Paper AS-18

Synthetic Data: A Process Paradigm

Sundaresh Sankaran, SAS Institute, Cary, NC, USA
Sherrine Eid, SAS Institute, Cary, NC, USA

ABSTRACT

Organisations face challenges to analytics innovation when confronted with data that's either sensitive, restricted, imbalanced in proportion, or insufficient in volume. In short, a lot can go wrong with data! Synthetic data generation has been acknowledged to be a useful & effective tool to help mitigate these challenges. It provides opportunities to capitalise on trends and patterns suggested by real data, while not compromising privacy and mitigating risk of leakage of private and sensitive information. However, for synthetic data to truly realise value for life sciences organisations, it's important to view the concept of ***synthetic data generation*** at two levels – one, as an ***operation*** which provides usable, high-quality data, and the other is as a robust and comprehensive ***process***, in which the operation of synthetic data generation plays a vital part. In this article, we focus on the ***process paradigm*** through which we should view synthetic data generation.

INTRODUCTION

To begin with, let's note that the world is not lacking in data. Forbes estimates that 2.5 quintillion bytes of data are generated every day[1]. When decision makers at organisations lament about the lack of data, most of the time, it's not volumes they are concerned with, but lack of the *right* data.

To understand what right data connotes, let's look at what real data is. Competitive forces, a focus on customer-centric outcomes and rapid innovation in analytics has fuelled the demand for variety and larger volumes of data. Real data, however, poses challenges, which can be summarised as **Access, Sufficiency, or Quality-related challenges**. These challenges make it hard to train robust and statistically significant models and take confident decisions. Importantly, lack of the right (adequate and appropriate) data **hinders innovation & research.**

ENTER, SYNTHETIC DATA

So, enter Synthetic Data Generation. Synthetic data generation refers to a variety of techniques that generate data elements very similar in characteristics to real data. Generated elements contain close resemblance to original data and can be customized to provide more volume to a given dataset, bolster a minority segment, or provide data for analysis as a proxy to a dataset that was hitherto restricted.

Here are some examples of challenges that can be mitigated by synthetic data.

What's the challenge?	How does Synthetic Data help?
Real data is private, sensitive, and restricted. Internal organisational data security policies and regulations prohibit direct usage of data for analytics	Provides a proxy dataset which retains characteristics of original data distributions. This data does not retain identifiable or sensitive information and isn't a direct subset of real data either, therefore can be released for analytical purposes. Proxy dataset, after quality checks, is shared for downstream use.

[1] https://www.forbes.com/sites/forbestechcouncil/2021/08/02/understanding-generation-data/?sh=65f15ca036b7

Real data is imbalanced with the patterns exhibited by 'minority segments' getting dwarfed by segments containing more observations.	Augments data points and observations for specified levels of categorical (segment) variables.
Real data is infrequently collected and observed and therefore scant in distribution.	Augments data observations, which smoothens data distribution and surfaces more observed values.
Poor quality of real data renders analytic results inoperable or inefficient.	Compensates for the shortfall in observations of acceptable data quality by generating additional observations.
Demand for synthetic data is fuelled by increasingly complex analytical methods such as machine learning and deep learning, which tend to be data-hungry in nature.	Provides adequate volume and sufficient data per segments to allow deep learning methods to be effective.

Once you acknowledge that synthetic data has its merits, a question arises as to how we define a rigorous and effective approach to generating synthetic data. A common pitfall is to focus solely upon the core operation of synthetic data generation, which may be accomplished through a variety of algorithms. Relying on algorithms alone may yield suboptimal results, due to the working of the famous adage – *Garbage In, Garbage Out*. Further, it's also important to understand and closely align with the end application intended for synthetic data, because it's unlikely that synthetic data is generated purely for its own sake. The end objective should dictate the method and not vice versa.

Thus, we realise that synthetic data generation has certain upstream and downstream considerations.

