KS&R

DECISION SCIENCES & INNOVATION

A KS&R TECHNICAL WHITEPAPER

The ECHO Index

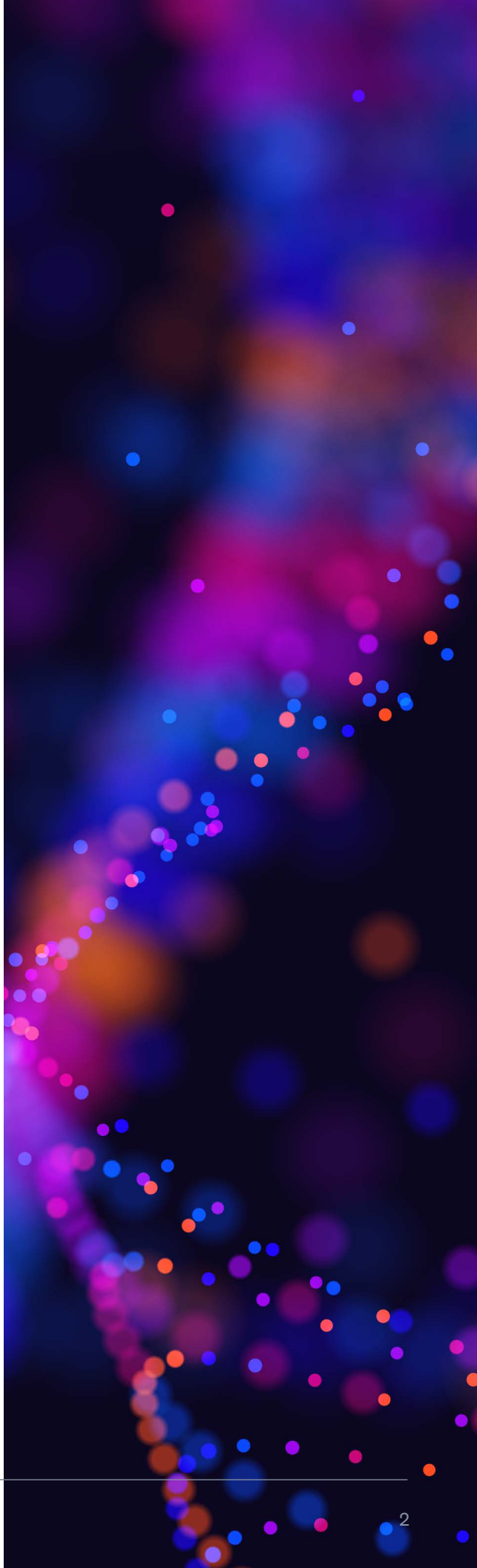
A rigorous, reproducible framework for validating synthetic data in market research

AUTHOR

Ben Cortese, PhD

SVP, Decision Sciences & Innovation

Contents



THE ARGUMENT IN BRIEF

Synthetic data has exploded in popularity, backed by clever, yet unjustified marketing claims, and the "trust us" model from AI-native platforms. Governance significantly lags usage, as evaluation is underdeveloped and difficult.

The ECHO Index closes that gap. It grounds every judgment in real human data, combines statistical and machine-learning evidence into a single interpretable score, and tells you not just *whether* a synthetic dataset is good enough, but exactly *where* it breaks.

The conversation around synthetic data has gotten ahead of the evidence.

Vendors can show that marginal distributions look close to the real world, reproduce a mean in a vacuum, and generate plausible respondents in seconds behind a polished interface. None of this is sufficient to support the claim that synthetic data leads to valid research conclusions. What the industry has lacked is a rigorous, transparent mechanism to decide when synthetic data is useful, when it is misleading, and what specific failure modes keep it from supporting real decisions.

This paper introduces KS&R's ECHO Index, an evaluation framework for synthetic data in market research. The focus is on the most defensible use case of **sample boosting**, the generation of additional rows of survey data for questions that already exist.

Through the ECHO Index, one can see that even in this well-established application, many synthetic models fail to align with real data when held to a proper standard.

This paper is intended for the people who build, buy, and stress-test these models: synthetic data providers, data scientists on insights teams, and the client- and agency-side researchers who need to know what a similarity score actually means before they implement it into their workflows.

A vocabulary, because evaluation depends on it

Synthetic data spans a spectrum from quantitative to qualitative, and the right way to evaluate it depends entirely on what is being generated, and more importantly, why. On the quantitative end sits **imputation** and **sample boosting**. Imputation, or filling in missing values in a data set, is a mature, well-understood problem with a long statistical history. Sample boosting extends it by applying similar techniques to solve a new problem. Instead of completing partially observed records, a model trained on real reference data generates net-new records with similar properties to the original data. To put it cleanly, imputation estimates missing entries conditional on observed rows, while sample boosting estimates an entire joint distribution. Both applications have historically relied on machine learning rather than generative AI, with sample boosting regarded as the most reliable form of synthetic generation primarily because it is rooted in established techniques.

