**Step Two: Build a Dataset**

**YouTube Link: *https://www.youtube.com/watch?time\_continue=187&v=PIlWp3w936s&feature=emb\_logo***

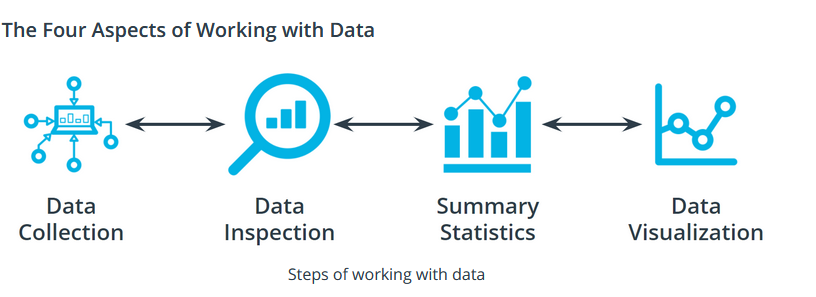
**Summary**

The next step in the machine learning process is to build a dataset that can be used to solve your machine learning-based problem. Understanding the data needed helps you select better models and algorithms so you can build more effective solutions.

**The most important step of the machine learning process**

Working with data is perhaps the most overlooked—yet most important—step of the machine learning process. In 2017, an O’Reilly study showed that machine learning practitioners spend 80% of their time working with their data.

**The Four Aspects of Working with Data**



**Summary**

The next step in the machine learning process is to build a dataset that can be used to solve your machine learning-based problem. Understanding the data needed helps you select better models and algorithms so you can build more effective solutions.

**The most important step of the machine learning process**

Working with data is perhaps the most overlooked—yet most important—step of the machine learning process. In 2017, an O’Reilly study showed that machine learning practitioners spend 80% of their time working with their data.

You can take an entire class just on working with, understanding, and processing data for machine learning applications. Good, high-quality data is essential for any kind of machine learning project. Let's explore some of the common aspects of working with data.

**Data collection**

Data collection can be as straightforward as running the appropriate SQL queries or as complicated as building custom web scraper applications to collect data for your project. You might even have to run a model over your data to generate needed labels. Here is the fundamental question:

Does the data you've collected match the machine learning task and problem you have defined?

**Data inspection**

The quality of your data will ultimately be the largest factor that affects how well you can expect your model to perform. As you inspect your data, look for:

* Outliers
* Missing or incomplete values
* Data that needs to be transformed or preprocessed so it's in the correct format to be used by your model

**Summary statistics**

Models can assume how your data is structured.

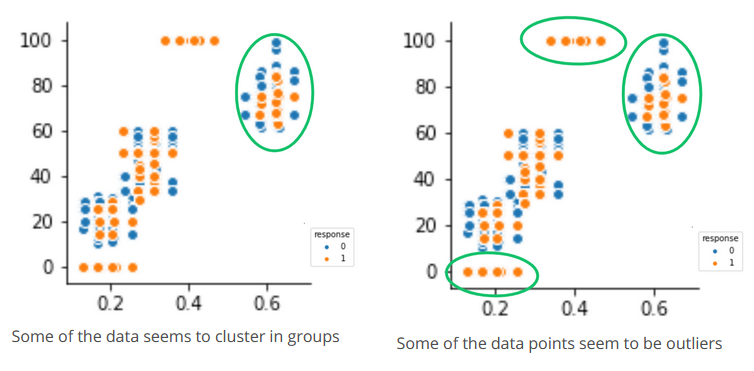
Now that you have some data in hand it is a good best practice to check that your data is in line with the underlying assumptions of your chosen machine learning model.

With many statistical tools, you can calculate things like the mean, inner-quartile range (IQR), and standard deviation. These tools can give you insight into the *scope*, *scale*, and *shape* of the dataset.

**Data visualization**

You can use data visualization to see outliers and trends in your data and to help stakeholders understand your data.

Look at the following two graphs. In the first graph, some data seems to have clustered into different groups. In the second graph, some data points might be outliers.



**Terminology**

* *Impute* is a common term referring to different statistical tools which can be used to calculate missing values from your dataset.
* *Outliers* are data points that are significantly different from others in the same sample.

**Additional reading**

* In machine learning, you use several statistical-based tools to better understand your data. The sklearn library has many examples and tutorials, such as this example demonstrating [outlier detection on a real dataset](https://sklearn.org/auto_examples/applications/plot_outlier_detection_housing.html#sphx-glr-auto-examples-applications-plot-outlier-detection-housing-py).