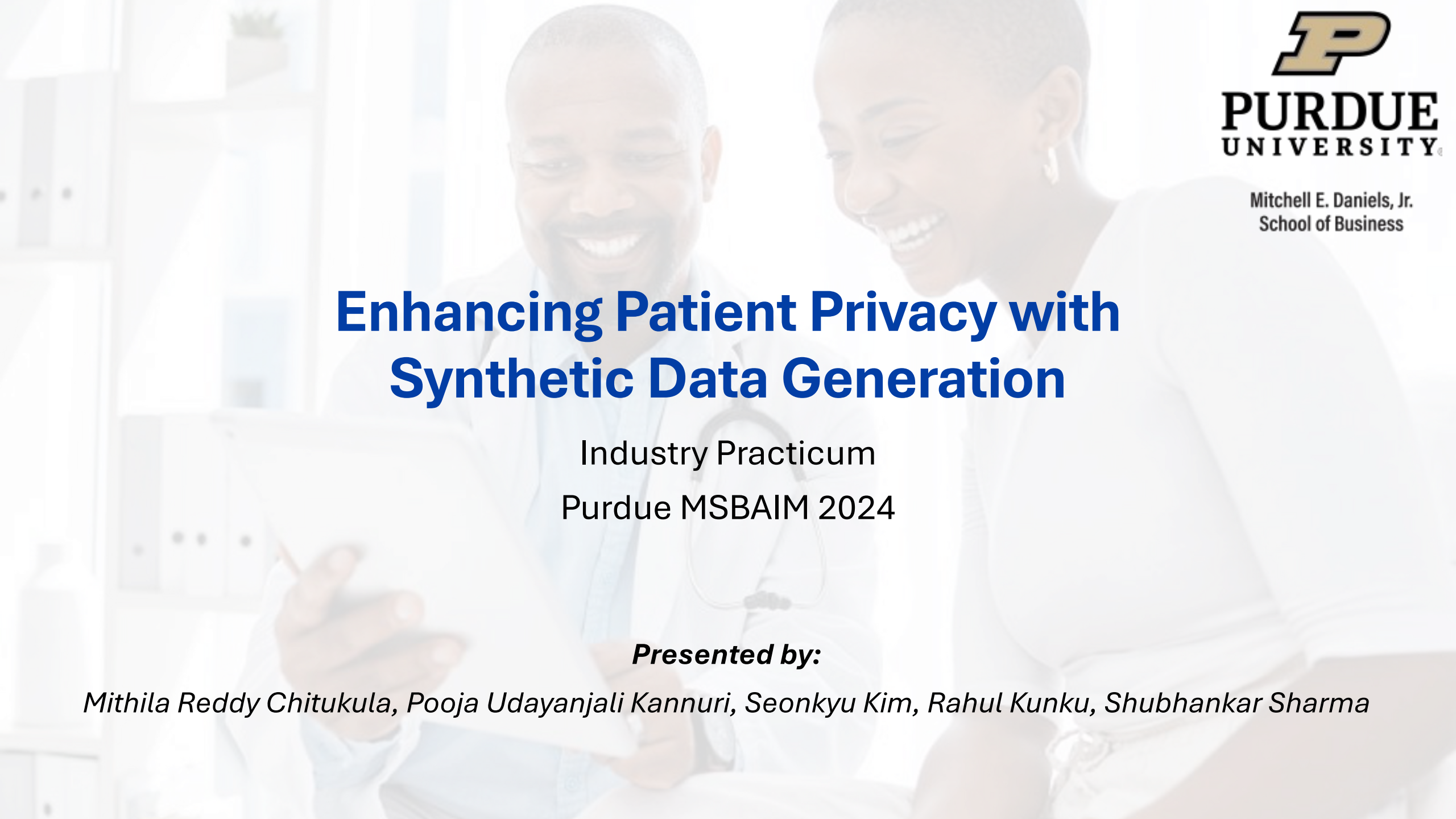


PURDUE UNIVERSITY

Mitchell E. Daniels, Jr. School of Business

Enhancing Patient Privacy with Synthetic Data Generation

Industry Practicum

Purdue MSBAIM 2024

Presented by:

Mithila Reddy Chitukula, Pooja Udayanjali Kannuri, Seonkyu Kim, Rahul Kunku, Shubhankar Sharma

The Team

Mithila Reddy Chitukula

Pooja Udayanjali Kannuri

Seonkyu Kim

Rahul Kunku

Shubhankar Sharma

Student Team
Purdue MSBAIM'24

Prof. Yang Wang

Mentor

Current scenario of healthcare industry

Practical implementation of health data privacy has become increasingly essential with the **de-siloing of digital health information**

Healthcare Insurance Portability and Accountability Act (HIPAA) marked the beginning of **modern privacy regulation**

Health data has tremendous research utility, but achieving high utility often requires making **sacrifices on patient privacy**

Regulation moves the industry along this **tradeoff curve** towards privacy, reducing data utility

Business Problem

Data Privacy Innovation: Navigating sharing restrictions with privacy-centric technology.

Growth & Expansion: Fueling growth and new services through technological leverage.

Responsible Healthcare Innovation: Ensuring security and responsibility in healthcare advancements.

Challenge

Safeguarding patient privacy against stringent regulations and high costs.

Solution

Implementing a synthetic data approach with California's health data network for privacy and cost-efficiency.

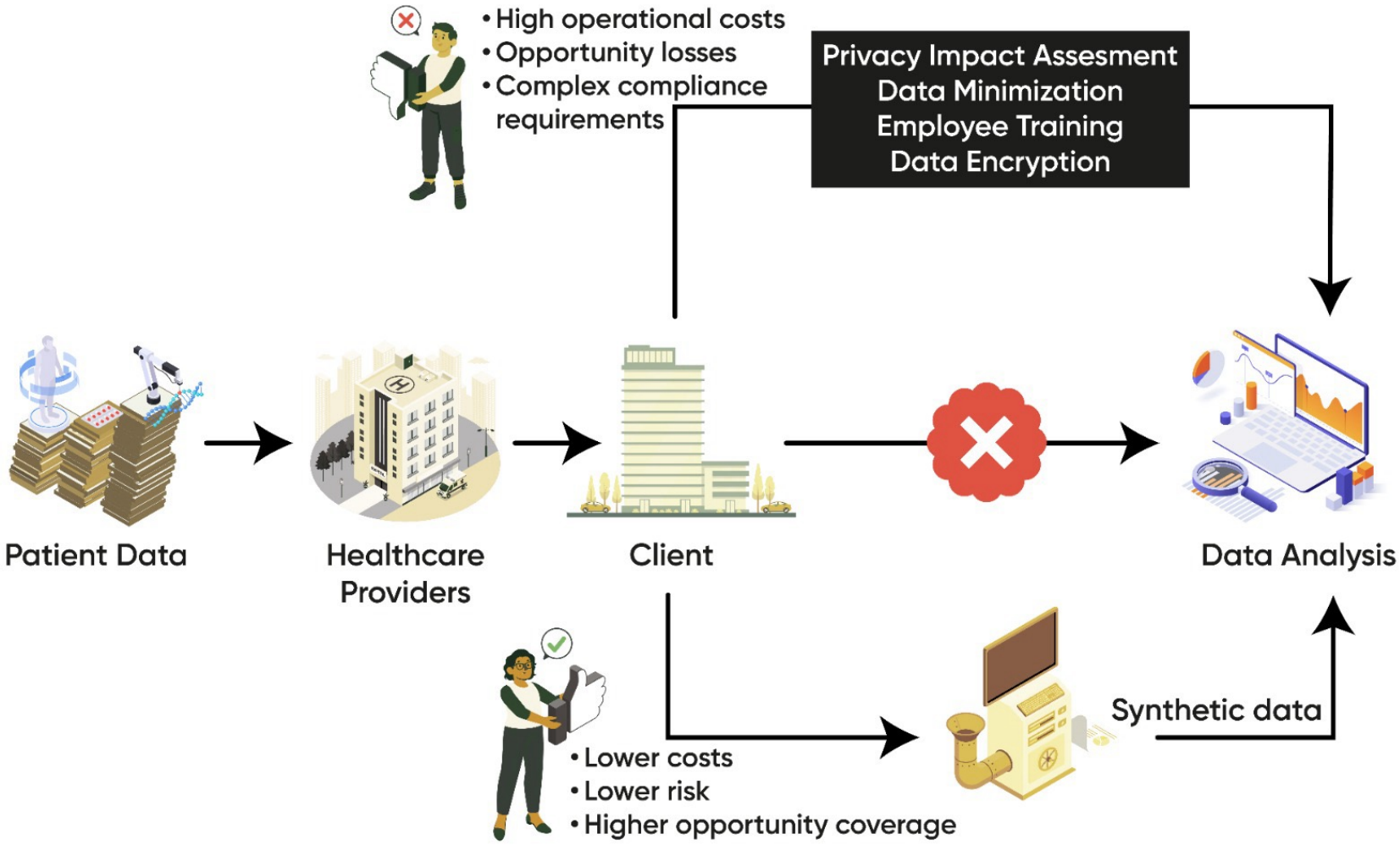
Benefits

Lowering costs while offering valuable data for innovation without risking privacy.

Impact

Leading the charge towards a new standard in privacy-focused, data-driven healthcare.

Patient Data Flow



Synthetic Data in Healthcare industry

Potential

Synthetic data has the potential to be a critical technology in this space, as it enables representative patient data with inherent privacy protection

Benefits

Enables organizations to access concrete and representative insights from sensitive data, while minimizing the risk to patient privacy and limiting governance requirements

Accessibility

Organizations can access and share synthetic data in circumstances where sharing identified, or de-identified data is too difficult or even impossible.

Synthetic data can hence be used to **shift the privacy-utility curve outward** and enable **broad access to high-utility data without making tradeoffs on patient privacy**.

