How do I to get started in & learn 'data science', 'data analytics'?

Fortune Walla

August 16, 2015

Created on Friday, February 21st, 2014 at 6:57 am

This is an open ended question that has created a lot of discussion on the Web. There are no right answers or approaches. It all depends on what you are interested in & want to achieve. If you choose the self study route, this would be my personal approach.

1) Decide your domain of interest: The fields of "data analytics", "data science" are very vast in their scope to learn everything. So although the general statistical & analytic principles are the same for all fields, it is best to find your industry of interest (say pharmaceutical, econometrics, finance, social & bio sciences, human resources, actuarial sciences, marketing, energy forecasting, predictive modeling, business statistical analysis, medical sciences, etc..) & develop domain specific knowledge for your interests. This will allow you to focus your learning & effort. Trying to master too many domains might get you confused & mentally drained in the long run.

2) Application or technical side: This is a broad generalization. Application side is where the understanding & application is more important than the implementation of the code or the software system. Typically for students of business, social sciences, pharmaceutical, etc...

Usually understanding & framing the research/business problem required to be solved precedes analysis. The important thing is to know the business impact of your analysis. For example, by changing the values of the variables or doing a "WHAT... IF" analysis, one should be able to interpret the change in output in terms of how it addresses the research/business problem that is required to be solved. This type of skill comes only with having the right domain knowledge.

Technical side usually interests students of mathematics, computer science, engineering & fields where focus is on developing new techniques, software & improving existing ones. The fields of data mining & "Big Data" are pretty technical in terms of the mathematical & programming knowledge required. This includes algorithms & equations for data mining, machine learning, pattern recognition, text processing. Database design using "Big Data" technologies like Hadoop, NOSQL, Hive, MongoDb, PIG, MapReduce etc... Also most existing databases like MS SQL Server & Oracle have data mining features.

As technical conferences & statistical journals discuss the latest techniques & methods in terms of mathematics, it would be best to get the mathematical background (usually calculus 1 & 2, linear algebra & probability at undergraduate level) as quickly as you can. This will not only enable you to understand but also express yourself in terms of mathematics.

The ability to understand the research/business problem & convert it into mathematical form & then choose or create an appropriate algorithmic method, that is relatively fast & with minimum error, for the specific software system are skills you should aim to acquire.

3) Books & reading material: The books & material you read from should match your level of expertise & also must be based on the software you plan to use. Your best option is to use academic websites, publisher websites & book review websites like Amazon.com to know the contents & the subject matter of the books. There are specialized statistical books for students of social sciences, marketing, computer science, pharmaceutical etc... Also books that teach data analytics/data science/statistics using particular software (such as R, SAS, SPSS, STATA, MS Excel, Python & many more).

Buy textbooks on the following criteria:

- Based on your mathematical level i.e. based on advanced math like calculus or simple math like algebra.

- That teaches & uses the software you plan to learn with. Using R, SAS, SPSS, STATA, MS Excel, etc. . . .
- That deal with the techniques & methods for your domain of interest i.e. finance, marketing, pharmaceutical, biostatistics etc. . .

4) Get the software running: Being familiar with the various functions & features of the software is almost as important as learning statistical theory. You want to be productive & not waste time looking up help/documentation all the time.

There are software specifically for statistical analysis. Some are suited for certain domains & industries. The hyperlink below shows a list of them.

http://www.amstat.org/careers/statisticalsoftware.cfm

The R project software is the ideal to begin learning with. It is extensive & has many user submitted packages for almost every kind of statistical analysis. It is available for free (as in gratis).

Since R is distributed under the GPL software license you might need to be familiar with the licensing issues of the various R packages especially if you plan to use R commercially & your code also contains other proprietary code using restrictive licenses.

Microsoft Excel is a good option as most office/college computers have machines running MS Windows & MS Office. It has many add-ins (e.g. XLMiner, neuroXL, Oracle Spreadsheet Add-In, Perfringens Predictor Excel Add-in, ADAPA Add-in for Microsoft® Office Excel®, RExcel, DataMinerXL, SAS Add-in for Rapid Predictive Modeler, Palisade Neuraltools Add-in, 11Ants Model Builder Microsoft Excel Add-in, etc..) available for analytics & data mining. It can also be used to interact with the data mining features of MS SQL Server.

Commercial statistical software like SAS, IBM SPSS & STATA are used widely in industry & academia. Those in academia should be able to get an academic license to access & use SAS/SPSS/STATA on their personal computers. SAS also offers an SAS OnDemand service to access SAS through the Internet for a fee.

5) Programming: Since each software package has its own programming environment. Learning a general programming course at undergraduate level will help you understand programming principles that each software uses. Those technically inclined would do well to do a course in computer algorithms & database design.

Knowing SQL is important since most of the data you will analyze or process will reside in a database system like MySQL, MS SQL Server, Oracle etc. . .

A major part of analytics & data science is modifying existing data into a particular format (especially dates, currency, telephone, number formats) for processing. Every software has built-in features for checking errors & missing values, replacing, searching, sorting, filtering & extracting text from a larger dataset. Text processing tools like grep, perl, and python are also used.

It is good idea, although not necessary, to get the basic certifications for commercial software like SAS, SPSS. The exams allow you to brush up your skills & your clients or the company would be somewhat assured of your competency.

Last but not least, it is best to explore websites & join academic/industry communities specific to your area of research or choice.

Summary:

- Decide on your area of interest & domain.
- Do you prefer to be on the technical/programming side or be application/business oriented?
- Get the right books & study material for your academic level & area of interest.
- Decide on the appropriate software used by your industry.
- Learn programming & algorithms.
- Explore the Internet for websites & join communities specific to your needs.