

Hybrid Human-machine Systems

Lecture 5

Gianluca Demartini

University of Sheffield

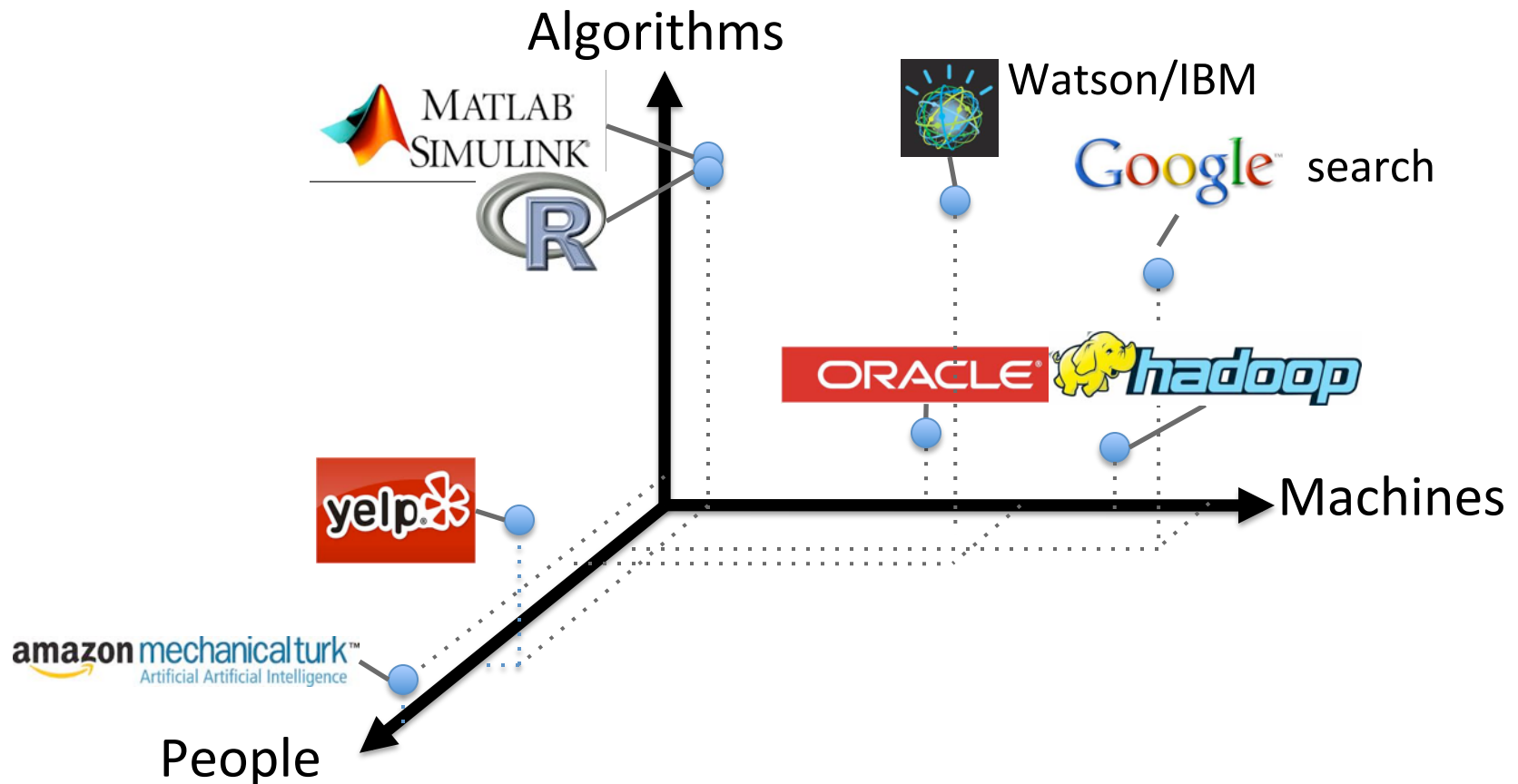
Background

- Hybrid systems
 - Combining the scalability of machines and the quality of human intelligence

