Data Preprocessing/Feature Engineering

Preprocessing

- Real-world databases are highly susceptible to noisy, missing and inconsistent data due their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources.

- Low quality data will lead to a low-quality mining results.

- *How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results? Moreover, how can the data be preprocessed so as to improve the efficiency and ease of mining process*

- Data preprocessing is very important and comprises majority of the work in a data mining/modeling process (about 80% of the work is done in this stage)

- There are a number of preprocessing techniques which can be applied to improve the quality of the data.

Data Preprocessing/Feature Engineering

Data preparation may be one of the most difficult and important steps in any machine learning project.

The reason is that each dataset is different and highly specific to the project. Nevertheless, there are enough commonalities across predictive modeling projects that we can define a loose sequence of steps and subtasks that you are likely to perform.

This process provides a context in which we can consider the data preparation required for the project, informed both by the definition of the project performed before data preparation and the evaluation of machine learning algorithms performed after.

Data Preprocessing/Feature Engineering

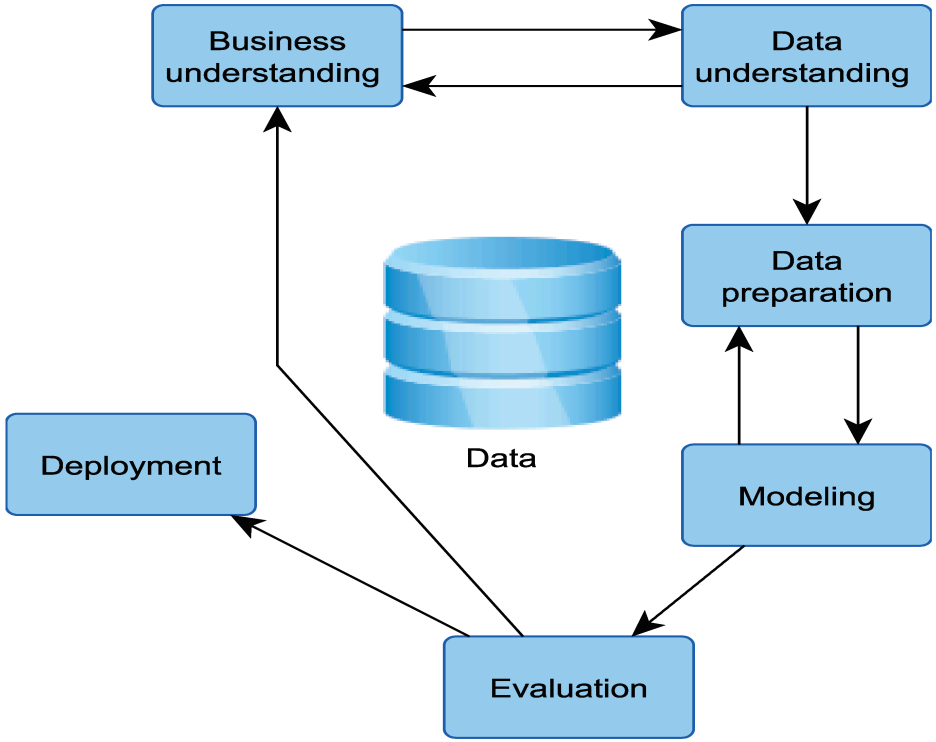
Each machine learning project is different because the specific data at the core of the project is different.

The right features can only be defined in the context of both the model and the data; since data and models are so diverse, it is difficult to generalize the practice of feature engineering across projects.

Even though project is unique, the steps on the path to a good or even the best result are generally the same from project to project. This is known as “data science process”.

The process consists of a sequence of steps. The steps are the same, but the names of the steps and takes performed may differ form description to description.

Data mining and Knowledge discovery



Data Preprocessing/Feature Engineering

The steps are written sequentially, but we will jump back and forth between the steps for any given project.

