EBOOK

Training Data Ops Software: Build or Buy?

BROUGHT TO YOU BY

EXPLORE V7 →

Table of Contents

Reliable infrastructure for training data management is crucial for AI-first product success. With the emergence of dedicated software for training data operations, ML teams must decide whether to build their own solution or buy a platform. This ebook examines the advantages and disadvantages of each approach, offering insights to help you make the best choice.

Introduction | 03

Buy vs Build | 05

Data Security | 09

Hosting Options | 10

Effectiveness | 12

Latest Technologies | 13

Customization | 15

Integrations | 17

Cloud Provider Compatibility | 18

Expert Support | 19

Costs | 20

Value Assessment | 21

Training Data Ops Software: Build or Buy?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

Costs

Value

Behind every successful AI product, there are three key components: a talented team of ML engineers, the right model architecture for the task at hand, and reliable infrastructure for training data management.

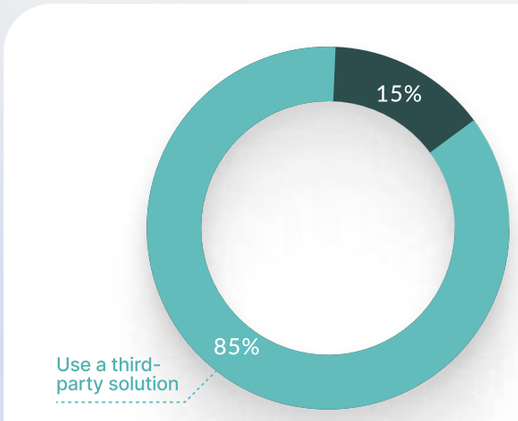
However, the scalability of the last component is often overlooked by many AI developers. And as projects progress, it becomes increasingly difficult to solve.

Setting up training data processes and making them work is the most challenging aspect of succeeding in a world of AI-first products. Training data is your edge, and ensuring its proper development is essential. From data storage to annotation quality assurance, managing and labeling your training data requires dedicated tools.

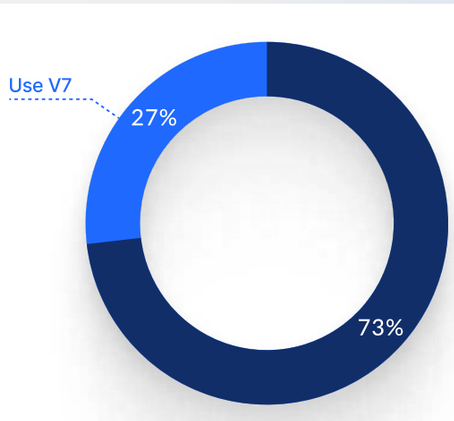
Now—

Most ML teams started building their own training data software some time from 2018-2021 because there were no procurable options. But, since then, dedicated software for managing training data operations as a mature piece of the modern ML Ops stack has emerged.

Today, the majority of companies use combinations of third-party solutions. It is easier than ever to access high-quality tools without having to build them from scratch.



The vast majority of Fortune 100 companies use third-party commercial training data platforms



One in four Fortune 100 companies uses V7 as their external training data management platform

EXPLORE V7 USE CASES →

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

Many ML teams still face the dilemma of whether to build a solution that meets their specific needs or just buy a training data platform.

The main considerations are:

Time to market

Developing in-house solutions can take significant time and resources, often resulting in delays that can be detrimental to a company's bottom line.

Ability to attract and retain top talent

Skilled professionals prefer to work on cutting-edge technologies and tools, rather than on internal systems that are not central to the company's core competencies.

Opportunity cost

By outsourcing non-core activities, companies can minimize their opportunity costs and maximize their returns on investment.

In this ebook, we explore the pros and cons of both approaches and provide insights into the key considerations for choosing the right solution.

Why companies decide to buy or build their ML Ops tools

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

Costs

Value

There are several reasons why companies choose to build their own tools. Historically, companies had to create their own frameworks and processes because there were no third-party options that met their requirements.

Why companies decide to build	Why companies decide to buy
Scarcity of external options	Engineering time can be better spent on their core product or other critical infrastructure tasks
Preference for keeping critical infrastructure as part of their own IP (or necessity to meet legal requirements)	Pace of innovation in AI is higher than their team can keep up with
Unique use case and more control	ML platforms have become more flexible

Some ML teams operating in industries such as healthcare had to develop their own solutions out of necessity to meet legal regulations.

For many businesses it wasn't a matter of building vs buying—they had to develop their own solution and now they weigh their options (and the sunk costs) between maintaining it or switching.

The pros and cons of both approaches boil down to whether you need a solution tailored to your specific use case or prefer instant access to state-of-the-art solutions. A platform will give you the ability to focus engineering resources on other tasks.

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

Build	Buy
PROS Tailored to specific use case Sense of ownership and control over the software In some cases, working on an existing solution instead of making a painful switch	**PROS** Instant access to state-of-the-art solutions Ability to focus engineering resources on core differentiators Professional support and a larger engineering team solely dedicated to improving this form of tooling
CONS Potential scaling challenges as the team grows Requires dedicated, scarce and highly competitive talent Time-consuming and costly	**CONS** It sometimes can be difficult to integrate software into existing systems Potential need for training to effectively use the software Lower flexibility in use-case customization

While building your own solution gives you more control, it can be a massive project that will drain resources. And it is a leap of faith—no one will promise you that your custom solution built in-house won't become obsolete by the time it is ready.

Let's dive into a crucial point: cost.

This includes the monetary investment as well as the opportunity cost of potential errors, falling behind with inferior capability, and distractions from core competencies. So, what are the actual costs we need to consider?

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

$250,000 for a minimum viable product

$100k to $1 million per year in cloud storage, image processing, and redundancy

Front-end, Back-end, and Deep Learning engineers requiring specialized skills for large-scale label rendering, inference infrastructure, and data versioning

To illustrate the costs and effort required, let's consider our solution.

V7 is a training data operations platform that includes everything you need to create and manage training data at scale, from AI-powered annotation tools to workflow management, an API toolkit and Python SDK.