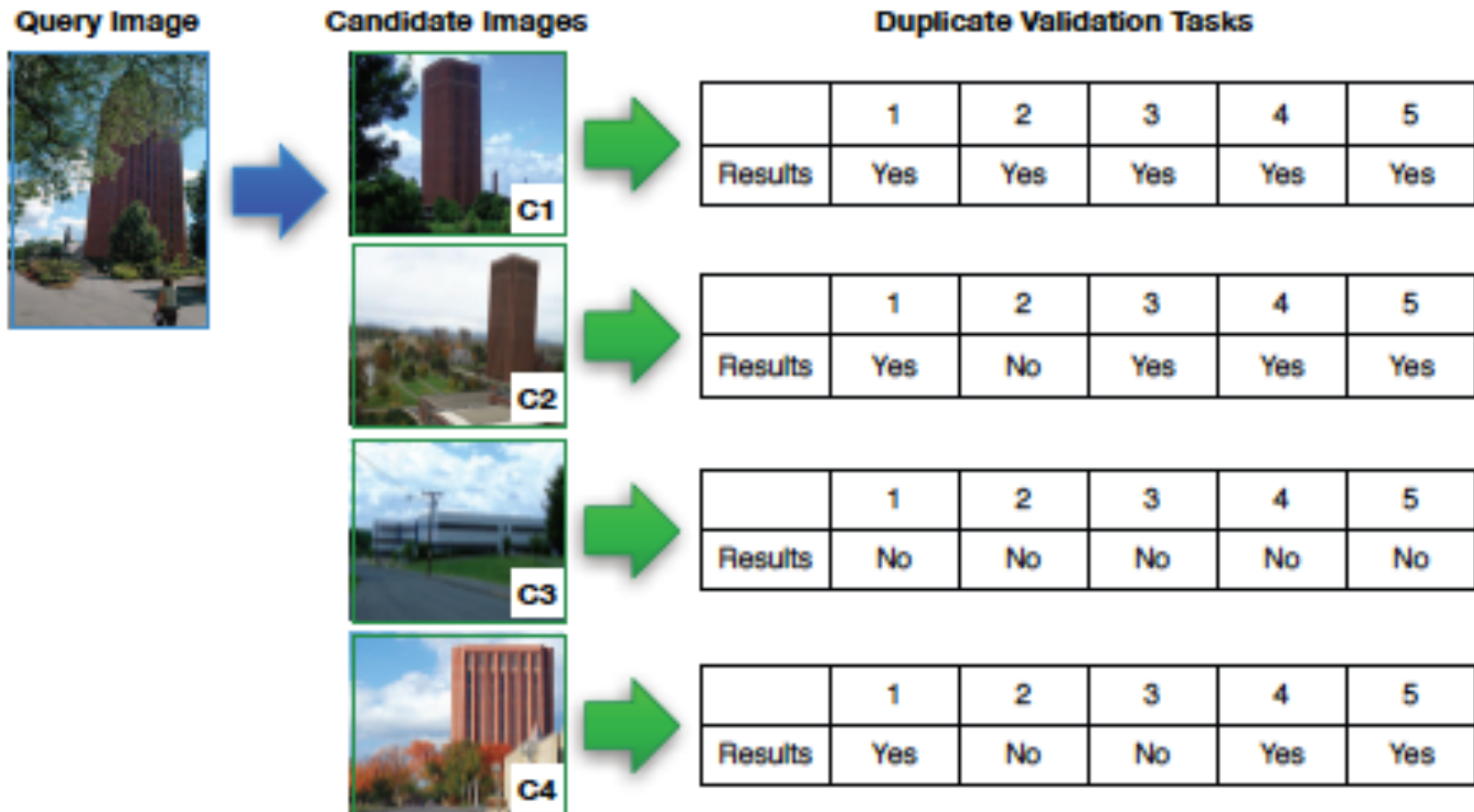
Hybrid Systems: Key Issues

- The role of machine (i.e., algorithm) and humans
 - use only humans? both? who's doing what?
- Quality control
- **Optimization: What to crowdsource**
- **Scalability: How much to crowdsource**

Thinking About Hybrid Systems



Example: Hybrid Image Search



Yan, Kumar, Ganesan, CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones, Mobisys 2010.

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

Example: Hybrid Query Processing

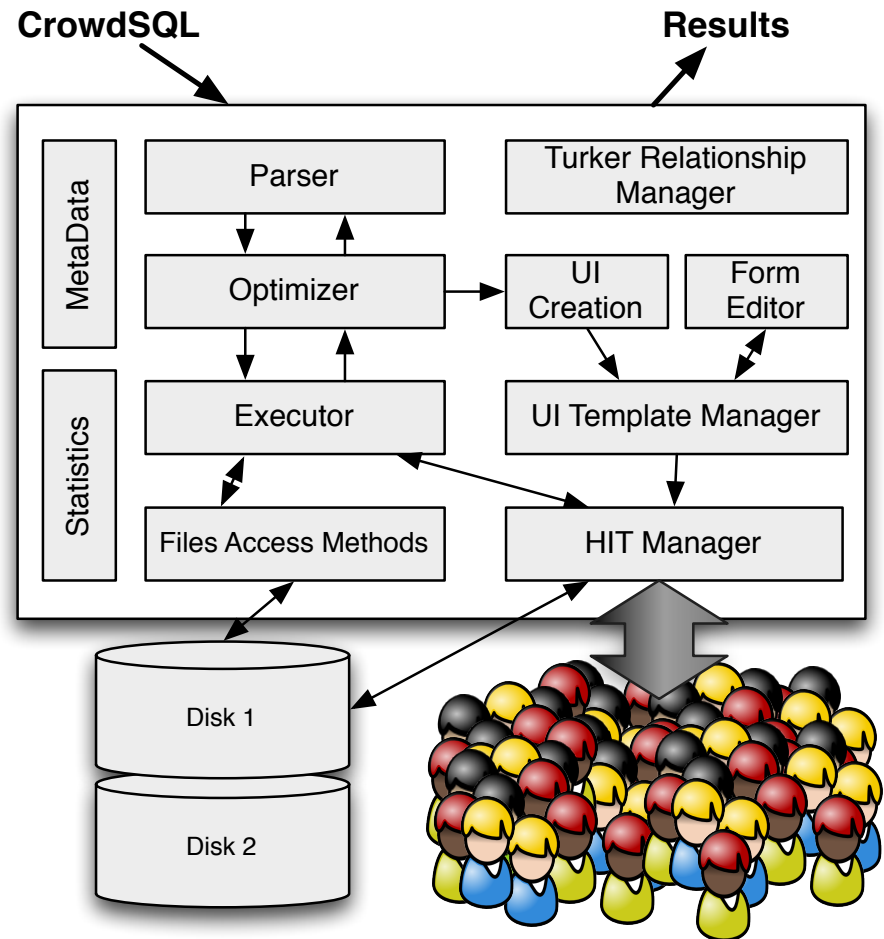
Use the crowd to answer DB-hard queries

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* ⁷

Crowdsourced Data Management Applications

- Relational
 - information extraction
 - schema matching
 - entity resolution
 - building structured KBs
 - sorting
 - top-k
 - ...
- Beyond relational
 - graph search
 - classification
 - mobile image search
 - social media analysis
 - question answering
 - NLP
 - text summarization
 - sentiment analysis
 - semantic wikis
 - ...

Qurk (MIT)

- Goal: crowd-source comparisons, missing data
- Basis: SQL3 + UDF
 - UDF encapsulate crowd input
 - special template language for crowd UDFs
 - specify UI, quality control, ... (possibly opt. hints)

Qurk Example [Markus et al. CrowdCrowd 2011]

- Task: Find all women in a “people” database
- Schema

```
CREATE TABLE people(  
    name varchar(256),  
    photo blob );
```

- Query

```
SELECT name  
FROM people p  
WHERE isFemale(p);
```


Qurk Example [Markus et al. CrowdCrowd 2011]

- Task: Find all women in a “people” database
- Schema

```
CREATE TABLE people(  
    name varchar(256),  
    photo blob );
```

- Query

```
SELECT name  
FROM people p  
WHERE isFemale(p);
```

TASK isFemale(tuple) TYPE: Filter

**Question: “is %s Female”,
tuple[“photo”]**

YesText: “Yes”

NoText: “No”

Qurk Example [Markus et al. CrowdCrowd 2011]

Is ____ Female?



