



28th European Summer School in  
Logic, Language and Information  
Bolzano-Bozen, 15-26 August, 2016

# Crowdsourcing Linguistic Datasets

Introductory Course at ESLLI, Bolzano, Italia, August 22-26, 2016

**Chris Biemann**

biem@cs.tu-darmstadt.de



---

## Some Words About The Crowds & Me

---

### Me

- from 10/2016: Professor for Language Technology, Computer Science Department, University of Hamburg, Germany
- 2011 – 9/2016: Assistant Professor for Language Technology, Computer Science Department, TU Darmstadt, Germany
- 2008 – 2010: Silicon Valley Search Engine Industry experience (Powerset, Bing)

### Crowds

- Successfully putting the crowds to work for me since Spring 2008
- Hands-On Crowdsourcing Workshop-Seminar in 2012 with 10 students
- 2015 Crowdsourcing Tutorial at AINL-FRUCT, St. Petersburg. Russia



---

## Let's Crowdfund!

### A Survey about you...

---

- European / non – European
- English native language?
- Ever worked as a crowdfunder?
- Ever employed crowdworkers?

---

# Crowdsourcing

---

## CROWDSOURCING (Howe 2006)

= **CROWD intelligence + outSOURCING**

“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”

Idea:

- Collective workforce of unskilled workers equals or beats the single expert (Wikipedia-principle, many eyes see more than two)
- Micropayments for microwork unlocks workforce that cannot work a regular job
- fast throughput through potentially large number of workers

B. Juska-Bacher, C. Biemann, U. Quasthoff (2012): Webbasierte linguistische Forschung: Möglichkeiten und Begrenzungen beim Umgang mit Massendaten. Linguistik Online,  
Howe, J. (2006): "The Rise of Crowdsourcing". Wired 14.6.



---

## Let's ask the Urban Dictionary

---

TOP DEFINITION



# Krautsource

to move your company overseas to Germany, to escape the inflation of the US dollar. also you are near countries that allow you to **launder** your money so it's isn't taxed by the Feds.

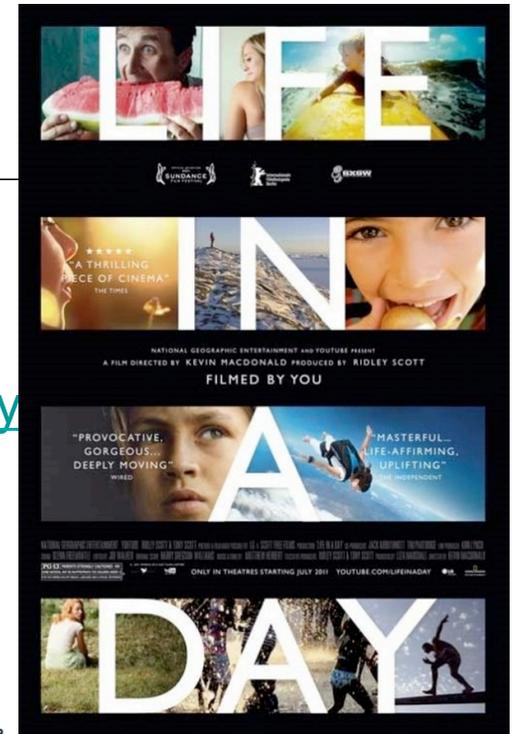
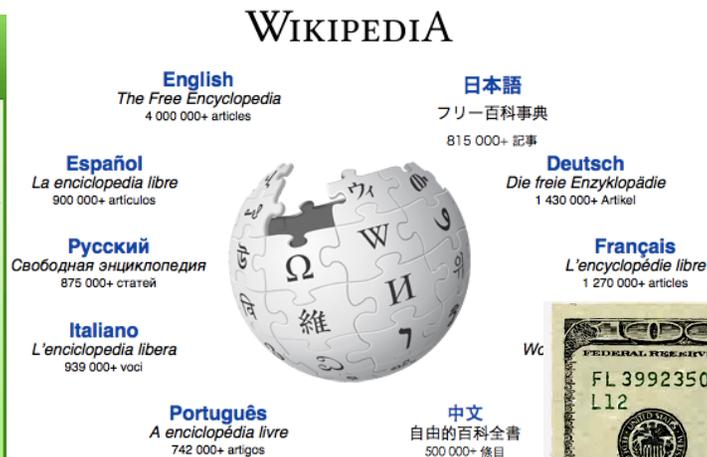
*While other companies are outsourcing their workers to china, my company KRAUTsourced by moving the HQ to Germany.*

by **unnp**y June 30, 2009



# Some Crowdsourcing Projects

- ESP Game (2004) (now Google Image Labeler)
- Wikipedia (2001 - ) <http://www.wikipedia.org/>
- Life in a Day (2010) <http://www.youtube.com/lifeinaday>



- Crowdsourcing 1 Image to 10000 Artists for 100 Dollars (2008) <http://www.labnol.org/internet/10000-artists-100-dollars-1-image/6103/>

---

## Invisible boy/girlfriend

# Finally. A **Girlfriend** your family can believe in.

Invisible Girlfriend gives you real-world and social proof that you're in a relationship - even if you're not - so you can get back to living life on your own terms.

MEET HER NOW

FEATURED ON

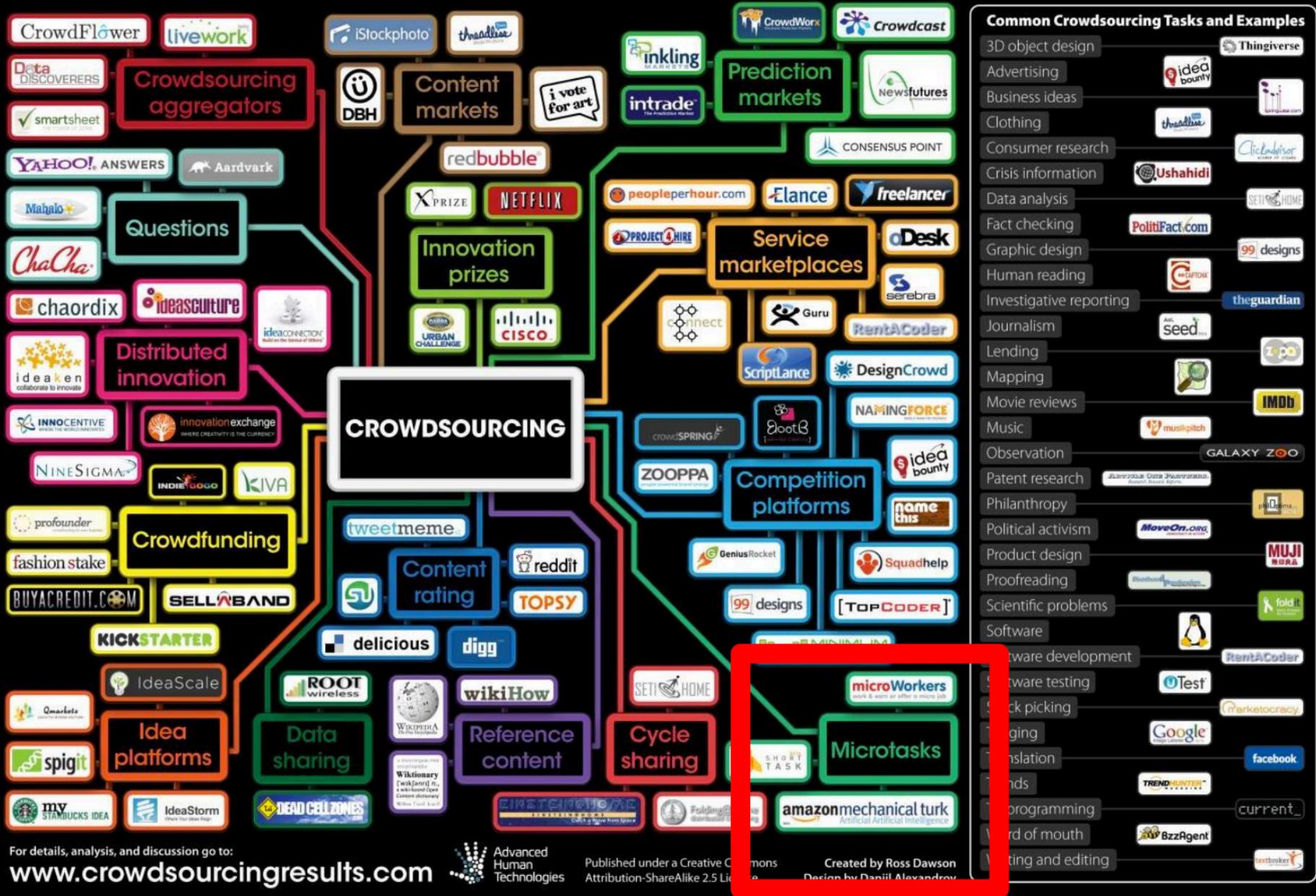
BUSINESS  
INSIDER

Buzz  
Feed

TIME

WIRED

# CROWDSOURCING LANDSCAPE Beta v1



For details, analysis, and discussion go to:  
[www.crowdsourcingresults.com](http://www.crowdsourcingresults.com)

