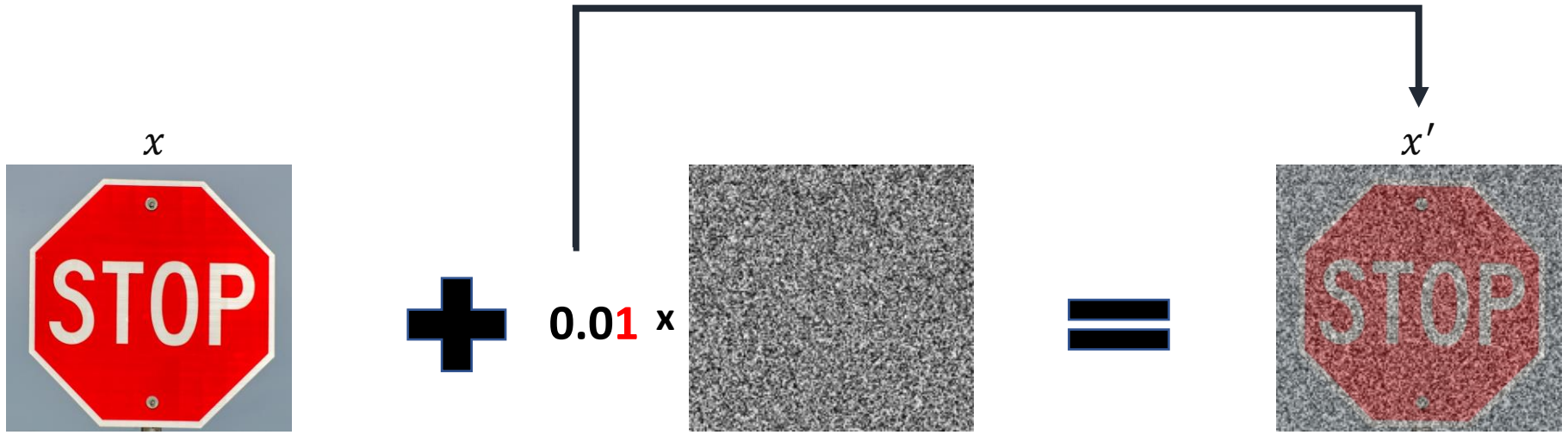
DEEP GENERATIVE MODELS AS AN ADVERSARIAL ATTACK STRATEGY FOR TABULAR MACHINE LEARNING

Salijona Dyrmishi[1], Mihaela Cătălina Stoian[2], Eleonora Giunchiglia [3], Maxime Cordy[1]

[1]University of Luxembourg, [2]University of Oxford, [3]Imperial College London




Adversarial attacks against ML



“Stop”

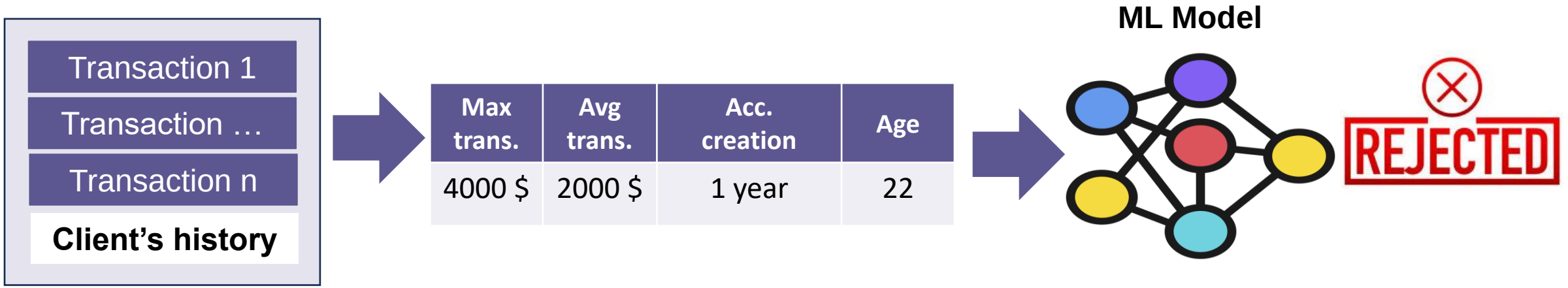
Carefully crafted adversarial perturbations

“Go”

Adversary’s strategy

minimize $||x - x'||$
subject to $h(x) \neq h(x')$
where h is the target model

Tabular ML models are prone to adversarial attacks too



Tabular ML models are prone to adversarial attacks too

Problem space

Transaction 1

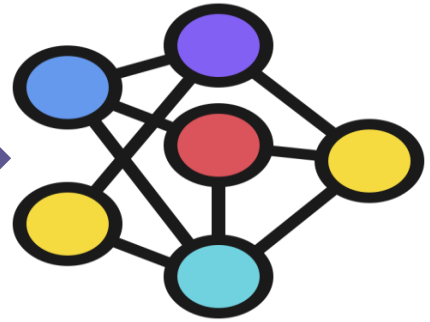
Transaction …

Transaction n

Client's history

Max trans.	Avg trans.	Acc. creation	Age
4000 $	2000 $	1 year	22

ML Model

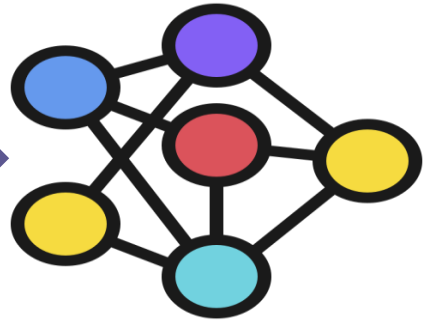


Tabular ML models are prone to adversarial attacks too

Feature space

Max trans.	Avg trans.	Acc. creation	Age
4000 $	2000 $	1 year	22

ML Model



Domain properties

$$acc\_creation < age$$

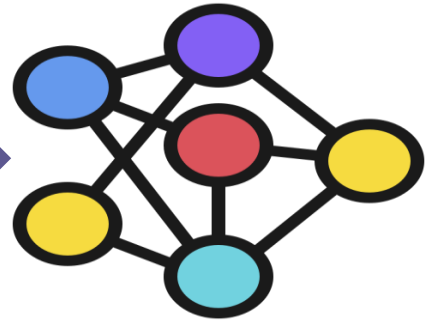
$$max\_transaction > avg\_transaction$$

Tabular ML models are prone to adversarial attacks too

Feature space

Max trans.	Avg trans.	Acc. creation	Age
4000 $	2000 $	**23 years**	22

ML Model



Domain properties

$$acc\_creation < age$$

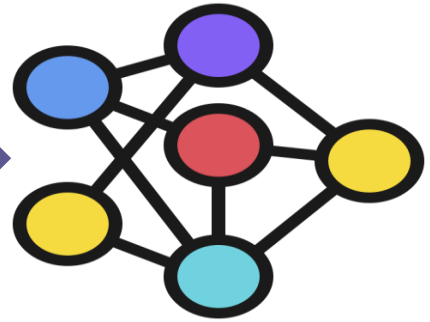
$$max\_transaction > avg\_transaction$$

Tabular ML models are prone to adversarial attacks too

Feature space

Max trans.	Avg trans.	Acc. creation	Age
4000 $	2000 $	**23 years**	22

ML Model



Unrealistic adversarial example

Domain properties

$$acc\_creation < age$$

$$max\_transaction > avg\_transaction$$

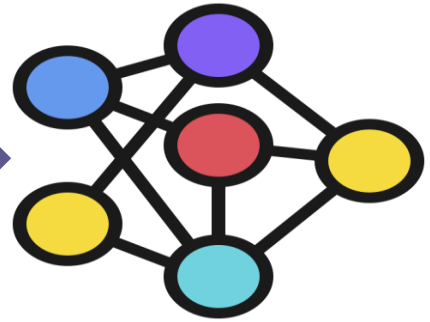
Tabular ML models are prone to adversarial attacks too

Realistic adversarial example

Feature space

Max trans.	Avg trans.	Acc. creation	Age
4000 $	**2020 $**	1 year	22

ML Model

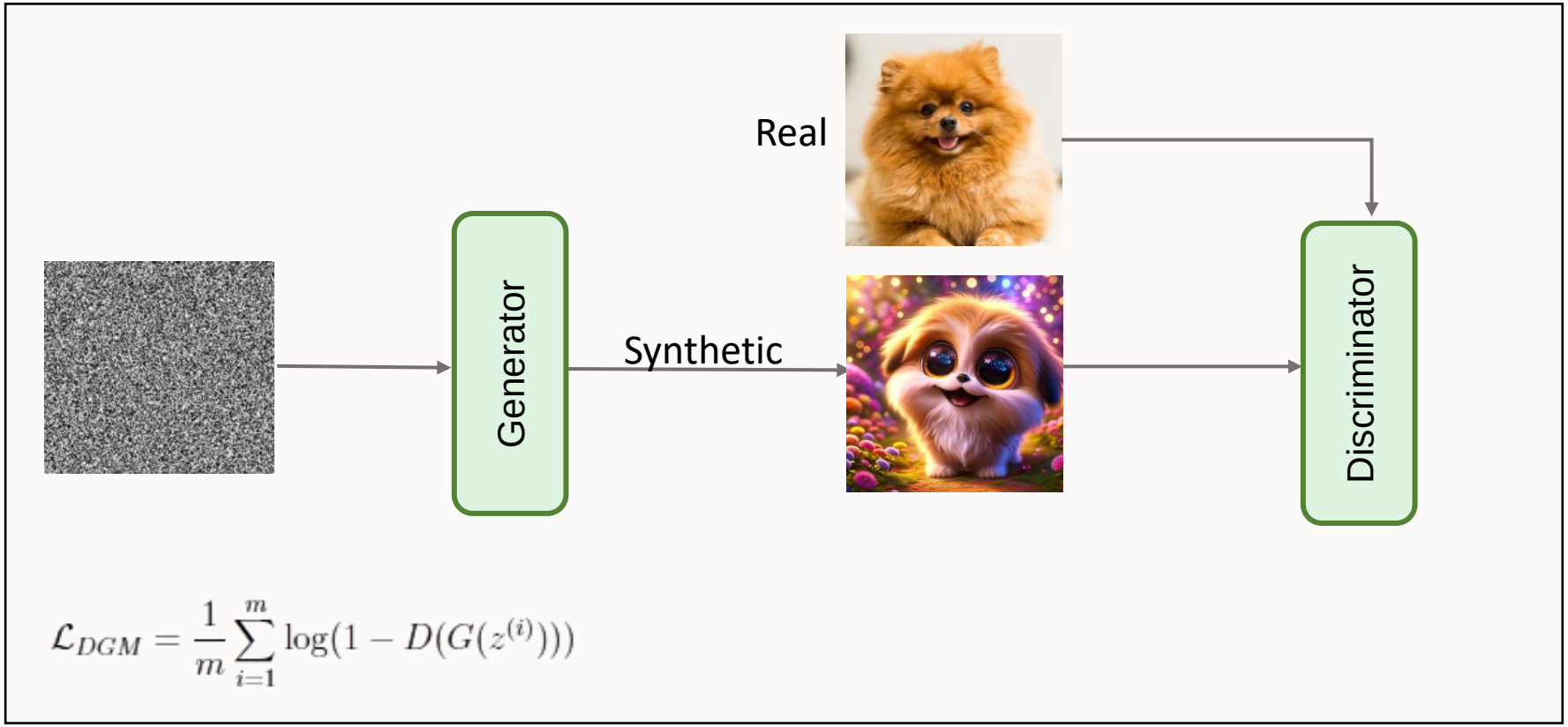


Domain properties

$$acc\_creation < age$$

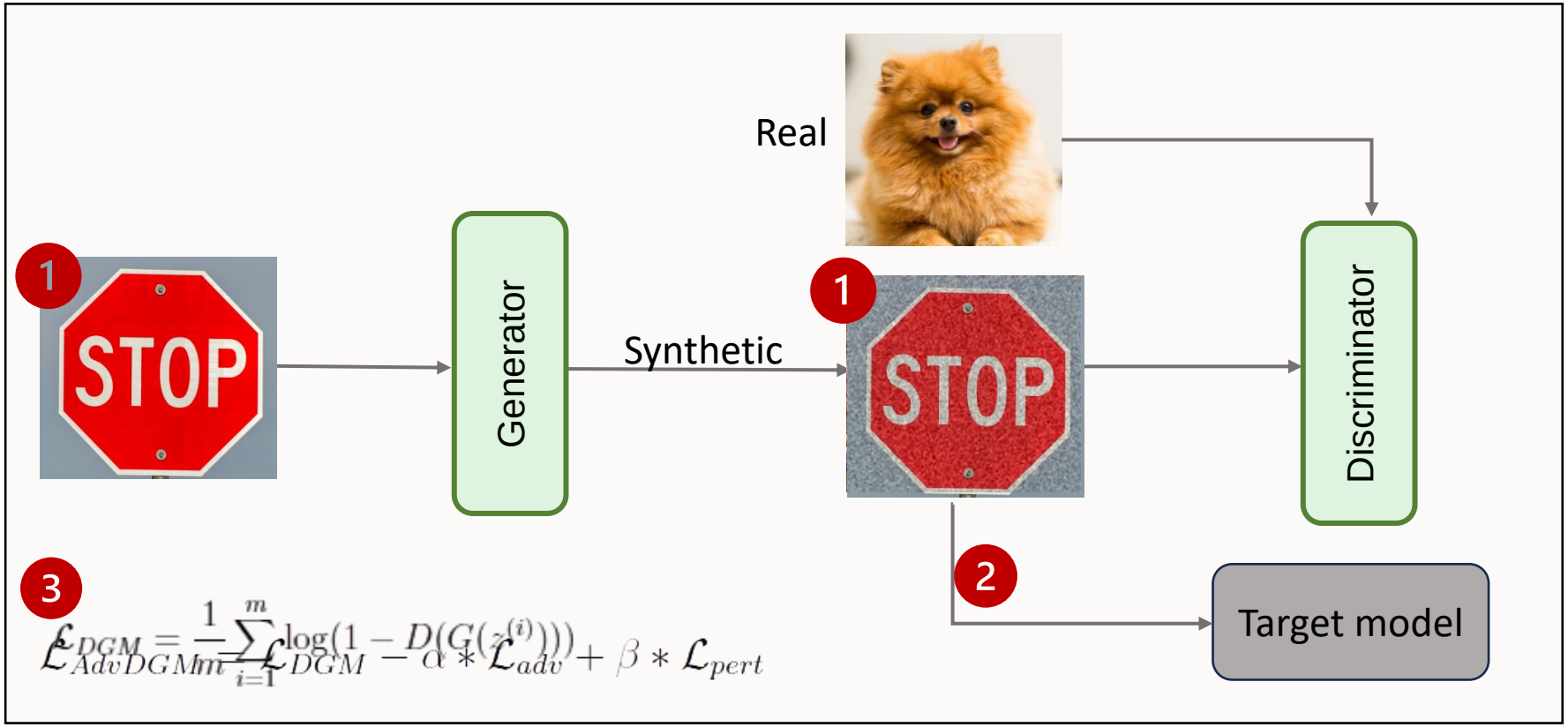
$$\boldsymbol{max\_transaction > avg\_transaction}$$

