RPI

Funding Acknowledgement

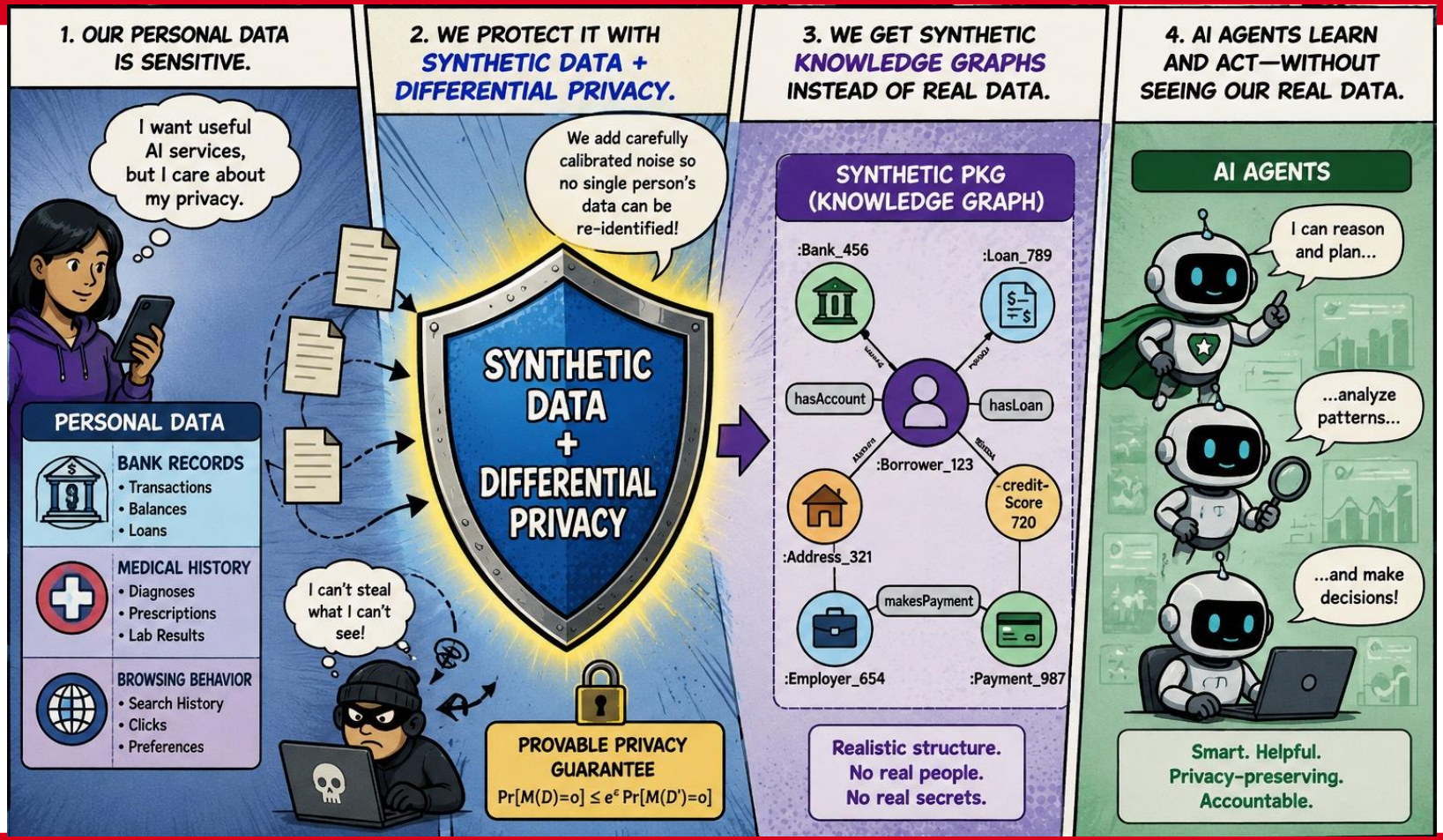
NSF

CRAFT
CENTER FOR RESEARCH TOWARD ADVANCING FINANCIAL TECHNOLOGIES

Trustworthy Agentic AI via Privacy-Preserving Synthetic Data: Lessons from Financial Tabular Data for PKG Systems

