ONE DOES NOT SIMPLY CROWDSOURCE THE SEMANTIC WEB: 10 YEARS WITH PEOPLE, URIS, GRAPHS AND INCENTIVES

Elena Simperl

Sematics 2018, Vienna

CROWDSOURCING = OUTSOURCING + CROWD

99designs

amazon mechanicalturk™
Artificial Artificial Intelligence

INNOCENTIVE®

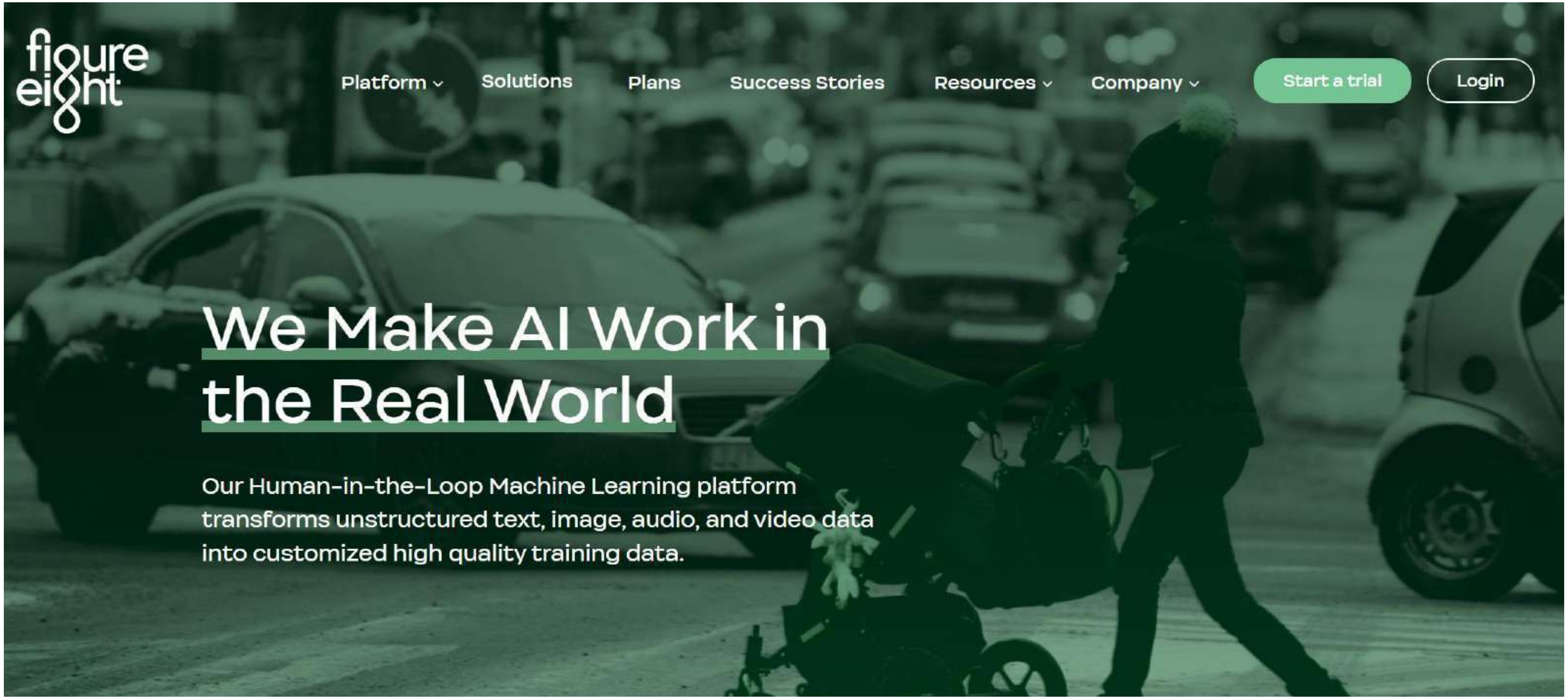
Ushahidi

KICKSTARTER

kaggle™

ZOONIVERSE
REAL SCIENCE ONLINE

AUGMENTED INTELLIGENCE:
CROWDSOURCING CREATES LABELLED DATA FOR ML ALGORITHMS



CROWDSOURCING AND THE SEMANTIC WEB

- Semantic applications developers use crowdsourcing to achieve a goal
- The Semantic Web is a giant crowdsourcing project

