PrvTel: Lightweight Models for Private and Accurate Telemetry Data Retention

Presenter: Lesley (Yajie) Zhou


Fuheng Zhao


Eric Wang


Ayse Coskun


Divyakant Agrawal


Amr El Abbadi

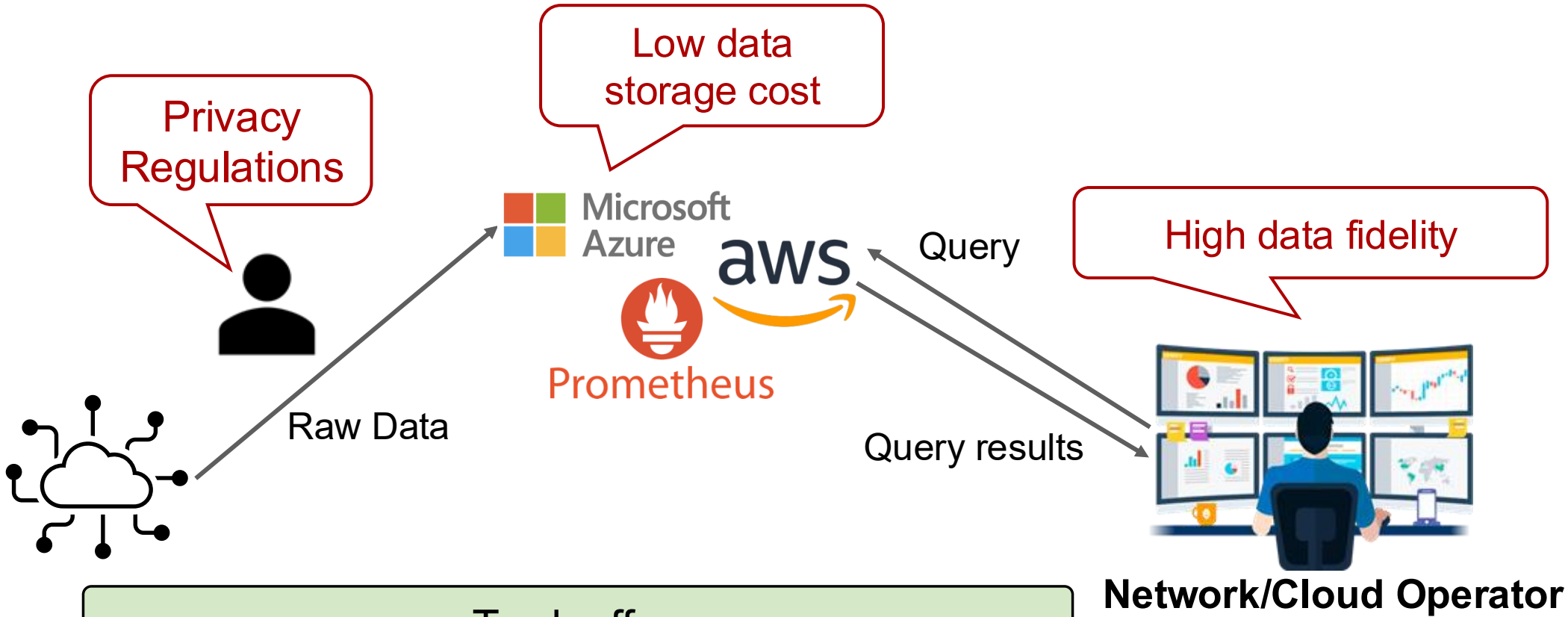

Alan Liu

Telemetry Data is Everywhere

Network Telemetry (flow-level) Example

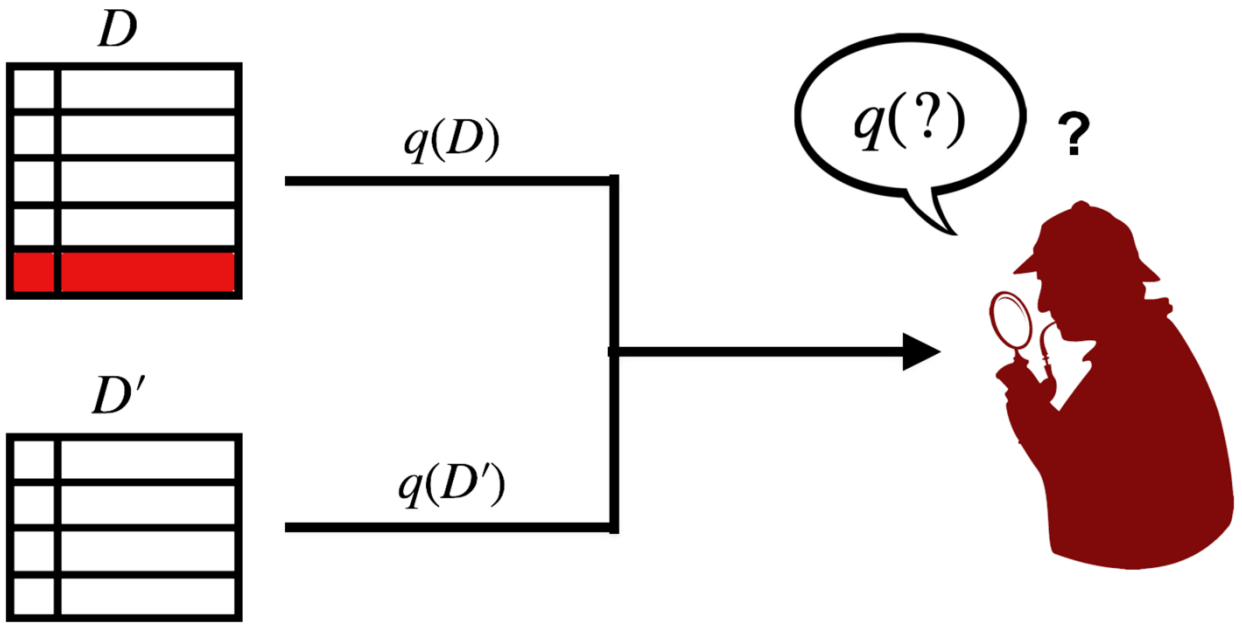
```
SRC_IP,DST_IP,SRC_PORT,DST_PORT,Protocol,IN_PKTS,OUT_PKTS,IN_BYTES,OUT_BYTES,FLOW_DURATION,Label
10.0.1.12,172.16.0.8,443,51524,TCP,18,22,12480,16340,2.3s,Benign
192.168.1.7,10.10.2.4,53,60122,UDP,2400,15,1850000,1200,0.8s,DDoS
...
```

