AstraZeneca

Synthetic Data Use:

Exploring use cases to optimize data utility

Stef James, Senior Director, Clinical Data Insights at AstraZeneca

9th March 2023

What will I cover today?

Peer reviewed paper: James, S., Harbron, C., Branson, J. Sundler, M. Synthetic data use: exploring use cases to optimise data utility. Discov Artif Intell 1, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

Acknowledgement: to the PSI Data Sharing Special Interest Group

The definition

'Synthetic data is data generated by simulation, based upon and mirroring properties of an original dataset.'

Uses of synthetic data

7 key use cases we will explore from a pharmaceutical context that will demonstrate the utility of synthetic data

Techniques to mask data

There are various methods to mask personal data, including anonymization and data synthesis.

Methods to produce synthetic data

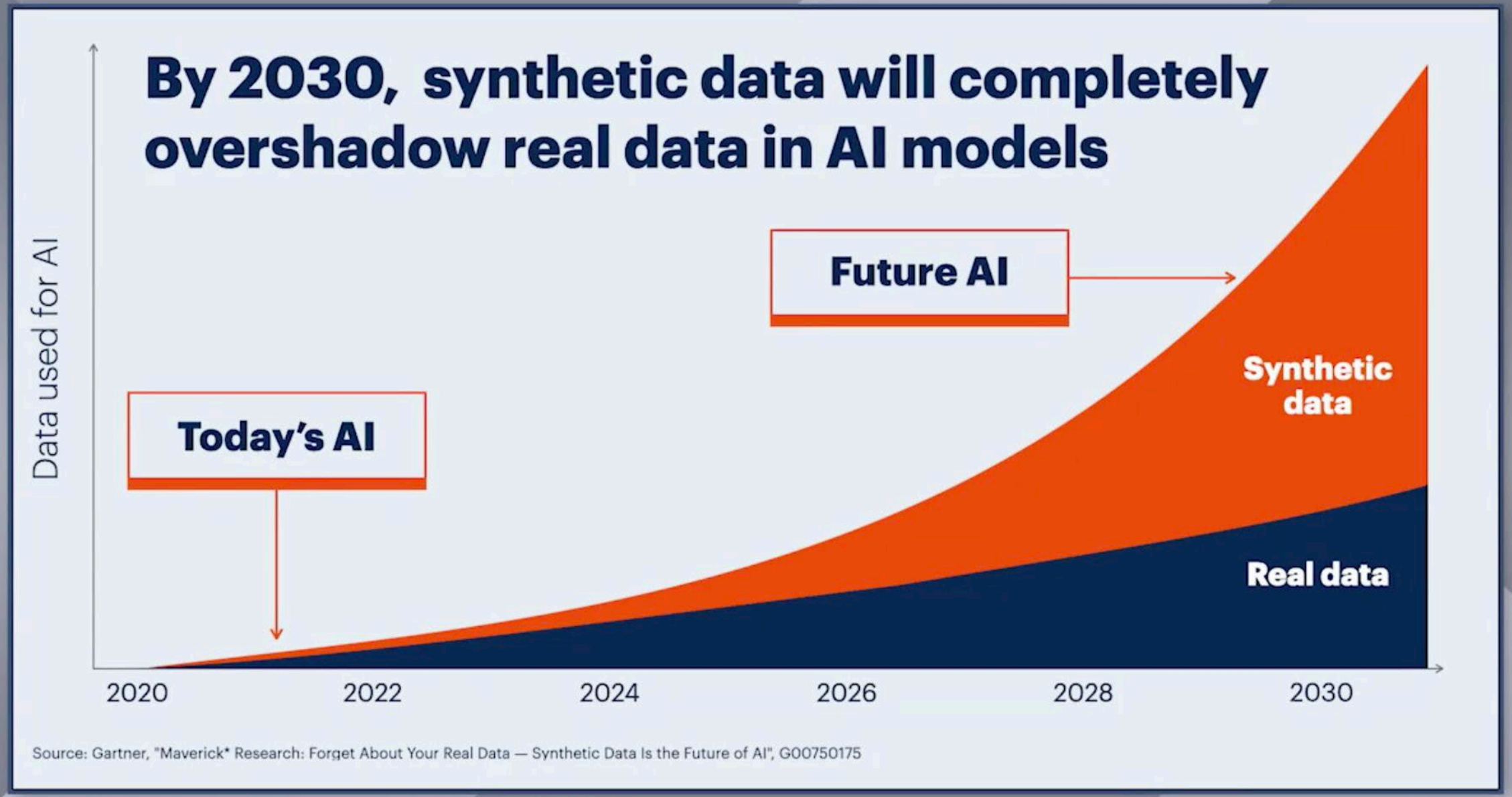
The different methods to produce synthetic data, which can be a combination, such as embedding or GAN.

