1. The data. We will focus on network traffic data. You can use tool “Bro” (<https://www.bro.org/>) to monitor live traffic, Bro will output log files into the working directory. There are many log files including http log, dns log, connection log. We will be analyzing the http log file obtained from “Bro”.
2. Machine Learning. Using machine learning to do anomaly detection will be the innovative part of the project. The http log file can be considered as the raw data and it has to be preprocessed in order to be used in python-based machine learning programs. You can use tool “Bat” (<https://github.com/SuperCowPowers/bat>) to transform the http log file from Bro into something that python-based machine learning library can handle. A detail example is here (<https://nbviewer.jupyter.org/github/SuperCowPowers/bat/blob/master/notebooks/Anomaly_Detection.ipynb>). You can do analysis on the data then. We will be using jupyter, python, etc.
3. Visualization. The tool we might want to use is “plotly” (https://plot.ly/). We might want to use this tool to visualize the raw data or the results obtained in Step 2 in some creative way.