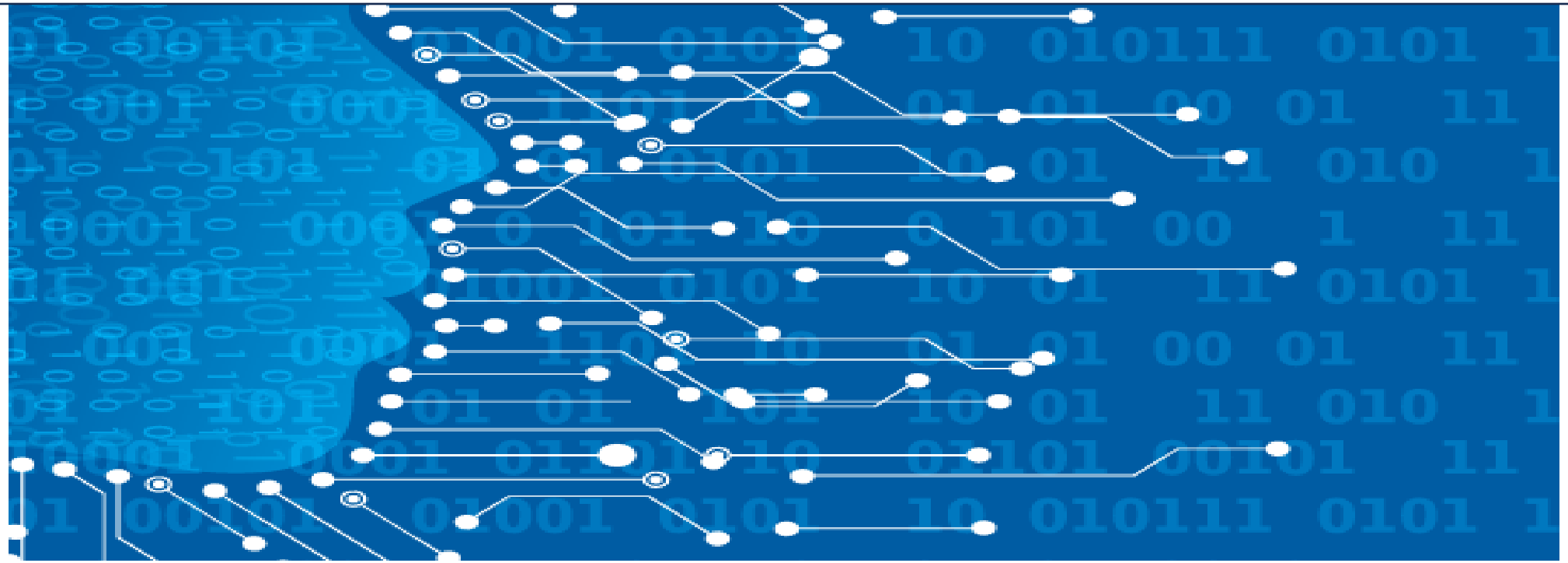
Synthetic Data for Official Statistics

& the UNECE Starter Guide



ESCAP Stats Café webinar

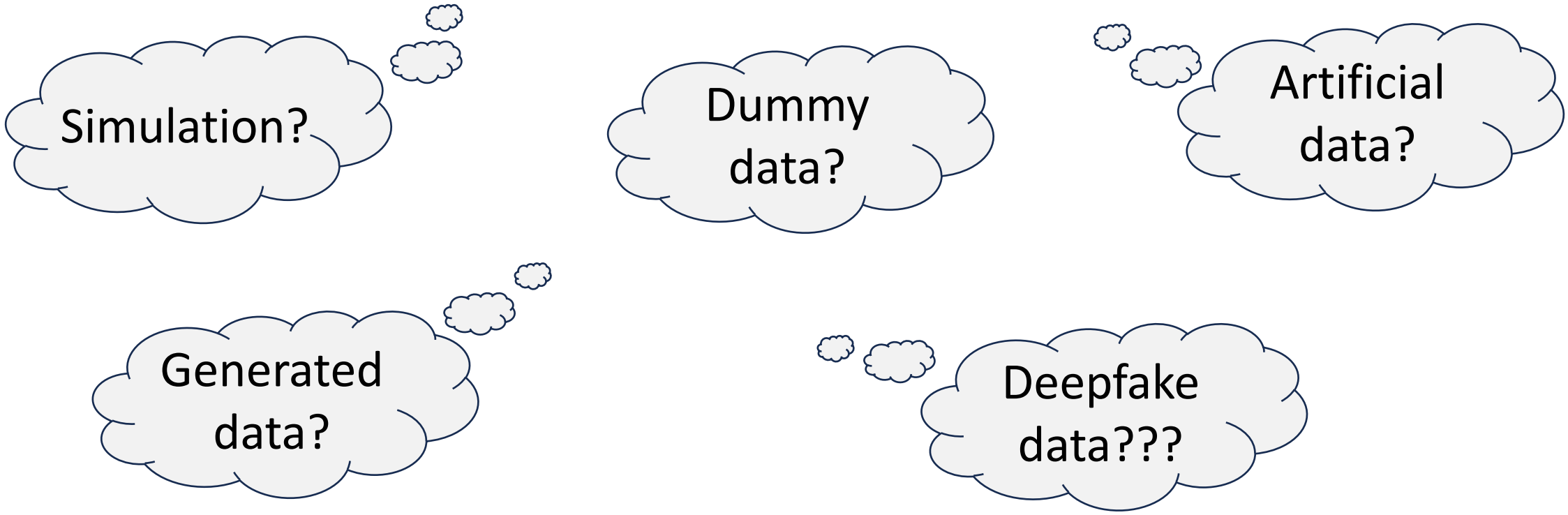
16 September 2024

