AXA VENTURE PARTNERS

White Paper

by Arthur Brasseur, Senior Associate

The Synthetic Data revolution: How does it fuel AI?

New York - London - Paris

www.axavp.com

Introduction

Ever since the early beginnings of civilization, people have consistently leveraged data for better decision-making or to secure a competitive edge. Today, data is the foundation of the modern economy and considered to be “the world’s most valuable resource” according to *The Economist*. The European Commission estimates indeed that the data economy will be worth $1tn in Europe by 2025, representing 6% of regional GDP.

While data, and especially, “big data”, was perceived as “the new oil” by Clive Humby in early 2000’s, recent developments in AI, “the new electricity”, have helped to unlock value out of it. According to PwC, this could create an additional $16tn of global GDP by 2030.

Real-world data is always the best source of insights and used as a fuel for AI. However, it is in fact often expensive, imbalanced, unavailable or unusable due to privacy and regulation.

This is where synthetic data comes in.

This whitepaper will divide into three main topics:

1. Defining synthetic data – benefits and risks for AI
2. Mapping the market – dynamics and key players
3. Assessing the opportunity – what it holds for entrepreneurs and investors

I. Defining synthetic data – Benefits and risks for AI

What is synthetic data?

Synthetic data is a simple and elegant concept that seems almost too good to be true. In short, this technology enables data generation on demand, at any volume, and can be tailored with precise specifications.

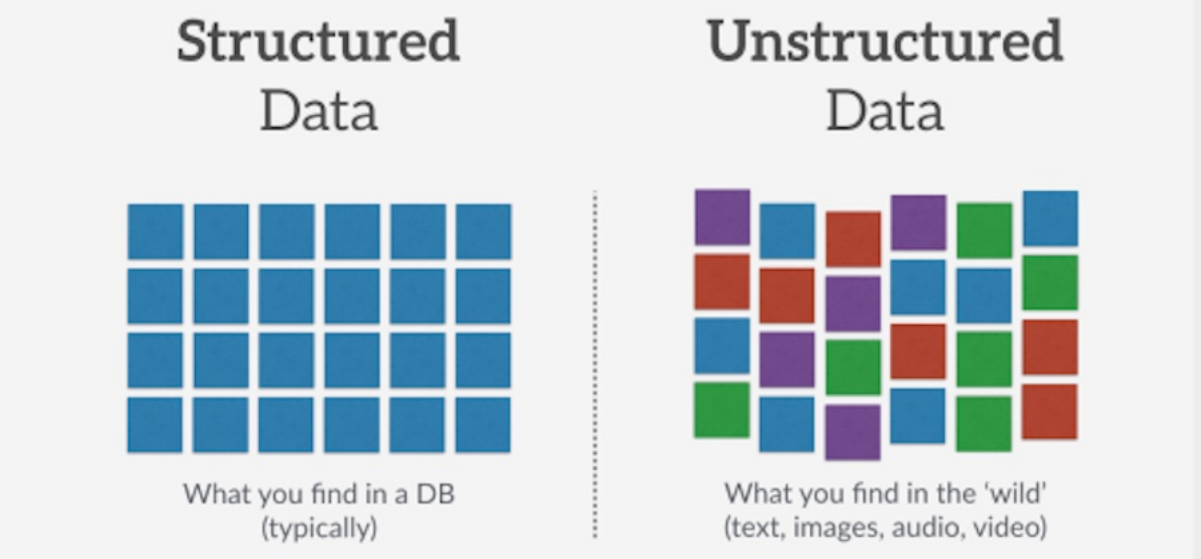
It is obtained by generating artificial data (from computer simulations or algorithms) that keep an original dataset's statistical properties and distributions, thus reflecting real-world data. This augmentation technique can be used instead of or in addition to original data to improve AI and Analytics projects and to solve different data-related problems.

Two types of data can be synthesized:

- **Structured data** – this has a standardized format, typically tabular with rows and columns: text, numerical data, time series, event data, etc.

- **Unstructured data** – this has an internal structure but does not contain a predetermined order or schema: images, audio, videos, 3D assets, etc.

Synthetic Data can be split into two categories: Structured Data or Unstructured Data



Few generative methods are used today:

- The simplest ones are based on **statistical distribution** (generated by Monte-Carlo methods) or on **an agent to model** (to fit actual data to the known distribution of data). These mostly apply to structured tabular data.

