Members: Haoen Huang, Liuyi Cui, Xiaoyuan Liu, Yujia Hong.

Project: XN project2: Find Novel Intrinsic Oncology Targets and Biology

Background information: This research is can be applied to detect new Oncology targets. It is importance as Oncology is an important topic about human’s health. We are trying to classify the lineages that are dependent on essential genes for tumorigenesis.

Data resources: [**https://depmap.org/portal/download/**](https://depmap.org/portal/download/). Each .csv document is about 1000 rows and 25000 columns.

Methods: We try to recover the result in the literature listed below. In this literature, we apply a procedure based on the “random forests” machine learning algorithm. Then we apply 5 different models to the data, including random forests, logistic regression, support vector machine, gradient boosting machine and artificial neural network.

Main work: We believe our main work lies in how to clean the huge data. It may take a long time for python to even just read the data, let alone computing. Before we clean the data, we must understand what each data means. It may take us some time on the Biology parts. After cleaning the data, applying each regression may be more easier.

References: [Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology.](https://northeastern.instructure.com/courses/138029/files/19338786?wrap=1)

Proposed timeline : In February the whole team read the relevant paper and download the data. Clean the data after fully understand it. In march, finish cleaning the data, apply each methods from basic to complicated ones. Learn random forest and apply to it. In April, finish all the paper work, prepare for presentation.

Division of labor:

Clean the data: Haoen Huang, Liuyi Cui.

Apply models to data: Haoen Huang, Liuyi Cui, Xiaoyuan Liu, Yujia Hong.

Paper wrting: Xiaoyuan Liu, Yujia Hong.

Presentation: Liuyi Cui, Haoen Huang.

Questions: It may be too late for us to learn about random forest in our class. Can professor provide some of the information and code for us to learn it ahead as it is a basic algorithm in our program.