# SYNTHEMA + ERN-EuroBloodNet

## Joint Training Programme on Synthetic Data Generation in SCD and AML

Funded by the European Union

# Introduction to SYNTHEMA & synthetic data in healthcare

Rudolf Mayer, SBA Research, Vienna, Austria

# Overview for this Session

What is personal data?

Disclosure risks in data sharing & analysis

Mitigations

# GDPR: Definition of Personal Data

Loan default dataset

| SSN | Name | Birthdate | Sex | Salary |
|---|---|---|---|---|
| 1541 | Tom | 02.03.1995 | Male | €3 950 |
| 7842 | Alex | 03.04.2006 | Male | €3 950 |
| 3651 | Tanja | 01.02.1994 | Female | €3 720 |

- **Personal data:** ”any information **relating** to an **identified or identifiable** natural person“ (Article 4(1) GDPR)
- Also, data which “could be attributed to a natural person by the use of additional information” (Recital 26 GDPR)

| Pseudonym | Birthdate | Sex | Salary |
|---|---|---|---|
| 5t4o55 | 02.03.1995 | Male | €3 950 |
| 8A4le4 | 03.04.2006 | Male | €3 950 |
| 45T23a | 01.02.1994 | Female | €3 720 |

| Pseudonym | SSN |
|---|---|
| 5t4o55 | 1541 |
| 8A4le4 | 7842 |
| 45T23a | 3651 |

➔ **Pseudonymised data is personal data**

➔ **Anonymised data is not personal data**

- Machine learning model: aggregated data (relation to individuals should be cut off)

➔**Aggregated data is not personal data**

➔ **Where is the boundary?** Relation to individuals/identification theoretically impossible vs. practically impossible

# GDPR: Definition of Personal Data

- "To determine whether a natural person is **identifiable**, account should be taken of **all the means reasonably likely to be used** […] either by the controller or by another person **to identify the natural person** directly or indirectly. To ascertain whether means are **reasonably likely to be used** to identify the natural person, account should be taken of **all objective factors**, such as the **costs** of and the amount of **time** required for identification, taking into consideration the available **technology** at the time of the processing and **technological developments**. The principles of data protection should therefore not apply to anonymous information […]"
  (Recital 26 GDPR)

- **Reasonably unlikely is something that will not happen in practice** (measured by difficulty vs. potential gain)

# General threats to privacy in published data

- Identity disclosure (re-identification):
  - individual can be linked to a specific data entry

    *-> Implies attribute and membership disclosure!*

- Attribute disclosure
  - Requires: knowing values of some attributes of a record
  - May be achieved even without linking to a specific item in a dataset
  - Discloses sensitive attributes from the dataset with which individuals are not willing to be linked with, e.g. the salary of a person

- Membership disclosure
  - Inference allows an attacker to determine whether data about an individual is contained in a dataset
  - Does not directly disclose any information from the dataset itself → but allows an attacker to infer meta-information that could be sensitive (i.e. implicit sensitive attribute = attributes globally true for all/most records in the dataset)

**Loan default dataset**

| Pseudonym | Birthdate | Sex | Salary |
|---|---|---|---|
| 5t4o55 | 02.03.1995 | Male | €3 950 |
| 8A4le4 | 03.04.2006 | Male | €3 950 |
| 45T23a | 01.02.1994 | Female | €3 720 |

Tanja

| Name | Sex | Salary |
|---|---|---|
| Tom | Male | ? |

| Salary |
|---|
| €3 950 |

| Birthdate | Sex | Salary |
|---|---|---|
| 02.03.1995 | Male | €3 950 |
| 03.04.2006 | Male | €3 950 |
| 01.02.1994 | Female | €3 720 |

# General threats to privacy in published data

## GDPR Article 29 Working Party

| Name | Birthdate | Sex | Salary |
|---|---|---|---|
| Tom | **02.03.1995** | **Male** | €3 950 |
| Alex | **03.04.2006** | **Male** | €3 950 |
| Tanja | **01.02.1994** | **Female** | €3 720 |

