Human Factors and Bias in Crowdsourced Information Retrieval Evaluation

Gianluca Demartini

demartini@acm.org

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Create change

Gianluca Demartini

demartini@acm.org

- BSc MSc in CS at U. of Udine, Italy
- PhD at U. of Hannover, Germany
  - Entity Retrieval
- Worked at U. Sheffield iSchool (UK), the eXascale Infolab U. Fribourg (CH), UC Berkeley (on Crowdsourcing), Yahoo! (ES), L3S Research Center (DE)
- Senior Lecturer in Data Science at the School of ITEE, U. Queensland since 2017
- Tutorials on
  - Entity Search at ECIR 2012 and RuSSIR 2015
  - Crowdsourcing at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016, Facebook

www.gianlucademartini.net

Research Interests

- **Entity-centric Information Access** (2005-now)
    - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
    - NER in Scientific Docs (WWW 14), Prepositions (CIKM 14)
    - **IR Evaluation** (IRJ 2015, ECIR 16 Best Paper Award, CIKM 17, **SIGIR 18**)
- **Hybrid Human-Machine Systems** (2012-now)
    - ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
    - Hybrid systems overview (COMNET 15, FnT 17)
- **Better Crowdsourcing Platforms** (2013-now)
    - 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** (2015-now)
    - Malicious Workers (CHI 15), **Attack Schemes (HCOMP 18)**
    - Modus Operandi (UBICOMP 17), **Bias in Crowdsourcing (SIGIR 18)**
    - Timeout (HCOMP 16), Complexity (HCOMP 16)

Thanks to:

European Commission

EPSRC
Engineering and Physical Sciences Research Council

Outline

- Crowdsourcing
- Information Retrieval Evaluation
- Human Factors
    - Relevance Scales (SIGIR 2018)
    - Gender Bias and Sexism (SIGIR 2018)
    - Crowd Attack Schemes (HCOMP 2018)
- Joint work with

UNIVERSITÀ DEGLI STUDI DI UDINE

OPEN UNIVERSITY OF CYPRUS

The University Of Sheffield.

Crowdsourcing

- "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an **open call**. This can take the form of peer-production (when the job is performed **collaboratively**), but is also often undertaken by sole **individuals**. The crucial prerequisite is the use of the open call format and the **large network of potential laborers**.“

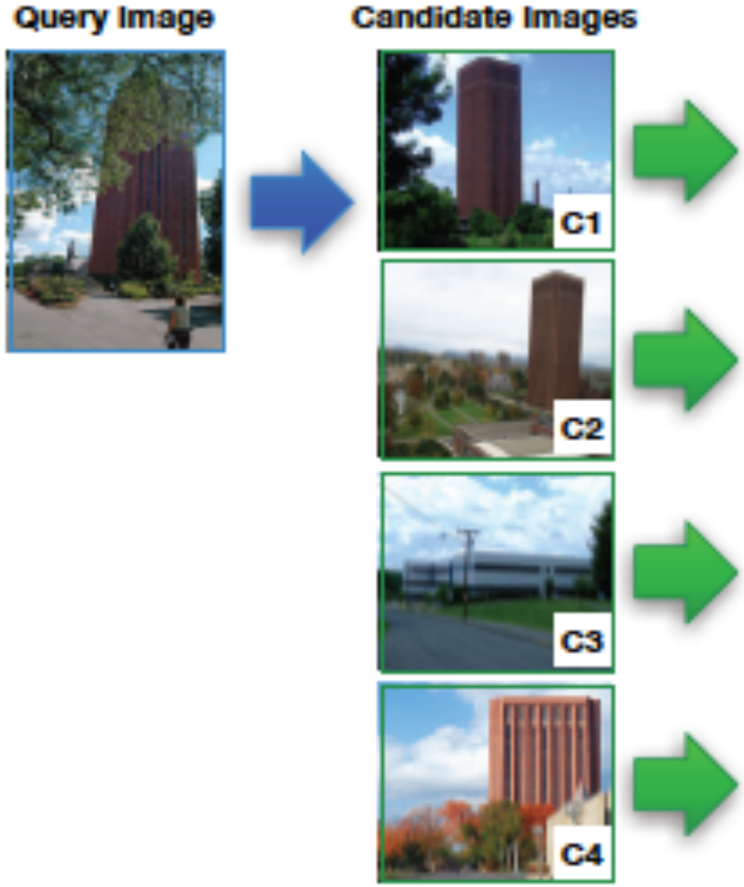
[Howe, 2006]

WIKIPEDIA
The Free Encyclopedia

GALAXY ZOO

amazon
beta
mechanical turk

Hybrid Image Search



Duplicate Validation Tasks

	1	2	3	4	5
Results	Yes	Yes	Yes	Yes	Yes

	1	2	3	4	5
Results	Yes	No	Yes	Yes	Yes

	1	2	3	4	5
Results	No	No	No	No	No

	1	2	3	4	5
Results	Yes	No	No	Yes	Yes

Yan, Kumar, Ganesan, CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones, Mobisys 2010.

CrowdDB

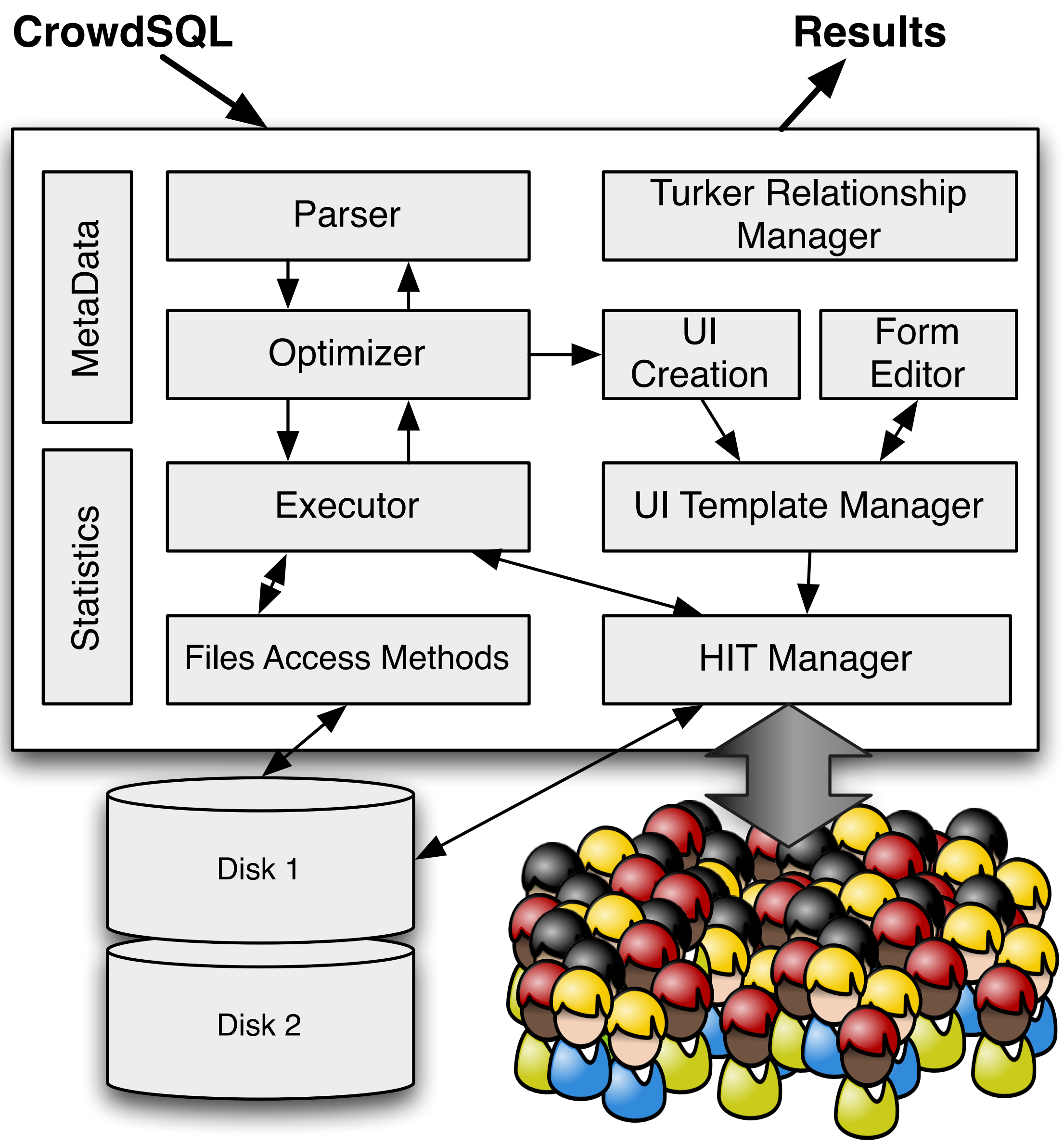
Use the crowd to answer DB-hard queries

Where to use the crowd:

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

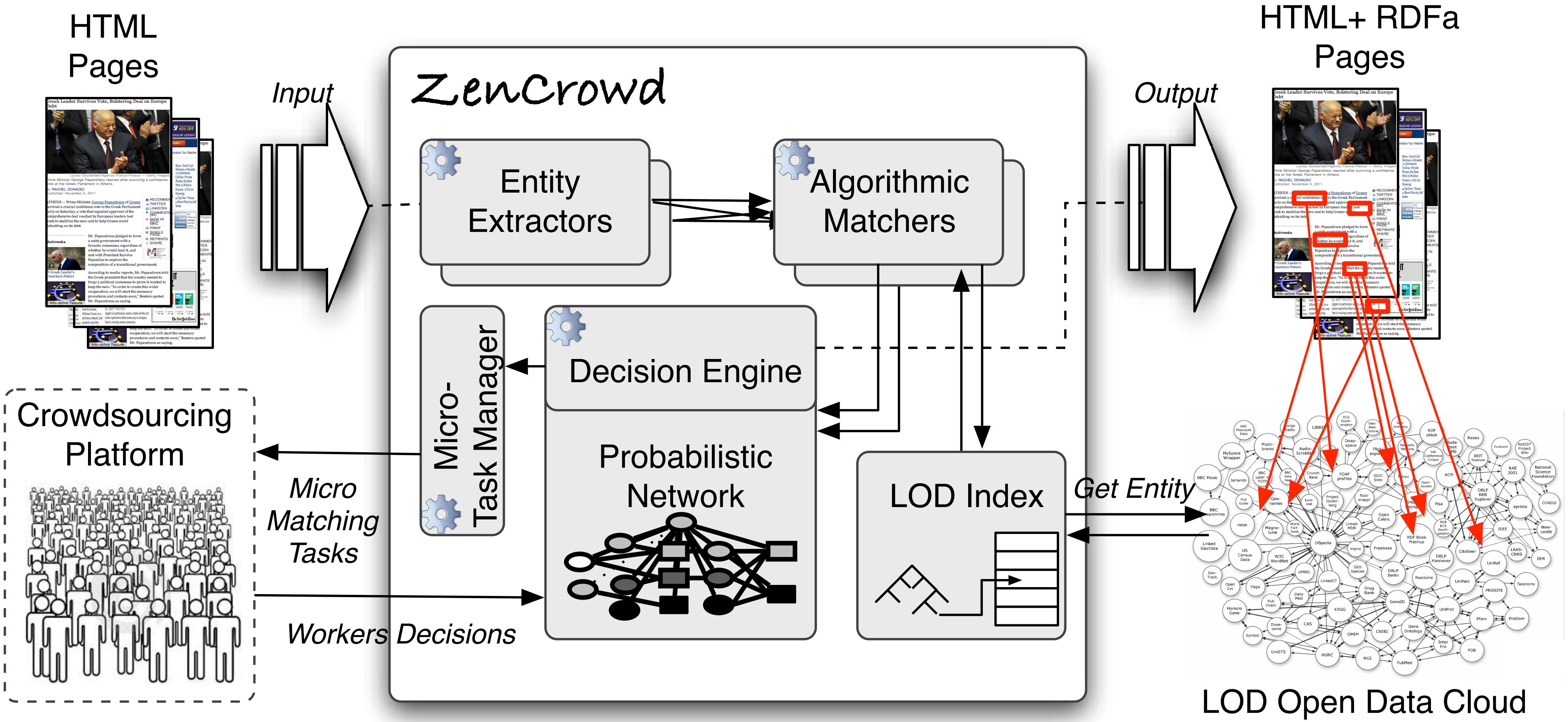
But not:

- Anything the computer already does well



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

ZenCrowd



Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In: 21st International Conference on World Wide Web (**WWW 2012**).

Human Computation 101 - Summary

- Crowdsourcing is growing in popularity
- It is used both in industry and academia
- For a number of applications across disciplines

Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, and Michele Catasta. **An Introduction to Hybrid Human-Machine Information Systems**. In: Foundation and Trends in Web Science Vol. 7: No. 1, pp 1-87. 2017.

- Open questions:
  - How to make sure we get quality results back from a crowdsourcing platforms? (**Effectiveness**)
  - Can we optimize the cost and execution in paid micro-task crowdsourcing? (Efficiency)

Outline

- Crowdsourcing
- **Information Retrieval Evaluation**
- Human Factors
  - Relevance Scales (SIGIR 2018)
  - Gender Bias and Sexism (SIGIR 2018)
  - Crowd Attack Schemes (HCOMP 2018)

Information Retrieval Evaluation

- Evaluate the effectiveness of search engines (how good the results are)
- Metrics: Precision, Recall, Average Precision (AP), NDCG

Query: Donald Trump

Results

V 1. Donald Trump – Wikipedia

X 2. The White House

X 3. Trump Tower

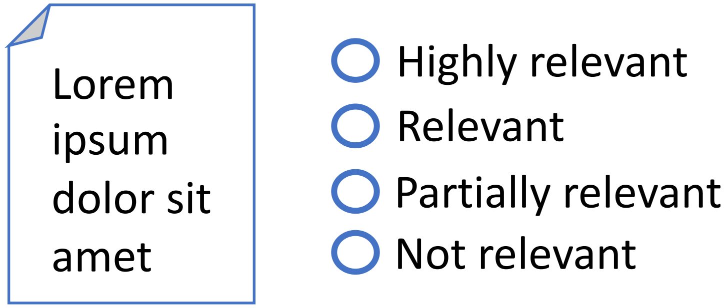
V 4. @realDonaldTrump - Twitter

Precision: 0.5
Recall: ??
AP: 0.75
NDCG: needs non-binary judgements

Crowdsourcing Relevance Judgements

- Task: Given a (Search query, Document) pair
  Is the document:
  highly relevant, relevant, partially relevant, not relevant?
- Ask multiple workers
- Aggregate answers to obtain one relevance label for the (query/doc)

Query: jaguar



Outline

- Crowdsourcing
- Information Retrieval Evaluation
- Human Factors
  - **Relevance Scales (SIGIR 2018)**
  - Bias and Sexism in Search Results (SIGIR 2018)
  - Crowd Attack Schemes (HCOMP 2018)

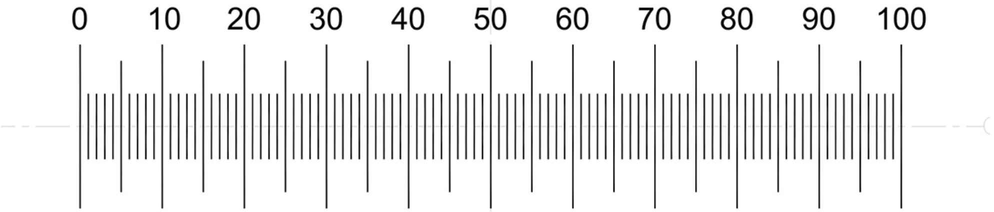
Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. **On Fine-Grained Relevance Scales**. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018). Ann Harbor, Michigan, July 2018.

