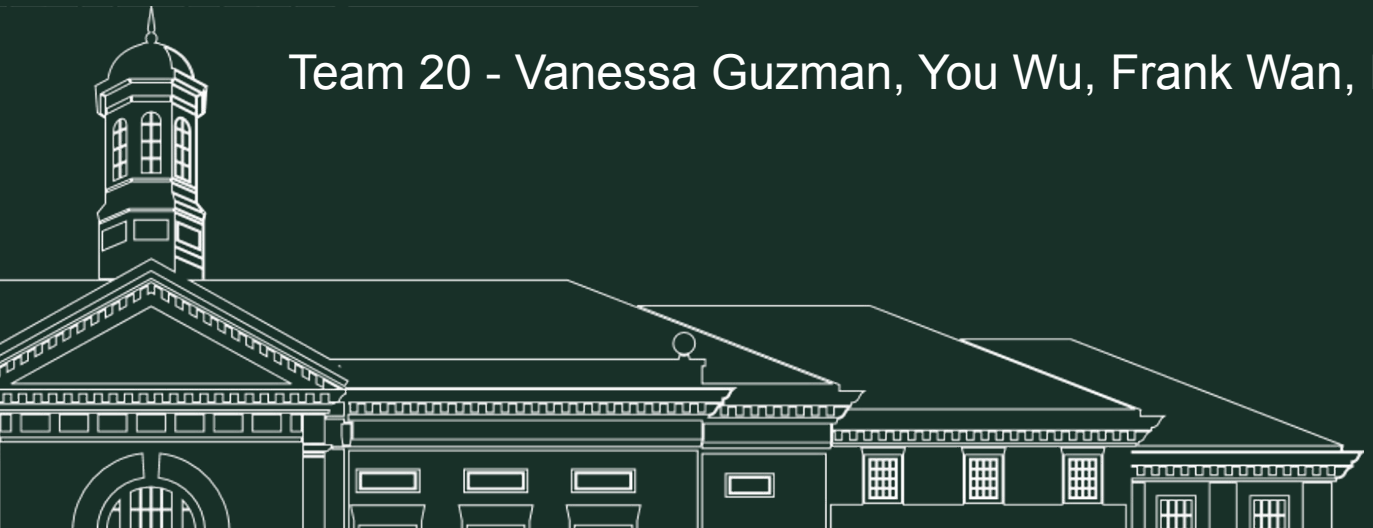
Raymond A. Mason
School of Business
WILLIAM & MARY

JP Morgan AI Research

Team 20 - Vanessa Guzman, You Wu, Frank Wan, Elie Baaklini



Agenda

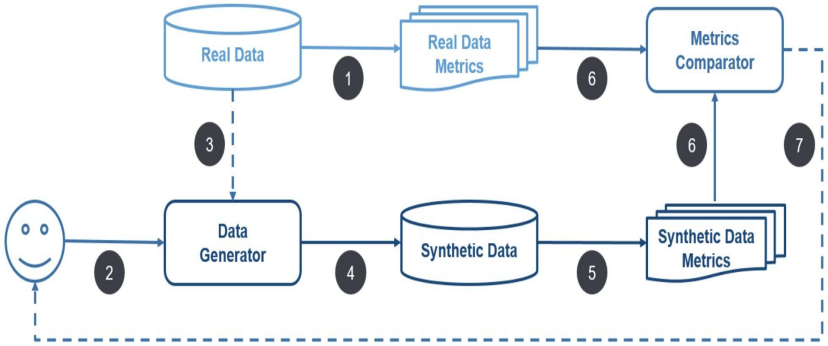
1. JP Morgan AI Research-Create Synthetic Data
    1.1. Background of Synthetic Data Generation and Fraud Detection
    1.2. Synthetic Data Generation Process in the Field of Fraud Detection
2. Exercise 12-9 from Textbook
3. Kahoot

What is Synthetic Data?

Motivation of Creating Synthetic Data

- Internal data use restrictions
- Lack of historical data
- Tackling class imbalance
- Training advanced Machine Learning model
- Data sharing

Process of Synthetic Data Generation



Creating Synthetic Data

Fraud

Imbalanced transaction dataset

Detection

Fraudulent activities constitute a small percentage of all activities

Background

Challenging for an ML model to learn from this type of dataset to detect new occasions of fraud

