Grounded Synthetic:
Synthetic Data's Missing Foundation

How Real Behavioural Data from Device Native AI Unlocks the Full Potential of Synthetic Intelligence

DataSapien | White Paper | 2026

Contents

1. Executive Summary

Sound research and knowledge is the foundation of all good strategy and decision-making. The quality of every decision an organisation makes is ultimately constrained by the quality of the data and insight that informs it.

For marketing to be successful, research must be grounded in truth: in real behaviour, real preferences, and real context from real people. Declared intentions, historical averages and statistical approximations each have their place, but they are not truth. They are proxies for it. The closer an organisation can get to genuine behavioural truth, the better its decisions become.

We are now moving into an era where insights can become exponentially sharper and more timely. New approaches to data collection, processing and intelligence are making it possible to understand individuals in context, in the moment, at scale. Synthetic data is accelerating this shift by offering speed, cost efficiency and privacy-safe generation. But synthetic data has a structural vulnerability that is becoming increasingly visible: without a foundation of verified, real-world behavioural data, it inherits bias from its source, homogenises toward the mean, and systematically underrepresents the outliers that carry disproportionate commercial and safety value.

In March 2026, this tension surfaced publicly. JPMorgan Chase halted a $5.3 billion debt sale tied to Qualtrics' acquisition of Press Ganey Forsta, with investors citing AI disruption risk.

In the same week, in coverage of the Qualtrics X4 event in March 2026, Gartner analyst Michael Maziarka outlined five validation requirements CX leaders should demand from any synthetic data vendor (Refs. 2, 3): data sourcing and quality, comparative testing against first-party VoC data, conclusion similarity testing, hallucination checks and bias mitigation. Organisations that skip this validation layer risk building decisions on simulations that feel confident but are not grounded in reality.

This white paper introduces the concept of **Grounded Synthetic**: synthetic data that is anchored in a verified foundation of real behavioural data, generated on-device, from known individuals, within known populations. We argue that the debate should not be "real versus synthetic" but "grounded versus ungrounded," and that Device Native AI provides the architectural foundation that makes synthetic data trustworthy, auditable and fit for high-stakes decision-making.
This paper provides a detailed, honest comparison of both approaches across data quality, fidelity, privacy, compliance, cost, and governance factors. Both approaches have genuine strengths, and for most organisations the optimal strategy is to deploy them complementarily.

2. The Positive Case for Synthetic Data

Synthetic data has earned its place in the modern data stack. Before examining its limitations, it is important to acknowledge, fairly, where it delivers genuine value.

2.1 Rapid prototyping and hypothesis testing

Dimension	Synthetic Data
Speed	Millions of records generated in hours, not weeks
Use cases	"What if" scenario modelling, pricing sensitivity, segment simulation
Example	Qualtrics Business Outcome Simulator (announced X4, March 2026)
Limitation	Outputs are hypotheses to be validated, not decisions to be acted upon

2.2 Privacy-safe dataset generation

Dimension	Synthetic Data
PII exposure	No personally identifiable information in output datasets
Use cases	Cross-border data sharing, third-party vendor testing, open research
Regulatory benefit	Reduces GDPR/CCPA compliance burden for data sharing

Limitation	Provenance questions remain: regulators scrutinise whether synthetic data derived from personal data inherits its legal obligations

2.3 Cost efficiency and democratisation

Dimension	Synthetic Data
Cost comparison	Fraction of traditional panel costs for directional research
Industry claim	Qualtrics synthetic panels at half the cost of traditional panels
Accessibility	Lowers barrier to evidence-based decisions for smaller organisations
Limitation	Cost advantage applies to broad, directional questions; not to individual-level precision

2.4 Training data augmentation

Dimension	Synthetic Data
Proven domains	Computer vision, NLP, classification tasks
Value	Fills volume gap when real datasets are small
Limitation	Edge cases and outlier patterns underrepresented or absent

2.5 Test environment population

Dimension	Synthetic Data
Use case	QA testing, staging environments, system integration
Status	Solved problem; sensible default for development workflows

Key insight: Synthetic data works best when the task requires volume over fidelity, direction over precision, and patterns over individuals. The problems begin when it moves beyond these strengths.

