Crowdsourcing

Alexandra Risch

Outline

- **Crowdsourcing**
  - HIT
  - Amazon Mechanical Turk
- **Paper 1**
  - The dynamics of microtask crowdsourcing: The case of amazon mturk
- **Paper 2**
  - In search of quality in crodwsourcing for search engine evaluation

Crowdsourcing

- Used if traditional approaches are inadequate
  - → Size of test collection
  - → Limitations on time
  - → Computers unable to perform task

- HIT: Human Intelligence Task

- Batch: a set of similar HITs published by a requester at a certain point in time

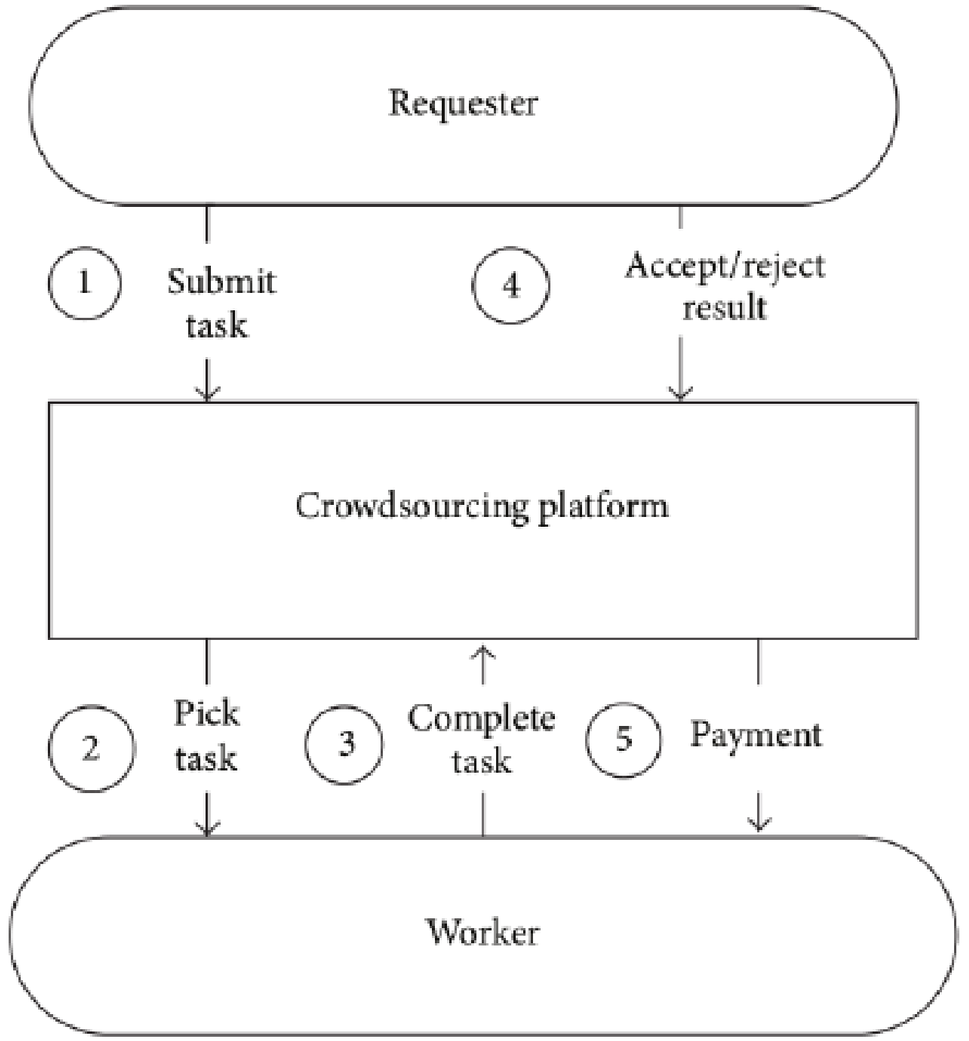
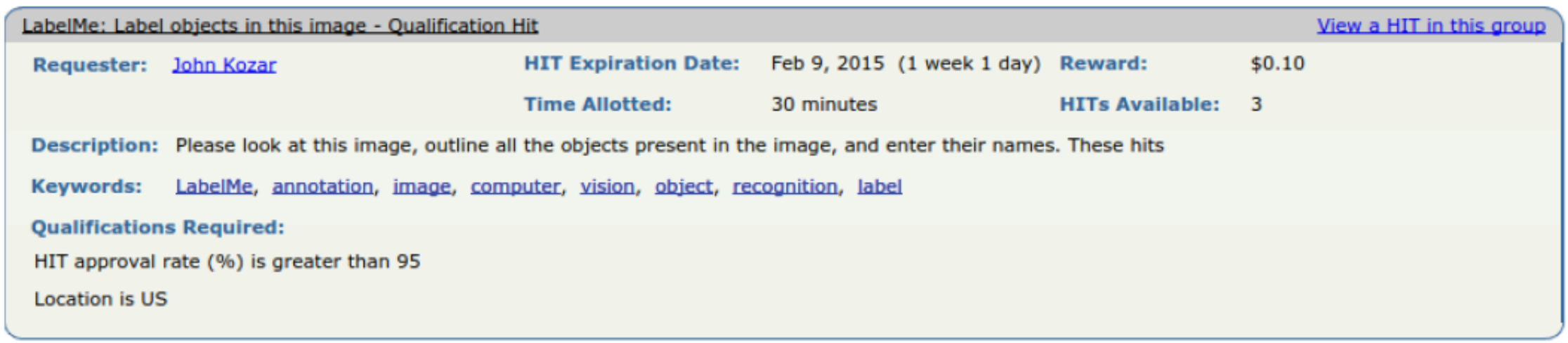


