CHAPTER 1

An Introduction to Data Analysis

In this chapter, you begin to take the first steps in the world of data analysis, learning in detail about all the concepts and processes that make up this discipline. The concepts discussed in this chapter are helpful background for the following chapters, where these concepts and procedures will be applied in the form of Python code, through the use of several libraries that will be discussed in just as many chapters.

Data Analysis

In a world increasingly centralized around information technology, huge amounts of data are produced and stored each day. Often these data come from automatic detection systems, sensors, and scientific instrumentation, or you produce them daily and unconsciously every time you make a withdrawal from the bank or make a purchase, when you record various blogs, or even when you post on social networks.

But what are the data? The data actually are not information, at least in terms of their form. In the formless stream of bytes, at first glance it is difficult to understand their essence if not strictly the number, word, or time that they report. Information is actually the result of processing, which, taking into account a certain dataset, extracts some conclusions that can be used in various ways. This process of extracting information from raw data is called *data analysis*.

The purpose of data analysis is to extract information that is not easily deducible but that, when understood, leads to the possibility of carrying out studies on the mechanisms of the systems that have produced them, thus allowing you to forecast possible responses of these systems and their evolution in time.

F. Nelli, *Python Data Analytics*, https://doi.org/10.1007/978-1-4842-3913-1_1

Starting from a simple methodical approach on data protection, data analysis has become a real discipline, leading to the development of real methodologies generating *models*. The model is in fact the translation into a mathematical form of a system placed under study. Once there is a mathematical or logical form that can describe system responses under different levels of precision, you can then make predictions about its development or response to certain inputs. Thus the aim of data analysis is not the model, but the quality of its *predictive power*.

The predictive power of a model depends not only on the quality of the modeling techniques but also on the ability to choose a good dataset upon which to build the entire data analysis process. So the *search for data*, their *extraction***,** and their subsequent *preparation***,** while representing preliminary activities of an analysis, also belong to data analysis itself, because of their importance in the success of the results.

So far we have spoken of data, their handling, and their processing through calculation procedures. In parallel to all stages of processing of data analysis, various methods of *data visualization* have been developed. In fact, to understand the data, both individually and in terms of the role they play in the entire dataset, there is no better system than to develop the techniques of graphic representation capable of transforming information, sometimes implicitly hidden, in figures, which help you more easily understand their meaning. Over the years lots of display modes have been developed for different modes of data display: the *charts*.

At the end of the data analysis process, you will have a model and a set of graphical displays and then you will be able to predict the responses of the system under study; after that, you will move to the test phase. The model will be tested using another set of data for which you know the system response. These data are, however, not used to define the predictive model. Depending on the ability of the model to replicate real observed responses, you will have an error calculation and knowledge of the validity of the model and its operating limits.

These results can be compared with any other models to understand if the newly created one is more efficient than the existing ones. Once you have assessed that, you can move to the last phase of data analysis—*deployment*. This consists of implementing the results produced by the analysis, namely, implementing the decisions to be taken based on the predictions generated by the model and the associated risks.

Data analysis is well suited to many professional activities. So, knowledge of it and how it can be put into practice is relevant. It allows you to test hypotheses and to understand more deeply the systems analyzed.

Knowledge Domains of the Data Analyst

Data analysis is basically a discipline suitable to the study of problems that may occur in several fields of applications. Moreover, data analysis includes many tools and methodologies that require good knowledge of computing, mathematical, and statistical concepts.

A good data analyst must be able to move and act in many different disciplinary areas. Many of these disciplines are the basis of the methods of data analysis, and proficiency in them is almost necessary. Knowledge of other disciplines is necessary depending on the area of application and study of the particular data analysis project you are about to undertake, and, more generally, sufficient experience in these areas can help you better understand the issues and the type of data needed.

Often, regarding major problems of data analysis, it is necessary to have an interdisciplinary team of experts who can contribute in the best possible way in their respective fields of competence. Regarding smaller problems, a good analyst must be able to recognize problems that arise during data analysis, inquire to determine which disciplines and skills are necessary to solve these problems, study these disciplines, and maybe even ask the most knowledgeable people in the sector. In short, the analyst must be able to know how to search not only for data, but also for information on how to treat that data.

