Reflection paper on BigDat 2019: 5th International Winter School on Big Data

Ambika Shrestha Chitrakar

BigDat 2019, 5th International Winter School on Big Data was held in Cambridge University, Department of Computer Science, the United Kingdom from January 7 to January 11, 2019. This winter school was divided into 3 different parallel sessions, because of which it was not possible to attend all the lectures. The lectures were mostly about data science and big data techniques and frameworks. This course was very intensive, from 9 o'clock morning to 7 o'clock in the evening, every day. Since I am interested in data science and big data, I found the lectures of this winter school very interesting and beneficial to me. Some sessions even had practical materials, which was useful to understand the methods in practice. We also got a certificate for attending the winter school on our last day. I am grateful to COINS for sponsoring me for this winter school. I really enjoyed the sessions and learned many things related to big data.

The keynotes speaker of the winter school was Kenji Takeda who is a Director of health and AI partnerships for Microsoft Research. He talked about the application of AI and machine learning to transform healthcare by exploiting data in the cloud. It was impressive to know how they are using machine learning algorithms to help the community in real life.

I learned about process mining in the first lecture, which was very new for me. The lecture was presented by Wil van der Aal who is the founder of the process mining discipline. Process mining is a missing link between model-based process analysis (process science) and data-oriented analysis (data science) techniques. With the help of process mining software ProM and data sets, we applied data science knowledge to analyze and improve the processes.

Another lecture was about big data algorithms that aren't machine learning. In this lecture, we learned about some algorithms that are useful for querying large datasets but are not machine learning algorithms. We learned about locality-sensitive hashing, PageRank, stream-processing algorithms (like counting occurrences, counting unique values, sampling), and graph-processing algorithms (counting neighborhoods, counting triangle) in this lecture.

Another lecture was about processing big data with Apache Spark. The lecturer introduced Apache Spark, which is an open-source cluster computing big data framework to handle big data. He talked about important components of the framework and gave a lot of practical examples to write codes in Spark. He also gave an example of using machine learning algorithms from Spark MLlib. Spark MLLib is one of the machine learning libraries in Spark. It includes parallel implementation of some of the known machine learning algorithms. This lecture was very interesting to me as I was doing research on big data handling and analyzing data in cybersecurity and digital forensics by using Spark.

There was another lecture on high-performance big data computing. It was about Twister2 tool, which will be released this year. This is a toolkit for big data solutions. The lecturer said that Hadoop, Apache Spark is slow. Twister 2 is a faster version. We will get the opportunity to validate his statement soon.

Apart from such interesting lectures, I got the opportunity to visit beautiful Cambridge and make new friends who have similar interests as mine.