SESSION 1: REGULATING PRIVACY RISKS FROM SYNTHETIC DATA

ASSESSING THE PRIVACY RISKS OF SYNTHETIC DATA AND OTHER PETS

uOttawa

Replica Analytics
AN AETION COMPANY

Presented by:

Paul Comerford,
Principal Technology Adviser,
Information Commissioner's Office (ICO)

Assessing the Privacy Risks of Synthetic Data

Paul Comerford
Principal Technology Adviser

Anonymisation and encryption
30th November 2023

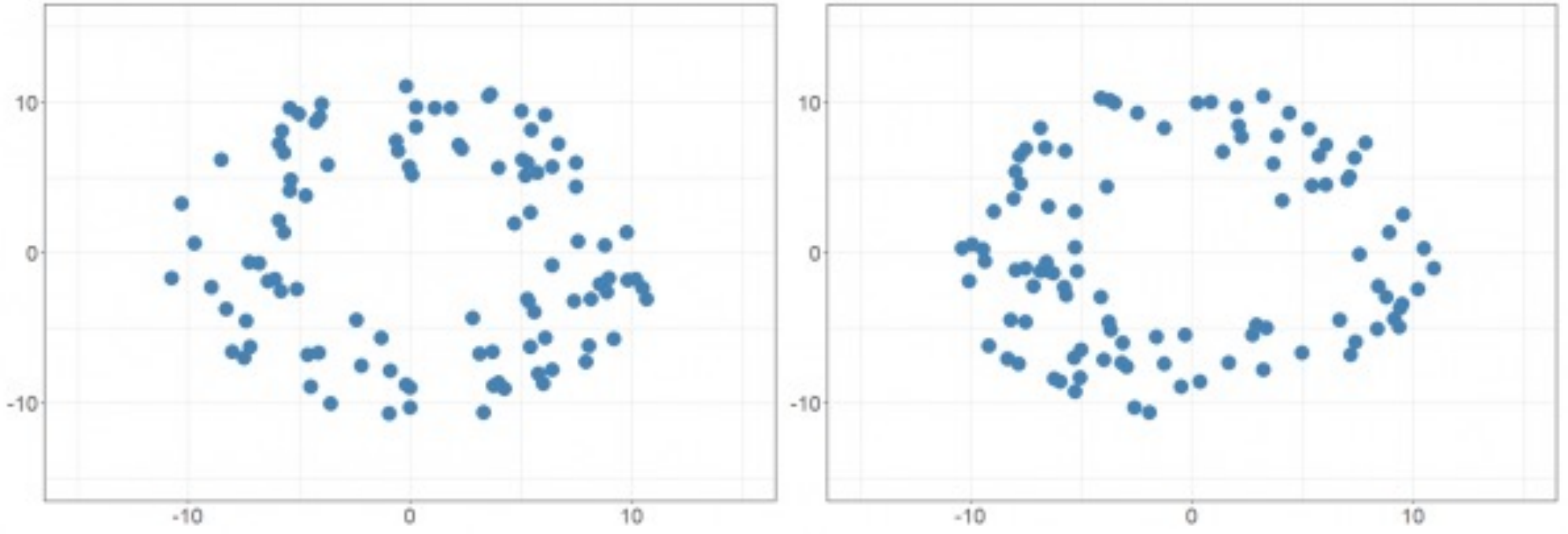
ico.
Information Commissioner's Office

ico.org.uk

Agenda

- How does synthetic data assist with data protection compliance?
- What are the risks associated with using synthetic data?
- Is synthetic data anonymous?
- What else are we doing in this space?
- Questions

How does synthetic data assist with data protection compliance?



Original data

Synthetic data

The synthetic data retains the structure of the original data but is not the same

What use cases are there for synthetic data?

