# Data Science Practicum
# Project 2
# Malware Classification

Teammates: Aishwarya Jagtap,
Ailing Wang,
Weiwen Xu.

# What we did ?

- Feature Engineering
  - Segment (frequency)
  - Bytes (unigram frequency, bigram frequency)
  - Opcodes (unigram frequency, bigram frequency, 4-gram frequency)
  - Combination of features
- Modeling
  - Random Forest
  - Logistic Regresssion
  - Multilayer Perceptron Classifier
  - Naïve Bayes

# Results

| Features | Accuracy on Small | Accuracy on Large |
|---|---|---|
| segment count | 73.21% | 94.85% |
| 1-gram Bytes | 89.94% | N/A |
| 1-gram & 2-gram Bytes | 90.53% | N/A |
| segment count & 1-gram bytes | 93.49% | N/A |
| segment count & 1-gram & 2-gram Bytes | 92.90% | N/A |
| 1-gram & 2-gram opcodes | 95.85% | N/A |
| segment count & 1-gram & 2-gram opcodes | 95.86% | N/A |
| segment count & 4-gram opcodes | 94.08% | N/A |

# Issues

- Features being too sparse!
  - Filter out opcodes with low IDF before n-gram
    - (opcodes with low IDF ≈ stopword ?)
    - too slow
  - Random Forest Feature Importance
  - Filter out less frequent features
    - Kept all the unigram features
    - only kept 1000 most frequent 2-gram and 1000 most frequent 4-gram features
- Memory issues
  - Only able to run one scenario on large data