Relevance Scales

	RELEVANT	NON-RELEVANT	
RETRIEVED	a	b	a + b
NOT RETRIEVED	c	d	c + d
	a + c	b + d	a + b + c + d = N (Total Collection)

FIGURE 1 2 x 2 CONTINGENCY TABLE

- Binary – Cranfield experiments, 1967
- Multi-level judgements – NDCG, SIGIR 2000
- Continuous judgements – Magnitude Estimation (ME), SIGIR 2015
  -]0, +∞[
- **S100 – SIGIR 2018**



Issues with ME

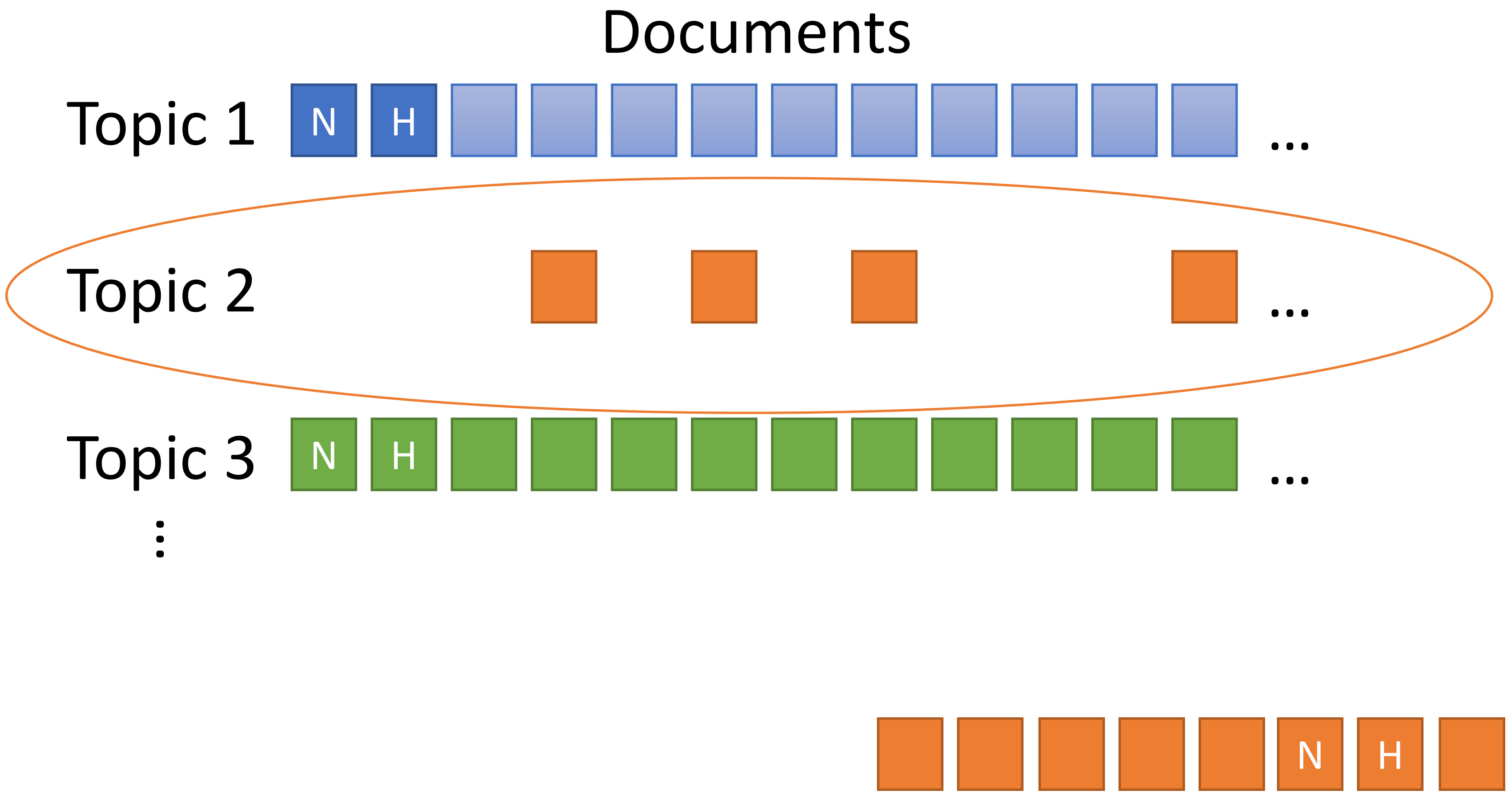
- My “inner scale” is different from yours
- Culture will affect which numbers are used
  - E.g., school marks over
    0–10 (in Italy) vs. 1-7 (in Australia) vs. 0–100 vs. …
  - Round number tendency

ME vs S100, in theory

- Pro ME
  - Ratio scale
  - New values always available
- Pro S100
  - No normalization issues
  - More familiar / similar to usual approaches (e.g., 5 stars)

Experimental Setup

- We compare S2 (R,N), S4 (H,R,M,N), S100 (0-100), and ME judgments over the same queries/documents
- ME and S100 judgments are collected by means of crowdsourcing
  - Randomized design to prevent potential ordering effects
  - Each set included a known ordinal “S4-H” and “S4-N” document for a topic; these were the same for every participant for that topic
  - 10 scores gathered for each of the 4,269 topic-document pairs
  - Total units: 7,059, ~50k judgments

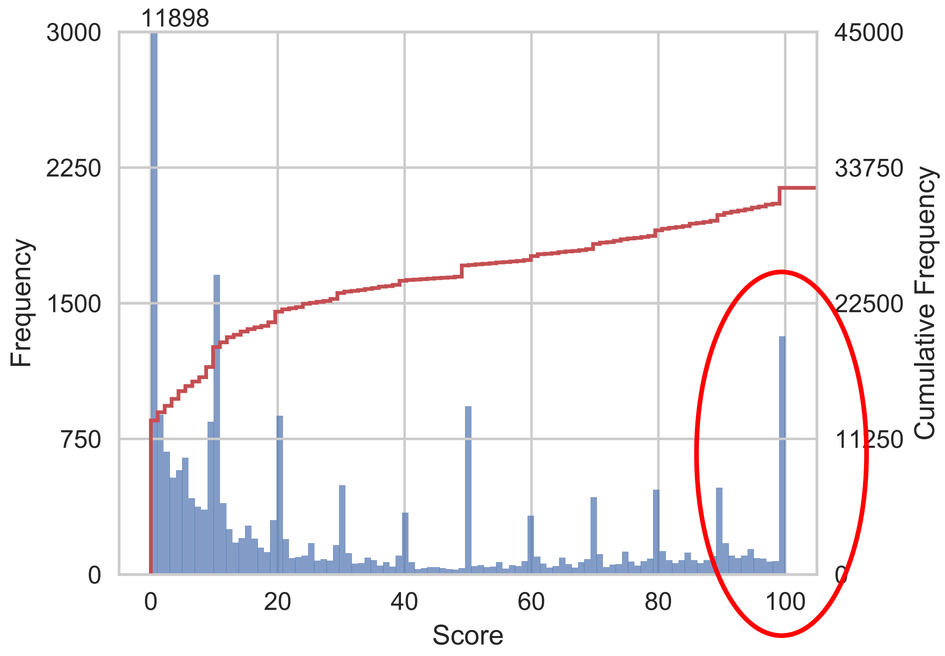


1. Worker chooses topic
2. Choose N, H, + 6 random documents
3. Shuffle randomly and present

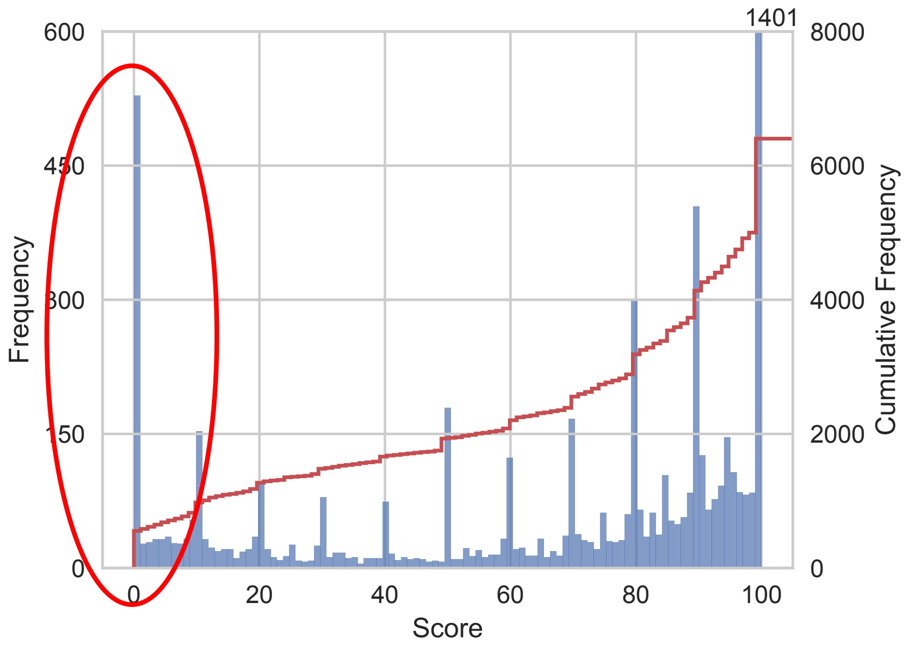
All workers for a topic get **the same N and H** docs

