Crowd vs. Managed Team:

A Study on Quality Data Processing at Scale

+

ABSTRACT

Whether you're training machine learning algorithms or using traditional analysis techniques, the quality of your data determines performance. Data-science tech developer Hivemind enlisted CloudFactory's managed workforce and a leading crowdsourcing platform's anonymous workers to complete a series of the same tasks, ranging from basic to more complicated, to determine which team delivered the highest-quality structured datasets and at what relative cost.

This report includes study results, lessons learned, and insights that will help you strategically deploy people to enhance the quality of your datasets, which can free up your highest-value team members to focus on machine learning performance and more complex data analysis.

IN THIS PAPER, YOU WILL LEARN

The difference in accuracy you might expect from using an anonymous crowdsourced team versus a managed team of data workers

The interesting behavioral impact of paying workers by the hour - rather than by the task - and how it can affect quality

Factors that can help you strategically deploy the right workforce for optimal results

WHY HIVEMIND CONDUCTED THIS STUDY

Hivemind provides software to assist teams tackling the challenges of working with messy or unstructured data. Our method involves breaking these problems down into conceptually simple bite-size chunks, or microtasks, which are then distributed as appropriate to computational methods or human workforces. Hivemind aggregates the results into a structured dataset ready for analysis.

To make the human part of this process accurate and efficient, it is vital to be able to select the appropriate workforce for the task at hand. Depending on the question you want your data to answer, that could be a crowdsourced solution, a managed service solution such as CloudFactory, or a workforce or individuals within a client's organisation who have expertise in a particular field.

Each of these workforces comes with advantages and disadvantages. We designed this study to understand those dynamics in more detail.

DEFINITION OF WORKFORCE TERMS

Crowdsourcing refers to using the cloud to send data tasks to a large number of anonymous workers at once. Crowdsourced workers typically are not screened and are paid by the task.

Managed cloud workers are recruited and screened in a more traditional way, and they are paid an hourly wage.

METHODOLOGY

We designed our study to compare two types of workforces, crowdsourced and managed, in terms of data quality and relative cost.

The Hivemind platform is well suited to running an experiment across multiple workforces because it's easy to route the tasks as required and allows you to monitor the output in granular detail - both in terms of accuracy and time taken.

The experiment consisted of three tasks representative of the kinds of problems our clients find they need workforces for. Tasks included:

1. Basic transcription,
2. Assess sentiment from text, and
3. Categorise an event from unstructured text.

To avoid potential bias, neither the crowdsourced workers nor the managed workers knew that they were participating in an experiment.

WORKFORCE & COSTS

We paid the managed workforce by the hour and paid the crowdsourced workforce per iteration, or task. To normalize costs across workforces, we expressed all costs in terms of the cost per minute of the managed workforce.

TASKS

TASK A: EASY TRANSCRIPTION

For each instance, contributors were asked to:

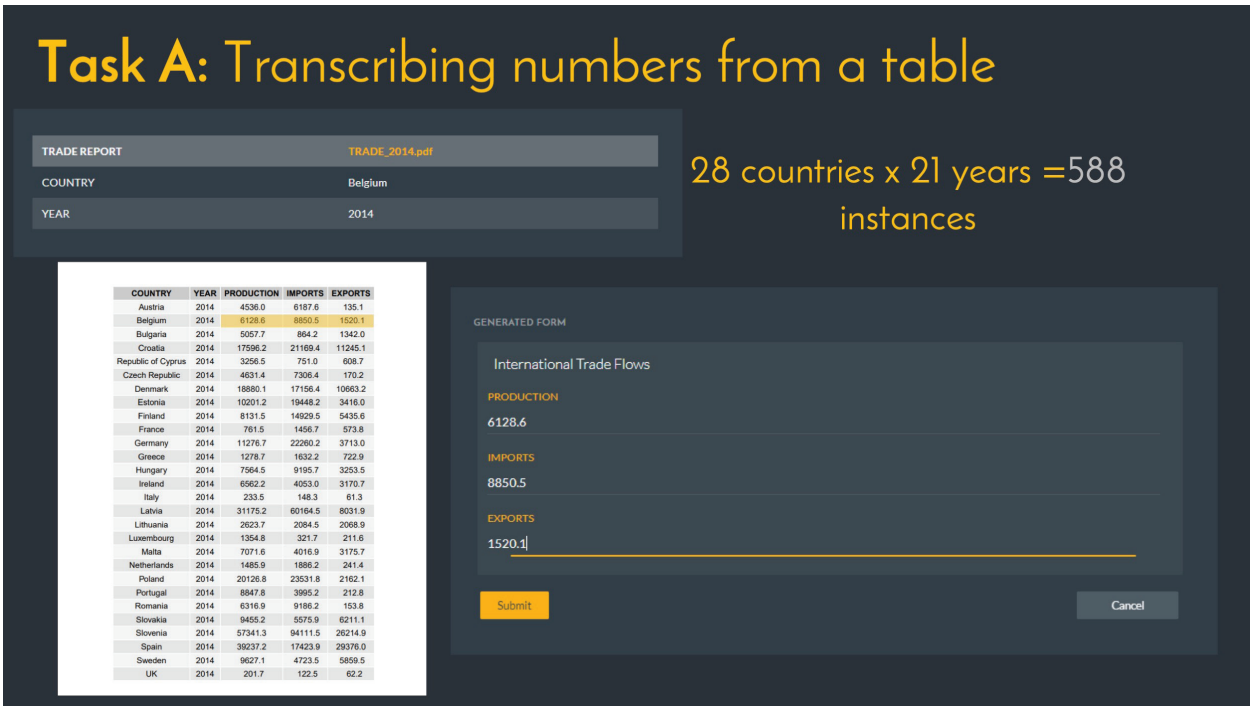
1. Open a PDF file containing a table of trade-flow figures for 28 European countries for a particular year.
2. Provide three trade numbers – production, imports, and exports for a specified country. In Figure A:1, they were asked to transcribe numbers for Belgium for 2014, which are highlighted in the graphic. (They were not highlighted in the actual task.)

Overall, each workforce completed 588 iterations of this task. In practice, this is the kind of task that could be done computationally but for the purposes of this study, it was designed to measure the basic transcription error rate of each workforce.

COST CONSIDERATIONS

Hivemind sent the task to the crowdsourced workforce at two different rates of compensation: 0.4 cost-units per iteration and 0.8 cost-units per iteration. The workers who received 0.4 units per iteration would have to do 2.5 iterations per minute to cost the same as the managed workers, while the workers receiving 0.8 units would only have to do 1.25 iterations per minute.

FIG. A:1



COUNTRY	YEAR	PRODUCTION	IMPORTS	EXPORTS
Austria	2014	4536.0	6187.6	135.1
Belgium	2014	6128.6	8850.5	1520.1
Bulgaria	2014	5057.7	864.2	1342.0
Croatia	2014	17596.2	21169.4	11245.1
Republic of Cyprus	2014	3256.5	751.0	608.7
Czech Republic	2014	4631.4	7306.4	170.2
Denmark	2014	18880.1	17156.4	10663.2
Estonia	2014	10201.2	19448.2	3416.0
Finland	2014	8131.5	14929.5	5435.6
France	2014	761.5	1456.7	573.8
Germany	2014	11276.7	22260.2	3713.0
Greece	2014	1278.7	1632.2	722.9
Hungary	2014	7564.5	9195.7	3253.5
Ireland	2014	6562.2	4053.0	3170.7
Italy	2014	233.5	148.3	61.3
Latvia	2014	31175.2	60164.5	8031.9
Lithuania	2014	2623.7	2084.5	2068.9
Luxembourg	2014	1354.8	321.7	211.6
Malta	2014	7071.6	4016.9	3175.7
Netherlands	2014	1485.9	1886.2	241.4
Poland	2014	20126.8	23531.8	2162.1
Portugal	2014	8847.8	3995.2	212.8
Romania	2014	6316.9	9186.2	153.8
Slovakia	2014	9455.2	5575.9	6211.1
Slovenia	2014	57341.3	94111.5	26214.9
Spain	2014	39237.2	17423.9	29376.0
Sweden	2014	9627.1	4723.5	5859.5
UK	2014	201.7	122.5	62.2

