



PAKDD 2014 Tutorial

Managing the Quality of Crowdsourced Databases

Reynold Cheng Yudian Zheng
Department of Computer Science
The University of Hong Kong
{ckcheng, ydzheng2}@cs.hku.hk

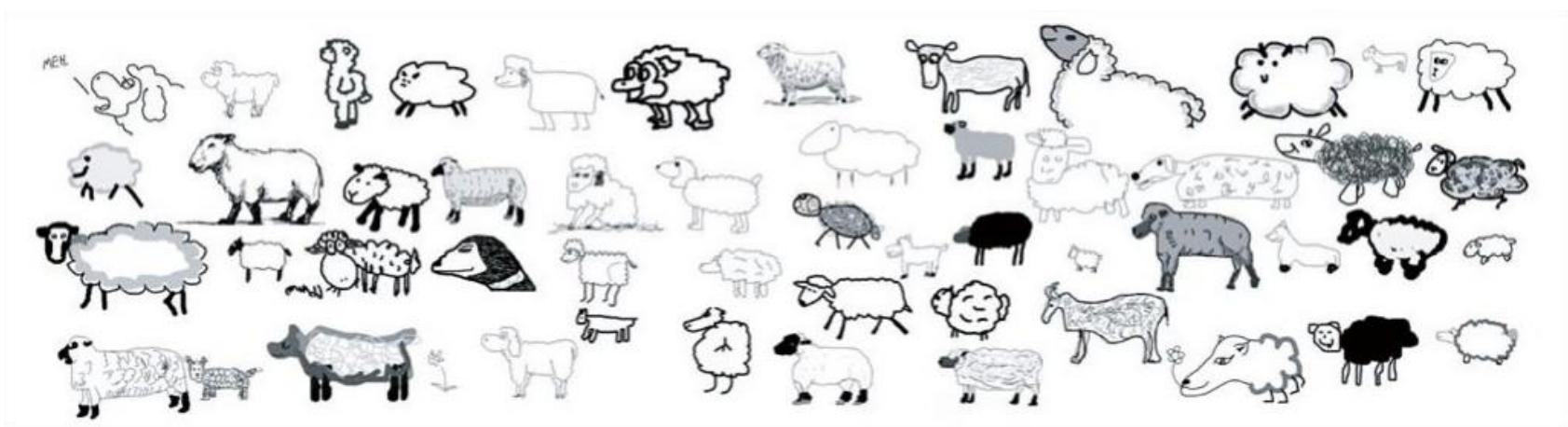
Sheep Market

2

- Draw a Sheep facing left

Pay each worker \$0.02

Collect 10,000 drawing sheep



Optical Character Recognition (OCR)

3

□ CAPTCHA

Completely Automated Public
Turing test to tell Computers
and Humans Apart



□ ReCAPTCHA

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.



Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum.
ReCAPTCHA: Human-Based Character Recognition via Web Security Measures.
Science, 321: 1465-1468, 2008

Entity Resolution (ER)

4

Are they the same?
iPad 2 = iPad Two

YES NO

SUBMIT

Find Duplicate Products In the Table. ([Show Instructions](#))

Tips: you can (1) SORT the table by clicking headers;
(2) MOVE a row by dragging and dropping it

Label	Product Name	Price ▾
1	iPad 2nd generation 16GB WiFi White	\$469
1	iPad Two 16GB WiFi White	\$490
2	Apple iPhone 4 16GB White	\$520
	iPhone 4th generation White 16GB	\$545

Reasons for Your Answers (Optional)

1
2
3
4

Submit (1 left)

J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. PVLDB, 5(11):1483-1494, 2012.

J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng. Leveraging transitive relations for crowdsourced joins. In SIGMOD Conference, pages 229-240, 2013.

Natural Language Processing (NLP)

5

□ Translation

Translate 3 lines from English to Russian (human translation needed).

Requester: Sergey Vasilyev

Reward: \$0.05 per HIT

HITs Available: 1

Duration: 15 minutes

Qualifications Required: HIT approval rate (%) is not less than 75

Translate a text between the markers below from English to Russian.

Human translation only! Machine translations will be rejected.

===== FROM HERE =====

Hello!

I am test text message to be translated from English to Russian.
If you ask me, I was born in a mind of a crazy web developer,
who tests the MTurk API to start a very promising service later.

===== TILL HERE =====

Any notes? Advices? Emotions? (Optional)

[1] C. Callison-Burch. “Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon’s Mechanical Turk”, EMNLP 2009.

[2] B. Bederson et al. Translation by Iter active Collaboration between Monolingual Users, GI 2010

Computer Vision (CV)

6

□ Painting Similarity



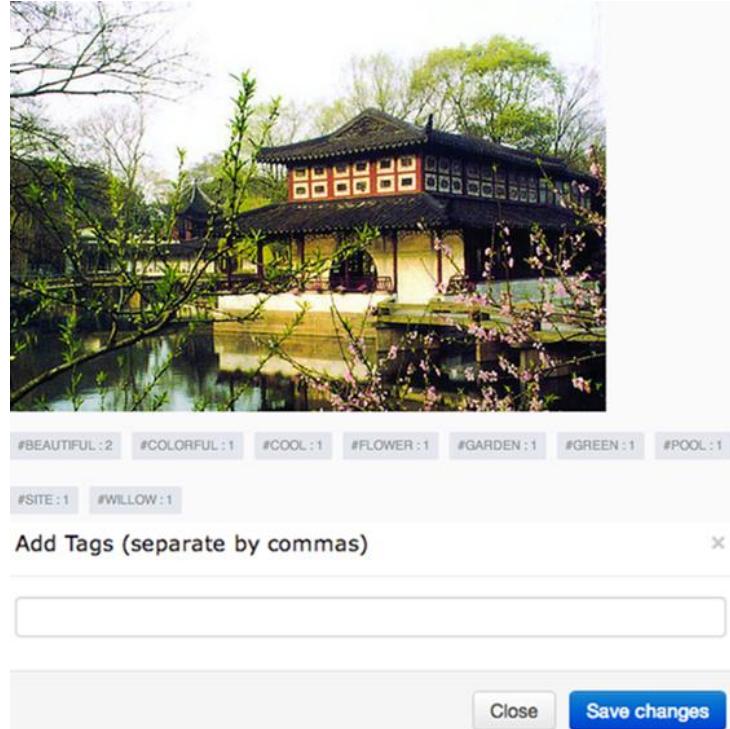
How similar is the artistic style in the paintings above?

- Very similar
- Somewhat similar
- Neither similar nor dissimilar
- Somewhat dissimilar
- Very dissimilar

A gradient based weighted averaging method for estimation of fingerprint orientation fields. Yi Wang et al. DICTA'05.

Collaborative Tagging

7



User interface for providing “tag” keywords

X. S. Yang, D. W. Cheung, L. Mo, R. Cheng, and B. Kao. On incentive-based tagging. In Proc. of ICDE, pages 685-696. 2013
Siyu Lei, Xuan S. Yang, Luyi Mo, Silviu Maniu, Reynold Cheng iTag: Incentive-Based Tagging. ICDE 2014 demo.

Why crowdsourcing?

8

Which picture visualizes better
"Golden Gate Bridge"



Submit

Please fill out the missing
department data

University	UC Berkeley
Name	EECS
URL	
Phone	(510) 642-3214

Submit

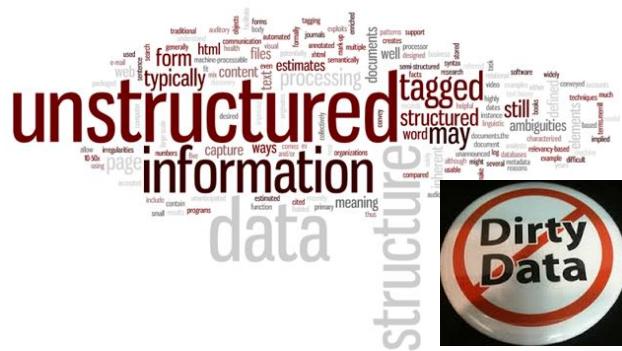


M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. CrowdDB: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Why crowdsourcing?

9

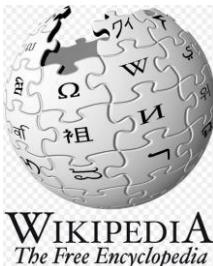
□ Every Minute



Crowdsourcing Platforms

10

- Voluntary



- Incentive-based



Crowdsourcing Model

11

A **requester** asks a **crowd** to do Human Intelligence Tasks (**HITs**) to solve **problems**.

problem:
entity resolution (ER) HIT:
comparison questions

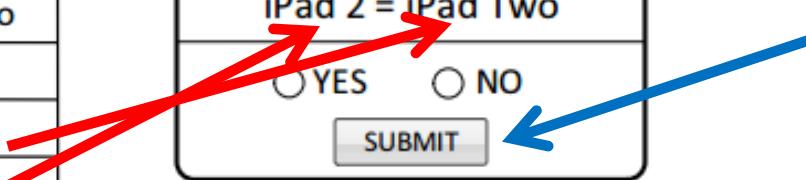
ID	Object
O_1	iPhone 2nd Gen
O_2	iPhone Two
O_3	iPhone 2
O_4	iPad Two
O_5	iPad 2
O_6	iPad 3rd Gen

Are they the same?

iPad 2 = iPad Two

YES NO

SUBMIT



crowd:
Internet users



Amazon Mechanical Turk (AMT)

12

□ Requesters

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. [Register Now](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account



Load your tasks



Get results

[Get Started](#)

□ HITs

Are they the same?

iPad 2 = iPad Two

YES NO

SUBMIT



#BEAUTIFUL:2 #COLORFUL:1 #COOL:1 #FLOWER:1 #GARDEN:1 #GREEN:1 #POOL:1

#SITE:1 #WILLOW:1

Add Tags (separate by commas)

Close

Save changes

□ Workers

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Find an interesting task



Work



Earn money



[Find HITs Now](#)

Human Intelligence Tasks (HITs)

13

□ HIT group on AMT

Quality estimation from Arabic to English

Requester: [Chris Callison-Burch](#)

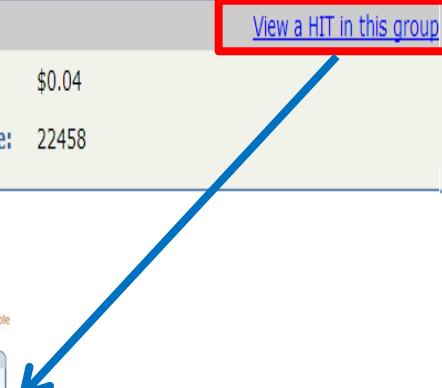
HIT Expiration Date: Jun 30, 2016 (123 weeks 4 days)

Reward: \$0.04

Time Allotted: 60 minutes

HITs Available: 22458

[View a HIT in this group](#)



□ A HIT

Timer: 00:00:00 of 2 hours

Want to work on this HIT? Want to see other HITs?

Accept HIT Skip HIT

Total Earned: Unavailable Total HITs Submitted: 0

Extract purchased items from a shopping receipt

Requester: Jon Breig Qualifications Required: None

Reward: \$0.06 per HIT HITs Available: 18941 Duration: 2 hours

NOTE: We have increased the long receipt bonus! Now 1 cent every 6 lines.

iPhone 2
iPad Two
Are they equal or not?

equal
 non-equal

iPhone 2
iPhone Two
Are they equal or not?

equal
 non-equal

Submit

AMT Statistics

14

- **New York Times (March, 2007)**

*Today, there are more than **100,000** “Turk Workers” in **more than 100 countries** who earn micropayments in exchange for completing a wide range of quick tasks called HITs, for human intelligence tasks, for various companies.*

- **Official Amazon Mechanical Blog (August, 2012)**

*more than **500,000 workers** in **190 countries***

<http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html>

<http://mechanicalturk.typepad.com/blog/2012/08/mechanical-turk-featured-on-aws-report.html>

AMT Statistics (HITs)

15

□ Types of tasks

From *January 2009 till April 2010*, we collected 165,368 HIT groups, with a total of 6,701,406 HITs, from 9,436 requesters. The total value of the posted HITs was \$529,259.

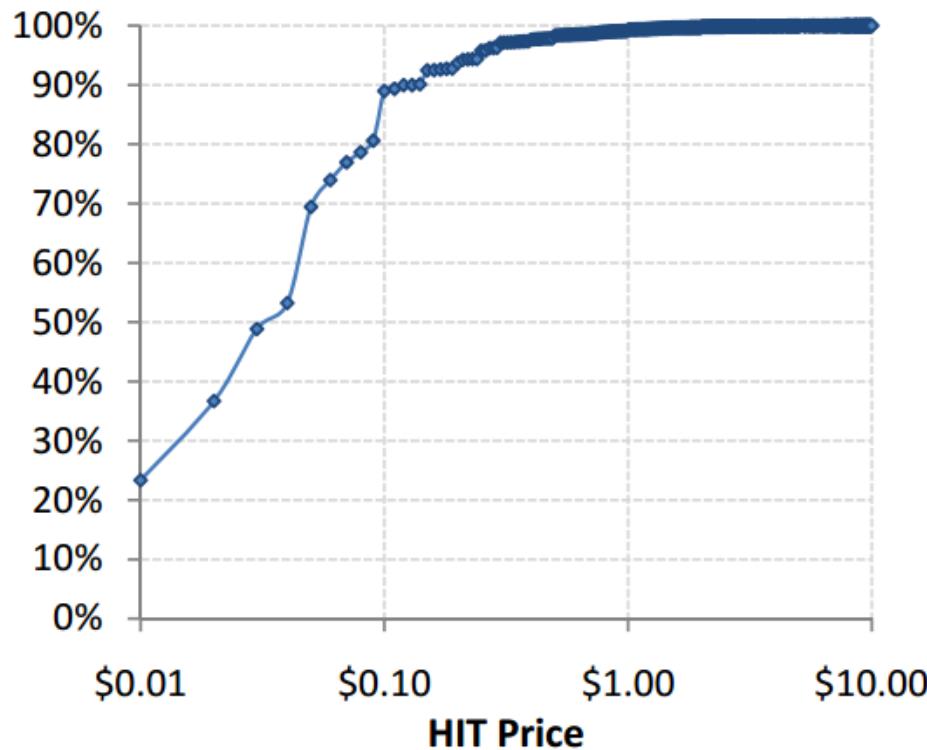
Requester ID	Requester Name	#HIT groups	Total HITs	Rewards	Type of tasks
A3MI6MIUNWCR7F	CastingWords	48,934	73,621	\$59,099	Transcription
A2IR7ETVOIULZU	Dolores Labs	1,676	320,543	\$26,919	Mediator for other requesters
A2XL3J4NH6JI12	ContentGalore	1,150	23,728	\$19,375	Content generation
A11970GL0WOQ3G	Smartsheet.com Clients	1,407	181,620	\$17,086	Mediator for other requesters
AGW2H4I480ZX1	Paul Pullen	6,842	161,535	\$11,186	Content rewriting
A1CTI3ZAWTR5AZ	Classify This	228	484,369	\$9,685	Object classification
A1AQ7EJ5P7ME65	Dave	2,249	7,059	\$6,448	Transcription
AD7C0BZNKYGYV	QuestionSwami	798	10,980	\$2,867	Content generation and evaluation
AD14NALRDSN9	retaildata	113	158,206	\$2,118	Object classification
A2RFHBFTZH7UN	ContentSpooling.net	555	622	\$987	Content generation and evaluation
A1DEBE1WPE6JFO	Joel Harvey	707	707	\$899	Transcription
A29XDCTJMAE5RU	Raphael Mudge	748	2,358	\$548	Website feedback

