Privacy-preserving tabular data generation: Systematic Literature Review

Pablo Sanchez-Serrano, Ruben Rios and Isaac Agudo

Network, Information and Security (NICS) Lab, Universidad de Málaga

19th International DPM Workshop on Data Privacy Management, 2024

NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

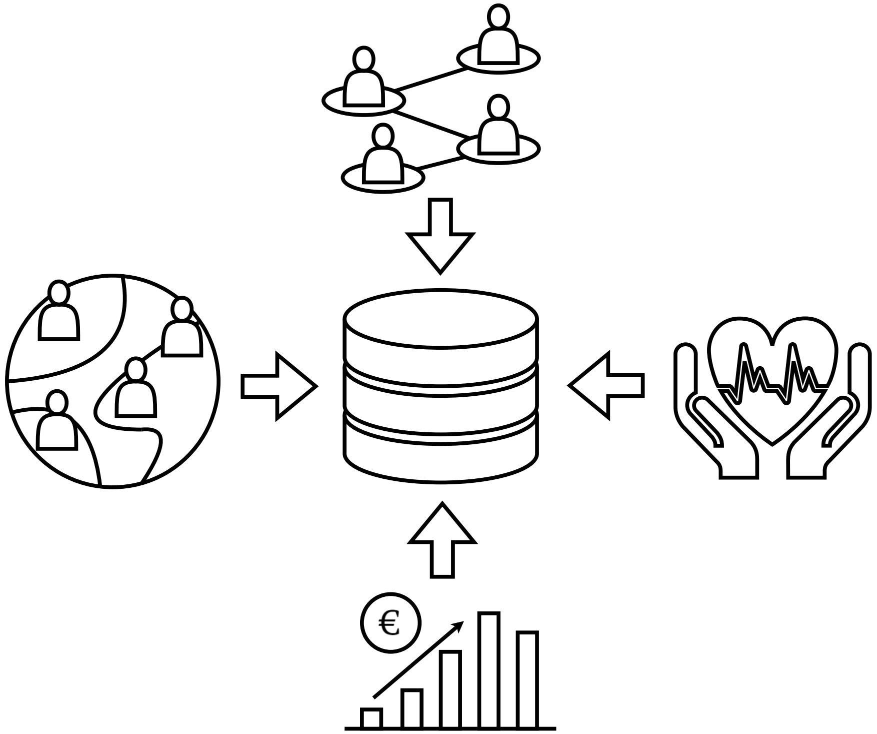
Agenda

- Introduction
- SLR methodology
- Tabular data generative models
- Conclusions

NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

The Need to Share



- Data hold immense value for **scientific, economic and social progress**
- The sharing of this data raises **privacy concerns**
- Traditional privacy techniques include:
  - k-anonymity
  - Differential Privacy
  - …

Generative models

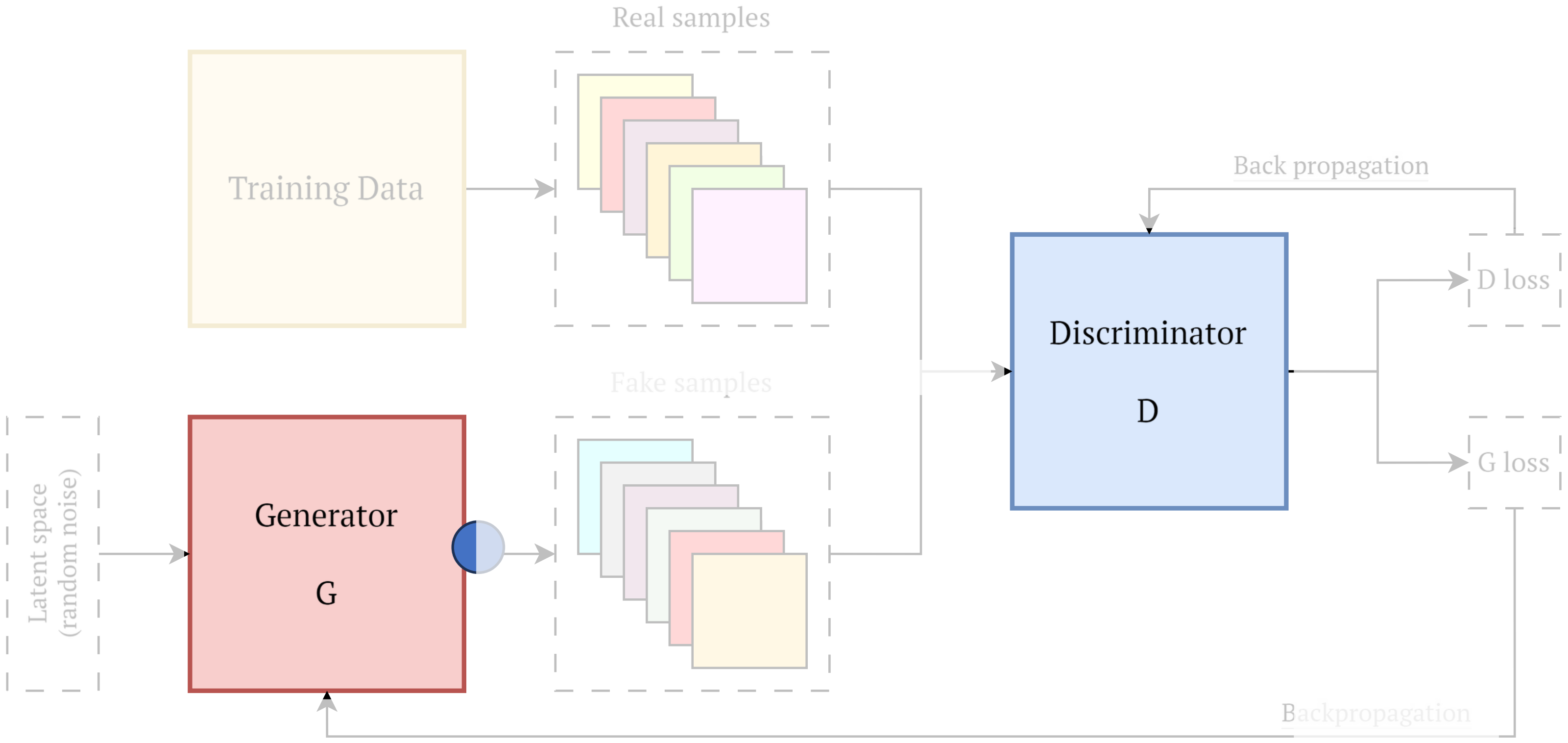
- Generative models have emerged as a **way to ensure the privacy** of tabular data

- Generative models **generate synthetic data** from real datasets
  - Images, video, text, tabular data, network traffic, etc

- There are many types of generative models:
  - Variational Autoencoders (VAEs)
  - Generative Adversarial Networks (**GANs**)
  - Recurrent Neural Networks (RNNs)
  - …

Generative Adversarial Networks (GANs)

- Two neural networks are trained in a zero-sum game setting:
  - **Generator (G)**: generates fake data that imitates a given dataset
  - **Discriminator (D)**: attempts to differentiate between real data samples and fake data samples generated by the generator

