**CSE-Data Science/Data Science, IV semester**

**CD404 INTRODUCTION TO DATA SCIENCE**

**UNIT 2**

**DATA PRE-PROCESSING**

**NOTES**

**SYLLABUS**

Unit – I: Introduction

Introduction to Data Science – Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues.

Unit – II: Data Collection and Data Pre-Processing Data Collection Strategies – Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization.

Unit – III: Exploratory Data Analytics Descriptive Statistics – Mean, Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA.

Unit – IV: Model Development Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and Decision Making.

Unit – V: Model Evaluation Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using Grid Search.

Data Pre processing

Data pre-processing/preparation/cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset, or and refers to identifying incorrect, incomplete, irrelevant parts of the data and then modifying, replacing, or deleting the dirty or coarse data.

**2.1 Overview of Data pre-processing**

Data pre-processing is an important step in the data science project process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data pre-processing is to improve the quality of the data and to make it more suitable for the specific data science project/ task.

* Data
  + Audio
  + Text
  + Video
  + Image

Data Pre-processing is a technique to convert raw data into meaning full data by using some different techniques.

**2.2 Why Data Pre-processing is important?**

* + Data in real world is impure.
    - * Incomplete
      * Inconsistent
      * Noisy
      * Duplicate

**2.3 Data pre-processing**

* Data pre-processing is the processing of preparing the data and making it suitable for a ML model.
* A quality of data affects the ability of model to learn , hence it is extremely important that we process data before giving as input to model.
* Whenever data is generated from different sources it is collected in raw format which is not suitable for analysis. So data pre-processing is a technique that is used to convert the raw data into a clean dataset.
* Data in raw format contains noises, missing values, not in usable format.
* Data pre-processing cleans the data and make it suitable for ML model which also increases the accuracy and efficiency of a ML model.

**2.4 Data Quality: Why Preprocess the Data?**

Data quality is a measure of a data set's condition based on factors such as accuracy, completeness, consistency, reliability and validity. Measuring data quality can help organizations identify errors and inconsistencies in their data and assess whether the data fits its intended purpose.

Organizations have grown increasingly concerned about data quality as they've come to recognize the important role that data plays in business operations and [advanced analytics](https://www.techtarget.com/searchbusinessanalytics/definition/advanced-analytics), which are used to drive business decisions

**2.4.1 Why is data quality so important?**

Low-quality data can have significant business consequences for an organization. Bad data is often the culprit behind operational snafus, inaccurate analytics and ill-conceived business strategies.

Measures for data quality: A multidimensional view

Accuracy: The data correctly represents the entities or events it is supposed to represent, and the data comes from sources that are verifiable and trustworthy.

Completeness: The data includes all the values and types of data it is expected to contain, including any metadata that should accompany the data sets.

Consistency: The data is uniform across systems and data sets, and there are no conflicts between the same data values in different systems or data sets.

Timeliness: The data is current (relative to its specific requirements) and is available to use when it's needed.

Believability/Uniqueness: The data does not contain duplicate records within a single data set, and every record can be uniquely identified.

Interpretability: Possibility to find its meaning or possible to find a particular meaning in it.

## **2.4.2 Data quality vs. data integrity**

The terms data quality and data integrity are sometimes used interchangeably, although they have different meanings. At the same time, some people treat [data integrity](https://www.techtarget.com/searchdatacenter/definition/integrity) as a facet of data quality or data quality as a component of data integrity. Others consider both data quality and data integrity to be part of a larger data governance effort, while still others consider data integrity to be a broader concept that combines data quality, data governance and [data protection](https://www.techtarget.com/searchdatabackup/definition/data-protection) into a unified effort for addressing data accuracy, consistency and security.

**2.5 Major Tasks in Data Preprocessing**

**Data collection**

Collect data from different sources

**Data cleaning**

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

**Data integration**

Integration of multiple databases or files

**Data reduction**

Dimensionality reduction

Data compression

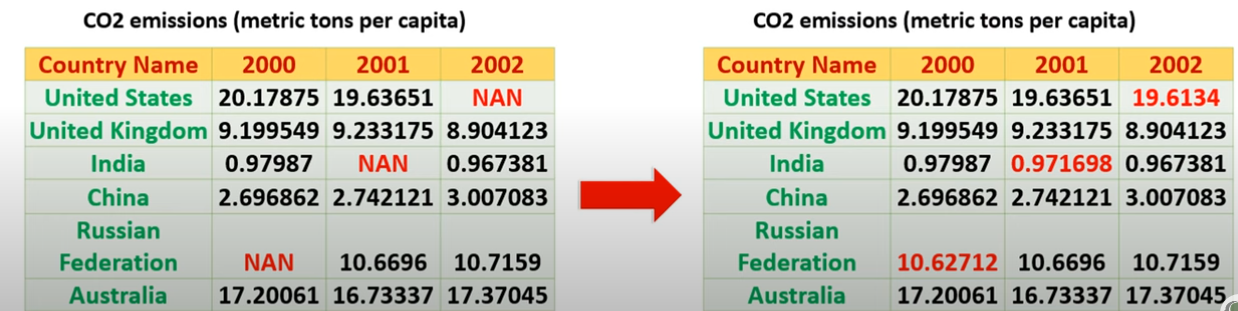
**Data transformation and data discretization**

Normalization

Concept hierarchy generation

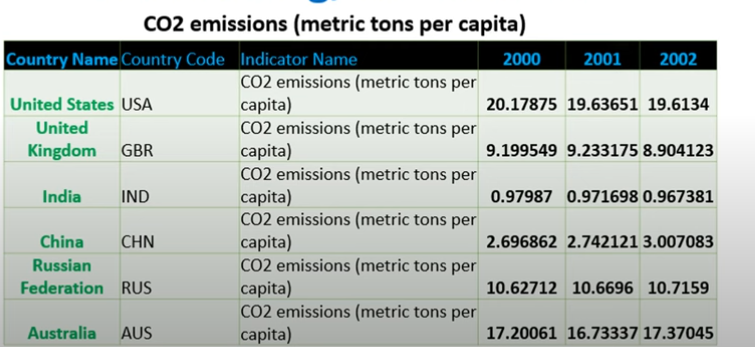
**Data Cleaning**

Data cleaning means fill in missing values, smooth out noise while identifying outliers and correct inconsistency in data.



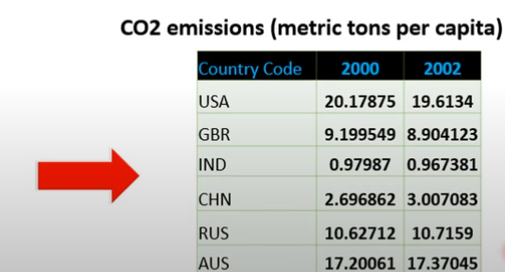
**Data Integration**

Data integration is a technique to merges data from multiple sources into a coherent data store, such as data warehousing.



**Data Reduction**

Data Reduction is a technique to reduce the data size by aggregating, eliminating redundant features, or clustering.



**Data Transformation**

Data transformation means data are transformed or consolidated into forms appropriate for ML model training, such as normalization, may be applied where data are scaled to fall with in a smaller range like 0.0 and 1.0

* Aggregation
* Feature type conversion
* Normalization
* Attribute/feature construction.



**Data Discretization**

Data discretization technique transforms the numeric data by mapping values to interval or concept labels.

It can be used to reduce the number of values for a given continuous attribute by dividing the range of attributes into intervals.

Data discretization technique include-

* Histogram analysis
* Cluster analysis
* Decision tree analysis
* Correlation analysis

**2.6 Data Collection**

Data collection is the essential step in data science project.

This step involves gathering data from various sources such as databases, files and external repositories. Before starting the data collection process, first we have to find out the problem solved by ML model. Knowing the objective to achieve by algorithm will assist in determining the type of data required.