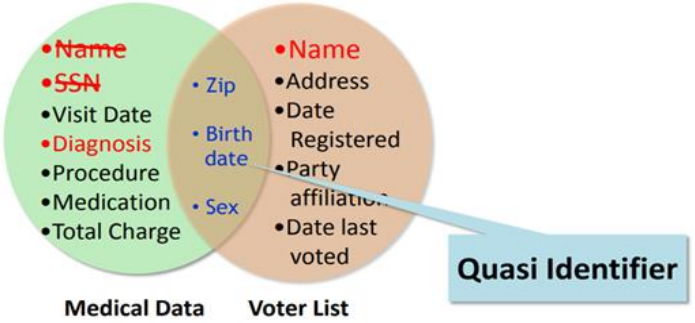
The following three criteria are proposed

- **Singling out**: the possibility of isolating some or all records that identify an individual in the dataset
  - **Enables** the isolation of a data unit
  - Providing control or **facilitating** other privacy attacks
- **Linkability**: the ability to link at least two records concerning the same data subject or a group of data subjects
  - Possible through singling out
  - Might **enable re-identification/identity disclosure** ➔ if the other record contains identifiable information (e.g. the name, or social insurance number, …)
- **Inference**: the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes (“attribute disclosure”)

- Confidentiality attacks also mentioned in **AI Act, Article 15 Cybersecurity**

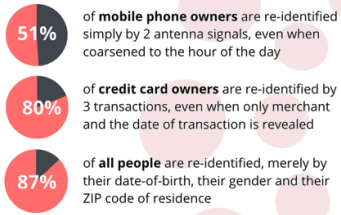
# Linkability: Example

- Match records from two databases togeth
  - Matches on **quasi-identifiers** (e.g. birthdate, ZIP code, sex, ...)



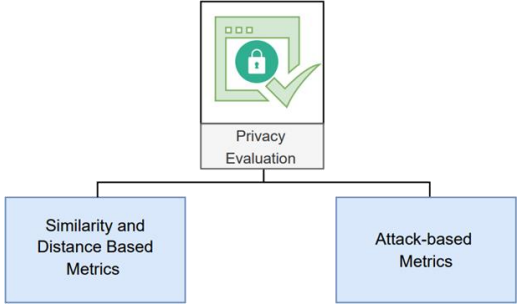

- Second dataset contains **personally identifiable information (PII)**

➔ Learn the identity of a record ➔ re-identification/identity disclosure!
  - Enabled via singling-out and linkability

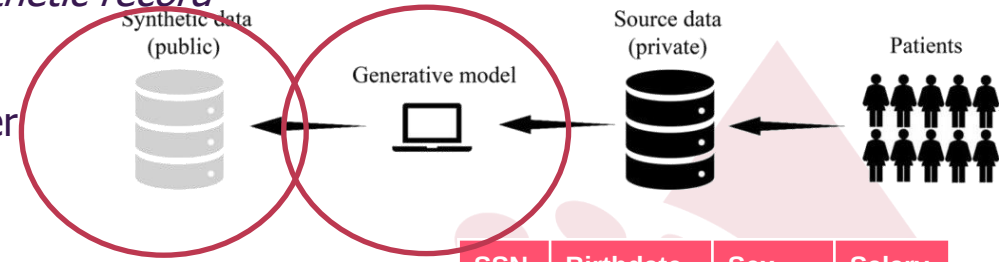



IMDb

# Disclosure risks in Synthetic Data




- Synthetic data: slightly different setting than before
- *Original data is the input to training a synthetic data generator*
- *We work with the synthetic data thereafter*
- *No 1:1 connection from original to synthetic record*




- Q: Which of the disclosure risks transfer to synthetic data? How?
  - Singling out
  - What is linkage / re-identification
  - Inference (attribute & membership)

  - What does the adversary have available?
    - Generator model? Generated data?

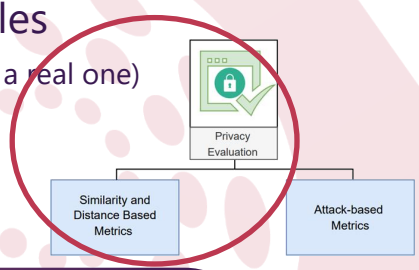
| SSN | Birthdate | Sex | Salary |
|---|---|---|---|
| 4216 | 01.02.1993 | Female | €4 050 |
| 3214 | 05.06.2007 | Male | €3 750 |
| 8463 | 08.07.1996 | Female | €3 850 |

