Work-in-progress: CANGen: Practical Synthetic CAN Traces Generation using Deep Generative Models

Yucheng Yin, CMU
Jorge Guajardo Merchan, Bosch Research
Pradeep Pappachan, Bosch Research
Vyas Sekar, CMU

July 8th, 2024

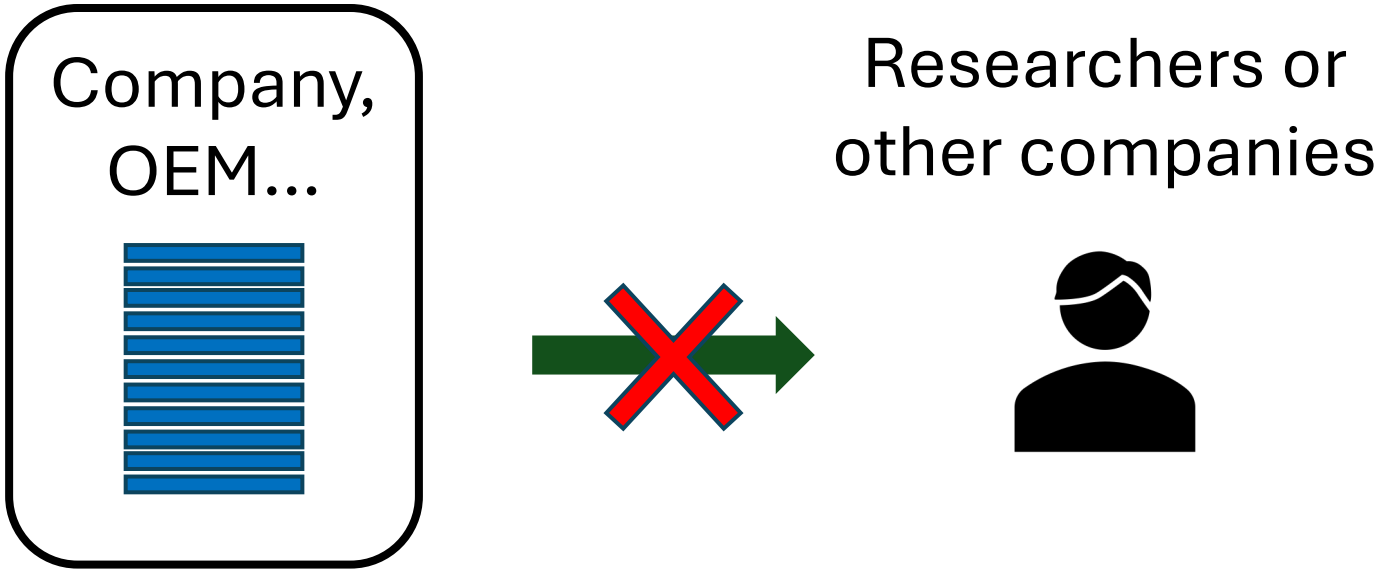
CAN traces enable many applications

Anomaly detection

Simulation and testing

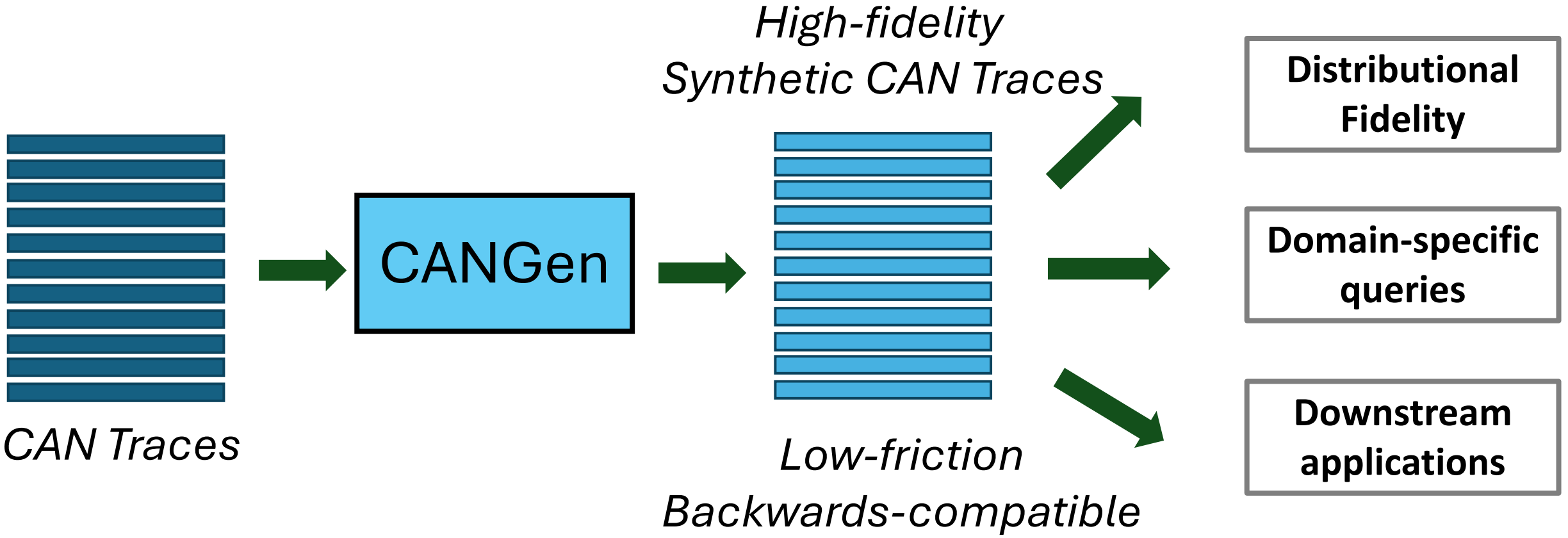
Smart transportation

Pain point: lack of real and diverse CAN traces



Bad outcomes:
Research is not reproducible,
Collaboration go untapped…

Our Vision: CANGen



Deep Generative Models (DGMs) are unprecedentedly popular and powerful

- GANs, Transformers, Diffusion models…
- Text, images, audio, video…

ChatGPT

Midjourney

Pika

Off-the-shelf DGMs are not applicable

- **Data generation**: model training inefficiency

	OpenXC [3]	SynCAN [20]	Car-hacking [41]
CTGAN [52]	✗	✗	✗
NetShare [54]	✗	✗	✗
RTF-Tab [42]	✓	✗	✓
RTF-Time [42]	✗	✗	✗
