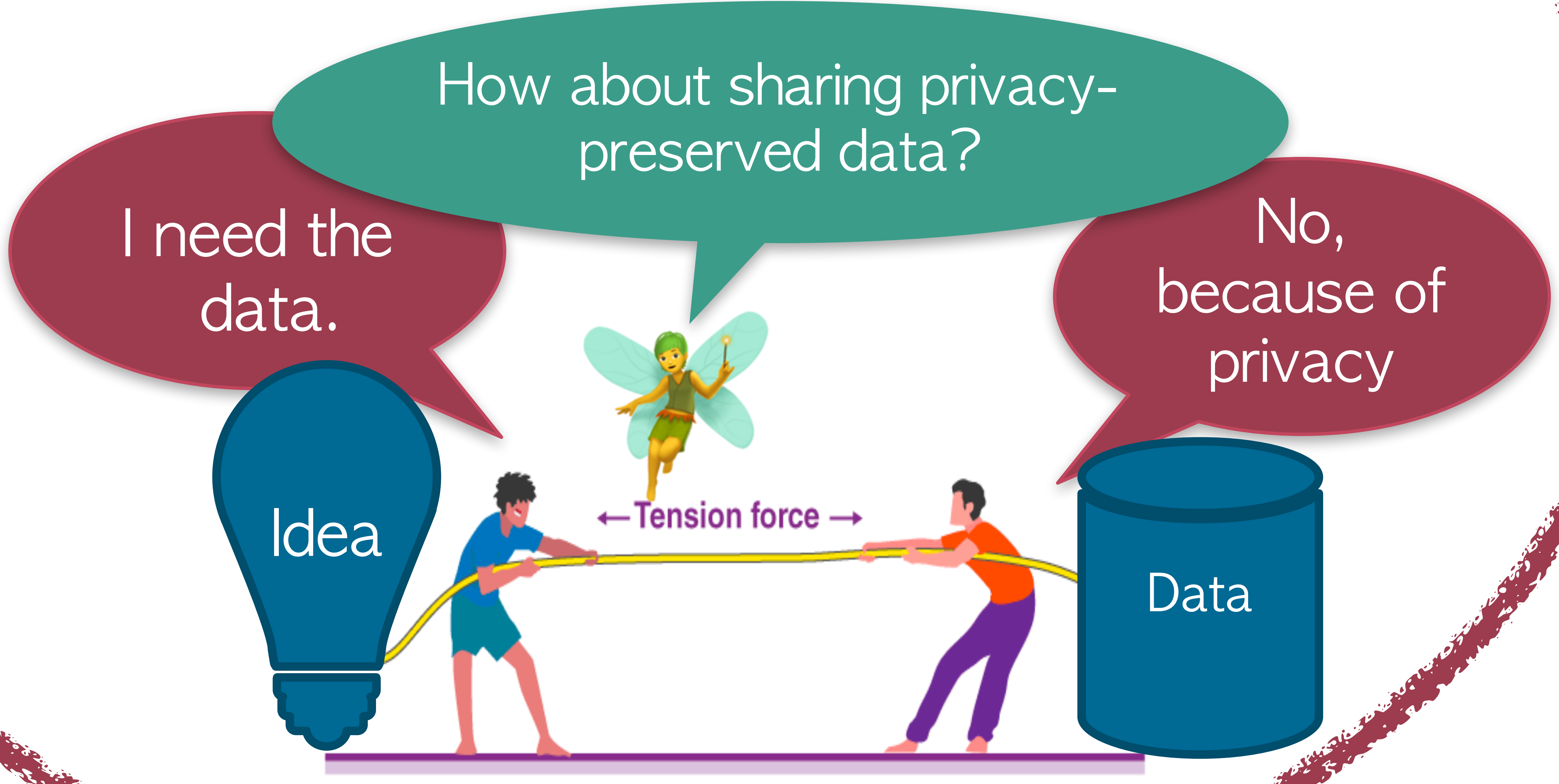
DistriNet

Towards Golden Standards for Quantifying Privacy of Synthetic Tabular Data

Qianying Liao, Dimitri Van Landuyt, Wouter Joosen

KU LEUVEN

Data Publishing



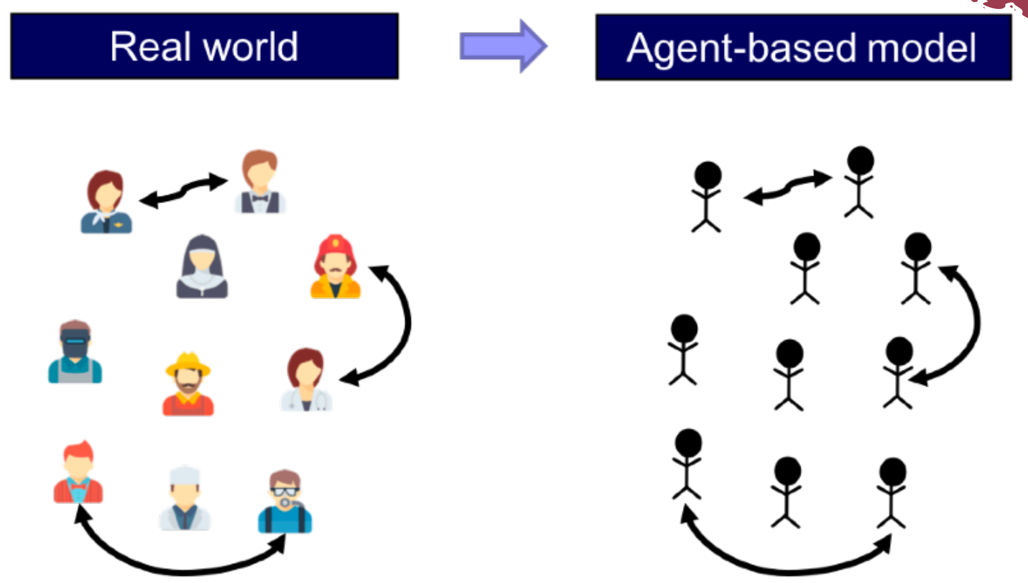
De-identified vs. Synthetic Data

- De-identified data

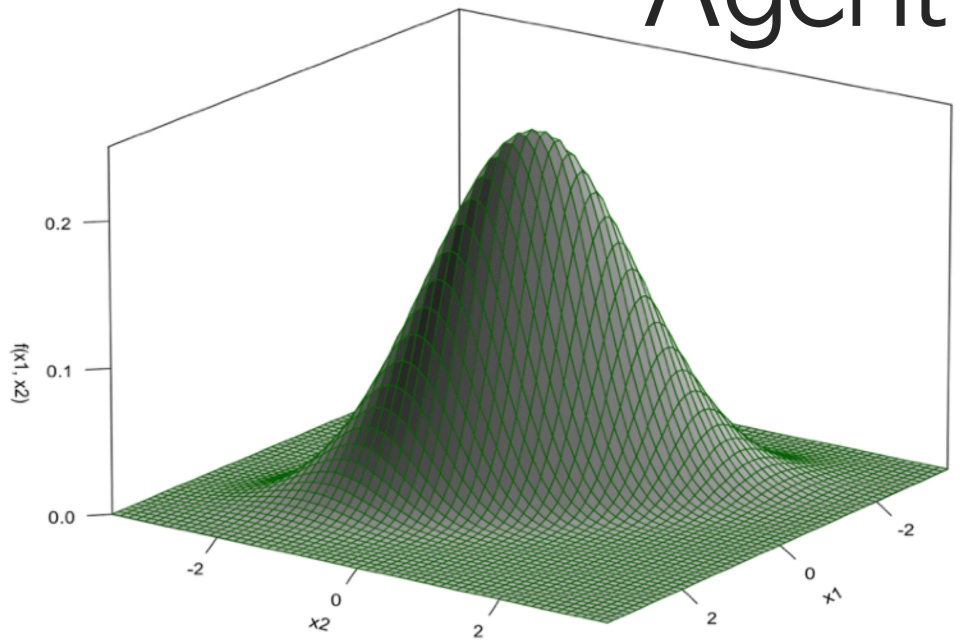
  Real data with PII removed/data fields scrambled