Figure source: http://dx.doi.org/10.1155/2014/135641

Amazon Mechnical Turk (MTurk)[1]

- Very popular crowdsourcing platform
- Launched on November 2, 2005
- More than 100,000 workers (2000 active workers)[2]



[1] https://www.mturk.com/
[2] Difallah, Djellel; Filatova, Elena; Ipeirotis, Panos. Demographics and Dynamics of Mechanical Turk Workers. 2018.

Paper 1

Difallah, Djellel Eddine, et al.

The dynamics of micro-task crowdsourcing:
The case of amazon mturk.
WWW '15.

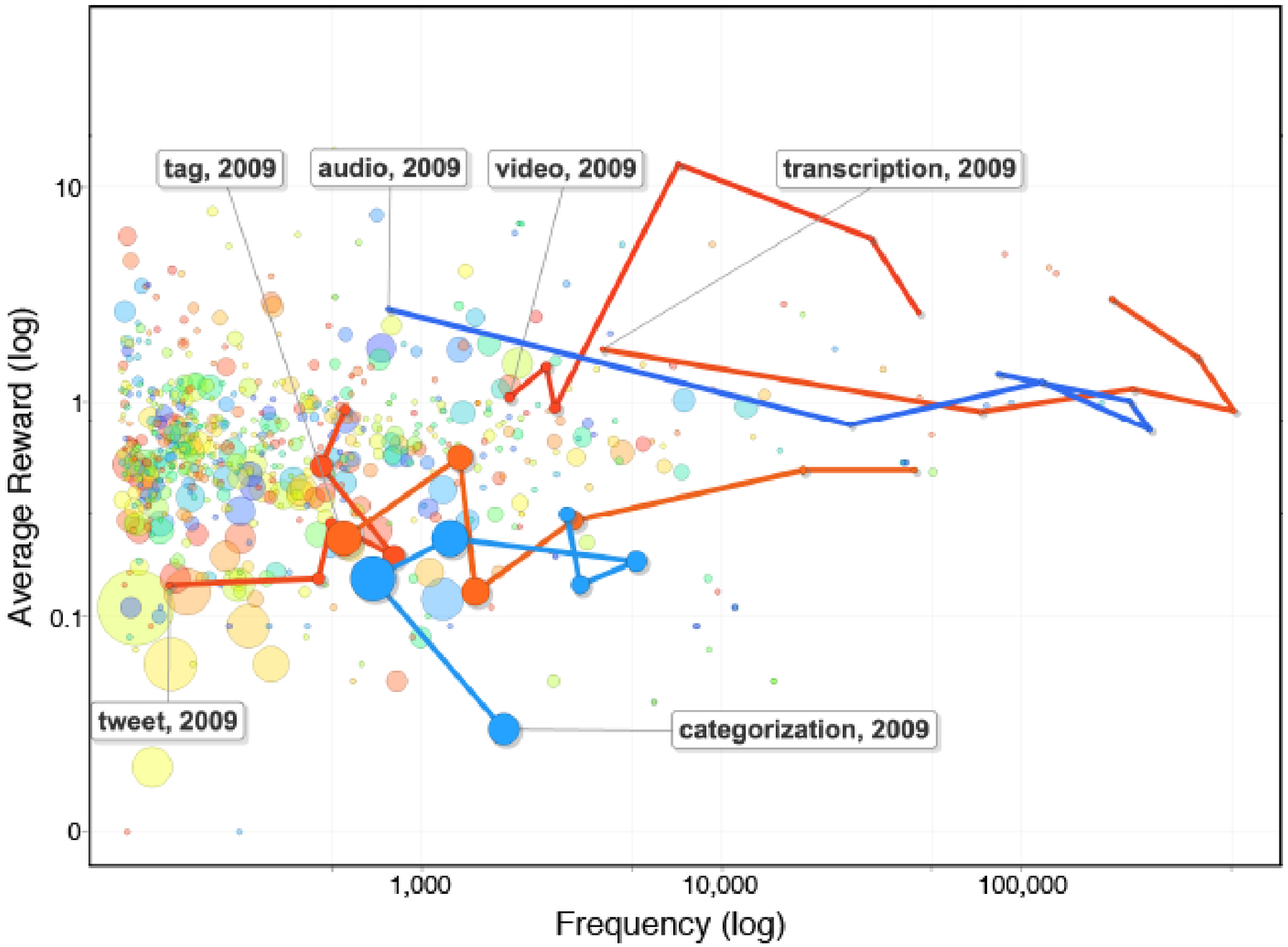
Paper 1

- Long-term analysis of MTurk
  → evolution of workers, requesters, tasks and platform
  → market behavior with regards to demand and supply
  → method to predict throughput

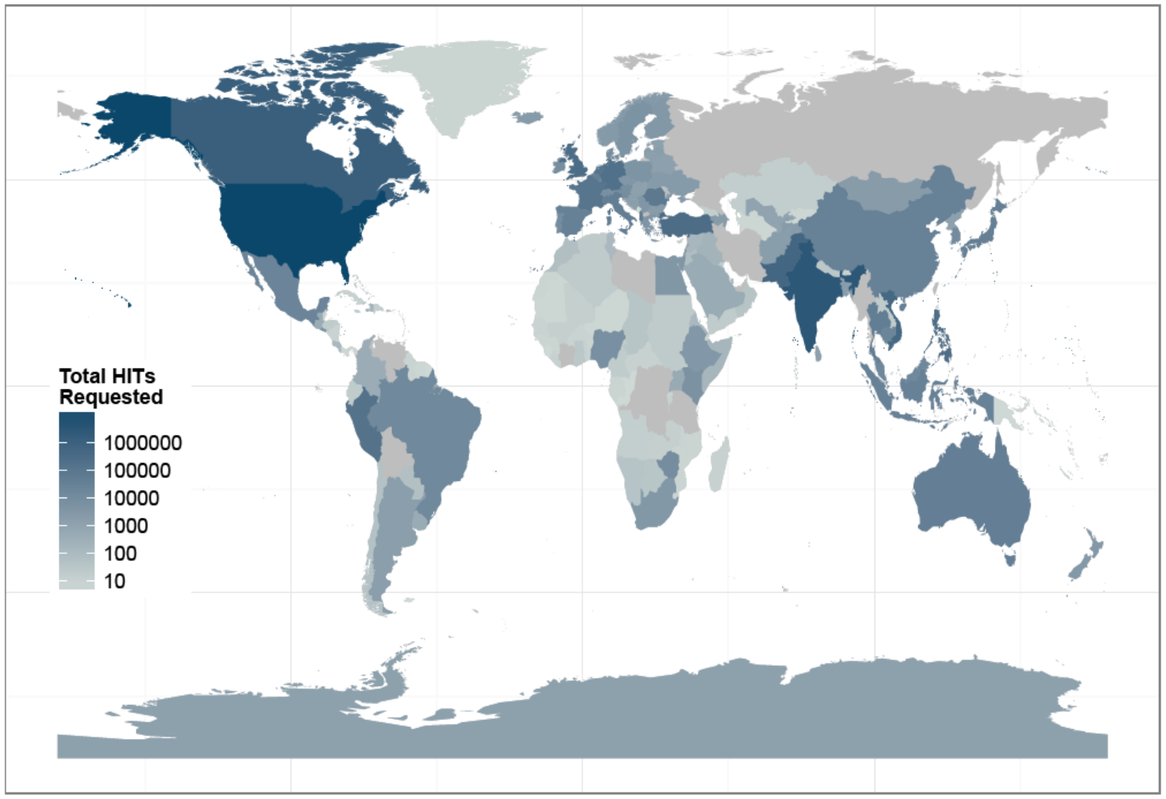
- Dataset[1]
  → periodically collected data about HITs from 2009 to 2014
  → hourly aggregated data about HIT batches and their metadata