THE DESIGN SPACE OF A CROWDSOURCING PROJECT

What Goal	**Who** Staffing
How Process	**Why** Incentives

CROWDSOURCING - WHAT

Tasks based on human skills, not easily replicable by machines

Editing knowledge graphs

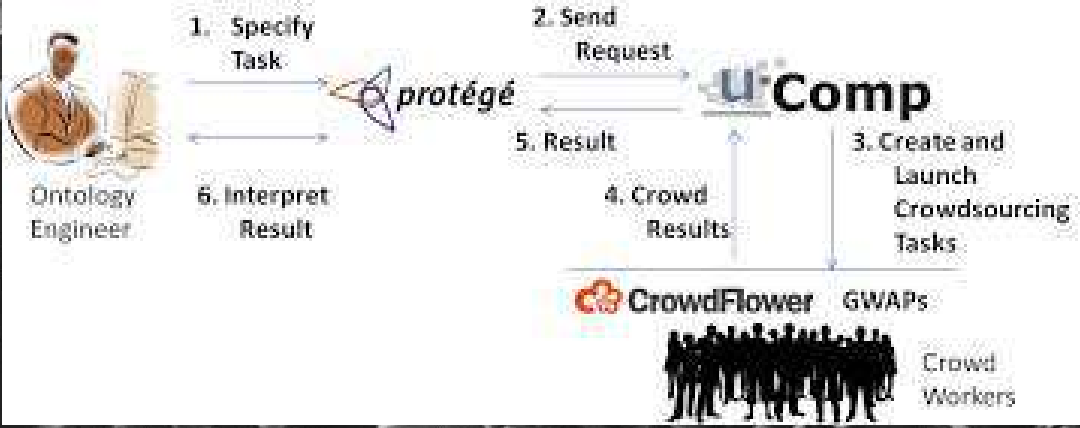
Adding semantic annotations to media

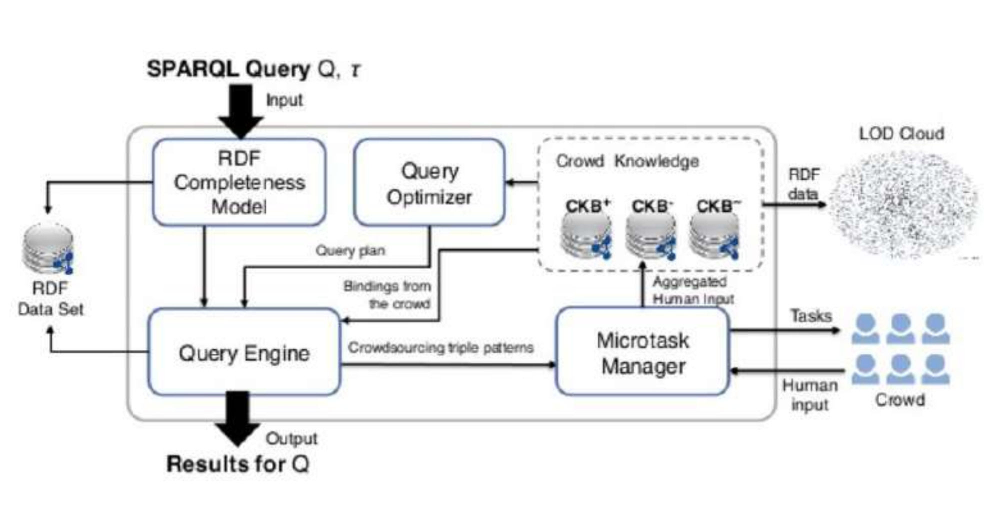
Adding multilingual labels to entities

Defining links between entities

…

Most effective when used at scale ('open call'), in combination w/ machine intelligence





CROWDSOURCING – **WHO**

There is more to crowdsourcing than Mechnical Turk

Use the right crowd for the right task

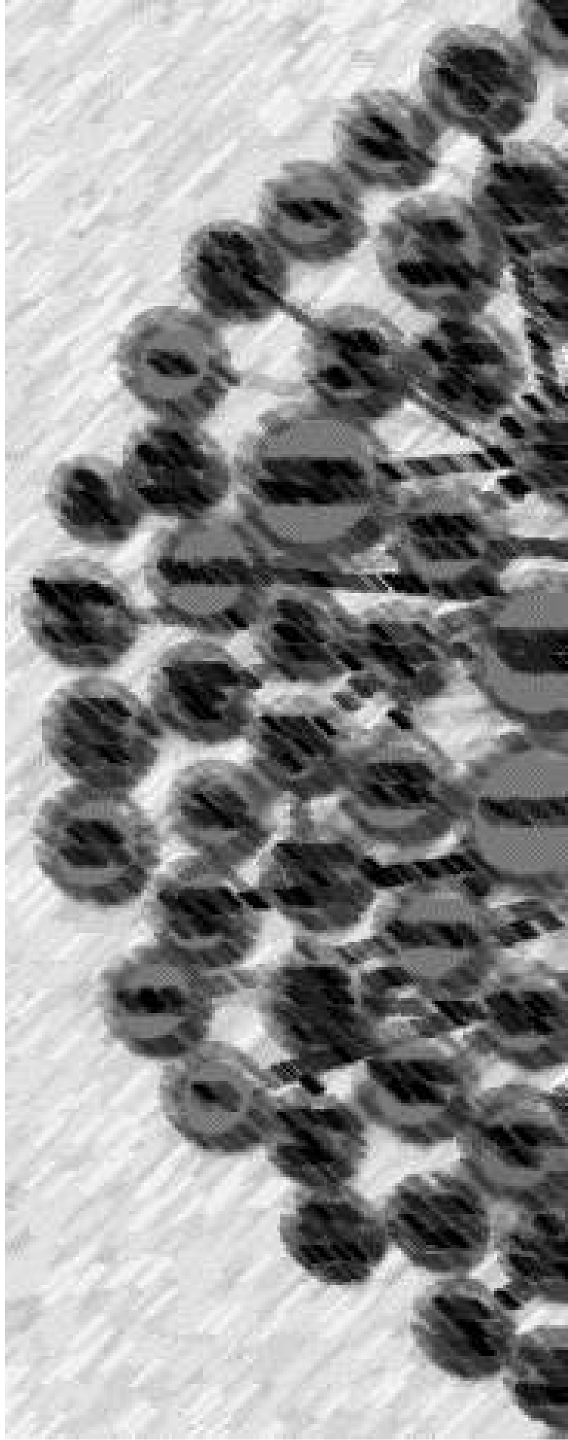
Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Flöck, F., & Lehmann, J. (2016). Detecting Linked Data quality issues via crowdsourcing: A DBpedia study. *Semantic Web Journal*, 1-34.

BACKGROUND

Varying quality of Linked Data sources

```
dbpedia:Dave_Dobbyn dbprop:dateOfBirth “3”.
```

