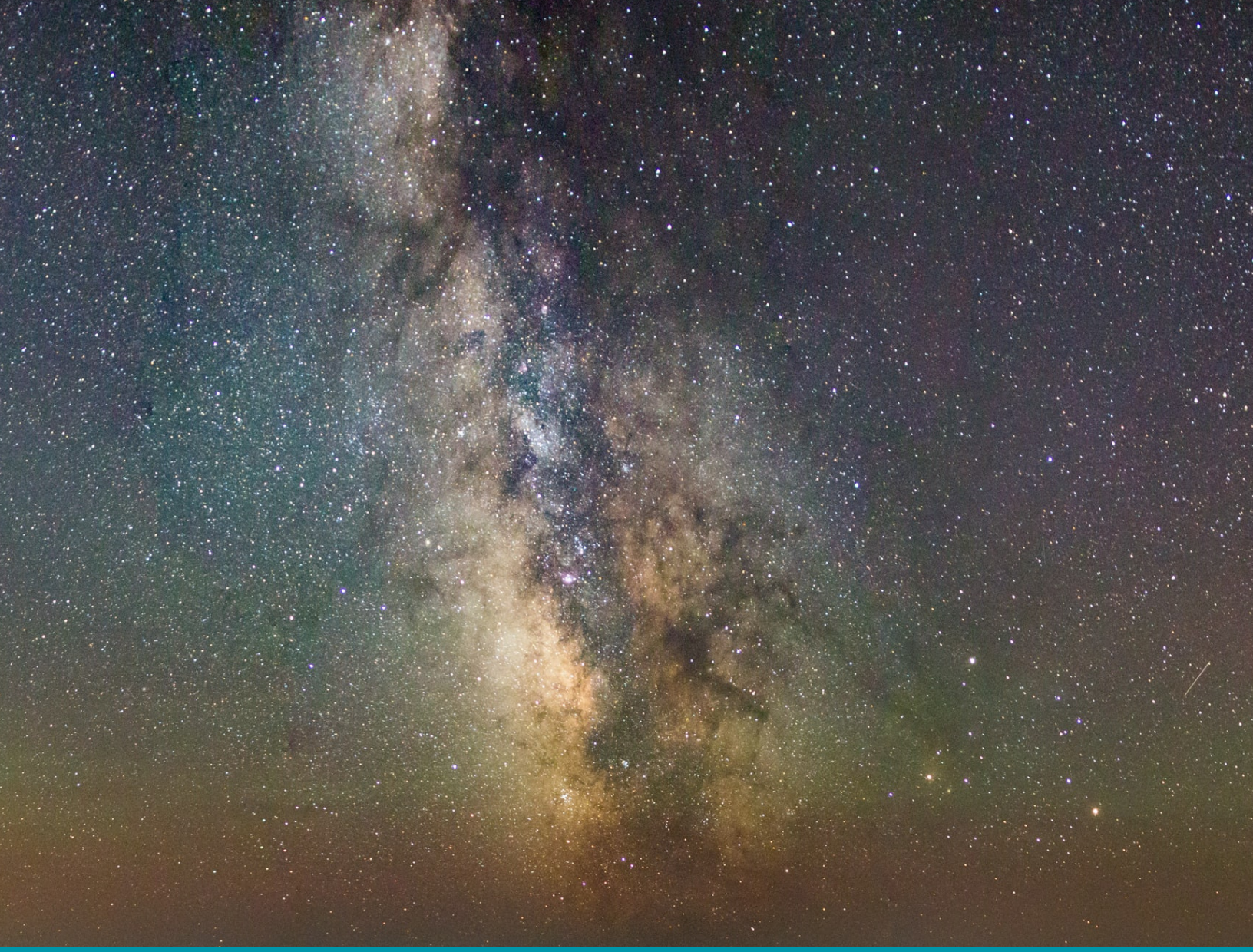


THE ALECTIO LABELING MARKETPLACE

A BETTER WAY TO LABEL YOUR DATA

WWW.ALECTIO.COM

INTRODUCTION

As machine learning and AI have gained broad, near ubiquitous adoption across organizations of all sizes, the ecosystem that makes machine learning possible has grown apace. Whether it's the abundant production of the graphical processing units (GPUs) that train state-of-the-art models, the server farms that store the data, or the data analytics and optimization solutions that make training more efficient, the continued growth of machine learning has both transformed and created myriad industries, both in tech and beyond.

And very few sectors have grown at the pace of data labeling. After all, if you're doing supervised or semi-supervised learning, chances are, you're going to need some data labeling to train functional models. NLP models frequently require labels for tone, parts of speech, entity extraction, sentiment, and more. Object detection models may require bounding boxes, semantic segmentation, or simple categorization labels among others. From audio to images to video to text to LIDAR–the market is broad and the need for labels is increasing year over year.

So what kind of companies are we looking at here? Broadly speaking, data labeling shops can be broken into two categories: **business process outsourcers (called BPOs) and crowdsourcing operations**.

Crowdsourcing companies rely on global workforces that log in from their home computers to label data, either as their primary job or a way to make a few extra dollars during downtime. The biggest of these companies can turn around thousands of data rows in minutes but you'll often sacrifice a little on accuracy and control. For big jobs without a lot of nuance, these are generally a great option.

BPOs, conversely, are smaller. They're a lot more like a typical office space you might be familiar with. They have actual staff instead of relying on a crowd, so you're looking at professional data labelers here, though generally a lot less of them than you'll see in one of the bigger crowdsourcing operations. Many BPOs have specialties (like semantic segmentation) and/or social missions (like employing underserved minorities in rural areas) and, because they generally have staff in a physical office, they're an ideal solution for particularly sensitive data you would never dream of surfacing to a global crowd you couldn't vet.

THREE PROBLEMS WITH THE STATUS QUO

The challenge for a lot of companies and machine learning teams is figuring out exactly which labeling solution is right for their project(s). Because different labeling companies really do vary quite significantly.

We've seen and worked with a lot of these labeling companies here at Alectio, both the bigger shops you've heard of and some of the smaller BPOs that might be new to even the most seasoned ML practitioners. We've interviewed our customers and we've talked to our peers about them. And we think there's a better way: **instead of locking yourself into a long, sometimes rather expensive arrangement with a single provider, we've created the Alectio Labeling Marketplace.**

Our marketplace is a group of some of the highest quality labeling providers in the industry. They have different specialties and so, no matter what labeling work you're doing now and no matter what labeling projects you're planning in six months, one of our providers will likely be an ideal fit.

Now, we'll get into the details of how the marketplace looks, how it functions, what you can expect from us and our labelers, and a whole lot more in the next section. But to start, we want to surface a few of the pervasive issues we see in the labeling space generally. We'll show you how we solved for each as we lay out how our marketplace works.

PROBLEM 1: SPEED

A lot of data labeling projects are slow. This is almost expected behavior for complex labeling tasks like LIDAR or semantic segmentation, but labeling providers are just like every other business: they experience seasonality, they have peak times, their tools break, and, most importantly, they have other clients that, well, aren't you.

In fact, there are a whole host of reasons some of us struggle getting labels in a reasonable time frame, Some companies require submissions by a certain deadline or they won't look at your order till next week. A more established client with a bigger order might take precedence over your project. If you're using a crowdsourcing model and a different company is paying more per label than you are for a similar job, the crowd will default to better pay for similar work.

PROBLEM 2: QUALITY & TRANSPARENCY

If you're in need of a data labeling provider, you likely need a lot of labels. It's not unusual to see tens of thousands, even hundreds of thousands of data rows for a single project. But ask yourself: how do you know the labels you get back are high quality?

Labeling providers often have their own internal metrics that boast of great accuracy. Many have multiple labelers per row so, in the event of a good faith error, additional labelers will be able to override that error by being accurate. But redundancy isn't fool proof and internal metrics can be tough to audit from your end.

Past that, crowd-based solutions have a problem that BPOs rarely (if ever) do: fraud. Because joining a crowd to label data is relatively easy to do, a small number of unscrupulous actors out there will build bots and run scripts to spam bad labels for a quick buck.

The issue is: you can't realistically audit these solutions. If you get a bit unlucky, you can fall victim to the classic garbage in, garbage out (GIGO) issue because you had no real way of knowing how good your labels were before you started training or retraining your model.

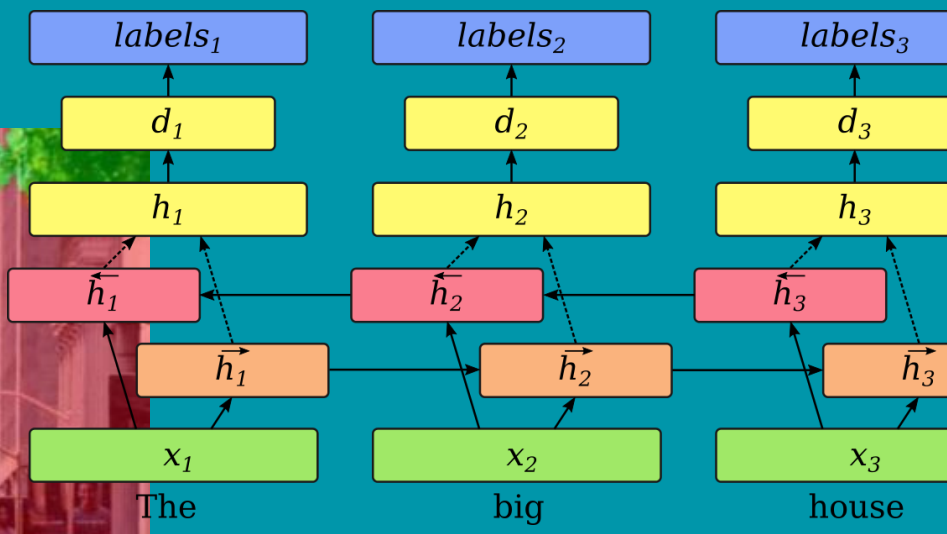
PROBLEM 3: CONVENIENCE & VERSATILITY

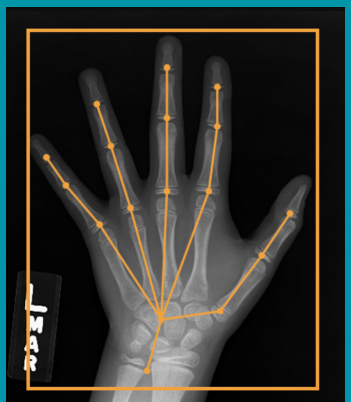
Some of us have a lot of different, sometimes competing, needs for our labeling providers. Part of your team might need transcription for an OCR model while a different part of your team just needs simple categorization of messages for the same project, for example. And anyone who's done their fair share of machine learning knows that sometimes, the priorities for your data, your project, your model, or your workflow will change. Your labels may need to change as well.

The problem is your provider might be expert in what one part of your team needs and not the other. They might struggle if you change priorities. They might not have the proper tooling for a new use case or the fluency you need for a specific language or task. Or, frankly, you might just find out that the provider you chose isn't quite as accurate as advertised.

Whatever the case, be it the fluidity of your work, changing management priorities, mediocre accuracy, or something else, getting locked into a long-term arrangement can really hurt the momentum of your machine learning projects.

There are smaller issues with data labeling as well, but those are three of the biggest ones. Thankfully, our marketplace solves each of them. We'll show you how on the next page.





HOW OUR LABELING MARKETPLACE SOLVES THOSE PROBLEMS

Alectio's labeling marketplace is a collection of the best labeling providers in the world, all at your fingertips for your next project. You can rely on our recommendation based on your specific project and needs or choose from a list we provide. We'll get into those when we walk you through the process in our next section, but for starters, let's talk about what you can expect from our providers and how they solve the issues with data labeling as it is today.

SOLUTION 1: SPEED

For starters, if you'll recall our introduction, all the providers in our marketplace are BPOs. That means they can guarantee privacy, speed, and a whole host of other important factors. Crowdsourcing operations simply can't do this. Because BPOs manage their own workload, you can count on them (and us) to be transparent about how long your labels will take.

In fact, at Alectio, when you choose a labeling provider, **we check to make sure they can start working on your project immediately.** You don't go into a queue five projects long while you wait for them to finish more lucrative contracts. If your first choice can't work on your project, we'll move to your second, then your third. It's our goal to get you accurate labels as quickly as possible so your team can stay nimble and agile.

SOLUTION 2: QUALITY & TRANSPARENCY

We vet every labeling provider in our marketplace. We check their quality and if they fall below our standards, they're removed from our marketplace. This is exceedingly rare but just know that we stand behind our labeling providers. They're in our marketplace because we trust them and we trust their quality.

We want to be as unbiased as possible here. Our feedback on both our marketplace and our labelers' performance is data-driven and dynamic. We factor their performance–their accuracy, their speed, and their cost–into our suggestions for providers. We want our best partners to work on your data.

Additionally, we give those providers feedback to help them improve as well. Say one is great at bounding boxes but a little shakier doing named entity extraction. We not only let them know this, but we quantify it so they can improve.

The point is: we don't play favorites with our providers. We help you find the best one for your specific project based on their accuracy, their expertise, their availability, and more. We don't take a cut of our labeler's profit either, so our only goal is to help you find the best partner for your immediate needs. There's no other incentive.

SOLUTION 3: CONVENIENCE & VERSATILITY

Having a whole host of different labeling companies in our marketplace means we can handle a wide breadth of tasks: far more than you'd get from any single provider, even the bigger, crowdsourcing operations.

So why is this important for you? Well, instead of getting locked into a long-term contract with a single provider, you can choose specific providers for each project, based on their availability, their expertise, and your budget. If you find one provider you love, don't worry! You can choose to forego our recommendation and work with them each time you need specific labels.

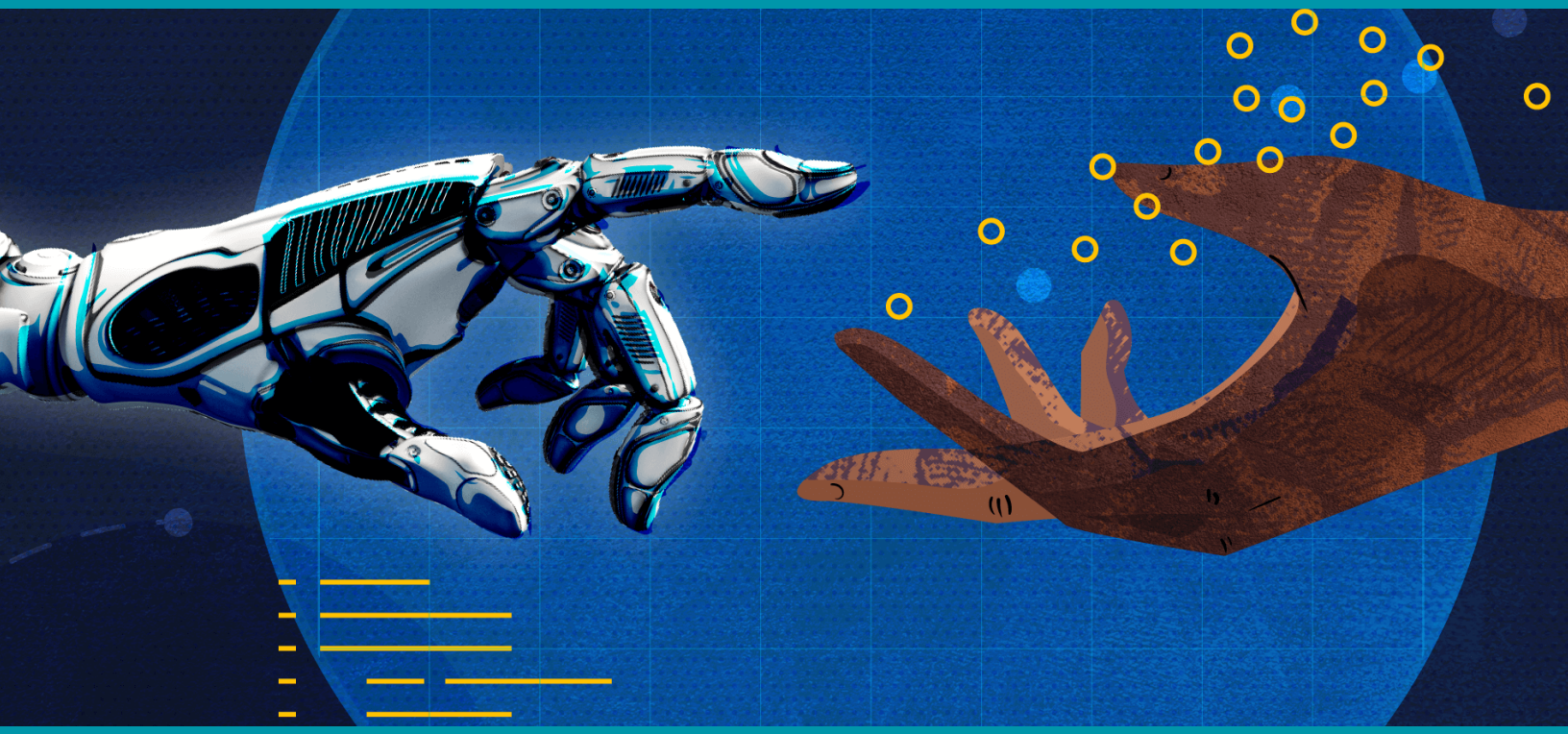
The idea is that the Alectio Labeling Marketplace gives you full control over each individual project. And for teams with changing priorities, teams that need fast turn-around, specific accuracy thresholds for specific projects, different labels for different projects (say, both LIDAR and segmentation), or anyone else who doesn't want to use the same provider for each and every project? Our marketplace is there for you.

HOW ALECTIO'S LABELING MARKETPLACE GIVES YOU FULL CONTROL

Our goal with the labeling marketplace is simple: to **pair you with the best labeling provider for that specific project**. That's it. No long-term contracts. Everything is on a project-by-project basis. For example, you can choose one provider for an image classification workflow and another for a bounding box task. They can work in tandem or one after the other. The point is that you'll have the ability to choose a new provider for every project––and we'll be there to help with suggestions along the way. (Just as an example here: if you need, say, a Japanese translation of English documents, we will only show you providers who can handle that specific task.)

And though our labeling marketplace works incredibly well with our SaaS data curation platform, anyone is free to use our labeling marketplace whether or not you're leveraging that data curation technology.

Our marketplace has a few big advantages over signing with a single labeling provider for all your projects but fundamentally they come down to one important thing: control. Control of your data, control who labels it, control over your priorities, your model, your data validation, all of it. And it starts with what data your provider should label––and what data a machine can help you label instead. It starts with true human-in-the-loop machine learning.



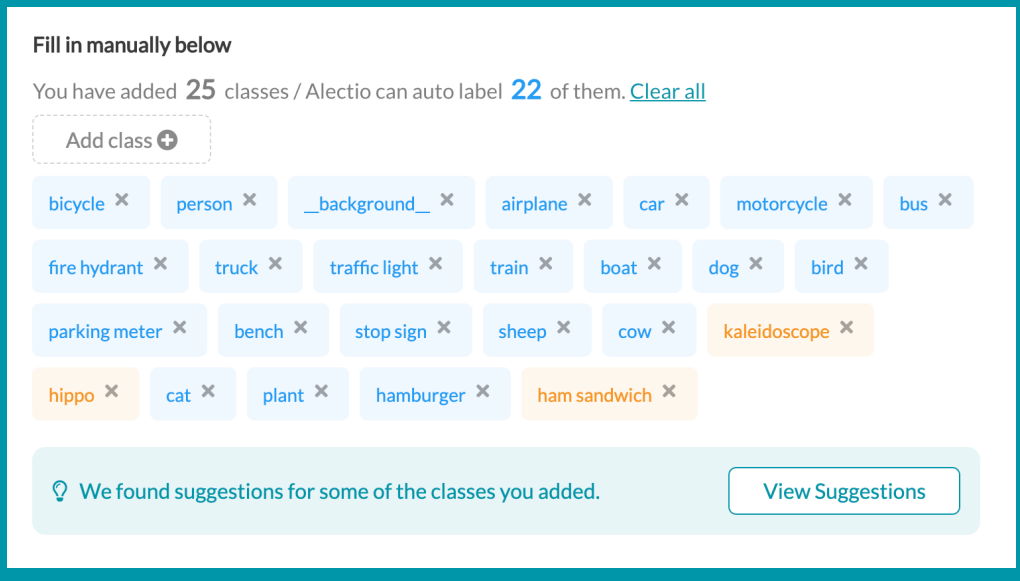
UNLOCK TRUE HUMAN-IN-THE-LOOP MACHINE LEARNING

At a fundamental level, there are two ways to label any piece of data: by a human or by a machine. Broadly speaking, machine labels are cheaper and faster but they may struggle with accuracy overall and edge-cases specifically. Boutique or newer use cases often have trouble finding algorithms to label data accurate. But machines can do yeoman's work with some of the easier labels, saving organizations significant time and money. Human labeling costs more, but can handle tougher data, subjective data, or those use cases where currently available algorithms simply don't understand.

Deciding which data goes to human labelers and which data gets autolabeled by a machine is something called **human-on-the-loop machine learning**. One of the most common ways to do this? A machine oracle evaluates a piece of data and its own confidence about it's accuracy on that piece of data. For items under the accuracy threshold, that data is sent to human labelers. For data the model is more confident about, it can provide the labels for you. In other words, the machine labels what it understands well and humans label what it doesn't.

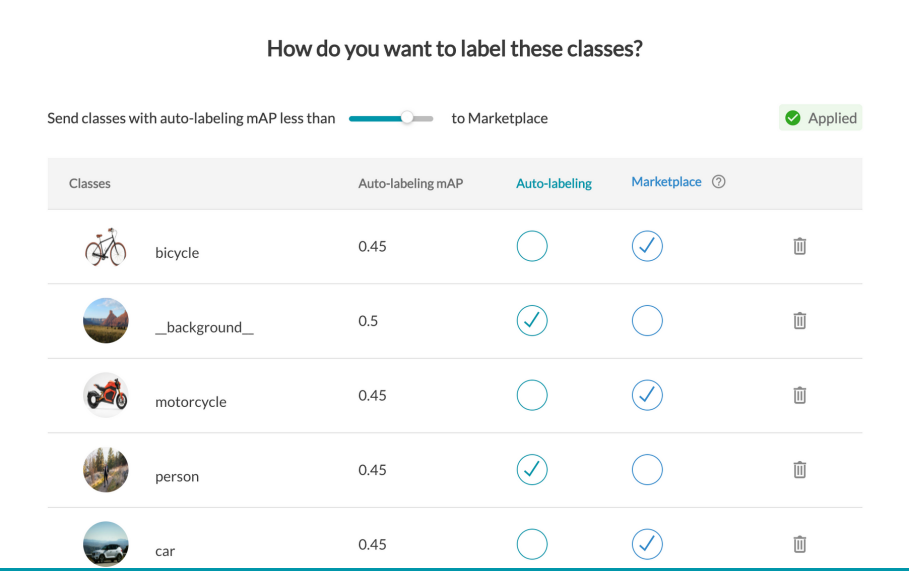
Alectio's platform and our labeling marketplace are ideally suited for this human-in-the-loop workflow. It works like this:

Say you're training an objection detection algorithm. As you enter the classes you need labeled, classes that appear in blue can be autolabeled by Alectio. Classes that appear in orange cannot be. For example, check out the image on the right:



Here, you can see we have models that can label some classes but not others. Any classes in orange will need to be sent to a labeling provider while the ones in blue can be autolabeled.

When you're actually paging through our workflow, you'll a page where you can choose where each class goes (we'll show that to you on the next page).



Here, we'll show you a relevant metric abouot how our autolabeling model will perform--in this case it's mAP or mean average precision, but it could be precision, accuracy, or a different metric depending on your specific use case. You can select which classes you'd like sent to a labeling provider and which you'd like to try autolabeling with. You can choose that on a class-by-class basis or with the slider above.

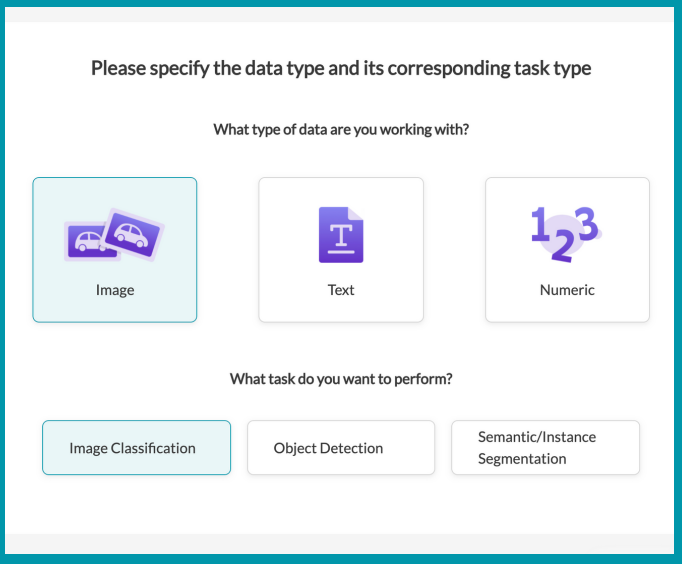
It's worth noting that, for certain use cases, even if an oracle says the autolabeling will be accurate, that particular class might be incredibly important to your model's performance, so you'd choose to have it hand-labeled. Something like accident data for an autonomous driving model is a great example here.

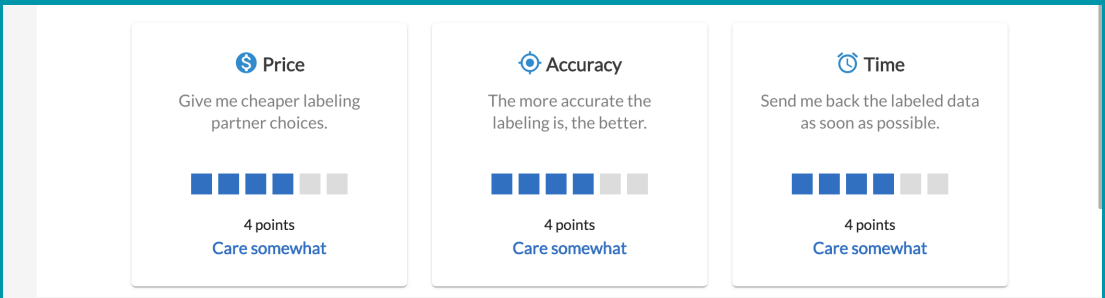
The point is: this is true human-in-the-loop machine learning in a single place. It's human-in-the-loop the way it was intended to be. We'll let you know what classes we can help with and which you should probably sent to a labeling provider. But again: this is about control. The choice is up to you. We're just doing our best to keep your labeling costs down.

DIFFERENT LABELERS FOR EVERY TASK

Every data labeling operation excels in different areas. Some could be great at certain language capabilities, others might have best-in-class tooling for LIDAR or audio, others might boast the fastest turnaround times or the best accuracy. Some labeling providers even have certain government clearances or are HIPAA complaint.

At Alectio, we're here to help you find the *right* labeling partner for every individual project. It's one of the reasons we ask you up front what kind of data you're working with. Selecting "image" and "image classification" in this case means, when we recommend providers to you later in the process, we're going to show you ones with great expertise working with image classification workflows.

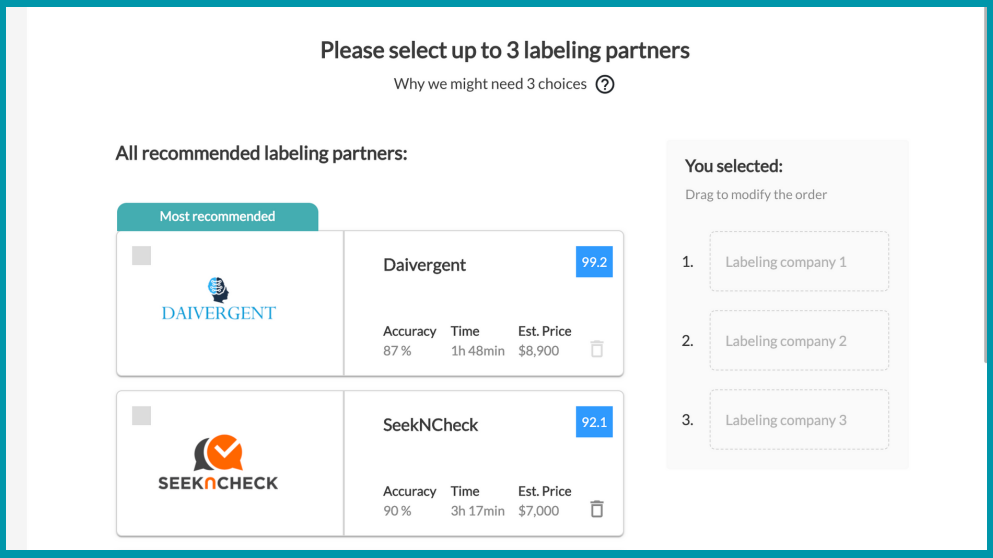




You have full control when choosing your provider as well. Once you've uploaded your data and your instructions for labeling, what you'll see next is a screen that will help us find the best provider for your current project.

The first option is that you use our recommendation. These recommendations are based on our internal quality metrics, what data each provider can excel with, and a whole host of other considerations. But we also ask you your priorities for this project. We're going to ask you to **rank what's most important for your project: price, accuracy, or speed**. Now, keep in mind: we've vetted all of these providers and we stand behind their pricing, their accuracy, and their velocity, but for some projects, you need labels to be turned around extremely quickly whereas for others, you might have budgetary restraints. You weigh which of these factors is most important for the job you're working on and we'll provide a list of the recommendations.

Based on historical data, your labeling task type, our own internal metrics, where the labeler is located (essentially, if they're operating when you launch your job), and a few other levers we'll estimate accuracy, time, and price for each provider. You click the ones that look good to you and then rank them yourself by simply dragging and dropping.



We ask you to pick several because we want you to get your labels quickly. What happens is we reach out to the first provider on your list and see if they're available. If they aren't or they're slow to respond, we'll move down the list to your second choice, and so on, down the line. This is in fact part of our contract and agreement with each labeler--that they have a limited time to approve each request--because they understand our clients are coming for speed and convenience.

DATA LABELING VALIDATION, INSIGHTS & TRANSPARENCY

You don't just get full control of who (or what) labels your data. You get a complete understanding of those labels themselves. See, at Alectio, when your labeling task is finished, we don't just provide labels back to you in a downloadable .csv.

We provide a robust, nuanced reports about those labels. You'll be able to see things like:

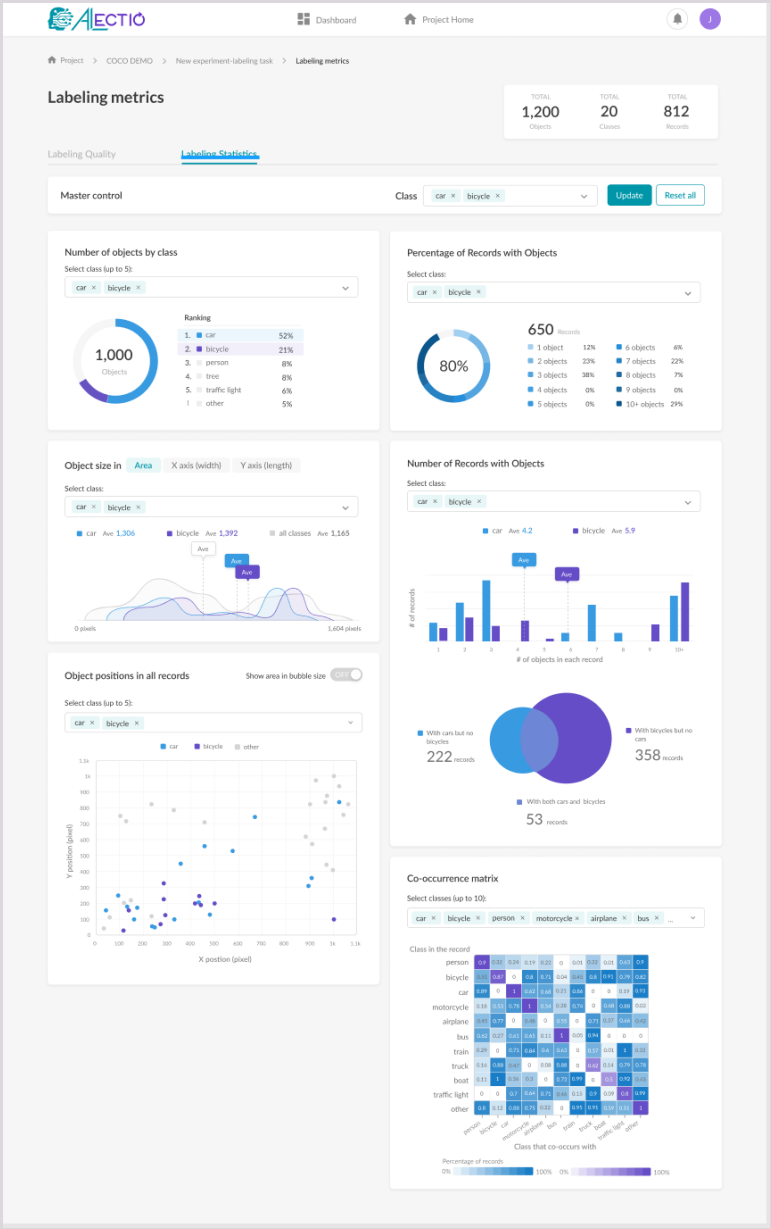
Accuracy predictions: We make smart predictions about which classes are most likely to have mislabels in them so you can spot check and validate immediately.

Filterable insights: Your reports allow you to filter by class and predicted accuracy of each label so you can see, at a glance, a class (like dog) with a predicted accuracy under 90%.

Co-occurrences: Your report also contains matrixes where you can easily view co-occurrences of certain classes so you can adjust your data collection and labeling practices. For example, if every picture of bicycle also contains a rider, you may need images of bicycles by themselves for your model.

Density of labels and classes: Your report shows the distribution of your labels over your data rows. In other words, at a glance you'll know if most of a certain class occurred many times in a few instances (perhaps one image has 100 bikes in it while your entire dataset only had 140 bikes overall).

Location of labels: For images, we can show you where certain classes most often appeared. For example, maybe most of your cat labels appeared in the bottom right hand quadrant of your images.



These insights can help you smartly collect your next batch of data or augment your current one, perhaps by collecting more instances of certain classes that are all clustered in a few discrete examples. They can help you understand your dataset generally, where it has enough examples and where you need more. Classes with lower accuracy can help you discover where your labeling instructions weren't quite as good so you can improve and iterate on those for the next round. They can help you validate, at a glance, every class by checking ones with lower predicted accuracy and seeing if they were in fact labeled to your liking.

Put simply: our labeling quality report gives you a deep and nuanced understanding of the labels you received, your dataset as a whole, and how they are likely to impact your model. How you choose to use the insights is up to you and your team.

PARTING THOUGHTS

Any individual labeling provider isn't right for every potential labeling project. Instead, at Alectio, we're committed to helping our customers find the perfect labeling partner for each individual project. Our growing community of expert labeling partners can handle every major data labeling project, regardless of your priorities, and, because they're BPOs, they can even handle your most sensitive labeling needs. There's no need to sacrifice on speed or flexibility or price or lock yourself into a long-term contract either. Just spend a few minutes finding the best labeling provider for whatever project you're working on now. It's that simple.

The Alectio Labeling Marketplace is available for both our SaaS data curation platform users as well as a stand alone service. If you'd like to try it out, just get in touch at info@alectio.com. We'll get you started.

ABOUT ALECTIO

At Alectio, we help the most innovative companies in the world train better machine learning models with less data. Our platform employs an ensemble approach that includes active learning, reinforcement learning, meta-learning, deep learning, and more to identify what data is actually helping a model learn and what data is holding it back. Alectio uncovers the smart data inside your big data, unlocking model performance and saving machine learning experts both time and money. Visit us at alectio.com