Operator Pain Points with Data Retention



Tradeoffs:
Privacy VS Fidelity; Compression VS Fidelity

Metric Definition: Differential Privacy (ϵ-DP)



$$\frac{\Pr[M(D) = o]}{\Pr[M(D') = o]} \leq e^{\epsilon}$$

Metric Definition: Data Fidelity

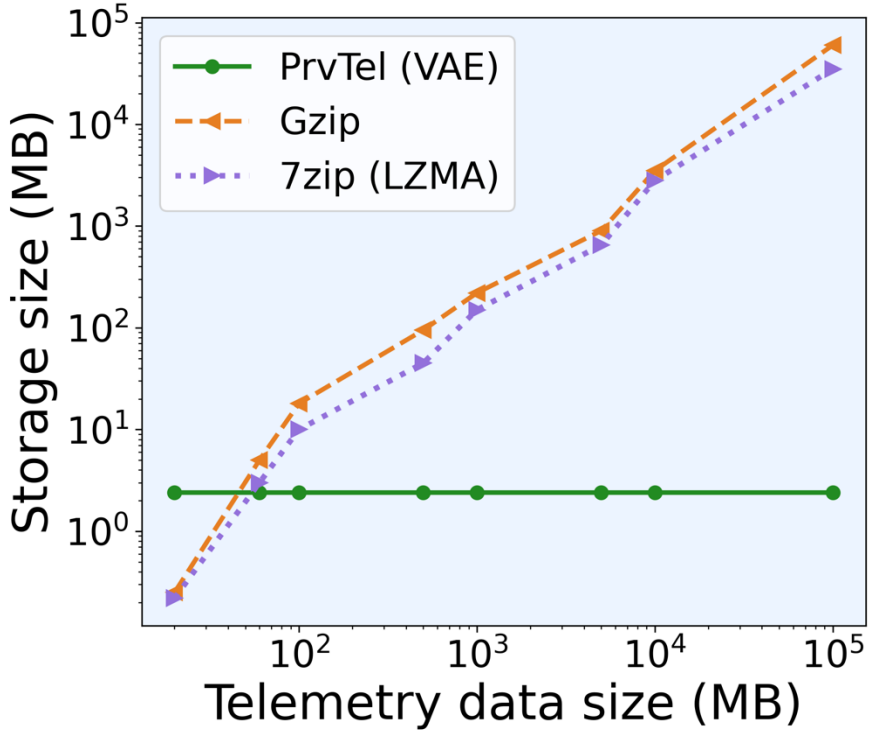
- Statistical summary
- Domain specific queries
- Downstream tasks

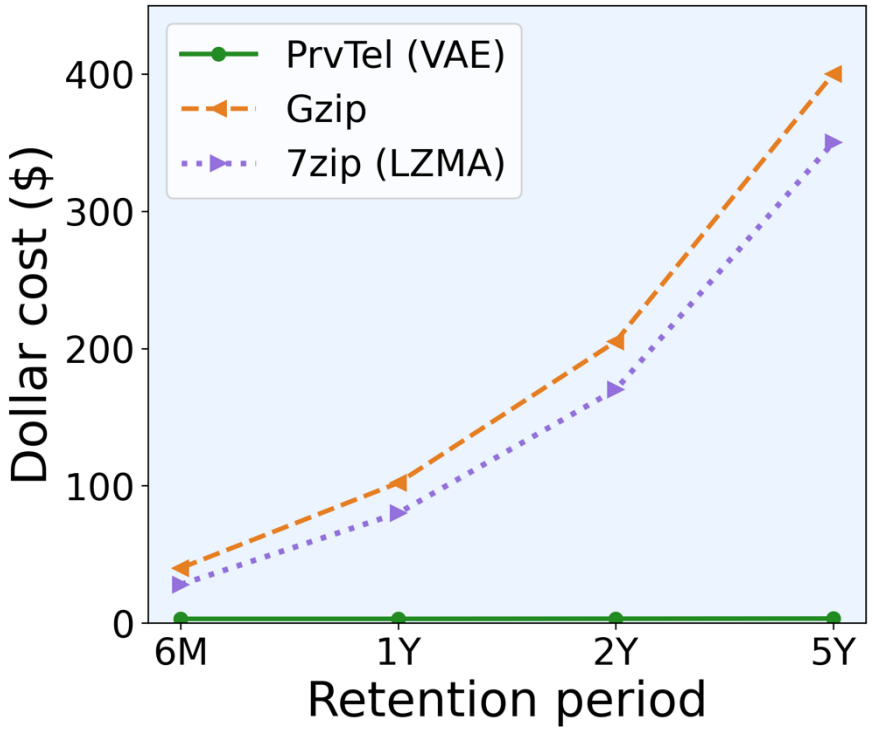
✓ Aggregated query
✓ Trend/pattern detection

✗ Exact flow/IP tracing
✗ Stateful query

Prior Work 1: Lossless Compression

1.1 TB of NetFlow data over 5 years





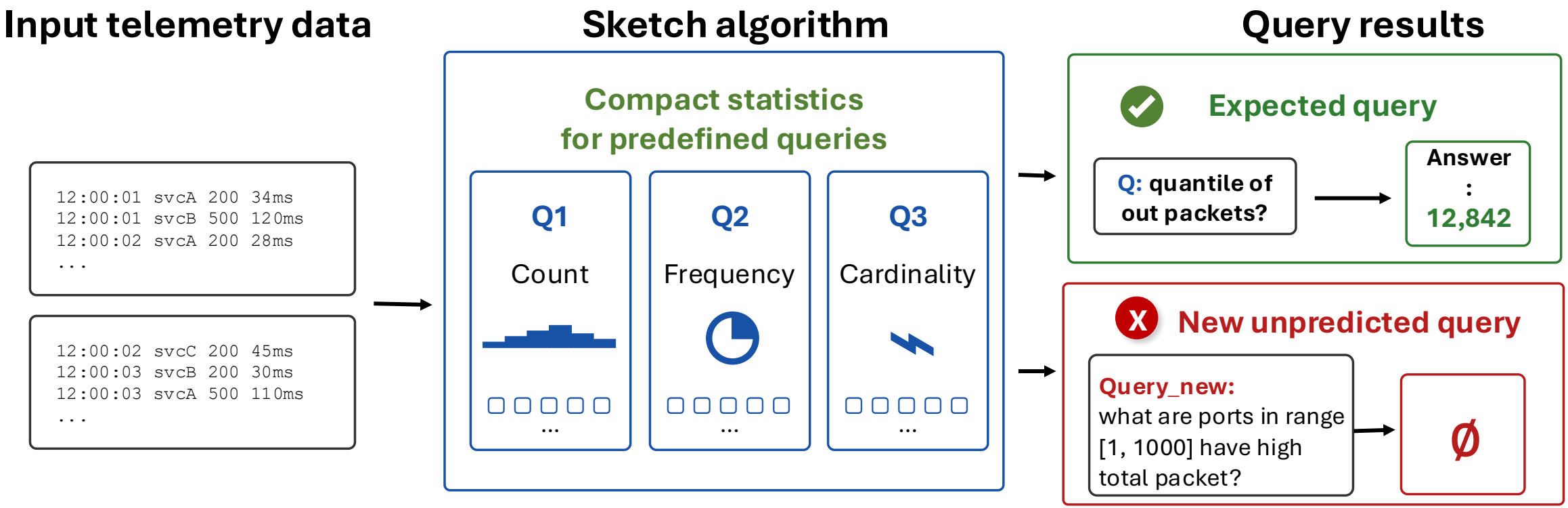
Estimated cumulative cost of retaining 1PB new data generated per week for one year: > $1M