AMT Statistics (HIT price)

16

□ HIT price

% of HITs vs HIT price



70 % HITs cost less than \$0.05

Percentage	Price
25%	[\$0 , \$0.01]
45%	[\$0.01 , \$0.05]
20%	[\$0.05 , \$0.10]
10%	[\$0.10 , \$10.00]

AMT APIs

17

□ Application Programming Interfaces

Official: C# Java Perl Ruby
Open-source: Python (boto)

GitHub



□ Java example (create a HIT)

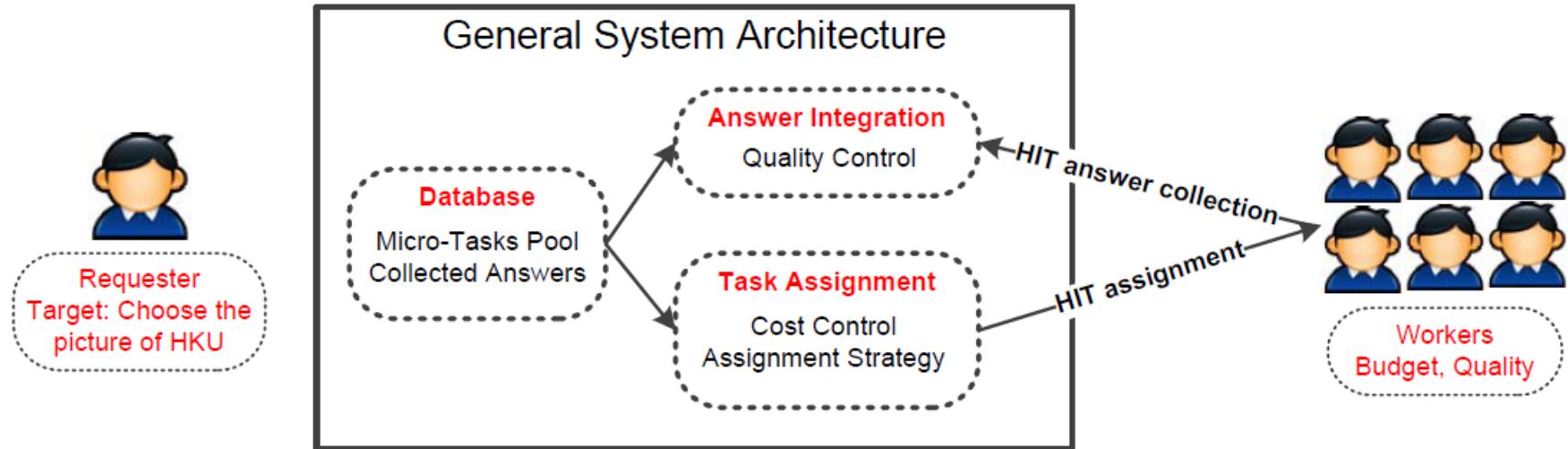
```
HIT hit = service.createHIT
(
    title,
    description,
    reward,
    RequesterService.getBasicFreeTextQuestion(
        "How many movies have you seen this month?"),
    numAssignments);
}
```

<https://github.com/boto/boto>

<http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMTurkAPI/Welcome.html>

Crowdsourcing Framework

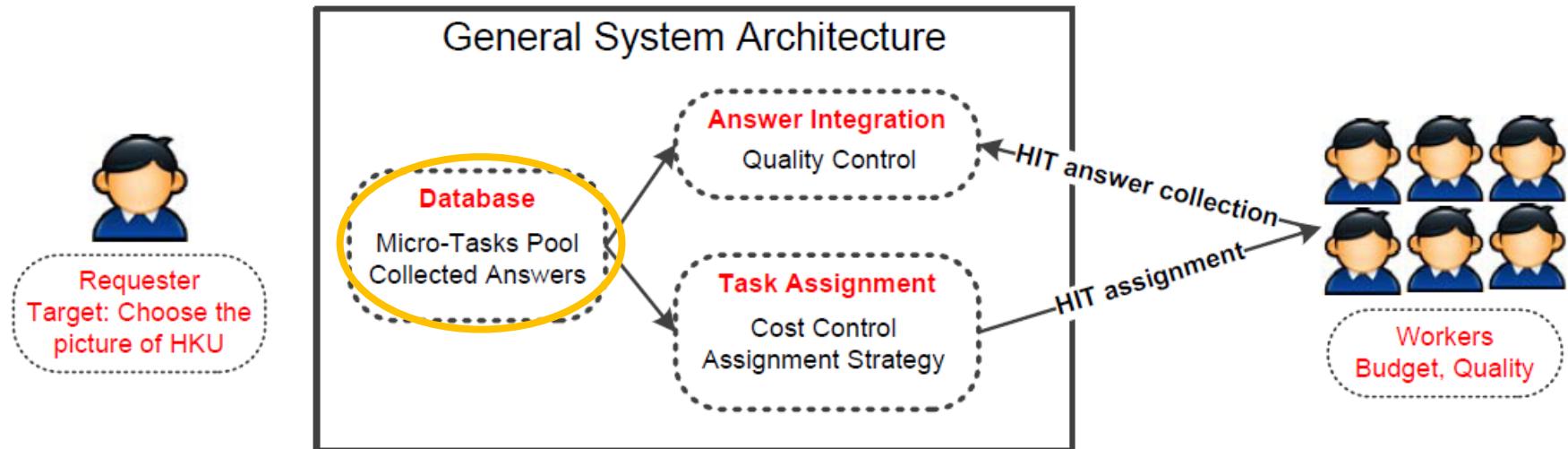
18



- **1. Answer Integration:**
How to integrate answers from workers ?
- **2. Task Assignment:**
Which tasks are chosen to assign to a worker ?
- **3. Database:**
How to store crowdsourced data?

Crowdsourcing Framework

19



- **1. Answer Integration:**
How to integrate answers from workers ?
- **2. Task Assignment:**
Which tasks are chosen to assign to a worker ?
- **3. Database:**
How to store crowdsourced data?

Crowdsourced Databases

20

Key difference from traditional DB:

The traditional **closed-world assumption** for query processing does not hold for human input



M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Query language for traditional DB

21

□ Simple selection query

```
SELECT market_capitalization FROM company  
WHERE name = "I.B.M.";
```

return NULL if “I.B.M.”
does not exist

□ Simple comparison query

```
SELECT image FROM picture  
WHERE topic = "Business Success"  
ORDER BY relevance LIMIT 1;
```

cannot decide the relevance

M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

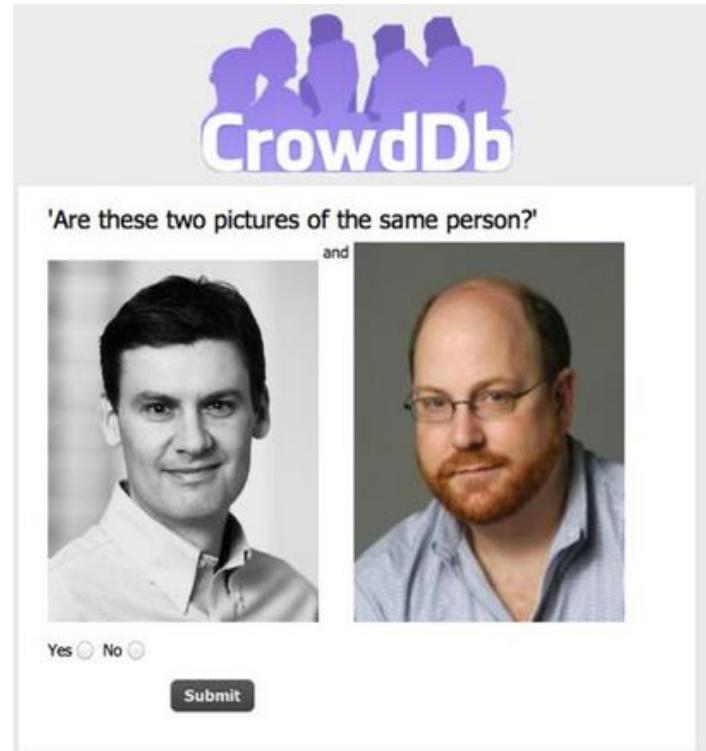
CrowdDB (SIGMOD'11)

22

□ Mobile APP



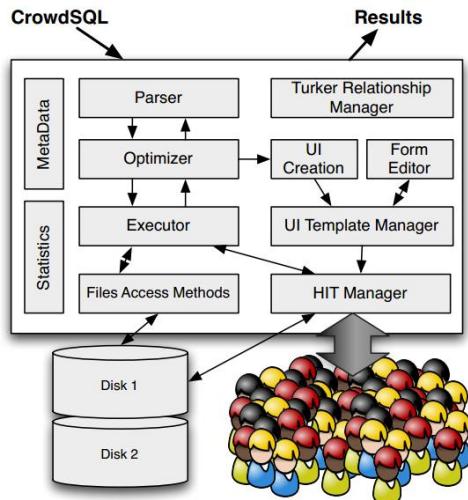
□ AMT Platform



M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.
<https://speakerdeck.com/rxin/crowddb-vldb-demo-award-talk>

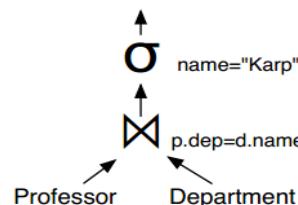
CrowdDB Architecture

23



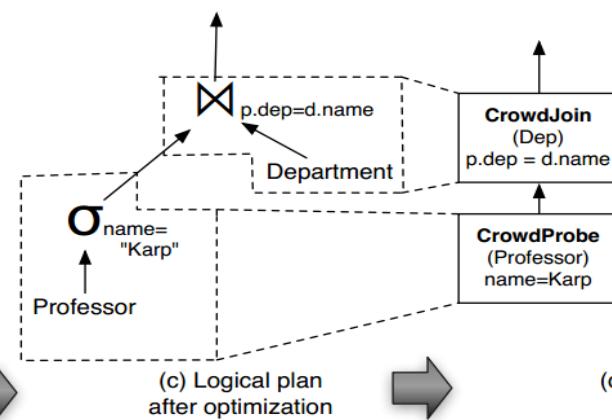
```
CREATE TABLE Department (
    university STRING,
    name STRING,
    url CROWD STRING,
    phone STRING,
    PRIMARY KEY (university, name) );
```

```
SELECT *
FROM professor p,
department d
WHERE p.department = d.name
AND p.university = d.university
AND p.name = "Karp"
```

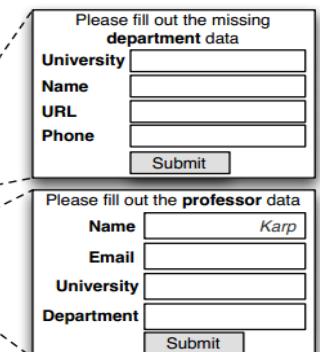


(a) PeopleSQL query

(b) Logical plan
before optimization



(c) Logical plan
after optimization

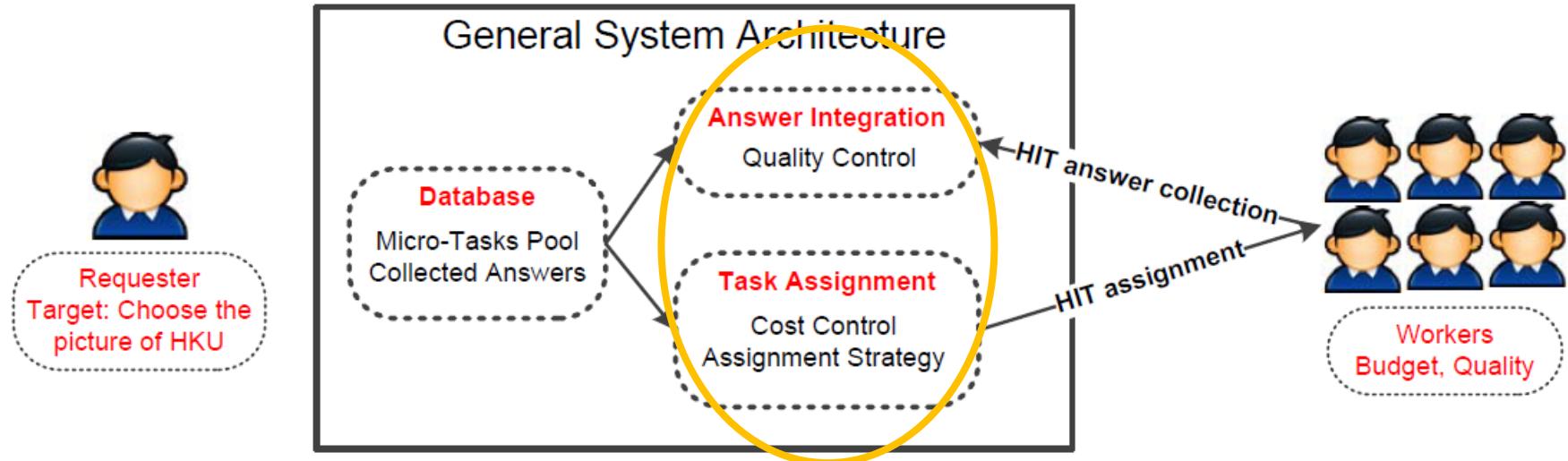


(d) Physical plan

M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. CrowdDB: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Crowdsourcing Framework

24



- **1. Answer Integration:**
How to integrate answers from workers ?
- **2. Task Assignment:**
Which tasks are chosen to assign to a worker ?
- **3. Database:**
How to store crowdsourced data?

Classification of HITs

25

- **Format of Questions**

How are the questions presented to workers?

- **Nature of Answers**

Does the question have a true answer?

