantities
Representation	Often represented using bar graphs	Usually represented using histograms or line graphs
Examples	Number of students, number of cars	Height, temperature, weight

9. Time Series Data

1. Time series data is data collected at regular time intervals.
2. The time interval must be equal and clearly defined.
3. It is mainly used to analyze trends and patterns over time.
4. Time series data helps in forecasting future values based on past data.
5. Examples include daily stock prices, monthly sales data, yearly population growth.

Characteristics of Time Series Data

1. **Equal Interval** — Data must always be collected at equal and consistent time intervals, not at random or uneven gaps.
2. **Trend** — A long-term increase or decrease in the data over time is called an uptrend or downtrend respectively.
3. **Seasonality** — Patterns that repeat regularly over a fixed period, such as higher sales in summer or festive seasons.
4. **Cyclicity** — Fluctuations that rise and fall over irregular time intervals, such as business cycles or economic booms.

---

10. Geographical Data

1. Geographical data refers to information that describes the physical location and features of places on Earth.
2. It includes information about cities, roads, rivers, mountains, and other geographic features.
3. This data is widely used in mapping, urban planning, transportation, and environmental studies.
4. Geographical data can be represented in 2D maps or 3D digital models.
5. Tools such as Geographic Information Systems (GIS) are used to store, analyze, and visualize geographical data.

Examples: Location of cities and towns, road networks, satellite maps, GPS location data.

---

11. Measures of Central Tendency

Measures of central tendency are methods used to find the central or typical value of a dataset. They help to summarize a large set of data into one representative value. The main measures are Mean, Median, and Mode.

Mean

1. Mean is the average value of a dataset.
2. It is calculated by adding all values and dividing by the number of values.
3. The mean uses every value in the dataset in the calculation.
4. It is commonly used for numerical and continuous data.
5. The mean can be affected by extreme values (outliers).

Formula: Mean = ΣX / N
Example: Data: 2, 4, 6, 8, 10 → Mean = (2+4+6+8+10) / 5 = **6**

Median

1. Median is the middle value of a dataset when the data is arranged in order.
2. It divides the dataset into two equal parts.
3. If the number of observations is odd, the middle value is the median.
4. If the number of observations is even, the median is the average of the two middle values.
5. The median is not strongly affected by extreme values.

Example (Odd): Data: 2, 5, 6, 7, 9 → Median = **6**
Example (Even): Data: 2, 4, 6, 8 → Median = (4+6) / 2 = **5**

Mode

1. Mode is the value that occurs most frequently in a dataset.
2. A dataset may have one mode, more than one mode, or no mode.
3. Mode is useful for both numerical and categorical data.
4. It helps to identify the most common value in the dataset.
5. Sometimes the mode may not represent the center of the dataset well.

Example: Data: 5, 4, 2, 3, 2, 1, 5, 4, 5 → Mode = **5** (appears 3 times)

How to Select the Correct Measure

1. If the data has a symmetrical distribution, all three measures can be used, but mean is usually preferred because it uses all values.
2. If the data is skewed or contains extreme values (outliers), the median is the better measure since it is less affected.
3. If the original raw data is available and you want to find the most typical value, both median and mode work well.
4. If the data is categorical or qualitative (like colors or names), the mode is the only appropriate measure.

5. In general, mean gives the most mathematical precision, median gives the most robustness, and mode gives the most common occurrence.

Comparison Table

Feature	Mean	Median	Mode
Definition	Average of all values	Middle value of ordered data	Most frequent value
Data used	Uses all values	Uses middle position	Uses most repeated value
Effect of outliers	Highly affected	Less affected	Usually not affected
Type of data	Numerical data	Numerical data	Numerical and categorical data
Main purpose	Find overall average	Find central value	Find most common value

---

12. Measures of Dispersion

Measures of dispersion are used to show how spread out the data values are in a dataset. While measures of central tendency give the center of the data, measures of dispersion show how far the values are from the center.

