**DS-UA 202, Responsible Data Science, Spring 2023**

**Homework 2: The data science lifecycle; privacy**

**Due at 11:59pm EDT on Friday, April 14**

## **Objectives and Learning Outcomes**

This assignment consists of written problems and programming exercises on the data science lifecycle and data protection. In the programming part of the assignment you will use the [DataSynthesizer](https://github.com/DataResponsibly/DataSynthesizer) library for privacy-preserving synthetic data generation.[[1]](#footnote-0)

After completing this assignment, you will:

1. Explore the interaction between the complexity of the learned model (a summary of the real dataset) and the accuracy of results of statistical queries on the derived synthetic dataset, under differential privacy
2. Understand the variability of results of statistical queries under differential privacy, by generating multiple synthetic datasets under the same settings (model complexity and privacy budget), and observing how result accuracy varies
3. Explore the trade-off between privacy and utility, by generating and querying synthetic datasets under different privacy budgets, and observing the accuracy of the results

**You must work on this assignment individually.** If you wish to clarify any parts of this assignment, please post Campuswire. For any other questions, please email all instructors.

## Grading

The homework is worth 80 points, or 10% of the course grade. Your grade for the programming portion will be significantly impacted by the quality of your written report for that portion. In your report, you should explain your observations carefully.

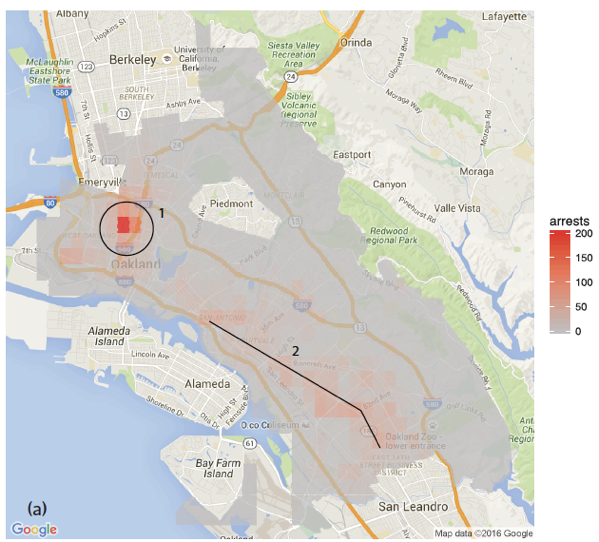
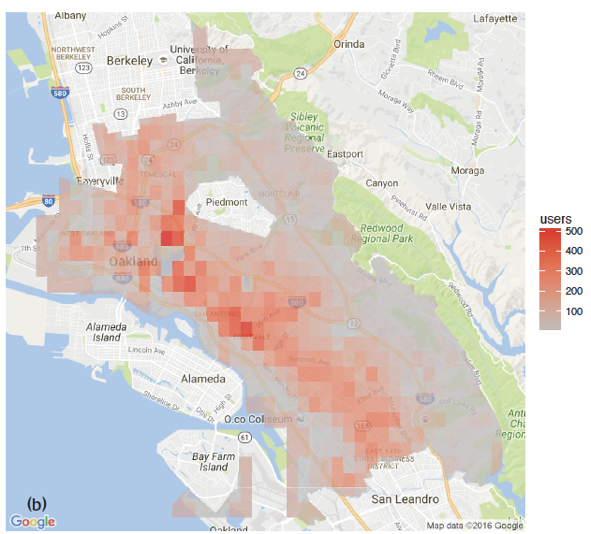
You are allotted 2 (two) late days over the term, which you may use on a single homework, or on two homeworks, or not at all. If an assignment is submitted at most 24 hours late -- one day is used in full; if it’s submitted between 24 and 48 hours late -- two days are used in full.

## Submission instructions

Provide written answers to Problems 1, 2, 3 and 4 in a **single PDF file**. (It is recommended that you [Google Docs](https://docs.google.com) to prepare this PDF, but you may instead use Word or LaTeX). Provide code in answer to Problem 5 in a **Google Colaboratory notebook**. Both the PDF and the notebook should be turned in as Homework 2 on Gradescope. Please clearly label each part of each question. Name the files in your submission *abc123*\_hw2.pdf and *abc123*\_hw2.ipynb (replace *abc123* with your UNI).

## Problem 1 (10 points): Racial disparities in predictive policing

A 2016 study by Lum and Isaac[[2]](#footnote-1) found a racial disparity in the number of drug arrests made in Oakland, CA: while estimated drug use does not differ by race, the number of drug-related arrests is much higher in Black neighborhoods. These findings are substantiated by Figure 1, which we reproduced here.