- Synthetic Data

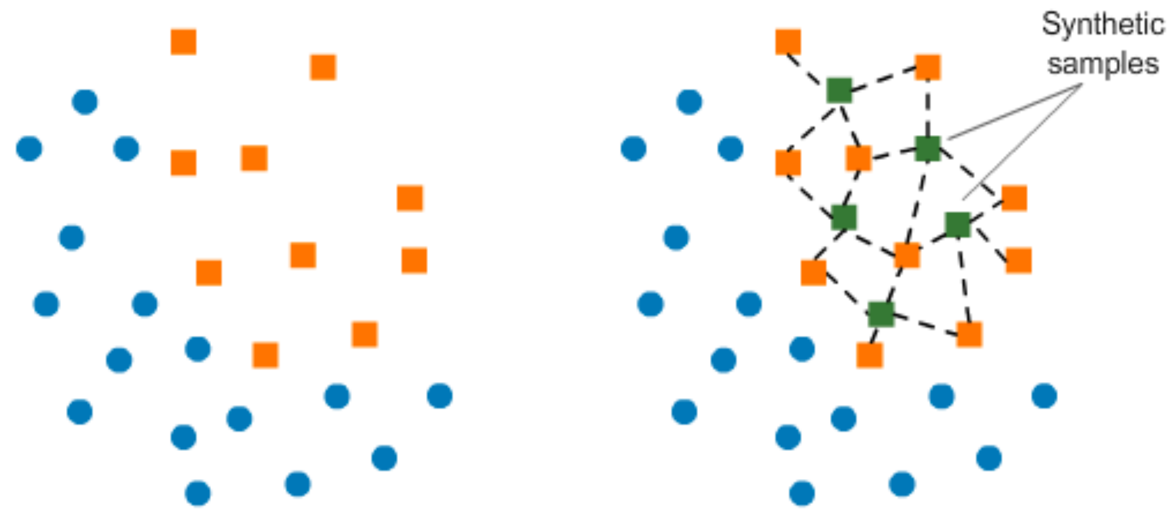
  Data created from a model



Agent-based Simulation



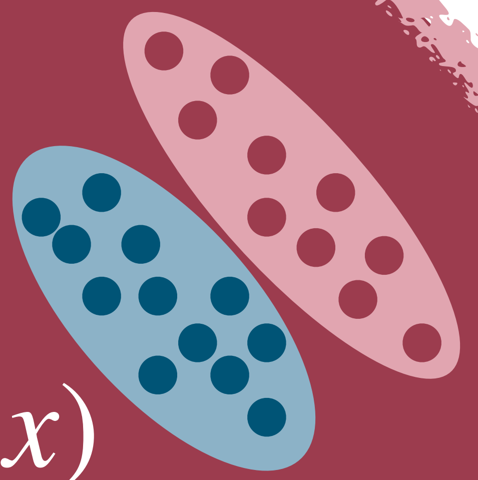
Multivariate Statistical Methods

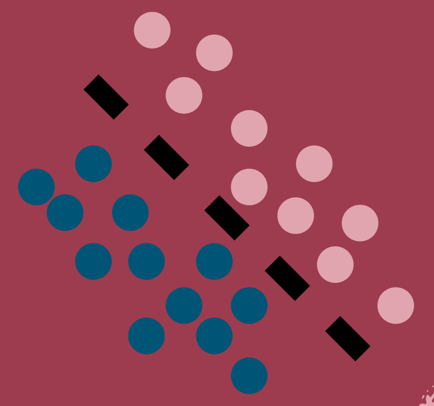


Oversampling

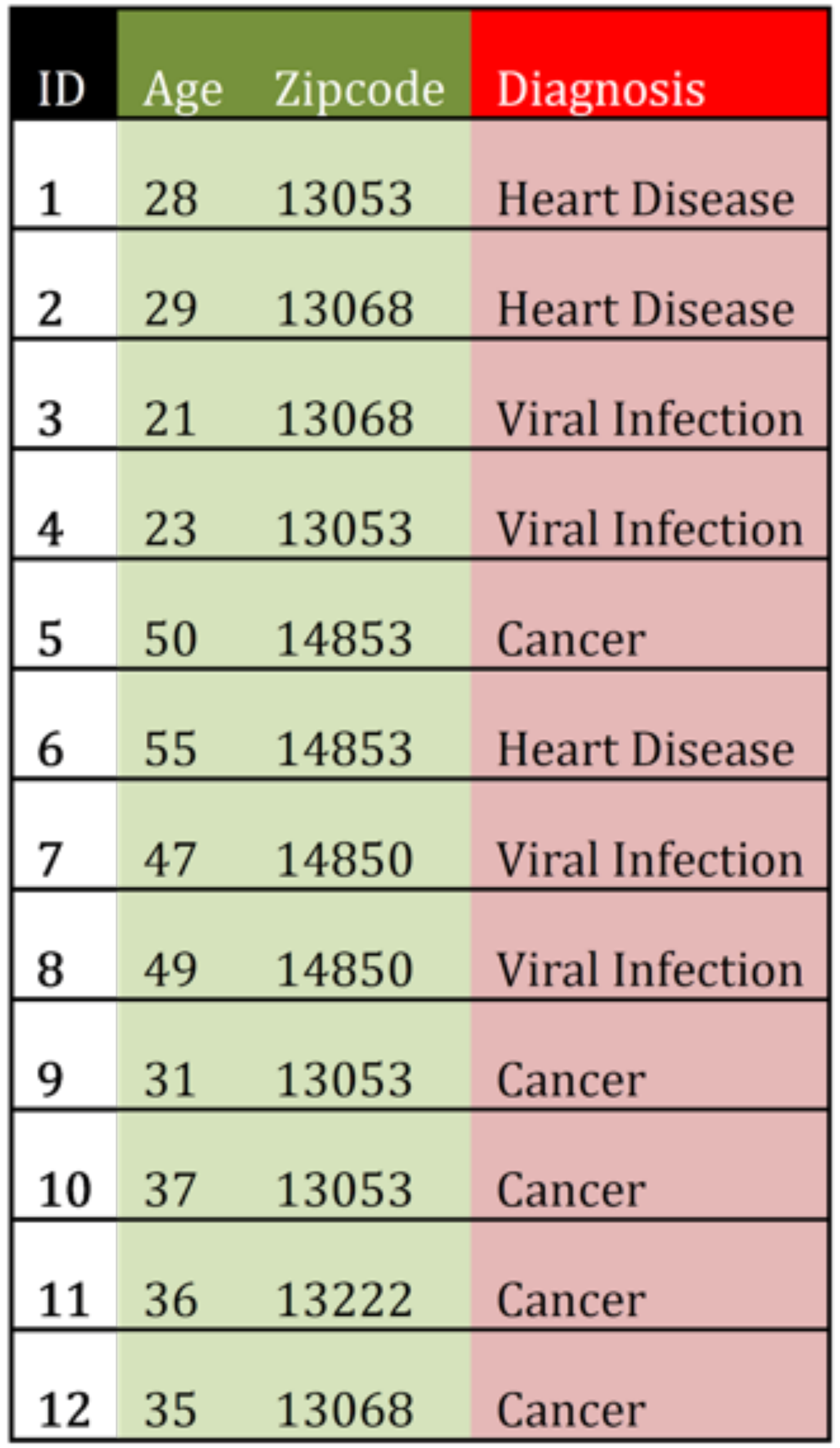
What is Generative Model?

