# Managing & Regulating Privacy Risks in Synthetic Data

Khaled El Emam & Lucy Mosquera

March 30, 2022

# Acknowledgements


This work wouldn’t have been possible without the thoughtful contributions of Anita Fineberg and Elizabeth Jonker

# Agenda

**Introduction to Synthetic Data**

1 General description of what synthetic data is and how it's used as a privacy enhancing technology

**Defining Privacy Risks in Synthetic Data**

2 An overview of state-of-the-art ways to measure privacy risks in synthetic data

**Regulation of Synthetic Data**

3 Overview of the Canadian regulatory landscape for synthetic data based on a review of current legislation and interviews with regulators

Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# Synthetic Data

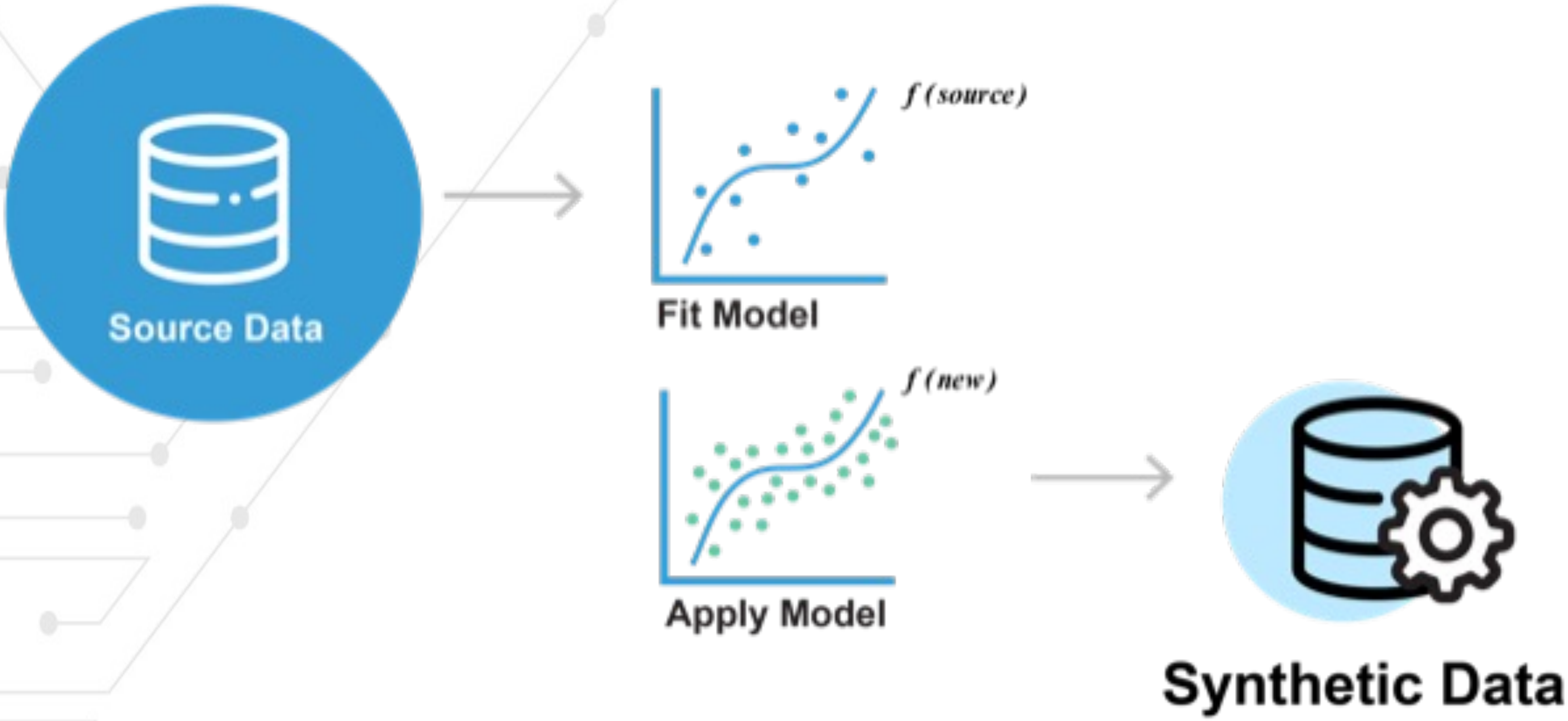
## Real Data

| COU1A | AGECAT | AGELE70 | WHITE | MALE | BMI | N |
|---|---|---|---|---|---|---|
| United States | 3 | 1 | 0 | 1 | 25.44585 | |
| United States | 3 | 1 | 1 | 0 | 24.09375 | |
| United States | 3 | 1 | 1 | 1 | 33.07829 | |
| United States | 2 | 1 | 1 | 0 | 33.64845 | |
| United States | 3 | 1 | 1 | 0 | 25.66958 | |
| United States | 3 | 1 | 1 | 0 | 25.85938 | |
| United States | 2 | 1 | 1 | 0 | 24.7357 | |
| United States | 5 | 0 | 0 | 0 | 27.75276 | |
| United States | 5 | 0 | 1 | 1 | 28.07632 | |

| COU1A | AGECAT | AGELE70 | WHITE | MALE | BMI |
|---|---|---|---|---|---|
| United States | 2 | 1 | 1 | 1 | 33.75155 |
| United States | 2 | 1 | 1 | 0 | 39.24707 |
| United States | 1 | 1 | 1 | 0 | 26.5625 |
| United States | 4 | 1 | 1 | 1 | 40.58273 |
| United States | 5 | 0 | 0 | 1 | 24.42046 |
| United States | 5 | 0 | 1 | 0 | 19.07124 |
| United States | 3 | 1 | 1 | 1 | 26.04938 |
| United States | 4 | 1 | 1 | 1 | 25.46939 |

## Synthetic Data

# The Synthesis Process



# Synthetic Data as a Privacy Enhancing Technology

Synthetic data looks real and has the same relationships and patterns as real datasets.

Since the individuals in the data are not real, the privacy implications are different than with real data, requiring different strategies to assess risk

# Traditional Risk Assessments

In Canadian law, **identity disclosure** is the main risk associated with de-identified data

Reidentification risk is the probability of being able to correctly match a record in a microdata sample to a real person

# Traditional Reidentification Risk

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Quasi-identifiers

Step 1: Identify quasi-identifiers an attacker may use

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Step 2: Compare microdata records to population using quasi-identifiers

Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Step 2: Compare microdata records to population using quasi-identifiers

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Step 2: Compare microdata records to population using quasi-identifiers

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Risk: 1/1

Step 3: Calculate risk for each record in the microdata as 1 divided by the number of records that match in the population

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Risk: 1/1

Step 3: Calculate risk for each record in the microdata as 1 divided by the number of records that match in the population

Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Risk: 1/2

Step 3: Calculate risk for each record in the microdata as 1 divided by the number of records that match in the population

# Traditional Reidentification Risk

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

Average risk:

$1/3 \times (1/1 + 1/1 + 1/2) = 0.83$

Step 4: Average the risk across all records in the microdata




Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# What's Different in Synthetic Data?

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata or Real Training Data

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

## Synthetic Data

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 55566-2110 |
| Male | 1986 | 0023-3670 |
| Female | 1987 | 54868-6348 |

Individuals in the synthetic dataset may or may not be present in the real population

# What's Different in Synthetic Data?

## Population

| Sex | Year of Birth | NDC |
|---|---|---|
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

## Microdata or Real Training Data

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |

## Synthetic Data

| Sex | Year of Birth | NDC |
|---|---|---|
| Female | 1983 | 55566-2110 |
| Male | 1986 | 0023-3670 |
| Female | 1987 | 54868-6348 |

Even if you match a synthetic record with a real person on the quasi-identifiers, the information learned may not be correct

Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# Records in the Synthetic Data

Fall into the following categories:

1) Duplicate real individuals in their entirety due to overfitting in the synthesis model or a simple dataset
2) Correspond with real individuals when considering QIs only
3) Do not correspond with real individual when considering QIs only




Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# Extension to Synthetic Data

**Real Data** ⟷ **Synthetic Data**

Every record is a real individual in the population

A small proportion of records correspond to real individuals in the population

# Extension to Synthetic Data

**Real Data**

**Synthetic Data**

Reidentification Risk

Attribution Disclosure

Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# Attribution disclosure: find a similar record in the synthetic data and learn something new

| Quasi-identifiers | | Sensitive variables |
|---|---|---|
| Sex | Year of Birth | NDC |
| Male | 1985 | 009-0031 |
| Male | 1988 | 0023-3670 |
| Male | 1982 | 0074-5182 |
| Female | 1983 | 0078-0379 |
| Female | 1989 | 65862-403 |
| Male | 1981 | 55714-4446 |
| Male | 1982 | 55714-4402 |
| Female | 1987 | 55566-2110 |
| Male | 1981 | 55289-324 |
| Female | 1986 | 54868-6348 |
| Male | 1980 | 53808-0540 |

# Learning Something New

| | | Similarity in Real Sample | |
|---|---|---|---|
| | | Individual is Similar to Others | Individual is an Outlier |
| Similarity Between Real & Synthetic Samples | Individual's Synthetic Information Similar to Real Information | Low Attribution Risk | High Attribution Risk |
| | Individual's Synthetic Information Different from Real Information | Low Attribution Risk | Low Attribution Risk |
| Note: This table only applies to records that match between the synthetic and real data, and hence have passed the first test for what is defined as meaningful identity disclosure. | | | |

A synthetic record matching a real individual is harmful if and only if it allows an attacker to learn something new about a real individual; that could not be learned through inference on a complete dataset

Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# Attribution Risk Results

Published risk assessment results for synthetic data generated using sequential tree synthesis method:

| | Synthetic Data Risk | |
|---|---|---|
| | Population-to-sample risk | Sample-to-population risk |
| Washington State Inpatient Database | 0.00056 | 0.0197 |
| Canadian COVID-19 cases | 0.0043 | 0.0086 |

El Emam K, Mosquera L, Bass J. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. J Med Internet Res 2020;22(11):e23139, doi: 10.2196/23139.

# Attribution Disclosure

Key traits:

- Conveys average risk within a synthetic dataset
- Converges to reidentification risk for duplicated records
- Accounts for the uncertainty introduced by synthesis