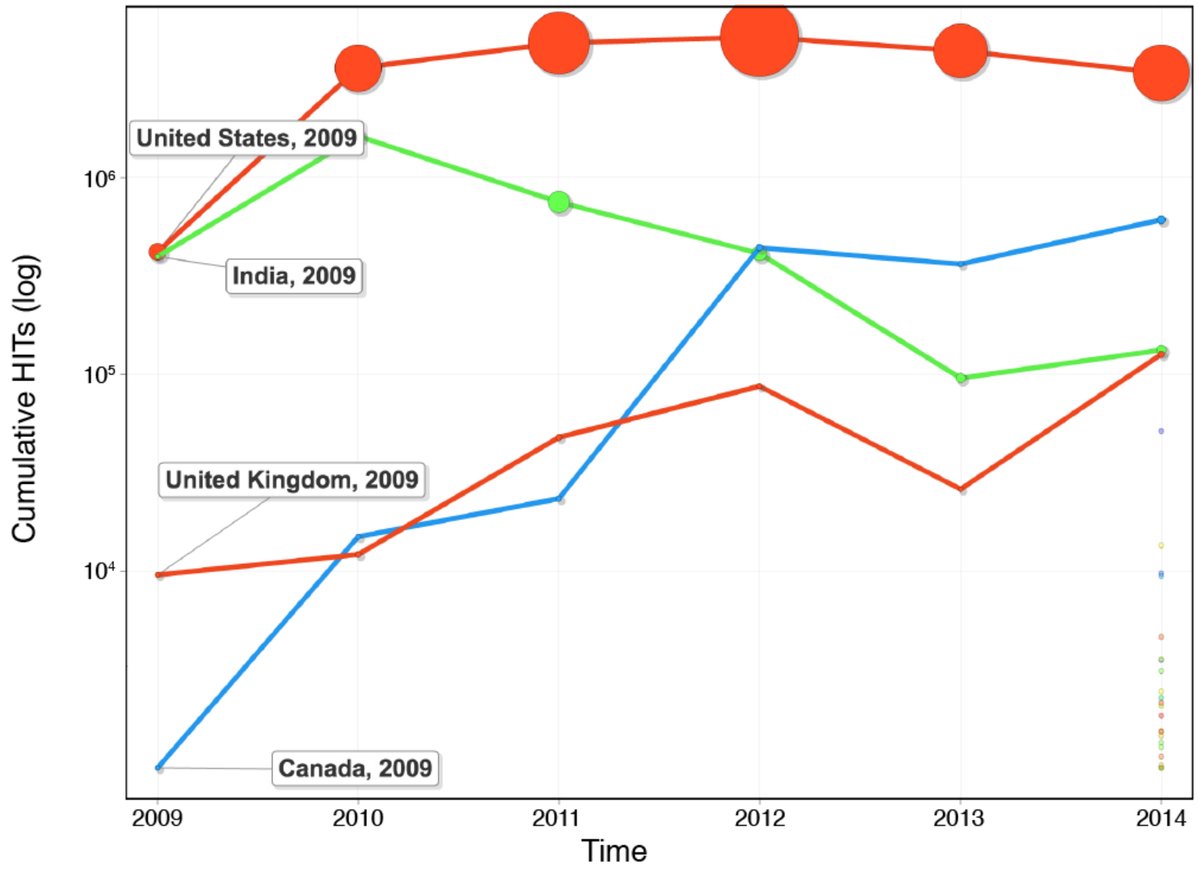
[1]http://www.mturk-tracker.com

Paper 1: Topics over Time



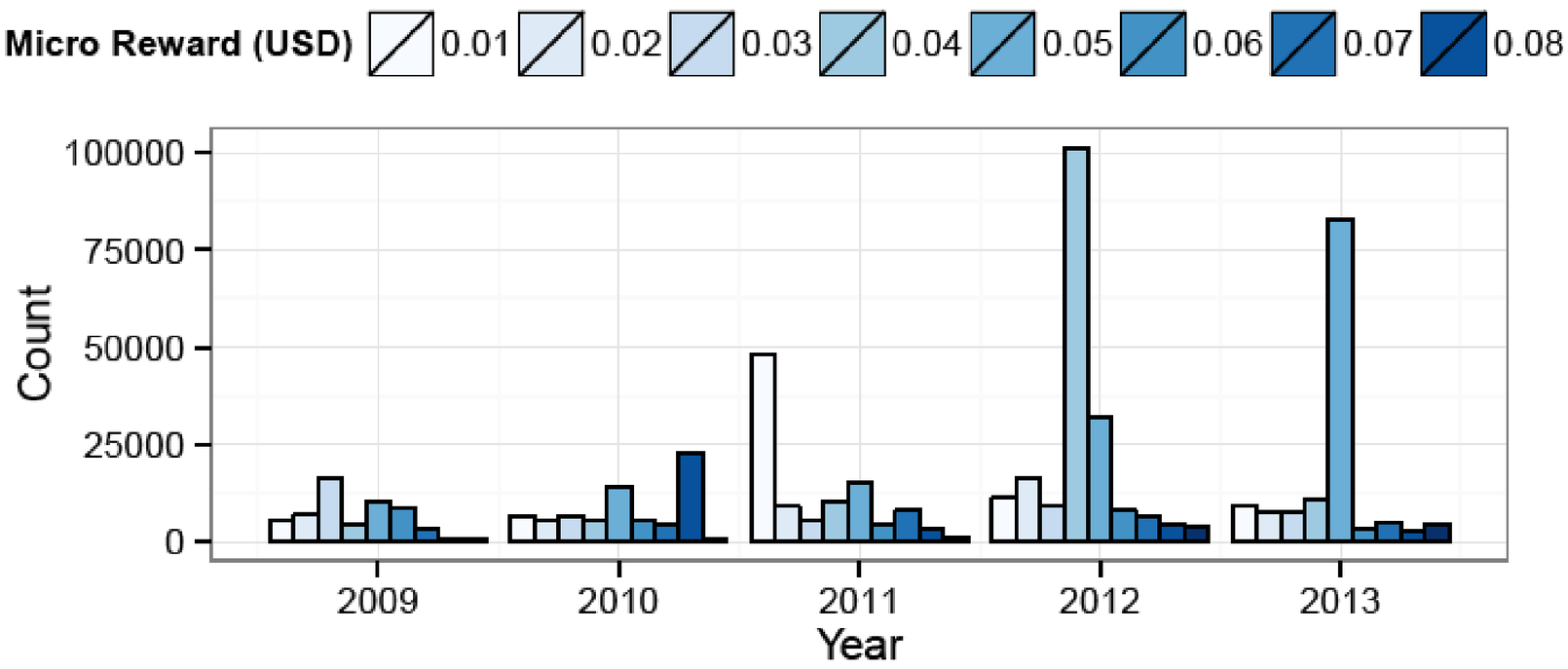
Paper 1: Preferred Countries by Requesters





Paper 1: HIT Rewards

- Higher rewards due to more complex HITs
- Linear increase in the total reward



Paper 1: HIT Batch Size and Requester

- Most batches are small (about 1,000 HITs)
- Average batch size has slightly decreased
- In 2014 very large batches (more than 200,000 HITs) appeared

- Increasing number of active requesters
  → More than 3,000
- Constant number of new requesters
  → About 1,000 per month

Paper 1: HIT Types

- Top-level tasks[1]
  - → Content Access (CA)
  - → Content Creation (CC)
  - → Interpretation and Analysis (IA)
  - → Information Finding (IF)
  - → Surveys (SU)
  - → Verification and Validation (VV)

[1]U. Gadiraju, R. Kawase, and S. Dietze. A taxonomy of microtasks on the web. In Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT '14, pages 218{223, New York, NY, USA, 2014. ACM.

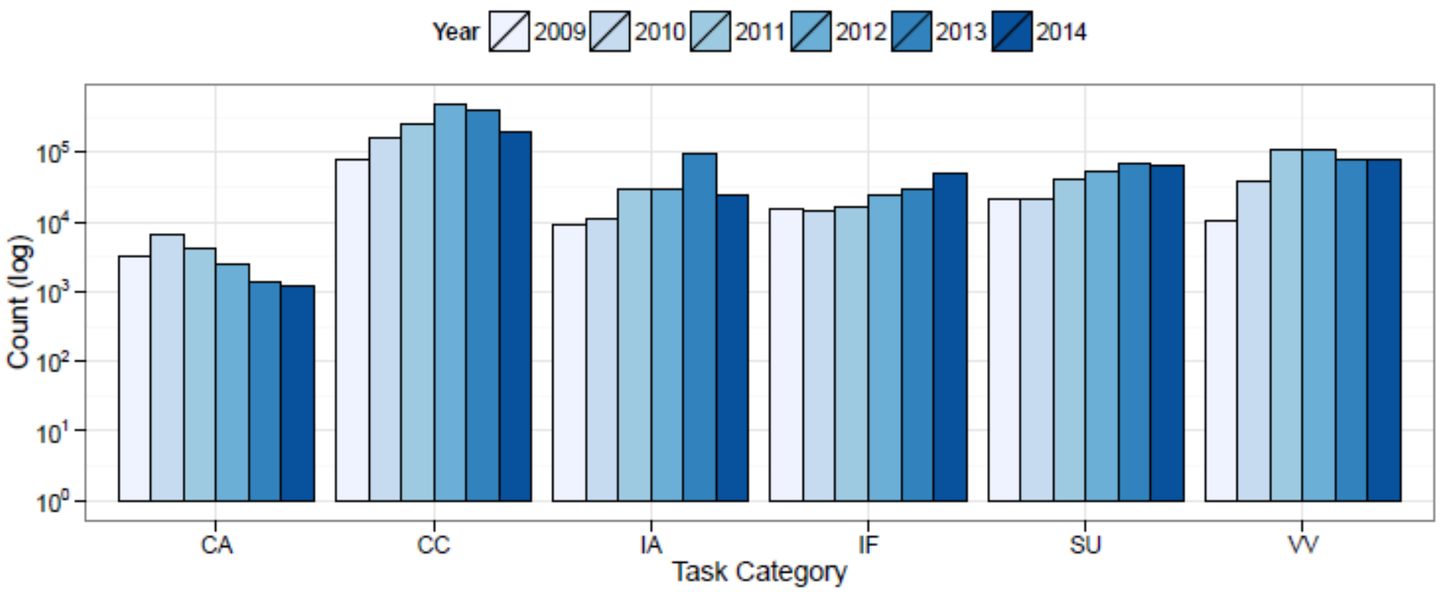
Paper 1: Task Type Popularity

- SVM classifier

  → features: title, description, keywords, reward, date, allotted time, batch size

→ Training data of 5,000 HITs labeled by crowdsourcing

→ Best features based on information gain: allotted time and reward



CA – Content Access
CC – Content Creation
IA – Interpretation and Analysis
IF – Information Finding
SU – Surveys
VV – Verification and Validation

Paper 1: Throughput Prediction

- Random Forest Regression using 29 features
  → Data from June to October 2014
  → HIT_available and Age_minutes have the biggest impact on prediction
    → Larger batches tend to attract more workers
    → But becomes less important after longer periods

Paper 1: Market Analysis

- Increase in arrived HITs is positively associated with a higher percentage of completed HITs

  → About 20% of the new demand gets completed within an hour

  → Most recently published HITs are presented first to workers

Paper 1: Conclusion and Future Work

- Overview of MTurk and its evolution
- Throughput prediction model
  → Publish large batches
- Worker-centered or requester-centered analysis

Paper 2

Kazai, Gabriella.

In search of quality in crowdsourcing for search engine evaluation.

ECIR '11

Paper 2

- Effects of pay, worker qualification and required effort on quality

- Series of controlled crowdsourcing experiments

→ designed several batches of HITS on MTurk with varied parameters

→ part of the INEX Book Track test collection

→ Initiative for the Evaluation of XML Retrieval

→ 4,490 judged pages

→ 8 different topics

→ 40% relevant and 60% irrelevant label ratio

Paper 2: HIT Design

- Judge whether book page is relevant to topic
  → ‘Relevant’, ‘Not relevant’, ‘Broken link’, ‘Don’t know’
  → comment field

- Captcha to reduce attractiveness of random clicking
  → enter last word on book page

- Questions regarding worker’s perception of the task
  → Familiarity, Task Difficulty, Interest in the task, Pay

Paper 2: Task Parameters

- 8 batches of HITs with different combinations of task parameter values
  → 800 book pages per batch with 100 book pages for each topic

Task parameter	Setting 1	Setting 2
Pay	$0.10 per HIT	$0.25 per HIT
Worker qualification	No required qualification	Over 95% approval rate with over 100 approved HITs
Effort	5 book pages per HIT	10 book pages per HIT

- 3 workers requested per HIT

Paper 2: Results

- Spam

  → less than 30% of captcha fields + less than 20s judging a page
- Cleaned

→ removed spam and HITs with unusable labels

	Batch/Subset	All Gathered Labels			No Spam			Cleaned		
		#Wkrs	Time	Acc.	#Wkrs	Time	Acc.	#Wkrs	Time	Acc.
1.	10-noQ-5	70	42	59.74%	66	51	61.79%	65	48	66.38%
2.	10-noQ-10	69	26	34.98%	63	42	59.29%	61	40	67.26%
3.	10-yesQ-5	66	42	60.78%	62	58	62.62%	62	58	66.97%
4.	10-yesQ-10	35	42	59.22%	33	59	59.75%	33	58	62.90%
5.	25-noQ-5	71	51	52.03%	68	61	54.79%	66	59	57.48%
6.	25-noQ-10	58	41	52.34%	54	49	63.50%	54	48	72.56%
7.	25-yesQ-5	43	61	71.50%	43	61	71.50%	43	61	73.58%
8.	25-yesQ-10	54	33	67.04%	48	38	69.83%	48	38	74.01%

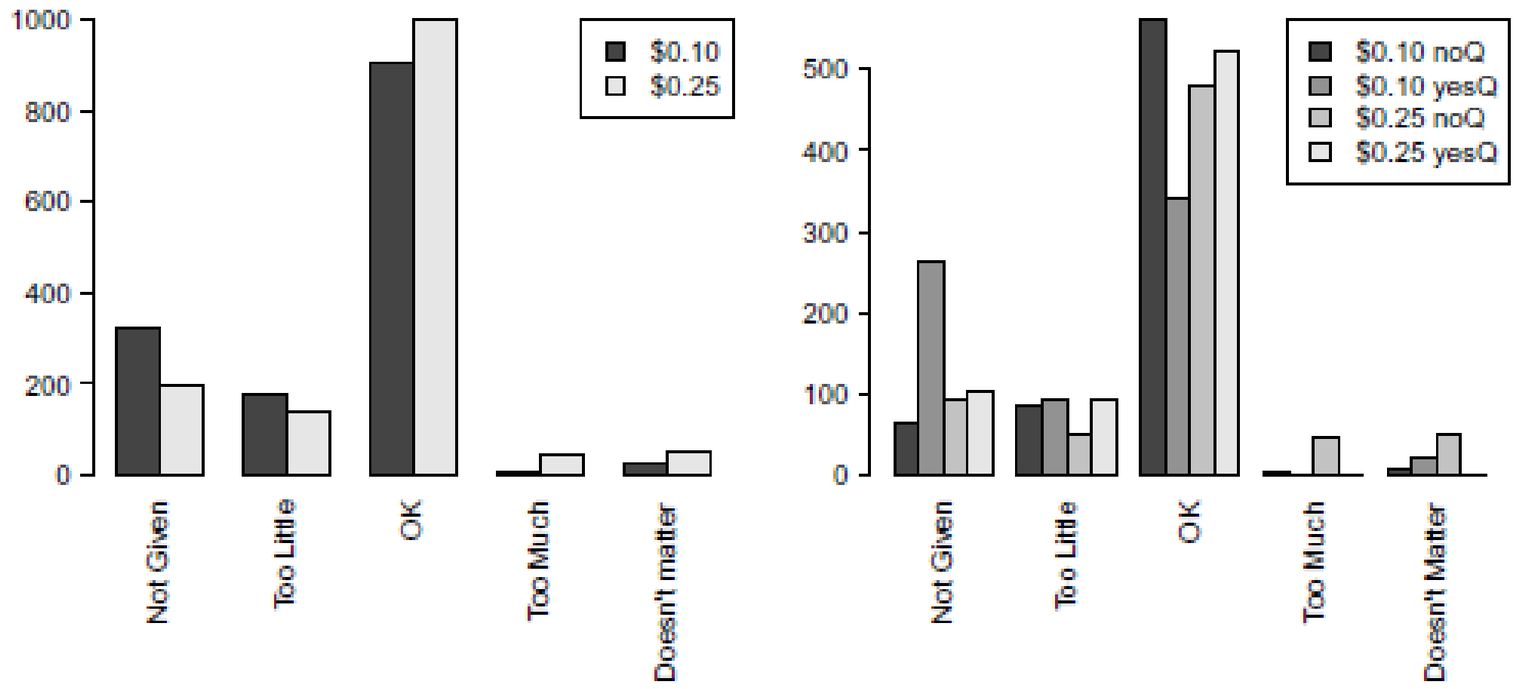
Paper 2: Pay

- Higher pay leads to better label quality
  - → Two sample t-test shows significant difference
  - → more usable labels and less spam
- But higher pay also attracts more unethical workers

