Generative models for synthetic panel data

Jessie Lamontagne, Paul Edwards (Scotiabank)
André dos Santos, Denilson Barbosa (University of Alberta)

March 25, 2021

UNIVERSITY OF ALBERTA

Scotiabank®

Who we are

Jessie Lamontagne

Data Scientist

André dos Santos

Postdoctoral Fellow

Introduction

Motivation

Data Privacy Will Be The Most Important Issue In The Next Decade

-Forbes, 2019

We are interested in synthetic data because it allows us to:

1. Develop a tool for analyzing and modelling data for which there are privacy concerns
2. Accelerate innovation by sharing representative, but synthetic data
3. Gain more insight into transactional behaviour

Can generative models allow us to learn about transactional behaviour at the same time as we create a fully private, synthetic dataset?

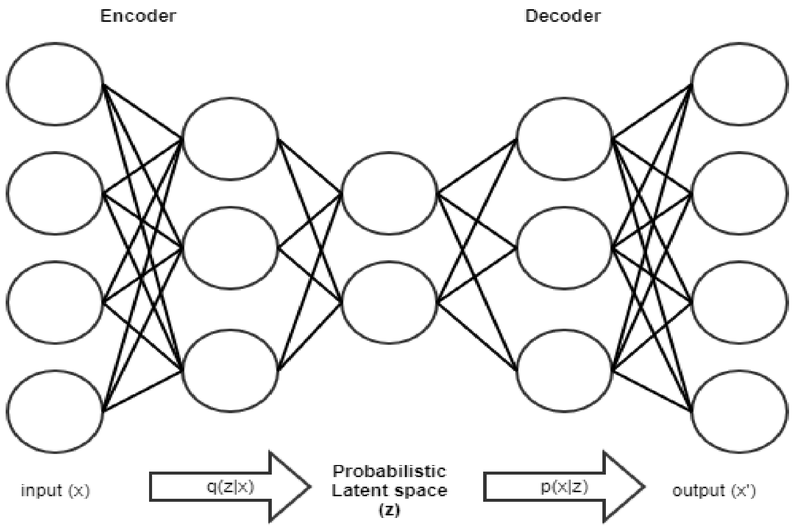


Fig. 1: A Variational autoencoder is one kind of generative model

Generative models

A generative model learns about a domain (commonly text or image) by fitting a function that minimizes the difference between the data it generates, and the true data.

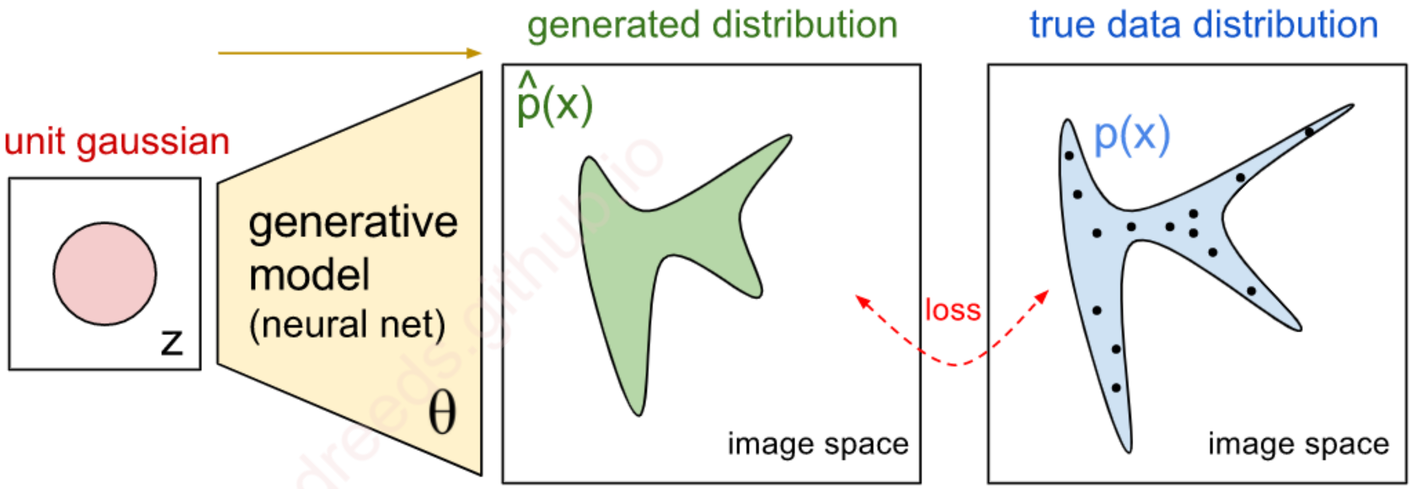


Fig 2: Overview of Generative Models

Advances in generative models came from Natural Language Processing and Computer Vision…

