Human Factors in Crowdsourcing

Gianluca Demartini
University of Sheffield / University of Queensland
http://gianlucademartini.net
@eglu81

Gianluca Demartini

g.demartini@sheffield.ac.uk

- B.Sc., M.Sc. at U. of Udine, Italy
- Ph.D. at U. of Hannover, Germany
  - Entity Retrieval
- Worked at the eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
- Senior Lecturer in Data Science at the iSchool, **U. of Sheffield** since 2014
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016, Facebook

www.gianlucademartini.net

Research Interests

- **Entity-centric Information Access** (2005-now)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  - NER in Scientific Docs (WWW 14), Prepositions (CIKM 14)
  - IR Evaluation (ECIR 16 Best Paper Award, IRJ 2015)
- **Hybrid Human-Machine Systems** (2012-now)
  - ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Human Memory based Systems (WWW 14, PVLDB)
  - Hybrid systems overview (COMNET, 2015)
- **Better Crowdsourcing Platforms** (2013-now)
  - Platform Dynamics (WWW 15)
  - Pick-a-Crowd (WWW 13), **Malicious Workers** (CHI 15)
  - Scale-up Crowdsourcing (HCOMP 14), Scheduling (WWW 16)
  - **Timeout** (HCOMP 16), **Complexity** (HCOMP 16)

Thanks to:

European Commission

EPSRC
Engineering and Physical Sciences Research Council

The University Of Sheffield.

Crowdsourcing

Crowdsourcing

- "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an **open call**. This can take the form of peer-production (when the job is performed **collaboratively**), but is also often undertaken by sole **individuals**. The crucial prerequisite is the use of the open call format and the **large network of potential laborers**.“

[Howe, 2006]

Incentives in Crowdsourcing

- **Extrinsic motivation** if task is considered boring, dangerous, useless, socially undesirable, dislikable by the performer.
  - Paid Crowdsourcing
- **Intrinsic motivation** is driven by an interest or enjoyment in the task itself.
  - Fun (enjoyment) / Games with a purpose
  - Community (belonging, desire to help)
  - Citizen Science

Dimensions of Human Computation

[Quinn & Bederson, 2012]

What is outsourced

- Tasks based on human skills not easily replicable by machines (visual recognition, language understanding, knowledge acquisition, basic human communication etc)

Who is the crowd

- Open call
- Call may target specific skills and expertise
- Requester typically knows less about the workers than in other work environments

How is the task outsourced

- Explicit vs. implicit participation
- Tasks broken down into smaller units undertaken in parallel by different people
- Coordination required to handle cases with more complex workflows
- Partial or independent answers consolidated and aggregated into complete solution

Dimensions of Human Computation (2)

[Quinn & Bederson, 2012]

How are the results validated

- Solutions space closed vs. open
- Performance measurements/ground truth
- Statistical techniques employed to predict accurate solutions
- May take into account confidence values of algorithmically generated solutions

How can the process be optimized

- Incentives and motivators
- Assigning tasks to people based on their skills and performance (as opposed to random assignments)
- Symbiotic combinations of human- and machine-driven computation, including combinations of different forms of crowdsourcing

Paid Micro-Task Crowdsourcing

A Crowdsourcing Platform allows **requesters** to publish a crowdsourcing request (***batch***) composed of multiple tasks (***HITs***)

Programmatically Invoke the crowd with APIs or using a website

Workers in the crowd complete tasks and obtain a monetary reward

Amazon MTurk

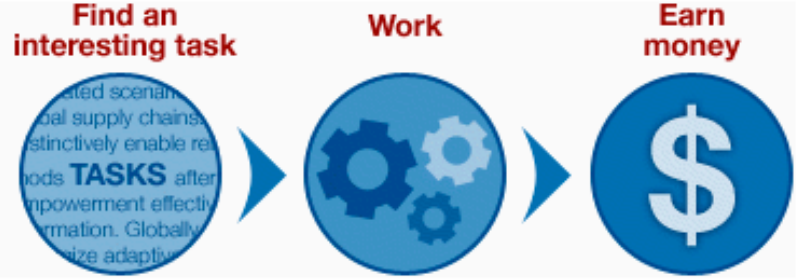
amazonmechanical turk
beta
Artificial Artificial Intelligence

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Find HITs Now

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Get Started

MTurk is a Marketplace for HITs

All HITs
1-10 of 3454 Results
Sort by: HITs Available (most first)

Show all details | Hide all details

1 2 3 4 5 › Next » Last

Provide Information about a Product				View a HIT in this group
Requester: requester	**HIT Expiration Date:** May 23, 2015 (4 weeks 1 day) **Time Allotted:** 25 minutes		**Reward:** $0.05 **HITs Available:** 11526	
Product Attribute Tagging - April 17th Please read the instructions				View a HIT in this group
Requester: slee	**HIT Expiration Date:** May 23, 2015 (4 weeks 2 days) **Time Allotted:** 60 minutes		**Reward:** $0.03 **HITs Available:** 23887	
Inv_B_2				View a HIT in this group
Requester: rohzit0d	**HIT Expiration Date:** May 22, 2015 (4 weeks 1 day) **Time Allotted:** 48 minutes		**Reward:** $0.00 **HITs Available:** 19740	
Geo Result Relevance-Tue Apr 21 10:40:14 PDT 2015				View a HIT in this group
Requester: Amazon Requester Inc.	**HIT Expiration Date:** May 22, 2015 (4 weeks 1 day) **Time Allotted:** 60 minutes		**Reward:** $0.00 **HITs Available:** 10734	
Type the text from the images, carefully. Productivity and bonuses guaranteed.				View a HIT in this group
Requester: CopyText Inc.	**HIT Expiration Date:** Apr 30, 2015 (6 days 23 hours) **Time Allotted:** 10 minutes		**Reward:** $0.01 **HITs Available:** 10590	
Transcribe up to 25 Seconds of Media to Text - Earn up to $0.12 per HIT!				View a HIT in this group
Requester: Crowdsurf Support	**HIT Expiration Date:** Apr 21, 2016 (51 weeks 6 days) **Time Allotted:** 15 minutes		**Reward:** $0.08 **HITs Available:** 6702	
Fun and Fast Fashion Tagging				View a HIT in this group
Requester: gavin	**HIT Expiration Date:** Apr 28, 2015 (5 days 11 hours) **Time Allotted:** 60 minutes		**Reward:** $0.02 **HITs Available:** 6460	
Geo Result Relevance-Wed Apr 08 14:30:08 PDT 2015				View a HIT in this group
Requester: Amazon Requester Inc.	**HIT Expiration Date:** May 10, 2015 (2 weeks 2 days) **Time Allotted:** 60 minutes		**Reward:** $0.00 **HITs Available:** 6182	
Transcribe up to 25 Seconds of General Content to Text - Earn up to $0.14 per HIT!				View a HIT in this group
Requester: Crowdsurf Support	**HIT Expiration Date:** Apr 21, 2016 (51 weeks 6 days) **Time Allotted:** 15 minutes		**Reward:** $0.09 **HITs Available:** 6043	
!Whac-a-mole by Gaze (hard mode) ! Play a 1min eye tracking game in the web browser! 0416				View a HIT in this group
Requester: px	**HIT Expiration Date:** Apr 23, 2015 (8 hours 40 minutes) **Time Allotted:** 60 minutes		**Reward:** $0.10 **HITs Available:** 4682	

1 2 3 4 5 › Next » Last

amazon mturk Worker | HITs | Dashboard | Skip | Accept

HIT Details | Reward: $0.15 | Time Allotted: 00:05:00

You must accept this HIT before working on it.

Data Collection Instructions!

Find the postal address for this Australian company.

- Search on Google, the company's website, YellowPages or Facebook to find the correct postal address for the company below.
- Enter the **full Australian postal address** for the business.
- You may use the research links provided to help.
- **Do not enter incomplete or incorrect details!**

Company name:	Stellar Electrical And Solar Systems
Location:	Australia
Company website:	
Company YellowPages:	
Company Facebook:	
Google search:	https://www.google.com.au/search?q=%22Stellar Electrical And Solar Systems%22+Australia+postal+address

Australian Street Address (ONLY this field is required if complete):

Start typing Australian Street Address...

amazon mturk beta Worker | HITs | Dashboard

HIT Details

Reward: $0.01

You must accept this HIT before working on it.

