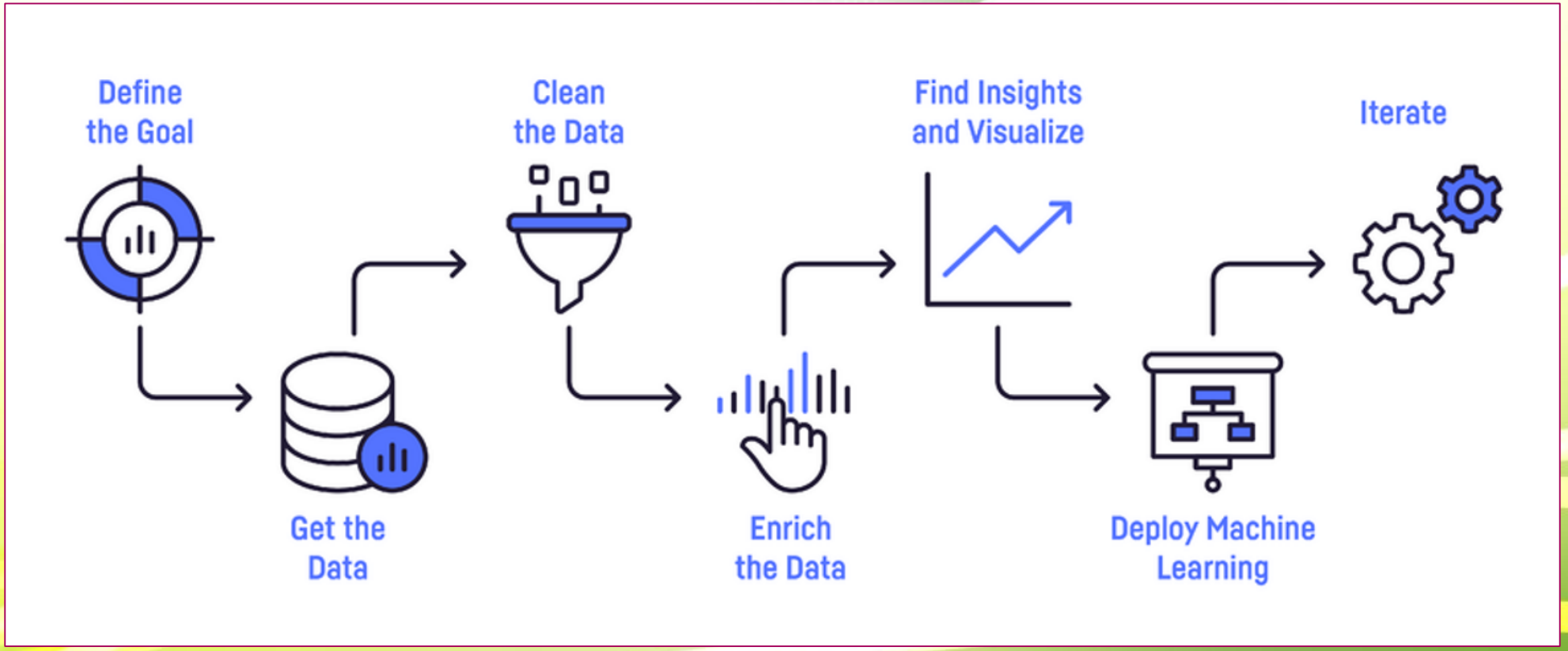
INTRODUCTION TO DATA PREPARATION

2nd Sem, MCA

CONTENT

- Introduction and overview
  - Steps in Data Analytics projects
    - Problem definition stage
    - Data Preparation
      - Data integration,
      - Data cleaning,
      - Missing values, Noisy data,
      - Data transformations,
      - Data partitioning.

STEPS IN DATA ANALYTICS



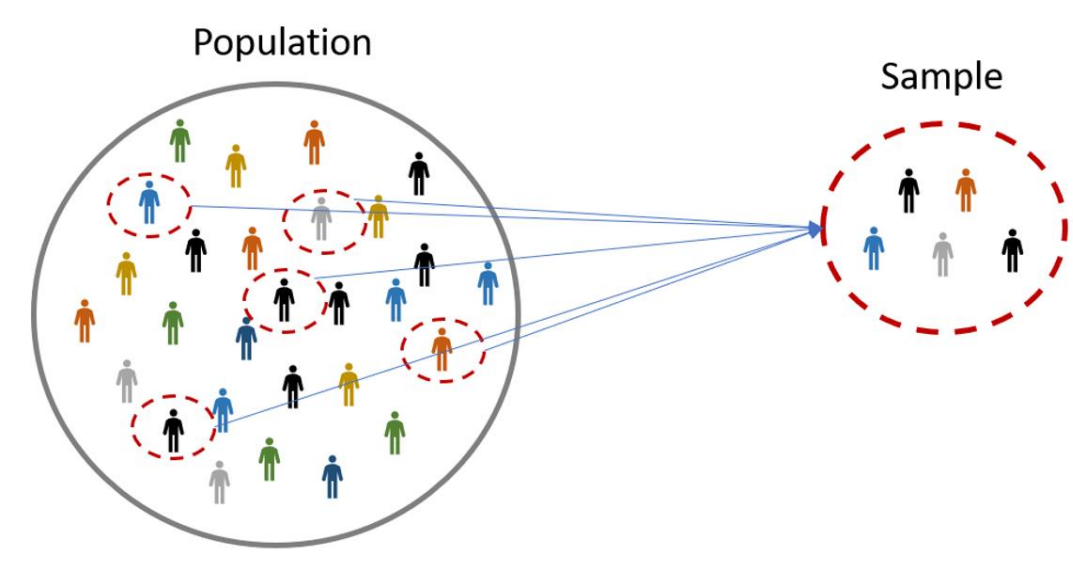
STEPS IN DATA ANALYTICS

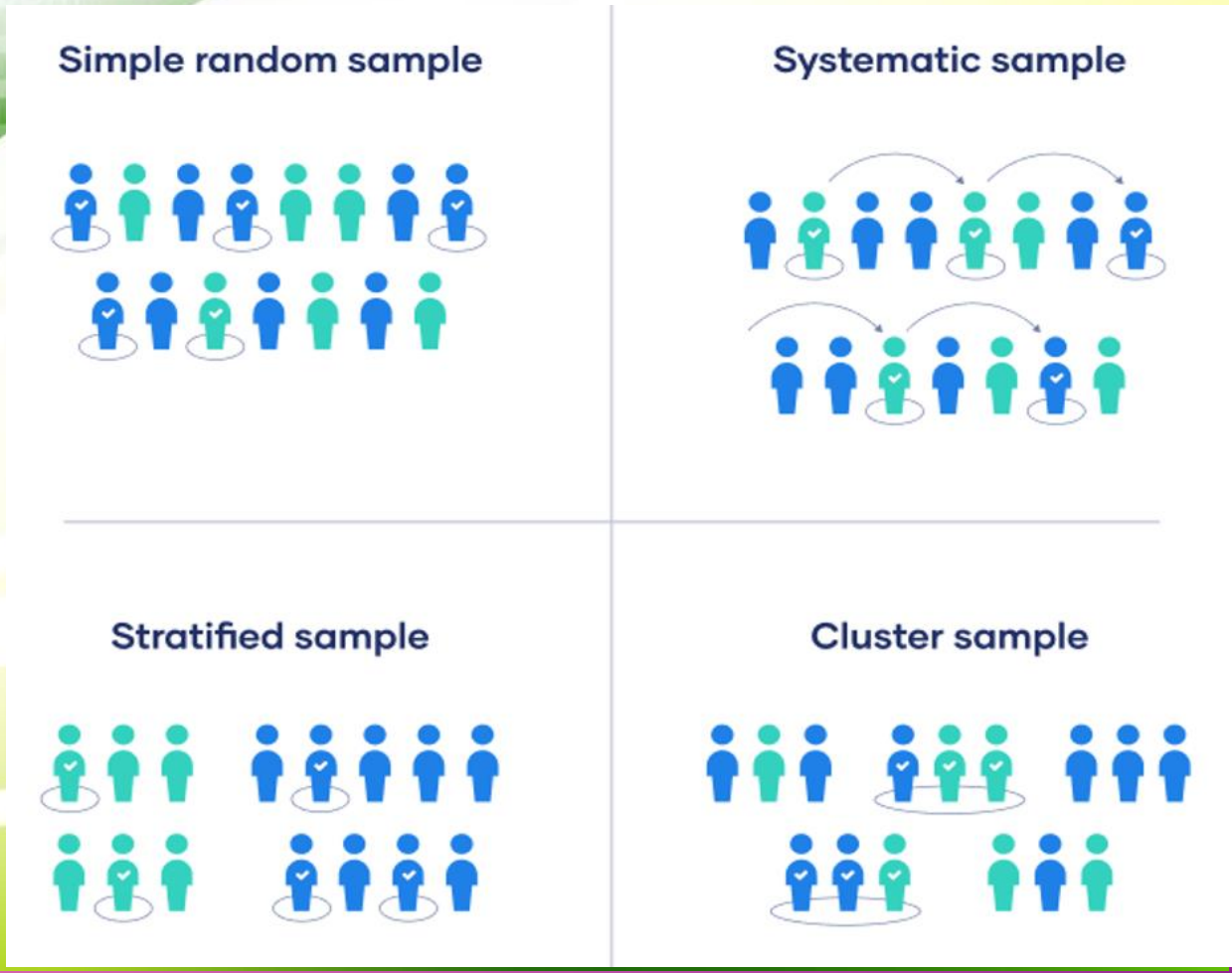
SEMMA Methodology:

- **Sample** – data sampling.
  - large enough dataset to contain sufficient information to retrieve, yet small enough to be used efficiently.
- **Explore** – understanding data by discovering anticipated and unanticipated relationships between variables, and also abnormalities, with help of data visualization.
- **Modify** – methods to select, create and transform variables in preparation for data modeling.
- **Model** – applying various modeling techniques on the prepared variables in order to create models that possibly provide the desired outcome.
- **Assess** – evaluation of the modeling results shows the reliability and usefulness of created models

