Project 1
Re-evaluate 5 RobustBench models with another attack algorithm (e.g., FMN) and identify samples for which one attack works and the other doesn't. Explain the results - i.e., provide some motivations on why one of the attacks did not work properly, while the other did.

Project 2
Consider 3 models from RobustBench (CIFAR10, L-inf) and craft universal (and untargeted) adversarial examples aimed to fool the 3 models at the same time. Evaluate transferability of such adversarial examples to other 7 models.

Project 3
Pick a transformer-based text classification model (e.g. from HuggingFace), apply on it both black-box (e.g. SHAP) and white-box (e.g. transformers-interpret) explanations methods on a small set of samples. Compare the attribution scores provided by different methods and discuss the results. **Hint:** Try Captum - https://github.com/pytorch/captum

Project 4
Implement "Indicators of Attack Failure" in secml-torch. Take 5 models from RobustBench, including at least two that report gradient obfuscations (column: "AA eval. potentially unreliable"), and re-evaluate the effectiveness of IoAF.
Github IoAF: https://github.com/pralab/IndicatorsOfAttackFailure
Github secml-torch: https://github.com/pralab/secml-torch
RobustBench: https://robustbench.github.io/