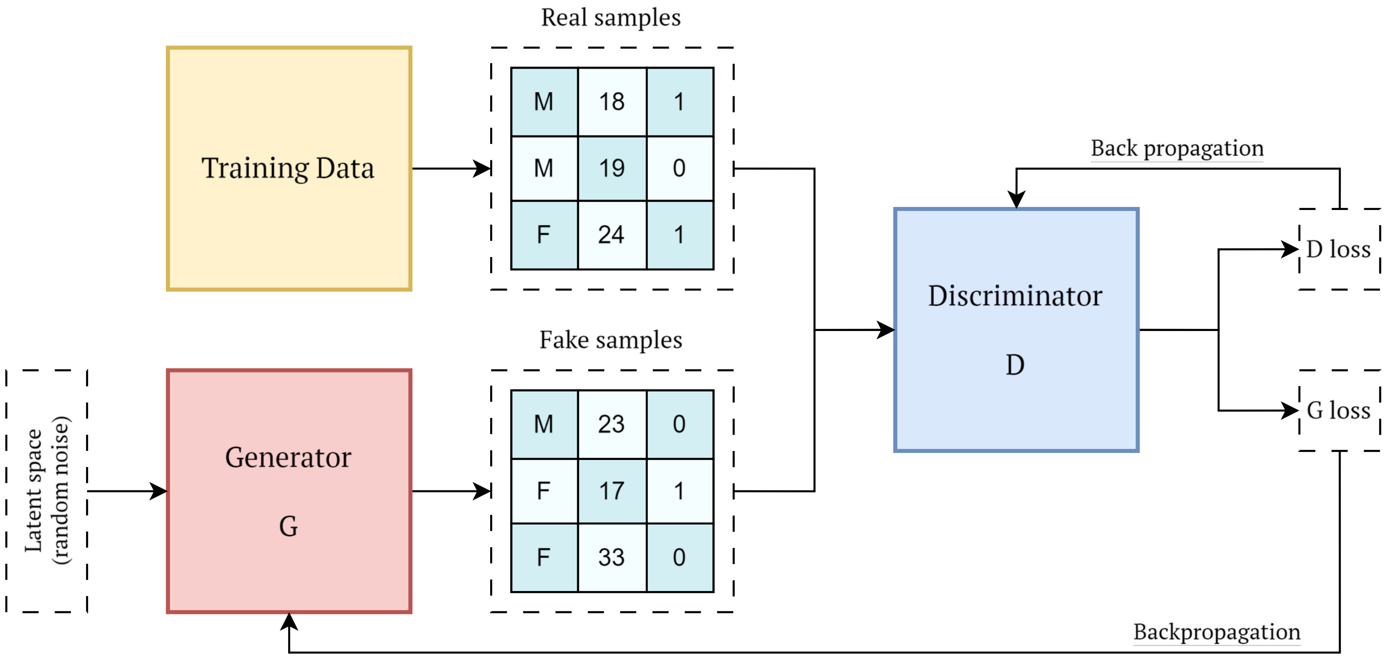


NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

Generative Adversarial Networks (GANs)

- There are several types of GANs
  - **Conditional** GANs: A condition can control the data generation process
  - **Deep Convolutional** GANs
  - …
- GANs can also be used to generate **tabular data**

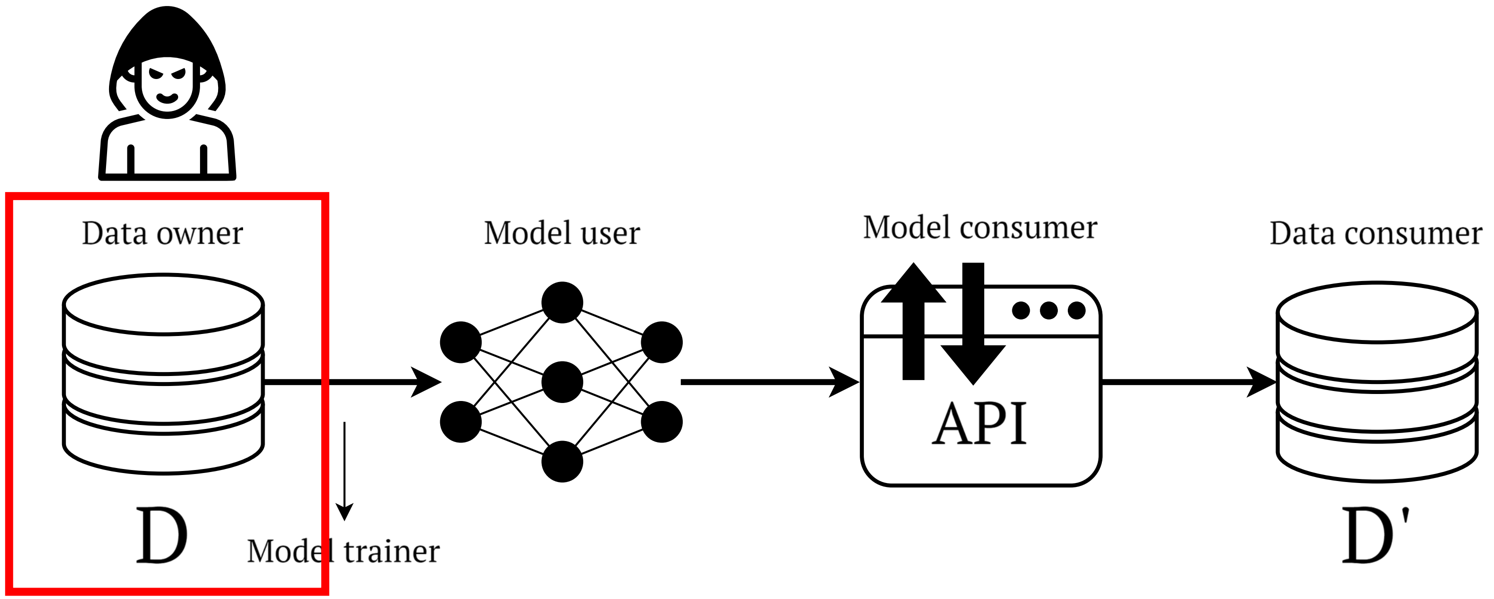


NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

Where is privacy involved? How do we measure it?

- When dealing with synthetic datasets, there are significant differences in the amount of knowledge and access available to different users
  - A range of privacy challenges need to be considered

