

Successful AI Development Means Fielding The Best Team

F forbes.com/sites/marksears1/2019/01/07/successful-ai-development-means-fielding-the-best-team/

January 7,
2019



Mark Sears

Contributor

AI & Big Data I write about AI, data and the future and science of work

If AI

Breakthroughs in AI happen with a team of people working behind the scenes to make it all possible.
Getty

development were a sport, it'd be closer to baseball than boxing. Headlines might make it seem like AI breakthroughs happen with a big knockout punch, but the reality is more akin to a baseball team grinding through a 162-game season. It's a process that involves having the right people in place over a long stretch, and fielding the best team is essential for success.

The root cause for this can be found in the data. As AI continues to mature and new technologies are introduced, they are only as good as the datasets that feed them. This translates into workers spending a combined hundreds or even thousands of hours cleaning, structuring and organizing the data used to develop and train AI systems.

Given the challenges of hiring and managing a team of people to complete this arduous data work, many development and operations leaders turn to outside help to recruit and manage this workforce. And, much like a baseball team trying to make a World Series push, some approaches to building the right team can yield better results than others.

The Scouting Report on AI Data Workers

It's a problem of speed and scale. Hiring for AI is on the rise, according to Indeed. And there's more data to work with than ever before. An increasing number of sensors, systems and devices are producing more complex data, and on a massive scale. According to IBM research, 80 percent of data growth is generated from video, images and documents. And 90 percent of this data is unstructured, so it needs to be analyzed and organized by people to be useful for technologies like AI. If these tasks aren't completed carefully, systems run the risk of using "dirty data" - or datasets that are organized incorrectly, inaccurately or with algorithmic bias - and that can create problems with AI models.

Historically, when it comes to AI development, in-house teams have managed data needs with reasonably good quality. But over time, these processes have grown more difficult and costly to manage internally, and more companies are turning to contractors, freelancers and gig workers to get the dirty work of AI development done.

Crowdsourcing has emerged as perhaps the most common alternative to an in-house team. Workforce vendors use the cloud to send out data tasks to many workers and pay them based on the number of tasks they complete. But AI models require high accuracy and consistency that crowds can't deliver. To build high-quality AI, you need high-quality data, and that means your workers must have domain knowledge - or a strong understanding of your rules and edge cases - so they can make informed, subjective decisions. They also must process your data with both accuracy and consistency across the team. When you use the same team, workers get more proficient with your domain and your data over time. When you throw the tasks out to the crowd, you discard these benefits of a dedicated team.

Imagine you are training a natural language processing (NLP) system to classify news articles that reference solar energy on a five-point scale of highly relevant to highly irrelevant, to create AI that delivers critical information to energy executives. When you use the same data team, their accuracy and consistency in classifying that data improves over time. They also can incorporate process changes faster.

This is typical of complex environments and learning algorithms in general, not just those for NLP. Workers who are better at classifying, labeling, or annotating data deliver higher quality data and better model performance. It's that simple. Having fast and easy access to the same workers who classified your original dataset, for example, can make a substantial difference in your overall speed and quality.

Using an anonymous crowd might feel like a baseball team trying to select a pitcher or right fielder without a scouting report. You may luck into the right skill set, or you may not; there's no way to know what you're getting. And in a business where the winners are the ones who get technology to market fast, this uncertainty can create significant risk. Anonymous workers are simply not the ace you want to have on the mound for important moments.

Building for the Playoff Stretch

As an alternative to crowdsourcing, managed workforce solutions are gaining momentum. In this scenario, companies recruit partners who act as an extension of their internal teams. Vetted workers can be matched and trained for each project based on its specific data needs. This model not only provides the benefits of a trained, internal team, it gives AI project leaders more visibility into the quality of the work and makes it easier to communicate changes to workers. Using the same workers on a project over time makes it easier to adjust data requirements as needed.

Like a successful professional baseball team, a managed workforce can deliver strong results through the grind of an AI development project. A managed team provides a deeper bench for AI teams, taking the guesswork out of complex data work and avoiding the rework that crowdsourcing can create.

Headed for Home

While the onus is on each ballclub...er, business... to decide how to field their team, it's important not to take data workforce decisions lightly. The right process for assigning complex data work can make the difference between a winning or losing season – or a successful or failed AI project. Having the right game plan in place to manage workforce demands will help put companies in the best position for on-field success.