IPSOS VIEWS

SYNTHETIC DATA BOOSTING

Unlock the transformative potential of data augmentation

David Priestley
Maciek Ozorowski
Jon Puleston

At Ipsos, we champion the unique blend of Human Intelligence (HI) and Artificial Intelligence (AI) to propel innovation and deliver impactful, human-centric insights for our clients.

Our Human Intelligence stems from our expertise in prompt engineering, data science, and our unique, high quality data sets – which embeds creativity, curiosity, ethics, and rigor into our AI solutions, powered by our Ipsos Facto Gen AI platform. Our clients benefit from insights that are safer, faster and grounded in the human context.

#IpsosHiAi

> Synthetic data has become an important part of Ipsos' research practice, helping us generate deeper insights.

The evolving uses of synthetic data

Synthetic data has become an important part of Ipsos' research practice, helping us generate deeper insights when real data, such as in questionnaire-based consumer surveys, is limited or unevenly distributed.

To advance this fast-moving field, Ipsos continues to invest in new methods and technologies that make synthetic data more accurate, transparent, and useful[1]. Our goal is to lead with science – ensuring every approach we use is carefully tested and scientifically sound.

To support this, we created a dedicated **Synthetic Data Research Department**, bringing together a team of AI and machine-learning experts and working closely with leading academics through our Scientific Council. Their focus is to test, refine, and apply synthetic data solutions responsibly across our business.

Working with both industry and academic partners – including a continuing collaboration with Stanford University – Ipsos has developed new techniques such as tabular diffusion models for market research data, built the 4D Integrity Framework (SURE) to evaluate data quality, and created a **synthetic data workbench** that brings these methods into everyday operational use – standardizing and productionizing our data boosting capabilities.

This paper focuses on **synthetic data boosting, which is one of several key applications of synthetic data** that Ipsos has been leading the way in applying, alongside uses such as imputation (to help shorten surveys), fusion (blending datasets together), and simulating responses (using AI personas and digital twins).

Each of these approaches raises its own set of methodological questions and practical considerations. This paper is the first in a planned series exploring the different roles and uses of synthetic data, outlining their value, limitations, and the questions they raise for researchers and clients alike.

What is synthetic data boosting?

Synthetic data boosting is the process of expanding an existing dataset by modeling the relationships within the original data and generating new synthetic cases that preserve its key patterns and constraints. Its purpose is to increase data availability and strengthen analytical power.

> "It's basically a way of teaching a model what your data looks like so it can create extra data points that behave the same way."

The use of synthetic sample boosting raises several important questions for the research industry, many of which are outlined in the ESOMAR paper, *Five Topics of Discussion to Help Buyers of Augmented Data*[2].

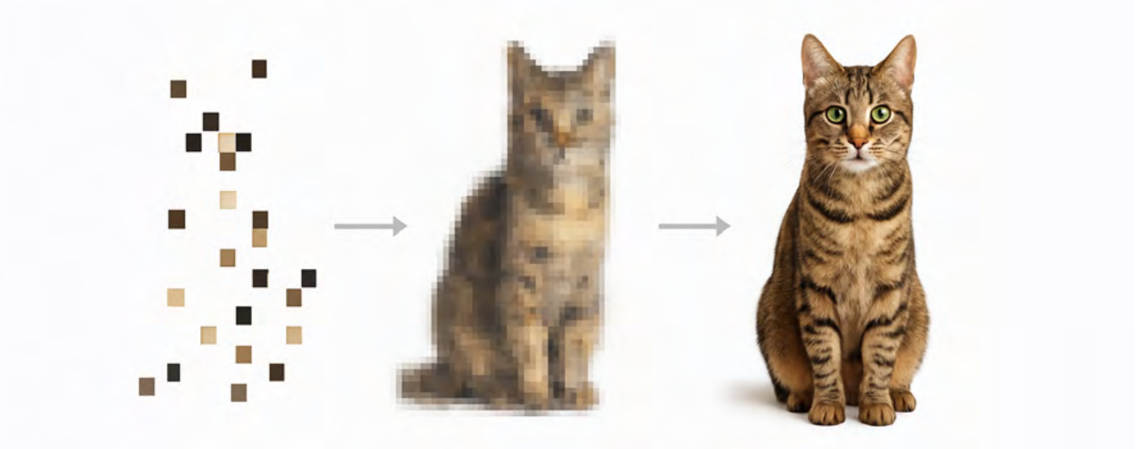
In this paper, *Synthetic Data Boosting*, we address some of key questions based on our current best knowledge, testing, and applied experience at Ipsos.

Specifically, we explore:

- How do we boost data at Ipsos?
- How do we prepare data to be synthesized?
- How can we evaluate the reliability of synthetically generated data we create?
- How much training data and sample is needed to reliably boost data?
- How much can you boost by?
- What is the value of synthetically boosting your data?

Similarly, the rise of synthetic data in market research will also take time. We should see this as an opportunity to harness its enormous potential and create the means by which we can use it safely.

> **The use of synthetic sample boosting raises several important questions for the research industry.**



How do we synthetically boost data at Ipsos?

There are many established ways to expand or "boost" data, from traditional weighting and bootstrapping methods to more advanced statistical and machine-learning techniques. More recently, neural-network-based generative models – such as diffusion models, transformers, and generative adversarial networks – have opened up new possibilities for enhancing limited or imbalanced datasets.

In recent years, much of the cutting-edge work has centred on Generative Adversarial Networks (GANs) - models that learn to generate new data by pitting two neural networks against each other: a generator, which produces synthetic examples, and a discriminator, which evaluates how real or fake they appear. While powerful, GANs are fundamentally unstable to train, hard to control, and prone to memorizing – leading to middling synthetic data quality. To address these limitations, Ipsos has focused on developing a new approach called **tabular diffusion** by adapting diffusion transformer (DiT) techniques – originally created for visual simulation – to work effectively with tabular data.

Why would image-generation algorithms be relevant for quantitative data? Any image is simply a vector (an ordered collection) of pixel values, each representing a color. Survey data follow the same structure: each respondent provides a vector of answers across the questionnaire.

Just as a well-trained neural network can learn the statistical relationships between pixels to generate new, realistic images – the types of generative AI images we are now all familiar with – another neural network can learn the relationships between variables in a dataset to generate new, coherent survey observations that mirror the structure of the real data.

Another useful parallel comes from the way images can be recognized even when only a modest number of pixels are present. A few well-placed pixels are enough to suggest a shape, which a generative model can refine by inferring the missing details. In much the same way, a dataset with a limited – though not too limited – number of observations can still contain enough structure for a model to identify underlying patterns and synthesize additional rows that complete the "picture" in a statistically consistent way.

The application of diffusion architectures has enabled hyper-realistic images that humans are no longer able to recognize from real[3]. At Ipsos we learned that with necessary adjustments, it's capable of doing the same for survey data. To understand how that happens, let's dive into how diffusion works.

Technical Note:

What is tabular diffusion

Diffusion is a process in which random noise is gradually added and then removed from data, allowing a neural network to learn to separate signal from noise - by learning to "clean" the noise step-by-step, it's only ever removing a small amount of noise at a time, so it can start from pure static to produce realistic, high-fidelity outputs.

This shift in methodology mirrors the transformation seen in image generation: early **GAN-based** models often produced distorted or implausible results - people with extra fingers or misaligned features, whereas **diffusion-based** methods enabled a step change in realism and fidelity. GANs learn through adversarial competition, capturing surface correlations without enforcing structural coherence, which explains why their outputs can look globally plausible yet locally impossible. Transformer-Diffusion models, by contrast, reconstruct structure iteratively through a denoising process, ensuring internal consistency at every stage and largely eliminating such artefacts.

When applied to **attitudinal or behavioral data**, the same risk persists but becomes far less visible. A GAN can generate statistically plausible yet psychologically incoherent respondents, combinations of views that could never coexist in real populations. Unlike an image with six fingers, these inconsistencies leave no obvious trace, making them difficult to detect. This is why **tabular diffusion** methods, which progressively learn and restore relationships between variables rather than confrontationally guessing them, provide a much more reliable foundation for synthetic survey data generation.

Ipsos is the first in the market research industry to develop and operationalize a **tabular diffusion methodology**, capable of delivering similar breakthroughs in data quality for synthetic data boosting.

