MSI at the ARF SUMMARY

MSI / ARF 2025 Analytics & Forecasting

How AI Analytics and Synthetic Data Can Reimagine Consumer Research

Keith Smith, PhD (Managing Director, MSI) and Tracy Adams, PhD. (Senior Director, Research & Insights, ARF), report on their major takeaways from the recent MSI/ARF Analytics & Forecasting conference.

At the MSI/ARF Analytics & Forecasting conference, AI was framed not as a substitute for human research but as a force reshaping how consumers are represented and understood.

Day 1 opened with debates on AI-generated reviews, underscoring the blurred line between AI-enhanced expression and fabricated fakes. The day's panel on synthetic data pushed deeper, asking whether "truth" in research should be judged against noisy, biased human data or by the business outcomes synthetic models enable. Panelists warned about reproducibility, with results shifting across model versions, and about homogeneity, as synthetic respondents tend to regress to the mean. Still, synthetic approaches were seen as powerful accelerators of qualitative insights if findings are validated.

Digital twins became a focal point. Columbia's Twin-2K-500 dataset systematically compared humans and their twins across demographic, psychological, and behavioral measures. Twins proved hyper-rational and knowledgeable, but skewed liberal, optimistic, and pro-technology, often erasing the conflict and inconsistency central to real behavior. Virginia Tech's e-commerce demonstration showed how twins trained on Amazon review data could predict purchases and generate semantically faithful reviews, illustrating both their potential and their tendency to over-smooth human messiness.

Day 2 moved to applications. Harvard's Ayelet Israeli showed that LLMs can conduct conjoint studies consistent with economic theory, though novelty bias emerged. TU Munich introduced synthetic personalities replicating traits like extraversion and neuroticism, but consistently more open and progressive than real populations.

The debates revealed more than methodological tweaks: they pointed to a paradigm shift in how audiences are constructed. Just as online panels once transformed the industry, synthetic respondents and digital twins are setting the terms of a new consumer imaginary — one at once more rational, more progressive, and more uniform than lived reality.

“AI isn’t just reshaping analytics — it’s forcing us to rethink how insights are discovered, trusted, and translated into decisions.”. Tracy Adams, PhD., Sr. Director, Research & Insights, ARF

Building on these discussions, one theme that emerged was that synthetic data is no longer experimental. It is becoming central to how marketing professionals think about the future of research. Across two days of presentations, we saw not only rapid innovation but a growing awareness that synthetic data is not just a technical tool. It represents a shift in how we simulate and represent consumers, raising new questions about what we consider valid, accurate, and actionable.

Several common themes emerged. One was definitional. What exactly is synthetic data? Can it be treated like traditional data? Do core statistical concepts still apply, or are new frameworks needed? Even a basic measure like standard deviation may require rethinking.

Ground truth was another recurring concern. As traditional human-based data collection becomes more limited or flawed, especially in survey contexts, what should synthetic outputs be compared to in order to evaluate accuracy?

We also saw how much depends on the inputs and methods used. The way a question is framed, whether directly or indirectly, can significantly affect responses. The source data used to train a model, whether survey responses, behavioral data, or constructed personas, may lead to different outcomes. The volume of training data matters, too little can lead to underperformance, while too much may reduce accuracy. Biases in synthetic data were also observed, including those tied to demographic and attitudinal patterns, reminding us that these systems can reinforce existing distortions.

Synthetic data holds tremendous promise, but its impact depends on thoughtful validation and a clear understanding of its limitations. Its usefulness is often proportional to the stakes of the decision. As adoption accelerates, the real challenge is not just knowing when synthetic data works, but understanding how it shapes the insights we choose to act on.

“The real challenge with synthetic data isn’t just deciding where to use it. It’s learning how its limitations scale with the complexity and consequences of the decisions we’re trying to make.” Keith Smith, PhD, Managing Director, MSI