# Project Instructions

We have seen two kinds of privacy-preserving data publishing techniques: generalization (e.g., *k*-anonymity) and perturbation. Unfortunately, it is not straightforward to mining data published by generalization, as the values become ranges/sets and could overlap, which traditional machine learning algorithms do not recognize as input. See here to get a feeling of how machine learning tools need a non-trivial adaption in order to mine such data, and you may try it out using UTD Anonymization ToolBox.

In this project, we will look at differential privacy which is more popular these days. In differential privacy, we have a privacy budget ε, and in order to keep it each operation needs perturbation. We will get an intuition on how the privacy budget ε affects the result quality and thus data utility.

We require you to have some background knowledge in (1) Python programming and (2) Jupyter Notebook. If you do not already have the background, please check the following links for self-study:

- A quick tutorial on Python: http://cs231n.github.io/python-numpy-tutorial/
- A quick tutorial on Jupyter Notebook: http://cs231n.github.io/ipython-tutorial/
- How to set up your Python and Jupyter Notebook environment: http://cs231n.github.io/setup-instructions/

We will use the dp-stats library which has implemented a number of operations satisfying differential privacy, including: Mean, Variance, Histogram, Principal Component Analysis (PCA), Support Vector Machines (SVM), Logistic Regression (LR).

In the assignment folder, we have a folder "dp_stats" implementing all the operations; the folder "docs" contain a list of notebook files illustrating how to use each operation. Please open run.ipynb in the assignment folder, and following the instructions step-by-step to complete this notebook file. Most of the codes are provided but you are required to complete the code where you see

```
###############################################################################
# TODO: your code here ...                                                    #
###############################################################################
```

You may also see some questions that you need to answer, marked by:

[Your Answer]

The notebook file provides instructions that are self-explanatory, and can be divided into 3 steps:

1. Step 1: The code for comparing the true mean and the perturbed mean that guarantees differential privacy is provided, and you may run with different privacy budget to see how it affects the results, and please answer what you observe.
2. Step 2: The code for comparing PCA with the perturbed version that guarantees differential privacy is provided, which performs dimension reduction over the iris dataset from 5 dimensions to 2 dimensions. You may run with different privacy budget to see how it affects the results, and please answer what you observe.
3. Step 3: Now it's your turn to predict iris types using SVM (e.g., sklearn.svm.SVC) with its perturbed version in dp-stats. You may self-study the related notebook files in the "docs" folder for how to use privacy-preserving SVM in dp-stats.

You are expected to answer all questions and complete all codes to fill in, and save the notebook file as an HTML file: