Crowdsourced Data Management Survey

Group 13
The University of Melbourne

Overview

1. Introduction
   a. Our Team
   b. Our Project Goal
   c. Our Project Timeline
2. Approach
   a. Taxonomy
   b. Literature Review
      i. Application & Platforms
      ii. Techniques
3. Conclusion
   a. Discussion
   b. Future Directions

Team & Project Goal

Project Goal:

1. Compose a survey to help the general public understand Crowdsourced Data Management。
2. Illustrate the survey results through presentation.

Communication & Research Tools

Manage weekly meetings & tasks

Confluence

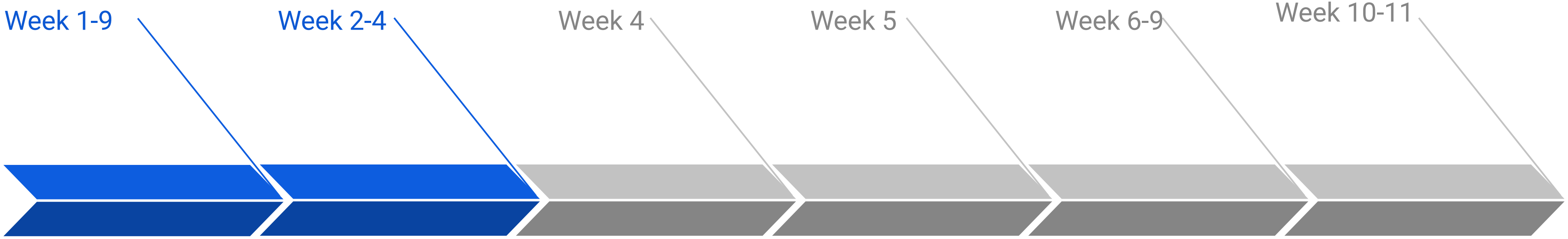
Manage report and reference

Overleaf

zotero

Project Timeline

Approx.10 hours per week



Week 1-9

Literature Review

General literature review to understand the 'big picture'

Week 2-4

Building Taxonomy

More specific literature review under classified topics

Week 4

Storytelling Strategy Development

Learn how to write storytelling surveys through literature review and academic workshops

Week 5

Assign Work

Grace: Architecture of CDM; John: Application of CDM; Shirley & Chris: Techniques of CDM

Week 6-9

Survey Composing

Use zotero and overleaf to compose academic survey. Peer review for each other's work.

Week 10-11

Presentation Preparation

Preparing slides and rehearse.

What is Crowdsourced Data Management? (Ellis, 2004; Howe, 2006)

1714 Longitude Prize

1936 Toyota Logo Design

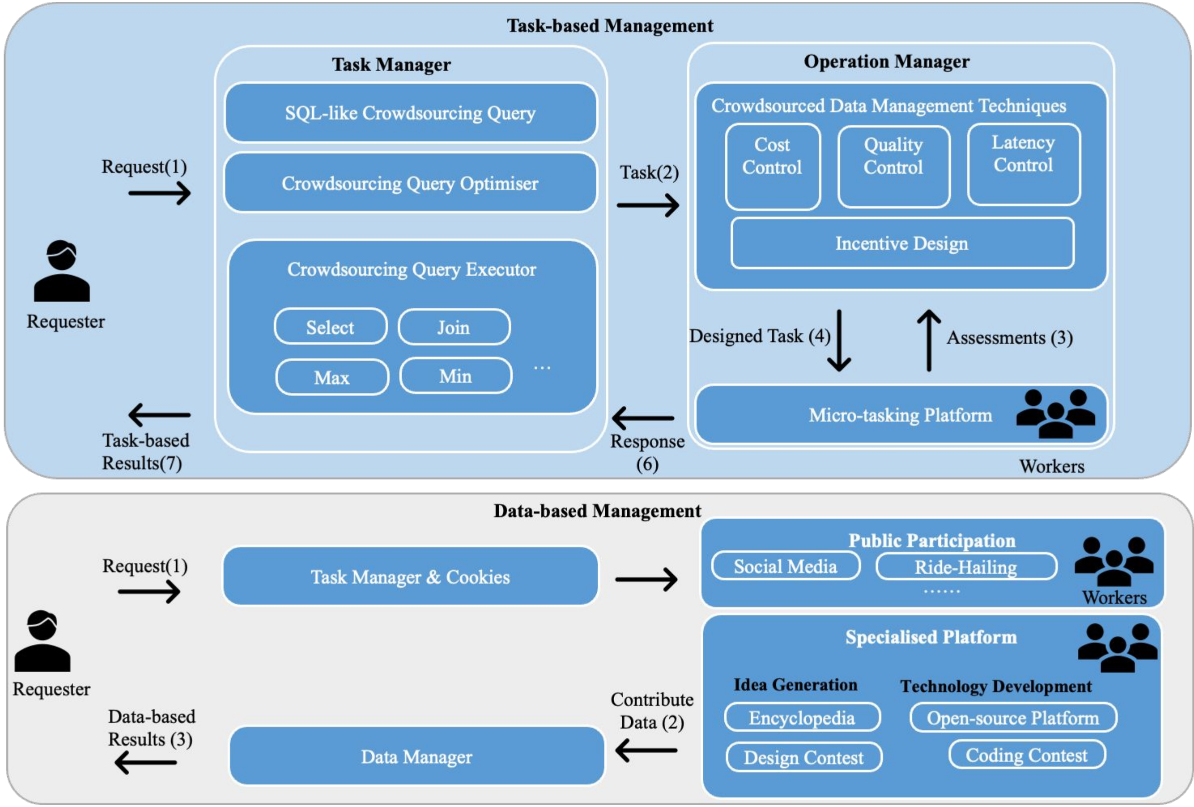
2005 Launch of Amazon Mechanical Turk

2006 Conceptualisation of “Crowdsourcing”

2000s Crowdsourcing Applications & Technology Explosion

2010s Crowdsourced Data Management Applications & Techniques

2020s Nowadays Application



Task-based Management

Requester

Request(1)

Task Manager

SQL-like Crowdsourcing Query

Crowdsourcing Query Optimiser

Crowdsourcing Query Executor

Select, Join, Max, Min …

Task(2)

Operation Manager

Crowdsourced Data Management Techniques

Cost Control

Quality Control

Latency Control

Incentive Design

Designed Task (4)

Assessments (3)

Micro-tasking Platform

Workers

Response (6)

Task-based Results(7)

Data-based Management

Requester

Request(1)

Task Manager & Cookies

Public Participation

Social Media

Ride-Hailing

……

Workers

Specialised Platform

Idea Generation

Encyclopedia

Design Contest

Technology Development

Open-source Platform

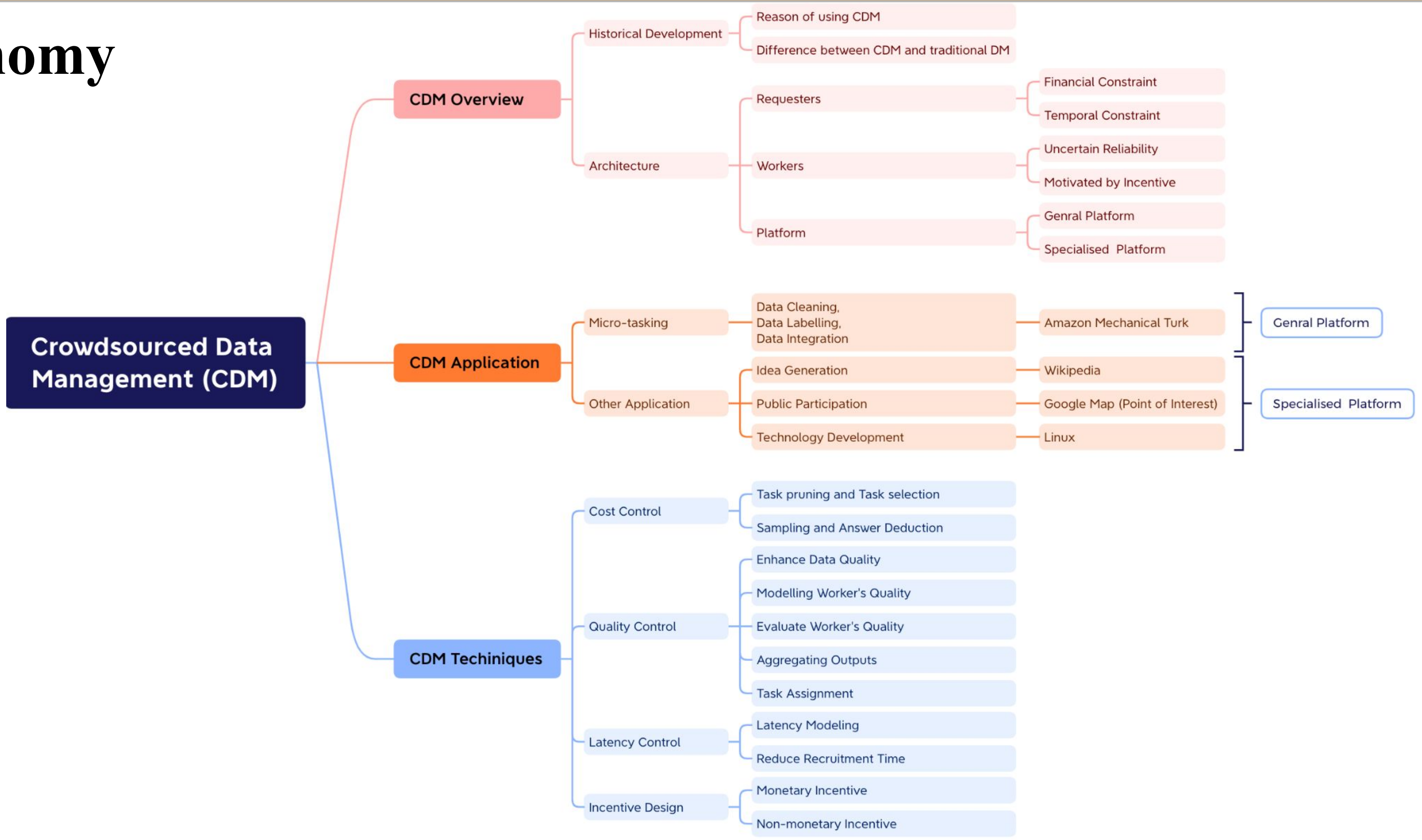
Coding Contest

Contribute Data (2)

Data Manager

Data-based Results (3)

Taxonomy



- Crowdsourced Data Management (CDM)
  - CDM Overview
    - Historical Development
      - Reason of using CDM
      - Difference between CDM and traditional DM
    - Architecture
      - Requesters
        - Financial Constraint
        - Temporal Constraint
      - Workers
        - Uncertain Reliability
        - Motivated by Incentive
      - Platform
        - Genral Platform
        - Specialised Platform
  - CDM Application
    - Micro-tasking
      - Data Cleaning, Data Labelling, Data Integration
        - Amazon Mechanical Turk
    - Other Application
      - Idea Generation
        - Wikipedia
      - Public Participation
        - Google Map (Point of Interest)
      - Technology Development
        - Linux
  - CDM Techiniques
    - Cost Control
      - Task pruning and Task selection
      - Sampling and Answer Deduction
    - Quality Control
      - Enhance Data Quality
      - Modelling Worker's Quality
      - Evaluate Worker's Quality
      - Aggregating Outputs
      - Task Assignment
    - Latency Control
      - Latency Modeling
      - Reduce Recruitment Time
    - Incentive Design
      - Monetary Incentive
      - Non-monetary Incentive

Platform (Chittilappilly et al., 2016)

