HLG-MOS Synthetic Data Project

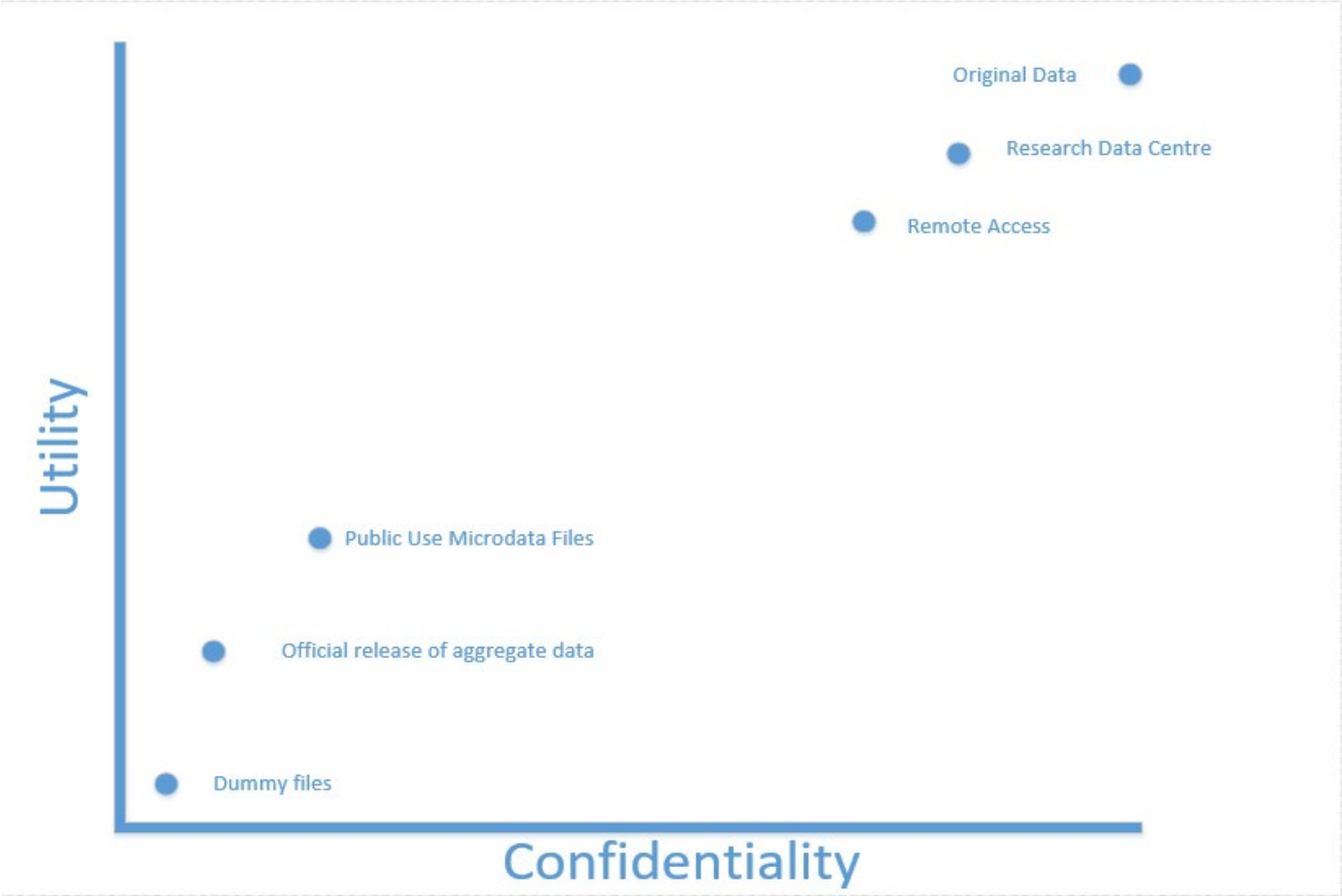
Kate Burnett-Isaacs, Statistics Canada

April 2021

Delivering insight through data for a better Canada

What Problem Would **Synthetic Data** Solve?

