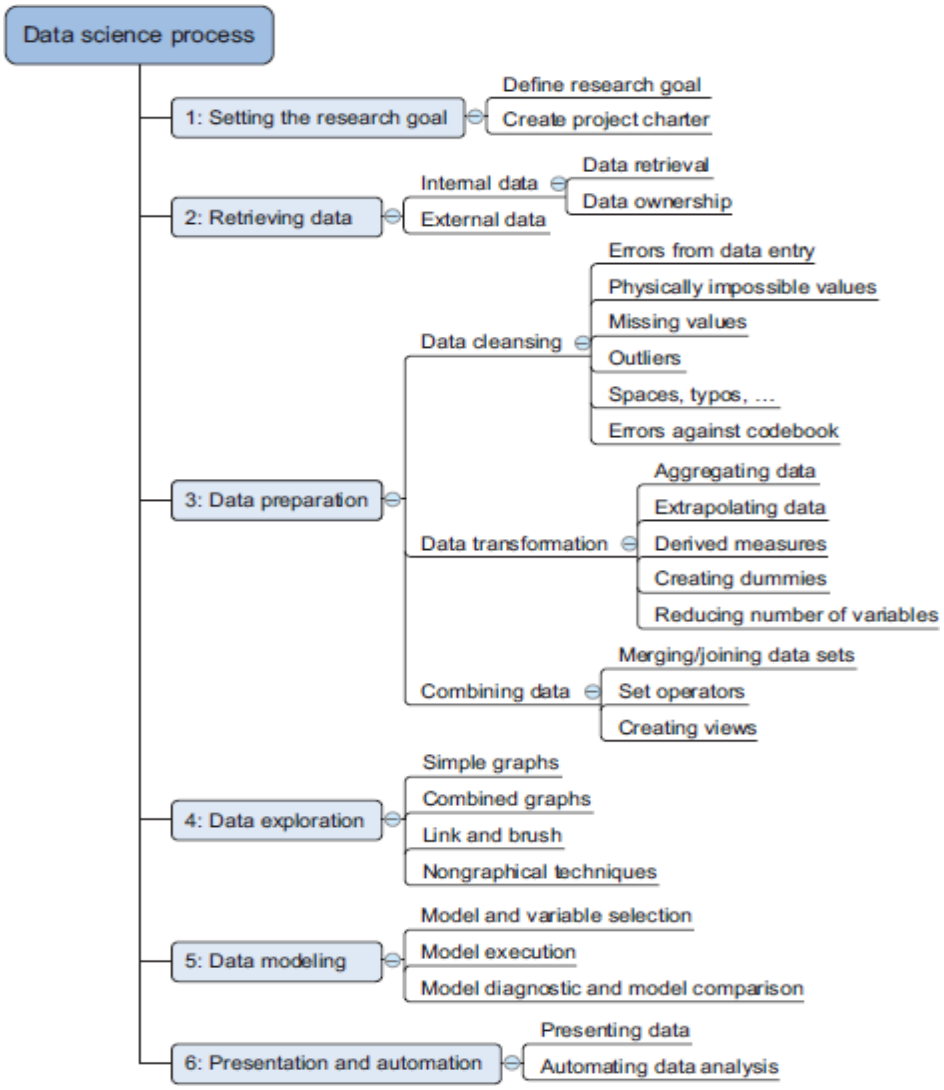
OVERVIEW OF THE DATA SCIENCE PROCESS

The typical data science process consists of six steps through which you'll iterate, as shown in figure



- The first step of this process is setting a research goal. The main purpose here is making sure all the stakeholders understand the what, how, and why of the project. In every serious project this will result in a project charter.
- The second phase is data retrieval. You want to have data available for analysis, so this step includes finding suitable data and getting access to the data from the data owner. The result is data in its raw form, which probably needs polishing and transformation before it becomes usable.
- Now that you have the raw data, it's time to prepare it. This includes transforming the data from a raw form into data that's directly usable in your models. To achieve this, you'll detect and correct

different kinds of errors in the data, combine data from different data sources, and transform it. If you have successfully completed this step, you can progress to data visualization and modeling.

- The fourth step is data exploration. The goal of this step is to gain a deep understanding of the data. You'll look for patterns, correlations, and deviations based on visual and descriptive techniques. The insights you gain from this phase will enable you to start modeling.
- Finally, we get to model building (often referred to as "data modeling" throughout this book). It is now that you attempt to gain the insights or make the predictions stated in your project charter. Now is the time to bring out the heavy guns, but remember research has taught us that often (but not always) a combination of simple models tends to outperform one complicated model. If you've done this phase right, you're almost done.
- The last step of the data science model is presenting your results and automating the analysis, if needed. One goal of a project is to change a process and/or make better decisions. You may still need to convince the business that your findings will indeed change the business process as expected. This is where you can shine in your influencer role. The importance of this step is more apparent in projects on a strategic and tactical level. Certain projects require you to perform the business process over and over again, so automating the project will save time.

DEFINING RESEARCH GOALS

A project starts by understanding the what, the why, and the how of your project. The outcome should be a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable. This information is then best placed in a project charter.

Spend time understanding the goals and context of your research:

- An essential outcome is the research goal that states the purpose of your assignment in a clear and focused manner.
- Understanding the business goals and context is critical for project success.
- Continue asking questions and devising examples until you grasp the exact business expectations, identify how your project fits in the bigger picture, appreciate how your research is going to change the business, and understand how they'll use your results

Create a project charter

A project charter requires teamwork, and your input covers at least the following:

- A clear research goal
- The project mission and context
- How you're going to perform your analysis
- What resources you expect to use
- Proof that it's an achievable project, or proof of concepts

- Deliverables and a measure of success
- A timeline

RETRIEVING DATA

- The next step in data science is to retrieve the required data. Sometimes you need to go into the field and design a data collection process yourself, but most of the time you won't be involved in this step.
- Many companies will have already collected and stored the data for you, and what they don't have can often be bought from third parties.
- More and more organizations are making even high-quality data freely available for public and commercial use.
- Data can be stored in many forms, ranging from simple text files to tables in a database. The objective now is acquiring all the data you need.

Start with data stored within the company (Internal data)

- Most companies have a program for maintaining key data; so much of the cleaning work may already be done. This data can be stored in official data repositories such as databases, data marts, data warehouses, and data lakes maintained by a team of IT professionals.
- Data warehouses and data marts are home to pre-processed data, data lakes contain data in its natural or raw format.
- Finding data even within your own company can sometimes be a challenge. As companies grow, their data becomes scattered around many places. The data may be dispersed as people change positions and leave the company.
- Getting access to data is another difficult task. Organizations understand the value and sensitivity of data and often have policies in place so everyone has access to what they need and nothing more.
- These policies translate into physical and digital barriers called Chinese walls. These "walls" are mandatory and well-regulated for customer data in most countries.

External Data

- If data isn't available inside your organization, look outside your organizations. Companies provide data so that you, in turn, can enrich their services and ecosystem. Such is the case with Twitter, LinkedIn, and Facebook.
- More and more governments and organizations share their data for free with the world.
- A list of open data providers that should get you started.