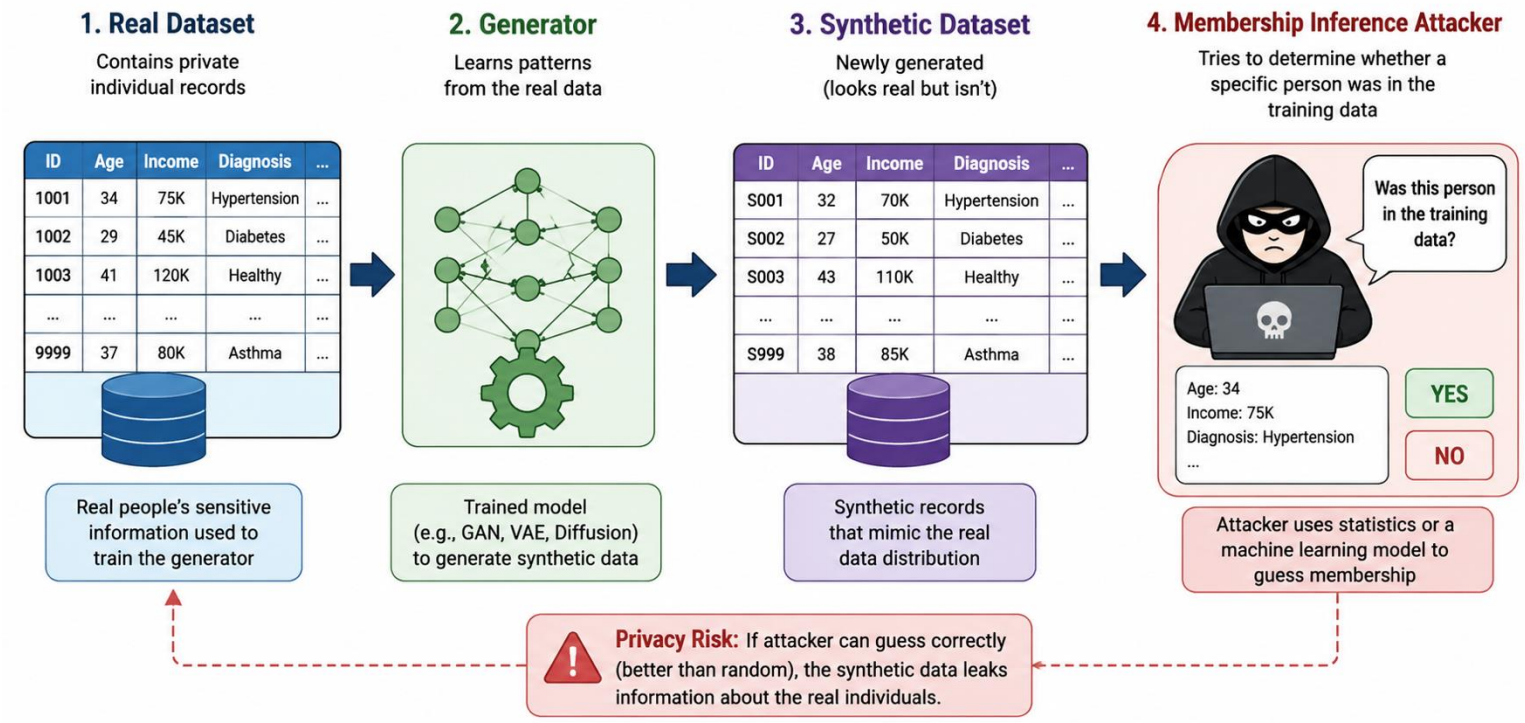
Oshani Seneviratne, Michael Zuo, Inwon Kang and Stacy Patterson

This presentation includes images generated using GenAI tools.



Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Synthetic Data Is Not Automatically Private



Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Differential Privacy: The Intuition

Hide the individual. Keep the insights.

1. Real World

Lots of people contribute their data (e.g., device metrics, usage, health stats...).

2. Analysis

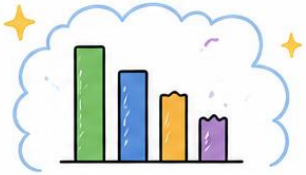
We compute some useful statistic or train a model.

3. Add a Little Noise

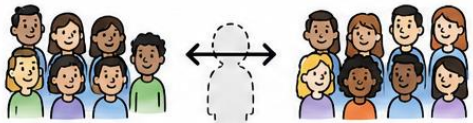
We add a small amount of random noise to the result.

4. Private Result

The result is still useful, but no one can tell whether any single person's data was included or not.



What Does It Guarantee?

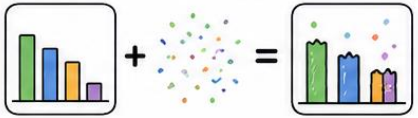


Dataset with you — Dataset without you

An observer cannot tell (with high confidence) whether your data was part of the dataset. Your impact on the output is "negligible."

