What Can Data Science Do For Me?

Data science is a fancy way to say **using numbers and names to answer a question**. You can start with videos, measurements, recordings, or text, but by the time the data scientist gets down to business, they've all been turned into data in the form of numbers and names. All the powerful things that data science can do eventually boil down to that. Estimating somebody's age from a photo. Recommending you a movie that you might enjoy. Identifying the creep who is using your credit card. They all start with a question and some numbers and names.

But it's not just any question or any collection of data that can get you what you want. Before data science can build the solution to simplify your life or make you lots of money, you have to give it some high quality raw materials to work with. The better the materials, the better the final product. You'll know you're ready when:

1. Your Question is Sharp

When choosing your data science question, imagine that you are approaching an oracle that can tell you anything in the universe, as long as the answer is a number or a name. But it's a mischievous oracle, and its answer will be as vague and confusing as it can get away with. You want to pin it down with a question so airtight that the oracle can't help but tell you what you want to know. Examples of poor questions are "What can my data tell me about my business?", "What should I do?", or "How can I increase my profits?" These leave a mile of wiggle room for useless answers. In contrast, clear answers to questions like "How many Model Q Gizmos will I sell in Montreal during the third quarter?" or "Which car in my fleet is going to fail first?" are impossible to avoid.

2. Your Data Measures What You Care About

Once you have a question it's important to make sure your data is relevant. If you want to answer the question "Which of my customers is most likely to leave me for a competitor?", you'll need information about which customers have left for competitors in the past. On top of that, you'll also need information on the factors that led to their leaving. If all you have is their birthdate and shoe size, you're not likely to do a good job finding patterns in customers that left you. But if you have relevant data, like their purchase history and responses to customer satisfaction surveys, then you have a much better chance.

3. Your Data is Accurate

The biggest myth in data science is that you can make up for bad data by having a lot of it. In this imaginary world, it doesn't matter if numbers are missing decimals, sensors fail, or names have been fat-fingered. Sheer volume of data mysteriously compensates. It is true that specialized tools like spell-checking can automatically correct for some errors. It is also true that unavoidable noise in otherwise carefully recorded data can sometimes be removed or worked around. But the type of errors introduced by sloppiness and neglect can't be automatically fixed. They contaminate a data set and make it more difficult to find the patterns. Often, a small set of carefully collected data is more valuable to a data scientist than a very large set of carelessly collected data.

https://blogs.technet.microsoft.com/machinelearning/2015/08/26/what-can-data-science-do-for-me/

4. Your Data is Connected

Looking at the age of drivers in one state against the accident rate of drivers in another tells you nothing about how driver age and accident rate are related. For that, you need to look at the age and accident rates of the *same group* of drivers. Mountains of accurate, relevant data will do you no good if they aren't describing the same data points. This is also known as the missing values problem. It's the same as if your drivers provided you information by filling out a form, but half of them left the age field blank and the other half didn't tell you about their accident history. Some missing values are normal and all but unavoidable. Too many missing values leaves your data collection looking like Swiss cheese and seriously degrades its usefulness.

5. You Have a Lot of Data

As in many things, data quality matters more than quantity. But quantity matters too. A small top-quality data set is more valuable than a very noisy behemoth. But a large high quality data set can do far more than either of them. More data is usually better, since it can let you see more details, ask more specific questions, and have greater confidence in the findings. Don't assume that you have enough because there are already millions of data points or because you've already filled up a one terabyte hard drive. As long as it doesn't degrade relevance, accuracy, or completeness, more data is always better.

If you have all five of these points covered – **a large, relevant, accurate, complete data set** and **a sharp question** – then congratulations. You're ready to do some major league data science. If you're weak on a couple of points, don't worry. That just means you're working in the real world where things are gritty. You can still probably get what you need. But if only one or two of these statements are true, then you need to do some more legwork. Gather more data, sharpen your question, organize your data, remove the noisiest points—whatever you can do to boost the quality of your raw materials. If you go ahead and crunch the numbers as they are, any answers you come up with will be highly suspect.

https://blogs.technet.microsoft.com/machinelearning/2015/08/26/what-can-data-science-do-for-me/