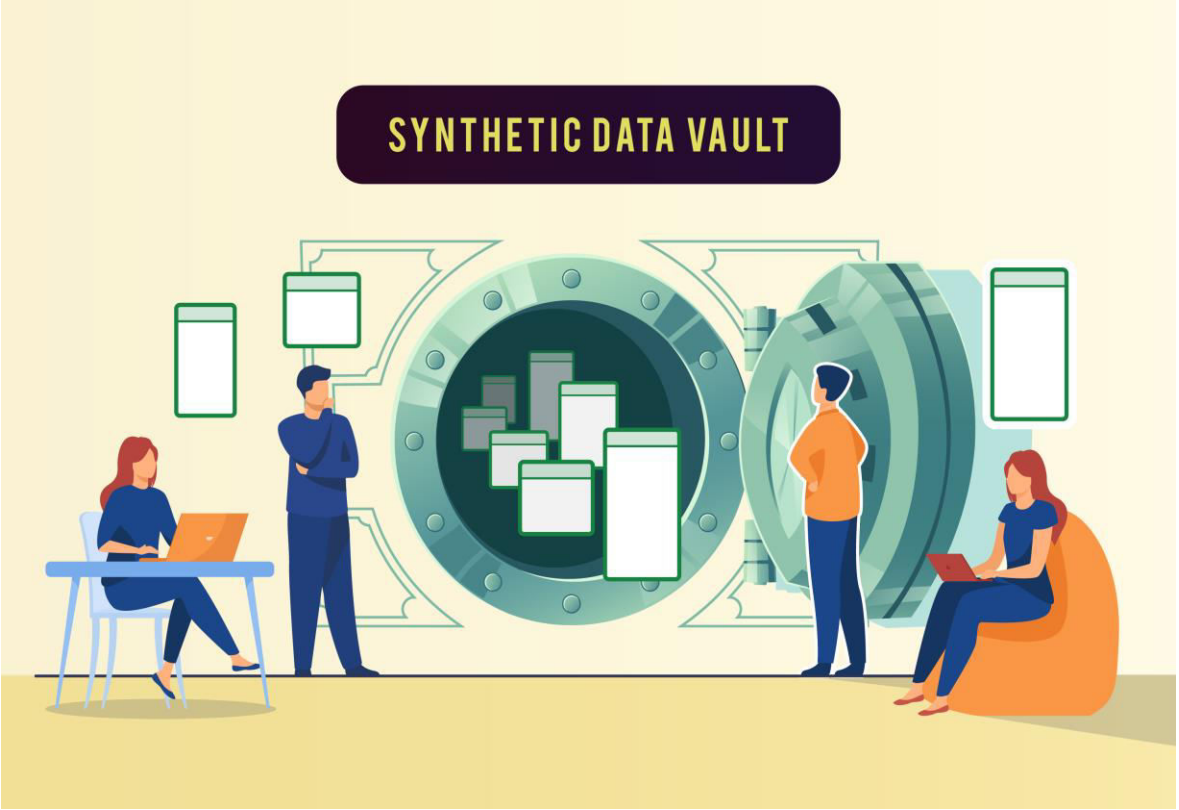
REPORT-1

(INTERNSHIP)



PRESENTED TO: Mrs SHOBHARANI MA'AM (BEL)

FROM: BHOOMIKA K (AIML, MS RAMAIAH)

TOPIC: SYNTHETIC DATA

CONTENTS:

INTRODUCTION

In the era of data-driven decision-making, the demand for high-quality data is paramount. However, accessing and utilizing real-world data comes with challenges such as privacy concerns, data scarcity, and the need for diverse datasets. Enter synthetic data – a powerful solution that bridges the gap between data availability and privacy preservation.

Synthetic data generated from computer simulations or algorithms provides an inexpensive alternative to real-world data that's increasingly used to create accurate AI models.

WHAT IS SYNTHETIC DATA?

Synthetic data is annotated information that computer simulations or algorithms generate as an alternative to real-world data.