Synthetic Data Use Case 1

US Based Pharma Company
Commercial Treatment for Cancer

General Data Protection Regulation
(GDPR)
Sharing patient-level data with the US company requires enhanced safeguards, contractual terms, and governance

EU Market Patient Data
For Cancer Treatment and Outcomes

Generate Synthetic Data
Creating a synthetic version of EU patient data allows the US company to access it with

- **Less overhead, saving time and money**
- **Ensures data protection regulations compliance**

Synthetic Data Use Case 2

Lab Testing Company
(A)

1 Shares complete Patient Records

Violates data protection and privacy laws

2 Shares de-identified Patient Records

Patient Records − Personal and Identifiable Information = De-identified Patient Records

- ✓ Protects privacy
- × Value add?

3 Shares "presale" synthetic records

De-identified Patient Records → Synthetic Data Generator → De-identified Synthetic Records

- ✓ Protects privacy
- ✓ Head start on tools
- ✓ Evaluate value-add

Analytics Company
(B)

Analytical Problem

Find the best synthetic data generator model

Evaluate accuracy, compliance with policy, and likelihood of reverse engineering

Identify balance point between accuracy / richness vs. privacy protection

Theoretical support:
Synthetic Data Generation

Generative Adversarial Networks (GANs)

- Generator and discriminator networks compete to produce realistic synthetic data
- Enables capturing complex data distributions and patterns

Variational Autoencoders (VAEs)

- Probabilistic approach based on Bayesian inference
- Learn latent representations
- Allows control over the generation process through latent space manipulation

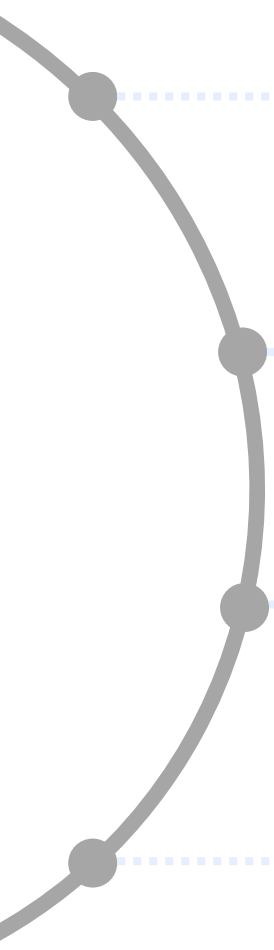
Copula based Methods

- Generator and discriminator networks compete to produce realistic synthetic data
- Enables capturing complex data distributions and patterns

Privacy-Preserving Synthetic Data Generation

- Theoretical foundations in differential privacy
- Techniques like PATE-GAN, DP-WGAN, and secure data synthesis
- Privacy guarantees + Data utility

Theoretical support:
Synthetic Data Evaluation



- Fairness Assessment
- Data Quality and Privacy
- Synthetic Data Techniques (Synthea)
- De-identification Efficacy and Realism

Theoretical support:
Data Privacy

To ensure synthetic healthcare data generation **maintains privacy and utility** while preventing discrimination and protecting sensitive attributes.

To evaluate de-identification methods for healthcare data to **protect privacy** while **preserving data utility** for analysis

To examine **privacy retention** in realistic synthetic data, ensuring ethical usage by **preventing sensitive information disclosure** and **re-identification**.

HL7 FHIR:

Revolutionizing Healthcare Data Exchange

FHIR (Fast Healthcare Interoperability Resources) is a standard framework created by HL7.

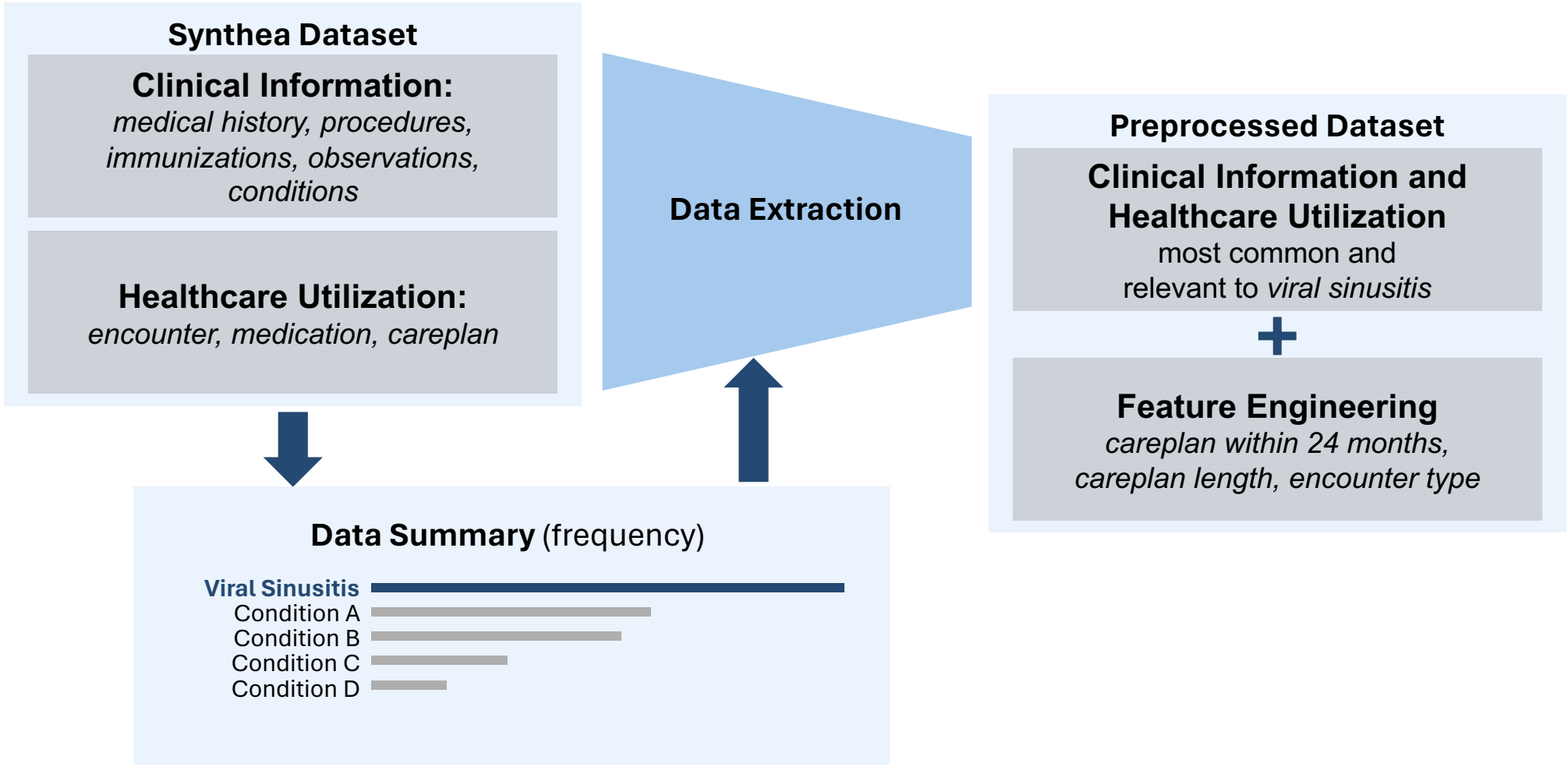
Simplifies data sharing across different healthcare systems.

Designed for the digital era with a focus on ease of implementation and web-based data exchange.

Supports RESTful architectures, making it well-suited for mobile and cloud applications.

Allows for seamless integration with existing healthcare protocols and standards.

Summary of Data Preparation

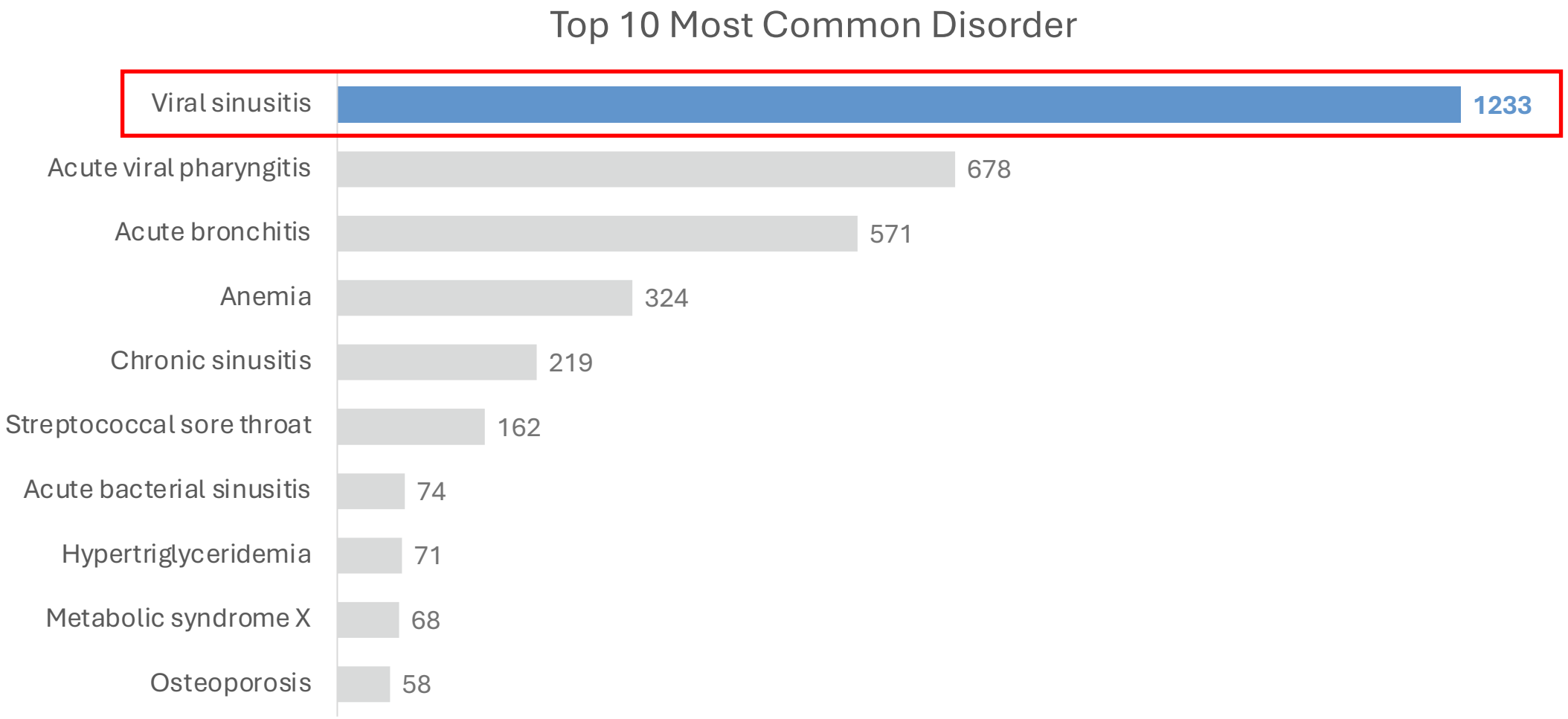


Synthea dataset was assumed to be real data

- 1K Sample Synthetic Patient Records, FHIR R4 (2019-2021)
- 9 CSV files

allergies	careplans	conditions
encounters	immunizations	medications
observations	patients	procedures