Therefore, JPM decided to create synthetic data to detect fraud

Sample Payments Data for Fraud Detection

Transaction_Id	Sender_Id	Sender_Account	Sender_Country	Sender_Sector	Sender_lob	Bene_Id	Bene_Account	Bene_Country	USD_Amount	label	Transaction_Type
PAY-BILL-3589	CLIENT-3566	ACCOUNT-3578	USA	21264	CCB	COMPANY-3574	ACCOUNT-3587	GERMANY	492.67	0	MAKE-PAYMENT
WITHDRAWAL-3591	CLIENT-3566	ACCOUNT-3579	USA	18885	CCB				388.92	0	WITHDRAWAL
MOVE-FUNDS-3528	CLIENT-3508	ACCOUNT-3520	USA	4809	CCB	COMPANY-3516	ACCOUNT-3527	GERMANY	280.7	0	MOVE-FUNDS
WITHDRAWAL-3529	CLIENT-3508	ACCOUNT-3519	USA	7455	CCB				118.14	0	WITHDRAWAL
QUICK-DEPOSIT-3471						CLIENT-3442	ACCOUNT-3461	USA	105.16	0	DEPOSIT-CASH
QUICK-DEPOSIT-3473						CLIENT-3442	ACCOUNT-3460	USA	164.97	0	DEPOSIT-CASH
PAY-BILL-3404	CLIENT-3384	ACCOUNT-3395	USA	36316	CCB	COMPANY-3392	ACCOUNT-3401	GERMANY	456.89	0	MAKE-PAYMENT
QUICK-DEPOSIT-3406						CLIENT-3384	ACCOUNT-3396	USA	413.17	0	DEPOSIT-CASH
PAY-CHECK-3347	CLIENT-3330	ACCOUNT-3341	USA	36194	CCB	CLIENT-3333	ACCOUNT-3338	CANADA	377.65	0	PAY-CHECK
PAY-CHECK-3348	CLIENT-3330	ACCOUNT-3340	USA	20626	CCB	CLIENT-3333	ACCOUNT-3338	CANADA	338.03	0	PAY-CHECK
MOVE-FUNDS-3292	CLIENT-3272	ACCOUNT-3284	USA	21568	CCB	CLIENT-3275	ACCOUNT-3291	CANADA	100.85	0	MOVE-FUNDS
MOVE-FUNDS-3294	CLIENT-3272	ACCOUNT-3284	USA	29040	CCB	CLIENT-3273	ACCOUNT-3289	USA	276.66	0	MOVE-FUNDS
PAY-BILL-3232	CLIENT-3203	ACCOUNT-3222	USA	27393	CCB	COMPANY-3210	ACCOUNT-3218	GERMANY	234.88	0	MAKE-PAYMENT
QUICK-DEPOSIT-3234						CLIENT-3203	ACCOUNT-3222	USA	945.22	0	DEPOSIT-CASH
DEPOSIT-CASH-3163						CLIENT-3139	ACCOUNT-3154	USA	655.09	0	DEPOSIT-CASH
PAY-BILL-3162	CLIENT-3139	ACCOUNT-3153	USA	25066	CCB	COMPANY-3147	ACCOUNT-3160	GERMANY	675.37	0	MAKE-PAYMENT
WITHDRAWAL-3100	CLIENT-3075	ACCOUNT-3090	USA	22778	CCB				319.95	0	EXCHANGE
QUICK-PAYMENT-3099	CLIENT-3075	ACCOUNT-3091	USA	39013	CCB	CLIENT-3078	ACCOUNT-3087	TAIWAN	771.54	0	QUICK-PAYMENT
PAY-BILL-3036	CLIENT-3016	ACCOUNT-3028	USA	43951	CCB	COMPANY-3022	ACCOUNT-3033	GERMANY	730.69	0	MAKE-PAYMENT

Techniques for Fraud Detection Synthetic Data Generation

CABBOT

Classification of Agents' Behavior Based on Observation Traces

a learning technique that allows JPM to perform on-line classification of the type of planning agent whose behavior is observing.

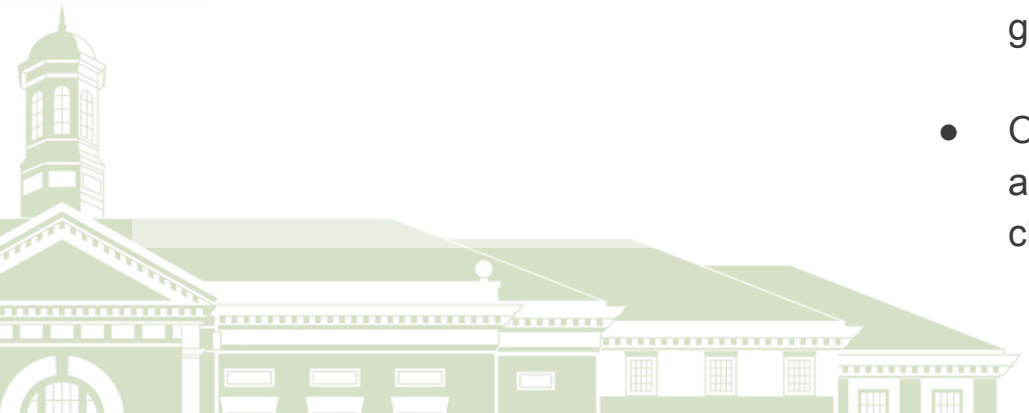
JP Morgan AI Research-Create Synthetic Data

