Bias in Human-in-the-loop Artificial Intelligence

Gianluca Demartini
demartini@acm.org
@eglu81
The University of Queensland, Australia

Research Interests

- **Entity-centric Information Access** (since 2005)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  - Entity Extraction (WWW 14), Prepositions (CIKM 14), Entity Cards (SIGIR 19)
  - IR Evaluation (IRJ 2015, ECIR 16 Best Paper, CIKM 17, SIGIR 18, CIKM 19)
- **Human-in-the-loop Information Systems** (since 2012)
  - Entity Linking (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - HumI systems overview (COMNET 15, FnT 17)
- **Better Crowdsourcing Platforms** (since 2013)
  - Platform Dynamics (WWW 15), Wikidata (CSCWJ 18)
  - Pick-a-Crowd (WWW 13), Scheduling Tasks (WWW 16)
  - Agreement (ICTIR 17, HCOMP 17), Pricing Tasks (HCOMP 14)
- **Human Factors in Crowdsourcing** (since 2015)
  - Malicious Workers (CHI 15), Attack Schemes (HCOMP 18 Best Paper, JAIR)
  - Modus Operandi (UBICOMP17, HT19, WSDM20), Bias (SIGIR18, ECIR20)
  - Time (HCOMP 16), Complexity (HCOMP 16), Abandonment (WSDM19, TKDE)
- **Data Quality** (since 2019)
  - Data Workers (SIGIR 20), Misinfo (SIGIR 20, CIKM 20), Know Graphs (ISWC 19)
  - Remove noise (WWW 19), Unknown Unknowns (ECAI 20)
  - User Behavior Embeddings (CIKM 20)

Thanks to:

Australian Government
Australian Research Council

European Commission

EPSRC
Engineering and Physical Sciences Research Council

facebook research

Outline

- Bias in Crowd-generated Data
  - Quality Control and Adversarial Attacks (HCOMP 2018 best paper + JAIR)
  - Wikidata editors and graph (CSCWJ + ISWC 2019)
  - Political bias (ECIR 2020, SIGIR 2020, CIKM 2020)
- Modelling Behavior
  - Logging Behaviors
  - Task Abandonment (WSDM 2019 + TKDE)
  - Experienced crowd workers (WSDM 2020)
  - Behavior embeddings (CIKM 2020)

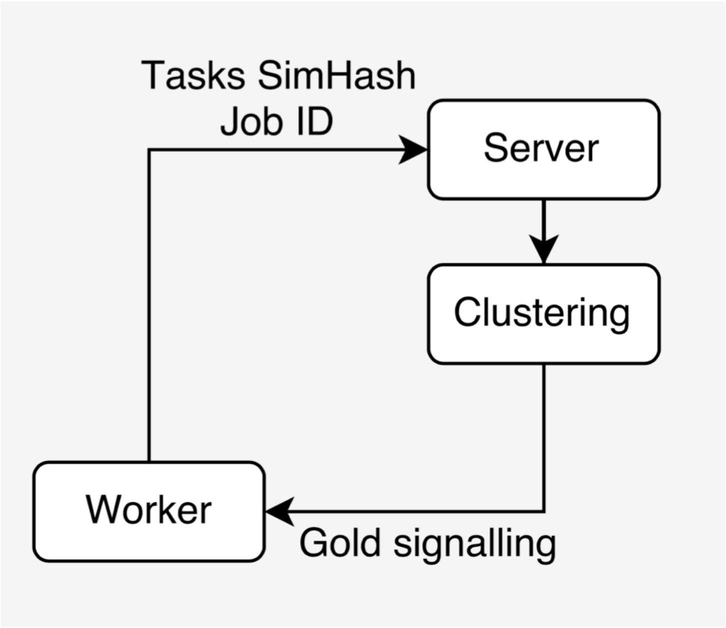
Crowdsourcing Quality control: Gold Questions

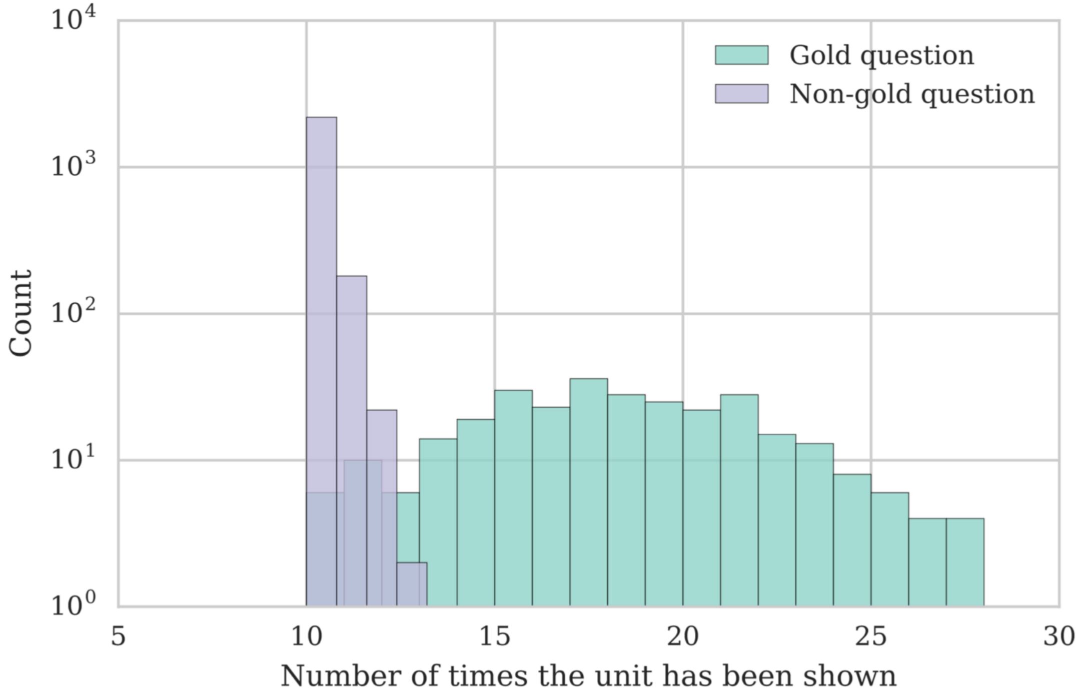
- Quality Control in Crowdsourcing
- Use known (ground truth) answers to check crowd answers
- If they answer correctly
  - we trust the other answers and use them
  - otherwise we discard them
- Randomly distributed
- **Indistinguishable by workers**
- **Very few available! (Expensive to generate) -> Repeated across different workers**

