A BLUEPRINT FOR

Preparing Your Own ML Training Data

ALEGION™

INTRODUCTION

We are in the business of preparing machine learning training datasets for Fortune 1000 organizations.

Our first meeting with a prospective client is pretty predictable. Most of the time we meet with the data science team responsible for the organization's first serious machine learning project.

They tell us that:

- Their project is potentially strategic and highly visible.
- The board is paying attention.
- Their internal proof of concept – a mix of home-grown and off-the-shelf algorithms and training data – went well.
- Now they are trying to get the model to a level of confidence that will let them to put it into production.
- They're preparing the training dataset themselves, and the volume of data and the complexity of the process has become overwhelming.
- As a result, they've burned through a greater-than-expected percentage of their budget, they're behind schedule, and their algorithm is far from being production-ready.

If this describes your situation this guide is for you. Think of it as a pre-flight checklist for data science teams that are contemplating preparing their own training data.

As we share with you the capabilities that we're confident you'll need, use this checklist to measure your own level of preparedness.

ML Training Data Preparation Checklist

- ☐ Do you know how much training data you need?
- ☐ Have you established a data prep manufacturing process?
- ☐ Do you have the specialized resources you need?

Tools

- ☐ Do you have a task and workflow management platform?
- ☐ Do you have the labelling and annotation tools you need?
- ☐ Do your tools maintain quality assurance at scale?

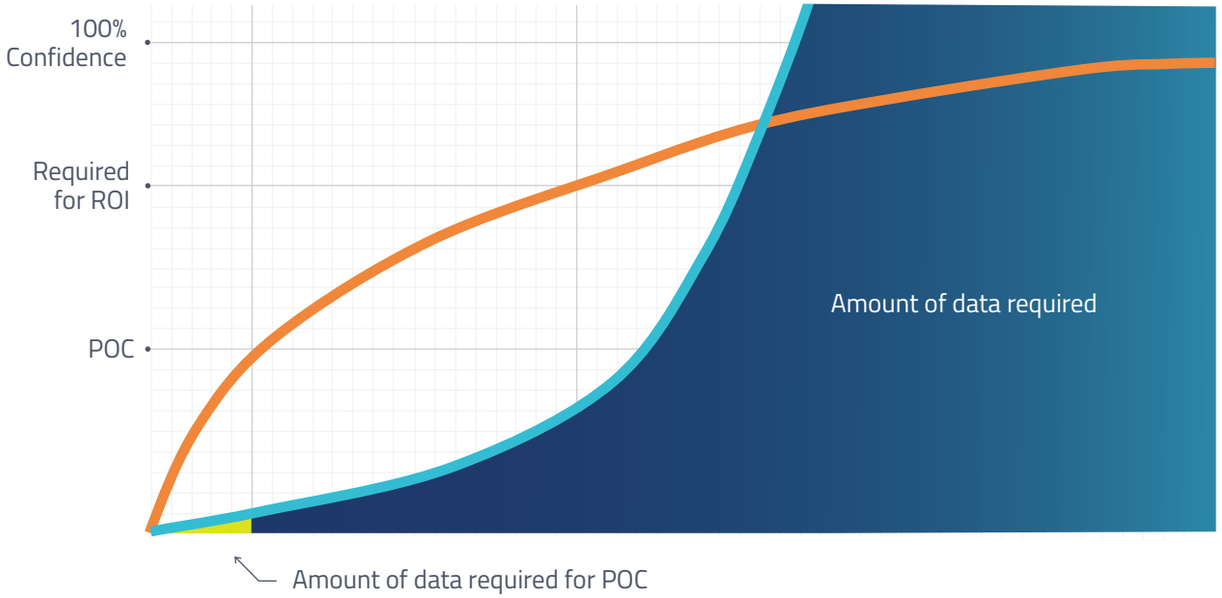
People

- ☐ Do you know how many data specialists you need?
- ☐ Can you find specialists with the right skills and clearances?
- ☐ Who will recruit, educate and manage your workforce?

Skills

- ☐ Does your team have task and workflow design skills?
- ☐ Are your team members skilled in training data pitfalls?
- ☐ Who will orchestrate, accelerate and de-risk your project?

Training Data Requirement vs. Model Confidence



Do you know how much training data you need?

In our experience, data science teams have a hard time convincing the organization of the enormity of the training data challenge.

Initial model training, where the goal is to win internal approval for an ML project, does not require a lot of data. At this stage data scientists often label the required data themselves, or they acquire pre-labeled data.

This kind of internal proof of concept usually addresses only an element or two of a larger business problem. The next step is much harder. Now the team must expose the algorithm to more - often many more - use cases.

The stakes are high. The model can't go into production if it isn't able to navigate the greater complexity and diversity of this second stage. And training on all of this complexity requires much more data than the POC. As a rule of thumb, you can count on each additional use case requiring as much, or more, data than the single use case in the POC.

For example, when clients ask us to prepare the training data required to get to ROI, it is not uncommon for us to label and annotate hundreds of thousands or even millions of data items.

