

# Microtask Crowdsourcing: Fundamentals (Part II)

Maribel Acosta, Amrapali Zaveri



Maastricht University

Institute of Data Science

# Quality Control

# Quality Control

---

- Extremely important part of the experiment
- Approach as “overall” quality; not just for workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester.
  - Do check the worker forums!

CrowdFlower has an option to directly chat with a live worker !

# Quality Control

---

- Approval rate: easy to use, & just as easily defeated
- Mechanical Turk Masters (since June 2011)
  - Only for specific tasks
- Qualification test
  - Pre-screen workers' ability to do the task (accurately)
- Assess worker quality as you go
  - Trap questions with known answers (“honey pots”)
  - Measure inter-annotator agreement between workers

# Qualification tests: pros and cons

---

- Advantages
  - Great tool for controlling quality
  - Adjust passing grade
- Disadvantages
  - Extra cost to design and implement the test
  - May turn off workers, hurt completion time
  - Refresh the test on a regular basis
  - Hard to verify subjective tasks like judging relevance
- Try creating task-related questions to get worker familiar with task *before* starting task in earnest

# Methods for measuring agreement

---

- What to look for
  - Agreement, reliability, validity
- Inter-agreement level
  - Agreement between judges
  - Agreement between judges and the gold set
- Some statistics
  - Percentage agreement
  - Cohen's kappa (2 raters)
  - Fleiss' kappa (any number of raters)
- With majority vote, what if 2 say relevant, 3 say not?
  - Use expert to break ties
  - Collect more judgments as needed to reduce uncertainty

# Quality Control & Assurance

---

- Filtering
  - Approval rate (built-in but defeatable)
  - Geographic restrictions (e.g. US only, built-in)
  - Worker blocking
  - Qualification test
    - Con: slows down experiment, difficult to “test” relevance
    - Solution: create questions to let user get familiar *before* the assessment
  - Does not guarantee success
- Identify workers that *always* disagree with the majority
- Ask workers to rate the difficulty of a task

# Other quality heuristics

---

- Justification/feedback as quasi-captcha
  - Should be optional
  - Automatically verifying feedback was written by a person may be difficult (classic spam detection task)
- Broken URL/incorrect object
  - Leave an outlier in the data set
  - Workers will tell you
  - If somebody answers “excellent” for a broken URL => *probably* spammer



# Dealing with bad workers

---

- Pay for “bad” work instead of rejecting it?
  - Pro: preserve reputation, admit if poor design at fault
  - Con: promote fraud, undermine approval rating system
- Use bonus as incentive
  - Pay the minimum \$0.01 and \$0.01 for bonus
  - Better than rejecting a \$0.02 task
- If spammer “caught”, block from future tasks
  - May be easier to always pay, then block as needed

# Answer justification

---




- Why settle for a label?
- Let workers justify answers
- Has to be optional for good feedback


# Build Your Reputation as a Requestor


---


- Word of mouth effect
  - Workers trust the requester (pay on time, clear explanation if there is a rejection)
  - Experiments tend to go faster
  - Announce forthcoming tasks (e.g. tweet)


# Crowd Worker Communities

Rating [info]	Description
FAIR: 5 / 5 	No need to contact, HITs approved next day.
FAST: 5 / 5 	Jan 21 2013   <a href="#">rjsc...@g...</a>   <a href="#">flag</a>   <a href="#">comment</a>
PAY: 5 / 5 	
COMM: NO DATA	

communicativity:  5 / 5

generosity :  5 / 5

fairness :  5 / 5




promptness :  4.71 / 5

[What do these scores mean?](#)

Scores based on [7 reviews](#)

[Report your experience with this requester »](#)