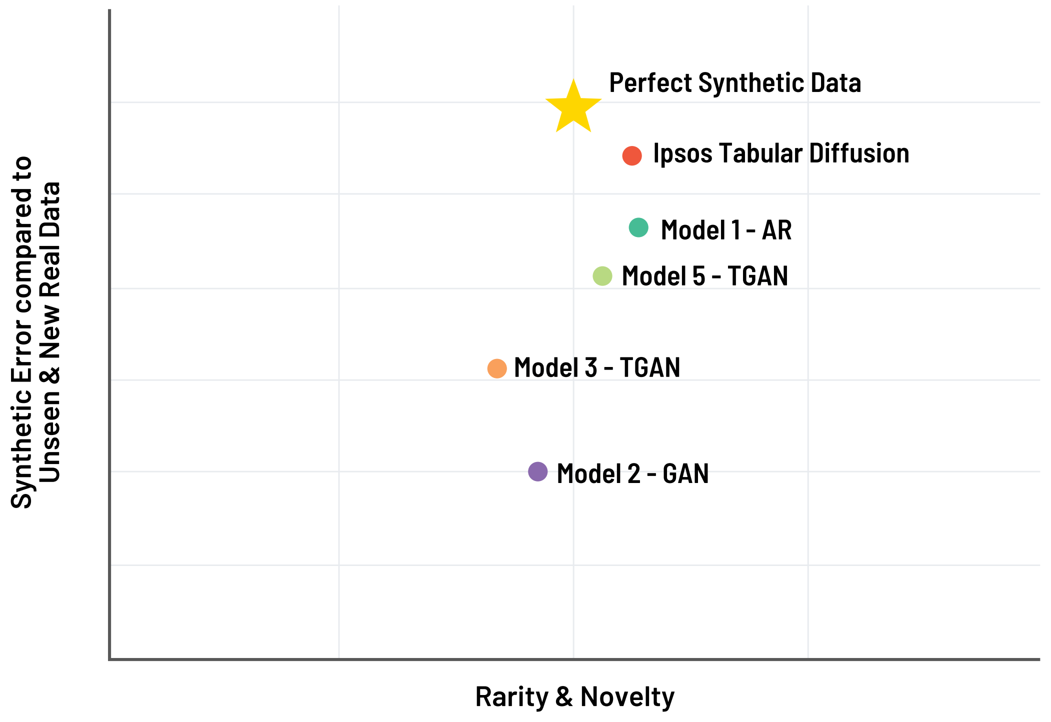
Ipsos' tabular diffusion model has been built specifically for market research applications - natively handling studies of all types: longitudinal, ad hoc, trackers, diaries and more.

Our tests show that this approach produces synthetic samples that are more faithful to the original data distribution, preserving both central tendencies and minority patterns. This results in datasets that are analytically robust, distributionally representative, and of high overall fidelity – a major advance in the creation of high-integrity boosted datasets. Our findings align with academic work on diffusion-based synthesis and data integrity[4].

Its application has enabled us to achieve step-change improvements in the boosted samples we produce for clients, delivering data with higher fidelity and enhanced privacy protection, while maintaining sufficient novelty to add analytical value without compromising respondent confidentiality.

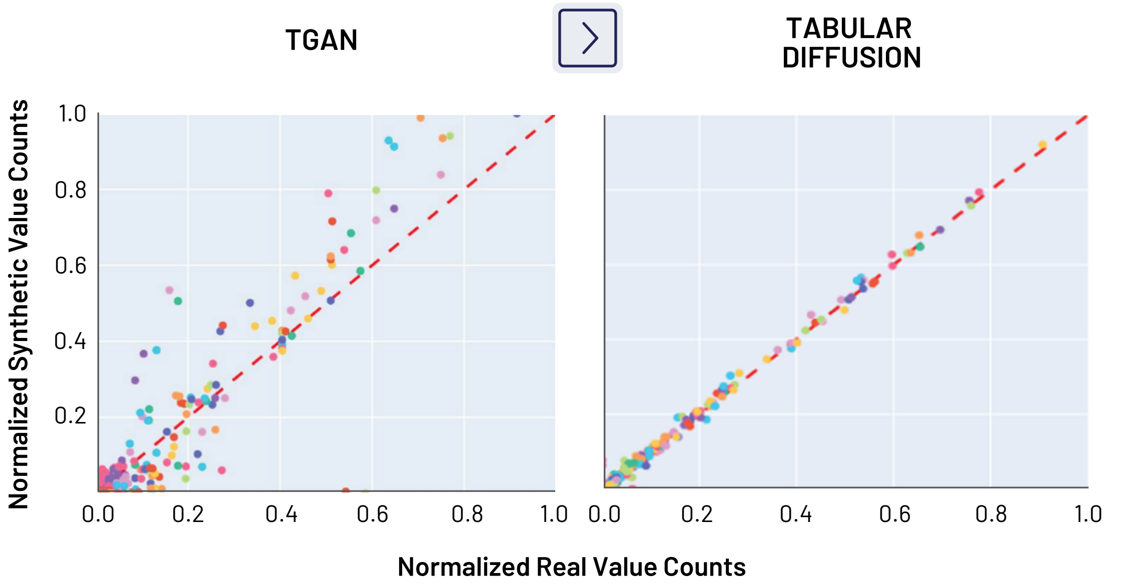
While our diffusion model is currently state-of-the-art, we recognize that as synthetic-data research progresses, diffusion methodologies will eventually be surpassed by new best-in-class generative AI paradigms. Ipsos' Synthetic Data Research Department is therefore committed to staying at the cutting edge of research, development and improvement, continually building, testing and benchmarking new methodological approaches.

Figure 1: Synthetic data rarity & novelty vs. synthetic data difference from new & real data



Source: Ipsos

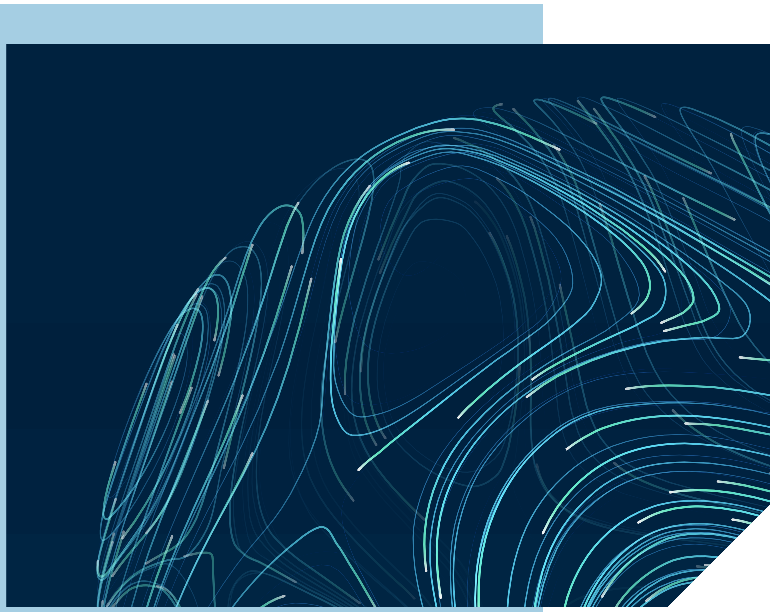
Figure 2: TGAN vs. tabular diffusion



Source: Ipsos

The Ipsos synthetic data workbench

To support this, we have spent the past year researching, developing and deploying our synthetic data workbench. This workbench is a suite of tools, procedures and frameworks, underpinned by a **curated and continually expanding portfolio of generative models** purpose-built for market-research data. The portfolio includes approaches tailored for rules-based survey structures, methods that integrate multiple related datasets, lightweight fast-learning models suited to smaller samples, universal imputation-and-prediction systems that ensure self-consistent outputs, and specialized time-series models designed for longitudinal trackers.



These approaches are the result of our own research and engineering work – designed specifically for the structure, logic and demands of market-research tabular and longitudinal data, rather than adapted from generic off-the-shelf generative frameworks.

How do we prepare data for synthesis at Ipsos

Thinking beyond the generalized approach to synthesizing data, one of the vital steps to undertaking reliable data synthesis is the preparation of the data, the cleaning and optimization and structuring process before synthesis takes place. Poor data input quality is one of the main failure points in this process.

> **"Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?"**

At Ipsos, as part of our synthetic data workbench we have developed a comprehensive data cleaning and optimization toolkit to ensure that the datasets used to train our models are as well-structured and representative as possible. This includes:

- **Standardizing variable formats** and coding schemes (e.g. aligning category labels, harmonizing scales, and ensuring consistent use of missing values).
- **Resolving logical inconsistencies,** such as respondents reporting behaviors that contradict demographic information or impossible answer combinations.
- **Detecting and managing outliers** or noise that could distort model learning.
- **Balancing and weighting** subgroups so the training data reflects the true structure of the population.
- **Encoding variables** appropriately to preserve relationships (e.g. treating ordinal, categorical, and numerical features differently).
- **Feature optimization** – selecting, transforming, or engineering variables to maximize the model's ability to capture meaningful patterns. For instance, consolidating sparse categorical features, normalizing skewed distributions, or creating derived variables that express latent relationships. Poorly optimized features can make models unstable or overly sensitive to random variation.
- **Checking correlation structures and dependencies** to ensure the model learns realistic, coherent relationships between variables.

While all these steps might not sound too exciting for someone who is not a data practitioner, they are absolutely crucial to enable reliable modeling and quality data generation – there is a reason why "garbage in, garbage out" is such a famous phrase in machine learning.

These processes also highlight how highly **structured** the training data must be for boosting and related algorithms to perform effectively. Unlike many machine-learning approaches that operate on unstructured data (e.g. language), boosting methods apply exclusively to **structured, tabular datasets**.

Ensuring data integrity through structured preprocessing and bootstrapped inputs

In addition to these data-preparation steps, we also apply bootstrapping techniques* to generate multiple variant training datasets. This approach mitigates the risk of overfitting to idiosyncrasies in any single sample and allows the model to learn relationships that generalize more robustly across the data space. Whereas many modeling frameworks focus on producing multiple modeled outputs (ensembles), our method also emphasizes the value of generating multiple, systematically varied inputs, ensuring that model learning reflects stable, underlying patterns rather than artefacts of a specific dataset.

** Resampling technique that generates multiple variant versions of the training dataset. This helps models learn relationships that generalize across samples rather than overfitting to the quirks of a single dataset.*

How do we evaluate the reliability of synthetic data at Ipsos?

Too often, the evaluation of synthetic data is reduced to basic distributional checks and claims of "error reduction" in simple holdout tests. Such assessments are overly simplistic and leave the most important questions unanswered: Is the data useful? Does it add new information? And does it appear authentic to an expert eye?