On the qualitative end are chatbots, **digital twins** (one-to-one models representing an individual, system, or process), and **synthetic personas** (aggregate generators that simulate the collective voice of an audience). Generative AI has blurred the line between qualitative and quantitative, but these approaches share a common challenge: when a model simulates answers to questions it was never trained on, there is no real-world data to check them against.

Two primary modes of tabular synthetic generation

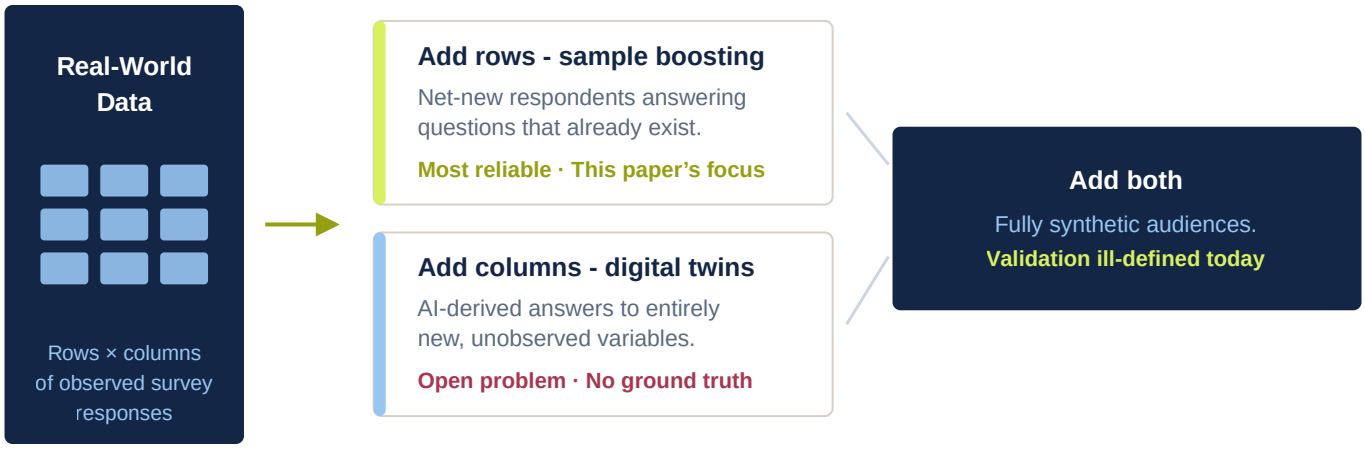


Figure 1. From real-world data, a model may generate additional rows (net-new respondents on existing questions), additional columns (AI-derived variables), or both. Row generation is the most defensible starting point as it can be viewed as an extension of imputation, a well-established technique. KS&R's ECHO Index begins here.

With these definitions in place, one principle holds true across use cases. Generating synthetic data that merely resembles the real data is not a sufficient standard. Synthetic data can look realistic but may carry artifacts that impact downstream analysis. Marginal distributions can be preserved while the relationships between variables are distorted. In market research, where the goal is rarely the dataset itself but the decisions it informs, that distinction is everything.

WHY START WITH SAMPLE BOOSTING

Row generation is both the most reliable form of synthetic data and the most amenable to rigorous statistical evaluation. While not entirely new, it is still an unsolved problem which makes evaluation of sample boosting the right place to start. Column generation and synthetic audiences show promise, however, rigorous evaluation of them sits on the roadmap, not in this paper.

02 THE PROBLEM WITH EVALUATION

The promise gap and why evaluation is hard

The promise gap is the misalignment between what synthetic data is marketed to do and what it actually delivers when applied to real research. It exists because synthetic data is easy to sell and hard to govern.

A provider can demonstrate that marginal distributions look close to the real world by focusing only on univariate comparisons. AI-native platforms can build plausible respondents in seconds, leaving validation as an afterthought. Neither of those approaches demonstrate that leveraging synthetic data actually leads to valid conclusions.

> The industry needs a **rigorous, consistent evaluation framework** for synthetic data.
>
> BEN CORTESE, PHD, SVP, KS&R

Three recurring challenges in how the industry currently evaluates synthetic data continue to fuel the promise gap.

Challenge One: Reliance on a Single Metric

A single divergence score, correlation, or distributional-similarity score is useful, but limiting evaluation to one metric alone is not validation. This narrow approach leaves ample room to cherry-pick the metric, overfit to "beat" a specific criterion, or hand pick a subset of questions that "proves" a model works. With limited transparency and little consistency across methods and vendors, persuasive claims become difficult to verify.

Challenge Two: Statistical Similarity Hides Artifacts

Statistical similarity can be preserved while a synthetic model leaves behind artifacts that only a classification algorithm can detect. When present, those artifacts introduce confounding information into any downstream analysis. The most reliable evaluation requires a machine-learning component that answers the question, "Can a classifier tell the synthetic data apart from the real data it was trained on?"

