Electronics and
Computer Science

UNIVERSITY OF Southampton

The role of humans in crowdsourced semantics

Elena Simperl, University of Southampton*

WIC@WWW2014

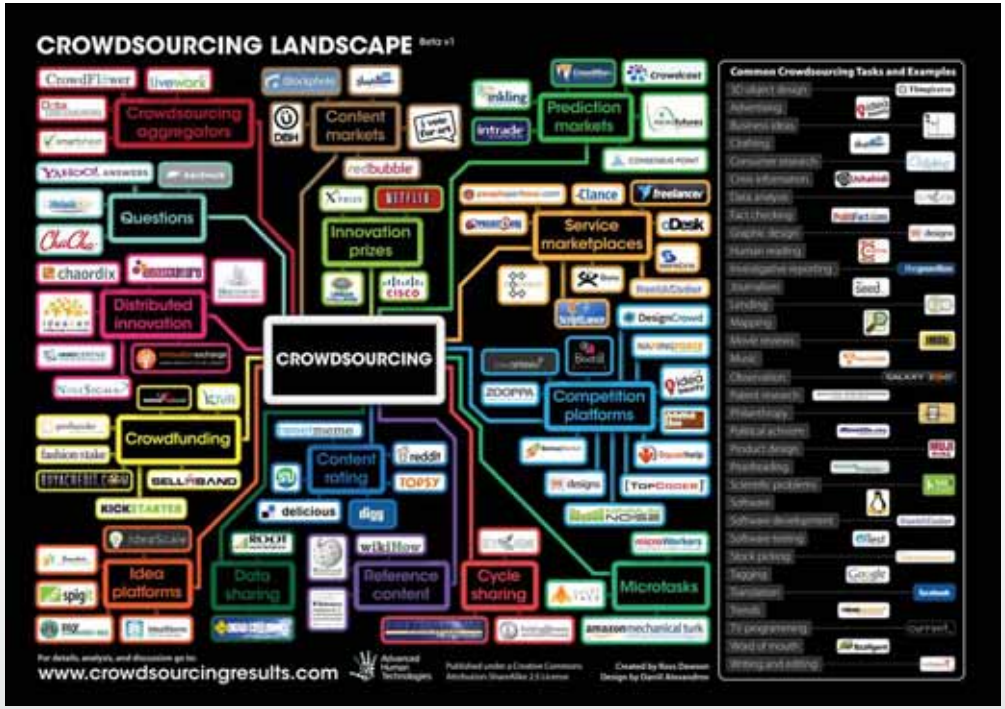
*with contributions by Maribel Acosta, KIT