- Define Problem
- Prepare Data
- Evaluate Models
- Finalize Model

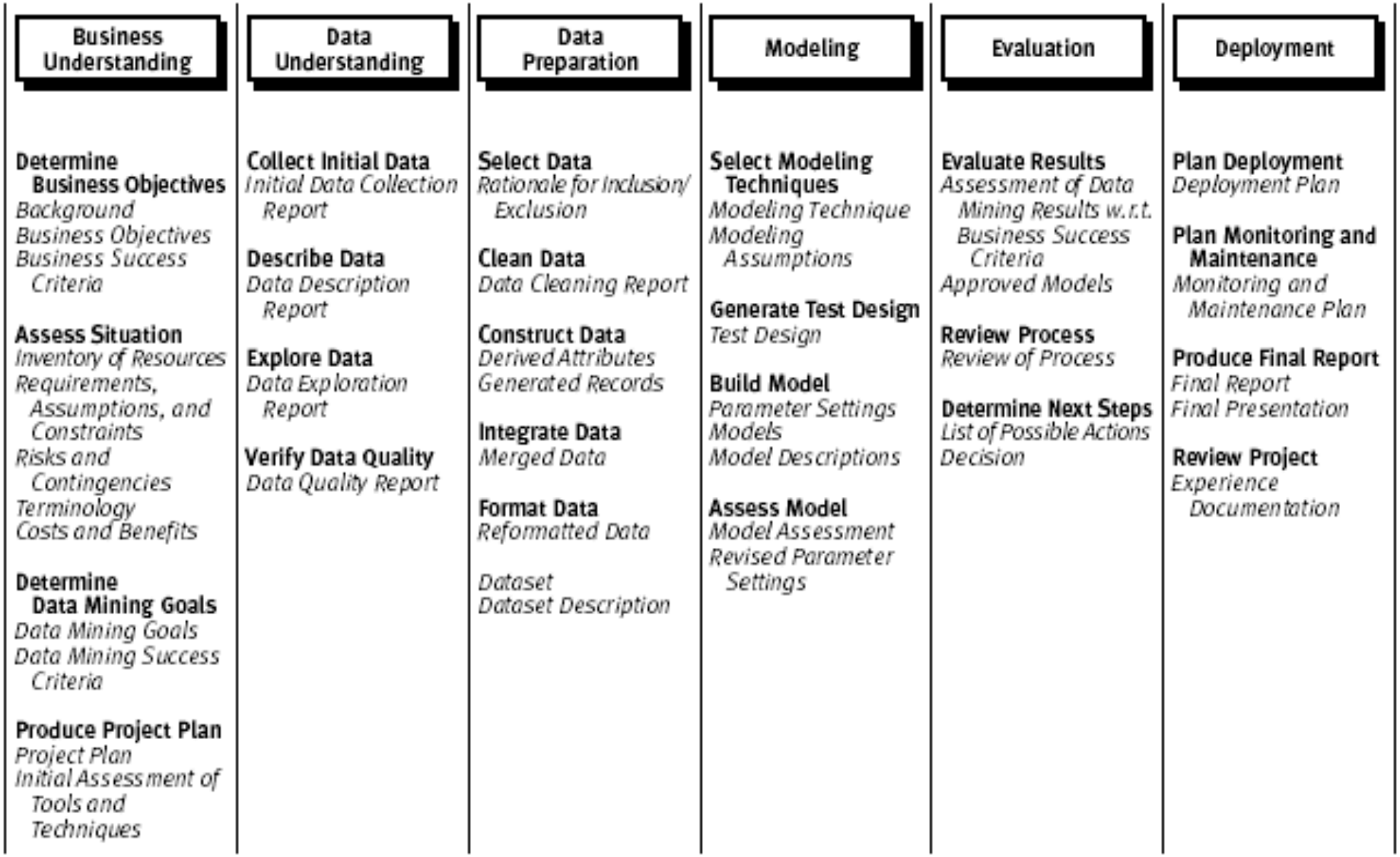


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Data Preprocessing/Feature Engineering

Define Problem

This step is concerned with learning enough about the project t
select the framing or framings of the prediction task. For example,
it classification or regression, or some other higher-order proble
type?

It involves collecting the data that is believed to be useful in making
prediction and clearly defining the form that the prediction will tak
It may also involve talking to project stakeholders and other peop
with deep expertise in the domain.

This step also involves taking a close look at the data, as well a
perhaps exploring the data using summary statistics and dat
visualization.

What is data?

- Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things
- Collection of data objects and their attributes/variable, characteristic or feature
- An attribute is a property of an object, e.g., color of a car, or temperature of furnace
- A collection of features describe an object. Object is also known as record, observation, sample etc.

Attributes

Observations

Sample No.	Thickness (cm)	Temperature (°C)	Concentration (g/L)
1	2.1740228	82	0.066
2	1.8774501	77	0.071
3	1.8774704	77	0.072
4	1.9762727	79	0.069
5	2.0266303	80	0.071
6	2.0994529	81	0.066
7	1.9468132	78	0.067
8	1.8972298	77	0.071
9	1.9169798	77	0.07
10	2.0692626	80	0.066
11	2.1292363	82	0.067
12	2.0479427	80	0.067
13	2.0479598	80	0.069
14	1.8972463	77	0.071
15	1.8774795	77	0.066

What is feature and why we need engineering of it

All machine learning algorithms use some input data to create outputs. This input data comprise features, which are usually in the form of structured columns. Algorithms require features with some specific characteristic to work properly.

Here, the need for feature engineering arises.

Data Representation

- Data has form: $\{(x_1,y_1),...,(x_n,y_n)\}$ (labeled), or $\{x_1,...,x_n\}$ (unlabeled)
- What the label y looks like is task-specific
- What about x which denotes a real-world object (e.g., image or text document)?
- Each example x is a set of (numeric) features/attributes/dimensions
- Features encode properties of the object which x represents
- x is commonly represented as a D × 1 vector
- Representing a 28 × 28 image: x can be a 784 × 1 vector of pixel values
- Representing a text document: x can be a vector of word-counts of words appearing in that document

Cont'd

Attirbutes						
Observations		x_1	x_2	x_3	…	x_p
	1	x_{11}	x_{12}	x_{13}	…	x_{1p}
	2	x_{21}	x_{22}	x_{23}	…	x_{2p}
	3	x_{31}	x_{23}	x_{33}	…	x_{3p}
	…	…	…	…	…	…
	n	x_{n1}	x_{n2}	x_{n3}		x_{np}

- You got your data: what's next:

What kind of analysis do you need which model is more appropriate for it? …

Most statistical method focuses on data modeling, prediction and statistical inference while it is usually assumed that data are in the correct state for data analysis. In practice, a data analyst/modeler spends much if not most of his time on preparing the data before doing any statistical operation.

It is very rare that the raw data one works with are in the correct format, are without errors, are complete and have all the correct labels and codes that are needed for analysis.

Data cleaning, or data preparation is an essential part of statistical analysis. In fact, in practice it is often more time-consuming than the statistical analysis itself. These techniques cover technical as well as subject-matter aspects of data cleaning.

Technical aspects include data reading, type conversion and string matching and manipulation. Subject-matter related aspects include topics like data checking, error localization and an introduction to imputation methods.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge during the training phase is more difficult. Data preparation can take considerable amount of time. The product of the data pre-processing is the final training data set.
The Python provides a good environment for reproducible data cleaning since all cleaning actions can be scripted and therefore reproduced.

Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to improve the quality of the data and, consequently of the mining results raw data is pre-processed so as to improve the efficiency and ease of the mining process. Data preparation is one of the most steps in data science project which deals with the preparation and transformation of raw dataset.

Need For Data Pre-Processing

- You want to get the best accuracy from machine learning algorithms on your datasets.

- Some machine learning algorithms require the data to be in a specific form. Whereas other algorithms can perform better if the data is prepared in a specific way, but not always.

- It is important to prepare your data in such a way that it gives various different machine learning algorithms the best chance on your problem.

- You need to pre-process your raw data as part of your machine learning project.

- Some data preparation is needed for all mining tools. The purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool

- Preparing data also prepares the scientist so that when using prepared data the scientist produces better models, and faster.

- Good data is a prerequisite for producing effective models of any type.

- Several data mining methods are sensitive to the scale and/or type of the variables

- Different variables (columns of our data sets) may have rather different scales

- Some methods are not able to handle either nominal or numeric Variables

- We may need to “create” new variables to achieve our objectives

- Sometimes we are more interested in relative values (variations) than absolute values

- We may be aware of some domain-specific mathematical relationship among two or more variables that is important for the task

- Frequently we have data sets with unknown variable values

- Our data set may be too large for some methods to be applicable

- If the user believe that the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output.

