Crowdsourced Data Management Survey: Overivew of Applications and Techniques
Group 13

He, Yonglin (1118835)
yonglinh@student.unimelb.edu.au

Liu, Jiahe (1214235)
jiahe3@student.unimelb.edu.au

Park, Jongho (1152502)
jonghop@student.unimelb.edu.au

Xu, Yuanbo (1118904)
yuanbo@student.unimelb.edu.au

Abstract. This survey provides an overview of crowdsourced data management techniques and applications by identifying gaps in current studies. Crowdsourced data management related techniques are compared in terms of quality control, cost control, latency control, and incentive design. Limitations and future directions are discussed. These will be presented in the following structure.

1 Introduction

Data management involves the process of collecting, cleaning, and organising data to support operations, decision-making processes, and enhance productivity. Crowdsourced data management (CDM) is the process where collective intelligence and labour force of a crowd is utilised for data management [1]. The main reason for applying crowdsourcing in data management is to utilise the crowd to perform computer-difficult tasks such as entity resolution and data preprocessing where data are in different formats. Hence, CDM was introduced to enhance scalability and efficiency by harnessing the workforce of the crowd when computer-based algorithms were insufficient to solve the tasks.

While previous surveys have covered various topics on CDM, they have not provided a comprehensive review of its chronological development or compared the various approaches used. In light of this gap, the aim of this survey is to systematically review the development, application, and optimisation of CDM by distinguishing the gaps between studies, and to categorise and compare existing approaches. The survey will be structured as follows: the historical development, general application, and architecture of CDM will be introduced in *Overview*. The application and optimisation approach will be reviewed in the *Related Work* section respectively. Comparisons of the pros and cons for the mentioned approaches will be presented in the *Comparison* section.

1.1 Overview

With the development of open call technology, Howe coined the term "crowdsourcing", referring to outsourcing a job traditionally done by employees to an unknown large crowd through an open invitation [2].

Table 1: Application of Crowdsourcing in Data Management

Application	Methods	Example	Platform
Micro-tasking	Reward-based Task Completion	Data Management Tasks	Amazon's Mechanical Turk
Idea Generation	Idea Competition	Find better product design	Kaggle
	Collective Ideation	Online Encyclopaedia	Wikipedia
Public Participation	Citizen Science	Contribute crowd's data for research	Google Map (POI)
	Citizen Journalism	Contribute crowd's opinions on news	CNN iReport
Technology Development	Open-source Software	Open-source operating system	Linux
		Open-source web server	Apache

Crowdsourcing can be classified into four main fields of application, including idea generation, micro-tasking, public participation, and technology developement. As demonstrated in Table 1, CDM is primarily involved in micro-tasking. During this process, a query from crowdsourcing requester will be processed, optimised, and executed to create human-intelligent tasks. These tasks are then published on crowdsourcing platforms, enabling workers to access and produce answers [3]. However, since data management also includes data collection, the CDM process is also involved when the crowd contributes data via platforms. For instance, on Google Map, the crowd contributes Point of Interest (POI) data for research and analysis purposes.

1.2 Crowdsourcing Data Management Architecture

As demonstrated in Figure 1, CDM consists of task-based management and data-based management. Micro-tasking jobs such as data cleaning and data labelling are mainly realised through task-based management where a the requester's query will be processed by task manager and published on Micro-tasking platform. A set of CDM techniques will be performed to guarantee the result quality and operation efficiency. In this process, workers will get designed task and be evaluated by their performance. Based on worker's result, the crowdsourcing executor will evaluate the query to return the results to requester [4].

For data-based management, the interest of requester would mainly be collecting data contributed by the crowd. In this process, tasks will be published on specialised platform for idea generation or technology development. Crowd can contribute their idea via the specialised platforms such as Wikipedia or Kaggle. Additionally, crowd's can also contribute personal data via websites or mobile devices. These data also can be crowdsourced with the crowd's consent as they agree to cookies.

