Smart Data Foundry, Bayes Centre, 47 Potterrow, Edinburgh EH8 9BT

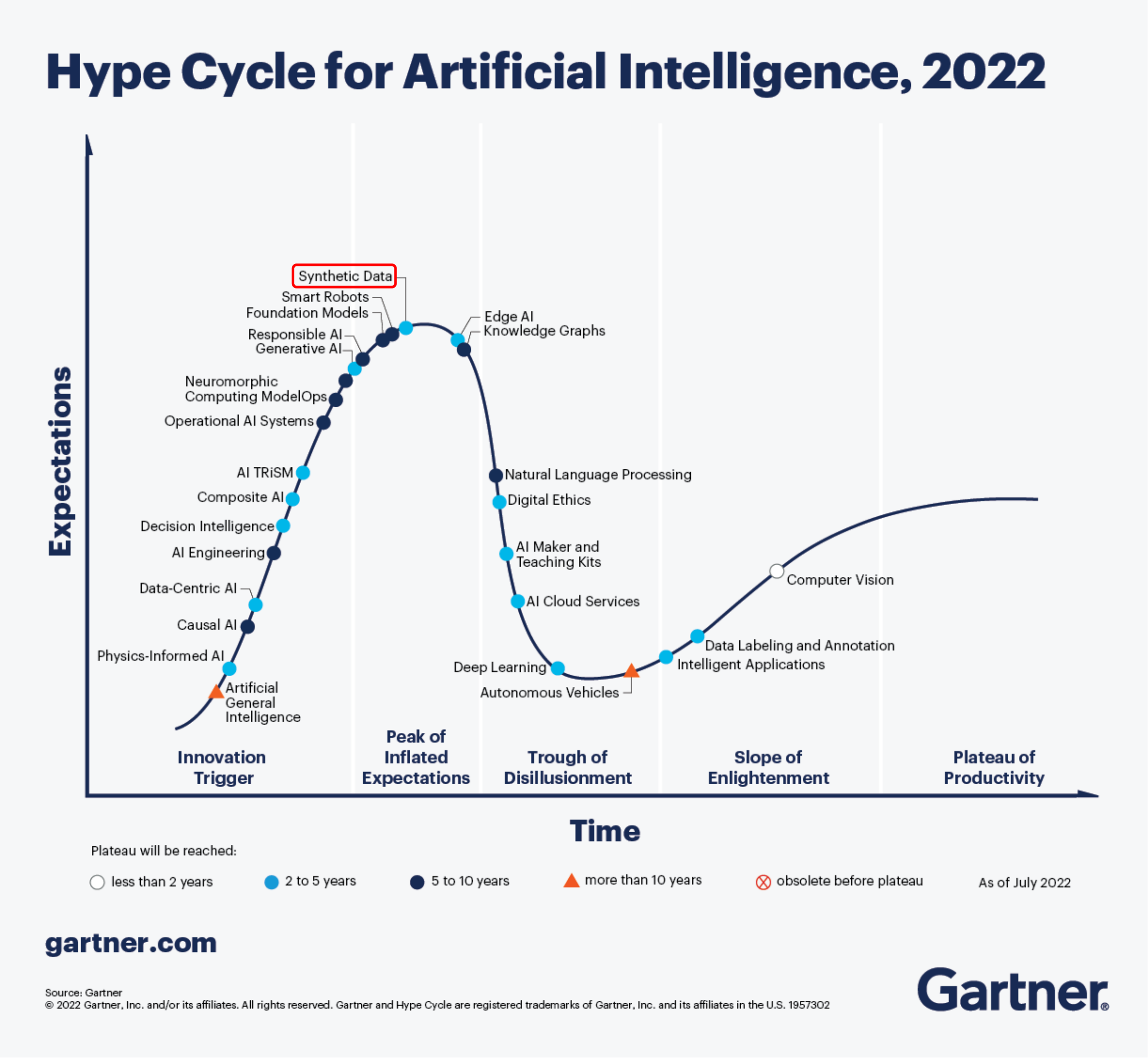
A Comparison of Synthetic Data Generation Approaches

SIMULATION & LEARNING-BASED SYNTHESIS

3rd APRIL 2023

SMART DATA FOUNDRY

There is a great deal of excitement at the moment around synthetic data. In fact, synthetic data is literally higher on the peak of inflated expectations than any other technology in Gartner's recent Hype Cycle, shown below.[1]



Hype Cycle for Artificial Intelligence, 2022

Source: Gartner

© 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1957302

But what is synthetic data? Why is it useful? How is it generated? And is it really all hype, or is it ready for mainstream use? We'll address these challenges and try to answer those questions in this paper.

[1] Source: https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle

What is Synthetic Data?

Real data contains information about real people or entities, real events, and real interactions. Synthetic data is the same, except some or all the people, entities, events, or interactions are artificial.

Artificial data is nothing new and pre-dates computers: it's routine to generate dummy data – usually having the right structure but not the right patterns – for the purposes of testing software for correctness and performance. Data like this is also sometimes constructed for use in product demonstrations and proof-of-concepts, or for new-user on-boarding.

The reason synthetic data is such a hot topic is that ever-increasing computing power and data volumes have combined with progress in machine learning, AI, and other algorithmic approaches to make possible much higher quality artificial data than ever before. This can:

- facilitate use cases that would be problematic using real data,
- enable applications that would otherwise be hard or impossible,
- augment real data to enable analyses that would otherwise be more difficult.

Why is Synthetic Data useful?

There are multiple reasons for the growing interest in synthetic data, and the different applications and motivations make different approaches to generation suitable, depending on the problem to be solved.

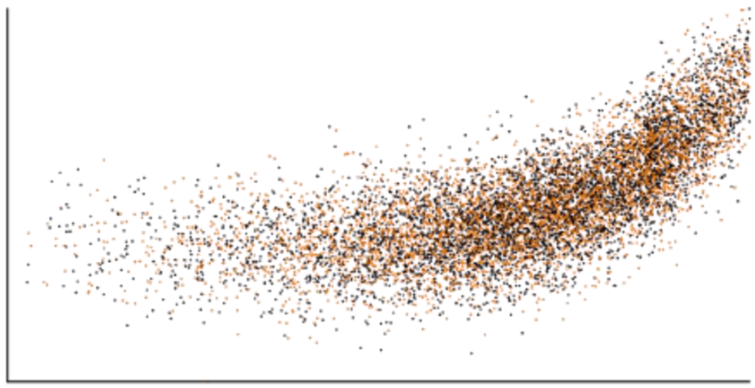
Privacy, confidentiality, and disclosure

One common set of reasons for using synthetic data concerns privacy, confidentiality, ethics, and disclosure risk. In these cases, there is usually a real dataset that has restrictions on use, retention, access, or distribution. For example:

- Any form of personal data (known in the US as personally identifiable information, or PII data) is normally hard to share or disclose. Data protection legislation, such as GDPR, often permits data only to be used for the purpose for which it was collected. Best-practice data minimisation also dictates that data should be retained only as long as it is required and should only be shared with people who need to access it. Even within a bank, for example, it would be best practice to minimise the number of staff with access to personal data. This is especially true for bulk access, as opposed to the ability to look up a single customer at a time, as, for example, a bank teller needs to be able to do. There are also often restrictions about moving personal data from the jurisdiction in which the data subject resides.

- Even for non-personal data, there can be concerns around confidentiality. For example, an organisation may wish to engage a third party to work on sensitive data without disclosing the real data.

In these circumstances, an ideal synthetic dataset would contain the same meaningful patterns as the original dataset but would not include any real information about any person or other sensitive entity. What this means is discussed further in Figure 1 below, but one good way to think about it is that if you build predictive models (such as credit scores, propensity scores or attrition scores) on the synthetic data and on the real data, the models built on the synthetic data would be as good as (and probably very similar to) those built on the real data. Increasingly, this approach is referred to as generating Synthetic Doubles.



Suppose you randomly assign each point in a large, (real) dataset to one of two partitions—here shown as black and orange. Since the dataset is large, if you build models on the black data and on the orange data, you would expect them to be similar, and to work similarly well on the whole dataset, even though every point is in only one or other of the partitions.

An ideal synthetic dataset (a 'synthetic double') would be 'exactly' as similar to each of the black and orange partitions as they are to each other, but would have no points from either. So a model built on the synthetic data would typically be just as good as one built on either the black or orange partition

Figure 1: Ideal Synthetic Doubles

Data is non-existent, unavailable, or exhibits a systematic problem such as bias

A very different set of motivations for synthetic data exists when no real data is available for the task at hand, or when the available data is inadequate. Examples of this are:

- The goal is to explore one or more scenarios that have not occurred, so no data about them exists (e.g., the launch of a revolutionary new product, the replacement of a traditional currency with a cryptocurrency, or the impact of a decision like Brexit before Brexit had occurred).

- The goal is to tackle algorithmic bias, which often results from non-representative training data or training data that reflects flawed and unfair data collection and model-fitting, pre-existing expectations, or other biases.

- Data exists, but its volume needs to be higher, or its coverage needs to be better.

In these situations, the goal is either to produce data that does not exist at all, or to produce a dataset that augments or replaces an inadequate existing dataset.

What are the most important qualities of Synthetic Data?

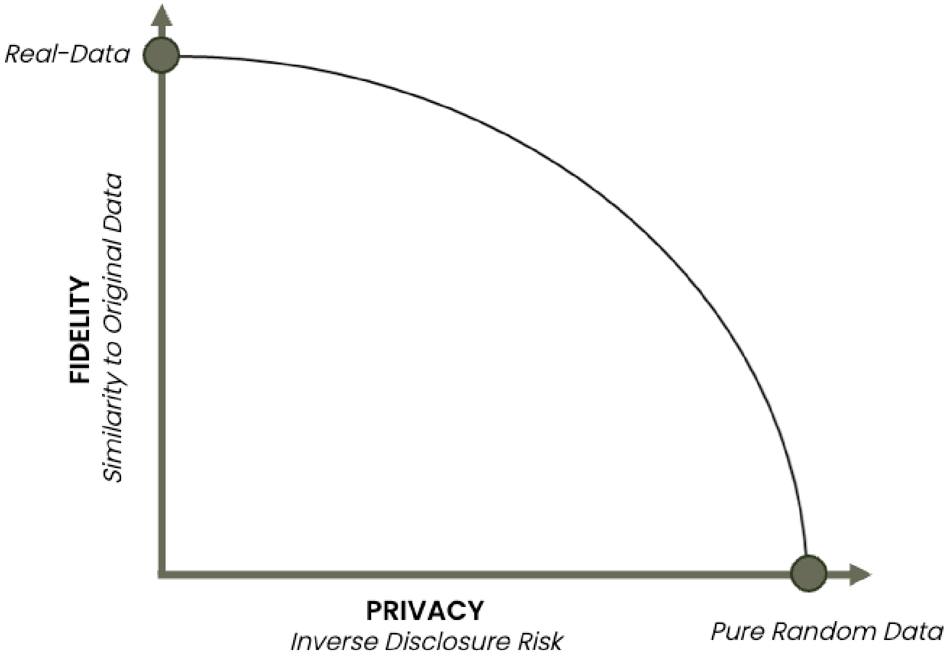
At Smart Data Foundry, we think of three factors when assessing how 'good' Synthetic Data is:

- Fidelity – how similar the synthetic dataset is to the 'real' data. For example, how representative is the data of the people or events you're trying to synthesise? Does it reflect the correlations between variables, and/or how they change over time?

- Privacy – the risk of unintended disclosure of information pertaining to any individual data subjects (or other confidential or sensitive information) that may have been used to manufacture the synthetic data.

- Utility – the 'usefulness' of the synthetic data, and how readily it can be applied to accomplish whatever objectives the user of the synthetic dataset has.

In practice, there is a limit on achieving both high fidelity and privacy, given their inverse relationship, and so it is important to consider and understand the potential trade-off, most likely with an eye to maximising utility at an acceptable level of fidelity and privacy. On the one hand, we may want the synthetic data to be 'as similar as possible' to the original data, in exactly the sense that we described in Figure 1. Measures of fidelity tend to focus on comparing properties of reference data (for example, in learning methods this might be the training data, or some hold-out sample) and the synthetic data, for example: univariate and multivariate distributions; correlations between features; or the ability of a machine learning classifier to discriminate between the datasets.



© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

Figure 2: Privacy/Fidelity Trade-off

On the other hand, the purpose of using synthetic data rather than the real data in many cases is to protect privacy or confidentiality. Example privacy measures might consider the 'distance' between synthetic and real data, with the important context of the relative density or sparseness of data in the region around the datapoint(s) being evaluated. Because learning-based approaches begin by learning from the real data (or, if masking approaches like differential privacy[2] are used, an altered version of the real data), there is always a need to assess the privacy of data generated this way, and we normally expect to see a Pareto-style trade-off curve like the illustration in Figure 2.

Generally, the more similar the synthetic data is to the real data, the higher its utility (assuming the goal is to create a double of the original data), but the privacy disclosure risk increases in turn, potentially to a point where utility is drastically lowered as it becomes too difficult to use or share negating the original benefit of using synthetic data. As we discuss later, the utility of the synthetic data, generated by whatever method, depends on the use to which it will be put, and the real test is the progress the user of the data can make in pursuing the outcomes they would like to achieve.