StyleGAN (2018)

Yet in a circle pallid as it flow,
by this bright sun, that with his light display,
roll'd from the sands, and half the buds of snow,
and calmly on him shall infold away

Deep-speare (2018)

Can we transfer architectural frameworks from these domains to the transaction domain?

Data

~4 million

IDs

¼ billion

transactions

24

months of records

Transaction data

What is a transaction?

*Table 1: A short sequence of transactions for customer "A"**

Customer	Card type	Transaction amount	Datetime stamp	Merchant category type
A	Debit	200.00	2019-05-08 02:15:22	ABM
A	Debit	22.50	2019-05-09 12:22:46	Grocery
A	Credit	78.72	2019-05-09 22:12:23	Restaurant
A	Debit	200.00	2019-05-11 09:23:34	ABM

*not a real customer

amount

card type

e-commerce

merchant type

time

Multi-objective optimization

Privacy-Utility: a pareto frontier

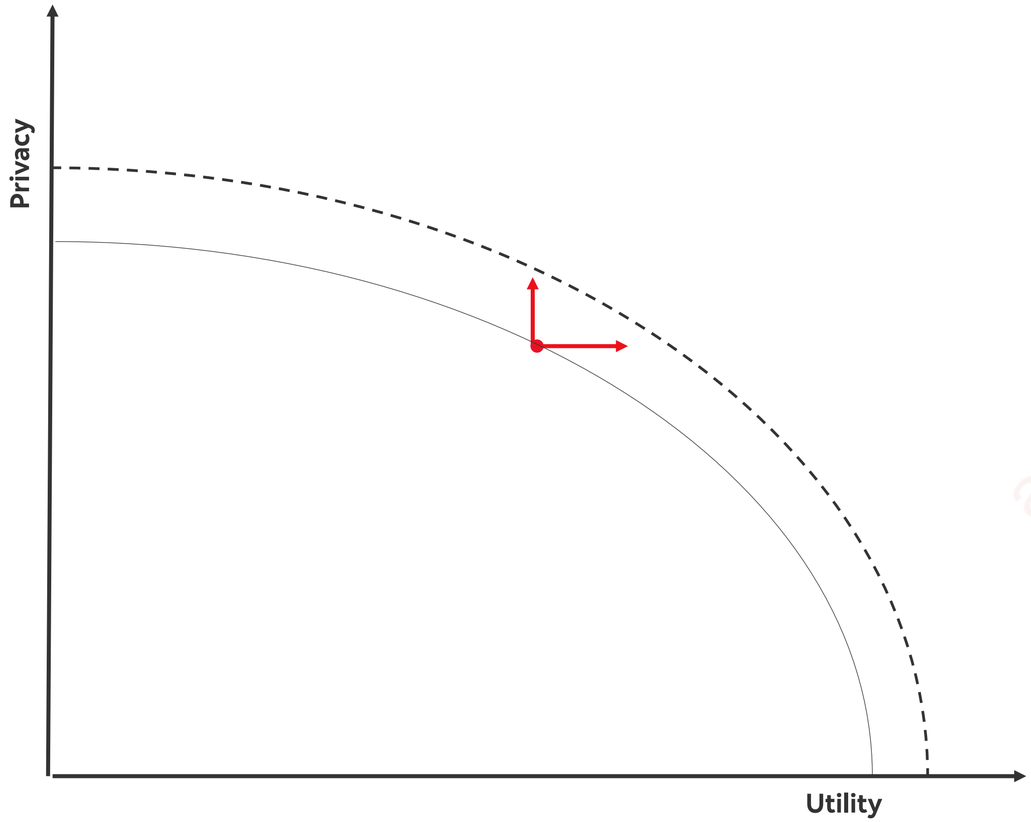


Fig 3: The utility-privacy optimization, and the pareto frontier

- A synthetic dataset has high **privacy** if none of the records can be mapped back to a real customer

- A synthetic dataset has high **utility** if it can be used to train a model that makes accurate prediction when faced with the real data

Developing a new technology for synthesizing data can help us **expand** the pareto frontier: for a given privacy level, more utility/for a given utility level, more privacy.