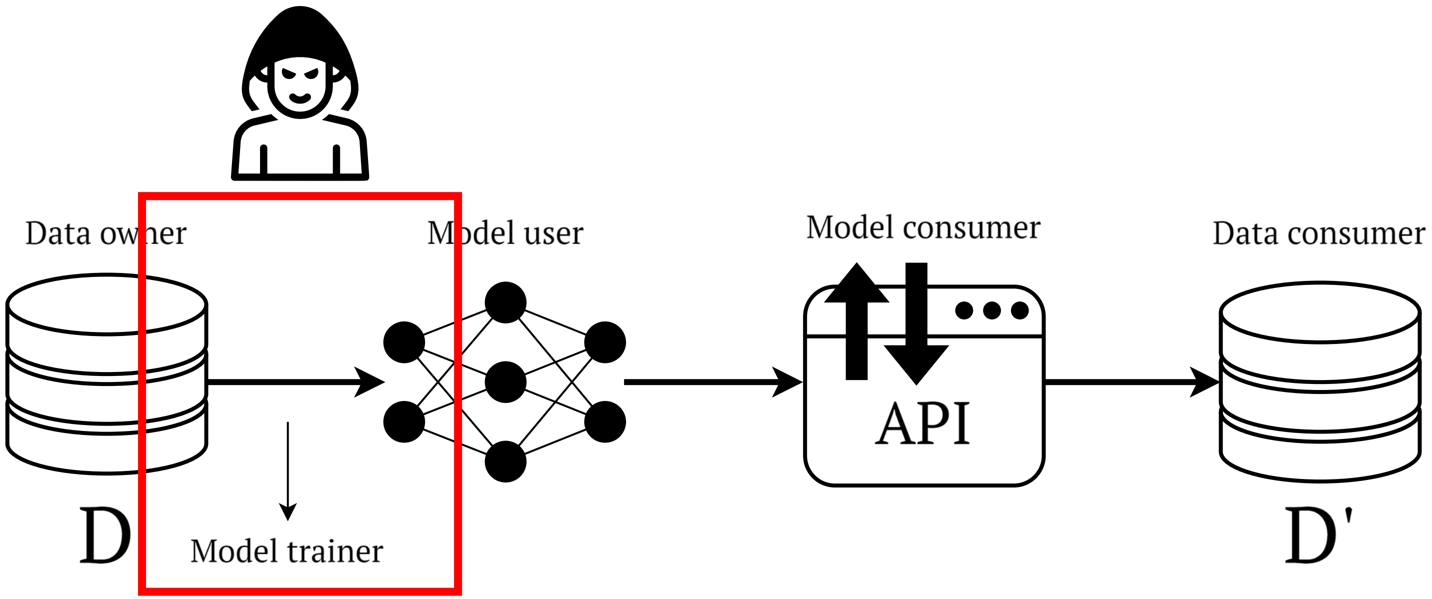


NICS
Network, information and security Lab

UNIVERSIDAD
DE MÁLAGA

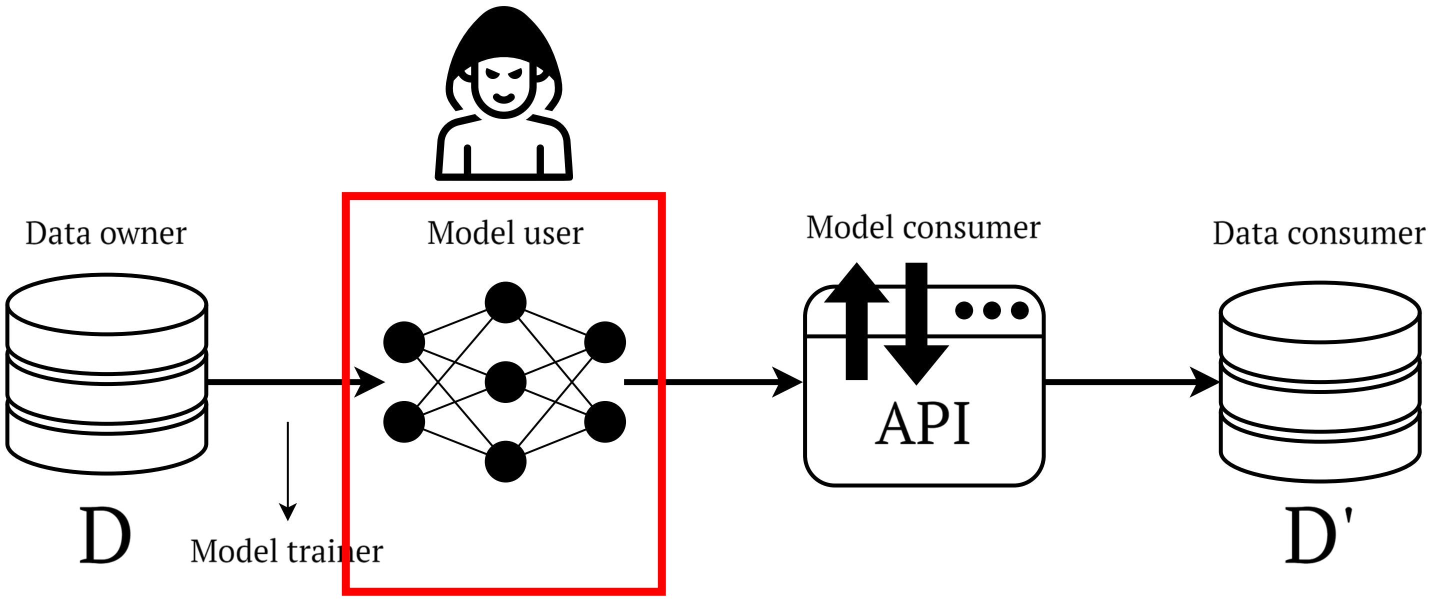
Where is privacy involved? How do we measure it?

- When dealing with synthetic datasets, there are significant differences in the amount of knowledge and access available to different users
  - A range of privacy challenges need to be considered



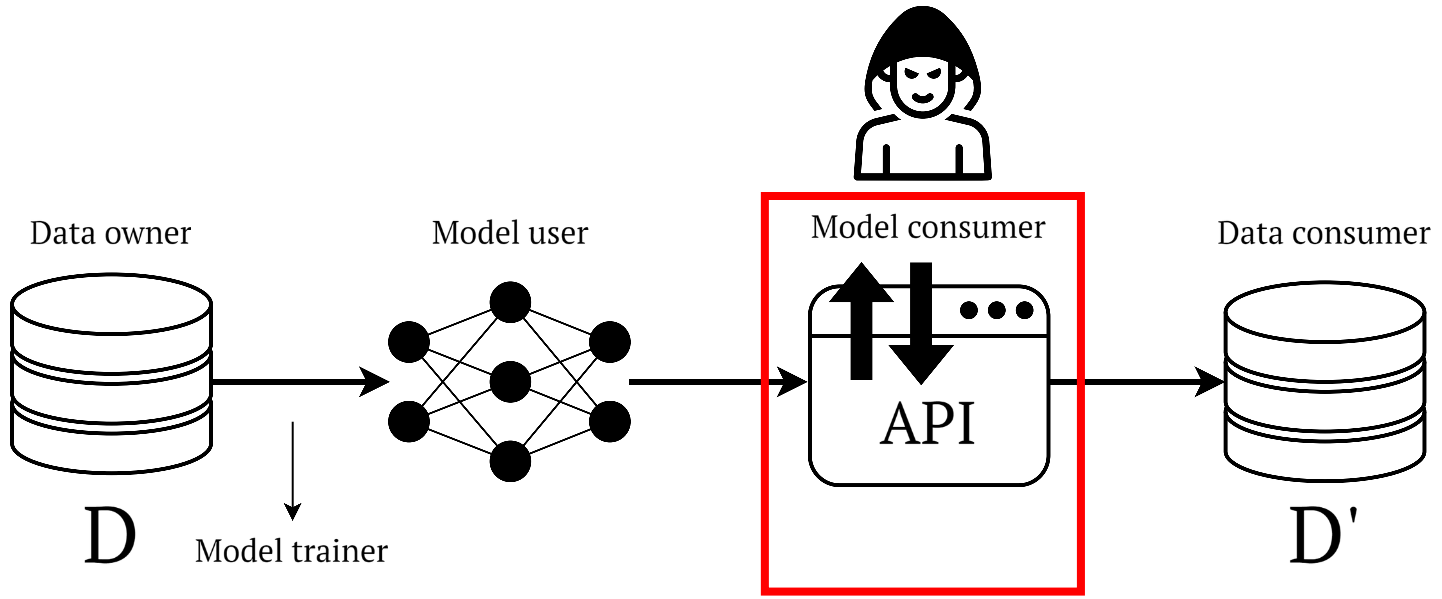
Where is privacy involved? How do we measure it?

- When dealing with synthetic datasets, there are significant differences in the amount of knowledge and access available to different users
  - A range of privacy challenges need to be considered



Where is privacy involved? How do we measure it?

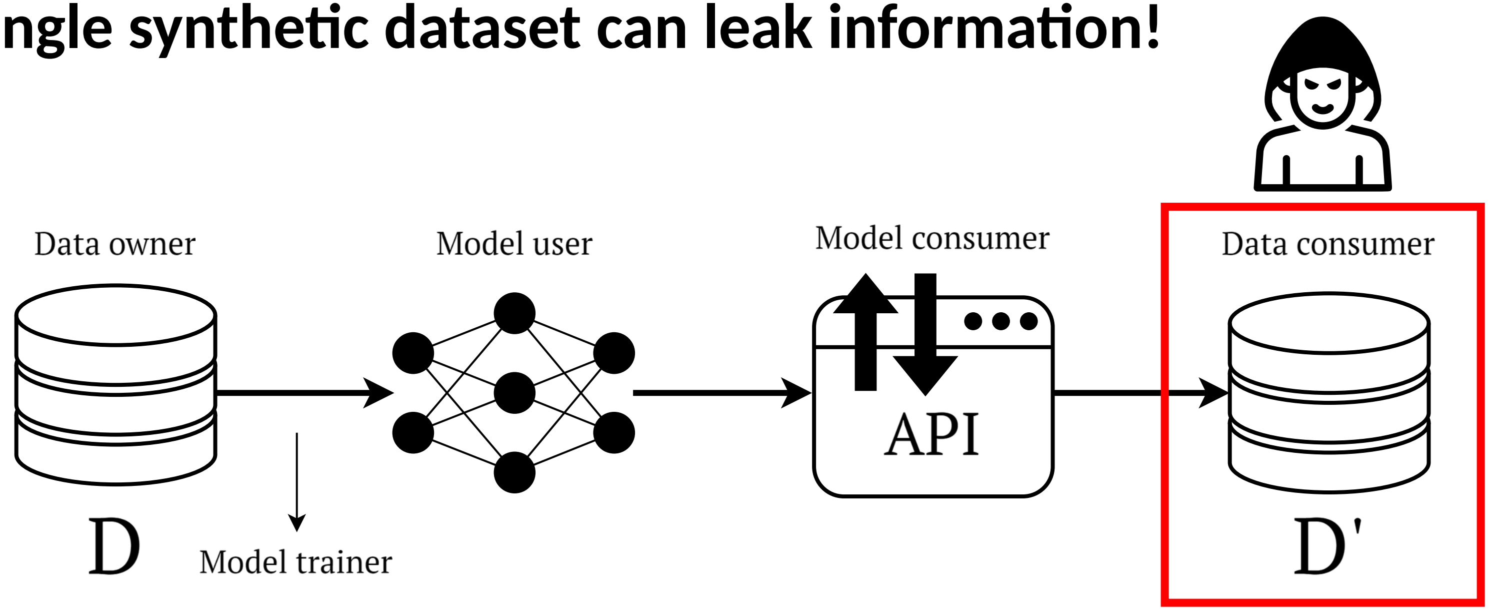
- When dealing with synthetic datasets, there are significant differences in the amount of knowledge and access available to different users
  - A range of privacy challenges need to be considered



Where is privacy involved? How do we measure it?