[2] Differential Privacy is an approach to adding noise—often Laplacian noise—to data in a controlled manner that allows certain statistical privacy guarantees to be given about the resulting data. When using differential privacy, the strength of the privacy guarantee is controlled by a parameter epsilon (ε). Lower values of epsilon cause more noise to be added, increasing privacy, but reducing the fidelity of the synthetic data to the original, reducing its utility. Inevitably, good privacy requires low values of epsilon (probably under 1), but this adds a lot of noise of making datasets of limited value. The original foundational paper on differential privacy is Calibrating Noise to Sensitivity in Private Data Analysis, Cynthia Dwork, Frank McSherry, Kobbi Nissim, & Adam Smith, in Proceedings of the Third conference on Theory of Cryptography (TCC'06), Springer-Verlag, Berlin, Heidelberg, 265–284. Available at https://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf,. There have been many more recent expositions.

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

How is Synthetic Data generated?

There are many approaches to synthetic data generation, and they can be classified in different ways.

At Smart Data Foundry, we focus on two primary approaches:

- **Simulation (primarily Agent-Based Modelling).** With simulation, or agent-based models, software is written that simulates key aspects of the systems that generate (or would generate) real data. Synthetic data is then captured from the simulation. Familiar examples of simulations include weather and climate simulations used for weather forecasting and climate scenario modelling; herd movement and crowd dynamics models; queuing, demand, and wait-time models; and socio-economic simulations such as the Bank of England/Office for Budget Responsibility model of the UK economy.[3] Many simulations[4] are well developed, but ones simulating people and social-economic systems are less well developed than physical simulations using well-established laws of physics.

- **Learning-based synthesis (often referred to as the creation of synthetic doubles)**. Much of the publicity and excitement around synthetic data today focuses on the use of various forms of machine learning for data synthesis. To begin, a model is trained to learn the patterns in an existing real dataset; and then the learned patterns are used to generate synthetic data with 'the same' multi-dimensional distributions as the original dataset. We will discuss this in the next section.

Simulation-based data synthesis (Agent-Based Modelling)

For physical and engineering simulations, highly accurate equations are often able to model systems with high fidelity, for example, spaceflight, models of nuclear reactors and structural models of buildings. For 'softer' systems, particularly socio-economic systems, there are no (and perhaps can never be) accurate formulae available, so we instead use more heuristic rules and adapt them over time by calibrating the outputs from systems against historical data. A good example of this is the Treasury/Office for Budget Responsibility macro model of the UK economy,[5] which tracks (at the time of writing) 627 variables.[6]

In socio-economic systems, and many other domains, simulation is often known as agent-based modelling. Here, a set of agents, with often stochastic behaviours, are simulated and can interact with each other. Data is collected from the simulation in the same way data is collected from corresponding real-world systems. Usually, there is input data that can be used to give appropriate distributions to the agents and other

[3] https://obr.uk/forecasts-in-depth/obr-macroeconomic-model/
[4] e.g. https://smartdatafoundry.com/news/leading-edge-synthetic-data-for-fca-psr-app-fraud-techsprint
[5] https://obr.uk/forecasts-in-depth/obr-macroeconomic-model/
[6] 627 variables as 20th December 2022. https://obr.uk/download/obr-macroeconomic-model-variables/

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

parameters in the system, and there is known data for the measurements taken from the real system that can be used for calibration. Agent-based simulations can be developed, adapted, and calibrated until the statistical distributions of the outputs match those from the real-world systems to an acceptable level for the task at hand.

