DSE 203 (Fall 2020)

Human in the Loop Data Integration (Crowdsourcing)

Entity Resolution Again

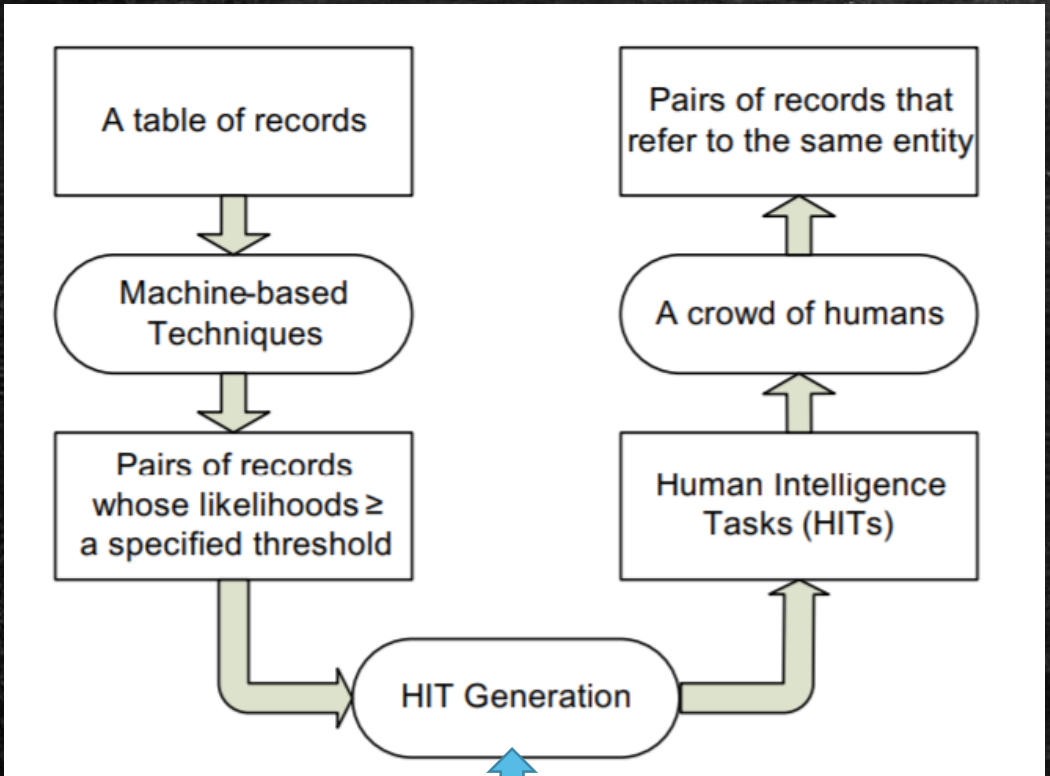
ID	Product Name	Price
r_1	iPad Two 16GB WiFi White	$490
r_2	iPad 2nd generation 16GB WiFi White	$469
r_3	iPhone 4th generation White 16GB	$545
r_4	Apple iPhone 4 16GB White	$520
r_5	Apple iPhone 3rd generation Black 16GB	$375
r_6	iPhone 4 32GB White	$599
r_7	Apple iPad2 16GB WiFi White	$499
r_8	Apple iPod shuffle 2GB Blue	$49
r_9	Apple iPod shuffle USB Cable	$19

Records r1 and r2 in the table have different text in the Product Name field, but refer to the same product

