## UNIT 5

## Data enhancement

**Data Enhancement Definition**

Data enhancement generally refers to the process of improving your or your client’s, first-party data with context pulled from additional, external sources.

This process is also referred to as data append, which generally describes adding external data to your records to expand them. A wide variety of data can be appended depending on your marketing needs, whether contact, demographic, firm graphic, or geographic information.

The data enhancement process follows this general format



## Business case

## A business case captures the reasoning for initiating a project or task. It is often presented in a well-structured written document, but may also come in the form of a short verbal agreement or presentation.

## Five elements of a business case

A common way of thinking about a **business case** is using these five elements:

1. **Strategic context**: The compelling case for change.
2. **Economic analysis**: Return on investment based on investment appraisal of options.
3. **Commercial approach**: Derived from the sourcing strategy and procurement strategy.
4. **Financial case**: Affordability to the organisation in the time frame.
5. **Management approach**: Roles, governance structure, life cycle choice, etc.

**Types of Data Enhancement**

### Data Enhancement

###  In a general sense, data enhancement involves taking your existing customer files and augmenting them to widen your data set. In a data enhancement scenario, you will maximize internal data by adding relevant and important external information as well.Data Enhancement Definition

### Data enhancement will simplify your data, making it easier to work with and run reports. In the customer-centric world, we live in, data enhancement allows this type of marketing approach to occur effectively.

### Maximizing Existing Data

You must maintain a clean database to get the most of your existing data. A clean database will improve communication efficiency and alleviate the potential of missing out on important opportunities. Bad data creates inefficient email marketing.

Deleting or merging duplicate subscribers is one of the most important data cleansing activities. Data enhancement encourages merging duplicates and normalizing your matches as well. Duplicate data burdens databases and wastes [lead generation](https://www.frescodata.com/demand-generation/) activities. These changes will automatically increase the reliability of your database.

More efficient, consistent and accurate reporting can occur when fields like job titles, phone numbers, and addresses are standardized within your CRM. The type of information you can fix during a data enhancement includes:

* Inaccurate record details
* Incorrect email addresses
* Wrong phone numbers
* Improper formatting

### Benefits of External Data

One of the biggest benefits you can receive from data enhancement is the introduction of external data to your data lists. External data can greatly enhance in-house data. You must acquire relevant and meaningful data to complement your existing data. This will strengthen the success of your email campaigns.
Some of the benefits of external data enhancement include:

* Greater understanding of target markets
* More precise targeting with segmented email lists
* Create on-point marketing messages
* Save money and reduce wastage

### The Relevance of External Data

It is important to gather external data that matches your objectives and needs. Keep in mind what your segmentation strategies are and how you can gather data to strengthen list segments. Do not gather list data without a plan. This will clog your database with unnecessary and irrelevant details.
Many marketers are interested in gathering behavior data, such as how a potential customer navigates through websites and purchases products online. The details you gather about your prospective clients should increase the effectiveness of your targeted email campaigns.

### How Often to Perform Data Enhancement?

Data enhancement should be an ongoing project within your business or company. Maintaining an accurate, up-to-date and relevant email list is of utmost importance. Cleaning your database should happen routinely at least every 3-6 months while appending external data should occur at least once yearly.

### Building Targeted Email Campaigns

With a fully optimized database, it is easier to build targeted email campaigns through list segmentation. Your list segments will be more accurate because the data that drives the segmentation is more complete. This way your segmented campaigns will be more meaningful to your audience.
People tend to ignore, delete, unsubscribe or block email messages that do not relate to them. With data enhancement, you are given the opportunity to truly reach your audience with creative, on-point messaging that will create more engagement and revenue.



## Incremental improvement is an approach to process improvement in which you and your staff focus efforts on smaller solutions that slowly but surely move the business toward success. These ideas are typically low-cost and low-risk, and are implemented by employees throughout the entire organization.

**Batch Processing** is one such method to effectively handle massive amounts of data and send data in batches to the target system. It is a flexible technique that provides you with more control and assists you in efficiently transferring data with the already available computational resources. Though there are methods like Stream Processing for quickly transferring data in real-time, they are highly ineffective when it comes to dealing with huge volumes of data.



Batch processing is an efficient way of running a large number of iterative data jobs. With the right amount of computing resources present, the batch method allows you to process data with little to no user interaction.

After you have collected and stored your data, the batch processing method allows you to process it during an event called a “batch window“. It provides an efficient workflow layout by prioritizing processing tasks and completing the given data jobs when it makes the most sense.