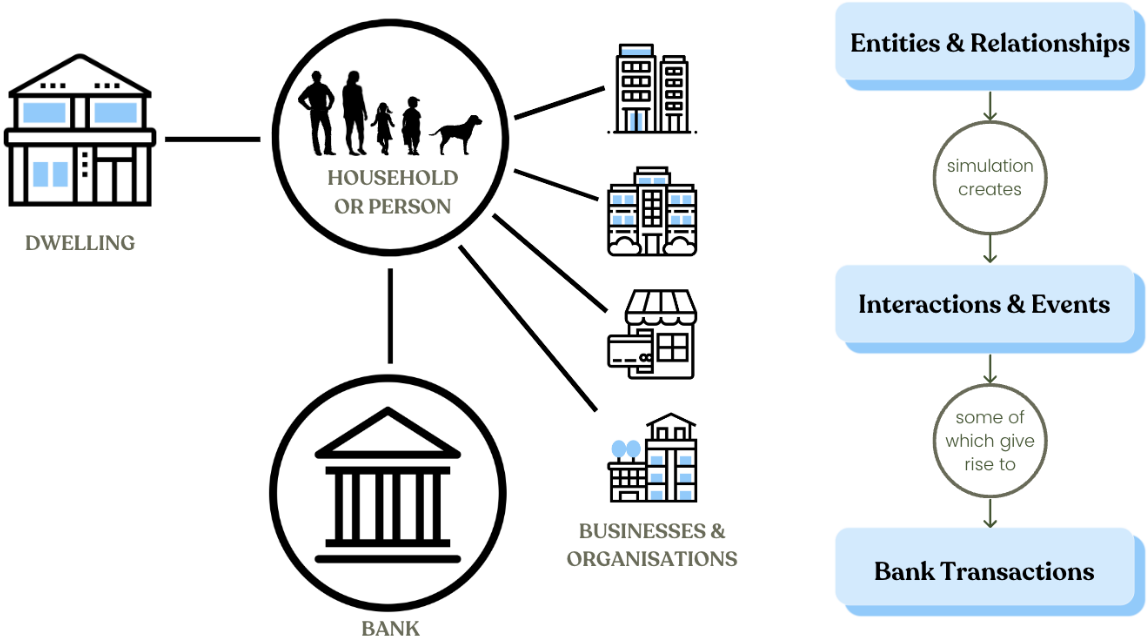


Figure 3: Agent-Based Simulation.

For example, to generate realistic looking synthetic personal banking data, as a minimum you usually need to model a person or household, with a home location, a set of businesses and other organisations they interact with, and a bank. The agent simulation models the people's behaviour over a defined period of time, resulting in a set of interactions and events. Some of those events then lead to financial transactions, many of which result in a bank account transaction. The transactions can be written out as output from the simulation. Of course, you would typically simulate not a single person or household but a large number of people, who would interact with each other and with people and organisations not simulated (i.e. outside the system).

Figure 3, above, shows a basic system for creating synthetic bank transactions focusing on a single account holder from a simulation. A more complete overview, including reference data and calibration, is illustrated in Figure 4 below.

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

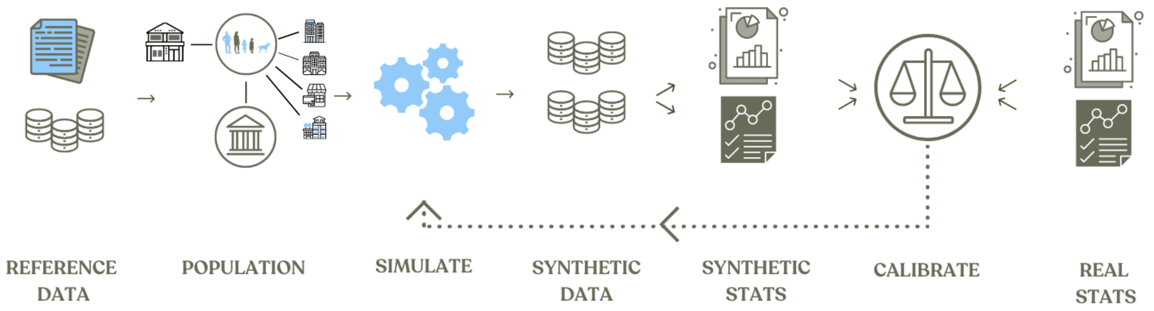


Figure 4: Agent-Based Simulation

A more complete account of agent-based simulation involves using reference data (often public data) to construct a suitably representative population of entities that would include people, organisations, relationships, etc., and then running the simulation for that population multiple times, with different parameter settings and gathering the synthetic outputs. Statistics can then be computed over these outputs. If similar statistics are available from real data (possibly public statistics), the synthetic outputs can be sense checked and/or calibrated against the real stats. This potentially leads to adjustments to the simulation to increase its accuracy. Interestingly, some data that is hard or practically impossible to collect in the real world can be generated easily from the simulation, e.g., cash transactions, giving an extra data point in the synthetic dataset which would be very difficult to record in real data.

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

Learning-based data synthesis (Synthetic Doubles)

Learning-based approaches to synthetic data generation start by feeding real data (or a modified version of real data) into a machine learning system, which tries to learn the meaningful patterns in the data, as illustrated in Figure 5 below. The model is then transformed into a data generator, in slightly different ways for each different kind of model, in such a manner as to allow its learned knowledge of the patterns in the original data to be used to generate synthetic data. The new synthetic data should exhibit 'the same' patterns but should not recreate any of the input data in full, nor pose any material privacy risk to anyone or any entity whose real data was included in the training data.