Challenge Three: Baseline Ambiguity

Scores are rarely reported against a reference point. Suppose a similarity of 0.93 is recorded and deemed "fit for use." What comparison justifies this interpretation? Without a baseline, one cannot say whether that number reflects genuine alignment, and furthermore, cannot directly compare two synthetic models.

This is the failure at the heart of most synthetic data evaluation, and it is the one the ECHO Index is built to fix.

These challenges point to what an evaluation framework must do. It should ground its judgments in real data; combine multiple complementary criteria rather than over-indexing on one; and enable genuine apples-to-apples comparison across generation mechanisms. Beyond that, it should not only flag *where* synthetic data fails to align with reality but point toward *how* it can be improved.

Arguably the most important aspect of evaluation is interpretability. It should stay interpretable for the mixed technical and non-technical audiences who leverage synthetic data to make decisions.

These requirements motivate the ECHO Index.

- **Ground every judgment in real data.** Synthetic data should be measured against real human observations, never scored on its own internal coherence.
- **Combine complementary criteria.** No single statistic is sufficient. Statistical and machine-learning evidence provide a holistic viewpoint of performance.
- **Make comparison apples-to-apples.** A common, per-dataset reference point lets any number of models be ranked on the same terms.
- **Stay interpretable, and diagnostic.** One headline number for stakeholders, with a drill-down to the individual question for the analyst iterating on the model.

The ECHO Index: Four classes of evidence

An *echo* starts with something real. A human creates a sound which travels and returns as something recognizable, but never an exact duplicate. Synthetic models work the same way. They are meant to echo real behaviors, preferences, and decisions. The ECHO Index quantifies how much confidence we can place in that echo.

The framework directly answers a seemingly simple question, “Can we trust data generated from a synthetic model, and if so, where, when, and how?” It groups evaluation into four distinct classes of evidence. Two are mature, scalable, and central to current evaluation, while two are essential to market research but remain harder to scale.

The four classes of evaluation evidence

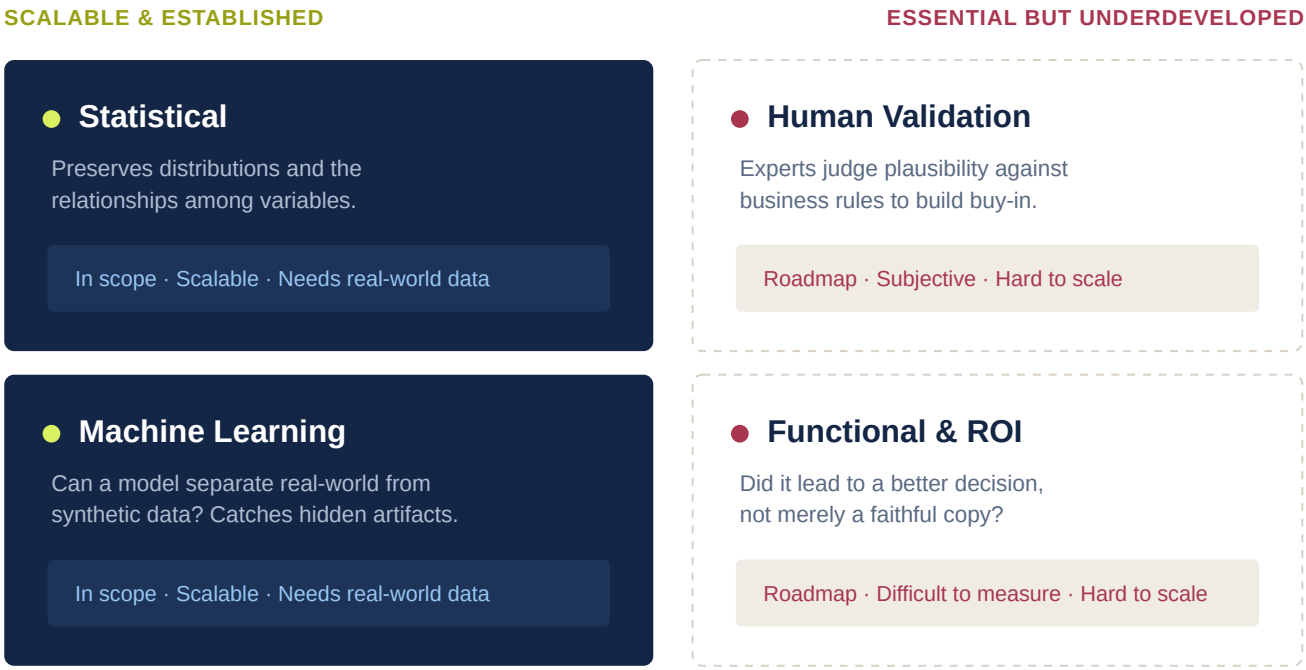


Figure 2. The “Statistical” and “Machine Learning” classes are scalable and interpretable but require real-world data for validation, while the “Human Validation” and “Functional & ROI” classes are essential to market research yet remain underdeveloped, subjective, and difficult to scale. This paper operationalizes the first two.