Viral Sinusitis was the most common disorder



Attribute Selection & Feature Engineering

Extract Latest Value from Each Patient

Select Attributes Relevant to Viral Sinusitis

Patient Information

- Exclude dead people
- Exclude personal information irrelevant to viral sinusitis (*Passport, Firstname, Lastname*, etc.)

Healthcare Information

(*allergies, immunizations, medications, observations, procedures*)

- Extract the most common and relevant to viral sinusitis

Feature Engineering

careplan_within_24	*careplan_length*	*encounter_type*	*Viral_sinusitis_present*
If a patient is in the care plan for past 24 months	The length of a care plan	inpatient, outpatient, or emergency visit	If a patient has viral sinusitis

1000 Rows × 66 Columns of Selected Data

Data was not enough for accurate models

Too Many Missing Values

Not Enough Data for Accurate Models
- Synthetic Data Generation Model
- Prediction Model

Need

Larger Dataset

New Dataset & Preprocessing

Data Source

- **100K** Sample Synthetic Patient Records, FHIR (2017)
- 9 CSV files

EDA

- Most Common Condition: **Viral Sinusitis** (77,926)

Feature Engineering

- **99,539 Rows × 73 Columns** of Selected Data

Preprocessing

- **Impute Values** for Continuous Variables
- **Remove Attributes** with Feature Importance
- **One Hot Encoded** Categorical Variables
- **Remove Missing Values**

