What data scientists tell us about AI model training today

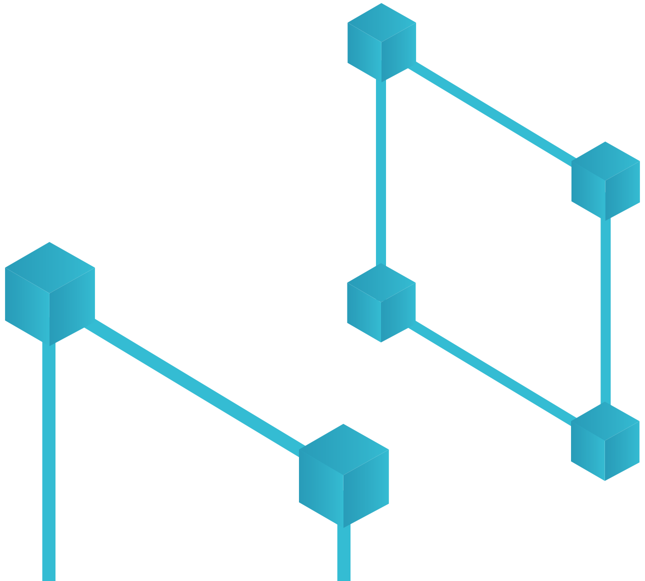
ALEGION™

INTRODUCTION

While no longer totally novel, AI is still nascent.

METHODOLOGY

Dimensional Research conducted a survey of 277 data scientists and other AI professionals in large companies across nearly 20 industries to determine their use and development of AI and ML projects. Their findings suggest that for many of their organizations it's still early days for the technology.



RESEARCH GOAL

The primary research goal was to gauge the maturity of machine learning (ML) in the enterprise, to understand today's ML project challenges, as well as the tools and resources used in these projects.

FINDINGS

The survey confirmed the existence of broad challenges that we see in our customer organizations. Labeling and annotating training data for machine learning projects is a serious problem for data science teams, and a significant obstacle to getting those projects into production.

4 out of 5 respondents admit that training AI with data is more difficult than expected.

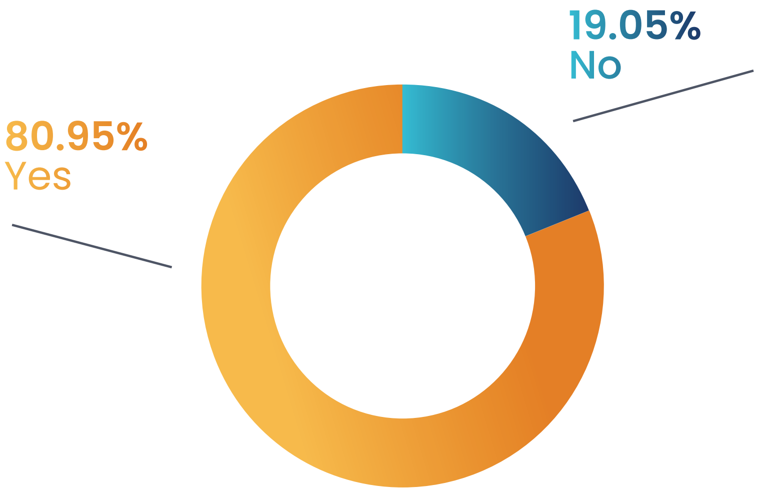
Nearly two thirds of data scientists we surveyed indicated that their ML project has progressed beyond POC and algorithm development to the training data phase. This is already a major win.

But for most, this next phase of feeding the algorithm enough training data to be ready for validation in the real world presents a host of challenges. In fact, 80% report that training their algorithm has proved more challenging that they expected.

Reasons cited include:

- **Bias or errors in the data**
- **Not enough data**
- **Data not in a usable form**
- **Don’t have the people to label data**
- **Don’t have the tools to label the data**

Has training the AI with data been more challenging than expected?



96% encountered data quality and labeling challenges

BIAS OR ERRORS IN THE DATA

AI systems do precisely what they are taught to do. They are only as good as their underlying mathematics and the data on which they are trained. When things go wrong with AIs for one of two reasons: either the model of the world at the heart of the AI is flawed, or the algorithm driving the model has been insufficiently or incorrectly trained. Bias in one form or another is behind many algorithm and data issues. If not mitigated, bias will cause the model to behave - or misbehave - in ways that reflect the bias. For more check out our white paper 4 Types Of ML Bias.