Tradeoff Comparison Summary

	Query Generality	Cost VS. Fidelity	Privacy VS. Fidelity
Lossless	✓	X	X

Lossless give 100% query accuracy, but is not private, and have high retention cost over time.

Prior Work 2: Sketches



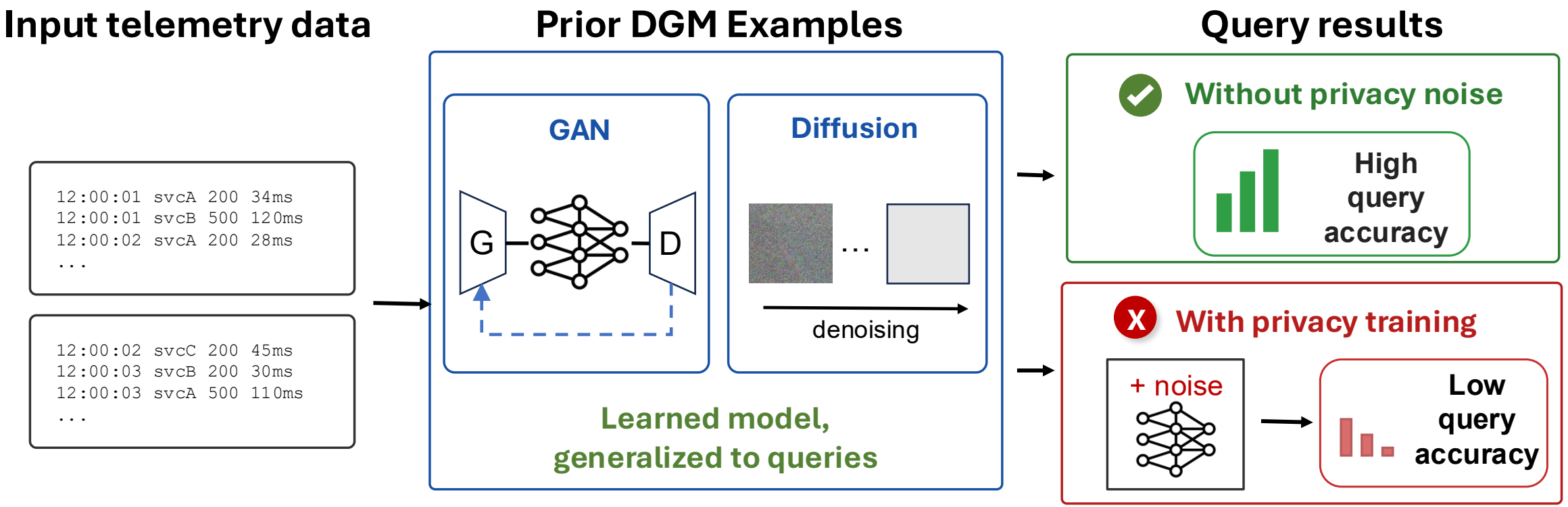
Sketches are efficient, but only for queries chosen ahead of time.

Tradeoff Comparison Summary

	Query Generality	Cost VS. Fidelity	Privacy VS. Fidelity
Lossless	✓	X	X
Sketches	X	✓	?

Adding privacy to Sketch need specific design adaptation on each Sketch algo

Prior Work 3: Deep Generative Models (DGM)



Adding privacy noise in DGM training degrades query accuracy

Tradeoff Comparison Summary

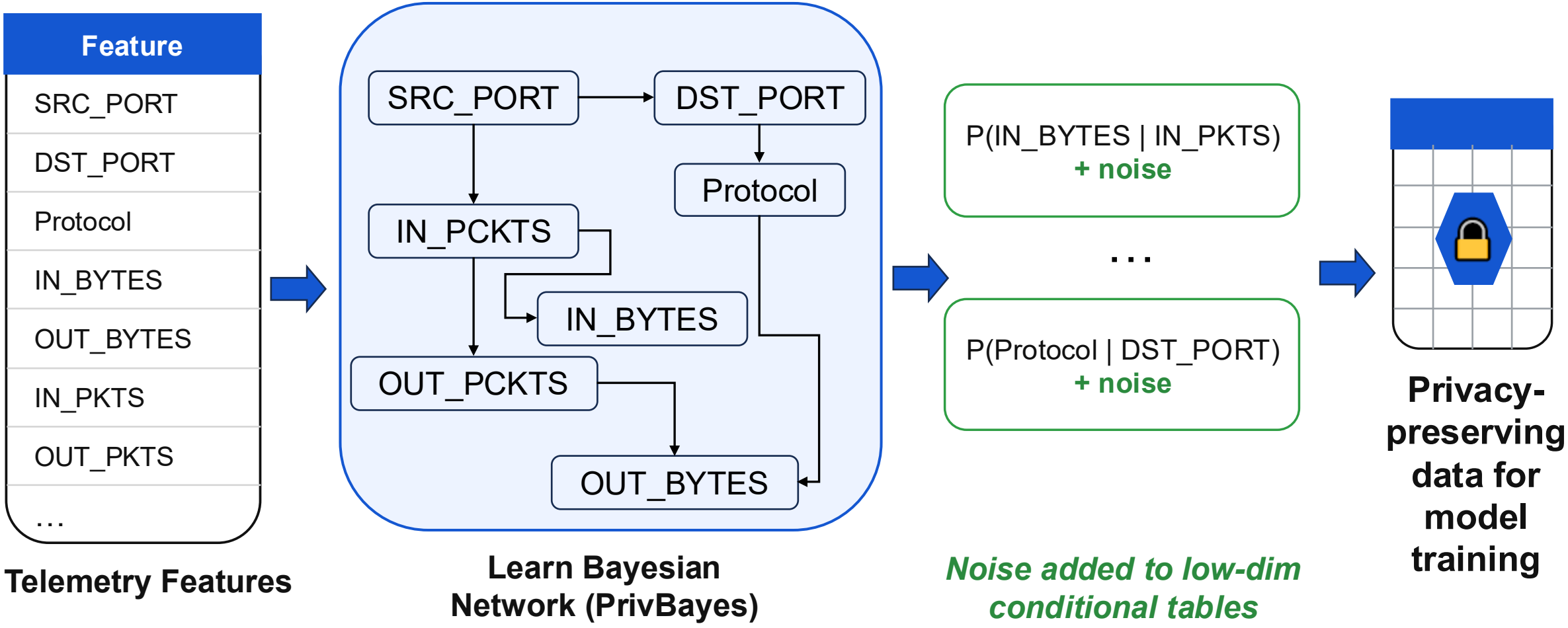
	Query Generality	Cost VS. Fidelity	Privacy VS. Fidelity
Lossless	✓	X	X
Sketches	X	✓	?
Prior DGM	✓	X	X