Format of Questions (1)

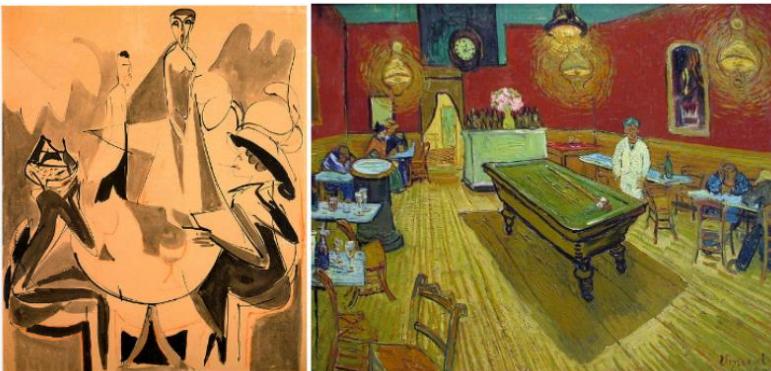
26

□ Binary Choice Question (BCQ)

Are they the same?	
iPad 2 = iPad Two	
<input type="radio"/> YES	<input type="radio"/> NO
SUBMIT	

ER (Entity Resolution):
Are two entities equal?

□ Multiple Choice Question (MCQ)



How similar is the artistic style in the paintings above?

- Very similar
- Somewhat similar
- Neither similar nor dissimilar
- Somewhat dissimilar
- Very dissimilar

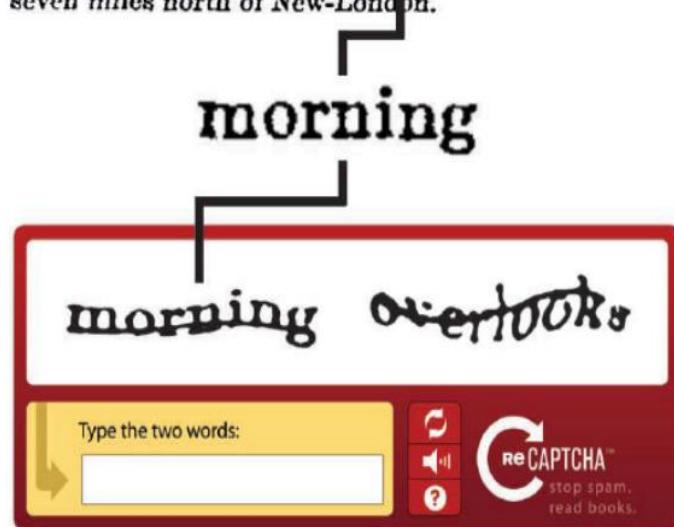
CV (Computer Vision):
similarity between two
paintings

Format of Questions (2)

27

□ Open questions with text answers

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.



OCR:
reCAPTCHA

Translate 3 lines from English to Russian (human translation needed).
Requester: Sergey Vasilyev Reward: \$0.05 per HIT HITs Available: 1 Duration: 15 minutes
Qualifications Required: HIT approval rate (%) is not less than 75

Translate a text between the markers below from English to Russian.
Human translation only! Machine translations will be rejected.

===== FROM HERE =====

Hello!
I am test text message to be translated from English to Russian.
If you ask me, I was born in a mind of a crazy web developer,
who tests the MTurk API to start a very promising service later.

===== TILL HERE =====

Any notes? Advices? Emotions? (Optional)

NLP:
language translation

Nature of Answers (1)

28

□ Ground truth to a question

Single ground truth answer:

Are they the same?
iPad 2 = iPad Two

YES NO

SUBMIT

Steve Jobs is great.
Choose the sentiment of the sentence.

positive
 neutral
 negative

Multiple ground truth answers:

Select the founder of Google Company

Larry Page
 Steve Jobs
 Sergey Brin

Decomposition →

Is Larry Page the founder of Google?

yes
 no

Is Steve Jobs the founder of Google?

yes
 no

Is Sergey Brin the founder of Google?

yes
 no

Nature of Answers (2)

29

- No ground truth to a question



#BEAUTIFUL:2 #COLORFUL:1 #COOL:1 #FLOWER:1 #GARDEN:1 #GREEN:1 #POOL:1

#SITE:1 #WILLOW:1

Add Tags (separate by commas)

A Classification of HITs

30

- TWO dimensions

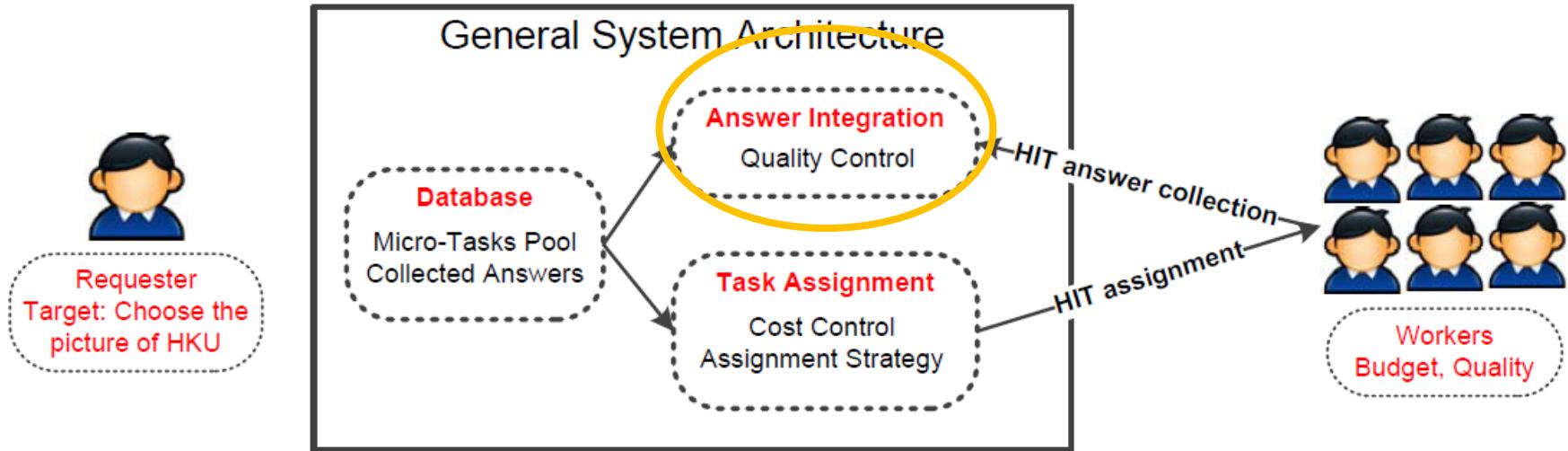
Questions: BCQ, MCQ / Open

Answers: With / Without Ground Truth

Answer Question	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

Crowdsourcing System Framework

31



- **1. Answer Integration:**

- How to integrate answers from workers ?

- **2. Task Assignment:**

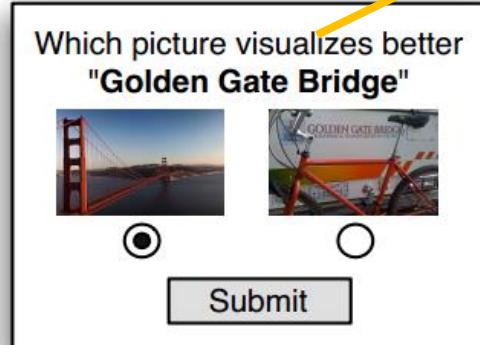
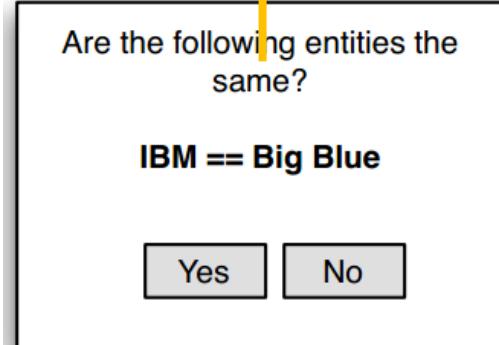
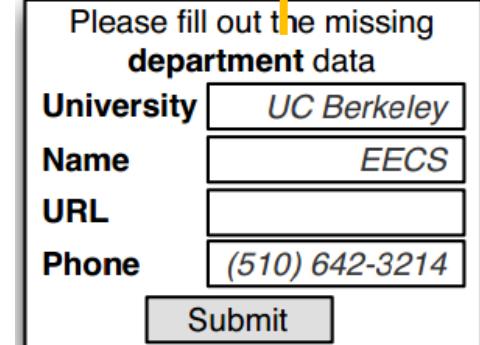
- Which tasks are chosen to assign to a worker ?

- **3. Database:**

- How to store crowdsourced data?

Answers with Ground Truth

32

Answer question	BCQ	MCQ	Open
With Ground Truth			
Without Ground Truth			

M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Voting Strategies

33

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth			
Without Ground Truth			

- Half Voting**
- Majority Voting**
- Bayesian Voting**

Example

34

Steve Jobs is great.
Choose the sentiment of the sentence.

positive
 neutral
 negative

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

- Half Voting Output ‘NULL’
- Majority Voting
Output ‘neutral’

Example

35

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

quality		positive	neutral	negative
0.7	worker 1	0.7	0.15	0.15
0.5	worker 2	0.25	0.5	0.25
0.6	worker 3	0.2	0.6	0.2
0.4	worker 4	0.3	0.3	0.4
0.3	worker 5	0.35	0.35	0.3
0.8	worker 6	0.8	0.1	0.1
0.6	worker 7	0.2	0.6	0.2

□ Bayesian Voting

positive: $0.7 * 0.25 * 0.2 * 0.3 * 0.35 * 0.8 * 0.2$
neutral: $0.15 * 0.5 * 0.6 * 0.3 * 0.35 * 0.1 * 0.6$
negative: $0.15 * 0.25 * 0.2 * 0.4 * 0.3 * 0.1 * 0.2$

normalized distribution:

(positive, neutral, negative) = (66%, 32%, 2%)

Observation

36

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

Half Voting
Output ‘NULL’
Majority Voting
Output ‘neutral’

Bayesian Voting

Output a distribution (66%, 32%, 2%)

- 
- ★ Bayesian Voting outputs a distribution by considering workers' qualities

Quality is important !

37

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

- How to represent worker's quality ?
- How to derive worker's quality ?

How to represent worker's quality ?

38

- A simple parameter q in $[0,1]$

$q=0.7$ indicates that the person will have 70% to correctly answer a question.

- Confusion Matrix M

	positive	neutral	negative
positive	0.6	0.3	0.1
neutral	0.15	0.8	0.05
negative	0.1	0.2	0.7

	positive	neutral	negative
positive	0.7	0.15	0.15
neutral	0.15	0.7	0.15
negative	0.15	0.15	0.7

X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. Cdas: A crowdsourcing data analytics system. PVLDB, 5(10):1040-1051, 2012.

P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In SIGKDD workshop, pages 64-67, 2010.

How to learn worker's quality ?

39

□ Golden Questions

Hire some experts to give the answers for a subset of questions, called “golden questions”.

7 correctly answered questions in 10 golden ones. $q = 7/10 = 0.7$

□ Learning (Expectation Maximization)

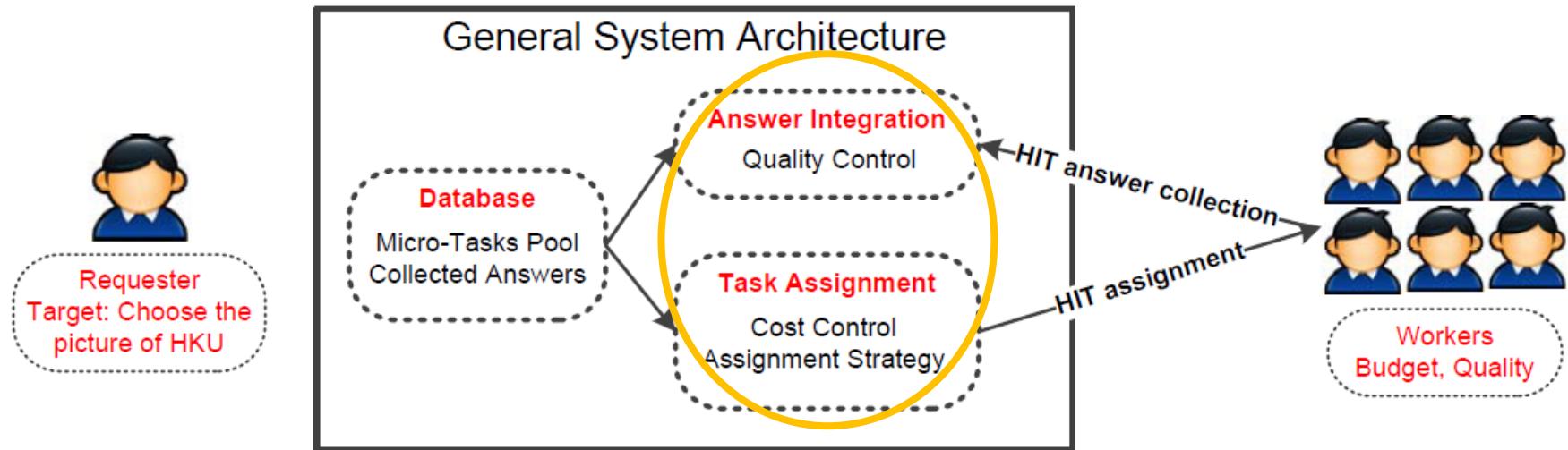
$L()$ function is not convex/concave, iteratively construct a concave upper bound function(E-step) and update the parameters(M-step).

X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. Cdas: A crowdsourcing data analytics system. PVLDB, 5(10):1040-1051, 2012.

P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In SIGKDD workshop, pages 64-67, 2010.

Crowdsourcing Framework

40



- **1. Answer Integration:**
How to integrate answers from workers ?
- **2. Task Assignment:**
Which tasks are chosen to assign to a worker ?
- **3. Database:**
How to store crowdsourced data?