Yes

No

TASK isFemale(tuple) **TYPE:** Filter

Question: “is %s Female”,
tuple[“photo”]

YesText: “Yes”

NoText: “No”

The magic is in the Templates

- Templates generate UIs for different kinds of crowd-sourcing tasks
 - filters: Yes / No questions
 - joins: comparisons between two tuples (equality)
 - order by: comparisons between two tuples (gt?)
 - generative: crowd-source attribute values
- Templates also specify quality control; e.g.,
COMBINER: MajorityVote

Crowdsourcing DB Systems

- Fundamentally new way of tackling data management issues using large networks of anonymous users
- At this point, first interesting systems and results, but still more questions than answers
 - Hot research topic
- Unique, unexpected issues
 - *“My database hates me”*

Problem: Populate Infoboxes

W Ernest Hemingway - Wik...
W Gail Caldwell - Wikipedia, x W Ray Bradbury - Wikiped...
W Gail Caldwell - Wikipedia, x
en.wikipedia.org/wiki/Gail_Caldwell
Log in / create account

Article Discussion Read Edit View history Search

Gail Caldwell

From Wikipedia, the free encyclopedia

Gail Caldwell (born 1951) was the chief book critic for *The Boston Globe*, where she was on staff from 1985 to 2009. Caldwell was the winner of the **2001 Pulitzer Prize for Criticism**. The award was for eight Sunday reviews and two other columns written in 2000. According to the Pulitzer Prize board, those columns were noted for "her insightful observations on contemporary life and literature."

Caldwell was born and raised in **Amarillo, Texas**. After graduating from **Tascosa High School**, she attended **Texas Tech University** for a while but transferred to **University of Texas at Austin** and obtained two degrees in **American studies**. She was an instructor at the University of Texas until 1981. Before joining the *The Boston Globe*, Caldwell taught feature writing at **Boston University**, worked as the arts editor of the *Boston Review* and wrote for the publications *New England Monthly* and *Village Voice*.

She lives in **Cambridge, Massachusetts** and wrote the 2006 memoir, *A Strong West Wind : A Memoir* (ISBN 1-4000-6248-9) and the 2010, *Let's Take the Long Way Home*, a memoir of her friendship with author **Caroline Knapp**. She has a Samoyed named Tula.

Gail Caldwell was born in 1951. At the age of 6 months, she caught polio.

External links

Solution: IE Using Machine + Human

Hemingway was an American author ... Infobox

The American readers ...

Born July 21, 1899
Nationality American

 Train a “nationality” extractor

Apply extractor to new pages
to extract nationalities.

Verify with crowd



“ray bradbury”

Google



- 17



WIKIPEDIA
The Free Encyclopedia

navigation

- [Main Page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

interaction

- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Donate to Wikipedia](#)
- [Help](#)

search

toolbox

- [What links here](#)
- [Related changes](#)
- [Special pages](#)
- [Printable version](#)
- [Permanent link](#)
- [Cite this page](#)

languages

[article](#)[discussion](#)[edit this page](#)[history](#)

Ray Bradbury

From Wikipedia, the free encyclopedia

Ray Douglas Bradbury (born [August 22 1920](#)) is an [American literary](#), [fantasy](#), [horror](#), [science fiction](#), and [mystery](#) writer best known for *[The Martian Chronicles](#)*, a 1950 book which has been described both as a [short story](#) collection and a novel, and his 1953 [dystopian](#) novel *[Fahrenheit 451](#)*. He is widely considered to be one of the greatest and most popular [American](#) writers of [speculative fiction](#) during the twentieth century.

Contents [\[show\]](#)

Beginnings

Bradbury was born in [Waukegan, Illinois](#), to a [Swedish](#) and telephone [lineman](#).^[1] His [paternal](#) grandfather and

Bradbury was a reader and writer throughout his youth, in [Waukegan](#). He used this library as a setting for much of his novel *[Something Wicked This Way Comes](#)*, and depicted Waukegan as "Green Town" in some of his other semi-[autobiographical](#) novels — *[Dandelion Wine](#)*, *[Farewell Summer](#)* — as well as in many of his short stories.^[3]

He attributes his lifelong habit of writing every day to an incident in 1932 when a carnival entertainer, Mr. Electrico^[4], touched him with an electrified sword, made his hair stand on end, and shouted, "Live forever!"

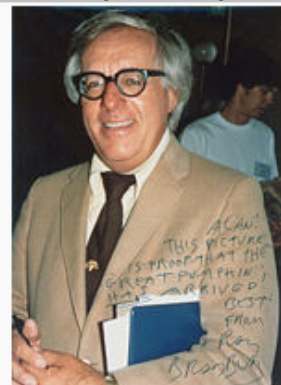
The Bradbury family lived in [Tucson, Arizona](#), in 1926–27 and 1932–33 as his father pursued employment, each time returning to Waukegan, and eventually settled in [Los Angeles](#) in 1934, when Ray was thirteen.

Bradbury graduated from the [Los Angeles High School](#) in 1938 but chose not to attend college. Instead, he sold newspapers at the corner of South Norton Avenue and Olympic Boulevard. He continued to educate himself at the local library, and having been influenced by [science fiction](#) heroes like [Flash Gordon](#) and [Buck Rogers](#), he began to publish science fiction stories in [fanzines](#) in 1938. Ray was invited by [Forrest J Ackerman](#) to attend the now legendary Clifton's Cafeteria Science Fiction Club. Here Ray met the writers [Robert A. Heinlein](#) [Emil Petaja](#) [Fredric Brown](#) [Henry Kuttner](#) [Leigh Brackett](#) and [Jack Williamson](#)

Wikipedia needs your help to improve the quality of this summary.

See the popups on this page.

