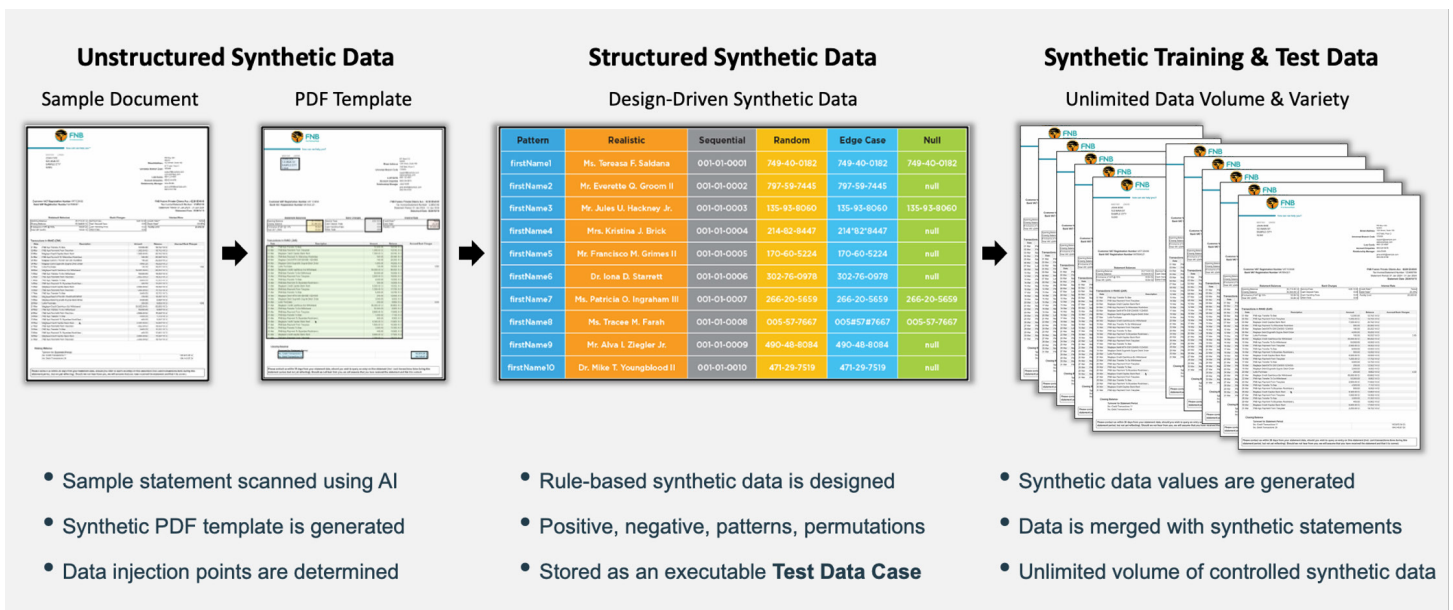


SOLUTIONS GUIDE

THE UNSTRUCTURED DATA ACCELERATOR (UDA)

UDA Capabilities Overview

The GenRocket Unstructured Data Accelerator (UDA) is a breakthrough enhancement for the GenRocket Synthetic Data generation platform. **UDA allows unstructured data or media to be combined with structured synthetic data that is controlled and conditioned to meet any training or test data requirement.** Now enterprises can quickly and easily generate unstructured data in the form of documents, ID cards, audio clips, or conversational text and blend it with structured data for testing existing OCR and document processing systems or training new and advanced AI-driven applications.



For organizations in regulated industries like financial services and healthcare, UDA makes it possible to safely test, train, and automate document-heavy workflows without accessing or exposing sensitive customer or patient data.

Design-Driven Structured Synthetic Data

One of the key capabilities of the genrocket platform is the ability to generate controlled and conditioned structured (tabular) synthetic data to provide complete coverage of all training and test data scenarios. This capability, called **Design-Driven Data**, allows control over the volume, variety, and format of structured synthetic data that is generated on-demand and in real-time. By leveraging GenRocket’s more than 750 intelligent data generators, synthetic data can be produced with any or all of the following conditions.

The chart below shows several examples of how Design-Driven Data can maximize coverage using both structured and unstructured data for both positive and negative test case scenarios. With GenRocket, synthetic data is always generated with referential integrity and in any volume required.

