DataEng: Data Ethics In-class Assignment

This week you will use various techniques to construct synthetic data.

**Submit**: Make a copy of this document and use it to record your responses and results (use colored highlighting when recording your responses/results). Store a PDF copy of the document in your git repository along with your code before submitting for this week.

## A. [MUST] Discussion Questions

A ride-share company (similar to Lyft or Uber) decides to publish detailed ride data to encourage researchers to develop ideas and open source software that might someday enhance the company’s products. The company’s data engineer publishes the complete set of ride trips for a single year. Data for each trip includes start location, end location, GPS breadcrumb data during trip, price charged, mileage, number of riders served, and information about make, model and year of the vehicle that serviced the trip. All personal information (names, ages, addresses, birthdates, account information, payment information, credit card numbers, etc.) is stripped from the data before sharing.

Can you see a problem with this approach? How might an attacker re-identify some of the real passengers? Insert your responses here and discuss with your group members.

If the attacker is able to breach the data that is not published, as in the original data set that the company had stripped all of the personal information from, then they would be in possession of extremely personal information. But for the sake of the question, let's assume that the company did their due diligence and disposed of that information in a responsible manner. Just the information that was publicized could still be very valuable to the attacker. Given that start and end locations were publicized, the attacker could possibly coordinate the locations with publicly available information like home addresses. They could use this information to potentially identify personal information about the riders. They could then combine this information with the GPS breadcrumb data to identify movement patterns of the individual. Furthermore, the attacker could coordinate the timestamp information with publicly available information via social media, to determine the individuals who took each uber trip.

Search the internet and provide a URL of one article that describes one data breach that occurred during the previous 5 years. The breach must be one in which the attacker obtained personal, private information about customers or employees of the attacked enterprise.

Briefly summarize the breach here, Which of the techniques discussed in the lecture might help to prevent this sort of problem in the future? Describe your chosen breach and your recommendations with your group members.

[Clouded Judgment: How a Former Amazon Employee Hacked Capital One | Seattle Met](https://www.seattlemet.com/news-and-city-life/2023/04/how-a-former-amazon-web-services-employee-hacked-capital-one-paige-thompson-seattle)

In 2019, Capital One was breached by 37 year-old Paige Thompson, an AWS employee from Seattle. She exploited a misconfigured web application firewall in order to gain access to millions of customers' data. The personal information of approximately 106-million customers' data was leaked in the attack, in which Paige posted about on Github. A Github user saw the post, and reported the post to Capital One. Therefore, Capital One contacted the FBI about the data breach. Paige Thompson only served a few-months of prison time due to claims that due to her Transgender status, she would struggle in prison. She was released shortly after, and has about five years of probation along with computer monitoring as a result of her crimes.

## B. [MUST] Model Based Synthesis

Your job is to synthesize a data set based on [the employees.csv data set](https://drive.google.com/file/d/1Ki1um7a22Gw2iXC37q4FcnB8qL4pQW_k)

This startup company of 320 employees intends to go public and become a 10,000 employee company. Your job is to produce an expanded 10K record synthetic database to help the founders understand personnel-related issues that might occur with th] expanded company.

Use the Faker python module to produce a 10K employee dataset. Follow these constraints:

* All columns in the current data set must be preserved. It is not necessary to preserve any of the actual data from the current database
* Need to keep track of social security numbers
* The database should keep track of the languages (other than English) spoken by each employee. Each employee speaks 0, 1 or 2 languages in addition to English.
* To grow, the company plans to sponsor visas and hire non-USA citizens. So your synthetic database should include names of employees from India, Mainland China, Canada, South Korea, Philippines, Taiwan and Mexico. These names should be in proportion to [the 2019 percentages of H1B petitions from each country](https://www.uscis.gov/sites/default/files/document/data/h-1b-petitions-by-gender-country-of-birth-fy2019.pdf).
* 88
* Salaries in each department must mimic the typical salaries for professionals in each field. You can find appropriate data for each type of profession at salary.com For example, see this page to find a model estimate for your synthetic marketing department: <https://www.salary.com/research/salary/benchmark/marketing-specialist-salary>
* The current startup company (as represented by the employees.csv data) is skewed toward male employees. Our goal for the new company is to make the numbers of men and women approximately equal.

Save your new database to your repository alongside your code that synthesized the data.

## C. [SHOULD] Analyze the Synthetic Company

* How many men vs. women will we need to hire in each department?
* How much will this new company pay in yearly payroll?
* Other than hiring from non-US countries, how else might the company grow quickly from size=320 to size=10000?
* How much office space will this company require?
* Does this new dataset preserve the privacy of the original employees listed in employees.csv?

## D. [ASPIRE] Quality of the Synthetic Dataset

Use ydata-profiling to explore your synthetic data set: <https://pypi.org/project/ydata-profiling/>

Use ydata-profiling with the original employees.csv as well to compare.

In what ways does the synthetic data set appear to be obviously synthetic and/or not representative of the current company?

How might you improve the synthetic data to make it more realistic?

## E. [SHOULD] Sampling

Use the DataFrame sample() method to produce a 20 element sample of the data. Use the “weights” parameter of the sample() method to synthetically bias the sample such that employees with ages 40-49 are three times as likely to be sampled as employees in other age ranges.

## F. [SHOULD] Anonymization

Anonymize the name (both first and last names), email, and phone number information in the employee data.

## G. [SHOULD] Perturbation

Perturb the age, salary and years of experience attributes of the employees data using Gaussian noise. How should we choose the standard deviation parameter for the noise? Should we choose the same standard deviation for all three of the perturbed attributes? If not, then how should we choose?