The two left-hand classes of “Statistical” and “Machine Learning” are well established, scalable, and straightforward to interpret, but they require real-world data for validation. These enable evaluation of *fidelity*, that is, the agreement between synthetic and real data. This is not to be confused with *utility*, the performance of synthetic data on downstream analyses, which is covered by the two right-hand classes.

Measurement of utility is critically important, particularly for market research, but is far less scalable and more subjective. Functional and ROI evaluation is especially difficult, where the downstream decision process is complex, and research functions as only one of the many inputs to the end result. Human validation builds trust but is a slow, manual process that lacks a consistent framework. For these reasons, the current framework focuses on fidelity, with the explicit acknowledgment that meaningful work remains to address utility.

Defining the index

KS&R’s ECHO Index is a combination of a statistical score and a machine-learning score. It is critical that both contribute equally as they measure separate, yet critical, success criteria. Of note, the raw statistical and machine-learning scores have very different centers and shapes, and if implemented as-is, one would dominate the overall index. We therefore introduce a **score-alignment scalar** ω that rescales the machine-learning score so the two contribute equally.

$$EI = \beta_S + \omega\beta_{ML}$$

where β_S is the statistical score, β_{ML} is the machine-learning score, and ω is the score-alignment scalar.

The scalar ω is computed from the baseline as the ratio of the two baseline means (described in Section 6). By design, lower ECHO Index scores indicate closer alignment to real data and higher-quality synthetic data overall.

04 COMPONENT ONE

The statistical score

Does the synthetic data behave like the real data, both for individual questions and relationships between questions?

Three question types are in scope for the statistical score, covering a majority of questions in a typical survey:

- **Likert and other quasi-numeric** questions, treated as continuous,
- **Single-choice questions**, where one outcome is selected from a fixed set, and
- **Multi-select questions**, where more than one may be selected.

Open-ended, ranking, and other less common types are intentionally omitted in this version and planned for later iterations.

Every question is scored on two complementary criteria. **Distribution preservation** measures how well the univariate distribution of the synthetic data aligns with the real data. **Relationship and dependency** measures how well the pairwise relationships between variables are preserved. Both univariate and multivariate relationships carry downstream implications, hence the necessity for two types of evaluation.

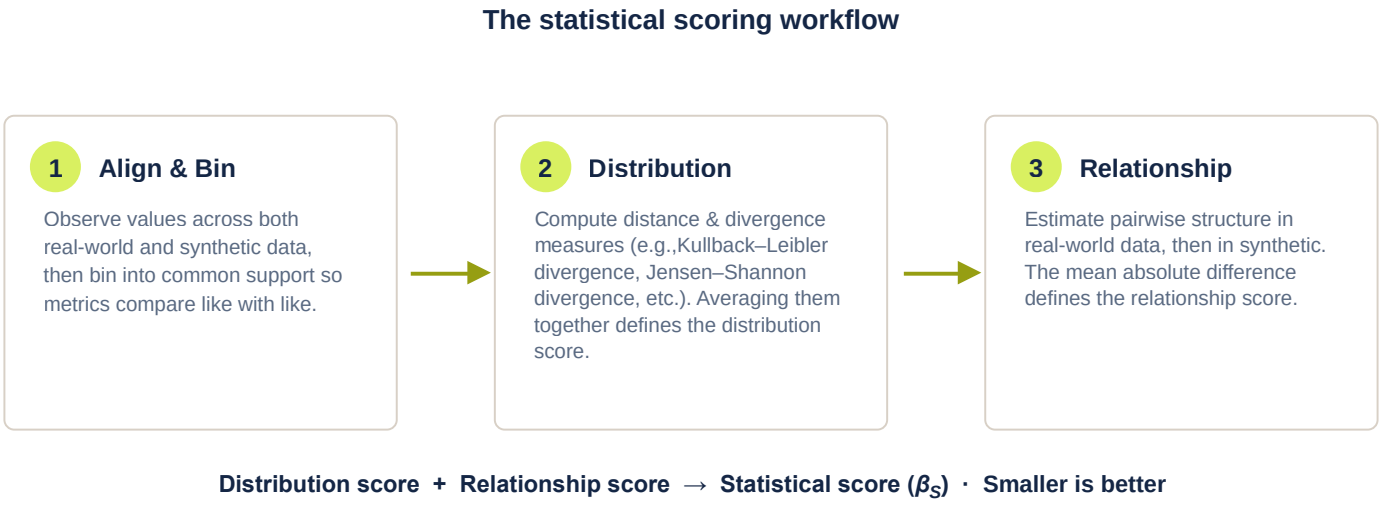


Figure 3. An example of the statistical scoring workflow for numeric data. Responses are aligned and binned as needed, scored for distributional similarity and relationship preservation, and combined into a single statistics score. Single-select questions swap Pearson correlation for Cramér's V, while multi-select questions add a check on the number of items selected.

