# — PL 9: Privacy-Preserving Data Publishing Assignment

---

**ASSIGNMENT #4: Anonymization of a Dataset with Risk and Utility Analysis**

Due date: June 3, 23:59

Grading: Assignment #4 is worth **3 points**

TO BE DONE IN **GROUPS OF TWO (MANDATORY)**

---

The goal of this assignment is to perform the anonymization of a larger dataset. You will resort to the adult dataset, which is a common choice for evaluating anonymization models: https://archive.ics.uci.edu/ml/datasets/adult.

1. Create a new ARX project and import the adult dataset.
2. Analyze the attributes and classify them into Identifying, QIDs, Sensitive or Insensitive. Justify your choice.

   This should take into consideration the values for distinction and separation.
3. You should now characterize/analyze the privacy risks of the dataset in original form.

   Re-identification risk of the original Dataset
4. You should **apply two anonymization models** to your dataset and conduct an analysis of the performance of each model applied to the dataset.

   For the analysis you should consider:

   - The re-identification risk of the anonymized dataset vs the original dataset
   - The utility level, measured through appropriate utility metrics and/or analysis of data distributions
   - The effect of the level of suppression and coding model (favor generalization vs suppression) on the results
5. If the results are not satisfying, you should perform extra iterations, by either adapting the parameters of the privacy models (e.g. suppression limit, coding model, attribute weights, etc) or considering other privacy models to apply.

Write a report that describes the procedures taken in the previous steps, including the **reasoning behind the choices made** and the analysis of the results/outcomes.

**Submit your report in moodle.**

## Evaluation Criteria

- Classification of attributes and justification [15%]
- Coding Models [20%]
  - Definition and configuration of coding model (e.g. hierarchies)
- Privacy models: utility and risk assessment [50%]
  - Analysis of the utility and privacy (re-identification risk) levels (before and after anonymization)
  - Analysis of privacy models' results according to varying privacy model's parameters
- Structure and organization of the report [15%]