

The role of humans in crowdsourced semantics

Elena Simperl, University of Southampton*

WIC@WWW2014

*with contributions by Maribel Acosta, KIT

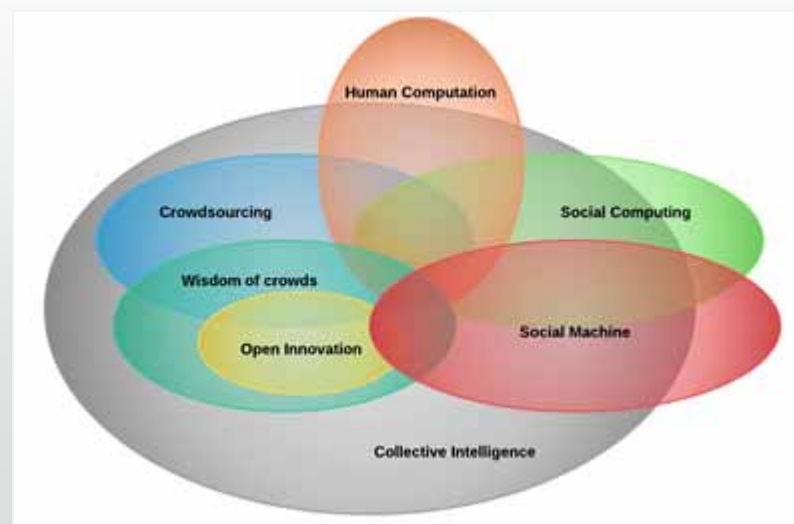
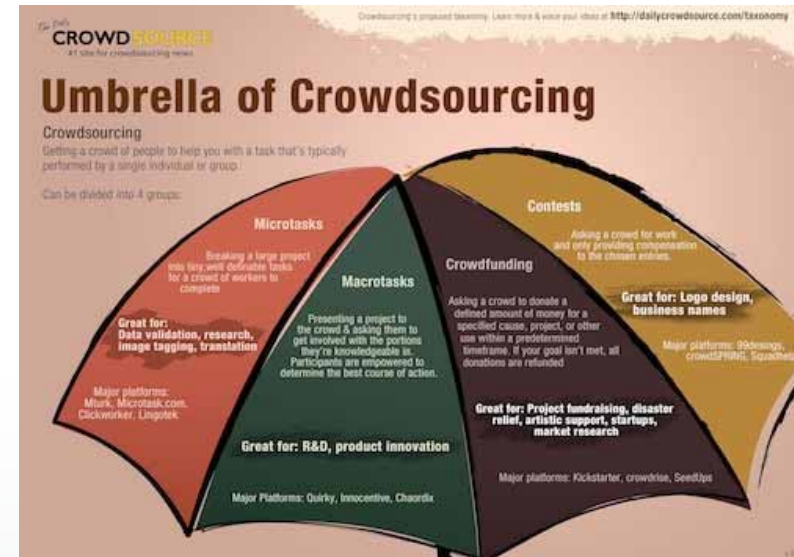
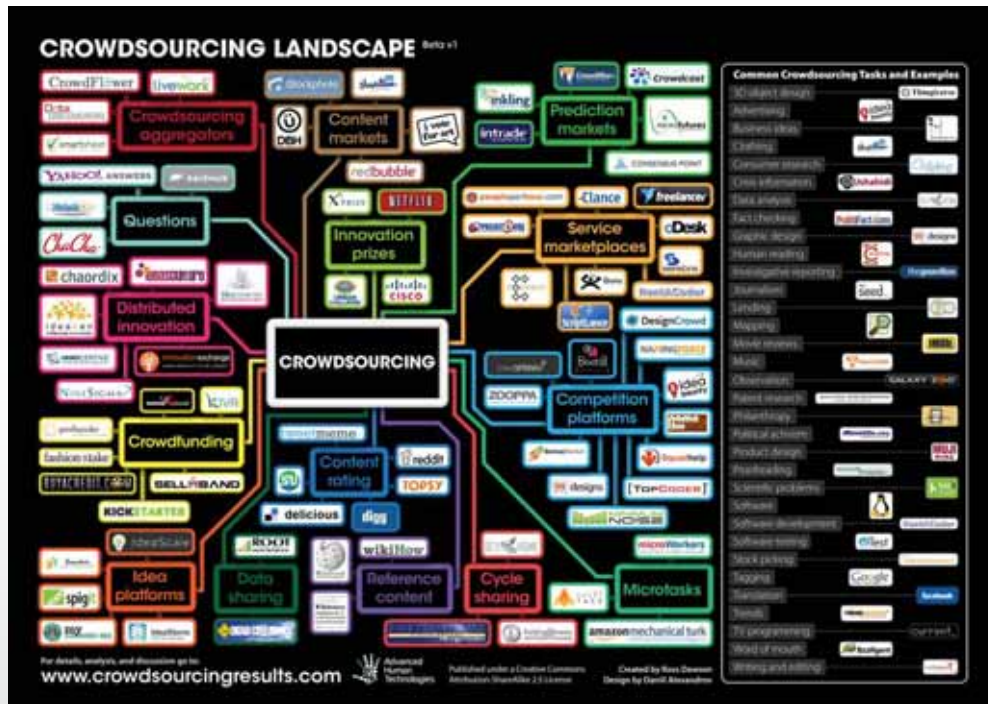
07 April 2014

