

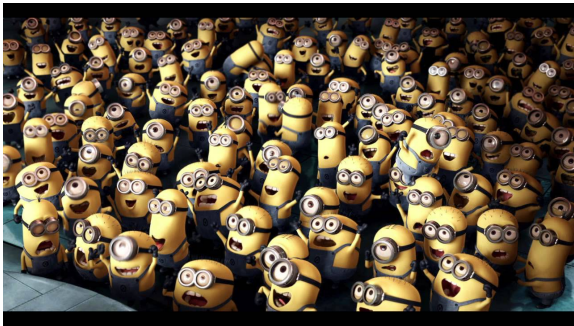
# Leveraging Knowledge Graphs for Web Search

**Part 4 - Crowdsourcing for Knowledge Graphs**

Gianluca Demartini

University of Sheffield

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



# 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, reCaptcha
- Incentives
  - Financial, fun, visibility
- See also my tutorial at ESWC 2013 and ISWC 2013



**Slides: [gianlucademartini.net/kg](http://gianlucademartini.net/kg)**

# Types of Crowdsourcing Tasks

Task Granularity	Examples
Complex Tasks	<ul style="list-style-type: none"><li>• Build a website</li><li>• Develop a software system</li><li>• Overthrow a government?</li></ul>
Simple Projects	<ul style="list-style-type: none"><li>• Design a logo and visual identity</li><li>• Write a term paper</li></ul>
Macro Tasks	<ul style="list-style-type: none"><li>• Write a restaurant review</li><li>• Test a new website feature</li><li>• Identify a galaxy</li></ul>
Micro Tasks	<ul style="list-style-type: none"><li>• Label an image</li><li>• Verify an address</li><li>• Simple entity resolution</li></ul>

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

# 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



# Case-Study: Amazon MTurk

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

# Amazon MTurk



## Make Money by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

### As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



## Get Results from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Register Now](#)

### As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



# MTurk is a Marketplace for HITs

## All HITs

1-10 of 3454 Results

Sort by:  

[Show all details](#) | [Hide all details](#)

[1](#) [2](#) [3](#) [4](#) [5](#) [Next](#) [Last](#)

<b>Provide Information about a Product</b> Requester: <a href="#">requester</a>	<b>HIT Expiration Date:</b> May 23, 2015 (4 weeks 1 day) <b>Time Allotted:</b> 25 minutes	<b>Reward:</b> \$0.05 <b>HITs Available:</b> 11526	<a href="#">View a HIT in this group</a>
<b>Product Attribute Tagging - April 17th Please read the instructions</b> Requester: <a href="#">slee</a>	<b>HIT Expiration Date:</b> May 23, 2015 (4 weeks 2 days) <b>Time Allotted:</b> 60 minutes	<b>Reward:</b> \$0.03 <b>HITs Available:</b> 23887	<a href="#">View a HIT in this group</a>
<b>Inv_B_2</b> Requester: <a href="#">rohzi0d</a>	<b>HIT Expiration Date:</b> May 22, 2015 (4 weeks 1 day) <b>Time Allotted:</b> 48 minutes	<b>Reward:</b> \$0.00 <b>HITs Available:</b> 19740	<a href="#">View a HIT in this group</a>
<b>Geo Result Relevance-Tue Apr 21 10:40:14 PDT 2015</b> Requester: <a href="#">Amazon Requester Inc.</a>	<b>HIT Expiration Date:</b> May 22, 2015 (4 weeks 1 day) <b>Time Allotted:</b> 60 minutes	<b>Reward:</b> \$0.00 <b>HITs Available:</b> 10734	<a href="#">View a HIT in this group</a>
<b>Type the text from the Images, carefully. Productivity and bonuses guaranteed.</b> Requester: <a href="#">CopyText Inc.</a>	<b>HIT Expiration Date:</b> Apr 30, 2015 (6 days 23 hours) <b>Time Allotted:</b> 10 minutes	<b>Reward:</b> \$0.01 <b>HITs Available:</b> 10590	<a href="#">View a HIT in this group</a>
<b>Transcribe up to 25 Seconds of Media to Text - Earn up to \$0.12 per HIT!</b> Requester: <a href="#">Crowdsurf Support</a>	<b>HIT Expiration Date:</b> Apr 21, 2016 (51 weeks 6 days) <b>Time Allotted:</b> 15 minutes	<b>Reward:</b> \$0.08 <b>HITs Available:</b> 6702	<a href="#">View a HIT in this group</a>
<b>Fun and Fast Fashion Tagging</b> Requester: <a href="#">gavin</a>	<b>HIT Expiration Date:</b> Apr 28, 2015 (5 days 11 hours) <b>Time Allotted:</b> 60 minutes	<b>Reward:</b> \$0.02 <b>HITs Available:</b> 6460	<a href="#">View a HIT in this group</a>
<b>Geo Result Relevance-Wed Apr 08 14:30:08 PDT 2015</b> Requester: <a href="#">Amazon Requester Inc.</a>	<b>HIT Expiration Date:</b> May 10, 2015 (2 weeks 2 days) <b>Time Allotted:</b> 60 minutes	<b>Reward:</b> \$0.00 <b>HITs Available:</b> 6182	<a href="#">View a HIT in this group</a>
<b>Transcribe up to 25 Seconds of General Content to Text - Earn up to \$0.14 per HIT!</b> Requester: <a href="#">Crowdsurf Support</a>	<b>HIT Expiration Date:</b> Apr 21, 2016 (51 weeks 6 days) <b>Time Allotted:</b> 15 minutes	<b>Reward:</b> \$0.09 <b>HITs Available:</b> 6043	<a href="#">View a HIT in this group</a>
<b>!Whac-a-mole by Gaze (hard mode) ! Play a 1min eye tracking game in the web browser! 0416</b> Requester: <a href="#">px</a>	<b>HIT Expiration Date:</b> Apr 23, 2015 (8 hours 40 minutes) <b>Time Allotted:</b> 60 minutes	<b>Reward:</b> \$0.10 <b>HITs Available:</b> 4682	<a href="#">View a HIT in this group</a>