75,739 Rows × 66 Columns of Finalized Data

Methodology

Tools

python

pandas

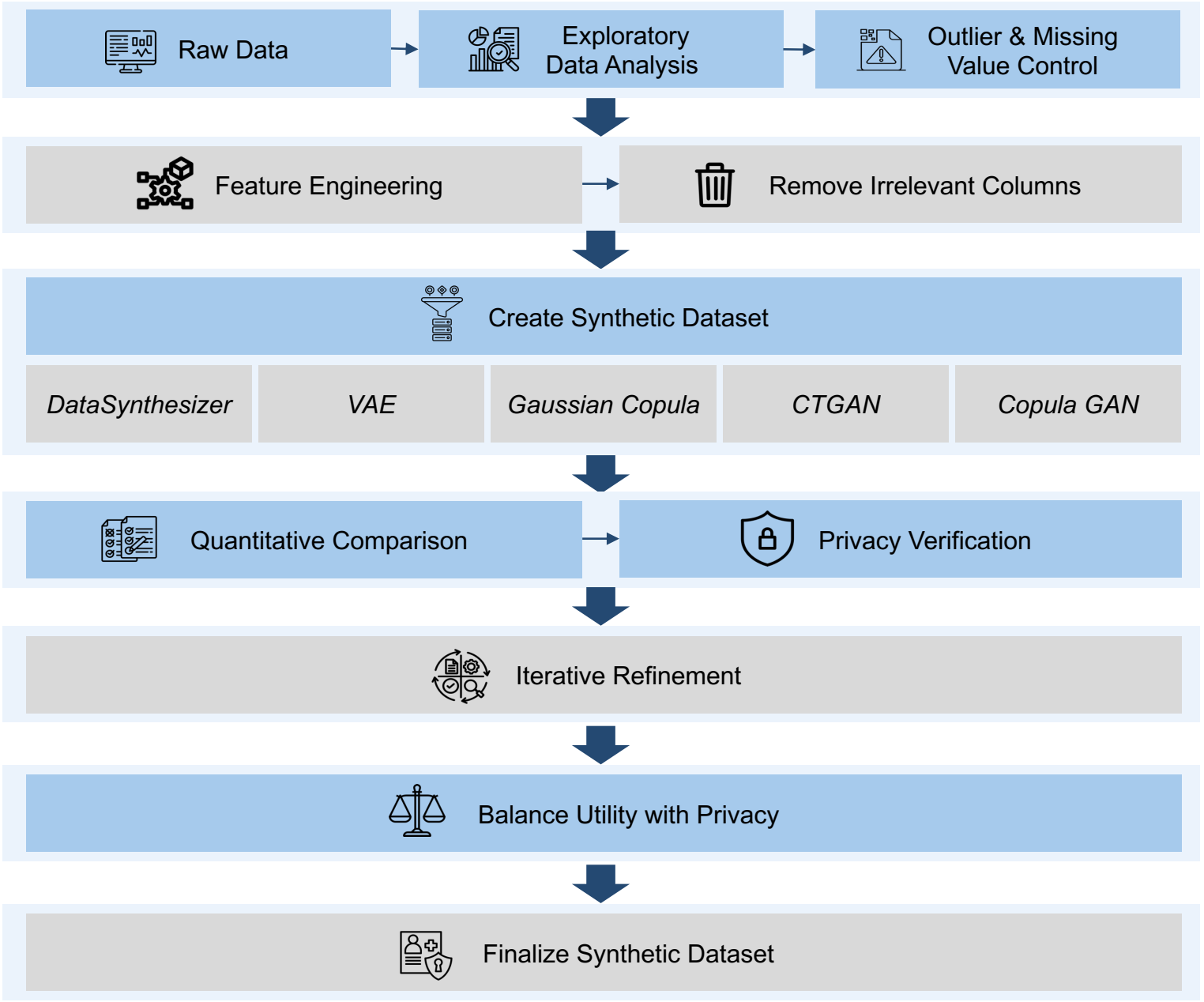
NumPy

scikit learn

datacebo CTGAN

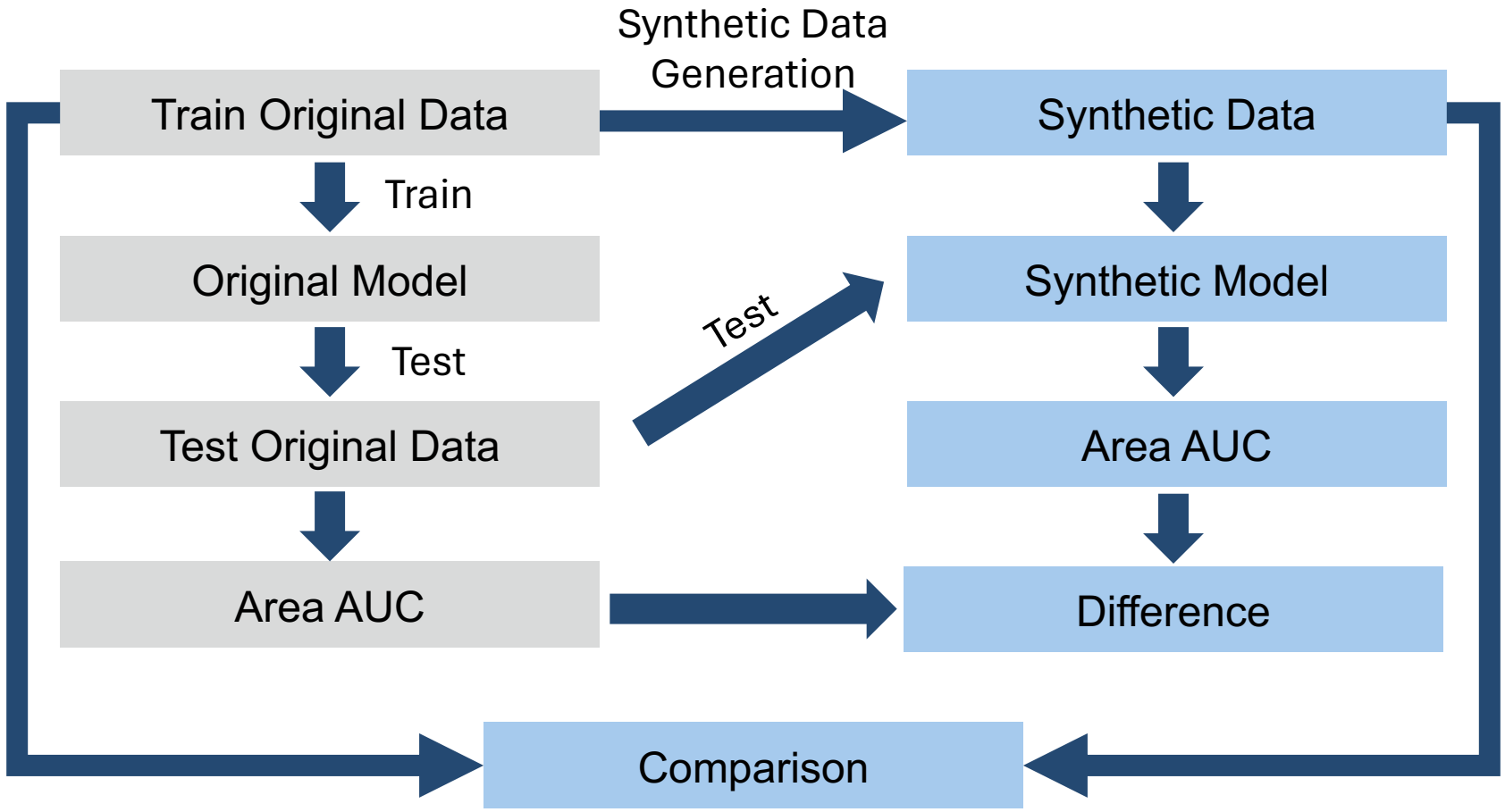
datacebo Copulas

SDV
The Synthetic Data Vault



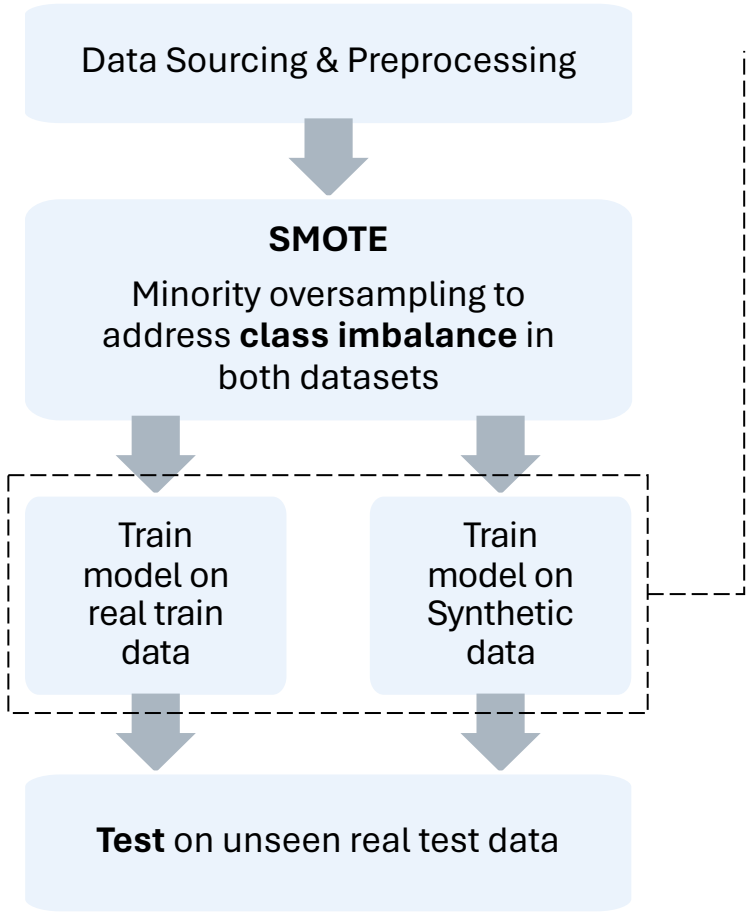
Model Building Framework

Train : Test = 80 : 20



Predictive Model

Objective To evaluate the effectiveness of synthetic data in training ML models compared to real data



Logistic Regression

'max_iter = 10000' to ensure convergence

Using the original data post SMOTE:

ROC = 0.73

Accuracy = 0.80

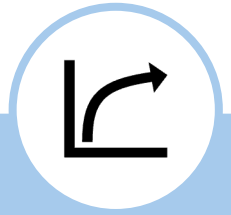
Recall for class 0 = 0.97

Recall for class 1 = 0.25

Potential improvements:

- Advanced oversampling techniques such as ADASYN
- More complex models to capture the complexity better
- Hyperparameter tuning
- Ensemble methods
- Feature Selection

Model Evaluation Metrics



ROC AUC Curve

- Compares model performance between Synthetic data and Real data
- Higher similarity in ROC between synthetic and real data indicates better synthetic data representation

Epsilon

- Epsilon measures the trade-off between privacy and utility
- Smaller epsilon increases privacy but may degrade the utility or accuracy of the synthetic data.

Correlation Matrix

- Provides insight into the multivariate structure of the data
- Comparing correlation matrices of original and synthetic data helps assess the preservation of dependencies

Noise Addition for Differential Privacy

- Noise addition ensures that the dataset cannot be used to identify individuals.
- The epsilon parameter controls the trade-off between privacy and data accuracy.
- Epsilon is inversely proportional to the noise added.

Lower ϵ → Higher noise added → Higher privacy

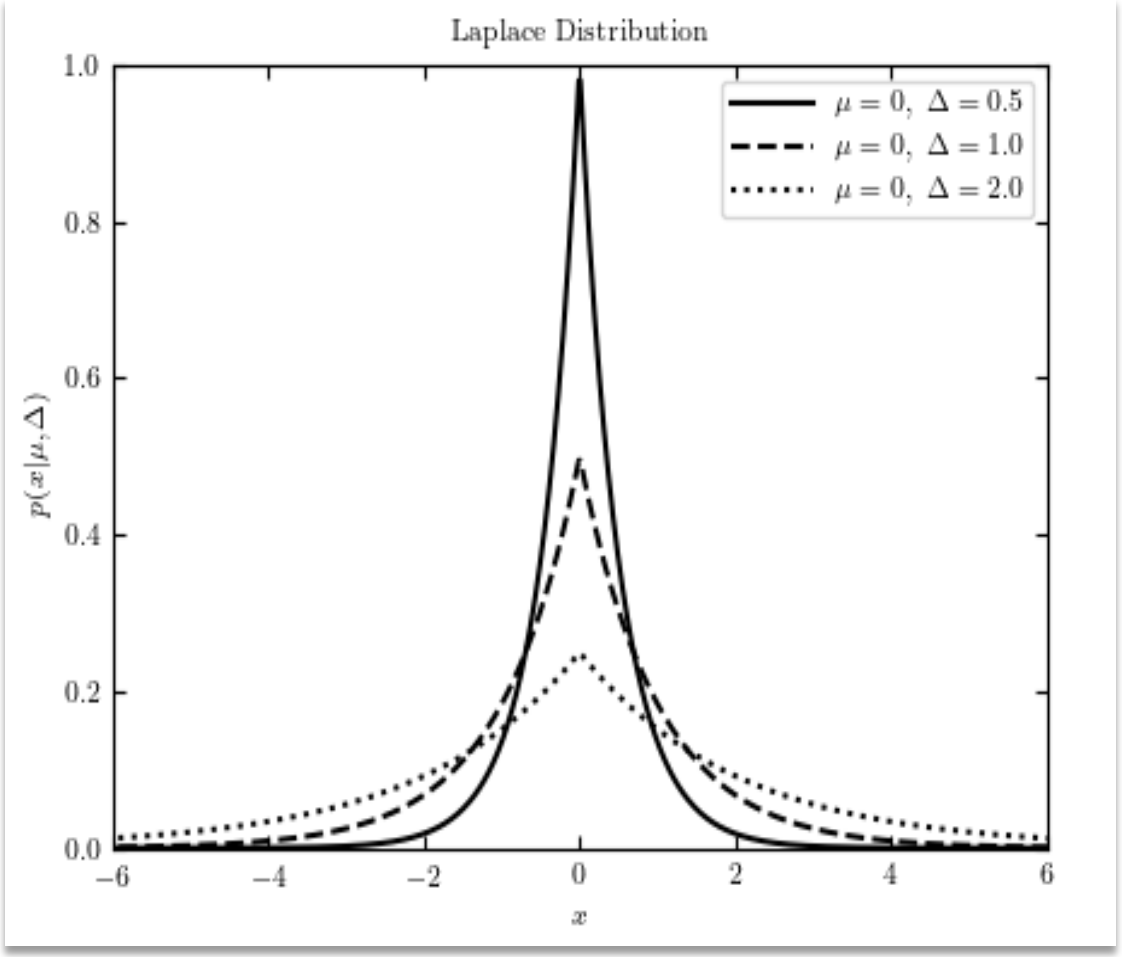
Techniques used

- Laplace Noise for Continuous Columns
- Salt-and-Pepper Noise for Binary Columns
- Adaptive Noise Addition Using Feature Importance

Noise Addition for Differential Privacy

Laplace Noise for Continuous Columns

- Symmetric ‘double exponential’ distribution centered around the mean.
- Ideal for numerical data; ensures strong differential privacy.
- Applied to continuous columns, allowing us to maintain statistical properties like mean and variance.



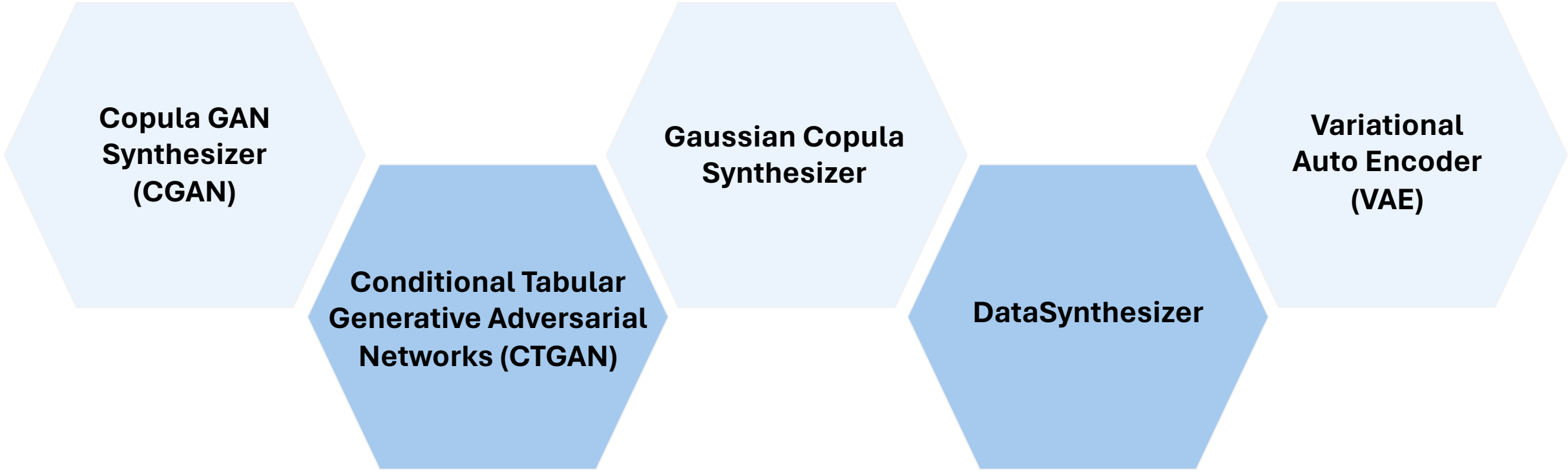
Salt-and-Pepper Noise for Binary Columns

- Randomly flips binary values to introduce variability, simulating natural data discrepancies.
- Provides a simple yet effective method for anonymizing binary/categorical data.

