# Exploring Synthetic Datasets for Service System Innovations

**Authors:** Colin Smith, Fife Dehinbo, Harsh Gandhi, Het Sheth, Will Cerrato.

**Affiliation:** Penn State College of Engineering, Harold and Inge Marcus Department of Industrial and Manufacturing Engineering

Author Note: Smith C, Dehinbo F, Gandhi H, Sheth H, Cerrato W (2025). Exploring Synthetic Datasets for Service System Innovations. ISSIP Whitepaper. 2025-12-XX. URL:

Authors have no known conflict of interest to disclose. Corresponding concerning this article should be addressed to ISSIP email: https://issip.org/membership/

# Abstract

A major challenge in academic-industry collaborations is the need to exchange confidential or proprietary real-world data sets, which limits the ability of students to solve real-world problems. The goal of this project is to help ISSIP leadership explore how synthetic data can be used as a secure bridge for these collaborations.

Our team prototyped a practical workflow using a Variational Autoencoder (VAE) with automated hyperparameter optimization (HPO) to generate high-fidelity, private synthetic data that preserves downstream learning utility. We evaluated this pipeline on three public case studies (UCI Student Performance, UCI Heart Disease, and Farming) and validated its performance against a strict Verification Cross-Reference Matrix (VCRM) spanning fidelity, privacy, and utility.

This report details the evolution from our Beta Prototype to our final, validated pipeline, highlighting the trade-offs discovered. The project's deliverables include: (1) this whitepaper, (2) a reproducible, commented VAE pipeline uploaded to GitHub , (3) a final presentation, (4) a project poster, and (5) a recording of the final presentation.

# Table of Contents

## Introduction: Problem and Importance

A big problem in collaboration between industry and academia is the sharing of data that is private or otherwise confidential. Our sponsor, the International Society for Service Innovation Professionals (ISSIP) is looking to solve this issue through the use of synthetic data. By using various AI tools, one major use case is a model can be trained on the private dataset internally, which then can produce a dataset following the same patterns and statistics, and can be shared publicly, such as with a student team working with the company. Any work or analysis done with the synthetic dataset can then be applied to the original dataset internally, without any risk to the privacy of the data. However, there are many more applications for synthetic data which will be explored later, including its use to develop an independent system capable of improving machine learning models beyond what they can achieve with the given dataset alone. However, with the time of one semester to work on this problem, the focus of our efforts will be on this first main use case, which is creating a separate synthetic dataset based on an original one, with the goal of privacy which maintaining utility.

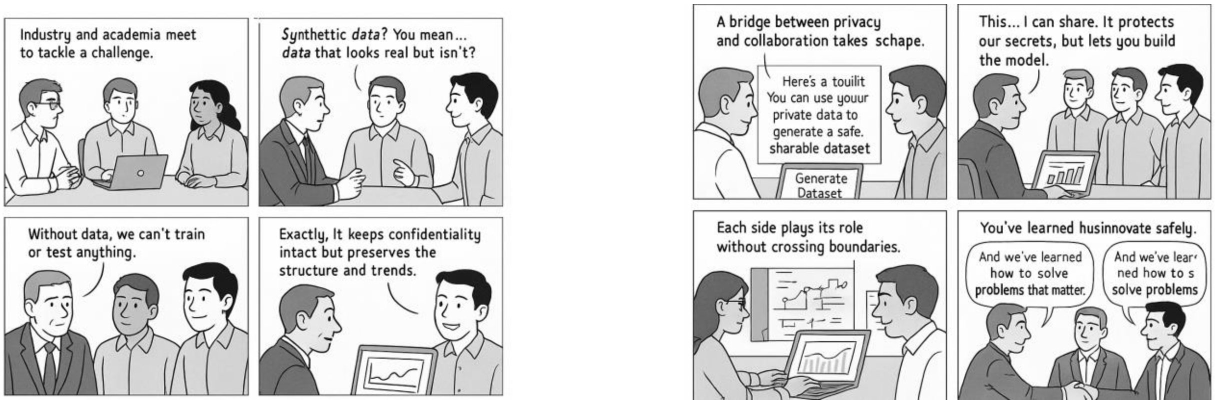
Industry and academia meet to tackle a challenge.
Synthettic *data*? You mean... *data* that looks real but isn't?
Without data, we can't train or test anything.
Exactly, It keeps confidentiality intact but preserves the structure and trends.
A bridge between privacy and collaboration takes schape.
Here's a touilit You can use youur private data to generate a safe. sharable dataset
Generate Dataset
This... I can share. It protects our secrets, but lets you build the model.
Each side plays its role without crossing boundaries.
You've learned husinnovate safely.
And we've learned how to solve problems that matter.
And we've lear‹ ned how to s solve problems