Also it helps in getting a clear picture of all data available, required and missing. The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data collection is the process of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Before an analyst begins collecting data, they must answer three questions first:

* + What’s the goal or purpose of this research?
  + What kinds of data are they planning on gathering?
  + What methods and procedures will be used to collect, store, and process the information?

**2.6.1 Different Data Collection Methods**

There are two types of data collection methods

* Primary data collection
* Secondary data collection

Primary and secondary methods of data collection are two approaches used to gather information for research or analysis purposes.

**Primary data collection:**

Primary data collection involves the collection of original data directly from the source or through direct interaction with the respondents.

Various techniques for primary data collection, including:

* Surveys and Questionnaires:- Researchers design structured questionnaires or surveys to collect data from individuals or groups. These can be conducted through face-to-face interviews, telephone calls, mail, or online platforms
* Interviews:-Interviews involve direct interaction between the researcher and the respondent. They can be conducted in person, over the phone, or through video conferencing. Interviews can be structured (with predefined questions), semi-structured (allowing flexibility), or unstructured (more conversational).
* Observations:- Researchers observe and record behaviors, actions, or events in their natural setting. This method is useful for gathering data on human behavior, interactions, or phenomena without direct intervention.
* Experiments:- Experimental studies involve the manipulation of variables to observe their impact on the outcome. Researchers control the conditions and collect data to draw conclusions about cause-and-effect relationships.
* Focus Groups:- Focus groups bring together a small group of individuals who discuss specific topics in a moderated setting. This method helps in understanding opinions, perceptions, and experiences shared by the participants.

**Secondary Data Collection:**

Secondary data collection involves using existing data collected by someone else for a purpose different from the original intent.

Researchers analyze and interpret this data to extract relevant information.

Secondary data can be obtained from various sources, including:

* Published Sources: Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data
* Online Databases: Numerous online databases provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys.
* Government and Institutional Records: Government agencies, research institutions, and organizations often maintain databases or records that can be used for research purposes
* Publicly Available Data: Data shared by individuals, organizations, or communities on public platforms, websites, or social media can be accessed and utilized for research.
* Past Research Studies: Previous research studies and their findings can serve as valuable secondary data sources. Researchers can review and analyze the data to gain insights or build upon existing knowledge.

**2.6.2 Data Collection Tools**

* Social Media Listening Tool
* Web Analytics Tools
* Data Logging Devices
* Mobile Data Collection Apps
* IoT Platforms

## **2.6.3 Issues Related to Maintaining the Integrity of Data Collection**

In order to assist the errors detection process in the data gathering process, whether they were done purposefully (deliberate falsifications) or not, maintaining data integrity is the main justification (systematic or random errors).

Quality assurance and quality control are two strategies that help protect data integrity and guarantee the scientific validity of study results.

Each strategy is used at various stages of the research timeline:

* Quality control - tasks that are performed both after and during data collecting
* Quality assurance - events that happen before data gathering starts

## **2.6.4 Common Challenges in Data Collection**

There are some prevalent challenges faced while collecting data, let us explore a few of them to understand them better and avoid them.

### **Data Quality Issues**

The main threat to the broad and successful application of machine learning is poor data quality. Data quality must be your top priority if you want to make technologies like machine learning work for you. Let's talk about some of the most prevalent data quality problems in this blog article and how to fix them.

### **Inconsistent Data**

When working with various data sources, it's conceivable that the same information will have discrepancies between sources. The differences could be in formats, units, or occasionally spellings. The introduction of inconsistent data might also occur during firm mergers or relocations. Inconsistencies in data have a tendency to accumulate and reduce the value of data if they are not continually resolved. Organizations that have heavily focused on data consistency do so because they only want reliable data to support their analytics.

### **Data Downtime**

Data is the driving force behind the decisions and operations of data-driven businesses. However, there may be brief periods when their data is unreliable or not prepared. Customer complaints and subpar analytical outcomes are only two ways that this data unavailability can have a significant impact on businesses. A data engineer spends about 80% of their time updating, maintaining, and guaranteeing the integrity of the data pipeline. In order to ask the next business question, there is a high marginal cost due to the lengthy operational lead time from data capture to insight.

Schema modifications and migration problems are just two examples of the causes of data downtime. Data pipelines can be difficult due to their size and complexity. Data downtime must be continuously monitored, and it must be reduced through automation.

### **Ambiguous Data**

Even with thorough oversight, some errors can still occur in massive databases or data lakes. For data streaming at a fast speed, the issue becomes more overwhelming. Spelling mistakes can go unnoticed, formatting difficulties can occur, and column heads might be deceptive. This unclear data might cause a number of problems for reporting and analytics

### **Duplicate Data**

Streaming data, local databases, and cloud data lakes are just a few of the sources of data that modern enterprises must contend with. They might also have application and system silos. These sources are likely to duplicate and overlap each other quite a bit. For instance, duplicate contact information has a substantial impact on customer experience. If certain prospects are ignored while others are engaged repeatedly, marketing campaigns suffer. The likelihood of biased analytical outcomes increases when duplicate data are present. It can also result in ML models with biased training data.

### **Too Much Data**

While we emphasize data-driven analytics and its advantages, a data quality problem with excessive data exists. There is a risk of getting lost in an abundance of data when searching for information pertinent to your analytical efforts. Data scientists, data analysts, and business users devote 80% of their work to finding and organizing the appropriate data. With an increase in data volume, other problems with data quality become more serious, particularly when dealing with streaming data and big files or databases.

### **Inaccurate Data**

For highly regulated businesses like healthcare, data accuracy is crucial. Given the current experience, it is more important than ever to increase the data quality for COVID-19 and later pandemics. Inaccurate information does not provide you with a true picture of the situation and cannot be used to plan the best course of action. Personalized customer experiences and marketing strategies underperform if your customer data is inaccurate.

Data inaccuracies can be attributed to a number of things, including data degradation, human mistake, and data drift. Worldwide data decay occurs at a rate of about 3% per month, which is quite concerning. Data integrity can be compromised while being transferred between different systems, and data quality might deteriorate with time.

### **Hidden Data**

The majority of businesses only utilize a portion of their data, with the remainder sometimes being lost in data silos or discarded in data graveyards. For instance, the customer service team might not receive client data from sales, missing an opportunity to build more precise and comprehensive customer profiles. Missing out on possibilities to develop novel products, enhance services, and streamline procedures is caused by hidden data.

2.7. Data Augmentation

**Data Augmentation (A method of data collection)**

In some cases, data augmentation might be required to expand the size of existing dataset without gathering more data. If we have more data, the better will be our ML model. But every data collection process is associated with a cost. This cost may be in terms of dollars, human efforts, computational resources and time consumed in the process. Therefore we may need to augment existing data to increase the data size that we feed to our ML model.

In machine learning, data augmentation is a common method for manipulating existing data to artificially increase the size of a training dataset. In an attempt to enhance the efficiency and flexibility of machine learning models, data augmentation looks for the boost in the variety and volatility of the training data.

Data augmentation can be especially beneficial when the original set of data is small as it enables the system to learn from a larger and more varied group of samples. By applying arbitrary changes to the information, the expanded dataset can catch various varieties of the first examples, like various perspectives, scales, revolutions, interpretations, and mishappenings. As a result, the model can better adapt to unknown data and become more resilient to such variations.

Techniques for data augmentation can be used with a variety of data kinds, including time series, text, photos, and audio.