Detecting and correcting errors may require manual inspection



Approach

Find

Contest
LD Experts
Difficult task
Final prize

TripleCheckMate

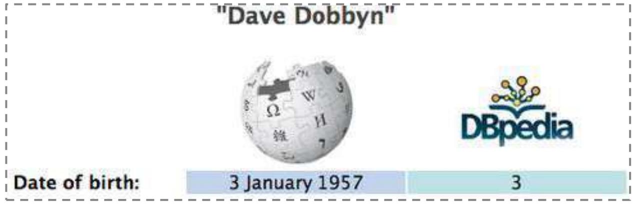
Verify

Microtasks
Workers
Easy task
Micropayments

MTurk

MTurk interfaces

Incorrect object



Incorrect data type

Incorrect outlink

Results: Precision

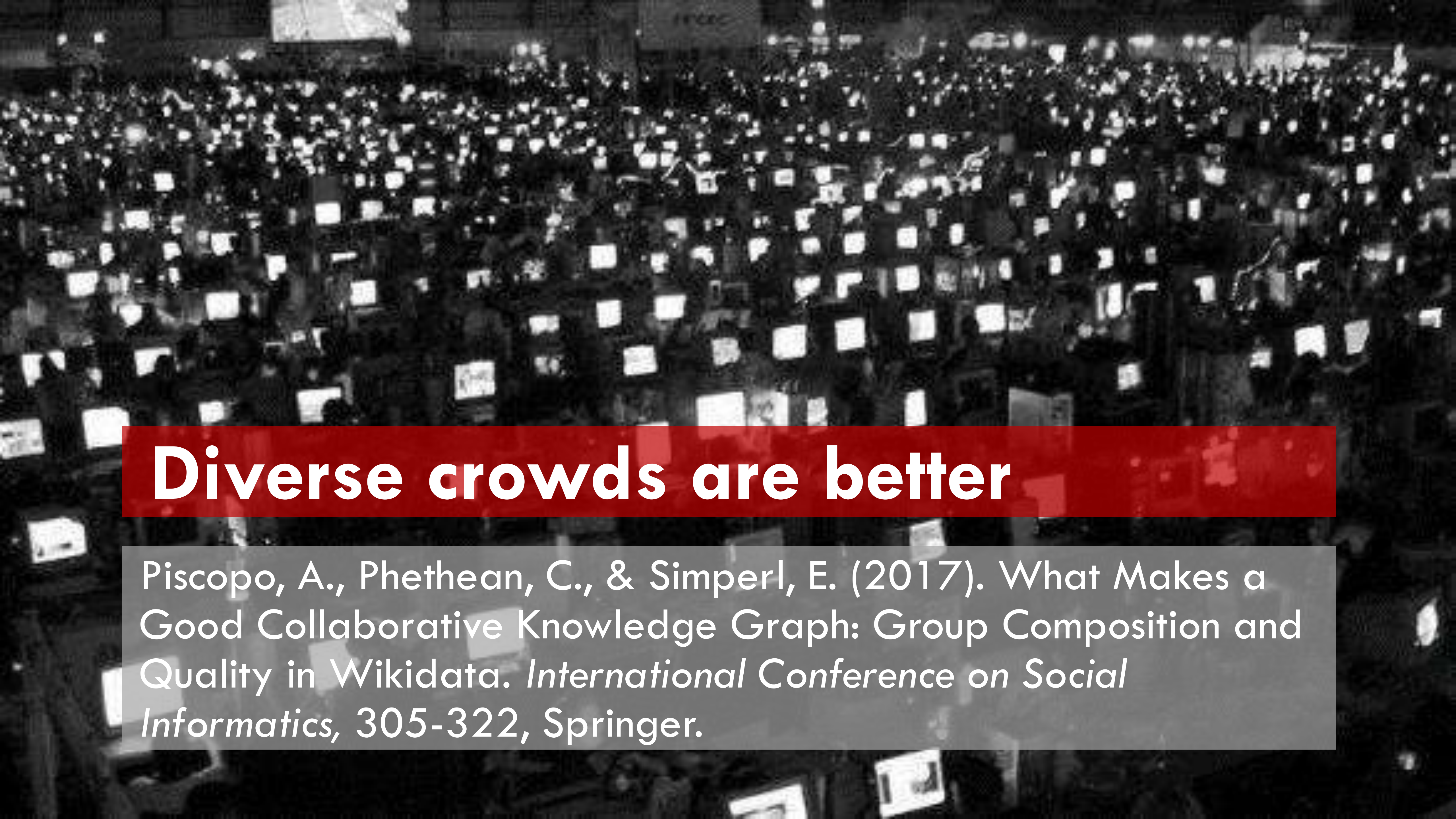
	Object values	Data types	Interlinks
Linked Data experts	0.7151	0.8270	0.1525
MTurk (majority voting)	0.8977	0.4752	0.9412

Findings

Use the right crowd for the right task

Experts detect a range of issues, but will not invest additional effort

Turkers can carry out the three tasks and are exceptionally good at data comparisons



Diverse crowds are better

Piscopo, A., Phethean, C., & Simperl, E. (2017). What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. *International Conference on Social Informatics,* 305-322, Springer.

BACKGROUND

Items and statements in Wikidata are edited by teams of editors

Editors have varied tenure and interests

Group composition impacts outcomes

- Diversity can have multiple effects
- Moderate tenure diversity increases outcome quality
- Interest diversity leads to increased group productivity

Chen, J., Ren, Y., Riedl, J.: The effects of diversity on group productivity and member withdrawal in online volunteer groups. In: Proceedings of the 28th international conference on human factors in computing systems - CHI '10. p. 821. ACM Press, New York, USA (2010)

STUDY

Analysed the edit history of items

- Corpus of 5k items, whose quality has been manually assessed (5 levels)*
- Edit history focused on community make-up
  - Community is defined as set of editors of item
  - Considered features from group diversity literature and Wikidata-specific aspects

*https://www.wikidata.org/wiki/Wikidata:Item_quality

HYPOTHESES

	Activity	Outcome
H1	Bots edits ↑	Item quality ↑
H2	Bot-human interaction ↑	Item quality ↑
H3	Anonymous edits ↑	Item quality ↓
H4	Tenure diversity ↑	Item quality ↑
H5	Interest diversity ↑	Item quality ↑

RESULTS

ALL HYPOTHESES SUPPORTED

		Model 1			Model 2			Model 3			Model 4		
		Coef.	SE	P	Coef.	SE	P	Coef.	SE	P	Coef.	SE	P
Label > = D		−.0715	.0609		−1.3024	.1037	***	−1.1739	.1779	***	−2.6487	.2125	***
Label > = C		−1.2553	.0642	***	−2.5499	.1081	***	−2.3874	.1815	***	−4.1062	.2175	***
Label > = B		−4.4452	.1028	***	−5.7677	.1361	***	−5.8900	.2145	***	−7.5732	.2450	***
Label > = A		−6.2173	.1320	***	−7.6024	.1628	***	−7.4843	.2262	***	−9.2759	.2573	***
Item age		.0003	.0001	***	.0001	.0001		.0002	.0001		−.0008	.0001	***
Group size		.0279	.0014	***	.0330	.0015	***	.0152	.0015	***	.0248	.0016	***
# Edits		.0029	.0003	***	.0033	.0003	***	.0039	.0003	***	.0040	.0003	***
p Bot edits	**H1**				1.4005	.1029	***				2.4695	.1237	***
Bot X Human	**H2**				4.6909	.3377	***				3.7688	.3618	***
p Anonymous edits	**H3**				−3.8258	1.2218	**				−3.6628	1.2403	
Tenure diversity	**H4**							1.5502	.1104	***	2.8043	.1166	***
Interest diversity	**H5**							1.0104	.1972	***	1.1004	.1999	***

SUMMARY AND IMPLICATIONS

01
The more is not always the merrier

02
Bot edits are key for quality, but bots and humans are better

03
Registered editors have a positive impact

04
Diversity matters

01
Encourage registration

02
Identify further areas for bot editing

03
Design effective human-bot workflows

04
Suggest items to edit based on tenure and interests

CROWDSOURCING - **HOW**

There are different ways to carry out a task using crowdsourcing

They will produce different results

Bu, Q., Simperl, E., Zerr, S., & Li, Y. (2016). Using microtasks to crowdsource DBpedia entity classification: A study in workflow design. *Semantic Web Journal*, 1-18.