How do you show that a synthetic dataset is Private? Useful?

Privacy

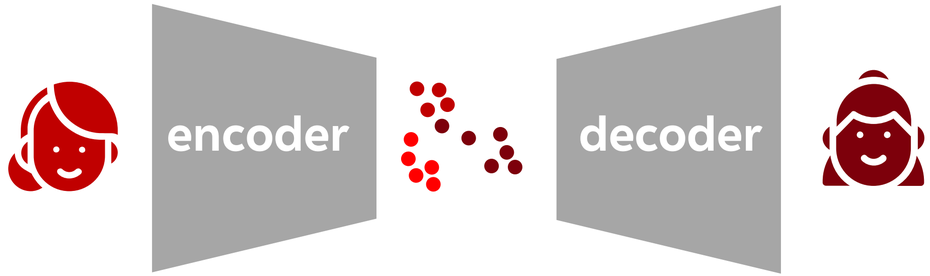
- Algorithmically
- Statistically
- Privacy challenge

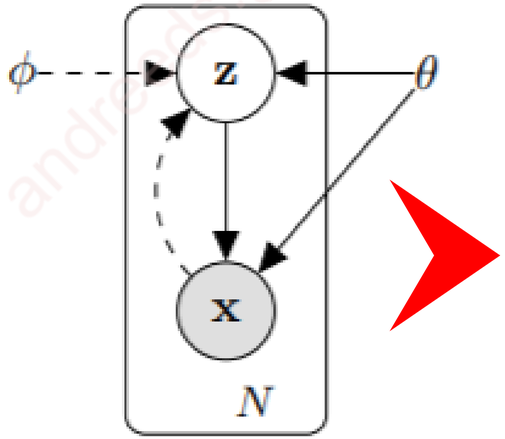
Utility

- Analysis of distributions
- Training a model

Technology

Variational Auto-Encoders





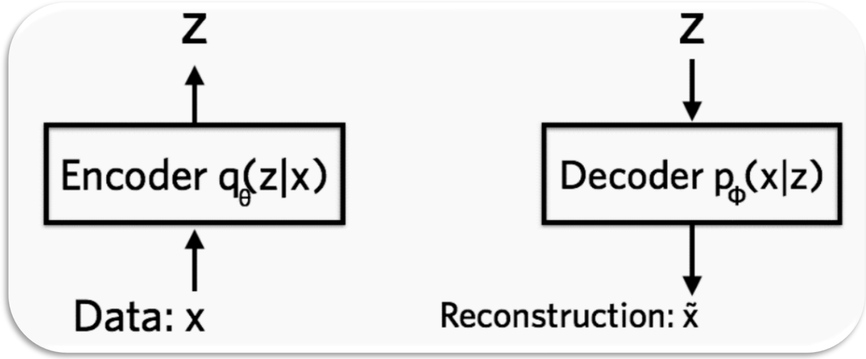
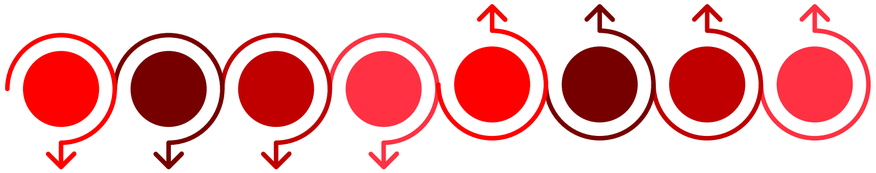


Fig 4: Understanding VAE as a directed graph, and the corresponding neural net representation. The encoder learns an approximation of the intractable posterior $p_\vartheta(z|x)$.

How do transactions tell a story?