- The most sophisticated ones include **neural network techniques** like **VAEs** (Variational Autoencoders), **GANs** (Generative Adversarial Networks) and **diffusion models**. These can handle much richer distributions of data than traditional algorithms such as decision trees and can synthesize unstructured data like images and video.

What are the key benefits?

As companies start to accelerate their AI adoption within their business processes, they face more and more issues with the data needed for the models, including:

- Governance aspects like access and sharing of privacy sensitive data
- Poor quality preventing successful model outcomes
- Retention problems

Compared to classic anonymization techniques (including pseudonymized data), synthetic generation fixes these challenges. **Five key drivers** of corporate adoption have been identified:

1. **Privacy:** Data protection laws are strict globally (especially in Europe), notably in critical industries (governments and defense, financial institutions, healthcare, energy). A lot of companies struggle to unlock value from data they collect as they are under the scope of regulation. Synthetic generation can be key in this case,

preserving the structure of original data, but amending sensitive information, improving de-identification and creating data sandboxes to enable easy sharing;

2. **Volume:** Sometimes, companies do not have enough data for their use case. Synthetic generation is very effective in this case to augment the datasets with the generation of endless variations of labeled data to prevent data scarcity, improve ML models' and solve imbalance problems;

3. **Cost:** Collecting, labeling and validating real-world data into usable data is very expensive – large companies need armies of data scientists to do that. The ROI is clear for businesses using AI: projects costs related to data collection (especially when it is unstructured) and processing are reduced immediately;

4. **Inclusivity:** A real-world dataset can be biased and unfair towards a certain group of people, and could result in discriminatory AI systems. For instance, a healthcare AI system trained predominantly on data from one gender may not perform as well for the other gender. This inequality in AI systems can have severe consequences, including exclusion and discrimination. Synthetic AI data generation can help reduce AI bias just by generating artificial data or modifying the original data before it enters the AI / ML algorithms. This ensures a more inclusive and fair AI application.
   However, this technique should be used with caution! It is key to keep in mind that synthetic data is made in a computer instead of coming from actual events, leading to potential fake representations of the world. To this regard, its usage should be as minimized as possible.

5. **Testing:** It improves and automates software testing, for instance in data migration cases or as test data generation.

What are the risks and limitations?

"*With great power comes great responsibility*" – a classic! Synthetic data is not magical, data scientists training models with this method must be cautious and aware that it can lead to false results, implying a few risks:

1. **Lack of realism and limited generalization:** The complexity and nuances of real-world data might be inaccurately synthesized, leading to poor performance of AI models with real-world scenarios.
   - Without careful generation of synthetic data, the data may indeed introduce biases (e.g. not capturing anomalies) into the training data, resulting in faulty AI models.
   - There is therefore a necessity for extra data validation phases, which could add complexity to data pipelines and prolong the overall training process.

2. **Model degradation:** Training a model using synthetic data only, without a periodical synchronization to underlying real-world data, could cause the AI model to degrade quickly and collapse over time.
    - Synthetic data may indeed lack the diversity and distribution of actual data, exaggerating imperfections – this could result in lower-quality training set data over time.
    - It is therefore critical to make sure synthetic data keep reflecting the characteristics of its real-world sibling set, using sound data-generation techniques.

    As discussed in the previous part, regarding its benefit towards inclusivity, synthetic data should be used with caution and minimized – you cannot rely only on fake data or it will get you even more biases in your model – finding the right balance between real-world data and synthetic data is key.

3. **Privacy:** Albeit designed to keep privacy by creating artificial data, there is still a risk that some synthetic data points might unintentionally resemble real individuals or reveal sensible information.

4. **Ethics:** Some applications in critical industries (e.g. medical diagnoses in healthcare) may raise ethical concerns due to potential risks from inaccurate models.

5. **Transparency:** As synthetic data gains broader adoption, business leaders may raise questions on the openness of the data generation techniques, especially when it comes to transparency.

Developers and data scientists will need to consider the impact of synthetic data on the training of the AI model and account for the potential missteps that could be introduced by a lack of data diversity or context.

