Understanding Crowd Worker Behaviors

Gianluca Demartini

demartini@acm.org

@eglu81

The University of Queensland, Australia

Research Interests

- **Entity-centric Information Access** (2005-now)
  - 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** (2012-now)
  - Entity Linking (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Remove noise (WWW 19), Unknown Unknowns (ECAI 20)
  - Huml systems overview (COMNET 15, FnT 17)
- **Better Crowdsourcing Platforms** (2013-now)
  - Platform Dynamics (WWW 15), Wikidata (CSCWJ 18, ISWC 19)
  - Pick-a-Crowd (WWW 13), Scheduling Tasks (WWW 16)
  - Agreement (ICTIR 17, HCOMP 17), Pricing Tasks (HCOMP 14)
- **Human Factors in Crowdsourcing** (2015-now)
  - 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)

Thanks to:

European Commission

Australian Government
Australian Research Council

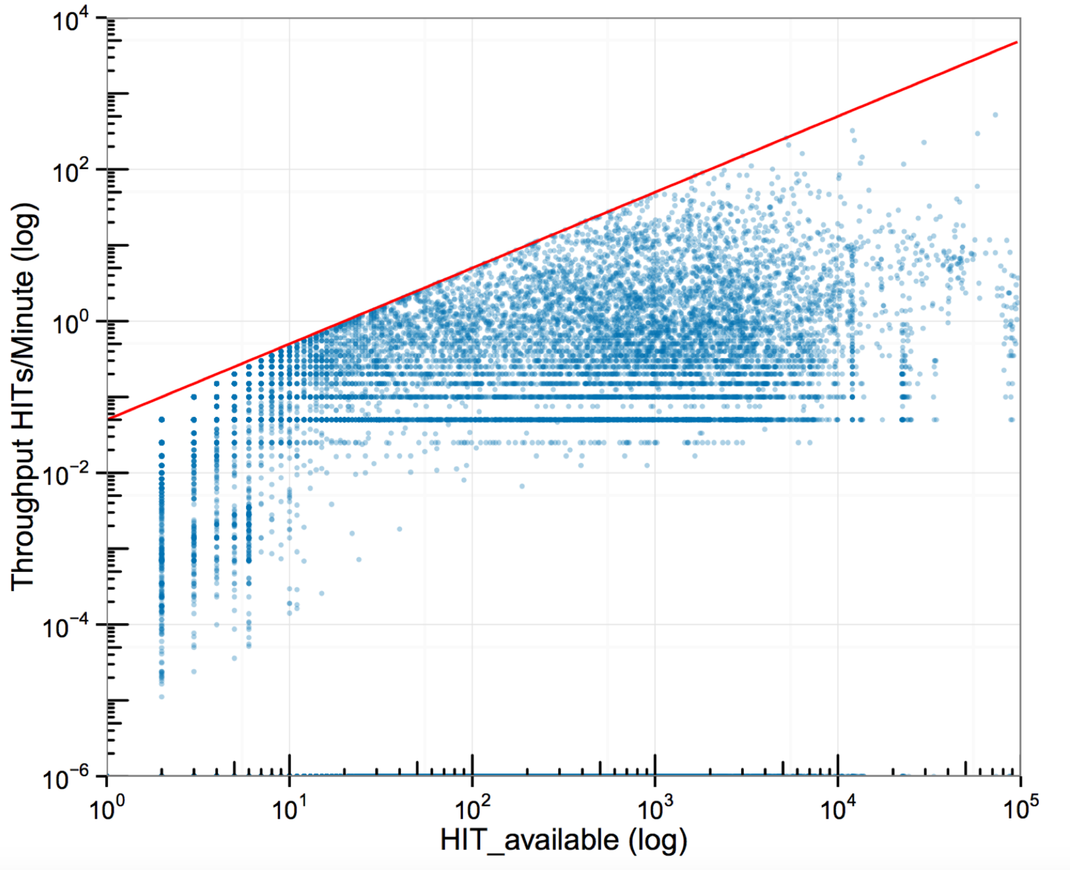
EPSRC
Engineering and Physical Sciences Research Council

facebook research

Outline

- Crowdsourcing and Crowd Workers
  - 5 years into Amazon Mturk (WWW 2015)
  - Malicious behaviours (CHI 2015)
  - Adversarial Attacks (HCOMP 2018 best paper + JAIR)
  - Wikidata editors and graph (CSCWJ + ISWC 2019)
- Worker Behaviors
  - Logging Behaviors
  - Modus operandi (UBICOMP 2017)
  - Task Abandonment (WSDM 2019 + TKDE)
  - Experienced workers (WSDM 2020)
- Worker Bias
  - Gender bias (SIGIR 2018)
  - Political bias (ECIR 2020)