3. Three Structural Constraints

3.1 Source data quality compounds with scale

Every synthetic dataset is only as good as the real data it was trained on. If the source data contains biases, gaps or errors, synthetic generation does not correct them. It propagates them. And with each generation cycle, the distortions compound.

Factor	Impact
Selection bias	Who opted in to the source dataset? Who was excluded?
Measurement bias	What was captured, and how? What was missed?
Representation gaps	Which populations are underrepresented or absent?
Model collapse risk	Iterative training on synthetic outputs narrows distribution until data converges on an average that no longer represents the real population
Visibility	Traditional data collection makes biases visible and correctable. Synthetic generation obscures them behind clean-looking output

Key insight: The critical question is not whether source data is perfect (it never is) but whether imperfections are being amplified or corrected downstream. Device Native AI provides clean, structured source data that reduces compounding error from the start.

3.2 Homogeneity and the loss of outliers

Generative models are optimisation machines. They learn the central tendency of a distribution because that is what minimises their loss function. The result is synthetic datasets that faithfully reproduce common patterns but systematically underrepresent the tails.

Domain	What outliers signal	Consequence of missing them
Clinical research	Rare adverse drug reactions, atypical symptom clusters	Drug reaches market with incomplete safety profile
Financial services	Unusual transaction patterns	Fraud model trained to miss the fraud that matters
Insurance	Atypical claim clusters	Emerging risk goes undetected

Consumer behaviour	Early-adopter behaviour, high-value switchers	Leading edge of market shift is invisible
Retail	Price-sensitive niche segments	Loyalty strategy misses highest-value customers

Key insight: In every high-stakes domain, the outlier carries disproportionate commercial or safety value. Synthetic generation systematically underrepresents it. Real behavioural data from Device Native AI preserves what matters most: the unusual.

3.3 The paradox: most needed where most dangerous

The domains where synthetic data is most attractive are precisely the domains where its weaknesses are most consequential.

Factor	Why synthetic data is attractive	Why its failure modes are dangerous
Data scarcity	Real data is scarce, expensive, ethically constrained	Less real data available to validate synthetic output
Regulatory pressure	Synthetic eliminates PII handling burden	Regulators will not accept synthetic confidence at decision points
Outlier consequence	—	Rare events carry highest safety and commercial consequence
Source bias	—	Historical data already contains systemic underrepresentation
Homogeneity risk	—	Central-tendency datasets mask the signals that matter most

Key insight: The industries rushing toward synthetic data because real data is hard to access are the industries that can least afford synthetic data's failure modes. This is the paradox that the market is beginning to acknowledge publicly.

4. Three Paradigms of Customer Data

The framing of "real versus synthetic" is a false choice. The customer data landscape has evolved through three distinct paradigms. Each solves the limitations of the last but introduces its own.

Dimension	Traditional Panel Data	Synthetic Data	Device Native AI
Approach	Ask people what they think	Model what people might do	Observe what people actually do, privately
Strengths	Decades of validation. Captures sentiment, motivation, context. Institutional acceptance	Rapid scale, low cost. Privacy-safe output. ML augmentation. Democratises access	Ground truth behaviour. Individual fidelity. Outlier preservation. Continuous, real-time signal. Privacy by architecture
Weaknesses	Intention-action gap. Expensive. Slow. Panel fatigue. Centralised PII	Inherits source bias. Homogenises. No individual fidelity. Provenance opacity	Requires SDK integration. Scale tied to user base. Cannot capture attitudinal "why" alone
Best for	Exploratory research, attitudinal measurement, concept testing	Volume research, hypothesis generation, ML augmentation, test environments	Personalisation, engagement, behavioural insight, ground truth validation

Key insight: Each paradigm addresses the predecessor's core limitation, but none makes the others obsolete. The strongest data strategy combines all three, with real behavioural data as the ground truth that validates and anchors the other two.

