THE UNIVERSITY OF HONG KONG

DEPARTMENT OF COMPUTER SCIENCE

PAKDD 2014 Tutorial

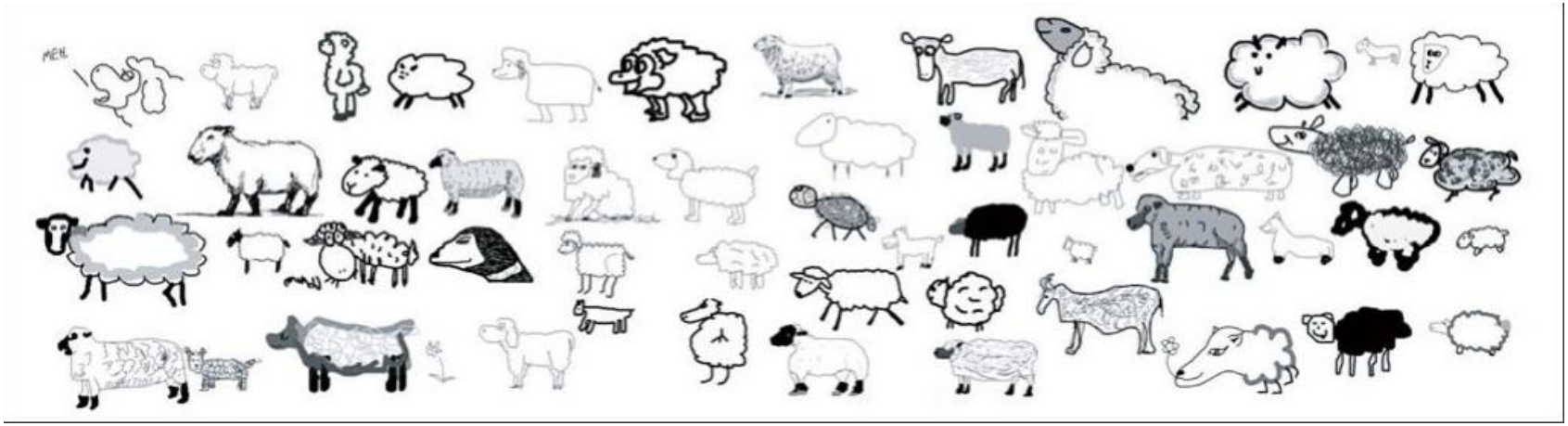
Managing the Quality of Crowdsourced Databases

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

Sheep Market

- Draw a Sheep facing left

Pay each worker $0.02
Collect 10,000 drawing sheep



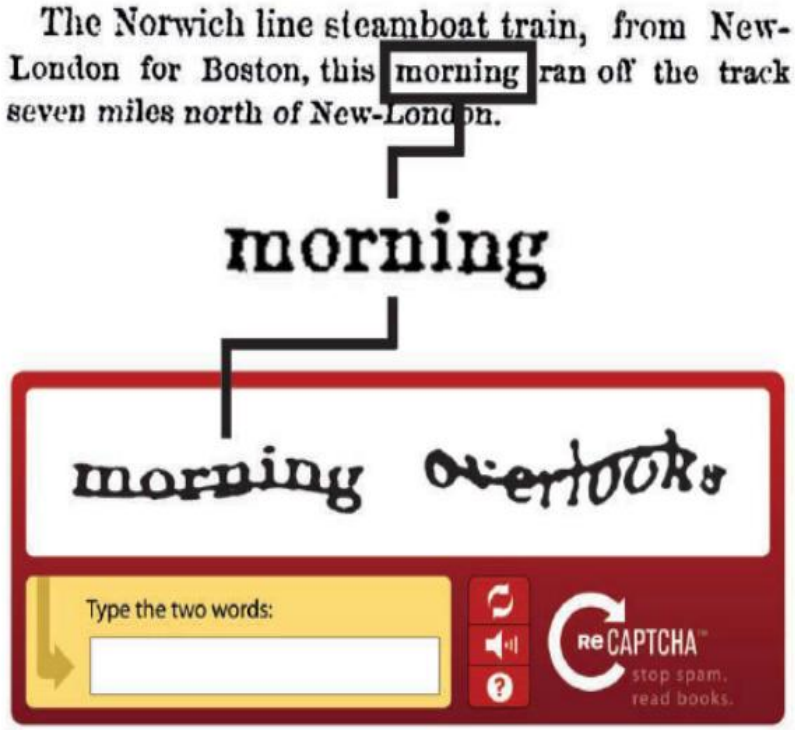
http://www.thesheepmarket.com/

Optical Character Recognition (OCR)

