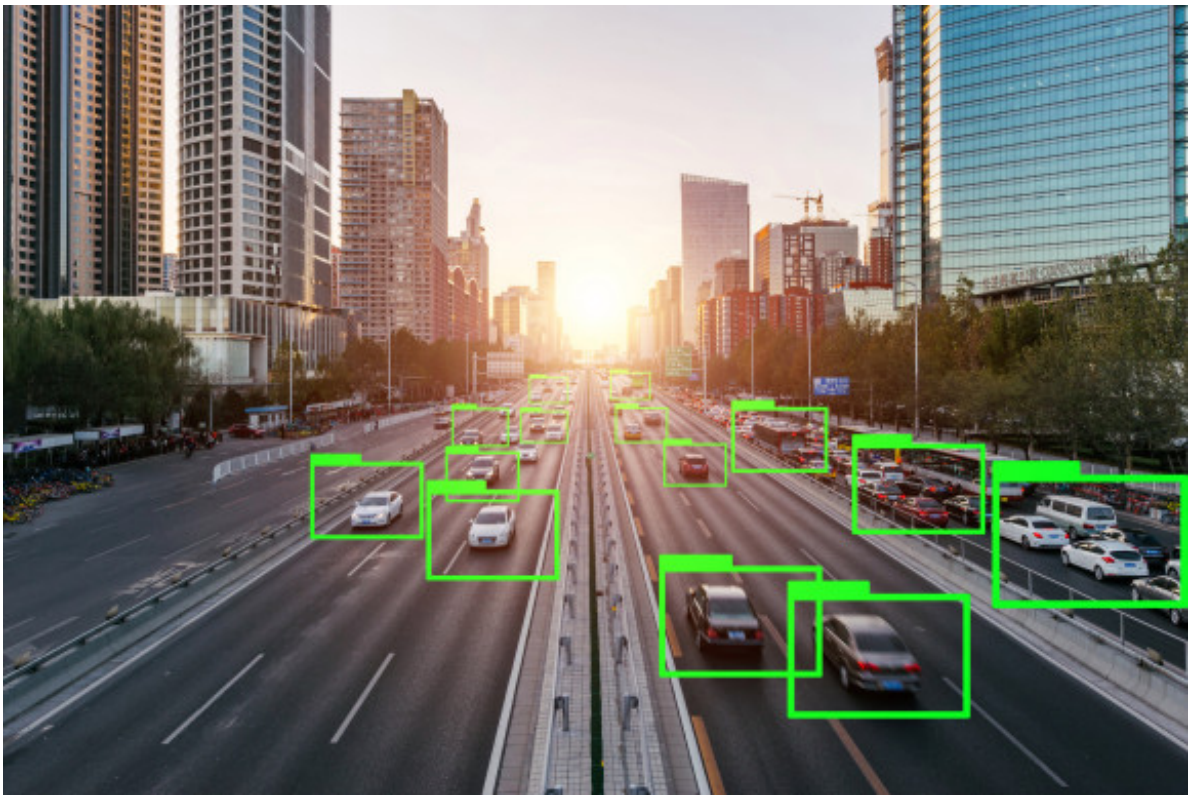
Essential tips for scaling quality AI data labeling

VB venturebeat.com/2019/06/12/essential-tips-for-scaling-quality-ai-data-labeling/

Damian Rochman, CloudFactory

June 12, 2019



Presented by CloudFactory

Across every industry, engineers and scientists are in a race to clean and structure massive amounts of data for AI. Teams of computer vision engineers use labeled data to design and train the deep learning algorithms that self-driving cars use to recognize pedestrians, trees, street signs, and other vehicles. Data scientists are using labeled data and natural language processing (NLP) to automate legal contract review and predict patients who are at higher risk of chronic illness.

The success of these systems depends on skilled humans in the loop, who label and structure the data for machine learning (ML). High-quality data yields better model performance. When data labeling is low quality, an ML model will struggle to learn.

According to a report by analyst firm Cognilytica, about 80 percent of AI project time is spent on aggregating, cleaning, labeling, and augmenting data to be used in ML models. Just 20 percent of AI project time is spent on algorithm development, model training and tuning, and ML operationalization. These tasks are at the heart of AI development and require

strategic thinking, along with a more advanced set of engineering or computer science skills. It's best to deploy more expensive human resources — such as data scientists and ML engineers — on tasks that require expertise, collaboration, and analytical skills.