Are you 'SURE' about your synthetic data?

S U R E

Statistical Similarity

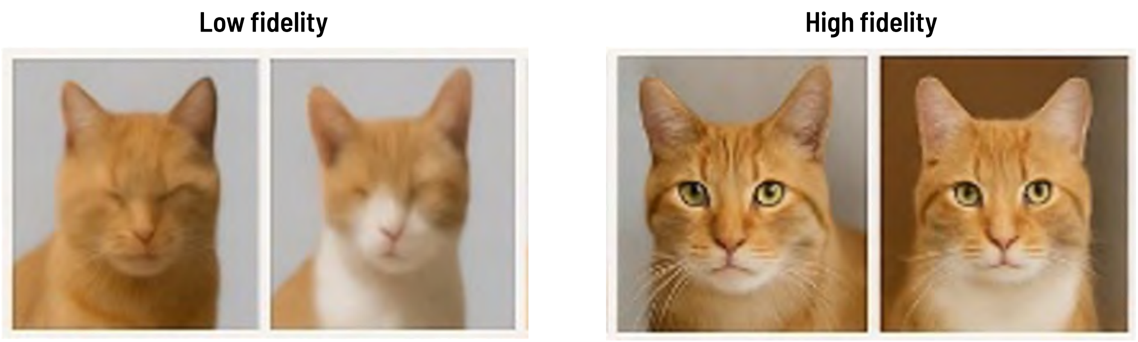
Utility & Fairness

Rarity & Novelty

Expert Validation

Synthetic data shouldn't just look plausible; it should behave, perform, and be judged as useful and trustworthy. **To evaluate that, Ipsos has developed a pioneering 4D evaluation framework: 'SURE'.**

SURE is a comprehensive system that tests synthetic data across four critical dimensions, each of them grounded in statistical validation and expert judgment.



S

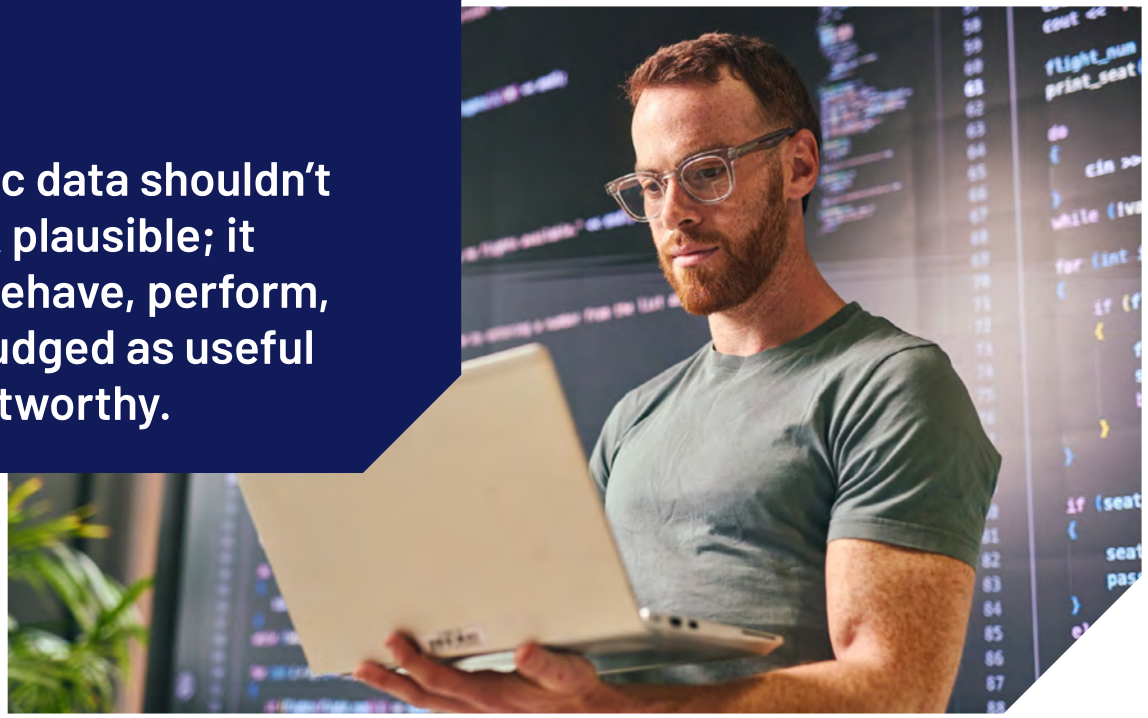
Statistical Similarity

Statistical Similarity – is it statistically faithful?

This measure concerns the **fidelity** of the synthetic data – how closely it reproduces the statistical properties of the **original (real-data) dataset**. We assess whether distributions, correlations, cross-tabulations, and logical constraints are preserved, both overall and within key subgroups. High fidelity means the generated data capture the same structural relationships and statistical patterns as the source, even though the individual records are newly created and the original data themselves may contain some bias.

Using an image analogy: a poorly performing generator may produce something that *resembles* a cat but lacks the fine detail and realism needed for recognition. Conversely, a well-performing image generator can create new images so convincing that humans cannot reliably tell whether they are real or AI-generated. In the same way, high-fidelity synthetic data are statistically indistinguishable from the source dataset, reproducing the level of detail required for **robust modeling, boosting, and decision-making.**

> Synthetic data shouldn't just look plausible; it should behave, perform, and be judged as useful and trustworthy.



Technical Note:

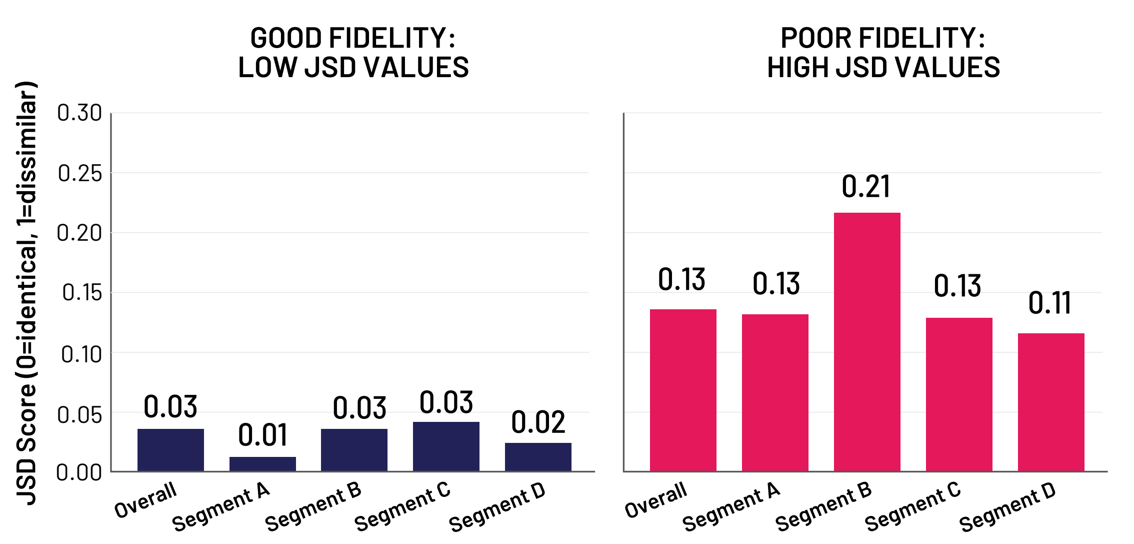
How Ipsos does it

We use a suite of techniques such as Jensen-Shannon Divergence, context aware scores, density estimations, PCA decomposition and correlation preservation analysis to quantify similarity between synthetic and holdout datasets - both globally and within decision-driving segments. These help us to, at-a-glance, pick out data anomalies.

This statistical error approach to validation is standard in the appraisal of synthetic datasets, in industry and academia, but it's where most stop. Statistical similarity alone isn't good enough.

Figure 3: Jensen-Shannon Divergence (JSD) comparison across segments.

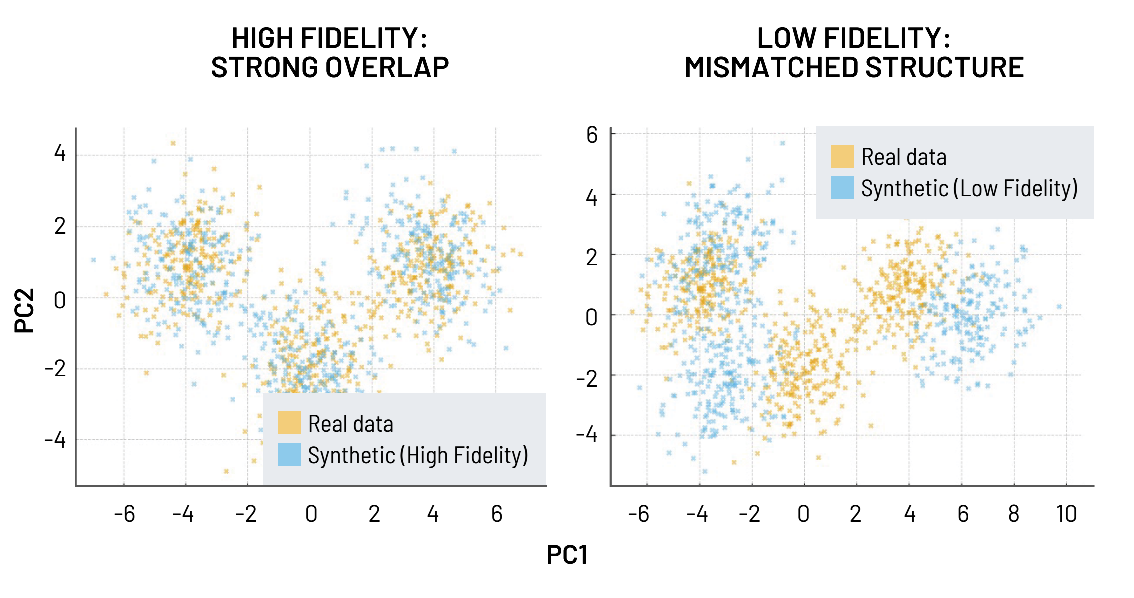
Left: low JSD scores (0.01–0.04) indicate high fidelity between real and synthetic distributions. *Right:* high JSD values (0.10–0.25) signal poor alignment and potential anomalies requiring review.