Adaptive Noise Addition Using Feature Importance

- Noise is adaptively scaled according to the importance of each feature, ensuring crucial data characteristics are less perturbed to preserve predictive power.
- Allows us to fine-tune privacy levels for different attributes based on their relevance to the analysis.

Synthetic Data Generation Models



CGAN *CopulaGAN Synthesizer*

Properties

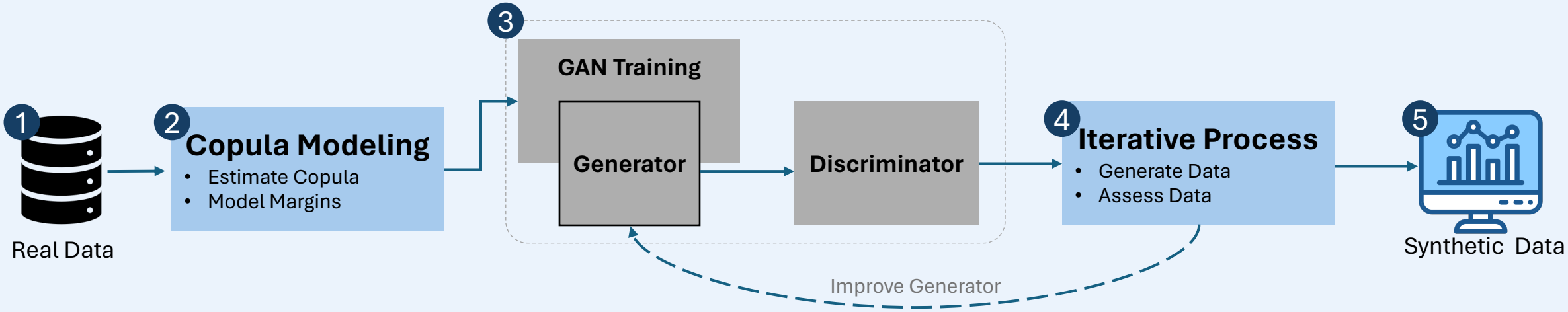
- **Hybrid Model:** Merges copulas with Generative Adversarial Networks (GANs) for detailed data patterns.
- **Realistic Synthesis:** Achieves high-quality, believable synthetic data promoting realism in synthetic healthcare data.
- **Preserves Distributions:** Maintains true variable distributions accurately.
- **Learns Dependencies:** Captures complex variable relationships effectively.

Applications

- ✓ **Risk Simulation:** Crafts synthetic datasets for nuanced financial risk assessment.
- ✓ **Medical Data Privacy:** Generates realistic, privacy-compliant healthcare datasets.
- ✓ **Logistics Optimization:** Provides synthetic data for supply chain efficiency studies.
- ✓ **Environmental Modeling:** Synthesizes datasets with complex ecological dependencies.

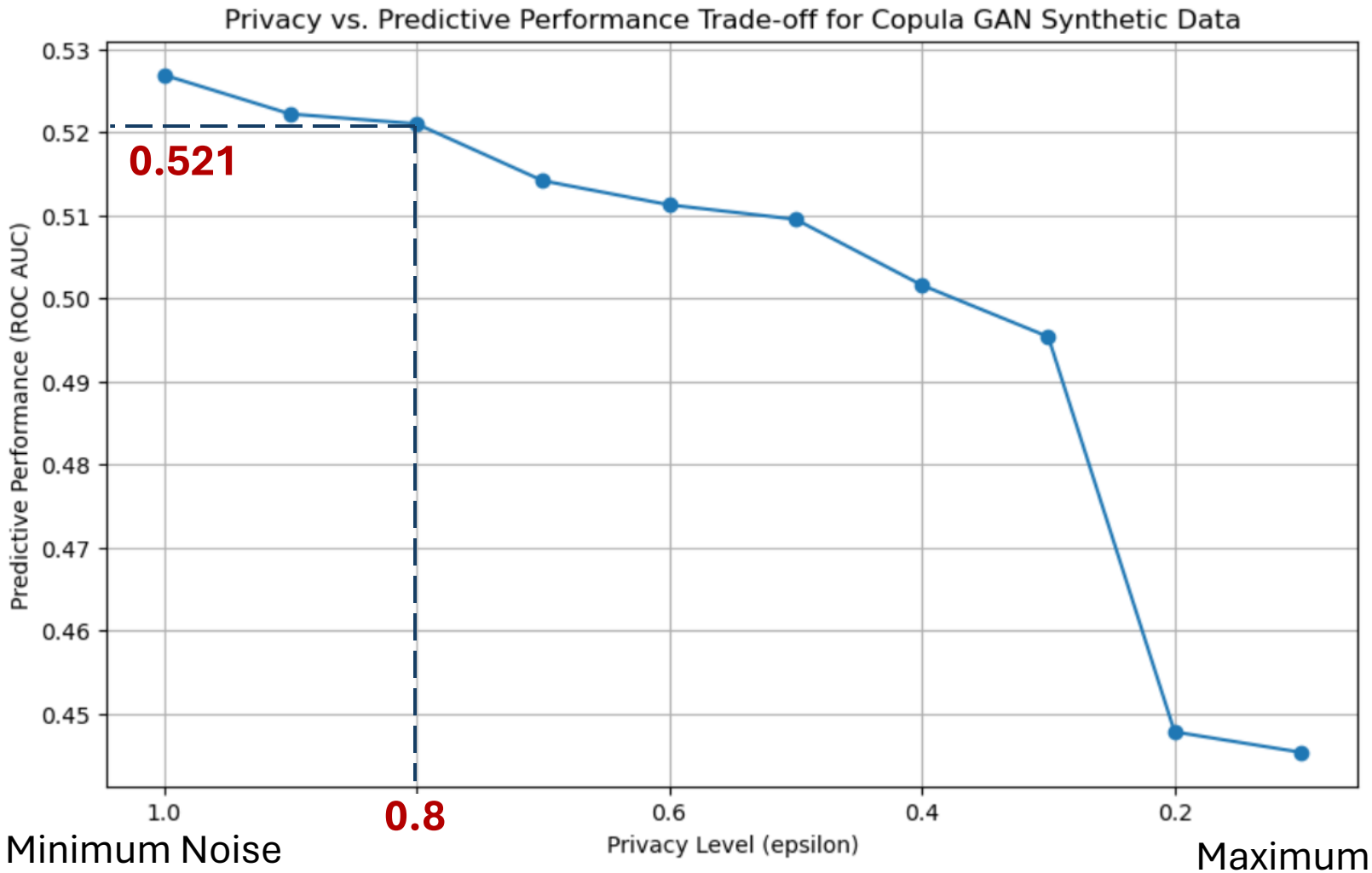
CGAN Architecture

Simplified Diagram of CGAN architecture



1. **Real Data:** *Authentic dataset for reference.*
2. **Copula Modeling:** *Analyzes dependencies within real data to guide synthetic generation.*
3. **GAN Training:**
    - **Generator:** *Crafts synthetic data based on real data patterns and input conditions.*
    - **Discriminator:** *Judges if samples are real or synthetic, considering their conditions.*
4. **Iterative Process:** *Refines synthetic data generation through continuous feedback.*
5. **Synthetic Data:** *The final, realistic data produced, matching specified conditions.*

Predictive Power starts to drop drastically at epsilon 0.8 for Copula GAN



CTGAN *Conditional Tabular Generative Adversarial Network*

Properties

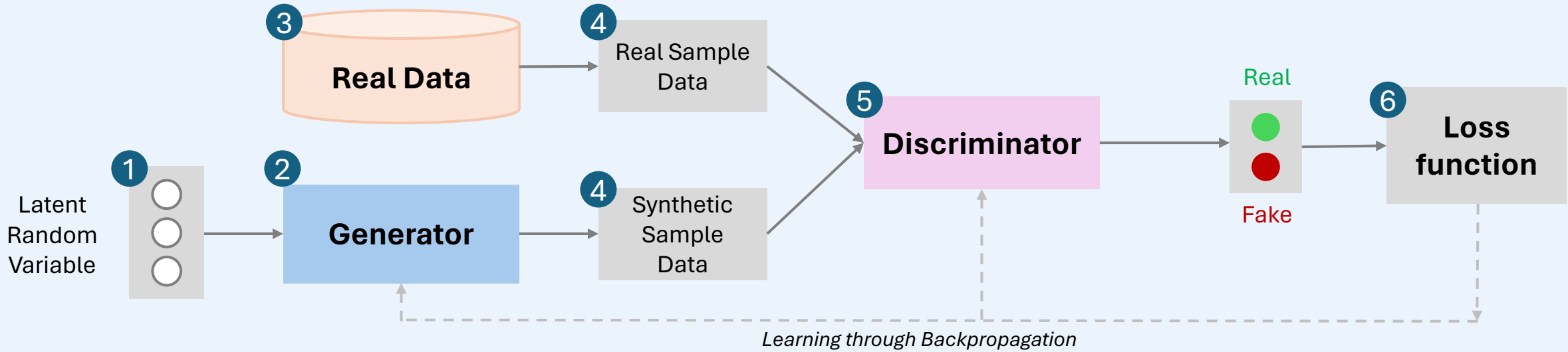
- **Deep Learning Foundation:** Optimized for synthetic data creation.
- **Specialized in Tabular Data:** Handles both continuous and categorical variables.
- **Customizable Outputs:** Allows tailored synthetic datasets.
- **Accurate Statistical Representation:** Ensures data integrity.

