Tutorial: Using Crowdsourcing Effectively for Social Media Research
17th May 2016
ICWSM 2016, Cologne,Germany

Part II
Quality Control in Crowdsourcing

Djellel Eddine Difallah
University of Fribourg, Switzerland

Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. Aspects that Affect the Quality of Results
3. Understanding Worker Malicious Behavior
4. Typical Quality Control Measures
5. Best Practices and Design Patterns

Introduction /1

- Google, Microsoft Bing.
  - Relevance judgment.
  - Image search.
- Twitter.
  - Understand new queries and hashtags.
- Amazon, LinkedIn.
  - Data curation.

Introduction /2

UTest Elance CloudCrowd GRABCAD liveops oDesk Changing How the World Works. mygengo

amazonmechanical turk CrowdFlower agentanything crowdcontent socialvibe

CTR Boost DESIGN RIDER CAPEO samasource SimpleWorkers Micro Job Projects! SpeakerText

trada guru.com microtask taskus DoNanza Make Money Working From Home Lingotek DeskLancer

peopleperhour.com Spool Source clickworker PEER TO PATENT tagasauris the smart way to tag!

freelancer expertplanet Truveli MOB4HIRE servio mediapiston

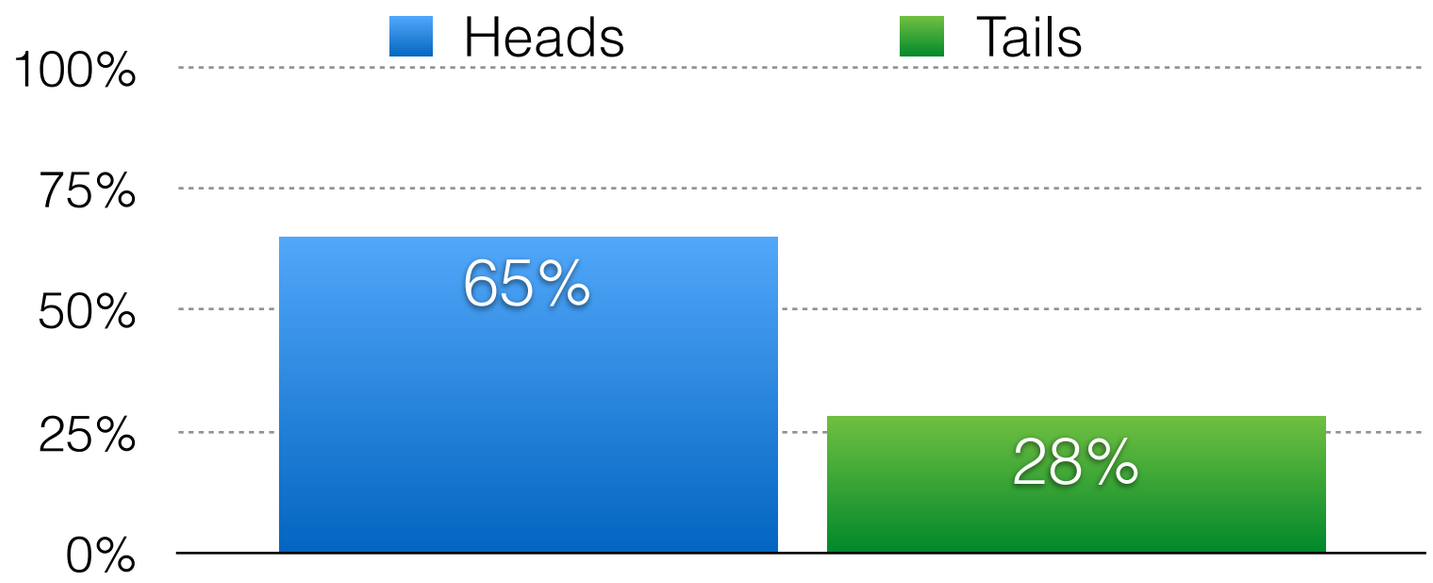
Paid Micro-Task Crowdsourcing

- Offer small monetary reward in exchange of completing short tasks online
    - Entertainment-driven workers primarily seek diversion by taking up interesting, possibly challenging tasks.
    - Money-driven workers mainly attracted by monetary incentives.
- A crowdsourcing platform acts as a marketplace for such tasks (Amazon Mechanical Turk)
- About five million tasks are completed per year at 1-5 cents each
- Some jobs can contain more than 300K tasks

Panos Ipeirotis.
Analyzing the amazon mechanical turk marketplace.
XRDS 2010.

Cheating or Genuine Errors?

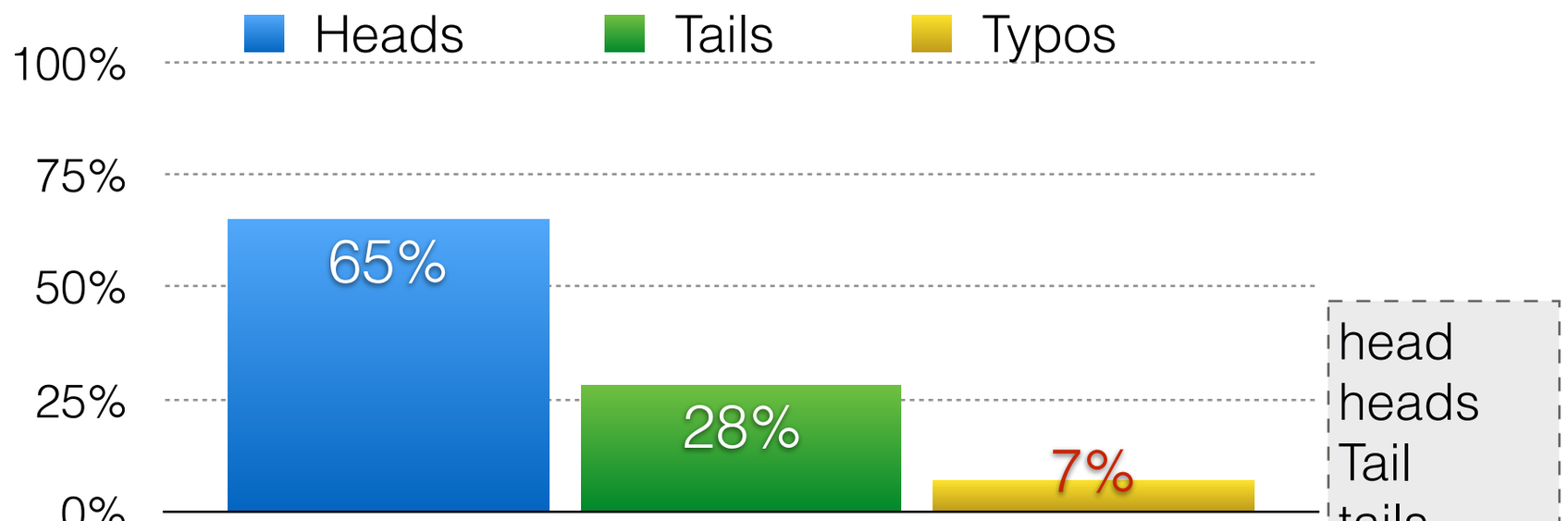
1,000 participants on Amazon Mechanical Turk flip a coin and report "**h**" (heads) or "**t**" (tails)