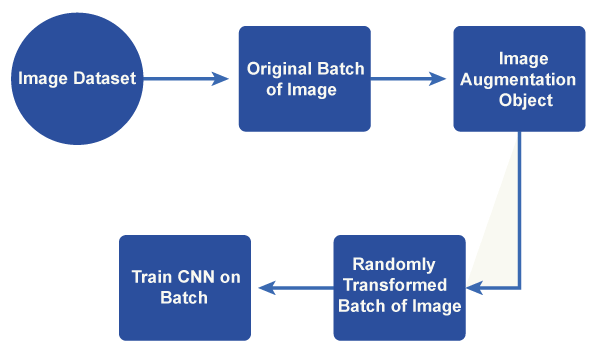


Fig 2.1: Process of data augmentation

* For example if a dataset of images are collected they can be augmented by
  + - Rotating the original image
    - Crop the original image differently
    - Altering the light condition

So for one image you can generate different sub samples

2.7.1 Techniques for data augmentation

Techniques for data augmentation can be used with a variety of data, including time series, text, photos, and audio.

* Images can be rotated at different angles and flipped horizontally or vertically to create alternative points of view.
* Random cropping and padding: By applying random cropping or padding to the photos, various scales, and translations can be simulated.
* Scaling and zooming: The model can manage various item sizes and resolutions by rescaling the photos to different sizes or zooming in and out.
* Shearing and perspective transform: Changing an image's shape or perspective can imitate various viewing angles while also introducing deformations.
* Colour segmentation: By adjusting the color characteristics of the images, including their brightness, contrast, saturation, and hue, the model can be made to be more resilient to variations in illumination.
* Gaussian noise: By introducing random Gaussian noise to the images, the model's resistance to noisy inputs can be strengthened.

**2.7.2 Types of Data Augmentations**

**1. Real Data Augmentation**

**2. Synthetic Data Augmentation**

**Real Data Augmentation: -** The process of modifying real-world data samples to enhance the base of training for artificial intelligence models is referred to as "real data augmentation." Real data augmentation, as compared to synthetic data augmentation produces new samples based on existing data and also modifies the original data in a way that accurately depicts fluctuations and disturbances that occur in the real world.

By capturing the inherent diversity in the data distribution, real data augmentation approaches strive to strengthen the model's adaptability to various scenarios, noise levels, or environmental factors.

Real Data augmentation approaches as examples:

* **Sensor noise:** By adding noise to sensor data, measurement errors or other flaws in the data collection process can be simulated. For instance, adding random Gaussian noise to camera-taken pictures can simulate the sensor noise found in actual image data.
* **Occlusion:** Blocking or partially occluding specific areas of an image might imitate the presence of objects or barriers that are hiding certain areas of the scene. With the aid of this augmentation technique, models are made more resistant to occlusions and are better equipped to deal with partial or blocked visual information.
* **Weather:** Simulating various weather conditions, including snow, rain, or fog, might make the model more resistant to changes in exterior settings. For instance, adding filters or overlays to photographs might make it appear as though it is raining or foggy.
* **Time series:** perturbations can imitate temporal changes and uncertainties in the actual world by altering time series data by adding variations like shifts, scaling, or warping. For activities involving sequential data, such as readings from sensors or financial data, this augmentation strategy can be helpful.
* **Label smoothing:** In some circumstances, real data enhancement may also entail introducing noise to the labels or target values connected to the data samples. Label smoothing supports more reliable predictions by preventing models from overfitting to certain values.

### **Synthetic Data Augmentation:-**In machine learning, synthetic data augmentation creates additional artificial data samples based on current data to increase the training set. It is a method for broadening the variety and volume of data accessible for model training. When a dataset is scarce or more variations are required to boost a model's performance, synthetic data augmentation can be especially helpful

* **Image synthesis:** When dealing with computer vision problems generative models like Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs) can be employed to create new images by combining old ones, using filters or transformations, or even using other techniques. By producing new versions of objects, scenes, or textures, this technique can create duplicates of the original data.
* **Text generation:** In natural language processing tasks, synthetic data augmentation can entail generating new phrases or text samples from existing data. Language models, sequence-to-sequence models, and rule-based approaches can all help with this. Synthetic text data can help improve the model's grasp of diverse sentence forms by increasing the diversity of language patterns.
* **Oversampling and undersampling:** When dealing with imbalanced classification situations in which certain classes are underrepresented in the training data, synthetic data augmentation may include oversampling the minority class or undersampling the majority class. To balance the class distribution, synthetic examples are constructed by duplicating or generating new instances. This reduces the model's bias towards the majority class and enhances its capacity to handle imbalanced data.
* **Data interpolation and extrapolation:** By interpolating or extrapolating existing data samples, synthetic data can be formed. Interpolation involves the generation of new samples that sit between existing data points, whereas extrapolation generates samples that are outside the original data's range. This strategy can assist models learning to predict in previously undiscovered regions of the input space.
* **Feature perturbation:** In synthetic data augmentation, the features or input variables of current data samples can be changed. This can be accomplished by using random noise, transformations, or modifying certain feature values within a legal range. Feature perturbation makes models more resistant to fluctuations in input and increases generalization.
  + 1. **Challenges Faced by Data Augmentation**
* **Maintaining label integrity:** It is critical to guarantee that the labels or ground truth information associated with the enhanced data stay valid when using data augmentation techniques. For example, if a picture is flipped horizontally as part of augmentation, the related label should also reflect the object's flipped version. Maintaining label integrity can be difficult, especially when performing sophisticated transformations or working with more complex data formats.
* **Excessive or incorrect data augmentation can result in overfitting:** Excessive or incorrect data augmentation can result in overfitting, in which the model becomes very specialized in recognizing augmented samples but performs poorly on real-world, unmodified data. If not sufficiently regulated, augmentation can generate false patterns or biases that did not exist in the original data distribution. Models trained on augmented data may struggle to generalize to previously unseen examples.
* **Data augmentation can dramatically increase the size of the training dataset:** Data augmentation can dramatically increase the size of the training dataset, necessitating additional computer resources and time for both data preparation and training. Using complicated augmentation techniques or dealing with huge datasets can be computationally expensive, especially when training deep learning models that require a lot of processing power.
* **Data security and privacy:** Augmentation may entail modifying or producing new data based on current samples. This presents privacy and security problems, especially when working with sensitive or personally identifiable information. It is critical to guarantee that any augmented data generated does not break privacy or ethical standards.
* **Interpretability and explain ability:** Data augmentation can complicate and obscure the model's decision-making process. Variations introduced by augmentation approaches may influence the interpretability of the model's internal representations. Understanding and describing how the model arrived at its predictions can be difficult, especially in crucial situations where interpretability is critical.

**2.8 Data Labelling**

Data labelling in supervised ML models, might be also a part of data preparation process. It can be done manually by crowd workers or automatically using specialized framework.

As the data collected in steps1 may be in undesired format, unorganized or extremely large so further steps are needed to enhance the quality of data. Data labeling is the way of identifying the raw data and adding suitable labels or tags to that data to specify what this data is about, which allows ML models to make an accurate prediction.Data labeling is the way of identifying the raw data and adding suitable labels or tags to that data to specify what this data is about, which allows ML models to make an accurate prediction.

Labels are also known as tags, which are used to give an identification to a piece of data and tell some information about that element. Labels are also referred to as the final output for a prediction. For example, as in the below image, we have labels such as a cat and dog, etc. For audio, labels could be the words that are said. This set of labels lets the ML model learn the dataset, as when we train a model with supervised techniques, we provide the model with a labeled dataset. With this labeled training dataset, the ML model easily predicts the accurate result when given the test dataset.

