HARNESSING ARTIFICIAL DATA GENERATION TO PROTECT SENSITIVE DATA

Mitchell E. Daniels, Jr. School of Business

Pratik Kamat, Anusha Reddy, Naveen Shaji, Prashanth Suresh, Amisha Turkel, Matthew A. Lanham

Purdue University, Mitchel E Daniel, Jr, School of Business

kamat1@purdue.edu; reddy118@purdue.edu; shaji@purdue.edu; suresh80@purdue.edu; turkel@purdue.edu; lanhamm@purdue.edu

BUSINESS PROBLEM

Firms collect and analyze sensitive consumer data to gain insights about their business and develop cutting edge strategies. With increasing regulations and risk of sensitive data leakage, firms employ several **stringent practices** to ensure data privacy. However, these practices drive-up **operational costs** and **opportunity losses**. Synthetic data allows firms to relax these practices at the cost of predictive power. In collaboration with a national timeshare firm, our solution generates synthetic data that provides a **high level of data privacy without compromising on model performance**.

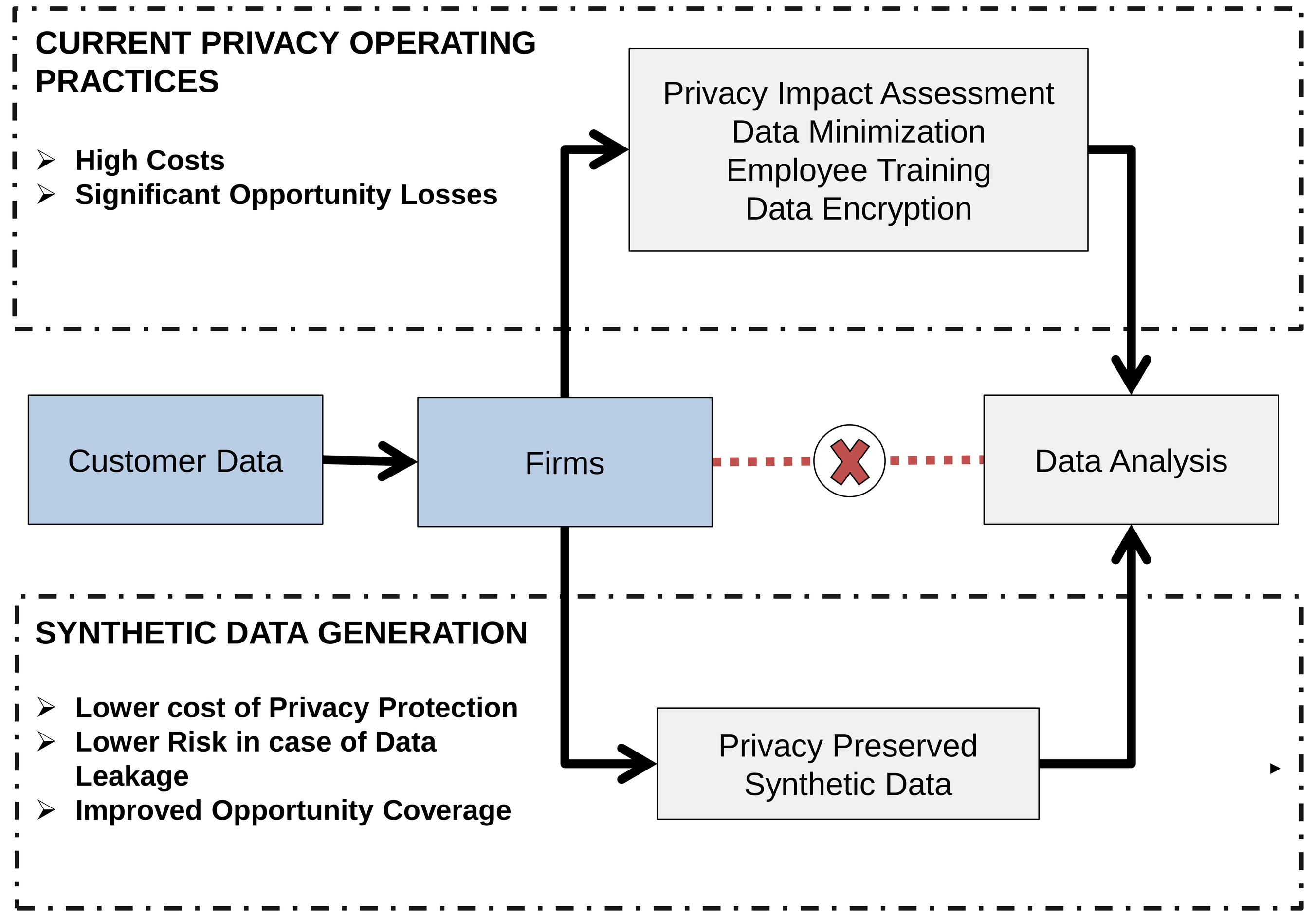

Fig 1. Current VS Recommended Privacy Operating Practice

Synthetic data generation is the process of **artificially generating data** that preserves **data privacy** while retaining information for meaningful analysis. Firms using synthetic data can expect **lower cost** of privacy operations, **lower risk** in case of data leakage and improved opportunity coverage. However, with increasing levels of privacy, synthetic data loses its ability to retain information and may impact model performance.

