# Synthetic Data For National Statistical Organizations: A Starter Guide


HLG-MOS Project 2021

Kate Burnett-Isaacs & Kenza Sallier, Statistics Canada

December 2nd, 2021

# OUTLINE

- Context: The HLG-MOS Synthetic Data Project
- Data Synthesis: Concepts
- Purpose of the Guide and How to Use It
- Next Steps

# Context

- **HLG-MOS: High-Level Group for the Modernization of Official Statistics within the UNECE**
  - Work collaboratively to identify opportunities in modernising statistical organisations
  - Adopt a service oriented approach
  - Ensuring that priorities are community driven

Source: https://statswiki.unece.org/pages/viewpage.action?pageId=187891840

- **The Synthetic Data Project:**
  - Collaborative work since January 2020
  - 50 participants from 15 NSOs, one academia institute and 3 private sector participants
  - Focus on modernizing data access solutions via the use of synthetic datasets

# What Problem Would Synthetic Data Solve?

- National statistical offices (NSOs) are striving to provide greater transparency and openness:
  - Open by default

- Be more user-centric and facilitate access to relevant data to external users
  - Need to disseminate quality data sets to support analyses, training, development purposes and testing
  - Find ways to disseminate more disaggregated data

- Data revolution: Advancement in technology and computer capacity made it possible to implement innovative solutions

- **Confidentiality remains a top priority**

# What Problem Would Synthetic Data Solve?

- National statistical offices (NSOs) are striving to provide greater transparency and openness:
  - Open by default

- Be more user-centric and facilitate access to relevant data to external users
  - Need to disseminate quality data sets to support analyses, training, development purposes and testing
  - Find ways to disseminate more disaggregated data

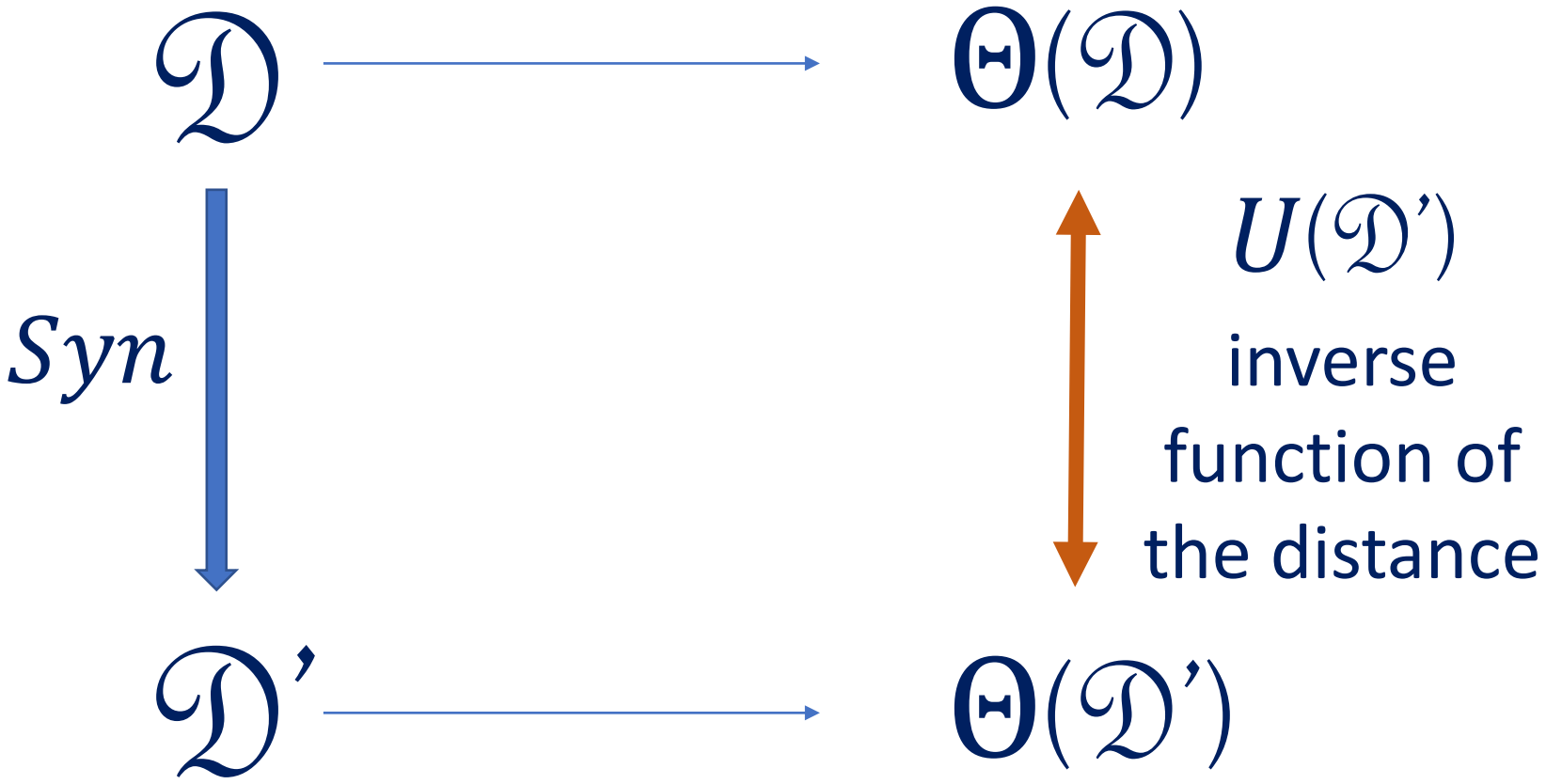
- Data revolution: Advancement in technology and computer capacity made it possible to implement innovative solutions

