

Micro-task Management and Quality Control

Gianluca Demartini
gianlucademartini.net
exascale.info

Types of Crowdsourcing Tasks

Task Granularity	Examples
Complex Tasks	<ul style="list-style-type: none">• Build a website• Develop a software system• Overthrow a government?
Simple Projects	<ul style="list-style-type: none">• Design a logo and visual identity• Write a term paper
Macro Tasks	<ul style="list-style-type: none">• Write a restaurant review• Test a new website feature• Identify a galaxy
Micro Tasks	<ul style="list-style-type: none">• Label an image• Verify an address• Simple entity resolution

Inspired by the report: “Paid Crowdsourcing”, Smartsheet.com, 9/15/2009

Outline

- Micro-task Crowdsourcing Challenges
 - Design the User Interfaces
 - Define the right Incentives
 - Task Patterns
 - Scalability
 - Quality

Design of a Task on MTurk

A Task on MTurk

Choose the best category for this image



- ☐ kitchen
- ☐ living
- ☐ bath
- ☐ bed
- ☐ outside

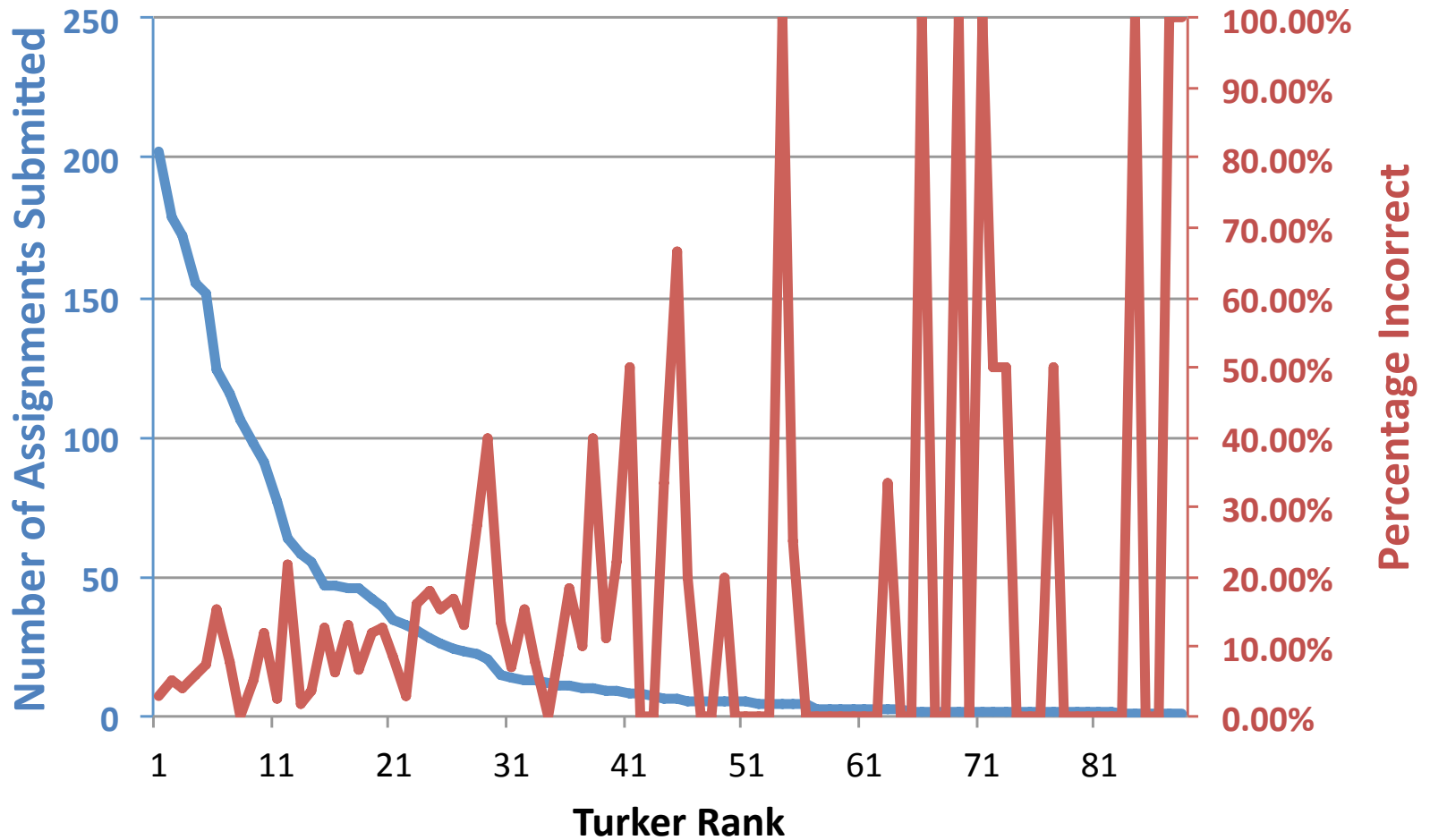
[View Instructions](#)↓

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

High-level Issues in Crowdsourcing

- Process
 - Experimental design, annotation guidelines, iteration
- Choose crowdsourcing platform (or roll your own!)
- Human factors
 - Payment / incentives, interface and interaction design, communication, reputation, recruitment, retention
- Quality Control / Data Quality
 - Trust, reliability, spam detection, consensus labeling

Turker Affinity and Errors



Typical Workflow

- Define and design what to test
- Sample data
- **Design the experiment**
- Run experiment (see later session by Maribel)
- Collect data and analyze results
- **Quality control**

Task Design

- One of the most important parts
- Part art, part science
- Instructions are key
- Prepare to iterate

Task Design

- Ask the right questions
- Workers may not be experts: don't assume the same understanding in terms of terminology
- Show examples
- Hire a technical writer
 - Engineer writes the specification
 - Writer communicates

Task Design - UI

- Generic tips
 - Experiment should be self-contained.
 - Keep it short and simple. Brief and concise.
 - Be very clear with the relevance task.
 - Engage with the worker. Avoid boring stuff.
 - Always ask for feedback (open-ended question) in an input box.

Task Design - UI

- Presentation
- Document design
- Highlight important concepts
- Colors and fonts
- Need to grab attention
- Localization

Other design principles

- Text alignment
- Legibility
- Reading level: complexity of words and sentences
- Attractiveness (worker's attention & enjoyment)
- Multi-cultural / multi-lingual
- Who is the audience (e.g. target worker community)
 - Special needs communities (e.g. simple color blindness)
- Cognitive load: mental rigor needed to perform task

Bad Example

- Asking too much, task not clear, “do NOT/reject”
- Worker has to do a lot of stuff

Help us describe How-To Videos! Earn \$2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the **Film & Theater** how-to video ==> [332492 Get a 35mm film look with a depth of field adapter](#)
2. Write a description of the video linked in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete **25 video descriptions** in order to earn the \$2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is **not necessary** to repeat the headline in your entry. It will **NOT** count toward your word count.
9. Do **NOT** describe the following: the format, where the video comes from, or how long the video is. This information is **IRRELEVANT**.
10. Do **NOT** describe the video in the following manner: "She turns around to face the camera. Then she faces left." Follow the examples below.

Current Word Count: 0 Current Character Count: 0 / 2000

Criteria for REJECTION:


1. Entries with obvious and multiple spelling or grammatical errors will be **rejected**.
2. Entries with fewer than 100 words will be automatically **rejected**.
3. Text copied from the web or other places will be **rejected**. Multiple plagiarized answers will lead to being **BLOCKED**. You may use a quotation, but the majority of your content must be **ORIGINAL**.
4. Incomplete and blank answers will be **rejected**. Multiple blank answers will result in being **blocked**.
5. Tasks submitted without descriptions will be **rejected**.
6. Tasks submitted with inaccurate descriptions will be **rejected** as well.
7. Do **NOT** add any personal opinions. Entries with personal opinions or reviews will be automatically **REJECTED**.
8. If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in **rejection**.
9. Entries that transcribe the video will be **REJECTED**.

Good Example

- All information is available
 - What to do
 - Search result
 - Question to answer

Task

Please evaluate the relevance of the following document for the query **milton keynes**.



The screenshot shows a Bing search results page. At the top, there are navigation links for Web, Images, Videos, Shopping, News, Maps, More, MSN, and Hotmail. The search bar contains the text 'milton keynes'. Below the search bar, there are links for 'Milton Keynes Map', 'Milton Keynes Restaurants', and 'Milton Keynes Hotels'. The main search results section shows 'ALL RESULTS' for '1-20 of 7,020,000 results'. The first result is 'Milton Keynes - Wikipedia, the free encyclopedia', which describes Milton Keynes as a large town in Buckinghamshire, England, about 45 miles (72 km) north-west of London. To the right of the search results, there is a 'Sponsored sites' section with a link to 'Milton Keynes Hotels' and a promotional message about saving up to 50% on hotels.

Please rate the above document according to its relevance to **milton keynes** as follows. Note that the task is about how relevant to the topic the document is.

☐ **Relevant.** A relevant document for the topic.

☐ **Not relevant.** The document is not good because it doesn't contain any relevant information.

Form and Metadata

- Form with a close question (binary relevance) and open-ended question (user feedback)
- Clear title, useful keywords
- Workers need to find your task

Describe your HIT

Title

Pick the best category

Describe the task to workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so workers know what to expect.

Description

Pick the best category

Give more detail about this task. This gives workers a bit more information before they decide to view your HIT.

Keywords

category, categorize

Provide keywords that will help workers search for your HITs.

How Much to Pay?

- Price commensurate with task effort
 - Ex: \$0.02 for yes/no answer + \$0.02 bonus for optional feedback
- Ethics & market-factors
 - e.g. non-profit SamaSource contracts workers refugee camps
- Uptake & time-to-completion vs. Cost & Quality
 - Too little \$\$, no interest or slow
 - too much \$\$, attract spammers
- Accuracy & quantity
 - More pay = more work, not better (W. Mason and D. Watts, 2009)

Development Framework

- Similar to a UX
- Build a mock up and test it with your team
 - Yes, you need to judge some tasks
- Incorporate feedback and run a test on MTurk with a very small data set
 - Time the experiment
 - Do people understand the task?
- Analyze results
 - Look for spammers
 - Check completion times
- Iterate and modify accordingly

Development Framework

- Introduce quality control
 - Qualification test
 - Gold answers (honey pots)
- Adjust passing grade and worker approval rate
- Run experiment with new settings & same data
- Scale on data
- Scale on workers

Summary

- Micro-task Crowdsourcing Challenges
 - Design the User Interfaces
 - Define the right Incentives
 - **Task Patterns**
 - **Scalability**
 - Quality (more in the next session)

Crowdsourcing Patterns

- Majority Vote Aggregation
 - Select the answer among a set of candidates
 - Pick the most popular answer
- Find-Fix-Verify
 - Creative process
 - Three-steps iterative crowdsourcing
- Interaction Protocol (for hybrid human-machine systems)
 - Upfront
 - Iterative

Interaction Protocol

How often can we refer to the crowd?

1. **Upfront:** Ask all the B queries at once
2. **Iterative:** Ask K queries to the crowd and use them to improve the system. Repeat this B/K times

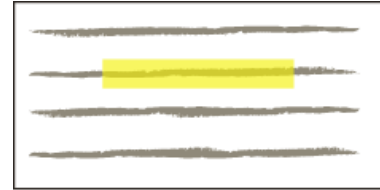
Measures Used for Selection

- **Uncertainty:** Asking hardest (most ambiguous) questions
- **Explorer:** Ask questions with potential to have largest impact on the system

Soylent: Find-Fix-Verify

Find

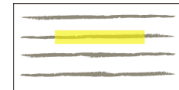
“Identify at least one area that can be shortened without changing the meaning of the paragraph.”



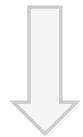
Independent agreement to identify patches

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...

Find-Fix-Verify

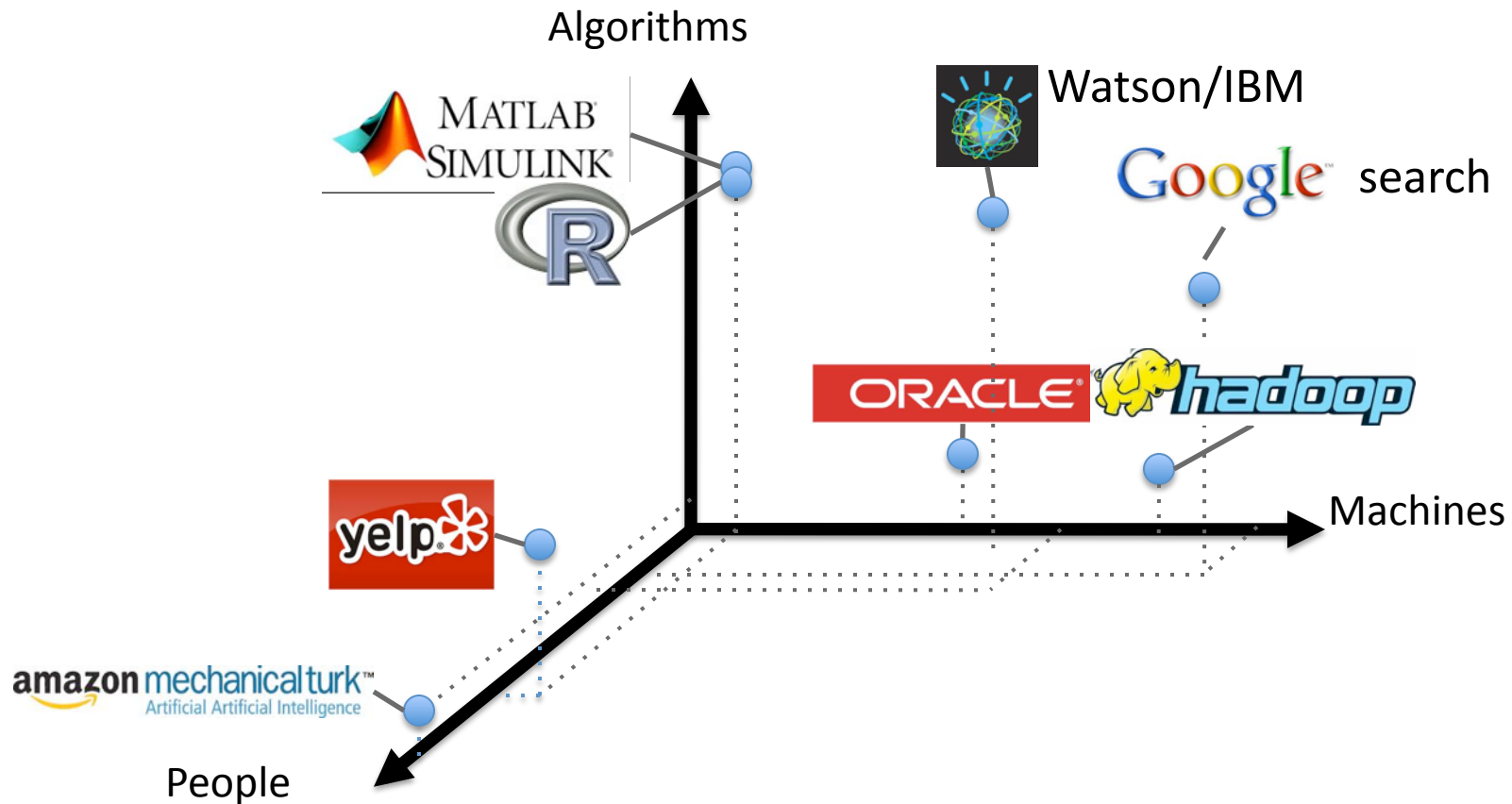
- Machine Translation example
- Find
 - Show automatically translated text
 - Ask if they are grammatically correct
- Fix
 - Ask to translate those which contain errors (multiple times)
- Verify
 - Select the best translation among the available ones

Micro-task Automation: Hybrid Human Machine Systems

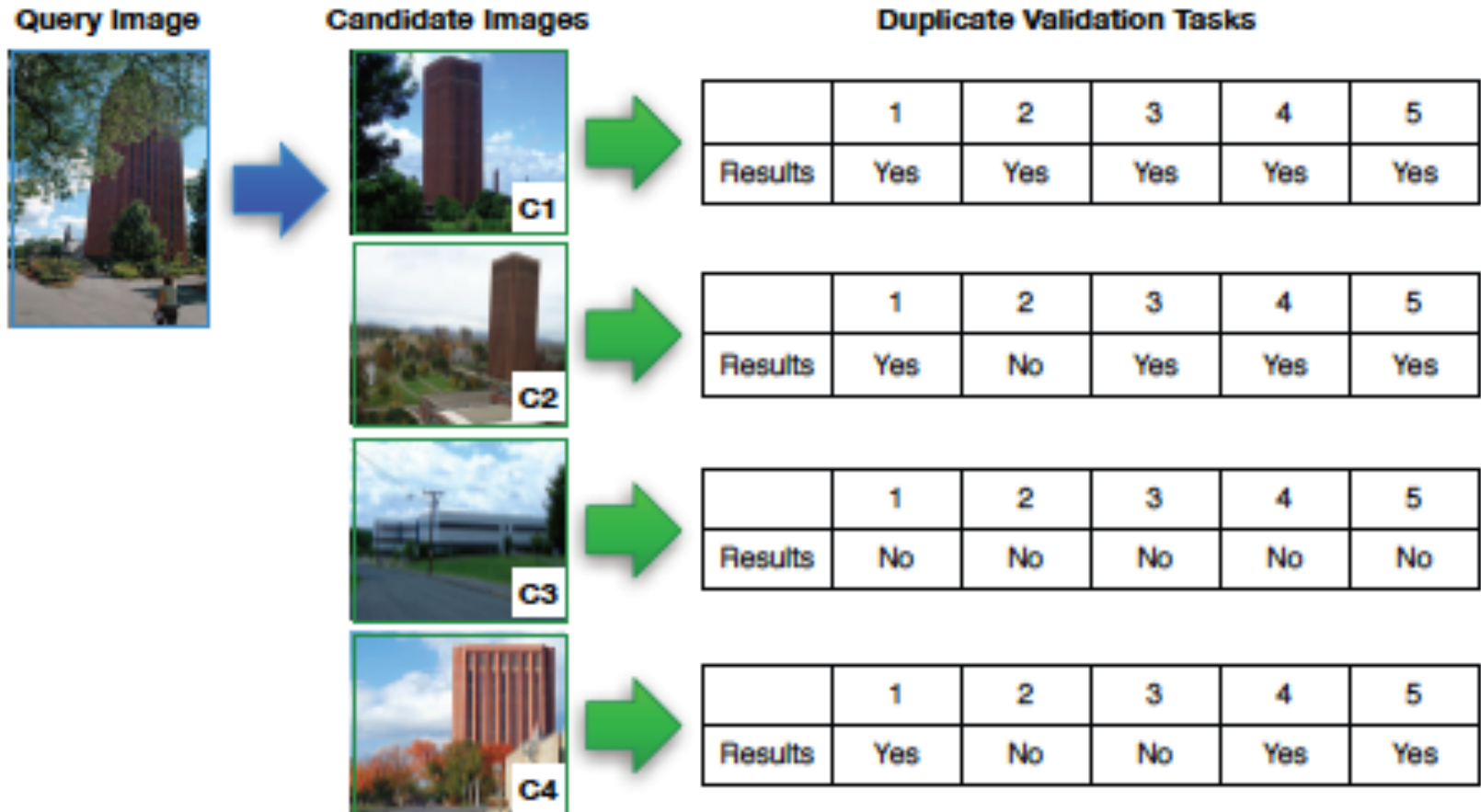
Hybrid Systems: Key Issues

- The role of machine (i.e., algorithm) and humans
 - use only humans? both? who's doing what?
- Quality control
- **Optimization: What to crowdsource**
- **Scalability: How much to crowdsource**

Thinking About Hybrid Systems



Example: Hybrid Image Search



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

Example: Hybrid Data Integration

paper	conf
Data integration	VLDB-01
Data mining	SIGMOD-02

title	author	email	venue
OLAP	Mike	mike@a	ICDE-02
Social media	Jane	jane@b	PODS-05

- **Generate plausible matches**

- paper = title, paper = author, paper = email, paper = venue
- conf = title, conf = author, conf = email, conf = venue

- **Ask users to verify**

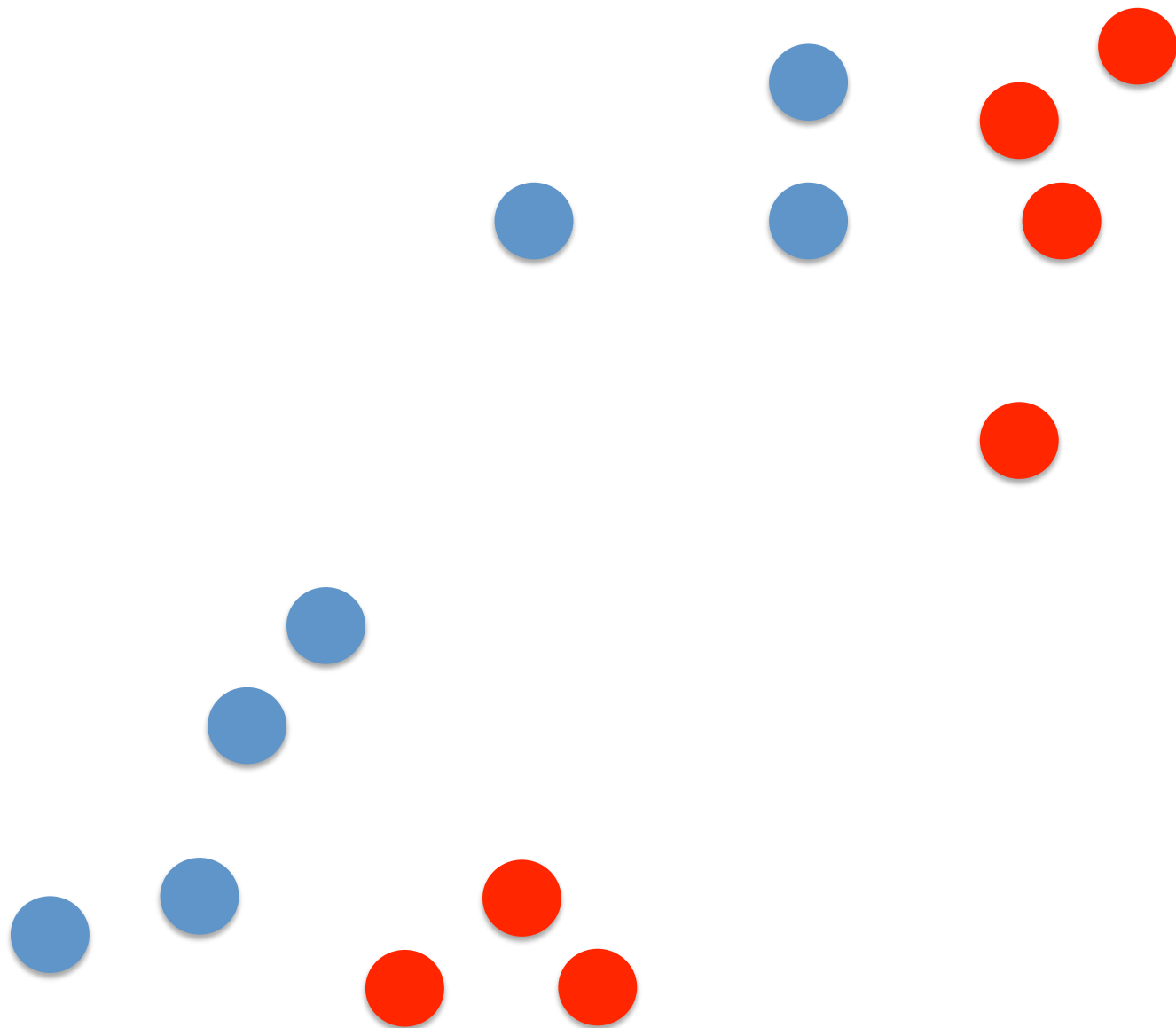
Does attribute **paper** match attribute **author**?

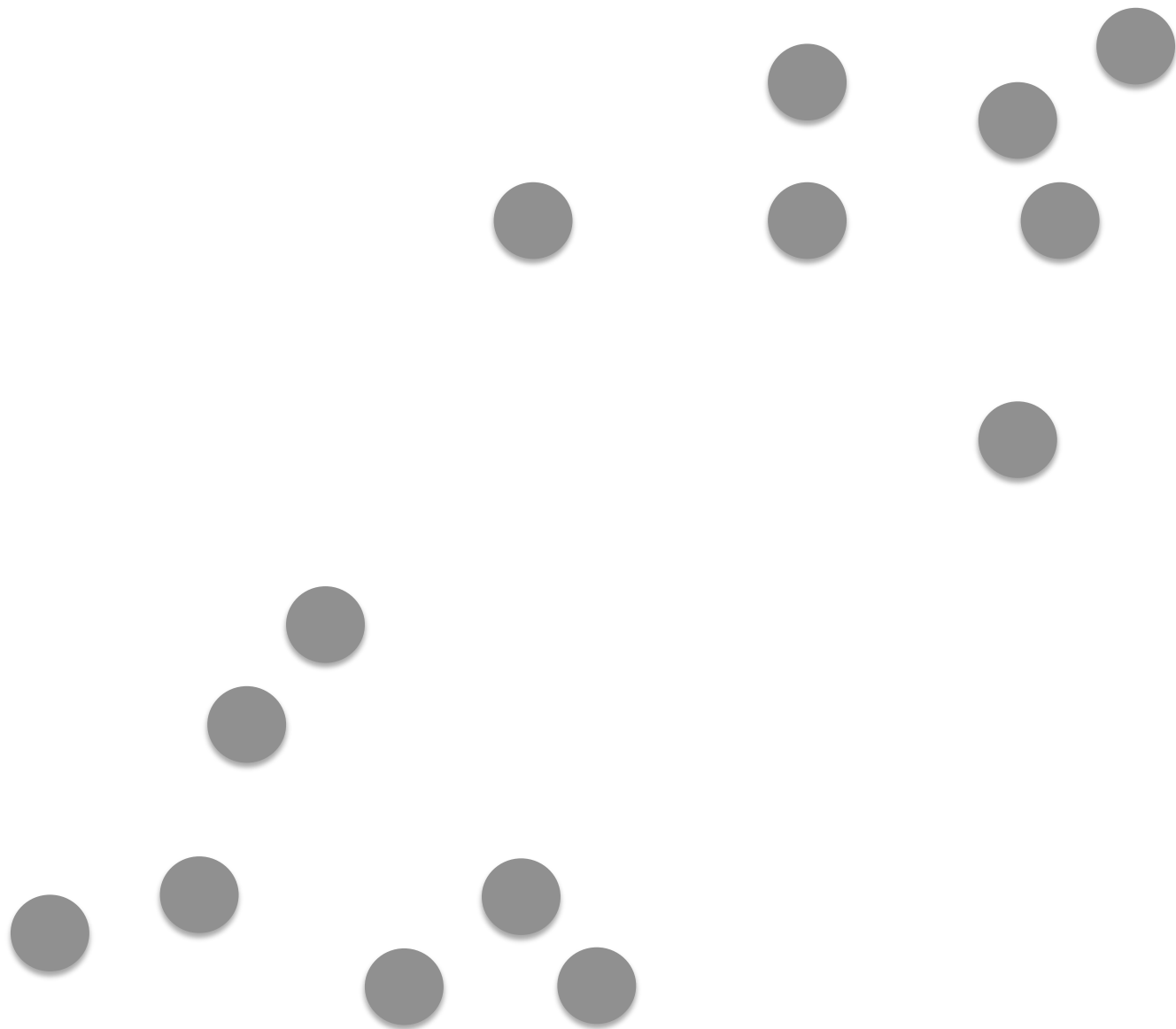
paper	conf
Data integration	VLDB-01
Data mining	SIGMOD-02