Crowdsourcing Web semantics: the great challenge

- Crowdsourcing is increasingly used to augment the results of algorithms solving Semantic Web problems
- Research questions
 - Which form of crowdsourcing for what task?
 - How to design the crowdsourcing exercise?
 - How to combine different human- and machine-driven approaches?



There is crowdsourcing and crowsourcing...



Microtask crowdsourcing

Work is broken down into smaller (,micro') pieces that can be solved independently

Two promotional banners. The left one for Servio says "Need to get work done?" with a "Get Started" button. The right one for CloudCrowd says "Want to earn money now?" with a "Learn More" button.

A screenshot of the CrowdSource website. It features a header with navigation links and a main section titled "500,000+ Hyper-Specialized Workers On Demand". Below this, there are sections for "What is crowdsourcing?" and "How can CrowdSource help you?" with various icons and text.

A diagram illustrating the microtask workflow. It shows three steps: "you send samasource a project", "samasource breaks it down into microwork", and "work is allocated to our service partners".

A screenshot of the Microtask website. It features a navigation bar with "HOME", "COMPANY", "SOLUTIONS", and "BLOG". The main content area has a large green banner that says "One Billion" and a "Follow us" section with social media icons.

An advertisement for Mechanical Turk. It says "Make Money by working on HITs" and "Get Results from Mechanical Turk". It lists benefits for workers: "Can work from home", "Choose your own work hours", and "Get paid for doing good work".

A screenshot of the CrowdFlower website. It features a navigation bar and a large headline: "The World's Largest Workforce". Below this, it says "Instantly hire millions of people to collect, filter, and enhance your data." There are several service cards, including "RTFM Real Time Fato Moderator" and "Sentiment Analysis".