Turkopticon.com  
Mturkforum.com  
Turkernation.com

FAIR: 5 / 5 	Small batch and mega bubbles. Not sure if I'm going in....
FAST: 4 / 5 	<b>Title:</b> Which is the most appropriate type?
PAY: 5 / 5 	<b>Requester:</b> <a href="#">Philippe Cudre-Mauroux</a> [A28PIN9Y6KHR3H] (TO)
COMM: NO DATA	<b>Description:</b> Please read the text and select the most appropriate description for each of the proposed entities.
	<b>Reward:</b> \$0.10
	<b>Qualifications:</b> HIT abandonment rate (%) is less than 51, HIT approval rate (%) is greater than 25, Location is US
	<b>Link:</b> <a href="https://www.mturk.com/mturk/preview?groupId=2ZSQUQIHPCGJ2FZIT6N51H1LQYU60M">https://www.mturk.com/mturk/preview?groupId=2ZSQUQIHPCGJ2FZIT6N51H1LQYU60M</a>

Powered by non-[amazonian](#) script monkeys ♦♦

To many bubbles but YMMV with your patience level.





“best collective decisions are  
result of *disagreement*,  
not consensus or compromise”

James Surowiecki



# BINARY WORLD

## 7 MYTHS ABOUT HUMAN ANNOTATION

**One truth:** knowledge acquisition for the semantic web assumes one correct interpretation for every example

**All examples are created equal:** triples are triples, one is not more important than another, they are all either true or false

**Disagreement bad:** when people disagree, they don't understand the problem

**Experts rule:** knowledge is captured from domain experts

**One is enough:** knowledge by a single expert is sufficient

**Detailed explanations help:** if examples cause disagreement - add instructions

**Once done, forever valid:** knowledge is not updated; new data not aligned with old

"Truth is a Lie: 7 Myths about Human Annotation", *AI Magazine* 2014, L. Aroyo, C. Welty



A photograph of a group of Japanese men in dark suits and ties, seated at a long wooden table in a formal meeting room. They are all raising their right hands, palm facing forward, in a gesture of agreement or voting. The men have various expressions, some looking towards the camera and others looking down. The background is a plain, light-colored wall. A semi-transparent grey banner with white text is overlaid across the middle of the image.

**disagreement = *signal***



# CrowdTruth

The framework for crowdsourcing ground truth data.

<http://crowdtruth.org/>

 Download

 Documentation

 Experiments

 Games

## disagreement is signal for the natural **ambiguity** of language and **diversity & perspectives** of human interpretation

Lora Aroyo, Chris Welty: **Truth is a Lie: 7 Myths about Human Annotation**, AI Magazine 2014.

Lora Aroyo, Chris Welty: **The Three Sides of CrowdTruth**. J. Human Computation. 1(1). 2014.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Arne Rutjes, Jelle van der Ploeg, Lora Aroyo, Robert-Jan Sips, Anca Dumitrache and Lukasz Romaszko: **Crowd Truth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data**. ISWC-RBDS 2014.



# CrowdED

---

Can we a-priori estimate optimal workers and tasks' assignment to obtain maximum accuracy on all tasks?

# CrowdED



---

a two-staged statistical  
**Crowdsourcing Experimental  
Design**





# CrowdED

---

## Stage 1:





- Train all workers 
- On a proportion of tasks 
- Identify best workers & 
- Hard tasks 

## Stage 2:

- Assign best workers to 
- Hard tasks 
- Remaining tasks 
- Calculate Overall Accuracy 