Deep Generative Models (DGM)



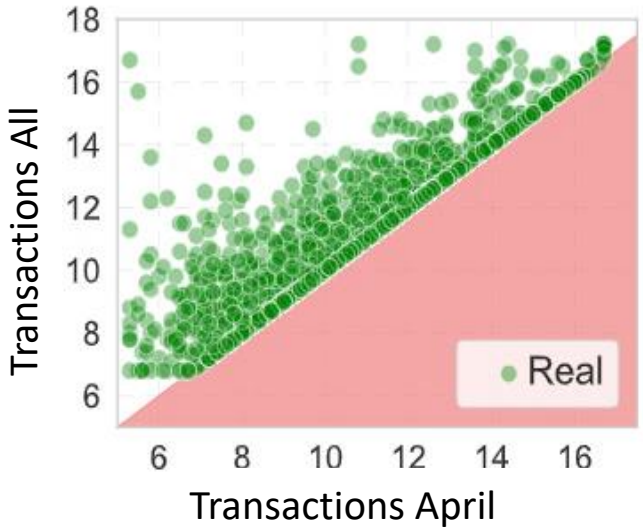
$$\mathcal{L}_{DGM} = \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(z^{(i)})))$$

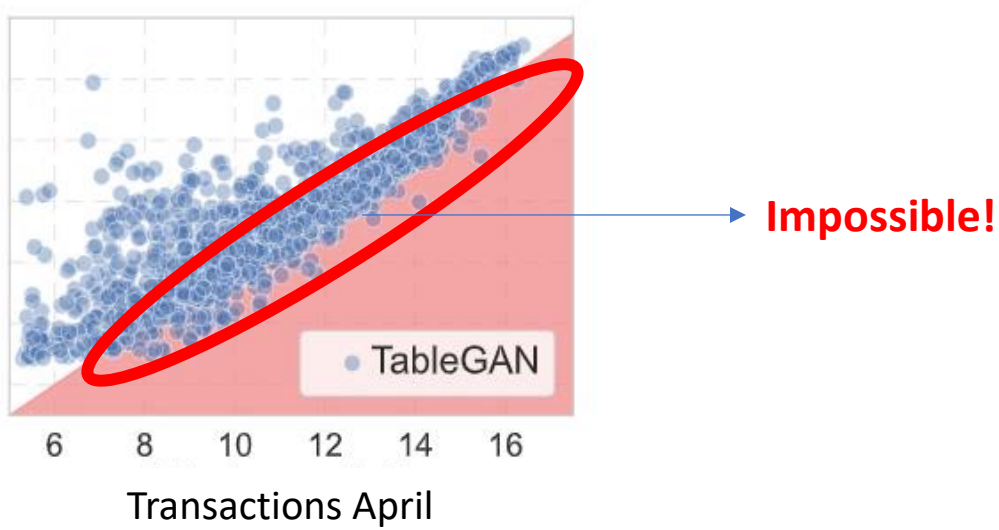
Adversarial Deep Generative Models (AdvDGM)



Xiao, Chaowei, et al. *Generating adversarial examples with adversarial networks,* IJCAI 2018