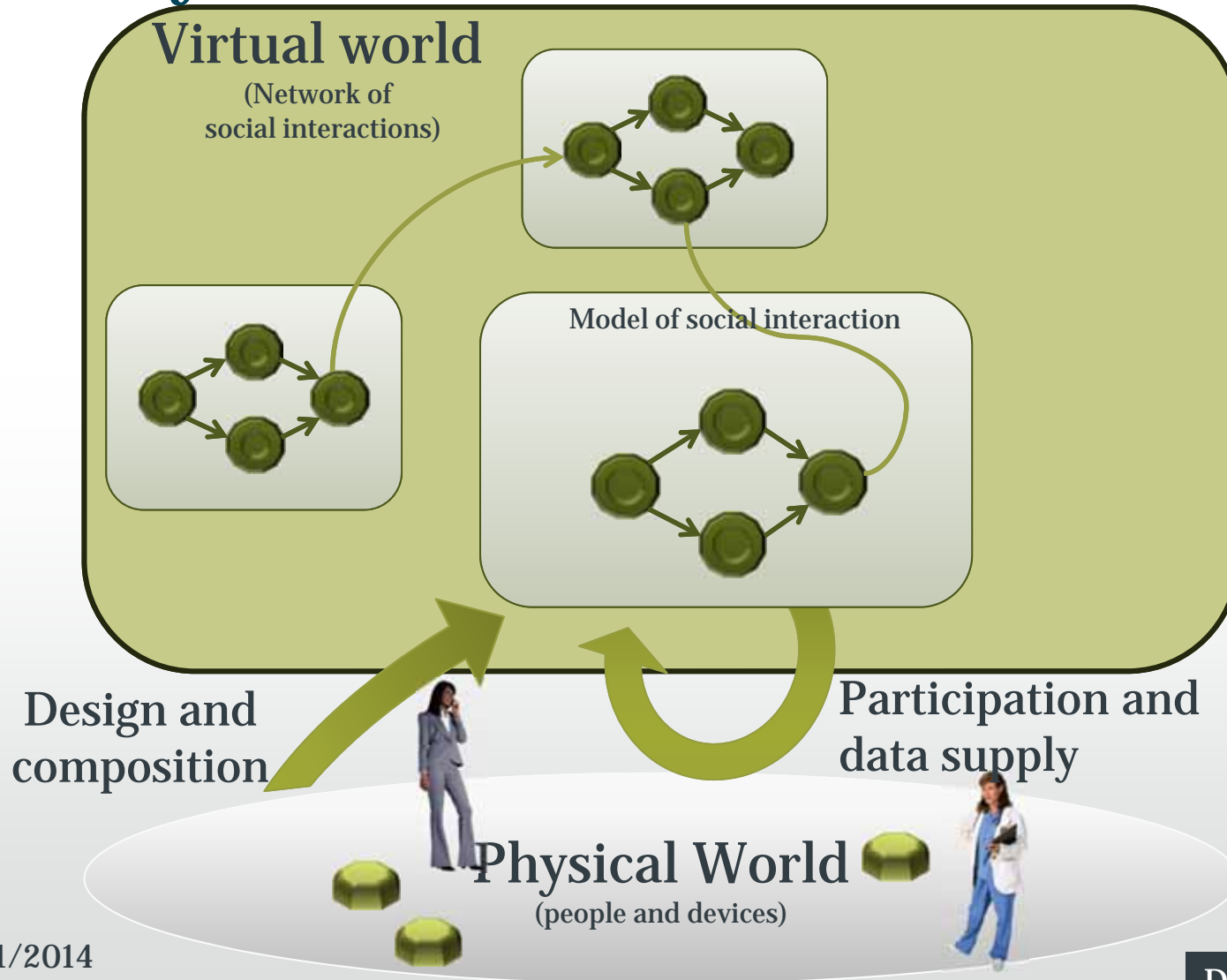
A screenshot of the MobileWorks website. It features a navigation bar and a headline: "Crowdsourcing On Demand. Faster. More accurate. Painless." Below this, there are several icons representing different services and a "Learn more" button.

An advertisement for Freebase. It shows three glass jars containing different colored powders (blue, brown, green). A diagonal banner says "Freebase".

The CloudCrowd logo, which consists of a stylized blue and white circular graphic followed by the text "CloudCrowd" and the tagline "We're working on it. Lots of us."

7/21/2014

Hybrid systems (or ,social machines‘)



