Data Preparation

Data Issues

- Quality
- Accuracy
- Consistency
- Irrelevant data

Data Preparation

- Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence (BI), analytics and data visualization applications.

- The components of data preparation include data preprocessing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources.

Why Data Preparation

- There are several reasons why we need to prepare the data.
  - By preparing data, we actually prepare the miner so that when using prepared data, the miner produces better models faster.
  - Good data is essential for producing efficient models of any type.
  - Data should be formatted according to required software tool.
  - Data need to be made adequate for given method.

Benefits of Data Preparation

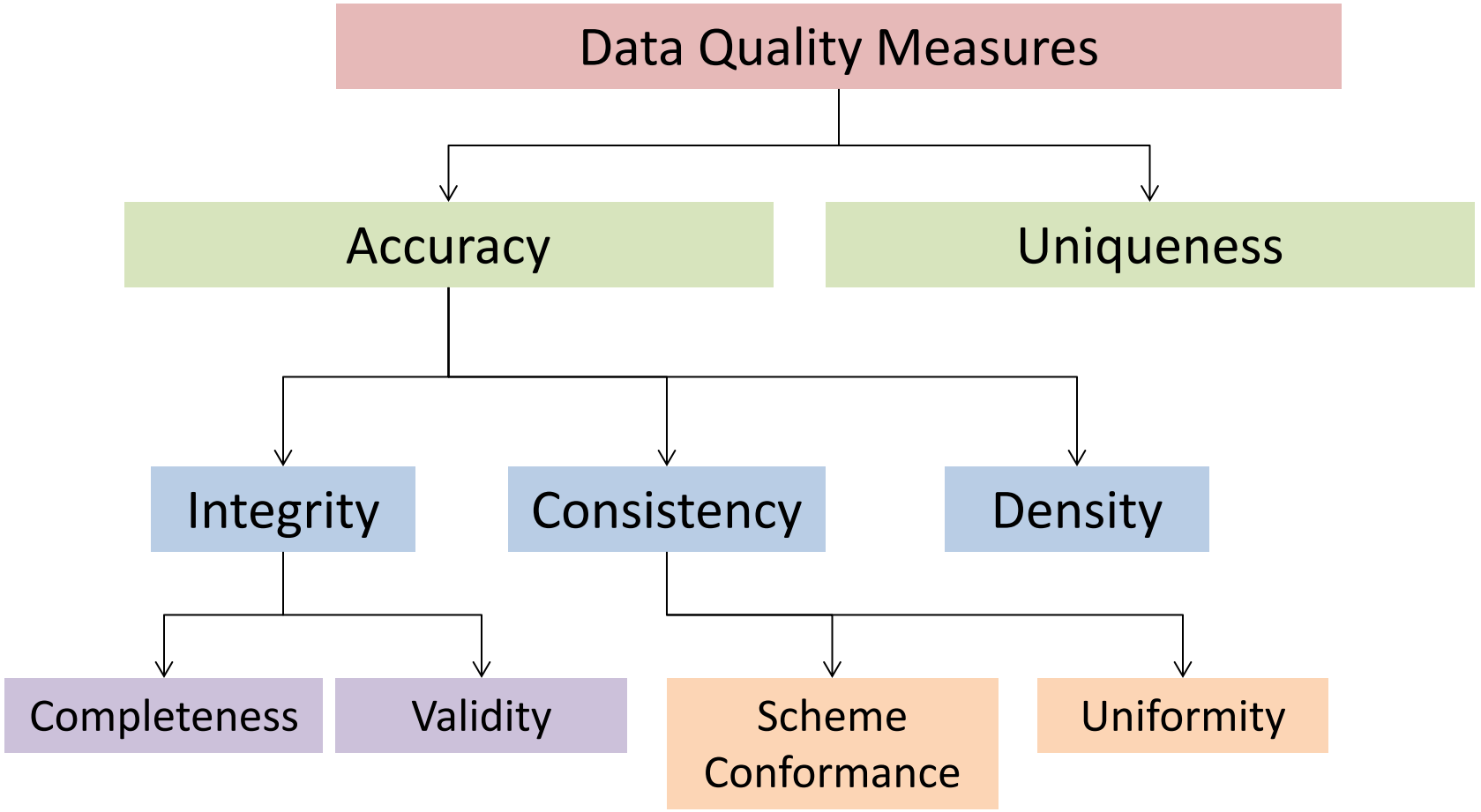
Data preparation helps:

- **Fix errors quickly** — Data preparation helps catch errors before processing. After data has been removed from its original source, these errors become more difficult to understand and correct.
- **Produce top-quality data** — Cleaning and reformatting datasets ensures that all data used in analysis will be high quality.
- **Make better business decisions** — higher quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient and high quality business decisions.