Results from Michael Bernstein:
https://forum.stanford.edu/events/2013/2013slides/plenary/michael%20bernstein-13.04.17-computerforum-30min.pdf

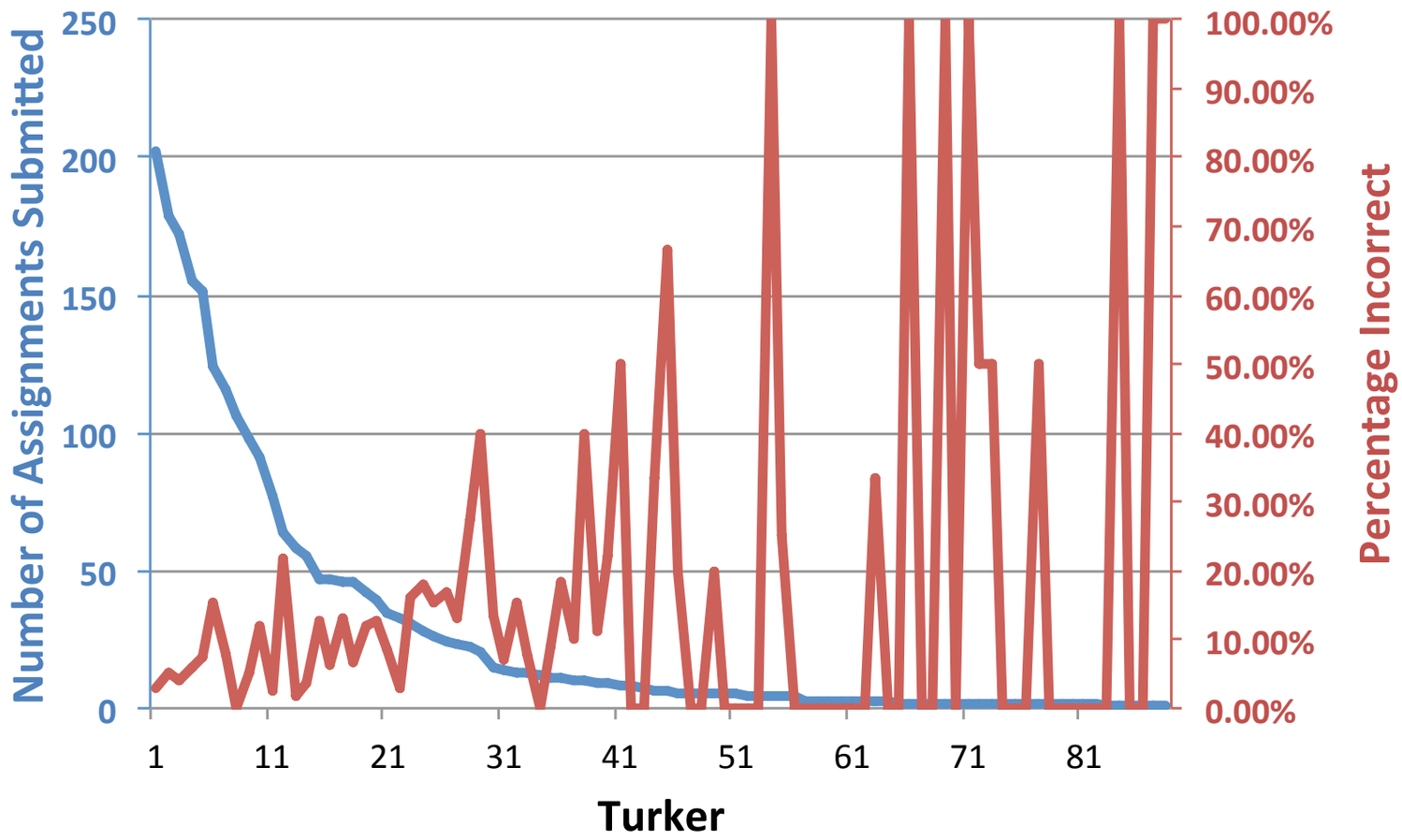
Cheating or Genuine Errors?

1,000 participants on Amazon Mechanical Turk flip a coin and report "**h**" (heads) or "**t**" (tails)



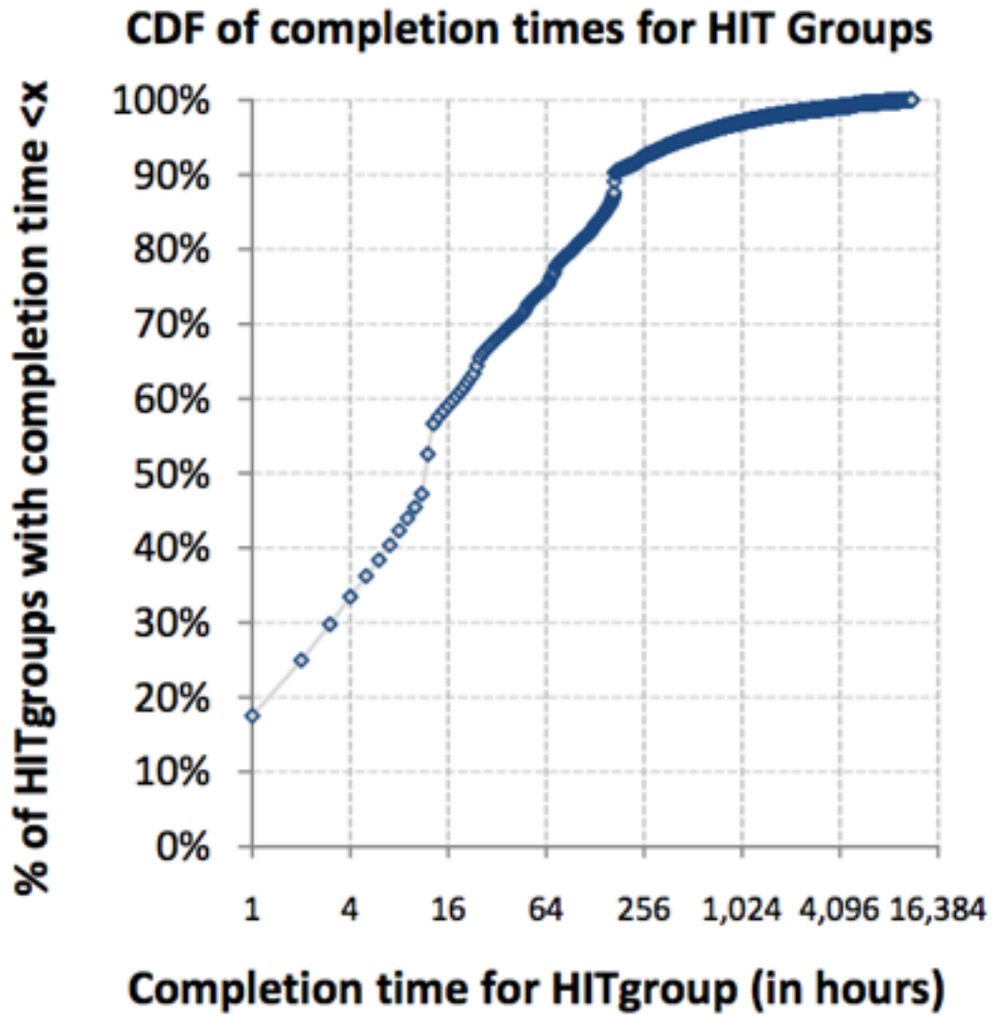
Results from Michael Bernstein:
https://forum.stanford.edu/events/2013/2013slides/plenary/michael%20bernstein-13.04.17-computerforum-30min.pdf