Receipt Transcription Instructions (Click to expand)

9× Subscription to Quip Business - Monthly	$108.00
Subtotal	$108.00
Coupons	-$30.00
Credit for first five users - QUIP3120 ($30.00 off)	
Total	$78.00

Is the receipt legible?

◯ Legible ◯ NOT legible

Issuer name:

Company Inc.

Invoice number:

IV2348977374

Invoice Date:

2017-05-13

Currency (3 digits):

USD / EUR / ...

Content and Cost:

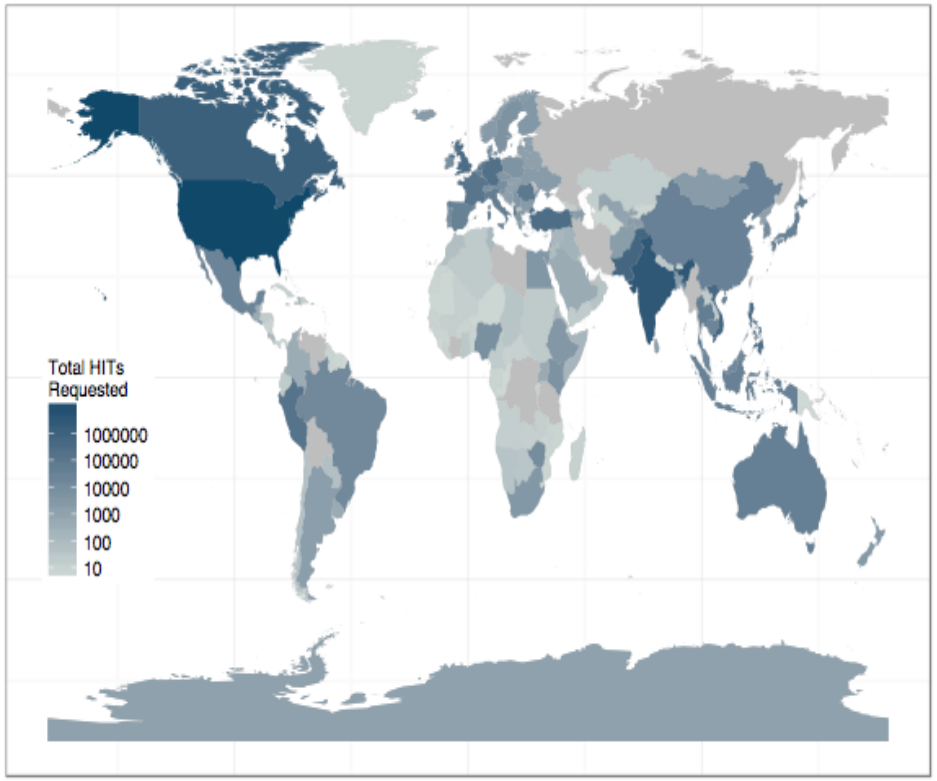
Content 1	0.00

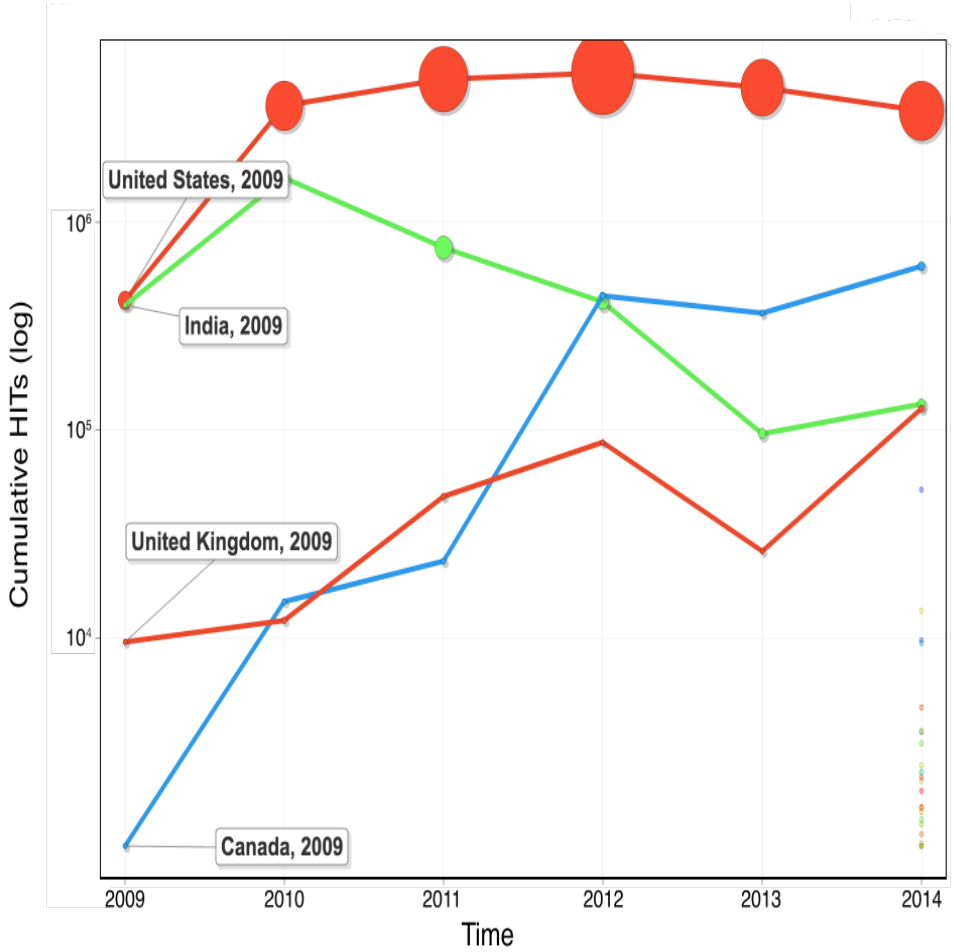
MTurk is a Marketplace for HITs

Top-1000 Requesters, report for April 16, 2016 to May 16, 2016

Requester name	hits	reward
Speechpad	23857	$172,994.63
Percy Liang	883	$7,320.48
Princeton Vision	51187	$5,762.44
Stanford GSB Behavioral Lab	3749	$2,110.70
Chris Callison-Burch	8157	$2,064.29
RC.org Mechanical Turk	6591	$2,011.33
VacationrentalAPI	399	$1,373.50
Med Expertise	869	$1,303.50
Bluejay Labs	13613	$1,288.59
YL Testing	1051	$1,236.83