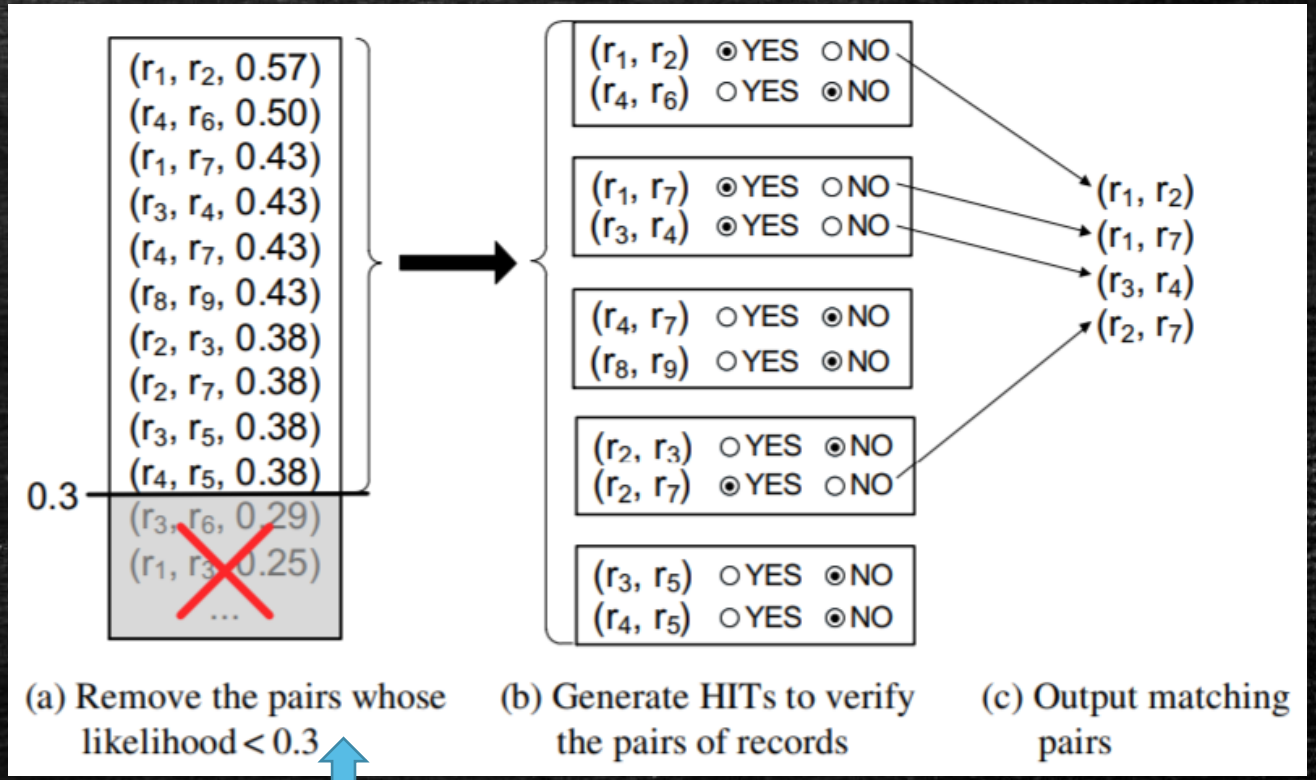
Our Conceptual Goal

SELECT p.id, q.id FROM product p, product q WHERE p.product_name ~= q.product_name;

The Human-Machine Solution



Human Intelligence Task



Likelihood computed by Jaccard similarity between Product Names

Amazon Mechanical Turk (AMT)

- AMT supports microtasks
  - Microtasks usually do not require any special training and typically take no longer than one minute to complete
  - A requester defines a price/reward (minimum $0.01) that the worker receives if the task is completed satisfactorily
- HIT
  - A Human Intelligent Task, or HIT, is the smallest entity of work a worker can accept to do. HITs contain one or more jobs.
- Assignment
  - Every HIT can be replicated into multiple assignments.
  - AMT ensures that any particular worker processes at most a single assignment for each HIT, enabling the requester to obtain answers to the same HIT from multiple workers
  - Odd numbers of assignments per HIT enable majority voting for quality assurance – 3 or 5 per HIT
- HIT Group
  - AMT automatically groups similar HITs together into HIT Groups based on the requester, the title of the HIT, the description, and the reward.
  - A HIT Group could contain 50 HITs, each HIT asking the worker to classify several pictures.
  - Workers typically choose which work to perform based on HIT Groups.

AMT API – (1)

