BECE352E - IoT Domain Analyst

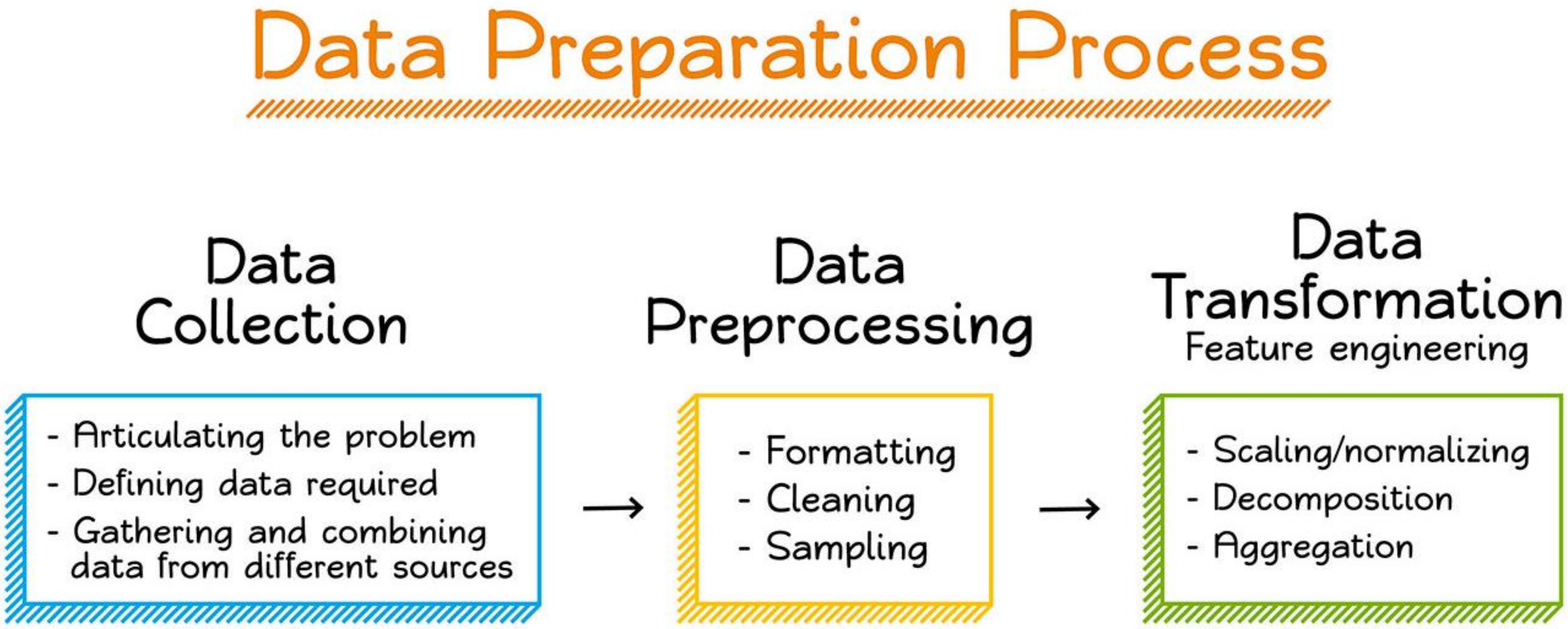
Dr.B.Nagajayanthi

SENSE

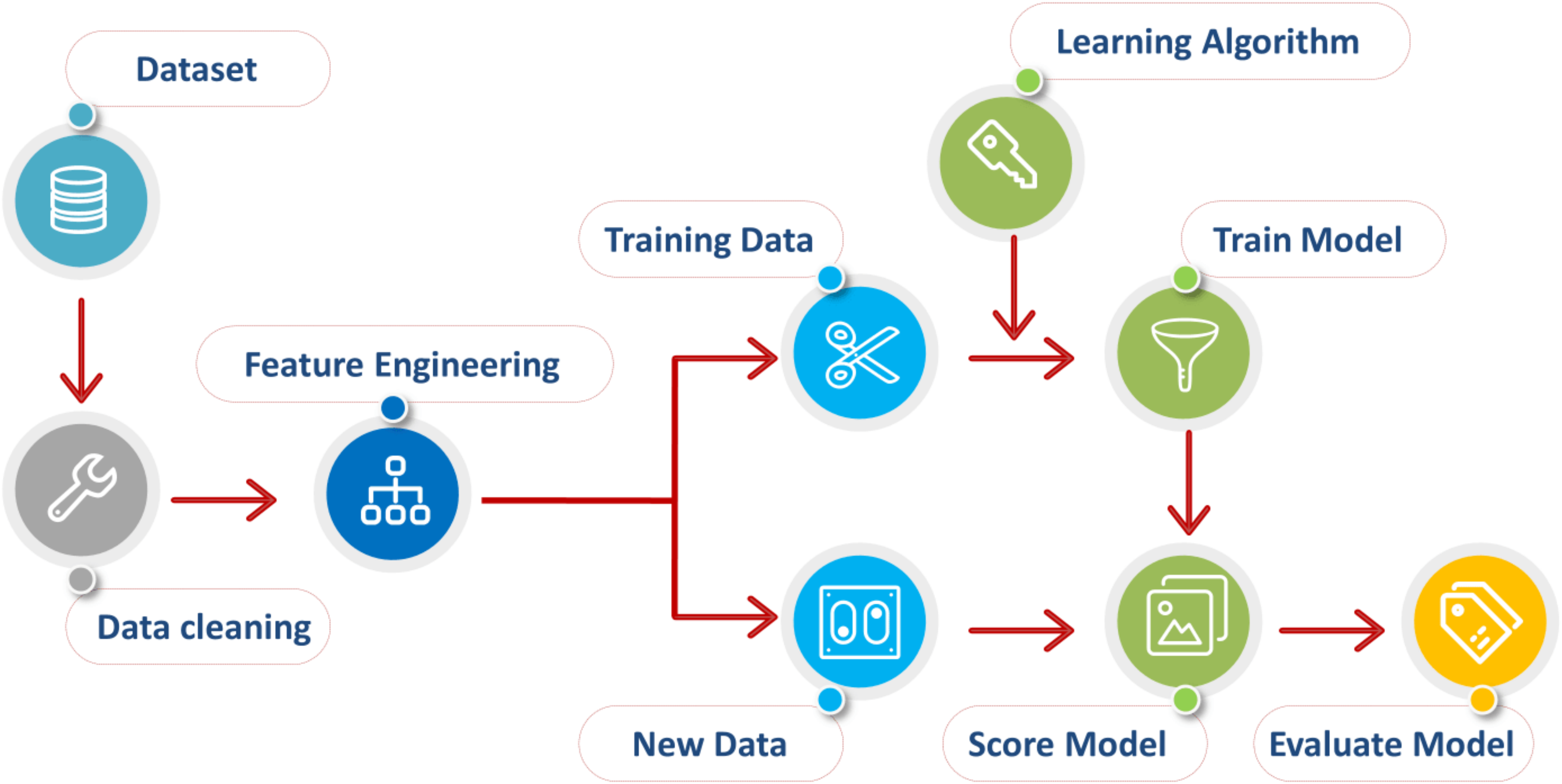
Associate Professor

Module-2-Data Preprocess and EDA

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Significance of Exploratory Data Analysis
- Making sense of Data.

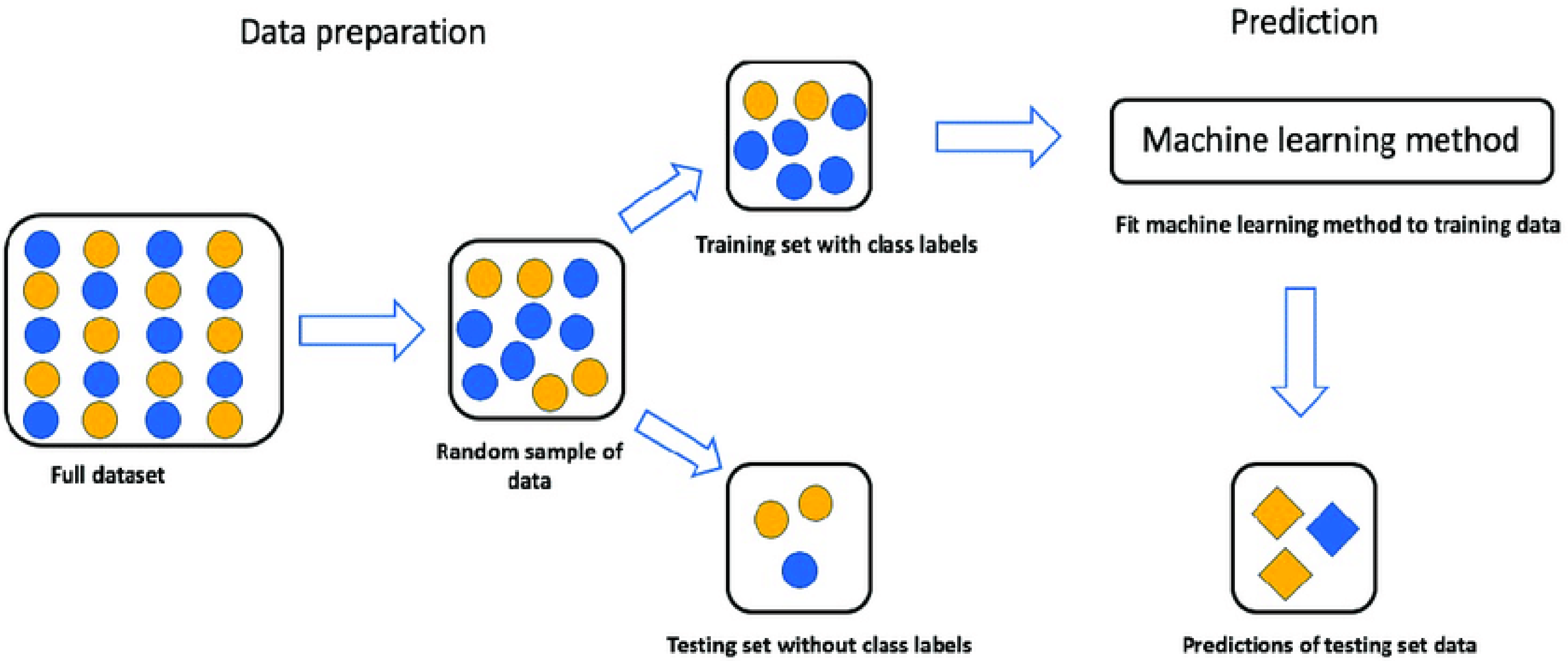


What is Data preprocessing?



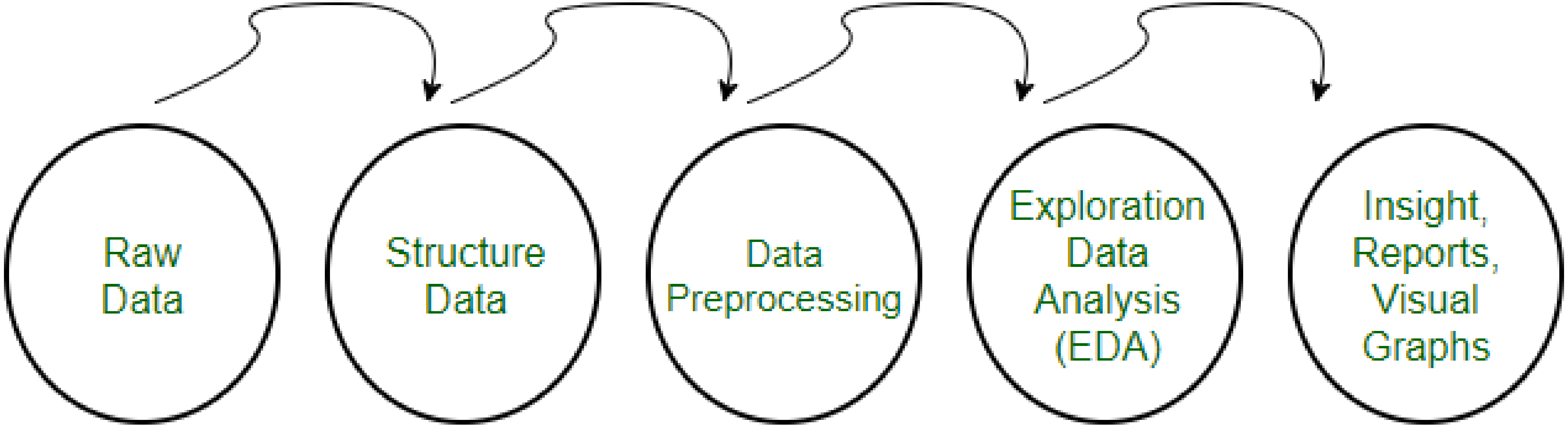
- **Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.**
- Whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

What is Data preprocessing?


Data preparation
Prediction
Machine learning method
Fit machine learning method to training data
Training set with class labels
Full dataset
Random sample of data
Testing set without class labels
Predictions of testing set data

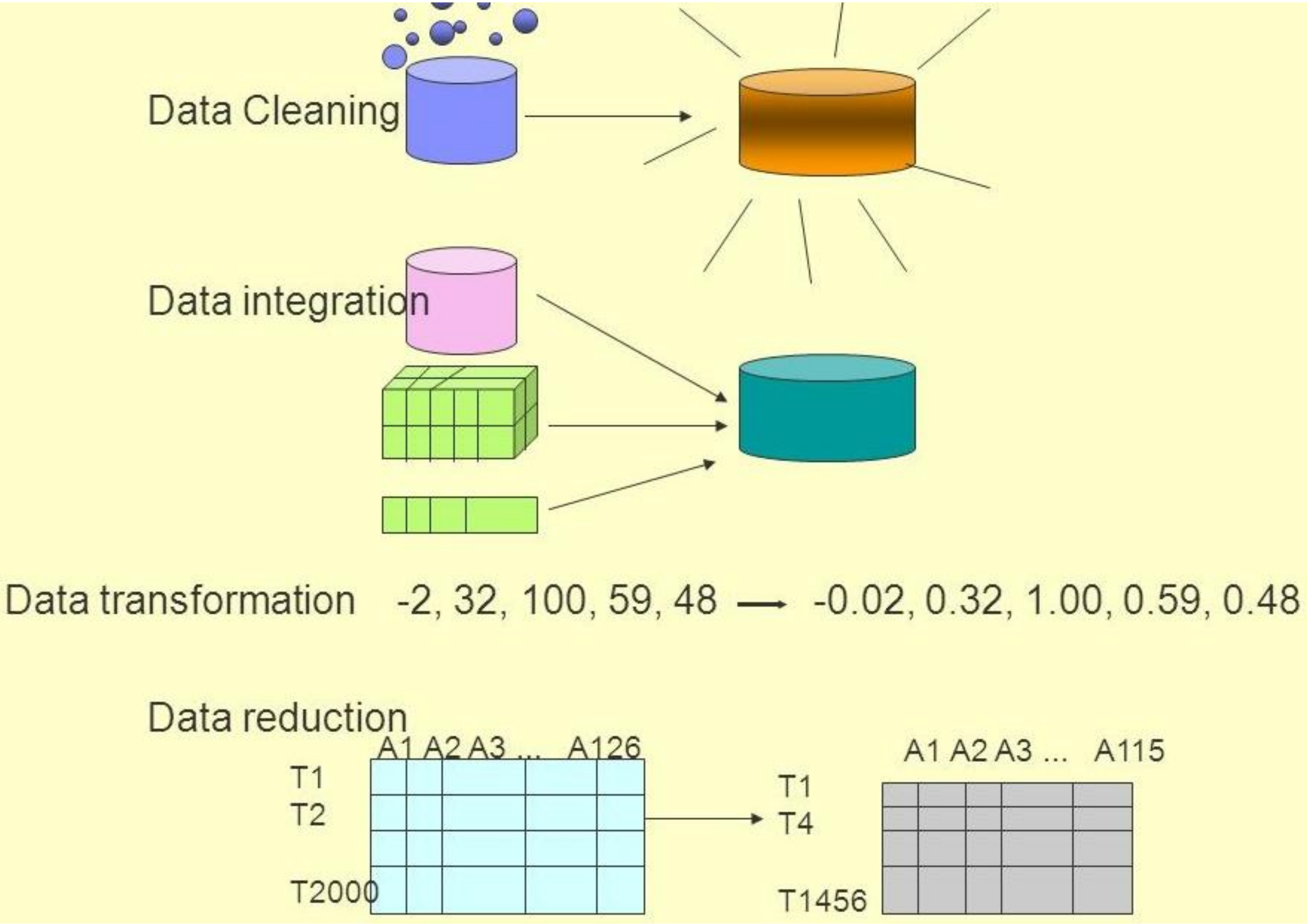
- **A technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.**
- Any type of processing performed on raw data to prepare it for another data processing procedure.

Need for Data preprocessing?

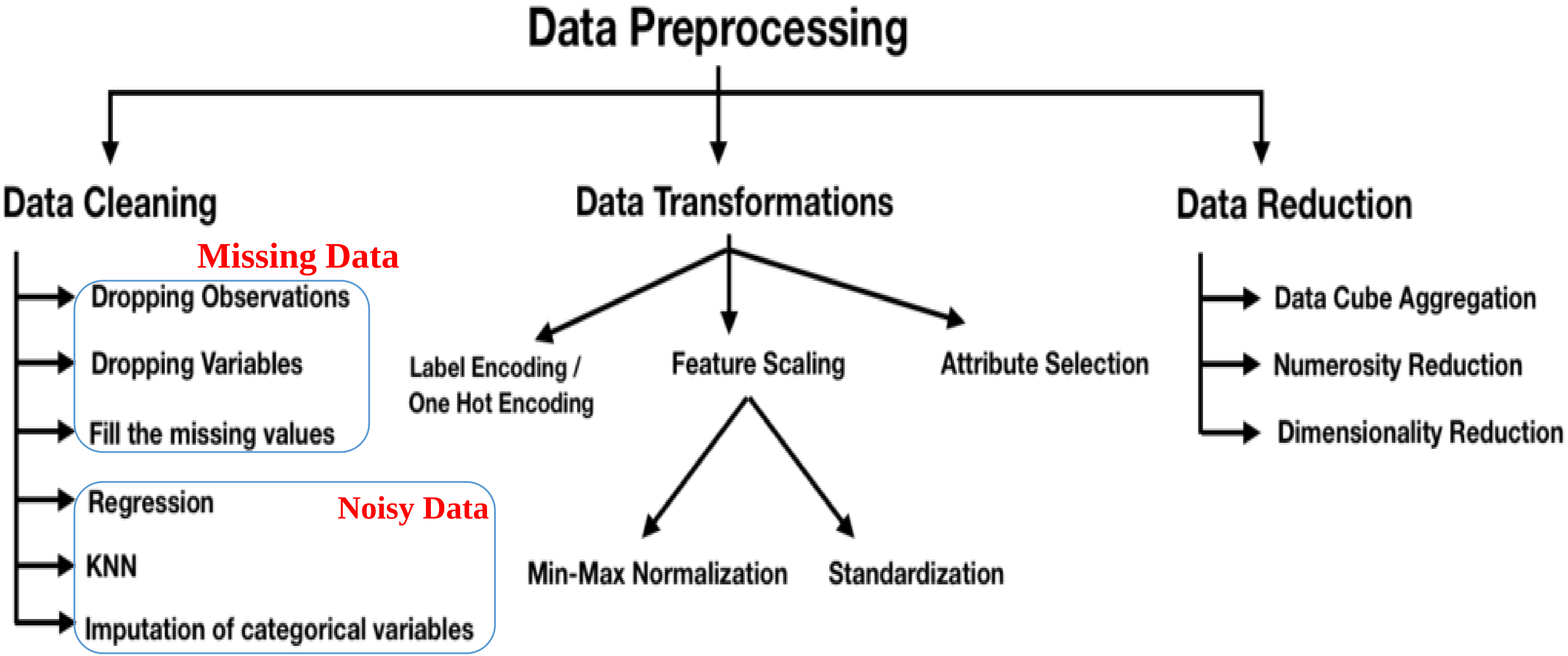

Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

- For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner.
- Some specified **Machine Learning model needs information in a specified format,** for example, **Random Forest algorithm does not support null values**, therefore to execute random forest algorithm null values have to be managed from the original raw data set.
- Another aspect is that the **data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and best out of them is chosen**.

