# Introduction to Micro-task Crowdsourcing Platforms

Lecture 2

Gianluca Demartini

University of Sheffield

# Outline

- Micro-tasks
- Micro-task Crowdsourcing platforms
- Amazon MTurk demographics and dynamics
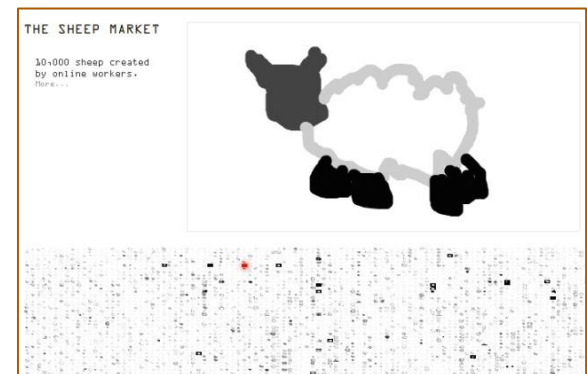- Examples of micro-task crowdsourcing

# Types of Tasks

| Task Granularity | Examples |
|---|---|
| Complex Tasks | • Build a website<br>• Develop a software system<br>• Overthrow a government? |
| Simple Projects | • Design a logo and visual identity<br>• Write a term paper |
| Macro Tasks | • Write a restaurant review<br>• Test a new website feature<br>• Identify a galaxy |
| Micro Tasks | • Label an image<br>• Verify an address<br>• Simple entity resolution |

Inspired by the report: "Paid Crowdsourcing", Smartsheet.com, 9/15/2009

# Example use of micro-task crowdsourcing

- Relevance judgments
- Ontologies
- Museums
- http://www.thesheepmarket.com/

'draw a sheep facing to the left.'

THE SHEEP MARKET

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

# Background

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

# Microtask Aggregators

# Samasource.org

# 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



**amazon mechanical turk**
*Artificial Artificial Intelligence*
beta

## 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 an interesting task | Work | Earn money |
| --- | --- | --- |

**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

| Fund your account | Load your tasks | Get results |
| --- | --- | --- |

**Get Started**

# Amazon MTurk

- Requesters create tasks (HITs)
- The platform takes a fee (30% of the reward)
- Workers preview, accept, submit HITs
- Requesters approve, download results

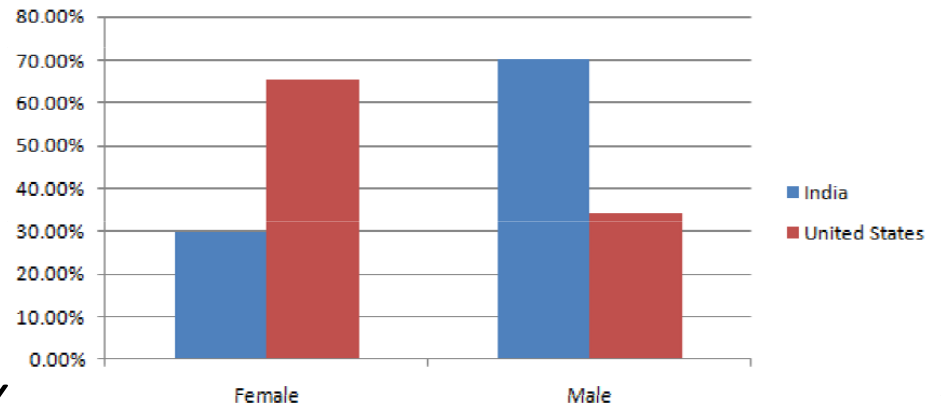- If the results are approved, workers are paid
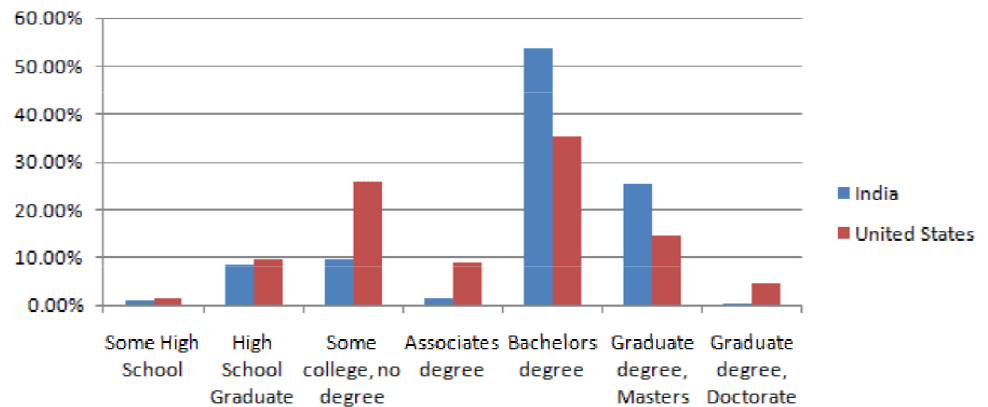
# Demographics of MTurk workers in 2009

y of residence

Country of residence
- United States: 46.80%
- India: 34.00%
- Miscellaneous: 19.20%
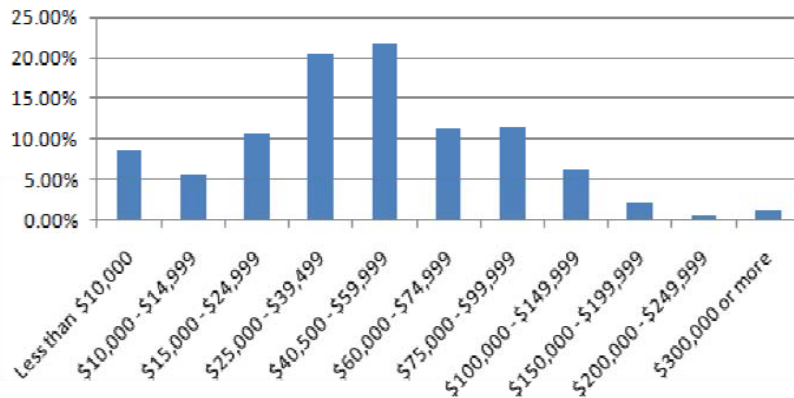
2013 Statistics:
1M workers
10% active

