# Examining the Utility of Synthetic Trace Data

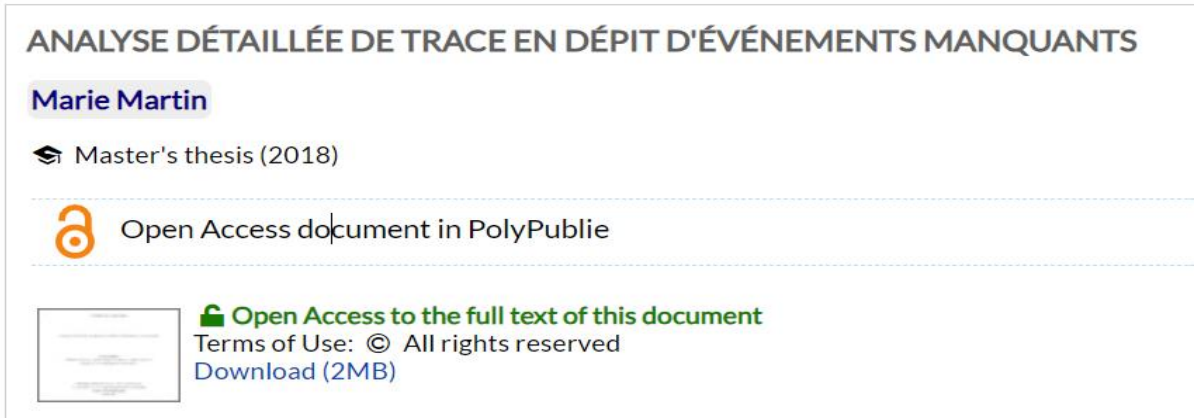
Madeline Janecek (mj17th@brocku.ca)

Kasra Darvishi (kdarvishi@brocku.ca)

Naser Ezzati-Jivan (nezzatijivan@brocku.ca)

# Introduction

- **Generative models** learn the structure of training data to produce new data with similar characteristics [1]
- Is it possible to generate synthetic trace events using generative techniques? If yes, this could mean:
  1. Enhancing datasets used to train machine learning models [2, 3, 4]
  2. **Reconstructing lost trace events**


ANALYSE DÉTAILLÉE DE TRACE EN DÉPIT D'ÉVÉNEMENTS MANQUANTS

Marie Martin

Master's thesis (2018)

Open Access document in PolyPublie

Open Access to the full text of this document

Terms of Use: © All rights reserved

Download (2MB)

# Goals

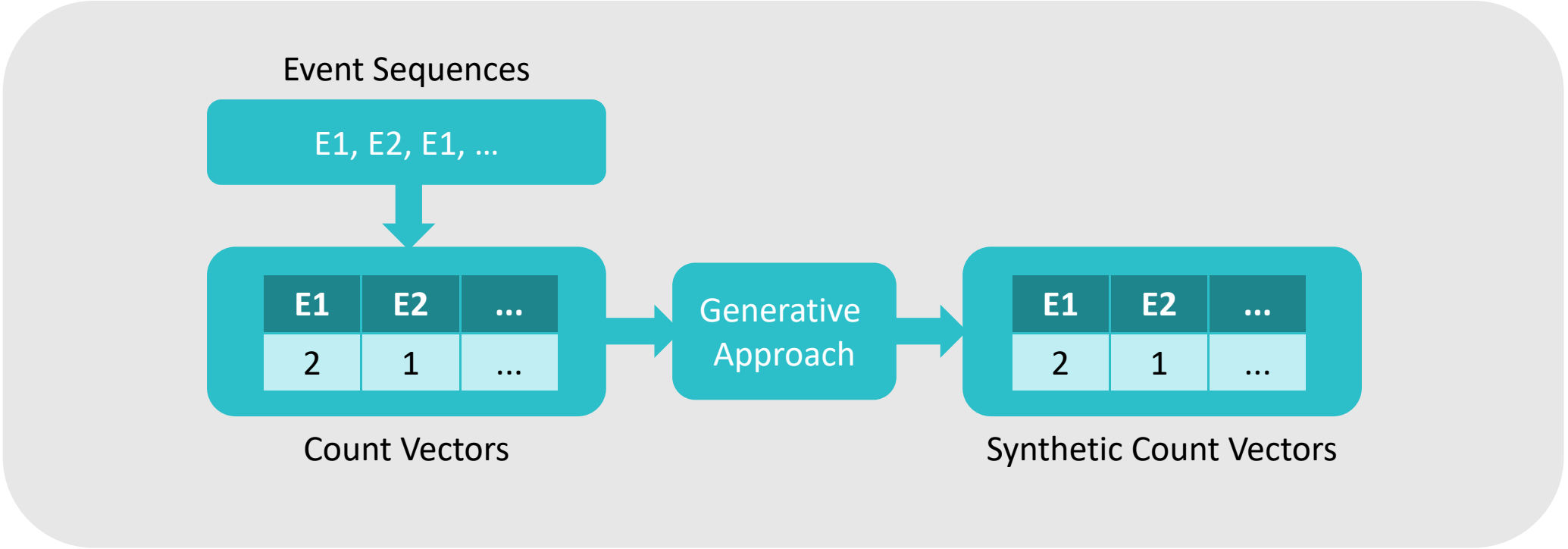
The goal of this stage is to explore different generative methods' efficacy using vector representations of trace events. Specifically, we look at:

1) Quality
   - Novelty: Synthetic data should not just be a copy of authentic data
   - Diversity: Generative models should aim to introduce as much variability as possible without sacrificing fidelity
   - Fidelity: Synthetic data should be indistinguishable from authentic data
2) Utility
   - How useful can synthetic data be in enhancing existing trace analysis methods?

Dataset Enhancement

# Framework

We start by looking at synthesizing event count vectors



# Models

We analyze and compare various approaches that can be broadly classified into three categories:

1) **Probabilistic Models** – Statistical models created to model authentic data that are then sampled
2) **Genetic Algorithms** – Evolutionary techniques that search for optimal samples by adhering to specified criteria
3) **Neural Networks** – Models like GANs, VAEs, and Transformers that are trained to produce realistic samples

Dataset Enhancement

# Quality Testing: Datasets

For two datasets, we quantify each method's quality when tasked with generating 2000 synthetic samples for each class.

1) ADFA-LD[1] dataset
   - Class 1: System call sequences exhibiting normal behaviour
   - Class 2: System call sequences collected during malicious attacks
2) HDFS_v1[2] dataset
   - Class 1: Log sequences exhibiting normal behaviour
   - Class 2: Log sequences collected during injected performance anomalies

[1] https://research.unsw.edu.au/projects/adfa-ids-datasets
[2] https://github.com/logpai/loghub/blob/master/HDFS/README.md