TabDDPM [25]	✗	✗	✗

”X” denotes inability to generate

- **Data evaluation**: Lack of systematic fidelity metrics

Key insight #1: adaptive preprocessing

- Different representations of CAN traces


```
{"name":"brake_pedal_status","value":true,"timestamp":1364310855.004000}
{"name":"transmission_gear_position","value":"first","timestamp":1364310855.004000}
{"name":"accelerator_pedal_position","value":0,"timestamp":1364323939.012000}
{"name":"engine_speed","value":772,"timestamp":1364323939.027000}
```


Time	ID	Signal1	Signal2	Signal3	Signal4
2088.41338746	id5	0.0	0.9587	-	-
2089.55410634	id8	0.2468	-	-	-
2119.05278128	id10	0.4545	0.1111	0.9478	0.1704
2133.51200647	id2	0.0	0.0	0.2426	-

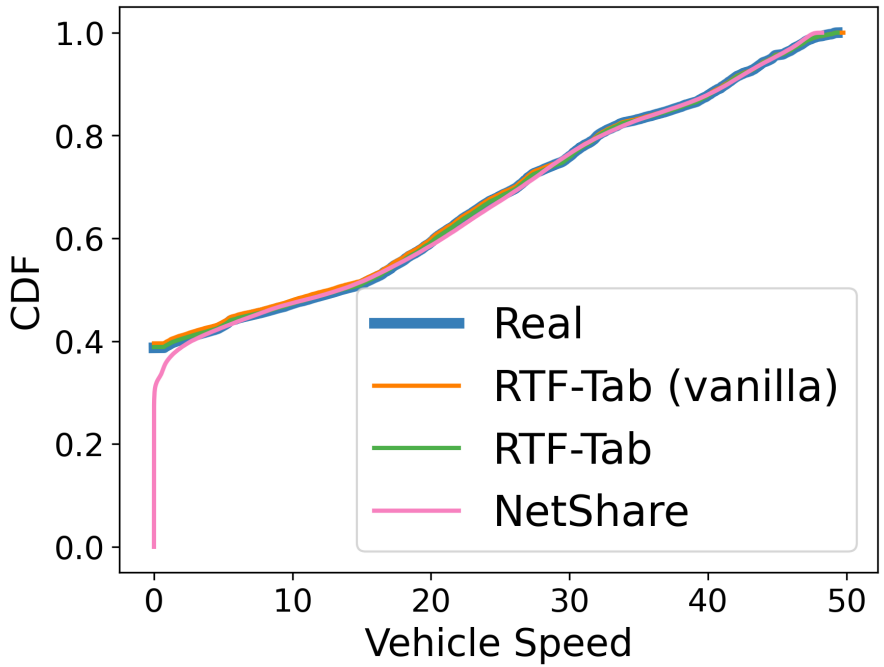
Timestamp	CAN ID	DLC	Data
1478198376.389427	0316	8	052168092121006f
1478198376.389636	018f	8	fe5B0000003C0000
1478198376.389864	0260	8	19212230088e6d3a
1478198376.409484	05f0	2	0100