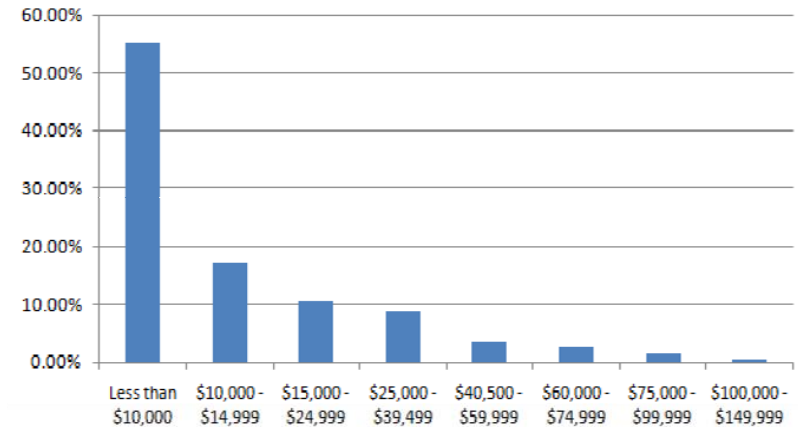**Gender Breakdown**



**Education Level**
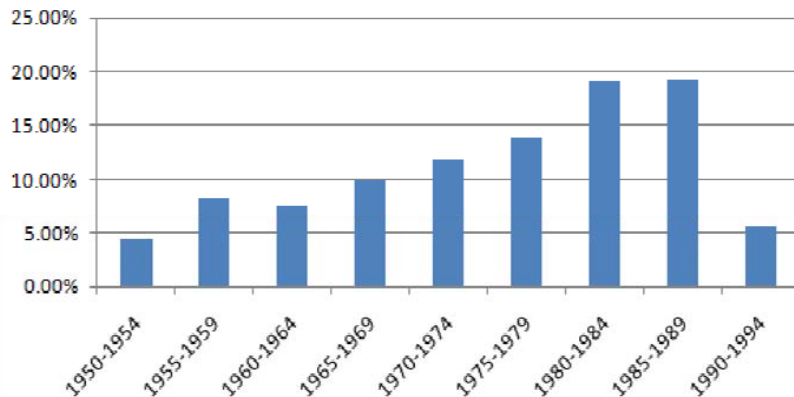
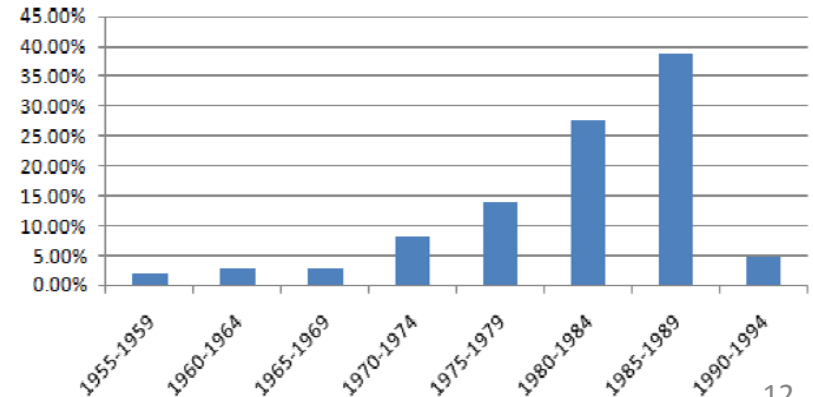# Demographics of MTurk workers in 2009



**Household Income for US workers**


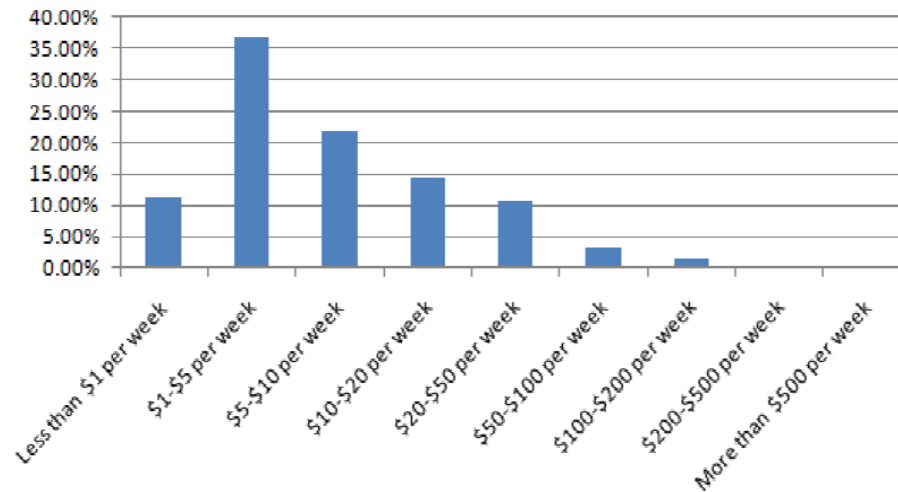
**Household Income for Indian workers**
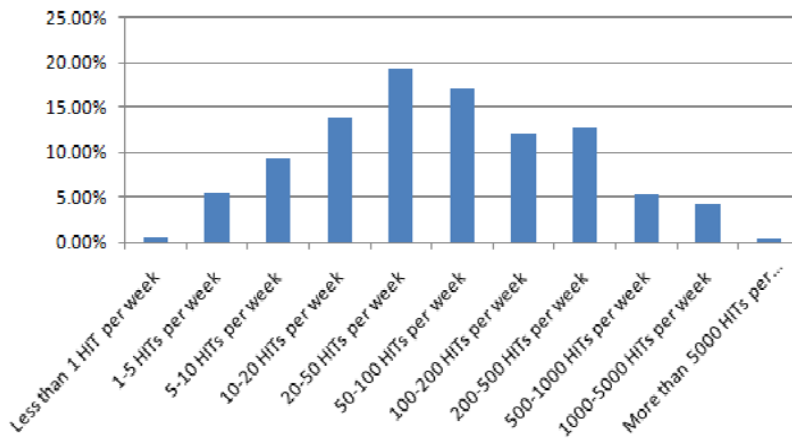