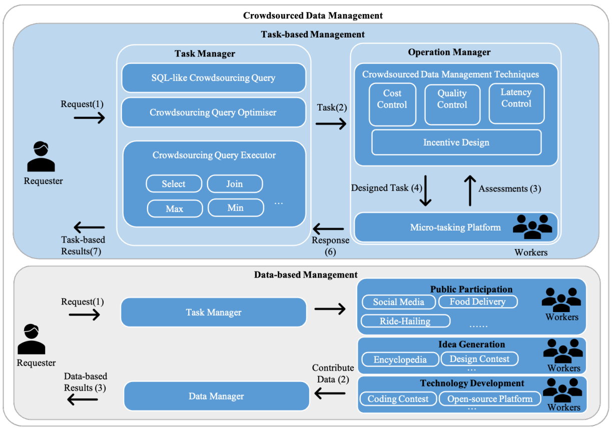

Figure 1: Crowdsourced Data Management Architecture

2 Related Work

As suggests by Table 1, CDM can be utilised with various techniques in multiple fields, this section aims to provide an overview of the application and techniques of CDM.

2.1 Crowdsourced Data Management Applications

There are various forms of services to provide the tasks or projects to crowd workers. Crowdsourcing platform allows people to interact with products and rewards such as tasks and monetary rewards. Platforms can be divided into general and specialised platforms based on their purpose. General-purpose crowdsourcing platforms provide a broad range of services for a wide variety of projects or tasks, while specialised platforms focus on specific types of projects or industries and often require deep knowledge of specific areas.

2.1.1 Platforms

Between the platforms, expected data aspects are different. General platforms often asks people's general behaviours and often required large amounts of data with quick responses. However specialised platforms need quality of tasks other than data size. The specialised platform normally asks for the best solution for a specific topic.

2.1.1.1 General-purpose platform

Amazon Mechanical Turk (AMT) is the most widely used crowdsourcing platform. Introduced in 2005 as an integral component of Amazon Web Services (AWS), it offers readily available, scalable, and cost-effective labor solutions [4]. AMT enables users to crowdsource tasks of varying complexity to a large pool of workers. It is accessible to Requesters, Workers, and Developers.

Requesters refer to the person who publishes tasks on the platform using a software application to access AMT services. A Requester Console is given to them for the purpose of task management and tracking their progress. If the task requires specific knowledge or skills, the requester can create a qualification test, and they can filter workers based on their quality using a parameter called Qualification Requirement.

Workers, on the other hand, are individuals who carry out tasks that have been posted by requesters. They can choose tasks based on their type and reward. Workers are paid only if their completed task is accepted by the requester.

Developers can use various APIs provided by AWS to build their own applications for crowdsourcing. In AMT, sandbox for developers is provided to simulate their environments and test their application or tasks.

2.1.1.2 Specialised platform

Kaggle is also a specialised crowdsourcing platform for idea competition, and technology development. Founded in 2010, it has become a crowdsourcing platform for companies and organisations seeking solutions to complex problems that require data analysis and modelling [5]. A large pool of talented workers with specific skill sets can participate in competitions and challenges posted on the platform and can also collaborate on projects and share knowledge through discussion forums and other resources.

Lego Ideas is a crowdsourcing platform for idea generation in commercials. It was developed as a collaboration between CUUSOO and The Lego Group in 2008. Users can submit their LEGO designs along with descriptions and features, and receive feedback through voting and comments. Ideas with

10,000 supporters undergo expert review and production procedures, and the submitter becomes a Lego designer and receives 1% of net revenue as a reward.

2.1.2 Data Management Application

Data collection can be done efficiently and effectively by crowdsourcing, where transportation data can be collected without adding financial burdens. People are willing to contribute to cycling issue-report systems such as CycleTracks and OneBusAway because they are motivated by improving their own transportation experience or the cycling infrastructure in their community.

Data cleaning and preprocessing is traditionally done by statistics and machine learning. However, this classic approach has limitations in terms of cleaning accuracy. Chu et al. introduced KATARA [6], a data cleaning system that utilises crowdsourcing to align semantics in a table and a knowledge base, and to identify correct and incorrect data. This approach is effective in data cleaning and repairing inaccurate data.