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”, Age=“250”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Data Preparation

- Preparing the data is one of the most time-consuming parts of data analysis.
- The way in which the data is collected and prepared is critical to the confidence with which decisions can be made.
- The data needs to be merged into a table and this may involve integration of the data from multiple sources.
- Once the data is in a tabular format, it should be fully characterized.
- The data should be cleaned by resolving ambiguities and errors, removing redundant and problematic data, and eliminating columns of data irrelevant to the analysis.
- This whole process is called *pre-processing. It goes by other names, such as "data wrangling", "data cleaning", and feature engineering*
- Data pre-processing techniques generally refer to the addition, deletion, or transformation of the dataset.

Data Preprocessing/Feature Engineering

Data Cleaning is the process of transforming raw data into consistent data that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability

This is highly specific to data, to the goals of the project, and to the algorithm that will be used to model the data.

The broader philosophy of data preparation is to discover how to best expose the underlying structure of the problem to the learning algorithm. This is the guiding light.

It can be more complicated than it appears at first glance. For example, different input variables may require different data preparation methods.

Measure of Data Quality

- Raw data is often not useful without some kind of organization or manipulation. Raw data seems to be just a bunch of meaningless values without any context or some level of organization.
- In recent years, data quality has gained more an more attention due to extended use of data warehouse systems and a higher relevance of customer relationship management.
- Due to this fact for decision makers, the benefits of data depends heavily on their completeness, correctness, and timeliness, respectively. Such properties are known as data quality dimensions.
- The consequences of poor data quality are manifold: They range from worsening customer relationships and customer satisfaction by falsely addressing customers to insufficient decision support for managers.

Measure of Data Quality

- The following measure can be used to test the quality of data
  - *Completeness: The proportion of stored data against the potential of "100% complete"*
  - *Uniqueness: No observation will be recorded more than once based upon how that observation is identified.*
  - *Velocity: The rate at which data is coming especially for streaming data*
  - *Accuracy: The degree to which data correctly describes the "real world" event being described*
  - *Consistency: The absence of difference, when comparing two or more representations of a thing against a definition.*
  - *Accessibility: How easy it is to access the data.*

Data Preprocessing/Feature Engineering

How do we know what data preparation techniques to use in our data?

As with many questions of statistics, the answer to “which feature engineering methods are the best?” is that it depends. Specifically, it depends on the model being used and the true relationship with the outcome.

On the surface, this is a challenging question, but if we loo at the data preparation step in the context of the whole project, it becomes more straightforward.

The step before data preparation involves defining the problem

Data Preprocessing/Feature Engineering

As part of defining the problem, this may involve many sub-tasks, such as:

- Gather data from the problem domain.
- Discuss the project with subject matter experts.
- Select those variables to be used as inputs and outputs for a predictive model.
- Review the data that has been collected.
- Summarize the collected data using statistical methods.
- Visualize the collected data using plots and charts.

Information known about the data can be used in selecting and configuring data preparation methods.

Major tasks involved

- **Integration of data**
  - Integration of data from multiple databases, or files
- **Data Cleaning**
  - Fill in missing values, smooth noisy data, remove outliers and resolve inconsistencies
- **Feature Selection**
  - Identifying features that are most relevant to the task
- **Feature Engineering**
  - Deriving new features from available data
- **Data Transformation**
  - Scaling/normalization and aggregation. Normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
- **Data reduction**
  - Optimize the features/attributes and obtain reduced representation in volume.

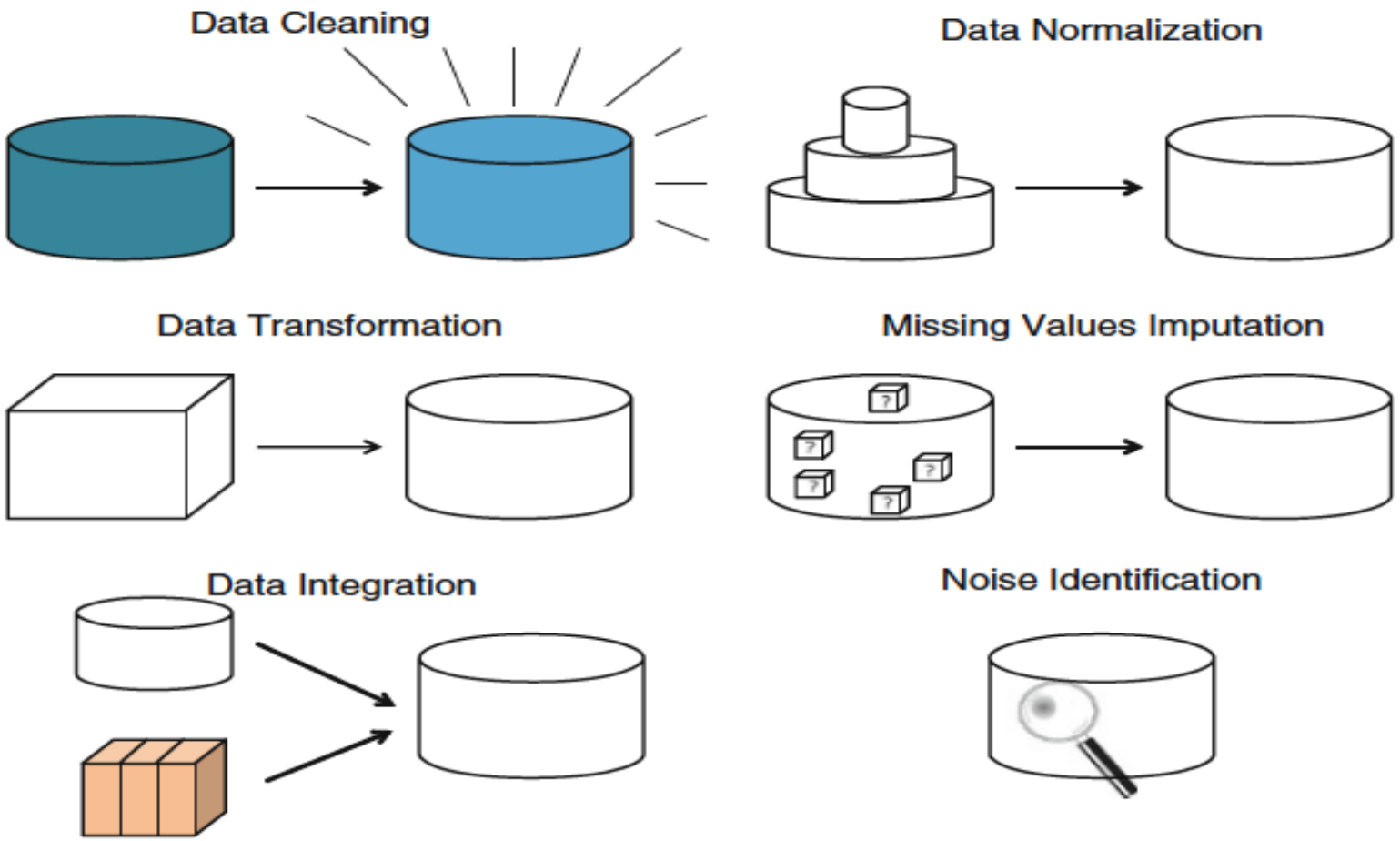
Data Preprocessing/Feature Engineering

There may also be interplay between the data preparation step and the evaluation of models.