Ray Bradbury



Ray Bradbury in 1975

Born	Check our guess.
	Check our guess.
Died	Check our guess.
Occupation	Writer , Playwright
Nationality	Check our guess.
Genres	Check our guess.
Influences	
Influenced	

[Official website] [Official website](#)



WIKIPEDIA
The Free Encyclopedia

[article](#)
[discussion](#)
[edit this page](#)
[history](#)

Ray Bradbury

From Wikipedia, the free encyclopedia

Ray Douglas Bradbury (born **August 22 1920**) is an **American literary, fantasy, horror, science fiction, and mystery** writer best known for *The Martian Chronicles*, a 1950 book which has been described both as a **short story** collection and a novel, and his 1953 **dystopian** novel *Fahrenheit 451*. He is widely considered to be one of the greatest and most popular **American** writers of **speculative fiction** during the twentieth century.

Contents [\[show\]](#)

Beginnings

Bradbury was born in **Waukegan, Illinois** and telephone **lineman**.^[1] His **paternal**

Bradbury was a reader and writer the **Waukegan**. He used this library as a *Comes*, and depicted Waukegan as *Dandelion Wine*, *Farewell Summer*

He attributes his lifelong habit of writing **Mr. Electrico**^[4], touched him with a **forever!**"

The Bradbury family lived in **Tucson** employment, each time returning to **Ray** was thirteen.

Bradbury graduated from the **Los Angeles High School** in 1938 but chose n

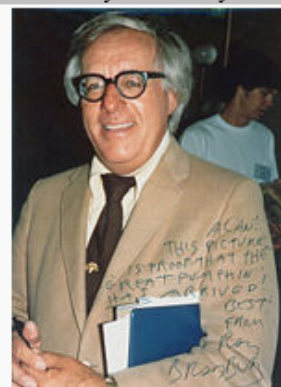
Instead, he sold newspapers at the corner of South Norton Avenue and Olympic Boulevard. He

continued to educate himself at the local library, and having been influenced by **science fiction** heroes like **Flash Gordon** and **Buck Rogers**, he began to publish science fiction stories in **fanzines** in 1938. Ray was invited by **Forrest J Ackerman** to attend the now legendary Clifton's Cafeteria Science Fiction Club. Here Ray met the writers **Robert A. Heinlein** **Emil Petaja** **Fredric Brown** **Henry Kuttner** **Leigh Brackett** and **Jack Williamson**

Wikipedia needs your help to improve the quality of this summary.

See the **highlighted** text on page.

Ray Bradbury



Ray Bradbury in 1975

Born	Check our guess.
	Check our guess.
Died	Check our guess.
Occupation	Writer, Playwright
Nationality	Check our guess.
Genres	Check our guess.
Influences	
Influenced	

[Official website] [Official website](#)

We think the article includes Ray Bradbury's **birth_date**.

Does the article say?

☐ **August 22 1920**

It seems likely the article says this in the sentence:

"Ray Douglas Bradbury (born **August 22 1920**) is an American literary, fantasy, horror, science fiction, and mystery writer best known for The Martian Chronicles, a 1950 book which has been described both as a short story collection and a novel, and his 1953 dystopian novel Fahrenheit 451."

☐ **No**

[Submit](#)

[Cancel](#)

navigation

- [Main Page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

interaction

- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Donate to Wikipedia](#)
- [Help](#)

search

[Go](#)

[Search](#)

toolbox

- [What links here](#)
- [Related changes](#)
- [Special pages](#)
- [Printable version](#)
- [Permanent link](#)
- [Cite this page](#)

languages

Other examples of hybrid systems

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>

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>

<http://dbpedia.org/resource/Instagram>

owl:sameAs

fbase:Instagram

RDFa
enrichment

<p><cite property="rdfs:label">Facebook</cite> 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 <cite property="rdfs:label">Instagram</cite>, the popular photo-sharing application, for about \$1 billion in cash and stock, the company said Monday.</p>

CNET > News > Mobile

Instagram for Android is now available

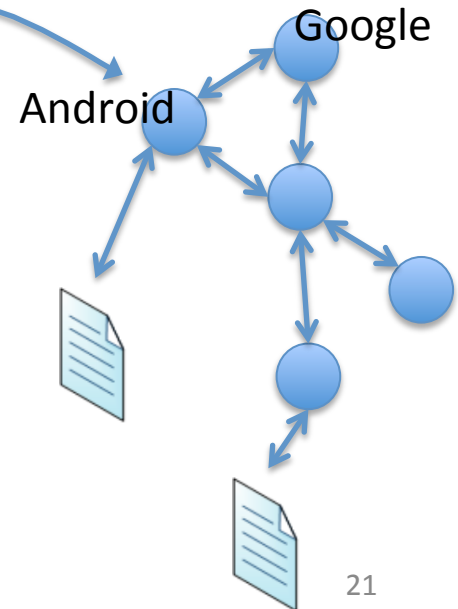
At long last, Instagram finally releases the Android version of its app.



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

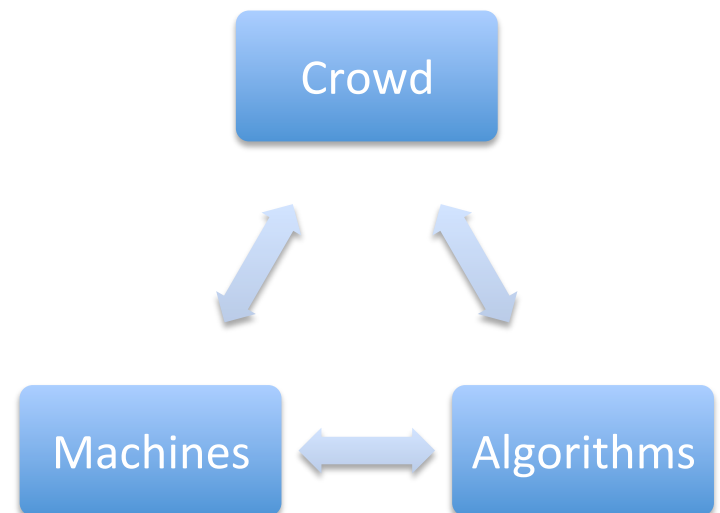
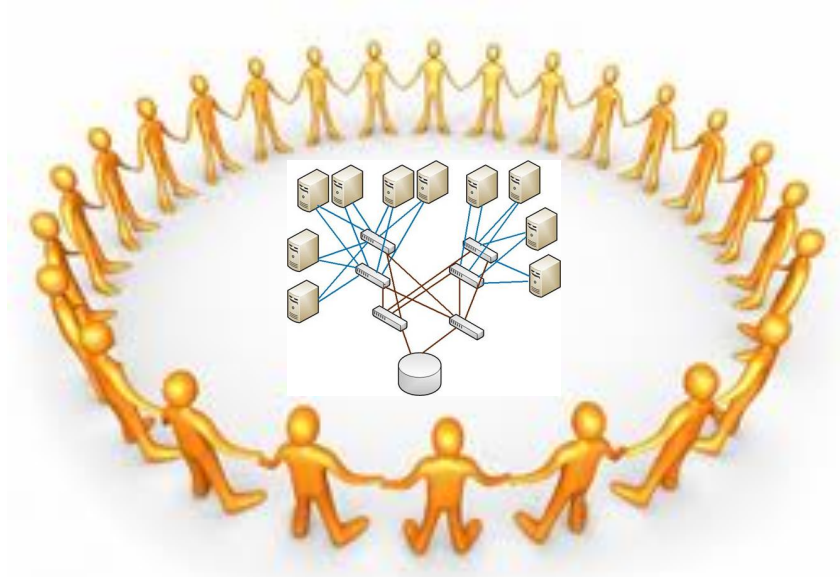
Follow

