Slides Available here:
www.gianlucademartini.net/crowdsourcing/searchsolutions

Crowdsourcing for Data Processing and Search

Dr Gianluca Demartini

University of Sheffield

http://gianlucademartini.net

Gianluca Demartini

g.demartini@sheffield.ac.uk

- BSc, MSc at U. of Udine, Italy
- PhD 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)
- **Lecturer in Data Science** at the iSchool, U. of Sheffield
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 13, ISWC 13, SearchSolutions 2015

www.gianlucademartini.net

Research Interests

- **Entity-centric Information Access** (2005-now)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13)
  - NER in Scientific Literature(WWW 14) Prepositions (CIKM 14)
- **Hybrid Human-Machine Systems** (2012-now)
  - ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Memory-based Information Systems (WWW 14, PVLDB)
- **Better Crowdsourcing Platforms** (2013-now)
  - Pick-a-Crowd (WWW 13), Malicious Workers (CHI 15)
  - Scale-up Crowdsourcing (HCOMP 14), Dynamics (WWW 15)

Learning Objectives

- Demonstrate an understanding of **crowdsourcing** applications to search problems with its **opportunities** as well as its **limitations**;
- Demonstrate knowledge of the **common techniques** to be used in crowdsourced task design to **improve the quality** of the collected data;
- Discuss how crowdsourcing can be leveraged in **combination with machine-based algorithms** for data processing problems and to answer complex search queries;
- Discuss the benefits and challenges of applying crowdsourcing solutions for **search within the enterprise**.

Slides Available here:

www.gianlucademartini.net/crowdsourcing/searchsolutions

Introductions

- Name, role
- Interest / experience in Crowdsourcing / Data Processing / Search

Tutorial Outline

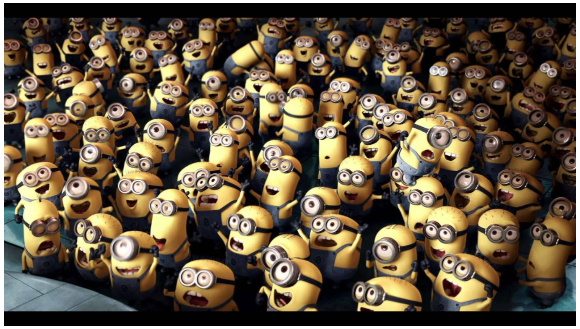
- Part 1
  - *Introduction to Crowdsourcing (30min)*
  - *Ensuring Quality in Paid Crowdsourcing (60min)*
- Part 2
  - *Hybrid Human-Machine Data Integration (30min)*
  - *Crowd-Powered Search (30min)*
  - *Enterprise Crowdsourcing for Search (30min)*

Introduction to Crowdsourcing

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

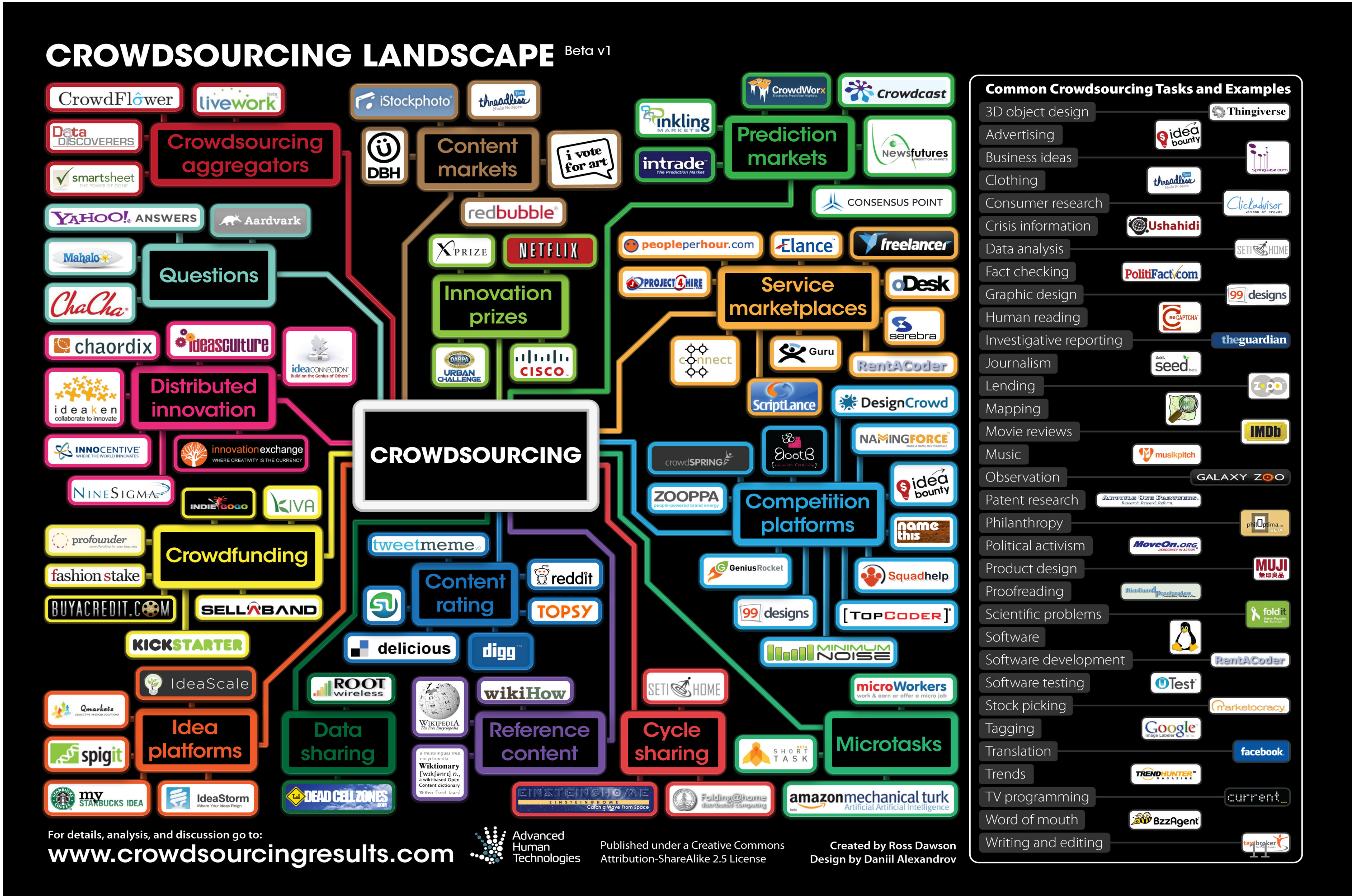
- Leverage human intelligence at scale to solve
  - Tasks simple for humans, complex for machines
  - With a large number of humans (the Crowd)
  - Small problems: micro-tasks (Amazon MTurk)
- Examples
  - Wikipedia, Image tagging
- Incentives
  - Financial, fun, visibility
- See also longer tutorial at ISWC 2013

mechanical turk
Artificial Artificial Intelligence
amazon
beta

Crowdsourcing Incentives

- Paid Crowdsourcing
- Fun (enjoyment)
  - Gamification
- Community (belonging, desire to help)
  - For example, Wikipedia

The Way Industry Looks At It



CROWDSOURCING LANDSCAPE Beta v1

CROWDSOURCING

Crowdsourcing aggregators: CrowdFlower, livework beta, Data Discoverers, smartsheet THE POWER OF DONE

Questions: YAHOO! ANSWERS, Aardvark, Mahalo, ChaCha

Distributed innovation: chaordix, ideasculture, ideaCONNECTION Build on the Genius of Others, ideaken collaborate to innovate, INNOCENTIVE WHERE THE WORLD INNOVATES, innovationexchange WHERE CREATIVITY IS THE CURRENCY, NINESIGMA

Crowdfunding: INDIEGOGO, KIVA, profounder, fashion stake, BUYACREDIT.COM, SELLABAND, KICKSTARTER

Idea platforms: IdeaScale, Qmarkets, spigit, my STARBUCKS IDEA, IdeaStorm Where Your Ideas Reign

Content markets: iStockphoto, threadless, DBH, i vote for art, redbubble

Innovation prizes: X PRIZE, NETFLIX, DARPA URBAN CHALLENGE, CISCO

Content rating: tweetmeme, reddit, TOPSY, su, delicious, digg

Data sharing: ROOT wireless, DEAD CELL ZONES .com

Reference content: WIKIPEDIA The Free Encyclopedia, wikiHow, Wiktionary

Cycle sharing: SETI@HOME, EINSTEIN@HOME Catch a Wave From Space, Folding@home

Microtasks: SHORT TASK beta, microWorkers work & earn or offer a micro job, amazon mechanical turk Artificial Artificial Intelligence

Prediction markets: CrowdWorx, Crowdcast, inkling markets, intrade The Prediction Market, Newsfutures, CONSENSUS POINT

Service marketplaces: peopleperhour.com, Elance, freelancer, PROJECT4HIRE, oDesk, serebra, connect, Guru, RentACoder, ScriptLance

Competition platforms: DesignCrowd, crowdSPRING, BootB, NAMINGFORCE, idea bounty, ZOOPPA people-powered brand energy, name this, GeniusRocket, Squadhelp, 99designs, TopCoder, MINIMUM NOISE

Common Crowdsourcing Tasks and Examples

3D object design	Thingiverse
Advertising	idea bounty
Business ideas	springwise.com
Clothing	threadless
Consumer research	Clickadvisor wisdom of crowds
Crisis information	Ushahidi
Data analysis	SETI@HOME
Fact checking	PolitiFact.com
Graphic design	99 designs
Human reading	reCAPTCHA
Investigative reporting	theguardian
Journalism	Aol. seed beta
Lending	zopa
Mapping	
Movie reviews	IMDb
Music	musikpitch
Observation	GALAXY ZOO
Patent research	Article One Partners Research. Reward. Reform.
Philanthropy	philOptima BETA
Political activism	MoveOn.ORG DEMOCRACY IN ACTION
Product design	MUJI 無印良品
Proofreading	Distributed Proofreaders
Scientific problems	fold it Solve Puzzles for Science
Software	
Software development	RentACoder
Software testing	uTest
Stock picking	marketocracy
Tagging	Google Image Labeler BETA
Translation	facebook
Trends	TRENDHUNTER MAGAZINE
TV programming	current
Word of mouth	BzzAgent
Writing and editing	textbroker

For details, analysis, and discussion go to:
www.crowdsourcingresults.com

Advanced Human Technologies

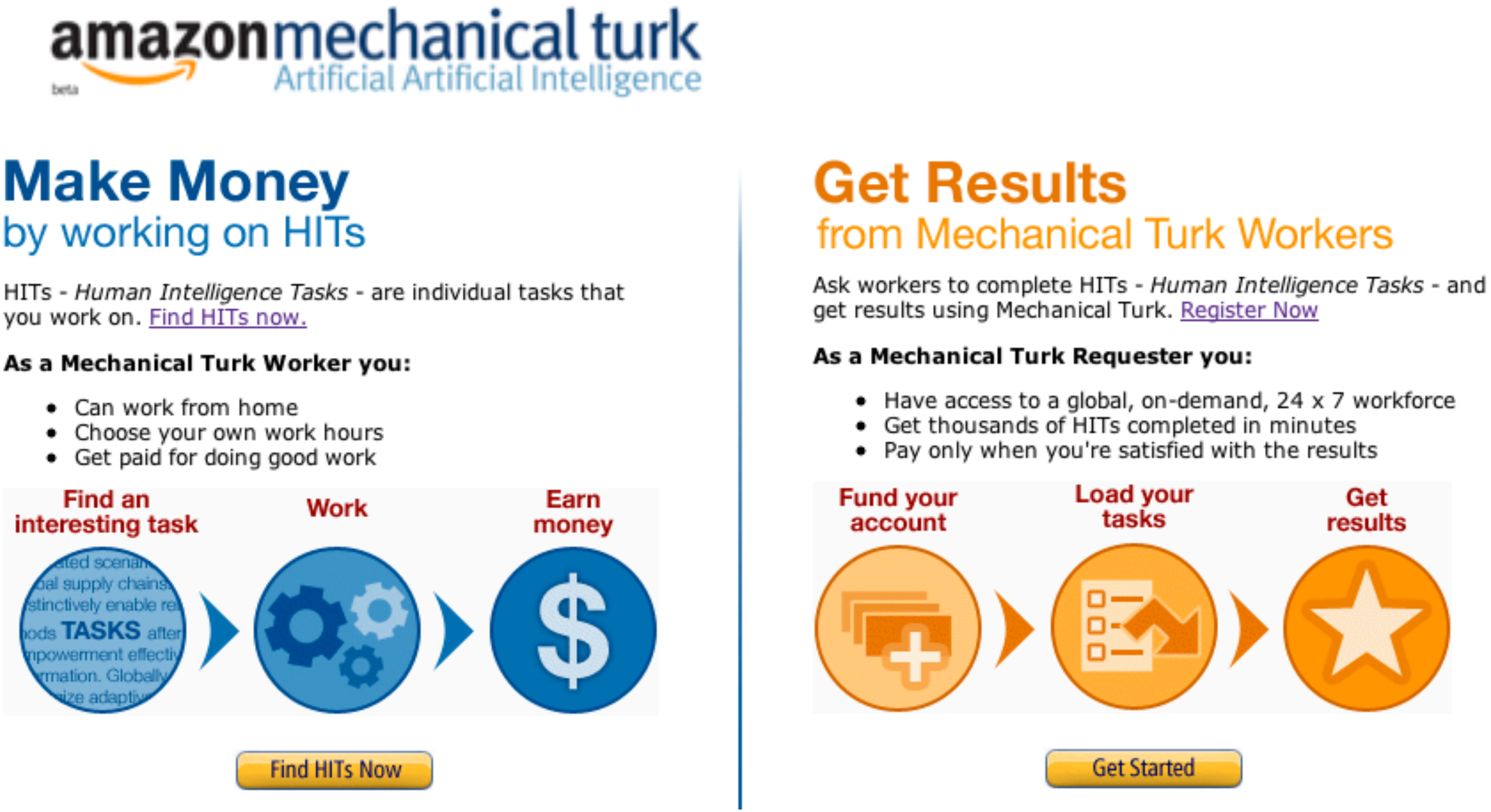
Published under a Creative Commons Attribution-ShareAlike 2.5 License

Created by Ross Dawson
Design by Daniil Alexandrov

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



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

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

Amazon MTurk

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

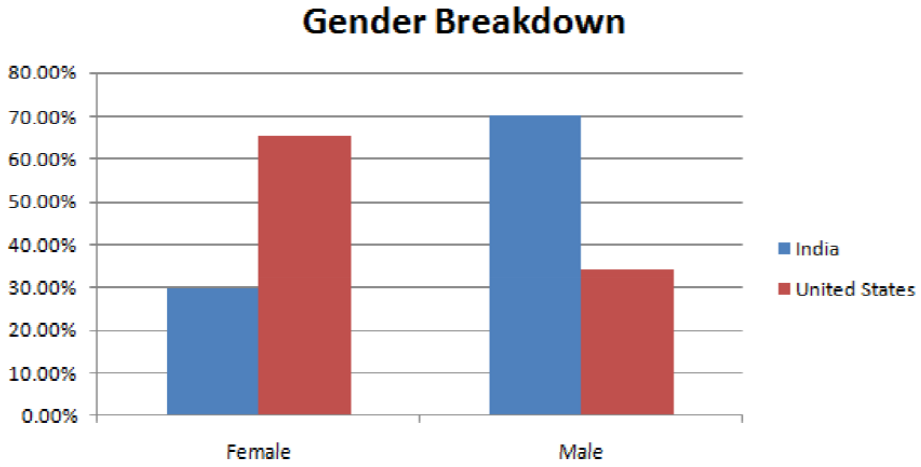
- If the results are approved, workers are paid