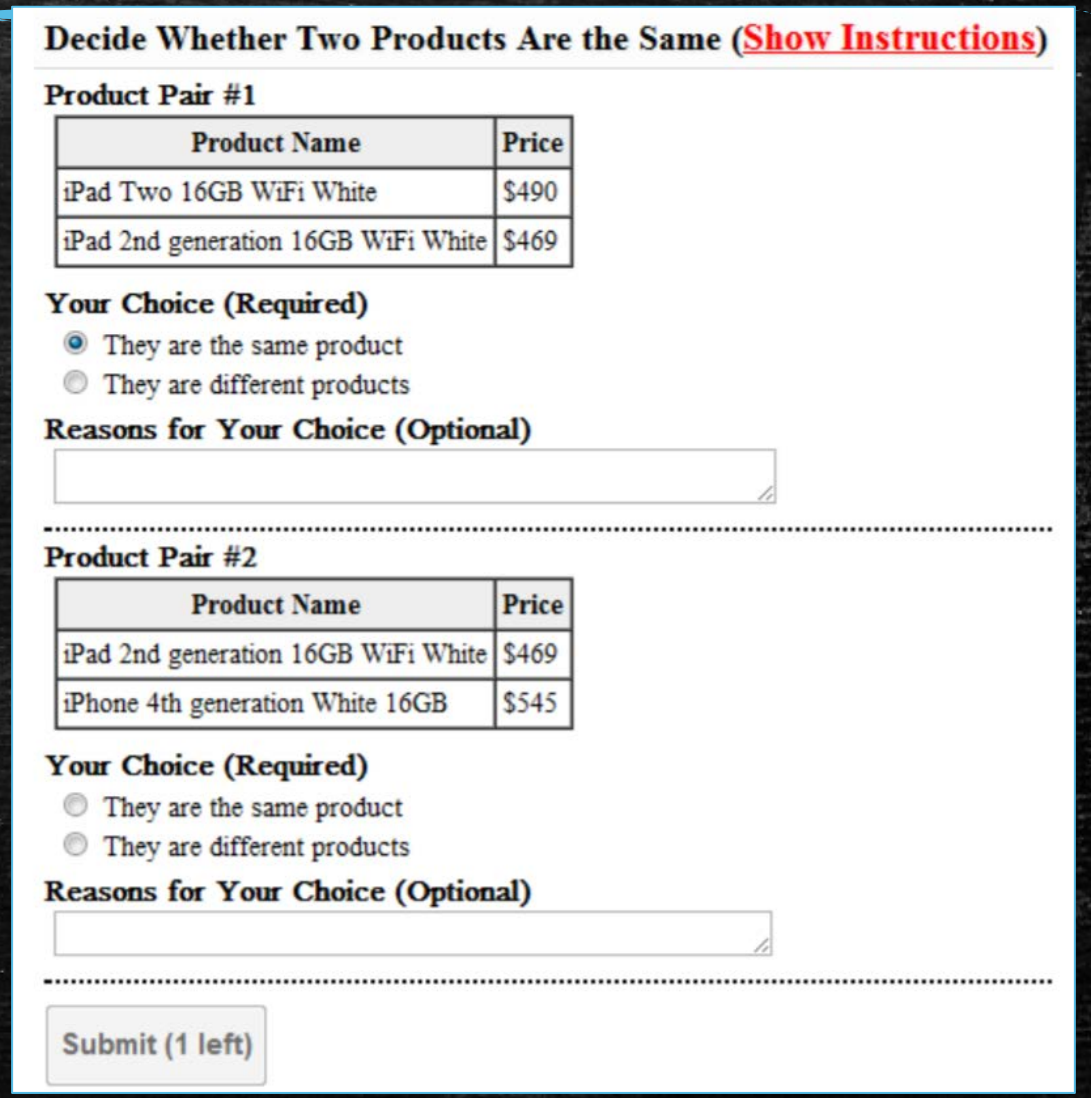
- createHIT(title, description, question, keywords, reward, duration, maxAssignments, lifetime) → HitID
- Calling this method creates a new HIT on the AMT marketplace. The createHIT method returns a HitID to the requester that is used to identify the HIT for all further communication. The title, description, and reward and other fields are used by AMT to combine HITs into HIT Groups.
- The question parameter encapsulates the user interface that workers use to process the HIT, including HTML pages.
- The duration parameter indicates how long the worker has to complete an assignment after accepting it. The lifetime attribute indicates an amount of time after which the HIT will no longer be available for workers to accept.
- Requesters can also constrain the set of workers that are allowed to process the HIT.

