# CS401: Data Privacy

Instructor: Dr. Stacey Truex
Due: Thursday September 16, 2021 5:00pm

## Programming Project #1

The goal of this programming project is to expand your familiarity with $k$-anonymity through hands on experience. Projects should be completed in pairs (and 1 group of 3); group assignments are posted on notebowl.

Identify a publicly available dataset containing microdata. Your data should contain **at least** 10 attributes and **at least** 1,000 records. Use the programming language of your choice to $k$-anonymize your data for at least 3 different $k$ values. Be sure to clearly state how you defined the $QID$ for your dataset and why. Your code should create csv files containing your anonymized data. Using your anonymized datasets, demonstrate how a changing $k$ value impacted the utility of the data.

There should be 1 submission per group containing the following:

- Compressed folder containing: Code used to anonymize data, a README file detailing how to reproduce your results using the provided code, and either the original data you worked with or a file detailing how to access the data from its original source.
- A write up explaining the following: what values of $k$ you used, how you defined your $QID$ and why, how you measured utility, results demonstrating how this utility measure was impacted by the choice of $k$.

If you are having trouble finding data, the UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets.php and Kaggle Dataset Repository https://www.kaggle.com/datasets are both good places to search. Note that your code does not need to be dataset agnostic. That is to say, you can use your knowledge of the features in your dataset to write your code. The code should simply be executable on the dataset you choose so that Dr. Truex can reproduce your $k$-anonymized data.