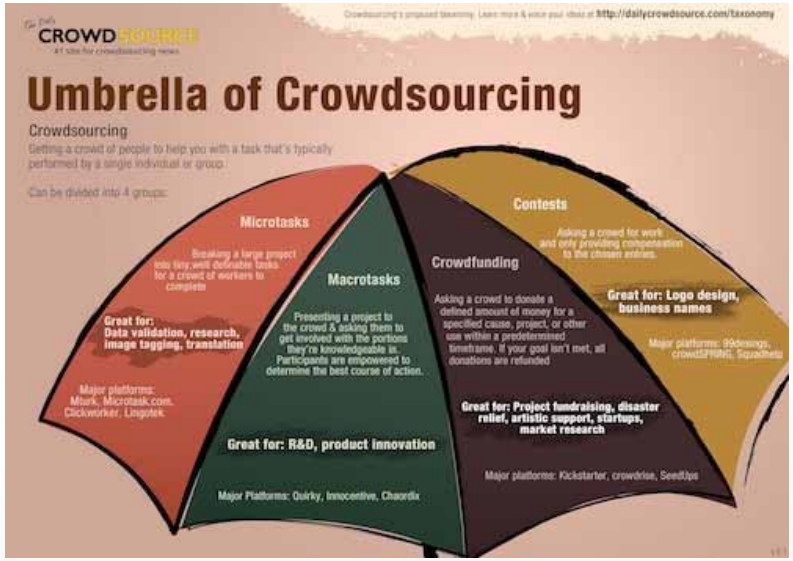
07 April 2014

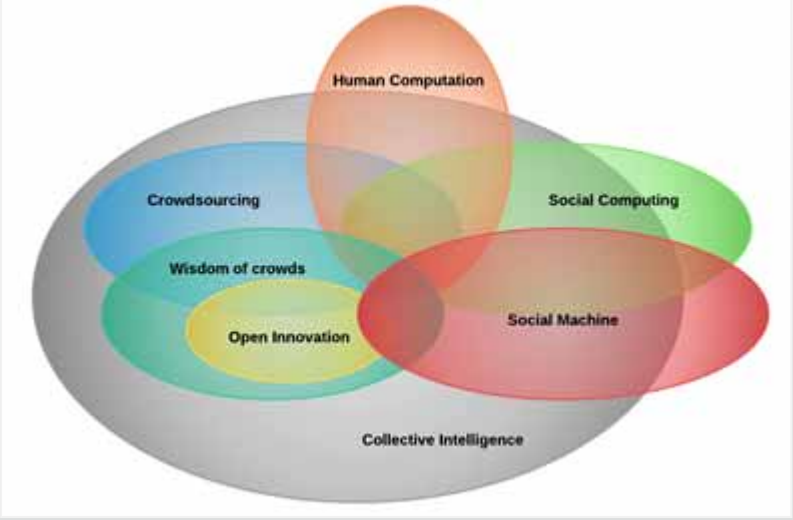
Crowdsourcing Web semantics: the great challenge

- 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...



