Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI

Published 24 June 2021 - ID G00750175 - 23 min read

By Analyst(s): Leinar Ramos, Jitendra Subramanyam

Initiatives: Artificial Intelligence

> Synthetic data is often seen as a lower-quality substitute, useful only when real data is inconvenient to get, expensive or constrained by regulation. This misses the true potential of synthetic data. The fact is you won't be able to build high-quality, high-value AI models without synthetic data.

Additional Perspectives

- Summary Translation: Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI
  (16 July 2021)

More on This Topic

This is part of an in-depth collection of research. See the collection:

- Maverick* Research: Push Yourself to Think Beyond Conventional Wisdom

Overview

Specific Maverick Caution

This Maverick* research contradicts the prevailing wisdom that real datasets will be the main driver for AI progress, arguing that synthetic data will become much more important. Its findings and advice should therefore be treated with caution.

Maverick Findings

- Real datasets are typically incomplete, imbalanced and not fully representative of the business domain. Synthetic data is designed to address these shortcomings.

- The next leap in AI efficacy cannot be achieved without injecting domain knowledge into training datasets using synthetic data techniques.

- Synthetic data is the *only* viable approach for solving two large, prevalent classes of business-critical problems: portfolio optimization and initiative sequencing.

Maverick Recommendations

- Develop capabilities in simpler data generation techniques like statistical modeling, and mature to more complex ones like simulation and agent-based approaches.

- Take advantage of synthetic data to supplement small, existing datasets that are currently being ignored.

- Use synthetic data to test and improve the quality of your AI models.

Maverick Research

This is "Maverick" research, designed to spark new, unconventional insights. Maverick research is unconstrained by our typical broad consensus-formation process to deliver breakthrough, innovative and disruptive ideas from our research incubator. We are publishing a collection of several Maverick research lines this year, all designed for maximum value and impact. We'll explore each of these lines of research to help you be ahead of the mainstream and take advantage of trends and insights that could impact your IT strategy and your organization (see Note 1).

Analysis

What You Really Need to Know

### The Synthetic Data Opportunity: It's Not What You Think

Synthetic data, a class of data that is artificially generated, is not a new concept. [1] It has been around for years, but it's often seen as a lower-quality substitute, useful only in certain situations where real data is inconvenient to get, expensive or constrained by regulation. Its adoption is limited to specific use cases in a narrow set of industries — for instance, it is used to preserve privacy in finance and healthcare datasets. [2]

Although these are valid use cases, this conventional wisdom completely misses the true potential of synthetic data — far from fake, an afterthought or a necessity due to privacy regulations, the fact is *you won't be able to build high-quality, high-value AI models without synthetic data*.

Synthetic data is on a trajectory to go from a sideshow to becoming the main force behind the future of AI (see Figure 1).

Figure 1: By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models
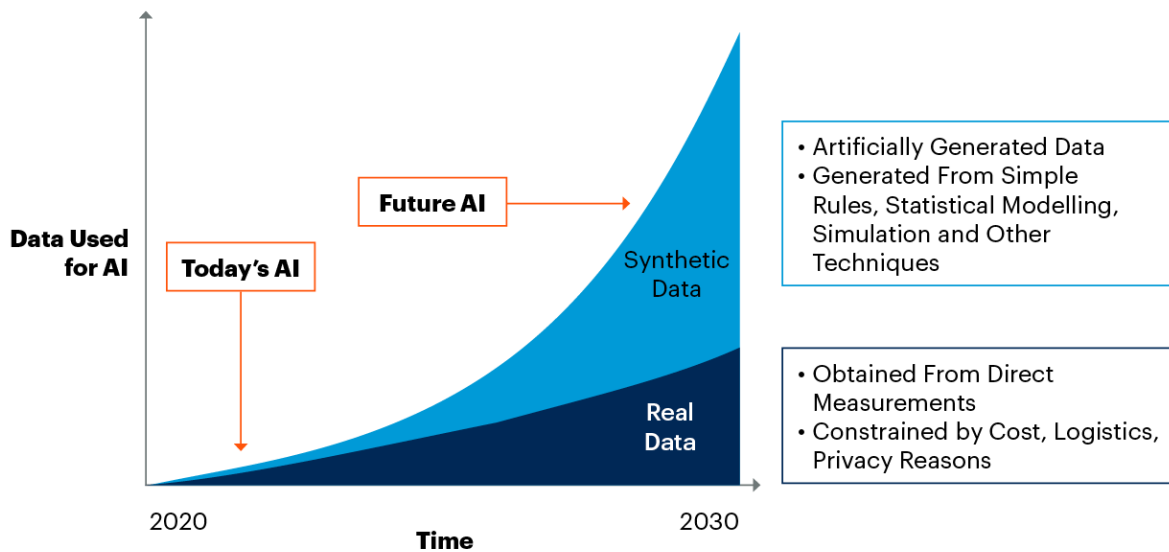
Source: Gartner
750175_C

Currently, most organizations believe that *big data* is the way to drive AI value. Consider all of your organization's investments in this space: data warehouses, data lakes, data pipelines, and data quality programs. What they all have in common is an assumption: *Your data has an untapped value significantly beyond its current use*.

But this is often false, as evidenced by the poor *return on investment* we typically see for these initiatives. Real data is not the new oil, nor is it the invaluable asset you believe it to be. Why not? Well, for a start, your data has several problems:

- **Real data is incomplete:** AI needs both large and diverse datasets, but real data is often incomplete, excluding infrequent scenarios that are critical for AI performance.

- **Real data is expensive:** It is hard to collect, integrate, store and maintain.

- **Real data is biased**: Even if data perfectly reflects reality, it can encode biases present in the real world that we would like to remove.