II. Mapping the market – Dynamics and key players

Synthetic Data has a long history but surged with the AI & ML model boom

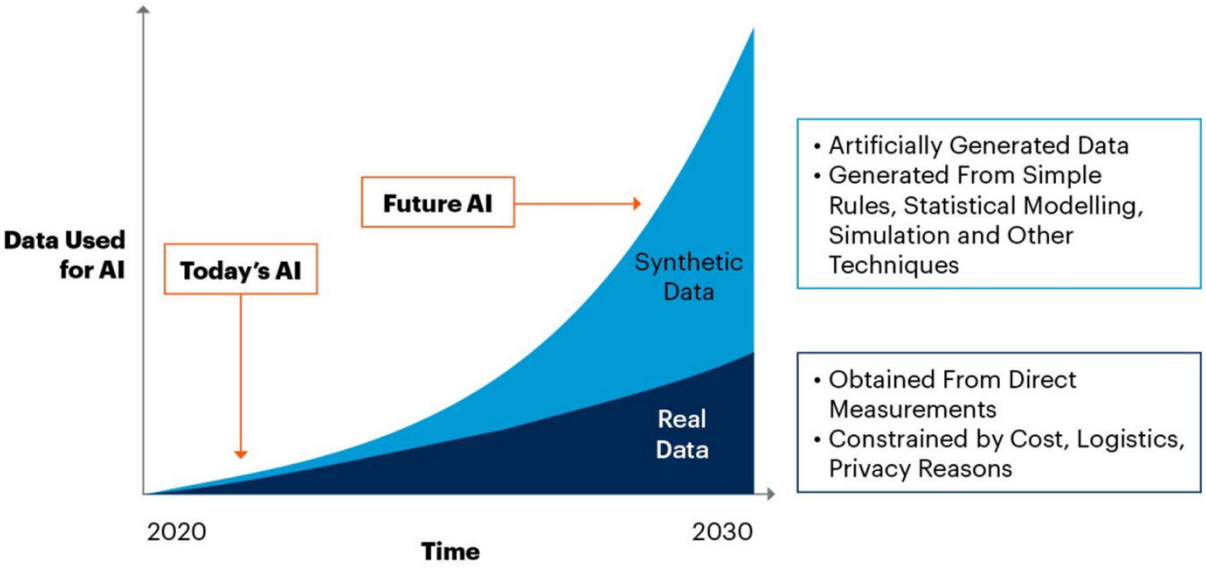
Synthetic Data is not a new idea. It has been around since the beginning of computer vision and way before today's golden era of deep learning, with examples of first artificial drawings dating back from the 1960s or 1970s. The idea of fully synthetic data was first put into practical application by Dean Pomerleau in 1989 when he tried to address the challenges associated with collecting vast amounts of on-road data for self-driving vehicles - Pomerleau ended up using image generation to simulate various road conditions. A few years later in 1993, the term 'synthetic data' was first formalized by Harvard statistics professor Donald Rubin who employed the concept to address undercounting and privacy issues in census datasets[1].

[1] https://www.aufaitai.com/data/synthetic-data-early-days/

However, it has recently triggered growing attention, fueled by advancements in AI & ML. These technologies require massive sets of data for model training, and synthetic data offers multiple benefits to that regard in terms of improved model accuracy, mitigation of privacy concerns and access to rare data instances such as credit card frauds or car accidents.

According to Gartner, "by 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated". It is even expected to become the main source of training data in AI models by 2030, which is well represented on the graph below[2].