**Figure 1(a): Number of drug arrests, 2010. Figure 1(b): Estimated number of drug users, 2011.**

*Figure 1(a) reports the number of drug arrests made by the Oakland police department in 2010, and Figure 1(b) reports the estimated number of drug users, based on the 2011 National Survey on Drug Use and Health. Reproduced from Lum & Isaac, Significance, 2016.*

Consider a hypothetical machine learning system, such as the one scrutinized by Lum and Isaac, that uses historical data to determine which neighborhoods would be likely targets for drug-related policing activities. The system may use past crime data along with other potentially useful datasets (e.g., time of year, weather, number of clubs and bars in the neighborhood, etc.).

1. **(3 points)** Give **three distinct reasons** why racial disparities might arise in the predictions of such a system.
2. **(3 points)** Propose **two mitigation strategies** to counteract racial disparities in the predictions of such a system. Note: It is insufficient to state that we could use a specific pre-, in- or post-processing technique that we covered in class when we discussed fairness in classification. Additional details are needed to demonstrate your understanding of how the ideas from fairness in classification would translate to this scenario.
3. **(4 points)** Give **two reasons** why we may be skeptical about the conclusions drawn by Lum & Isaac in their paper.

## Problem 2 (10 points) : Data science lifecycle

Alex, a data scientist in the human resources department of a technology company *MegaSoft*, is running job applicants through an already trained model that ranks applicants for job openings at the company. The applicant dataset has four features: *sex*, *race*, *experience* (measured in years), and performance on a *skills test*. Alex conducts data profiling on the applicant dataset and produces the table below. She finds that all demographic groups perform comparably on the skills test (results are not shown), but that experience differs across groups. Further, she knows that every applicant must take a skills test to apply for a job at *MegaSoft*, but that some applicants do not report their experience.

|  | **EXPERIENCE (years)** | | | | |
| --- | --- | --- | --- | --- | --- |
| **male** | | **female & non-binary** | | **all** |
|  | **white** | **other** | **white** | **other** |
| **N** | 2001 | 1065 | 634 | 300 | 4000 |
| **min** | 1 | 1 | 1 | 1 | 1 |
| **mean** | 5.70 | 5.66 | 7.40 | 7.91 | 6.12 |
| **median** | 5 | 5 | 6 | 6 | 5 |
| **max** | 26 | 35 | 44 | 40 | 44 |
| **NULL** | 80 | 63 | 60 | 42 | 225 |

**(a) (3 points)** To prepare the data to run through the ranking model, Alex replaces missing values (NULL) in the experience feature with the **overall mean value** for that feature in the dataset. She expects the ranking model to prioritize (rank higher) those applicants who score higher on the skills test and have more years of experience. Looking at the data profiling table above, which applicant group(s) may be disadvantaged by Alex’s imputation method and why?

**(b) (3 points)** Propose an alternative data imputation method that may improve the ranking of individuals from the group(s) you identified in **(a)**.

**(c) (4 points)** The data imputation method described in **(a)** can introduce technical bias. Explain how this type of technical bias relates to pre-existing and emergent bias in *MegaSoft*’s hiring example. *Be concise and concrete.*

## Problem 3 (10 points): Database reconstruction

Read the article [*Understanding Database Reconstruction Attacks on Public Data*](https://queue.acm.org/detail.cfm?id=3295691), and answer the following questions, briefly justifying your answer.

1. **(3 Points)** How many possible age value assignments are there for respondents in group 2D? Assume no one is 0 years old.
2. **(2 Points)** List one possible combination of ages for respondents in group 2D.
3. **(3 Points)** Explain why dropping statistics about groups 2A **and** 2B increases the number of feasible satisfying assignments, while dropping either 2A **or** 2B does not.

