# REPORT-2

# (INTERNSHIP)


**PRESENTED TO: Mrs SHOBHARANI MA'AM (BEL)**

**FROM: BHOOMIKA K (AIML, MS RAMAIAH)**

## CONTENTS:

# VARIETIES/TYPES OF SYNTHETIC DATA.

**Synthetic data can be categorized into different types based on its purpose, application, and generation methods. Here are some common types of synthetic data:**

i. **Structured Synthetic Data:**

Tabular Data: Synthetic data generated in tabular form, often resembling spreadsheet-like structures with rows and columns.

Relational Data: Data that mimics relationships between tables in a relational database.

ii. **Time Series Synthetic Data:**

Temporal Data: Synthetic data that represents values over time, commonly used in applications like financial forecasting, stock market analysis, and IoT.

iii. **Textual Synthetic Data:**

Natural Language Text: Synthetic text data generated to simulate language patterns and semantics, often used for training natural language processing (NLP) models.

iv. **Image Synthetic Data:**

Computer Vision Data: Synthetic images generated for tasks like image recognition, object detection, and segmentation.

v. **Graph Synthetic Data:**

Network Data: Synthetic data representing relationships and connections between entities in a network, used in graph-based applications.

vi. **Spatial Synthetic Data:**

Geospatial Data: Synthetic data that mimics geographical information, commonly used in applications like GIS (Geographic Information System) and mapping.

vii. **Multimodal Synthetic Data:**

Combined Data Types: Synthetic data that incorporates multiple data types, such as text, images, and numerical data, to simulate real-world scenarios.

viii. **Healthcare Synthetic Data:**

Medical Data: Synthetic data generated to resemble healthcare records, including patient information, diagnostic data, and medical images, while maintaining privacy.

ix.**Financial Synthetic Data:**

Transaction Data: Synthetic financial data simulating transactions, account balances, and other financial activities for applications like fraud detection and risk analysis.

x. **Sequential Synthetic Data:**

Event Sequences: Synthetic data that represents sequences of events or actions, often used in applications like recommendation systems or process modeling.

xi. **Unstructured Synthetic Data:**

Randomized Data: Synthetic data with no specific structure, often used for testing and stress-testing systems.

xii. **Privacy-Preserving Synthetic Data:**

Anonymized Data: Synthetic data generated with privacy-preserving techniques to protect sensitive information while maintaining statistical validity.

xiii. **Domain-Specific Synthetic Data:**

Specialized Data: Synthetic data tailored for specific industries or domains, such as manufacturing, retail, or education.

xiv.**Meta-Synthetic Data:**

Data About Data: Synthetic data generated to mimic metadata, often used for testing and validation of data processing pipelines.

-The choice of the type of synthetic data depends on the specific requirements of the application or task at hand. Different types of synthetic data are used in diverse.

## WAYS TO GENERATE SYNTHETIC DATA (GENERATION).

Generation involves creating artificial data that mimics real-world data. This can be useful in various fields such as machine learning, data analytics, and privacy protection. Here are some common methods and kinds of synthetic data generation:

**Random Sampling:**

Uniform Random Data: Generating data with random values following a uniform distribution.

Normal Distribution: Creating data with values following a normal (Gaussian) distribution.

Other Distributions: Using other probability distributions like exponential, Poisson, etc., based on the characteristics of the real data.

**Rule-Based Generation:**

Functional Dependency: Creating data based on known functional dependencies between variables.

Constraints and Rules: Applying specific rules and constraints to ensure the synthetic data meets certain criteria.

**Data Augmentation:**

Adding Noise: Introducing random noise to existing data to create variations.

Data Smoothing: Applying techniques to smooth out data fluctuations.

Rotation and Transformation: Modifying data through rotations, translations, or other transformations.

**Bootstrapping:**

Resampling with Replacement: Creating new samples by randomly selecting from the existing data with replacement.

**Generative Models:**

Generative Adversarial Networks (GANs): Using neural networks to generate synthetic data by training a generator to produce data that is indistinguishable from real data by a discriminator.

**Copulas:**

Copula-Based Models: Modeling the dependence structure between variables to generate synthetic data that preserves the multivariate relationships.

**Time Series Simulation:**

Autoregressive Models: Creating time series data by modeling the dependence on past values.

Seasonal Decomposition: Simulating time series data by decomposing it into trend, seasonal, and residual components.

**Domain-Specific Generators:**

Text Generation Models: Generating synthetic text data using techniques like Markov models, recurrent neural networks (RNNs), or transformer models.