TASK A RESULTS

Figure A:2 shows the results from Task A. The horizontal axis gives the percent error rate of each workforce. At 0.4 units per iteration, the crowdsourced workforce has an error rate of just over 7% per instance. That means that for each instance, which involved transcribing three numbers, at least one of the numbers was incorrect in 7% of cases. When the compensation was doubled to 0.8 units per iteration, this error rate fell to just under 5%, which is a significant improvement. However, even the more highly compensated crowdsourced workers had an error rate of more than 10 times what the managed workforce achieved. The managed workers only made a mistake in 0.4% of cases, a significant difference, both in a statistical sense and in a practical sense, given its implication for data quality.

Crowdsourced workers had an error rate of more than 10x the managed workforce.

This large difference in performance raises questions about why the managed workforce was so much better at this relatively straightforward task. Figure A:3 shows the average number of seconds taken for each iteration by the different workforce groups. It shows that in this task, the lowest compensated crowdsourced workers were on average the slowest and the managed workforce was the fastest. However, the differences in average time between workforces are not large and are within the uncertainty of the estimates. Example A:3 shows the managed workforce did not achieve higher accuracy simply by taking more time.

FIG. A:2

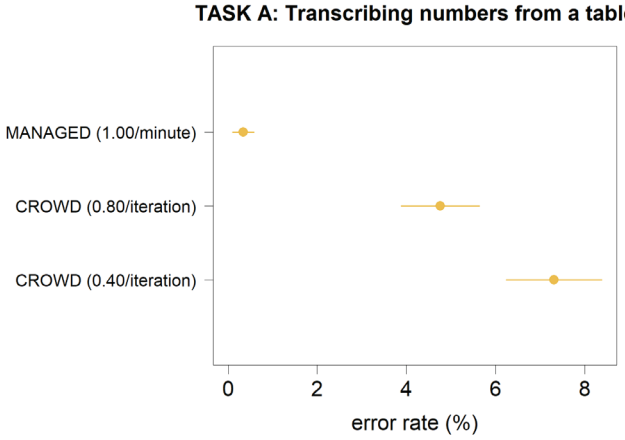


FIG. A:3

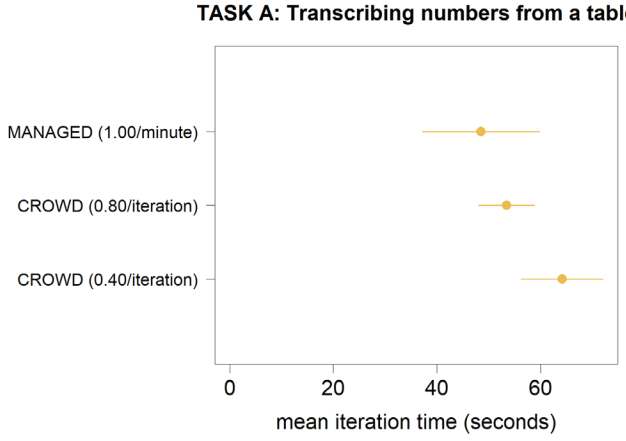


Table 1 shows the number of different types of errors each workforce made. The most common type of error was where workers keyed in accurate numbers for the correct country but for the wrong year. The crowdsourced workers receiving 0.40 units per iteration did this in 24 of the 588 instances. If workers were careless, they may have entered the numbers from the wrong document. In six instances, the crowdsourced workers used the document for the right year but entered a row corresponding to the wrong country. In another six cases, the same workers entered numbers from the wrong row of the wrong document. In yet another six cases, there was no row in any of the documents that corresponded to what the workers entered. The managed workforce only made errors in two instances.

TABLE 1

ERROR TYPE	CROWD (0.40/iteration)	CROWD (0.80/iteration)	MANAGED (1.00/minute)
non–numeric	1	0	0
year incorrect	24	9	1
country incorrect	6	7	0
both incorrect	6	4	0
no match	6	8	1

COST CONSIDERATIONS

The managed workers took an average of 51 seconds per task, which translates to a compensation rate of 0.85 units per iteration. This is only 6% more than the compensation crowdsourced workers working at the higher rate received, so it is unlikely that the large difference in accuracy can be explained simply by the managed workers' having received more.

