Universiti Malaysia PAHANG
Engineering • Technology • Creativity

CHAPTER 2
Data Preprocessing

DATA MINING (BSD3533)

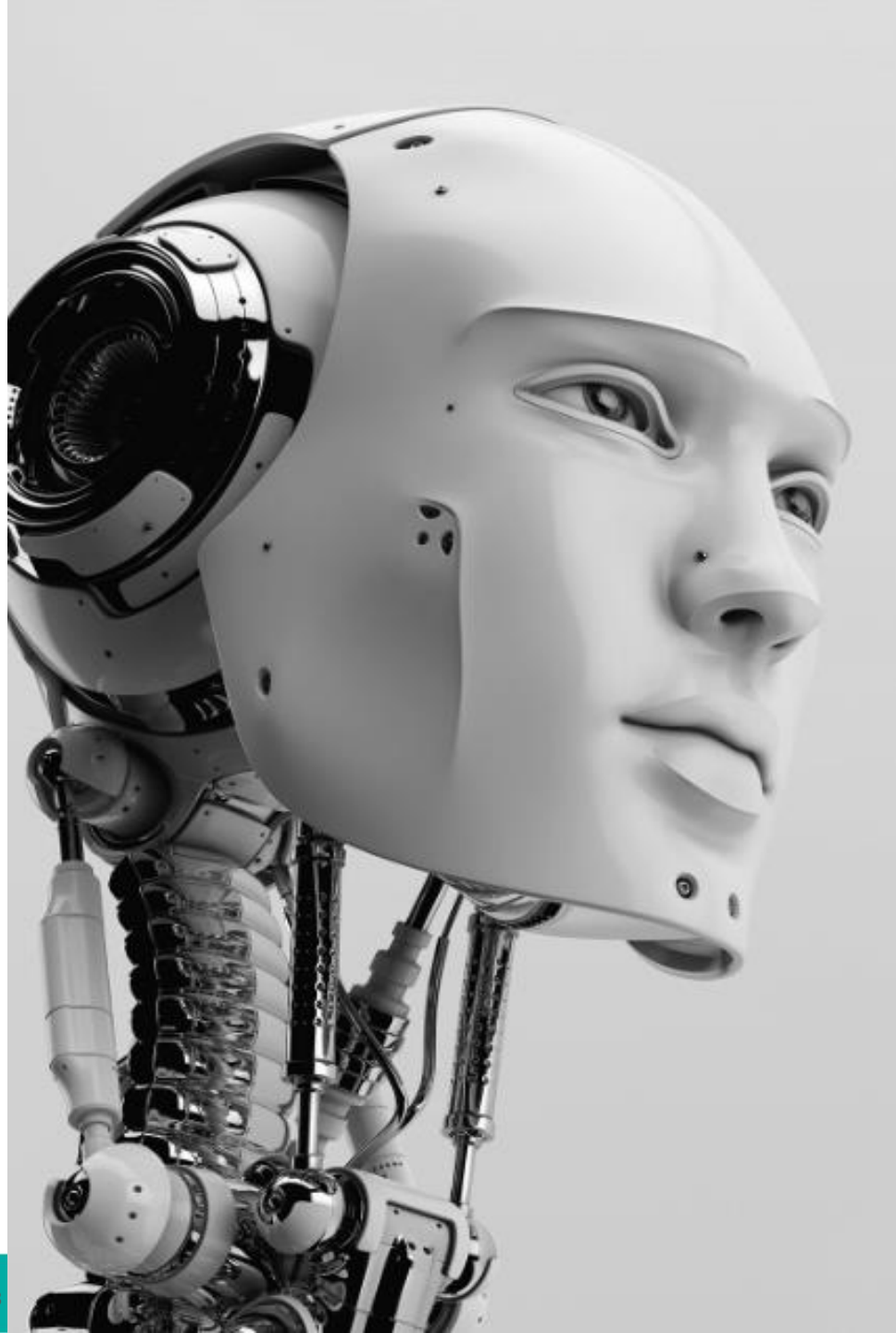
DR. KU MUHAMMAD NA'IM KU KHALIF

Content

Chapter 2.1: Data Access

By the end of this topic, you should be able to:

- understand the concept of data processing and how it's applied in the real world.
- understand the needs and why to study data preprocessing.
- acquire the data access knowledge to read data from a different format.



Need for Data Preprocessing

- Data is one of the most significant considerations for every data analyst. As a matter of fact, the representation and quality of the data employed in analysis is the first and main concern of every analyst. "Garbage in, garbage out" is a frequent idiom in the area of data mining and machine learning when working with enormous quantities of data.
- Typically, we end up with a great deal of noisy data; for instance, income: -400, or negative income. Sometimes, we may have ridiculous and impossible data combinations, such as Gender-Male as Pregnant-Yes in a record. Clearly ludicrous! because guys do not get pregnant.
- Additionally, we incur losses owing to missing numbers and other data anomalies. Analyzing such data sets, which have not been pre-screened, can produce false results.

- Consequently, data preprocessing is the initial stage in any data mining procedure. Data preprocessing is a data mining approach that involves transforming raw data into a comprehensible format, as data from the actual world are frequently incomplete, inconsistent, or even incorrect. Data preprocessing eliminates such problems. It guarantees that subsequent data mining processes are error-free. It is a precondition for data mining and prepares the raw data for the core procedures.
- Data comes from multitude of sources; it can be high in volume and have variety of attributes. Real-world data is generally noisy, incomplete and inconsistent. It implies that raw data tends to be corrupt, have missing values or attributes, outliers or conflicting values. Data preprocessing stage resolves such kinds of data issues to ensure the dataset used for modelling stage is acceptable and of improved quality.
- Analytical models fed with poor quality data can lead to misleading predictions.

Dirty Data

Dirty data refers to any data that takes away the data integrity of the entire dataset. Below are some examples.

- ✓Data errors include misspelt data, typos, duplicate data, and erroneously parsed data.
- ✓Data that violate business rules may not be easily fixed even if it is identified. More often than not, the business needs to review such data.
- ✓Data can be consistently generated by systems that have entity constraint issues, bugs, and legacy “patching” placed inside the systems. Often, this data will look “consistent” with all other data. But, upon close inspection, this data is simply wrong.
- ✓Data can be collected using the wrong method, or the wrong population. This data often comes from asking the wrong business question.
- ✓Data can be calculated using inconsistent codebase, modules, and Application Programming Interface (API). The magnitude of the data errors might not be large for individual transactions.

Is this an acceptable dataset?