Individual scores

- Decimal tendency
- "Wrong" scores…



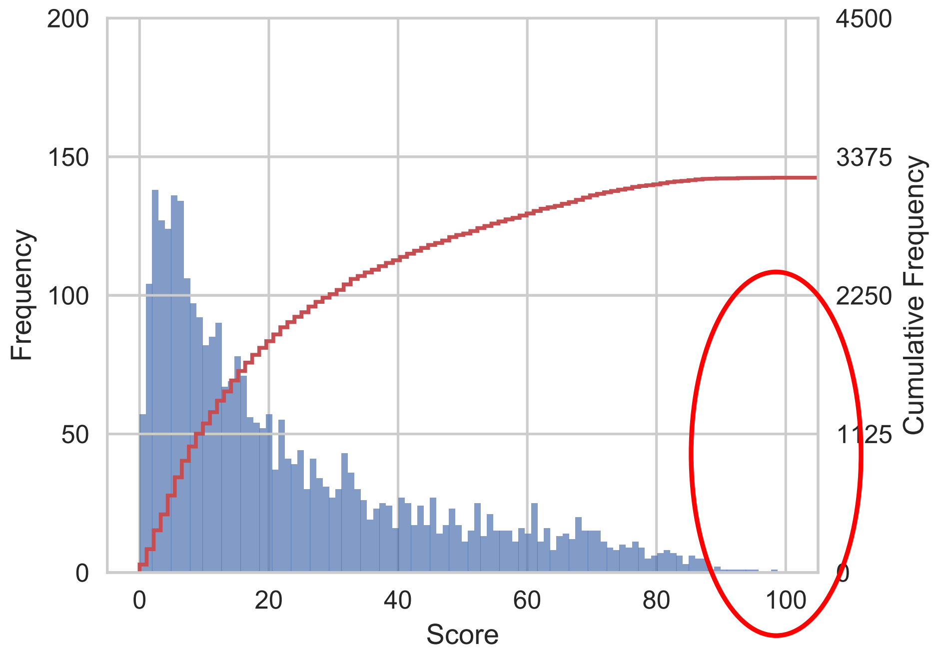
Non-rel



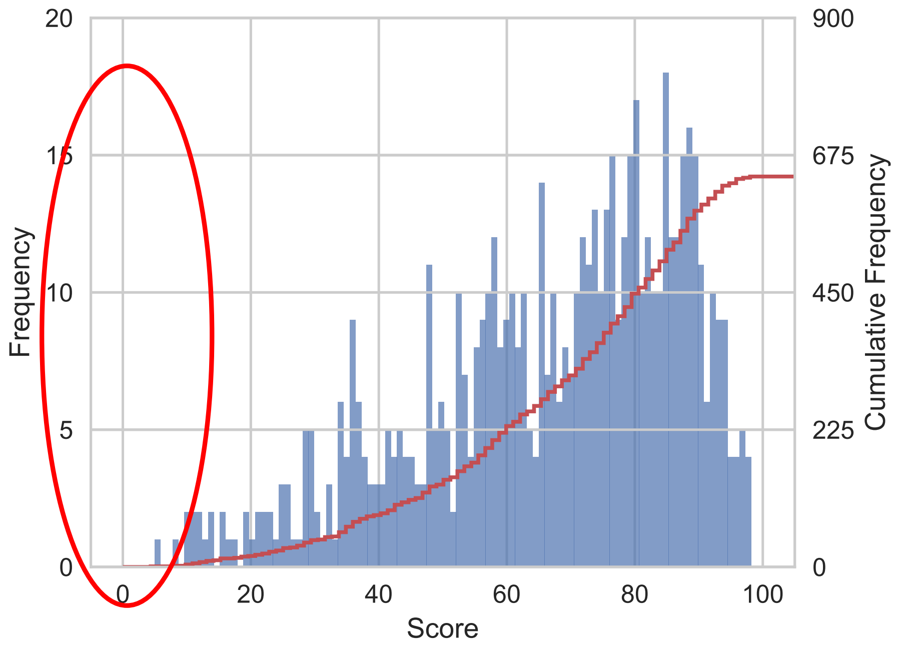
Rel

Aggregated scores

- Decimal tendency gone
- No more "wrong" scores…

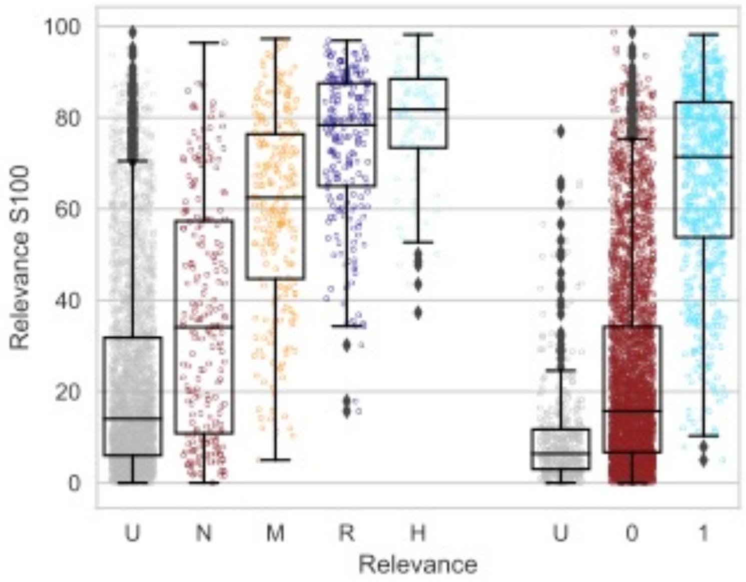


Non-rel

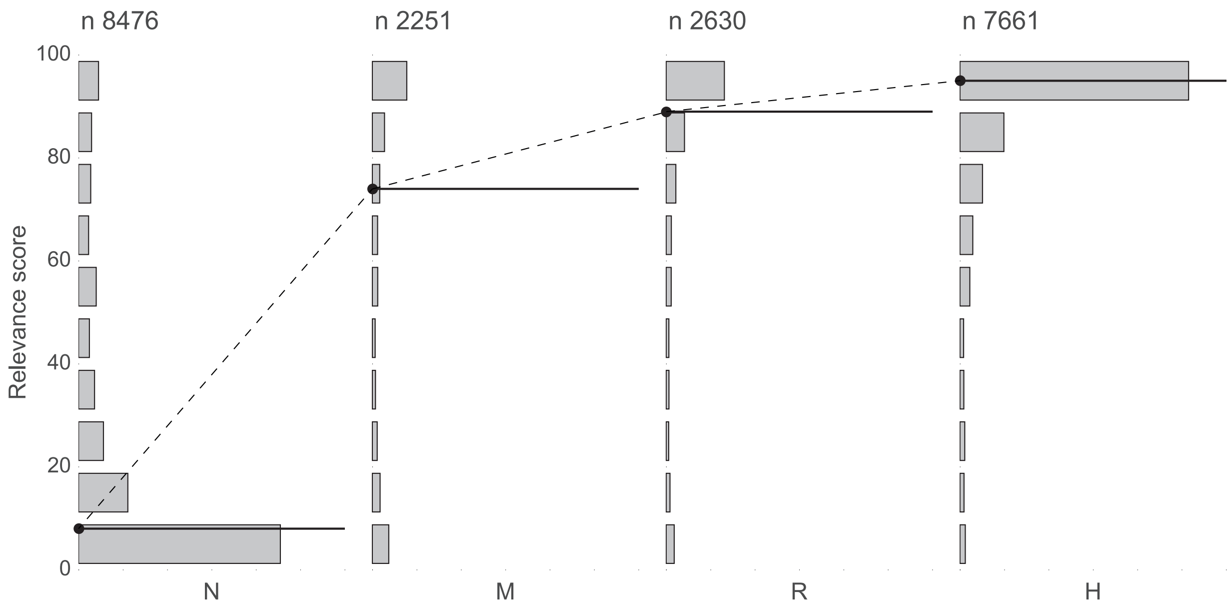


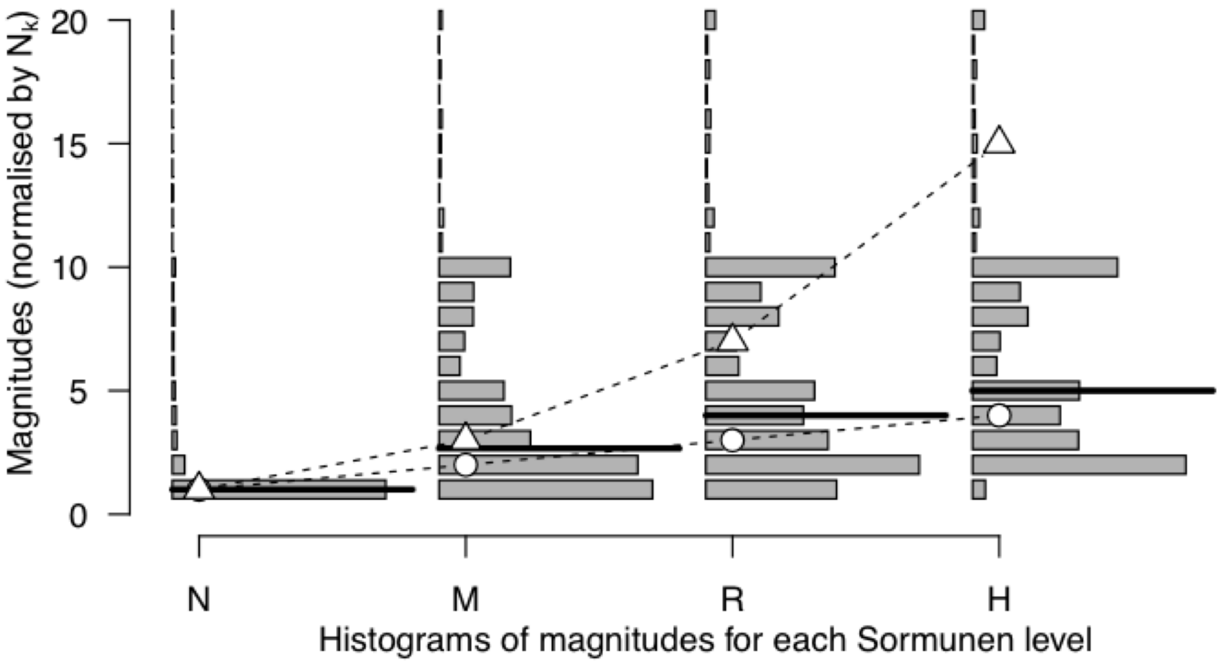
Rel

Consistency of S100 and Ordinal / Binary Relevance



Gain Profiles