4/21/2014

Dave Robertson

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

Yes

No

Not sure

Example: Hybrid query processing

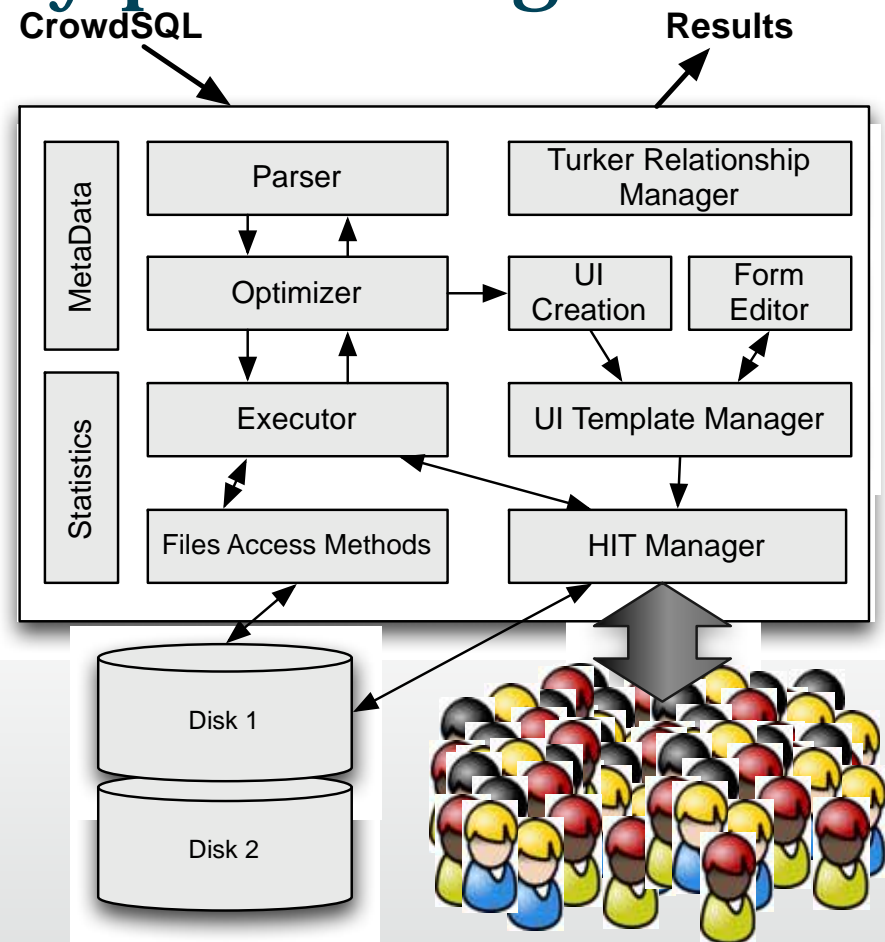
Use the crowd to answer
DB-hard queries

Where to use the crowd:

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

But not:

- Anything the computer already does well



Crowdsourcing Linked Data Quality Assessment

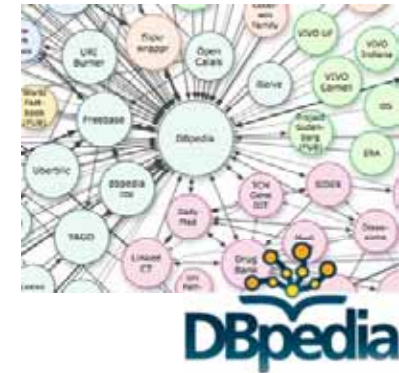
M Acosta, A Zaveri, E Simperl, D Kontokostas, S Auer, J Lehmann
The Semantic Web–ISWC 2013, 260-276

CROWDSOURCING LINKED DATA CURATION

Tasks to be crowdsourced

- **Incorrect object**

- Example: `dbpedia:Dave_Dobbyn dbprop:dateOfBirth "3"`.



- **Incorrect data type or language tags**

- Example: `dbpedia:Torishima_Izu_Islands foaf:name "鳥島"@en`.

- **Incorrect link to “external Web pages”**

- Example: `dbpedia:John-Two-Hawks dbpedia-owl:wikiPageExternalLink <http://cedarlakedvd.com/>`

Combination of approaches



Find

Contest

LD Experts

Difficult task

Final prize



TripleCheckMate
[Kontoskostas2013]