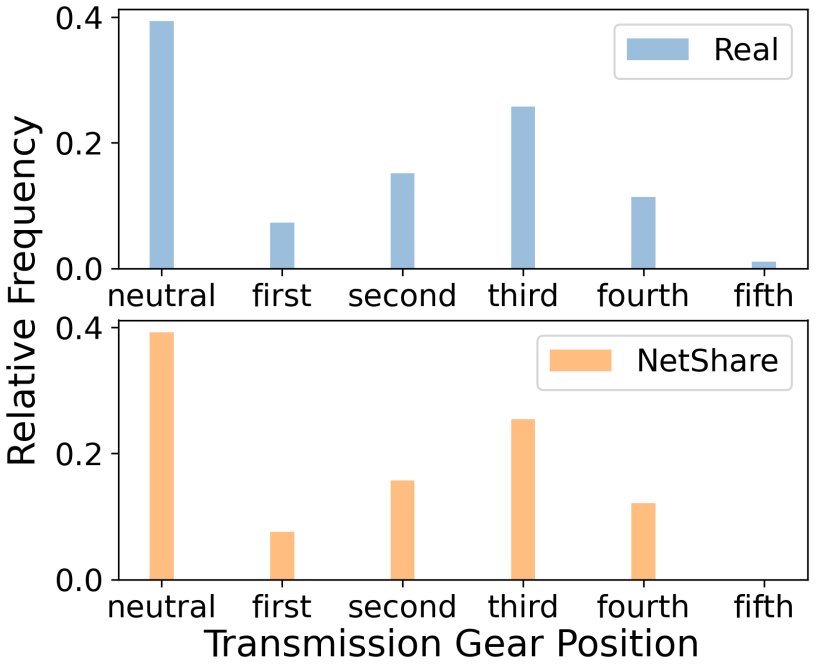
Decoded Signal-based:
Forward fill missing values

Raw Signal-based:
Add missing signal indicator

Byte-based:
Convert hex strings to bits

Adaptive preprocessing enables efficient learning of DGMs





Recall that most of the models are not even able to generate data before applying adaptive preprocessing!

Key insight #2: a suite of fidelity metrics

- Distributional metrics
  - Statistical similarity between real and synthetic data
- Domain-specific queries
  - E.g., do the engine speed and vehicle speed match?
- Use-case related
  - Accuracy preservation
  - Rank preservation

Implementation and Evaluation

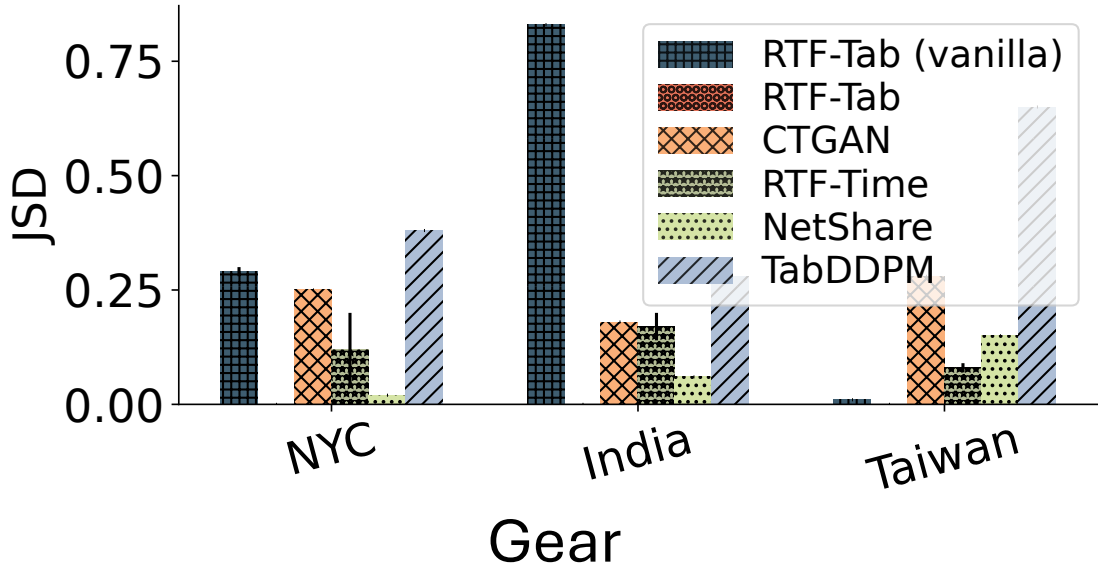
- Datasets

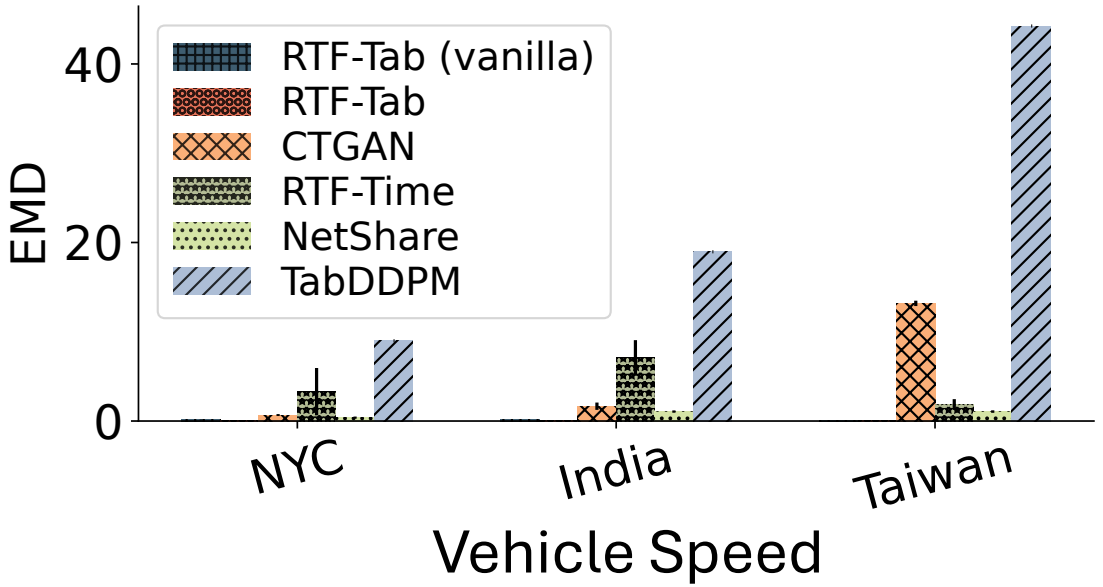
Dataset	Year	Type	Attack/Normal
OpenXC [3]	2013	Signal (decoded)	Normal
SynCAN [20]	2019	Byte	Normal, Attack
Car-Hacking [41]	2018	Signal (raw)	Normal, Attack

- DGMs
  - RTF-Tab (vanilla), RTF-Tab, CTGAN, RTF-Time, NetShare, TabDDPM
  - Covering GANs, diffusion models, transformers

- CANGen is open-source at https://github.com/netsharecmu/CANGen

Evaluation – distributional fidelity





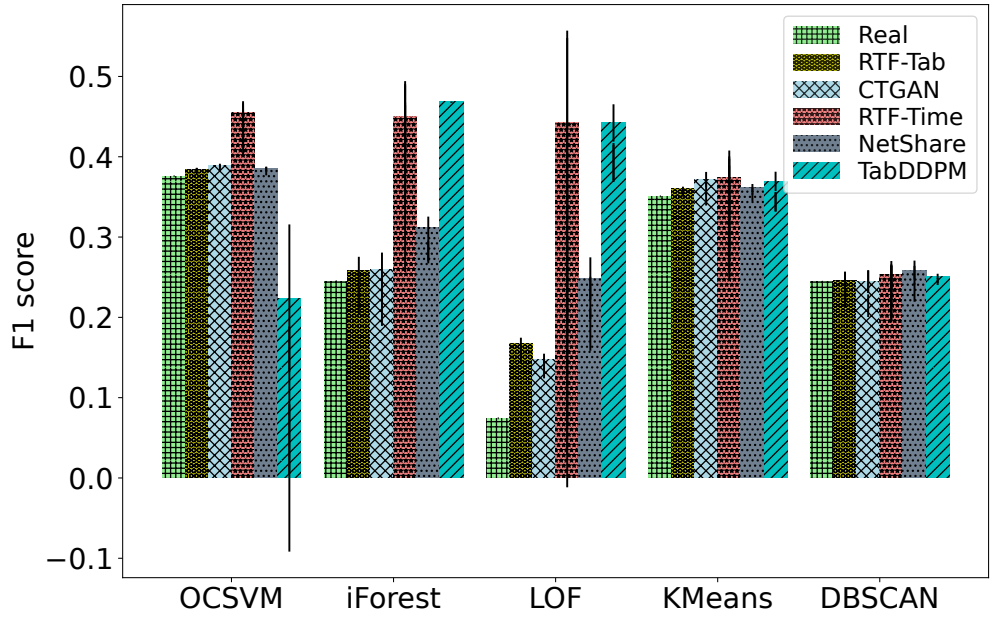
Across decoded-signal datasets,
RTF-Tab with adaptive preprocessing enabled achieves
50% better distributional fidelity on average compared to
the second-best baseline (RTF-Tab (vanilla)).