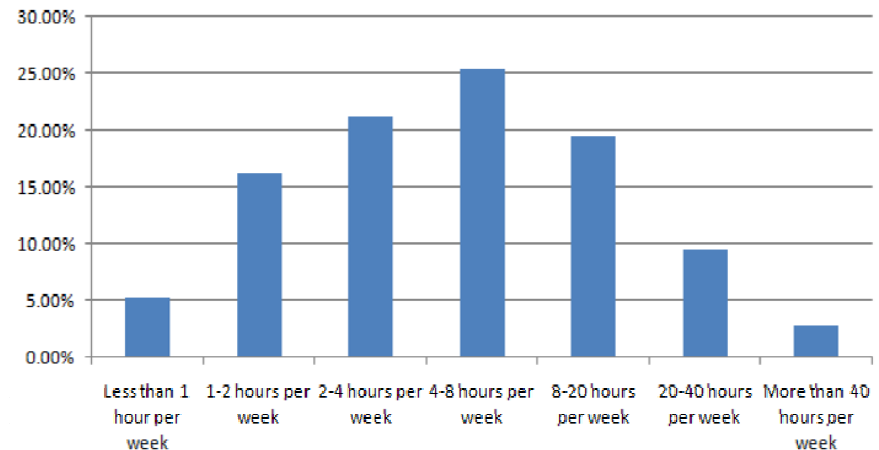**Year of Birth for US workers**



**Year of Birth for Indian workers**

# Demographics of MTurk workers in 2009





Number of HITs completed per week



Time spent on Mechanical Turk per week

http://www.mturk-tracker.com/

mturk tracker

General

Arrivals

Completions

Top requesters

Demographics

Search

API

Old version

03/23/2015     04/23/2015

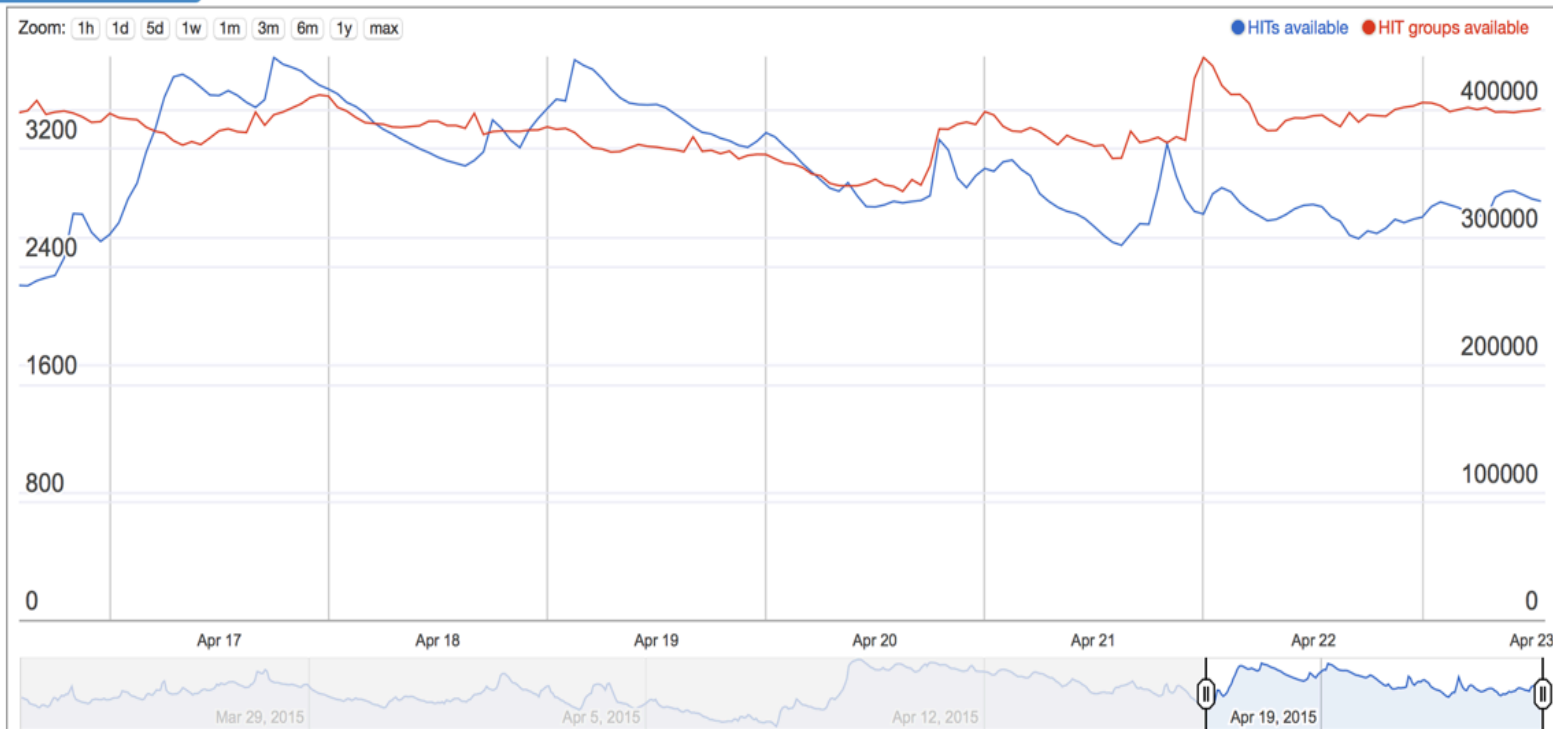Market statistics     HIT groups posted     HITs posted     Rewards posted

Zoom: 1h  1d  5d  1w  1m  3m  6m  1y  max

● HITs available  ● HIT groups available

For bugs reports or feature requests, please contact Panos Ipeirotis
If you want to cite this website, please cite the paper Analyzing the Amazon Mechanical Turk Marketplace, P. Ipeirotis, ACM XRDS, Vol 17, Issue 2, Winter 2010, pp 16-21.

14

# MTurk is a Marketplace for HITs

# Top requesters

## Top-1000 Requesters, report for May 14, 2015 to June 13, 2015

| Requester name | hits | reward |
|---|---|---|
| Turk Experiment | 9794 | $25,969.90 |
| video7789 | 87826 | $18,963.40 |
| Visual Genome | 48935 | $16,822.55 |
| VisionTurk | 69990 | $16,244.14 |
| Kevin Dodds | 139744 | $15,928.04 |
| Jon Brelig | 290131 | $13,001.36 |
| Tagasauris | 274148 | $7,872.97 |
| Amazon Requester Inc - browse classification | 127864 | $7,352.67 |

# Paid Microtask Crowdsourcing
# scales-out but remains highly unpredictable
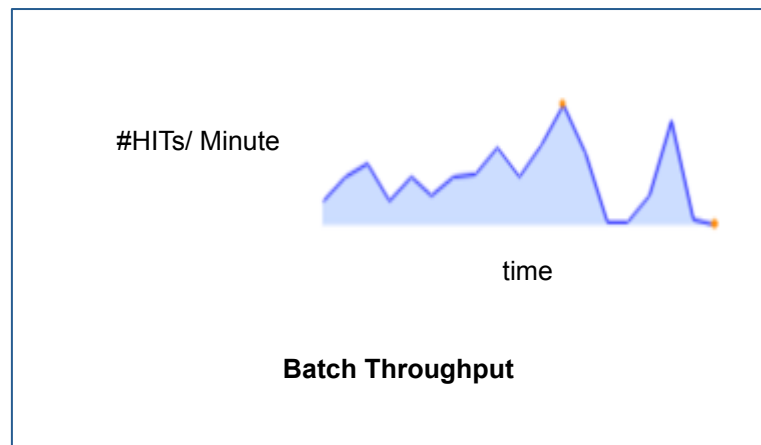


#HITs/ Minute

time

**Batch Throughput**

# SLAs are expensive

# mturk-tracker.com

- Collects metadata about each visible batch (Title, description, rewards, required qualifications, HITs available etc)

- Records batch progress (every ~20 minutes)

We note that the tracker reports data periodically only and does not reflect fine-grained information (e.g., real-time variations)

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.