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

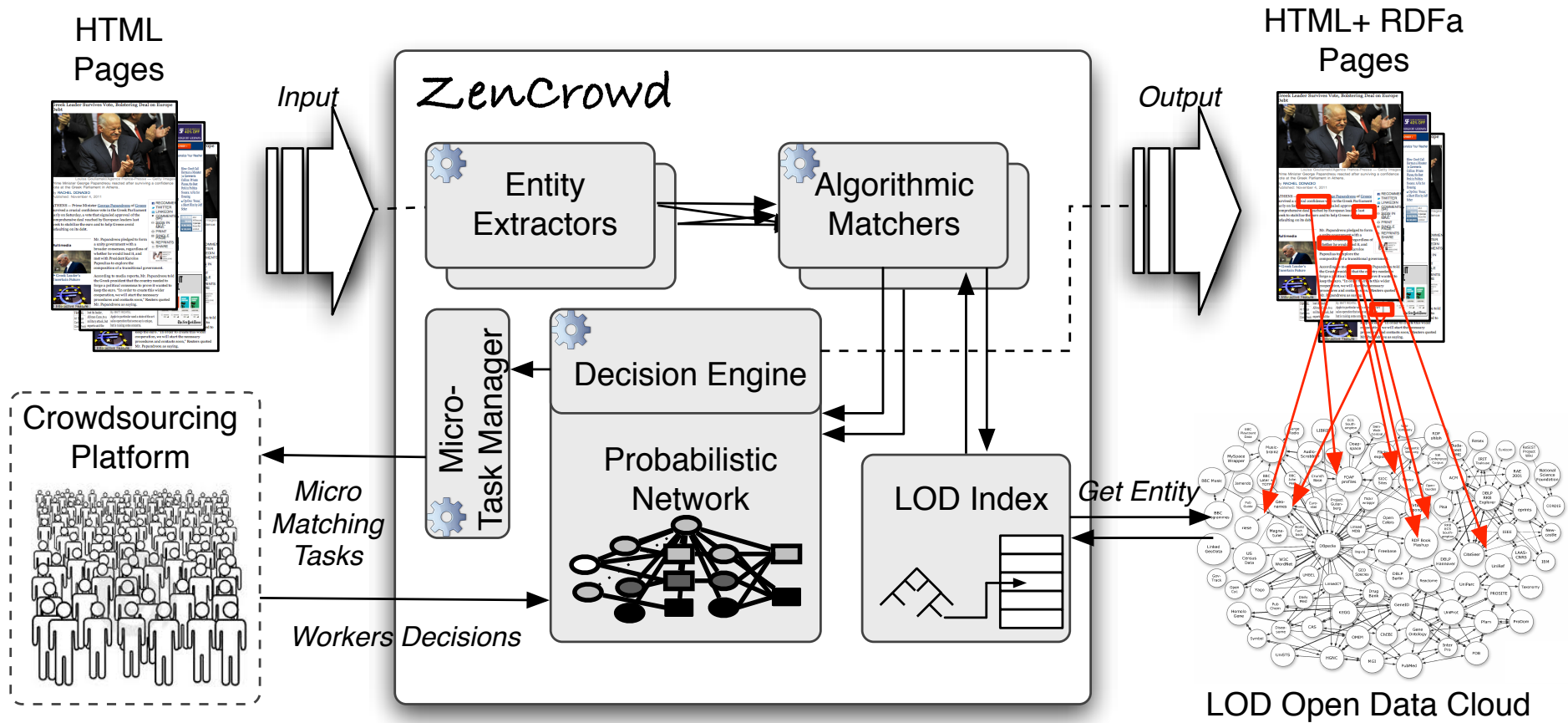


ZenCrowd

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



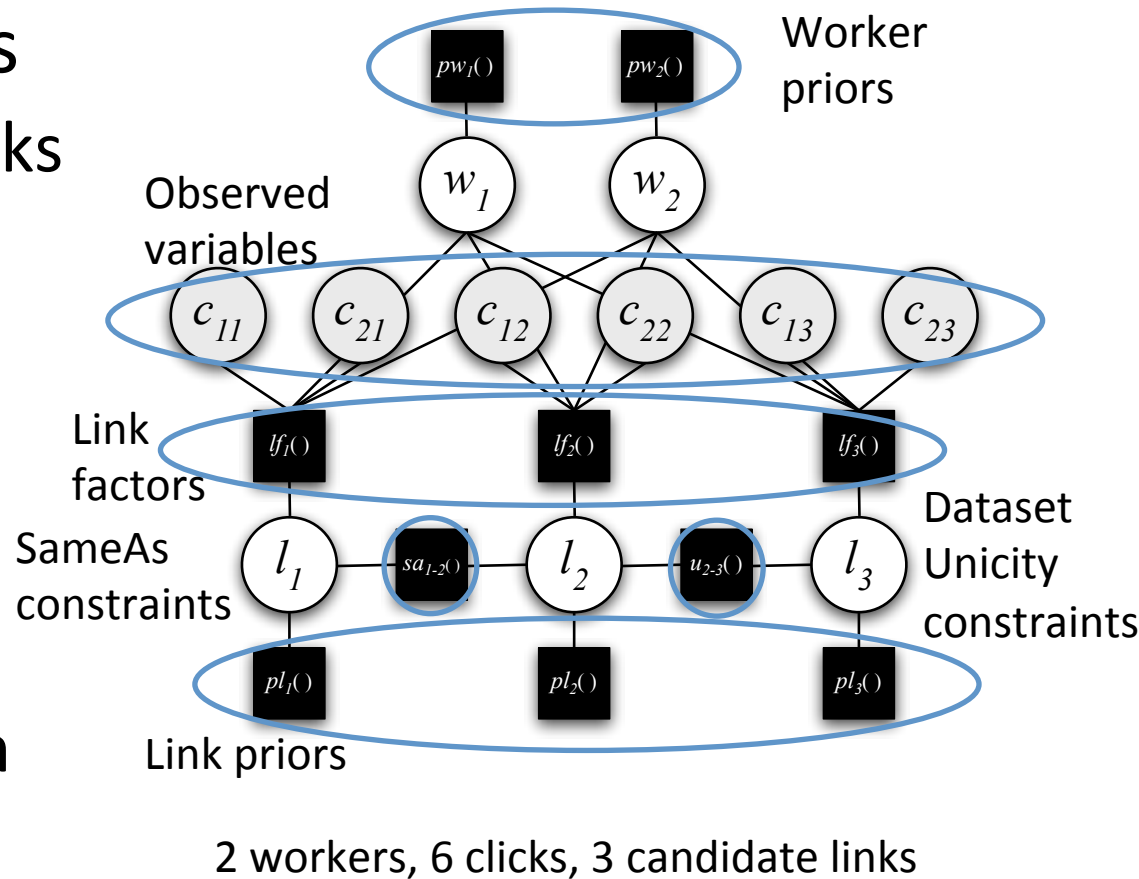
ZenCrowd Architecture



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

Entity Factor Graphs

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



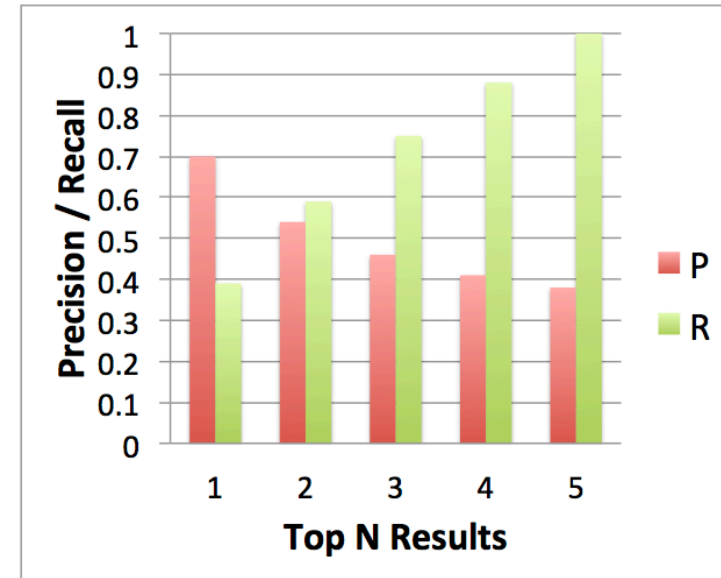
