These Slides are here:
http://www.gianlucademartini.net/crowdsourcing/adc2017

# Crowdsourcing for Data Management

Gianluca Demartini

University of Queensland

http://gianlucademartini.net

@eglu81

# Gianluca Demartini

g.demartini@uq.edu.au

- B.Sc., M.Sc. at U. of Udine, Italy
- Ph.D. at U. of Hannover, Germany
    - Entity Retrieval
- Worked at the University of Sheffield (UK), eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
- Senior Lecturer in Data Science at the University of Queensland, since 2017.
- 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 (CIKM 2017, 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.

# Course Outline

- Micro-task Crowdsourcing
    - Examples
    - Dimensions
    - Platforms (Amazon MTurk)
- Hybrid human-machine systems
    - Examples
    - Challenges: Efficiency / Effectiveness
- Effectiveness
    - Quality control
    - Task assignment
- Efficiency
    - Scheduling
    - Pricing / Timeouts

# Crowdsourcing

from http://www.bbc.co.uk/news/magazine-32993891

# Crowdsourcing

- *Portmanteau* of "crowd" and "outsourcing," first coined by Jeff Howe in a June 2006 Wired magazine article

- [Merriam-Webster] the practice of obtaining needed services, ideas, or content by soliciting **contributions from a large group of people** and especially from the online community rather than from traditional employees or suppliers

# 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

# Games with a Purpose

- Tasks leveraging common human skills, appealing to large audiences
    - Selection of domain and task more constrained in games to create typical UX
- Tasks decomposed into smaller units of work to be solved independently
- Complex workflows
    - Creating a casual game experience vs. patterns in microtasks
    - Single vs. multi-player

# 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

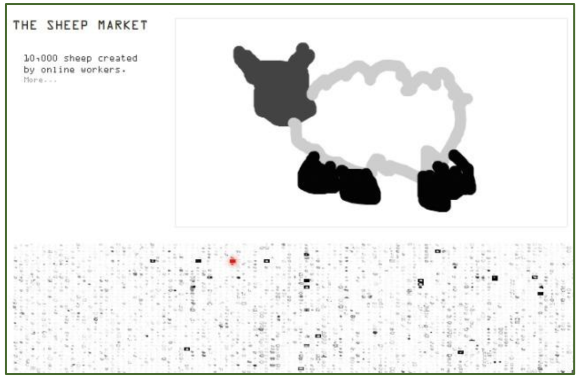
**Workers** in the crowd complete tasks and obtain a monetary reward

The **platform** takes a fee (30% of the reward)

# Example use of micro-task crowdsourcing

- Relevance judgments
- Ontologies
- Sentiment Analysis in Social Media
- http://www.thesheepmarket.com/

'draw a sheep facing to the left.'



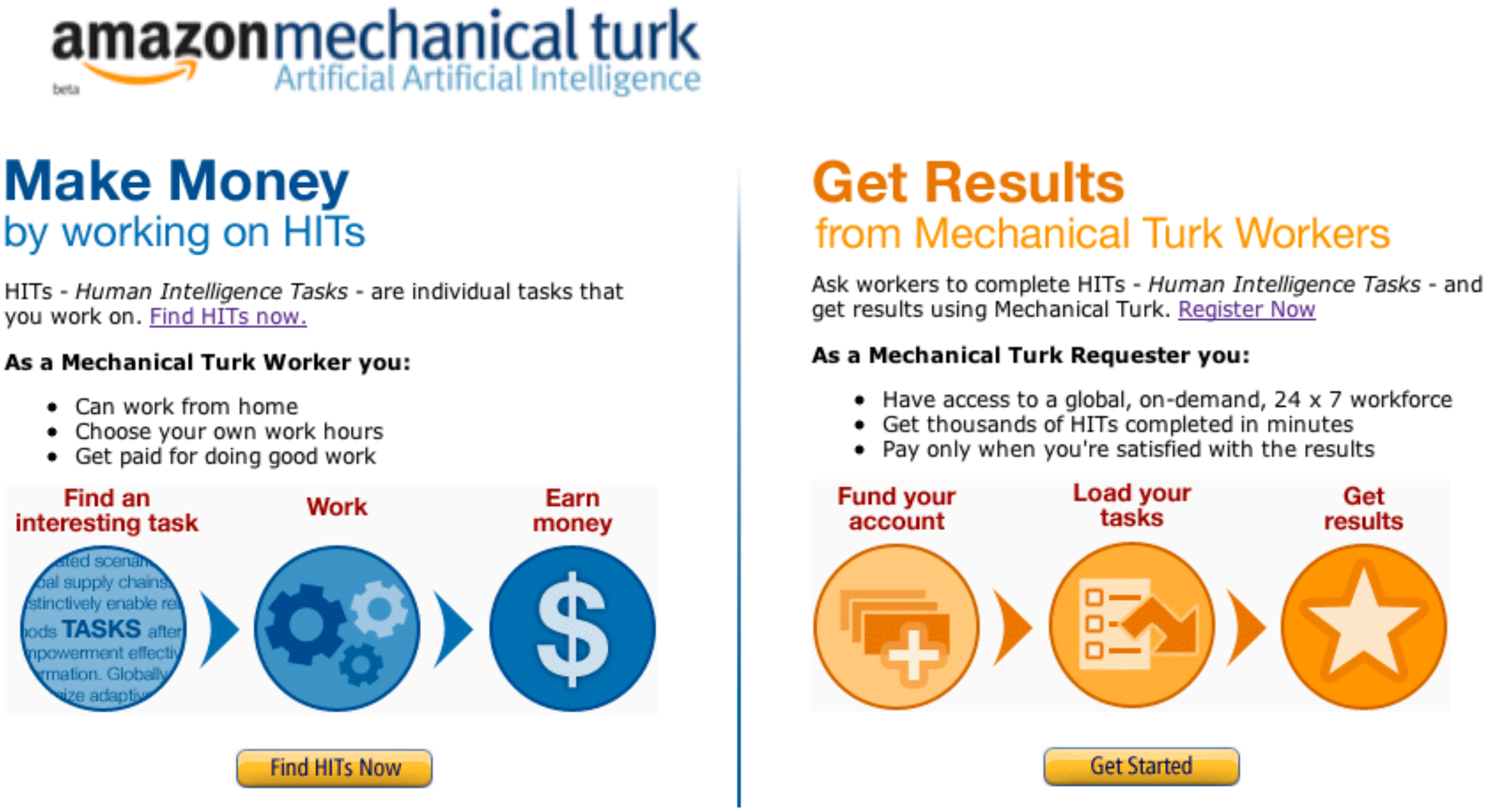
THE SHEEP MARKET

10,000 sheep created
by online workers.
More...

# Case-Study: Amazon MTurk

- Micro-task crowdsourcing marketplace
- On-demand, scalable, real-time workforce
- Online since 2005 (still in “beta”)
- Currently the most popular platform
- Developer’s API as well as GUI

# Amazon MTurk



## 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

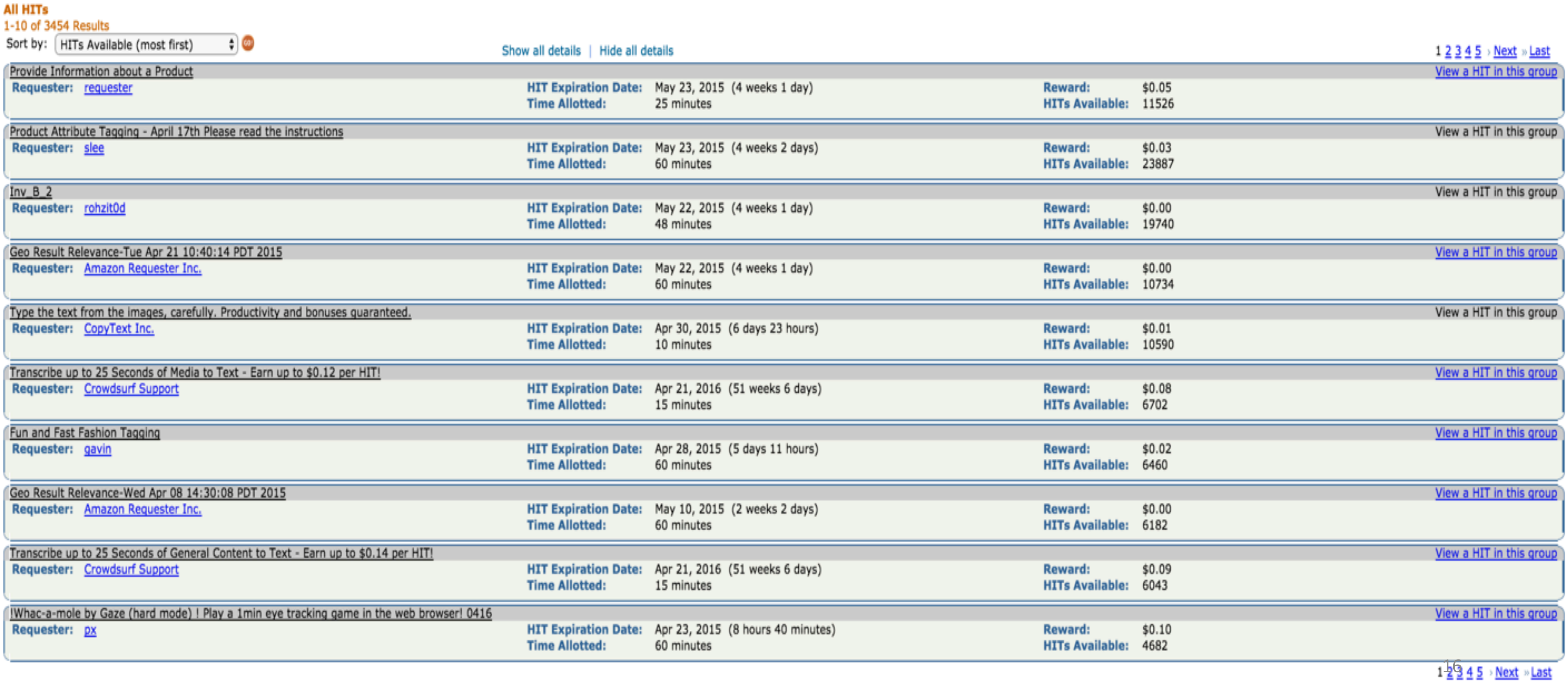
## Get Results
from Mechanical Turk Workers

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

# 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) | Reward: $0.05
Time Allotted: 25 minutes | 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) | Reward: $0.03
Time Allotted: 60 minutes | HITs Available: 23887

Inv_B_2 — View a HIT in this group
Requester: rohzit0d | HIT Expiration Date: May 22, 2015 (4 weeks 1 day) | Reward: $0.00
Time Allotted: 48 minutes | 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) | Reward: $0.00
Time Allotted: 60 minutes | 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) | Reward: $0.01
Time Allotted: 10 minutes | 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) | Reward: $0.08
Time Allotted: 15 minutes | HITs Available: 6702

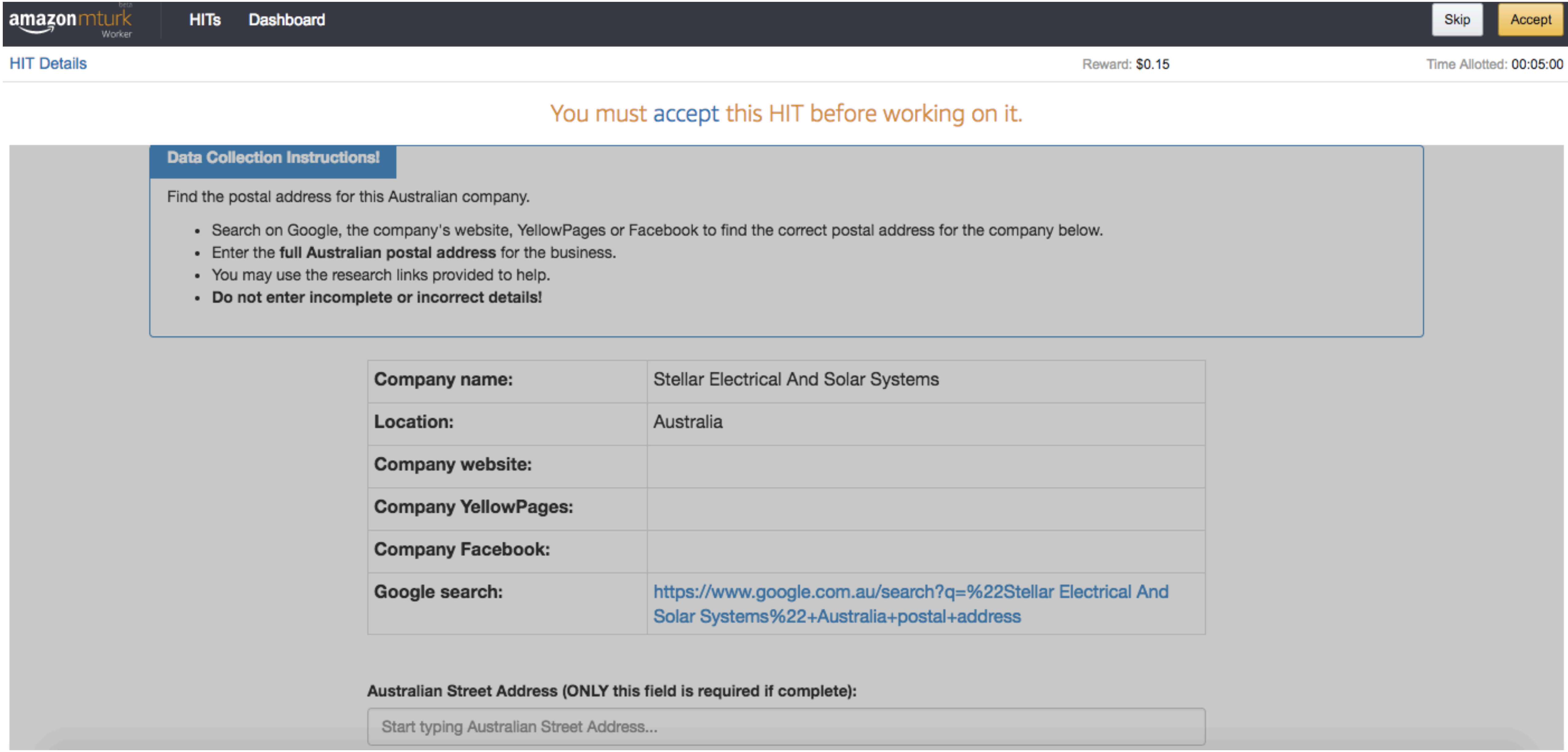
Fun and Fast Fashion Tagging — View a HIT in this group
Requester: gavin | HIT Expiration Date: Apr 28, 2015 (5 days 11 hours) | Reward: $0.02
Time Allotted: 60 minutes | HITs Available: 6460

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) | Reward: $0.00
Time Allotted: 60 minutes | 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) | Reward: $0.09
Time Allotted: 15 minutes | 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) | Reward: $0.10
Time Allotted: 60 minutes | HITs Available: 4682

1 2 3 4 5 › Next » Last



HIT Details

Reward: $0.15

Time Allotted: 00:05:00

You must accept this HIT before working on it.