Computer Science

Knowledge of computer science is a basic requirement for any data analyst. In fact, only when you have good knowledge of and experience in computer science can you efficiently manage the necessary tools for data analysis. In fact, every step concerning data analysis involves using calculation software (such as IDL, MATLAB, etc.) and programming languages (such as C ++, Java, and Python).

The large amount of data available today, thanks to information technology, requires specific skills in order to be managed as efficiently as possible. Indeed, data research and extraction require knowledge of these various formats. The data are structured and stored in files or database tables with particular formats. XML, JSON, or simply XLS or CSV files, are now the common formats for storing and collecting data, and many applications allow you to read and manage the data stored on them. When it comes to extracting data contained in a database, things are not so immediate, but you need to know the SQL query language or use software specially developed for the extraction of data from a given database.

Moreover, for some specific types of data research, the data are not available in an explicit format, but are present in text files (documents and log files) or web pages, and shown as charts, measures, number of visitors, or HTML tables. This requires specific technical expertise for the parsing and the eventual extraction of these data (called *web scraping*).

So, knowledge of information technology is necessary to know how to use the various tools made available by contemporary computer science, such as applications and programming languages. These tools, in turn, are needed to perform data analysis and data visualization.

The purpose of this book is to provide all the necessary knowledge, as far as possible, regarding the development of methodologies for data analysis. The book uses the Python programming language and specialized libraries that provide a decisive contribution to the performance of all the steps constituting data analysis, from data research to data mining, to publishing the results of the predictive model.

Mathematics and Statistics

As you will see throughout the book, data analysis requires a lot of complex math during the treatment and processing of data. You need to be competent in all of this, at least to understand what you are doing. Some familiarity with the main statistical concepts is also necessary because all the methods that are applied in the analysis and interpretation of data are based on these concepts. Just as you can say that computer science gives you the tools for data analysis, so you can say that the statistics provide the concepts that form the basis of data analysis.

This discipline provides many tools to the analyst, and a good knowledge of how to best use them requires years of experience. Among the most commonly used statistical techniques in data analysis are

- Bayesian methods
- Regression
- Clustering

Having to deal with these cases, you'll discover how mathematics and statistics are closely related. Thanks to the special Python libraries covered in this book, you will be able to manage and handle them.

Machine Learning and Artificial Intelligence

One of the most advanced tools that falls in the data analysis camp is machine learning. In fact, despite the data visualization and techniques such as clustering and regression, which should help you find information about the dataset, during this phase of research, you may often prefer to use special procedures that are highly specialized in searching patterns within the dataset.

Machine learning is a discipline that uses a whole series of procedures and algorithms that analyze the data in order to recognize patterns, clusters, or trends and then extracts useful information for data analysis in an automated way.

This discipline is increasingly becoming a fundamental tool of data analysis, and thus knowledge of it, at least in general, is of fundamental importance to the data analyst.

Professional Fields of Application

Another very important point is the domain of competence of the data (its source—biology, physics, finance, materials testing, statistics on population, etc.). In fact, although analysts have had specialized preparation in the field of statistics, they must also be able to document the source of the data, with the aim of perceiving and better understanding the mechanisms that generated the data. In fact, the data are not simple strings or numbers; they are the expression, or rather the measure, of any parameter observed. Thus, better understanding where the data came from can improve their interpretation. Often, however, this is too costly for data analysts, even ones with the best intentions, and so it is good practice to find consultants or key figures to whom you can pose the right questions.

Understanding the Nature of the Data

The object of study of data analysis is basically the data. The data then will be the key player in all processes of data analysis. The data constitute the raw material to be processed, and thanks to their processing and analysis, it is possible to extract a variety of information in order to increase the level of knowledge of the system under study, that is, one from which the data came.

When the Data Become Information

Data are the events recorded in the world. Anything that can be measured or categorized can be converted into data. Once collected, these data can be studied and analyzed, both to understand the nature of the events and very often also to make predictions or at least to make informed decisions.

When the Information Becomes Knowledge

You can speak of knowledge when the information is converted into a set of rules that helps you better understand certain mechanisms and therefore make predictions on the evolution of some events.

