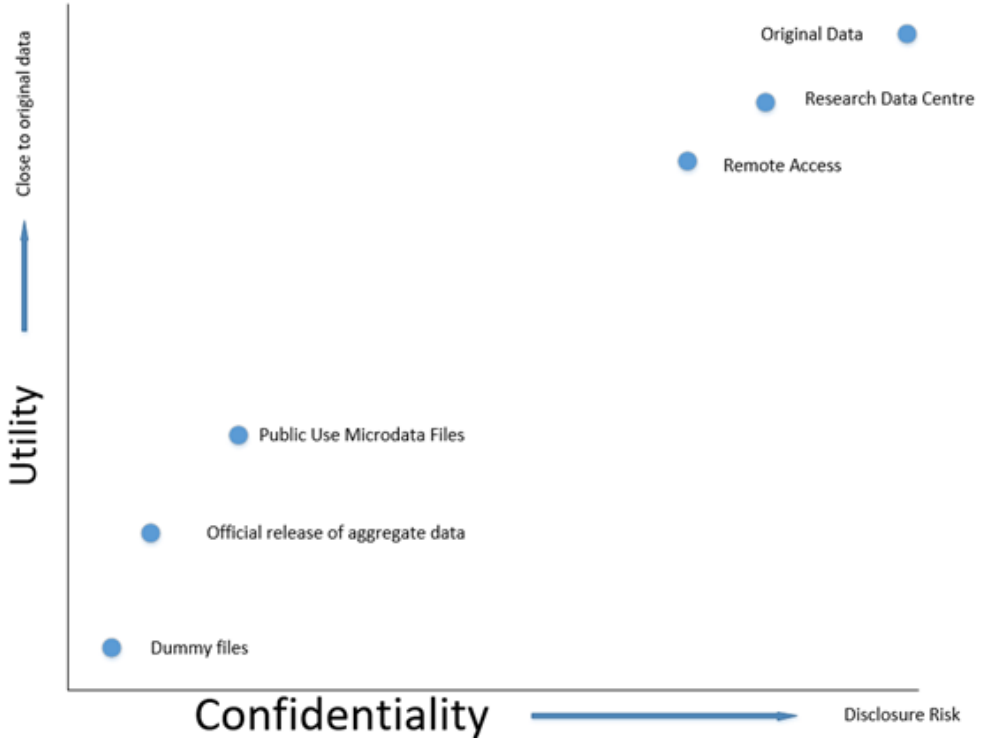
Synthetic Data For National Statistical Organizations: A Starter Guide