Information known about the choice of algorithms and the discovery of well-performing algorithms can also inform the selection and configuration of data preparation methods.

For example, the choice of algorithms may impose requirements and expectations on the type and form of input variables in the data. This might require variables to have a specific probability distribution, the removal of correlated input variables, and/or the removal of variables that are not strongly related to the target variable.

- Forms of data preparation



Cont'd

- These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

- Each of these tasks is a whole field of study with specialized algorithms.

- Data preparation is not performed blindly

- Data cleaning techniques, when applied before mining, can substantially improve the overall quality of the modeling.

- Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying o removing outliers, and resolving inconsistencies.

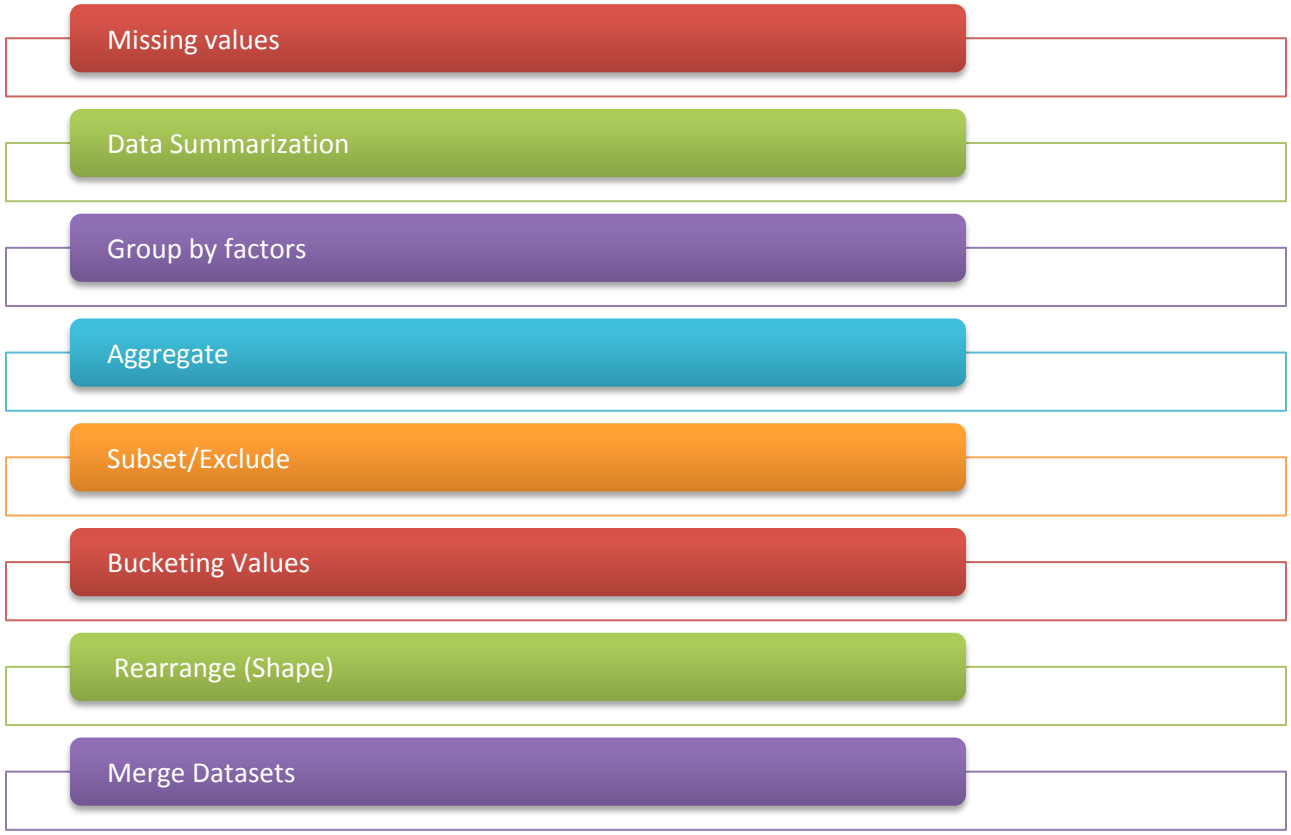
How to Prepare Data

Below are the steps involved to understand, clean and prepare the data for building the predictive model:

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

What are the ways to manipulate data



Data Integration

- Data integration is the process of combining data from multiple sources. These sources may include multiple databases, data cubes or flat files
- The data may also need to be transformed into forms appropriate for mining.
- Integrate metadata from different sources
- Removing duplicates and redundant data
- Detect and resolve data value conflicts
  - For the same real world entity, attribute values from different sources are different, e.g., different scales, different units (miles vs. km)

Data Cleaning

Data cleaning involves fixing systematic problems or errors in “messy” data.

The most useful data cleaning involves deep domain expertise and could involve identifying and addressing specific observations that may be incorrect.

There are many reasons data may have incorrect values, such as being mistyped, corrupted, duplicated, and so on. Domain expertise may allow obviously erroneous observations to be identified as they are different from what is expected, such as a person’s height of 200 feet.

Once messy, noisy, corrupt, or erroneous observations are identified, they can be addressed. This might involve removing a row or a column. Alternately, it might involve replacing observations with new values.