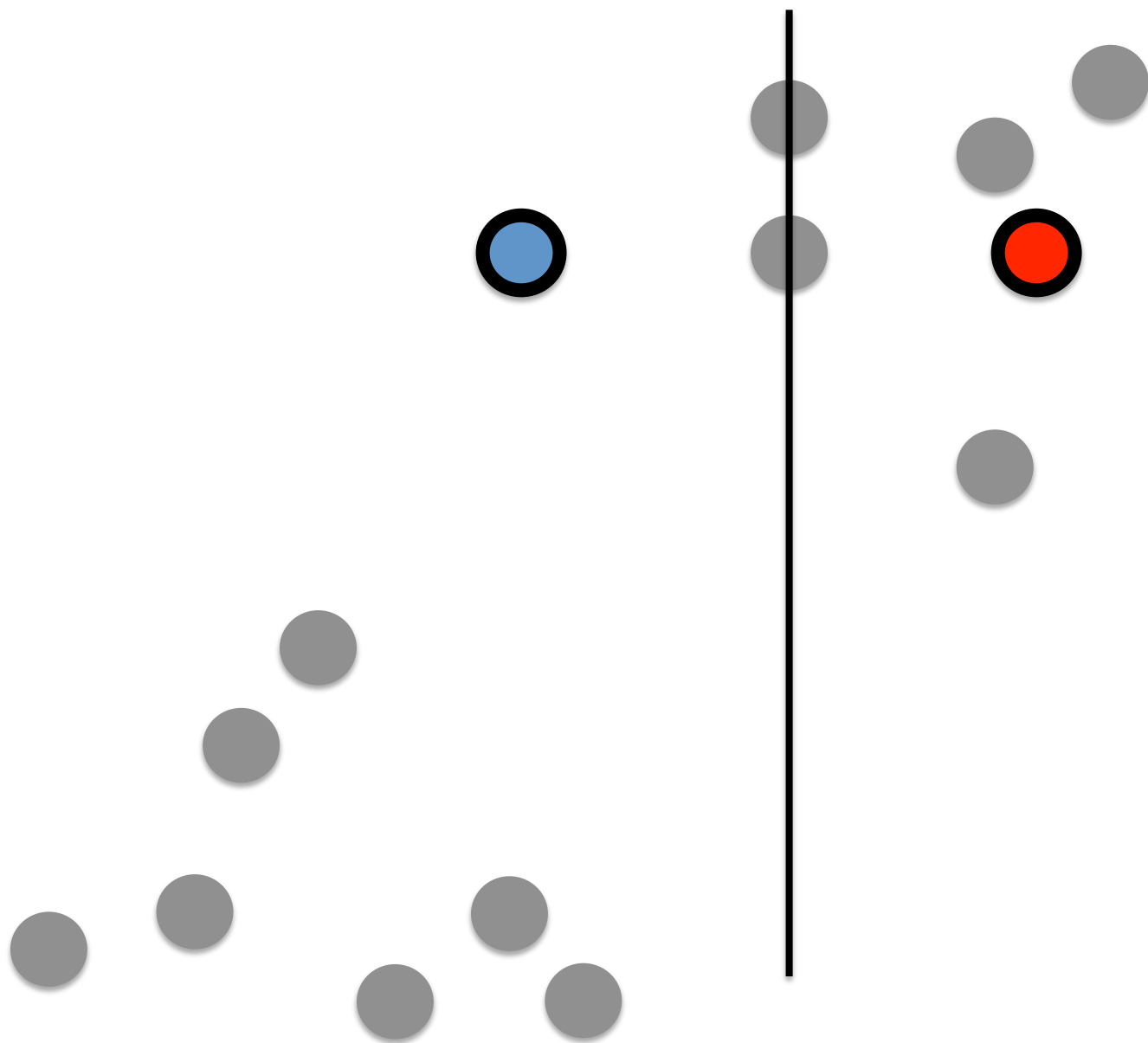
title	author	email
OLAP	Mike	mike@a
Social media	Jane	jane@b