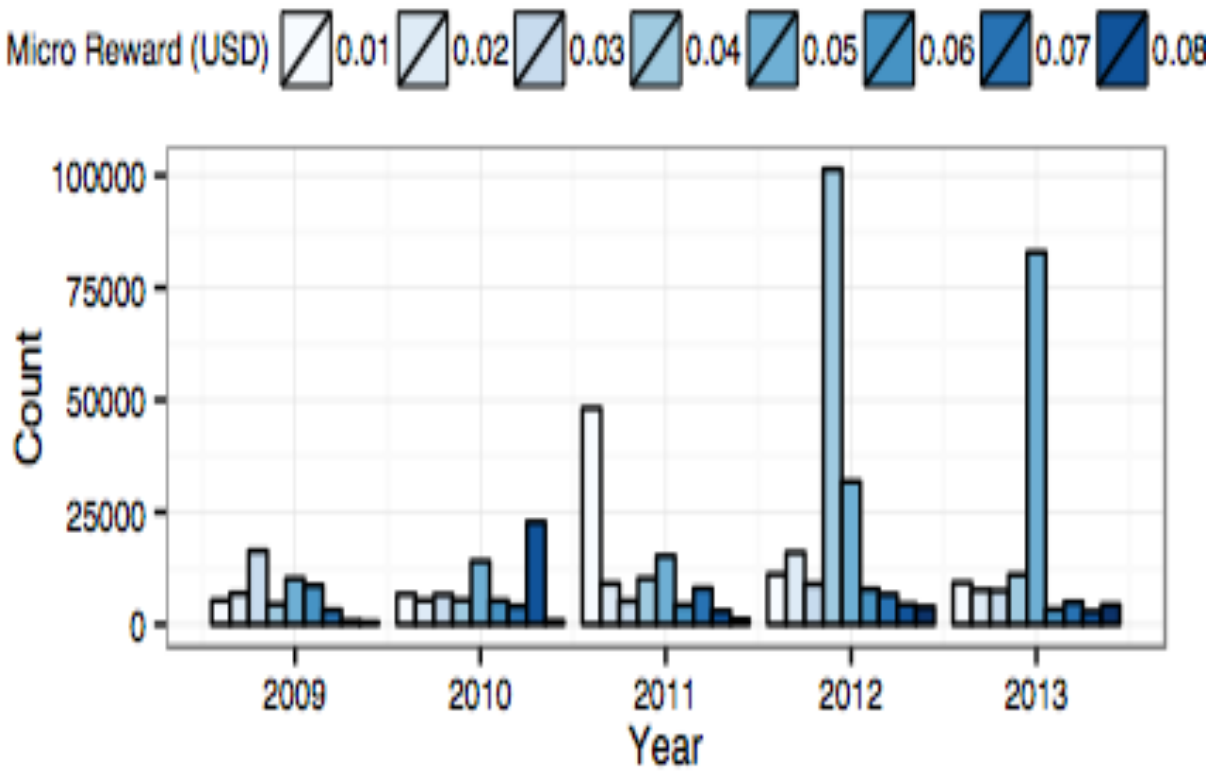
Requested Workers

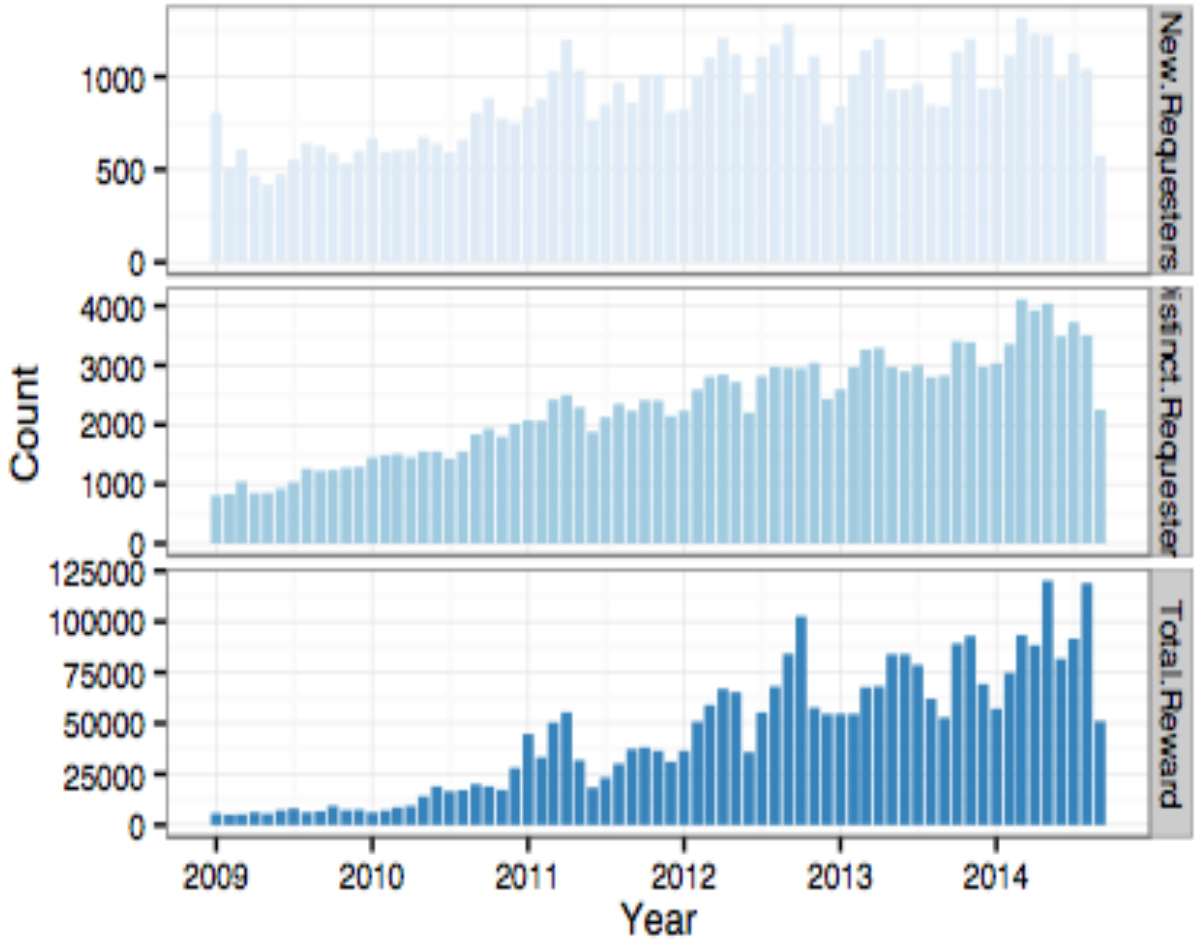




Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. **The Dynamics of Micro-Task Crowdsourcing -- The Case of Amazon MTurk**. In: 24th International Conference on World Wide Web (WWW 2015), Research Track. Firenze, Italy, May 2015.

Reward Distribution





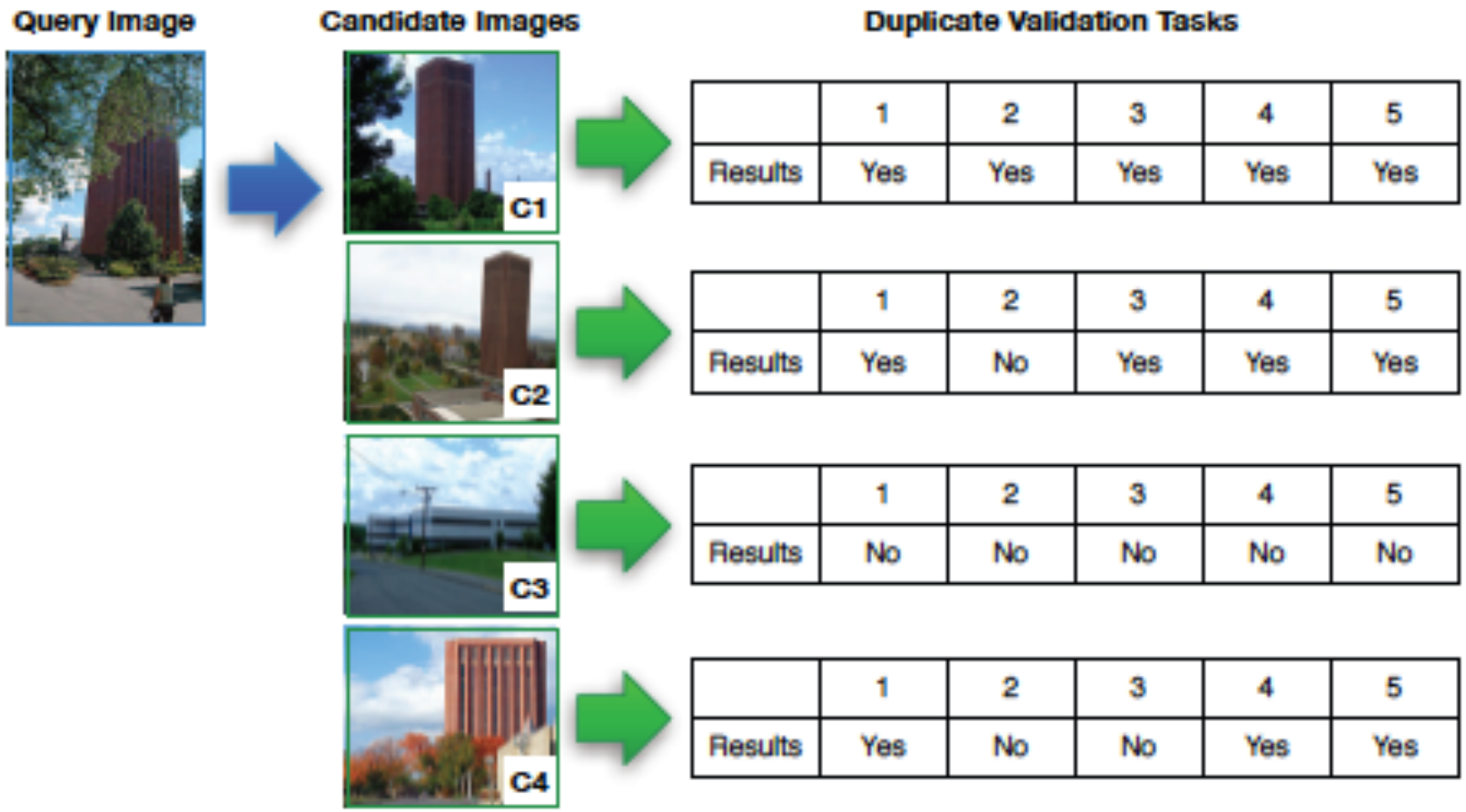
Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. **The Dynamics of Micro-Task Crowdsourcing -- The Case of Amazon MTurk**. In: 24th International Conference on World Wide Web (WWW 2015), Research Track. Firenze, Italy, May 2015.