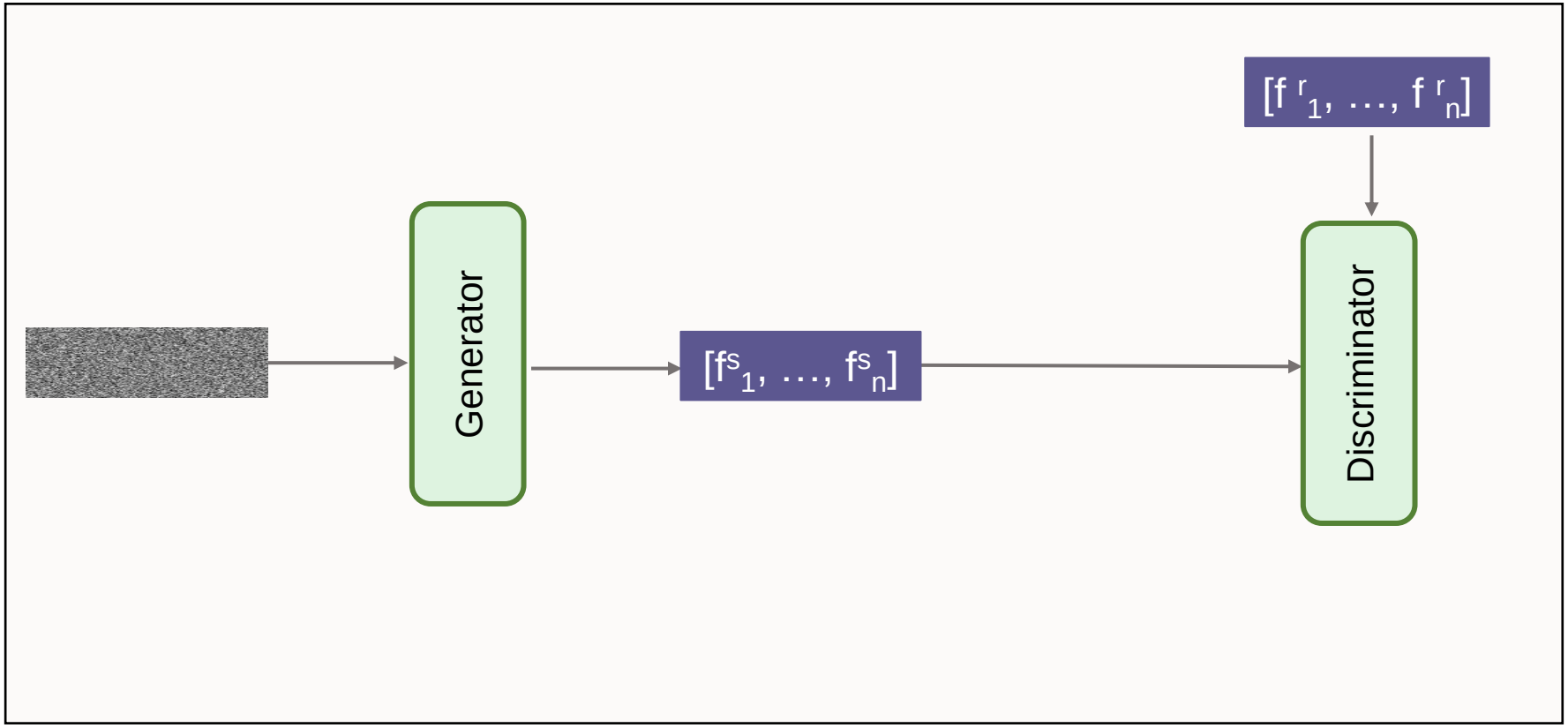
Tabular DGM Failures





Domain constraint: *Nr. Transactions all > Nr. Transactions past month*

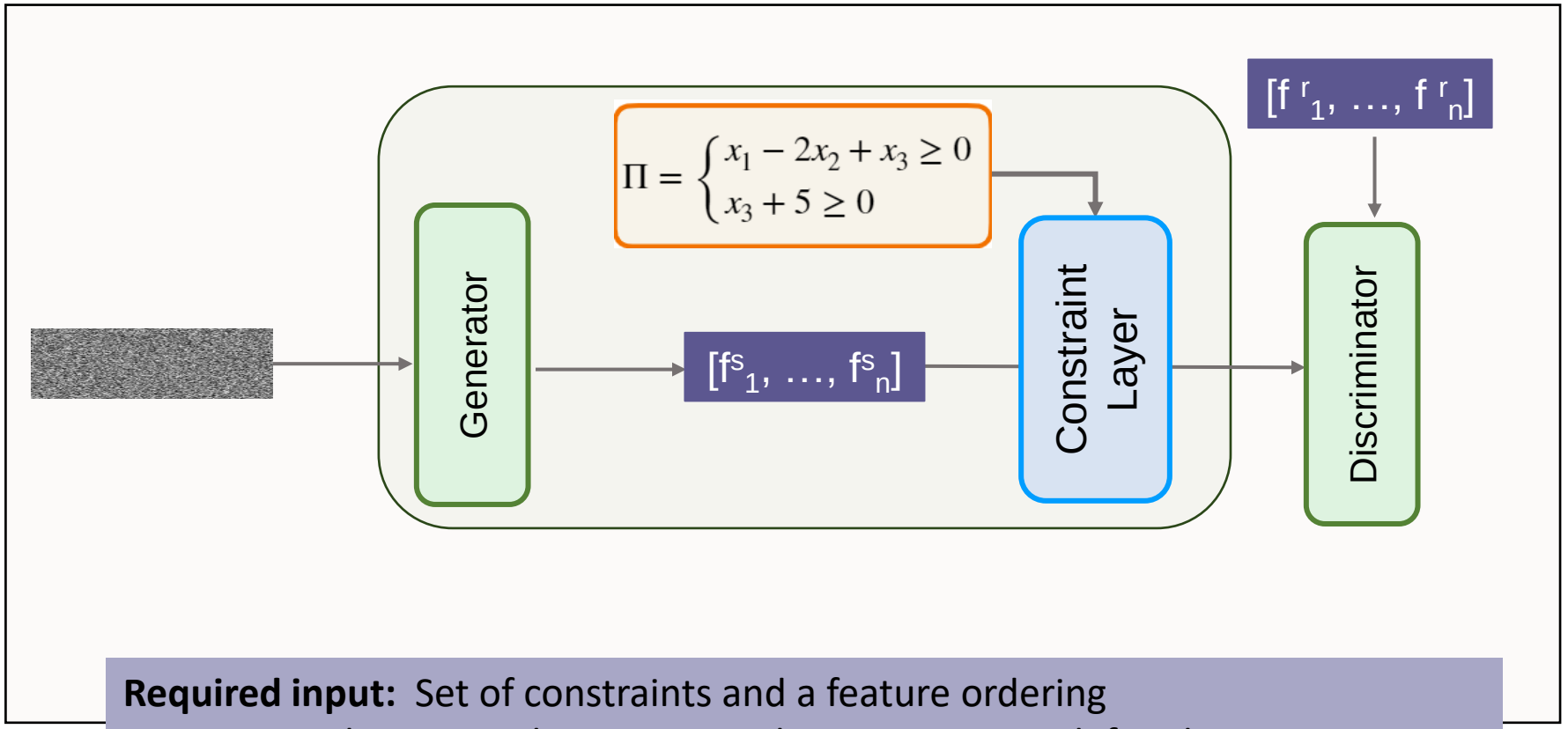
Constrained Deep Generative Models (C-DGM)



Constrained Deep Generative Models (C-DGM)