How It Works (High-Level)

Add carefully calibrated random noise to the results.

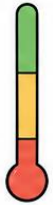


Noise hides individuals, but overall patterns remain.

Privacy Budget (ε)

Smaller ε (Stronger Privacy)

Larger ε (Weaker Privacy)



- ε controls how much noise is added.
- Smaller ε → more noise → stronger privacy.
- Larger ε → less noise → higher accuracy.
- We choose ε based on the desired privacy-utility trade-off.

Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Research Questions

1. *Can we generate high-quality synthetic data with formal privacy guarantees?*
2. *How do we audit actual privacy leakage?*
3. *Can these methods be used to build synthetic PKGs?*

Connection to TAAPAAI Themes

Theme	Contribution
Trust	Formal differential privacy
Autonomy	Safe training and simulation for agents
Accountability	Auditable privacy budgets and attack testing

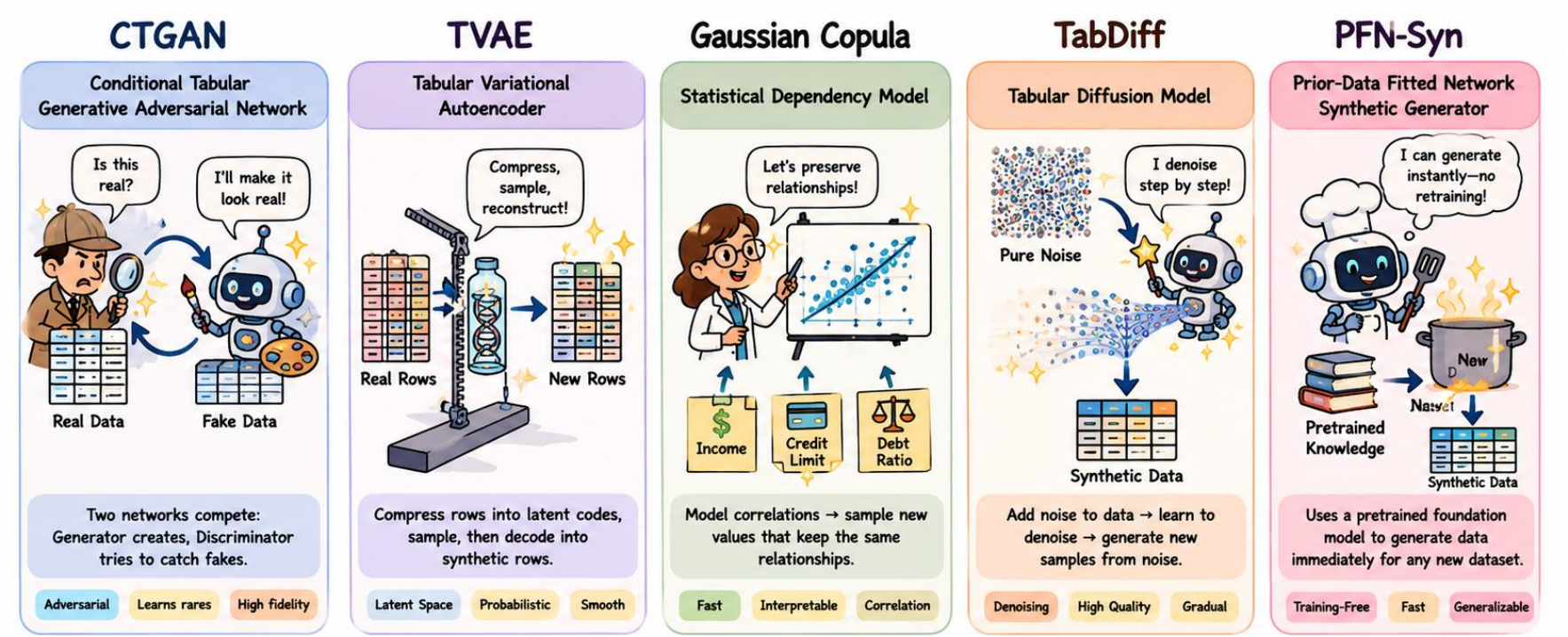
Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Benchmark Overview

- *New DP implementations of CTGAN and TVAE*
- *Privacy auditing framework*
- *Six financial datasets*

