Learn Something Every Day

Data Preparation

The data collected from the respondents is generally not in the form to be analyzed directly. After the responses are recorded or received, the next stage is that of **preparation of data** i.e. to make the data amenable for appropriate analysis.

Data preparation **includes editing, coding, and data entry** and is the activity that **ensures the accuracy of the data** and their **conversion from raw form** to reduced and classified forms that are more appropriate for analysis. Preparing a descriptive statistical summary is another preliminary step leading to an understanding of the collected data

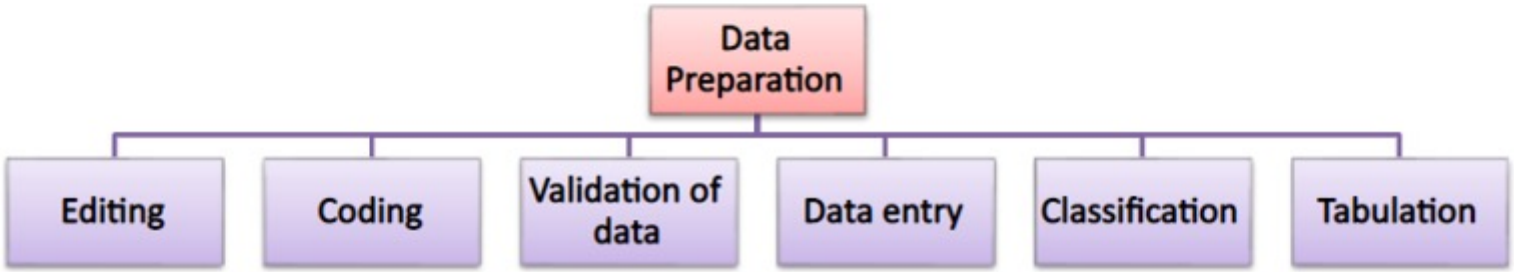
Why Data Preprocessing?

Data in the real world is dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- noisy: containing errors or outliers
- inconsistent: containing discrepancies in codes or names

No quality data, no quality mining results!

- Quality decisions must be based on quality data
- Data warehouse needs consistent integration of quality data



Major Tasks in Data Preprocessing

Data cleaning

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

Data integration

Integration of multiple databases, data cubes, or files

Data transformation

Normalization and aggregation

Data reduction

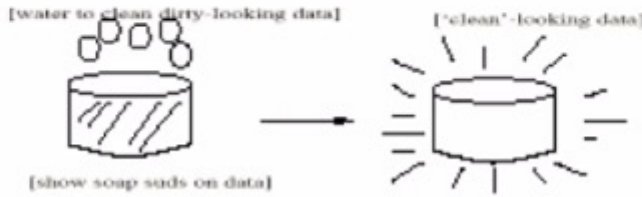
Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

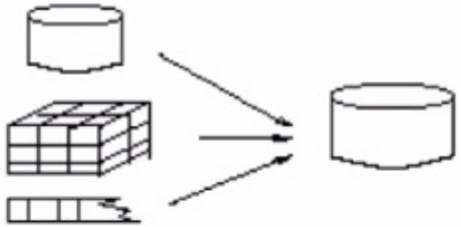
Part of data reduction but with particular importance, especially for numerical data

Forms of data preprocessing

Data Cleaning



Data Integration



Data Transformation

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

Data Reduction

Data Cleaning

Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

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.

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.

How to Handle Missing Data?

Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably)

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 most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Noisy Data

Noise: random error or variance in a measured variable

Incorrect attribute values may 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

How to Handle Noisy Data?

Binning method:

- first sort data and partition into (equi-depth) bins
- then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Clustering

- detect and remove outliers

Combined computer and human inspection

- detect suspicious values and check by human

Regression

- smooth by fitting the data into regression functions

Simple Discretization Methods: Binning

Equal-width (distance) partitioning:

- It divides the range into *N* intervals of equal size: uniform grid
- if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
- The most straightforward
- But outliers may dominate presentation
- Skewed data is not handled well.