# Disclosure risks in Synthetic Data

## Singling out / Linkability

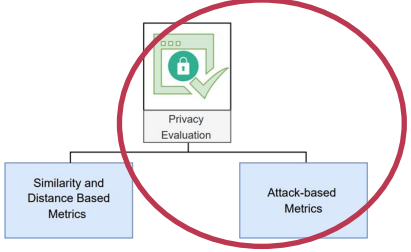
| SSN | Birthdate | Sex | Salary |
|---|---|---|---|
| 4216 | 01.02.1993 | Female | €4 050 |
| 3214 | 05.06.2007 | Male | €3 750 |
| 8463 | 08.07.1996 | Female | €3 850 |
| **3651** | **01.02.1994** | **Female** | **€3 720** |

- Singling out is (of course) possible
  - It is a (kind-of) necessary condition to perform linkage
  - But by itself not yet a disclosure!
- Linkability
  - Link record to other, externally available data, which contains **personally identifiable information (PII)**
  - How do we link a synthetic record with non-accurate value to real data?
  - Alternative: measure how close synthetic data is to real examples
    - Identify **exact matches** (a synthetic record has the exact same values as a real one)
    - **Distance to closest** real **record**, **outlier similarity**, …
- *Overall: no consensus if that constitutes a disclosure*



# Disclosure risks in Synthetic Data

## Inference (**attribute**, membership)

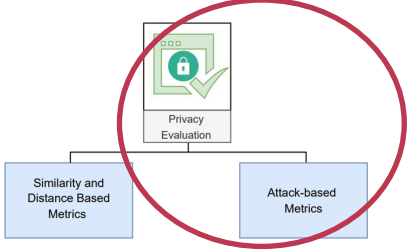



- Attribute inference: estimate value of a missing attribute

| Birthdate | Sex | Salary |
|---|---|---|
| 02.03.1995 | Male | €3 950 |
| 03.04.2006 | Male | €3 950 |
| 01.02.1994 | Female | ? |

- Attacker has available either
  1. Generated synthetic data records
     ➔ Utilise that to make a statistical prediction of the likely value
     - Simple mean/median, neighbourhood averages, machine-learning models, …
  2. Trained generator model
     ➔ Generate synthetic samples, ➔ 1.
     ➔ Otherwise use the model for more information on the records
- Attack is easy to perform, correctness measurable (but; what is exact/similar?)
- Success is not that easy to put in perspective: how good could we guess the values w/o the synthetic data / generator?

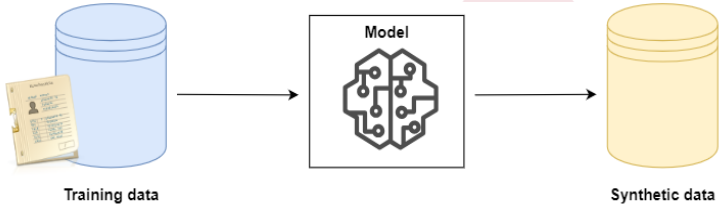
# Disclosure risks in Synthetic Data



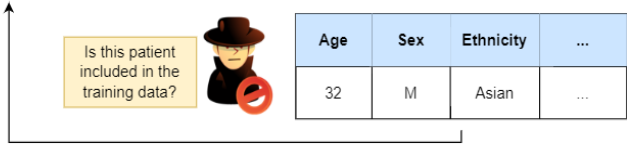

## Inference (attribute, **membership**)

- Membership inference: decide whether a (real) record was part of a synthetic data generation training process
- Disclosure: can infer meta-information from the dataset (e.g. a specific disease)
- Attacks utilise the phenomenon that machine learning models overfit – they learn better the data that they have seen during training, than other data

- Attacker can train a similar synthetic data generator, and use it to learn patterns of (non) membership
- Attack is relatively costly, requires knowledge




- Interpretation: when is membership inference an issue?
  - If the attacker is correct 2/3 of the time? 90%? …



# Mitigations to disclosure risks in synthetic data

- Operational / deployment
  - Limit access to generator model, to generated data, …
    ➔ Might limit user experience
  - Filter out records at risk from training data (e.g. outliers), …
  - Filter out records at risk from generated data, …
    ➔ Might reduce quality / utility of generated data

- Model training: make the synthesizer differentially private
  - Main idea: add noise / imprecision to make disclosure more difficult