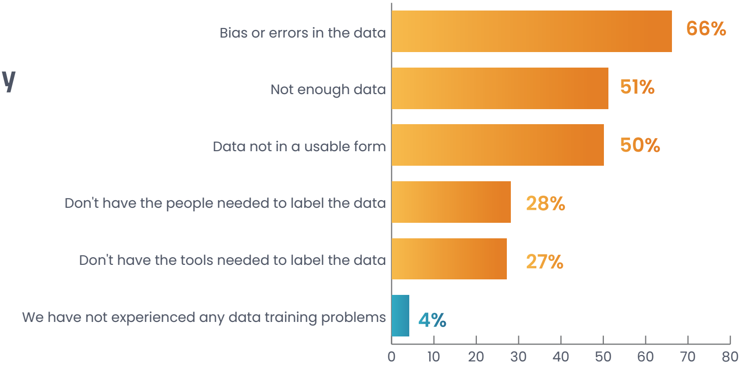
NOT ENOUGH DATA OR DATA NOT IN A USABLE FORM

Aggregating data entails its own unique obstacles, including differing formats, incompatible databases, and security concerns (the data might be classified, for example, so another team in your own company won't let you have it). Then the data has to be properly labeled and classified in a taxonomy. Without properly labeled data, training can't even begin.

DON'T HAVE THE PEOPLE AND/OR THE TOOLS TO LABEL DATA

These projects don't have the people or the tools because they are trying to do it all themselves. Data labeling is not a manual process and it is impossible to properly scale without technological assistance. Designing a platform on which to manage people and design tasks, as well as build the tools for them to use requires a whole slew of projects in and of itself.

Which problems have your company experienced with AI training data specifically?



The DIY approach - doesn’t leave much time for the task at hand

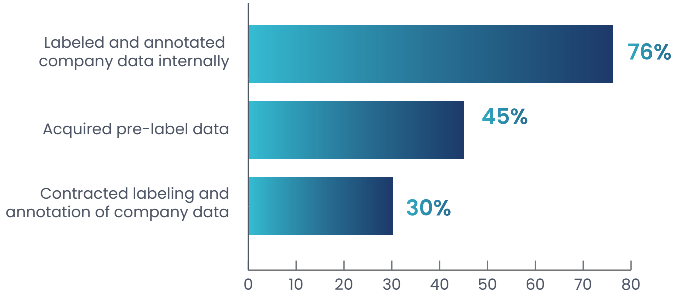
80% of data scientist’s time is currently spent preparing and managing data, which is not surprising since 76% of companies attempt at least some data labeling in house. This is not only problematic for companies because data scientists are expensive and in high demand, but also dissatisfying for data scientists who take jobs to do interesting, challenging, and strategic work, not to draw boxes. We also see here, however, that companies are also dabbling with outsourcing some or all of their data prep.

On top of doing their own labeling, 63% of enterprises are spending resources building their own labeling and annotation tools.

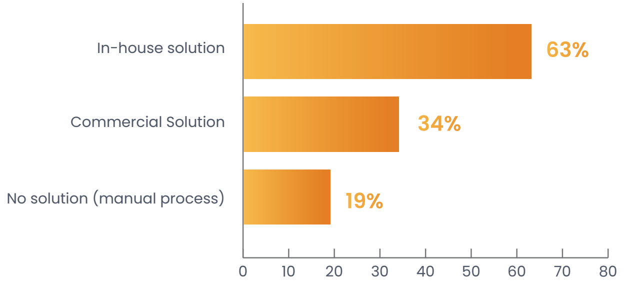
- **Computer Vision** makes use of keypoint, polygon, bounding box, object detection and classification, parts ID and landmark detection, instance and semantic segmentation, as well as actions and interaction identification.
- **NLP** requires complex query classification, pattern recognition, sentiment analysis, information extraction, intent recognition, semantic enrichment, and identity recognition.
- **Entity Resolution** demands record linkage and ontology resolution.

Between building tools and labeling data data scientists are not left with much time to do what they were presumably hired to do - to use machine learning to improve the business and, potentially, to carve out a position of industry dominance through innovation.

What approach is being used to train the initial algorithm?



What solution was used in your most recent project to label and annotate your training data?



78% of AI or ML projects stall at some point before deployment

A third of AI/ML projects stall right off the bat with their **proof of concept (POC)**. POCs are a litmus test for an idea. Can you demonstrate that this problem can be solved with tech? For example, “I think I could teach a computer to identify strawberries and how ripe they are, so I know when to pick them.” Ideally, necessary resources are allocated to the POC to see if it flies or fails as fast as possible.

Pitfalls of DIY approach to the AI training lifecycle

1. TRAINING DATA LABELING (PREPARATION)

To build model confidence, a corpus of training data is fed to the model. Scaling this effort, however, is often overwhelming due to deficiency of tools, lack of people, & insufficient knowledge of how to manage a data labeling project. For more check out our [Blueprint for preparing your own ML training data.]

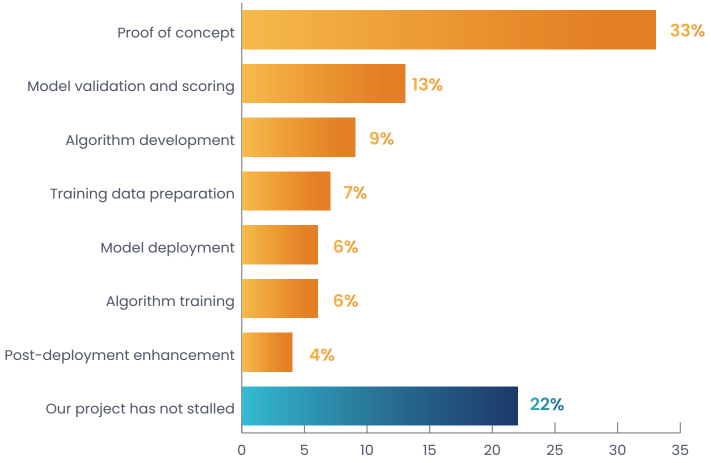
2. MODEL VALIDATION & SCORING

Once the model has achieved a high level of confidence with training data, it is then exposed to the universe for validation, ie does it work in the real world? At this point models are introduced to edge cases, new, unique, or extreme examples outside of the scope of model recognition. This presents another training opportunity. An augmented reality (AR) tool can help score the algorithm and continue to train it as it is exposed to a bigger universe.

3. POST-DEPLOYMENT ENHANCEMENT

Finally your model is ready for production - Even then it will still be exposed to nuance and new aspects of the universe, to which it has never been exposed and it doesn’t know how to handle. To remain competitive the model needs to continually improve confidence and accuracy, which means it will need geometrically more training data throughout the AI lifecycle.

In which phase did your AI or ML project stall?



Outsourcing - 71% of companies have outsourced some AI/ML activities

To combat the pitfalls discussed above, 70% of companies have opted to outsource at least some of their AI/ML activities including:

DATA COLLECTION

Companies with ML projects that don't have the correct data or enough of it, often outsource the collection or generation of data they need to train their algorithm.