ANALYTICS PROBLEM

➢ Study the trade-off between privacy offered by synthetic data and its predictive power.

➢Our scope is limited to studying the impact of synthetic data generation for imbalanced binary classification problems and model performance will be evaluated using ROC score.

➢Identifying the right methodology to generate synthetic data that offers high levels of privacy for a negligible loss in opportunity coverage for our client, denotes the success of this engagement.

DATA

Dataset is sourced from a timeshare firm and contains customer membership and past transaction details.

Highly imbalanced dataset - ROC score is used instead of accuracy to evaluate model performance.
Presence of outliers: Outliers were identified and capped in the original dataset.
Constraints exist between features: Some features are restricted in value by other features.

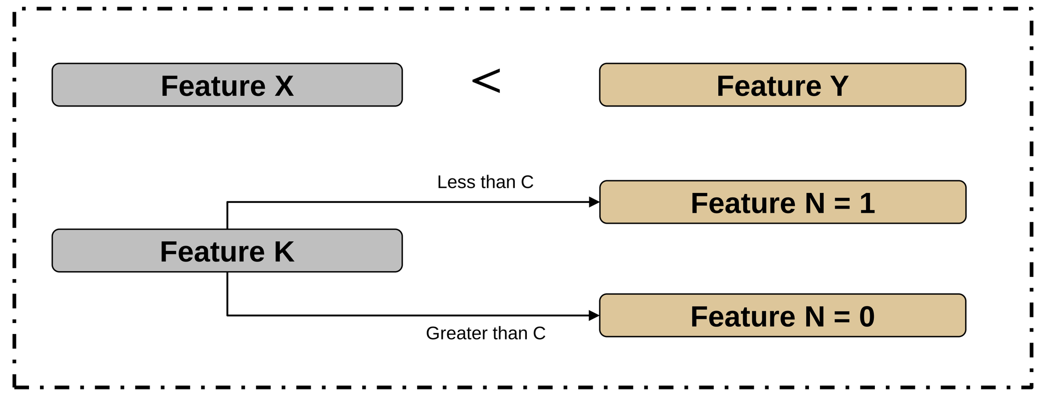

Fig 2. Data Constraints

Measures to ensure that the synthetic data and real data conform to the same template:
- Identifying datatypes of features to define metadata.
- Formulating data constraints that need to be fed into the synthetic data generators.