Worker Affinity and Errors



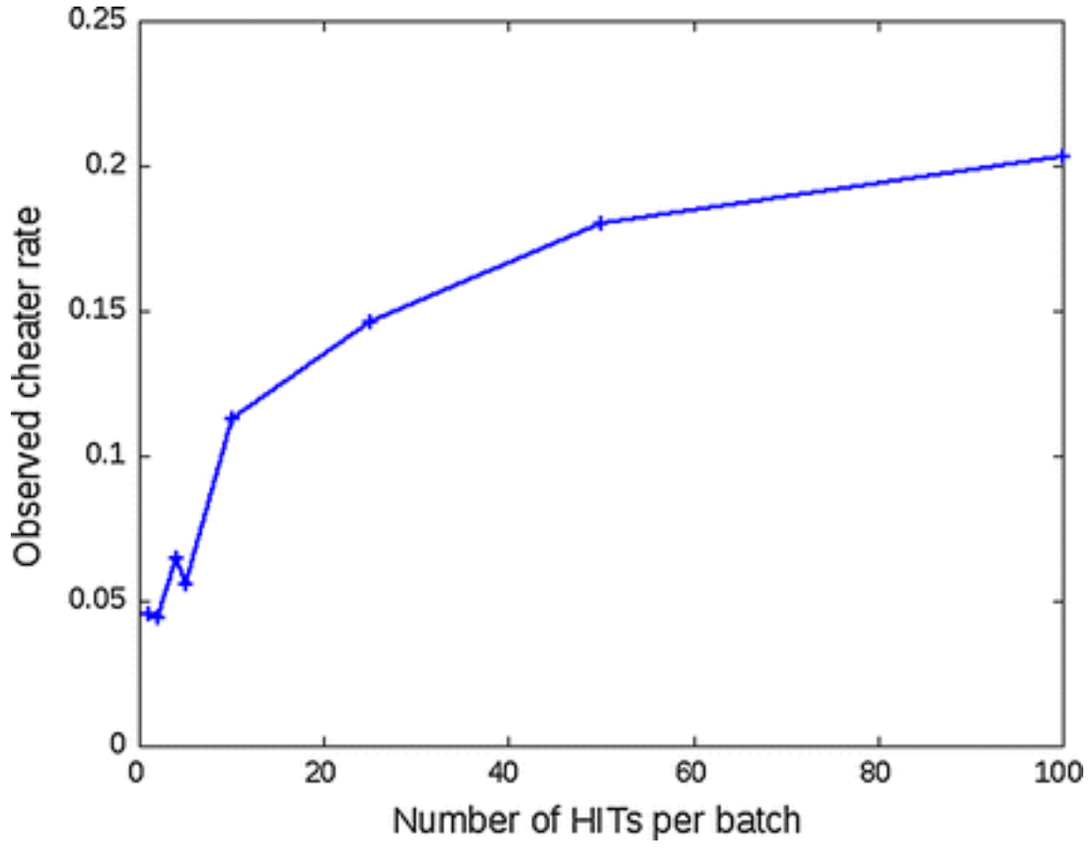
Franklin, Kossmann, Kraska, Ramesh, Xin
CrowdDB: Answering Queries with Crowdsourcing.
SIGMOD,2011

Task Arrival vs Completion Time



Panos Ipeirotis.
Analyzing the amazon mechanical turk marketplace.
XRDS 2010.

Batch Size vs Error Rate



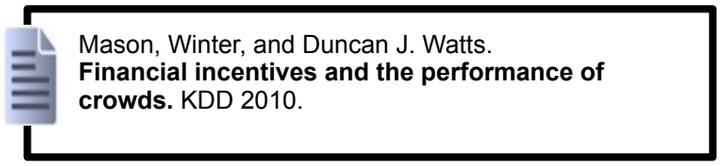
Eickhoff, Carsten, and Arjen P. de Vries. **Increasing cheat robustness of crowdsourcing tasks.** Information retrieval 2013.

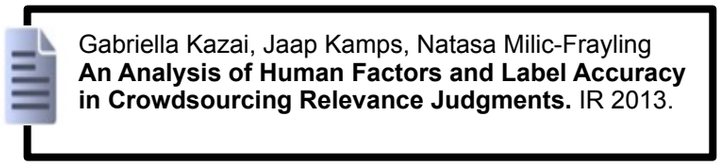
Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. **Aspects that Affect the Quality of Results**
3. Understanding Worker Malicious Behavior
4. Typical Quality Control Measures
5. Best Practices and Design Patterns

Task Pricing

- Too little reward leads to “sloppy” work (no commitment from the workers).
- Paying more increases the quantity of responses and the throughput, but not the quality.
- Encourages good workers.
- Attract bad workers with sophisticated cheating schemes (automated scripts, sharing answers).





Workers Screening

HIT Totals (What's this?)