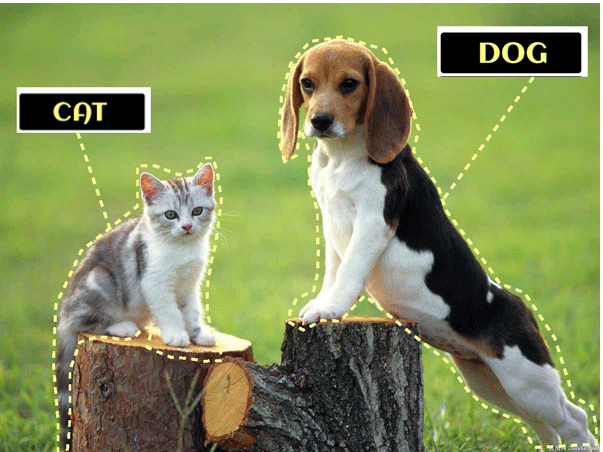


Fig 2.2: Data Labelling

If we input a vast amount of raw data to a Machine Learning model and expect it to learn from it, then it is not enough. As it will give an abrupt result, so it is necessary to pre-process the data, and one of the parts of the pre-processing data stage is Data Labelling. In the data labeling process, we provide some identification to raw data (that may include an image, audio, text) and add some tags to it. These tags tell which class of object the data belongs to, which helps the ML model learn from this data and make the most accurate prediction.

Hence, we can define it as, "Data labelling is a process of adding some meaning to different types of datasets, so that it can be properly used to train a Machine Learning Model. Data labelling is also called as Data Annotation (however, there is minor difference between both of them)."

Data Labelling is required in the case of Supervised Learning, as in supervised learning techniques, we input the labeled data set into the model.

**2.8.1 Labelled Data vs. Unlabelled Data**

* Labelled data is data that has some predefined tags such as name, type, or number. For example, an image has an apple or banana. At the same time, unlabelled data contains no tags or no specified name.
* Labelled data is used in Supervised Learning techniques, whereas unlabelled data is used in Unsupervised Learning.
* Labelled data is difficult to get, whereas Unalabled data is easy to acquire.

**Features in Machine Learning**

Features are the individual independent variables that work as input for the ML system. For an ML dataset, a column can be understood as a feature. ML models use these features to make predictions. Further, we can also get the new features from the old features using feature engineering methods.

We can understand *the difference between both of them with a simple example of an image dataset of animals. So, height, weight, color, etc., are the features. Whereas it is a cat or dog, these are the labels.*

**2.8.2 Approaches to Data Labelling**

Data labeling is an important step while building the high-performance Machine Learning Model. Although the process of data labeling appears easy and simple, it is a bit critical to implement. Therefore, in order to use data labeling techniques, companies should consider multiple factors to find the best approach to labeling. Some common data labeling approaches are given as follows:

* + **Internal/In-house data labelling:** In-house data labeling is performed by data scientists or data engineers of the organization. It is one of the highest quality possible labeling approaches with greater accuracy and simplified tracking. However, it is a time-consuming approach, and it is suitable for companies with extensive resources.
  + Synthetic Labeling: In this approach, new project data is generated from the pre-existing dataset, which increases the quality of the data, and also time efficiency. However, this approach needs high computing power and resources, which enhances the overall cost.
  + **Programmatic Labeling:** Programmatic labeling is an automated process that reduces time consumption and requirement for human annotation as it uses a script. However, besides an automated process, it needs HITL as a part of the QA process to check the possible technical problems.
  + **Outsourcing:** Outsourcing is another popular approach to data labeling, where a team of external labelers is put together in which; most of them are freelancers. This approach can be the best choice for high-level temporary projects; however, it may be a time-consuming approach to develop and manage the freelance-oriented workflow. Although there are various freelancing platforms, such as Upwork, which provide complete candidate information to make the selection process easier, hiring managed data labeling teams provides pre-assessed staff and pre-built data labeling tools.
  + **Crowdsourcing:** Crowdsourcing is one of the fastest and most cost-effective approaches, as it has micro-tasking capabilities and web-based distribution. It obtains annotated data from a large number of freelancers who are registered on a crowdsourcing platform. The datasets that need to be annotated mostly contain data such as images of plants, animals, natural environment, which do not need additional expertise for the annotation. One of the popular examples of crowdsourced data labeling is Recaptcha.

**2.8.3 Tools for Data Labelling**

* + Super Annotate
  + Encord
  + Kili
  + Appen
  + Dataloop
  + Labellers
  + Scale Rapid
  + Ango hub
  + Trainingdat.io

**2.8.4 Benefits and Challenges of Data Labelling**

**Benefits**

* **Precise Predictions:** With accurate data labeling, models can be trained with better quality data and hence generate the expected output. Otherwise, if we provide poor data to the model, then it will generate abrupt results.
* **Better Data Usability:** Data labeling techniques make the data more usable within a model**.** For example, the categorical variables can be reclassified as binary variables to make them more consumable for a model. Therefore, with the aggregation of data, the model can be optimized by reducing the number of variables. Further, high-quality data is always a top priority, whether it is to build computer vision models (i.e., putting bounding boxes around objects) or NLP models (i.e., classifying text for social sentiment).

**Challenges**

* **Costly and time-consuming:** Being one of the crucial steps of building Machine Learning models, data labeling is time consuming and costly process. For a completely automated process also, engineering teams will still need to set up data pipelines prior to data processing, and manual labeling will almost always be costly and time-consuming.
* **Possibilities of Human-Error:** The labeling processes and approaches are prone to human errors, including coding errors or manual entry errors, which degrades the quality of data. The low-quality data leads to inaccurate data processing and modeling. Hence, in order to maintain data quality, quality assurance checks are essential.
  1. **Data Cleaning**

Data cleaning is an essential step in the data science process.

Data cleaning is the process of correcting or deleting inaccurate, damaged, improperly formatted, duplicated, or insufficient data from a dataset.

Even if results and algorithms appear to be correct, they are unreliable if the data is inaccurate.

Data cleaning is an essential step in the data mining process. It is crucial to the construction of a model. The step that is required, but frequently overlooked by everyone, is data cleaning. The major problem with quality information management is data quality. Problems with data quality can happen at any place in an information system. Data cleansing offers a solution to these issues.

Data cleaning is the process of correcting or deleting inaccurate, damaged, improperly formatted, duplicated, or insufficient data from a dataset. Even if results and algorithms appear to be correct, they are unreliable if the data is inaccurate. There are numerous ways for data to be duplicated or incorrectly labeled when merging multiple data sources.

In general, data cleaning lowers errors and raises the caliber of the data. Although it might be a time-consuming and laborious operation, fixing data mistakes and removing incorrect information must be done. A crucial method for cleaning up data is data mining. A method for finding useful information in data is data mining. Data quality mining is a novel methodology that uses data mining methods to find and fix data quality issues in sizable databases. Data mining mechanically pulls intrinsic and hidden information from large data sets. Data cleansing can be accomplished using a variety of data mining approaches. To arrive at a precise final analysis, it is crucial to comprehend and improve the quality of your data. To identify key patterns, the data must be prepared. Exploratory data mining is understood. Before doing business analysis and gaining insights, data cleaning in data mining enables the user to identify erroneous or missing data. There are numerous ways for data to be duplicated or incorrectly labeled when merging multiple data sources. In general, data cleaning lowers errors and improves the quality of the data.

**2.9.1 Steps for Cleaning Data**

* **Remove duplicate or irrelevant observations:** Remove duplicate or pointless observations as well as undesirable observations from your dataset. The majority of duplicate observations will occur during data gathering. Duplicate data can be produced when you merge data sets from several sources, scrape data, or get data from clients or other departments. One of the most important factors to take into account in this procedure is de-duplication. Those observations are deemed irrelevant when you observe observations that do not pertain to the particular issue you are attempting to analyse.
* **Fix structural errors:** When you measure or transfer data and find odd naming practices, typos, or wrong capitalization, such are structural faults. Mislabelled categories or classes may result from these inconsistencies. For instance, "N/A" and "Not Applicable" might be present on any given sheet, but they ought to be analyzed under the same heading.
* **Filter unwanted outliers:** There will frequently be isolated findings that, at first glance, do not seem to fit the data you are analyzing. Removing an outlier if you have a good reason to, such as incorrect data entry, will improve the performance of the data you are working with. However, occasionally the emergence of an outlier will support a theory you are investigating. And just because there is an outlier, that doesn't necessarily indicate it is inaccurate. To determine the reliability of the number, this step is necessary. If an outlier turns out to be incorrect or unimportant for the analysis, you might want to remove it.
* **Handle missing data:** Because many algorithms won't tolerate missing values, you can't overlook missing data. There are a few options for handling missing data. While neither is ideal, both can be taken into account, for example:

Although you can remove observations with missing values, doing so will result in the loss of information, so proceed with caution.