- Discriminative Model $p(y\,|\,x)$
- Generative Model $p(x, y)$



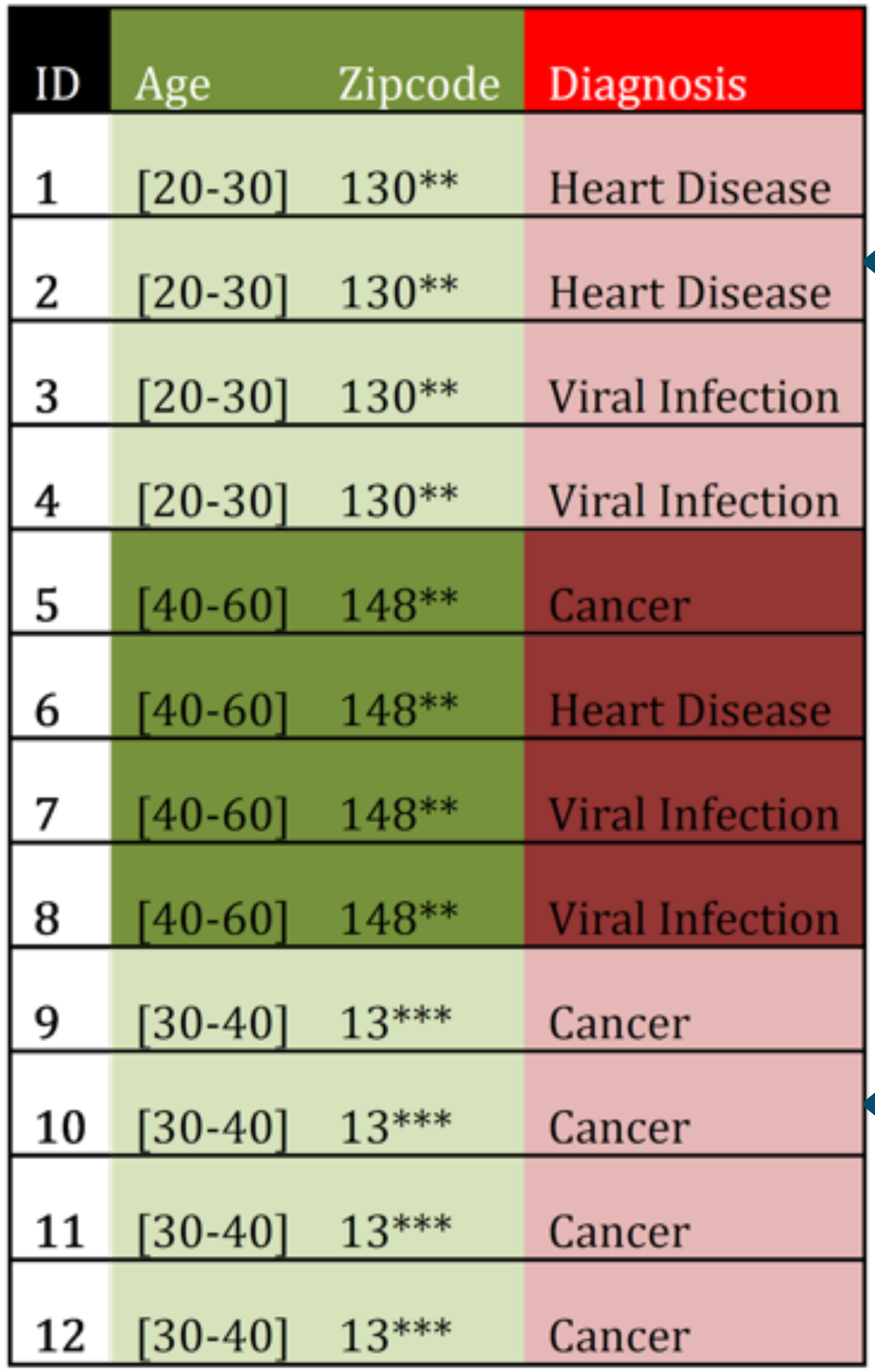


Is de-identification the silver bullet?



ID	Age	Zipcode	Diagnosis
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer

k-anonymization



ID	Age	Zipcode	Diagnosis
1	[20-30]	130**	Heart Disease
2	[20-30]	130**	Heart Disease
3	[20-30]	130**	Viral Infection
4	[20-30]	130**	Viral Infection
5	[40-60]	148**	Cancer
6	[40-60]	148**	Heart Disease
7	[40-60]	148**	Viral Infection
8	[40-60]	148**	Viral Infection
9	[30-40]	13***	Cancer
10	[30-40]	13***	Cancer
11	[30-40]	13***	Cancer
12	[30-40]	13***	Cancer

De-identified data is not the way - Dwork

"The arms race of re-identification and de-identification" - Bellovin et al

Is synthetic data generation the silver bullet?