THREE WORKFLOWS TO CROWDSOURCE ENTITY TYPING

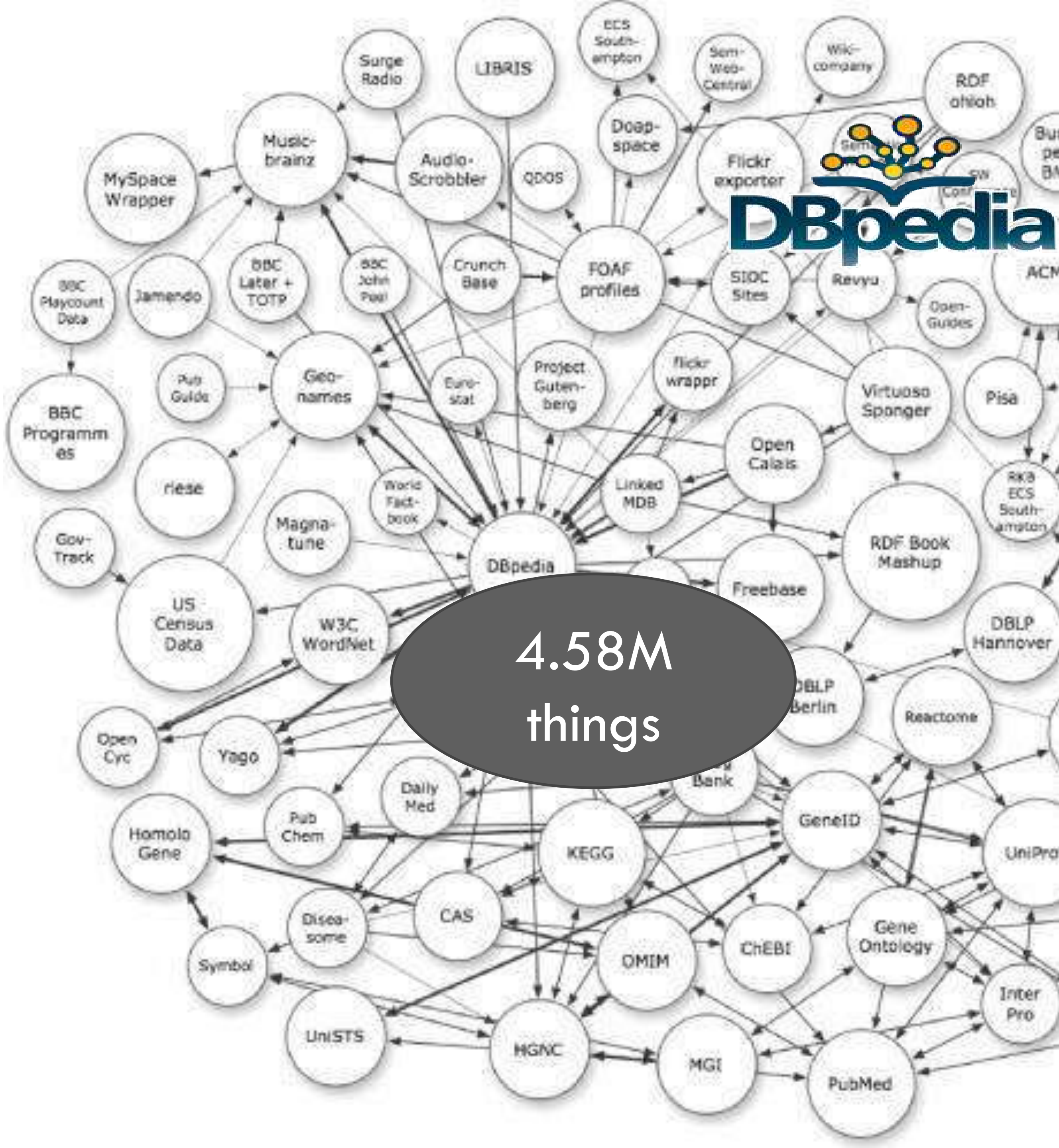
Free associations

Validating the machine

Exploring the DBpedia ontology

Findings

- **Shortlists are easy & fast**
  - Popular classes are not enough
  - Alternative ways to explore the taxonomy
- **Freedom comes with a price**
  - Unclassified entities might be unclassifiable
  - Different human data interfaces
- **Working at the basic level of abstraction achieves greatest precision**
  - But when given the freedom to choose, users suggest more specific classes



Crowds need human-readable interfaces to KGs

Kaffee, L. A., Elsahar, H., Vougiouklis, P., Gravier, C., Laforest, F., Hare, J., & Simperl, E. (2018). Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders. In *European Semantic Web Conference* (pp. 319-334). Springer.

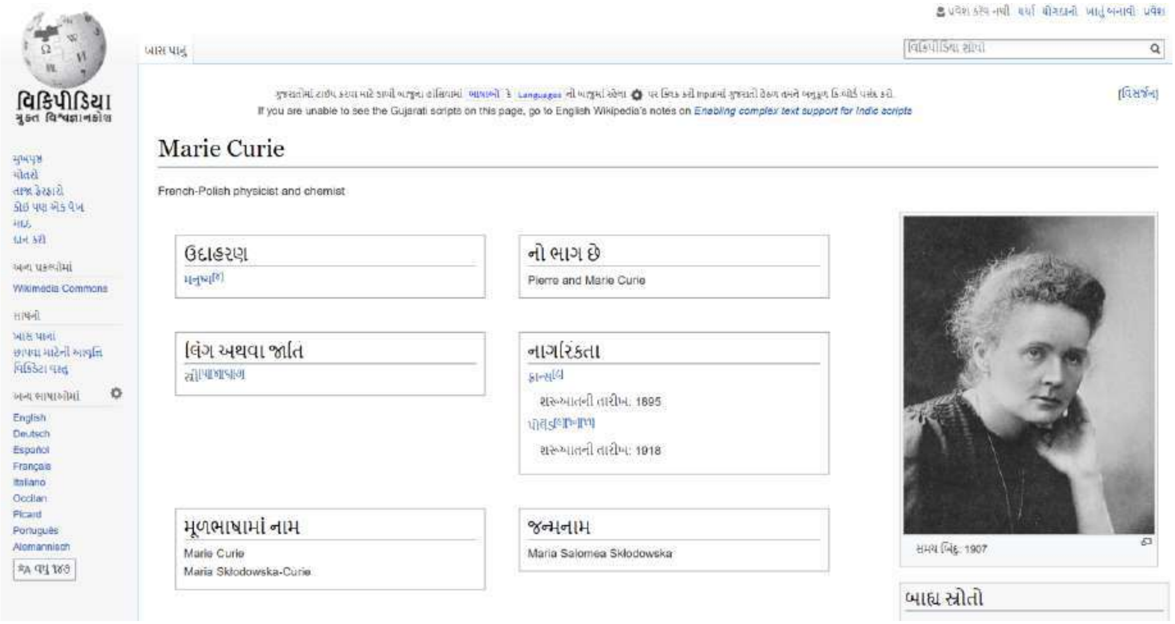
BACKGROUND

Wikipedia is available in 287 languages, but content is unevenly distributed

Wikidata is cross-lingual

ArticlePlaceholders display Wikidata triples as stubs for articles in underserved Wikipedia's

Currently deployed in 11 Wikipedia's



STUDY

Enrich ArticlePlaceholders with **textual summaries** generated from Wikidata triples

Train a **neural network** to generate one sentence summaries resembling the opening paragraph of a Wikipedia article

Test the approach on two languages, **Esperanto** and **Arabic** with readers and editors of those Wikipedia's