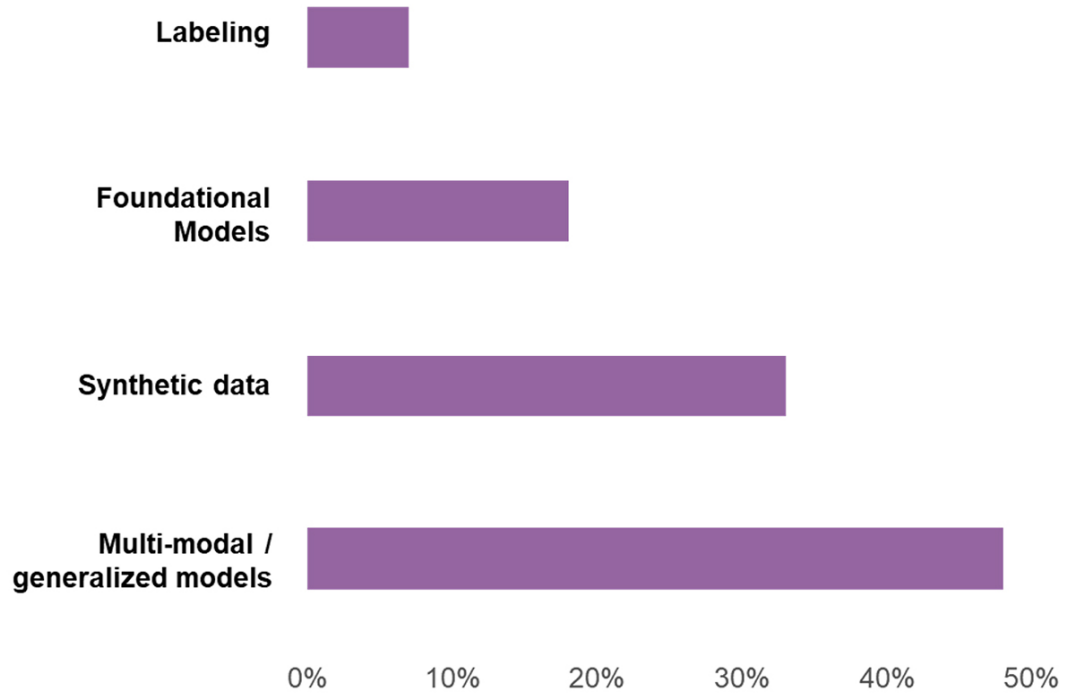
Synthetic Data will completely overshadow real data in AI models by 2030



Synthetic data appears indeed as a top priority for AI & ML teams within organizations. Indeed, the figure below shows that synthetic data situates just behind multi-modal/generalized models as one of the areas of AI/ML infrastructure that is expected to have the biggest breakthroughs in the next 3 to 5 years, according to a study by AI infrastructure alliance.

[2] As per Gartner

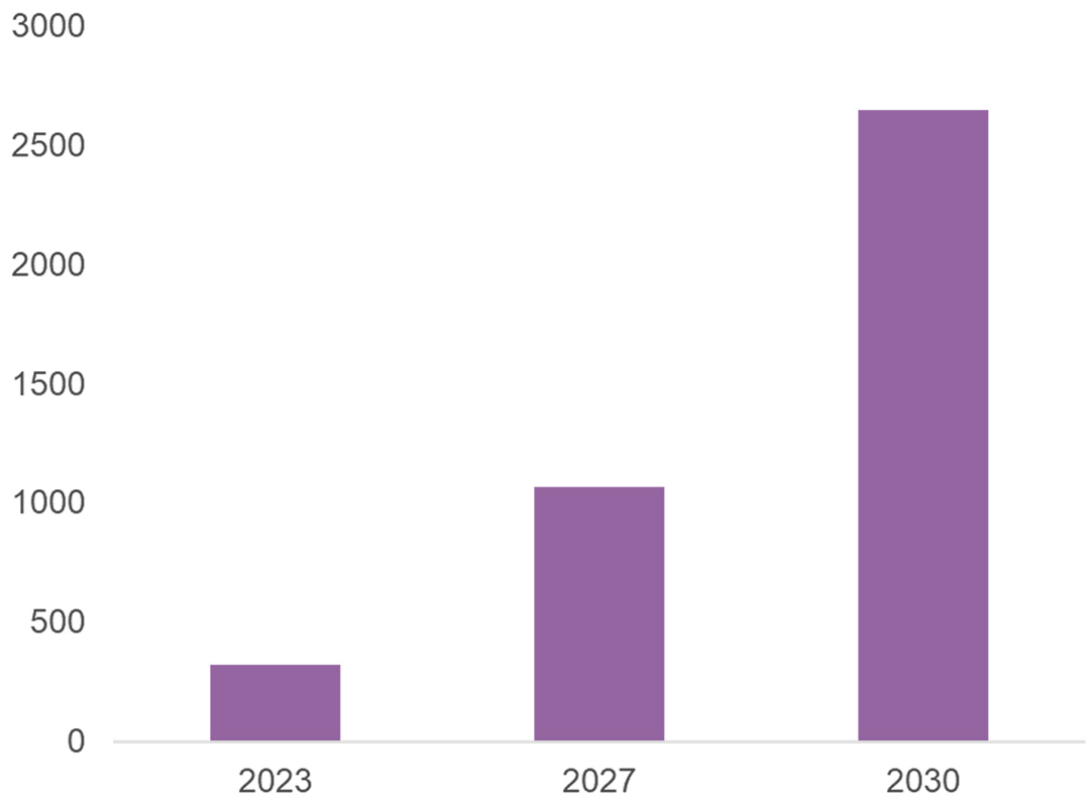
Areas of AI/ML infrastructure expected to have the biggest breakthroughs in the next 3 to 5 years