[1](#) [2](#) [3](#) [4](#) [5](#) [Next](#) [Last](#)

# 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



# 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)

General

03/23/2015



04/23/2015

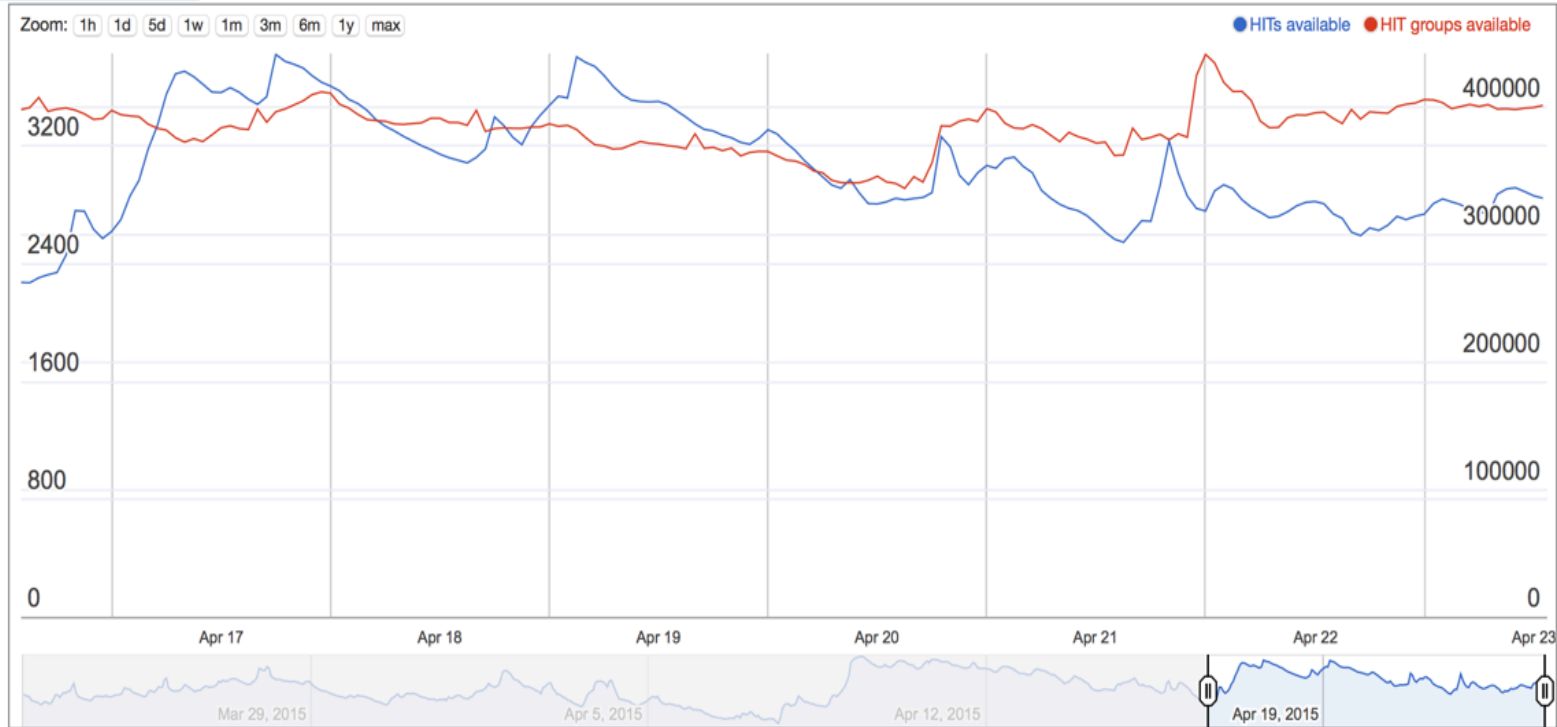


Market statistics

HIT groups posted

HITs posted

Rewards posted



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.

