- *Who is in your team (max 2 without permission from instructor)?*

Team Member: Kaihua Cai (kc3182), Weiqi Yao (wy697)

- *Problem explained & the objective of the project*

Intrusion detection is a common network security research problem, whose aim is to build a predictive model (i.e. a classifier) capable of distinguishing between "bad" connections, called intrusions or attacks(generated from hackers), and "good" normal connections(generated from normal users).

An ideal intrusion detection system should be able to detect intrusions from hackers fast and precisely, i.e. the administrator is informed of attacks the moment it reaches the network, or more formerly, detection at line-rate.

However, due to the mass size of the data to be processed in real time network, most intrusion detection system right now is offline. So ***our objective is to come up with a simple and rather precise machine learning classifier that is possible to deploy online***.

- *Where are you getting the data?*

KDD '99, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html is a dataset acquired from nine weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks.

- *Review of past techniques*

Since network security and intrusion detection has always been a hot topic, and due to the lack of traces/data, KDD '99 is a very common dataset to perform analysis on evaluating the property of classifier.

The baseline approach in Intrusion Detection couple years and now back is often SVM and C4.5 decision tree. But given the popularity of neural network, some recent work has also been done using deep learning.

Some example code from other groups that we used in our project is listed below:

1. K-means clustering in Spark. In S. Ryza, U. Laserson, S. Owen, & J. Wills (Authors), *Advanced analytics with Spark: patterns for learning from data at scale* (pp. 81-97). Sebastopol, CA: OReilly.
2. J. (n.d.). Jeffheaton/t81_558_deep_learning. Retrieved December 16, 2017, from https://github.com/jeffheaton/t81_558_deep_learning/blob/master/tf_kdd99.ipynb

- *Do you have ideas of how you may solve this?*

We will use the most popular method in Machine Learning, which is neural network, but with simpler model design than usual image classification model. Only 3 hidden layers, each with 10, 50 and 10 neurons assigned to them, model borrowed from (jeffheaton, 2017).

As a comparison, to speed up the speed of detection, *we would apply feature selection to reduce the number of inputs to the classifier*. This way we will achieve better speed in detection.

The feature selection method to be used would be PCA, we would downsize the features from 120 to 20 Principle Components. Through this procedure, we hope to greatly reduce the

time to train and to validate the data. However, we do expect a drop in accuracy, from training and testing as well.

Then we would borrow the features selected from a recent paper (*Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets*) using advanced feature selection method for KDD '99 dataset. We will apply the features selected there and apply it our own model and compare the results.

- *Citations:*

1. J. (n.d.). Jeffheaton/t81_558_deep_learning. Retrieved December 16, 2017, from https://github.com/jeffheaton/t81_558_deep_learning/blob/master/tf_kdd99.ipynb
2. K-means clustering in Spark. In S. Ryza, U. Laserson, S. Owen, & J. Wills (Authors), *Advanced analytics with Spark: patterns for learning from data at scale* (pp. 81-97). Sebastopol, CA: OReilly.
3. Suthaharan, S., & Panchagnula, T. (2012). Relevance feature selection with data cleaning for intrusion detection system. *2012 Proceedings of IEEE Southeastcon*. doi:10.1109/secon.2012.6196965