- **Real data is restricted**: Regulation is increasingly limiting data use for AI. [3]
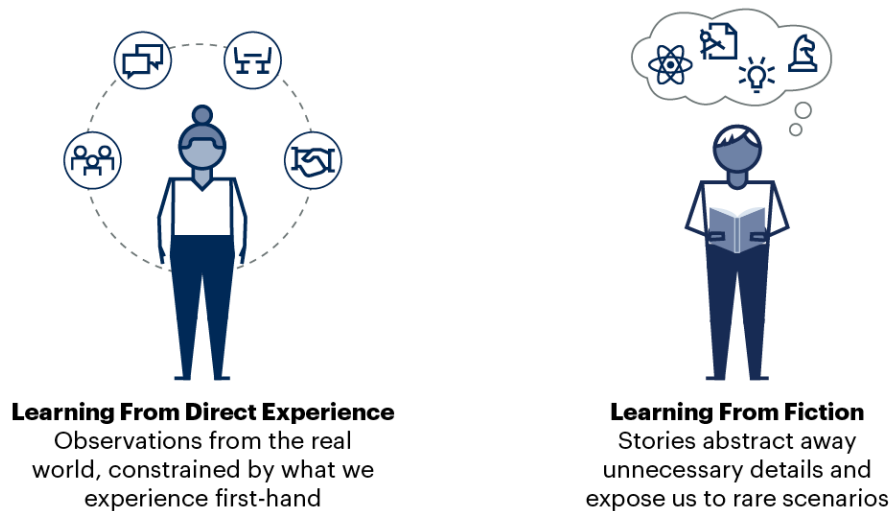
This research explores how synthetic data addresses these challenges. But first, it's key to ask: *Why would synthetic data even work?*

The core intuition is that synthetic data — by virtue of being artificially generated — allows us to introduce knowledge into AI models that we wouldn't be able to incorporate via real data. Synthetic data is not "fake" data; it is derived from reality.

Consider how humans learn. We learn from experiencing the world, but we also learn from literature, myths, fables, anecdotes and countless other stories. These stories are powerful not because they accurately reflect the world, *but precisely because they don't:* They abstract away unnecessary details and expose us to scenarios that we will rarely, if ever, experience firsthand.

Our ability to simulate and explore hypothetical worlds lies at the core of what makes us intelligent. Like fiction, synthetic data allows us to inject information into our AI models that we wouldn't be able to transmit otherwise. And like fiction, synthetic data can be more valuable than direct observation (see Figure 2).

Figure 2. Analogy to Human Learning

Analogy to Human Learning

Learning From Direct Experience
Observations from the real
world, constrained by what we
experience first-hand

Learning From Fiction
Stories abstract away
unnecessary details and
expose us to rare scenarios

Source: Gartner
750175_C

Gartner

A lot of business value is riding on this shift toward synthetic data: The future of AI itself will depend on its adoption. Progress in AI has so far been an interplay between creating more sophisticated models, training with increasingly larger datasets and using higher computational power. For instance, deep learning models (like AlexNet) [4] only became feasible after the adoption of larger datasets (ImageNet) and new computing architectures (GPUs). So it's worth asking: *What will be the fundamental constraint of future advances in AI?*

Most of AI's current research goes into creating larger and more complex models. This has given us exciting developments in transformer architectures, [5] self-supervised learning [6] and many other model types. [7] Some of these models already have hundreds of billions of parameters. [8]
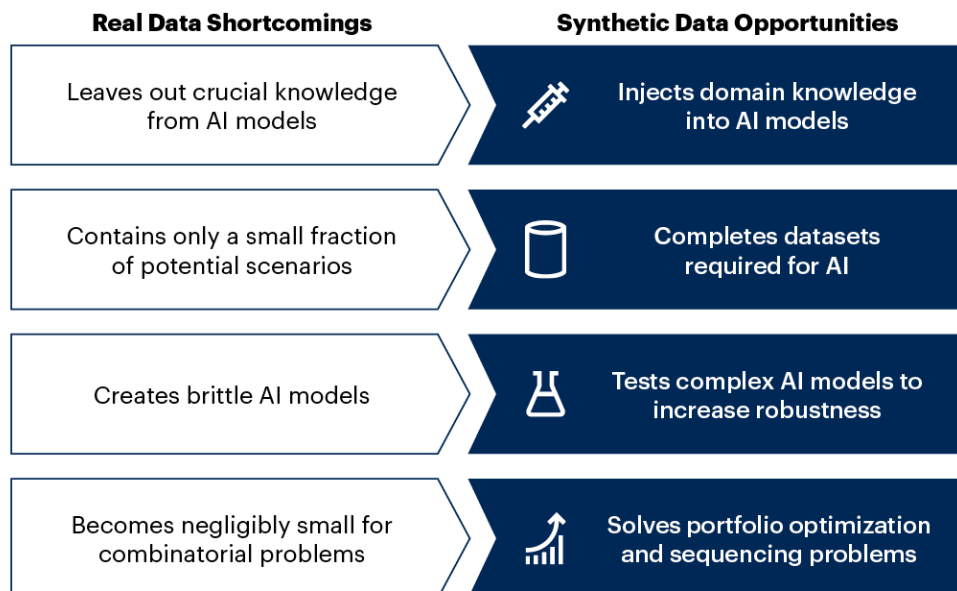
But as model sizes increase, having a lot of data matters more than model architecture. [9] At a large data scale, relatively simple model architectures achieve performance on par with more complex architectures. [10] So the fundamental constraint of AI progress will be data, not model architecture or computing.

Therefore, synthetic data is a must-have, but not for the reasons we might expect. Progress in AI, and the business value this progress can generate, requires synthetic data. As Figure 3 illustrates, there are four main reasons why it will be essential:

1. Injecting domain knowledge into the training of AI models to improve the quality of the model's predictions

2. Completing the datasets we need by generating infrequent or unknown scenarios to train AI models

3. Effectively testing complex AI models by illuminating edge cases that would otherwise never be found

4. Solving two classes of prevalent business-critical problems — optimizing a portfolio of initiatives and finding the right sequence in which to execute on these initiatives.

Figure 3: Four Nonobvious Reasons Synthetic Data Is a Must-Have for AI

Four Nonobvious Reasons, Synthetic Data Is a Must-Have for AI

Real Data Shortcomings	Synthetic Data Opportunities
Leaves out crucial knowledge from AI models	Injects domain knowledge into AI models
Contains only a small fraction of potential scenarios	Completes datasets required for AI
Creates brittle AI models	Tests complex AI models to increase robustness
Becomes negligibly small for combinatorial problems	Solves portfolio optimization and sequencing problems

Source: Gartner
750175_C

Gartner

Expanded Treatment

Four Reasons Synthetic Data Is a Must-Have

Reason No. 1: Synthetic Data Injects Domain Knowledge Into the AI Model's Training Dataset

In AI, our goal is not to give our models as much real data as we can; our goal is to give them as much information as possible. This information might come from real data, but also from useful domain knowledge that we have about the world in which the AI model must operate.

Domain knowledge consists simply of things we already know that are relevant to the context in which the AI model is working. In image recognition, for example, we use augmentations (a type of synthetic data) like image rotation to teach AI models useful domain knowledge — namely, that rotation shouldn't matter for classification (i.e., a cat is still a cat even when we rotate its picture). Similarly, in deciding whether machine data patterns of computer activity are normal or malicious, it pays to know more about the context in which the computer is being used. [11] Without this domain knowledge, AI models in complex problem domains will figuratively just spin their wheels.

Figure 4 illustrates how both real data and domain knowledge can be fed into a model that generates synthetic data for AI training.

Figure 4: Increase Information, Not Data

Increase Information, Not Data

Examples
• Data Distributions
• Laws of Physics
• Augmentations
  (Crop, Pose, etc.)
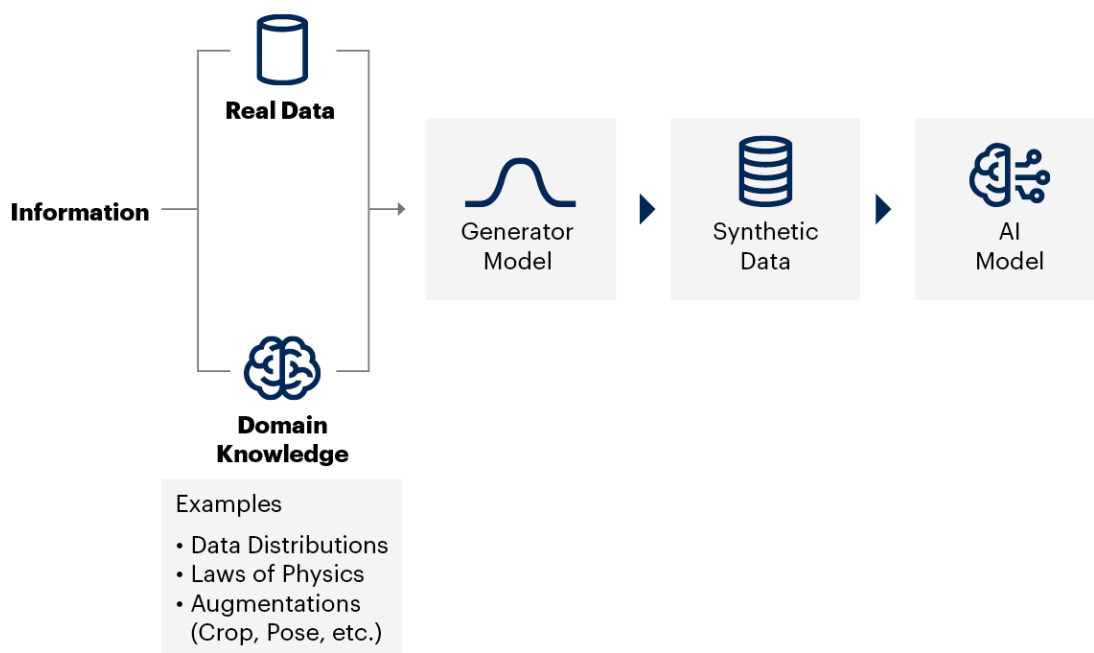
Source: Gartner
750175_C

Gartner

At first, introducing domain knowledge appears to be missing the point — after all, we want AI to learn by itself and not to rely on hard-coded information. But this is a false assumption. The fact is *humans are not blank slates either.* We are born with core knowledge such as the notion of objects, actions, agents, basic physics, geometry and more. [12] These priors make us much more efficient in learning from the data we encounter. [13]

Thus, the best way to make progress in AI is to both 'teach' the AI model directly with real data and to also incorporate additional knowledge about the world. Synthetic data is a scalable way to transmit this additional knowledge — not by hard-coding rules into AI models directly, but by modifying the data generation process and letting machine learning models extract these patterns. This is an example of the broader trend of composite AI, combining different AI techniques to improve the efficiency of learning (see Emerging Technologies and Trends Impact Radar: Artificial Intelligence).
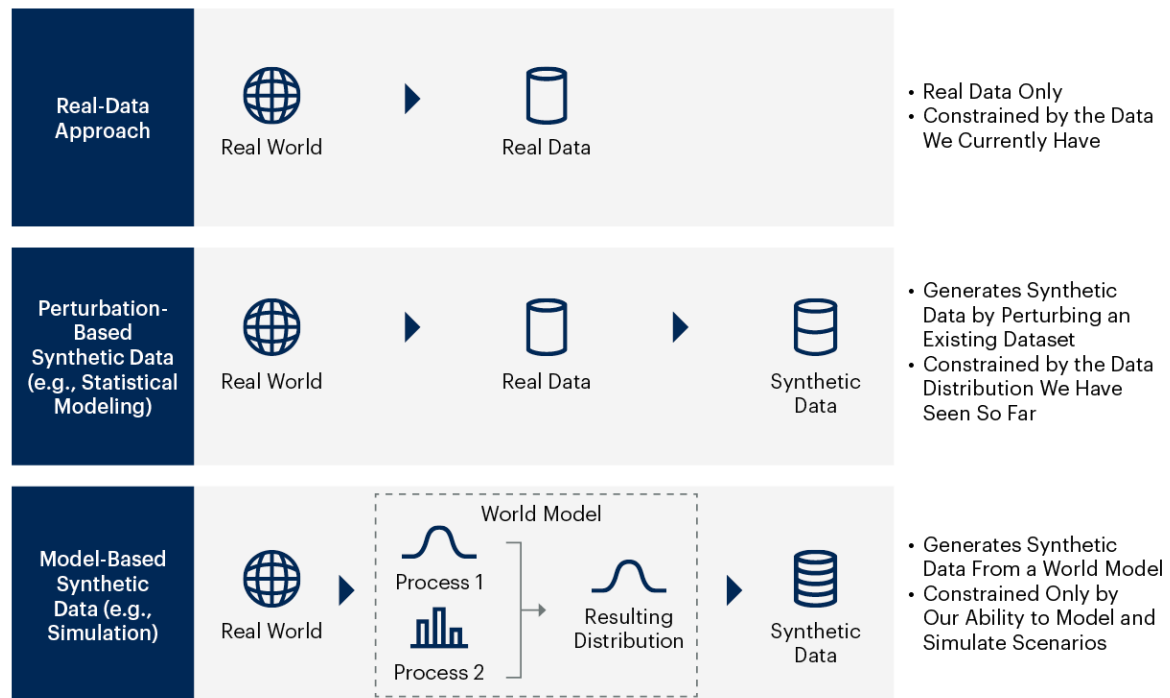
Consider the example of detecting rare product defects in manufacturing. We can't really train an AI model with the very few images we have (defects are measured in parts per million). Instead, we can synthesize images, projecting known defects onto the surfaces of different products, changing their rotation, size and shape. [14] In this way, we are introducing domain knowledge: what features matter (e.g., product cracks, blemishes, etc.) and what features don't matter (e.g., specific surface, rotation, etc.) for the object detection task.

