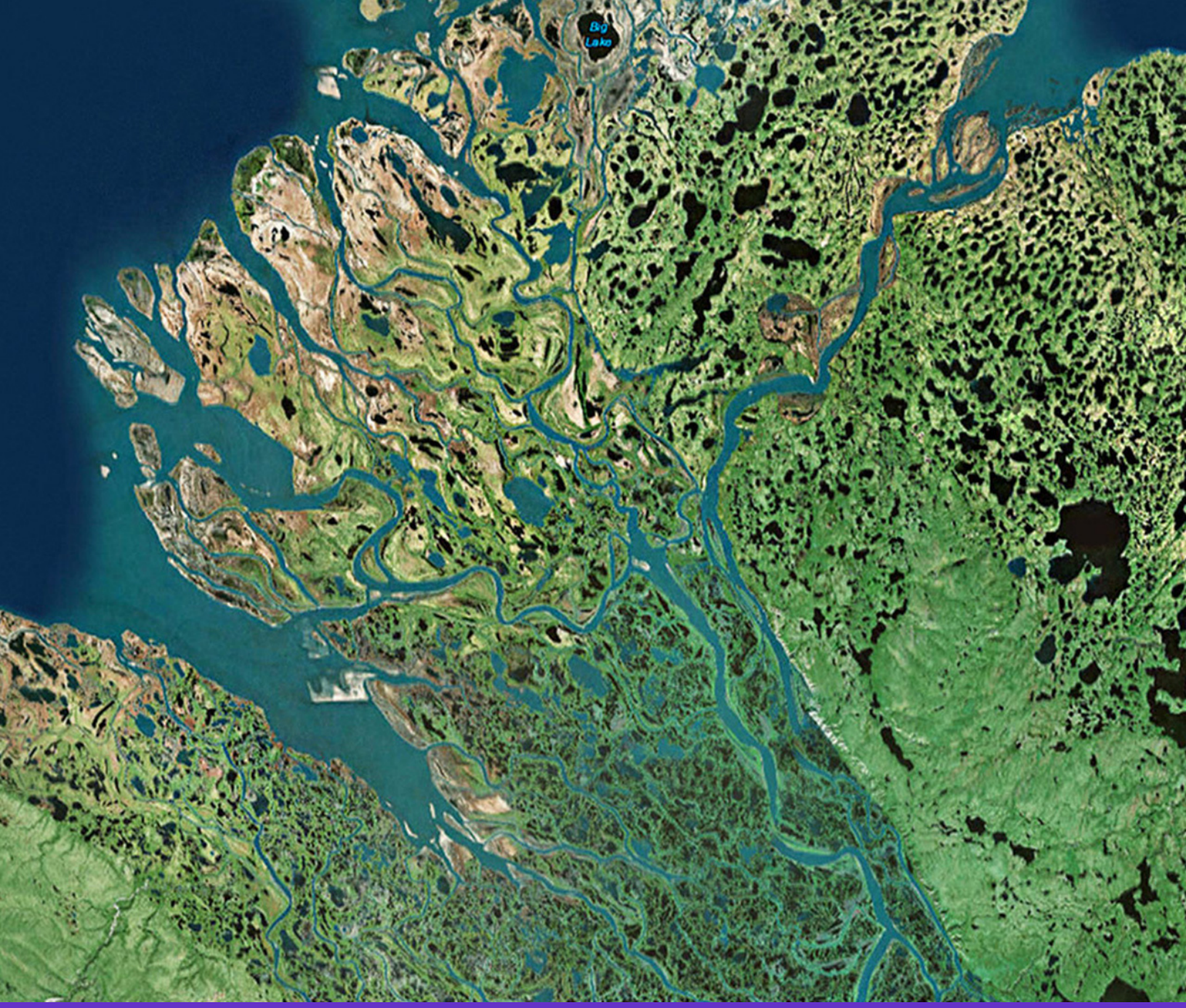


HOW ALECTIO CAN TRANSFORM YOUR MACHINE LEARNING PROJECTS

WWW.ALECTIO.COM

HOW WE THINK ABOUT DATA AT ALECTIO

There's this pervasive misconception in machine learning that more data is always better. Is your model not converging? Add more data. Is your model overfitting? Add more data. Is your model below your accuracy threshold? Add more data.

You get the idea.

Now, we all understand that more data really can help. If your model hasn't seen enough of a class to make confident predictions about that class, chances are, you really do need more data. If you're in a domain that evolves constantly or faces seasonality challenges, you're going to need to retrain with new data to stay relevant. If you need to add a new class or behavior to your model, you'll likely need extra data for that. Again, you get the idea here too.

But there's something hidden in this concept that we all understand and don't think about nearly as much as we should. And that's that **all data is not created equal**.

Instinctually, nearly every data scientist and machine learning practitioner understands this. They know that datasets in the real world are rarely the carefully curated datasets you'd find in academia. They know that sometimes, data can be redundant, images can be blurry, third-party labeling companies can mess up. Not every data row you feed your model will help it. Some will actually do the exact opposite--here, it's helpful to remember the old "garbage in, garbage out" (GIGO) adage.

In fact, at Alectio, we believe data can largely be broken into three categories: **helpful, useless, and harmful**.

Helpful data helps your model. It's well labeled and has utility. When you're adding data to a training set, this is your ideal. This is the good stuff.

Useless data is a wider category. This data isn't going to help your model but it isn't going to necessarily hurt it either. This data might be redundant. It might be images or text with little informational content. It could be subjective corner cases or just simply blurry pictures or other low relevance data. In fact, useless data is often high quality data–it's just that it's irrelevant to your model's performance.

And then there's **harmful** data. Harmful data makes your model worse. It could be data that's corrupted from a bad sensor. It could be a harmful synthetic example. It could be data that's bad quality. It could be data that was mislabeled during the labeling process. But whatever it is, harmful data reduces your model's effectiveness. It's garbage in, which means garbage out. It takes you backward.

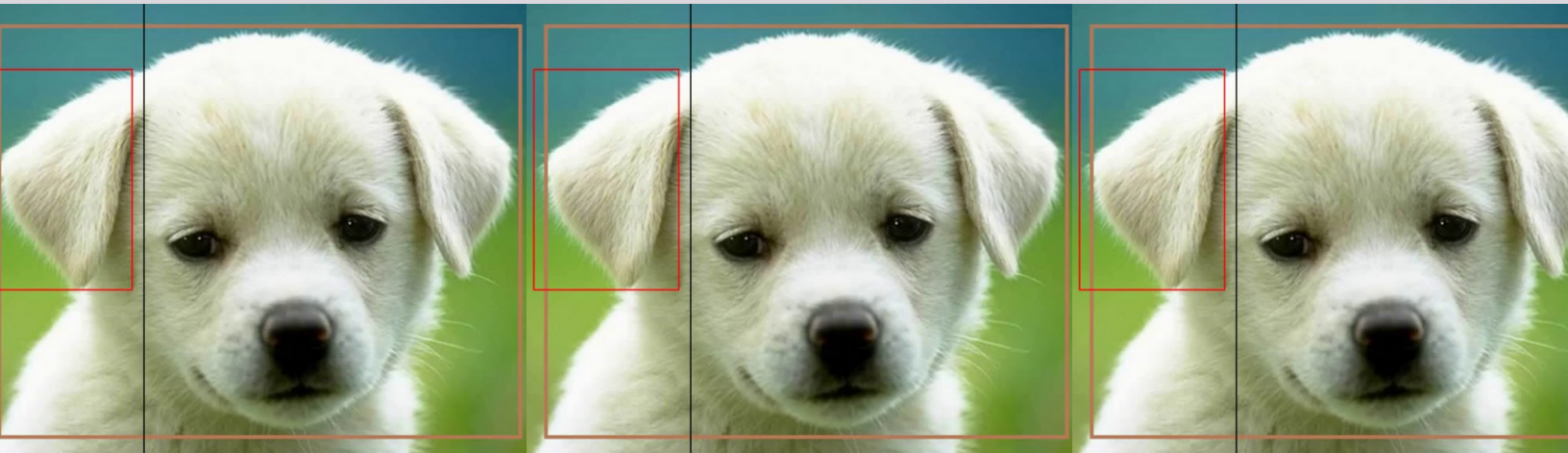
Which brings us back to where we started: if you're simply adding data to your models, do you know what kind of data you're adding? Because useless or harmful data isn't going to help, you can't simply add more and more of that and hope your models will converge. They just won't.

This is all fundamental to our goal at Alectio: to help you find the data your model needs. To help you diagnose your data so you can **find the smart data inside your big data**.

BUT IT'S ABOUT MORE THAN JUST QUALITY

Now, there's a bit more nuance that simply defining data as useful, useless, and harmful. In fact, within that bucket of good and useless data, there's something more to think about. You want to understand not just data quality but data value.

See, a piece of data can be high quality but not actually valuable for your model. Think about a redundant image, for example.



Those dogs are all well labeled. They're clear. They're high quality. But are they going to be **valuable** to your model? (It is worth noting that *some* redundancy can be valuable–data augmentation practices prove this–but those images are in fact augmented, not identical. Past that, the idea we want to drive home here is the distinction between quality and value and that high quality data is not necessarily high value data to your model.)

That said, the question isn't just about redundancy. It's about value, generically. Data or batches of data that improve your model are valuable. And to be sure, all data that's valuable needs to be high quality. But all data that's high quality isn't necessarily valuable. Not to mention that both value and usefulness are contextual––the same data row labeled the same way will have a different effect based on your model and use case.

That high value data is exactly what we help you find here at Alectio. Knowing which data is driving model performance is crucial to training and retraining your models faster, cheaper, and with better accuracy.

Which really is the driving force behind what we do here: we make sure your models are trained with the best possible data. Think of it like this: if you're just pushing all your data into your model, it's just eating everything in sight. It's having both its veggies and dozens of bags of Doritos. Alectio puts your models on a diet, making sure they ingest only the best food so they're lean, lithe, and effective.

THE ALECTIO MARKETPLACE

So far, we've discussed our core technology, the Alectio data curation platform. But we also have another part of our solution that we're really proud of: the Alectio Labeling Marketplace. Here's how it works:

Our marketplace is a collection of labeling providers and BPOs we've vetted and found are ideal partners for our solution. When an Alectio customer needs labels, they can access this marketplace. They tell us what their biggest priorities are (cost, speed, or accuracy), what data they need labeled and how, and provide instructions, all very similar to how things would work in a typical labeling partnership.

But instead of being locked into a long term contract with a big labeling company, Alectio customers can choose from our marketplace of providers on a project-by-project basis instead. There's an important reason why we like this approach and that's this: Alectio works best with more training runs with smaller batches of labels. Big labeling companies cannot always be relied on to provide these.

But our marketplace can be. It's part of our agreement with our partners, in fact. When you select the one that fits your criteria, labeling providers get a limited amount of time to take the job. If they can't, we move to your next preferred candidate. This ensures your labels are turned around quickly and used to improve your models.

Now, you can certainly use our service just to get your data labeled. This is a great option for smaller companies who might not have the need for a long contract with a bigger company or ones who are still new to labeling. But Alectio's Labeling Marketplace works hand-in-hand with our platform to run multiple training runs quickly, so you can understand and improve your data and your models.

THE PROBLEMS ALECTIO SOLVES

Understanding what data truly matters solves a whole host of really pervasive, tenacious problems for machine learning teams. Broadly, when we think of what our technology does, we break it out into five big buckets. There's some overlap here, certainly, but we'll be going over each in detail below. Here are the ways our clients are benefiting from our platform right now:

Saving money
Saving time
Building better models
Collecting and preparing better data
Understanding their data and their models

We'll start with one we know your finance team will be happy to hear about:

SAVING MONEY WITH ALECTIO

Let's face it: machine learning with big data is expensive. It costs a lot to collect data, to store data, to label data, to hire the best talent, and to retrain models in an ever-changing world. But when you know what data is truly valuable, you can really bring your cost outlay under control.

Here are a few ways we can help:

REDUCED LABELING COSTS

Because Alectio can identify the data that really matters to your model, that means you can prioritize labeling the data with the most utility, not the data that might hurt your model or the data that won't much help it.

In other words, instead of sending large batches of data to your labeling provider (or your in-house labelers), you can send them only the information that really matters.

How do we do this? Well, as we mentioned before, Alectio works by "listening" to your model and your data as you train. Because it does this, it can identify the types of data that are most meaningful to your model. You're in charge of making the determination about what to do with that information, but labeling less data (and only the best of it) is a sure way to reduce your budget and move that spend to something a little more glamorous.

Past that, users of our marketplace can benefit from reduced labeling costs as well. Firstly, they can choose to prioritize cost over, say, speed. This will help us find vendors who work for their budget. Additionally, they won't be locked into long term contracts and instead will enjoy smaller, more focused partnerships. We also vet labels as they come in, reducing the need for additional labeling redundancy.

(We could write pages upon pages about this and, in fact, we already have. If you're curious about how to reduce your data labeling costs, download our whitepaper at https://alectio.com/data-labeling-white-paper)

REDUCED COMPUTE COSTS

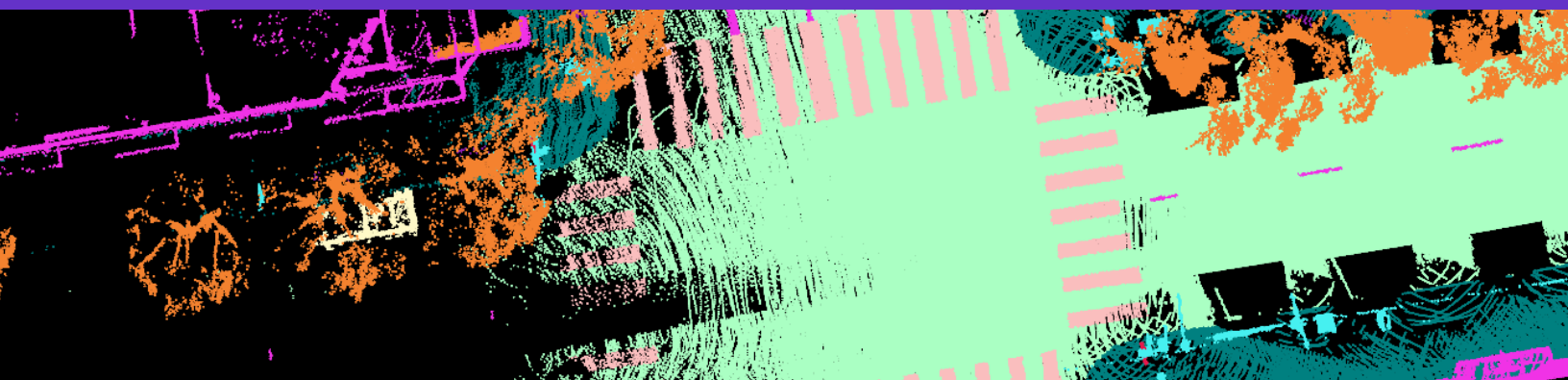
Of course, operational machine learning costs don't come just from data labeling. Compute costs can really add up too.

So what drives these compute costs? Training and retraining big, data greedy models is certainly a culprit. Since Alectio helps you identify the best data for your models, that means each training run you do requires less data than it would've before. We've found that a **majority** of our customers' data often isn't helping or is even actively hurting their models. Doing away with that data improves not just model performance but it reduces compute costs.

SAVING TIME WITH ALECTIO

Training production-ready models with big data isn't just expensive--it's time consuming. From the actual time training (and retraining) to the attendant issues like waiting for your labeling partner, doing machine learning right simply takes time. But it doesn't need to take as much time as you're likely spending right now.

Here's how we can help:



USING LESS DATA MEANS LESS LABELS. AND LESS LABELS MEANS LESS TIME WAITING ON YOUR PROVIDER

Let's face it: data labeling can be slow. This is especially true if the data you're getting labeled is complex (think LIDAR labels, semantic segmentation, crowded bounding box jobs, and the like) and generally because most labeling is still done by hand.

Add to that that data labeling companies have different priorities than you do. If they have a crowdsourced model, labelers generally choose jobs that are easy or that pay more than you may be able to afford. Smaller shops and BPOs, well, they may prioritize other clients over you or simply not be able to turn around labels at the speed you need to get your models into production. If a labeling job needs a specific expertise (for example, translating a document from French to Japanese), you could be waiting for the experts to finish as well.

When you send only the best, most beneficial data to labelers, well, you're sending less data than you otherwise would. That means their turn-around time will be faster. And so will any validation of these labels.

Now, add to *this* that you can use Alectio's Data Labeling Marketplace for projects on our platform. As we mentioned above, our marketplace providers have a limited amount of time to accept your job or we move to the next one. We vet them for these reasons specifically! They understand that quality and speed are paramount for our process so you'll get your labels back quickly and at high levels of accuracy.

LESS DATA MEANS FASTER TRAINING TOO

Put simply, smaller batches of data typically translate to quicker model training. Moreover, if you have a model that needs constant updates or significant amounts of retraining, each subsequent round of retraining goes far more quickly as well.

That means you'll be faster pushing models to production, faster retraining them, iterate faster on the models themselves, and be more *agile* as an entire machine learning team. It means less waiting on retraining cycles before you can make changes. It means increased efficiency in addition to the monetary savings we outlined in our last section.

But enough about saving time and money. Let's get into how your models will improve too.

YOUR MODEL IS ONLY AS GOOD AS THE DATA YOU TRAIN IT WITH.

LET US HELP YOU FIND THE SMART DATA INSIDE YOUR BIG DATA.

BUILDING BETTER MODELS WITH ALECTIO

Cleaner, better data directly impacts model performance. Not only does it make training cheaper and faster, but it helps ensure that your model isn't polluted with harmful or inconsequential data. And that leads to faster convergence.

Here's a bit more about how Alectio's platform can improve your models.

LEVERAGE MODEL DIAGNOSTICS

Alectio's technology can help you diagnose your overall approach to a particular model. This has been something we've seen really resonate with some of our most curious customers.

Say, for example, while using us that you find out a lot of the data you're feeding your model isn't particularly helpful (i.e. it's "useless" in the good/useless/harmful breakdown). What exactly does that mean? Well, it could mean that your model is too shallow or inappropriate for the use case you're currently working on. Or it could mean that your model doesn't need additional data for certain classes (we'll get into this in a second).

The point is, by understanding what your model wants to see, you can make intelligent decisions about how to tweak its hyperparameters, find new approaches or build new models to handle classes your model might not be handling, and simply understand your model in ways that would be far more difficult without Alectio.

USE CURATION AND FILTERING TO SOLVE OVERFITTING AND OTHER ISSUES

As mentioned above, some useless data is just data your model already understands well. But if you're constantly feeding it this useless or redundant data, you're probably aware that this can cause overfitting. Alectio gives you a programmatic way to make this determination.

Additionally, say you discover that there's a preponderance of **harmful** data in your training data, data that makes your model less accurate or hurts whichever other metric you're currently optimizing for. You can simply remove that and retrain. You can look at the data itself and diagnose if it's poorly labeled. You can decide if it's even data your model should be ingesting at all.

UNDERSTAND YOUR TRICKIEST DATA

In a similar vein, Alectio's technology does not only help you prioritize the most impactful data, but also rank which data is most likely to be mislabeled, or, for that matter, which data has already been mislabeled.

Knowing this allows you to take all sorts of measures to deal with that issue. You can create real human-in-the-loop workflows, sending your most complex data to your most relevant labeling provider. You can add additional layers of redundancy in your labeling processes to make certain that labels are accurate. You can collect more of that data type to offset issues. You can make smart predictions about similar types of data you should take special care of in the future. You can fix potentially bad labels

But what it all gets you is better models. See, it's not enough to train your models on any kind of data. You want to focus on training with the **right** data. You'll get better models, faster, and cheaper. But still, you can do more with Alectio.

THE MAJORITY OF A TYPICAL TRAINING DATASET IS EITHER USELESS OR ACTIVELY HARMFUL TO MODEL PERFORMANCE.

WE CAN HELP.

COLLECTING AND PREPARING BETTER DATA WITH ALECTIO

From data collection and generation to training and filtering, Alectio's platform unlocks better data practices across your entire machine learning organizations. Here's how:

LET MACHINES UNCOVER BIAS, NOT HUMANS

When you're currently deciding what data you need to collect or retrain your models with, how are you doing it? Chances are, you're relying on human experts. And though people are very good at building models, we're not always great at understanding what their biases are. Here, it's instructive to think about the fact that deep neural networks often rely more on texture than shape when predicting objects (in other words, the texture of a dog's fur vs. the shape of a dog). This isn't necessarily intuitive and, if you let humans choose which data a machine wants to see most, well, they might choose the wrong data.

Alectio helps solve that. Our tools show you which data your model learns most from and which data your model wants to see next. You don't have to guess at its biases because we're listening for what those biases actually are and how they change over time as you train and retrain your models.

The point is: as humans, we're actually not all that good at anticipating what a model's biases might be. With Alectio, you don't have to do that. You can just let the model tell you what it needs to learn next.

MODIFY YOUR DATA COLLECTION PROCESSES

Since Alectio can help you uncover which data is helping your model and which data isn't, you can make real strides in your data collection efforts. Here's an example about what we mean:

Say you're working on autonomous driving models. You use Alectio to diagnose your data and the results tell you that your model is doing quite well with highway data but is hungry for data from rural roads. Suddenly, you have real insights you can use to change your strategy. Your drivers don't need to sit in traffic jams to get data that's not helping your models. They can take a scenic drive down a country road instead.

Alectio can also help you find co-occurences of specific objects in your data (like bikes and riders) so you can tweak your data collection practices accordingly, and in this specific case, collect more bikes without riders to avoid biases. We can show you the size of objects of a specific class, the density of information on a record-by-record basis, the evenness of distributions, and a whole lot more.

Essentially, if you have to make active efforts to collect data, Alectio can point you towards the data you should be collecting right now and the sort of data you already have enough of. And that can improve model performance immensely.

FILTER DATA ON THE EDGE

Sticking with the autonomous driving example, Alectio can be really powerful for something we call “on-edge filtering.”

See, autonomous vehicles ingest a lot of data. That data is stored, sent to the cloud, and eventually, hopefully, labeled and used in model training. But experts know a lot of that data just isn’t that helpful. A car idling at an empty intersection, for example, may not have much utility for the machine learning team. But the data will still be stored and, when you’re dealing with a fleet of vehicles, that’s very costly.

With Alectio, you can filter out this data at the point of collection. For data we’ve predicted will hurt or be useless to your model, you can decide to simply not store it. You can choose not collect it. You can choose to not get it labeled. You can ignore it because it won’t help. By filtering a lot of that data out, you could actually transfer the important data as you collect it instead of requiring wired connections and uploads later.

Granted, this is a feature that isn’t relevant to everyone, but for autonomous driving, robotics, sensor data, etc., it can be a big driver of success and savings for your team. Essentially, if you have an active data collection process vs. a passive one, this can be a real game changer.

UPDATE YOUR SYNTHETIC DATA GENERATION PRACTICES

It follows that having intimate knowledge of your data's value and your model's needs also opens up new avenues for data *generation* as well.

Think about a use case like facial recognition. Your model might be very confident predicting white or European faces but need more data for people of African descent. Alectio might show you that those non-white faces are the most helpful data for its continued improvement, in fact. This can inform your synthetic data generation efforts without relying on your model to fail before you notice the problem. You'll know as you train what data is really helping your model and you can create (or augment!) data to fit that need.

Synthetic data generation can be both very valuable and done incredibly well. There are some companies out there now doing a fantastic job creating that data. But it's often harder to know which data to create than to create great data. Alectio can solve the "which" problem so you can focus on the creation.

UNDERSTANDING YOUR DATA AND YOUR MODELS WITH ALECTIO

Lastly, we'd like to discuss explainability, active learning, and how Alectio can do more than just simply tell you which data is useful: we can tell you why that data is useful.

Alectio didn't invent active learning; in fact, active learning has been the subject of advanced academic research for decades.

Active learning is an elegant paradigm designed to prioritize data rows by order of usefulness, usually to reduce the amount of labeling. Instead of "injecting" the entire training dataset into the model at once, AL consists in selecting a smaller subset of data to train on, and reassessing the "state" of the model by analyzing how "confused" the model is on the rows it hasn't seen yet.

The problem is that not everything that the model is confused by is necessarily useful: imagine you're building a sentiment analysis model for english, but some of your data is in german. Your model will be 'confused' by german rows but it won't mean that those are useful. In other terms, model confusion isn't always a good proxy for data usefulness.

So why does this matter? Well, Alectio actually goes one step further and actually analyzes the reason why the model seems confused, and doesn't assume confusion equals usefulness. We try to understand why the model got confused. This allows us to separate bad data from useless (redundant or irrelevant) data.

See, usefulness is a use case-specific concept. It might seem obvious enough, but it's incredibly important. An image of an empty highway might be super important for a lane detection model but wouldn't help much for a pedestrian detection model. In other words, **data shouldn't be curated in a vacuum,** even though we don't necessarily realize it as humans since we naturally make assumptions about how the data is likely going to be used.

At Alectio, we use your model or a proxy model to select data, both in the context of data curation (dynamic discovery of useful data) and data filtering (prediction of usefulness). Our tools help diagnose why data is useful, harmless, or harmful, as well as helping you catalog data for its usefulness in specific use cases.

CONCLUSION

The premise behind our company is deceptively simple: **not all data is created equal**. But if you can act on that simple premise, if you can identify the data that actually helps your model and the data that doesn't, you can train and retrain better models in less time, you can reduce your labeling budgets, you can explain your model's behavior and what it needs to see next. You can rid yourself of dark data that's wasting valuable storage space. You can hone new data collection strategies, optimize your synthetic data generation practices, and keep your models up-to-date without incurring massive compute costs.

It all comes from separating the good data from the bad, from finding the smart data inside your big data. And if you ever have a question or want to try our service, we'd love to get you started. Email us at info@alectio.com and we'll get you up and running in no time flat.

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