Applications

- ✓ **Data Privacy**: Generates anonymized datasets.
- ✓ **Model Training**: Enhances ML with diverse data.
- ✓ **Healthcare & Finance**: Offers risk-free simulation.
- ✓ **Market Analysis**: Provides synthetic consumer insights.

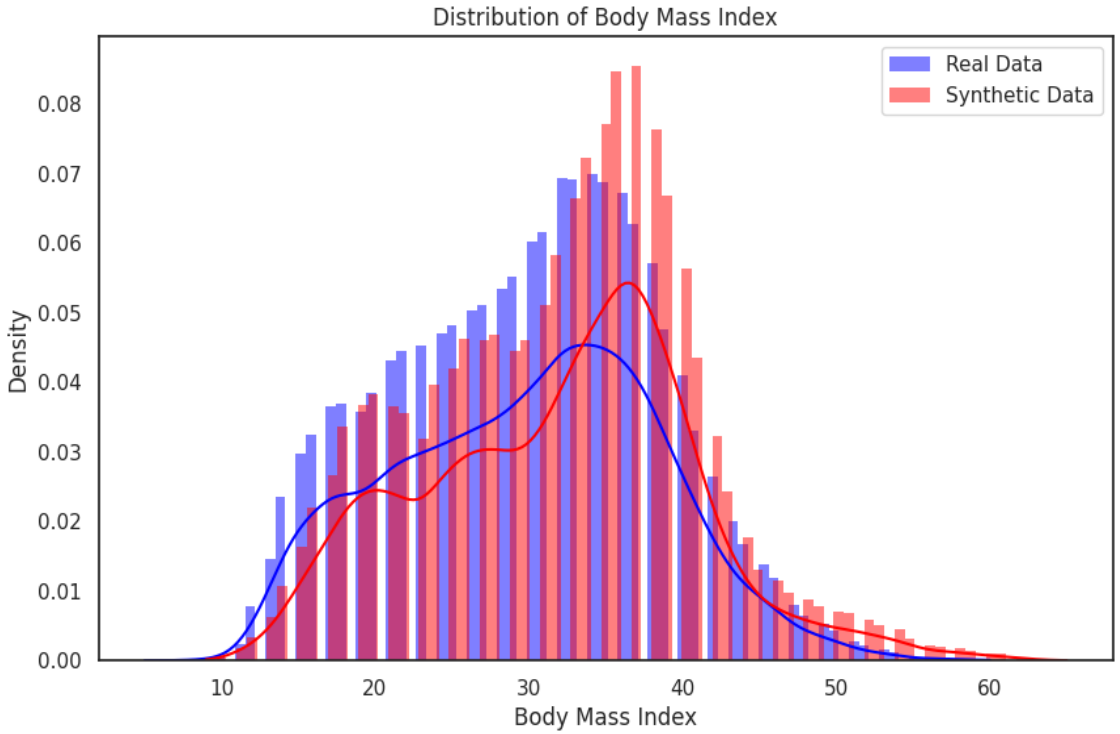
CTGAN Architecture

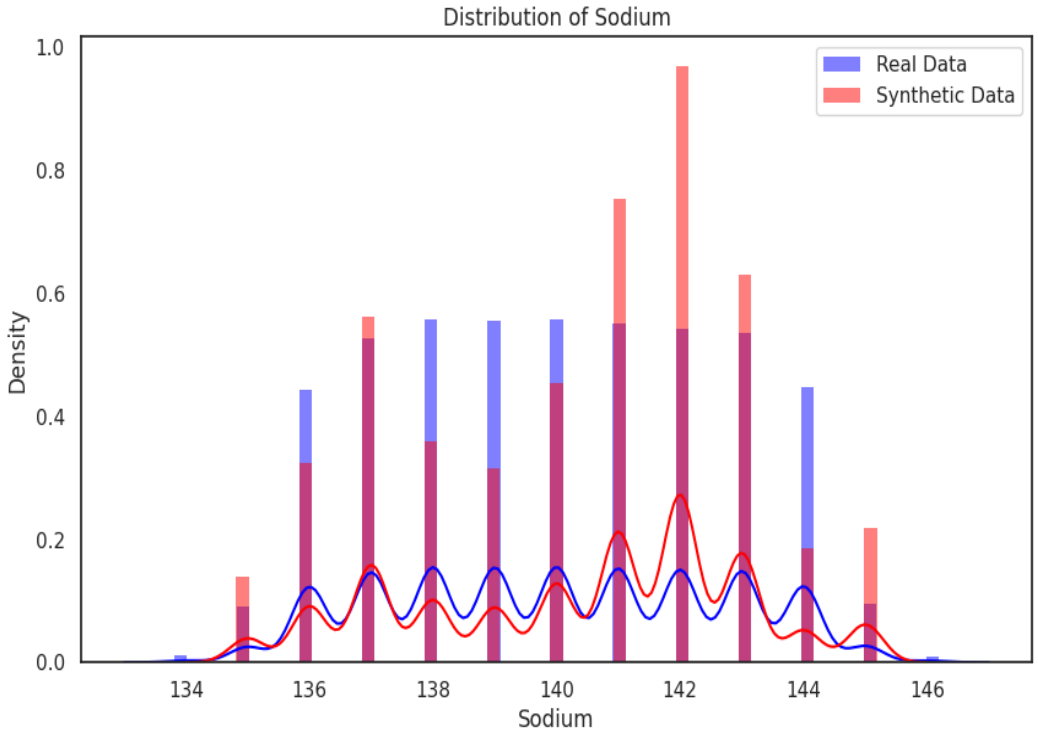
Simplified Diagram of CTGAN architecture



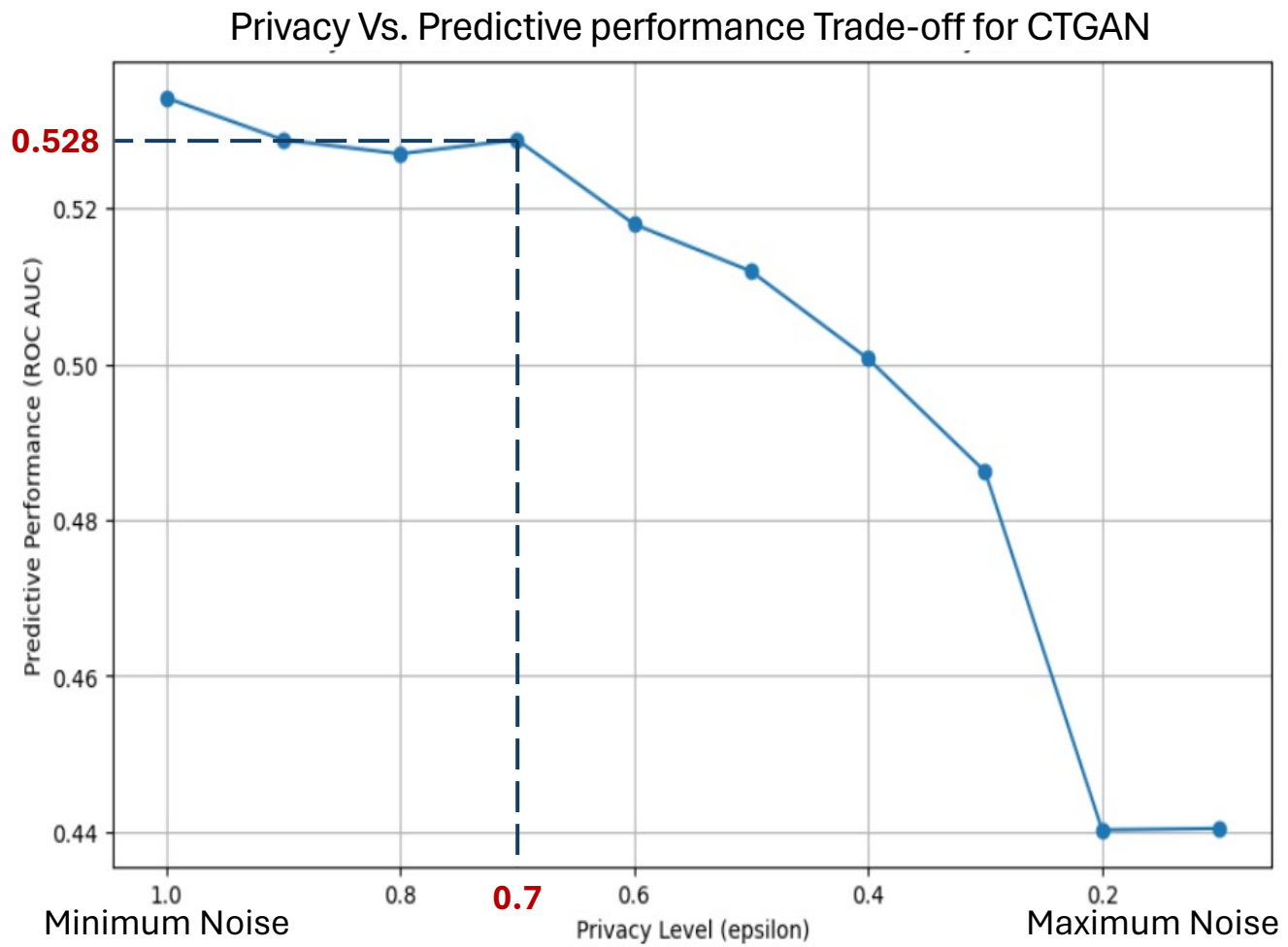
1. **Latent Random Variable**: *Serves as the input to the Generator, representing potential data patterns or features.*
2. **Generator**: *Produces synthetic data samples aiming to mimic the real data's statistical properties.*
3. **Real Data**: *The authentic dataset used for training, providing the Discriminator with examples of genuine data.*
4. **Sample Data**: *Both the Generator and Real Data provide sample data for the Discriminator to assess.*
5. **Discriminator**: *Evaluate samples to determine if they are real (from the actual dataset) or fake (generated by the Generator).*
6. **Loss**: *The feedback mechanism that guides the training of both the Generator and Discriminator, optimizing their performance over time.*