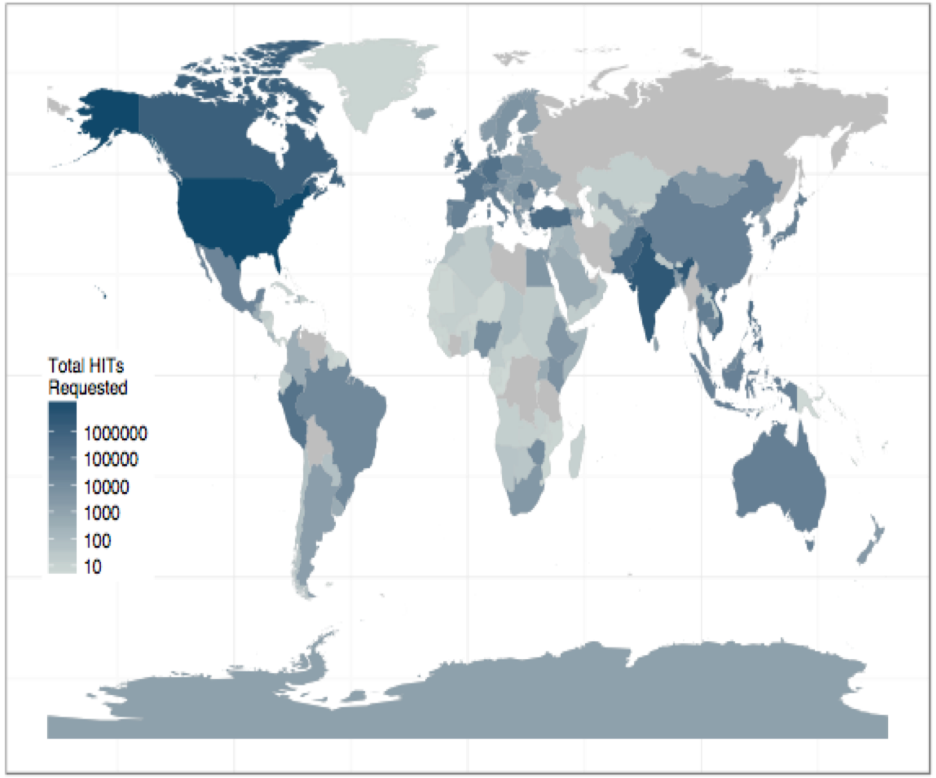
Amazon MTurk – A longitudinal study

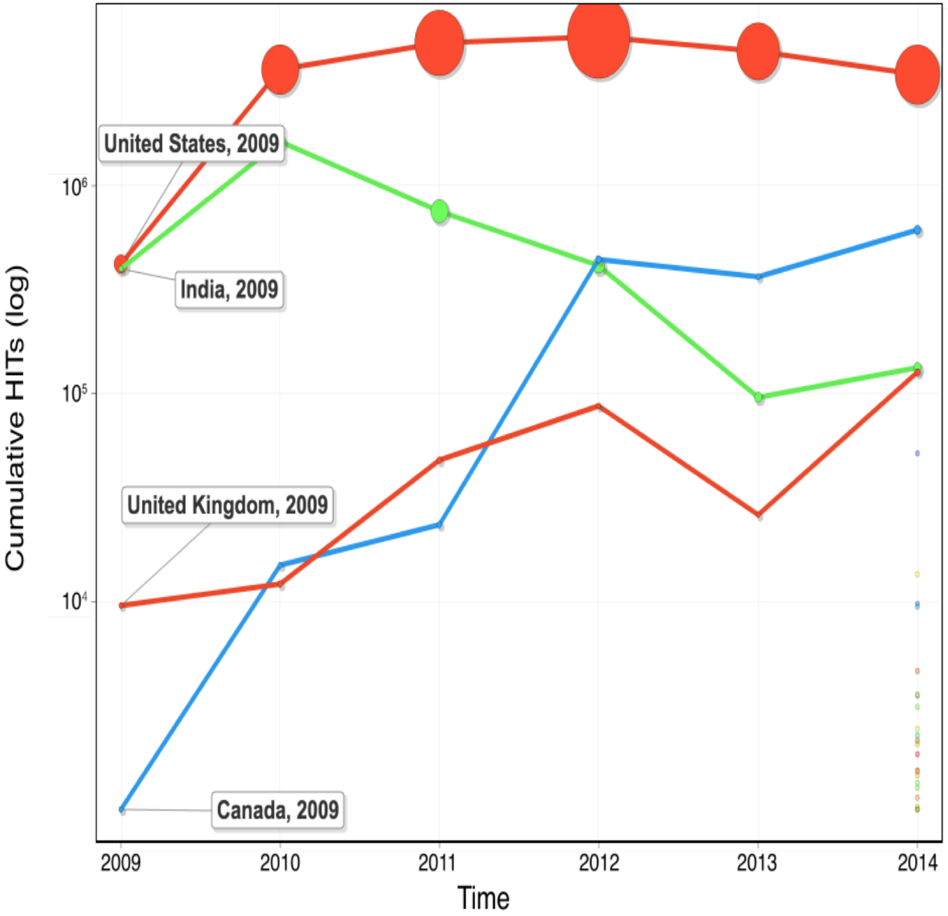


- Analyzed 130M Crowdsourcing Tasks
- Hourly aggregated data over 5 years (2009-2014)
- Reward, task types, platform throughput, market dynamics
  - 5-cents is the new 1-cent
  - Increasing number of new requesters
- Check #mturkdynamics for a summary

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.

Requested Workers





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.

Malicious workers

- CrowdFlower Platform to deploy survey
- Survey questions
  - Demographics
  - Educational & general background
- 34 Questions in total
  - Open-ended
  - Multiple Choice
  - Likert-type
- Responses from 1000 crowd workers
  - Monetary Compensation per worker : 0.2 USD

Gadiraju, Kawase, Dietze, and 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**)

RQ1 - Behavioral Patterns

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!

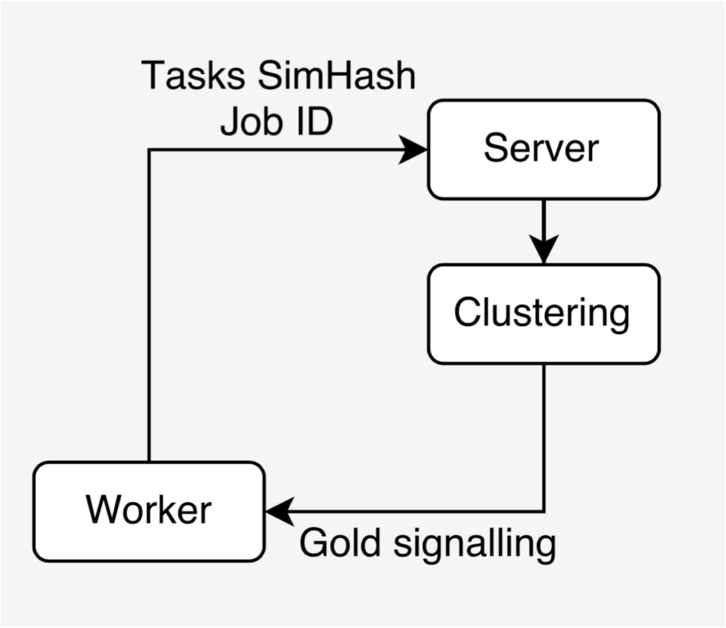
Crowdsourcing Quality control: Gold Questions