- Use cases that require access to large amounts of information (eg model training, research and development).
- Synthetic data provides realistic datasets in environments where access to large real datasets is not possible.
- synthetic data can be used at the training stage to reduce the amount of personal information used to train artificial intelligence.
- CDEI use case repository
- Repository of Use Cases | PETs Adoption Guide (cdeiuk.github.io)

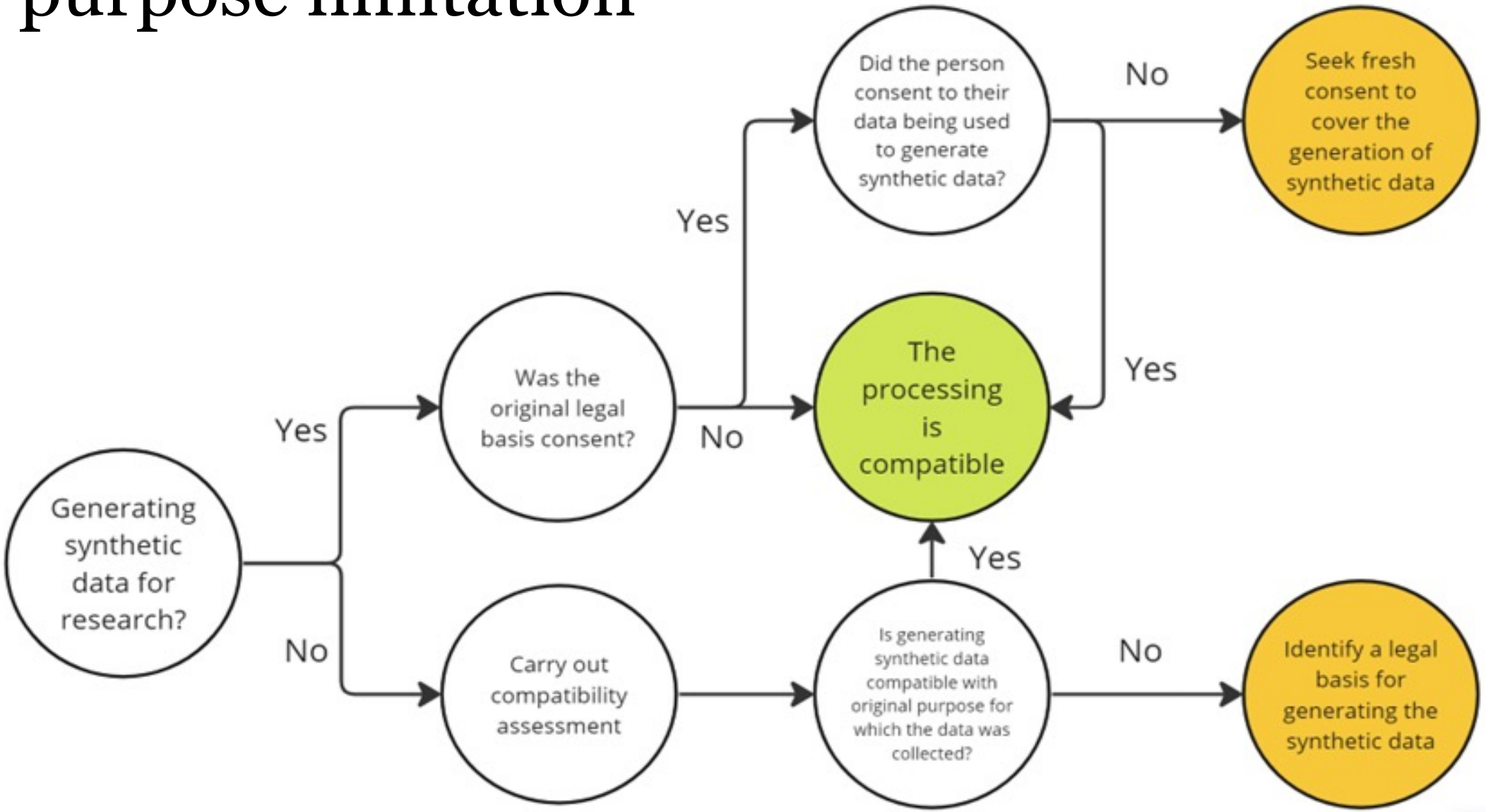
What data protection considerations are there?