Image Synthesis Models: Creating synthetic images with techniques like DeepDream, neural style transfer, or conditional GANs.

**Privacy-Preserving Techniques:**

Differential Privacy: Injecting noise to ensure that the statistical properties of the synthetic data do not significantly differ from the real data.

Data Masking: Redacting or transforming specific attributes to protect sensitive information.

**Hybrid Approaches:**

Combining Techniques: Using a combination of methods to generate synthetic data that closely resembles the characteristics of real data.

-The choice of synthetic data generation method depends on the specific requirements of the task, the nature of the data, and the goals of the data generation process.

# REAL DATA V/S AI GENERATED DATA.

NOW, WHAT IS REAL DATA? AND WHAT IS AI GENERATED DATA?

**REAL DATA:**

Real data refers to information that is collected from actual observations, measurements, or events in the real world. It is the raw, authentic information that is obtained through various means, such as sensors, surveys, observations, transactions, or any other methods of data collection. Real data reflects the characteristics, patterns, and behaviors present in the natural environment or system under study. It is often used as the foundation for training machine learning models, conducting statistical analyses, and gaining insights into real-world phenomena.

**AI GENERATED DATA:**

AI generated data, on the other hand, is synthetic data that is artificially created by algorithms, machine learning models, or other artificial intelligence (AI) techniques. This data is not directly observed or measured from the real world; instead, it is generated to simulate real-world scenarios. The purpose of AI-generated data can vary and may include tasks such as training and testing machine learning models, augmenting datasets, or preserving privacy.

| REAL LIFE: | SYNTHETIC DATA: |
|---|---|
| Precision | Quality |
| Wider Insights | Scalability |
| Crucial for Models | Data Privacy |
| Prediction Accuracy | Overcoming Challenges |
| | Filling Data Gaps |

## Drawbacks of Real-World Data:

It Can Be Expensive and Hard to Get: Getting real-world data can cost a lot of money and take a lot of effort. Gathering, storing, and managing data often needs a big budget and a lot of people.

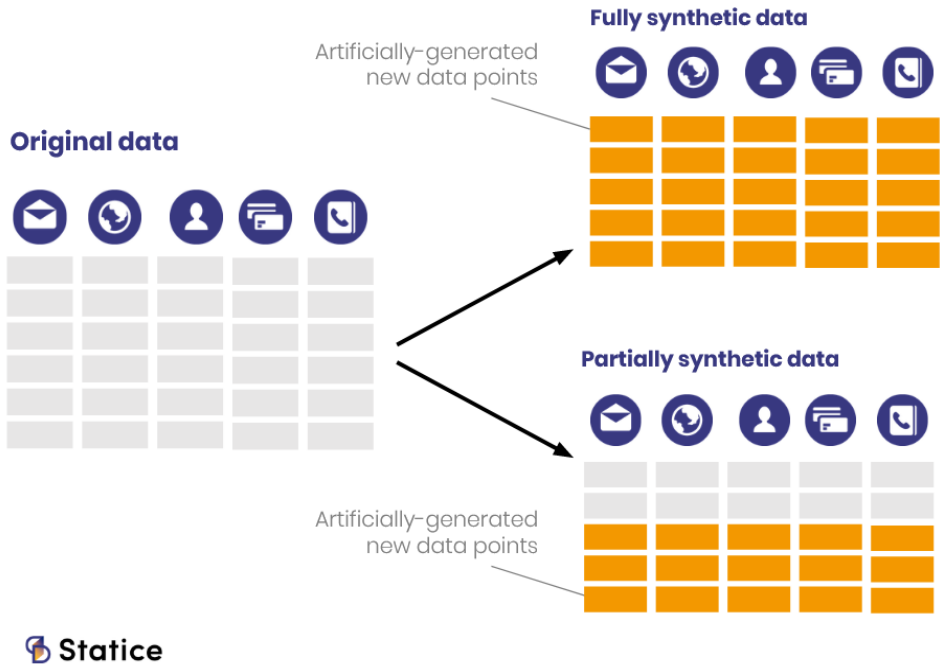
Protecting It Can Be Challenging: Keeping sensitive information in actual data safe requires strong security measures to avoid legal and ethical problems. Mishandling sensitive data can lead to legal and ethical troubles.

Risk of Picking Data with a Bias: When you collect real-world data, there's a chance you might end up with a biased sample. This means the data you get might not represent the whole group or situation you're studying.

## THE CONCEPT OF 30% REAL DATA AND 70% GENERATED:

IN CASE,

if you have a dataset consisting of 30% real data, the remaining 70% would be generated data. This approach is often used in situations where obtaining a sufficient amount of real data is challenging, expensive, or limited. The generated data is designed to supplement the real data, providing a larger and more diverse dataset for tasks such as training machine learning models.



**30% Real Data:**

This portion of the dataset comprises actual observations, measurements, or events obtained from the real world.

It reflects the authentic characteristics, patterns, and behaviors present in the domain of interest.

Real data serves as the foundation for training and testing machine learning models, as it represents the ground truth.

**70% Generated Data:**

This portion is artificially created using algorithms, generative models, or other techniques.

The generated data is designed to simulate the patterns and characteristics of the real data.

It helps to augment the dataset, increase its size, and introduce variations that might not be present in the original real data.

By combining real data with generated data, you can create a more comprehensive dataset, potentially mitigating issues related to data scarcity. However, it's crucial to ensure that the generated data accurately captures the key features and variations present in the real data.

# GAN

**Generative Adversarial Network (GAN)— an algorithm based on two neural networks, working together to generate fake yet realistic data points. One neural network attempts to generate fake data points while the other learns to differentiate fake and real samples. GAN models are complex to train and computationally intensive, but can generate highly detailed, realistic synthetic data points.**

GAN is being widely used to generate photorealistic images and videos, and its application to synthetic data is compelling. However, using GAN for synthetic data generation presents a few challenges:

Models can be difficult to control and might not generate images that fit the requirements of the researcher.

Training GAN models is time consuming and requires specialized expertise. It is also computationally intensive, requiring an investment in computing resources.

GAN models can fail to converge—this means the generator and discriminator do not reach a balance and one overpowers the other, resulting in repetitive output.

GAN models can collapse—this means GAN produces a small set of images with minor changes between them.

Training Process:

Initialization:

The generator and discriminator are initialized with random weights.

b. Adversarial Training:

The generator produces synthetic samples, and the discriminator evaluates both real and synthetic samples.

The discriminator is trained to correctly classify real and synthetic samples.

The generator is trained to fool the discriminator, aiming to generate samples that are difficult to distinguish from real data.

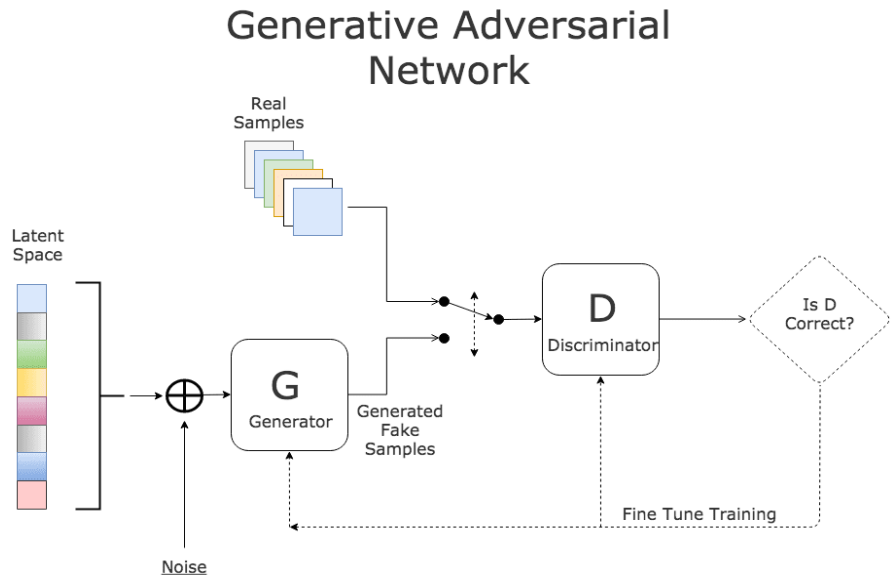
c.Feedback Loop:

The training process iterates in a feedback loop, with the generator and discriminator improving their capabilities over time.

As the generator improves, it becomes more challenging for the discriminator to differentiate between real and synthetic samples.

d. Equilibrium:

Ideally, the training process reaches equilibrium, where the generator generates data that is statistically similar to real data, and the discriminator is unable to reliably distinguish between real and synthetic samples.




Applications:

Image Generation: GANs are widely used for generating realistic images, faces, and artwork.

Data Augmentation: GANs can be employed to augment datasets by generating additional samples for training machine learning models.

Style Transfer: GANs can be used for transferring artistic styles between images.

Super-Resolution: GANs can generate high-resolution images from low-resolution inputs.

Image-to-Image Translation: GANs can convert images from one domain to another (e.g., turning satellite images into maps).

**EX: Synthetic images and videos.**

None of these individuals are real. These synthetic images were artificially generated by the Generative Adversarial Network, StyleGAN2 (Dec 2019) from the work of Karras et al. and Nvidia. The system learned properties of real-life people's pictures in order to generate realistic images of human faces.

# FRAMEWORKS, LIBRARIES, TOOLS.

## FRAMEWORK FOR SYNTHETIC DATA:

**GANs (Generative Adversarial Networks):**

Description: GANs consist of a generator and a discriminator. The generator creates synthetic data, and the discriminator tries to distinguish between real and synthetic data. Through adversarial training, GANs improve the quality of generated data.

**Frameworks: TensorFlow, PyTorch**

VAEs (Variational Autoencoders):

Description: VAEs are generative models that learn a probabilistic mapping between input data and latent space. They can be used to generate new data samples by sampling from the learned latent space.

Frameworks: TensorFlow, PyTorch

**SMOTE (Synthetic Minority Over-sampling Technique):**

Description: Primarily used for imbalanced datasets, SMOTE generates synthetic instances for the minority class by interpolating between existing instances.

Frameworks: Available in various libraries and can be implemented using scikit-learn in Python.

**CTGAN (Conditional Tabular GAN):**

Description: CTGAN is designed for generating synthetic tabular data. It learns the conditional distribution of each column given other columns in the dataset.

Frameworks: PyTorch

**DataWig:**

Description: DataWig is a Python library for missing data imputation and synthetic data generation. It uses deep learning models to predict missing values and generate synthetic data.

Frameworks: TensorFlow, PyTorch

**Faker:**

Description: Faker is a Python library that generates fake data for various types of information, such as names, addresses, and dates. It's useful for creating synthetic datasets for testing purposes.

Frameworks: Pure Python

**Scikit-Multiflow:**

Description: Scikit-Multiflow is a machine learning framework for online and streaming data. It includes tools for generating synthetic datasets for evaluating stream learning algorithms.

Frameworks: scikit-muliflow in Python

**IBM DataSynthesizer:**

Description: An open-source tool by IBM for synthesizing structured data. It uses a Bayesian network to model the relationships between attributes and generate synthetic data.

Frameworks: Python

**Synthetic Data Vault:**

Description: A framework designed for generating synthetic data for healthcare. It includes tools for generating synthetic patient records while preserving privacy.

Frameworks: Synthea (a component of Synthetic Data Vault)

**MOGAN (Mode-Seeking Generative Adversarial Networks):**

Description: MOGAN is an extension of GANs that focuses on generating data with diverse modes, which can be useful for creating diverse synthetic datasets.

Frameworks: TensorFlow, PyTorch.

-When using these frameworks, it's essential to validate the generated synthetic data to ensure that it accurately represents the characteristics of the real data.

## Generating synthetic data using Python-based libraries:

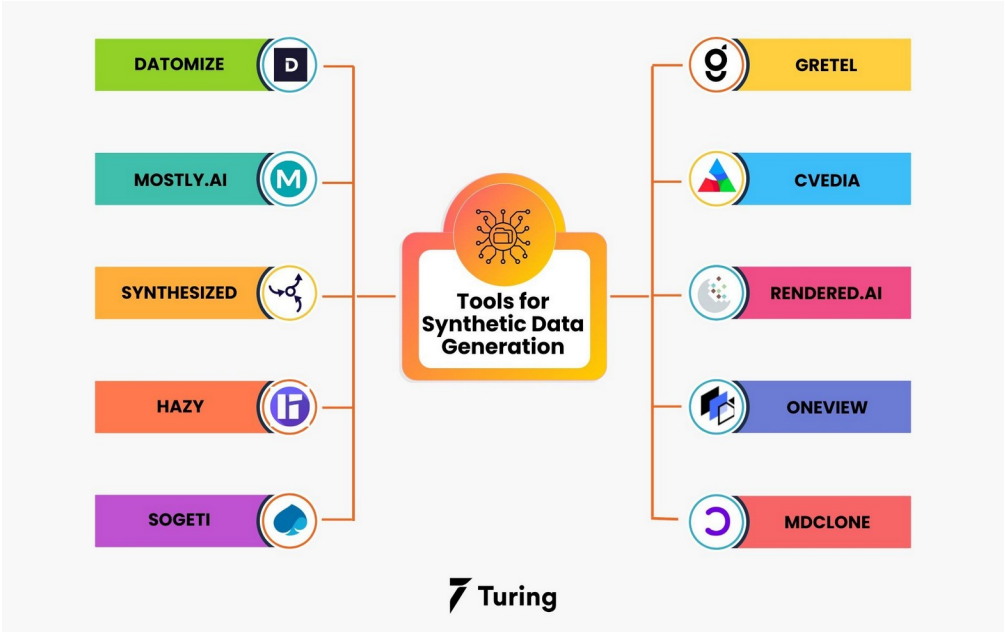
A few Python-based libraries can be used to generate synthetic data for specific business requirements. It is important to select an appropriate Python tool for the kind of data required to be generated.