Id	Name	DoB	Age	Gender	Phone	Country
1801	Shah Rukh Khan	11/12/1984	34	Male	5551212	India
1802	Roselinda	14/13/1986	32	Female	4568765	Kuala Lumpur
1803	Muhammad Ali Jannah	31/08/1983	35	M	5678900	Jordan
1804	Lynda Carter	12/30/1980	38	Female	9999999	America
1805	Smith, Tracy	23/08/1981	37	2	6856262	UK
1806	Ng Chee Chin	3/10/1989	19	Fenale	3209876	Malaysia
1807	Fatoush Olkan	18/07/1982	36	-	2348765	Turkey
1808	John Doe	20/11/1987	31	Male	7735075	USA
1809	Tracy Smith	23/08/1981	37	2	8356753	UK
1809	Ibrar Yaacob	18/09/1974	44	Male	6544321	Pakistan

Issues to highlight

Id	Name	DoB	Age	Gender	Phone	Country
1801	Shah Rukh Khan	11/12/1984	34	Male	5551212	India
1802	Roselinda	14/13/1986	32	Female	4568765	Kuala Lumpur
1803	Muhammad Ali Jannah	31/08/1983	35	M	5678900	Jordan
1804	Lynda Carter	12/30/1980	38	Female	9999999	America
1805	Smith, Tracy	23/08/1981	37	2	6856262	UK
1806	Ng Chee Chin	3/10/1989	19	Fenale	3209876	Malaysia
1807	Fatoush Olkan	18/07/1982	36	-	2348765	Turkey
1808	John Doe	20/11/1987	31	Male	7735075	USA
1809	Tracy Smith	23/08/1981	37	2	8356753	UK
1809	Ibrar Yaacob	18/09/1974	44	Male	6544321	Pakistan

There are some issues raise here:

1. DoB column, there are unacceptable for 14/13/1986 and 12/30/1980.
2. Gender column, doesn't synchronously written for gender type.
3. Country column, Kuala Lumpur doesn't correctly represent the Country.

Answer: Unacceptable

Common Types of Dirty Data

a. **Incomplete data:** Most common occurrence of dirty data. Important fields on master data records, useful to the business, are often left blank. For example, if you haven't classified your customers by industry, you cannot segment your sales and marketing initiatives by industry.

b. **Duplicate data:** Very common. Most companies deal with issues such as duplicate customer records, but duplicate materials are also very common. This can be costly to companies due to excess in inventory and sub-optimal procurement decisions.

c. **Incorrect data:** Incorrect data can occur when field values are created outside of the valid range of values. For example, the value in a month field should range from 1 to 12 or a street address should be a real address.

Common Types of Dirty Data

d. **Inaccurate data:** It is possible or data to be technically correct but inaccurate given the business context. Costly business interruptions are often rooted in inaccurate data. For example, minor errors in customer addresses can result in deliveries at wrong locations even though the addresses are actual addresses.

e. **Business rule violations:** There are often large collections of poorly documented business rules associated with master data that are specific to the industry or business context. For example, beverage products should have a Unit of Measure in ‘fl. oz.’ or payment terms for a certain type of customers should always be ‘Net 30.’

f. **Inconsistent data:** Data redundancy–i.e., the same field values stored in different places-often leads to inconsistencies. For example, most companies have customer information in multiple systems and the data is often not kept in sync.

Data Access

- ✓ Dealing with data science project, you've probably spent a lot of time browsing the internet looking for interesting data sets to analyze.
- ✓ It can be fun to sift through dozens of data sets to find the perfect one, but it can also be frustrating to download and import several csv files, only to realize that the data isn't that interesting after all.
- ✓ Luckily, there are online repositories that curate data sets and (mostly) remove the uninteresting ones.

Data Repositories for Data Science Project

Kaggle

Kaggle is a data science community that hosts machine learning competitions. There are a variety of externally-contributed interesting data sets on the site. Kaggle has both live and historical competitions. You can download data for either, but you have to sign up for Kaggle and accept the terms of service for the competition.

UCI Machine Learning Repository

It is one of the oldest sources of data sets on the web. Although the data sets are user-contributed, and thus have varying levels of documentation and cleanliness, the vast majority are clean and ready for machine learning to be applied. UCI is a great first stop when looking for interesting data sets.

data.world

describes itself at ‘the social network for data people’, but could be more correctly describe as ‘GitHub for data’. It’s a place where you can search for, copy, analyze, and download data sets. In addition, you can upload your data to data.world and use it to collaborate with others.

data.gov.us

It is a relatively new site that’s part of a US effort towards open government. Data.gov makes it possible to download data from multiple US government agencies. Data can range from government budgets to school performance scores. Much of the data requires additional research, and it can sometimes be hard to figure out which data set is the “correct” version. Anyone can download the data, although some data sets require additional hoops to be jumped through, like agreeing to licensing agreements.

Github

Github has an API that allows you to access repository activity and code. You can get started with the API here. The options are endless — you could build a system to automatically score code quality, or figure out how code evolves over time in large projects.

Quandl

Quandl is a repository of economic and financial data. Some of this information is free, but many data sets require purchase. Quandl is useful for building models to predict economic indicators or stock prices. Due to the large amount of available data sets, building a complex model that uses many data sets to predict values in another is possible.

data.gov.my

One-stop centre to browse Malaysia's wealth of open data. Whether you're a regular citizen looking for information, a researcher looking for material, or an app developer looking for an API, we've got you covered.

AWS Public Data sets

Amazon makes large data sets available on its Amazon Web Services platform. You can download the data and work with it on your own computer, or analyze the data in the cloud using EC2 and Hadoop via EMR.

Amazon has a page that lists all of the data sets for you to browse. You'll need an AWS account, although Amazon gives you a free access tier for new accounts that will enable you to explore the data without being charged.

Google Public Data sets

Much like Amazon, Google also has a cloud hosting service, called Google Cloud Platform. With GCP, you can use a tool called BigQuery to explore large data sets.

Google lists all of the data sets on a page. You'll need to sign up for a GCP account, but the first 1TB of queries you make are free.

File Format

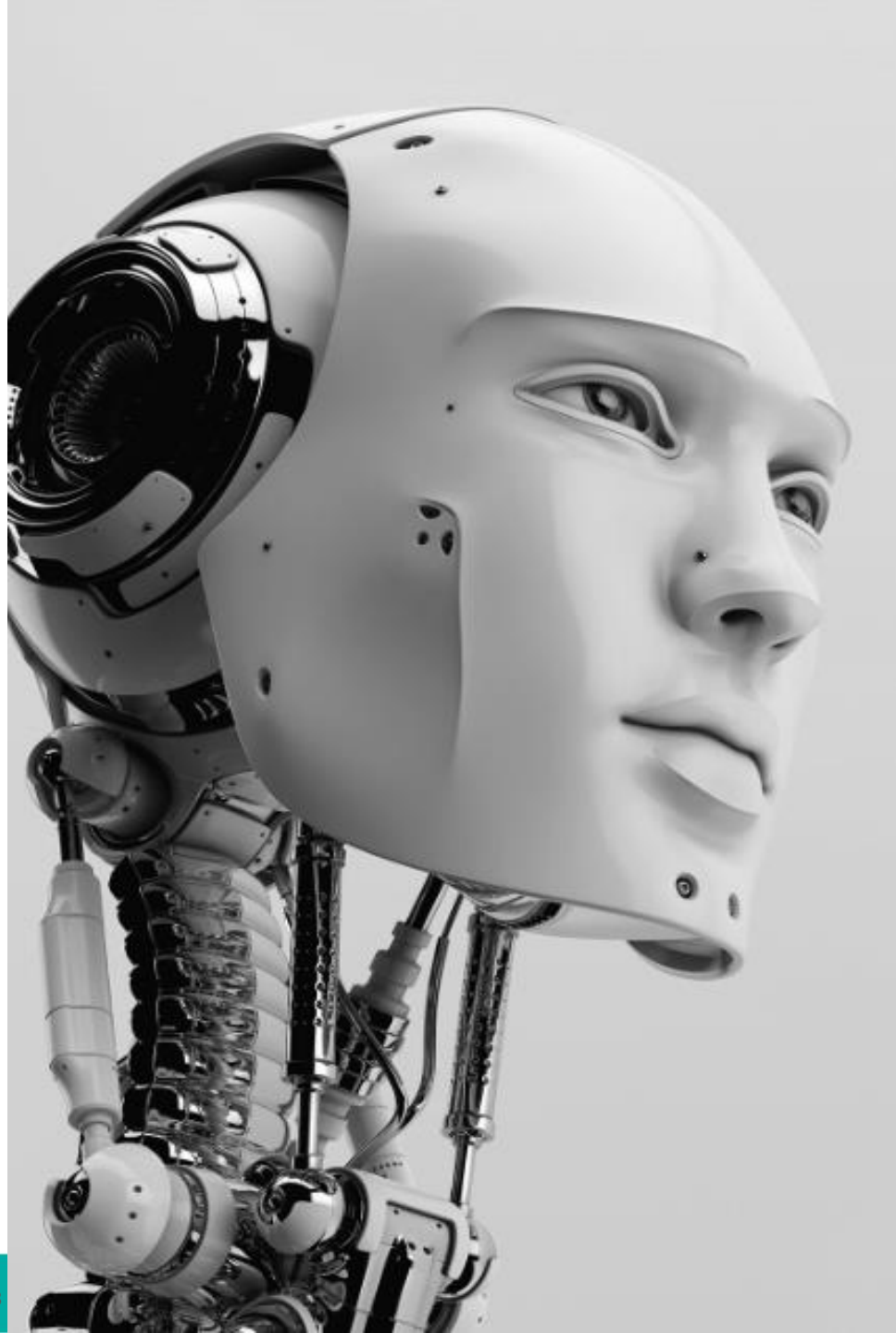
- ✓A file format is a standard way in which information is encoded for storage in a file.
- ✓First, the file format specifies whether the file is a binary or ASCII file. Second, it shows how the information is organized. For example, comma-separated values (CSV) file format stores tabular data in plain text.
- ✓To identify a file format, you can usually look at the file extension to get an idea. For example, a file saved with name “Data” in “CSV” format will appear as “Data.csv”. By noticing “.csv” extension we can clearly identify that it is a “CSV” file and data is stored in a tabular format.
- ✓Usually, the files you will come across will depend on the application you are building. For example, in an image processing system, you need image files as input and output. So you will mostly see files in jpeg, gif or png format.