Data preparation steps

1) Data Profiling
2) Data discretization
3) Data cleaning
4) Data integration
5) Data transformation
6) Data reduction

Data Profiling: Sourcing, selecting and auditing appropriate data



Data Profiling: Sourcing, selecting and auditing appropriate data

- Assuring and improving data quality are two of the primary reasons for data preprocessing.

- There are common criteria to measure and evaluate the quality of data, which can be categorized into two main elements; accuracy and uniqueness.

Data Profiling: Sourcing, selecting and auditing appropriate data

- **Accuracy** is described as an aggregated value over the quality criteria: Integrity, Consistency, and Density.

- Intuitively this describes the extent to which the data are an exact, uniform and complete representation of the *mini-world: the aspects of the world that the data describe.*

Data Profiling: Sourcing, selecting and auditing appropriate data

- **Integrity**: An integral data collection contains representations of all the entities in the mini-world and only of those.
- Access data from any source — no matter the origin, format or narrative and integrating them together. Increased access to data means less manual work, faster insights and faster time to value realized by your organization.
- Integrity requires both completeness and validity.

Data Profiling: Sourcing, selecting and auditing appropriate data

- **Completeness**: Complete data give a comprehensive representation of the mini-world and contain no missing values.
- We achieve completeness within data cleansing by correcting anomalies and not just deleting them.
- It is also possible that additional data are generated, representing existing entities that are currently unrepresented in the data.
- A problem with assessing completeness is that *you don't know what you don't know*. As a result, there are no known gold standard data, which can be used as a reference to measure completeness.

Data Profiling: Sourcing, selecting and auditing appropriate data

- **Validity:** Data are valid when there are no constraints violated.
- There are numerous mechanisms to increase validity including mandatory fields, enforcing unique values, and data schema/structure.

Data Profiling: Sourcing, selecting and auditing appropriate data

- **Consistency:** This quality concerns syntactic anomalies as well as contradictions. The main challenge concerning data consistency is choosing which data source you trust for reliable agreement among data across different sources.
  - **Schema conformance:** This is especially true for the relational database systems where the adherence of domain formats relies on the user.
  - **Uniformity:** is directly related to irregularities.

Data Profiling: Sourcing, selecting and auditing appropriate data

- **Density:** This criterion concerns the quotient of missing values in the data. There still can be non-existent values or properties that have to be represented by null values having the exact meaning of not being known.

The above three criteria of Integrity, Consistency, and Density collectively represent the accuracy measure.

Data Cleansing

- Where the data contain noise or anomalies it may be desirable to identify and remove outliers and other suspect data points, or take other remedial action.
- Data cleansing is defined as the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Data cleansing can also be referred to as data cleaning, data scrubbing, or data reconciliation.

Data Cleansing

- More precisely, the process of data cleansing could be explained as a four-stage process:
  - Define and identify errors in data such as incompleteness, incorrectness, inaccuracy or irrelevancy.
  - Clean and rectify these errors by replacing, modifying, or deleting them
  - Document error instances and error types; and finally
  - Measure and verify to see whether the cleansing meets the user's specified tolerance limits in terms of cleanliness.

Data anomalies

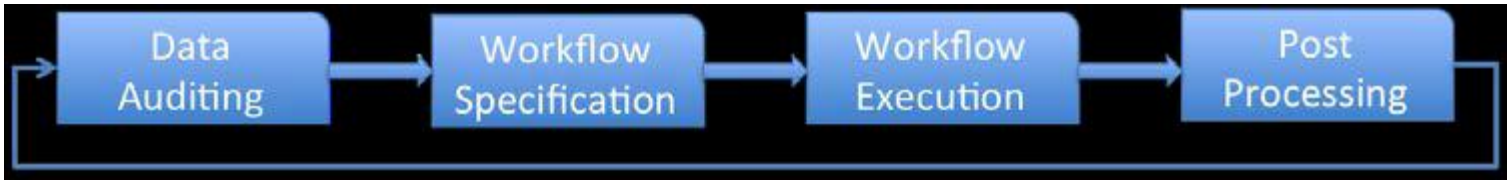
- The term data anomaly describes any distortion of data resulting from the data collection process.
- From this perspective, anomalies include duplication, inconsistency, missing values, outliers, noisy data or any kind of distortion that can cause data imperfections.