Reason No. 2: Synthetic Data Completes the Datasets We Need for AI

AI datasets are incomplete: We often don't have sufficient data to train models on a long tail of scenarios. This is crucial as AI performance is increasingly driven by how well models behave in these infrequent scenarios. Synthetic data can solve this.

Yet, there is often a misconception that synthetic data can only generate known scenarios, given that data generation is modeled from what we have seen in the past. What this misses is that simple, known distributions can interact to create previously unknown scenarios. Figure 5 illustrates how this is possible. In particular, simulation and other model-based approaches can help generate situations that rarely, if ever, appear on real datasets.

Figure 5: Synthetic Data Generates Unknown Scenarios
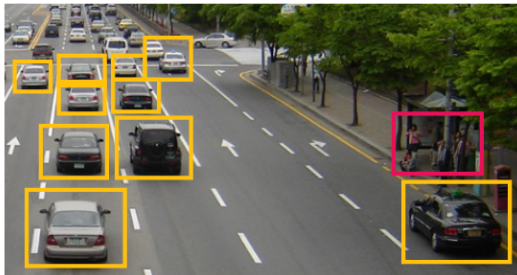

Synthetic Data Generates Unknown Scenarios

Real-Data Approach	Real World → Real Data		• Real Data Only • Constrained by the Data We Currently Have
Perturbation-Based Synthetic Data (e.g., Statistical Modeling)	Real World → Real Data → Synthetic Data		• Generates Synthetic Data by Perturbing an Existing Dataset • Constrained by the Data Distribution We Have Seen So Far
Model-Based Synthetic Data (e.g., Simulation)	Real World → World Model (Process 1, Process 2 → Resulting Distribution) → Synthetic Data		• Generates Synthetic Data From a World Model • Constrained Only by Our Ability to Model and Simulate Scenarios

Source: Gartner
750175_C

Gartner

This ability to generate unknown scenarios is crucial. As complexity increases, the scenarios that we haven't seen far outpace the ones we have. And even when we perfectly capture all the data from what *has* occurred, we still leave behind all the data that *could* occur. Synthetic data is the only way to train AI for these new scenarios we haven't seen. It goes back to the way we learn from stories — ways in which the world *could* be — and from hypothetical reasoning.

Consider, for example, video analysis. AI is now used for many video applications: inventory control in retail stores, real-time threat detection in defense and object detection in self-driving cars.

But video is challenging. AI performance is sensitive to even small changes in factors like lighting conditions, so large datasets need to be collected to capture all these scenarios. Additionally, the cost of labelling videos can quickly become prohibitive and the resulting labels are not always accurate. This limits AI video analysis to a handful of use cases in organizations with large budgets. Further, privacy regulations can severely restrict the use of video. In the EU, for instance, organizations need permission from every person in street scenes, rendering real video unusable for self-driving and other use cases (see Figure 6).

Figure 6: Real Versus Synthetic Data for AI Video Analysis

Real Versus Synthetic Data for AI Video Analysis

| Real Data for Video Analysis	Synthetic Data for Video Analysis
• Limited examples for infrequent scenarios (e.g., collisions, extreme weather) • High cost of obtaining, processing, storing and labeling • Inaccurate labeling (bounding boxes) • Privacy issues	• Diverse scenarios can be generated (e.g., collisions, extreme weather) • Zero marginal cost of obtaining additional data (fixed cost to create a data generator) • High-granularity labeling (pixel level) • No privacy issues

Source: Gartner
750175_C

Gartner

Instead, consider how synthetic data can be used for video: Simulation engines can create photo-realistic models of streets, [15] retail stores [16] or manufacturing plants. [17] Then, data can be generated for a variety of conditions (e.g., lighting, weather, textures) with no privacy issues. Models can generate perfectly accurate, pixel-level labels at zero-marginal cost. This is not science fiction: Photo-realistic synthetic data is now used for self-driving cars, satellite imagery, [18] weapon detection [19] and more.

Reason No. 3: Synthetic Data Is Required for Testing Complex AI Models

Underspecification is one of AI's deepest problems. [20] It arises whenever there are several AI model configurations that are equally good at predicting test data, but have a significantly different performance in the real world. This is different from "data drift": Even if test data perfectly resembles real data, models with equally good test performance can fail once deployed, as illustrated in Figure 7.

Figure 7: Machine Learning Underspecification

Machine Learning Underspecification
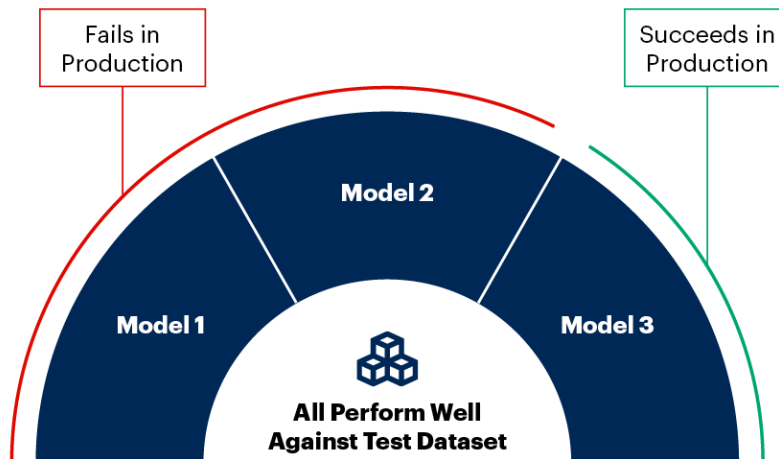
Source: Gartner
750175_C

Gartner

This challenge becomes critical for large model architectures: As model complexity increases, more configuration solutions can be found, without a clear way of telling which ones will work once deployed. In these circumstances, we say that the AI pipeline is underspecified, because it can produce many invalid solutions.

