TRUATA.

Not All Synthetic Data Is Created Equal

The privacy risk contained within a synthetic dataset can be objectively quantified so that more informed decisions may be made.

One of the biggest impediments to innovation in commercial and academic spheres is access to data.

Building novel algorithms and technologies requires access to realistic data, so that business logic, predictive model accuracy, and algorithm performance can be tested and validated.

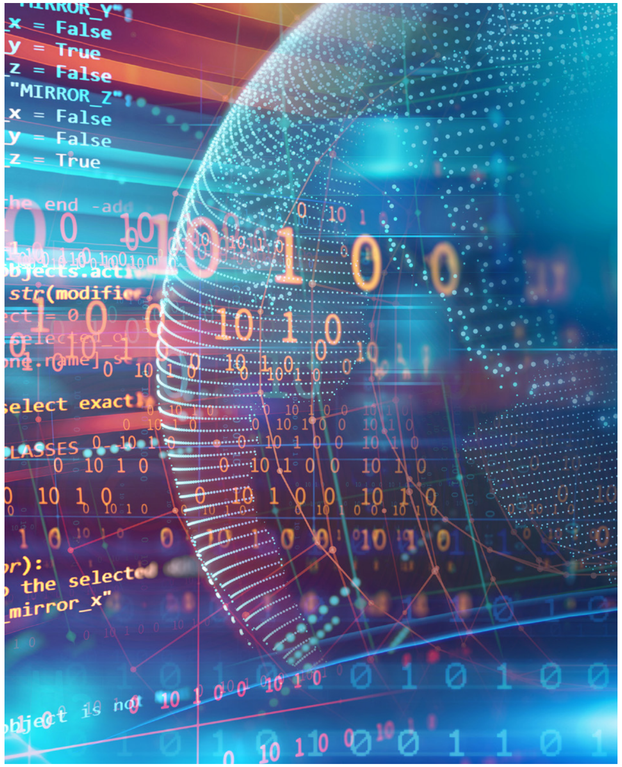
However, the original data is typically highly confidential and may contain personal data such that it also represents a compliance risk to repurpose the data for testing purposes. Even transferring this data to secure development environments within an organisation increases the company's risk profile.

Michael Fenton, Imran Khan, Maurice Coyle, Aoife Sexton | Truata

Recently, methods for generating synthetic data that resemble the original data have received a lot of attention and investment, often with claims that the synthesised dataset does not contain personal data. These techniques vary in terms of how they balance the analytic similarity and privacy characteristics of the synthetic data.

As we will describe in this paper, maximum privacy and utility cannot be achieved simultaneously; a trade-off must always be made.

Data synthesis techniques that claim to preserve general analytic equivalence while simultaneously making re-identification of individuals impossible are highly unlikely to achieve this. They are more likely to contain considerable privacy risk by underestimating the risk of re-identification.



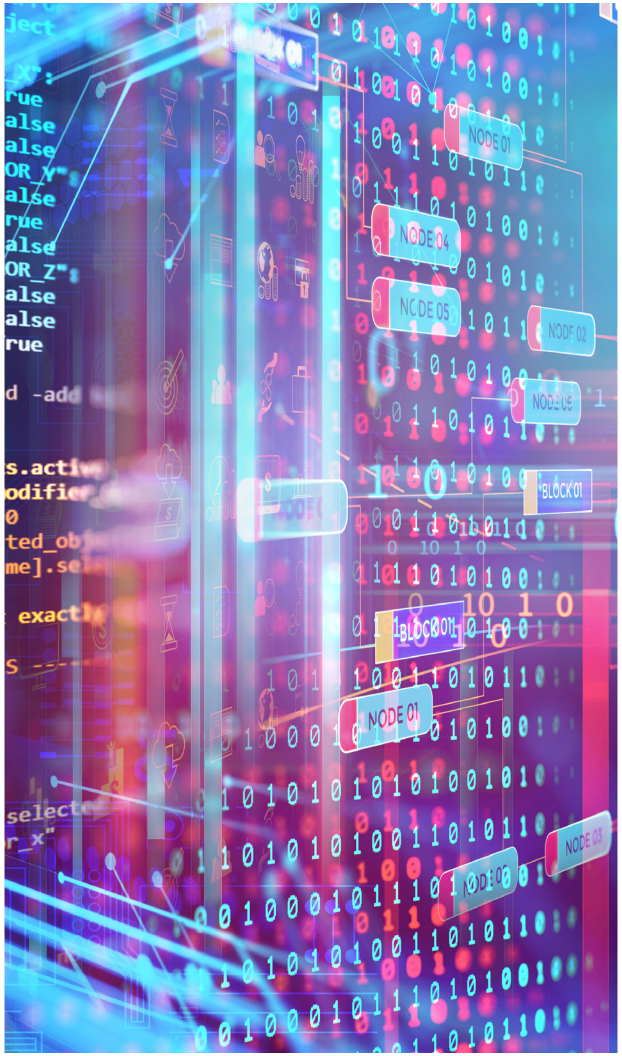
In this paper, we will define synthetic data and describe a number of different data synthesis techniques. We will describe how privacy and analytic utility are conflicting goals, limiting the use cases to which a given synthetic dataset may be applied.

We will detail how the privacy risk contained within a synthetic dataset can be objectively quantified so that better, more informed decisions may be made, leading to increased confidence in the appropriate use of synthetic data.

What is synthetic data?

There has been a lot of media attention and investments in techniques for generating synthetic data and some bold claims have been made regarding its usefulness.

Some claim that synthetic data can be generated that has identical statistical or general analytical performance to the original, while making it impossible to re-identify an individual. We will outline why we believe this claim is difficult to substantiate and potentially could be misleading, since it typically ignores hidden "fingerprints" or quasi-identifiers within the data which may be often reproduced in a synthesized dataset. Others claim that synthetic data is inherently safe since it is generated to reflect the original dataset instead of being a transformed version of the original.



In many instances, claims are being made that synthetic data, even where it preserves statistical properties of the original data, is anonymous data and as such, the GDPR won't apply, as the data is no longer considered to be personal data. Achieving GDPR grade anonymization is very difficult as you have to be able to demonstrate irreversibility and that a person is not identifiable from the synthetic data.

This means you have to show how you have addressed the risk of singling out, linkability, and inference, and how you have taken account of all means reasonably likely to be used to identify someone.

Failure to do this means that the data will still be considered personal data and subject to the GDPR. We believe that objective quantification using statistical analyses of the dataset should be the norm for demonstrating that the risk of re-identification has been reduced to an insignificant level.

Is synthetic data "real" or "fake"?

We feel that this is the wrong question to ask, in the context of data protection at least. That is, the answer doesn't tell us anything about the re-identification risk a synthetic dataset may contain. It's like asking whether a photograph is the same as the actual subject. It clearly isn't, but from a privacy point of view, you can determine sensitive information about the subject by examining a photograph.

The same is true of synthetic data, i.e. depending on how the data model was created and how values are generated, the resulting dataset may contain the ability to re-identify individuals as we will discuss below. The simple fact is that synthetic data may contain re-identification risk and should be considered as personal data unless objective proof can be provided that it contains no re-identification risk. It cannot be considered as inherently safe in all cases simply because it is "fake".

The privacy / utility trade-off in Privacy-Enhancing Technologies

Where privacy-enhancing technologies (PETs) are concerned, privacy and utility are generally opposing objectives and synthetic data is no different. That is, if some processing or transformation of a dataset optimises privacy, its analytic utility will be affected and vice versa. This makes intuitive sense, since most privacy-enhancing technologies operate by adding noise, changing values to be less distinguishable or otherwise altering the dataset's contents so as to be unrecognisable to the original.

Figure 1 depicts this trade-off, where at the privacy extreme of the graph (top left), the data exhibits minimum utility but maximum privacy. In effect, this is complete deletion of the data.

There is zero utility as there is zero data, but on the other hand, it is the most privacy-preserving data possible as there is no risk whatsoever if there are no data.