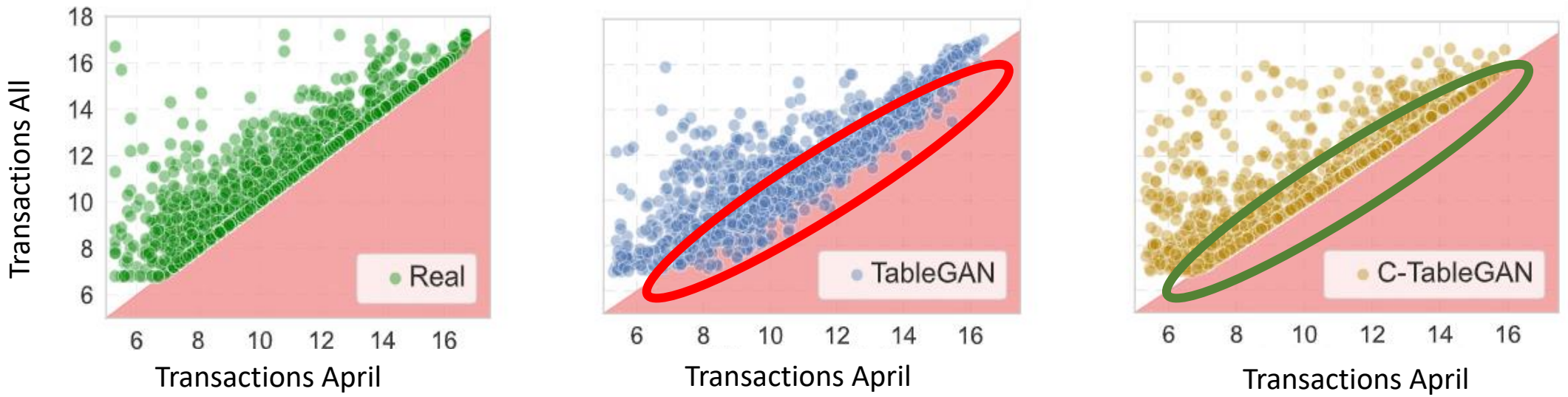
C-DGM when applied during training

P-DGM when applied during sampling



Required input: Set of constraints and a feature ordering
Output: Synthetic samples guaranteed to respect user defined constraints