General-purpose platform

- Amazon Mechanical Turk

Specialised platform

- Kaggle
- Lego Ideas

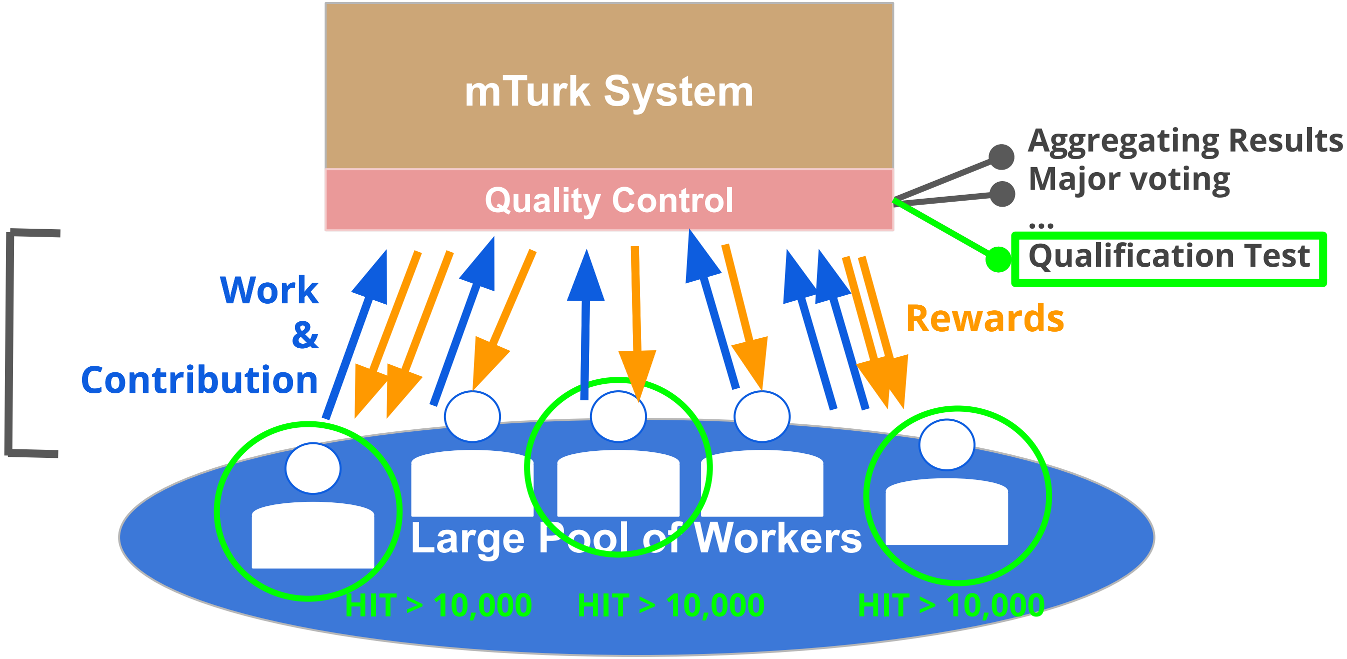
General-purpose Platform

Amazon Mechanical Turk

- Requester
- Worker
- Developer

Cost Control
- Multiple assignments
- Iterative Learning
- …

Latency Control
- Monetary Rewards



General-purpose Platform

amazon mechanical turk

Microtasks in AMT

- Data Labelling
- Data Cleaning