Again, there is a chance to undermine the integrity of the data since you can be working from assumptions rather than actual observations when you input missing numbers based on other observations.

* + - Delete Row
    - Delete column
    - Fill missing value with mean, medium, and mode.
    - Replace missing data with arbitrary value.
* **Validate and QA:** As part of fundamental validation, you ought to be able to respond to the following queries once the data cleansing procedure is complete:
  + Are the data coherent?
  + Does the data abide by the regulations that apply to its particular field?
  + Does it support or refute your working theory? Does it offer any new information?
  + To support your next theory, can you identify any trends in the data?
  + If not, is there a problem with the data's quality?

False conclusions can be used to inform poor company strategy and decision-making as a result of inaccurate or noisy data. False conclusions can result in a humiliating situation in a reporting meeting when you find out your data couldn't withstand further investigation. Establishing a culture of quality data in your organization is crucial before you arrive. The tools you might employ to develop this plan should be documented to achieve this.

## **2.9.2 Techniques for Cleaning Data**

1. **Ignore the tuples:** This approach is not very practical because it is only useful when a tuple has multiple characteristics and missing values.
2. **Fill in the missing value:** This strategy is also not very practical or effective. Additionally, it could be a time-consuming technique. One must add the missing value to the approach. The most common method for doing this is manually, but other options include using attribute means or the most likely value.
3. **Binning method:** This strategy is fairly easy to comprehend. The values nearby are used to smooth the sorted data. The information is subsequently split into several equal-sized parts. The various techniques are then used to finish the assignment.
4. **Regression:** With the use of the regression function, the data is smoothed out. Regression may be multivariate or linear. Multiple regressions have more independent variables than linear regressions, which only have one.
5. **Clustering:** This technique focuses mostly on the group. Data are grouped using clustering. After that, clustering is used to find the outliers. After that, the comparable values are grouped into a "group" or "cluster".

## **2.9.3 Process of Data Cleaning**

The data cleaning method for data mining is demonstrated in the subsequent sections.

1. **Monitoring the errors:** Keep track of the areas where errors seem to occur most frequently. It will be simpler to identify and maintain inaccurate or corrupt information. Information is particularly important when integrating a potential substitute with current management software.
2. **Standardize the mining process:** To help lower the likelihood of duplicity, standardize the place of insertion.
3. **Validate data accuracy:** Analyse the data and spend money on data cleaning software. Artificial intelligence-based tools were utilized to thoroughly check for accuracy.
4. **Scrub for duplicate data:** To save time when analyzing data, find duplicates. By analyzing and investing in independent data-erasing technologies that can analyze imperfect data in quantity and automate the operation, it is possible to avoid again attempting the same data.
5. **Research on data:** Our data needs to be vetted, standardized, and duplicate-checked before this action. There are numerous third-party sources, and these vetted and approved sources can extract data straight from our databases. They assist us in gathering the data and cleaning it up so that it is reliable, accurate, and comprehensive for use in business decisions.
6. **Communicate with the team:** Keeping the group informed will help with client development and strengthening as well as giving more focused information to potential clients.

**2.9.4 Tools for Data Cleaning**

Data Cleansing Tools can be very helpful if you are not confident of cleaning the data yourself or have no time to clean up all your data sets.

* OpenRefine
* Trifacta Wrangler
* Drake
* Data Ladder
* Data Cleaner
* Cloudingo
* Reifier
* IBM Infosphere Quality Stage
* TIBCO Clarity
* Winpure

**2.9.5** **Benefits of Data Cleaning**

* Removal of inaccuracies when several data sources are involved.
* Clients are happier and employees are less annoyed when there are fewer mistakes.
* The capacity to map out the many functions and the planned uses of your data.
* Monitoring mistakes and improving reporting make it easier to resolve inaccurate or damaged data for future applications by allowing users to identify where issues are coming from.
* Making decisions more quickly and with greater efficiency will be possible with the use of data cleansing tools.

2.10 **Data Integration**

Data integration is the process of merging data from several disparate sources.

While performing data integration, you must work on data redundancy, inconsistency, duplicity, etc.Data integration is the process of combining data from many sources. Data integration must related with issues such as duplicated data, inconsistent data, old systems, etc.Data integration is a record pre-processing method that includes merging data from a couple of the different data sources into coherent data to retain and provide a unified perspective of the data. These assets could also include several record cubes, databases, or flat documents.

It has been an integral part of data operations because data can be obtained from several sources. It is a strategy that integrates data from several sources to make it available to users in a single uniform view that shows their status. There are communication sources between systems that can include multiple databases, data cubes, or flat files. Data fusion merges data from various diverse sources to produce meaningful results. The consolidated findings must exclude inconsistencies, contradictions, redundancies, and inequities.Data integration is important because it gives a uniform view of scattered data while also maintaining data accuracy. It assists the data-mining program in meaningful mining information, which in turn assists the executive and managers make strategic decisions for the enterprise's benefit. It has been an integral part of data operations because data can be obtained from several sources. It is a strategy that integrates data from several sources to make it available to users in a single uniform view that shows their status. Data fusion merges data from various diverse sources to produce meaningful results. The consolidated findings must exclude inconsistencies, contradictions, redundancies, and inequities.

**2.10.1** **Why it is Important?**

Data integration is important because it gives a uniform view of scattered data while also maintaining data accuracy. It assists the data-mining program in meaningful mining information, which in turn assists the executive and managers make strategic decisions for the enterprise's benefit.

The data integration methods are formally characterized as a triple (G, S, M), where;

* G represents the global schema,
* S represents the heterogeneous source of schema,
* M represents the mapping between source and global schema queries.

**2.10.2 Data Integration Approaches**

* Tight Coupling
* Loose Coupling

### **Tight Coupling**

It is the process of using ETL (Extraction, Transformation, and Loading) to combine data from various sources into a single physical location.

### **Loose Coupling**

Facts with loose coupling are most effectively kept in the actual source databases. This approach provides an interface that gets a query from the user, changes it into a format that the supply database may understand, and then sends the query to the source databases without delay to obtain the result.