The answer to underspecification is to deliberately stress-test AI models, observing their output distribution for carefully designed inputs. *Synthetic data is the main way to design these test inputs.*

Figure 8 shows how synthetic data can be used to solve for underspecification: We first define variables for stress-testing, creating synthetic data with perturbations along these variables. We then determine the model's distribution of output variables and choose robust models with less variation for these perturbations.

Figure 8: Testing AI Models With Synthetic Data

Testing AI models With Synthetic Data
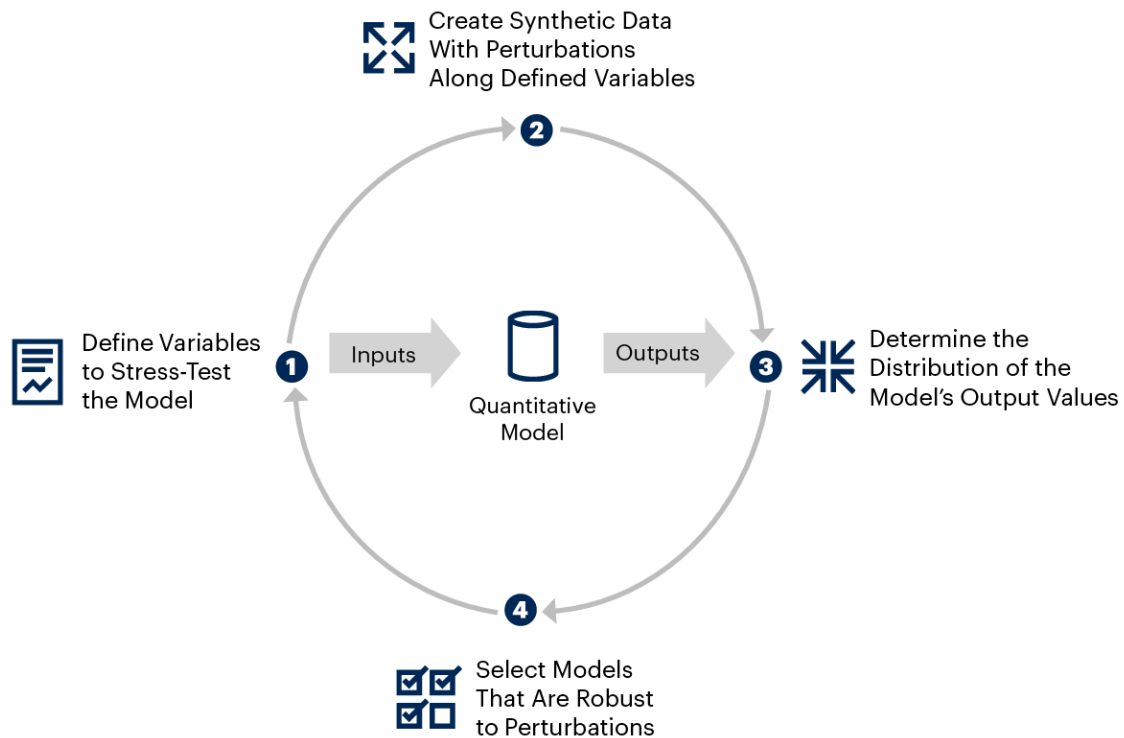
Source: Gartner
750175_C

For instance, one might construct an "unfair" synthetic dataset to then test for model fairness performance, helping disambiguate which models would be robust once deployed along this dimension.

In sum, quality assurance will become increasingly important for AI. And without synthetic data, it would be impossible to stress-test and increase model robustness.

Reason No. 4: Synthetic Data Is the Only Solution for Business Problems That Manifest Combinatorial Explosion

There is a large class of frequently occurring business-critical problems for which synthetic data is the only alternative. Collecting real data for these problems is a nonstarter due to combinatorial explosion. While these problems don't always involve AI, once synthetic data is generated to solve them, they become suitable training datasets for AI models.

Combinatorial explosion is simply the rapid growth in complexity due to the number combinations or permutations. For instance, three elements generate six permutations. But Figure 9 shows what happens as we further increase the input size: After just 26 elements, we get to a number of permutations of 10 to the power of 24, which is roughly equal to the number of stars in the observable universe. [21]

Figure 9: Combinatorial Explosion

Combinatorial Explosion
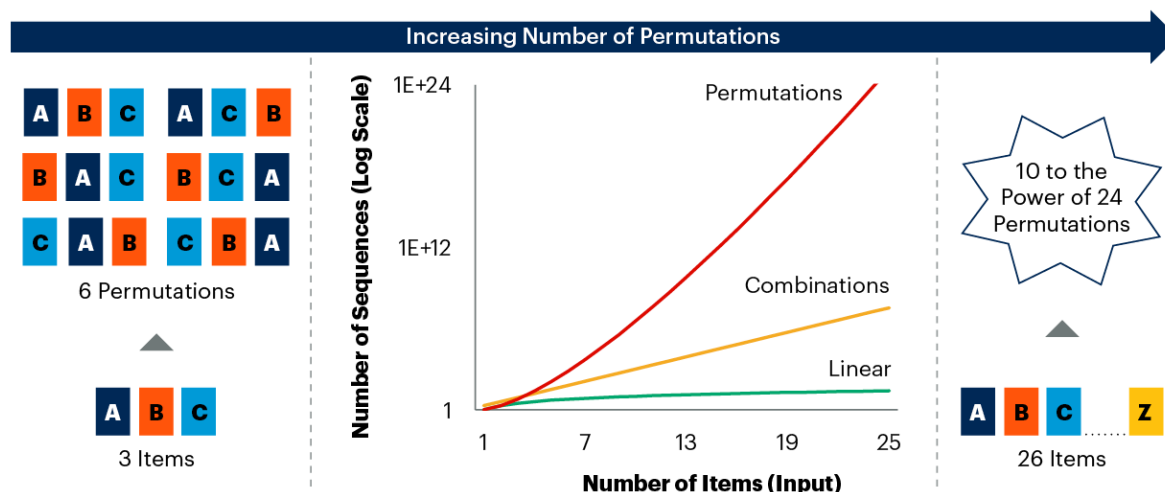
Source: Gartner
750175_C

