Problem Set 2

Chang Gao

Main Tools of a Data Scientist

Programming languages like R, Python, Julia and softwares like Stata. SPSS, SAS are used by data scientist. R, Python, Julia are scripted languages that were designed based on compiled languages such as C++, they could not read by CPUs directly but more like human language.

Web scraping is important for mining data, we could do this with the languages above by applying correct packages. In practice, we can do the web scraping by using API or download the html document. When use API, we need to care about the limit added by the source provider.

R, Python and Julia has limited memory so we need to use Resilient Distributed Datasets (RDDs) to chop huge data sets into chunks to manage.

The data visualization is also very important, we can use some packages that was built for visualization under different programming environment. 'matplotlib' for Python, 'ggplot2' for R and 'Plots.jl' for Julia.

After we have access and could manipulate the data, we can let data tell us something with these tools. The data can be used for testing theory or making predictions.