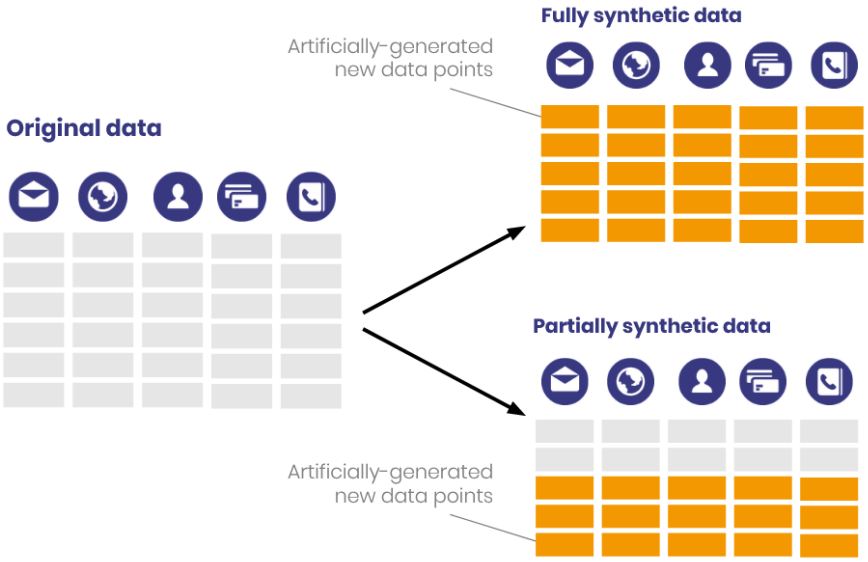
Synthetic Data: Hype, Hope, or Hard Proof?

The last year has seen a remarkable increase in research companies talking about adoption of synthetic data. It seems to be taking the research world by storm with a host of arguments for and against the use and application of synthetic data. A Conversion Alchemy report suggests that there will be a fundamental shift in how organisation derive insights to make decision as the synthetic data market is projected to grow from $267 million in 2025 to over $4.6 billion in 2032 which seems an exponential growth and adoption of this source of data.

It is known that synthetic data is created artificially based on patterns within existing or freshly captured datasets to mimic and project consumer behaviour patterns, actions and characteristics as compared to the reality of primary surveys which capture actual behaviours, feelings and emotions of consumers. We see that synthetic data is used mainly for low incidence, hard to reach audiences to amplify and mushroom the sample for a project. The critical element here is the quality of the minimal data that is used to generate synthetic data as deep validation is needed to ensure high quality is maintained and inherent sample biases do not get bigger. This would create doubt in the user's mind on the quality of synthetic data as business decisions rest on insights and outcomes being generated from this data set.

With AI democratising data processing and analysis and creating a sort of level playing field within research operations, synthetic data becomes another form of generative AI to create operational efficiencies and hence reduce sample acquisition costs for agencies. Of course, this will ease bottomline pressures which would make it favourable for research agencies to look into synthetic data solutions and possible adopt it.



This brings us to possible use cases for synthetic data. On the face of it, synthetic data can be used for concept testing but the key question here is if we test 5-6 concepts and randomly remove a portion of the sample to make way for synthetic data, at what point do the results start deviating from the main sample results? Does a reduced sample make sense? Is it also stimulus dependent especially with an new innovation concept is new to the market and is a disruptor for whom benchmarking data is limited? This is where strong validation is needed for results to be valid and replaceable from synthetic data and convince users that they can take up the results with confidence.

Another possible use case is for communication and message testing where synthetic data usage follows the same logic as mentioned for concept testing as long as the results are validated against relevant norms and benchmarks.