- Data integration

Bounding Box

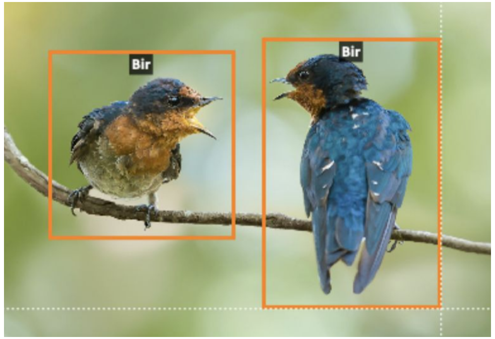
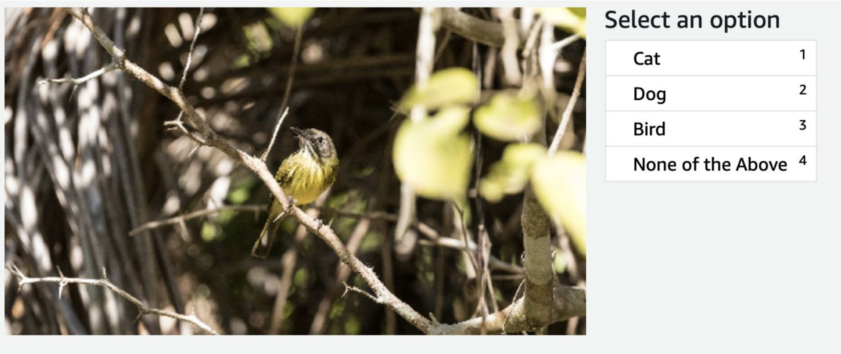
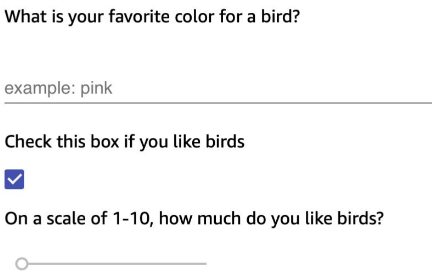


Image Labelling


Select an option

Cat	1
Dog	2
Bird	3
None of the Above	4

Survey


What is your favorite color for a bird?

example: pink

Check this box if you like birds

☑

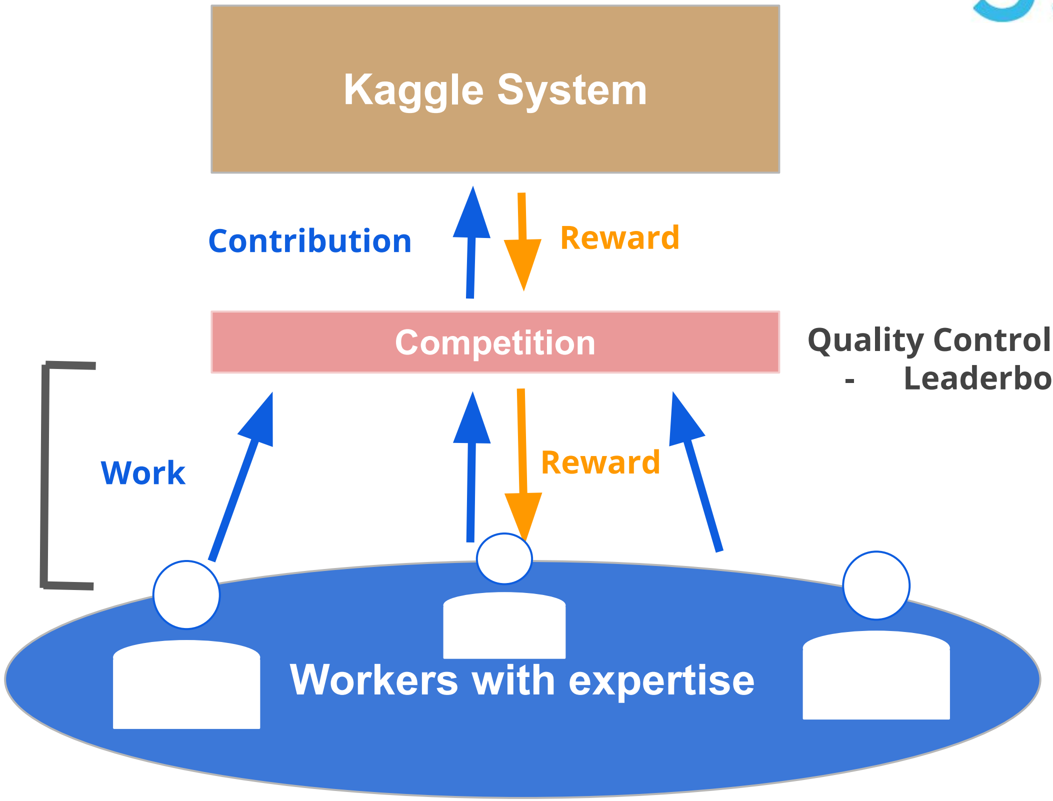
On a scale of 1-10, how much do you like birds?