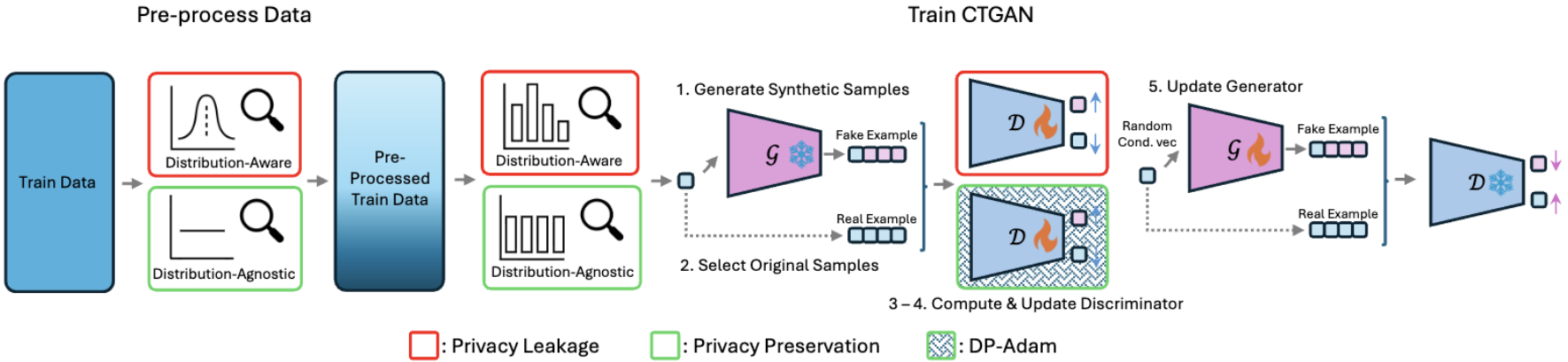
Dataset	Num. Samples	Num. Features	% Categorical Features	% Minority Class
Adult Income (AD)	48,842	14	57.14	23.9
Bank Customer Churn (BC)	10,000	11	45.45	20.3
Bank Marketing (BM)	45,211	13	57.14	11.7
Credit Card Defaults (CC)	30,000	23	39.13	22.12
German Credit Data (CR)	1,000	20	85.00	30.00
Give Me Some Credit (GM)	150,000	10	40.00	6.68

Generators



Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Synthetic Data Generation Algorithms



- *We constructed **differentially private versions of CTGAN and TVAE**.*
- *Performed extensive benchmarking of data quality and utility, across non-private and private-generators.*
- *Data utility metric: train XGBoost model and measure balanced accuracy:* $\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$.

Results

Dataset	Original	Non-DP				DP-CTGAN			
		Gauss.	TabDiff	CTGAN	TVAE	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$
AD	0.818	0.516	0.684	0.718	0.769	0.681	0.709	0.734	0.784
BC	0.721	0.539	0.707	0.670	0.650	0.523	0.556	0.562	0.648
BM	0.596	0.506	0.576	0.586	0.603	0.528	0.574	0.579	0.580
CC	0.658	0.546	0.642	0.609	0.635	0.554	0.614	0.613	0.673
CR	0.686	0.533	0.614	0.507	0.552	0.501	0.494	0.505	0.555
GM	0.594	0.502	0.583	0.628	0.666	0.541	0.633	0.591	0.583

DP version competitive with non-private generator.
Sometimes even improves performance over real data!

Membership Inference Attack

- *Customized Membership Inference Attack based on "Synthetic Data - Anonymisation Groundhog Day", by Stadler et al.*
- *Generate synthetic dataset pairs: half trained with real data; half trained with real data + "canary".*
- *Train distinguisher to determine type of training dataset: with canary or without.*
- *Run distinguisher on new synthetic dataset pairs: high distinguisher success rate indicates low data privacy.*
- *Evaluated attack on our new DP-CTGAN implementation.*

Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Results on the Membership Inference Attack

Dataset	Non-DP				DP-CTGAN			
	Gauss.	TabDiff	CTGAN	TVAE	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$
AD	0.491	0.507	0.505	0.496	0.513	0.497	0.512	0.475
BC	0.536	0.49	0.426	0.501	0.504	0.499	0.499	0.514
BM	0.523	0.482	0.505	0.515	0.491	0.487	0.49	0.488
CC	0.498	0.549	0.412	0.489	0.494	0.499	0.471	0.494
CR	0.530	0.502	0.497	0.488	0.517	0.495	0.516	0.485
GM	0.48	0.528	0.453	0.522	0.532	0.516	0.477	0.474

0.5 = random chance.

All generators exhibit **low success rate**; indicates gap between theoretical privacy and empirical attack risk.
Theoretical privacy tells us what is guaranteed under all circumstances.
Empirical privacy tells us what we observed against today's known attacks.