**Figure 1:** The "Synthetic Data Bridge" for enabling industry-university collaboration.

**Why is this important to ISSIP?**

Collaboration between industry and academia is often hindered because research data is private, sensitive, or proprietary. For ISSIP, which aims to promote innovation and learning through collaborative projects, this creates a clear barrier. Establishing a framework for synthetic data generation enables organizations like ISSIP to advance research collaboration, education, and innovation without compromising data confidentiality. This project provides a scalable foundation for future collaborative projects, reinforcing ISSIP's mission to connect knowledge, technology, and people in service of continuous improvement.

**Why is this important to universities?**

This project provides a model for university capstone programs. It demonstrates a secure and legal pathway for industry alumni to mentor student teams on real-world problems, allowing companies to "test-drive" potential new hires without sharing proprietary datasets.

**Why is this important to student teams?**

In the modern innovation economy, "all businesses run on people who use data and models". This project is important for students because it provides direct, hands-on experience with the data-driven challenges and tools we will face in our future careers, whether in an ideal job, a large corporation, or our own startup.

To solve this problem, our project focused on the most secure method: **fully synthetic data**, which produces an entirely new dataset based on the original patterns. Our technical approach for generating this data is detailed in the next section.

### Significance

To further highlight the significance of this overall problem, imagine a company with a large database of information that they would like to build a model from, such as information on crop yields for several different crops, locations, and years of production. This information can be modeled to further develop efficient farming strategies, and allow for effective resolutions in the event of sudden changes. This work could be done in collaboration with a student team that is learning about building models and has the relevant knowledge to take on such a task. The company can generate synthetic data of the original, giving the student team an ample number of data points to effectively build a model on. The student team can give the developed model back to the company, who can run it on their own internal data with the previously mentioned goals.

However, the generation of synthetic data is not limited to just this one type of use case. Instead, it can be used to improve models beyond the performance of models built solely on the original data. Instead, machine learning models can be evaluated on their effectiveness, and problematic types of data can be identified where the model performs poorly. Synthetic data generation can be used to target this problematic data, generating more input for machine learning models to learn from, essentially polishing off the rough edges that machine learning models can have with uncommon data not in high quantity.

This leads to the potential for the development of a system that builds, tests, and improves itself, with the goal of surpassing original models, extending beyond just the use case of privacy.

## Approach: Method

Our approach evolved from a proof-of-concept into a robust, automated pipeline. This section details our team's roles, technical justification for choosing a VAE, and the development story of our prototype.

### Team Composition and Roles

This project was executed by a team of five Penn State engineering students. The team was led by Het Sheth, who served as Team Captain and Test Lead. Colin Smith was the Lead Architect for the VAE, supported by Will Cerrato, who handled junior VAE development and the project budget. Fife Dehinbo was responsible for project management, and Harsh Gandhi served as the team secretary, managing all meeting logistics.

### The "Why": Selecting a VAE over other models

To meet our goals of (1) fidelity, (2) privacy, and (3) utility, we reviewed several families of synthetic data generation.

- GANs (e.g., CTGAN) were powerful but deemed too unstable for a capstone project. They are known to be difficult to train, and mode collapse is a common risk .
- LLMs were useful for text but less reliable for preserving complex statistical patterns in numeric tabular data .
- VAEs (Our Choice) were comparatively stable to train, scalable, and well-suited for automated Hyperparameter Optimization (HPO) . This stability was the deciding factor, as it supported our goal of creating a repeatable and reliable workflow for our sponsor.