Christopher Jones

UNECE

modernstats BY HLG - MOS

What is Synthetic Data?



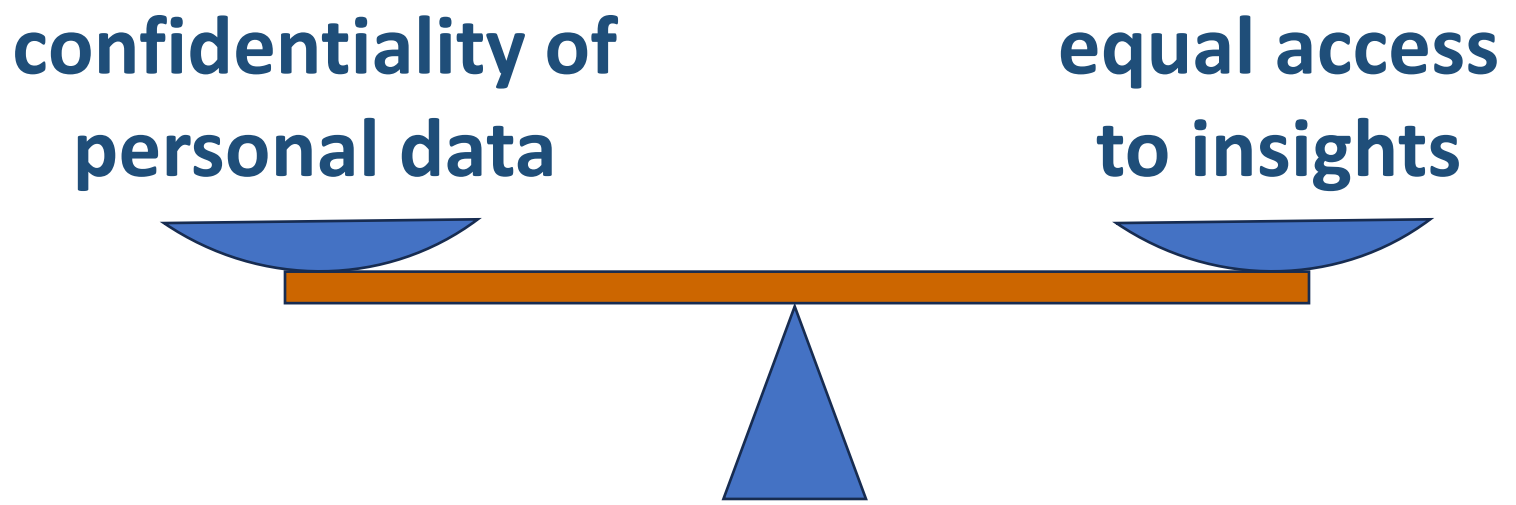
Think of it as: Artificial data that mimics the properties of real data