- When dealing with synthetic datasets, there are significant differences in the amount of knowledge and access available to different users
  - A range of privacy challenges need to be considered

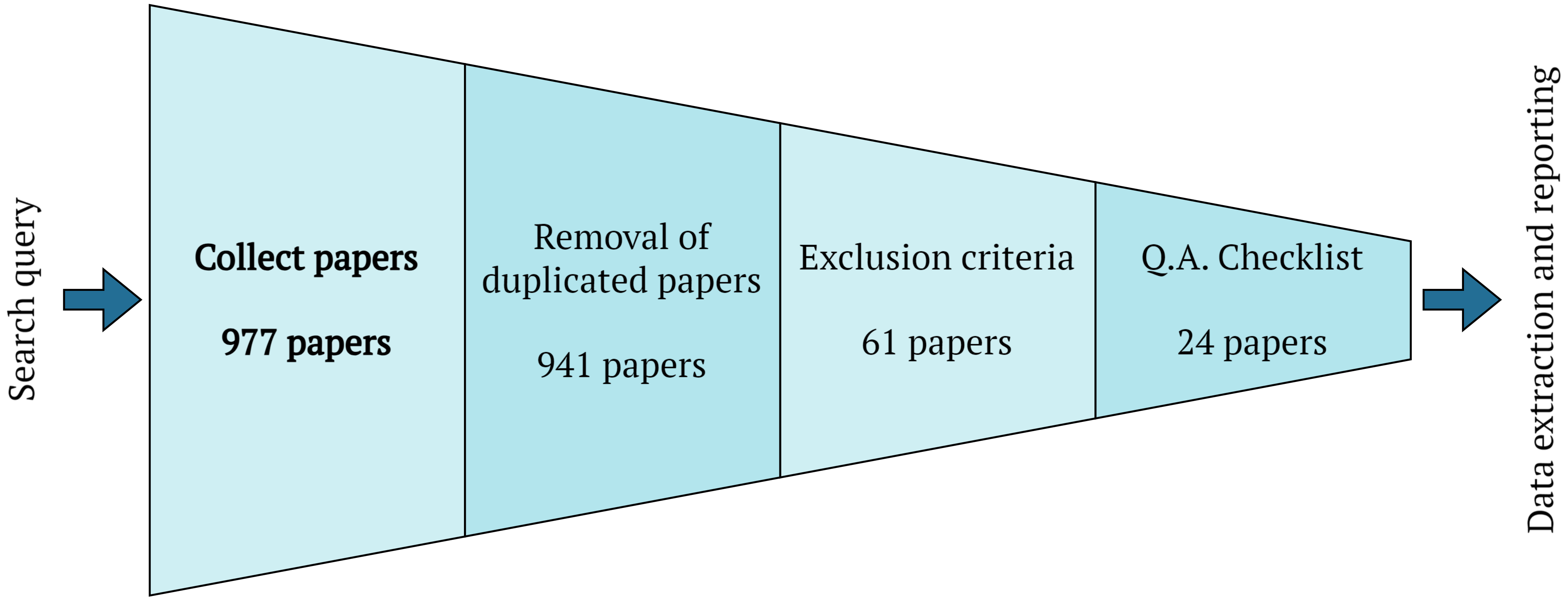
  - **A single synthetic dataset can leak information!**