Demographics of MTurk workers in 2009

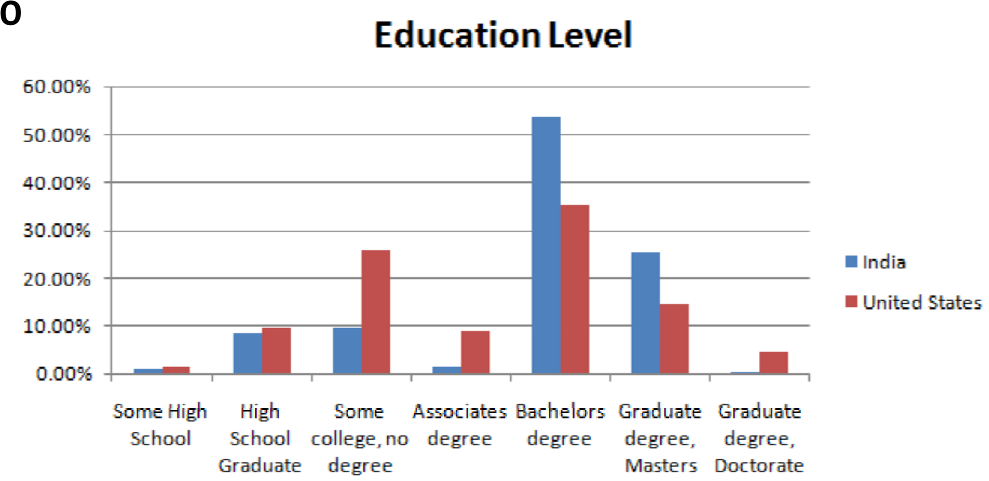
Country of residence

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

2013 Statistics:
1M workers
10% active

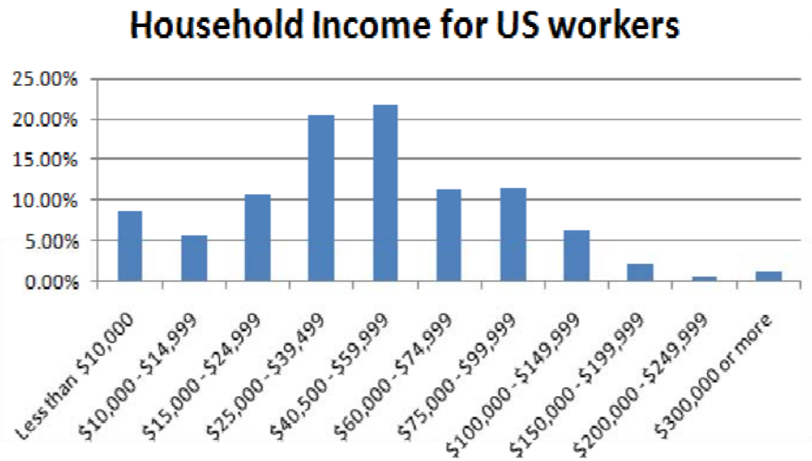


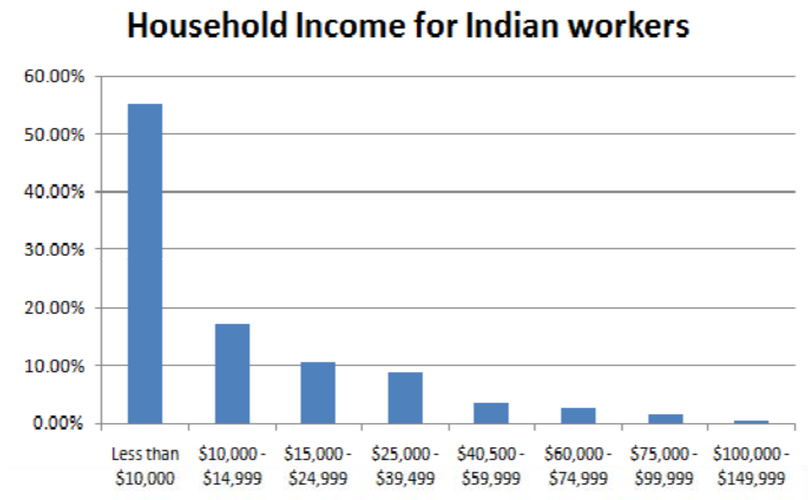
Gender Breakdown

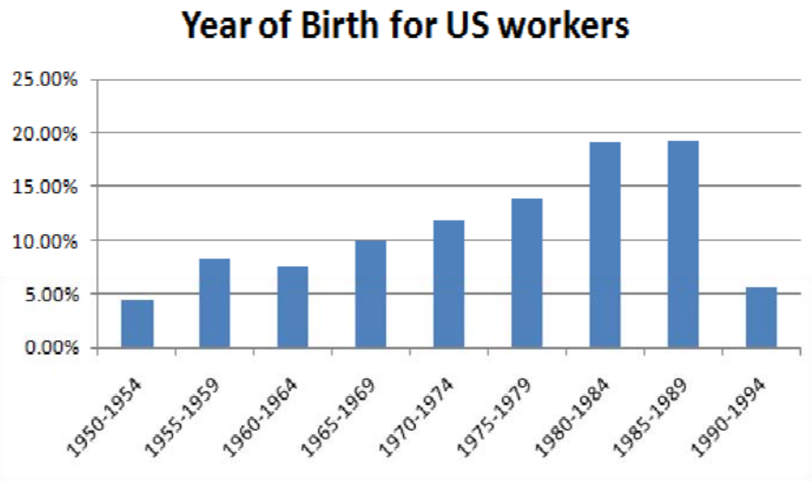


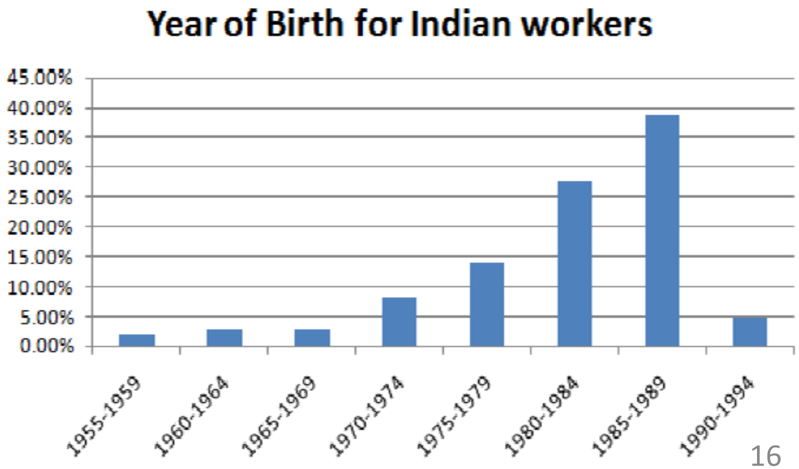
Education Level

Demographics of MTurk workers in 2009

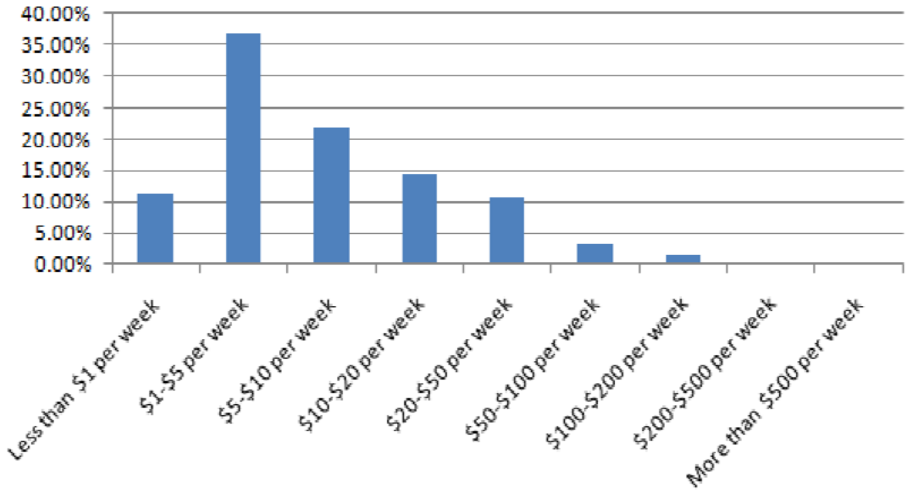

Household Income for US workers

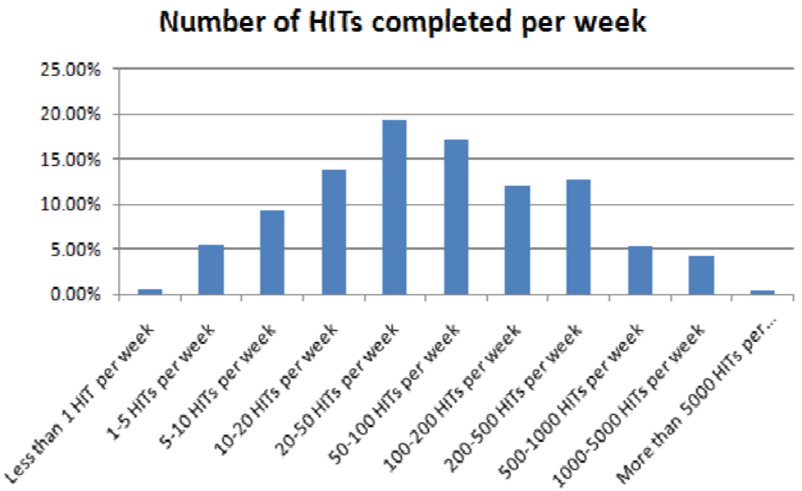

Household Income for Indian workers

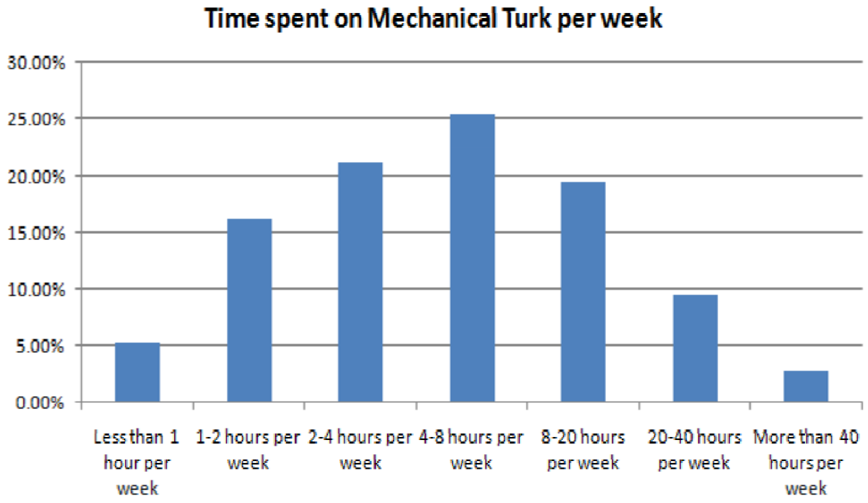

Year of Birth for US workers


Year of Birth for Indian workers

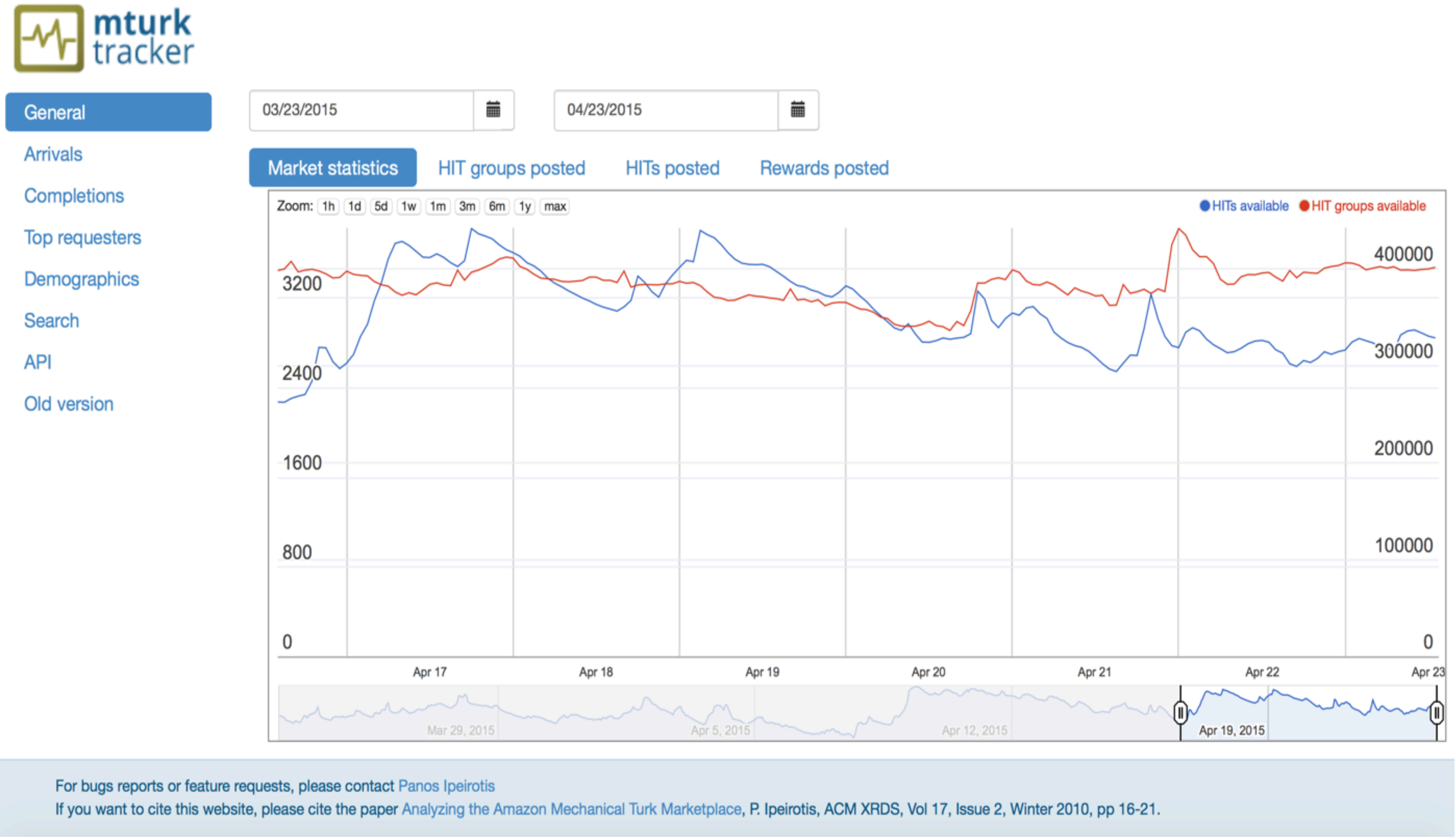
Demographics of MTurk workers in 2009







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



03/23/2015

04/23/2015

For bugs reports or feature requests, please contact Panos Ipeirotis

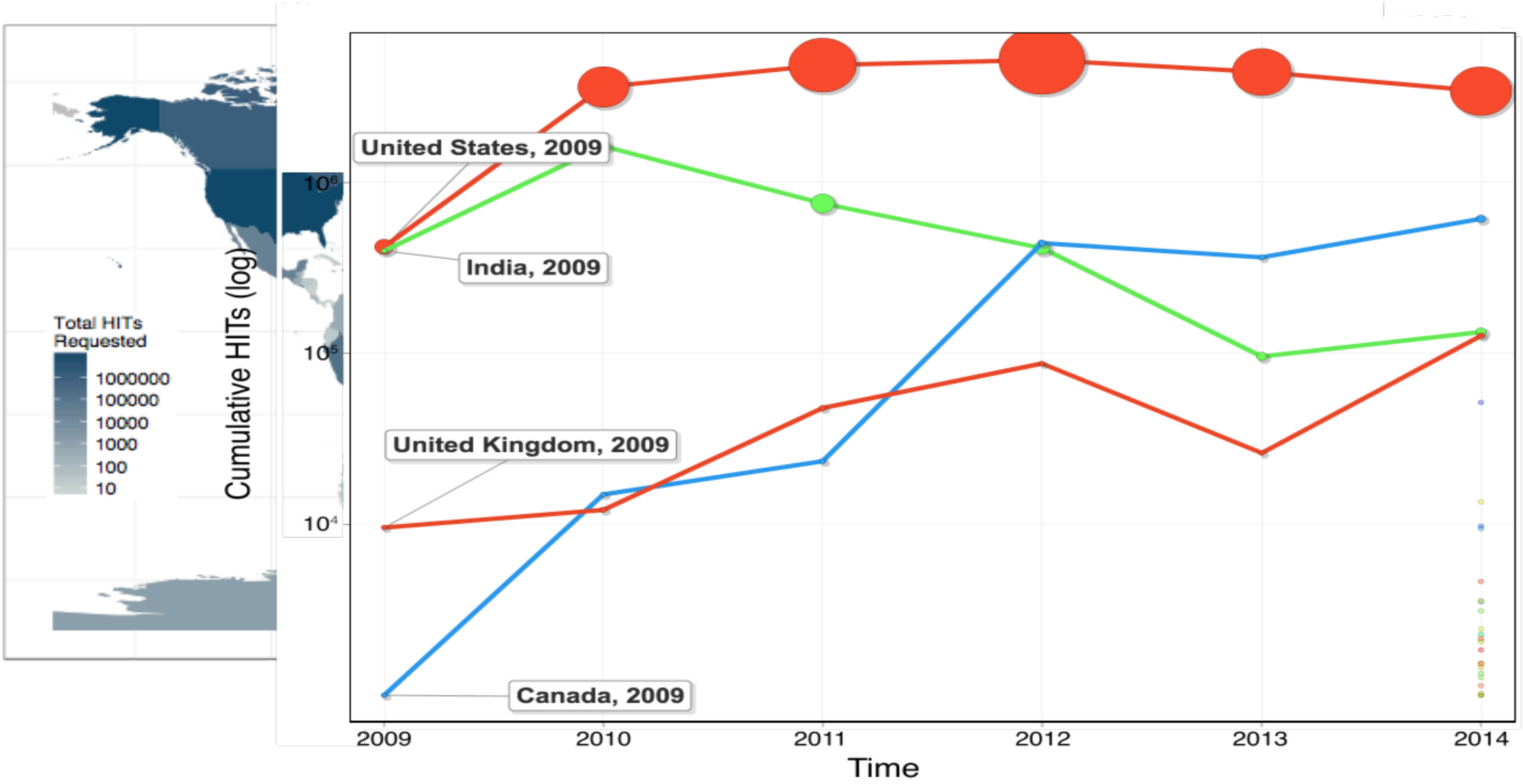
If you want to cite this website, please cite the paper Analyzing the Amazon Mechanical Turk Marketplace, P. Ipeirotis, ACM XRDS, Vol 17, Issue 2, Winter 2010, pp 16-21.

5-year Analysis of MTurk workload

- 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

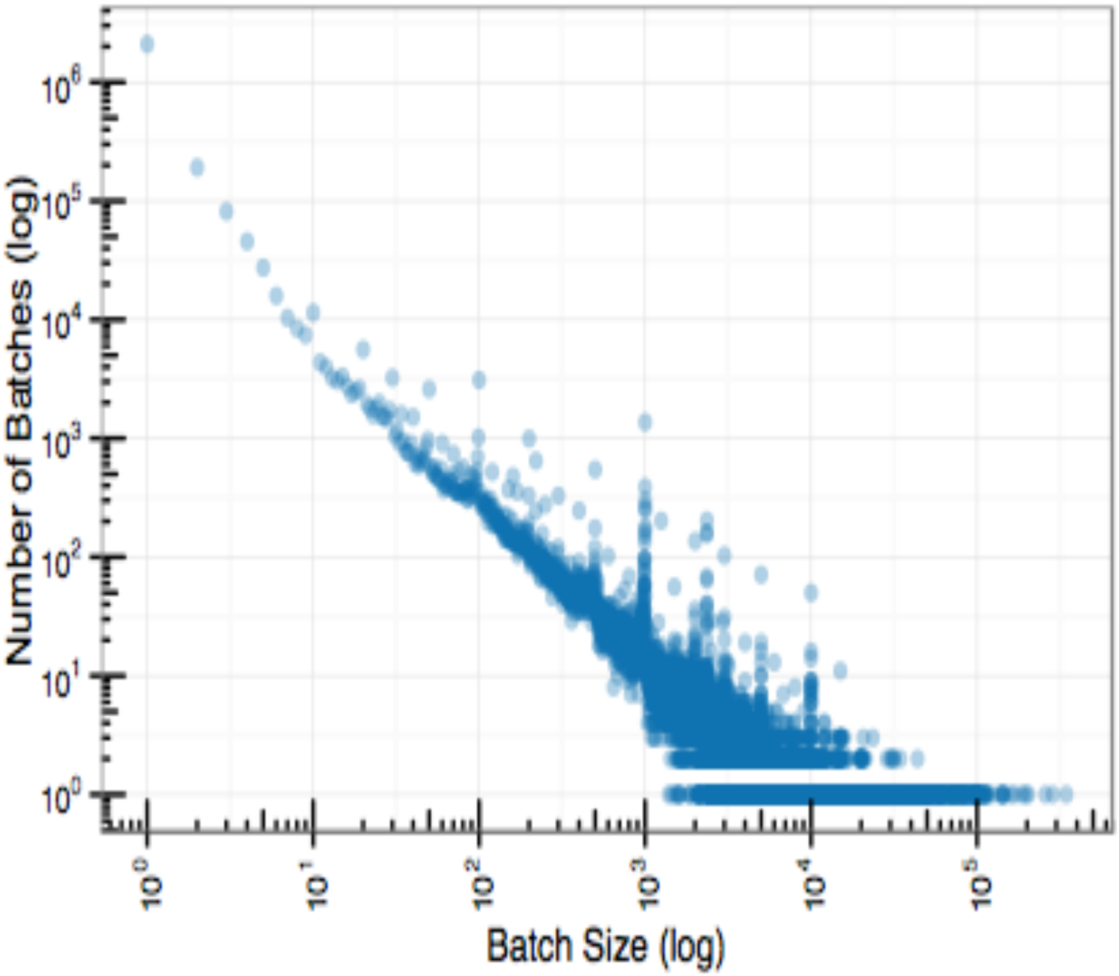
Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. **The Dynamics of Micro-Task Crowdsourcing -- The Case of Amazon MTurk**. In: 24th International Conference on World Wide Web (**WWW 2015**), Research Track. Firenze, Italy, May 2015.

Country-Specific HITs



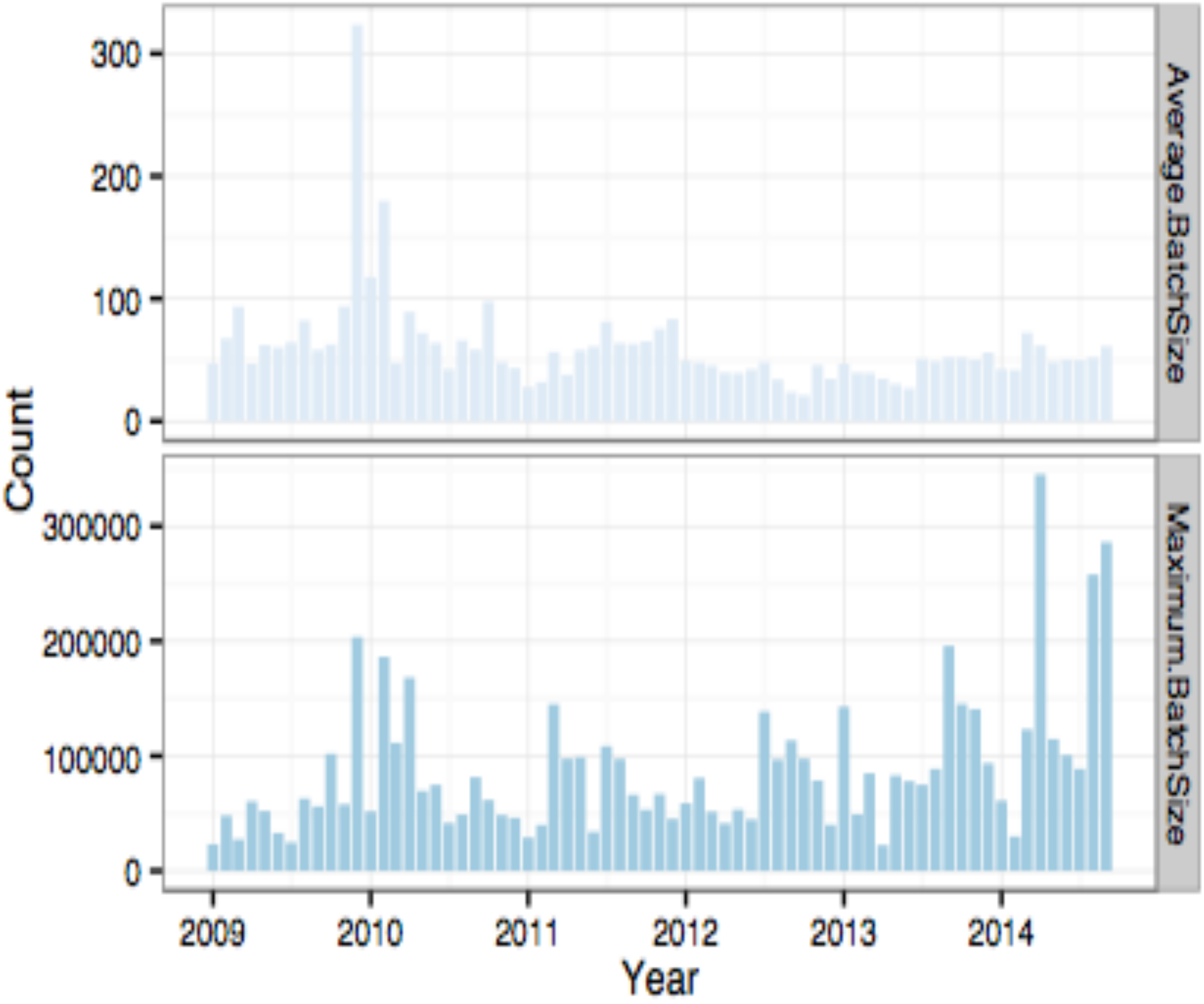
Workers from US, India and Canada are the most sought after.

Distribution of *Batch Size*



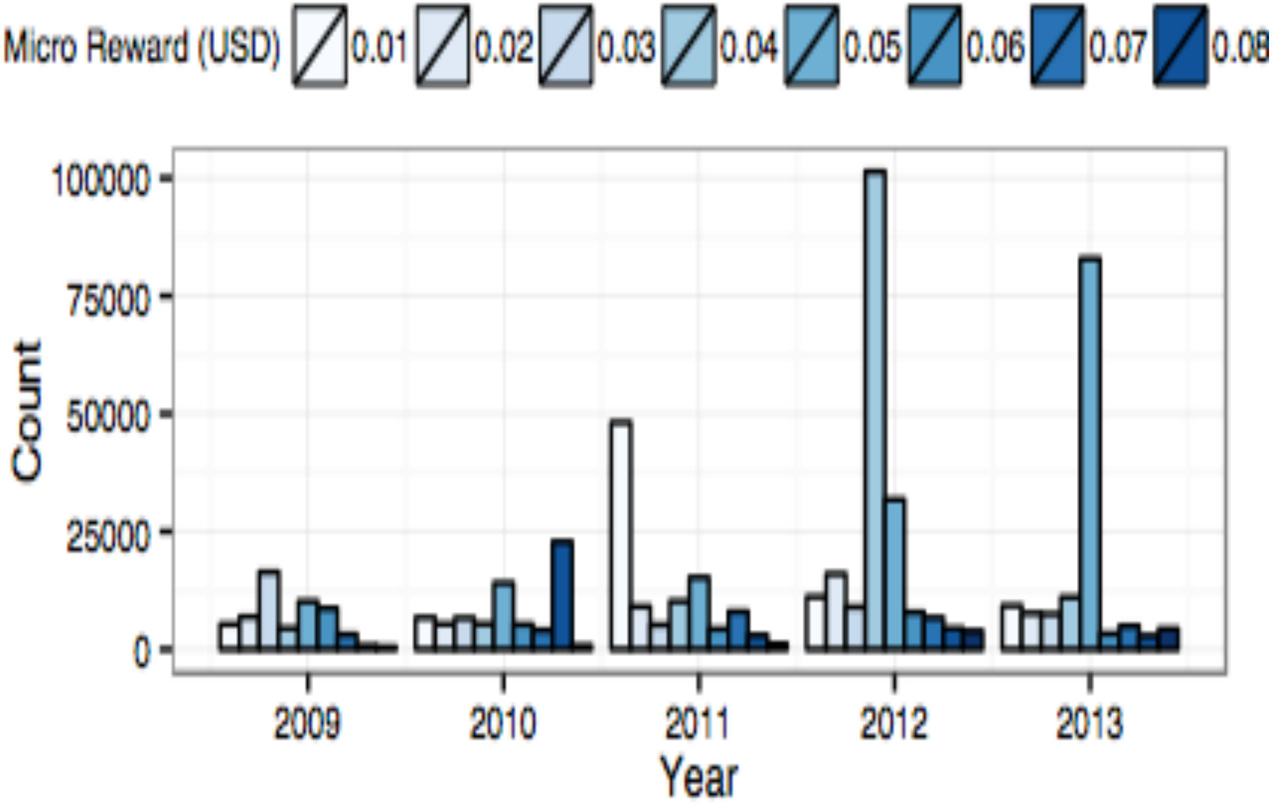
"Power-law"

Batch Size over time



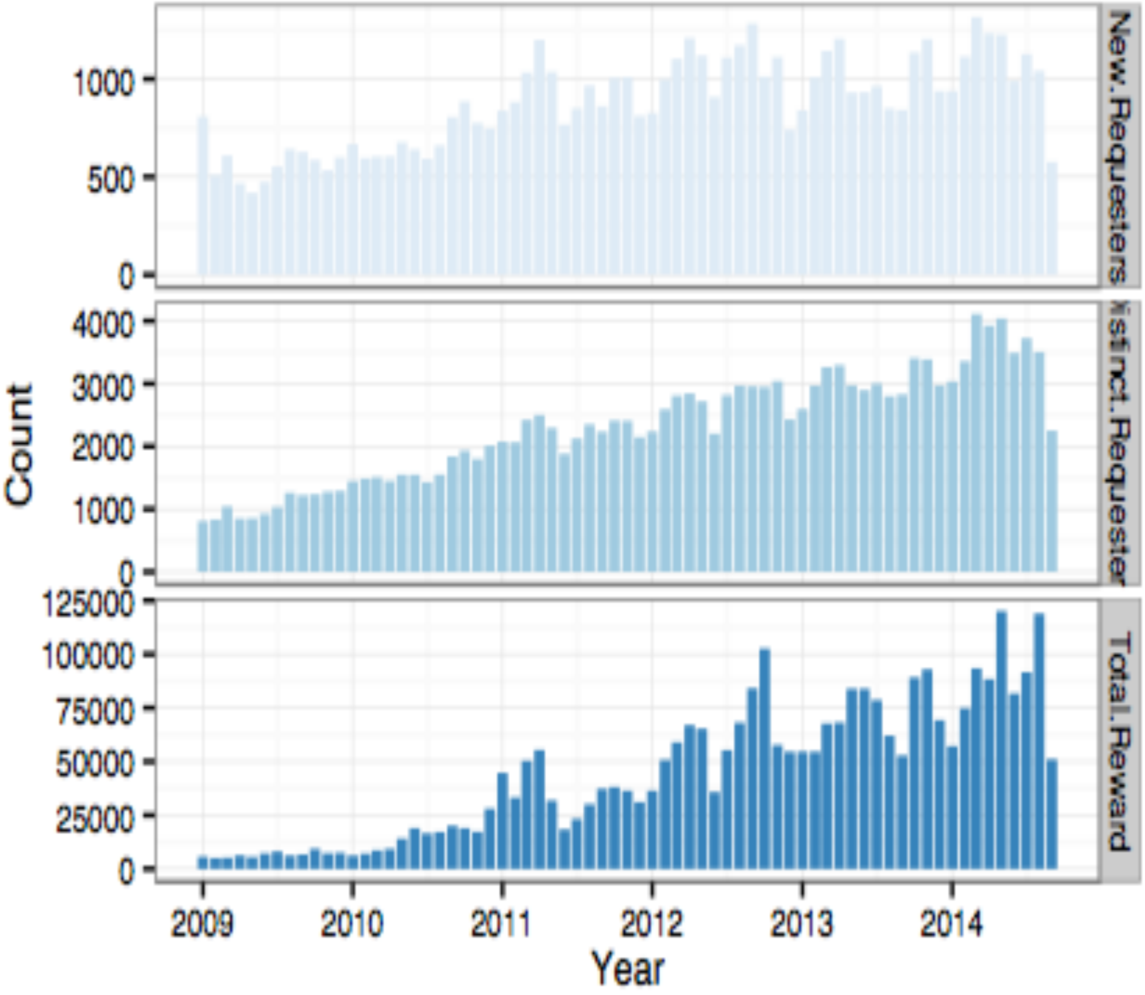
Very large
batches
start to appear

How much are HITs paid?