HITs You Have Submitted	Value	Rate
HITs Submitted	1104	—
... Approved	1022	99.9%
... Rejected	1	0.1%
... Pending	81	—

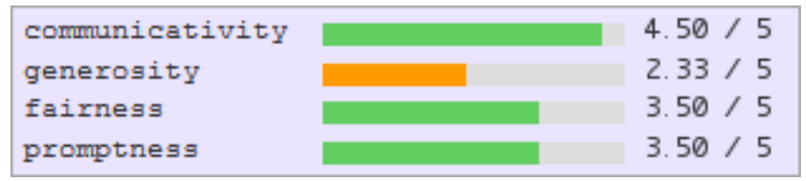
- Current tools for workers selection (or blocking) are based on statistics which are not necessarily indicative of the worker's skills.

- Unique IDs can be used to track performance for current an future experiments.

Panos Ipeirotis
http://www.behind-the-enemy-lines.com/2010/10/be-top-mechanical-turk-worker-you-need.html

Requester Reputation

- Workers express their dissatisfaction on forums and specialized platforms.



- Underpaying requesters.

- Poor task design or instructions.

- Unclear policy of rejection.

Paolacci, Chandler, Ipeirotis.
Running experiments on amazon mechanical turk.
Judgment and Decision making 2010.

rani, Lilly C., and M. Silberman.
Turkopticon: Interrupting worker invisibility in amazon mechanical turk. CHI 2013.

Task Packaging

- HIT Meta information (pay, title, description, instructions).
- Task granularity.
  - Small tasks can attract workers who are motivated by fun.
  - Task formulation.
- The user interface of the HITs.
  - “This took me about half an hour. **Mega bubble** hell though” — a worker.

Gabriella Kazai, Jaap Kamps, Natasa Milic-Frayling
An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments. IR 2013.

Eickhoff, Carsten, and Arjen P. de Vries.
Increasing cheat robustness of crowdsourcing tasks. Information retrieval 2013.

Framing and Priming

- Workers seem to respond better when they know what the task results will be used for.

Chandler, Dana, and Adam Kapelner. **Breaking monotony with meaning: Motivation in crowdsourcing markets.** Journal of Economic Behavior & Organization 2013.

- Inter-tasks content affect the answers provided by the crowd.

Edward Newell, Derek Ruths.
How One Microtask Affects Another. ACM Conference on Human Factors in Computing Systems, 2016

Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. Aspects that Affect the Quality of Results
3. **Understanding Worker Malicious Behavior**
4. Typical Quality Control Measures
5. Best Practices and Design Patterns

Challenges

Diverse pool of crowd workers with different behavior and various motivations

Malicious Workers: *workers with ulterior motives, who either simply sabotage a task, or provide poor responses in an attempt to quickly attain task completion for monetary gains.*

Untrustworthy: *workers who provide wrong answers in response to one or more simple and straightforward attention-check or gold standard questions.*

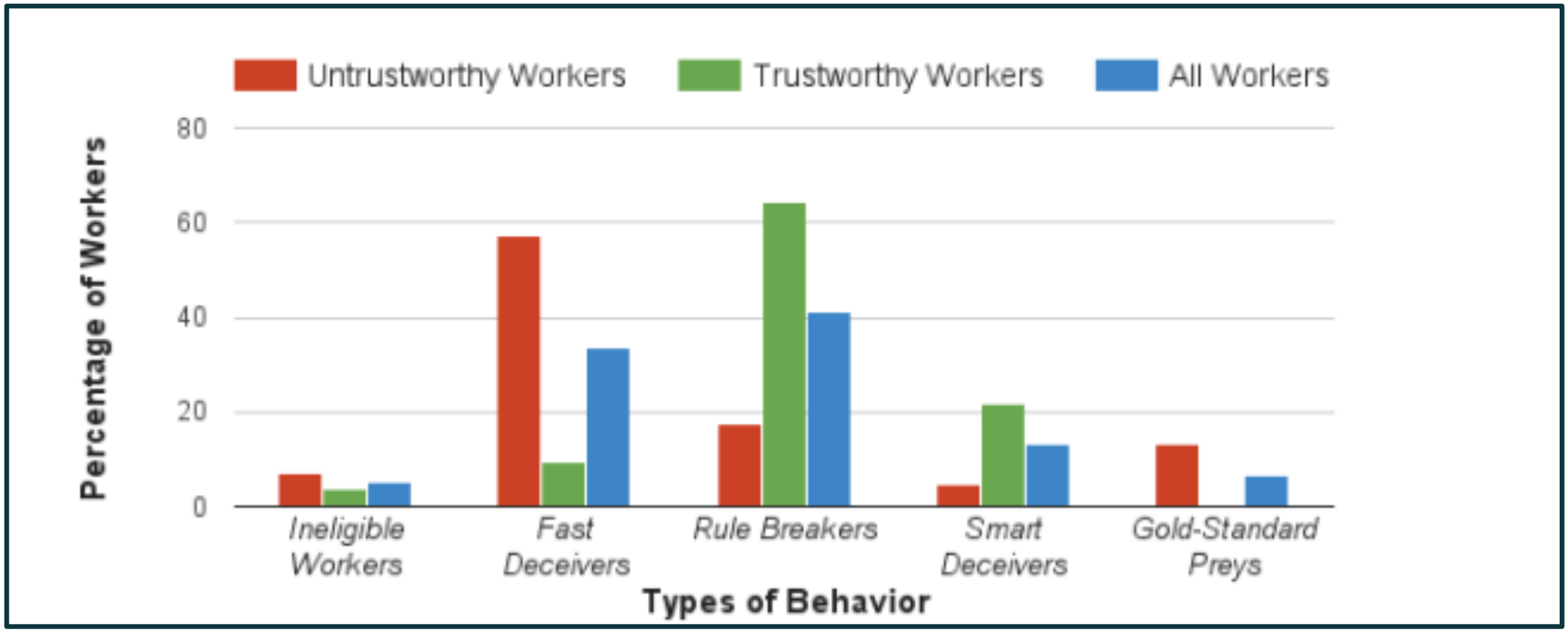
Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. **Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys.** CHI 2015.

Worker 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!

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. **Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys.** CHI 2015.

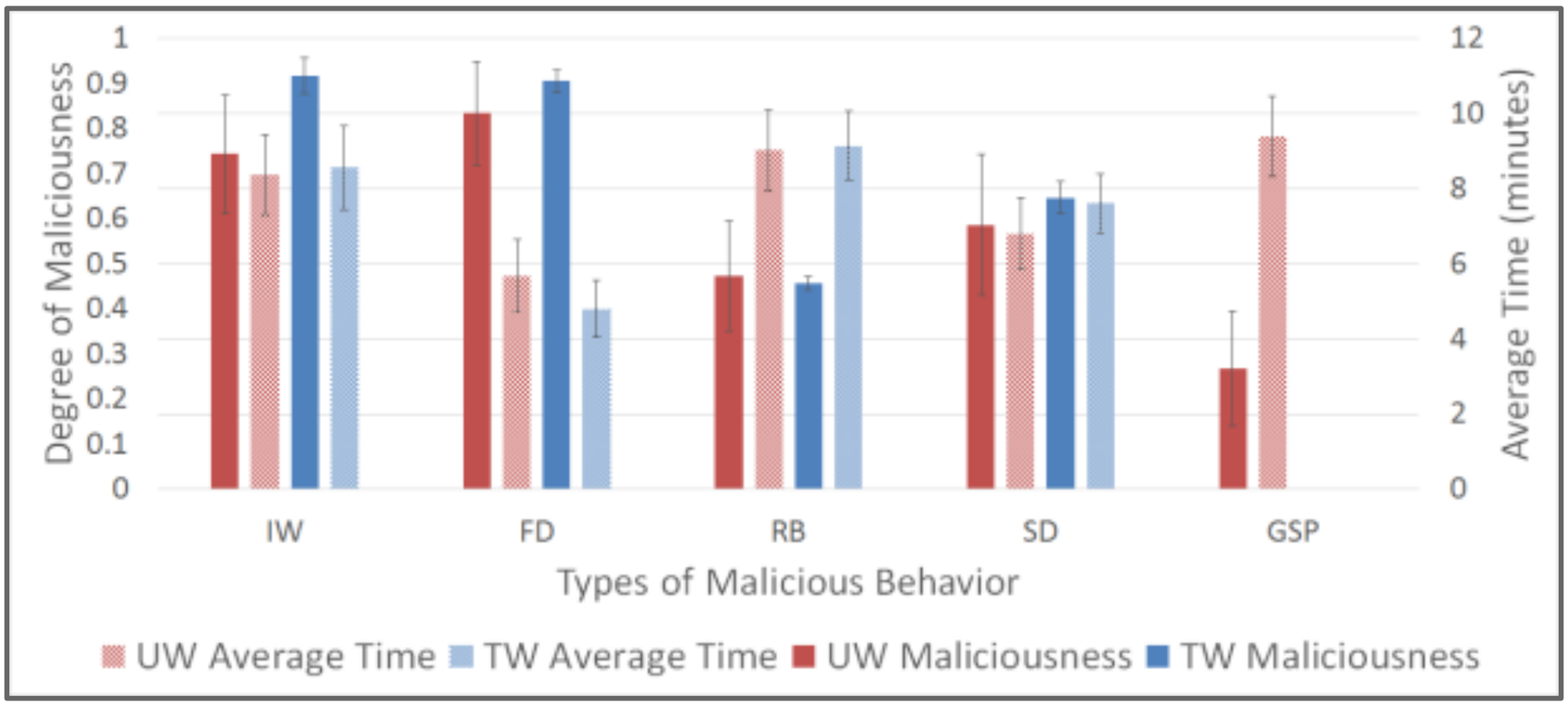
Worker Behavioral Patterns in a Survey Experiment



1000 workers, 34 questions: multiple choice, open ended and likert scale.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. **Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys.** CHI 2015.

Task Completion Time vs Worker Maliciousness



1000 workers, 34 questions: multiple choice, open ended and likert scale.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. **Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys.** CHI 2015.

Cheating Techniques

- Individual Attacks:
  - Random Answers.
  - Educated guess.
  - Automated Answers.
  - Semi-Automated Answers.

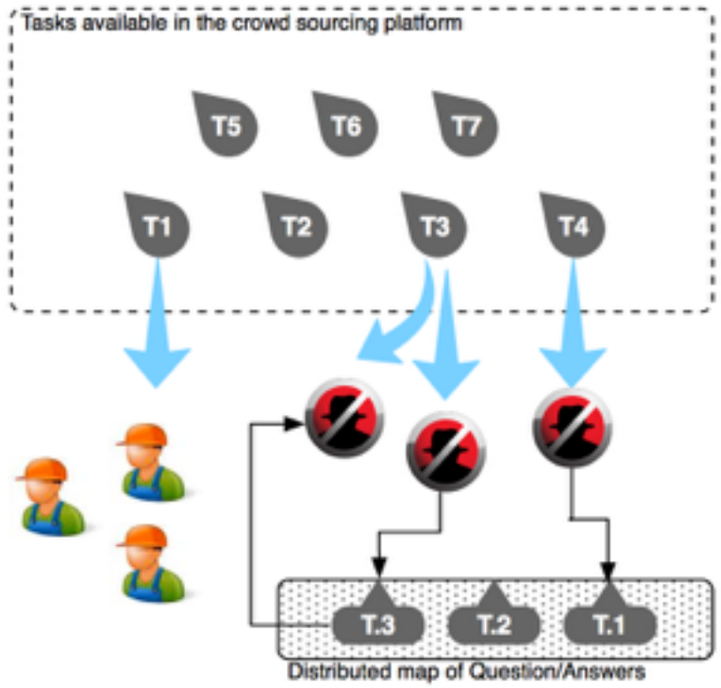
Eickhoff, Carsten, and Arjen P. de Vries. **Increasing cheat robustness of crowdsourcing tasks.** Information retrieval 2013.

Difallah, Djellel Eddine, Gianluca Demartini, and Philippe Cudré-Mauroux. **Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms.** CrowdSearch 2012.

Trushkowsky, Beth, Tim Kraska, Michael J. Franklin, and Pradyut Sarkar. **Crowdsourced enumeration queries.** ICDE 2013.

Cheating Techniques

- Group Attacks.
- Agree on Answers.
- Answer Sharing.
- Multiple bots.



Difallah, Djellel Eddine, Gianluca Demartini, and Philippe Cudré-Mauroux. **Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms.** CrowdSearch 2012.

Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. Aspects that Affect the Quality of Results
3. Understanding Worker Malicious Behavior
4. **Typical Quality Control Measures**
5. Best Practices and Design Patterns

Typical Quality Control Measures

- Preventive measures.
    - Prevent malicious workers from participating in you task.
- Post-hoc filtering.
    - Eliminating unreliable responses after paying for and acquiring the required responses from workers.

Preventive measures

Workers Pre-selection

Tools provide by the platform.

- Qualification tasks: Using a sample/simulating real data.

- Demographic filtering e.g., language, region.

Preventive measures

Incentive Design /1

- Using game elements to engage crowd workers, improve their reliability and the overall quality of responses [1, 2, 3].

- Badges, Leaderboards, Levels, Access, Power and Bonuses as furtherance incentives [4].

- ‘Survival probability’, dynamic task allocation with dynamic goals [5].