Agenda

- Introduction
- SLR methodology
- Tabular data generative models
- Conclusions

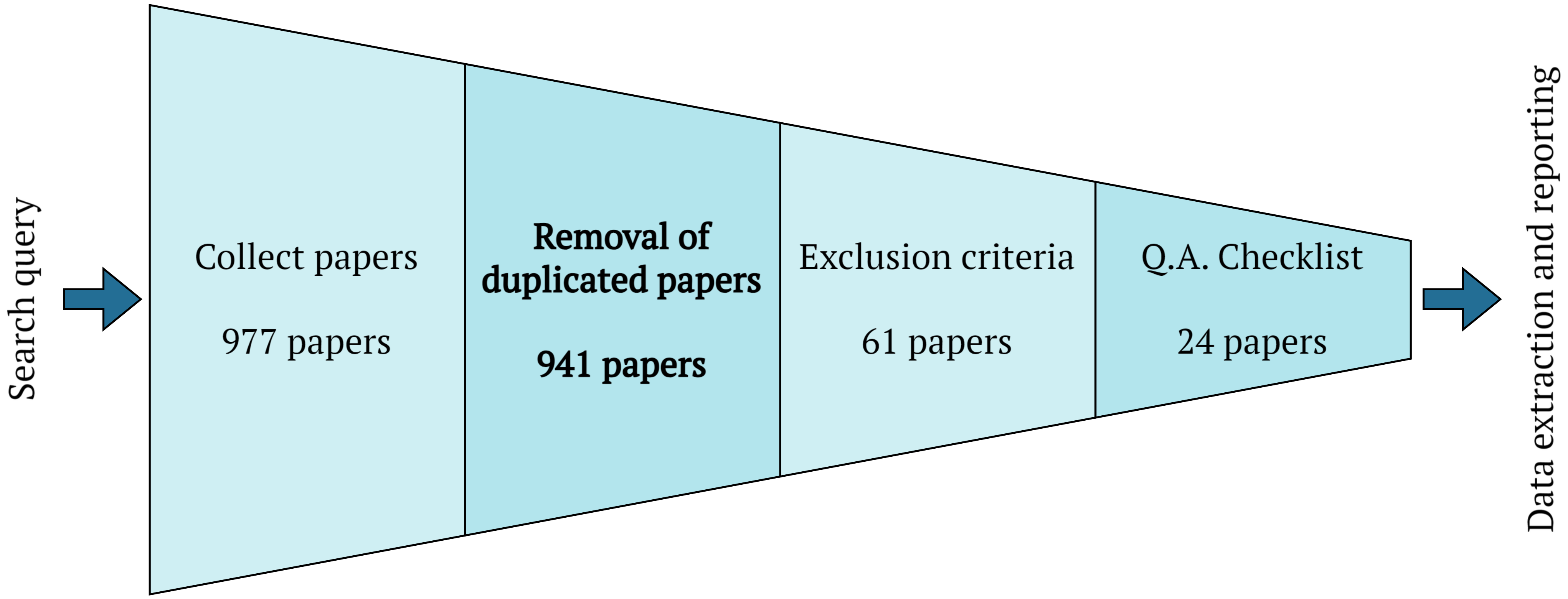
SLR: Methodology



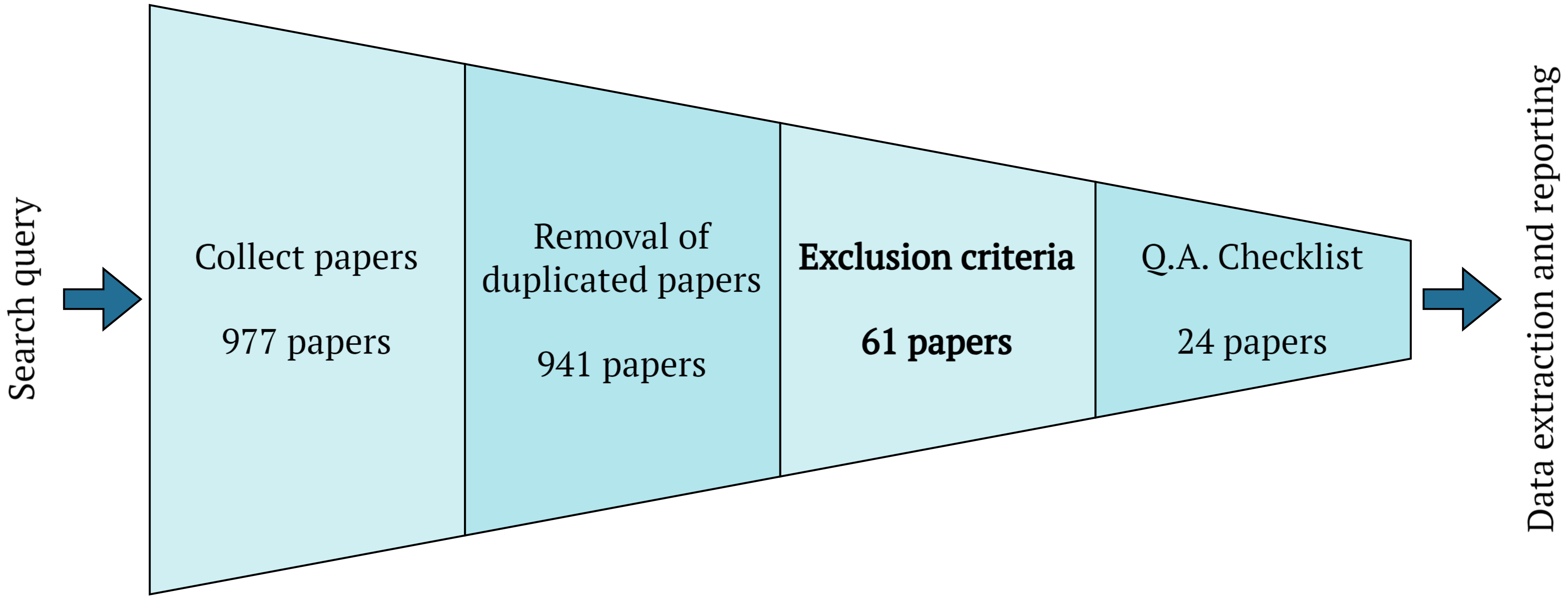
NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

