Contents

1 From Real to Synthetic: A Comparative Analysis of Python Packages Data Generators

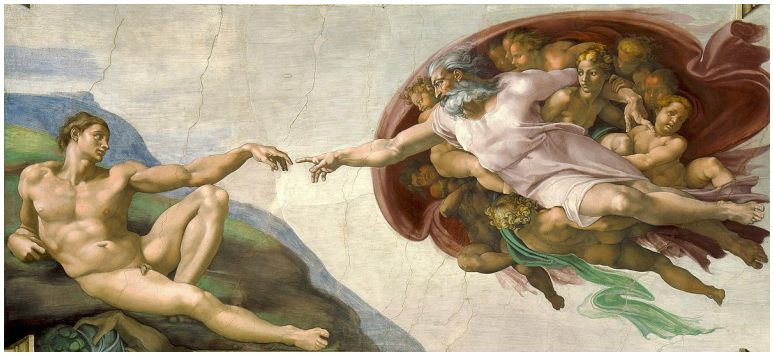

Figure 1: "The Creation of Adam" by Michelangelo

2 Brief Research Description

In recent years, the rise of Large Language Models (LLMs) has required an increased need for high-quality data [1]. However, for a variety of reasons, including regulations, difficulty in access, and privacy concerns [2], obtaining this kind of high-quality real data has become increasingly challenging. A

possible solution to this problem is to use data generated synthetically from real-world data.

This project aims to compare different Python packages, specifically Synthetic Data Vault (SDV) [3] and YData [4], used to create synthetic data.

The main question of the study is:

How closely does the synthetic data resemble the real data when generated by existing Python libraries?

This question is broad and requires clarification on how the "similarity" of the data will be assessed:

1. **Statistical Similarity**: Matching distributions, correlations, and general statistical properties, such as mean, median, and standard deviation.

2. **Predictive Utility**: Evaluating the synthetic data by training and testing Machine Learning models and comparing their performance on real and synthetic data.

Figure 2: "Narciso" by Caravaggio

3 Reasons

This project is of significant interest at this moment due to the exponential need for data in the training of ML models, particularly LLMs. Comparing

the best models for data generation with predictive utility tests on ML models could reveal interesting results regarding the differences (or similarities) between synthetic and real data.

Moreover, this project will allow us to dive deeper into a rapidly expanding topic while also implementing ML models for testing purposes. As an exercise, we could create our own small Generative Adversarial Network (GAN) and compare it with the larger models (just for fun).

Furthermore, the potential for a publishable outcome adds value to this research, offering the opportunity to contribute to a highly active research area at the moment.

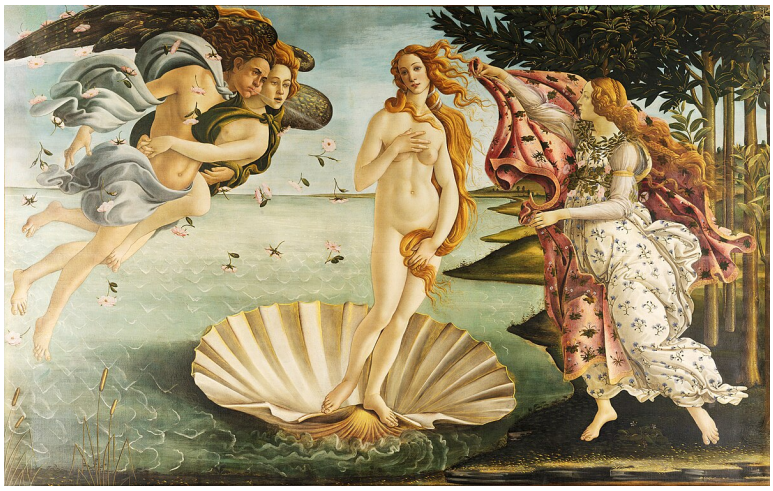

Figure 3: "Nascita di Venere" by Sandro Botticelli

4 Constraints

Difficulties in the project may include:

1. **Computational Resources:** Generating synthetic data, particularly with advanced models like Generative Adversarial Networks (GANs), can require significant computational power. For this reason we will focus on small datasets.

2. **Evaluation Complexity:** Assessing the "similarity" between real and synthetic data involves multiple dimensions, Designing robust and meaningful evaluation metrics to capture these differences may pose challenges.

3. **Tool Limitations:** Python packages like Synthetic Data Vault (SDV) and YData have predefined functionalities. Exploring the nuances of these tools and understanding their constraints will require time and experimentation.

4. **Generalization of Results:** The findings of this study might be dataset-specific. Results obtained on one dataset may not generalize to other types of data, such as time-series, images, or unstructured data.

Figure 4: "Les Amants II" by René Magritte

5 Goals and Non-Goals

Goals for this project include:

1. **Compare different Python packages** (SDV, YData) for synthetic data generation based on statistical similarity metrics (mean, median, standard deviation, correlation, etc.).

2. **Evaluate the predictive utility** of synthetic data by training and testing Machine Learning models on both real and synthetic datasets.

3. **Implement a small Generative Adversarial Network** (GAN) as an exploratory exercise to compare its performance against established synthetic data generation packages.

4. **Assess potential limitations** and trade-offs between different synthetic data generation methods.

5. **Discuss the feasibility of using synthetic data** as a substitute for real-world data in ML applications, considering performance and practical constraints.

Non-Goals include:

1. **Developing a state-of-the-art synthetic data generation model from scratch**
2. **Addressing all types of data** (e.g., images, time-series, unstructured text)
3. **Optimizing ML models for maximum performance**
4. **Comprehensive analysis of privacy risks in synthetic data**
5. **Evaluating all synthetic data generation libraries**

Figure 5: "Dama con l'ermellino" by Leonardo Da Vinci

6 Metrics

The success of the project will be measured using the following criteria:

1. **Implementation and Understanding of Synthetic Data Generation**

    - Successfully utilizing at least two synthetic data generation packages (SDV and YData).
    - Understanding their functionalities and differences in how they generate synthetic data.
    - Comparing real and synthetic datasets using both statistical metrics (mean, median, standard deviation, correlations) and predictive utility through ML models.

2. **Evaluation of Synthetic Data in Machine Learning Models**

    - Training and testing ML models on both real and synthetic data.
    - Measuring and comparing model performance to determine how well synthetic data replicates real data for predictive tasks.
    - Development of a Small Generative Adversarial Network (GAN)

3. **Implementing a basic but functional GAN to generate synthetic data.**

    - Comparing its outputs with those generated by SDV and YData to assess its effectiveness.

Figure 6: "Ceci n'est pas une pipe" by René Magritte

7 References

[1] Chen, Hao, et al. "On the Diversity of Synthetic Data and Its Impact on Training Large Language Models." arXiv preprint, 22 Oct. 2024, arXiv:2410.15226.

[2] "Data Privacy Regulations: Compliance Challenges and Best Practices." Ironhack Blog, 2023, https://www.ironhack.com/us/blog/data-privacy-regulations-compliance-challenges-and-best-practices

[3] "SDV: The Synthetic Data Vault." SDV Developers, 2023, https://sdv.dev.

[4] YData Synthetic: Synthetic Data Generation for Machine Learning." YData Developers, 2023, `https://github.com/ydataai/ydata-synthetic`.