Survey Report

Synthetic Data

Key to Production-Ready AI in 2022

datagen

Computer Vision teams double down on synthetic data to unlock the potential to large-scale applications

Datagen, the world leader in simulated synthetic data recently commissioned Wakefield Research to conduct an online survey of 300 computer vision professionals to better understand how they obtain and use AI/ML training data for computer vision systems and applications, and how those choices impact their projects.

The findings portray computer vision as a field in a temporary state of growing pains as paradigms shift, industry standards struggle to take shape, and development lags behind major market potential. The scarcity of high-quality, domain-specific datasets for the testing and training of Computer Vision (CV) applications has left teams scrambling for alternatives. Most in-house approaches require teams to collect, compile, and annotate their own "DIY" data — further compounding the aforementioned problems with the potential for biases, inadequate edge-case performance (i.e. poor generalization), and privacy violations.

However, a saving grace appears to already be at hand for this transitioning industry — advances in synthetic data. This computer-generated, photo-realistic image data intrinsically offers solutions to practically every item on the list of mission-critical problems CV teams currently face.

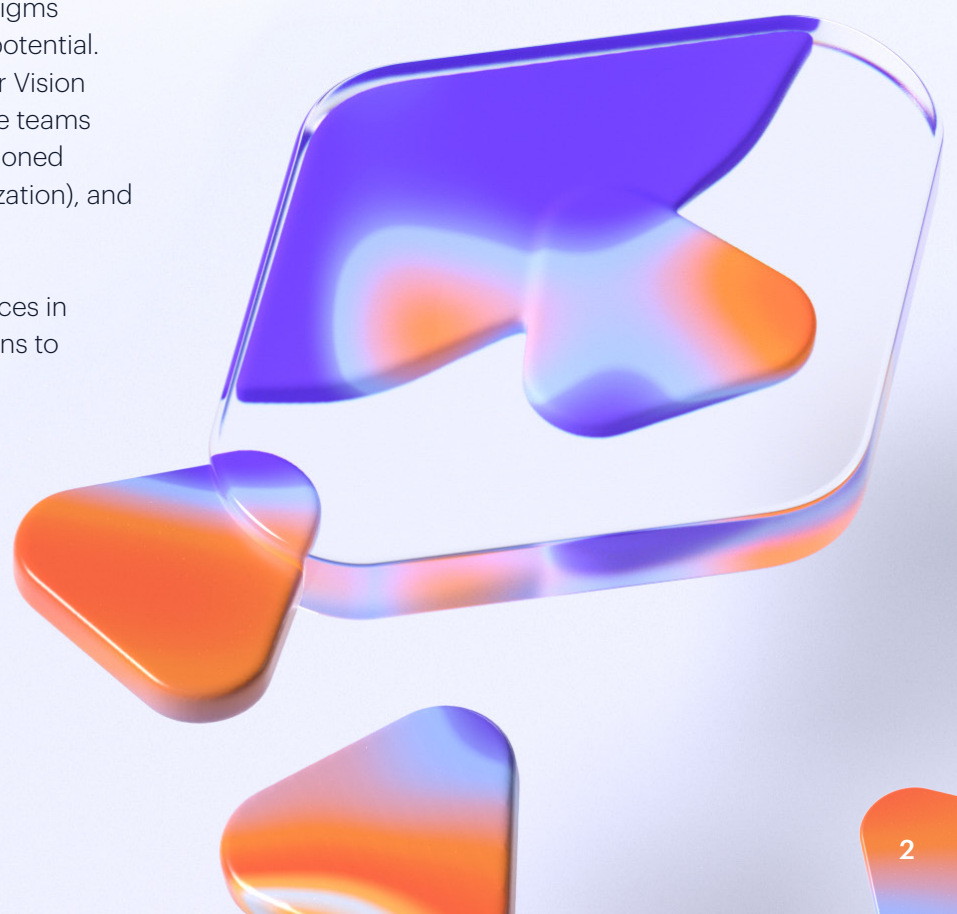


Table of Contents

Executive Summary

Organizations are investing heavily into a broad spectrum of markets that are dependent upon computer vision, including extended reality (XR), robotics, smart vehicles, and manufacturing. However, the commercialization of computer vision technologies is currently being hampered by a lack of standards, best practices, and resources for corporate teams.

The issue of training has become a clear choke point in the development of computer vision applications. As these applications grow in complexity and sophistication, so do their data requirements. Datasets must be larger, more variable, more consistently annotated, and more focused on the application's intended domain. Because of these increased demands, relying on manual data for training such applications becomes exceedingly impractical, if not downright impossible. With traditional approaches to training data now having reached the limits of their utility, computer vision teams are going in search of alternatives. And thankfully, there is good reason to be optimistic: synthetic data.

Synthetic data offers a broad host of benefits by allowing for quicker, less resource-intensive generation of high-quality, targeted (i.e. domain-/application-specific) datasets for machine learning (ML) model training. As a result, teams are able to adopt a data-centric approach to ML development, iterating quickly with increasingly robust and refined datasets, well-equipped for edge-case generalization and well-optimized for reliable performance. What's more, synthetic data generation eliminates the need for human annotation, mitigates bias, and removes privacy concerns.

This survey's findings reveal that synthetic data has already started to gain widespread acceptance among computer vision professionals. However, how it is sourced and to what extent it is being employed leaves a great deal of room for ML development to catch up with ML opportunities. Relying on publicly available datasets, be they real-world or synthetic, is no longer adequate for most computer vision applications, and teams are often left utilizing a grab-bag of poorly optimized data, blending manual collection, publicly-available datasets, and in-house synthetic data generation to compensate. Unsurprisingly, roadblocks persist.

Thanks to the advent of technologies like the metaverse, smart automotives, and the rise of new paradigms in the field, such as Andrew Ng's "Data-Centrism," there will be no scarcity of demand for high-quality, targeted synthetic data.

"One consistent theme in the history of deep learning is that larger and more diverse datasets lead to stronger models. Synthetic data is an incredibly promising way to increase dataset size and diversity and allow us to build stronger models across all computer vision use cases."

Anthony Goldbloom,
Founder and CEO, Kaggle

Key Findings

1

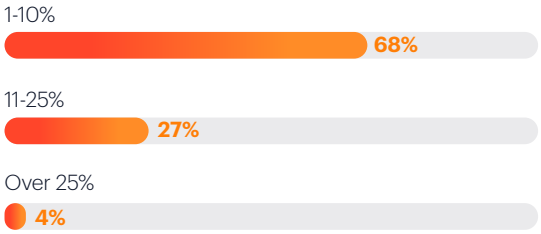
The Problem

DIY data delays & derails development

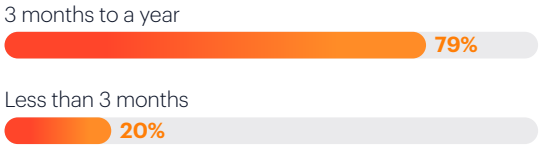
When it comes to the current state of the industry, the results of this survey leave little room for debate: computer vision projects are plagued by inefficiencies, delays, and cancellations -- and inadequate data is often to blame. A stunning 99% of respondents reported having had an ML project completely canceled due to insufficient training data. And nearly a third (31%) of those respondents reported that such cancellations affected more than 10% of their projects.

While 3 of the 300 respondents were lucky enough to avoid outright project cancellation due to insufficient data, no one was able to escape the plague of data-driven project delays. The entire 100% of respondents reported experiencing project delays as a result of insufficient training data. 80% experienced delays lasting at least three months, while nearly a third (33%) reported experiencing delays lasting seven months or more.

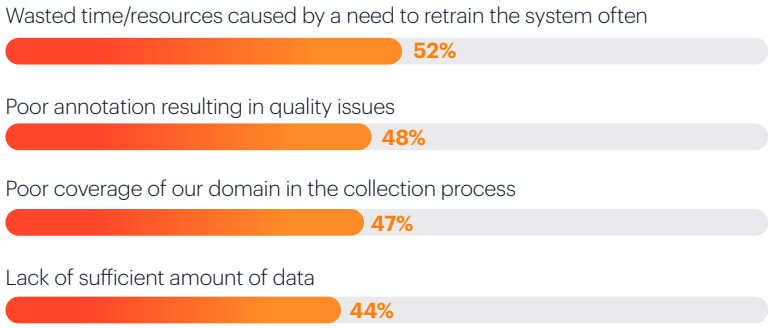
Approximately what percent of your active projects have ever been canceled due to a lack of sufficient data for training your networks?



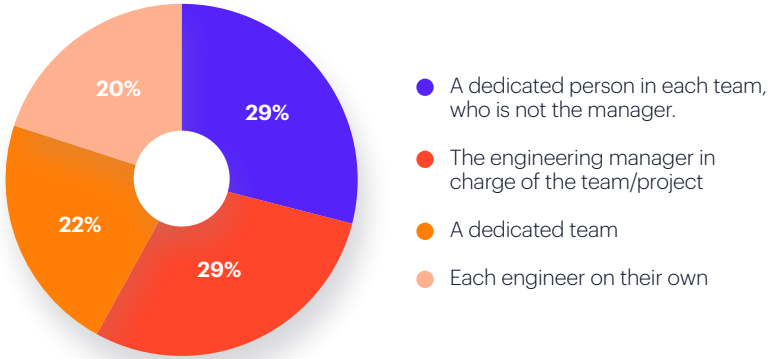
What was the longest period a project has been delayed because of issues in training data for your networks?



Has your team experienced issues with training models?



Who is primarily responsible for gathering synthetic data within your organization?



The primary culprit behind the ubiquitous delays and cancellations is inadequate training data. While insufficient amounts of training data will surely compromise the training process, datasets can be inadequate in more ways than one. The survey revealed that everything from poor or inconsistent annotation to inadequate domain coverage is compromising the training process for computer vision teams, and these problems are by no means mutually exclusive. In fact, **each issue affects nearly half of all computer vision teams** (44-52%).

In addition to the multivariate issue of inadequacy, computer vision teams seem to lack a consensus around procedures and best practices for collecting and curating training data. When asked who is primarily responsible for gathering synthetic data within their organization, **20% of respondents report each engineer was "on their own" in gathering training data within their organization. Meanwhile, less than a quarter (22%) reported having a dedicated team responsible for synthetic data collection.**

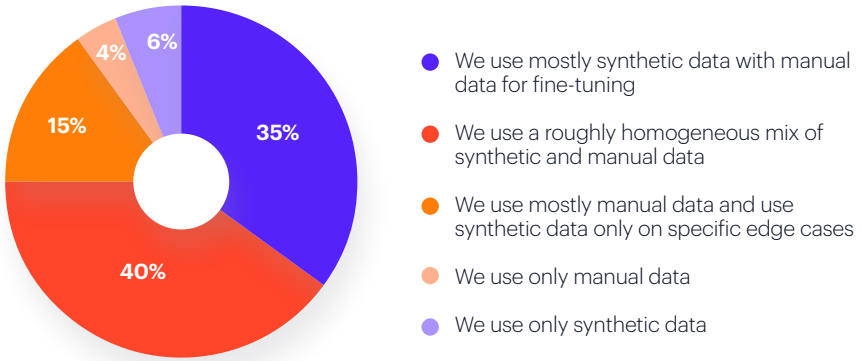
2

The Solution

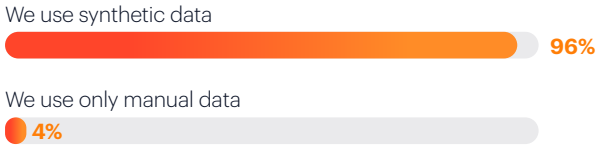
The synthetic data solution is already underway

Synthetic data is already gaining widespread adoption by computer vision teams. **96% of teams report using synthetic data** in some proportion for training computer vision models.

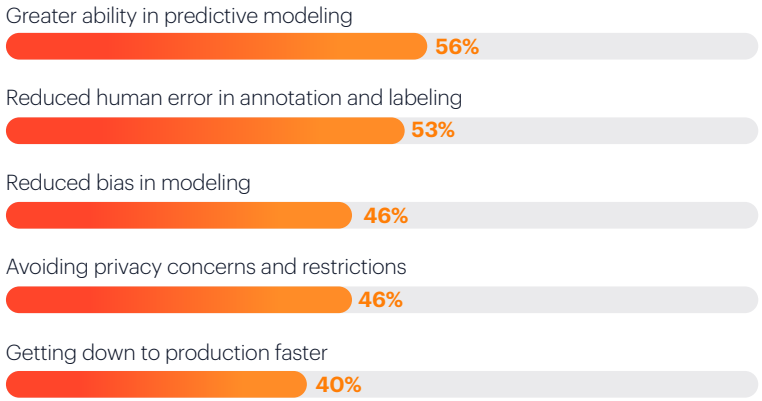
Which of the following best describes how your organization currently uses synthetic and manual data for modeling machine learning?



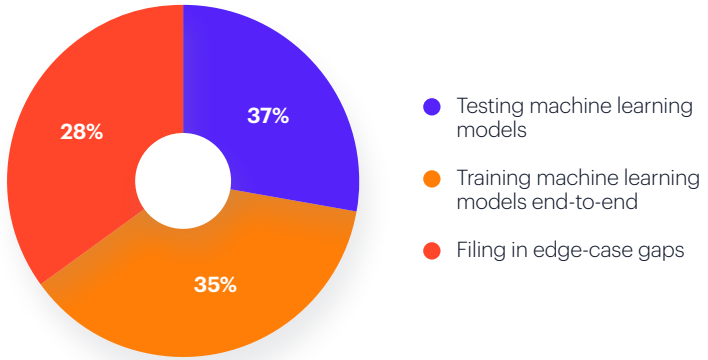
Synthetic Data Adoption



Which of the following benefits has your organization experienced in using synthetic data?



What is the primary reason your organization uses synthetic data?



The benefits of synthetic data are both broadly understood and broadly experienced. For example, when asked what the primary motivation was behind their organization's use of synthetic data, CV teams revealed a surprisingly even distribution between testing (38%), training (37%), and addressing edge-cases (28%) for machine learning models. Similarly, when asked which benefits have their organizations experienced from the use of synthetic data, respondents showed near-uniform support for reduced time-to-production (40%), elimination of privacy concerns (46%), reduced bias (46%), fewer annotation and labeling errors (53%), and improvements in predictive modeling (56%).

The consistent, across-the-board recognition of such a broad list of benefits and motivations for the adoption of synthetic data leaves little doubt that its use will increase in the coming years, as third-party synthetic data platforms and providers scale to meet this demand.

"Today, we have democratized sharing code, and with the generation of synthetic data, soon we will democratize training production models and accelerate the adoption of computer vision algorithms into our daily lives."

Fernando De La Torre, Research Associate Professor at CMU

Section 1

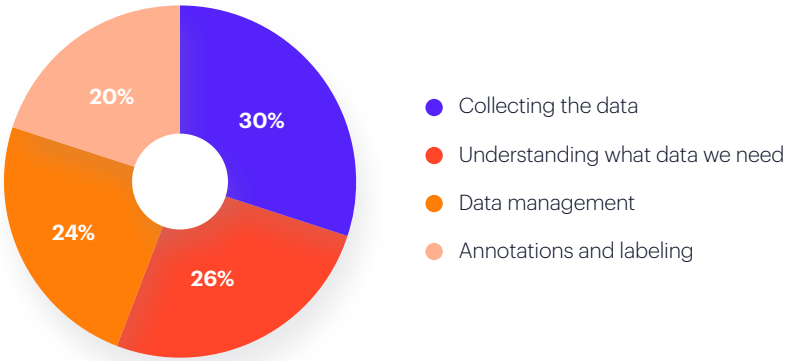
The Current State of Computer Vision

This report pulled the curtain back on computer vision teams' internal operations to reveal a field of missing standards and best practices. A large number of delayed or canceled projects caused by a confluence of factors is abundant, beginning with a dearth of readily-available, high-quality, domain-specific training data. They also indicate a lack of operational standards and practices, and a knowledge gap in the proper compilation and curation of training data for an intended application or use case, among other issues. **Siloed and squirreled away from the core, day-to-day business operations and decision-making, CV teams are being marooned on poorly-resourced islands, removed from the organizational mainland.**

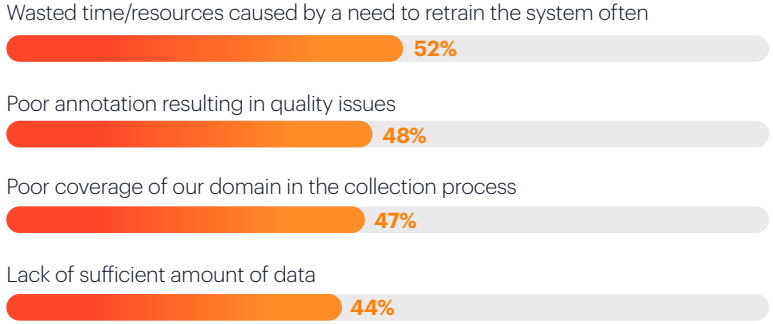
How Is Training Data Being Collected? In every which way, and no way in particular...

It is readily apparent the computer vision community has yet to establish clear standards or best practices around the matter of training data. When asked how training data is typically gathered at their organizations, CV teams report a patchwork of sources and methodologies employed both across the field and within individual organizations. A narrow majority (53%) reported still manually collecting and annotating some of their own training data, and nearly the same share (50%) reported working with synthetic data from publicly available datasets. Whether synthetic or real, collected in-house or sourced from public datasets, it seems organizations are utilizing any and all data they can in order to train their computer vision models.

What is the primary challenge when gathering data to train your network?



Has your team experienced issues with training models?



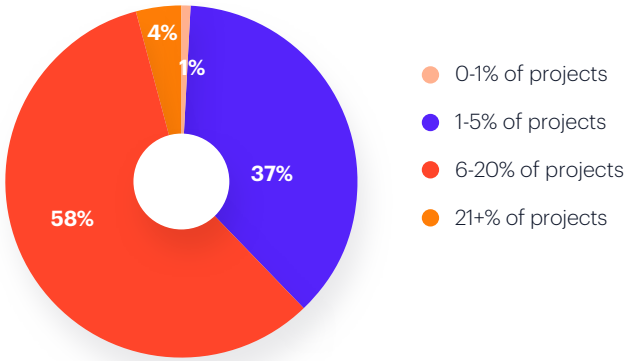
Training Troubles Are Plentiful in Both Frequency and Kind

The training phase of computer vision development appears to be a minefield of complications in today's enterprise environment. **97% of respondents share having one or more training-related complications** affecting their team. For each of the four most recognizable roadblocks jeopardizing computer vision training, nearly half of the organizations had experienced them at least once. From insufficient amounts of training data, to poor representation of the intended application's domain, and all the way to simple annotation errors/inconsistencies, CV teams experience a plethora of challenges while trying to train computer vision models in an enterprise setting. The burden of limited and poorly-targeted training data is placing computer vision teams at a severe disadvantage as they work to bring the next generation of artificial intelligence to life.

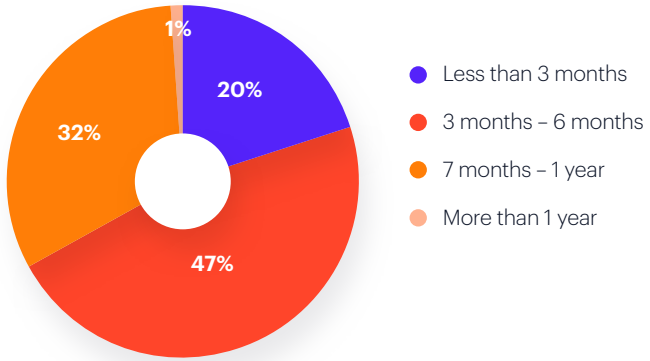
Section 2

The Costs of Confusion

Approximately what percent of your active projects have ever been canceled due to a lack of sufficient data for training your networks?



What was the longest period a project has been delayed because of issues in training data for your networks?



What's at stake when it comes to the current complications faced by the computer vision community is the field's scientific and commercial progress and innovation acceleration. Projects are being delayed or canceled altogether, and untold numbers of applications have yet to reach the market as a result. Whether it's delays or outright cancellations, computer vision teams are being held back by data constraints in spectacular fashion.

99% of respondents reported having had a computer vision project canceled due to insufficient or inadequate training data. Worse yet, 31% of those respondents reported that over 10% of the computer vision projects at their organization ended in cancellation.

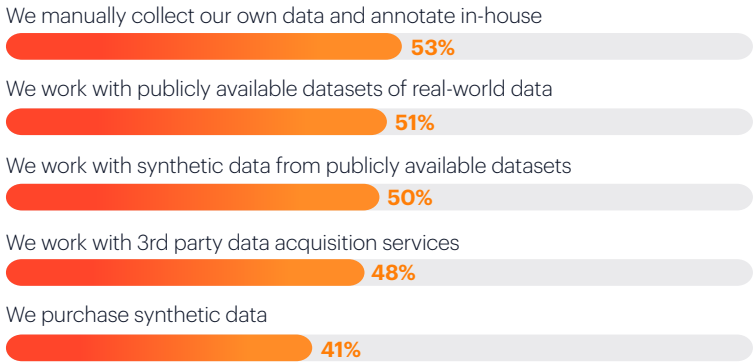
Meanwhile, all 100% of respondents reported experiencing project delays as a result of training data limitations. 80% reported having experienced delays lasting three months or more, while nearly a third (33%) reported experiencing delays stretching all the way to seven months and beyond.

Finally, to make these matters worse, the survey revealed just how time-consuming the process of annotating and otherwise preparing manual datasets can be. When asked, "On average, how long does it take you to start working with new data for your machine learning models?" 91% of respondents reported needing at least two weeks to prepare, and over 52% reported needing anywhere from one to three months.

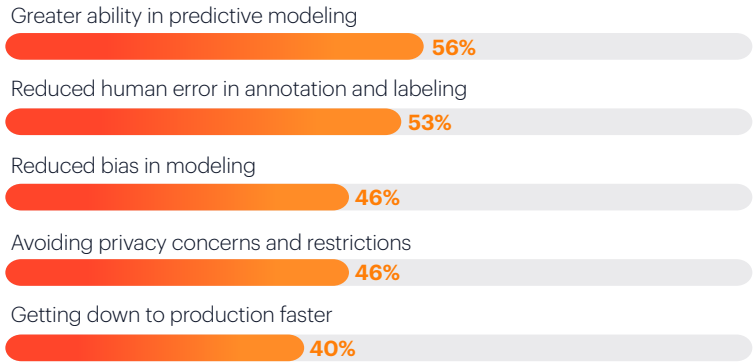
Section 3

Synthetic Data Earns Widespread Adoption

How does your organization typically gather data to train your networks?



Which of the following benefits has your organization experienced in using synthetic data?

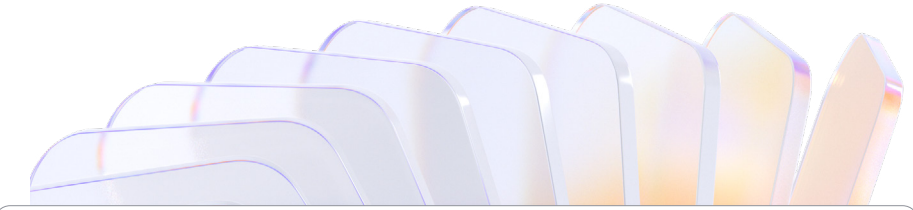


Although the methods of sourcing training data showed variation, the presence of synthetic data was clear — **96% of respondents reported using synthetic data** in some proportion for the training of their computer vision models. What's more, of the remaining 4% that only use manually gathered data, **over 38% plan on using synthetic data in 2022.**

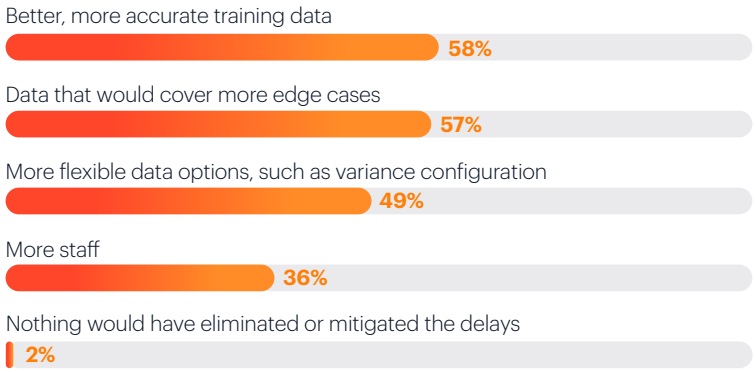
That means 293 out of 300 respondents are either currently using synthetic data or plan to in the next year.

While the vast majority of respondents reported currently using some combination of both types of data, a clear trend has emerged in which synthetic data has begun to overtake manual data as the predominant type of data being used for training and testing purposes by computer vision teams. **41% of CV teams reported most, if not all, of their training data was synthetic, whereas only 19% reported using manual data in greater proportion than synthetic.** Until recently, synthetic data was used exclusively to supplement real-world datasets. Now, it appears to be the rule rather than the exception. For a relatively new technology in a young field, the level of adoption of synthetic data among computer vision teams is remarkably high. This becomes less surprising, however, when taking into consideration the number and variety of synthetic data benefits the respondents cited.

Of the 96% of respondents that already use synthetic data in their development of computer vision applications, over 45% reported improvements in predictive modeling, reduced human error in annotation and labeling, reduced bias in modeling, and the mitigation of privacy concerns. Not only are CV teams benefiting from more accurate data for their models - they are able to get to production faster - a crucial business benefit cited by 40% of CV teams.



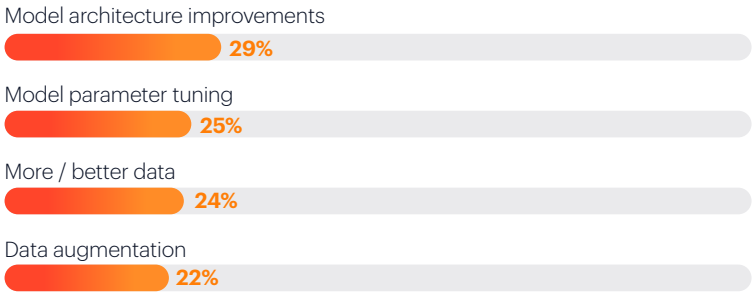
Thinking about the delays from issues in training data for your network, which of the following would have eliminated or mitigated those delays?



Finally, when asked what had the most significant impact on the performance of their machine learning models over time, the leading response was better data at 46% (inclusive of better, more, and augmented data). Better data was followed by model architecture improvements at 29% and the ability to tune the model parameters at 25%, in which better data is also key to the model tuning.

These findings support, and even go beyond, recent expectations put forth by Gartner, which forecast that 60% of the data used for AI and data analytics projects will be synthetic by 2024, and by 2030, synthetic data will eclipse real-world data entirely.

Which of the following has the most significant impact on your machine learning models over time?



Section 4

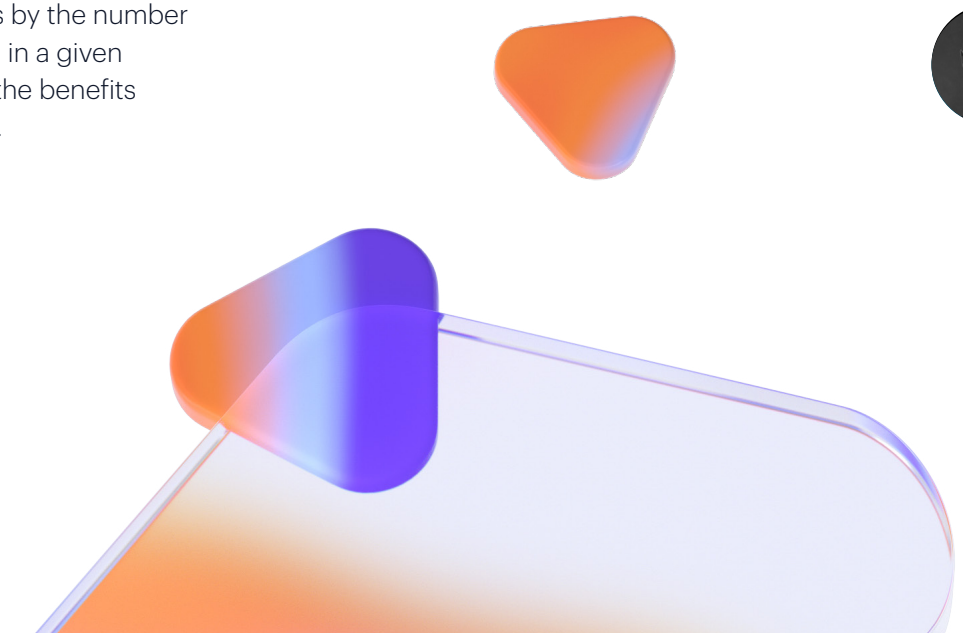
Data Centrism and Other Drivers

Beyond the benefits, synthetic data is also uniquely effective in mitigating edge-case failures, as one is able to make fast, inexpensive, and targeted additions to one's dataset with each iteration. For example, if a CV team discovers that their autonomous vehicle is frequently mistaking sandstorms for rain (and needlessly reducing speed and engaging traction control in response), rather than flying to Dubai with a team of professional photographers and waiting for the right wind conditions, they can simply generate synthetic, computer-generated images with sandstorm conditions in place. Plus, using synthetic data would spare the team the additional weeks or months of preparation and annotation that manually collected data require. Now amplify those discrepancies by the number of edge cases likely encountered in a given project development cycle, and the benefits become all the more compelling.

These sentiments were reflected in the study, with **57% of respondents sharing that their training delays could have been mitigated by "data that covered more edge cases."** What's more, over a quarter of respondents **(28%) reported their organization's primary reason for using synthetic data was to "fill in edge case gaps"** (Note, this is not the only reason, just the primary motivator).

"Synthetic data is fast, safe, sustainable, and can be engineered to be free of bias. This is a new era of data-centric AI development where you can have full control of your data to ensure successful AI applications."

Danny Lange, SVP,
Artificial Intelligence, Unity



Recent market trends, along with recent advancements in the field of artificial intelligence, are also fueling the continued adoption of synthetic data. Obviously, the significance of synthetic data's advantages scales in direct relation to the number of computer-vision-dependent applications in production. And perhaps no other phenomenon has had such a significant knock-on effect as this year's rise of the metaverse. After Facebook's initial July announcement of its massive metaverse initiative, interest in the mixed-reality medium has skyrocketed. In a November 2021 report, Grayscale Head of Research, David Grider, and research analyst, Matt Maximo, estimated that the metaverse would soon be a $1 trillion per year market. While such a valuation is clearly speculative, it is far from baseless. One figure the co-authors cited in support of the valuation was that from January 2020 to June 2021, the rate of active metaverse users increased by 1000%.

The metaverse's reliance on XR technologies makes it a wellspring of opportunity for computer vision developers. And as various organizations compete for first-to-market status with countless XR applications, the increased speed of iteration and superior model performance that synthetic data affords will undoubtedly cause a steep increase in its demand. Combine that with growing concerns surrounding privacy and bias manifesting itself in technology, and synthetic data becomes not only an attractive option for CV engineers but a necessary one.

At the same time, a recent paradigm shift in machine learning methodologies has led to an increased focus on and valuation of data. After Andrew Ng's March 2021 presentation on MLOps, "From Model-Centric to Data-Centric AI," the machine learning community has adopted a new outlook on training and developing AI applications that places the onus of iteration on improved datasets.

These trends are reflected in the current survey, as well. When asked what factors had the most significant impact on their machine learning models over time, 46% of CV teams cited either data quantity or data quantity or data quality as most important. Meanwhile, both annotation and data distribution were seen as equally significant factors in training efficacy.

46%

of CV teams cited either data quantity or data quality as being the most important

Annotation

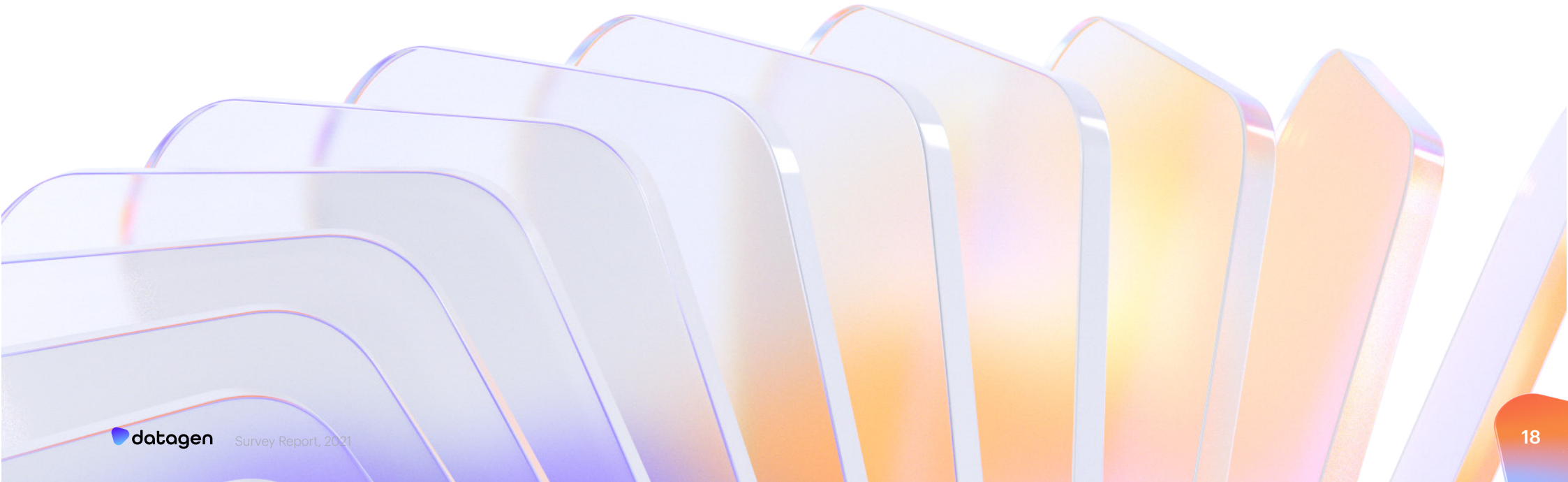
&

data distribution were seen as equally significant factors in training efficacy

Conclusion & Closing Thoughts

The current state of upheaval seen among computer vision teams may be alarming at first blush, but all signs point to this being a temporary byproduct of the changing technological landscape, rather than anything endemic to the field of computer vision itself. Even more encouraging is the range of benefits afforded by synthetic data itself. Not only does synthetic data increase teams' ability to iterate quickly and intelligently, it also drives down the influence of biases and eliminates any and all concerns surrounding privacy.

That it is a logical solution to many of the current challenges faced by CV and ML practitioners is self-evident. What this survey reveals is that the advantages are becoming evident to the teams and organizations as well. With more of the industry taking notice and more synthetic data options available, CV teams will, at last, have access to adequate quantities of high-quality, domain-specific training data. That it will be synthetic only matters in so much as it makes the latter possible.



We are Datagen

Datagen is powering the AI revolution by providing high-performance synthetic data, with a focus on data for human-centric computer vision applications.

We developed the first self-serve synthetic data platform that generates visual data, which is both photorealistic and high-variance. Our platform allows CV Engineers to create high-fidelity synthetic data with granular control and in a scalable manner. Fortune 500 companies rely on Datagen to develop their future products in the worlds of AR/ VR/ Metaverse, In-cabin Vehicle Safety, Robotics, IoT Security, and more. Founded in 2018, Datagen is led and backed by world-renowned AI experts.

Let's talk

Research Methodology

This Datagen Survey was conducted by Wakefield Research among 300 employees working in the roles of data managers, data acquisition managers, algorithms engineers, computer vision engineers, AI/ML engineers, data engineers, data scientists, and deep learning engineers at their companies, between November 9th and November 22rd, 2021, using an email invitation and an online survey.

Results of any sample are subject to sampling variation. The magnitude of the variation is measurable and is affected by the number of interviews and the level of the percentages expressing the results. For the interviews conducted in this particular study, the chances are 95 in 100 that a survey result does not vary, plus or minus, by more than 5.7 percentage points from the result that would be obtained if interviews had been conducted with all persons in the universe represented by the sample.