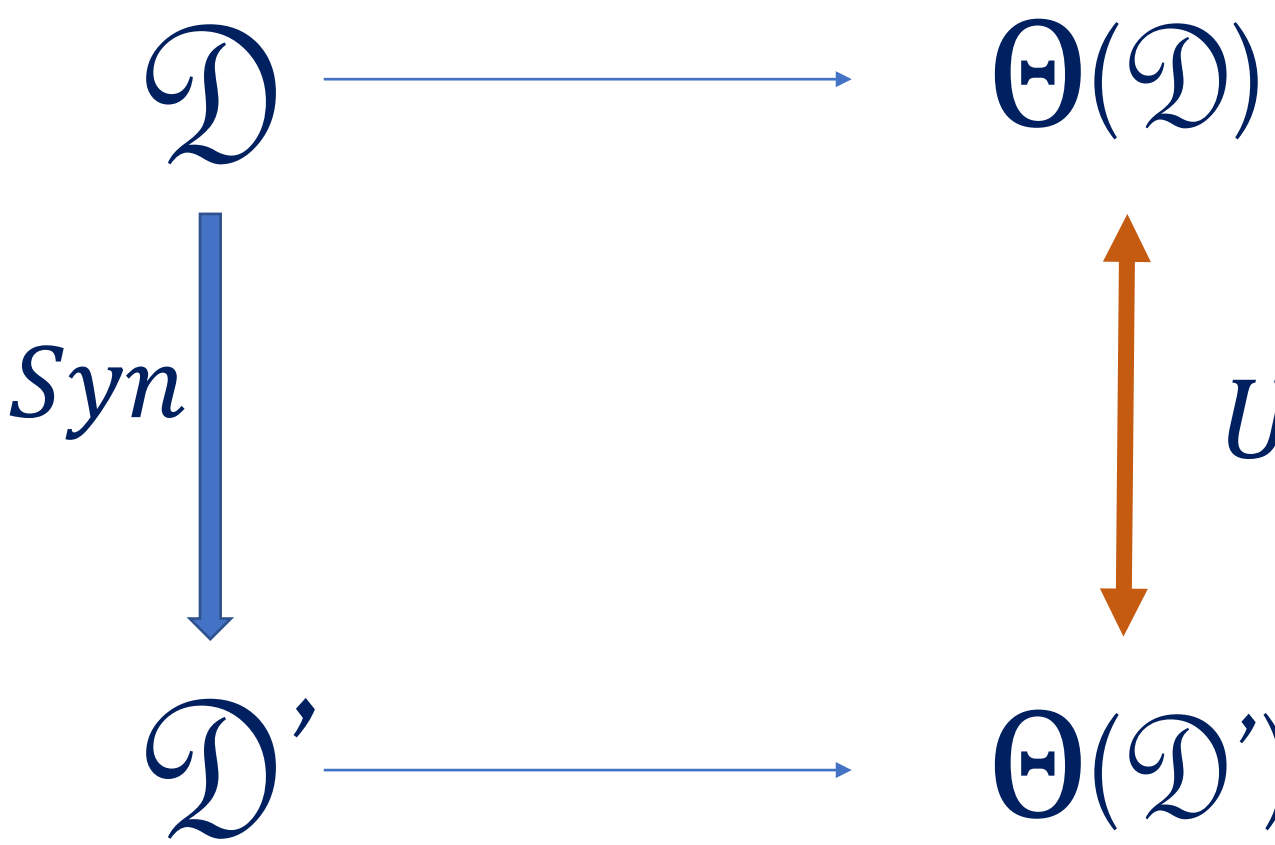
- National statistical offices (NSOs) are striving to provide greater transparency and openness
- Need to disseminate quality data sets to support testing, evaluation, education and development purposes
- **Output Privacy Method:** Confidentiality remains a top priority
- Synthetic data can be a solution to providing rich data while respecting integrity and confidentiality imperatives.



What is Synthetic Data?