TECHNICAL NOTE

HOW THE STATISTICAL SCORE IS COMPUTED

DISTRIBUTION PRESERVATION

For each question, alignment between the synthetic and real distributions is measured using several distance and divergence metrics including **Kullback–Leibler (KL) divergence, Jensen–Shannon divergence,** and **total variation distance (TVD)**. Averaging the raw values together defines the distribution score. Note that no normalization occurs during this step despite KL divergence presenting as an unbounded metric with both Jenson-Shannon and TVD bounded by 1, as the primary goal of the ECHO Index is to identify when synthetic data is not fit for use.

RELATIONSHIP AND DEPENDENCY

Pairwise relationships are estimated from the real data, then again via the synthetic data. The mean absolute difference between these estimates defines the relationship score. The metric depends on the question type: **Pearson correlation** for numeric questions, **Cramér's V** for categorical questions, and the **phi coefficient** for binary questions. Mixed-type relationships are reserved for a later iteration.

THE MULTI-SELECT TRAP

Multi-selects need one extra step, as it is known that synthetic models engage with multi-selects differently than humans, choosing either more or fewer options. Rather than treating a "select all that apply" as a single distribution, it is broken down into a binary yes-or-no distribution for each option. The divergence calculation is repeated on the distribution of the number of items selected.

05 COMPONENT TWO

The machine-learning score

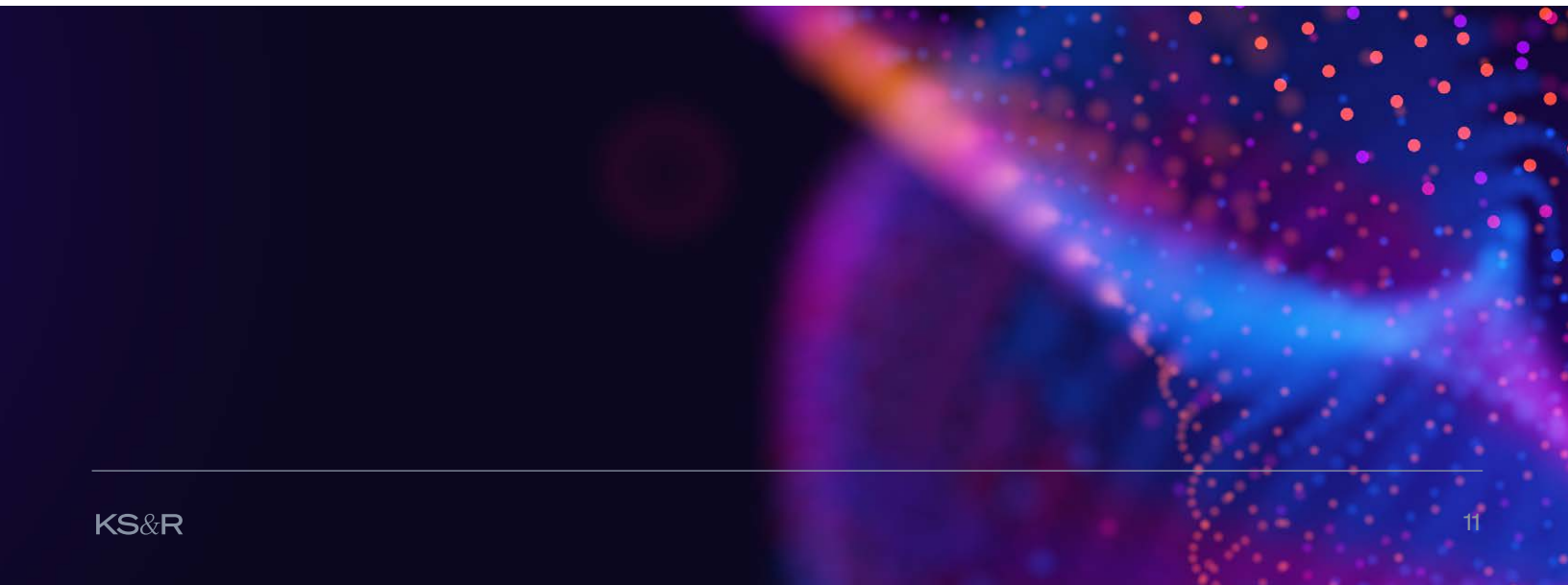
If a classifier can reliably tell synthetic from real, the model has left behind artifacts that can mislead downstream analytics and reporting.

The machine-learning score measures how difficult it is for classifiers to separate real world data from synthetic data after generation. This is the essential complement to statistics-only evaluation as it identifies any systematic artifacts that distributional tests miss. Several classifiers are used, spanning historically successful and more modern techniques such as random forest, logistic regression, and XGBoost. They are chosen because they handle mixed variable types, scale with minimal tuning, and produce consistent results without the need for customization.

A ROADMAP FOR SYNTHETIC MODEL ENHANCEMENT

While machine learning scoring leverages the entire dataset for evaluation, importance scores are produced from each model indicating which variables are driving a failed synthetic run. This, combined with question level statistical scoring, creates a roadmap for iteration and enhancement of future synthetic data generation. This diagnostic capability is what makes the ECHO Index more than a verdict, turning a failed run into a prioritized path toward higher-fidelity synthetic data.