DGM training cost on GPU >> storage cost

Goal: Balance Tradeoffs of 3 Pain Points

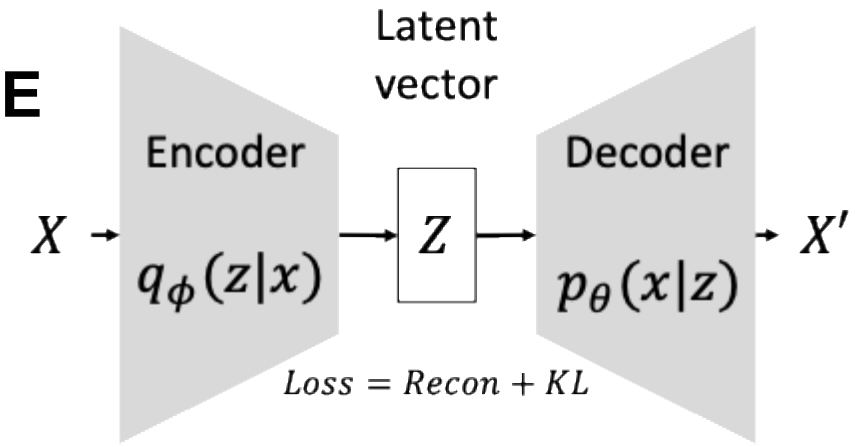
	Query Generality	Cost VS. Fidelity	Privacy VS. Fidelity
Lossless	✓	X	X
Sketches	X	✓	?
Prior DGM	✓	X	X
PrvTel	✓	✓	✓

Insight 1: Minimize added privacy noise before training

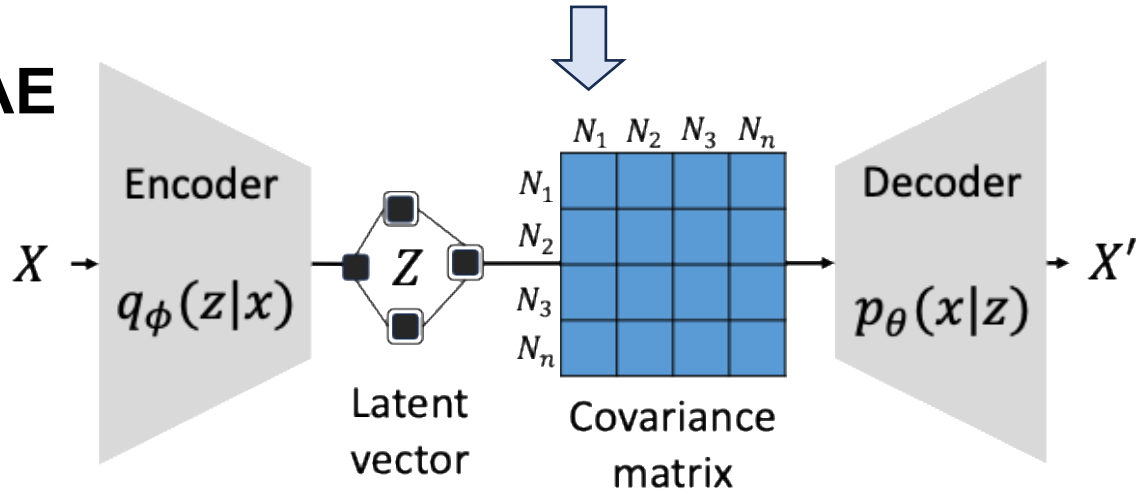


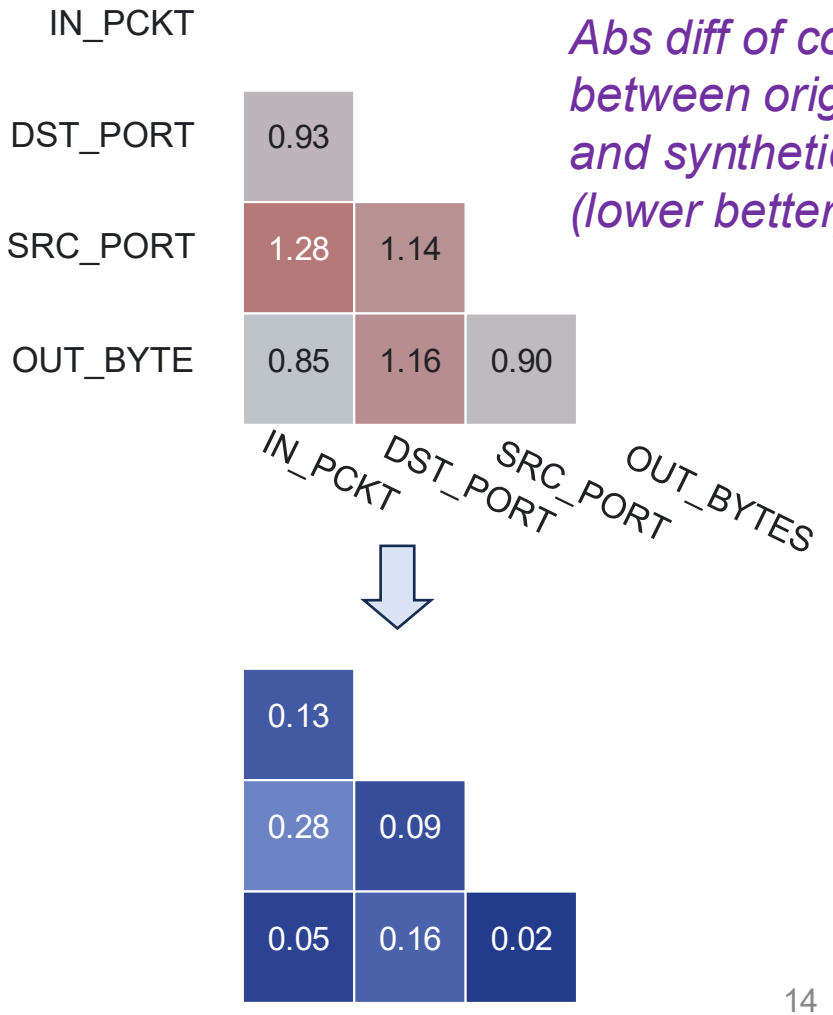
Insight 2: Use a lightweight VAE model and capture field correlations