Verify

Microtasks

Workers

Easy task

Micropayments

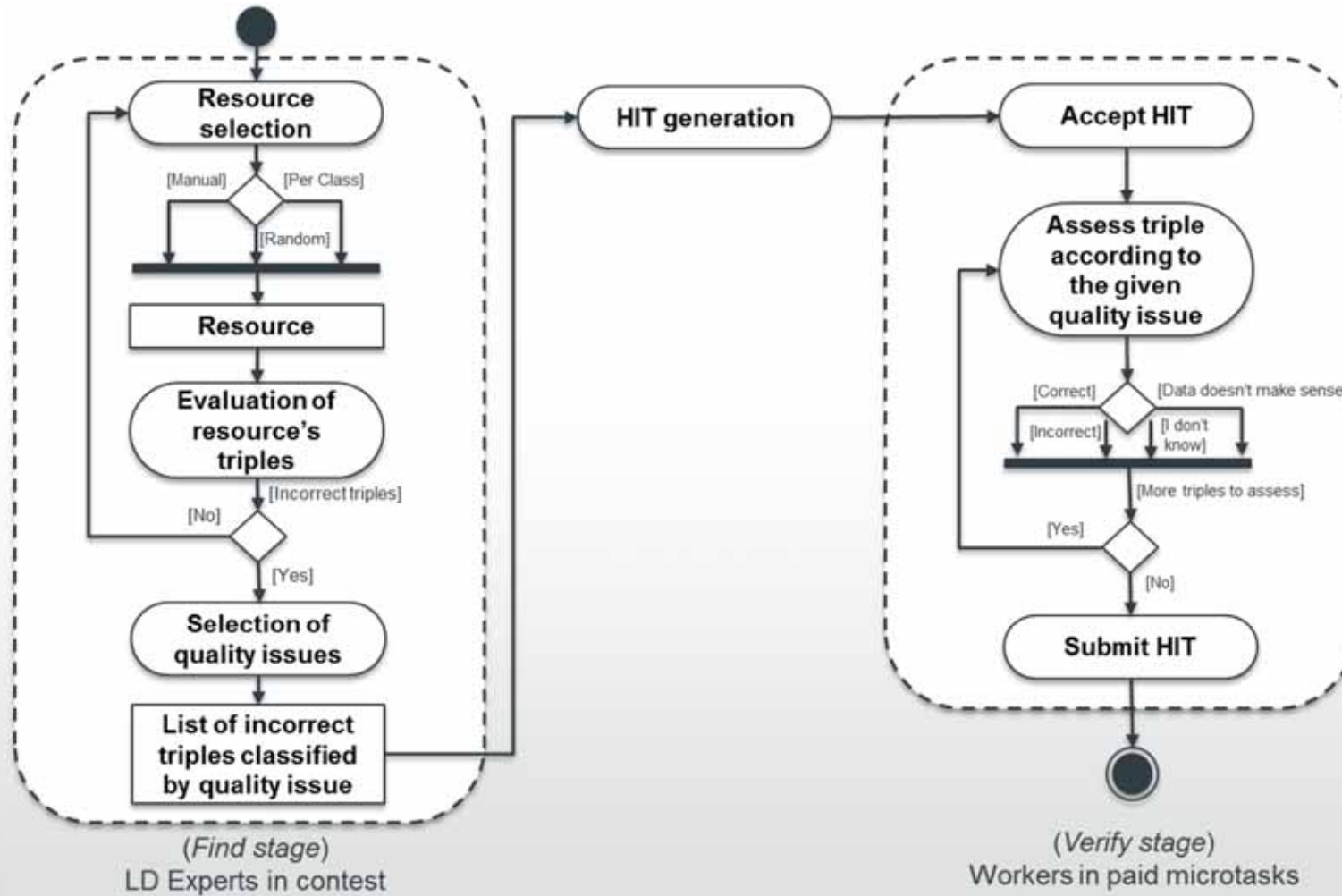


MTurk

<http://mturk.com>

Adapted from [Bernstein2010]

Workflow



Microtask design

- Selection of `foaf:name` or `rdfs:label` to extract human-readable descriptions
- Values extracted automatically from **Wikipedia infoboxes**
- Link to the Wikipedia article via `foaf:isPrimaryTopicOf`
- Preview of external pages by implementing HTML `iframe`

Incorrect object

"Dave Dobbyn"



Date of birth: 3 January 1957 3

This block illustrates an incorrect object. It shows a snippet from a DBpedia infobox for "Dave Dobbyn". The "Date of birth" property is highlighted with a blue background, and the value "3" is also highlighted, indicating an error in the data type or language tag.