- Modelling or generation process in order to target both the preservation of the analytical value and confidentiality protection.
- The advantage is that synthetic data seeks to preserve the analytical value (or utility) of the original file while minimizing disclosure risk.
- Fully and partial synthetic data files

Data Synthesis: the Concept



$\mathcal{D}$ the original dataset

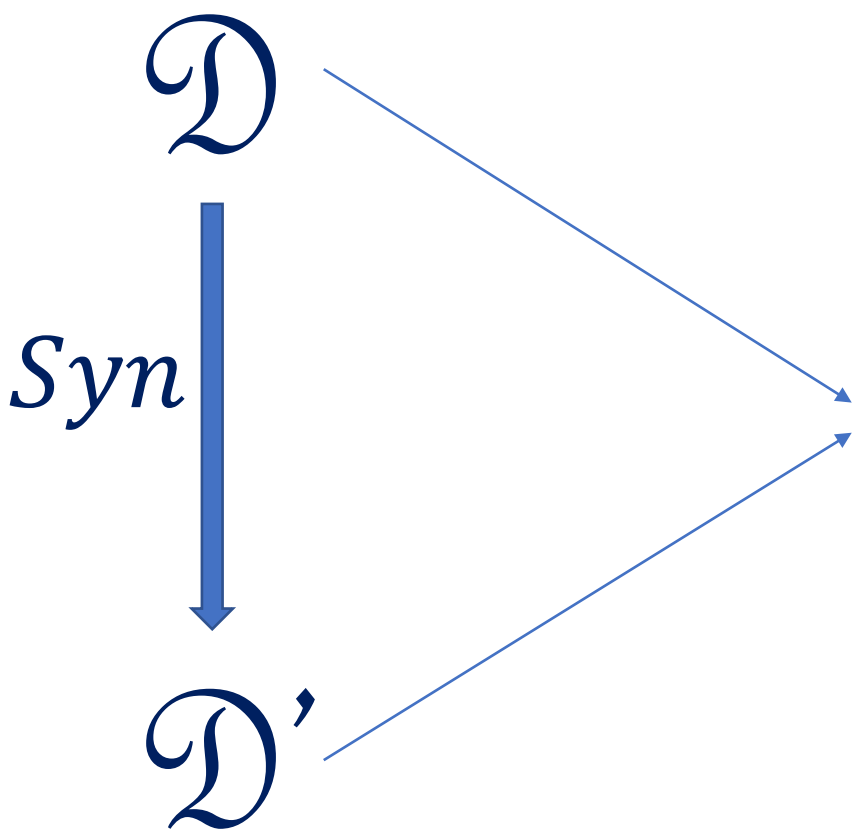
$\mathcal{D}'$ the synthetic dataset

Syn Process creating synthetic data

Θ Results of analyses

U the utility

Data Synthesis: the Concept



$\mathcal{D}$ the original dataset

$\mathcal{D}'$ the synthetic dataset

Syn Process creating synthetic data

Θ Results of analyses

U the utility

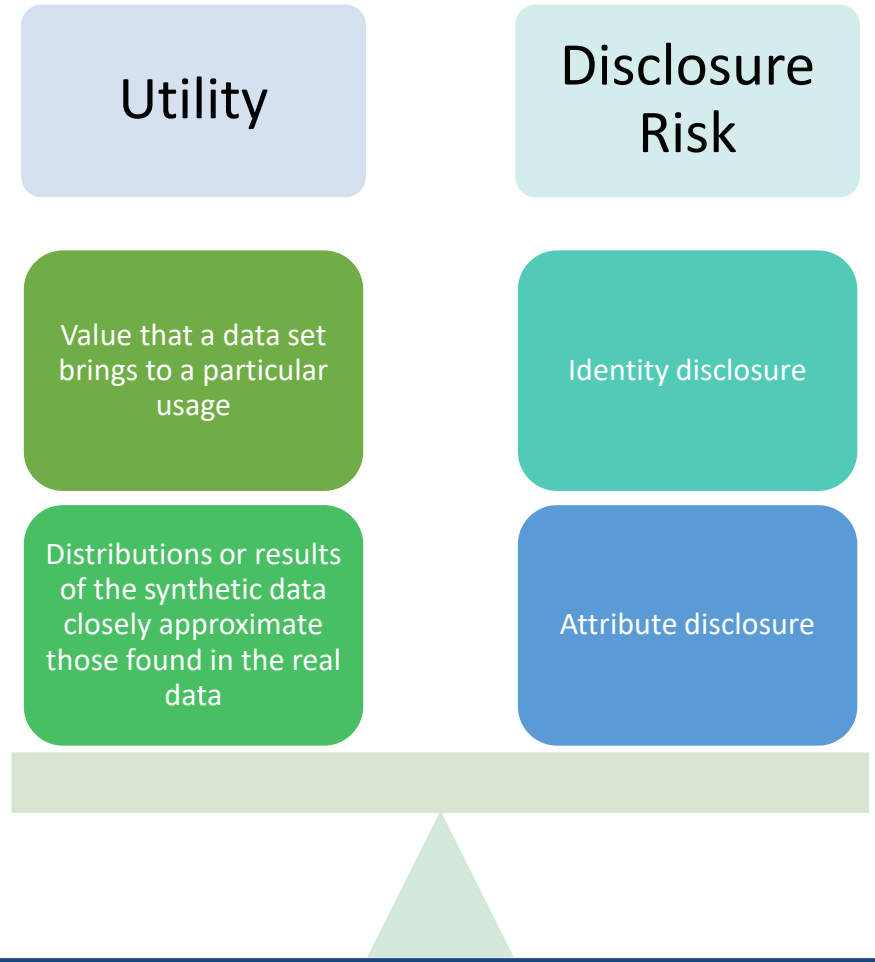
Minimal distance, maximum utility

Θ should not be known in advance

Why is a **Practical Guide to Synthetic Data** Needed?

New methods are emerging for generating and evaluating confidentiality of synthetic data, and **more guidance is needed to maximize utility while ensuring confidentiality**.

The utility and level of risk accepted is entirely dependent on the purpose for the synthetic data



Description of the **Practical Guide to Synthetic Data** Project

Develop a hands-on guide for creating and using synthetic data for data protection and disclosure control geared towards NSOs and their data users.

Work Package 1: Use cases for synthetic data

Work Package 2: Recommended methods for creating synthetic data

Work Package 3: Measuring the analytical value and/or disclosure risk of synthetic data sets

Work Package 4: Experimenting with the recommendations

What have we done so far?

Use Cases