Range

1. Range is the difference between the highest value and the lowest value in a dataset.
2. It is the simplest measure of dispersion.
3. Range helps us understand the overall spread of the data.
4. It uses only two values: maximum and minimum values of the dataset.
5. Range can be strongly affected by extreme values (outliers).

Formula: Range = Maximum Value − Minimum Value
Example: Finishing times: 16, 21, 28, 12, 30, 21, 24 → Range = 30 − 12 = **18**

Variance

1. Variance measures how far the data values are from the mean.
2. It calculates the average of the squared differences between each value and the mean.
3. Variance helps to understand the spread or variability of the dataset.
4. A larger variance means the data is more spread out.
5. The unit of variance is the square of the unit of the original data.

Formula: Variance = $\Sigma(x - \bar{x})^2 / N$

Standard Deviation

1. Standard deviation shows how far the data values are from the mean on average.
2. It is calculated as the square root of variance.
3. Standard deviation is always a positive value.
4. A large standard deviation means the data values are widely spread.
5. A small standard deviation means the data values are close to the mean.

Formula: SD = √Variance

Comparison Table

Feature	Range	Variance	Standard Deviation
Definition	Difference between maximum and minimum value	Average squared deviation from the mean	Square root of variance
Data used	Uses only two values (max and min)	Uses all data values	Uses all data values
Complexity	Very simple to calculate	More complex calculation	Moderate calculation
Effect of outliers	Highly affected	Affected by extreme values	Affected but more interpretable
Unit	Same unit as data	Square of data unit	Same unit as data

---

13. Data Preprocessing — Overview

1. Data preprocessing is the process of converting raw data into a clean and understandable format.
2. It is an important step in data mining and machine learning because raw data cannot be used directly.
3. The process checks the quality and reliability of the data before analysis.
4. It prepares the data so that data mining or machine learning algorithms can work properly.
5. The main goal of data preprocessing is to improve the quality of data and make it suitable for analysis.

Why Data Preprocessing is Important

1. Real-world data often contains noise, missing values, and incorrect information.

2. Raw data may be incomplete or in an unusable format, so it cannot be used directly for analysis.
3. Data preprocessing cleans and organizes the data before applying machine learning models.
4. It helps to improve the accuracy and efficiency of the analysis or model.
5. Proper preprocessing makes the data mining process easier and more reliable.

Data Quality Parameters

1. **Accuracy** — Checks whether the data entered is correct and truly represents the real-world value it is supposed to capture.
2. **Completeness** — Checks whether all required data values are present and that no important fields have been left empty or skipped.
3. **Consistency** — Ensures that the same data is stored and maintained correctly across all places where it appears, with no conflicting versions.
4. **Timeliness** — Ensures that the data is updated correctly and is available when it is needed, without being outdated or delayed.
5. **Believability** — Ensures that the data is reliable, trustworthy, and comes from a credible source that users can depend on.
6. **Interpretability** — Ensures that the data is easy to understand, with clear definitions, labels, and documentation so users know what each field means.

Major Tasks in Data Preprocessing

1. **Data Cleaning** — Remove noise and incorrect data values, handle missing data values, and correct inconsistencies.
2. **Data Integration** — Combine data from multiple sources into a single unified dataset or data warehouse.
3. **Data Reduction** — Reduce the size of the dataset while keeping important information.
4. **Data Transformation** — Convert data into a suitable format for analysis using normalization, discretization, or attribute selection.

---

14. Data Cleaning

Data cleaning is an important step in data preprocessing. It is used to improve data quality by handling missing values, noise, and inconsistencies.

Missing Values

1. Missing values occur when some data fields are empty or not recorded.

2. They may appear as blank cells, NULL values, N/A, or special symbols like "?".
3. Missing values can occur due to equipment malfunction or data entry errors.
4. Sometimes the data may not be recorded because it was considered unimportant.
5. Missing values do not always indicate an error — for example, optional fields in forms.