Here are some key facts:

- Developing a basic version of V7 took approximately $250,000 in engineering costs., and a stable production version would cost approximately $2 million.
- We employ engineers with specialized skills in deep learning for labeling automation, back-end for large scale dataset management, and front-end for label rendering and drawing, which are scarce and highly competed on among few machine learning companies.
- We develop through a dedicated tech stack from the back end (Elixir) to front end (Vue.js, HTML cavas, and several custom libraries) plus several deep learning implementations dedicated to automated labeling.
- We also spend millions of dollars per year on cloud infrastructure costs and can maintain a dedicated team for our 99.9% uptime, massive data transfers, and lightning fast loading speeds.

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

Now, imagine trying to create a robust tool with the same functionalities from scratch.

Still, before making a decision, it's essential to have a clear understanding of your specific needs and the available options.

So-

Here are some of the most important issues that arise when deciding whether to build or buy training data software.

- Data Security. Will your data be kept secure?
- Hosting. Should you choose cloud or on-prem implementation?
- Effectiveness. How does the software stack up against your current solution?
- State-of-the-Art. What are the latest technologies available in training data management?
- Customization. Will the software fit your unique use case?
- Integration. Will the software integrate with your existing technology stack?
- Compatibility. Should you consider using your cloud provider's default proprietary software?
- Cost. What is the total cost and how do you calculate it?
- Expertise. Will you get help and assistance from experts in the field?
- Value. Is it worth investing in a training data platform?

Let's explore each of these factors to ensure you make the right decision for your company and your team.

1. Will my data be secure?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

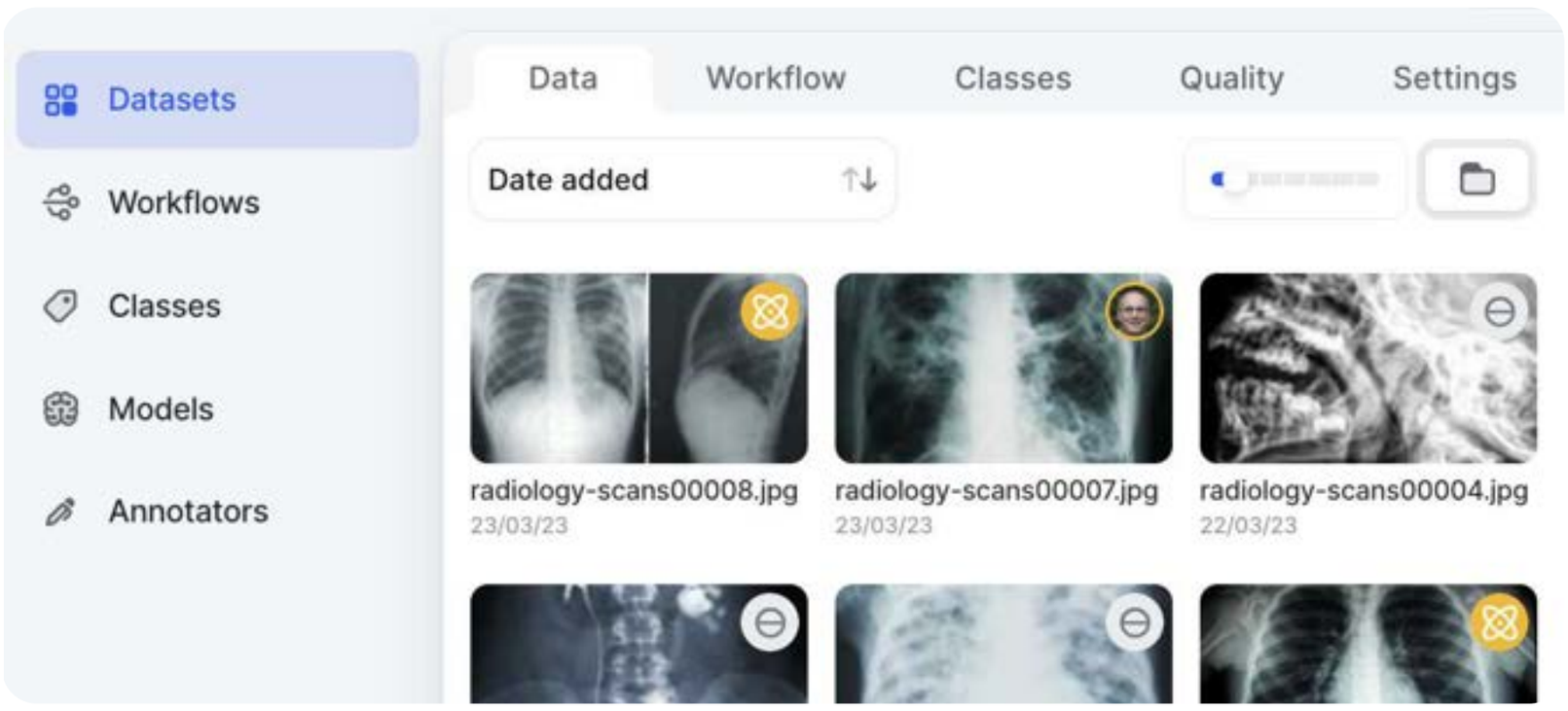
Expert Support

Costs

Value

Security is undoubtedly a top concern for companies when it comes to handling their data. Many businesses believe that internal software is the safest option because the data remains within their own architecture. However, modern advances in cloud computing have made it possible for data to remain within a company's own network or storage architecture while still integrating with external services.

For instance, the V7 training data platform offers both cloud and object storage integrations. This allows users to store data wherever they prefer while still visualizing it in V7's interface.



This approach is in line with HIPAA and FDA guidance for healthcare, and it is generally a requirement for all use cases. In situations where integrating with cloud storage platforms is not feasible, V7 can integrate with storage from behind a firewall or within a VPC, making it possible to use the V7 platform even in previously inaccessible cases.

Find out more: Accelerating AI Product Development in Digital Pathology with Advanced Workflows

2. Where should I host my training data?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

Costs

Value

Next up, you need to consider where to host your training data. In the past, companies preferred to keep their technology on-premises, but with the benefits of cloud connections becoming more apparent, a cloud-first approach is becoming increasingly popular. This is particularly true in the realm of ML-Ops where deployed models “close the loop” with their training data in the cloud for continual learning.

One of the primary advantages of cloud-based training data operations software is the ease with which data can be shared between different stakeholders and collected from various sources. With cloud implementation, data can be accessed and analyzed from anywhere, at any time, making collaboration among team members more efficient and effective.

Cloud-based training data management solutions also offer more flexibility in terms of scalability and cost-effectiveness. With cloud implementation, companies can easily scale their data management solutions up or down as their needs change, without having to make significant investments in additional hardware or infrastructure. This can be particularly advantageous for companies that experience rapid growth or fluctuations in demand.

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

By adopting a third-party platform, enterprises can easily scale their training data operations to ensure that all teams involved in developing visual training data follow a consistent and controlled process. In contrast, teams that develop their own tools for this purpose tend to create inefficiencies across the company by working in their own unique ways with disparate class schemas, labeling formats, and produce incompatible assets. With V7, companies can provide a horizontal solution to their entire enterprise, unifying all developers, models, and datasets under one roof and promoting a "center of excellence" for their training data operations.

Additionally, choosing an external solution can minimize downtime or the the risk of data loss. ML-Ops is a new and complex sector with novel technologies that often break. Experienced machine learning teams carry a greater ratio of engineering support than normal developer teams due to the fickle nature of ML tools. Working with an external vendor you can outsource this maintenance to a team familiar with the codebase and dedicated to building redundancy. Many ML teams experienced label loss or corruption and it can bring product development to a halt for weeks.

3. How to benchmark it against my current solution?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

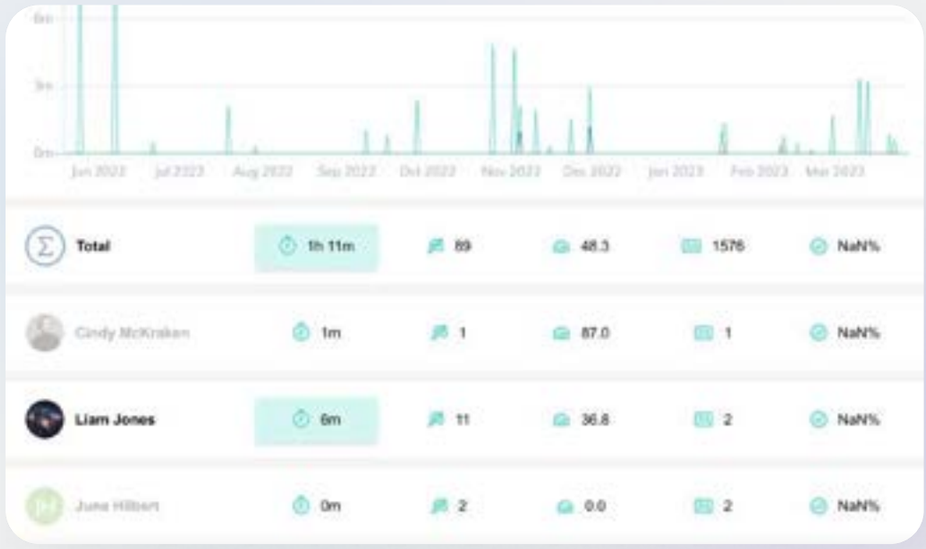
Expert Support

Costs

Value

There are two sets of criteria to consider when benchmarking: qualitative factors, such as support quality and user interface, and quantitative factors that can be measured.

A good approach is to select a limited number of projects that are varied across annotation, workflow, and data types, then evaluate them against key measurable criteria such as speed and accuracy of annotation, speed of administration, and end-to-end project time. Gather existing benchmarks across these areas, and use them to test both your existing in-house solution and a SaaS platform.



Most teams look at annotation accuracy versus a defined gold standard set, speed of annotation, time taken from requesting new data to receiving perfectly labeled data back, and model accuracy measured by MaP or similar metrics.

Designing a test

Select a set of images that represent the type of data you work with. Make sure that the images are diverse in terms of complexity, size, and content. The annotation tasks should involve the same annotation class types, for instance keypoint skeletons.

4. What is considered state-of-the-art in the world of ML data annotation?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

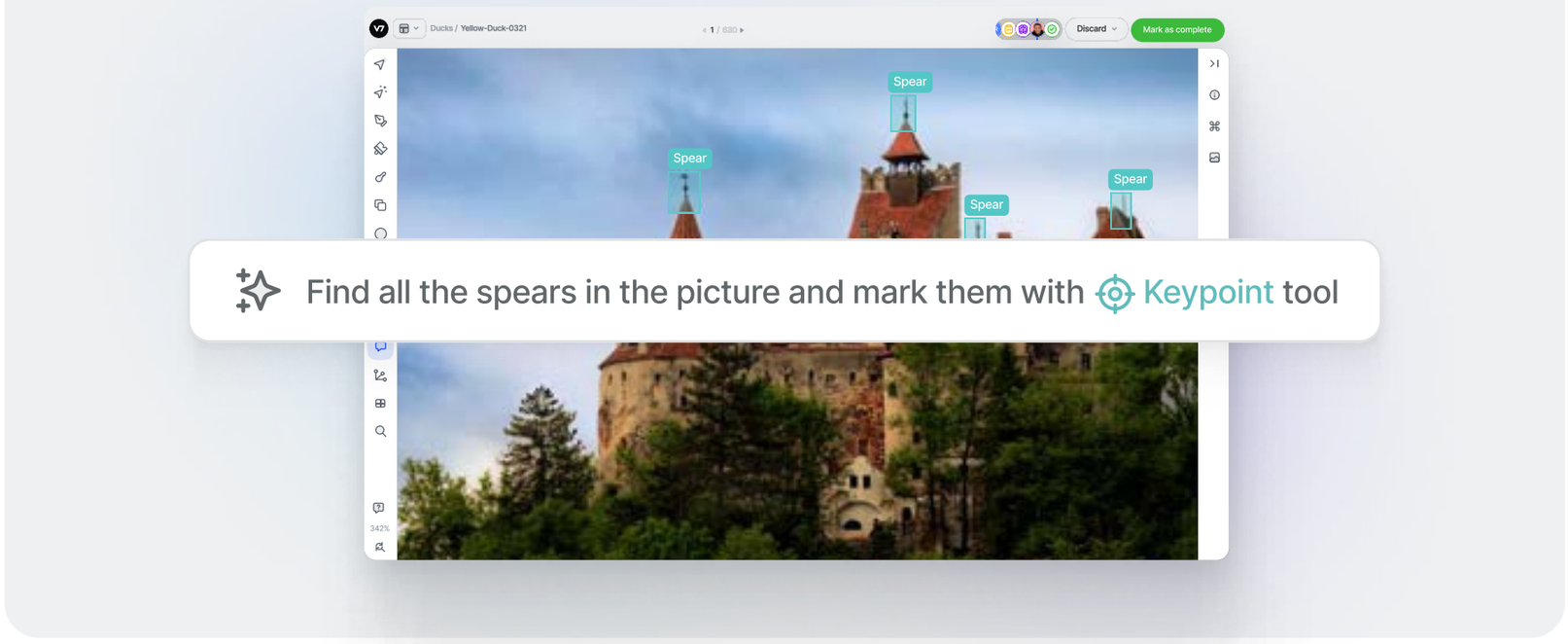
Costs

Value

Data annotation, especially areas such as image annotation, is evolving fast. Keeping up with the latest advancements is crucial to staying ahead of the competition.

One recent breakthrough in this field is the integration of models and humans using cloud-native workflow solutions. This allows for seamless collaboration between the two and results in faster and more accurate annotation.

Auto-annotation and using pre-trained public computer vision models can save time and resources. Additionally, features like label interpolation can help with annotating videos. To be truly "data-centric" and experiment with datasets, the software should allow for easy management of datasets and data annotations.

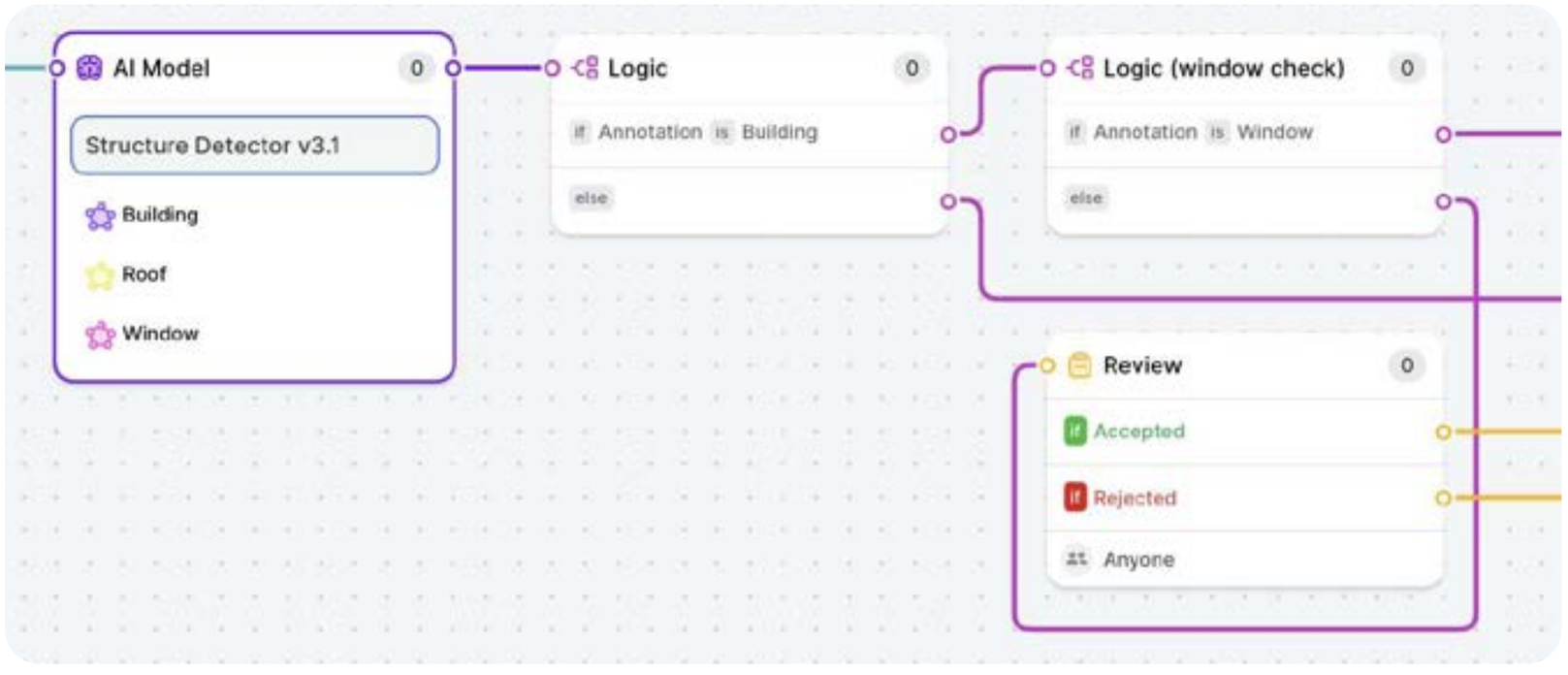


A recent trend that has gained traction in the field of data annotation is the integration of Large Language Models (LLMs) with computer vision models through conversational interfaces. By integrating your training data platform with an LLM model, you can provide instructions or prompts to annotate your data, such as "label all windows in the image." The LLM model then interprets these instructions and uses the appropriate computer vision model to solve the labeling task.

EXPLORE V7 USE CASES →

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

Designing custom workflows is another essential feature for modern ML teams. Workflows can be simplified or made more complex with advanced automations, multiple review stages, or nested conditions and logic rules. These workflows start as your labeling pipeline and evolve into the continual learning infrastructure for your product.



The last key component is analytics. It is important to have access to clear metrics and visualizations to monitor project progress, worker performance, and data quality. This can also include metrics such as acceptable disagreement thresholds for parallel blind tests where multiple annotators or models label the same images.

5. Will a training data ops platform work for my use case?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

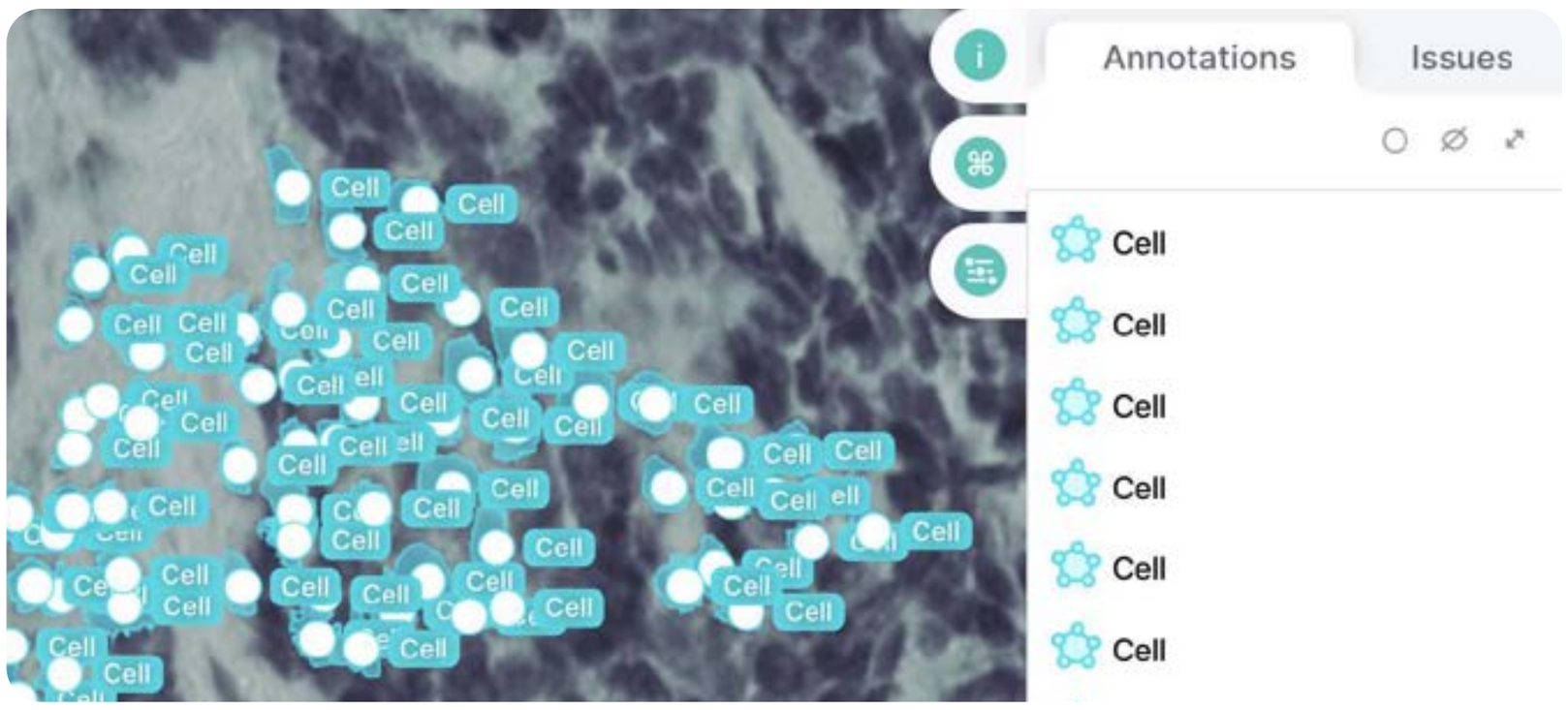
Costs

Value

In the overwhelming majority of cases the answer is a resounding "yes". However, as with most things, it ultimately depends on your specific needs and requirements.

Most third-party training data platforms cover 95% of use cases with drawbacks. But if you have unique features or customization options that your company requires, it's essential to ask potential vendors about their ability to accommodate those specific needs.

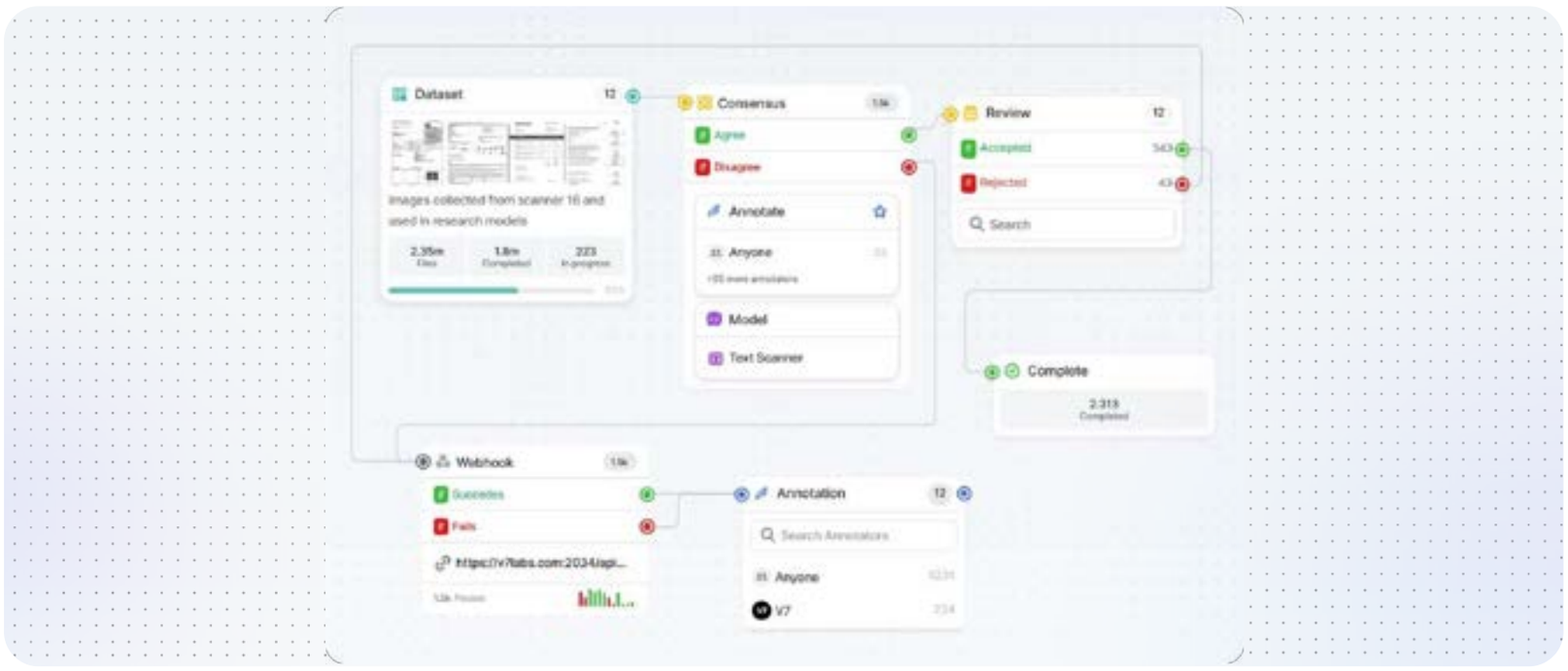
First of all, consider the data types you work with. Evaluate the capabilities of any tool you are considering to ensure that it can effectively handle your datasets. If you are looking for a solution for Natural Language Processing tasks or working with sound files, you may want to consider finding a vendor who specializes in those areas.



In order to effectively train machine learning models on WSI scans (Whole Slide Imaging), research teams need data annotation tools that support thousands of annotations on a single file.

Secondly, depending on the size of your team and the complexity of your data management needs, you may require different levels of support and user roles. Make sure to evaluate these features and capabilities to ensure that the tool can effectively support your team's workflow.

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value



It's important to do your due diligence and thoroughly evaluate any potential training data platform to ensure that it can effectively meet your needs. Don't hesitate to ask for a demo or trial period to test the tool with your own data and workflows.

6. Will it integrate with our current tech stack?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

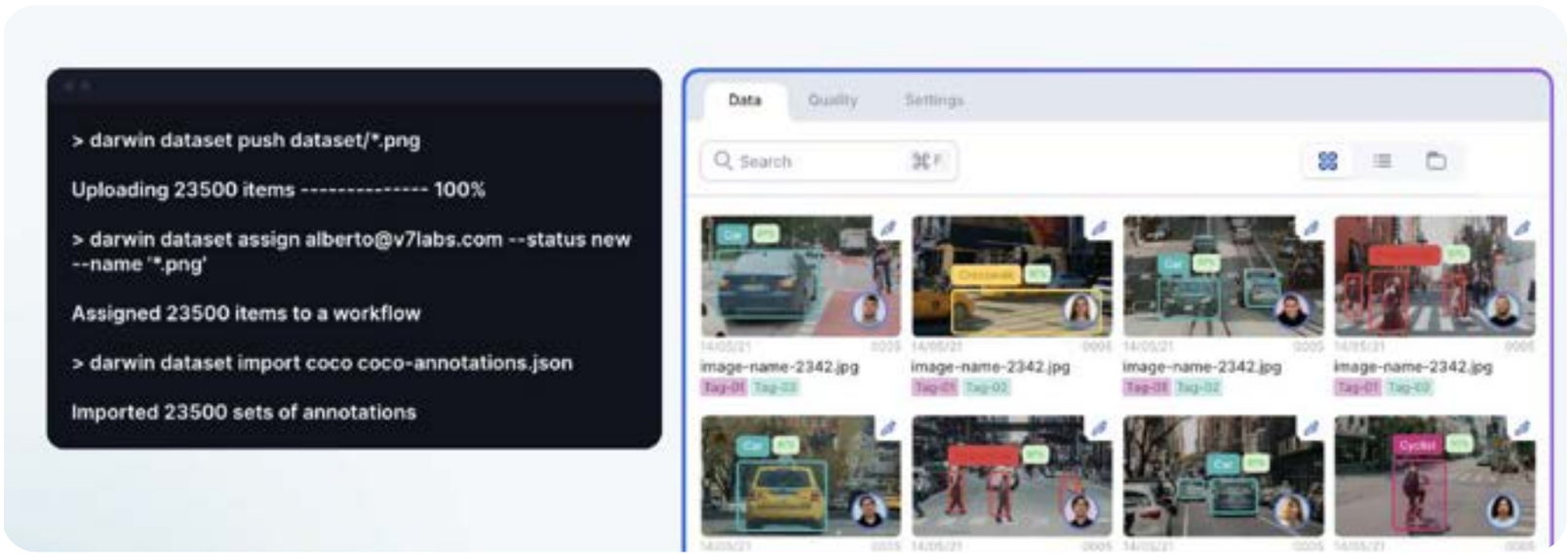
Expert Support

Costs

Value

The good news is that most training data platforms are designed with integration in mind. But it's also important to evaluate the specifics of how they will work with your existing systems.

For 50% of cases, the switch from in-house solution to a third-party tool will be almost instantaneous. In general, 90% of companies solve their integration problems within the first two weeks. For example, for new V7 users, we recommend having 1 engineer available for 2 hours per day for 2 weeks to help migrate onto the API. This will help to ensure a smooth transition and minimize any potential issues.



When evaluating different training data platforms, it's important to look for those that offer flexibility in their API options. Many platforms offer a REST API, which allows for easy integration with a wide variety of systems. Additionally, some platforms may offer SDKs (such as a dedicated Python SDK). For deciding if the pre-processed files with annotation will be suitable for your model training architecture, it is a good idea to check the JSON schema of exported labels too.

Most training data softwares integrate with all major cloud providers and various other parts of the cloud ecosystem. However, you'll want to make sure that the platform you choose is compatible with the infrastructure you use on a daily basis, which brings us to our next point —

7. Should I just go with my cloud provider's software?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

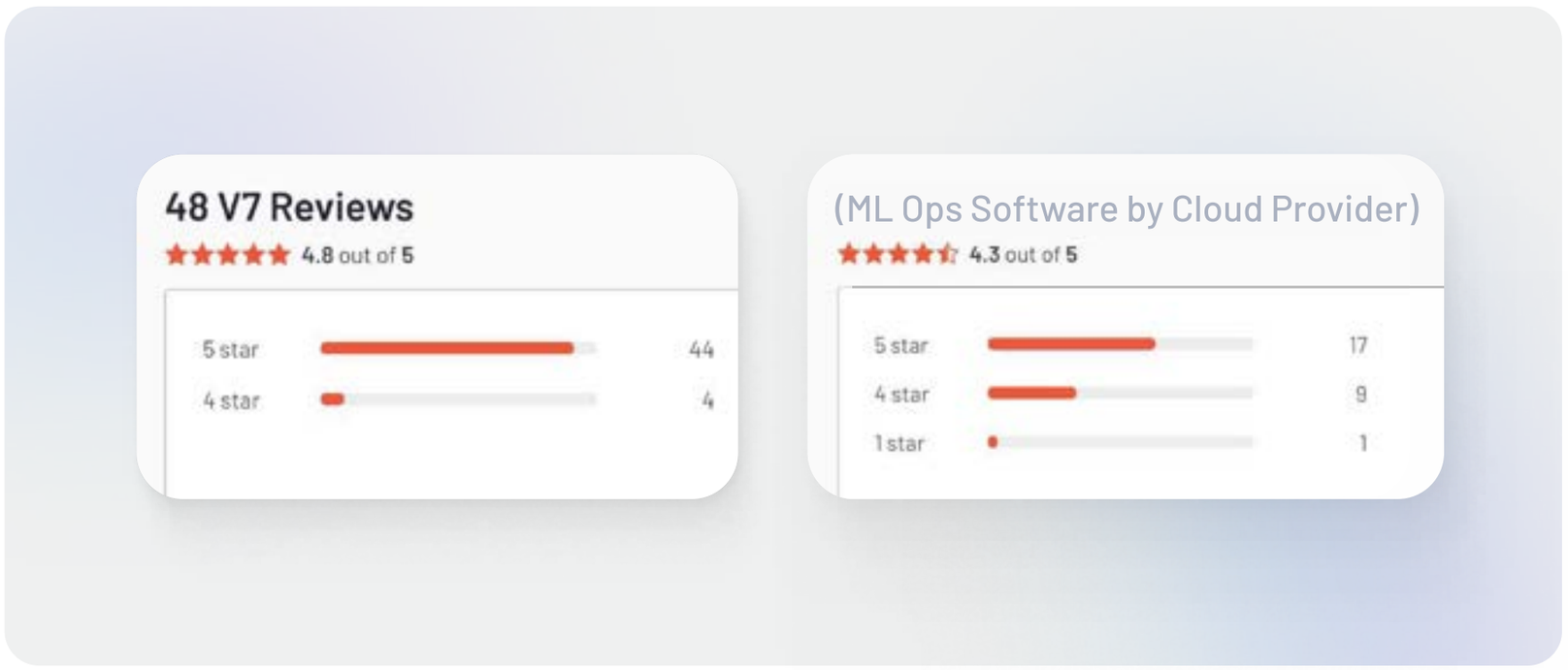
Compatibility

Expert Support

Costs

Value

When it comes to buying training data management software, one question that often arises is whether or not to use the tool offered by your cloud provider. For example, Amazon offers Amazon SageMaker and Ground Truth, its data labeling service.



While using a tool offered by your cloud provider may seem like a logical choice, it's important to evaluate the tool's capabilities and limitations before committing. Cloud providers strategically compete on compute, not on software platforms, often referring their clients to specialized third party vendors. They build tools that let customers get started in ML to secure them within their cloud offering, and expect them to outgrow them thereafter. The training data management space is highly specialized and requires a specific set of features and capabilities.

Using a tool from your cloud provider can be a good starting point for teams with basic requirements. It allows for a halfway house between a premium offering and an in-house offering, providing some functionality without large investment in building your own specialized platform.

Find out more: SageMaker AWS vs V7 Data Management Platform

8. What assistance, resources, and expertise will I get?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

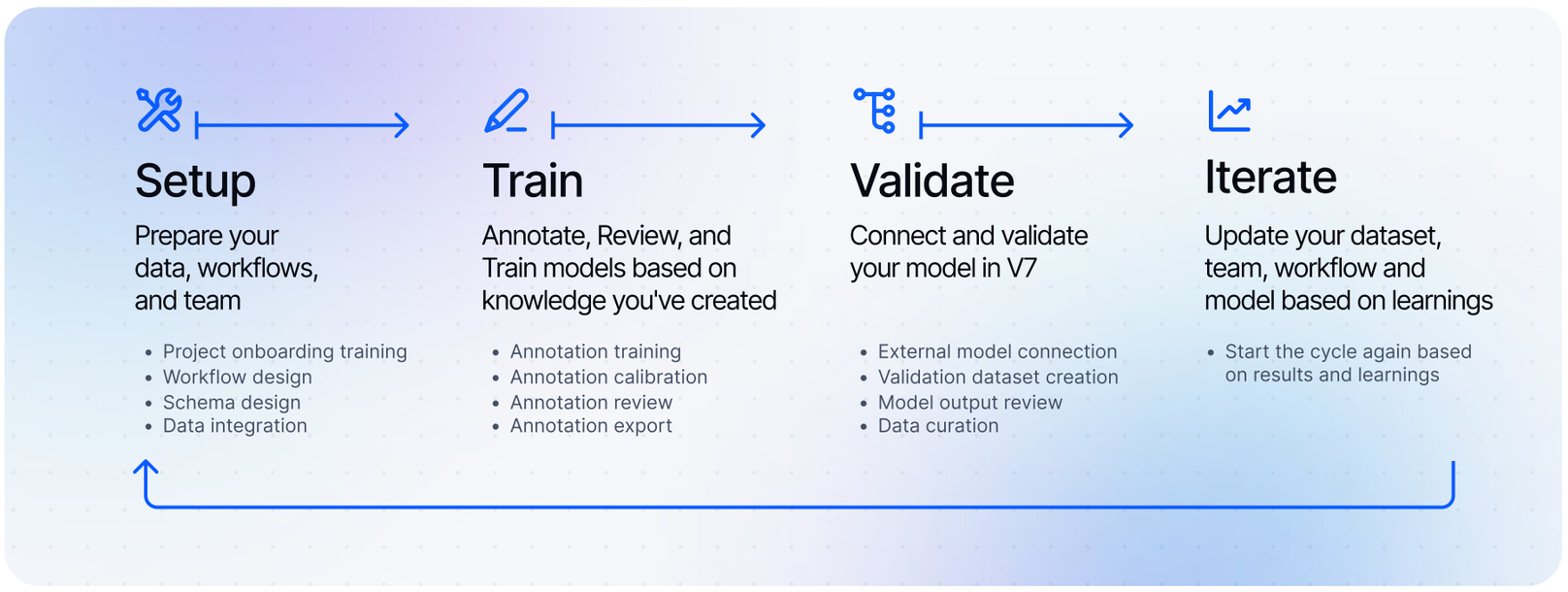
Expert Support

Costs

Value

Building your own framework for data annotation from ground up gives you direct control over your project. But, when you develop an in-house solution, your team is also solely responsible for addressing any issues that may arise. Consulting it with experts may be costly and time consuming as they need to investigate your particular use case, your tech stack and the tool itself.

On the other hand, with an external provider like V7, you have access to a dedicated support team, as well as a wealth of external resources and expertise. This ensures that your team can efficiently address any challenges they encounter, promoting the successful implementation of the platform.



Personalized support, access to webinars, and guides available out of the box are just some of the perks of picking an external training data platform

External support teams are typically staffed with experts in the field, who have a deep understanding of the complexities of machine learning and data management. They can provide valuable insights and suggestions for improving your data workflows and optimizing your training data for better machine learning outcomes.

With a third party platform you get instant access to detailed documentation, tutorials, and other training materials.

9. What are the costs and where will the budget sit?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

Costs

Value

Cost of building or buying your training data platform is undoubtedly another major factor to consider. Companies must weigh the expenses of developing an in-house tool against the potential ROI of outsourcing to a specialized provider like V7.

To make a well-informed decision, most teams weigh the cost of a SaaS subscription against the potential gains in their core product's abilities, ability to generate new areas of R&D and new product lines, speed to release new products, current cost of generating training data, and the current cost of maintaining and updating their internal tool. This can help determine whether investing in a training data management software will make economic sense for your company.

In terms of budget, most companies view training data software as sitting in either Core IT, R&D, Infrastructure or verticalized software spend. Often, teams will source discretionary budget by reducing human spend for creating training data in order to pool it for the purchase of a software that will make their whole team more efficient.

	In-House Development	Platform
Initial Cost	High	Lower
Ongoing Maintenance	High	Included in SaaS Subscription
Scalability	Varies	Highly Scalable
Speed to Release	Slower	Faster
R&D/New Features	Limited	Greater Potential

10. Is building my own training data management tool worth it?

Introduction

Buy vs Build

Data Security

Hosting Options

Effectiveness

Latest Tech

Customization

Integrations

Compatibility

Expert Support

Costs

Value

Ultimately, this is something only you and your team can know.

However, the cost of switching your solution has never been lower, and the potential benefits gained from using a SaaS platform are greater than ever.

No matter which option you are leaning towards, here is a checklist:

	Buy	Build
1	Assess your team's technical expertise and available resources.	Assess your team's technical expertise and available resources.
2	Determine if an off-the-shelf solution meets your team's needs.	Identify the specific features and functionalities your team requires.
3	Research available vendors and evaluate their offerings.	Research the available open-source tools and libraries.
4	Compare the costs of purchasing and maintaining the software with the cost of building it in-house.	Evaluate the costs of building the software, including development time, ongoing maintenance, and potential opportunity costs.
5	Consider the time-to-market for a purchased solution versus a built solution.	Consider the time-to-market for building the solution in-house.
6	Evaluate the level of customization and flexibility required for your team's needs.	Determine if an off-the-shelf solution can be customized to meet your team's needs.
7	Consider the level of support and training provided by the vendor.	Determine if your team has the necessary expertise to support and maintain the software in-house.
8	Make a decision based on a cost-benefit analysis of the options available.	Make a decision based on a cost-benefit analysis of the options available.

Introduction
Buy vs Build
Data Security
Hosting Options
Effectiveness
Latest Tech
Customization
Integrations
Compatibility
Expert Support
Costs
Value

Before starting the evaluation, it's essential to consult your annotation team, engineering team, data operations, and executive stakeholders. To ensure a rigorous evaluation, consult each group and find out what they need from a tool and what they like or dislike about the current system (if you already have one).

We hope that this ebook was helpful and that you will be able to make a well informed decision that is best for your situation.

If you want to check how other companies solved their issues, here's a link to some of our case studies:

- How CurbFlow Improved Model Accuracy and Reduced Time to Delivery by 50%
- Build an Organ Volume Estimation Model Achieving 97% Accuracy
- How MTC is Using V7 to Build AI for Sorting and Segregating Nuclear Waste
- How to Double the Speed of Your Training Data Pipeline

EXPLORE V7 USE CASES →