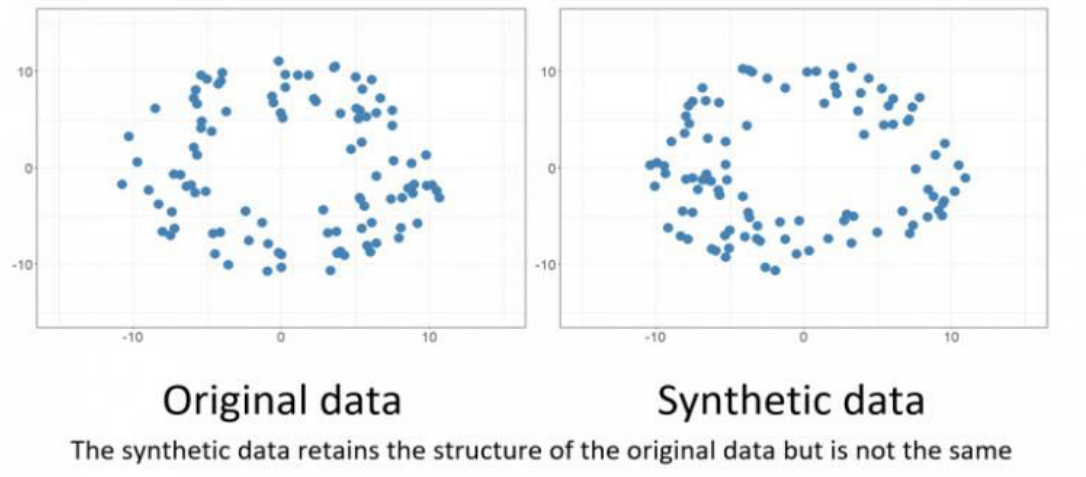
Synthetic data is information that's artificially generated rather than produced by real-world events. Typically created using algorithms, synthetic data can be deployed to validate mathematical models and to train machine learning models.

Data generated by a computer simulation can be seen as synthetic data. This encompasses most applications of physical modeling, such as music synthesizers or flight simulators. The output of such systems approximates the real thing, but is fully algorithmically generated.

LET'S UNDERSTAND THIS WITH AN EASY EXAMPLE :

Imagine you have a secret recipe for your favorite cookies, but you don't want to share it with anyone. However, your friends want to know how delicious your cookies are. What do you do? You create a "fake" recipe that looks and tastes similar to your real one but doesn't reveal the actual ingredients.

In the world of data, it's kind of like that. Sometimes, we have important information that we need to keep private—like personal details in a database. Instead of using the real data, which could be sensitive, we make up pretend data that still behaves like the real stuff. This made-up data lets us work, study, or test things without risking someone's privacy.



The synthetic data retains the structure of the original data but is not the same

WHY SYNTHETIC DATA?

Now, why is it so important?

Developers need large, carefully labeled datasets to train neural networks. More diverse training data generally makes for more accurate AI models.

The problem is gathering and labeling datasets that may contain a few thousand to tens of millions of elements is time consuming and often prohibitively expensive.

Enter synthetic data. A single image that could cost $6 from a labeling service can be artificially generated for six cents, estimates Paul Walborsky, who co-founded one of the first dedicated synthetic data services.

Cost savings are just the start. Synthetic data can address privacy issues and reduce bias by ensuring users have the data diversity to represent the real world.Because synthetic datasets are automatically labeled and can deliberately include rare but crucial corner cases, it's sometimes better than real-world data.

Here are few points on why synthetic data?

Privacy Protection: In many situations, real data involves sensitive or private information. Synthetic data provides a way to create a substitute that maintains statistical properties without revealing the details of individuals.

Data Diversity: Synthetic data can be designed to cover a wide range of scenarios, ensuring that machine learning models encounter a variety of situations during training, which can improve their robustness and generalization.

Data Augmentation: When you need more data to train a machine learning model but collecting additional real data is impractical or costly, synthetic data can be generated to supplement the existing dataset.

Testing and Development: In scenarios where obtaining real data for testing or development purposes is challenging, synthetic data offers a solution. It allows developers to simulate different scenarios and test the performance of algorithms or systems.

Anonymization: Synthetic data can be shared publicly without disclosing real details. This is useful for researchers, developers, and organizations that want to collaborate or provide datasets without compromising privacy.

HOW IS SYNTHETIC DATA GENERATED?

"There are a bazillion techniques out there" to generate synthetic data.

Generative Models:

Generative Adversarial Networks (GANs): GANs consist of a generator and a discriminator. The generator creates synthetic data, and the discriminator evaluates how well it resembles real data. This process continues iteratively until the generator produces data that is difficult for the discriminator to distinguish from real data.

Variational Autoencoders (VAEs): VAEs learn the underlying structure of the data and generate synthetic samples by sampling from the learned distribution.

Simulation Techniques:

Physics-based Simulations: In fields like engineering or physics, simulations can be used to generate synthetic data based on known principles. For example, simulating the behavior of a car in different driving conditions.

Agent-Based Modeling: Simulating interactions among individual agents according to specific rules can generate synthetic data representing complex systems, such as traffic flow or economic markets.

Statistical Methods:

Monte Carlo Methods: These involve random sampling to obtain numerical results. In the context of synthetic data, Monte Carlo methods can be used to generate samples that match the statistical properties of real data.

