This chapter begins by talking about how we are entering the era of big data, where there is data present in many facets of everyday life including web pages, videos, genome sequencing, and transactions. The sheer amount of data that is interacted with and processed calls for new automated methods of data analysis, known as machine learning. Machine learning can be defined as a set of methods that can automatically detect patterns in data, then use these discovered patterns to predict future data or otherwise to perform other kinds of decision making that happen under uncertainty, even including the decision of how to collect more data.

The author asserts that probability theory is the best tool with which to solve the problems incurred by machine learning, given that it can be applied to any problem involving uncertainty, which can present itself in a variety of different forms, including the best way to predict the future based on past data, the best model to explain a certain set of data, or the best measurement to collect in order to explain some data. While probability theory as an approach does have some similarities to statistics, it slightly differs in the terms that it uses and the concepts that it focuses on.

Even in very large data sets, the amount of data points that are actually useful might be significantly smaller, a property which is known as long tail, where a few things are common but most things are rare. An example of long tail is the fact that the vast majority of Google searches each day have never been searched previously.

Machine learning can be divided into two main types. The first of these is called predictive or supervised learning, which involves learning a mapping from inputs to outputs given a labeled set of input-output pairs, called a training set, which has a certain number of training examples. Each training input in this set is generally a vector of numbers representing values called features, attributes, or covariates, although they can also represent more complex objects like images, sentences, or graphs. The output, also known as a response variable, can also be any type of data, but it is generally assumed that it is a categorical or nominal variable from a finite set, or that it is a real-valued scalar. When the output is categorical, the problem is called classification or pattern recognition, and when it is real-valued, it is called regression. When there are two class labels, mapping from inputs to outputs is called binary classification, but if there are more than two, it is called multi-label classification, which is best viewed as a multiple output model.

The second type of machine learning is called descriptive or unsupervised learning, which is where we are only given inputs and the goal is to find interesting patterns in the data, which is also sometimes called knowledge discovery, and is overall less well-defined since it does not specify which patterns to look for and there is no obvious error metric as there is in supervised learning.