[1]. Eickhoff, Carsten, et al. "**Quality through flow and immersion: gamifying crowdsourced relevance assessments**." ACM *SIGIR 2012.*

[2]. Ipeirotis, Panagiotis G., and Evgeniy Gabrilovich. "**Quizz: Targeted crowdsourcing with a billion (potential) users.**" ACM *WWW 2014.*

[3]. Rokicki, Markus, Sergej Zerr, and Stefan Siersdorfer. "**Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions**." ACM *WWW 2016.*

[4]. Feyisetan, Oluwaseyi, et al. "**Improving paid microtasks through gamification and adaptive furtherance incentives**." ACM *WWW 2015.*

[5]. Kobren, Ari, et al. **"Getting more for less: optimized crowdsourcing with dynamic tasks and goals."** ACM *WWW 2015.*

Preventive measures

Incentive Design /2

- Pricing Schemes
  - How much ? *“The best way to determine the appropriate level of pay is to estimate the price per unit of effort”* [1].

[1] Gabriella Kazai, Jaap Kamps, Natasa Milic-Frayling. **An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments**. IR 2013.

- Worker retention using periodic bonuses [2].

[2] Difallah, Catasta, Demartini, Cudré-Mauroux. **Scaling-up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement.** HCOMP 2014

Post-hoc Analysis

Aggregation

- Repetition: assign the same task to multiple workers [1].
- Majority Voting : Based on agreement between multiple independent judgments.
- Weighted vote (individual performance, community based) [2,3].
- SQUARE: A benchmark for crowd answer aggregation [4]
  - Binary choices (e.g., sentiment).
  - Multiple-choices (e.g., relevance, word-sense disambiguation).

[1] Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis. **Get another label? improving data quality and data mining using multiple, noisy labelers.** KDD 2008.

[2] Demartini, Gianluca, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. **ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking**. WWW 2012.

[3] Venanzi, Matteo, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. **Community-based bayesian aggregation models for crowdsourcing.** WWW 2014.

[4] Aashish Sheshadri and Matthew Lease. **SQUARE: A Benchmark for Research on Computing Crowd Consensus.** HCOMP 2013. http://ir.ischool.utexas.edu/square/

Post-hoc Analysis

Direct Assessment /1

- Gold-standard Data.
  - Relying on questions with priorly known answers to filter out low quality workers.

Oleson, David, et al. **"Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing."** Human computation (2011).

- Attention check questions.
  - Captchas.
  - Simple tasks (result of a sum).

Eickhoff, Carsten, and Arjen P. de Vries. **Increasing cheat robustness of crowdsourcing tasks.** Information retrieval 2013.

Post-hoc Analysis

Direct Assessment /2

- Continuous testing and feedback
  - Initial training phases followed by the sporadic insertion of test data (gold standard data) [1, 2].
  - Providing expert feedback and allowing workers to assess their work, improves quality of crowd work [3].

[1]. Le, John, et al. "**Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution**." *Crowdsourcing for search evaluation. SIGIR 2010.*

[2].Gadiraju, Ujwal, Besnik Fetahu, and Ricardo Kawase. "**Training Workers for Improving Performance in Crowdsourcing Microtasks**." *EC-TEL 2015.*

[3]. Dow, Steven, et al. **Shepherding the crowd yields better work**. *CSCW 2012*.

Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. Aspects that Affect the Quality of Results
3. Understanding Worker Malicious Behavior
4. Typical Quality Control Measures
5. **Best Practices and Design Patterns**
    1. Pricing Using Error Time Area
    2. Task Design Patterns
    3. Hybrid Human-Machine Aggregation

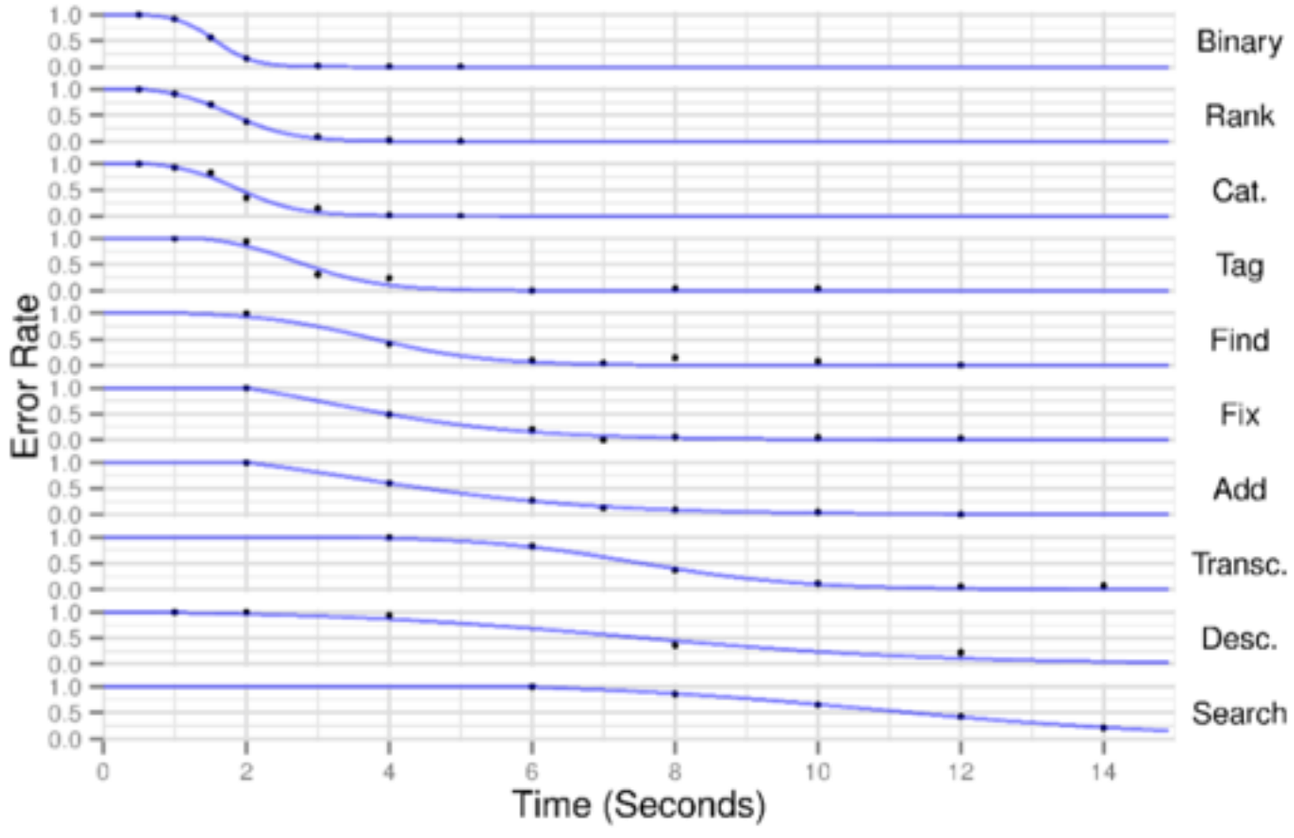
Pricing Using Error Time Area (ETA)