Data labelling and annotation often requires experts in their analytic domains, which can be costly and time-consuming. Crowdsourcing divides the annotation process into smaller sections and distributes to volunteers from different knowledge backgrounds. Su et al. stated crowdsourcing is very cost-effective in tasks like visual object detection annotation, which saves more than 30% of total cost [7].

Data quality assurance can often be unstable and vary due to the different processing methods, task difficulty, worker's bias. Integrated Data Labelling Engine (IDLE) shows an example of a hybrid system that combines a group of domain experts with workers on a crowdsourcing platform to maintain and improve the quality of data labelling [8].

2.2 Crowdsourced Data Management Techniques

In recent years, all kinds of research focusing on crowdsourcing data management have mostly focused on the following four core issues: quality control, cost control, Latency control and Incentive Design. The findings are categorised and presented by introducing gaps between studies.

2.2.1 Quality Control

Quality control is a core issue in the current crowdsourcing data management research, the main challenges of managing crowdsourced data is the potential for errors and inconsistencies introduced by the crowd. Crowdsourced data can be influenced by different factors, such as the quality of instructions provided to workers, and the professionalism and competence of workers. Quality control can be accomplished through the quality of the data itself, selection of workers, how to effectively aggregate feedback from different crowdsourcing participants to form high-quality task results, and task assignment.

2.2.1.1 Worker Modelling and Selection

A common quality control method proposed in many papers is to eliminate low-quality workers, which first requires modeling the quality of the workers. The quality of workers can be modeled through probability, probability with confidence interval and confusion matrix. Some papers also consider the different professional levels of workers in different fields [9].

Based on the estimation workers' quality, spammers and low quality workers could be eliminated. The filtering methods include iterative learning [10], finding outlier, workers' skills and expertise, as well as their reputation and past performance on crowdsourcing platforms [11].

2.2.1.2 Enhancing Data Quality

In addition to the workers' quality, the quality of the data provided to them is also crucial for achieving satisfactory outcomes in the tasks. Data can be preprocessed in advance before being provided to workers, such as applying data cleansing and preprocessing steps or applying computer vision techniques to the graphics, or being cleaned after completed by workers, with expert feedback, arbitration and peer review mechanisms, and whether basic facts [12] provided can be used to filtering output. Moreover, the bias introduced by workers can be further eliminated through techniques such as active learning [13].

2.2.1.3 Golden Task

The dynamic nature of online workers in crowdsourcing platforms often poses challenges when attempting to evaluate the quality of workers. As a result, studies have suggested the use of golden tasks as a means to tackle the issue of worker quality. Golden tasks involve utilising a small number of tasks that have known labels or answers. The golden tasks can be performed before being assigned to any normal tasks, known as qualification testing and the golden tasks can also mixed with the normal tasks, refereed as hidden testing [14].

2.2.1.4 Result Aggregation

This type of research aims to effectively aggregate feedback from different crowdsourcing participants to form high-quality task results. Different approaches have been proposed, such as majority voting and weighted voting, where weights can be based on different things, such as the reliability of workers [15] or their quality. In addition, some studies have proposed methods for iteratively measuring worker quality, such as using the EM algorithm [16]. The EM algorithm consists of two steps, first aggregating worker labels, and then estimating the weight of workers by comparing worker labels and aggregated labels.

2.2.1.5 Task assignment

The task allocation problem assigns the most suitable crowdsourcing participants for each task. Due to the unknown nature of workers, research has mainly focused on developing algorithms to infer their reliability and match the most suitable tasks. The following is a survey of the main contributions.

Initially, worker's reliability was not investigated with multiple assignments strategy where one task was assigned to many workers to generate aggregated results for higher accuracy. However, to improve efficiency, worker's reliability was inferred from comparing their answers with others or their previous work. With the development of technology, people have conducted research on more complex application environments, including real-time dynamic task allocation, collaborative task allocation, and AI task allocation.

