Department for Science, Innovation & Technology

Government Digital Service

AI Insights

Spotlight: Synthetic Data

What is synthetic data?

Synthetic data refers to data that we create to mimic the properties and patterns of real-world data. We maintain the statistical properties, relationships and distributions of that original data, often with some corrections or modifications that make it especially beneficial in the development and tuning of machine learning models.

Why do we need it?

It's not because we're running out of data, or we've used up all of the internet already! Our problems are generally not quantity, but quality related. Synthetic data may also remove risks often associated with real-world data, like bias or privacy concerns.

We also create synthetic data due to scaling problems. We may wish to model weather events, and extreme weather records may be insufficiently represented in real-world datasets as a percentage of the entirety.

The same applies to modeling systems for medical analysis. It can be difficult to model for rare illnesses when they are especially sparse as a proportion of the general population. The desirable data effectively gets drowned out by the vast majority.

Bias is another concern in AI modeling, specifically algorithmic bias. This is where data may reflect historical inequities, or cultural bias, with sampling errors baked in. With synthetic data we can equitably rebalance datasets – provided, of course, that we do so correctly.

We also create synthetic datasets where we have privacy concerns, for example where data may have personally identifiable characteristics. Anonymised data may be

insufficient, and can potentially be reconstructed under specialised circumstances. Read more about this in Narayanan and Shmatikov's work on the Netflix anonymised dataset.

There are usage scenarios where, to guarantee privacy, confidentiality or secrecy, we generate entirely novel artificial datasets which reflect our target distributions. These could never be used to correlate against other known datasets, as they cannot possibly be members of those sets.

How do we create it?

Synthetic data is created through a variety of methods, like Monte Carlo Simulations or quantitative methods based on Gaussian (normal) distributions. The latter is especially popular and takes a similar approach to sampling for research purposes. To reflect the population, we ensure that we know the feature distributions. We know what percentage of each feature is present as a proportion of the source data.

From this we generate artificial datasets with these same percentage representations. We can also generate 'what if' scenarios if needed, such as altering the scale to emphasise lower populations, or to increase variability in subsets with high similarity.

This approach works very well for generating facsimiles of many real-world datasets. There are some cases where our data is more elaborate: non-linear, or multi-modal distributions. These can be accommodated by many specialised algorithms suited to these tasks depending on our requirements.

Responsible synthetic data development

Poorly generated synthetic data can introduce unrealistic patterns and distributions within the data. The model will learn these patterns, yielding unsatisfactory performance.

This is hardly an ideal outcome. It is important to ensure sufficient testing of models to give us confidence in the results. In the case of large language models (LLMs), refer to our LLM Testing AI Insights article for more information.

If you are using synthetic data in the Test dataset, and it has been configured similarly to the training set, you may end up with a model which has passed training and testing configurations but fails on real-world data in production. Synthetic data should support the different usage scenarios expected from the different phases of model development.

Over-estimating the contribution of synthetic data as part of model training or fine-tuning is another concern we must bear in mind. This is the false and misleading notion that we are not exposed to bias, ethical concerns, or extremes of scaling or misalignment simply because we are using synthetic data. This is quite a common misconception.

Synthetic data is just as vulnerable to weakness, bias, omission and so on, as real-world data. It must be subject to the same evaluations as real-world data. Engineering rigour applies, as for any data.

The quality of artificially created data is important. Mistakes in sampling, estimation or measurement of the source data will be reflected in the synthetic dataset. These can be subtle, especially where there are hidden, or at least unknown or unmeasured, dependencies between attributes of the underlying set.

Removing features from source data can introduce unexpected bias where dependencies or relationships are harmfully affected. In retail or financial data, removing features such as gender, age or marital status may remove key identifiers which qualify subsets of that data meaningfully, often with unfavourable consequences. This is not to say that we cannot remove features, simply that we must know and measure the consequences and the relationships.

Attempts to rebalance biased data also require careful attention. Corrective efforts can negatively skew the distributions and introduce spurious bias in other dimensions of the dataset. Bias correction in synthetic data should involve continuous validation against fairness metrics. Independent validation datasets and cross-validation can help ensure that synthetic data does not inadvertently bias model outputs. Refer to the Evaluation Metrics AI Insights article for more information.

Rigorous testing and validation can make it easier to incorporate synthetic data in model development. Given our precise knowledge of the distributions in our artificial data, we can target testing and evaluation knowing the expected results. Distribution similarity metrics (divergence or distance measurements) allow us to quantify our results and adjust accordingly. We must ensure that the inclusion of artificially generated data has not unintentionally biased our model towards artificial patterns not reflected in real-world distributions.

It is important to remember to document and version control any synthetic datasets used in model development, so they may be evaluated and recreated if needed.

Final thoughts

Synthetic data is very useful under the right circumstances. It is not a replacement, but a powerful complement to real-world data. We use it to fill gaps in our datasets, and to address specific concerns, but we use it combined with real-world data and real-world testing and validation. Synthetic data can be a tremendous addition to our toolkit when used responsibly.