# CISC 5352 MiniProject: Machine learning in Implied volatility Pricing and Credit risk analytics

# Selective Learning analytics (80 points)

- dataset: NBoption.csv
- Visualize OTM, ITM, and ATM by using PCA, t-SNE and UMAP
- Apply selective learning for this dataset and compare its results with original machine learning methods
  - Learning machines: k-NN, SVM, Random forest, DNN, anad Gradieent boosting
  - Training/test partition: 80% training data and 20% test data
  - Train-train/train-test partition: 80%: 20%
  - Bad guys: the samples whose values $error = |Volatility - predictedVolatility|$ are at the bottom 10% of all train-test samples
  - The number of nearest neighbors: $k = 10$
  - Note: you should try selective learning both two stages: training clean and test clean
  - It's your freedom
    - to choose kernels in SVM
    - to choose distance measures in k-NN
  - Only do adaptive learning for SVM and the dataset
  - Why do OTM options fit machine learning better?
  - Apply adaptive learning to your option data.

# Revisit Credit Risk Analytics (I) (40 points)

- Credit risk analytics is key in personal loan decision making for banks. Using credit risk analytics, banks are able to analyze previous lending data, along with associated default rates, to create an effective predictive model in loan decision making.
- The file *credit_risk_small_data_0.02.csv* has 2405 credit records with the following variables
  - 'Index': case number
  - 'Delinquency': this is a binary variable 1: means bad credit and 0: means good credit
  - 'Revolving Credit Percentage',
  - 'Capital Reserves',
  - 'Num Late 60',
  - 'Debt Ratio'
  - 'Monthly Income' ($)
  - 'Num Credit Lines' ($1000)
  - 'Num Late Past 90',
  - 'Num Real Estate',
  - 'Num Late 90',
  - 'Num Employees' (should not be more than 10 for a personal credit analytics)

1. Visualize data at least using t-SNE
2. Partition data as 80% for training and 20% for testing and use K-NN and SVM, to conduct credit risk analytics, i.e. do classification to determine good or bad credit records.
3. Compute all classification measures and F1-measure
4. Find all samples in 'TP'/'TN'/'FP'/'FN' class

   (a) TP class: the positive samples that are correctly predicted
   (b) TN class: the negative samples that are correctly predicted
   (c) FP class: the negative samples that are falsely predicted
   (d) FN class: the positive samples that are falsely predicted

5. Develp your own mehod to overcome the imbalance issue (extra credits 20 points)

# Credit Risk Analytics (III) (30 points)

- We have the following data set for credit ranking for 12 different industry sections (it is a simulated data):
  - *credit_sim_ data.csv*, where the first **1540** samples (rows) are labeled as 'good credit' (label type: '1' ), i.e., whose credit rankings are 'AAA', 'AA', or 'A'
  - and the remaining **130** samples are labeled as ' bad credit', (label type: '0') whose credit ranks are 'CCC'.
- There are six variables (columns) in this data set:

```
variable 1: Working capital / Total Assets (WC_TA)
variable 2: Retained Earnings / Total Assets (RE_TA)
variable 3: Earnings Before Interests and Taxes / Total Assets (EBIT_TA)
variable 4: Market Value of Equity / Book Value of Total Debt (MVE_BVTD)
variable 5: Sales / Total Assets (S_TA)
variable 6: Industry sector labels from 1-12
```

- Complete the following problems
  - Conduct k-fold (k=10) cross validation for the data and use the following prediction to conduct classifications and compare their results
    - SVM with 'linear', 'rbf', 'poly', and 'sigmoid' respectively
    - Compare the support vectors under different kernels.
    - Compare the eigenvalues of kernel matrices under different kernels (Extra credits: 10 points)

# What should you turn in?

- 1. A folder that contains
  - A report to show details of your analytics (at least 40 pages)
  - your data
  - source files
  - corresponding related output.
- 2. Submt your zipped folder to BB