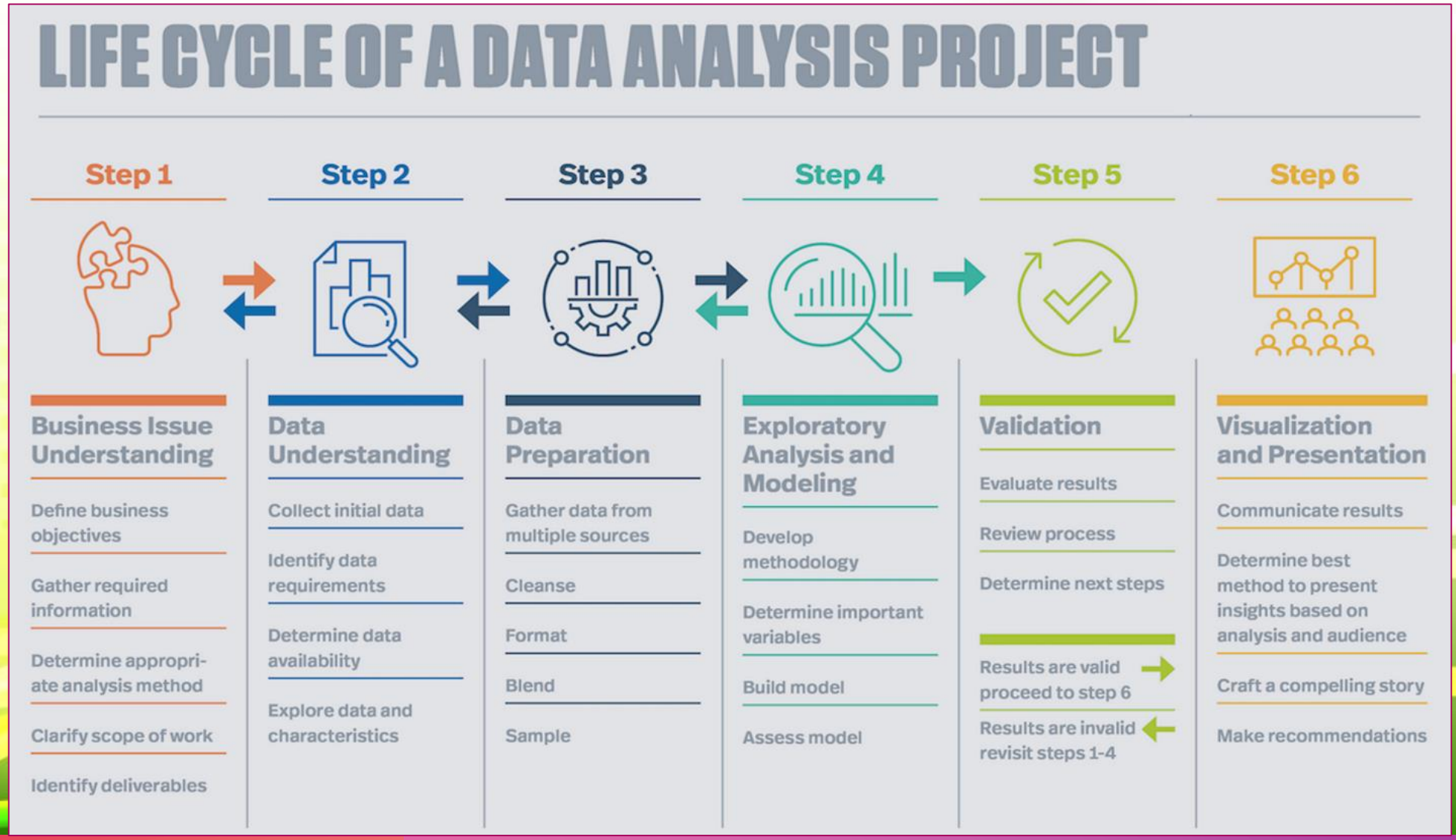
STEPS IN DATA ANALYTICS

- **Population:** group of items whose properties are to be analyzed.
- **Sample:** (suitable) subset of population.
- **Sampling:** process of picking sample





STEPS IN DATA ANALYTICS



STEPS IN DATA ANALYTICS

Understand Business Goals & Expectations

- Begin by understanding the business's vision.
- What are the pain point that they are facing?
- What resources are available?
  - Infrastructure, Pre-requisite features, Transactions, Business Results.
- What are the potential benefits?
- What risks are there in pursuing the project?
- Determine if the expected benefits are realistic and attainable from a data point of view.
- Determine the duration of the project.
- Perspective, see the problem from business point of view and from their client's point of view.
- Ensure to obtain domain knowledge required for particular problem.

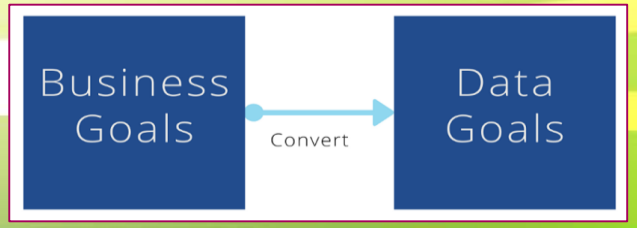
STEPS IN DATA ANALYTICS

Translate Business Goals to Data Analysis Goals

- *Example:* *A Café franchise has outlet A & B.*

  Cafe A wants to increase their profits as much that of Cafe B.

  - What product is more/less popular? (Popularity categorization)
  - How does the price of items at Cafe A compare to that of their competitors at Cafe b? (Price chart analysis)
  - How many customers does Cafe A have as compared to Cafe b?
  - What is the footfall to each café on a timely basis (hourly/daily/weekly)?
  - What are the peak hours at Cafe A and cafe B, is there any convergence?
  - What is the average age of the customers at each cafe?
  - What is the number of repeat customers to each cafe.