Causes of Missing Data:

1. Equipment malfunction during data collection — sensors or devices may fail and not record the value at that point.
2. Human errors while entering data — the person filling in the form may accidentally skip a field or enter the wrong value.
3. Deletion of inconsistent data during processing — values that don't match other records may be removed, leaving a gap.
4. Data not entered due to misunderstanding of the question — the respondent may not understand what is being asked and leaves it blank.
5. Historical data not recorded or stored properly — older records may simply not have captured certain fields that were added later.

Methods to Handle Missing Values:

1. Ignore the tuple — simply remove the entire record containing the missing value; works only when very few values are missing and the dataset is large enough.
2. Fill values manually using human input — a person with domain knowledge fills in the missing value; impractical for large datasets.
3. Use a global constant such as "Unknown" or "Not Available" — all missing values are replaced with the same placeholder value.
4. Use statistical measures like mean, median, or mode to fill values — the missing value is replaced with the average or most common value of that attribute.
5. Use predictive methods such as regression, Bayesian methods, or decision trees to estimate the most probable value based on other attributes in the record.

Noisy Data

1. Noisy data refers to data that contains random errors or incorrect values.
2. Noise can occur due to measurement errors or incorrect data entry.
3. Noisy data can affect the accuracy of data analysis and machine learning models.
4. Data smoothing techniques are used to reduce noise in the dataset.
5. Common methods include binning, regression, and clustering.

Binning Method:

1. Binning is a technique used to smooth noisy data.
2. The data values are sorted and divided into groups called bins.
3. Each bin contains values that are close to each other.
4. The values in each bin are replaced by the mean, median, or boundary values.
5. This method helps reduce the effect of noise in the dataset.

Regression Method:

1. Regression is used to smooth data by fitting values into a mathematical function.
2. It finds the best line or curve that represents the relationship between variables.
3. Linear regression uses one variable to predict another.
4. Multiple linear regression uses more than one variable for prediction.
5. Regression helps estimate missing values and reduce noise in data.

Example: In the Titanic dataset, missing age values can be predicted using features like Pclass, Fare, SibSp, and Parch.

Outliers

1. Outliers are data values that are very different from other values in the dataset.
2. They may occur due to measurement errors or unusual events.
3. Outliers can distort statistical results and analysis.
4. Detecting and handling outliers is an important step in data cleaning.
5. Techniques such as statistical methods and clustering algorithms are used for outlier detection.

Methods for Outlier Detection:

1. Z-score method — a value is considered an outlier if its Z-score is greater than +3 or less than −3, meaning it is more than 3 standard deviations away from the mean.
2. Standard deviation method — any data value that lies more than 2 or 3 standard deviations from the mean is flagged as a potential outlier and investigated further.
3. Clustering methods — data is grouped into clusters, and any point that does not belong clearly to any cluster or sits far from all cluster centers is considered an outlier.
4. K-means algorithm — assigns each data point to the nearest cluster center; points that are consistently far from their assigned center across iterations are treated as abnormal values.
5. DBSCAN algorithm — identifies dense regions of data as clusters; points that do not fit into any dense region are labeled as noise or outliers.

Data Cleaning as a Process

1. The first step in data cleaning is discrepancy detection.
2. Discrepancies occur when two or more datasets contain conflicting values.
3. These errors may be caused by poor data entry forms, human errors, or outdated data.
4. Metadata (data about data) is used to understand the structure and rules of the dataset.
5. Statistical methods like mean, median, and standard deviation help identify anomalies.

Tools Used for Data Cleaning:

1. Data scrubbing tools detect and correct errors using domain knowledge such as postal addresses and spell-checking.
2. Data auditing tools analyze data to detect inconsistencies and rule violations using statistical analysis or clustering.
3. ETL tools (Extract, Transform, Load) allow users to specify transformation rules through a graphical interface.
4. Custom scripts may also be written for specific data transformation and cleaning tasks.
5. Interactive tools such as Potter's Wheel help users detect and fix data discrepancies visually and interactively.

