LAB L5 LIVE DATA GENERATION & PRIVACY

Synthetic Data Foundry

GANs - Diffusion - Tabular - Differential Privacy

High-fidelity synthetic datasets that train, test and audit AI without touching real customer data.

Thesis: If your team needs prod data to ship, you've built a privacy debt machine. Synthetic-first is faster, safer, cheaper.

COLD-START ACCURACY	TIME-TO-TEST-DATA	PRIVACY DISCLOSURE RISK
+ 22 pts	6w 2h	1.0

MANIFESTO

Why this lab exists

Real customer data is a liability and a bottleneck. The Foundry generates statistically faithful, privacy-preserving synthetic data - tabular, text, image, time-series - so teams can train, test and audit models without ever touching production records.

KPIS

Outcomes we measure

- Cold-start accuracy: + 22 pts
- Time-to-test-data: 6w 2h
- Privacy disclosure risk: 1.0

ACTIVE EXPERIMENTS

What the lab is testing now

> Tabular diffusion vs CTGAN
Benchmarking on 18 enterprise schemas for utility, fidelity and privacy.

> Privacy-preserving text
DP fine-tuning + retrieval-conditioned generation for support transcripts and clinical notes.

> Time-series fidelity
Synthetic transaction streams that pass downstream fraud-model evals within 2 pts of real.

> Audit-by-replay
Synthetic populations for stress-testing models on edge cases and protected classes.

SHIPPABLE ARTEFACTS

Everything that ships

> Generator library
Tabular, text, image and time-series generators with utility + privacy reports per release.

> DP toolkit
/ accounting, privacy-budget tracking, formal disclosure-risk certificates.

> Utility evals
TSTR, downstream-task and population-statistics test suites - pass/fail in CI.

> Audit datasets
Curated synthetic populations for fairness, robustness and edge-case stress.

> Schema-aware connectors
Snowflake, BigQuery, Databricks, Postgres - discover schema, generate, write back.

LAB TEAM

Who you'll work with

- Synthetic Data Principal
- Generative Modelling Researcher
- Privacy & DP Specialist
- Eval / Utility Engineer

ENGAGEMENT TIMELINE

Weeks 1-8 - first synthetic corpus shipped by week 3

1. **Weeks 1-2 - Schema + utility baseline**
   Discover schemas, agree utility tests, set DP budget.
2. **Weeks 2-5 - First synthetic corpus**
   Generate, evaluate, sign DP certificate, ship to dev / test environments.
3. **Weeks 5-8 - Production loop**
   CI integration, audit datasets, automated regeneration on schema drift.

FLAGSHIP PODS

Squads that productionise this lab

- Cold-Start Model Pod
- Privacy & DP Pod
- Audit Population Pod
- Synthetic Time-Series Pod

PARTNERS

Who we collaborate with

Snowflake - Databricks - Microsoft Fabric - OpenAI - Hugging Face - MOSTLY AI

PUBLICATIONS

Receipts

Tabular diffusion outperforms CTGAN on enterprise schemas
ICML Workshop on Synthetic Data - 2025

Audit-by-replay: stress-testing fairness with curated synthetic populations
AXP Internal Whitepaper - 2026

FAQS

What partners actually ask

Q. Is synthetic data really safe?
A. Under formal differential privacy with 1.0 and signed certificates - yes. We publish the disclosure-risk number, every time.

Q. Do downstream models suffer?
A. Utility evals (TSTR, population stats) gate every release. We don't ship corpora that fail.

Q. What schemas are supported?
A. Tabular (Snowflake / BQ / Databricks / Postgres), text, image, time-series. Schema-aware connectors discover and write back.

Q. Can we audit fairness with this?
A. Yes - that's a primary use case. Curated synthetic populations stress-test on protected classes without touching production.

Partner with Synthetic Data Foundry

Outcome-priced. Sovereign by default. Refund-backed if the contracted KPI isn't hit.

Apply: alfaxprienz.com/labs/synthetic-data-foundry#partner