Advanced Human Technologies

Published under a Creative Commons Attribution-ShareAlike 2.5 License

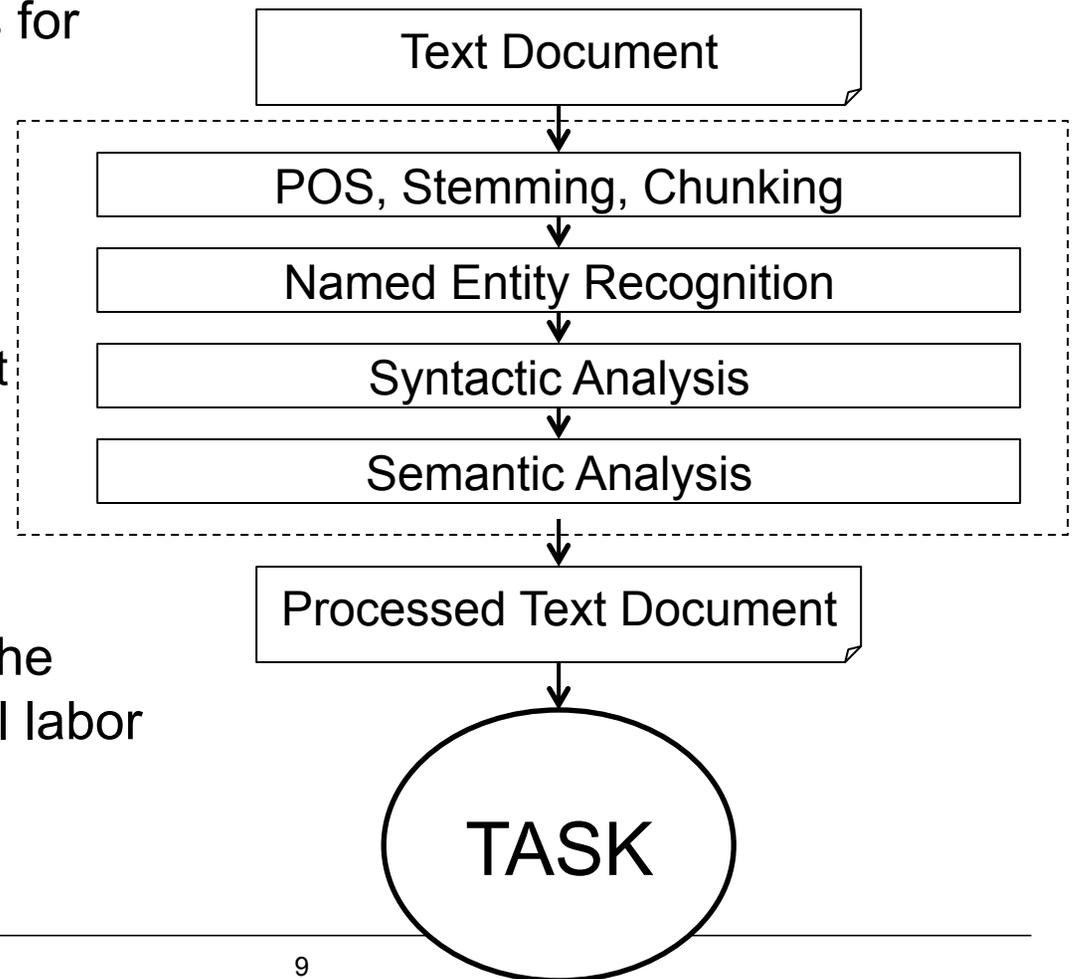
Created by Ross Dawson Design by Danil Alexandrov

---

# Language Technology Pipelines

---

- NLP-Pipeline: several linguistically motivated preprocessing steps for achieving a given task
- Single steps:
  - rule-based
  - relying on lexical resources
  - machine learning on training text
  - task definition through data



**Acquisition Bottleneck:** all of the above require significant manual labor

---

## Solution to Acquisition Bottleneck: Crowdsourcing!

---

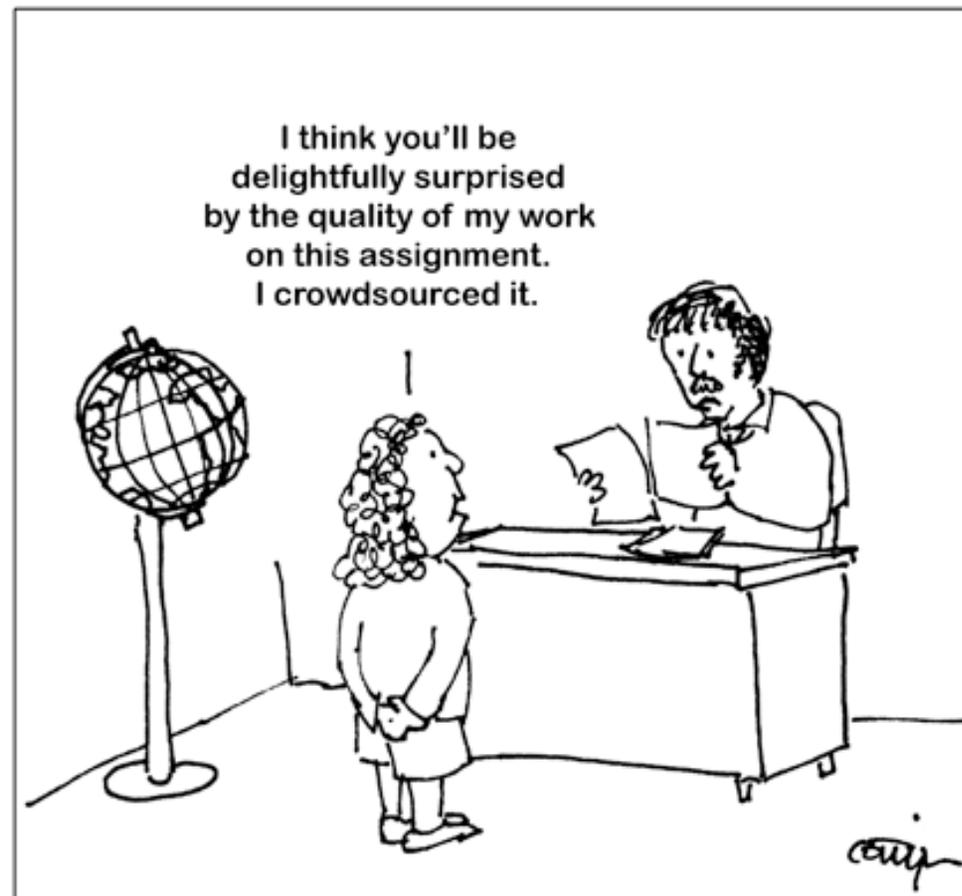
Crowdsourcable parts:

- tasks that are easy to explain
- tasks that are fun
- tasks where people reach high agreement
- tasks that require many small datapoints