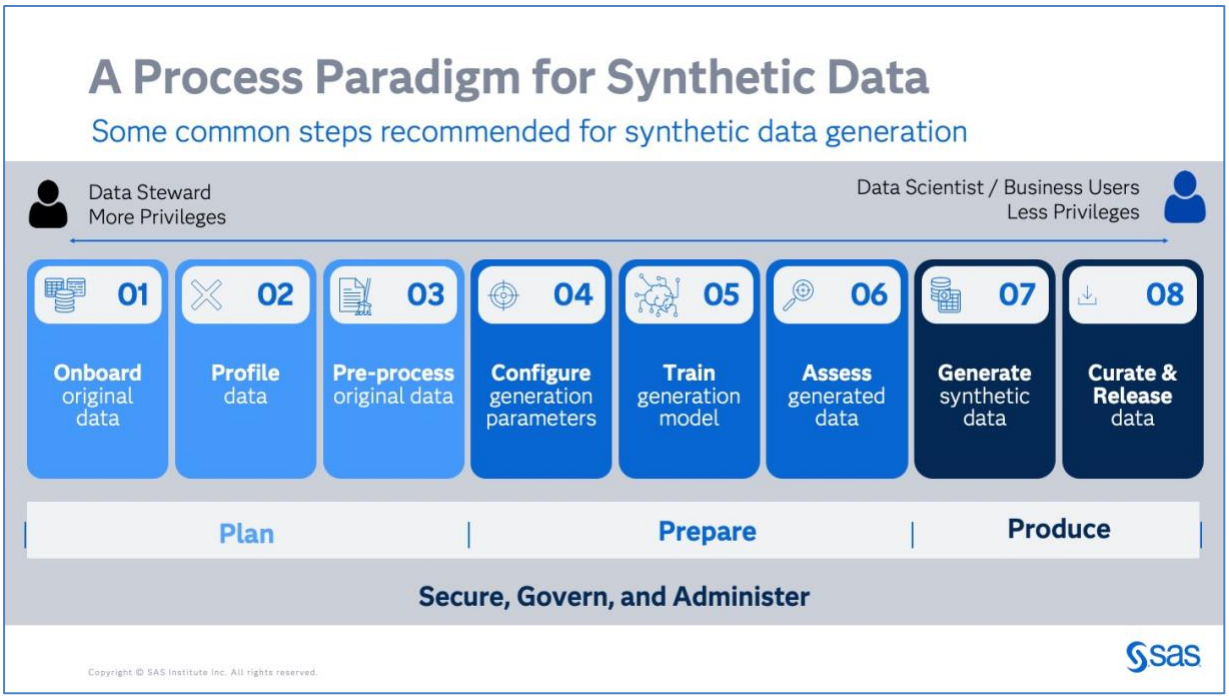


Figure 1: The Process Paradigm for Synthetic Data Generation

Let's take a deeper look at the core aspects of the process paradigm.

DATA PROFILING AND ANALYSIS

A first step in any data-centric process is to "know your data". This can be understood at different levels, including **structure, content and sensitivity**. Understanding the data structure is important because it helps informs the generation in terms of defining the grain at which data needs to be generated. It also informs the choice of generation technique down the line. For example, generating data for a snapshot dataset (patient current conditions and status) is different from a transactional dataset (or even a dataset containing time series), for example, the visits of a patient over the course of treatment. This is due to the way observations and measures are constructed.

Another important consideration is to detect **classes which may contribute to bias** in original data. The bias could arise due to inherent structural reasons, such as a majority group (e.g. male patients of a dominant race in a geography) crowding out sub-groups and smaller segments. The intention is to identify the possible areas where synthetic data will also reflect the bias from original data and, to mitigate this risk, take steps to amplify the minority groups.

Finally, we also mention the need for **profiling of data content**, which helps surface quality issues as well as candidate sensitive data elements. This stage would address data cleansing and standardization activities such as imputation to replace missing values for numeric and categorical columns. There is also scope for transformation of numeric variables with low cardinality that could be changed into categorical variables. For better efficiency, such imputations and curations can also be carried out upstream in the database or data warehouse which hosts real data.

Regardless of tool, the important consideration is to be able to easily understand the nature of data considered for synthetic data generation so that you can form a benchmark regarding what to expect after generating synthetic data. Tools like SAS's Information Catalog help fulfil this purpose by providing such profile.