**2.10.3 Issues in Data Integration**

* **Entity Identification Problem:** As you understand, the records are obtained from heterogeneous sources, and how can you 'match the real-world entities from the data'. For example, you were given client data from specialized statistics sites. Customer identity is assigned to an entity from one statistics supply, while a customer range is assigned to an entity from another statistics supply. Analyzing such metadata statistics will prevent you from making errors during schema integration.
* Structural integration is completed by guaranteeing that the functional dependency and referential constraints of a character in the source machine match the functional dependency and referential constraints of the identical character in the target machine. For example, assume that the discount is applied to the entire order in one machine, but in every other machine, the discount is applied to each item in the order. This distinction should be noted before the information from those assets is included in the goal system.
* **Redundancy and Correlation Analysis:** One of the major issues in the course of data integration is redundancy. Unimportant data that are no longer required are referred to as redundant data. It may also appear due to attributes created from the use of another property inside the information set. For example, if one truth set contains the patronage and distinct data set as the purchaser's date of the beginning, then age may be a redundant attribute because it can be deduced from the use of the beginning date.
* Inconsistencies further increase the level of redundancy within the characteristic. The use of correlation analysis can be used to determine redundancy. The traits are examined to determine their interdependence on each difference, consequently discovering the link between them.
* **Tuple Duplication:** Information integration has also handled duplicate tuples in addition to redundancy. Duplicate tuples may also appear in the generated information if the denormalized table was utilized as a deliverable for data integration.
* **Data warfare Detection and backbone:** The data warfare technique of combining records from several sources is unhealthy. In the same way, that characteristic values can vary, so can statistics units. The disparity may be related to the fact that they are represented differently within the special data units. For example, in one-of-a-kind towns, the price of an inn room might be expressed in a particular currency. This type of issue is recognized and fixed during the data integration process

## **2.10.4 Data Integration Techniques**

### **Manual Integration**

This method avoids using automation during data integration. The data analyst collects, cleans, and integrates the data to produce meaningful information. This strategy is suitable for a mini organization with a limited data set. Although, it will be time-consuming for the huge, sophisticated, and recurring integration. Because the entire process must be done manually, it is a time-consuming operation.

### **Middleware Integration**

The middleware software is used to take data from many sources, normalize it, and store it in the resulting data set. When an enterprise needs to integrate data from legacy systems to modern systems, this technique is used. Middleware software acts as a translator between legacy and advanced systems. You may take an adapter that allows two systems with different interfaces to be connected. It is only applicable to certain systems.

### **Application-based integration**

It is using software applications to extract, transform, and load data from disparate sources. This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding. This strategy saves time and effort, but it is a little more complicated because building such an application necessitates technical understanding.

### **Uniform Access Integration**

This method combines data from a more disparate source. However, the data's position is not altered in this scenario; the data stays in its original location. This technique merely generates a unified view of the integrated data. The integrated data does not need to be stored separately because the end-user only sees the integrated view.

### **Data Warehousing**

This technique is related to the uniform access integration technique in a roundabout way. The unified view, on the other hand, is stored in a different location. It enables the data analyst to deal with more sophisticated inquiries. Although it is a promising solution and increased storage costs, the unified data's view or copy requires separate storage and maintenance costs.

**2.10.5 Integration tools**

* + On-promise data integration tool: An on-premise data integration tool integrates data from local sources and connects legacy databases using middleware software.
  + Open-source data integration tool: If you want to avoid pricey enterprise solutions, an open-source data integration tool is the ideal alternative. Although, you will be responsible for the security and privacy of the data if you're using the tool.
  + Cloud-based data integration tool: A cloud-based data integration tool may provide an **'integration platform as a service'**.
  1. **Data Reduction**

Data reduction techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume.

Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data.

By reducing the data, the efficiency of the data analysis process is improved, which produces the same analytical results.

**Importance of Data Reduction**

Data reduction aims to define it more compactly. When the data size is smaller, it is simpler to apply sophisticated and computationally high-priced algorithms.

The reduction of the data may be in terms of the number of rows (records) or terms of the number of columns (dimensions).

**2.10.1** **Techniques of Data Reduction**

* Dimensionality Reduction
* Numerosity Reduction
* Data Cube Aggregation
* Data Compression

1. **Dimensionality Reduction**

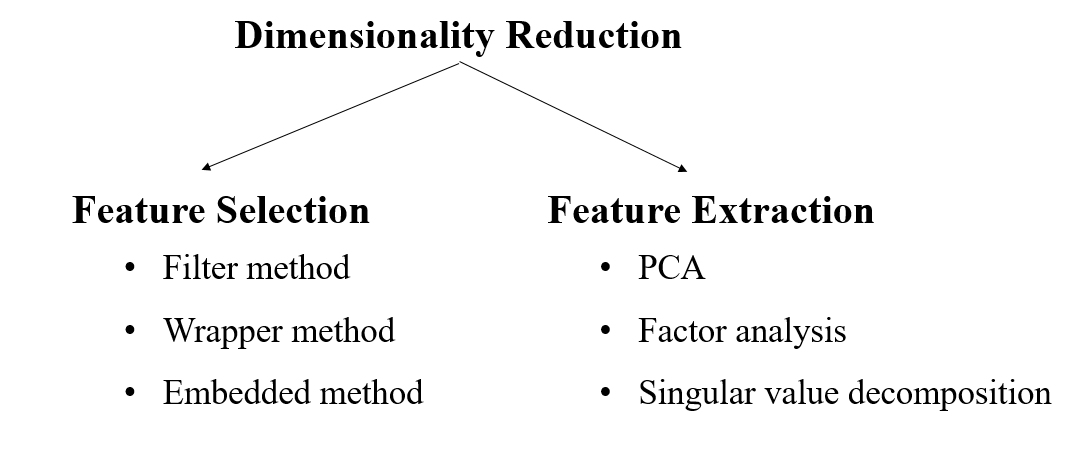
Dimensionality reduction eliminates the attributes from the data set under consideration, thereby reducing the volume of original data. It reduces data size as it eliminates outdated or redundant features. It is the process of reducing the number of attributes or random variables under consideration. The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.

**Definition**

Dimensionality reduction technique can be defined as, "It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information." These techniques are widely used in machine learning for obtaining a better predictive model while solving the classification and regression problems.It is commonly used in the fields that deal with high-dimensional data, such as speech recognition, signal processing, bioinformatics, etc. It can also be used for data visualization, noise reduction, cluster analysis, etc. A dataset contains a huge number of input features in various cases, which makes the predictive modeling task more complicated. Because it is very difficult to visualize or make predictions for the training dataset with a high number of features, for such cases, dimensionality reduction techniques are required to use.

**Methods of Dimensionality Reduction**

* Feature Selection
* Feature Extraction
* Wavelet Transform
* Attribute Subset Selection



**Fig 2.3** **Dimensionality Reduction**

**Feature Selection:-**It is the process of selecting the subset of relevant features and leaving the irrelevant features present in a dataset to build a model of high accuracy.

Filter method:-Filter out the irrelevant features and redundant columns from the model by using different metrics.

Wrapper method:-Evaluates the usefulness of features

Embedded method:-finds the error and adds necessary features and remove unnecessary features.

Feature Extraction:-It is the process of transforming space containing many dimensions into space with fewer dimensions.

[**Principal Component Analysis**](https://www.geeksforgeeks.org/principal-component-analysis-with-python/) **(PCA)** technique was introduced by the mathematician **Karl Pearson** in 1901**.** It works on the condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

**How Principal Component Analysis (PCA) Works?**

Principal Component Analysis (PCA) is used to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retaining most of the sample’s information, and useful for the regression and classification of data.**Principal Component Analysis (PCA)**is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables.PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, Principal Component Analysis (PCA) is an [unsupervised learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) algorithm technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit. The main goal of Principal Component Analysis (PCA) is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables. Suppose we have a data set to be analyzed that has tuples with n attributes. The principal component analysis identifies k independent features with n attributes that can represent the data set. In this way, the original data can be converted into a much smaller space, and dimensionality reduction can be achieved. Principal component analysis can be applied to sparse and skewed data.

**Factor Analysis**

* Technique to reduce a large number of variables into fewer number of factors.
* There will not be any outliers in the dataset.
* Used to explore interactions between variables.

**Singular Value Decomposition**

* SVD is like putting on special glasses that help us see the essential patterns in our data more clearly, making it easier to understand and work with.
* SVD helps us simplify the data. It finds the most important patterns in the data and focuses on them, ignoring the less important stuff.

**Wavelet Transform**

* In the wavelet transform, suppose a data vector A is transformed into a numerically different data vector A' such that both A and A' vectors are of the same length.
* Then how it is useful in reducing data because the data obtained from the wavelet transform can be interrupted.
* Wavelet transform can be applied to data cubes, sparse data, or skewed data.

