Reading Reflection for Data Science courses

1. The key word in "Data Science" is not Data, it is Science:

Data science, it is a combination of “data” and “science”. However, it is more about science then the data. Why? It is because what you want is what we can get from the data. We want to use the data to answer our questions. By doing that, we do the science work because we need some science tool and science knowledge to answer our question. Of course, science without data is not a science anymore. To me, I would like to say both data and science have equal meaning in practice and in philosophy for the word “Data Science”

1. Data Science and Prediction:

It was estimate that data amount increasing 50% each year or around 40 folds for a decade. There is huge amount of data that people collected. The questions are how we organize the data, put it in structure and use it with or without structure. To make data meaningful, tools have been developed to extract information from data to obtain some useful predictions. The very first tools was from statistics methods, they mainly use statistics to justify whether the prediction is good or useful. Later on, with the developing of computation power, Machine Learning tools was developed by algorithm and computation power with the purpose with better accuracy predictions even though there is some doubt about from statistical points of view. However, there are some area (e.g. economy), both statistical method or machine learning tool do not have good prediction, often go wrong. With more and more data are collected and the new methods and algorithm are developing, the way we predict the results are more and more better. But we also have some obstacles that most data are not structured, how do we interpret and drive those information to extract meaning and useful knowledge is still a problem.

1. 5 Steps of a Data Science Project Lifecycle:

For a data science project, we always start with a question that we want to answer. In order to get an answer, we need to collect data, processing data and do science work to make a conclusion. Here are some majors steps we should consider:

1. Obtain Data: We need to collect data from many different sources and format, therefore, we need more skill and programing language to accomplish.
2. Scrub Data: This step is what we need to do to transform the collected data into some structure that we may use later to answer our questions. For examples, we may need to slip, separate, join, ... data into a format that we need for our model.
3. Explore Data: Before we do with the data, we may need to understand some basic facts about the data we collected. From that, we can decide which models or method we may use for further investigation
4. Models: From the previous step, we had some ideas what kinds of models we may use. In the step, we justify and find the best models that fit our questions.
5. Interpret: Once we have the results from our model, it remains to understand what the results from the model says. That is the part we need to interpret the meaning of the results.

We may start again from beginning if the answers from our data or models does not satisfy us. We may need to repeat and redo the process until we satisfy the results.

1. Data wrangling and visualization:

One of the key part of data processing is to understand basic about the collected data. In order to get it, we need to make some visualization for the data that we obtained. However, in order to do that, we may need to make the data in some forms that is useful for us for people who will use the data later. That is what we call data wrangling. We may need to use tools to clean the data, organize them, and put them into a format that is easy to understand and manipulate for using later on different purpose.

1. DataWrangling: The Challenging Journey from the Wild to the Lake

We collected more and more data, massive data every day. However, the challenge is how we organize them and make it more useful for different type of researches and even across different institutions. This paper show how IBM created a “Data Lake Architecture” that can be shared for use among different institutions and projects. Moreover, it is organize in a structure that is easy to use, to change, protected copyrights and some basis forms of data analysis ready to use.