Source : AI Infrastructure Alliance

In fact, 2/3 of the companies that have synthetic data as their core business, that is 100 companies out of 148, were founded in 2017 or later (e.g. Mostly.AI, Gretel, Datacebo for structured data, Rendered AI, Synethesis AI for unstructured data). 2023 has even showcased a strong deal activity with 43 deals for a total of $262.4 million invested, at a median post-money valuation of $36.5 million. And this recent growth is not expected to slow down anytime soon. On the figure below, we can observe that while today still a native market, estimated at over $300m in 2023, it is expected to grow at a CAGR of 35% to reach more than $1 Bn in 2027 and more than double in 2030.

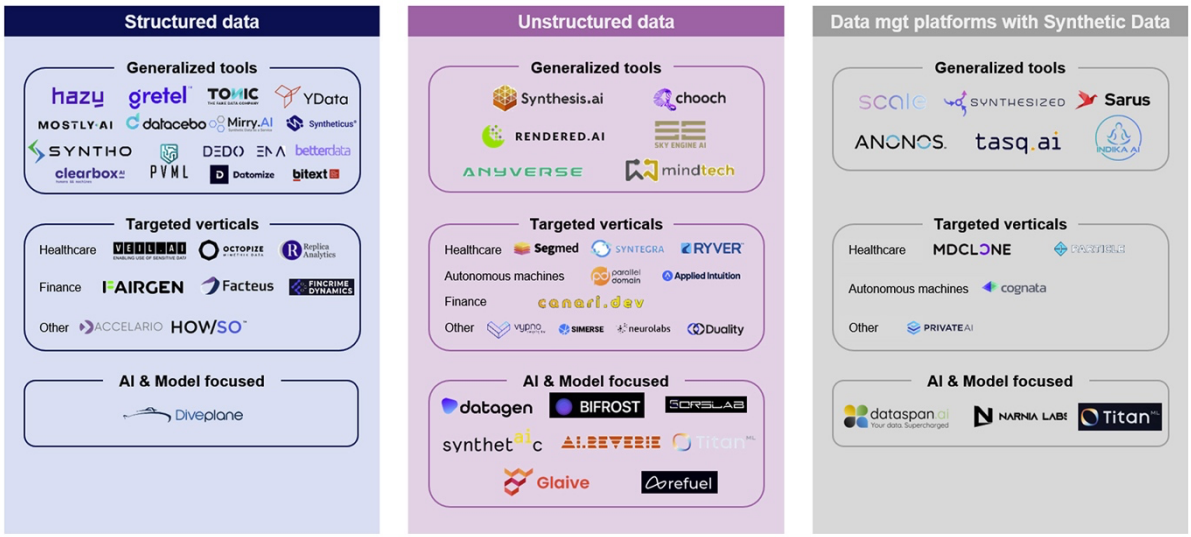
Synthetic data market size estimate ($M)



Source: Precedence Research

This is therefore inevitable: massive opportunities will come from it. As Ofir Zuk, cofounder and CEO of synthetic data startup Datagen put it, “we can simply say that the total addressable market of synthetic data and the total addressable market of data will converge”. While growing fast, synthetic data can be a complex landscape to navigate with more and more companies joining the field and offering various use cases. In order to better understand, we have decided to split it into three parts; structured data, unstructured data and broader data management platforms. The following market mapping collects data from Pitchbook, Crunchbase, LinkedIn.

Synthetic data market mapping



Source: Pitchbook, Crunchbase, LinkedIn

As a brief recap from before, Structured or Tabular data is data that is organized in a specific format or schema, typically in rows or columns. Examples of structured data include sales data, customer information or financial statements. It typically represents 20% of enterprise data.

The other 80%, however, are unstructured data, that is data that does not have a defined structure or format, such as images, audio and video. It is increasingly important in industries like gaming, retail or automotive, where computer vision and other advanced analytical tools are used to extract meaningful insights. It typically needs more storage.