Bootstrapping: This statistical resampling technique involves creating multiple datasets by drawing samples with replacement from the original dataset. It helps capture the variability in the data.

Data Masking and Perturbation:

Data Masking: Sensitive information in real data can be replaced, masked, or encrypted to generate synthetic data while preserving the overall structure.

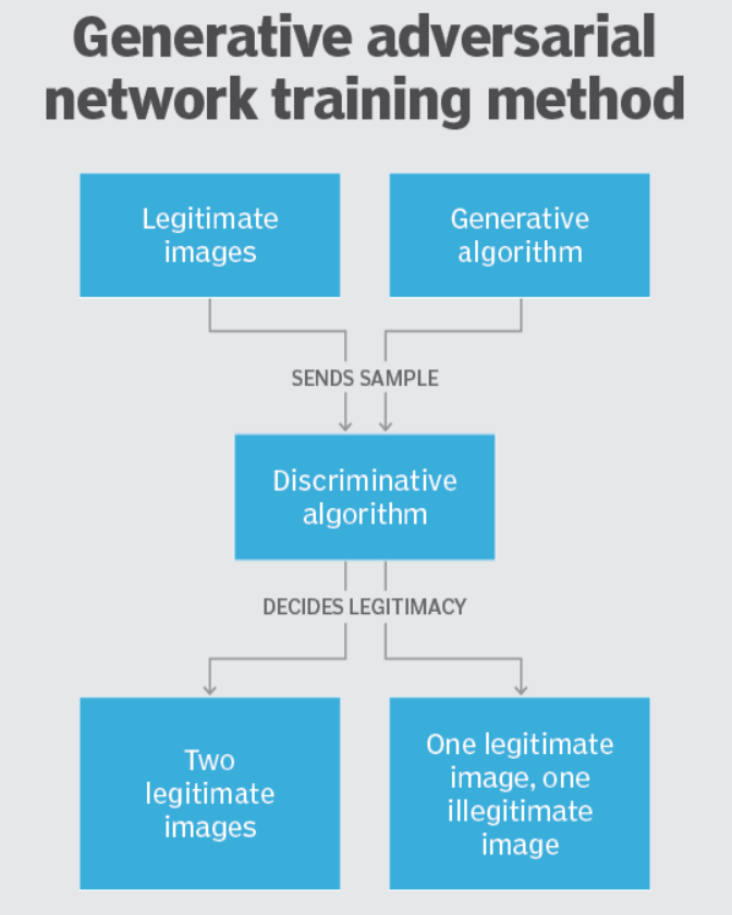
Perturbation: Adding random noise to the original data can create synthetic samples with similar statistical characteristics but with some level of variation.

Rule-Based Approaches:

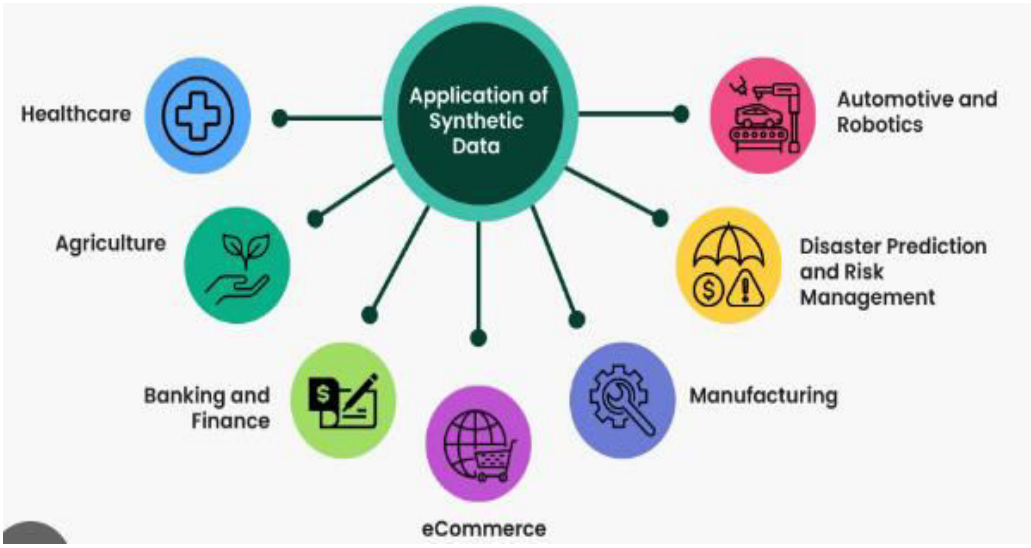
Domain-Specific Rules: In some cases, synthetic data is generated by applying domain-specific rules. For example, in finance, synthetic data might be created by following rules related to market trends, interest rates, and economic indicators.