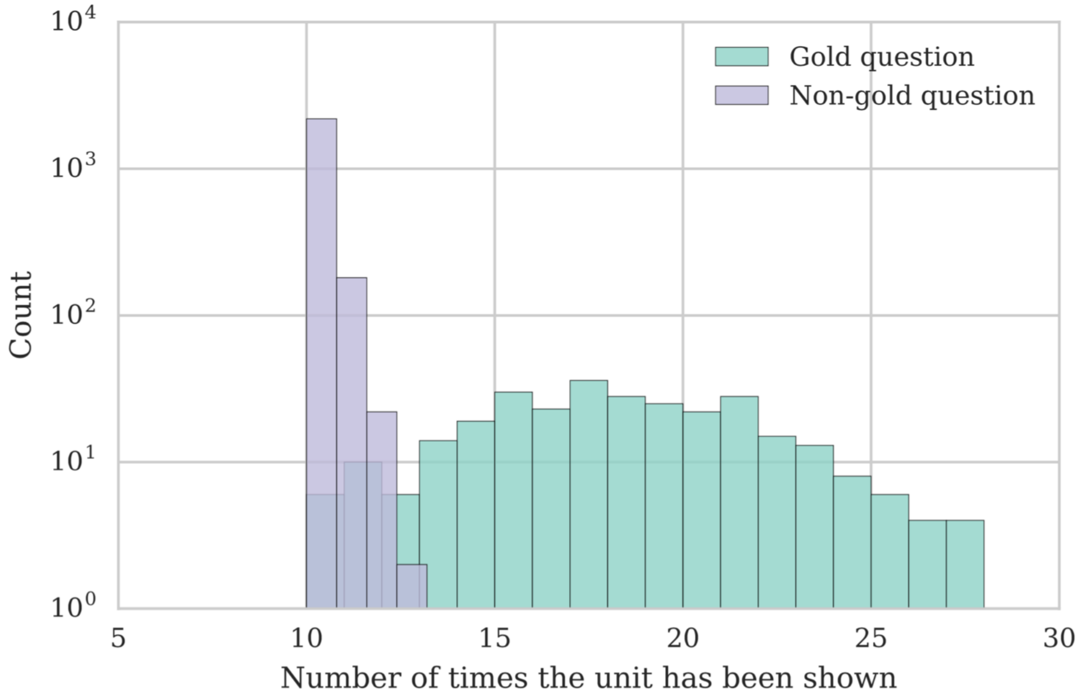
- Quality Control in Crowdsourcing
- Use known (ground truth) answers to check crowd answers
- If they answer correctly
  - we trust the other answers and use them
  - otherwise we discard them
- Randomly distributed
- **Indistinguishable by workers**
- **Very few available! (Expensive to generate) ->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

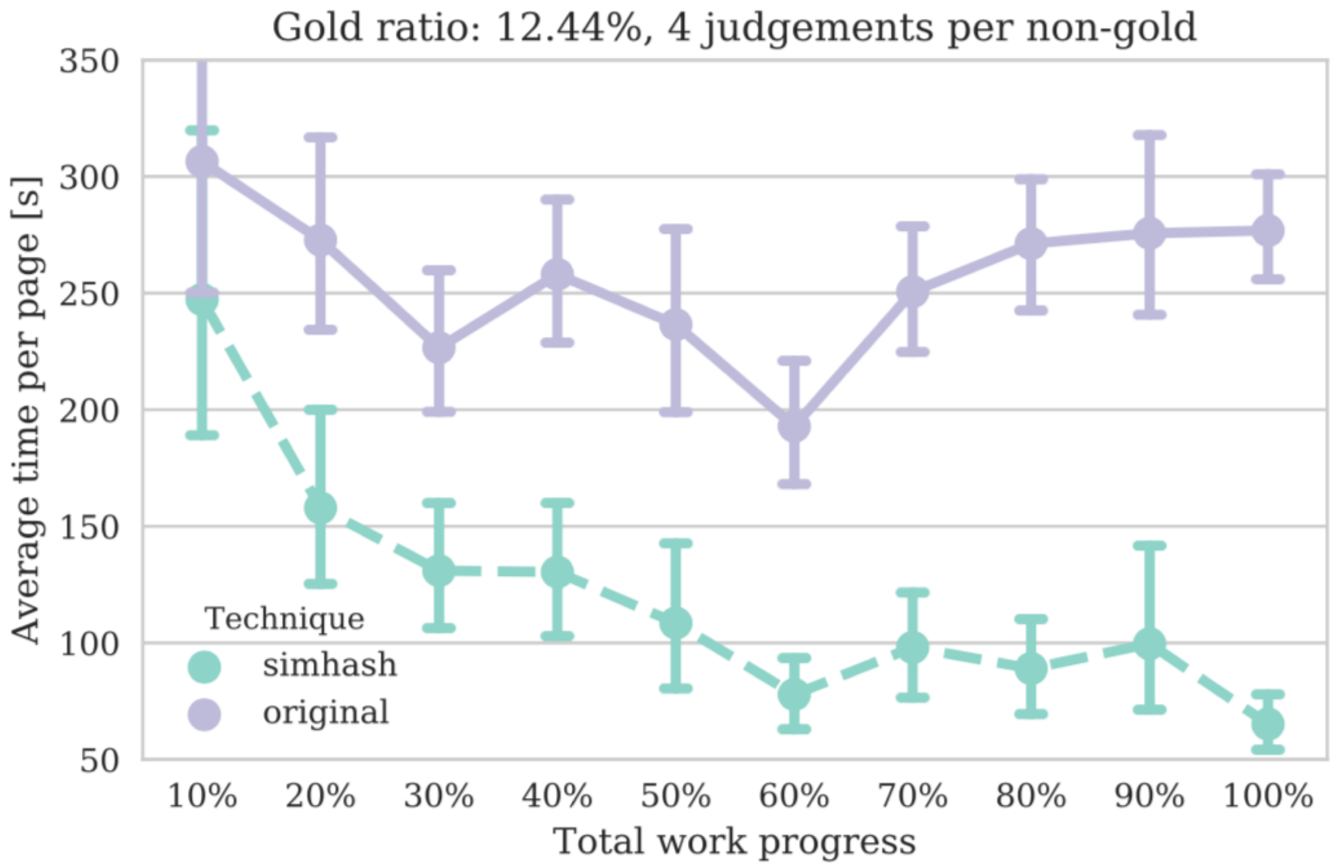




Alessandro Checco, Jo Bates, and Gianluca Demartini. Adversarial Attacks on Crowdsourcing Quality Control. In: **Journal of Artificial Intelligence Research** (JAIR). March 2020.

