

I believe that data science, at its core, is a mindset for solving problems. Our world is full of complex issues that go beyond the theoretical knowledge of humanity and have stood as challenges for centuries. While many of these issues have resisted theoretical approaches, data science gives us a new lens in which we can understand and approach them. With data science, the focus is not on fully solving a problem, but rather on gathering enough information to help us benefit from our still imperfect understanding. With the immense amounts of data in our world, and with the ever-increasing power of computers, we as data scientists can discover insights, patterns, and solutions that have remained hidden throughout generations. While everyone works with data differently, these are what I believe are the most important points to keep in mind when tackling problems with a data science approach.

### **1. Explore the data**

Working with data is a difficult task, and often one of the hardest parts is figuring out where to start. With the sheer amount of data that exists today, it's hard not to become overwhelmed. It's for this reason that I think the first step of any good data science project should be to simply explore the data. When you find a potentially-helpful dataset, it works well to load the data into your tool of choice (for me, this tool is Python, with the help of libraries like Pandas) and simply play around with it for a bit. Look at the data, sort it, look at it again, do some normalization, look at it yet again, and keep the loop going. While this step can seem fruitless, I would argue that it's quite possibly the most important one, and sets you up for a successful project. This step allows you to get close to the data, knowing what it contains, what it doesn't, the questions it could answer, the questions it couldn't, and more than anything helps you toward the next step: understanding the data.

### **2. Understand the data**

Before you can utilize a data set to its full potential, you need to understand it. As the data scientist, it's important that you know your dataset inside and out, as this will allow you to help other people understand it too. Understanding the data doesn't stop at a surface level; it's not enough to know the min, mean, and max values. Your goal here is to understand the data in a way that will help you tell its story. This requires knowing where it came from, in what context it was collected, what the values truly mean, and as much more as you can figure out. This step can be difficult, especially when the data is not from an area in which you have extensive knowledge. But without a true understanding of the data, the insights you present from it could range from superficial to misleading.

### **3. Work with the data (not against it)**

When working as a data scientist, it's important to remember that the data we use is not just a tool, it's the medium through which we do all our work. Just as an artist must work with the paints they have, a data scientist must work with the data. An area where this principle is violated all too often is with people trying to tell a story that the data simply does not support. It can be easy when beginning a project to be drawn toward a certain conclusion, either from

personal belief or pressure from a higher-up. However, it's critical to have a sense of integrity as a data scientist. Purposefully leaving out data points, scaling axes in misleading ways, and a slew of other tricks that people have come up with over the years can be tempting, but we must resist. The greatest scientists care about uncovering the truth more than being correct, and data scientists are no exception. The need to work with the data can also be seen in how data is a messy being. There is no one size fits all technique to data science, as all data has its unique characteristics. Instead of forcing our data into what we're comfortable with, we must embrace the defining characteristics of our data to tell its story, not the story we want it to tell.

#### **4. Apply the data**

Finally, the ultimate goal of data science is to apply what we learned from our data to the real world. If data science is a mindset for solving problems, then this step is the actual problem solving. We must take our results and deliver them in some actionable way, whether it's solving a problem, telling a story, making an argument, or simply expanding the knowledge of others. While the other steps of data science are worthwhile in their own right, their usefulness really starts to come alive in the application. The value of the preceding steps relies on the project's ability to make some difference, whether it be groundbreaking or simply interesting information. Without an application stage, the efforts of the data scientist remain hidden, underscoring the powerful role that data science can play in furthering our knowledge.

There are many ways in which one can "do" data science, and by no means is one of these ways the correct one. It is my belief, however, that the value of a data scientist lies in their ability to take raw data, typically useless on its own, and turn it into something that benefits us. I believe that the four principles listed above can serve as guiding steps for the practice of data science, helping to cultivate a mindset in which one works with data to develop a greater understanding and spread that understanding to more people. Each step, from exploration to application, is a critical part of unlocking the full potential that data has in addressing the complex issues of our world.