- Such details understanding may derive conclusion that ***Cafe A sells less coffee than Cafe B in peak hours***.
  - This changes the problem statement from "**How do we increase profits**?" to " **How to sell more coffee**?"

STEPS IN DATA ANALYTICS

Frame the Problem Statement

- Write statement that describes the problem, why solving the problem is important and a starting point to begin solving it.

  "The problem P. . .": problem as defined by company.

  ". . . has the impact I ." negative impacts/pain points of the problem.

  ". . . which affects B. . ." parties that are affected (business, customers or a third party).

  "…, so a good starting point would be S." benefits of solving the problem.

- *"The problem of low coffee sales, has the impact of decreased profits, which affects Cafe A, so a good starting point would be to compare their coffee price with that of their competitors."*

STEPS IN DATA ANALYTICS

Success Metric:

- Objective of problem statement should be to **generate business insights** and **drive actionable plans.**
- Success of problem statement need to be evaluated at the end.
- Achievement should be measurable.
- Common metrics are:
  - Model assessment: Accuracy, Performance etc.
  - Benchmarks:
    - *Increase coffee sales by at least 10% in first month of solution implementation.*

STEPS IN DA - DATA PREPARATION

- *Data processing* is collecting raw data and translating it into usable information.
  - Raw data is *collected, filtered, sorted, processed, analyzed, stored*, and presented in a readable format.
- Data Processing can be: **Manual, Mechanical, Electronic.**
- Types of Data Processing:
  - **Batch Processing:** data is collected and processed in batches (for large amounts of data).
  - **Single User Programming Processing:** done by single person for personal use (for smaller data).
  - **Multiple Programming Processing:** simultaneously storing and executing multiple program; processed using two or more CPUs. parallel processing.
  - **Real-time Processing:** processing, which always remains under execution.
  - **Online Processing:** entry and execution of data directly (no need to store; to reduce data entry errors)
  - **Time-sharing Processing:** one form of online data processing that facilitates several users to share resources on time-based manner.
  - **Distributed Processing:** remote systems remain interconnected forming network for processing.

STEPS IN DA - DATA PREPARATION

- Data preparation: Process of gathering, combining, structuring and organizing data so it can be used in business intelligence.

- **Steps in Data Preparation:**
  - Collect
  - Discover
  - Clean
  - Transform
  - Validate

STEPS IN DA - DATA PREPARATION

- **Data discovery and profiling.** explore the collected data to better understand what it contains and what needs to be done to prepare it for intended uses.
  - identifies patterns, relationships and other attributes in data, as well as inconsistencies, anomalies, missing values and other issues that can be addressed.
  - ***Data Inspection:*** Detect unexpected, incorrect, and inconsistent data.
- **Data cleansing.** Correct the identified data errors and issues to create complete and accurate data sets.
  - faulty data is removed or fixed,
  - missing values are filled in
  - inconsistent entries are harmonized.
- **Data structuring.** data is modeled and organized to meet analytics requirements.
- **Data transformation and enrichment.** transformed into a unified and usable format.
- **Data validation and publishing.** data is validated for its consistency, completeness and accuracy.

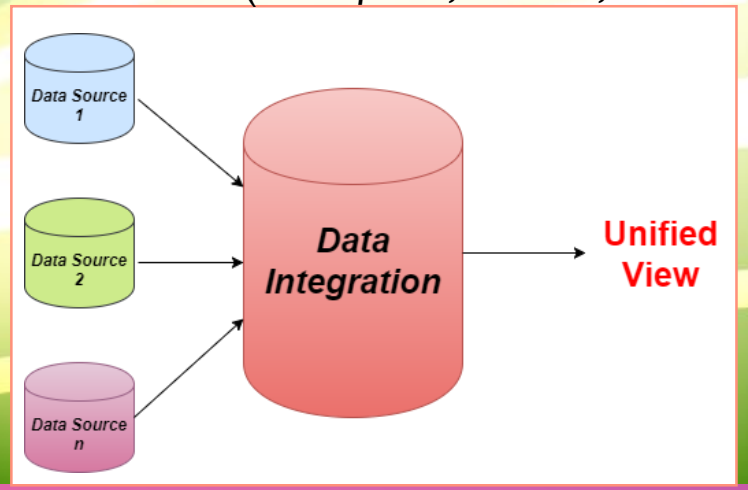
STEPS IN DA - DATA PREPARATION

Challenges in Data (need for Data preparation)**:**

- **Inadequate or nonexistent data profiling.** errors, anomalies and other problems might not be identified, which can result in flawed analytics.
- **Missing or incomplete data.** must be fixed to ensure analytics accuracy
- **Invalid data values.** Misspellings, other typos and wrong numbers.
- **Name & address standardization.** may be inconsistent with variations that can affect accuracy in analysis
- **Inconsistent data across enterprise systems.**
- **Data enrichment.**
- **Maintaining and expanding data prep processes.** Data preparation work often becomes a recurring process that needs to be sustained and enhanced on an ongoing basis.

STEPS IN DA - DATA PREPARATION

- **Data integration/Data consolidation**: process of combining data from different sources into a single, unified view.
- Data consolidation techniques
  - **Hand-coding.** Using manual process for small, uncomplicated data collection *(time-consuming for exploding volumes of data)*.
  - **ETL software.** ETL applications can pull data from multiple sources, transform it into necessary format and then transfer it to final data storage location.
  - **ELT tools.** Data from cloud may not support ETL much. With ELT, first extract data from sources and load it into data warehouse and then transform that data (*much faster, scalable, and more cost-effective)*.