Types of Data

Data can be divided into two distinct categories:

- Categorical (nominal and ordinal)
- Numerical (discrete and continuous)

Categorical data are values or observations that can be divided into groups or categories. There are two types of categorical values: *nominal* and *ordinal*. A nominal variable has no intrinsic order that is identified in its category. An ordinal variable instead has a predetermined order.

Numerical data are values or observations that come from measurements. There are two types of numerical values: *discrete* and *continuous* numbers. Discrete values can be counted and are distinct and separated from each other. Continuous values, on the other hand, are values produced by measurements or observations that assume any value within a defined range.

The Data Analysis Process

Data analysis can be described as a process consisting of several steps in which the raw data are transformed and processed in order to produce data visualizations and make predictions thanks to a mathematical model based on the collected data. Then, data

analysis is nothing more than a sequence of steps, each of which plays a key role in the subsequent ones. So, data analysis is schematized as a process chain consisting of the following sequence of stages:

- Problem definition
- Data extraction
- Data preparation - Data cleaning
- Data preparation - Data transformation
- Data exploration and visualization
- Predictive modeling
- Model validation/test
- Deploy - Visualization and interpretation of results
- Deploy - Deployment of the solution

Figure 1-1 shows a schematic representation of all the processes involved in the data analysis.

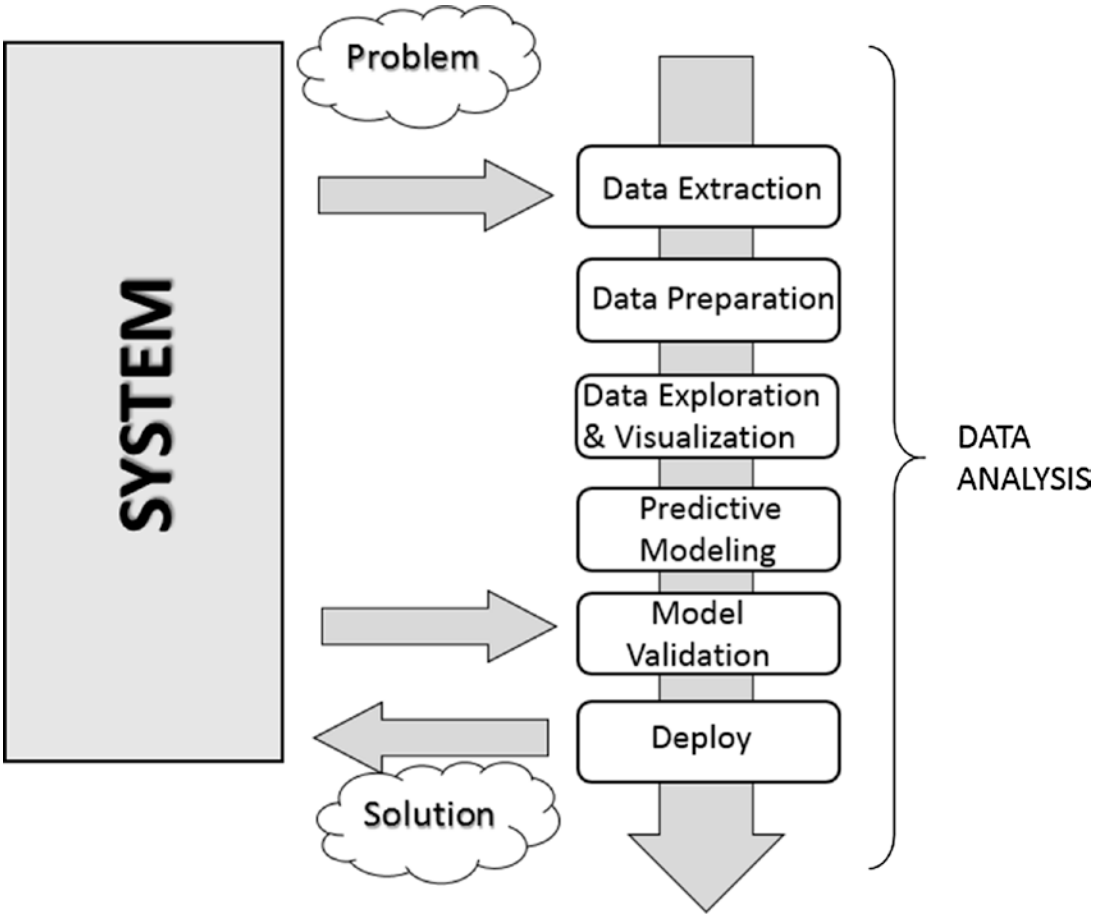

Figure 1-1. *The data analysis process*

Problem Definition

The process of data analysis actually begins long before the collection of raw data. In fact, data analysis always starts with a problem to be solved, which needs to be defined.