- ✓As a data scientist, you need to understand the underlying structure of various file formats, their advantages and dis-advantages. Unless you understand the underlying structure of the data, you will not be able to explore it. Also, at times you need to make decisions about how to store data.
- ✓Now, we will look at the following file formats and how to read them in Python:
    a. Comma-separated values
    b. XLSX
    c. ZIP
    d. Plain Text (txt)
    e. JSON
    f. XML
    g. HTML
    h. Images
    i. Hierarchical Data Format
    j. PDF
    k. DOCX
    l. MP3
    m. MP4

Chapter 2.2:
Data Quality from Data Preprocessing

By the end of this topic, you should be able to:

- understand the concepts of data quality in data mining.
- understand the data quality dimension and how to improve data quality in organisation.
- distinguish data quality and data integrity.



What is Data Quality?

- Data quality refers to the development and implementation of operations that use quality management techniques to data to ensure that the data is suitable for serving the unique demands of an organisation in a given context.
- Considered to be of good quality is data that is judged suitable for its intended purpose.
- Examples of data quality difficulties include duplicated data, incomplete data, inconsistent data, erroneous data, poorly specified data, poorly organised data, and inadequate data security.
- Data quality analysts conduct data quality evaluations by evaluating and interpreting each unique data quality measure, aggregating a score for the overall quality of the data, and providing enterprises with a percentage representing the precision of their data.

- A low data quality scorecard implies poor data quality, which is of low value, is misleading and can result in poor decision-making that could be detrimental to the organisation.
- Data quality rules are a vital part of data governance, which is the act of defining and implementing a set of rules and standards by which all data within an organisation is managed.
- Effective data governance should unify data from diverse data sources, define and monitor data usage policies, and eradicate inconsistencies and mistakes that might otherwise have a detrimental influence on the accuracy of data analytics and regulatory compliance.

Data Quality Dimension

- By which metrics do we measure data quality?

a. **Accuracy**: The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.

b. **Completeness**: Completeness is a measure of the data's ability to effectively deliver all the required values that are available.

c. **Consistency**: Data consistency refers to the uniformity of data as it moves across networks and applications. The same data values stored in different locations should not conflict with one another.

d. **Validity**: Data should be collected according to defined business rules and parameters, and should conform to the right format and fall within the right range.

e. **Uniqueness**: Uniqueness ensures there are no duplications or overlapping of values across all data sets. Data cleansing and deduplication can help remedy a low uniqueness score.

f. **Timeliness**: Timely data is data that is available when it is required. Data may be updated in real-time to ensure that it is readily available and accessible.

How to Improve Data Quality?

- Data quality measures can be accomplished with data quality tools, which typically provide data quality management capabilities such as:

  a. **Data profiling** - The first step in the data quality improvement process is understanding your data. Data profiling is the initial assessment of the current state of the data sets.

  b. **Data Standardization** - Dissimilar data sets are conformed to a common data format.

  c. **Geocoding** - The description of a location is transformed into coordinates that conform to U.S. and worldwide geographic standards

  d. **Matching or Linking** - Data matching identifies and merges matching pieces of information in big data sets.

  e. **Data Quality Monitoring** - Frequent data quality checks are essential. Data quality software in combination with machine learning can automatically detect, report, and correct data variations based on predefined business rules and parameters.

  f. **Batch and Real time** - Once the data is initially cleansed, an effective data quality framework should be able to deploy the same rules and processes across all applications and data types at scale.

