I am delighted that aspired data scientists have told me that they have read my blog posts. Since it seems that people read my blog, I am writing another blog post to summarize the advice I often repeatedly give to job seekers.

Different types of Data Scientists

The term data scientist is very vague and not well defined. I have seen many data scientists job postings that are machine learning engineer job postings in disguise. It is important to first understand what the job entails.

{% include figure.liquid loading="eager" path="assets/img/types_of_ds/surprised_pikachu.jpg" class="img-fluid rounded z-depth-1" %}

I came across a blog post a while ago that defines two types of data scientists. Type A (for Analysis) and Type B (for build) data scientist and I agree with the distinction:

> Type A Data Scientist: The A is for Analysis. This type is primarily concerned with making sense of data or working with it in a fairly static way. The Type A Data Scientist is very similar to a statistician (and maybe one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.
>
> Type B Data Scientist: The B is for Building. Type B Data Scientists share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. The Type B Data Scientist is mainly interested in using data "in production." They build models that interact with users, often serving recommendations (products, people you may know, ads, movies, search results).

Universities often produce statistics graduates who are trained in being a type A data scientist and most graduates tend to think that type A data scientist is the only kind. However, in reality, having a type A data scientist is a luxury to most companies as they are still focusing on data foundations. As companies become more data mature, they will hire more type A data scientists.

Just to elaborate a bit more on each type:

Type A Data Scientist (Traditional Data Scientists)

These are the "traditional" kind of data scientists and are often what people think of when they say "data scientists". They do data analysis and are individuals who consume data pipelines and produce actionable data-driven insights for business stakeholders. They skilled in drawing analytical insights from data by methods of exploratory data analysis and machine learning/statistical modeling. They are often the same as data analysts (with higher pay).

Type B Data Scientist (Machine Learning Engineers)

Type B data scientists are often called Machine Learning Engineers. Machine learning engineers are individuals who deploy machine learning applications into production as mentioned above. They are often a hybrid of data scientists, data engineers, and software engineers as well.

A traditional statistics program would train type A data scientists (without the database skills), As a result, I encourage job seekers to explore whether they are more interested in performing data analysis (type A) or build machine learning softwares (type B), as they can dictate the skills job seekers may want to acquire.

1. Be Big-data-enabled

This is often the first advice I give to job seekers. Many job seekers have a great understanding of machine learning and have performed extensive data analysis on small data sets in CSV files using Pandas in Python as part of their course work. While Pandas is a great library in Python, it is not the most scalable tool in data science; in practice, most companies store their data in their database and SQL is the way to query most of these databases. Hence, being proficient in SQL is a must. Fortunately, SQL is not very difficult to pick up and most of what you do in Pandas has an analog in most SQL based languages, such as group by, where, join clauses.

Another popular big data tool is Apache Spark. Spark offers an in-memory distributed computing framework; the most fundamental data structure in Spark is called an RDD (Resilient Distributed Dataset) and there is also a Dataframes framework which is an abstraction on RDD, very similar to Pandas data frames from a user's perspective. Learning spark is more than simply learning the syntax; due to the complexity of distributed computing, it is also important to understand the underlying concepts such as partitioning, caching, narrow vs wide transformations for optimization purposes.

2. Being a data scientist is not all about machine learning

Being a data scientist is about drawing actionable insights from data. Machine learning is a tool to draw insights but should not be the primary focus for both type A and type B data scientists. It is great for data scientists to know machine learning, but it is not the core of data science. Most successful data science projects do not emphasize the technique, but the outcome. Therefore the most important part is to understand the pros and cons of each algorithm to optimize results.

To further reinforce this idea, there are many auto ML algorithms out there to automate the machine learning model building process. This means data scientists of the future can focus more on drawing insights from data, as opposed to working on tedious tasks such as model selection and hyper-parameter tuning.

3. Expect to learn a lot

{% include figure.liquid loading="eager" path="assets/img/types_of_ds/pikachuwithshirt.png" class="img-fluid rounded z-depth-1" %}

The data science trend keeps on changing. Ten years ago everyone was obsessed with Neural Networks, deep learning, xgboost, and now there are many auto ML engines out there automating model selection. In terms of technology, distributed file storage systems like Hadoop have been very popular for the past decade, and in recent years, people have moved a lot of computing and storage onto various cloud services. In order to stay competitive in this market, one must keep up with the latest trend. I suggest that one should be very passionate about technology in order to succeed in the data science field or else learning can become dreadful.