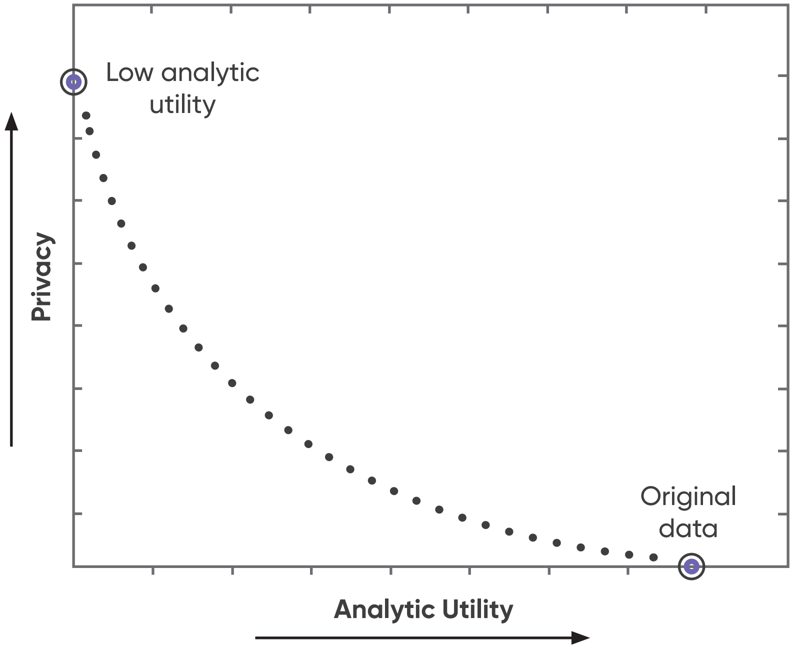


Figure 1

At the utility extreme of the graph (bottom right), the data exhibits minimum privacy but maximum utility. In effect, this is complete and unmodified original data. If nothing at all has been done to the data, then there is no reduction in utility at all. However, the data contains maximum privacy risk as there has been no action to mitigate that risk.

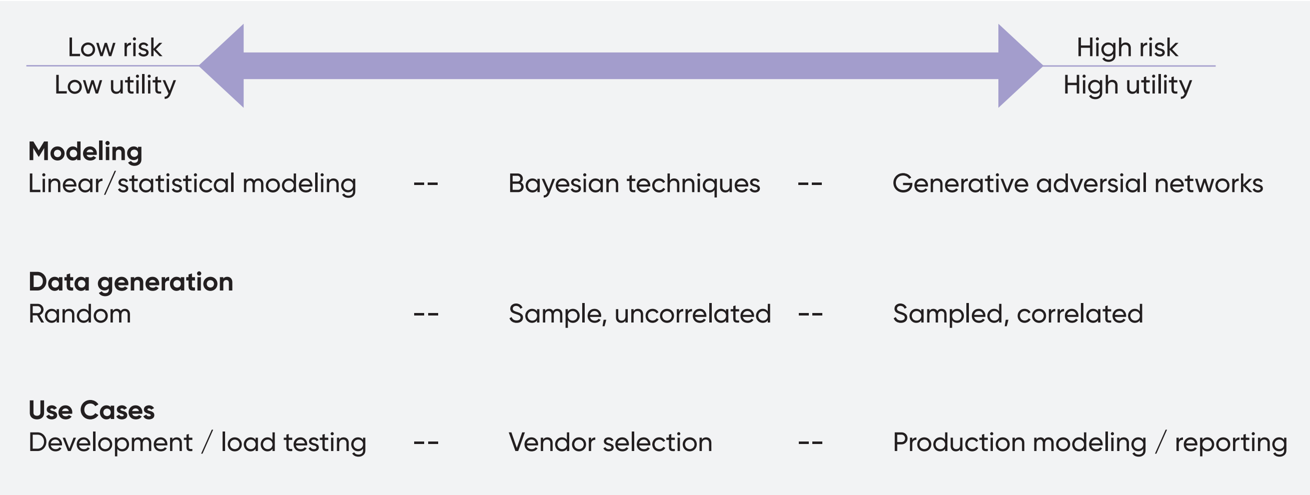
According to some positioning of products we have observed on the market today, there are claims that a synthetic version of any given dataset will allow you to travel directly upwards on this graph, essentially retaining all of the original utility, but increasing the privacy level. In reality, synthetic datasets (like all PETs) will merely allow you to travel along this curve to greater or lesser extents, sacrificing analytical similarity as privacy protection increases.

It is concerning that there are claims on the market that the generated data retains all analytical utility while making it impossible to re-identify an individual, or claiming that synthetic data is "fake" and therefore inherently safe to use for any purpose. This is very misleading.

Synthetic data types and uses

There are many techniques for producing synthetic versions of a dataset available and more sophisticated approaches are being developed all the time.

Rather than providing an exhaustive review of the literature and landscape, here we will describe the broad categories of techniques for creating a model of the data to be synthesized and for generating values that exist, discuss how they balance analytical utility and privacy and are suitable for particular categories of use case.