HLG-MOS Project 2021

Kate Burnett-Isaacs

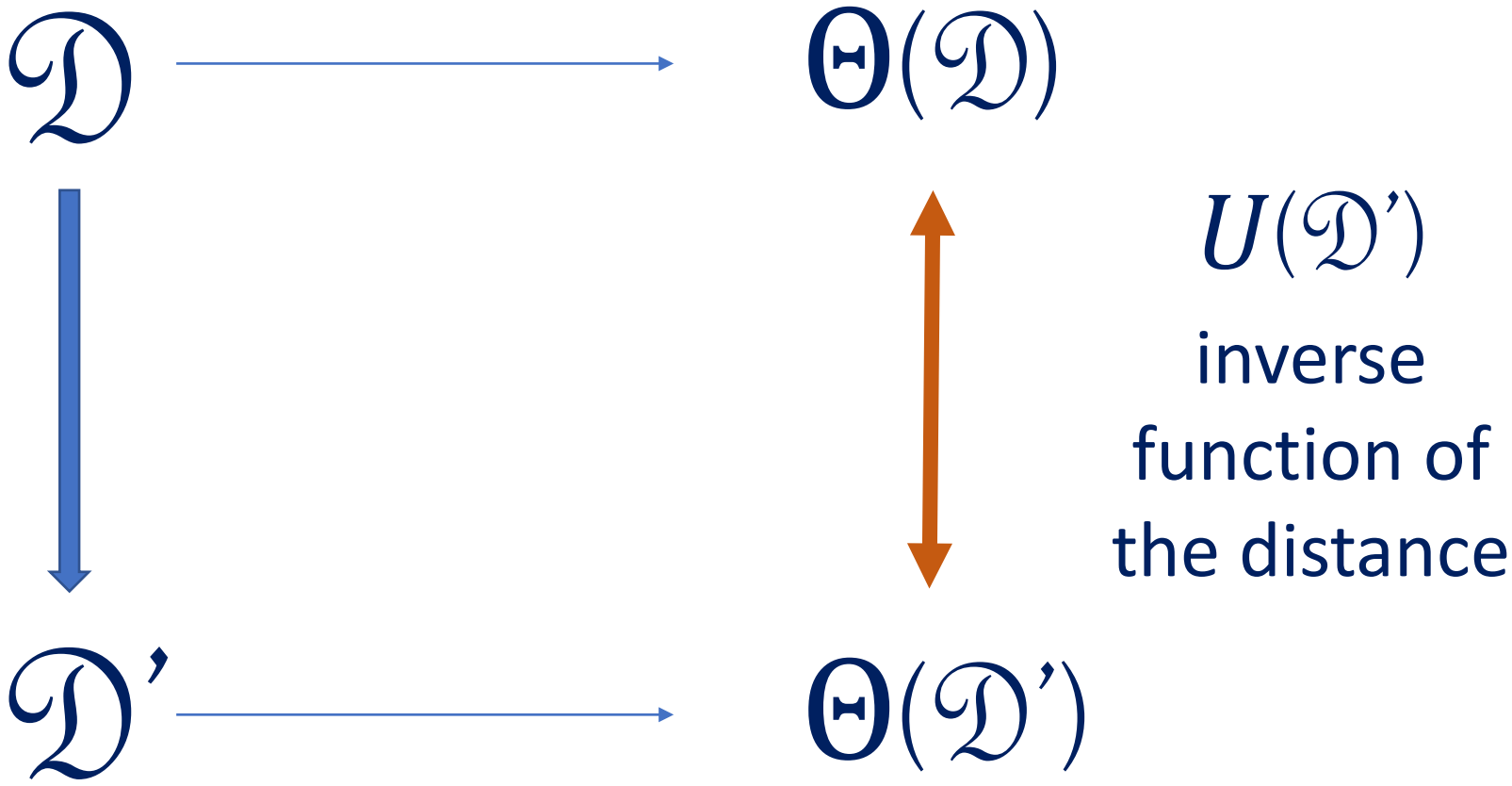
What Problem Would **Synthetic Data** Solve?

- 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.