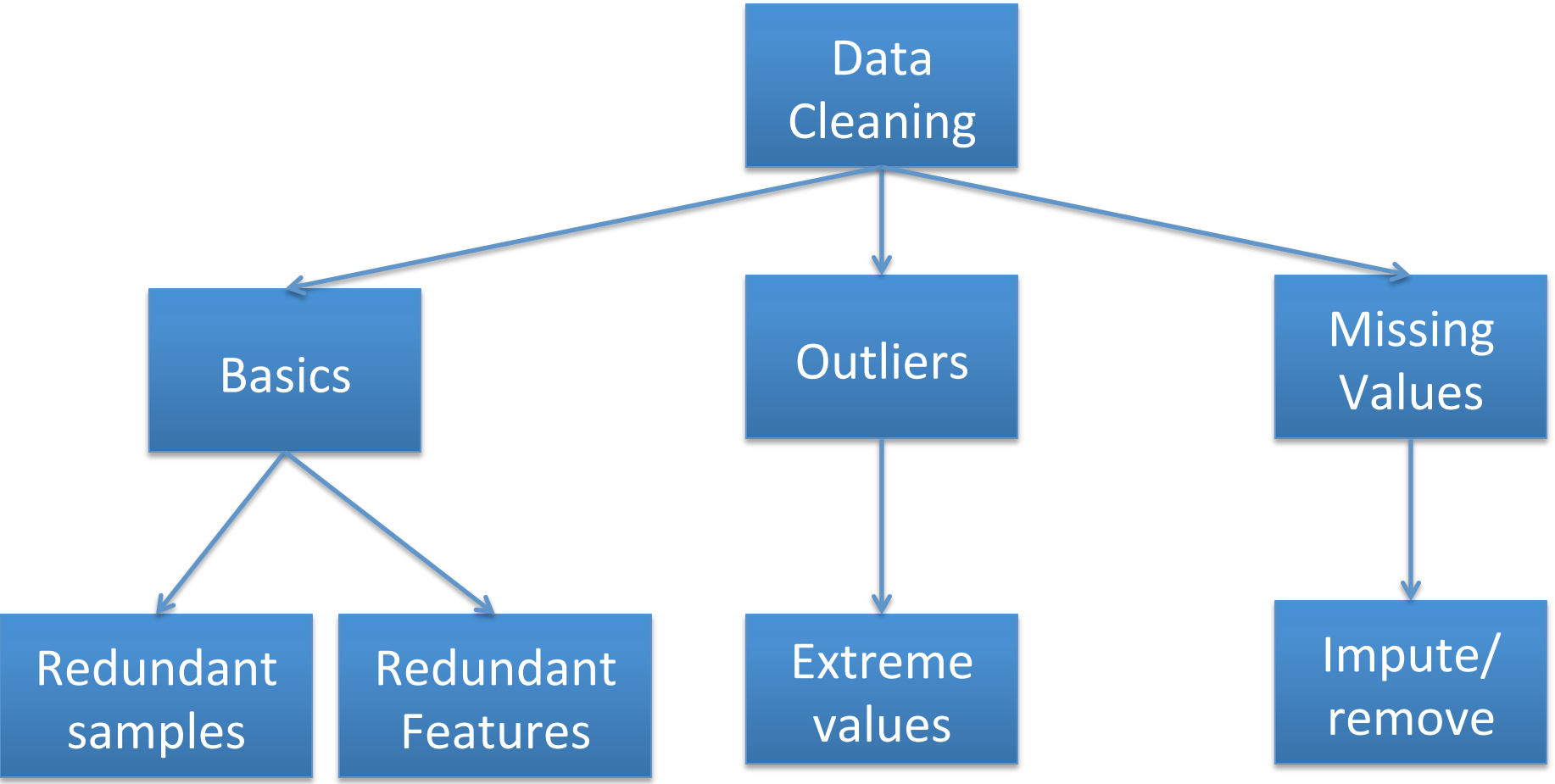
Data Cleaning

Nevertheless, there are general data cleaning operations that can be performed, such as:

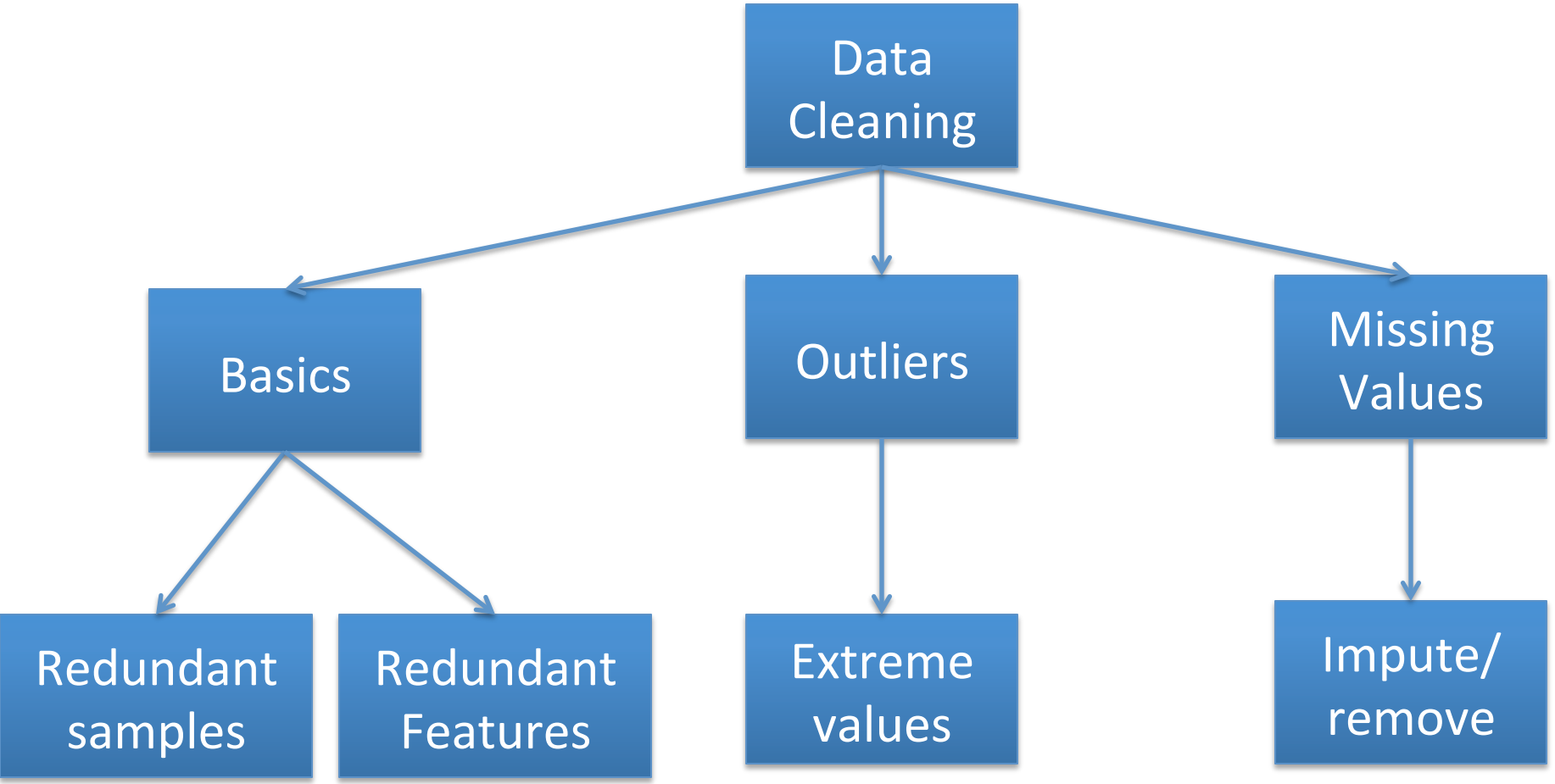
- Using statistics to define normal data and identify outliers.
- Identifying columns that have the same value or no variance and removing them.
- Identifying duplicate rows of data and removing them.
- Marking empty values as missing. Imputing missing values using statistics or a learned model.