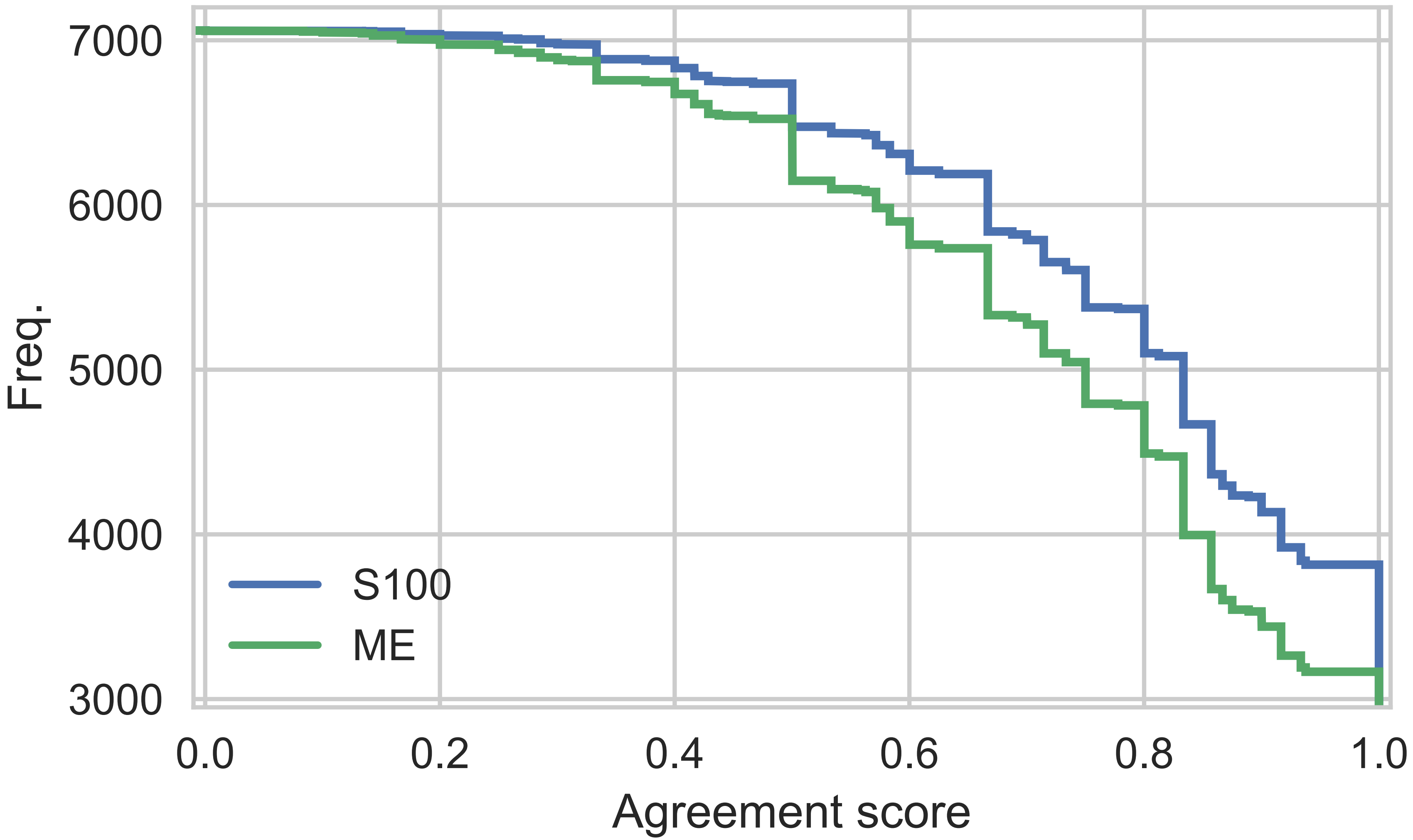




S100: sublinear

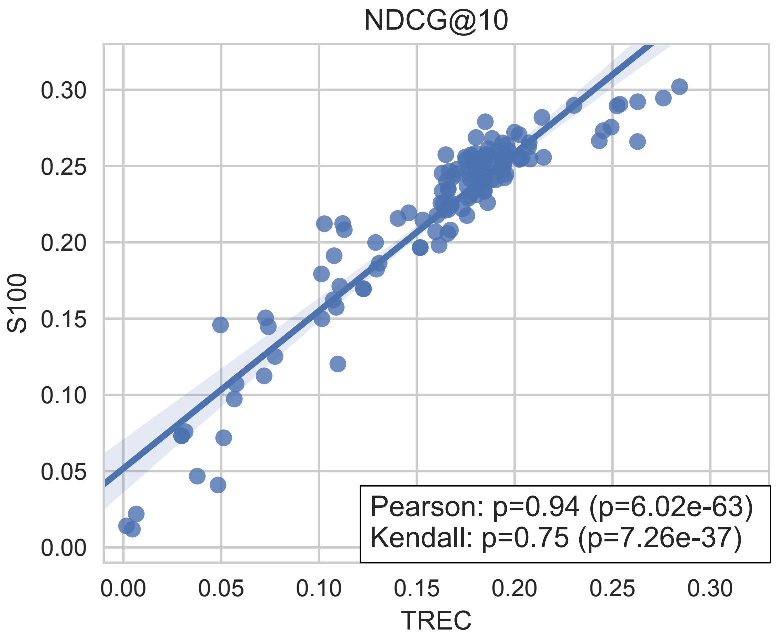
ME: superlinear

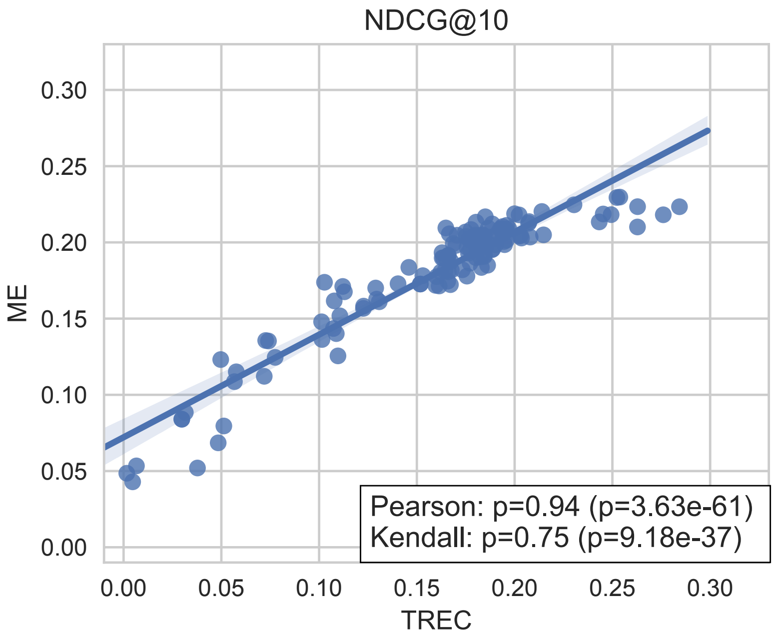
Agreement with Editors (S2): S100 > ME

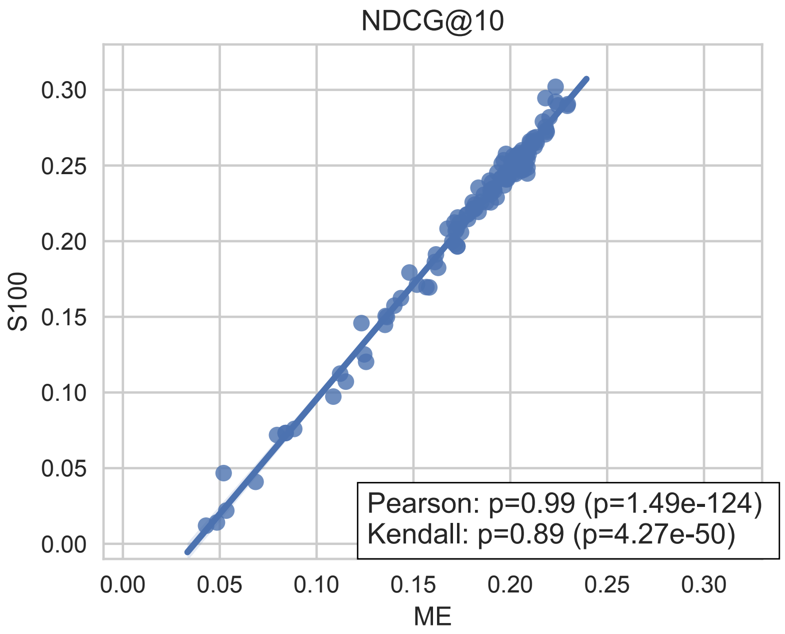


Effect on system ranking

- S100 and ME generate very similar results

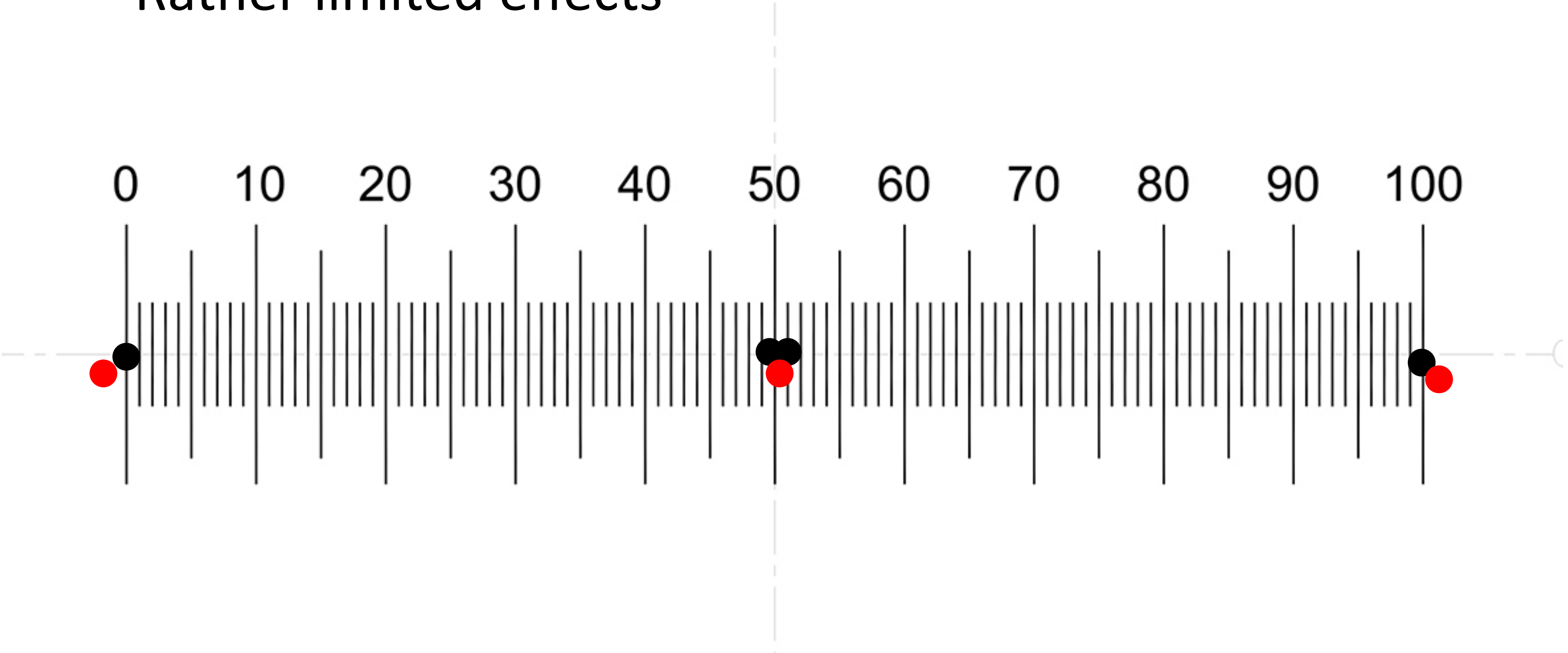






S100 running out of values?

- Scale boundaries
- Discrete vs. Continuous scale