# Country-Specific HITs
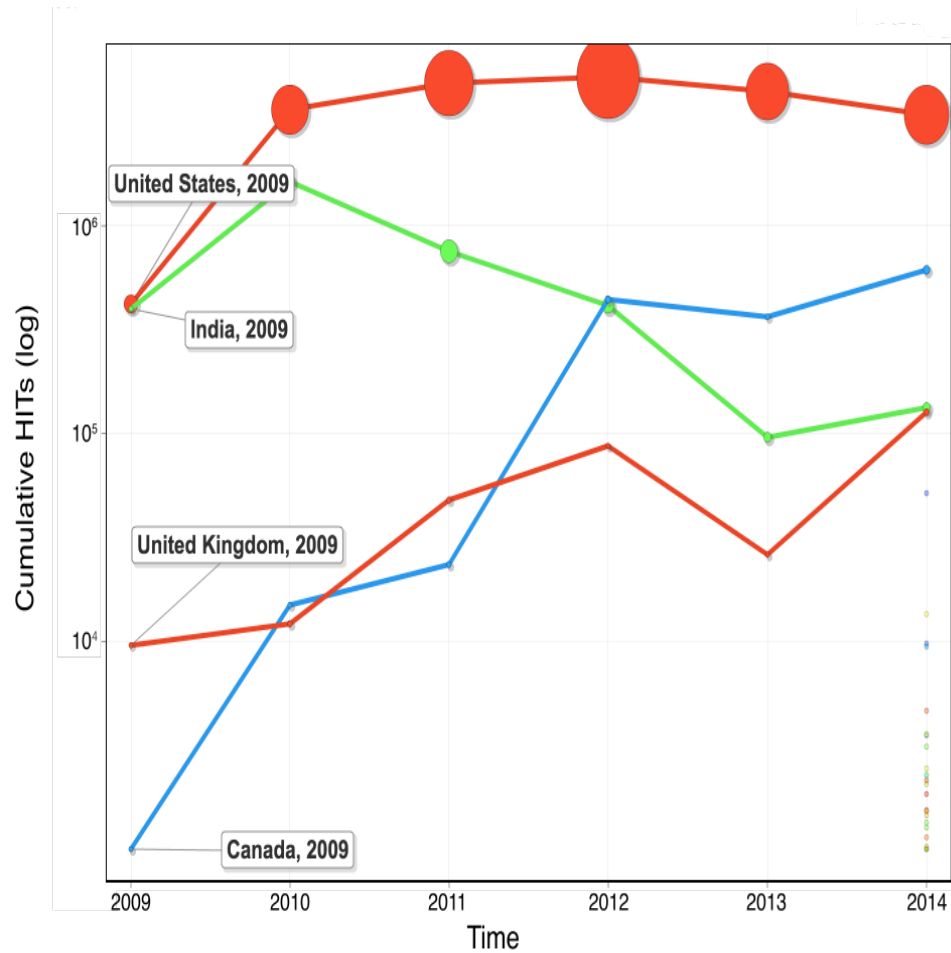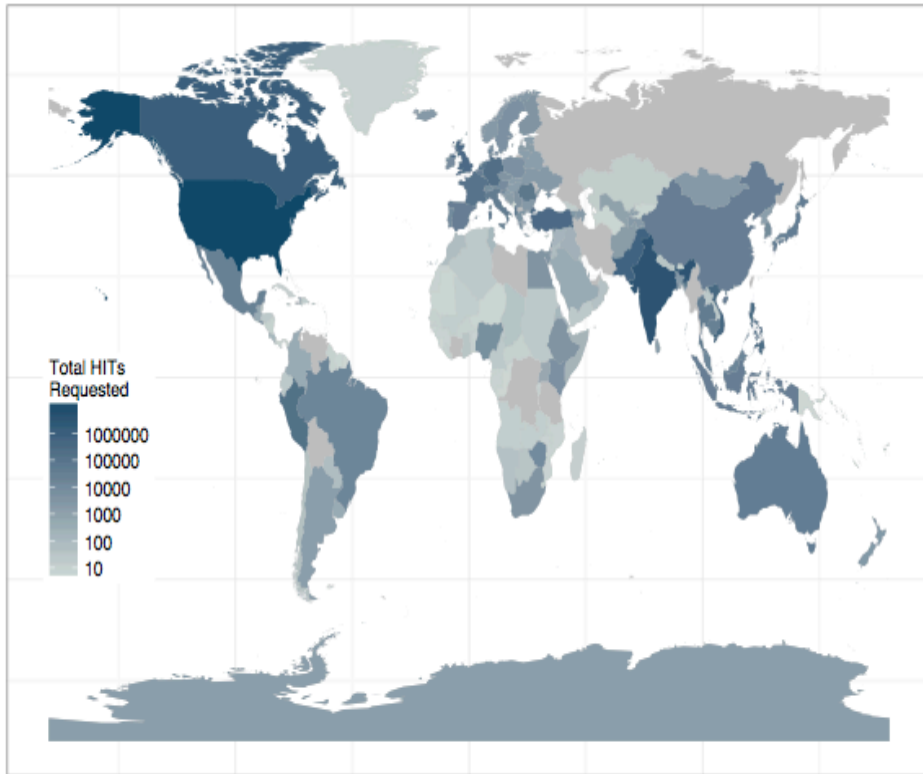


Workers from US, India and Canada are the most sought after.

# Distribution of *Batch Size*
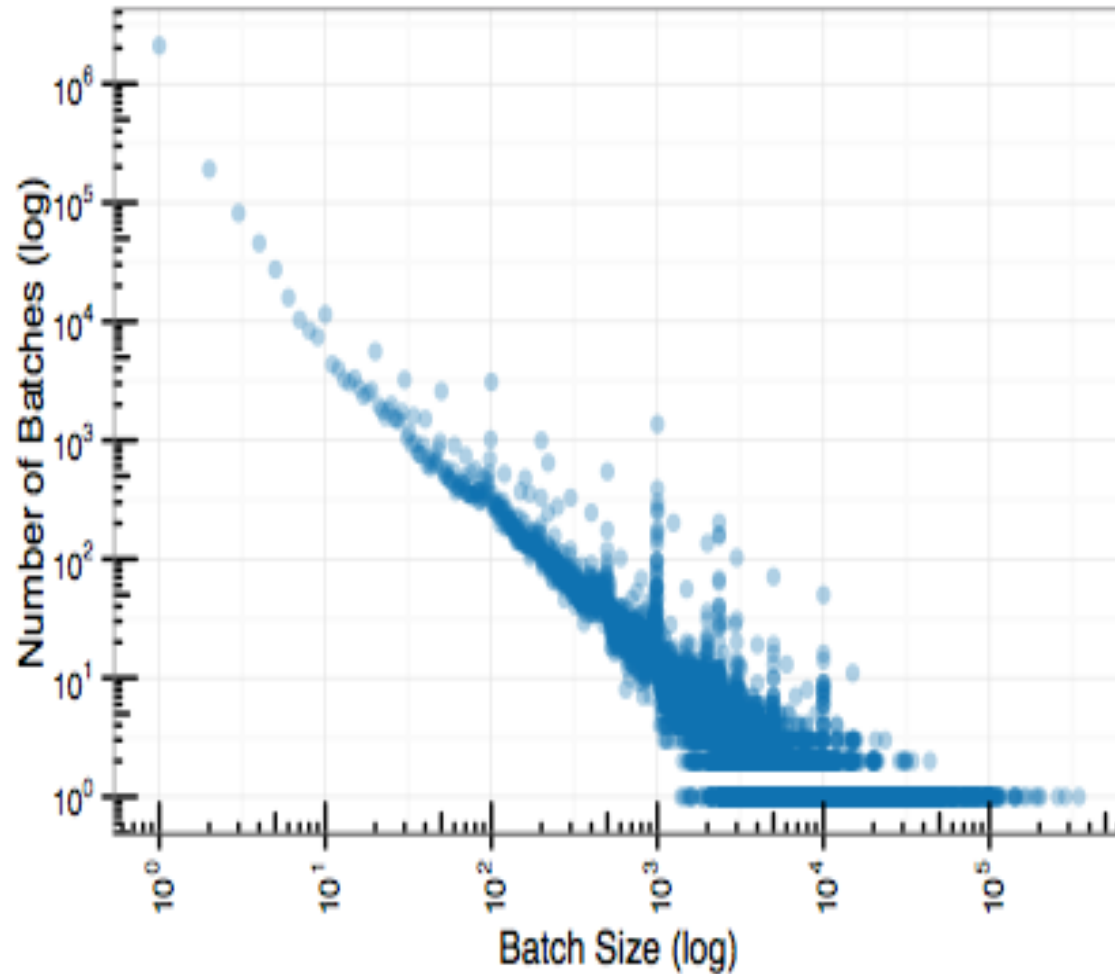


"Power-law"