Page Statistic	Esperanto	Arabic	English	Wikidata
Articles	241,901	541,166	5,483,928	37,703,807
Avg edits/page	11.48	8.94	21.11	14.66
Active users	2,849	7,818	129,237	17,583
Vocab. size	1.5M	2.2M	2.0M	–

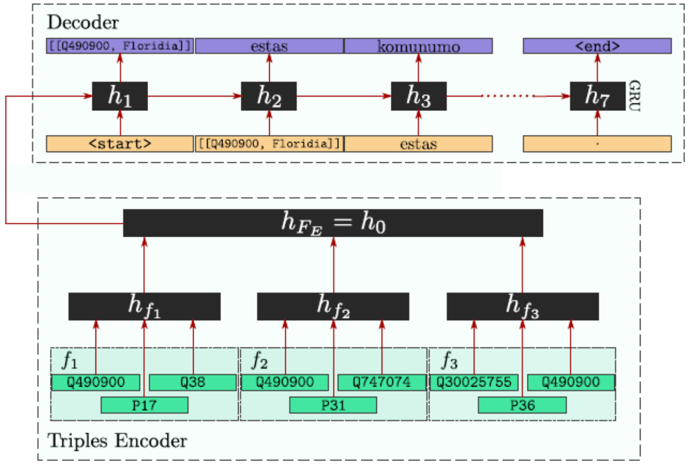
APPROACH

NEURAL NETWORK TRAINED ON WIKIDATA/WIKIPEDIA

Feed-forward architecture encodes triples from the ArticlePlaceholder into vector of fixed dimensionality

RNN-based decoder generates text summaries, one token at a time

Optimisations for different entity verbalisations, rare entities etc.



Article-Placeholder Triples	f_1 : Q490900 (Floridia)	P17 (ŝtato)	Q38 (Italio)
	f_2 : Q490900 (Floridia)	P31 (estas)	Q747074 (komunumo de Italio)
	f_3 : Q30025755 (Floridia)	P1376 (ĉefurbo de)	Q490900 (Floridia)
Textual Summary	Floridia estas komunumo de Italio.		
Vocab. Extended Summary	[[Q490900, Floridia]] estas komunumo de [[P17]].		

AUTOMATIC EVALUATION
APPROACH OUTPERFORMS BASELINES

Trained on corpus of Wikipedia sentences and corresponding Wikidata triples (205k Arabic; 102k Esperanto)

Tested against three baselines: machine translation (MT) and template retrieval (TR, TR_{ext})

Using standard metrics: BLEU, METEOR, $ROUGE_L$

	Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		$ROUGE_L$		METEOR	
		valid.	test	valid.	test	valid.	test	valid.	test	valid.	test	valid.	test
Arabic	MT	31.12	33.48	19.31	21.12	12.69	13.89	8.49	9.11	29.96	30.51	31.05	30.1
	TP	41.39	41.73	34.18	34.58	29.36	29.72	25.68	25.98	43.26	43.58	32.99	33.33
	TP_{ext}	49.87	48.96	42.44	41.5	37.29	36.41	33.27	32.51	51.66	50.57	34.39	34.25
	Ours	**53.18**	**52.94**	**45.86**	**45.64**	**40.38**	**40.21**	**35.7**	**35.55**	**57.9**	**57.99**	**39.22**	**39.37**
Esperanto	MT	5.35	5.47	1.62	1.62	0.59	0.56	0.26	0.23	4.67	4.79	0.66	0.68
	TP	43.01	42.61	33.67	33.46	28.16	28.07	24.35	24.3	46.75	45.92	20.71	20.46
	TP_{ext}	52.75	51.66	43.57	42.53	37.53	36.54	33.35	32.41	58.15	57.62	**31.21**	**31.04**
	Ours	**56.51**	**56.96**	**47.72**	**48.1**	**41.8**	**42.13**	**37.24**	**37.52**	**64.36**	**64.69**	28.35	28.76

USER STUDIES

SUMMARIES ARE USEFUL FOR THE COMMUNITY

		Fluency		Appropriateness
		Mean	SD	Part of Wikipedia
Arabic	Ours	4.7	1.2	77%
	Wikipedia	4.6	0.9	74%
	News	5.3	0.4	35%
Esper.	Ours	4.5	1.5	69%
	Wikipedia	4.9	1.2	84%
	News	4.2	1.2	52%

Readers study, 15 days, mixed corpus of 60 articles

Editors study, 15 days, 30 summaries

	Category	Examples	%
Arabic	WD	خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة(كلمة ناقصة)، ويكون على شكل بلورات بيضاء. خماسي كلوريد الزرنيخ هو مُركب كميائي له الصيغة(AtClu2085)، ويكون على شكل بلورات بيضاء.	45.45%
	PD	بيتش باتوم أوهايو (بالإنجليزية (كلمة ناقصة) Ohio)هي منطقة سكنية تقع في الولايات المتحدة في(كلمة ناقصة.) بيتش باتوم (بالإنجليزية:Beach Batom)هي قرية تقع في الولايات المتحدة الامريكية في بروك كاونتي.	33.33%
	ND	دير علا هي بلدة تقع في جنوب غرب إيران. دير علا، أو بيثر، هي قرية أردنية	21.21%
Esperanto	WD	Zederik estas komunumo en la nederlanda provinco Zuid-Holland. Zederik estas komunumo en la nederlanda provinco Zuid-Hooland kaj estas ĉirkaŭata de la municipoj Lopik kaj Zederik.	78.98%
	PD	Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando, kiu havis (manka nombro) loĝantojn en (jaro). Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando.	15.79%
	ND	Ibiúna estas municipo de la brazila subŝtato San-Paŭlio, kiu taksis (manka nombro) enloĝantojn en (jaro). Ibiúna estas brazila [[municipo]] kiu troviĝas en la administra unuo [[San-Paŭlo]].	5.26%