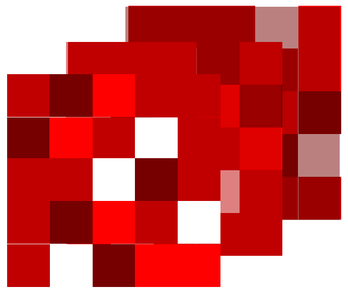
Transactions

1



LSTM-based
encoder-decoder networks

2



CNN-based
encoder-decoder networks

CNN-based encoder-decoder networks

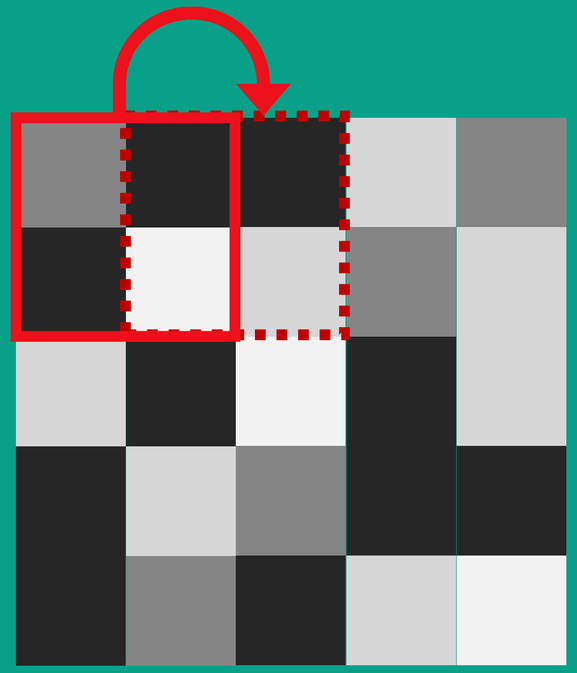
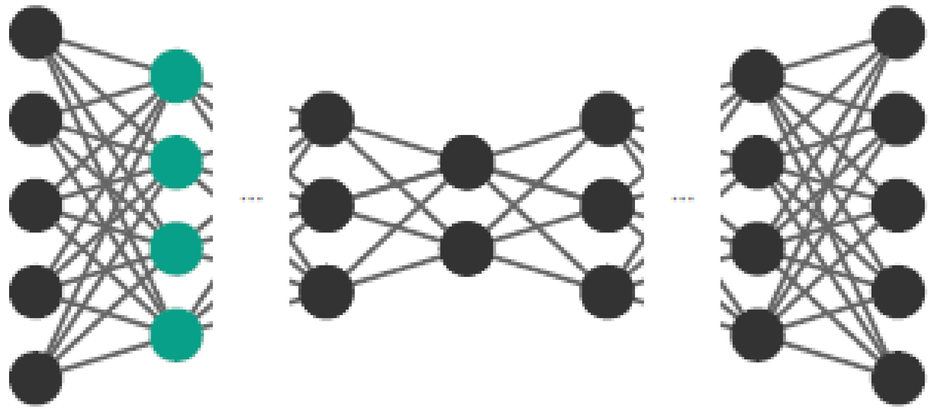
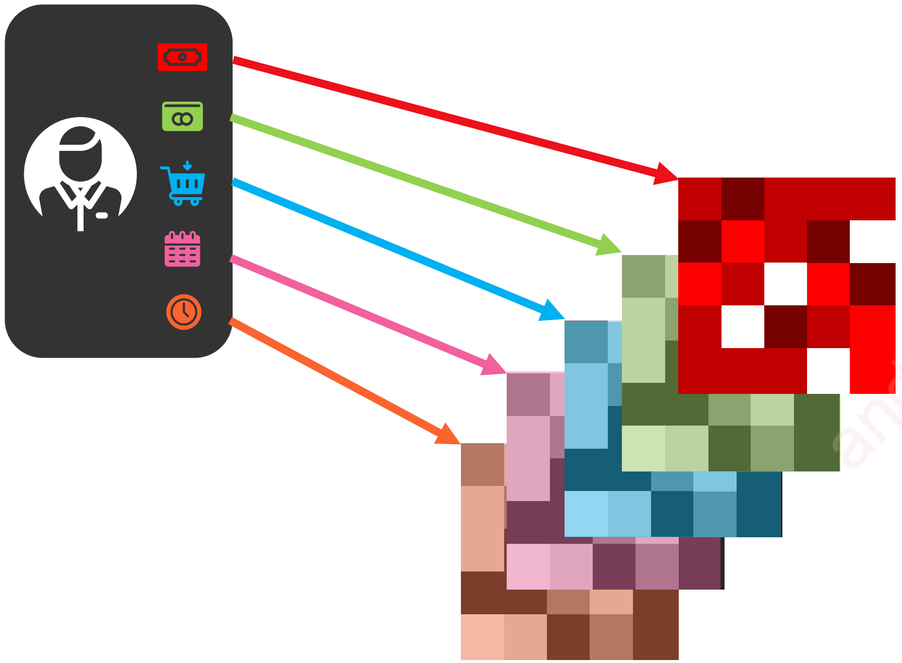