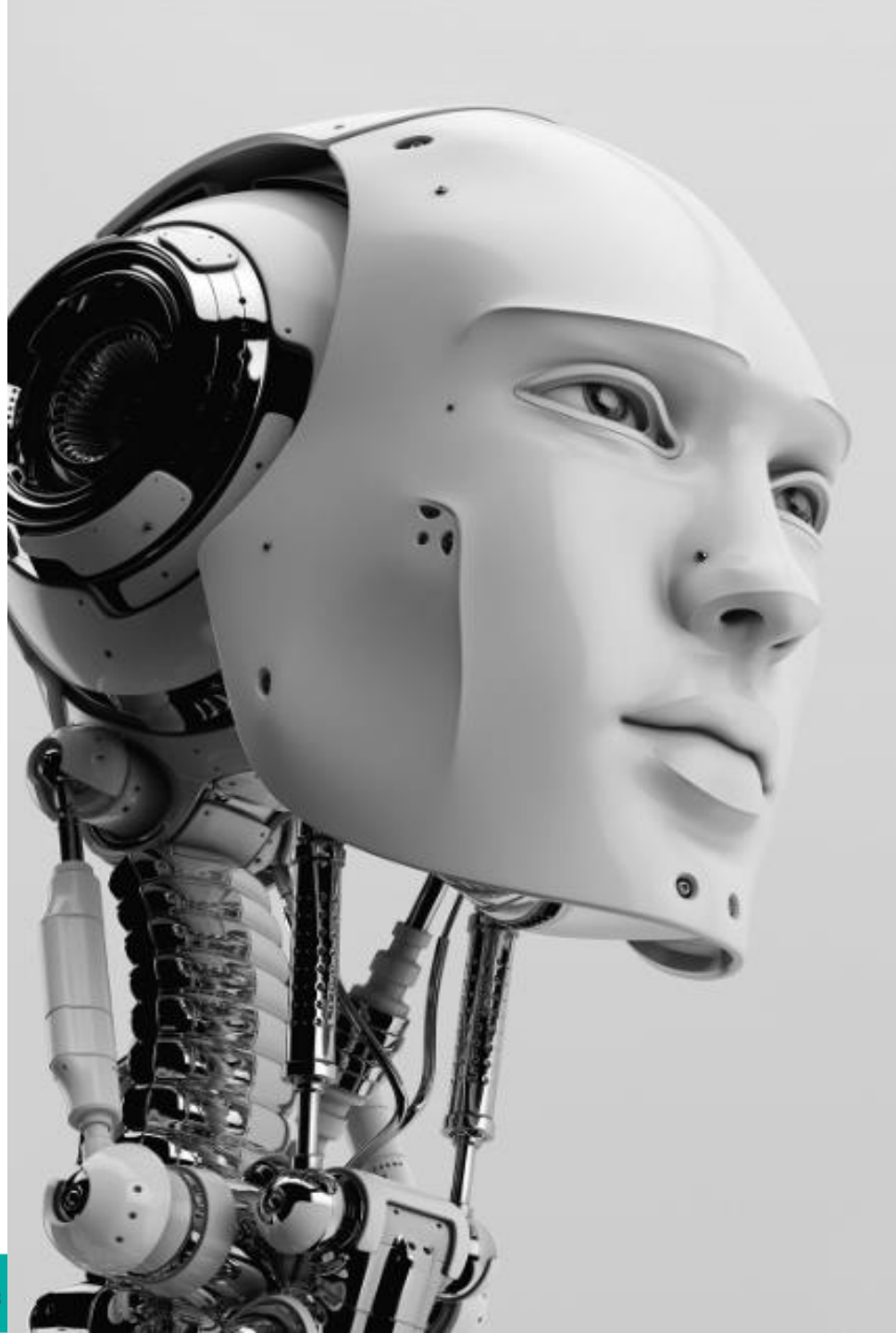
Data Quality vs Data Integrity

- Data quality oversight is just one component of data integrity. Data integrity refers to the process of making data useful to the organization. The four main components of data integrity include:
    a. **Data Integration:** Data from disparate sources must be seamlessly integrated.
    b. **Data Quality:** Data must be complete, unique, valid, timely, consistent, and accurate.
    c. **Location Intelligence:** Location insights add a layer of richness to data and make it more actionable.
    d. **Data Enrichment:** Data enrichment adds a more complete, contextualized view of data by adding data from external sources, such as customer data, business data, location data, etc.

Chapter 2.3:
ETL and ELT Concepts

By the end of this topic, you should be able to:

- understand the concepts of data ETL and ELT.
- implement how to visualise the data interactively using analytics tools.



ETL

- As the volume of data, data sources, and data types at organizations grow, the importance of making use of that data in analytics, data science and machine learning initiatives to derive business insights grows as well.
- The need to prioritize these initiatives puts increasing pressure on the data engineering teams because processing the raw, messy data into clean, fresh, reliable data is a critical step before these initiatives can be pursued.
- ETL, which stands for Extract, Transform, and Load, is the process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.

- ETL tools are used to fetch data from one database and put it into another one after transformation and quality checks.
- The first step called Extraction involves pulling out data from a data source. During this phase, the data is read and gathered, often from numerous and diverse kinds of sources, such as on-premise and cloud databases, enterprise applications, file systems, and more.
- During Transformation, the data extracted is then converted into a format that is acceptable for another database. In this stage, data transformation is done using rules or lookup tables or by merging one data set with another.
- The last step is Loading which is the procedure of writing or stacking the data into the targeted database or data warehouse.

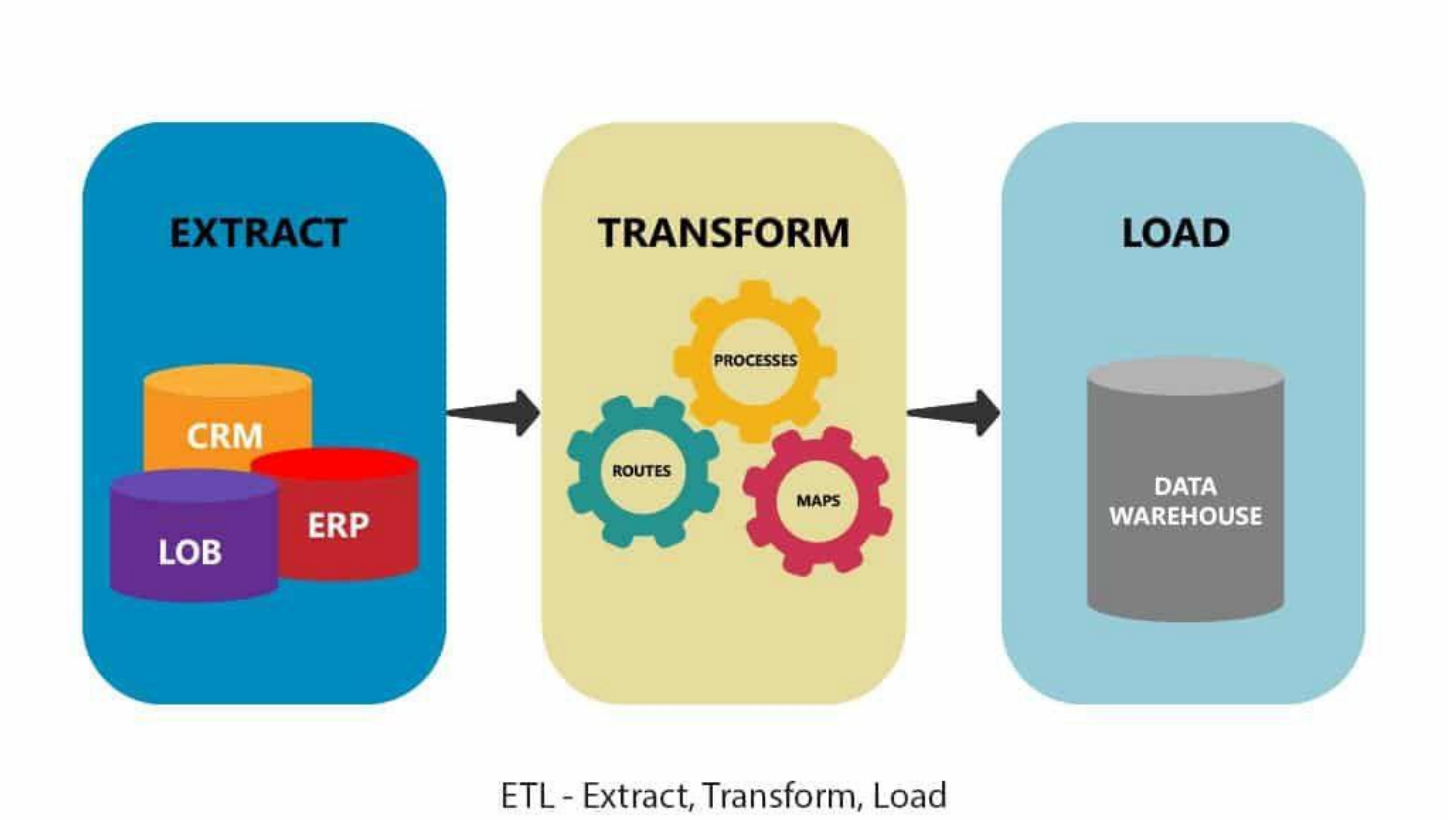


Figure 2.1: ETL – Extract, Transform and Load.