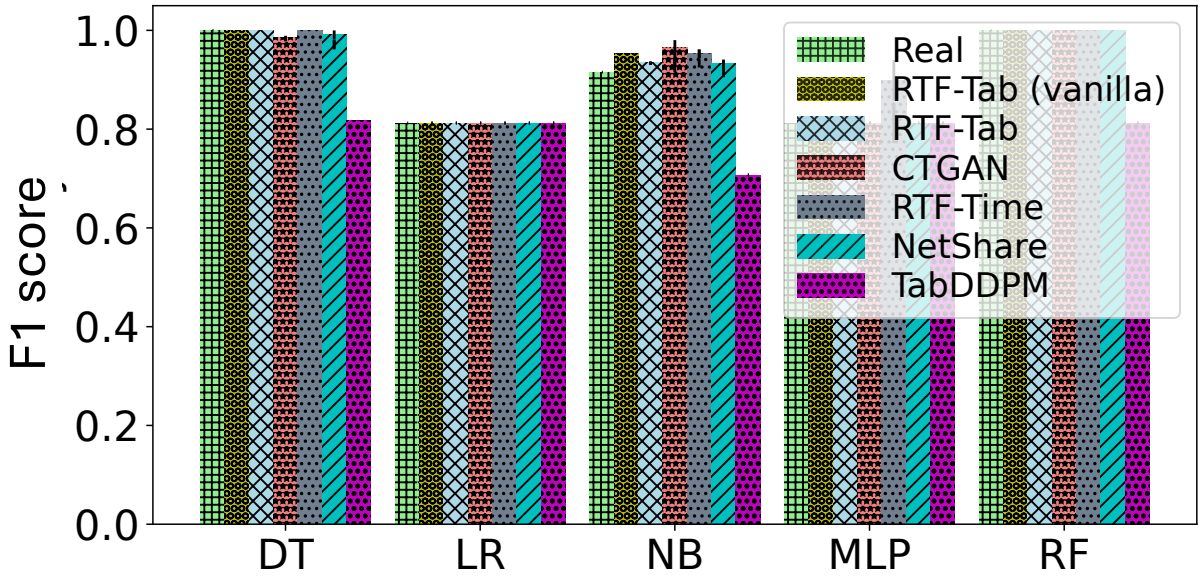
Evaluation – domain specific queries

- Query 1: Is vehicle speed within a reasonable range?
  - Single sensor, distributional
- Query 2: Is gear change reasonable?
  - Single sensor, timeseries
- Query 3: At any given timestamp, do vehicle speed - Engine speed - gear position match?
  - Multi-sensor, distributional
- Query 4: Do vehicle speed and engine speed change simultaneously?
  - Multi-sensor, timeseries

Across decoded-signal datasets, RTF-Tab with adaptive preprocessing enabled achieves the highest average valid ratio (92.8%) on selected queries.

Evaluation – anomaly detection





Across all datasets, RTF-Tab with adaptive preprocessing enabled preserves the accuracy and rankings of different downstream tasks best.

Discussion and Future work

- Protocol-complaint generation
- More comprehensive and realistic evaluation
- Privacy-preserving data sharing

Takeaways

- Pain points due to lack of real and diverse CAN traces
- New opportunity: **synthetic** CAN trace generation by Deep Generative Models (DGMs)
- CANGen: An end-to-end framework that utilizes different state-of-the-art DGMs to ***generate*** and ***evaluate*** different kinds of CAN traces
- Join us and contribute ☺
  - https://github.com/netsharecmu/CANGen
  - https://github.com/netsharecmu
  - https://users.ece.cmu.edu/~vsekar/projects/datafuel/