CAPTCHA

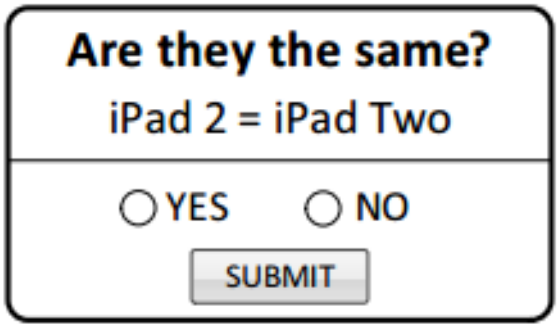
Completely Automated Public Turing test to tell Computers and Humans Apart

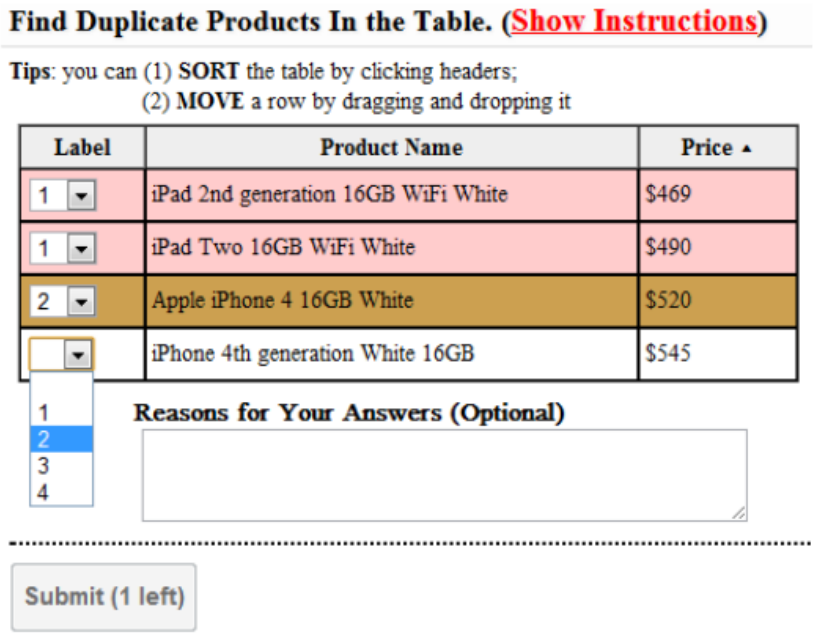
ReCAPTCHA



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)

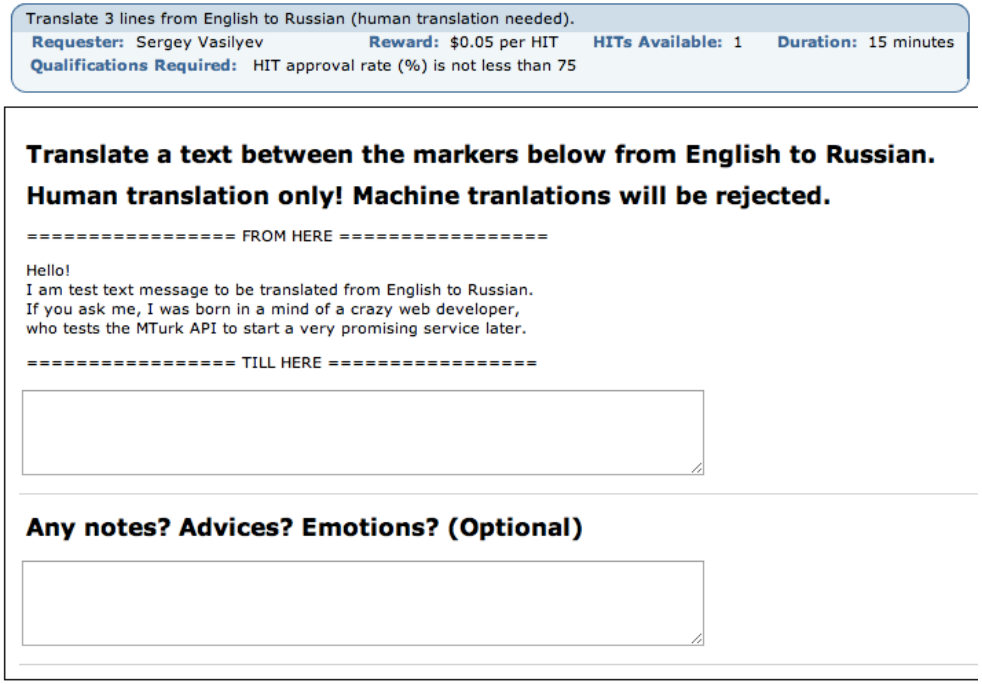




J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution.PVLDB, 5(11):1483-1494,2012.
J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng. Leveraging transitive relations for crowdsourced joins. In SIGMOD Conference, pages 229-240, 2013.