Gartner

This is more than an intellectual curiosity. Organizations across all industries deal with combinatorial explosion challenges. Two particularly impactful business problems are:

- **Portfolio optimization:** Selecting a combination of investments out of several options, which results in exponential growth (yellow line in Figure 9)

- **Initiative sequencing:** Selecting a sequence of investments when the order matters, which results in combinatorial growth (red line in Figure 9)

These problems come up frequently in business settings. Following are two examples.

Suppose you have a set of 10 data and analytics initiatives, and you've quantified the cost, risk and value of each initiative. Question: What is the best portfolio of 10 or fewer initiatives, where "best" means the portfolio is under a certain cost and risk and above a certain total value?

A brute force way to answer this question requires creating all one-initiative portfolios, two-initiative portfolios, three-initiative portfolios and so on, and calculating the overall value of each one. In general, the number of such subset portfolios you can create from n distinct initiatives is 2^n. [22] When n is 10, the total number of possible portfolios is 1,024, which is a manageable number. But 10 initiatives is a small number if your goal is to discover an optimal portfolio; the more initiatives you start with, the better the chance of discovering the optimal portfolio. However, as n increases, the number of possible portfolios quickly grows out of control:

- n = 10, number of all possible portfolios = 1,024

- n = 30, number of all possible portfolios = 1,073,741,824

- n = 50, number of all possible portfolios = 1,125,899,906,842,624

- n = 100, number of all possible portfolios = 1,267,650,600,228,229,401,496,703,205,376

With even a relatively small number of total initiatives, say 100, the number of possibilities is a n*onillion* (i.e., 10^30 or 1 followed by 30 numerals). [23]

The only way to make progress on this type of portfolio optimization problem is to synthetically generate a large but manageable number of initiatives — say, a trillion — and use efficient algorithms to sift through to find the optimal clusters of initiatives. An alternative approach would be the use of quantum computing algorithms, but these are still in the far future (see Predicts 2021: Disruptive Potential During the Next Decade of Quantum Computing).

A similar type of uncontrolled growth occurs in problems where you're trying to find the best way to sequence investments. In many situations, the order in which we invest can make a difference. The outcome of a series of investments is likely to be path-dependent. But how much difference is there between investing in one order versus another? Can we quantify the extent to which the return on an investment depends on which investments are made before it?
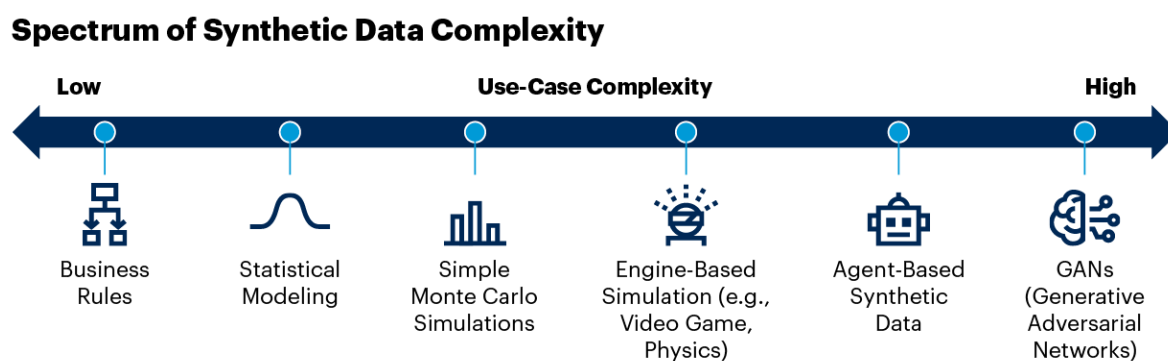
Suppose we tried to use benchmark data to answer these questions. If we're looking to make statistical inferences about how to sequence three distinct investments from a possible set of 14 investments (we choose these numbers because this reflects an actual dataset we have at Gartner), there are 14*13*12 ways, or 2,184 unique ways, to pick an ordered sequence of three investments. If we are to have an n of roughly 30 for every investment sequence (for making statistical inferences that pass muster), then the size of the benchmark dataset must be at least 2184*30 = 65,520 data points. Lacking a dataset of this size, synthetic data can be used to take the original dataset as an input and carefully construct a much larger synthetic one based on a series of business rules and prior knowledge, resulting in a valuable dataset that can help guide investment sequencing decisions.

How to Get Started With Synthetic Data

As we've seen so far, synthetic data is essential for the future of AI. But this future can often feel overwhelming for organizations, so it's important to ask: *How do we get started with synthetic data?*

First, a common misconception on synthetic data is that it requires complex data generation. In reality, it doesn't always need advanced techniques like simulation or generative adversarial networks (GANs). In fact, there is a spectrum of techniques, depending on the complexity of the use case and the underlying phenomenon that we are modelling, as seen in Figure 10.

