Crowdsourcing for Entity-Centric Information Access

Gianluca Demartini
University of Sheffield, UK
gianlucademartini.net

Entity-Centric Information Access

tom cruise

Web News Images Videos Shopping More Search tools

About 153,000,000 results (0.26 seconds)

In the news

Readers' Poll: The 10 Best Tom Cruise Movies
RollingStone.com - 13 hours ago
Tom Cruise's recent string of bombs coupled with embarrassing revelations about his role in ...

Tom Cruise & Suri: Why He Hasn't Seen Her In Two Years — Report
Hollywood Life - 16 hours ago

Mission Impossible 5 review: Tom Cruise offers us insane fun
Hindustan Times - 3 hours ago

More news for tom cruise

Official Tom Cruise: Edge Of Tomorrow, Movies, Bio, News ...
www.**tomcruise**.com/
OFFICIAL **TOM CRUISE** SITE: View the latest EDGE OF TOMORROW trailer! Watch career movie trailers, videos, and retrospective. Read the **Tom Cruise** ...

Tom Cruise - IMDb
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: Mission: Impossible. If you had told fourteen-year-old Franciscan seminary student Thomas Cruise Mapother IV that one day in the ...

Tom Cruise

Actor · tomcruise.com

Tom Cruise is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Endless Love. Wikipedia

Born: July 3, 1962 (age 53), Syracuse, New York, United States

Height: 1.70 m

Full name: Thomas Cruise Mapother IV

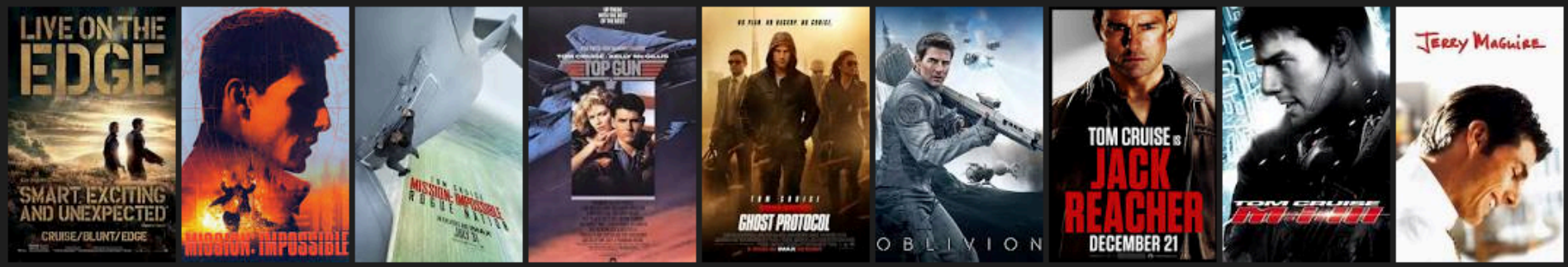
Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)

Children: Suri Cruise, Connor Antony Cruise, Isabella Jane Cruise

tom cruise movies

Web | Videos | Images | Shopping | News | More | Search tools

Tom Cruise > Movies



Edge of Tomorrow	Mission: Impossible	Mission: Impossible – Ro…	Top Gun	Mission: Impossible – Gh…	Oblivion	Jack Reacher	Mission: Impossible III	Jerry Maguire
2014	1996	2015	1986	2011	2013	2012	2006	1996

Tom Cruise filmography - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Tom_Cruise_filmography

Tom Cruise is an American actor and producer who made his film debut with a minor role ... As of 2015, Cruise has reprised his role as Hunt in four more **films** in the ..... "**Movie** Review : Utility Vehicle : '**Days of Thunder**': The NASCAR racing footage and **Tom Cruise's** grin are fine. ... "**The Last Samurai Movie** Review (2003)".

Rock of Ages - Ask the Dust (film) - Losin' It

Tom Cruise - IMDb

www.imdb.com/name/nm0000129/

Tom Cruise and Arnold Schwarzenegger in Terminator Genisys (2015) **Tom Cruise** at ... Visit our IMDb Picks section to see our recommendations of **movies** and TV ..... **movie** preview - · **Mission: Impossible - Ghost Protocol** -- An exclusive clip ...

Tom Cruise

Actor

Tom Cruise is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Endless Love. Wikipedia

Born: July 3, 1962 (age 53), Syracuse, New York, United States

Height: 1.70 m

Full name: Thomas Cruise Mapother IV

- Entity-seeking queries make up 40-50% of the query volume
  - Jeffrey Pound, Peter Mika, Hugo Zaragoza: Ad-hoc object retrieval in the web of data. WWW 2010: 771-780
  - Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, Ariel Fuxman: Active objects: actions for entity-centric search. WWW 2012: 589-598
- Show a summary of the most likely information-needs
  - Including related entities for navigation
  - *Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, Nicolas Torzec: Entity Recommendations in Web Search. ISWC 2013*

Matthew Paige "Matt" Damon is an American actor, voice actor, screenwriter, producer, and philanthropist whose career was launched following the success of the drama film Good Will Hunting (1997) from a screenplay... wikipedia.org

Born: October 8, 1970 (age 43), Cambridge, Massachusetts, USA

Height: 5' 10" (1.78m)

Spouse: Luciana Barroso (m. 2005-present)

Partner: Winona Ryder (1998-2000)