Data cleaning is an operation that is typically performed first, prior to other data preparation operations.

Data Cleaning



Feature Selection



Data Cleaning

Data cleaning is a critically important step in any machine learning project.

In tabular data, there are many different statistical analysis and data visualization techniques you can use to explore your data in order to identify data cleaning operations you may want to perform.

Before jumping to the sophisticated methods, there are some very basic data cleaning operations that you probably should perform on every single machine learning project. These are so basic that they are often overlooked by seasoned machine learning practitioners, yet are so critical that if skipped, models may break or report overly optimistic performance results.

Data Cleaning

Data cleaning refers to identifying and correcting errors in the dataset that may negatively impact a predictive model.

Data cleaning is used to refer to all kinds of tasks and activities to detect and repair errors in the data.

Although critically important, data cleaning is not exciting, not does it involve fancy techniques. Just a good knowledge of the dataset.

Cleaning up your data is not the most glamorous of tasks, but it's an essential part of data wrangling. Knowing how to properly clean and assemble your data will set you miles apart from others in your field.

Data Cleaning

There are many types of errors that exist in a dataset, although some of the simplest errors include columns that don't contain much information and duplicated rows.

- **Identify columns that contain a single value**
  - Columns that have a single observation or value are probably useless for modeling
  - These columns are referred to zero-variance features because there truly is no variance displayed by the feature.
  - Columns that have a single value do not contain any information for modeling

You can detect this by using ***unique()*** function and can be removed by using ***drop()*** function.

Data Cleaning

- **Identify columns that contain very few values**
  - There may be columns in the dataset which have few unique values. This might make sense for ordinal or categorical features. In this case, the dataset only contains numerical variables.
  - These columns are near-zero variance features, as their variance is not zero, but a very small number close to zero.

These columns may or may not contribute to the model. We can't assume that they are useless to modeling. This does not mean that these columns should be deleted, but they require further attention. For example:

- The unique values can be encoded as ordinal values
- Unique values can be encoded as categorical values
- Compare model with each variable removed from the dataset.

Identify rows that contain Duplicate Data

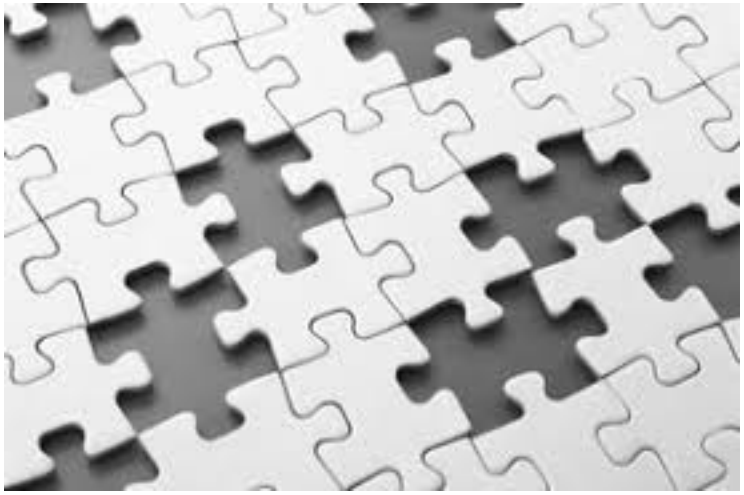
Rows that have identical data are probably useless

If you have used raw data that may have duplicate entries, removing duplicate data will be an important step in ensuring that data can be used accurately.

The pandas function ***duplicated()*** can be used to find the duplicated values and ***drop_duplicates()*** function can be used to remove duplicate rows.

Missing Data

- Missing data in the data set can reduce the power/fit of a model or can lead to a biased model. It can lead to wrong prediction or classification.
- Data is not always available
  - e.g., many records have not values fro attribute, such as customer age or income in sales data



Missing Data

- Missing data may be due to the following reasons:
  - Equipment malfunction
  - Data not entered properly
  - Certain data may not be considered important at the time of data entry or collection
  - Deleted due to inconsistent
  - Information is not collected (e.g. people decline to give their age)
  - NA, Inf, NaN, NULL
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

Handling missing values

- *Ignore the Missing Value During Analysis*: This is usually done when class label is missing (assuming the mining task involves classification). This method is not very effective, unless the record contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
- *Fill in the missing value manually:* In general, this approach is time-
- consuming and may not be feasible given a large data set with many missing values
- Use a global constant/mean or median to fill in the missing value: Replace all missing feature values by the same constant/mean or median of the attribute

- Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

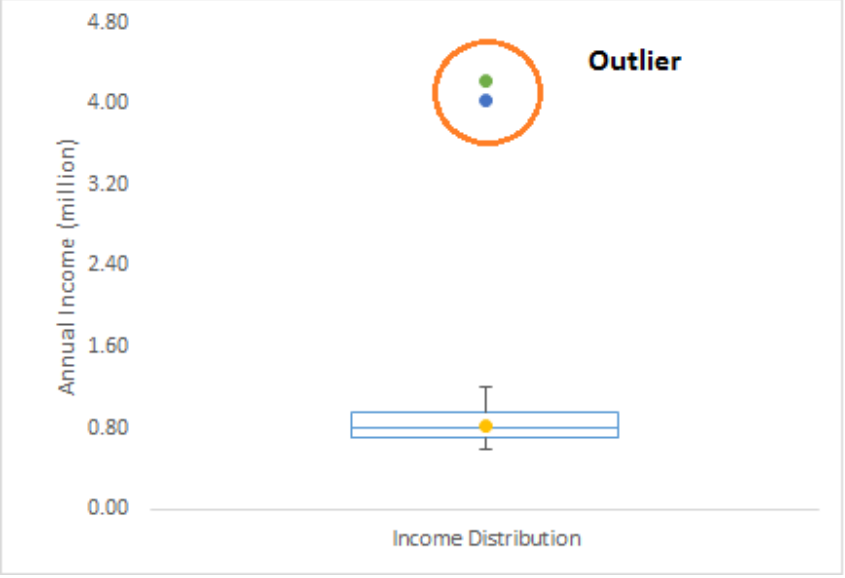
- It is important to note that, in some cases, a missing value may not imply error in the data, e.g., in case of CC, driver license number.