	Batch/Subset	All Gathered Labels			No Spam			Cleaned		
		#Wkrs	Time	Acc.	#Wkrs	Time	Acc.	#Wkrs	Time	Acc.
15.	$0.10 noQ	121	32	44.83%	113	46	60.54%	110	44	66.81%
16.	$0.10 yesQ	90	42	60.00%	84	58	61.19%	84	58	64.93%
17.	$0.25 noQ	121	46	52.18%	114	55	59.15%	112	53	64.70%
18.	$0.25 yesQ	92	45	69.01%	86	49	70.67%	86	49	73.79%

Paper 2: Pay

- Majority were content with offered pay
- Qualified workers have higher expectations on pay
- Pay becomes secondary when interest is high



Workers‘ feedback on amount of pay

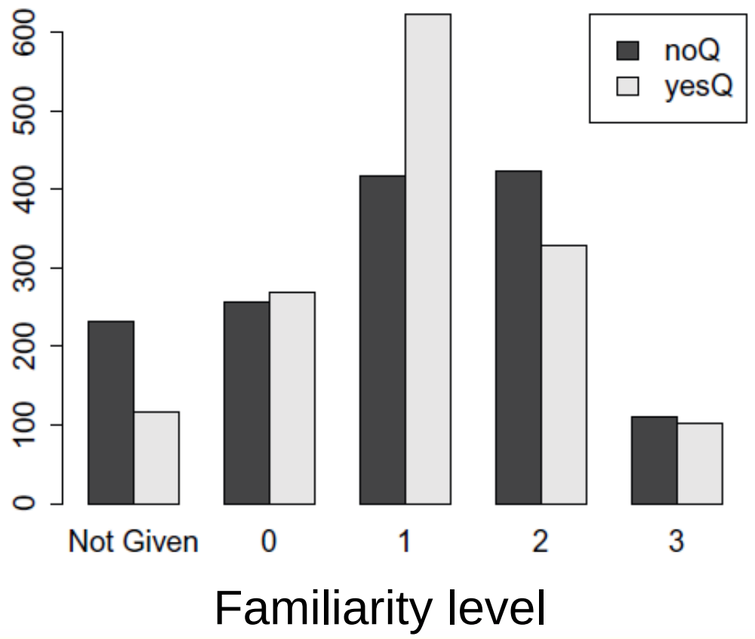
Paper 2: Worker Qualification

- Pre-selecting workers lead to better accuracy
  → Two sample t-test shows significant difference
- Non-qualified workers contribute more spam and unusable labels
  → Tend to overestimate their knowledge
- Best accuracy is obtained when workers reported minimal knowledge of the topic

	Batch/Subset	All Gathered Labels			No Spam			Cleaned		
		#Wkrs	Time	Acc.	#Wkrs	Time	Acc.	#Wkrs	Time	Acc.
19.	noQ	225	38	48.05%	211	51	59.84%	207	49	65.75%
20.	yesQ	155	43	63.88%	148	54	65.93%	148	54	69.40%

Paper 2: Worker Qualification

- Pre-selecting workers lead to better accuracy
  → Two sample t-test shows significant difference
- Non-qualified workers contribute more spam and unusable labels
  → Tend to overestimate their knowledge
- Best accuracy is obtained when workers reported minimal knowledge of the topic



Paper 2: Effort of HIT

- Better results when workers are not overloaded
  → Significant difference in accuracy in the unfiltered set
- Workers seem to be less motivated if more effort is required
  → More unusable and spam labels
  → On average less time spent judging a page with high effort

Paper 2: Conclusion and Future Work

- Network of influences between the different task parameters and the output quality

  → increase pay, reduce effort and use qualification requirements to raise quality and reduce undesirable behavior

- More detailed exploration
- Include new task parameters: clarity, emotion, aesthetics, etc.
- Goal: provide a framework to guide the design of crowdsourcing tasks to maximize quality

Discussion