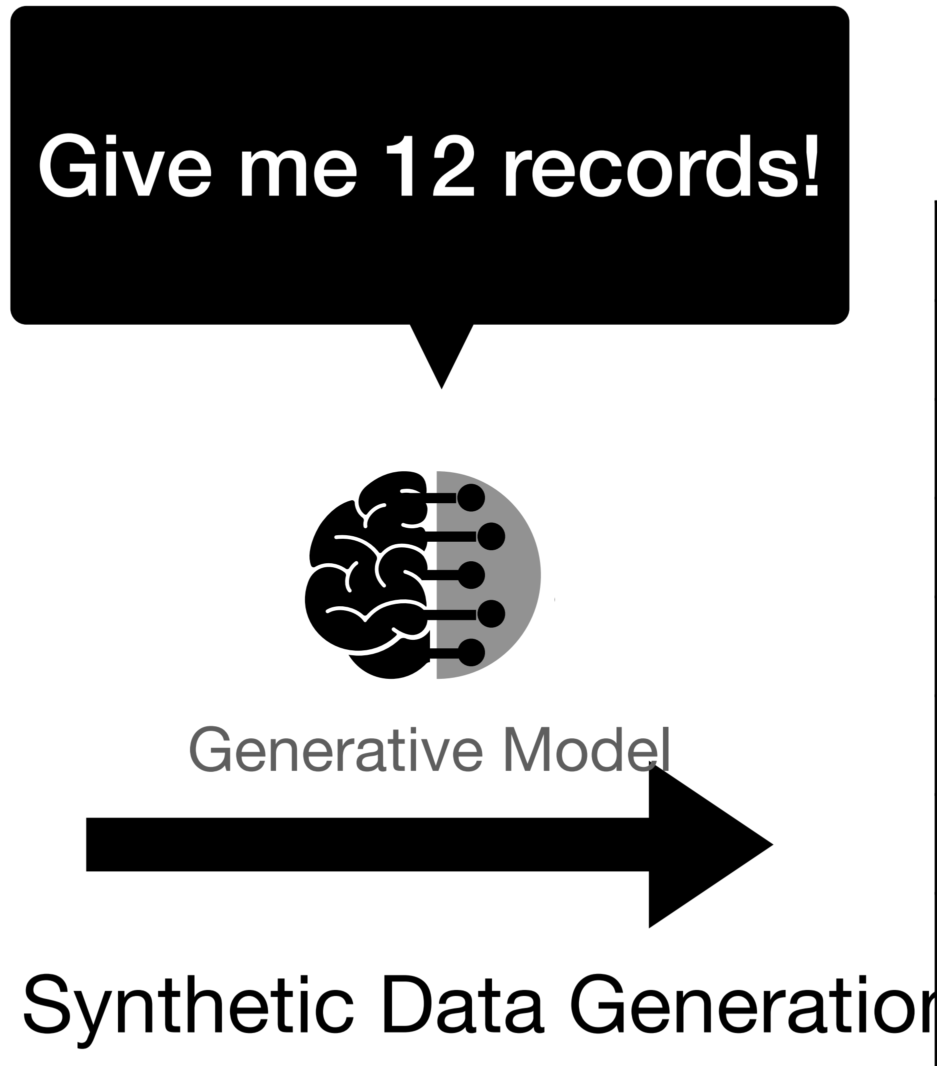
“First-Glance” Similarity

Easy to be evaluated

Privacy Friendly

Flexible

ID	Age	Zipcode	Diagnosis
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer


Give me 12 records!

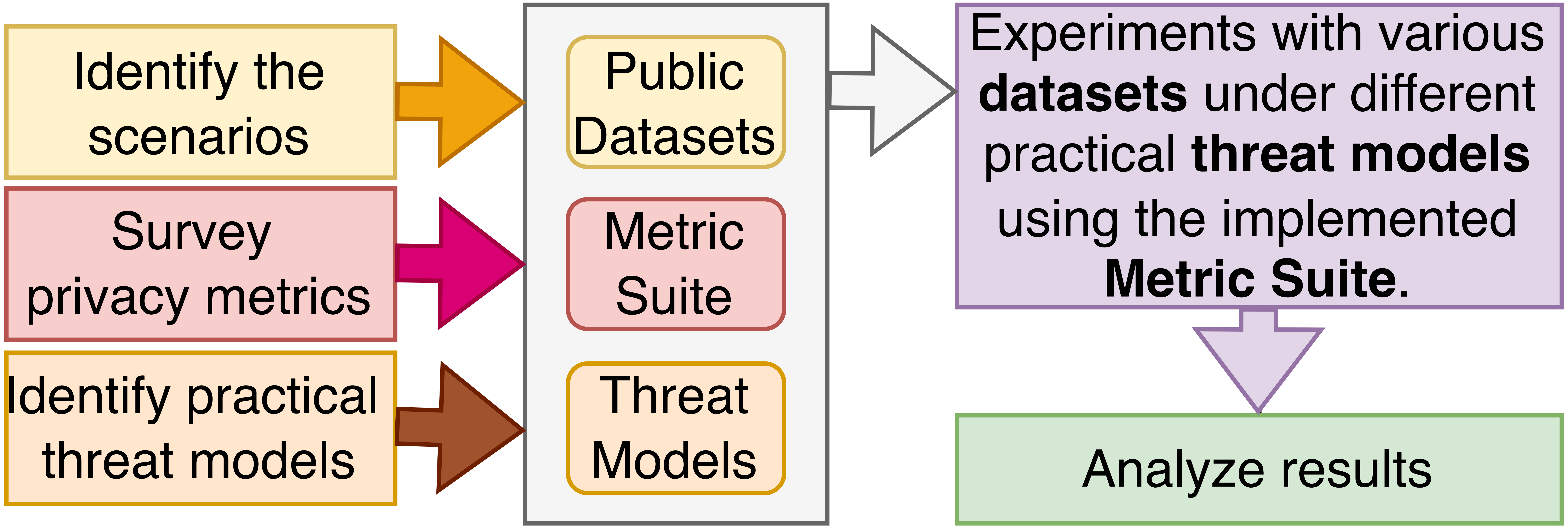
Synthetic Data Generation

ID	Age	Zipcode	Diagnosis
1	25	14651	Cancer
2	55	16546	Heart Disease
3	28	16544	Viral Infection
4	30	16545	Cancer
5	78	16160	Cancer
6	55	14410	Heart Disease
7	33	14564	Cancer
8	26	14646	Heart Disease
9	38	16464	Viral Infection
10	36	19845	Cancer
11	22	16444	Heart Disease
12	28	16545	Viral Infection

Is synthetic data a really better alternative to de-identified data?

What are the gold standards for evaluating synthetic data?

Current State of Synthetic Tabular Data Evaluation



Identify the scenarios

Survey privacy metrics

Identify practical threat models

Public Datasets

Metric Suite

Threat Models

Experiments with various **datasets** under different practical **threat models** using the implemented **Metric Suite**.

