Form to fill out by Monday:
https://docs.google.com/forms/d/1soNlxh2i86dEGBbTPZuPMiDsMQaYtM5yuRz7cnlWqOs/viewform

Title: Anti-Money Laundering
Team: Ryan Lee, Brandon Sim, Wesley Chen
Advisor: Pavlos Protopapas

Data: From Bank of America, order of 2M transaction records (more if we find good results)

Project Description: Goals *

Discuss questions and/or goals of the project and the expected outcome. Describe the best case scenario and worst case scenario.

Problem #1: Create a better point system for suspicious activity
Referring to p.72 of "the Bank Secrecy Act/ Anti-Money Laundering Examination Manual" http://www.ffiec.gov/bsa_aml_infobase/documents/BSA_AML_Man_2010.pdf, as part of the methodology outlined, "establishing and applying expected activity or profile filtering criteria" is essential to increasing successful detection rate. This research request is to develop and assess potential new ways of defining expected activity or profile. We will have a sample set with ground truth.

Problem #2: Time-dependent machine learning of transaction monitoring
As part of the transaction monitoring methodology, detection rules and thresholds were developed from the point in time activities deemed most suspicious/unusual. Over time, the underlying pattern of suspicious/unusual activities could have be shifted and the existing rules and thresholds will no longer be effective. Even though periodic reviews can resolve the problem, it is time- and resource-consuming when hundreds of rules are involved. Machine learning could be leveraged to monitor the effective of rules, detect the early sign of the shift and provide direction. This request is to explore some basic approaches and access effectiveness.

BEST CASE SCENARIO
- Take multiple approaches to identify suspicious activity given millions of transaction records from consumer and business bank accounts.
- Strive to beat existing risk scoring systems on new data
- If time permits, detect and describe changes in transaction patterns over time and seasonally. Incorporate this information into the algorithm

WORST CASE SCENARIO
- Our classification does not beat the current scoring method.
- We do not have enough time or resources to tackle time-variant patterns in transaction behavior.

[PENDING POTENTIAL PROJECT: Contact at Google who is a Engineering Manager doing research for optimization and visualization problems. Lots of operations research.]

Performance Goals *

Describe the minimum achievement necessary for you to consider the project a success.

1. Take multiple approaches to identify suspicious activity given millions of transaction records from consumer and business bank accounts.
2. Strive to beat existing risk scoring systems on new data
3. Be able to understand the banking fraud detection industry, and how data science is used in the field.

Project Description:Methodology *

Give an overview of the methodology, software and technologies you will use.Describe the computational elements and techniques to be used in the project. Projects should be mainly computational.

Important to first extract features that might be useful. We can do this by making features that might be logical (i.e. location of transaction vs. location of other transactions, etc.) as well as through dimensionality reduction (PCA).

Next, use classifiers, such as:
- SVM
- random forests (seem like they would work really well here)
- stochastic gradient descent classifier
- KNN classifier

Finally, we perform stochastic optimization of parameters.

If data/time permits, we can create transaction networks to analyze clusters of suspicious persons, include into the algorithm.

Schedule *

You should plan your work so that you can avoid a big rush right before the end of the semester. Please provide milestones with dates, including two progress reports to be turned in on 10/11 and 11/15..

~9/15 - Meet with Bank of America, detailed project scoping and planning, get corporate computers will access to dataset
9/25 - Explore data and have excellent understanding of the features provided.
10/11 - Progress Report, Have important features or key components identified
10/20 - At least 2 classifiers written and evaluated
11/15 - Have a presentable scoring method based on best classifier
11/25 - Project Report, Start exploring and understanding time-series approaches to rare/new events or shifts in pattern
12/10 - END, project presentation, write-ups