For the very first time, batch processing was brought into use in the 19th century by Herman Hollerith, an American inventor who built the first tabulating machine. This device pioneered the latest computers that can count and sort data organized in the form of punch cards. The card and the information it contains are collected in batches and processed together. This innovation has made it possible to process large amounts of data faster and more accurately than any of the traditional manual input methods.

**KDD (Knowledge Discovery in Databases)** is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process in data mining typically involves the following steps:

1. **Selection**: Select a relevant subset of the data for analysis.
2. **Pre-processing:** Clean and transform the data to make it ready for analysis. This may include tasks such as data normalization, missing value handling, and data integration.
3. **Transformation:**Transform the data into a format suitable for data mining, such as a matrix or a graph.
4. **Data Mining:** Apply data mining techniques and algorithms to the data to extract useful information and insights. This may include tasks such as clustering, classification, association rule mining, and anomaly detection.
5. **Interpretation**: Interpret the results and extract knowledge from the data. This may include tasks such as visualizing the results, evaluating the quality of the discovered patterns and identifying relationships and associations among the data.
6. **Evaluation**: Evaluate the results to ensure that the extracted knowledge is useful, accurate, and meaningful.
7. **Deployment**: Use the discovered knowledge to solve the business problem and make decisions.

The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.



1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
	* Cleaning in case of ***Missing values***.
	* Cleaning ***noisy*** data, where noise is a random or variance error.
	* Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.
2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
	* Data integration using ***Data Migration tools***.
	* Data integration using ***Data Synchronization tools***.
	* Data integration using ***ETL***(Extract-Load-Transformation) process.
3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
	* Data selection using ***Neural network***.
	* Data selection using ***Decision Trees***.
	* Data selection using ***Naive bayes***.
	* Data selection using ***Clustering***, ***Regression***, etc.
4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two step process:

* + ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
	+ ***Code generation***: Creation of the actual transformation program.
1. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
	* Transforms task relevant data into ***patterns***.
	* Decides purpose of model using ***classification*** or ***characterization***.
2. ***Pattern Evaluation***: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
	* Find ***interestingness score*** of each pattern.
	* Uses ***summarization*** and ***Visualization*** to make data understandable by user.
3. ***Knowledge representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
	* Generate ***reports***.
	* Generate ***tables***.
	* Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

# Data Mining Vs Data Warehousing

**Data warehouse** refers to the process of compiling and organizing data into one common database, whereas **data mining** refers to the process of extracting useful data from the databases. The data mining process depends on the data compiled in the data warehousing phase to recognize meaningful patterns. A data warehousing is created to support management systems.

## Data Warehouse:

A **Data Warehouse** refers to a place where data can be stored for useful mining. It is like a quick computer system with exceptionally huge data storage capacity. Data from the various organization's systems are copied to the Warehouse, where it can be fetched and conformed to delete errors. Here, advanced requests can be made against the warehouse storage of data.



Data warehouse combines data from numerous sources which ensure the data quality, accuracy, and consistency. Data warehouse boosts system execution by separating analytics processing from transnational databases. Data flows into a data warehouse from different databases. A data warehouse works by sorting out data into a pattern that depicts the format and types of data. Query tools examine the data tables using patterns.

**Data warehouses** and **databases** both are relative data systems, but both are made to serve different purposes. A data warehouse is built to store a huge amount of historical data and empowers fast requests over all the data, typically using **Online Analytical Processing** (OLAP). A database is made to store current transactions and allow quick access to specific transactions for ongoing business processes, commonly known as **Online Transaction Processing** (OLTP).

Play Video

### Important Features of Data Warehouse

The Important features of Data Warehouse are given below:

**1. Subject Oriented**

A data warehouse is subject-oriented. It provides useful data about a subject instead of the company's ongoing operations, and these subjects can be customers, suppliers, marketing, product, promotion, etc. A data warehouse usually focuses on modeling and analysis of data that helps the business organization to make data-driven decisions.

**2. Time-Variant:**

The different data present in the data warehouse provides information for a specific period.

**3. Integrated**

A data warehouse is built by joining data from heterogeneous sources, such as social databases, level documents, etc.

**4. Non- Volatile**

It means, once data entered into the warehouse cannot be change.

### Advantages of Data Warehouse:

* More accurate data access
* Improved productivity and performance
* Cost-efficient
* Consistent and quality data

# The Virtuous Cycle of Data Mining

#### Consists of four major business processes (success in data mining requires all four):



A virtuous cycle is **a recurring cycle of events**. The result of each one increases the beneficial effect of the next. For example, in the economy, higher wages increase consumption, which leads to higher prices and a fattening up of the corporate bottom line.