Forms of Data preprocessing



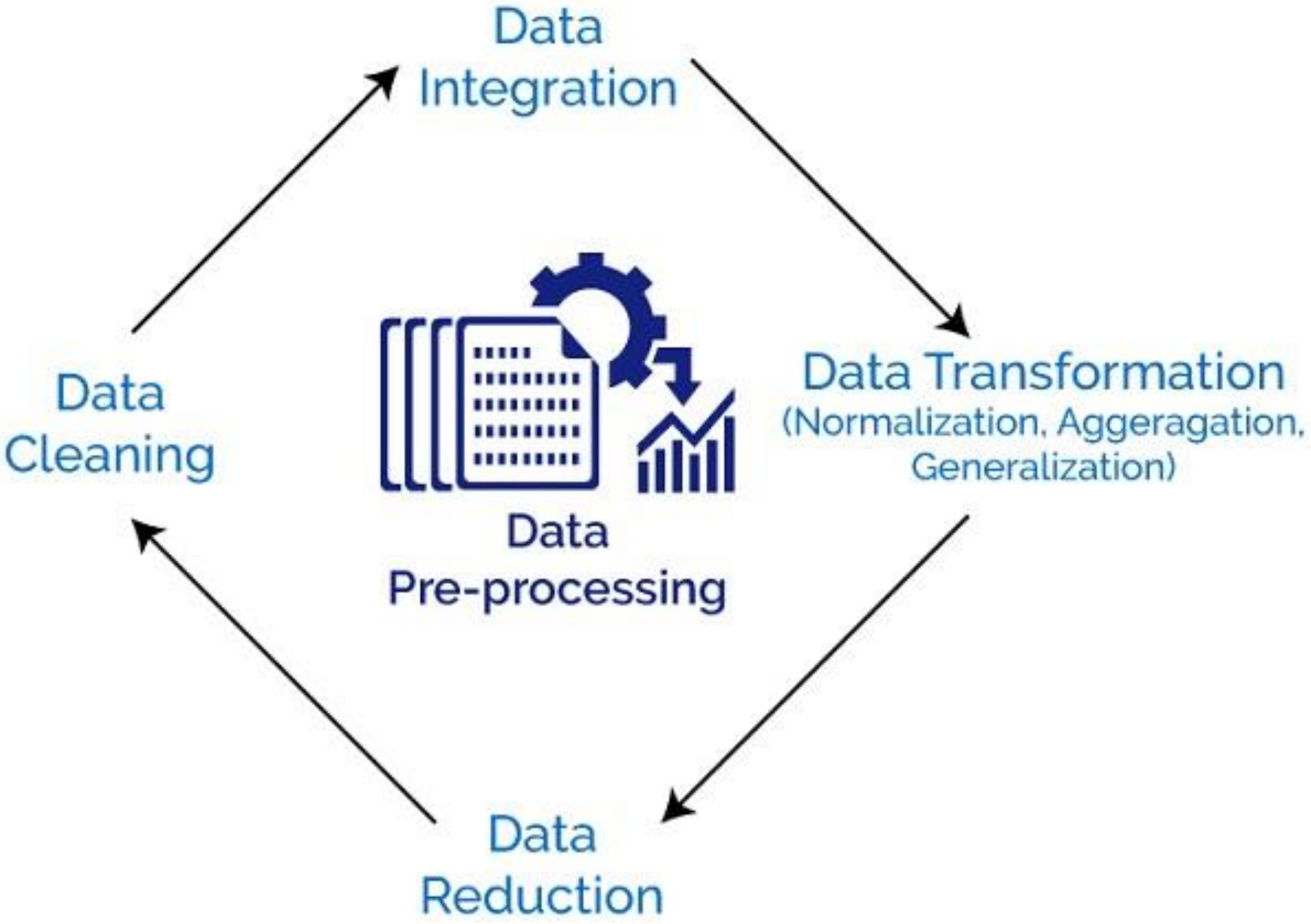
Forms of Data preprocessing



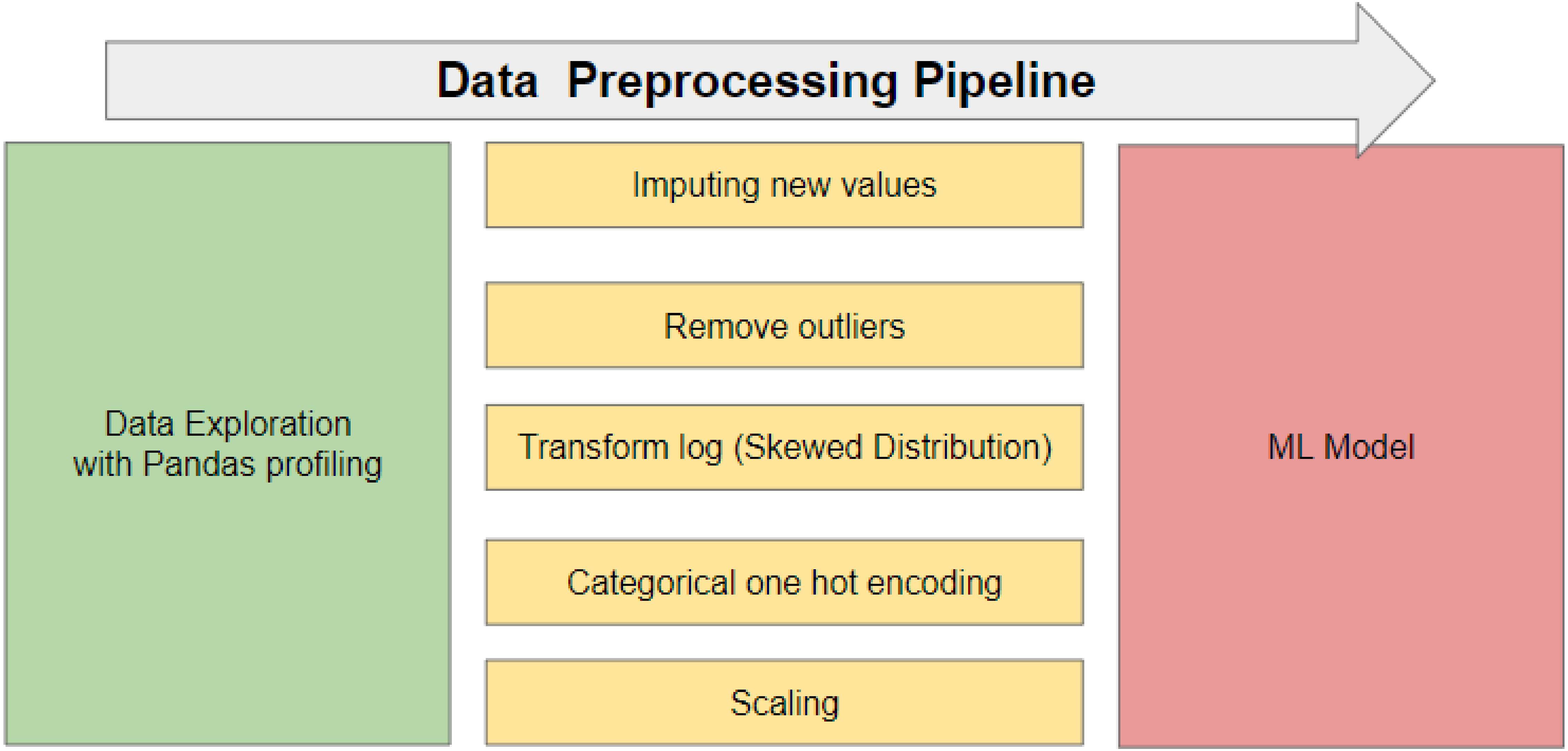
Data Preprocessing

- Data Cleaning
  - Missing Data
    - Dropping Observations
    - Dropping Variables
    - Fill the missing values
  - Noisy Data
    - Regression
    - KNN
    - Imputation of categorical variables
- Data Transformations
  - Label Encoding / One Hot Encoding
  - Feature Scaling
    - Min-Max Normalization
    - Standardization
  - Attribute Selection
- Data Reduction
  - Data Cube Aggregation
  - Numerosity Reduction
  - Dimensionality Reduction

Sequence of Data preprocessing


Data Integration
Data Transformation (Normalization, Aggeragation, Generalization)
Data Reduction
Data Cleaning
Data Pre-processing

Effective Data preprocessing for a ML model



Data Preprocessing Pipeline

Data Exploration with Pandas profiling

Imputing new values

Remove outliers

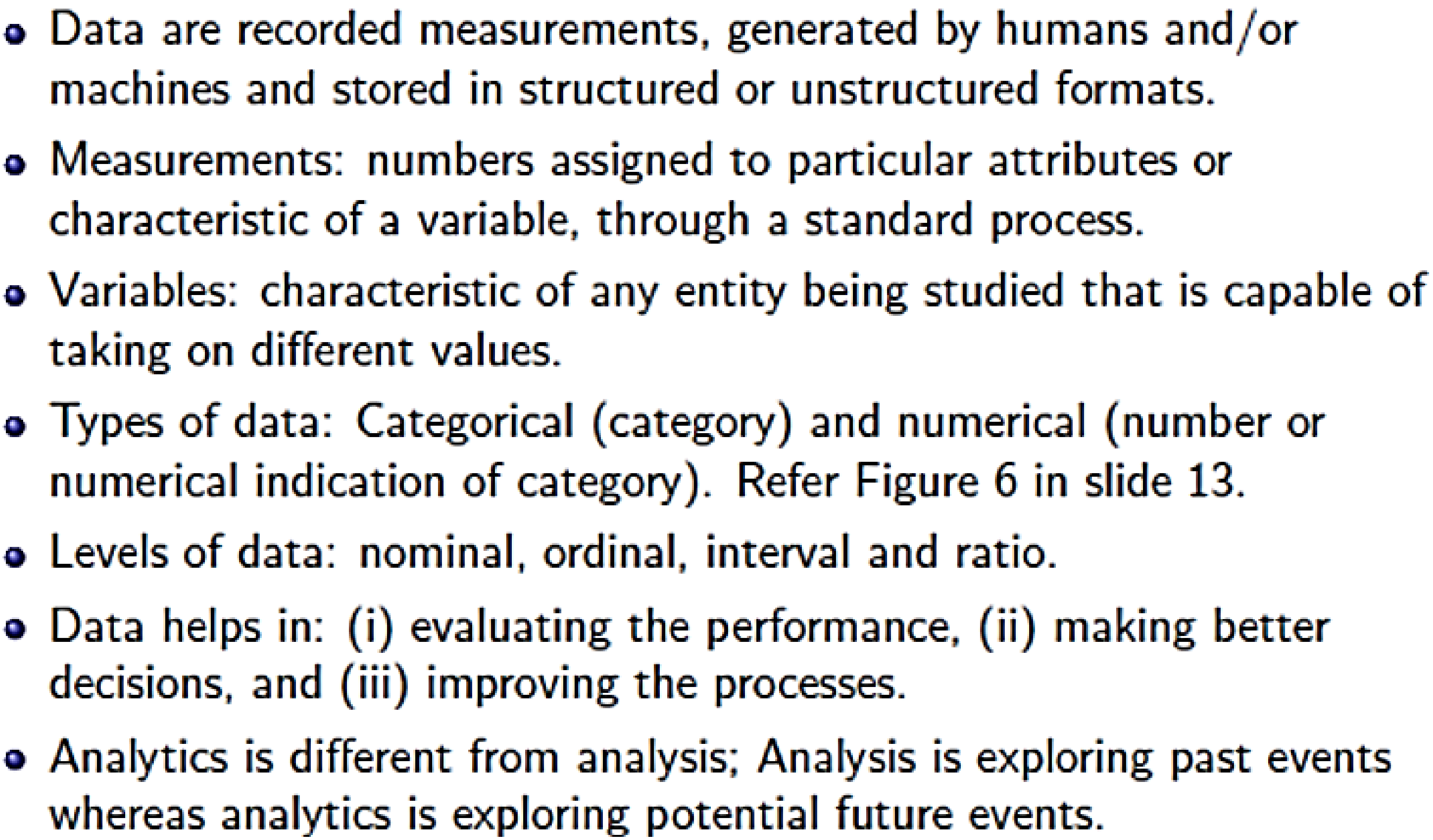
Transform log (Skewed Distribution)

Categorical one hot encoding

Scaling

ML Model

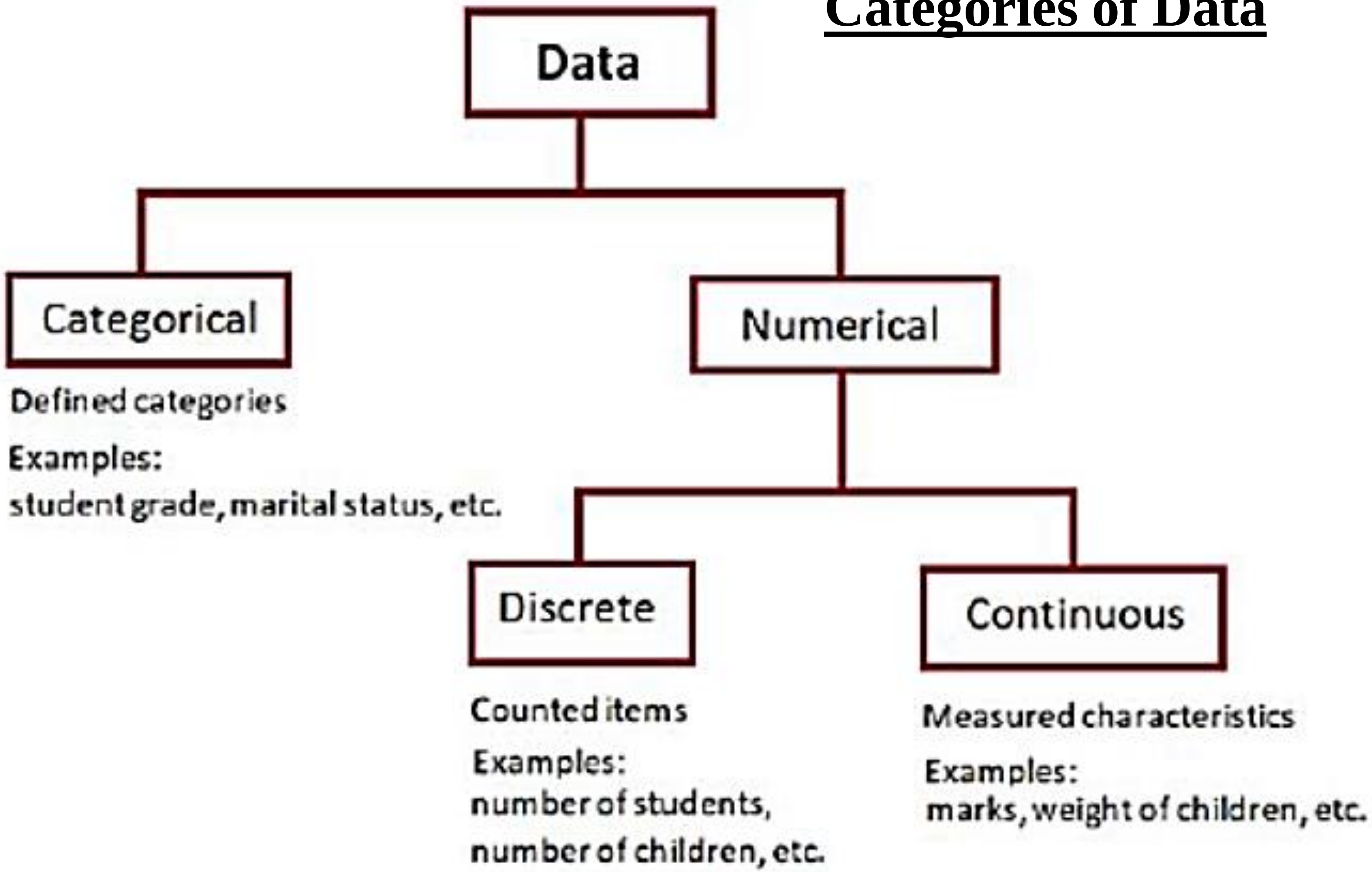
Data and related definitions



- Data are recorded measurements, generated by humans and/or machines and stored in structured or unstructured formats.
- Measurements: numbers assigned to particular attributes or characteristic of a variable, through a standard process.
- Variables: characteristic of any entity being studied that is capable of taking on different values.
- Types of data: Categorical (category) and numerical (number or numerical indication of category). Refer Figure 6 in slide 13.
- Levels of data: nominal, ordinal, interval and ratio.
- Data helps in: (i) evaluating the performance, (ii) making better decisions, and (iii) improving the processes.
- Analytics is different from analysis; Analysis is exploring past events whereas analytics is exploring potential future events.

Data and related definitions

Categories of Data



Data

Categorical

Defined categories

Examples:
student grade, marital status, etc.

Numerical

Discrete

Counted items

Examples:
number of students,
number of children, etc.

Continuous

Measured characteristics

Examples:
marks, weight of children, etc.

Data Cleaning and Data Integration

- **Data cleaning** is a technique that is applied to remove the noisy data and correct the inconsistencies in data[12].
- **Data integration** is a data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store[134]. This process can involve cleaning and transforming the data, as well as resolving any inconsistencies or conflicts that may exist between the different sources

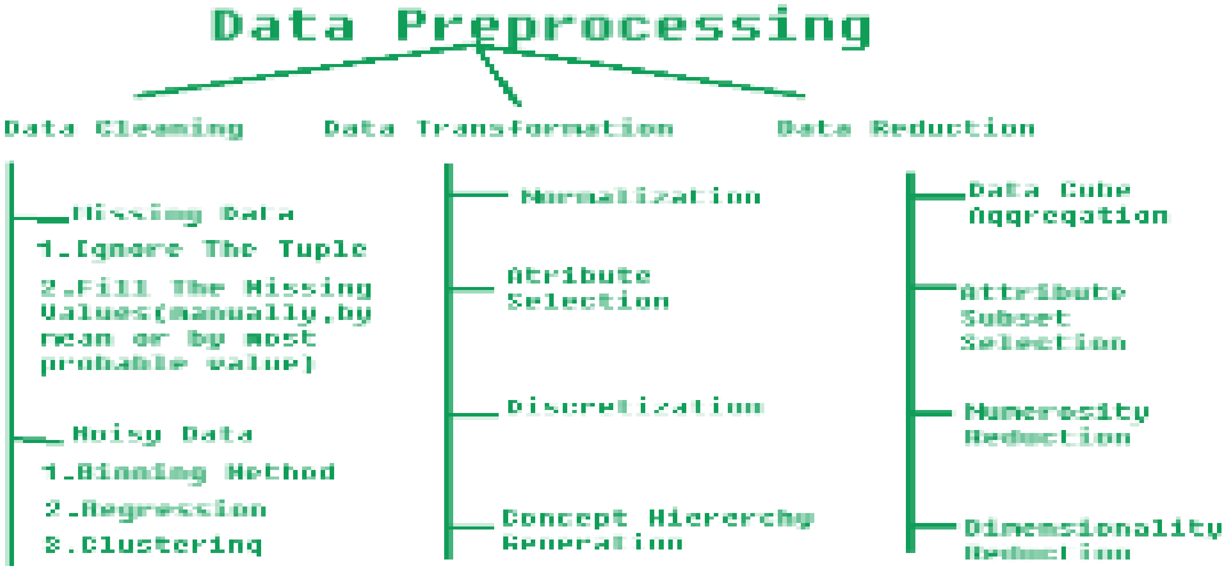
- Data cleaning is an essential step in the data mining process. It is crucial to the construction of a model. The step that is required, but frequently overlooked by everyone, is data cleaning. The major problem with quality information management is data quality. Problems with data quality can happen 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.

Data preprocessing

- Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis.
- The goal of data preprocessing is to improve the quality of the data .Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.
- **Some common steps in data preprocessing include:**
- Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:
- **Data Cleaning:** This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.
- **Data Integration:** This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

- **Data Transformation:** This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.
- **Data Reduction:** This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.
- **Data Discretization:** This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.
- **Data Normalization:** This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.
- Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.
- By performing these steps, the data mining process becomes more efficient and the results become more accurate.

- **Preprocessing in Data Mining:**
  Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



- **Steps Involved in Data Preprocessing:**
- **1. Data Cleaning:**
  The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

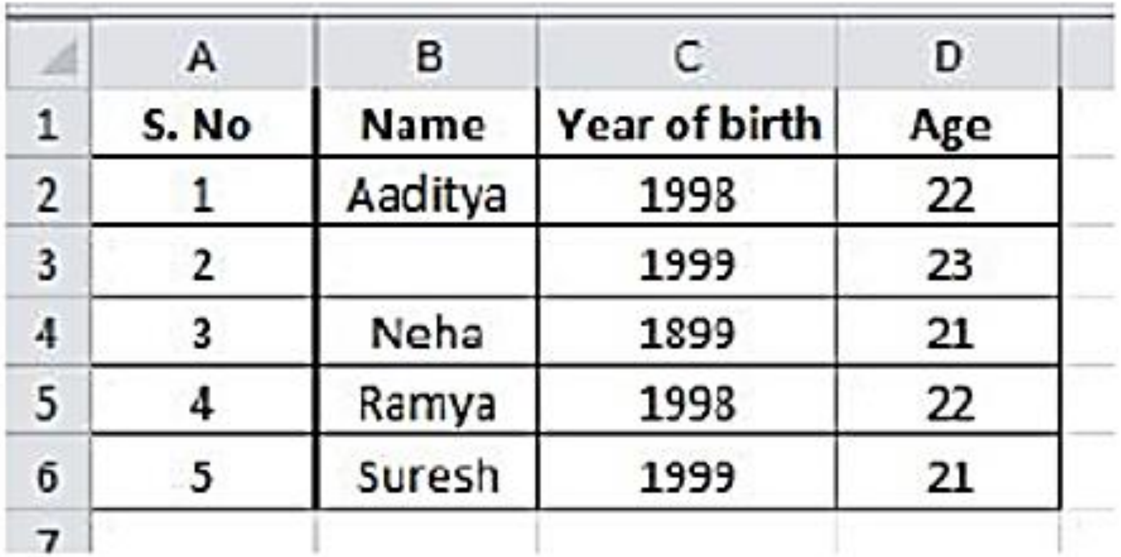
- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  - **Ignore the tuples:**
    This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

  - **Fill the Missing values:**
    There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:**
  Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  - **Binning Method:**
    This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  - **Regression:**
    Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  - **Clustering:**
    This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Data Cleaning

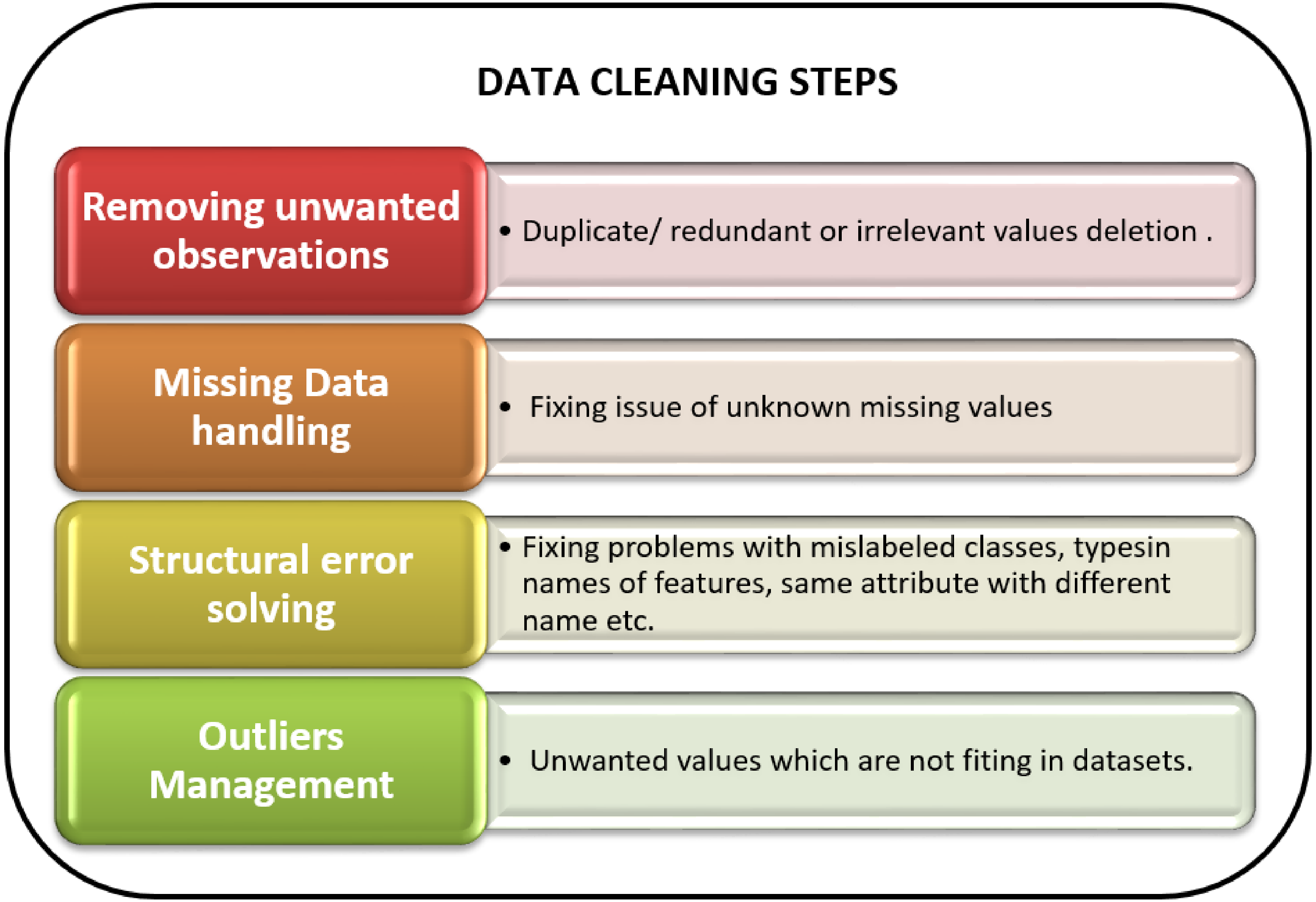
- *Data is the new oil.* It should be refined according to objective/goal.
- Real world data is considered *dirty* because it might be
  (i) incomplete, (ii) noisy, and/or (iii) inconsistent.
- Identify the reason(s) why the data shown in Figure 7 is not clean.



	A	B	C	D
1	S. No	Name	Year of birth	Age
2	1	Aaditya	1998	22
3	2		1999	23
4	3	Neha	1899	21
5	4	Ramya	1998	22
6	5	Suresh	1999	21
7				

Sample data - incomplete, erroneous and inconsistent.

Data Cleaning



DATA CLEANING STEPS

Removing unwanted observations

- Duplicate/ redundant or irrelevant values deletion .

