**Data Preprocessing summary**

**DATA PREPROCESSING:**

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

### 1.Data quality assessment:

**Mismatched data types:** When you collect data from many different sources, it may come to you in different formats. While the ultimate goal of this entire process is to reformat your data for machines

**Mixed data values:** Perhaps different sources use different descriptors for features

**Missing data:** Take a look for missing data fields, blank spaces in text, or unanswered survey questions.

### 2. Data cleaning:

Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set.

#### Missing data

**Ignore the tuples:** A tuple is an ordered list or sequence of numbers or entities. If multiple values are missing within tuples, you may simply discard the tuples with that missing information.

**Manually fill in missing data:** This can be tedious, but is definitely necessary when working with smaller data sets.

#### Noisy data

#### This is data that includes unnecessary data points, irrelevant data, and data that’s more difficult to group together.

1. Remove URLs, symbols, emojis, etc., that aren’t relevant to your analysis
2. Translate all text into the language you’ll be working in
3. Remove HTML tags
4. Remove boilerplate email text
5. Remove unnecessary blank text between words
6. Remove duplicate data

**3.**Data profiling****

[Data profiling](https://searchdatamanagement.techtarget.com/definition/data-profiling) is the process of examining, analyzing and reviewing data to collect statistics about its quality.

There are several different tools and methods used for preprocessing data, including the following:

* sampling, which selects a representative subset from a large population of data;
* transformation, which manipulates raw data to produce a single input;
* denoising, which removes [noise](https://whatis.techtarget.com/definition/noise) from data;
* imputation, which synthesizes statistically relevant data for missing values;
* [normalization](https://searchsqlserver.techtarget.com/definition/normalization), which organizes data for more efficient access; and
* feature extraction, which pulls out a relevant feature subset that is significant in a particular context.