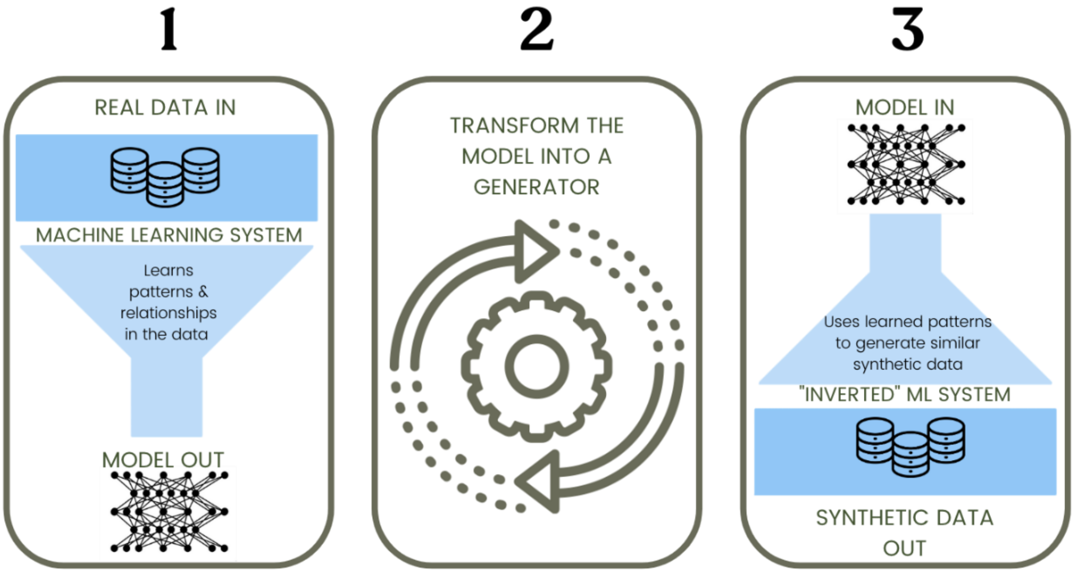

1
REAL DATA IN
MACHINE LEARNING SYSTEM
Learns patterns & relationships in the data
MODEL OUT
2
TRANSFORM THE MODEL INTO A GENERATOR
3
MODEL IN
Uses learned patterns to generate similar synthetic data
"INVERTED" ML SYSTEM
SYNTHETIC DATA OUT

Figure 5: Learning-Based Synthesis

With learning-based simulation, the input is the real data someone would ideally like to use, but cannot (usually for reasons of privacy, confidentiality, ethics, or data retention). In some cases, noise is added to the real data, often in the manner prescribed by differential privacy, before passing it into some form of machine learning system. The goal is to configure the machine learning system so that it learns all the meaningful, general patterns in the real data, while not overfitting, i.e., not learning the particularities of individual cases. The learned patterns are then used to generate new synthetic data that, ideally, is statistically indistinguishable from the real data, but does not contain any of the same records. More generally, the synthetic data should not allow anything to be learned about any individual in the original data.[7]

[7] Differential privacy formalises this by giving a statistical guarantee about the maximum difference between two differentially private datasets differing only by the presence or absence of any single individual in the original data.

Different approaches depending on the job to be done

All synthetic data is not created equal, there is no plausible one-size-fits-all approach, and different methods solve different use cases. Agent-based and learning-based synthesis have rather complementary strengths and weaknesses, are at different stages of maturity, and are best suited to different application areas.

Agent-based simulation is the more suitable approach if the goal is to explore scenarios that have not occurred, to model sensitivities and seek different ways to solve complex problems – or if real data with suitable patterns and characteristics don't exist, or are not available for learning. Here, there may be few, if any, challenges around ethics, privacy, or confidentiality, but in many cases simulation software, reference data and calibration data will need to be found or created. The challenges here are centred around simulating useful, sufficiently realistic, and diverse behaviours for the domain of interest and validating results in situations where reference data is available to build confidence that the underlying model is sound when it is extended into areas for which no data exists. Today, simulation capabilities are best tailored to specific domains on a case-by-case basis rather than using a general approach.

Where the goal is to make available a safe synthetic double of existing data to which the provider has access while protecting privacy or confidentiality, a learning-based approach is probably most suitable. Here, the required level of privacy will have to be carefully determined, and the actual privacy of the output data assessed. The key question is likely to be whether sufficiently high utility can be achieved at the target level of privacy. Off-the-shelf open-source and commercial tools may require 'only' tuning and assessment, but skilled practitioners will be needed, and the ethical, technical and operational risks inherent in creating synthetic data from real data, particularly real personal data, will need to be carefully considered and managed.