**Attribute Subset Selection**

* The large data set has many attributes, some of which are irrelevant and some are redundant.
* The core attribute subset selection reduces the data volume and dimensionality.
* The attribute subset selection reduces the volume of data by eliminating redundant and irrelevant attributes.
* The attribute subset selection ensures that we get a good subset of original attributes even after eliminating the unwanted attributes.
* The resulting probability of data distribution is as close as possible to the original data distribution using all the attributes.

1. **Numerosity Reduction**

* The numerosity reduction reduces the original data volume and represents it in a much smaller form.
* It reduces the data volume by chossing alternative smaller forms of data representation.

**Types of numerosity reduction**

**Parametric numerosity reduction**

Parametric numerosity reduction incorporates storing only data parameters instead of the original data. One method of parametric numerosity reduction is the regression and log-linear method.

Regression and Log-Linear: Linear regression models a relationship between the two attributes by modeling a linear equation to the data set. Suppose we need to model a linear function between two attributes.

y = wx +b

Here, y is the response attribute, and x is the predictor attribute. If we discuss in terms of data mining, attribute x and attribute y are the numeric database attributes, whereas w and b are regression coefficients. Multiple linear regressions let the response variable y model linear function between two or more predictor variables. Log-linear model discovers the relation between two or more discrete attributes in the database. Suppose we have a set of tuples presented in n-dimensional space. Then the log-linear model is used to study the probability of each tuple in a multidimensional space. Regression and log-linear methods can be used for sparse data and skewed data.

**Non-parametric numerosity reduction.**

A non-parametric numerosity reduction technique does not assume any model. The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric. There are at least four types of Non-Parametric data reduction techniques, Histogram, Clustering, Sampling, Data Cube Aggregation, and Data Compression.

**Histogram:** A histogram is a graph that represents frequency distribution which describes how often a value appears in the data. Histogram uses the binning method to represent an attribute’s data distribution. It uses a disjoint subset which we call bin or buckets.  
A histogram can represent a dense, sparse, uniform, or skewed data. Instead of only one attribute, the histogram can be implemented for multiple attributes. It can effectively represent up to five attributes.

**Clustering:** Clustering techniques groups similar objects from the data so that the objects in a cluster are similar to each other, but they are dissimilar to objects in another cluster.  
How much similar are the objects inside a cluster can be calculated using a distance function. More is the similarity between the objects in a cluster closer they appear in the cluster.  
The quality of the cluster depends on the diameter of the cluster, i.e., the max distance between any two objects in the cluster. The cluster representation replaces the original data. This technique is more effective if the present data can be classified into a distinct clustered.

**Sampling:** One of the methods used for data reduction is sampling, as it can reduce the large data set into a much smaller data sample.

1. **Data Cube Aggregation**

This technique is used to aggregate data in a simpler form. Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction. The data cube aggregation is a multidimensional aggregation that eases multidimensional analysis. The data cube present precomputed and summarized data which eases the data mining into fast access.

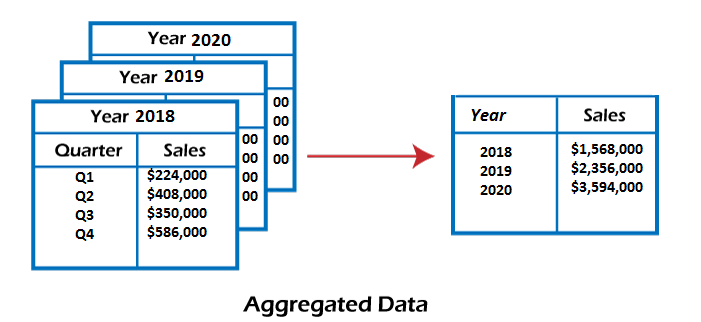


Fig 2.4: Data Cube Aggregation

**4. Data Compression**

Data compression employs modification, encoding, or converting the structure of data in a way that consumes less space. Data compression involves building a compact representation of information by removing redundancy and representing data in binary form. Data that can be restored successfully from its compressed form is called Lossless compression. In contrast, the opposite where it is not possible to restore the original form from the compressed form is Lossy compression. Dimensionality and numerosity reduction method are also used for data compression.

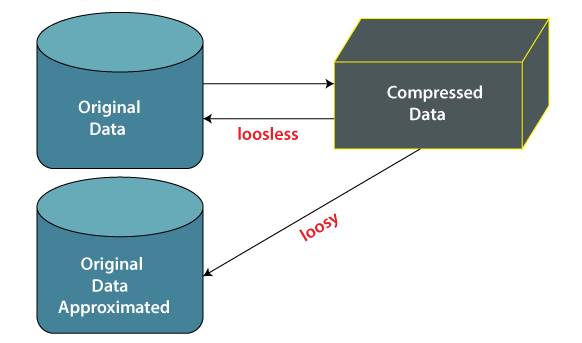


Fig 2.5: Data Compression

**Data compression techniques**

**Lossless Compression: -** Encoding techniques (Run Length Encoding) allow a simple and minimal data size reduction.

Lossless data compression uses algorithms to restore the precise original data from the compressed data.

**Lossy Compression: -** In lossy-data compression, the decompressed data may differ from the original data but are useful enough to retrieve information from them.

For example, the JPEG image format is a lossy compression, but we can find the meaning equivalent to the original image.

Methods such as the Discrete Wavelet transform technique PCA (principal component analysis) are examples of this compression.

* 1. **Data Transformation**

Raw data is difficult to trace or understand. That's why it needs to be preprocessed before retrieving any information from it. Data transformation is a technique used to **convert** the raw data into a suitable format that efficiently eases data mining and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

Data transformation is an essential data preprocessing technique that must be performed on the data before data mining to provide patterns that are easier to understand.

Data transformation changes the format, structure, or values of the data and converts them into [clean, usable data](https://www.zuar.com/blog/data-cleaning-the-benefits-and-steps-to-creating-and-using-clean-data/). Data may be transformed at two stages of the data pipeline for data analytics projects. Organizations that use on-premises data warehouses generally use an ETL (extract, transform, and load) process, in which data transformation is the middle step. Today, most organizations use cloud-based data warehouses to scale compute and storage resources with latency measured in seconds or minutes. The scalability of the cloud platform lets organizations skip preload transformations and load raw data into the data warehouse, then transform it at query time.

*Data integration, migration, data warehousing, data wrangling* may all involve data transformation. Data transformation increases the efficiency of business and [analytic processes](https://www.zuar.com/blog/data-automation-improve-analysis-productivity/), and it enables businesses to make better data-driven decisions. During the data transformation process, an analyst will determine the structure of the data. This could mean that data transformation may be:

* Constructive: The data transformation process adds, copies, or replicates data.
* Destructive: The system deletes fields or records.
* Aesthetic: The transformation standardizes the data to meet requirements or parameters.
* Structural: The database is reorganized by renaming, moving, or combining columns.

### **2.11.1 Data Transformation Techniques**

**Data Smoothing**

Data smoothing is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

We have seen how the noise is removed from the data using the techniques such as binning, regression, clustering.

* Binning: This method splits the sorted data into the number of bins and smoothens the data values in each bin considering the neighborhood values around it.
* Regression: This method identifies the relation among two dependent attributes so that if we have one attribute, it can be used to predict the other attribute.
* Clustering: This method groups similar data values and form a cluster. The values that lie outside a cluster are known as outliers.

**Data Normalization**

Normalizing the data refers to scaling the data values to a much smaller range such as [-1, 1] or [0.0, 1.0]. There are different methods to normalize the data, as discussed below.

Consider that we have a numeric attribute A and we have n number of observed values for attribute A that are V1, V2, V3, ….Vn.