Equal-depth (frequency) partitioning:

- It divides the range into *N* intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

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, e.g., A.cust-id ≡ B.cust-#

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

Handling Redundant Data

Redundant data occur often when integration of multiple databases

- The same attribute may have different names in different databasesCareful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

Smoothing: remove noise from data

Aggregation: summarization, data cube construction

Generalization: concept hierarchy climbing

Normalization: scaled to fall within a small, specified range

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Data Reduction Strategies

Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

Data reduction

- Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Data reduction strategies

- Data cube aggregation
- Dimensionality reduction
- Numerosity reduction
- Discretization and concept hierarchy generation

Data Cube Aggregation

The lowest level of a data cube

- the aggregated data for an individual entity of interest
- e.g., a customer in a phone calling data warehouse.

Multiple levels of aggregation in data cubes

- Further reduce the size of data to deal with

Reference appropriate levels

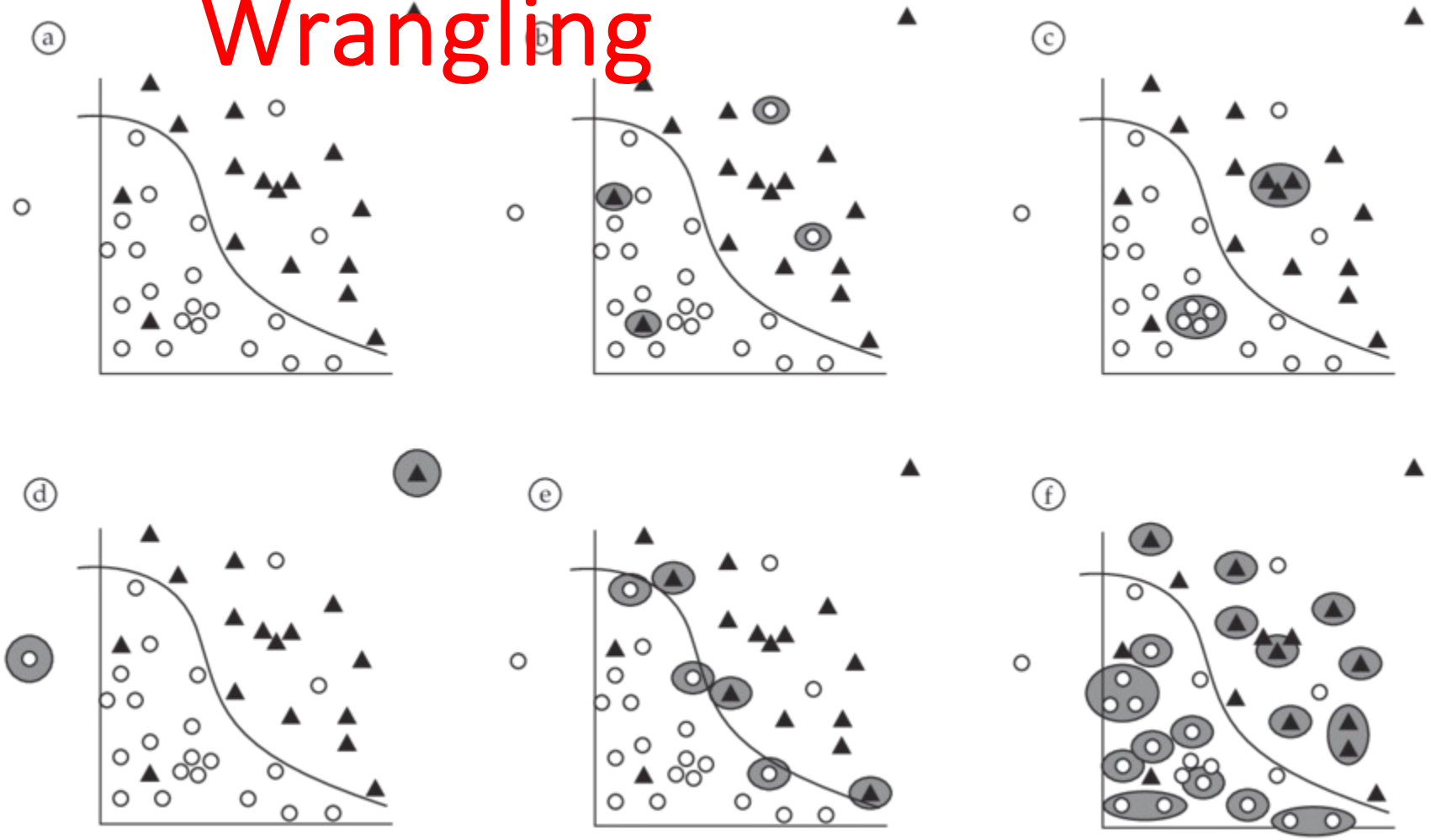
- Use the smallest representation which is enough to solve the task