Comparing data labelers for machine learning

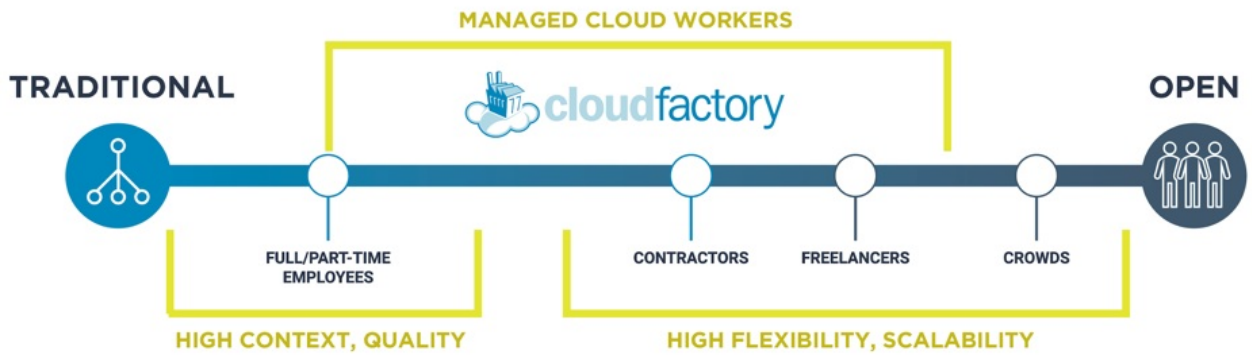
A growing number of organizations are using one or more of these four options to source data labelers for AI projects. Each choice brings benefits and challenges, depending on project needs.

1. Full-time and part-time employees can manage data labeling with good quality, and this approach works fine until it's time to scale. There will be some worker churn, and the existing team will have to bring each new worker up to speed, adding cost and management burden.

2. Contractors and freelancers are another option. It takes time to source and manage a contracted team. If human resources is not involved in hiring contractors, workers may not be subject to the same cultural and skills assessments used for full-time employees. That can be a problem when it comes to quality labeling, so it will require additional time for training and management.

3. Crowdsourcing uses the cloud to send data tasks to a large number of people at once. Quality is established using consensus: several people complete the same task, and the answer provided by the majority of workers is chosen as correct. We've used this model in the past for data work at CloudFactory and our client success team found consensus models cost about 200 percent more per task than processes where quality standards can be met from the first pass. The burden is on the AI team to manage workers' data outputs at scale. Crowdsourcing is a good option for short-term projects.

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 high-quality data labeling, a task that often requires workers to understand the context. Labelers on a managed team increase their understanding of your business rules, edge cases, and context over time, so they can make more accurate subjective decisions that result in higher quality data.



After a decade of data labeling, transcription, and annotation for organizations around the globe, we've learned that it is critical to establish a closed feedback loop between AI project teams and data labelers. Tasks can change as development teams train and tune their models, so labeling teams must be able to adapt and make changes in the workflow quickly.

Workforce solutions that charge by the hour, rather than by the task, are designed to support these iterations. A 2019 Hivemind study shows that paying by task can incentivize workers to complete tasks quickly at the expense of quality.

Critical questions to ask when sourcing a data labeling team

We encourage organizations to ask workforce vendors these questions as they compare data labeling workforce options:

- Scale: Can your labeling team increase or decrease the number of tasks they do for us, based on demand?
- Quality: Can you provide us with visibility into work quality and worker productivity?
- Speed: What is your track record for on-time delivery of data labeling work?
- Tool: Do we have to use your tool or can we build our own?
- Agility: What happens if our tools or processes change?
- Contract terms: What happens if we need to cancel our work with your labeling team?

To further explore how to choose a data labeling workforce for quality, speed, and scale, download this report: Scaling Quality Training Data: Optimize Your Workforce and Avoid the Cost of the Crowd.

Damian Rochman is VP of Products and Platform Strategy, CloudFactory.

Sponsored articles are content produced by a company that is either paying for the post or has a business relationship with VentureBeat, and they're always clearly marked. Content produced by our editorial team is never influenced by advertisers or sponsors in any way. For more

information, contact sales@venturebeat.com.