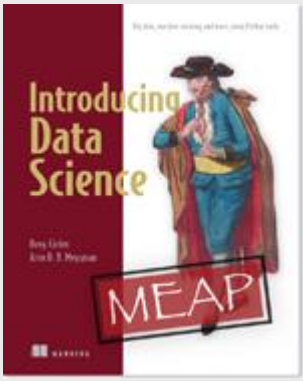

Overview of the data science process

By Davy Cielen

In this article, I'll discuss the six steps involved in the typical data science process.

Following a structured approach to data science helps you to maximize your chances of success in a data science project at the lowest cost. It also makes it possible to take up a project as a team, with each team member focusing on what he or she does best. Take care however: this approach might not be suitable for every type of project or the only way to do good data science.

The typical data science process consists of six steps through which you will iterate, as shown in figure 1.

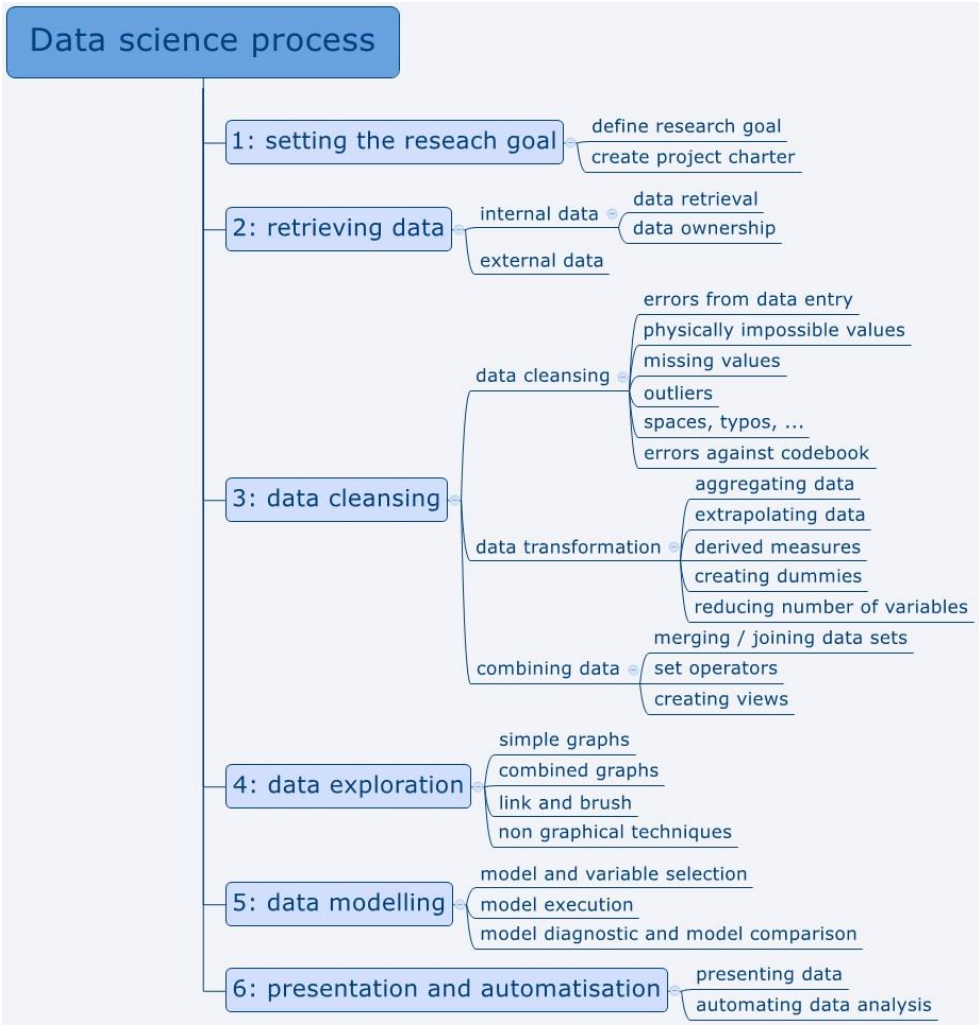

Figure 1 The six steps of the data science process

Figure 1 summarizes the data science process and shows the main steps and actions you will take during a project.