Data Wrangling

- What Is Data Wrangling?
- Importance of Data Wrangling
- Benefits of Data Wrangling
- Data Wrangling Tools
- Data Wrangling Examples

Data Wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze.

Need of Data Wrangling

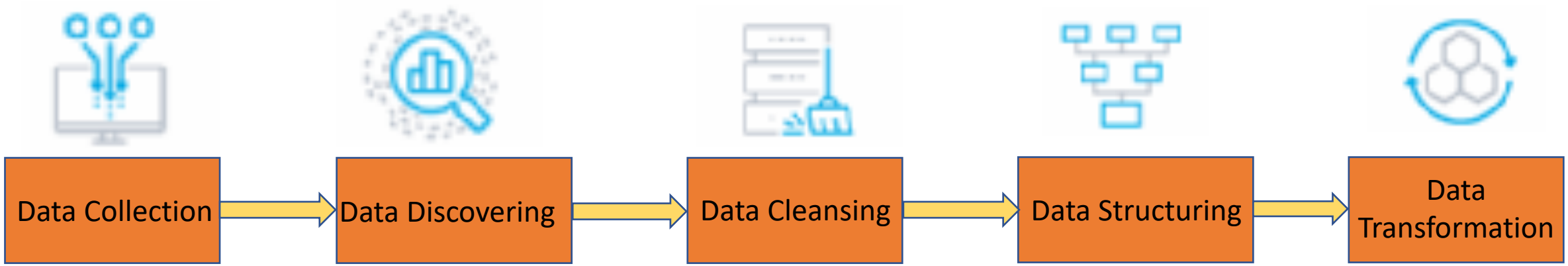


Importance of Data Wrangling

- Making raw data usable. Accurately wrangled data guarantees that quality data is entered into the downstream analysis.
- Getting all data from various sources into a centralized location so it can be used.
- Piecing together raw data according to the required format and understanding the business context of data
- Automated data integration tools are used as data wrangling techniques that clean and convert source data into a standard format that can be used repeatedly according to end requirements. Businesses use this standardized data to perform crucial, cross-data set analytics.
- Cleansing the data from the noise or flawed, missing elements
- Data wrangling acts as a preparation stage for the data mining process, which involves gathering data and making sense of it.
- Helping business users make concrete, timely decisions