Parents: Kent Damon, Nancy Carlsson-Paige

Children: Isabella Damon, Alexia Barroso, Gia Zavala Damon, Stella Damon

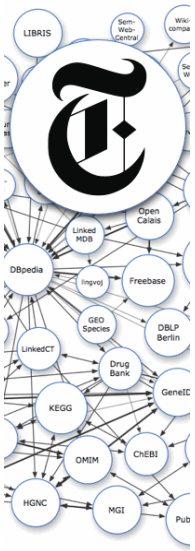
Movies & TV Shows

The Zero Theorem | Elysium | The Departed | We Bought a Zoo | Good Will Hunting

Feedback

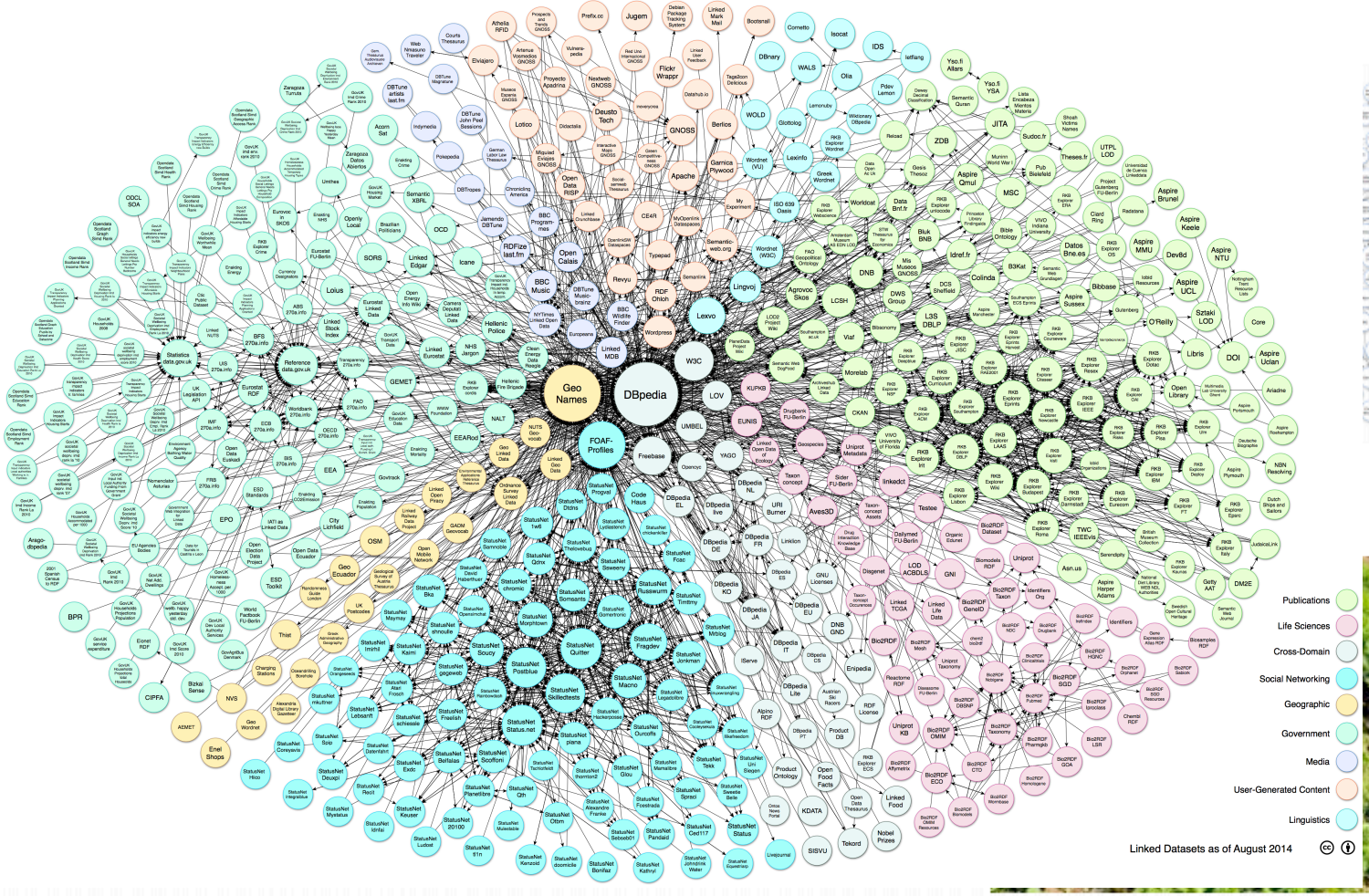
Web of Data

- Freebase
  - Acquired by Google in July 2010.
  - Knowledge Graph launched in May 2012.
  - Read-only in December 2014 -> WikiData
- Schema.org
  - Driven by major search engine companies
  - Machine-readable annotations of Web pages
- Linked Open Data
  - 31 billion triples, Sept 2011
  - 90 billion triples, Aug 2015 (stats.lod2.eu)



Linked Open Data

2007



Z. Kaoudi and I. Manolescu, ICDE seminar 2013

Today I will talk about

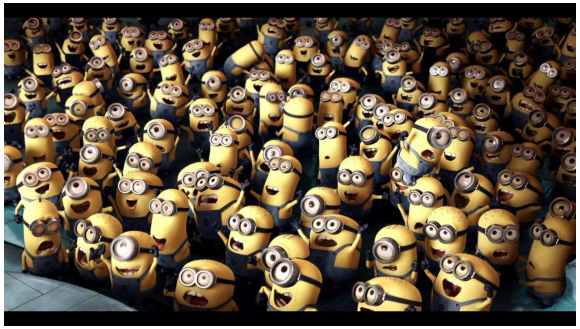
- Crowdsourcing
  - Amazon MTurk as a crowdsourcing platform
- Entity Linking on the Web
  - With the crowd
- Finding the best type for an entity appearing in Web pages

Overview

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 my tutorial at ESWC 2013 and ISWC 2013

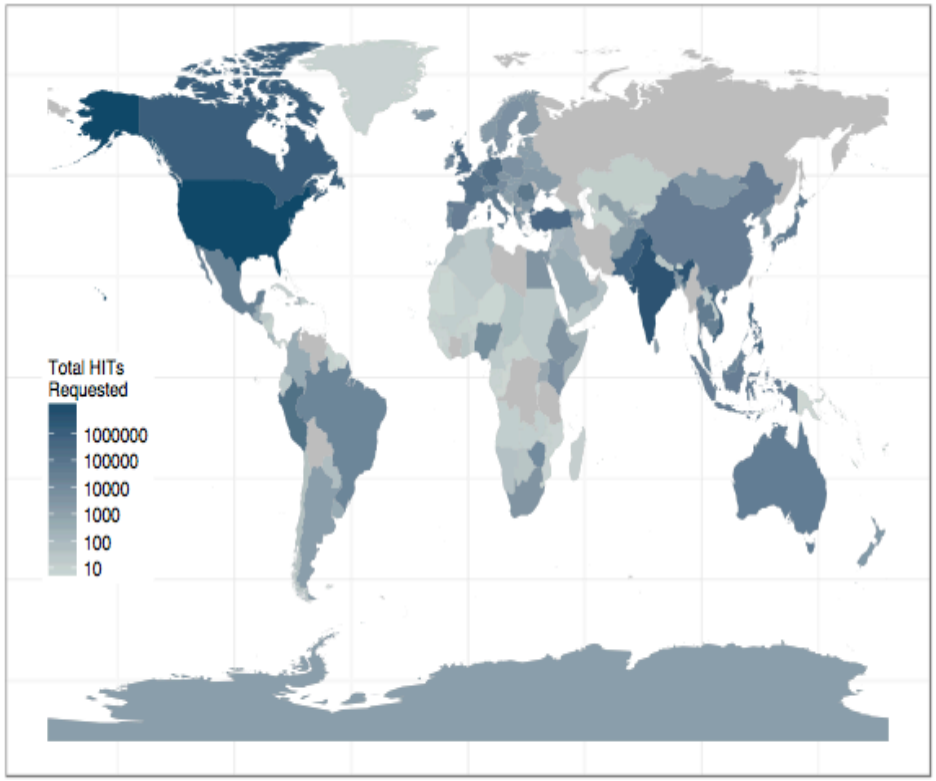
mechanical turk
Artificial Artificial Intelligence
amazon
beta

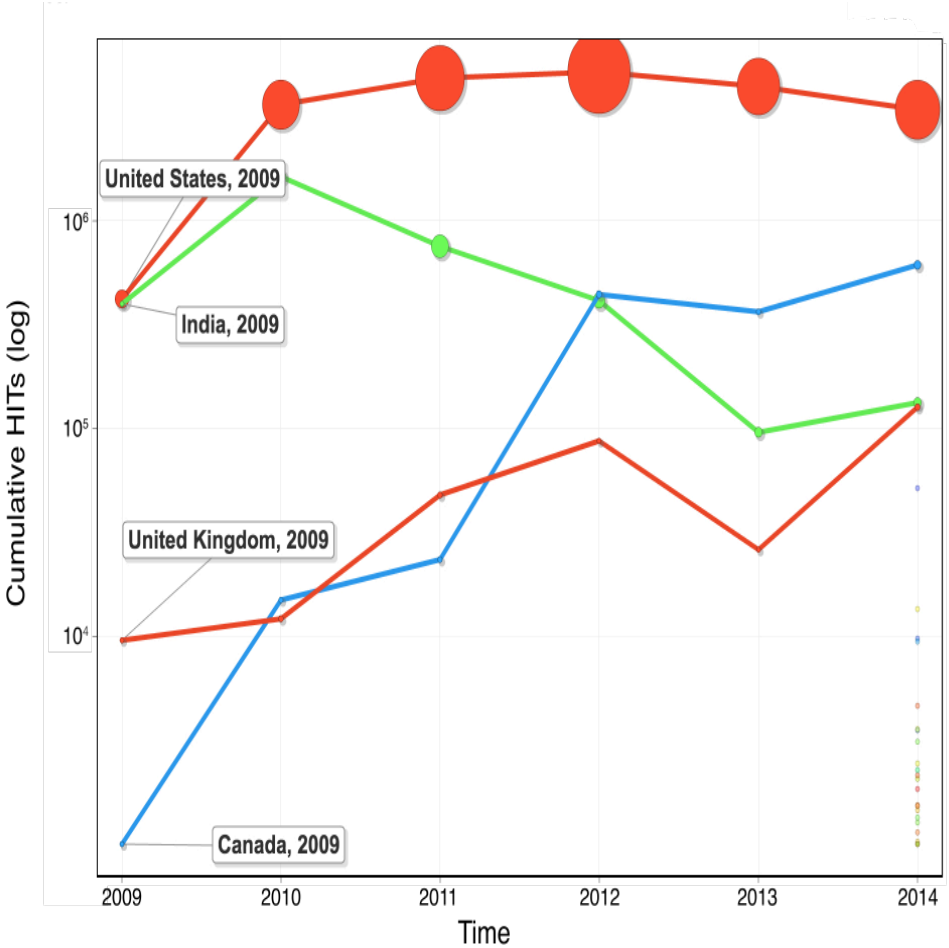
5-year Analysis of MTurk workload

- 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.