## A 5-years analysis of the Amazon MTurk market evolution:

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),

# Top requesters last week

Top-1000 Requesters, report for July 25, 2015 to August 24, 2015

Requester name	hits	reward
Speechpad	39932	\$270,902.29
WorkFusion	2215	\$9,496.00
CastingWords	10229	\$6,405.64
VidAngel	193	\$2,757.55
Amazon Requester Inc - browse classification	39795	\$2,387.70
University of California, Berkeley	170	\$2,210.00
p9r	20440	\$1,853.52
Mark Yatskar	19206	\$1,456.75
Mediaeval Search Hyper	13926	\$1,392.60
World Vision International	19769	\$1,262.00

# SLAs are expensive



[HOW IT WORKS](#) | [SERVICES](#) | [BUZZ](#) | [PRICING](#) | [FAQ](#) | [JOBS](#)

[REGISTER](#) | [SIGN IN](#)

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

# Why Crowdsourcing for IR?

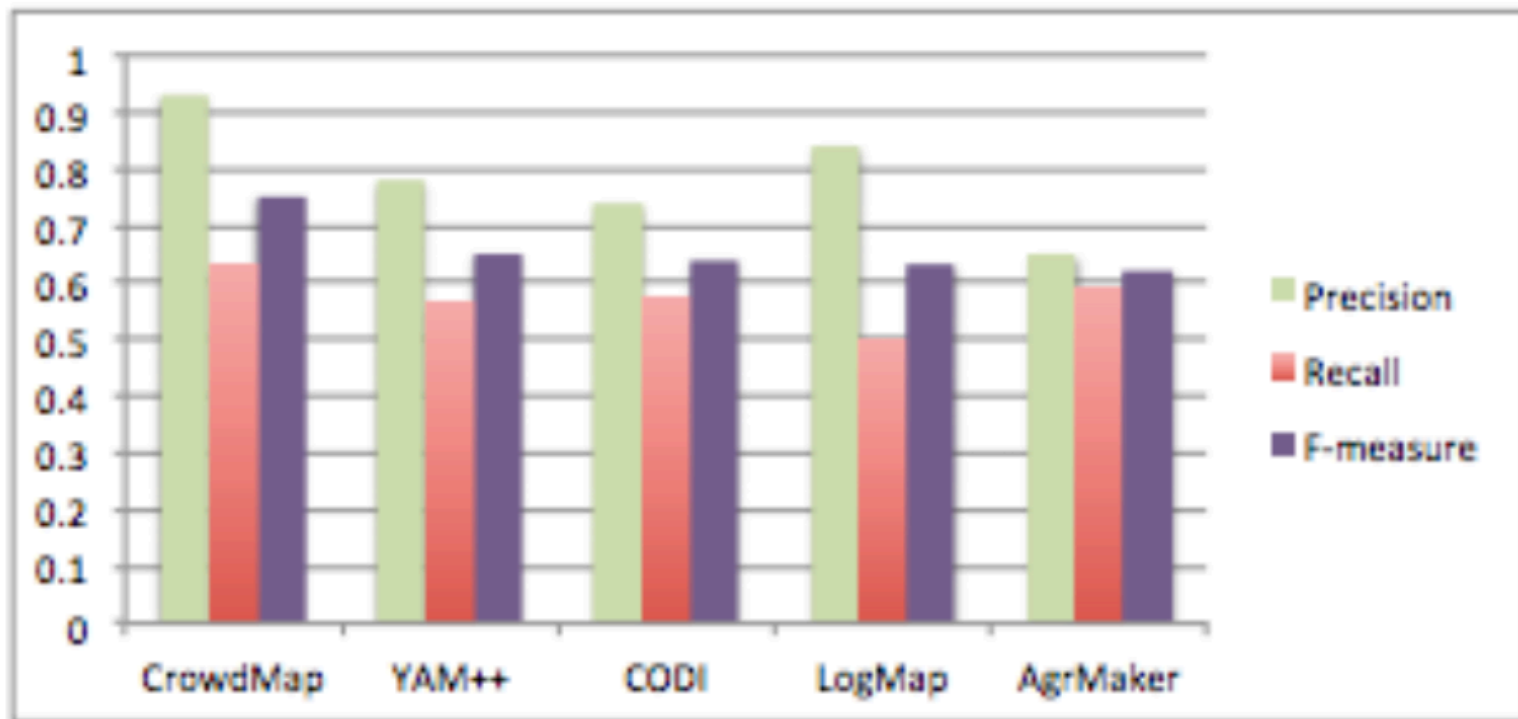
- Easy, cheap and fast labeling
- Ready-to use infrastructure – MTurk payments, workforce, interface widgets – CrowdFlower quality control mechanisms, etc.
- Allows early, iterative, frequent experiments – Iteratively prototype and test new ideas – Try new tasks, test when you want & as you go
- Proven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