As shown on the mapping above, the structured data segment is the one that captures most of the activity. These vendors target industries subject to high pressure from regulators like healthcare, or financial services, and often struggle to meet data privacy or compliance requirements. Typical use cases include simulation and prediction research in healthcare, fraud identification in financial services and public release of datasets for research and education purposes. For example:

- In healthcare, as pointed out by Mostly AI, a generalized tool for synthetic data, tabular data include electronic health records (EHR) of patients and populations, electronic medical records generated by patient journeys (EMR), lab results, and data from monitoring devices. These highly sensitive structured data types can all be synthesized for ethical, patient-protecting usage without utility loss. Synthetic health data can then be stored, shared, analyzed, and used for building AI and machine learning applications without additional consent, speeding up crucial steps in drug development and treatment optimization processes[3].

- In finance, some tools offering synthetic financial data like Mimic from Facteus can be useful for customers for (i) internal initiatives like analytics, machine learning and AI, marketing and segmentation activities, and (ii) new revenue streams through external data monetization. At the same time, consumer privacy

[3] https://mostly.ai/blog/healthcare-data-platform

information is protected from being exposed and the statistical relevancy of the data is maintained!

Structured data today however offers low tech defensibility, as some commercial tools today still rely on open-source technology which provides them with limited intellectual property and barriers to entry. On another note, while value privacy is the main use case advertised by these companies, it is unlikely to be a sufficient driver of adoption for enterprises. That is largely due because privacy preserving is a business need, while most platforms today focus on developers, which question the product-market fit and customer readiness for long-term adoption.

Moreover, tech leaders or large corporates are seeking to train their model internally using structured synthetic data. For example, Google's Waymo uses it to train its self-driving cars while American Express & J.P. Morgan are using synthetic financial data to improve fraud detection[4]. If these large organizations opt for in-house development, it raises a challenge around the market size. Finally, some structured synthetic data companies propose on-demand datasets for non-recurring usage, which can threaten the scalability and sustainability of their business model. As a result, we see more companies in the space trying to expand to multiple use cases, targeting different customers from software engineers to business developers. Likely, with broader data management platforms offering synthetic data as an extra feature, we are likely to see consolidation between vendors.

Unstructured data, on its end, is probably the most promising segment of this industry. Today, use cases are mostly focused on computer vision, such as edge cases in autonomous vehicles to improve object detection performance with startups like Applied Intuition. Another emerging application is in Natural Language Processing (NLP), especially since data can be limited and subject to privacy in such domains. Typically, Amazon's Alexa and similar systems leverage synthetic data to improve natural language understanding, enabling adaptability to new accents and aiding low-resource languages, with companies like Refuel, Titan, and Glaive exploring this avenue. However, providing training data for ML models in areas where traditional data is insufficient or non-scale also means greater technical challenges through complex generation methods, thus the need for vendors to precisely target use cases and industries with immediate and continuous need for data.

Finaly, the last type of companies we have identified gather the broader data management platforms (e.g. Tasq, Sarus, Scale) that, in addition to features like data privacy, model evaluation, computer vision, can also offer data enrichment with synthetic data. In the long term, we are likely to see broader data management platforms offer synthetic data, structured or unstructured, challenging pure-players and leading to more consolidation.

[4] https://www.statice.ai/post/types-synthetic-data-examples-real-life-examples

III. Assessing the opportunity – what it holds for entrepreneurs and investors

What are the most relevant applications?

1. It all started with AV

As we saw in our brief timeline in the previous chapter, synthetic data has been around for decades, with its theoretical underpinnings dating back to the 1960s. In the mid-2010s, however, technology showed solid commercial traction in the autonomous vehicle (AV) industry and is still a prominent use case today.

Synthetic data was meant to be in AV. It is in fact simply impossible to collect real-world driving data for every scenario an autonomous vehicle might encounter on the road – it would literally take hundreds of years! So, instead, AV companies developed simulation engines to generate a certain volume of synthetic data that would be large enough to produce millions of possible driving scenarios with infinite permutations: playing with the locations of other cars, their speed, adding or removing pedestrians, adjusting the weather, etc. This is why:

- Leading AV players (Waymo, Cruise, Aurora, Zoox) invested heavily in synthetic data and simulation as a core part of their technology stack;
- AV was one of the first use cases tackled by synthetic data startups like Applied Intuition (valued at $3.6bn in 2021), Parallel Domain and Cognata.