Hybrid Human-Machine Systems

- Use Machines to scale over large amounts of data
- Keep humans in the loop
  - By means of Crowdsourcing
  - To make sure the quality of the data processing is good
- Crowd for Pre-processing vs Post-processing

G Demartini. Hybrid human–machine information systems: Challenges and opportunities. In: **Computer Networks**, 90, 5-13. 2015

Hybrid Image Search



Yan, Kumar, Ganesan, CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones, Mobisys 2010.

CrowdDB

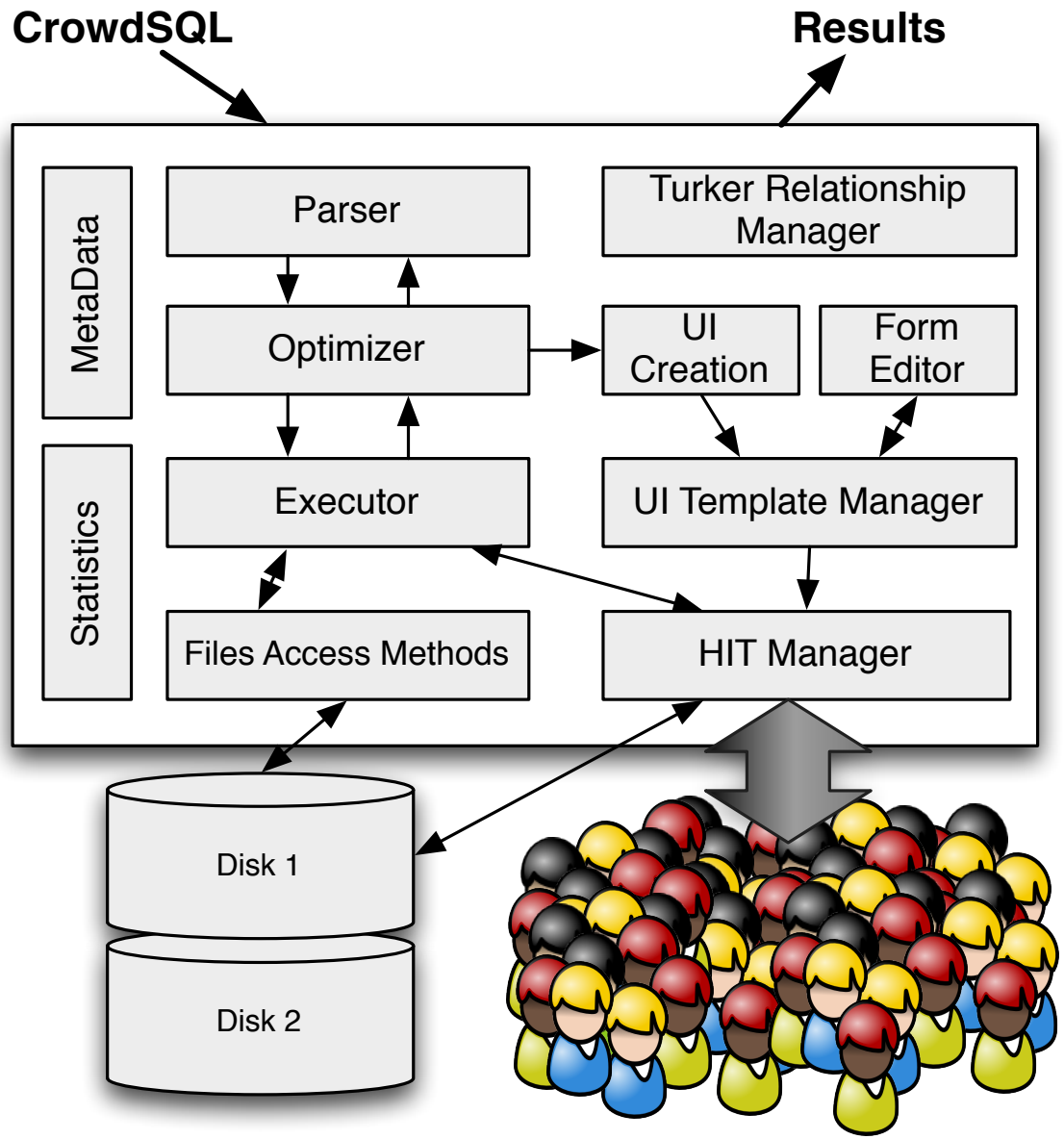
Use the crowd to answer DB-hard queries

Where to use the crowd:

- **Find missing data**
- **Make subjective comparisons**
- **Recognize patterns**

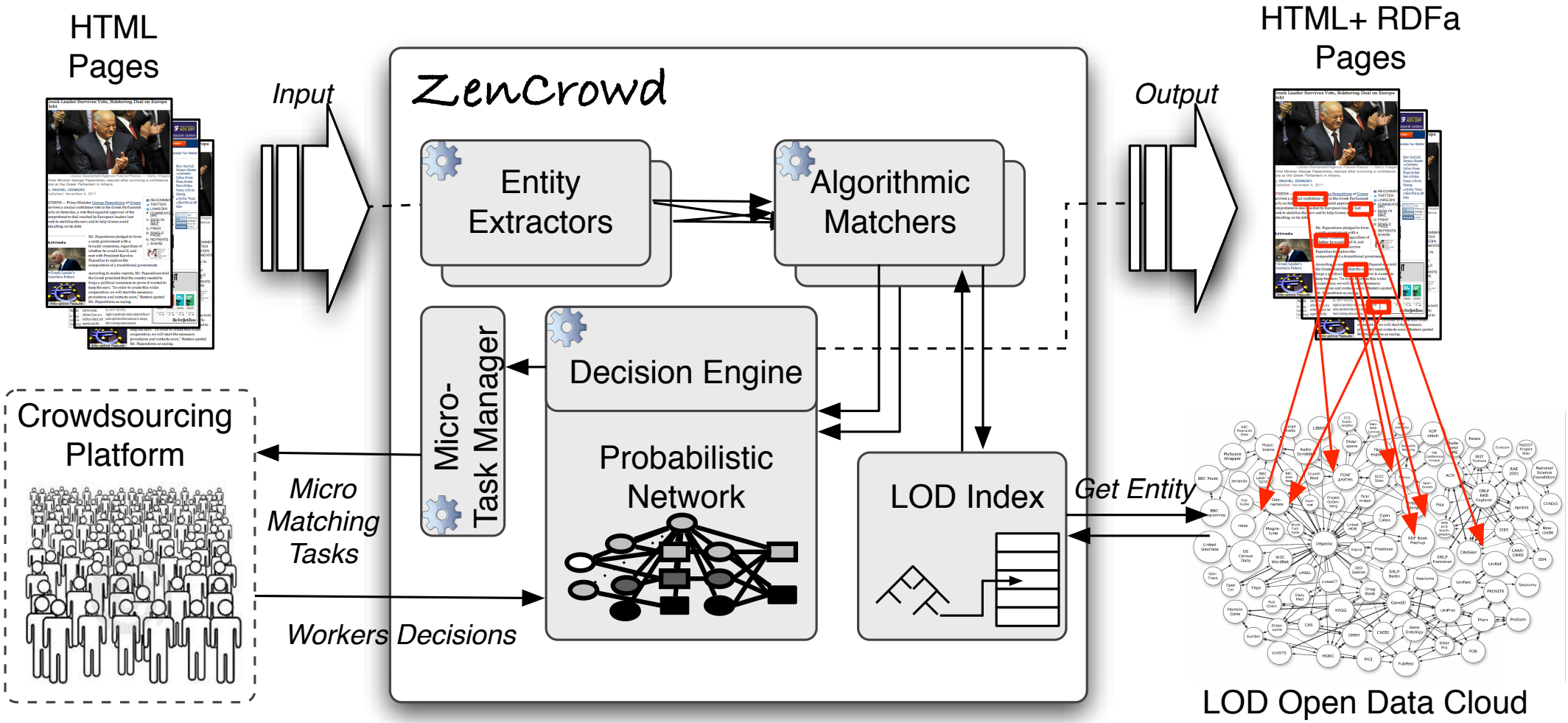
But not:

- Anything the computer already does well



M. Franklin, D. Kossmann, T. Kraska, S. Ramesh and R. Xin.
CrowdDB: Answering Queries with Crowdsourcing, *SIGMOD 2011*

ZenCrowd



Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In: 21st International Conference on World Wide Web (**WWW 2012**).