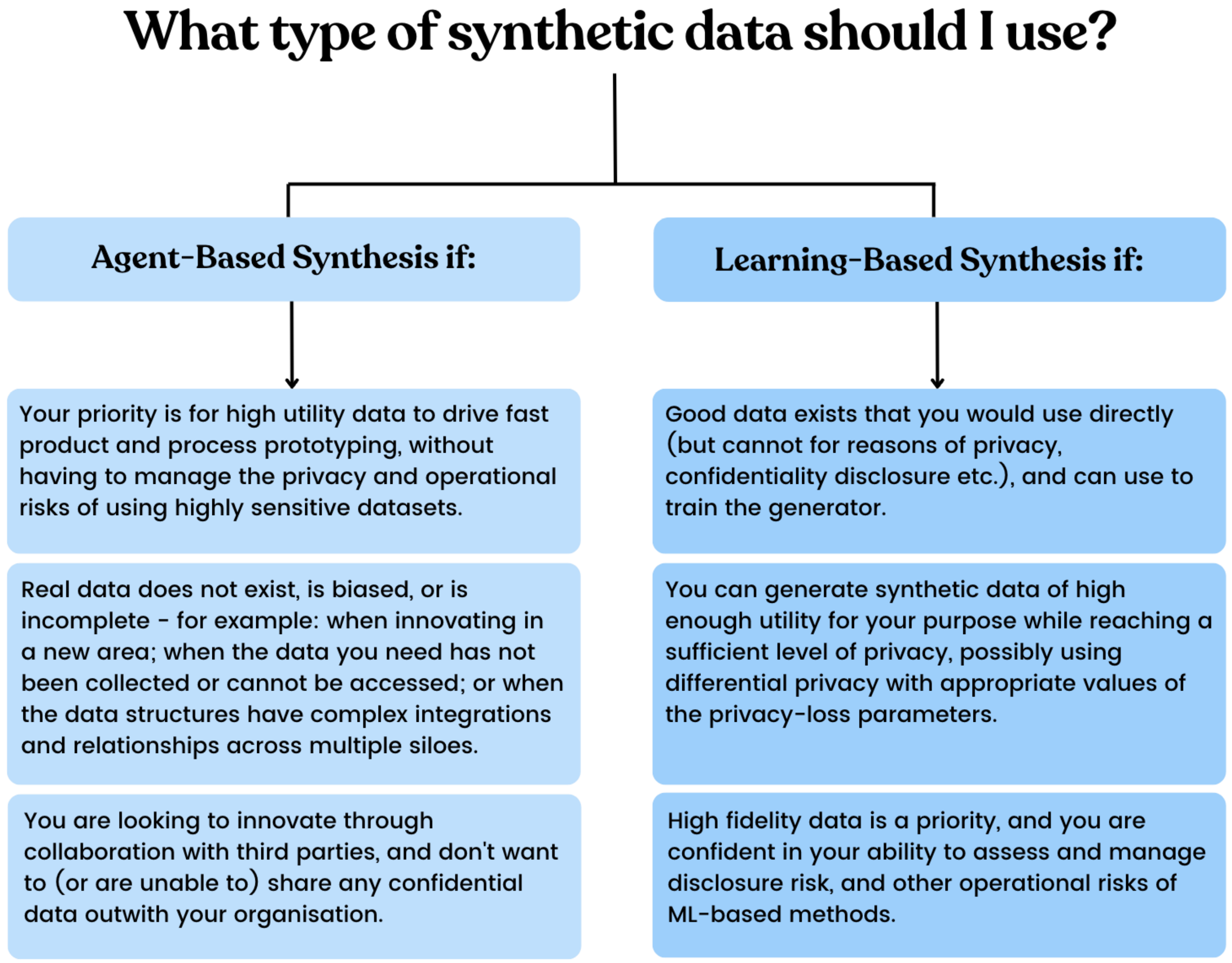
Synthetic data 'quality' in the fidelity sense can to some extent be quantitatively measured, as discussed previously – however focussing solely on this can be counter-productive, not least because privacy disclosure risk will almost certainly constrain the maximum fidelity achievable. It is crucial to consider the utility (or, more plainly, usefulness) of the synthetic data output, which is driven by the context of the use case, and how the synthesis helps the end user achieve it. Unless your goal or requirement critically depends on a very detailed double of the data, then it is likely that lower fidelity data can still have excellent utility, particularly in innovation and whitespace areas, and this is where agent-based simulation approaches excel.

Hybrid approaches are also possible. These may be appropriate in cases where some real data is available for modelling, but it has known deficiencies or sensitivities. In those cases, it may be possible to create a synthetic double of the defective data, using a learning-based data synthesis, and to augment that with simulation-based data incorporating patterns designed to counteract the deficiencies and limitations of the real data.

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

Summary

Agent-based simulation and learning-based synthetic data generation approaches are both powerful and largely complementary ways of creating synthetic data. In most cases, one will be more applicable than the other. We have tried to summarise these differences in the illustration below.