Traditional VAE



PrvTel VAE

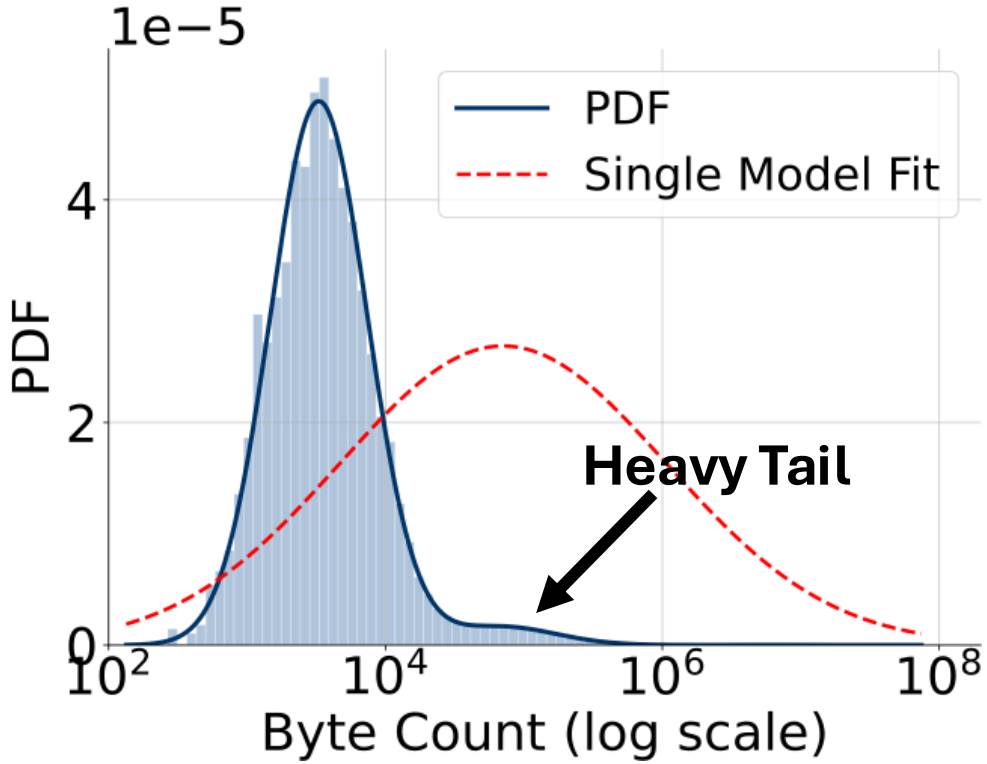


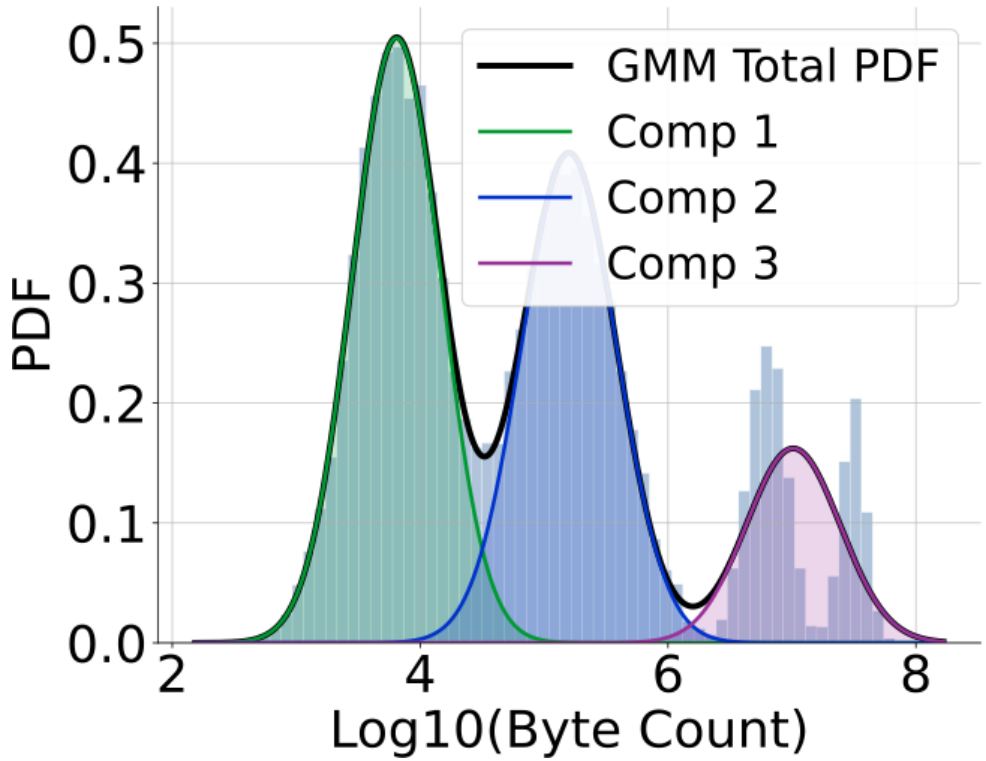


Abs diff of corr between origin and synthetic (lower better)