Figure 10: Spectrum of Synthetic Data Complexity

Spectrum of Synthetic Data Complexity

Low — Use-Case Complexity — High

Business Rules | Statistical Modeling | Simple Monte Carlo Simulations | Engine-Based Simulation (e.g., Video Game, Physics) | Agent-Based Synthetic Data | GANs (Generative Adversarial Networks)
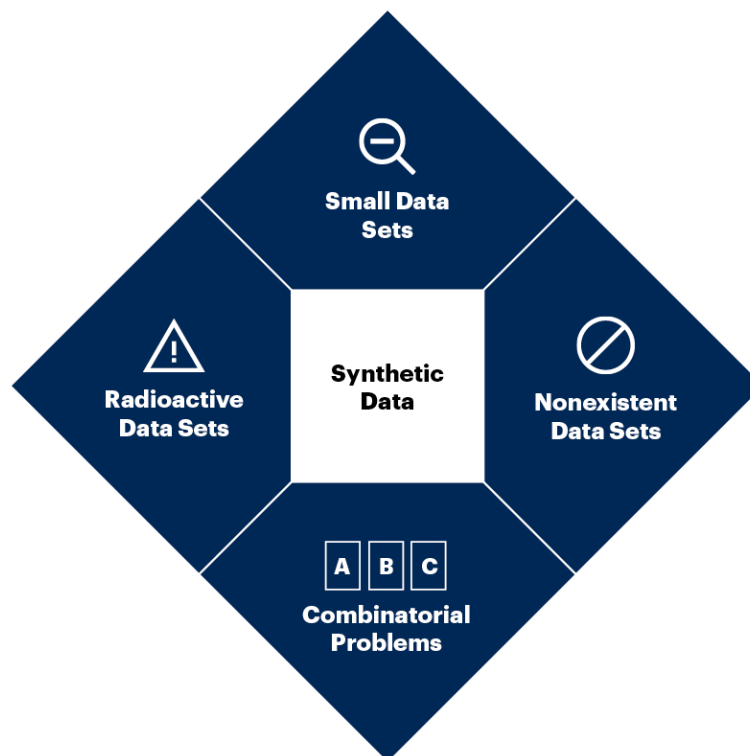
Source: Gartner
750175_C

Given this, we recommend starting by developing capabilities in lower-complexity techniques like business rules and statistical modelling, then maturing to simulation and agent-based approaches (see Hype Cycle for Data Science and Machine Learning, 2020).

Similarly, as Figure 11 illustrates, we recommend identifying AI use cases where your existing datasets are considered too small, radioactive (due to privacy risks) or even nonexistent (no access to the data needed). Similarly, consider synthetic data for portfolio optimization and sequencing problems, as discussed in the previous section. This approach will help you identify initial use cases and start building synthetic data generation capabilities that will be essential for future AI value.

Figure 11: Synthetic Data Opens New Possibilities

Synthetic Data Opens New Possibilities

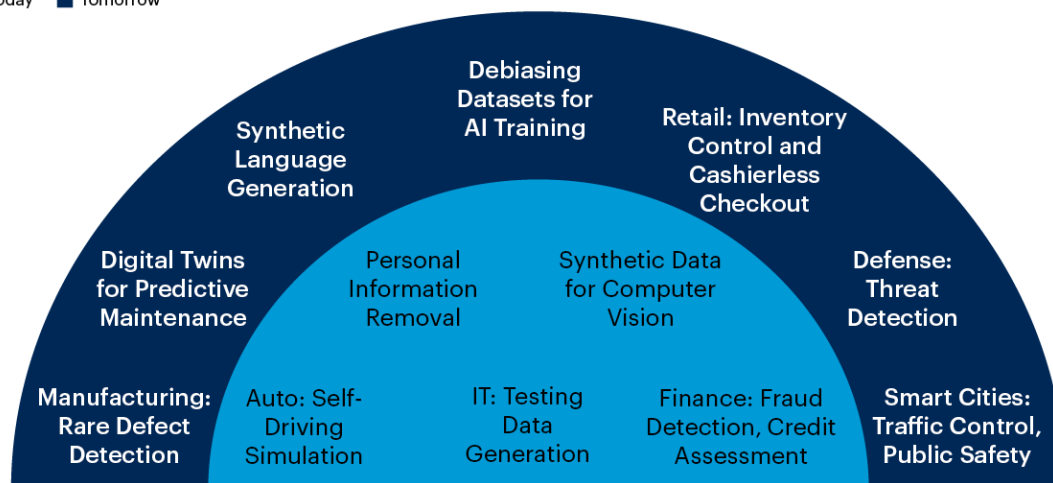Source: Gartner
750175_C

Gartner

Finally, it is important to be mindful of current synthetic data limitations. [24] In particular, the data generation process can introduce bias into AI models and inadequately represent the underlying real-world phenomena. These potential risks need to be evaluated on a case-by-case basis and recalibrated with time; use cases that are not feasible today might become feasible in a few years. This is because the technologies underpinning synthetic data, like engine-based simulation, are rapidly improving due to advances in other areas like the video game industry.

To this end, Figure 12 illustrates some of the use cases in different industries that are common today, as well as some of the emerging ones that could mature in a few years. We advise you to explore and monitor these and other relevant use cases as synthetic data generation techniques evolve.

Figure 12: Illustration of Synthetic Data Use Cases

Illustration of Synthetic Data Use Cases

■ Today ■ Tomorrow

Debiasing Datasets for AI Training

Synthetic Language Generation

Retail: Inventory Control and Cashierless Checkout

Digital Twins for Predictive Maintenance

Personal Information Removal

Synthetic Data for Computer Vision

Defense: Threat Detection

Manufacturing: Rare Defect Detection

Auto: Self-Driving Simulation

IT: Testing Data Generation

Finance: Fraud Detection, Credit Assessment

Smart Cities: Traffic Control, Public Safety

Source: Gartner
750175_C

Gartner.

Synthetic Data Changes the AI Game

If synthetic data is the future of AI, *what changes and why should we care?*

Synthetic data has a practical effect on AI strategy, fairness, technology and sourcing as Figure 13 illustrates:

Figure 13: The Long-Term Promise of Synthetic Data

The Long-Term Promise of Synthetic Data
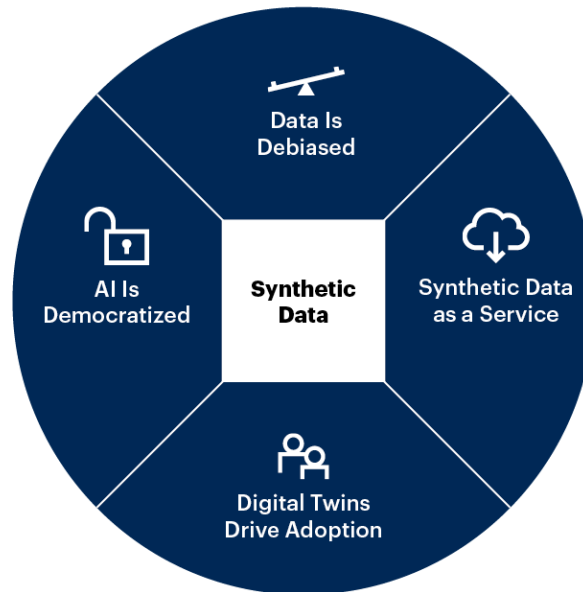
Source: Gartner
750175_C