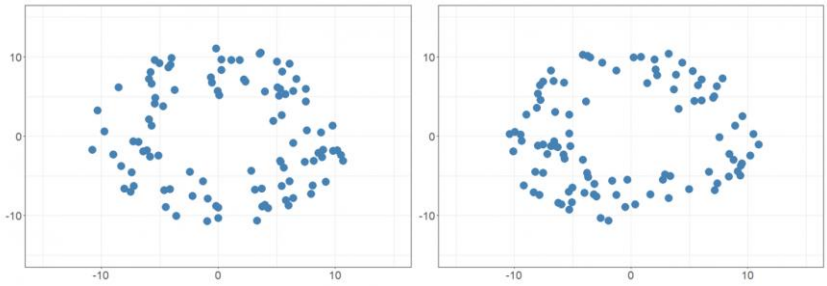
An area where synthetic data could fall short is when we look into experience-based research i.e. brand, consumer and shopper experience as human reaction and behaviours here are very individualistic and needs to be captured in reality rather than projected through synthetic data which might paper over some of nuances we see in the experience world. Some synthetic data models will attempt to mimic consumer behaviour and experiences but in a rapidly evolving marketplace with the online and digital world linked to instant gratification driving key consumer moments, replicating consumer or customer experiences will be quite a challenge.

An couple of Insight leaders at global FMCG company mentioned to me that they are looking at synthetic data for the flexibility it offers for early stage research but will not risk using it when making big business decisions There could be a thinking here that synthetic data could find it's place in the run of the mill tests and projects taking over the heavy lifting of sample acquisition and providing flexibility to enhance sub groups. It is telling, though, that synthetic data will not be used for major investment and high stakes outcomes

Another point to note for synthetic data also is how it represents the target market or the audience which it may do on the surface, but when it comes to sub-group analysis, there could be marked differences as evidenced in a validation exercise done by Vera Sight in US.[1]

While the LLM data generated has been used in the political / social world in this case, would there be similar parallels in the consumer sector where the key question is if the model can accurately cover and report on sub groups and niche audiences and also how the models react to questions they have not been trained on. Apparently not, with synthetic data not able to accurately represent smaller brands and smaller target groups and also an in-built bias towards more generic preferences which could be misleading especially for innovation, messaging and comms testing.[2]

A recent Research Live article mentions that 'Some claim that synthetic data is upto 80% accurate. Is that a reason to believe or is it an embarrassment?' The true test for synthetic data will be it's performance on hard-to-reach audiences and also heterogenous targets as it does well where there is enough data available and homogeneity is evident.



The synthetic data retains the structure of the original data but is not the same

[1] https://www.verasight.io/reports/synthetic-sampling

[2] https://www.verasight.io/reports/coffee-llm

Therefore, users of synthetic data i.e. the insight teams at the marketer's end who the research agencies will offer synthetic data solutions will increasingly ask the question – how has the synthetic data been generated or will be generated? They will want to know more about the details around the synthetic data black box and how the data has been built by the LLMs or other approaches especially since business decisions will rest on the insights and outcomes from this data set.

There will be a cost-benefit analysis done to assess the application of synthetic data based on the risk appetite of business needs. In terms of investment, there will be different actions as regard to these data sets depending on low risk v/s high risk objectives. Having said that, it will still be the human who translates this data to insight and now the LLM machine.

When the dust from synthetic data clears, there will be only one winner and it will not be based just on LLM or AI tech being applied but the ones who deliver maximum business impact through empathy and liking research outcomes to actual business needs whether it's through the use of humans, AI and synthetic data or a combination of any of these as long as representation is guaranteed and accurate.

The winners won't be the ones adopting or applying synthetic data first but ones who can imbed commercial insights through use of technology without losing the voice of the consumer

There is an overarching feeling amongst clients that what real humans say matters more than ever in the world where humans use technology more and more. Will clients maintain their ages old developed patience of traditional research that takes more time and more money when the same could be done quicker and with lesser investment using synthetic data? Will there is a greater need to prove accuracy and ensure reality of consumer truths are outcomes of research?

Of course, this leads to insight users and insight agencies looking at some elements of systematic change when it comes to synthetic data adoption in the midst of possible change fatigue. It will be up to us as sector custodians to be able to manage this change for both sides as we try to get closer to consumer truths with higher levels of accuracy. The challenge we all face is presenting insights to our client who in turn act on this synthetic precision posing as human signals, as data driven insights. Is that right or wrong? That judgement will sit in the hands of the clients who see if synthetic data is fit for business decisions or not or is it a situation where synthetic data works in synergy with traditional methods by being complementary.

In sum, the jury is out, we are all waiting to see if the adoption of synthetic data will play as is being projected.