UNECE

modernstats BY HLG - MOS

What problem does Synthetic Data solve?

Producers of statistics have a difficult balancing act…

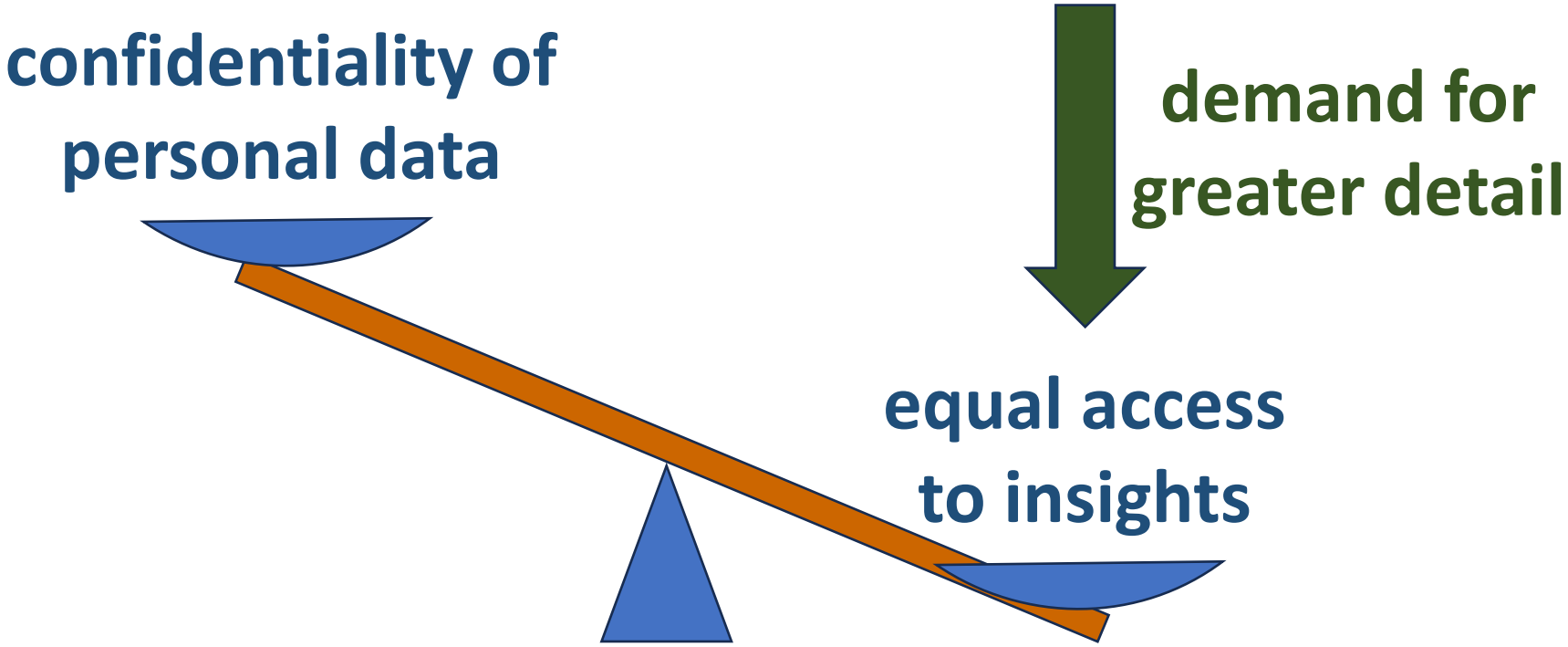


UNECE

modernstats BY HLG - MOS

What problem does Synthetic Data solve?