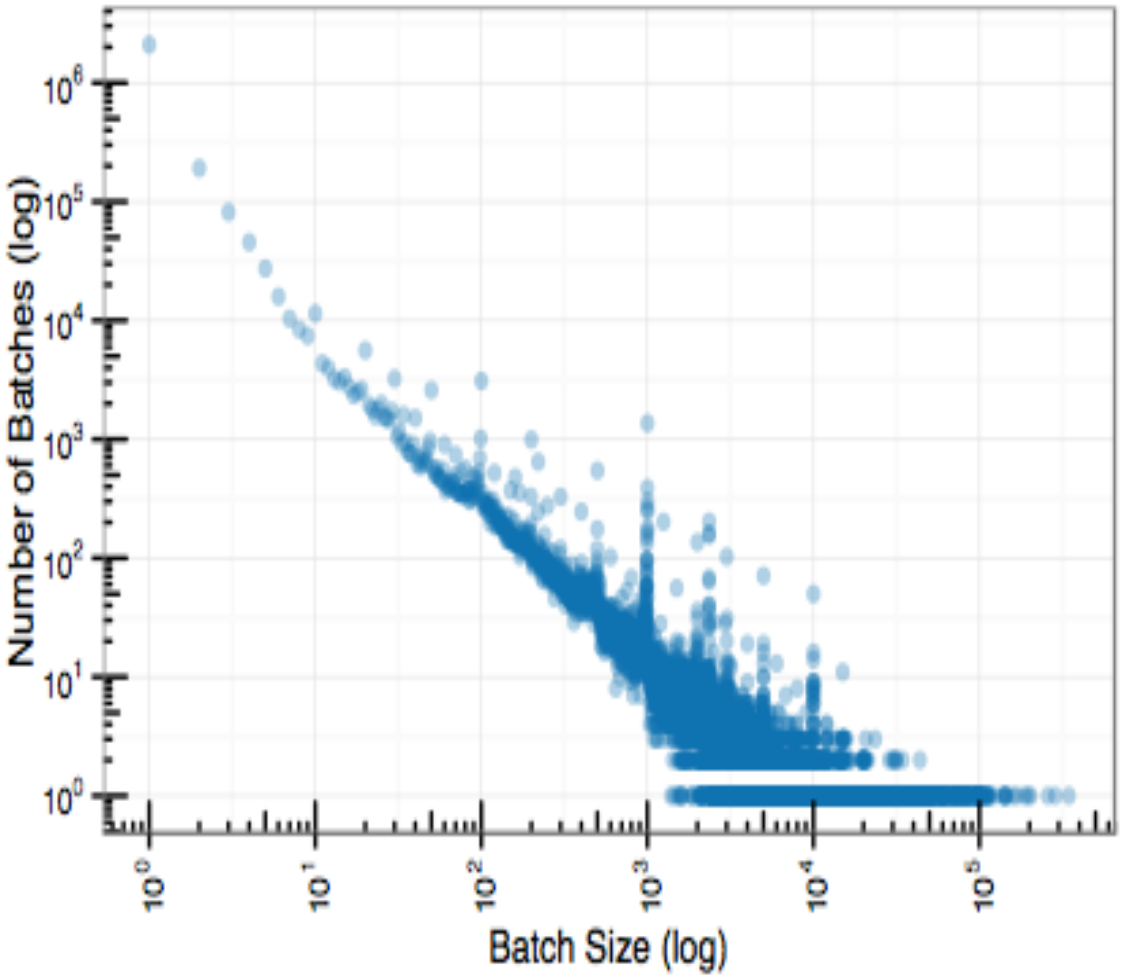
Country-Specific HITs





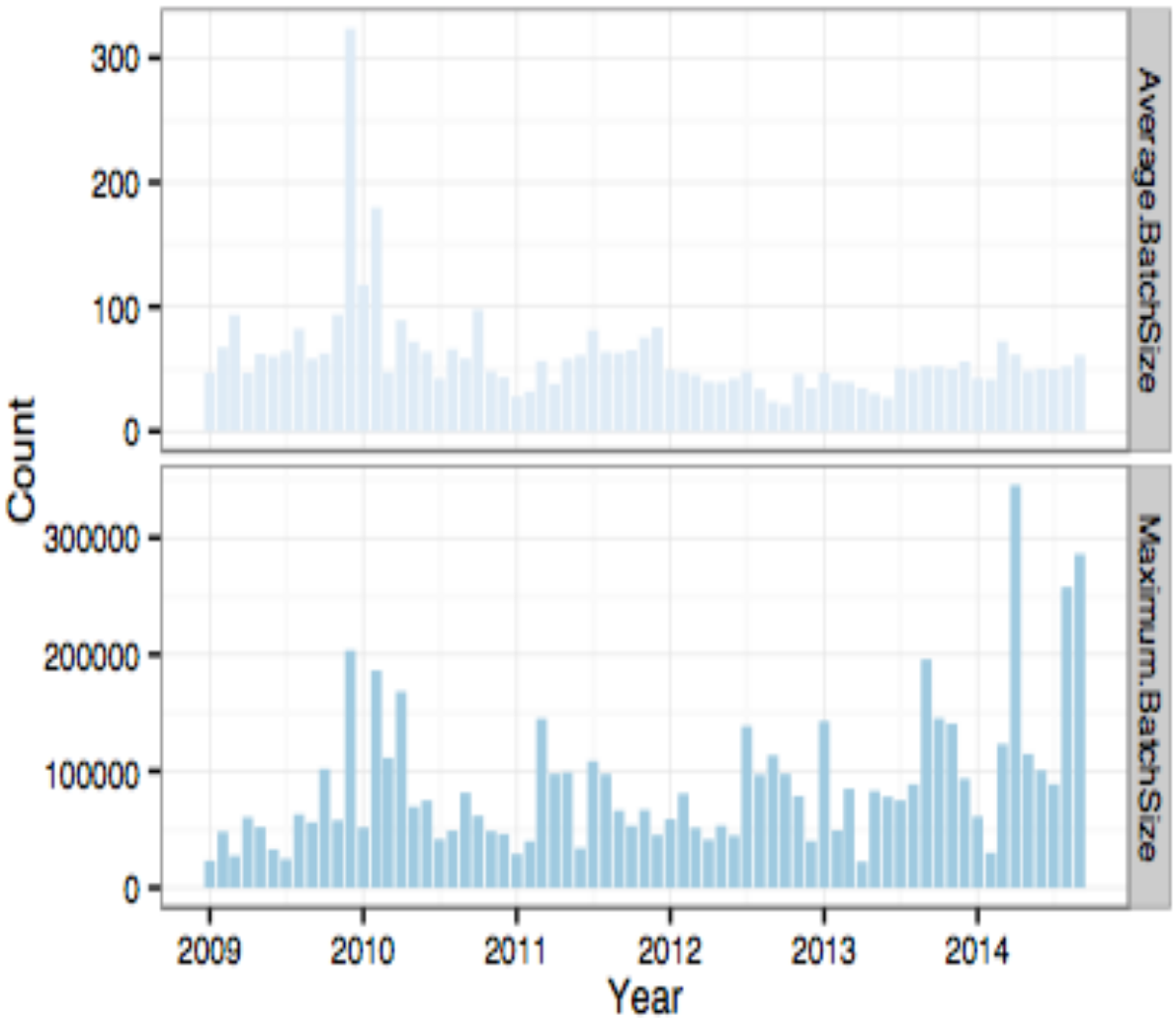
Workers from US, India and Canada are the most sought after.