Human Computation 101 - Summary

- Crowdsourcing is growing in popularity
- It is used both in industry and academia
- For a number of applications across disciplines

- Open questions:
  - How to make sure we get quality results back from a crowdsourcing platforms? (**Effectiveness**)
  - Can we optimize the cost and execution in paid micro-task crowdsourcing? (**Efficiency**)

Human Factors - Outline

- Understanding **malicious behaviors** in paid crowdsourcing (CHI 2015)
- The effect of limiting **task time** (HCOMP 2016)

Understanding Malicious Behaviors

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding Malicious Behaviour in Crowdsourcing Platforms: The Case of Online Surveys. In**: Proceedings of the ACM Special Interest Group on Computer Human Interaction (CHI 2015)**. Seoul, South Korea, April 2015

Quality Control in Paid Crowdsourcing

- Diverse pool of crowd workers
- Wide range of behavior
- Various motivations

➢Typically adopted solution to prevent/flag malicious activity :
Gold-Standard Questions

Research Questions

RQ1: Do untrustworthy workers adopt different **methods to complete tasks**, and exhibit different kinds of behavior?

RQ2: Can **behavioral patterns** of malicious workers in the crowd be identified and quantified?

Design

- CrowdFlower Platform to deploy survey
- Survey questions
  - Demographics
  - Educational & general background
- 34 Questions in total
  - Open-ended
  - Multiple Choice
  - Likert-type
- Responses from 1000 crowd workers
  - Monetary Compensation per worker : 0.2 USD

RQ1 - Behavioral Patterns

Ineligible Workers (IW)

Instruction: Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.
Response: *'this is my first task'*

Fast Deceivers (FD)

eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.
Response: *'What's your task?'* , *'adasd'*, *'fgfgf gsd ljlkj'*

Rule Breakers (RB)

Instruction: Identify 5 keywords that represent this task (separated by commas).
Response: *'survey, tasks, history'* , *'previous task yellow'*

Smart Deceivers (SD)

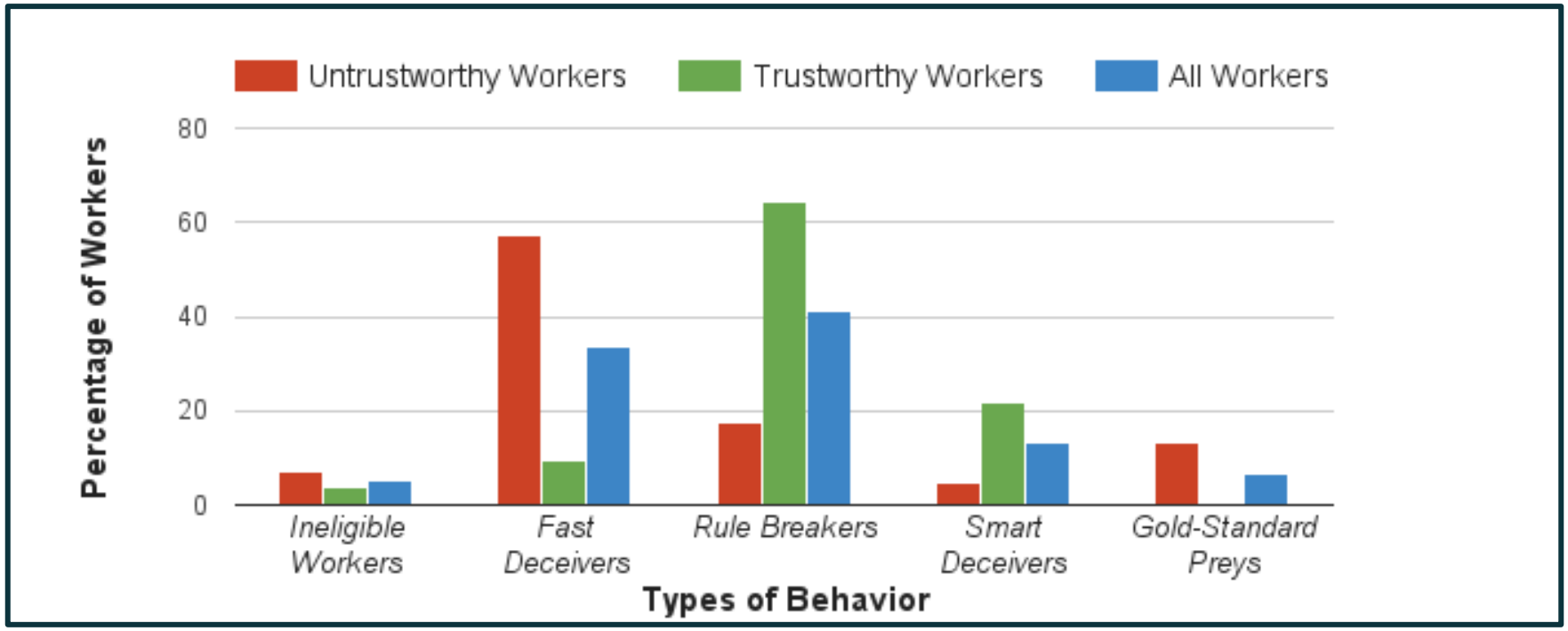
Instruction: Identify 5 keywords that represent this task (separated by commas).
Response: *'one, two, three, four, five'*

Gold Standard Preys (GSP)

These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions!

RQ2 - Distribution of Low-quality Workers

- passed the gold-standard: Trustworthy workers (TW)
- failed to pass the gold-standard: Untrustworthy workers (UW)



Tipping Point

"the first point at which a worker begins to exhibit malicious behavior after having provided an acceptable response"