# Evolution of Batch Sizes
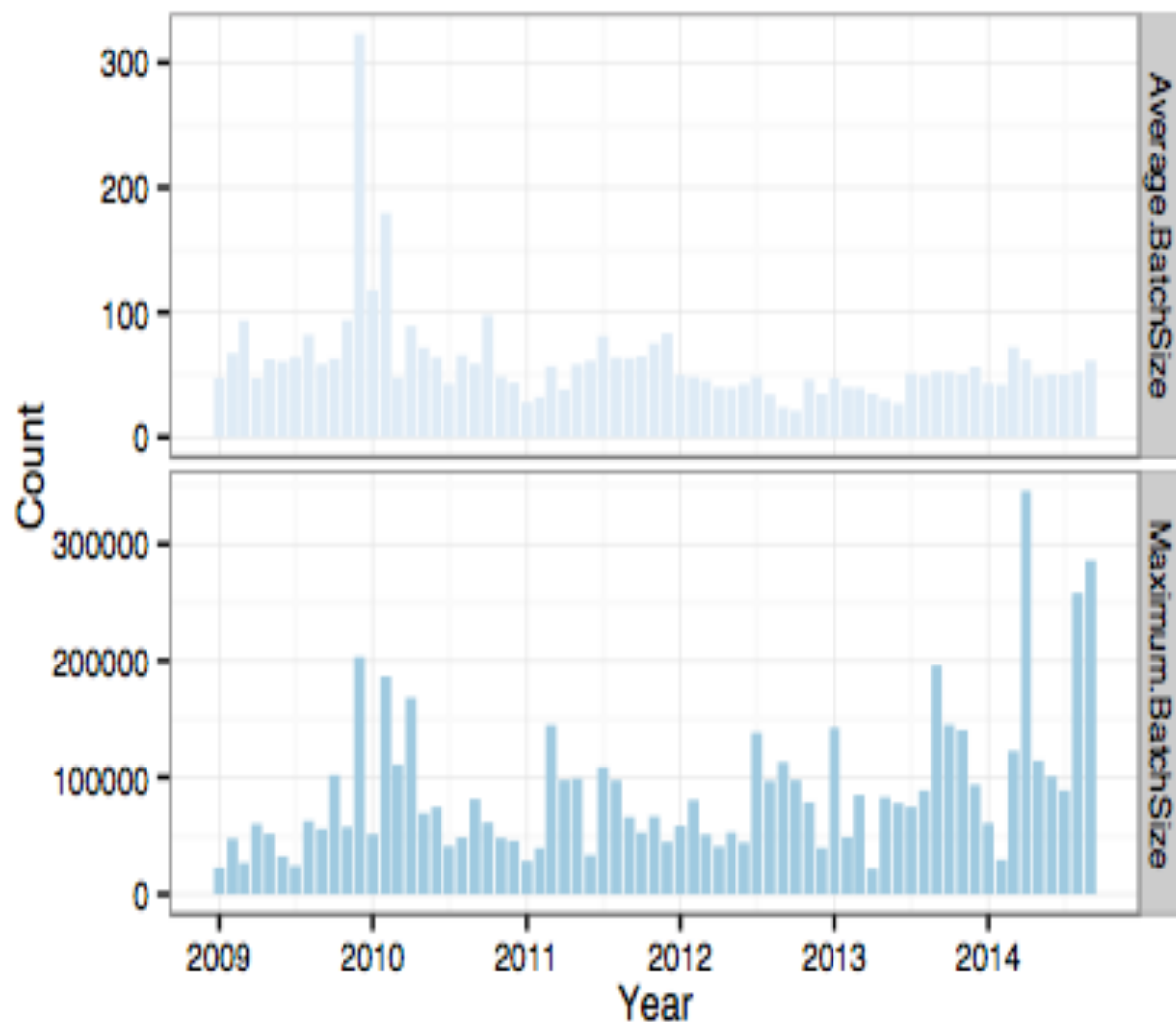


Very large batches start to appear

# HIT Pricing
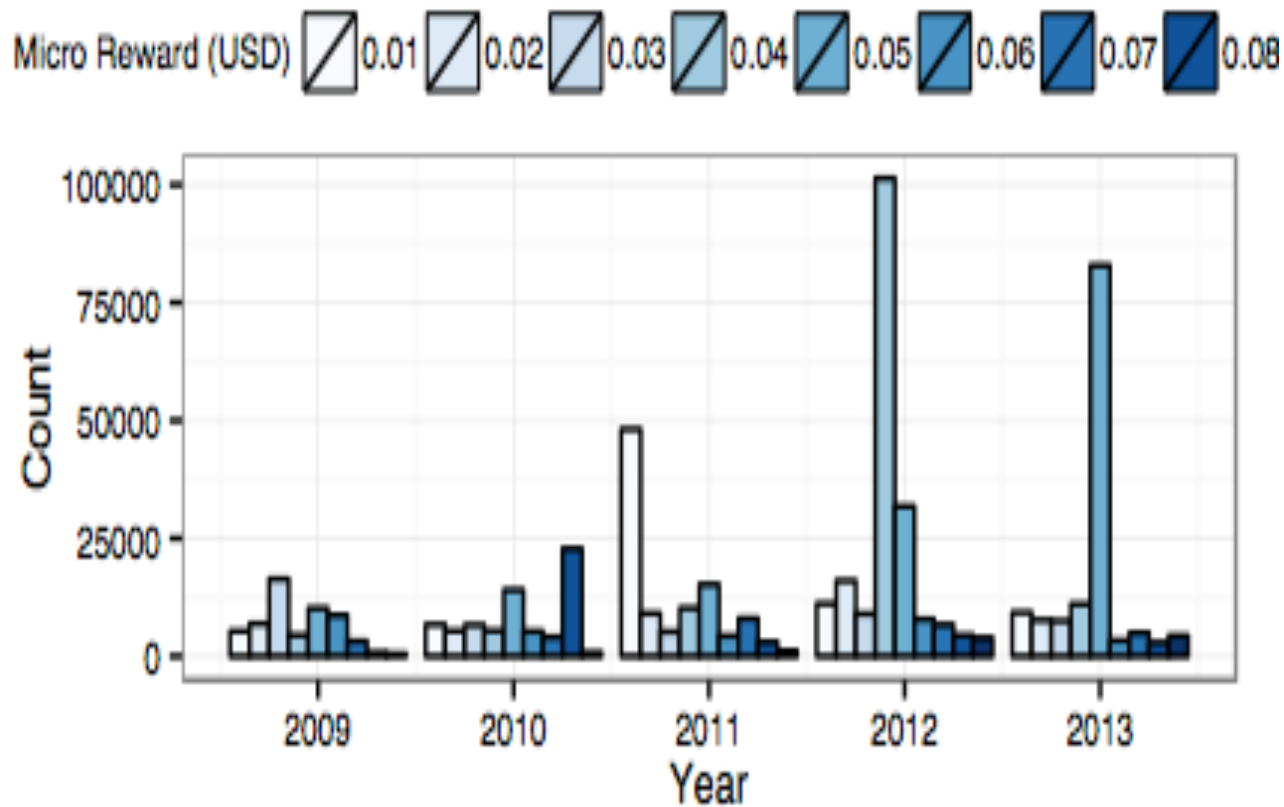


5-cents is the new 1-cent

# Requesters and Reward Evolution
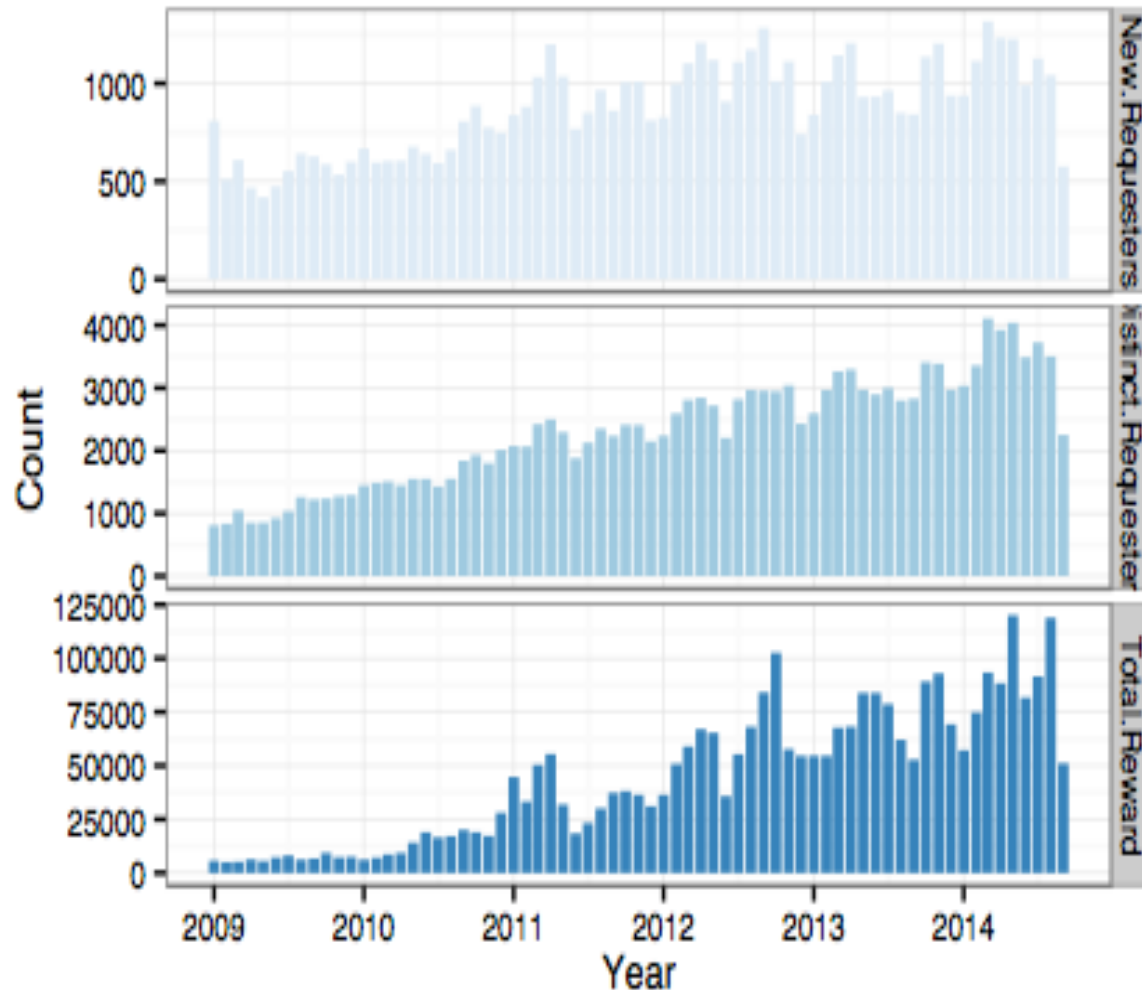


Increasing number of New and Distinct Requesters

# HIT Classes

Classify HITs into types (Gadiraju et. al 2014)
- Information Finding (IF)
- Verification and Validation (VV )
- Interpretation and Analysis (IA)
- Content Creation (CC)
- Surveys (SU)
- Content Access (CA)

# Supervised Classification With the Crowd

We trained a Support Vector Machine (SVM) model
- HIT title, description, keywords, reward, date, allocated time, and batch size
- Created labeled data on Mturk for 5,000 HITs uniformly sampled HITs