Analyze results

Research Methodology

Research Objectives

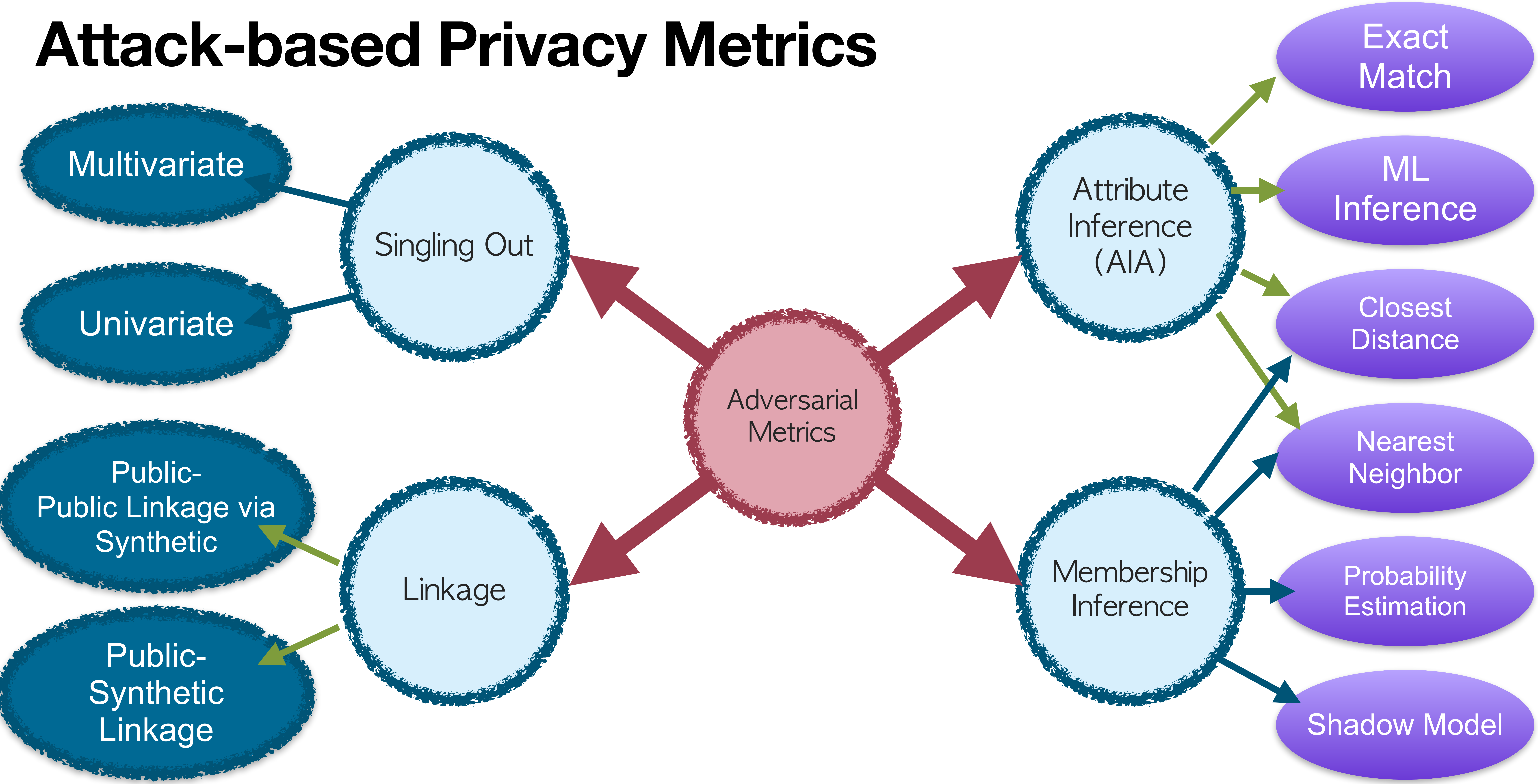
- **RO1:** Survey of Privacy Metrics
- **RO2:** Effectiveness and Efficiency of Privacy Metrics
- **RO3:** Cut-off Values for Privacy Metrics
- **RO4:** A Gold Standard for Privacy Assessment

Similarity-based Privacy Metrics

Non-adversarial Metrics
Memoriza-tion
Hitting %
Similarity %
Neighbor Dist.
Close Value %
Neighbor %
Re-identifiability
Data Likelihood
Detection
Distinguish-ability
k-anonymity
l-diversity
t-closeness

Qianying Liao, Dimitri Van Landuyt, Wouter Joosen, 2025. Pick Your Enemy: Pick Your Enemy: A Survey on Privacy Threat Models of Synthetic Tabular Data

Attack-based Privacy Metrics



Qianying Liao, Dimitri Van Landuyt, Wouter Joosen, 2025. Pick Your Enemy: Pick Your Enemy: A Survey on Privacy Threat Models of Synthetic Tabular Data

Threat Models in Privacy Metrics

- No-Box: Black-box that only returns synthetic data with no specific prompts
- Real Black-Box: Black-box with conditional prompts
- Grey-Box: Model hyper-params
- White-Box: Model params