ELT

- ELT is an acronym for Extract, Load, and Transform. It's a process that transfers raw data from a source system to a target system and the information is then transformed for downstream applications.
- Unlike ETL, where data transformation occurs on an intermediate server prior to being loaded into the target system, ELT loads raw data directly into the target system and converts it over there.
- In this way, ELT is most beneficial for handling enormous datasets and use them for business intelligence and big data analytics.

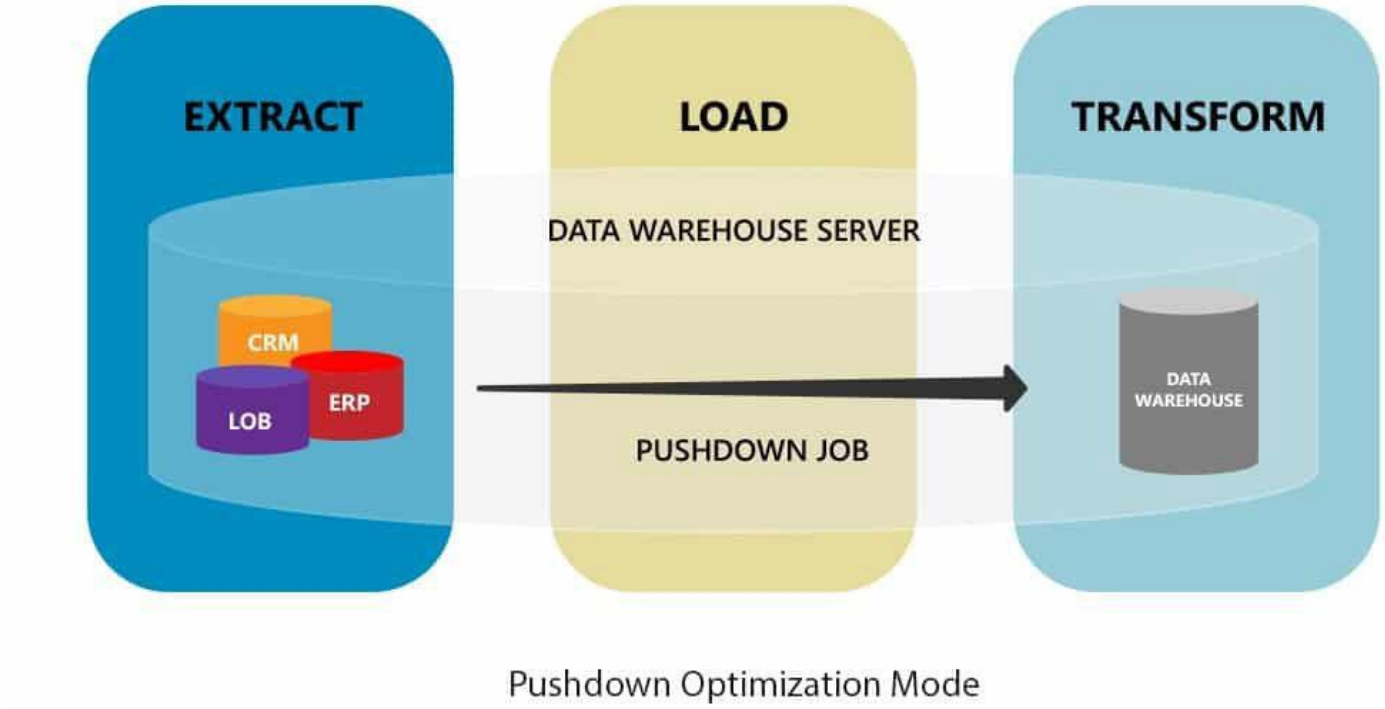


Figure 2.2: ELT – Extract, Load and Transform.

- As compared to the ETL process, ELT considerably reduces the load times. It's a more resource-efficient process as it leverages the processing capability developed into a data warehousing setup, decreasing the time spent in data transfer.

ETL vs. ELT: Which Approach Should You Choose?

Whether you should use ETL or ELT for a data management use case depends primarily on three things; the fundamental storage technologies, your data warehouse architecture, and the application of data warehouse for your business.

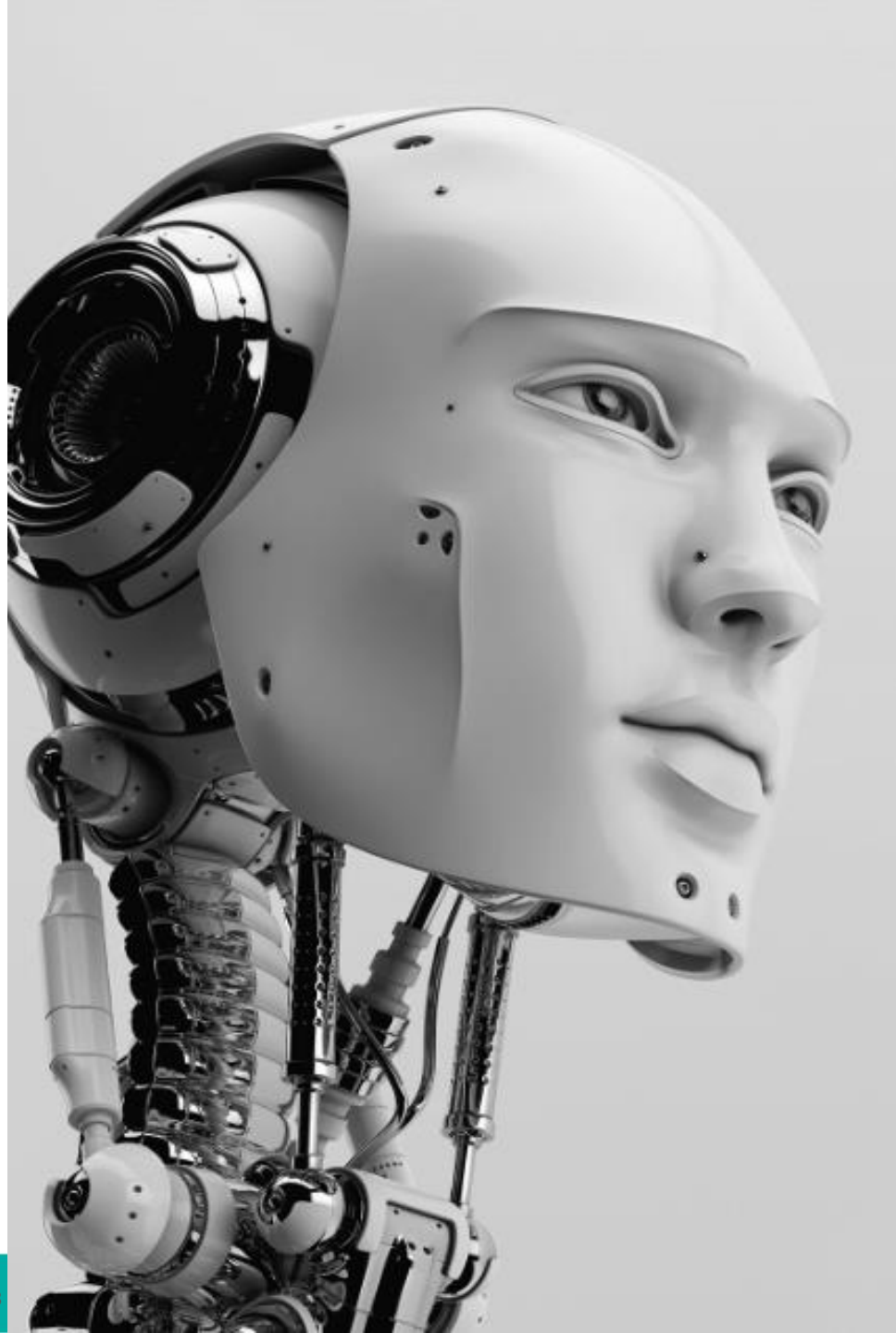
Parameters	ETL	ELT
Process	Data is transformed at staging server and then transferred to Datawarehouse DB.	Data remains in the DB of the Datawarehouse.
Code Usage	Used for Compute-intensive Transformations Small amount of data	Used for High amounts of data
Transformation	Transformations are done in ETL server/staging area.	Transformations are performed in the target system
Time-Load	Data first loaded into staging and later loaded into target system. Time intensive.	Data loaded into target system only once. Faster.
Time-Transformation	ETL process needs to wait for transformation to complete. As data size grows, transformation time increases.	In ELT process, speed is never dependant on the size of the data.
Time- Maintenance	It needs highs maintenance as you need to select data to load and transform.	Low maintenance as data is always available.
Implementation Complexity	At an early stage, easier to implement.	To implement ELT process organization should have deep knowledge of tools and expert skills.
Support for Data warehouse	ETL model used for on-premises, relational and structured data.	Used in scalable cloud infrastructure which supports structured, unstructured data sources.
Data Lake Support	Does not support.	Allows use of Data lake with unstructured data.
Complexity	The ETL process loads only the important data, as identified at design time.	This process involves development from the output-backward and loading only relevant data.
Cost	High costs for small and medium businesses.	Low entry costs using online Software as a Service Platforms.
Lookups	In the ETL process, both facts and dimensions need to be available in staging area.	All data will be available because Extract and load occur in one single action.
Aggregations	Complexity increase with the additional amount of data in the dataset.	Power of the target platform can process significant amount of data quickly.
Calculations	Overwrites existing column or Need to append the dataset and push to the target platform.	Easily add the calculated column to the existing table.
Maturity	The process is used for over two decades. It is well documented and best practices easily available.	Relatively new concept and complex to implement.
Hardware	Most tools have unique hardware requirements that are expensive.	Being Saas hardware cost is not an issue.
Support for Unstructured Data	Mostly supports relational data	Support for unstructured data readily available.

