Synthetic Data: Balancing Utility, Fidelity, and Privacy

Definition, Use Cases, Challenges, and Vendor Landscape

Jim Henegan

CDA

January 27, 2025

Overview

1. Definition and History
2. Use Cases
3. Challenges and Key Questions
4. The Utility–Fidelity–Privacy Triangle

What Is Synthetic Data?

- **Artificial Generation**: Synthetic data is generated by computational models: simple statistical processes, agent-based simulations, or machine learning techniques like Generative Adversarial Networks (GANs).
- **Approximation of Real-World Patterns**: Despite not containing actual records, high-quality synthetic datasets preserve many of the relationships, distributions, and characteristics seen in real data.
- **No Direct Personal Identifiers**: Because these datasets do not replicate real individuals' records, they often mitigate or eliminate direct privacy risks.

- Tabular Data
- Text
- Images
- Audio
- Video

Brief History

The concept of simulating data has been around for decades (e.g., in flight simulators and audio synthesis), but **modern synthetic data** for privacy and analytics emerged in the early 1990s when Donald Rubin and Roderick Little proposed adapting multiple imputation techniques (originally used to fill in missing survey responses) to replace or "synthesize" whole sections of sensitive microdata, thus:

- Protecting individual respondents' identities in **public-use datasets** (e.g., national census or household surveys).
- Preserving enough statistical detail for researchers to perform meaningful analyses.

Key milestones include the U.S. Federal Reserve Board's use of synthetic data in the **Survey of Consumer Finances** in 1997 and the U.S. Census Bureau's **SIPP Synthetic Beta** in 2007. Since then, government agencies worldwide have integrated synthetic data approaches to promote data sharing while respecting confidentiality obligations.

Use Cases

- Privacy-Preserving Data Sharing
- System and Software Testing
- Data Augmentation for Machine Learning
- De-biasing and Fairness
- Rapid Prototyping and Product Development
- Data Monetization and Sharing
- Risk Management and Rare Event Simulation
- Regulatory Compliance and Auditing
- Specialized Industry Use Cases

Challenges

- **Data Quality vs. Privacy**. High privacy constraints can distort data distributions, reducing utility.
- **Embedded Bias** Generative models can replicate existing biases from the real dataset.
- **Overfitting and Leakage**: If the model "memorizes" actual records, re-identification risk increases.
- **Complex Data Structures** Multi-table, time-series, or high-dimensional data are harder to synthesize accurately.

Key Questions

Can synthetic data be used in place of real data to do the same tasks (training models, hypothesis testing, data analysis)?

Yes in principle, but specialized methods (e.g., Bayesian updates for model parameters, bias-correction strategies) often improve accuracy when dealing with synthetic data.

Can synthetic data be treated exactly like real data (e.g., linking records from different datasets)?

Linking different synthetic datasets independently generated from real data can break 1-to-1 correspondence across records. Specialized approaches—or regenerating a single, joint synthetic dataset—may be needed.

The Utility–Fidelity–Privacy Triangle

- **Utility**. How well does the synthetic data serve the target task (e.g., training an ML model, performing analytics)?
- **Fidelity**. Statistical similarity to the original dataset (distributions, correlations, outliers).
- **Privacy**. Degree to which re-identification risks and sensitive details are eliminated.

This leads to the following **balancing act**:

- Increasing **fidelity** can reduce **privacy**.
- Overly strict **privacy** can lower **utility** and **fidelity**.
- Finding the "sweet spot" depends on your **use case** and **risk appetite**.

Differential Privacy

Evaluating Fidelity