- [Voiceover] In any attempt to define the field of Data Science, one of the best places to startis with something called the Data Science Venn Diagram. Now, this was created by Drew Conway back in 2013, and it consists of three separate circles, which represent different areas, and taken together, they constitute Data Science. The first one, on the top left, is Hacking Skills, or computer programming and coding. The ability to retrieve and manipulate data. On the right is Math and Statistics, the ability to actually make sense of that data.

And on the bottom, a crucial third component is Substantive Expertise, or familiarity of working in any particular applied domain. Now, let's take a look at each of those things very briefly.Hacking Skills, or computer programming, is there to gather and prepare data. Once you get that data, it's often in unusual data formats, things that don't fit well into the rows and columns of a spreadsheet or a database. And consequently, substantial creativity is required in the hacking skills.

That's why they call it hacking, because it's a creative endeavour. What makes data science data science is that each project brings with it new challenges. The next step is Mathematics, or Math and Statistics. The important thing here is not necessarily to be the world's leading expert in math and statistics, but to know how to choose a useful procedure to answer the questions that you have at hand, and also, how to diagnose problems. Similarly, one of the things that goes on in Data Science is the need to develop and improve procedures as needed to confront new data challenges.

Next is Substantive Expertise, and it's important to understand whatever your field is, what constitutes value. In that particular field, what are the goals? What is used as the common tool? You need to know the goals. You need to know the methods, and especially the constraints of a particular domain. Certain things are possible. Certain things aren't, and that will help you frame your analysis in the most useful way that can be easily implemented.

Now, in our Venn Diagram, we had our three circles here, that collectively make Data Science, but because this is a Venn diagram, there's also these three other areas that involve two circles at a time. The first of these is what's called Machine Learning. The second one is Traditional Research, and the third one is what Drew Conway called the Danger Zone, but I'll have something to say about that. Let's look at each of these in turn. First, Machine Learning.The idea here is that you can make what's called a Black Box predictive model.

All you have to know is, these are the variables that go into it. This is what we're trying to predict, and we get a model that puts it together. Next is Traditional Research. Now, this is possible because in most traditional fields, the datasets and the analyses are structured and they have some sort of continuity across them. Next is what Conway called the Danger Zone.This may be unlikely to happen, because the person who has substantial computer programming ability and substantive expertise probably also developed some math and statistics along the way.

So, what are our conclusions here? First, Data Science combines domains, three domains, hacking, math stats, and domain expertise. Second, diverse skills are involved. There's a lot of different things that you need to be able to do, and do them all at least reasonably well, in order to do good Data Science, and the complement of that is that there are many different roles in Data Science. There's a lot of different skills that people can bring, a lot of different backgrounds, and a lot of different emphases, and that's what we'll talk about next.