

# 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.

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.

A) Yes, there are several issues with this approach:

Linkage attacks can occur when attackers match ride data with public information, like social media posts, to re-identify passengers. Regular trip patterns, such as commutes, can expose identities. Geospatial analysis using start and end locations, along with GPS data, can pinpoint individuals. Additionally, vehicle information (make, model, year) can further narrow down possible matches when combined with other data. These risks highlight the vulnerability of anonymized data to re-identification.

<https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>

The Facebook-Cambridge Analytica data breach in 2018 revealed that Cambridge Analytica harvested data from millions of Facebook profiles without user consent to influence voter decisions in the 2016 U.S. election. This breach highlighted significant privacy issues and the need for better data protection measures. Prevention techniques include differential privacy, which adds noise to the data to protect individual privacy, and data anonymization, ensuring all identifiable information is effectively removed or obfuscated. Implementing strict access controls and monitoring for unusual activity can also help prevent unauthorized data extraction. Additionally, establishing ethical data sharing policies that prioritize user privacy and obtaining explicit consent for data use are crucial measures to prevent similar breaches in the future

## 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.