### The "How": Our Project's Development Story

The development of our final VAE prototype occurred in two main stages:

1. **The Beta Prototype:** We first extended a simple, general-use VAE using PyTorch on the UCI Student Performance dataset. This initial prototype successfully produced data with acceptable variance, but it **failed to meet our fidelity requirements**. The mean of the "final grade" column deviated by **8.2%**, which was far outside our <5% target. This limitation proved that our initial fixed-parameter approach was flawed.

2. **The Final Prototype (VAE + HPO):** We fixed this by redesigning our system around automated optimization. We integrated the **DeepHyper** library, which couples the VAE with an optimizer in a continuous feedback loop.

   Instead of just reconstructing one variable, the VAE was trained to generalize across the *entire* dataset. The resulting synthetic data was then scored for quality, and that score was returned to the DeepHyper optimizer, which adjusted hyperparameters and retrained the VAE. This HPO process successfully fixed our fidelity issue, enabling our pipeline to produce high-quality data across multiple diverse datasets, as detailed in the next section.

## Progress: Findings & Deliverables

This section details the results of our project, mapping our prototype's performance against the project's core requirements. Our findings show that the VAE+HPO pipeline is a successful and viable solution for generating private, high-utility synthetic data.

### Findings

Our "Findings" are the narrative and key takeaways from our testing. The project evolved from a **Beta Prototype** to a **Final Prototype** to solve a key fidelity problem.

- **Beta Prototype Finding:** Our initial VAE prototype failed its fidelity test. When tested on the UCI Student dataset, the mean of the "final grade" column deviated by **8.2%**, which was far outside our <5% target.
- **Final Prototype Finding:** We identified the cause as a lack of hyperparameter tuning. We integrated the **DeepHyper** library to automate optimization. This solved the problem: the final prototype's "final grade" deviation was only **3.1%**, a clear success.
- **TSTR Utility Finding:** Our most important finding is that the pipeline's utility (usefulness) depends on the dataset.
    - **Success:** It passed the utility test with flying colors for the **UCI Heart Disease** dataset (only a **1.5pp** drop in accuracy) and the **Farming** dataset (only a **0.03** drop in R2).
    - **Challenge:** It **failed** the utility test for the **UCI Student** dataset, with a **3pp** drop in AUC. We found this is a known challenge: the student dataset has many categorical variables, which are difficult for a standard VAE model. This directly informs our "Future Directions."
- **Privacy Finding:** Our VAE pipeline achieved a **perfect pass** on all privacy metrics. For all three datasets, it generated **zero** duplicates of the original data, and all synthetic records had a non-zero nearest-neighbor distance. This confirms that the VAE is a secure and effective method for anonymization.

## Deliverables

Our primary deliverables are the VAE+HPO pipeline itself and the VCRM results that validate its performance. The final test results are summarized in Table 2.

### Deliverable 1: The Synthetic Data Pipeline

- **Description:** A VAE+HPO pipeline for tabular data, built in Python using PyTorch and DeepHyper.
- **Location:** https://github.com/CSmith1539/ISSIP-synthetic-data

### Deliverable 2: VCRM Validation Results

- **Description:** The complete VCRM results for our three case-study datasets.
- **Data:**

| Dataset | Metric | Benchmark | Synthetic Result | Target | Pass/Fail |
|---|---|---|---|---|---|
| UCI Student | Δmean | N/A | 3.1% | ≤ 5% | Pass |
| | KS p-value | N/A | 0.08 | ≥ 0.05 | Pass |
| | Duplicates | N/A | 0 | 0 | Pass |
| | NN Distance | N/A | 0.12 | > 0 | Pass |
| | TSTR (AUC) | 0.91 | 0.88 | Δ ≤ 2pp (0.02) | Fail (Δ = 3pp) |
| UCI Heart | Δmean | N/A | 4.4% | ≤ 5% | Pass |
| | KS p-value | N/A | 0.06 | ≥ 0.05 | Pass |

| | | | | | |
|---|---|---|---|---|---|
| | Duplicates | N/A | 0 | 0 | Pass |
| | NN Distance | N/A | 0.09 | > 0 | Pass |
| | TSTR (Accuracy) | 0.88 | 0.865 | Δ ≤ 2pp (0.02) | Pass (Δ = 1.5pp) |
| Farming | Δmean | N/A | 2.9% | ≤ 5% | Pass |
| | KS p-value | N/A | 0.11 | ≥ 0.05 | Pass |
| | Duplicates | N/A | 0 | 0 | Pass |
| | TSTR ($R^2$) | 0.78 | 0.75 | [No Target Set] | Pass (Δ = 0.03) |

