AI GUEST

What IBM looks for in a data scientist

SETH DOBRIN, IBM ANALYTICS JEAN-FRANÇOIS PUGET, IBM NOVEMBER 30, 2017 2:10 PM

Image Credit: ra2studio / Shutterstock

Job seekers sometimes ask how IBM defines “data scientist.” It’s an important question since more and more would-be data scientists are fighting for attention in an increasingly lucrative labor market.

The first step is to distinguish between what we see as true data scientists and other professionals working in adjacent roles (for instance, data engineers, business analysts, and AI

application developers). To make that distinction, let's first define what we mean by data science.

At its core, data science is applying the scientific method to solve business problems.

You can further expand on the definition by understanding that we solve those business problems using artificial intelligence to create predictions and prescriptions and to optimize processes.

The definition demonstrates that to achieve the true potential of data science, we need data scientists with very particular experiences and skills — specifically, we need people with the experiences and skills required to run and complete data science projects:

1. Training as a scientist, with an MS or PhD
2. Expertise in machine learning and statistics, with an emphasis on decision optimization
3. Expertise in R, Python, or Scala
4. Ability to transform and manage large data sets
5. Proven ability to apply the skills above to real-world business problems
6. Ability to evaluate model performance and tune it accordingly

Let's look at those qualifications in the context of our definition of data science.

1. Training as a scientist, with a Masters of Science or Doctorate

This is less about the degree itself and more about what you learn when you get an advanced degree. In short, you learn the scientific method, which starts with the ability to take a complex yet abstract problem and break it down into a set of testable hypotheses. This continues with how well you design experiments to test your hypotheses, and how you analyze the results to see whether the hypotheses are confirmed or contradicted. A determined person can learn these skills outside of academia or via the right mix of online training and practice — so there's some flexibility around having the actual degree — but direct experience applying the scientific method is a must.

Another advantage of an advanced degree is the rigor of the peer review process and publishing requirements that the degree programs impart. To get published, candidates have to present their work in a way that allows others to review and reproduce it. You must also

deep understanding of the difference between probabilistic and deterministic factors as well as the value and curse of the correlation. It's possible to get an abstract sense of those values, but there's no substitute for the negative and positive reinforcement from mentors or the rejection or acceptance of journals and reviews.

2. Expertise in machine learning and statistics, with an emphasis on decision optimization

Applying the scientific method to business problems lets us make better decisions by predicting what will happen next. Those predictions are the product of artificial intelligence and more specifically machine learning. For a true data scientist, the core technical skillsets of machine learning and statistics are simply non-negotiable.

In addition, decision optimization (aka operations research) is a fast-growing aspect of data science. Indeed, the goal of data science is to help make better decisions by probabilistically estimating what's likely to occur in the future. Carefully applying decision optimization lets data scientists prescribe or determine the next best action for the best business outcome.

3. Expertise in R, Python, or Scala

Being a data scientist doesn't require you to be as good at programming as professional developers, but the ability to create and run code that supports the data science process is mandatory — and that includes the ability to use statistical and machine learning packages in one of the popular data science languages.

Python, R, and Scala are the fastest-growing languages for data science, along with Julia, another upcoming language in the space, though Julia isn't yet fully mature. Like Python, R, and Scala, the core of Julia is open source. But it's important to note that the reason to use these languages isn't that they're free, but for the innovation and the freedom to take them where you want to go.

4. Ability to transform and manage large data sets

The fourth skill is sometimes called big data. Here, the ability to use distributed data processing frameworks like Apache Spark is key. The true data scientist will know how to pull data sets

team. The data itself might be a combination of structured, semi-structured, and unstructured data living on multiple clouds.

The data management process consists of finding and collecting the data, exploring the data, transforming the data, identifying features (data elements important in the prediction), engineering the features, and making the data accessible to the model for training. A priority for any data scientist will be streamlining this process, which can easily eat up 80 percent of their time.

5. Proven ability to apply the skills above to real-world business problems

Fifth on the list is a soft skill set. It's the ability to communicate with non-data scientists in order to make sure that data science teams have the data resources they need and that they're applying data science to the right business problems. Mastering this skill also means ensuring that the results of data science projects — for instance, predictions about the probable evolution of the business — are fully understood and actionable by business people. This requires good storytelling skills, and in particular, the ability to map mathematical concepts to common sense.

6. Ability to evaluate model performance and tune it accordingly

To some, this sixth skillset is an aspect of the second skillset: expertise in machine learning in general. We wanted to call it out separately because, all too often, it's what distinguishes a good data scientist from a dangerous one. Data scientists who lack this skill can easily believe that they've created and deployed effective models when in fact their models are badly over-fit to the available training data.

Be a true data scientist

If you want to be a true data scientist — as opposed to an aspiring data scientist or a data scientist in title only — we encourage you to master each of these six competencies. A data scientist is fundamentally different from a business analyst or data analyst, who often serve as product owners on data science teams, with the important role of providing subject matter

That's not to say business analysts, data analysts, and others can't transition to become true data scientists — but understand that it takes time, commitment, mentoring, and applying yourself again and again to real and difficult problems.

Seth Dobrin is vice president and chief data officer at IBM Analytics.

Jean-François Puget is an IBM distinguished engineer in machine learning and optimization.