The problem is defined only after you have focused the system you want to study; this may be a mechanism, an application, or a process in general. Generally this study can be in order to better understand its operation, but in particular the study will be designed to understand the principles of its behavior in order to be able to make predictions or choices (defined as an informed choice).

The definition step and the corresponding documentation (*deliverables*) of the scientific problem or business are both very important in order to focus the entire analysis strictly on getting results. In fact, a comprehensive or exhaustive study of the

system is sometimes complex and you do not always have enough information to start with. So the definition of the problem and especially its planning can determine the guidelines to follow for the whole project.

Once the problem has been defined and documented, you can move to the *project planning* stage of data analysis. Planning is needed to understand which professionals and resources are necessary to meet the requirements to carry out the project as efficiently as possible. So you're going to consider the issues in the area involving the resolution of the problem. You will look for specialists in various areas of interest and install the software needed to perform data analysis.

Also during the planning phase, you choose an effective team. Generally, these teams should be cross-disciplinary in order to solve the problem by looking at the data from different perspectives. So, building a good team is certainly one of the key factors leading to success in data analysis.

Data Extraction

Once the problem has been defined, the first step is to obtain the data in order to perform the analysis. The data must be chosen with the basic purpose of building the predictive model, and so data selection is crucial for the success of the analysis as well. The sample data collected must reflect as much as possible the real world, that is, how the system responds to stimuli from the real world. For example, if you're using huge datasets of raw data and they are not collected competently, these may portray false or unbalanced situations.

Thus, poor choice of data, or even performing analysis on a dataset that's not perfectly representative of the system, will lead to models that will move away from the system under study.

The search and retrieval of data often require a form of intuition that goes beyond mere technical research and data extraction. This process also requires a careful understanding of the nature and form of the data, which only good experience and knowledge in the problem's application field can provide.

Regardless of the quality and quantity of data needed, another issue is using the best *data sources.*

If the studio environment is a laboratory (technical or scientific) and the data generated are experimental, then in this case the data source is easily identifiable. In this case, the problems will be only concerning the experimental setup.

But it is not possible for data analysis to reproduce systems in which data are gathered in a strictly experimental way in every field of application. Many fields require searching for data from the surrounding world, often relying on external experimental data, or even more often collecting them through interviews or surveys. So in these cases, finding a good data source that is able to provide all the information you need for data analysis can be quite challenging. Often it is necessary to retrieve data from multiple data sources to supplement any shortcomings, to identify any discrepancies, and to make the dataset as general as possible.

When you want to get the data, a good place to start is the Web. But most of the data on the Web can be difficult to capture; in fact, not all data are available in a file or database, but might be content that is inside HTML pages in many different formats. To this end, a methodology called *web scraping* allows the collection of data through the recognition of specific occurrence of HTML tags within web pages. There is software specifically designed for this purpose, and once an occurrence is found, it extracts the desired data. Once the search is complete, you will get a list of data ready to be subjected to data analysis.

Data Preparation

Among all the steps involved in data analysis, data preparation, although seemingly less problematic, in fact requires more resources and more time to be completed. Data are often collected from different data sources, each of which will have data in it with a different representation and format. So, all of these data will have to be prepared for the process of data analysis.

The preparation of the data is concerned with obtaining, cleaning, normalizing, and transforming data into an optimized dataset, that is, in a prepared format that's normally tabular and is suitable for the methods of analysis that have been scheduled during the design phase.

Many potential problems can arise, including invalid, ambiguous, or missing values, replicated fields, and out-of-range data.

Data Exploration/Visualization

Exploring the data involves essentially searching the data in a graphical or statistical presentation in order to find patterns, connections, and relationships. Data visualization is the best tool to highlight possible patterns.