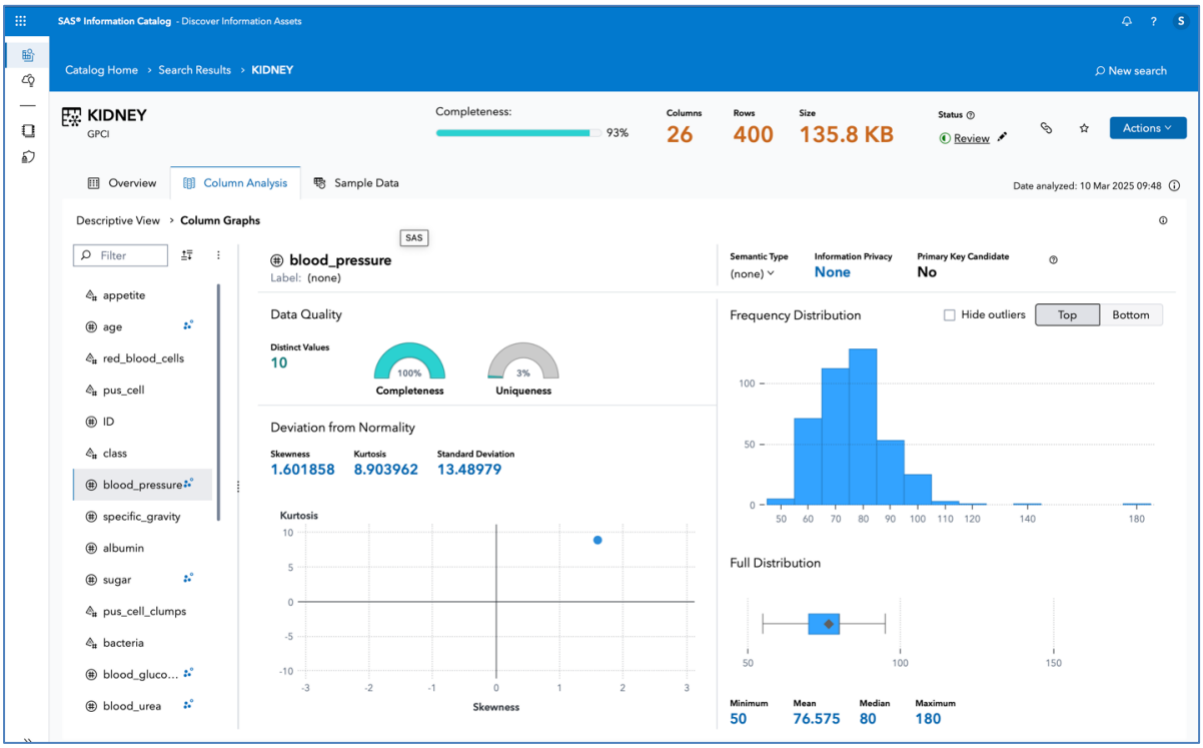


Figure 2: An example of profiling various columns in a candidate dataset through SAS Information Catalog

DATA SELECTION AND TREATMENT

Enterprises, Regulators and users are strongly inclined to protect personal data/privacy. This is driven by existing policies, practices and regulation around the ownership, security and access privileges concerned with data elements. Prior to generating synthetic data, it is important to ensure that identified source data has gone through a process where sensitive confidential information has been tagged and undergone a process of deidentification. This is keeping the extremely sensitive nature of certain regulated industries – example, life sciences, clinical trials, rare diseases, adverse events of sensitive nature, healthcare, financial services - in mind, and to make sure that any private information found in such data does not 'leak' or extend to the dataset used for synthetic data generation. Consequences of not paying attention to this aspect are possible data breaches and a backdoor loss of privacy which may be exposed in downstream synthetic data.

Privacy protection mechanisms are many and range from simpler methods such as "masking" or "anonymisation", to more sophisticated methods such as differential privacy, or k-anonymisation. Of all these methods, differential privacy has been found to be more robust, with masking and anonymisation considered as relatively 'weaker' techniques. Masking and anonymisation of data only change the labels of identifier information in the data, but does not change the data observation's indicator values, thus allowing a sufficiently motivated (bad) actor to be able to reverse engineer back to the original information. Relatively stronger techniques such as differential privacy introduce some noise within the original data in such a way that the original statistical pattern or trend is retained, but the data itself is rendered sufficiently different. Note also, that the choice of privacy protection technique determines the stage at which it's applied. Masking and anonymisation are applied right on the data, while differential privacy is at an algorithmic level.

To mitigate the risk of data leakage, knowledge about potential sensitive elements and inherent risk in the dataset is useful. Metadata glossaries and detection algorithms help in identifying such candidate variables for screening prior to generation.