CROWDSOURCING - **WHY**

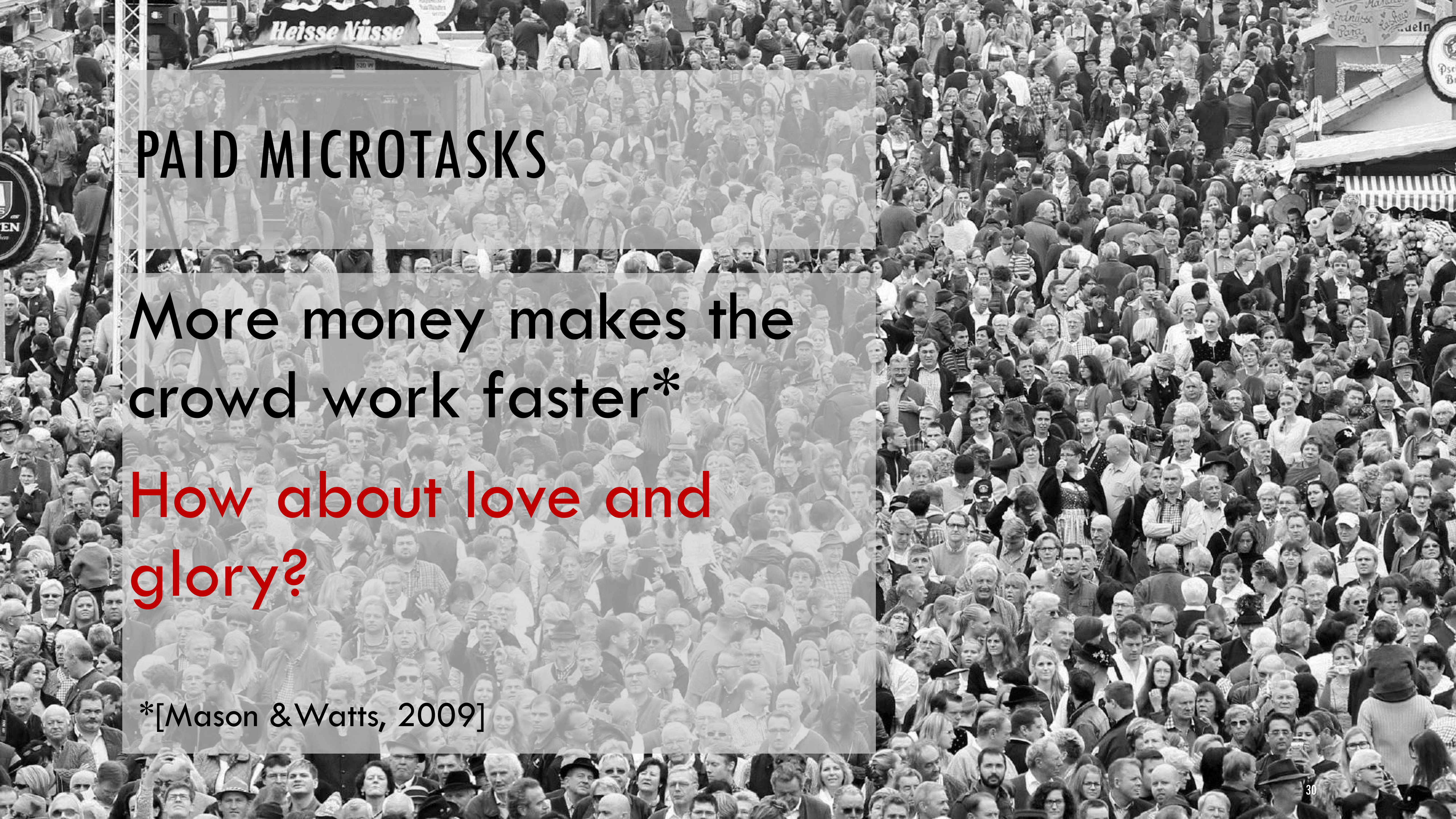
THEORY OF MOTIVATION

People do things for three reasons

Love and glory keep costs down

Money and glory deliver faster

LOVE
MONEY
GLORY



PAID MICROTASKS

More money makes the crowd work faster*

How about love and glory?

*[Mason & Watts, 2009]

EXPERIMENT 1

Make paid microtasks more cost-effective w/ gamification

Workers will perform better if tasks are more engaging

- Increased accuracy through higher inter-annotator agreement
- Cost savings through reduced unit costs

Micro-targeting incentives when players attempt to quit improves retention

MICROTASK DESIGN

Image labelling tasks, published on microtask platform
- Free-text labels, varying numbers of labels per image, taboo words
- Workers can skip images, play as much as they want

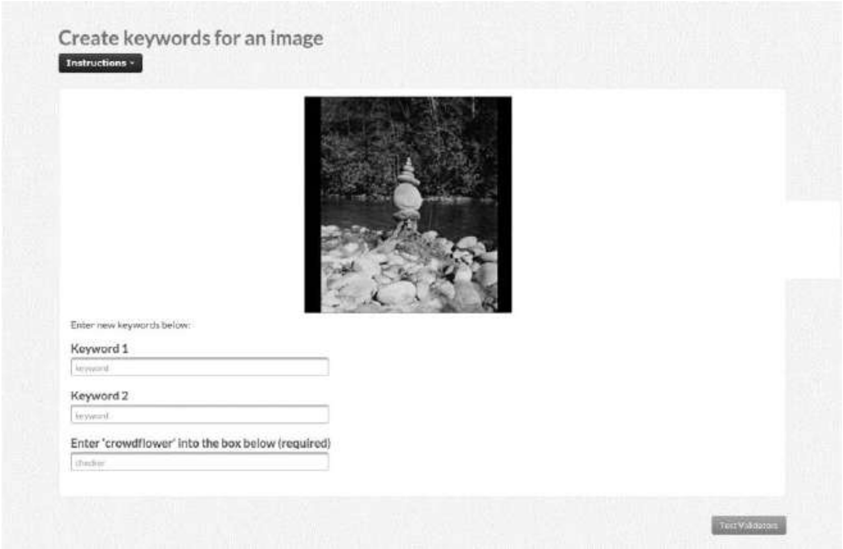
Baseline: 'standard' tasks w/ basic spam control

vs

Gamified: same requirements & rewards, but crowd asked to complete tasks in Wordsmith

vs

Gamified & furtherance incentives: additional rewards to stay (random, personalised)



EVALUATION

- ESP data set as gold standard
- #labels, agreement, mean & max #labels/worker
- Three tasks
  - Nano: 1 image
  - Micro: 11 images
  - Small: up to 2000 images
- Probabilistic reasoning to predict worker exit and personalize furtherance incentives

RESULTS (GAMIFICATION, 1 IMAGE)

BETTER, CHEAPER, BUT FEWER WORKERS

Metric	CrowdFlower	Wordsmith
Total workers	600	423
Total keywords	1,200	41,206
Unique keywords	111	5,708
Avg. agreement	5.72%	37.7%
Avg. images/person	1	32
Max images/person	1	200

RESULTS (GAMIFICATION, 11 IMAGES)

COMPARABLE QUALITY, HIGHER UNIT COSTS, FEWER DROPOUTS

Metric	CrowdFlower	Wordsmith
Total workers	600	514
Total keywords	13,200	35,890
Unique keywords	1,323	4,091
Avg. agreement	6.32%	10.9%
Avg. images/person	11	27
Max images/person	1	351

RESULTS (WITH FURTHERANCE INCENTIVES)
MORE ENGAGEMENT, TARGETING WORKS

Increased participation

- People come back (20 times) and play longer (43 hours vs 3 hours without incentives)
- Financial incentives play important role

Targeted incentives work

- 77% players stayed vs. 27% in the randomised condition
- 19% more labels compared to no incentives condition

