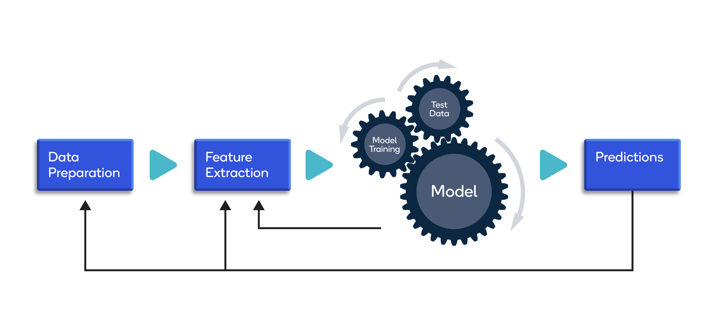
Within the data preparation stage are the data collection and data pre-processing stages.



**Data collection**

Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained. Following are some of the problems that can arise in data collection:

* Inaccurate data. The collected data could be unrelated to the problem statement.
* Missing data. Sub-data could be missing. That could take the form of empty values in columns or missing images for some class of prediction.
* Data imbalance. Some classes or categories in the data may have a disproportionately high or low number of corresponding samples. As a result, they risk being under-represented in the model.
* Data bias. Depending on how the data, subjects and labels themselves are chosen, the model could propagate inherent biases on gender, politics, age or region, for example. Data bias is difficult to detect and remove.

Several techniques can be applied to address those problems:

* Pre-cleaned, freely available datasets. If the problem statement (for example, image classification, object recognition) aligns with a clean, pre-existing, properly formulated dataset, then take advantage of existing, open-source expertise.
* Web crawling and scraping. Automated tools, bots and headless browsers can crawl and scrape websites for data.
* Private data. ML engineers can create their own data. This is helpful when the amount of data required to train the model is small and the problem statement is too specific to generalize over an open-source dataset.
* Custom data. Agencies can create or crowdsource the data for a fee.

**Data pre-processing**

Real-world raw data and images are often incomplete, inconsistent and lacking in certain behaviors or trends. They are also likely to contain many errors. So, once collected, they are pre-processed into a format the machine learning algorithm can use for the model.

Pre-processing includes a number of techniques and actions:

* Data cleaning. These techniques, manual and automated, remove data incorrectly added or classified.
* Data imputations. Most ML frameworks include methods and APIs for balancing or filling in missing data. Techniques generally include imputing missing values with standard deviation, mean, median and k-nearest neighbors (k-NN) of the data in the given field.
* Oversampling. Bias or imbalance in the dataset can be corrected by generating more observations/samples with methods like repetition, bootstrapping or [Synthetic Minority Over-Sampling Technique (SMOTE)](https://towardsdatascience.com/a-deep-dive-into-imbalanced-data-over-sampling-f1167ed74b5), and then adding them to the under-represented classes.
* Data integration. Combining multiple datasets to get a large corpus can overcome incompleteness in a single dataset.
* Data normalization. The size of a dataset affects the memory and processing required for iterations during training. Normalization reduces the size by reducing the order and magnitude of data.