Specialised Platform

kaggle

Kaggle

- Sponsors
- Workers



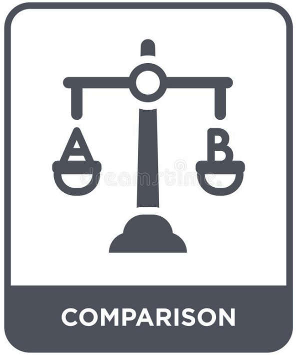
Quality Control
- Leaderboard

Cost Control
- Multiple assignments

Latency Control
- Monetary Rewards

Quality Control

- Data quality may be influenced by several factors
- Various approaches have been proposed to guarantee high quality results
    - Enhancing data quality
    - Selection of workers
    - Aggregating outputs of different workers
    - Assigning suitable tasks to high quality workers



Quality Control	Pros	Cons
Aggregating Result	● Produce more accurate and reliable overall results ● Assist identifying malicious or low-quality workers	● Requires more workers than other methods ● Higher cost
Majority Voting	● Easy to implement ● Robust to noise	The professional level of different workers and task difficulty is ignored
Confusion Matrix	Capture more information	● Not applicable in all situations ● Biases in ground truth.
Golden Task	Simplify the process of assessing worker quality.	● Higher cost from hiring experts ● Difficulty in deciding sufficient number of golden tasks ● If the answer is leaked, or many requesters use the same golden task, the mechanism will fail
Qualification Task	Improves confidence of result's quality from reliable workers	● Many workers are unwilling to answer “extra” tasks for free. ● Poses potential risk that spammers could carefully label these golden tasks to increase their reputation.