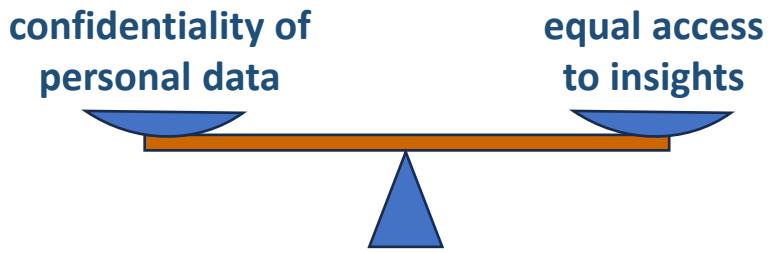
Producers of statistics have a difficult balancing act…



UNECE

modernstats BY HLG - MOS

The Problem: Confidentiality

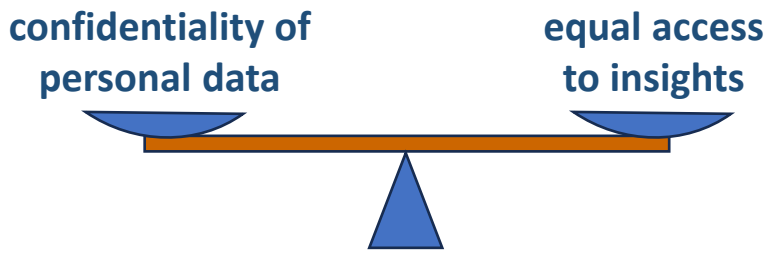


Main techniques for avoiding disclosure:

1. Removing names etc. from record-level datasets

2. Tabulating aggregates (averages, totals,..) with breakdowns by other variables.

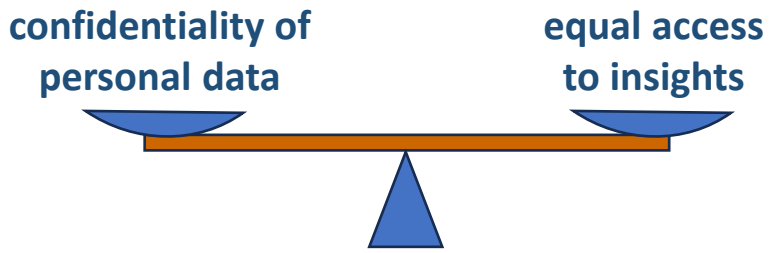
The Problem: Confidentiality



Main techniques for avoiding disclosure:

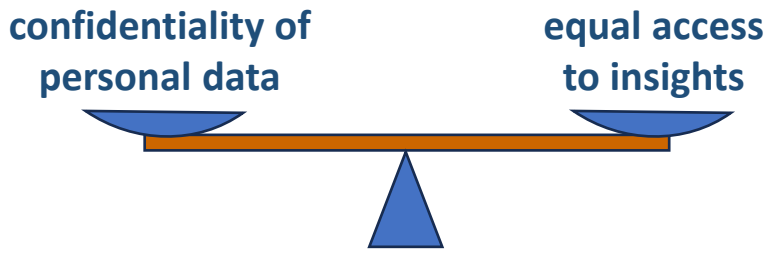
1. Removing names etc. from record-level datasets
    - Might not prevent a record being linked to a person
2. Tabulating aggregates (averages, totals,..) with breakdowns by other variables.
    - A user might figure out more detail than they should
    - Less useful for some analyses

The Problem: Equal Access



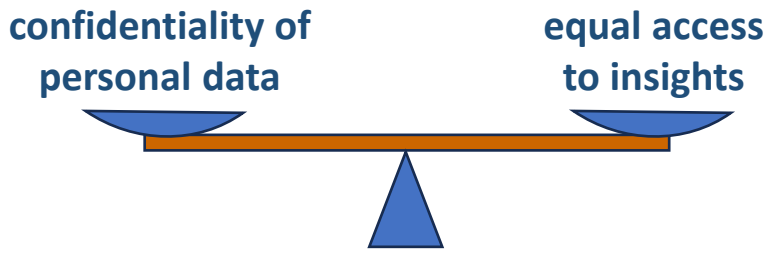
- Often not feasible to grant **equal access** to analytical insights
  - Instead trusted researchers given access to confidential microdata
  - Others can only access the published tables (less detailed)