- Legal basis considerations, including questions of purpose limitation
- Transparency considerations
- Data minimisation considerations
- Security considerations
- Accountability considerations
- Identifiability assessments

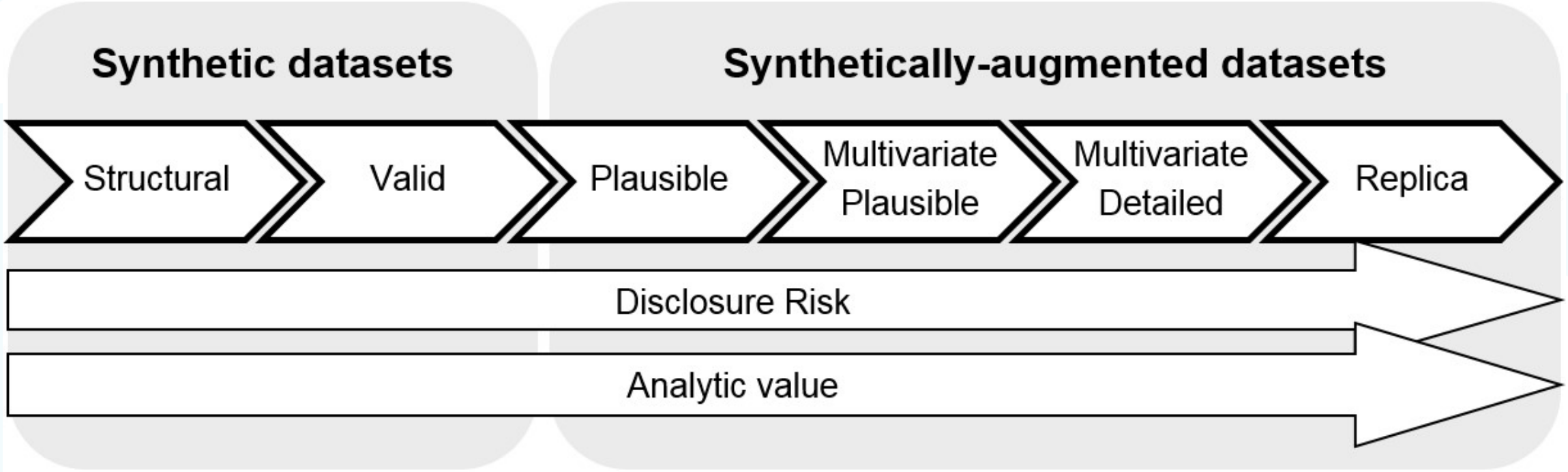
Anonymisation as a processing activity

- anonymous synthetic data is not subject to the GDPR
- the process of transforming identifiable data into anonymous data is processing
- "*any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction*"

Synthetic data generation: legal basis and purpose limitation



What are the risks associated with using synthetic data?



Source: Office for National Statistics

Synthetic data and bias

- If you are generating synthetic data from personal information, any inherent biases in the information will be carried through.
- You **should:**
- Ensure that you can detect and correct bias in the generation of synthetic data, and ensure that the synthetic data is representative;
- Consider whether you are using synthetic data to make decisions that have consequences for people (ie legal or health consequences). If so, you **must** assess and mitigate any bias in the information.
- De-biasing

Is synthetic data anonymous?