AMT API (2)

- getAssignmentsForHIT(HitID)→list(asnId, workerId , answer)
  - Returns the results of all assignments of a HIT that have been provided by workers
    - At most, maxAssignments answers as specified when the requester created the HIT
  - Each answer of an assignment is given an asnID which is used by the requester to approve or reject that assignment
- approveAssignment(asnID) / rejectAssignment(asnID)
  - Approval triggers the payment of the reward to the worker and the commission to Amazon
- forceExpireHIT(HitID)
  - Expires a HIT immediately. Assignments that have already been accepted may be completed

Pair-Based HIT Generation

- A pair-based HIT should contain at most *k* pairs
- Given a set of pairs, *P*, we need to generate $\lceil |P| / k \rceil$ pair-based HITs



Decide Whether Two Products Are the Same (Show Instructions)

Product Pair #1

Product Name	Price
iPad Two 16GB WiFi White	$490
iPad 2nd generation 16GB WiFi White	$469

Your Choice (Required)

- (•) They are the same product
- ( ) They are different products

Reasons for Your Choice (Optional)

Product Pair #2

Product Name	Price
iPad 2nd generation 16GB WiFi White	$469
iPhone 4th generation White 16GB	$545

Your Choice (Required)

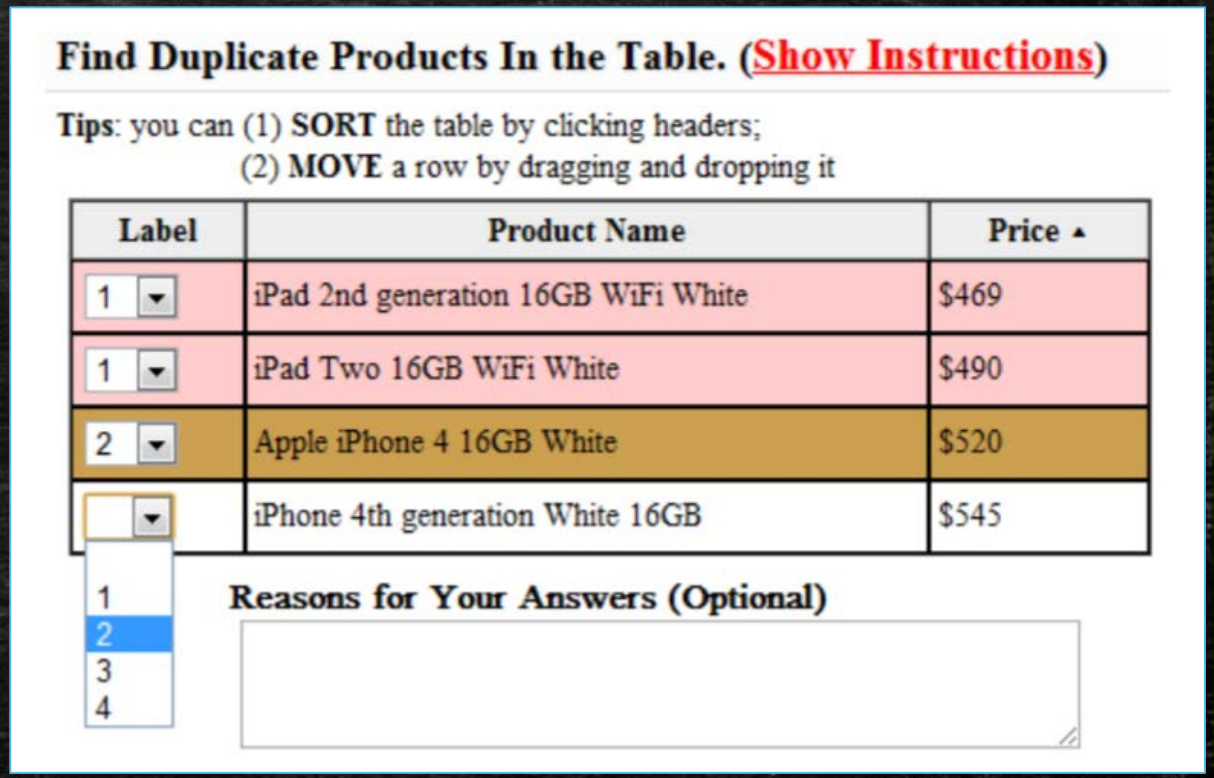
- ( ) They are the same product
- ( ) They are different products

Reasons for Your Choice (Optional)

Submit (1 left)