Design-Driven Synthetic Data Scenarios

Data Type	Positive	Negative	Volumes
Structured Data	- Realistic Data - Patterned Data - Combinations - Permutations - Edge Cases - Statistical Profiles	- Illegal Data - Inaccurate Data - Null/Empty Data - Invalid Formats - Rule Violations - Anomaly Conditions	Unlimited
Unstructured Data	- Field Boundaries - Checkbox Marking - Handwriting Styles - Variable Ink Colors - Variable Font Styles - Facial Image Variation	- Crumpled Forms - Skewed Images - Ink & Coffee Stains - Torn & Cropped Edges - Illegible Handwriting - Overwritten Text	Unlimited

Flexible Unstructured Synthetic Data

The power of the GenRocket platform is the ability to combine structured data with virtually unlimited use cases for unstructured synthetic data in the form of PDF documents, images, text and audio files. If physical documents are being digitized for critical processing functions such as check cashing or insurance claims processing, then it's essential these documents be templated intelligently and synthetic versions of these documents generated on demand.

In the case of check cashing, applications can be trained with a variety of positive and negative and edge cases in terms of image recognition. Below is an example of synthetic bank checks that were designed and generated for variety of challenging use cases to train image processing and OCR systems for accuracy and predictability.

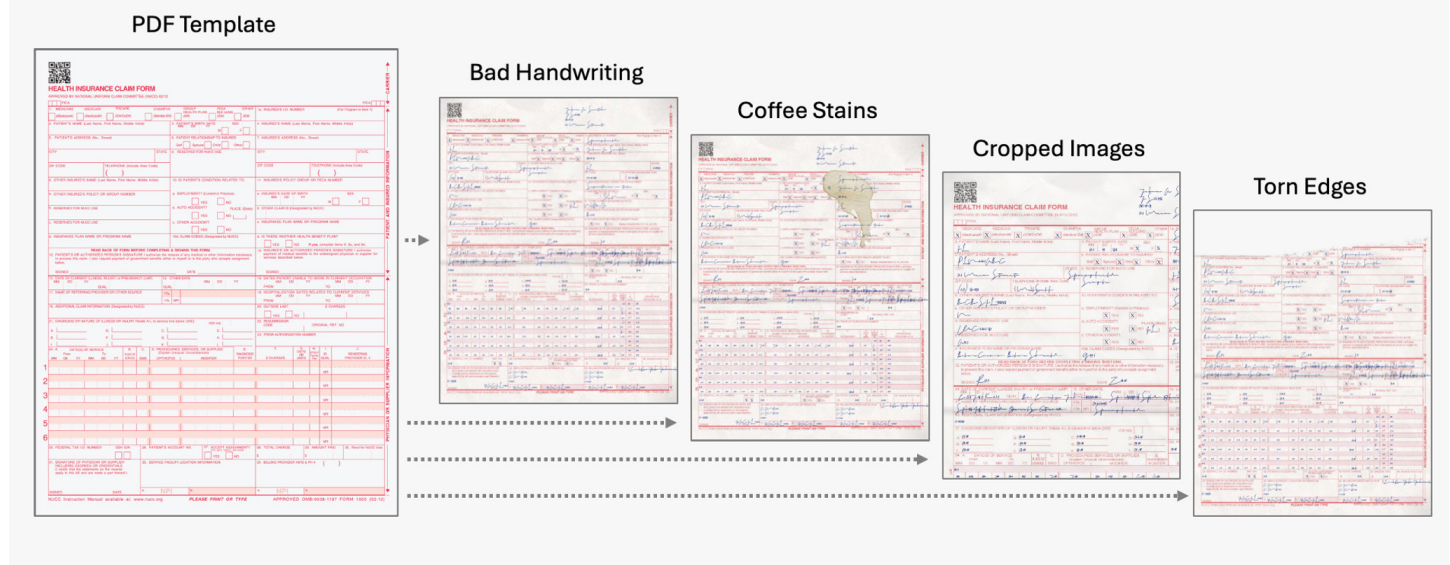


Bank checks simulating positive and negative image capture scenarios

In the case of insurance claims, paper forms can be simulated with a variety of image quality defects and positive/negative data values so they can be handled accurately and efficiently by the systems being trained and tested.

Typical scenarios might include images of crumpled paper, torn pages, coffee stains and ink blots that compromise the form data. Similarly, a variety of handwriting styles, legibility, and ink colors may be encountered by these systems. GenRocket's UDA solution has the ability to simulate all of these positive and negative scenarios at scale.