Optimizing Plurality for HITs (CIKM'13)

41

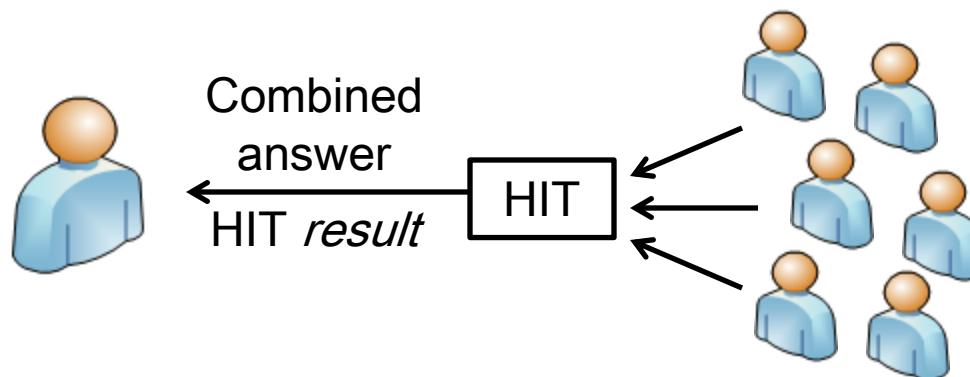
Answer Question	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- How to set **plurality** for HITs?
- How to develop effective and efficient plurality setting algorithms for HITs

Plurality of HITs

42

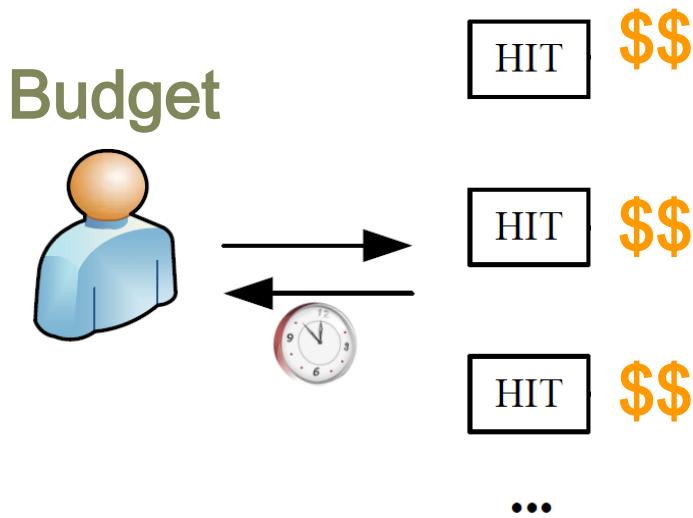
- Imperfect answers from a single worker
 - make careless mistakes
 - misinterpret the HIT requirement
- Specify *sufficient plurality* of a HIT (*number* of workers required to perform that HIT)



Plurality Assignment Problem (PAP)

43

- Plurality has to be *limited*
 - A HIT is associated with a **cost**
 - Requester has limited **budget**
 - Requester requires time to verify HIT results



PAP:
wisely assign the right
pluralities to various
HITs to achieve overall
high-quality results

Our Goal

44

- Manually assigning pluralities is tedious if not infeasible
 - AMT on 28th October, 2012
 - 90,000 HITs submitted by Top-10 requesters
- Algorithms for automating the process of plurality assignment are needed!

Multiple Choice Questions (MCQs)

45

- Most popular type
 - AMT on 28th Oct, 2012
 - About three quarters of HITs are MCQs
- Examples
 - Sentiment analysis, categorizing objects, assigning rating scores, etc.

Sentiment Review for Youtube Comment

I literally cannot watch this for longer than a few seconds without pausing it because i am laughing so hard and cant pay attention.

Positive Neutral Negative

Data Model

46

- Set of HITs $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$
- For each HIT t_i
 - contains a single MCQ
 - plurality k_i (i.e., k_i workers are needed)
 - cost c_i (i.e., c_i units of reward are given for completing t_i)

Quality Model

47

- Capture the goodness of HIT result's for t_i
- MCQ quality
 - likelihood that the result is correct after it has been performed by k workers
- Factors that affect MCQ quality
 - plurality: k
 - Worker's accuracy (or accuracy): p_i for HIT t_i
 - probability that a randomly-chosen worker provides a correct answer for t_i
 - estimated from similar HITs whose true answer is known

$$\zeta_i(k) = \sum_{l=\lceil \frac{k}{2} \rceil}^k \binom{k}{l} p_i^l (1-p_i)^{(k-l)}$$

Problem Definition

48

- **Input**

- budget B
 - Set of HITs \mathcal{T} , and $\zeta_i(k), c_i$ for $t_i \in \mathcal{T}$

- **Output**

- plurality for every HIT $\vec{k} = \{k_1, k_2, \dots, k_n\}$

- **Objective**

- maximize overall average quality

$$\mathcal{Q}(\mathcal{T}, \vec{k}) = \frac{1}{n} \sum_{i=1}^n \zeta_i(k_i) \quad \zeta_i(k) = \sum_{l=\lceil \frac{k}{2} \rceil}^k \binom{k}{l} p_i^l (1-p_i)^{(k-l)}$$

Solutions

49

given B, \mathcal{T} and p_i, c_i for $t_i \in \mathcal{T}$

maximize $\mathcal{Q}_t(\mathcal{T}, \vec{k})$

subject to $\sum_{i=1}^n k_i c_i \leq B$ **and**

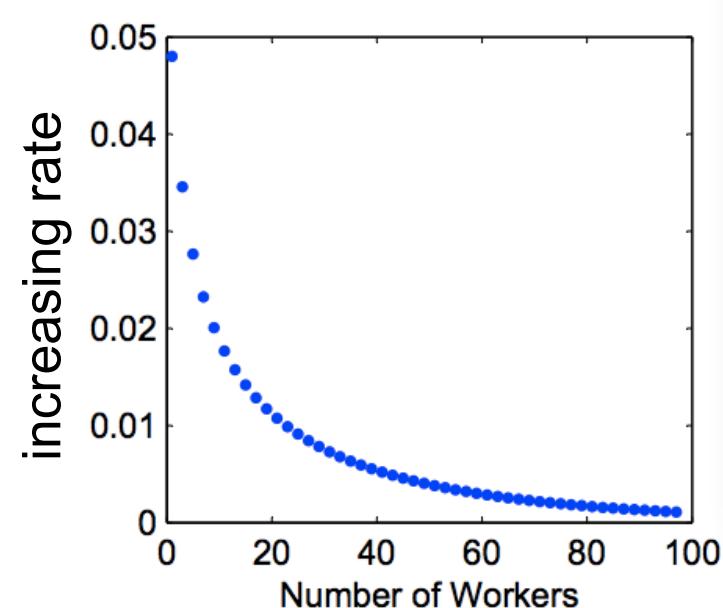
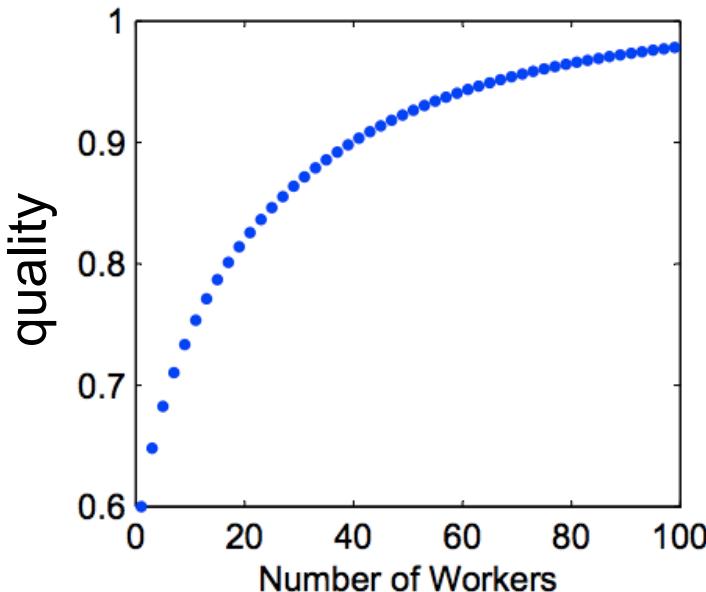
$k_i = 0$ or a positive odd number

- Optimal Solution
 - Dynamic Programming
 - Not efficient for HIT sets that contain thousands of HITs
 - 60,000 HITs extracted from AMT
 - Execution time: 10 hours

Properties

50

- Monotonicity and Diminishing Return
 - the quality function increases with plurality;
 - the rate of quality improvement drops with plurality.



Greedy: 2-approximate algorithm

51

- Properties of MCQ quality function
 - Monotonicity
 - Diminishing return
 - PAP is approximable for HITs with these two properties
- Greedy
 - Select the “best” HIT and increase its plurality until budget is exhausted
 - Selection criteria: the one with largest *marginal gain*
 - Theoretical approximation ratio = 2

Grouping techniques

52

- Observations
 - Many HITs submitted by the same requesters are given the same cost and of very similar nature
- Intuition
 - Group HITs of the same cost and quality function
 - More or less the same plurality for HITs in one group
- Main idea
 - Select a “representative HIT” from each group
 - Evaluate its plurality by **DP** or **Greedy**
 - Deduce each HIT’s plurality from the representative HIT

Experiments

53

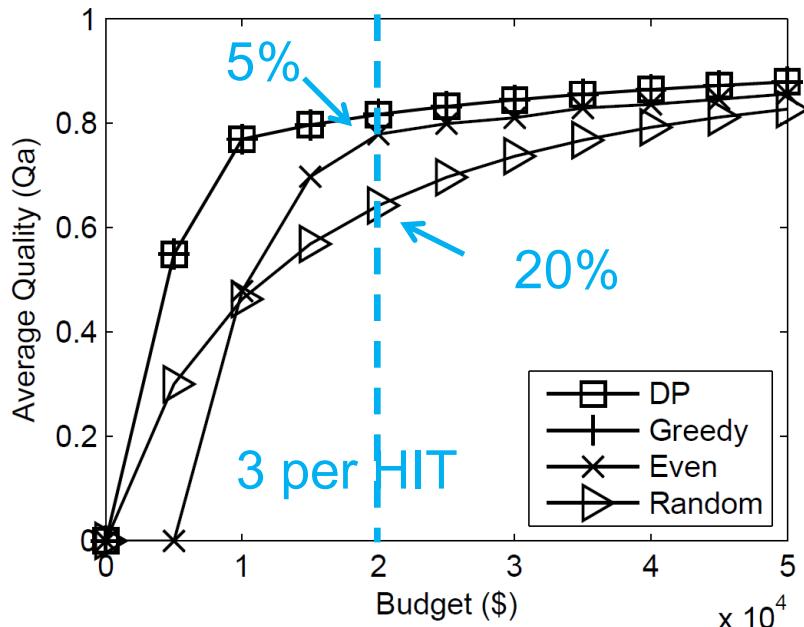
- Synthetic
 - Generated based on the extraction of an AMT requester's HITs information on Oct 28th, 2012
 - Statistics
 - 67,075 HITs
 - 12 groups (same cost and accuracy)
 - Costs vary from \$0.08 to \$0.24
 - Accuracy of each group is randomly selected from [0.5, 1]

Effectiveness

54

□ Competitors

- *Random*: arbitrarily pick a HIT to increase its plurality until budget is exhausted
- *Even*: divide the budget evenly across all HITs

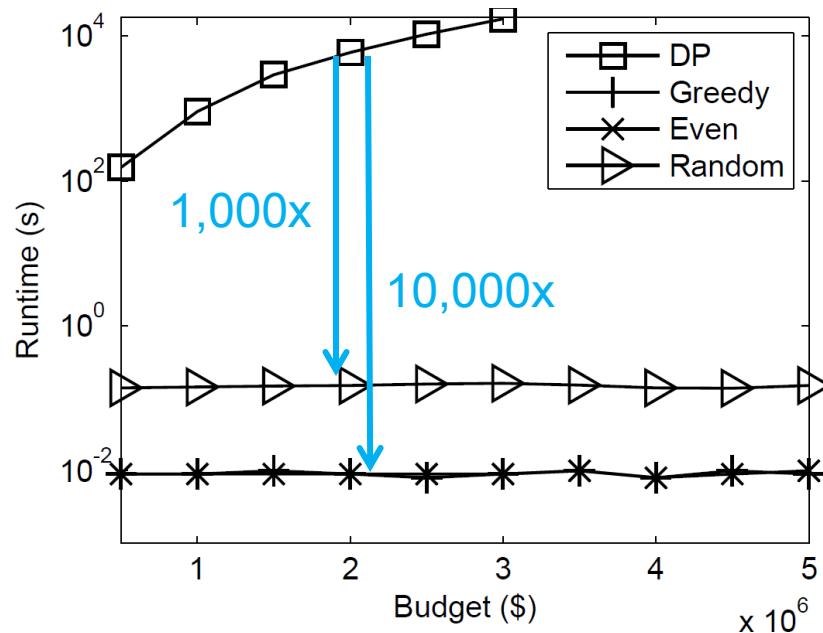


Greedy is close-to-optimal in practice

Performance (1)

55

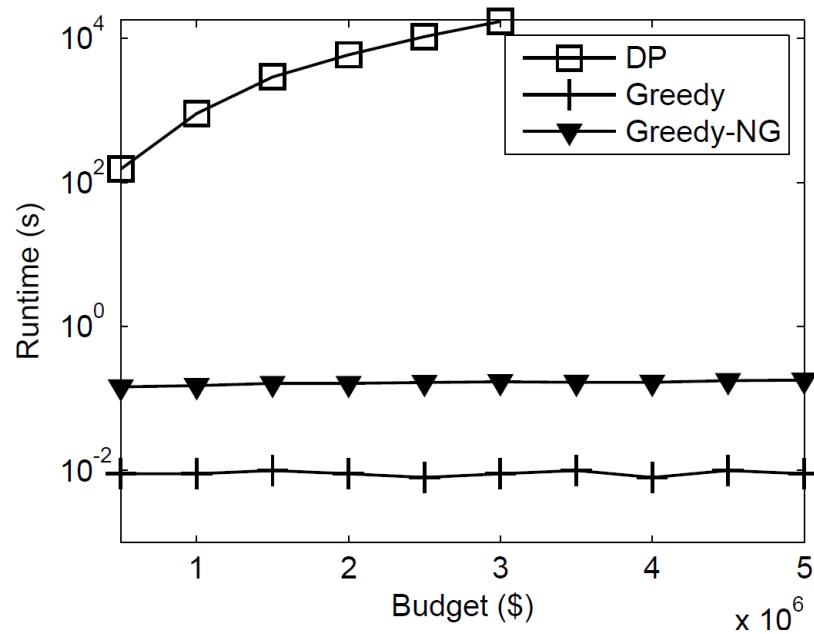
- DP and Greedy are implemented using grouping techniques
- Greedy is efficient!