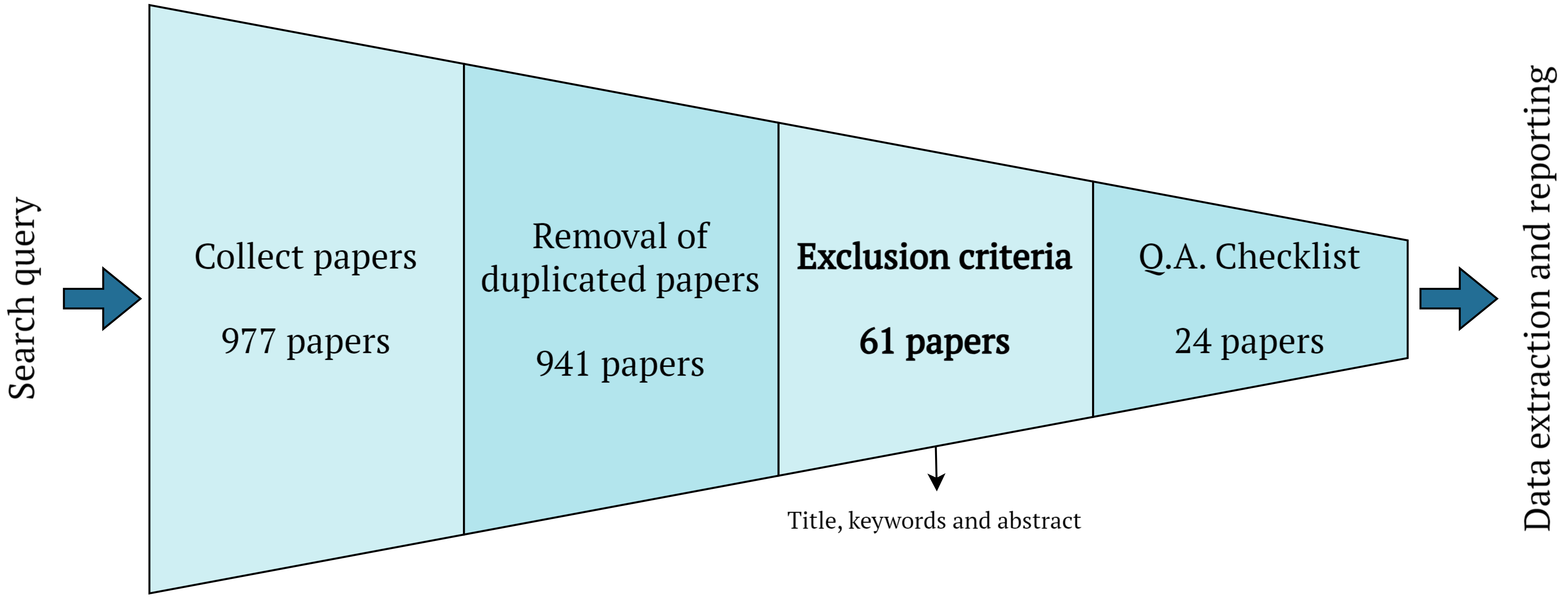
SLR: Methodology



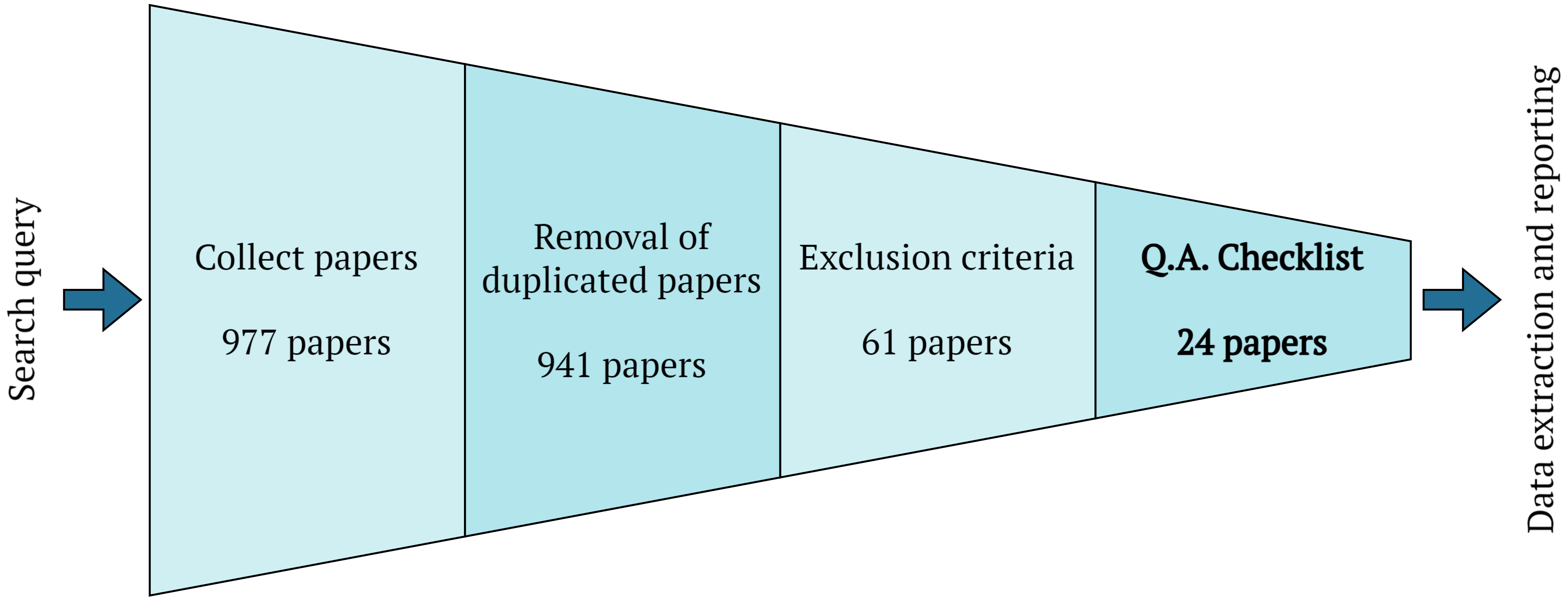
SLR: Methodology



SLR: Methodology



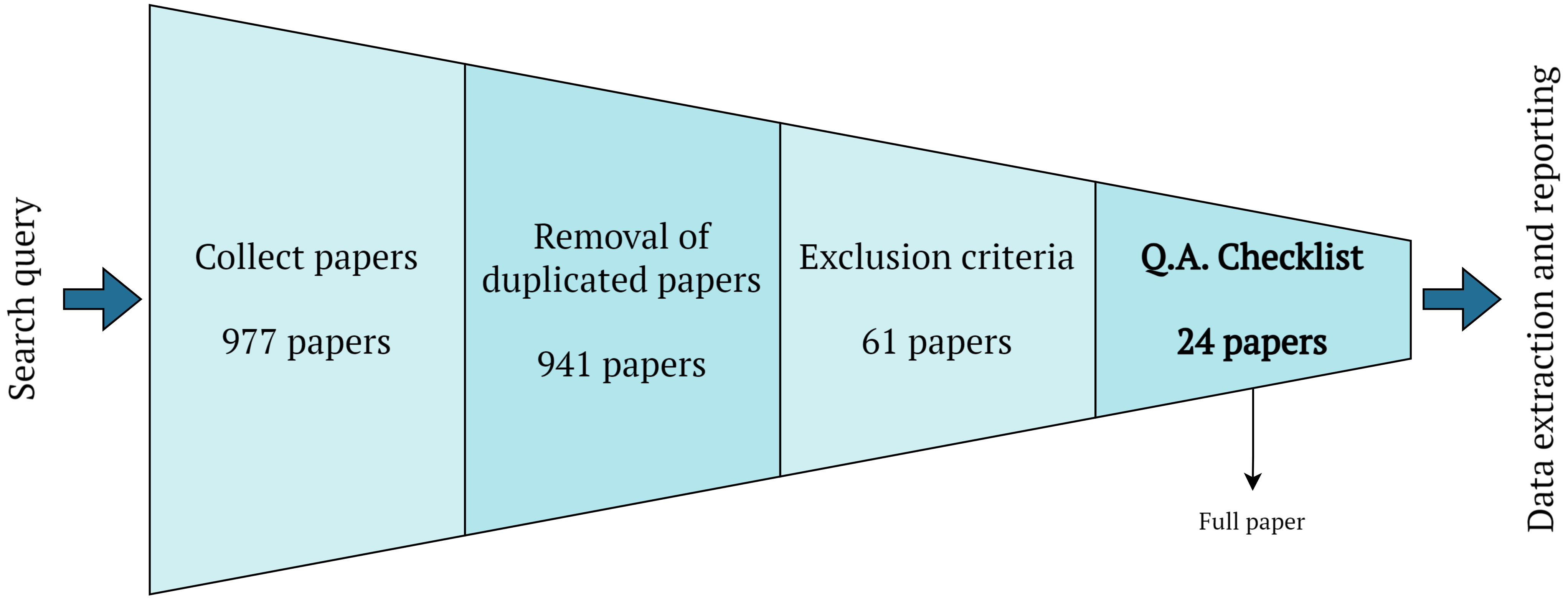
SLR: Methodology



NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

SLR: Methodology



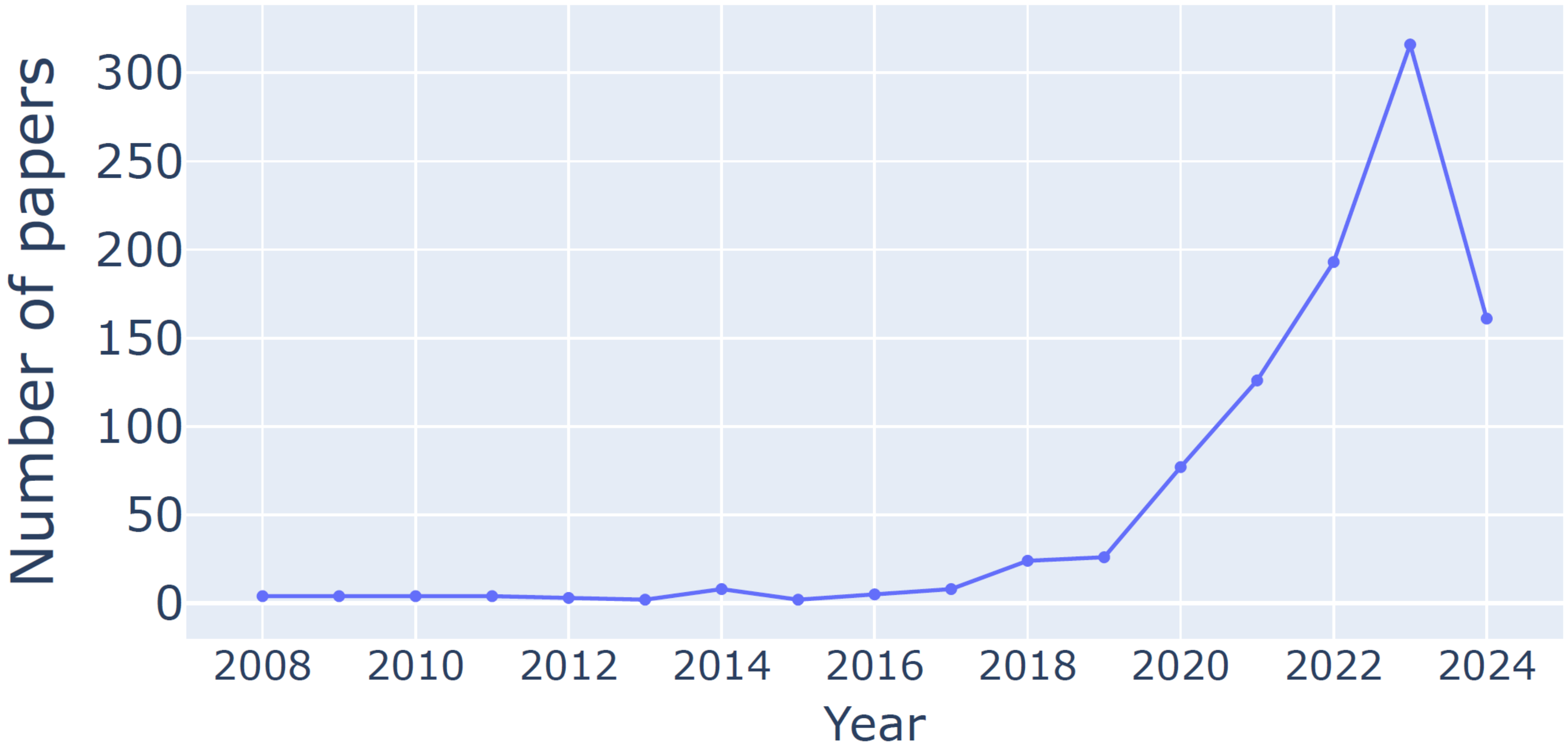
SLR: What questions do we want to answer?

- We aim to answer the following questions:
  - What are the main **techniques** used to guarantee **privacy** in generative models for tabular data?
  - How can we **measure the privacy** of generative models for tabular data?
- Following the PICOC approach, we came out with this query

("Tabular data" OR "Database" OR "Dataset") ***AND***
("Privacy techniques" OR "Data masking" OR "Differential privacy" OR "Masked data" OR "Privacy approach" OR "Privacy methods" OR "Privacy-preserving" OR "k-anonymity" OR "l-diversity" OR "t-closeness") ***AND***
("Generative model" OR "Data synthesis" OR "Synthesizer" OR "Synthetic data generation" OR "Synthetic generator") ***AND***
("Benchmark" OR "Privacy metric" OR "Anonymity metric" OR "Utility metric" OR "Data quality" OR "Data utility" OR "ML efficacy" OR "Usefulness of data")