**Detailed description withheld for confidentiality.*

METHODOLOGY

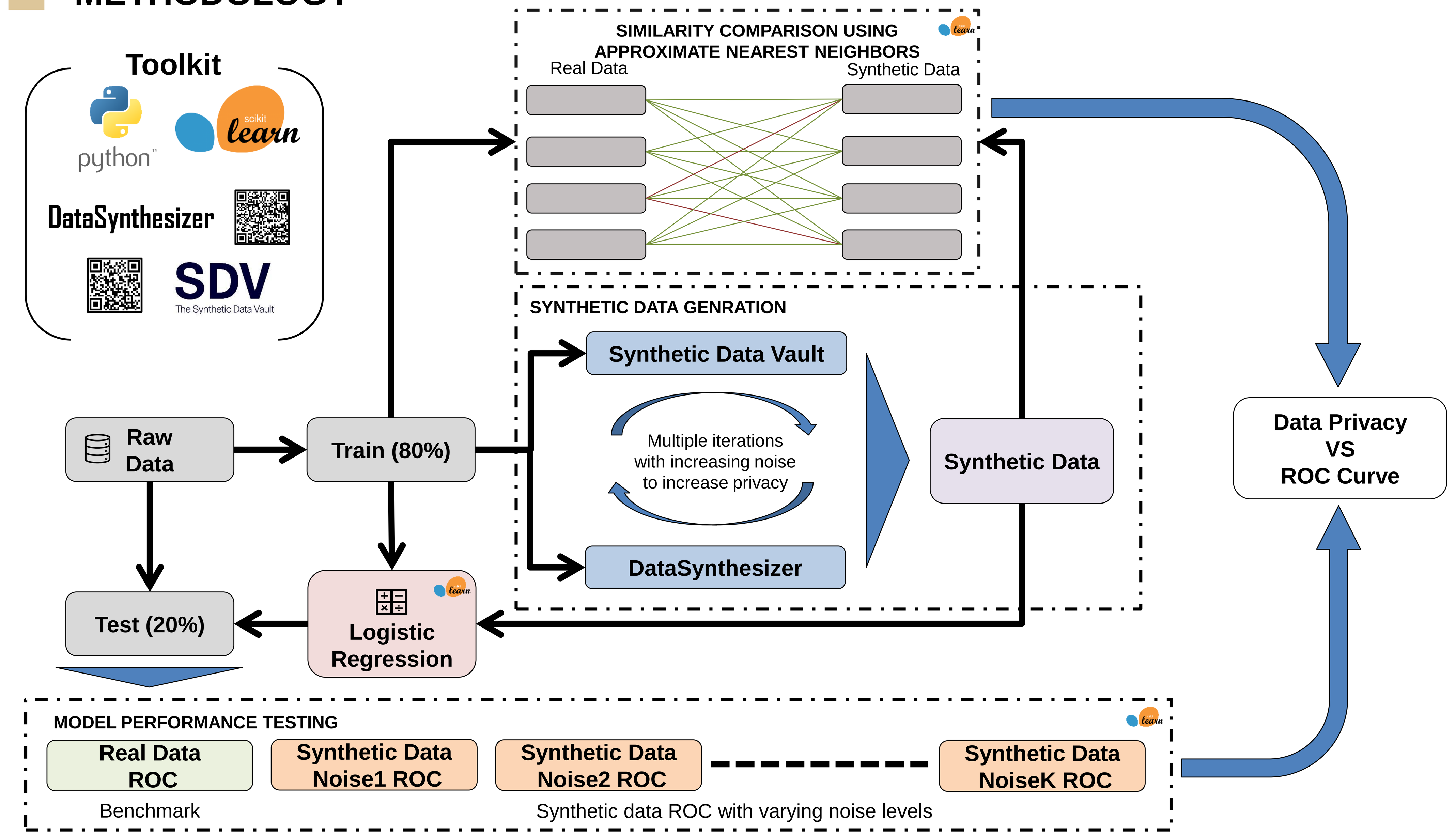

Fig 3. Methodology

STATISTICAL RESULTS

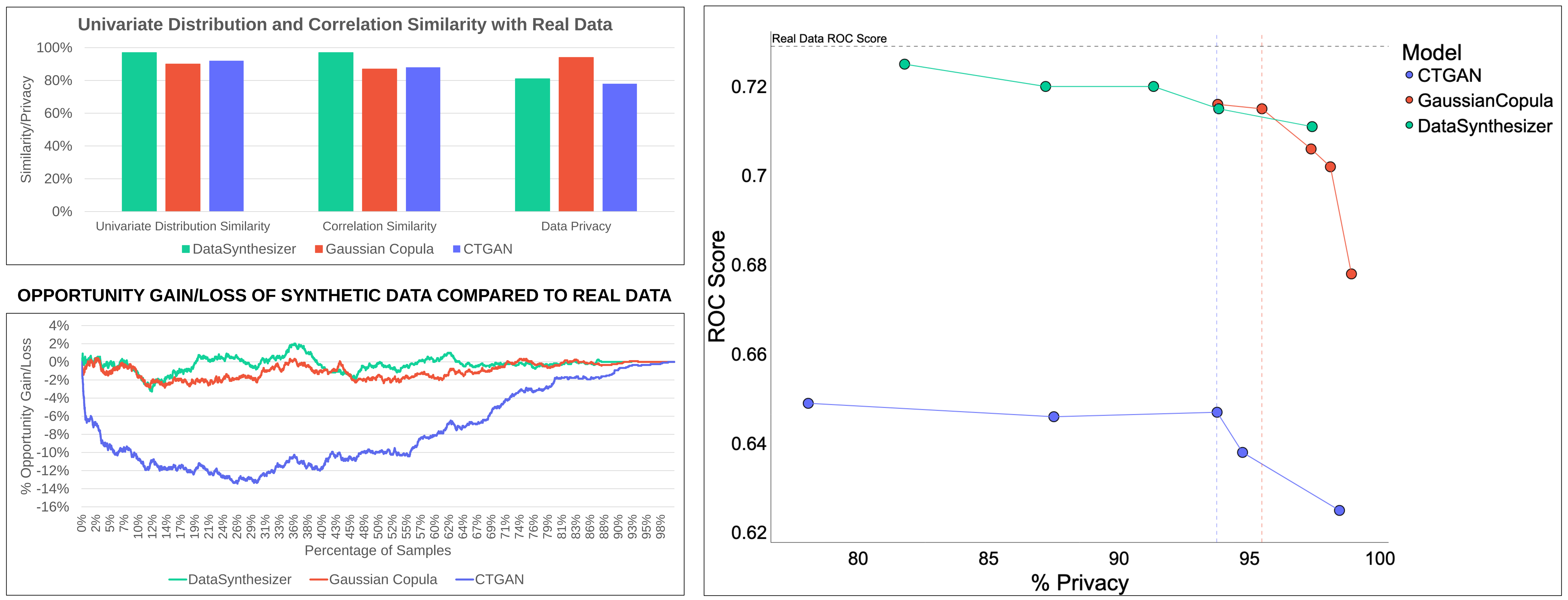

Fig 4. Model Results

BUSINESS IMPACT ASSESSMENT

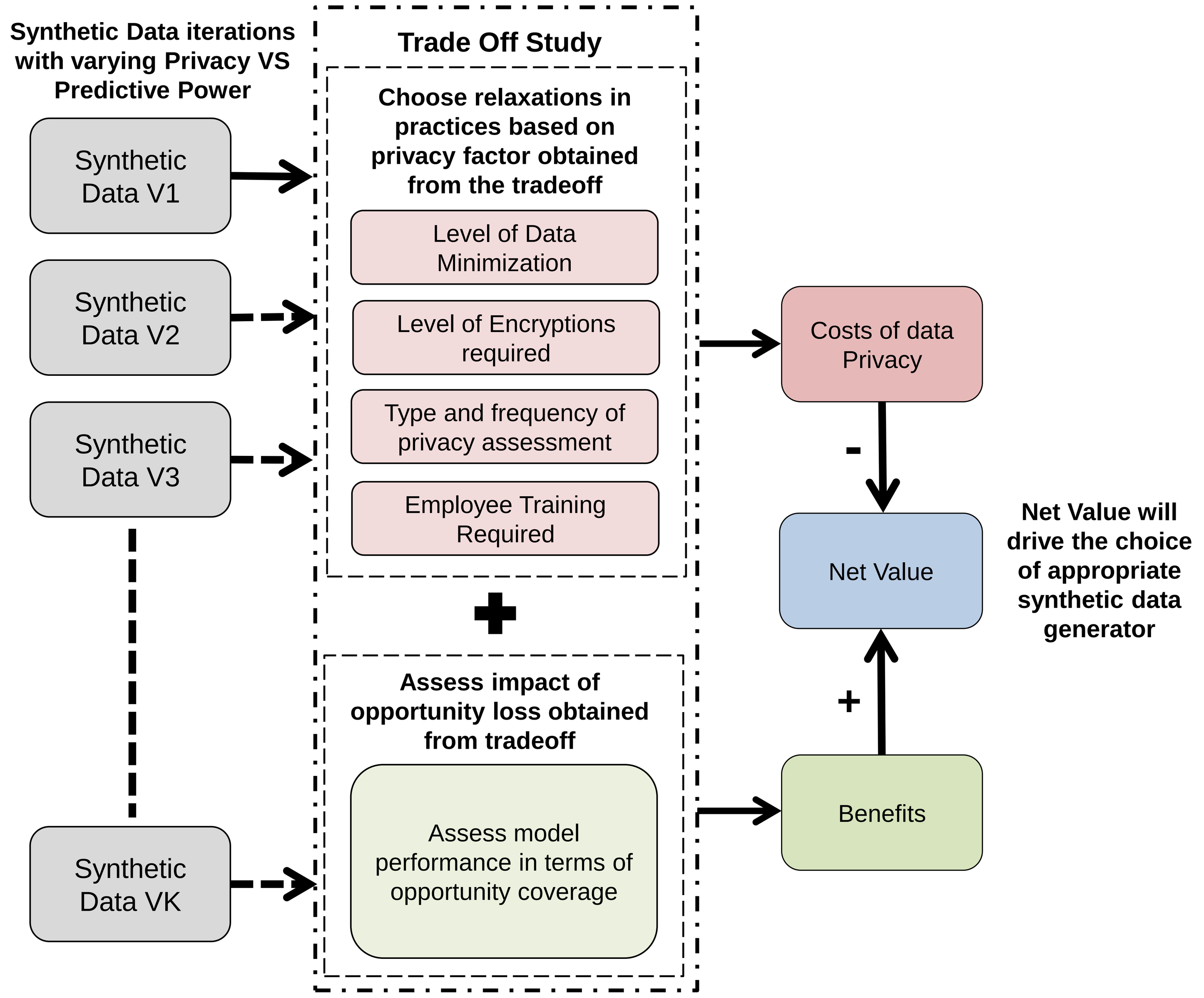

Fig 5. Net Value Assessment Process

Firms can leverage the tradeoff profile generated in the study to assess the expected **net value** in adopting various **synthetic data generators**. The generation method with the highest net value can then be compared with the net value under current practices to arrive at the **optimal strategy** for adoption of **data privacy practices**.

DEPLOYMENT

Business Priority	DEPLOYMENT RECOMMENDATIONS		
	DataSynthesizer	Gaussian Copula	CTGAN
Model Performance	✓		
Privacy and model performance	✓	✓	
Data has many features/columns	✓		
Maintaining feature constraints		✓	✓

- ➢ In most settings, we recommend DataSynthesizer which provides **81% data privacy with the same opportunity coverage** as real data.
- ➢ Adding small amounts of noise is a powerful technique to improve data privacy with negligible drops in opportunity coverage. However, adding noise beyond a point can adversely affect model performance

ACKNOWLEDGEMENTS

We would like to express our gratitude to Professor Matthew Lanham and our industry partners for this opportunity and their support throughout this project.