The analytical value or utility and the measure of quality are highly dependent on what you are using the synthetic data for.

- Dissemination to the public
- Testing analysis
- Testing ML algorithms
- Education
- Testing systems

Disseminating to the Public

- Want to provide microdata with high analytical value to all users
- Challenge:
  - No knowledge of the type of analysis being conducted
  - High need for confidentiality

High Utility and High Confidentiality

US Census Bureau

- The release of the 2020 US Census uses mostly differential privacy methods, these methods are not suitable for the Island Area Census.
- The Island Area Census contains more demographic information
- These data will be released to the public with a combination of swapping and synthetic data

Testing Analysis

- Provide synthetic data to researchers or other users while they wait to get access to real data
- Considerations:
  - Prior knowledge of analysis being conducted and variables of interest
  - Researchers may already of some level of security clearance
- New synthetic data file for every researcher?

High utility and moderate confidentiality

Statistics Canada

- Statistics Canada is creating a synthetic version of a census-modified database in order to make the data accessible to a broader audience outside of the traditional Research Data Centers.
- The target of the synthetic dataset is to test and run the New Dynamic Microsimulation Model of Retirement Income to provide preliminary results

Testing ML Algorithms

- Similar use case to testing analysis
- Considerations
  - Must have prior knowledge of relationships that need to be preserved
  - How real does the data need to be?