Background

- Make assumptions on agents' rational behavior
- Use the automated planning formalism to describe the tasks

Learning to Classify Behavior

- Given: (1) a set of classes of behavior (labels) $C = \{C1, C2, \ldots, Cn\}$; (2) a set of labeled observed traces $T_{B,Ci}$, $\forall Ci \in C$; and (3) a partially observable domain model of each Ci given by $\Pi_{B,Ci}$
- Obtain: a classifier that takes as input a new (partial) trace t (with unknown class) and outputs the predicted class



Generation of Synthetic Behavior

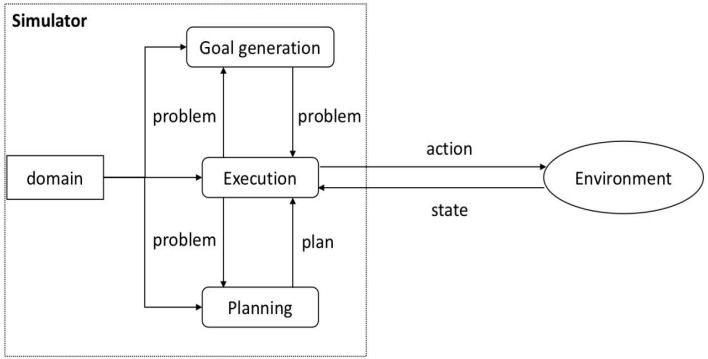


Figure 1: High level view of the simulator.

The components of the simulator for the planning agents are:

- Execution
- Planning
- Goal generation

JP Morgan AI Research

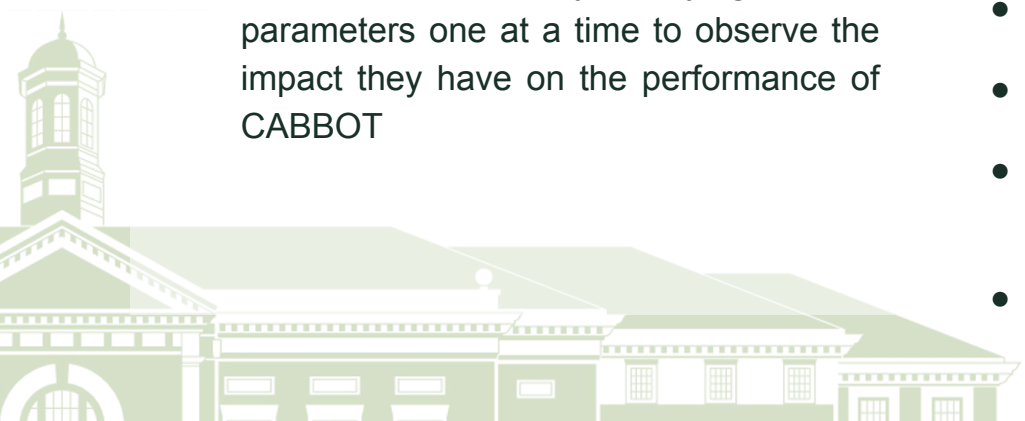
Experiments

- Defined or use the domain in the planning community
- Randomly generated traces of each type of behavior for training and some for test
- Present results by varying these parameters one at a time to observe the impact they have on the performance of CABBOT

Results

Classification accuracy in the fraud detection varying:

- the probability of appearing goals for the two kinds of customers, fraud or non-fraud
- the length of the trace
- the similarity function
- the probability of partial observability individual action execution failure
- the similarity function in several domains



References

1. Generating Synthetic Data in Finance: Opportunities, challenges and pitfalls. S Assefa, D Dervovic, M Mahfouz, R Tillman, P Reddy, T Balch and M Veloso. Proceedings of the 1st International Conference on AI in Finance (ICAIF), 2020. Also in https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3634235

2. Domain-independent generation and classification of behavior traces. D Borrajo and M Veloso. https://arxiv.org/abs/2011.02918

3. 4 Synthetic data applications to enable finance innovation in ’22 https://research.aimultiple.com/synthetic-data-finance/