The interpretation runs opposite to the usual reading of classifier performance. Normally, high accuracy is the goal, however, here high accuracy means the classifier can tell real from synthetic, which is an indicator of lower-quality synthetic data. A perfect classifier is actually a red flag.



TECHNICAL NOTE

READING THE CLASSIFIER THE RIGHT WAY

Performance is summarized with several success metrics:

- ROC AUC
- F1 score
- Balanced accuracy

Together, these capture both predictive accuracy and any bias introduced. Performance scores are averaged across **ten-fold cross-validation.** No true holdout set is used: a conventional train-test split does not align with this use case, where the goal is not to predict a future state but to detect whether the two samples are separable at all.

The scale typically runs from 0.5 to 1.0. A score near 0.5 means the classifier cannot distinguish real from synthetic, which is the desired outcome. A score near 1.0 means the synthetic data is trivially separable from the real data, and any model trained on the combined set will learn the artifact rather than the signal. This is the critical component that statistical similarity cannot replace.

06 THE KEY IDEA

The baseline: The idea that makes it work

A score of 0.93 means nothing until you know how real data scores against itself. The ECHO Index builds that unique reference point for each dataset in a vacuum, and that is what grounds evaluation and comparison.

The novelty of the ECHO Index is anchoring to a baseline distribution. First, the index is estimated using the real-world data alone to build an understanding of what a model should achieve if it perfectly mimicked reality without exact duplication.

The procedure works as follows:

1. Partition the real dataset in two
2. Label one partition “synthetic” and the other “real”
3. Run the ECHO Index on the partitions
4. Repeat hundreds of times

The result is a baseline distribution of the ECHO Index and of its statistical and machine-learning components.

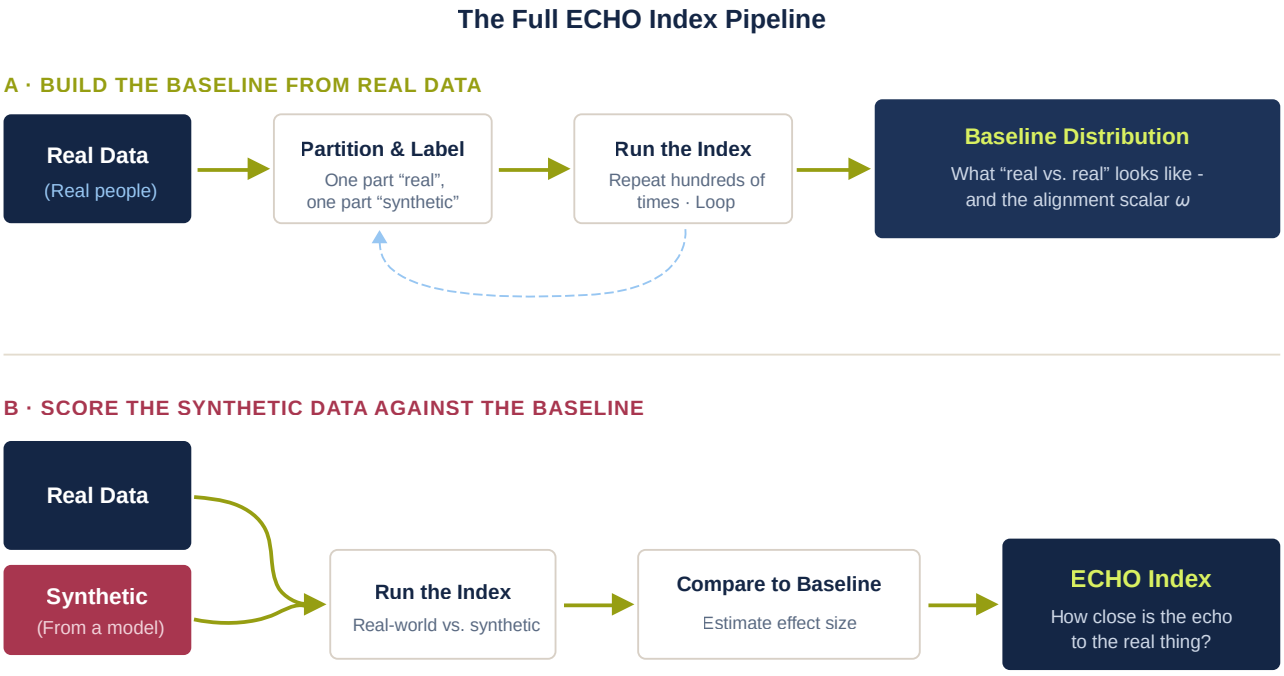


Figure 4. A baseline distribution is built by repeatedly partitioning the real-world data and labeling one partition “synthetic.” The alignment scalar ω is estimated from that baseline. Synthetic performance is then judged by how close its index sits to the baseline via effect size.