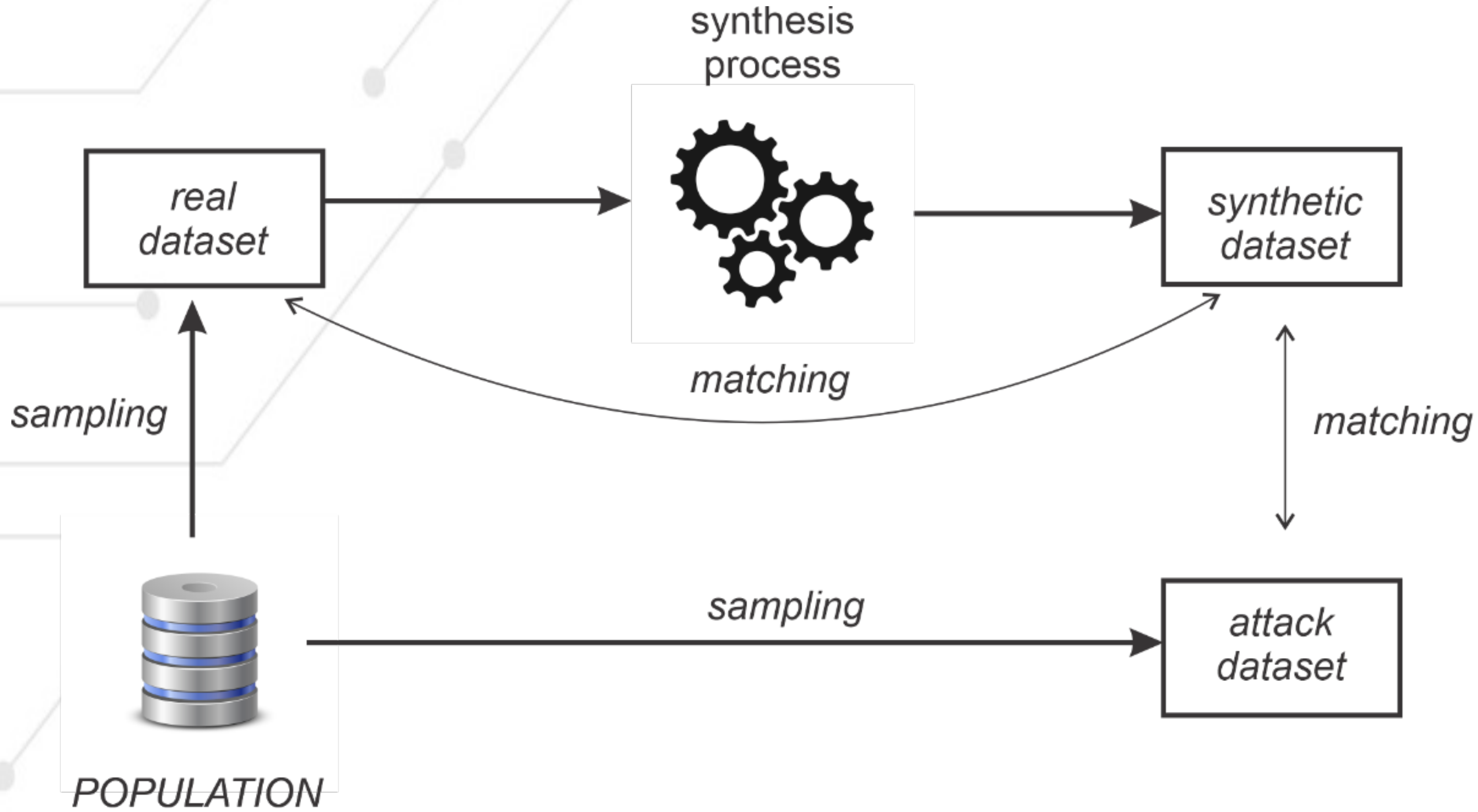
# Membership Disclosure

Consider a situation where a synthetic dataset has been generated for all individuals who received care at your local hospital in 2021 who were diagnosed with COVID.

If you could use the synthetic dataset to find an individual in your community who was in the training dataset, you would learn sensitive information about that individual: even if you learned nothing else about them