Handling missing values

- Note that, it is as important to avoid adding bias and distortion to the data as it is to make the information available.
    - bias is added when a wrong value is filled-in
- No matter what techniques you use to conquer the problem, it comes at a price. The more guessing you have to do, the further away from the real data the database becomes. Thus, in turn, it can affect the accuracy and validation of the mining results.

Outliers

- Outlier is a commonly user term as it needs attention otherwise it can result in wildly wrong estimations.
- Simply speaking, outlier is an
- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
- Data points inconsistent with the majority of data
- Outlier detection can be used for fraud detection or data cleaning



Impact of Outliers

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavourable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

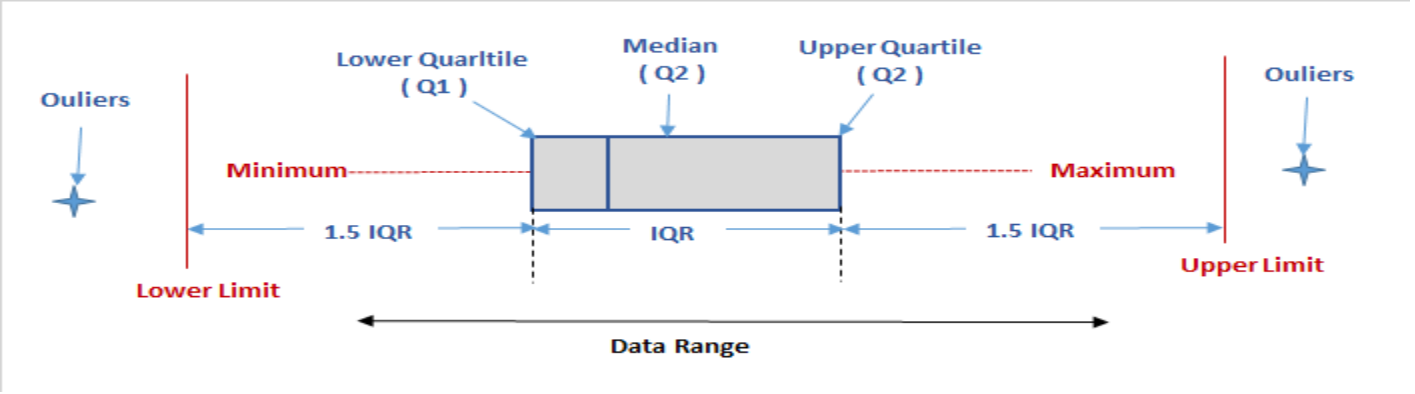
To understand the impact deeply, let's take an example to check what happens to a data set with and without outliers in the data set.

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

How to detect Outliers

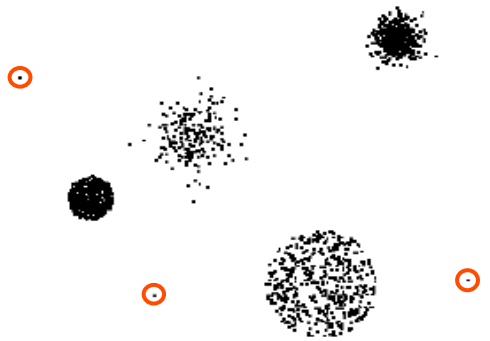
Most commonly used method to detect outliers is visualization. We use various visualization methods, like *Box-plot, Histogram, Scatter plot*. Some analysts also various thumb rules to detect outliers. Some of them are:

- Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR
- Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier
- Data points, three or more standard deviation away from mean are considered outlier
- Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding

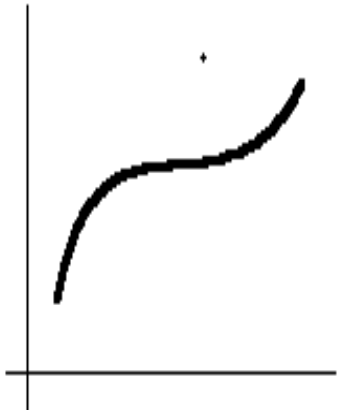


How to detect outliers

- Clustering
  - Outliers may be detected by clustering, where similar values are organized into groups, or “clusters”, Intuitively, values that fall outside of the set of clusters may be considered outliers or very small clusters are outliers



- Combined computer and human inspection
  - Tedious and time consuming
- Curve fitting



How to remove outliers

Imputing: Like imputation of missing values, we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyse if it is natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use statistical model to predict values of outlier observation and after that we can impute it with predicted values.

Treat separately: If there are significant number of outliers, we should treat them separately in the statistical model. One of the approach is to treat both groups as two different groups and build individual model for both groups and then combine the output.

Feature Selection

Feature selection refers to techniques for selecting a subset of input features that are most relevant to the target variable that is being predicted.

This is important as irrelevant and redundant input variables can distract or mislead learning algorithms possibly resulting in lower predictive performance. Additionally, it is desirable to develop models only using the data that is required to make a prediction, e.g. to favor the simplest possible well performing model.

Data Transformation

- The final step is to transform the data.
- Sometimes data analysis programs have difficulty processing data in its raw form.

For these cases, certain mathematical transformation can be applied to the data.

- Smoothing: remove noise from data
- Normalization: The preprocessed data may contain attributes with a mixture of scales for various quantities such as nautical miles, height. Many learning algorithms like data attributes to have the same scale such as between 0 and 1 for the smallest and largest value for a given feature.
- New attributes constructed from the given ones
- Decomposition: There may be that represent a complex concept parts. An example is a date that may have day and time components that in turn could be split further.
- Aggregation: There may be features that can be aggregated into a single feature that would be more meaningful to the problem you are trying to solve. For example, there may be a data instances for each minute a customer's usage of energy, that could be aggregated over an hr.