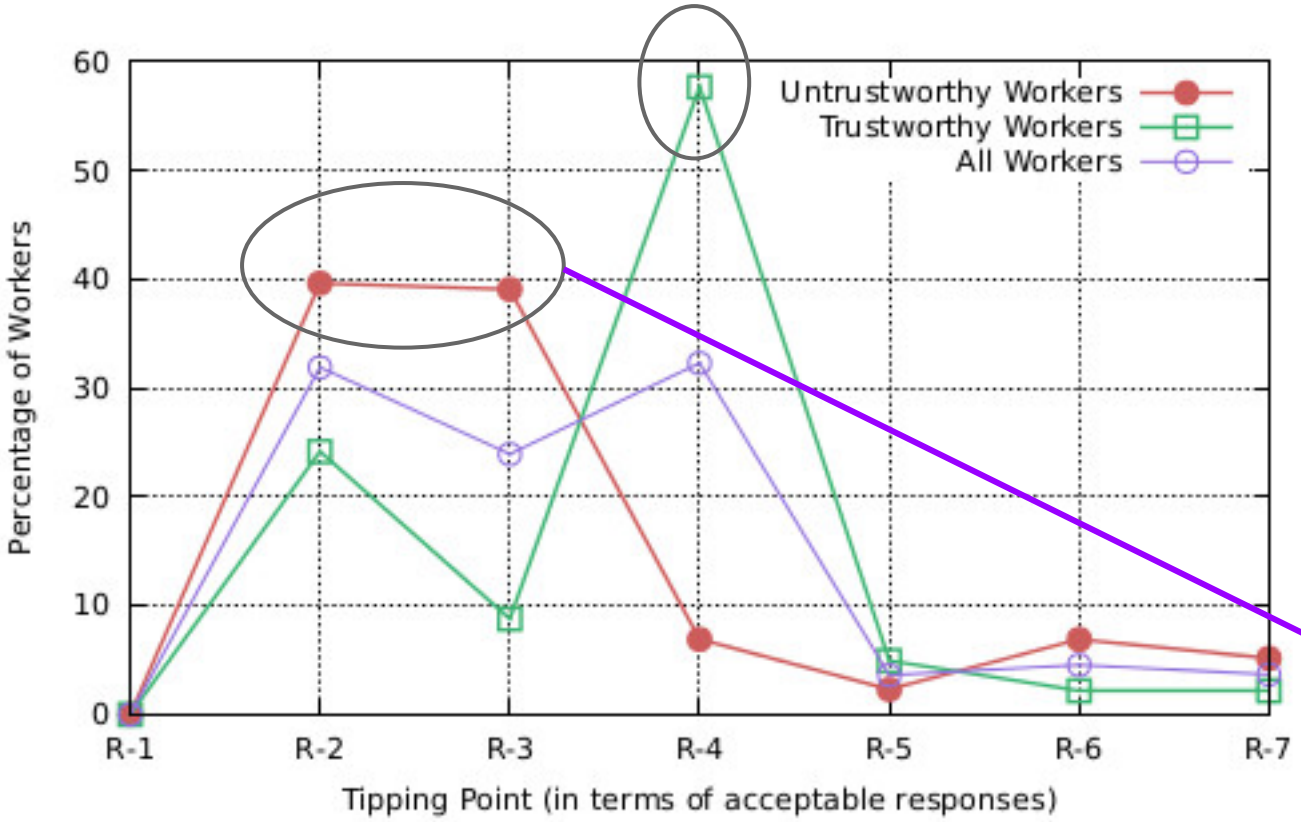


Table 1. Relationship between the Maliciousness and Tipping Point of untrustworthy and trustworthy workers (percentage of workers having tipping point @R).

Maliciousness	UW	TW
$0 < M \leq 0.2$	40.9% @ R-7 31.8% @ R-6	28.5% @ R-7 28.5% @ R-5
$0.2 < M \leq 0.4$	43.47% @ R-3 21.73% @ R-6	30% @ R-5 30% @ R-3
$0.4 < M \leq 0.6$	66.19% @ R-3 25.35% @ R-2	88% @ R-4 5.1% @ R-3
$0.6 < M \leq 0.8$	71.05% @ R-2 28.95% @ R-3	60% @ R-3 40% @ R-2
$0.8 < M \leq 1$	100% @ R-2	100% @ R-2

Findings

- Identified different types of malicious behavior exhibited by crowd workers.
- Measuring ‘maliciousness’ of workers to quantify their **behavioral traits**, and ‘**tipping point**’ to further understand worker behavior.
- This understanding helps requesters in effective task design, ensures adequate utilization of the crowdsourcing platform(s).
- Guidelines for efficient design of Surveys by limiting malicious activity.
  - Pre-screening (ineligible)
  - Validators (fast deceivers, rule breaker)
  - Psychometric approaches (smart deceivers)

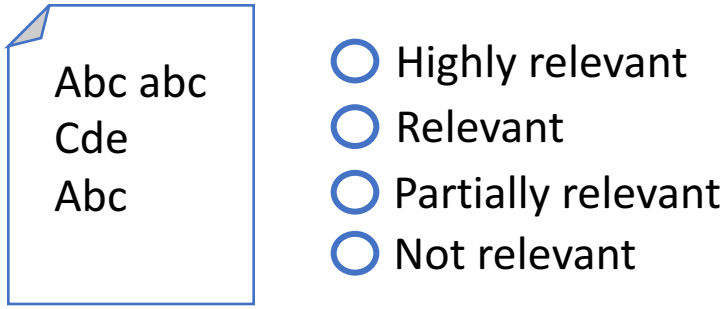
The Unexpected Benefits of Limiting the Time to Judge

Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In: **The 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016).** Austin, Texas, October 2016.

Crowdsourcing Relevance Judgements

- Task: Given a Query, Document pair
  Is the doc
  highly relevant, relevant, partially relevant, not relevant?
- Ask multiple workers
- Aggregate answers to obtain a relevance label

Query: jaguar



Our Research Question

Can we **limit the time to judge** to **reduce the cost ($$)** of creating IR test collections?

Hypothesis

Yes, but with quality loss

Our Experimental Setup

- **TREC8** Topics and documents (binary and 4-level expert judgements)
- **CrowdFlower**, repeated for USA and IND
- Majority vote aggregation
- Quality control: topic understanding question + high quality workers
- HIT Reward adapted based on the expected completion time

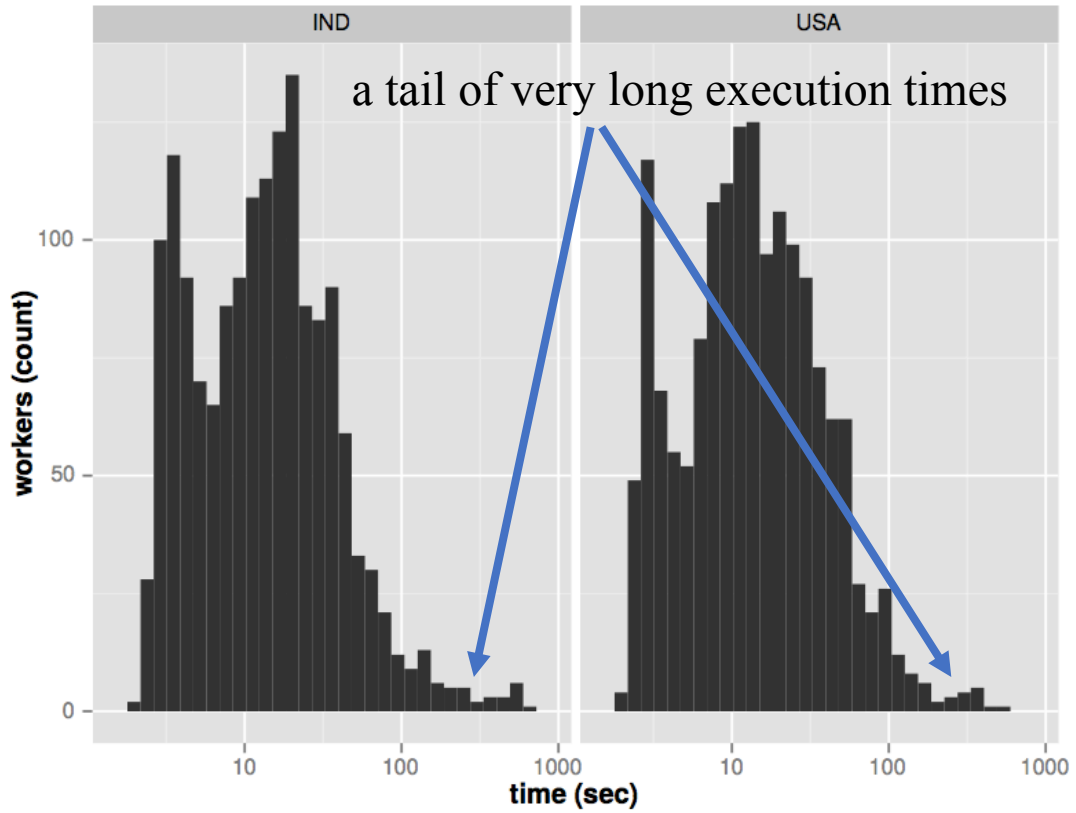
- Quality of a judgement: **Agreement** with editorial judgements
  - Cohen's Kappa and distance with 4-level labels