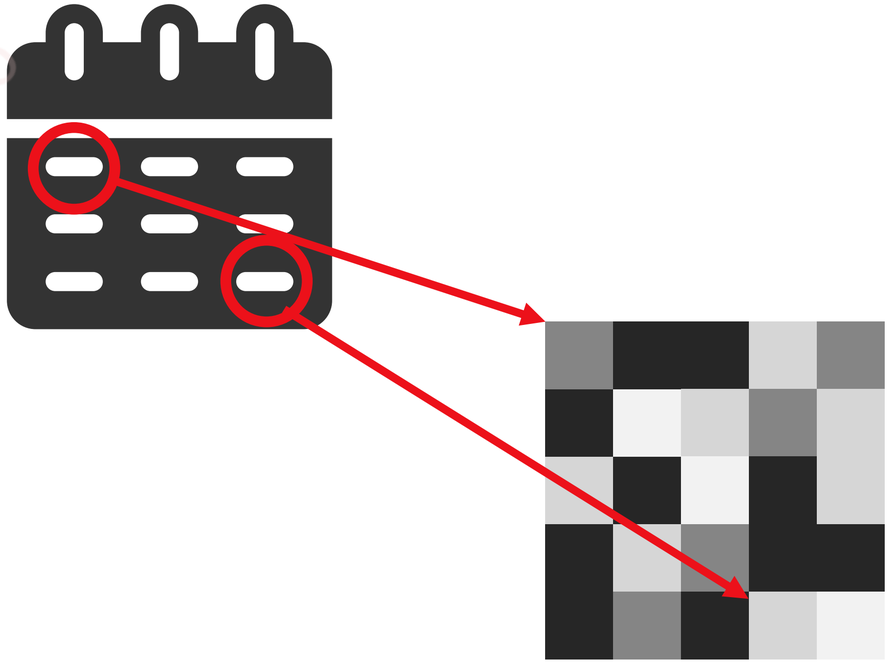
Fig 6: Convolutional window and stride



CNN-based encoder-decoder networks

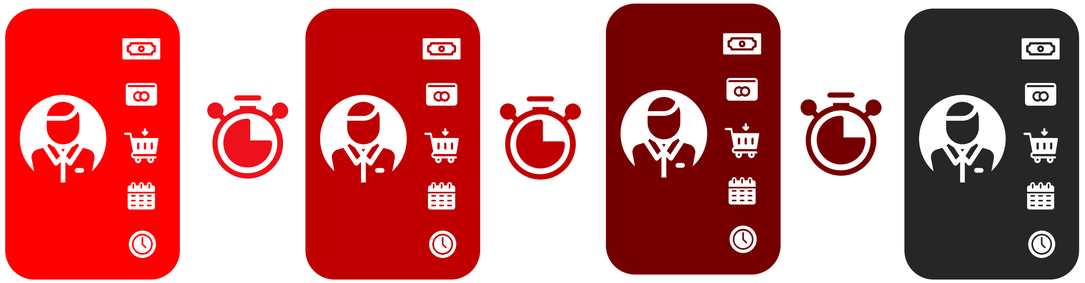


