ML for CS Project Proposal:
Malware Detection netML Leaderboard Result Reproduction

Rhea Madhogarhia, Anushka Agrawal, Rohan Madhogarhia

- **Project summary: What is the problem, and why is it important?**
  - The problem is to determine whether we can train a model to detect malware flows in a network. This is important because effective malware detection helps protect systems and users from security threats, and applying machine learning could improve accuracy and scalability in identifying malicious activity.
- **Data: What data will you be using for the project?**
  - The netML leaderboard gives us this linked dataset that we will use to try and improve on our replication of the results on the malware detection leaderboard.
- **Machine learning: What models do you expect to try?**
  - We may start with models such as Random Forest (as recommended as a great baseline model in the textbook), Catboost (Gradient Boosting), and XGBoost, since they tend to perform well on structured network flow data and can handle complex feature interactions. But we expect to experiment with different models. For example, one of the papers linked in the netML leaderboard (coauthored by Prof Feamster) uses weighted ensemble models like AutoGluon-Tabular, an Automated Machine Learning tool that combines multiple high performing models. Furthermore, we will experiment with neighbor based classification (k nearest neighbors), and even potentially deep neural networks (if possible). We expect that a combination of these methods (tree based methods, neighbor based classification, and neural networks) will lead to a replication of the leaderboard results.
- **Evaluation: How will you evaluate your models?**
  - Evaluation will be based on how accurately the model is able to detect traffic flows that are malware attacks. We expect that it is more prudent and valuable to stop all malware attacks than it is to ensure we don’t get false positives (labeling a non-malware as malware). As such, we are likely to place greater weight on metrics like recall and sensitivity during our evaluations.
- **Learning objective: What are you expecting to learn from your project?**
  - We hope to learn what features and behaviors distinguish malware attacks from normal network traffic. This will also help us learn how ML techniques can be applied to real-world security problems, and to compare the efficacy of different models on the traffic data.