- Q1
- Q2
- Q3
- Q4
- Q5
- Q6
- Q7 <- Gold Question
- Q8
- Q9
- Q10

Power Imbalance - Gold Question Attacks

- Colluding workers sharing the questions they see can identify gold

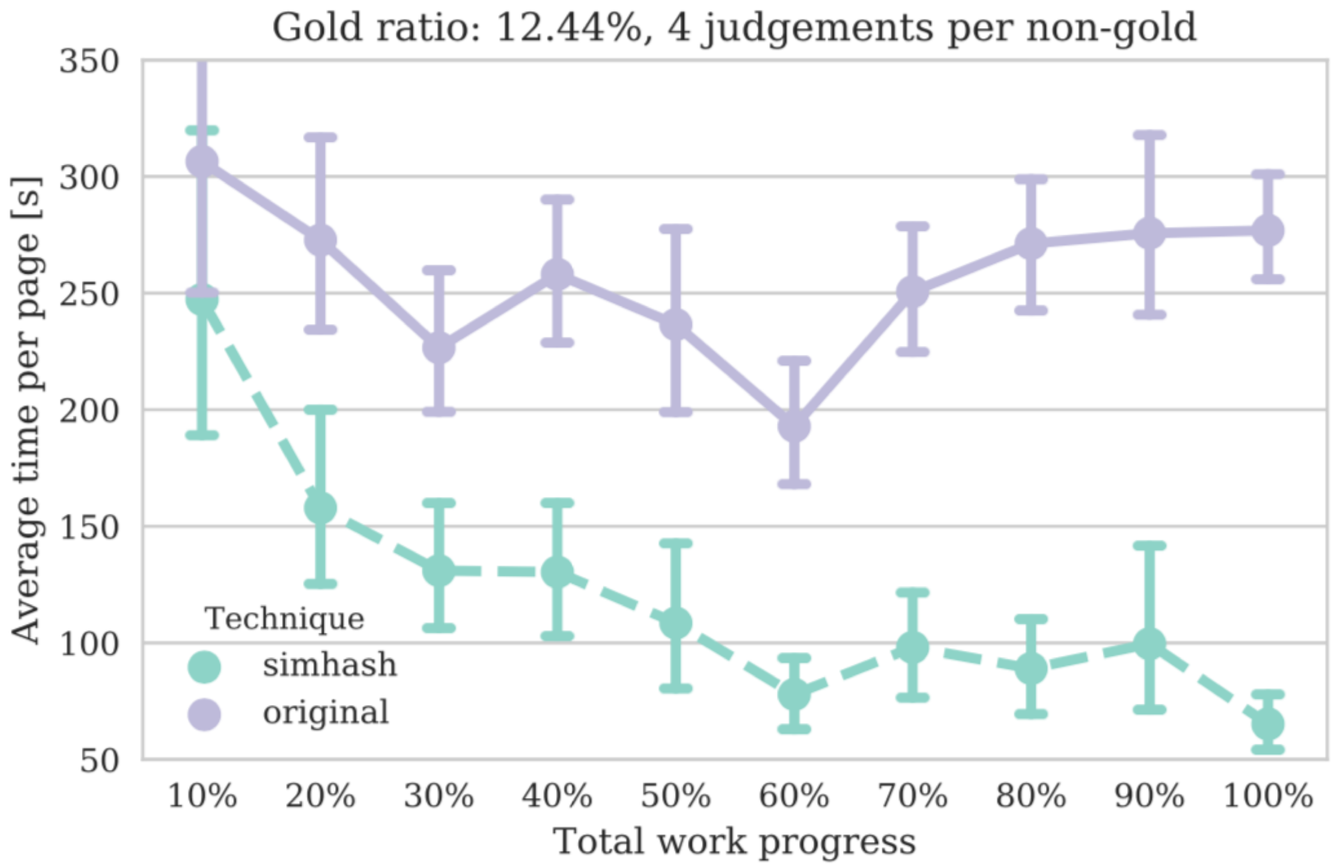




Alessandro Checco, Jo Bates, and Gianluca Demartini. Adversarial Attacks on Crowdsourcing Quality Control. In: **Journal of Artificial Intelligence Research** (JAIR). March 2020.