5. Detailed Comparison: Synthetic Data vs Device Native AI

5.1 Data Quality and Fidelity

Dimension	Synthetic Data	Device Native AI
Source fidelity	● Approximation of statistical patterns. Captures central tendencies but smooths individual-level variance	✓ Ground truth from actual individual behaviour. What people do, not what models predict
Variability & outliers	✗ Generative models optimise toward the mean. Rare events underrepresented or absent	✓ Captures genuine edge cases and rare behaviours. Atypical symptoms, early-adopter signals preserved
Bias propagation	✗ Inherits and amplifies source biases. Errors compound (model collapse risk)	● Reflects genuine behavioural biases, but errors are observable and correctable
Auditability	✗ Provenance opaque. Difficult to trace to real-world signal	✓ Full provenance trail: known individual, known context, known timestamp

Key insight: Synthetic data captures patterns. Device Native AI captures reality. The combination, with DNA as the foundation, delivers both scale and fidelity.

5.2 Privacy and Compliance

Dimension	Synthetic Data	Device Native AI
PII in output	✓ No direct PII. Genuine strength for cross-border sharing, vendor testing	✓ Zero-shared data: processed on-device, never centralised. GDPR compliant by design
Provenance	● Regulators scrutinise whether synthetic data inherits source obligations	✓ Clean provenance: data generated on individual's own device under their control
Breach exposure	✓ Low (no PII in synthetic output)	✓ Low (most sensitive PII never leaves device; cloud holds only consented inferences)
Consent model	● Consent relates to source data collection, not synthetic output	✓ Granular: individuals control which insight categories they share

5.3 Personalisation and Engagement

Dimension	Synthetic Data	Device Native AI
Personalisation depth	✗ Segment-level at best. Cannot personalise to an individual	✓ Individual-level, contextual, real-time. Genuine one-to-one personalisation

Engagement impact	● Enables faster research cycles to inform strategy	✓ Up to 44x engagement effectiveness uplift (see Note 8). 25%+ activation uplift (Verizon PoC)

5.4 Cost and Operations

Dimension	Synthetic Data	Device Native AI
Cost per record	✓ Low marginal cost. Cheaper than traditional panels for volume research	● Requires SDK integration. Marginal cost decreases with scale
Cost at scale	✓ Scales cheaply	✓ On-device processing: zero marginal AI costs (no cloud tokens)
ROI mechanism	Cost saving on research	Revenue uplift through engagement + cost saving on cloud AI

5.5 High-Stakes Domain Suitability

Dimension	Synthetic Data	Device Native AI
Healthcare	✗ Most attractive where data is scarce, but outlier loss carries highest consequence	✓ On-device processing bypasses data access barriers. Real data without centralisation
Financial services	✗ Risks training fraud models that miss atypical patterns	✓ Real transaction behaviour via Open Banking APIs, processed on-device
Consumer insights	● Useful for directional research and hypothesis generation	✓ Behavioural ground truth from known populations. Known-universe balancing
Sample representativeness	● Accuracy depends entirely on source data quality	✓ Partial samples from known universes can be balanced to match the whole population

Key insight: The domains where synthetic data is most needed are the domains where Device Native AI's ground truth is most valuable. Deployed together, the combination addresses the paradox.

6. Grounded Synthetic: The Missing Foundation

The concept of Grounded Synthetic resolves the "real versus synthetic" debate by reframing the question. The issue is not whether synthetic data is useful (it clearly is) but whether it is anchored in verified reality.

6.1 The architectural principle

Component	Role
Device Native AI	Provides clean, structured, privacy-safe life-stream data from known individuals within known populations. The ground truth foundation
Synthetic data generation	Extends the reach of real data for scenario modelling, segment exploration, ML training, and hypothesis testing
Grounded Synthetic	The combined approach: synthetic data calibrated, validated and constrained by a verified foundation of real behavioural data

6.2 Known-universe balancing

Consider a retailer with 2 million loyalty members, 150,000 of whom are generating on-device behavioural data through the DataSapien SDK embedded in the retailer's app.