- Rather limited effects



Observations

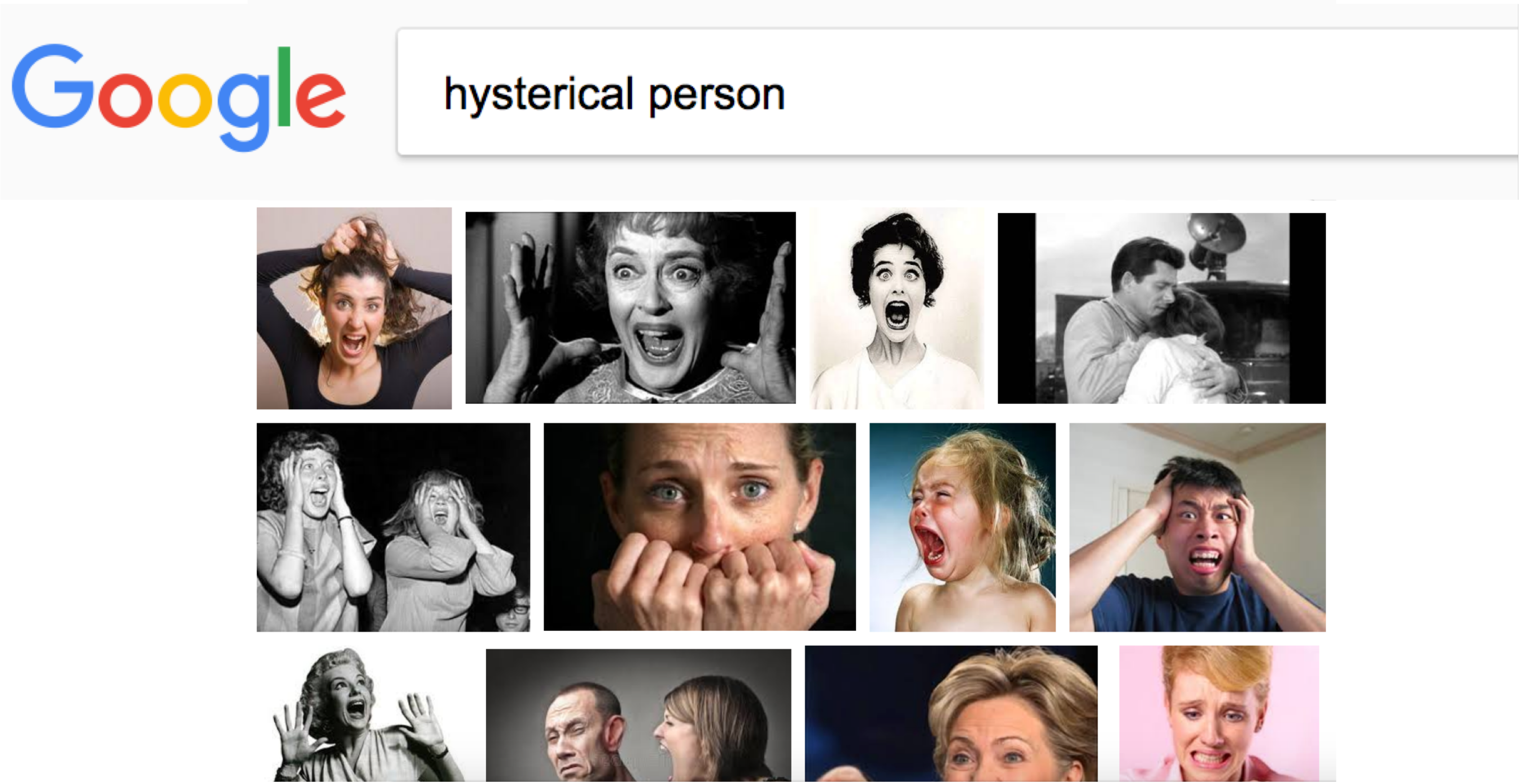
- S100 has many of the advantages of ME
- S100 is better w.r.t.:
  - Agreement with TREC/S2
  - Familiarity for human assessors (looking at time taken to judge)
  - More robust to fewer data (not shown)
- Disadvantages look only theoretical
  - "running out of values" rarely an issue
- S100 looks a good compromise
- Current work: S2 and S4 from the crowd (S10 as well)