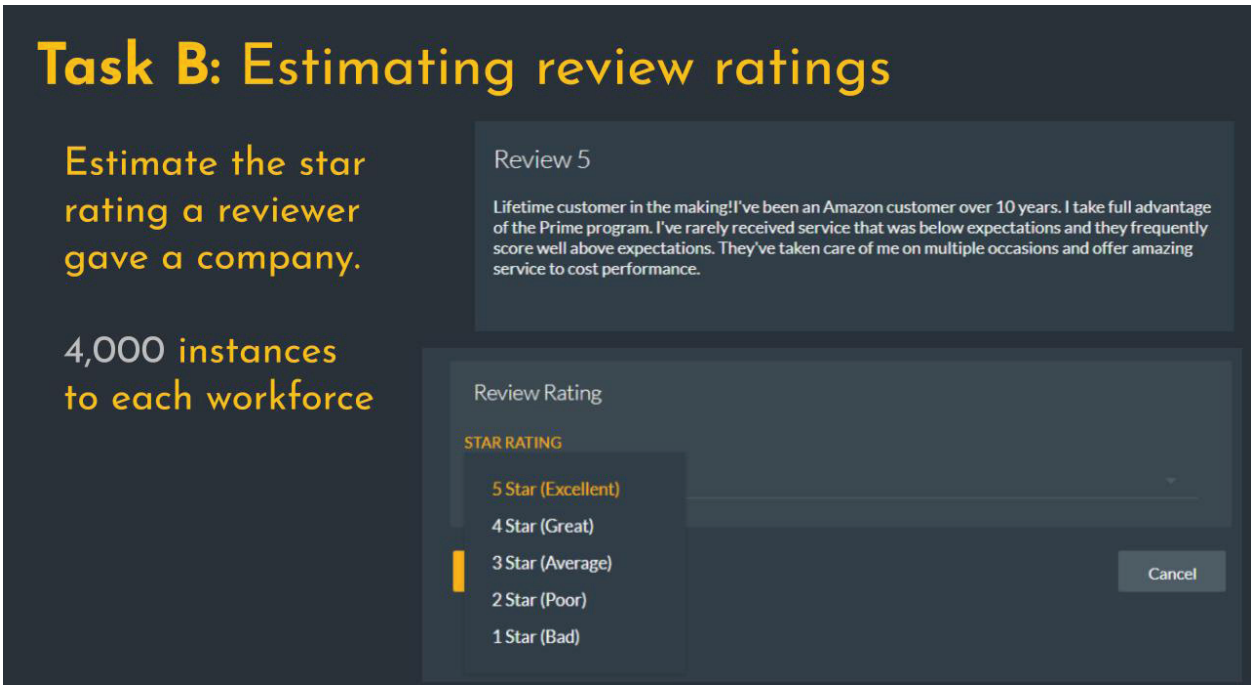
TASK B: SENTIMENT ANALYSIS

The second task we gave each workforce was much more subjective.

Workers were presented with the text of a company review from a review website. The original review came with a rating from 1 to 5 stars but we stripped these from the reviews and asked the workers to estimate what rating they think the reviewer had given the company, based solely on the text of the review.

In Figure B:1, the text of the review is very positive so one might reasonably guess the reviewer gave a 5-star rating. We sent each workforce the same 4,000 reviews, then compared their estimates with the actual ratings given by the person who wrote the review.

FIG. B:1



TASK B RESULTS

As shown in Figure B:2, we broke down the accuracy of the estimates by the rating from the original data. When we did this, we found the managed workers had fairly consistent accuracy, getting the rating (out of 5 options) correct in about 50% of cases, irrespective of whether the reviews were good or bad.

However, the crowdsourced workers seemed to have a problem, particularly with poor reviews. Their accuracy was almost 20%, essentially the same as guessing, for 1- and 2-star reviews. For 4- and 5-star reviews, there was little difference between the workforce types on accuracy.

The crowdsourced workers' accuracy on 1- and 2-star reviews was almost 20% - essentially the same as guessing.