Medium utility and medium confidentiality

Australian Bureau of Statistics

- Needed synthetic data to test ML methods and give demos
- Needed data that included entity-entity relationships in a realistic way that allows for testing methods dependent on these relationships.
- Must be "fully synthetic": no dependence on non-public data.
- Must be scaleable

Education

- High quality data is needed in order for students, academic and users in general to learn new concepts and methods.
- The more complex the methods, the more important it is that the data used in this training can provide realistic results and emulate what students will be facing in the real world.

Low utility and high confidentiality

Scottish Centre for Administrative Research

- Synthetic data provided for a course on the use of administrative data for social and health research
- Original data from the linked Census and administrative records on youth employment and school attendance
- This allowed students on course to get exposure to real data and their problems.

Testing Systems

- Traditionally use dummy files
- However, more and more systems will need to be tested where the outputs and analysis of those outputs need to mirror real life.

Low utility and medium confidentiality

Office of National Statistics

- The ONS Census team was developing the processing platform for the 2021 UK Census
- Data Science Campus made a synthetic version of the previous Census to test the 2021 platform
- The synthetic data were initially generated within a secure environment for use within the organisation but is being expanded with the inclusion of privacy preserving guarantees.

Methods Gathered So Far

- Sequential multimodel multivariate model method
- Fully conditional specification
- Special cases of a fully conditional specification
- Microsimulation for generating synthetic data
- Information preserving statistical obfuscation
- Pseudo likelihood method
- Method by Fleishman and Vale & Maurelli to simulate multivariate non-normal random numbers from the original data
- GANS

Upcoming

- Decision Trees
- Differentially private synthetic data
- Special cases of fully conditional specification
- Tuning/considerations

Utility and Disclosure Risk Measure

Utility Measures

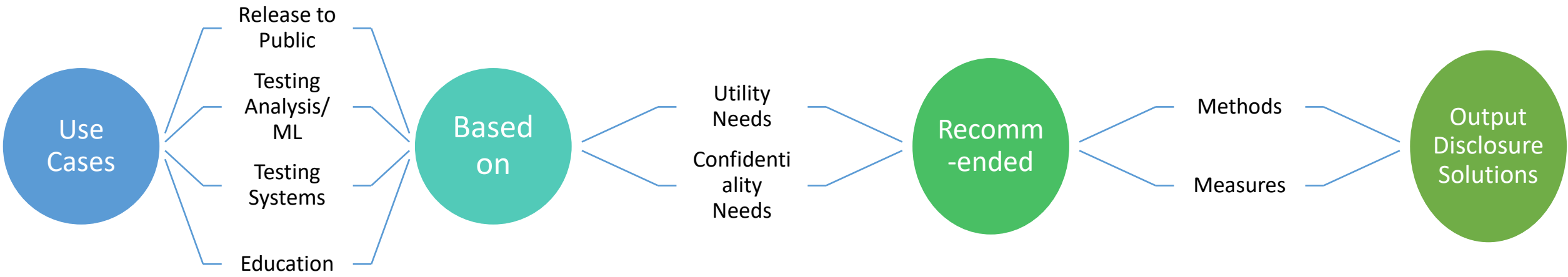
- Feature Mean Scaled Variance (FMSV)-based actual-to-synthetic unit record similarity analysis
- Confidence interval overlap
- Mahalanobis distance ratio
- Voas-Williamson and Freeman-Tukey measures for tables
- Propensity score mean-square error
- Histogram comparison
- Compare Pearson correlation heatmaps for real and synthetic data
- Identify a task relevant to the dataset (e.g. classification) and compare task accuracy for real and synthetic data.
- Marginal distribution metrics

Disclosure Risk Measure

- Peer review audit of data linking, sampling, and modeling methodologies
- Feature Mean Scaled Variance (FMSV)-based actual-to-synthetic unit record similarity analysis
- Rates related to database reconstruction (reconstruction, putative re-identification, confirmation)
- Bayesian estimation using different priors representing adversary knowledge
- Bounding of the ratio of probabilities between output synthetic data sets that differ in one record, called epsilon-synthetic privacy
- Correct attribute probability in which the adversary searches all records in the synthetic dataset that match the key values known by them, and calculate the distribution of the occurring values of the target attribute

What's Next?