# 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

# 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



# Application of Crowdsourcing to Knowledge Graphs

- Entity Linking (Demartini et al., WWW2012)
- Search Query Understanding (Demartini et al., CIDR2013)
- Search Result Extraction (Bernstein et al., CHI2012)
- KG enrichment (Ipeirotis and Gabrilovich, WWW2014)

# 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

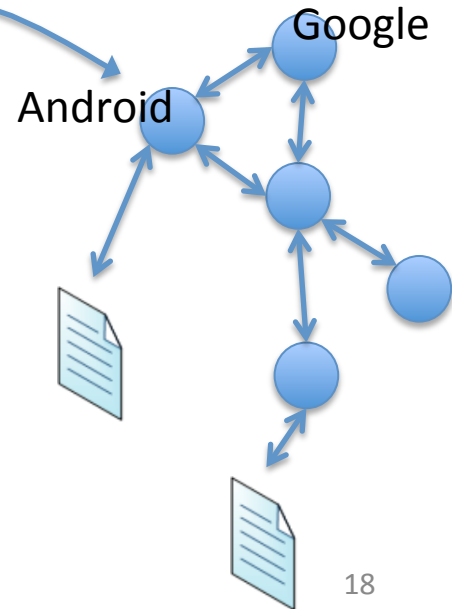
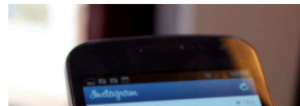
At long last, Instagram finally releases the Android version of its app.



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

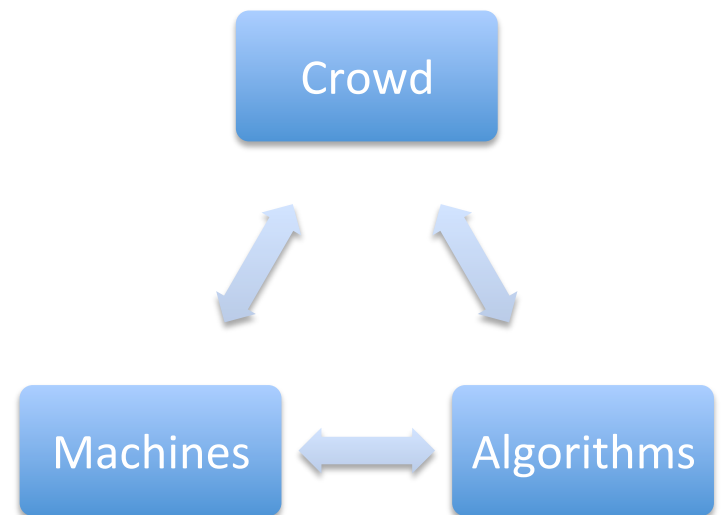
Follow

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.

