Addressing the Challenges of Healthcare Data Analytics

Enhancing Patient Privacy with Synthetic Data Generation

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Mithila Reddy Chitukula, Pooja Udayanjali Kannuri, Seonkyu Kim, Rahul Kunku, Shubhankar Sharma, Prof. Yang Wang
Purdue University, Mitchell E. Daniels, Jr. School of Business
mchituku@purdue.edu; pkannuri@purdue.edu; kim4377@purdue.edu; rkunku@purdue.edu; sharm842@purdue.edu; yangwang@purdue.edu

BUSINESS PROBLEM

In the healthcare sector, the utilization of AI and machine learning for advanced data analysis is imperative for **innovation**. Yet, it faces the critical challenge of maintaining **patient privacy** amidst stringent regulations and high data protection **costs**. Partnering with California's premier health data network, we propose a **synthetic data** solution that **ensures privacy**, **reduces operational expenses**, and supplies valuable data for research. This approach not only addresses the privacy concerns of patients, healthcare providers, and policymakers but also paves the way for **cost-effective, data-driven healthcare advancements**.

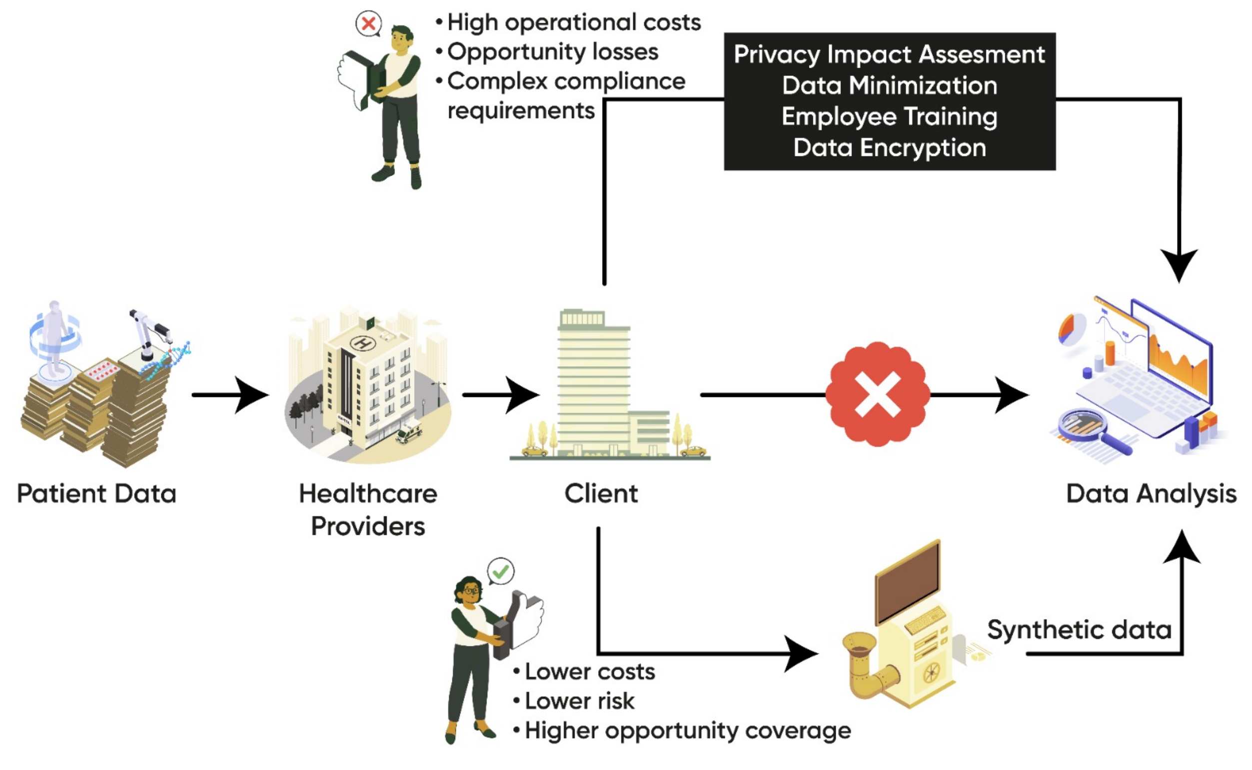

Fig 1. Overview of Business Problem

ANALYTICS PROBLEM

The analytical challenge lies in **evaluating** various synthetic data generation methods to determine the optimal approach for healthcare analytics, adhering to **HL7 FHIR** standards.