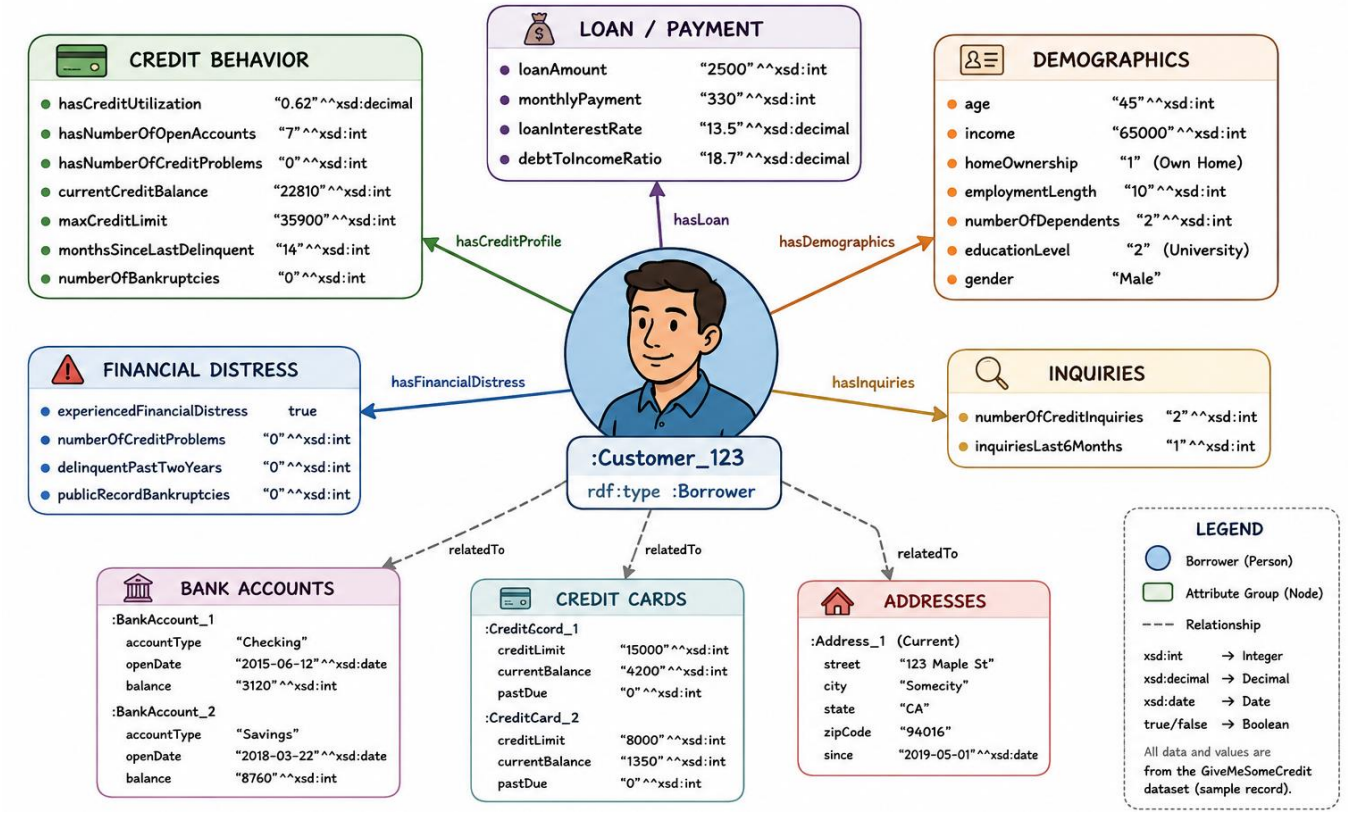
Theory vs Practice

- *Differential privacy gives provable measure of risk of determining whether any particular real record was used to produce synthetic dataset.*
  - *ϵ quantifies privacy leak.*
- *Our generators give a guaranteed ϵ for any input data.*
- *Analysis is conservative; privacy may be stronger in practice.*
- *Can convert membership inference attack success rates into empirical estimate of ϵ.*

From Tables to Personal Knowledge Graphs

Tabular Concept	**PKG Concept**
Row	Entity
Column	Property or Relation
Category	Linked Node
Numeric Value	Datatype Property

Example Synthetic PKG



Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Future Directions

- *Graph-native synthetic PKG generators*
- *Privacy auditing for graph data*
- *Integration with Solid Project pods*
- *Federated synthetic data generation*

Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)

Questions?

Oshani Seneviratne (https://oshani.info | senevo@rpi.edu)