Whether the review was good or bad was not the main driver of accuracy. This was a confounding variable, since there was a strong relationship between how good the review was and the length of the review. As shown in Figure B:3, positive reviews tended to be short, with just a few words, such as "excellent customer service." Negative reviews often included lengthy explanations or assertions about what went wrong, in the reviewer's opinion.

FIG. B:2

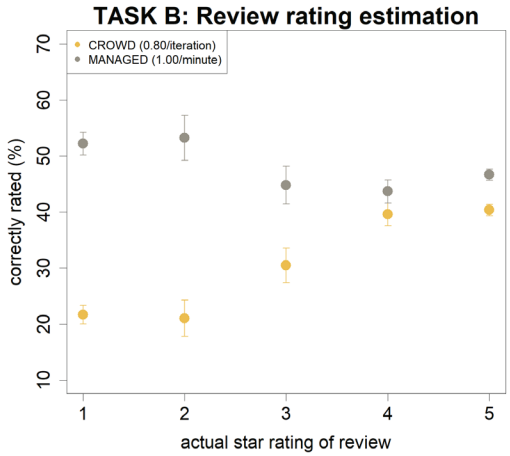
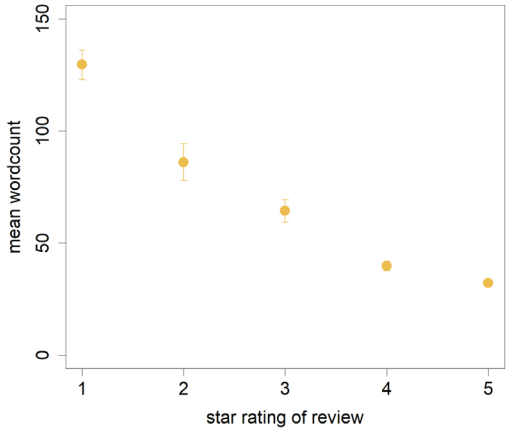


FIG. B:3



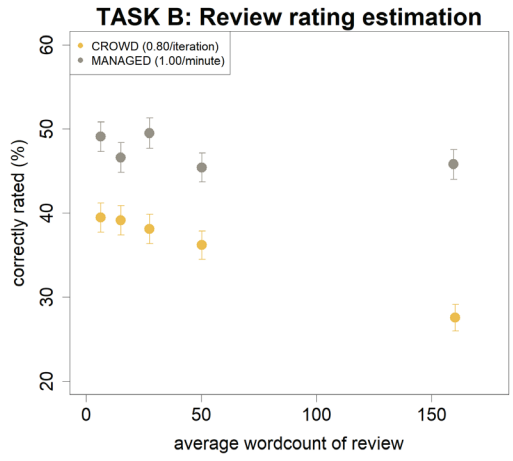
When we break up the reviews into quintiles, by the number of words, we see that as reviews get longer, the accuracy of the crowdsourced workers declines. For the shortest 20% of reviews, crowdsourced workers estimated the rating correctly 40% of the time, and for the longest 20% of reviews, accuracy fell to less than 30%.

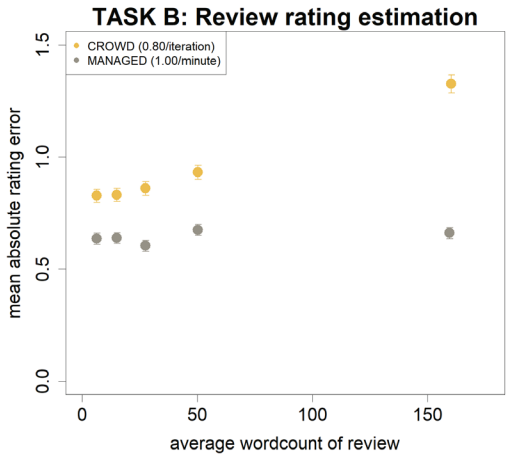
In contrast, the accuracy of the managed workers was insensitive to the length of the review. To understand what might be driving this, we can look at how long the different workforces took per iteration as a function of the length of the review.