Incorrect data type or language tag

About:

Kyoto University

Given the property "name", is the value "京都大?" of type "english"?

This block illustrates an incorrect data type or language tag. It shows a snippet from a DBpedia infobox for "Kyoto University". The "name" property is highlighted with a blue background, and the value "京都大?" is also highlighted, indicating an error in the data type or language tag.

Incorrect outlink

About:



John Two-Hawks

External page: <http://www.cedarlakedvd.com>





This block illustrates an incorrect outlink. It shows a snippet from a DBpedia infobox for "John Two-Hawks". The "External page" property is highlighted with a blue background, and the value "http://www.cedarlakedvd.com" is also highlighted, indicating an error in the outlink. Below the text is a video player showing a cedar lake nature scene.

Experiments

- **Crowdsourcing approaches:** 
 - *Find* stage: Contest with LD experts
 - *Verify* stage: Microtasks (5 assignments)
- **Creation of a gold standard:** 
 - Two of the authors of this paper (MA, AZ) generated the gold standard for all the triples obtained from the contest
 - Each author independently evaluated the triples
 - Conflicts were resolved via mutual agreement
- **Metric: precision**
$$p = \frac{TP}{TP+FP}$$

Overall results

	LD Experts		Microtask workers	
Number of distinct participants	50		80	
Total time	3 weeks (predefined)		4 days	
Total triples evaluated	1,512		1,073	
Total cost	~ US\$ 400 (predefined)		~ US\$ 43	

Precision results: Incorrect object task

- MTurk workers can be used to reduce the error rates of LD experts for the *Find* stage

Triples compared	LD Experts	MTurk (majority voting: n=5)
509	0.7151	0.8977

- 117 DBpedia triples had **predicates related to dates** with incorrect/incomplete values:

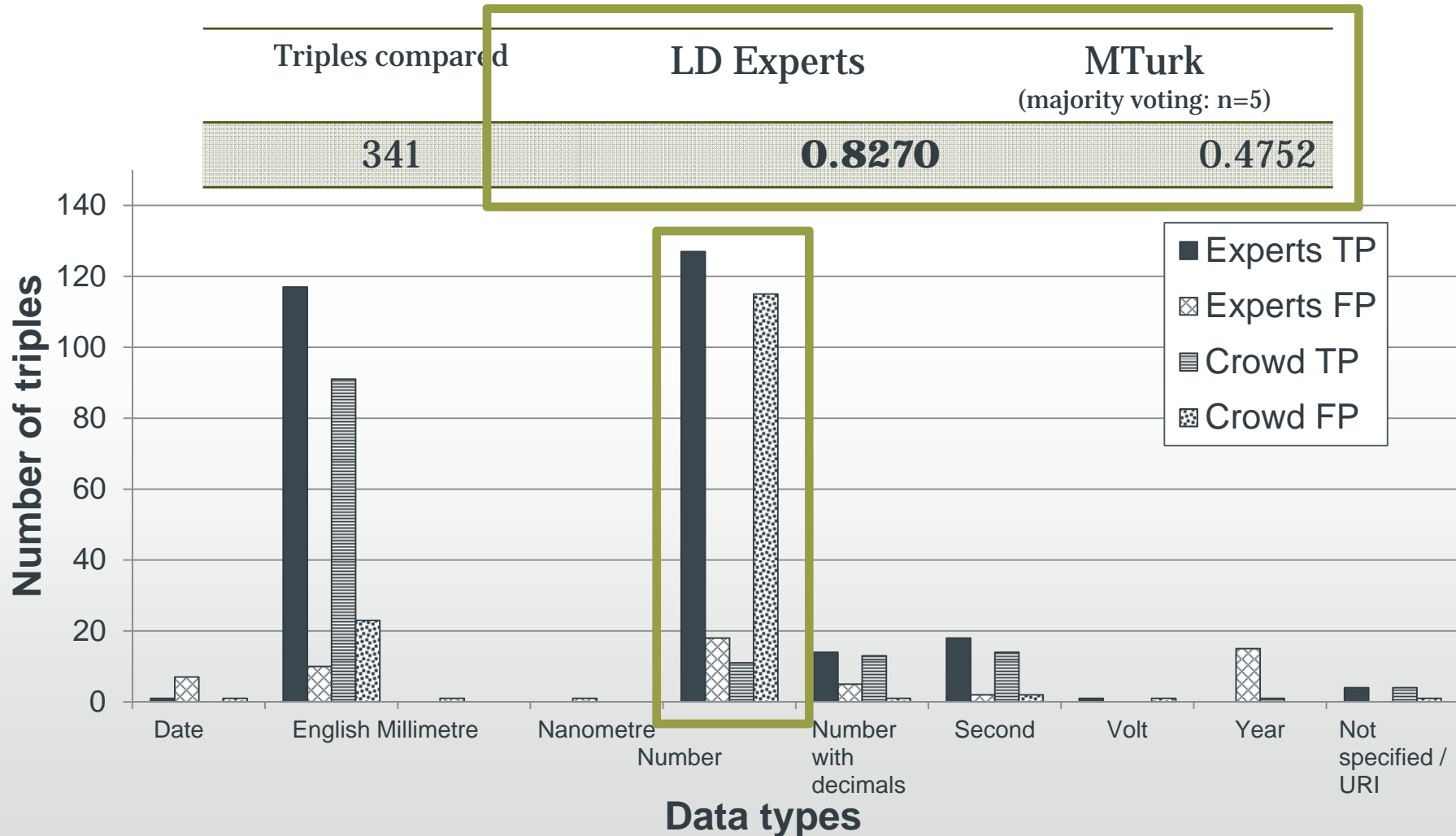
"2005 Six Nations Championship" Date 12 .