Distribution of *Batch Size*



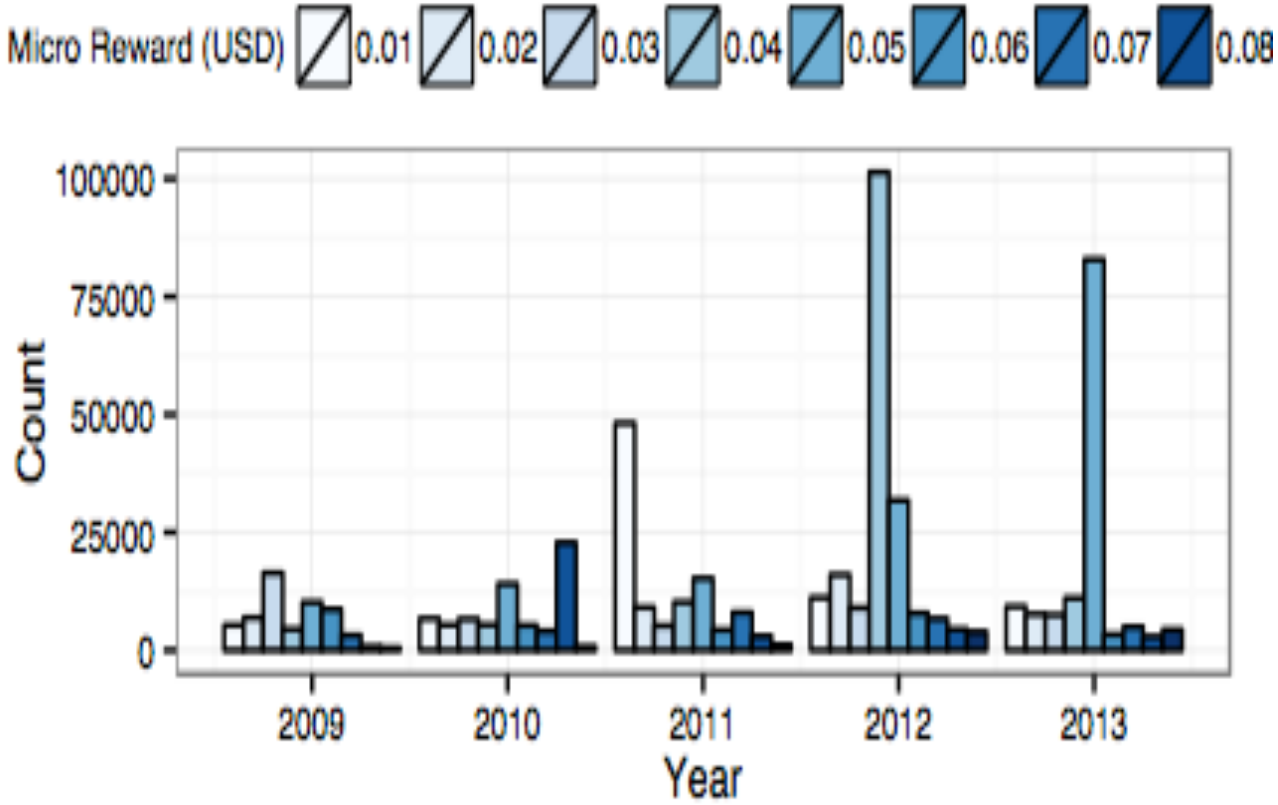
"Power-law"

Evolution of Batch Sizes



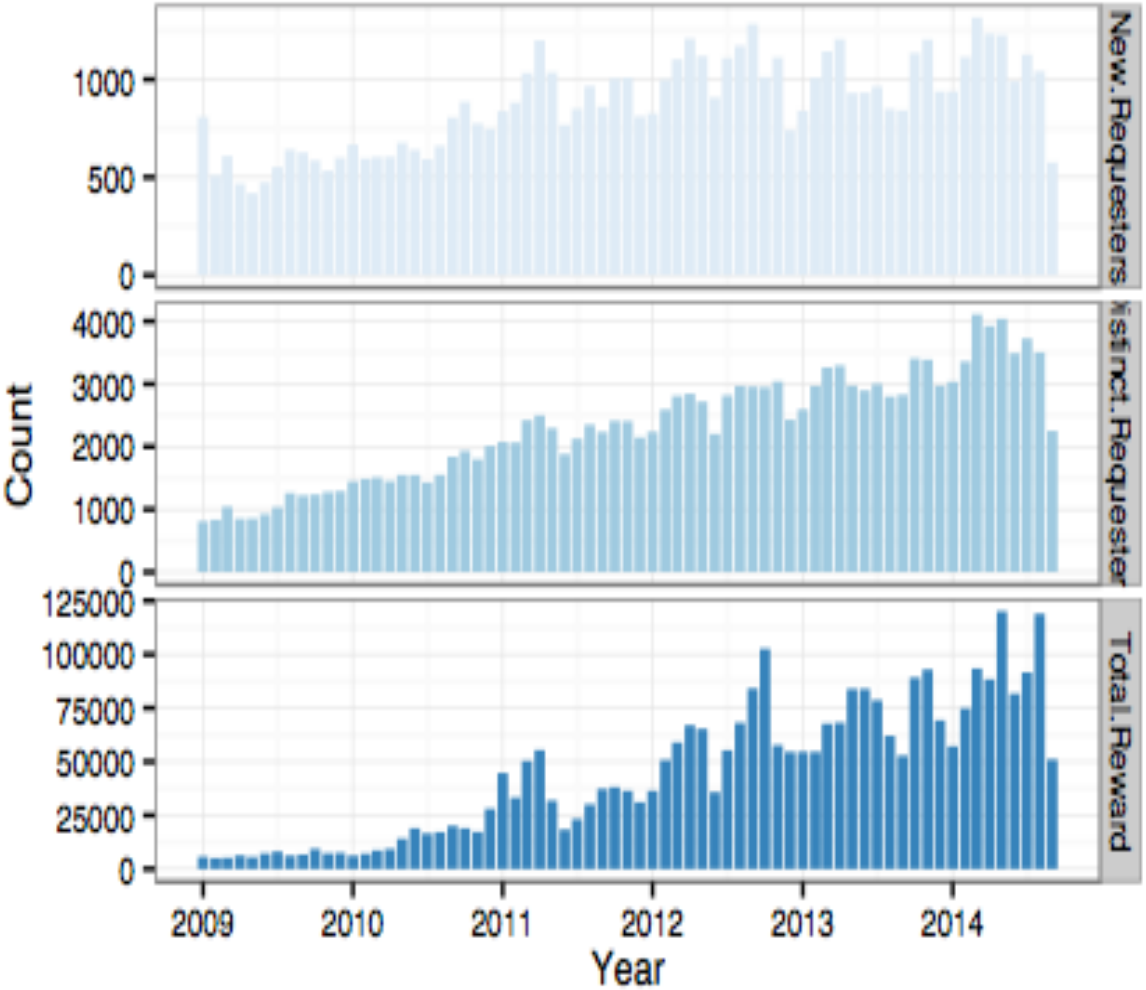
Very large batches start to appear

HIT Pricing



5-cents is the new 1-cent

Requesters and Reward Evolution



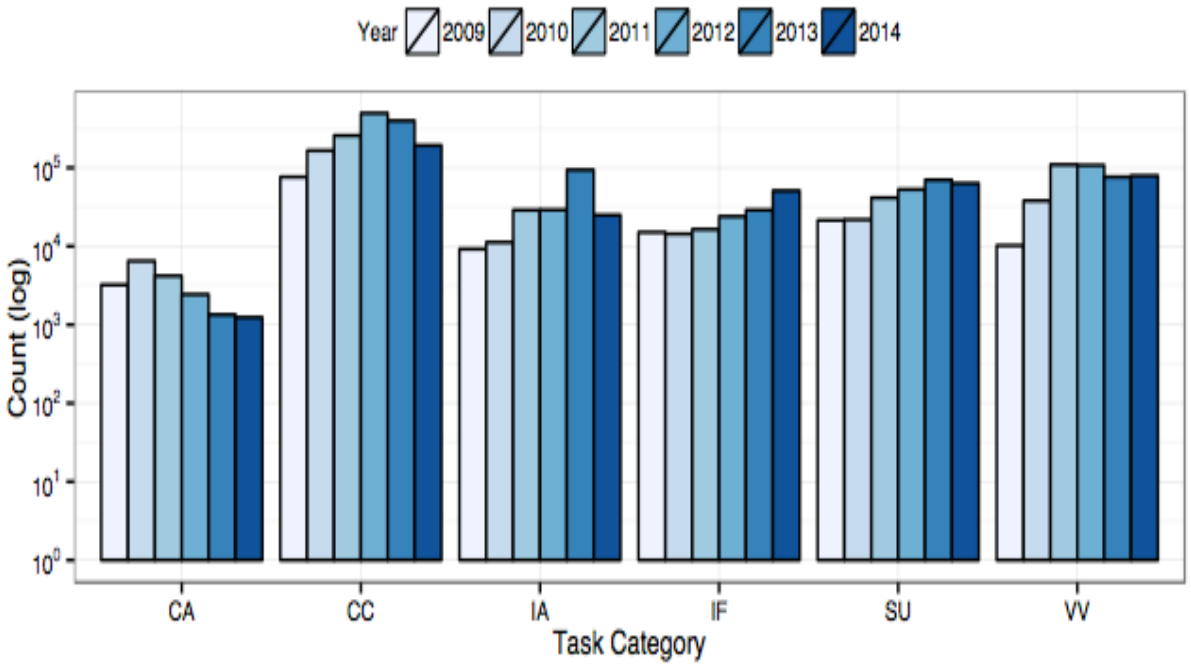
Increasing number of New and Distinct Requesters