simhash – Gold Detection

- Time saved by workers with Gold Detection



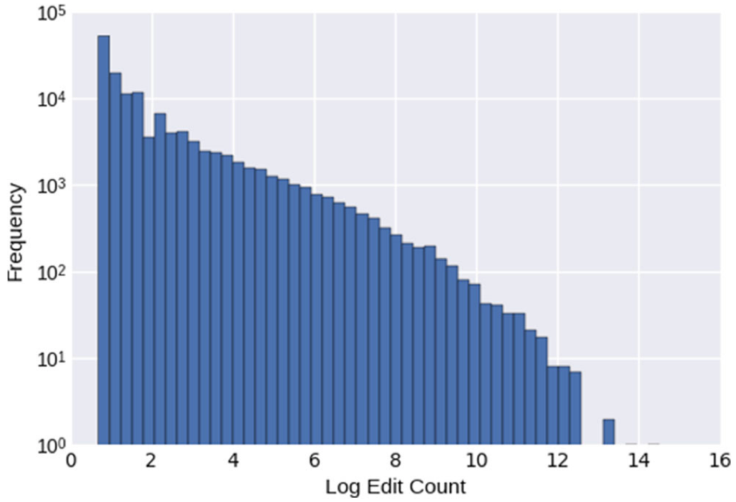
Countermeasures and implications

- Countermeasures
  - Increase gold set size
  - Increase worker retention (probability to see gold questions with high multiplicity is low)
  - Non uniform selection from the gold set
  - Programmatic gold questions (with distant simhashes)
- Implications - the future of crowd work
  - A shift towards different quality assurance approaches
  - Re-balancing in part the digital power imbalance
  - Trust between requesters and crowd workers
  - Bias in collected data

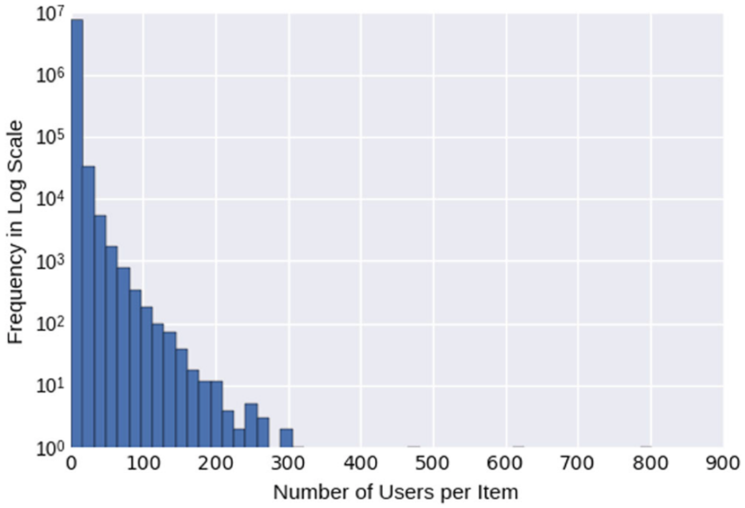
Knowledge Graph Editors

Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel Difallah, Michael Feldman, and Lydia Pintscher. **The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits**. In: Computer Supported Cooperative Work (CSCW) Special Issue on Crowd Dynamics: Conflicts, Contradictions, and Cooperation Issues in Crowdsourcing, Springer, 2018.

- The Wikidata edit history (2012-2016)
  - 35M (human) edits, 8M items, 140K editors
- In Wikidata we find shorter times between edits than in Wikipedia
- Why do certain editors have a lifetime longer than others?
  - **It's a habit**: Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not
  - **It's not boring**: Editors with a long lifespan tend to increase the diversity of the type of their edits



Total number of edits done by each Wikidata user.



Histogram of editors per item.

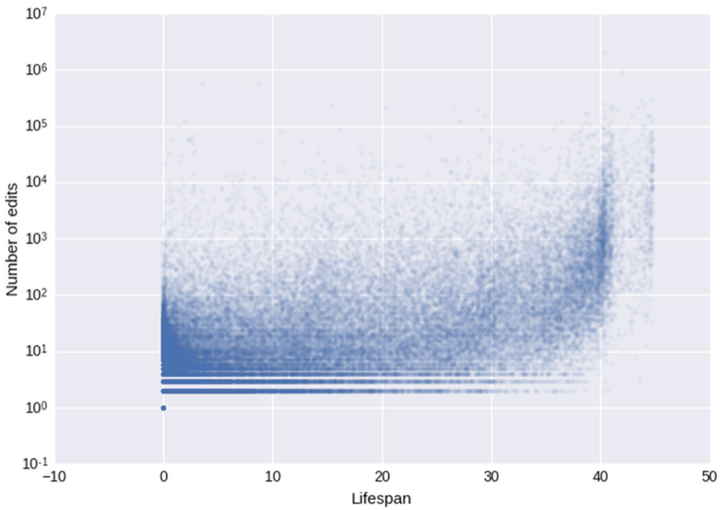
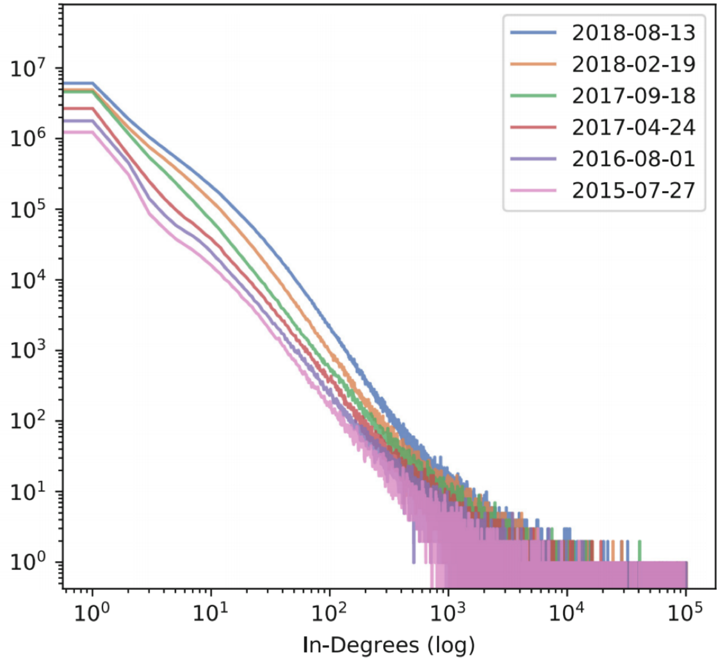