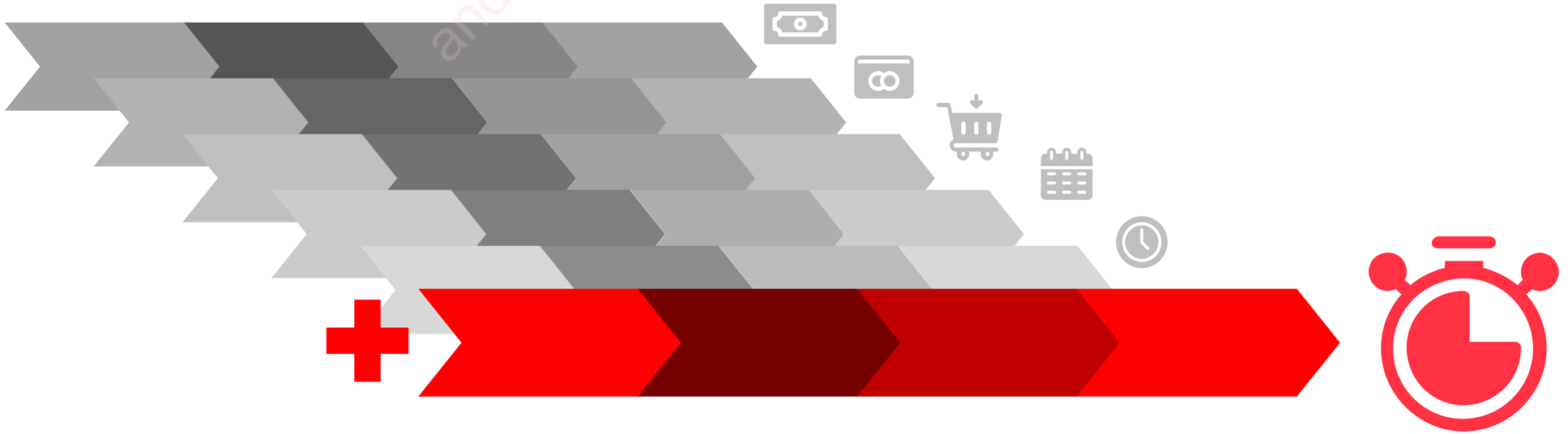
Features as channels



Time as pixel position

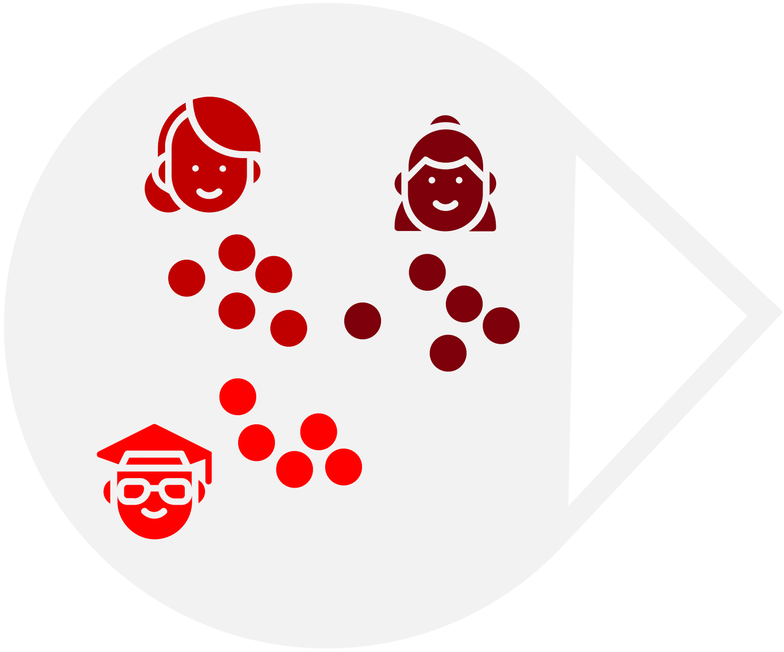
Focusing on the transaction intervals was a better solution