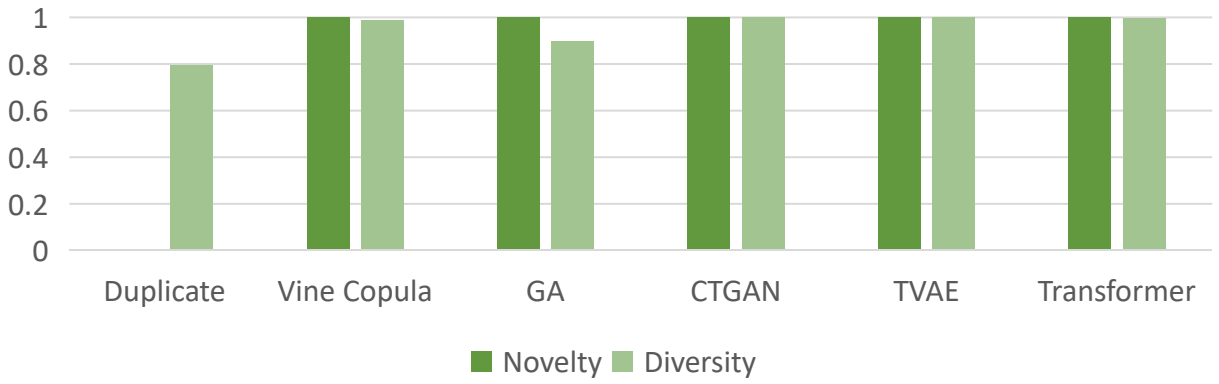
# Quality Testing: Metrics

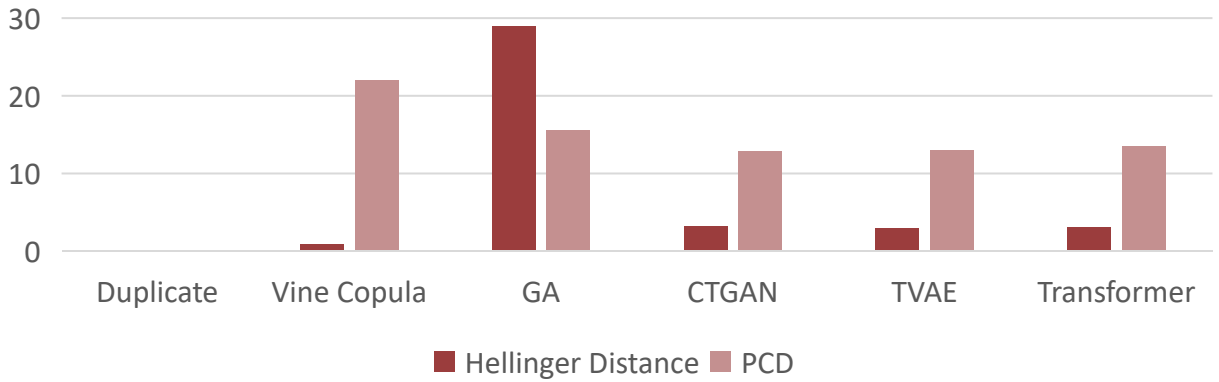
- **Novelty** is computed as the proportion of synthetic samples that are not already seen in the training set
- **Diversity** is computed as the proportion of synthetic samples that are only found once in the generated set
- **Fidelity** is computed in two ways:
  1) The univariate fidelity is computed using the variables' average Hellinger distance [5, 6]
  2) The bivariate fidelity is computed using the Pairwise Correlation Difference (PCD) [5, 7]

Dataset Enhancement

# Quality Testing: Findings

- When working with the diverse training data (> 0.6 score), every model can achieve high novelty and diversity scores
- Network-based models achieve the best fidelity in these cases