Source:
Ipsos

Figure 4: PCA decomposition examples.

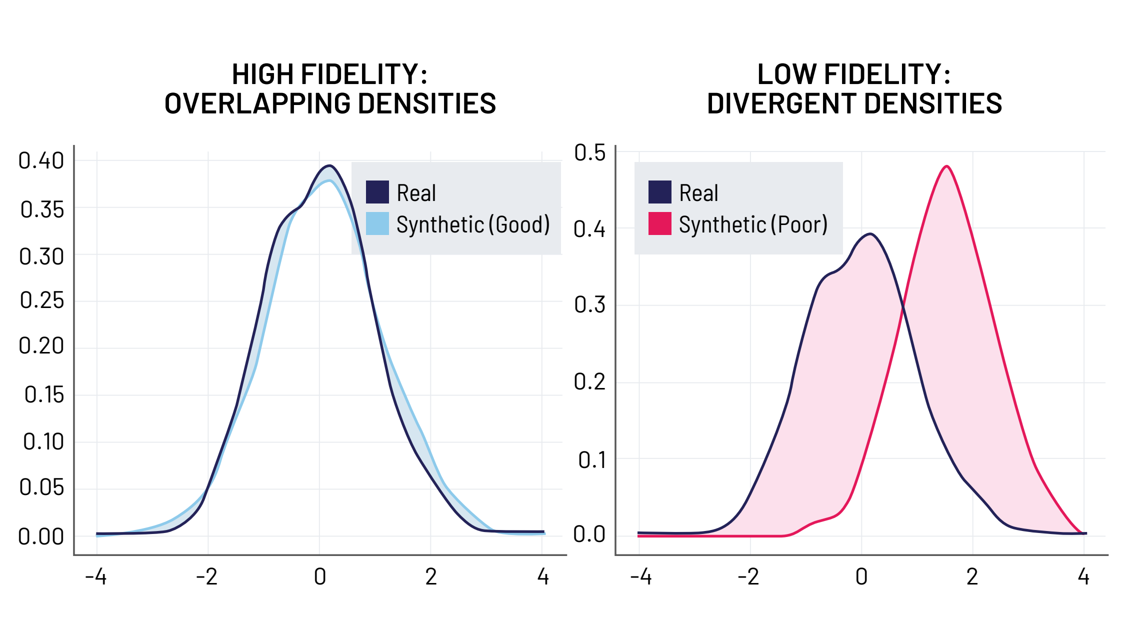
Left: High-fidelity synthesis where real and synthetic data align closely across principal components, preserving cluster structure. *Right:* Low-fidelity synthesis showing clear distributional drift and loss of structure, indicating model or data-preparation issues.



Source:
Ipsos

Figure 5: Density estimation comparison (KDE).

Left: high-fidelity synthesis shows close overlap with real distributions.
Right: low-fidelity reveals drift and shape distortion.



Source:
Ipsos

> For statistical analysis, the aim is to strengthen precision without distorting genuine variability.

U

Utility & Fairness

Utility – is it analytically useful?

The key question is whether the synthetic data enhances confidence and analytical power, for example, by **reducing random noise or sampling error** and thereby increasing the effective sample size. Importantly, this does not mean eliminating meaningful variance: excessive smoothing can mask real effects and increase Type II error. For statistical analysis, the aim is to strengthen precision *without distorting genuine variability.* For machine learning utility, the important question is whether the inclusion of synthetic data into model training improves the performance of our classifiers and regressors.

Utility is, in many ways, the critical pillar of evaluation – it answers the key question: *is our synthetic data actually useful?* In many **industry and competitor frameworks**, this question is seldom examined in a mathematically robust way. The risks of neglecting it are significant: if synthetic data are treated 1:1 as real in statistical testing, they can **inflate false positives (Type I errors)** dramatically. Depending on the generation method and dependence structure, such inflation can exceed ten-fold in simulation settings – effectively making small, random differences appear statistically significant.

In our framework, by contrast, we explicitly quantify **utility** in one of two ways: 1) using analytically driven formula to explore the amount of extra information our real dataset supports and 2) by doing bootstrapping procedures to identify the variability process, ensuring an unbiased estimate of the standard error(s). This allows us to compute meaningful **utility metrics** that reflect the improvement in statistical power, rather than an artefactual one. This helps safeguard against false confidence and ensures that decisions are based on authentic, replicable patterns in the data.

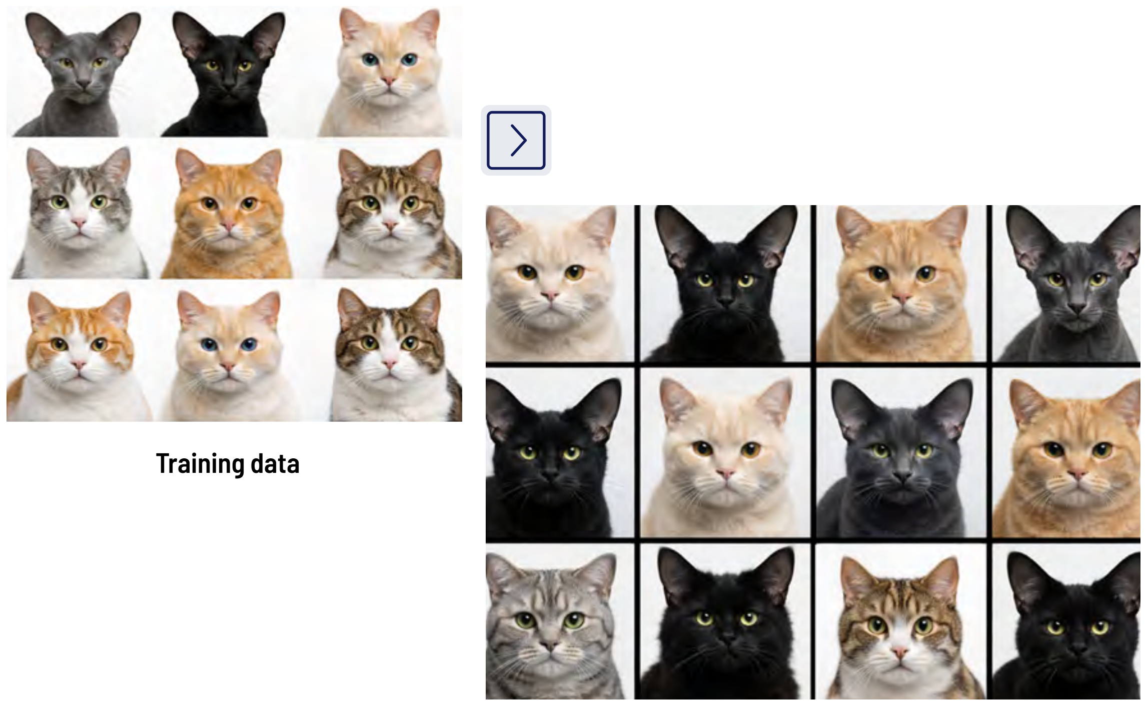
In many **industry and competitor frameworks**, the current most common approach involves finding a synthetic sample that simply minimizes an *error margin*, typically defined as the **accuracy error between each variable in the synthetic data and its counterpart in a real-data holdout.** But this says nothing about the information content of your data, or the variance equivalence of your sample size (crucial for not inflating Type I errors), it only tells you that you have found some data that will decrease this arbitrary variable's error margin.

Here's a simple way to think about it. Imagine you're boosting your cat dataset with lots of synthetic cats - but the model has a subtle flaw: it struggles to reproduce ears accurately. Before long, your expanded dataset is full of one-eared cats. When you analyze the data, you might proudly announce, "Our new research shows that most modern cats only have one ear!"

What's happened is that the modeling process has **amplified a systematic bias** – a small generation error – until it dominates the sample. The model has treated this artefact as truth, giving you a statistically confident but completely false conclusion. This is the essence of **Type I error inflation** in synthetic data: when a model's internal bias or random quirk becomes over-represented through data expansion, producing an illusion of significance where none exists.

Figure 6: Also face the issue of amplifying biases in the training data (Type I errors)

Boosting the number of black cats produces a population in which all black cats have large ears, as the training data contained only two black-cat breeds — both with large ears — turning a sampling artefact into a spurious structural pattern (**a Type I-like error**).