2.2.2 Cost Control

Finding ways to keep costs under control without compromising the quality of the results is one of the major hurdles in CDM as the expenses accumulate quickly when dealing with a large amount of work. Simplest cost management is to fix the number of workers per task based on the requester budget [9]. Reducing the number of tasks performed by workers, is another cost control method presented as follows.

2.2.2.1 Task Pruning and Task selection

In order to decrease the task volume, studies recommend prioritising essential tasks and eliminating unnecessary ones. Researches have proposed several ways to eliminate the number of comparisons for entity resolution (ER) problems. For instance, evaluating the similarity of entities and cluster them [17], or identify optimal transitive relations to reduce comparisons. Tasks can also be chosen by identifying the most advantageous ones for workers. For instance, active learning is used for labeling tasks, allowing to label much larger datasets at lower costs [18].

2.2.2.2 Sampling and Answer Deduction

Some paper propose cost control methods inferring the task results from previous tasks's results. For example, OASIS infer the label of tasks using the labels collected for a subset of tasks [19]. Task sampling is a cost control method where a small portion of tasks is executed to represent the entire dataset. A sampling and cleaning framework can reduce the impact of dirty data on aggregated query results by cleansing small subsets of data [20].

2.2.3 Latency Control

Controlling latency is crucial under specific time constraint. Reasons leading to delays in crowdsourcing may involve employees having difficulty completing tasks and not being interested in completing them. Therefore, if there are time constraints, managing latency becomes crucial. To address this issue, different methods have been proposed, including adjusting prices, modeling delays, and reducing worker recruitment time.

2.2.3.1 Price Setting

The price of a task is often closely related to its completion time. Increasing the price can attract a larger pool of workers, and lead to a reduction in latency. More sophisticated pricing approach including adjust task price based on the deadline and monetary. Recently, a more advanced methods is proposed which analyzed various sources of latency and each were minimized through different approaches [21]. After researching various bonus schemes, it was found that milestone bonus schemes improves worker's concentration and reaction time most [22].

2.2.3.2 Latency Modelling

Some papers introduce the use of modelling to reduce latency, where statistical and round models are introduced [23]. CrowdSearch builds a dynamic model based on deadlines and observed validation results to describe latency, accuracy, and cost behaviour [24]. Tasks can be performed in parallel to reduce latency, and the round model breaks down the task into smaller sub-tasks. A dynamic budget allocation algorithm with polynomial time is introduced to minimize latency when formulating the problem each round [25].

2.2.3.3 Reduce Recruitment Time

The recruitment time is a mainly source of latency, which refers to the time from a task is posted on crowdsourcing platform until it was accepted by a worker, some papers focusing on minimising it to control latency. A group of workers can be recruited in advance so that they can be readily available when new tasks arise. A recruitment model is built that notifies crowdworkers available on the crowdsourcing platform when a new task arrives, and rewards a small amount of money for accepting the task [26].

2.2.4 Incentive Design

Incentive design is crucial in in CDM as workers are primarily driven by incentives for doing a job and individual may no longer be interested in doing certain tasks with the absence of incentives. Incentives are classified into monetary and non-monetary, and below is a detailed description of both.

2.2.4.1 Monetary Incentives

Monetary incentives are financial benefits given to individuals as payment for completing certain tasks or jobs. Financial benefit is widely used and the easiest way to motivate and reward workers. The reward amount usually varies according to the worker's performance, completion time, work difficulty, as well as higher completion rates [27].

2.2.4.2 Non-Monetary Incentives

In fact, a growing number of crowdsourcing projects are shifting towards non-monetary, as more and more participants are seeking some intrinsic and extrinsic incentives rather than material benefits. These motivations can be categorised as individual development, public good, reputation, and happiness from game.

Individual Development. Crowdsourcing provides opportunities for personal skill development as individuals can work collectively. Knowledge-sharing websites including Wikipedia and Stack Overflow provide collective ideation platforms where individuals could not only develop their skills but also learn from experts with different backgrounds [28].