Synthetic Data closely follows real data's distribution in CTGAN model

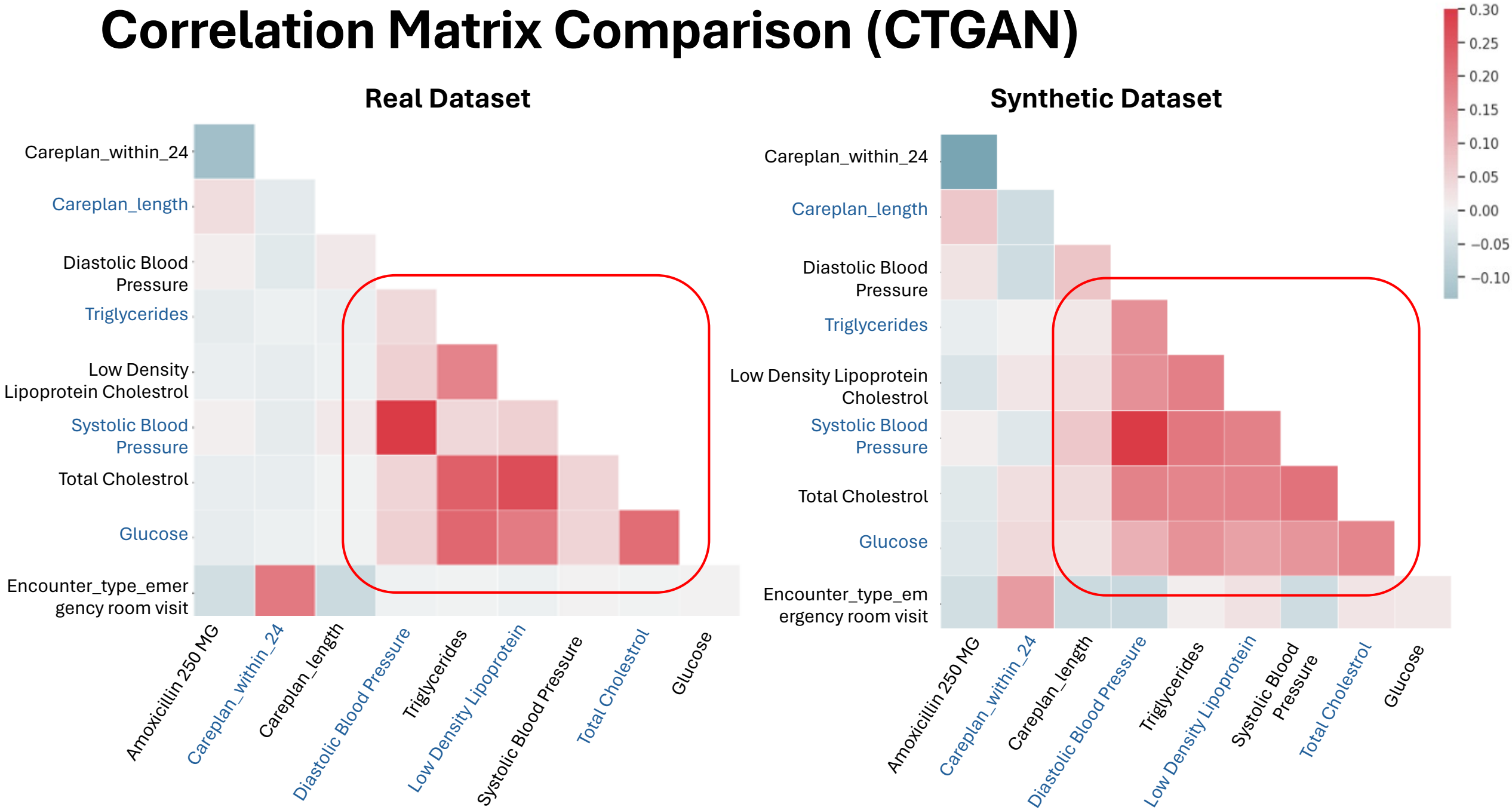




Predictive Power starts to drop drastically at epsilon 0.7 for CTGAN



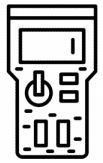
Correlation Matrix Comparison (CTGAN)



Gaussian Copula Synthesizer

Classical Machine Learning Algorithm

Leverages classical machine learning algorithms and statistical models to estimate the distribution and dependencies of the data



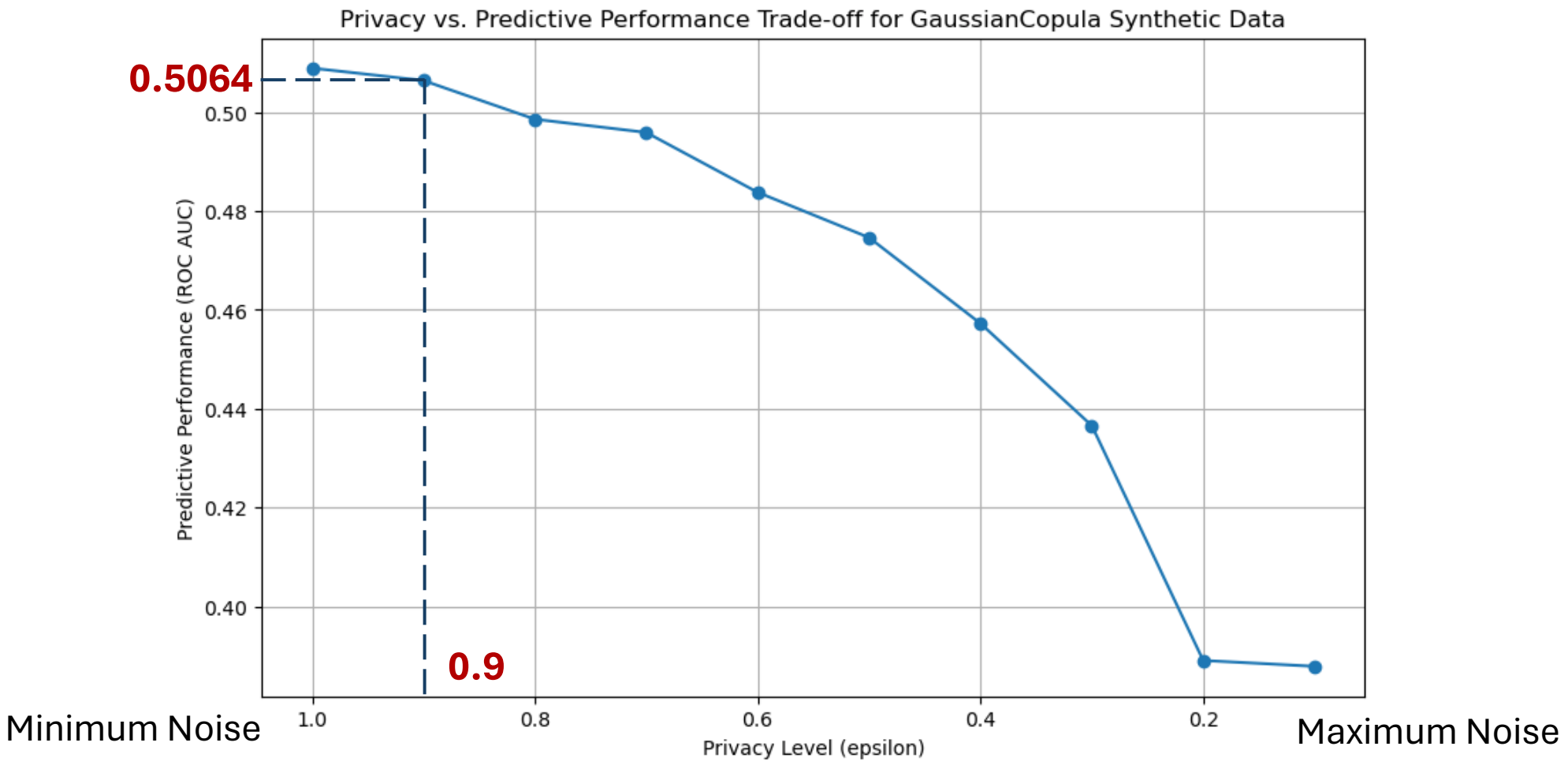
Customizable Synthesizer

Suited for multivariate data, enabling the generation of synthetic datasets that accurately reflect the joint distribution of multiple variables

Faster Performance

It is computationally more efficient. Its reliance on statistical models rather than neural networks means it can often generate synthetic data faster

Predictive Power starts to drop drastically at epsilon 0.9 for Gaussian Copula



DataSynthesizer

Data Describer

Examines the real dataset to understand its characteristics without storing individual details.

Data Generator

Constructs a statistical model based on the analyzed data, capturing its overall structure and patterns.

Model inspector

Uses the model to create a new dataset that resembles the original in statistics and relationships but contains no real personal information.