Experimental Evaluation

- Datasets

- 25 news articles from

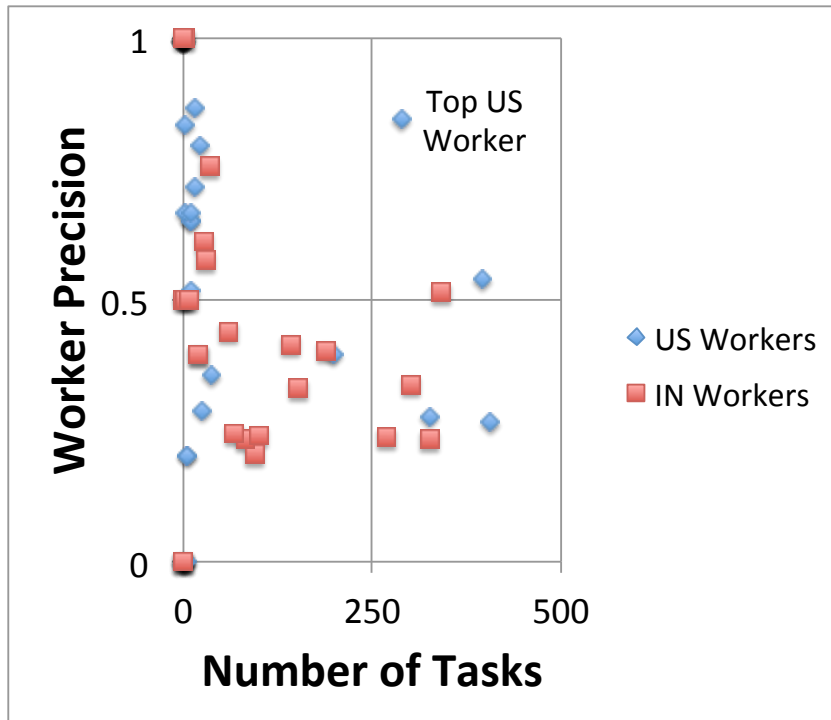
- CNN.com (Global news)
 - NYTimes.com (Global news)
 - Washington-post.com (US local news)
 - Timesofindia.indiatimes.com (India news)
 - Swissinfo.com (Switzerland local news)

- 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

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)