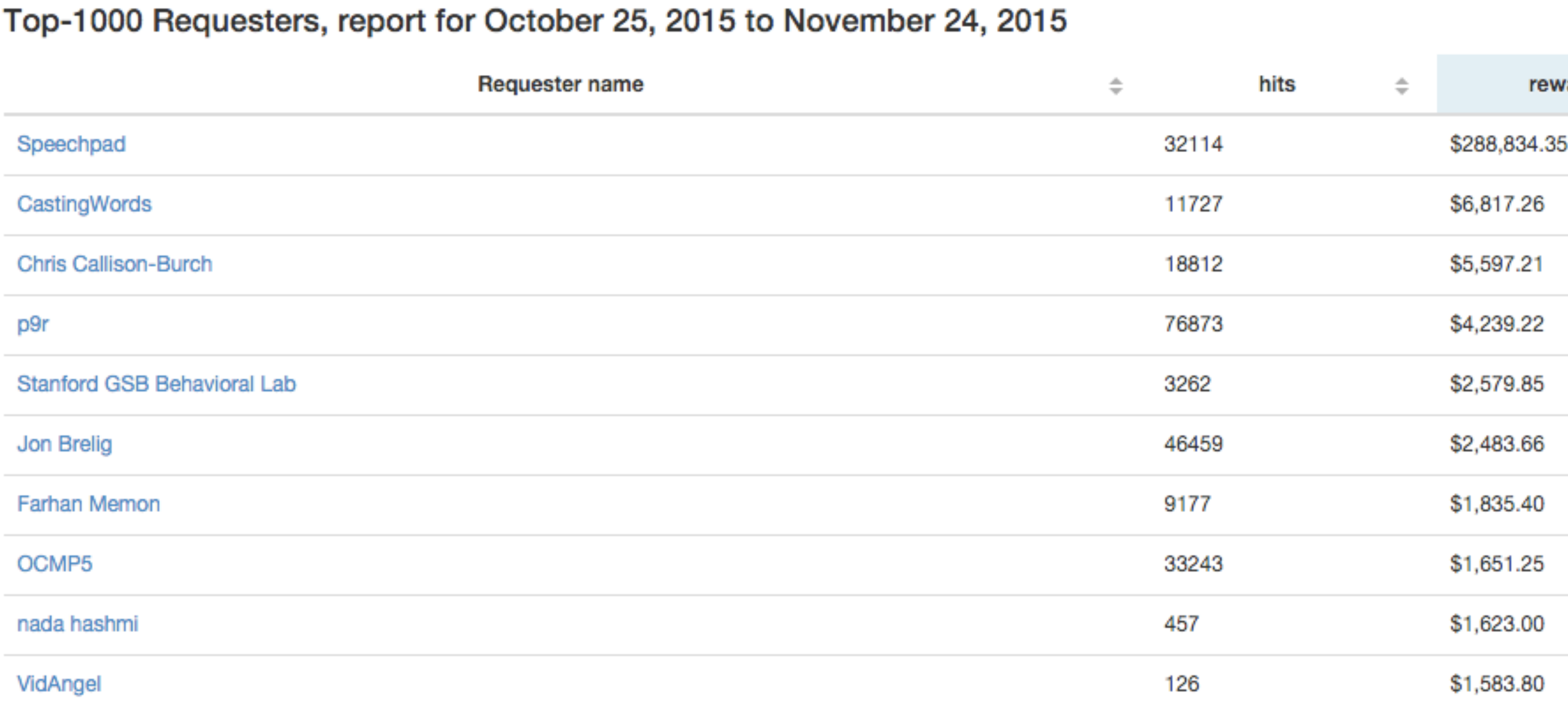
5-cents is the new 1-cent

Requesters and Reward over time



Increasing number of New and Distinct Requesters

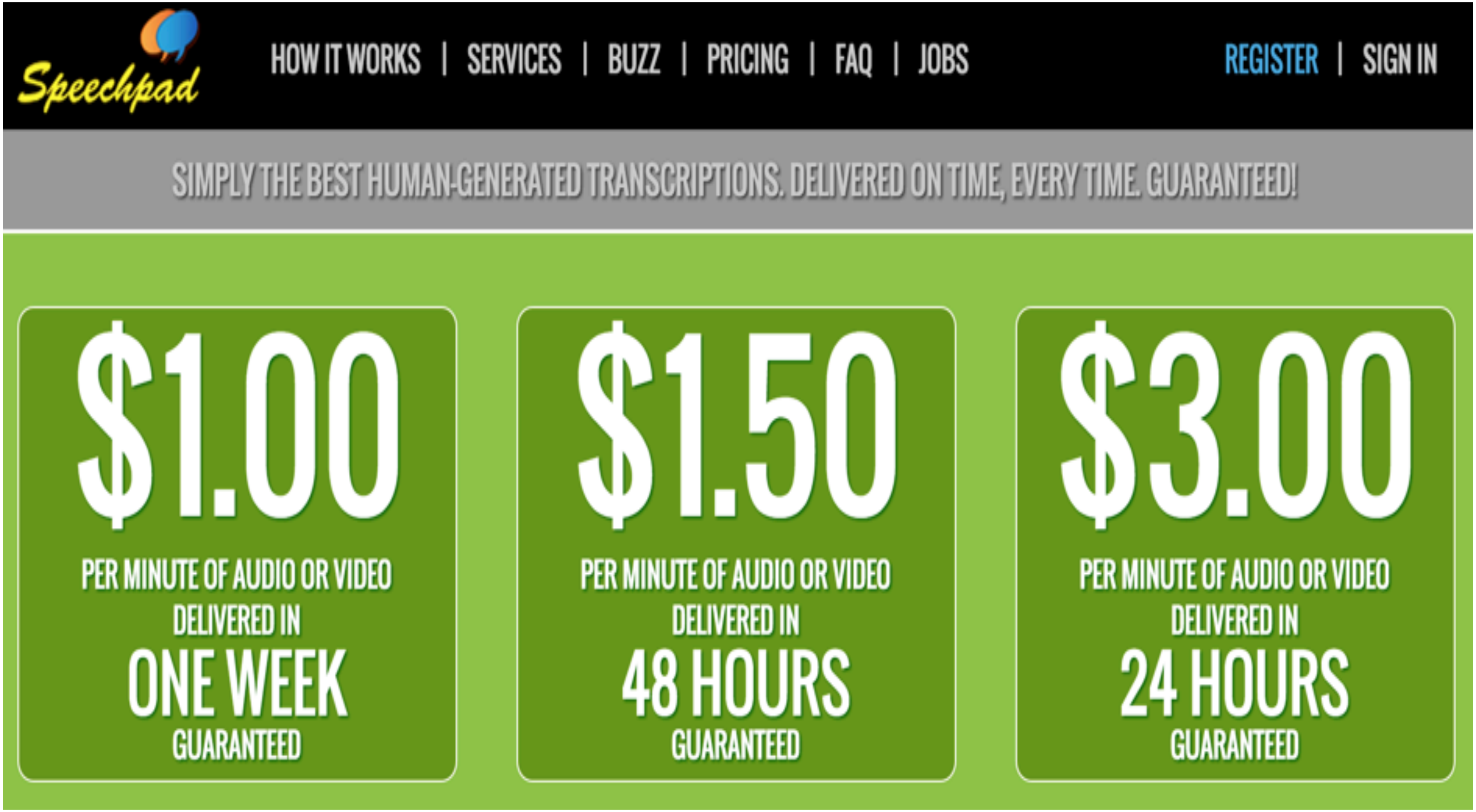
One month of MTurk Requesters



Top-1000 Requesters, report for October 25, 2015 to November 24, 2015

Requester name	hits	rew
Speechpad	32114	$288,834.35
CastingWords	11727	$6,817.26
Chris Callison-Burch	18812	$5,597.21
p9r	76873	$4,239.22
Stanford GSB Behavioral Lab	3262	$2,579.85
Jon Brelig	46459	$2,483.66
Farhan Memon	9177	$1,835.40
OCMP5	33243	$1,651.25
nada hashmi	457	$1,623.00
VidAngel	126	$1,583.80

Top requesters



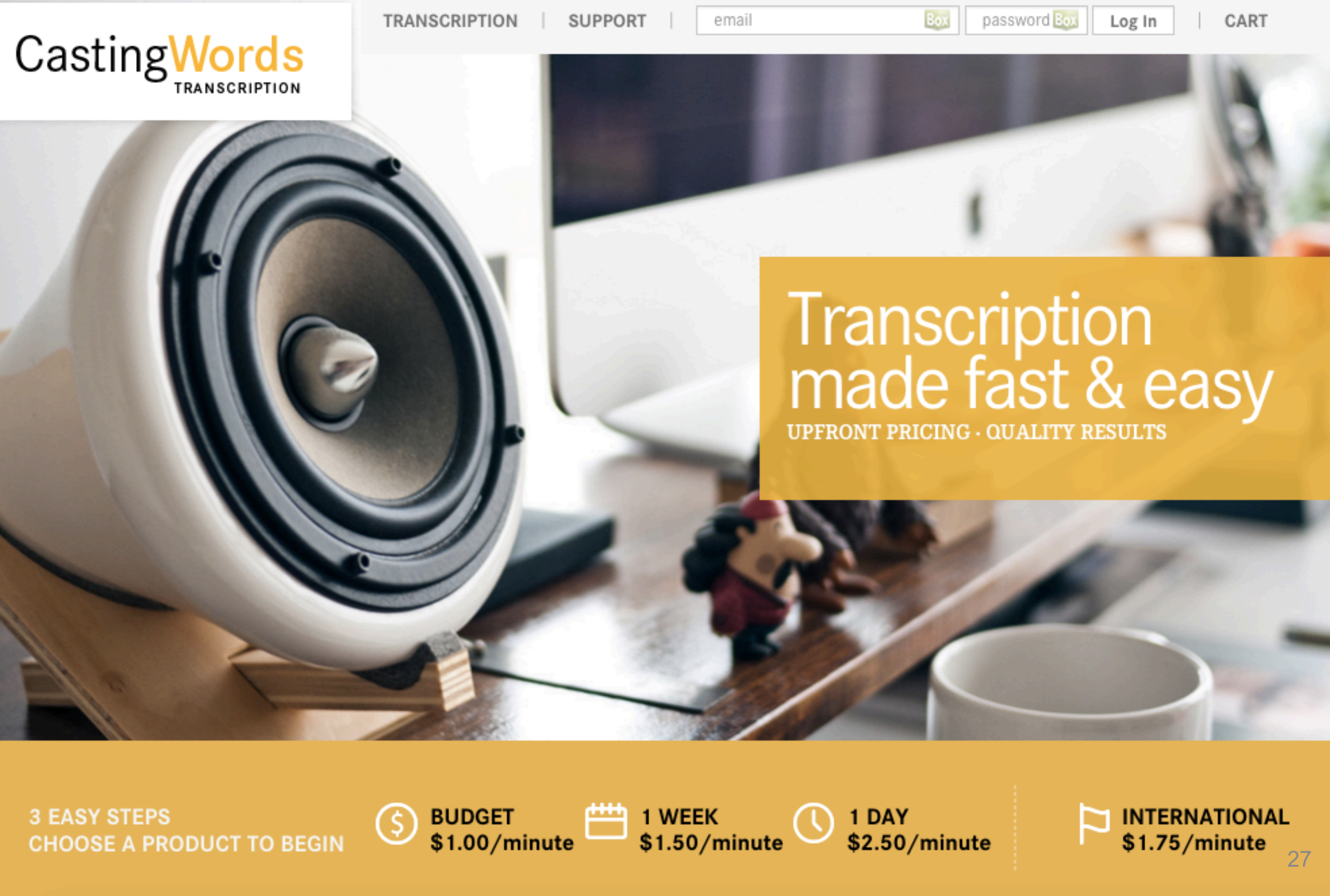
Speechpad

SIMPLY THE BEST HUMAN-GENERATED TRANSCRIPTIONS. DELIVERED ON TIME, EVERY TIME. GUARANTEED!

$1.00 PER MINUTE OF AUDIO OR VIDEO DELIVERED IN ONE WEEK GUARANTEED

$1.50 PER MINUTE OF AUDIO OR VIDEO DELIVERED IN 48 HOURS GUARANTEED

$3.00 PER MINUTE OF AUDIO OR VIDEO DELIVERED IN 24 HOURS GUARANTEED



Transcription made fast & easy

UPFRONT PRICING · QUALITY RESULTS

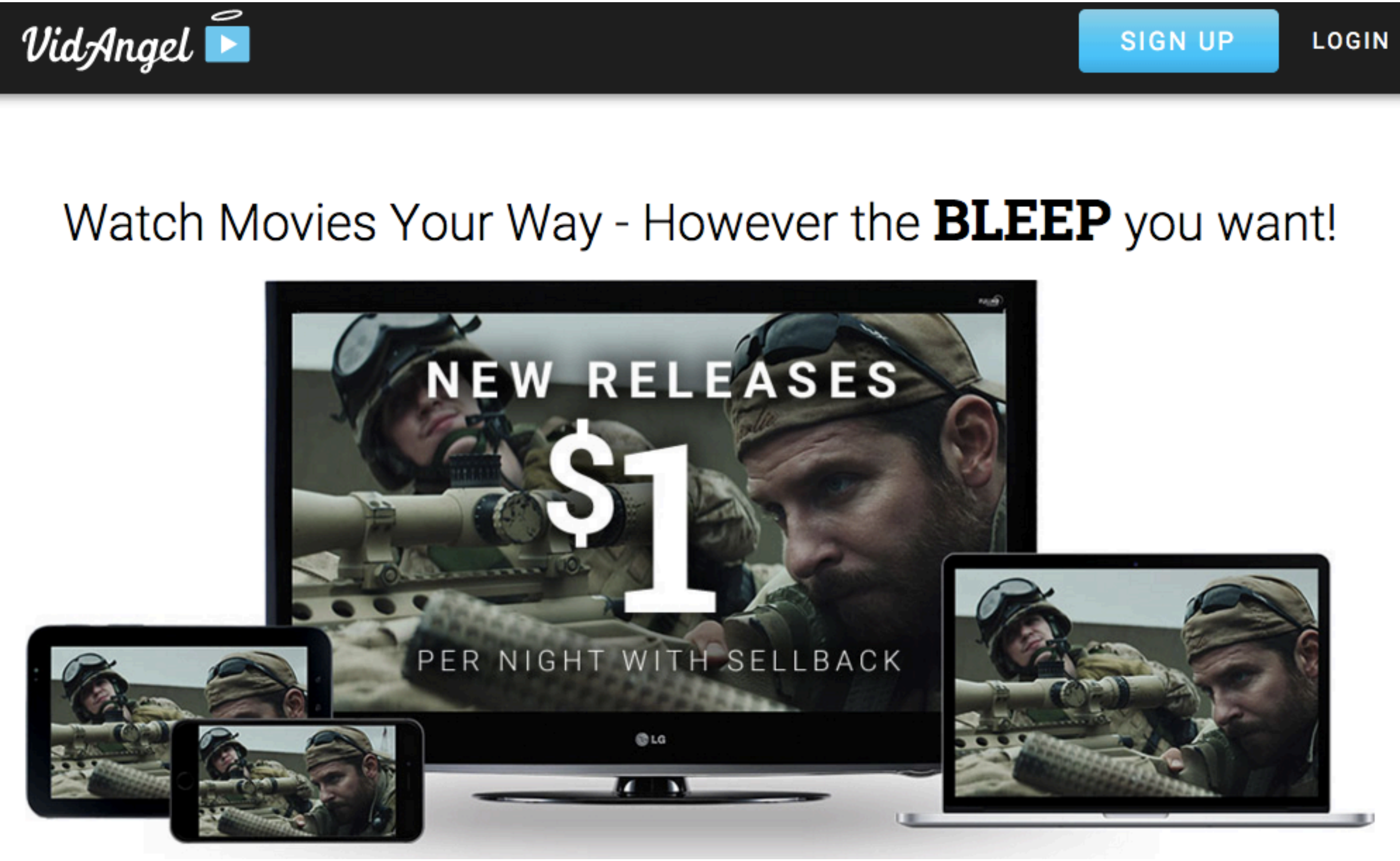
3 EASY STEPS
CHOOSE A PRODUCT TO BEGIN

BUDGET
$1.00/minute

1 WEEK
$1.50/minute

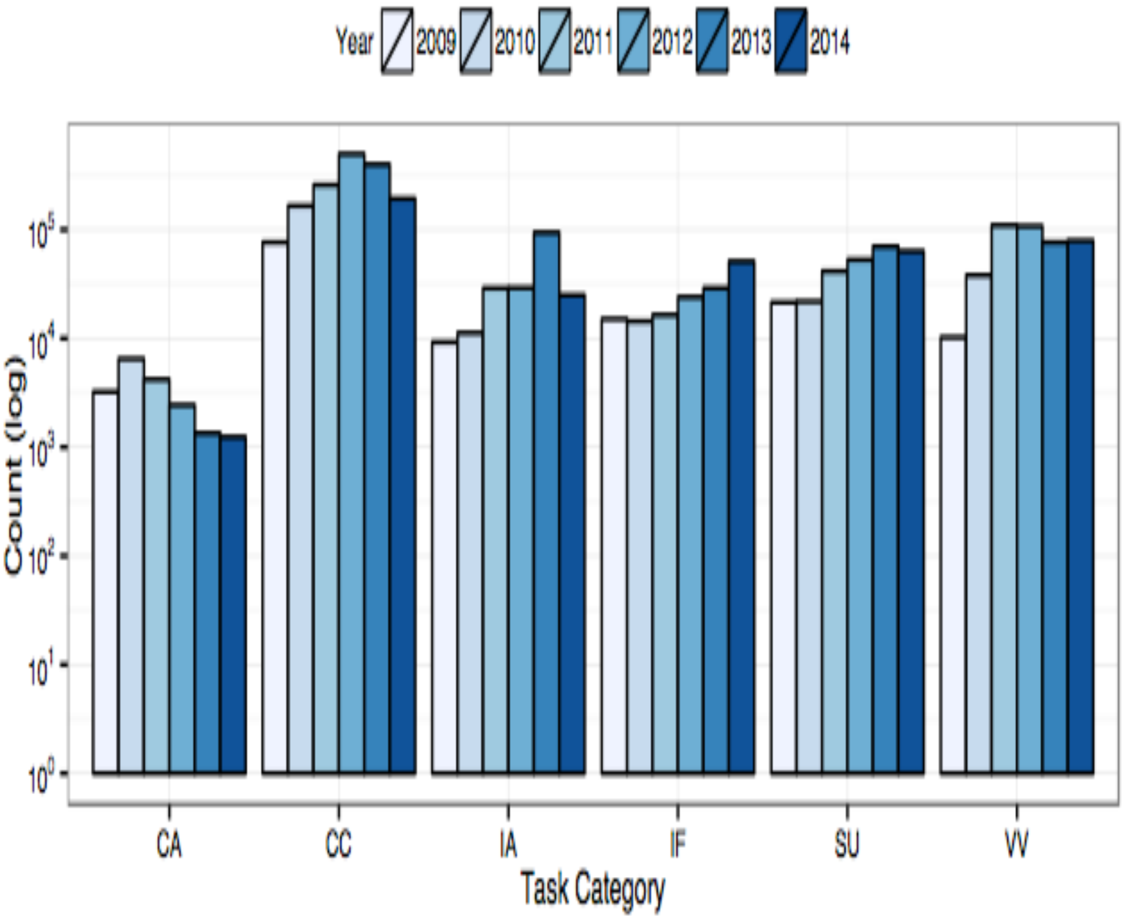
1 DAY
$2.50/minute

INTERNATIONAL
$1.75/minute



Watch Movies Your Way - However the **BLEEP** you want!

Distribution of HIT Types



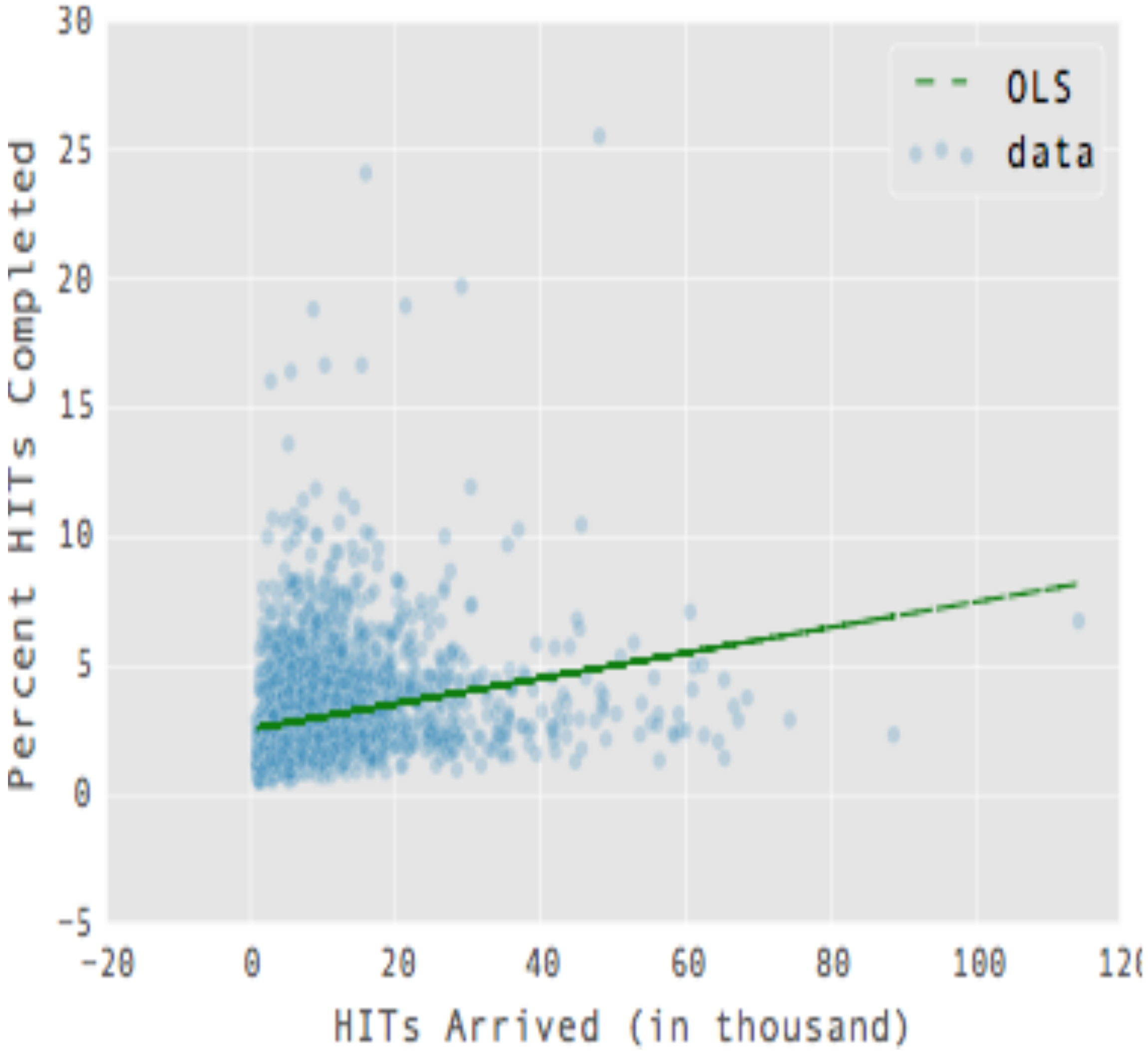
Less *Content Access* batches

Content Creation: the most popular

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)

Is the Market Elastic?



Intercept = 2.5
Slope = 0.5%

20% of new work gets completed within an hour

Summary

- HIT reward has increased over time
- **Audio transcription**: 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 more findings

Why Crowdsourcing for IR Evaluation?

- 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

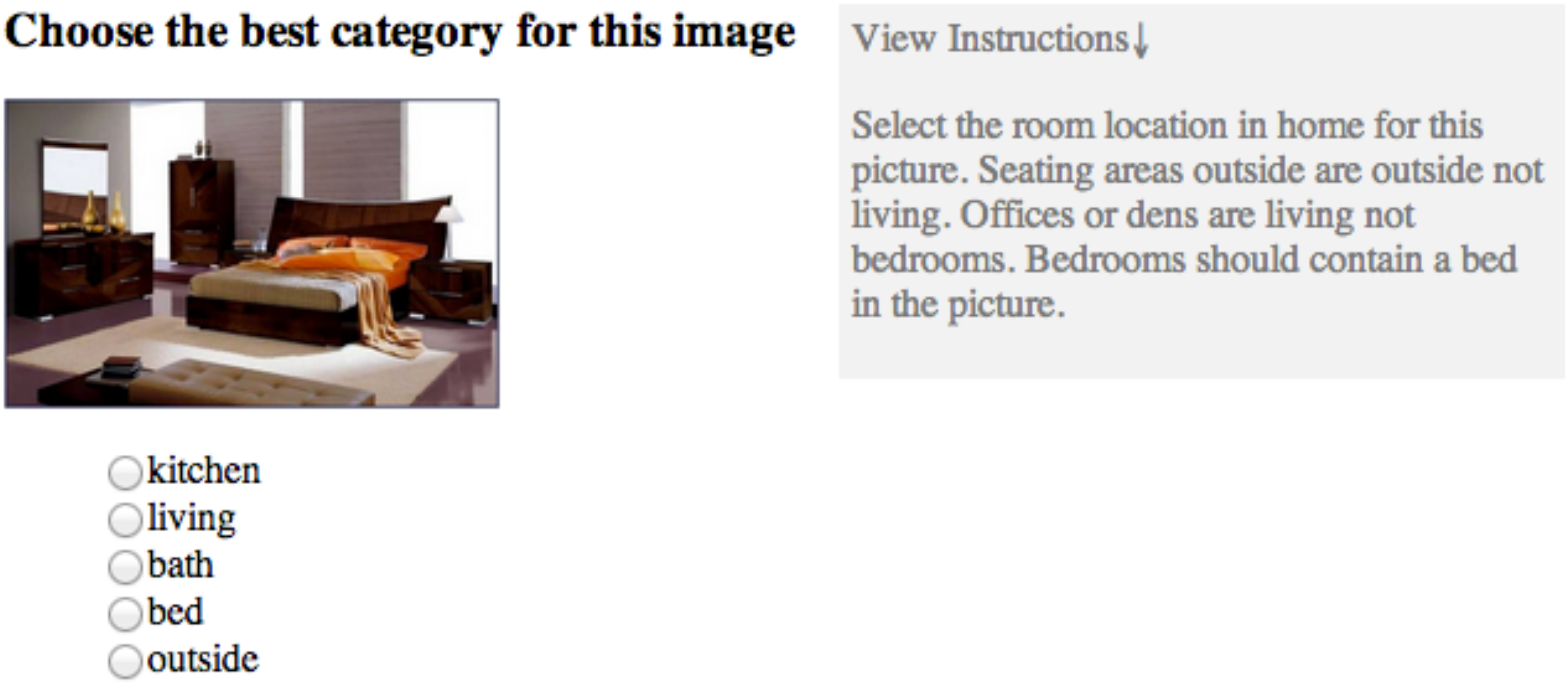
Quality through Flow and Immersion: **Gamifying Crowdsourced Relevance Assessments**. Eickhoff, C., C. G. Harris, A. P. de Vries, and P. Srinivasan. SIGIR 2012.

Tutorial Outline

- Part 1
  - *Introduction to Crowdsourcing (30min)*
  - ***Ensuring Quality in Paid Crowdsourcing*** *(60min)*
- Part 2
  - *Hybrid Human-Machine Data Integration (30min)*
  - *Crowd-Powered Search (30min)*
  - *Enterprise Crowdsourcing for Search (30min)*

Ensuring Quality in Paid Crowdsourcing

A Crowdsourcing Task



Choose the best category for this image

○ kitchen
○ living
○ bath
○ bed
○ outside

View Instructions↓

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

High-level Issues in Crowdsourcing

- Process
  - Experimental design, annotation guidelines, iteration
- Choose crowdsourcing platform (or roll your own!)
- Human factors
  - Payment / incentives, interface and interaction design, communication, reputation, recruitment, retention
- Quality Control / Data Quality
  - Trust, reliability, spam detection, consensus labeling

Task Design

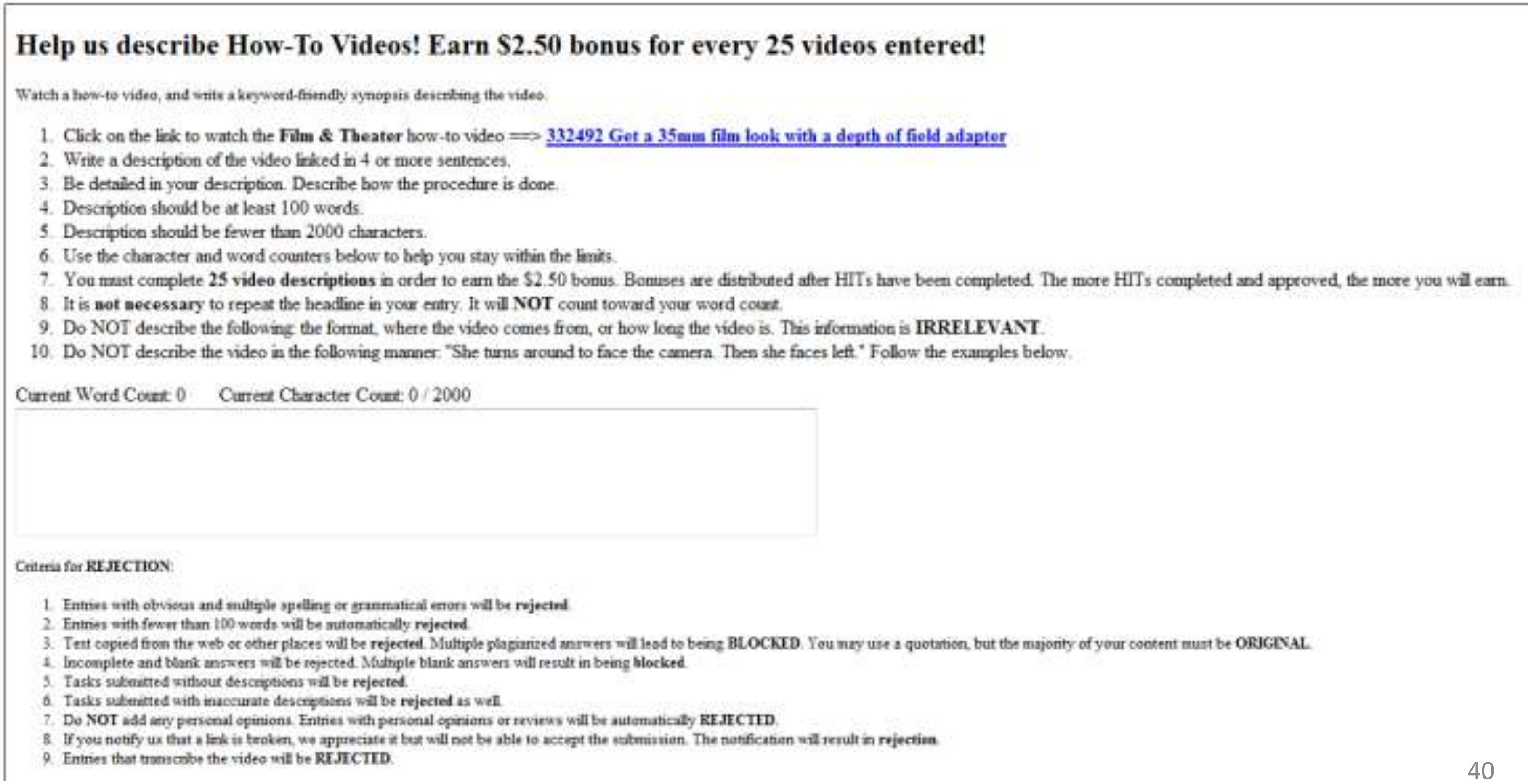
- Ask the right questions
- Workers may not be experts so don't assume the same understanding in terms of terminology
- Instructions matter!
- Show examples
- Hire a technical writer
  - Engineer writes the specification
  - Writer communicates

Task Design - UI

- Generic tips
  - Experiment should be self-contained.
  - Keep it short and simple. Brief and concise.
  - Be very clear with the task.
  - Engage with the worker. Avoid boring stuff.
  - Always ask for feedback (open-ended question) in an input box.

Bad Example

- Asking too much, task not clear, “do NOT/reject”
- Worker has to do a lot of stuff



Help us describe How-To Videos! Earn $2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the **Film & Theater** how-to video ==> 332492 Get a 35mm film look with a depth of field adaptor
2. Write a description of the video linked in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete **25 video descriptions** in order to earn the $2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is **not necessary** to repeat the headline in your entry. It will **NOT** count toward your word count.
9. Do NOT describe the following: the format, where the video comes from, or how long the video is. This information is **IRRELEVANT**.
10. Do NOT describe the video in the following manner: "She turns around to face the camera. Then she faces left." Follow the examples below.

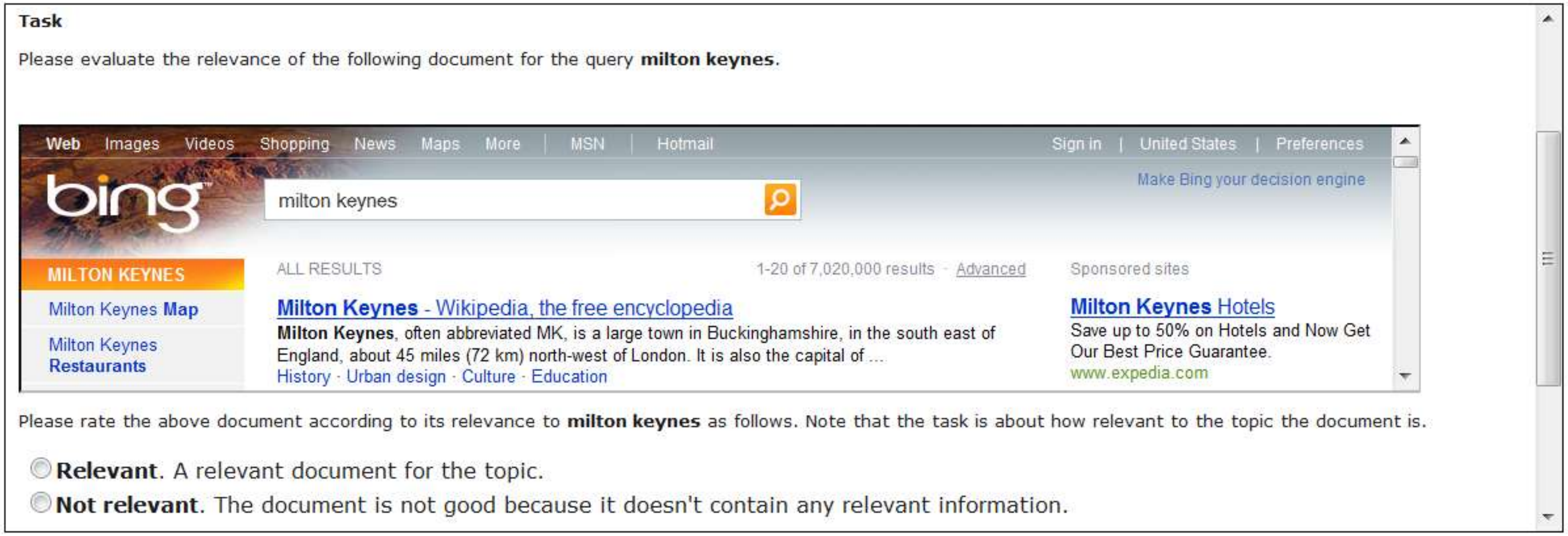
Current Word Count: 0 Current Character Count: 0 / 2000

Criteria for **REJECTION**:

1. Entries with obvious and multiple spelling or grammatical errors will be **rejected**.
2. Entries with fewer than 100 words will be automatically **rejected**.
3. Text copied from the web or other places will be **rejected**. Multiple plagiarized answers will lead to being **BLOCKED**. You may use a quotation, but the majority of your content must be **ORIGINAL**.
4. Incomplete and blank answers will be rejected. Multiple blank answers will result in being **blocked**.
5. Tasks submitted without descriptions will be **rejected**.
6. Tasks submitted with inaccurate descriptions will be **rejected** as well.
7. Do **NOT** add any personal opinions. Entries with personal opinions or reviews will be automatically **REJECTED**.
8. If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in **rejection**.
9. Entries that transcribe the video will be **REJECTED**.

Good Example

- All information is available
  - What to do
  - Search result
  - Question to answer



Task

Please evaluate the relevance of the following document for the query **milton keynes**.

Web Images Videos Shopping News Maps More | MSN | Hotmail
Sign in | United States | Preferences
bing
milton keynes
Make Bing your decision engine
MILTON KEYNES
Milton Keynes **Map**
Milton Keynes **Restaurants**
ALL RESULTS
1-20 of 7,020,000 results · Advanced
Sponsored sites

[**Milton Keynes** - Wikipedia, the free encyclopedia]
Milton Keynes, often abbreviated MK, is a large town in Buckinghamshire, in the south east of England, about 45 miles (72 km) north-west of London. It is also the capital of ...
History · Urban design · Culture · Education

Milton Keynes Hotels
Save up to 50% on Hotels and Now Get Our Best Price Guarantee.
www.expedia.com

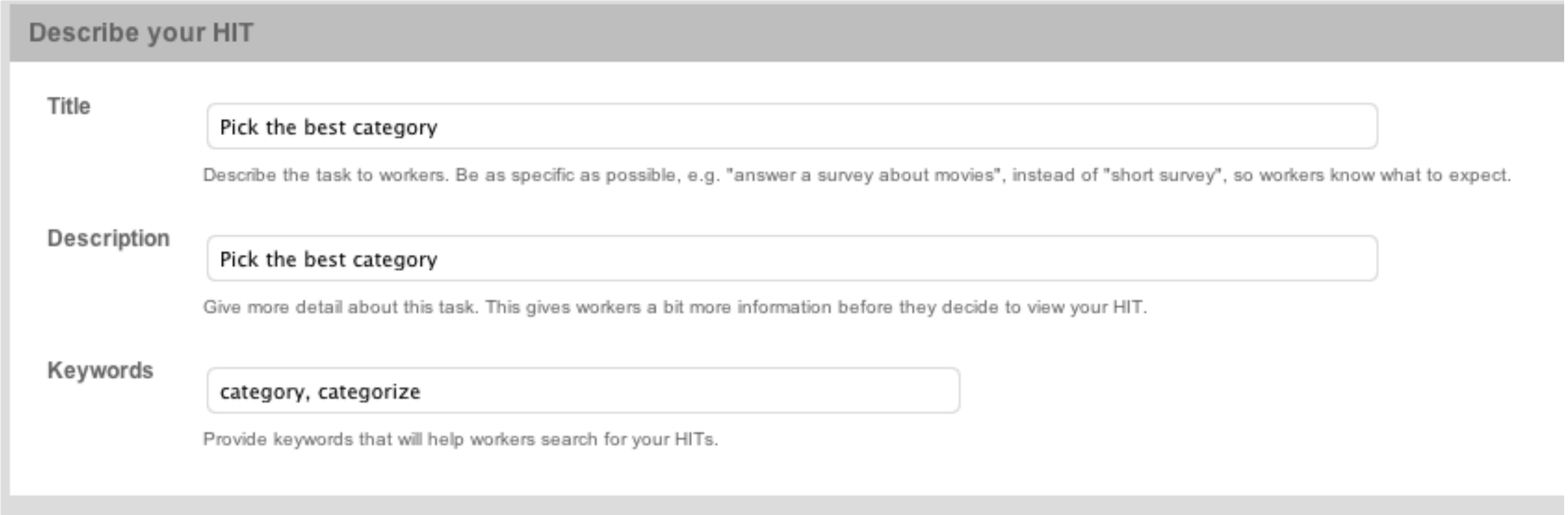
Please rate the above document according to its relevance to **milton keynes** as follows. Note that the task is about how relevant to the topic the document is.

○ **Relevant**. A relevant document for the topic.

○ **Not relevant**. The document is not good because it doesn't contain any relevant information.

Form and Metadata

- Form with a close question (binary relevance) and open-ended question (user feedback)
- Clear title, useful keywords
- Workers need to find your task



Describe your HIT

Title: Pick the best category

Describe the task to workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so workers know what to expect.

Description: Pick the best category

Give more detail about this task. This gives workers a bit more information before they decide to view your HIT.

Keywords: category, categorize

Provide keywords that will help workers search for your HITs.

How Much to Pay?

- Price commensurate with task effort
  - Ex: $0.02 for yes/no answer + $0.02 bonus for optional feedback
- Ethics & market-factors
  - e.g. non-profit SamaSource contracts workers refugee camps
- Uptake & time-to-completion vs. Cost & Quality
  - Too little $$, no interest or slow
  - too much $$, attract spammers
- Accuracy & quantity
  - More pay = more work, not better (W. Mason and D. Watts, 2009)

Quality Control

- Extremely important part of the experiment
- Approach as “overall” quality; not just for workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester.

Quality Control

- Approval rate: easy to use, & just as easily defeated
- Mechanical Turk Masters
  - Recent addition, only for specific tasks
- Qualification test
  - Pre-screen workers' ability to do the task (accurately)
- Assess worker quality as you go
  - Trap questions with known answers ("honey pots")
  - Measure inner-annotator agreement between workers

Qualification tests: pros and cons

- Advantages
  - Great tool for controlling quality
  - Adjust passing grade
- Disadvantages
  - Extra cost to design and implement the test
  - May turn off workers, hurt completion time
  - Refresh the test on a regular basis
  - Hard to verify subjective tasks like judging relevance
- Try creating task-related questions to get worker familiar with task *before* starting task in earnest

Other quality heuristics

- Justification/feedback as quasi-captcha
  - Should be optional
  - Automatically verifying feedback was written by a person may be difficult (classic spam detection task)
- Broken URL/incorrect object
  - Leave an outlier in the data set
  - Workers will tell you
  - If somebody answers “excellent” for a broken URL => *probably* spammer

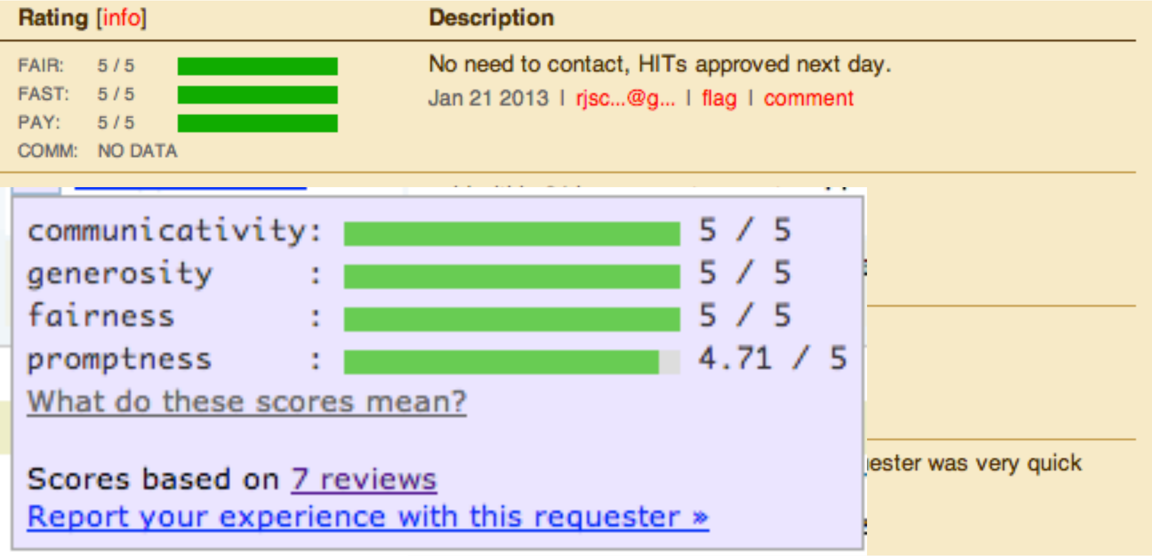
Dealing with bad workers

- Pay for “bad” work instead of rejecting it?
  - Pro: preserve reputation, admit if poor design at fault
  - Con: promote fraud, undermine approval rating system
- Use bonus as incentive
  - Pay the minimum $0.01 and $0.01 for bonus
  - Better than rejecting a $0.02 task
- If spammer “caught”, block from future tasks
  - May be easier to always pay, then block as needed

Build Your Reputation as a Requestor

- Word of mouth effect
  - Workers trust the requester (pay on time, clear explanation if there is a rejection)
  - Experiments tend to go faster
  - Announce forthcoming tasks (e.g. tweet)

Crowd Worker Communities



Rating [info]			Description
FAIR:	5 / 5	🟩	No need to contact, HITs approved next day. Jan 21 2013 | rjsc...@g... | flag | comment
FAST:	5 / 5	🟩	
PAY:	5 / 5	🟩	
COMM:	NO DATA		

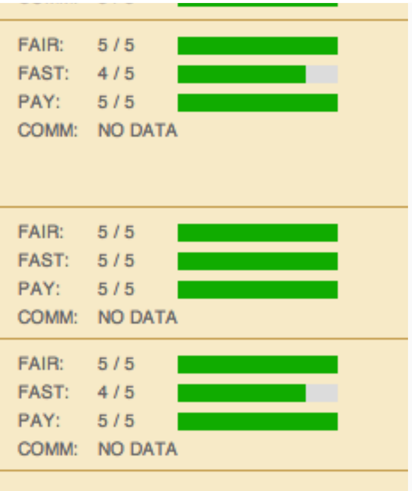
communicativity:	🟩	5 / 5
generosity :	🟩	5 / 5
fairness :	🟩	5 / 5
promptness :	🟩	4.71 / 5

What do these scores mean?