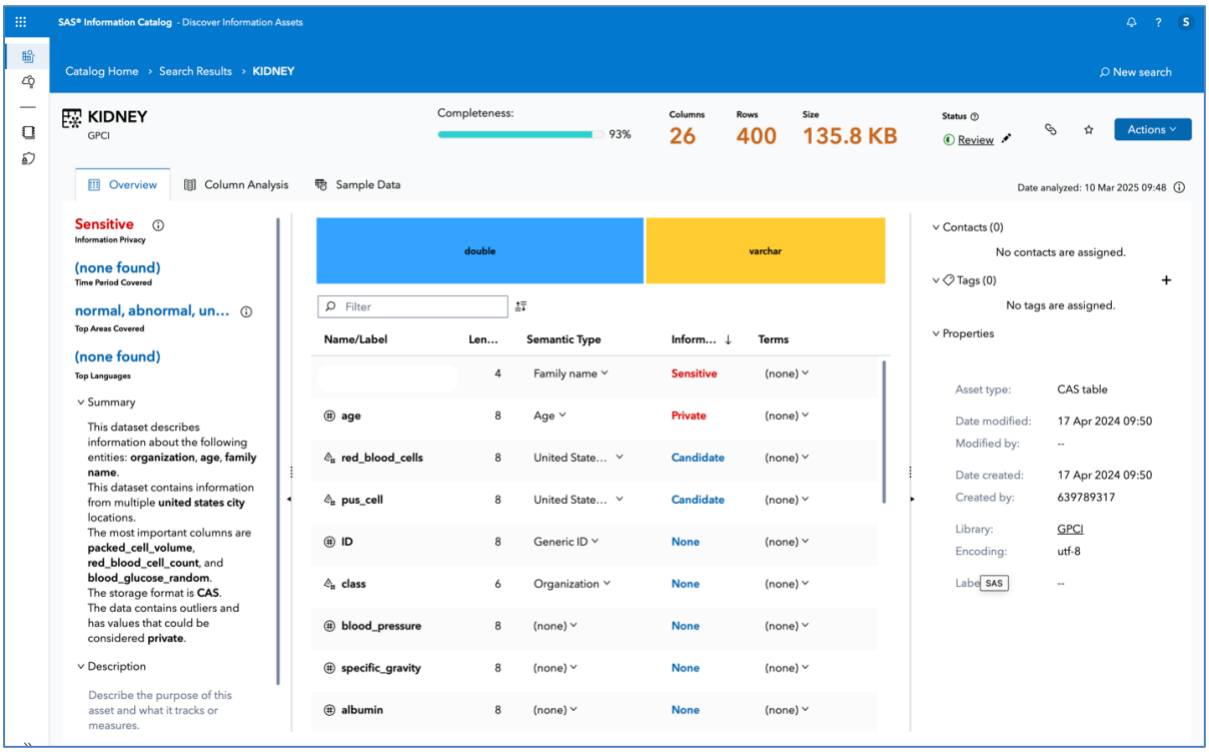


Figure 3: A screenshot of SAS Information Catalog showing automatically detected sensitive variables which form part of a original data table slated for synthetic data generation.

SYNTHETIC DATA METHOD

A variety of methods exist for the core operation of generating synthetic data, each presenting their own pros and cons. We focus on a couple which have been established in practice for some time and are sufficiently different in their approach to be illustrative of the many approaches followed.

Generative Adversarial Networks: A machine learning-based technique which uses a special form of architecture involving two deep learning networks: a generator network, which creates synthetic data candidates which were learnt from real data as part of a training exercise, and a discriminator network, whose job is to assess and identify if the generated data has high probability of getting identified as 'fake' data. The objective would be to generate data which is able to convince the discriminator network that it does in fact emerge from real data.

Synthetic Minority Oversampling TEchnique (SMOTE): A statistical, distance-based technique which identifies candidate synthetic data points that lie between close (nearest) neighbours of real data observations. Based on the provision of a parameter k, the algorithm identifies k-nearest neighbours to original data points spread across a spectrum of observations and then interpolates candidate synthetic data points that lie between the two.

The choice of algorithms to use depends on a range of factors such as the complexity of the source datasets, speed-accuracy trade-offs, and available computing resources to explore these algorithms. Irrespective of the algorithm chosen, the generated data should undergo a process of assessment as described in the next section.

Prior to using the algorithms above, it's recommended to implement techniques that make sure latent trends in the original dataset are captured in terms of their joint probability distribution. Information about the cardinality of variables is also captured and the data goes through a process of clustering, so that appropriate centroid points act as the "representative" observation which is used by the synthetic data generator as an input. Other considerations and decisions that need to be made are whether transformations such as normalization and encoding are warranted, given the advantages they offer in terms of standardization and the problems they pose in terms of loss of interpretability.