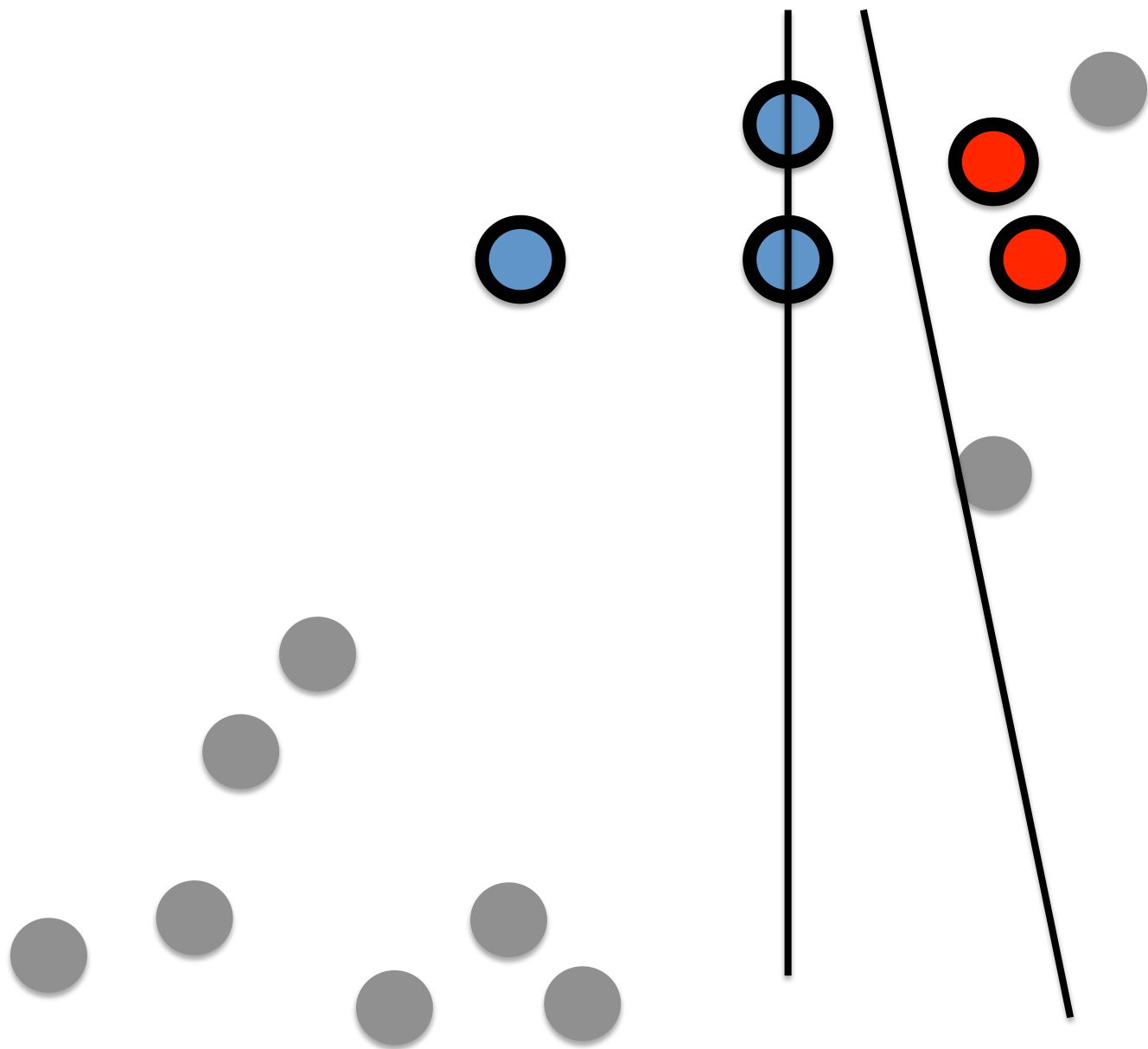
Crowdsourcing Scalability

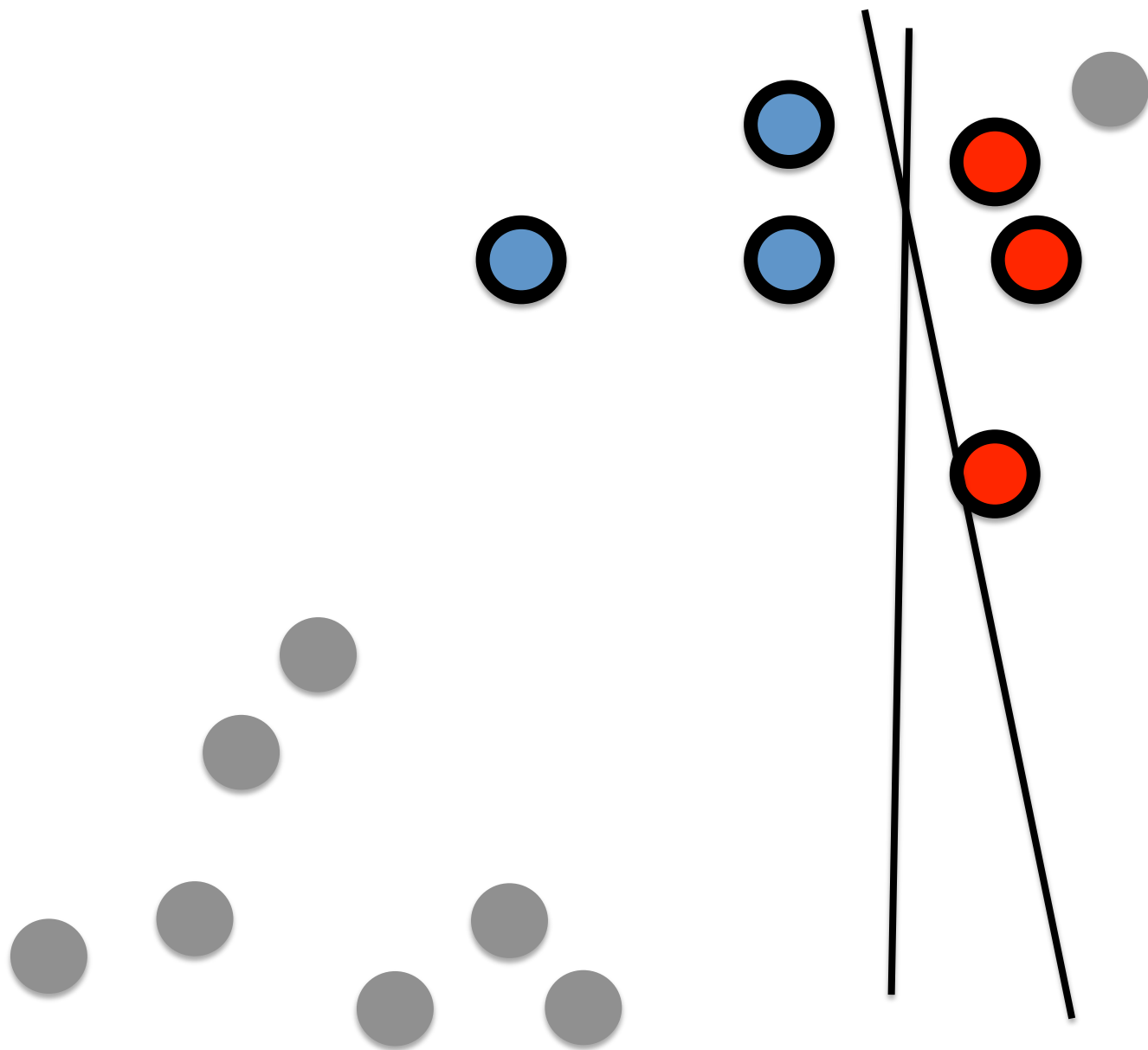
- Crowd-sourcing is becoming an indispensable method of collecting labeled data, e.g., Machine Learning
- BUT crowd-sourcing can be expensive, slow, and noisy
- All Human Intelligent Tasks (HIT) *are NOT equally difficult for the machine*
- To achieve scalability, we need to know when and how to use machines along with humans