In recent years, data visualization has been developed to such an extent that it has become a real discipline in itself. In fact, numerous technologies are utilized exclusively to display data, and many display types are applied to extract the best possible information from a dataset.

Data exploration consists of a preliminary examination of the data, which is important for understanding the type of information that has been collected and what it means. In combination with the information acquired during the definition problem, this categorization will determine which method of data analysis will be most suitable for arriving at a model definition.

Generally, this phase, in addition to a detailed study of charts through the visualization data, may consist of one or more of the following activities:

- Summarizing data
- Grouping data
- Exploring the relationship between the various attributes
- Identifying patterns and trends
- Constructing regression models
- Constructing classification models

Generally, data analysis requires summarizing statements regarding the data to be studied. *Summarization* is a process by which data are reduced to interpretation without sacrificing important information.

Clustering is a method of data analysis that is used to find groups united by common attributes (also called *grouping*).

Another important step of the analysis focuses on the *identification* of relationships, trends, and anomalies in the data. In order to find this kind of information, you often have to resort to the tools as well as perform another round of data analysis, this time on the data visualization itself.

Other methods of data mining, such as decision trees and association rules, automatically extract important facts or rules from the data. These approaches can be used in parallel with data visualization to uncover relationships between the data.

Predictive Modeling

Predictive modeling is a process used in data analysis to create or choose a suitable statistical model to predict the probability of a result.

After exploring the data, you have all the information needed to develop the mathematical model that encodes the relationship between the data. These models are useful for understanding the system under study, and in a specific way they are used for two main purposes. The first is to make predictions about the data values produced by the system; in this case, you will be dealing with *regression models.* The second purpose is to classify new data products, and in this case, you will be using *classification models* or *clustering models.* In fact, it is possible to divide the models according to the type of result they produce:

- *Classification models*: If the result obtained by the model type is categorical.
- *Regression models*: If the result obtained by the model type is numeric.
- *Clustering models*: If the result obtained by the model type is descriptive.

Simple methods to generate these models include techniques such as linear regression, logistic regression, classification and regression trees, and k-nearest neighbors. But the methods of analysis are numerous, and each has specific characteristics that make it excellent for some types of data and analysis. Each of these methods will produce a specific model, and then their choice is relevant to the nature of the product model.

Some of these models will provide values corresponding to the real system and according to their structure. They will explain some characteristics of the system under study in a simple and clear way. Other models will continue to give good predictions, but their structure will be no more than a "black box" with limited ability to explain characteristics of the system.

Model Validation

Validation of the model, that is, the test phase, is an important phase that allows you to validate the model built on the basis of starting data. That is important because it allows you to assess the validity of the data produced by the model by comparing them directly with the actual system. But this time, you are coming out from the set of starting data on which the entire analysis has been established.

Generally, you will refer to the data as the *training set* when you are using them for building the model, and as the *validation set* when you are using them for validating the model.

Thus, by comparing the data produced by the model with those produced by the system, you will be able to evaluate the error, and using different test datasets, you can estimate the limits of validity of the generated model. In fact the correctly predicted values could be valid only within a certain range, or have different levels of matching depending on the range of values taken into account.

This process allows you not only to numerically evaluate the effectiveness of the model but also to compare it with any other existing models. There are several techniques in this regard; the most famous is the *cross-validation*. This technique is based on the division of the training set into different parts. Each of these parts, in turn, will be used as the validation set and any other as the training set. In this iterative manner, you will have an increasingly perfected model.

Deployment

This is the final step of the analysis process, which aims to present the results, that is, the conclusions of the analysis. In the deployment process of the business environment, the analysis is translated into a benefit for the client who has commissioned it. In technical or scientific environments, it is translated into design solutions or scientific publications. That is, the deployment basically consists of putting into practice the results obtained from the data analysis.

There are several ways to deploy the results of data analysis or data mining. Normally, a data analyst's deployment consists in writing a report for management or for the customer who requested the analysis. This document will conceptually describe the results obtained from the analysis of data. The report should be directed to the managers, who are then able to make decisions. Then, they will put into practice the conclusions of the analysis.