Application of utility measurements/ parameters

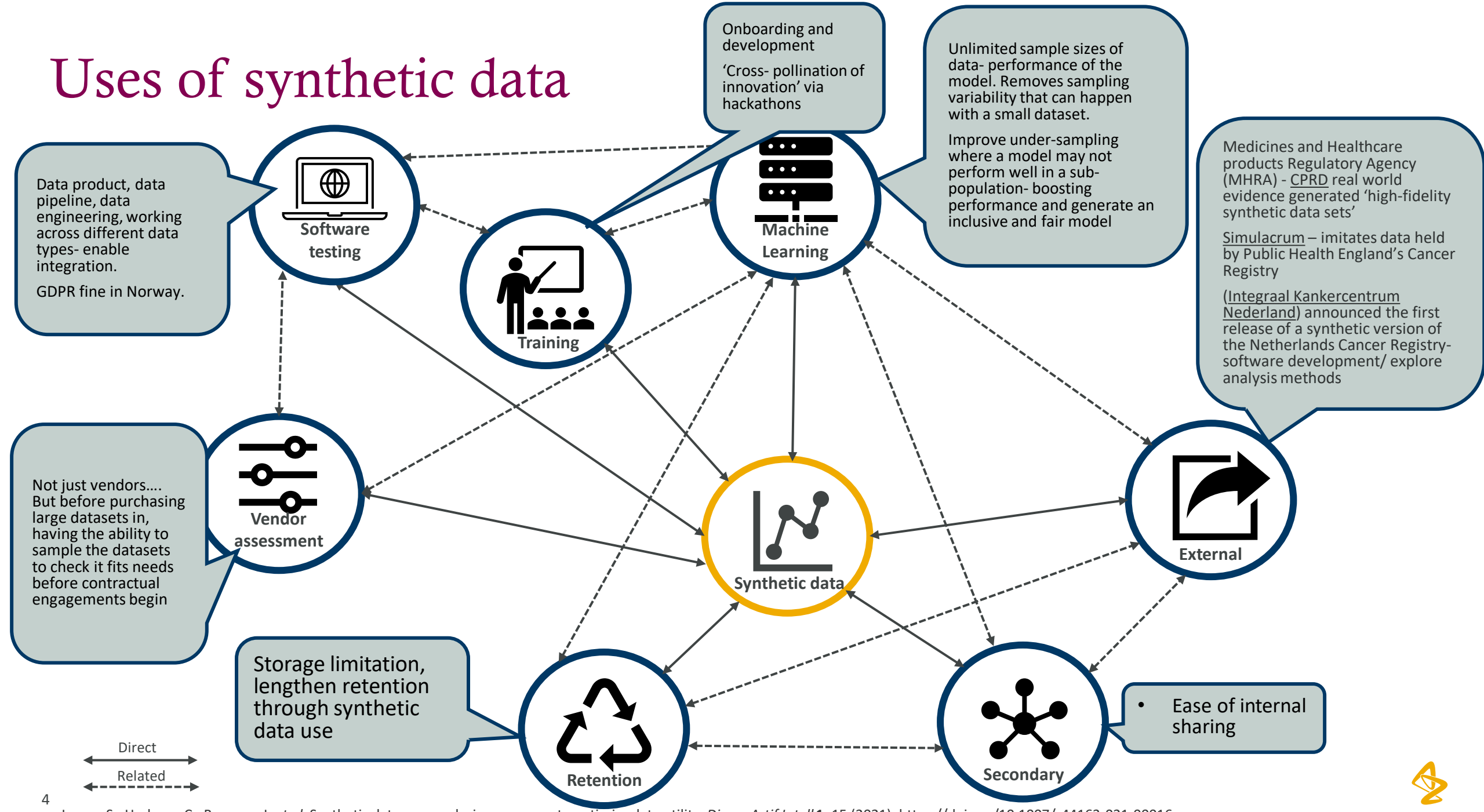
The measurement, check and balance on the effectiveness and quality of synthetic data generation.

Consideration of the challenges to adoption

The considerations specifically for the pharma industry, although this may transfer to other industries also where sensitive data is a consideration.