Controlled Synthetic Images Improve Test & Training Accuracy



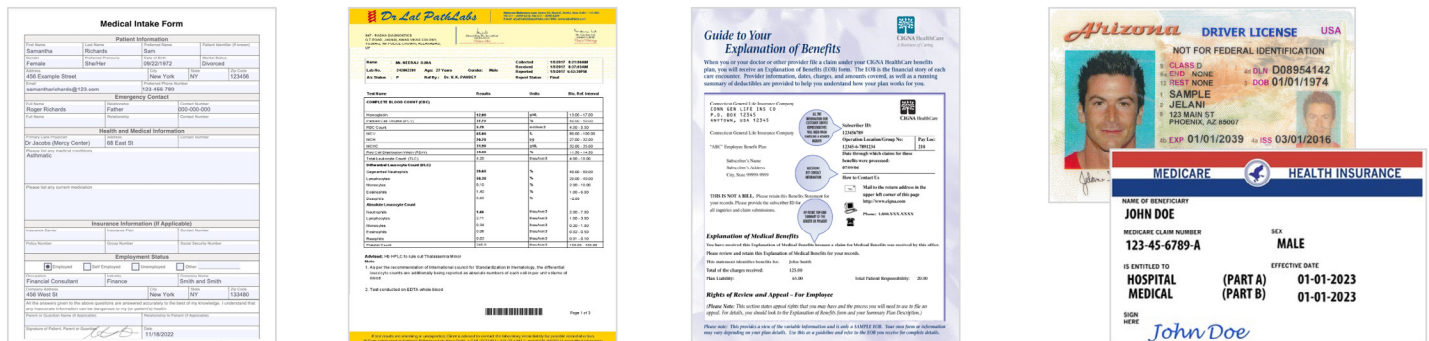
Health insurance claims forms with positive and negative image capture scenarios

The unique power of UDA is its ability to combine *design-driven* structured data with flexible unstructured data to produce high quality training and test data. The UDA solution can generate an unlimited volume of synthetic data through the use of executable *Test Data Projects* that can be categorized, stored, reused, shared, repurposed, and version controlled.

- Generate synthetic bank statements, trade confirmations, and wire transfer forms to test transaction workflows without exposing customer data.
- Create synthetic claims forms and loan applications to validate document capture and processing systems.
- Produce regulator-required formats safely in test environments for compliance testing.
- Generate synthetic audio recordings to train call center staff and validate speech analytics tools.

Healthcare

In healthcare, any combination of documents and images can be simulated synthetically. PDF templates can simulate insurance claims and laboratory documents. Scanned identification cards can be generated for testing patient verification and eligibility systems. And audio clips can be used to train and test text-to-speech or speech-to-text systems.



- Create synthetic patient intake forms, billing statements, and insurance ID cards to test EHR and billing workflows in a HIPAA-compliant manner.
- Generate synthetic documents to validate EMR/EHR system workflows without using real patient data.
- Produce synthetic ID cards combined with structured data to support secure identity verification systems.

Customer Challenges Solved with UDA

Compliance risk: Production documents often contain PII/PHI and can't be safely used.

Limited coverage: Real data rarely provides edge cases and negative testing scenarios.

Scaling needs: Training and testing AI/ML systems require massive volumes of data.

Integration gaps: Integration gaps: Data must be delivered seamlessly across the DevOps environment.

Benefits Realized

- Accelerates testing and training by providing large, diverse, and realistic datasets.
- Eliminates compliance risk by replacing real documents with synthetic documents.
- Ensures full coverage with the ability to generate unlimited data variations.
- Scales with CI/CD pipelines to support automated and integrated DevOps workflows.
- Boosts AI/ML model accuracy with controlled, balanced, and conditioned data.

How UDA Delivered

UDA is a modular add-on to the GenRocket platform, not a standalone tool. It extends synthetic data into the world of unstructured data with software and professional services that complement the existing GenRocket synthetic data platform. UDA will be configured with the required number of *Test Data Projects* and *Navigator Services* to ensure a successful implementation that meets your unstructured data requirements.

If your organization is addressing some of the challenges and use cases described in this document, we are ready to help you explore the most effective UDA solution. Please contact your GenRocket account director for additional information. They will schedule a discovery session with our UDA experts to discuss your specific use case and how GenRocket's UDA can meet your requirements.