SLR: Collected papers over the years

- There is an increase in the number of papers found over the years

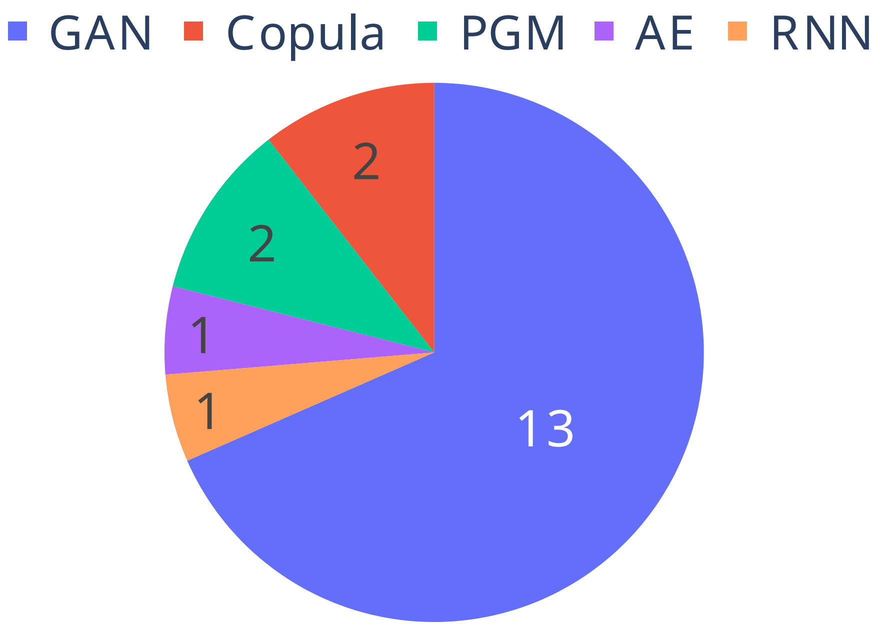


SLR: Quality Assessment Checklist

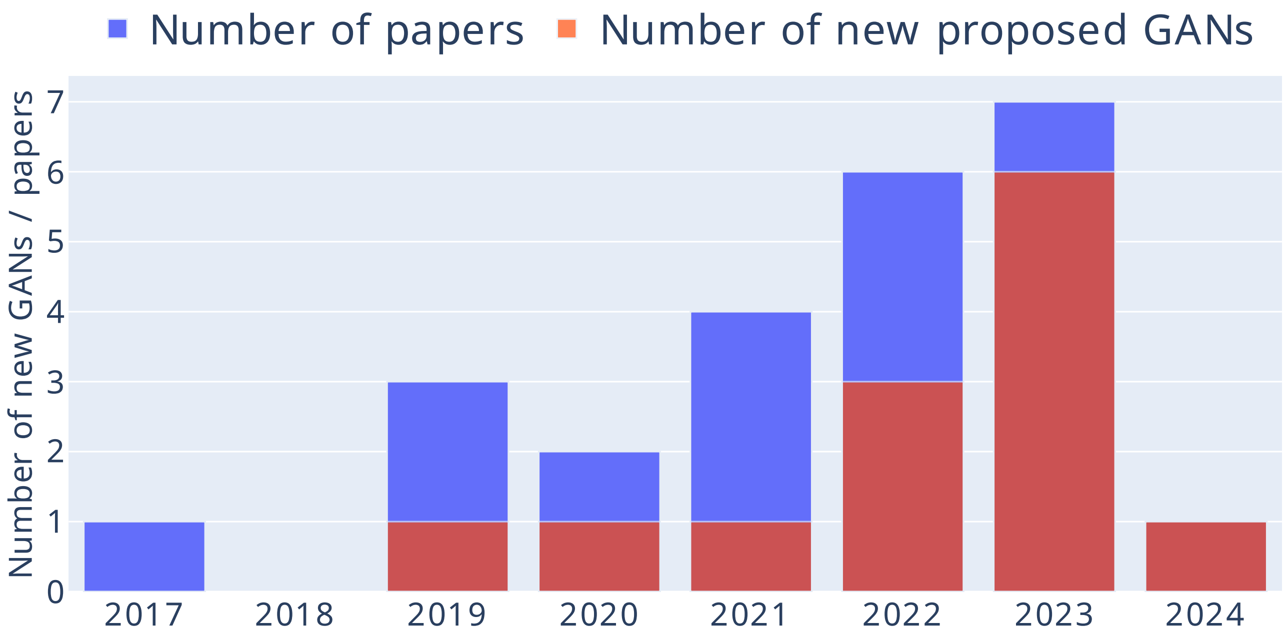
- After an initial filtering using the exclusion criteria, a checklist of questions is established
  - Does the article propose a **new AI model** for tabular data generation?
  - Does the article propose **new attacks** to privacy in generative models?
  - Does the paper propose a model **practical implementation**?
  - Does the model **include** techniques to provide **privacy**?
  - Does the article discuss how to **measure privacy** for tabular generative data models? Does it also include a way to **measure utility**?
- Papers with a minimum score of **3/5** are finally selected:
  - Yes (1 point), partially (0,5 points), no (0 points)