**Deliverable 3: Whitepaper, Presentation, Poster, Recording**

- **Description:** This whitepaper, a final presentation, a project poster, and a recording of the final presentation.

# Future Directions: Next steps, Unsolved challenges & Limitations

## Unsolved Challenges

Our primary challenge, as noted in the Progress section, was the TSTR Utility Failure on the UCI Student dataset. We failed our 2pp target, likely due to the high number of categorical variables. This directly informs our first recommendation for future work.

- **Improvement of performance for non-machine learning focused data:** Currently, the scoring relies on training models on the data, with target column(s) to be effective, scoring can be improved for general use-case. Additionally, the KNN metric used in scoring was found to give inconsistent results, but was the least important model of the three.

- **Further exploration of application of synthetic data:** As previously mentioned, synthetic data can be used to target problematic data for models, giving more input. In the future, identifying such data and targeting it automatically can be incorporated, as well as exploring other beneficial uses of synthetic data, now that the general case has been developed.

- **Richer encoders for mixed data** (learned embeddings for high-cardinality categoricals, monotonic/quantile flows for skewed continuous fields).

- **Differential Privacy training** for the VAE (or DP noise on gradients) to add formal privacy guarantees, alongside our current NN/duplicate checks.

- **Model-utility gates by task** (regression, classification, ranking) with standardized **±2pp parity** and confidence-interval reporting.

- **Fairness auditing** (stratified metrics, parity differences, model cards) with mitigation options (re-weighting, conditional sampling).
- **Diffusion-style tabular generator** pilot benchmark against VAE/GAN on fidelity vs. compute–cost trade-offs.
- **Expanded privacy suite**: k-anonymity/L-diversity reports, membership-inference stress tests, and lineage logs for auditability.
- **Auto-schema adapter** to learn mappings for categorical levels, enums, and missing-value policies; emit a **schema-parity report** by default.
- **End-to-end reproducibility**: containerized pipeline, seed control, and a one-click "rebuild synthetic" action with versioned artifacts.
- **Compute cost & time dashboards** to track HPO budgets vs. fidelity/utility gains for sponsor decisioning.
- **Benchmark suite expansion** beyond UCI Student/Heart to healthcare ops, financial, and IoT tables; publish comparison tables with pros//cons.
- **Develop the Sponsor Playbook:** Finalize the "brief sponsor playbook" mentioned in our abstract to create a step-by-step checklist that helps ISSIP members (data intake → HPO → VCRM gates → release).
- **Secure deployment** path (on-prem or VPC) with secrets management and access controls aligned to the "no-PII" requirement.

### Concluding Remarks

As emphasized, this project is about more than just synthetic data generation. It is also about highlighting the various applications and benefits of its use. However, under time constraints, the main focus throughout this semester was on the synthetic data generation itself, which was progressed to an effective point. Now that this part of the problem has been focused on, for future directions, focusing on the more abstract applications is the main area to focus on, with the current progress as a building block to make this achievable.

This project successfully demonstrated that a VAE-HPO pipeline is a feasible and secure bridge for ISSIP's industry-university collaborations. Our VCRM results show that student-built models, trained on synthetic data, can achieve high utility on real-world problems. We provide this pipeline as a proven asset for future capstone projects and recommend ISSIP adopt the 'Sponsor Playbook' concept to accelerate future innovation.

## Acknowledgements

We would like to thank our sponsor, Jim Spohrer, for his invaluable guidance and support throughout this project. We also thank our faculty advisor, Dr. Anirudh Subramanyam, and our capstone coach, Dr. Hannah Nolte, for their feedback.