STEPS IN DA - DATA PREPARATION

- **Data integration approaches:**
  - **Extract, Transform and Load**: copies of datasets from disparate sources are gathered together, harmonized, and loaded into a data warehouse or database.
  - **Extract, Load and Transform**: data is loaded into a big data system and transformed at later time for particular analytics uses.
  - **Change Data Capture**: identifies data changes in databases in real-time and applies them to a data warehouse or other repositories.
  - **Data Replication**: data in one database is replicated to other databases to keep the information synchronized to operational uses and for backup as well.
  - **Data Virtualization**: data from different systems are virtually combined to create a unified view rather than loading data into a new repository.
  - **Streaming Data Integration**: a real time data integration method.

STEPS IN DA - DATA PREPARATION

2 types of data integration *(tight & loose Coupling)*

- **Tight Coupling:**
  - data is combined from different sources into a single physical location through the process of ETL.
- **Loose Coupling:**
  - data remains only in the actual source databases.
  - an interface is provided that takes query from user/system, transforms it in a way the source database can understand, and then sends query directly to source databases to obtain ONLY the result.

Issues in Data Integration:

- Schema Integration,
- Redundancy,
- Data value conflicts.

STEPS IN DA - DATA PREPARATION

- **Data profiling:** explore the collected data to better understand what it contains and what needs to be done to prepare it for intended uses.
- A **summary statistics** about the data is helpful to give a general idea about the quality of data.
- **Data Inspection:** Detect unexpected, incorrect, and inconsistent data.
  - Quality of data is critical for final analysis.
  - Data which tend to be incomplete, noisy and inconsistent can effect result.
  - **Data cleaning** is the process of detecting and removing corrupt/inaccurate records from record, table or database.
- By analyzing and **visualizing** data using statistical methods such as *mean, standard deviation, range, quantiles, etc.* one can find values that are unexpected and thus erroneous.

STEPS IN DA - DATA PREPARATION

- **Data cleaning:** process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
  - Check and maintain Data Quality
    - **Validity:** degree to which data conform to defined business rules or constraints.
    - **Accuracy:** degree to which data is close to the true values *(error margin)*
    - **Completeness:** degree to which all required data is known.
    - **Consistency:** degree to which data is consistent.
    - **Uniformity:** degree to which data is specified using same unit of measure.
  - May happen due to wrong data, or during data integration.
  - If data is incorrect, outcomes and algorithms are unreliable.

STEPS IN DA - DATA PREPARATION

- Data cleaning: process to remove/correct data that does not belong in dataset.
- Data transformation: process of converting data from one structure into another (data wrangling/munging).
- Incorrect data is either ignored, removed, corrected, or imputed.

- **Removal of unwanted observations**
  - deleting duplicate/redundant/irrelevant values from dataset.
- **Fixing Structural errors**
  - errors due to measurement, data transfer, type conversion, etc..
  - include typos in name of features, same attribute with a different name, mislabeled classes, separate classes that should really be the same, or inconsistent capitalization. *(e.g. "N/A" and "Not Applicable")*
- **Managing Unwanted outliers & missing value**
  - Outliers & missing value can cause problems with certain types of models

STEPS IN DA - DATA PREPARATION

- **Handling missing data** (drop / fill / flag)
  - *Missing data is like missing a puzzle piece. If you drop it, that's like pretending the puzzle slot isn't there. If you impute it, that's like trying to squeeze in a piece from somewhere else in the puzzle.*

1. **Dropping** observations with missing values.
2. **Imputing** the missing values with *global constant/statistical observation/popular value.*
   - Using **statistical values** like *mean, median, mode*, etc.
   - Mean is most useful when original data is not skewed, while median is more robust, not sensitive to outliers (used when data is skewed).
   - Using a **linear regression** (with best fit line between two variables)
   - Manual or automated (depending on data size)
3. **Hot-deck**: Copying values from other similar records.
4. **Flag:** Filling missing values leads to a loss in information.
   - **Ignore** tuple/record: not very effective, unless tuple contains several attributes with missing values.
   - Missing data is informative in itself, and algorithm should know about it.

STEPS IN DA - DATA PREPARATION

- **Noisy data** is meaningless data (corrupt data).
  - Data that cannot be understood or interpreted correctly by machines (e.g. unstructured text).
  - Spelling errors, industry abbreviations and slang can also impede machine reading.
- Noisy data unnecessarily increases data storage space required and can also adversely affect analysis results.
  - Noisy data can be caused by *faulty data collection instruments, data entry problems, technology limitation, hardware failures, programming errors, gibberish input from speech or optical character recognition (OCR) programs, etc.*
- **Example** :
  - *Unknown encoding:* *Marital Status — Q*
  - *Out of range values:* *Age — -10*
  - *Inconsistent Data:* *DoB — 4th Oct 1999, Age — 50*
  - *inconsistent formats:* *DoJ — 13th Jan 2000, DoL — 10/10/2016)*

STEPS IN DA - DATA PREPARATION

- **Regression:** Data can be smoothed by fitting the data into a regression functions.
- **Clustering:** Outliers may be detected by clustering (similar values are organized into groups/clusters).
  - Values that fall outside of the set of clusters may be considered outliers.
  - may be smoothed or removed.