The Hivemind platform provides metrics on how long it took workers to complete each case so we can generate a graph like Figure B:4, which shows the average number of seconds taken for each quintile of review length. For the shortest 20% of reviews, both workforces took an average of about 20 to 25 seconds to estimate the rating. As the reviews get longer, the time spent by the managed workforce quickly increases, while the maximum time spent by the crowdsourced workforce is about 35 seconds.

The managed workers spent an average of 80 seconds on the longest 20%, more than double the time the crowdsourced workers spent.

FIG. B:4





COST CONSIDERATIONS

Interestingly, when we look at the average time spent per iteration by the managed workers, we find the costs per iteration are very comparable between the two workforces: 0.80 units per iteration for the crowd and 0.77 units per iteration for the managed workers. But these averages conceal the fact that the crowdsourced workers vary the amount of time per iteration much less than the managed workers, who spent substantially more time reading the longer reviews.

When we consider how these workforces are paid, this difference in behavior makes sense. The crowdsourced workers are being paid per iteration. That could incentivize them to do as many iterations as quickly as possible. Managed workers are paid for their time, so they seem to be more willing to spend longer on the more difficult, or in this case longer, cases. This difference has important implications for data quality.

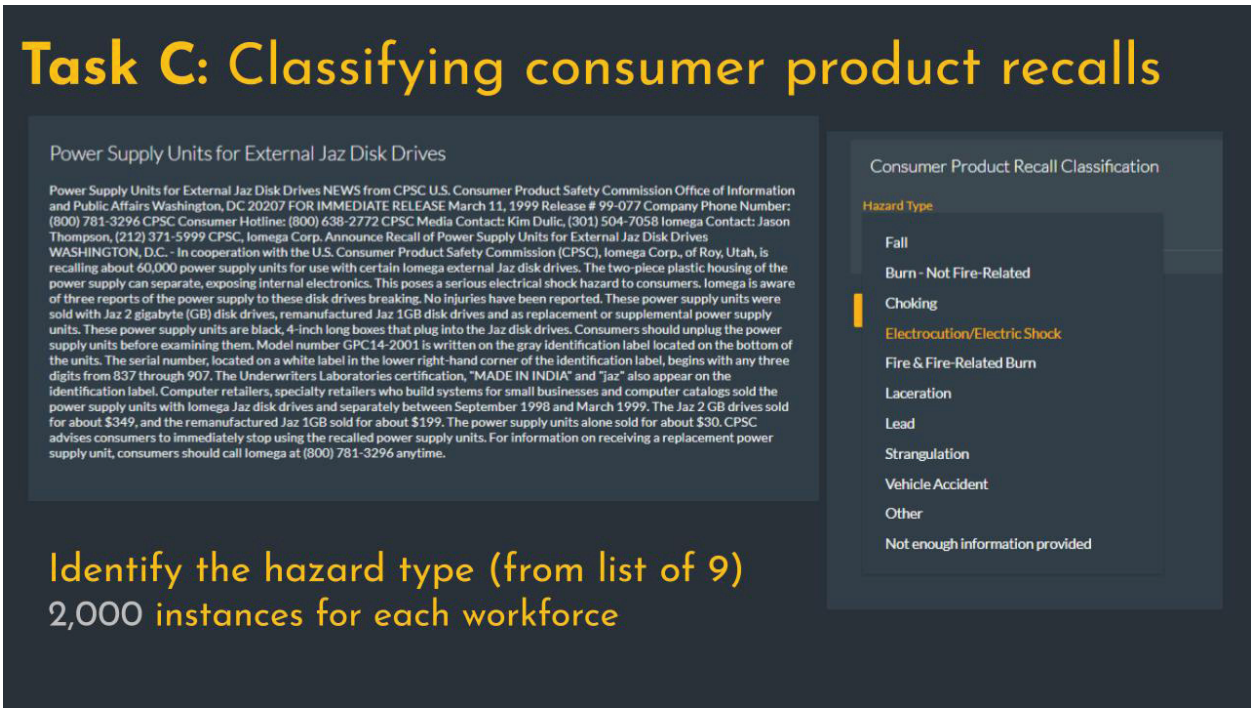
TASK C: EXTRACTING INFORMATION FROM UNSTRUCTURED TEXT

For the final test, Hivemind presented workers with the title and description of a product recall issued by the U.S. Consumer Product Safety Commission. The Commission gives a hazard type classification for every product recall but we did not provide these to workers.