DataSynthesizer

Correlated Attribute Mode	• Learns a differentially private Bayesian network to capture attribute correlations • Draws samples from the model to create the dataset
Independent Attribute Mode	• Derives histograms for each attribute and adds noise for privacy • Draws independent samples for each attribute
Random Mode	• Generates type-consistent random values for each attribute • Used for extremely sensitive data

Comparison of Synthetic data generation methods

	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

- Prediction Model with Logistic Regression
- DataSynthesizer, VAE: Could **not generate binary columns** well

Exploring Innovative Paths Forward

Differential Privacy methods

- **Diffprivlib**
  IBM's Differential Privacy Library is a software toolkit developed by IBM to facilitate the implementation of differential privacy techniques in various applications and systems
- **PySyft**
  PySyft is an open-source Python library for privacy-preserving machine learning (PPML) and secure multi-party computation (MPC)

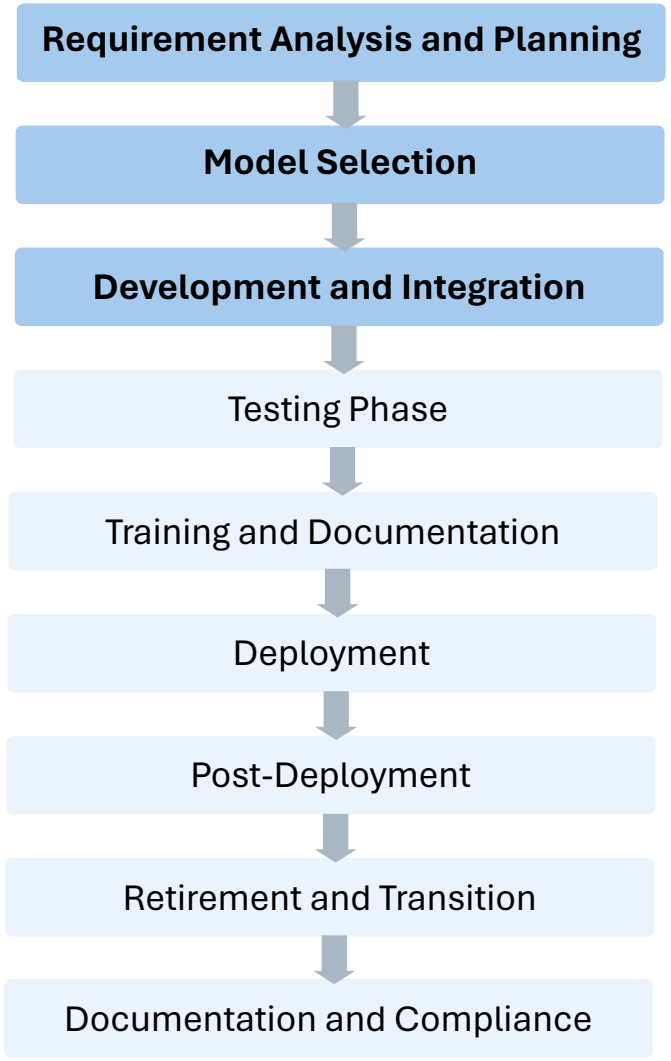
Explore and Experiment:

- SMOTE before and after noise addition
- Parameter tuning in Synthetic Data Generator Models (Epochs, Learning rate)

Product based solutions:

gretel

Deployment and Lifecycle Management



Requirement Analysis and Planning

- Identify Goals
- Regulatory compliance
- Stakeholder involvement and resource allocation

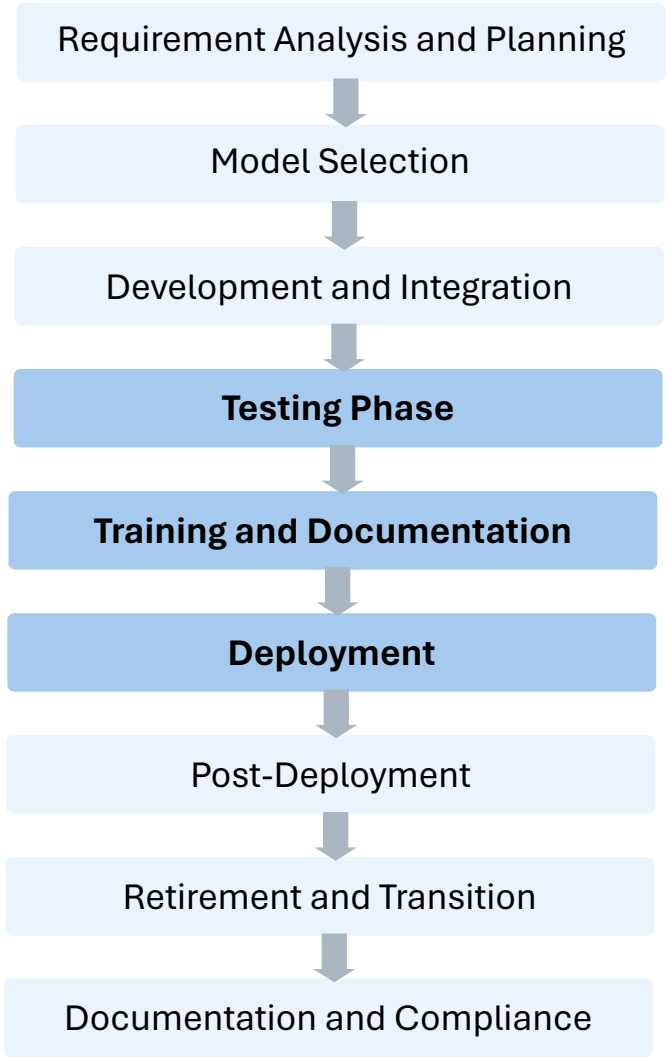
Model Selection

- Evaluation and testing
- Validation
- Final Selection

Development and Integration

- Customization
- Integration
- Security measures

Deployment and Lifecycle Management



Testing Phase

- Unit and integration testing
- Compliance testing
- User acceptance testing

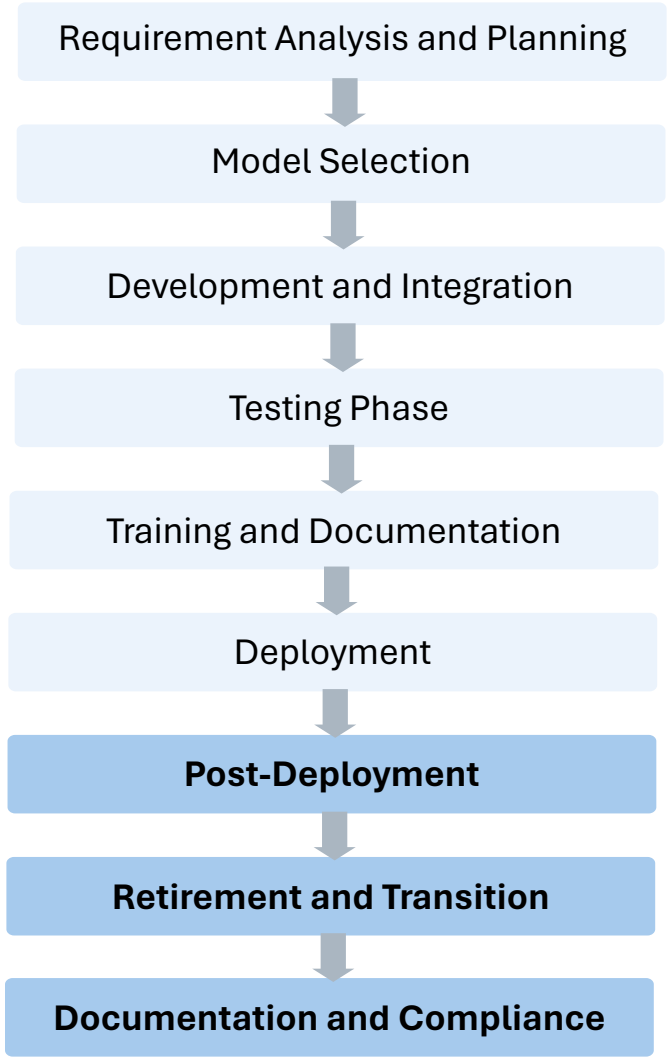
Training and Documentation

- Staff training
- Documentation

Deployment

- Staging deployment
- Monitoring and support
- Rollout

Deployment and Lifecycle Management



Post-Deployment

- Continuous monitoring
- Feedback loop
- Updates and maintenance

Retirement and Transition

- Assessment
- Decommissioning Plan
- Transition

Documentation and Compliance

- Record-keeping
- Review and Audit

Revolutionizing Healthcare Research

Enable studying disease trends without compromising privacy

Improve medication effectiveness analysis across demographics

Personalize treatment plans using safe data

Fueling Company Growth

Meeting the increasing demand for healthcare insights

Gaining competitive edge with unique data solutions

Expanding into new markets and research domains

Unlocking Predictive Healthcare

AI models predicting disease outbreaks and at-risk populations

Personalizing preventative care based on advanced analytics

Shaping public health strategies with predictive insights

Learnings (1/2)

Complex Ecosystem

The healthcare industry is a vast and intricate ecosystem comprising hospitals, clinics, research institutions, insurance companies, and regulatory bodies.

Regulatory Environment

A key learning is the significant role of regulations such as HIPAA in the US, GDPR in Europe, which govern data privacy, patient rights, and data sharing across the sector.

Patient-Centric Models

A shift towards patient-centric models emphasizes personalized healthcare, preventive measures, and patient engagement through digital platforms.

Innovation and Technology Adoption

The rapid adoption of technologies like EHRs, telemedicine, and AI-driven diagnostics is reshaping care delivery but also presents integration and standardization challenges.

Learnings (2/2)

Synthetic Data Generation

Understanding and deploying advanced models to generate synthetic data that mimics real patient data while ensuring privacy and compliance with healthcare regulations

Model Evaluation and Validation

Learned the processes for evaluating synthetic data models for accuracy and regulatory compliance. This included techniques for validating the utility of synthetic data in real-world scenarios

Applications of Synthetic Data

Explored innovative applications of synthetic data in healthcare, such as improving disease prediction models, enabling risk-free data sharing for research, and enhancing the development of personalized medicine.

Future Directions

Recognized the potential for future advancements in synthetic data technology to further revolutionize healthcare research, privacy preservation, and the development of AI-driven healthcare solutions.

Conclusion

Foundation for Personalized Healthcare Solutions

*The project we undertook, set up the **groundwork for our client** to venture into setting up new product lines that provide analyses and personalized recommendations at the patient level*

Advancing Research with Synthetic Data

*Evaluation of Synthetic Data Generators and their pitfalls will prove crucial in helping the client make datasets available for **research quicker and more efficient***


THANK YOU
NETWORK DOCTOR
EKG
MEDICAL
MEDICAL
ECC
MRI