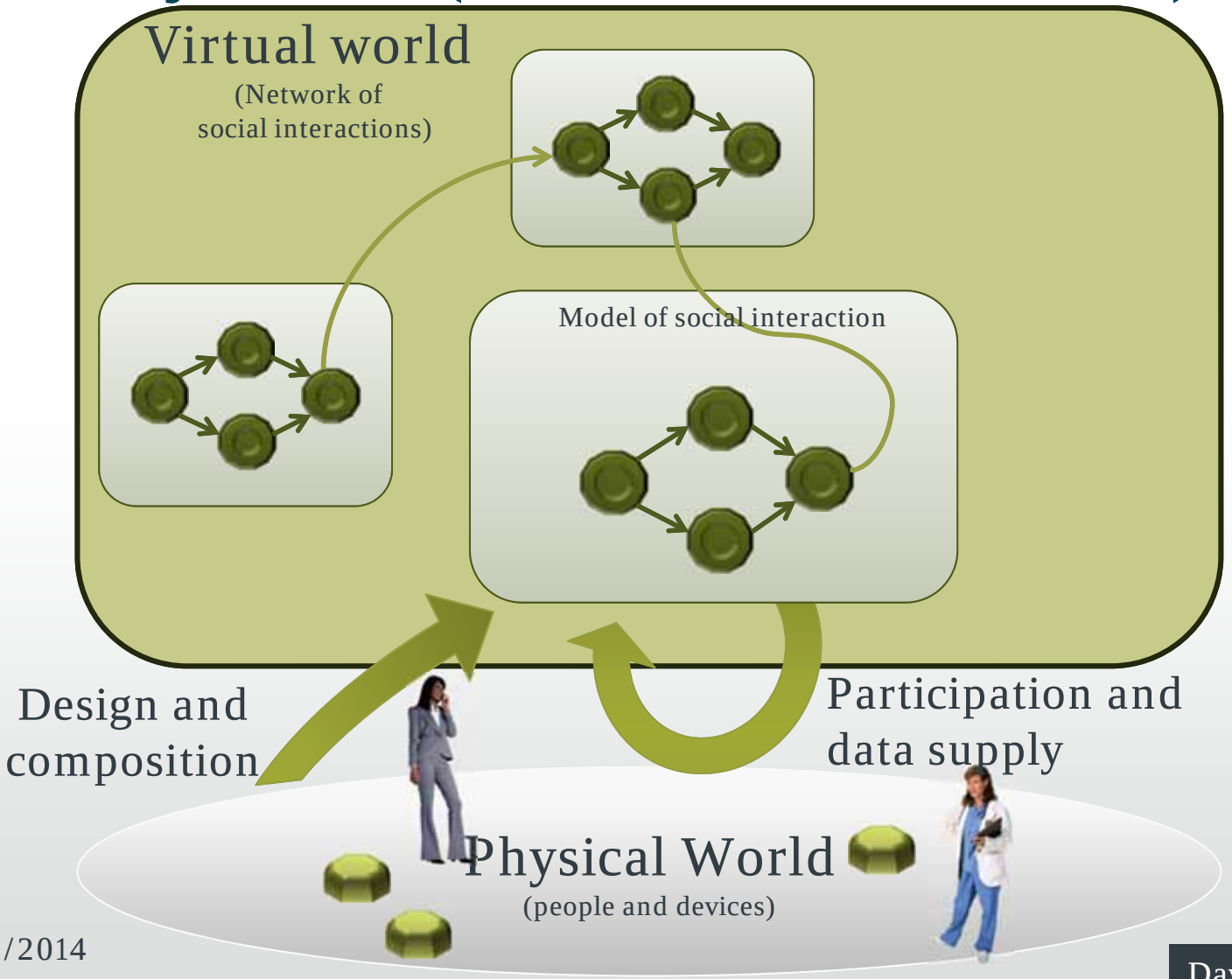




Microtask crowdsourcing

Work is broken down into smaller (‚micro‘) pieces that can be solved independently

Hybrid systems (or ‚social machines')



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

McCann, Shen, Doan: Matching Schemas in Online Communities. ICDE, 2008

Example: Hybrid query processing

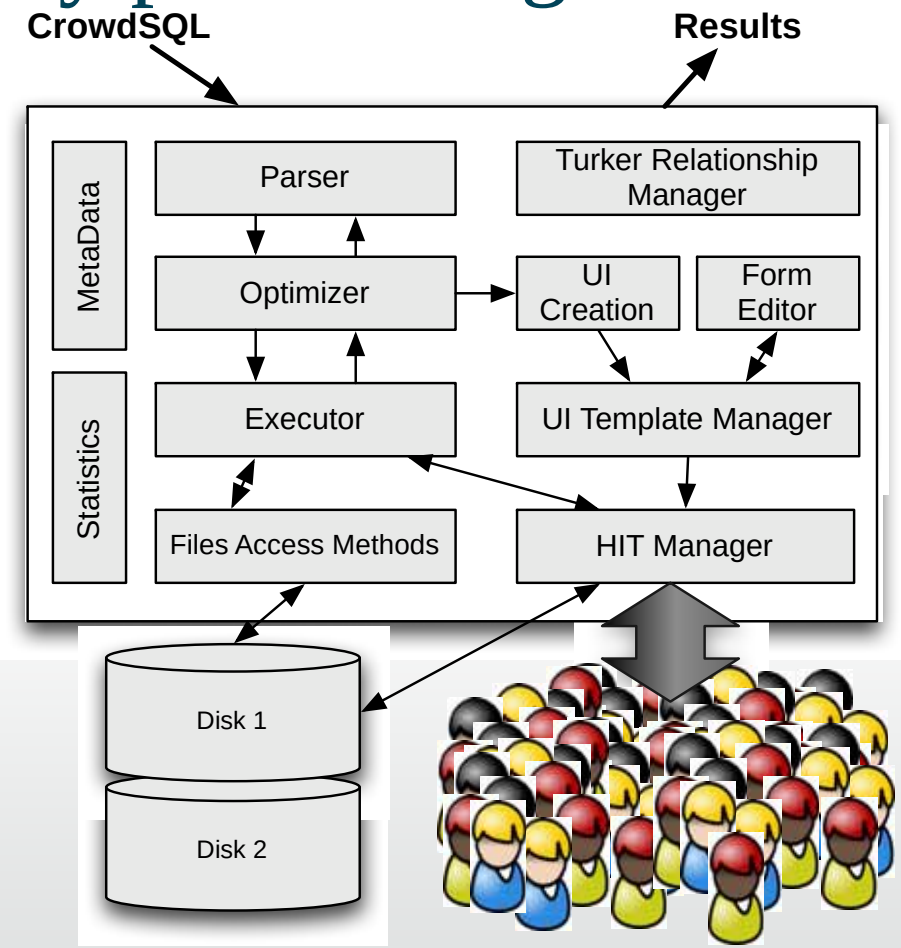
Use the crowd to answer DB-hard queries

Where to use the crowd:

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

But not:

- Anything the computer already does well

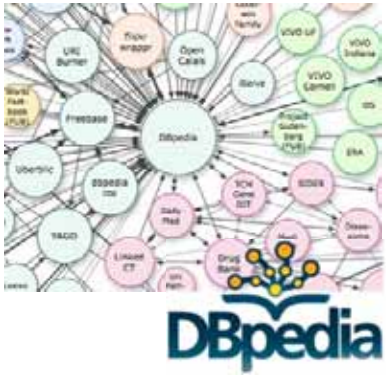


M. Franklin, D. Kossmann, T. Kraska, S. Ramesh and R. Xin . CrowdDB: Answering Queries with Crowdsourcing, SIGMOD 2011

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

Verify

Contest

LD Experts

Difficult task

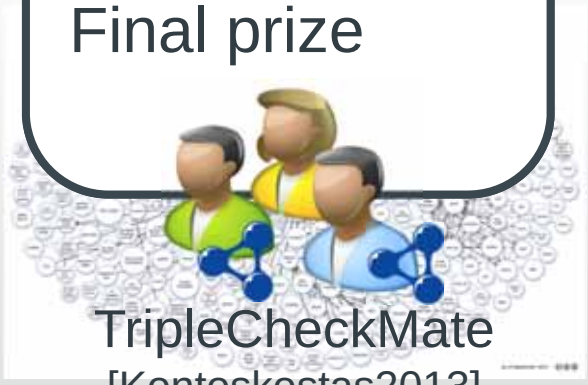
Final prize

Microtasks

Workers

Easy task

Micropayments

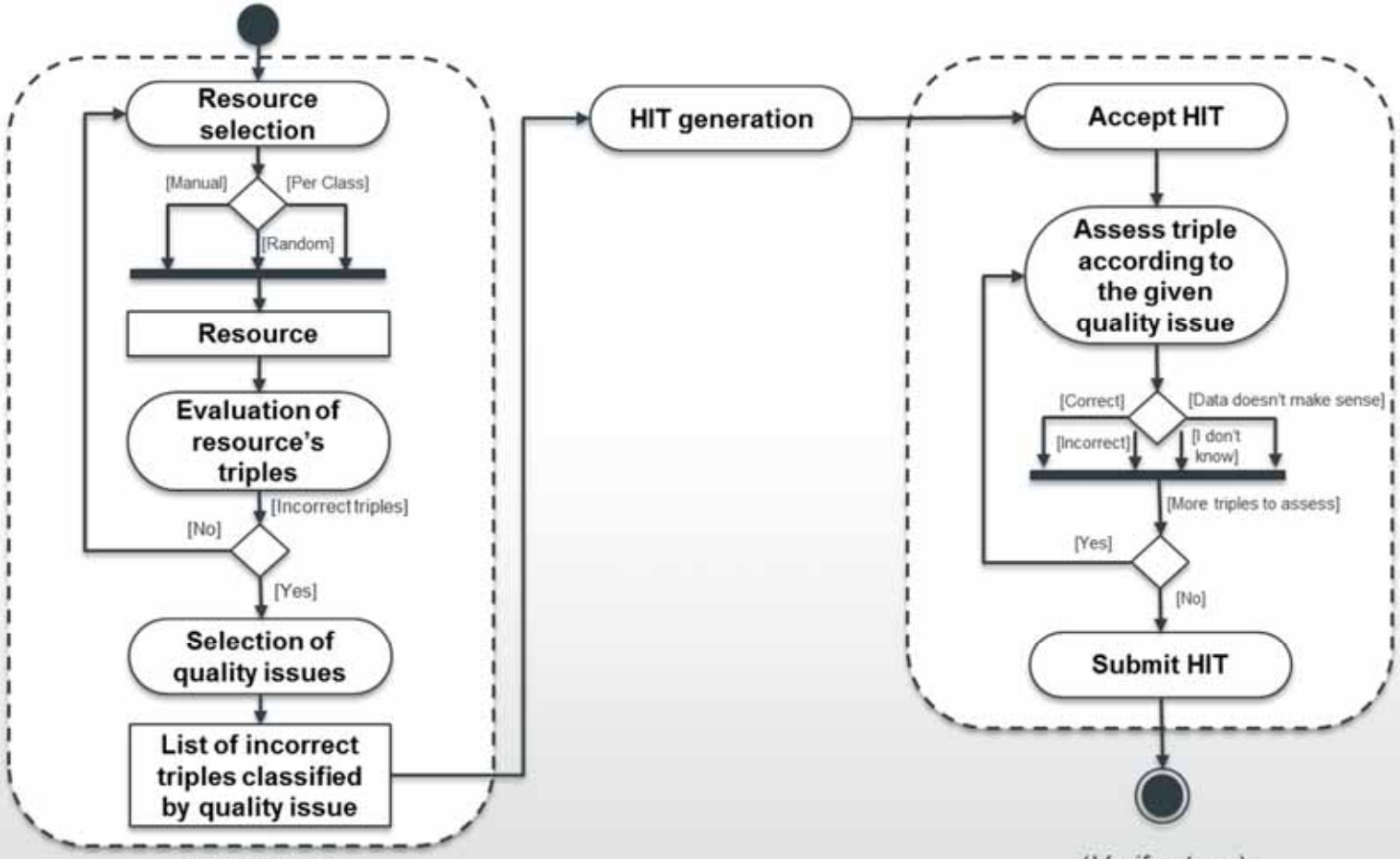


TripleCheckMate
[Kontoskostas2013]