- **Confidentiality remains a top priority**

Synthetic data can be a solution to providing **analytically rich microdata** while respecting integrity and **confidentiality imperatives**

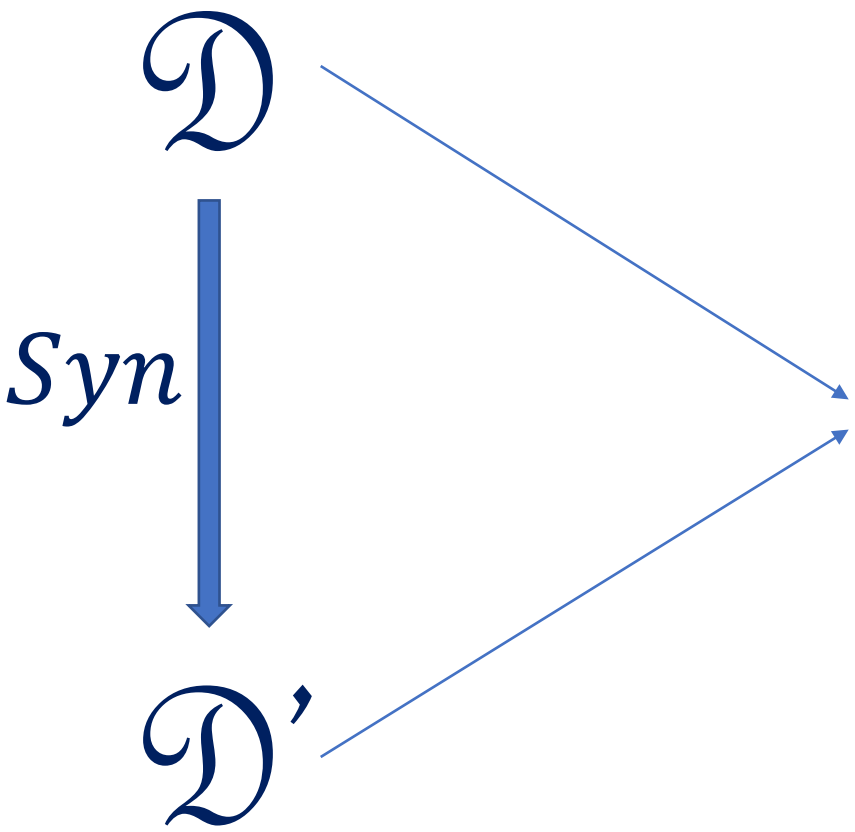
# Concepts Behind Data Synthesis

- $\mathcal{D}$ the original dataset
- $\mathcal{D}'$ the synthetic dataset
- $Syn$ Process creating synthetic data
- $\Theta$ Results of analyses
- $U$ Utility



# Concepts Behind Data Synthesis

- $\mathcal{D}$ the original dataset
- $\mathcal{D}'$ the synthetic dataset
- $Syn$ Process creating synthetic data
- $\Theta$ Results of analyses
- $U$ Utility




Minimal distance = maximum analytical value

# What Problem Would Synthetic Data Solve?

- Many methods exist: which one to use?
  - ➢Various types of synthetic datasets exist with more or less analytical value/disclosure risks
  - ➢What does it take to implement the methods?
- How the evaluate the analytical value? Disclosure risks?