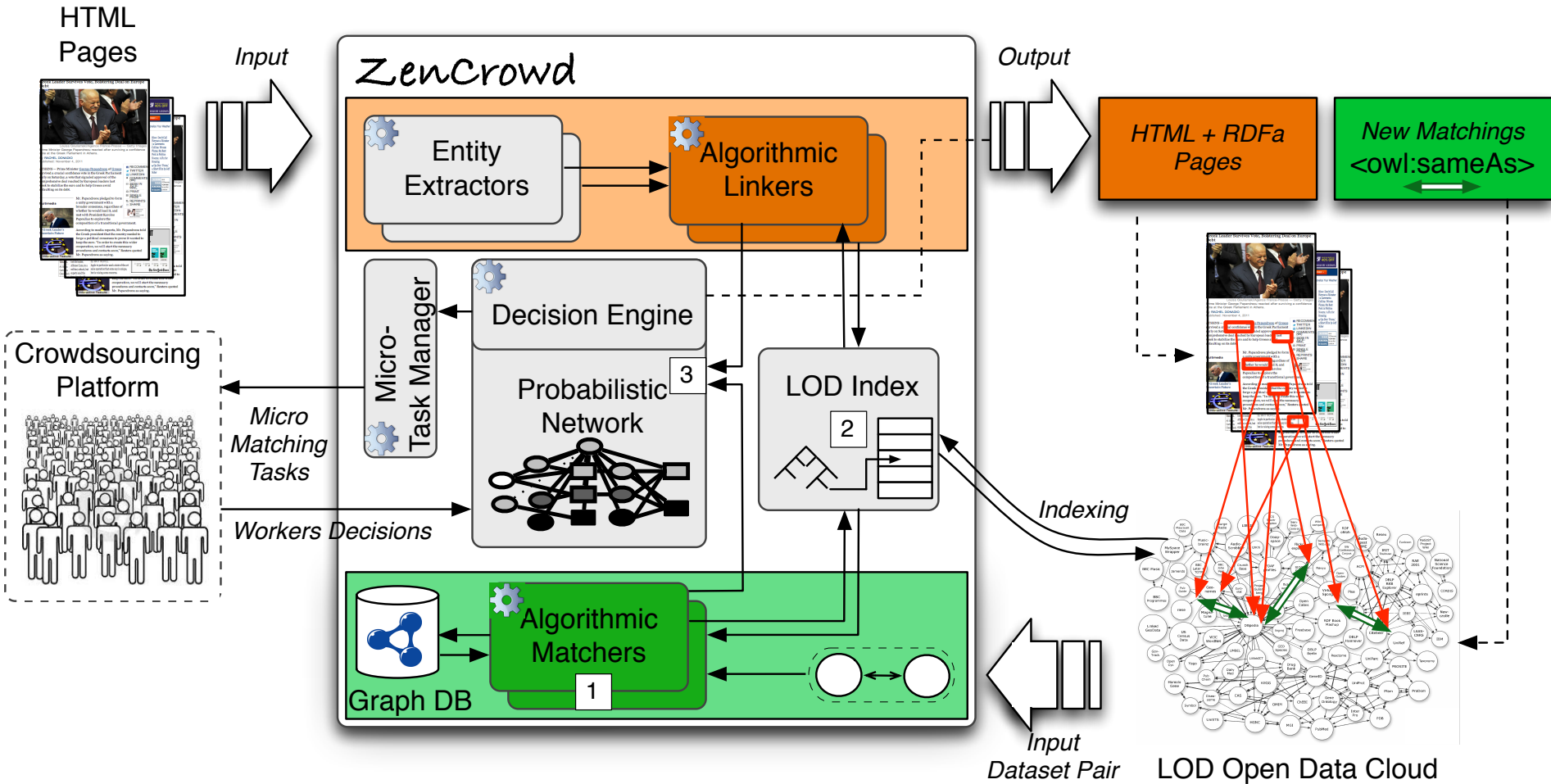
ZenCrowd Summary

- 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

<http://exascale.info/zencrowd/>

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

ZenCrowd Architecture



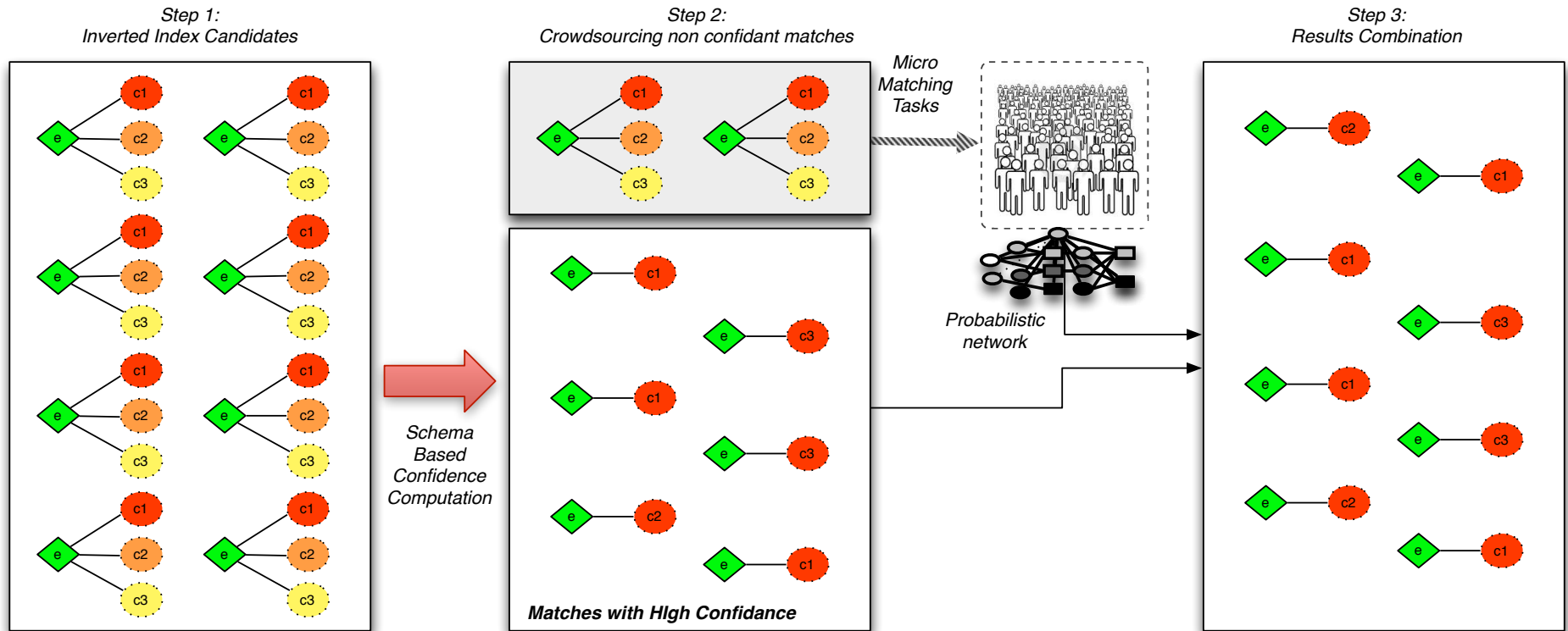
Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. **ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking.** In: 21st International Conference on World Wide Web (WWW 2012)