Technical Note:

How Ipsos does it

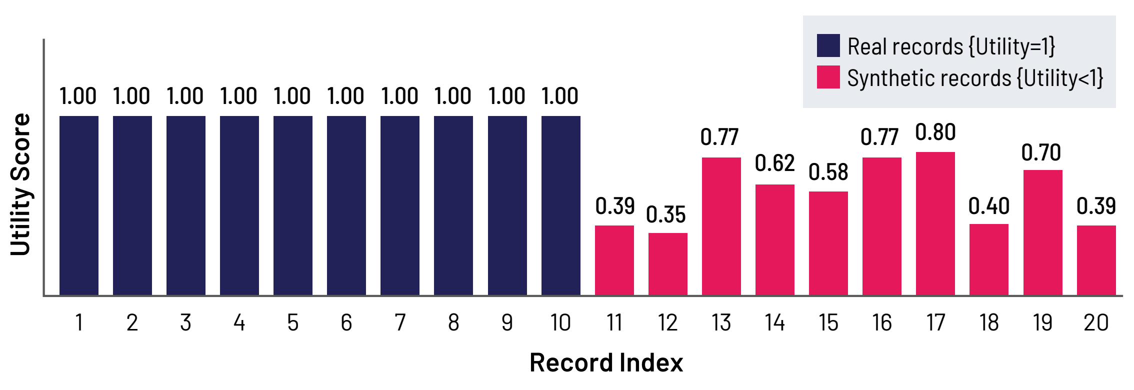
Statistical Utility: For inferential applications, we assess whether the inclusion of synthetic data improves analytical power and confidence without inflating error. In particular, we focus on two approaches: 1) analytically derived variance equivalence, and 2) empirical estimation of the standard error of synthetic data and effective sample size calculations (ESS). This ensures that any increase in apparent precision reflects genuine information rather than artefactual stability.

Our analytical derivation of a variance equivalent synthetic sample allows us to quickly, and computationally cheaply, understand how much 'fresh' information our synthetic sample is bringing. In this case, the real data defines the upper limit of utility – each real sample identically has a perfect utility score of 1. A synthetic record will have a score in the range of 0-0.999... This is a directional measure that does not control the confidence inflation and thus does not make the synthetic data immediately fit for statistical testing purposes.

The only way to truly understand the controlled gains in statistical utility and to ensure our statistical testing does not suffer from overconfidence is by empirical testing. To do this, we estimate an ESS that maintains type I error under strict control through bootstrapping. It is critical the bootstrapping is done on the training sample and the models are retrained each time - not on the generated sample - to capture all sources of uncertainty. More broadly, each synthetic dataset is assessed for the incremental informational value it contributes, ensuring that any observed gain in power is both valid and meaningful.

Figure 7: Illustration of record-level statistical utility.

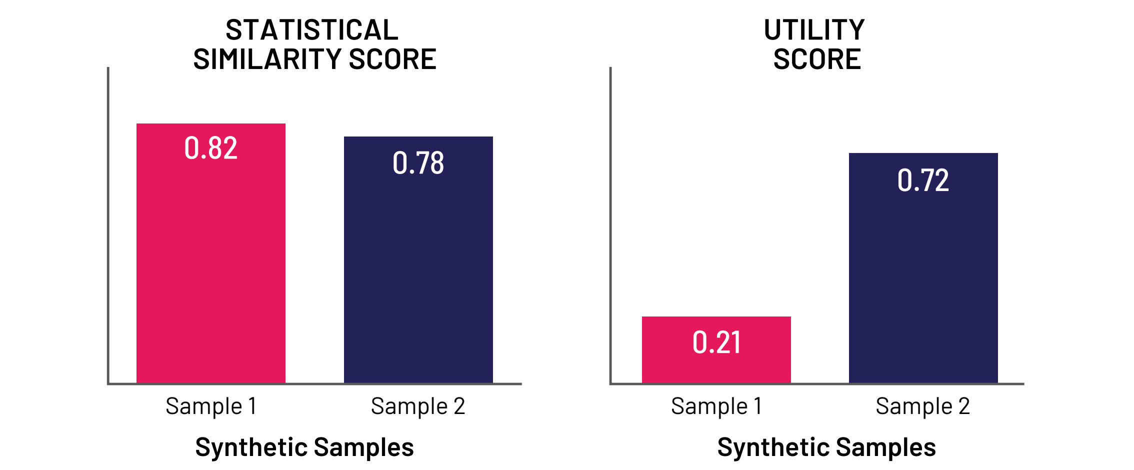
Real records (blue) each contribute full information (utility = 1). Synthetic records (pink) contribute partial information, with utility scores below 1. The effective sample size (ESS) is the sum of these weighted utilities, preserving correct inference power without inflating Type I errors.



Source: Ipsos

Figure 8: This figure shows two synthetic datasets from the same original data.

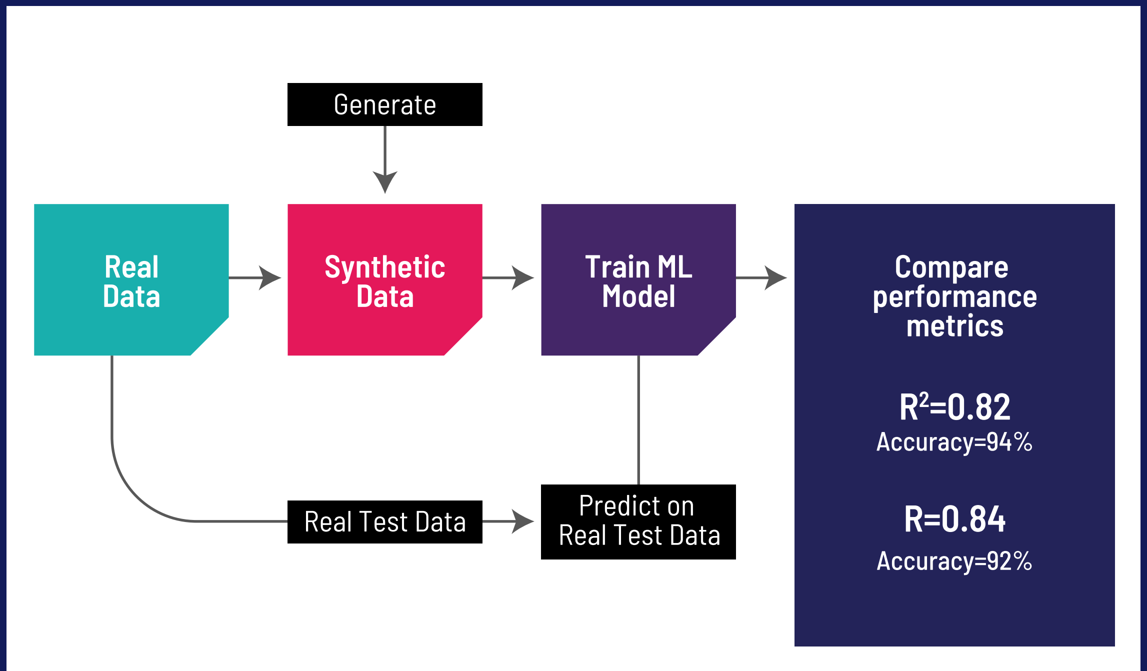
Despite both synthetic datasets having "good" statistical similarity, we can see the left one contains a much less useful set of information, Whereas the other sample, with slightly lower fidelity, has much higher utility scores. Statistical testing with the left one would have given us a huge sense of false confidence!



Source: Ipsos

For machine learning utility, we do Train-Synthetic-Test-Real (TSTR) to understand if our synthetic data is useful in machine learning tasks and improves our algorithm's performance on regression and classification tasks.

Figure 9: Train-Synthetic-Test-Real (TSTR) evaluation



> Synthetic data's novelty lies in generating new, realistic combinations of attributes that enrich the information space, rather than simply replicating existing cases.

R

Rarity & Novelty

Rarity & Novelty – does it bring anything new?

Does the synthetic data extend the real data by filling gaps, revealing rare patterns, or representing the "long tail" of real behavior - without duplicating near neighbors or fabricating unrealistic combinations?

Novelty ensures that synthetic data isn't just a copy or "fancy weighting" - it's a meaningful enrichment of what exists.

Again, looking to image generation for analogy: a well-performing algorithm can create images of cats it has **never seen before**, yet they remain hyper-realistic and consistent with what "catness" looks like. There is little value, by contrast, in producing more images of the same cats already present in the training data. The same applies to synthetic data, its **novelty** lies in generating new, realistic combinations of attributes that enrich the information space, rather than simply replicating existing cases. While concepts such as **fidelity, utility,** and **novelty** inevitably overlap, they are not identical and often interact in complex ways: high fidelity ensures realism, novelty provides diversity, and utility determines whether that diversity adds genuine analytical value.