- 52 DBpedia triples had **erroneous values from the source**:

"English (programming language)" Influenced by ? .

- Experts classified all these triples as incorrect
- Workers compared values against Wikipedia and successfully classified this triples as "correct"

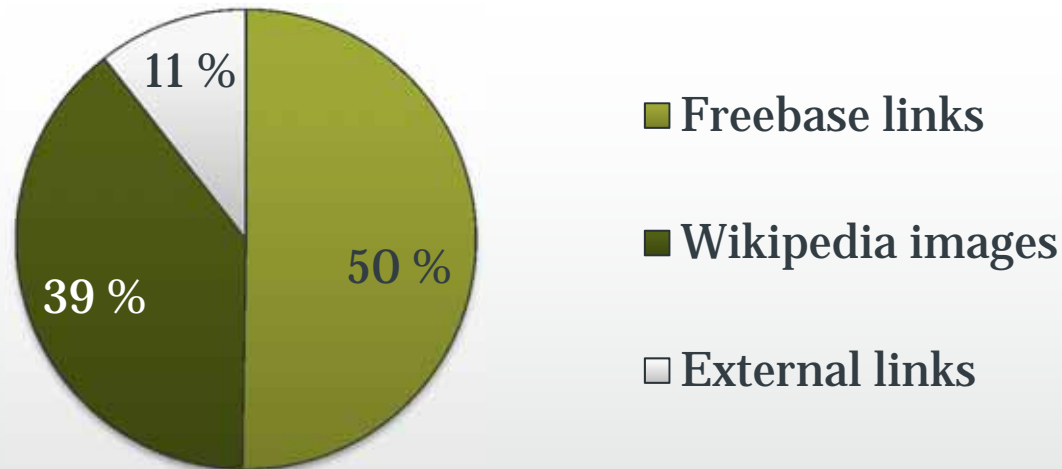
Precision results: Incorrect data type task



Precision results: Incorrect link task

Triples compared	Baseline	LD Experts	MTurk (n=5 majority voting)
223	0.2598	0.1525	0.9412

- We analyzed the **189 misclassifications** by the **experts**:



- The **6% misclassifications** by the **workers** correspond to pages with a language different from English.

Summary of findings

- The effort of LD experts must be applied on those tasks demanding specific-domain skills.
- MTurk crowd was exceptionally good at performing data comparisons
- Lay users do not have the skills to solve domain-specific tasks, while experts performance is very low on tasks that demand an extra effort (e.g., checking an external page)