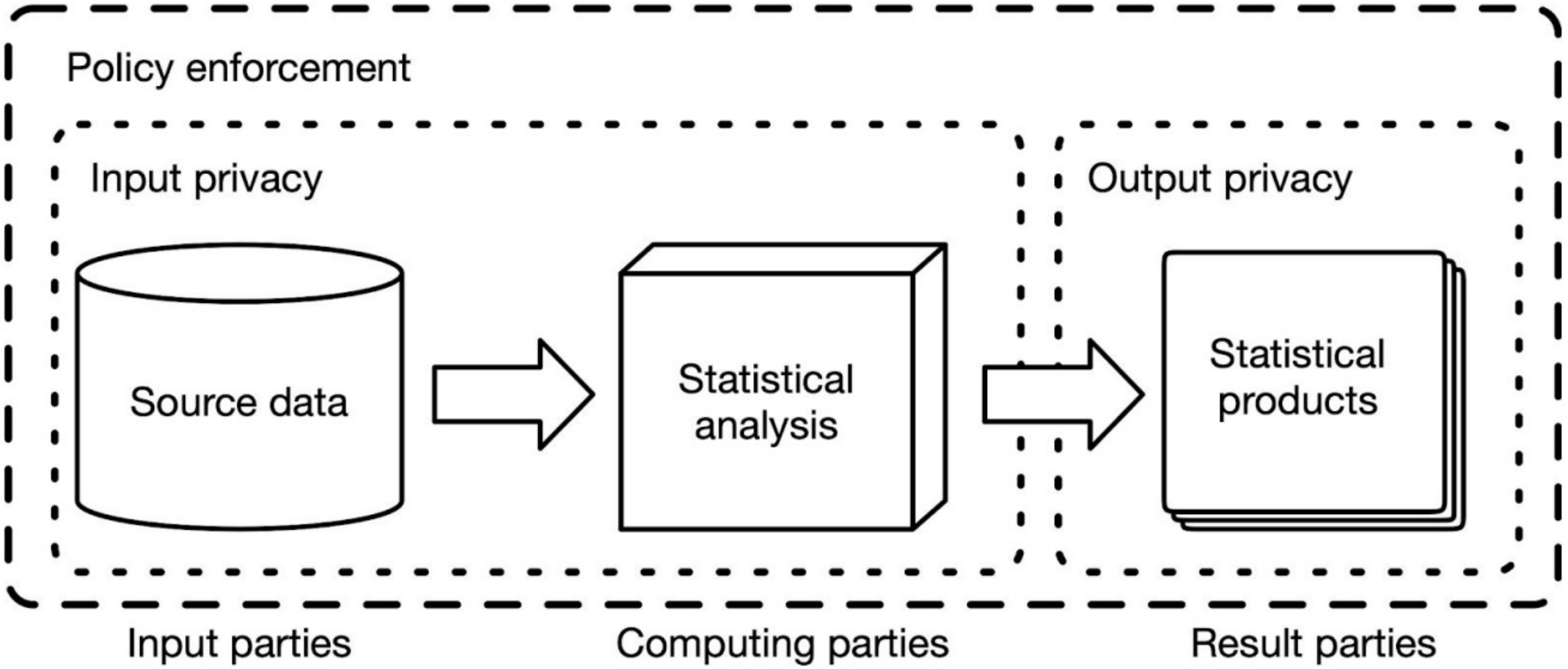
- *"To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person, to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."*

- *"The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."*

Is synthetic data anonymous?

- Some synthetic data generation methods have been shown to be vulnerable to model inversion attacks, membership inference attacks and attribute disclosure risk.
- You **could** protect any records containing outliers from these types of linkage attacks with other information through:
- suppression of outliers (data points with some uniquely identifying features); or
- differential privacy with synthetic data.
- However, it may reduce the utility of the information and introduce a degree of unpredictability in the characteristics of the information.

What are the privacy properties of synthetic data?