The Problem: Equal Access



- Often not feasible to grant **equal access** to analytical insights
  - Instead trusted researchers given access to confidential microdata
  - Others can only access the published tables (less detailed)
- Much of the value contained in the microdata may not be utilised
  - I.e. for **all** insights contained being used to inform society's decisions
  - Harder for those outside academia/govt to access microdata

The Problem: Equal Access



- Often not feasible to grant **equal access** to analytical insights
  - Instead trusted researchers given access to confidential microdata
  - Others can only access the published tables (less detailed)
- Much of the value contained in the microdata may not be utilised
  - I.e. for **all** insights contained being used to inform society's decisions
  - Harder for those outside academia/govt to access microdata
- Due to the duty to protect confidentiality

A third option: Separate ***data*** from their analytical ***insights***?

This is what Synthetic Data can do…

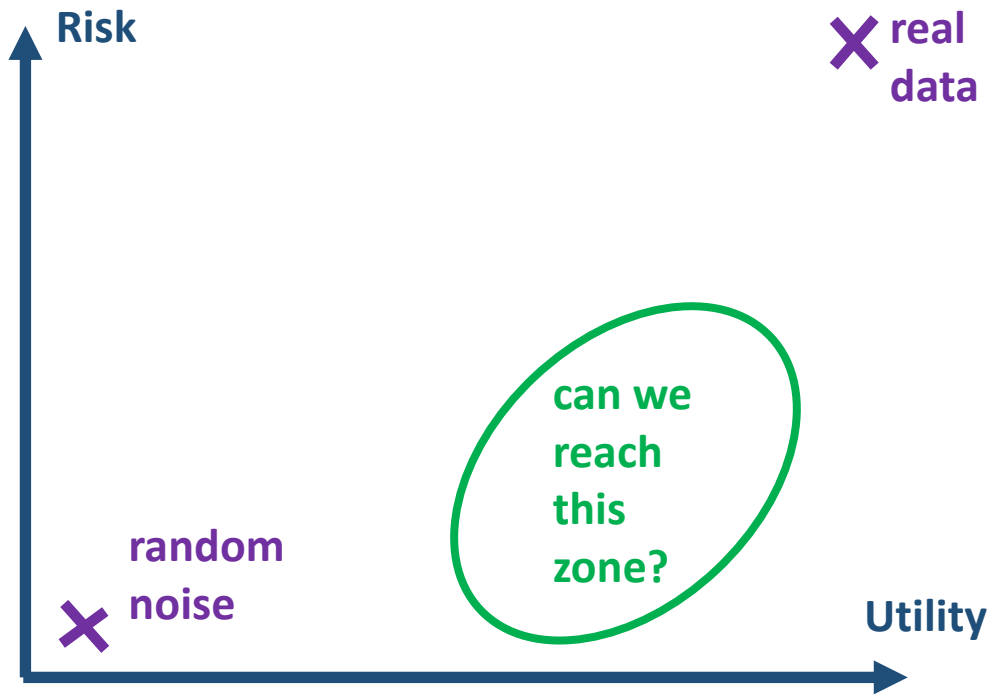
…sometimes

…to a certain extent

- The basic idea is:
  - Making insights derived from synthetic data **as close as possible** to those obtained from real data
  - But also making the synthetic data **as hard as possible** to identify people from

A third option: Separate ***data*** from their analytical ***insights***?

