# The Data Science Process

By Springboard.

What does a day in the data science life look like? Here is a very helpful framework that is both a way to understand what data scientists do, and a cheat sheet to break down any data science problem.

Key words: CRISP-DM, Data Science, Methodology, Springboard

At Springboard, our data students often ask us questions like “what does a Data Scientist do?”. Or “what does a day in the data science life look like?”

These questions are tricky. The answer can vary by role and company.

Uma imagem contendo captura de tela, cartão de negócios

Descrição gerada automaticamente

Fig 1: Data Science Process, credit: Wikipedia

So we asked Raj Bandyopadhyay, Springboard’s Director of Data Science Education, if he had a better answer.

Turns out, Raj employs an incredibly helpful framework that is both a way to understand what data scientists do, and a cheat sheet to break down any data science problem.

Raj calls it “the Data Science Process”, which he outlines in detail in a short 5-day email course. Here’s a summary of his insights.

## Step 1: Frame the problem

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

You’ll often get ambiguous inputs from the people who have problems. You’ll have to develop the intuition to turn scarce inputs into actionable outputs–and to ask the questions that nobody else is asking.

Say you’re solving a problem for the VP Sales of your company. You should start by understanding their goals and the underlying why behind their data questions. Before you can start thinking of solutions, you’ll want to work with them to clearly define the problem.

A great way to do this is to ask the right questions.

You should then figure out what the sales process looks like, and who the customers are. You need as much context as possible for your numbers to become insights.

You should ask questions like the following:

1. Who are the customers?
2. Why are they buying our product?
3. How do we predict if a customer is going to buy our product?
4. What is different from segments who are performing well and those that are performing below expectations?
5. How much money will we lose if we don’t actively sell the product to these groups?

In response to your questions, the VP Sales might reveal that they want to understand why certain segments of customers have bought less than expected. Their end goal might be to determine whether to continue to invest in these segments, or de-prioritize them. You’ll want to tailor your analysis to that problem, and unearth insights that can support either conclusion.

It’s important that at the end of this stage, you have all of the information and context you need to solve this problem.

## Step 2: Collect the raw data needed for your problem

Once you’ve defined the problem, you’ll need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you’ll need and finding ways to get that data, whether it’s querying internal databases, or purchasing external datasets.

You might find out that your company stores all of their sales data in a CRM or a customer relationship management software platform.You can export the CRM data in a CSV file for further analysis.

## Step 3: Process the data for analysis

Now that you have all of the raw data, you’ll need to process it before you can do any analysis. Oftentimes, data can be quite messy, especially if it hasn’t been well-maintained. You’ll see errors that will corrupt your analysis: values set to null though they really are zero, duplicate values, and missing values. It’s up to you to go through and check your data to make sure you’ll get accurate insights.

You’ll want to check for the following common errors:

1. Missing values, perhaps customers without an initial contact date
2. Corrupted values, such as invalid entries
3. Timezone differences, perhaps your database doesn’t take into account the different timezones of your users
4. Date range errors, perhaps you’ll have dates that makes no sense, such as data registered from before sales started

You’ll need to look through aggregates of your file rows and columns and sample some test values to see if your values make sense. If you detect something that doesn’t make sense, you’ll need to remove that data or replace it with a default value. You’ll need to use your intuition here: if a customer doesn’t have an initial contact date, does it make sense to say that there was NO initial contact date? Or do you have to hunt down the VP Sales and ask if anybody has data on the customer’s missing initial contact dates?

Once you’re done working with those questions and cleaning your data, you’ll be ready for exploratory data analysis (EDA).

## Step 4: Explore the data

When your data is clean, you’ll should start playing with it!

The difficulty here isn’t coming up with ideas to test, it’s coming up with ideas that are likely to turn into insights. You’ll have a fixed deadline for your data science project (your VP Sales is probably waiting on your analysis eagerly!), so you’ll have to prioritize your questions. ‘

You’ll have to look at some of the most interesting patterns that can help explain why sales are reduced for this group. You might notice that they don’t tend to be very active on social media, with few of them having Twitter or Facebook accounts. You might also notice that most of them are older than your general audience. From that you can begin to trace patterns you can analyze more deeply.

## Step 5: Perform in-depth analysis

This step of the process is where you’re going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to crunch the data and find every insight you can.

In this case, you might have to create a predictive model that compares your underperforming group with your average customer. You might find out that the age and social media activity are significant factors in predicting who will buy the product.

If you’d asked a lot of the right questions while framing your problem, you might realize that the company has been concentrating heavily on social media marketing efforts, with messaging that is aimed at younger audiences. You would know that certain demographics prefer being reached by telephone rather than by social media. You begin to see how the way the product has been has been marketed is significantly affecting sales: maybe this problem group isn’t a lost cause! A change in tactics from social media marketing to more in-person interactions could change everything for the better. This is something you’ll have to flag to your VP Sales.

You can now combine all of those qualitative insights with data from your quantitative analysis to craft a story that moves people to action.

## Step 6: Communicate results of the analysis

It’s important that the VP Sales understand why the insights you’ve uncovered are important. Ultimately, you’ve been called upon to create a solution throughout the data science process. Proper communication will mean the difference between action and inaction on your proposals.

You need to craft a compelling story here that ties your data with their knowledge. You start by explaining the reasons behind the underperformance of the older demographic. You tie that in with the answers your VP Sales gave you and the insights you’ve uncovered from the data. Then you move to concrete solutions that address the problem: we could shift some resources from social media to personal calls. You tie it all together into a narrative that solves the pain of your VP Sales: she now has clarity on how she can reclaim sales and hit her objectives.

She is now ready to act on your proposals.

Throughout the data science process, your day-to-day will vary significantly depending on where you are–and you will definitely receive tasks that fall outside of this standard process! You’ll also often be juggling different projects all at once.

It’s important to understand these steps if you want to systematically think about data science, and even more so if you’re looking to start a career in data science.

If that’s the case, you may want to check out our free, 40-page guide to Getting your First Job in Data Science!

Even if you’re not looking to break into the field, your career in data science will only get better by getting back to the basics and understanding them thoroughly. We’d love any feedback you have on the data science process.