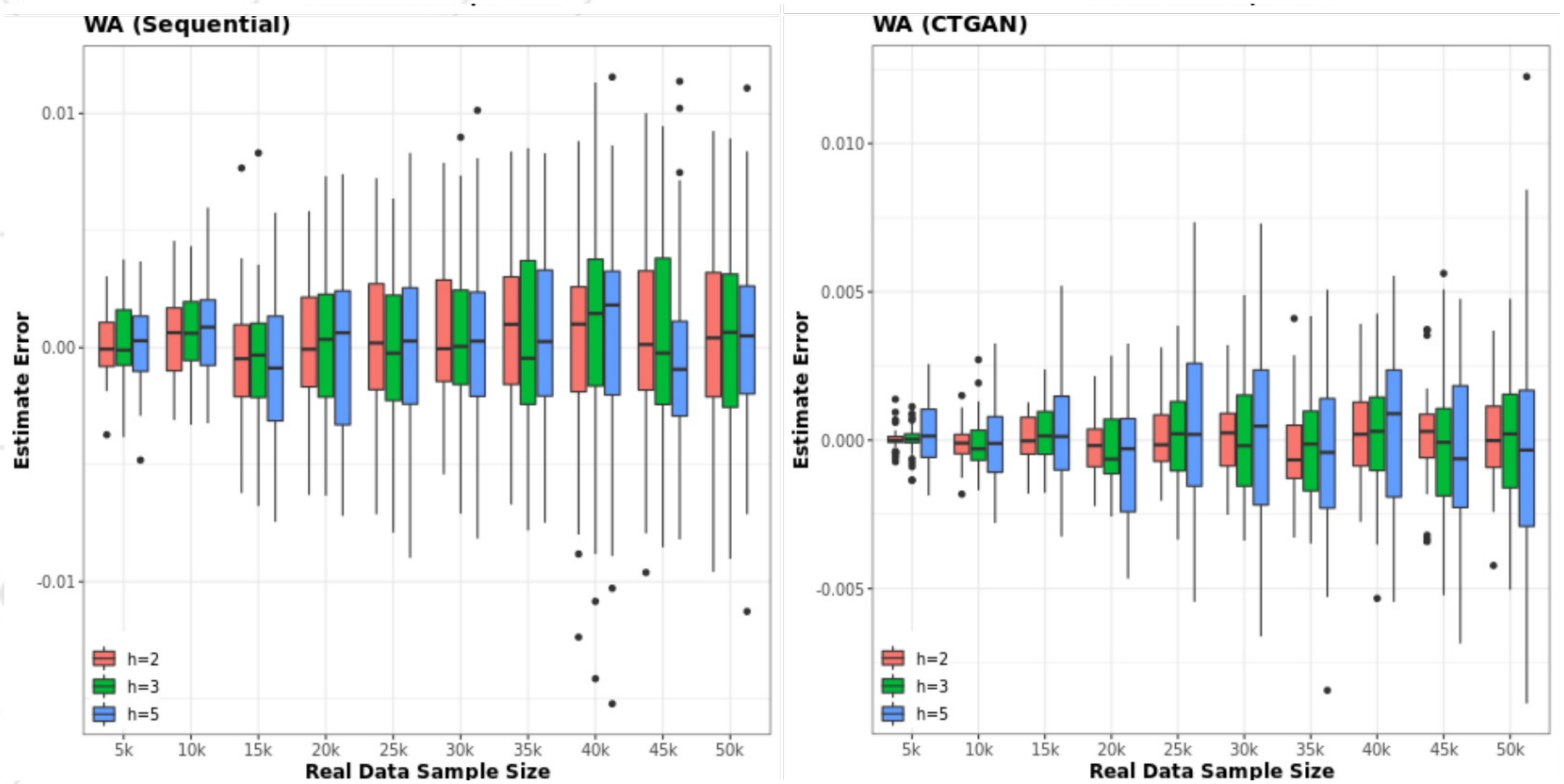
Membership disclosure quantifies this risk

# Membership Disclosure




Considering a target individual, from the population, what is the probability that this individual was in the dataset used to train the generative model, given that the target individual matches a record in the synthetic data on the quasi-identifiers

# Membership Disclosure Results




Simulation study to compare true membership disclosure to estimated within Washington hospital discharge dataset where the population is available.





Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# Unified Risk Assessment

First unified risk assessment for synthetic data:

$$max \begin{Bmatrix} attribution\ disclosure, \\ membership\ disclosure \end{Bmatrix}$$

- Model can be extended to take into account different types of attacks (deliberate, inadvertent, and breach) as well as verification of matches
- Can compare risk values to acceptable risk thresholds




Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# What If the Risk In My Synthetic Data Is High?