Natural Language Processing (NLP)

- Translation



[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)

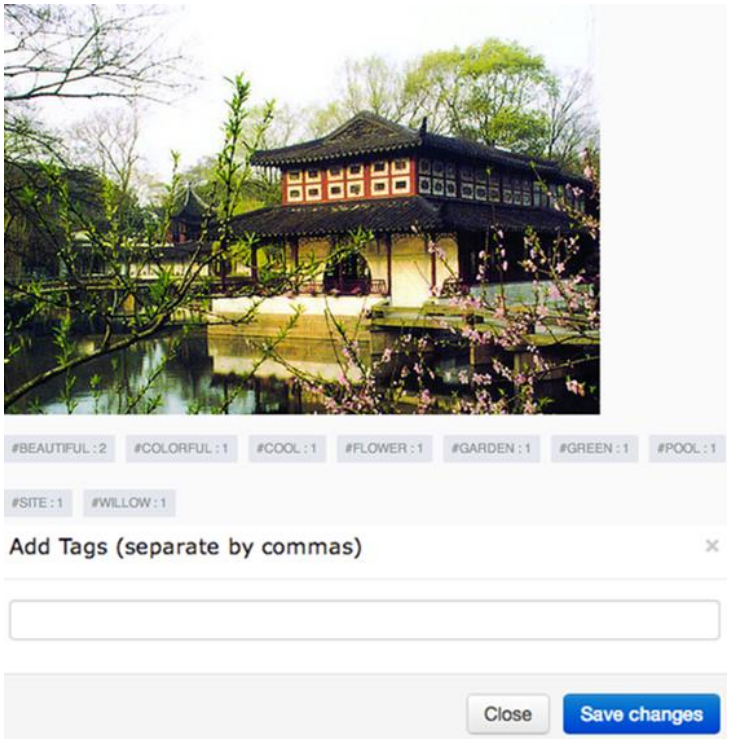
- **Painting Similarity**

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

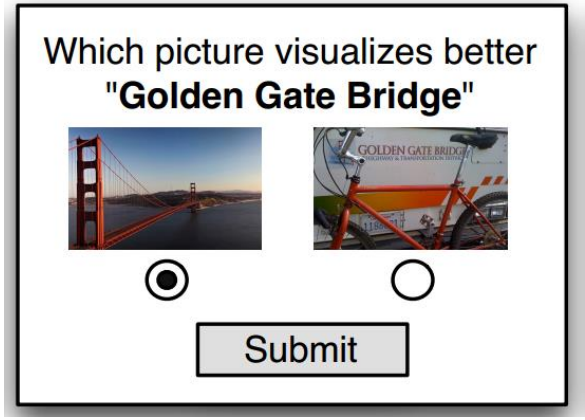


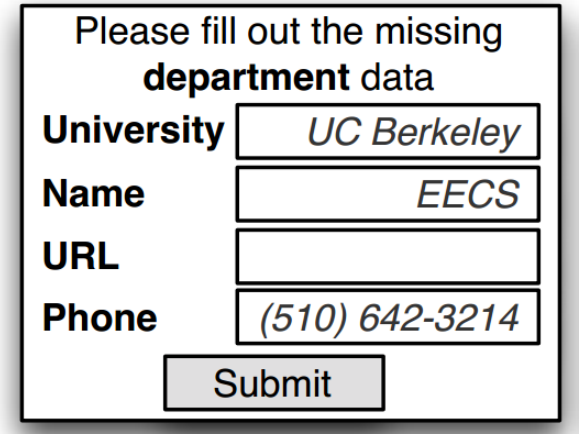
User interface for providing "tag" keywords

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?