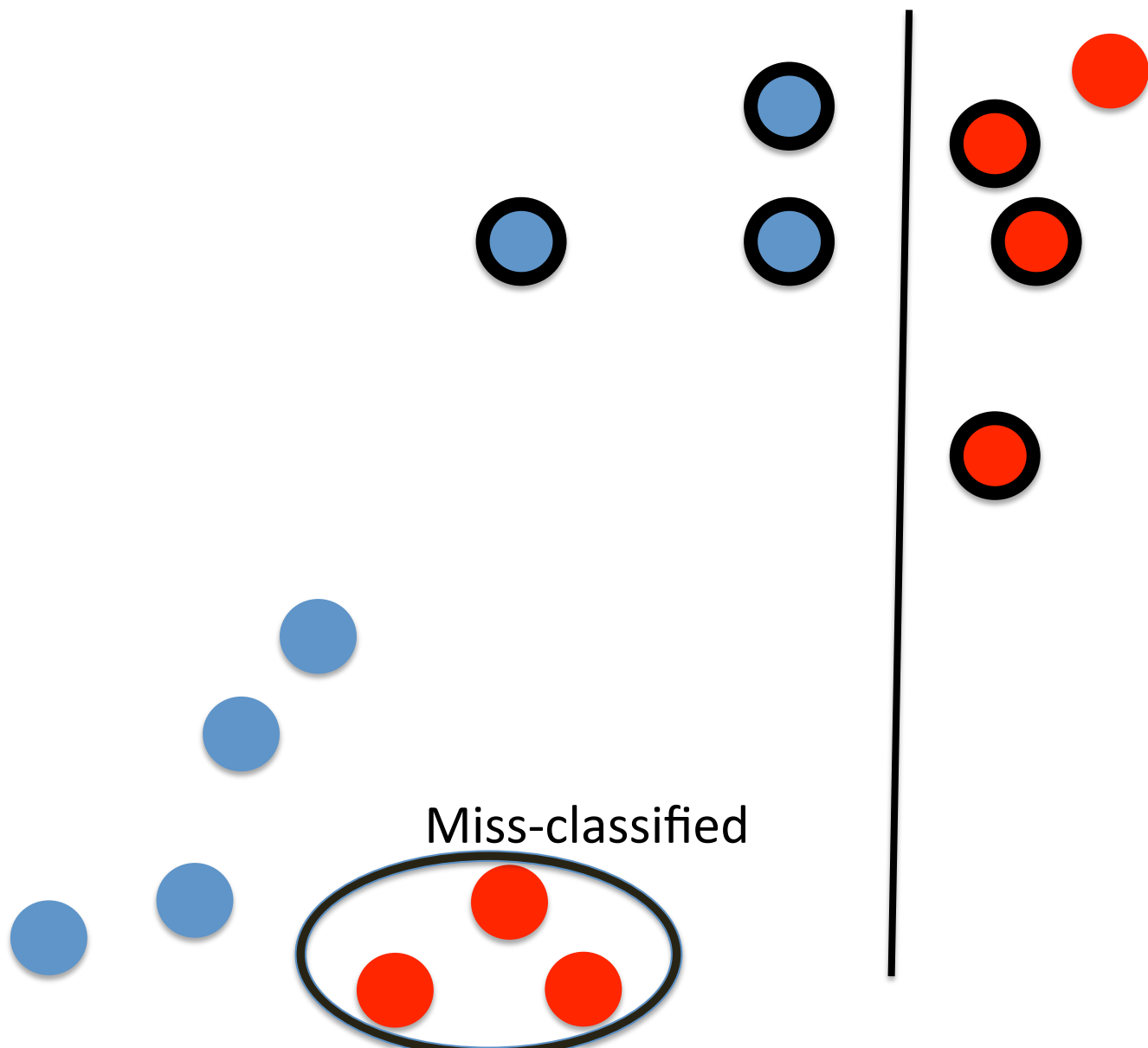


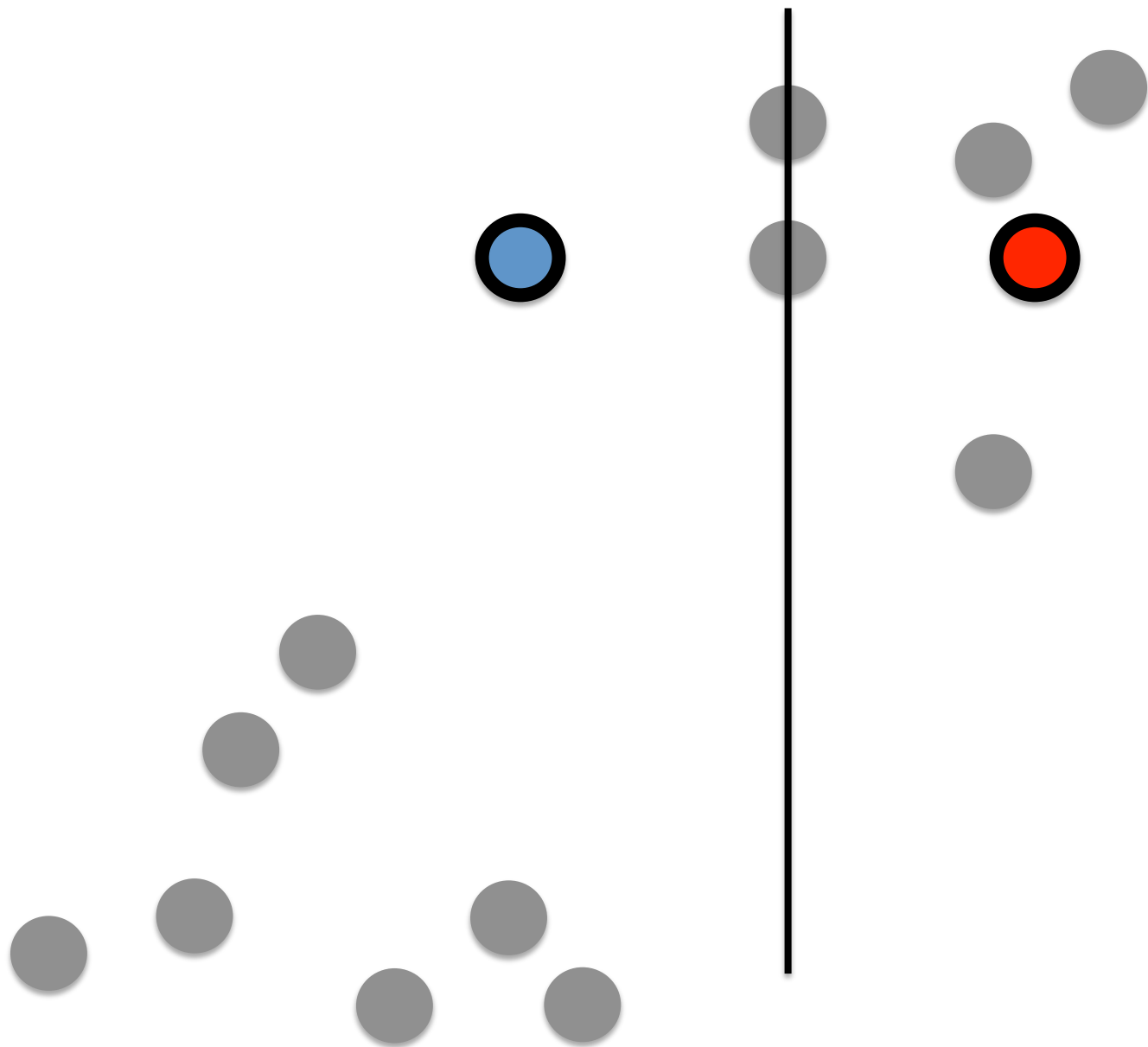


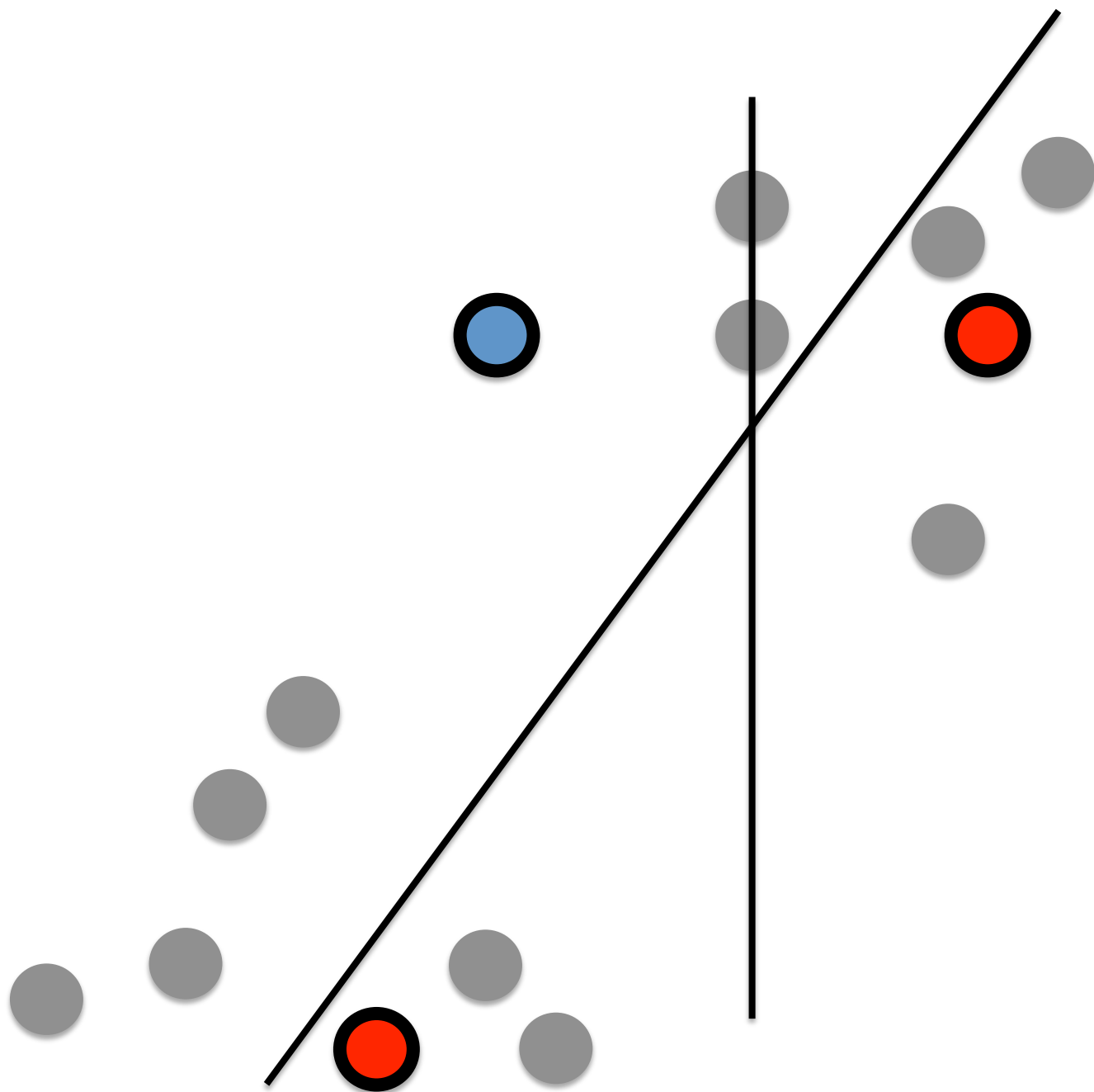


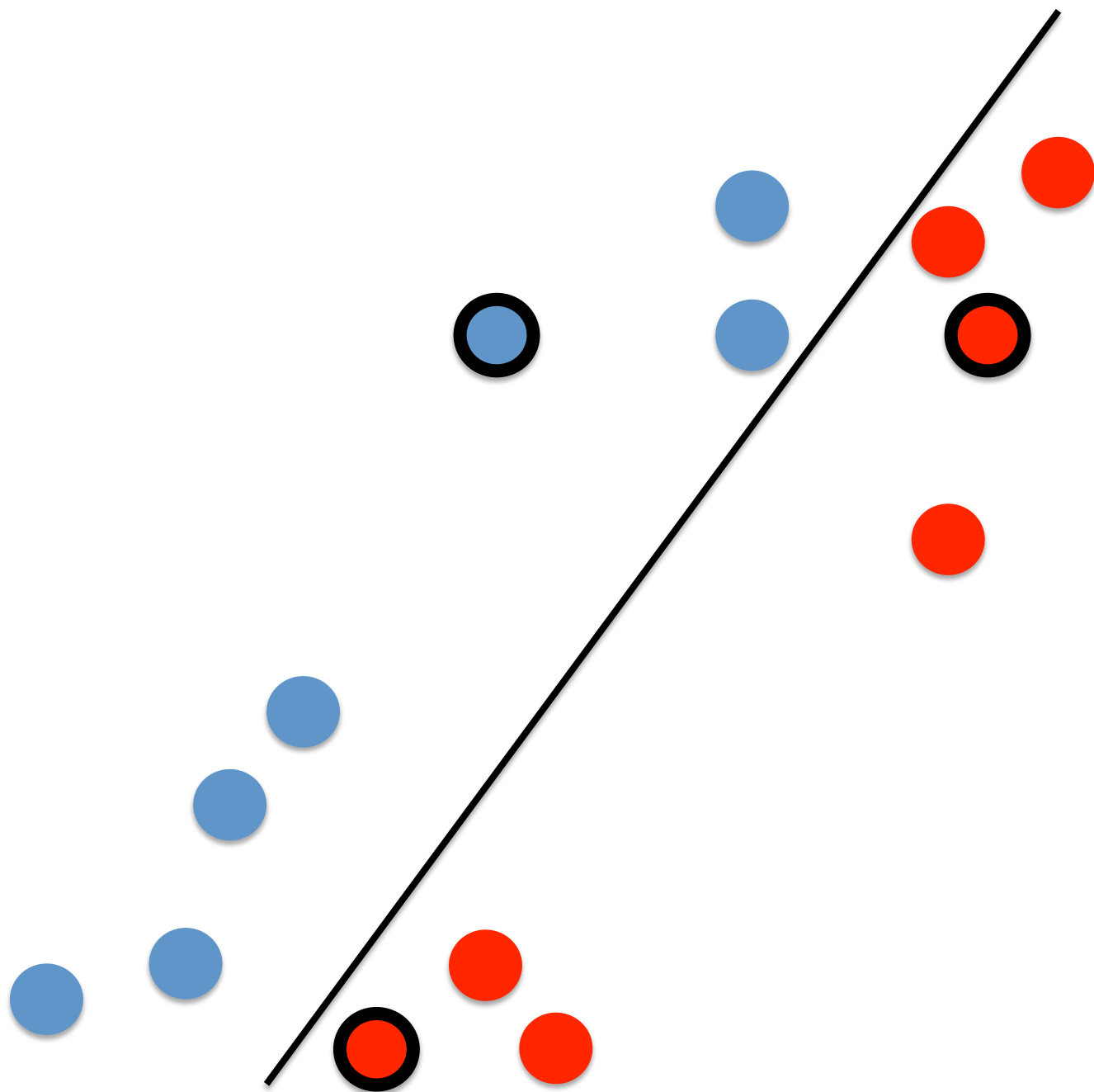












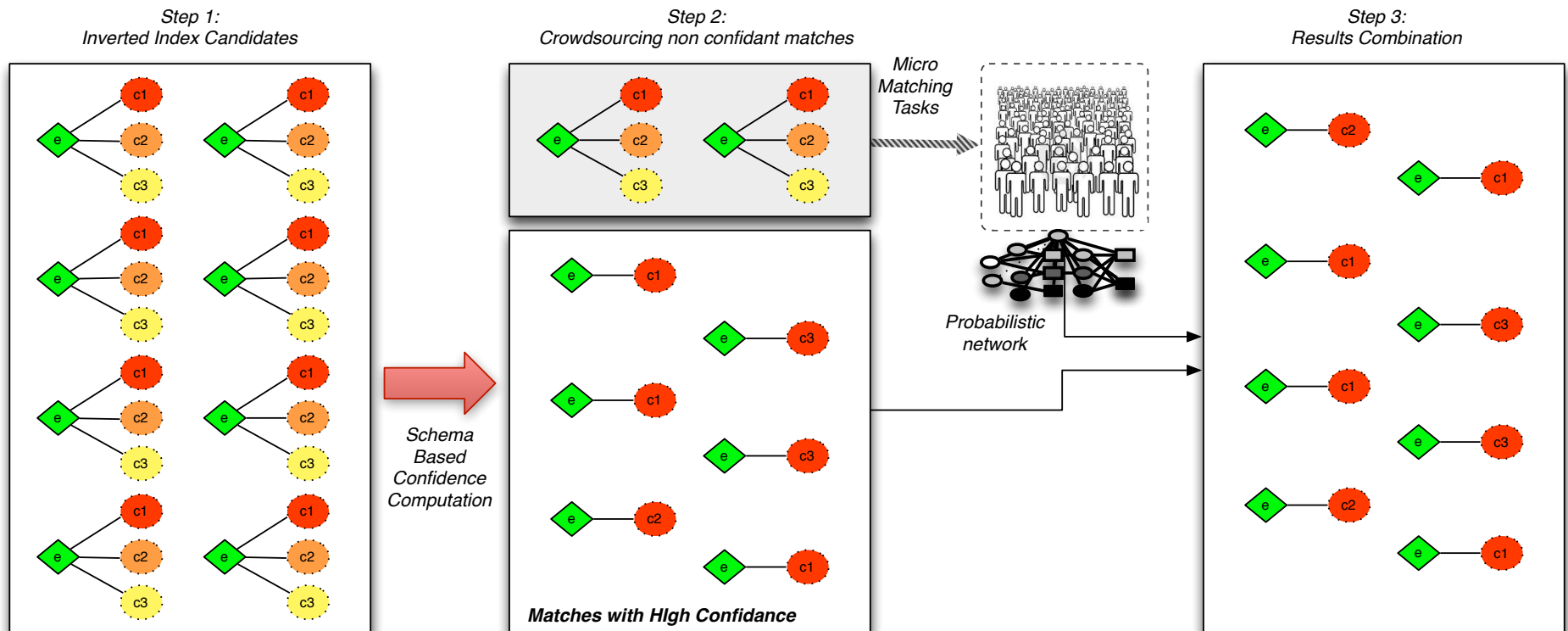
Blocking for Instance Matching

- Find the instances about the same real-world entity within two datasets
- Avoid Comparison of all possible pairs
 - Step 1: cluster similar items using a cheap similarity measure
 - Step 2: $n*n$ comparison within the clusters with an expensive measure

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Large-Scale Linked Data Integration Using Probabilistic Reasoning and Crowdsourcing. In: VLDB Journal, Volume 22, Issue 5 (2013), Page 665-687, Special issue on Structured, Social and Crowd-sourced Data on the Web. October 2013.

3-steps Blocking with the Crowd

- Crowdsourcing as the most expensive similarity measure



Conclusions




- Carefully design the User Interface
- Define the right Incentives
- Use Task Patterns
- Enable Scalability
- Quality (more in the next session)


Quality Control in Micro-Task Crowdsourcing


Quality Control


- Extremely important part of the experiment
- Approach as “overall” quality; not just for workers
- Bi-directional channel
 - You may think the worker is doing a bad job.
 - The same worker may think you are a lousy requester.
 - Do check the worker forums!


Crowd Worker Communities

Rating [info]	Description
FAIR: 5 / 5 	No need to contact, HITs approved next day.
FAST: 5 / 5 	Jan 21 2013 rjsc...@g... flag comment
PAY: 5 / 5 	
COMM: NO DATA	

communicativity:  5 / 5

generosity :  5 / 5

fairness :  5 / 5

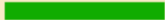


promptness :  4.71 / 5

[What do these scores mean?](#)

Scores based on [7 reviews](#)

[Report your experience with this requester »](#)

Turkopticon.com
Mturkforum.com
Turkernation.com

FAIR: 5 / 5 	Small batch and mega bubbles. Not sure if I'm going in....
FAST: 4 / 5 	Title: Which is the most appropriate type?
PAY: 5 / 5 	Requester: Philippe Cudre-Mauroux [A28PIN9Y6KHR3H] (TO)
COMM: NO DATA	Description: Please read the text and select the most appropriate description for each of the proposed entities.
	Reward: \$0.10
	Qualifications: HIT abandonment rate (%) is less than 51, HIT approval rate (%) is greater than 25, Location is US
	Link: https://www.mturk.com/mturk/preview?groupId=2ZSQUQIHPCGJ2FZIT6N51H1LQYU60M

Powered by non-amazonian script monkeys ♦♦

To many bubbles but YMMV with your patience level.

Quality Control

- Approval rate: easy to use, & just as easily defeated
- Mechanical Turk Masters (since June 2011)
 - Recent addition, only for specific tasks
- Qualification test