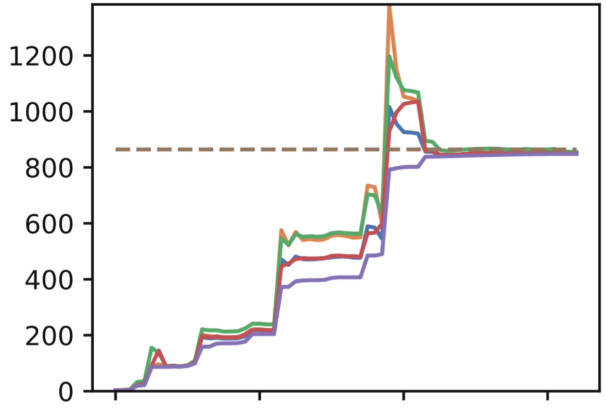
Figure 7. Number of edits vs lifespan.

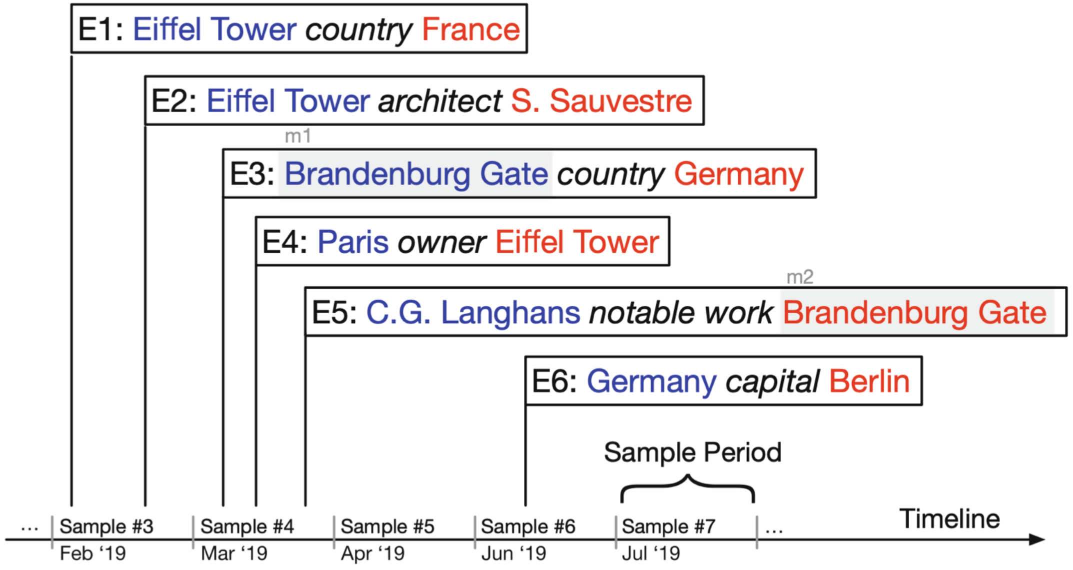
Knowledge Graph - Completeness



- Estimating Class Completeness
  - Do we have all the cities of Germany in the KG?
- Need to know class cardinality
  - Easy for US States, difficult for others (need to estimate)
- Estimation based on capture/recapture
  - Need sampling/mentions over time

(h) Paintings by Vincent van Gogh





Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs. In: The **International Semantic Web Conference** (ISWC 2019 - Research Track).

Crowdsourcing Truthfulness Judgements

- ~600 MTurk US workers
- To assess truthfulness of
  - US political statements (Politifact)
  - non-US political statements (ABC)
- 3 scales (3, 6, and 100 levels)

