DataGenesis

Synthetic population
powered by AI

Synthetic population data for automated testing and fraud detection

December 2025

Federal tax administration agencies face a critical "Compliance Paradox": modernizing legacy systems requires high-fidelity testing data, yet security mandates demand an absolute Zero-Trust posture regarding Production PII (Personally Identifiable Information).

Traditional testing methods often rely on masking or obfuscating production data. This approach is costly, risky, and frequently results in "data swamps"—datasets that retain the structural messiness of production without the verifiable logic needed for precise testing. Furthermore, strict regulations (FISMA, NIST) often trap this data in siloed environments, preventing the agility required for modern DevOps Automation.

DataGenesis resolves this paradox by delivering **Greenfield Provenance**. Rather than scrubbing real records, the engine generates synthetic entities from scratch using authoritative population heuristics (e.g., US Census data). This ensures that the resulting data is:

- **Zero-Trust:** Contains no direct or indirect PII information at any step, eliminating re-identification risk.
- **Functionally isomorphic:** Mathematically indistinguishable from production data, preserving the statistical shape and correlations of the real world.
- **Longitudinally consistent:** Capable of "aging" entities over time to support multi-year tax code validation.

Sensitive Personal data (production environment)	Synthetic data (test automation environment)
High risk: Retains re-identification vectors, restricted by privacy laws	**Zero risk:** synthetic “Greenfield” origin. Safe for lower environments
Static snapshots: Represents a single moment, fails to model progression	**Temporal dynamics:** Ages data to test carryovers, depreciation, events, multi-year tax code support
Reactive: Limited to historical fraud patterns found in past audits	**Proactive:** Supports adding new fraud patterns for model training
Siloed: Difficult to move between Dev, Test, and QA due to security protocols	**Portable:** Containerized architecture deployable in GovCloud / OpenShift

The Synthetic Digital Twin

DataGenesis operates as a **socioeconomic simulation engine** rather than a simple data generator. It builds a **synthetic Digital Twin** of the taxpayer population, layered to provide depth and context.

- **Baseline Reality Layer:** Ingests statistical heuristics from 140+ public sources (Census, Bureau of Labor Statistics) to establish the ground truth.
- **Synthetic Population Layer:** Generates millions of distinct entities (People, Households, Businesses) with over 131 dimensionalized socioeconomic factors. Synthetic data is indistinguishable from real data.

- **Temporal Logic Layer:** Applies statistical aging coefficients. A "Person" entity isn't just a static record; they get married, change jobs, accumulate assets, and age out of dependent status over simulated years.
- **Micro Scenario Layer (Individuals and Businesses):** Ability to increase probabilities of life scenarios like octuplets, fraud,
- **Macro Scenario Layer (The Multiverse):** Allows analysts to branch reality. One dataset can simulate a "Business as Usual" economy, while another simulates a "Recession" scenario with higher unemployment and lower capital gains.

Use cases: IRS SimoN and Fraud Detection

The **Simulation of the Nation (SimoN)** project at the IRS represents the gold standard for enterprise-scale synthetic data for test automation. As the prime contractor, Skymantics has evolved SimoN from an initial proof-of-concept into a fully operational system over four contract phases.

- **Enterprise scale:** The DataGenesis system installed at the IRS simulates real-life events for a population of over **13,000,000 individuals**, 6.4M households and 360,000 businesses.
- **MeF compliance:** It generates over 40 high-fidelity tax returns compliant with **Modernized e-File (MeF)** schemas and internal integration formats.
- **System integration:** SimoN feeds synthetic data directly into critical IRS downstream systems, including the **Individual Master File (IMF) and Business Master File (BMF)**, to support performance testing and Enterprise System Testing.
- **Ecosystem simulation:** Beyond tax forms, the engine simulates the broader federal data ecosystem, creating **Social Security Administration (SSA) DM1 records** for SSN assignment and **IRS Form SS-4** records for EIN assignment.

DataGenesis empowers agencies to move from reactive auditing to proactive fraud prevention by using **Intentional Data Poisoning**—injecting specific, controlled anomalies into the dataset.

Multi-year identity theft

A fraudster uses a synthetic identity to file legitimate low-value returns for years to build trust before filing a massive refund request.

DataGenesis uses the Age function to generate a compliant 3-year history for an entity, then uses the Merge function to inject a high-value refund request in Year 4.

This allows frauds models to be trained on longitudinal anomalies that static data would miss.

Shell-company tax evasion

A business reports high gross receipts but offsets them with fictitious expenses to zero out tax liability.

DataGenesis generates business activity records where expenses (e.g., massive advertising costs for a small local service business) are statistically inconsistent with the entity's NAICS code.

These records verify if the automated business rules (IMF) correctly flag the discrepancy.

Core technology and enterprise integration

- **Generative AI engine:** Utilizes AI/ML techniques to analyze public datasets and create a simulated population that evolves.

- **High-dimensional sanity:** Enforces strict logic across variables. For example, it ensures that a specific "Occupation" correlates with the appropriate "Income Range" and "Asset Class," maintaining the statistical validity required for reliable analytics. To date, DataGenesis incorporates over 400 dimensions of data ensuring that data is realistic.

- **Automated MeF validation and verification:** The engine includes a built-in validation suite. Every record is checked against 14,000+ business rules (e.g., IMF logic) to ensure schema compliance before it ever reaches the test environment.

- **Automated Synthetic Data Validation and Verification:** DataGenesis contains multiple levels of QA spanning statistical inputs, AI models and synthetic data level quality controls to ensure that the data has been generated and aged, year after year.

DataGenesis is engineered to fit seamlessly within the modern federal technology stack, ensuring that synthetic data generation is an automated utility, not a manual bottleneck.

- **Infrastructure portability:** The solution is deployed in the Enterprise Cloud Platform (ECP) PaaS environment. Its containerized architecture is also fully compliant with AWS GovCloud, Azure and OpenShift.

- **CI/CD pipeline integration:** SimoN integrates directly with the IRS-owned CI/CD pipeline. It supports headless execution, allowing Test Automation teams to trigger the generation of fresh, validated data for every nightly build or sprint cycle.

- **Agile delivery:** The platform is delivered using the IRS One Solution Delivery Lifecycle (OneSDLC) framework, ensuring all data artifacts meet strict federal quality control plans (QCP).

- **Downstream ETL:** The engine handles the complex ETL required to format synthetic entities into regulatory standards, exporting data in XML (MeF schemas), JSON, and CSV for immediate consumption by legacy and modernized systems.

DataGenesis
by skymantics

SBA 8(a)-certified Disadvantaged Small Business

We provide AI, data and geospatial solutions in
Government, healthcare and financial industries

www.skymantics.com

info@skymantics.com