The concept of data synthesis

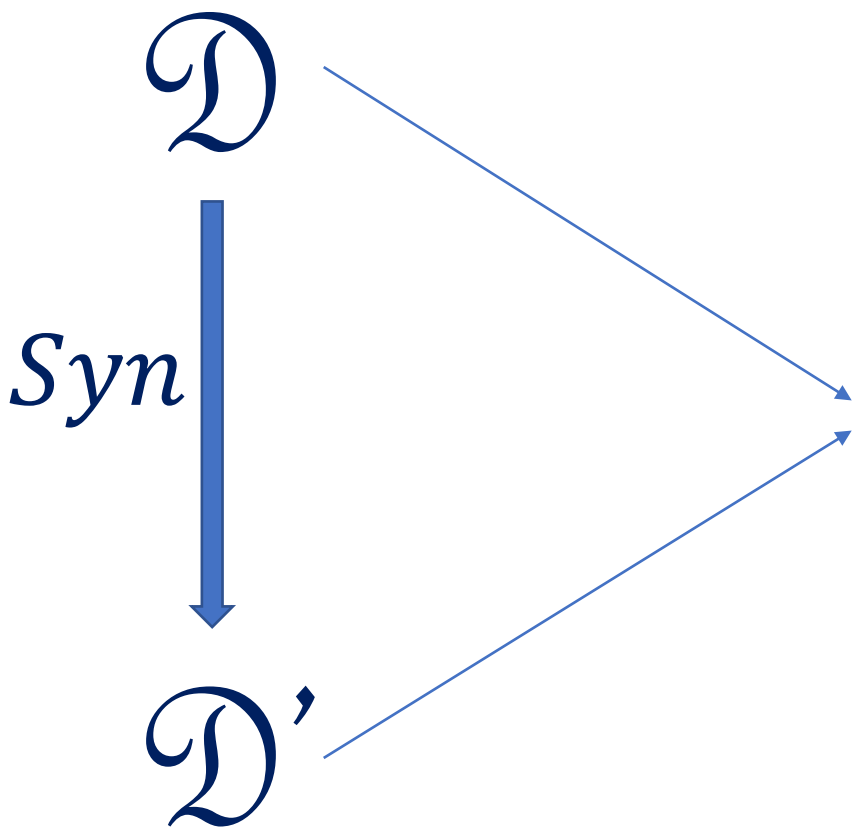
$\mathcal{D}$ the original dataset
$\mathcal{D}'$ the synthetic dataset
Syn Process creating synthetic data
Θ Results of analyses
U the utility



The concept of data synthesis

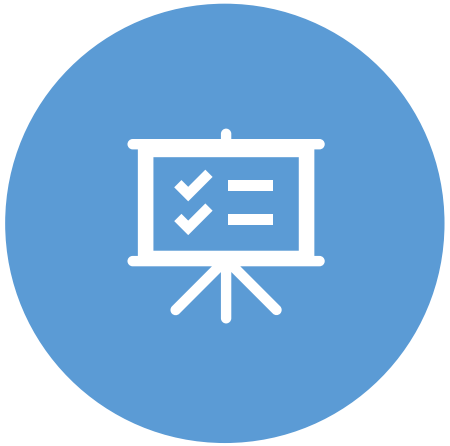
$\mathcal{D}$ the original dataset
$\mathcal{D}'$ the synthetic dataset
Syn Process creating synthetic data
Θ Results of analyses
U the utility

Θ should not be known in advance



Minimal distance, maximum utility

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 IS THE RIGHT SOLUTION TO THEIR DATA DISCLOSURE PROBLEM.**

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.

How to Use the Guide

01

What data access problem are you facing?

02

Mobilize **synthetic data** appropriately in order to **solve your problem**

03

Assess the quality of your synthetic data: Disclosure risk and Utility

Chapter 2: What data access problem are you facing?