Cont'd

- *Normalization* uses a mathematical function to transform numeric columns to a new range. Normalization is important in preventing certain data analysis methods from giving some variables undue influence over others because of differences in the range of their values. In other words, normalization helps to prevent that attributes with large ranges out-weight attributes with small ranges

- There are various ways to do this.

Min-max normalization

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

Cont'd

- *Z-Score normalization*

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- *Normalization by decimal scaling*

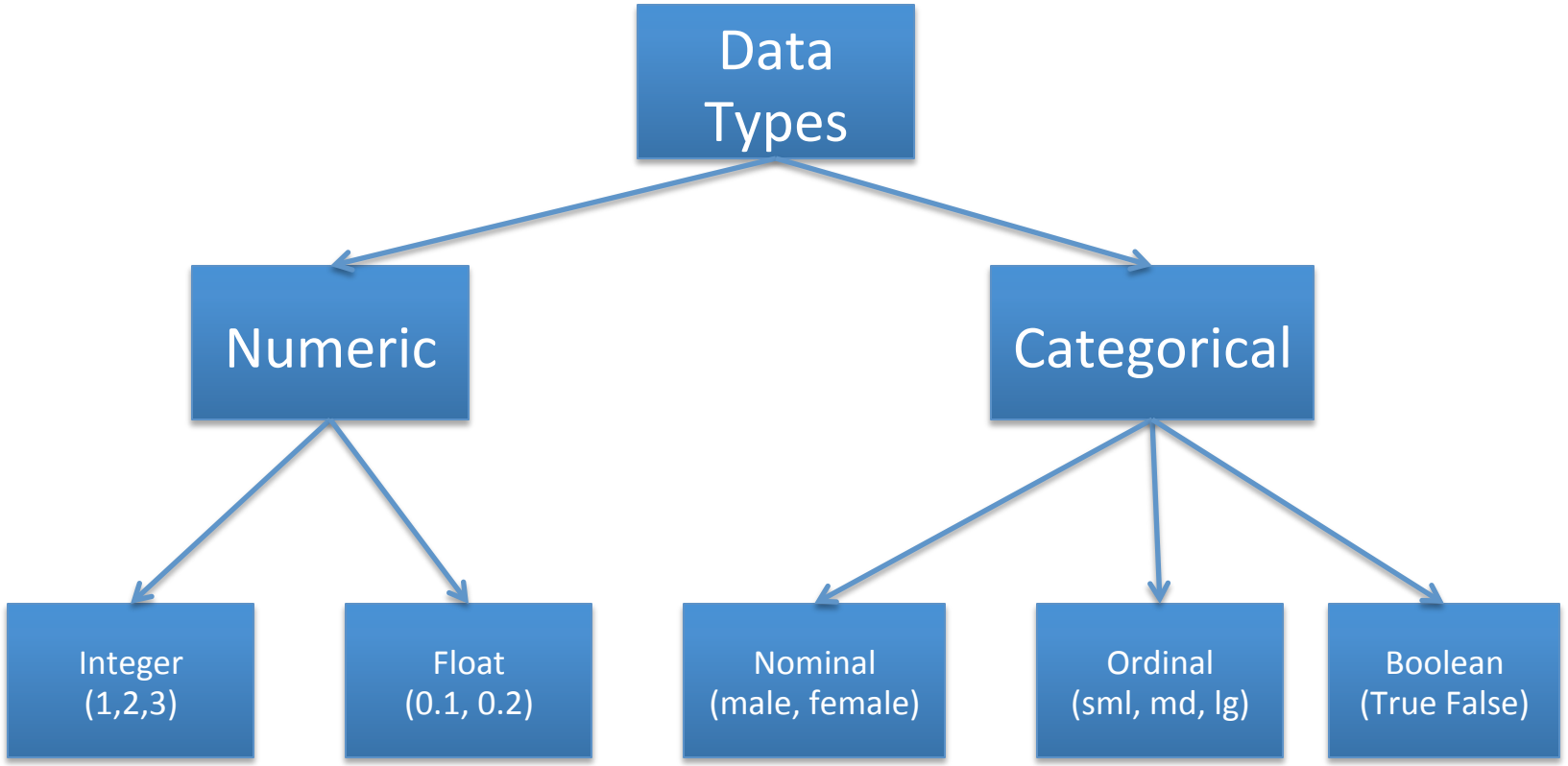
$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that $\text{Max}(|v'|)<1$

Cont'd

Data Transforms

Data transforms are used to change the type or distribution of data variables. Recall that the data may have one of a few types, such as numeric or categorical, with subtypes for each, such as integer and real-valued for numeric, and nominal, ordinal, and boolean for categorical.



Cont'd

We may wish to convert a numeric variable to an ordinal variable in a process called discretization. Alternatively, we may encode a categorical variable as integers or boolean variables, required on most classification tasks.

Discretization Transform: Encode a numeric variable as an ordinal variable.
Ordinal Transform: Encode a categorical variable into an integer variable.
One-Hot Transform: Encode a categorical variable into binary variables.

Cont'd

A variable may not conform to a normal frequency distribution, however, certain data analysis methods may require that the data follow a normal distribution.

To transform the data so that it more closely approximates a normal distribution, it may be necessary to take the *log, exponential* or *Box-Cox* transformation.

Data Reduction

- Data is too big to work with. Complex data analysis and mining on huge amounts of data can take a long time. Making such analysis impractical or infeasible.

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data, i.e, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

- Data reduction strategies
  - Dimensionality reduction — remove unimportant attributes
  - Aggregation and clustering

Understanding Relationship

- A critical step in making sense of data is an understanding of the relationships between different variables. For example, is there any relationship between weather and price of gas.?

- The existence of an association between variables does not imply that one variable causes another.

- These relationships or associations can be established through an examination of different summary tables and data visualizations as well as calculations that measure that measure the strength and confidence in the relationship.

Summary

- Data preparation is a big issue for data mining
- Data preparation includes
  - Data cleaning and data integration
    - Data reduction and feature selection
    - Discretization
- Many methods have been proposed but still an active area of research

- Data preparation is a large subject that can involve a lot of iterations, exploration and analysis. Getting good at data preparation will make life easy for machine learning.