# Differential Privacy

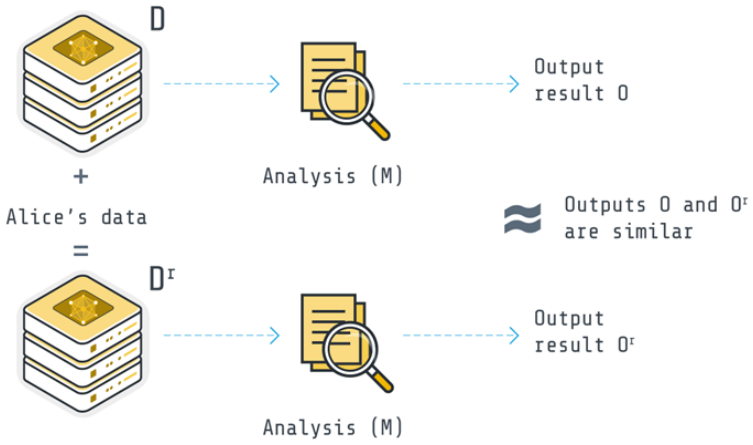
## Definition

Differential Privacy (DP) is **a formal mathematical framework** that ensures statistical analysis does not compromise individual privacy.

- Adds noise to queries or training process
- Controlled by privacy parameter ε --> smaller ε = stronger privacy

**DP guarantee: presence or absence of any individual has a minimal impact on the output of the analysis.**

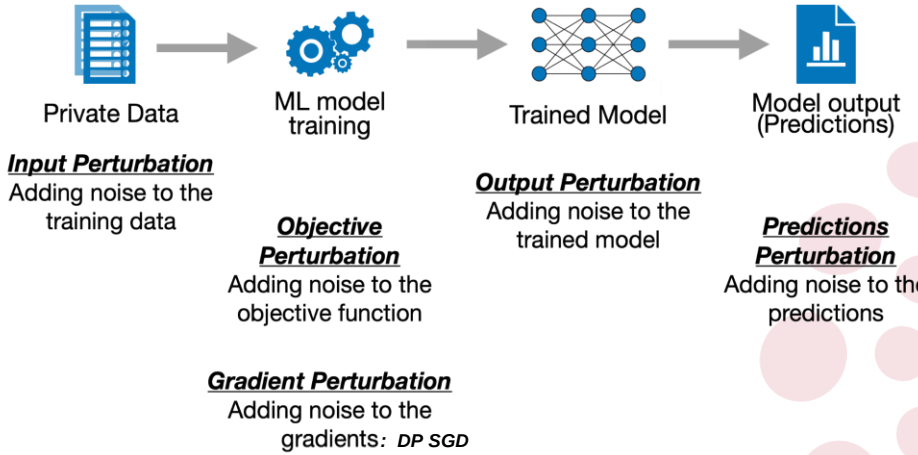
E.g. *percentage of smokers in the room? --> very similar output whether we exclude from the analysis any person from the room.*



# Differential Privacy

## Application

- Widely applied in Machine Learning
- Achieved by adding a controlled amount of noise to the data, model parameters or output
- Cost: computational overhead, reduced data quality / utility



# Summary

- Data publishing carries disclosure risks (identity, attribute, membership / singling out, linkability, inference)
- Synthetic data removes the 1:1 connection from original to sanitized record
  - But: that **does not solve all problems**
  - Need to still consider inference attacks, maybe re-identification
  - Still very active discussion on impact, legal status, etc..
  - Mitigations: e.g. differential privacy, output treatments, ..

- SYNTHEMA outputs: toolbox for privacy assessment & DP
  - https://github.com/synthema-project

- Further reading: *Toward practical anonymity: A white paper on privacy risk, metrics, and governance in synthetic data*
- https://ieeexplore.ieee.org/document/11197237

# Thanks!

## Any questions?

**Keep in touch!** rmayer@sba-research.org https://www.linkedin.com/in/rudolf-mayer/

eurobloodnet.eu /ERNEuroBloodNet @ERNEuroBloodNet @erneurobloodnet.bsky.social

synthema.eu /synthema @SYNTHEMA_EU @synthemaeu.bsky.social

Funded by
the European Union

# Acknowledgements

European Reference Network
for rare or low prevalence complex diseases
Network Hematological Diseases (ERN EuroBloodNet)

Funded by the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN-EuroBloodNet)-Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

SYNTHEMA

Funded by the European Union

SYNTHEMA is an initiative funded by the European Union's Horizon Europe Research and Innovation programme under grant agreement No. 101095530.