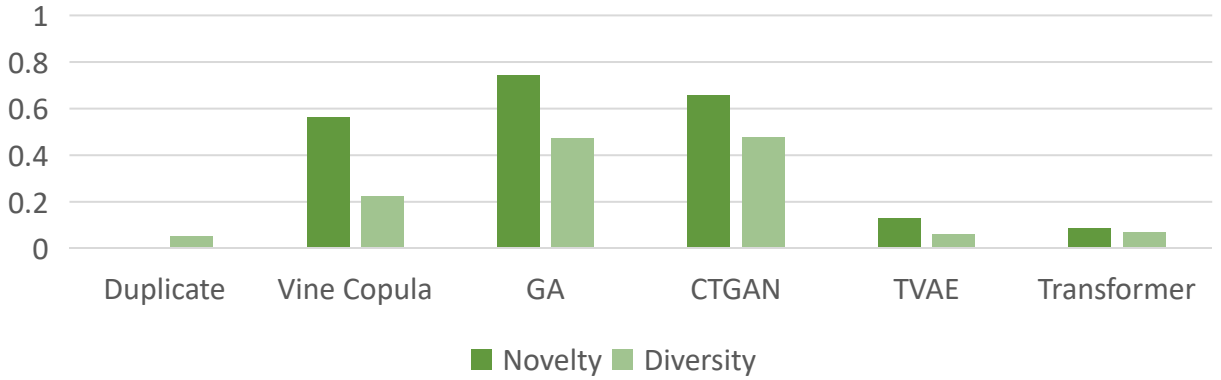


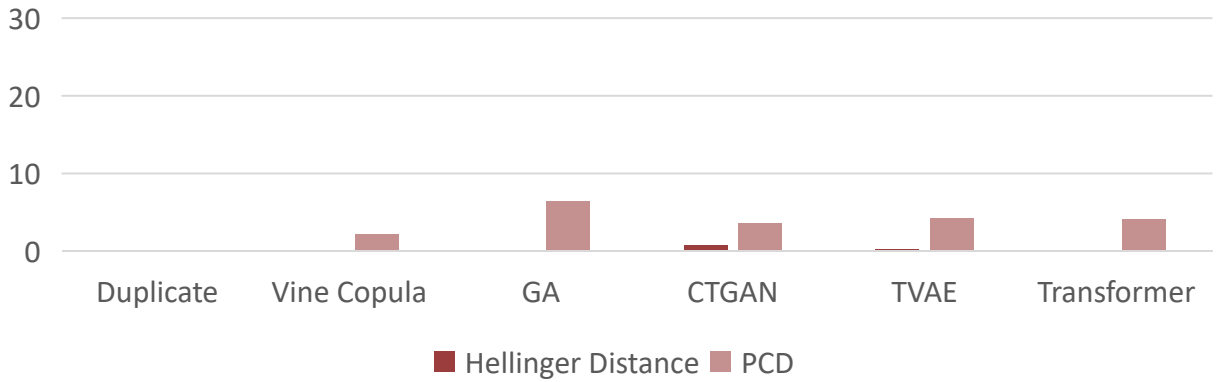



Dataset Enhancement

# Quality Testing: Findings

- When working with training data that has a low diversity (< 0.1 score), network-based approaches sacrifice novelty and diversity to maintain fidelity
- In these cases, data created using Genetic Algorithms have the best overall quality



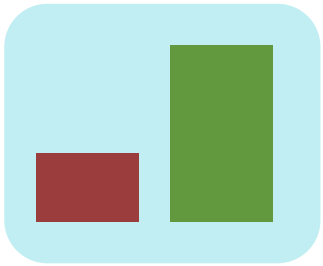
Dataset Enhancement

# Utility Testing

- The performance of machine learning models is directly linked with the quality, diversity, and relevance of their training data
- Anomalies are rare, which often leads to severe class imbalances when training supervised learning models to perform anomaly detection
- To measure utility, we examine if performance improves when including synthetically generated anomalies in a classifier's training data

Dataset Enhancement

# Utility Testing: Framework

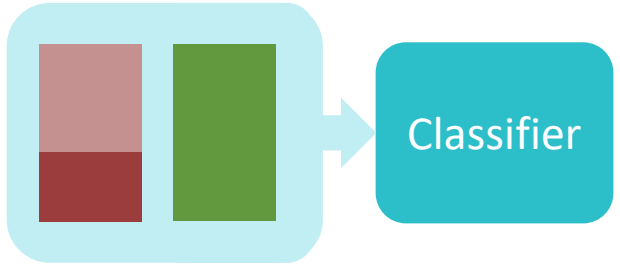


**1) Training Data**

Select a training dataset


**2) Data Synthesis**

Use a generative approach to learn the structure of the underrepresented class (or classes) and create synthetic samples

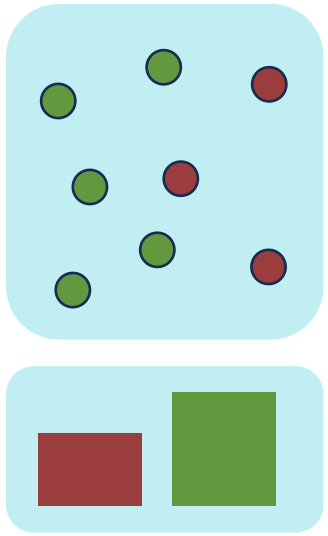



**3) Classifier Training**

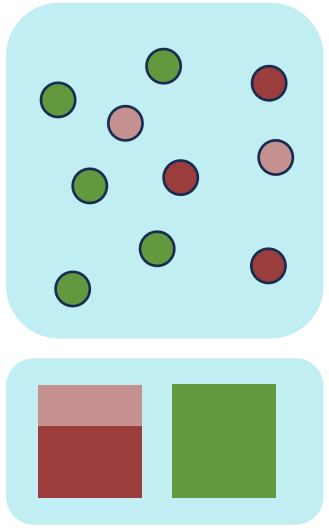
Train a classification model using the artificially balanced dataset

