endava

Unleashing the Power of Synthetic Data and Machine Learning

How synthetic data supports machine learning training to help organisations make better decisions, improve efficiency, and find new market opportunities.

www.endava.com

This guide introduces the rapidly growing field of synthetic data, and how it solves the common challenges of training machine learning models.

It illustrates the value of visual synthetic data, and the business benefits it can deliver for organisations across industries.

This guide also highlights the importance of a trusted partner to help businesses sidestep the barriers of adoption so they can unleash the power of synthetic data and machine learning.

The synthetic data opportunity

For decades, businesses have been using artificial intelligence (AI) and machine learning (ML) to gain a competitive advantage. Today, it has an astonishing array of applications, and the scope for use cases is broadening fast.

But there is a problem: the amount and quality of data we use to train ML models has not caught up. That's because this data all comes from the real world. And in many cases, its sophistication, availability and reliability are inconsistent.

Synthetic data offers a solution. By artificially generating data that simulates the properties of real-world occurrences, we can overcome today's data challenges. It's no surprise that synthetic data is predicted to grow by more than 30% over the next seven years1.

However, there is one more problem: while the use cases keep growing, and the product-focused synthetics market expands, few companies can develop the in-house capability to deliver synthetic data due to widening gaps in funding, resources, and expertise.

Read on to find out how you can access the value of synthetic data and enhance your AI and ML capabilities.

Machine learning opens new value streams

An introduction to artificial intelligence and the importance of high-quality data to machine learning.

Before talking about synthetic data, we need to explain the vital role of AI and ML.

AI refers to technologies that enable systems to copy human cognitive functions, such as seeing, reading and interpreting data, before making recommendations.

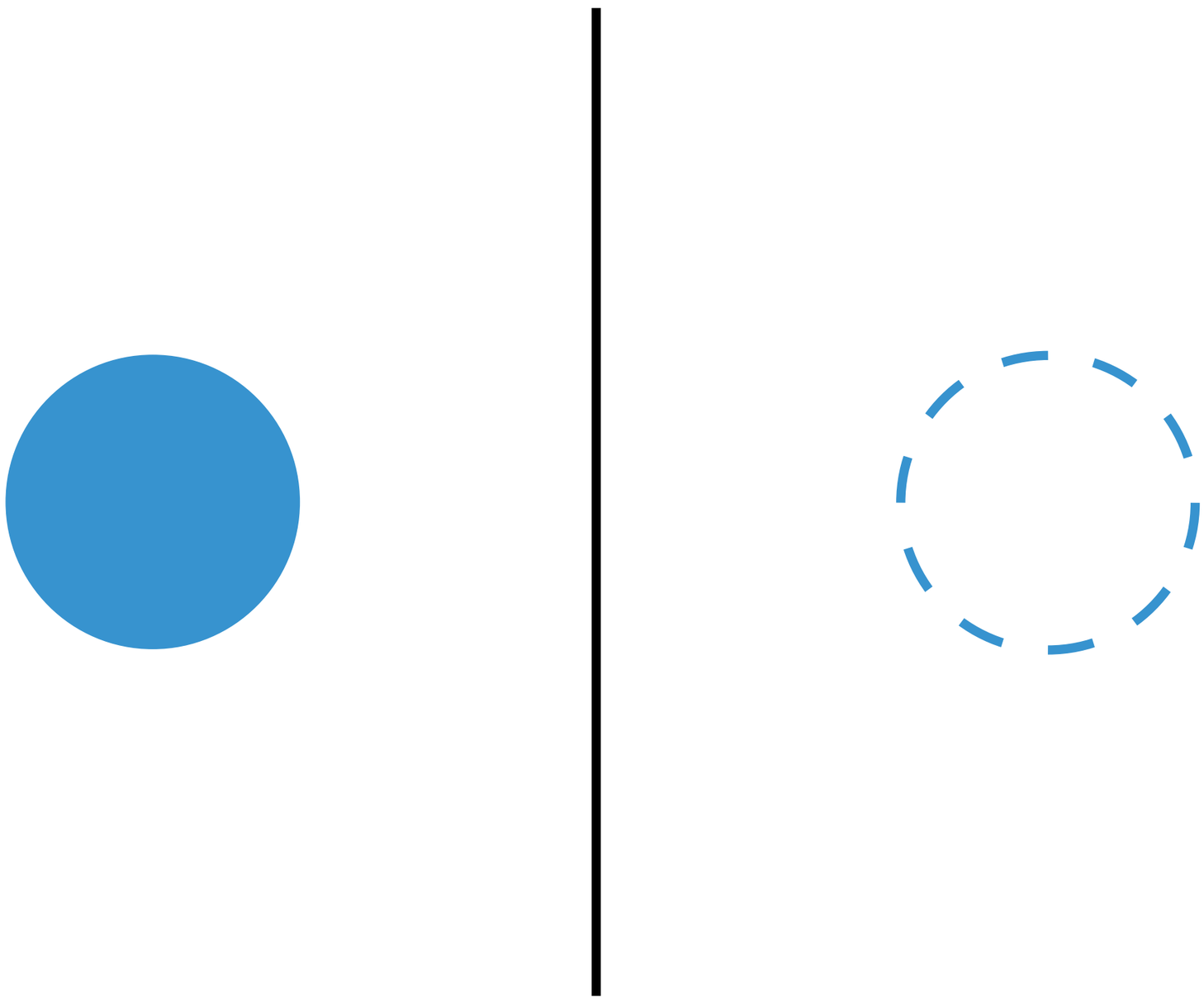
ML is a subgroup of AI that enables systems to learn from the data they receive. ML uses algorithms to analyse large amounts of data and improve its cognitive functions over time.

AI enables faster, cheaper and stronger problem-solving and more informed decision-making while opening new, previously unreachable value streams.

It does this in two ways.

First, AI enables computers to use a wider array of inputs/outputs and be more interactive – giving computers the power to visualise imagery, listen and read natural language, then answer back in a similar way that humans would. These capabilities help businesses automate processes and decision-making, reducing operational costs and improving scalability.

Second, with this improved understanding, we can present information in an enhanced way: by augmenting data with contextual relevance – for example, identifying trends by looking at a line in a monthly report, automatically fetching past reports, and comparing the documents. Algorithms can also make use of inputs from visual and spatial sources and be deployed at scale.



AI works like our brains; it recognises patterns after repeated exposure to help better define concepts. But it needs far more data to learn than the human brain. More to the point, ML needs high-quality data to train the model and validate how well it behaves.

Because ML lives off training data that empowers it to learn efficiently, systems work best when driven by top-flight information that provides the full picture. To make ML work well, we need to feed it the right data.

This is where the problem of real-world data arises, and where synthetic data steps in to solve it.

An introduction to synthetic data

An introduction to artificial intelligence and the importance of high-quality data to machine learning.

Synthetic data is artificially generated ML training data that mimics the characteristics of real-world phenomena. This data simply doesn't exist until we create it. Visual synthetic data – the subject of this guide – is a subset of this vast area and refers to images, videos, and 3D model data. The are other types of synthetic data: audio, text and tabular data.

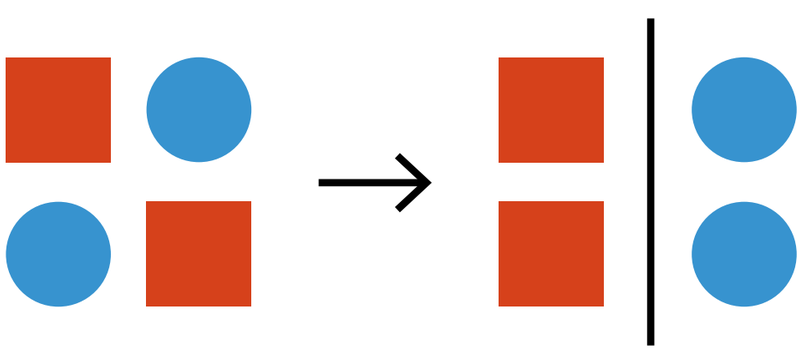
Visual synthetic data enables scenarios where it's impractical, complex, or simply impossible for humans to label and annotate images from the real world accurately and efficiently. More accurate data tagging is crucial for creating precise and diverse learning for ML models.

Skilled artists craft synthetic data using visual effects tools to simulate realistic, high-quality images and environments. The outputs are digital doubles that represent how subjects of interest look and behave.

In this way, we can think of visual synthetics as the point where visual effects meet data science. It's this new field which combines two skills – technical artistry and ML expertise – that has many wide-ranging applications and multiple business benefits.

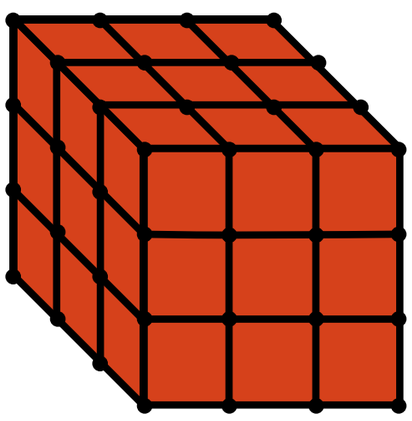
Synthetic data processes – the machine learning training journey

These are the most common computer vision tasks; the fundamental building blocks for how computers extract meaning from visual data (imagery):



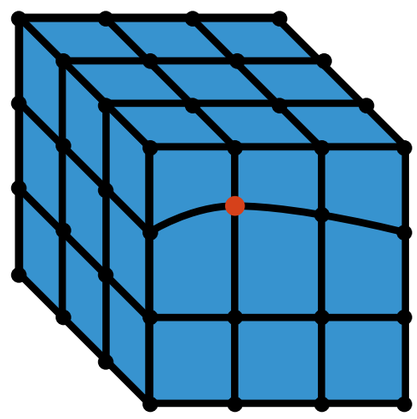
Classification: The neural network assesses the contents of the image and categorises it using predefined labels.

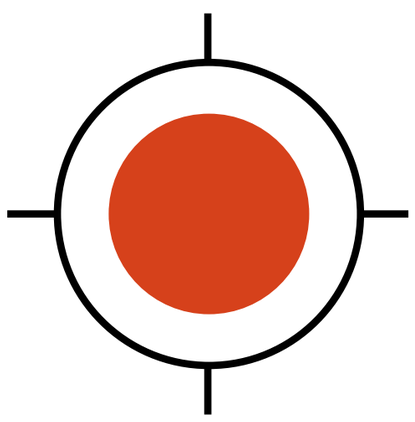
Detection: The system learns to identify the general region or "bounding box" (i.e., a geometric shape that encloses an object to define its location and size) within the digital image.



Segmentation: The computer defines a polygon encompassing an object's shape with more detail and precision. This keeps less relevant background information out of the annotation, allowing for more detail in complex shapes.

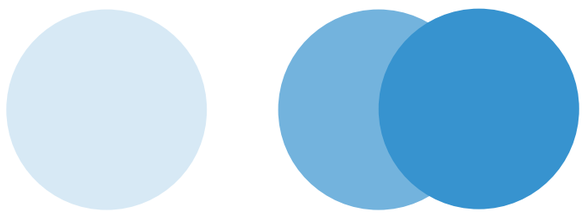
Understanding: The network can identify specific intrinsic properties of an object, such as the presence of a defect.





Targeting: The system can comprehend an object's position within 3D space.

Tracking: The computer can interpret an object's motion properties.



In many cases, these essential computer vision tasks provide value directly in automatic decision-making. For example, a collection of algorithmic systems in an autonomous vehicle could perform these tasks to detect a pedestrian and brake.

Or an emergency rescue system camera could identify a forest fire in the corner of the image and dispatch firefighters to the exact location. Both applications add value by saving time, resources – and even lives.

In other cases, advanced tasks could recognise exact pressure values by reading a pressure gauge or computing a person's gaze direction based on key points detected within a face.

Crafting synthetic data with visual effects tools – an example

Popular tools such as Houdini, Maya and Blender (commonly used to produce 3D visuals for film, TV and gaming) are often the primary starting point for creating synthetic data. With these tools, we can produce photorealistic renderings that accurately represent the real scenario, and then embed them with the ground truth (GT) information that the ML model will use to learn.

Images act as reference points; any GT data can be computed mathematically, and this ensures accuracy. Relevant aspects can be controlled, while potentially important remaining aspects can be randomised.

Other types of measurements can always be simulated by computation because the 3D scene is the real asset (X-rays, lidar, IR, and near IR are on the table); akin to what movies and games do with 3D in terms of repurposing.

How does it arrive at this conclusion? AI/ML recognises what patterns are relevant to define a concept. With synthetic data, you can explicitly craft and refine what you want to teach the model in a controllable way.

Gartner estimates that by 2030, the majority of the data used to build models will be synthetic data.

“The most valuable data will be the data we create, not the data we collect.” [1]

$2,339.8 million - The projected value of the global synthetic data generation market in 2030, a CAGR of 31.1%[2]

1 **Gartner Data & Analytics Summit 2022 Orlando: Day 1 Highlights**, Gartner, August 2022.
2 **Synthetic Data Generation Market**, Fortune Business Insights, June 2023.

The real-world data problem – and why synthetic data is the solution

Synthetics helps us resolve issues with the data gathered from the real world.

Many organisations enhance solutions with ML to improve decision-making, automate tasks and create new revenue streams. But ML models are only as good as the data you feed them, and therein lie several big problems which all ML teams struggle to overcome.

Real-world data is often skewed, hard to obtain and expensive to procure – or it simply doesn't exist. And because most of the use cases relevant for business are human-centric, most of the time the data relates to individuals, so the use of this data is governed by strict privacy regulations. There are other problems, too.

Here are the main challenges of using real-world data in ML model training, and how synthetics can help you avoid each one.

Filling in rare data gaps

Organisations in several industries, such as healthcare insurance, rely on visual data about rare events like uncommon conditions, which means there's a lack of available datasets; this low volume isn't enough to train an ML system. Synthetically generated imagery of these rare cases can fill in these gaps, enabling businesses to train ML vision models in ways that would not otherwise be possible.

Protecting personal privacy

Building ML models for companies in many industries – including automotive, retail and healthcare – can be a long and complicated process. That's often because data needs to be cleaned or anonymised to remove any personally identifiable information, such as medical history. Synthetic data allows these organisations to create visual data not associated with real individuals.

Ensuring data precision

In sectors like manufacturing that deal with millions of data points, human error is common. With synthetic data, analysts can create pixel-perfect annotations, maximising efficiency and unlocking new capabilities, such as improving the accuracy of anomaly detection systems.

Visualising products before prototyping

Often in a hardware production cycle, physical devices may not exist when the data is needed. With synthetic data, you can produce accurate representations of future products and the data that they will produce, unblocking algorithm development and better informing the hardware prototyping phase.

Adapting training model data

Because synthetic data can be replicated on demand, training data can adapt to project specifications as requirements change. This reproducibility also enables targeted parameter refinement to produce the most effective training set for a scenario. When training data models, you can also turn certain features on or off to see how they affect evaluation metrics, meaning you can see the role of that feature in the decision. This is an important capability in understanding bias, running ablation studies, and explaining how models come to a decision.

Avoiding biased data

Collecting data from a biased sample, such as a single geographic region, means your contextual insights won't provide a truly global representation. But this is often the only data that's available in many industries. With synthetic data, you can generalise the data to create a robust, flexible model that's trained on an unbiased collection of diverse data, giving you explicit control over distribution parameters.

Validating model performance

Collecting enough real-world data to validate model performance can be challenging for organisations in many industries. Synthetic data can be used to validate edge case scenarios in existing models, enabling a more proactive approach to model iteration.

What you can do with synthetic data

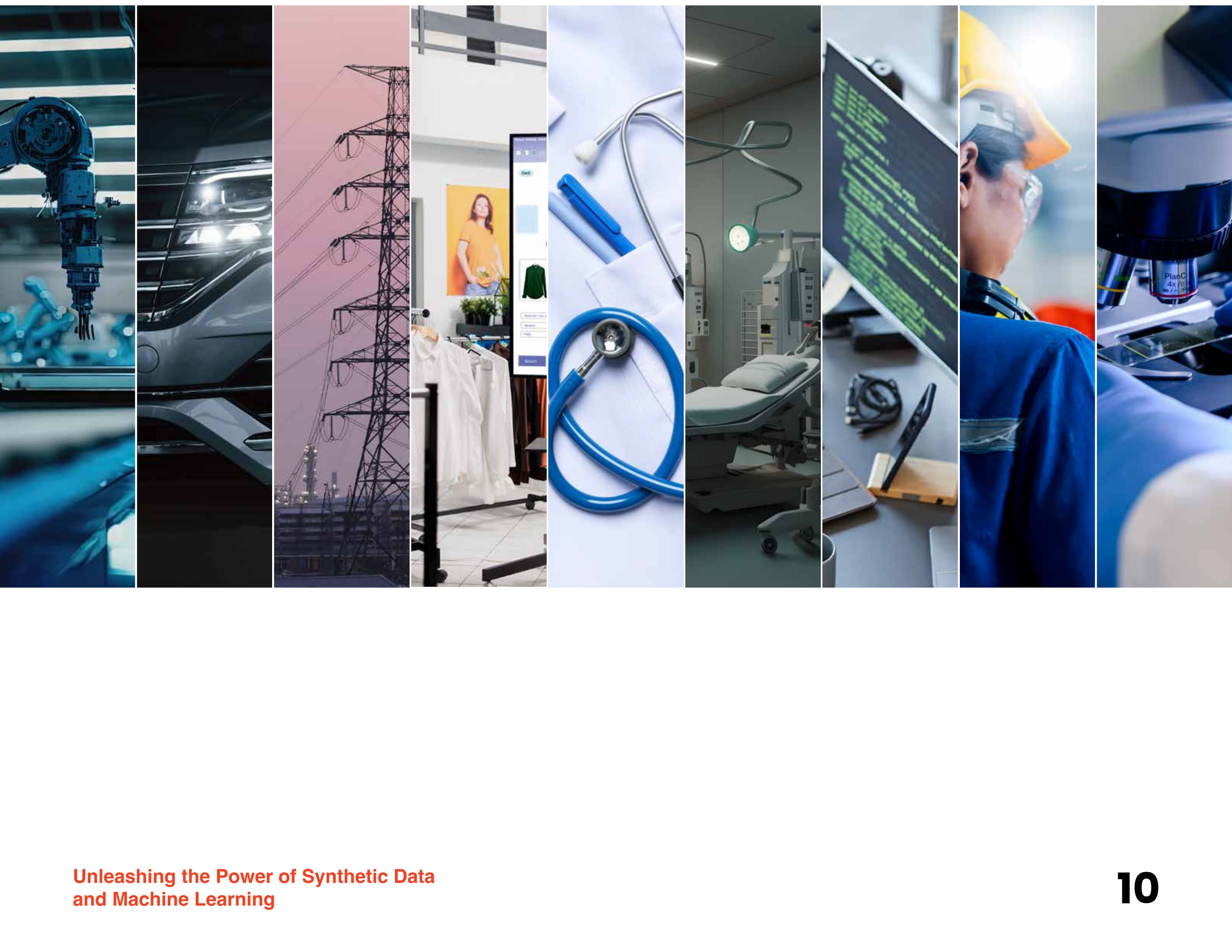
Using synthetic data for training ML models helps you overcome the challenges of real-world data because it enables you to:

- **Enhance visual models** by improving their performance and reliability through training with diverse and realistic synthetic datasets.
- **Ensure superior visual experiences** by refining visual interactions under various conditions.
- **Comply with visual data privacy standards** by using realistic, non-sensitive data not associated with real individuals.
- **Unlock innovative feature applications** by expanding capabilities beyond the constraints of real-world data.
- **Accelerate development cycles** by swiftly iterating and refining new features and enhancements – reducing your reliance on acquiring and labelling cumbersome real-world data.

Fewer costs and more opportunities across industries

By solving the problems of real-world data, synthetics brings many industry-specific business outcomes:

- Reduced quality control costs for **manufacturers**
- Safer and more reliable autonomous vehicles for **automotive** companies
- Lower asset management costs for **energy** companies
- Higher product availability and lower returns costs for **retailers**
- Improved patient monitoring and engagement for **US healthcare** organisations
- Enhanced operations and patient flow in **hospitals**
- Higher-quality product capabilities for **technology** companies
- Greater reliability and functionality for **robotics** companies
- Smarter, lower-cost pollution reduction for **science** organisations



Synthetics applications across industries

Synthetics and ML can help organisations overcome a range of industry-specific challenges so they can bring several high-value benefits.

Most organisations are always looking for new ways to cut costs, improve efficiency, scale operations and extend into new markets. Today even more so. By using synthetic data to train ML models, they can achieve many of these strategic objectives – and achieve a range of other business benefits in the process.

Here are just some of the applications across industries as diverse as manufacturing and healthcare insurance.

Manufacturing – improving quality control and reducing waste

Quality assurance is critical in manufacturing, and anomalies and defects must be identified quickly and accurately. However, these incidents are relatively infrequent, making the data a rarity.

By generating synthetic images representing various scenarios and defect types, ML models can be trained to recognise quality issues.

This improves the accuracy of anomaly detection systems, reducing quality control costs and increasing process scalability. And, by introducing quality control checkpoints early in the pipeline to catch defects promptly, manufacturers can also reduce waste, which helps them achieve their sustainability goals.

Automotive – enhancing safety and compliance

Developing safe and reliable vehicles requires extensive data to train, test and validate ML models. However, collecting real-world data about diverse driving conditions, scenarios and environments is expensive and time-consuming. Auto manufacturers also need a deep understanding of driver behaviour – something that's hindered by data privacy concerns.

By generating synthetic images representing various defect types and driving scenarios, manufacturers can train ML models, which will improve system performance especially in edge scenarios, which are difficult to find and capture in the real world. Synthetic data can also generate imagery representing diverse driver demographics and behaviours to develop ML solutions that don't infringe on customer privacy.

This improves quality and lowers costs for developing and validating vision and spatial capabilities. It also improves manufacturing operations by introducing quality control checkpoints in parts of the assembly lines where it's too dangerous for humans to perform inspections.

Energy companies – improving asset management

Optimising equipment lifespan and performance while predicting asset failures is a constant challenge for energy companies due to the often harsh and hard-to-reach environments they deal with. This makes manual maintenance hazardous, expensive and labour-intensive. Defects and anomalies are expensive and dangerous, especially in the oil and gas sector, so AI is often used by these companies to improve operations and perform preventative maintenance.

Synthetics can create representative failure scenarios enabling the development of robust predictive maintenance models. This enhances maintenance schedules to improve operational efficiency and reduce costs, while extending equipment lifecycles and reducing waste. It also helps increase safety and compliance.

One application of this is where autonomous legged robots or flying drones are sent into dangerous environments to perform frequent and repetitive equipment inspection, checking gauges and values for example, to detect degrading trends as early as possible.

Another application is for electricity distribution companies which use satellite imagery to identify low-carbon technology use, such as solar power supplies or electric vehicles. By aggregating this data and checking trends for different areas, they can see where they need to prioritise infrastructure upgrades to reduce the chance of failures due of overloads.

Retail – better inventory management

Retailers' challenges include gathering insights from customer data, complying with regulations – especially privacy standards – and ensuring accuracy in manual inventory management.

With synthetic data, retail organisations can simulate scenarios involving sales trends, supply chain disruptions and consumer demand shifts to provide a more enriched dataset for developing sophisticated inventory management models. They can also mimic real customer data that doesn't contain any sensitive information, enabling secure data analysis and compliance with privacy regulations.

This helps online and brick-and-mortar retailers maintain ideal stock levels to reduce holding costs and ensure product availability. They can also optimise store layouts, product assortments and customer loyalty while reducing returns and improving marketing.

US healthcare insurance – more accurate diagnosis and better patient monitoring

Physicians rely on medical images to diagnose patients and continually monitor their condition. Health insurance companies also need this information to process claims. But to do this effectively, both need diverse, high-quality datasets of every illness in existence – and every variation. This simply isn't possible due to the rarity of certain diseases and data privacy issues preventing the use of visual data.

Synthetics can generate diverse examples of medical images, from X-rays to MRIs. These can represent the entire range of conditions to augment existing datasets, improving diagnostic models' training. Synthetic videos or images can simulate various patient conditions and behaviours, enhancing models designed for patient monitoring and engagement.

This can create new or improved advanced diagnostic suggestion capabilities directly in the medical imagery device or platform. This helps physicians identify and monitor conditions, improving the efficiency of treatment – and, ultimately, patient care – while speeding insurance company claims processes.

Hospitals – improved operational efficiency and security

For hospitals, the variability in real-world scenarios makes it hard to optimise operations and patient flow using visual data, like CCTV feeds.

Synthetic data enables hospitals to simulate different patient flow scenarios and facility layouts visually. Enabling monitoring of offline monitoring equipment, via CCTV, improves operational efficiency and security through a better understanding of patient and visitor behaviour.

Technology companies – extended reach into new markets

Developing robust and efficient algorithms requires high-quality, diverse datasets. These are often scarce or expensive to procure, and they're bound by strict data privacy regulations. Ensuring the quality and reliability of software, hardware and systems also requires extensive testing under diverse conditions.

Synthetics can simulate varied scenarios, user inputs and environmental conditions to create high-quality datasets. This allows for comprehensive, efficient algorithm training, testing,validation and quality assurance processes, reducing dependency on real-world data acquisition.

And because synthetic data retains the statistical properties of real data but doesn't contain sensitive information, it ensures privacy compliance and secure data analysis.

All this means technology organisations can develop services and products with high-quality capabilities which are applicable across more industries – extending their reach into new markets. They can also reduce operational costs by eliminating or significantly lowering the amount of human effort needed to source labelled training data.

Robotics – increased safety, performance and cost-effectiveness

Deploying robots in real-world scenarios for training can be prohibitively expensive, especially when considering wear and tear, energy costs and potential damages. Robots in learning phases could also pose risks to humans, themselves, or the environment, especially in unpredictable scenarios. There is also limited diversity in training scenarios as real-world training can be constrained by location, availability of objects, environmental conditions and other variables.

Synthetics offers cost-effective training as robots can be taught in virtual environments, reducing wear and tear and eliminating the risk of damage in real-world scenarios. Virtual training also eliminates risks to humans and the environment, while allowing robots to learn without real-world consequences. It can ensure comprehensive training by simulating a wide variety of controlled situations, environments and interactions.

By reducing the need for physical trials and data labelling, organisations can make significant savings in training and deployment costs. Plus, training in risk-free environments ensures robots are safer when deployed, decreasing potential liabilities associated with mishaps. And training on diverse synthetic datasets ensures robots are prepared for a wide array of scenarios, leading to enhanced performance, reliability and functionality.

Science – a smarter way to reduce pollution

Synthetics has many applications for scientific research organisations. One example is tracking rubbish in the world's waterways, which is currently labour-intensive and costly using human effort alone. Even automated systems rely on training, testing and validating computer vision models using vast amounts of labelled images that cover an extensive array of objects.

By generating synthetic images which can be added to training datasets – and using AI-based image recognition to track litter – scientific research organisations can build effective and scalable monitoring systems.

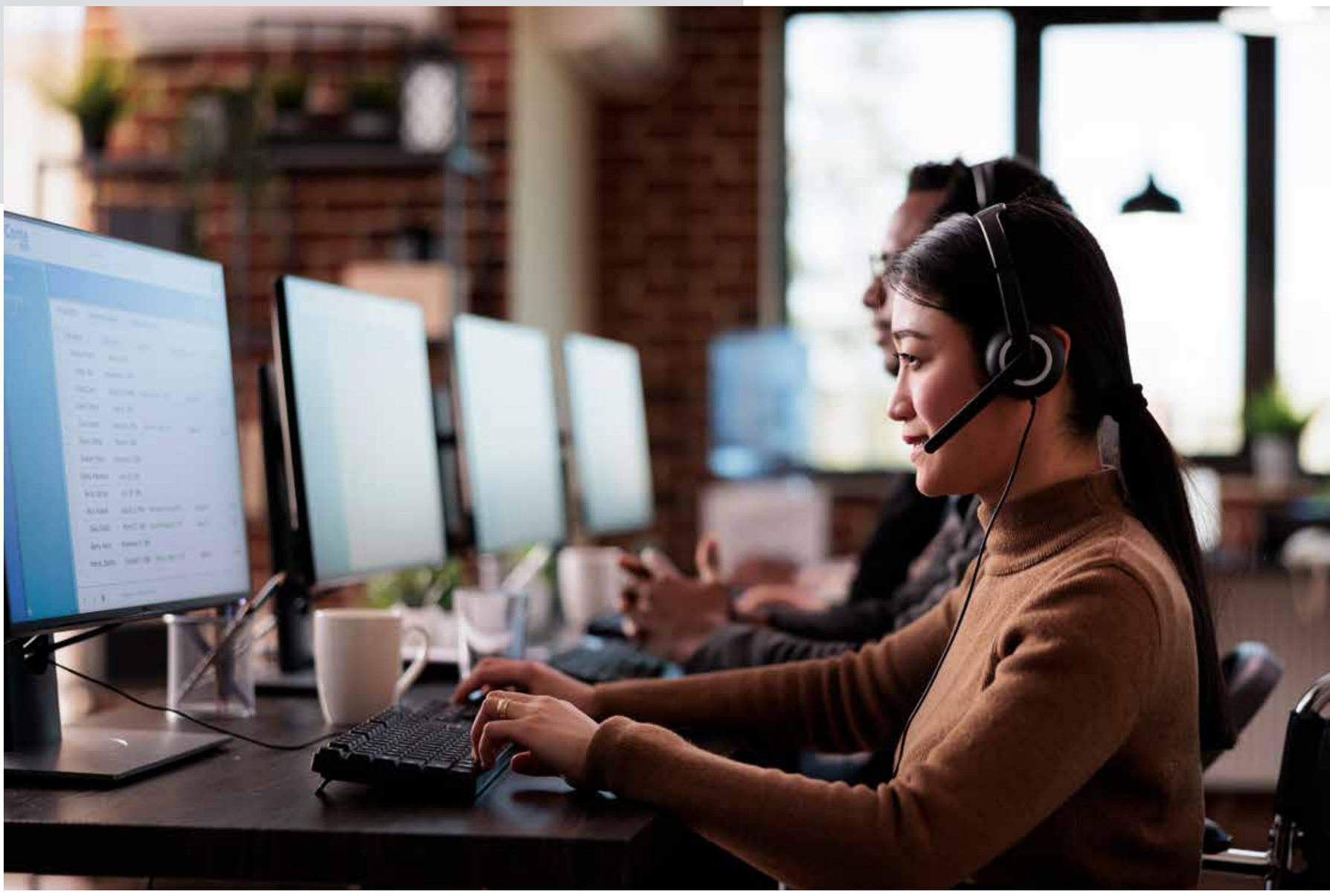
This use of synthetic data in this way would help to predict trends and identify areas of high pollution density, accelerating the clean-up response far beyond the capabilities of manual processes. Ultimately it would support a more intelligent, lower-cost process for reducing damage to the environment.

How synthetics accelerates insurance claims

Synthetic data and ML improve insurance claims processing for automotive, property damage, medical and workplace safety.

ML, trained on synthetics can process claims faster and at a lower cost by using predictive models and automation to triage claims, assign adjusters and determine the best method of inspection. Data analytics can also help insurers optimise their workflows and reduce manual admin burden.

These technologies complement claims adjusters' intuition and experience by providing data-supported insights and recommendations that can enhance decision-making processes. Data analytics can also help adjusters identify potential fraud, subrogation, litigation or other complex issues that require special attention.



Not just for organisations with next-gen technology

The use of synthetic data to train ML models is not restricted to businesses with the latest technology systems in cloud-native industries. Companies that still run on homegrown or on-premises technology can also extract value from this new field to solve their industry-specific challenges.

Endava Synthetics – your strategic partnership for success

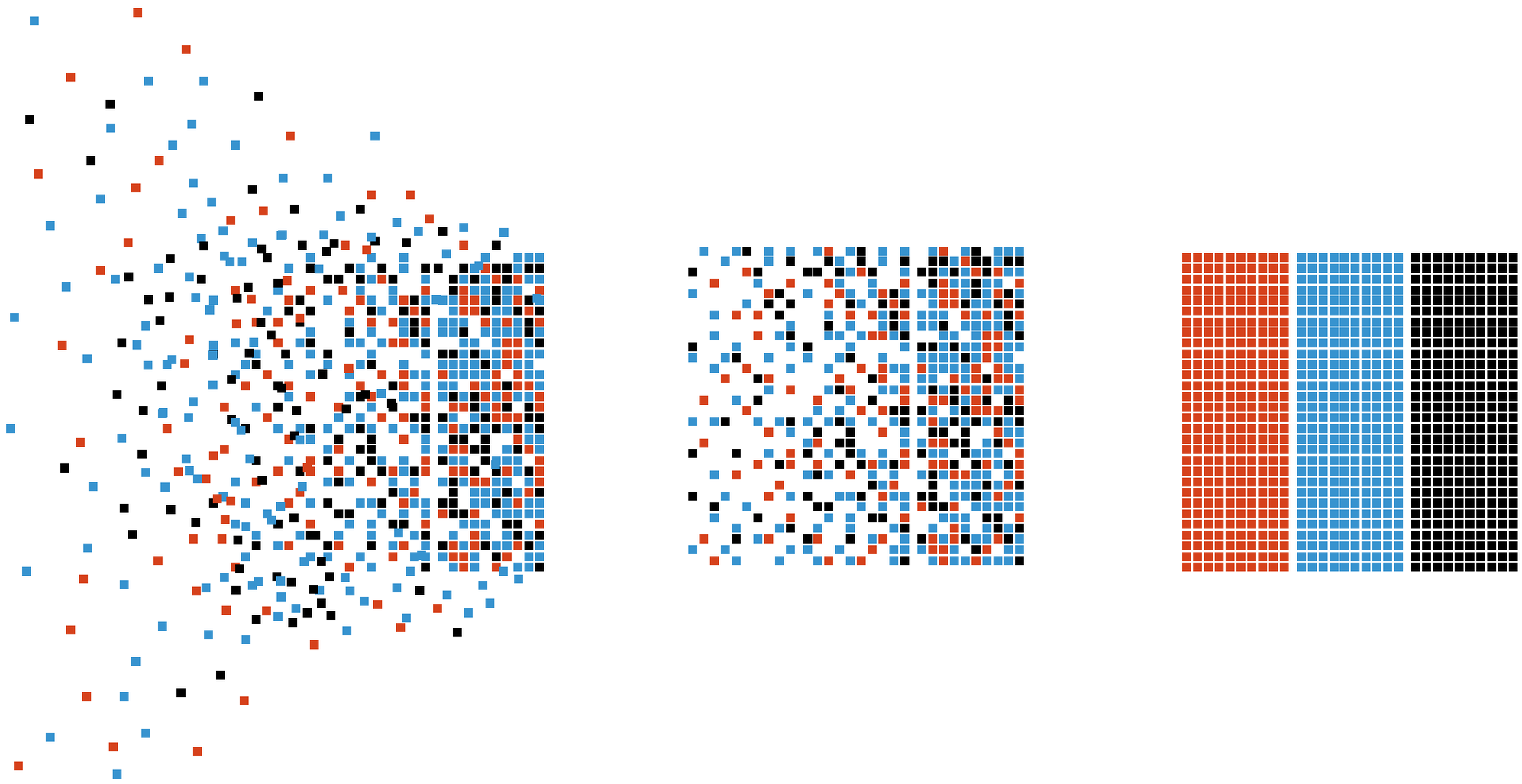
Numerous companies already use synthetic data, but implementation is being hindered for many others. Endava Synthetics can help you accelerate adoption.

Large tech companies have been successful using synthetic data for many years. However, to access the business benefits, you need teams with a combination of technical artistry skills and machine learning expertise.

This is because achieving the reliability, cost-effectiveness and value of synthetics requires experts in the fields of 3D pipeline, graphics, data science and machine learning.

Few companies have the budget, resources or expertise to create an internal team and build the solutions – nor scale them up. Furthermore, due to the abundant use cases of synthetic data, many of the companies that do have internal teams often lack the bandwidth to meet all their project needs.

These factors have prevented organisations in many industries from exploring the use of synthetic data. This is where Endava comes in.



Introducing Endava Synthetics

Endava Synthetics specialises in generating artificial visual data – including images, videos, and 3D data – to enhance your AI and ML solutions.

We help you build robust ML solutions by providing tailored synthetic data to tackle complex challenges, optimise performance and unlock new opportunities. We also offer consultancy, working collaboratively with you to help make sense of the data and apply it.

We have specialists in technical artistry and ML with over 15 years of combined experience in the field of synthetic data. They manage the complete lifecycle of ML operations, from data generation and model training to performance evaluation and deployment.

Our team delivers data capabilities in a broad range of domains, including manufacturing, automotive, energy, retail, robotics, defence, insurance, healthcare and science.

We offer a range of services to help you create an effective approach to developing ML solutions. And you can customise these offerings to your unique requirements.

Regardless of your technological maturity, we can help empower your teams to take advantage of synthetics.

Endava Synthetics offerings

Dataset generation

We design a strategy tailored to your unique scenario. We then deliver datasets that you own; your internal ML teams use these for algorithmic development. This improves your model performance through an iterative process of pipeline refinement, dataset generation, model training and evaluation to craft the ideal dataset for your use case..

Pipeline development

Expanding on our dataset generation service, we develop the end-to-end synthetic data pipeline for your teams to use in-house. This is accompanied by stafftraining and continued support that ensures you can efficiently self-serve your ongoing synthetic data needs.

Component development

Our specialists design artist-centric tools, procedural 3D assets and ML-ready pipeline components that you can integrate into your existing synthetic data pipeline.

ML operations

With this offering, we deliver a range of services designed to maximise the effectiveness of synthetic data solutions. These include dataset management, pipeline development, model training and evaluation, as well as deployment into a production environment.

Staff augmentation

When you need further support, you can bring our experts in-house to work on your own synthetic data projects and pipelines.

Endava Synthetics capabilities

Our synthetic data is engineered to mimic the essential qualities of a given scenario while maximising generalisation by varying dynamic properties. This improves upon strict realism by allowing precise control over diversifying factors. Here are the capabilities of our offerings:

- **Advanced ground truth:** Labelling, bounding boxes, semantic segmentation, key points, arbitrary metadata and output variables allow for deep labelling of all frame elements.
- **Targeted distribution:** Precise control over data characteristics enables the fine-tuning of specific parameters to help generalise and reduce bias
- **Hybrid data:** Augment real imagery with synthetic elements reduces the domain gap – discrepancies between synthetic and real data – and content production timeline.
- **Sensor modelling:** Model sensor noise, distortion, grading, vignetting and blur to characterise real-world hardware more accurately.
- **Multi-spectral data:** Produce data in non-visible light bands such as infrared and X-ray.
- **Procedural variability:** Introduce a range of realistic conditions and scenarios that enhance the robustness of machine learning models.
- **Deformations and defects:** Represent rare and difficult-to-collect imperfections or anomalies.
- **Crowd behaviours:** Analyse complex group dynamics, movement patterns and interactions.

The business value of Endava Synthetics

- **Remove human error:** Endava Synthetics data generation offers pixel-perfect annotations which ensure data precision and improve learning capabilities.

- **Protect privacy:** By generating synthetic data that mimics the properties of real data, we can help you preserve the privacy of individuals, reducing the risk of data breaches and complying with privacy regulations.

- **Fill in rare data gaps:** Our services can help you simulate rare events, future scenarios or conditions which are not present in the original data.

- **Add variety and volume:** With Endava Synthetics offerings, you can augment existing datasets to provide additional diversity and volume for ML model training.

- **Save product development time:** By precisely previsualising hardware performance and interaction scenarios, you can make iterative improvements and optimisations before physical prototyping or deployment, potentially avoiding the need to return to the start of the product development process.

- **Validate edge case scenarios:** Endava can help you adopt a more proactive approach to model iteration by validating edge case scenarios in existing models.

- **Create a reusable resource:** We can help you regenerate synthetic data as needed, providing a stable and reproducible resource for testing and development.

- **Avoid biased data:** By creating an unbiased collection of diverse and varied synthetic data, we can help you train ML models to generalise better to unseen data.

Unlock new opportunities

To discover how Endava Synthetics can help you get the synthetic data you need to tackle complex ML challenges, optimise performance and unlock new opportunities, **get in touch.**