Data anomalies

Anomalies can be classified at a high level into three categories:

- **Syntactic Anomalies:** describe characteristics concerning the format and values used for the representation of the entities. Syntactic anomalies include lexical errors, domain format errors, syntactical errors, and irregularities.
- **Semantic Anomalies**: hinder the data collection from being a comprehensive and non-redundant representation of the mini-world. These types of anomalies include integrity constraint violations, contradictions, duplicates and invalid tuples.
- **Coverage Anomalies:** decrease the number of entities and entity properties from the mini-world that is represented in the data collection. Coverage anomalies are categorized as missing values and missing tuples.

Data cleansing process



1. **Data Auditing**: This first step mainly identifies the types of anomalies that reduce data quality. Data auditing checks the data using validation rules that are pre-specified, and then creates a report of the quality of the data and its problems. We often apply some statistical tests in this step for examining the data.
2. **Workflow specification**: The next step is to detect and eliminate anomalies by a sequence of operations on the data. The information collected from data auditing is then used to create a data-cleaning plan. It identifies the causes of the dirty data and plans steps to resolve them.
3. **Workflow execution**: The data cleaning plan is executed, applying a variety of methods on the data set.
4. **Post-processing and controlling**: The post-processing or control step involves examination of the workflow results and performs exception handling for the data mishandled by the workflow.

Dealing with Missing values

- One major task in data cleansing is dealing with missing values. It is important to determine whether the data have missing values and, if so, to ensure that appropriate measures are taken to allow the learning system to handle this situation.
- Handling data that contain missing values is crucial for the data cleansing process and data wrangling in general. In real-life data, most of existing data sets contain missing values that were not introduced or were lost in the recording process for many reasons.

Handling outliers

- An outlier is another type of data anomaly that requires attention in the cleansing process. Outliers are data that do not conform to the overall data distribution.
- Outliers can be seen from two different perspectives; first, they might be seen as glitches in the data. Alternatively, they might be also seen as interesting elements that could potentially represent significant elements in the data.

Data Enrichment/Integration

- Existing data may be augmented through data enrichment. This commonly involves sourcing of additional information about the data points on which data are already held. For example, customer data might be enriched by obtaining socio-economic data about individual customers.
- Data integration is a crucial task in data preparation. Combining data from different sources is not trivial especially when dealing with large amounts of data and heterogeneous sources. Data are typically presented in different forms (structured, semi- structured or unstructured) as well as from different sources (web, database) that could be stored locally or distributed.

