# Practical Guide to Synthetic Data

HGL-MOS project proposal

Kate Burnett-Isaacs, Statistics Canada

November 18, 2020

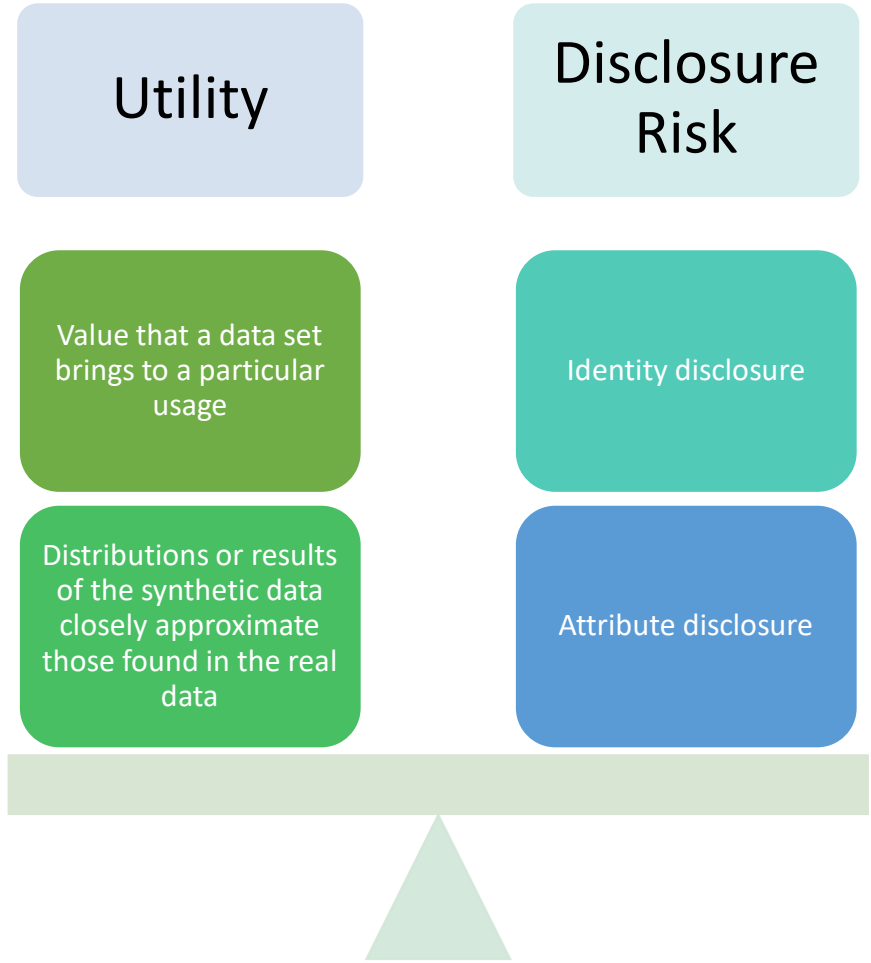
Delivering insight through data for a better Canada

# What Problem Would a **Practical Guide to 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
- Synthetic data can be a solution to providing rich data while respecting integrity and confidentiality imperatives.

## Why is a Guide 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
- Once properly explored and understood, synthetic data can play an important role in the way that NSOs share data while maintaining public trust.



# 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.

**WP1: Use cases for synthetic data**

**WP2: Recommended methods for creating synthetic data**

**WP3: Measuring the analytical value and/or disclosure risk of synthetic data sets**

**WP4: Experimenting with the recommendations**

## Description of the Project (continued)

- Building on the success of the BSTN Working Group on Synthetic Data
- Serve as the foundation for future standards as synthetic data is more broadly adopted within NSOs and by their users.
- Resources towards this project would involve NSOs to contribute in kind
- The project is targeted for one year with the potential of earlier results with sufficient participation

# Why the **Practical Guide** needs to be a project?

- The BSTN working group on synthetic data is at 30 members
- This initiative has outgrown the scope of the BSTN.
- In addition, the target audience for synthetic data and the planned products are beyond the current methods of BSTN communication.
- A formal project would provide the proper scope, oversight and communications for the intended deliverables.

# Importance of Preserving Privacy – Examples from member NSOs

For public release

## 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

## 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

For testing analysis

For training

## 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.

## 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.

For training

# Thank you