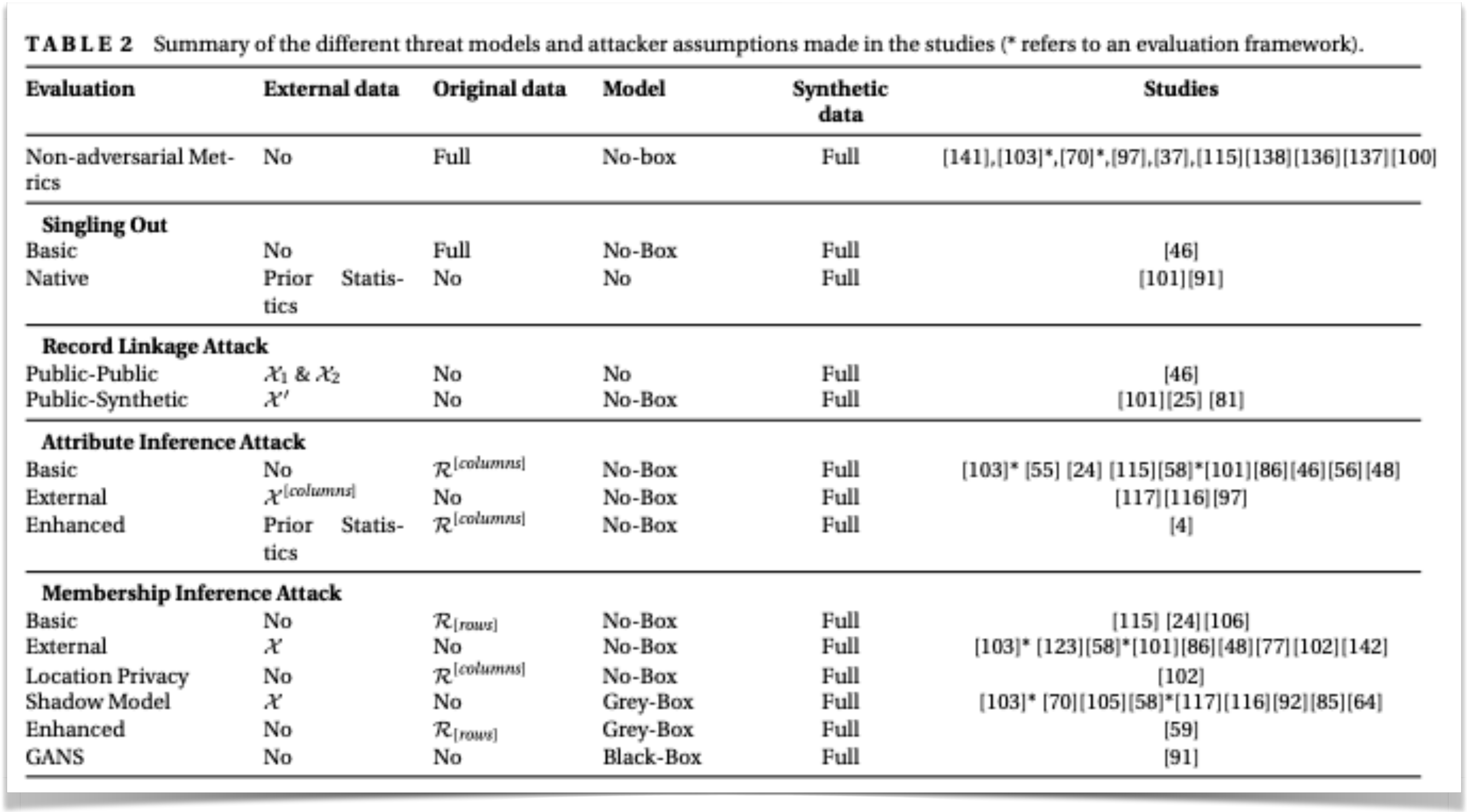
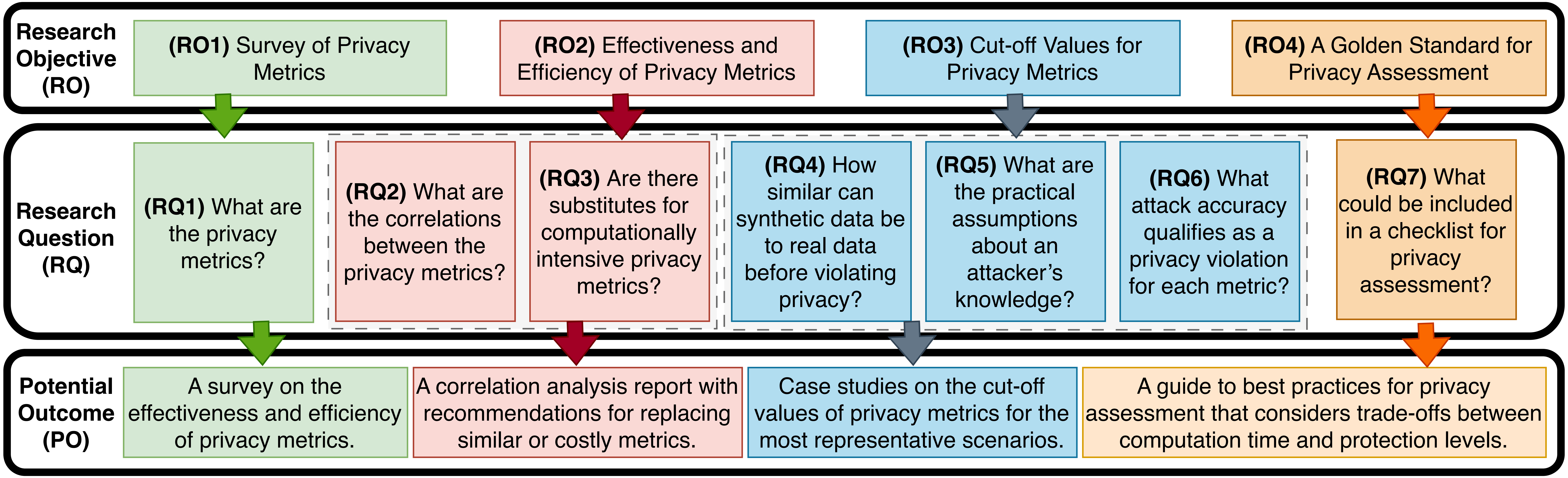


TABLE 2 Summary of the different threat models and attacker assumptions made in the studies (* refers to an evaluation framework).

