Good morning, everyone. Thanks for taking your time for attending my thesis defense today. It is getting cold in Brookings. I don’t know how the weather in Columbus and Rapid City. So, hope you all stay warm. Today’s condition is a little bit special. Besides my academic advisor, Dr. Ge. I have two committee members joining online. Dr. Ma is my previous research advisor and he currently is an associate professor in Ohio state university. Dr. Burrows is SDSU extension horticulture specialist in Rapid city. We are all in a different time zone. So, thanks again for your effort to attend this presentation. Thank you. Dr. Ma and Dr. Burrows, if you cannot hear me clearly, please stop me immediately. I will try my best to fix any internet issue. Well, today my presentation topic will be Machine learning algorithms and computational tools development for the analysis of biological sequencing and medical records. Machine learning is a popular topic recently and I can hear many many people talk about it. What is machine learning. In a brief words, machine learning is to train computer with a lot of data and let it make prediction for newer dataset. The most important things for machine learning is data. It does require a lot of data to make accurate prediction. Luckily, during past two years, I had enough data to work with. These data are sequencing data and electronic medical records.

I split my presentations to four different chapters. In chapter 1, I will provide some of background about sequencing data and medical data, and give a brief idea that why machine learning can be used for data analysis. In the chapter 2, I will introduce two of my research topics, sequencing data analysis and EMR data analysis. In the chapter 3, I would like to introduce some of my collaborative efforts for other research projects, which are all valuable experience for me. In the last chapter, I will give a brief conclusion about my thesis.

As I mentioned above, machine learning require training data. A lot of training data. Therefore, for applying machine learning algorithms, we have to find accurate or high-resolution data. You throw treasure in machine learning algorithms, you will receive treasure. You throw garbage in, you will get garbage. Next-generation sequencing technology provide us a friendly cost, high speed, high resolution and accuracy sequencing data. The basic NGS process involves fragmenting DNA/RNA into multiple pieces, adding adapters, sequencing the libraries and reassembling them to form a genomic sequence. The most critical advantage is NGS sequence millions of fragments parallels. Therefore, it is no doubt that we can apply machine learning algoritms on it. Through transcription process, DNA can be converted to RNA. Then RNA can translate to protein sequencing. NGS trigger explosive growth of sequencing data because its advantages. As Jan of 2019, NCBI Genebank have 200 million sequencing data.

Electronic medical record (EMR) data is another my research interest. Medical record stores longitudinal health information which include