What type of synthetic data should I use?

Agent-Based Synthesis if:

Your priority is for high utility data to drive fast product and process prototyping, without having to manage the privacy and operational risks of using highly sensitive datasets.

Real data does not exist, is biased, or is incomplete - for example: when innovating in a new area; when the data you need has not been collected or cannot be accessed; or when the data structures have complex integrations and relationships across multiple siloes.

You are looking to innovate through collaboration with third parties, and don't want to (or are unable to) share any confidential data outwith your organisation.

Learning-Based Synthesis if:

Good data exists that you would use directly (but cannot for reasons of privacy, confidentiality disclosure etc.), and can use to train the generator.

You can generate synthetic data of high enough utility for your purpose while reaching a sufficient level of privacy, possibly using differential privacy with appropriate values of the privacy-loss parameters.

High fidelity data is a priority, and you are confident in your ability to assess and manage disclosure risk, and other operational risks of ML-based methods.

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

About Smart Data Foundry

Smart Data Foundry is a data innovation organisation, serving the public, private and third sectors. Our purpose is to inspire financial innovation and improve people's lives by unlocking the power of financial data. We aim to be the leading provider of data for research and innovation — providing access to real data for research and supplying synthetic data for innovation.

Smart Data Foundry was one of the first major data-themed organisations launched within the University of Edinburgh's Data-Driven Innovation (DDI) initiative, part of the City Region Deal. Under DDI, the University is creating a network of hubs to help public, private and third-sector organisations improve products and services through research and high-powered data science. Smart Data Foundry has expertise and experience in all major approaches to synthetic data.

Contact details:

Richard Seabrook
Head of Marketing, Smart Data Foundry
richard.seabrook@smartdatafoundry.com

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved

Our Experience

Some notable experiences working with the various methods include:

- Developing data for the 2021 Financial Conduct Authority (FCA) ESG Sprint[8] using a mixture of constrained randomised generation and some custom heuristic construction.

- Winning the Synthetic Data Challenge[9] run by the United Nations Economic Commission for Europe's High-Level Group for the Modernisation of Official Statistics[10] (UNECE HLG-MOS), using a range of techniques, including Generative Adversarial Networks,[11] Fully Conditional Specification,[12] Constrained Random Generation, Gaussian Copulas[13] and a modified K-Nearest Neighbours/Surprisal-based approach,[14] using Synthetic Data Vault[15] (from MIT), Synthpop[16] from University of Edinburgh, Miró[17] from Stochastic Solutions and GEMINAI[18] from Diveplane Inc.

- Developing data for the 2022 FCA Authorised Push Payment Fraud TechSprint[19] using agent-based modelling.

[8] Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot, Financial Conduct Authority, 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

[9] HLG-MOS Synthetic Data Test-Drive, United Nations Economic Council for Europe, January 2022. https://pages.nist.gov/HLG-MOS_Synthetic_Data_Test_Drive

[10] Modernization of Official Statistics https://unece.org/statistics/modernization-official-statistics

[11] Modelling Tabular Data using Conditional GAN, Lei Xu, Maria Skoularidou, Alfredo Cuesta, Infante, Kalyan Veeramachaneni. arXiv:1907.00503v2 [cs.LG] 28 Oct 2019, https://arxiv.org/pdf/1907.00503.pdf

[12] synthpop: Bespoke Creation of Synthetic Data in R. Nowok, B., Raab, G. M., & Dibben, C. (2016). Journal of Statistical Software, 74(11), 1–26. https://doi.org/10.18637/jss.v074.i11

[13] GaussianCopula Model. https://sdv.dev/SDV/user_guides/single_table/gaussian_copula.html

[14] Natively Interpretable Machine Learning and Artificial Intelligence: Preliminary Results and Future Directions. C. J. Hazard, C. Fusting, M. Resnick, M. Auerbach, M. Meehan, V. Korobov. arXiv:1901.00246 [cs.LG] January 2019. https://arxiv.org/abs/1901.00246

[15] SDF: The Synthetic Data Vault, https://sdv.dev/

[16] synthpop: R package for generating synthetic versions of sensitive microdata for statistical disclosure control. https://synthpop.org.uk/.

[17] Miró: https://stochasticsolutions.com/miro/

[18] DIVEPLANE GEMINAI™. https://diveplane.com/geminai/

[19] Authorised Push Payment Fraud TechSprint. 27-29 September 2022. https://www.fca.org.uk/events/authorised-push-payment-fraud-techsprint

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

© SMART DATA FOUNDRY, 2023 Commercial in Confidence. All Rights Reserved