simhash – Gold Detection

- Time saved by workers with Gold Detection



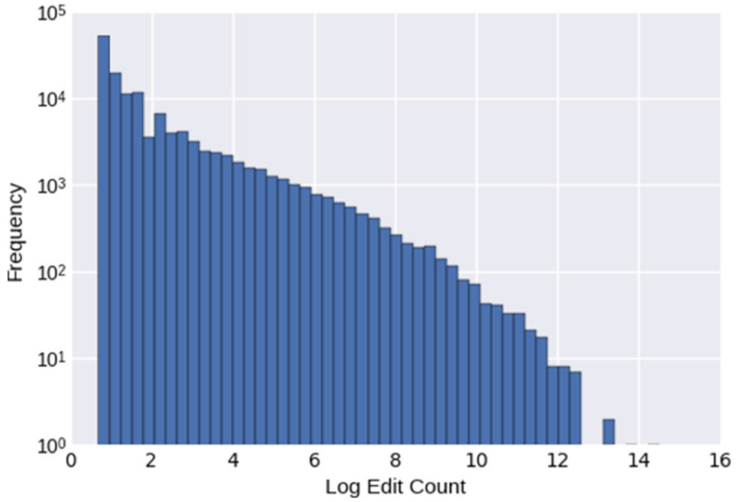
Countermeasures and implications

- 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

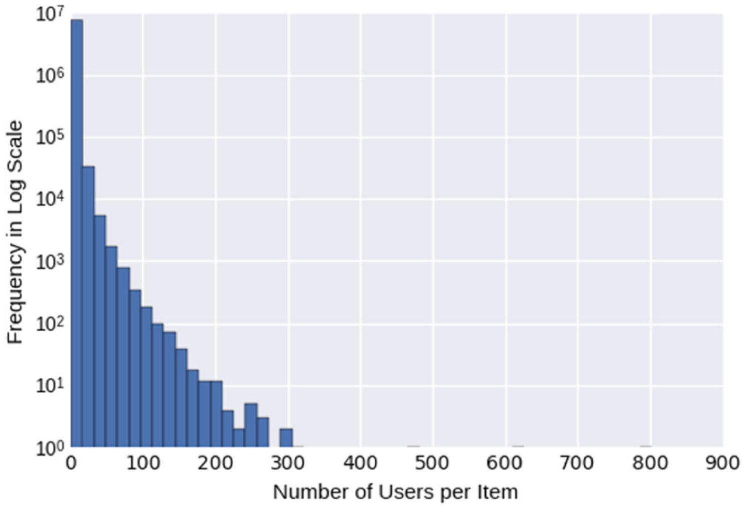
Knowledge Graph Editors

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.

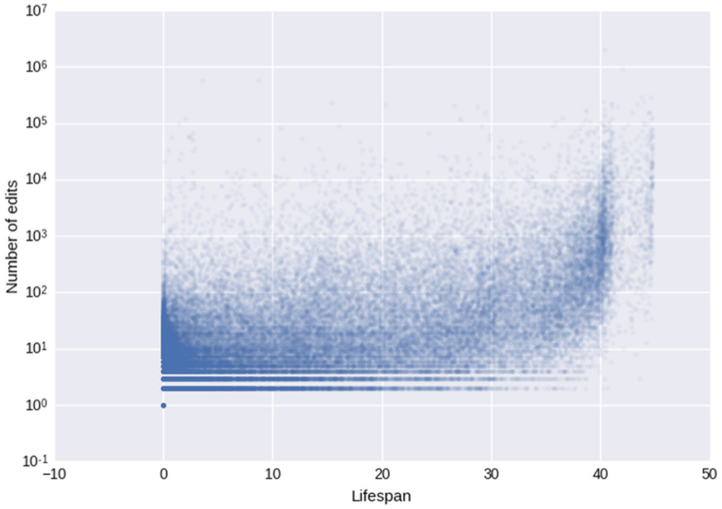
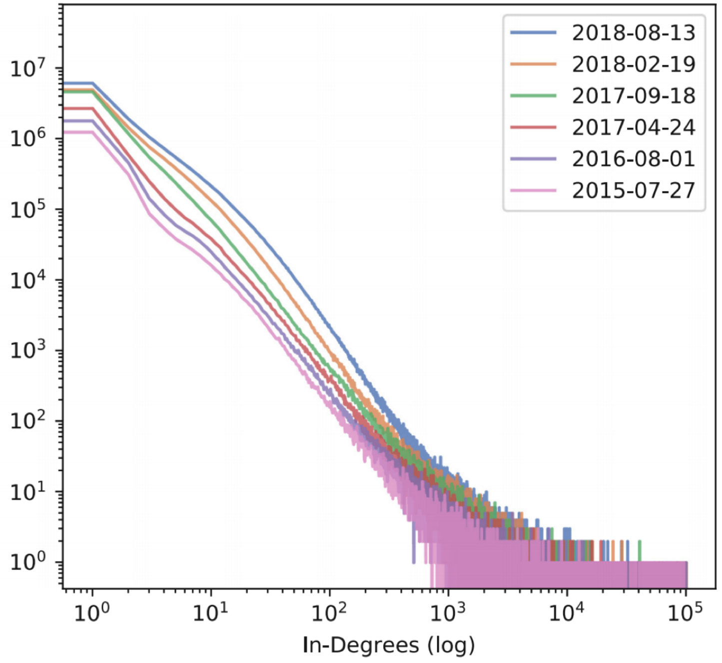
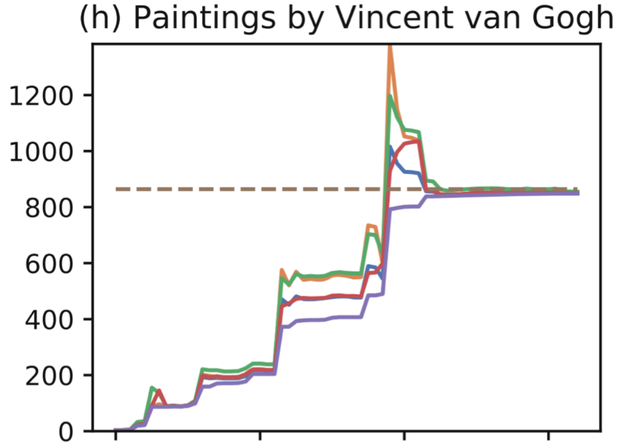


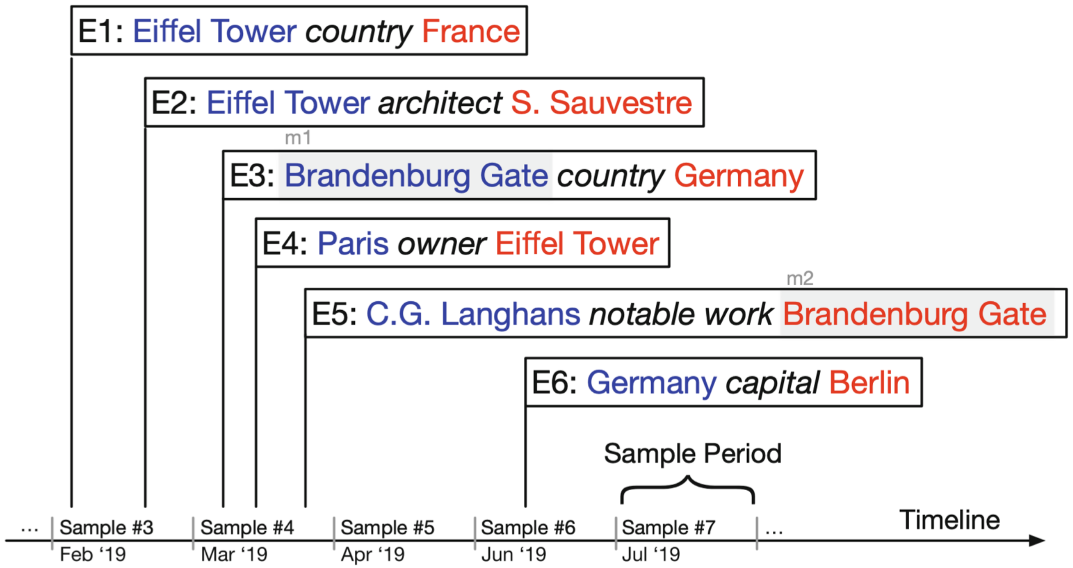
Figure 7. Number of edits vs lifespan.