Public Good. Some individuals participate in crowdsourcing projects for the greater benefit of the community as a whole. To incentive this motivation, the public transportation crowdsourcing system emphasised the improved accuracy of bus arrival and travel times for the entire community, utilising the crowdsourced people's travelling data [29].

Reputation. Crowdsourcing can also enhance an individual's reputation. People may establish themselves as reliable and significant contributors by taking part in projects and showcasing their abilities and expertise, which may result in appreciation and respect from others [29]. This incentive design can be applied in the context where people aiming to establish reputation in a certain industry.

Gamification. Crowdsourcing projects have started to deploy game-like concepts such as badges, rank systems, and news tickers to make the work more engaging, enjoyable, and entertaining for the workers [30]. These elements provide the contributors a sense of accomplishment and competitiveness, which motivates them to work longer and harder on the project.

3 Comparison

Among the CDM optimisation strategies, comparative analysis was conducted to evaluate the pros and cons on quality control, cost control, latency control, and incentive design. It should be noted that the applications of CDM were not included in the comparative analysis as each application serve a different purpose. Therefore, direct comparison of the pros and cons of applications is not meaningful. However, it is worth noting that CDM has unique advantages and limitations on the specific use.

3.1 Pros and Cons of Quality Control Approaches

For quality control approaches as demonstrated in Table 2, the main advantages lies in easy implementation and helping identify worker's reliability. However, the main disadvantages include higher cost due to more labor force and bias in ground truth during evaluation process, as there may be situations where ground truth is not limited to one answer. Some approach are limited in scope as it can only be applied in certain situations.

Additionly, for task assignment techniques as demonstrated in Table 3, the main advantages include easy implementations, leveraged resource usage, and flexibility. On the other hand, as most algorithms

Table 2: Pros and Cons for Quality Control Techniques

Techniques	**Techniques**	**Pros**
Cons		
Result Aggregation [31]	• Produce more accurate and reliable overall results. • Assist identifying malicious or low-quality workers.	• Requires more workers than other methods. • Higher cost.
Majority Voting [14]	• Robust to noise. • Easy to implement.	• The professional level and of different workers and task difficulty is ignored.
Confusion Matrix [32]	• Capture more information compared to using a single value to model workers	• Biases in ground truth. • Only used when tasks have a fixed optional label set, so not applicable in all situations.
Golden Task [14] [33]	• Simplify the process of assessing worker quality.	• Higher cost from hiring experts to label golden tasks. • Difficulty in deciding sufficient number of golden tasks to reveal workers' domain knowledge. • If the answer is leaked, or many requesters use the same golden task, the mechanism will fail.
Qualification Tests [34]	• Improves confidence of result's quality from reliable workers.	• Many workers unwilling to answer "extra" tasks for free. • Poses potential risk that spammers could carefully label these golden tasks to increase their reputation.

infers worker's reliability on historical data, disadvantages mainly include difficulty in implementation when there is lack of historical data.

3.2 Pros and Cons of Cost Control Approaches

Without prior knowledge of the quality of online workers, fixing the number of workers can easily control the cost but may lead to wasted expenses, and may lead to low quality of task results because the quality of online workers is not clear. Different methods of reducing the number of tasks performed by workers have different characteristics. Although pruning technology can effectively control costs and significantly save labor costs while maintaining high quality [17]. However, task pruning only considers the relationships between tasks, so it cannot reduce costs at each task level. In addition, many pruning techniques are limited to certain types of tasks. Another method to reduce tasks is the task selection method. Although it can reduce costs, it may sacrifice some quality and introduce some delay, as it requires the use of iterative queries to determine which tasks can be selected next [23]. The use of deduction and sampling techniques can help avoid crowds from doing many unnecessary tasks. However, the drawbacks are obvious, as the derivation of answers will increase human error, and sampling techniques will fail in many cases, such as finding the largest number in the dataset.