We asked them to determine what the hazard type was from the text. As shown in Figure C:1, workers could choose from a drop-down menu of nine hazard-type classifications used by the Commission. We provided two additional options: "other" and "not enough information provided."

In some recalls, the hazard type is buried in the text while in other cases, it is explicitly stated in the title. We sent 2,000 recalls to the crowdsourced workforce and the same 2,000 recalls to the managed workforce.

FIG. C:1



TASK C RESULTS

As with Task B, we can break down the results based on the length of the recall. As shown in Figure C:2, when we do this, we do not see decreasing accuracy as the text gets longer for either workforce, as we saw with the crowdsourced workers on Task B. The crowdsourced workers achieved accuracy of 50% to 60%, irrespective of the word count of the recall, while the managed workers achieved higher accuracy, 75% to 85%.

Why is the accuracy of the crowdsourced workers lower than the managed workers? As shown in Figure C:3, if we look at the distribution of responses from each of the workforces, we see that while both workforces chose the "not enough information" response with the same frequency, the crowdsourced workers were much more likely to answer "other" - in fact, 4 times more likely.

FIG. C:2

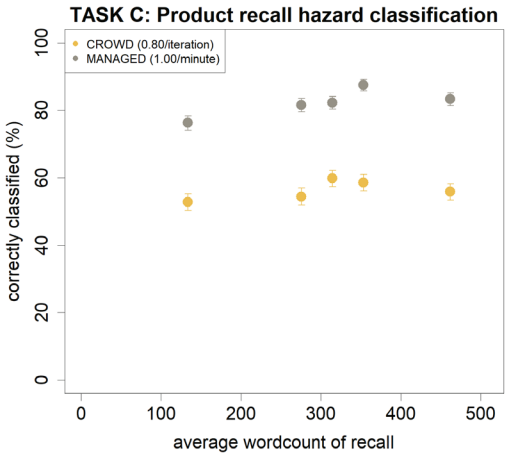
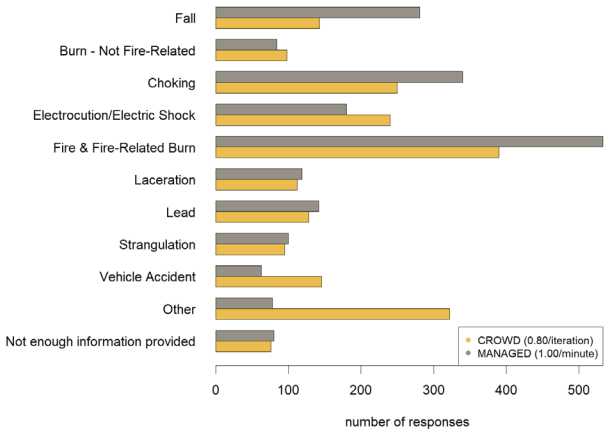


FIG.C:3



Managed workers achieved 25% higher accuracy than crowdsourced workers.

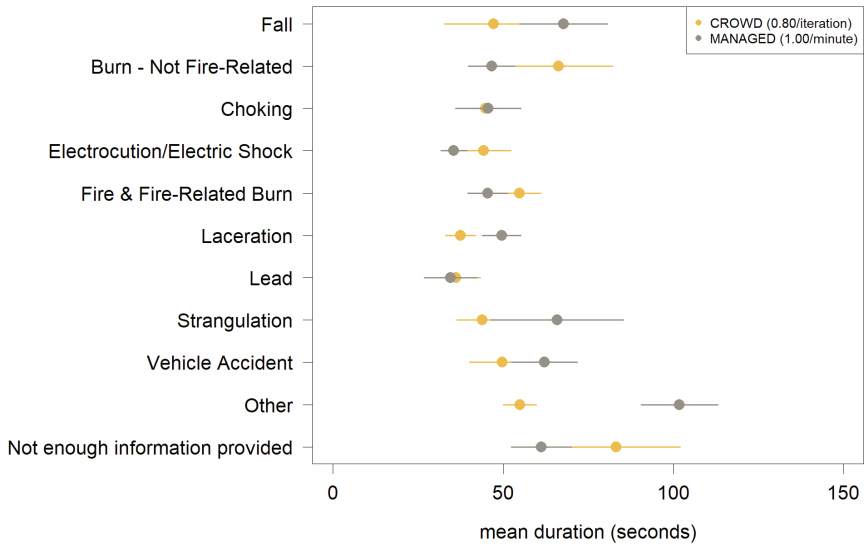
Crowdsourced workers were 4x more likely than managed workers to default to "other" as a task answer, even when the text provided the information to make a more accurate classification.