Scores based on 7 reviews

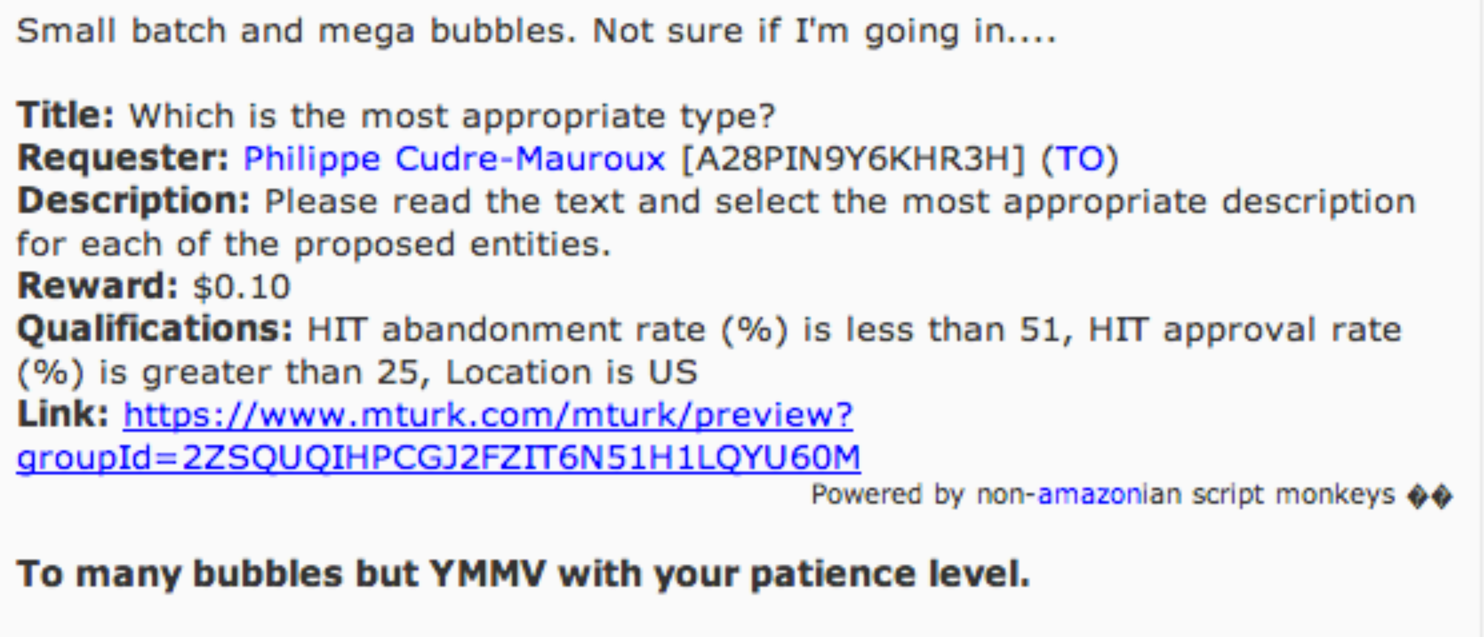
Report your experience with this requester »

ester was very quick



FAIR:	5 / 5	🟩
FAST:	4 / 5	🟩
PAY:	5 / 5	🟩
COMM:	NO DATA	
FAIR:	5 / 5	🟩
FAST:	5 / 5	🟩
PAY:	5 / 5	🟩
COMM:	NO DATA	
FAIR:	5 / 5	🟩
FAST:	4 / 5	🟩
PAY:	5 / 5	🟩
COMM:	NO DATA	

Turkopticon.com
Mturkforum.com
Turkernation.com

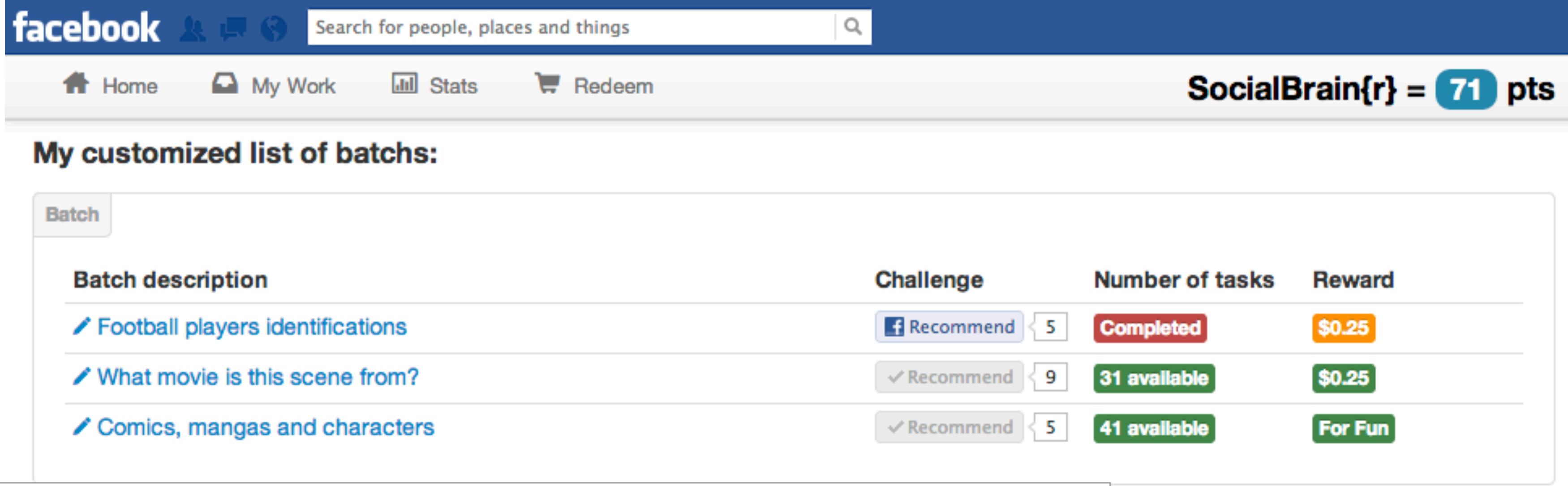


Small batch and mega bubbles. Not sure if I'm going in....

Title: Which is the most appropriate type?
Requester: Philippe Cudre-Mauroux [A28PIN9Y6KHR3H] (TO)
Description: Please read the text and select the most appropriate description for each of the proposed entities.
Reward: $0.10
Qualifications: HIT abandonment rate (%) is less than 51, HIT approval rate (%) is greater than 25, Location is US
Link: https://www.mturk.com/mturk/preview?groupId=2ZSQUQIHPCGJ2FZIT6N51H1LQYU60M

Powered by non-amazonian script monkeys ◆◆

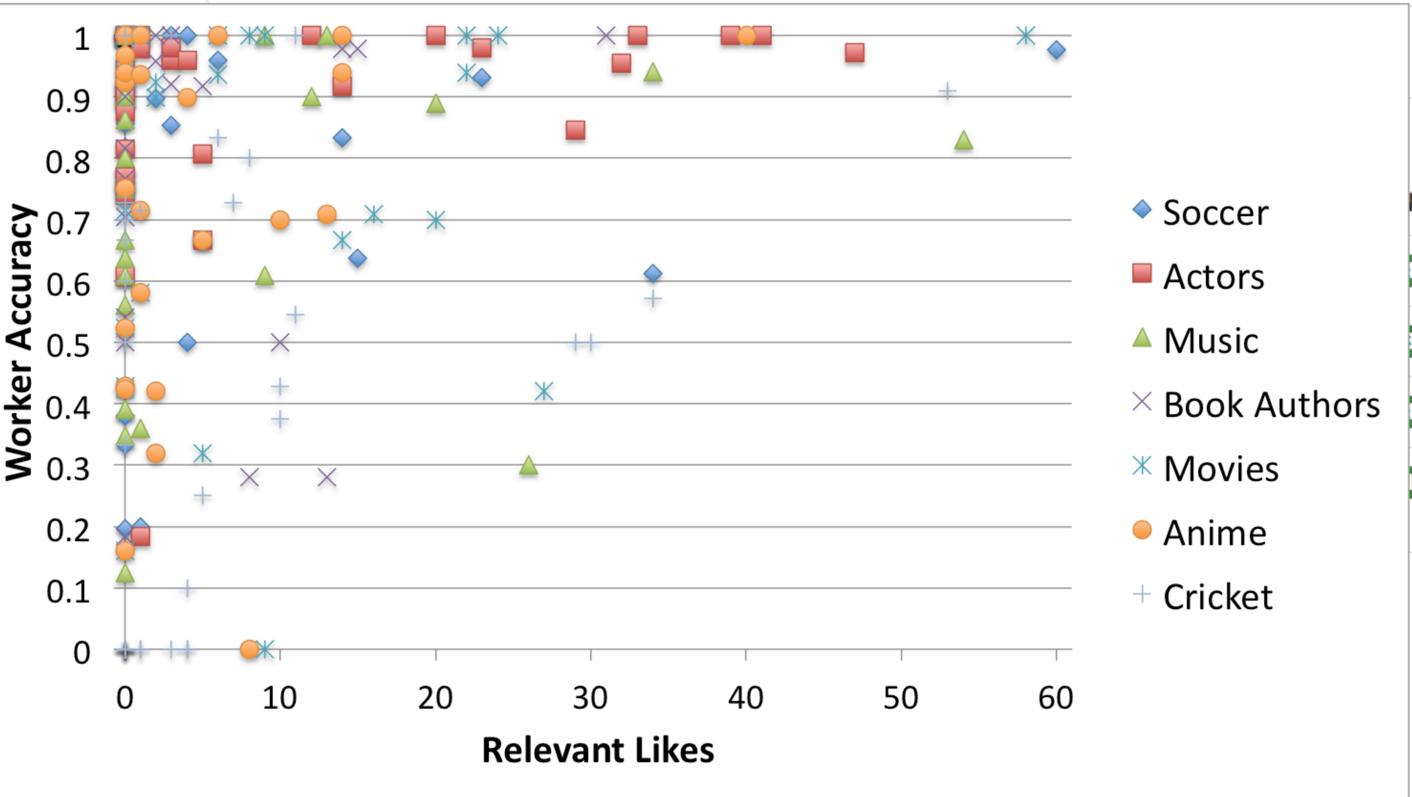
To many bubbles but YMMV with your patience level.

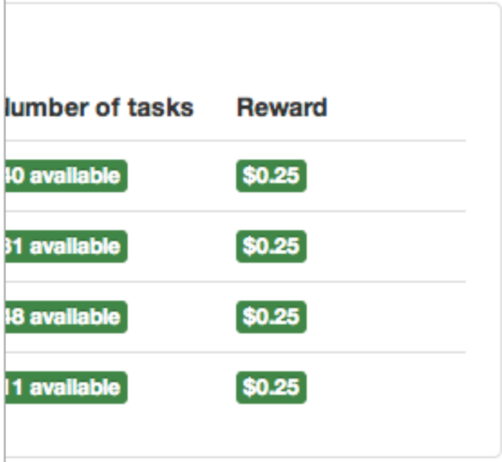


My customized list of batchs:

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





umber 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**).

OpenTurk.com

- Yet another a platform? Build on top of Mturk!
- Chrome Extension for push / notification
- 400+ users
- http://bit.ly/openturk-extension
- Open source: https://github.com/openturk/extension

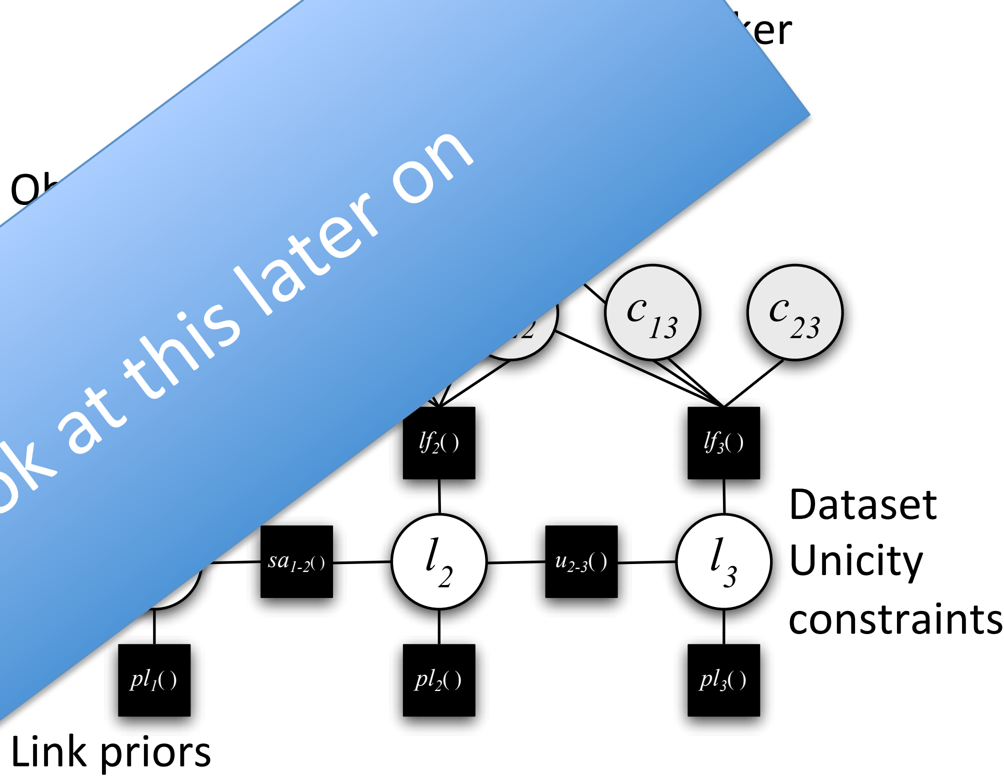
Majority Vote

- Ask N workers and pick the most popular answer
- Works for multiple-choice questions
  - Relevance judgments
  - Sentiment analysis / supervised machine learning
- For other task use **iterations**
  - Audio transcription
  - Ask one worker to transcribe, the next to correct, etc.
- Learning weights for workers

Entity Factor Graphs

- Graph components
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints
- Probabilist
  Inferen
  - Se
    post

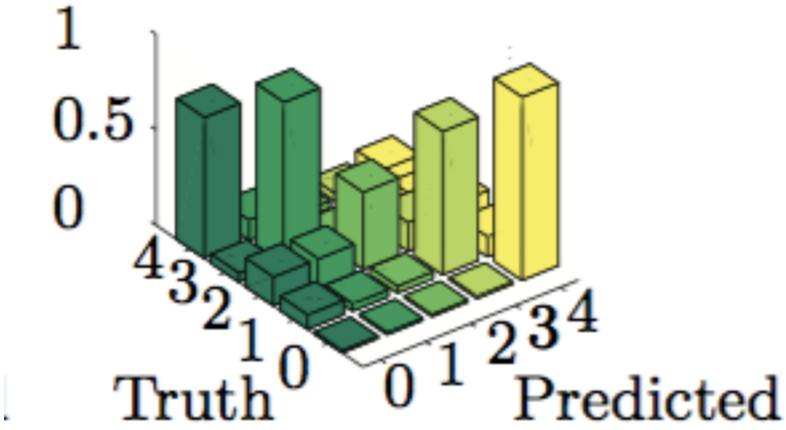
We will look at this later on

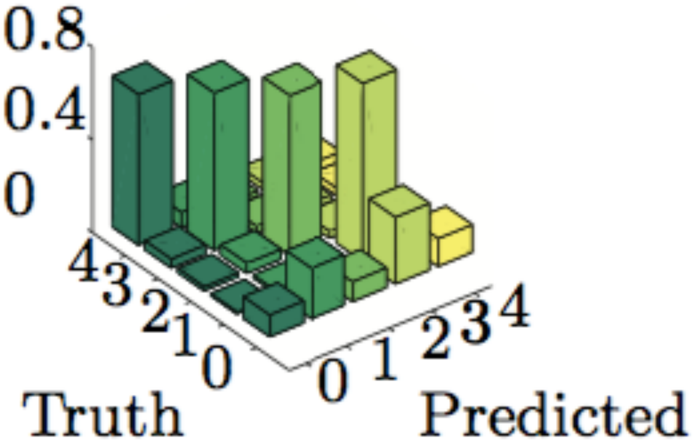


2 workers, 6 clicks, 3 candidate links

Aggregation based on worker similarity

- "Community-Based Bayesian Aggregation Models for Crowdsourcing", Venanzi et al., WWW2014.
- Community-based Bayesian aggregation model
- Group workers by the type of errors they do





SQUARE

- A benchmark for crowd answer aggregation
  - Binary choices (e.g., sentiment)
  - Multiple-choices (e.g., relevance, word-sense disambiguation)
- Compares a number of aggregation techniques over a number of tasks

http://ir.ischool.utexas.edu/square/

Other benchmarks

- Simulations
  - BATC - A Benchmark for Aggregation Techniques in Crowdsourcing
  - Understand effect on efficiency and effectiveness
  - Set aggregation parameters

Tutorial Outline

- Part 1
  - *Introduction to Crowdsourcing (30min)*
  - *Ensuring Quality in Paid Crowdsourcing (60min)*
- Part 2
  - ***Hybrid Human-Machine Data Integration*** *(30min)*
  - *Crowd-Powered Search (30min)*
  - *Enterprise Crowdsourcing for Search (30min)*

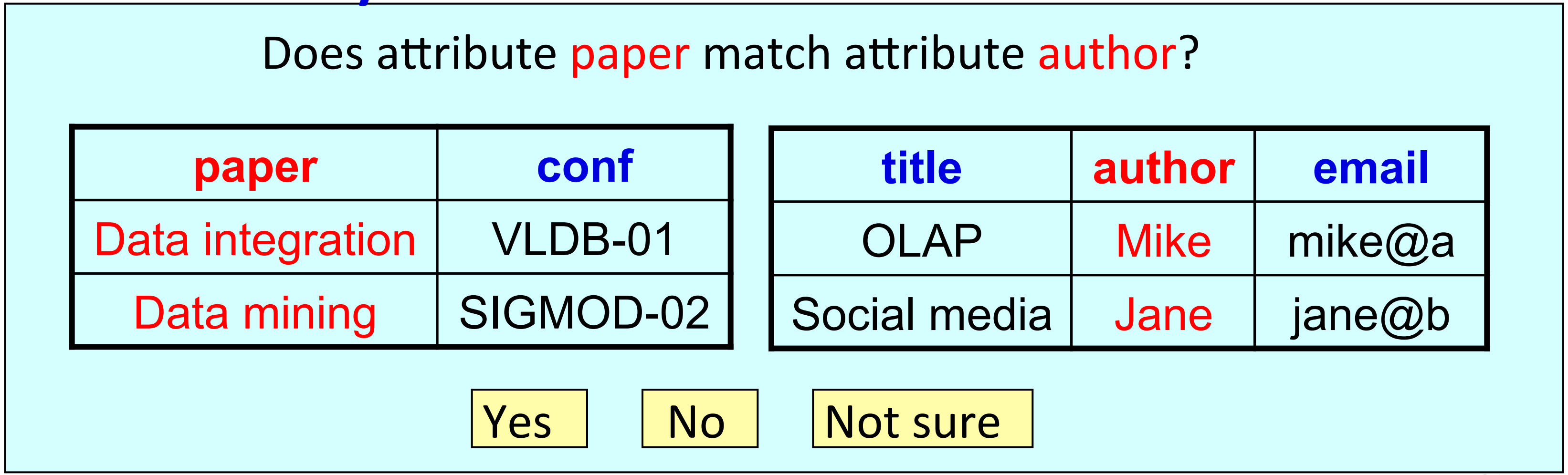
Hybrid Human-Machine Data Integration

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

Example: Hybrid Query Processing

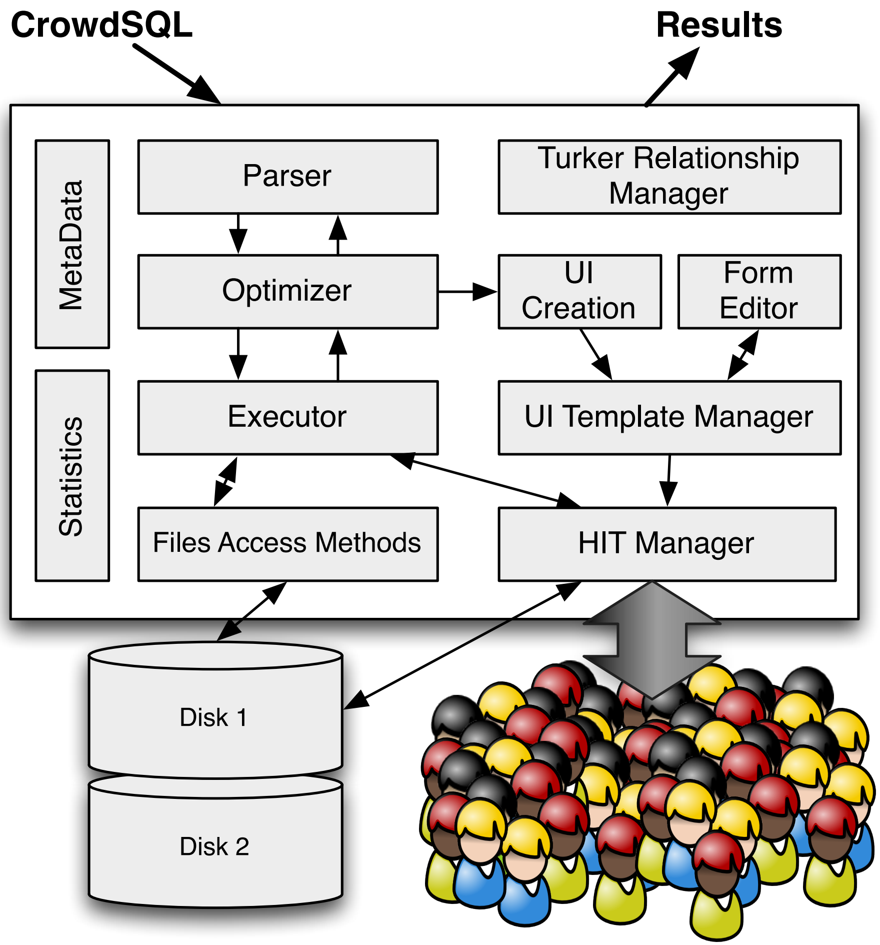
Use the crowd to answer DB-hard queries

Where to use the crowd:

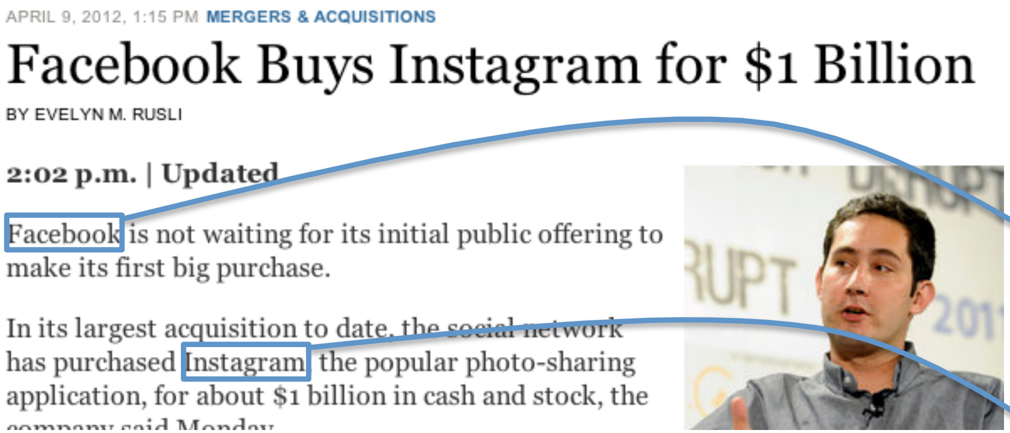
- **Find missing data**
- **Make subjective comparisons**
- **Recognize patterns**

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*



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

owl:sameAs

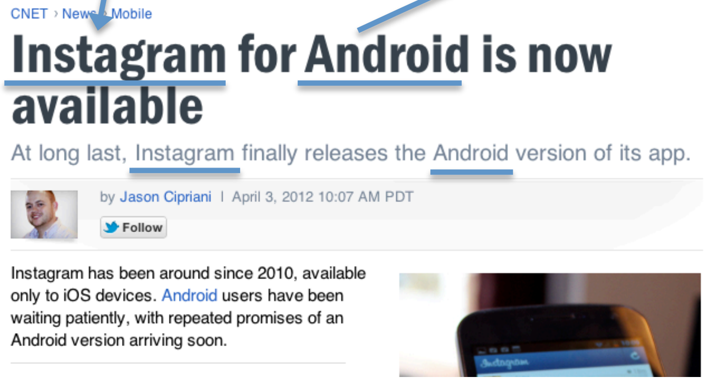
fbase:Instagram

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 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

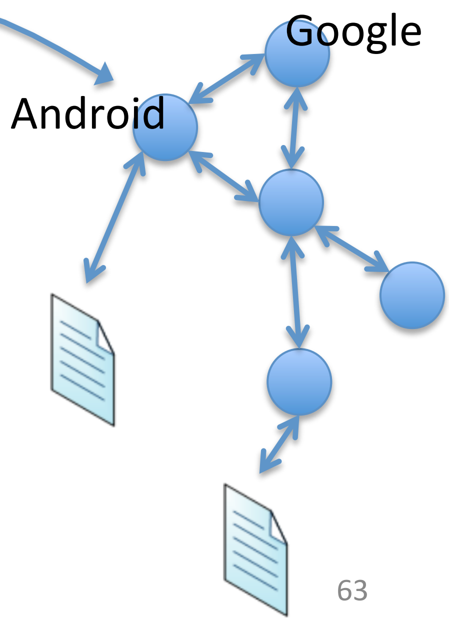
Instagram for Android is now available

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

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

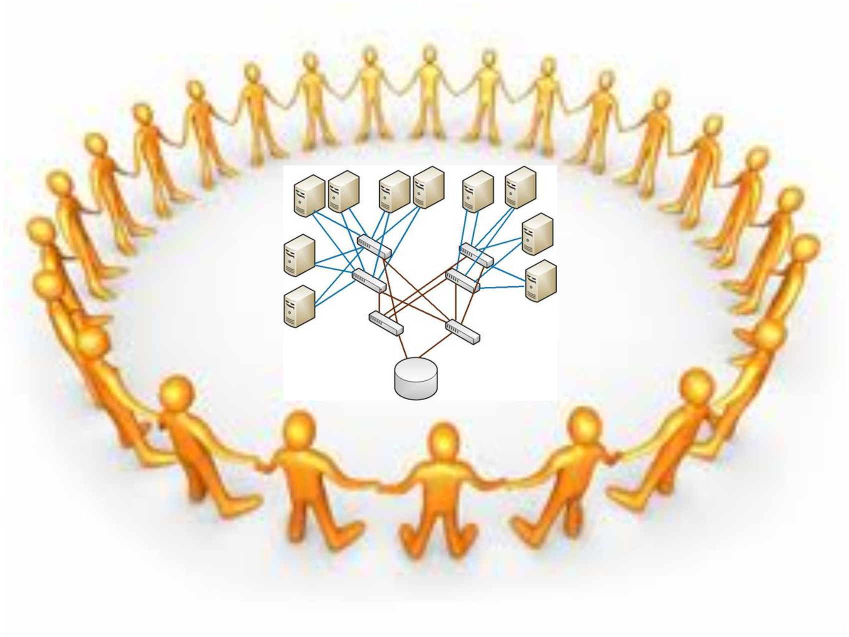
Follow

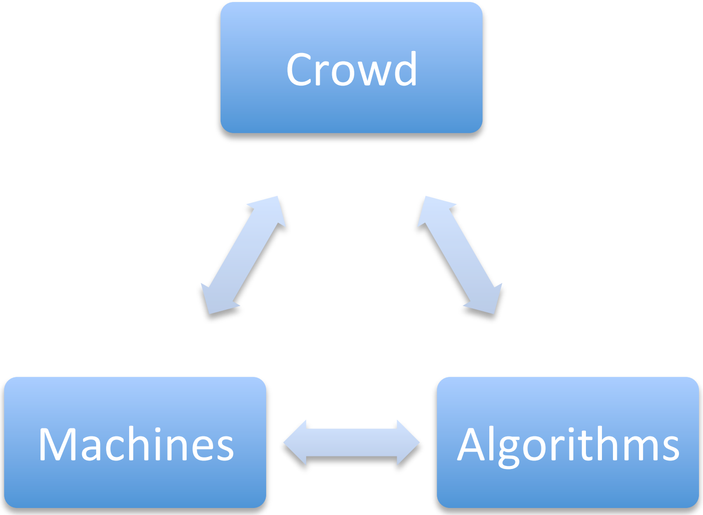
Instagram has been around since 2010, available only to iOS devices. Android users have been waiting patiently, with repeated promises of an Android version arriving soon.



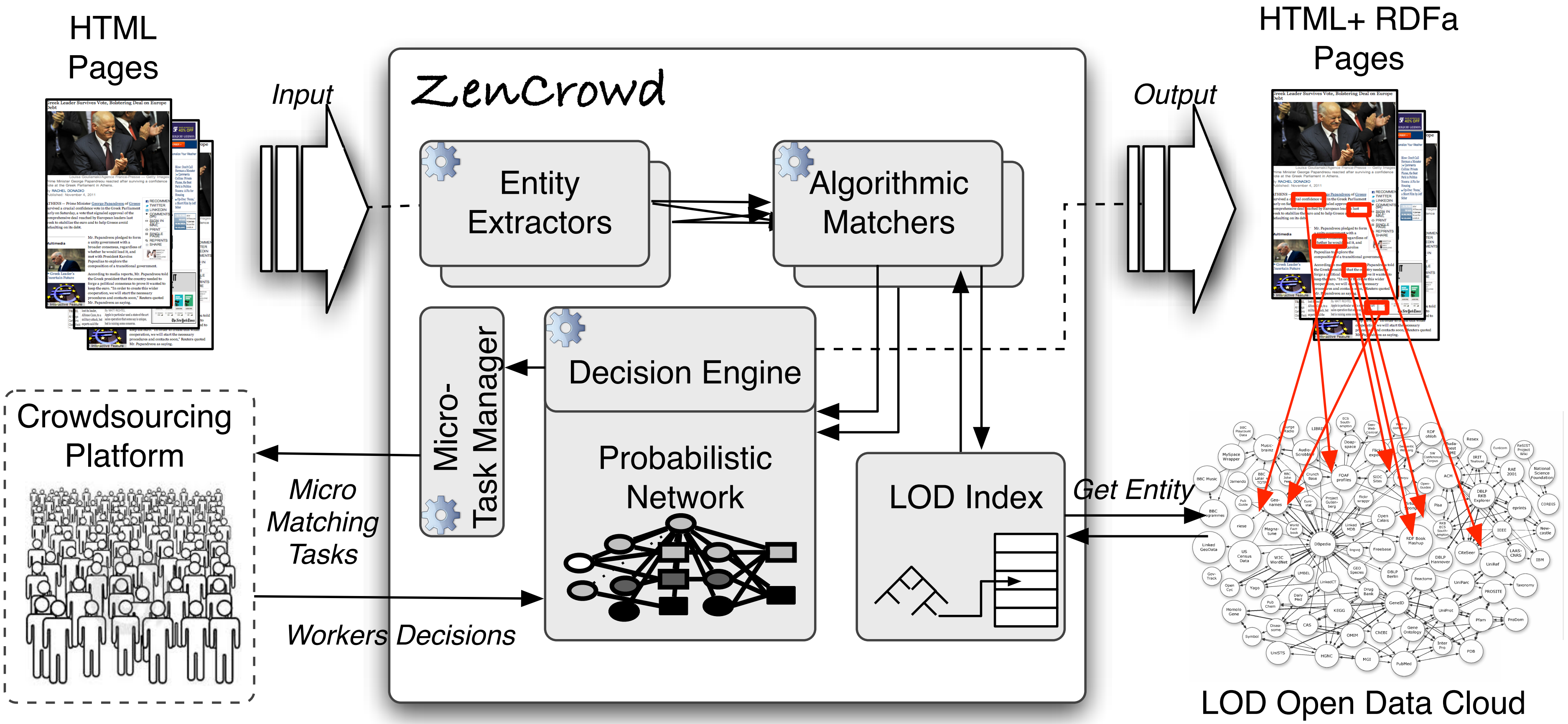
ZenCrowd

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





ZenCrowd Architecture



HTML Pages

Input

ZenCrowd

Entity Extractors

Algorithmic Matchers

Micro-Task Manager

Decision Engine

Probabilistic Network

LOD Index

Output

HTML+ RDFa Pages

Crowdsourcing Platform

Micro Matching Tasks

Workers Decisions

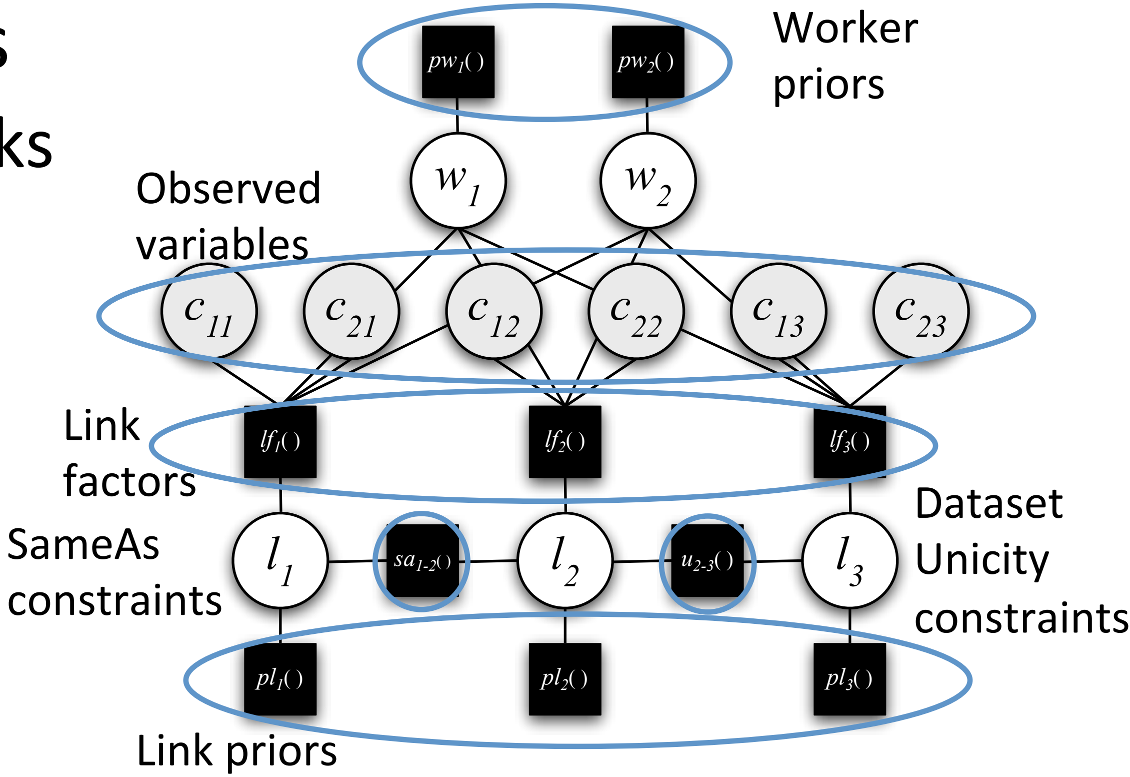
Get Entity

LOD Open Data Cloud

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

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 WWW13 paper)

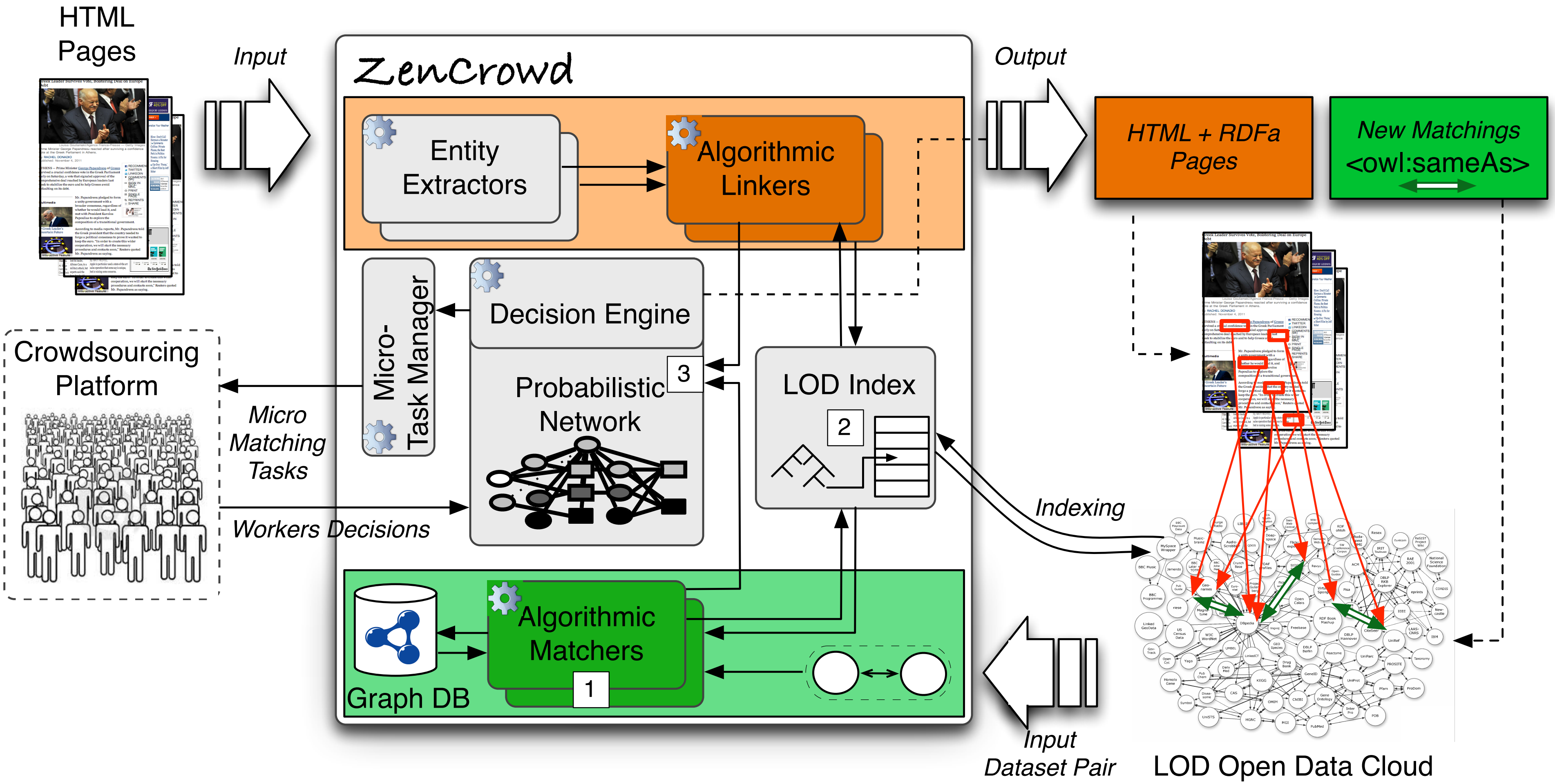
ZenCrowd Summary

- 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

http://exascale.info/zencrowd/

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

ZenCrowd Architecture



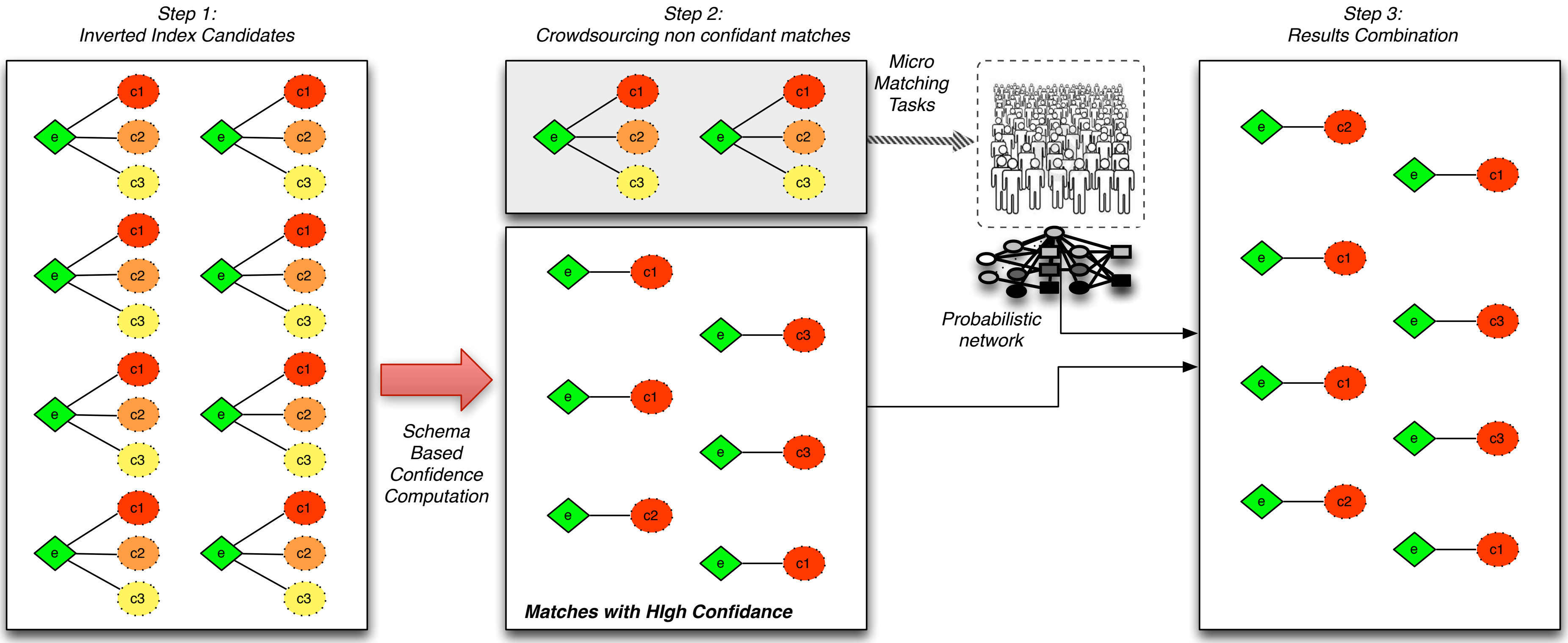
Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. **ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking.** In: 21st International Conference on World Wide Web (WWW 2012)

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

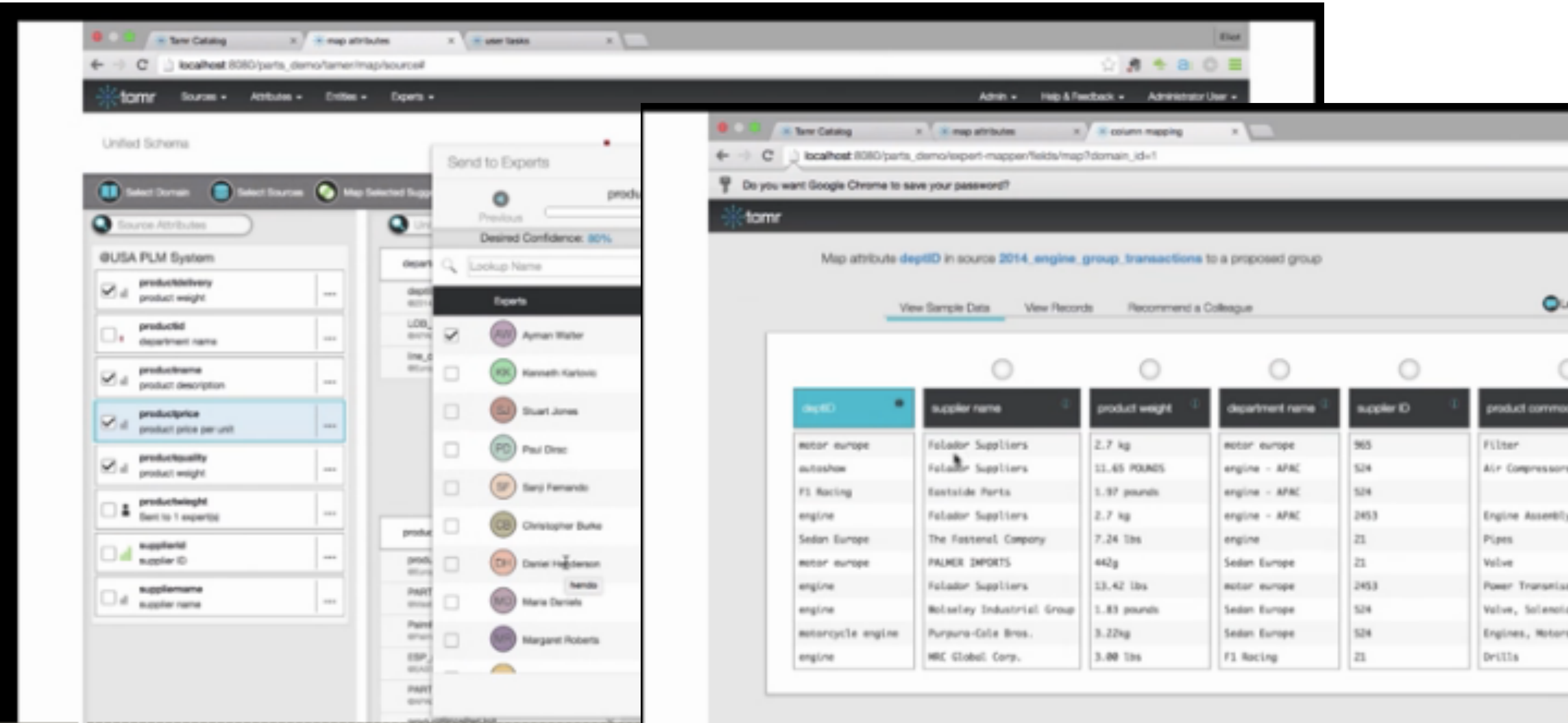
- Crowdsourcing as the most expensive similarity measure