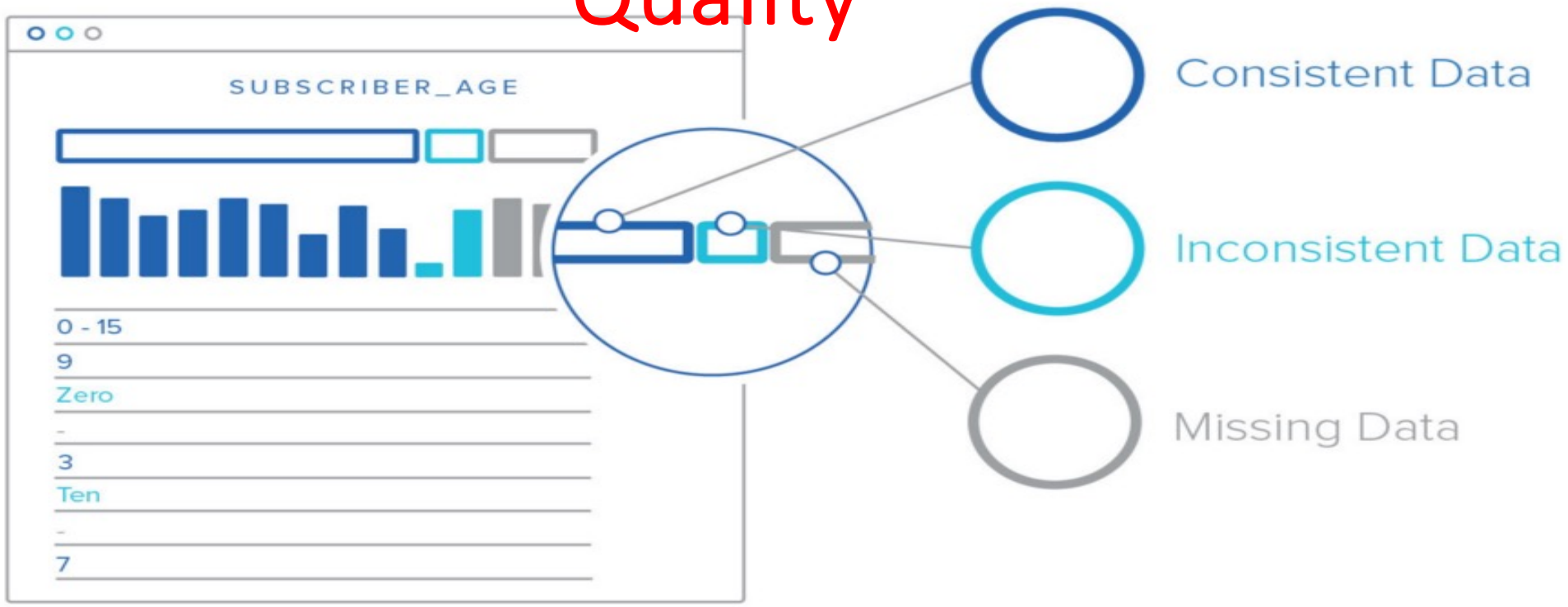
Data Preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis.