Outline

- Crowdsourcing
- Information Retrieval Evaluation
- Human Factors
  - Relevance Scales (SIGIR 2018)
  - **Bias and Sexism in Search Results (SIGIR 2018)**
  - Crowd Attack Schemes (HCOMP 2018)

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.

Search results are biased/imbalanced (CHI17)



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
  - Part4 (ASI)

Experimental Setup

- 281 different users equally split across the three regions and 10 unique queries
- Queries

Query	Trait	Bias
smart person	+	M
aggressive person	-	M
warm person	+	F
anxious person	-	F
hot air balloon	=	na

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**

Outline

- Crowdsourcing
- Information Retrieval Evaluation
- Human Factors
  - Relevance Scales (SIGIR 2018)
  - Bias and Sexism in Search Results (SIGIR 2018)
  - **Crowd Attack Schemes (HCOMP 2018)**

Alessandro Checco, Jo Bates, and Gianluca Demartini. **All That Glitters is Gold -- An Attack Scheme on Gold Questions in Crowdsourcing**. In: The 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018). Zurich, Switzerland, July 2018.

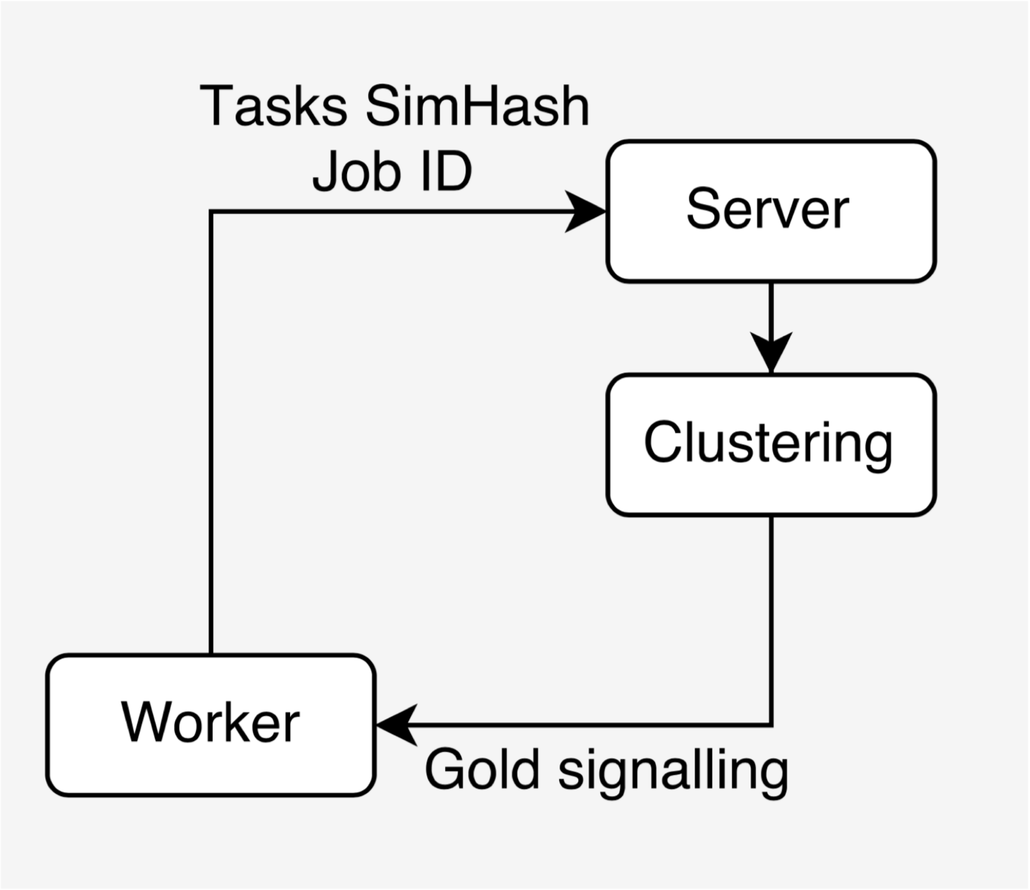
Gold Questions

- 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

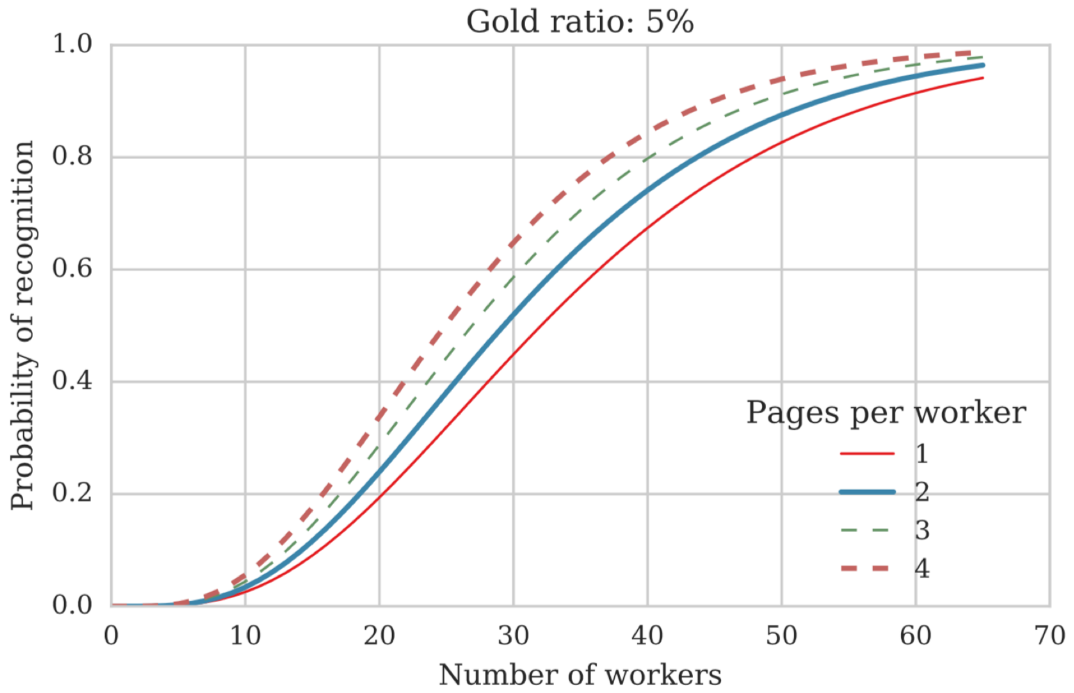
Sharing Information to Spot Gold Questions