Constrained Deep Generative Models (C-DGM)



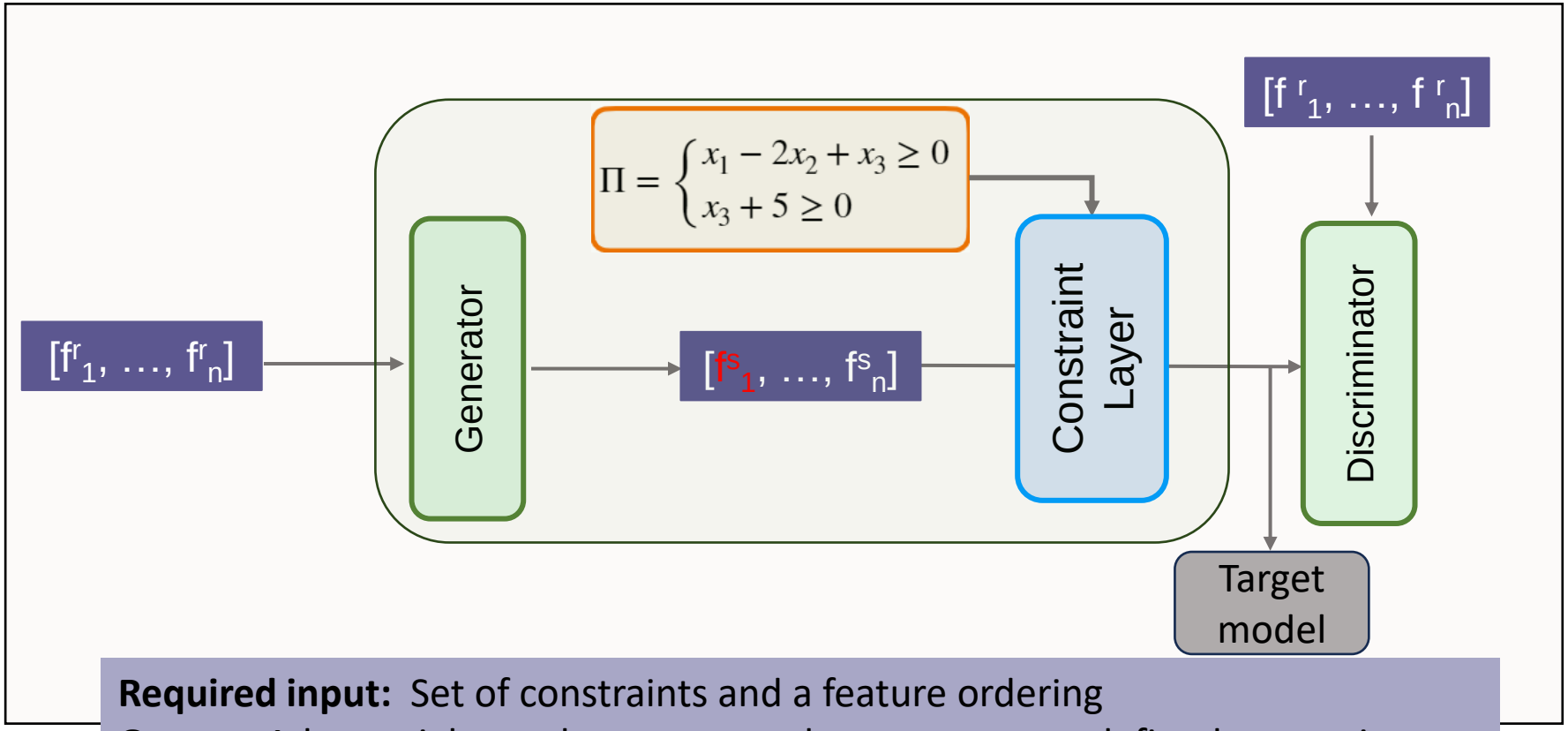
Model/Dataset	URL	WiDS	LCLD	Heloc	FSP
WGAN	11.1±1.6	98.2±0.2	100.0±0.0	57.0±13.0	70.7±8.3
TableGAN	4.9±1.4	96.4±2.4	6.1±0.9	45.6±16.3	71.6±8.7
CTGAN	3.1±2.6	99.9±0.0	11.8±2.7	41.6±12.1	74.3±5.2
TVAE	3.0±0.7	99.9±0.0	3.9±0.5	55.5±1.4	66.4±3.0
GOGGLE	5.9±6.6	78.2±11.6	13.1±2.9	47.3±7.0	63.7±17.6
All C-models	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0±0.0**	**0.0 ±0.0**

Table 1. Constraint Violation Rate: Percentage of generated samples violating at least one constraint in the set of linear constraints.