- Our HIT used 3 repetitions
  - Consensus reached for 89% of the tasks
- 10-fold cross validation
  - Precision of 0.895
  - Recall of 0.899
  - F-Measure of 0.895
- Then, a large-scale classification for all 2.5M HITs

# Distribution of HIT Types
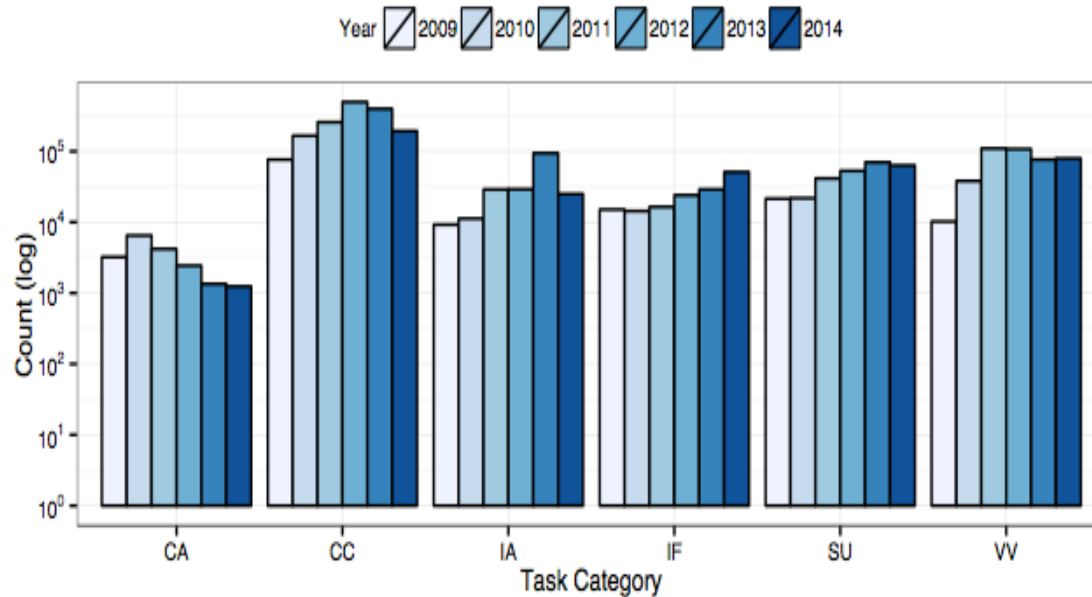


Less Content Access batches

Content Creation being the most popular

# Batch Throughput Prediction

- Predict batch throughput at time T by training a Random Forest Regression model with samples taken in [T-delta, T) time span
- 29 Features (including the Type of the Batch)
- Hourly Data in range [June-October] 2014
- We sampled 50 times points for evaluation purposes

*delta*

*time*

T
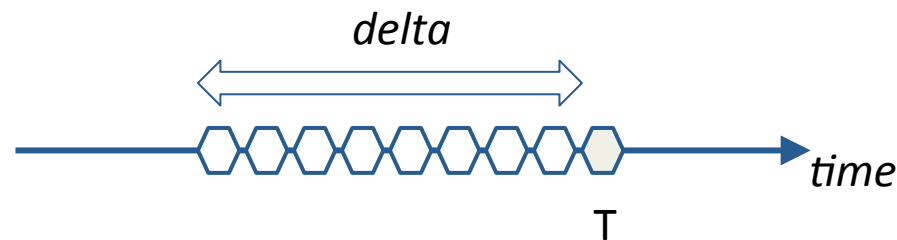
We are interested in cases where prediction works reasonably