Data synthesis can be combined with other PETs and risk mitigation strategies including:

- Pre–synthesis:
  - Training data could be de-identified
- During synthesis:
  - Differential privacy integrated with model training to prevent overfitting
- Post generation:
  - Data transforms and implementing controls regarding data access

Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# Canadian Regulatory Stance on Synthetic Data

- Legal analysis of the treatment of data synthesis under Part I of PIPEDA and Bill C-11 – the Consumer Privacy Protection Act (CPPA)
- Analysis goes through the life cycle of synthetic data and the implications at each stage
- Assumption is that the data has a very small risk of identity disclosure

Note that Bill C-11 died on the order paper when Parliament was prorogued on August 15, 2021, so it informs how regulators aim to amend PIPEDA to promote innovation and protect Canadians' data

# Generation of Synthetic Data

**Is the use of the original (real) dataset to generate and/or evaluate a synthetic data set restricted or regulated?**

- PIPEDA – unclear
- CPPA – addressed but could be improved

# Use & Disclosure of Synthetic Data

**Do the statutes regulate or affect, if at all, the resulting use and disclosure of the synthetic data set? In other words, is synthetic data personal information ?**

- PIPEDA: Addressed, but could be improved to better protect privacy rights
- CPPA: Addressed, but could be improved to better protect privacy rights

# Canadian Regulator Perspectives

Interviewed 13 Canadian regulators on synthetic data with the aim of assessing perspectives in four areas:

- Adoption of synthetic data generation practices
- Consent requirements for synthetic data generation
- Regulation of synthetic data
- Implementation of good synthetic data generation practices



Electronic Health Information Laboratory

Replica Analytics

AN AETION COMPANY

# Adoption of Synthetic Data Generation (SDG) Practices

- Many had no direct experience with synthetic data generation
- Experience with synthetic data was limited to:
  - A complaint launched against Facebook for using synthetic data in software testing
  - An ongoing project in Alberta regarding synthetic data generation that consulted the privacy commissioner

# Consent Requirements for SDG

4 perspectives identified:

1. The legislation makes it clear that the creation of non-identifiable datasets is a permitted use, and therefore no additional consent is required.
2. The use of PETs further protect the rights of the data subjects by generating non-identifiable data, which should be encouraged. If the SDG was appropriately executed, consent is not required.
3. The purposes for which the synthetic data will be used or disclosed are consistent with the initial consent for which the dataset was collected. Consent is not required.
4. Explicit consent is required to create synthetic datasets (or any other form of non-identifiable data).

Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# Regulation of Synthetic Data

Given that the synthetic data generation has been done properly, how should it be regulated?

1. Organizations processing synthetic data should have fewer obligations than organizations processing personally identifiable information. Some obligations on synthetic data would be very difficult to operationalize (e.g., deletion and access as synthetic records cannot be linked to a specific individual).
2. Synthetic data falls outside privacy legislation as it would have a very small identity disclosure risk. No additional obligations or constraints required
3. Additional obligations (e.g., obtaining additional consent) to use synthetic data for secondary purposes would not be required if the secondary purpose was deemed to be a socially beneficial purpose or a legitimate commercial purpose.
4. Given that synthetic data does not have precisely zero identity disclosure risks, regulation of synthetic data would be necessary.

Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY

# Implementation of Good Synthetic Data Generation Practices

If synthetic data is considered “not personal information”, conditional on SDG being implemented properly, then there is a need to proactively define codes of practice for SDG.

Subsequent concerns include:

- Who should approve SDG codes of practice?
- Need to harmonize SDG codes of practice nationwide or adopt international standards

# Conclusions from Regulator Discussions

- Synthetic data adoption is in its early phases
- Perspectives vary among regulators across Canada, but there is consensus on key issues
- There is a need to establish consistent codes of practice nationwide
- Inconsistent interpretations (uncertainty) will lead to inaction and slower adoption of synthetic data

# Key Take-Aways

- Synthetic data, if generated using good practices, could legitimately be characterized as having low identity disclosure risk
- Privacy risk in synthetic data can be comprehensively quantified using our unified privacy metric
- The establishment of codes of practice for synthetic data generation would provide confidence that practices are sound and could help reduce regulatory obligations
- There is some consensus among privacy regulators across Canada on how synthetic data should be regulated, but there is also some divergence




Electronic Health Information Laboratory

Replica Analytics
AN AETION COMPANY


QUESTIONS
cc: an untrained eye - https://www.flickr.com/photos/