Performance (2)

56

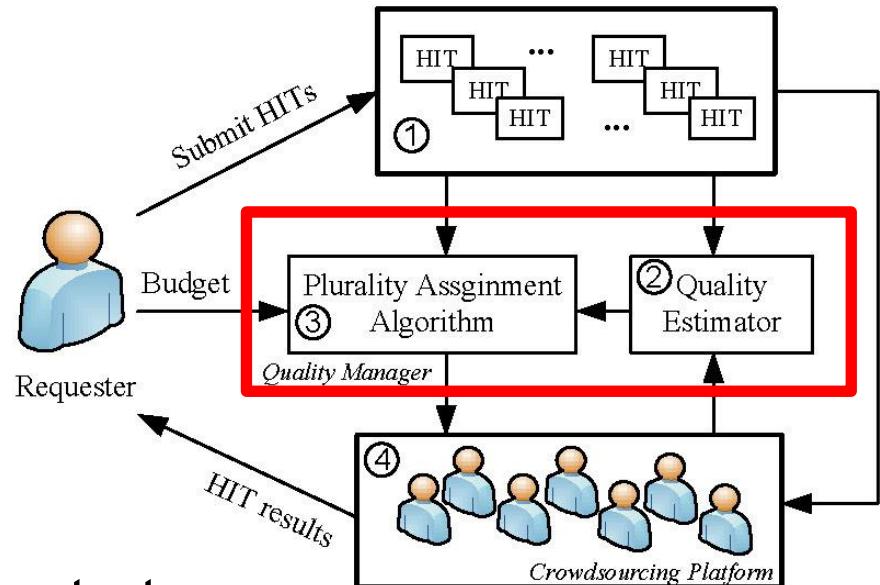
- Grouping techniques
 - 20 times faster than non-group solutions
 - 12 groups vs. 60,000 HITs



Other HIT Types

57

- Examples
 - Enumeration Query
 - Tagging Query
- Solution framework
 - Quality estimator
 - Derive $\zeta_i(k)$
 - accuracy p_i in MCQ
 - PA algorithm
 - Greedy for HITs demonstrate monotonicity and diminishing return



Enumeration Query

58

- Objective

- obtain a complete set of distinct elements for a set query

- Quality function

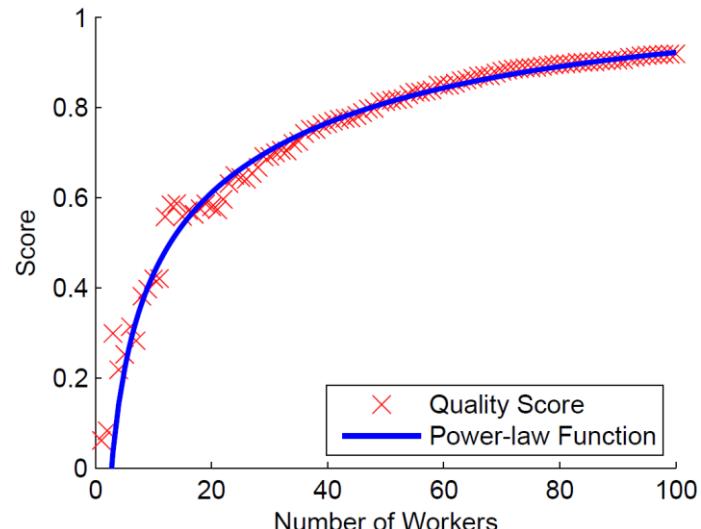
$$\zeta_i(k) = \hat{f}_0 \left[1 - \left(1 - \frac{1 - \hat{C}}{\hat{f}_0} \right)^{k-k_0} \right]$$

- Satisfy monotonicity and diminishing return
 - Greedy can be applied

Tagging Query

59

- Objective
 - obtain keywords (or tags) that best describe an object
- Empirical results illustrate **power-law** relationship with number of workers' answers



Summary

60

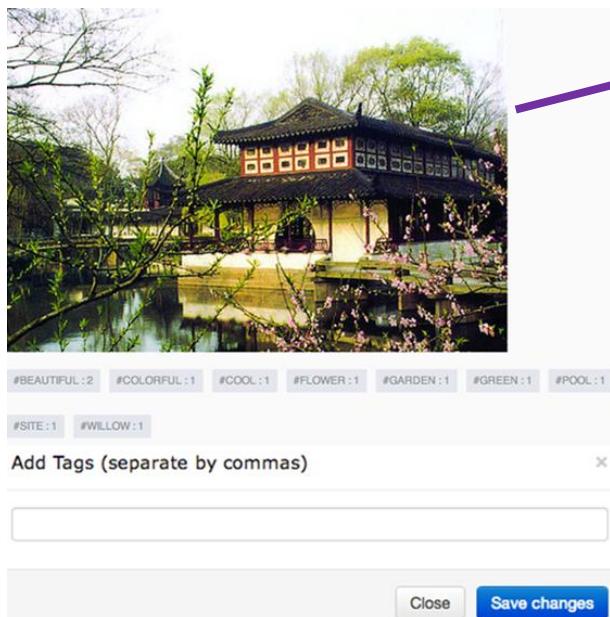
Answer/Question	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- Problem of setting plurality for HITs
- Develop effective and efficient plurality algorithms for HITs

Tagging

61

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging



Worker inputs tags for the picture
Open question without ground truth

X. S. Yang, D. W. Cheung, L. Mo, R. Cheng, and B. Kao. On incentive-based tagging. In Proc. of ICDE, pages 685-696. 2013

Siyu Lei, Xuan S. Yang, Luyi Mo, Silviu Maniu, Reynold Cheng iTAG: Incentive-Based Tagging. ICDE 2014 demo.

On incentive-based tagging (ICDE'13)

62

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- How to define Tag Data Quality
- Incentive-Based Tagging
- How to select pictures for workers to tag

Collaborative Tagging Systems

63

- Example:
 - Delicious, Flickr
- Users / Taggers
- Resources
 - Webpages
 - Photos
- Tags
 - Descriptive keywords
- Post
 - Non-empty set of tags



Applications with Tag Data

64

- Search
- Recommendation
- Clustering
- Concept Space Learning

Optimizing web search using social annotations. S. Bao et al. WWW'07

Can social bookmarking improve web search? P. Heymann et al. WSDM'08

Structured approach to query recommendation with social annotation data. J. Guo CIKM'10

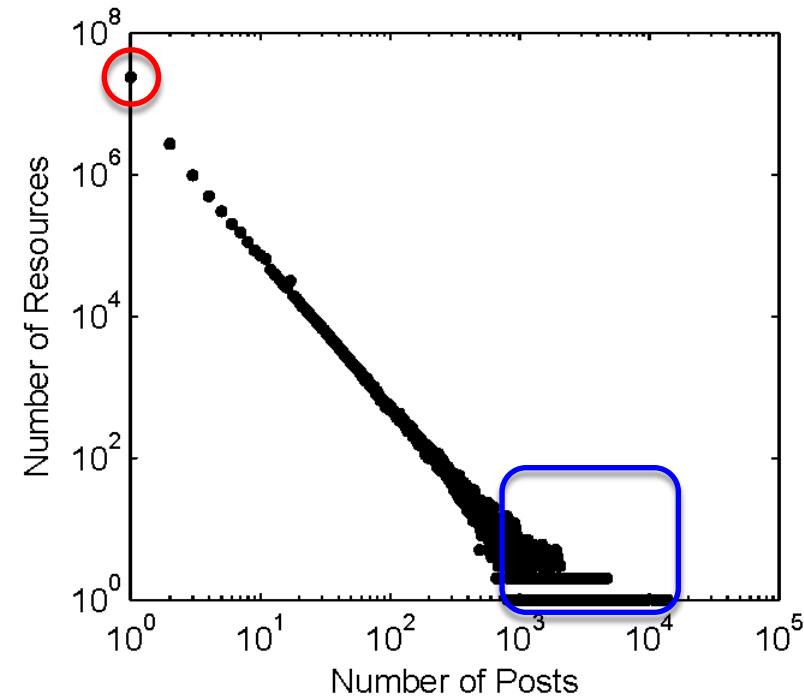
Clustering the tagged web. D. Ramage et al. WSDM'09

Exploring the value of folksonomies for creating semantic metadata. H. S. Al-Khalifa IJWSIS'07

Problem of Collaborative Tagging

65

- **Most** posts are given to **small** number of highly popular resources
- dataset from delicious
 - All ***30m*** urls
 - Over ***10m*** urls are just tagged ***once***
 - Under-Tagging
 - **39%** posts vs. ***1%*** urls
 - Over-Tagging



Under-Tagging

66

- Resources with **very few** posts have **low quality** tag data
- Low quality of one single post
 - ▣ Irrelevant to the resource
 - {3dmax}
 - ▣ Not cover all the aspects
 - {geography, education}
 - ▣ Don't know which tag is more important
 - {maps, education}



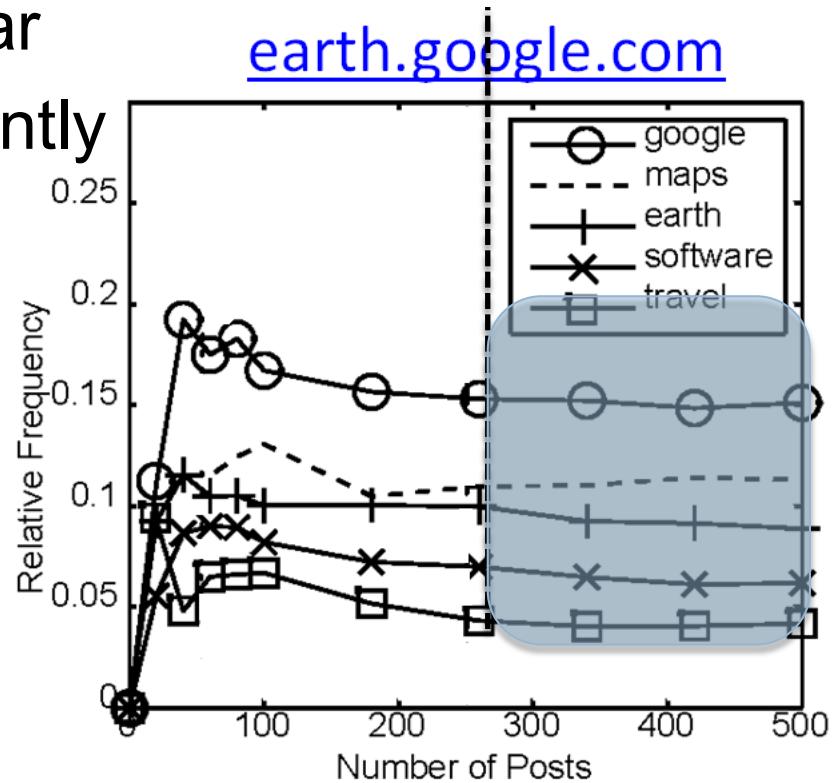
Improve tag data quality for under-tagged resource by giving it **sufficient number** of posts

Relative Frequency

67

- Relative occurrence frequency of tag t can reflect its importance
 - Irrelevant Tags rarely appear
 - Important tags occur frequently
- Relative Frequency vs. no. of posts
 - ≥ 250 , stable

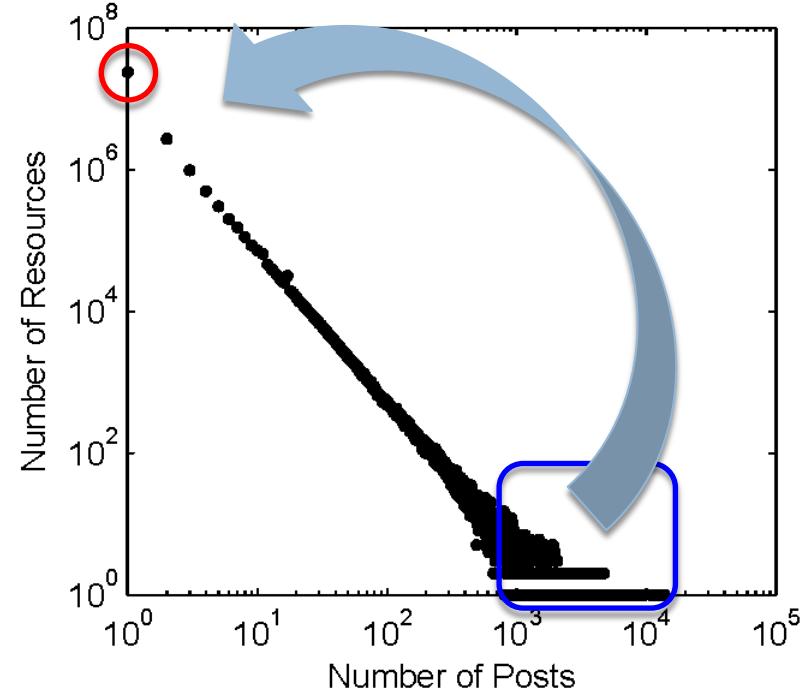
Tagging Efforts are
Wasted!



Incentive-Based Tagging

68

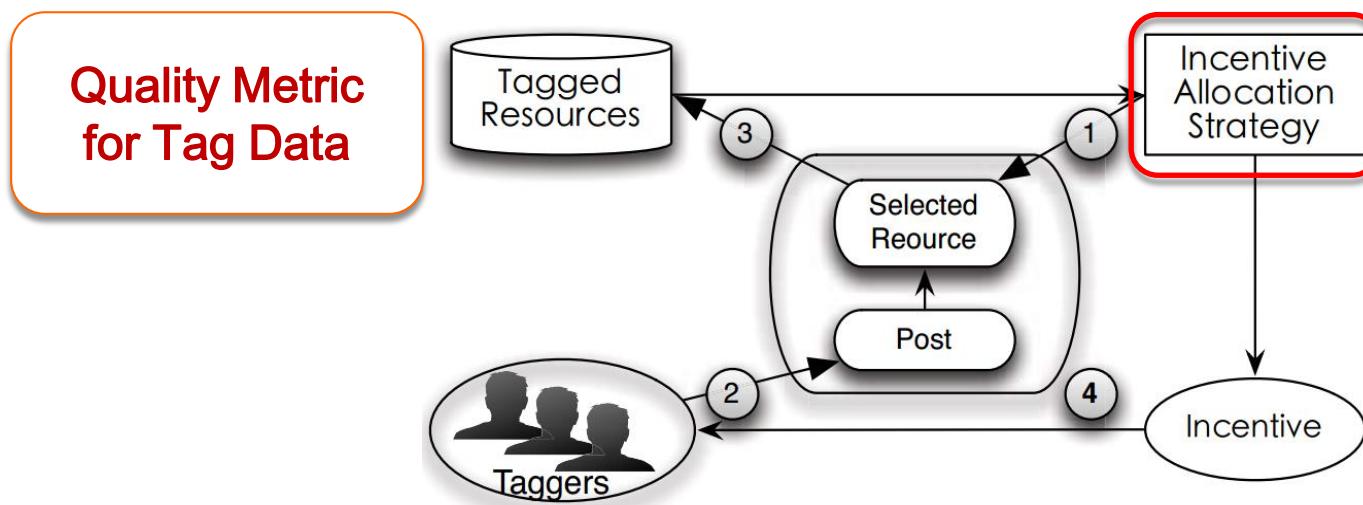
- Guide users' tagging effort
 - **Reward** users for annotating under-tagged resources
- Reduce the number of under-tagged resources
- Save the tagging efforts wasted in over-tagged resources



Incentive-Based Tagging (cont'd)

69

- Limited Budget
- Incentive Allocation
- Objective: Maximize Quality Improvement



Data Model

70

- Set of Resources \mathcal{R}
- For a specific r_i
 - Post: a set of tags
 - Post Sequence $\{p_i(k)\}$
 - Relative Frequency Distribution (rfd)
 - After r_i has k posts