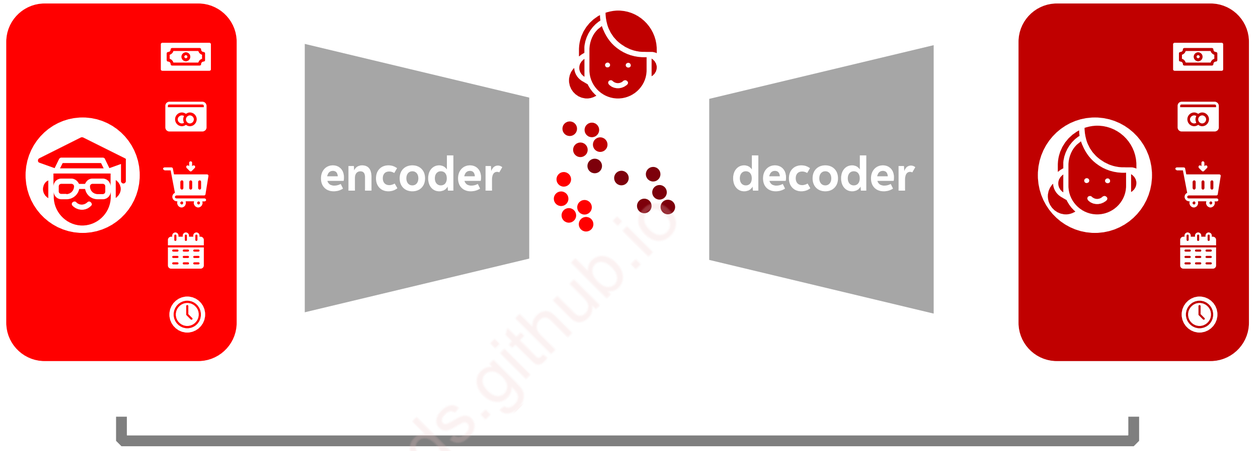


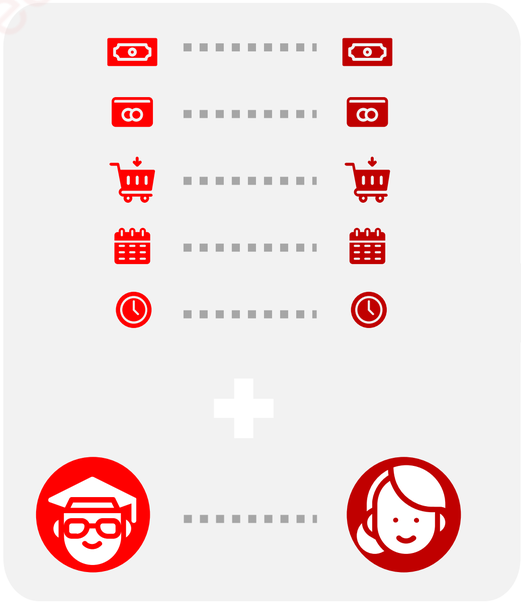
Data cleaning and Pre-Processing only possible with GPU help