Our Experimental Setup

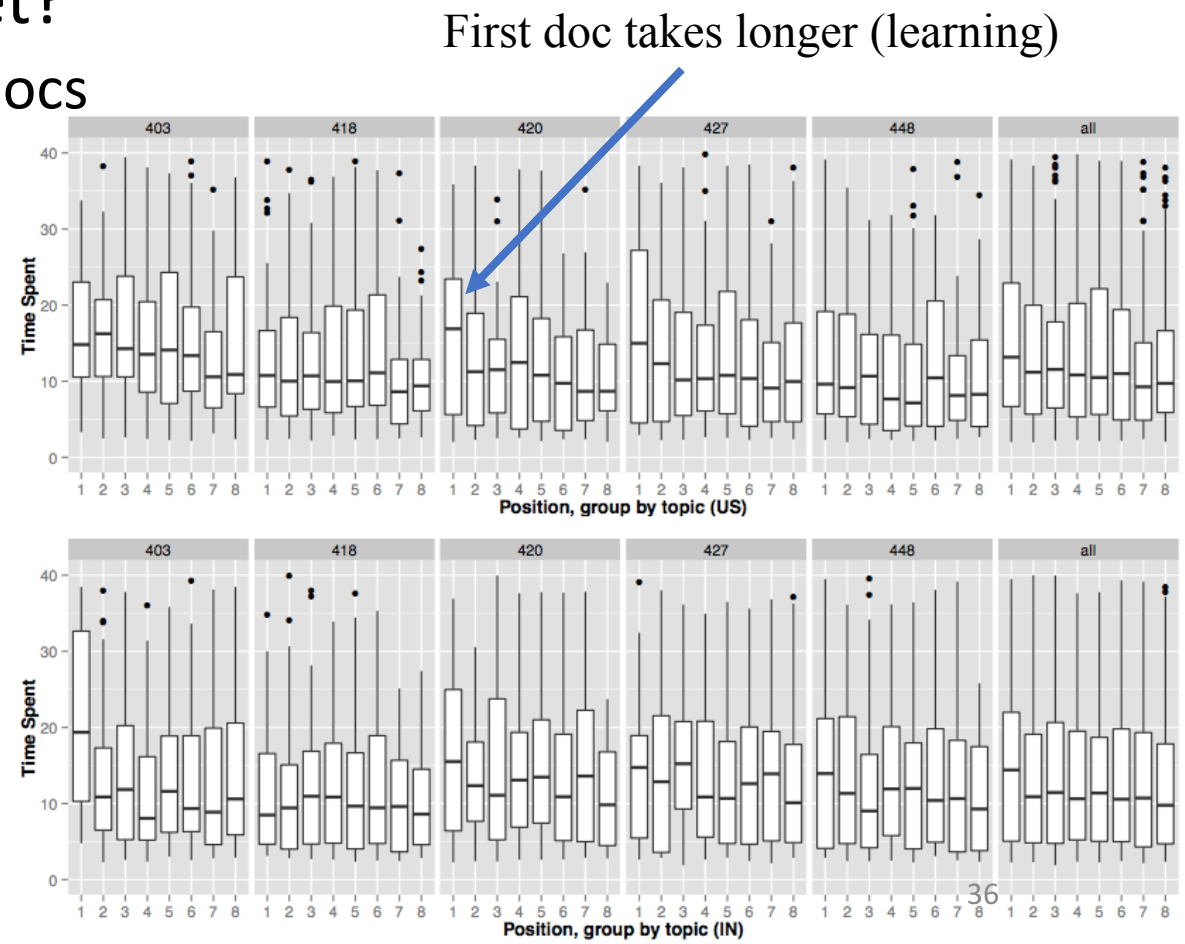
- E1 **Unbound time** (i.e., the standard approach)
  - 5 judgements per doc, 8 documents, 5 topics, 2 crowds = 400 workers
- E2 Document shown for a **predefined amount of time**
  - **30, 15, 7, 3 seconds**. Each worker to judge 8 docs
- E3 **Same timeout** for all 8 documents (**15 or 30 sec**)
- E4 **Fixed budget**: comparison between
  - more quick judgements
  - few slow judgements

E1: We Have All the Time in the World

- RQ: **How much time** do crowd workers take to judge the relevance of a document **if no time constrain** is set?
  - 5 workers to judge a permutation of 8 docs

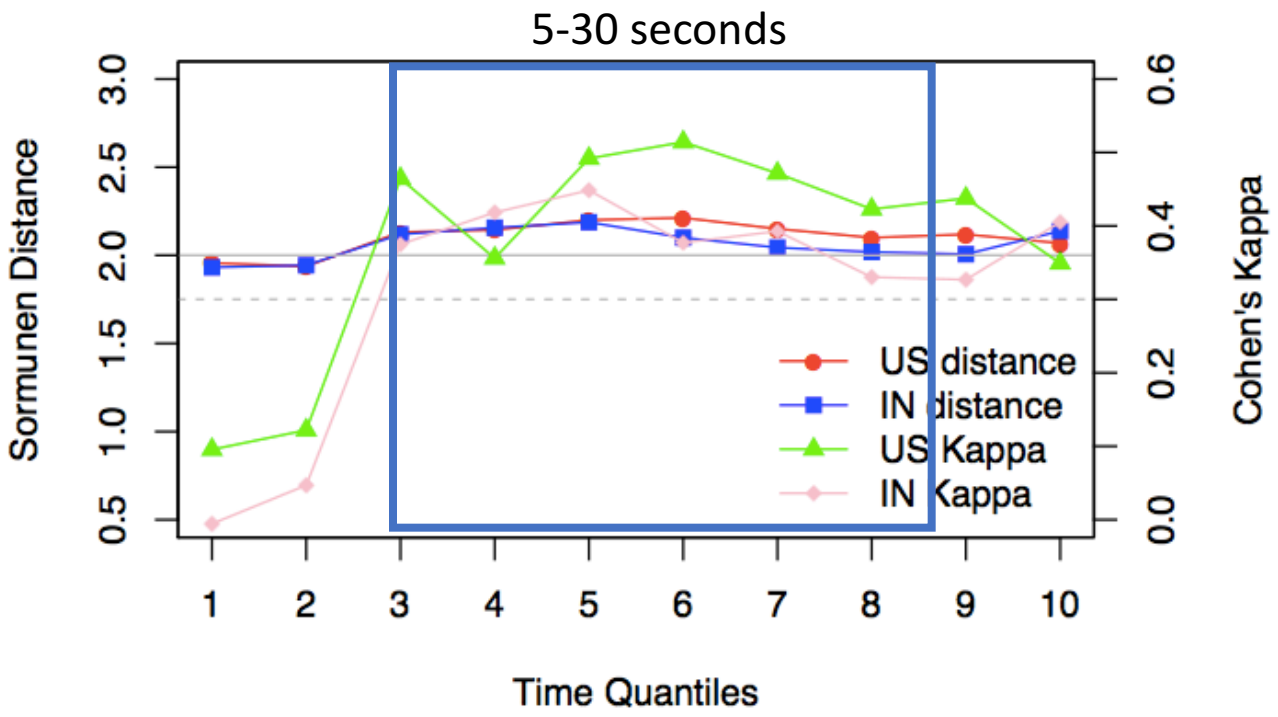


Median: 13 sec
Mean 24-25 sec



E1: We Have All the Time in the World

- No correlation of time with
  - Doc length
  - Doc readability
  - Topic
  - Relevance level
- Time vs Quality



	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
U.S.	2.0	3.2	**5.1**	7.6	10	13	17	**23**	32	51	580
IN	1.9	3.4	**4.5**	7.0	9.9	13	17	**22**	31	46	630