Low risk Low utility				High risk High utility
Modeling Linear/statistical modeling	--	Bayesian techniques	--	Generative adversial networks
Data generation Random	--	Sample, uncorrelated	--	Sampled, correlated
Use Cases Development / load testing	--	Vendor selection	--	Production modeling / reporting

Figure 2: Choice of modelling and generation technique have varying privacy risk and utility characteristics and are suitable for different categories of use case

Data Modelling and Generation

Synthetic data is data that is generated based on information that has either been derived from an existing dataset or created to serve a particular purpose. To generate synthetic data, there must be some information about the dataset that is to be synthesized, i.e. a “model” of what that data should look like. This can be based on the characteristics of what a desired dataset should look like or the characteristics of an actual dataset.

We will limit our discussion to the latter, since the former does not contain re-identification risks. At a high level, a data model can include:

- Column names and types
- Statistics about each column, including frequency distribution information, maximum, minimum, cardinality, etc.
- Correlation information that describes the relationship between columns.

This model may be defined manually, extracted using data mining techniques, or learned using advanced machine learning algorithms. The sophistication of the generation method determines how close the generated dataset is to the original in statistical or analytical terms, and also how much re-identification risk is retained.

The values within a synthetic dataset can be generated using a set of techniques that preserve different levels of privacy risk and analytic utility, namely:

Random: attribute values are randomly selected without regard for the actual values within the original dataset. The basic schema (names and types of attributes) is observed but no other similarities exist. Randomly generated data contains minimal re-identification risk.

Uncorrelated data: attribute values are sampled from the distribution of the original values. However, the correlations between columns are not retained. This makes it less likely that large amounts of the original dataset will be reproduced. The privacy risk within such a dataset depends on factors such as how many records are synthesized, or the number of possible attribute combinations the schema supports.

Correlated data: attribute values are sampled as above, and the statistical correlation between attributes is retained. Correlated, sampled data retains significant re-identification risk, as large amounts of the original dataset are likely to be reproduced.

For data synthesis techniques that sample values from the actual distributions or a model generated from the original dataset, a data model is required. Data models derived from an existing dataset can be generated in a number of ways, with some examples listed here in ascending order of complexity, privacy risk and analytical utility:

- Simple statistical models: Models of data based on linear, statistical properties of underlying distributions [4,5].
- Bayesian techniques: conditional dependencies between attributes are modelled using a probabilistic graph-based approach [3].
- Generative adversarial networks (GANs): many modern approaches use this technique, based on neural networks that "compete" with each other [1,2].

Uses of Synthetic Data

Each of the techniques above produces a synthetic dataset with particular privacy and utility characteristics and each can serve a specific category of use case:

- High utility, high risk. If critical business decisions will be made using the synthetic dataset, thus requiring a lot of data utility, a technique that preserves analytical similarity should be used, such as GANs. These techniques reproduce a lot of personal data and so while they can be somewhat privacy-enhancing, they must be treated as having a high risk of re-identification.
- Low utility, low risk. Where realistic data is required to develop software, perform load or integration testing, a simple uncorrelated approach to data synthesis can be adopted. In this case, the accuracy of outputs is not important and therefore a higher degree of privacy may be preserved.

Note that these two examples represent extreme opposing ends of the privacy/utility scale, and that there can exist many more use cases that lie somewhere between these two.

General Analytic Utility -vs- Specific Use Case

Special attention needs to be paid to the distinction between data synthesized for general analytic purposes and that generated for a specific use case. There is something of a disconnect between some of the promising work described in recent literature around privacy-preserving synthetic data and the commercial applications of it.

Much of the literature around privacy-preserving synthetic data supports a singular defined use case (or set of cases), rather than general analytic utility. For these predefined use cases, there is evidence it may well be possible to produce synthetic data that is privacy-preserving but that retains a level of analytic utility.

However, the recent glut of commercial applications and techniques are more often than not touting general analytic utility rather than a singular defined use case, which is the focus of this paper. The research community and the commercial world need to work together to bridge this gap.