…but still a trade-off between risk and utility for a given Synthetic Data set



Easier to optimise utility / risk when we know how Synthetic Data will be used

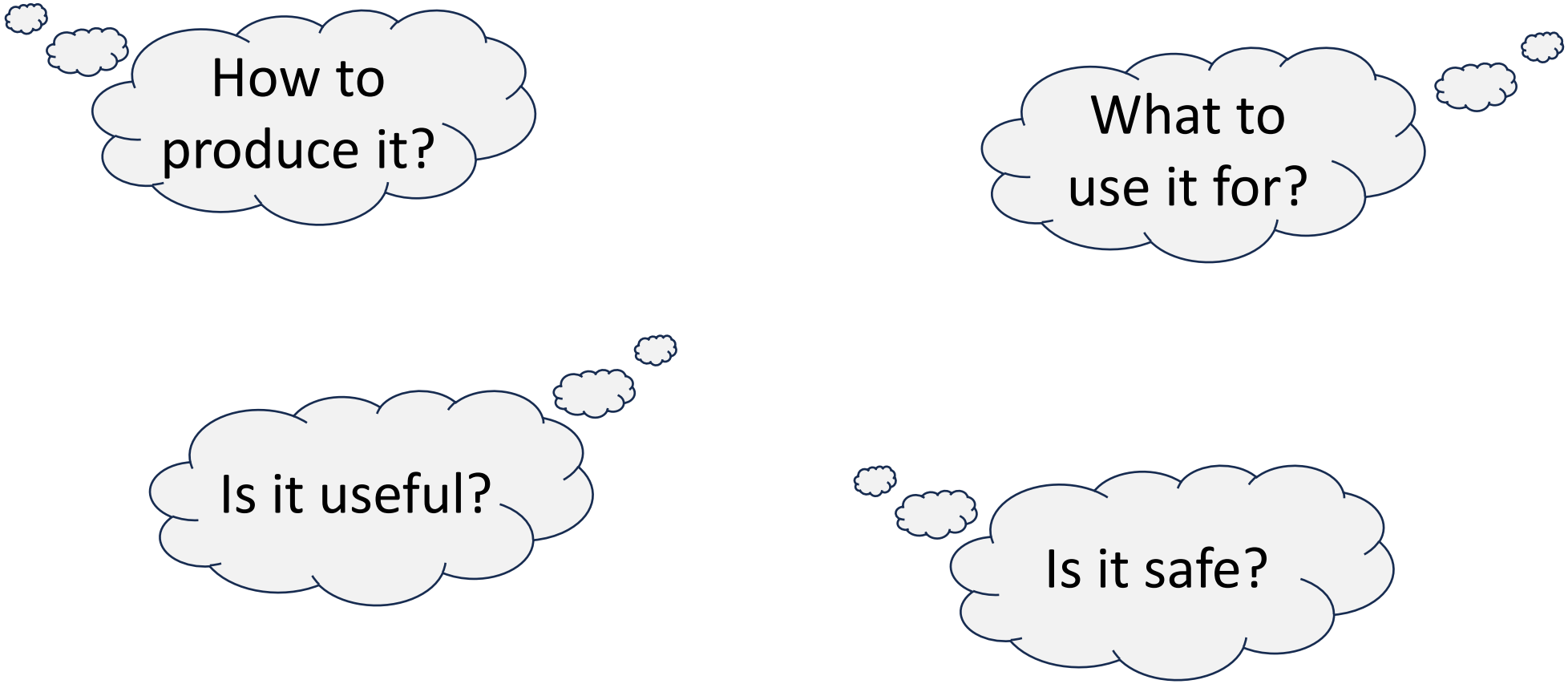
…because there are different ways they can be measured, depending on the use case

But what if we don't know in advance how it will be used?

