Lecture-1 Notes

1. What is data science?
   →Involves
   Computer science
   Statistics
   Making science
   Making models
   Making predictions
   Getting insights from the data
   Making sense of data ourselves
   Communicating results through visualizations
2. How to merge multiple data sets taken from different sources?
   Keeping in mind that all these datasets were being collected in very different settings and definitely there will some bias in whole of data ranging minimum to maximum values. Now how you will consider those biases. For this you have to take in to account the Bayesian inference as in looking at prior models and historical values to predict the current weighting scheme for given data set. Definitely this can't be done in excel you have to lot of scripting. E.g. python
3. In some cases data will be so much complicated that simple visualizations on linear scales will not help to communicate meaningful essence sophisticated visualization schemes have to be adopted. E.g genomic data
4. Dimensionality incurs a big challenge both computationally as well as statistically.
5. Missing data and sparse data is another issue in data science as we have to make some assumptions in order to account for missing data.
6. Train data: Portion of data on which you have trained your model/algo to get some predictions or inferences out of your dataset.
7. Test data: Portion of data you have kept aside to validate your predictions and inferences.
8. Big data: a cubic mm of brain is about petabyte of data to compute its brain structure. Mouse full brain is cubic cm is about hexabyte of data nearly equal google's one data center.
9. Domain knowledge is a major requirement apart from CS and Stats background. As it augments your thinking ability while you are trying to derive some inferences from a given dataset.
10. How to select a model and how to validate you results is pretty much important.