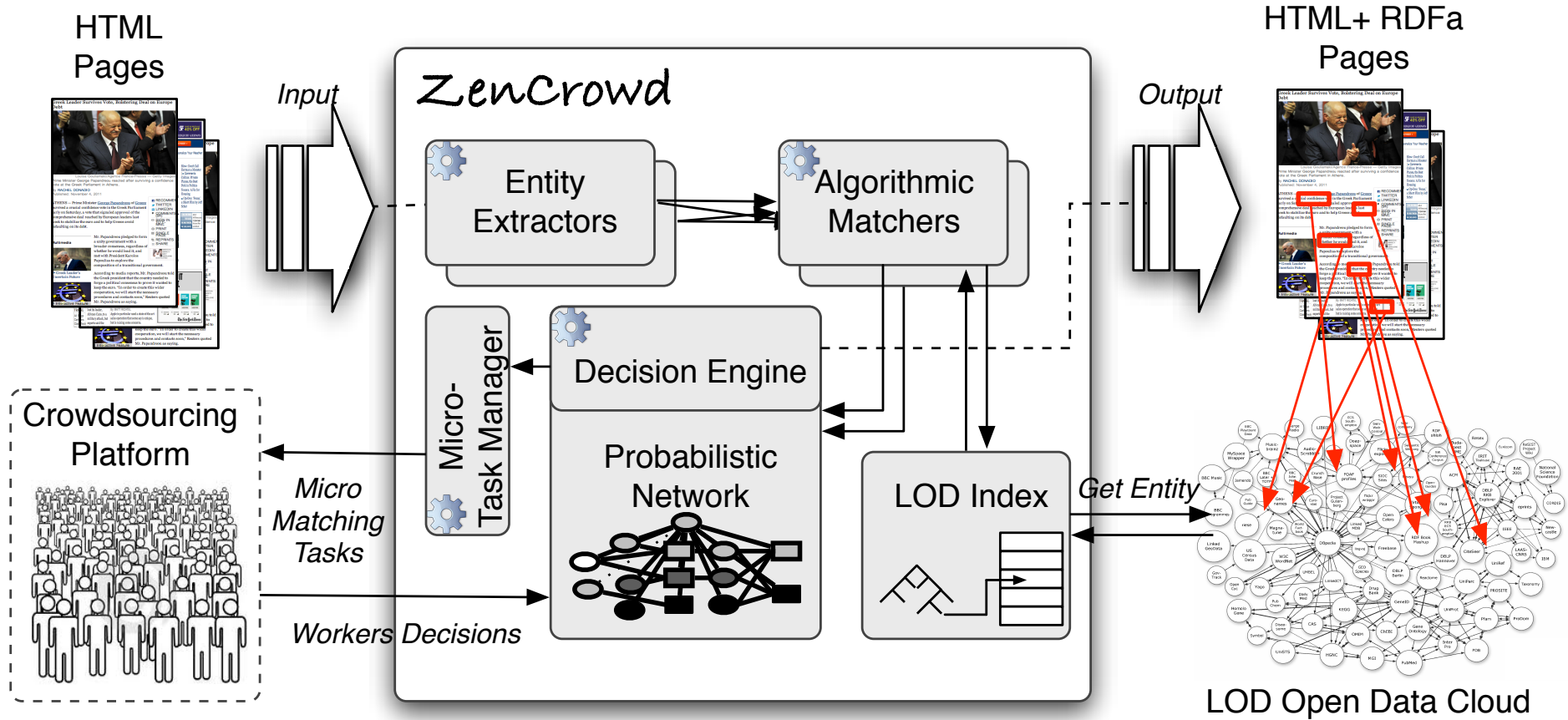


# ZenCrowd

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



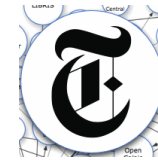
# ZenCrowd Architecture



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**).

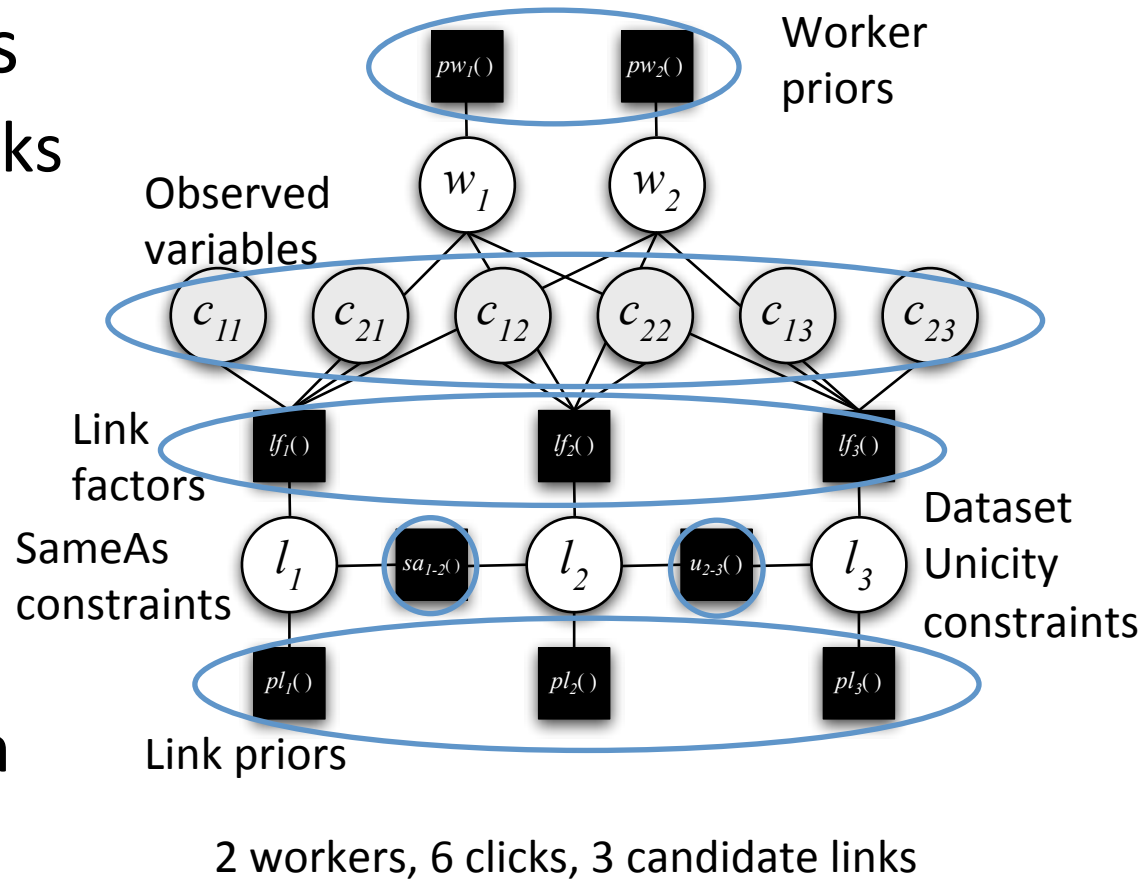
# Algorithmic Matching

- Inverted index over LOD entities
  - DBpedia, Freebase, Geonames, NYT
- TF-IDF (IR ranking function)
- Top ranked URIs linked to entities in docs
- Threshold on the ranking function or top N