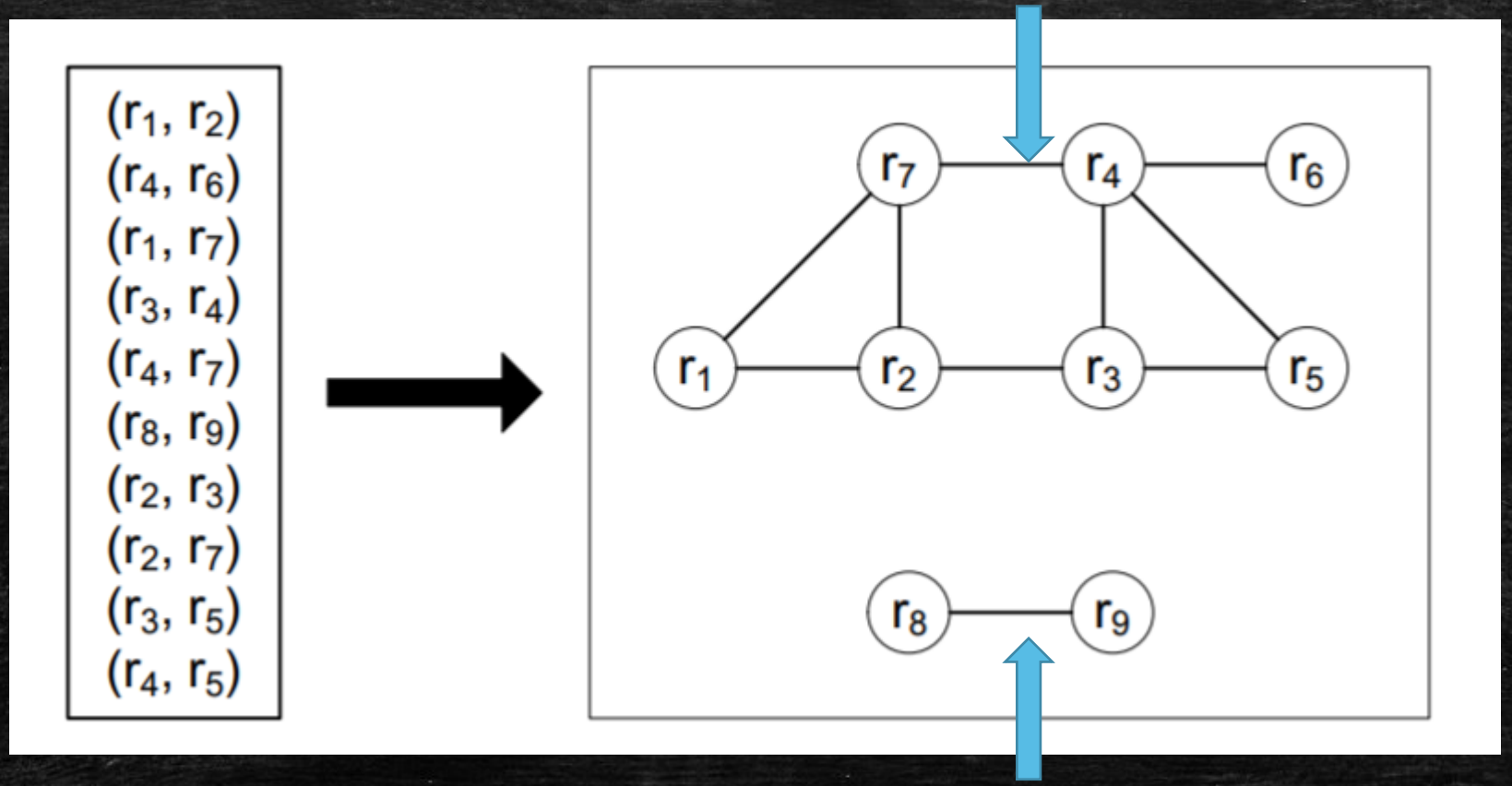
Cluster-Based HIT Generation

- A set of pairs of records P, and a cluster-size threshold, k
- The cluster-based HIT generation problem is to generate the minimum number of cluster-based HITs $H_1, H_2, \cdots, H_h$, that satisfy
  - $|H_\ell| \leq k$ for any $\ell \in [1, h]$, where $|H_\ell|$ denotes the number of records in H_ℓ
  - for any $(r_i, r_j) \in P$, there exists H_ℓ ($\ell \in [1, h]$) s.t. $r_i \in H_\ell$ and $r_j \in H_\ell$
- Hard Problem!!



A Graph-Based Approach

Large Connected Component -- partition the LCC



Small Connected Component -- output the SCC

A Graph-Based Approach

(indegree, outdegree)