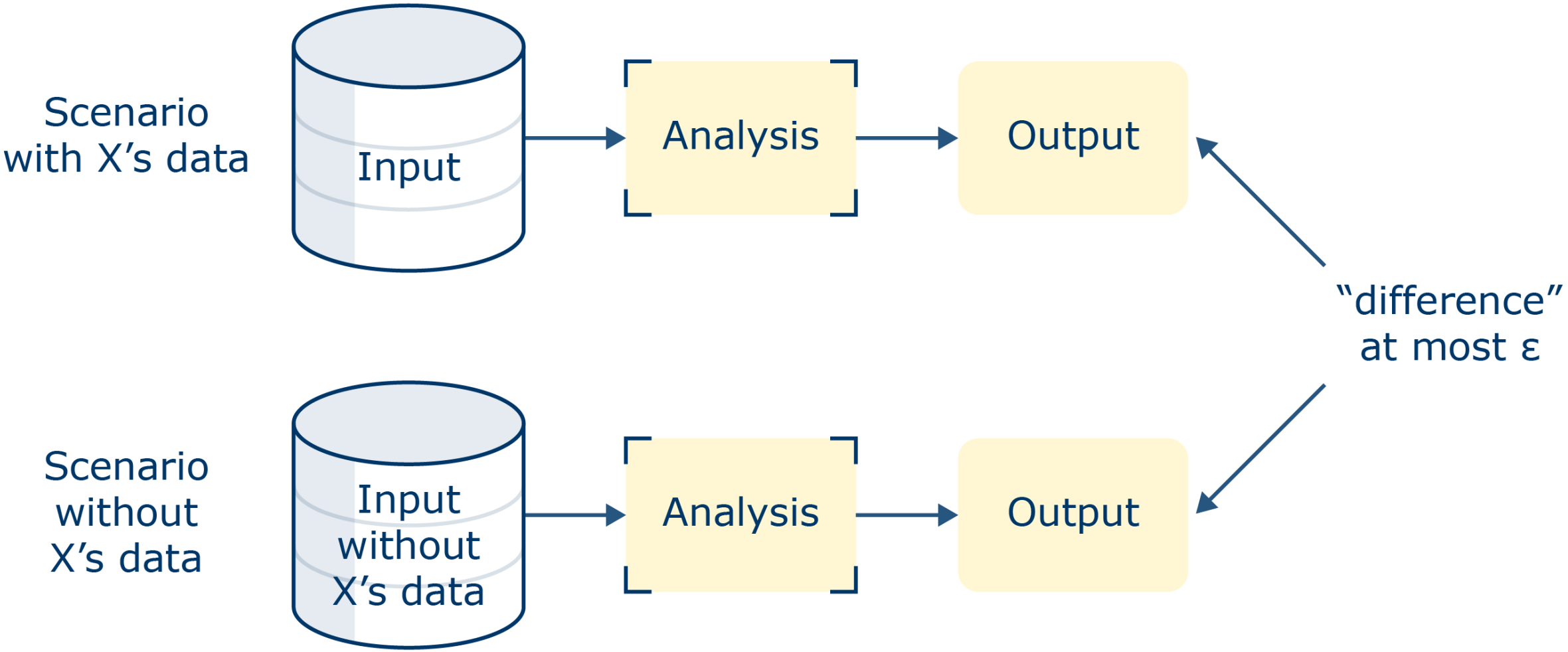


UN Guide on Privacy-Enhancing Technologies for Official Statistics // Task Team on Privacy Enhancing Techniques — UN-CEBD

Mitigating the risk of identification

- only include the properties needed to meet the specific use case,
- suppression of outliers (data points with some uniquely identifying features);
- Additional technical and organisational measures
- differential privacy with synthetic data

What about differential privacy?