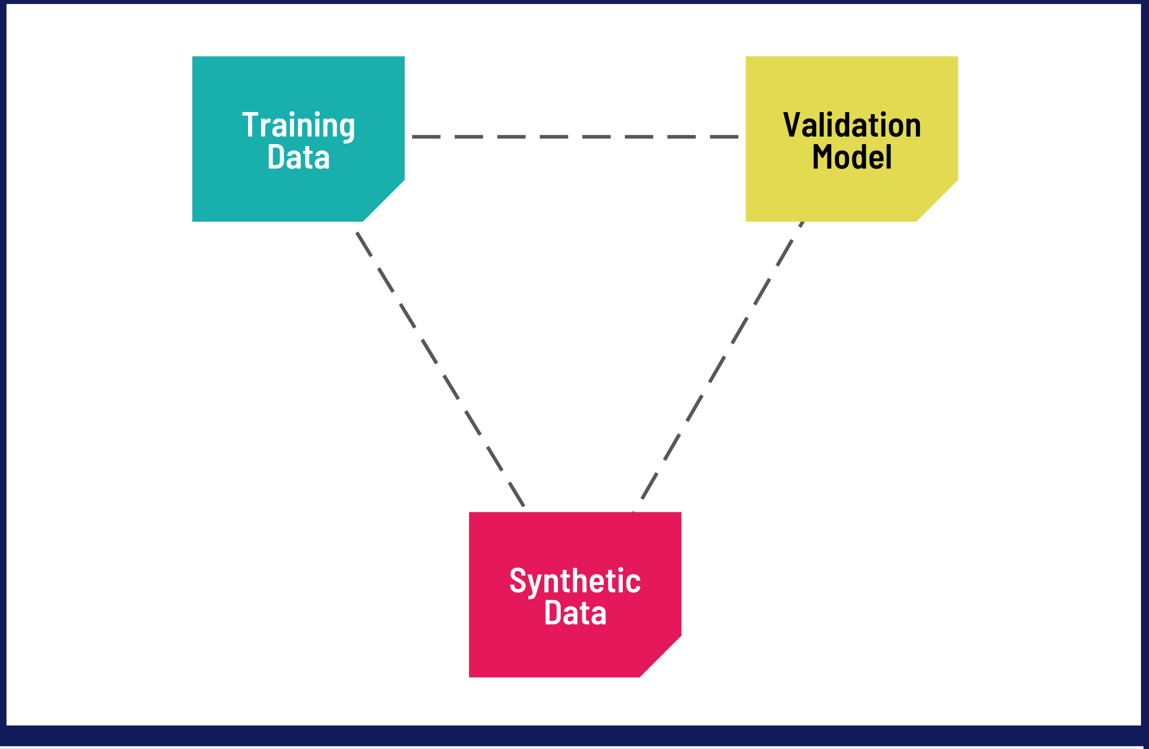
Technical Note:

How Ipsos does it

We apply pairwise distance analysis, nearest-neighbour redundancy checks, and coverage ratio metrics to identify how much new ground is being covered.

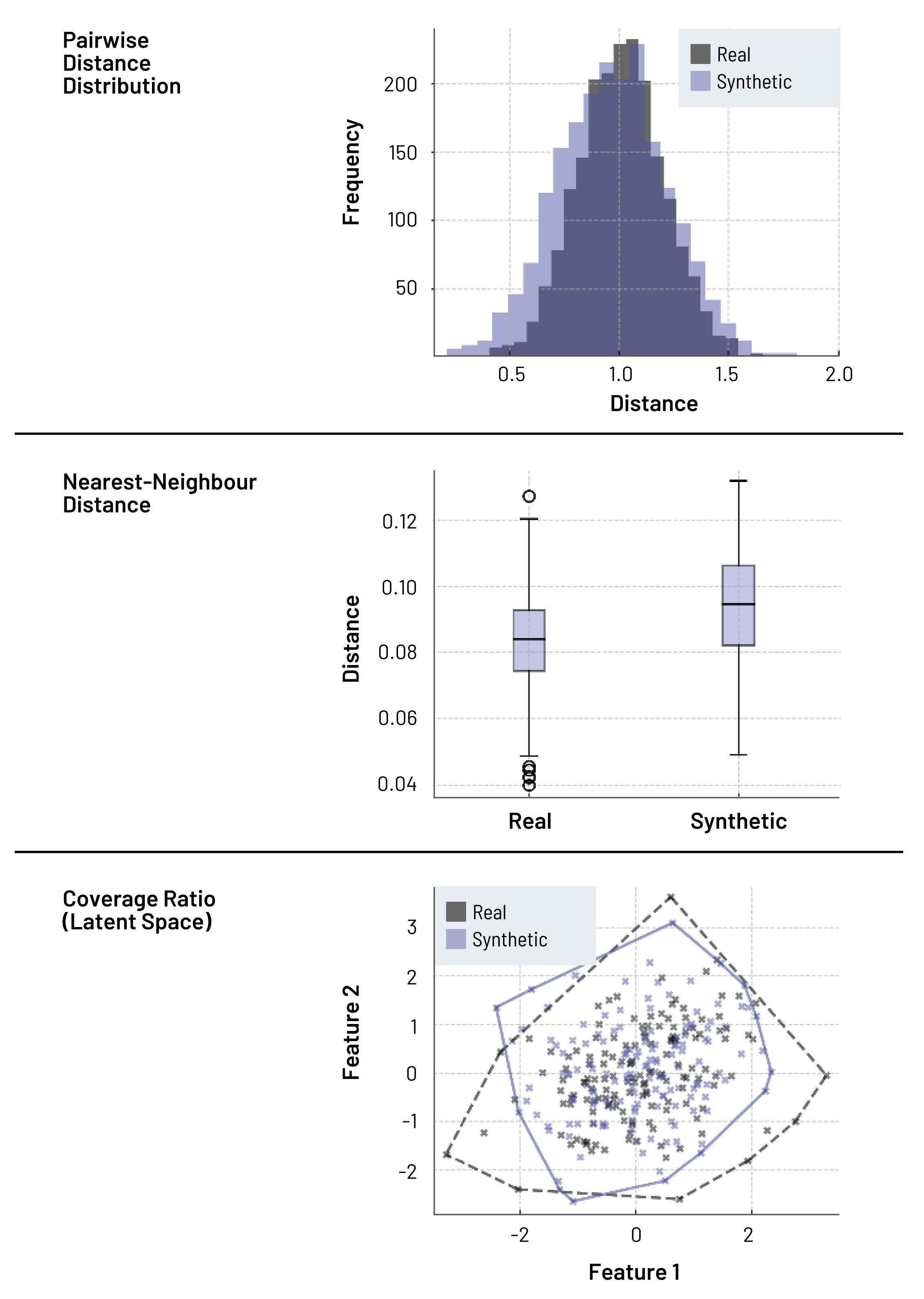
Diversity is quantified through entropy measures and latent-space dispersion to ensure the model expands, not repeats, reality.

Figure 10: Synthetic data that is equally likely to be closer to the training data or validation data suggests it is novel. That is to say, it looks equally like and dislike the data it was trained on and unseen data – so it is bringing new information.



Source: Ipsos

Figure 11: Diversity diagnostics for synthetic data.



Source: Ipsos

Expert Validation – does it make sense to experts?

Even the most elegant model must pass a human plausibility test. We ask whether subject-matter experts find the data and resulting insights credible, ethical, and feasible reality ensures that synthetic data doesn't just satisfy algorithms – it satisfies common sense and lived experience.

This final stage ensures our synthetic data pass the real-world test. Hyper-realistic pink cats might pass all numerical and statistical checks yet still be flagged by human experts during validation. To prevent such implausible cases from arising in the first place, domain experts may also need to be involved from the outset, helping to define the structural and logical constraints that guide the algorithms during generation. Their expertise may also be applied again at the review stage, verifying that outputs remain realistic and contextually credible. This dual layer of potential expert input – both preventive and evaluative – helps ensure that synthetic data faithfully reflect plausible real-world patterns while avoiding "pink-cat" anomalies.

At Ipsos, we believe in the unique synergy between Human Intelligence (HI) and Artificial Intelligence (AI) to drive innovation and deliver impactful, human-centric insights for our clients. These principles are embedded into all our AI solutions, including synthetic data boosting. With this combination of HI and AI, we provide insights that are safer, faster, and always grounded in the human context, ensuring relevance and value for our clients.

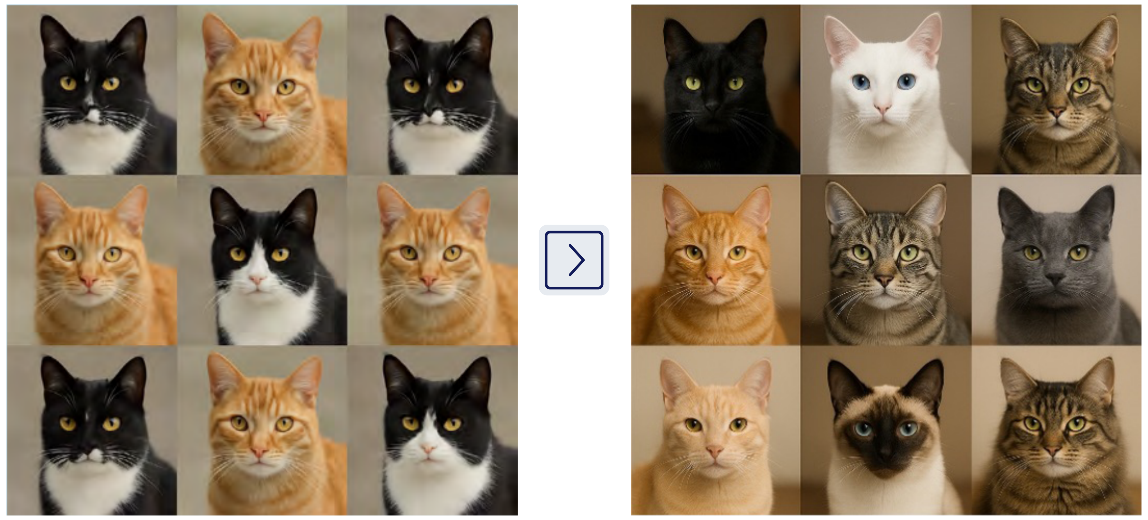
If we were to ask a machine, the images above could all seem like cats, and would behave statistically like cats, but you might hope an expert would spot that one or two are not actually very cat-like.