# Predicted vs. Actual Batch Throughput (delta=4 hours)
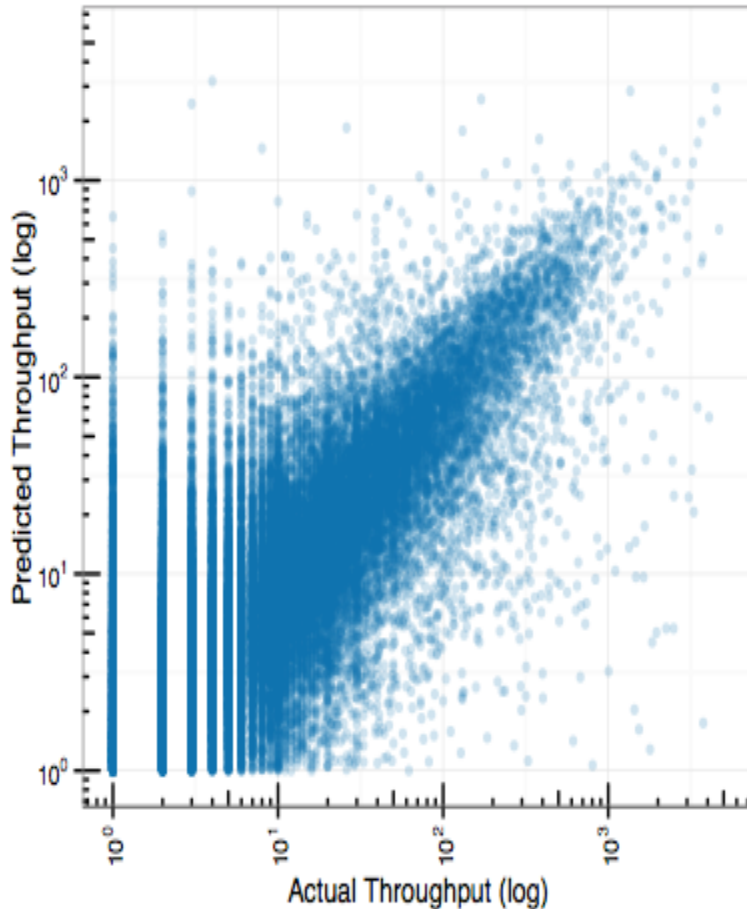


Prediction Works best for larger batches having large momentum

# Significant Features

- What features contribute best when the prediction works reasonably
- We proceed by feature ablation
  - Re-run prediction by removing 1 feature at a time.
  - 1000 samples

*HITs_Available* (Number of tasks in the batch)
*Age_Minutes* (how long ago the batch was created)

# Supply Elasticity

How does the market reacts when new tasks arrive on the platform?

# Supply Elasticity
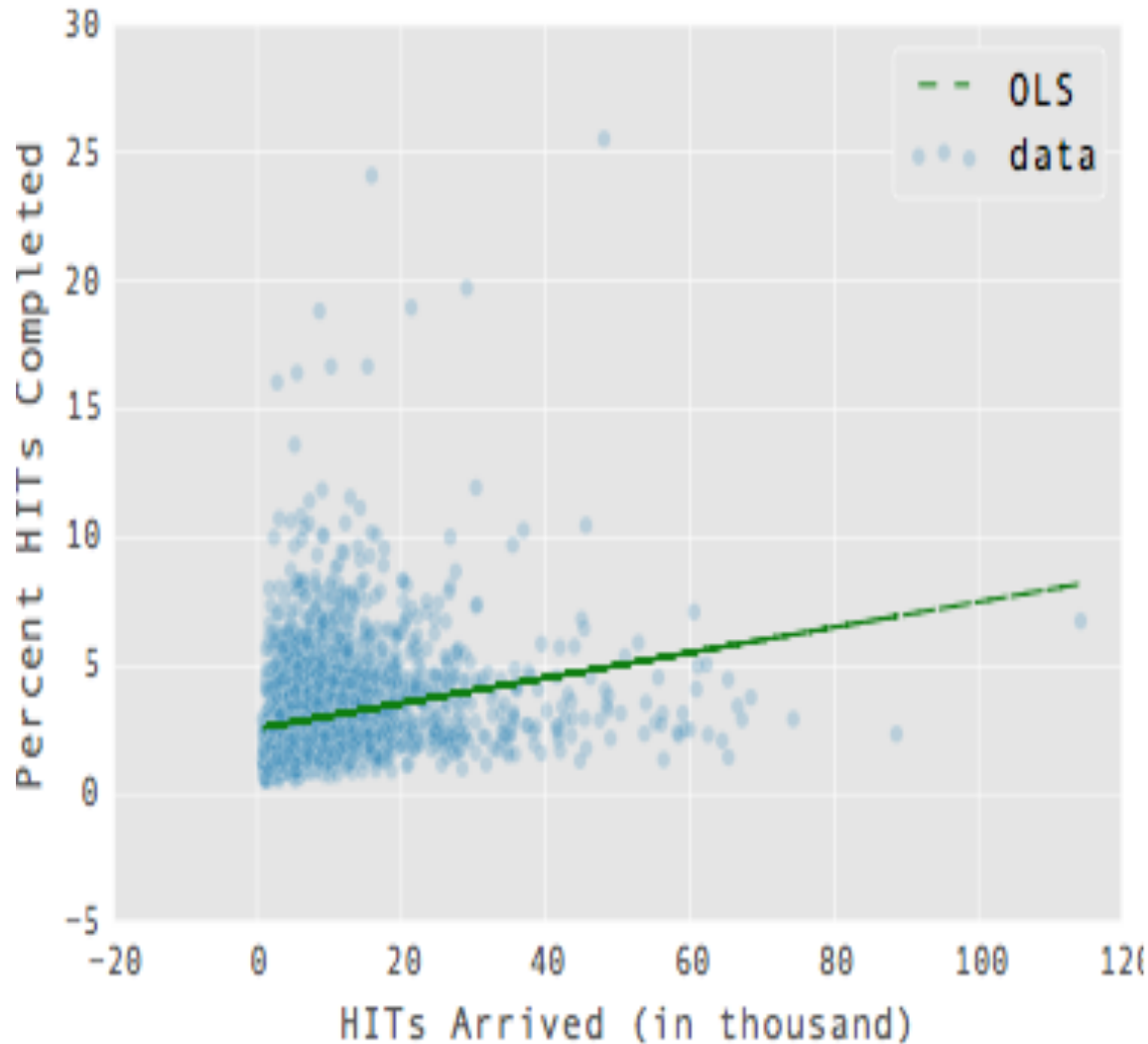


We regressed the percentage of work done (within 1 Hour) against the number of new HITs