- Noise can be handled using **binning**.
  - Smoothing of sorted data is done using the values around it.

1. *Sorted data is placed into bins or buckets.*
2. *Bins can be created by equal-width (distance) or equal-depth (frequency) partitioning.*
3. *On these bins, smoothing can be applied (by bin mean, bin median or bin boundaries).*
4. *Outliers can be treated by using binning and then smoothing it.*

STEPS IN DA - DATA PREPARATION

Example

Data = 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin a: 4, 8, 15

Bin b: 21, 21, 24

Bin c: 25, 28, 34

Smoothing by bin means:

each value in a bin is replaced by the mean value of the bin.

Bin a: 9, 9, 9

Bin b: 22, 22, 22

Bin c: 29, 29, 29

Smoothing by bin boundaries:

each bin value is replaced by the closest boundary value.

Bin a: 4, 4, 15

Bin b: 21, 21, 24

Bin c: 25, 25, 34

STEPS IN DA - DATA PREPARATION

- **Outlier** extreme values that deviate from other observations on data; indicate variability in measurement.

- Common causes of outliers on data set:
  - Data entry errors (human errors)
  - Measurement errors (instrument errors)
  - Experimental errors (data extraction or experiment planning/executing errors)
  - Intentional (dummy outliers made to test specific methods)
  - Data processing errors (data manipulation or data set unintended mutations)
  - Sampling errors (extracting or mixing data from wrong or various sources)
  - Natural (not an error, novelties in data))

STEPS IN DA - DATA PREPARATION

Types of Outlier:

- Type 1 – Global/Point outlier
- Type 2 – Contextual outlier
- Type 3 – Collective outlier

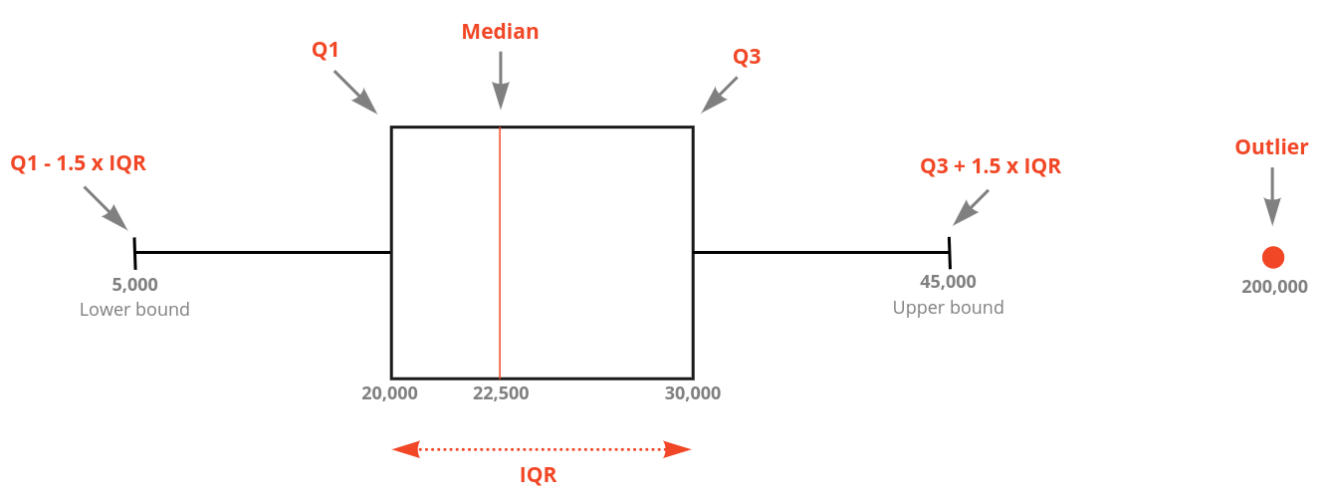
- **Univariate outliers** found when looking at distribution of values in single feature space.
- **Multivariate outliers** found in n-dimensional space (n-features).

- Outlier detection methods:
  - Numeric Outlier
  - Z-Score Analysis
  - DBSCAN
  - Isolation forest

STEPS IN DA - DATA PREPARATION

Numerical outlier Detection Method:

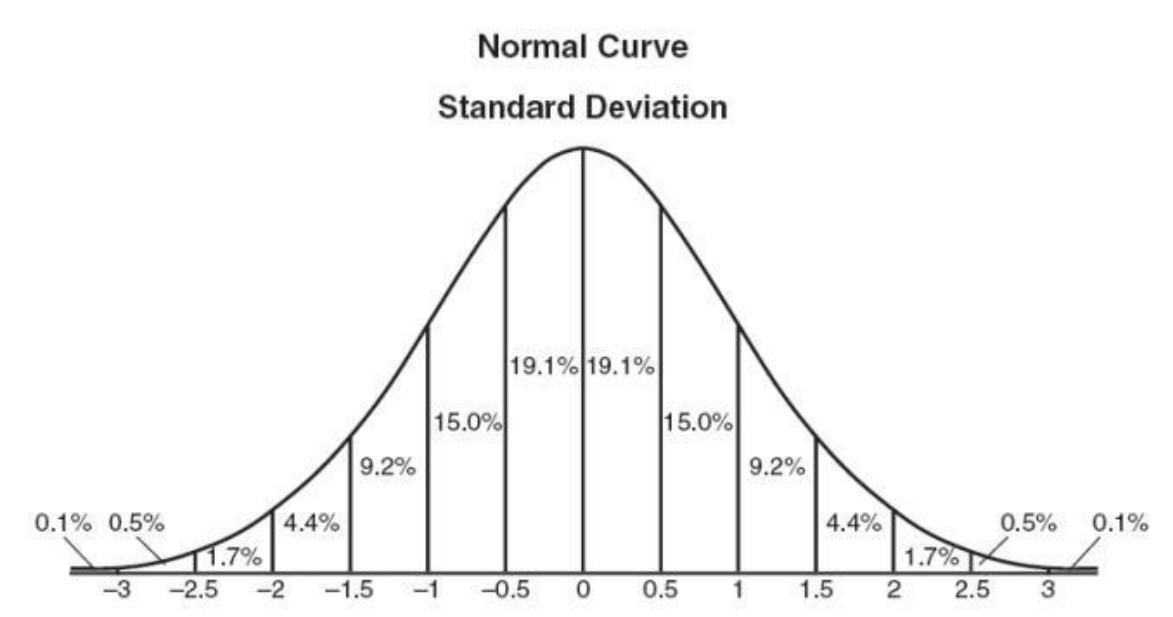
- Simple, non-standard outlier detection technique in one-dimensional feature space.
- Exteriors are calculated by IQR (Inter Quartile Range).
- Analyze with upper & lower bounds using interquartile amplifier value k = 1.5.
    - lower bound/fence = Q1 - 1.5(IQR)
    - outlier < lower bound/fence
    - upper bound/fence = Q3 + 1.5(IQR)
    - outlier > upper bound/fence



STEPS IN DA - DATA PREPARATION

Z-Score outlier Detection Method:

- Z-score tells how many standard deviations away a given observation is from mean.
  - 68% of the data points lie between +/- 1 standard deviation.
  - 95% of the data points lie between +/- 2 standard deviation
  - 99.7% of the data points lie between +/- 3 standard deviation



STEPS IN DA - DATA PREPARATION

Z-Score outlier Detection Method:

- Z-score technique considers Gaussian distribution of data.
- Outliers are data points on tail of the distribution and are therefore far from average.
- Z-score is a ***parametric measure*** and it takes two parameters — *mean and standard deviation.*

  Z score = (x -mean) / std. deviation

- Z-score tells how many standard deviations away a given observation is from mean.
- Limit must be specified in data set → good 'thumb rule' limits may be fixed deviations of 2.5, 3, 3.5, or more.
  - *Example, Z-score of 2.5 means data point is 2.5 standard deviation far from mean.*
  - *Since it is very (2.5 S.D. times) far from center, it's flagged as outlier/anomaly.*

STEPS IN DA - DATA PREPARATION

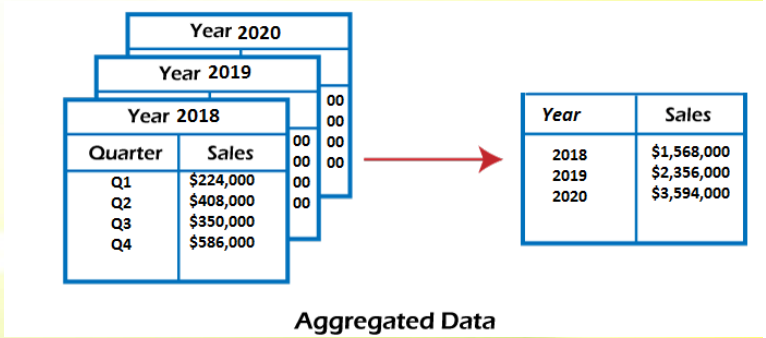
Transformation:

- Processed data are transformed from one format to another format, that is more appropriate for analysis.
- Data transformation may be:
  - **Constructive**: adds, copies, or replicates data.
  - **Destructive**: deletes fields or records.
  - **Aesthetic**: standardizes the data to meet requirements or parameters.
  - **Structural**: reorganized by renaming, moving, or combining columns.
- **Benefits of data transformation**:
  - *Better Organization*
  - *Improved Data Quality*
  - *Perform Faster Queries*
  - *Better Data Management*
  - *More Use Out of Data*

STEPS IN DA - DATA PREPARATION

Data Reduction

- Process of reducing volume of original data to represent in much smaller volume by maintaining integrity of original data.
  - Reducing number of attributes/columns/dimension; and/or number of records/tuples/rows.
  - Efficiency of data mining process is improved, while producing same analytical results.
  - Necessary when processing the entire data set is expensive and time consuming.
- **Data cube aggregation:** aggregation at various levels of data in a simpler form.
- **Dimensionality reduction:** Not all attributes are required for data analysis.
  - Keep most suitable subset of attributes from a large number of attributes
  - techniques like forward selection, backward elimination, decision tree induction or combination these.
- **Data compression:** large volumes of data is compressed (number of bits used to store data is reduced).
  - In *loss compression,* quality of data is compromised for more compression.
  - In *lossless compression,* quality of data is not compromised for higher compression level.
- **Numerosity reduction :** reduces volume of data by choosing smaller forms for data representation.
  - using clustering or sampling of data.