 - Pre-screen workers' ability to do the task (accurately)
- Assess worker quality as you go
 - Trap questions with known answers (“honey pots”)
 - Measure inner-annotator agreement between workers

Qualification tests: pros and cons

- Advantages
 - Great tool for controlling quality
 - Adjust passing grade
- Disadvantages
 - Extra cost to design and implement the test
 - May turn off workers, hurt completion time
 - Refresh the test on a regular basis
 - Hard to verify subjective tasks like judging relevance
- Try creating task-related questions to get worker familiar with task *before* starting task in earnest

Methods for measuring agreement

- What to look for
 - Agreement, reliability, validity
- Inter-agreement level
 - Agreement between judges
 - Agreement between judges and the gold set
- Some statistics
 - Percentage agreement
 - Cohen's kappa (2 raters)
 - Fleiss' kappa (any number of raters)
- With majority vote, what if 2 say relevant, 3 say not?
 - Use expert to break ties
 - Collect more judgments as needed to reduce uncertainty

Quality Control & Assurance

- Filtering
 - Approval rate (built-in but defeatable)
 - Geographic restrictions (e.g. US only, built-in)
 - Worker blocking
 - Qualification test
 - Con: slows down experiment, difficult to “test” relevance
 - Solution: create questions to let user get familiar *before* the assessment
 - Does not guarantee success
- Identify workers that *always* disagree with the majority
- Ask workers to rate the difficulty of a task

Other quality heuristics

- Justification/feedback as quasi-captcha
 - Should be optional
 - Automatically verifying feedback was written by a person may be difficult (classic spam detection task)
- Broken URL/incorrect object
 - Leave an outlier in the data set
 - Workers will tell you
 - If somebody answers “excellent” for a broken URL
=> *probably* spammer

Dealing with bad workers

- Pay for “bad” work instead of rejecting it?
 - Pro: preserve reputation, admit if poor design at fault
 - Con: promote fraud, undermine approval rating system
- Use bonus as incentive
 - Pay the minimum \$0.01 and \$0.01 for bonus
 - Better than rejecting a \$0.02 task
- If spammer “caught”, block from future tasks
 - May be easier to always pay, then block as needed

Build Your Reputation as a Requestor

- Word of mouth effect
 - Workers trust the requester (pay on time, clear explanation if there is a rejection)
 - Experiments tend to go faster
 - Announce forthcoming tasks (e.g. tweet)

Answer justification

- Why settle for a label?
- Let workers justify answers
- INEX (Initiative for the Evaluation of XML Retrieval)
 - 22% of assignments with comments
- Has to be optional for good feedback