---

15. Data Integration

Data integration is the process of combining data from multiple sources into a single unified dataset.

1. Data integration merges data from different databases or data sources into a coherent data store.
2. It helps reduce data redundancy and inconsistencies.
3. Proper integration improves the accuracy and speed of data mining processes.
4. It ensures that data from different systems can work together.
5. However, integration is challenging because data sources may have different formats and structures.

Entity Identification Problem

1. The same real-world object may appear with different names in different databases.
2. For example, "Bill Clinton" and "William Clinton" may represent the same person.
3. Identifying such entities is called the entity identification problem.
4. Metadata and attribute information are used to match entities correctly.

5. Proper entity identification prevents duplicate records in the integrated dataset.

Redundancy and Correlation Analysis

Handling Redundancy:

1. Redundant data occurs when the same information appears multiple times in different databases.
2. Redundant data increases storage requirements and processing time.
3. It can lead to inconsistent or conflicting results.
4. Redundancy is detected using correlation analysis and covariance analysis.
5. Removing redundant attributes improves data quality and mining performance.

Correlation Analysis (Numeric Data):

1. Correlation shows whether two numeric variables change together — when one changes, does the other follow?
2. If both variables increase together, it is called positive correlation — for example, study hours and exam scores.
3. If one variable increases while the other decreases, it is called negative correlation — for example, TV watching time and exam marks.
4. Correlation helps detect redundant attributes — if two attributes are highly correlated, one may be removed without losing information.
5. Pearson's correlation coefficient is commonly used to measure the strength and direction of linear relationships in numeric data.

Pearson Correlation Coefficient:

1. Pearson correlation measures the strength of the linear relationship between two numeric variables using a single number.
2. Its value always lies between −1 and +1, making it easy to compare relationships across different datasets.
3. A value of +1 indicates a perfect positive correlation — as one variable increases, the other increases by a proportional amount.
4. A value of −1 indicates a perfect negative correlation — as one variable increases, the other decreases by a proportional amount.
5. A value of 0 indicates no linear correlation between the two variables — they are statistically independent of each other.

Chi-Square Test (Categorical Data):

1. The Chi-square test is used specifically for categorical or nominal data to check if two attributes are related.
2. It helps determine whether two attributes are statistically related or whether they are independent of each other.
3. It compares observed values (what is actually in the data) against expected values (what would be expected if the attributes were independent).
4. If the calculated chi-square value is greater than the critical value from the table, we reject the null hypothesis — the attributes are correlated.
5. This test is commonly used for data redundancy detection in categorical datasets, such as checking if gender and product preference are related.

Covariance Analysis:

1. Covariance measures how two numeric attributes change together — whether they move in the same direction or opposite directions.
2. A positive covariance means both variables increase together — for example, income and spending tend to rise at the same time.
3. A negative covariance means one variable increases while the other decreases — for example, exercise hours and body weight may move in opposite directions.
4. Covariance helps understand the relationship direction between two variables and is used to identify potential redundancies.
5. However, its magnitude depends on the units of the variables, making it hard to compare covariances across different pairs of attributes.

Covariance vs. Pearson Correlation:

Feature	Covariance	Pearson Correlation
Range	Can be any value from $-\infty$ to $+\infty$	Always between −1 and +1
Units	Depends on units of variables	Unitless
Interpretation	Harder to interpret magnitude	Easy to interpret relationship strength
Purpose	Shows direction of relationship	Shows strength and direction

Tuple Duplication

1. Tuple duplication occurs when the same record exists multiple times across different databases.
2. This can happen when data is collected from multiple sources that overlap.