UNECE

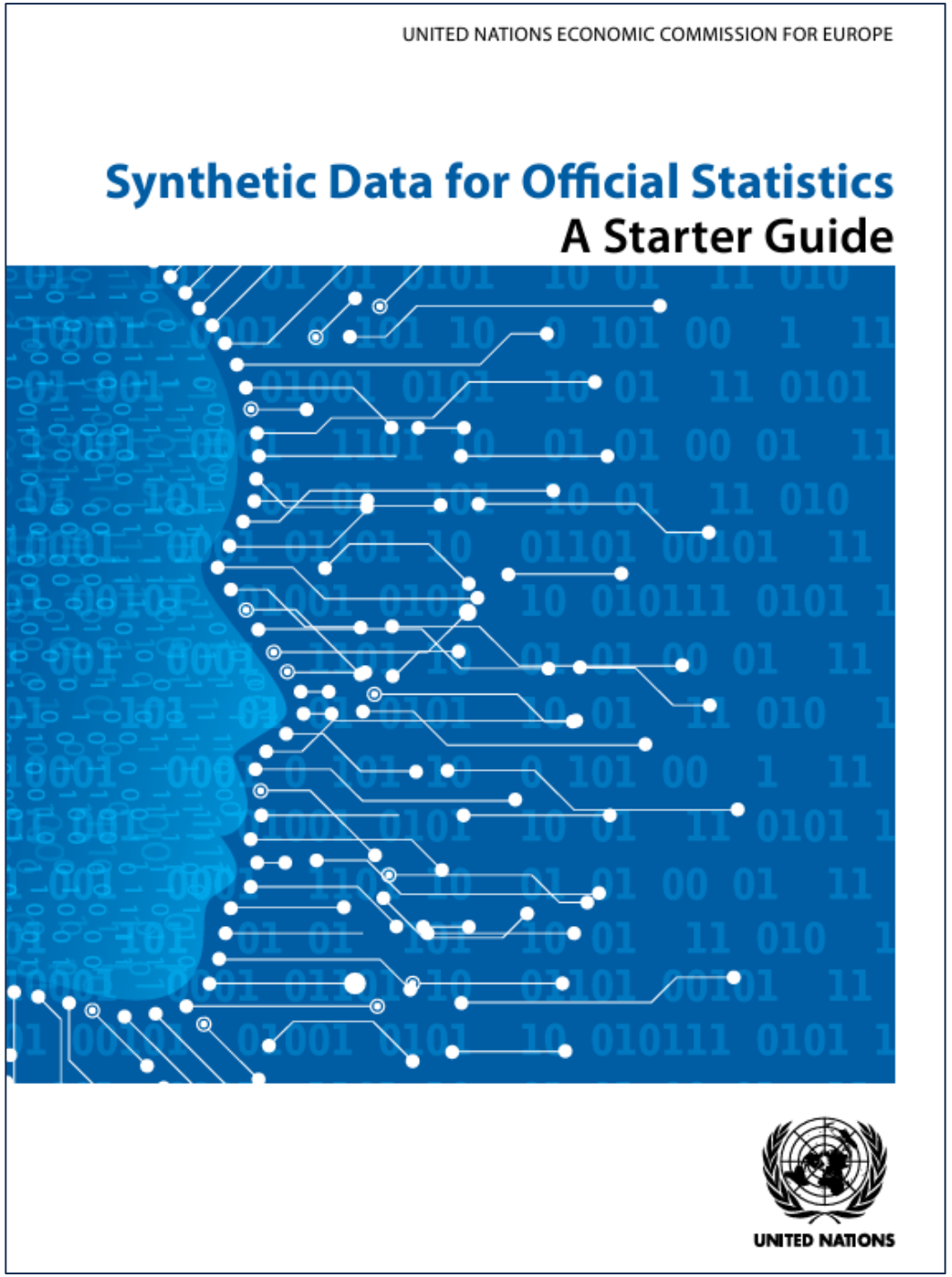
modernstats BY HLG - MOS

Synthetic Data: The need for Guidance



UNECE

modernstats BY HLG - MOS



- Introduces concepts
- Entry points into wider literature
- Use Case examples
- Methodologies for making SD, e.g.:
  - Pseudo-likelihood, Deep learning/GANs,....
- Pros and cons of different methods
- Ways of measuring risk and utility
- Software applications

modernstats
BY HLG - MOS

Thanks for your interest!

If you'd like a **free printed copy** of the Starter Guide, or info about future synthetic data activities, this can be indicated via the QR code: