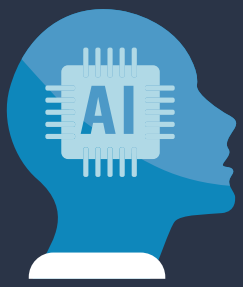
cloudfactory



SCALING QUALITY TRAINING DATA

OPTIMIZE YOUR WORKFORCE AND AVOID THE COST OF THE CROWD

Scaling Quality Training Data

If you need people to process some portion of the big data that feeds your artificial intelligence, you need a reliable workforce. You're not alone: more businesses are using in-house staff, contractors, and crowdsourcing to get this kind of work done, and industry analysts expect that trend to increase significantly over the next two years.

IN THIS PAPER, YOU WILL LEARN

How to determine what kind of AI data work to outsource

Why anonymous crowdsourcing adds cost to your data operations

Best practices to accelerate and scale high-quality data for AI

CHOOSING PEOPLE FOR YOUR AI TECH STACK

Bringing artificial intelligence (AI) to life in the real world is a lot like the 20th-century "space race" for dominance in spaceflight capability. Few can fathom the level of innovation and hard work it takes. From model development and data prep to testing and deployment, AI requires a pioneering spirit, sharp minds, and hard work. AI innovators encounter countless challenges and frustrating defeats.

One of those challenges is access to talent that is in short supply. More than half (54%) of leaders report skill shortage as the biggest challenge facing their organizations, **according to Gartner**. Another is dirty data, which data scientists say is their number-one problem, according to a **Kaggle survey**. If you want to strategically deploy your team, you probably don't want your prized data scientists doing the tedious, time-consuming work of data labeling or annotation.

But they're likely mired in it. A massive amount of data must be gathered, structured, and quality-checked in the process of machine learning (ML). For example, to develop computer vision for a self-driving car, you'll need people in the loop to annotate, or label, countless frames of driving video to teach the algorithm to "see" objects such as people, signs, trees, and vehicles. For each hour of video, there's a staggering **800 hours** of annotation work to do.

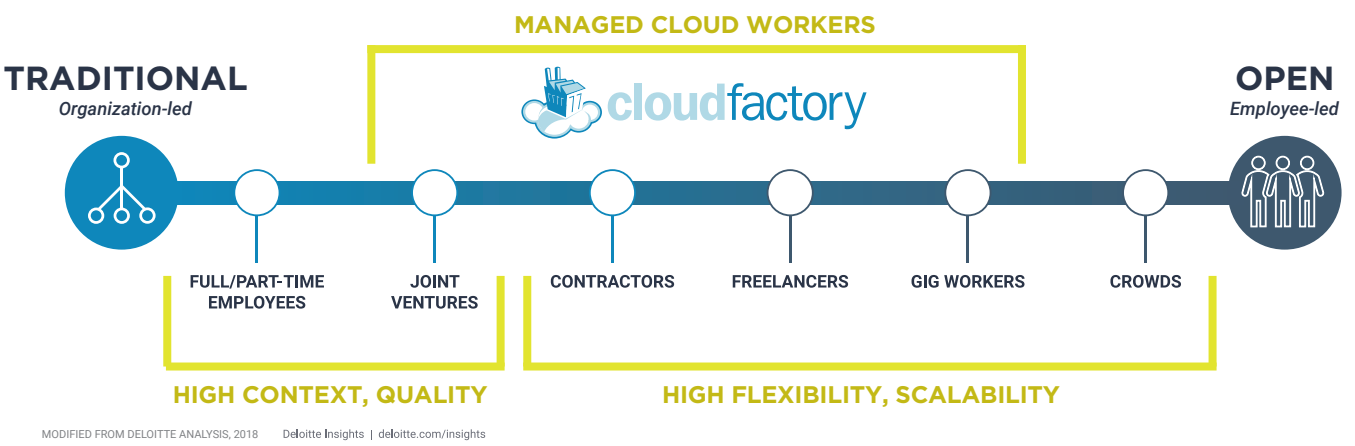
To process the big data that feeds your artificial intelligence, you need a reliable workforce with relevant domain expertise and high standards for quality. A growing number of innovators are using in-house staff, freelancers, contractors, and gig workers to get this massive amount of data work done, and as ML models proliferate, **Deloitte predicts** that trend will increase significantly over the next few years.

Not All Contractors Have Context

New workforce models are emerging. Last year, the U.S. National Aeronautics and Space Administration (NASA) researched disruptors driving the future of work so it could evolve its talent strategies. The result was its **Future of Work Framework**, which encourages leaders to design for agility and focus on impact because "work today requires fluid talent to meet ever increasingly complex work, requiring multidisciplinary skills, delivered by teams of people, networked together that have overarching goals tied to organizational performance and productivity."

In **Deloitte's 2018 Global Human Capital Trends report,** half of respondents said they have a large number of contractors in their workforces. Deloitte maps the workforce ecosystem from traditional, full-time workers who have strong organizational context to open, crowd workforces who have less understanding of the organization's overall strategy.

And that's the challenge with open sourcing for AI development: domain expertise and context. Workers need more than the ability to tag, label, or annotate your data. They must understand the needs of your end user, the rules for your data, and context for the tasks they are doing if you want them to return quality data to train your ML algorithms. And quality is king, because when the people who tag, label, or annotate your data provide low-quality work, your model struggles to learn. So while each task may be simple, how it fits into the larger picture of your end user's experience isn't as easy to teach someone quickly. **That's difficult to scale.**



The right workforce gives you the flexibility to respond to changes in market conditions, product development, and business requirements. On the left side, you'll shoulder the burden of management with an in-house team. On the right side, quality work is likely to be a hurdle with crowds.

Cleaning and Structuring Data for AI

1 **In-house employees** can manage your data needs with reasonably good quality, and this approach works fine until it's time to scale your model. Over time, these processes will grow more difficult and costly to manage, so you're likely to join the growing list of companies that are turning to contractors, freelancers, and gig workers to structure data for AI development.

2 **Contractors and freelancers** are another option but be sure to factor in the time it will take you to source and manage your team. One-third of Deloitte's survey respondents said their human resources departments are not involved in sourcing (39%) or hiring (35%) decisions for contract employees, which "suggests that these workers are not subject to the cultural, skills, and other forms of assessments used for full-time employees." That can be a problem when it comes to quality work, so allocate additional time for sourcing, training, and management.

3 **Crowdsourcing** leverages the cloud to send data tasks to a large number of people at once. Quality is established using consensus, which means several people complete the same task, and the answer provided by the majority of the workers is chosen as the correct one. Crowd workers are paid based on the number of tasks they complete on the platform provided by the workforce vendor, so you could spend, on average, double the time processing data with a crowd than you would with an in-house team. The burden is on you to manage workers' data outputs at scale.

4 **Managed cloud workers** have emerged as an option over the last decade. This approach combines the quality of a trained, in-house team with the scalability of the crowd. It's ideal for data work because dedicated teams are steeped in your business rules and they stick with projects long-term, enabling them to increase their throughput and accuracy while providing consistent labeling quality. This model also provides a team that is in direct communication with you, enabling agile process iterations necessary for sustainably creating high-quality datasets. To learn more, start by reviewing these **five steps to sourcing great-fit cloud labor.**

TAKEAWAYS

- It takes up to 800 hours to annotate one hour of video.
- To annotate your data with quality, workers must understand context.
- Your workforce choice might determine your AI success.

THE HIDDEN COSTS OF THE CROWD

NASA estimated that it took **400,000 engineers, scientists, and technicians** to send astronauts to the moon on the Apollo missions. The massive workforce was comprised of people from four major enterprise companies and a host of subcontractors who worked for them.

Like sending astronauts to the moon, building AI requires access to a large number of people to gather, process, and structure training data. Speed to market is a priority, as AI development teams contend with the challenges of innovating fast in an increasingly competitive marketplace. As more companies seek fast access to talent that is in short supply, crowdsourcing has emerged as an alternative to an in-house team.

Crowdsourcing uses the cloud to send data tasks to a large number of anonymous people, who are paid based on the number of tasks they complete. While it offers a cheap option for training ML algorithms, it's rarely as inexpensive as it seems.

Measuring Quality

We've explored the importance of quality. The better your data, the better your model will perform. And when the people who tag, label, or annotate your data provide low-quality work, your model struggles to learn.

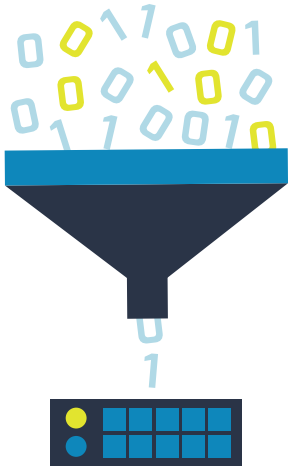
There are three methods we use in the workforce industry to measure work quality. At CloudFactory, we use one or more of these methods to measure the quality of our own vetted, managed workforce on every project.

3 METHODS FOR MEASURING WORK QUALITY

1. CONSENSUS
We assign several people do the same task, and the correct answer is the one that comes back from the majority of workers. This is the crowdsourcing model.

2. GOLD STANDARD
There is a correct answer for the task, and we measure quality based on correct and incorrect tasks.

3. SAMPLE REVIEW
We select a random sample of completed tasks, and a more experienced worker, such as a team lead or project manager, reviews the sample for accuracy.



Our client success team found consensus models cost at least 200% more per task than processes where quality standards can be met from the first pass

3 Hidden Costs of the Crowd

Over a decade of processing critical business data for companies around the globe, we've learned that applying the crowdsourcing model to data processing for AI applications can get you access to a large number of workers. But it can create other issues that affect your speed to market. Here are some of the hidden costs of the crowd.

1. POOR DATA QUALITY
Anonymity is a bug, not a feature, when it comes to crowdsourcing. Workers have little accountability for poor results. When task answers aren't straightforward and objective, crowdsourcing requires double-entry and consensus models to be used as control measures. If you're unsatisfied with the work, often you must send the work through again, hoping for a different result, placing more of the quality assurance (QA) burden on your team. Each time a task is sent to the crowd, costs rack up.

"Re-working poorly labeled data is very expensive," **said Brian Rieger**, COO of **Labelbox**, a California-based company that provides tools for labeling and managing training data.

At CloudFactory, we have a microtasking platform that can distribute a single task to multiple workers, using the consensus model to measure quality. Our client success team found consensus models cost at least 200% more per task than processes where quality standards can be met from the first pass. Managed teams are better suited to tasks requiring high quality because they can handle more nuanced tasks and get them right the first time.

2. LACK OF AGILITY

In AI development, tasks can change as you train your models, so your workforce must be able to adapt. That requires a tight communication and feedback loop with your workforce. And if there isn't continuity in the workforce, it's more difficult to acquire learned domain expertise and context that make it possible to adapt to changes in the workflow quickly. As a result, your process will be inefficient and your models will struggle to learn.

"Labelers get better at annotation tasks over time, as they get familiar with the source imagery and the nuances of the interpretation desired. Labelers who are better at labeling lead to better training data, and that leads to better model performance," said Rieger.

Crowdsourcing limits that agility to modify and evolve your process, creating a barrier to worker specialization, or the proficiency with your data and process that grows over time. Workers are ever-changing, few overcome the learning curve, and you are likely to see less continuity in your data. Any changes in your process can create bottlenecks.

Data workers on a managed team can increase their domain expertise - or understanding of your rules and edge cases - over time, so they can make informed subjective decisions that are more accurate and result in higher quality data.

3. MANAGEMENT BURDEN

When you crowdsource your data production, you can expect worker churn. As new workers join the crowd, you'll have to rely on the business rules you created and task new workers with training themselves on how to do the work. If your team is bringing each new worker up to speed, be sure to allocate time for that management responsibility.

With some crowdsourcing options, you are responsible for posting your projects, reviewing and selecting candidate submissions, and managing worker relationships. You'll need to factor in your costs to attract, train, and manage a disconnected group of workers.

If you're considering a crowd model, look into who owns your data as part of your agreement. In addition to platform and transaction fees, some crowdsourcing vendors stake ownership on the data that passes through their platforms, which means they're allowed to use your data to train their own algorithms or serve their own customers.

While it can be difficult to determine the end-to-end cost of a crowdsourced project, you can plan for the crowd to cost more per task as you send low-quality data back to the crowd for reprocessing. Watch for hidden fees in technology, onboarding, and training.

TAKEAWAYS

- Consensus models cost 200% more per task than processes where quality standards can be met from the first pass.
- Managed teams increase their context and domain expertise over time for more accurate decisions, resulting in higher quality datasets.

DESIGNING DATA OPERATIONS FOR QUALITY

"Houston, we've had a problem." Astronaut Jack Swigert **made the words famous** when he communicated to NASA mission control that an explosion had rocked the Apollo 13 capsule that was transporting him and two other people to the moon in April 1970. To get the astronauts home safely, the engineers at Johnson Space Center in Houston, Texas would have to do something they had never attempted before: use the descent engines on the lunar lander to send it home.

The algorithms for calculating the maneuver had been written only months earlier. The two young programmers who had written them sprung into action to check every possible parameter to see if the maneuver would work. Thanks to their hard work - and that of hundreds of people in the loop on the data, the algorithms, the required calculations, and other critical factors - they returned all three astronauts safely to Earth.

Machines + People in the Loop

In AI development, similar urgent challenges abound. Teams of computer vision engineers are training the algorithms that self-driving cars use to recognize pedestrians, trees, street signs, and other vehicles. **Researchers** are using data and natural language processing (NLP) to detect psychiatric patients who are at a higher risk for suicide. The success of these systems depends on massive pipelines of data and the skilled people in the loop who structure the data for AI use.

A growing number of teams are using in-house staff and contractors to do this mission-critical work. We've explored the hidden costs of using anonymous crowdsourcing to process data and structure it for AI use. Now, we'll take a closer look at how you can design your training data operations to support quality, speed, and scale.

Your Data Production Line

In many ways, your training data operations are a lot like the assembly lines of yesterday's industrial factories: data is your raw material, and you have to get it through multiple processing and review steps to structure it for ML. Like the Apollo astronauts, you need skilled people on the ground - or, in the loop on your data - who can help you make changes when you run into a problem or your process evolves.

If you want to develop a high-performing ML model, you need smart people, tools, and operations that can consistently deliver high accuracy. Here are four critical elements to consider when you design your data production line for quality, speed, and scale.

1. APPLY TECHNOLOGY

Think of your data production line as your tech-and-human stack, combining people and machines in a workflow that directs tasks where they are best suited for high performance. That means assigning to people the tasks that require domain expertise, context, and adaptability - and giving machines the tasks that require repetition, measurement, and consistency.

Technology is important for communication with your workforce too. Direct contact with your team will give you visibility into the quality of work. It also will allow workers to share insights that will help you make adjustments as your business requirements evolve.

2. USE A TRAINED, MANAGED WORKFORCE

Managed teams deliver higher skill levels, engagement, accountability, and accuracy. Unlike an anonymous crowdsourced team, managed teams can improve in their quality and expertise over time as they grow more familiar with the source data and the nuances of the interpretation for your model.

That means they will get better at making decisions about your data, based on their experience with your domain, context, and edge cases.

It's critical here to have a tight feedback loop with your workers via direct communication with a single point of contact on the ground. This person should be an expert in your data and business rules who can provide feedback, speed change requests, and train new team members.

3. MEASURE QUALITY

The quality of your data will determine the performance of your model. There are three methods we use in the workforce industry to measure quality: consensus, gold standard, and sample review. We use one or more of these methods to check the quality of our teams' work, and quality is a top driver for many of our clients.

Labelbox, a company that provides tools for labeling and managing training data, **distinguishes accuracy from quality**. "Accuracy measures how close a label is to the ground truth, or a subset of the training data labeled by your expert. Consistency is the degree to which labeler annotations agree with one another," said Brian Rieger, COO at Labelbox.

As you build your data production line, look for a workforce provider that is transparent with quality metrics. Also consider how important quality is for your tasks today and how that could evolve over time.

4. DESIGN FOR AGILITY

The keys here are training and technology to scale your data work seamlessly. This is about more than getting workers to label, annotate, or categorize data faster. It's about designing your production line for use-case progression as your AI model develops. As you move through the development process, you'll want flexibility to add higher-level features that can advance your AI application.

For one CloudFactory client, our teams label images to train algorithms that identify counterfeit retail products. The combination of the labeling tool, our technology platform, and our managed-team approach made it possible to iterate the process, resulting in better team morale, higher productivity, and 99.3% accuracy.

Workforce solutions that charge by the hour, rather than by the task, are designed to support iteration in the work. Paying by task can incentivize workers to complete tasks quickly, without high quality. Look for options that get more cost effective as you scale and add more work.

TAKEAWAYS

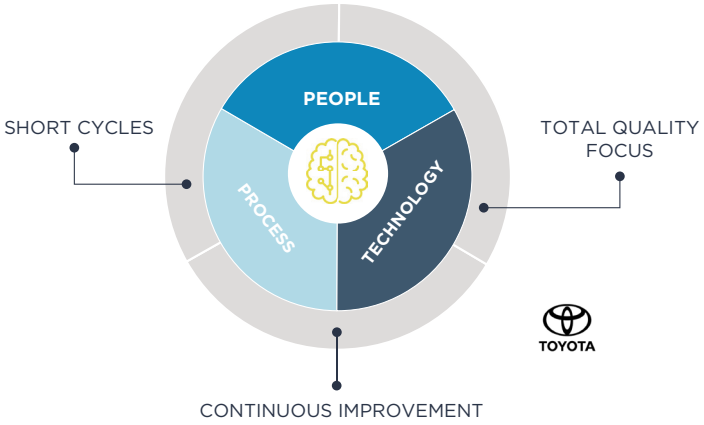
- Design your data production line to deploy tasks strategically to people and machines.
- Make sure your data team can provide feedback on your process, speed change requests, and train new team members.
- Ask about quality metrics and see if options become more cost effective as you scale and add work.

ANATOMY OF A SUCCESSFUL DATA PRODUCTION LINE

The primary objective of your data production line is to make the complex simple.

If we look at the production system developed by Toyota Motor Corporation, it offers a great example for us to talk about AI data production. Toyota's lean process goes beyond the assembly line by organizing manufacturing and logistics for optimal production - including its own interactions with partners and suppliers. The model makes the complex simple by using short production cycles to generate a hyper-focus on quality and continuous improvement measures.

Toyota's system goes even further - reducing the number of difficult jobs on the assembly line by ranking each into three categories: green, yellow, and red. The goal is to improve each job to the easier green level, essentially eliminating difficult jobs altogether.



Designing Your Data Production Line

Start by mapping your process and targeting waste

1. Break the work into steps and fix bottlenecks
2. Eliminate defects
3. Reduce costs
4. Introduce flexibility
5. Partner with trusted suppliers

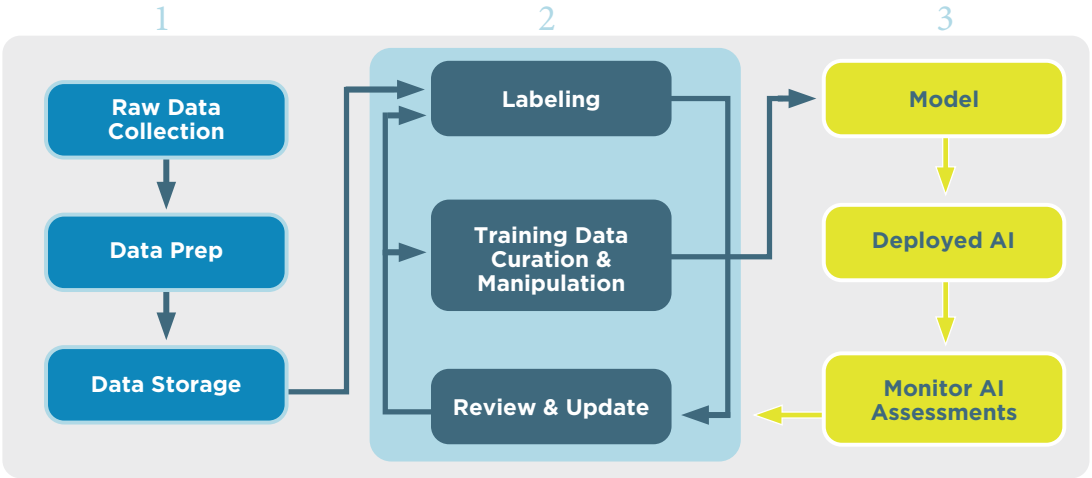
1 Your AI data production starts with raw data collection, data preparation, and data storage. This is the first stage of ML, digitizing the world around you so you can apply human expertise to structure this raw data, then model it so machines can make mission-critical decisions and predictions for your business.

2 The middle stage is labeling, reviewing, and updating. Related QA and QC processes ensure that you're accurately and consistently interpreting digitized reality or ground truth so you can produce training data that powers models applicable for your domain, whether you're building self-driving cars or performing sports analytics. This stage includes curating and manipulating training data before it goes into a model training environment.

3 The third stage - or modeling - is at the heart of ML. Here, your objective is to arrive at a mathematical model that accurately represents your training data at scale.

As with any mission-critical system that makes decisions for your business, this system must be monitored and its decisions must be assessed. Bad decisions are often fed back into the training-data process, so you will want to continuously improve and evolve the model over time to ensure it performs well in its domain. This is a natural part of deploying production systems into the marketplace.

YOUR AI DATA PRODUCTION LINE



As you scale your data operations for AI development, you'll need access to skilled people in the loop who can transform messy information into structured data with high accuracy and consistency across your datasets. Your workforce choice might be the factor that determines your success. A managed team combines technology and people to support direct communication, flexibility to iterate your use cases and tasks, and increases in domain knowledge and context over time.

For more information about how CloudFactory's managed teams can accelerate and scale quality training data for your AI application, visit us at cloudfactory.com.

ABOUT CLOUDFACTORY

CloudFactory combines people and technology to create your workforce in the cloud. Our managed teams process data for artificial intelligence and mission-critical business operations. With use-case experience across industries, we serve innovators whose projects require accuracy, agility, and scale. Our workers are experts in the art and science of digital work and can use virtually any tool, even the ones our clients build. Headquartered in Reading, UK with a globally distributed workforce, we put disruption within reach.

- TRUSTED BY 130+ COMPANIES
- EXPERIENCE WITH 150+ AI PROJECTS
- CLIENTS INCLUDE 11 OF THE WORLD'S TOP AUTONOMOUS VEHICLE COMPANIES
- MILLIONS OF TASKS PROCESSED A DAY
- 3 MILLION HOURS OF MEANINGFUL WORK IN 2018

CLOUDFACTORY.COM

UK • USA • NEPAL • KENYA

CONTACT@CLOUDFACTORY.COM