Top MTurk Requesters last week

Top-1000 Requesters, report for August 4, 2015 to September 3, 2015

Requester name	hits	reward
Speechpad	41092	$261,239.88
WorkFusion	2668	$11,552.00
Jon Brelig	115360	$6,000.16
CastingWords	11037	$5,953.96
Mark Yatskar	45008	$3,364.97
VidAngel	181	$2,881.54
p9r	41754	$2,502.03
Amazon Requester Inc - browse classification	35227	$2,113.62
University of California, Berkeley	152	$1,976.00

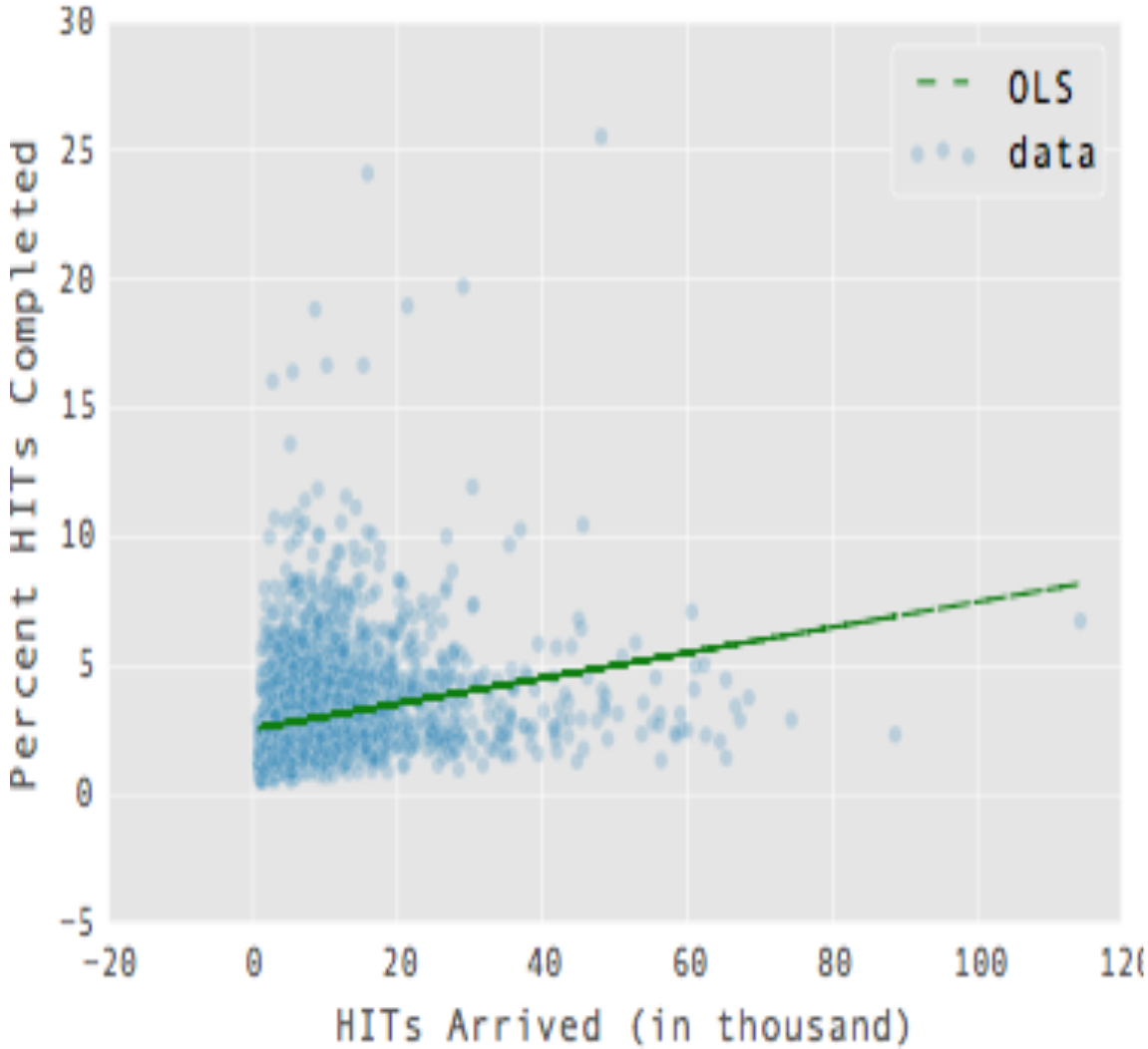
Distribution of HIT Types



Less Content Access batches

Content Creation being the most popular

Supply Elasticity



Intercept = 2.5
Slope = 0.5%

20% of new work gets completed within an hour

Summary

- HIT reward has increased over time
- **Audio transcription** is the most popular task
- Demand for Indian workers has decreased
- **Surveys** are most popular for US workers
- 1000 new requesters per month join
- 10K new HITs arrive and 7.5K HITs get completed every hour

- Check #mturkdynamics for the main findings

Crowdsourcing for Entity Linking

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

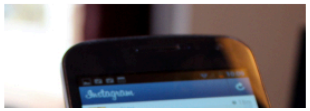
Instagram for Android is now available

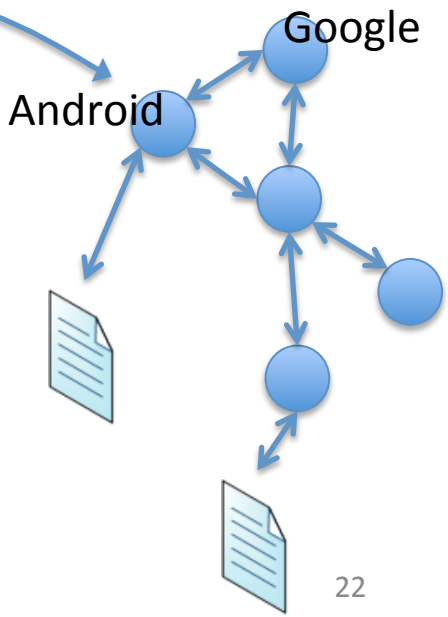
At long last, Instagram finally releases the Android version of its app.

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

Follow

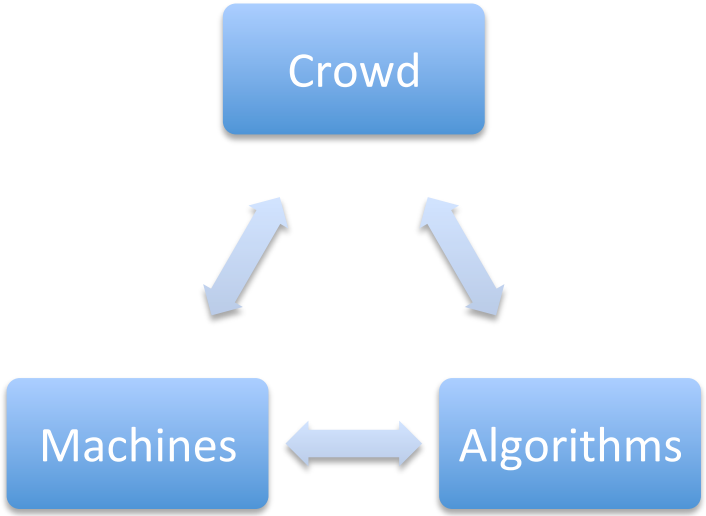
Instagram has been around since 2010, available only to iOS devices. Android users have been waiting patiently, with repeated promises of an Android version arriving soon.