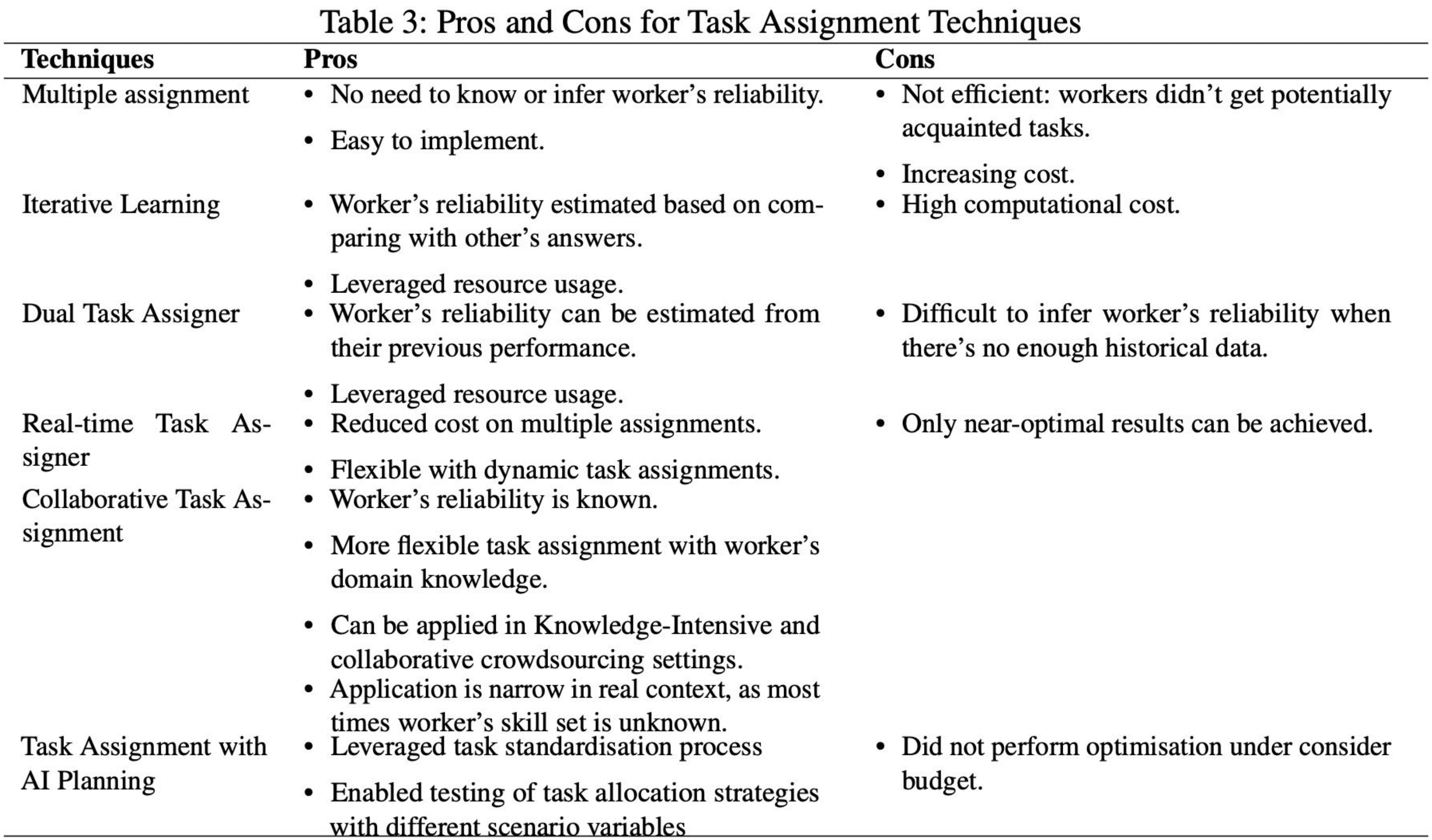


Table 3: Pros and Cons for Task Assignment Techniques

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.

Cost Control

- Keeping costs under control without compromising the quality of the results
- Techniques:
  - ❖ Fix the number of workers per task based on the requester budget
  - ❖ Reduce the number of tasks performed by workers
    - Task Pruning
    - Task Selection
    - Sampling
    - Answer Deduction

Comparison

Cost Control	Pros	Cons
Fixing number of workers	Easy to implement	Wasted expenses
Task Pruning	Significantly saves labour costs while maintaining high quality	● Cost cannot be reduced on a per-task level ● Limited to certain types of tasks
Task Selection	Sufficiently reduce the number of tasks	● Sacrifice some quality ● Introduce some delay
Answer Deduction and Sampling	Avoid crowds doing a lot of unnecessary work	● Introduce human error ● Sampling fail under certain situation

Latency Control

- If the requester has a time limit, controlling latency is important
- Several strategies to address this issue:

❖ Adjusting price

❖ Latency modelling

❖ Reducing Recruitment Time

Comparison:

❖ Increasing task price will greatly increase the cost

❖ Dynamic budget allocation may cause confused or dissatisfied among workers

❖ Latency modelling provides more objective decisions but

❖ The retainer model can efficiently reduce latency, but may introduce low quality results and more costs

Incentive

- **Monetary incentives**
  - ❖ Financial Benefits
  - ❖ Straightforward & Easy Implementation

- **Non-monetary incentives**
  - ❖ Individual Development
  - ❖ Public Good
  - ❖ Reputation
  - ❖ Gamification

Comparison

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

Discussion & Future Direction

- Uncertainty
- Bias
- Ethical issue

Reference

Chittilappilly, A. I., Chen, L., & Amer-Yahia, S. (2016). A Survey of General-Purpose Crowdsourcing Techniques. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2246–2266. https://doi.org/10.1109/TKDE.2016.2555805

Ellis, S. (2014). A History of Collaboration, a Future in Crowdsourcing: Positive Impacts of Cooperation on British Librarianship. *Libri*, *64*(1), 1–10. https://doi.org/10.1515/libri-2014-0001

Jeff, H. (2006, June). *The Rise of Crowdsourcing*. *14.06*. https://www.wired.com/2006/06/crowds/