# Entity Factor Graphs

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



# Entity Factor Graphs

- Training phase
  - Initialize worker priors
  - with  $k$  matches on known answers
- Updating worker Priors
  - Use link decision as new observations
  - Compute new worker probabilities
- Identify (and discard) unreliable workers

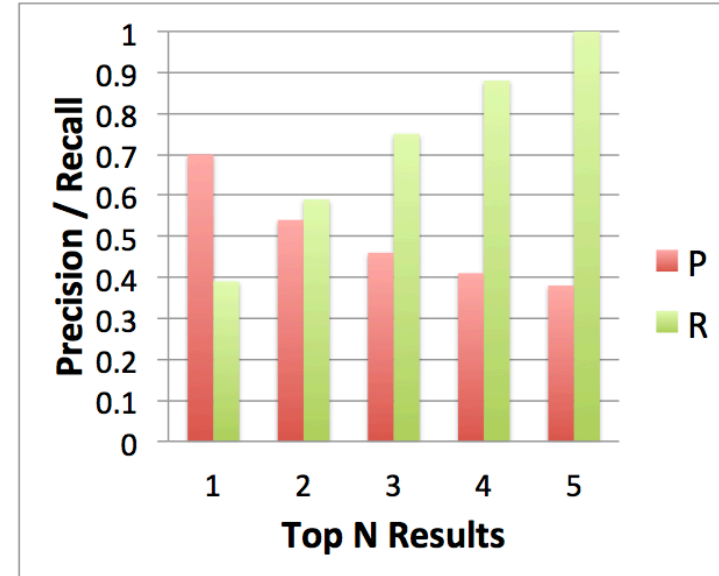
# Experimental Evaluation

- Datasets

- 25 news articles from

- CNN.com (Global news)
- NYTimes.com (Global news)
- Washington-post.com (US local news)
- Timesofindia.indiatimes.com (India news)
- Swissinfo.com (Switzerland local news)

- 40M entities (Freebase, DBPedia, Geonames, NYT)



	US Workers			Indian Workers		
	P	R	A	P	R	A
GL News	0.84	0.87	0.90	0.67	0.64	0.78
US News	0.64	0.68	0.78	0.55	0.63	0.71
IN News	0.84	0.82	0.89	0.75	0.77	0.80
SW News	0.72	0.80	0.85	0.61	0.62	0.73
All News	0.80	0.81	0.88	0.64	0.62	0.76



# Experimental Evaluation

- Entity Linking with Crowdsourcing and agreement vote (at least 2 out of 5 workers select the same URI)

	US Workers			Indian Workers		
	P	R	A	P	R	A
GL News	0.79	0.85	0.77	0.60	0.80	0.60
US News	0.52	0.61	0.54	0.50	0.74	0.47
IN News	0.62	0.76	0.65	0.64	0.86	0.63
SW News	0.69	0.82	0.69	0.50	0.69	0.56
All News	0.74	0.82	0.73	0.57	0.78	0.59

Top-1 precision: 0.70

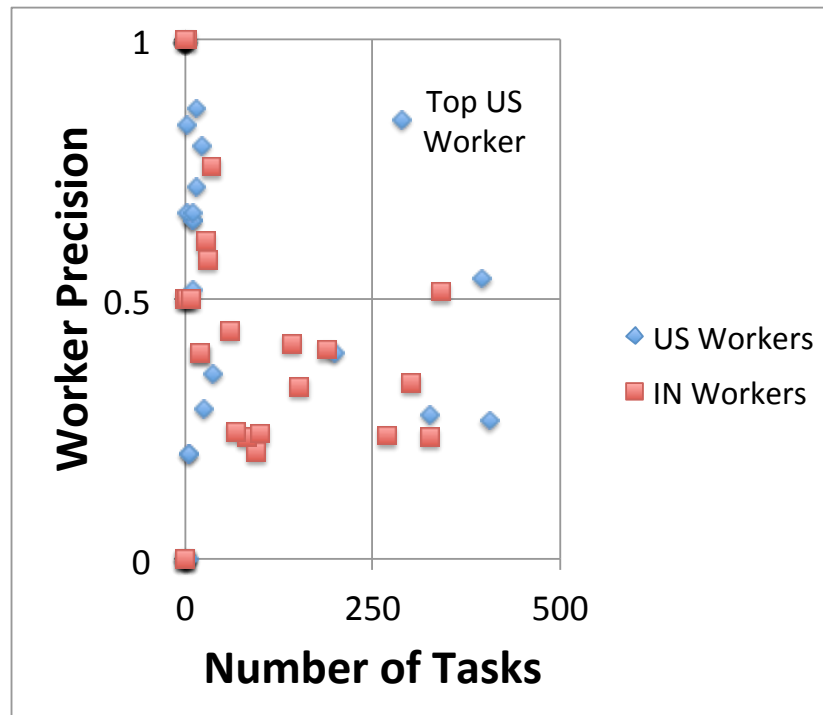
# Experimental Evaluation

- Entity Linking with ZenCrowd
  - Training with first 5 entities + 5% afterwards
  - 3 consecutive bad answers lead to blacklisting