SLR: Reporting

Most contributions are based on GANs



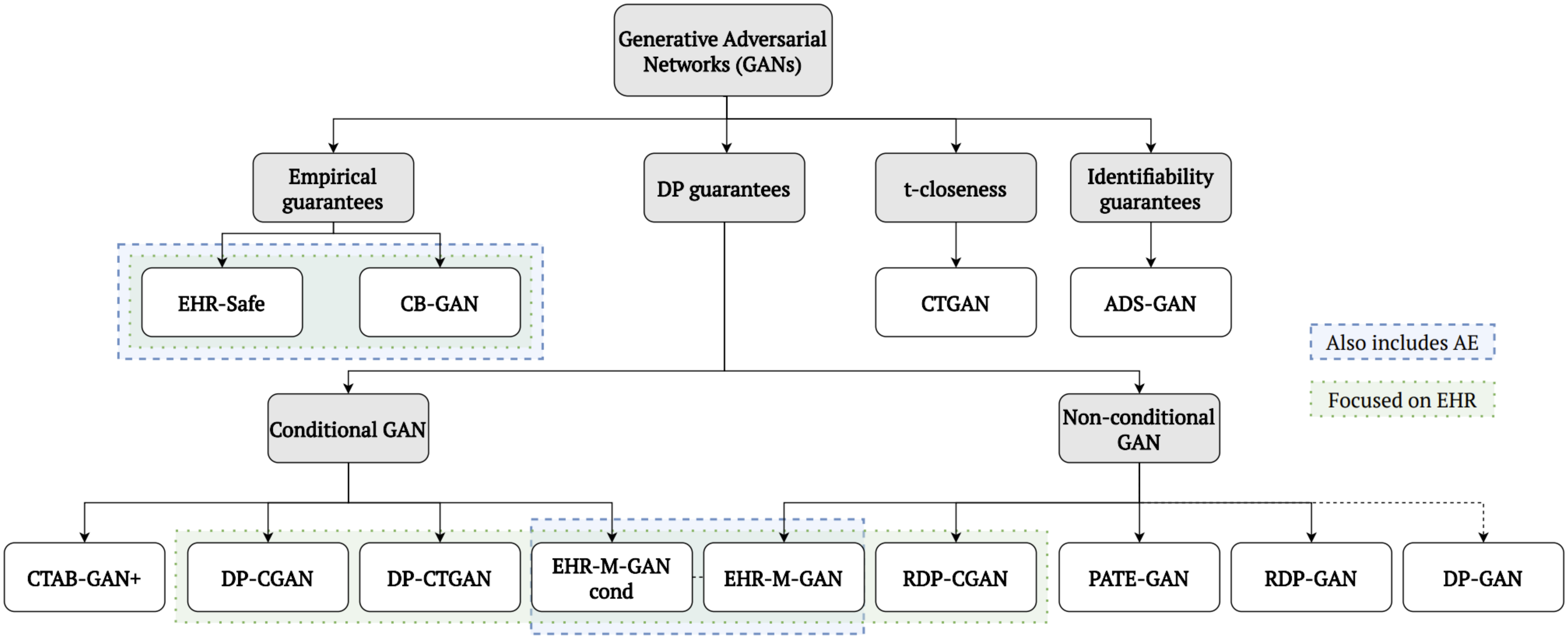
Exponential increase in last years



Agenda

- Introduction
- SLR methodology
- Tabular data generative models
- Conclusions

Privacy-preserving tabular data GAN taxonomy



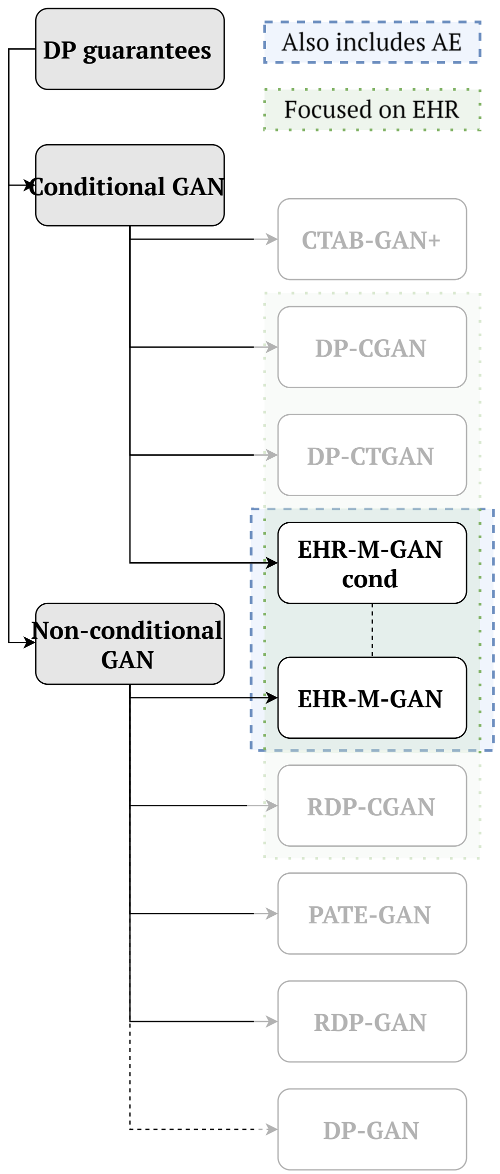
NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

Differential privacy GANs

Differential privacy: *"An algorithm, M, satisfies (ε ,δ)-differential privacy if for any pair datasets D, D' (which differs from D in only one entry)"*

- Most of the papers analyzed fall into this category
  - Approximately **half** of them use **conditional** GAN
- Most common scenario is **EHR** (Electronic Health Records)
- Two of them make use of **AEs** (Autoencoders)



NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

t-Closeness GANs

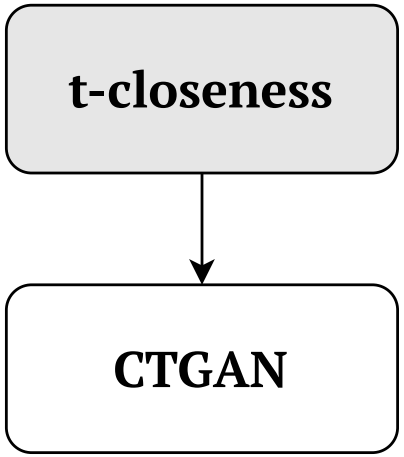
t-closeness: “*The distance between the distributions of a sensitive attribute in shared vs original dataset should be no more than a threshold t*”

- k-anonymity protects against **identity disclosure**
  - But it can not prevent **attribute disclosure**…
  - l-diversity is neither necessary nor sufficient to prevent it

- t-closeness is an extension of k-anonymity that solves this problem

- CTGAN generates samples until t-closeness requirements are met
  - Finally, sensitive attributes are encoded

- t-closeness and ε-differential privacy are strongly related



Identifiability guaranteed GANs

- **Identifiability** aims to measure and limit the risk of re-identification

- Synthetic samples should be “**different enough**” from the original elements
  - ADS-GAN use **weighted Euclidean distance**

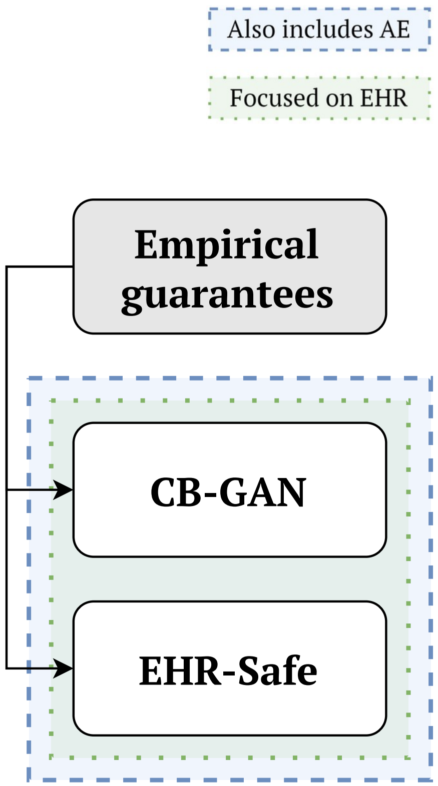
- ε-identifiability: percentage of non-identifiable patients
  - 0-identifiability → perfectly non-identifiable dataset
  - 1-identifiability → fully identifiable dataset

Identifiability guarantees

↓

ADS-GAN

Empirically guaranteed GANs



- Some models do **not focus on theoretical privacy guarantees**
  - Rather focus on an **empirical** approach to measure privacy
- Instead of proving privacy properties, they measure resistance to known attacks
  - The most common is the **Membership Inference Attack (MIA)**
- Similar to how car crash tests work…

Agenda

- Introduction
- SLR methodology
- Tabular data generative models
- Conclusions

Conclusions

- We provide an overview of the **state of the art** in privacy-preserving tabular data generation

- The SLR produced 24 papers to **answer two research questions**
  - GAN is the predominant model for synthetic tabular data
  - Most models focus on providing differential privacy guarantees
    - Some models does not theoretically guarantee privacy

- Future work:
  - Design a common framework to **evaluate** the models
  - Identify **other generative models** that can incorporate privacy guarantees

NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA

Privacy-preserving tabular data generation: Systematic Literature Review

Pablo Sanchez-Serrano, Ruben Rios and Isaac Agudo

Network, Information and Security (NICS) Lab, Universidad de Málaga

19th International DPM Workshop on Data Privacy Management, 2024

NICS
Network, information and security Lab

UNIVERSIDAD
DE MÁLAGA

SLR: PICOC terms and keywords

- We define PICOC terms. They help to define a list of keywords:

Keywords	Synonyms	PICOC
Tabular data	Database, Dataset	Population
Privacy techniques	Data masking, Differential privacy, Masked data, Privacy approach, Privacy methods, Privacy-preserving, k-anonymity, l-diversity, t-closeness	Intervention
Generative model	Data synthesis, Synthesizer, Synthetic data generation, Synthetic generator	Comparison
Benchmark	-	Outcome
Privacy metric	Anonymity metric	Outcome
Utility metric	Data quality, Data utility, ML efficacy, Usefulness of data	Outcome

SLR: Exclusion criteria

- To refine the search and ensure the inclusion of high-quality and relevant studies, the following exclusion criteria are applied:
  - The paper **does not discuss privacy**
  - The paper **does not discuss AI**
  - The paper **does not focus on tabular data**
  - It is a **survey/review**
  - It is not an article, conference paper, proceeding or journal
  - It has **not enough citations**
    - Papers published before 2022 must have at least 20 citations
    - Papers from 2022 must have at least 10 citations
    - Papers from 2023 or 2024 must have at least 5 citations
  - It is **not** published in **English**

Tabular data generative models

- Other type of models were found:
  - **Autoencoders (AE)**: DP-SYN
    - Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., & Sweeney, L. (2019). Privacy preserving synthetic data release using deep learning.
  - **Probabilistic Graphical Models (PGMs)**: PrivMRF and PrivIncr
    - Cai, K., Lei, X., Wei, J., & Xiao, X. (2021). Data synthesis via differentially private markov random fields.
    - Liu, G., Tang, P., Hu, C., Jin, C., Guo, S., Stoyanovich, J., ... & Mühlig, J. (2023). Multi-Dimensional Data Publishing With Local Differential Privacy
  - **Recurrent Neural Networks (RNNs)**: Conditional-LSTM
    - Mosquera, L., El Emam, K., Ding, L., Sharma, V., Zhang, X. H., Kababji, S. E., ... & Eurich, D. T. (2023). A method for generating synthetic longitudinal health data
  - **Copula-based models**: LoCop and DR_LoCop
    - Wang, T., Yang, X., Ren, X., Yu, W., & Yang, S. (2019). Locally private high-dimensional crowdsourced data release based on copula functions

NICS
Network, information and security Lab

UNIVERSIDAD DE MÁLAGA