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





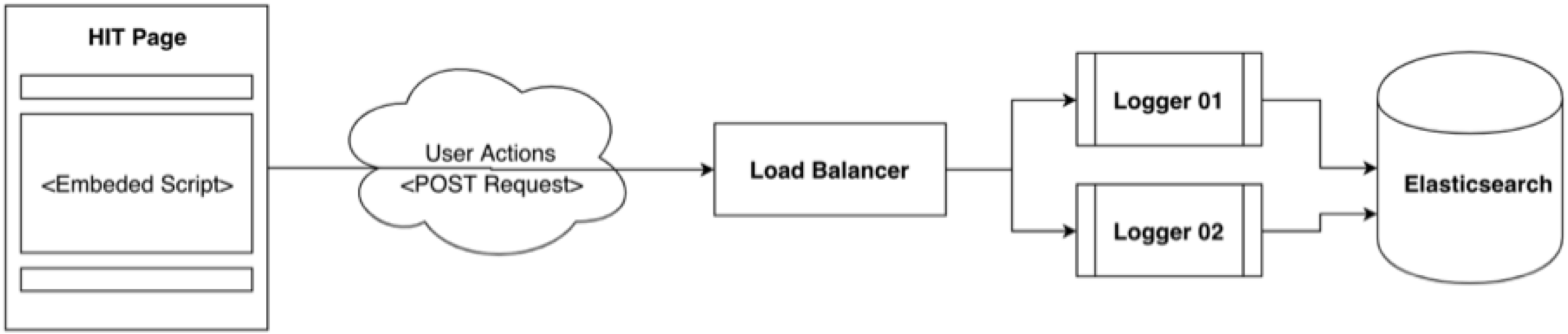
Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs. In: The **International Semantic Web Conference** (ISWC 2019 - Research Track).

Outline

- Crowdsourcing and Crowd Workers
  - 5 years into Amazon Mturk (WWW 2015)
  - Malicious behaviours (CHI 2015)
  - Adversarial Attacks (HCOMP 2018 best paper + JAIR)
  - Wikidata editors and graph (CSCWJ + ISWC 2019)
- Worker Behaviors
  - Logging Behaviors
  - Modus operandi (UBICOMP 2017)
  - Task Abandonment (WSDM 2019 + TKDE)
  - Experienced workers (WSDM 2020)
- Worker Bias
  - Gender bias (SIGIR 2018)
  - Political bias (ECIR 2020)

Logging User Behaviors

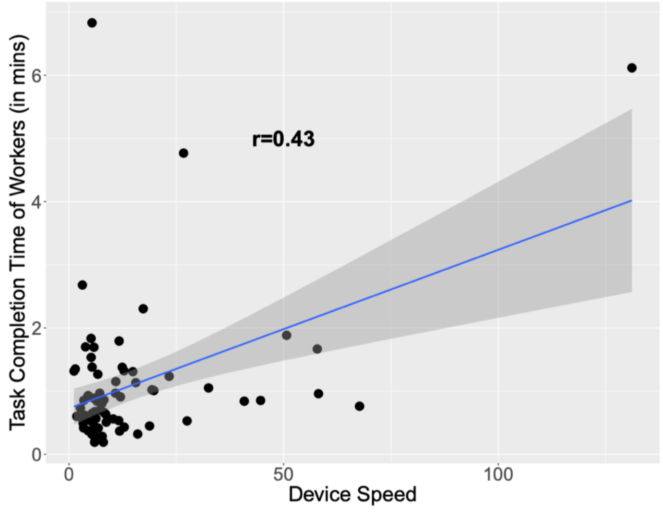
- UQCrowd Logging System
  - JS code embedded in the crowdsourcing tasks
  - Send msg (for every click, keystroke, scroll, new tab, etc.) to our server



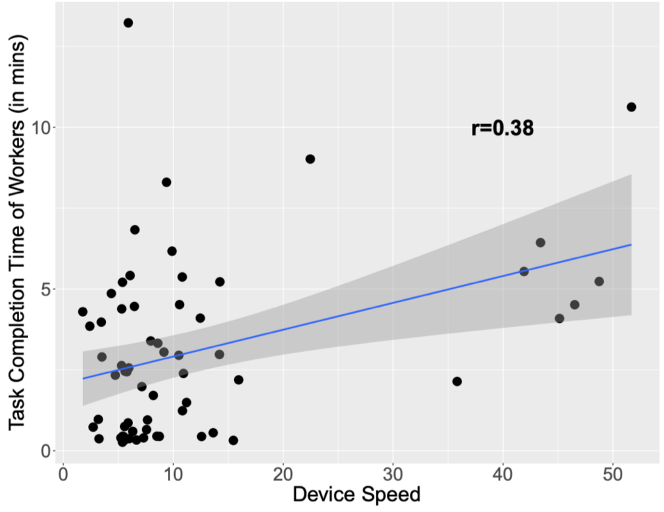
- Observe user/worker online behaviors while they complete tasks