GENERATE, CURATE AND ASSESS QUALITY

Synthetic Data Generation algorithms do not use a validation set since the focus is to generate data rather than predict an outcome. A key question to answer is regarding **how well does generated data resemble the original**? A complementary question would be to establish that generated data **only looks like but does not bear exact attributes** of an original record. And finally, it is also important to establish if the synthetically generated data **makes logical sense**.

To answer the above questions, a common assessment approach involves visualising different assessment metrics which help the data steward get a sense of how similar and obfuscated (from a privacy protection perspective) the synthetic observations are. In addition, we also recommend a semi-supervised approach with a human-in-the-loop element, where the generated data would be exposed to a team of clinicians who provide their expertise in curating the data and act as stewards to establish that the data is 'fit for use'. Note that 'fit for use' is a subjective term which depends on the end objective to which synthetic data is applied. A typical post-processing decision system performs multiple checks (through rules and model thresholds) which are applied to the data before it is 'released' for downstream usage.

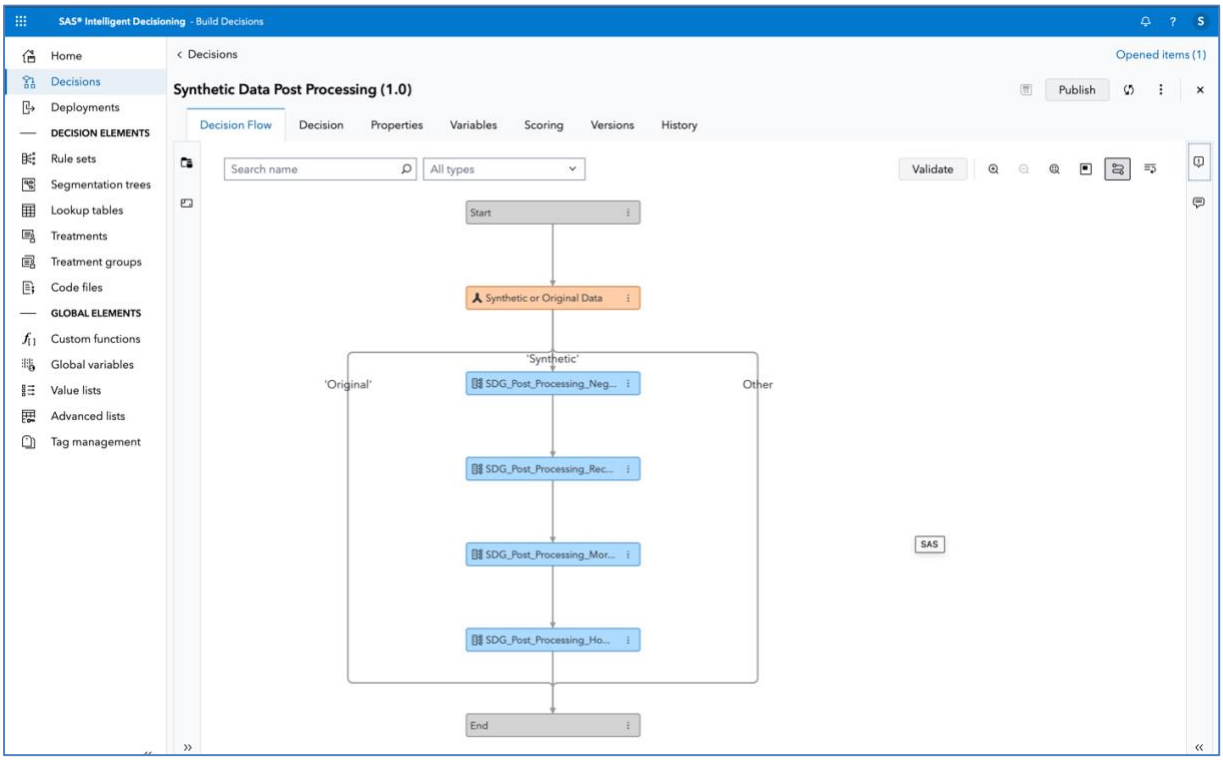
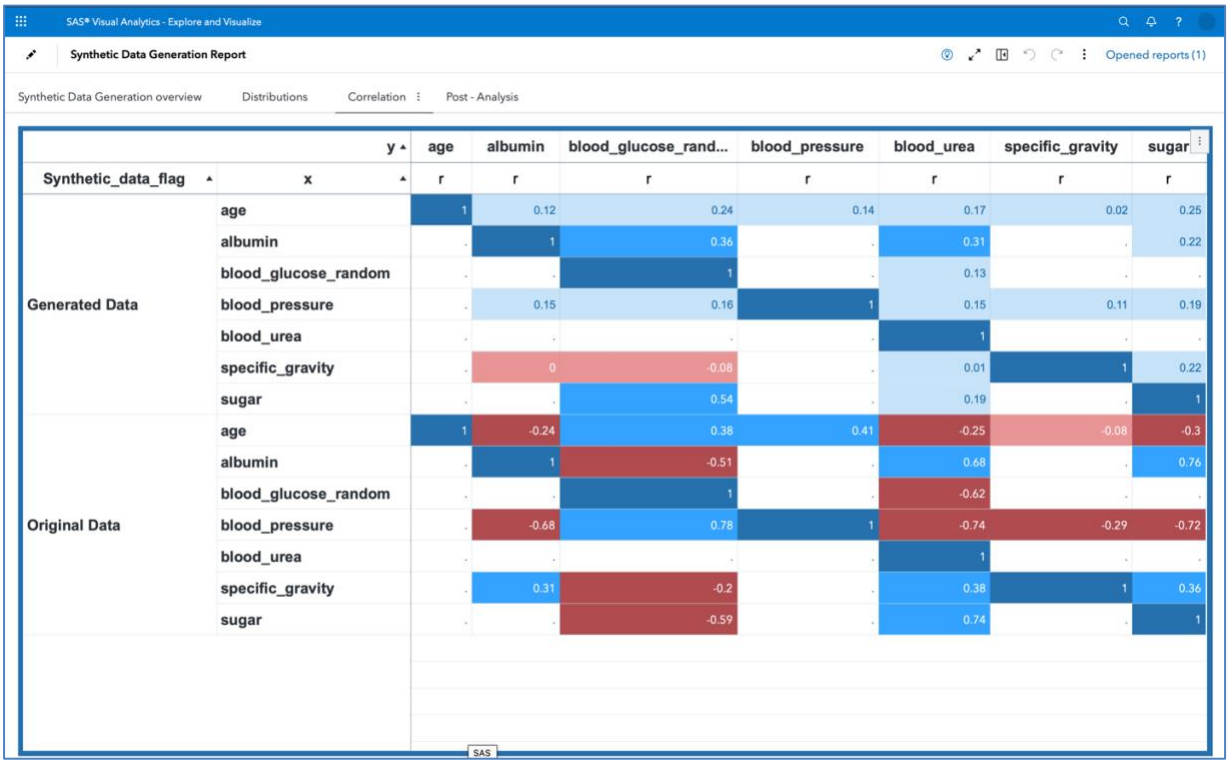