Incentive	C3: Randomised	C4: Targeted
Power	26.09%	30.16%
Money	19.65%	46.17%
Leaderboard	16.59%	5.71%
Levels	13.01%	7.34%
Badges	13.04%	5.98%
Access	11.61%	4.35%

EXPERIMENT 2

Make paid microtasks more cost-effective w/ social incentives

Working in pairs is more effective than the baseline

- Increased higher inter-annotator agreement
- Higher output

Social incentives improve retention past payment threshold

MICROTASK DESIGN

Image labelling tasks published on microtask platform

- Free-text labels, varying numbers of labels per image, taboo words

Baseline: ‘standard’ tasks w/ basic spam control

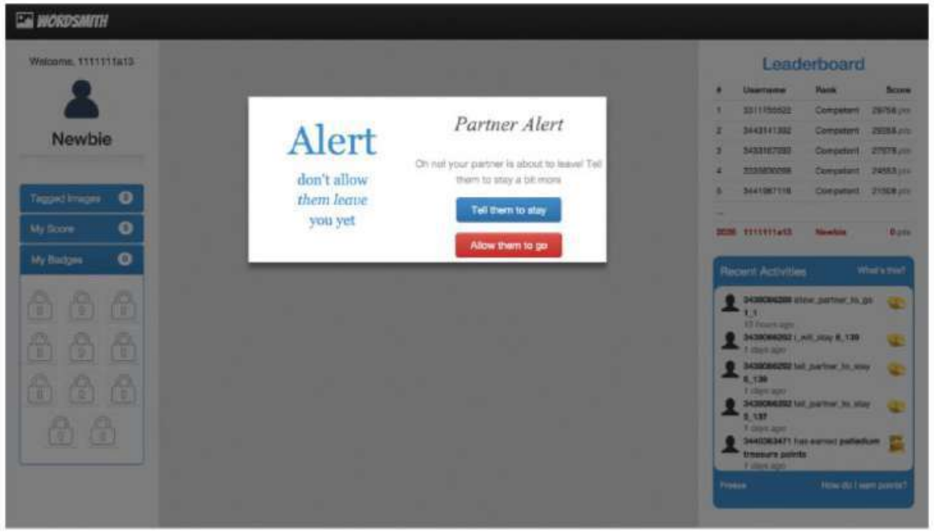
vs

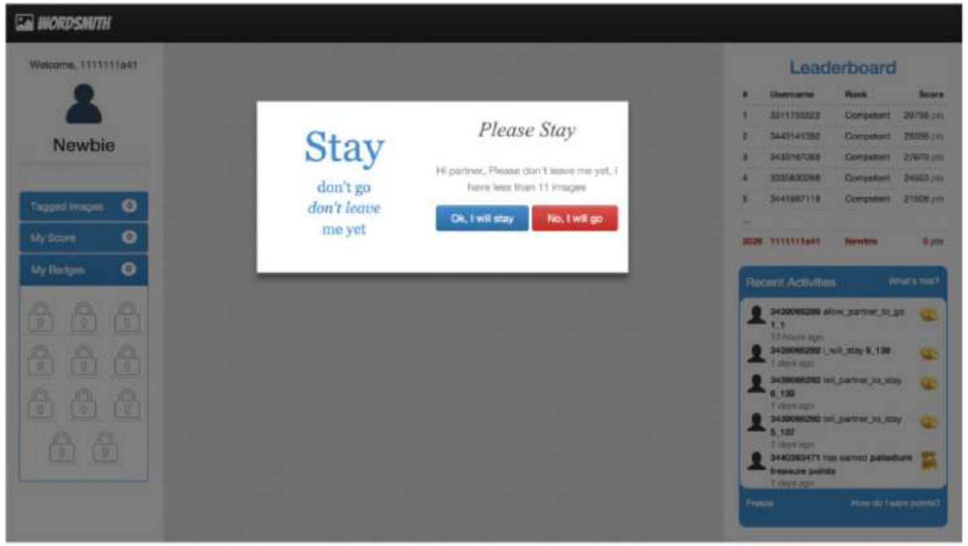
Pairs: Wordsmith-based, randomly formed pairs, people join and leave all the time, in time more partner switches

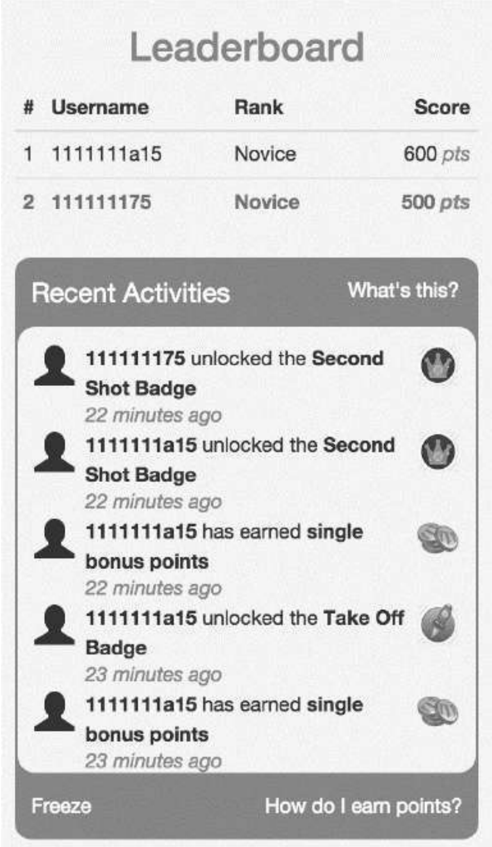
vs

Pairs & social incentives: let’s play vs please stay offered to worker when we expect their partner to leave

INCENTIVES







No global leaderboard

Empathic social pressure: stay (and help your partner get paid)

Social flow: keep playing and having fun together

EVALUATION

- ESP data set as gold standard
- Evaluated #labels, agreement, avg/max #labels/worker
- Two tasks
  - Low threshold: 1 image
  - High threshold: 11 images
- Probabilistic reasoning to predict worker exit* and offer social incentive