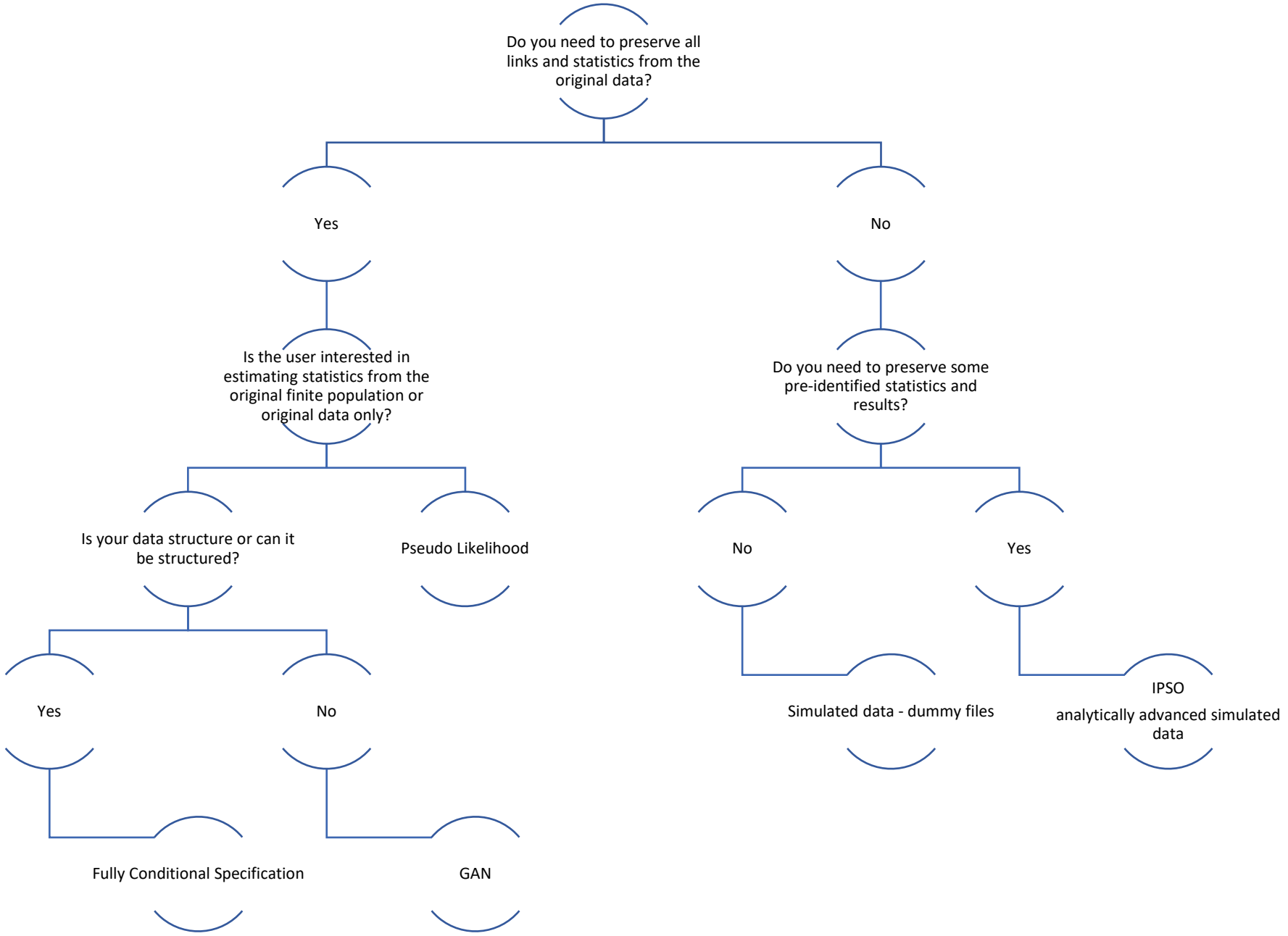
DISSEMINATING TO THE PUBLIC

TESTING ANALYSIS

EDUCATION

TESTING SYSTEMS

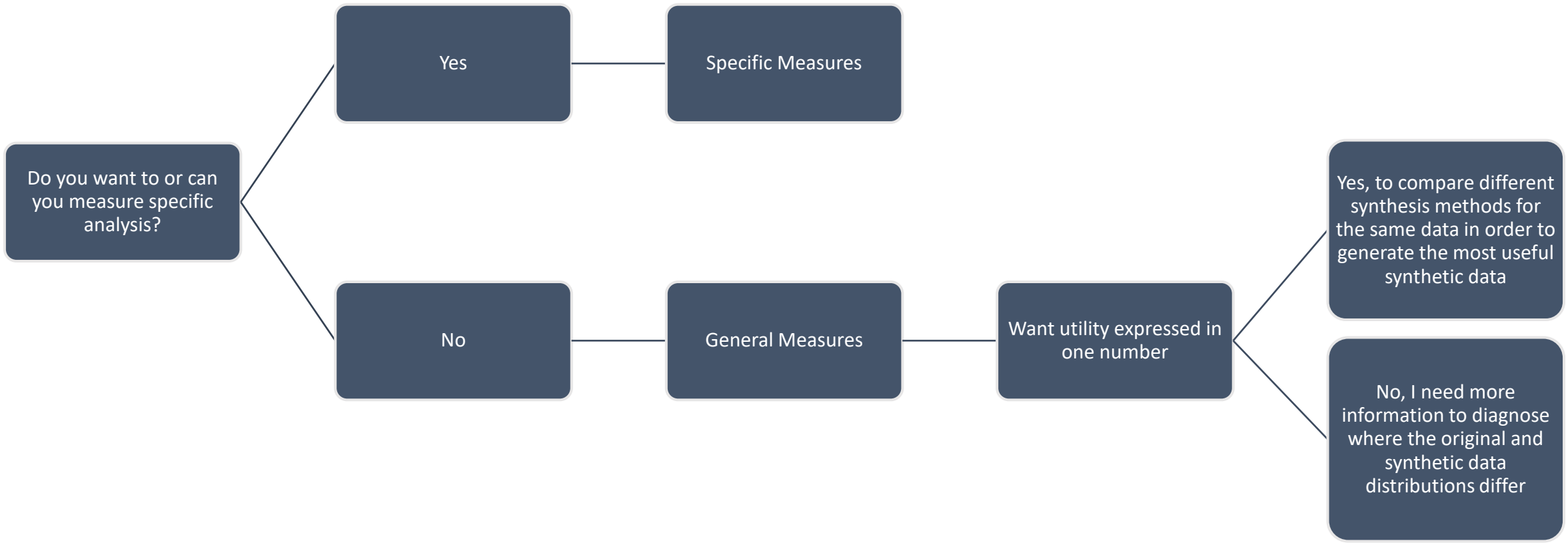
Chapter 3: Choose your Method based on the properties of your original data and the desired properties of your synthetic data



Chapter 4: ensure your synthetic data does not present any disclosure risk

- Disclosure risk is the risk of inappropriate release of data or attribute information of a record (often individual).
- Although no record in a (fully) synthetic data file corresponds to a real person or household, there is concern that attribute and identification disclosure risk could still be present.
- **Recommendation:** NSOs should choose additional disclosure controls based on their own legislative and operational frameworks.
- The purpose of this chapter is to present disclosure control options available to NSOs and their synthesizers

Chapter 5: Choose utility measures to ensure you produce the most useful synthetic data possible



Next Steps

Plans for 2022 and beyond

Gather Feedback

- Workshop on Synthetic Data, November 17, 2021
- Meeting on Statistical Data Confidentiality, December 2, 2021
- Data Challenge – test drive the guide!
  - Dates: January 24 to January 28, 2022
  - Problem: you are a NSO that is facing one of 4 disclosure problems. 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.
- Feedback from these events will be integrated into the final publication will be finalized and is targeted for a formal printed UN publication in 2022/2023.

Early impact of this project

This project and the collaboration that it entailed has already impacted the implementation of synthetic data in the industry.

New information and research that has occurred in this forum is making it's way into the popular open source package *synthop* that generates and evaluates synthetic data

Thank you!

Project had 50 participants from 15 NSOs, one academia institute and 3 private sector participants.

A great big thank you to the technical committee: Gillian Raab, Kenza Sallier, Christine Task, Jia Xin Chang,

Key guide contributors: Héloïse Gauvin, Claude Girard, Ioannis Kaloskampis, Allistair Ramsden, Rolando Rodriguez, Manel Slokom and Steven Thomas

Presenters over the course of last 2 years - their content has been included in this guide: Ovyind Langsrud, Gillian Raab, Kenza Sallier, Christine Task, Joerg Drechsler, Geoffrey Brent, Nicolas Grislain, Ioannis Kaloskampis, Rolando Rodriguez and Joseph Chien.