The Importance of Correlations in Synthetic Data

> "Inferences may still be drawn from the dataset, especially if attributes are correlated or have strong logical relationships."
>
> "As part of such residual risks, take into account the identification potential of the non-anonymised portion of a dataset (if any), especially when combined with the anonymised portion, plus of possible correlations between attributes"

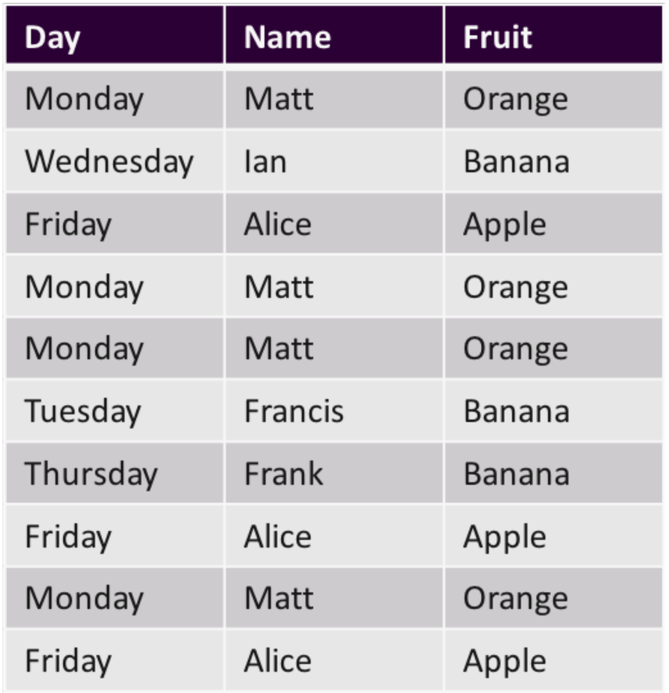
Article 29 Working Party Opinion on Anonymization Techniques.

The two primary factors that determine how much re-identification risk is present in a synthetic dataset are (i) whether the generation technique samples from the distribution of real values and (ii) whether correlations between columns are preserved. We will provide a simple example by way of illustration.

Figure 3 shows a dataset containing the fruit purchasing behaviours of a group of individuals on different days that we want to synthesise. A common first step with such a dataset is to alter or remove the direct identifier column entirely, in this case "Name". Figure 4 shows the "de-identified" version of this dataset, where the Name column has been simply removed.

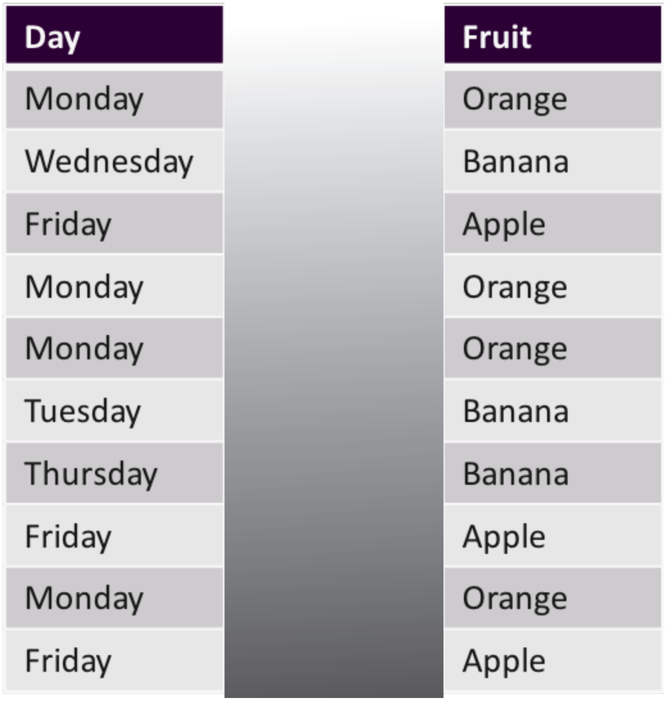
A data synthesis technique that preserves correlations in the original dataset and samples actual values from the original dataset's distributions could learn a model such as that depicted in Figure 5. If the exact distributions contained in this model are used to generate a new, synthetic dataset we will end up with 4 rows where an orange is bought on Monday, 1 row where a banana is bought on Tuesday, etc.