Our objective is to find the perfect **balance** between maintaining patient **privacy** and ensuring the **utility** of the data for research purposes. Through collaboration with **industry experts** and comprehensive **literature review**, we have developed a synthetic data methodology that upholds scientific integrity and is recognized by healthcare analytics professionals.

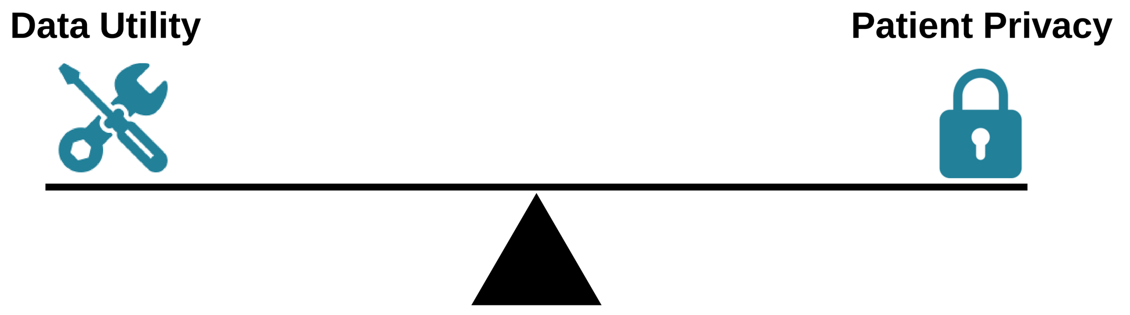

Fig 2. Balance between Privacy and Utility

Assumptions: *We assume the use of a publicly available synthetic HL7 FHIR dataset, mimicking real patient data, to test our methodology under realistic conditions. Data Source: https://synthea.mitre.org/downloads*