RAPIDS



clustering

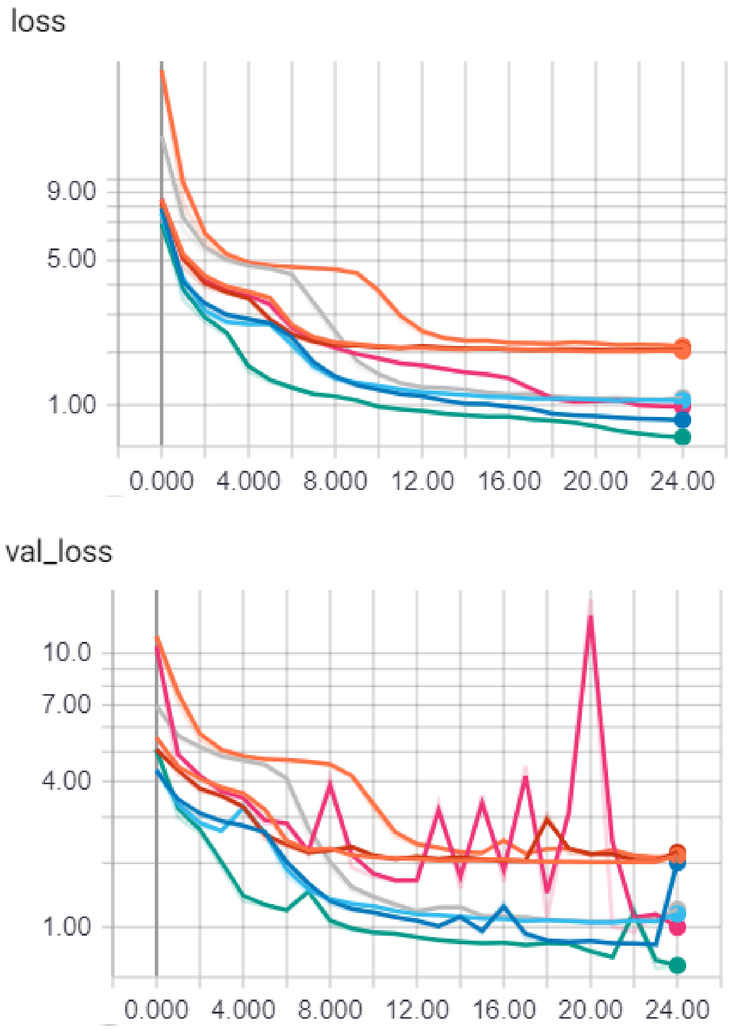




loss

rough clustering
to help initialize the training

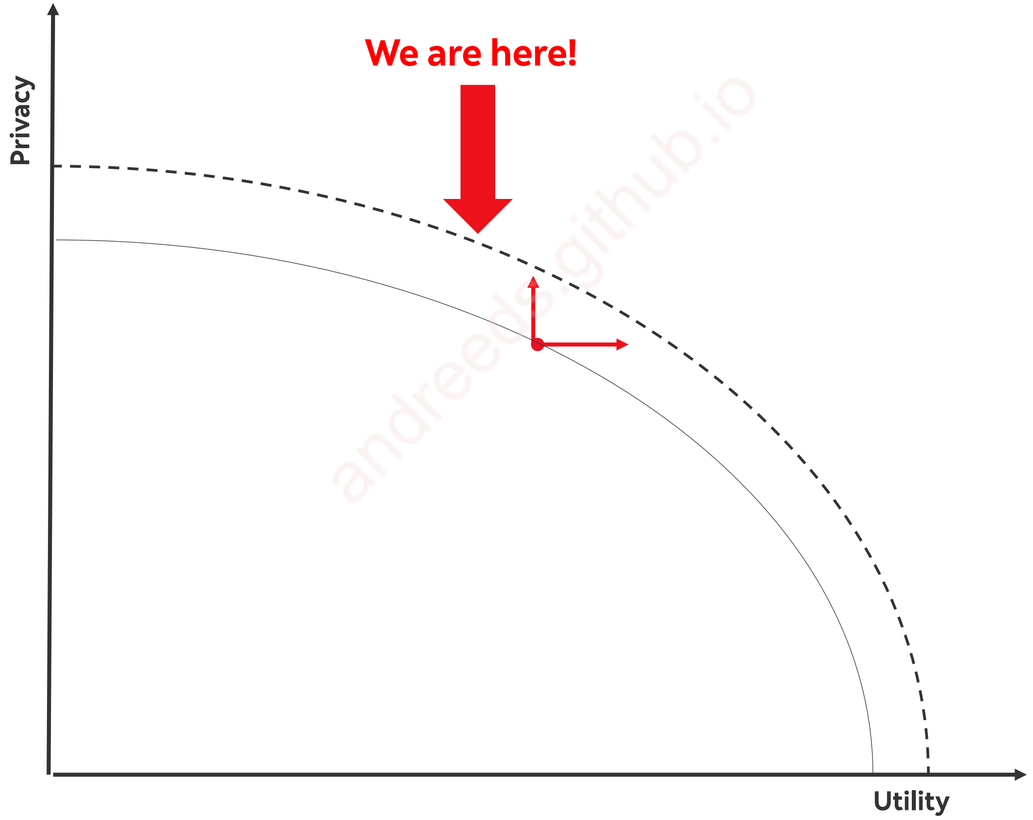
Training went like...



VAE variations

Users clusters

How well is that multi-objective optimization going?



Thank You!

Questions? Suggestions?

Jessie.Lamontagne1@Scotiabank.com

UNIVERSITY OF ALBERTA

Scotiabank®