MTurk
http://mturk.com

Adapted from [Bernstein2010]

Workflow

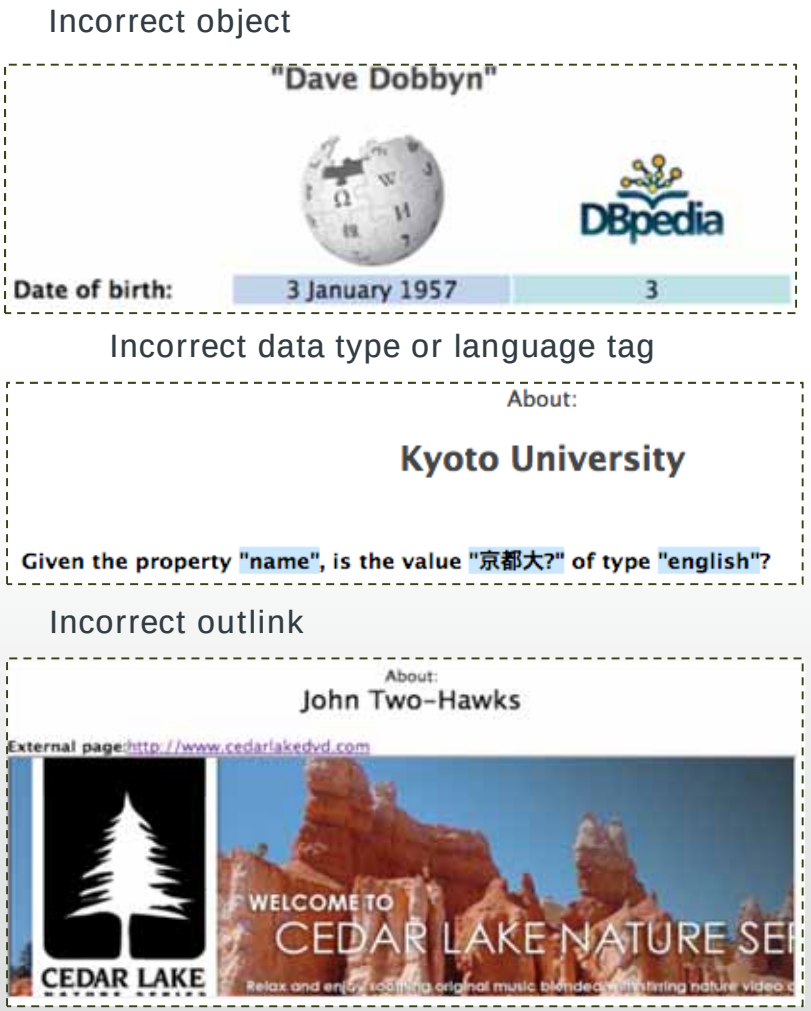


(*Find stage*)
LD Experts in contest

(*Verify stage*)
Workers in paid microtasks

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`



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 .
```