Figure 4: A decision flow applying multiple rules for identifying and filtering out infeasible operations prior to release of synthetic data.

To enable assessment of results, reporting templates typically assess distributions of variables for both original and synthetic data and identifies points of difference and variation. Correlation among variables in synthetic and original data is also measured. This correlation is not confined to singular distributions alone but also takes interaction effects into consideration.



	y	age	albumin	blood_glucose_rand...	blood_pressure	blood_urea	specific_gravity	sugar
Synthetic_data_flag	x	r	r	r	r	r	r	r
Generated Data	age	1	0.12	0.24	0.14	0.17	0.02	0.25
	albumin	.	1	0.36	.	0.31	.	0.22
	blood_glucose_random	.	.	1	.	0.13	.	.
	blood_pressure	.	0.15	0.16	1	0.15	0.11	0.19
	blood_urea	.	.	.	.	1	.	.
	specific_gravity	.	0	-0.08	.	0.01	1	0.22
	sugar	.	.	0.54	.	0.19	.	1
Original Data	age	1	-0.24	0.38	0.41	-0.25	-0.08	-0.3
	albumin	.	1	-0.51	.	0.68	.	0.76
	blood_glucose_random	.	.	1	.	-0.62	.	.
	blood_pressure	.	-0.68	0.78	1	-0.74	-0.29	-0.72
	blood_urea	.	.	.	.	1	.	.
	specific_gravity	.	0.31	-0.2	.	0.38	1	0.36
	sugar	.	.	-0.59	.	0.74	.	1

Figure 5: An example correlation diagram showing cross-variable correlations for original and synthetic data

Pertinent decisions taken at this stage include whether the generated data is of the right sample size, the strategy to follow in terms of whether to carry out analysis purely on synthetic data alone or to join it with original data and assess whether augmenting with synthetic data either enhances or negatively affects downstream analysis and prediction results. The role of curators (clinicians and statisticians) is important here as they may like to impose limits on values which go beyond the realm of possibilities in generated data.