- Worker Collusion
- Worker to share the questions they receive to identify the gold

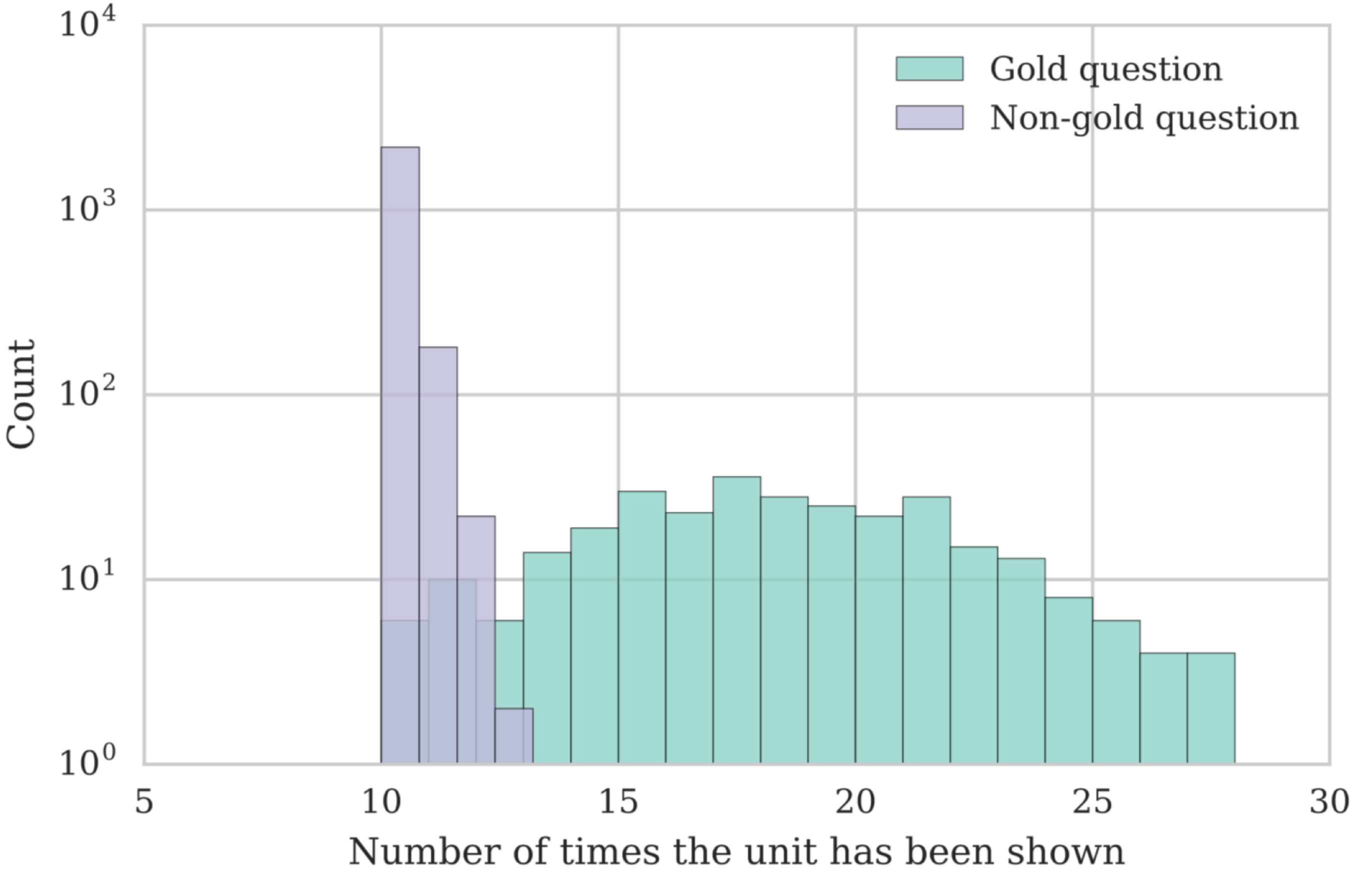


Example

- 5% gold (Answer known for 5 questions each 100 we crowd-source)
- 10 questions per task
- 50 workers, 30 questions per worker -> 90% detection probability

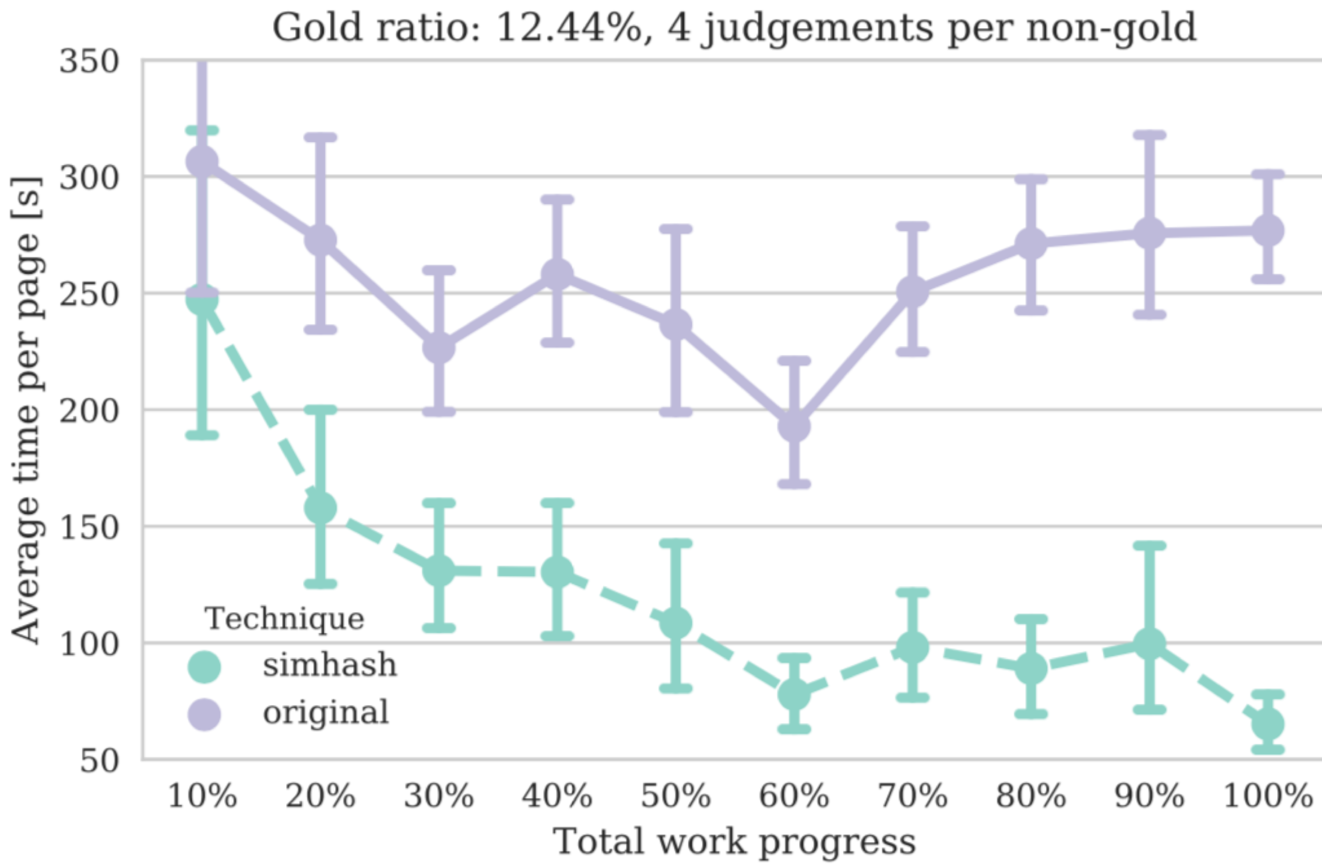


Real data – Repetition of Gold Questions



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

Conclusions

- Human-in-the-loop systems can solve complex tasks at scale
- Humans come with challenges!
- How to best **ask questions** (Relevance Scales)
- How to deal with **implicit biases** in collected data
  - that is then used to train ML
  - than is then used to make decisions
- How to **guarantee quality** (if workers collude to attack quality control)

gianlucademartini.net

demartini@acm.org

@eglu81