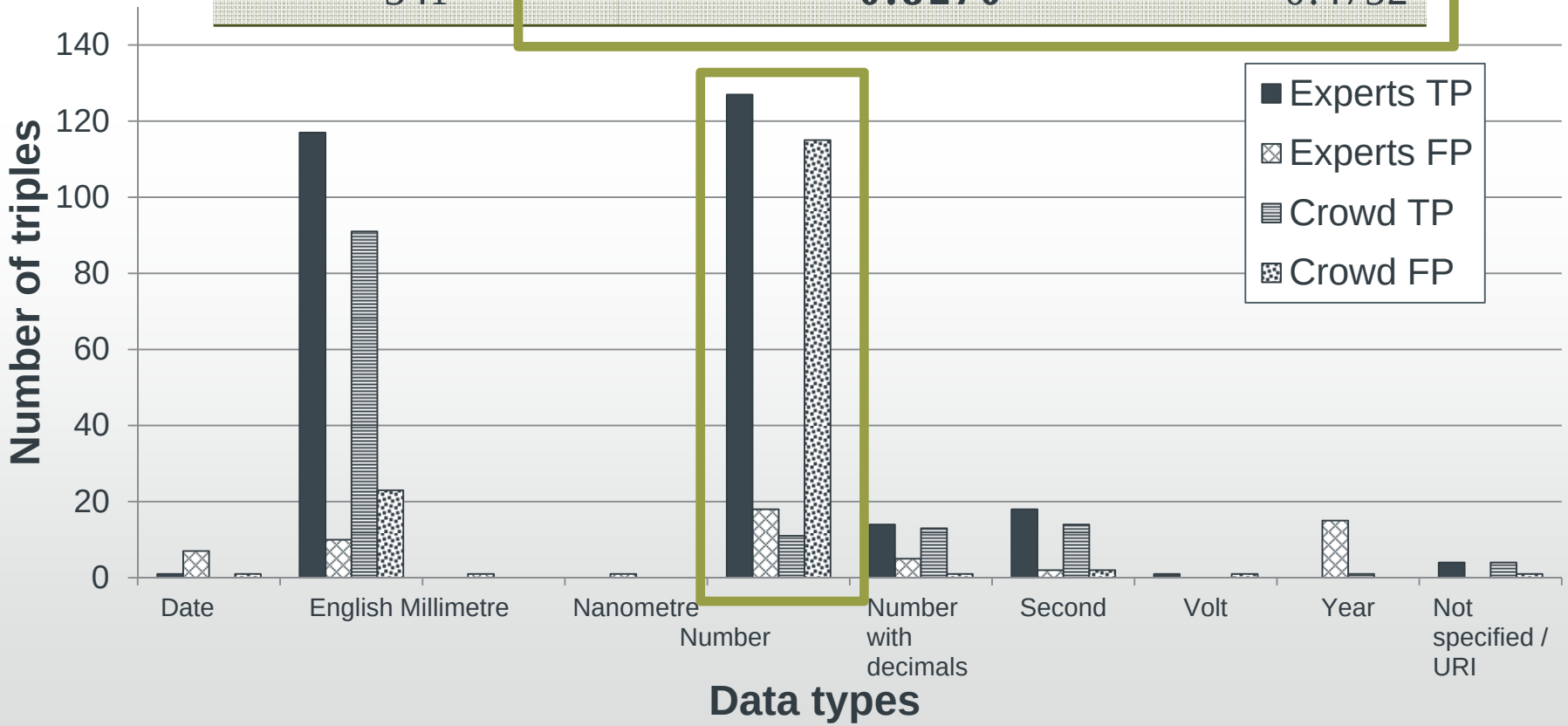
- 52 DBpedia triples had **erroneous values from the source**:

```
"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

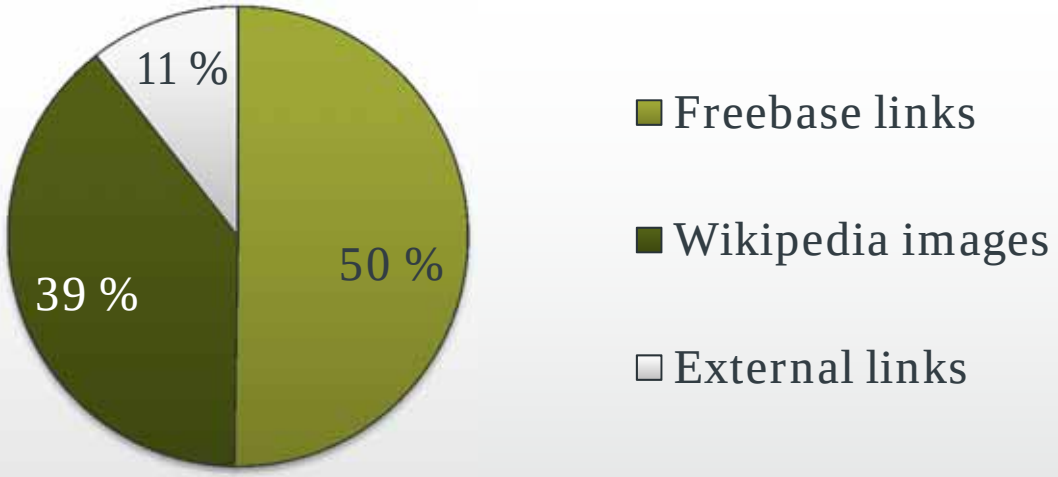
Triples compared	LD Experts	MTurk (majority voting: n=5)
341	**0.8270**	0.4752



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)