- Estimate the effort to complete a task
  - Requester: Price and structure their task
  - Worker: Decide whether the task is worth
- ETA is a data-driven effort metric
  - Empirically model relationship between time and quality

Cheng, Teevan, Bernstein.
Measuring Crowdsourcing Effort with Error-Time Curves. CHI 2015

Error Time Area (ETA)

- Perform a task under time constraints
- Recommendation: at least seven time limits and 10 workers
- HIT Price= *Time@10* * Hourly Wage



Cheng, Teevan, Bernstein.
Measuring Crowdsourcing Effort with Error-Time Curves. CHI 2015

ETA
Pros and Cons

- ✦ Price can be computed easily (and potentially explained to workers)
- Requires gold answers
- Allows Limited response variability, inter-tasks and across workers

Cheng, Teevan, Bernstein.
Measuring Crowdsourcing Effort with Error-Time Curves. CHI 2015

Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. Aspects that Affect the Quality of Results
3. Understanding Worker Malicious Behavior
4. Typical Quality Control Measures
5. **Best Practices and Design Patterns**
    1. Pricing Using Error Time Area
    2. **Task Design Patterns**
    3. Hybrid Human-Machine Aggregation

Design Patterns

Quality Control for Free Text

- Repetition and aggregation is often used for multiple choice questions. How about:
- Open ended questions.
- Multiple correct versions.
- Good but can do better answers.
- Subjective.

Design Patterns

Find-Fix-Verify

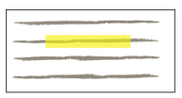
Find

"Identify at least one area that can be shortened without changing the meaning of the paragraph."

Independent agreement to identify patches

Fix

"Edit the highlighted section to shorten its length without changing the meaning of the paragraph."



```
Soylent, a prototype...
```

Randomize order of suggestions

Verify

"Choose at least one rewrite that has style errors, and at least one rewrite that changes the meaning of the sentence."

- ☐ Soylent ~~is~~, a prototype...
- ☐ Soylent ~~is a~~ prototypes...
- ☑ Soylent is a ~~prototype~~test...

Bernstein, Michael S., et al.
Soylent: A Word Processor with a Crowd Inside.
UIST, 2010.

Find-Fix-Verify Use-Cases

- Text editing (proof reading, summarization)[1].

[1] Bernstein, Michael S., et al. **Soylent: A Word Processor with a Crowd Inside.** UIST, 2010.

- Fixing reviews (Well written reviews lead to higher sales) [2].

[2] Ghose, Anindya, and Panagiotis G. Ipeirotis. **Designing novel review ranking systems: predicting the usefulness and impact of reviews.** ICEC 2007.

- Translation.

- Improving textual content for machine learning.

Agenda

1. Introduction to Quality Issues in Crowdsourcing
2. Aspects that Affect the Quality of Results
3. Understanding Worker Malicious Behavior
4. Typical Quality Control Measures
5. **Best Practices and Design Patterns**
    1. Pricing Using Error Time Area
    2. Task Design Patterns
    3. **Hybrid Human-Machine Aggregation**

Hybrid Human-Machine Aggregation

- A Hybrid Human-Machine system combines the results of machine based problem solvers (algorithms) and the the crowd (when necessary).

- Natural Language Processing, Image captioning, Speech processing etc.

- Leverage the output of the algorithm in the quality control process.

- Use-case: *Entity Linking.*

Effective Entity Linking Architecture

Input

HTML Pages

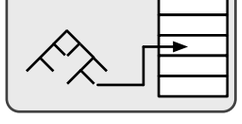
SOTA Entity Extraction

Example

….of Bern and the city of Fribourg, part of the country..

Automatic Linking

Index



- DBPedia
- Freebase
- Geonames
- NYTimes

Example

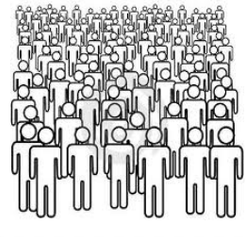
- http://sws.geonames.org/7285870/
- http://dbpedia.org/page/Canton_of_Fribourg
- http://dbpedia.org/page/Fribourg
- http://sws.geonames.org/2660717/
- http://www.freebase.com/m/01qtgw
- http://www.freebase.com/m/01tvfk