DATA COLLECTION

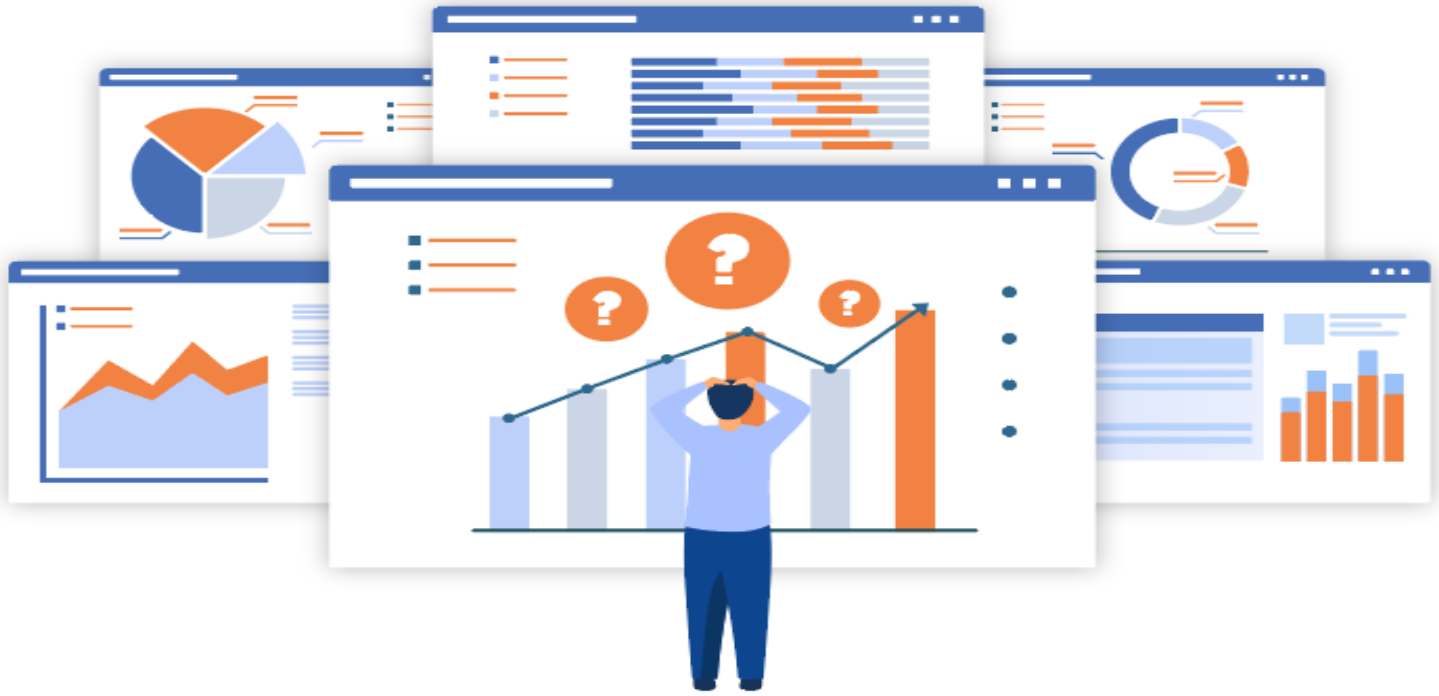
Data Quality



Data Quality includes examining data accuracy, consistency, completeness, and relevance.

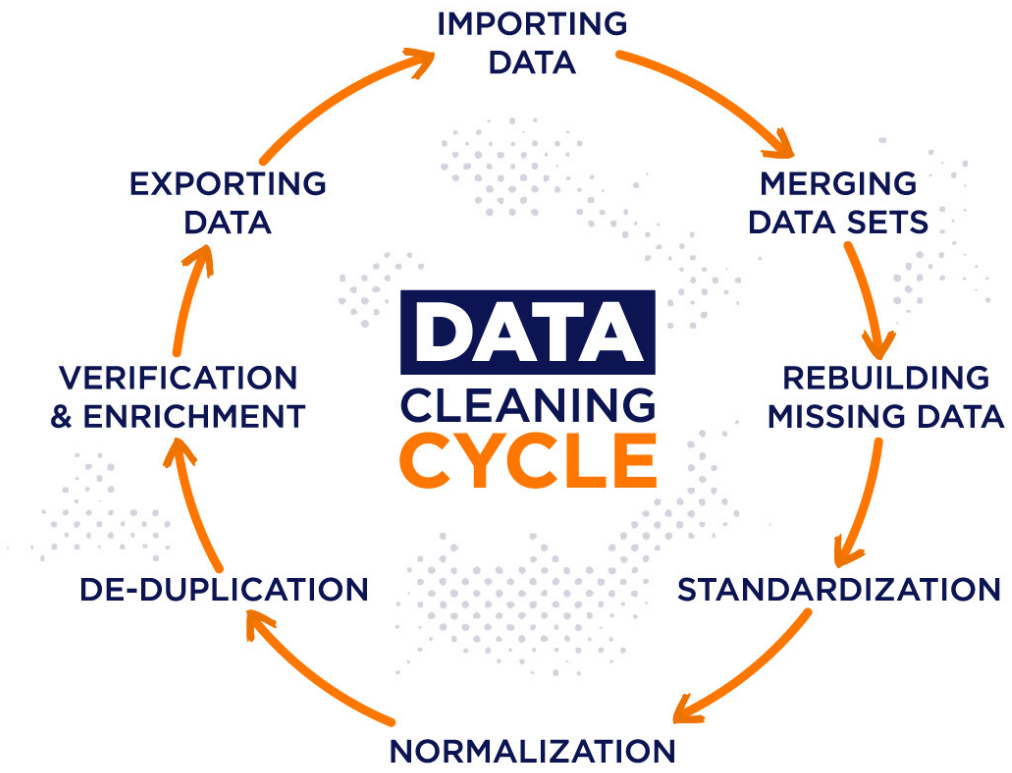
Data Discovery

The second step in the Data Wrangling process is **Discovery.** This is an all-encompassing term for understanding or getting familiar with your data.