Evaluation	External data	Original data	Model	Synthetic data	Studies
Non-adversarial Metrics	No	Full	No-box	Full	[141],[103]*,[70]*,[97],[37],[115][138][136][137][100]
Singling Out					
Basic	No	Full	No-Box	Full	[46]
Native	Prior Statistics	No	No	Full	[101][91]
Record Linkage Attack					
Public-Public	$\mathcal{X}_1$ & $\mathcal{X}_2$	No	No	Full	[46]
Public-Synthetic	$\mathcal{X}'$	No	No-Box	Full	[101][25] [81]
Attribute Inference Attack					
Basic	No	$\mathcal{R}^{[columns]}$	No-Box	Full	[103]* [55] [24] [115][58]*[101][86][46][56][48]
External	$\mathcal{X}^{[columns]}$	No	No-Box	Full	[117][116][97]
Enhanced	Prior Statistics	$\mathcal{R}^{[columns]}$	No-Box	Full	[4]
Membership Inference Attack					
Basic	No	$\mathcal{R}_{[rows]}$	No-Box	Full	[115] [24][106]
External	$\mathcal{X}$	No	No-Box	Full	[103]* [123][58]*[101][86][48][77][102][142]
Location Privacy	No	$\mathcal{R}^{[columns]}$	No-Box	Full	[102]
Shadow Model	$\mathcal{X}$	No	Grey-Box	Full	[103]* [70][105][58]*[117][116][92][85][64]
Enhanced	No	$\mathcal{R}_{[rows]}$	Grey-Box	Full	[59]
GANS	No	No	Black-Box	Full	[91]

Qianying Liao, Dimitri Van Landuyt, Wouter Joosen, 2025. Pick Your Enemy: Pick Your Enemy: A Survey on Privacy Threat Models of Synthetic Tabular Data

Research Objectives - Next Steps



Research Objective (RO)

(RO1) Survey of Privacy Metrics

(RO2) Effectiveness and Efficiency of Privacy Metrics

(RO3) Cut-off Values for Privacy Metrics

(RO4) A Golden Standard for Privacy Assessment

Research Question (RQ)

(RQ1) What are the privacy metrics?

(RQ2) What are the correlations between the privacy metrics?

(RQ3) Are there substitutes for computationally intensive privacy metrics?

(RQ4) How similar can synthetic data be to real data before violating privacy?

(RQ5) What are the practical assumptions about an attacker's knowledge?

(RQ6) What attack accuracy qualifies as a privacy violation for each metric?

(RQ7) What could be included in a checklist for privacy assessment?

Potential Outcome (PO)

A survey on the effectiveness and efficiency of privacy metrics.

A correlation analysis report with recommendations for replacing similar or costly metrics.

Case studies on the cut-off values of privacy metrics for the most representative scenarios.

A guide to best practices for privacy assessment that considers trade-offs between computation time and protection levels.

Thank you.

qianying.liao@kuleuven.be

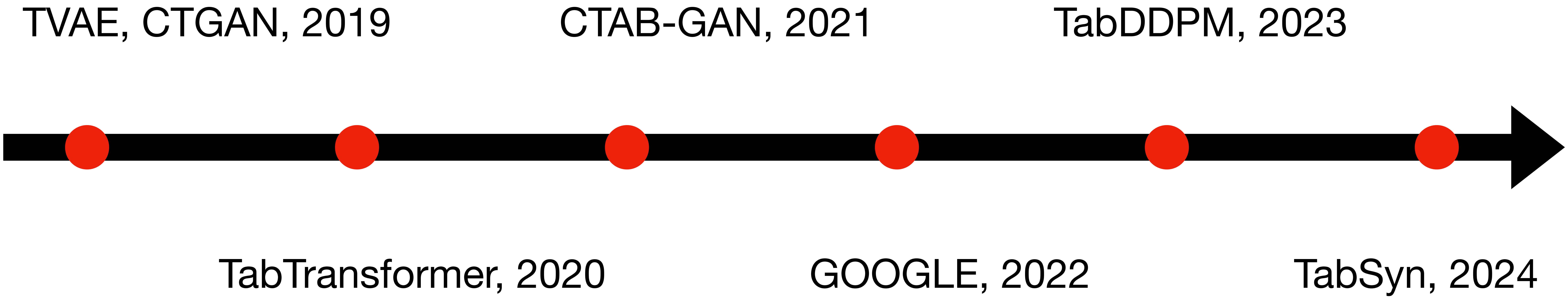
Relevant Literature on Privacy Evaluation

- Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2024). A unified framework for quantifying privacy risk in synthetic data, Proceedings of Privacy Enhancing Technologies Symposium.

- Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2024). Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data. *ACM Computing Surveys*, *57*(4), 1-38.

Seminal Synthetic Tabular Data Generation Approaches

Dev of GenAI for Tabular Data



TVAE, CTGAN, 2019

TabTransformer, 2020

CTAB-GAN, 2021

GOOGLE, 2022

TabDDPM, 2023

TabSyn, 2024

GenAI-based synthetic data generation is an active area of research, with new generators featuring greater generative capabilities published every year.