Data Collection Instructions!

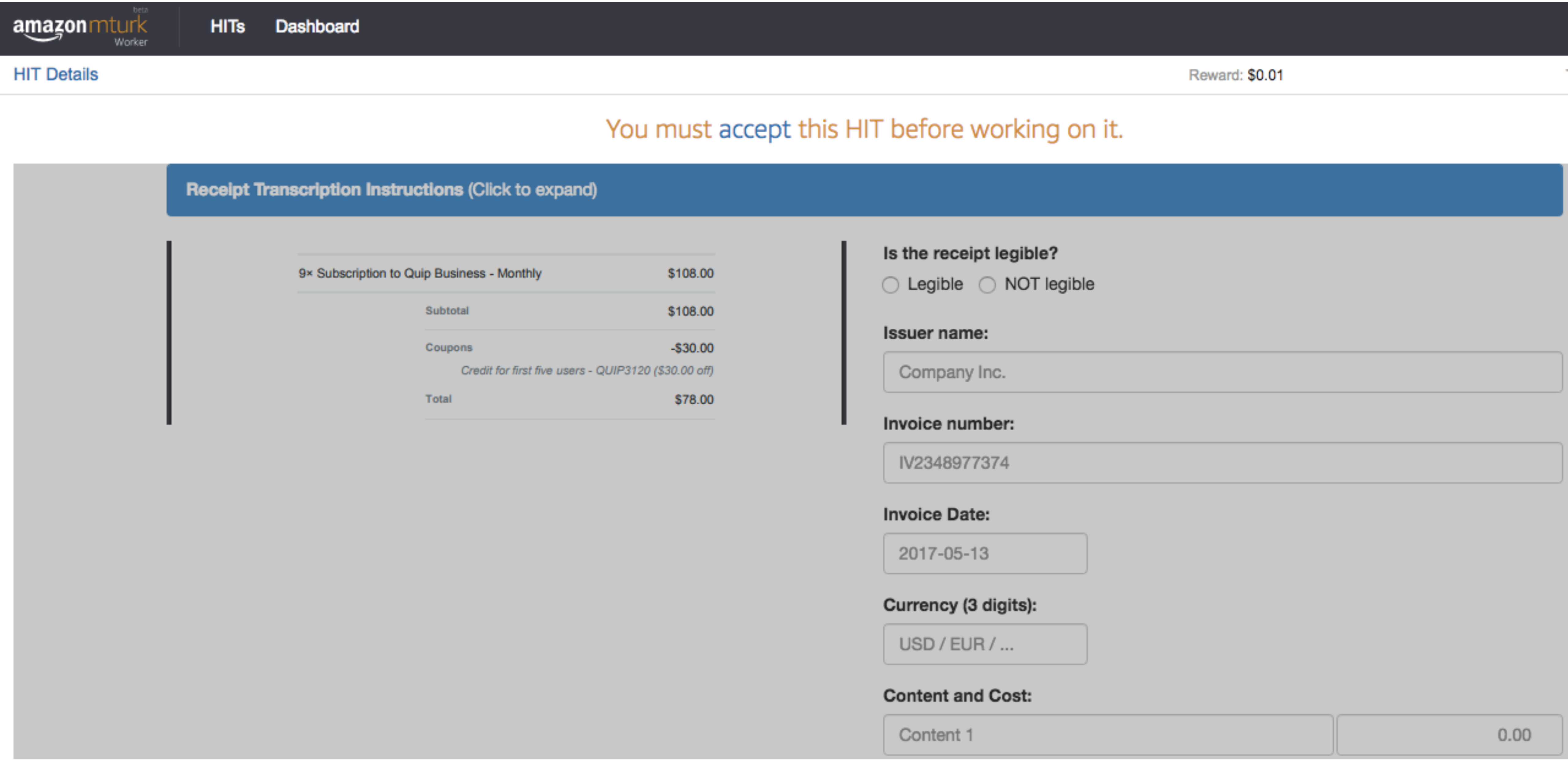
Find the postal address for this Australian company.

- 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...



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

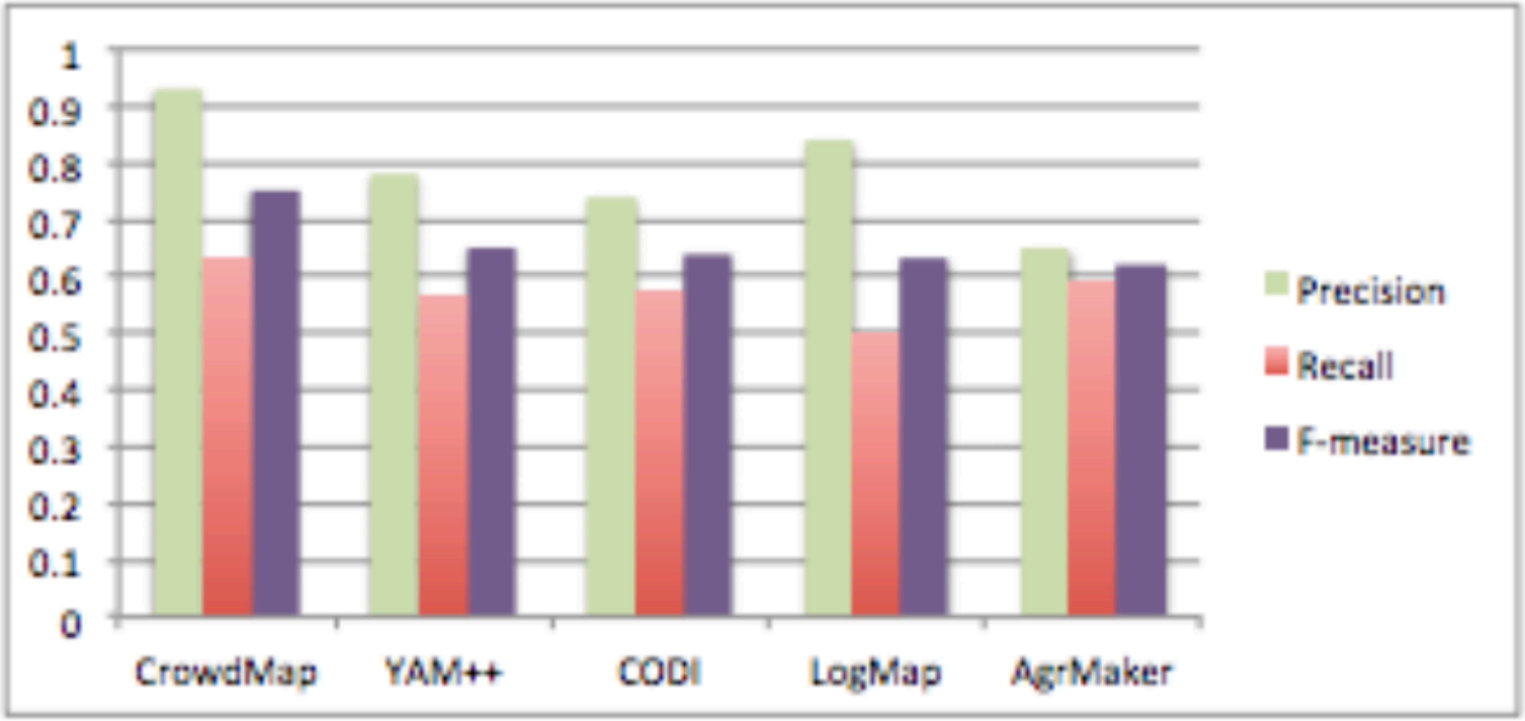
# Crowdsourcing Ontology Mapping

- Find a set of mappings between two ontologies
- Micro-tasks:
    - Verify/identify a mapping relationships:
        - Is concept A the same as concept B
        - A is a kind of B
        - B is a kind of A
        - No relation

Cristina Sarasua, Elena Simperl, and Natalya F. Noy. CROWDMAP: Crowdsourcing Ontology Alignment with Microtasks. In: International Semantic Web Conference 2012, Boston, MA, USA.

# Crowdsourcing Ontology Mapping

- Crowd-based outperforms purely automatic approaches

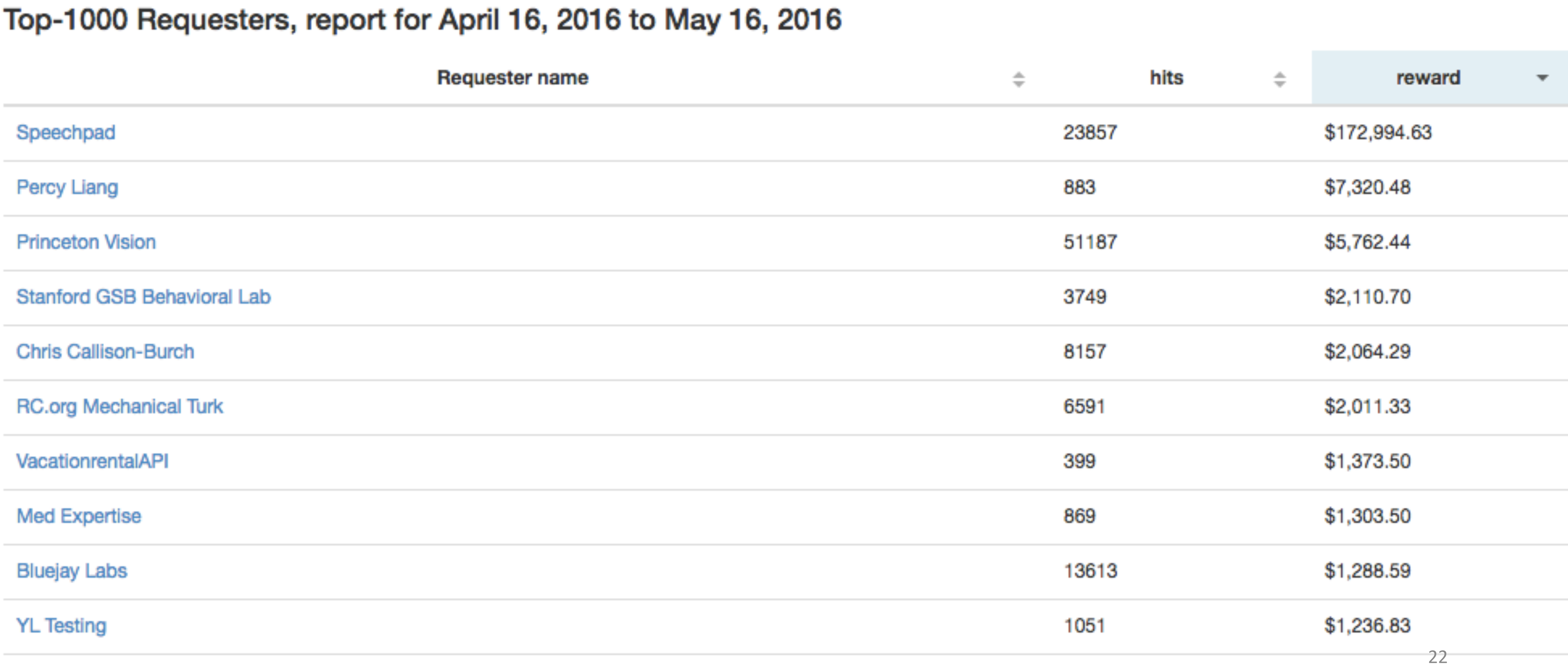


# Crowdsourcing Ontology Engineering

- Ask the crowd to create/verify subClassOf relations
  - "Car" is a "vehicle"
- Does it work for domain specific ontologies?
  - A "protandrous hermaphroditic organism" is a "sequential hermaphroditic organism"
- Workers perform worse than experts
- Workers presented with concept definitions perform as good as experts

Jonathan Mortensen, Mark A. Musen, Natasha F. Noy: Crowdsourcing the Verification of Relationships in Biomedical Ontologies. AMIA 2013