	US Workers			Indian Workers		
	P	R	A	P	R	A
GL News	0.84	0.87	0.90	0.67	0.64	0.78
US News	0.64	0.68	0.78	0.55	0.63	0.71
IN News	0.84	0.82	0.89	0.75	0.77	0.80
SW News	0.72	0.80	0.85	0.61	0.62	0.73
All News	0.80	0.81	0.88	0.64	0.62	0.76

# Experimental Evaluation

- Worker Selection



# 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

# *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

# Three-stage blocking with the Crowd for Data Integration


- 1. Cheap clustering/inverted index selection of candidates
- 2. Expensive similarity measure
- 3. Crowdsourced low confidence matches

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Large-Scale Linked Data Integration Using Probabilistic Reasoning and Crowdsourcing. In: **VLDB Journal**, Volume 22, Issue 5 (2013), Page 665-687, Special issue on Structured, Social and Crowd-sourced Data on the Web. October 2013.

# Crowd-powered Direct Answers and Query Understanding



# Extract Direct Answers w/ Crowdsourcing

dog temperature 

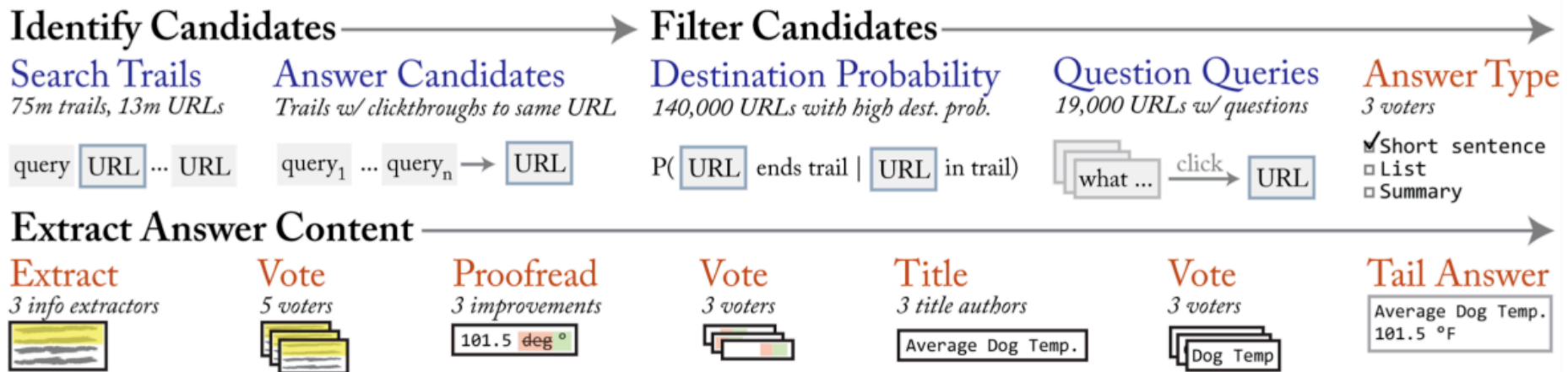
**Normal Body Temperature for Dogs**

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



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

# birthdate of the mayor of the capital city of italy

Web

Shopping

News

Images

Maps

More ▾

Search tools

About 3,830,000 results (0.46 seconds)

## Asmara - Wikipedia, the free encyclopedia

[en.wikipedia.org/wiki/Asmara](https://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](https://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](https://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



Web

Images

Maps

Shopping

Videos

More ▾

Search tools

About 123,000,000 results (0.29 seconds)



## Rome

Italy, Capital



Feedback

# mayor of rome

mayor of rome



**Web**

Images

Videos

News

Maps

More ▾

Search tools

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 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)

[https://en.wikipedia.org/wiki/Rome\\_municipal\\_election,\\_2013](https://en.wikipedia.org/wiki/Rome_municipal_election,_2013)

*Feedback*

# birthdate of ignazio marino

birthdate of Ignazio Marino



**Web**

News

Images

Videos

Maps

More ▾

Search tools

About 1,140,000 results (0.34 seconds)

**March 10, 1955 (age 60 years)**

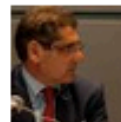
Ignazio Marino, Date of birth



**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 held Italian Senate from 2006 u as mayor of Rome. [Wikip](#)

**Born:** March 10, 1955 (ag Italy

**Education:** Catholic Univ Sacred Heart (1979)

**Party:** Democratic Party

[Feedback](#)

# 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.italylogue.com/italian-cities](http://www.italylogue.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](#)