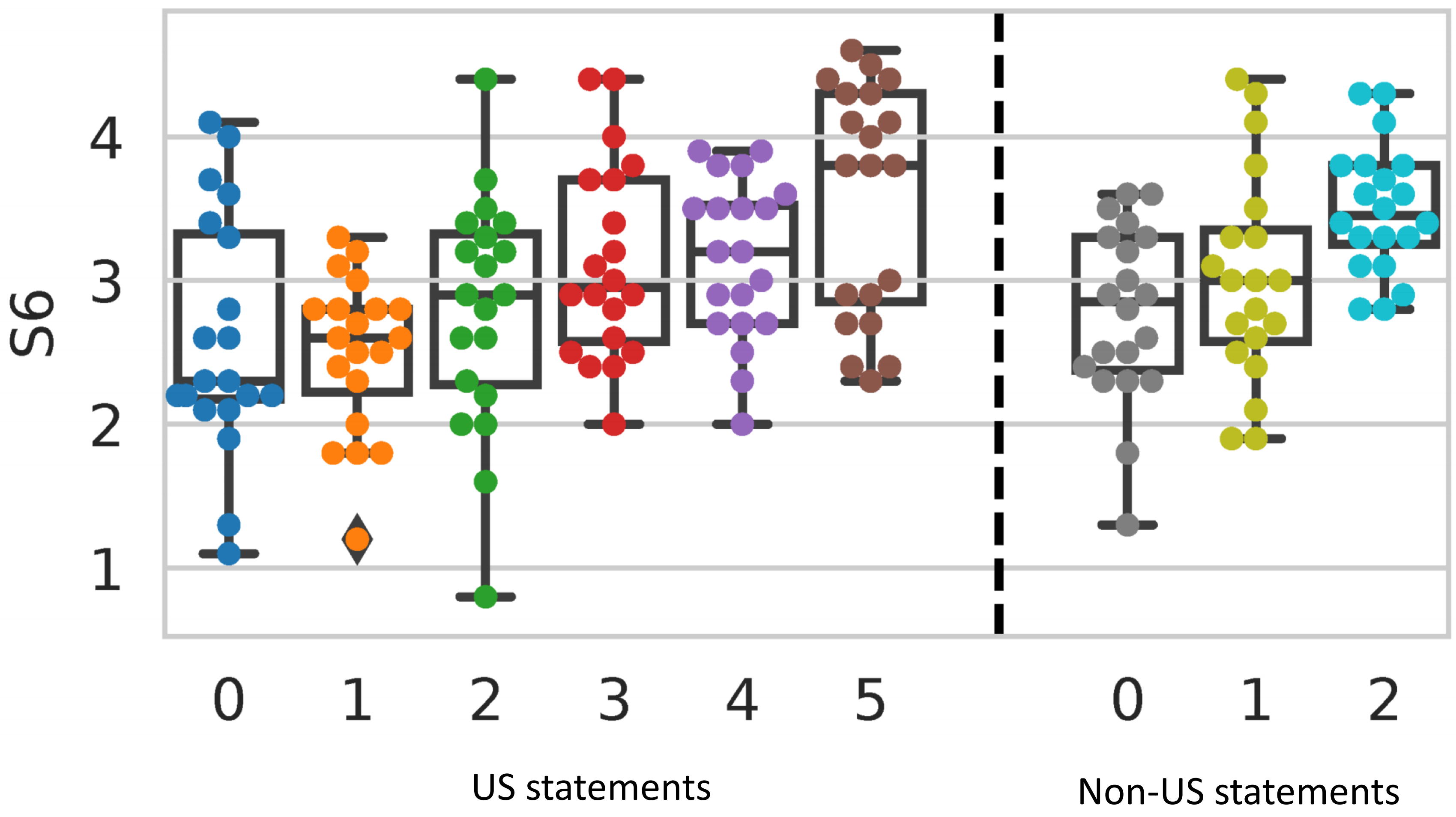
Table 1: Example of statements in the `PolitiFact` and `ABC` datasets.

	Statement	Speaker, Year
`PolitiFact` Label: `mostly-true`	"Florida ranks first in the nation for access to free prekindergarten."	Rick Scott, 2014
`ABC` Label: `in-between`	"Scrapping the carbon tax means every household will be $550 a year better off."	Tony Abbott, 2014

- All data:
- https://github.com/kevinRoitero/crowdsourcingTruthfulness

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro and Gianluca Demartini. **Can The Crowd Identify Misinformation Objectively? The Effects of Judgments Scale and Assessor's Bias**. In: The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)

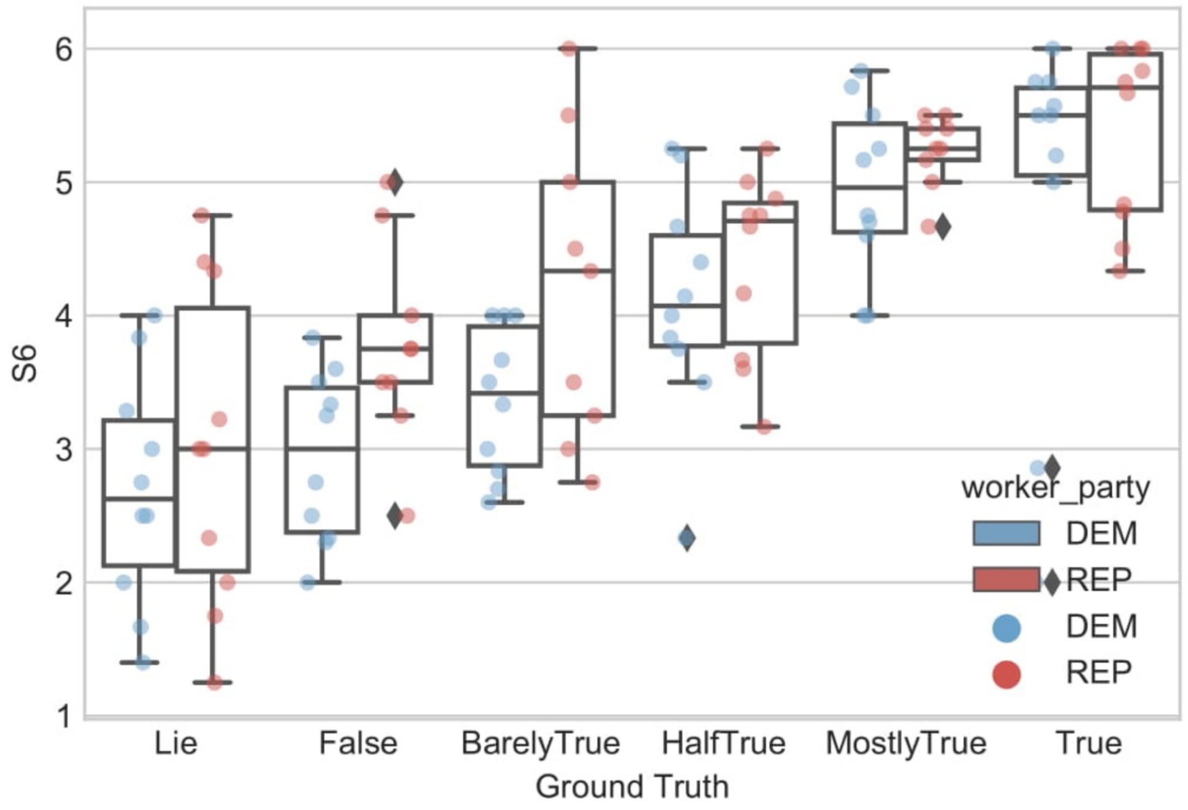
Crowd Performance VS Expert Ground Truth