3. Duplicate records must be identified and removed during integration.
4. Keeping duplicates leads to incorrect analysis results.
5. Proper deduplication ensures data integrity in the integrated dataset.

---

16. Data Reduction

Data reduction reduces the size of the dataset while keeping important information.

1. It helps improve the efficiency of data mining algorithms.
2. Techniques include data aggregation and feature selection.
3. It removes redundant or unnecessary data attributes.
4. It simplifies the dataset while maintaining the original meaning of the data.
5. A database may store terabytes of data — complex analysis may take very long on the full dataset, so reduction is necessary.

Attribute Subset Selection

1. Attribute subset selection removes irrelevant or redundant attributes (dimensions) from the dataset.
2. The goal is to find a minimum set of attributes such that the probability distribution of classes is as close as possible to the original.
3. Reducing the number of attributes makes patterns easier to understand and speeds up mining algorithms.
4. Fewer attributes also means less storage, less computation, and simpler models that are easier to interpret.
5. It avoids the "curse of dimensionality" — where too many features make it hard to find meaningful patterns.

Heuristic (Greedy) Methods:

1. Heuristic methods use a step-by-step strategy to efficiently search for a good subset of attributes without trying every possible combination.
2. **Stepwise Forward Selection** — Start with an empty set of attributes; at each step, add the single best-performing attribute until no further improvement is found.
3. **Stepwise Backward Elimination** — Start with all attributes; at each step, remove the single worst-performing attribute until no further improvement is found.
4. **Combination** — Combine forward selection and backward elimination together — at each step, both the best attribute is added and the worst is removed simultaneously.

5. **Decision Tree Induction** — Build a decision tree on the data; any attributes that do NOT appear anywhere in the tree are considered irrelevant and are discarded.

Histograms

1. Histograms are a graphical method for summarizing the distribution of a given attribute.
2. The range of values is partitioned into disjoint consecutive subranges called buckets or bins.
3. The height of each bar indicates the frequency (count) of values in that range.
4. Typically, buckets are of equal width — this is called equal-width partitioning.
5. Equal-frequency (equal-depth) partitioning: each bucket has approximately the same number of data points.

Example: A price attribute with range $1–$200 partitioned into $1–$20, $21–$40, $41–$60, etc.

Sampling

1. Sampling obtains a small representative sample from the whole dataset instead of analyzing all the data.
2. It allows a mining algorithm to run in complexity that is sub-linear to the size of the data, making it much faster.
3. Useful when working with very large datasets where analyzing every record is impractical or too slow.
4. A good sample should represent the overall distribution of the data accurately to give reliable results.
5. The key challenge in sampling is ensuring the sample does not miss important subgroups or rare patterns.

Types of Sampling:

1. **Simple Random Sampling** — Every item in the dataset has an equal probability of being selected; simple to implement but may perform poorly when data is heavily skewed.
2. **Sampling without Replacement** — Once an object is selected, it is removed from the population and cannot be selected again, ensuring no duplicates in the sample.
3. **Sampling with Replacement** — The selected object is not removed from the population, so the same item can be picked more than once across multiple draws.
4. **Stratified Sampling** — The dataset is divided into subgroups (strata) based on shared characteristics such as age or region, and random samples are taken from each stratum; this ensures all subgroups are fairly represented, especially useful with skewed data.

---

17. Data Transformation

Data transformation is the process of converting data into a suitable format for data mining and analysis.

1. Data transformation changes the structure or format of data so that it can be analyzed easily.
2. It helps improve data quality and efficiency of mining algorithms.
3. Raw data is often scaled, aggregated, or generalized before analysis.
4. It is an important step in data preprocessing.
5. Common techniques include smoothing, aggregation, generalization, normalization, and attribute construction.

Normalization — Min-Max

1. Min-Max normalization transforms values into a specified range, usually 0 to 1.
2. It performs a linear transformation on the original data.
3. It preserves the relationships between data values.
4. It is widely used in machine learning preprocessing.
5. Limitation: if a new value falls outside the original [min, max] range, it causes an out-of-bounds error.