# MTurk is a Marketplace for HITs



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

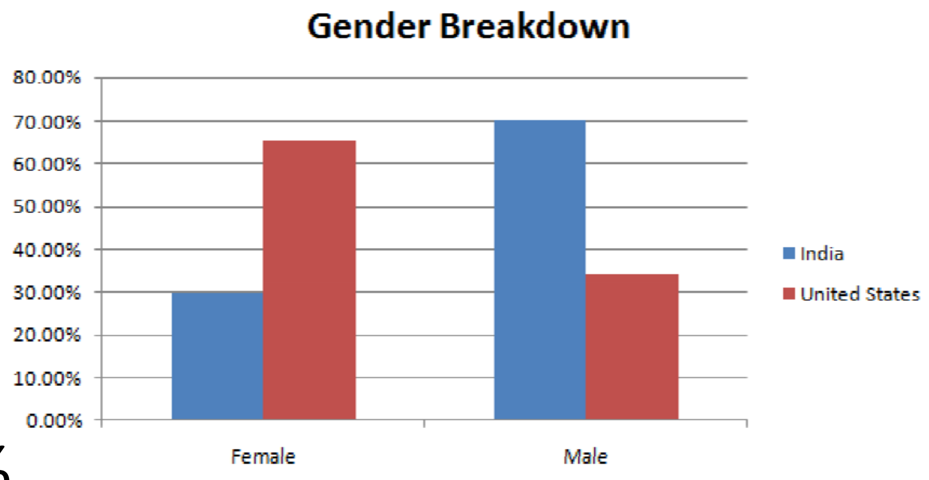
| Requester name | hits | reward |
|---|---|---|
| Speechpad | 23857 | $172,994.63 |
| Percy Liang | 883 | $7,320.48 |
| Princeton Vision | 51187 | $5,762.44 |
| Stanford GSB Behavioral Lab | 3749 | $2,110.70 |
| Chris Callison-Burch | 8157 | $2,064.29 |
| RC.org Mechanical Turk | 6591 | $2,011.33 |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

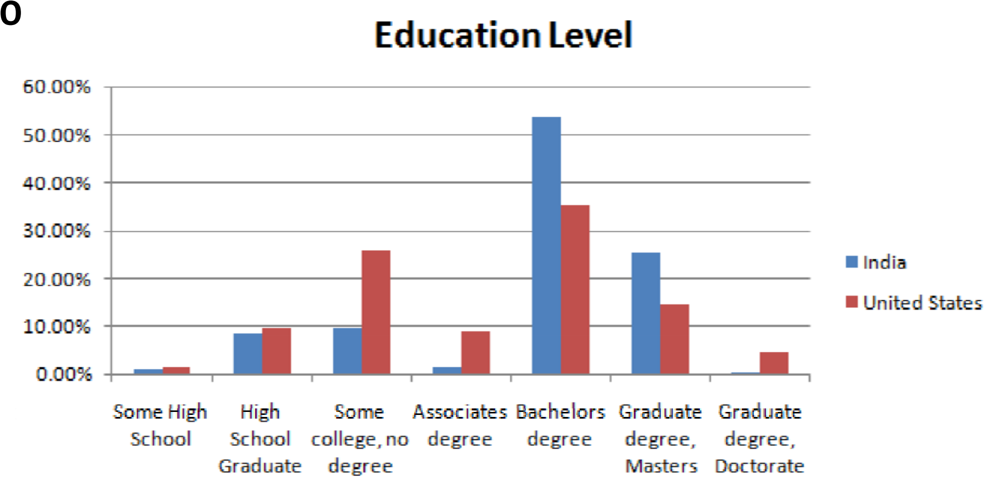
# Demographics of MTurk workers in 2009

Country of residence

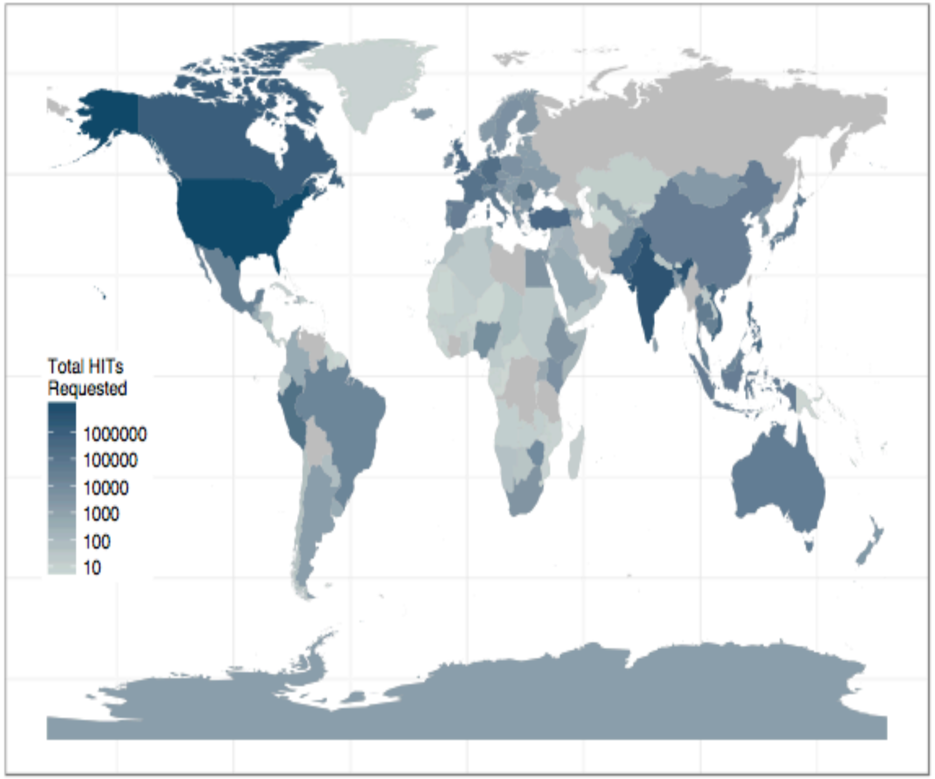
- United States: 46.80%
- India: 34.00%
- Miscellaneous: 19.20%

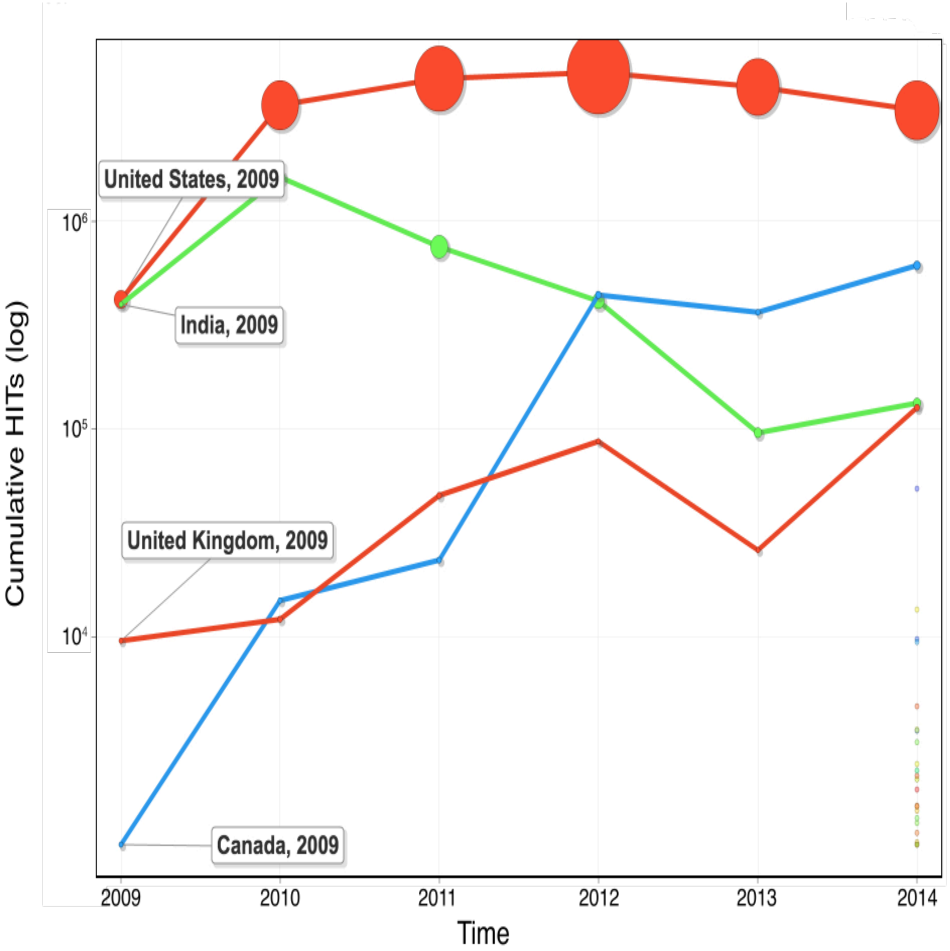
2013 Statistics:
1M workers
10% active





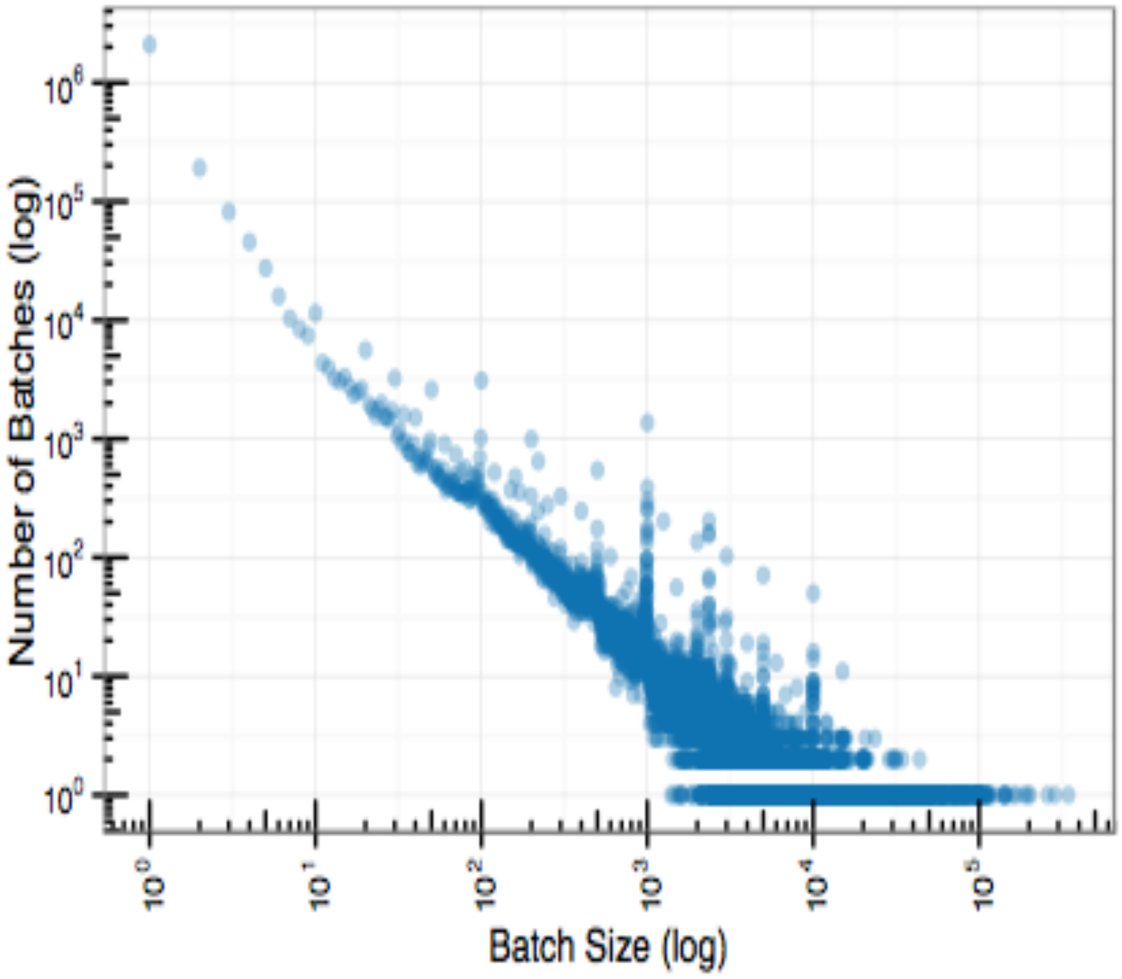
# Requested Workers





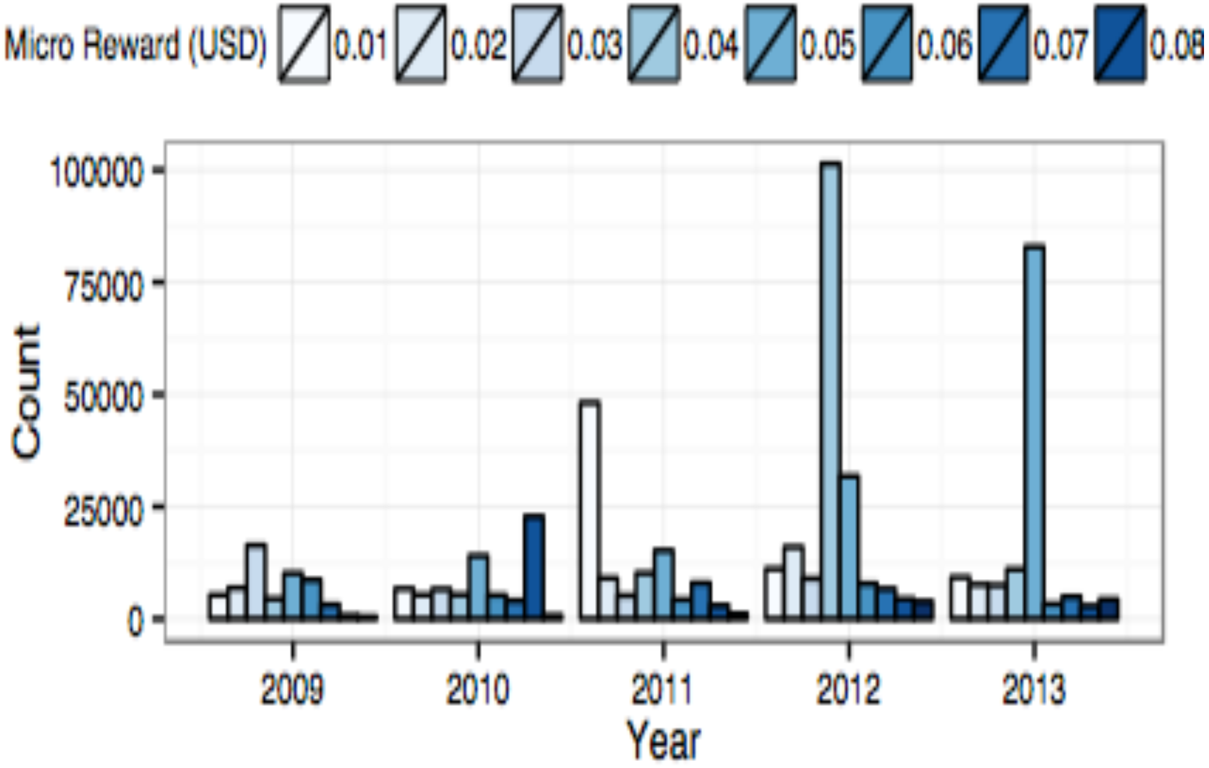
Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. **The Dynamics of Micro-Task Crowdsourcing -- The Case of Amazon MTurk**. In: 24th International Conference on World Wide Web (WWW 2015), Research Track. Firenze, Italy, May 2015.

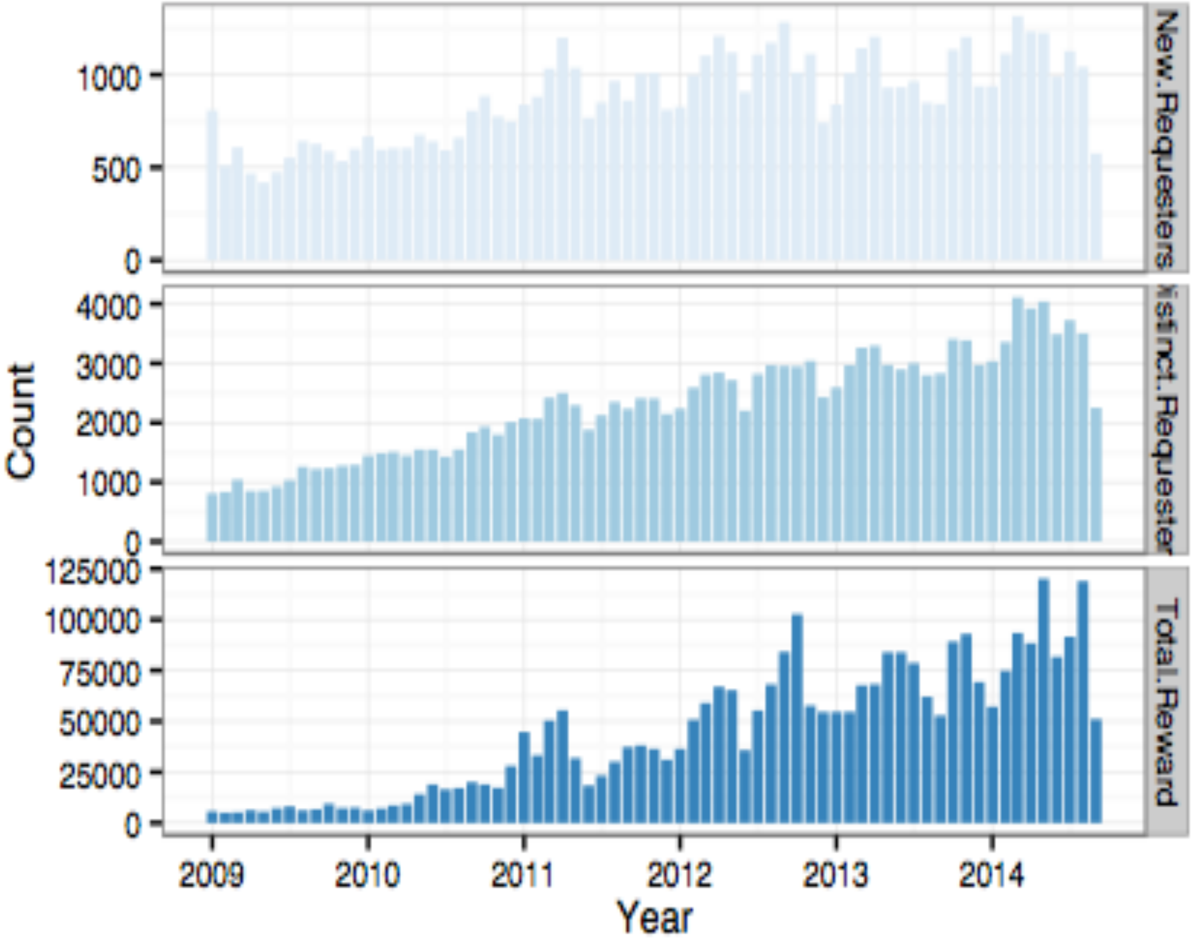
# Distribution of *Batch Size*



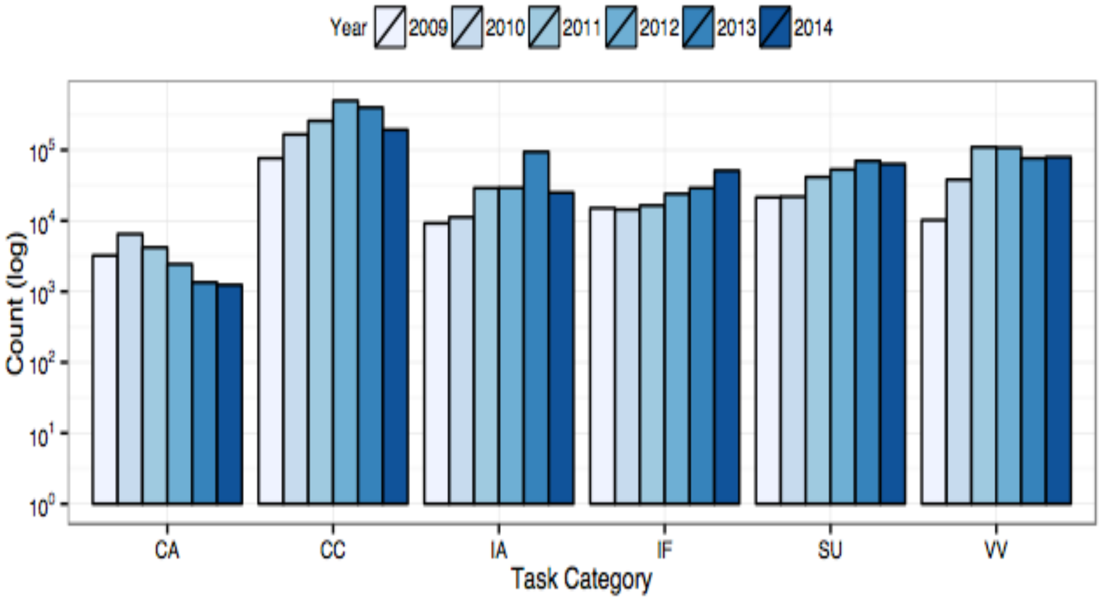
“Power-law”

# Reward Distribution





# Distribution of HIT Types



Less Content Access batches

Content Creation being the most popular

# Some findings of our longitudinal study

- HIT reward has increased over time
- Audio transcription is the most popular task
- Demand for Indian workers has decreased
- Surveys are most popular for US workers
- 1000 new requesters per month join
- 10K new HITs arrive and 7.5K HITs get completed every hour

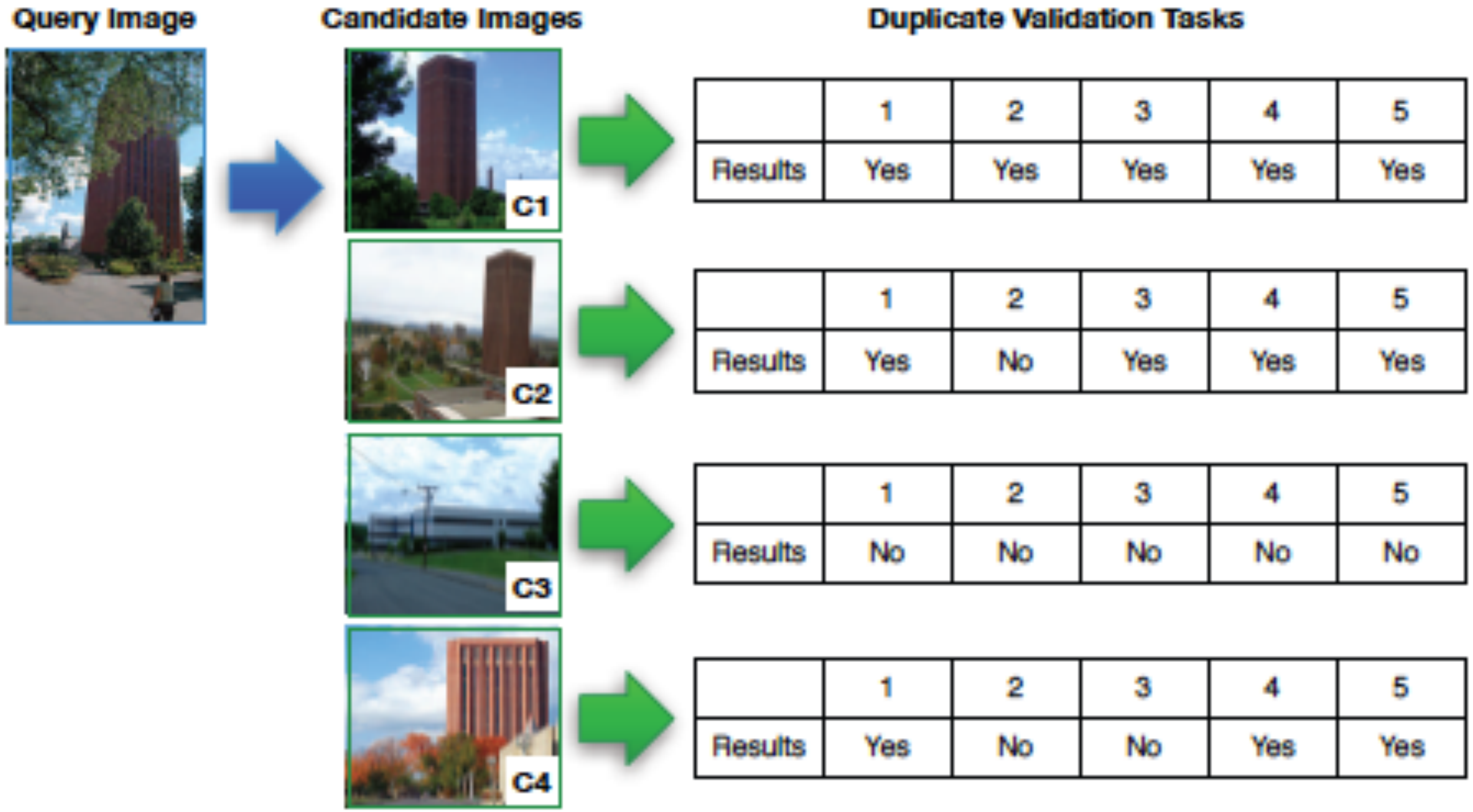
- Check **#mturkdynamics** for the main findings

# 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



| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Results | Yes | Yes | Yes | Yes | Yes |

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Results | Yes | No | Yes | Yes | Yes |

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Results | No | No | No | No | No |

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Results | Yes | No | No | Yes | Yes |

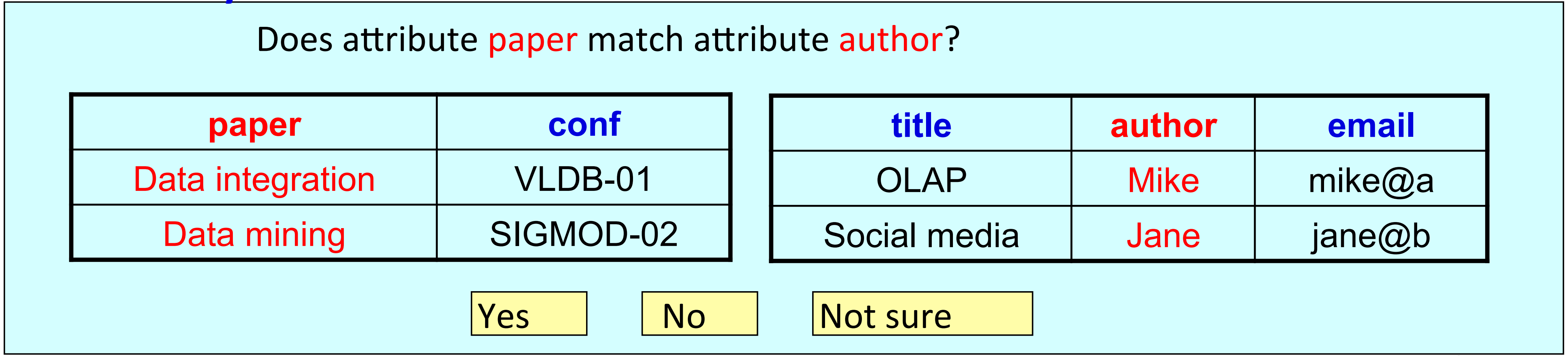
Yan, Kumar, Ganesan, CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones, Mobisys 2010.

# Example: Hybrid Data Integration

| paper | conf |
|---|---|
| Data integration | VLDB-01 |
| Data mining | SIGMOD-02 |

| title | author | email | venue |
|---|---|---|---|
| OLAP | Mike | mike@a | ICDE-02 |
| Social media | Jane | jane@b | PODS-05 |

- **Generate plausible matches**
  - paper = title, paper = author, paper = email, paper = venue
  - conf = title, conf = author, conf = email, conf = venue
- **Ask users to verify**



Does attribute paper match attribute author?

| paper | conf |
|---|---|
| Data integration | VLDB-01 |
| Data mining | SIGMOD-02 |

| title | author | email |
|---|---|---|
| OLAP | Mike | mike@a |
| Social media | Jane | jane@b |

McCann, Shen, Doan: Matching Schemas in Online Communities. ICDE, 2008

# CrowdDB

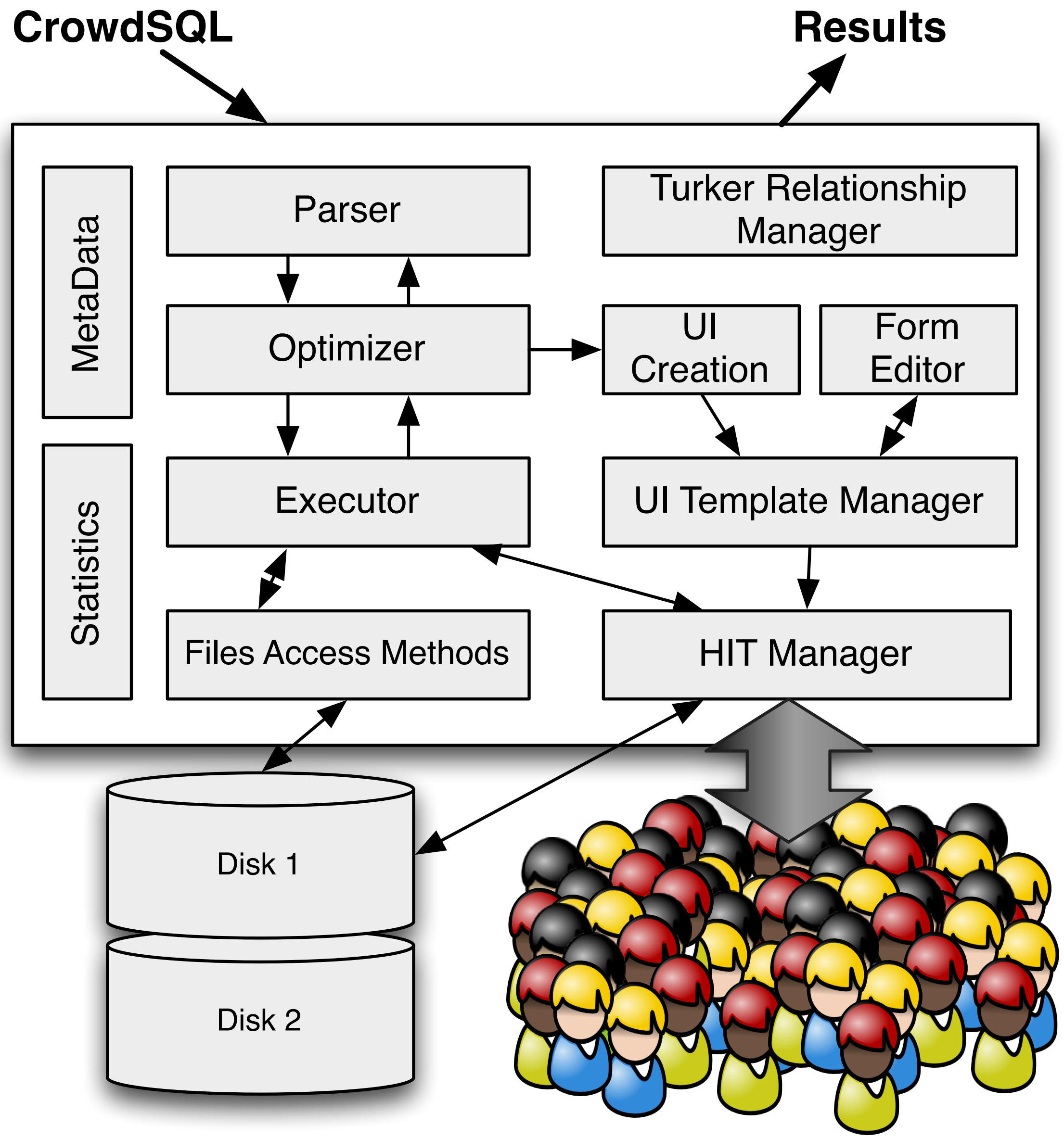
**Use the crowd to answer DB-hard queries**

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*

# Crowd DB query

The Vovlo S80 is the flagship model of this brand…

Is the review positive?

Which one is better?

```
SELECT review
FROM car_review
WHERE sentiment ~= "pos";
```

```
SELECT image i
FROM car_image
WHERE subject = "Volvo S60"
ORDER BY CROWDORDER(”clarity");
```

# CrowdDB – Missing Data

## Missing Columns

| review | make | model | sentiment |
|---|---|---|---|
| xxx | Volvo | S80 | ? |

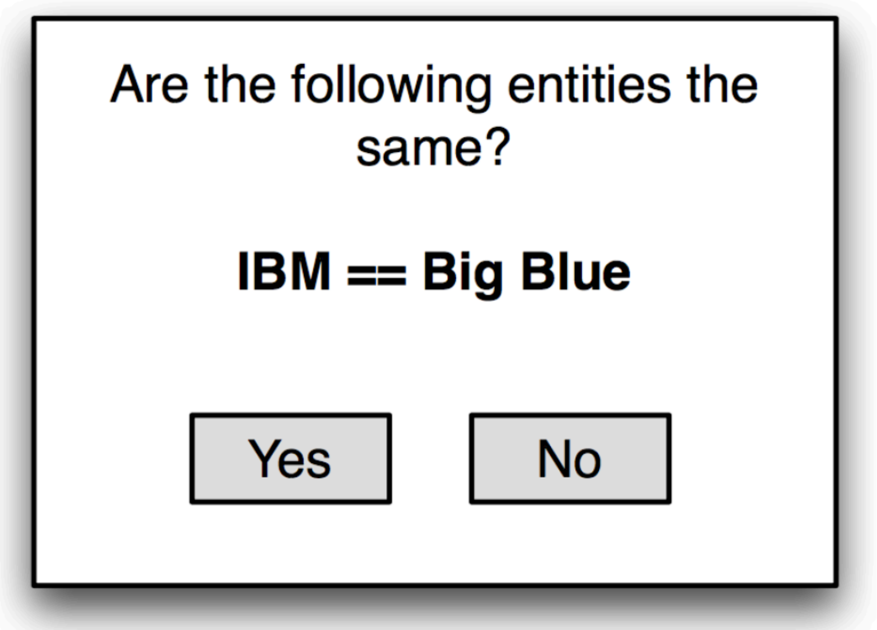
```
CREATE TABLE car_review
(
  review STRING,
  make CROWD STRING,
  model CROWD STRING,
  sentiment CROWD STRING
);
```

## Missing Tuples

| make | model | style | color |
|---|---|---|---|
| ? | ? | ? | ? |

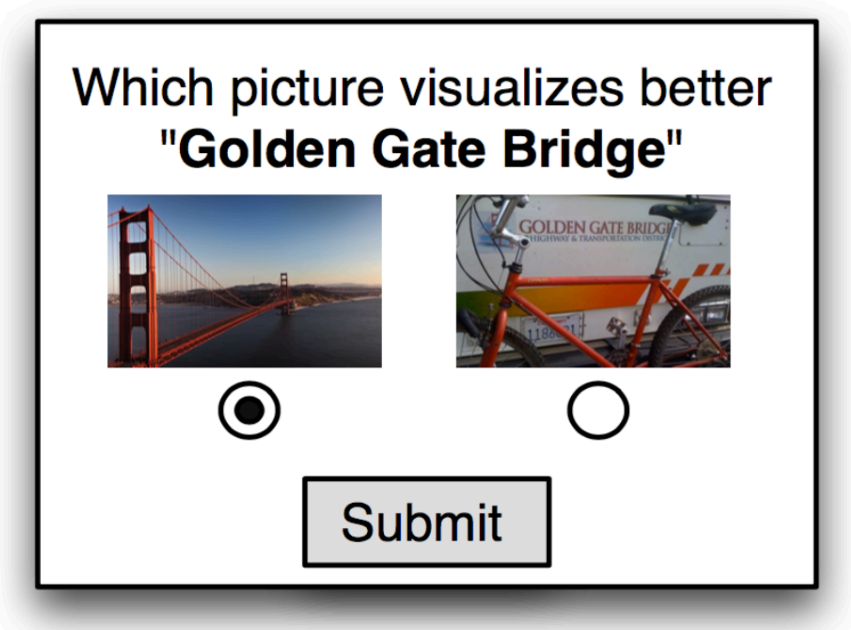
```
CREATE CROWD TABLE car
(
  make STRING,
  model STRING,
  color STRING,
  style STRING,
  PRIMARY KEY (make, model)
);
```

# CrowdDB - Joins and Sorts



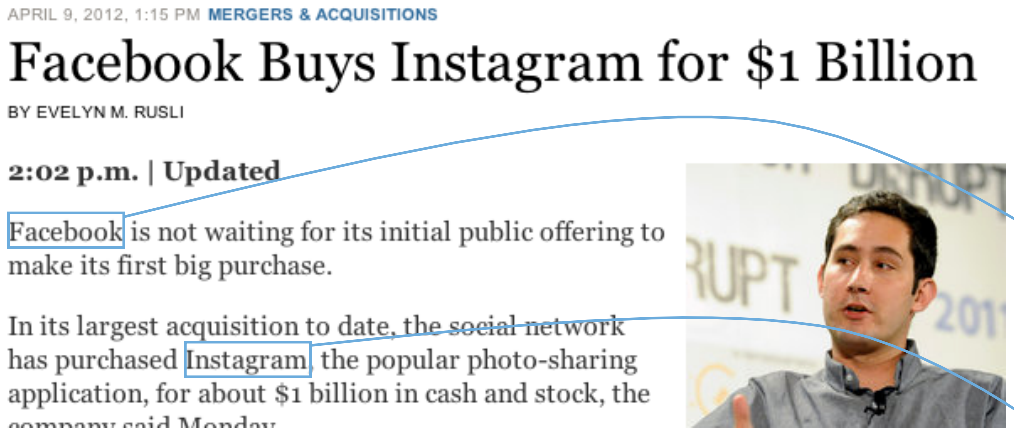
Are the following entities the same?

**IBM == Big Blue**



Which picture visualizes better "**Golden Gate Bridge**"

# Crowdsourcing for Entity Linking



APRIL 9, 2012, 1:15 PM MERGERS & ACQUISITIONS

Facebook Buys Instagram for $1 Billion

BY EVELYN M. RUSLI

**2:02 p.m. | Updated**

Facebook is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.

http://dbpedia.org/resource/Facebook

http://dbpedia.org/resource/Instagram

HTML:

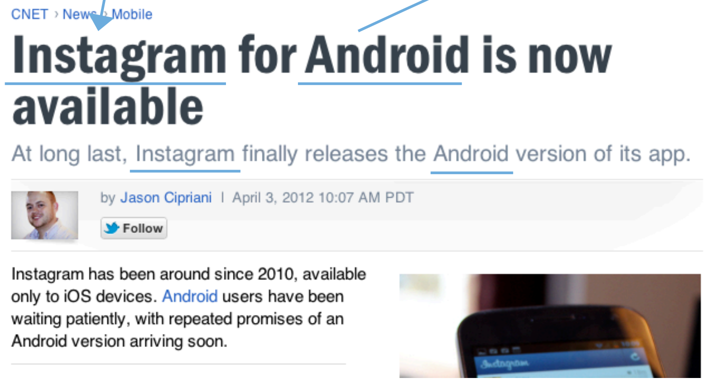
```
<p>Facebook is not waiting for its initial
public offering to make its first big
purchase.</p><p>In its largest
acquisition to date, the social network
has purchased Instagram, the popular
photo-sharing application, for about $1
billion in cash and stock, the company
said Monday.</p>
```

RDFa enrichment

```
<p><span about="http://dbpedia.org/resource/
Facebook"><cite property=”rdfs:label">Facebook</
cite> is not waiting for its initial public offering to
make its first big purchase.</span></p><p><span
about="http://dbpedia.org/resource/Instagram">In
its largest acquisition to date, the social network has
purchased <cite property=”rdfs:label">Instagram</
cite> , the popular photo-sharing application, for
about $1 billion in cash and stock, the company said
Monday.</span></p>
```

owl:sameAs

fbase:Instagram



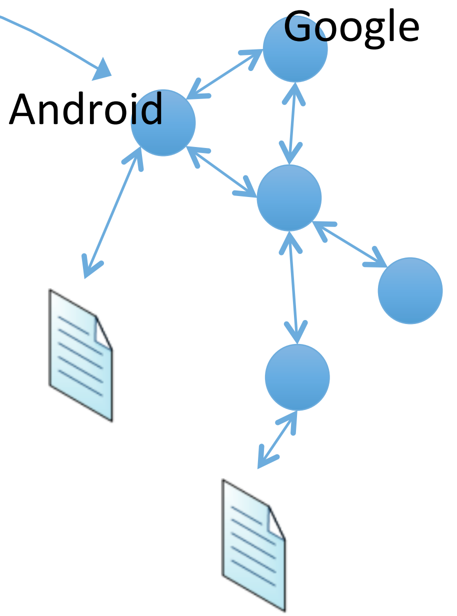
CNET › News › Mobile

Instagram for Android is now available

At long last, Instagram finally releases the Android version of its app.

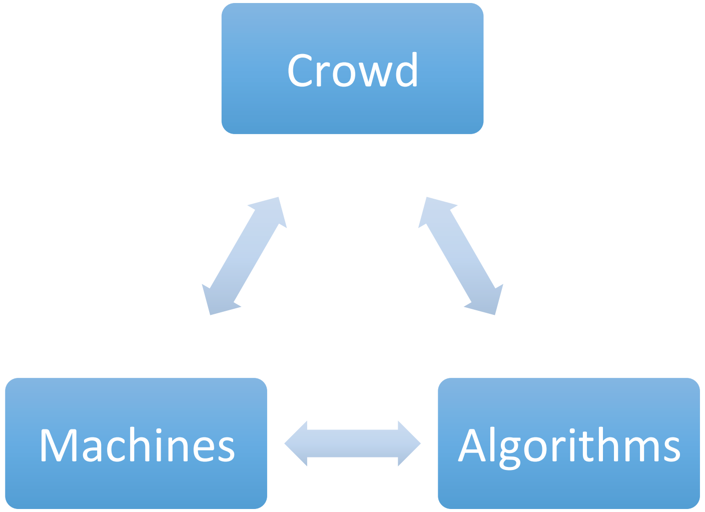
by Jason Cipriani | April 3, 2012 10:07 AM PDT

Instagram has been around since 2010, available only to iOS devices. Android users have been waiting patiently, with repeated promises of an Android version arriving soon.



# ZenCrowd

- Combine both algorithmic and manual linking
- Automate manual linking via crowdsourcing
- Dynamically assess human workers with a probabilistic reasoning framework

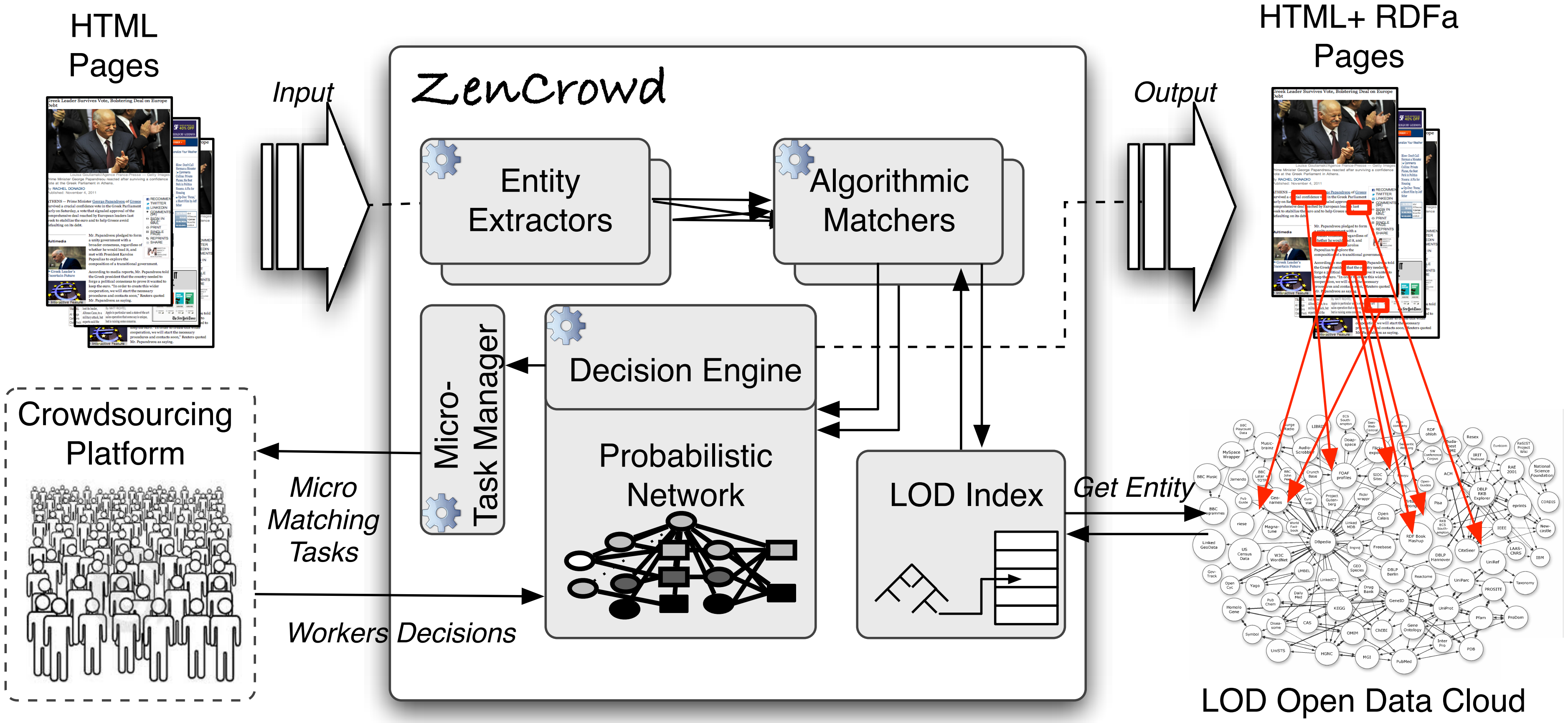


Crowd

Machines

Algorithms

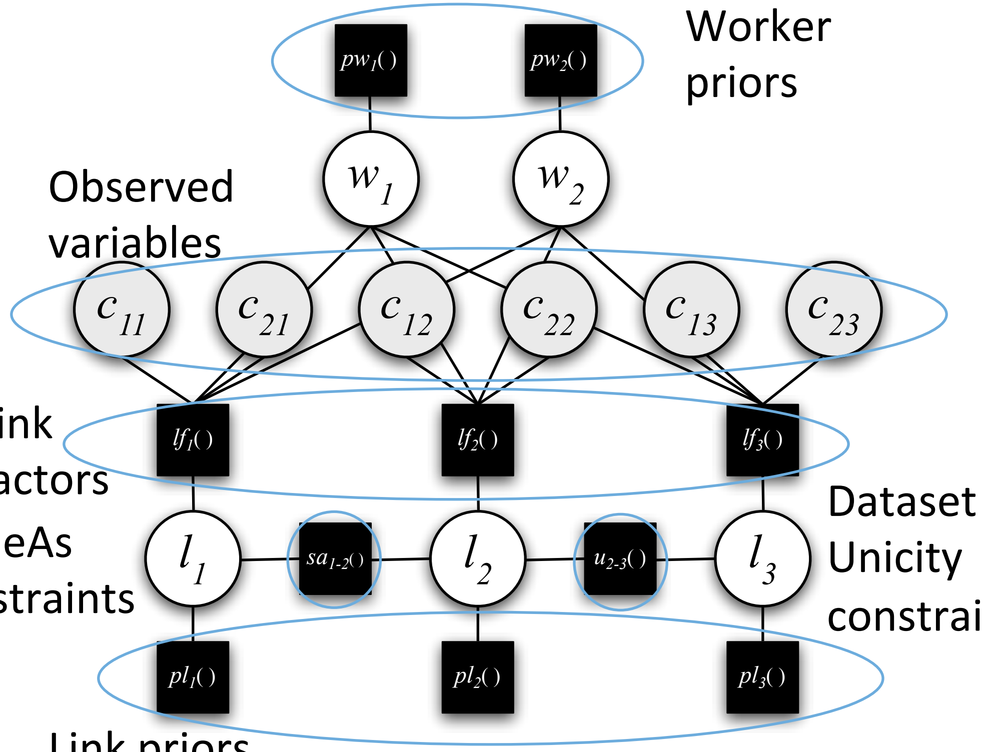
# ZenCrowd Architecture



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**).

# Entity Factor Graphs

- Graph components
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints
- Probabilistic Inference
  - Select all links with posterior prob >τ



Worker priors

Observed variables

Link factors

SameAs constraints

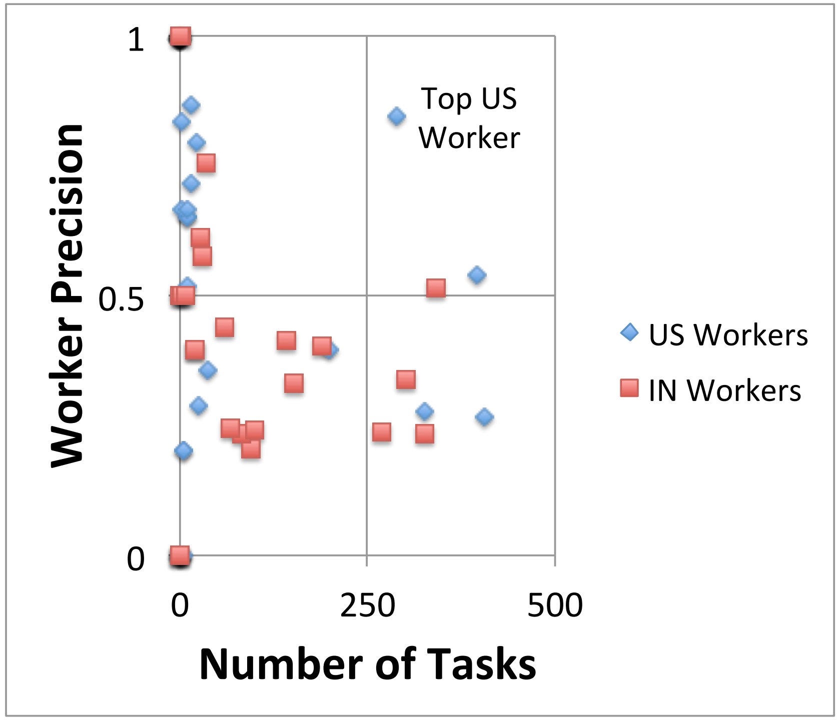
Dataset Unicity constraints

Link priors

2 workers, 6 clicks, 3 candidate links

# Experimental Evaluation

- Worker Selection



# ZenCrowd Summary

- ZenCrowd: Probabilistic reasoning over automatic and crowdsourcing methods for entity linking
- Standard crowdsourcing improves 6% over automatic
- 4% - 35% improvement over standard crowdsourcing
- 14% average improvement over automatic approaches

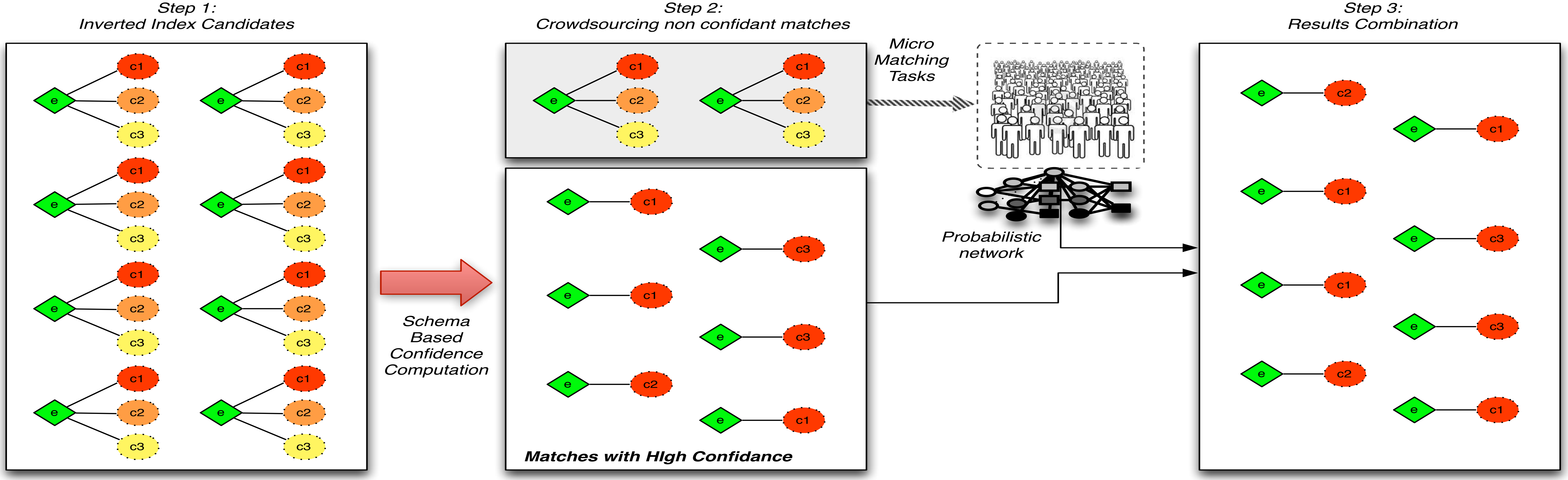
- Follow up-work (VLDBJ, 2013):
  - Also used for **instance matching** across datasets
  - 3-way blocking with the crowd

# *Blocking* for Instance Matching

- Find the instances about the same real-world entity within two datasets
- Avoid Comparison of all possible pairs
  - Step 1: cluster similar items using a cheap similarity measure
  - Step 2: n*n comparison within the clusters with an expensive measure

# 3-steps Blocking with the Crowd

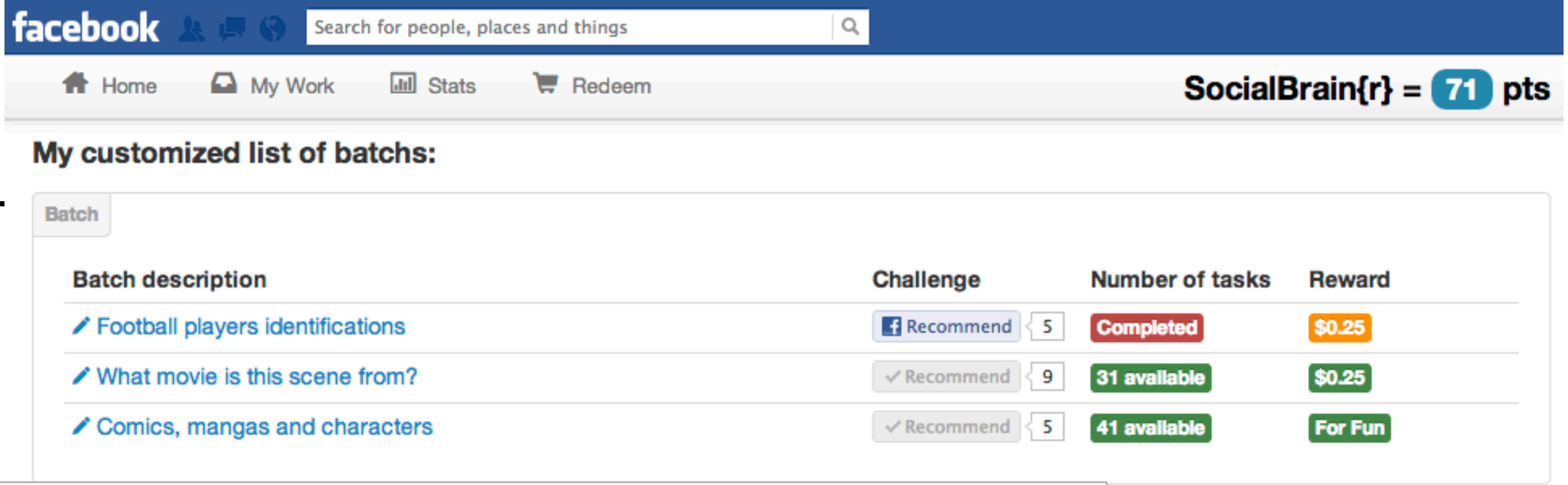
- Crowdsourcing as the most expensive similarity measure


Step 1:
Inverted Index Candidates
Step 2:
Crowdsourcing non confidant matches
Step 3:
Results Combination
Micro Matching Tasks
Probabilistic network
Schema Based Confidence Computation
Matches with HIgh Confidance
e
c1
c2
c3

# Lessons Learnt

- Crowdsourcing + Prob reasoning works!
- But
  - Different worker communities perform differently
  - Many low quality workers
  - Completion time may vary (based on reward)
- Need to **find the right workers** for your task (see WWW2013 and CHI2015 papers)
- Need to make sure **high priority tasks** are completed fast (see WWW2016 paper)

# Pick-A-

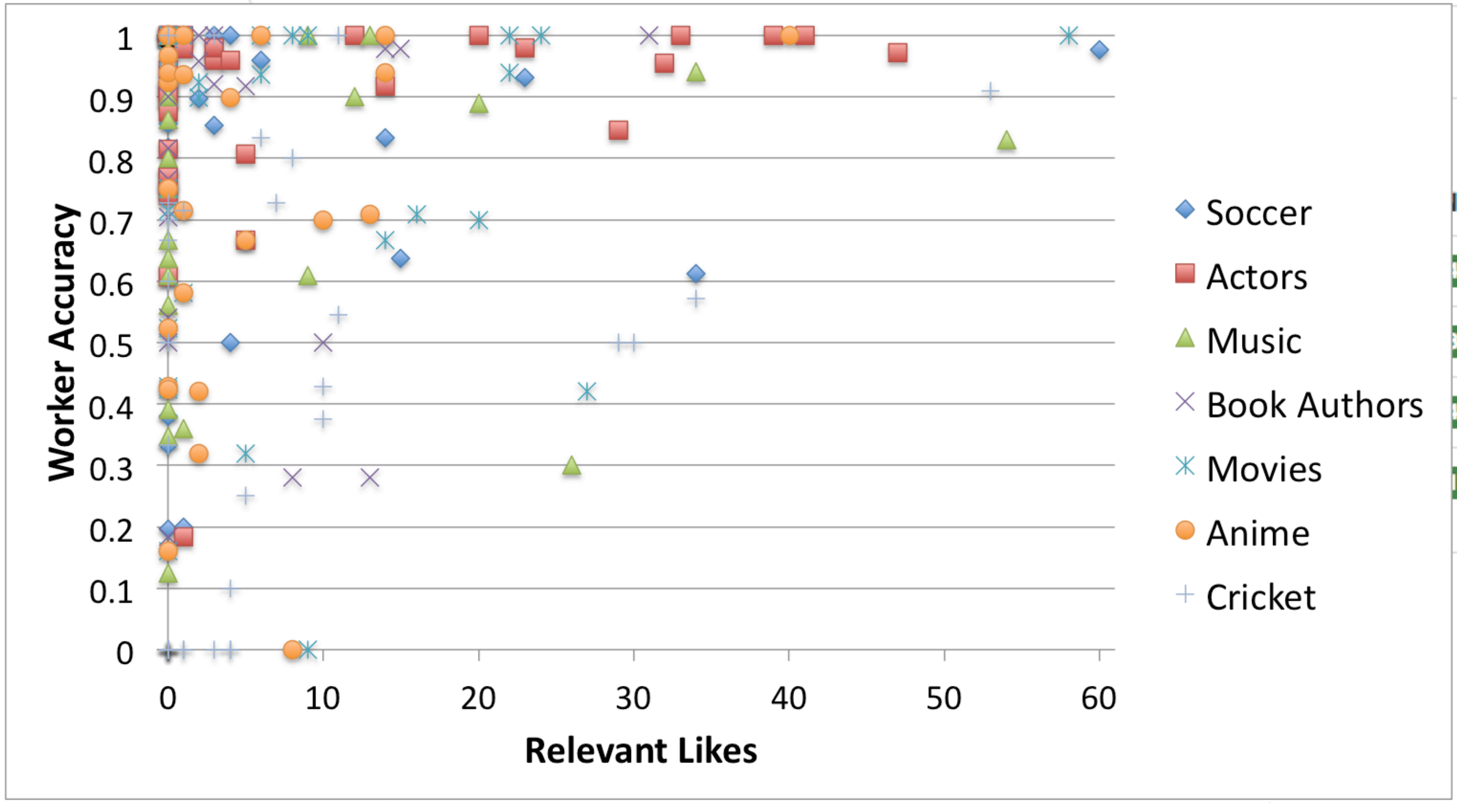


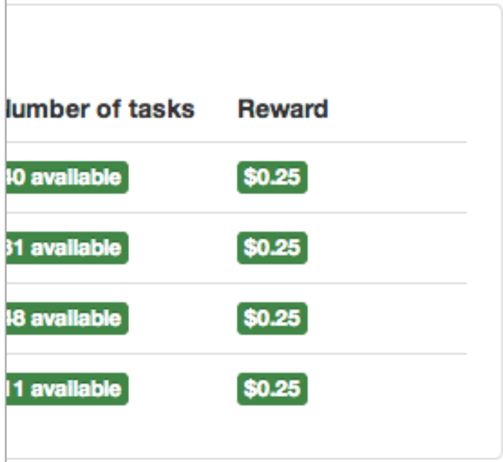
SocialBrain{r} = 71 pts

My customized list of batchs:

Batch

| Batch description | Challenge | Number of tasks | Reward |
|---|---|---|---|
| Football players identifications | Recommend 5 | Completed | $0.25 |
| What movie is this scene from? | ✓ Recommend 9 | 31 available | $0.25 |
| Comics, mangas and characters | ✓ Recommend 5 | 41 available | For Fun |





| lumber of tasks | Reward |
|---|---|
| 0 available | $0.25 |
| 1 available | $0.25 |
| 8 available | $0.25 |
| 1 available | $0.25 |

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do. In: **WWW2013**

# Behavioral Patterns of Malicious Workers

**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).

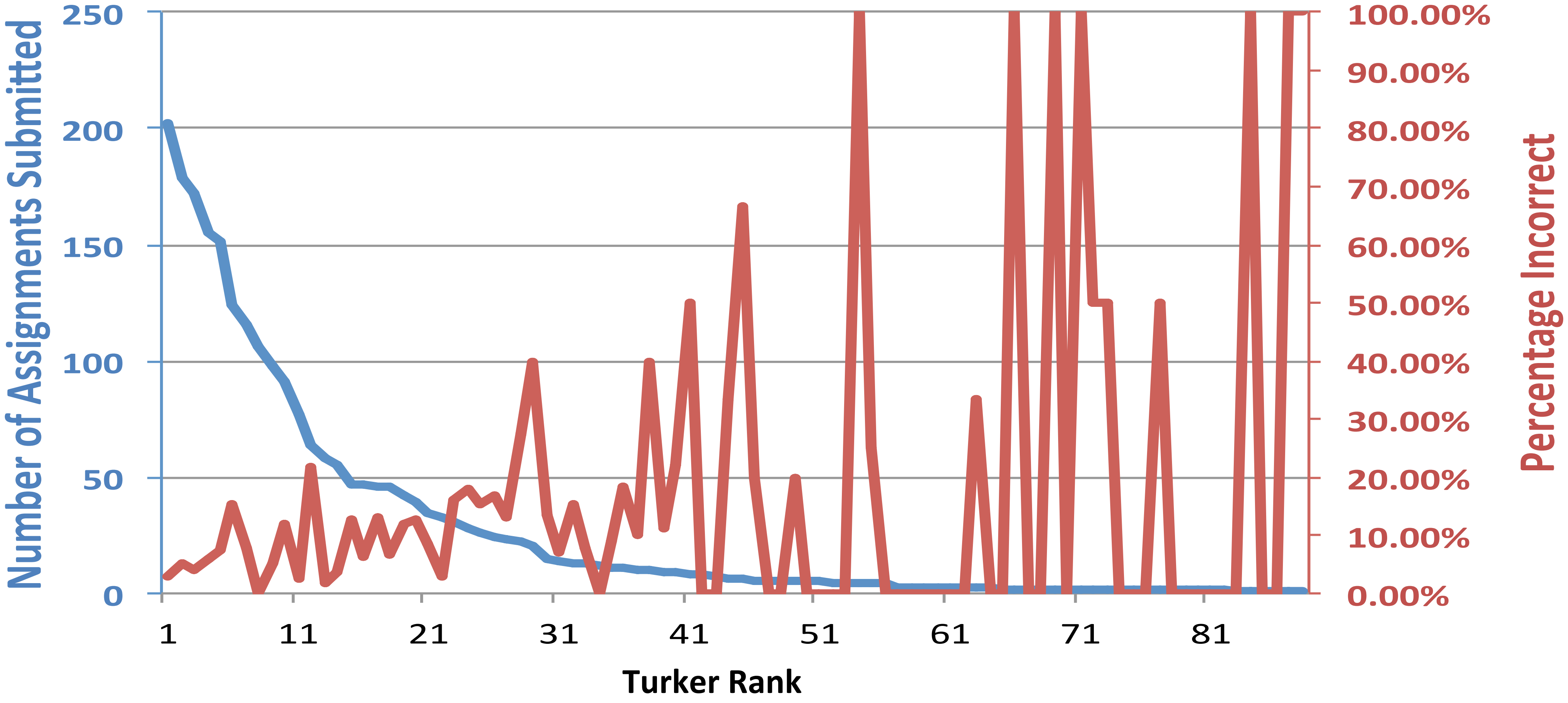
Response: *'one, two, three, four, five'*

**Gold Standard Preys (GSP)**

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

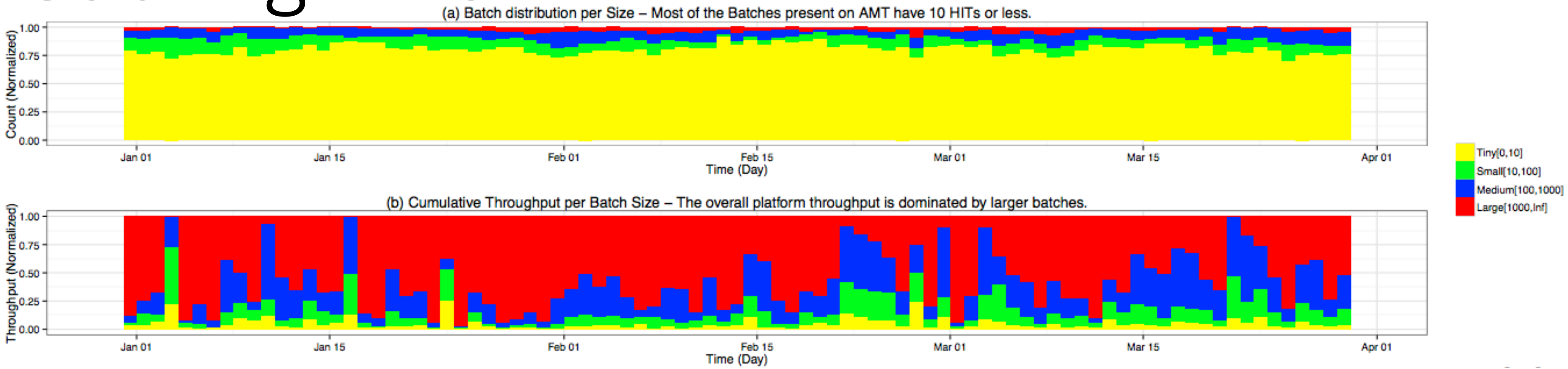
Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. **Understanding Malicious Behaviour in Crowdsourcing Platforms**: The Case of Online Surveys. In: Proceedings of the ACM Special Interest Group on Computer Human Interaction (**CHI 2015**).

# Turker Contribution and Errors

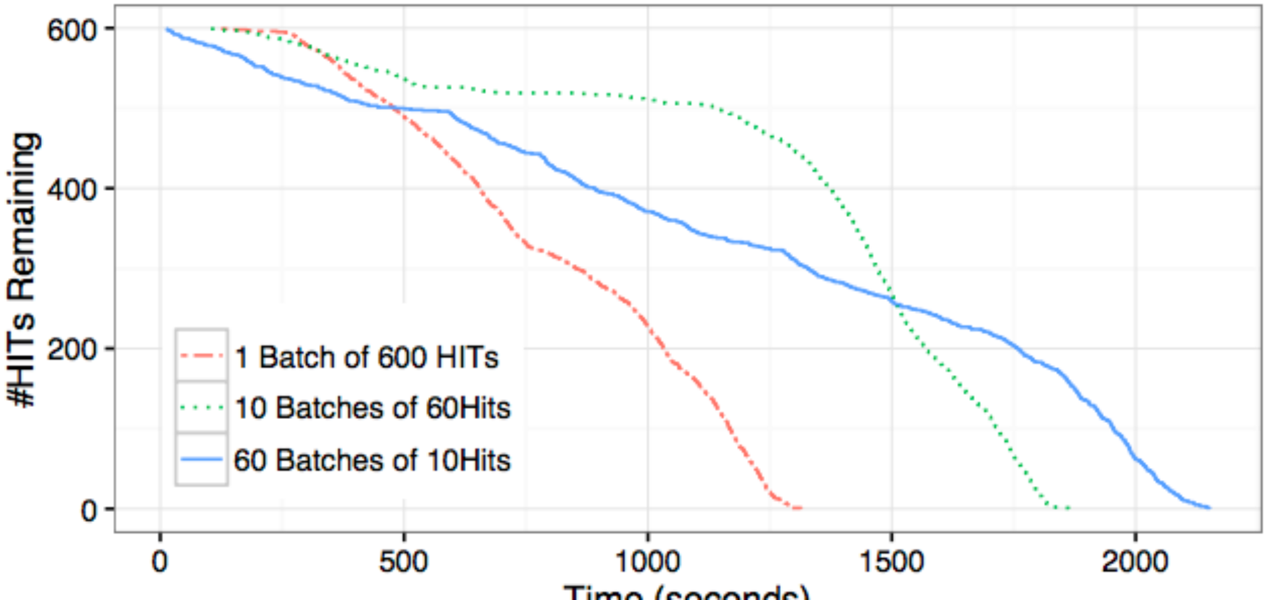


[Franklin, Kossmann, Kraska, Ramesh, Xin: CrowdDB: Answering Queries with Crowdsourcing. *SIGMOD*,2011]