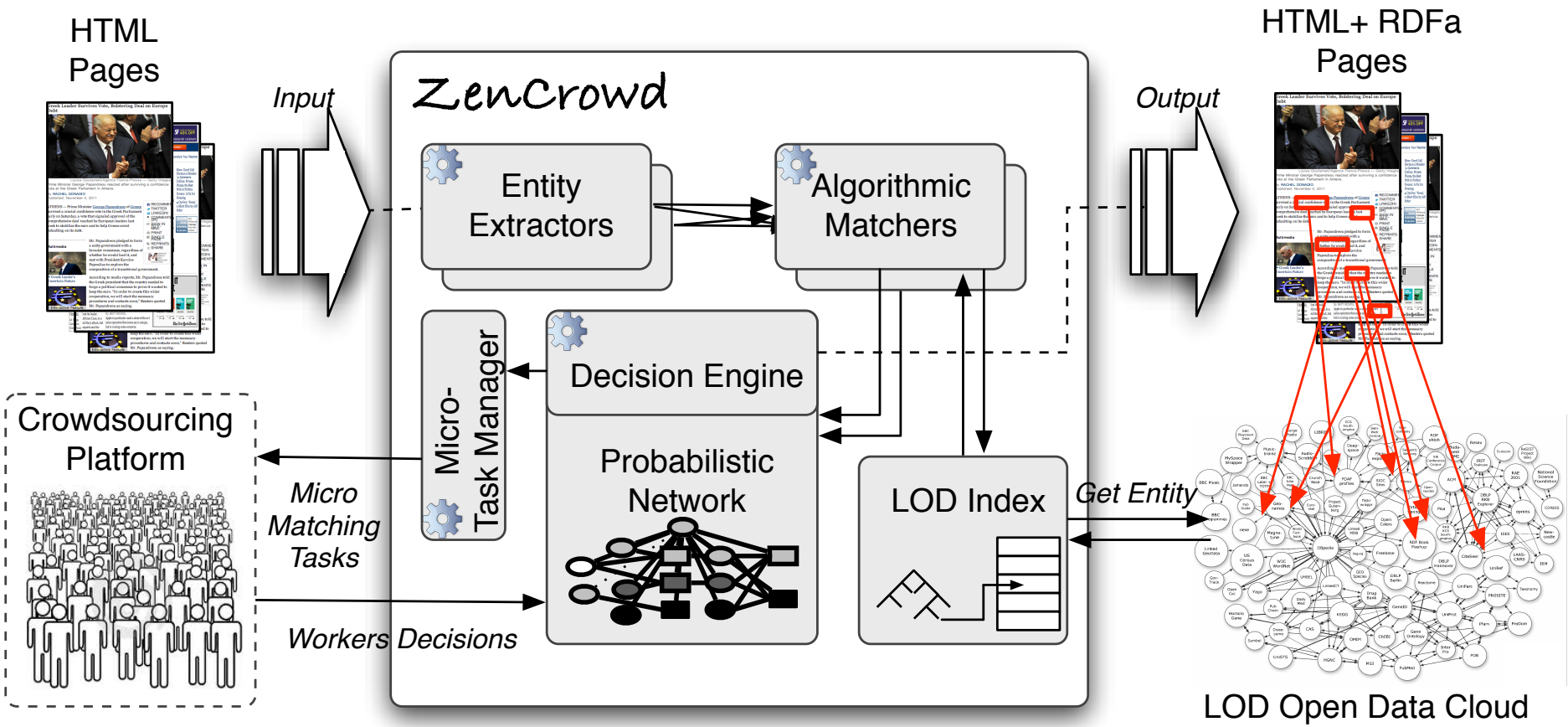


ZenCrowd

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



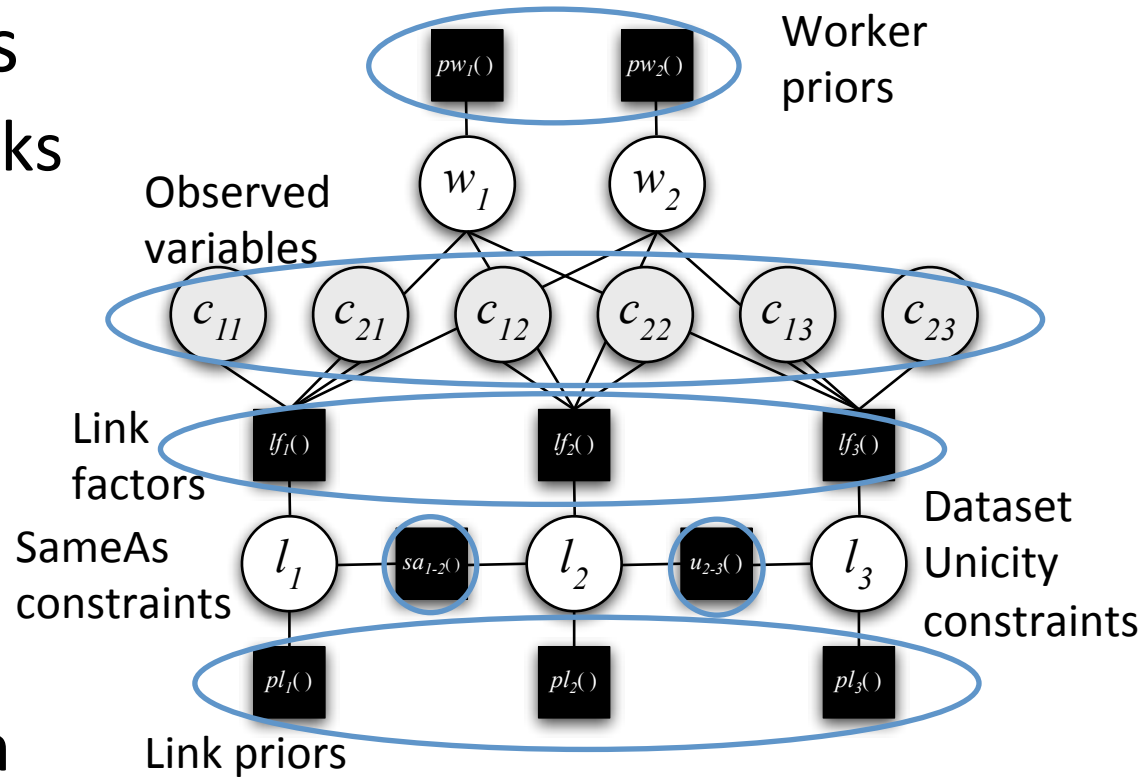
ZenCrowd Architecture



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



2 workers, 6 clicks, 3 candidate links

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

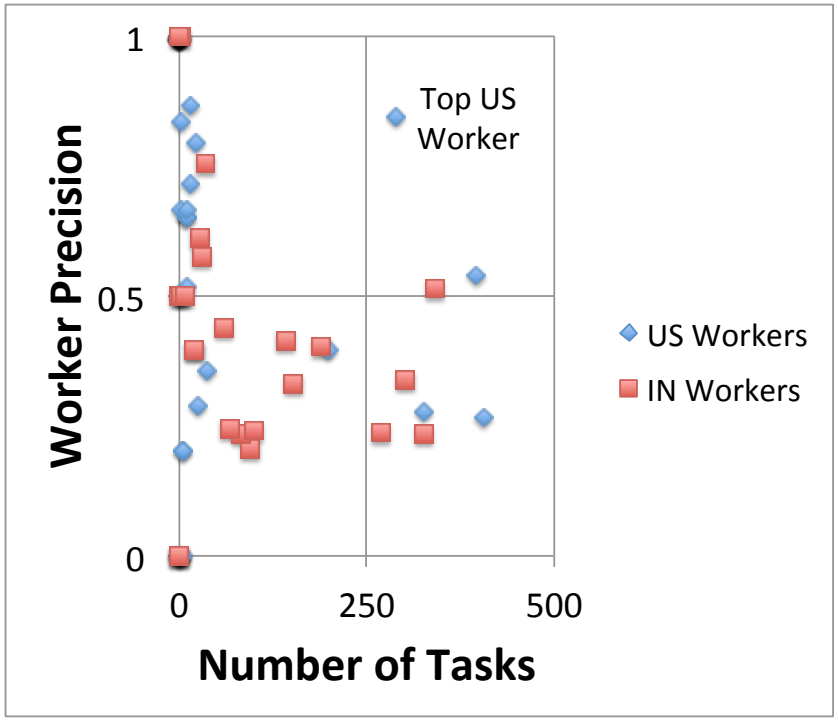
Experimental Evaluation

- 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

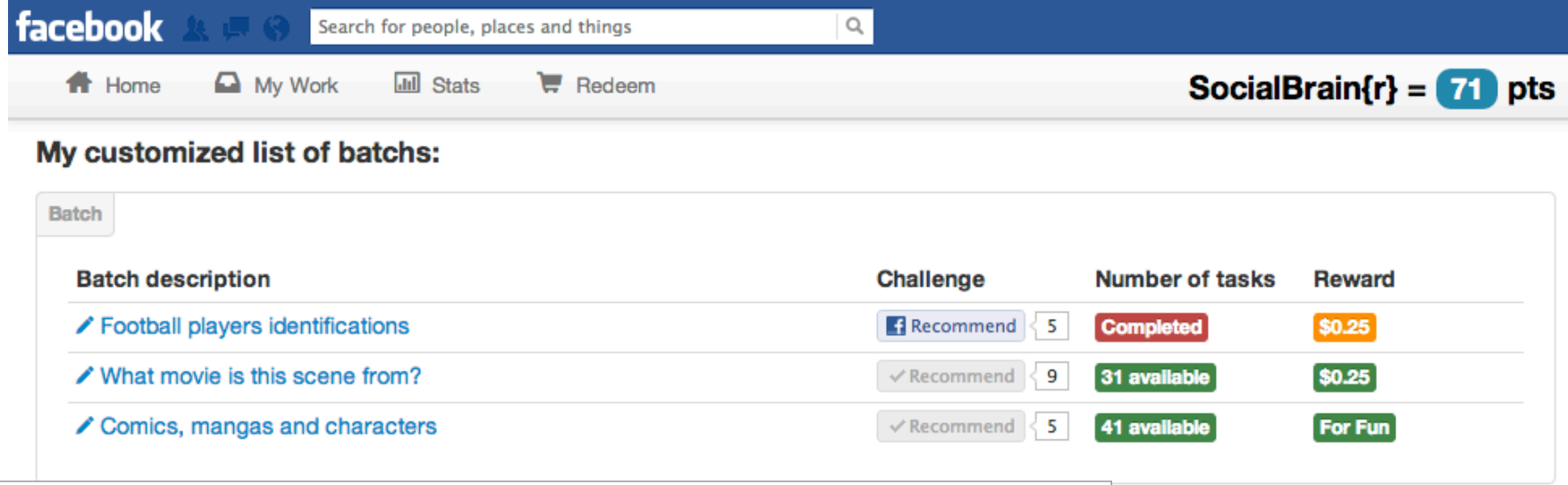
Experimental Evaluation

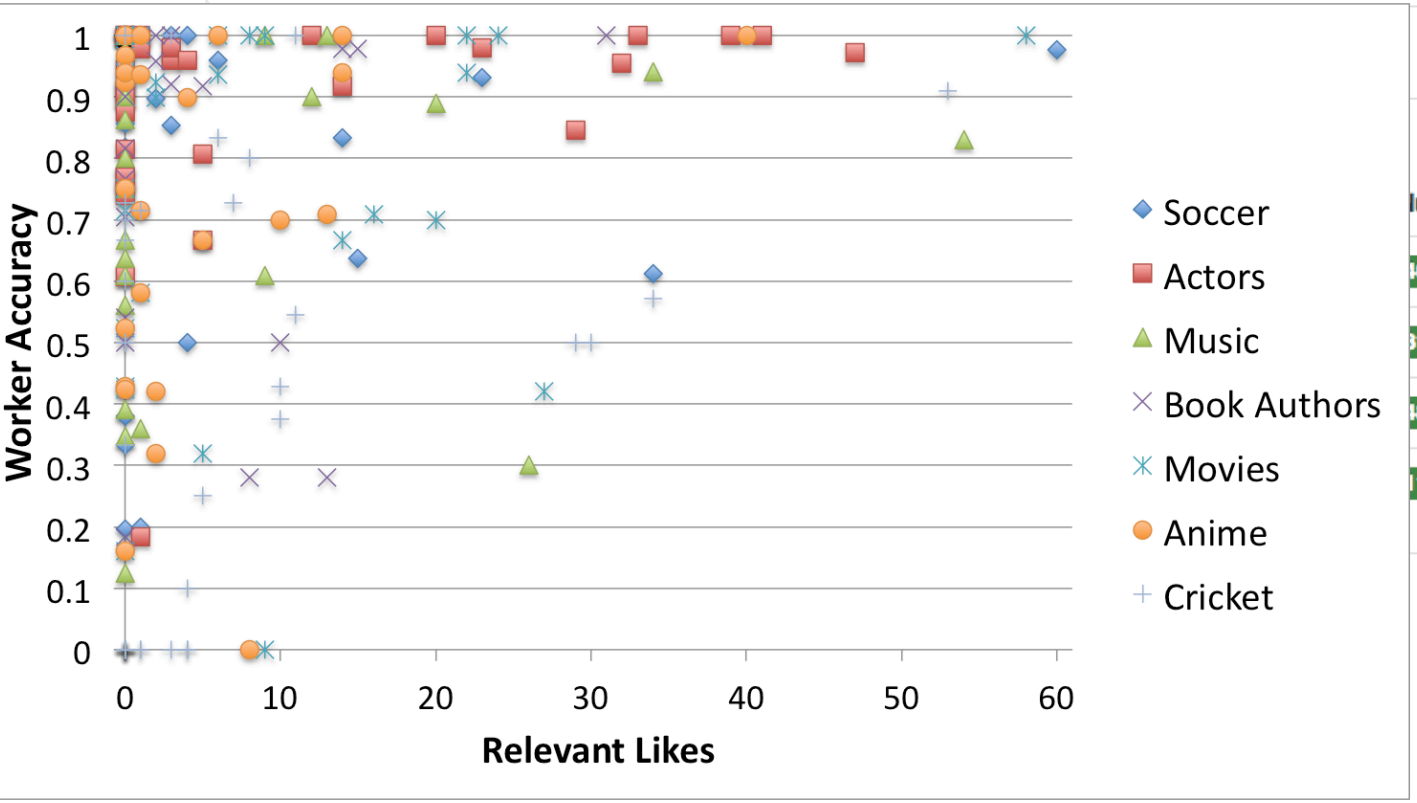
- Worker Selection