The Impact of Crowd Work Environment

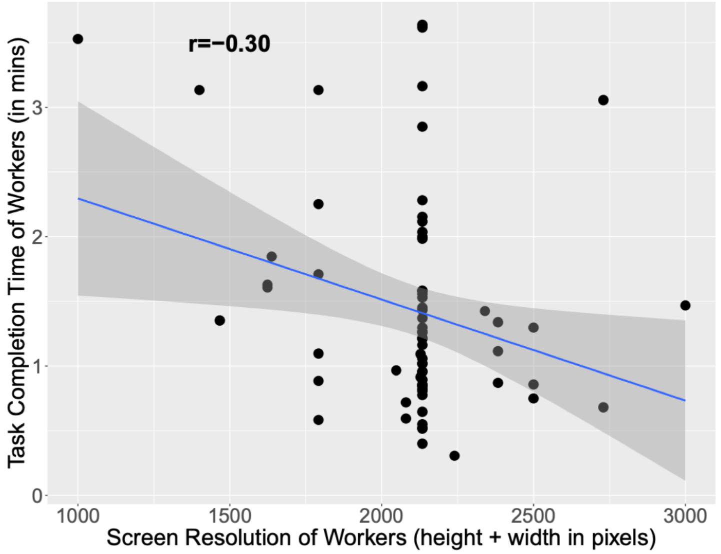
- Crowd workers use a diversity of devices and the quality of their working conditions varies dramatically (survey + interviews)
- How do microtask crowdsourcing work environments influence the quality of work produced by crowd workers? (data)



(a) *TCT* and *device speed* of American workers who completed tasks with *text area* variations.



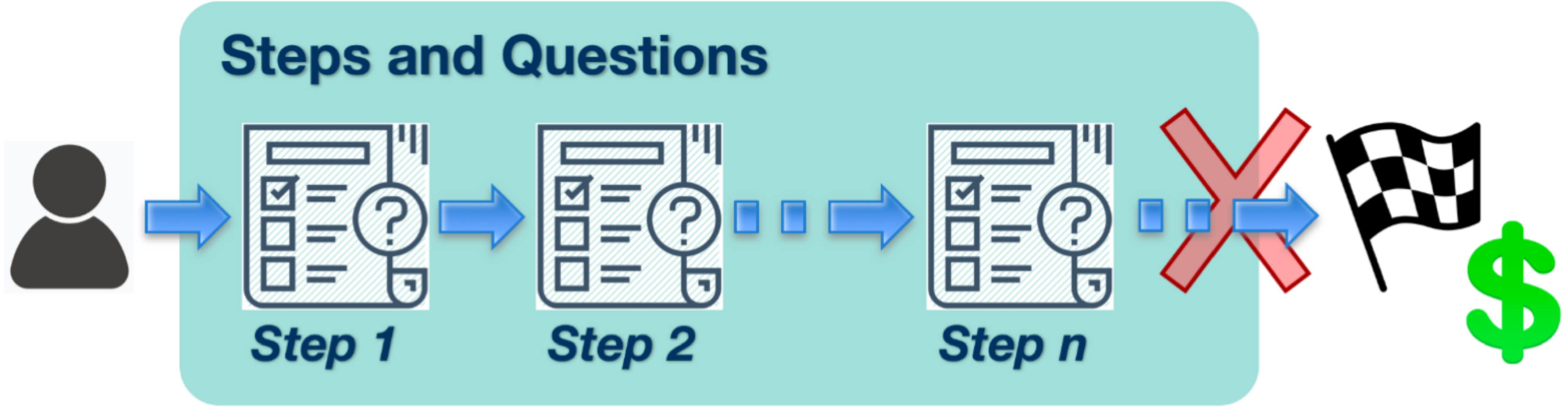
(b) *TCT* and *device speed* of American workers who completed tasks with *audio* variations.



Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) presented at The ACM International Joint Conference on Pervasive and Ubiquitous Computing (**UBICOMP 2017**)

Task Abandonment in Crowdsourcing

- Quantify task abandonment (i.e., workers who start but don't finish a task)
- 5265 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

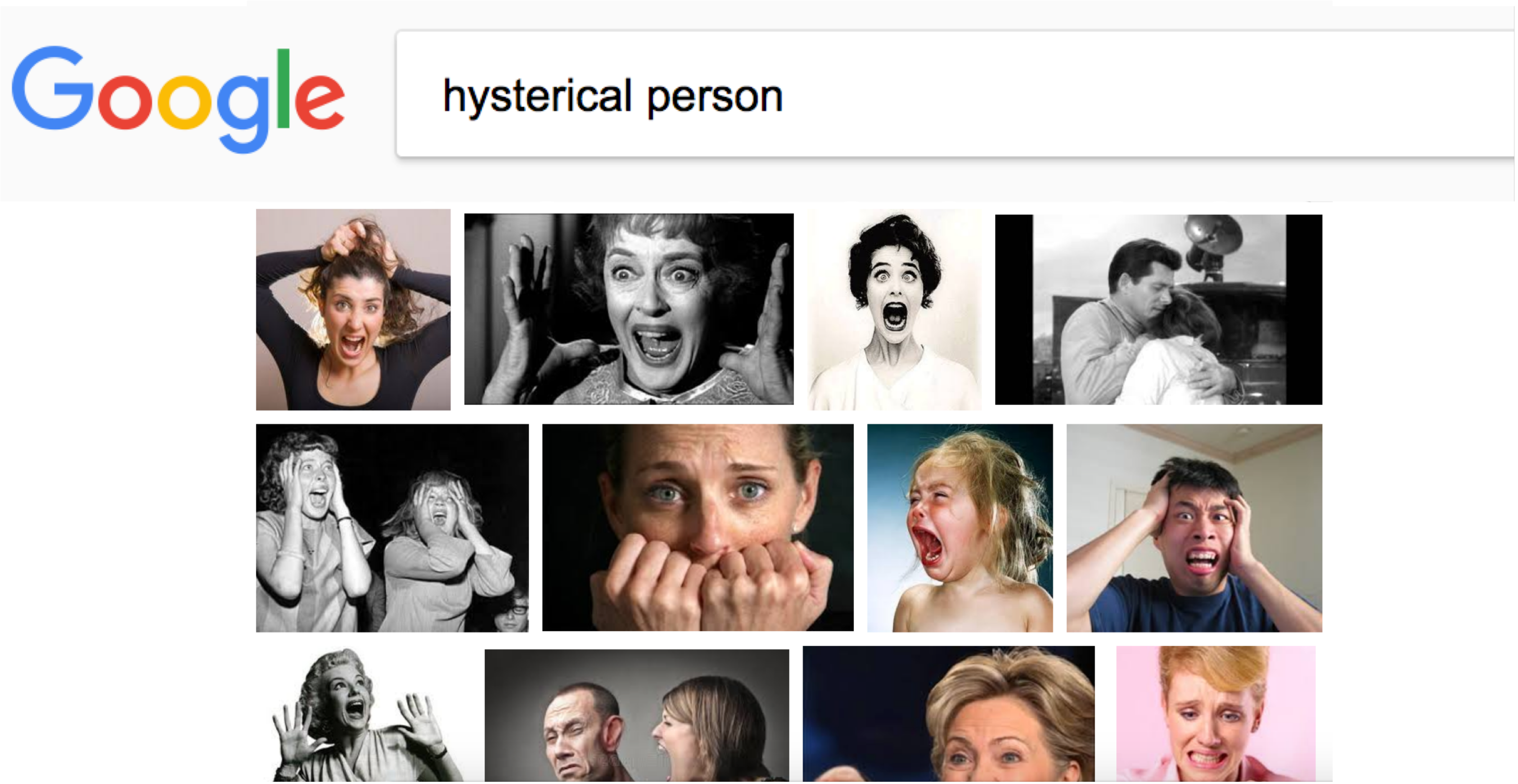
- Survey + Interviews + Crowdsourcing (1200 judgments, 154 workers)
- Findings:
  - Shortcuts (copy/paste) and reusing existing text -> reduce task time, increase wages!
  - Ctrl (Cmd) + F helps finding relevant keywords -> It’s not popular!
- Experienced workers:
  - reuse previous text more
  - are 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.