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

tamr.com

- Data Integration solutions: algorithms+experts



@USA PLM System

☑ productdelivery
product weight

□ productid
department name

☑ productname
product description

☑ productprice
product price per unit

☑ productquality
product weight

□ productweight
Sent to 1 expert(s)

□ supplierid
supplier ID

□ suppliername
supplier name

☑ Ayman Walter
□ Kenneth Karlovic
□ Stuart Jones
□ Paul Dirac
□ Sanji Fernando
□ Christopher Burke
□ Daniel Henderson
□ Maria Daniels
□ Margaret Roberts

Map attribute deptID in source 2014_engine_group_transactions to a proposed group

deptID	supplier name	product weight	department name	supplier ID	product commo
motor europe	Folador Suppliers	2.7 kg	motor europe	965	Filter
autoshow	Folador Suppliers	11.65 POUNDS	engine - APAC	524	Air Compressor
F1 Racing	Eastside Parts	1.97 pounds	engine - APAC	524	
engine	Folador Suppliers	2.7 kg	engine - APAC	2453	Engine Assembl
Sedan Europe	The Fastenal Company	7.24 lbs	engine	21	Pipes
motor europe	PALMER IMPORTS	442g	Sedan Europe	21	Valve
engine	Folador Suppliers	13.42 lbs	motor europe	2453	Power Transmis
engine	Wolseley Industrial Group	1.83 pounds	Sedan Europe	524	Valve, Solenoi
motorcycle engine	Purpura-Cole Bros.	3.22kg	Sedan Europe	524	Engines, Motor
engine	MRC Global Corp.	3.00 lbs	F1 Racing	21	Drills

http://www.tamr.com/oreilly-webinar-taming-data-variety/

Tutorial Outline

- Part 1
  - *Introduction to Crowdsourcing (30min)*
  - *Ensuring Quality in Paid Crowdsourcing (60min)*
- Part 2
  - *Hybrid Human-Machine Data Integration (30min)*
  - ***Crowd-Powered Search** (30min)*
  - *Enterprise Crowdsourcing for Search (30min)*

Crowd-Powered Search

Slow Search

- “Not All Searches Need to Be Fast”
  - Planning a vacation
  - Medical diagnosis
- Use additional time for human computation

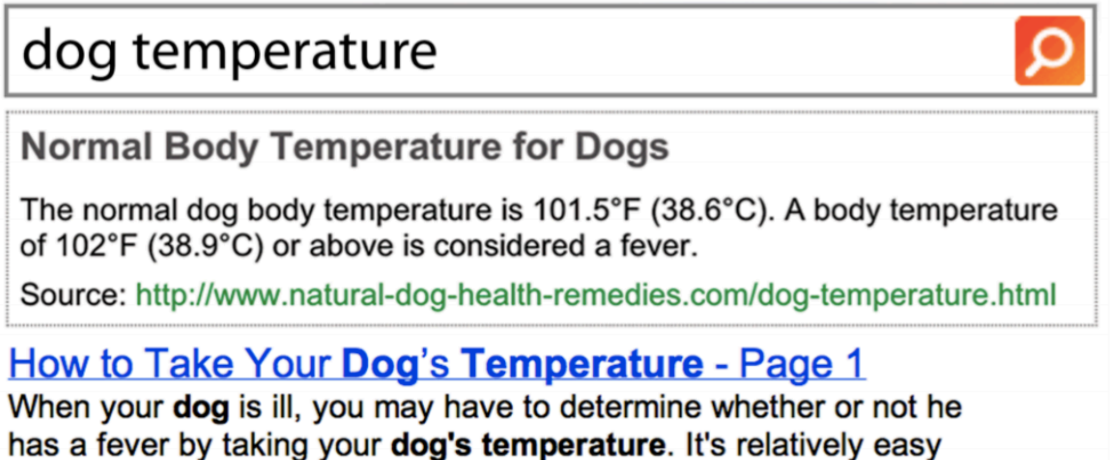
Jaime Teevan. “Slow Search: Improving Information Retrieval Using Human Assistance”, CIKM 2015.

Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, and Susan Dumais. “SlowSearch”. CACM, 57-8, Aug 2014.

Crowd-powered Search

- Search process
  - Understand query
  - Retrieve
  - Understand results
- Machines are good at operating at scale
- People are good at understanding

Extract Direct Answers w/ Crowdsourcing



dog temperature

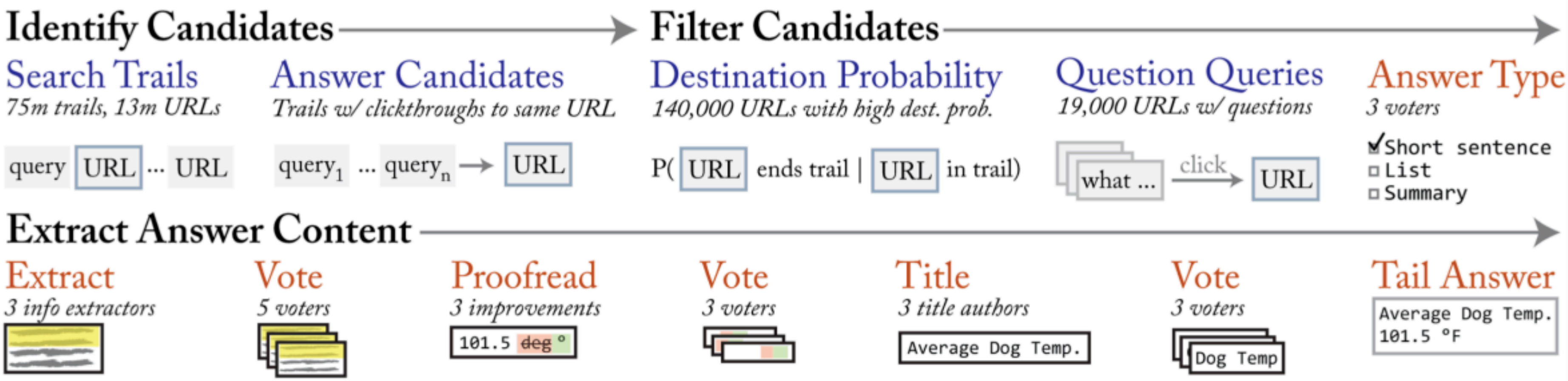
Normal Body Temperature for Dogs

The normal dog body temperature is 101.5°F (38.6°C). A body temperature of 102°F (38.9°C) or above is considered a fever.

Source: http://www.natural-dog-health-remedies.com/dog-temperature.html

How to Take Your **Dog**'s **Temperature** - Page 1

When your **dog** is ill, you may have to determine whether or not he has a fever by taking your **dog's temperature**. It's relatively easy



Identify Candidates

Search Trails
75m trails, 13m URLs
query URL ... URL

Answer Candidates
Trails w/ clickthroughs to same URL
query_1 ... query_n → URL

Filter Candidates

Destination Probability
140,000 URLs with high dest. prob.
P(URL ends trail | URL in trail)

Question Queries
19,000 URLs w/ questions
what ... click → URL

Answer Type
3 voters
✓Short sentence
□List
□Summary

Extract Answer Content

Extract
3 info extractors

Vote
5 voters

Proofread
3 improvements
101.5 ~~deg~~ °

Vote
3 voters

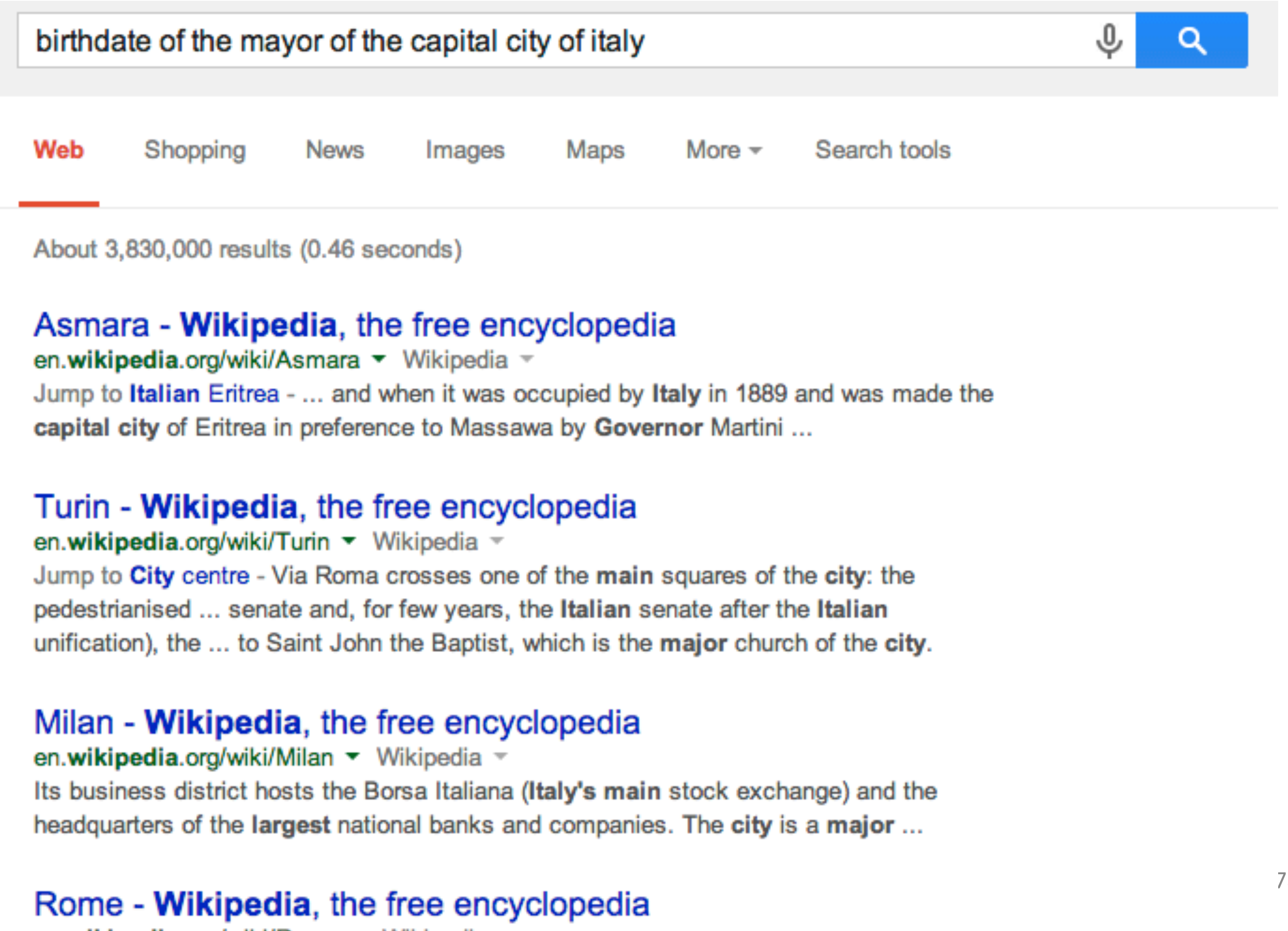
Title
3 title authors
Average Dog Temp.

Vote
3 voters
Dog Temp

Tail Answer
Average Dog Temp.
101.5 °F

Bernstein et al., Direct Answers for Search Queries in the Long Tail, CHI 2012.

birthdate of the mayor of the capital city of italy



birthdate of the mayor of the capital city of italy

About 3,830,000 results (0.46 seconds)

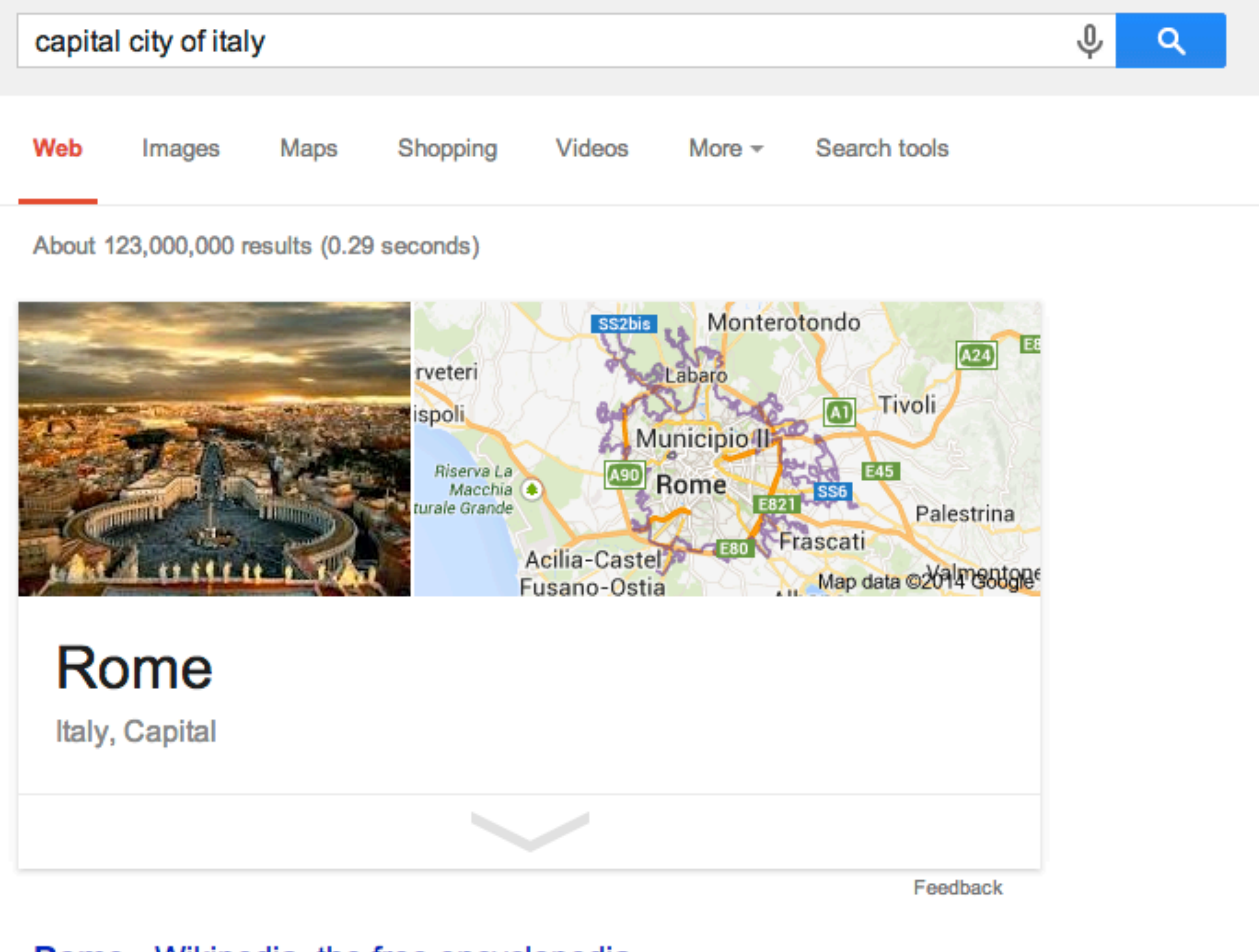
Asmara - **Wikipedia**, the free encyclopedia
en.**wikipedia**.org/wiki/Asmara ▾ Wikipedia ▾
Jump to **Italian** Eritrea - ... and when it was occupied by **Italy** in 1889 and was made the **capital city** of Eritrea in preference to Massawa by **Governor** Martini ...

Turin - **Wikipedia**, the free encyclopedia
en.**wikipedia**.org/wiki/Turin ▾ Wikipedia ▾
Jump to **City** centre - Via Roma crosses one of the **main** squares of the **city**: the pedestrianised ... senate and, for few years, the **Italian** senate after the **Italian** unification), the ... to Saint John the Baptist, which is the **major** church of the **city**.

Milan - **Wikipedia**, the free encyclopedia
en.**wikipedia**.org/wiki/Milan ▾ Wikipedia ▾
Its business district hosts the Borsa Italiana (**Italy's main** stock exchange) and the headquarters of the **largest** national banks and companies. The **city** is a **major** ...

Rome - **Wikipedia**, the free encyclopedia

capital city of italy



capital city of italy

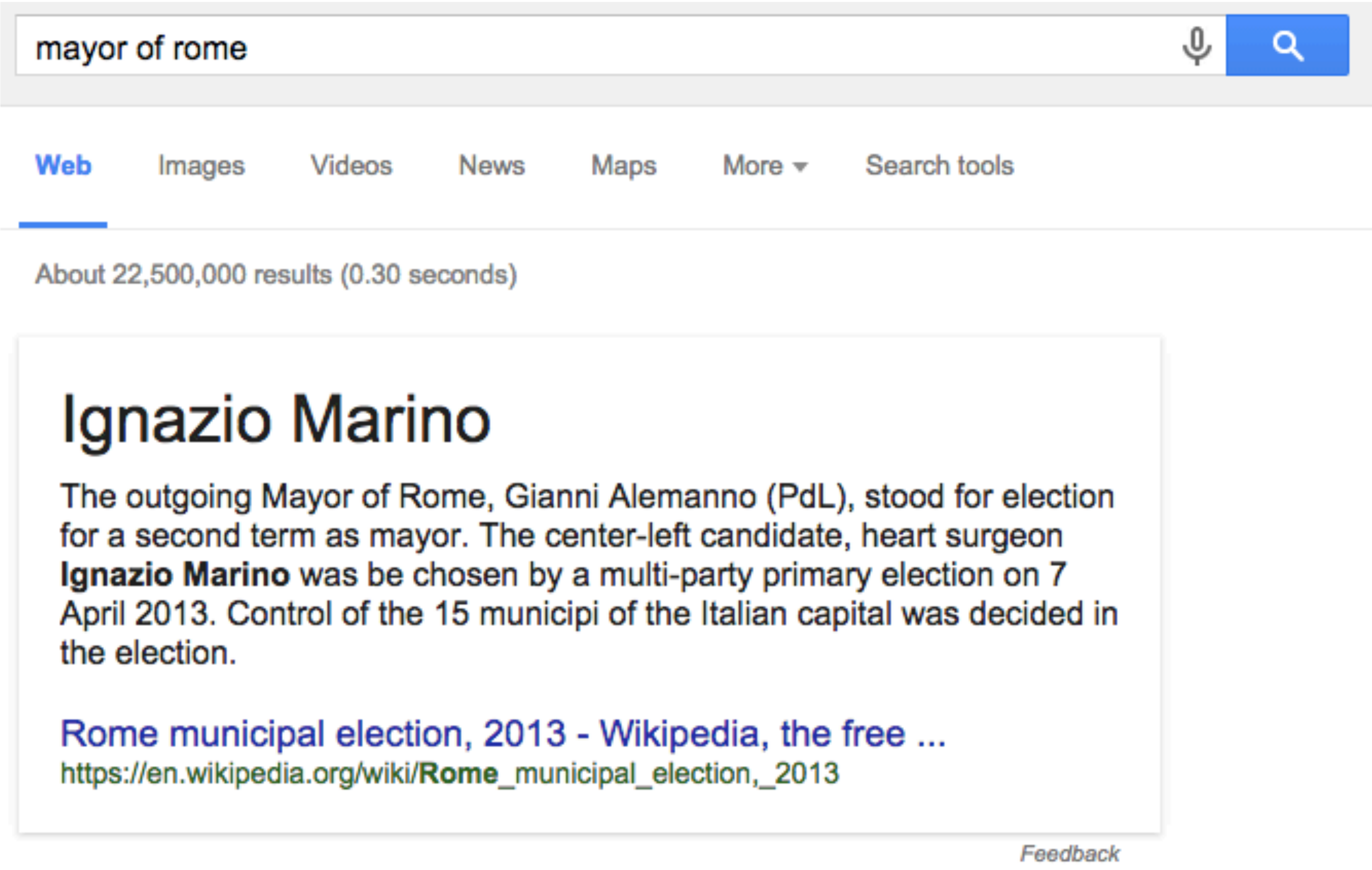
About 123,000,000 results (0.29 seconds)

Rome

Italy, Capital

Rome - Wikipedia, the free encyclopedia

mayor of rome



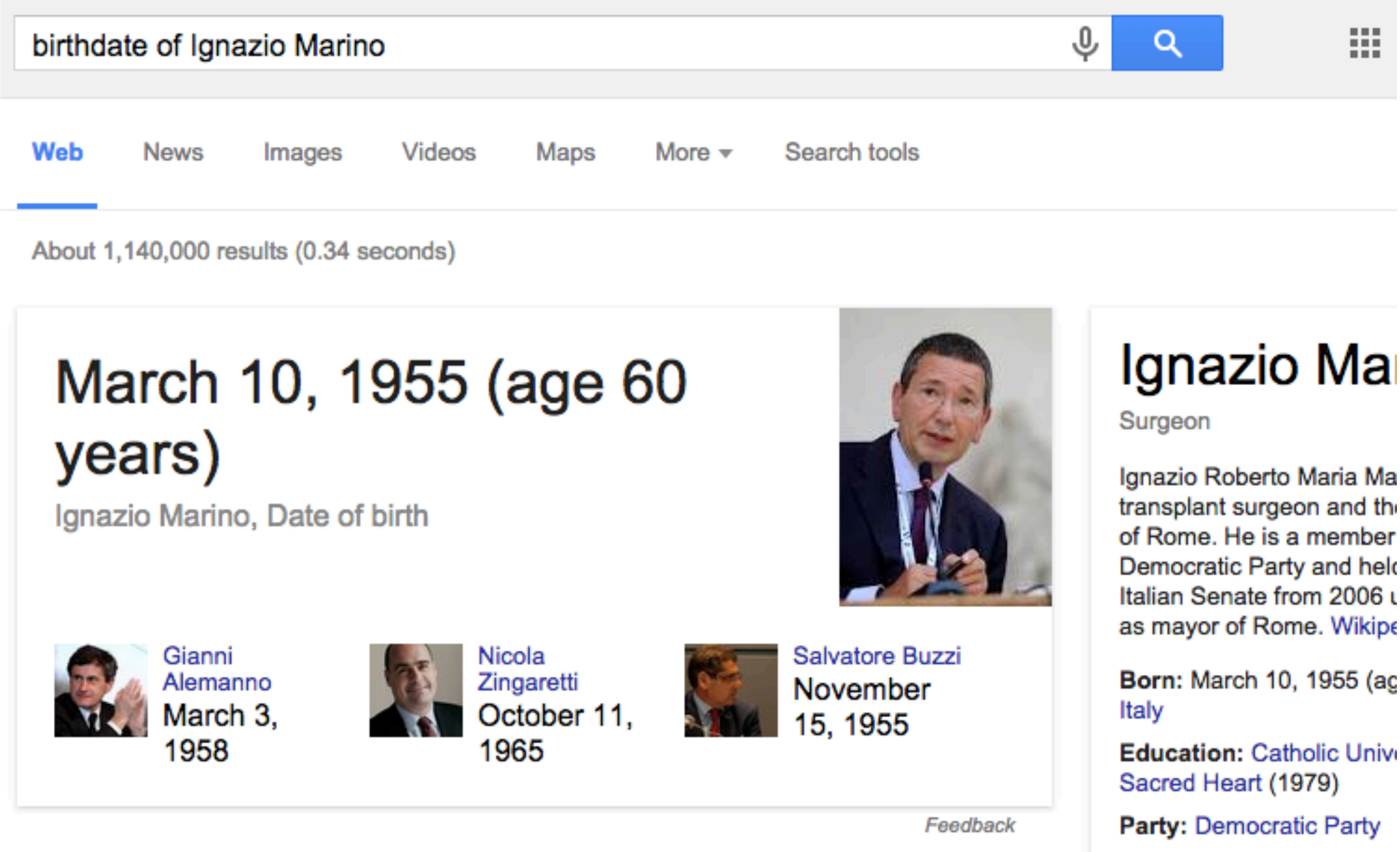
About 22,500,000 results (0.30 seconds)

Ignazio Marino

The outgoing Mayor of Rome, Gianni Alemanno (PdL), stood for election for a second term as mayor. The center-left candidate, heart surgeon **Ignazio Marino** was be chosen by a multi-party primary election on 7 April 2013. Control of the 15 municipi of the Italian capital was decided in the election.