The baseline does two jobs at once. Its primary role is to provide reference data for the real-versus-synthetic evaluation, where “close” is determined through an **effect size** – the number of standard deviations the ECHO Index score for synthetic data is from the baseline mean score. Effect size provides a single value that is easy to communicate and interpret. Second, it supplies the alignment scalar ω, calculated as the ratio of the baseline mean of the statistical score to the baseline mean of the machine-learning score, which blends the two components so each contributes equally.

WHY THIS CHANGES THE GAME

Because the reference distribution is constructed from the dataset itself, the ECHO Index can score almost any standalone dataset rather than relying on rules of thumb about what a given distance metric "typically" means. It also yields effect sizes that make comparisons across many models genuinely apples-to-apples. This relativistic, per-dataset baseline is the single feature that most distinguishes the ECHO Index from conventional scoring.

07 BENEFITS OF THE ECHO INDEX

Four advantages over traditional scoring

It should come as no surprise that **alignment between synthetic and real data is data-dependent**. KS&R's ECHO Index presents a portable solution applicable to almost any standalone dataset

1. **The baseline grounds every evaluation.** It adds an interpretability layer that estimates the magnitude of similarity to real data, rather than leaving an analyst to infer what a raw distance typically means.

2. **Effect sizes make comparison apples-to-apples** across any number of generators, adding value beyond what single-score reporting offers.

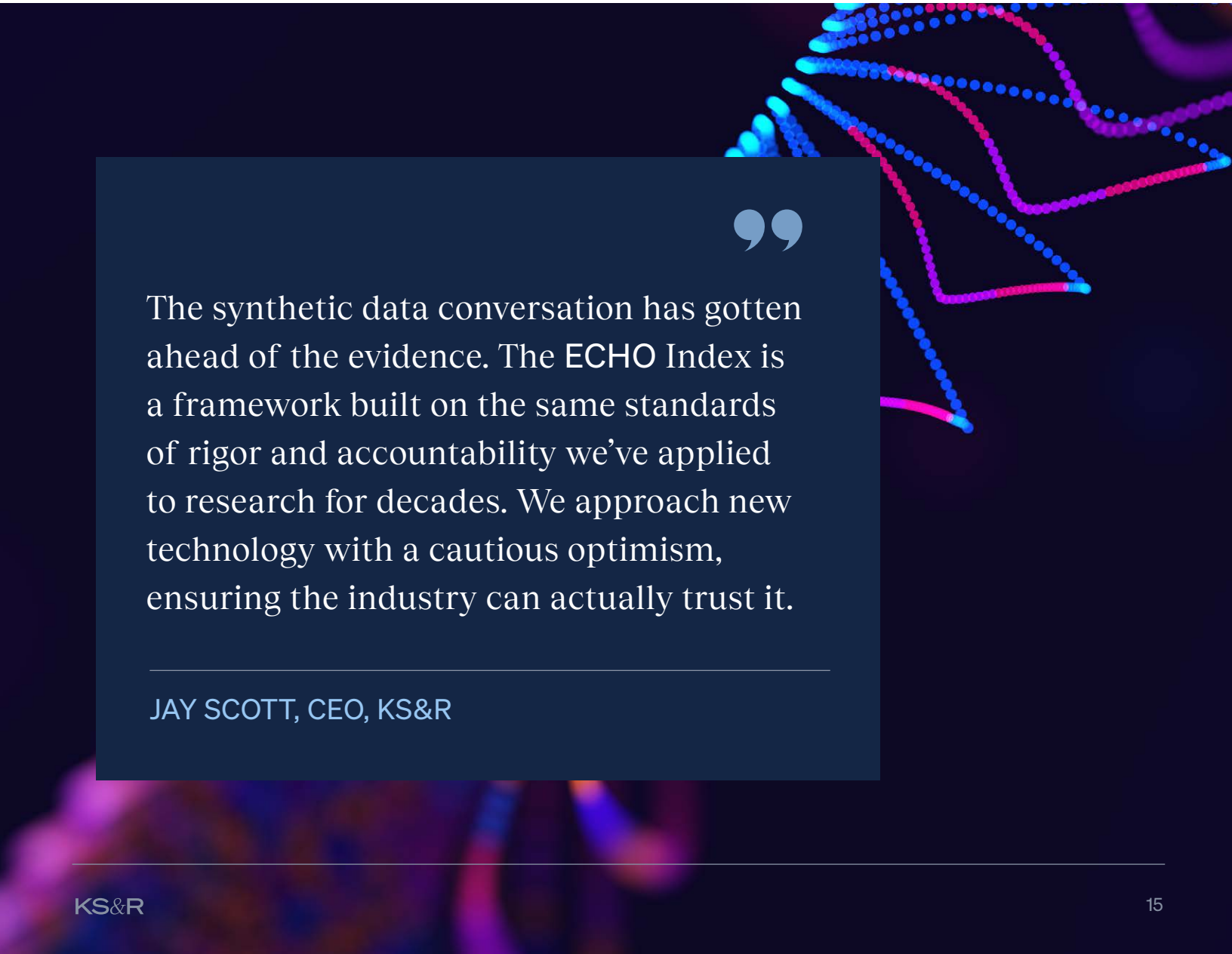
3. **Combining statistical and machine-learning criteria increases rigor.** The machine-learning component is not new on its own, but it has not been combined with statistical scoring this way before, and it guards against the cases where statistical similarity alone misleads.