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 WWW2013 and CHI2015 papers)





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).

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, 2013):
  - Also used for instance matching across datasets
  - 3-way blocking with the crowd

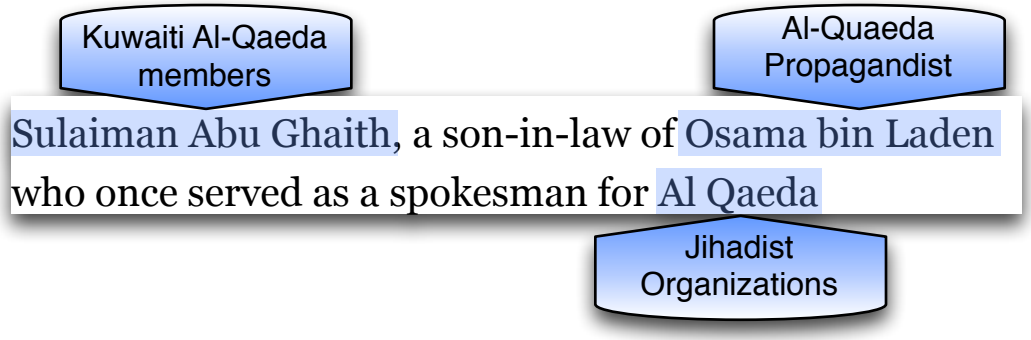
Entity Types

…and Why Types?

- "Summarization" of texts

Article Title	Entities	Types
Bin Laden Relative Pleads Not Guilty in Terrorism Case	Osama Bin Laden Abu Ghaith Lewis Kaplan Manhattan	Al-QaedaPropagandists Kuwaiti Al-Qaeda members Judge Borough (New York City)