Google™ earth

{maps,
education}

{geography
, education}

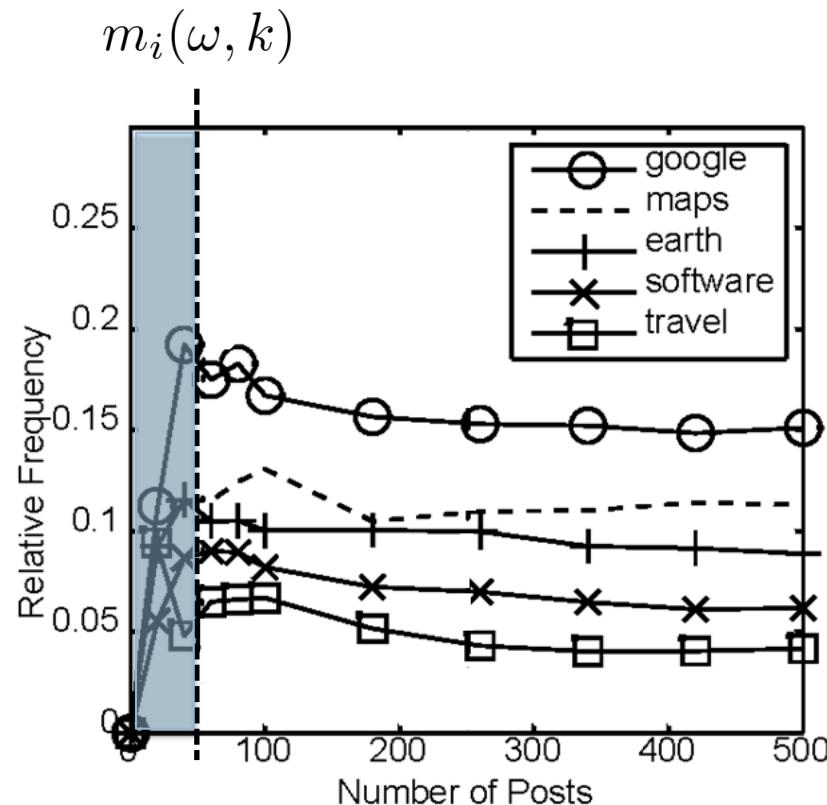
{3dmax}

Tag	Frequency	Relative Frequency
Maps	1	0.2
Geography	1	0.2
Education	2	0.4
3dmax	1	0.2

Quality Model: Tagging Stability

71

- Stability of rfd $m_i(\omega, k)$
 - Average Similarity between ω rdfs', i.e.,
 $(k-\omega+1)$ -th, ..., k -th rfd
- Stable point
 - Threshold τ
 - Stable rfd $\hat{\varphi}_i(\omega, \tau)$



Quality

72

- For one resource r_i with k posts
 - Similarity between its current rfd and its stable rfd
$$q_i(k) = s(\vec{F}_i(k), \hat{\varphi}_i)$$
- For a set of resources \mathcal{R}
 - Average quality of all the resources

$$q(\mathcal{R}, \vec{k}) = \frac{1}{n} \sum_{i=1}^n q_i(k_i)$$

Incentive-Based Tagging

73

- Input
 - A set of resources \mathcal{R}
 - Initial posts \vec{c}
 - Budget B
 - Output
 - Incentive assignment
 - how many new posts should r_i get
 - Objective
 - Maximize quality
- $q(\mathcal{R}, \vec{c} + \vec{x})$
- $\vec{c} = [3, 2, 4]$ Current Time

r_1 r_2 r_3

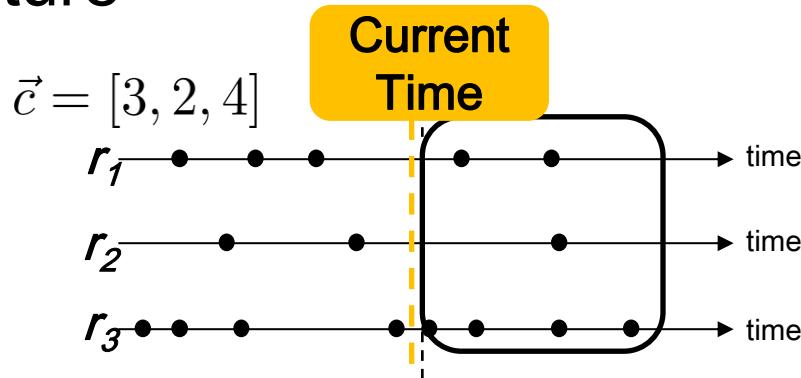
time time time

$B = 7$ $\vec{x} = [2, 1, 4]$ $q(\mathcal{R}, \vec{c} + \vec{x})$

Incentive-Based Tagging (cont'd)

74

- Optimal Solution
 - Dynamic Programming
 - Best Quality Improvement
 - Assumption: know the stable rfd & posts in the future



Strategy Framework

75

Require: Budget B , Resources \mathcal{R} , Initial no. of posts \vec{c}

- 1: **for** $i \leftarrow 1$ to n **do** $x[i] \leftarrow 0$
 - 2: INIT()
 - 3: **while** $B > 0$ **do**
 - 4: $i_0 \leftarrow \text{CHOOSE}()$
 - 5: Present r_{i_0} for a tagger to tag
 - 6: The tagger completes a post task on r_{i_0}
 - 7: UPDATE()
 - 8: $x[i_0] \leftarrow x[i_0] + 1$, $B \leftarrow B - 1$
- return** \vec{x}

Implementing *CHOOSE()*

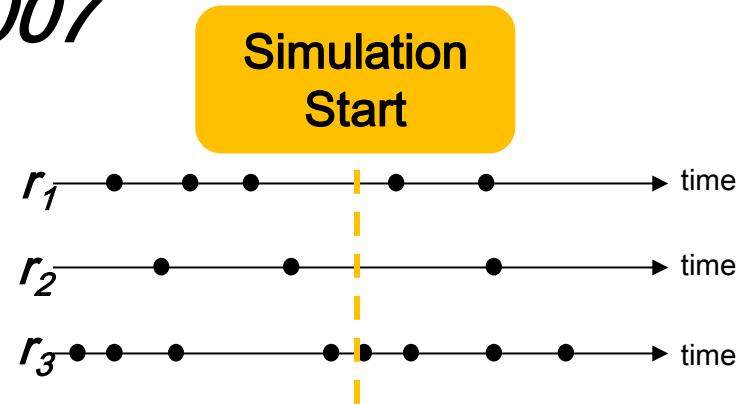
76

- **Free Choice (FC)**
 - Users freely decide which resource they want to tag.
- **Round Robin (RR)**
 - The resources have even chance to get posts.
- **Fewest Post First (FP)**
 - Prioritize Under-Tagged Resources
- **Most Unstable First (MU)**
 - Resources with unstable rdfs' need more posts
 - Window size
- **Hybrid (FP-MU)**

Setup

77

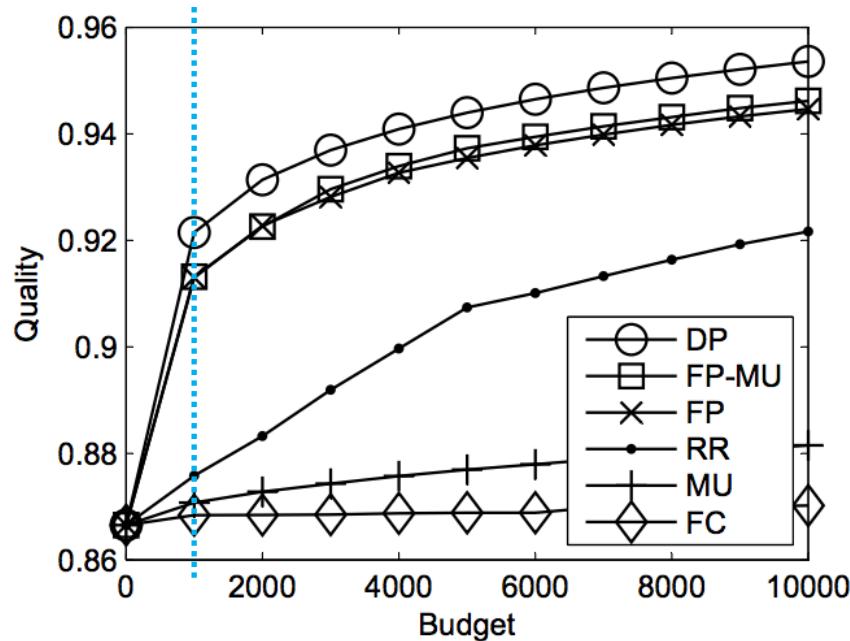
- *Delicious* dataset during year *2007*
- **5000** resources
 - Passed their stable point
 - Know the entire post sequence
- Simulation from *Feb. 1 2007*
 - **148,471** Posts in total
 - **7%** passed stable point
 - **25%** under-tagged
(# of Posts < **10**)



Quality vs. Budget

78

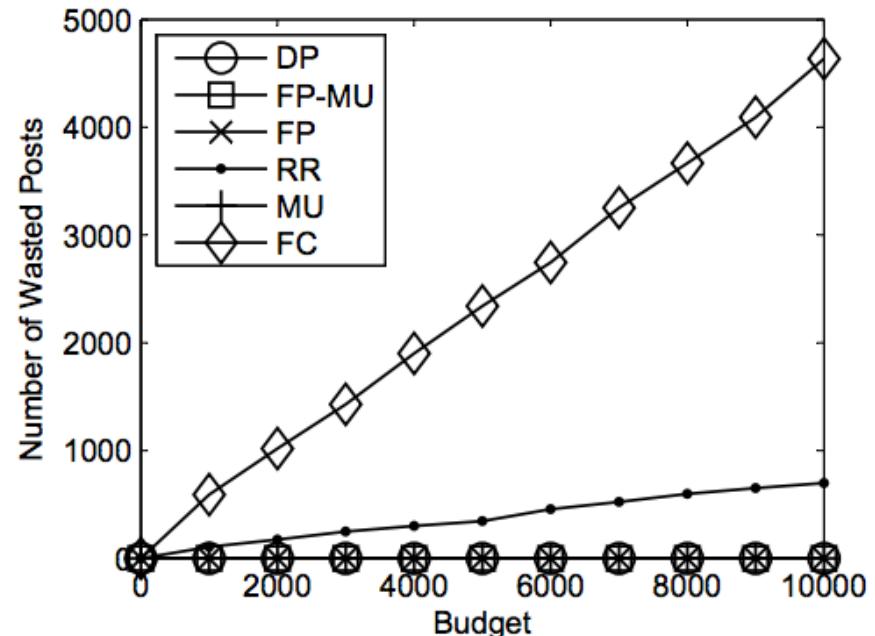
- FP & FP-MU are close to optimal
- FC does NOT increase the quality
- Budget = **1,000**
 - 0.7% more posts comparing with initial no.
 - **6.7%** quality improvement
- Make all resources reach stable point
 - FC: over **2 million** more posts
 - FP & FP-MU: **90%** saved



Over-Tagging

79

- Free Choice: **50%** posts are over-tagging, **wasted**
- FP, MU and FP-MU: **0%**



Top-10 Similar Sites (Cont'd)

80

- On *Feb. 1 2007*
 - www.myphysicslab.com
 - 3 posts
 - Top-10 all java related
- **10,000** more posts by FC
 - get 4 more posts
 - **4/10** physics related

Rank	$3 + 4(\text{by FC}_{10000}) = 7 \text{ posts}$	
1	physicsclassroom.com	✓
2	hyperphysics.phy-astr. gsu.edu/hbas/hframe.html	✓
3	practicalphysics.org	✓
4	hyperphysics.phy-astr. gsu.edu/hbas/hframe.html	✓
5	javaranch.com	
6	robocode.sourceforge.net	
7	www.onjava.com	
8	java.sun.com	
9	jguru.com	
10	kickjava.com	

Top-10 Similar Sites (Cont'd)

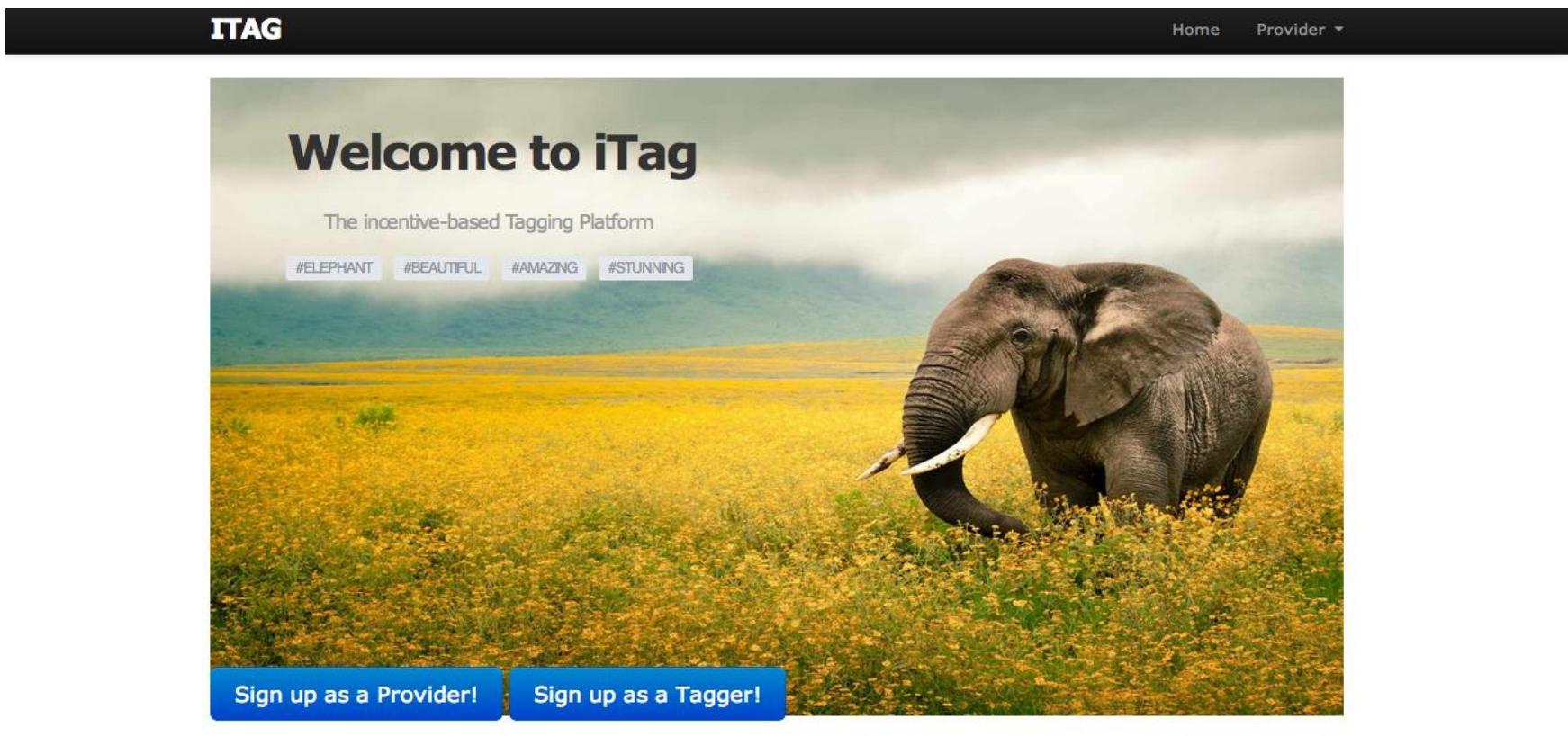
81

- On *Dec. 31 2007*
 - 270 Posts
 - Top-10 all physics related
 - ***Perfect Result***
- **10,000** more posts by FP
 - get **11** more posts
 - Top **9** physics related
 - **9** included in Perfect Result
 - Top **6** same order with Perfect Result

Rank	
1	$3 + 11 \text{ (by } FP_{10000}\text{) } = 14 \text{ posts}$ physicsclassroom.com
2	practicalphysics.org
3	hyperphysics.phy-astr. gsu.edu/hbas/hframe.html
4	hyperphysics.phy-astr. gsu.edu/hbas/hph.html
5	physicsforums.com
6	upscale.utoronto.ca/ GeneralInterest/Harrison/Flash
7	bethe.cornell.edu
8	grc.nasa.gov/WWW/K-12/Numbers/ Math/Mathematical_Thinking/index.htm
9	micro.magnet.fsu.edu/primer/ java/scienceopticsu/powersof10
10	sodaplay.com

iTag (1) [ICDE'14 demo]

82



- Welcome page for provider

iTag (2)

83

The screenshot shows the iTag web application interface. At the top, there is a navigation bar with the 'ITAG' logo on the left and 'Home' and 'Provider' buttons on the right. Below the navigation bar, there are two project cards.

