

# Human Computation for Entity-Centric Information Access

Gianluca Demartini  
University of Sheffield  
[gianlucademartini.net](http://gianlucademartini.net)

# Gianluca Demartini



- 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**
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016, Facebook




[g.demartini@sheffield.ac.uk](mailto:g.demartini@sheffield.ac.uk)



[www.gianlucademartini.net](http://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)

# Entity-Centric Information Access

tom cruise   Gianluca 

[All](#) [News](#) [Images](#) [Videos](#) [Shopping](#) [More ▾](#) [Search tools](#)  


About 78,300,000 results (0.47 seconds)

**Official Tom Cruise: Edge Of Tomorrow, Movies, Bio, News ...**  
[www.tomcruise.com/ ▾](http://www.tomcruise.com/)  
OFFICIAL TOM CRUISE SITE: View the latest EDGE OF TOMORROW trailer! Watch career movie trailers, videos, and retrospective. Read the **Tom Cruise ...**

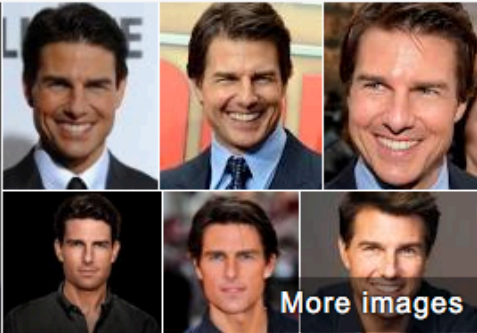

**Tom Cruise - IMDb**  
[www.imdb.com/name/nm0000129/ ▾](http://www.imdb.com/name/nm0000129/)  
**Tom Cruise**, Actor: Top Gun. If you had told fourteen-year-old Franciscan seminary student Thomas Cruise Mapother IV that one day in the not-too-distant future ...


**Tom Cruise - Wikipedia, the free encyclopedia**  
[https://en.wikipedia.org/wiki/Tom\\_Cruise ▾](https://en.wikipedia.org/wiki/Tom_Cruise)  
**Tom Cruise** is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his ...  
[Tom Cruise filmography](#) - [Mimi Rogers](#) - [Katie Holmes](#) - [Nicole Kidman](#)

**In the news**


 **Scientologist who worked with Tom Cruise condemned to horrific work camp over lesbian kiss**  
[PinkNews](#) - 2 days ago  
A former Scientologist, who worked with celebrities like **Tom Cruise** and John Travolta, has ...

[Jerry Bruckheimer confirms Tom Cruise is signed up for Top Gun 2](#)

 **More images**

**Tom Cruise** 

Actor

 [tomcruise.com](http://tomcruise.com)

Tom Cruise is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Endless Love.  
[Wikipedia](#)

**Born:** July 3, 1962 (age 53), [Syracuse, New York, United States](#)

**Height:** 1.7 m

**Spouse:** [Katie Holmes](#) (m. 2006–2012), [Nicole Kidman](#) (m. 1990–2001), [Mimi Rogers](#) (m. 1987–1990)

- Entity-seeking queries make up 40-50% of the query volume
  - Jeffrey Pound, Peter Mika, Hugo Zaragoza: Ad-hoc object retrieval in the web of data. WWW 2010: 771-780
  - Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, Ariel Fuxman: Active objects: actions for entity-centric search. WWW 2012: 589-598
- Show a summary of the most likely information-needs
  - Including related entities for navigation
  - *Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, Nicolas Torzec: Entity Recommendations in Web Search. ISWC 2013*



Matthew Paige "Matt" Damon is an American actor, voice actor, screenwriter, producer, and philanthropist whose career was launched following the success of the drama film *Good Will Hunting* (1997) from a screenplay... wikipedia.org

**Born:** October 8, 1970 (age 43), [Cambridge, Massachusetts, USA](#)

**Height:** 5' 10" (1.78m)

**Spouse:** [Luciana Barroso \(m. 2005-present\)](#)

**Partner:** [Winona Ryder \(1998-2000\)](#)

**Parents:** [Kent Damon](#), [Nancy Carlsson-Paige](#)

**Children:** [Isabella Damon](#), [Alexia Barroso](#), [Gia Zavala Damon](#), [Stella Damon](#)

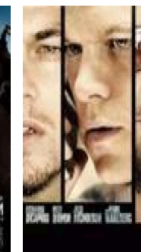
#### Movies & TV Shows



[The Zero Theorem](#)



[Elysium](#)



[The Departed](#)



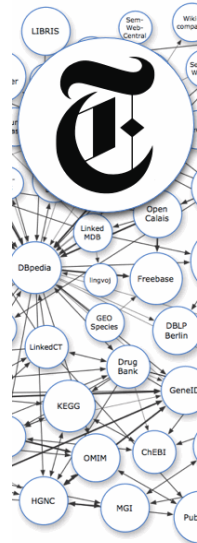
[We Bought a Zoo](#)



[Good Will Hunting](#)

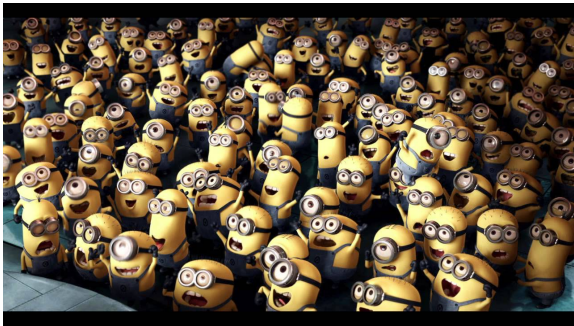
# Web of Data

- Freebase
  - Acquired by Google in July 2010.
  - Knowledge Graph launched in May 2012.
  - Read-only in December 2014 -> WikiData
- Schema.org
  - Driven by major search engine companies
  - Machine-readable annotations of Web pages
- Linked Open Data
  - 31 billion triples, Sept 2011
  - 90 billion triples, Aug 2015 (stats.lod2.eu)



# Today I will talk about

- Human Computation
  - Amazon MTurk as a crowdsourcing platform
  - Hybrid Human-Machine Information Systems
- Entity Linking on the Web
  - With the crowd
- Efficiency/effectiveness of Human Computation



# Crowdsourcing

- Leverage human intelligence at scale to solve
  - Tasks simple for humans, complex for machines
  - With a large number of humans (the Crowd)
  - Small problems: micro-tasks/ HITs (Amazon MTurk)
- Examples
  - Wikipedia, Image tagging
- Incentives
  - Financial, fun, visibility
- See my tutorials at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016 and upcoming FnTWeb paper.