E2: Faster! Faster! Sorry, Too Late

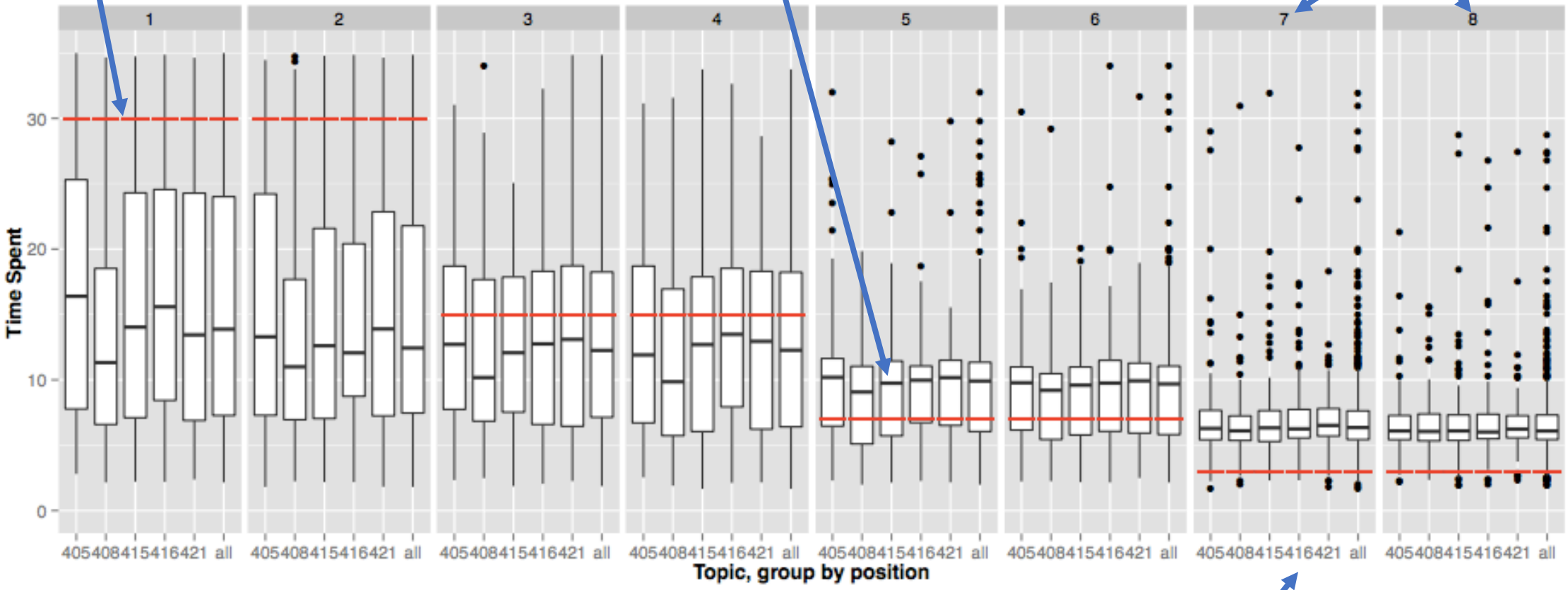
- Understand which is the **minimum amount of time required** to perform relevance judgments
- (max) timeouts: 30, 15, 7, 3 seconds
- Each worker to judge 8 docs, 2 for each timeout (one long, one short)
- Looking at Quality measures:
  - 3 and 7 secs are not enough
  - 15 slightly better than 30 (learning bias for position 1-2?)

E2: Faster! Faster! Sorry, Too Late

Time when document disappears

Time when judgement is made

Position of the document judged (1-8)



Variance across topics

E3: Selecting the Best Timeout

- We repeated E1 using 15 and 30 sec timeouts

- 15 seconds timeouts yield consistently better quality judgements
  - Than 30 seconds timeouts
  - Than no timeouts (E1 quality values)

Our Research Question

Can we **limit the time to judge** to **reduce the cost ($$)** of creating IR test collections?

Hypothesis **Yes, and it improves the quality!**

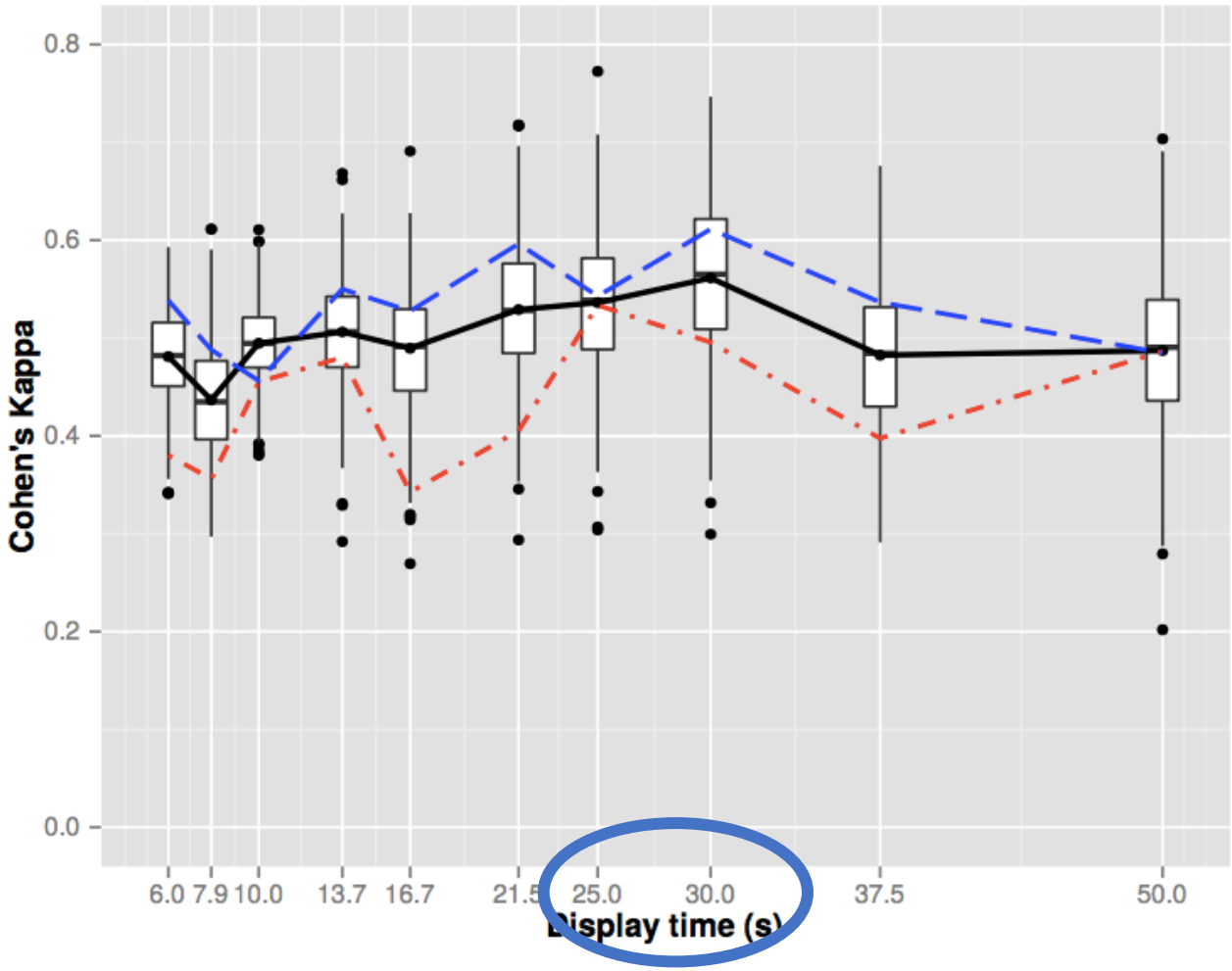
~~Yes, but with quality loss~~

E4: Many Fast&Furious or a Few Laid-Back?

- **Fixed budget**:
  - small timeout, more workers
  - Long timeout, less workers
- We compared 10 combinations with the same budget

Timeslot(sec)	6	7.9	10	13.7	16.7	21.5	25	30	37.5	50
Assignments	25	19	15	11	9	7	6	5	4	3

- **Highest quality at 25-30 sec**



Findings

- The **first** couple of judgments done by a worker are of **lower quality**
- Judgements that take **more than 30** seconds are of **lower quality**
- **Time-outs** in relevance judgements HITs can **increase quality**
- The **best timeout** to be used lies in the interval of **25-30 seconds** and does not depend on topic, document, or crowd.

Discussion

- Crowdsourcing Relevance Judgements for IR Evaluation can be **expensive to scale**
- **Limiting the time** to judge can **control the cost**
- But can also **increase the quality**!
  - By inducing workers to look at the document for a predefined amount of time
- Why? (Hypotheses)
  - With a balance between boredom and stress -> “in the flow”
  - System I and System II thinking

Conclusions

- Paid micro-task crowdsourcing to build hybrid human-machine systems
- Human-in-the-loop systems means human factors!

- Malicious behaviors
  - Supervised worker type classification
- Timeouts to increase efficiency and effectiveness of crowd work
  - Does it generalize to other task types?
- Predicting task complexity
  - Build recommender systems / order tasks based on complexity (gamification)