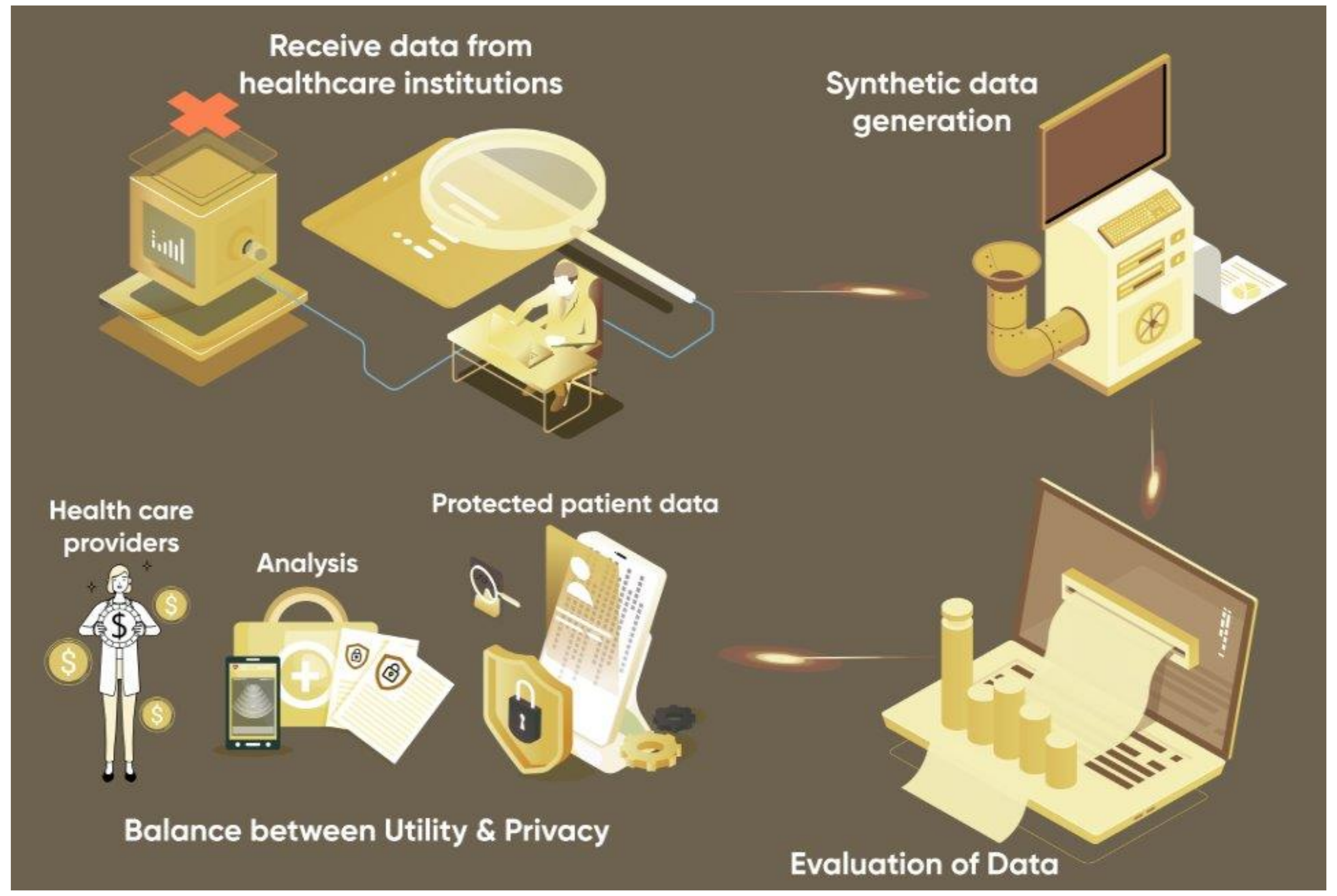

Fig 3. Overview of Analytics Problem

Success Metrics

- **Multivariate Distribution Comparison**: The synthetic data should have multivariate distributions that closely match those of the presumed real dataset.
- **Receiver Operating Characteristic (ROC) Analysis**: Ability of the synthetic data to replicate the predictive performance of models trained on real data, as measured by ROC curves.
- **Differential Privacy Guarantees**: Ensuring that the method provides strong privacy guarantees, ideally quantifiable through differential privacy metrics.