The following table highlights available Python libraries for specific tasks.

| Purpose | Python Library |
|---|---|
| Increasing data points | DataSynthesizer, SymPy |
| Create fake names, addresses, contact, or date information | Fakeer, Pydbgen, Mimesis |
| Create relational data | Synthetic Data Vault (SDV) |
| Create entirely fresh sample data | Platipy |
| Timeseries data | TimeSeriesGenerator, Synthetic Data Vault |
| Automatically generated data | Gretel Synthetics, Scikit-learn |
| Complex scenarios | Mesa |
| Image data | Zpy, Blender |
| Video data | Blender |

All these libraries are open-source and free to use with different Python versions. This is not an exhaustive list as newer tools get added frequently.

## Synthetic data generation tools.



Datomize: Datomize has an Artificial Intelligence or Machine Learning model which is majorly used by world-class banks all over the globe. With Datomize, you can easily connect your enterprise data services and process high-intensity data structures and dependencies with different tables. This algorithm will help you in extracting behavioral features from the raw data and you can create identical data twins with the original data.

MOSTLY.AI: MOSTLY.AI is a synthetic data tool that enables AI and high-priority privacy while extracting structures and patterns from the original data for preparing completely different datasets.

Synthesized: Synthesized is an all-in-one AI dataOps solution which will help you with data augmentation, collaboration, data provisioning, and secured sharing. This tool generates different versions of the original data, and also tests them with multiple test data. This helps in identifying the missing values and finding sensitive information.

Hazy: Hazy is a synthetic data generation tool that aims to train raw banking data for fintech industries. It will let the developers ramp up their analytics workflows by avoiding any fraudulence while collecting real customer data. You can generate complex data during financial service generations and store it in silos within the company. But, sharing real financial data for research purposes is severely limited and restricted by the government.

Sogeti: Sogeti is a cognitive-based solution that helps you with data synthesis and processing. It uses Artificial Data Amplifier technology which reads and reasons with any data type, whether it's structured or unstructured. ADA uses deep learning methods to mimic recognition capabilities and sets it apart.

Gretel: Gretel is the tool that is specifically built to create synthetic data. It is a self-proclaimed tool that generates statistically equivalent datasets without giving out any sensitive customer data from the source. While training the model for data synthesis, it compares the real-time information by using a sequence-to-sequence model for enabling the prediction while generating new data.

CVEDIA: Packed with different machine language algorithms, CVEDIA provides synthetic computer vision solutions for improved object recognition and AI rendering. It is used for a variety of tools, and IoT services for developing AI applications and sensors.

Rendered.AI: Rendered.AI generates physics-based synthetic datasets for satellites, robotics, healthcare, and autonomous vehicles. It is a no-code configuration tool and API for engineers to make quick changes and analytics on datasets. They can perform data generation on the browser and it will enable easy operation on ML workflows without much computing power.

Oneview: Oneview is a data science tool that uses satellite images and remote sensing technologies for defense intelligence. Using mobiles, satellites, drones, and cameras,

this algorithm will help object detection even where there are blurred images or lower resolutions. It will provide accurate and detailed annotations on the virtually created imagery which will closely resemble the real-world environment.

MDClone: MDClone is a dedicated tool that is majorly used in healthcare businesses for generating an abundance of patient data which will allow the industry to harness the information for personalized care. But, for accessing clinical data, researchers should depend on mediators and the process was slow and limited.

## EXAMPLES OF REAL LIFE APPLICATIONS.

Amazon is using synthetic data to train Alexa's language system

Google's Waymo uses synthetic data to train its self driving cars

Health insurance company Anthem works with Google Cloud to generate synthetic data

American Express & J.P. Morgan are using synthetic financial data to improve fraud detection

Roche is using synthetic medical data for clinical research

German insurance company Provinzial tests synthetic data for predictive analytics

# REFERENCES:

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


**Wikipedia.**

**statice.ai**

**turing.com**

**towardsdatascience.com**

**datagen.tech**

**questionpro.com**

**arxiv.org/pdf/2**