**Sarkar, T. & Roychowdhury, S. (2019) Data Wrangling with Python. 1st ed. Packt.**

**Chapter 1**

**Introduction to Data Wrangling with Python**

**Importance of Data Wrangling**

Most data scientists spend the majority of their time data wrangling.

Data wrangling is generally done at the very first stage of a data science/analytics pipeline. After the data scientists identify useful data sources for solving the business problem (for instance, in-house database storage or internet or streaming sensor data), they then proceed to extract, clean, and format the necessary data from those sources.

Generally, the task of data wrangling involves the following steps:

* Scraping raw data from multiple sources (including web and database tables)
* Imputing, formatting, and transforming – basically making it ready to be used in the modeling process (such as advanced machine learning)
* Handling read/write errors
* Detecting outliers
* Performing quick visualizations (plotting) and basic statistical analysis to judge the quality of your formatted data

The process of data wrangling includes first finding the appropriate data that's necessary for the analysis. This data can be from one or multiple sources. This data needs to be cleaned. If there is missing data, we will either delete or substitute it, with the help of several techniques. If there are outliers, we need to first detect them and then handle them appropriately. If data is from multiple sources, we will have to perform join operations to combine it.

In an extremely rare situation, data wrangling may not be needed. For example, if the data that's necessary for a machine learning task is already stored in an acceptable format in an in-house database, then a simple SQL query may be enough to extract the data into a table, ready to be passed on to the modelling stage.

**Python for Data Wrangling**

There is always a debate on whether to perform the wrangling process using an enterprise tool or by using a programming language and associated frameworks. There are many commercial, enterprise-level tools for data formatting and pre-processing that do not involve much coding on the part of the user.

However, programming languages such as Python provide more flexibility, control, and power compared to these off-the-shelf tools.

As the volume, velocity, and variety (the three Vs of big data) of data undergo rapid changes, it is always a good idea to develop and nurture a significant amount of in-house expertise in data wrangling using fundamental programming frameworks so that an organization is not beholden to the whims and fancies of any enterprise platform for as basic a task as data wrangling:

A few of the obvious advantages of using an open source, free programming paradigm such as Python for data wrangling are the following:

* General purpose open source paradigm putting no restriction on any of the methods you can develop for the specific problem at hand
* Great ecosystem of fast, optimized, open source libraries, focused on data analytics
* Growing support to connect Python to every conceivable data source type
* Easy interface to basic statistical testing and quick visualization libraries to check data quality
* Seamless interface of the data wrangling output with advanced machine learning models

Python is the most popular language of choice of machine learning and artificial intelligence these days.