1. **Directed graphs** - The edges are orderedd pair ie.(u,v) and (v,u) have different meaning where (u,v) is read as edge from u to v. Hence there is set direction where information can flow. This direction is represented by arraows used as edges. The following graph can only be traversed from **A** to **B**

2. **Undirected graph** - The edges are unordered pair. Hence, (u,v) and (v,u) mean the same thing . The flow of information is in both directions simultaneously. Since there is no set direction undirected arcs represent edges. The following graph can be traversed from **A** to **B** as well as from **B** to **A**.


Now, we will further discover the differences between these two graphs

## Edges

In undirected graphs the edges are bidirectional, with no direction associalted with them the absence of arraow differenciates them from directed graphs. In directed graphs since the edges can only be traversed in only 1 direction in pictoral depiction arrows are used as eedges in directed graphs with arrow head pointing to Endpoint of relationship.


## Nodes

Even though the pictorial representation of the graphs usually do not change whether it be undirected or directed graph but, the directed graph has concept of source and sink vertices. Wherein the node from which the edge orignates is called source vertex while the node at which the edge terminates is known as sink vertex. while in undirected graphs since the arcs are bidirectional the two nodes joined by edges are simply known as end points.


## Adjacency Matrix

Since the edges in undirected graph are bi-directional it leads to formation of of an adjacency matrix that is symmetrical.


However in case of directed graphs no such symmetry is seen hence it is a usre way of knowing that if adjacency matrix is not symmetrical it will be a directed graph.
== Symmetrical directed graphs are undirected graphs.==


Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining −

* Descriptive
* Classification and Prediction

## Descriptive Function

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions −

* Class/Concept Description
* Mining of Frequent Patterns
* Mining of Associations
* Mining of Correlations
* Mining of Clusters

### Class/Concept Description

Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways −

* **Data Characterization** − This refers to summarizing data of class under study. This class under study is called as Target Class.
* **Data Discrimination** − It refers to the mapping or classification of a class with some predefined group or class.

### Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns −

* **Frequent Item Set** − It refers to a set of items that frequently appear together, for example, milk and bread.
* **Frequent Subsequence** − A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
* **Frequent Sub Structure** − Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

### Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

### Mining of Correlations

It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

### Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

## Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms −

* Classification (IF-THEN) Rules
* Decision Trees
* Mathematical Formulae
* Neural Networks

The list of functions involved in these processes are as follows −

* **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
* **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
* **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
* **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

## Data Mining Task Primitives

* We can specify a data mining task in the form of a data mining query.
* This query is input to the system.
* A data mining query is defined in terms of data mining task primitives.

**Note** − These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives −

* Set of task relevant data to be mined.
* Kind of knowledge to be mined.
* Background knowledge to be used in discovery process.
* Interestingness measures and thresholds for pattern evaluation.
* Representation for visualizing the discovered patterns.

### Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following −

* Database Attributes
* Data Warehouse dimensions of interest

### Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are −

* Characterization
* Discrimination
* Association and Correlation Analysis
* Classification
* Prediction
* Clustering
* Outlier Analysis
* Evolution Analysis

# Data Mining Techniques

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction.

Depending on various methods and technologies from the intersection of machine learning, database management, and statistics, professionals in data mining have devoted their careers to better understanding how to process and make conclusions from the huge amount of data, but what are the methods they use to make it happen?

In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression.

**1. Classification:**

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria, as follows:

1. **Classification of Data mining frameworks as per the type of data sources mined:**
This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..
2. **Classification of data mining frameworks as per the database involved:**
This classification based on the data model involved. For example. Object-oriented database, transactional database, relational database, and so on..
3. **Classification of data mining frameworks as per the kind of knowledge discovered:**
This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together..
4. **Classification of data mining frameworks according to data mining techniques used:**
This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.
The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

## 2. Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view, clustering plays an extraordinary job in data mining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more.

In other words, we can say that Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

## 3. Regression:

Regression analysis is the data mining process is used to identify and analyze the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

## 4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.

The way the algorithm works is that you have various data, For example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.

These are three major measurements technique:

* **Lift:**
This measurement technique measures the accuracy of the confidence over how often item B is purchased.
                  **(Confidence) / (item B)/ (Entire dataset)**
* **Support:**
This measurement technique measures how often multiple items are purchased and compared it to the overall dataset.
                  **(Item A + Item B) / (Entire dataset)**
* **Confidence:**
This measurement technique measures how often item B is purchased when item A is purchased as well.
                  **(Item A + Item B)/ (Item A)**

## 5. Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outilier mining. The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

## 6. Sequential Patterns:

The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

## 7. Prediction:

Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyses past events or instances in the right sequence to predict a future event.

##