A simplistic example, provided as an illustration, would be traits that show for a patient of a particular gender, which are usually found in other genders (e.g. a male with a pregnancy indicator). Or indicators (e.g. creatinine levels) may follow limits which are more likely in deceased patients rather than live patients. Cases like this need to be tackled at multiple levels to ensure that they are mitigated. First, the algorithm is flexible and can be tuned to learn specific patterns from correlation in the original data (and reduce possibilities for unrealistic results). It is iterative and parameters can be adjusted to ensure that accuracy is maintained. Finally, at the curation level, clinicians have facility to execute decisions and rules which impose realistic limits and conditions.

A critical element to consider here is to provide trustworthy AI solutions. This requires that the analytical output (generated synthetic data) does not reflect the bias latent in the original data, and even if it exists, controls are available to mitigate the adverse events due to the same. For this purpose, methods for bias mitigation may be applied. Note that bias in data science can occur at various stages starting from data collection all the way till deployment, some of which are controllable, and some others which are not. Similarly, synthetic data can reflect inherent bias in original data but also amplify bias further because of synthetic data generation. Machine Learning, applied downstream, helps measure and identify any marked change in bias for these two datasets separately, providing an additional dimension to assess bias risk. Data scientists should harness model bias assessment and mitigation functionality from machine learning packages, which identifies how different categories within a dataset may provide predictions reflecting bias (either negative or positive) towards these segments. The bias is presented as prediction bias (indicating difference in

average prediction per identified segment) and performance bias which measures prediction accuracy differences across identified segments.

SECURE, GOVERN, AND ADMINISTER

The output of the training phase is a **model (also called a generator)**, which is a set of instructions in machine language optimized to carry out generation efficiently. Note that for some techniques that fall on the 'simpler' end of the spectrum, the model could very well be implementation code (i.e. instructions). The model represents a machine learning asset which needs a residence (i.e. accessible from an endpoint, preferably through REST APIs) and required metadata for informational and versioning purposes. The metadata can also contain assessment statistics that permit comparison of different generator models, which can be reviewed by a model steward to decide the generator model that will be used for a given dataset and situation.

As part of enterprise data governance processes which release data assets to stakeholders with the right access controls, generators for synthetic data should be part of a repository and store of models and be subject to governance and versioning. The intensity and practices followed by such governance mechanisms depend upon the policies followed per organisation but should be treated with the same seriousness as other analytical artifacts generated within an enterprise.

The model repository also facilitates a process of continuously monitoring and assessing synthetic data generation models. Upon decisions to retrain or replace a designated generator model, project metadata and versions can be appropriately updated. Project structures for synthetic data generation are therefore crucial as they provide structure and easy management.