Project Card 1: This card is titled "Create Project". It displays a blue square icon with a white power symbol. To its right, the text "Project Name: Tagging photos" is shown, along with "Project Details" and "Stop Project" buttons. Below this, there are three small images: a fluffy orange cat, a white piglet, and a traditional building surrounded by trees. A progress bar at the bottom indicates "0%".

Project Card 2: This card is titled "Project Name: Tag Delicious Data". It displays a green square icon with a white question mark symbol. To its right, the text "Project Name: Tag Delicious Data" is shown, along with "Project Details" and "Stop Project" buttons. Below this, there are three URLs: <http://www.cycling74.com/>, <http://www.jpl.nasa.gov/>, and <http://www.openwebdesign.org/>. A progress bar at the bottom indicates "Quality Score".

iTag

HKU CS

- Existing projects created by the Provider

iTag (3)

84

The screenshot shows the 'New project' form on the iTag platform. The form fields are as follows:

- Name: An empty input field.
- Description: An empty input field.
- Budget (\$): An empty input field.
- Pay/Task (\$): An empty input field.
- Resource Type: A dropdown menu set to 'Images'.
- Choose Strategies: A dropdown menu set to 'Hybrid Strategy (FP-MU)'.

A blue 'Create Project' button is located at the bottom of the form. The top navigation bar includes the 'ITAG' logo, 'Home', and 'Provider' dropdown.

- Create a new project

iTag (4)

85

The screenshot shows the iTag web application interface. At the top, there is a navigation bar with the 'ITAG' logo on the left and 'Home' and 'Provider' buttons on the right. Below the navigation bar, there are two project cards.

Project Card 1: This card is titled 'Create Project'. It features a blue square icon with a white power symbol. Below it, the text 'Name: Provider' and 'Email: r@g.com' are displayed. A large blue button labeled 'Create Project' is positioned above the project details. The project details section includes the text 'Project Name: Tagging photos' and two buttons: 'Project Details' (green) and 'Stop Project' (orange). Below this, there are three thumbnail images: a fluffy orange cat, a white piglet, and a traditional building.

Project Card 2: This card is titled 'Project Name: Tag Delicious Data'. It also features a blue square icon with a white power symbol. Below it, the text 'http://www.cycling74.com/' and 'http://www.jpl.nasa.gov/' are listed. A green progress bar labeled 'Quality Score' is shown at the bottom. The project details section includes the text 'Project Name: Tag Delicious Data' and two buttons: 'Project Details' (green) and 'Stop Project' (orange).

iTag

HKU CS

- A previously created project

iTag (5)

86

ITAG

Home Provider ▾

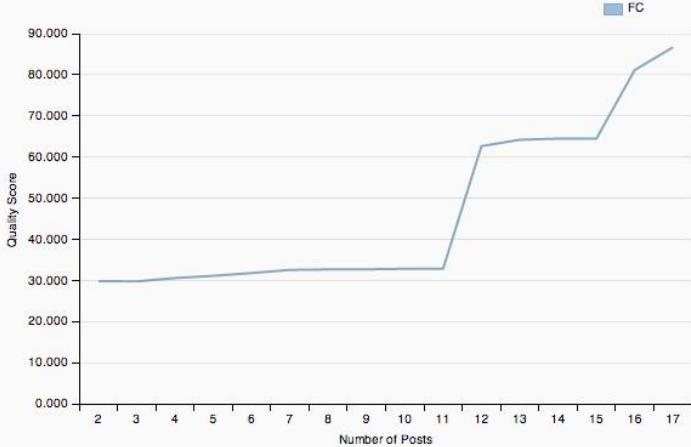
Project Name: Tagging photos
Description: Tagging photos
Budget: 10.0
Pay/Task (\$): 1.0
Quality:


Resource Type: Images
Strategy:
Free Choice (FC)
Save changes

Upload Resource (.jpg, .png)

Quality Details



Number of Posts	Quality Score (FC)
2	30,000
3	30,000
4	30,000
5	30,000
6	30,000
7	30,000
8	30,000
9	30,000
10	30,000
11	30,000
12	62,000
13	62,000
14	62,000
15	62,000
16	82,000
17	85,000

- The quality score by FC

iTag (6)

87

ITAG

Home Provider ▾

Project Name: Tagging photos
Description: Tagging photos
Budget: 10.0
Pay/Task (\$): 1.0
Quality: 

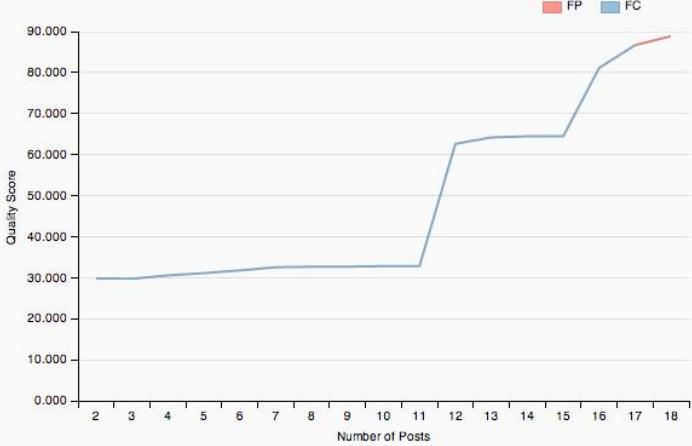
Resource Type: Images
Strategy: Fewest Post First (FP) 

Save changes

Upload Resource (.jpg, .png)



Quality Details



Number of Posts	FP Quality Score	FC Quality Score
2	30,000	30,000
3	30,500	30,500
4	31,000	31,000
5	31,500	31,500
6	32,000	32,000
7	32,500	32,500
8	33,000	33,000
9	33,500	33,500
10	34,000	34,000
11	34,500	34,500
12	35,000	65,000
13	65,000	65,000
14	65,000	65,000
15	65,000	65,000
16	65,000	80,000
17	85,000	85,000
18	85,000	85,000

- The quality score with different strategies

iTag (7)

88



#CAT : 6 #CUTE : 4 #CUTIE : 2 #DOG : 3 #FLUFFY : 3

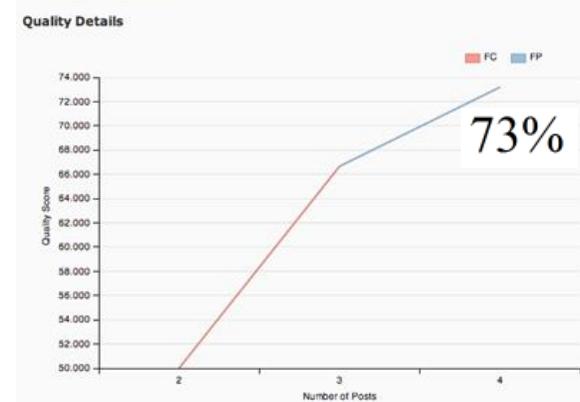
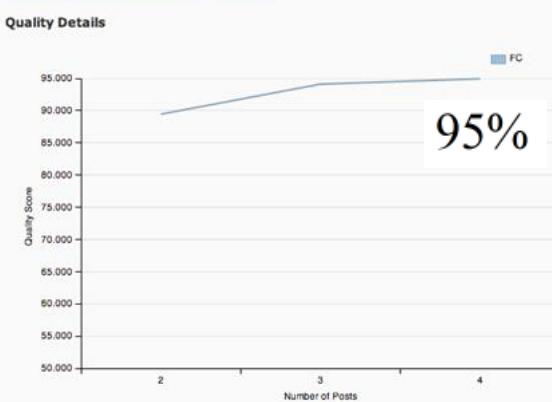
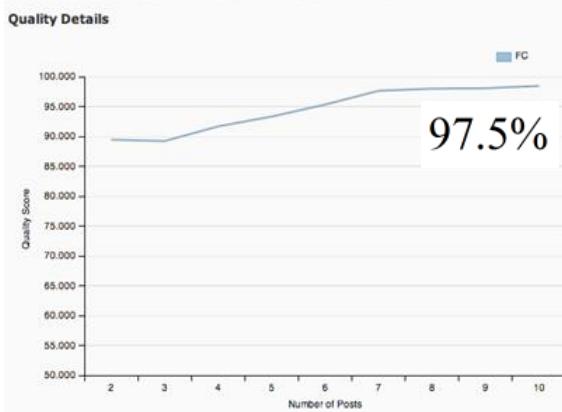


#CUTE : 2 #GREEN GRASS : 1 #PIG : 4



#BEAUTIFUL : 2 #COLORFUL : 1 #COOL : 1 #FLOWER : 1 #GARDEN : 1 #GREEN : 1 #POOL : 1

#SITE : 1 #WILLOW : 1



- The quality scores for three pictures

iTag (8)

89

ITAG

Home Provider ▾

Project Name: Tagging photos
Description: Tagging photos
Budget: 10.0
Pay/Task (\$): 1.0
Quality: 

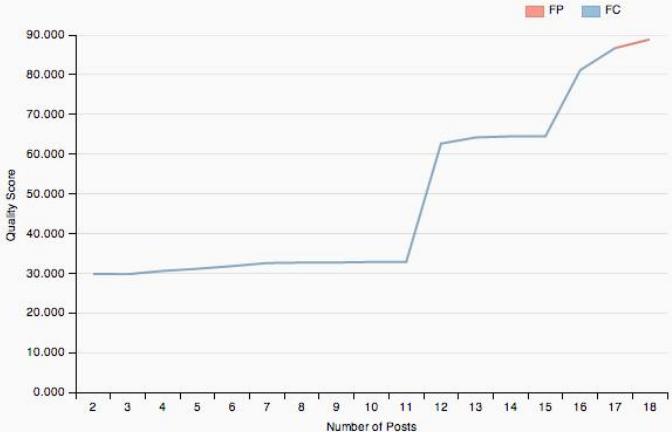
Resource Type: Images
Strategy: Most Unstable First (MU) 

Save changes

Upload Resource (.jpg, .png)



Quality Details



Number of Posts	FP (Quality Score)	FC (Quality Score)
2	30.000	30.000
3	30.000	30.000
4	30.000	30.000
5	30.000	30.000
6	30.000	30.000
7	30.000	30.000
8	30.000	30.000
9	30.000	30.000
10	30.000	30.000
11	30.000	30.000
12	62.000	62.000
13	62.000	62.000
14	62.000	62.000
15	62.000	62.000
16	82.000	82.000
17	82.000	82.000
18	88.000	88.000

- Tagging sequence for MU

iTag (9)

90

The screenshot shows the iTAG website's homepage. At the top, there is a dark header bar with the word "ITAG" in white. To the right of the header are two links: "Home" and "Tagger". Below the header is a large, scenic photograph of an elephant standing in a field of yellow flowers under a cloudy sky. Overlaid on the left side of the image is the text "Welcome to iTAG" in a large, bold, dark font, followed by the subtitle "The incentive-based Tagging Platform" in a smaller, lighter font. Below the subtitle are four small, light-colored rectangular buttons containing the hashtags "#ELEPHANT", "#BEAUTIFUL", "#AMAZING", and "#STUNNING". At the bottom of the page, there are two blue call-to-action buttons: "Sign up as a Provider!" on the left and "Sign up as a Tagger!" on the right. The bottom of the page features a thin horizontal footer bar with the "iTAG" logo on the left and the "HKU CS" logo on the right.

- Welcome page for Tagger

iTag (10)

91

The screenshot shows a web application interface for 'iTAG'. At the top, there is a blue header bar with the number '91' on the left. Below it is a black navigation bar with the word 'ITAG' in white. On the right of the navigation bar are links for 'Home' and 'Tagger' with a dropdown arrow. The main content area has a light gray background with a dotted border. It displays two project cards. The first card on the left contains a blue square icon with a white power symbol. To its right, the project details are listed: 'Project Name: Tagging photos | Provider: Provider', 'Pay/Task: 1.0 | Project Type: Photos', and a green 'View in Details' button. Below this card, user information is shown: 'Name: Tagger' and 'Email: p@p.com'. The second card follows the same structure, listing 'Project Name: Tag Delicious Data | Provider: Provider', 'Pay/Task: 1.0 | Project Type: URLs', and a green 'View in Details' button. At the bottom of the page, there is a horizontal footer bar with the 'iTAG' logo on the left and the 'HKU CS' logo on the right.