Dataset Enhancement

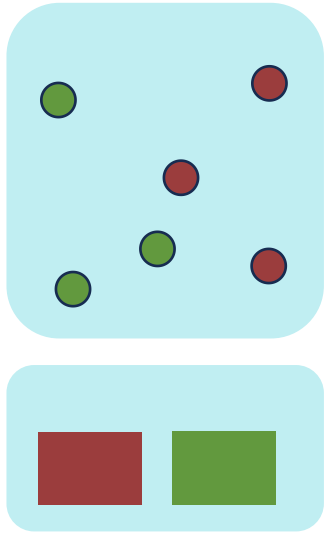
# Utility Testing: Baselines



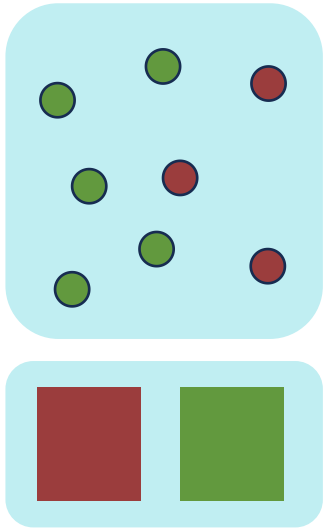
**1) Original Data**



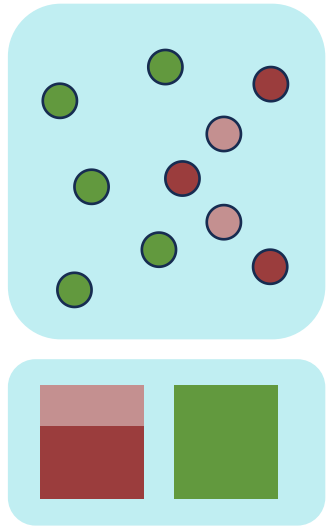
**2) Random**
Add samples that are randomly generated



**3) Random Undersampling (RUS)**
Randomly remove samples from the larger class



**4) Random Oversampling (ROS)**
Randomly duplicate samples from the smaller class

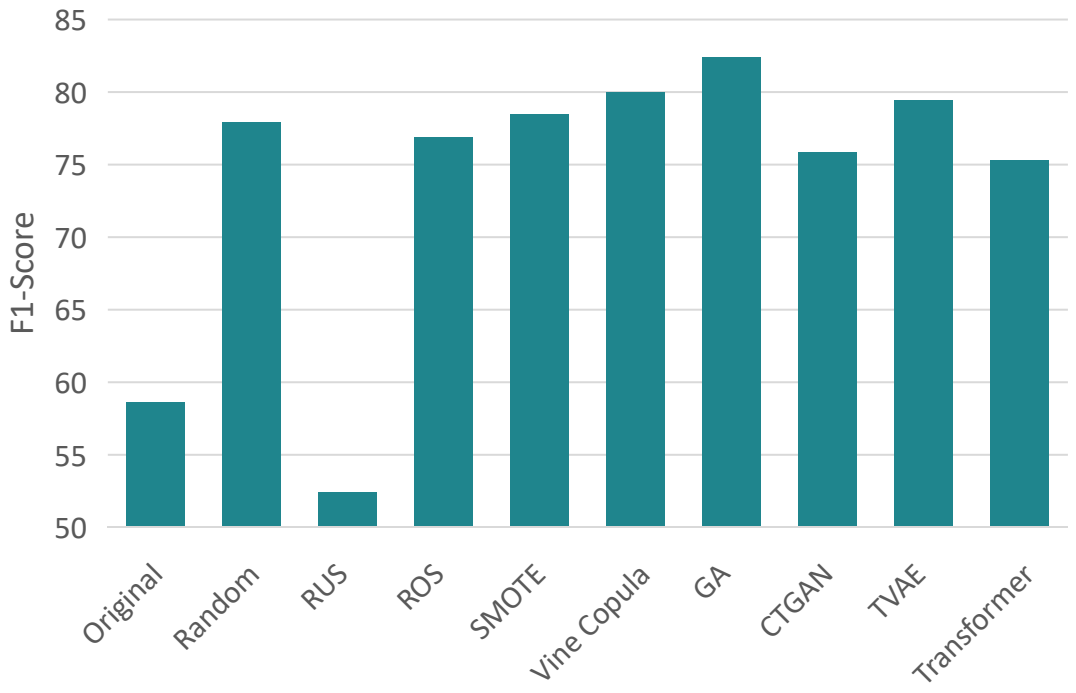


**5) SMOTE [8]**
An oversampling technique that randomly adds samples between authentic samples based on their nearest neighbours

Dataset Enhancement

# Utility Testing: Results

| Type | Approach | Accuracy | F1-Score |
|---|---|---|---|
| Baseline | Original Data | 88.130 | 58.594 |
| | Random | 94.681 | 77.958 |
| | RUS | 78.891 | 52.459 |
| | ROS | 93.057 | 76.866 |
| | SMOTE | 94.065 | 78.455 |
| Probabilistic | Vine Copula | 94.345 | 80.00 |
| **Genetic Algorithm** | **Genetic Algorithm** | **95.577** | **82.40** |
| Neural Network | CTGAN | 93.169 | 75.889 |
| | TVAE | 94.793 | 79.470 |
| | Transformer | 92.497 | 75.277 |