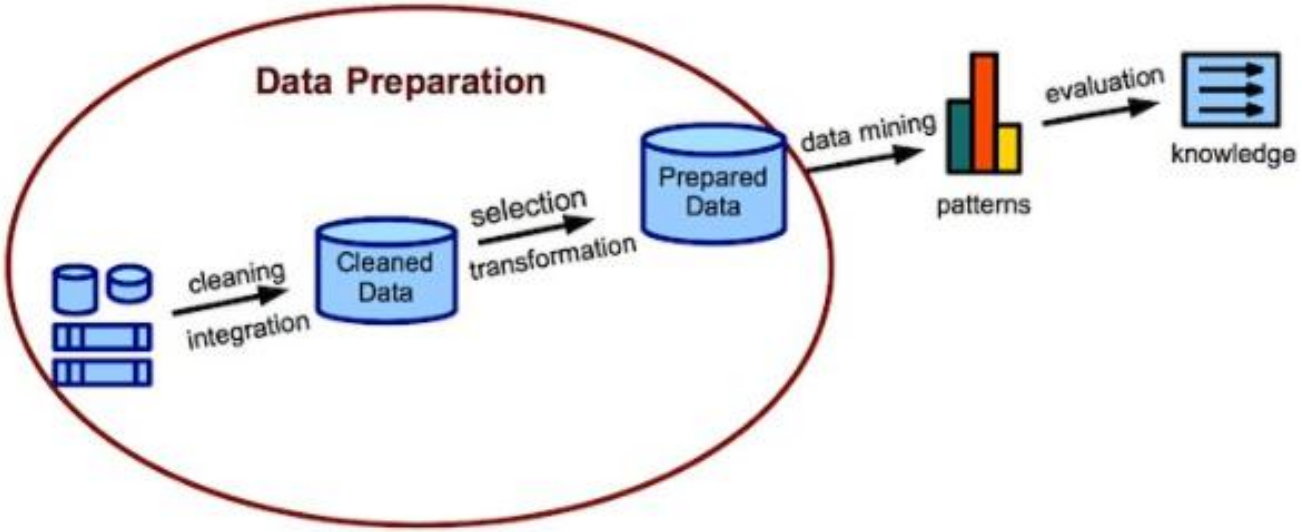
Data Transformation

- It is frequently necessary to transform data from one representation to another. There are many reasons for changing representations:
  - To generate symmetric distributions instead of the original skewed distributions:
  - Transformation improves visualisation of data that might be tightly clustered relative to a few outliers
  - Data are transformed to achieve better interpretability.
  - Transformations are often used to improve the compatibility of the data with assumptions underlying a modelling process, for example, to linearize (straighten) the relation between two variables whose relationship is non-linear. Some of the data mining algorithms require the relationship between data to be linear.

Discretization

- Discretization transforms continuous data into a discrete form. This is useful in many cases for: better data representation, data volume reduction, better data visualization and representing data at a various level of granularity for data analysis. Data discretization approaches are categorized as supervised, unsupervised, bottom-up or top down. Approaches for data discretization include Binning, Entropy based, Nominal to numeric, 3-4-5 rule and Concept hierarchy.

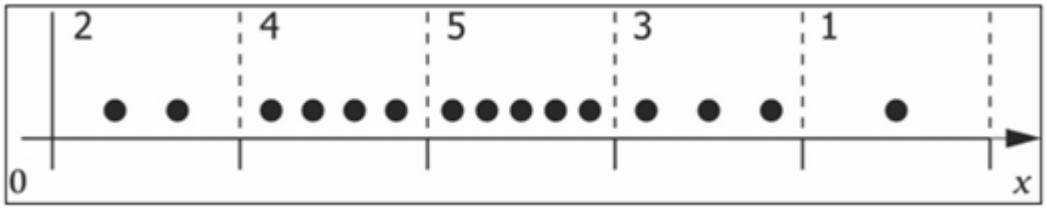
Data preparation example



- There are multiple values that are commonly used to represent the virus. A virus like COVID-19 could be represented by ‘SAR-Cov2’, ‘Corona’, ‘Covid’ or ‘Covid-19’ to name a few.
- A data preparation tool could be used in this scenario to identify an incorrect number of unique values (in the case of virus, a unique count greater than suitable number in case of covid would raise a flag, as there are only few names aligned with virus). These values would then need to be standardized to use only an abbreviation or only full spelling in every row.

Data binning

- Data binning, bucketing is a data pre-processing method used to minimize the effects of small observation errors. The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin. This has a smoothing effect on the input data and may also reduce the chances of over fitting in case of small datasets.



- There are 2 methods of dividing data into bins.
- **Equal Frequency Binning:** bins have equal frequency.
- **Equal Width Binning:** bins have equal width with a range of each bin are defined as [min + w], [min + 2w] …. [min + nw] where w = (max — min) / (no of bins).

Importance of Data Binning: -

- Binning is used for reducing the cardinality of continuous and discrete data.
- Binning groups related values together in bins to reduce the number of distinct values.
- Binning can improve resource utilization and model build response time dramatically without significant loss in model quality.
- Binning can improve model quality by strengthening the relationship between attributes.
- Supervised binning is a form of intelligent binning in which important characteristics of the data are used to determine the bin boundaries.
- In supervised binning, the bin boundaries are identified by a single-predictor decision tree that takes into account the joint distribution with the target. Supervised binning can be used for both numerical and categorical attributes.

Advantages (Pros) of data smoothing

- Data smoothing clears the understandability of different important hidden patterns in the data set.
- Data smoothing can be used to help predict trends. Prediction is very helpful for getting the right decisions at the right time.
- Data smoothing helps in getting accurate results from the data.

Cons of data smoothing

- Data smoothing doesn’t always provide a clear explanation of the patterns among the data.
- It is possible that certain data points being ignored by focusing the other data points.