However, upon comparison with the original dataset in Figure 3, we can see that every time an orange is bought on Monday it is a unique "fingerprint" for Matt, just as the purchase of apples every Friday is a "fingerprint" of Alice. In fact, each day-fruit pair corresponds to exactly one name in the original dataset. Thus, the column combination of "Day" and "Fruit" forms a fingerprint, or quasi-identifier for the data subject. The synthetic data we have generated can be linked with the original dataset or another source of fruit-purchasing information because the unique fingerprints have been reproduced. Thus, we have produced a synthetic dataset with high analytical utility and a correspondingly high re-identification risk.



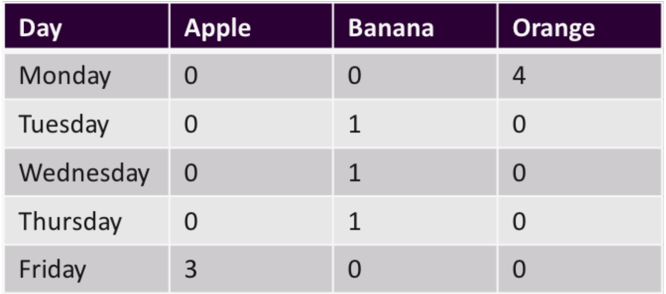
Day	Name	Fruit
Monday	Matt	Orange
Wednesday	Ian	Banana
Friday	Alice	Apple
Monday	Matt	Orange
Monday	Matt	Orange
Tuesday	Francis	Banana
Thursday	Frank	Banana
Friday	Alice	Apple
Monday	Matt	Orange
Friday	Alice	Apple

Figure 3:A dataset containing information about fruit purchasing patterns.



Day	Fruit
Monday	Orange
Wednesday	Banana
Friday	Apple
Monday	Orange
Monday	Orange
Tuesday	Banana
Thursday	Banana
Friday	Apple
Monday	Orange
Friday	Apple

Figure 4: De-identified fruit purchasing dataset.



Day	Apple	Banana	Orange
Monday	0	0	4
Tuesday	0	1	0
Wednesday	0	1	0
Thursday	0	1	0
Friday	3	0	0

Figure 5: Distribution and correlations of the fruit purchasing patterns.

The Importance of Measuring Re-identification Risk

While the example above is contrived for illustrative purposes, in practice, it may be difficult to determine if a data synthesis technique has replicated high-risk portions of the dataset. Consideration should be given to the types of re-identification risk defined in the Article 29 Working Party opinion, namely singling out, linkability and inference. Each of these can be quantified objectively using statistical analyses of the dataset. We believe that objective quantification using statistical analyses of the dataset should be the norm for demonstrating that the risk of re-identification has been reduced to an insignificant level.

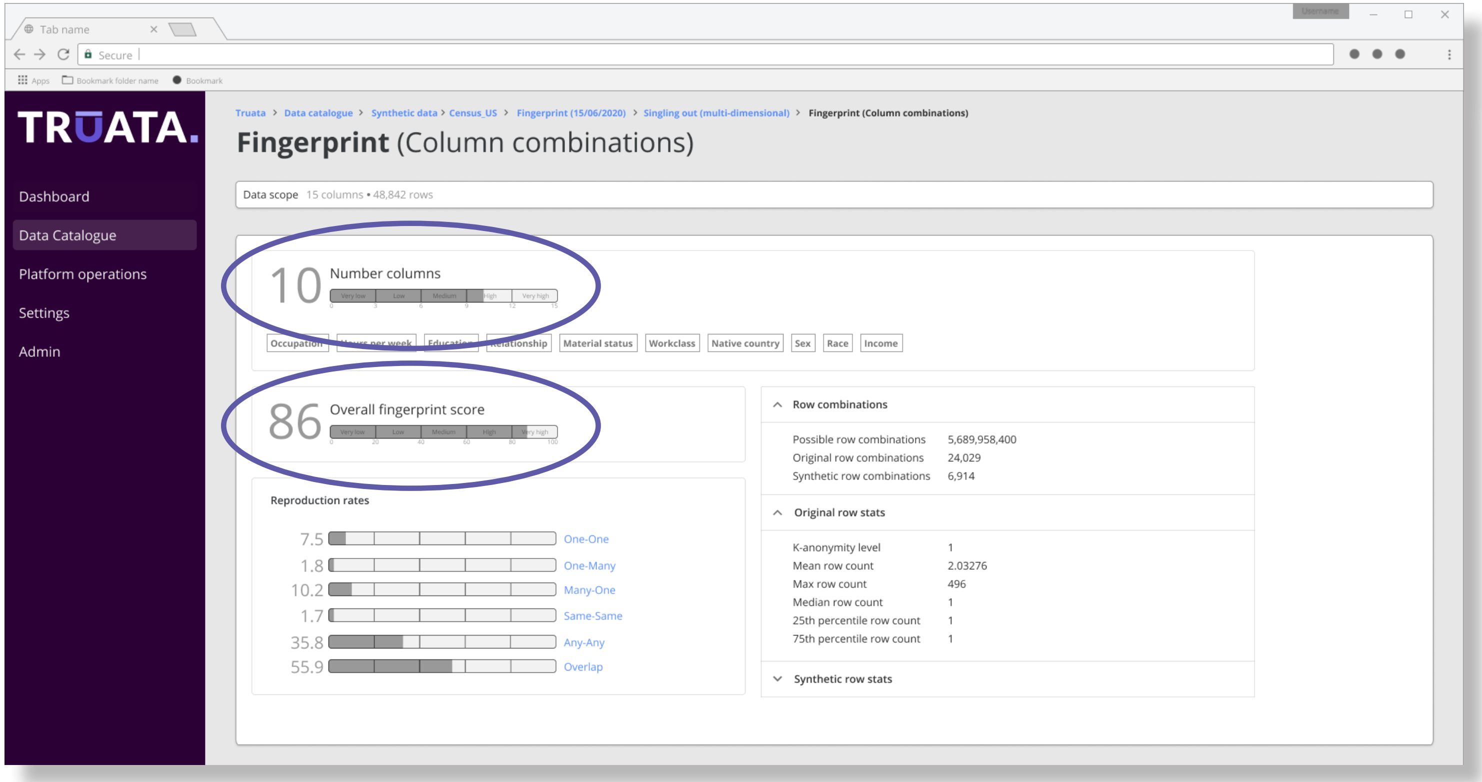
The first task is to determine which attributes combine to form the strongest unique fingerprints or quasi-identifiers that may enable singling out of individuals in the original dataset. These sources of re-identification risk may be reproduced by the selected synthesis technique and so they need to be measured in advance. Once the dataset has been synthesized, the level of overlap with the quasi-identifiers in the original dataset must be measured to determine the linkability risk. Depending on the desired outcome, any overlapping quasi-identifiers may be generalized, perturbed or removed entirely to further mitigate the privacy risk profile of the synthesized dataset. Without an understanding of these sources of re-identification risk, it is difficult or impossible to have confidence that the synthesized dataset can be used for the chosen purpose. Even where correlations between columns are not explicitly preserved, large amounts of the original dataset may be reproduced in the synthesized dataset. Depending on the characteristics of the dataset, this reproduced data may have very high re-identification capability [7].

By way of an example, Figure 6 shows the re-identification risk report produced by Truata's Calibrate tool for a synthetic dataset produced using a well-known technique that is used commercially by certain companies on the market.

It is claimed that this technique makes it impossible to identify individuals and hence the resulting data is safe to use or "anonymous". The figure concerns a 10-attribute quasi-identifier taken from a census dataset. The "Fingerprint score" for this quasi-identifier is highlighted, showing a score of 86% for this set of attributes, meaning in effect 86% of records in the original dataset can be uniquely identified using this set of attributes.

The "Comparison scores" reflect the extent to which the riskiest quasi-identifiers in the original dataset are reproduced in the synthetic dataset. As we can see, 56% of the rows for this quasi-identifier are reproduced in the synthetic dataset, meaning that it is far from impossible to re-identify individuals and therefore would not pass the threshold for GDPR grade anonymization.

This indicates that it would be risky to release the synthetic dataset publicly and without additional security and privacy protections.



Truata > Data catalogue > Synthetic data > Census_US > Fingerprint (15/06/2020) > Singling out (multi-dimensional) > Fingerprint (Column combinations)

Fingerprint (Column combinations)

Data scope 15 columns • 48,842 rows

10 Number columns

Occupation | Hours per week | Education | Relationship | Material status | Workclass | Native country | Sex | Race | Income

86 Overall fingerprint score

Reproduction rates

7.5	One-One
1.8	One-Many
10.2	Many-One
1.7	Same-Same
35.8	Any-Any
55.9	Overlap

Row combinations

Possible row combinations	5,689,958,400
Original row combinations	24,029
Synthetic row combinations	6,914

