Seminar Writeup
Nan Tang

More Data More Problem

Daniela Witten

3/2 2:30 - 3:30

Name: Kathleen Champion
Affiliation: She has finished her master's degree in UW Applied Math, and her Ph.D. study will begin later this year.
Facts: She spent three years at Johns Hopkins University applied physics laboratory, working on methods that use laser light to target distance, and scan the three-dimensional representations of the target object.

The seminar introduced challenges we encountered in the big data era, the more data we can collect, store and process, the more challenge ambitions we will have. Professor Daniela used three examples on issues that come along with big data. The first challenge is about the identification of genetic variants, how specific variants do harm to human health. As she said, three million forms of genetic sequence, of which each sequence is consisted of multiple gene pairs, and the variant(s) can be any number of gene pairs in the sequence that different from reference sequence. The combination she calculated, about total types of variants, is nine billion. The number is so big that it is impossible to test each of the case if it is harmful. The second example displayed, is decoding neural activity from calcium imaging data, of which, in my point of view, the functionality of each position in brain. People study such activity by estimating the spikes in time series and find the best fitted curve for spikes. Profess Daniela has developed an algorithm that can more efficiently predict the time of next few spikes based on recurse past spikes. The last challenge she mentioned is big data's characteristic that the data set always contains more features than observation, in the aspect of matrix, larger colon number than row number. For example, the DNA sequence for one thousand patients. In such cases, it is difficult to calculate the unique coefficients for regression model. One of the solutions for this problem is by regulation, i.e. minimizing the features.