Rome municipal election, 2013 - Wikipedia, the free ...
https://en.wikipedia.org/wiki/**Rome**_municipal_election,_2013

birthdate of ignazio marino



birthdate of Ignazio Marino

About 1,140,000 results (0.34 seconds)

March 10, 1955 (age 60 years)

Ignazio Marino, Date of birth

Gianni Alemanno
March 3, 1958

Nicola Zingaretti
October 11, 1965

Salvatore Buzzi
November 15, 1955

Ignazio Ma

Surgeon

Ignazio Roberto Maria Ma
transplant surgeon and th
of Rome. He is a member
Democratic Party and hel
Italian Senate from 2006
as mayor of Rome. Wikipe

Born: March 10, 1955 (ag
Italy

Education: Catholic Univ
Sacred Heart (1979)

Party: Democratic Party

Motivation

- Web Search Engines can answer simple factual queries directly on the result page

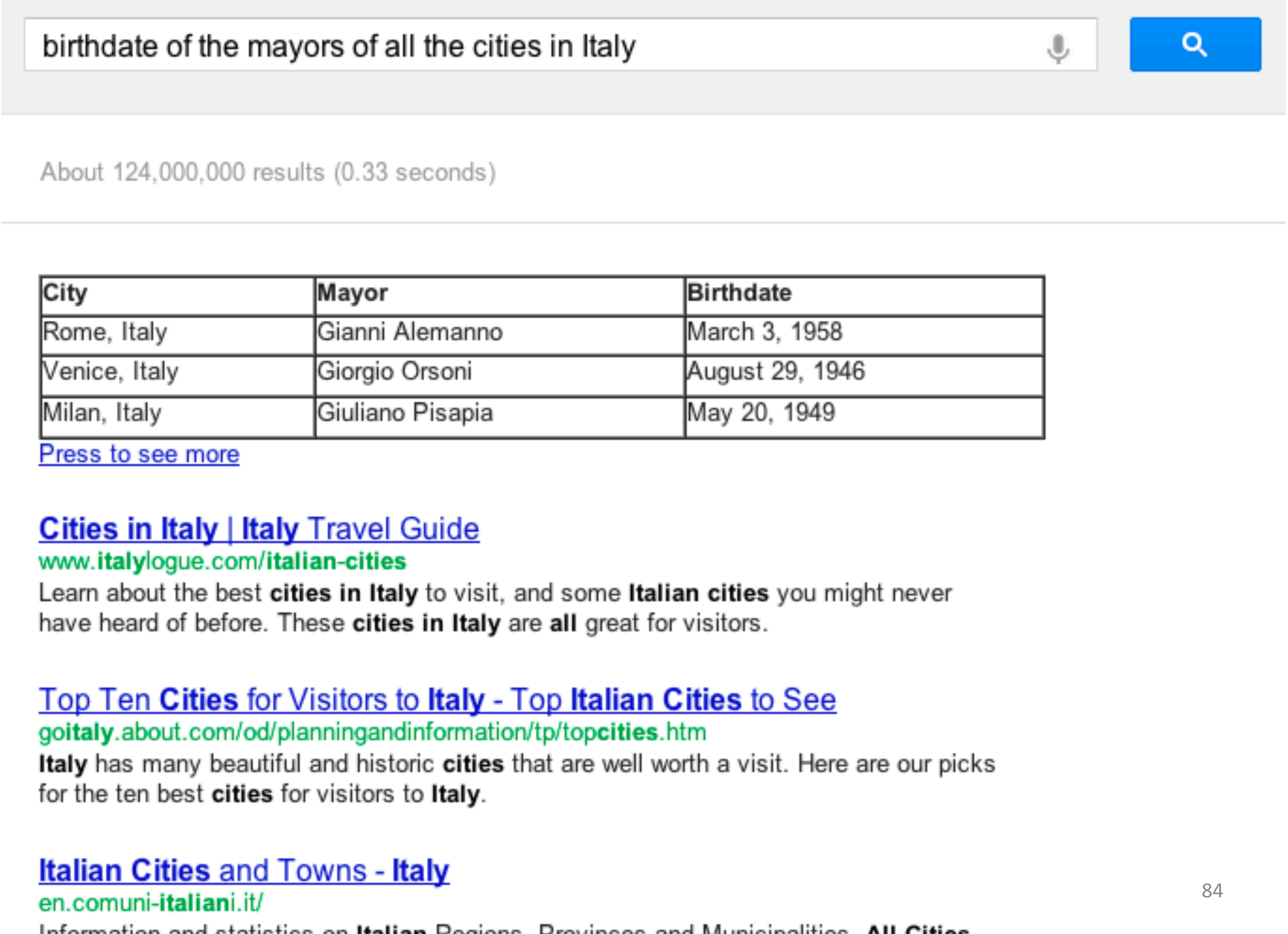
- Users with complex information needs are often unsatisfied

- Purely automatic techniques are not enough
- We want to solve it with Crowdsourcing!

CrowdQ

- CrowdQ is the first system that uses crowdsourcing to
  - *Understand* the intended meaning
  - *Build* a structured query template
  - *Answer* the query over Linked Open Data

Gianluca Demartini, Beth Trushkowsky, Tim Kraska, and Michael Franklin. **CrowdQ: Crowdsourced Query Understanding**. In: 6th Biennial Conference on Innovative Data Systems Research (**CIDR 2013**).



birthdate of the mayors of all the cities in Italy

About 124,000,000 results (0.33 seconds)

City	Mayor	Birthdate
Rome, Italy	Gianni Alemanno	March 3, 1958
Venice, Italy	Giorgio Orsoni	August 29, 1946
Milan, Italy	Giuliano Pisapia	May 20, 1949

Press to see more

Cities in Italy | Italy Travel Guide
www.**italy**logue.com/**italian-cities**
Learn about the best **cities in Italy** to visit, and some **Italian cities** you might never have heard of before. These **cities in Italy** are **all** great for visitors.

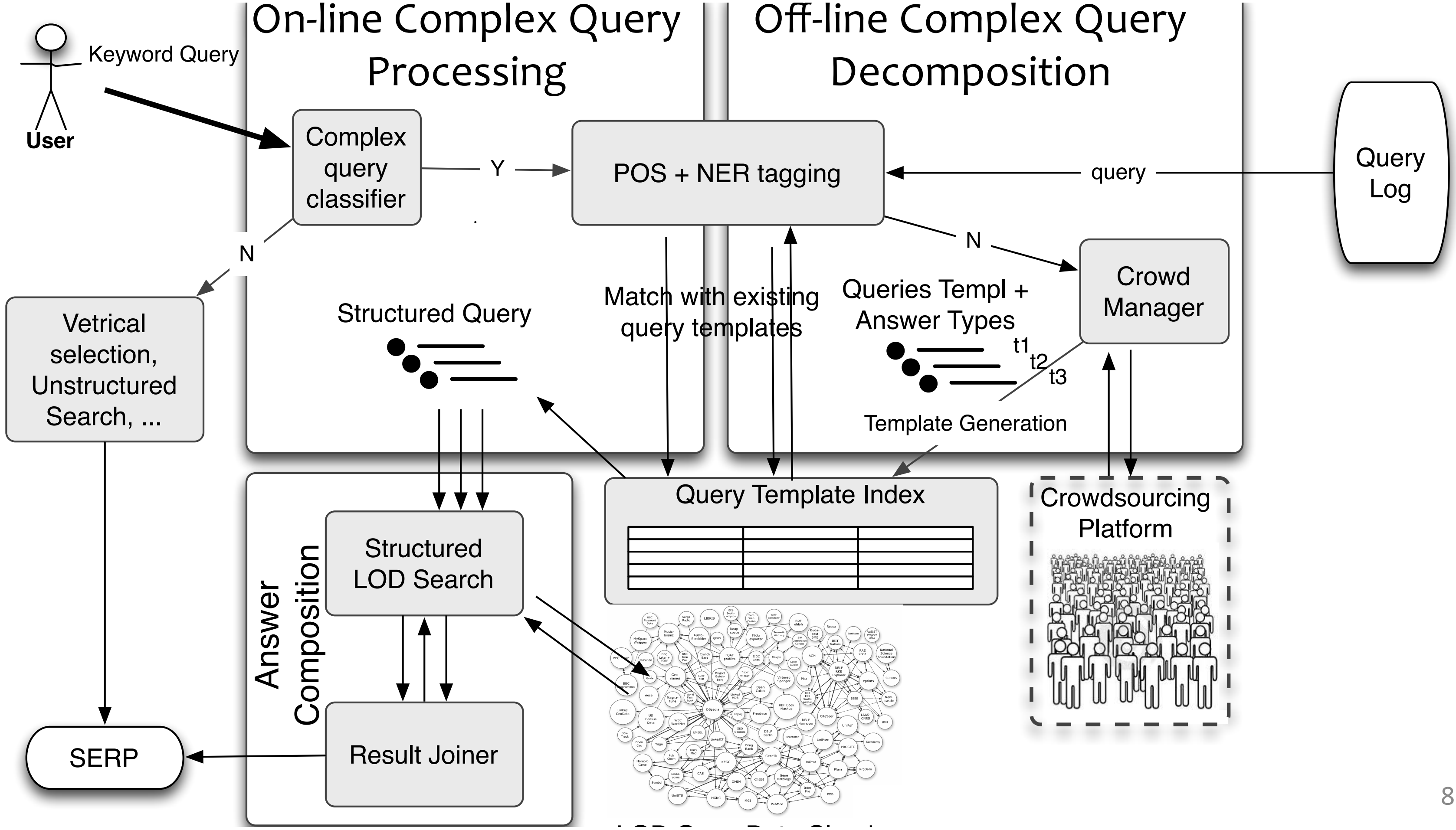
Top Ten **Cities** for Visitors to **Italy** - Top **Italian Cities** to See
go**italy**.about.com/od/planningandinformation/tp/top**cities**.htm
Italy has many beautiful and historic **cities** that are well worth a visit. Here are our picks for the ten best **cities** for visitors to **Italy**.

Italian Cities and Towns - **Italy**
en.comuni-**italian**i.it/
Information and statistics on **Italian** Regions, Provinces and Municipalities. **All Cities**

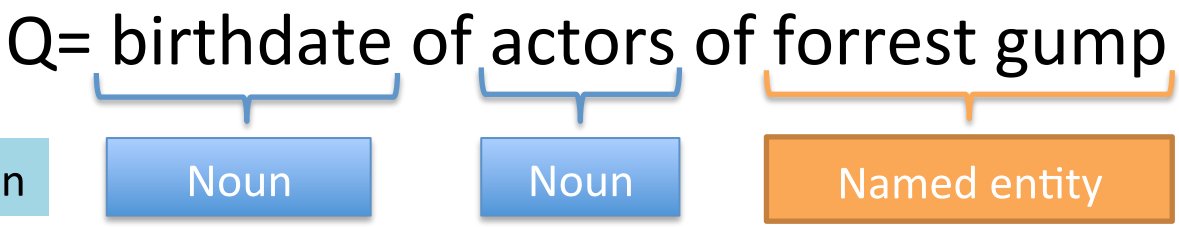
CrowdQ Architecture

Off-line: query template generation with the help of the crowd

On-line: query template matching using NLP and search over open data



Hybrid Human-Machine Pipeline



Query annotation

Verification

Is forrest gump this entity in the query?

Entity Relations

Which is the relation between: actors and forrest gump → starring

Schema element

Starring → <dbpedia-owl:starring>

Verification

Is the relation between:
Indiana Jones – Harrison Ford
Back to the Future – Michael J. Fox
of the same type as
Forrest Gump - actors

Structured query generation

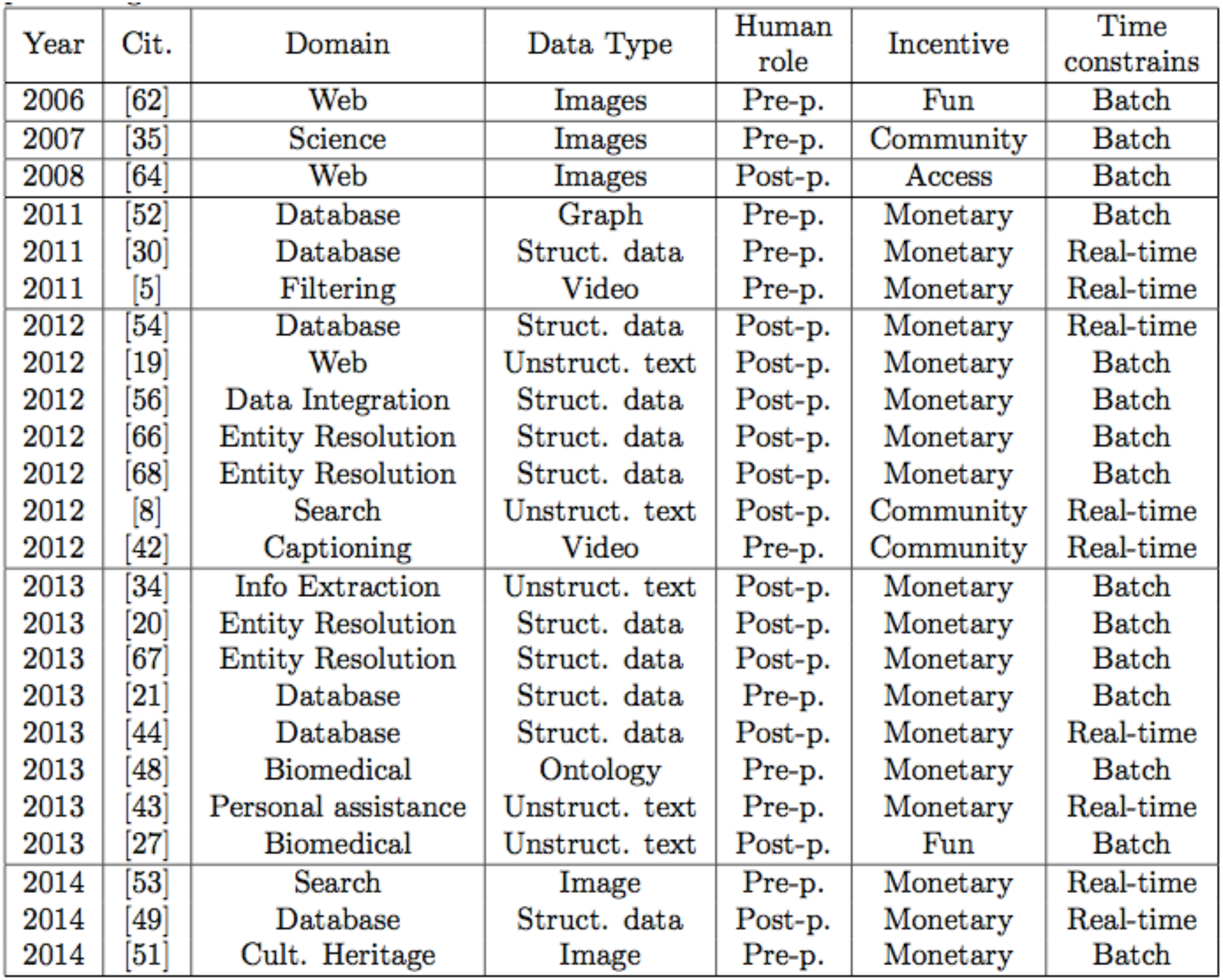
Q= birthdate of actors of fo MOVIE np

```
SELECT ?y ?x
WHERE { ?y <dbpedia-owl:birthdate> ?x .
        ?z <dbpedia-owl:starring> ?y .
        ?z <rdfs:label> ‘Fo MOVIE mp’ }
```

Results from BTC09:

```
<http://dbpedia.org/resource/Robin_Wright_Penn> 1966-04-08
<http://dbpedia.org/resource/Tom_Hanks> 1956-07-09
<http://dbpedia.org/resource/Sally_Field> 1946-11-06
<http://dbpedia.org/resource/Gary_Sinise> 1955-03-17
<http://dbpedia.org/resource/Mykelti_Williamson> 1960-03-04
```

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

Summary

- Crowdsourcing big data can make you go bankrupt! -> hybrid systems
- When to ask a human, when to trust the machine
- Hybrid systems (human in the loop)
  - Pre-processing: training data for ML
  - Post-processing: based on confidence scores
  - Mix: active learning

Gianluca Demartini**. Hybrid Human-Machine Information Systems: Challenges and Opportunities**. In: Computer Networks, Special Issue on Crowdsourcing, Elsevier.

Tutorial Outline

- Part 1
  - *Introduction to Crowdsourcing (30min)*
  - *Ensuring Quality in Paid Crowdsourcing (60min)*
- Part 2
  - *Hybrid Human-Machine Data Integration (30min)*
  - *Crowd-Powered Search (30min)*
  - ***Enterprise Crowdsourcing for Search** (30min)*

Enterprise Crowdsourcing for Search

Enterprise Crowdsourcing

- Internal crowd
  - Employees of the company
  - Full-time annotators
  - Casual crowd workers
- Pro: Trust, Domain Knowledge
- Contra: Incentives, Load-balancing

Crowds for Enterprise Crowdsourcing

- Internal Crowd
  - IBM
  - Microsoft
  - Google
- External Crowd
  - Amazon MTurk
  - Yandex Toloka toloka.yandex.com

Crowds for Enterprise Crowdsourcing

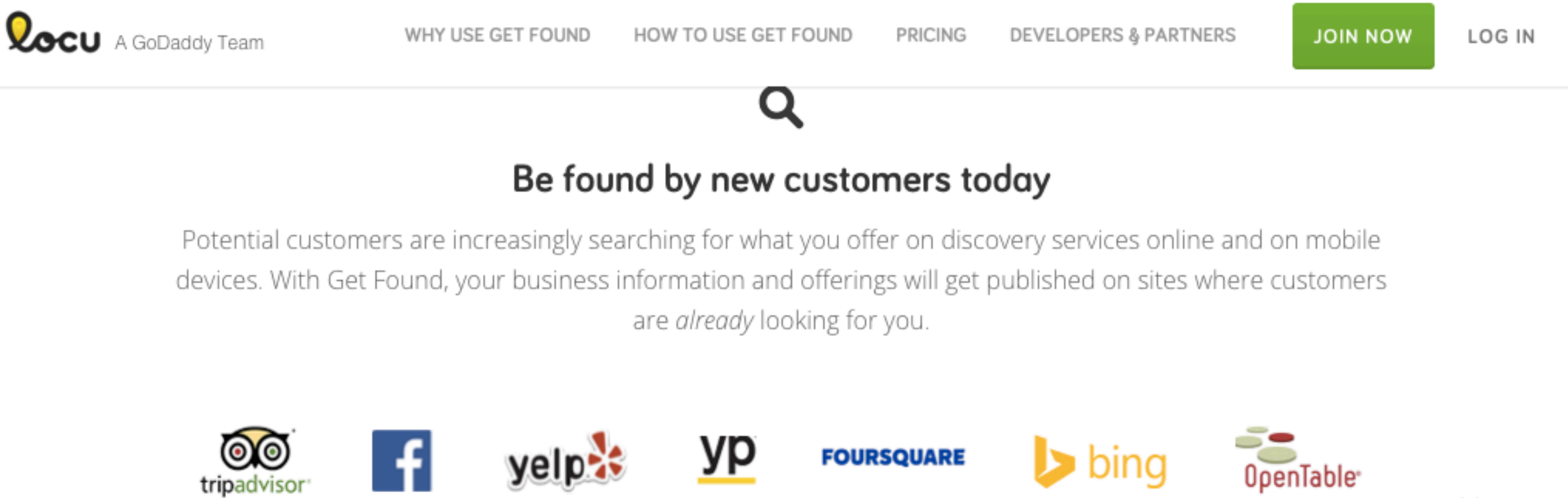
- Mixed

info.crowdflower.com/nda-contributors

  - NDA Crowds by Crowdflower
  - Top Tolokers become Yandex employees
- Tamr.com
  - Internal Expert-sourcing for data integration

Use of Crowdsourcing for data cleaning / extraction

- Locu / GoDaddy
  - http://www.oreilly.com/pub/e/3298
  - “learnings from 17 conversations with companies that make heavy use of crowd work”



Be found by new customers today

Potential customers are increasingly searching for what you offer on discovery services online and on mobile devices. With Get Found, your business information and offerings will get published on sites where customers are *already* looking for you.

Conclusions

- Crowdsourcing: a way to get manual data annotation / cleaning / processing at scale
- Applications to search
  - Evaluation / relevance judgments
  - Complex query understanding
  - Information Finding (e.g., customer care phone no)
  - Result extraction and aggregation in tabular format

Conclusions

- Challenges
  - **Quality** if public crowds are used
  - Many techniques can be used to guarantee high quality, commercial services are coming up
  - **Deadlines:** it is difficult to predict crowd execution time
  - Task reward can be used as a means to speed-up execution
  - **Cost:** can be reduced thanks to hybrid human-machine systems