## 

## Problem 4 (15 points): Randomized response

The simplest version of randomized response involves flipping **a single fair coin** (50% probability of heads and 50% probability of tails). As in the example we saw in class, an individual is asked a potentially incriminating question, and flips a coin before answering. If the coin comes up tails, he answers truthfully, otherwise he answers “yes”. Is this mechanism differentially private? If so, what epsilon value does it achieve? *Carefully justify your answer.*

*Note: This mechanism is different from the mechanism we discussed in class.*

## 

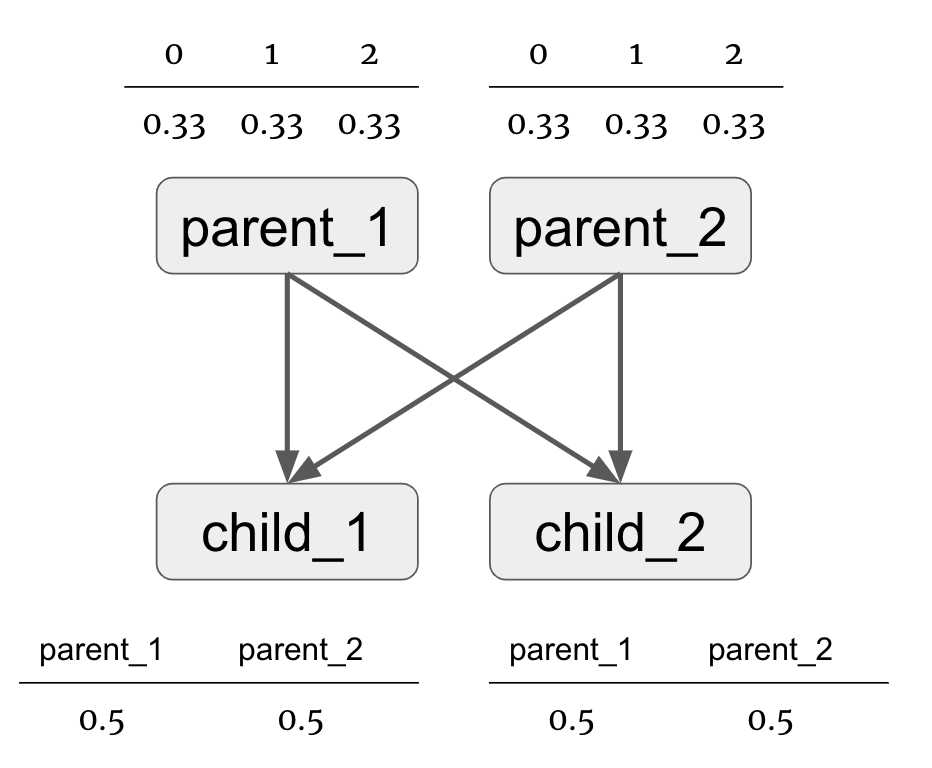
## Problem 5 (35 points) : Privacy-preserving synthetic data

In this problem, you will take on the role of a data owner, who owns two sensitive datasets, called **hw\_compas** and **hw\_fake**, and is preparing to release differentially private synthetic versions of these datasets.

The first dataset, **hw\_compas** is a subset of the dataset released by ProPublica as part of their [COMPAS investigation](https://propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing). The **hw\_compas** dataset has attributes age, sex, score, and race, with the following domains of values: age is an integer between 18 and 96, sex is one of ‘Male’ or ‘Female’, score is an integer between -1 and 10, race is one of 'Other', 'Caucasian', 'African-American', 'Hispanic', 'Asian', 'Native American'.

The second dataset, **hw\_fake**, is a synthetically generated dataset. We call this dataset “fake” rather than “synthetic” because you will be using it as *input* to a privacy-preserving data generator. We will use the term “synthetic” to refer to privacy-preserving datasets that are produced as *output* of a data generator.

We generated the **hw\_fake** dataset by sampling from the following Bayesian network:



In this Bayesian network, **parent\_1**, **parent\_2**, **child\_1**, and **child\_2** are random variables. Each of these variables takes on one of three values {0, 1, 2}.

* Variables **parent\_1** and **parent\_2** take on each of the possible values with an equal probability. Values are assigned to these random variables independently.
* Variables **child\_1** and **child\_2** take on the value of one of their parents. Which parent’s value the child takes on is chosen with an equal probability.

To start, use the [Data Synthesizer library](https://github.com/DataResponsibly/DataSynthesizer) to generate 4 synthetic datasets for each sensitive dataset **hw\_compas** and **hw\_fake** (8 synthetic datasets in total), each of size N=10,000, using the following settings:

* A: random mode
* B: independent attribute mode with **epsilon = 0.1**.
* C: correlated attribute mode with **epsilon = 0.1**, with Bayesian network degree **k=1**
* D: correlated attribute mode with **epsilon = 0.1**, with Bayesian network degree **k=2**

For guidance, you can use the [HW2\_Template](https://drive.google.com/file/d/1xXm9bWKSYnBxR5VEJyyb4TN0qyZ-mFuC/view?usp=sharing) here. Please make sure to duplicate this file rather than put your code directly here.

