Goals

- Implement the *Privelet* algorithm that perturbs histograms while satisfying ϵ-differential privacy.
- Perform an experimental evaluation of your implementation.

Expectations

- All questions starting with the $\star$ symbol are optional (bonus).
- **Groups :** two students per group, five supervised sessions, homework needed.
- **Deliverables :** a `tar.gz` archive containing :
  - **Q1** **Your final report** (5p max without appendices[1], PDF format)
  - **Q2** **Your code** (with comments, with a `readme` file (setup, install, usage), with the possible required libraries)
- **Evaluation criteria (generic, the scale may change) :** `Consistency and completeness of the plan of the report` (/2), `Presentation : style, spelling, readability, figures, sources` (/2), `Explaination of you code/technical achievements` (/6), `Analysis` (/6), `readme file` (/2), `Code quality` (/2)
- **Deadlines :**
  - **Intermediary version** (2p max, current state of your work and planning for the final version): Sunday 6th April, 11:59 PM.
  - **Final version**: Sunday 27th April, 11:59 PM.

Environment

⚠ You will work with Python. There is a lot of documentation online if needed (e.g., here and here).

- You can either use Jupyter on your computer or through online services such as CoCalc and Google Colab.

⚠ If you are running Windows we recommend you install Python using Conda.

- You will need to install the following dependencies:
  - either Jupyter's jupyterlab or notebook to run the notebook,
  - pandas to process the data,
  - matplotlib or seaborn to plot your experiments.
- This project uses the Adult data set (only `adult.data`).

Queries

H1 Distribution (histogram) of the education level.

1 Privelet Algorithm

Q1 Implement the 1-dimensional `Privelet` algorithm described in the Section 4 "Privelet for one-dimensional ordinal data"[2] [1]).

Q2 Test your code with `Privelet` to compute query H1 with $\epsilon \in \{0.01, 0.1, 1\}$ (all other parameters set to default values).

[1] Please note that the appendices are supplementary materials and might or might not be read.
[2] Available here: `http://www.cs.cornell.edu/%7Eguoz/Guozhang%20Wang%20publications/privelet_tkde2010.pdf`

2 Experimental Evaluation

Q3 Study the quality of `Privelet` on query H1 :

- Consider the following ϵ values : $\epsilon \in \{0.01, 0.1, 1, 10\}$.
- First compute H1 on raw data without any noise.
- Second, compute 20 times the query H1, and, each time, measure the *Wasserstein distance*[3] between the non-perturbed histogram and the freshly computed perturbed histogram.
- Plot on a graph the average/min/max Wasserstein distances for each ϵ value. The x-axis will be the ϵ values and the y-axis the average/min/max Wasserstein measures.
- Comment the graph by discussing the variation of the Wasserstein distance according to the value of ϵ.

References

[1] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. "Differential Privacy via Wavelet Transforms". In: *IEEE Trans. Knowl. Data Eng.* 23.8 (2011), pp. 1200–1214. DOI: 10.1109/TKDE.2010.247. URL: https://doi.org/10.1109/TKDE.2010.247.

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html