In the documentation supplied by the analyst, each of these four topics will be discussed in detail:

- Analysis results
- Decision deployment
- Risk analysis
- Measuring the business impact

When the results of the project include the generation of predictive models, these models can be deployed as stand-alone applications or can be integrated into other software.

Quantitative and Qualitative Data Analysis

Data analysis is completely focused on data. Depending on the nature of the data, it is possible to make some distinctions.

When the analyzed data have a strictly numerical or categorical structure, then you are talking about *quantitative analysis*, but when you are dealing with values that are expressed through descriptions in natural language, then you are talking about *qualitative analysis*.

Precisely because of the different nature of the data processed by the two types of analyses, you can observe some differences between them.

Quantitative analysis has to do with data with a logical order or that can be categorized in some way. This leads to the formation of structures within the data. The order, categorization, and structures in turn provide more information and allow further processing of the data in a more mathematical way. This leads to the generation of models that provide *quantitative predictions,* thus allowing the data analyst to draw more objective conclusions.

Qualitative analysis instead has to do with data that generally do not have a structure, at least not one that is evident, and their nature is neither numeric nor categorical. For example, data under qualitative study could include written textual, visual, or audio data. This type of analysis must therefore be based on methodologies, often *ad hoc,* to extract information that will generally lead to models capable of providing *qualitative predictions***,** with the result that the conclusions to which the data analyst can arrive may also include *subjective interpretations*. On the other hand, qualitative analysis

can explore more complex systems and draw conclusions that are not possible using a strictly mathematical approach. Often this type of analysis involves the study of systems such as social phenomena or complex structures that are not easily measurable.

Figure 1-2 shows the differences between the two types of analysis.

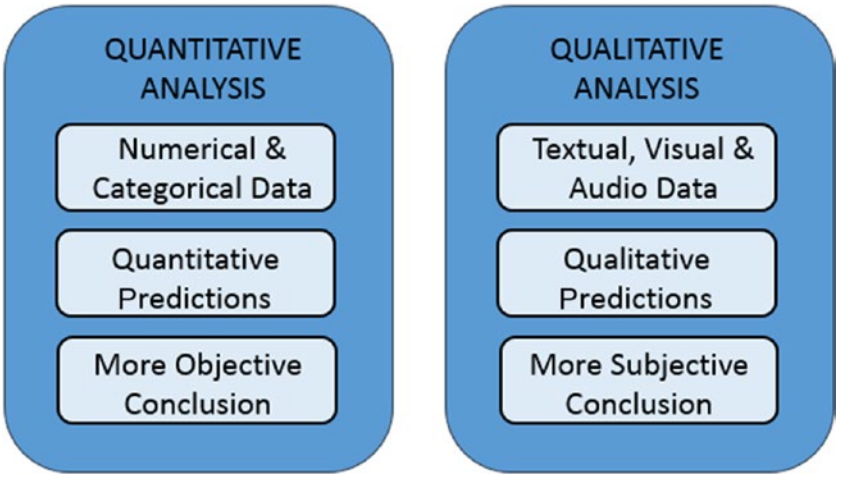

Figure 1-2. *Quantitative and qualitative analyses*

Open Data

In support of the growing demand for data, a huge number of data sources are now available on the Internet. These data sources freely provide information to anyone in need, and they are called *open data*.

Here is a list of some open data available online. You can find a more complete list and details of the open data available online in Appendix B.

- DataHub (`http://datahub.io/dataset`)
- World Health Organization (`http://www.who.int/research/en/`)
- Data.gov (`http://data.gov`)
- European Union Open Data Portal (`http://open-data.europa.eu/en/data/`)
- Amazon Web Service public datasets (`http://aws.amazon.com/datasets`)
- Facebook Graph (`http://developers.facebook.com/docs/graph-api`)

- Healthdata.gov (http://www.healthdata.gov)
- Google Trends (http://www.google.com/trends/explore)
- Google Finance (https://www.google.com/finance)
- Google Books Ngrams (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html)
- Machine Learning Repository (http://archive.ics.uci.edu/ml/)

As an idea of open data sources available online, you can look at the *LOD cloud diagram* (http://lod-cloud.net), which displays the connections of the data link among several open data sources currently available on the network (see Figure 1-3).

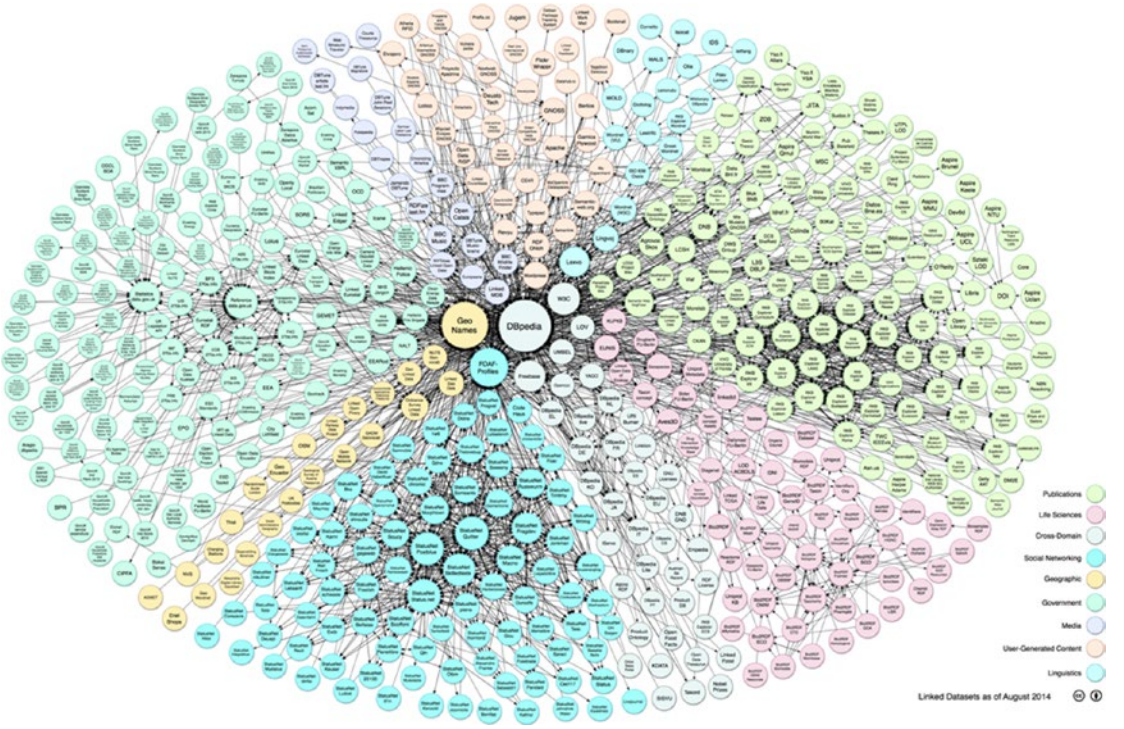

Figure 1-3. *Linking open data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch, and Richard Cyganiak.* http://lod-cloud.net/ *[CC-BY-SA license]*

Python and Data Analysis

The main argument of this book is to develop all the concepts of data analysis by treating them in terms of Python. The Python programming language is widely used in scientific circles because of its large number of libraries that provide a complete set of tools for analysis and data manipulation.

Compared to other programming languages generally used for data analysis, such as R and MATLAB, Python not only provides a platform for processing data, but also has features that make it unique compared to other languages and specialized applications. The development of an ever-increasing number of support libraries, the implementation of algorithms of more innovative methodologies, and the ability to interface with other programming languages (C and Fortran) all make Python unique among its kind.

Furthermore, Python is not only specialized for data analysis, but also has many other applications, such as generic programming, scripting, interfacing to databases, and more recently web development, thanks to web frameworks like Django. So it is possible to develop data analysis projects that are compatible with the web server with the possibility to integrate it on the Web.

So, for those who want to perform data analysis, Python, with all its packages, is considered the best choice for the foreseeable future.

Conclusions

In this chapter, you learned what data analysis is and, more specifically, the various processes that comprise it. Also, you have begun to see the role that data play in building a prediction model and how their careful selection is at the basis of a careful and accurate data analysis.

In the next chapter, you will take this vision of Python and the tools it provides to perform data analysis.