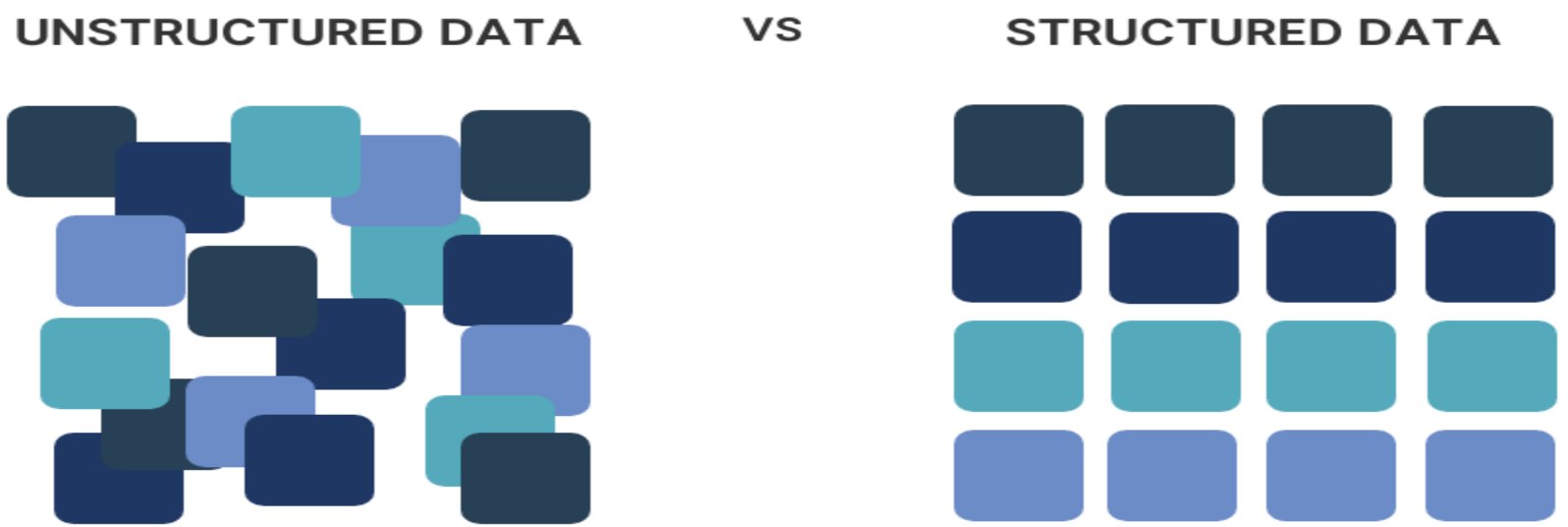
Data Cleaning

Data Cleaning involves Tackling Outliers, Making Corrections, Deleting Bad Data completely, etc. This is done by applying algorithms to tidy up and sanitize the dataset.