**Need to centralize the information, share NSOs past experiences, best practices, methods, tools, evaluation metrics, references**

# Purpose of the Guide

PRESENT **THEORETICAL METHODS** TO CREATE SYNTHETIC DATA AND PROVIDE AN **INTERNATIONAL CONSENSUS** ON PRACTICAL APPLICATIONS AND **BEST PRACTICES** TO PROMOTE **CONSISTENCY**, **TRANSPARENCY** AND **COMPARABILITY** WITHIN AND ACROSS STATISTICAL AGENCIES, AS WELL AS AMONG USERS IN ACADEMIA AND THE PRIVATE SECTOR.

PROVIDE **COHERENT** GUIDANCE TO DECISION MAKERS WORKING AT ANY LEVEL IN NSOS SO THAT THEY CAN DETERMINE IF SYNTHETIC DATA FIT THEIR DATA ACCESS NEEDS

SCOPE: THE GUIDE IS **INTENTIONALLY DESIGNED FOR PRACTICAL APPLICATION:** IT IS NOT AN EXHAUSTIVE TEXTBOOK. **RESOURCES TO SUPPORT** FURTHER EXPLORATION OF TECHNICAL CONCEPTS ARE HIGHLIGHTED IN THE GUIDE.

# Overview of the Guide

## Chapter 01
Introduction

## Chapter 02
Uses of synthetic data: What data access problem are you facing?

## Chapter 03
Choose the method and tool to produce your synthetic data

## Chapter 04
Assess the quality of your synthetic data: disclosure considerations

## Chapter 05
Assess the quality of your synthetic data: analytical value

# Chapter 2:
## What data access challenge are you facing?

DISSEMINATING TO THE PUBLIC

TESTING ANALYSIS

EDUCATION

TESTING SYSTEMS

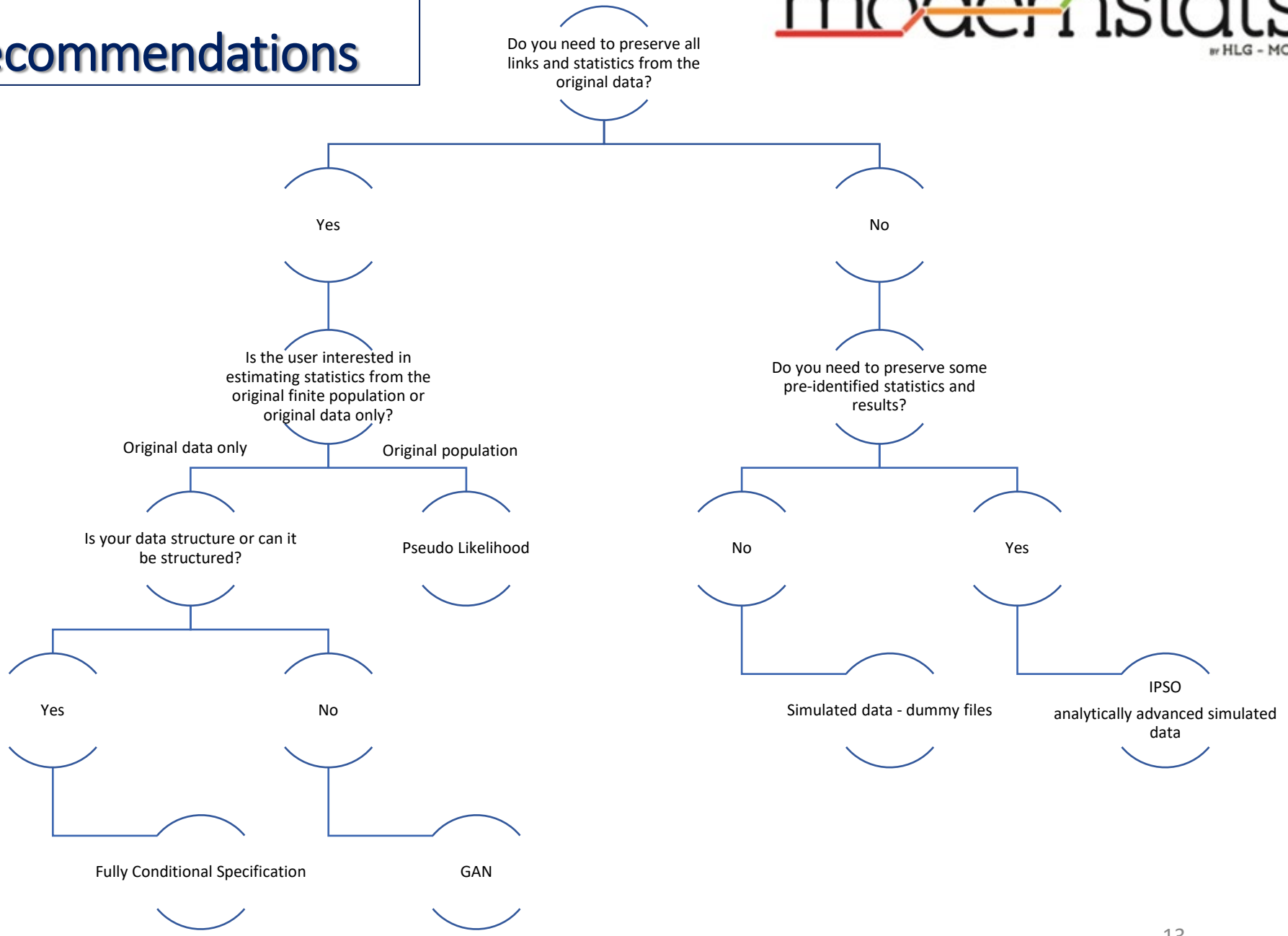
# Chapter 3:
# Methods, tools and recommendations

- 5 main methods: Fully Conditional Specification, IPSO, Simulated Data, Pseudo Likelihood, GANs

- Goal: help choosing your method based on the properties of your original data and the desired properties of your synthetic data

- Structure:
  1. Overview of the method
  2. Tools to implement the method
  3. Pros and cons
  4. Do we recommend this method for the use-cases from chapter 2
  5. References