1. The first step of this process is setting a *research goal*. The main purpose here is to make sure all the stakeholders understand the *what*, *how*, and *why* of the project. In every serious project this will result in a project charter.

2. 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.
3. The result is data in its raw form, which probably needs some polishing and transformation before it becomes usable.
4. Now that you have the raw data, it is time to *cleanse* it. This includes transforming the data from a raw form into data that is directly usable in your models. To achieve this, you will detect and correct different kinds of errors in the model, combine data from different data sources, and transform it. If you have successfully completed this step, you can progress to data visualization and modeling.
5. The fourth step is *data exploration*. The goal of this step is to gain a deep understanding of the data. You will 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.
6. Finally we get to the sexiest part: *data modeling*. It is now you attempt to gain the insights or make the predictions that were 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 have done this phase right, you are almost done.
7. 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 the process and/or make better decisions. You might 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. Some projects require you to perform the business process over and over again, so automating the project will save you lots of time.

In reality you will not progress in a linear way from step 1 to step 6; often you will regress and iterate between the different phases.

Following these six steps pays off in terms of a higher project success ratio and increased impact of research results. This process ensures you have a well-defined research plan, a good understanding of the business question, and clear deliverables before you even start looking at data. The first steps of your process focus on getting high-quality data as input for your models. This way your models will perform better later on. In data science there's a wellknown paradigm: *Garbage in equals garbage out*.

Another benefit of following a structured approach is that you work more in *prototype mode* while you search for the best model. When building a *prototype* you will probably try multiple models and won't focus heavily on things like program speed or writing code against standards. This allows you to focus on bringing business value instead.

Not every project is initiated by the business itself. Insights learned during analysis or the arrival of new data can spawn new projects. When the data science team generates an idea, it means some work has already been done to make a proposition and find a business sponsor.

Dividing a project into smaller stages also allows employees to work together as a team. It is impossible to be a specialist in everything. You would need to know how to upload all the data to all the different databases, find an optimal data scheme that works not only for your application but also for other projects inside your company, and then keep track of all the statistical and data-mining techniques while also being an expert in presentation tools and business politics. That is a hard task, and it's why more and more companies rely on a team of specialists rather than trying to find one person who can do it all.

The process we described in this section is best suited for a data science project that contains only a few models. It is not suited for every type of project. For instance, a project that contains millions of real-time models would need a different approach than the flow we describe here. A beginning data scientist should already get you a long way following this manner of working.

Don't be a slave to the process

Not every project will follow this blueprint because it is subject to the preferences of the data scientist, the company, and the nature of the project you work on. Some companies require you to follow a strict protocol whereas others have a more informal manner of working. In general, you will need a more structured approach when you work on a complex project or when many people or resources are involved.

The *agile* project model is an alternative to a sequential process with iterations. As this methodology wins more ground in the IT department and throughout the company, it is also being adopted by the data science community. Although the agile methodology is a suitable methodology for a data science project, many company policies will favor a more structured approach toward data science.

Planning every detail of the data science process upfront is not always possible, and more often than not you will iterate between the different steps of the process. For instance, after the briefing you start your normal flow until you are in the exploratory data analysis phase. Your graphs show a distinction in the behavior between two groups—men and women maybe? You are not sure because you don't have a variable that indicates whether the customer is male or female. You need to retrieve an extra dataset to confirm this. For this you need to go through the approval process, which indicates that you (or the business) need to provide some kind of project charter. In big companies, getting all the data you need to finish your project can be quite an ordeal.

You can learn more about data science in the book Introducing Data Science, offered by Manning Publications.

For source code, sample chapters, the Online Author Forum, and other resources, go to
http://www.manning.com/cielen/