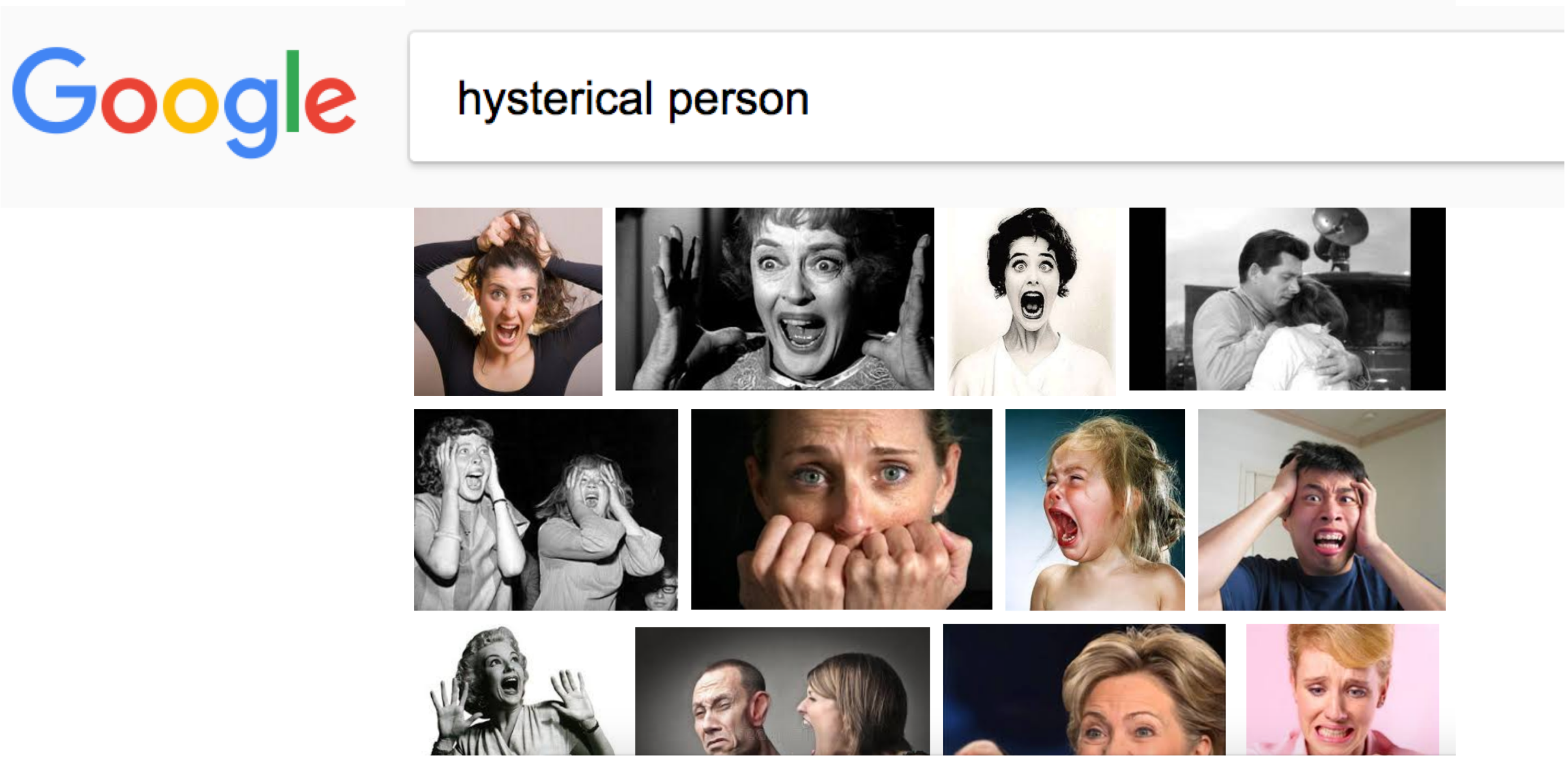
Outline

- Crowdsourcing and Crowd Workers
  - 5 years into Amazon Mturk (WWW 2015)
  - Malicious behaviours (CHI 2015)
  - Adversarial Attacks (HCOMP 2018 best paper + JAIR)
  - Wikidata editors and graph (CSCWJ + ISWC 2019)
- Worker Behaviors
  - Logging Behaviors
  - Modus operandi (UBICOMP 2017)
  - Task Abandonment (WSDM 2019 + TKDE)
  - Experienced workers (WSDM 2020)
- Worker Bias
  - Gender bias (SIGIR 2018)
  - Political bias (ECIR 2020)

Search results are biased/imbalanced (CHI 17)



How do users perceive them? – Gender bias



Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR 2018**). Ann Harbor, Michigan, July 2018.

Research Questions

- **RQ1**: Are **sexist/non-sexist people** less/more likely to evaluate a heavily gender-imbalanced result set as being subjective?
- **RQ2**: Is there evidence that sexist/non-sexist people **perceive a given image result set** differently?

