# Exercise 1 - Phase 2: Security, Privacy and Explainability in Machine Learning

**Student**: Marko Georgiev 12442103

**Mentor**: Prof. Rudolf Mayer

## Overview

This report serves as an overview of what I did during Phase 2 (P2) of Exercise 1 (E1) of the SPEML course at TU Wien. Unlike P1, where we are given a fingerprinted dataset and we must attack it, here we are given two datasets, one from P1 and another from Bob. We are also allowed to organize between each other, something which I took full advantage of during this phase. In contrast to P1, where I had to figure out how to measure data utility, here I can focus strictly on removing the fingerprint.

## Datasets Analysis

Before anything I need to analyze all the datasets, I’ve compiled for E1. I have a total of 5 datasets, two of my own and three of fellow students taking the course. Having access to multiple copies of the data is extremely important for my approach in terms of removing the report.

### Importance of Counting Unique Datasets

As pointed out to us by the exercise description, there is a chance that some of the Bob copies given out to different students might be the same. This makes it necessary to check whether every copy of the dataset that we have is completely unique. By analyzing the mechanism by which fingerprints are introduced to the datasets and by having multiple copies of the exact same fingerprinted dataset, I might introduce bias into the fingerprint removal process. For example, since I have 5 copies at the time of writing this report, if three copies of the data are the same, they will always have the majority in terms of changed values in the ‘cleaned’ dataset. I will explain in more detail why this bias occurs in the ‘[Attack Execution Details](#_Attack_Execution_Details)’ section.

### Dataset Difference Analysis

Before I worked on the attack, I needed to know how the datasets were different from each other. My goal is to identify positions in the dataset where values differ and mark these positions as the locations where fingerprint modification occurred. To assess this, I looked at every cell across all unique datasets and counted how many unique values occur. Results are saved in a file ‘diff\_map.csv’ and this difference map is then analyzed into a ‘summary.csv’ file that tells us how many cells in total will be modified. Lastly, I have a file titled ‘uniqueness.csv’ which tells me the number of times a dataset had unique values in a cell where multiple datasets disagree (the number of unique values was > 1 because of that dataset). The complete notebook that does these computations is in ‘dataset-diff-analysis.ipynb’.

## Attack Execution and Results

Now, it is finally time to discuss the attack. Since I have what I consider to be enough dataset versions, my attack only focuses on combining them all into a single one.

### Dataset Utils

An extremely important script in my approach is the ‘dataset\_uniqueness\_utils.py’. This script contains the function that

### Attack Execution Details

### How are Attack Results Saved

## E1 P2 Questions

#### 1. Having multiple datasets, how obvious do you think the fingerprint was before you started attacking it? (0-10)

#### 2. How eager were you to include Bob’s copy to design your attack? (0-10)

#### 3. How eager were you to include more copies to design your attack? (0-10)

#### 4. How confident are you that you broke the fingerprint and that no collaborator can be detected? (0-10)

#### 5. How confident are you that at least one collaborator cannot be detected? (0-10)

#### 6. How difficult did you find the task of disrupting the fingerprint? (0-10)

#### 7. How many collaborators (including Bob) did you work with? (Answer truthfully, there is no wrong answer here)

#### 8. Why do you believe your attack was effective (or not)?

#### 9. If you had more time, what would you do differently?

#### 10. Did you notice any patterns in the data or in the dataset differences that gave you hints about the fingerprinting method?

#### 11. What was your biggest challenge in balancing fingerprint disruption and data utility?

#### 12. How does having multiple datasets help or complicate your attack?