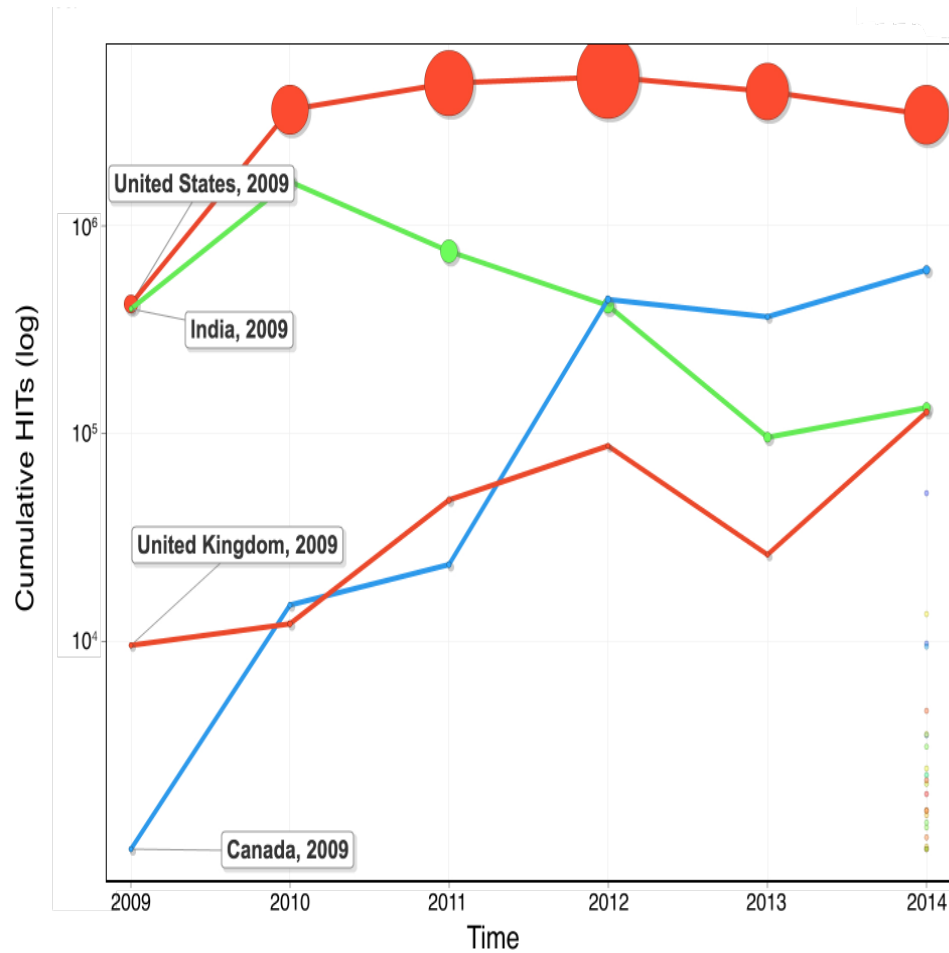
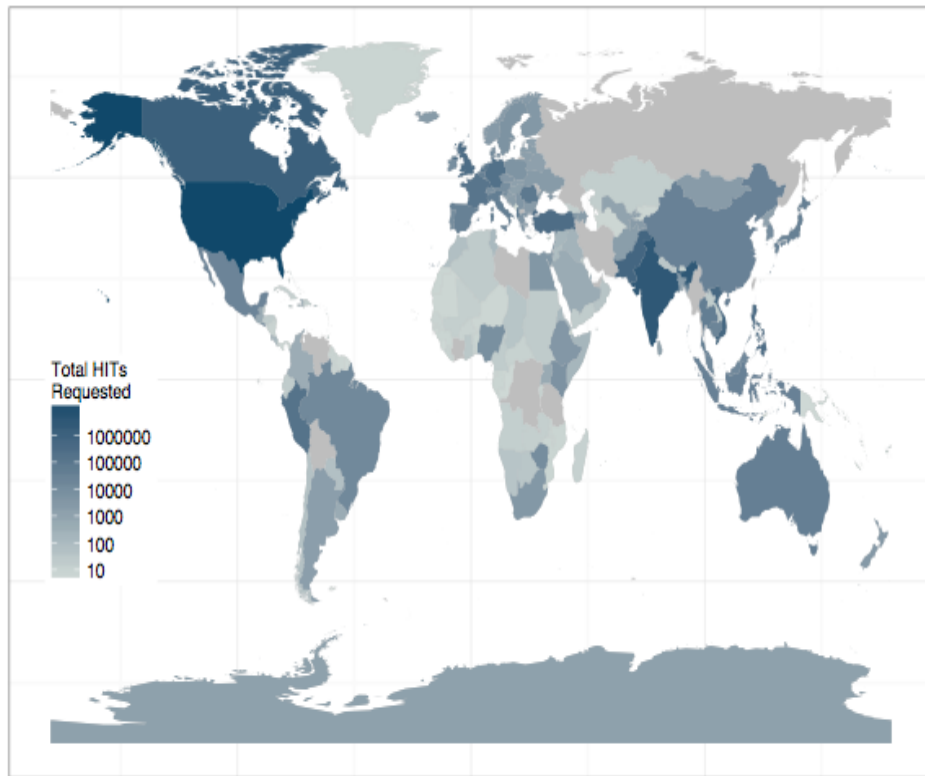


# 5-year Analysis of MTurk workload

- Mturk-tracker.com
  - Collects metadata about each visible **batch** (Title, description, rewards, required qualifications, HITs available, requester, etc), that is, set of similar tasks or **HITs**
  - Records batch progress (every ~20 minutes)
  - Covers 130M tasks
  - 2009-2014

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.

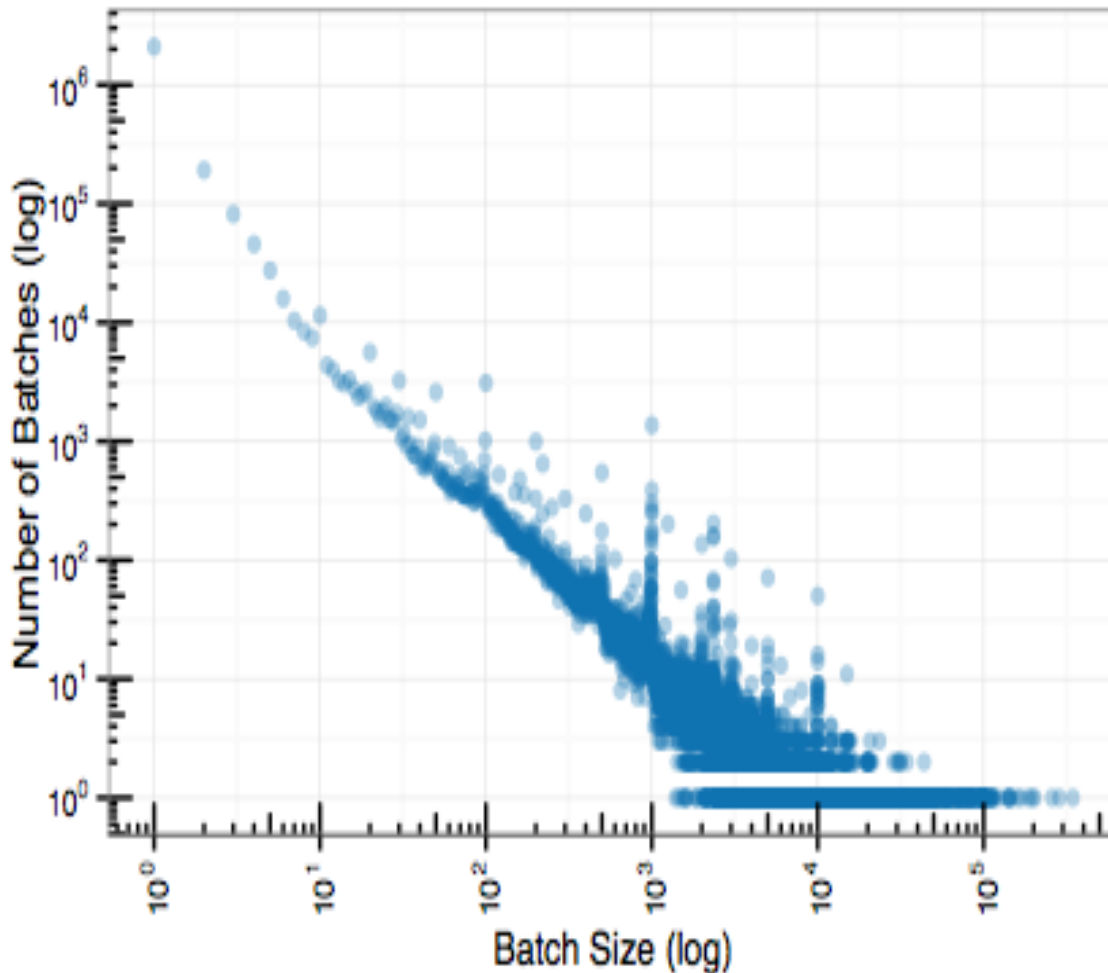
# Country-Specific HITs



Workers from US, India and Canada are the most sought after.

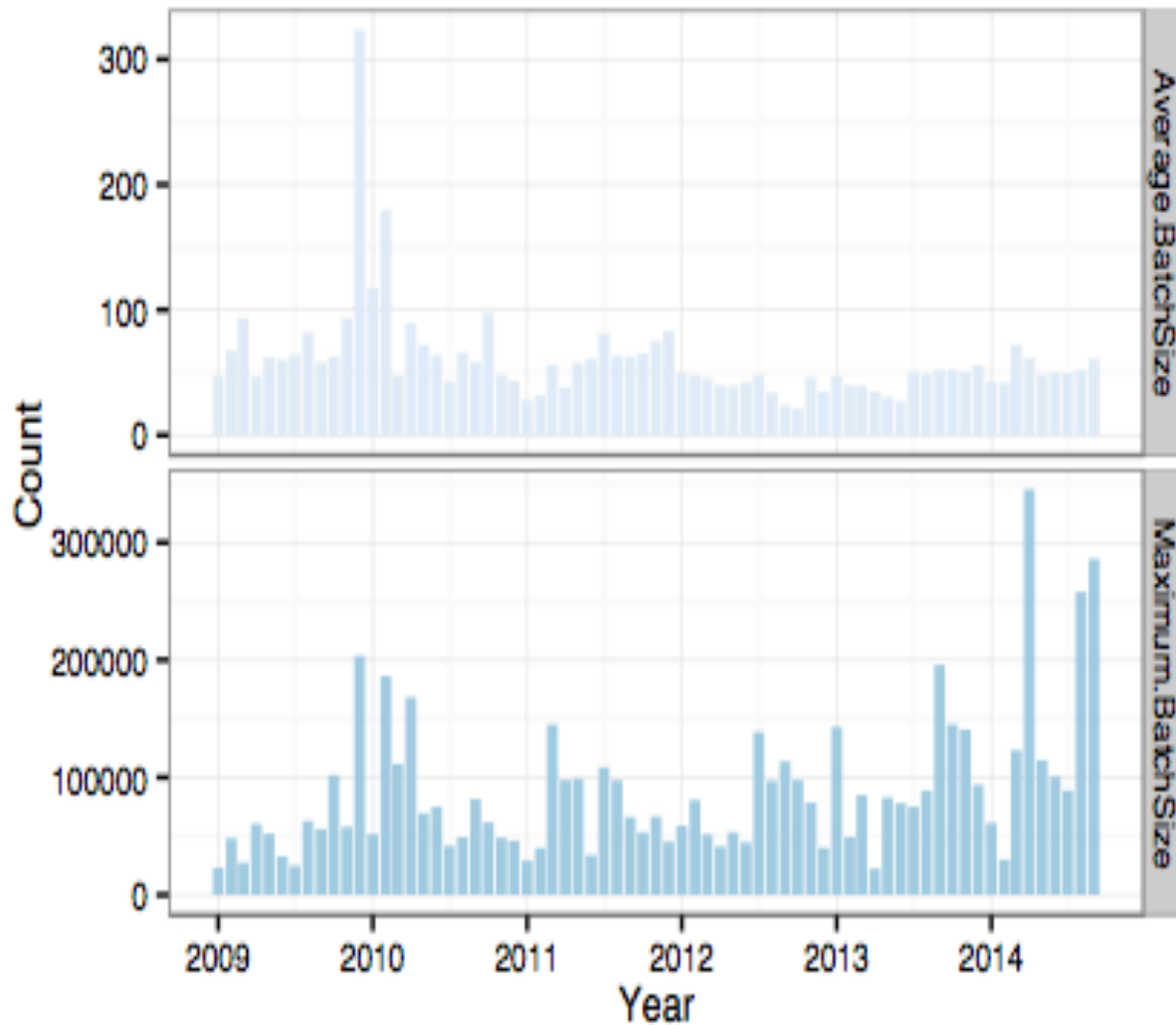
<http://exascale.info/mturk-mrkt/> for interactive visualizations

# Distribution of *Batch Size*



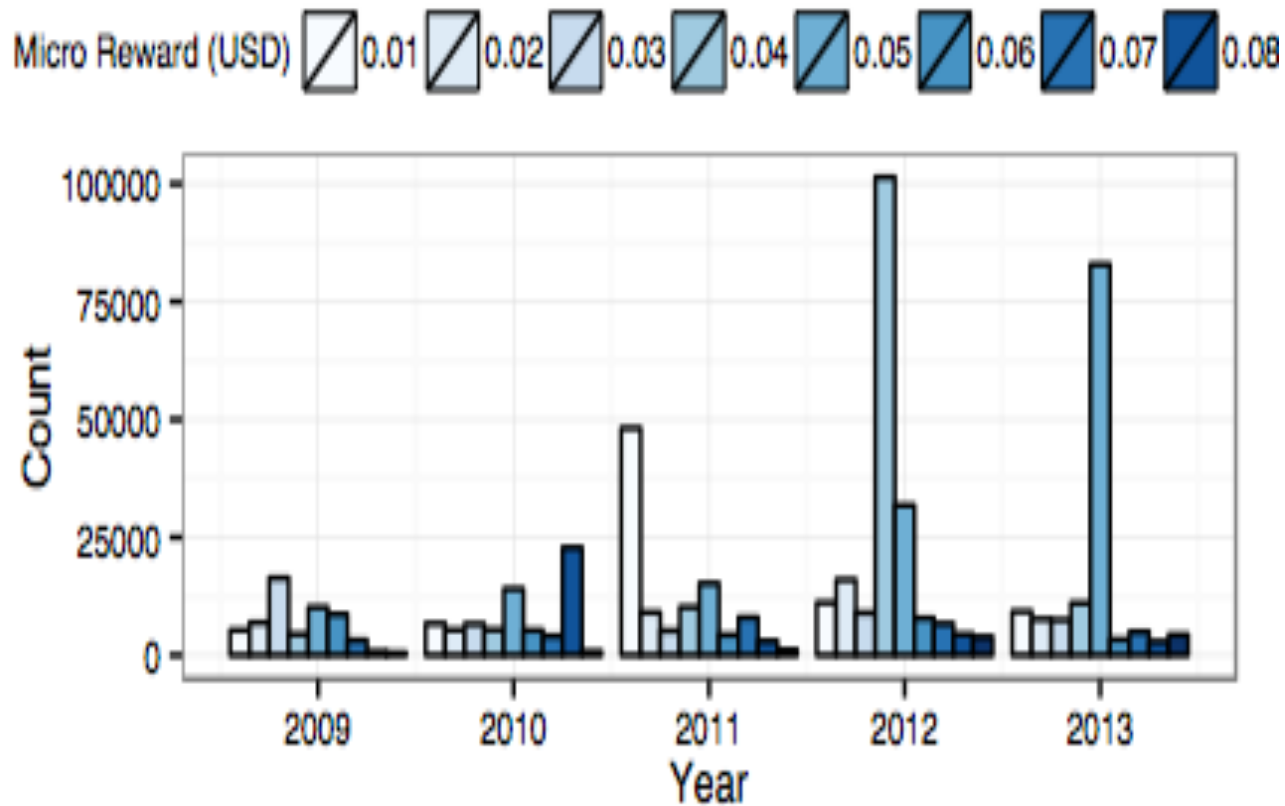
A **batch** is a collection of homogeneous HITs (e.g., Audio transcription from radio programs) by the same requester

# Evolution of Batch Sizes



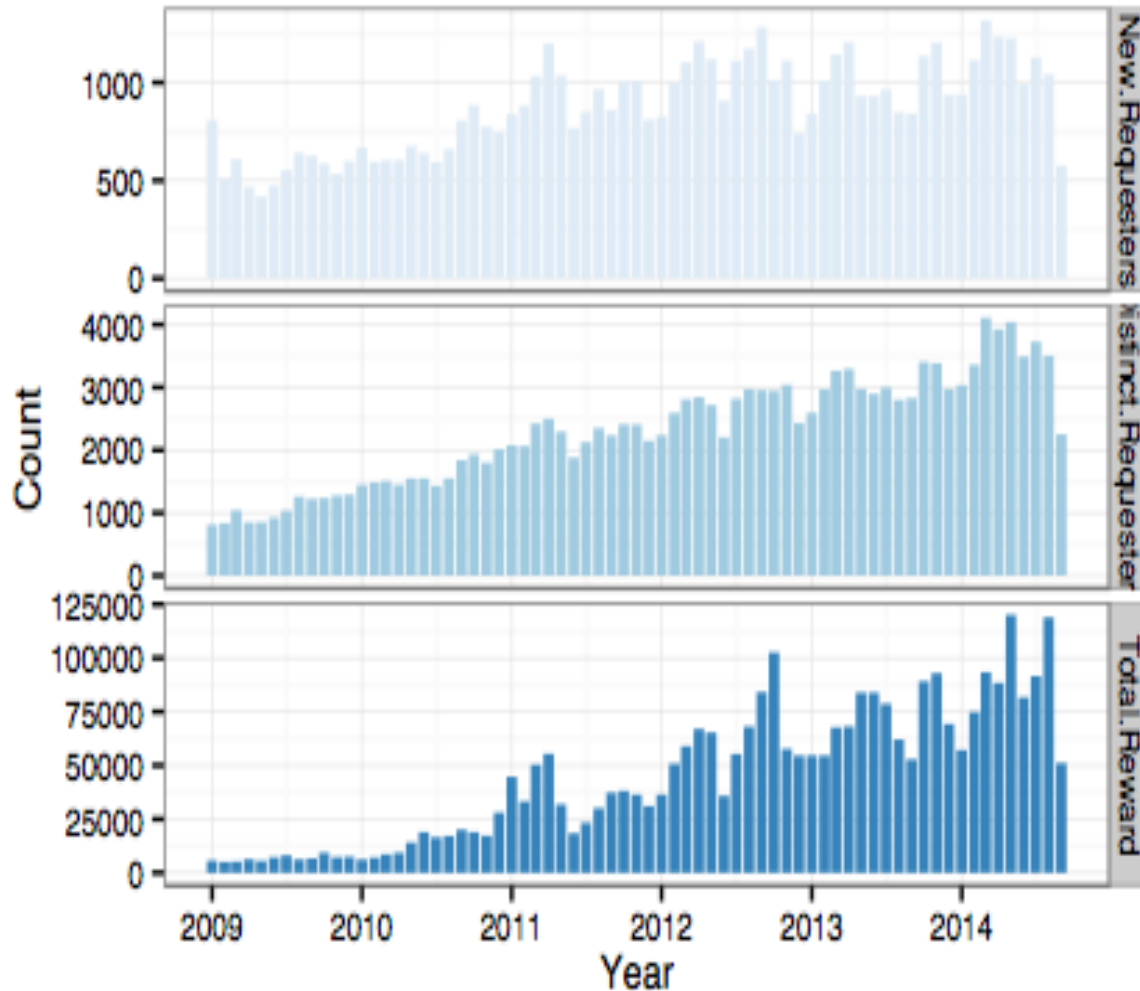
Very large batches  
start to appear

# HIT Pricing



5-cents is the new  
1-cent

# Requesters and Reward Evolution



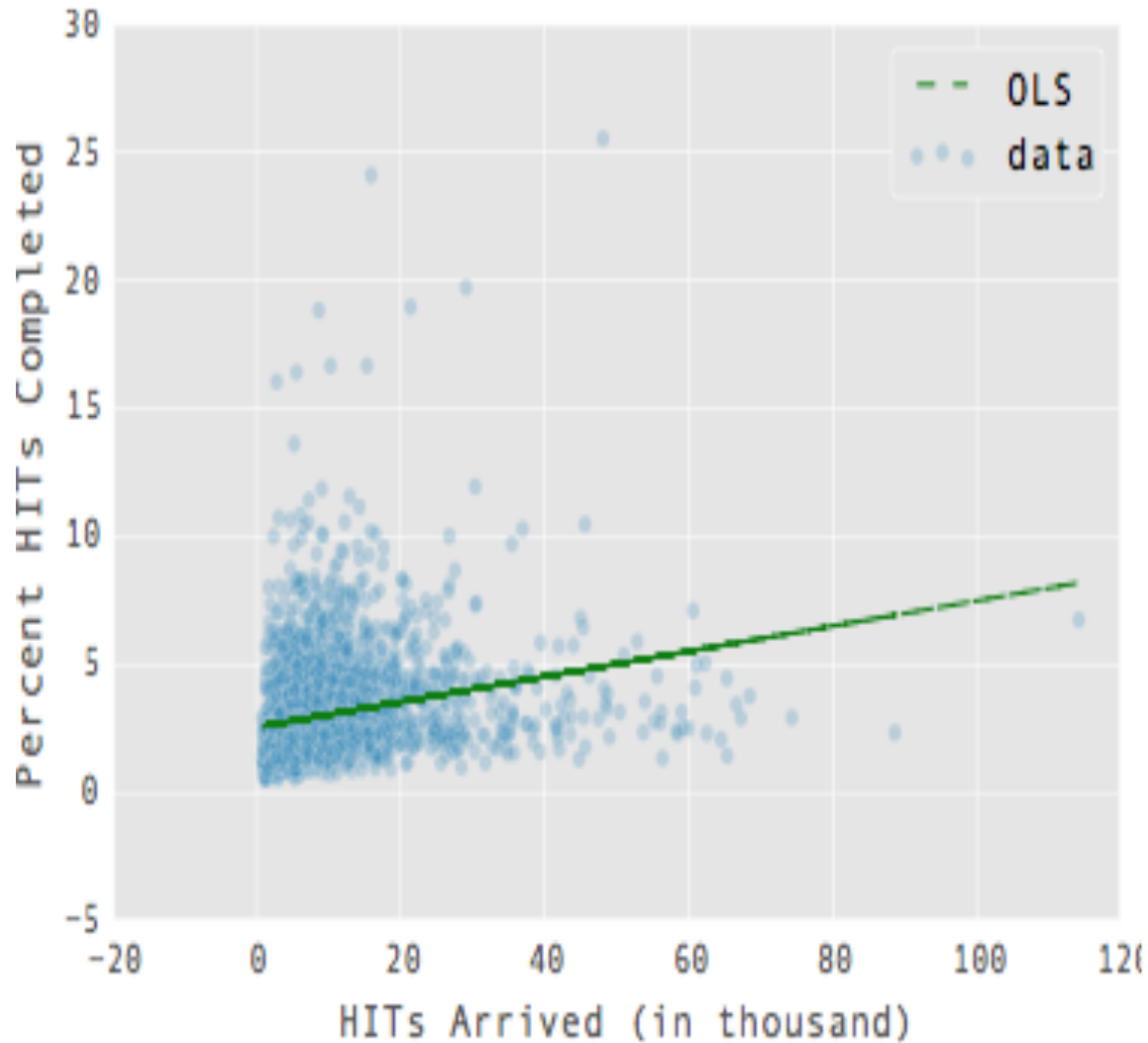
Increasing number of New and Distinct Requesters

# Top MTurk Requesters over 1 month

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

# Supply Elasticity



Intercept = 2.5  
Slope = 0.5%

20% of new work gets  
completed within an hour



# Summary

- 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

# Crowdsourcing for Entity Linking

# 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>

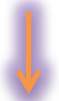
owl:sameAs

fbase: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 e 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>

CNET > News > Mobile

## Instagram for Android is now available

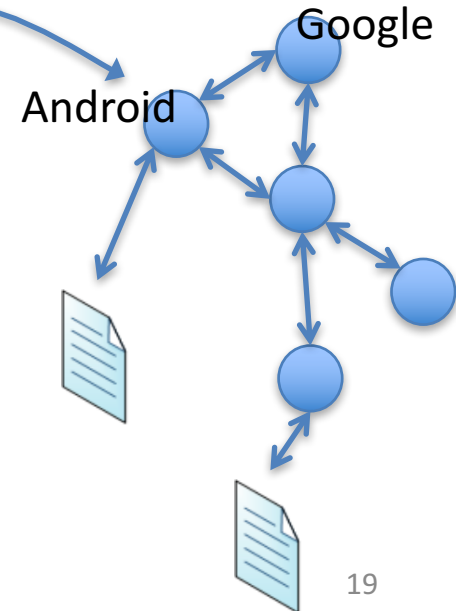
At long last, Instagram finally releases the Android version of its app.



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

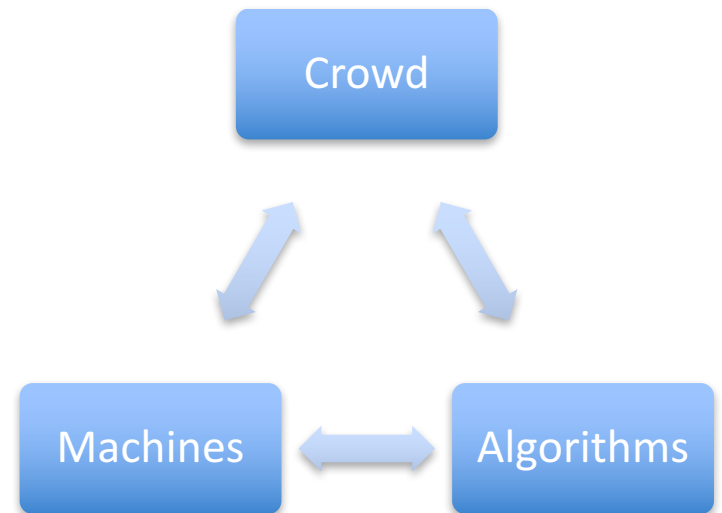
Follow

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.

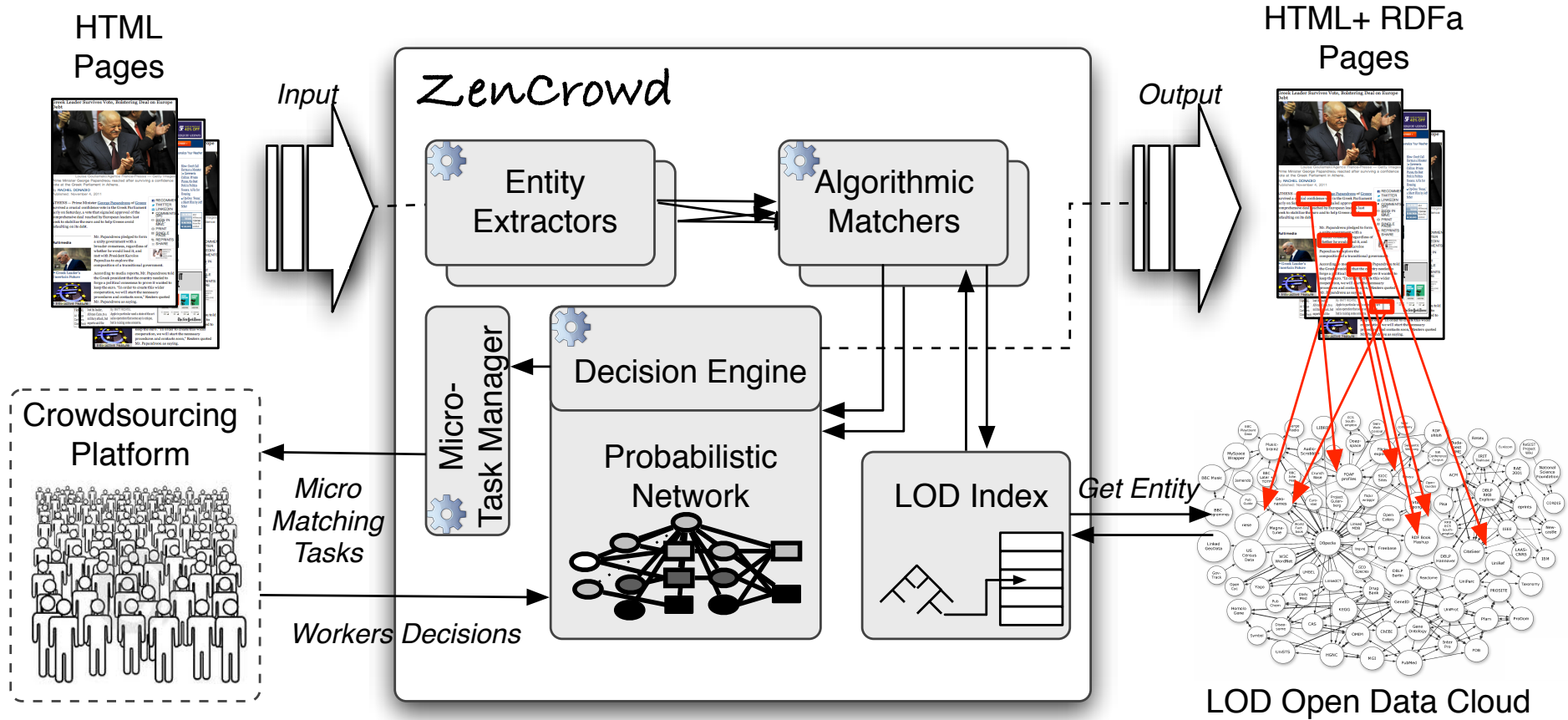


# ZenCrowd

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



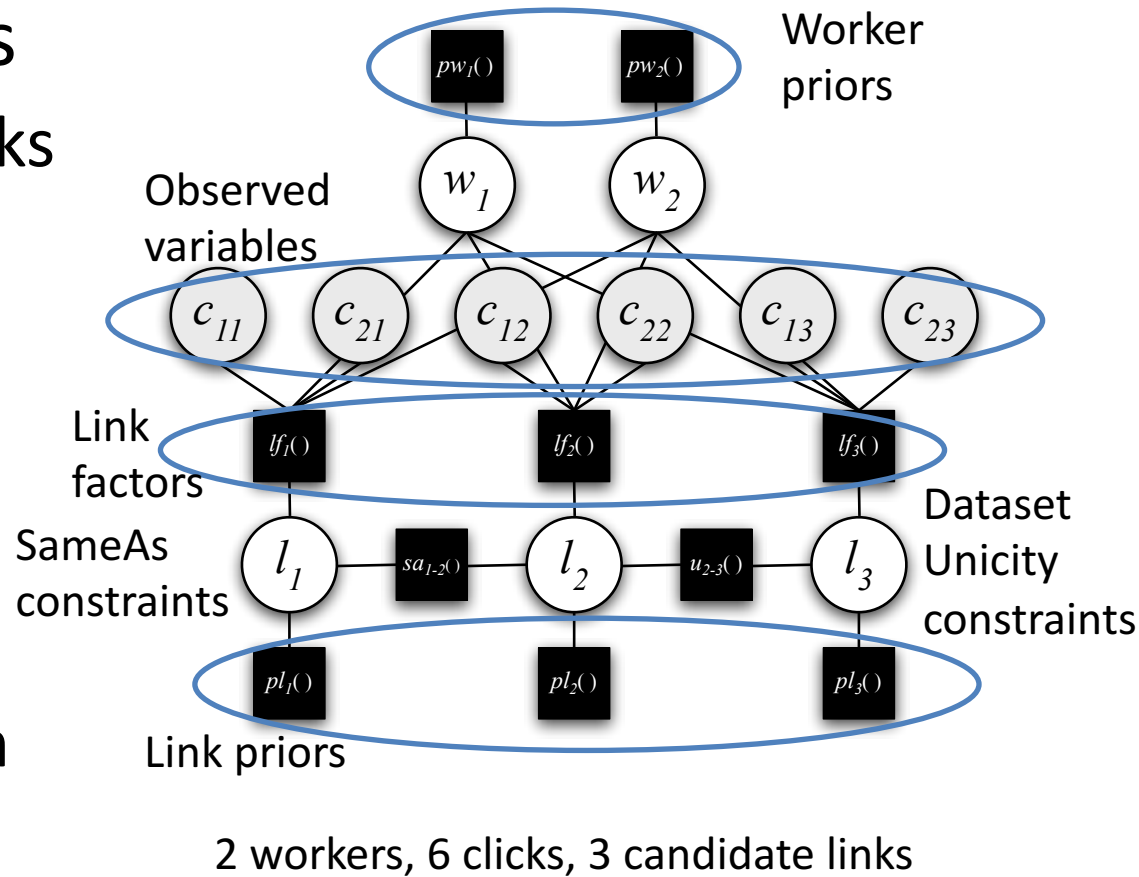
# ZenCrowd Architecture



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**).

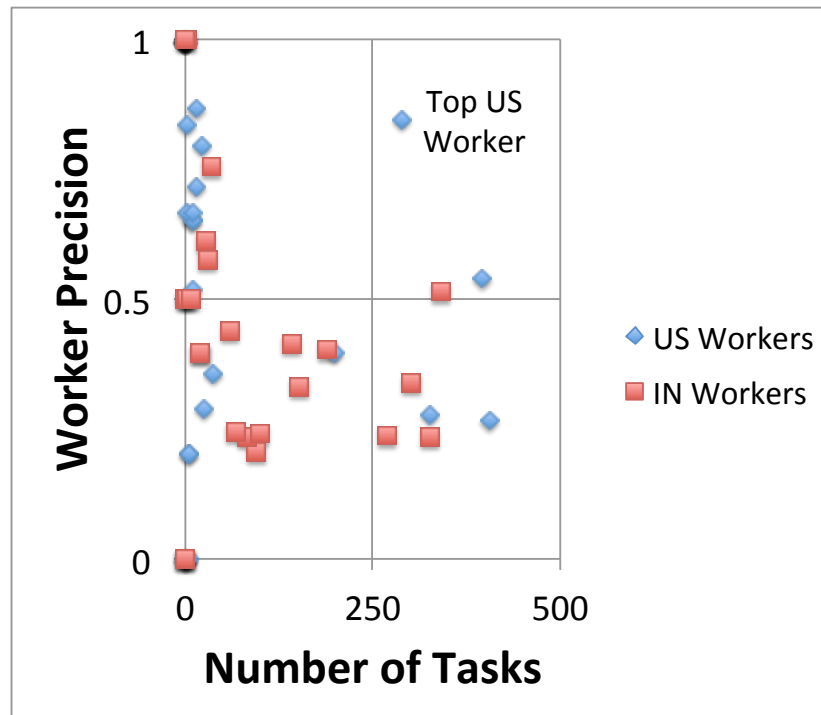
# Entity Factor Graphs

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



# 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

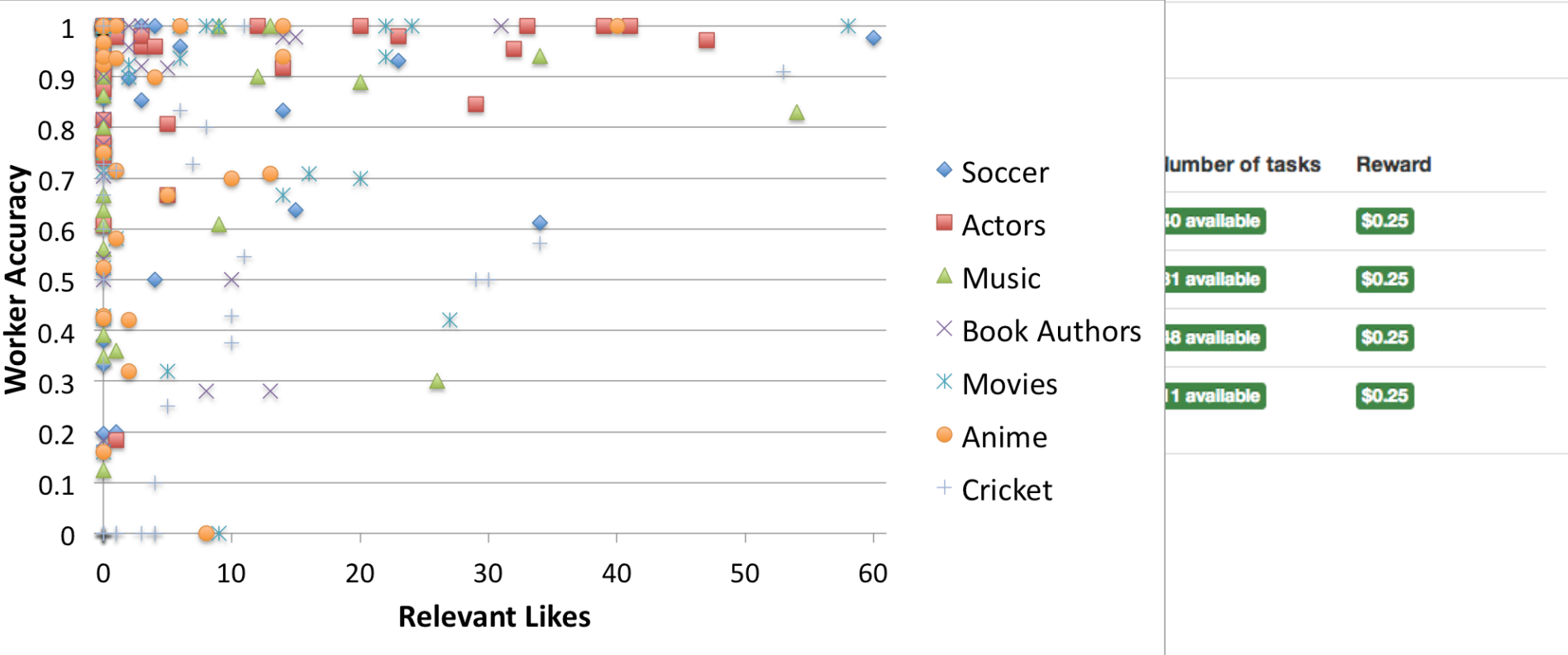


# Lessons Learnt

- 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)

My customized list of batches:

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



Number of tasks	Reward
10 available	\$0.25
31 available	\$0.25
18 available	\$0.25
11 available	\$0.25

# 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).

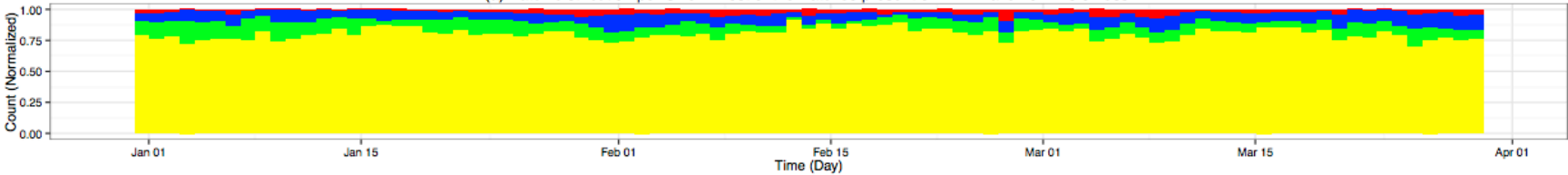
Response: *'one, two, three, four, five'*

Gold Standard  
Preys (GSP)

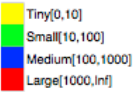
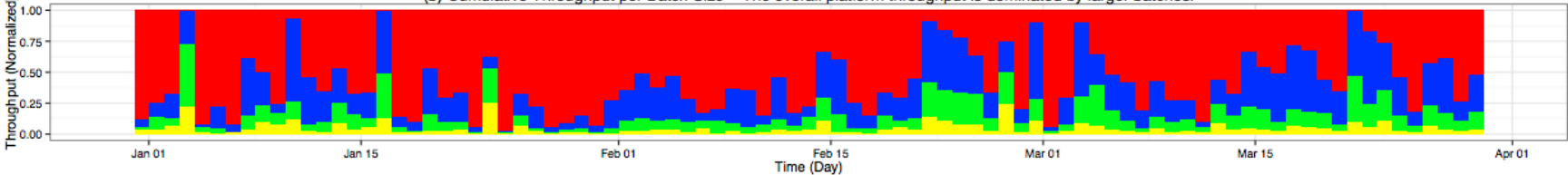
These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions!

# Scheduling HITs

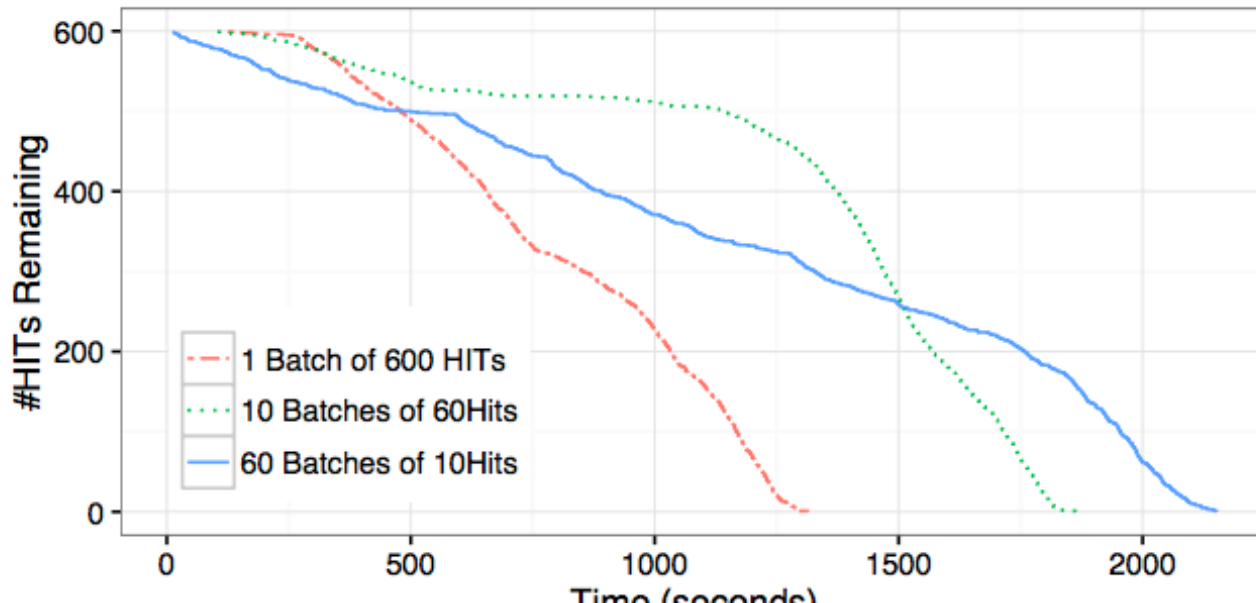
(a) Batch distribution per Size – Most of the Batches present on AMT have 10 HITs or less.



(b) Cumulative Throughput per Batch Size – The overall platform throughput is dominated by larger batches.



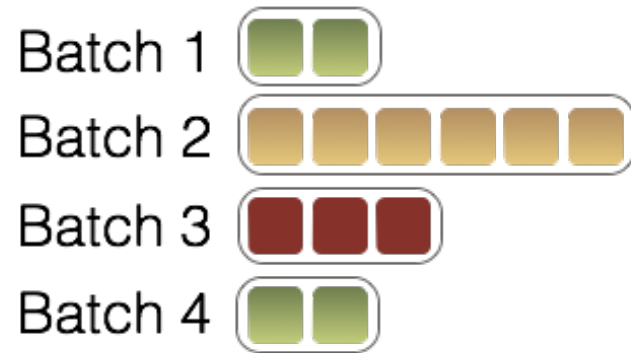
**Platform throughput (HITs/h) dominated by large HIT batches**



# HIT-Bundle

## Definition

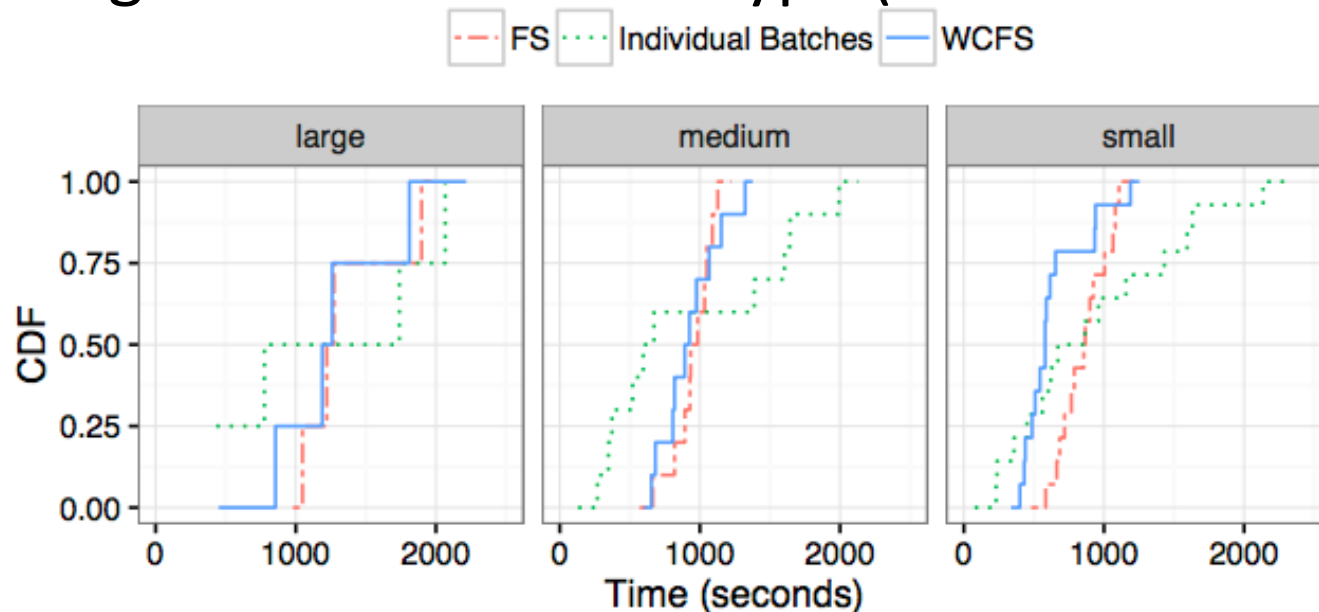
- Scheduling requires control over the serving process of tasks
- A **HIT-Bundle** is a batch that contains heterogeneous tasks
- All tasks that are generated by the system are published through the HIT-Bundle



HIT-Bundle

# Scheduling HITs

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



# 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