Are you prepared to treat data prep as a manufacturing process?

We often see teams fail to recognize the need to adjust the way they prepare training data as they move from internal POC to actual model training. They maintain some version of their simpler initial approach.

Is there something inherently wrong in an artisanal approach to turning unstructured data into structured data? Of course not. Obviously, there are lots of ways to add structure to data, just as there are lots of ways to assemble an automobile.

Aston Martin makes its cars largely by hand. But keep in mind that it only needs to make 5,000 cars per year. Ford Motor Co. made 6.5 million vehicles worldwide last year. Ford could not have accomplished this with Aston Martin's approach to manufacturing.

To have a better chance of succeeding, a team's dataset preparation should look more like an assembly line, where:

- The factory's raw materials comprise raw, unstructured data..
- The algorithm's requirements dictate what is done to the data as it moves through the manufacturing process.
- The manufacturing tolerances may be exact or loose, depending on the algorithm and its intended application.
- The factory floor is a mix of machines and humans, each employing specialized tools to add value to the final product.
- Quality is measured at many steps in the process, creating feedback for prior contributors and possibly routing the product through additional or remedial manufacturing steps.
- The final product that rolls off the line is structured data, produced to consistent, prescribed tolerances, designed to teach the algorithm about the environment in which it will operate.

Do you have the specialized resources you need?

An industrial approach to training data preparation requires resources that few organizations have at this early stage in the evolution of enterprise machine learning projects:

Training data preparation requires specialized technology.

Software tools for tracking and managing huge numbers of images, video frames, text fragments. Tools for defining data labeling or annotation tasks, and for assigning those tasks to people. Tools for evaluating human decisions and determining the need for further review or adjudication.

Training data preparation involves people.

Human judgement is essential to structuring ML training data. Large volumes of data call for many people, some or all of whom may need particular skills or qualifications.

Training data preparation demands specialized project skills.

Skills needed to root out model-wrecking bias in the data. Skills required to design tasks that promote efficiency, scalability and quality. Skills needed to oversee iterative and very fluid processes.

Tools

Do you have a task and workflow management platform?

ML training data volumes are far too large – and data labeling at scale is far too complex – to be managed in a spreadsheet or a generic database. If training data preparation resembles an assembly line, arguably what is required is the equivalent of an MRP system.

You will be at a significant advantage with a purpose-built platform for training data preparation, technology that lets project managers:

- keep a record of every data item that needs to be labeled
- track every data item through what may be numerous operations
- design and execute on multi-step and conditional data labeling tasks
- determine which workers can do a particular task
- follow worker performance
- optimize workflows to continually improve task efficiency
- enforce quality control and review escalation workflows

Do you have the labeling and annotation tools you need?

Each of the many ML use cases requires training data with particular labeling and annotation. And regardless of use case, there is always a demand for accuracy and efficiency in preparing ML datasets.

Computer vision use cases need tools for expediting annotations, pixel-level segmentation, and classifying objects into complex taxonomies.

NLP use cases rely on training data preparation tools that identify patterns, context, and intent.

And entity resolution use cases demand training data that has been labeled to accommodate pattern recognition, matching and clustering.

If your project involves	You may need tools that support
Computer vision	Keypoint Polygon Bounding Box Object Detection and classification Parts ID and Landmark Detection Instance and Semantic Segmentation Actions and Interaction identification
NLP	Complex query classification Pattern Recognition Sentiment Analysis Information extraction Intent Recognition Semantic Enrichment Identity Recognition
Entity Resolution	Record Linkage Ontology Resolution

Tools:

Do your tools automate quality assurance?

Quality is of paramount importance with regard to training data. And given the quantity of training data required for typical CV and NLP projects, the cost of attaining quality cannot be high.

ML project teams have numerous quality assurance techniques and options at their disposal. What they don't typically have is unlimited time, nor unlimited budget to pursue non-productive options.

Once training data preparation begins in earnest, teams need tools that rely on predictive algorithms to score human or machine judgements against a particular task. These tools need to dynamically determine if additional quality control parameters like judgment consensus, gold standard data, administrative reviews, or exception handling are needed.

People

Do you know how many data specialists you need?

It takes a lot of data, properly labeled, to train an algorithm. The most common mistake companies make when getting to this phase of their project is to underestimate the size of this challenge and the sheer number of people they will need to get this work done.

People supply human judgement to the data preparation process. These data specialists drive the tools, label the data, and even evaluate the work of other people.

These data specialists can come from your own private crowd, if your organization has one. You can engage a public crowd, like Amazon's Mechanical Turk. Or, you may need to assemble a project-specific crowd if you require particular domain expertise or security requirements.

Can you find specialists with the right skills and clearances?

Depending on the nature of the project, you may need people who do no more than make simple observations. On the other hand, you may require people with specific training or skills.

Regardless of their skills, the data specialists you engage will be handling and working with what may be confidential or sensitive corporate data. It may be necessary to conduct background checks on the data specialists you employ.

If the data are protected by government regulation, as in the case of healthcare data, you may need to employ data specialists with particular clearances or attestations.

And in extreme cases – as when the data are subject to federal or defense security constraints – you may need to ensure that the data specialists work in virtual or physical private crowds, or even in secure, badged facilities.

69% of enterprises are facing a "moderate, major or extreme" skills gap in finding skilled associates to staff their new AI-driven business models and projects

Who is going to recruit, educate and manage your workforce?

It's not by accident that the human judgement behind ML training data is referred to as a crowd. Depending on the volume of data the algorithm requires, and the number and complexity of the tasks needed to structure the data appropriately, an ML project team may need to find, train and then manage hundreds or thousands of people.

The team members need to take stock of their ability and willingness to be responsible for a crowd. In our experience, the team has neither the time nor the energy for this additional responsibility that ultimately has nothing directly to do with machine learning.

Skills

Does your team have task and workflow design skills?

There are many steps between understanding what an algorithm needs for training data, and actually labeling and annotating unstructured data in ways that deliver a model that produces ROI. Understanding these steps is key. A tool can do lots of things. How a tool should be used in a given context is the question.

Training data preparation requires specialized project management skills that must remain effective at great scale. The process of producing training data often relies on multi-step conditional workflows that involve large numbers of people and extraordinary volumes of data. It demands continual improvement of task efficiency, and deployment of tasks to trained data specialists based on skill and security requirements. And training data has to be of the highest quality, meaning that escalation workflow design and data specialist evaluation are also key skills.

Task and Workflow Design Skills

- Requirements consultation
- Custom task design
- Workflow configuration
- Task distribution
- Specialized workforce curation and training
- Human intelligence integration
- Judgement consensus
- Adjudication strategy

Skills:

Are your team members skilled in training data pitfalls?

With training data tasks and processes set up, there are still pitfalls to dataset preparation that experience and specific skills are required to detect and mitigate.

Bias in the training data can jeopardize model confidence. There are three distinct types of data bias you need to watch out for – sample bias, prejudicial bias, and measurement bias – that can distort or degrade algorithm performance. Not all data scientists have the training to detect or prevent this kind of bias in their training data.

Data specialist accuracy is another area that is fraught with risk. Some of your specialists will get things wrong despite their best efforts. Other workers do not come to the project with best intentions and may attempt to game the system.

There are a numerous approaches to detecting and mitigating these issues – exposing workers to gold data, for example – but this is another part of training data preparation that data scientists may not have fluency with.

Can your team orchestrate an enterprise ML project with an eye to both quality and scale?

Making tools and people work together to prepare extremely large volumes of high quality ML training data requires specialized skills. Managing all of this in a context of high expectations and scrutiny is even harder.

For first-time ML project teams, the pressure to accelerate the project's progress, while taking risk out of the endeavor, will be relentless.

Skills:

How did you do?

Are you ready to do this yourself? Did you check all the boxes? In our experience few organizations do. Because, even though they are fully committed to one or more machine learning projects, some of the capabilities these projects demand are not, and shouldn't be, core competencies of the organization.

Training data preparation is a good example. You need it. You will not put an ML model in production without it. But that doesn't mean you have to prepare it yourself.

We can do it for you. You can have all of this done for you with enterprise-experienced project managers to run your project. You can point us to your data, explain how it needs to be labeled for your algorithm, and get back high-quality, accurately prepared training data - without having to lift a finger.

And since there are plenty of people, with the right clearance, the right skills, working in a comprehensive suite of training data technologies - the Alegion Training Data Platform - you can use this data with confidence.

ML Training Data Preparation Checklist

- ☐ Do you know how much training data you need?
- ☐ Have you established a data prep manufacturing process?
- ☐ Do you have the specialized resources you need?

Tools

- ☐ Do you have a task and workflow management platform?
- ☐ Do you have the labelling and annotation tools you need?
- ☐ Do your tools maintain quality assurance at scale?

People

- ☐ Do you know how many data specialists you need?
- ☐ Can you find specialists with the right skills and clearances?
- ☐ Who will recruit, educate and manage your workforce?

Skills

- ☐ Does your team have task and workflow design skills?
- ☐ Are your team members skilled in training data pitfalls?
- ☐ Who will orchestrate, accelerate and de-risk your project?

About Alegion

We are the Fortune 500's machine learning training data partner. You know our clients. You know their logos. They are the world's most influential companies, and they are making huge, strategic bets on machine learning. They come to us because:

The scope of training data prep is overwhelming. They can't or don't want to shoulder the task internally. But they need it done their way, with meticulous accuracy, and quickly.

We offload the entire process, relying on a mix of our own AI-enhanced platform technology and our nearly million-member global pool of on-demand data specialists.

Our Customer Success team removes 100% of the burden of training data from clients' shoulders. They define the labeling tasks, curate the crowd, keep it all on schedule and ensure high accuracy across many thousands or millions of data items.

///ALEGION™

Want to learn more?

Reach out to Adam Elliott at aelliot@alegion.com