Hybrid Approaches:

Combining Methods: Some approaches combine multiple techniques. For instance, using generative models to capture high-level features and then applying statistical methods to refine the details.



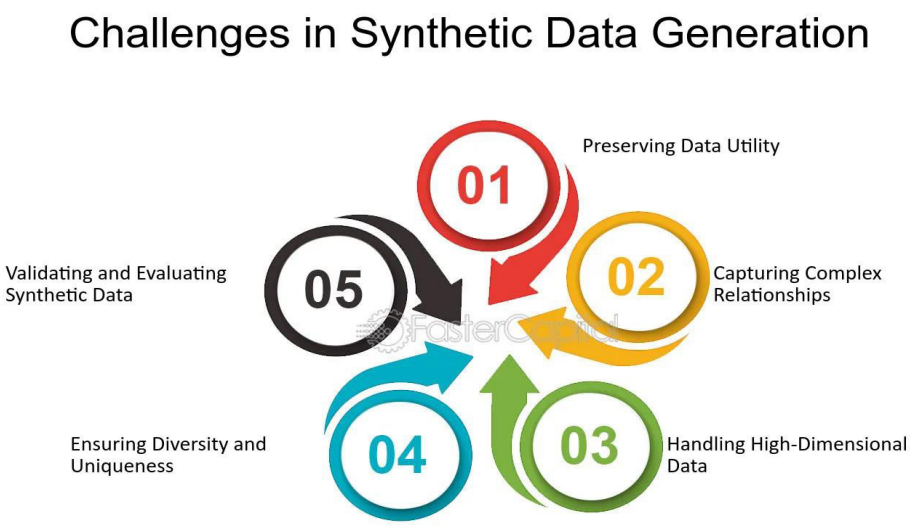
USAGE OF SYNTHETIC DATA



ADVANTAGES AND DISADVANTAGES

Advantages	Disadvantages
Privacy Protection	Limited Realism
Data Diversity	Model Generalization Challenges
Data Augmentation	Model Overfitting
Testing and Development	Ethical Concerns
Anonymization	Complexity in Generation
Cost Savings	Security Risks
Customization	Dependency on Data Generation Methods

CHALLENGES



Limitations:

Biased or deceptive results: Synthetic data can be misleading, limited or discriminatory due to its lack of variability and correlation.

Lack of accuracy: Another challenge with synthetic data is that it is often created using a computer algorithm, which may not always be accurate. As a result, synthetic data can occasionally produce inaccurate results.

Time-consuming steps: Relatedly, synthetic data requires additional verification steps, such as comparing model results with human-annotated, real-world information. Such efforts take time to complete and prolong the projects.

Losing outliers: Synthetic data may not cover some of the outliers present in the original dataset because it can only mimic but not replicate real data. However, outliers can be relevant for some research.

Dependency on the real data: Synthetic data quality often depends on the real model and the dataset that have been developed for creating synthetic data. Without a desirable and qualitative real dataset, various synthetic datasets that are generated in huge amounts by using the original dataset will end up functioning ineffectively and sometimes even incorrectly.

Consumer skepticism: As synthetic data use increases, businesses can face consumer skepticism, such as questioning the credibility of the data.

BENEFITS

Synthetic data is a lifesaver for organizations that work with confidential or sensitive data. Its power to replicate the characteristics and patterns of real-world data without exposing confidential information helps preserve data security while still allowing researchers, analysts, and decision-makers to gain valuable insights.

In addition, generating synthetic data offers several other benefits to organizations:

Lower costs associated with data management and analysis.

Faster turnaround time for development workflows and projects.

Greater control over the quality and format of the dataset.

Better performance in machine learning algorithms.

Greater flexibility and increased collaboration.

Reduced bias and improved data security.

HISTORY

Synthetic data dates back to the advent of computing in the 1970s. Most initial systems and algorithms depended on data to function. However, restricted processing capacity, challenges in collecting vast volumes of data and privacy concerns led to the creation of synthetic data.

In the wake of the ImageNet competition of 2012 -- commonly referred to as "the Big Bang of AI -- a group of researchers" led by Geoff Hinton succeeded in training an artificial neural network to win an image classification challenge with a startlingly large margin. Researchers began looking for artificial data seriously once it was revealed that neural networks could recognize items more quickly than humans.

REFERENCES

I STUDIED VARIOUS ARTICLES TO COME UP WITH THIS REPORT AND VISITED WEBSITES ONLINE.

Wikipedia.

nvidia & blogs.

aimultiple.

syntheticus.

IBM technology on youtube.