Step	What happens
1. Real data collection	150,000 active SDK users generate structured behavioural data on-device
2. Known-universe weighting	Sample is statistically balanced against the full 2 million population's known demographic and transactional profile
3. Ground truth established	Validated, representative behavioural dataset anchored in real individual behaviour
4. Synthetic extension	Synthetic generation extends the validated foundation for scenario modelling and hypothesis testing
5. Validation loop	Synthetic outputs continuously validated against the real life-stream data foundation

Key insight: This is not a synthetic approximation. It is statistical inference grounded in real behaviour. The anchor is real life-stream data from real individuals.

6.3 Addressing Gartner's five validation requirements

In the Gartner Voice of the Customer report, analyst Michael Maziarka outlined five validation requirements CX leaders should demand from any synthetic data vendor. Grounded Synthetic addresses them as follows:

Gartner Requirement	How Grounded Synthetic Addresses It
Data sourcing and quality	✓ Device Native AI generates structured, clean source data directly from individual behaviour
Comparative testing against first-party VoC data	✓ The real behavioural data layer provides exactly the first-party dataset for comparison
Bias mitigation	✓ Known-universe balancing corrects selection bias. Balanced foundation constrains synthetic generation
Conclusion similarity testing	● Requires process overlay, but significantly easier when ground truth exists
Hallucination checks	● Requires process overlay, but synthetic outputs can be validated against real data anchor

7. Implications for Enterprise Decision-Makers

7.1 For consumer insights and research leaders

The five validation requirements Gartner has outlined are not hypothetical. They represent the standard that procurement teams will increasingly demand from synthetic data vendors. Organisations that can demonstrate clean source data, independent validation and full provenance will be positioned as trusted partners. Those that cannot will face the same scepticism that credit markets are now applying to traditional SaaS platforms.

7.2 For enterprise technology buyers

The choice is not between synthetic and real data. It is between grounded and ungrounded synthetic data. When evaluating vendors, enterprise buyers should ask: where did the source data come from? Is it current? How was it validated? What happens to the outliers? If a vendor cannot answer these questions, the efficiency gains they promise may be built on assumptions that cannot be verified.

7.3 For brands with existing customer relationships

Brands with existing app user bases, loyalty programmes and engaged customer communities already possess the most valuable asset in this equation: a known population of real people, using a real app, generating real behavioural signals every day. Device Native AI transforms that

existing relationship into a continuous, privacy-safe, real-time life-stream data asset that no synthetic generator can replicate, because it is real.

The ground truth foundation is not something that needs to be built from scratch. It needs to be activated within the infrastructure that already exists.

Consider the practical architecture. A retailer with 2 million loyalty app users embeds a lightweight SDK into their existing app. No separate download. No new permissions screen. From that moment, every consenting user's device begins generating structured behavioural data: app usage patterns, health and wellness signals, location context, calendar rhythms, purchase behaviour via Open Banking APIs — all processed on-device, with only consented inferences shared back as Zero-Party Data.

Within weeks, the brand has a living, continuously updating behavioural dataset from a known, demographically profiled population. That dataset can be statistically balanced against the full user base using known-universe weighting (as described in Section 6.2). And it can serve as the verified foundation against which any synthetic extension is calibrated, validated and constrained.

This is where the competitive advantage crystallises. Third-party synthetic data vendors can generate plausible-looking personas at scale. But they cannot generate personas grounded in what your specific customers actually did yesterday. Only the brand that holds the first-party behavioural relationship can do that. The ground truth is not a commodity that can be purchased — it is an asset that accrues to the organisation that activates it.

For organisations evaluating their data strategy, the implication is direct: the longer the gap between having an engaged customer base and activating on-device behavioural intelligence within it, the longer synthetic outputs remain ungrounded — and the wider the window for competitors who move first.