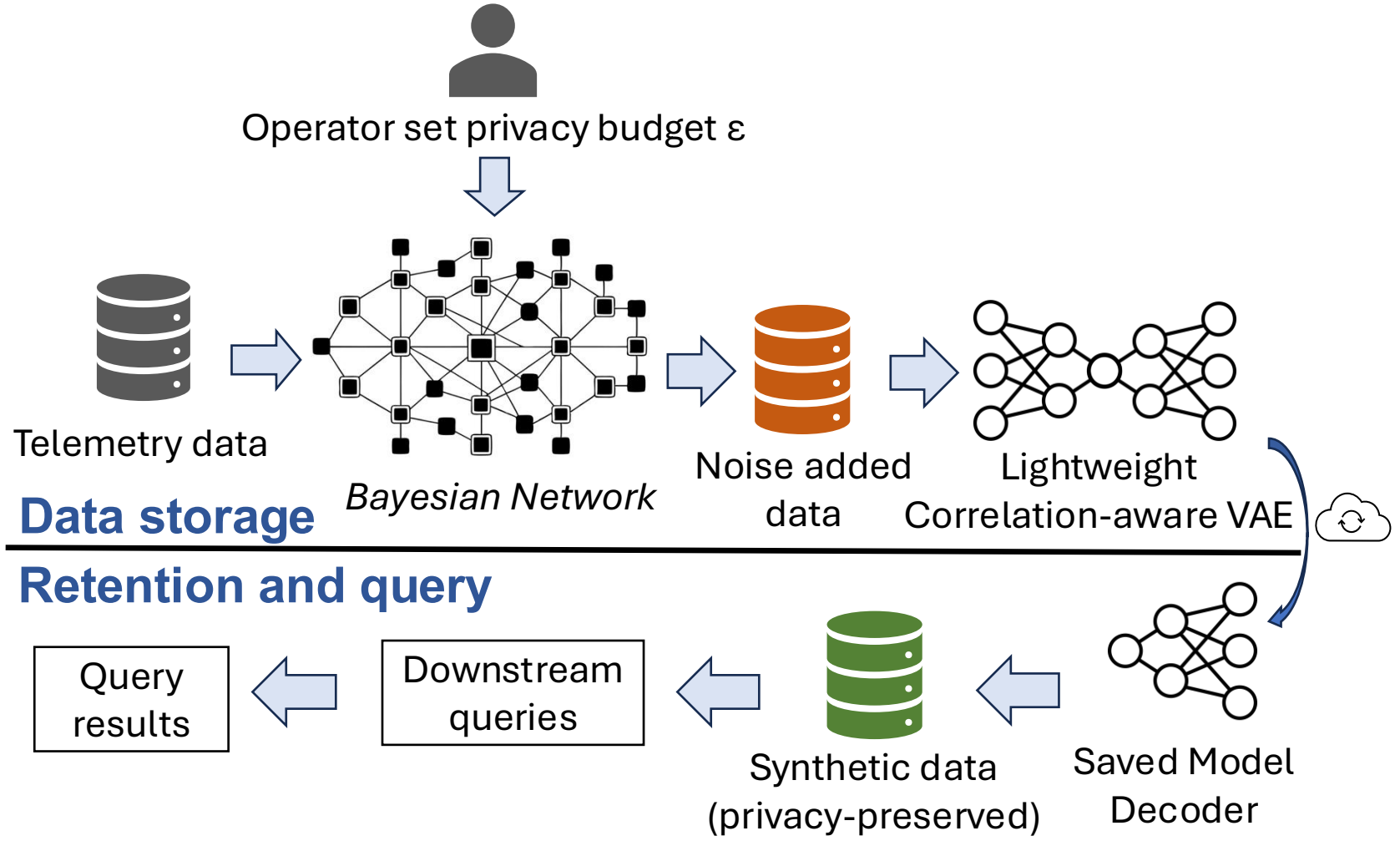
Insight 3: Domain-specific encoding

Standard reconstruction fail to capture telemetry data distributions





PrvTel System

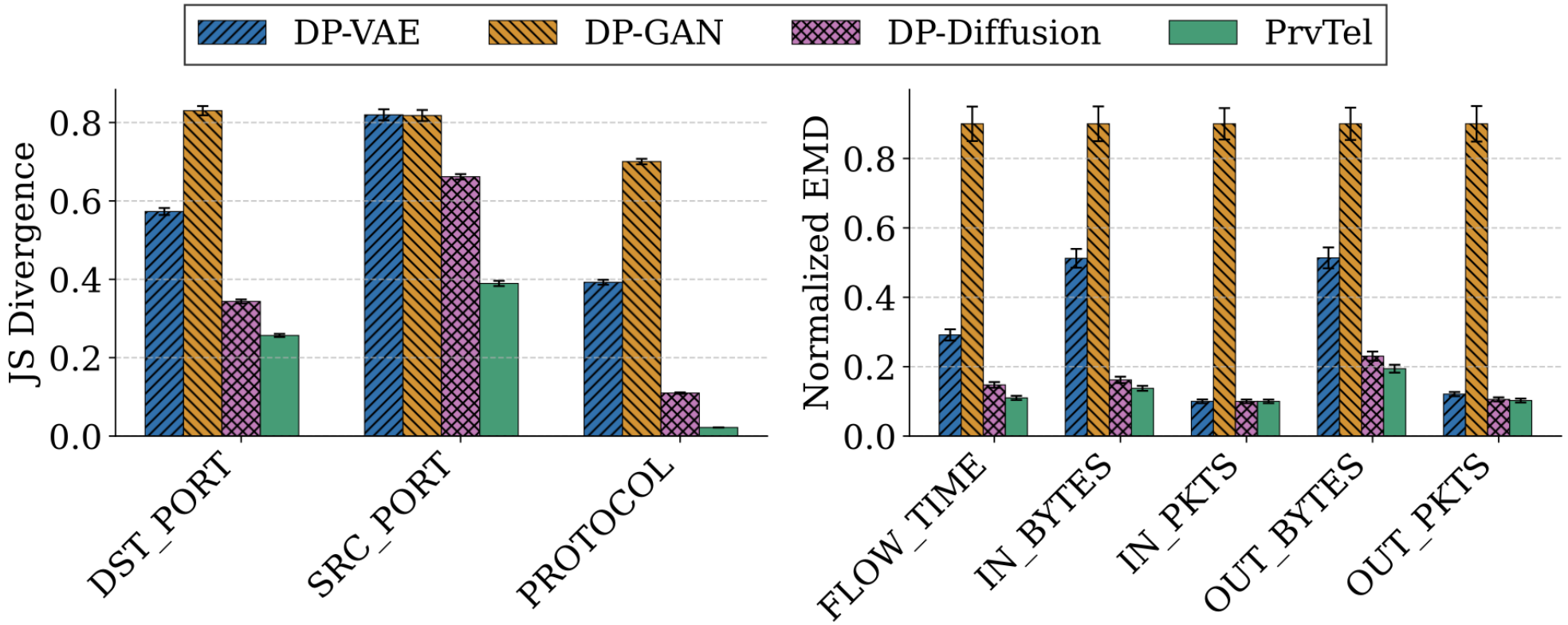


Evaluation Setup

6 baselines; 6 real dataset; 1 synthetic dataset

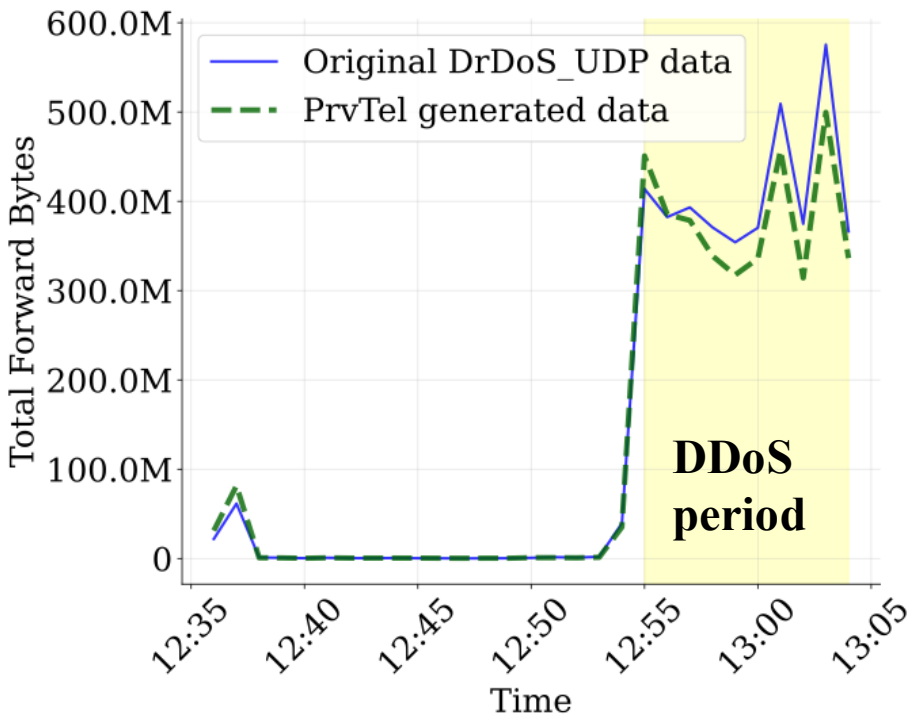
Category	Dataset	Size	Features	Duration	Feature Examples
NetFlow	Appraise [2]	820MB	15	2(d)	IN_BYTS, IN_PKTS
	NF-IoT [17]	4.10GB	10	2(d)	IN_BYTS, FLOW_TIME
