Excerpt from “Executive Data Science, a guide to training and managing the best data scientists” by B. Caffo, R. Peng and J. Leek (2018)

1. What is Data Science?

Note: Some of the material in this section appeared on the Simply Statistics Blog

Because this is a book about data science, it might be reasonable to first ask, “What is data science?”. Most people hyping data science have focused on the first word: *data*. They care about volume and velocity and whatever other buzzwords describe data that is too big for you to analyze in Excel. This hype about the size (relative or absolute) of the data being collected fed into the second category of hype: hype about *tools*. People threw around EC2, Hadoop, Pig, and had huge debates about Python versus R.

But the key word in data science is not “data”; it is *science*. Data science is only useful when the data are used to answer a question. That is the science part of the equation. The problem with this view of data science is that it is much harder than the view that focuses on data size or tools. It is much easier to calculate the size of a data set and say “My data are bigger than yours” or to say, “I can code in Hadoop, can you?” than to say, “I have this really hard question, can I answer it with my data?”

A few reasons it is harder to focus on the science than the data/tools are:

- John Tukey’s quote: “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.” You may have 100 Gb and only 3 Kb are useful for answering the real question you care about.

- When you start with the question you often discover that you need to collect new data or design an experiment to confirm you are getting the right answer.
- It is easy to discover structure or networks in a data set. There will always be correlations for a thousand reasons if you collect enough data. Understanding whether these correlations matter for specific, interesting questions is much harder.
- Often the structure you found on the first pass is due to a phenomeon (measurement error, artifacts, data processing) that isn't related to answer an interesting question.

The hype around big data/data science will flame out (it already is) if data science is only about "data" and not about science. The long term impact of data science will be measured by the scientific questions we can answer with the data.

Moneyball

One of the examples that you hear about a lot when you hear about data science is Moneyball. With Moneyball, the question was, can we build a winning baseball team if we have a really limited budget? They used quantification of player skills, and developed a new metric that's more useful to answer that question. But the key underlying question that they were asking, the key reason why this was a data science problem, was "Could we use the data that we collected to answer this specific question, which is *building a low budget baseball team*?"

Voter Turnout

A second question would be, "How do we find the people who vote for Barack Obama and make sure that those people end up at the polls on polling day?" And so this is an example from a study of Barack Obama's data team, where they went and they actually tried to run experiments and analyze the data to identify those people. They ended up being a surprising group of people that weren't necessarily the moderate voters that everybody thought they would be, that could be swayed to go out and vote for Barack Obama.

This is again an example where there was a high-level technical issue that had been used–A/B testing on websites and things like that–to collect and identify the data that they used to answer the question. But at the core, the data science question was "Can we use data to answer this question about voter turnout, to make sure a particular candidate wins an election.

Engineering Solutions

We've talked a lot about how data science is about answering questions with data. While that's definitely true there are also some other components to the problem. Data science is involved in formulating quantitative questions, identifying the data that could be used to answer those questions, cleaning it, making it nice, then analyzing the data, whether that's with machine learning, or with statistics, or with the latest neural network approaches. The final step involves communicating that answer to other people.

One component of this entire process that often gets left out in these discussions is the engineering component of

it.n A good example of where the engineering component is critical came up with the Netflix prize. With the Netflix prize, Netflix had a whole bunch of teams competing to try to predict how best to show people what movies to watch next. The team that won blended together a large number of machine learning algorithms. But it turns out that's really computationally hard to do, and so Netflix never actually ended up implementing the winning solution on their system, because there wasn't enough computing power to do that at a scale where they could do it for all their customers.

In addition to the actual data science, the actual learning from data and discovering what the right prediction model is, there's the implementation component (often lumped into data engineering) which is how you actually implement or scale that technology to be able to apply it to, say, a large customer base or to a large number of people all at once.

There are trade-offs that always come up in data science. The trade-offs between interpretability and accuracy or interpretability and speed, or interpretability and scalability, and so forth. You can imagine that there are all these different components to a model: whether it's interpretable, simple, accurate, fast, and scalable. You have to make judgments about which of those things are important for the particular problem that you're trying to solve.