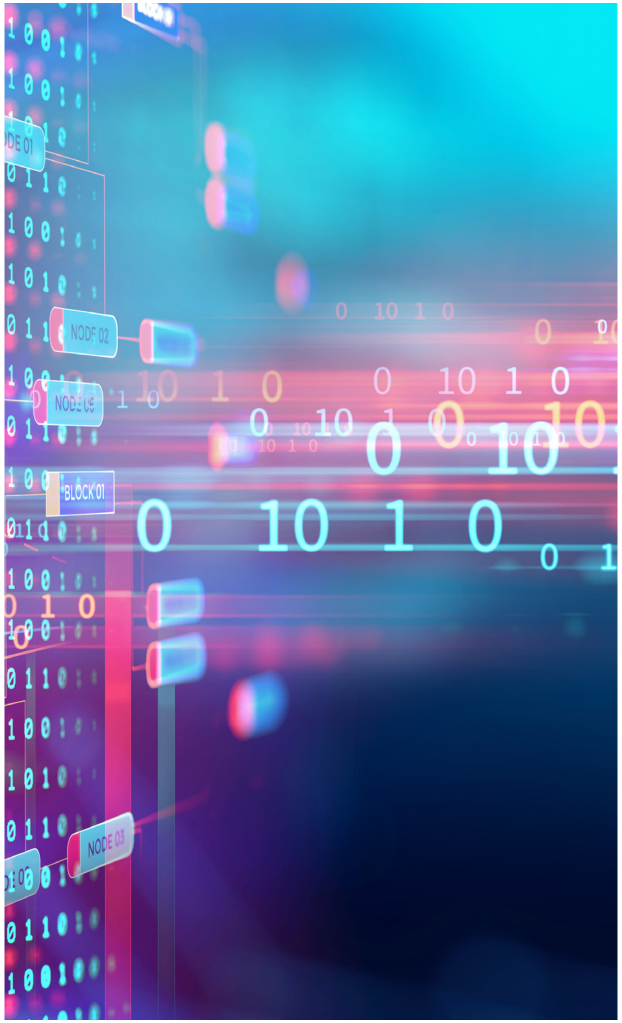
Original row stats

K-anonymity level	1
Mean row count	2.03276
Max row count	496
Median row count	1
25th percentile row count	1
75th percentile row count	1

Synthetic row stats

Figure 6:The re-identification risk within a synthetic dataset shows a high degree of reproduction of risky quasi-identifiers.

Conclusions



Data synthesis is a powerful approach to facilitate the production of large-scale datasets that are similar to an existing dataset. Many different approaches exist, each with very distinct privacy and utility characteristics, determined by how much of the original data is reproduced. This is very important in a world where a drive for companies to be data-driven conflicts with the need to abide by data protection regulations and engender trust with customers.

Synthetic data is increasingly being positioned as a means to perform general analytics in a privacy-conscious way. This has the potential to create confusion and a misunderstanding that such data is fake or non personal data and is inherently safe and therefore may be used for any purpose without additional safeguards.

In this paper, we have reviewed some common approaches to data synthesis, and noted how methods for generating synthetic data can produce datasets that are useful for specific purposes.

We have described how a synthesized dataset can contain significant re-identification risk and outlined how objective methods for quantifying privacy risk before and after can aid data controllers in making informed decisions about appropriate uses of the synthetic datasets they have produced.

Find out more at truata.com

References

1. Park, Noseong, et al. “Data synthesis based on generative adversarial networks.” Proceedings of the VLDB Endowment 11.10 (2018): 1071-1083.

2. Jordon, James, Jinsung Yoon, and Mihaela van der Schaar. “PATE-GAN: generating synthetic data with differential privacy guarantees.” (2018).

3. Zhang, Jun, et al. “Privbayes: Private data release via bayesian networks.” ACM Transactions on Database Systems (TODS) 42.4 (2017): 25.

4. Sue, Leurgans. “Linear models, random censoring and synthetic data.” Biometrika 74.2 (1987): pp. 301-309.

5. Rupert Miller, Jerry Halpern. “Regression with Censored Data”, Biometrika, 69.3 (1982): pp. 521-531.

6. Stefanie Koperniak. “Artificial data give the same results as real data — without compromising privacy”, MIT News (2017) https://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303

7. Rocher et al. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10, 3069 (2019). https://www.nature.com/articles/s41467-019-10933-3