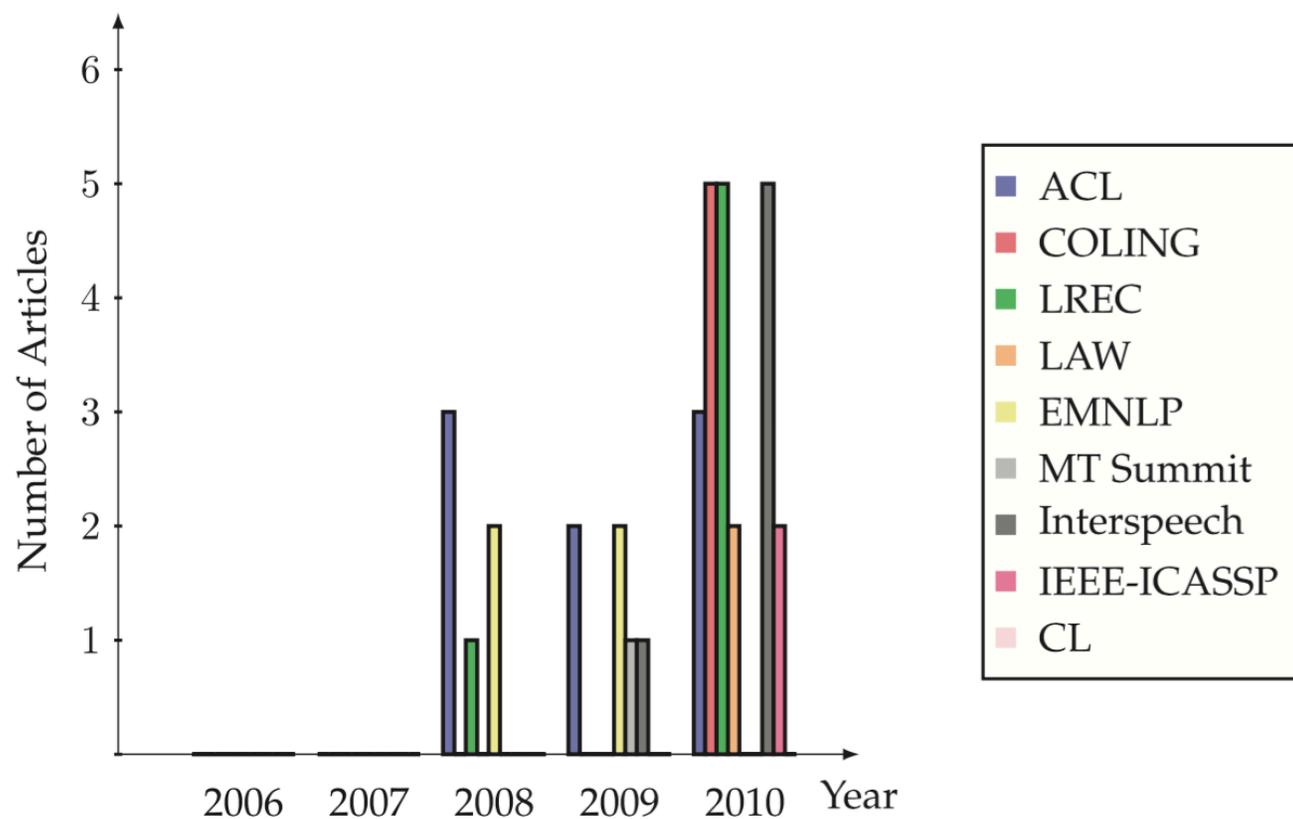
Not crowdsourcable:

- fuzzy tasks
- complex tasks
- boring tasks
- tasks where correct answer variety is high

NOISE TO SIGNAL  
Rob Cottingham



# MTurk goes NLP



**Figure 1**  
Evolution of MTurk usage in NLP publications.

Fort, K., Gilles A. and K. Bretonnel Cohen. "Amazon Mechanical Turk: Gold Mine or Coal Mine?." Computational Linguistics 37 (2011): 413-420.

---

## Outline of this Course

---

### 1. **What is Crowdsourcing?**

History and demographics, definitions, elementary concepts, example projects

### 2. **Crowdsourcing platforms**, esp. Amazon Mturk and Crowdfunder.

Technical possibilities, payment schemes, Do's and Don't's, schemes for ensuring quality

### 3. **Successful design patterns**

Crowdsourcing projects for language tasks

### 4. **Crowdsourcing projects for language tasks**

a variety of projects, and lessons learned

### 5. **Ethical considerations**

how to treat your crowdworkers, requester code of conduct, turker forums

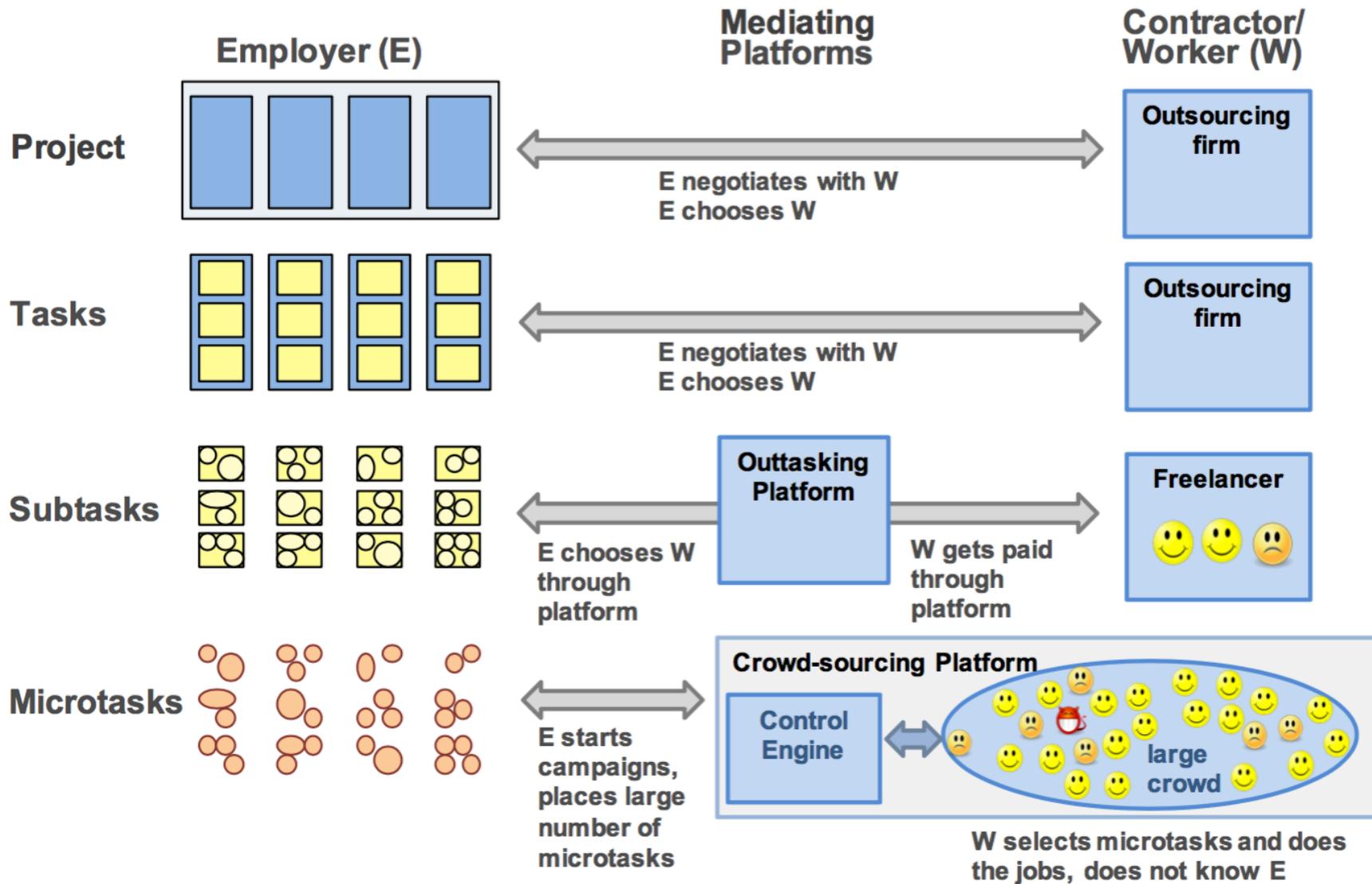
---

# Lesson 1: What is Crowdsourcing?

---

- Definitions
- Crowdsourcing Landscape
- Examples for Crowdsourcing
- Who are the Crowdworkers?
- Crowdsourcing Business Model

# Evolution of Work Organization



# Crowdsourcing Platforms

- **Amazon Mechanical Turk** <https://www.mturk.com>

- since 2005
- largest and most well-known platform
- highly flexible



Your Account

HITS

Qualifications

Introduction | Dashboard | Status | Account Settings

**Mechanical Turk is a marketplace for work.**

We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient.

**295,657 HITS** available. [View them now.](#)

- **Crowdfunder** <http://crowdfunder.com>

- automated quality control
- more formalized, easier to configure



- **MicroWorkers** <https://microworkers.com/>

- mostly SEO

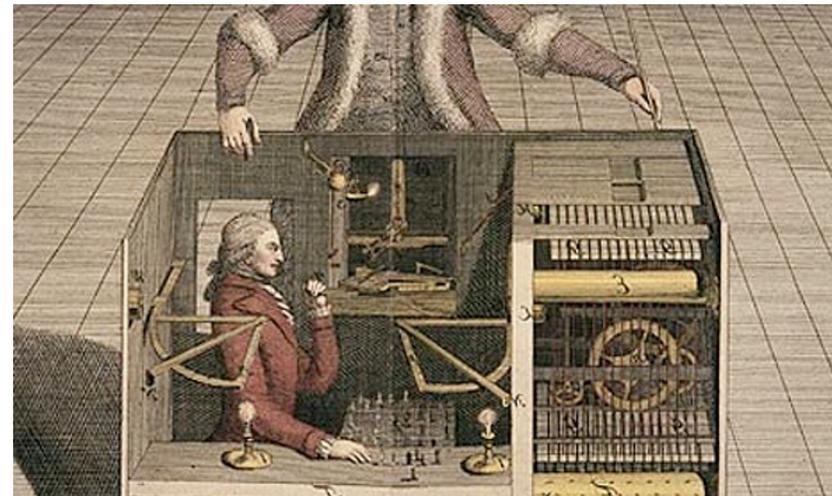
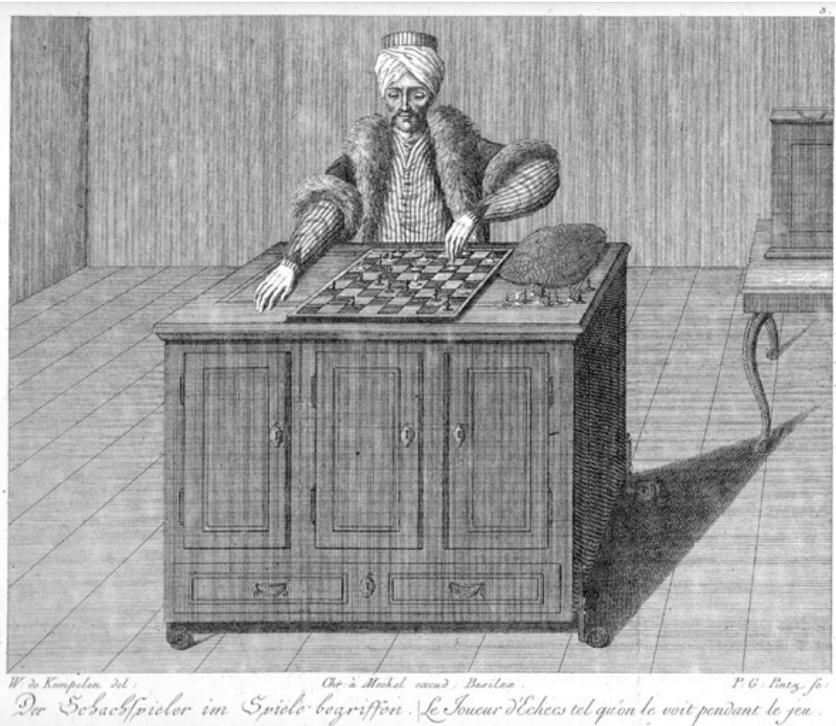
Most platforms sell 'campaign' packages or specialize in a few 'key' areas. We concentrate on those that allow the free definition of micro-tasks.



---

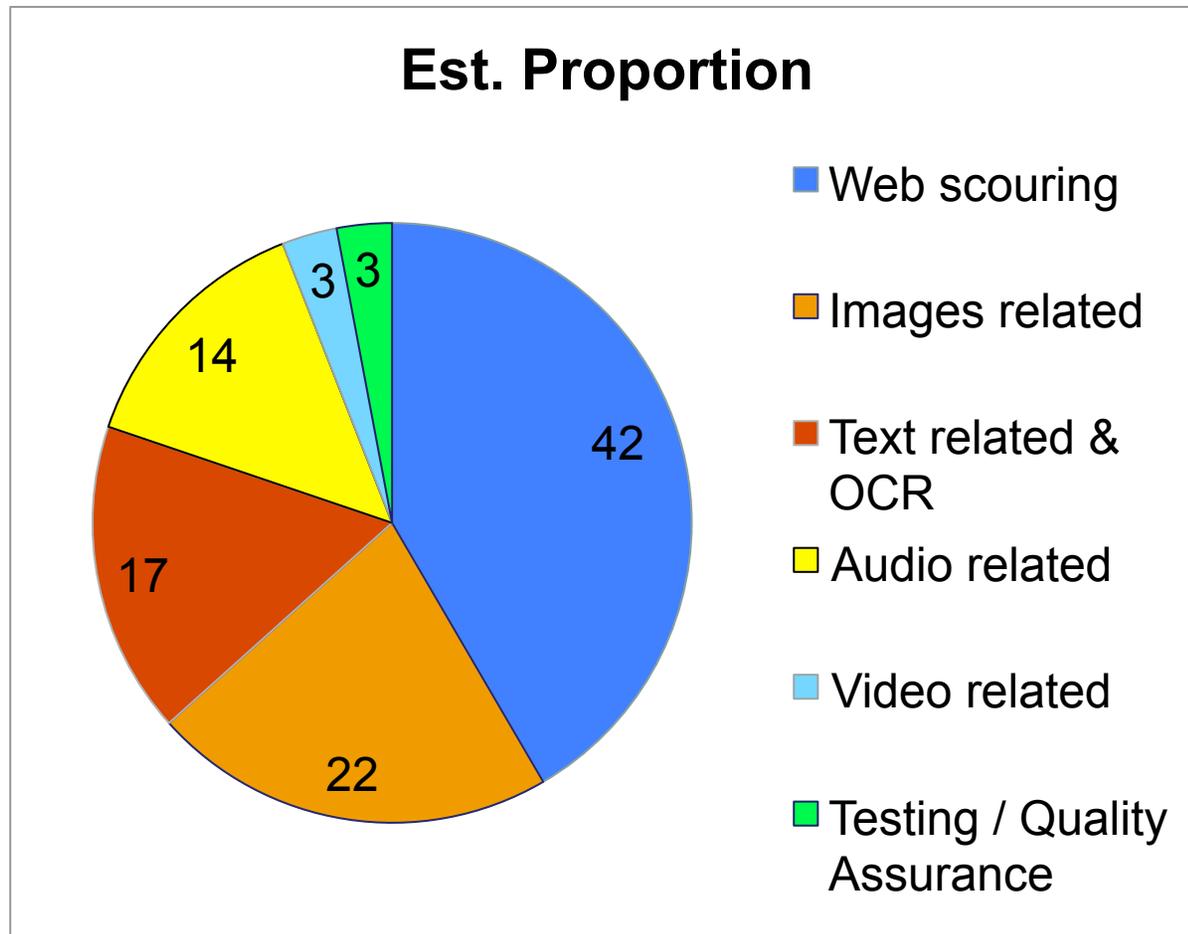
# The Mechanical Turk

---



- mechanical chess player (1770-1854) that was very hard to beat
- first use of a HPU (human processing unit) within an automaton

# Typical Microtasks on MTurk



## Web scouring:

- Verify a restaurant listing
- Match my products to Amazon products
- Find official websites for places, companies, ...
- Find the email addresses for wedding venues
- Find the school website and its school supply list
- Find Yelp reviews for businesses
- Categorize a Twitter search query
- Find a company name from an email domain
- Find main title, subtitle, and authors for a book
- Categorize web pages

---

# Popular Language Technology-Related Crowdsourcing Tasks

---

- Sentiment Classification (on Twitter)
- OCR/Transcription
- Search Relevance
- Data Collection / Information extraction templates
- Marketing Surveys
- Image and Content Tagging
- SEO: clicking on search engine results
- SEO: blog comments
- SEO: review writing

**For these kind of projects:** consider to outsource them to one of the providers.

**This tutorial:** focus on creating our own, specialized tasks.

---

---

# A sample Human Intelligence Task (HIT)

Contact information web research

**Requester:** Rob Tally

**Reward:** \$0.06 per HIT

**HITs Available:** 6742

**Duration:** 60 minutes

**Qualifications Required:** Total approved HITs is greater than 100, HIT approval rate (%) is not less than 85

## Contact information research

Your task is to find the contact information of: [Anne Loison](#)

(click on the name to find their contact information)

Additional Information about their workplaces (law firm, media agency, institute, bank etc.):

*Unité Mixte de Recherche n°5558 'Biométrie et Biologie Evolutive' Bâtiment Gregor Mendel Université Claude Bernard Lyon 1 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex France*

Please find their email address:

Please find their phone number (if available):

What is their professional country (if available):

Before you click on 'submit', make sure you found [Anne Loison's](#) contact details, and not their colleague's.

Thank you for your help!

You must ACCEPT the HIT before you can submit the results.

# Another sample HIT

## Classify Receipt

Requester: Jon Brellg

Reward: \$0.02 per HIT

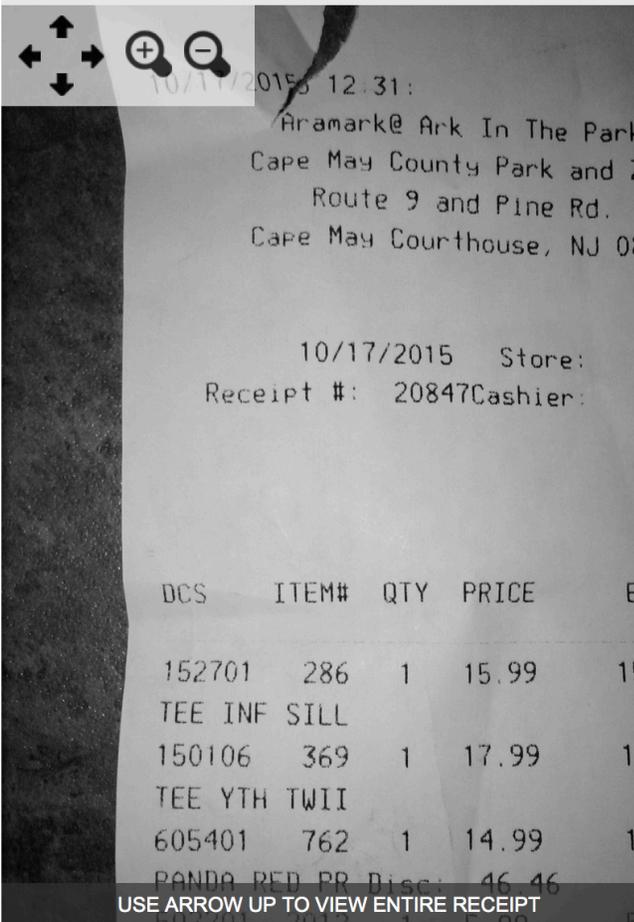
HITs Available: 64016

Duration: 20 minutes

Qualifications Required: Total approved HITs is not less than 1000; HIT approval rate (%) is not less than 97; Location is US

This is a preview of hit, click **ACCEPT HIT** above to complete

KEYBOARD SHORTCUTS: Scroll: Shift + up/down [Open Image](#)



## Classify Receipt

Hit Reward: \$0.02

Real readable original receipt

Not a receipt or not readable

The following details can often be found at the top or bottom of the receipt. Enter as much information as you can find.

Find and enter the business phone number:

Phone

Example: (888) 555-1234 or 8885551234

Find and enter the business address:

Address

City

State  Postal code

Example: 321 Fake Street, Los Angeles, CA, 90210

Next

USE ARROW UP TO VIEW ENTIRE RECEIPT

Want to work on this HIT?

Accept HIT

Want to see other HITs?

Skip HIT

---

## Early Crowdsourcing for Linguistic Annotations

---

- Grammaticality (Madnani et al. 2011)  
“is the following sentence grammatical? ‘It has a beautiful day!’”
- Prepositional Phrase Attachment (Jha et al. 2010)  
“I saw the man with the telescope”
- Semantic Alignment of Dictionary Definitions (Parent and Eskenazi, 2010)  
“pine cone” vs. “ice cream cone”
- Translation pairs (Zaidan & Callison-Burch, 2010)  
“translate ‘Ich bin ein Berliner’” into English”
- Paraphrase collection (Chen & Dolan, 2011)  
“describe what you see in the movie”
- Colors and Emotions of Words (Mohammad & Turney, 2011)
- Word Sense Inventory (Biemann & Nygaard, 2010)

---

## Workers per Country on AMT (2014)

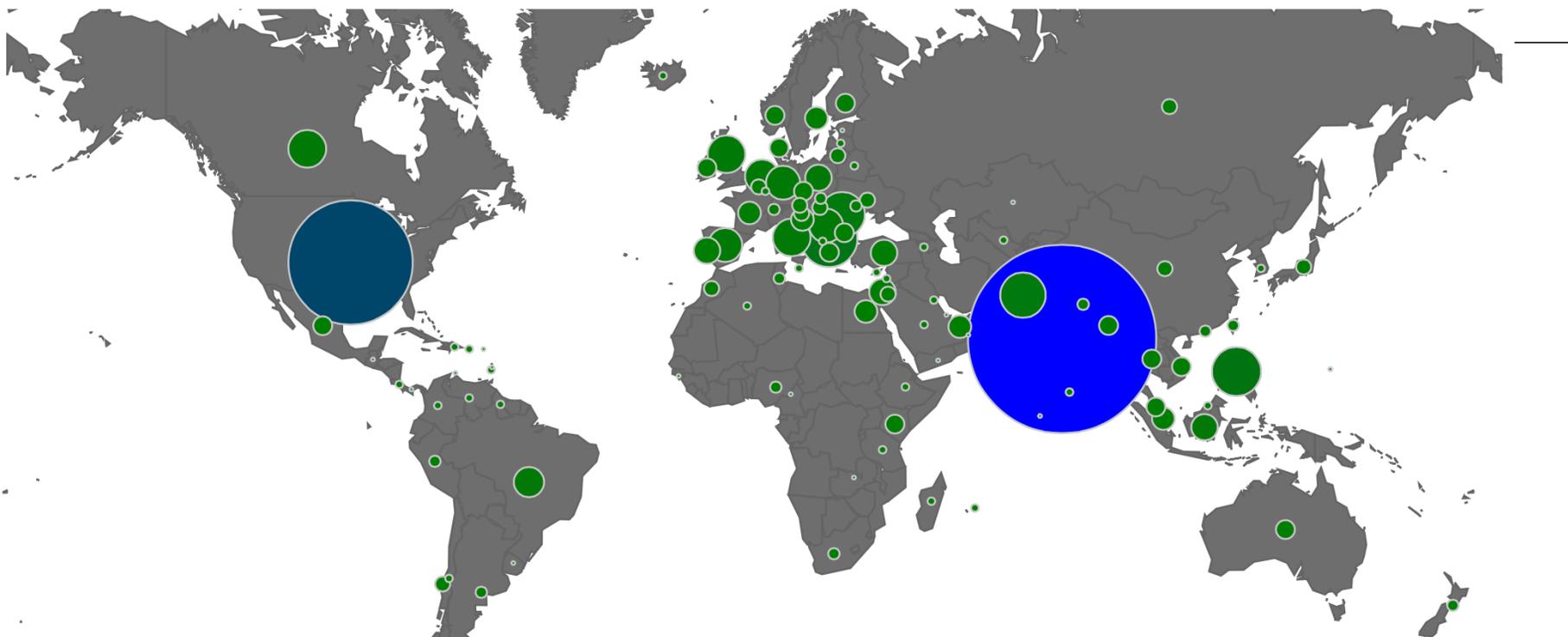
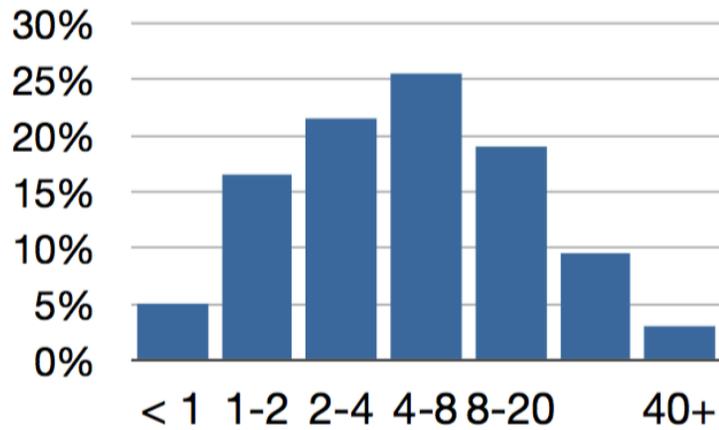


Figure 1: The number of workers per country. This map was generated based on geolocating the IP address of 4,983 workers in our study. Omitted are 60 workers who were located in more than one country during the study, and 238 workers who could not be geolocated. The size of the circles represents the number of workers from each country. The two largest are India (1,998 workers) and the United States (866). To calibrate the sizes: the Philippines has 142 workers, Egypt has 25, Russia has 10, and Sri Lanka has 4.

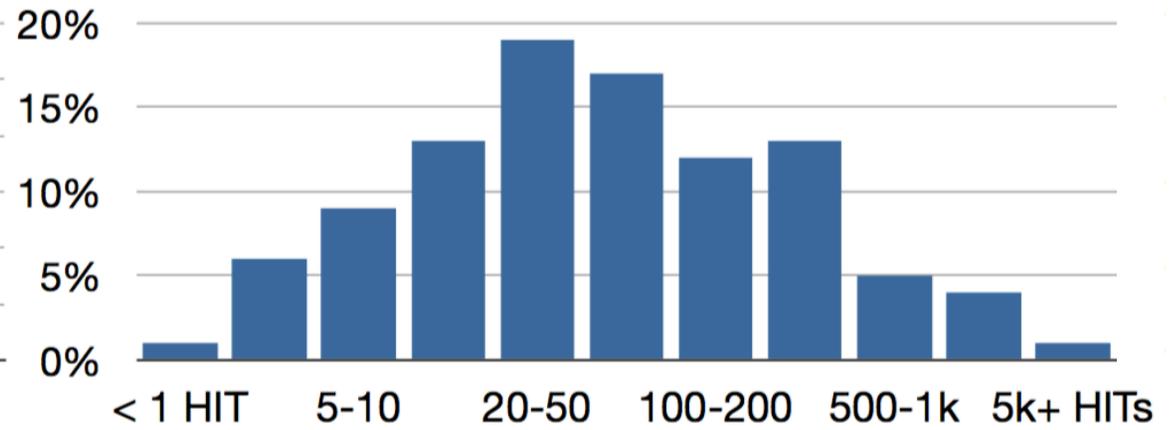
Pavlick, E., Post, M., Irvine, A., Kachaev, D., Callison-Burch, C. (2014): The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2 (2014) 79–92

# Hours, HITs, Wages and Reasons

**Hours spent on Mechanical Turk per week**

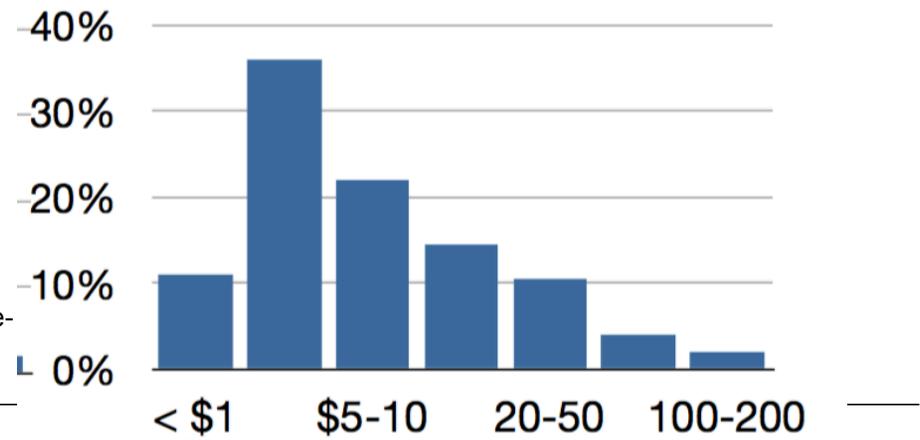


**Number of HITs completed per week**



Why do you complete tasks in MTurk?	US	India
To spend free time fruitfully and get cash (e.g., instead of watching TV)	70%	60%
For “primary” income purposes (e.g., gas, bills, groceries, credit cards)	15%	27%
For “secondary” income purposes, pocket change (for hobbies, gadgets)	60%	37%
To kill time	33%	5%
The tasks are fun	40%	20%
Currently unemployed or part time work	30%	27%

**Weekly income from Mechanical Turk**



Ipeirotis, P. (2010): New demographics of Mechanical Turk. <http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html>.

---

## Home Stories from 2006

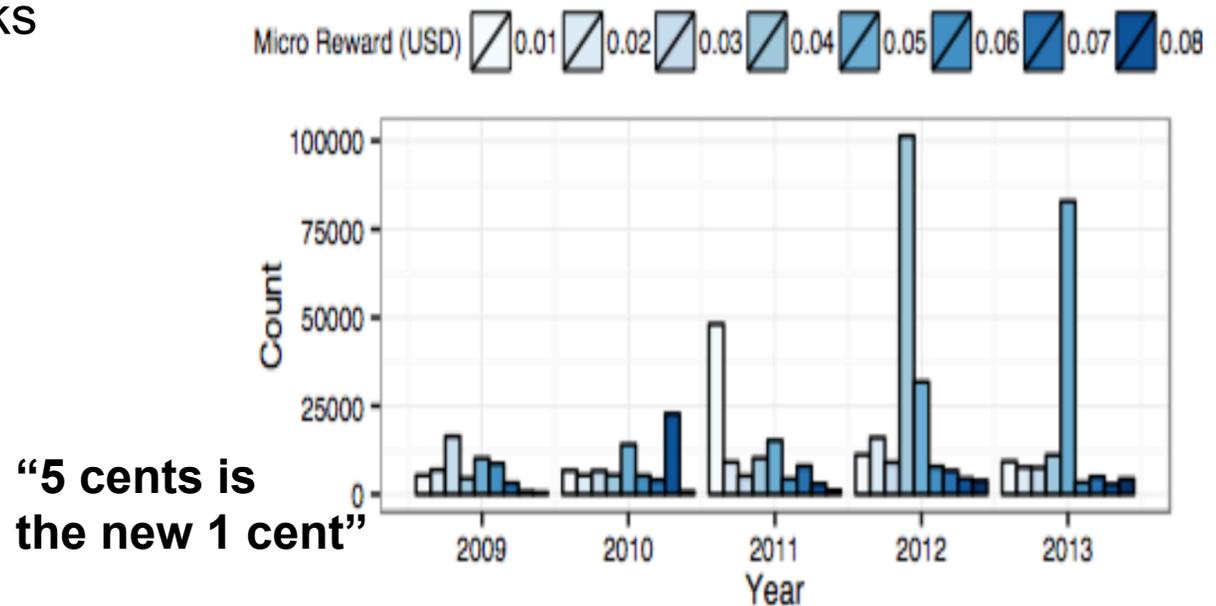
---

- One of those workers is Kristy Milland, 27, a mother of one [..] **“I have a lot of free time basically sitting at the computer while the kids play,”** she says. [..] When Mturk.com first began, Milland would churn through 3-cent HITs. [..] In an eight-hour day, when she didn’t have the kids to watch, Milland could go through 1,000 photos, making a cool \$30.
- Curtis Taylor, 50, [..] , who has earned more than \$345 on Mturk.com, doesn’t even think of turking as work. To him, it’s a way to kill time. **“I’m not in it to make money, I’m in it to goof off,”** he says. Taylor travels a lot for business and finds himself sitting around in hotel rooms at night. He doesn’t like to watch TV much, and says that turking beats playing free online poker. [..] . **“If I ever stop being entertained, I’ll stop doing it,”** he says. “I’ll just quit.”
- Why do people do it if the pay is so low? It’s a question Stranieri [a requester] wonders about himself. **“I think it’s something of a hybrid between trying to make money on the side and a diversion, a substitute for doing a crossword puzzle. It’s sort of a mental exercise.”**

[http://www.salon.com/2006/07/24/turks\\_3/](http://www.salon.com/2006/07/24/turks_3/)

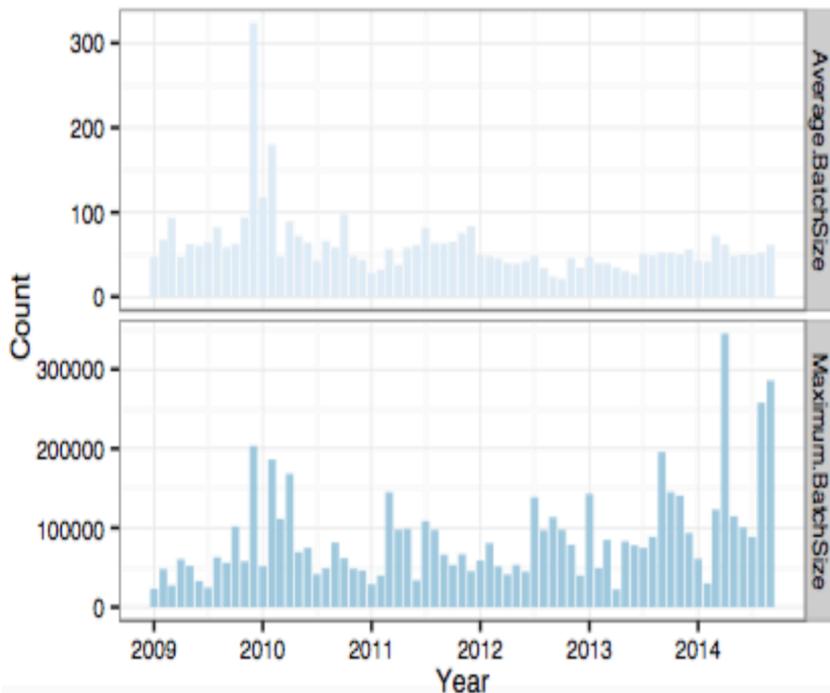
# Mturk-tracker.com

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

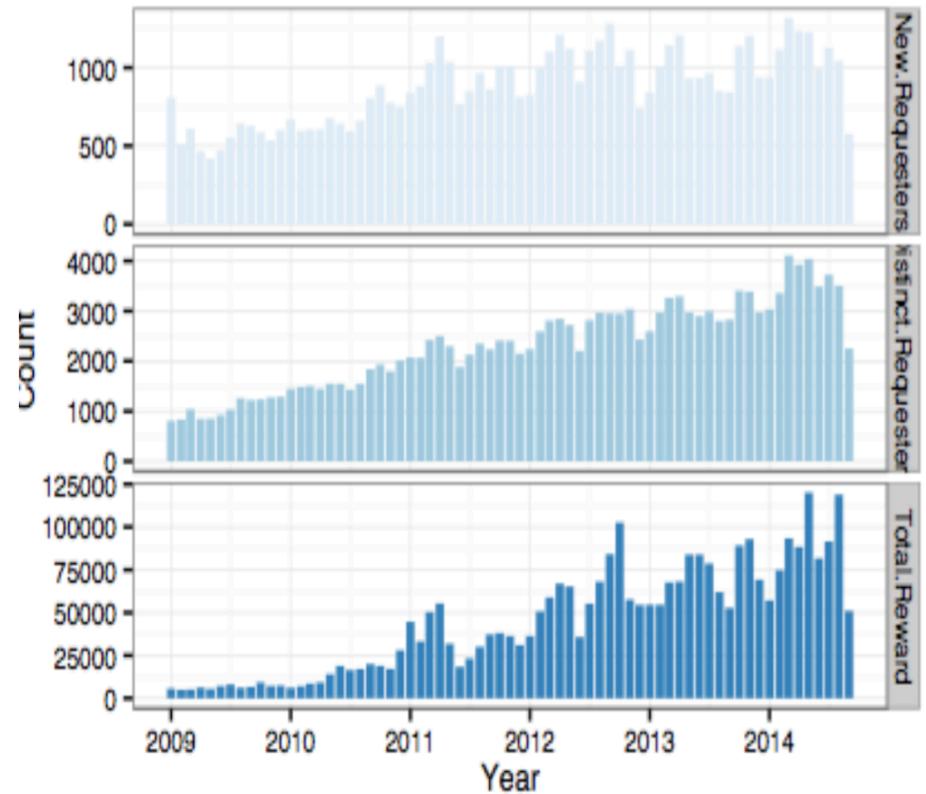


Difallah, D.E., Catasta, M., Demartini, G., Ipeirotis, P. G. and Cudré-Mauroux, P. (2015): The Dynamics of Micro-Task Crowdsourcing -- The Case of Amazon MTurk. In: 24th International Conference on World Wide Web (WWW 2015), Research Track. Firenze, Italy

# Mturk-tracker.com

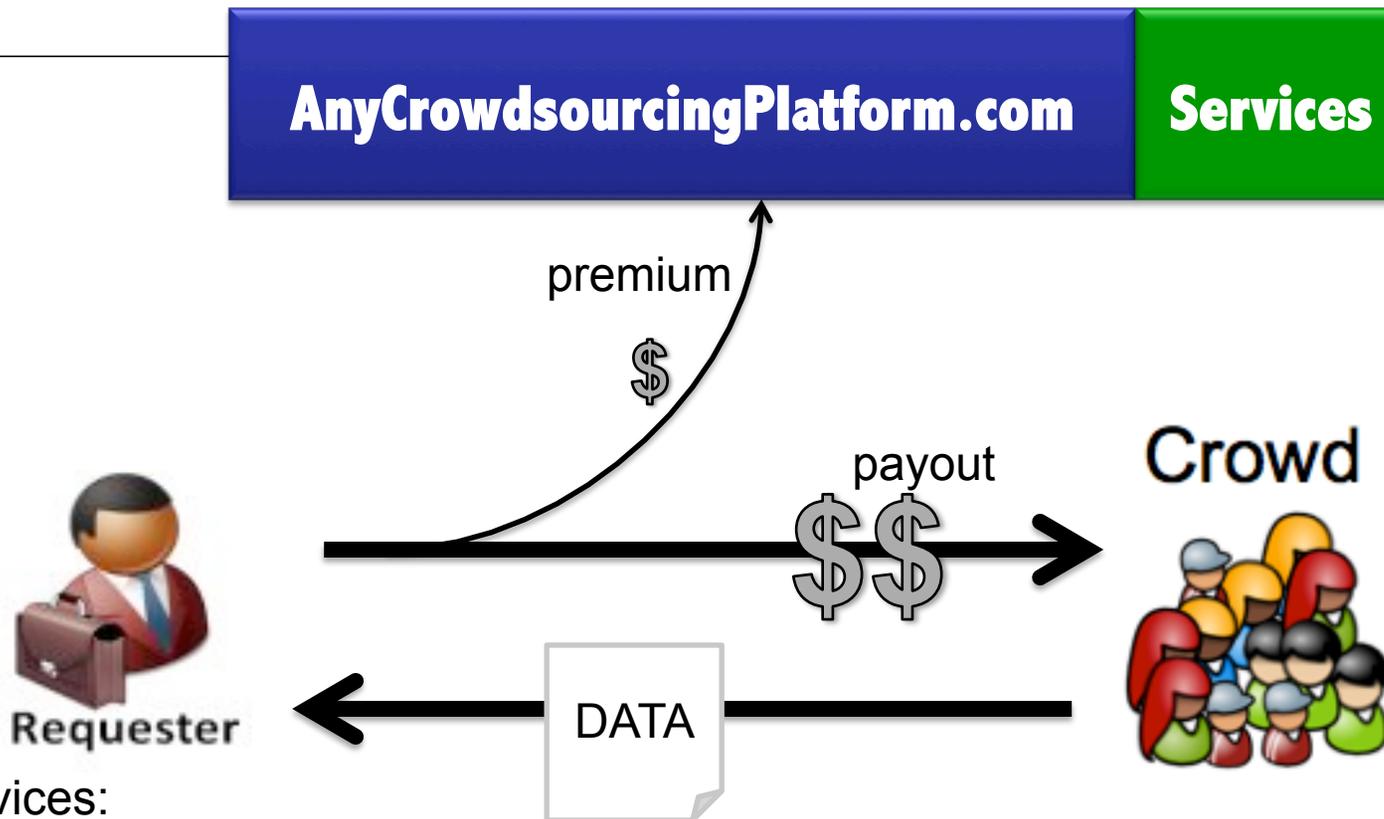


**very large batches start to appear**



**increasing number of new and distinct requesters**

# Business Model of Crowdsourcing Platforms



Services:

- connect requester and crowd
- hosting
- quality control mechanisms
- billing and taxes

Depending on services, the premium varies between 15% and 33%, plus extra fees

---

# Requesters and Workers

---

- Requester
  - defines a crowdsourcing task
  - accepts/rejects work
  - pays the bill
  - common requesters: search engines (relevance judgments), survey, data entry, transcription, marketing, researchers
- Worker
  - works on tasks
  - gets paid
  - tries to maximize her/his acceptance rate
  - common workers (MTurk):
    - in the beginning: mostly US population from all social statuses, slightly increased: young educated females
    - nowadays: mostly young males from India

---

## Main Issue: Quality Control

---

- workers try to get money quickly, using random work
- when checking the work of workers, there is no time saving as opposed to doing it yourself

### Mechanisms

- HIT acceptance rate, location
- qualifications: pre-tests
- test items with known answers
- redundancy
- verification through subsequent task

### Consequences (Eickhoff & de Vries, 2011)

- simple HITs
- quick HITs
- clear answers that allow redundancy
- no training required

**Never trust a  
single worker –  
trust the crowd!**

Eickhoff, C. & de Vries, A. (2011): "How Crowdsourcable is Your Task?". Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conf.

---

## A few words on Crowdsourcing Practice

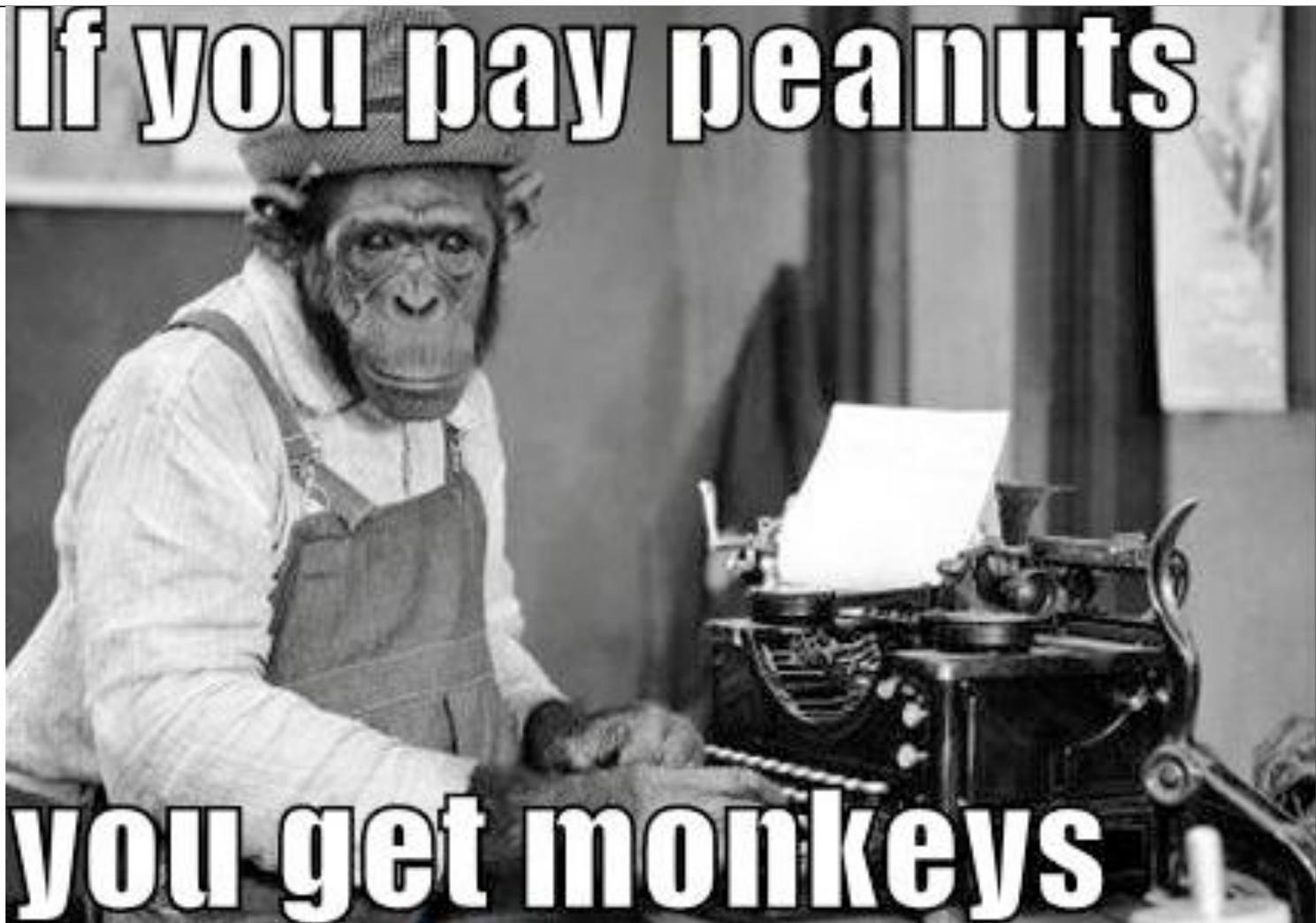
---

- In case of yes/no-answers, these have to be (more or less) balanced
  - workers learn quickly that always the same answer is 'right' and stop thinking
  - balance you dataset artificially by biasing test items, or appropriate filtering
- The benefit of doubt
  - don't get hung up on corner cases: strike the balance between easiness of task and well-definedness
- There is no 'best way'
  - try what works
  - try again
  - and make it work.
  - everything counts: the task is only as strong as its weakest link
- Don't underpay!
  - or do you want to be one of those bad capitalists that practice modern slavery?

---

**Quoting James Goldsmith (1933 – 1997)**

---



---

## **US Tax Laws – Form 1099**

**a law that was originally designed for highly-paid experts**

---

- IRS (US tax organization) Form 1099 “independent contractors”
  - “nonemployee” status reduces burden on employer, such as social security, health benefits, no statutory right to minimum wage, overtime pay, compensation for injuries sustained on the job, unemployment insurance if involuntarily separated from employment, or protection against discrimination.
  - contractors have to declare their income and pay taxes, employer does not have to withhold them
- Only yearly payouts of \$600 or more have to be declared by the employer, using Form 1099.

Law was originally designed for highly-paid experts.

This leads to serious complications for large-scale requesters, including Amazon themselves.

---

## Amazon Mechanical Turk: Characteristics

<https://www.mturk.com>

---

- old and popular platform
- bare-bone platform with very little in-built mechanisms for quality control
- free definition of tasks, can use tasks from your own server
- comparatively low premium of 20%, but at least 1 cent / HIT
- RESTful API
- need a US address for requester role
  - if you don't live in the US, get a forwarding address from services like <http://www.ipsparcels.com> if you are getting serious, or make up one if you're not
  - use your credit card with billing address from anywhere
- need to handle taxation bureaucracy on the requester side
  - payouts of more than \$600 / year for a single worker require a US tax form
  - avoid by tracking payouts and revoking qualifications

Summary: Full control (for the better or the worse), low premium, need some trickery for non-US-based requesters, beware the taxation issue.

---

---

## Crowdfunder: Characteristics

<http://crowdfunder.com>

---

- established meta-platform, distributes tasks over many platforms
- many mechanisms for quality control, main: gold items
- free definition of tasks, can use java script for complex interfaces
- comparatively high premium of 33%
- RESTful API
- works from anywhere with credit card
- taxation bureaucracy handling included in service
- academic accounts with minimal support, need to make data public
- pro accounts (with private data) need a minimum cash flow of several thousand USD / month!

Summary: Good quality control handling, high premium, high amount of services, access to many platforms, issues with proprietary data, good for business customers esp. outside of the US

---

---

## In a Nutshell: Learned in Lesson 1

---

- Crowdsourcing: distribution of microtasks to workers that get micropayments
- Crowdsourcing landscape
- Roles of requesters, crowdworkers, platform
- Demographics of crowdworkers: Mostly based in India and the US, mostly as a side job
- Examples of typical Human Intelligence Tasks (HITs)
- Suitable platforms for linguistic tasks