	DDoS2019 [85]	410MB	13	2(h)	Fwd Packet, Fwd Byte
	CAIDA [26]	1.10TB	86	10(d)	Fwd Packet, ACK Count
Cloud	Cisco-IE [6]	80MB	23	2(d)	load-interval, bytes-sent
	NERC [16]	390MB	8	10(d)	cpu_ratio, sent_bytes

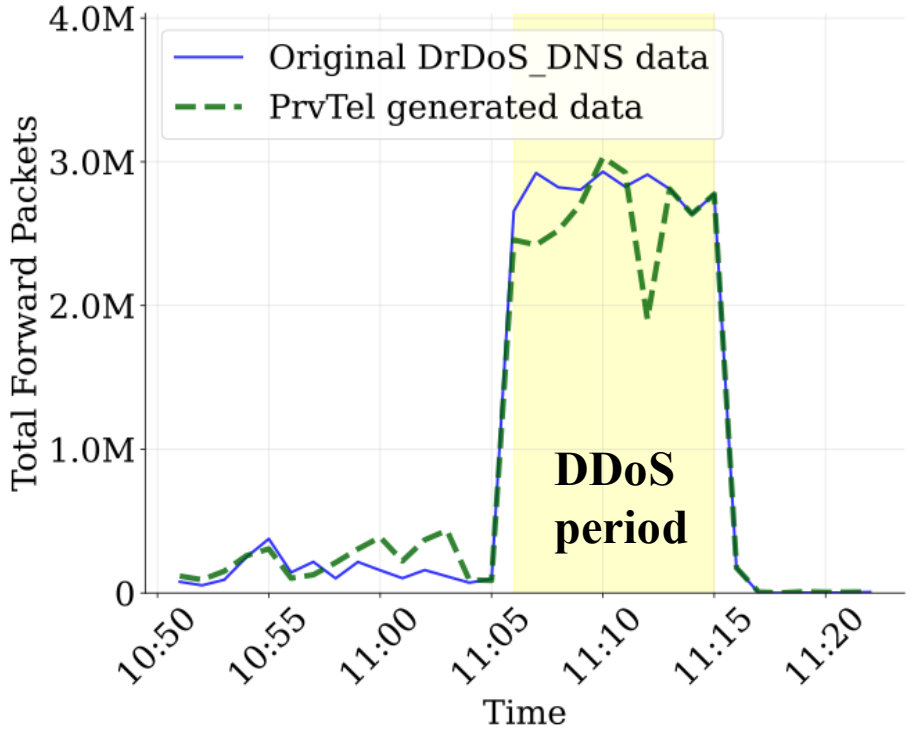
Result 1: PrvTel generated data is closer to real data



Adding noise before training minimize the fidelity degrades in the system

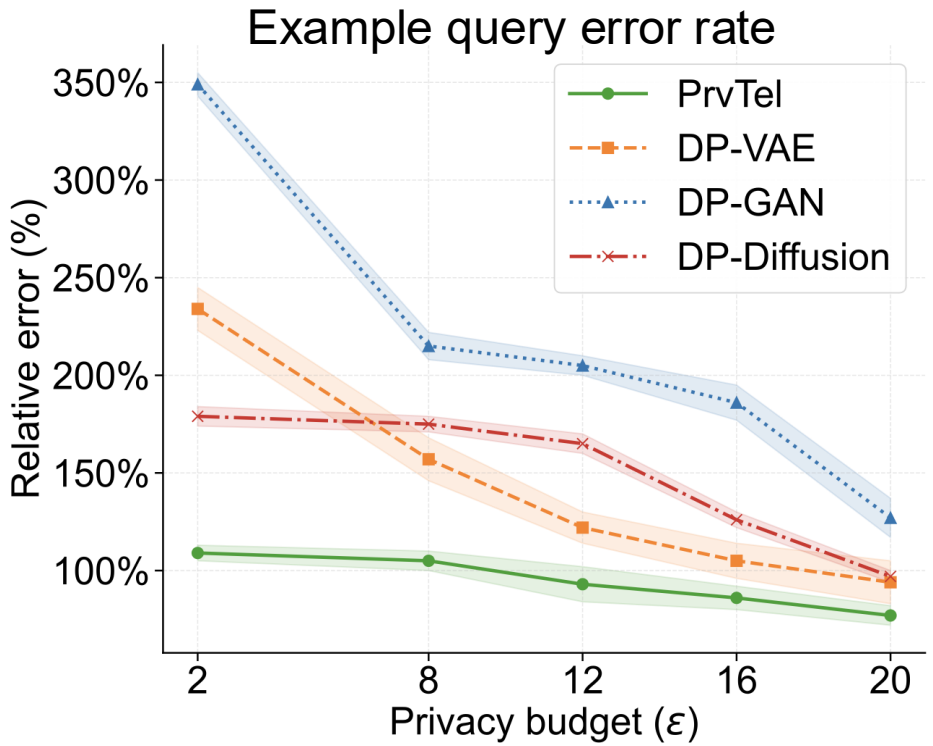
Result 2: PrvTel maintains downstream task utility