- Contextual entities summaries in Web-pages

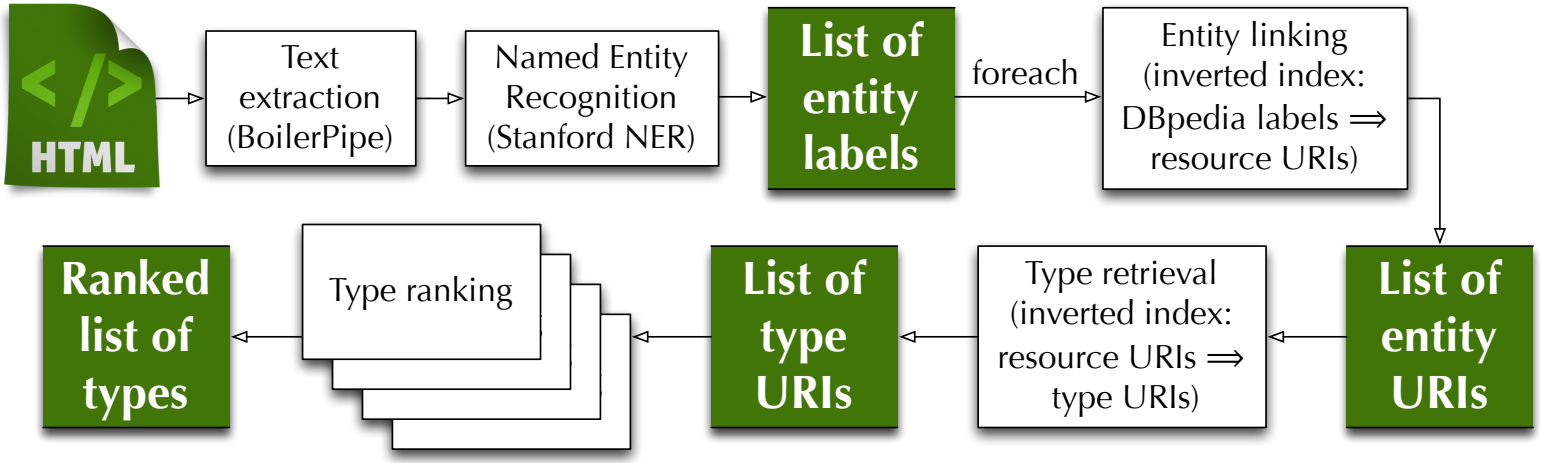


- Disambiguation of other entities

- Diversification of search results

Entities May Have Many Types

TRank Pipeline



Alberto Tonon, Michele Catasta, Gianluca Demartini, Philippe Cudré-Mauroux, and Karl Aberer. TRank: Ranking Entity Types Using the Web of Data. In: The 12th International Semantic Web Conference (**ISWC 2013**). Sydney, Australia, October 2013. (nominated for best paper award)

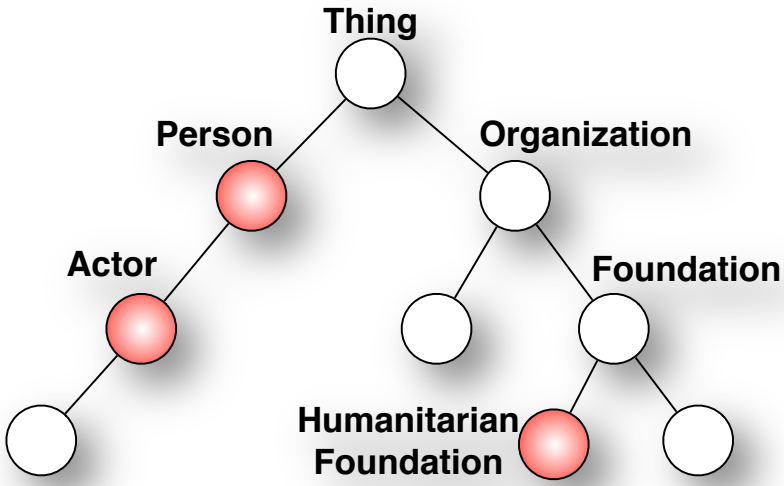
Ranking Algorithms

- **Entity centric**
- **Hierarchy-based**
- **Context-aware (featuring type-hierarchy)**
- **Learning to Rank**

Hierarchy-Based Approaches (An Example)

- **ANCESTORS**

Score(e, t) = number of *t*'s ancestors in the type hierarchy contained in T_e.

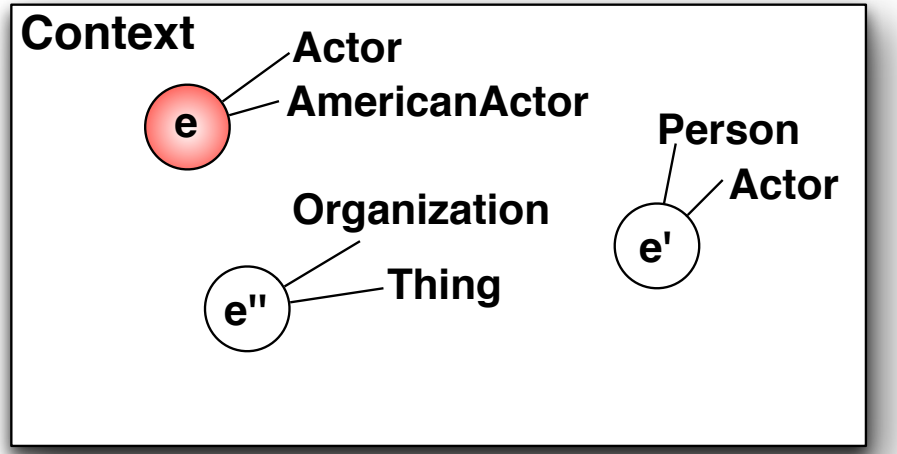


T_e often doesn't contain all super types of a specific type

Context-Aware Ranking Approaches (An Example)

- **SAMETYPE**

Score(e, t, c_T*)* = number of times *t* appears among the types of every other entity in c_T.



Learning to Rank Entity Types

Determine an optimal combination of all our approaches:

- Decision trees
- Linear regression models
- 10-fold cross validation

Datasets

- 128 recent NYTimes articles split to create:
  - *Entity Collection*
  - *Sentence Collection*
  - *Paragraph Collection*
  - *3-Paragraphs Collection*
- Ground-truth obtained by using crowdsourcing
  - *3 workers per entity/context*
  - *4 levels of relevance for each type*
  - *Overall cost: 190$*

Effectiveness Evaluation

	Entity-only		Sentence		Paragraph		3-Paragraphs	
Approach	**NDCG**	**MAP**	**NDCG**	**MAP**	**NDCG**	**MAP**	**NDCG**	**MAP**
FREQ	0.6284	0.4659	0.5409	0.3758	0.5315	0.3739	0.5250	0.3577
WIKILINK-OUT	0.6874	0.5406	0.6050	0.4521	0.6063	0.4550	0.6059	0.4444
WIKILINK-IN	0.6832	0.5342	0.5907	0.4213	0.5879	0.4254	0.5853	0.4143
SAMEAS	0.6848	0.5328	0.6049	0.4310	0.5990	0.4221	0.6172	0.4417
LABEL	0.6672	0.5067	0.6075	0.4265	0.5883	0.4104	0.5821	0.4034
SAMETYPE	-	-	0.6024	0.4452	0.5917	0.4327	0.5813	0.4256
PATH	-	-	0.6507	0.4956	0.6538	0.4974	0.6315	0.4742
DEPTH	0.7432	0.6128	0.6754	0.5385	0.6797	0.5475	0.6741	0.5354
ANCESTORS	0.7424	0.6154	**0.6967**†	**0.5637**†	**0.6949**†	**0.5662**†	**0.6879**†	**0.5562**†
ANC_DEPTH	**0.7469**†	**0.6236**†	0.6832	0.5488	0.6885	0.5546	0.6796	0.5423
DEC-TREE	**0.7614**	**0.6361**	**0.7373***	**0.6079***	**0.7979***	**0.7019***	**0.7943***	**0.6914***
LIN-REG	0.7373	0.6079	0.6906	0.5579	0.6987	0.5702	0.6899	0.5529

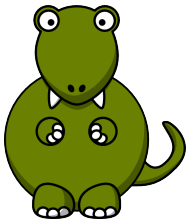
Use TRank:

Open Source (Scala)

https://github.com/MEM0R1ES/TRank

Web Service (JSON)

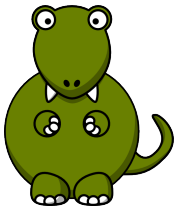
http://trank.exascale.info



Efficiency Evaluation

- Tested efficiency on a CommonCrawl sample of 1TB
  - 1,310,459 HTML pages
  - 23GB compressed
- Map/Reduce on a cluster of 8 machines with 12 cores, 32GB of RAM and 3 SATA disks
- On average, 25 min. processing time (> 100 docs/node x sec)

Text Extraction	NER	Entity Linking	Type Retrieval	Type Ranking
18.9%	35.6%	29.5%	9.8%	6.2%



Conclusions

- New task: ranking entity types.
  - Useful for: “summarization” of Web-documents, entity summaries, disambiguation.
- Various approaches: entity-centric, context-aware, hierarchy-based, learning to rank.
  - Hierarchy-based and learning to rank are the most effective.
- Hadoop, Map/Reduce, and inverted indices to achieve scalability.

Summary

- Hybrid human-machine systems can
  - Scale over large amounts of data
  - Reach high accuracy by keeping humans in the loop
- Entities are the new entry point to Web content
  - "Things not string"
  - Google Knowledge Graph (but also Bing, Yahoo!, Yandex)
- Users can benefit from entity-centric search, browsing, and exploration of the Web