DATA EXPLORATION

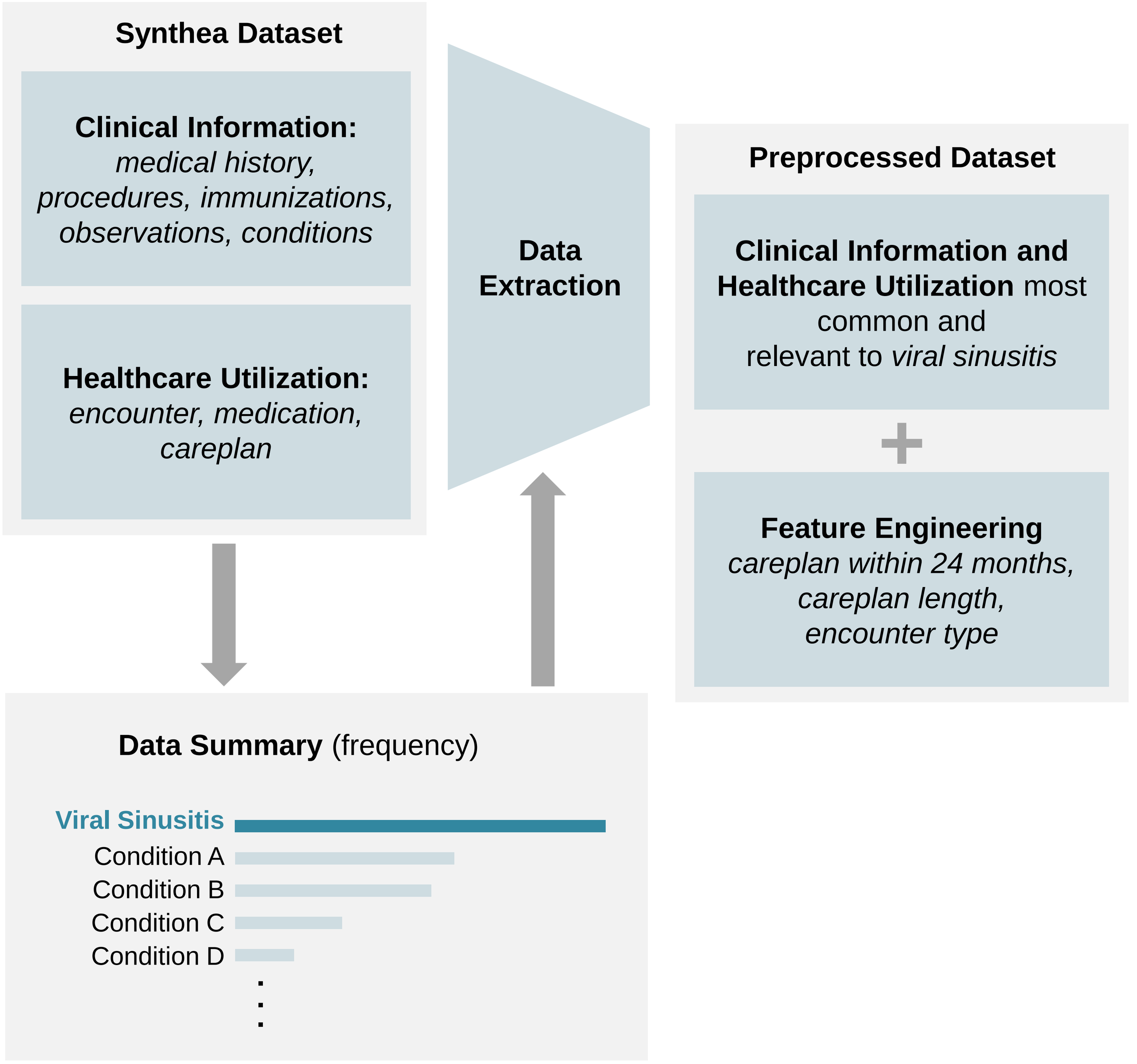

Fig 4. Data Structure and Preprocessing

Utilizing an artificial dataset provided by Synthea, we navigate the patient privacy regulations using this dataset as a **proxy** for real patient data. This dataset, conforming to **HL7 FHIR** standard, encompasses over **60** medical variables such as *conditions, encounters, procedures, and observations.*

Our data preprocessing involved analyzing **correlations** between the target condition and pertinent medical variables, including allergies and immunizations, guided by **insights** from our industry partner. This collaboration has been instrumental in decoding the complex **interplay** of medical data elements.