Gartner

1. **Data is debiased:** AI bias is increasingly a concern and a limiting factor for AI adoption. There are many places in our AI pipelines where bias can be introduced: from our data collection process and model architecture selection to the way we incorporate model's outputs into our decisions. But ultimately, even if we solve these problems, we are left with the fact that real datasets aren't fair because the real world isn't fair — *even datasets that perfectly reflect reality can be deeply biased.*

Synthetic data is, therefore, the only way to train AI for the world we want to live in, creating synthetic datasets that are fair with respect to legally-protected attributes and other dimensions we care about. AI fairness is an emerging field, but synthetic data will be the key tool to avoid perpetuating harmful patterns present in the real world. Indeed, there are already promising signs that synthetic data can be used to debias datasets, while maintaining high AI model accuracy. [25]

2. **AI is democratized:** Data availability is a difficult barrier to entry for startups and smaller organizations. Imagine, for instance, the data disparity that *neobank* startups face against established incumbents with millions of customers. Synthetic data democratizes the playing field by allowing smaller organizations to create AI models without a lot of data, effectively solving their *cold-start* problem.

On the other hand, synthetic data is also inherently more shareable, avoiding the privacy pitfalls that plague real datasets. This allows for more collaboration and innovation. Going back to the banking example, established incumbents can work more effectively with *FinTech* startups by giving them access to synthetic data, running proofs of concept to test their capabilities. [26]

Together, these factors will create more opportunity for AI startups to both compete and partner with established players (see Case Study: AI Innovation With Startups (Stora Enso)).

3. **Synthetic data as a service**: Synthetic data technology requires an initial fixed cost to build the data generator, but a near-zero marginal cost once it has been built — *the cost of generating an additional data point is negligible*. These unit economics will drive the emergence of central players investing in building complex simulations and offering broad access to their data or simulation as a service.

Further, complex synthetic data generation will require bringing together a wide variety of skills, from AI expertise and domain knowledge to motion capture and computer-generated image creation. For this reason, it is likely that most synthetic data will be created by specialized players and platforms.

However, individual organizations will have a crucial role to play in customizing synthetic data generation to their specific use cases, testing and monitoring its quality, and combining it with real datasets to extract value from their investments. It's not too soon to start thinking about how you'll govern your synthetic data assets.

4. **Digital twins drive adoption**: A digital twin is a virtual representation of an entity such as a physical asset, person or process. Digital twins are currently adopted to improve situational awareness and automate responses to patterns detected in the monitoring data (see Use 4 Building Blocks for Successful Digital Twin Design).

Digital twins are models of the world. As these models become more sophisticated, they will be increasingly used to simulate complex what-if scenarios and generate synthetic data to train AI models. This will help, for example, to predict and mitigate potential risks (e.g., predictive maintenance for utility companies).

In this way, we will be able to train AI not only on what has happened before, but also on what could happen. This ability to run scenarios for real entities will make digital twins a driving force behind synthetic data adoption.

In sum, synthetic data will transform the AI landscape. And the opportunities are vast: Organizations that invest early in this technology will be uniquely positioned to drive value from AI, tackling business problems that wouldn't be solvable otherwise, opening up use cases for which they have limited data and being able to reliably test their complex AI models to ensure they will work once deployed. *The future of AI is synthetic and it belongs to the ones that embrace synthetic data early.*

Evidence

[1] Related terms include: simulated data, anonymized data, test data, data masking, real data, data encryption, hashing and scrambling.

[2] Mayo Clinic Completes DeIdentification of Expansive Medical Dataset, Mayo Clinic News Network.

[3] Europe Fit for the Digital Age: Commission Proposes New Rules and Actions for Excellence and Trust in Artificial Intelligence, European Commission.

[4] ImageNet Classification With Deep Convolutional Neural Networks, University of Toronto.

[5] Attention Is All You Need, Google Research.

[6] Emerging Properties in Self-Supervised Vision Transformers, Facebook AI Research..

[7] The Neural Network Zoo, The Asimov Institute.

[8] Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, Google Brain.

[9] The Unreasonable Effectiveness of Data, Google Research.

[10] MLP-Mixer: An All-MLP Architecture for Vision, Google Research.

[11] Graph Node Embeddings Using Domain-Aware Biased Random Walks, Synaptic AI.

[12] Core Knowledge, Harvard University's Laboratory for Developmental Studies.

[13] On the Measure of Intelligence, Google Research.

14 Making AI Work with Small Data, Landing AI.

15 Off Road, But Not Offline: How Simulation Helps Advance Our Waymo Driver, Waymo.

16 Synthetic Data Generation, Kinetic Vision.

17 True Digital Twin, Kinetic Vision.

18 RarePlanes: Synthetic Data Takes Flight, AI.Reverie.

19 Defense and Government Data Generation, Labeling, and Enhancements, AI.Reverie.

20 Underspecification Presents Challenges for Credibility in Modern Machine Learning, Google Research.

21 How Many Stars Are In the Universe?, Space.com.

22 Number of Possible Subsets of All Sizes, Stack Overflow.

23 Numbers of Zeros in a Million, Billion, Trillion, and More, ThoughtCo.

24 Can You Fake It Until You Make It?: Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness, University of Toronto.

25 Diving Deep Into Fair Synthetic Data Generation (Fairness Series Part 5), MOSTLY AI.

26 Nationwide Unlocks Rapid Innovation With Synthetic Data, Hazy.

Note 1 Roots of the Word "Maverick"

Derived from the name of Texas rancher Samuel Maverick and his steadfast refusal to brand his cattle, "maverick" connotes someone who willfully takes an independent — and frequently disruptive or unorthodox — stand against prevailing modes of thought and action.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Hype Cycle for Data Science and Machine Learning, 2020