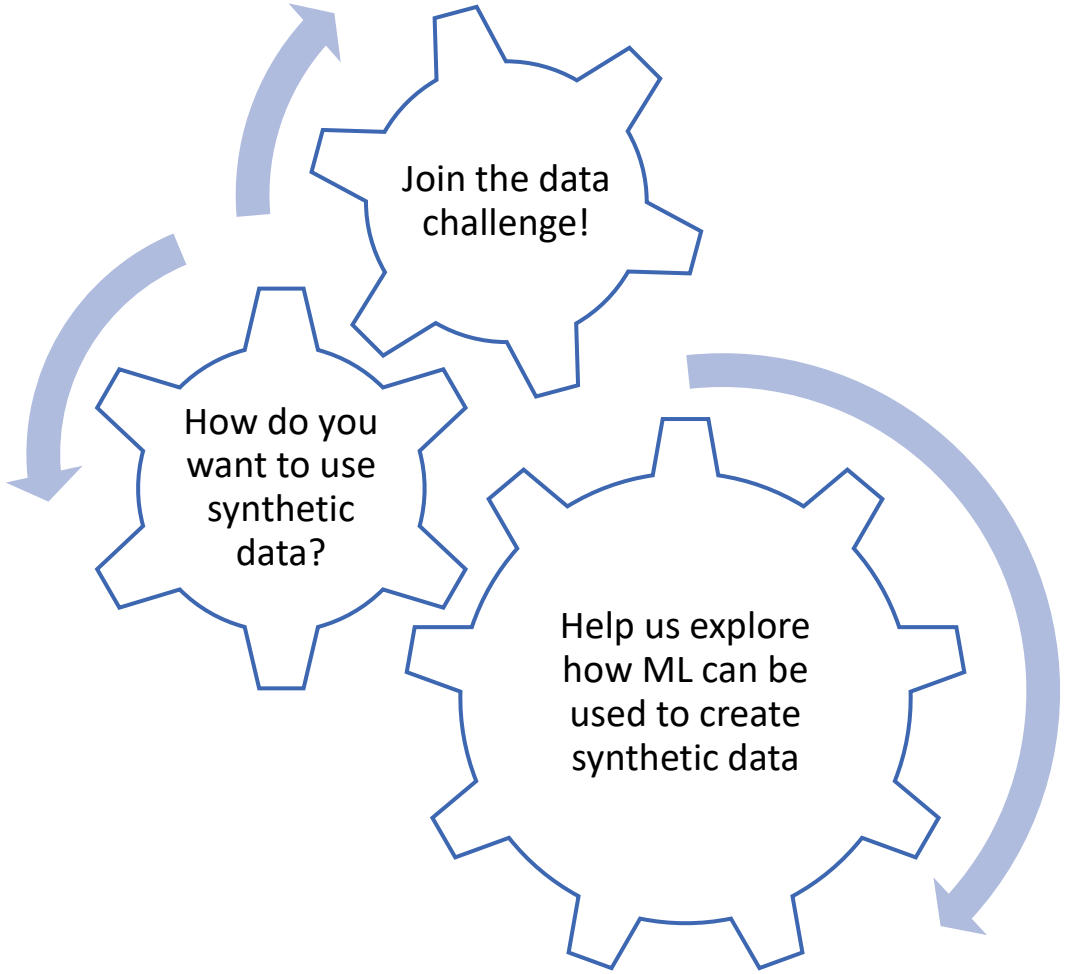
Recommendations



Synthetic Data Challenge: September 2021

- Do our recommendations hold up in real life scenarios?
- Virtual
- Multi-national teams
- You have to find the best solution given a chosen use case
- Results of the data challenge could impact the content of the guide

How can the HLG-MOS Synthetic Data Project and ML Group collaborate?



Contact: kate.burnett-isaacs@Canada.ca

Thank you