Missing Data handling

- Fixing issue of unknown missing values

Structural error solving

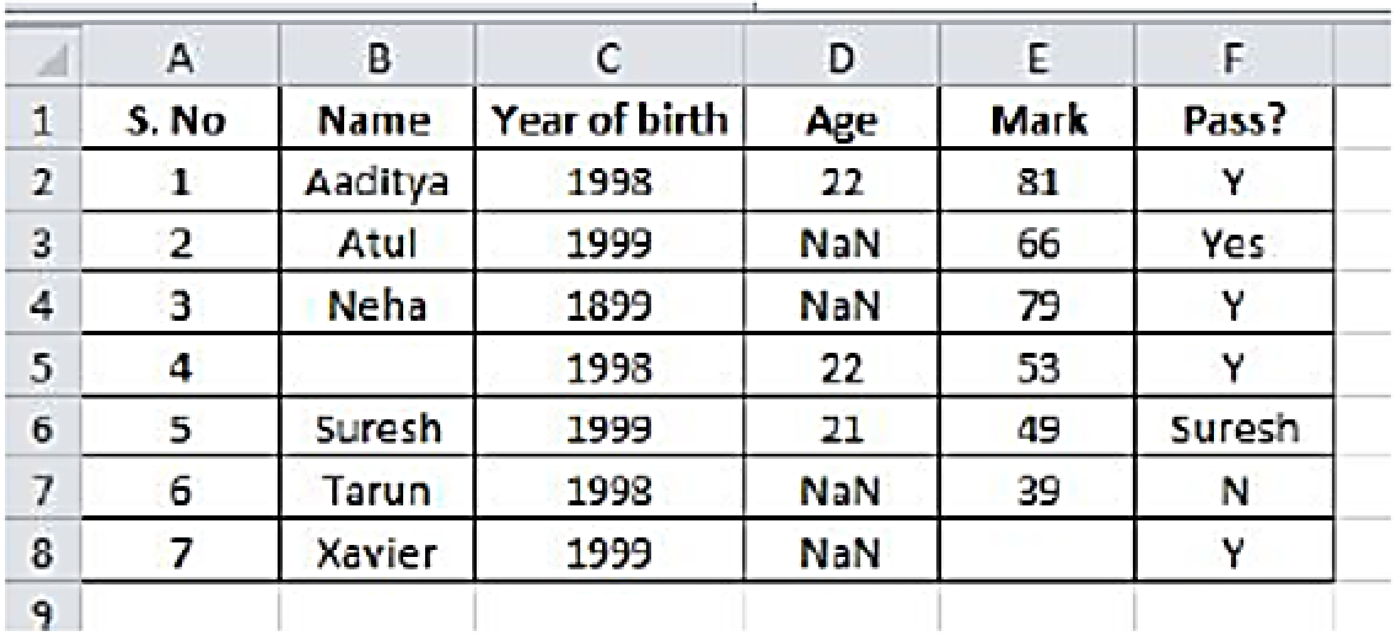
- Fixing problems with mislabeled classes, typesin names of features, same attribute with different name etc.

Outliers Management

- Unwanted values which are not fiting in datasets.

Data Cleaning

- Upon identifying the independent and dependent variables according to problem statement, unwanted columns can be easily deleted. Certain functions from pandas library will serve this purpose.
- Consider the data shown in Figure 9. Assume *df* is the variable assigned to this tabular data.



	A	B	C	D	E	F
1	S. No	Name	Year of birth	Age	Mark	Pass?
2	1	Aaditya	1998	22	81	Y
3	2	Atul	1999	NaN	66	Yes
4	3	Neha	1899	NaN	79	Y
5	4		1998	22	53	Y
6	5	Suresh	1999	21	49	Suresh
7	6	Tarun	1998	NaN	39	N
8	7	Xavier	1999	NaN		Y
9						

Sample data for data cleaning

Data Cleaning-Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

Data Cleaning-How to handle missing data?

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Data Cleaning-How to handle missing data?

- Ignore the tuple(**store multiple items in a single variable)**: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Data Cleaning-Noisy data?

- Random error in a measured variable.
    - Incorrect attribute values may be due to
    - faulty data collection instruments
    - data entry problems
    - data transmission problems
    - technology limitation
    - inconsistency in naming convention
- Other data problems which requires data cleaning
    - duplicate records
    - incomplete data
    - inconsistent data

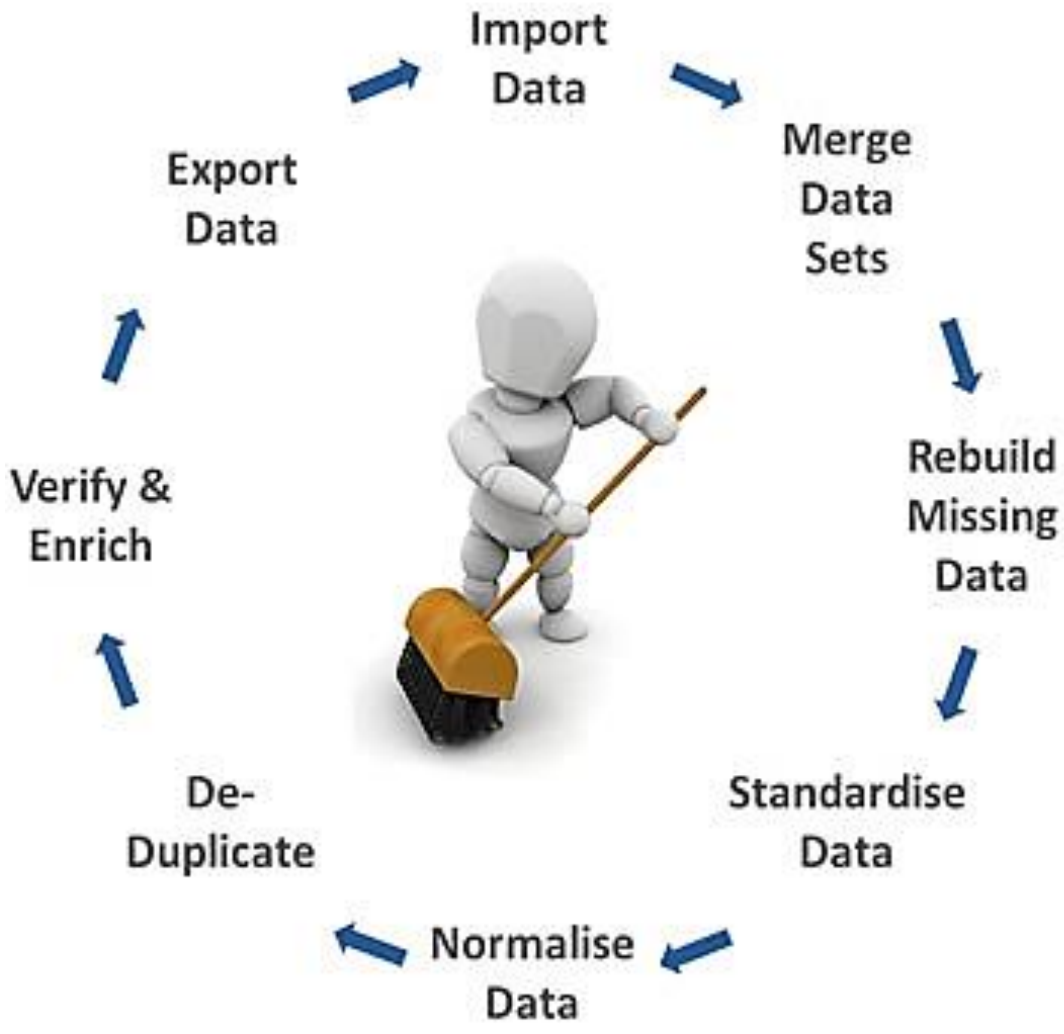
Data Cleaning-How to handle Noisy data?

- Binning method: (Bins- Smallest unit of space inside a **database)**
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - used also for discretization
- Clustering
  - detect and remove outliers
- Semi-automated method: combined computer and human inspection
  - detect suspicious values and check manually
- Regression
  - smooth by fitting the data into regression functions

Data Cleaning-How to handle Inconsistent data?

- Manual correction using external references
- Semi-automatic using various tools
    - To detect violation of known functional dependencies and data constraints
    - To correct redundant data

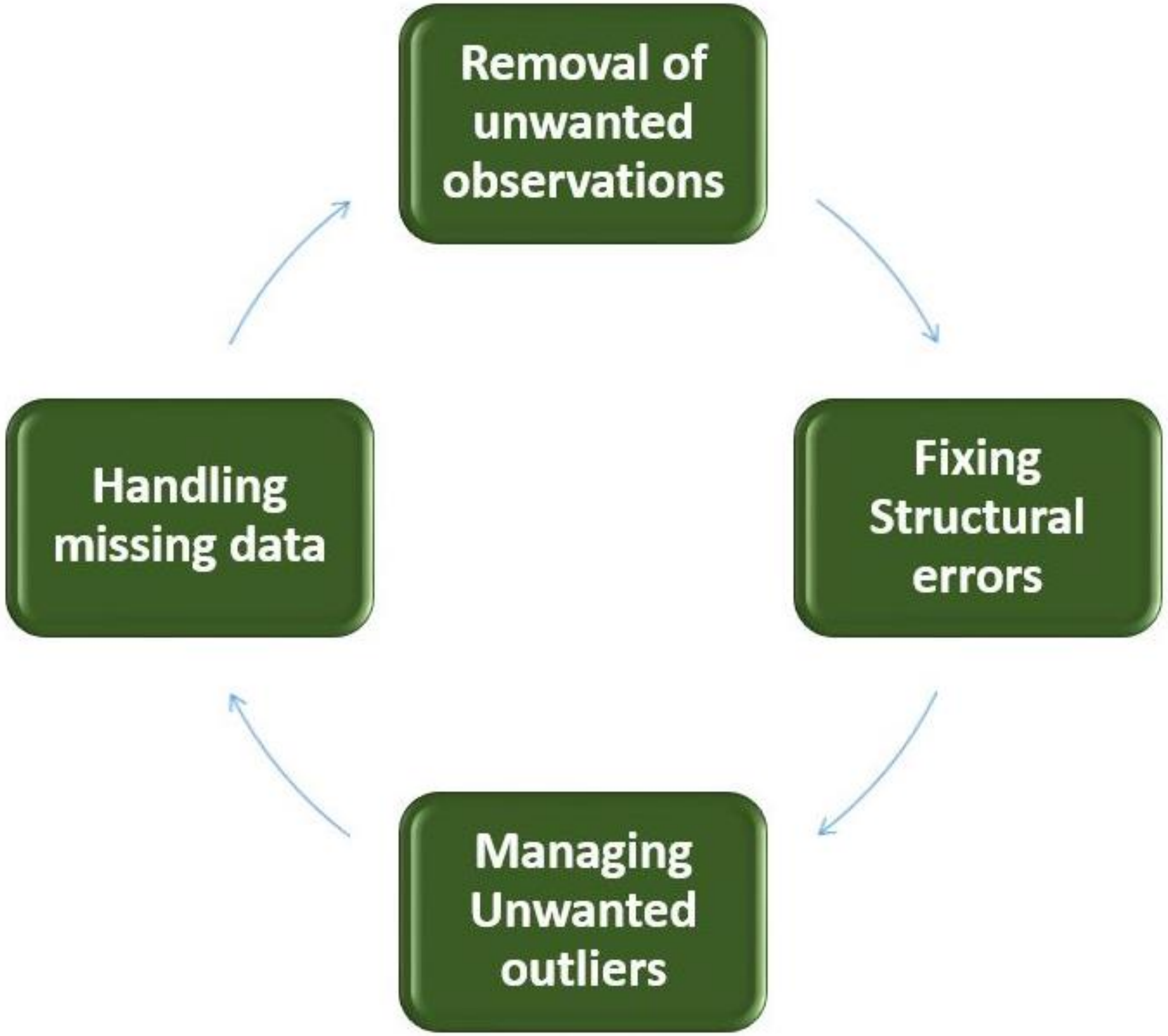
Data Cleaning



- Step 1: Remove duplicate or irrelevant observations. Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. ...
- Step 2: Fix structural errors. ...
- Step 3: Filter unwanted outliers. ...
- Step 4: Handle missing data. ...
- Step 5: Validate and QA.

Data Cleaning-Benefits

- Removal of errors when multiple sources of data.
- Ability to map the different functions and what your data is intended to do.
- Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
- Using tools for data cleaning will make for more efficient business practices and quicker decision-making.



Removal of unwanted observations

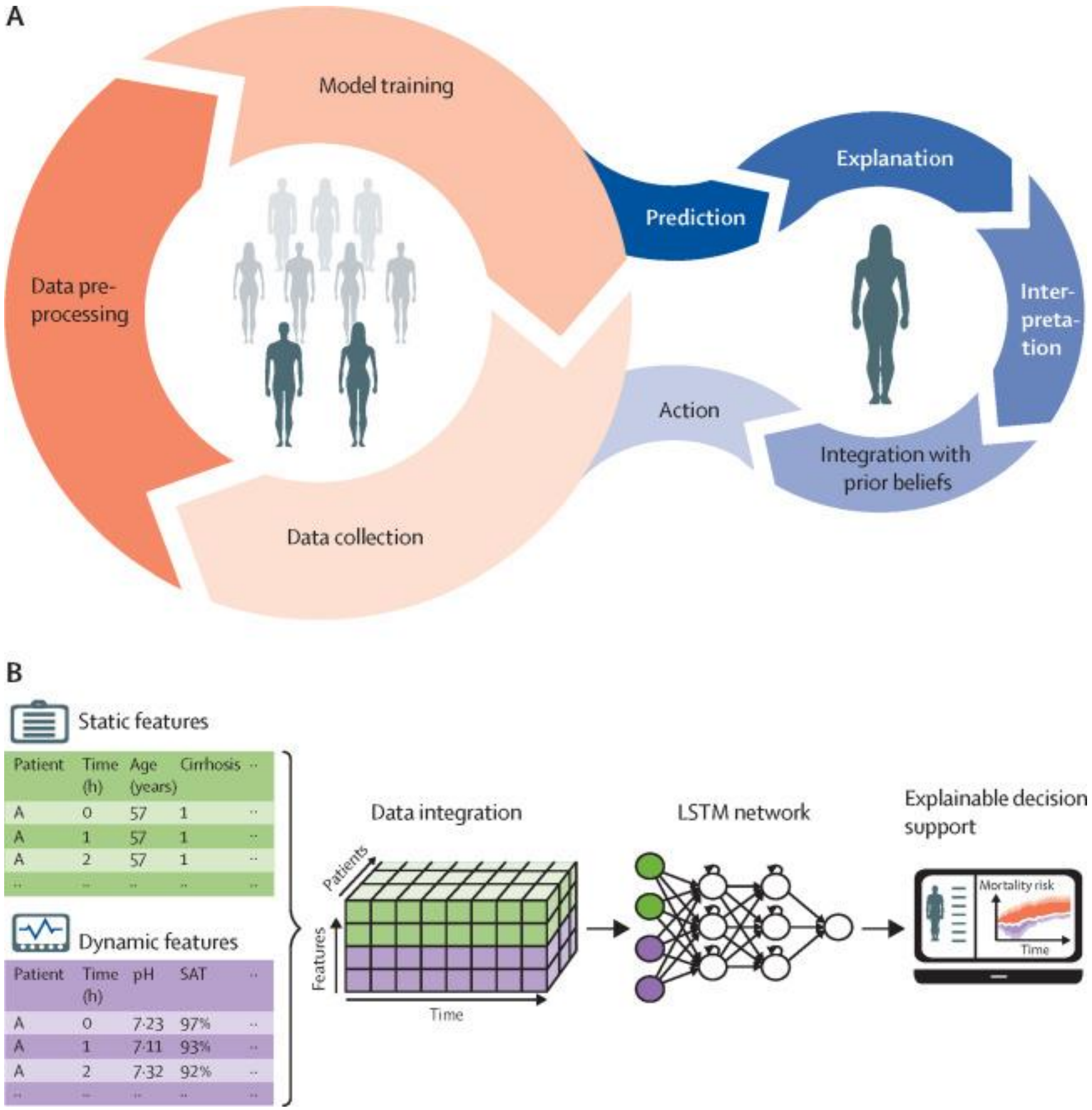
Fixing Structural errors

Managing Unwanted outliers

Handling missing data

Data Integration

- Data integration: combines data from multiple sources into a coherent store
- Schema integration:
  - Integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources

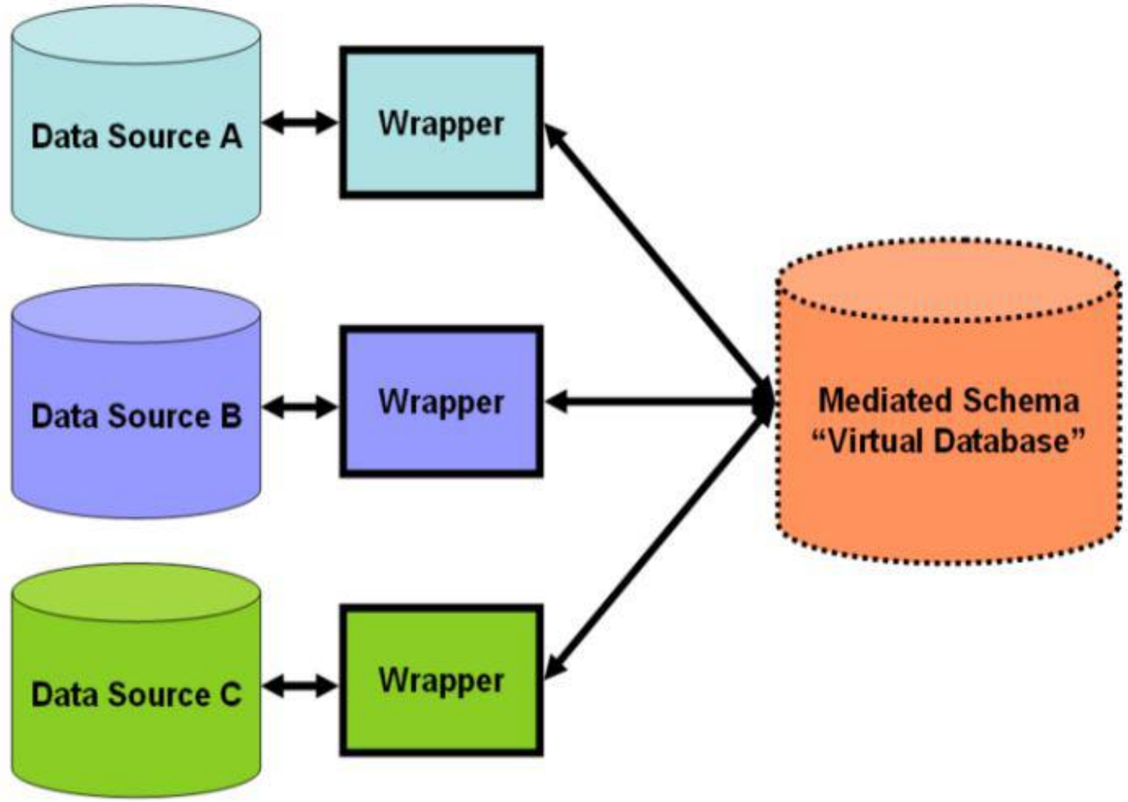


Patient	Time (h)	Age (years)	Cirrhosis	··
A	0	57	1	··
A	1	57	1	··
A	2	57	1	··
··	··	··	··	··

Patient	Time (h)	pH	SAT	··
A	0	7·23	97%	··
A	1	7·11	93%	··
A	2	7·32	92%	··
··	··	··	··	··

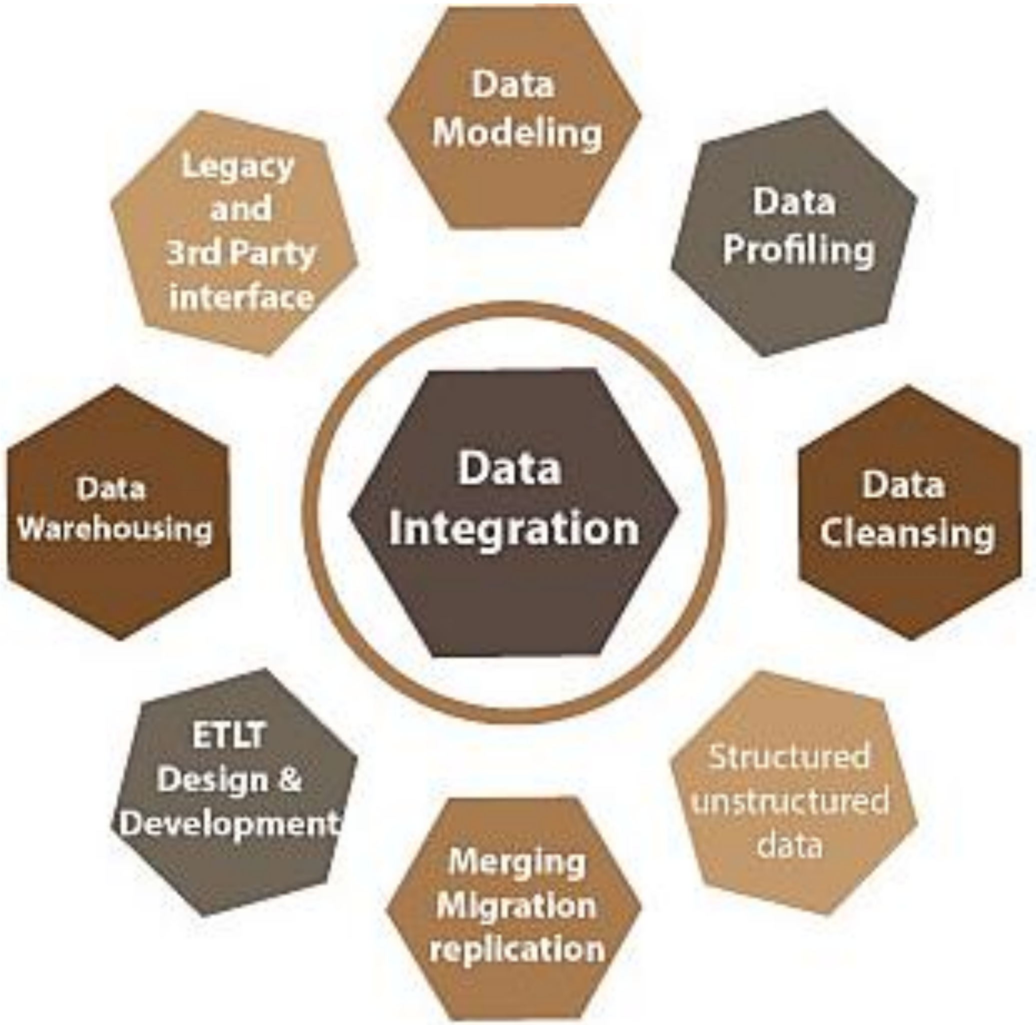
Data Integration

- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units, different currency

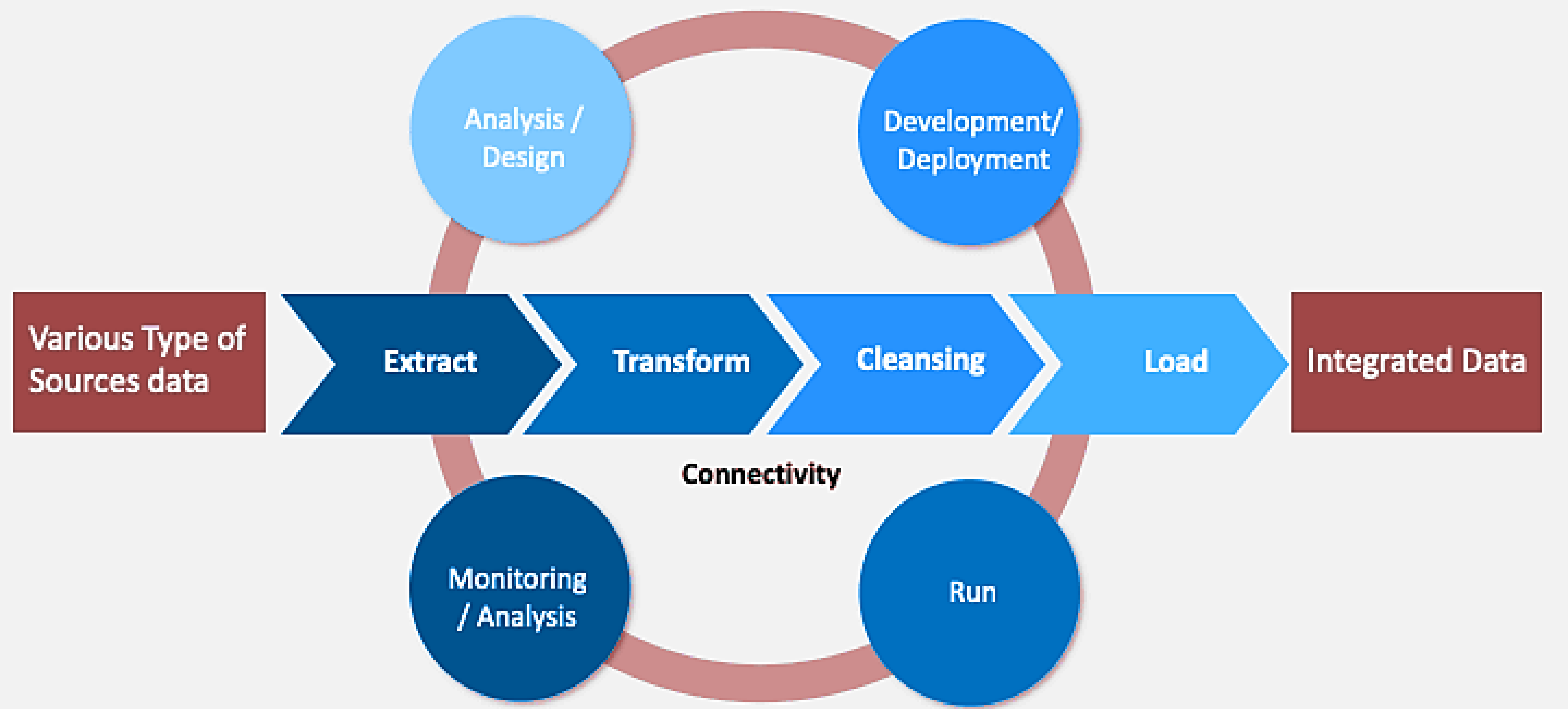


Why Data Integration is important?