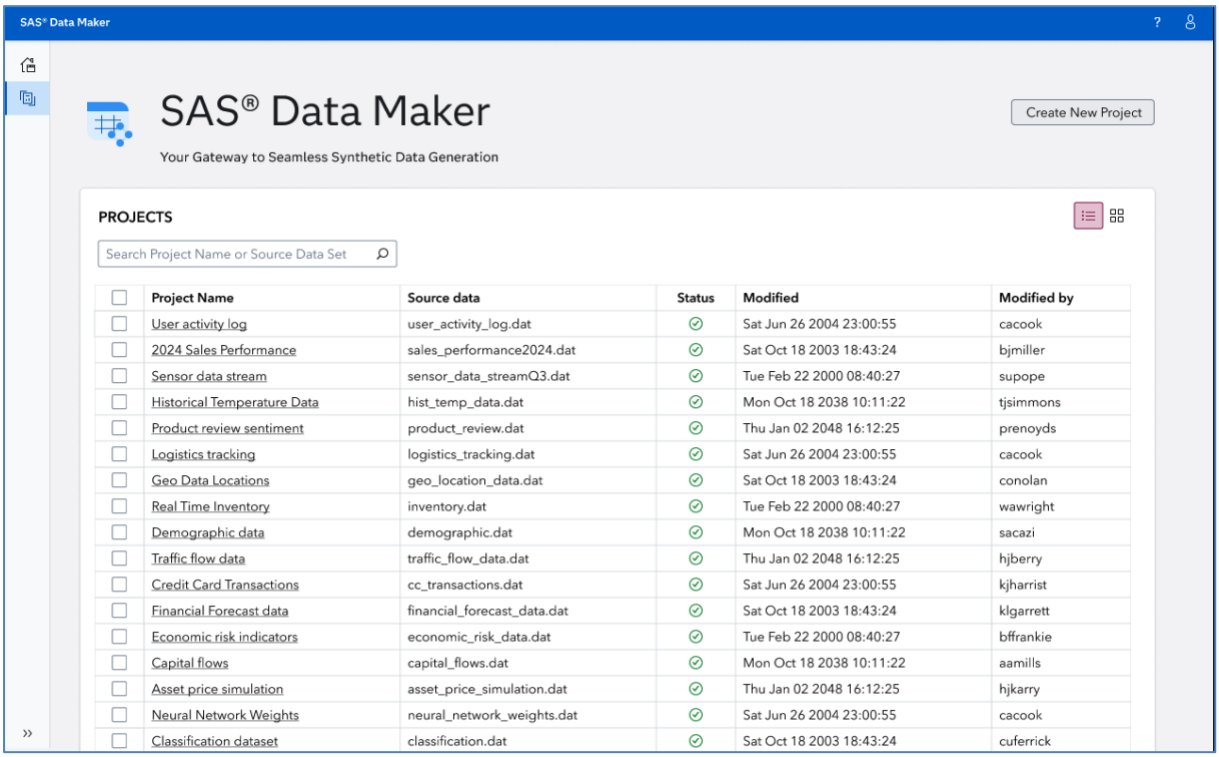


Figure 6: An example of a project-oriented synthetic data generation, SAS Data Maker. (SAS Data Maker will.be made Generally Available at a future date.)

CONCLUSION

Synthetic data is an effective tool which can augment data for research and innovation while honouring privacy requirements for various life sciences organisations, many of whom handle valuable, sensitive data, sometimes with identifiable characteristics beyond just patient labels. But, just like any analytical approach, synthetic data is not a one-click activity focussing on the operation of generation alone – upstream and downstream considerations for security, privacy, distribution and governance are prominent. Hence, the importance of a robust framework and a well-designed process which helps teams generate synthetic data in a governed manner, assisted by user-friendly low-code components which do not require programming knowledge.

Also, user personas are an important consideration, since synthetic data generation involves interaction with real data in most cases, and real (original) data requires careful policies regarding handling and use. Identifying and clarifying roles (preferably through role-based access controls in software platforms) helps ensure that the workflow is handled by the right persona at the right stage. Further, implementing robust governance mechanisms also help mitigate privacy and data quality risks.

Thinking in a process paradigm ensures that application of synthetic data techniques is supported by strong data management, exploratory and human-in-the-loop methods. Doing so will help ensure generated synthetic data is useful and relevant in enabling better patient outcomes.

REFERENCES

1. SAS Documentation on Generative Adversarial Networks, https://go.documentation.sas.com/doc/en/pgmsascdc/v_060/casml/casml_tabulargan_overview.htm
2. SAS Documentation on SMOTE (Synthetic Minority Oversampling TEchnique): https://go.documentation.sas.com/doc/en/pgmsascdc/default/casactml/casactml_smote_details01.htm
3. Generating Synthetic Data using Generative Adversarial Networks, Jaron Cones, SAS Communities, https://communities.sas.com/t5/SAS-Communities-Library/Generating-Synthetic-Data-Using-Generative-Adversarial-Networks/ta-p/903702

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Sundaresh Sankaran
Company: SAS Institute
Address: 100 SAS Campus Dr Cary NC 27513
Work Phone: +1 919 531 4921
Email: Sundaresh.sankaran@sas.com
Website: www.linkedin.com/in/sundareshsankaran

Author Name: Sherrine Eid
Company: SAS Institute
Address: 100 SAS Campus Dr Cary NC 27513
Work Phone: +1 919 531 3991
Email: Sherrine.Eid@sas.com
Website: https://www.linkedin.com/in/sherrineeid/