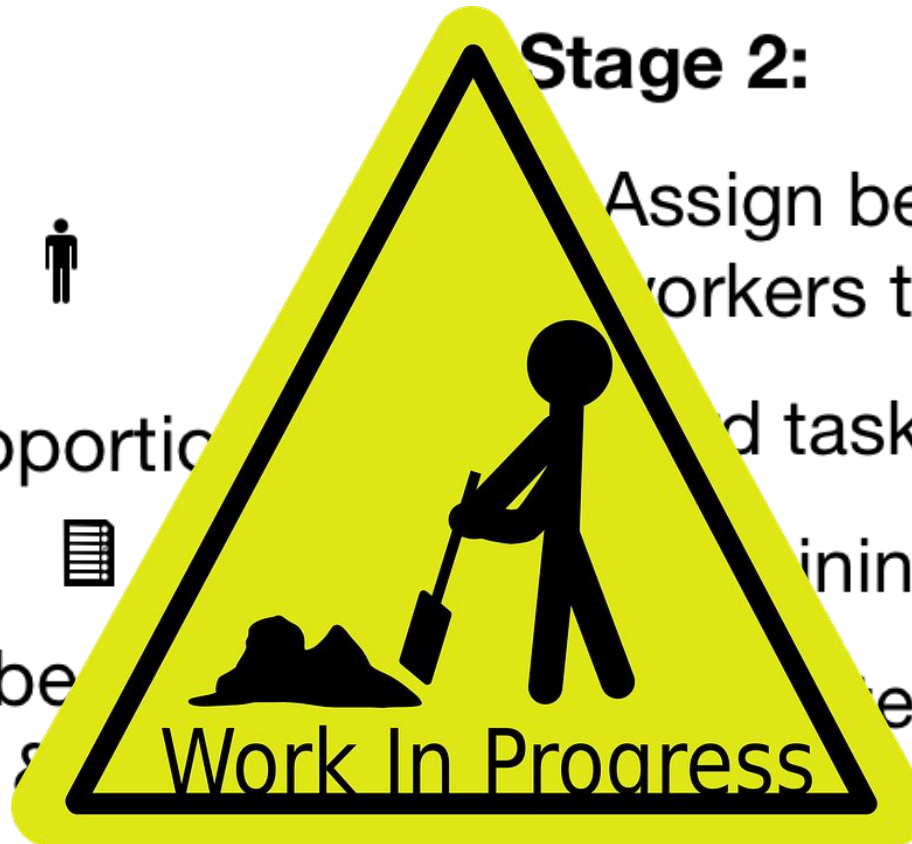
# CrowdED

## Stage 1:

- Train all workers 
- On a proportion of tasks 
- Identify best workers 
- Hard tasks 

## Stage 2:

- Assign best workers to 
- Hard tasks 
- Training tasks 
- Accuracy 



# Recommendations

---

- Reputation system for workers
- More than financial incentives
- Recognize worker potential (badges)
  - Paid for their expertise
- Train less skilled workers (tutoring system)

# Recommendations

---

- Promote workers to management roles
  - Create gold labels
  - Manage other workers
  - Make task design suggestions (first-pass validation)
- Career trajectory (based on reputation):
  1. Untrusted worker
  2. Trusted worker
  3. Hourly contractor
  4. Employee
- Platforms logs
  - Which kind of tasks attract skilled workers

# Summary

---

- Things that work
  - Qualification tests
  - Honey-pots
  - Good content and good presentation
  - Economy of attention
- Things to improve
  - Manage workers in different levels of expertise including spammers and potential cases.
  - Mix different pools of workers based on different profile and expertise levels.

# Summary

---

- Enforce Quality:
  - Task design
  - Iterate
  - Crowd incentives
  - Know your crowd: Model workers



# Hands-on II: Executing a task on CrowdFlower (Figure Eight)

<https://www.figure-eight.com/>

# Crowdsourcing Settings - 4 people per group

<b>Group</b>	<b>Task</b>	<b># Workers (Redundancy)</b>	<b>Contributor Level</b>
1	Data categorization	2	1
2	Data categorization	3	2
3	Data categorization	3	3
4	Data categorization	Dynamic	2
5	Image annotation	2	1
6	Image annotation	3	2
7	Image annotation	3	3
8	Image annotation	Dynamic	2

# Acknowledgements

---

Slides adapted from the tutorial “Microtask Crowdsourcing to Solve Semantic Web Problems” by Gianluca Demartini, Elena Simperl, and Maribel Acosta at ISWC 2013.

Source: <https://github.com/maribelacosta/crowdsourcing-tutorial>