By 2030, synthetic data will completely overshadow real data in AI models
Data used for AI
Future AI
Today's AI
Synthetic data
Real data
2020
2022
2024
2026
2028
2030
Source: Gartner, "Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI", G00750175

Uses of synthetic data



James, S., Harbron, C., Branson, J. *et al.* Synthetic data use: exploring use cases to optimise data utility. *Discov Artif Intell* **1**, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

Request form for CPRD database

What is your intended purpose for accessing the synthetic data? Please tick all that apply.	
Training of machine learning algorithms	
Testing/validation of machine learning algorithms	
Developing, populating or testing models	
Methodological research	
Sample size boosting	
Feasibility counts	
To support regulatory submissions	
Building medical software applications	
Trend simulation/analysis	
Sample dataset to understand CPRD Aurum data structure and utility	
Data management teaching/training resource	
Develop/validate/test analytics tools for use with CPRD Aurum	
Improve bespoke CPRD Aurum application interfaces/algorithms	
Develop machine learning workflows for application to anonymised CPRD Aurum data	
Other (please state below)	

Note:

- Training
- Testing
- Machine learning
- Sample size boosting
- Support regulatory submissions

Synthetic datasets are owned by the Medicines and Healthcare products Regulatory Agency (MHRA)

Common techniques adopted to mask data

Unmasked patient data

Real patient data contains sensitive health data which is **subject to GDPR**

Jill 25 years | Olive 42 years | Sara 31 years | Jack 62 years | Max 50 years

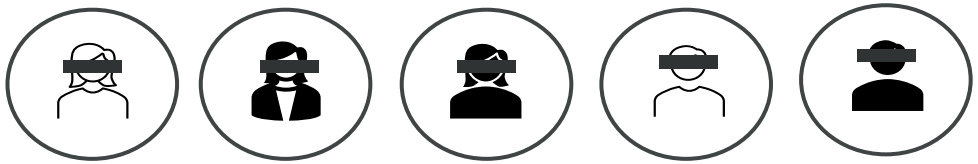
Pseudonymisation

Names and personal information removed or encrypted which is subject to GDPR- must have a **legal basis**

#001 25 years | #002 42 years | #003 31 years | #004 62 years | #005 50 years

Anonymisation

Aggregation and statistics: sanitation applied to remove personally identifiable information, **excluded from GDPR**



N= <11

Age range= 21-65

Synthetic data

Artificially generated patients, **excluded from GDPR**

Jim 32 years | Steven 21 years | George 57 years | Molly 44 years | Caroline 64 years

James, S., Harbron, C., Branson, J. *et al.* Synthetic data use: exploring use cases to optimise data utility. *Discov Artif Intell* **1**, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

Methods to produce synthetic data

Embedding

Uses two neural networks to help generate data. Neural networks are **encoder** and **decoder**. There is a need to train the two networks to encode and decode the data. The outcome provides synthetic data

Generative adversarial network (GAN)

Uses a **generator** and **discriminator**. The discriminator attempts to figure out what records are real which is produced by the generator. The generator and discriminator are trained over time. This provides a propensity score. This is a machine learning/ deep learning method.

Sequential synthesis

This method synthesizes datasets **variable** by **variable**. The more variables in a dataset, the more the sequence of the variables will need to be optimized

More methods exist e.g. tree based or copula based methods, one or several methods can be used and combined to produce a synthetic dataset. There would be a requirement to detail the methods used.

James, S., Harbron, C., Branson, J. *et al.* Synthetic data use: exploring use cases to optimise data utility. *Discov Artif Intell* **1**, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

Utility measurements/ parameters

Distribution comparison

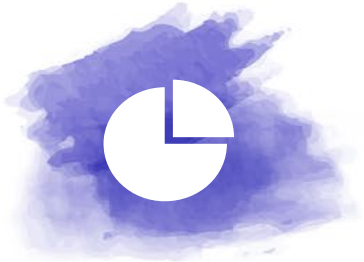
Measures the distribution comparison between the real and synthetic data to measure representation per variable

Hellinger distance

Measures the distance between the real dataset and the synthetic data.
0 = equal, 1 = far apart

Prediction accuracy

Provides a comparison of future modelling and prediction, comparing machine learning models used on real data versus the synthetic dataset



Distinguishability

A measure which tries to determine in the model developed, whether the data is real or synthetic.
0 = perfect synthesis 1 = easily identified

Area under the receiver operating characteristic (AUROC)

Determines the discrimination in the dataset. ROC measures the probability and AUC measures the degree of separability between the real and synthetic dataset. The higher the AUC the better the model is at distinguishing between patients with the disease and no disease

Lots in the literature

And others!

James, S., Harbron, C., Branson, J. *et al.* Synthetic data use: exploring use cases to optimise data utility. *Discov Artif Intell* **1**, 15 (2021). https://doi.org/10.1007/s44163-021-00016-y

The considerations

The over-arching consideration is to ensure the characteristics between the original and synthetic data set are comparable

Tiers of synthetic data

- **Tiered approach** towards synthetic data, where lower fidelity datasets can be produced for certain use cases.