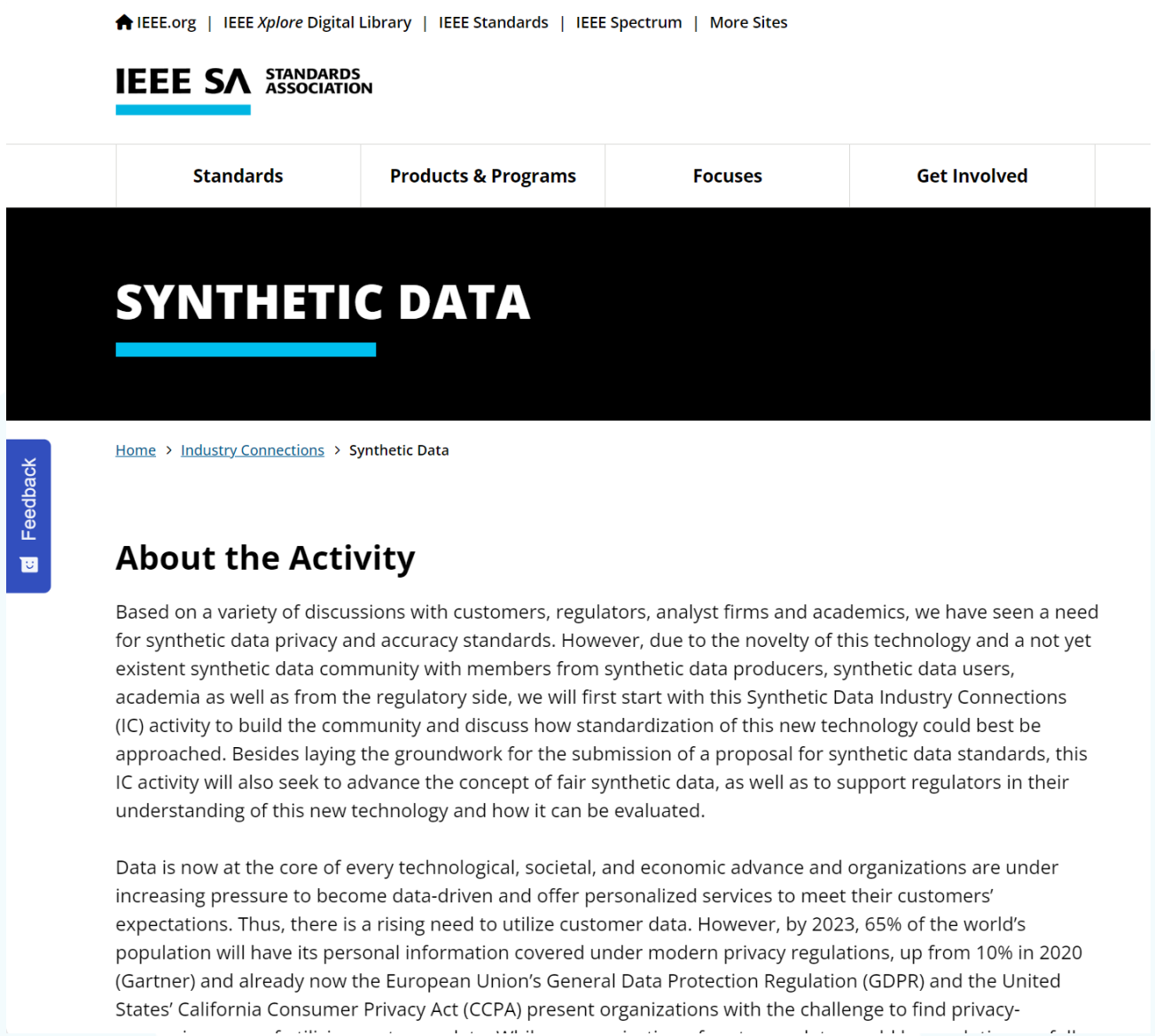


What risk mitigation measures are appropriate?

- You **should** consider the purposes and context of the processing when you decide what risk mitigation measures are appropriate
- You **could** consider less stringent measures (eg adding less noise than if you were releasing the information to the world at large when using differential privacy).
- If you don’t have expertise in house - consult an external expert in setting an appropriate privacy budget when using differential privacy.

Synthetic data case studies

Standardisation efforts



Thank you! - Questions

Paul Comerford
Anonymisation & Encryption team
paul.comerford@ico.org.uk