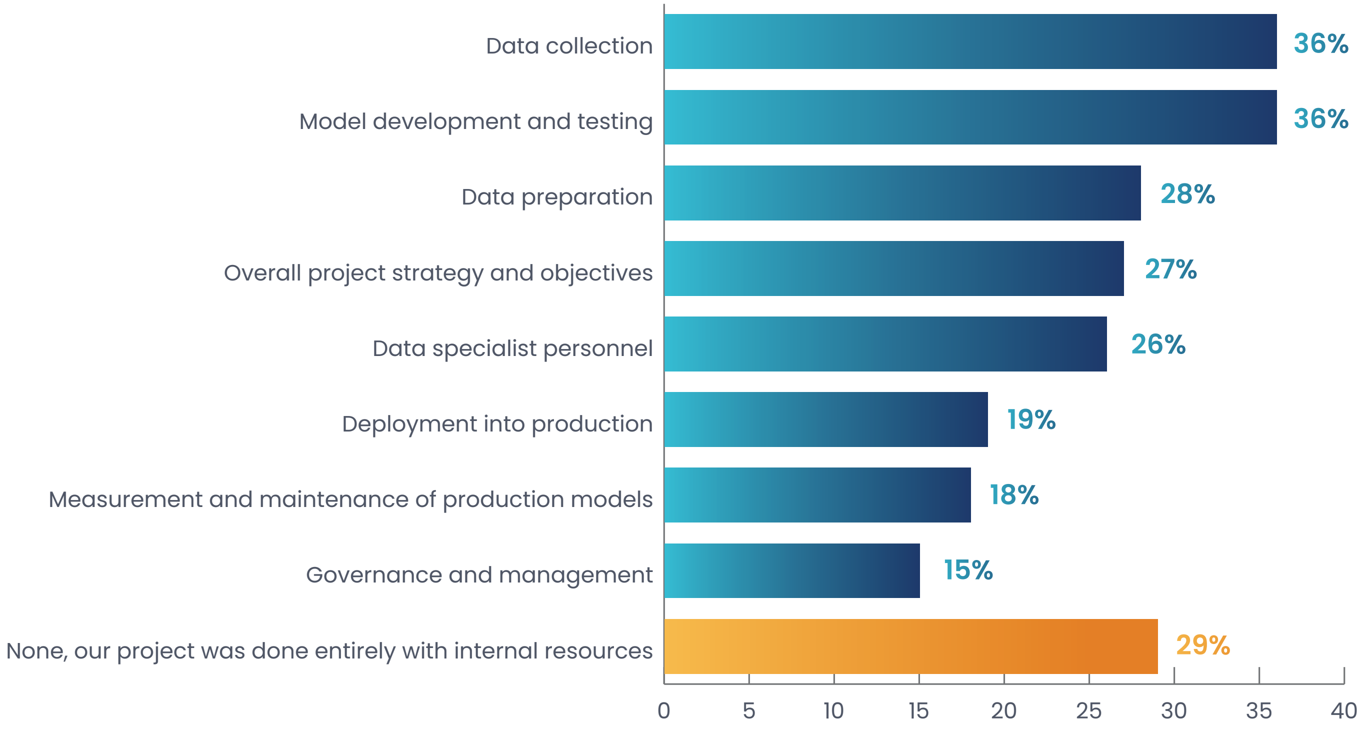
DATA SPECIALIST PERSONNEL

It takes a lot of properly labeled data to train an algorithm and people are needed to supply human judgment to the data preparation process. These data specialists drive the tools, label the data, and evaluate the work of other people. Depending on the nature of the project, people may do no more than make simple observations, but for others specific training and skills are required. Depending on the volume of data the algorithm requires as well as the number and complexity of the tasks needed to structure the data appropriately, an ML project team may need to find, train, and manage hundreds of people.

DATA LABELING (PREPARATION)

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. Teams need a task and workflow management platform to track every data item, design and execute on multi-step data labeling tasks, optimize workflows to continually improve task efficiency, and to enforce quality control. Once training data preparation begins in earnest, teams need tools that rely on predictive algorithms to score human or machine judgments 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.

Which of the following external services has your company used for its AI and ML projects?



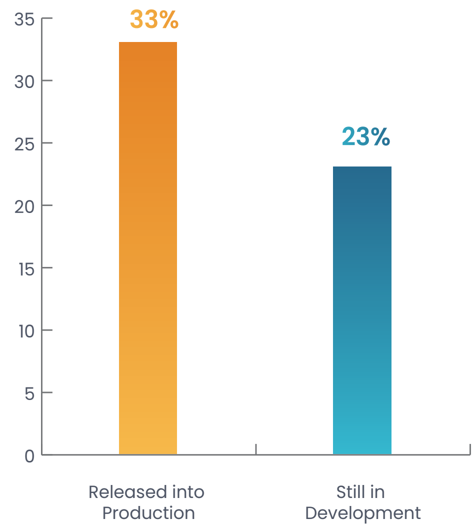
Companies that Outsource Data Labeling Are More Likely to Be in Production

Most of the data scientists in this survey indicate that they have attempted to label at least some of their training data in-house. The list of issues this approach engenders is documented, as is the frequency with which their projects stall.

The data also shows that data science teams are offloading a variety of ML development tasks, including data labeling and annotation. And there's evidence to suggest that offloading at least some tasks is beneficial.

In their answers, the data scientists surveyed show that offloading training data labeling and annotation is associated with a significantly higher rate of successful project deployment. This is not surprising, given the typical volume of training data involved, the relatively small team sizes and the numerous data quality issues the respondents described.

Outsourcing of Labeling and Annotation of Company Data



CONCLUSION

Data scientists report that more often than not their projects stall around issues of training data. They are clear that training data labeling and annotation presents numerous challenges. This is an activity that a sizable percentage of the survey audience has outsourced, with measurable effects on project outcomes. As you review your last ML project or contemplate your next, it's worth thinking about how best to avoid the data labeling obstacles that are so obviously common to machine learning.

ABOUT ALEGION

Alegion provides ground truth training data for machine learning initiatives. Our offering operates at massive scale, combining a data and task management software platform with a global pool of trained data specialists.

We assist data science teams throughout the AI life cycle, delivering custom training datasets, providing human-scored model testing, and making available human-in-the-loop exception handling. With our white glove level of service we completely offload these activities, freeing data professionals to focus on their areas of specialization.

We support machine learning projects broadly, with particular emphasis on Computer Vision, Natural Language Processing and Entity Resolution, in retail, financial services, defense, technology and manufacturing.

ALEGION™

Want to learn more?

Reach out to Adam Elliott at aelliot@alegion.com