Technical Note:

How Ipsos does it

We combine expert panel validation with plausibility scoring, rule-based consistency checks, and ethical compliance review (privacy, fairness, and legal risk).

Synthetic datasets must pass both quantitative plausibility metrics and qualitative expert review to be considered "reality-consistent".



How much training data and sample is needed to reliably boost data?

This is one of the central questions in any data boosting project – while there's no single rule, the honest answer is: we can't know for sure until we test it. Our practical guideline is that reliable boosting typically requires a training dataset of at least 300–500 cases, depending on the project. Below this threshold, modeling error tends to exceeds sampling error, and in such cases, traditional weighting or imputation methods are usually the safer choice.

To boost data reliably, you first need a **large and sufficiently diverse training set** to prime the model. It must capture the essential latent variety present in the original data. For example, to train a model that synthesizes cats, your training data should include a broad and representative mix of cat types, colors, and features, ideally reflecting their **real-world proportions** where relevant. The goal is not to include every extant variety, but to ensure that the dataset spans the meaningful variability of the population. Otherwise, the model will simply reproduce the narrow patterns it has already seen.

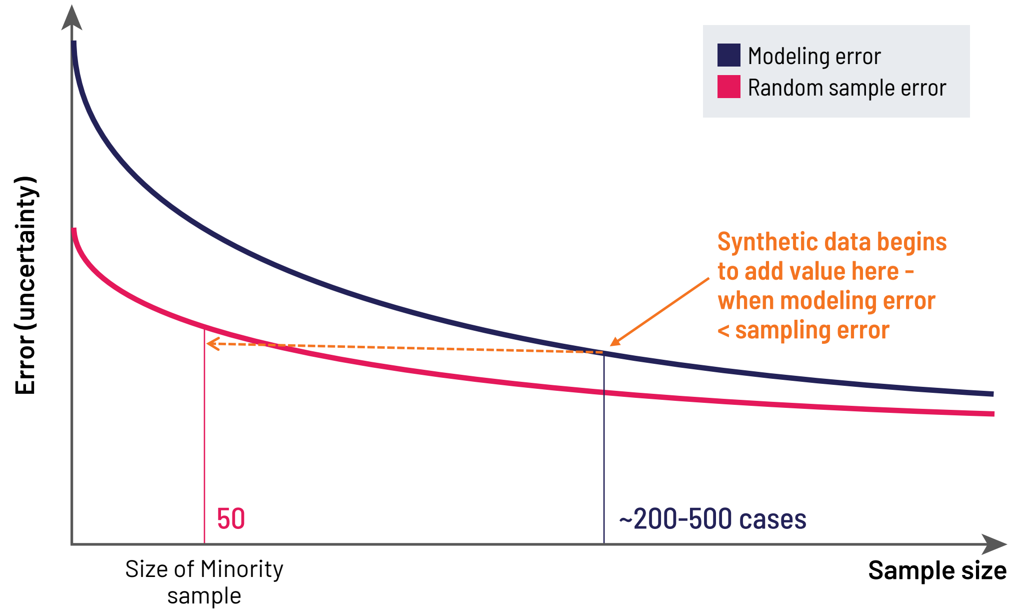
In short: you can't boost from nothing. Some datasets are naturally richer and more varied than others, and that determines how much can be reliably boosted.

We also have to contend with, and confront, the quality of data. If the raw data is polluted with a high level of noisy responses, this will reduce the starting sample size and has the potential to corrupt the modeling process.

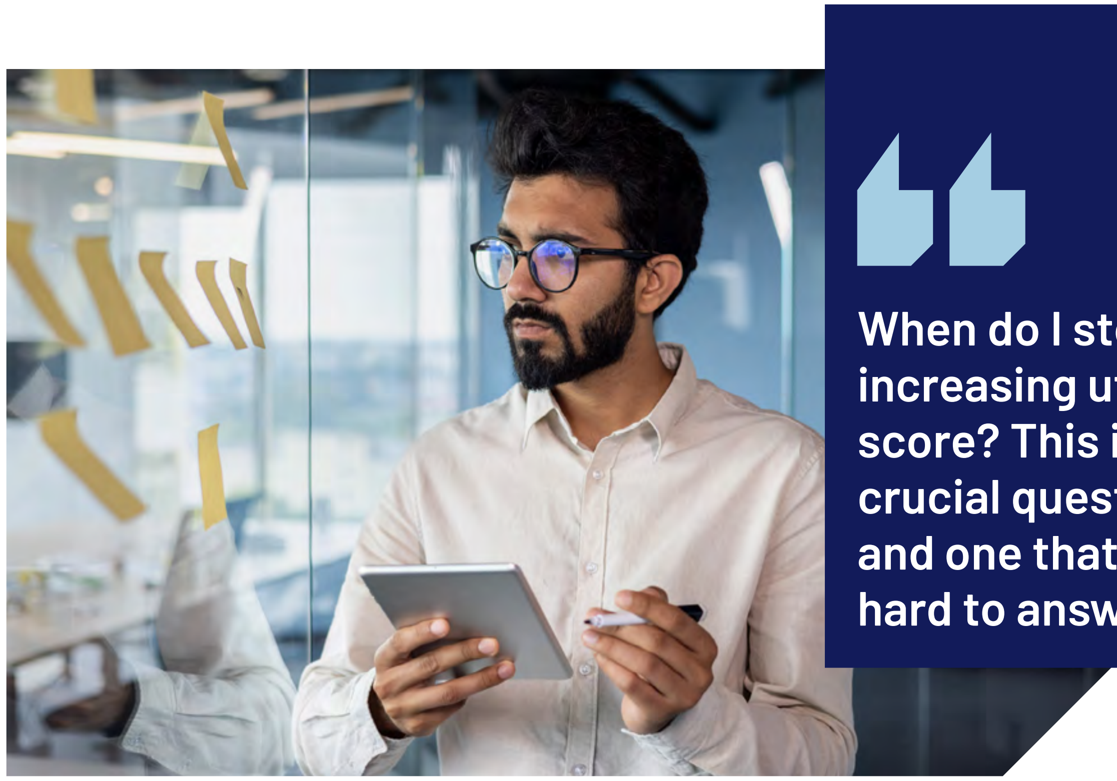
We must also consider both the **sampling error** inherent in the original data and the **modeling error** introduced during synthesis — the two compound each other. For instance, training a model on a sample of only 50 (with an inherent random sampling error of roughly ±15%) can easily add modeling error several times greater. With so few real observations to learn from, the synthetic results may appear plausible but remain statistically fragile and unreliable.

However, with a larger body of prior data, the picture changes. Imagine a tracker with 5,000 total respondents, including some from the subgroup we now wish to boost. When a new wave of 500 is collected and only 50 fall into that subgroup, a model trained on the earlier 5,000 can draw on those **previous observations** – capturing both general behavior and subgroup-specific patterns. In doing so, it effectively incorporates **historical information** that reduces uncertainty and improves precision beyond what conventional **RIM weighting** can achieve. Whereas weighting adjusts only marginal totals, synthetic boosting leverages **multivariate relationships** to generate additional, internally consistent cases. In this situation, the modeled results may have uncertainty lower than ±15%, meaning synthetic boosting genuinely adds value beyond traditional weighting approaches.

Figure 12: *Illustrative only.* How much training data is needed?



Source: Ipsos



> When do I stop increasing utility score? This is a crucial question and one that is very hard to answer.

In this purely illustrative example, random sampling error (orange) and modeling error (blue) both decrease, but at different rates. In this visual example, when the training data reach roughly 200–500 cases, the combined final error – the total uncertainty from both sources – drops below what would be expected from the unboosted random sample alone. Beyond this point, synthetic boosting provides a genuine reduction in overall error and greater analytical stability.

This threshold point isn't fixed, and it is important to add, not always reachable, the above example is purely illustrative. The crossover where synthetic data outperforms random error depends on three main factors:

01 **Underlying variance and required confidence:** Sampling error depends on both sample size and variability. For highly variable measures, or where tight precision is required (e.g. ±5%), larger real samples are needed before modeling becomes the more reliable option.

02 **How modelable the outcome is:** Some outcomes are easier to predict because they are strongly linked to other variables. Example: predicting lipstick use when you already know gender and age – high signal, low model error. Counterexample: predicting dog ownership from the same variables – weak signal, high model error. The stronger the underlying signal, the sooner model error falls below sampling error.

03 **How much noise is in the data:** Survey data can contain response noise from careless, inconsistent, or fabricated answers that blur real patterns. If the training data include too many such cases, the model ends up learning noise rather than structure.

At Ipsos, we use our **SURE framework** to assess how modelable the data is and to decide which projects are suitable for sample boosting.

How much can you boost by?

This is an area that our synthetic data research unit has investigated in detail by adding more and more synthetic data and seeing at what point does it stop adding value. That is to say, when do I stop increasing utility score? This is a crucial question and one that is very hard to answer. However, our work on our utility metrics provide us with a rigorous, mathematical way of understanding the limit.