- Gathering enormous volumes of data from various sources needs to be meaningful.
- Ease of accessible for analysis, when fresh data enters the database every second.
- Integrated data unlocks a layer of connectivity thereby improving the productivity.
- By connecting systems that contain valuable data and integrating them, ease of achieve data continuity and seamless knowledge transfer can be achieved.

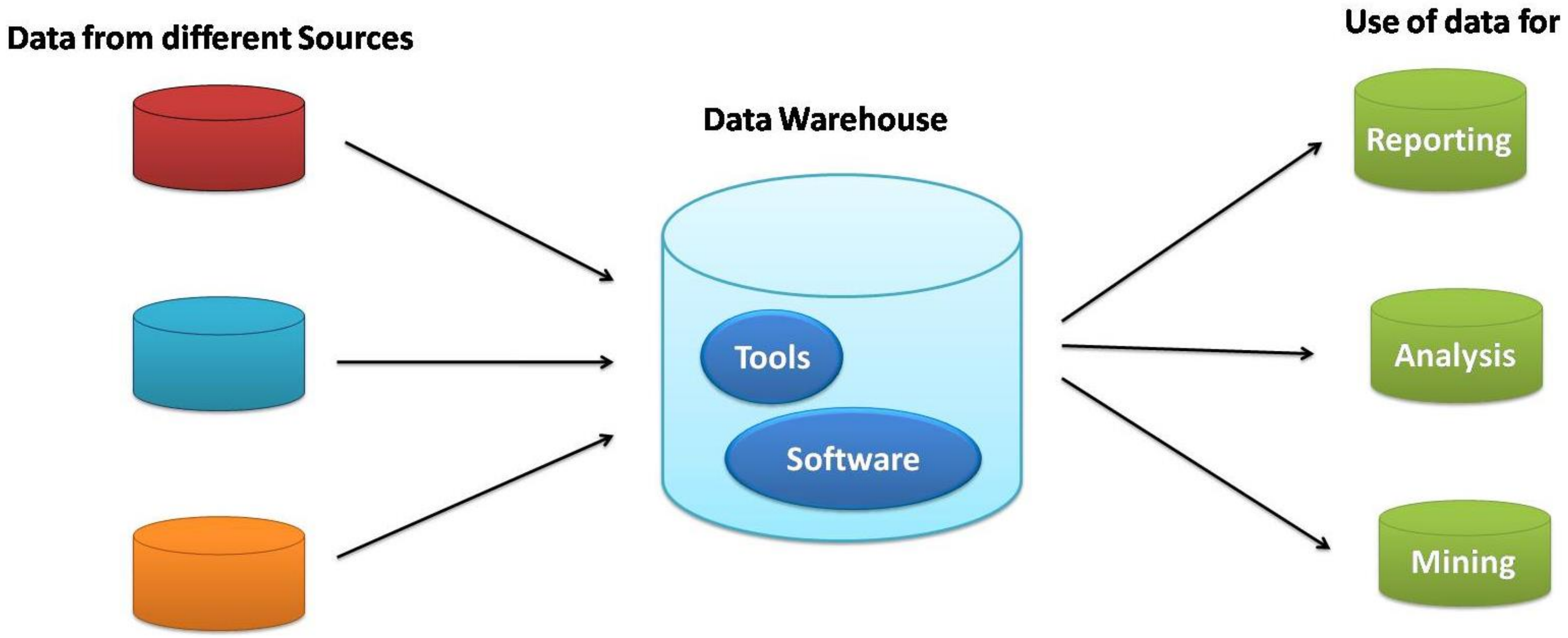

Data Integration
Data Modeling
Data Profiling
Data Cleansing
Structured unstructured data
Merging Migration replication
ETLT Design & Development
Data Warehousing
Legacy and 3rd Party interface

Data Integration



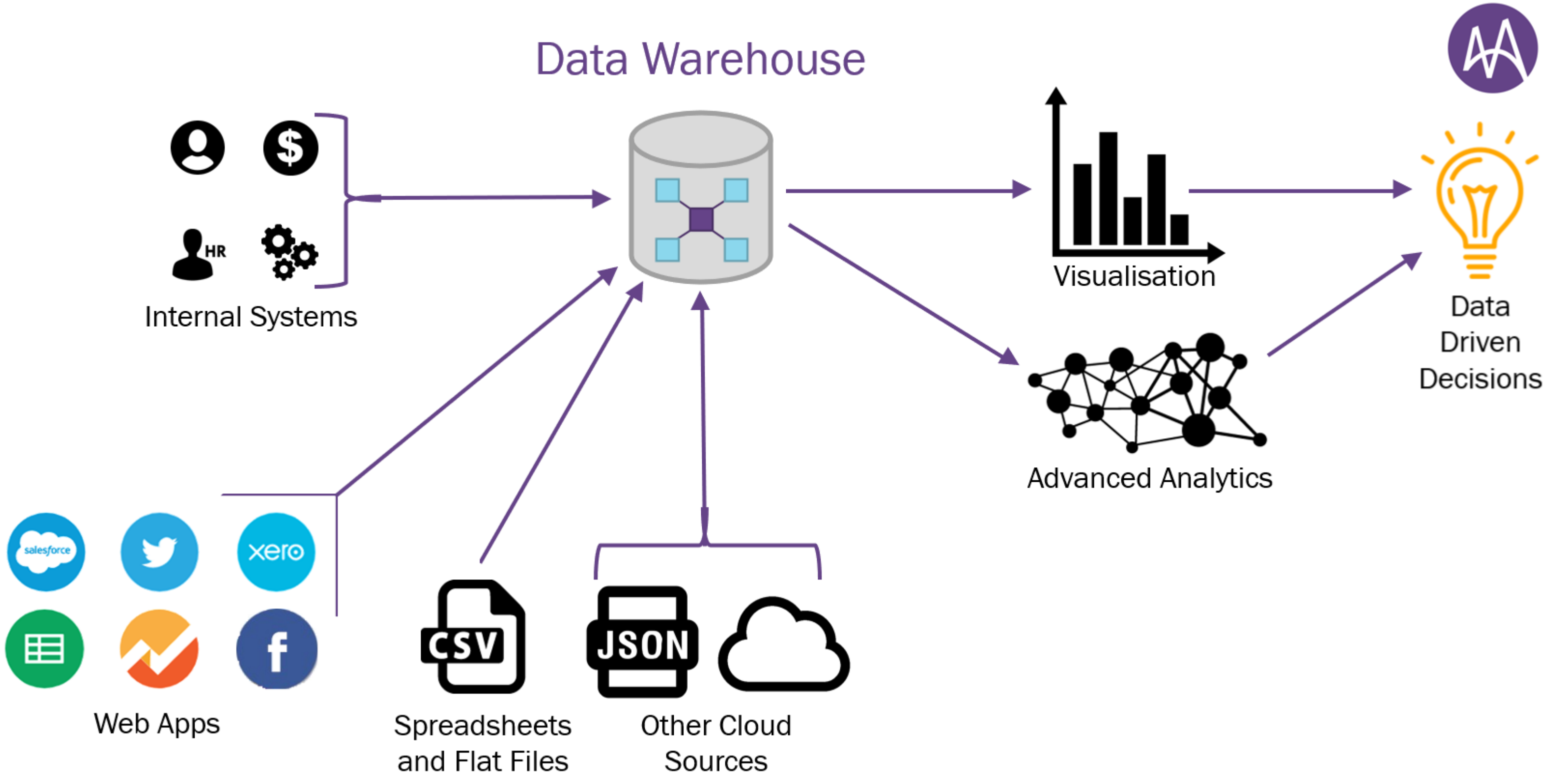
Ways to create Data Integration

- **Creating a data warehouse:** Data warehouses allow you to integrate different sources of data into a master relational database.
- When critical data is collected, stored and easily available, it’s much easier to assess micro and macro processes, manage operations and make strategic decisions based on this business intelligence.



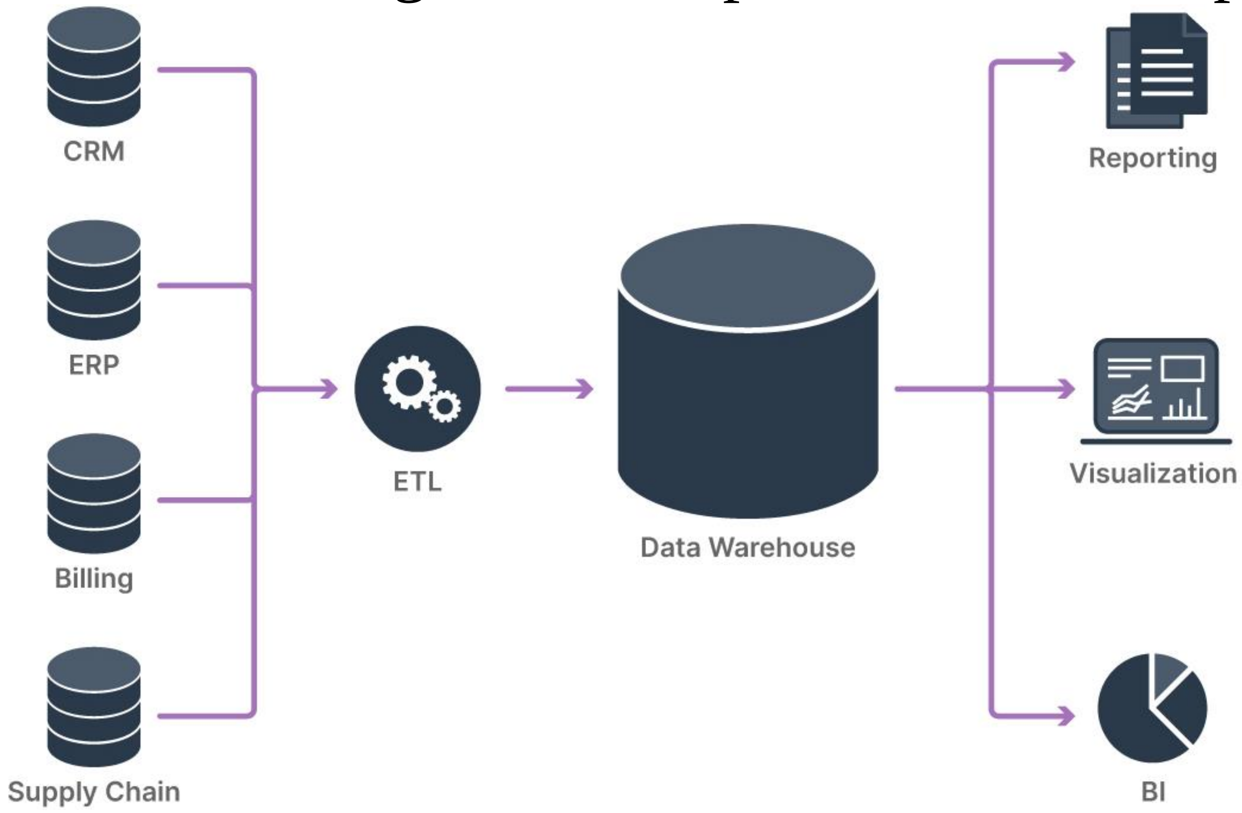
Ways to create Data Integration

- In this case, data integration works by providing a cohesive and centralized look at the entirety of an organization’s information, streamlining the process of gaining business intelligence insights.
- To achieve this, the managed service provider would a process called ETL.



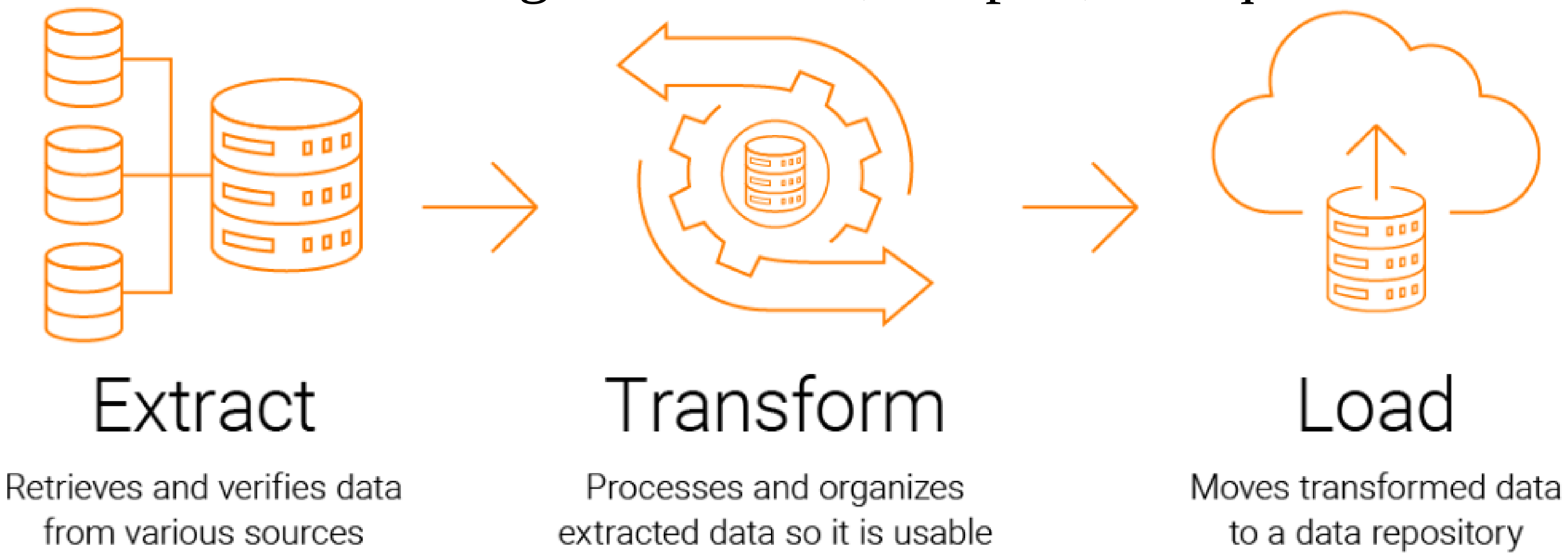
Ways to create Data Integration

- In this case, data integration works by providing a cohesive and centralized look at the entirety of an organization’s information, streamlining the process of gaining business intelligence insights.
- To achieve this, the managed service provider would a process called ETL.



Ways to create Data Integration

- **ETL (Extract, Transform, Load):** ETL is the process of sending data from source systems an organization possesses to the data warehouse where this information will be viewed and used.
- Most data integration systems involve one or more ETL pipelines, which make data integration easier, simpler, and quicker.



Extract

Retrieves and verifies data from various sources

Transform

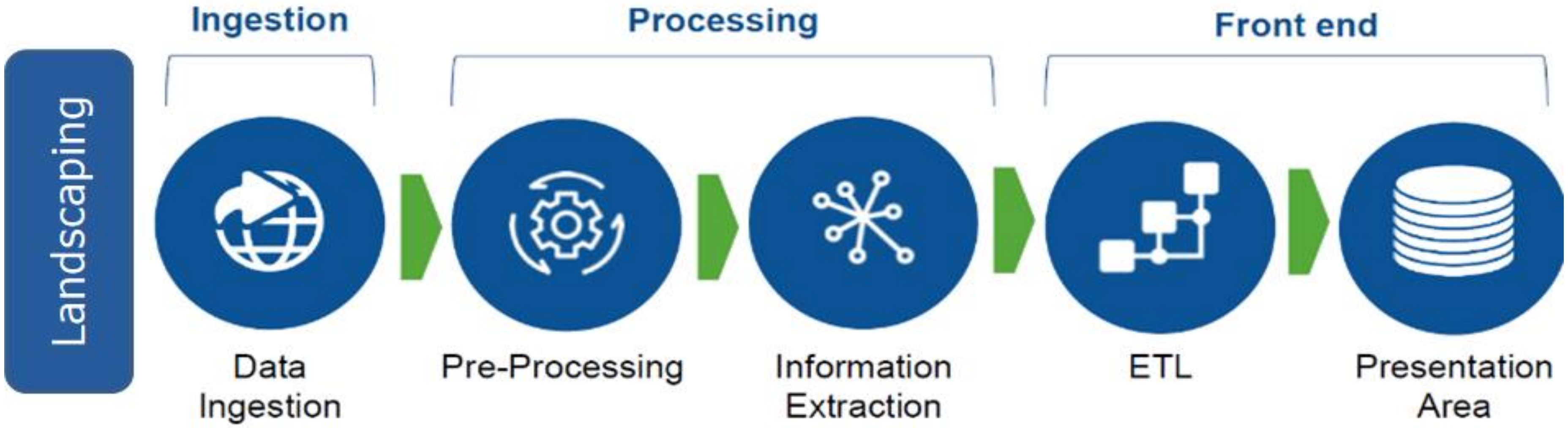
Processes and organizes extracted data so it is usable

Load

Moves transformed data to a data repository

Ways to create Data Integration

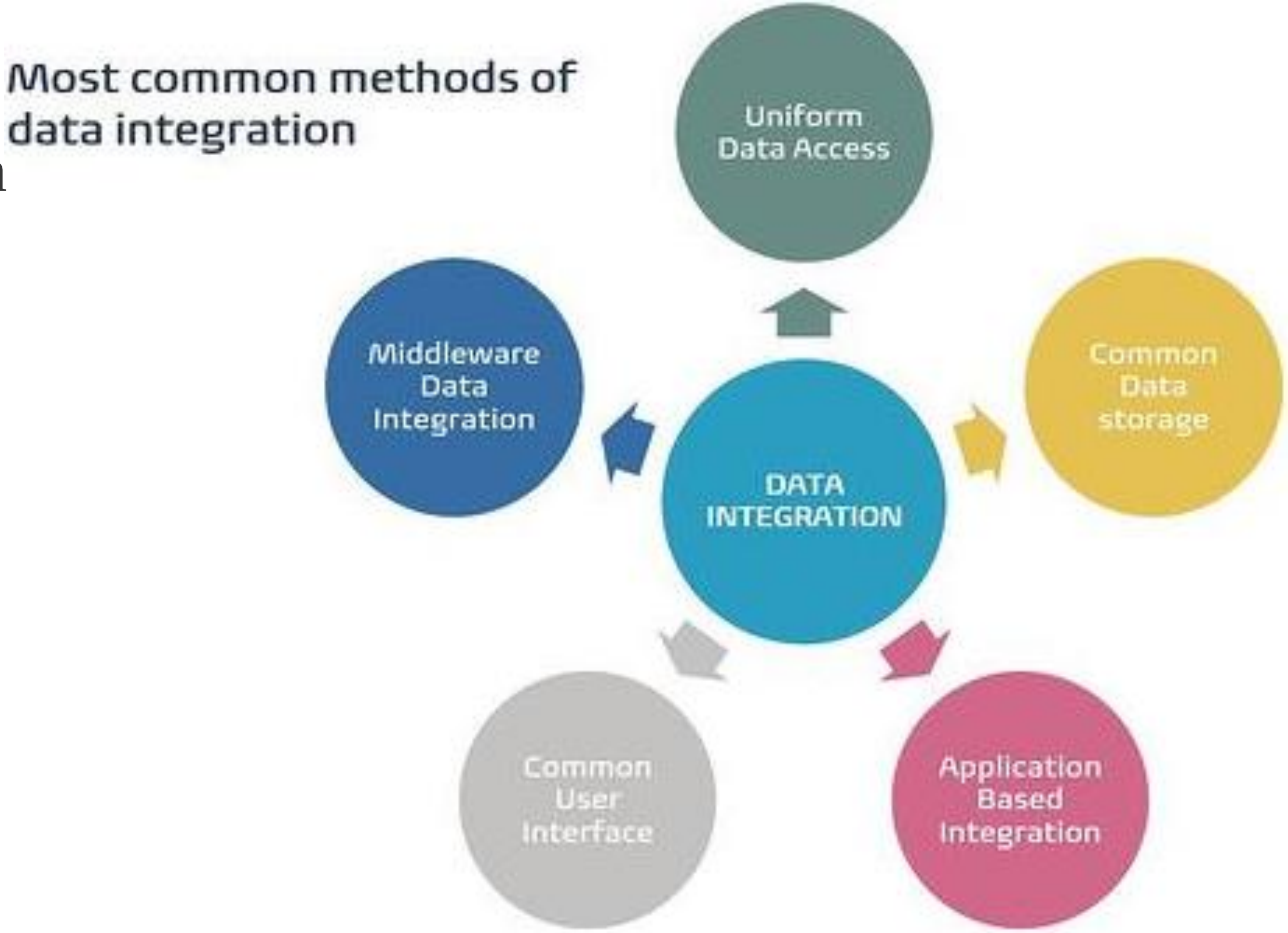
- **Building Data Pipelines:**
  - Avariety of built-in data connectors (for data ingestion), pre-defined transformations, and built-in job scheduler for automating the ETL pipeline.
  - Such tools make data integration easier, faster, and more cost effective by reducing the dependency on human expertise for manual operation.



Different types of Data Integration methods

Uniform Data Access:

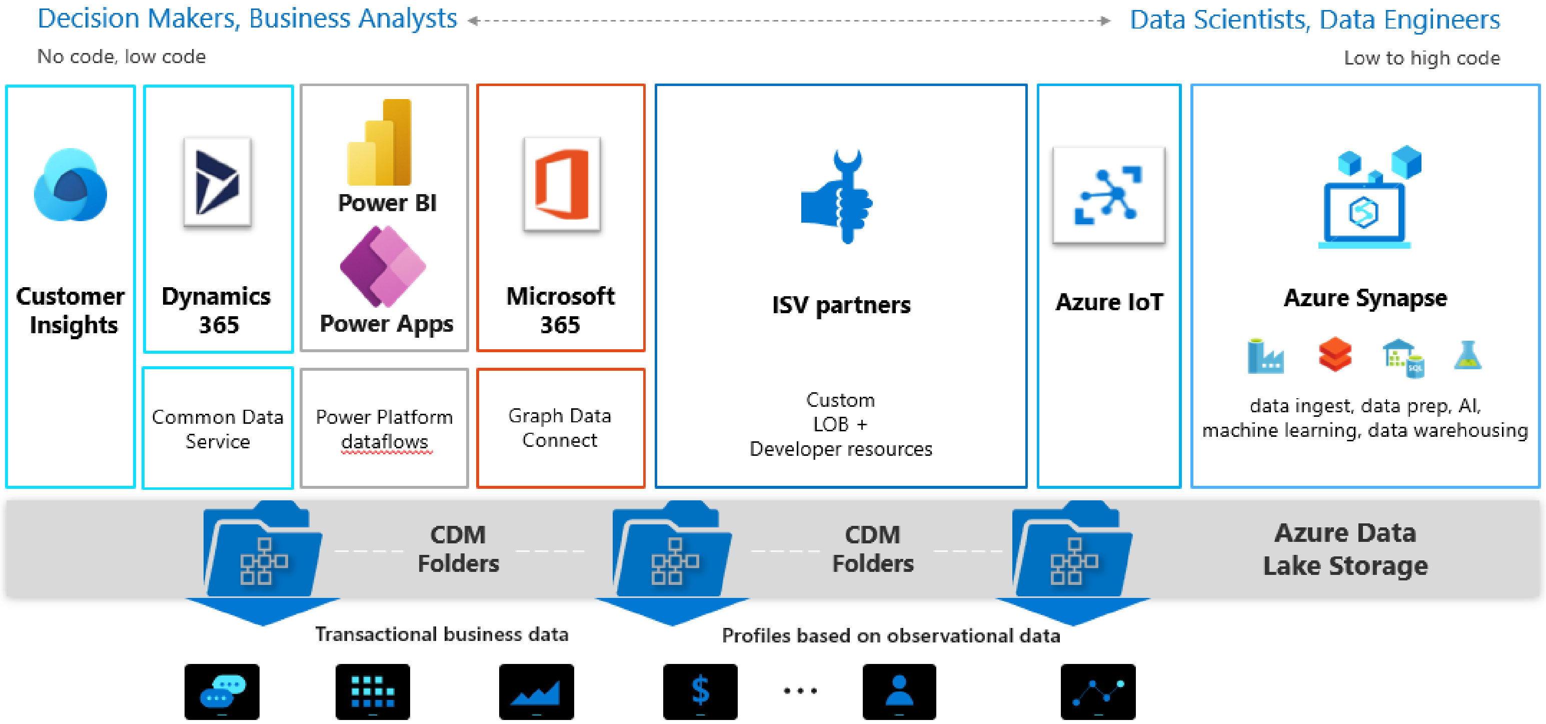
- With Uniform Data Access, enterprise data can be accessed from very disparate sets and present it uniformly.
- Uniform Data Access does this while allowing the data to stay in its original location.
- It leaves the data in the source system and defines a set that can provide a unified view to various customers across a platform.
- There is zero latency from the source system to the consolidated view.


Most common methods of data integration
Uniform Data Access
Middleware Data Integration
Common Data storage
DATA INTEGRATION
Common User Interface
Application Based Integration

Different types of Data Integration methods

Common Data storage:

- Common Data Storage (or CDS or Data Warehouse) is a storage space that enables you to manage and securely store data used by multiple applications or programs.
- Helps collecting data from various sources, combining them to a central space and management (Database files, mainframes, and flat files).



Different types of Data Integration methods

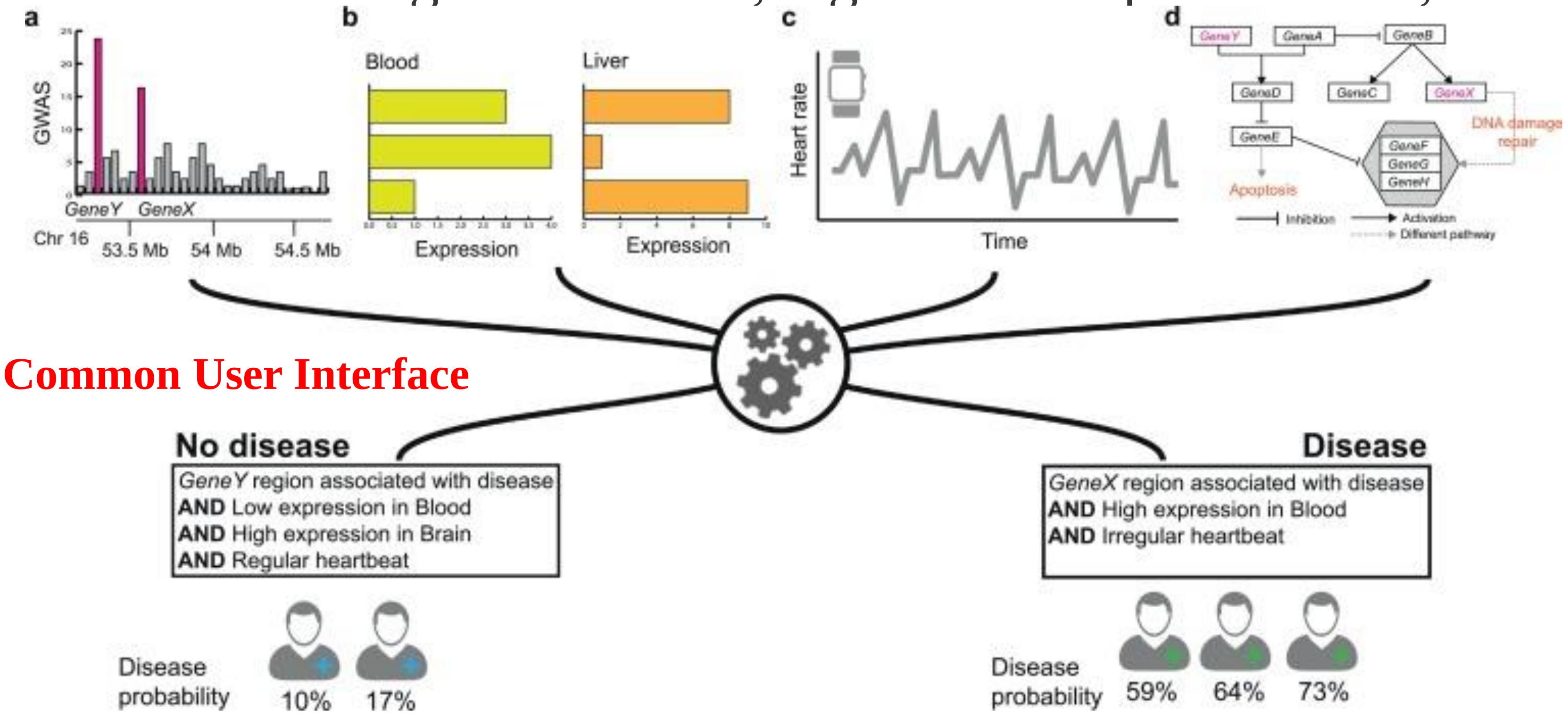
Application based Integration:

- Application Based Integration solutions are specialized programs that locate, retrieve and integrate your data.
- Application Based Integration accesses various data sources and returns integrated results to the user.
- It has limitations if you are handling large volumes of data and large numbers of sources because it requires the applications to implement all the integration efforts.

Different types of Data Integration methods

Common User Interface: Common User Interface means manually locating the information in each data source and comparing or cross-referencing them yourself in order to get the insight you need.

- The users must deal with different user interfaces and query languages and therefore need to have detailed knowledge on location, logical data representation, and data semantics.



An example for Common User Interface

Different types of Data Integration methods

Middle ware Data integration:

- Middleware is a layer of software that creates a common platform for all interactions internal and external to the organization—system-to-system, system-to-database, human-to-system, web-based, and mobile-device-based interactions.
- Middleware integration refers to applications that connect two or more applications.

Handling redundant data in Data Integration

- Redundant data occur often when integrating multiple DBs
- The same attribute may have different names in different databases
- One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis

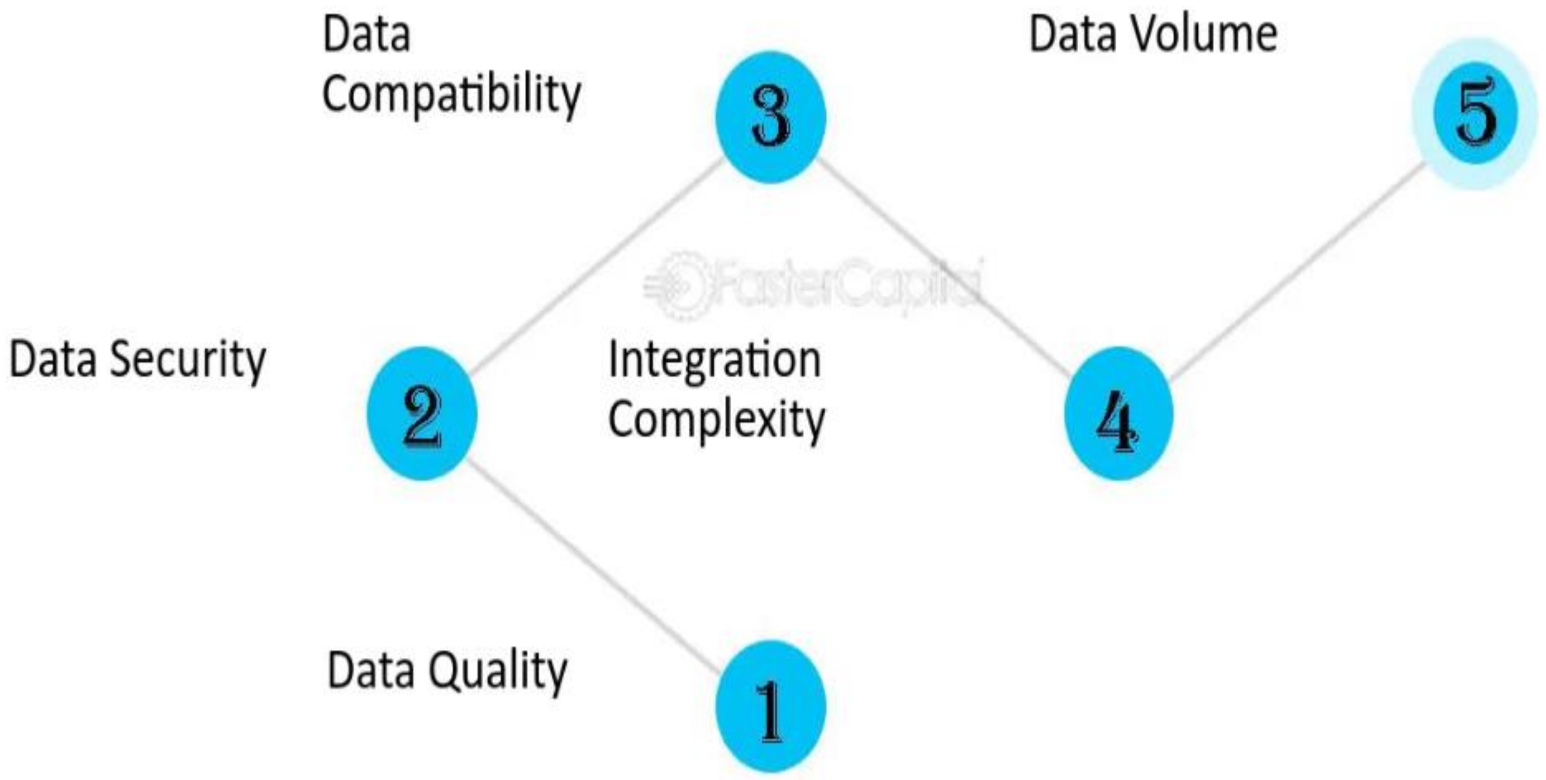
$$r_{A,B} = \frac{\Sigma(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B}$$

- Careful integration can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Advantages of Data Integration

- Elimination of errors
- Saves time
- Better sense of all the available information.
- Streamline the processes and improve the efficacy of data usage.
- Inter-system cooperation
- Seamless knowledge transfer between systems.
- Data integrity and data quality.

Five challenges and its solution of Data Integration



Solution to challenges:

- Clean up your data
- Introduce clear processes for data management
- Back up your data
- Choose the right software to assist you with data integration
- Manage and maintain your data

Data Transformation

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

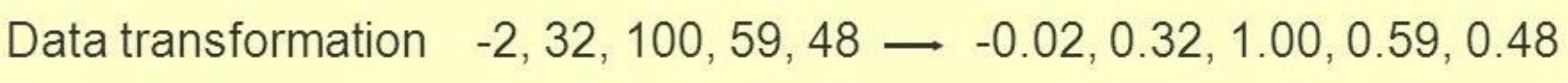
1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

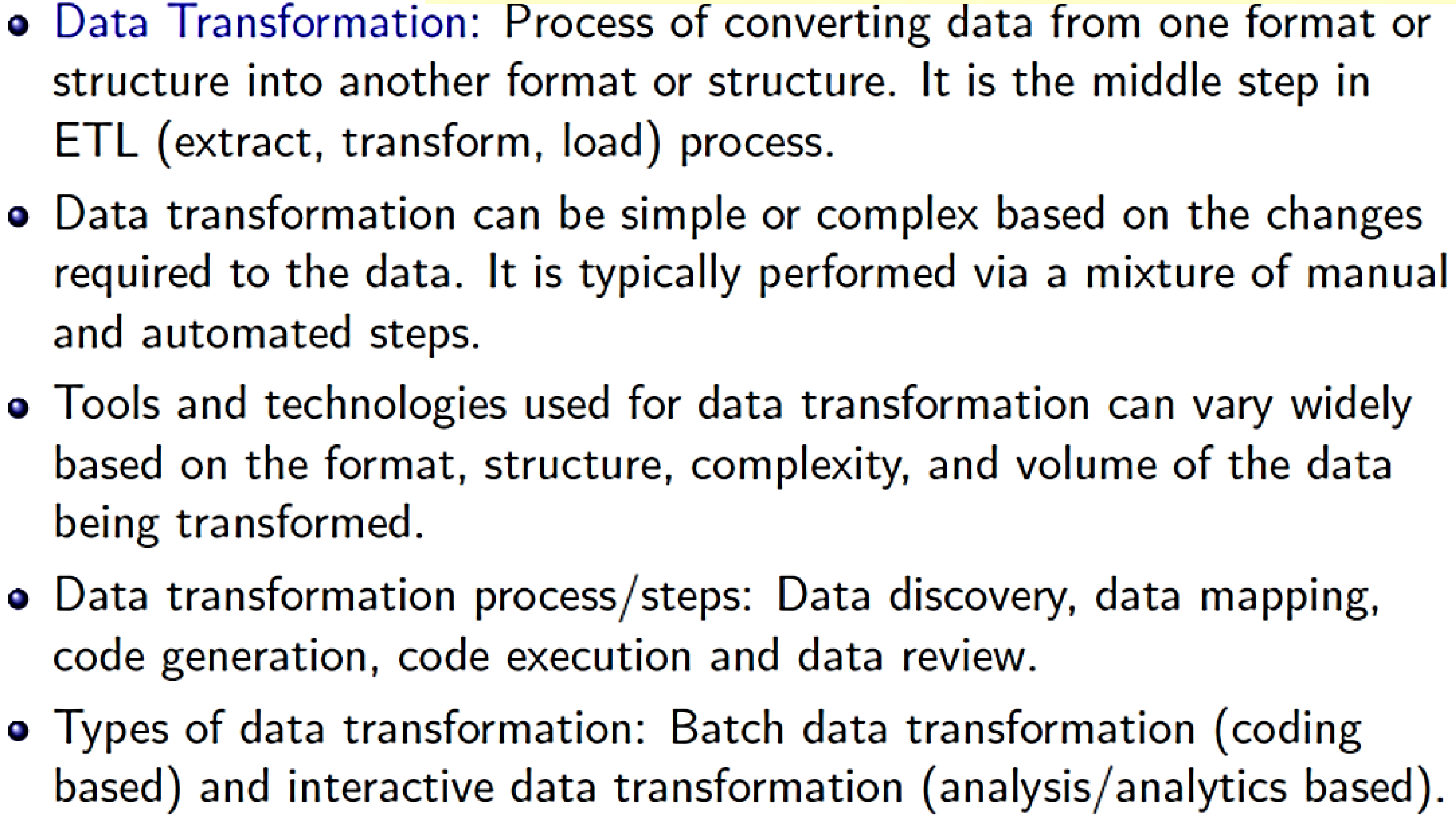
3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Data Transformation



Data transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48



- Data Transformation: Process of converting data from one format or structure into another format or structure. It is the middle step in ETL (extract, transform, load) process.
- Data transformation can be simple or complex based on the changes required to the data. It is typically performed via a mixture of manual and automated steps.
- Tools and technologies used for data transformation can vary widely based on the format, structure, complexity, and volume of the data being transformed.
- Data transformation process/steps: Data discovery, data mapping, code generation, code execution and data review.
- Types of data transformation: Batch data transformation (coding based) and interactive data transformation (analysis/analytics based).

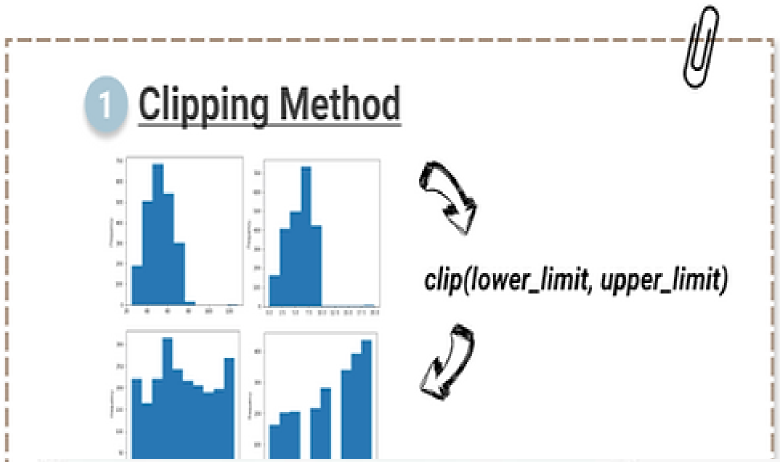
Data Transformation

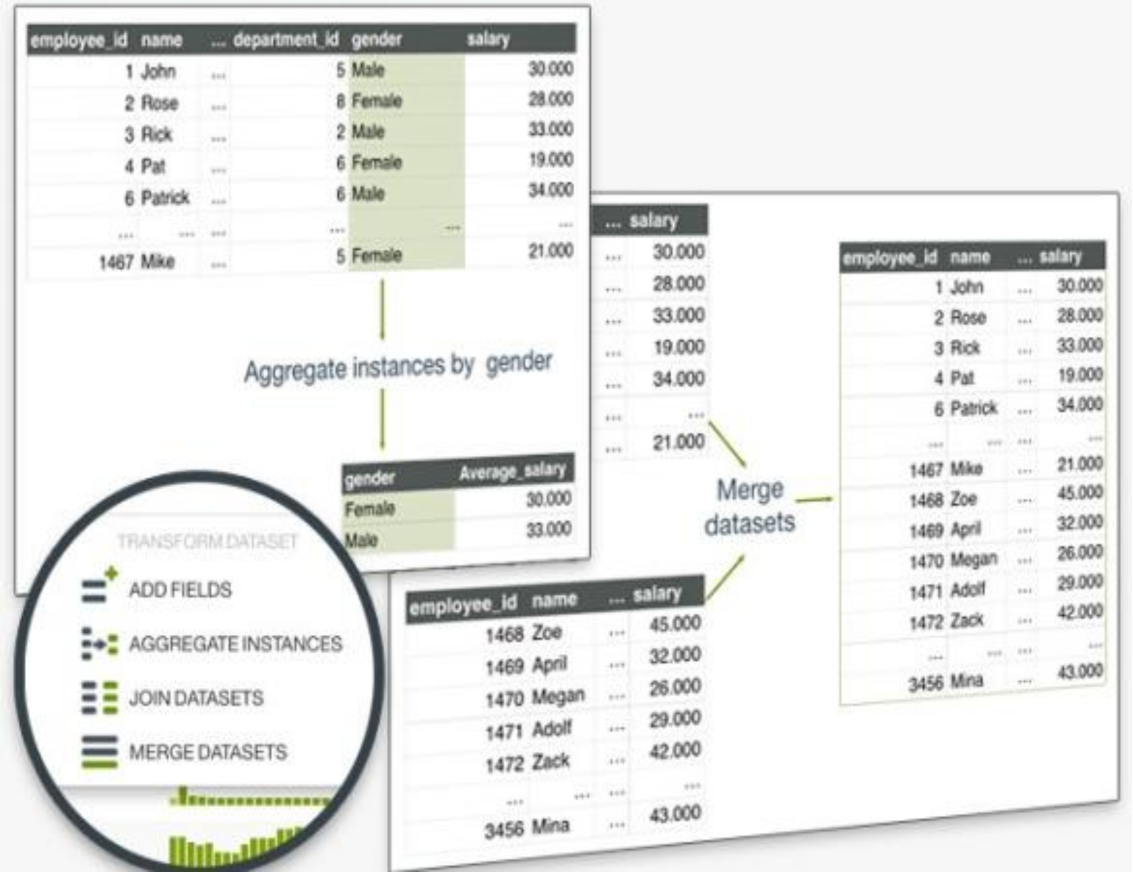
- Data transformation in data mining refers to the process of converting raw data into a format that is suitable for analysis and modeling. The goal of data transformation is to prepare the data for data mining so that it can be used to extract useful insights and make it suitable for an application. Data transformation typically involves several steps, including:

1. **Data cleaning:** Removing or correcting errors, inconsistencies, and missing values in the data.
2. **Data integration:** Combining data from multiple sources, such as databases and spreadsheets, into a single format.
3. **Data normalization:** Scaling the data to a common range of values, such as between 0 and 1, to facilitate comparison and analysis.
4. **Data reduction:** Reducing the dimensionality of the data by selecting a subset of relevant features or attributes.
5. **Data discretization**: Converting continuous data into discrete categories or bins.
6. **Data aggregation:** Combining data at different levels of granularity, such as by summing or averaging, to create new features or attributes.
7. **Data transformation** is an important step in the data mining process as it helps to ensure that the data is in a format that is suitable for analysis and modeling, and that it is free of errors and inconsistencies. Data transformation can also help to improve the performance of data mining algorithms, by reducing the dimensionality of the data, and by scaling the data to a common range of values.

Data Transformation

- **Smoothing:** 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.





employee_id	name	...	department_id	gender	salary
1	John	...	5	Male	30.000
2	Rose	...	8	Female	28.000
3	Rick	...	2	Male	33.000
4	Pat	...	6	Female	19.000
6	Patrick	...	6	Male	34.000
...	...	...	...	...	...
1467	Mike	...	5	Female	21.000

Aggregate instances by gender

gender	Average_salary
Female	30.000
Male	33.000

...	salary
...	30.000
...	28.000
...	33.000
...	19.000
...	34.000
...	...
...	21.000

employee_id	name	...	salary
1468	Zoe	...	45.000
1469	April	...	32.000
1470	Megan	...	26.000
1471	Adolf	...	29.000
1472	Zack	...	42.000
...	...	...	...
3456	Mina	...	43.000

Merge datasets

employee_id	name	...	salary
1	John	...	30.000
2	Rose	...	28.000
3	Rick	...	33.000
4	Pat	...	19.000
6	Patrick	...	34.000
...	...	...	...
1467	Mike	...	21.000
1468	Zoe	...	45.000
1469	April	...	32.000
1470	Megan	...	26.000
1471	Adolf	...	29.000
1472	Zack	...	42.000
...	...	...	...
3456	Mina	...	43.000

- **Aggregation:** Data collection or aggregation is the method of storing and presenting data in a summary format.
- The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.
- Combine and summarize data to create a compact representation.
- This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used
- Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.
- The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies. For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

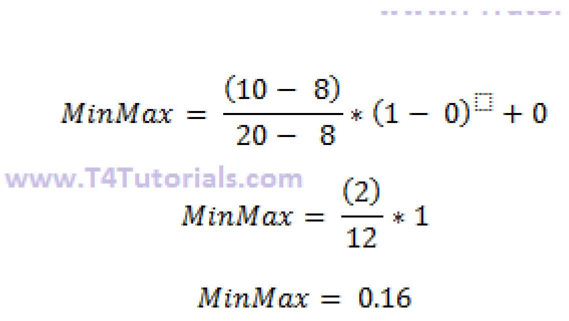
- **Discretization:** It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes. Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values. For **example**, (1-10, 11-20) (age:- young, middle age, senior). ibutes. Yet many of the existing data mining frameworks are unable to handle these attributes. Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values. For **example**, (1-10, 11-20) (age:- young, middle age, senior).
- **Attribute Construction:** Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.
- **Generalization:** It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old). For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country. .

- **Normalization:** Data normalization involves converting all data variables into a given range. It scales numerical values to fit within a specific range. Normalization is generally required when we are dealing with attributes on a different scale.Techniques that are used for normalization are:
- **Min-Max Normalization:**
  - This transforms the original data linearly.
  - It will scale the data between 0 and 1.
  - Suppose that: min_A is the minima and max_A is the maxima of an attribute, P
  - Where v is the value you want to plot in the new range.
  - v’ is the new value you get after normalizing the old value.

marks
8
10
15
20

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

For Marks as 10:



$$MinMax = \frac{(10 - 8)}{20 - 8} * (1 - 0)^{\square} + 0$$

$$MinMax = \frac{(2)}{12} * 1$$

$$MinMax = 0.16$$

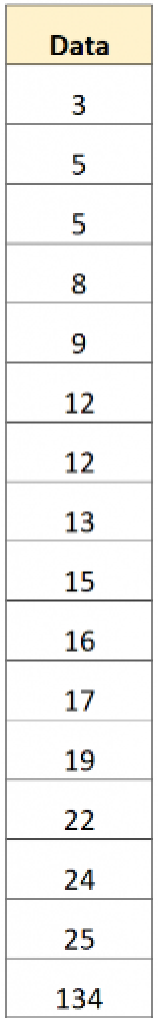
marks	marks after Min-Max normalization
8	0
10	0.16
15	0.58
20	1

- **Z-Score Normalization:**
  - In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation
  - A value, v, of attribute A is normalized to v' by computing
  - **Z-score normalization** refers to the process of normalizing every value in a dataset such that the mean of all of the values is 0 and the standard deviation is 1.

New value = (x – μ) / σ

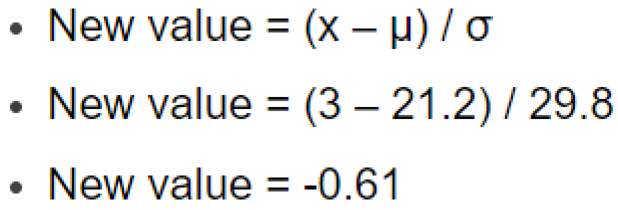
where:

- **x**: Original value
- **μ**: Mean of data
- **σ**: Standard deviation of data

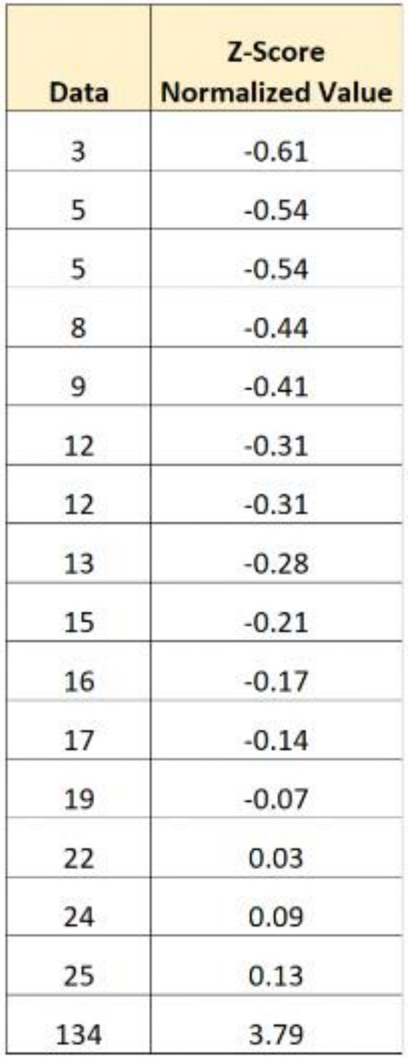


Data
3
5
5
8
9
12
12
13
15
16
17
19
22
24
25
134

- Using a calculator, we can find that the mean of the dataset is **21.2** and the standard deviation is **29.8**.



  - New value = (x – μ) / σ
  - New value = (3 – 21.2) / 29.8
  - New value = -0.61



Data	Z-Score Normalized Value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	-0.41
12	-0.31
12	-0.31
13	-0.28
15	-0.21
16	-0.17
17	-0.14
19	-0.07
22	0.03
24	0.09
25	0.13
134	3.79

Decimal Scaling:

- It normalizes the values of an attribute by changing the position of their decimal points.
- The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
  A value, v, of attribute A is normalized to v' by computing
  where j is the smallest integer such that Max(|v'|) < 1.
  Suppose: Values of an attribute P varies from -99 to 99.
  The maximum absolute value of P is 99.
  For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97.

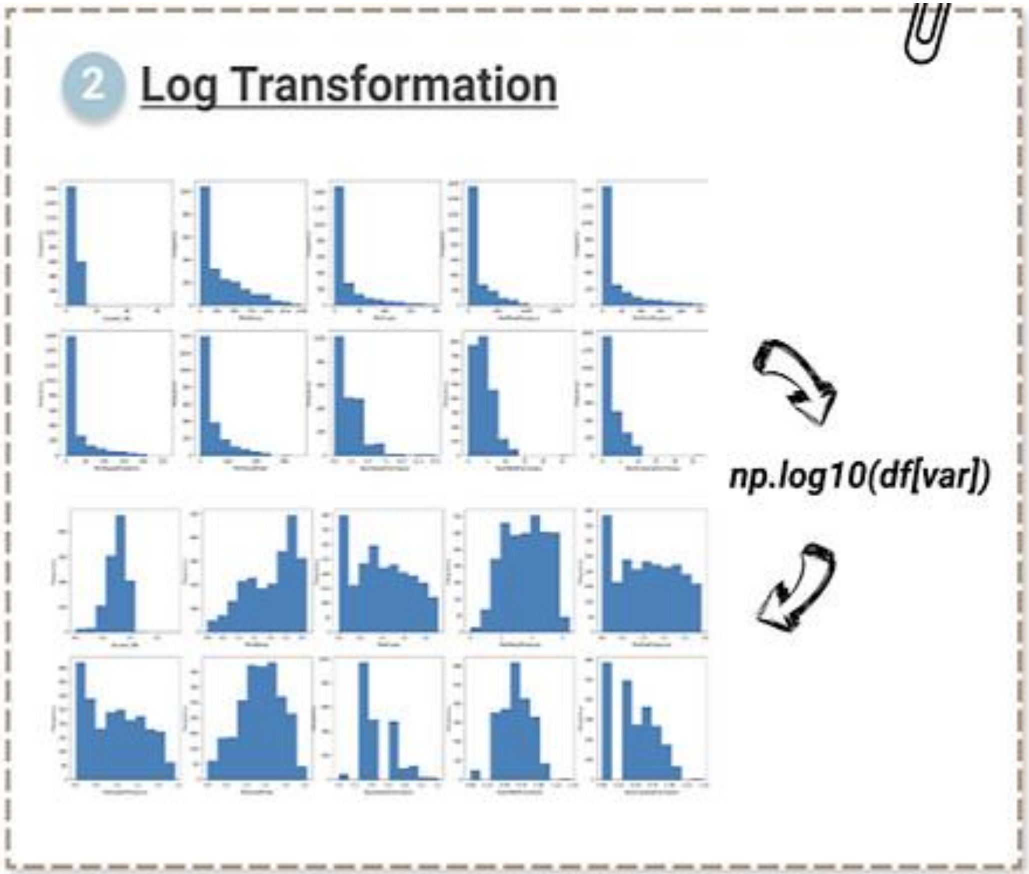
- The data is normalised by shifting the decimal point of its values. By dividing each data value by the maximum absolute value of the data, we can use this technique to normalise the data. The following formula is used to normalise the data value, v, of the data to v’:

$$\boxed{v' = \frac{v}{10^j}}$$

- *Let the input data is: -10, 201, 301, -401, 501, 601, 701 To normalize the above data,*
- ***Step 1:*** *Maximum absolute value in given data(m): 701*
- ***Step 2:*** *Divide the given data by 1000 (i.e j=3)*
- ***Result:*** *The normalized data is: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701*

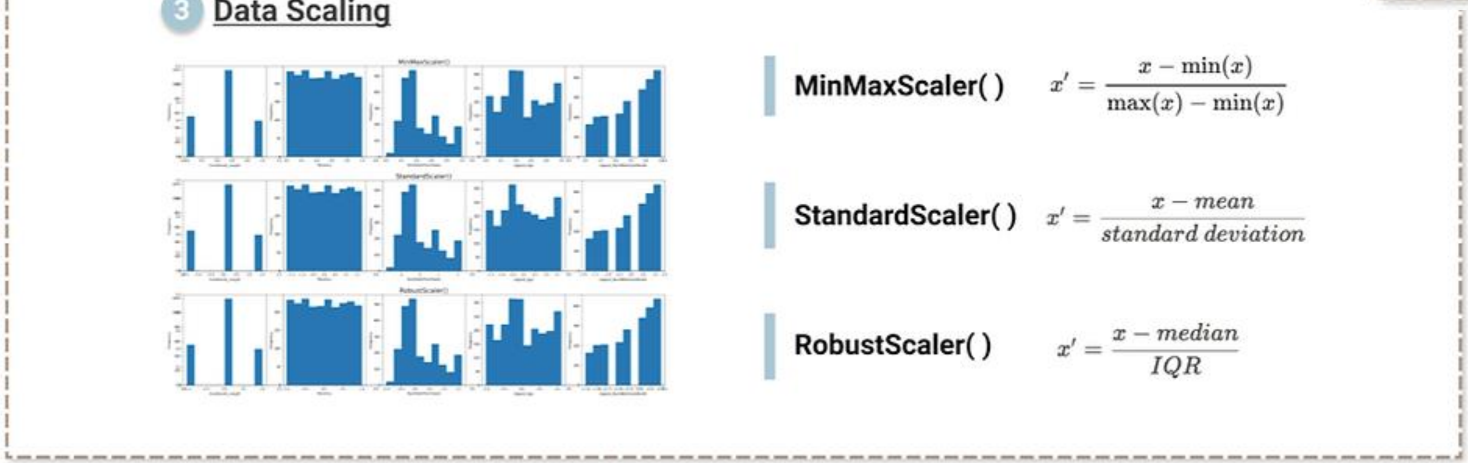
Data Transformation

- **Normalization:** Scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Attribute/feature construction**
  - New attributes constructed from the given ones



2 Log Transformation

np.log10(df[var])



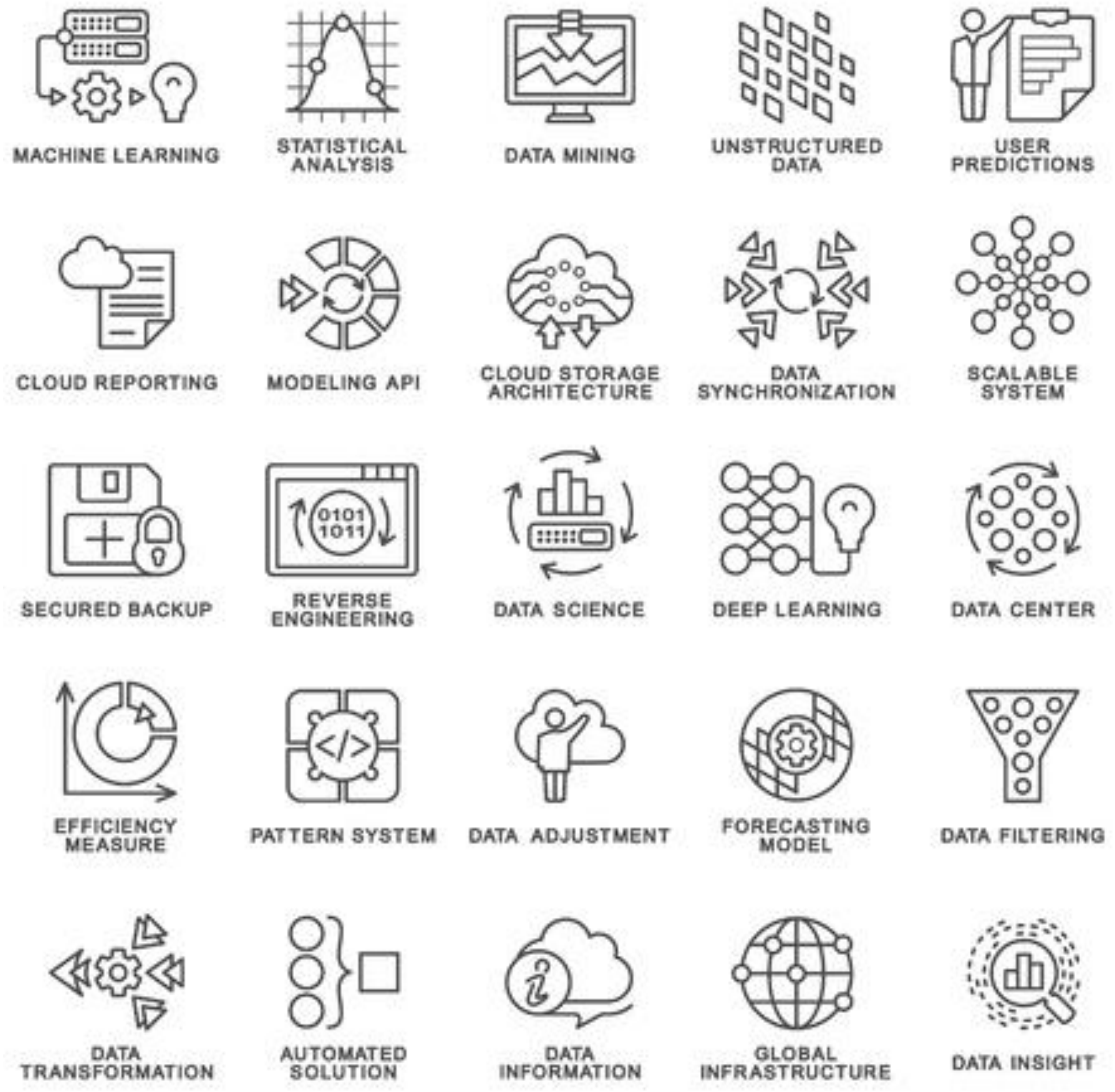
3 Data Scaling

MinMaxScaler( ) $x' = \dfrac{x - \min(x)}{\max(x) - \min(x)}$

StandardScaler( ) $x' = \dfrac{x - mean}{standard\ deviation}$

RobustScaler( ) $x' = \dfrac{x - median}{IQR}$

Uses of Data Transformation



- **Advantages of Data Transformation in Data Mining:**

1. Improves Data Quality: Data transformation helps to improve the quality of data by removing errors, inconsistencies, and missing values.
2. Facilitates Data Integration: Data transformation enables the integration of data from multiple sources, which can improve the accuracy and completeness of the data.
3. Improves Data Analysis: Data transformation helps to prepare the data for analysis and modeling by normalizing, reducing dimensionality, and discretizing the data.
4. Increases Data Security: Data transformation can be used to mask sensitive data, or to remove sensitive information from the data, which can help to increase data security.
5. Enhances Data Mining Algorithm Performance: Data transformation can improve the performance of data mining algorithms by reducing the dimensionality of the data and scaling the data to a common range of values.

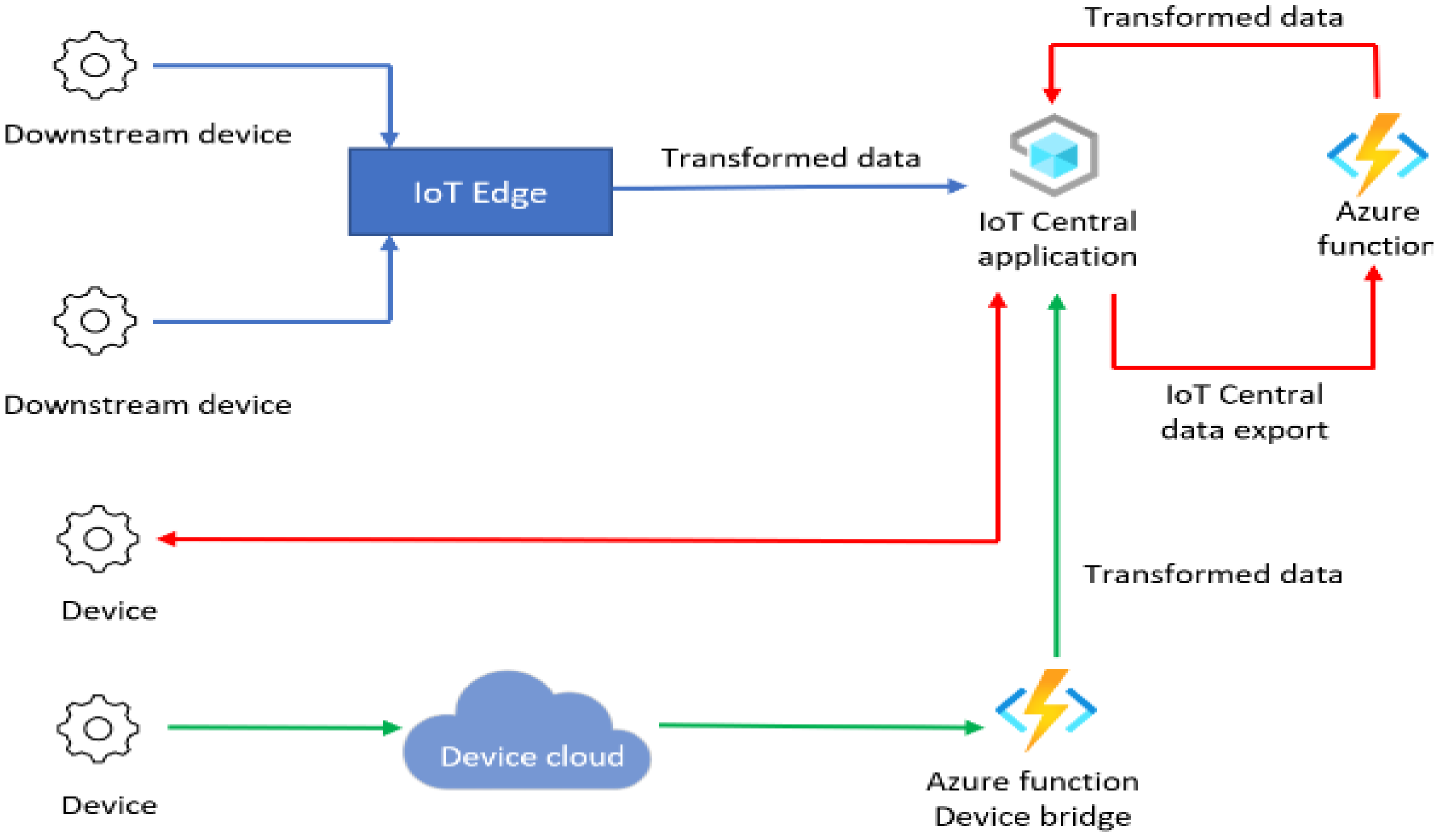
- **Disadvantages of Data Transformation in Data Mining:**

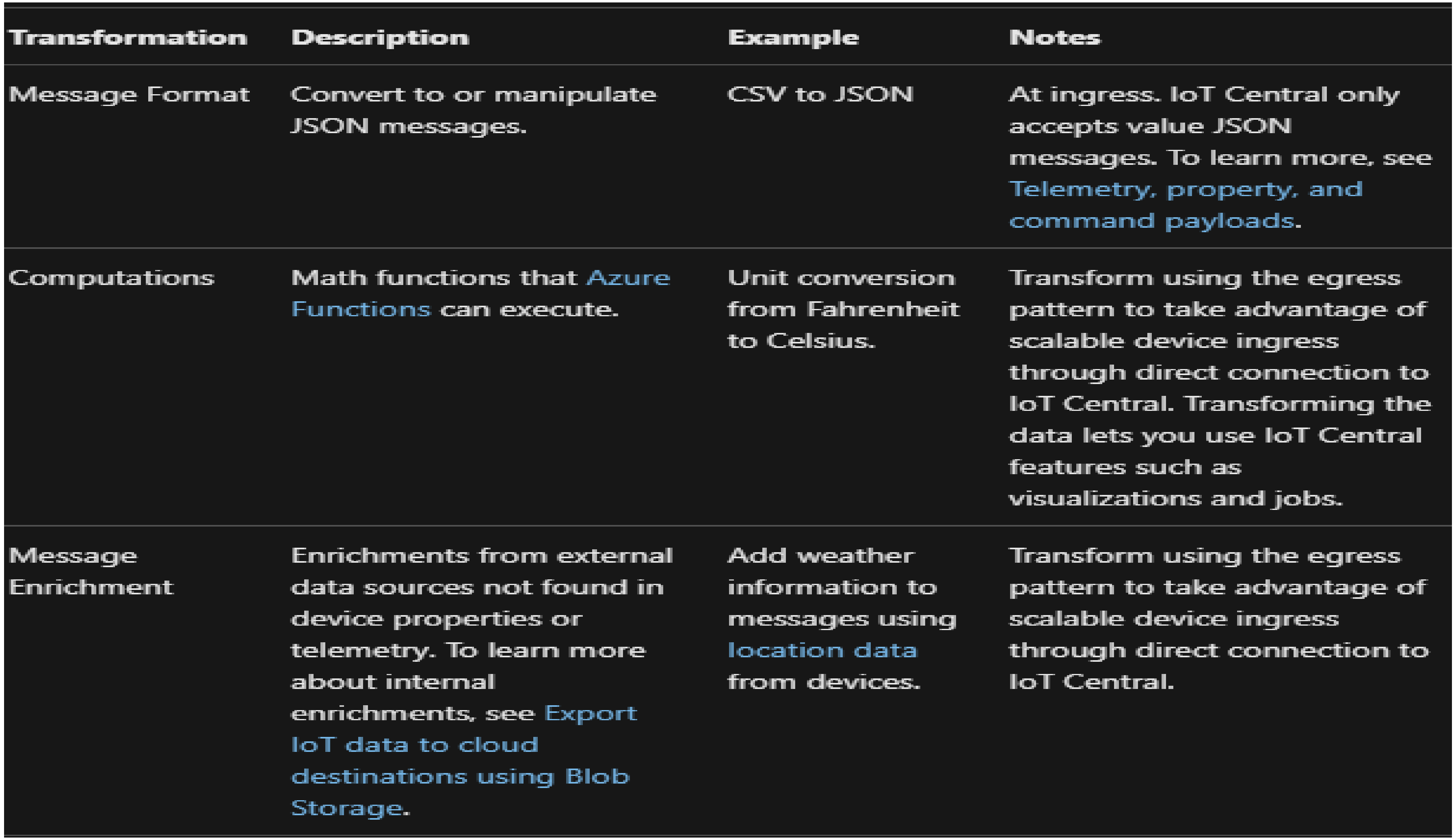
1. Time-consuming: Data transformation can be a time-consuming process, especially when dealing with large datasets.
2. Complexity: Data transformation can be a complex process, requiring specialized skills and knowledge to implement and interpret the results.
3. Data Loss: Data transformation can result in data loss, such as when discretizing continuous data, or when removing attributes or features from the data.
4. Biased transformation: Data transformation can result in bias, if the data is not properly understood or used.
5. High cost: Data transformation can be an expensive process, requiring significant investments in hardware, software, and personnel.

- Overfitting: Data transformation can lead to overfitting, which is a common problem in machine learning where a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new unseen data.

Data transformation

- IoT devices send data in various formats. To use the device data with your IoT Central application, you may need to use a transformation to:
- Make the format of the data compatible with your IoT Central application.
- Convert units.
- Compute new metrics.
- Enrich the data from other sources.