# Scheduling HITs



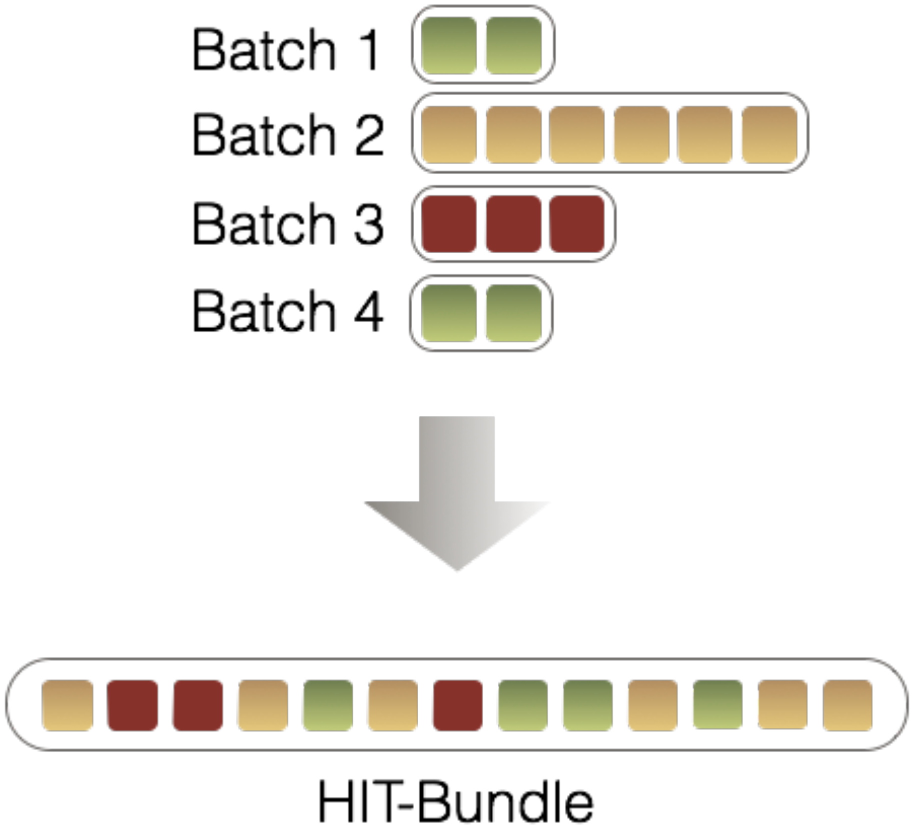
Platform throughput is dominated by large HIT batches



# HIT-Bundle

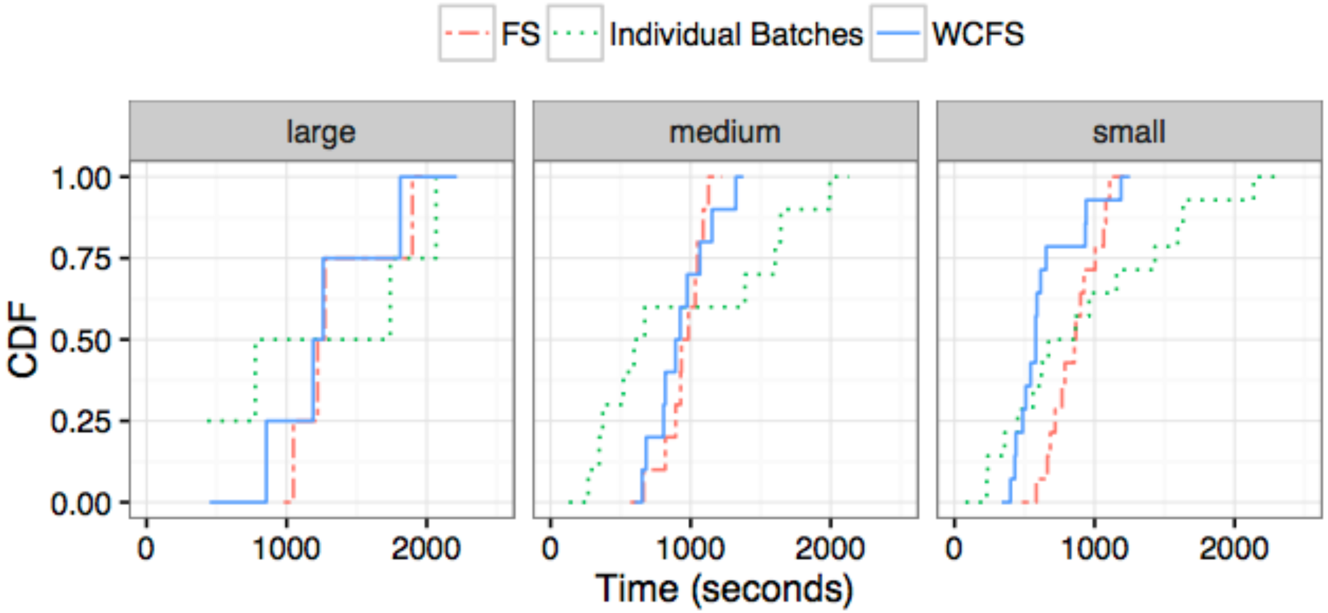
## Definition

- Scheduling requires control over the serving process of tasks
- A **HIT-Bundle** is a batch that contains heterogeneous tasks
- All tasks that are generated by the system are published through the HIT-Bundle



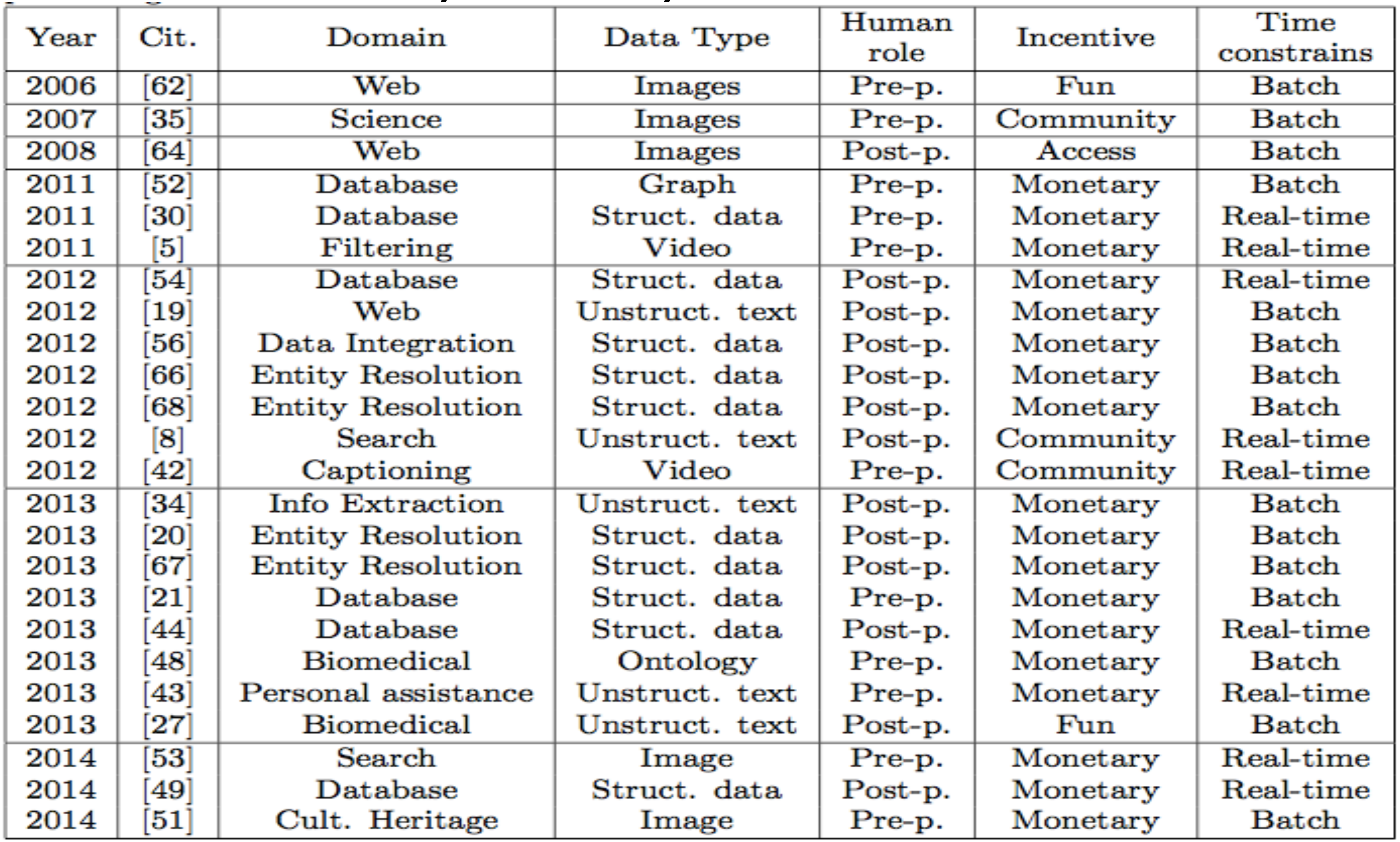
# Scheduling HITs

- Fair Scheduling
  - Priority of HITs but avoid starvation
  - Assign HITs of the same type (no context switch)



Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Scheduling Human Intelligence Tasks in Multi-Tenant Crowd-Powered Systems. In: 25th International Conference on World Wide Web (**WWW 2016**), Research Track.

# Overview of hybrid systems



| Year | Cit. | Domain | Data Type | Human role | Incentive | Time constrains |
|---|---|---|---|---|---|---|
| 2006 | [62] | Web | Images | Pre-p. | Fun | Batch |
| 2007 | [35] | Science | Images | Pre-p. | Community | Batch |
| 2008 | [64] | Web | Images | Post-p. | Access | Batch |
| 2011 | [52] | Database | Graph | Pre-p. | Monetary | Batch |
| 2011 | [30] | Database | Struct. data | Pre-p. | Monetary | Real-time |
| 2011 | [5] | Filtering | Video | Pre-p. | Monetary | Real-time |
| 2012 | [54] | Database | Struct. data | Post-p. | Monetary | Real-time |
| 2012 | [19] | Web | Unstruct. text | Post-p. | Monetary | Batch |
| 2012 | [56] | Data Integration | Struct. data | Post-p. | Monetary | Batch |
| 2012 | [66] | Entity Resolution | Struct. data | Post-p. | Monetary | Batch |
| 2012 | [68] | Entity Resolution | Struct. data | Post-p. | Monetary | Batch |
| 2012 | [8] | Search | Unstruct. text | Post-p. | Community | Real-time |
| 2012 | [42] | Captioning | Video | Pre-p. | Community | Real-time |
| 2013 | [34] | Info Extraction | Unstruct. text | Post-p. | Monetary | Batch |
| 2013 | [20] | Entity Resolution | Struct. data | Post-p. | Monetary | Batch |
| 2013 | [67] | Entity Resolution | Struct. data | Post-p. | Monetary | Batch |
| 2013 | [21] | Database | Struct. data | Pre-p. | Monetary | Batch |
| 2013 | [44] | Database | Struct. data | Post-p. | Monetary | Real-time |
| 2013 | [48] | Biomedical | Ontology | Pre-p. | Monetary | Batch |
| 2013 | [43] | Personal assistance | Unstruct. text | Pre-p. | Monetary | Real-time |
| 2013 | [27] | Biomedical | Unstruct. text | Post-p. | Fun | Batch |
| 2014 | [53] | Search | Image | Pre-p. | Monetary | Real-time |
| 2014 | [49] | Database | Struct. data | Post-p. | Monetary | Real-time |
| 2014 | [51] | Cult. Heritage | Image | Pre-p. | Monetary | Batch |

# Overview of hybrid systems

- Balance between systems that use the human component as **pre-processing or post-processing** of data (11 vs 13)
- Mostly **monetary reward**
- Majority of systems perform **batch** data processing rather than real-time jobs
- In 2014 we can observe a decreased number of hybrid human-machine systems being propose : focus on solving **core problems** rather than building new systems

# Want to know more?

- SIGMOD 2017 Tutorial (3 hours): http://www.cs.sfu.ca/~jnwang/ppt/sigmod17-tutorial-crowd.pdf

- Adam Marcus and Aditya Parameswaran. **Crowdsourced data management industry and academic perspectives**. Foundations and Trends in Databases, 2015.
- Gianluca Demartini. **Hybrid Human-Machine Information Systems: Challenges and Opportunities**. In: Computer Networks, Volume 90, page 5-13 (2015), Elsevier.
- Edith Law and Luis von Ahn. **Human Computation**. Synthesis Lectures on Artificial Intelligence and Machine Learning. June 2011
- "An introduction to Hybrid Human-Machine Information Systems" in Foundations and Trends® in Web Science (coming soon)

- **Open Research topics**: http://www.gianlucademartini.net/research/openquestions.html

# Summary

- **Hybrid human-machine systems** can
  - Scale over large amounts of data
  - Reach high accuracy by keeping humans in the loop
- Entities are the new entry point to Web content
  - “Things not string”
  - Google Knowledge Vault (but also Bing, Yahoo!, Yandex)
- Users can benefit from **entity-centric search**, browsing, and exploration of the Web

Gianluca Demartini
http://gianlucademartini.net
@eglu81

Gianluca Demartini. **Hybrid Human-Machine Information Systems: Challenges and Opportunities**. In: Computer Networks, vol 90, 5-13, Elsevier, 2015.