1. What is the project

   We start by building a classifier that outputs if a given sample is a malware or a benign program. Finally, we will utilize the representation power of digital DNA fingerprints to study malware samples, their collective behaviors and create predictive models based on it to identify malware. Inspired by its biological counterpart, in the digital DNA representation, we encode a malware program's behavioral lifetime through multiple dimensions (network activity, system calls, etc.) in a sequence of characters. Then, we define similarity measures for such digital DNA sequences. We build upon digital DNA and the similarity between groups of malware to study different families of malware. We leverage the built framework to characterize both benign software and malware.

2. Why is tackling this project important

   We plan to study behaviors of various malware through the lens of sequential data mining inspired by ideas from Biology. Leveraging such encoding of malware activity, we plan to investigate discriminatory traits among benign software and malware. We also believe we can better understand malware evolution (from simple / easily detectable ones to complex / evasive ones) with the formulation of digital DNA fingerprints.

3. How do you plan on tackling the project
   a. Include a brief timeline for the project steps

   Goals:

   Week 1:
   1. Select 2 malware from the same family and 2 from different families to start the analysis
      a. Malware samples can be found on https://giantpanda.gtisc.gatech.edu/malrec/dataset/
      b. We plan to use this resource as it provides the following data ; simulating running the malware of our choice:
         i. Network activity: PCAP file
         ii. System calls:
            a. Check how to get the file with readable data. At the moment, it is unreadable.
         iii. Malwords: look for signatures.
         iv. Instruction Usage
         v. Record/replay log
   2. Start making a machine learning classifier from the feature vectors (system calls, network calls).

   Week 2:
   1. Finish making a machine learning classifier from the feature vectors (system calls, network calls).
   2. Analyse the results.
   3. Generate DNA encoding scheme for system calls & Network calls by analyzing the PCAP file

i. Explore if it makes sense to encode all the syscalls or just security-sensitive syscalls

Week 3:

1. Analyze the generated DNA and find commonalities, repetitions among DNA sequences using analysis of common sub-sequences and substrings.
2. Analyze the DNA generated from a benign program
3. Use the established framework to understand the similarities and differences between different varieties of malwares and benign programs.
4. Build a predictive model using the DNA encoding framework to classify unseen samples.
5. Write the project report.

4. List of 3 sets of deliverables (I'll take these under advisement when grading, but I don't promise to strictly abide by them):
    a. Set of deliverables that will yield a passing grade
        i. Perform some exploratory analysis on the feature dimensions (network activity , system calls, instruction usages etc) and extract security relevant feature vectors from the malware samples as well as benign samples.
        ii. Try and test different machine learning methods to build a classifier for the benign program / malware. Study feature importance plots to explain the role of specific features.
    b. Set of deliverables that will yield an A grade
    The goal of this step would be to go beyond classification of an unknown program as malware and benign. We plan to use the DNA fingerprints to characterise different families of malwares from the dimension of different activity features.
        i. Generate digital-DNA fingerprints from our feature set and analyze the similarities and differences between the digital-DNA structures to gain insights amongst the 2 inter-families and intra-families malware.
        ii. Use the developed analysis framework to compare and contrast fingerprints of malware with benign programs. Analyze if the DNA fingerprints have any better detection capability than traditional feature vectors.
    c. Set of deliverables that shows work clearly beyond an A
        i. Analyse the DNA fingerprints of simple and complex malware belonging to the same malware families to understand malware evolution through time and activities.

5. Link to a git repository where you'll keep all the code, documentation, and development through the project.
    a. On github & gitlab you can create private repositories for free, If you do that please add megele@bu.edu to the repository
    https://github.com/jyotsna-penumaka/EC700-Bravo1

    b. If you chose a git provider other than gitlab/github, please contact me so I can get access to your code