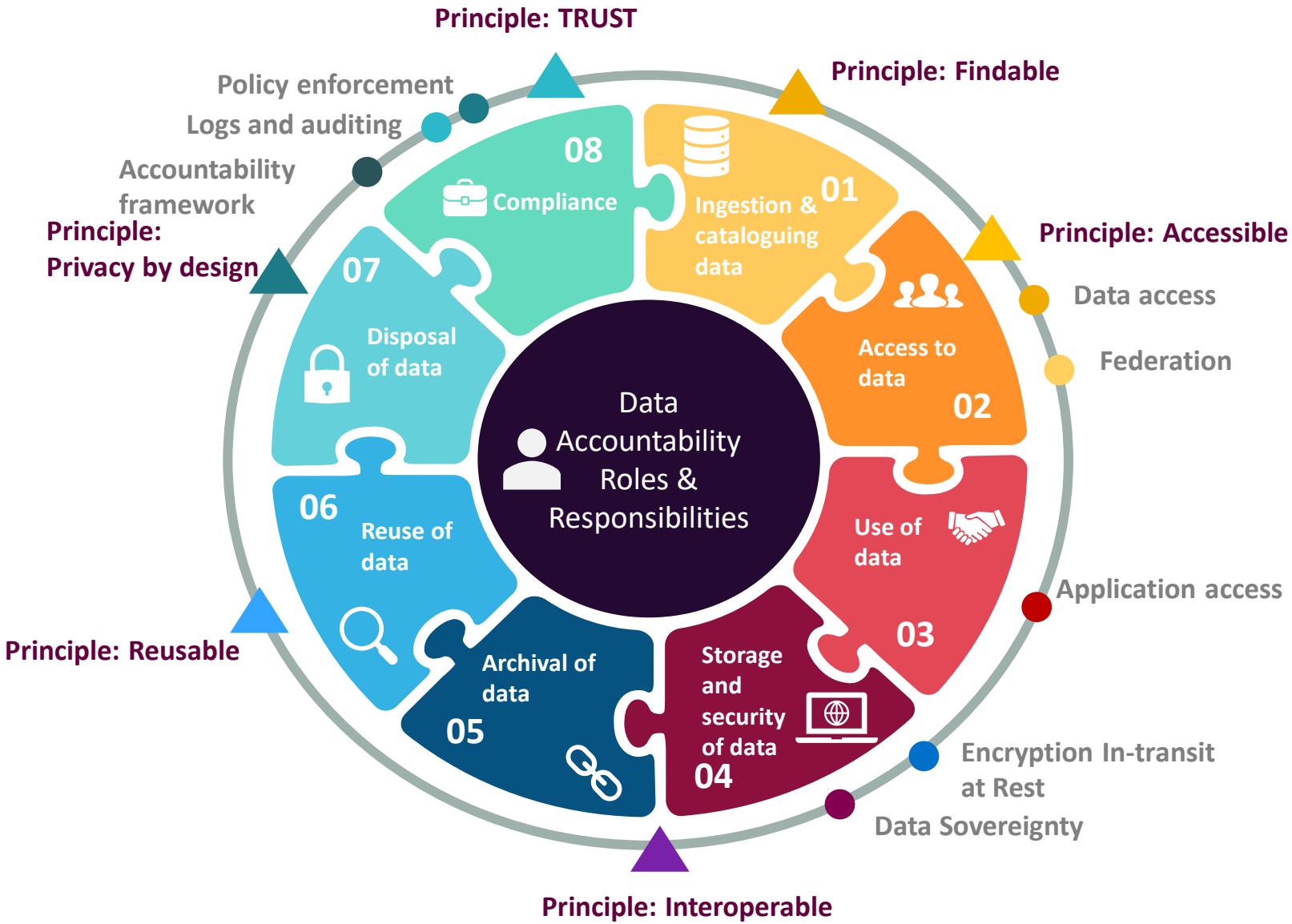
Technical and organizational measures

- **Business processes and skilled individuals** to avoid pitfalls when using synthetic data such as under/over-fitting models.
- **Bridge the gap between privacy and utility of the data**. This is a specialised skillset which is likely to need several different departments within an organisation.
- **Industry and regulatory acceptance**.
- **Privacy assurance assessment/report:** data protection by design approach.

Production considerations

- **Effort and potential expense** in producing representative synthetic data.
- **Scientific and medical acceptance-** could be used for de-centralised trials, to enlarge the overall cohort size. Aid recruitment in under-represented regions/populations to increase trial diversity as an example. Identification of vulnerabilities in trial design.
- **Agreed industry standard**s towards synthetic data, industry come together and create a central repository/ mechanisms for synthetic data sharing.

Data Protection in the Data Lifecycle: Business Processes



Remember where we began…

Gartner:

‘Synthetic data will **overshadow** using real data in AI and models’

2030

What do you think?

Will synthetic data overshadow the use of real data?