Fake News labelling - Political bias

- Fact checkers are expert journalists verifying sources and validating news

- Can we (non-experts) do the same?
- Non-expert people who vote REP are more likely to believe to statements by REP politicians



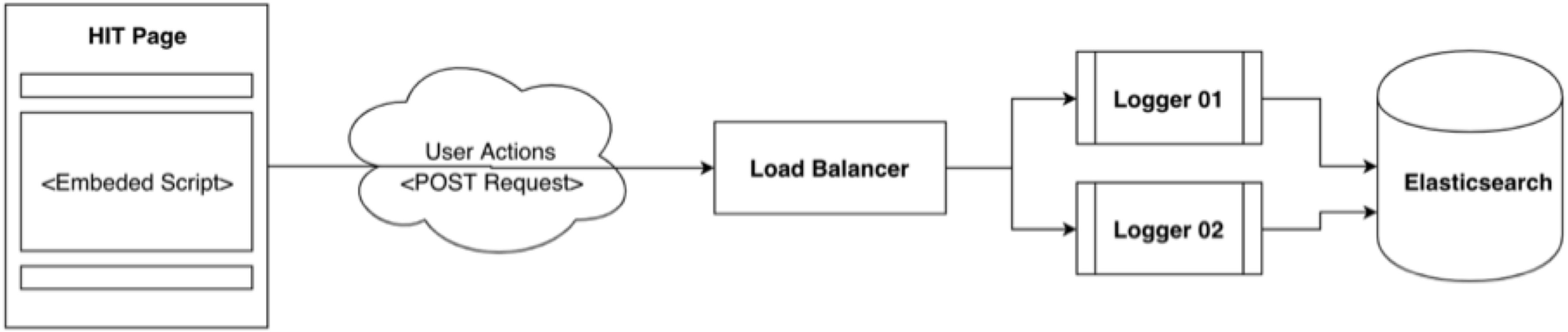
David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. **Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias**. In: The 42nd European Conference on Information Retrieval (ECIR 2020). Lisbon, Portugal, April 2020.

Outline

- Bias in Crowd-generated Data
  - Quality Control and Adversarial Attacks (HCOMP 2018 best paper + JAIR)
  - Wikidata editors and graph (CSCWJ + ISWC 2019)
  - Political bias (ECIR 2020, SIGIR 2020, CIKM 2020)
- Modelling Behavior
  - Logging Behaviors
  - Task Abandonment (WSDM 2019 + TKDE)
  - Experienced crowd workers (WSDM 2020)
  - Behavior embeddings (CIKM 2020)

Logging User Behaviors

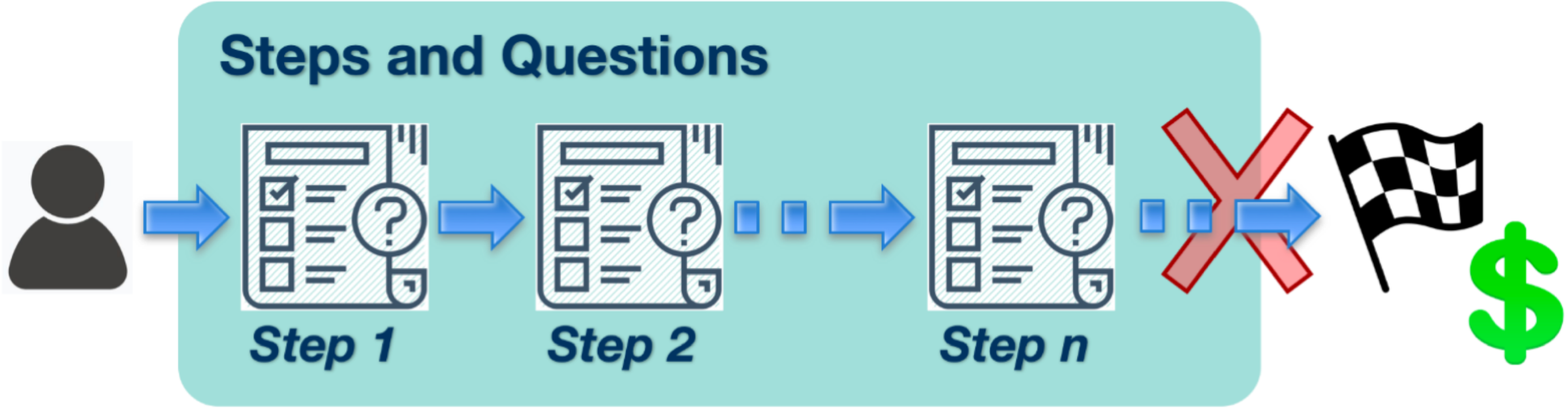
- UQCrowd Logging System
  - JS code embedded in the crowdsourcing tasks
  - Send msg (for every click, keystroke, scroll, new tab, etc.) to our server



- Observe user/worker online behaviors while they complete tasks
- https://github.com/d-lab/uqcrowd-log

Task Abandonment in Crowdsourcing

- Quantify task abandonment (i.e., **workers who start but don't finish a task**)
- 5K workers, 280K log entries over 4K documents
- Logged all actions and sent them to our external server before completion
- Total time not rewarded due to abandonment: 616 hours -> 3.5 months FTE



Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In: 12th ACM International Conference on Web Search and Data Mining (**WSDM 2019**). Melbourne, Australia, February 2019.