But if we just take these 322 cases where the crowd answered "other" we find that the managed workers only classed 10% of these as "other" and correctly classified 74% of the recalls, implying that the information required to make a correct classification was present in the recall text in the vast majority of these cases.

As shown in Figure C:4, if we break down the time spent on each instance conditioned on the response given we can see that the crowdsourced workers took an average of less than 50 seconds before responding "other" while managed workers would spend over twice as long before resorting to this response.

It appears that overuse of the "other" category explains some of the 25% accuracy gap between the workforces. However, even when we remove these cases, the managed workers still classify 16% more cases correctly than the crowdsourced ones.

FIG. C:4



COST CONSIDERATIONS

There was little dependency between the length of the recall text and the amount of time either workforce spent. There also does not appear to be a meaningful difference between the time it took each workforce to do the task. Both workforces took an average of about 50 seconds to classify each recall. As a result, the managed workers, who were paid by the hour, cost the equivalent of 0.87 units per iteration, slightly higher than the cost of the crowdsourced workers.

SUMMARY

In all three tasks, the managed workforce outperformed the crowdsourced workers in terms of accuracy, even when the effective costs per task were similar for each workforce. In the case of Task B, the difference in performance can be explained by the managed workers' being more willing to spend longer to read lengthier reviews because of the way they are compensated. With Tasks A and C, it's less obvious what might be behind the substantial differences in accuracy between the two types of workforces.

WORKFORCE FACTORS TO CONSIDER

What the study clearly demonstrates is that there can be large differences in the accuracy of data analyzed by different workforce types. When choosing a workforce, it is important to consider your data quality requirements and to what extent you can sacrifice data quality for other workforce characteristics, such as rapid scalability or fast turnaround.

When you are specifically considering whether to use a managed team or crowdsourcing, here are factors to consider:

In general, crowdsourcing can be a good model when you need a lot of people to do the work right away, task iteration is unlikely, and quality is not of paramount importance. The rapid data turnaround

can be helpful in establishing a process and defining business rules. At Hivemind, we have found **it can cost up to two times more to use a crowd**, because it distributes the same task to multiple people and often requires a consensus model with multiple people completing or reviewing tasks to achieve passable quality. So while crowdsourcing offers a cheap option for training machine learning models, it's rarely as inexpensive as it seems. Watch for hidden fees in technology, onboarding, and training with crowdsourcing.

A managed team is a better choice when quality is important, and you want to be able to iterate or evolve the work. With a managed team, you can create a closed feedback loop with your workers that makes it possible to evolve your tasks over time. This is especially important if you are developing AI because that process requires collaboration and strong communication across teams of people, many of whom are doing disparate work. AI models require high accuracy and consistency, which crowdsourcing can't deliver.

HIVEMIND

Hvmd.io
LONDON, UK
henrik@hvmd.io

Hivemind creates software that helps companies build, clean, and enrich datasets from messy or unstructured information. We fuse computational methods with distributed human intelligence, integrating to support internal, outsourced, and crowdsourced contributors. Our REST API makes it possible to seamlessly embed with ongoing workflows. Hivemind data scientists provide support on all of our client engagements.

cloudfactory

CloudFactory.com
UK • USA • NEPAL • KENYA
hello@cloudfactory.com

CloudFactory combines people and technology to create your workforce in the cloud. Our managed teams process data with high quality at scale for machine learning and mission-critical business operations using virtual any tool. Our technology puts you in contact with your data team on the ground for easier iteration of tasks and use cases. With expertise across 150+ AI projects, we put disruption within reach.