Event Reconstruction

# Goals

- The goal of this stage is to explore different generative methods' ability to synthesize sequences of trace events
- The results of this work could be used to reconstruct trace events that are overwritten or discarded due to a high volume of events

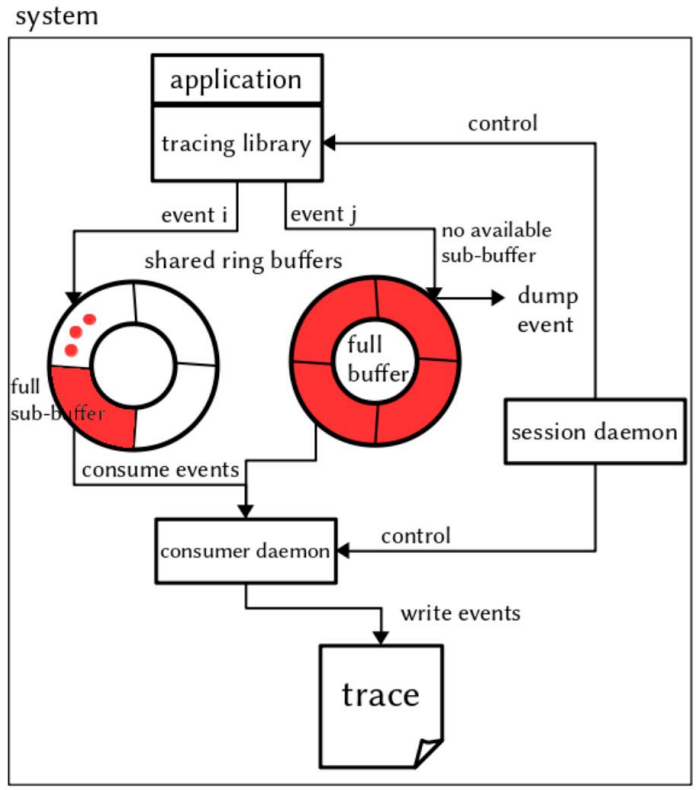
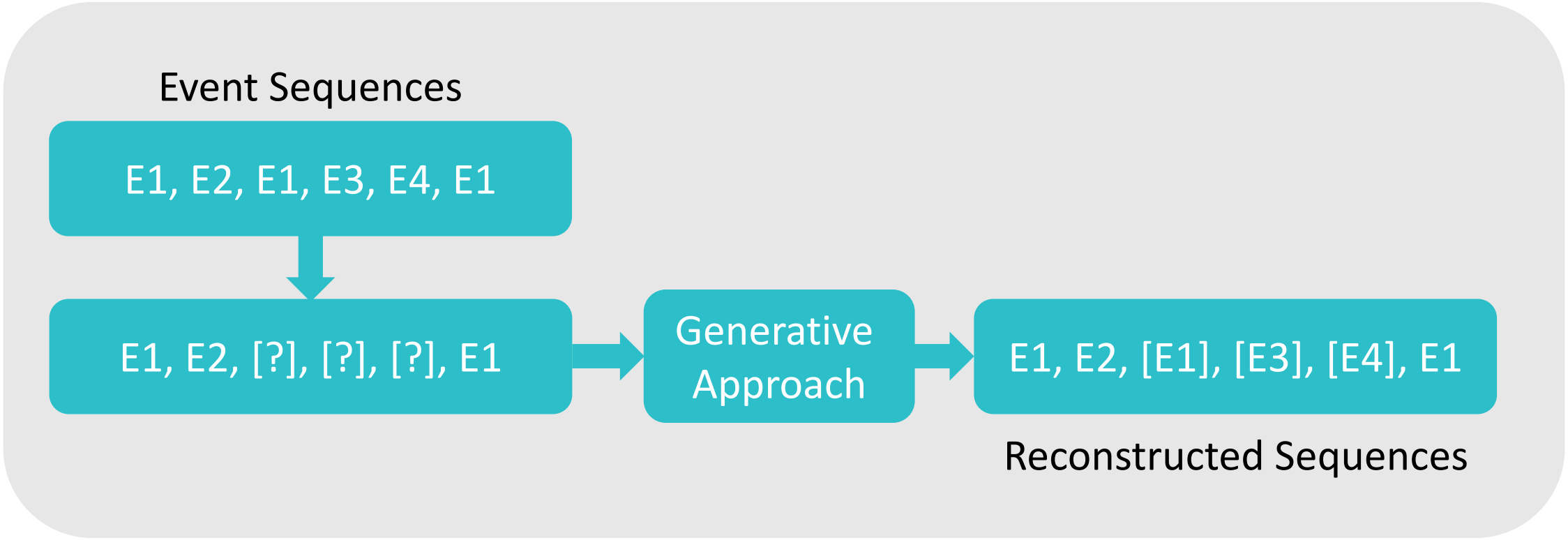



