Bing Huang

Data Analysis 101

Data Tools

The category, Data Tools, talks about a CSV file, which is the common format, which is purely barebone. It's a single file, which is limited by the computer's storage, which takes a much longer time to open and use. You have to use secondary software in order to fully use and comprehend the data at hand. From then on, it teaches about the functions of spreadsheets, things such as COUNT, MIN/MAX functions, and SUM. From filtering to more, it guides us through how to properly manipulate the CSV file.

Big Data

In Big Data, it's all about statistics, whether it be gradually increasing or decreasing, it talks about all things related. In datasets that are vaster than humans on Earth, NASA'S EOSDIS adds 23 terabytes daily. With that type of increment, you'd need a highly scalable network to keep up. Large companies such as IBM and NASA need server rooms/data centers, that have a plethora of petabyte storage systems, in order to keep up. In order for engineers to keep up with the intake, they'd need to pick up parallel computers, which would split up the processing power amongst other computers.

Bias in Machine Learning

By feeding neural networks massive amounts of specified data, it's able to continuously register and sharpen their intuition. When it comes to bias in machine learning, we can look at a prominent problem, within criminal justice. With the fed dataset, it can "predict" how likely a person would commit a crime again, with an overwhelming amount of repeat offenders being people of color. Without the proper feeding data, it'll be prejudice without fail, and it's not fair to minorities. The same goes with facial recognition, as it was more likely for the system to fail to recognize East Asian and African American faces. It's not advanced enough to be put into public use, and it falsely incriminates people.

Unit Test

While taking the Unit test and going over the three major topics, I realized that the answers were much more obvious than I had originally thought. I overthought a lot of it, but I was able to learn from my mistakes and understand that were the answers. Once you really thought about all the questions though, it was pretty simple. From how the dataset on a graph showed a story, to what certain terms were called that engineers needed to use in order to keep up with the intake of data. Overall, it was a good learning experience that every Computer Science major should go through at least once.