Gamification of IR Evaluation

- GeAnn: <http://www.geann.org/>
- Relevance judgments with Gamification:
 - Text relevance
 - Image relevance

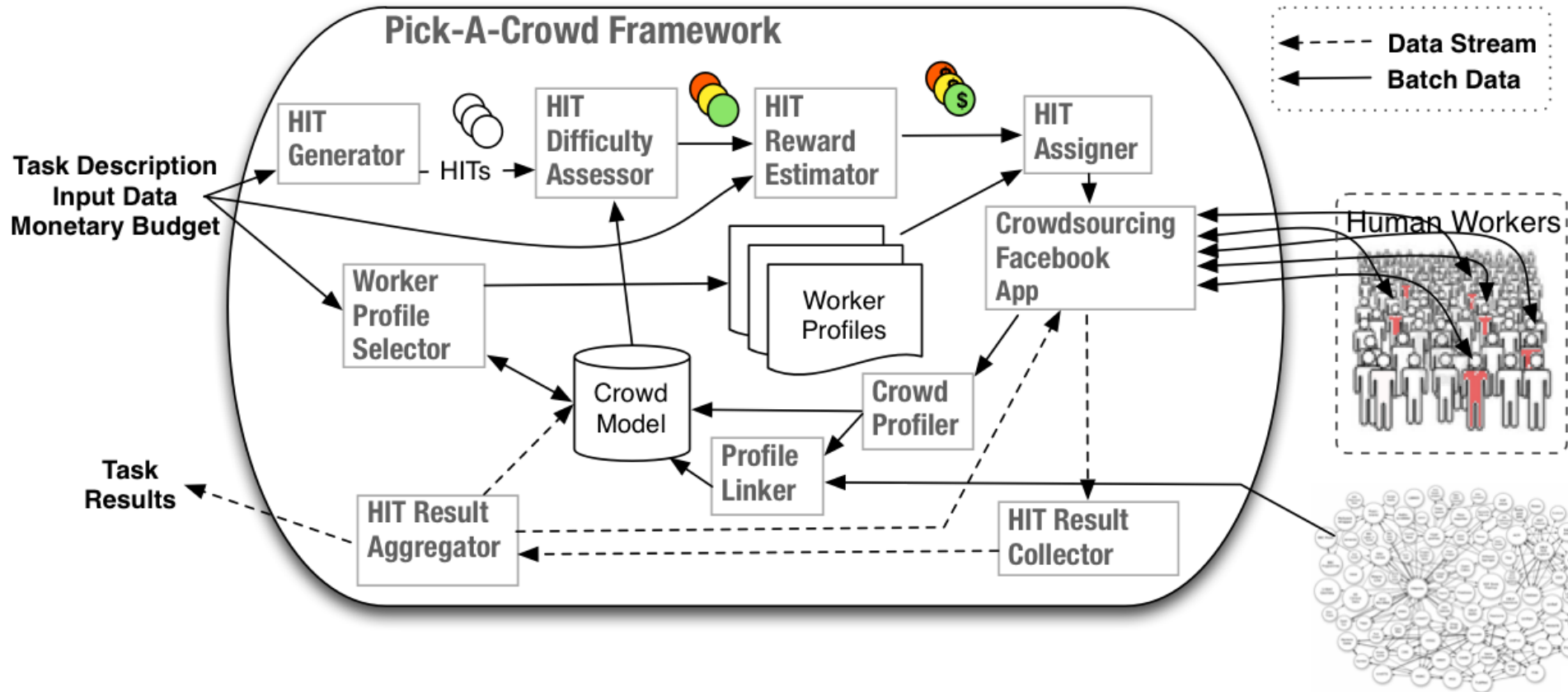
Quality through Flow and Immersion: **Gamifying Crowdsourced Relevance Assessments**. Eickhoff, C., C. G. Harris, A. P. de Vries, and P. Srinivasan. SIGIR 2012.

Summary

- Things that work
 - Qualification tests
 - Honey-pots
 - Good content and good presentation
 - Economy of attention
- Things to improve
 - Manage workers in different levels of expertise including spammers and potential cases.
 - Mix different pools of workers based on different profile and expertise levels.

Modeling Crowd Workers (via Social Network Profiles)

Pick-A-Crowd



Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux.
Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do.
 In: 22nd International Conference on World Wide Web (WWW 2013)

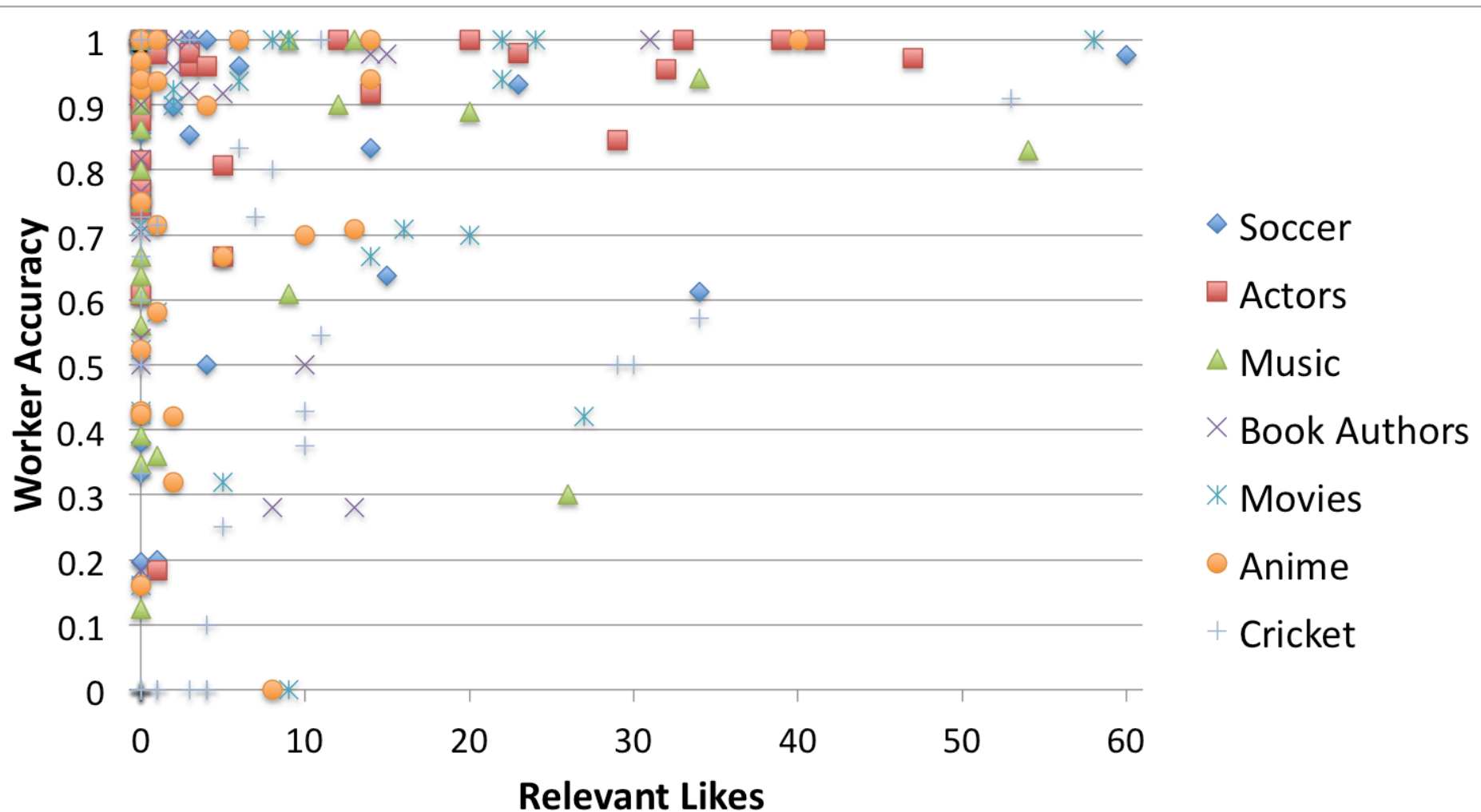
Batch

Batch description	Challenge	Number of tasks	Reward
🔗 Football players identifications	 Recommend 5	Completed	\$0.25
🔗 What movie is this scene from?	 Recommend 9	31 available	\$0.25
🔗 Comics, mangas and characters	 Recommend 5	41 available	For Fun

Batch

Batch description	Challenge	Number of tasks	Reward
 Actors identification	 Recommend { 8	40 available	\$0.25
 Music bands identification	 Recommend { 4	31 available	\$0.25
 Book authors identification	 Recommend { 5	48 available	\$0.25
 Cricket questions.	 Recommend { 8	11 available	\$0.25

Like vs Accuracy



References

- **“Crowdsourcing for Information Retrieval: Principles, Methods, and Applications” SIGIR 2011 Tutorial.**
- **“Crowdsourcing for Search Evaluation and Social-Algorithmic Search” SIGIR 2012 Tutorial.**
- **“When to Ask a Noisy Crowd: Active Learning Meets Crowd” Barzan Mozafari et al.**