The Impact of Crowd Work Experience

- Survey + Interviews + Crowdsourcing (1’200 judgments, 154 workers)
- Findings:
  - Shortcuts (copy/paste) and reusing existing text -> reduce task time, increase wages!
  - Ctrl (Cmd) + F helps finding relevant keywords -> It’s not popular!
- Experienced workers:
  - reuse previous text more
  - are faster (but not better quality)
  - complete more tasks (participation bias)

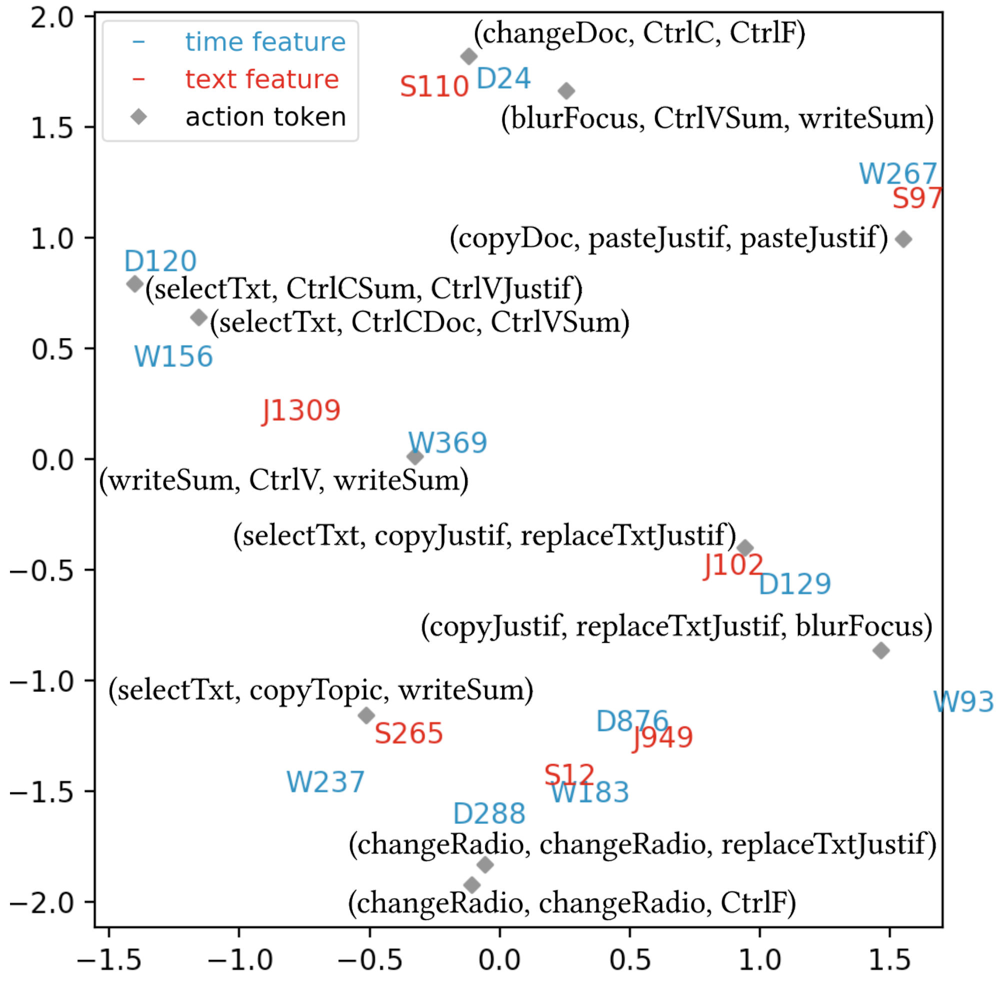
Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. Crowd Worker Strategies in Relevance Judgment Tasks. In: 13th ACM International Conference on Web Search and Data Mining (**WSDM 2020**). Houston, TX, USA, February 2020.

Behavior embeddings

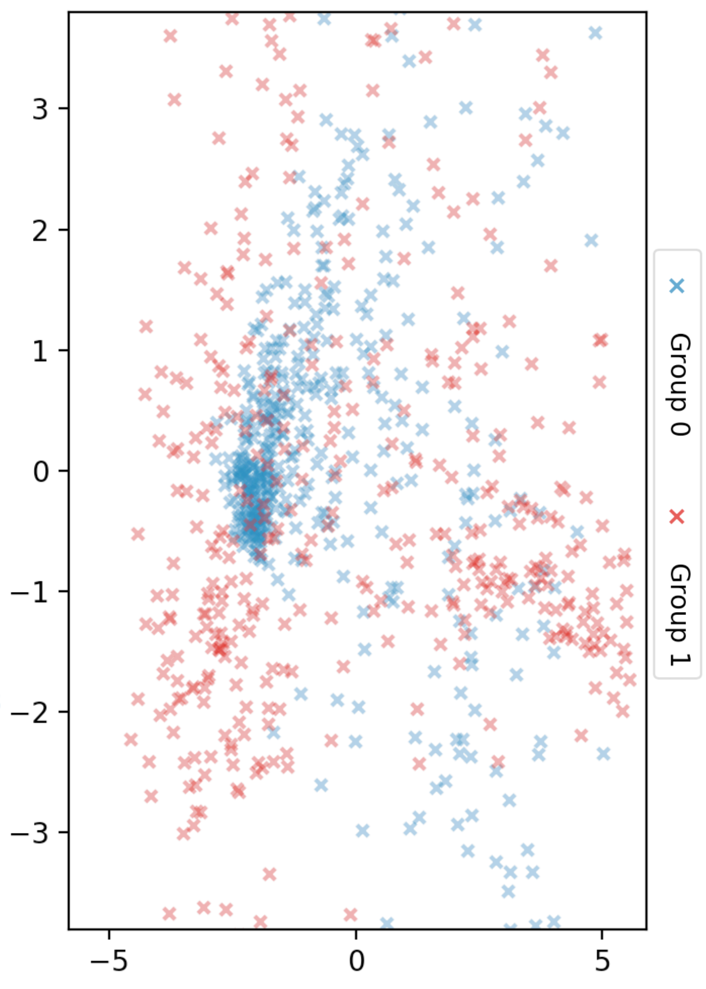
Order	Single Action	*n*-gram Token ($n = 2$)
1	`Ctrl+C`	(`Ctrl+C`, `Ctrl+V`)
2	`Ctrl+V`	(`Ctrl+V`, type characters)
3	type characters	(type characters, delete characters)
4	delete characters	(delete characters, click 'next')
5	click 'next'	—