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: $n \times n$ comparison within the clusters with an expensive measure



3-steps Blocking with the Crowd

- Crowdsourcing as the most expensive similarity measure



CrowdQ – Crowd-powered Query Understanding

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 ▾ 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



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

Videos

Maps

News

Images

More ▾

Search tools

About 28,800,000 results (0.21 seconds)

Marino ran the 2013 election for Mayor of Rome with the support of a centre-left alliance. After leading in the first round he was elected (on 10 June) Mayor of Rome at the second ballot, winning 63.9% of the votes in a run-off against the centre-right candidate, the outgoing mayor **Gianni Alemanno**.

[Ignazio Marino - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Ignazio_Marino Wikipedia ▾

Feedback

[Mayor of Rome - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Mayor_of_Rome ▾ Wikipedia ▾

The **Mayor of Rome** (Italian: **Sindaco di Roma Capitale**) is an elected politician who

birthdate of ignazio marino

birthdate of Ignazio Marino



Web

News

Images

Maps

Shopping

More ▾

Search tools

About 1,260,000 results (0.28 seconds)

March 10, 1955 (age 59 years)

Ignazio Marino, Date of birth



Ignazio

Surgeon

Ignazio Roberto
current Mayor of
Party and held
mayor of Rome

Party: [Democrat](#)



**Gianni
Alemanno**
March 3,
1958



**Nicola
Zingaretti**
October 11,
1965



Matteo Renzi
January 11,
1975

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

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

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](#)

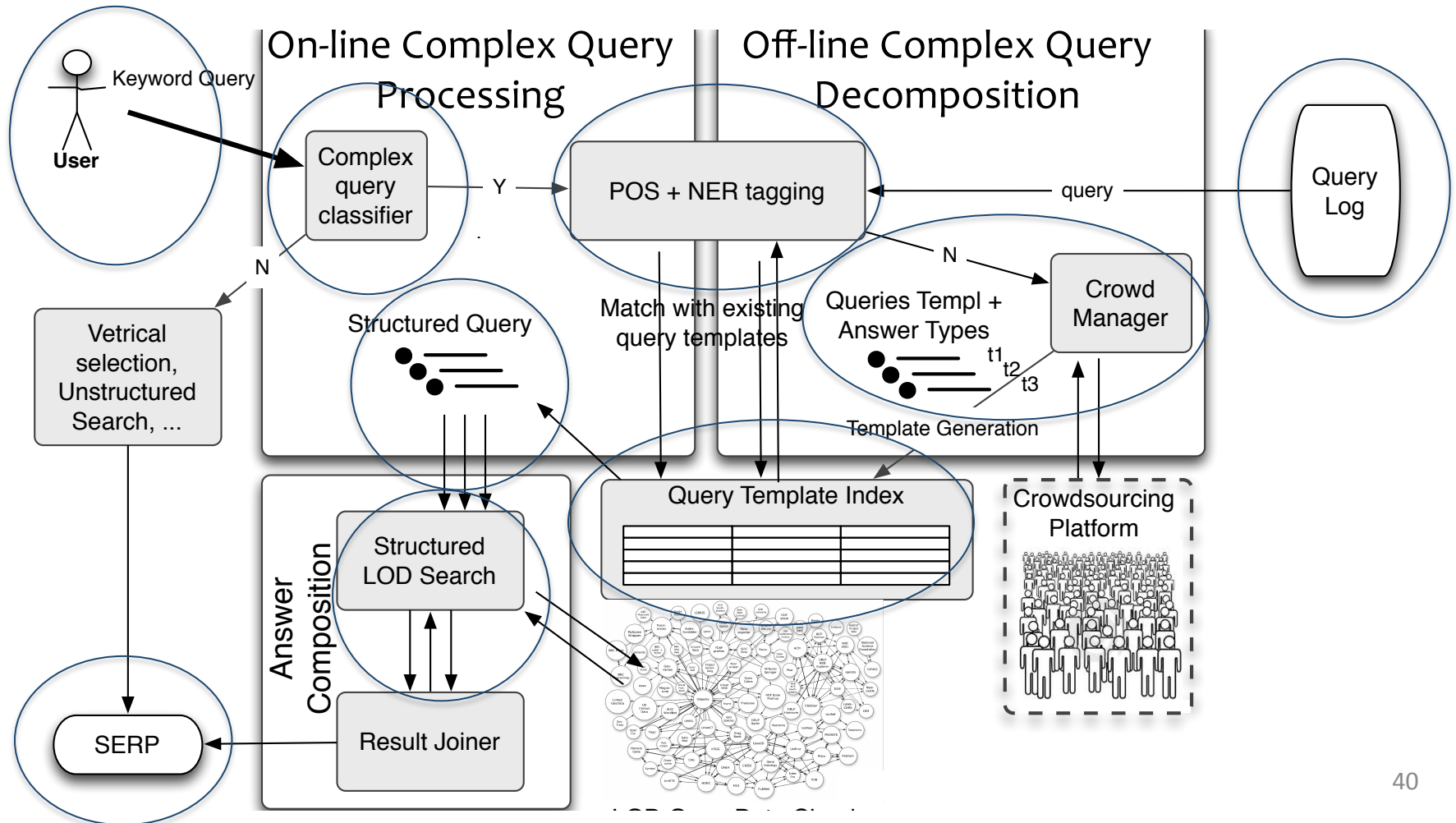
en.comuni-italiani.it/

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

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 constraints
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 ask a human, when trust the machine
- Hybrid (human in the loop)
 - Pre-processing: training data for ML
 - Post-processing: based on confidence scores
 - Mix: active learning