# Chapter 3:
# Methods, tools and recommendations



# Chapter 4:
# Disclosure considerations for synthetic data

- The purpose of this chapter is to present disclosure considerations for synthetic data and some disclosure risk measures
- Although a synthetic record is generated and do not correspond to a real person or household, **there is concern that attribute and identification disclosure risk could still be present**
- 3 main disclosure risk measures
- Discussion on differential privacy and differentially private data synthesis
- **Recommendation:** NSOs should elaborate their release strategy on their own legislative and operational frameworks and not only rely on these metrics

# Chapter 5:
## Utility measures for synthetic data

- 2 main uses:
  - Having a metric that assess the analytical value for quality evaluation purposes (when it comes time to release the data)
  - Have a score to improve your synthesis process (like a tuning process)
- 17 methods more or less complex in terms of implementation:
  - Evaluate simple results: marginal distributions
  - Evaluate overall multivariate distributions

# Next steps: Gather Feedback

- Link to the guide: https://statswiki.unece.org/x/UQTUE
- Permanent slido: https://app.sli.do/event/rtrdwu72/embed/polls/43080f18-e9c7-4826-880c-dfc66238acde
- Data Challenge – test drive the guide!
  - Dates: January 24 to January 24 to January 28, 2022
  - Problem: you are a NSO that is facing one of 4 use cases (from chapter 2). You must generate synthetic data and assess if it meets the disclosure and utility standards to release it.
  - You will be provided with an 'original' data file
  - Experts will be on hand to help.
  - Registration now open: https://indico.un.org/event/1000359/
- Feedback will included into the final publication that is targeted for a formal printed UN publication in 2022/2023.

# Thank you!
# Questions?

kate.burnett-isaacs@statcan.gc.ca
kenza.sallier@statcan.gc.ca