[goitaly.about.com/od/planningandinformation/tp/topcities.htm](http://goitaly.about.com/od/planningandinformation/tp/topcities.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-italiani.it/](http://en.comuni-italiani.it/)

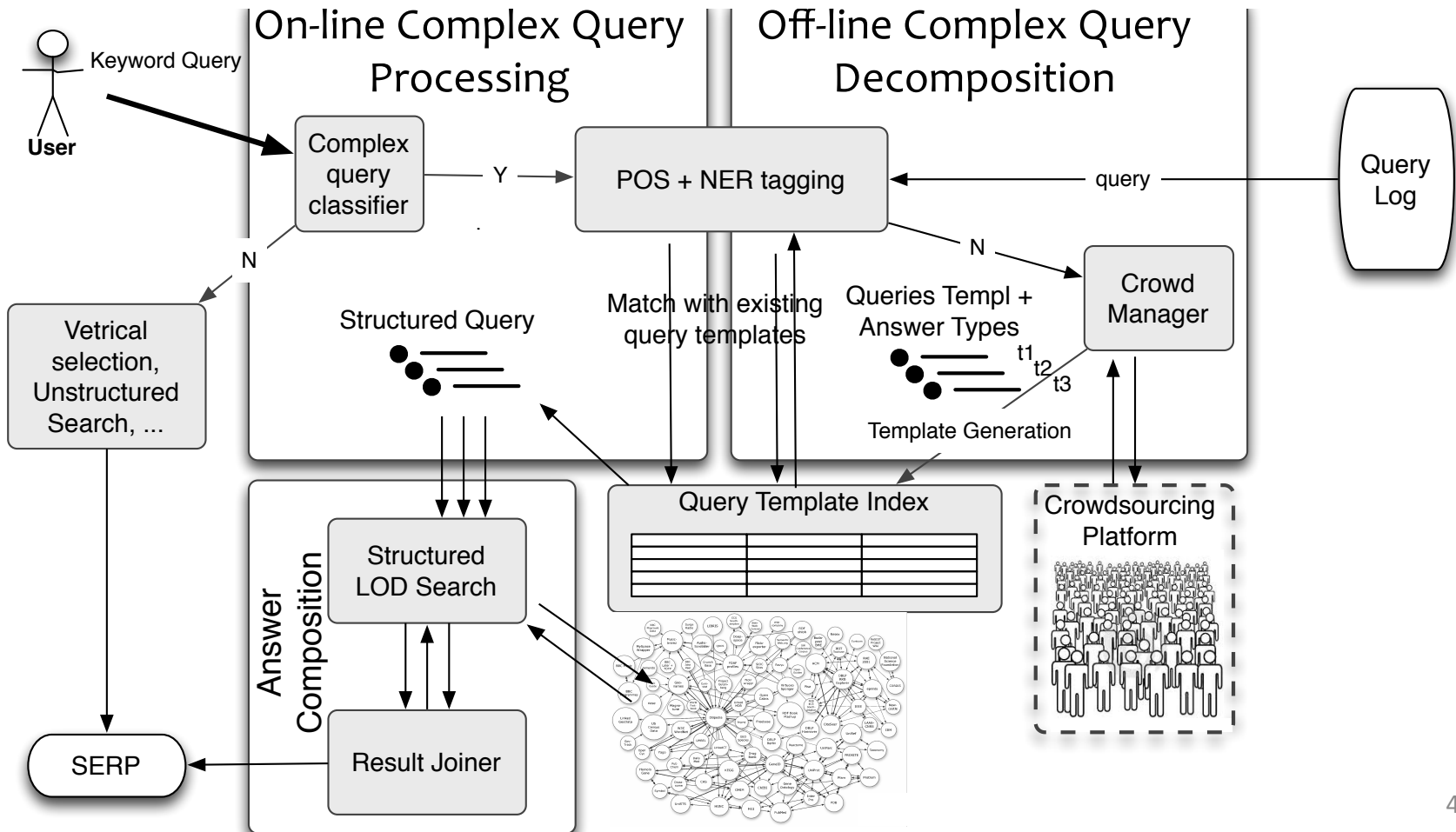
Information and statistics on **Italian Regions, Provinces, and Municipalities**. **All Cities**



# 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

Q= birthdate of actors of forrest gump

Query annotation

Noun

Noun

Named entity

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 mp

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
```

# Summary

- Crowdsourcing as a means to access a large number of on-line workers on-demand
- Hybrid human-machine systems to scale over large amount of data with high quality
  - Entity Linking
  - Data Integration
  - Answer extraction
  - Keyword query understanding

## **An overview of such systems:**

Gianluca Demartini. Hybrid Human-Machine Information Systems: Challenges and Opportunities. In: Computer Networks, Special Issue on Crowdsourcing, Elsevier, 2015.

# I'm hiring

- A post-doctoral researcher to start Jan 2016 (or later)
- On Crowdsourcing and Human Computation systems
- Get in touch
  - g.demartini@sheffield.ac.uk

**<http://bit.ly/bettercrowd>**