Figure 2.3: Difference between ETL vs ELT.

Chapter 2.4: Data Preprocessing Tasks

By the end of this topic, you should be able to:

- understand the data preprocessing tasks in data mining process.
- Implement data preprocessing tasks in preparing data for modelling purposes..



Data Preprocessing Tasks

- Raw data is highly vulnerable to missing values, noise and inconsistency and the quality of data affect the data mining results. So, there is a need to improve the quality of data, in order to improve mining results.
- For achieving better results, raw data is pre-processed so as to enhance its quality and make it error-free. This eases the mining process.
- As depicted in Figure 2.4 the various stages in which data preprocessing is performed.
    a. Data Cleaning
    b. Data Integration
    c. Data Transformation
    d. Data Reduction

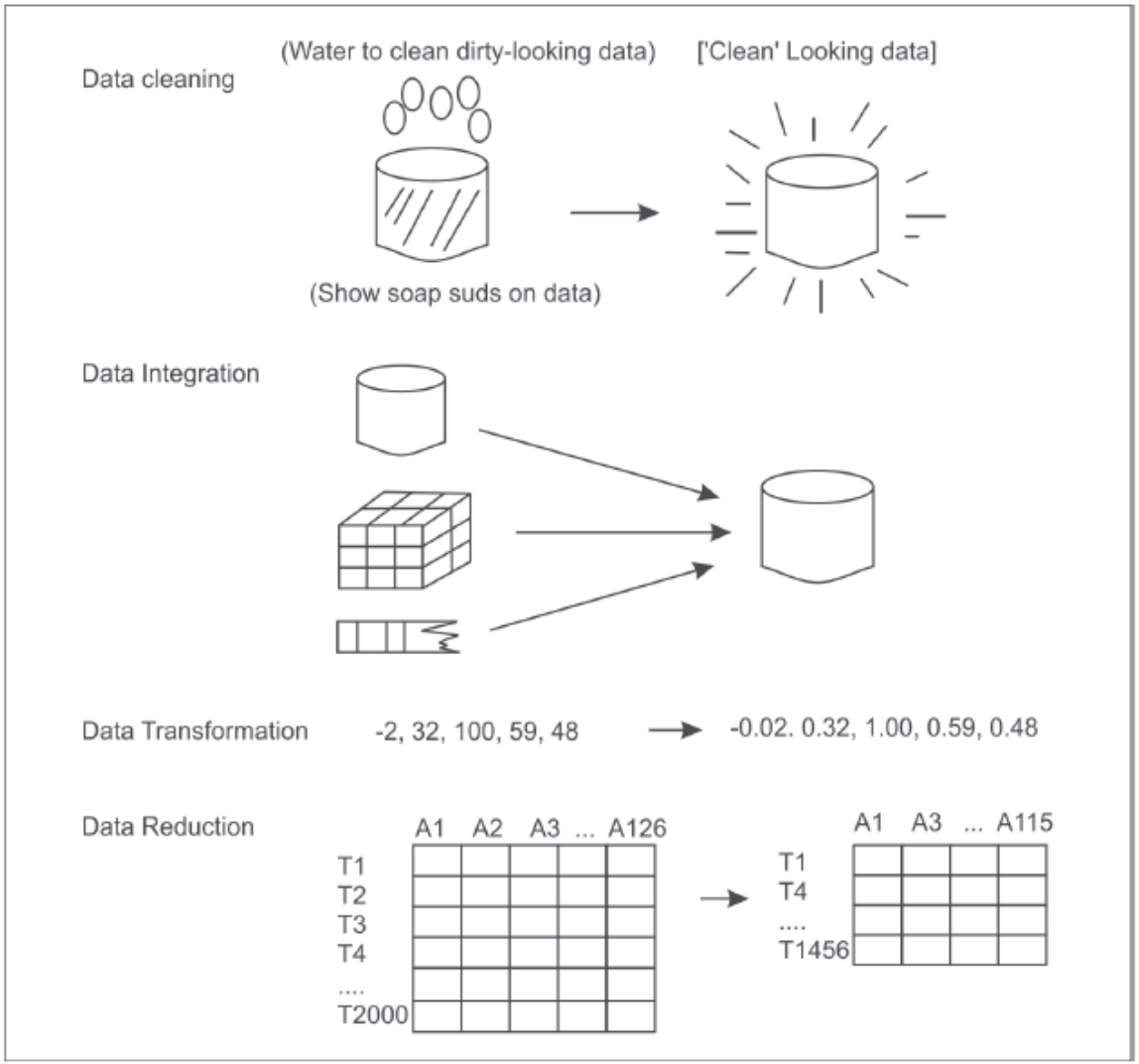


Figure 2.4: Various stages of preprocessing.

Data Cleaning

- First the raw data or noisy data goes through the process of cleansing.
- In data cleansing missing values are filled, noisy data is smoothened, inconsistencies are resolved, and outliers are identified and removed in order to clean the data.

Handling Missing Values

- It is often found that many of the database tuples or records do not have any recorded values for some attributes. Such cases of missing values are filled by different methods, as described below.
  i. Fill in the missing value manually: Naturally, manually filling each missing value is laborious and time-consuming and so it is practical only when the missing values are few in number. There are other methods to deal with the problem of missing values when the dataset is very large or when the missing values are very many.
  ii. Use of some global constant in place of missing value: In this method, missing values are replaced by some global label such as ‘Unkown’ or -∞. Although one of the easiest approaches to deal with the missing values, it should be avoided when the mining program presents a pattern due to repetitive occurrences of global labels such as ‘Unknown’. Hence, this method should be used with caution.
  iii. Use the attribute mean to fill in the missing value: Fill in the missing values for each attribute with the mean of other data values of the same attribute. This is a better way to handle missing values in a dataset.

iv. Use some other value which is high in probability to fill in the missing value: Another efficient method is to fill in the missing values with values determined by tools such as Bayesian Formalism or Decision Tree Induction or other inference-based tools. This is one of the best methods, as it uses most of the information already present to predict the missing values, although it is not biased like previous methods. The only difficulty with this method is the complexity in performing the analysis.

v. Ignore the tuple: If the tuple contains more than one missing value and all other methods are not applicable, then the best strategy to cope with missing values is to ignore the whole tuple. This is commonly used if the class label goes missing or the tuple contains missing values for most of the attributes. This method should not be used if the percentage of values that are missing per attribute varies significantly.

- So, we can conclude that using the attribute mean to fill in the missing value is the most common technique used by most data mining tools to handle the missing values. However, one can always use knowledge of probability to fill these values.

Handling Noisy Data

- Most data mining algorithms are affected adversely due to noisy data. The noise can be defined as unwanted variance or some random error that occurred in a measurable variable. Noise is removed from the data by the method of ‘smoothing’. The methods used for data smoothing are as follows:

i. Binning Methods

The Binning method is used to divide the values of an attribute into bins or buckets. It is commonly used to convert one type of attribute to another type. For example, it may be necessary to convert a real-valued numeric attribute like temperature to a nominal attribute with values cold, cool, warm, and hot before its processing. This is also called ‘discretizing’ a numeric attribute. There are two types of discretization, namely, equal interval and equal frequency. In equal interval binning, we calculate a bin size and then put the samples into the appropriate bin. In equal frequency binning, we allow the bin sizes to vary, with our goal being to choose bin sizes so that every bin has about the same number of samples in it. The idea is that if each bin has the same number of samples, no bin, or sample, will have greater or lesser impact on the results of data mining.

To understand this process, consider a dataset of the marks of 50 students. . The process divides this dataset on the basis of their marks into, for this example, 10 bins. In case of equal interval binning, we will create bins from 0-10, 10-20, 20-30, 30-40, 40-50, 50- 60, 60-70, 70-80, 80-90, 90-100. If most students commonly have marks between 60 to 80, some bins may be full and most bins may have very few entries e.g., 0-10, 10-20, 90-100. Thus, it might be better to divide this dataset on the equal frequency basis. It means that with the same 50 students in class and we want to put these into 10 bins on the basis of their marks then instead of creating the bins for marks like 0-10, 10-20 and so on, here we will first sort the records of students on the basis of their marks in descending order (or ascending order as we prefer). The first 5 students having highest marks will put into one bin and next 5 students on the basis of their marks will put into another and so on. If our boundary students have same marks then bin range can be shifted to accommodate students with the same marks into one common bin. For example, let us suppose that after arranging the data in descending order of marks and we found that marks of 5th and 6th students are same of 85. Then we cannot put one student in one bin and other in a different bin because both have the same marks.

So, we either shift our bin range may be up (i.e., 86 in this case) to accommodate the first 4 students in one bin and next 5 into another. Similarly, we can shift our bin range down to accommodate first 6 students in one bin (i.e., 85 in this case so that 5th and 6th student falls in same bin) and next 5 into another. Thus, in this case of equal frequency most of bins will have a count of approximately 5, while in case of equal interval some bins will be heavily loaded while most will be lightly loaded. Thus, the idea of having same number of samples in each bin works better as no bin, or sample, will have greater or lesser impact on the results of data mining.

ii. Clustering or outlier analysis

Clustering or outlier analysis is a method that allows the detection of outliers by clustering. In clustering, values which are common or similar are organized into groups or ‘clusters’, and those values which lie outside these clusters are termed as outliers or noise.

iii. Regression

Regression is another such method which allows data smoothing by fitting it to some function. For example, Linear Regression is one of the most used methods that aim at finding the most suitable line to fit values of two variables or attributes (i.e., best fit). The primary purpose of this is to predict the value of other variables using the first one. Similarly, Multiple Regression is used when more than two variables are involved. Regression allows data fitting which in turn removes noise from data and hence smoothens the dataset using mathematical equations.

iv. Combined computer and human inspection

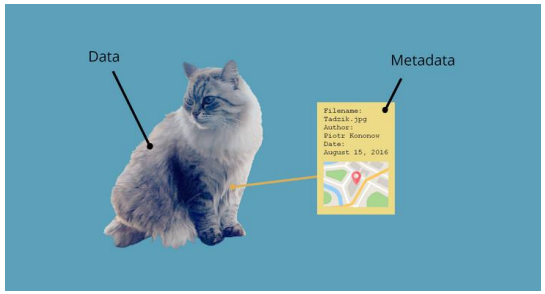
Using both computers and human inspection one can detect suspicious values and outliers.

Handling of Inconsistent Data

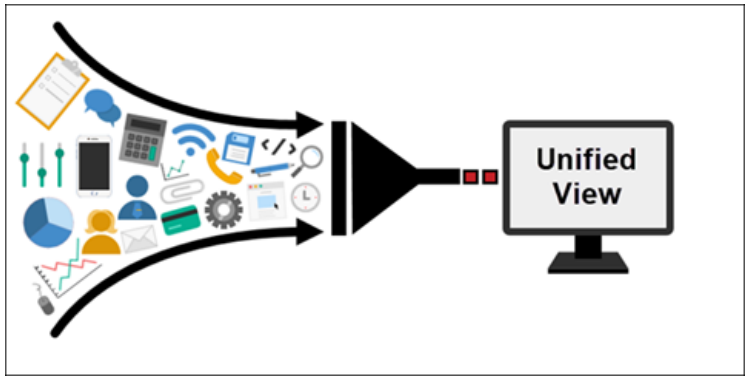
- Many times data inconsistencies are encountered when data is recorded during some transaction. Such inconsistencies can be manually removed by using external references.
- As an example: errors that have been made at the time of data entry be corrected manually by performing a paper trace operation.

Data Integration

- A most necessary step to be taken during data analysis is data integration. Data integration is a process which combines data from a plethora of sources (such as multiple databases, flat files or data cubes) into a unified data store.
- During data integration, several tricky issues have to be considered. For example, how does the data analyst or the analyzing machine be sure that the student_id of one database and student_number of another database refer to the same entity? This is referred to as the problem of entity identification. The solution to the problem lies with the term ‘metadata’.



- Databases consist of metadata, which is data about data. This metadata is taken as a reference and referred by the data analyst to avoid errors during the process of data integration.
- Another such issue which may be caused due to schema integration is redundancy. In the language of database, an attribute is said to be redundant if it is derivable from some other table (of the same database).
- Mistakes in attribute naming can also lead to data redundancies in the resulting dataset. We use a number of tools to perform data integration from different sources into one unified schema.



Data Transformation

- When the value of one attribute is small as compared to other attributes, then that attribute will not have much influence on the mining of information, since the values of this attribute were smaller than other attributes and the variation within the attribute will also be small. Thus, data transformation is a process in which data is consolidated or transformed into some other standard forms which are better suited for data mining.
- For example, the dataset given in Figure 2.5 is for the chemical composition of wine samples. Note that the values for different attributes cover a range of six orders of magnitude. It turns out that data mining algorithms struggle with numeric attributes that exhibit such ranges of values.

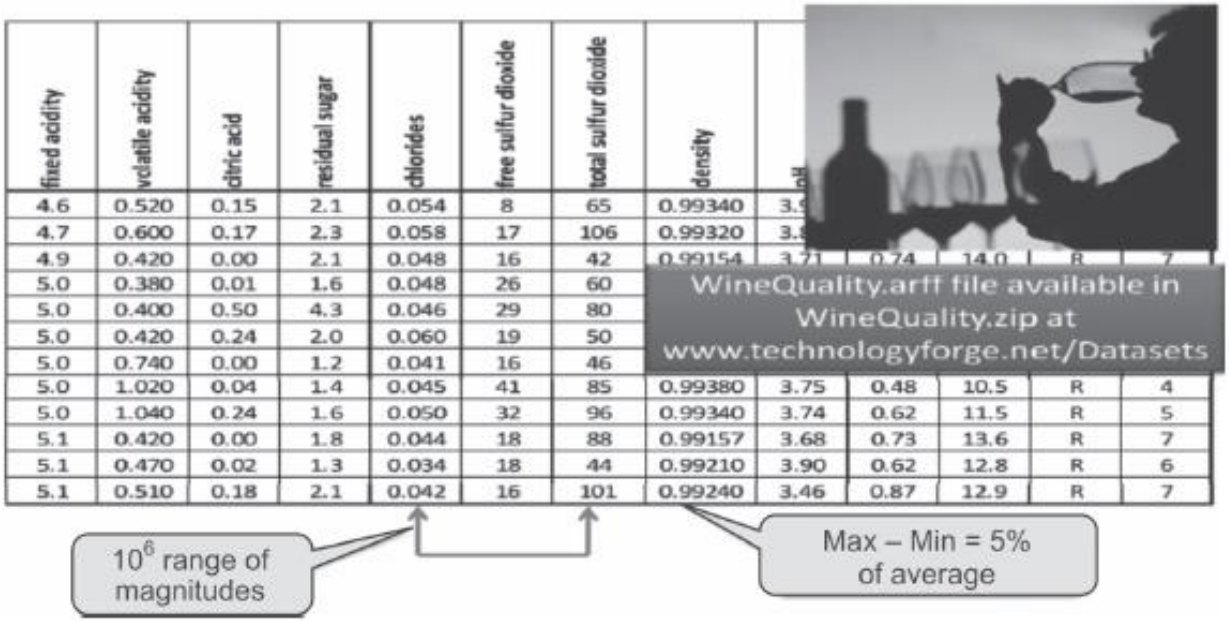


Figure 2.5: Chemical composition of wine samples.

- All attributes should be transformed to a similar scale for clustering to be effective unless we wish to give more weight to some attributes that are comparatively large in scale. Commonly, we use two techniques to convert the attributes: Normalization and Standardization are the most popular and widely used data transformation methods.

Normalisation vs Standardisation

- In developing the machine learning models, a model can be considered as good (or as bad) as the data you train the model with. The magnitude of different features affects different machine learning models for various reasons.
- For example, consider a data set containing two features, age, and income. Here age ranges from 0–100, while income ranges from 0 to a huge amount which is mostly higher than 100. Income is about 1,000 times larger than age. So, these two features are in very different ranges. When we do further analysis, like multivariate linear regression, for example, the attributed income will intrinsically influence the result more due to its larger value. But this doesn't necessarily mean it is more important as a predictor. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance.
- For this exact purpose, using Feature Scaling is essential.

Feature Scaling

- Feature scaling is a technique to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

Why use Feature Scaling?

- Gradient descent converges much faster with feature scaling than without it.
- Many classifiers (like KNN, K-means) calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. So the range of features should be scaled so that each feature contributes approximately proportionately to the final distance.
- However, every dataset does not require feature scaling. It is required only when features have different ranges.

Normalisation

Normalization (also called, Min-Max normalisation) is a scaling technique such that when it is applied the features will be rescaled so that the data will fall in the range of [0,1]

Normalized form of each feature can be calculated as follows:

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

Here ‘x’ is the original value and ‘x`’ is the normalized value.

Standardisation

Standardization (also called, Z-score normalisation) is a scaling technique such that when it is applied the features will be rescaled so that they'll have the properties of a standard normal distribution with mean, $\mu=0$ and standard deviation, $\sigma=1$; where μ is the mean (average) and σ is the standard deviation from the mean.

Standard scores (also called z scores) of the samples are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

This scales the features in a way that they range between [-1,1]

When to use what?

- “ Normalization or Standardization?” — There is no obvious answer to this question: it really depends on the application.
- Generally, there is no rule for when to use normalization versus standardization. However, if your data has outliers, use standardization, otherwise, use normalization.
- Using standardization tends to make the remaining values for all of the other attributes fall into similar ranges since all attributes will have the same standard deviation of 1.

Data Reduction

- It is often seen that when the complex data analysis and mining processes are carried out over humongous datasets, they take such a long time that the whole data mining or analysis process becomes unviable.
- Data reduction techniques come to the rescue in such situations. Using data reduction techniques a dataset can be represented in a reduced manner without actually compromising the integrity of the original data.
- Data reduction is all about reducing the dimensions (referring to the total number of attributes) or reducing the volume. Moreover, mining when carried out on reduced datasets often results in better accuracy and proves to be more efficient.

- There are many methods to reduce large datasets to yield useful knowledge. A few among them are:
  a. Dimension reduction - In data warehousing, ‘dimension’ equips us with structured labelling information. But not all dimensions (attributes) are necessary at a time. Dimension reduction uses algorithms such as Principal Component Analysis (PCA) and others. With the usage of such algorithms, one can detect and remove redundant and weakly relevant, attributes or dimensions.
  b. Numerosity reduction - It is a technique which is used to choose smaller forms of data representation for reducing the dataset volume.
  c. Data compression - We can also use data compression techniques to reduce the dataset size. These techniques are classified as lossy and lossless.

Links to Read

Data Preprocessing in Data Mining

https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Data Preprocessing in Data Mining -A Hands On Guide

https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

Data Quality

https://www.heavy.ai/technical-glossary/data-quality#:~:text=There%20are%20six%20main%20dimensions,confirmed%20with%20a%20verifiable%20source

Links to Read

Data Integration vs Data Migration: A Comparative Study

https://hevodata.com/learn/data-integration-vs-data-migration/

ETL vs. ELT: What's the difference?

https://medium.com/@asterasoftware1/etl-vs-elt-whats-the-difference-46107728cd7f

Normalization vs Standardization

https://towardsdatascience.com/normalization-vs-standardization-cb8fe15082eb