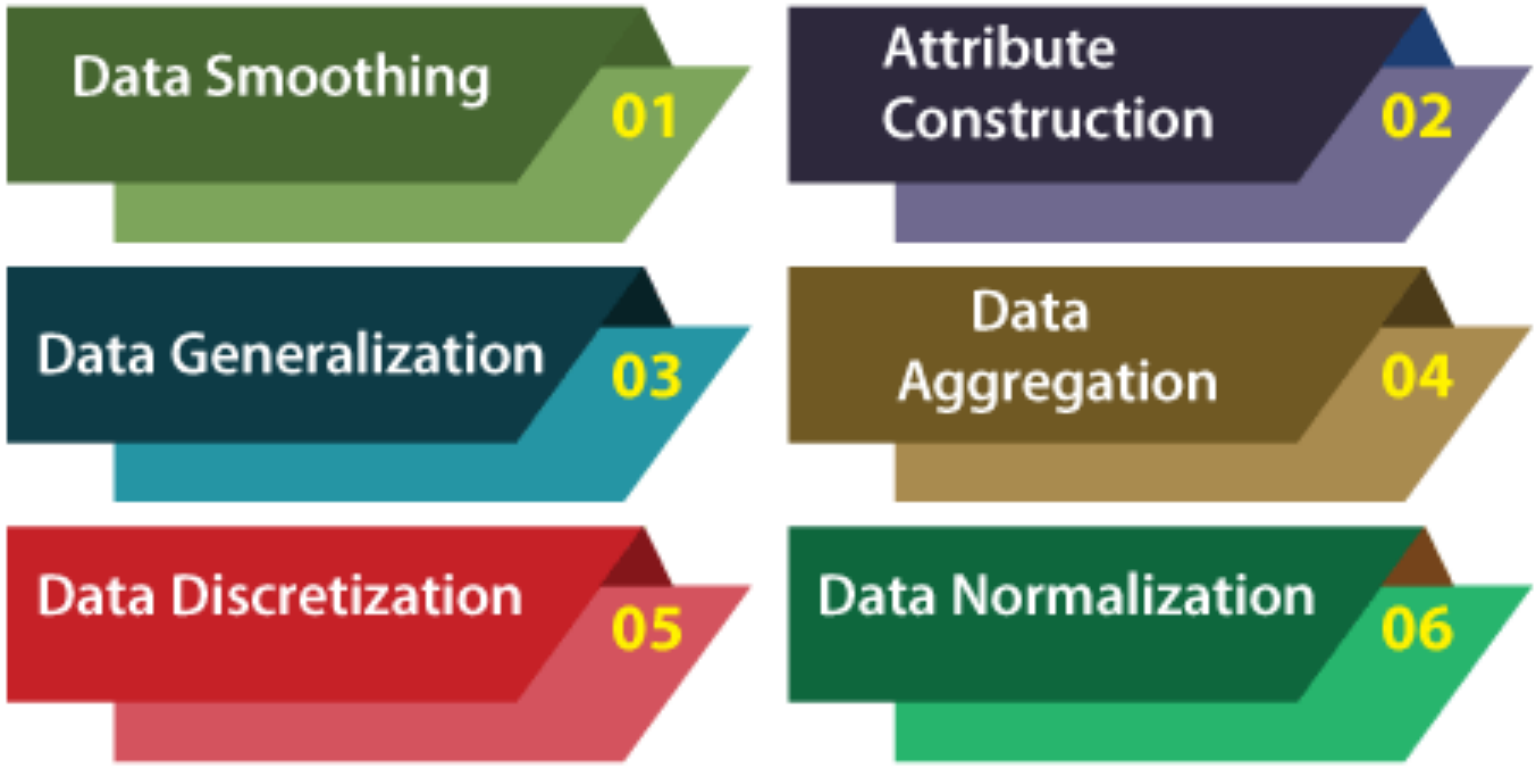
Data Structuring

When raw data is collected, it's in a wide range of formats and sizes. It has no definite structure, which means that it lacks an existing model and is completely disorganized. It needs to be restructured to fit in with the **Analytical Model** deployed by your business, and giving it a structure allows for better analysis.



Structured data	Unstructured data
Structured data stands for information that is highly organized, factual, and to-the-point.	Unstructured data doesn't have any predefined structure to it and comes in all its diversity of forms.
Quantitative	Qualitative
Data warehouses Relational databases	Data lakes Non-relational databases
Several predetermined formats	A huge array of formats

Data Transformation



THANK YOU!