METHODOLOGY

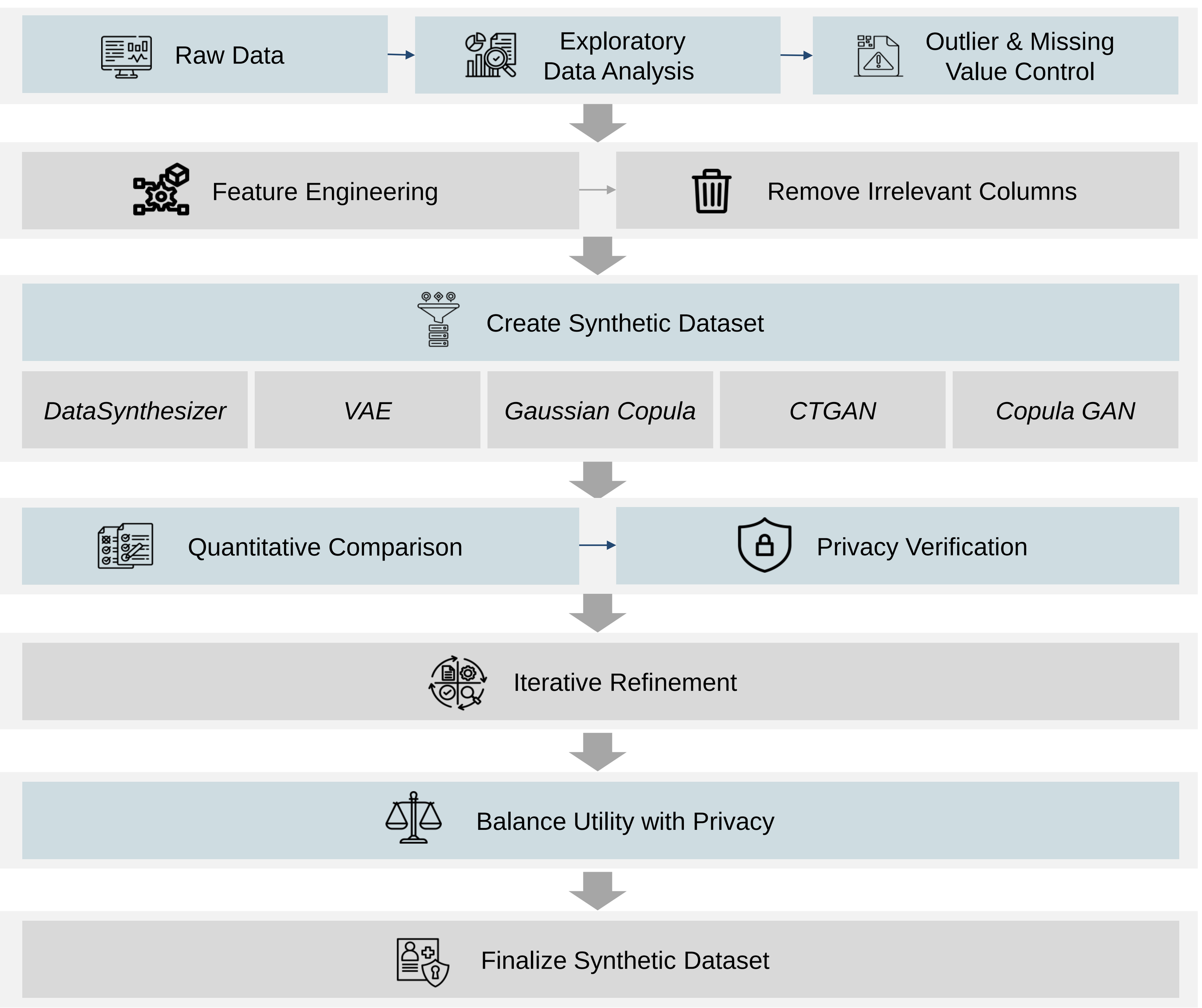

Fig 5. Methodology Workflow

TOOLS

python, pandas, NumPy, scikit-learn, CTGAN, Copulas, SDV

MODEL BUILDING

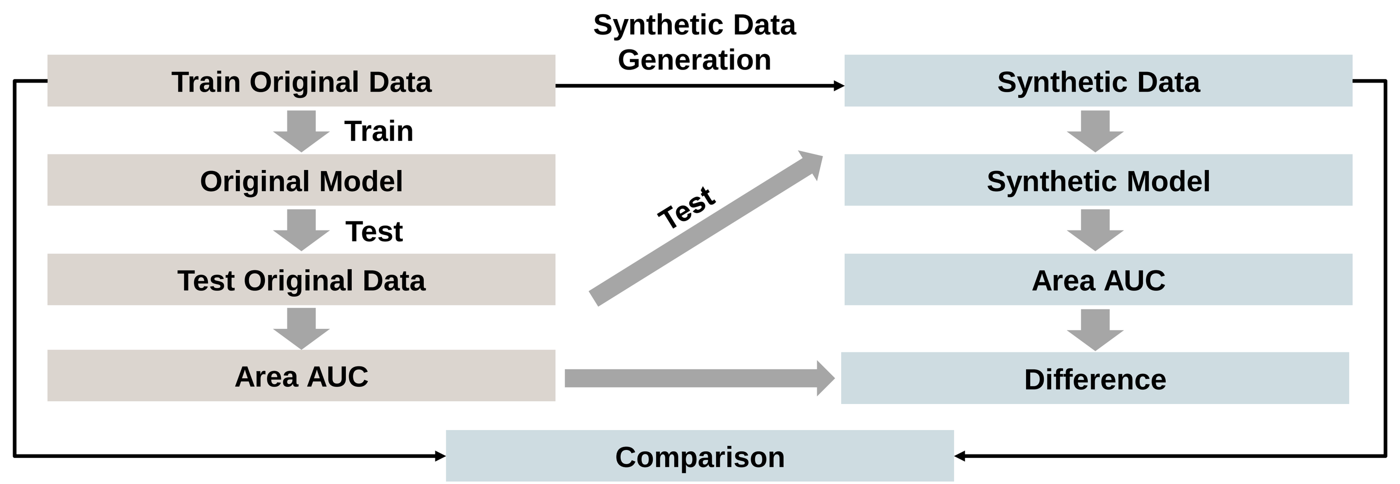

Fig 6. Model Setup

Model Training with Original Data: We start with training a robust predictive model on the original Synthea data, establishing a benchmark for performance comparison with synthetic data models.
Synthetic Data Generation: Using synthetic data generators, we create a dataset that statistically mirrors the original, prioritizing privacy preservation.
Model Training with Synthetic Data: A 'Synthetic Model' is trained on this synthetic dataset to assess its fidelity and utility compared to the original data model.
Performance Evaluation: Both models are tested against a separate original dataset, with the Area Under the Receiver Operating Characteristic (AUC) metric used to evaluate classification performance.
Areas for Improvement: Enhance Synthetic Data Quality, Optimize Privacy-Preserving Techniques, Increase Model Diversity, Expand Data Features.