4. **The index is well scoped for mixed audiences.** One headline number communicates to non-technical stakeholders, while the same framework drills down through every evaluation criterion, all the way to the individual questions driving a poor score.

Why this matters: Governance as infrastructure

The synthetic data debate has moved from “is it valid?” to “how do we govern it rigorously enough to use it with confidence?” That shift is the real story, and it is why an evaluation framework is the model the industry needs.

KS&R built the ECHO Index in response to the promise gap: the difference between marketing claims and realized performance. That gap is widened by inconsistent validation practices, selective use of high-performing metrics, and pressure to demonstrate that synthetic data works without fully understanding where it does and does not deliver. The fix is a rigorous, transparent evaluation framework that can quantify performance of synthetic data.



> The synthetic data conversation has gotten ahead of the evidence. The ECHO Index is a framework built on the same standards of rigor and accountability we’ve applied to research for decades. We approach new technology with a cautious optimism, ensuring the industry can actually trust it.

JAY SCOTT, CEO, KS&R

It is of note that KS&R is a research and consulting firm, not a platform vendor. During early efforts to investigate the utility of synthetic data, it became clear that governance infrastructure was, and still is, a blind spot in the maturity of synthetic data as a product. Platform vendors have yet to sufficiently solve this problem, hence the development of the ECHO Index by KS&R.

The ECHO Index provides a model framework as we move from questioning whether synthetic data is valid toward asking how to govern it rigorously enough to use with confidence.

Any supplier offering synthetic augmentation, synthetic respondents, or AI-generated personas should be able to produce a quantified similarity score against real human response data. Suppliers self-evaluating performance should no longer pass as sufficient for any buyer, and instead, buyers should demand transparency and third party governance. This is exactly what the ECHO Index was built to tackle by answering the question **"how similar is your synthetic output to real human response, and how do you know?"**

> **THE CONTROL POINT**
>
> As AI makes synthetic generation more accessible, it also highlights validation as critical path while the industry keeps inventing new ways to be confidently wrong. The advantage does not belong to whoever generates synthetic data fastest, but instead, it belongs to whoever can tell you, rigorously and reproducibly, whether that data is fit for purpose.

09 DISCUSSION & WHAT'S NEXT

An honest reframing and the future roadmap

The original goal was not to develop the ECHO Index, instead, it was to answer whether synthetic data can alleviate low feasibility and poor quality sample. During this investigation, it became clear that there was no consistent and transparent mechanism to evaluate synthetic data. Once the ECHO Index was developed and applied to synthetic datasets, it became apparent that producing usable synthetic data even for the tractable case of sample boosting (i.e., generating synthetic rows) is not always possible. The implication is not that synthetic data is hopeless, instead, it is that the industry should solve the comparatively "easy" problem of sample boosting before reaching for synthetic insights generated directly from a model.

A second implication concerns the cost of getting evaluation wrong. Traditional, statistics-only evaluation is flawed in ways that can manufacture false confidence. It gives vendors and practitioners agency to claim their methods work across a wide variety of use cases, and it enables cherry-picking the metrics that make a model look successful through limited transparency. The ECHO Index aligns disparate techniques into a single, transparent procedure to tell users rigorously and comparably whether the output is good enough, and when it is not, to provide a roadmap toward improvement.

Limitations we state plainly

- The framework depends on real-world reference data for benchmarking. No real data, no baseline.
- The Human Validation and Functional & ROI classes remain underdeveloped. ROI in particular is challenging because of a highly complex downstream decision making process, and human-in-the-loop review does not scale cleanly.
- The current scope excludes open-ended, ranking, and other less common question types, and mixed-type relationship measurement.
- Naively generated synthetic data can inherit or amplify biases present in the real data, so fairness-aware evaluation is an important extension of this work.

Above all, synthetic data is a tool, not the end result. Our role is not to manufacture synthetic inference but to use the tools at our disposal to inform better decisions. A rigorous, shared framework for synthetic quality is a prerequisite for doing that responsibly.

WHAT'S NEXT?

The ECHO Index lays the foundation for third party governance of synthetic data. The initial pass enables scoring and evaluation of sample boosting among common question types. There is room for advancing the index, expanding use cases and applications:

- Expand to enable evaluation of digital twin models
- Apply the index to open-text questions
- Develop more extensive relationship measurement beyond similar question types
- Evaluate synthetic providers, comparing performance across benchmark data

The ECHO Index

KS&R Technical White Paper · Decision Sciences & Innovation

Ben Cortese, PhD
SVP, Decision Sciences & Innovation
bcortese@ksrinc.com
linkedin.com/in/bcortese/

KS&R

Dynamic **Wisdom.** Better **Decisions.**

PO Box 2145, Syracuse, NY 13220 · ksrinc.com · hq@ksrinc.com