Methods

- Ambivalent Sexism Inventory (ASI) – 22 questions
  - Hostile Sexism (HS) and Benevolent Sexism (BS)
- Assess perceived bias
  - Reverse image search: we retrieve images through a search engine, and ask the users to describe them ("guess the query").
- Crowdsourcing Task
  - Part 1 (guess the query)
  - Part 2 (search engine opinions) – do search engines give biased results?
  - Part 3 (perceived bias) – compare the real query with yours
  - Part 4 (ASI)

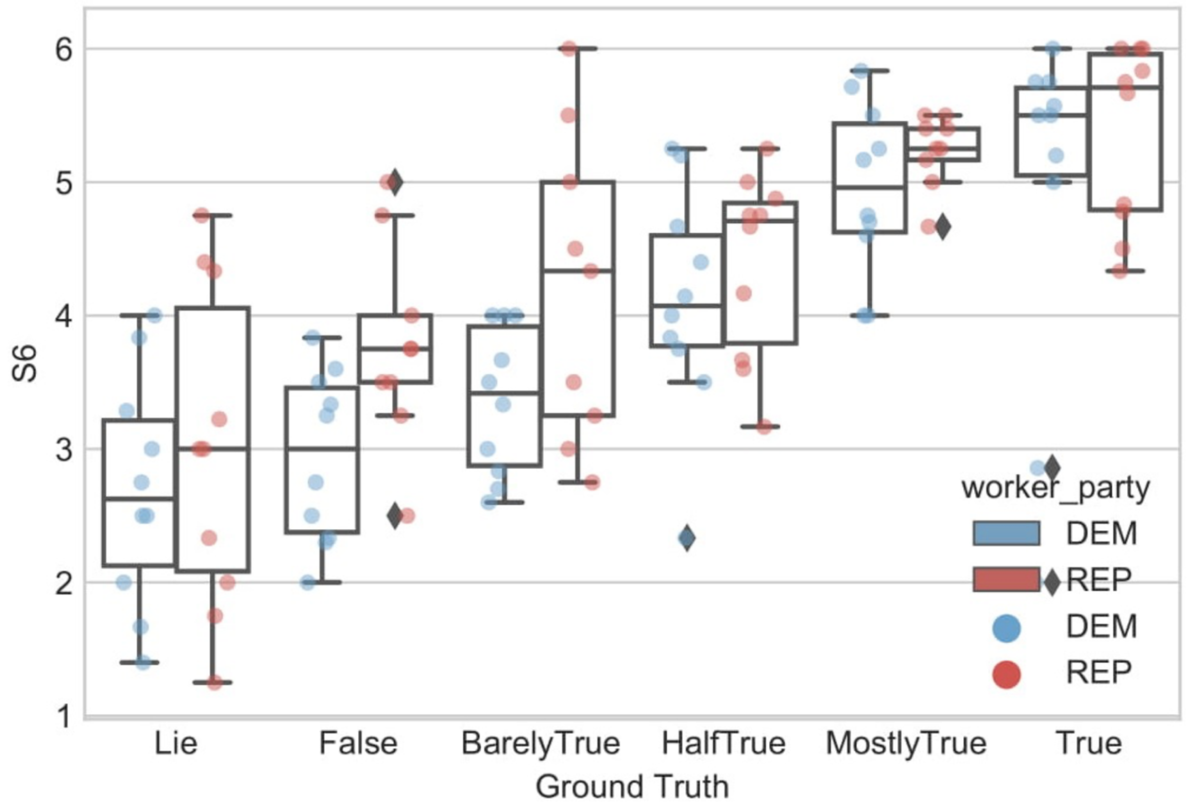
Experimental Results

- ASI: **Regional and gender differences**
  - Men scored higher than women on both BS and HS
  - India > US > UK
- Is sexism directly correlated to bias evaluation? Yes
  - Benevolent sexists are less likely to consider biased images for “smart person” or “warm person,” which primarily features images of men/women respectively
  - Benevolent sexists hold positive, yet traditional views of women
- Do sexists perceive results differently? Yes
  - Users who are more sexist, perceive image results differently than non-sexist people, and are less likely to perceive gender-biased results sets.
- **People who are more sexist are less likely to recognise gender biases in image search results and thereby reinforce social stereotypes**

Fake News labelling - Political bias

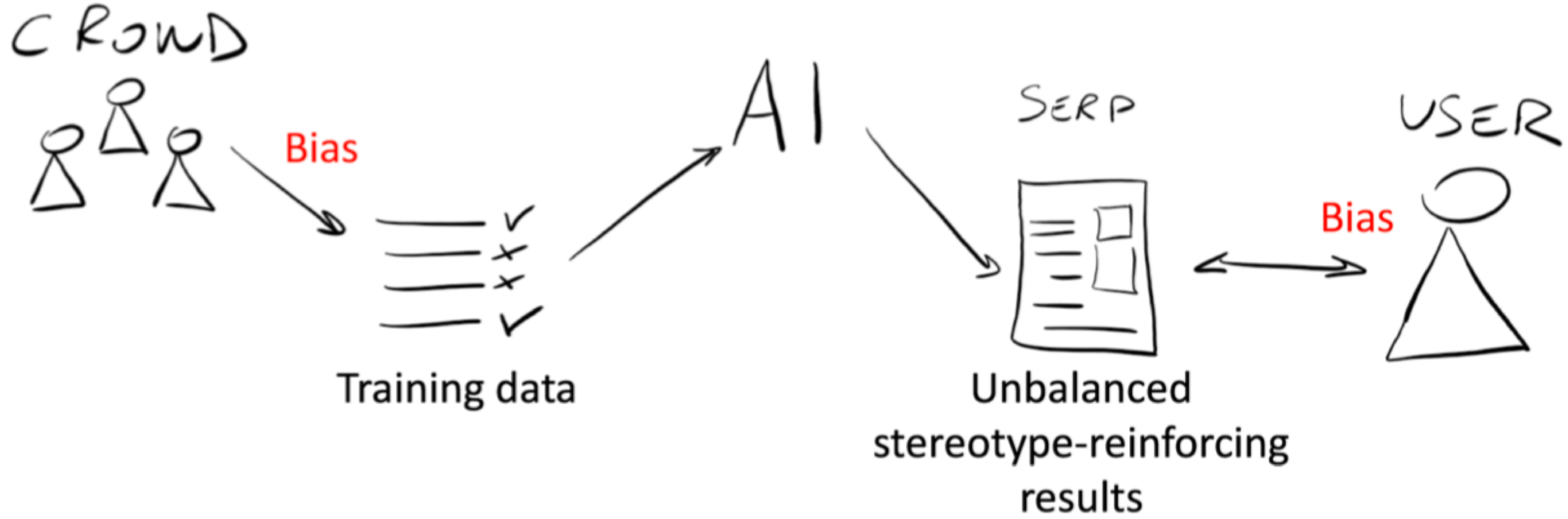
- Fact checkers are expert journalists verifying sources and validating news

- 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.

Should AI systems reinforce stereotypes or rather break the bubble?



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
  - Means to deal with **implicit bias** to **improve the quality of data** with humans in the loop