3.3 Pros and Cons of Latency Control Approaches

Although increasing task prices can easily reduce delays, it will greatly increase costs. Dynamic budget allocation adjusts prices based on real-time factors to ensure effective completion of tasks, but

Table 3: Pros and Cons for Task Assignment Techniques

Techniques	Pros	Cons
Multiple assignment	• No need to know or infer worker's reliability. • Easy to implement.	• Not efficient: workers didn't get potentially acquainted tasks. • Increasing cost.
Iterative Learning	• Worker's reliability estimated based on comparing with other's answers. • Leveraged resource usage.	• High computational cost.
Dual Task Assigner	• Worker's reliability can be estimated from their previous performance. • Leveraged resource usage.	• Difficult to infer worker's reliability when there's no enough historical data.
Real-time Task Assigner	• Reduced cost on multiple assignments. • Flexible with dynamic task assignments.	• Only near-optimal results can be achieved.
Collaborative Task Assignment	• Worker's reliability is known. • More flexible task assignment with worker's domain knowledge. • Can be applied in Knowledge-Intensive and collaborative crowdsourcing settings.	• Application is narrow in real context, as most times worker's skill set is unknown.
Task Assignment with AI Planning	• Leveraged task standardisation process • Enabled testing of task allocation strategies with different scenario variables	• Did not perform optimisation under consider budget.

if allocation standards are not transparent or communication is unclear, this may cause confusion or dissatisfaction among workers. Delayed modeling can generate more accurate predictions and more objective decisions. However, implementing a delay model may be a complex process that requires a large amount of data analysis and modeling expertise, and it may not be flexible enough to adapt to changes in the crowdsourcing environment. The hiring model used to reduce recruitment time can accept tasks faster, but it may bring low-quality results as random workers are assigned to tasks. In addition, this will introduce more costs as the model will pay for workers who receive tasks.

3.4 Pros and Cons of Incentive Design Approaches

For incentive design methods, the benefits include motivating individual participation, improving accuracy and speed. In addition, non monetary incentives will inherently motivate individuals in terms of reputation and personal development. However, the limitations of incentive design methods may manifest as excessive or insufficient payments, both of which can lead to problems. The pros and cons for incentive design are shown in Table 4.

4 Discussion and Future Direction

The general trend of crowdsourced data management is focused on its application and techniques in optimising management efficiency. As discussed in *Overview*, the application of CDM involves micro-tasking, idea generation, public participation, and technology development. Therefore, major challenges involve managing uncertainty and bias introduced by the crowd.

Table 4: Pros and Cons for Incentive Design

Incentive Types	Pros	Cons
Monetary	• Straightforward way to motivate individuals to participate. • Compensation for contributions. • Can attract a consistent and steady involvement of participants. • Better accuracy and speed compared to voluntary work.	• Pay rate should be carefully considered to match time and effort required. • Overpaying or underpaying can lead to issues. • Participants may attempt the project multiple times, leading to poor data quality. • Not feasible for project starters with no or low budget.
Non-monetary	• Cost-effective. • Increase motivation. • Provide better data quality. • Can offer opportunities for personal skill development. • Can contribute to public good. • Enhance an individual's reputation	• Fewer people may be willing to participate. • Not applicable for all workers. • More time and effort required for task design. • Different people value different types of incentive.

Uncertainty remains a major challenge in crowdsourced data management, as worker reliability, expertise, and motivation can vary and lead to unqualified responses. Despite the development of optimisation techniques for quality control, latency control, and incentive design, there is no universal solution due to the diverse contexts in which crowdsourcing is applied. Nevertheless, advances in CDM research allow for the inference of worker expertise and flexibility in combining existing management techniques based on budget and requirements.

Additionally, as CDM involves human participation, managing bias in various processes is crucial. Since humans are inherently biased due to diverse backgrounds, bias may be introduce by different perceptions during quality control process. This includes different perceptions of "fair division of tasks" in task assignment and bias in the "ground truth" during the expert evaluation phase since ground truth may not be limited to a single answer. Additionally, individual perceptions of incentives can also introduce bias, as some may prefer monetary incentives while others are motivated by non-monetary incentives. Thus, managing bias is a crucial topic in CDM.

