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.
Yes. Publishing ride-share data with start and end locations, GPS breadcrumbs, and vehicle details, even without personal information, poses re-identification risks. Attackers could cross-reference trip data with external sources to identify passengers by their travel patterns. Home and work addresses, along with routine trips, can reveal individual identities. Vehicle data might also be linked back to specific drivers or passengers. Privacy measures should include data generalization and noise addition to protect user anonymity.

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.
One significant data breach in the past five years occurred with 23andMe in October 2023. The genetic testing company experienced a breach in which attackers accessed customer accounts via a credential-stuffing attack. The stolen data included sensitive personal information such as names, email addresses, birth dates, and genetic ancestry information. This breach is particularly concerning due to the sensitivity of genetic data, which could potentially be used for identity theft or genetic discrimination.
URL : https://tech.co/news/data-breaches-updated-list

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.
In October 2023, genetic testing company 23andMe suffered a data breach that exposed sensitive personal and genetic information of over 5 million users due to a credential-stuffing attack. Attackers used previously leaked credentials from other sites to access 23andMe accounts. To prevent such breaches, techniques like multi-factor authentication (MFA), regular password changes, and monitoring for suspicious login activity are crucial. Educating users about the dangers of password reuse and employing robust cybersecurity practices could also help mitigate future risks.

B. [MUST] Model Based Synthesis

Your job is to synthesize a data set based on the employees.csv data set

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 the 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 40% employees who are non-USA citizens and 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.
- The expanded company will have additional departments include “Legal” (approximately 5% of employees), “Marketing” (10%), “Administrative” (10%), “Operations” (20%), “Sales” (10%), “Finance” (5%) and “I/T” (10%) to go along with the current “Product” (20%) and “Human Resource” (10%) departments.
- 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?