STEPS IN DA - DATA PREPARATION

Data Transformation methods:

- **Smoothing**: process of removing noise from data *(binning, regression, and clustering)*.
- **Aggregation**: process where summary or aggregation operations are applied to data.
  - *daily sales data may be aggregated to compute monthly and annual total amounts (sum, min, max, group, etc)*
- **Generalization**: low-level *(primitive/raw)* data are replaced with high-level data of categorical value.
  - *street_name → city or country;*
  - *Converting text to numbers; color (black, red, white) → 0,1,2*
  - *Converting continuous data to categories; age (15-30,30-50,50-70) → youth, middle-aged, and senior.*
- **Normalization**: Normalize & scale attribute data so as to fall within a small specified range (e.g. 0.0 to 1.0)
- **Attribute Construction**: New attributes are constructed from given set of attributes.
  - *New year and month column for original date column.*

STEPS IN DA - DATA PREPARATION

Data Transformation Process (ETL → Extract, Transform, and Load)

- **Data Discovery**: Understand and identify data in its source format. Decide what they need to do to get data into its desired format.
- **Data Mapping**: Determine how individual fields to be modified, mapped, filtered, joined, and aggregated.
- **Data Extraction**: Extract data from original source.
- **Code Generation and Execution**: Create a code to complete the transformation.
- **Review**: After transforming the data, check it to ensure everything has been formatted correctly.
- **Sending**: Send data to its target destination.

STEPS IN DA - DATA PREPARATION

- When attributes are on different ranges or scales, data modelling and mining can be difficult.
- When multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations.
  - They are normalized to bring all the attributes on the same scale.
- **Normalization** transforms the data to fall under a given range; hence helps in applying data mining algorithms and extracting data faster.
  - *Min-max normalization*
  - *Decimal scaling*
  - *Z-score normalization*

person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

STEPS IN DA - DATA PREPARATION

- **Min-max normalization** − Implements linear transformation on original data.
  - min_A and max_A are minimum and maximum values of an attribute, A.
  - Min-max normalization maps a value, v, of A to v' in the range [new_min_A , new_max_A]

$$v'_i = \frac{v_i - min_A}{max_A - min_A}(new_{max_A} - new_{min_A}) + new\_min_A$$

  - *Example, $1200 and $9800 are minimum, and maximum value for attribute income.*
  - *[0.0, 1.0] is the range in which we need to map a value of $73,600.*
  - *Datapoint $73,600 would be transformed using min-max normalization as follows:*

$$\frac{73600-1200}{9800-1200}(1.0 - 0.0) + 0.0 = 0.716$$

STEPS IN DA - DATA PREPARATION

- **Z-score (zero-mean) normalization** − Values for attribute 'A' are normalized based on mean and standard deviation of A.
- Datapoint, v, of A is normalized to v' by computing

$$v' = \frac{v - A'}{\sigma_A}$$

  - where A' and σ_A are mean and standard deviation of attribute A.
- Useful when actual minimum and maximum of attribute A are unknown, or when there are outliers.

  - *Example, mean and standard deviation for attribute A as $65,000 and $18,000.*
  - *Normalized value $85,800 using z-score normalization is;*

$$\frac{85,800-65,000}{18,000} = 1,156.$$

STEPS IN DA - DATA PREPARATION

- **Decimal Scaling** − Normalizes by changing the decimal point of values of attribute A. This movement of a decimal point depends on the maximum absolute value of A.
- Datapoint, v, of A is normalized to v′ by computing.

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

  - Where j is the smallest integer such that Max (|v′|)<1.

  - *Example, observed values for attribute A range from -986 to 917.*
  - *Maximum absolute value for attribute A is 986.*
  - *To normalize each value of attribute A using decimal scaling, divide each value of attribute A by 1000, i.e., j=3 (number of integers in the largest number).*
  - *So, the value -986 would be normalized to -0.986, and 917 would be normalized to 0.917.*

STEPS IN DA - DATA PREPARATION

- **Data Partitioning:** technique for physically dividing data during the loading of Master Data.
- **For Easy Management:**
  - Data volume in data warehouse can grow up to hundreds of gigabytes.
  - This huge size is very hard to manage as a single entity.
- **To Assist Backup/Recovery**
  - Partitioning allows to load only as much data required on a regular basis (instead of whole data always).
  - Reduces time to load and also enhances system performance.
  - Also beneficial for backup purpose.
- **To Enhance Performance**
  - Query performance enhances, having to process lesser/partitioned (required) data; not entire data..

STEPS IN DA - DATA PREPARATION

Horizontal Partitioning

- **Partitioning by Time into Equal Segments:** Data partitioned on basis of time period (of equal size)
- **Partition by Time into Different-sized Segments:** Implemented as a set of small partitions for relatively current data, larger partition for inactive data. (When specific/aged data is accessed infrequently)
- **Partition on Different Dimension:** Partition on basis of dimensions other than time (product group, region, supplier, etc).
- **Partition by Size of Table:** partition on basis of size (When no clear basis/dimension for partitioning).
  - Set a predetermined size/critical point. When data exceeds predetermined size, partition is created.
- **Partitioning Dimensions:** If a dimension contains large number of entries, then partition the dimensions.

STEPS IN DA - DATA PREPARATION

Vertical Partition

- Splits data vertically.
- Each partition holds subset of fields.
- **Normalization**
  - Standard relational method of database organization.
  - removing redundancies from database by splitting tables and linking them with foreign key.

Functional partitioning

- Data is aggregated according to how it is used by each bounded context in the system.
- *Example, e-commerce system might store invoice data in one partition and product inventory data in another.*