**(a) (15 points)**: Execute the following queries on synthetic datasets and compare their results to those on the corresponding real datasets:

* **Q1** (**hw\_compas** only): Execute basic statistical queries over synthetic datasets.

The **hw\_compas** has numerical attributes **age** and **score.** Calculate **Median, Mean, Min, Max** of **age** and **score** for the synthetic datasets generated with settings A, B, C, and D (described above). Compare to the ground truth values, as computed over **hw\_compas**. Present results in a **table**. Discuss the accuracy of the different methods in your report. Which methods are accurate and which are less accurate? If there are substantial differences in accuracy between methods - explain these differences.

* **Q2 (hw\_compas** only): Compare how well random mode (A) and independent attribute mode (B) replicate the original distribution.

Plot the distributions of values of **age** and **sex** attributes in **hw\_compas** and in synthetic datasets generated under settings A and B. Compare the **histograms** visually and explain the results in your report.

Next, compute cumulative measures that quantify the difference between the probability distributions over age and sex in **hw\_compas** vs. in privacy-preserving synthetic data. To do so, use the Two-sample Kolmogorov-Smirnov test (KS test) for the numerical attribute and Kullback-Leibler divergence (KL-divergence) for the categorical attribute, using provided functions **ks\_test** and **kl\_test**. Discuss the relative difference in performance under A and B in your report.

For Two-sample Kolmogorov-Smirnov test and Kullback-Leibler divergence, you might find functions such as ‘*entropy*’ and *‘ks\_2samp*’ from *scipy.stats* useful.

* **Q3** (**hw\_fake** only): Compare the accuracy of correlated attribute mode with k=1 (C) and with k=2 (D).

Display the pairwise mutual information matrix by heatmaps, showing mutual information between all pairs of attributes, in **hw\_fake** and in two synthetic datasets (generated under C and D). Discuss your observations, noting how well / how badly mutual information is preserved in synthetic data.

To compute mutual information, you can use functions from <https://github.com/DataResponsibly/DataSynthesizer/blob/master/DataSynthesizer/lib/utils.py>

For heatmaps, we suggest considering functions (*heatmap*) provided in the seaborn library (see example: <https://seaborn.pydata.org/examples/many_pairwise_correlations.html>) and remember to set up *vmax* and *vmin* when plotting.

**(b) (10 points, hw\_compas** only**)**: Study the variability in accuracy of answers to Q1 under part (a) for A, B, and C for attribute **age**.

To do this, fix epsilon = 0.1, generate 10 synthetic databases (by specifying different seeds) for each setting A, B, and C. Calculate the **mean** and **median** of age in each of the generated datasets. Then, for each setting, plot the 10 **median** values and the 10 **mean** values using a **box-and-whiskers** plot. Compare these metrics to the ground truth median and mean from the real data. Carefully explain your observations: which mode gives more accurate results and why? In which cases do we see more or less variability?

Specifically, you should generate 30 datasets in total: 10 under setting A, 10 under setting B, 10 under setting C. For the box-and-whiskers plots, we expect to see two subplots: one for each of the **median** and **mean** withthe three parameter settings (A, B and C) along the X-axis and age on the Y-axis. You should include these plots in your report.

**(c) (10 points, hw\_compas** only**)**: Study how well statistical properties of the data are preserved as a function of the privacy budget, epsilon. To see robust results, execute your experiment with 10 different synthetic datasets (with different seeds) for each value of epsilon, for each data generation setting (B, C, and D). Compute the following metrics, visualize results as appropriate with box-and-whiskers plots, and discuss your findings in the report.

* For each setting (B, C, and D), vary epsilon from 0.02 to 0.1 in increments of 0.02. Specifically, the epsilons are [0.02, 0.04, 0.06, 0.08, 1]. In total, you should generate 150 synthetic datasets (3\*10\*5) and calculate the KL-divergence for race in each dataset. Create three box-and-whiskers plots, one for each setting (B, C, D). Each plot should have epsilon on the X-axis and KL-divergence on the Y-axis. Discuss your findings in the report and include your plots.

1. Both methods are described in papers that are part of the [Data protection reader](https://dataresponsibly.github.io/rds/assets/protection_reader.pdf). [↑](#footnote-ref-0)
2. We discussed the study by Lum & Isaac during class. The article is available as part of the [Data Science Lifecycle reader](https://dataresponsibly.github.io/rds/assets/lifecycle_reader.pdf). [↑](#footnote-ref-1)