The ethical issue of crowdsourcing remains a topic of controversy in many fields. Ethical concerns involves the exploitation of workers with low wages. For instance, to guarantee ChatGPT speak politely, OpenAI used crowdsourcing for data filter and data labelling with Kenyan workers. However, workers were underpaid with less than $2 per hour [35]. Furthermore, crowdsourcing can result in the violation of privacy or the misuse of personal data. While website cookies are commonly used to collect personal data, there is still a need for greater transparency and accountability in this area.

5 Conclusion

In conclusion, The main reason for utilising CDM is to apply the crowd's intelligence and labor to perform computational-difficult tasks. This process involves the requester sending queries to crowdsourcing platforms for workers to view and respond. Platforms can be classified to general-purpose and specialised. The variability of the crowd introduces uncertainty, leading to inconsistent reliability, expertise, and motivation among workers, which can consequently yield inadequate answers. Additionally, managing budget and time constraints is critical due to the large number of people involved in the crowdsourcing process.

Therefore, CDM techniques are constantly developed in terms of cost, quality, and latency control strategies and incentive designs to assist requester to overcome the challenges. Quality control techniques involve preprocessing input, selecting reliable workers, task assignment optimisation, and result aggregation. Latency control techniques include worker motivation and performance modeling. Cost control techniques focus on task simplification through filtering and redundancy avoidance, as well as inferring answers. To motivate workers, both monetary and non-monetary incentives can be utilised.

As this survey reviewed and compared the major techniques and discovered that the trade-offs are inherent in the application. The techniques were analysed for their advantages and disadvantages in terms of implementation, efficiency, cost, flexibility, universality, and bias. Furthermore, there are trade-offs between management fields, as achieving cost, latency, and quality goals simultaneously may not be feasible.

References

[1] V. Crescenzi, A. A. A. Fernandes, P. Merialdo, and N. W. Paton, "Crowdsourcing for data management," *Knowledge and Information Systems*, vol. 53, no. 1, pp. 1–41, Oct. 2017.

[2] H. Jeff, "The Rise of Crowdsourcing," no. 14.06, Jun. 2006.

[3] C. Chai, J. Fan, G. Li, J. Wang, and Y. Zheng, "Crowdsourcing Database Systems: Overview and Challenges," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. Macao, Macao: IEEE, Apr. 2019, pp. 2052–2055.

[4] J. Fan, M. Zhang, S. Kok, M. Lu, and B. C. Ooi, "CrowdOp: Query optimization for declarative crowdsourcing systems," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. Helsinki, Finland: IEEE, May 2016, pp. 1546–1547.

[5] "What is Kaggle, Why I Participate, What is the Impact? | Data Science and Machine Learning," https://www.kaggle.com/getting-started/a.

[6] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye, "KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne Victoria Australia: ACM, May 2015, pp. 1247–1261. [Online]. Available: https://dl.acm.org/doi/10.1145/2723372.2749431

[7] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Citeseer, 2012.

[8] W. Lee, C.-W. Chang, P.-A. Yang, C.-H. Huang, M.-K. Wu, C.-C. Hsieh, and K.-T. Chuang, "Effective Quality Assurance for Data Labels through Crowdsourcing and Domain Expert Collaboration," 2018. [Online]. Available: https://openproceedings.org/2018/conf/edbt/paper-243.pdf

[9] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney NSW Australia: ACM, Aug. 2015, pp. 745–754.

[10] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *Journal of Machine Learning Research*, vol. 13, no. null, pp. 491–518, Feb. 2012.

[11] Z. Zhao, D. Yan, W. Ng, and S. Gao, "A transfer learning based framework of crowd-selection on twitter," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago Illinois USA: ACM, Aug. 2013, pp. 1514–1517.

[12] A. Marcus, D. Karger, S. Madden, R. Miller, and S. Oh, "Counting with the crowd," *Proceedings of the VLDB Endowment*, vol. 6, no. 2, pp. 109–120, Dec. 2012.

[13] F. L. Wauthier and M. Jordan, "Bayesian bias mitigation for crowdsourcing," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.