8. Summary: Where Each Approach Wins

Synthetic Data Strengths

Strength	Detail
Speed and scale	Millions of records generated rapidly. Ideal for prototyping, hypothesis testing, scenario modelling
Privacy-safe output	No direct PII. Enables cross-border sharing, vendor testing, open research
Cost efficiency	Fraction of traditional panel costs. Democratises access for smaller organisations
ML augmentation	Proven value in computer vision, NLP and classification tasks
Test environments	Sensible default for QA, staging and system integration

Device Native AI (Ground Truth) Strengths

Strength	Detail
Source fidelity	Ground truth from actual individual behaviour
Outlier preservation	Captures genuine edge cases and individual-level variance that synthetic generation underrepresents
Individual personalisation	Real-time, contextual, one-to-one. Cannot be replicated by segment-level synthetic data
Auditability	Full provenance trail: known individual, known context, known timestamp
Known-universe balancing	Partial samples from defined populations can be weighted to represent the whole
Privacy by architecture	Zero-shared data, processed on-device. GDPR compliant by design
Data veracity	Multi-source triangulation and Verifiable Credential integration
Engagement uplift	Up to 44x on contextual notifications. 25%+ activation uplift (Verizon PoC)

The Strategic Recommendation

Synthetic data and Device Native AI are not competing technologies. They are complementary layers of a modern data strategy. Synthetic data provides scale, speed and cost efficiency for directional research, hypothesis testing and ML training. Device Native AI provides the ground truth foundation that makes synthetic outputs trustworthy: clean source data, preserved outliers, known-universe balancing, and full provenance.

Organisations that deploy both, with Device Native AI as the foundation and synthetic data as the scaled extension, will build data strategies that are faster than real data alone, more trustworthy than synthetic data alone, and more defensible than either in isolation.

This is Grounded Synthetic.

DataSapien | Device Native AI: Synthetic Data's Missing Foundation
Built by the team behind CitizenMe (500K MAU), who have spent a decade building private, on-device data architecture for consumer brands and insight platforms.

Patent pending: GB2503062.8
Contact: hello@datasapien.com | datasapien.com

References

1. Insight Innovation Ventures (2026). "The Qualtrics Canary." Substack, March 22, 2026.

2. Maziarka, M. (2026). Maziarka, M., quoted in Becker, M. (2026). 'Insight Is Cheap. Qualtrics Is Selling Something Harder.' CMSWire, March 2026. https://www.cmswire.com/customer-experience/insight-is-cheap-qualtrics-is-selling-something-harder/. Maziarka is also co-author of the Gartner Magic Quadrant for Voice of the Customer Platforms (9 March 2026, ID G00836480), which notes that synthetic data services in VoC 'are still nascent.

3. Becker, M. (2026). "Insight Is Cheap. Qualtrics Is Selling Something Harder." CMSWire, March 2026. https://www.cmswire.com/customer-experience/insight-is-cheap-qualtrics-is-selling-something-harder/

4. De Haan, E., Wiesel, T. & Pauwels, K. "The Effectiveness of Different Forms of Online Advertising for Purchase Conversion." IJRM.

5. Dawes, J. "Advertising Effectiveness and the 95:5 Rule." Ehrenberg-Bass Institute / LinkedIn B2B Institute.

6. Cohen, S. (2026). "Synthetic data is doing too much work as a term." Fairgen.

7. Shumailov, I. et al. (2023). "The Curse of Recursion." arXiv:2305.17493.

8. **Note on the 44x engagement effectiveness figure.** The 44x engagement effectiveness figure draws on two independent bodies of research. First, De Haan, Wiesel and Pauwels (Ref. 4) established that traditional digital advertising conversion rates typically sit at approximately 1%, with most impressions served to consumers who are not in-market at the point of delivery. Second, Dawes (Ref. 5) demonstrated that at any given time, only approximately 5% of potential buyers are actively in-market. The 44x figure represents the ratio between traditional broadcast conversion

(~1%) and intent-driven activation conversion (~44%), where activation is triggered only when on-device intelligence detects that an individual has moved into the in-market window. The 44% conversion estimate is consistent with observed rates for intent-triggered activations such as Google Shopping ads (10–30%), first-party triggered emails (30–50%) and in-app contextual offers at the point of decision. Readers should note that the 44x figure describes the *effectiveness ratio* of intent-driven versus broadcast delivery.