* **Min-max normalization:** This method implements a linear transformation on the original data. Let us consider that we have minAand maxA as the minimum and maximum value observed for attribute A and Viis the value for attribute A that has to be normalized.  
  The min-max normalization would map Vito the V'i in a new smaller range [new\_minA, new\_maxA]. The formula for min-max normalization is given below

For example, we have 1200 and 9800 as the minimum, and maximum value for the attribute income, and [0.0, 1.0] is the range in which we have to map a value of 73,600.  
The value 73,600 would be transformed using min-max normalization as follows:

(73600-1200)/(9800-1200)\*(1.0-0.0)+0.0=0.716

**Z-score normalization:** This method normalizes the value for attribute A using the ***mean***and ***standard deviation***. The following formula is used for Z-score normalization:  
Data Transformation in Data Mining  
Here Ᾱ and σAare the mean and standard deviation for attribute A, respectively.  
For example, we have a mean and standard deviation for attribute A as $54,000 and $16,000. And we have to normalize the value $73,600 using z-score normalization.  
Data Transformation in Data Mining

**Decimal Scaling:** This method normalizes the value of attribute A by moving the decimal point in the value. This movement of a decimal point depends on the maximum absolute value of A. The formula for the decimal scaling is given below:  
Data Transformation in Data Mining  
Here j is the smallest integer such that max(|*v'i*|)<1  
For example, the observed values for attribute A range from -986 to 917, and the maximum absolute value for attribute A is 986. Here, to normalize each value of attribute A using decimal scaling, we have to divide each value of attribute A by 1000, i.e., j=3.  
So, the value -986 would be normalized to -0.986, and 917 would be normalized to 0.917.  
The normalization parameters such as mean, standard deviation, the maximum absolute value must be preserved to normalize the future data uniformly.

**Data Discretization**

This is a process of converting continuous data into a set of data intervals. Continuous attribute values are substituted by small interval labels. This makes the data easier to study and analyze. If a data mining task handles a continuous attribute, then its discrete values can be replaced by constant quality attributes. This improves the efficiency of the task.

This method is also called a data reduction mechanism as it transforms a large dataset into a set of categorical data. Discretization also uses decision tree-based algorithms to produce short, compact, and accurate results when using discrete values.

Data discretization can be classified into two types: **supervised discretization**, where the class information is used, and **unsupervised discretization**, which is based on which direction the process proceeds, i.e., 'top-down splitting strategy' or 'bottom-up merging strategy'. For example, the values for the age attribute can be replaced by the interval labels such as (0-10, 11-20…) or (kid, youth, adult, senior).

**Data Generalization**

It converts low-level data attributes to high-level data attributes using concept hierarchy. This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data. Data generalization can be divided into two approaches:

Data cube process (OLAP) approach.

Attribute-oriented induction (AOI) approach.

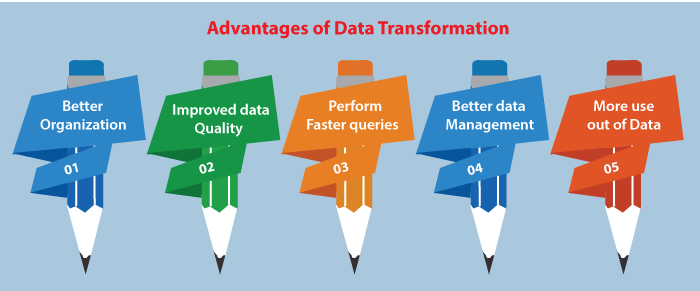
For example, age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical value (young, old).

### **2.11.2 Data Transformation Process**

The entire process for transforming data is known as [**ETL**](https://www.sqlshack.com/an-overview-of-etl-and-elt-architecture/) **(Extract, Load, and Transform)**. Through the ETL process, analysts can convert data to its desired format. Here are the steps involved in the data transformation process:

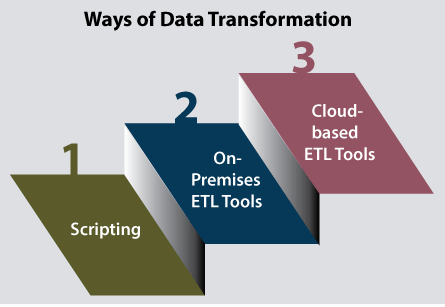
1. [**Data Discovery**](https://www.dataversity.net/analytics-vs-data-discovery/): During the first stage, analysts work to understand and identify data in its source format. To do this, they will use data profiling tools. This step helps analysts decide what they need to do to get data into its desired format.
2. **Data Mapping:** During this phase, analysts perform data mapping to determine how individual fields are modified, mapped, filtered, joined, and aggregated. Data mapping is essential to many data processes, and one misstep can lead to incorrect analysis and ripple through your entire organization.
3. **Data Extraction:** During this phase, analysts extract the data from its original source. These may include structured sources such as databases or streaming sources such as customer log files from web applications.
4. **Code Generation and Execution:** Once the data has been extracted, analysts need to create a code to complete the transformation. Often, analysts generate codes with the help of data transformation platforms or tools.
5. **Review:** After transforming the data, analysts need to check it to ensure everything has been formatted correctly.
6. **Sending:** The final step involves sending the data to its target destination. The target might be a data warehouse or a database that handles both structured and unstructured data.

### **2.11.3 Advantages of Data Transformation**



* **Better Organization:** Transformed data is easier for both humans and computers to use.
* **Improved Data Quality:** There are many **risks and costs associated with bad data**. Data transformation can help your organization eliminate quality issues such as missing values and other inconsistencies.
* **Perform Faster Queries:** You can quickly and easily retrieve transformed data thanks to it being stored and standardized**in a source location.**
* **Better Data Management:** Businesses are constantly generating data from more and more sources. If there are inconsistencies in the metadata, it can be **challenging to organize and understand it.** Data transformation refines your metadata, so it's easier to organize and understand.
* **More Use Out of Data:** While businesses may be collecting data constantly, a lot of that data **sits around unanalyzed**. Transformation makes it easier to get the most out of your data by standardizing it and making it more usable.

**2.11.4** **Disadvantages of Data Transformation**



**Data transformation can be expensive.** The cost is dependent on the specific infrastructure, software, and tools used to process data. Expenses may include licensing, computing resources, and hiring necessary personnel.

**Data transformation processes can be resource-intensive.** Performing transformations in an on-premises data warehouse after loading or transforming data before feeding it into applications can create a computational burden that slows down other operations. If you use a cloud-based data warehouse, you can do the transformations after loading because the platform can scale up to meet demand.

**Lack of expertise and carelessness can introduce problems during transformation**. Data analysts without appropriate subject matter expertise are less likely to notice incorrect data because they are less familiar with the range of accurate and permissible values.

**Enterprises can perform transformations that don't suit their needs.** A business might change information to a specific format for one application only to then revert the information to its prior format for a different application.

### **2.11.5 Ways of Data Transformation**

**Scripting:** Data transformation through scripting involves ***Python* or *SQL*** to write the code to extract and transform data. Python and SQL are scripting languages that allow you to automate certain tasks in a program. They also allow you to extract information from data sets. Scripting languages require less code than traditional programming languages. Therefore, it is less intensive.

**On-Premises ETL Tools:** ETL tools take the required work to script the data transformation by automating the process. On-premises ETL tools are hosted on company servers. While these tools can help save you time, using them often **requires extensive expertise and significant infrastructure costs**.

**Cloud-Based ETL Tools:** As the name suggests, cloud-based ETL tools are **hosted in the cloud**. These tools are often the easiest for non-technical users to utilize. They allow you to collect data from any cloud source and load it into your data warehouse. With cloud-based ETL tools, you can decide how often you want to pull data from your source, **and you can monitor your usage**.