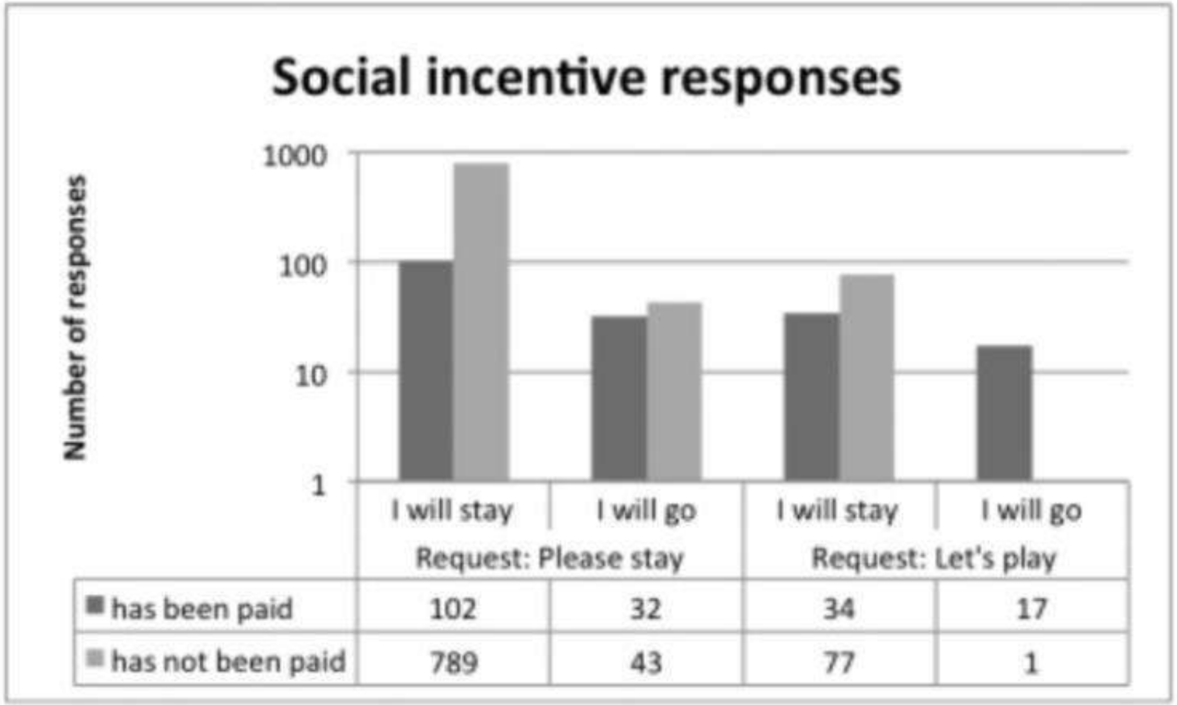
* [Kobren et al, 2015] extended w/ utility features

RESULTS (COLLABORATION)
BETTER, CHEAPER, FEWER WORKERS, ADDS COMPLEXITY

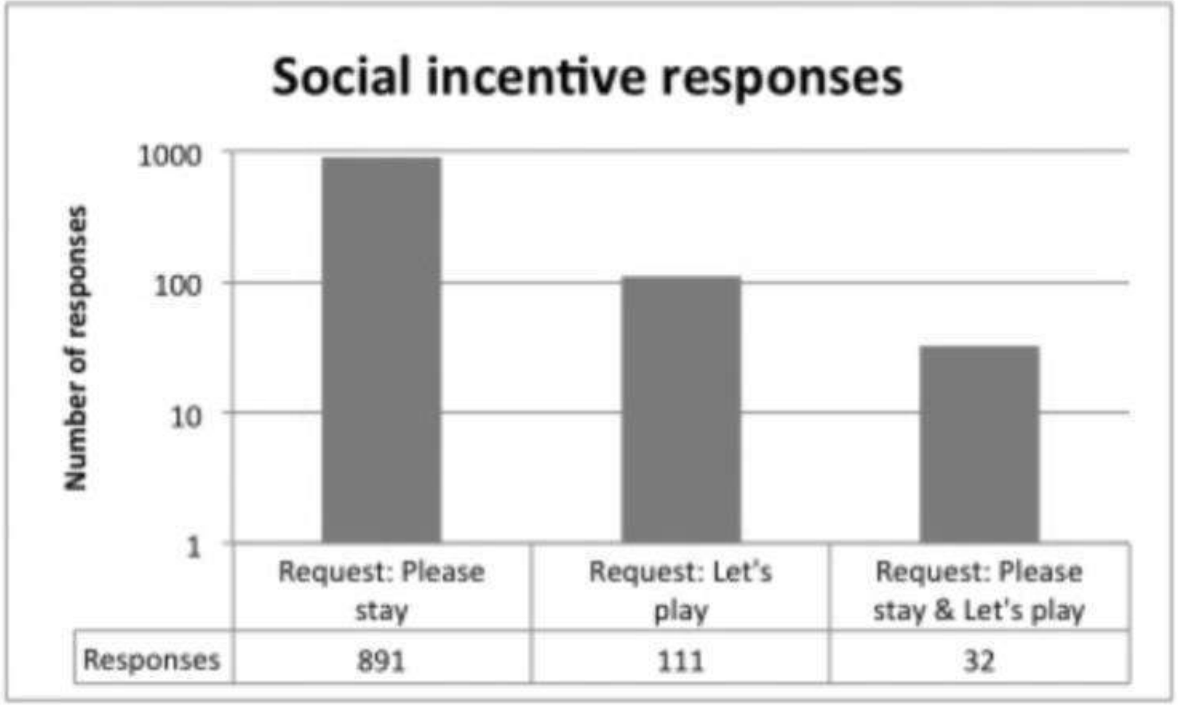
Experiment Results

	Low Threshold		High Threshold		
	Traditional	Collabo-rative	Traditional	Collabo-rative	Social Incentive
Total workers	402	365	514	499	508
Total tags	21,538	48,171	27,652	108,950	158,716
Unique images tagged	200	200	2,196	2,200	2,200
Inter-annotator	29.44%	**34.55%**	14.26%	25.82%	**29.35%**
ESP tags agreement	41.26%	25.39%	43.96%	37.94%	40.11%
Avg. images tagged / person	26.68 (SD=38.21)	9.77 (SD=13.23)	26.75 (SD=42.07)	25.05 (SD=17.92)	29.00 (SD=28.30)
Avg. tags / person	53.57	**131.97**	53.80	218.34	**312.43**
Avg. new tags / person	2.78 (1,117/402)	**8.69** (3,172/365)	1.80 (925/514)	11.83 (5,903/499)	**16.21** (8,236/508)

RESULTS (SOCIAL INCENTIVES)
IMPROVED RETENTION, PLEASE STAY MORE EFFECTIVE



(a) Breakdown of worker responses to *please stay* and *let's play* requests (on logarithmic scale)



(b) Number of *i will stay* responses including responses from both request types (on logarithmic scale)

SUMMARY OF FINDINGS

Social incentives generate **more tags** and **improve retention**

Social dynamics: **different responses if partner has been paid or not**

- Paid worker 76% more likely to stay after social pressure, unpaid worker: 95% more likely to stay
- Paid workers annotate more if they decide to stay than unpaid workers

Social flow more effective than **social pressure** in generating more tags: 99% of unpaid workers are likely to stay

Social pressure works more often overall

ONE DOES NOT SIMPLY CROWDSOURCE THE SEMANTIC WEB

CONCLUSIONS

With AI and ML on the rise, crowdsourcing is a critical for any Semantic Web developer

Explore the **what**, **who**, **how**, **why** design space

Use the full range of approaches and techniques to scale to large datasets