[14] Y. Zheng, G. Li, and R. Cheng, "DOCS: A domain-aware crowdsourcing system using knowledge bases," *Proceedings of the VLDB Endowment*, vol. 10, no. 4, pp. 361–372, Nov. 2016.

[15] C.-J. Ho, S. Jabbari, and J. W. Vaughan, "Adaptive task assignment for crowdsourced classification," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28. Atlanta, Georgia, USA: PMLR, Jun. 2013, pp. 534–542.

[16] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*. Washington DC: ACM, Jul. 2010, pp. 64–67.

[17] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "CrowdER: Crowdsourcing entity resolution," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1483–1494, Jul. 2012.

[18] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden, "Scaling up crowd-sourcing to very large datasets: A case for active learning," *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 125–136, Oct. 2014.

[19] Y. Amsterdamer, S. B. Davidson, T. Milo, S. Novgorodov, and A. Somech, "OASSIS: Query driven crowd mining," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. Snowbird Utah USA: ACM, Jun. 2014, pp. 589–600.

[20] J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraska, and T. Milo, "A sample-and-clean framework for fast and accurate query processing on dirty data," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. Snowbird Utah USA: ACM, Jun. 2014, pp. 469–480.

[21] D. Haas, J. Wang, E. Wu, and M. J. Franklin, "CLAMShell: Speeding up crowds for low-latency data labeling," *Proceedings of the VLDB Endowment*, vol. 9, no. 4, pp. 372–383, Dec. 2015.

[22] G. Demartini, P. Cudré-Mauroux, D. E. Difallah, and M. Catasta, "Scaling-up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement," 2004.

[23] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced Data Management: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2296–2319, Sep. 2016.

[24] T. Yan, V. Kumar, and D. Ganesan, "CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*. San Francisco California USA: ACM, Jun. 2010, pp. 77–90.

[25] V. Verroios, P. Lofgren, and H. Garcia-Molina, "tDP: An Optimal-Latency Budget Allocation Strategy for Crowdsourced MAXIMUM Operations," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne Victoria Australia: ACM, May 2015, pp. 1047–1062.

[26] M. S. Bernstein, J. Brandt, R. C. Miller, and D. R. Karger, "Crowds in two seconds: Enabling realtime crowd-powered interfaces," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. Santa Barbara California USA: ACM, Oct. 2011, pp. 33–42.

[27] L. Litman, J. Robinson, and T. Abberbock, "TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences," *Behavior Research Methods*, vol. 49, no. 2, pp. 433–442, Apr. 2017. [Online]. Available: http://link.springer.com/10.3758/s13428-016-0727-z

[28] A. I. Chittilappilly, L. Chen, and S. Amer-Yahia, "A Survey of General-Purpose Crowdsourcing Techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2246–2266, Sep. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7456302/

[29] A. Misra, A. Gooze, K. Watkins, M. Asad, and C. A. Le Dantec, "Crowdsourcing and Its Application to Transportation Data Collection and Management," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2414, no. 1, pp. 1–8, Jan. 2014. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2414-01

[30] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. Xi'an, China: IEEE, Sep. 2015, pp. 891–897. [Online]. Available: http://ieeexplore.ieee.org/document/7344680/

[31] S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann, "Shepherding the crowd yields better work," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. Seattle Washington USA: ACM, Feb. 2012, pp. 1013–1022.

[32] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd International Conference on World Wide Web*. Seoul Korea: ACM, Apr. 2014, pp. 155–164.

[33] D. Yuan, G. Li, Q. Li, and Y. Zheng, "Sybil Defense in Crowdsourcing Platforms," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Singapore Singapore: ACM, Nov. 2017, pp. 1529–1538.

[34] J. Fan, G. Li, B. C. Ooi, K.-l. Tan, and J. Feng, "iCrowd: An Adaptive Crowdsourcing Framework," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. Melbourne Victoria Australia: ACM, May 2015, pp. 1015–1030.

[35] B. Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," *TIME*, Jan. 2023.