Portions of this whitepaper were created with the help of AI systems; see Appendix A for details.

## References

IBM. (2023, January 31). Synthetic Data. Ibm.com. https://www.ibm.com/think/topics/synthetic-data

Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics, 13(17), 3509. https://doi.org/10.3390/electronics13173509

International Organization for Standardization. (2018). Privacy enhancing data de-identification terminology and classification of techniques (ISO Standard No. 20889:2018). https://www.iso.org/standard/69373.html

Delft University of Technology. (n.d.). Design specification (criteria). In Delft Design Guide. Retrieved from user-provided PDF: Delft Design Guide - Design Specification (Criteria).pdf.

National Aeronautics and Space Administration. (n.d.). How to write a good requirement [Presentation document]. Retrieved from user-provided PDF: NASA - How to Write a Good Requirement.pdf.

Penn State University Libraries. (n.d.). Finding & obtaining standards. In Engineering Standards & Specifications. Retrieved September 7, 2025, from https://guides.libraries.psu.edu/c.php?g=311177&p=2079775

Penn State University Libraries. (n.d.). What are standards? In Engineering Standards & Specifications. Retrieved September 7, 2025, from https://guides.libraries.psu.edu/c.php?g=311177&p=2080369

OpenAI. (2025). Custom vector illustration of interconnected industries (healthcare, EdTech, agriculture) with synthetic data network [Digital image]. Generated using ChatGPT (DALL·E model).

"CSMITH1539/ISSIP-Synthetic-Data: Repository Containing Progress on Building Synthetic Datasets for a Variety of Different Data Types." *GitHub*, 29 Nov. 2025, github.com/CSmith1539/ISSIP-synthetic-data.

# Appendix A: AI Tools Used

Following our sponsor's recommendation, our team used multiple modern AI tools to assist in the completion of this project, from initial research to final whitepaper drafting. Portions of this whitepaper were drafted, refined, and reviewed with the help of AI systems.

### A.1 Tools Used

Our team primarily used the following AI tools to brainstorm technical explanations, debug code, and refine the narrative of this paper:

- **OpenAI ChatGPT:** Used for initial drafts of technical descriptions and for structuring the paper's narrative.
- **Google Gemini:** Used to analyze and summarize project documents (like the SOW and past whitepapers) and to help refine the VCRM data into a clear story.
- **Microsoft CoPilot:** Used within Visual Studio Code to assist in writing and debugging PyTorch code for the VAE.

Our process involved using these tools to generate a first pass, and then as a team, we heavily rewrote, edited, and combined the best elements to ensure technical accuracy and a clear voice.

### A.2 Tool Problems

All AI systems today suffer from the 3 E Problems:

- **Energy:** AI systems use too much energy.
- **Errors:** AI systems make a lot of errors.
- **Ethics:** AI system vendors made ethically questionable choices, often violating copyright by training on material they did not have rights to train upon.

For these reasons (and other reasons), users should limit their use of today's AI systems and be very careful when using today's AI systems for any purpose.

### A.3 Prompts for This Project

The design of prompts was key to our success. We followed our sponsor's recommendation of asking for clarification to get better results. Here are examples of prompts we used for this project:

**Example 1: Technical Research Prompt**

"Please compare Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) for the specific task of generating synthetic *tabular* data. I need to know the pros and cons of each, focusing on training stability and preserving statistical correlations. Before starting, is the task clear? Any clarification questions for me before generating the response?"

**Example 2: Code Debugging Prompt**

"I am building a VAE in PyTorch. My Beta Prototype failed its fidelity test, and the mean of the 'final grade' column deviated by 8.2%. I think the issue is a lack of hyperparameter tuning. Can you show me how to integrate the 'DeepHyper' library to automate HPO on my VAE model? Is this task clear?"

**Example 3: Whitepaper Writing Prompt**

"You are a technical writer helping a team of engineering students write a final capstone report for ISSIP leadership. Take the following 'Findings' and 'Results' data and rewrite it as a single 'Progress: Findings & Deliverables' section. The goal is to tell a compelling story about our project's success, explaining how our 'Final Prototype' solved the problems of our 'Beta Prototype'. Is this task clear?"