Decision Engine

Probabilistic Network

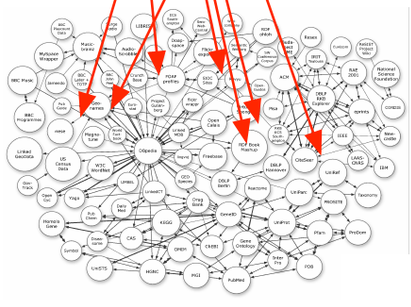
Micro-Tasks

Crowdsourcing Platform



Output

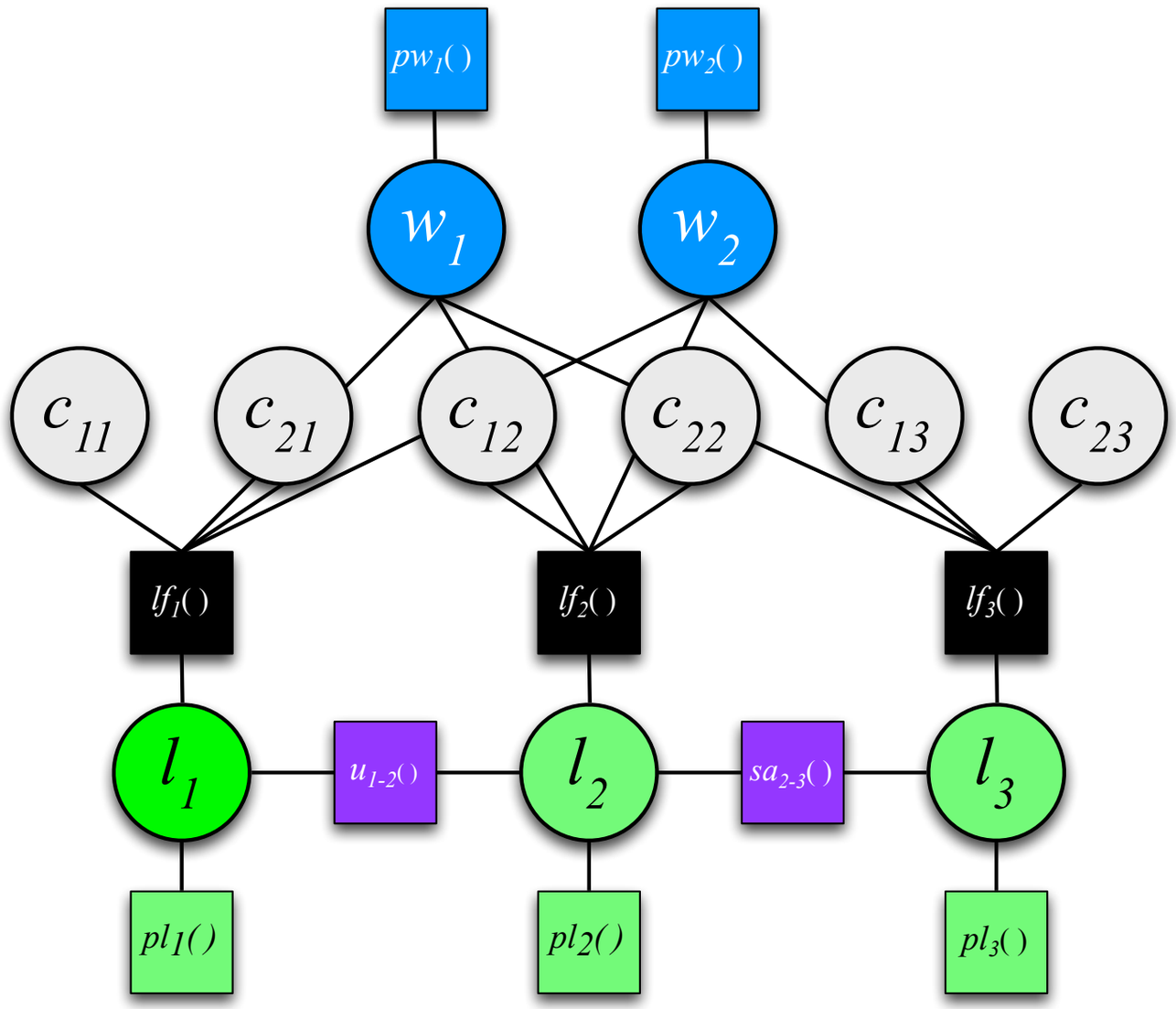
HTML+ RDFa Pages



LOD Open Data Cloud

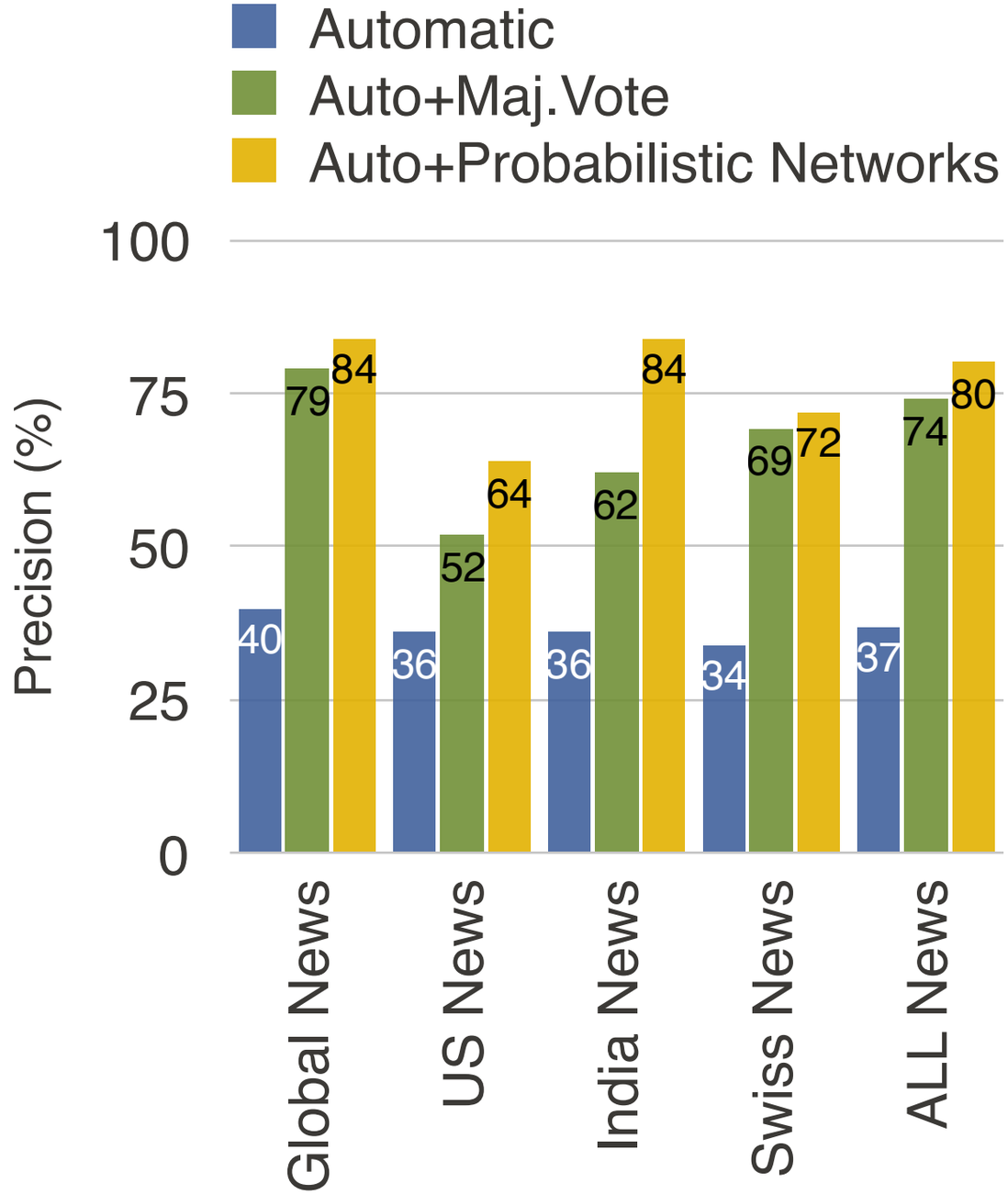
Probabilistic Network for Entity Linking

- Variables
  - Links (l_i)
  - Workers (w_i)
  - Clicks (c_{ij}) observed variable of w_i click for l_j
- Link Factors
- Priors
  - worker prior
  - link prior
- Constraints
  - SameAs Links
  - Unicity (per KB)



ZenCrowd Results

- Experiment
  - 25 news articles
  - Stanford-NER recognizes 383 out of 488 Linkable Entities
- On average, we achieve precision improvement over automatic linking when we use crowdsourcing
- an additional improvement with our probabilistic framework



Q&A

–Djellel Eddine Difallah