Constrained Adversarial Deep Generative Models (C-AdvDGM)

C-AdvDGM when applied during training

P-AdvDGM when applied during sampling



Required input: Set of constraints and a feature ordering
Output: Adversarial samples guaranteed to respect user defined constraints

Adversarial generation capability

Insights:

1. Adding the constrained layer during the training (C-AdvDGM) or sampling (P-AdvDGM) increases the performance of the attack in equal number of times
2. Except WGAN and its constrained counterparts, other models are not increasing the error rate of the model

ASR↑

Attack/Dataset	URL	WiDS	Heloc	FSP
-	0.04	0.19	0.28	0.24
AdvWGAN	**0.73**±0.10	0.03±0.00	0.31±0.16	0.30±0.19
P-AdvWGAN	**0.73**±0.10	0.07±0.08	**0.93**±0.04	0.70±0.04
C-AdvWGAN	0.52±0.16	**0.17**±0.14	0.46±0.33	**0.73**±0.06
AdvTableGAN	**0.14**±0.08	0.03±0.00	0.15±0.04	0.08±0.03
P-AdvTableGAN	**0.14**±0.08	**0.17**±0.01	**0.28**±0.02	**0.28**±0.03
C-AdvTableGAN	0.09±0.01	0.12±0.02	0.09±0.19	0.27±0.01
AdvCTGAN	0.01±0.00	0.01±0.01	0.18±0.03	0.02±0.03
P-AdvCTGAN	0.01±0.00	**0.19**±0.11	0.28±0.01	0.06±0.08
C-AdvCTGAN	**0.02**±0.00	0.16±0.01	**0.37**±0.06	**0.32**±0.02
AdvTVAE	0.00±0.00	0.00±0.00	0.18±0.01	0.06±0.02
P-AdvTVAE	0.00±0.00	**0.12**±0.01	0.32±0.02	0.23±0.01
C-AdvTVAE	**0.01**±0.00	0.10±0.00	**0.60**±0.04	**0.28**±0.01

Impact on runtime

Insight:

C-AdvDGMs are at most 2.7 times slower during training and 1.3 times during sampling compared to AdvDGMs.

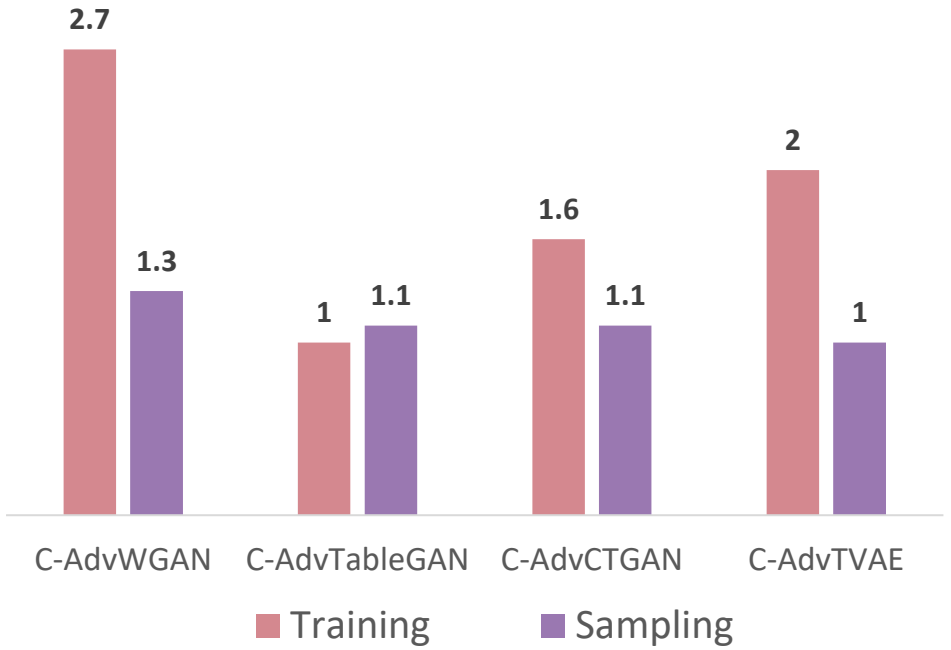


Fig 12. Relative run time compared to unconstrained models averaged over four dataset

Takeaway

Adding domain knowledge to DGMs brings improvements for adversarial attacks. However, the effect of the layer during training or sampling needs further investigation.

Thank you!