RESULTS

	Original Data	CTGAN	Gaussian Copula	Copula GAN
ROC Score Without Noise	0.73	**0.53**	0.64	0.51
Optimal Privacy Level (∑)	-	**0.7**	0.9	0.8
Closely Follows Distributions?	-	**Yes**	Yes	Yes
Correlation Matrix	-	**Best**	Good	Average

Fig 7. Comparison of Synthetic Data Generation Methods

According to our evaluation criterions and experiments, **CTGAN** emerged as the best synthetic data generation method. It was able to capture the **distribution** of the real data well. It also had the **best prediction power** of 0.528 at a reasonable privacy level of **0.7**. It was also able to capture **correlations** between variables, crucial to retaining the predictive power of the real dataset.

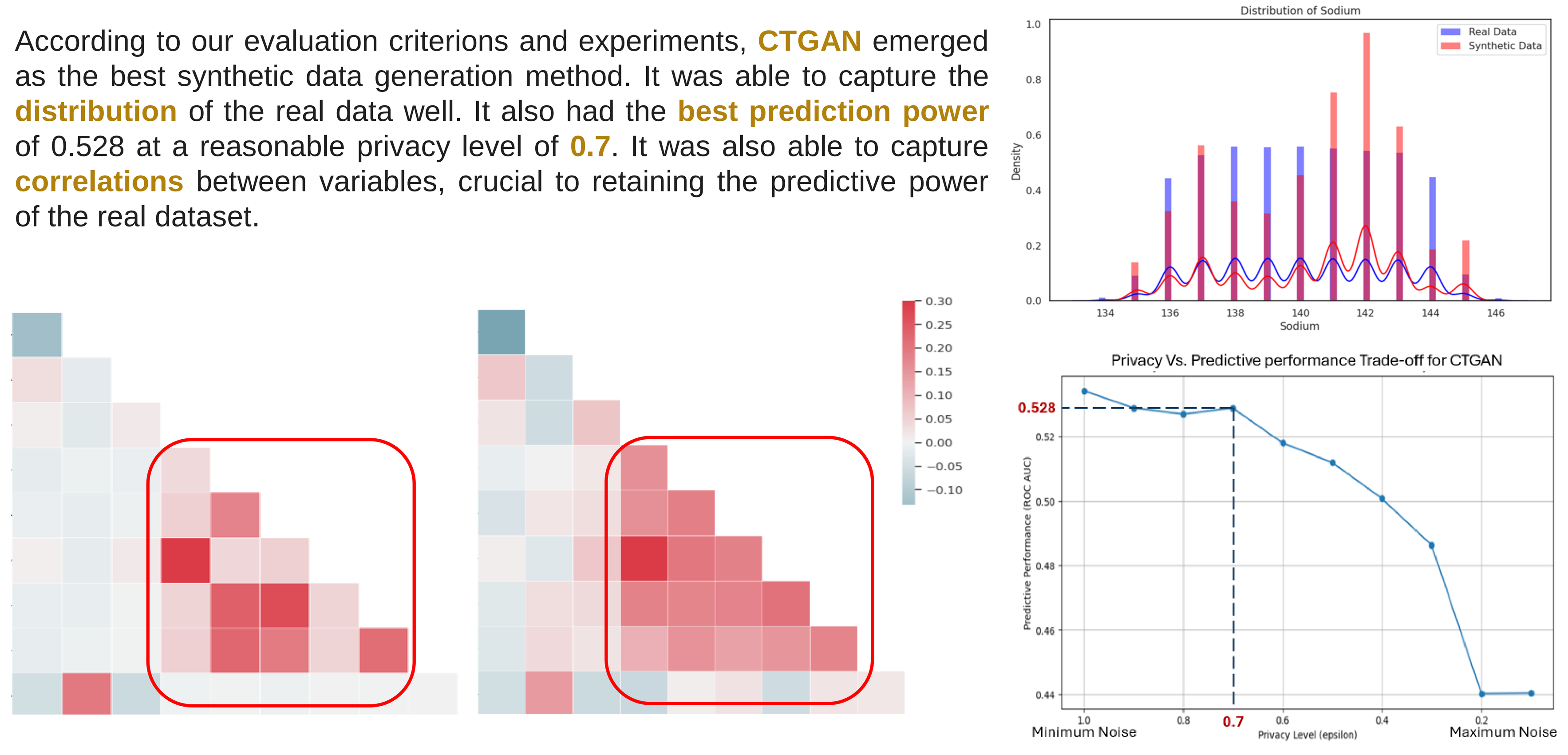

Fig 8. Results of CTGAN

DEPLOYMENT AND LIFECYCLE MANAGEMENT

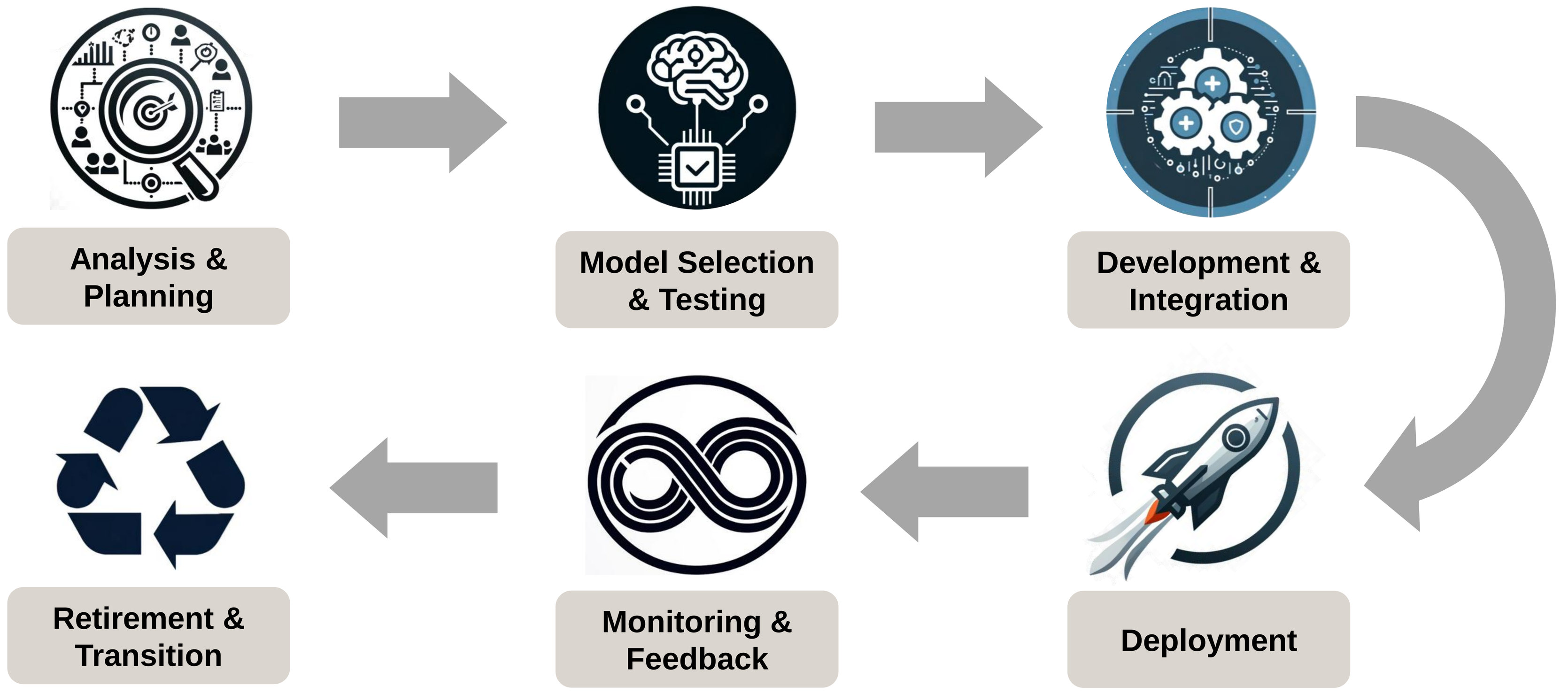

Fig 9. Deployment and Lifecycle Management

ACKNOWLEDGEMENTS

We would like to thank Daniels School of Business, Professor Yang Wang and our industry partner for this opportunity, their guidance, and support throughout this project.

informs ANALYTICS CONFERENCE
ORLANDO, FL | APRIL 14-16, 2024