AV was the beginning of something bigger for synthetic data. Many entrepreneurs recognized quickly that this tech could be generalized and applied to many other applications indeed.

2. Computer vision

From robotics to agriculture, from healthcare to defense and security, computer vision has found a large range of valuable applications in recent years. And for these use cases where input data is constantly changing and needs to be dealt in live, building AI models requires massive volumes of labeled image data.

Synthetic data represents a powerful solution here. But how is it possible to artificially generate such high-fidelity, photorealistic image data? Three important research advances have made this possible:

- **Generative Adversarial Networks (GANs)** – Invented by AI pioneer Ian Goodfellow in 2014, this method, based on two neural networks (the "generator" and the "discriminator"), aims at generating synthetic photos that are indistinguishable from real ones
- **Diffusion models** – With meaningful advantages over GANs and serving as the technological backbone of DALL-E2 (Open AI text-to-image model), these are expected to play an increasingly prominent role in Gen AI

- **Neural Radiance Fields (NeRF)** – A powerful new method to quickly and accurately turn 2D images into complex 3D scenes, which can then be manipulated and navigated to produce diverse, high-quality synthetic data

It is clear: computer vision applications require deep technical knowledge. And this is a huge opportunity for verticalized players (e.g. Datagen or Synthesis AI) to capture solid market shares.

3. Language (NLP)

Synthetic data is a game-changer for computer vision, but the technology will be even more ground-breaking for another use case facing well-known issues (bias, hallucinations, output reliability): language.

Indeed, while language models work properly for generic themes, they struggle to capture nuances in semantics that are important in domains such as legal. Synthetic data has a key role to play here and here are two examples to illustrate:

- **Anthem**, one of the largest health insurance companies worldwide, powers AI applications like automated fraud detection and personalized patient care with proprietary data from medical records and claims. Last year, the company announced a partnership with Google Cloud to generate large volumes of synthetic text data (medical histories, healthcare claims) to improve and scale these use cases. And, cherry on the cake, this addresses the data privacy concerns that have held back the development of AI in healthcare for years.

- **Illumina**, the world's leading genetic sequencing company, announced a partnership with Bay Area startup Gretel.ai to synthesize genomic datasets (which are textual data btw). This is a revolution as it can replicate the characteristics and signal of real genomic datasets while being compliant with data privacy regulations. And, another cherry on the cake, this enable researchers to develop a deeper understanding of disease, health and how life itself works.

A lot of promising companies like Gretel.ai, DataCebo, Syntegra or Tonic.ai have emerged in recent years to help companies tackle such use cases. While these businesses, focused on structured text, use statistical methods and traditional ML, an opportunity exists to build next-generation synthetic data, focused on unstructured text, by using LLMs to get unparalleled realism, originality, and diversity.

A few startups have recently emerged to provide such LLM-generated synthetic data like Discus AI and Refuel.ai, but we are still super early in development.

4. Verticalized applications

Generally, in addition to autonomous machines (AV and robots), synthetic data is becoming increasingly important across various industries, notably:

- **Healthcare**: It enables the simulation of patient data for clinical trials and the creation of biophysiological simulations. While ensuring patient privacy, it

optimizes treatment plans and personalizes medicine. Examples include MDClone and Syntegra.

- **Retail**: It is used to simulate shopper behavior, train autonomous checkout systems and improve inventory management. Enabling many use cases (e.g. demand prediction or churn model improvement), this leads to more personalized and efficient customer experiences while maintaining customer privacy. Examples include Bitext and Statice.
  Finance: Various use cases can be cited in the sector: risk assessment, fraud detection, algorithmic trading. Simulations on realistic transaction data, account balances and credit histories can be used to train AI models that identify suspicious activities, optimize investment strategies and ensure stability and compliance. Examples include Canari.dev and FinCrime Dynamics.

What are the key takeaways?

Clearly, there are multiple use cases, but what should we retain from it? Well, at AVP, we see the evolution of the market in 3 ways.

First of all, the fragmented structured data segment is poised to consolidate. Despite emerging in 2018, tabular synthetic data companies still struggle with product-market fit challenges because it is too “basic” and can be done by internal teams. Offering data privacy as a primary value proposition, they face obstacles related to tech defensibility, customer adoption and business models, limiting the potential for long-term growth. Vendor consolidation has already started with recent acquisitions of Statice by Anonos and Logiq.AI by Apica.