Let's use this example:

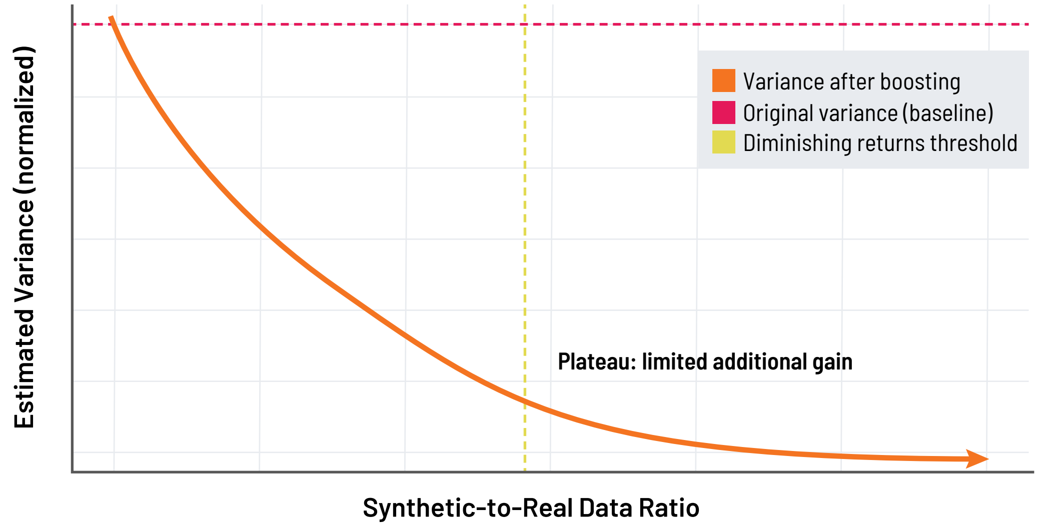
Statistical testing on synthetic data & effective sample size:

Suppose we begin with 1,000 real observations and generate an additional 500 synthetic ones. At first glance it may seem that our sample size for statistical testing is now 1,500. Not quite. Synthetic records violate the assumption of *independent, equiprobable sampling* that underpins classical statistical tests. They are not fully independent, each stems from a model trained on the original data and the model itself may not reproduce the underlying population distribution perfectly.

The result is that the *effective sample size* lies somewhere between the 1,000 original cases and the 1,500 total after boosting. This is precisely what the methodologies described in our Utility section of the SURE framework do. Through emperical testing we achieve a robust framework for adapting synthetic data to statistical testing. This framework allows us to avoid the false confidence that arises from counting synthetic data as real data.

Figure 13: ***Illustrative only.*** **Variance reduction as synthetic data are added.**

Initially, boosting expands analytical precision (lower variance), but gains taper off beyond a certain synthetic-to-real ratio. The plateau marks where additional synthetic data contribute little new information – the point of diminishing returns.



Source: Ipsos

Building on first principles, Ipsos has developed and validated a formal method to identify the plateau point at which synthetic data no longer contribute additional information. The approach, reviewed by independent statisticians and verified through empirical simulations, enables the rigorous demonstration of the nature of diminishing returns on utility when over-creating synthetic data.

The underlying important point to understand is that you cannot apply traditional statistical validation protocols with boosted sample, the dangers of naively statistical testing on synthetic data as if it were real are considerable.

> The underlying important point to understand is that the dangers of naively statistical testing on synthetic data as if it were real are considerable.

Summary of Ipsos' boosting approach

These four stages outline the **operational workflow** through which the SURE framework is applied in practice. They are not a separate model, but the practical steps that translate SURE's principles of *Statistical integrity, Utility, Rarity & Novelty,* and *Expert Validation* into action:

01 **Data appraisal – is the data synthesizable?** Assessing data suitability, quality, and representativeness before modeling.

02 **Data preparation – cleaning, alignment, optimization.** Standardizing formats, resolving inconsistencies, and ensuring the data are model-ready.

03 **Data modeling and generation.** Applying diffusion-based synthesis and boosting algorithms consistent with SURE standards.

04 **Data validation and integrity checking.** Testing the synthetic outputs against SURE's Fidelity, Utility, and Risk criteria to confirm robustness.

Together these steps form the practical implementation pathway for SURE, ensuring methodological consistency rather than introducing a new framework.

Frameworks and evidence - not promises or predictions

The market often features bold claims about synthetic data performance, such as "boost by 3-5x" or "reduce margins of error by 10-30%". We have to be clear that these are empty claims as these cannot be asserted a priori. Ipsos focuses instead on validation and testing to demonstrate true value empirically.

Let's take a simple analogy.

Imagine someone preparing for an exam. They're handed some new study material from an unknown source. It might be accurate and useful – or full of errors, **biases, or only partial coverage** of the syllabus. The student knows the subject of the exam but not the exact questions. So, can they really predict in advance how much this mystery material will improve their score? Of course not. To claim, "studying this unseen material for an unknown amount of time will raise my grade on an unknown exam by 10%", would sound absurd.

The same logic applies to synthetic data. Before testing, it isn't possible to know with certainty how much sample to generate and whether the generated sample will enhance analytical accuracy or model performance at all. In some cases, the synthetic data may even **reduce accuracy (a kind of "negative utility")** if the model learns misleading relationships or overfits to biased inputs. The extent of improvement depends on the quality of the underlying data, the modeling approach, and the analytical context.

What we can do is assure our clients that we approach the task with care and transparency, using a reliable framework to test assumptions and assess model quality. Rather than making promises about outcomes, we focus on **measuring** them, giving clients clear evidence of when synthetic data genuinely adds value and when it does not.

> Rather than making promises about outcomes, we focus on measuring them, giving clients clear evidence of when synthetic data genuinely adds value and when it does not.

In conclusion

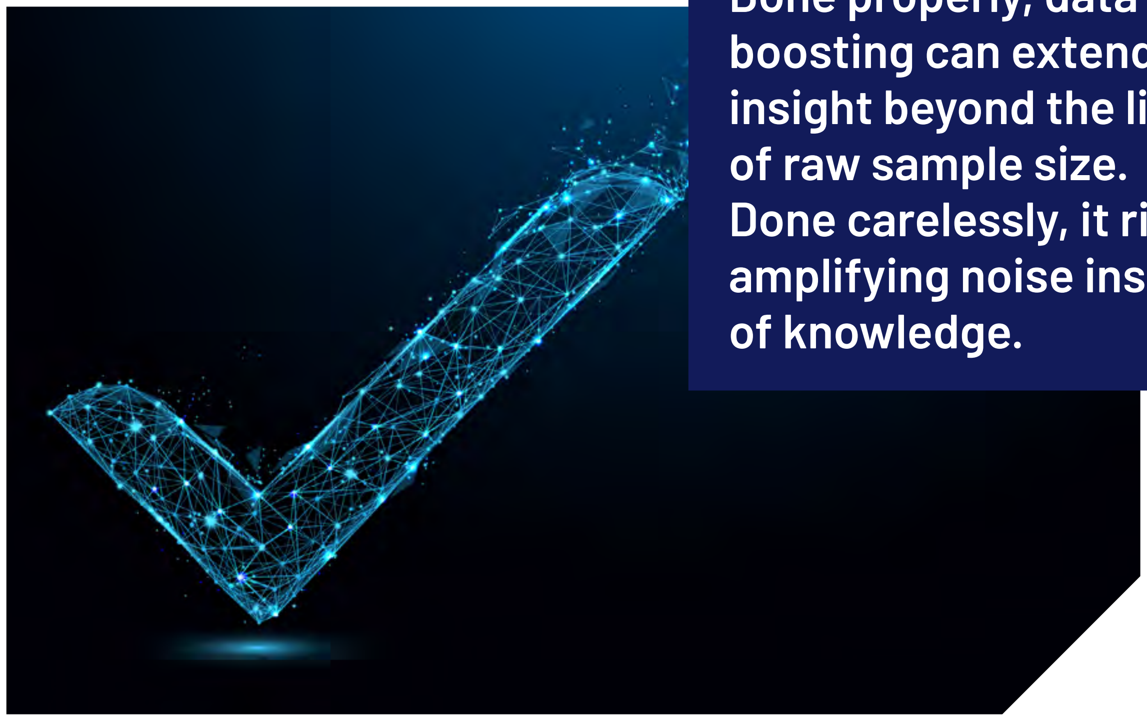
Modern tools can produce outputs that appear statistically sound, but without careful design, validation, and governance, results can be misleading.

The starting point is always the same: do we have enough high-quality real data to boost from? From there, our experts apply proven frameworks and diagnostic tests to ensure synthetic data genuinely adds value — not just volume.

Just as with segmentation, weighting, or sample design, successful data boosting depends on methodological rigor, domain knowledge, and expert judgment. Done properly, it can extend insight beyond the limits of raw sample size. Done carelessly, it risks amplifying noise instead of knowledge.

If you want to learn how your research projects can benefit from reliable synthetic data, please reach out to your Ipsos contact.

> Done properly, data boosting can extend insight beyond the limits of raw sample size. Done carelessly, it risks amplifying noise instead of knowledge.



Endnotes

1 Guidi, M., Hubert, B., & Sava, C. Synthetic Data: From Hype to Reality – A Guide to Responsible Adoption. Ipsos (2025).

2 Five Topics of Discussion to Help Buyers of Augmented Data. ESOMAR (2025).

3 Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. arXiv:2304.13023 [cs.AI] (2023).

4 Dhariwal, P., & Nichol, A. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG] (2021).

The cat visuals throughout this paper were created using Ipsos Facto - Ipsos' own generative AI-powered platform.

Further Reading

JANUARY 2026

SYNTHETIC DATA BOOSTING

Unlock the transformative potential of data augmentation

AUTHORS

David Priestley
Head of Data Science,
Synthetic Data, Ipsos

Maciek Ozorowski
Head of AI Transformation,
Ipsos

Jon Puleston
IIS Chief Methodologist,
Ipsos

The **IPSOS VIEWS** white papers are produced by the **Ipsos Knowledge Centre.**

www.ipsos.com
@Ipsos