# Supply Elasticity
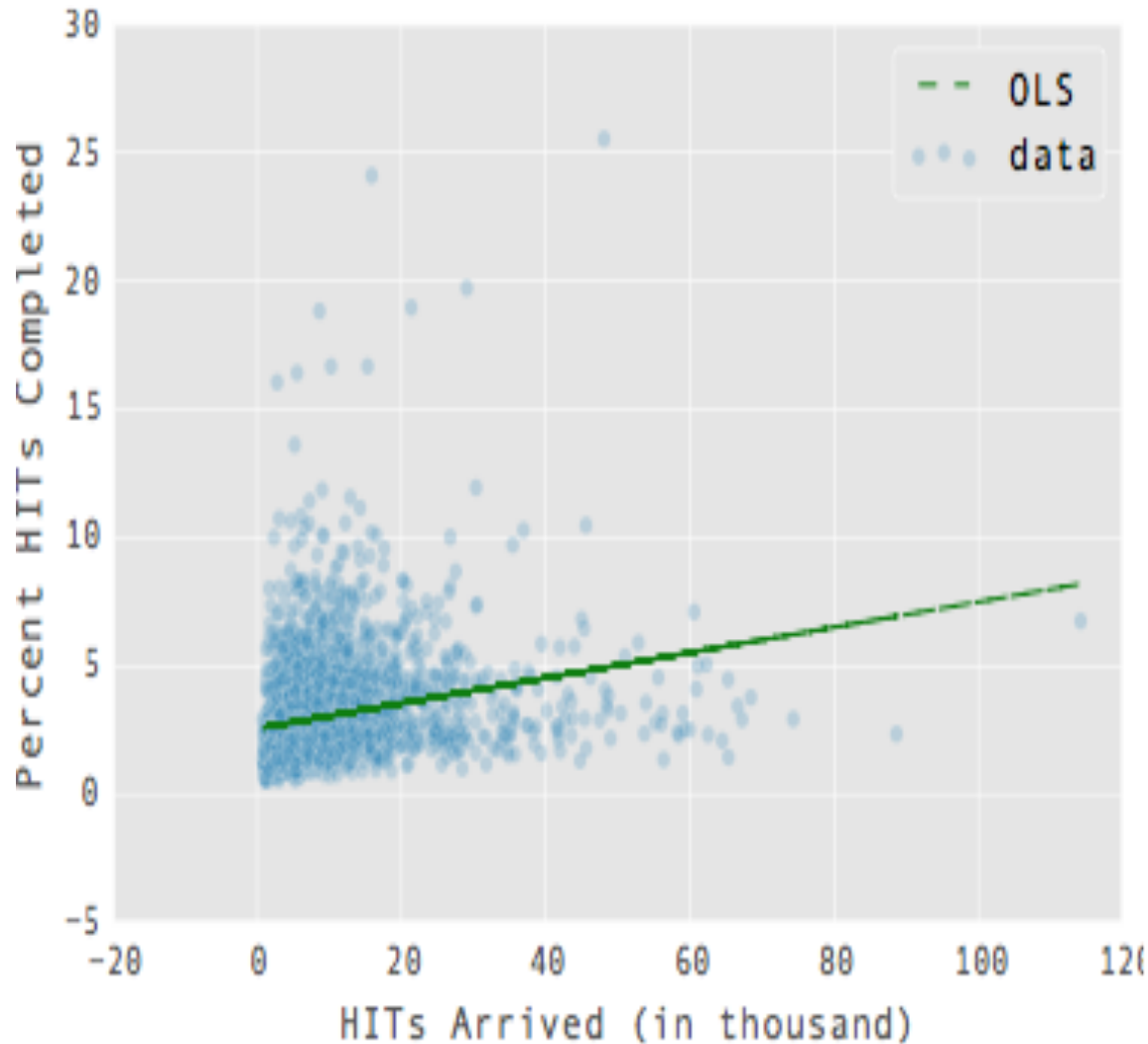


Intercept = 2.5
Slope = 0.5%

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

# Examples of micro-task crowdsourcing

# Information Retrieval

- A document collection

- A user with an information need

- The system has to rank documents by relevance to the user query

# IR Evaluation

- Test collections
  - A document collection
  - A set of queries
  - Manual (binary) relevance judgments

- Given the results of a search engine we can
  - Check which results are correct
  - Compute a quality measure

# Traditional Collections

- Setup data collection software / harness
- Recruit volunteers (often undergrads)
- Pay a flat fee for experiment or hourly wage
- Characteristics
  - Slow
  - Expensive
  - Tedious
  - Sample Bias

# IR Evaluation

- Relevance is hard to evaluate
  - Highly subjective
  - Expensive to measure
- Click data
- Professional editorial work

# Why Crowdsourcing for IR?

- Easy, cheap and fast labeling
- Ready-to use infrastructure – MTurk payments, workforce, interface widgets – CrowdFlower quality control mechanisms, etc.
- Allows early, iterative, frequent experiments – Iteratively prototype and test new ideas – Try new tasks, test when you want & as you go
- Proven in major IR shared task evaluations
  - CLEF image, TREC, INEX, WWW/Yahoo SemSearch

# Gamification of IR Evaluation

- GeAnn: http://www.geann.org/

- Relevance judgments with Gamification:
  - Text relevance
  - Image relevance

Quality through Flow and Immersion: **Gamifying Crowdsourced Relevance Assessments**. Eickhoff, C., C. G. Harris, A. P. de Vries, and P. Srinivasan. SIGIR 2012.

# 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
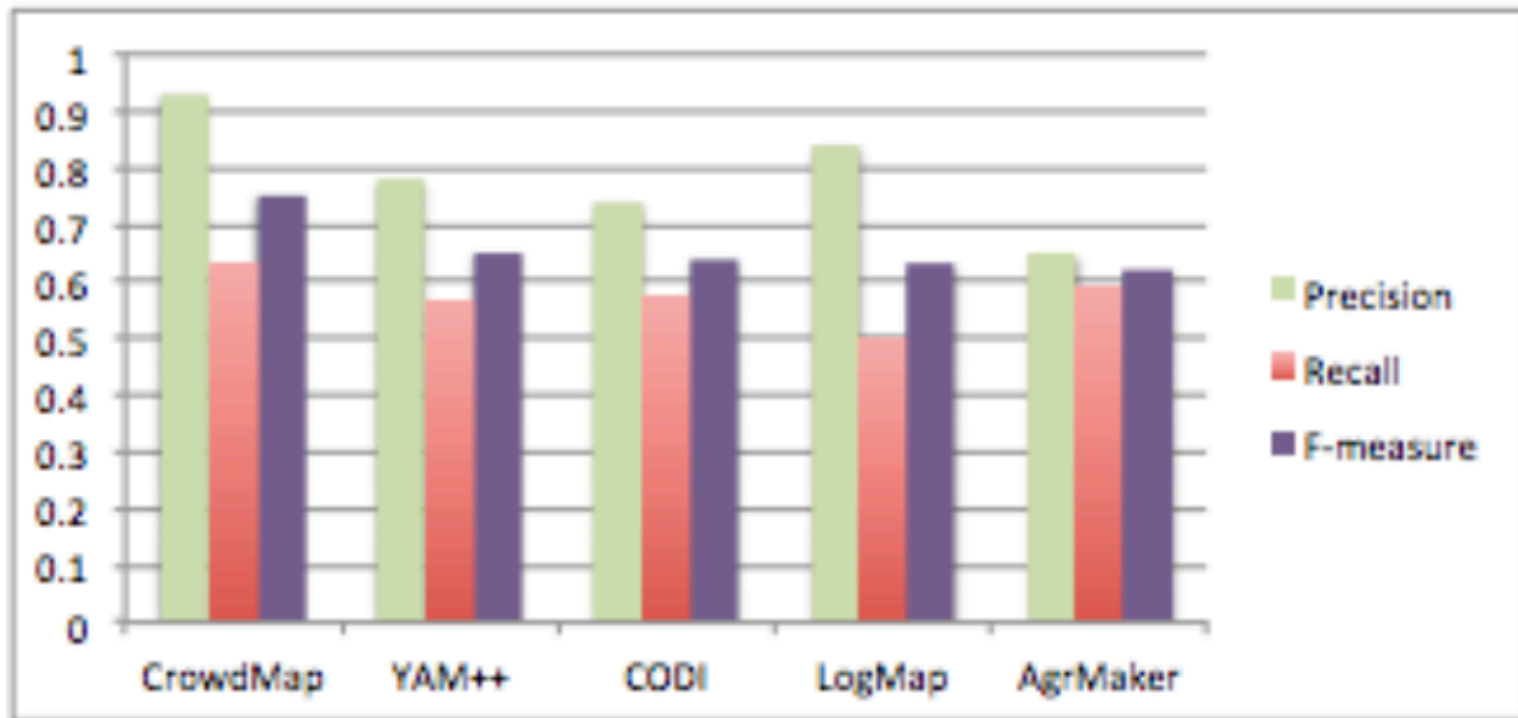
# Crowdsourcing Ontology Engineering

- 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

# Summary

- Micro-task crowdsourcing is a growing market
- Multiple platforms available
- Different ways of using it (more about this on Wednesday)