Separately, we believe verticalized companies in the unstructured data segment that own the end-to-end data pipeline will gain more traction as they will be experts in addressing the full customer value chain, from dataset creation to annotation, post-processing and curation for post-model deployment. Since data requirements are too complex and sometimes niche, we do not expect to see one vendor covering all data types for all verticals. Rather, we might observe more and more specialized vendors popping off such as Segmed or Particle in healthcare.

Finally, we believe some of the most successful companies in the field will not only provide synthetic data but will also offer foundational models that are trained on it. We believe this more complete offer might win the market, capturing more value for enterprise AI in the long run.

Conclusion

Without a doubt, data is the lifeblood of AI and synthetic data will be tied to its rise. It can well provide a helpful complement to real data, providing access to better annotated data to build accurate and extensible AI models.

However, synthetic data generation methods are not perfect yet (on quality, fidelity scalability, transparency) while it needs to answer this simple question: *"Can it be accurate enough to substitute with for real data?"*

Innovation is key to bridge the gap between simulation and reality. And there have been a lot of technological advances, including the introduction of massive foundation models from companies like Open AI and Mistral AI, the transformer architecture, the multimodal data generation, photorealistic game engines, diffusion engines, etc. Recent progress in LLMs has participated in the rise of synthetic data that is now often indistinguishable from human data.

To succeed, it's crucial for synthetic data pure players to ensure that their product integration aligns seamlessly with the intended use case and demonstrates a clear ROI (increase in model productivity and predictability, cost savings for ML engineers and data scientists, etc.).

Due to low barriers to entry as based on open-source technologies, companies in the structured data segment might suffer from lack of tech differentiation, leading to vendor consolidation in the sector. As to companies manipulating unstructured data, the most prominent use cases are in computer vision and NLPs offerings. However, as data must be specific and is sometimes scarce, companies will have to overcome complex generation methods and the ones the most likely to succeed will be those targeting at specific industries or verticals. We therefore expect synthetic data market to be growing, though the greatest opportunity might come from companies that will not only generate but also train their own models. As such, they will be able to propose an end-to-end data solution, tailored to customers' needs, favoring a better customer experience and long-term adoption.

As a result, as synthetic data technology will reach its full potential in AI, this calls into question the defensibility implied by data. While datasets become deeper and more unique as a business gains more customers, synthetic data accessibility will make differentiation much harder in early stage, forcing companies to outperform on model rather than relying on a supposed data edge.

Getting a technological edge in AI is more and more difficult as well though – even open-source algorithms are complicated to overcome with the same dataset. AI companies will need to pay attention to each and every detail to make the difference on the market, addressing their customer needs with the finest precision on the back of superior, stellar, tech teams.

Bibliography

- https://www.clearbox.ai/synthetic-data
- https://www.forbes.com/sites/forbestechcouncil/2023/11/20/the-pros-and-cons-of-using-synthetic-data-for-training-ai/
- https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/
- https://research.ibm.com/blog/what-is-synthetic-data
- https://mostly.ai/synthetic-data/what-is-synthetic-data
- https://mostly.ai/blog/how-to-benchmark-synthetic-data-generators
- https://mostly.ai/blog/synthetic-data-companies
- https://www.eweek.com/artificial-intelligence/best-synthetic-data-software/
- https://files.pitchbook.com/website/files/pdf/2023_Emerging_Space_Brief_Synthetic_Data.pdf#page=1
- https://www.aufaitai.com/data/synthetic-data-early-days/
- https://www.altexsoft.com/blog/synthetic-data-generation/
- https://www.turing.com/kb/synthetic-data-generation-techniques
- https://techcrunch.com/2022/05/10/the-market-for-synthetic-data-is-bigger-than-you-think/
- https://www.ibm.com/blog/synthetic-data-generation-building-trust-by-ensuring-privacy-and-quality/
- https://medium.com/datafabrica/exploring-synthetic-data-use-cases-6114935a54d1#:~:text=Several%20companies%20also%20use%20synthetic,to%2Dtext%20applications%20and%20more.
- https://ai-infrastructure.org/wp-content/uploads/2023/10/AIIA-Landscape-October-2023.pdf
- https://www.manot.ai/blog/the-future-of-synthetic-data-generation-with-manot

We invest in great entrepreneurs.

We support outstanding companies.

New York - London - Paris

www.axavp.com