Introduction

An overview of industry data science and the scikit-learn API.

In the **Data Preprocessing** section, you will learn data preprocessing techniques with scikit-learn, one of the most popular frameworks used for industry data science.

A. ML engineering vs. data science

In industry, there is quite a bit of overlap between machine learning engineering and data science. Both jobs involve working with data, such as data analysis and data preprocessing.

The main task for machine learning engineers is to first analyze the data for viable trends, then create an efficient input pipeline for training a model. This process involves using libraries like NumPy and pandas for handling data, along with machine learning frameworks like TensorFlow for creating the model and input pipeline. For more information on ML engineering and the NumPy and pandas libraries, check out the previous two sections in this course.

While the NumPy and pandas libraries are also used in data science, the **Data Preprocessing** section will cover one of the core libraries that is specific to industry-level data science: scikit-learn. Data scientists tend to work on smaller datasets than machine learning engineers, and their main goal is to analyze the data and quickly extract usable results. Therefore, they focus more on traditional data inference models (found in scikit-learn), rather than deep neural networks.

The scikit-learn library includes tools for data preprocessing and data mining. It is imported in Python via the statement `import sklearn`.