Project Name: Tagging photos | Provider: Provider
Pay/Task: 1.0 | Project Type: Photos
[View in Details](#)

Name: Tagger
Email: p@p.com

Project Name: Tag Delicious Data | Provider: Provider
Pay/Task: 1.0 | Project Type: URLs
[View in Details](#)

iTag HKU CS

- Select a project to tag

iTag (11)

92

ITAG

Home Tagger ▾

Pay/Task (\$): 1.0



#BEAUTIFUL : 2 #COLORFUL : 1 #COOL : 1 #FLOWER : 1 #GARDEN : 1 #GREEN : 1 #POOL : 1

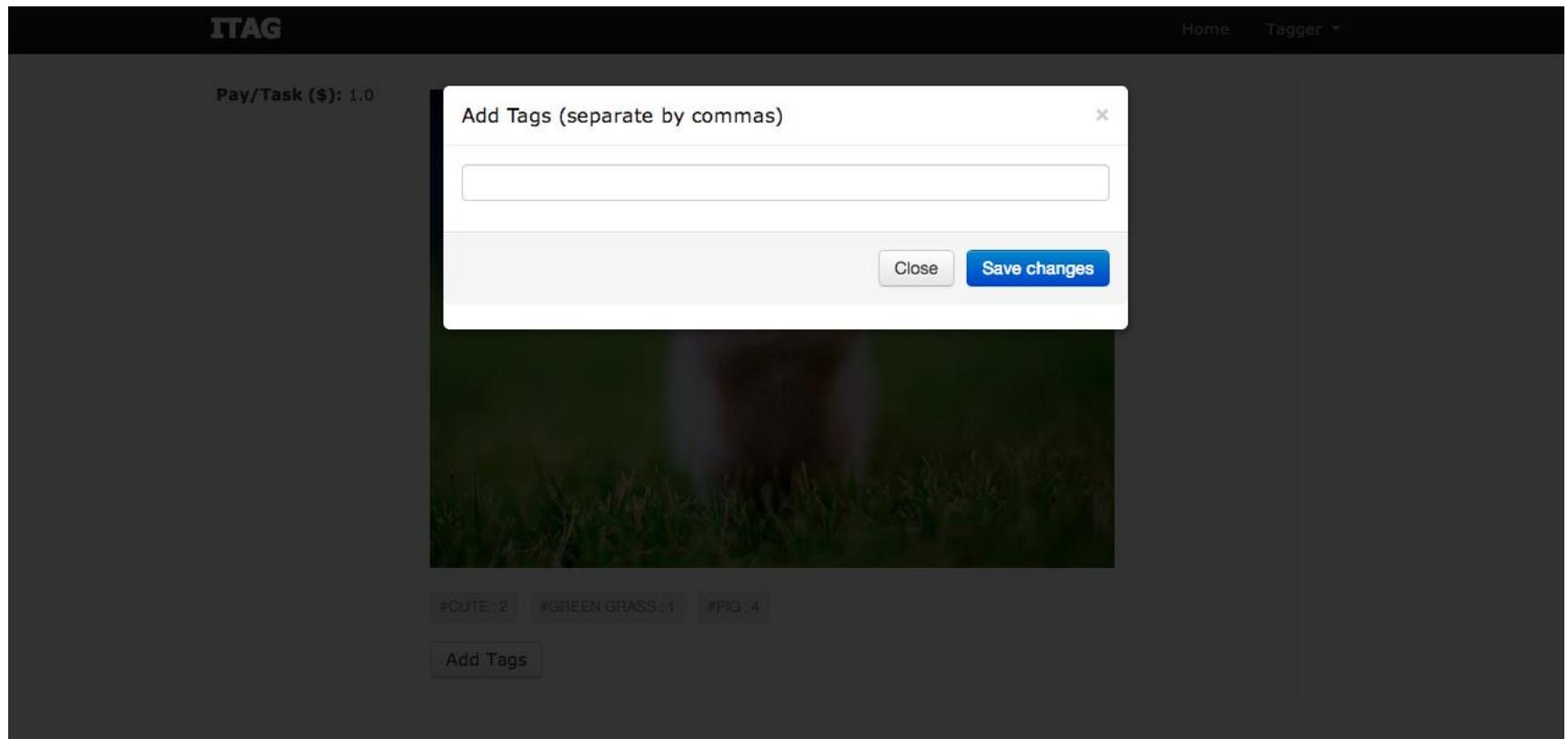
#SITE : 1 #WILLOW : 1

Add Tags

- A picture with existing tags

iTag (12)

93



- Add your tags for the picture

iTag (13)

94

ITAG

Home Tagger ▾

Pay/Task (\$): 1.0



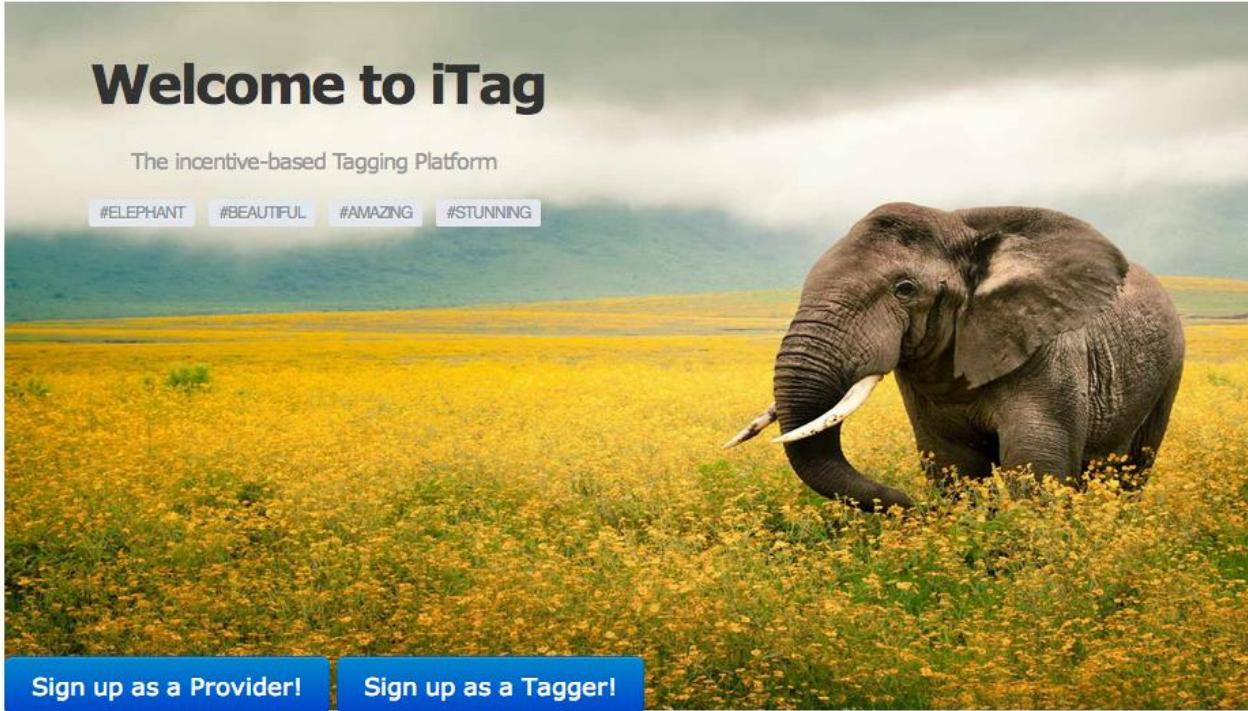
#CUTE : 2 #GREEN GRASS : 1 #PIG : 4

Add Tags

- Another picture for you to tag

iTag (14)

95



The screenshot shows the homepage of the iTAG platform. At the top, there is a dark header bar with the word "ITAG" in white. To the right of the header are two links: "Home" and "Provider". Below the header is a large, scenic photograph of an elephant standing in a field of yellow flowers under a cloudy sky. Overlaid on this image is the text "Welcome to iTag" in a large, bold, dark font. Below this, in a smaller font, is the subtitle "The incentive-based Tagging Platform". Underneath the subtitle are four small, light-colored rectangular buttons containing the hashtags "#ELEPHANT", "#BEAUTIFUL", "#AMAZING", and "#STUNNING". At the bottom of the page, there are two blue call-to-action buttons: "Sign up as a Provider!" on the left and "Sign up as a Tagger!" on the right.

Summary

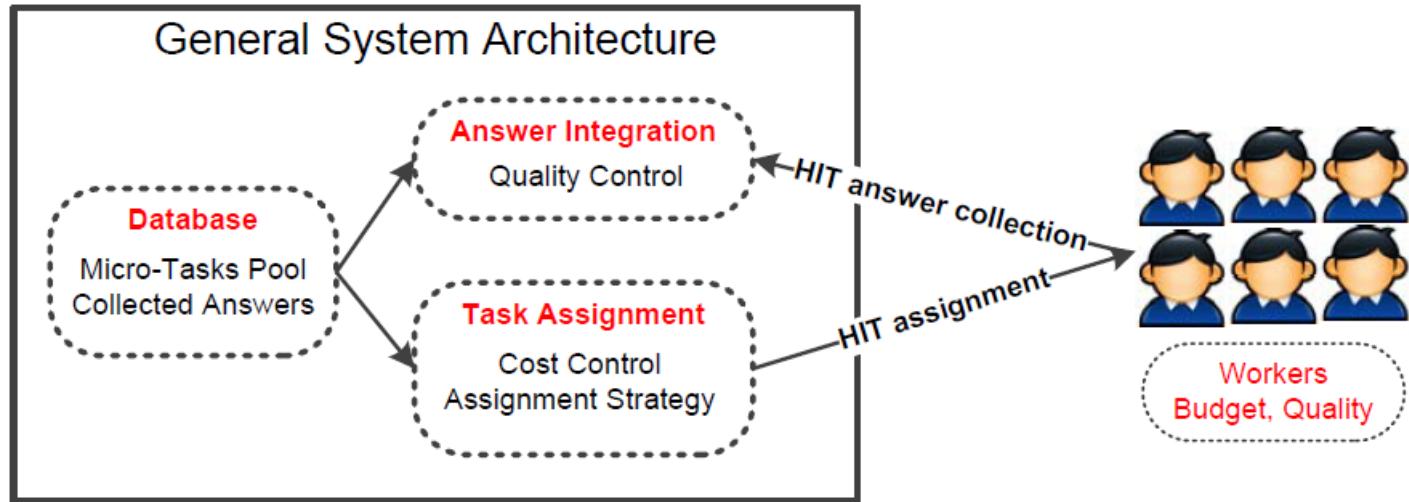
96

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- Define Tag Data Quality
- Incentive-Based Tagging
- Effective Assignment Strategies

Review : Framework

97



- **1. Answer Integration:**

How to integrate answers from workers ?

- **2. Task Assignment:**

Which tasks are chosen to assign to a worker ?

- **3. Database:**

How to store crowdsourced data?

Review: Answer Integration

98

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

□ Half Voting

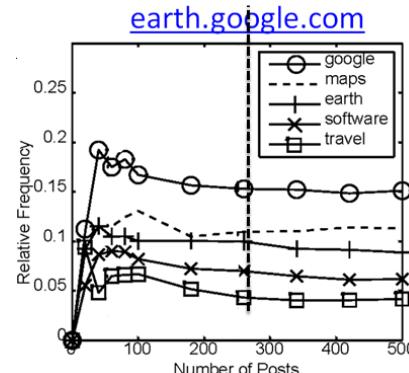
[CIKM'13]

Are they the same?
iPad 2 = iPad Two

YES NO

SUBMIT

□ Relative Frequency
[ICDE'13, ICDE'14]



Review: Task Assignment

99

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- **Assignment Strategy [ICDE'13, ICDE'14]**

Free Choice, Round Robin, FPF, ...

- **Cost Control [CIKM'13]**

Challenge 1: Manage answer integration & task assignment

100

2 dimensions (nature format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

How about majority voting and Bayesian voting?

Which kind of answer integration techniques are better?

Challenge 2: Uncertainty of Crowdsourced Data

101

Steve Jobs is great.
Choose the sentiment of the sentence.

positive
 neutral
 negative

Bayesian voting
(‘positive’:66%,
‘neutral’:32%,
‘negative’:2%)

★ How can database handle this uncertain data (or data with distribution) ?



Traditional and Crowdsourced database does not provide support for handling uncertain data

Open questions

102

- **Can we use a probabilistic database to support crowdsourcing technologies?**
- **How to choose a probability database, and how to populate it by crowdsourced data?**
- **How to update the probability database as more answers are collected?**

Solution: Probabilistic Database

103

Probabilistic databases are databases where the value of some *attributes* or the presence of some *records* are **uncertain** and known only with some **probability**.



**help manage uncertain/probabilistic data
the uncertainty exists in tuples/attributes**

Types of Uncertainty

104

□ Tuple-level uncertainty

A tuple is a random variable that has a Boolean domain: it is true when the tuple is present and false if it is absent

entity	entity	prob.
ipad2	Ipad two	0.8
ipad2	Iphone 2	0.3

□ Attribute-level uncertainty

An attribute represents a random variable, whose domain is the set of values that may take

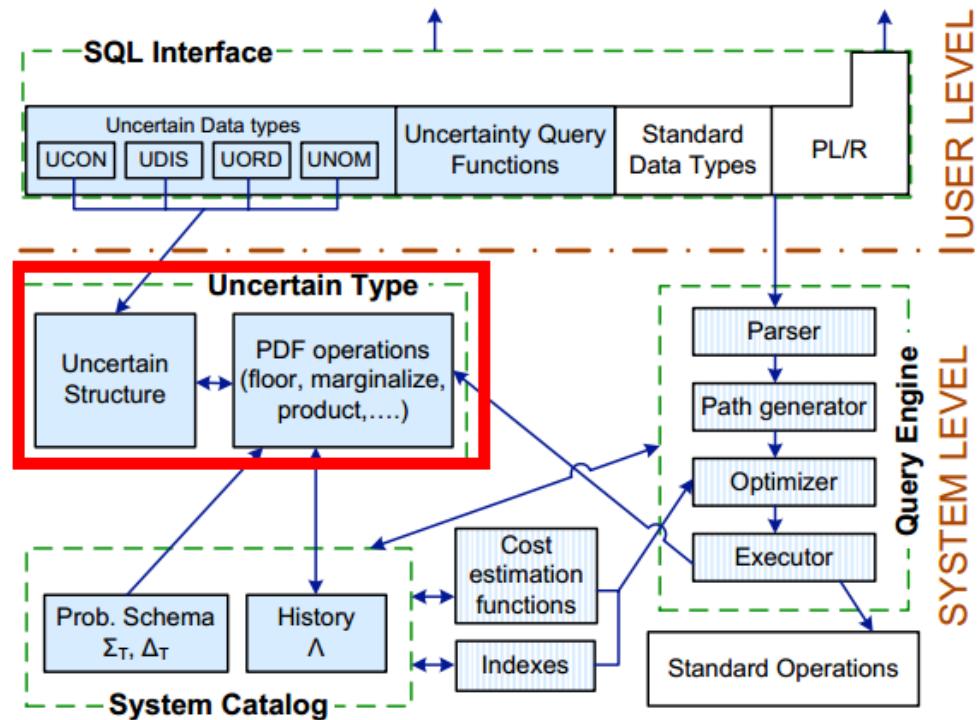
No.	sentiment		
	positive	Neutral	Negative
1	0.66	0.32	0.02

Orion

105

- A Database System For Managing Uncertain Data

- System Architecture
<http://orion.cs.purdue.edu/>



S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, and R. Shah, “Orion 2.0: native support for uncertain data,” in SIGMOD ’08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 1239–1242

Challenge 3: Machine Learning and Crowdsourcing

106

- **Can we develop a better method which gets the best from the two communities?**

- For example,
 - Apply machine learning to handle some clear cases
 - Use human effort to work on the difficult questions



Useful Links

<http://i.cs.hku.hk/~ckcheng/talks/ICDE2013-tagging.pptx>

<http://i.cs.hku.hk/~lymo/paper/cikm13.pptx>

<http://www.just.edu.jo/~amerb/teaching/2-12-13/cs728/Crowdsourcing.pdf>



Thank you!

Reynold Cheng
ckcheng@cs.hku.hk

<http://www.cs.hku.hk/~ckcheng>