Downstream device
IoT Edge
Transformed data
Downstream device
Transformed data
IoT Central application
Azure functior
IoT Central data export
Device
Transformed data
Device
Device cloud
Azure function Device bridge



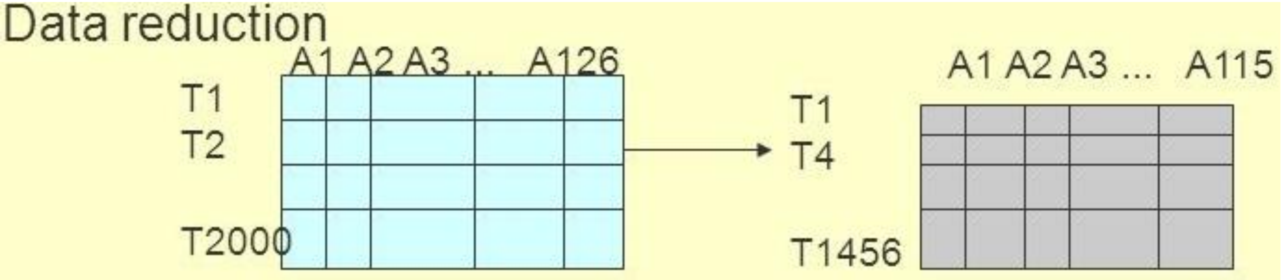
Transformation	Description	Example	Notes
Message Format	Convert to or manipulate JSON messages.	CSV to JSON	At ingress. IoT Central only accepts value JSON messages. To learn more, see Telemetry, property, and command payloads.
Computations	Math functions that Azure Functions can execute.	Unit conversion from Fahrenheit to Celsius.	Transform using the egress pattern to take advantage of scalable device ingress through direct connection to IoT Central. Transforming the data lets you use IoT Central features such as visualizations and jobs.
Message Enrichment	Enrichments from external data sources not found in device properties or telemetry. To learn more about internal enrichments, see Export IoT data to cloud destinations using Blob Storage.	Add weather information to messages using location data from devices.	Transform using the egress pattern to take advantage of scalable device ingress through direct connection to IoT Central.

- **Data transformation at ingress**
- To transform device data at ingress, there are two options:
- **IoT Edge**: Use an IoT Edge module to transform data from downstream devices before sending the data to your IoT Central application.
- **IoT Central device bridge**: The IoT Central device bridge connects other IoT device clouds, such as Sigfox, Particle, and The Things Network, to IoT Central. The device bridge uses an Azure function to forward the data and you can customize the function to transform the device data.

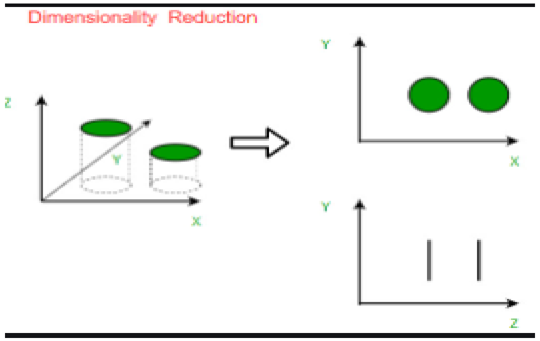
- In this scenario, an IoT Edge module transforms the data from downstream devices before forwarding it to your IoT Central application. At a high level, the steps to configure this scenario are:

1. **Set up an IoT Edge device**: Install and provision an IoT Edge device as a gateway and connect the gateway to your IoT Central application.
2. **Connect downstream device to the IoT Edge device:** Connect downstream devices to the IoT Edge device and provision them to your IoT Central application.
3. **Transform device data in IoT Edge:** Create an IoT Edge module to transform the data. Deploy the module to the IoT Edge gateway device that forwards the transformed device data to your IoT Central application.
4. **Verify**: Send data from a downstream device to the gateway and verify the transformed device data reaches your IoT Central application.

Data Reduction



- Data Reduction: Process of transforming numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form.
- The underlying concept is the reduction of vast amounts of data down to the meaningful parts.
- In the case of discrete data, reduction involves smoothing and interpolation. In the case of digital data, reduction involves some editing, scaling, encoding, sorting, collating, etc.
- Benefits of data reduction: (i) Reduced representation of data will be smaller in volume but produces the same (or similar) analytical results; (ii) Data analytics/mining on the reduced data set consumes much lesser time compared to that done on the complete data set.

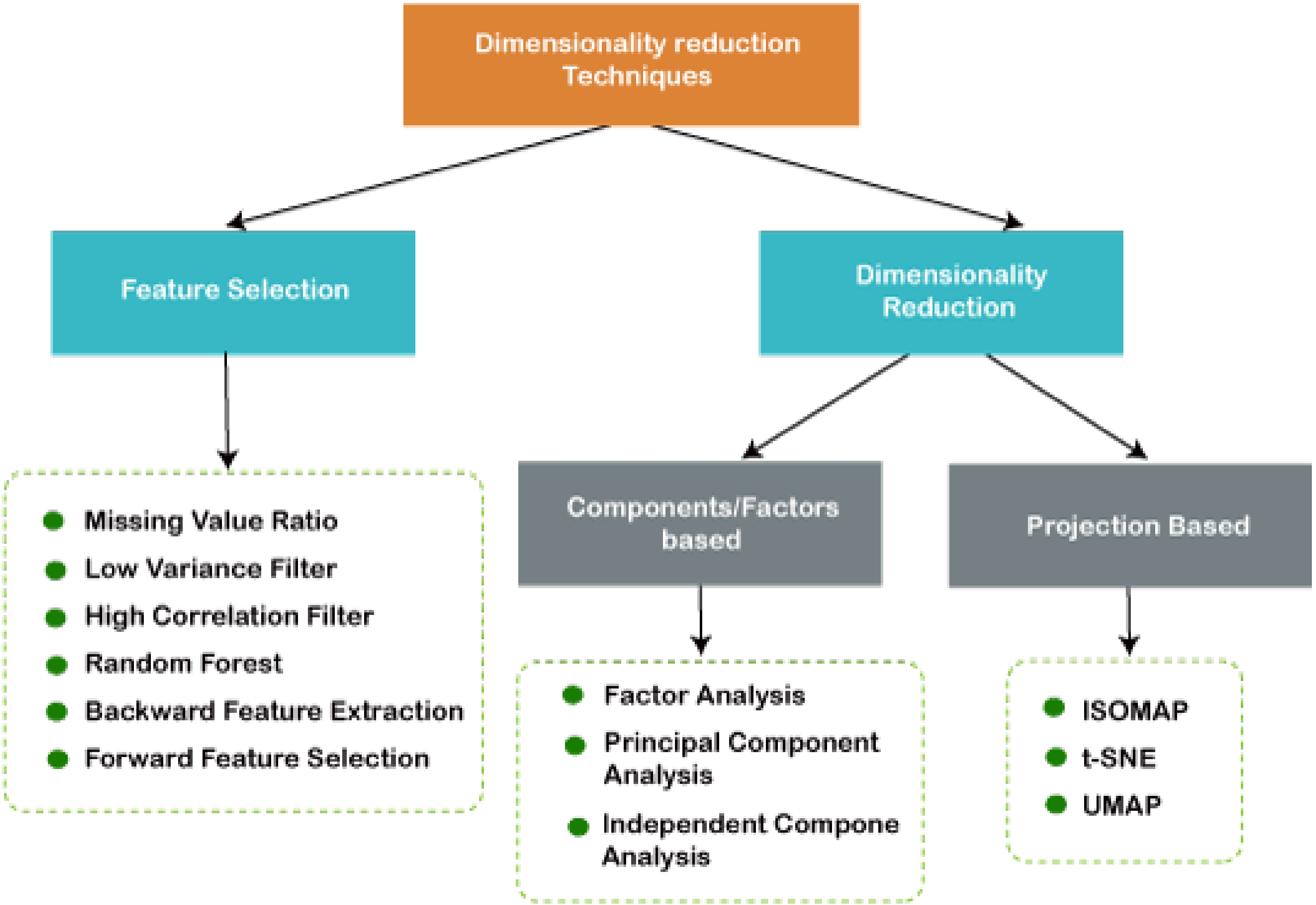


- In machine learning, high-dimensional data refers to data with a large number of features or variables.
- The curse of dimensionality is a common problem in machine learning, where the performance of the model deteriorates as the number of features increases.
- This is because the complexity of the model increases with the number of features, and it becomes more difficult to find a good solution.
- In addition, high-dimensional data can also lead to overfitting, where the model fits the training data too closely and does not generalize well to new data.
- Dimensionality reduction can help to mitigate these problems by reducing the complexity of the model and improving its generalization performance. There are two main approaches to dimensionality reduction: feature selection and feature extraction.

Data Reduction and Strategies :

- Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:
- **Feature Selection:** This involves selecting a subset of relevant features from the dataset based on a criteria. Feature selection is often performed to remove irrelevant or redundant features from the dataset.This transforms raw data into a new representation by creating new features, often combining with existing ones. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).
- **Dimensionality reduction: reduces** the number of features in the existing dataset while preserving the information. Example PCA
- **Feature Extraction:** This involves transforming the data into a lower-dimensional space while preserving the important information. This creates new features by transforming the original ones. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).
- **Data Aggregation**: Combining data points to create summary statistics.Simplifies data by retaining key information.
- **Sampling:** This involves working with a subset of the data rather than the entire dataset. This speeds up computation. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.
- **Binning or discretization**: Grouping continuous data into intervals or bins. Reduces complexity and handles noise.
- **Clustering:** This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.
- **Compression:** This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

Data Reduction-Strategies



Dimensionality reduction Techniques

Feature Selection

- Missing Value Ratio
- Low Variance Filter
- High Correlation Filter
- Random Forest
- Backward Feature Extraction
- Forward Feature Selection

Dimensionality Reduction

Components/Factors based

- Factor Analysis
- Principal Component Analysis
- Independent Compone Analysis

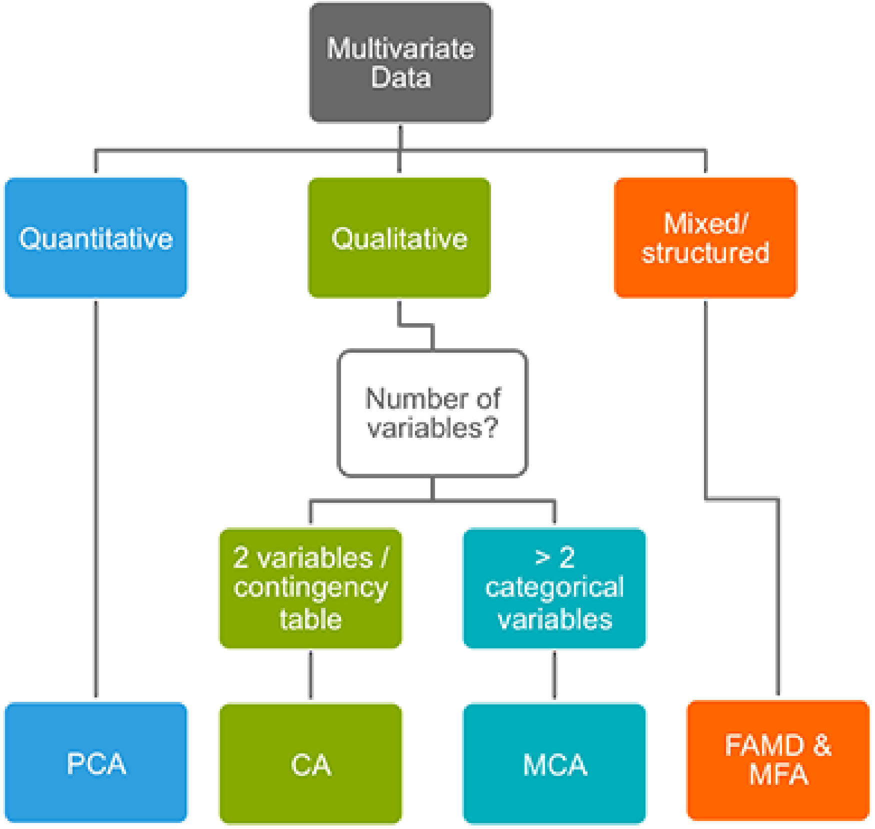
Projection Based

- ISOMAP
- t-SNE
- UMAP

Data Reduction-Strategies

DIMENSIONALITY REDUCTION

Methods to Summarize & Visualize Multivariate Data



- PCA: Principal Component Analysis
- (M) CA: (Multiple) Correspondence Analysis
- FAMD: Factor Analysis of Mixed Data
- MFA: Multiple Factor Analysis

Exploratory data analysis

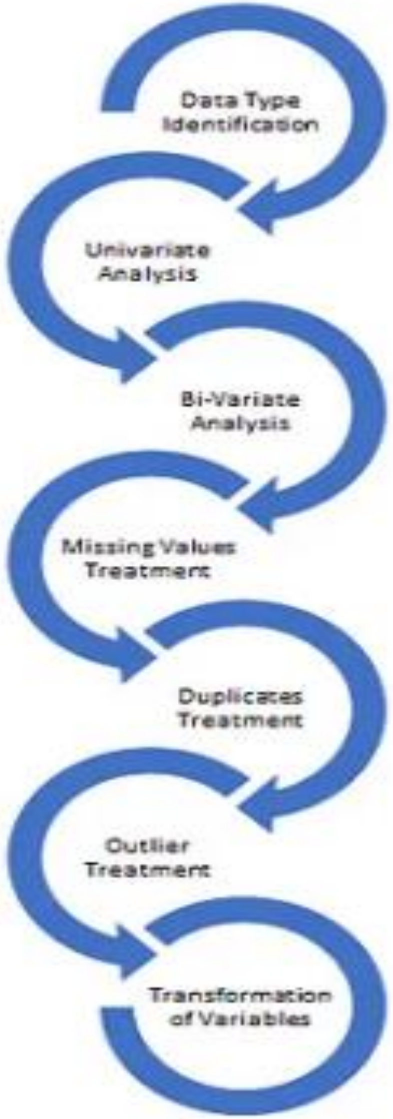
- Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.
- EDA is normally carried out as a preliminary step before undertaking extra formal statistical analysis or modelling.
- It is a process in data analysis where one visually and statistically examines a dataset to understand its main characteristics.
- **Exploratory Data Analysis (EDA)**, also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

Goals of EDA

- 1. Data Cleaning: EDA involves examining the information for errors, lacking values, and inconsistencies. It includes techniques including records imputation, managing missing statistics, and figuring out and getting rid of outliers.
- 2. Descriptive Statistics: EDA utilizes precise records to recognize the important tendency, variability, and distribution of variables. Measures like suggest, median, mode, preferred deviation, range, and percentiles are usually used.
- 3. Data Visualization: EDA employs visual techniques to represent the statistics graphically. Visualizations consisting of histograms, box plots, scatter plots, line plots, heatmaps, and bar charts assist in identifying styles, trends, and relationships within the facts.
- 4. Feature Engineering: EDA allows for the exploration of various variables and their adjustments to create new functions or derive meaningful insights. Feature engineering can contain scaling, normalization, binning, encoding express variables, and creating interplay or derived variables.
- 5. Correlation and Relationships: EDA allows discover relationships and dependencies between variables. Techniques such as correlation analysis, scatter plots, and pass-tabulations offer insights into the power and direction of relationships between variables.
- 6. Data Segmentation: EDA can contain dividing the information into significant segments based totally on sure standards or traits. This segmentation allows advantage insights into unique subgroups inside the information and might cause extra focused analysis.
- 7. Hypothesis Generation: EDA aids in generating hypotheses or studies questions based totally on the preliminary exploration of the data. It facilitates form the inspiration for in addition evaluation and model building.

- EDA, or Exploratory Data Analysis, refers back to the method of analyzing and analyzing information units to uncover styles, pick out relationships, and gain insights. There are various sorts of EDA strategies that can be hired:
- 1. Univariate Analysis: This sort of evaluation makes a speciality of analyzing character variables inside the records set. It involves summarizing and visualizing a unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Techniques like histograms, field plots, bar charts, and precis information are generally used in univariate analysis.
- 2. Bivariate Analysis: Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.
- 3. Multivariate Analysis: Multivariate analysis extends bivariate evaluation to encompass greater than variables. It ambitions to apprehend the complex interactions and dependencies among more than one variables in a records set. Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.
- 4. Time Series Analysis: This type of analysis is mainly applied to statistics sets that have a temporal component. Time collection evaluation entails inspecting and modeling styles, traits, and seasonality inside the statistics through the years. Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA (AutoRegressive Integrated Moving Average) fashions are generally utilized in time series analysis.
- 5. Missing Data Analysis: Missing information is a not unusual issue in datasets, and it may impact the reliability and validity of the evaluation. Missing statistics analysis includes figuring out missing values, know-how the patterns of missingness, and using suitable techniques to deal with missing data. Techniques along with lacking facts styles, imputation strategies, and sensitivity evaluation are employed in lacking facts evaluation.
- 6. Outlier Analysis: Outliers are statistics factors that drastically deviate from the general sample of the facts. Outlier analysis includes identifying and knowledge the presence of outliers, their capability reasons, and their impact at the analysis. Techniques along with box plots, scatter plots, z-rankings, and clustering algorithms are used for outlier evaluation.
- 7. Data Visualization: Data visualization is a critical factor of EDA that entails creating visible representations of the statistics to facilitate understanding and exploration. Various visualization techniques, inclusive of bar charts, histograms, scatter plots, line plots, heatmaps, and interactive dashboards, are used to represent exclusive kinds of statistics.

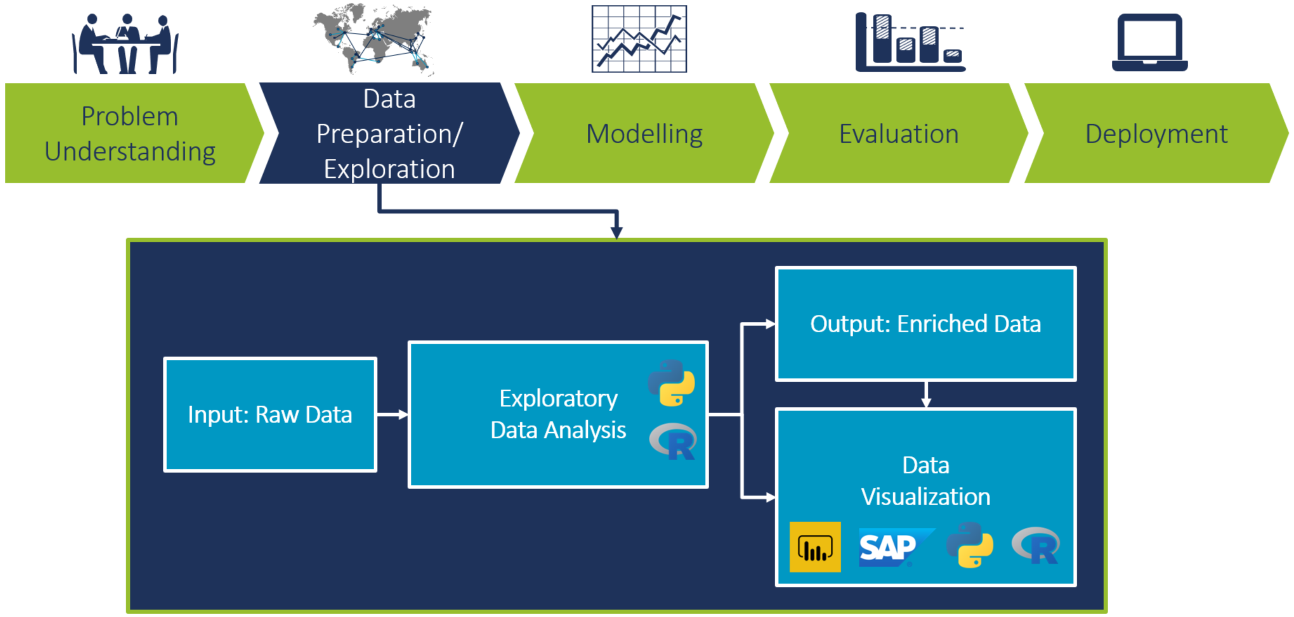
- Extracting important variables and leaving behind useless variables
- Identifying outliers, missing values, or human error
- Understanding the relationship(s), or lack of, between variables
- Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process
- Exploratory Data Analysis does two main things:
- It helps clean up a dataset.
- It gives you a better understanding of the variables and the relationships between them.

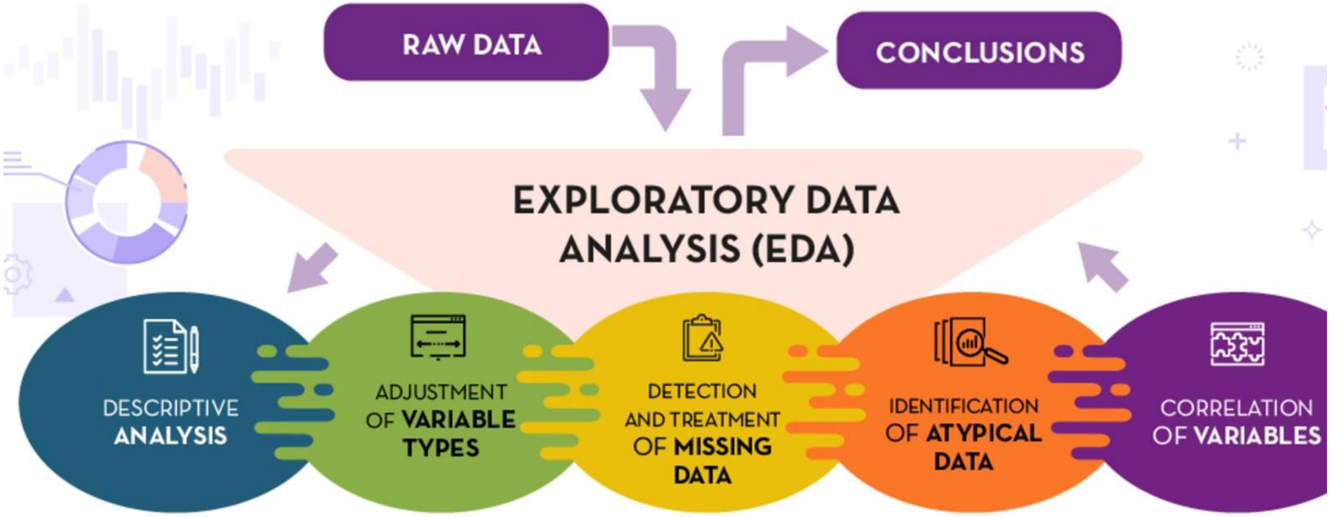


Steps in Exploratory Data Analysis

Significance of Exploratory Data Analysis (EDA)

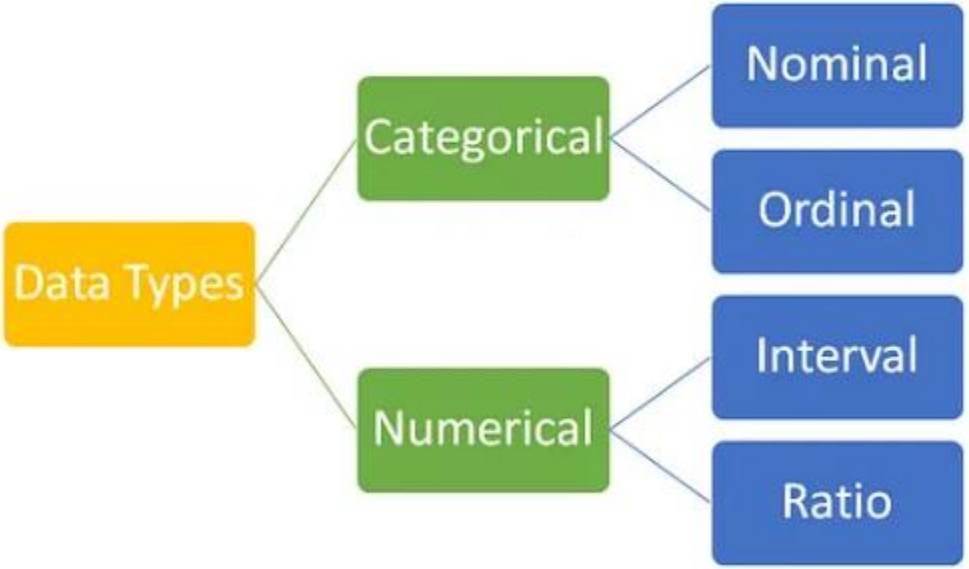
- ***Exploratory Data Analysis*** provides the context needed to develop an appropriate model – and interpret the results correctly.
- The purpose of Exploratory Data Analysis is essential to tackle specific tasks such as:
  - •Spotting missing and erroneous data
  - •Mapping and understanding the underlying structure of your data
  - •Identifying the most important variables in your dataset
  - •Testing a hypothesis or checking assumptions related to a specific model
  - •Establishing a parsimonious model (one that can explain your data using minimum variables)





RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

Understanding data types



Data Types

Categorical

Numerical

Nominal

Ordinal

Interval

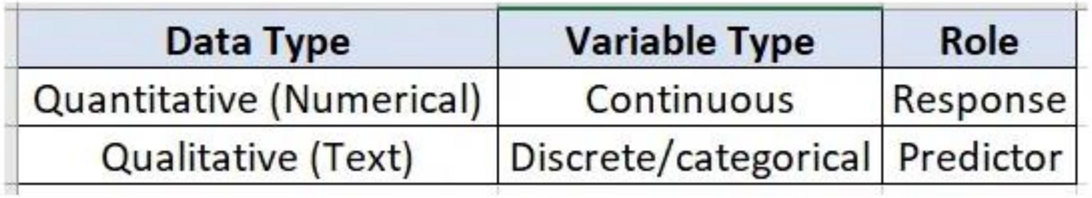
Ratio

Types of Data

1. **Categorical Data:** This is the type of data which has categories or characteristics. It could be something like, say , Gender, race, color, Likert Scale etc. There are two types of categorical data.

- a) *Nominal Data:* Here the data is something discrete & non-quantitative, like Gender, color, race etc. We cannot rank it.
- b) *Ordinal Data:* Here the data is discrete but it can be ordered. Example a Likert Scale or the educational background, organization structure(Associate, Senior Associate, Manager..etc.).
- 2. **Numerical Data:** This is a type of data as mentioned quantitative in nature. There are two types of numerical data.
- a) *Interval:* Here, the data is ordered & we can calculate the differences between two data points. Example — Temperature. But, these data points do not have a true ***ZERO*** value.
- b) *Ratio:* The data type here is also ordered & is same as interval data but it has a true ZERO value. Good example would be distance, weight etc., which cannot be negative & has a true zero.

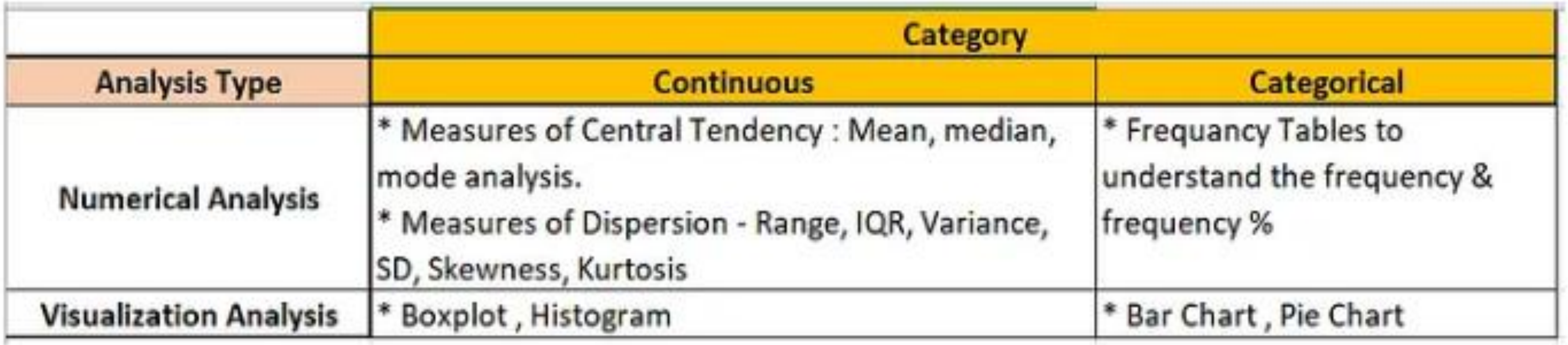
- Data Type Identification: Structured data is in the form of rows & columns. Columns in a dataset can represent two types: Predictor (Input), Response (Output), more so in Supervised Learning Model. In case of unsupervised learning all the variables act as inputs to be grouped. The below table shows data identification categories.



Data Type	Variable Type	Role
Quantitative (Numerical)	Continuous	Response
Qualitative (Text)	Discrete/categorical	Predictor

Profiling the variables.

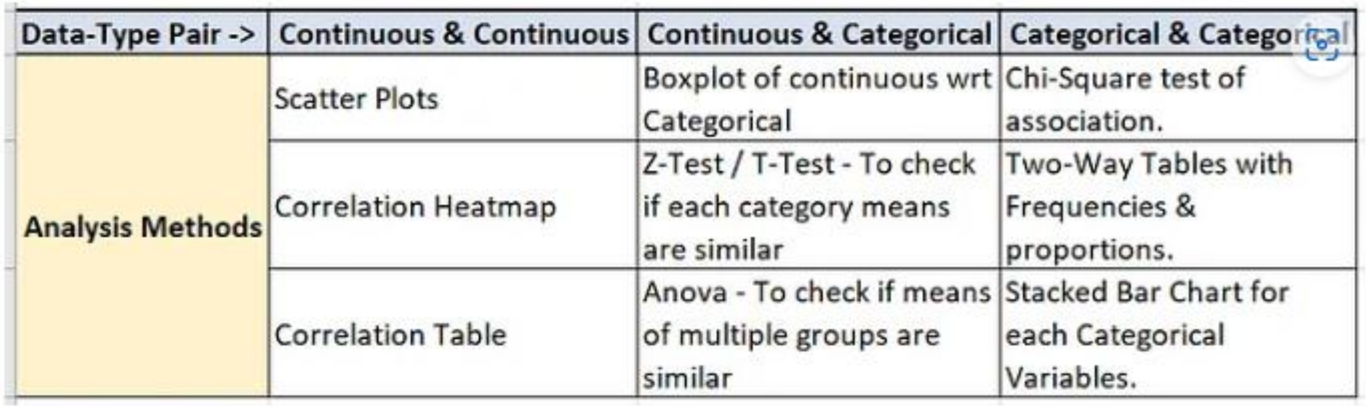
- ***Univariate Analysis:*** As the name suggests, this is analysis of individual variables. We have 2 types of variables i.e., Continuous & Categorical. The types of analysis to be done for each data type is given below.
- This focuses on examining the distribution and characteristics of a single variable in a dataset.



	Category	
Analysis Type	Continuous	Categorical
Numerical Analysis	* Measures of Central Tendency : Mean, median, mode analysis. * Measures of Dispersion - Range, IQR, Variance, SD, Skewness, Kurtosis	* Frequancy Tables to understand the frequency & frequency %
Visualization Analysis	* Boxplot , Histogram	* Bar Chart , Pie Chart

Univariate Analysis

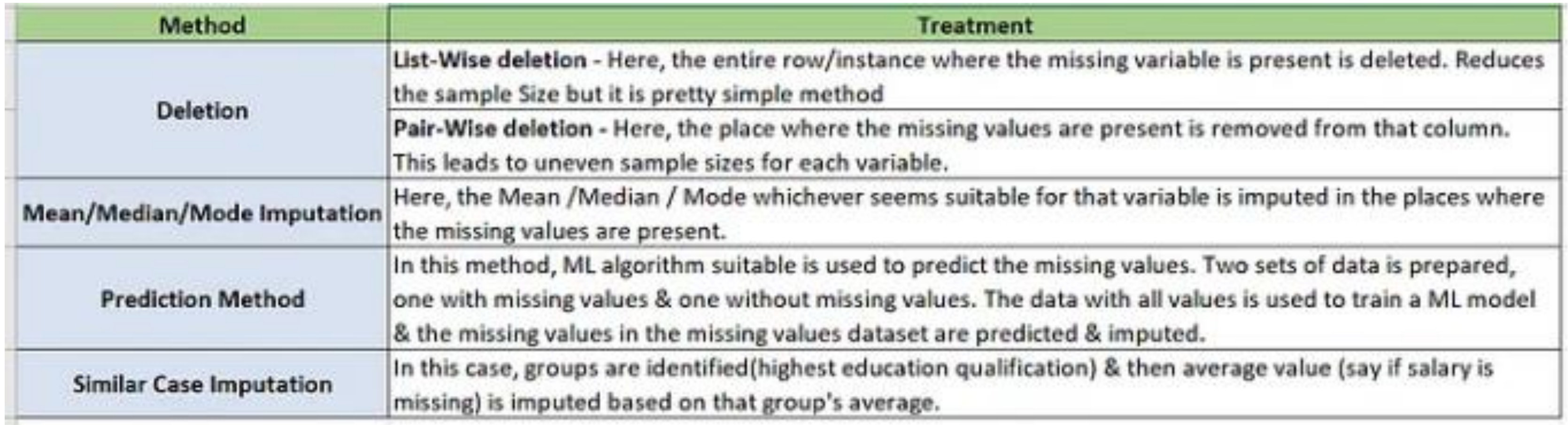
- ***Bi-Variate Analysis:*** Here, we analyze the relationship between any two variables in a dataset to determine patterns .
- Understand how one variable can affect the other variable.
- Knowing the relationship can be very important with respect to Target Variable or within the predictor variables. This can help us understand the variables which could cause un-necessary noise & reduction in performance of the model we build in a later stage. The analysis depends on the data-type of pair we choose.



Data-Type Pair ->	Continuous & Continuous	Continuous & Categorical	Categorical & Categorical
Analysis Methods	Scatter Plots	Boxplot of continuous wrt Categorical	Chi-Square test of association.
	Correlation Heatmap	Z-Test / T-Test - To check if each category means are similar	Two-Way Tables with Frequencies & proportions.
	Correlation Table	Anova - To check if means of multiple groups are similar	Stacked Bar Chart for each Categorical Variables.

Bivariate Data Analysis

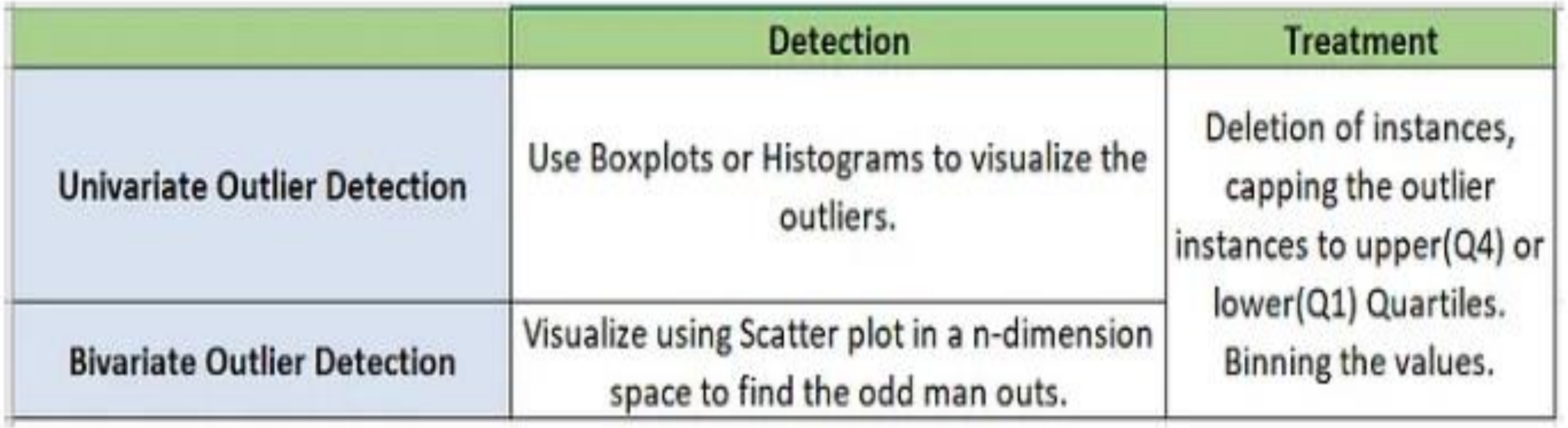
- ***Missing Value Treatment:*** Missing values in a dataset can occur due to unavailable data or manual errors. It is very important to consider & treat the missing values in any dataset since it is determinantal to model performance. The methods to treat missing values are summarized in table below,



Method	Treatment
Deletion	**List-Wise deletion** - Here, the entire row/instance where the missing variable is present is deleted. Reduces the sample Size but it is pretty simple method
	Pair-Wise deletion - Here, the place where the missing values are present is removed from that column. This leads to uneven sample sizes for each variable.
Mean/Median/Mode Imputation	Here, the Mean /Median / Mode whichever seems suitable for that variable is imputed in the places where the missing values are present.
Prediction Method	In this method, ML algorithm suitable is used to predict the missing values. Two sets of data is prepared, one with missing values & one without missing values. The data with all values is used to train a ML model & the missing values in the missing values dataset are predicted & imputed.
Similar Case Imputation	In this case, groups are identified(highest education qualification) & then average value (say if salary is missing) is imputed based on that group's average.

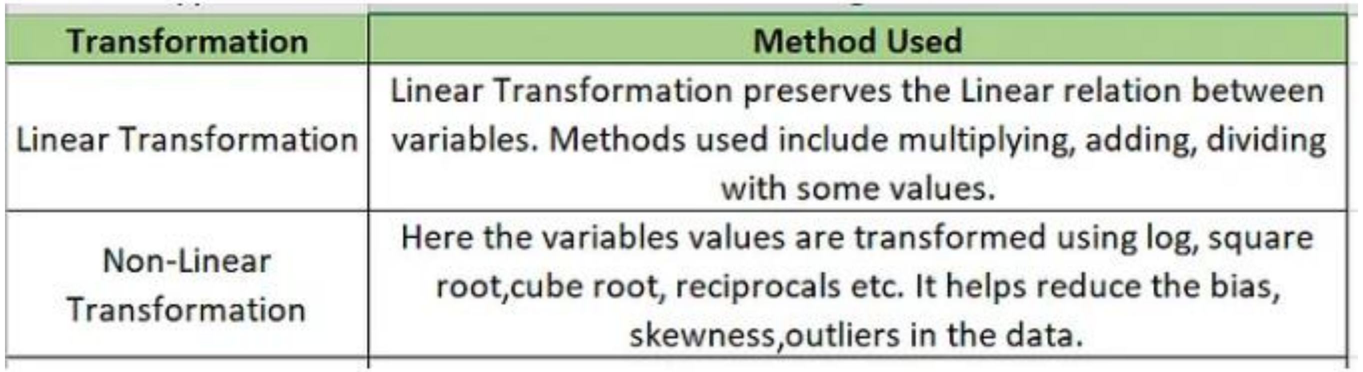
Missing Values Treatment

- ***Duplicate Value Treatment:*** Here the duplicate instances are just deleted from the dataset. Duplicates can reduce the effectiveness of the model created.
- ***Outlier Treatment:*** Outliers are un-naturally large or small values in variable column. If we take data of net-worth of some individuals in an area & find abnormally high value in the data collected then it might mean that, that particular case is of a rich businessman or executive living in that area. Outliers need not occur naturally always, it could also be clerical errors when the data-entry was done. Find the below table which summarizes outlier Treatment.



	Detection	Treatment
Univariate Outlier Detection	Use Boxplots or Histograms to visualize the outliers.	Deletion of instances, capping the outlier instances to upper(Q4) or lower(Q1) Quartiles. Binning the values.
Bivariate Outlier Detection	Visualize using Scatter plot in a n-dimension space to find the odd man outs.	

- ***Transformation of Variables:*** Why transformation? Not always do we get the variables in a right form. We might need to bring it to a common scale or reduce the variance in the column, make the relationship linear etc. Let us see some methods we could use to do the same.



Transformation	Method Used
Linear Transformation	Linear Transformation preserves the Linear relation between variables. Methods used include multiplying, adding, dividing with some values.
Non-Linear Transformation	Here the variables values are transformed using log, square root,cube root, reciprocals etc. It helps reduce the bias, skewness,outliers in the data.

How important Exploratory Data Analysis (EDA)

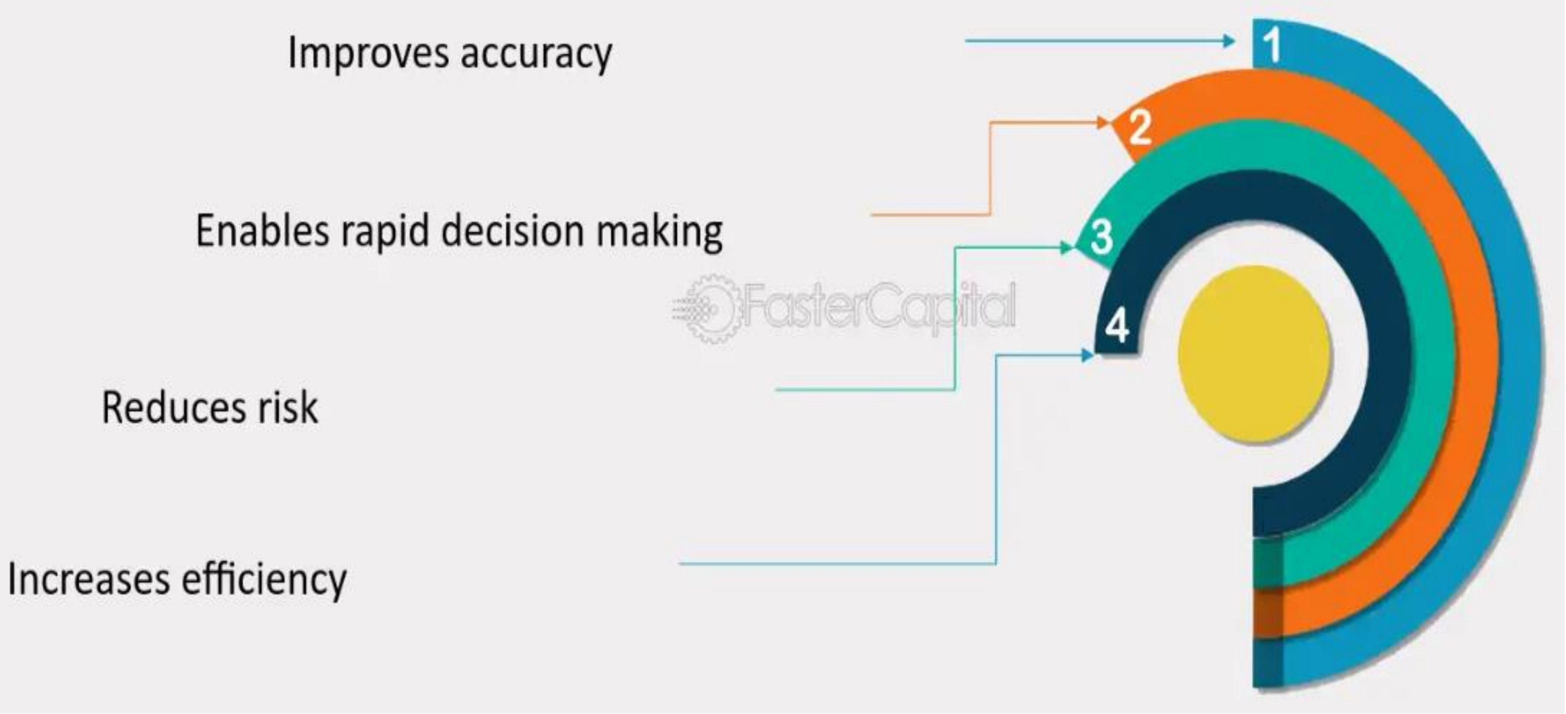
Data Understanding
Data Cleaning
Pattern Discovery
Data Visualization
Model Selection
Quality Control

Making sense of Data

5 characteristics of quality data

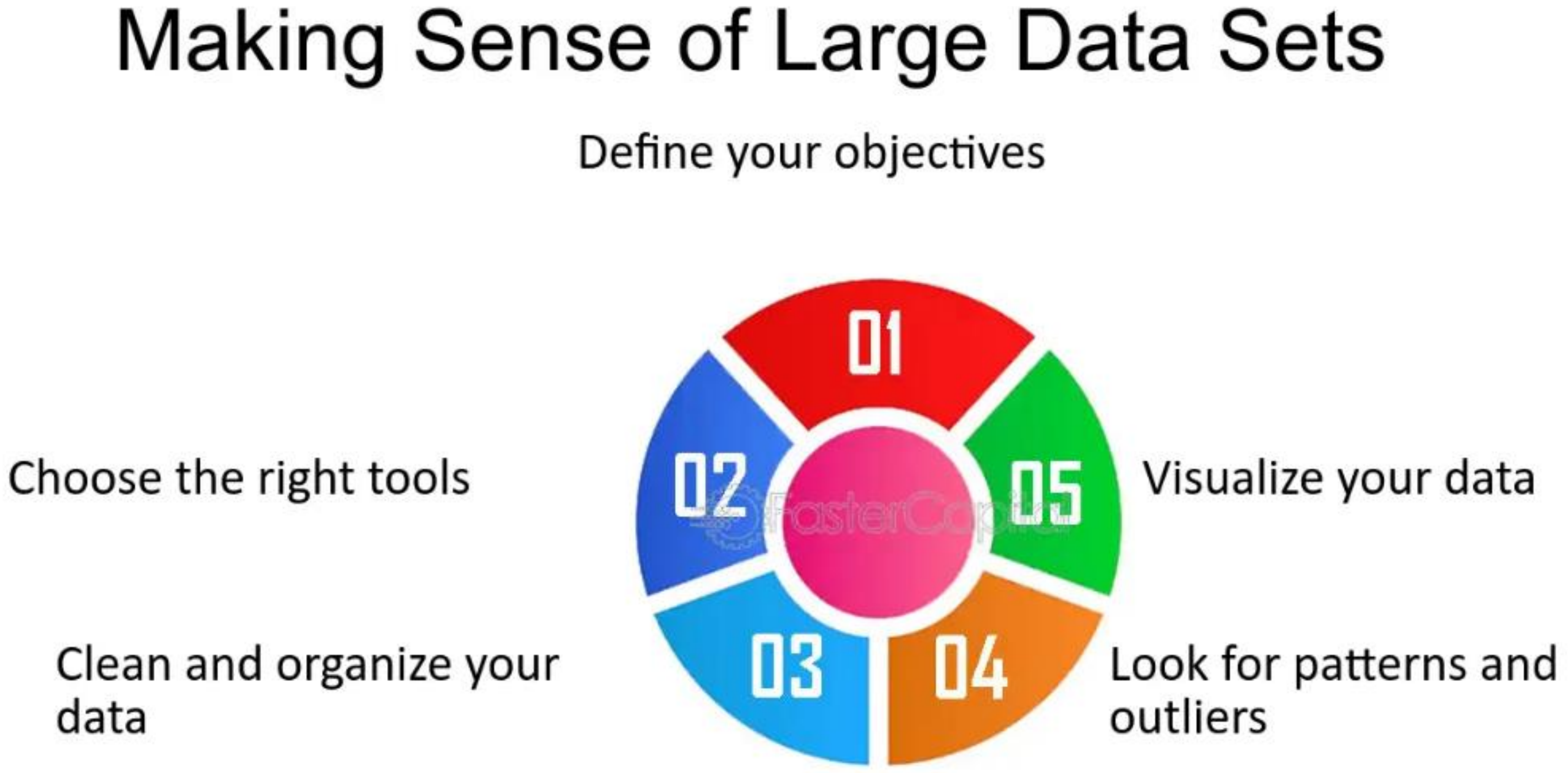
- **Validity.** The degree to which your data conforms to defined rules or constraints.
- **Accuracy.** Ensure your data is close to the true values.
- **Completeness.** The degree to which all required data is known.

Importance of quality data



- **Consistency.** Ensure your data is consistent within the same dataset and/or across multiple data sets.
- **Uniformity.** The degree to which the data is specified using the same unit of measure.

Making sense of Data



Making Sense of Large Data Sets

Define your objectives

Choose the right tools

Clean and organize your data

Look for patterns and outliers

Visualize your data

Making sense of Data

Use Case and Examples

Conversion Rate

Customer Acquisition Cost (CAC)

Average Order Value (AOV)

Customer Lifetime Value (CLV)