Figure 4.1 of Martin, M. (2018). *Analyse détaillée de trace en dépit d'événements manquants* [Master's thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/3248/

Event Reconstruction

# Framework



Event Reconstruction

# Model

- Structured state space diffusion (SSSD) models [9] combine the generative capabilities of diffusion models with structured state-space models (SSM) [10] reconstruct missing time series data
- Have seen success over other SOTA models with continuous data (ECG, electricity usage patterns, etc.)
  - May be adapted for kernel event sequence reconstruction

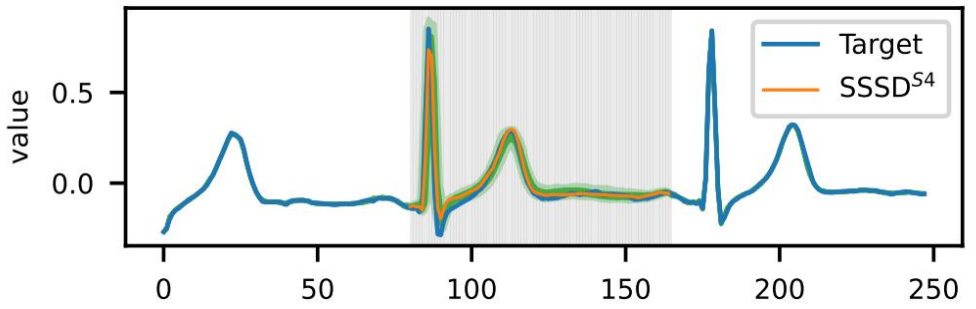



Figure 3 of Lopez Alcaraz, J.M., & Strodthoff, N. (2022). Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. *ArXiv, abs/2208.09399*.

# Results

- Preliminary results seen with kernel traces collected with LTTng on an Apache Web Server are promising
- When given 5000 sequences of 100 events:

| Length of Missing Portion | Method | Event-Level Accuracy | % of Perfectly Recreated Sequences |
|---|---|---|---|
| 5 | Random | 3.24 | N/a |
| | SSSD | 80.16 | 56.4 |
| 10 | Random | 2.38 | N/a |
| | SSSD | 80.38 | 38.8 |
| 20 | Random | 2.61 | N/a |
| | SSSD | 78.06 | 14.2 |

# Future Work

- Employ various methods to improve the accuracy of SSSD model
- Synthesis of trace event arguments, durations, etc.

# Questions?

Madeline Janecek (mj17th@brocku.ca)

Kasra Darvishi (kdarvishi@brocku.ca)

Naser Ezzati-Jivan (nezzatijivan@brocku.ca)

# References


[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, “A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT,” ArXiv, vol. abs/2303.04226, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257405349

[2] Lu, Yingzhou & Wang, Huazheng & Wei, Wenqi. (2023). Machine Learning for Synthetic Data Generation: a Review. 10.48550/arXiv.2302.04062.

[3] Fonseca, J., Bacao, F. Tabular and latent space synthetic data generation: a literature review. J Big Data 10, 115 (2023). https://doi.org/10.1186/s40537-023-00792-7

[4] Dankar, Fida & Ibrahim, Mahmoud. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Applied Sciences. 11. 2158. 10.3390/app11052158.

[5] F. K. Dankar, M. K. Ibrahim and L. Ismail, "A Multi-Dimensional Evaluation of Synthetic Data Generators," in IEEE Access, vol. 10, pp. 11147-11158, 2022, doi: 10.1109/ACCESS.2022.3144765.

[6] Hellinger distance. Encyclopedia of Mathematics. URL: http://encyclopediaofmath.org/index.php?title=Hellinger_distance&oldid=47206

[7] Goncalves, A., Ray, P., Soper, B. et al. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 20, 108 (2020). https://doi.org/10.1186/s12874-020-00977-1

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 1 (January 2002), 321–357.

[9] Lopez Alcaraz, J.M., & Strodthoff, N. (2022). Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. ArXiv, abs/2208.09399.

[10] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In International Conference on Learning Representations, 2022