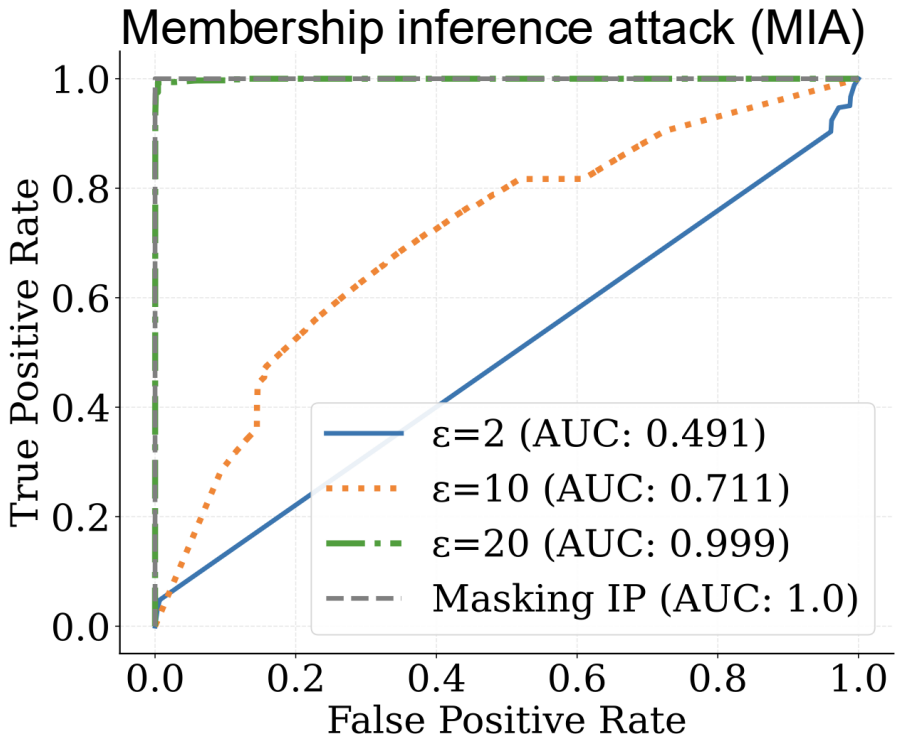




Correlation-aware design could capture some temporal patterns

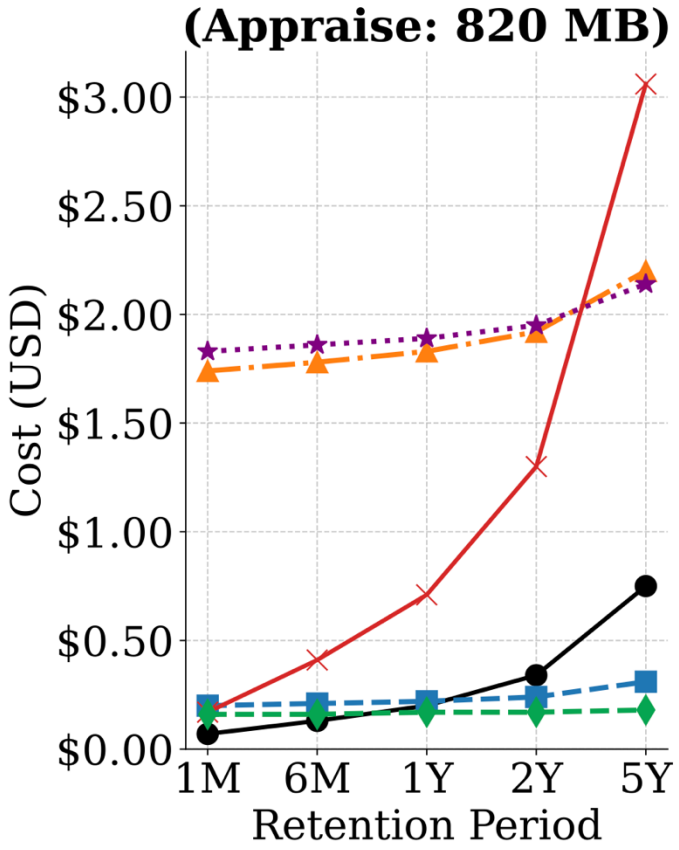
Result 3: PrvTel provides strong privacy even with a tight budget

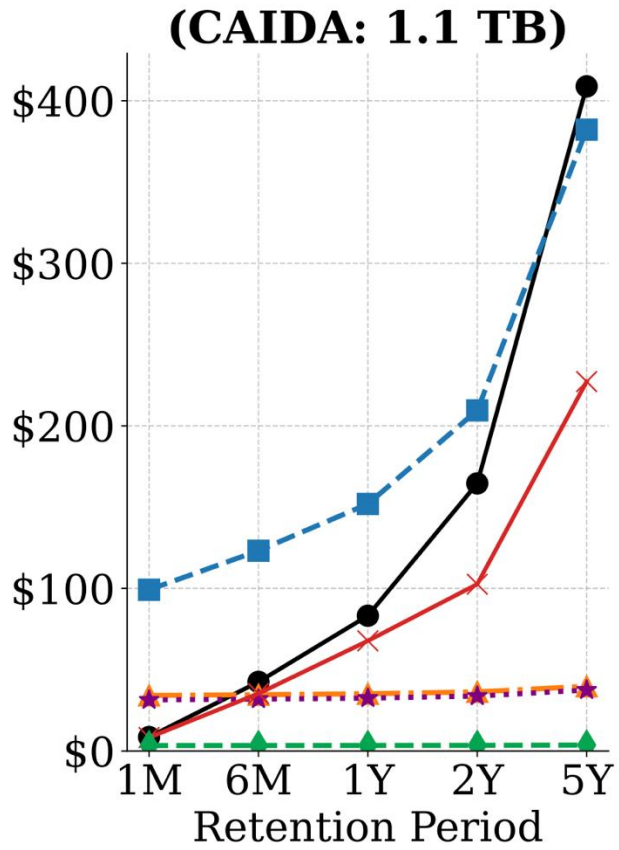




When privacy gets loose (↑), baseline fidelity increases, but the attack success rate also increase

Result 4: PrvTel has least retention cost over long time





PrvTel's VAE is cheap to train and to store

Takeaways

- Do NOT add noise during training if you want high fidelity

- Generative model can be a generalizable data retention approach, if well-adapted

Project page: https://github.com/Froot-NetSys/PrvTel