DATA CLEANING

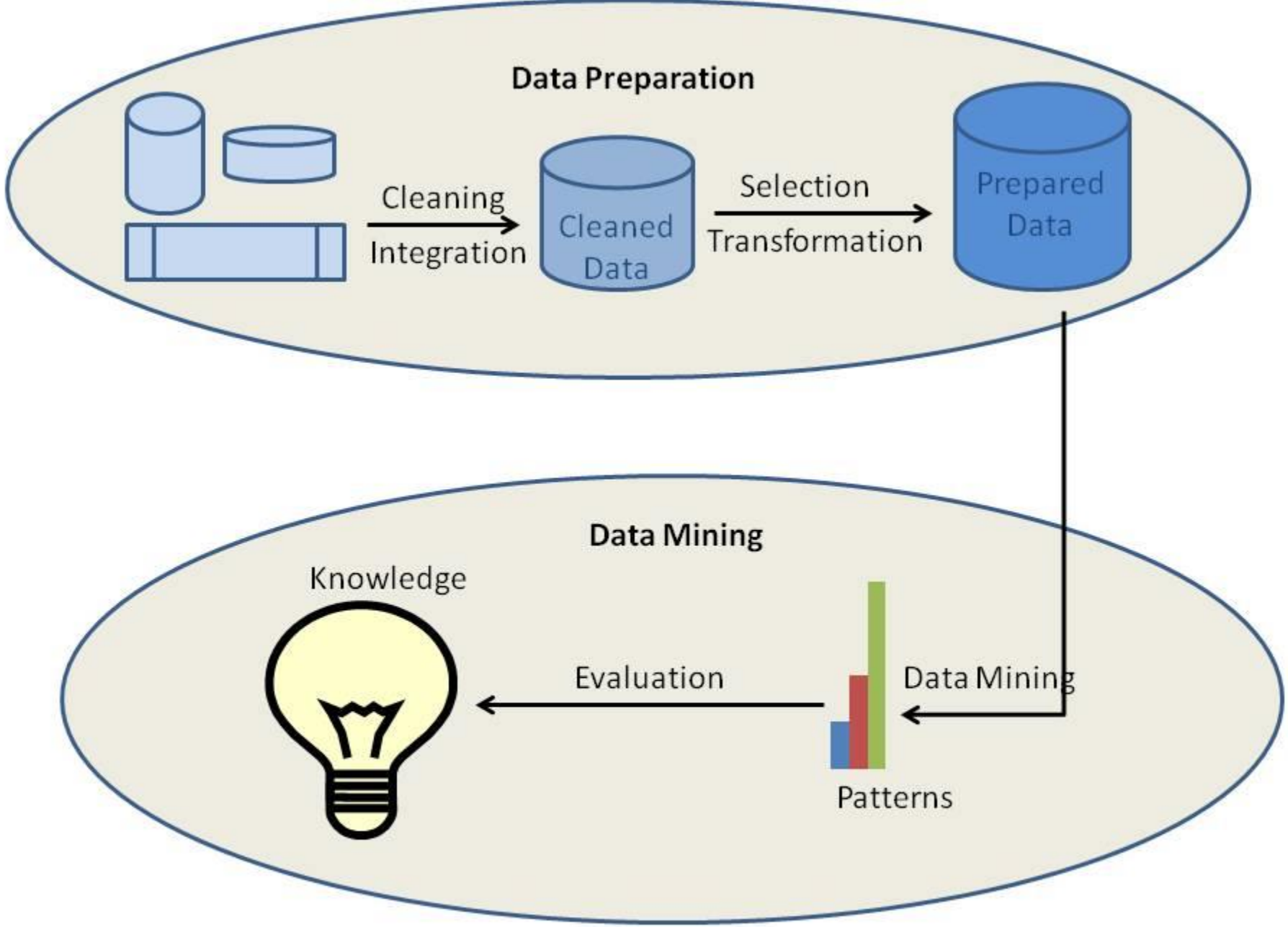
By :

M. Ashish Reddy

DSWP -> Batch-5

- Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model.
- It surely isn’t the fanciest part of machine learning and at the same time, there aren’t any hidden tricks or secrets to uncover.
- However, the success or failure of a project relies on proper data cleaning.
- Professional data scientists usually invest a very large portion of their time in this step because of the belief that **“Better data beats fancier algorithms”**.

- If we have a well-cleaned dataset, there are chances that we can get achieve good results with simple algorithms also, which can prove very beneficial at times especially in terms of computation when the dataset size is large.
- Obviously, different types of data will require different types of cleaning.
- However, this systematic approach can always serve as a good starting point.


Data Preparation
Cleaning
Integration
Cleaned Data
Selection
Transformation
Prepared Data
Data Mining
Knowledge
Evaluation
Data Mining
Patterns

Steps involved in Data Cleaning:

Removal of unwanted observations

This includes deleting duplicate/ redundant or irrelevant values from your dataset. Duplicate observations most frequently arise during data collection and Irrelevant observations are those that don’t actually fit the specific problem that you’re trying to solve.

1. Redundant observations alter the efficiency by a great extent as the data repeats and may add towards the correct side or towards the incorrect side, thereby producing unfaithful results.
2. Irrelevant observations are any type of data that is of no use to us and can be removed directly.

Fixing Structural errors

The errors that arise during measurement, transfer of data, or other similar situations are called structural errors. Structural errors include typos in the name of features, the same attribute with a different name, mislabeled classes, i.e. separate classes that should really be the same, or inconsistent capitalization.

1. For example, the model will treat America and America as different classes or values, though they represent the same value or red, yellow, and red-yellow as different classes or attributes, though one class can be included in the other two classes. So, these are some structural errors that make our model inefficient and give poor quality results.

Managing Unwanted outliers

- Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. Generally, we should not remove outliers until we have a legitimate reason to remove them. Sometimes, removing them improves performance, sometimes not. So, one must have a good reason to remove the outlier, such as suspicious measurements that are unlikely to be part of real data.

Handling missing data

Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important. The two most common ways to deal with missing data are:

1. Dropping observations with missing values.
    1. The fact that the value was missing may be informative in itself.
    2. Plus, in the real world, you often need to make predictions on new data even if some of the features are missing!
2. Imputing the missing values from past observations.
    1. Again, “missingness” is almost always informative in itself, and you should tell your algorithm if a value was missing.
    2. Even if you build a model to impute your values, you’re not adding any real information. You’re just reinforcing the patterns already provided by other features.

Conclusion:

- So, we have discussed four different steps in data cleaning to make the data more reliable and to produce good results.
- After properly completing the Data Cleaning steps, we'll have a robust dataset that avoids many of the most common pitfalls.
- This step should not be rushed as it proves very beneficial in the further process.