Formula: $v' = (v - \min(A)) / (\max(A) - \min(A)) \times (new\_max - new\_min) + new\_min$

Example: Income ranges from $12,000 to $98,000; a value of $73,600 becomes **0.716** after normalization to [0, 1].

Normalization — Z-Score

1. Z-score normalization uses the mean and standard deviation.
2. It converts values into standard deviation units.
3. This method is useful when minimum and maximum values are unknown.
4. It centers the data around the mean value.
5. It is commonly used in statistical analysis.

Formula: $Z = (x - \mu) / \sigma$

Example: Mean income = $54,000, SD = $16,000 → $73,600 normalizes to approximately **1.225**.

Normalization — Decimal Scaling

1. Decimal scaling moves the decimal point of values.
2. Each value is divided by a power of 10.

3. The value of j is chosen so that all normalized values are less than 1.
4. This method is simple and easy to implement.
5. It is used when data values have large numeric ranges.

Example: Values range from −986 to 917 → divide by 1000 → −986 becomes −0.986, 917 becomes 0.917.

Normalization Comparison Table

Normalization Type	Formula Used	Range After	Best When
Min-Max	Linear transformation	[new_min, new_max] e.g. [0, 1]	Known min/max, no extreme outliers
Z-Score	Mean and SD	No fixed range	Unknown min/max, or outliers present
Decimal Scaling	Divide by 10^j	(−1, 1) approximately	Simple scaling is needed

Generalization (Concept Hierarchy Climbing)

1. Generalization replaces low-level data with higher-level, more general concepts.
2. It helps simplify data for analysis.
3. It uses a concept hierarchy structure.
4. Lower-level values are grouped into more general categories.
5. Example: city → state → country hierarchy.

---

18. Data Discretization

Data discretization converts continuous numerical values into categories or intervals.

1. It divides data values into groups called bins or intervals.
2. It reduces the number of distinct values in the dataset.
3. It simplifies large datasets for analysis.
4. It helps handle noisy data and improve model performance.
5. Example: Age → Child, Teen, Adult, Senior.

Methods of Data Discretization

1. **Binning method** — divides sorted data values into equal-sized groups called bins and replaces values with the bin mean or boundary values; unsupervised and top-down.

2. **Histogram analysis** — uses frequency distributions to identify natural breakpoints in the data and partition values into intervals; unsupervised and top-down.
3. **Clustering analysis** — groups similar data values into clusters and treats each cluster as a separate interval; can be top-down or bottom-up and is unsupervised.
4. **Decision tree analysis** — uses class labels and entropy-based measures to find the best split points in the data; supervised and top-down.
5. **Chi-square analysis** — merges adjacent intervals that have similar distributions based on a statistical test; unsupervised and bottom-up.

Concept Hierarchy

1. Concept hierarchy organizes data in a tree-like structure where lower levels hold specific values and higher levels hold more general concepts.
2. Each higher level in the hierarchy represents more general, summarized information about the level below it.
3. It helps reduce data complexity by replacing many specific values with a smaller number of meaningful categories.
4. It improves data visualization and analysis by allowing users to view data at different levels of detail.
5. Example: City → State → Country → Continent — a street address can be rolled up all the way to a continent.

Applications of Concept Hierarchy

1. Data Warehousing — organizes data from multiple sources into a consistent, hierarchical structure for efficient analysis.
2. Business Intelligence — helps identify trends and patterns at different levels such as product category or sales region.
3. Online Retail — organizes products into categories and subcategories, making it easier to analyze sales by group.
4. Healthcare — groups patient data by diagnosis, treatment plan, or department to support medical analysis and research.
5. Fraud Detection — identifies unusual patterns in financial data by analyzing behavior at different levels of generalization.

End of Unit 1 Notes