- Model user behavior using embeddings
  - Raw actions from logs as sequences of tokens + CBOW
  - Vector representations of user behaviors
- Compare user behaviors (e.g., high performers / low performers)
- Changes over time
- Different time granularities

Lei Han, Alessandro Checco, Djellel E. Difallah, Gianluca Demartini, and Shazia Sadiq. Modelling User Behavior Dynamics with Embeddings. In: 29th ACM International Conference on Information and Knowledge Management (**CIKM 2020**).



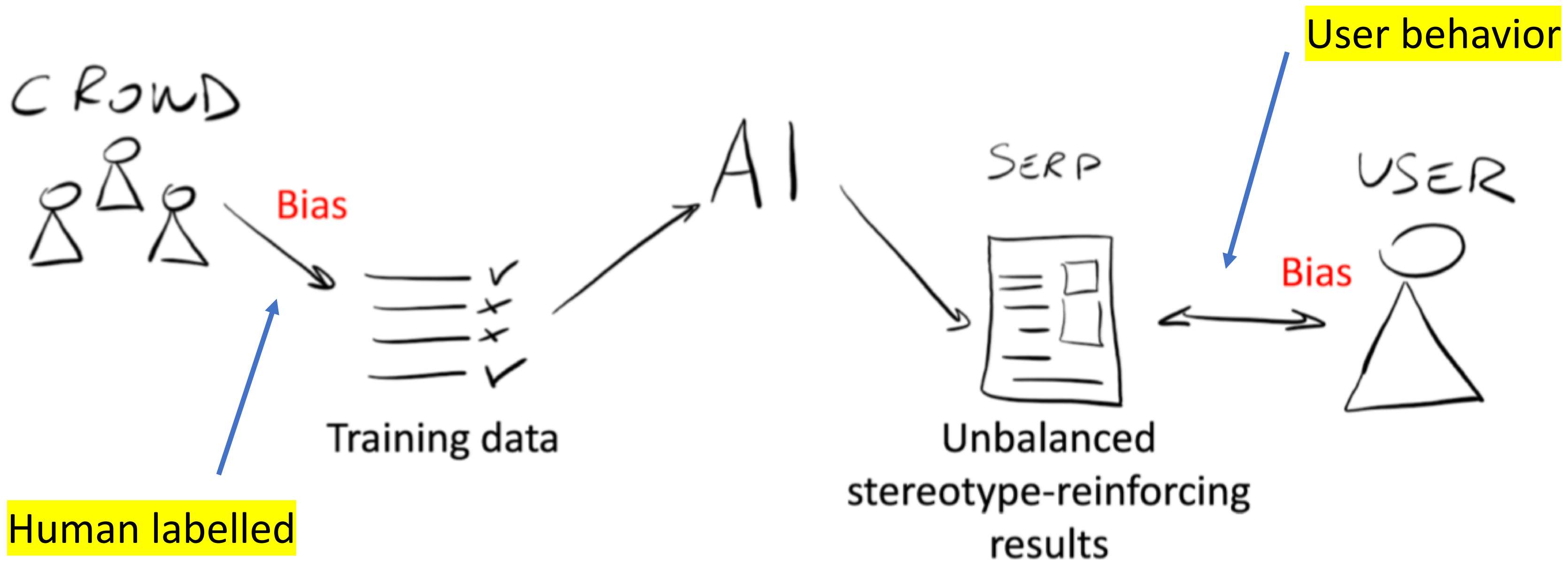
Crowdsourcing Task



WikiData: 0 less active; 1 more active

Datasets: https://github.com/tomhanlei/20cikm-behavior

Should AI systems reinforce stereotypes or rather break the bubble?



Hybrid Human-AI Approaches

- **Crowd workers** provide reliable (but not perfect) labels
- **AI** can provide reliable (but not perfect) labels
- **Experts** can provide perfect labels and justifications

- Can we leverage them all to work effectively and at scale?

Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. **Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities**. In: Data Engineering Bulletin, September 2020 issue.

Open Research Questions

- Who should do what?
  - Task allocation models
  - Cascade models: First AI to label at scale and quickly, then experts to “slowly” check the most important ones
- Urgency vs effectiveness
  - Identify difficult data items for expert to check and let “easy” ones for non-experts
- How would experts actually work when embedded in such a new framework
  - Trust in the hybrid system
  - Giving up levels of control: need for self-explainable human-in-the-loop AI tools

Summary

gianlucademartini.net
demartini@acm.org
@eglu81

- **Human-in-the-loop AI** systems can solve complex tasks at scale by combining
  - The ability of machines to scale over **very large amounts of data**
  - The quality of human intelligence and **manual content curation**
- Humans come with challenges
  - Data-driven (activity logging and log analysis) **behavior understanding**
  - System optimization (improving **efficiency and effectiveness**)
- Ongoing research
  - Better AI with humans to *pre-process* or *post-process* data
  - A combined expert-AI-crowd approach could provide the best scale/quality/urgency trade-off