Data analysis is hard, and part of the problem is that few people can explain how to do it. It's not that there aren't any people doing data analysis on a regular basis. It's that the people who are really good at it have yet to enlighten us about the thought process that goes on in their heads.

Imagine you were to ask a songwriter how she writes her songs. There are many tools upon which she can draw. We have a general understanding of how a good song should be structured: how long it should be, how many verses, maybe there's a verse followed by a chorus, etc. In other words, there's an abstract framework for songs in general. Similarly, we have music theory that tells us that certain combinations of notes and chords work well together and other combinations don't sound good. As good as these tools might be, ultimately, knowledge of song structure and music theory alone doesn't make for a good song. Something else is needed.

In Donald Knuth's legendary 1974 essay "Computer Programming as an Art", Knuth talks about the difference between art and science. In that essay, he was trying to get across the idea that although computer programming involved complex machines and very technical knowledge, the act of writing a computer program had an artistic component. In this essay, he says that *"Science is knowledge which we understand so well that we can teach it to a computer."*

Everything else is art.

At some point, the songwriter must inject a creative spark into the process to bring all the songwriting tools together to make something that people want to listen to. This is a key part of the *art* of songwriting. That creative spark is difficult to describe, much less write down, but it's clearly essential to writing good songs. If it weren't, then we'd have computer programs regularly writing hit songs. For better or for worse, that hasn't happened yet.

Much like songwriting (and computer programming, for that matter), it's important to realize that *data analysis is an art*. It is not something yet that we can teach to a computer. Data analysts have many *tools* at their disposal, from linear regression to classification trees and even deep learning, and these tools have all been carefully taught to computers. But ultimately, a data analyst must find a way to assemble all of the tools and apply them to data to answer a relevant question---a question of interest to people.

Unfortunately, the process of data analysis is not one that we have been able to write down effectively. It's true that there are many statistics textbooks out there, many lining our own shelves. But in our opinion, none of these really addresses the core problems involved in conducting real-world data analyses. In 1991, Daryl Pregibon, a prominent statistician previously of AT&T Research and now of Google, said in reference to the process of data analysis that "statisticians have a process that they espouse but do not fully understand".

Describing data analysis presents a difficult conundrum. On the one hand, developing a useful framework involves characterizing the elements of a data analysis using abstract language in order to find the commonalities across different kinds of analyses. Sometimes, this language is the language of mathematics. On the other hand, it is often the very details of an analysis that makes each one so difficult and yet interesting. How can one effectively generalize across many different data analyses, each of which has important unique aspects?

The purpose of this course is to describe the process of data analysis so that you can learn how to better manage that process. What we describe is not a specific "formula" for data analysis---something like "apply this method and then run that test"--- but rather is a general process that can hopefully be applied in a variety of situations. Through our extensive experience both managing data analysts and conducting our own data analyses, we have carefully observed what produces coherent results and what fails to produce useful insights into data.

Note: Parts of this reading were taken from *The Art of Data Science*, by Peng & Matsui.