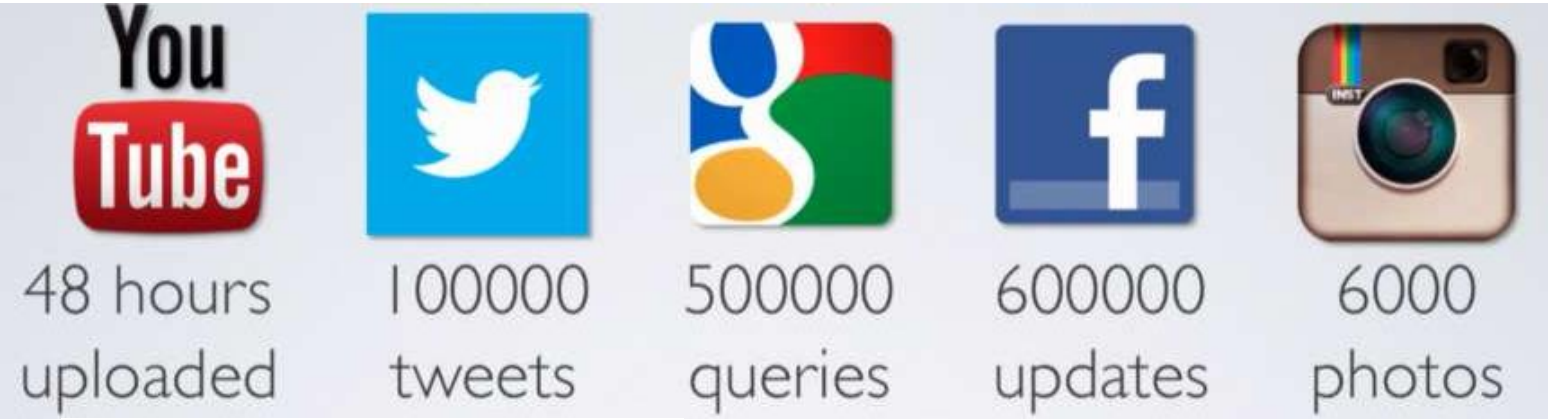




M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Why crowdsourcing?

- **Every Minute**



Aditya Parameswaran ,Human-Powered Data Management ,
http://msrvideo.vo.msecnd.net/rmcvideos/185336/dl/185336.pdf

Crowdsourcing Platforms

- Voluntary

- Incentive-based

Crowdsourcing Model

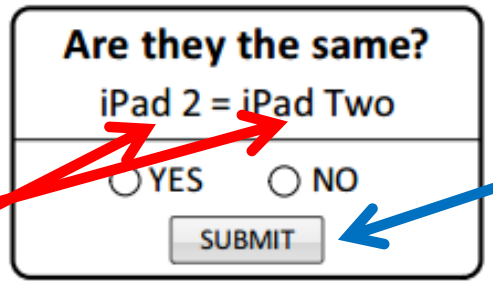
A requester asks a crowd to do Human Intelligence Tasks (HITs) to solve problems.

problem: entity resolution (ER)

HIT: comparison questions

crowd: Internet users

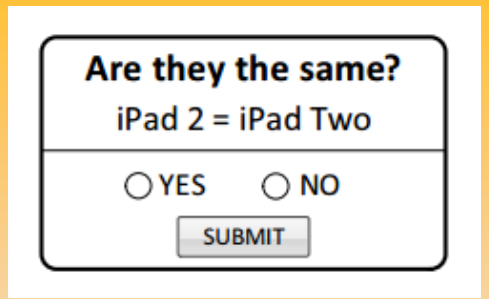
ID	Object
o_1	iPhone 2nd Gen
o_2	iPhone Two
o_3	iPhone 2
o_4	iPad Two
o_5	iPad 2
o_6	iPad 3rd Gen

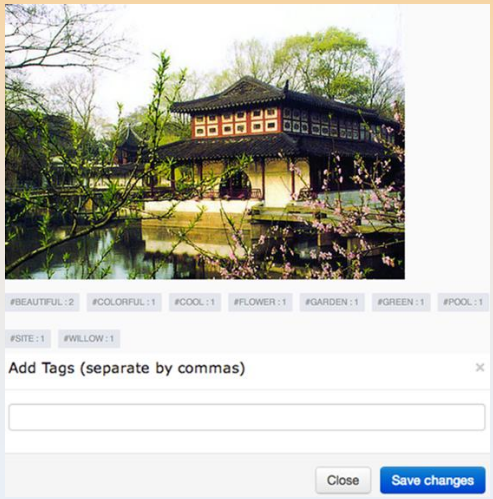


Amazon Mechanical Turk (AMT)

Requesters

HITs

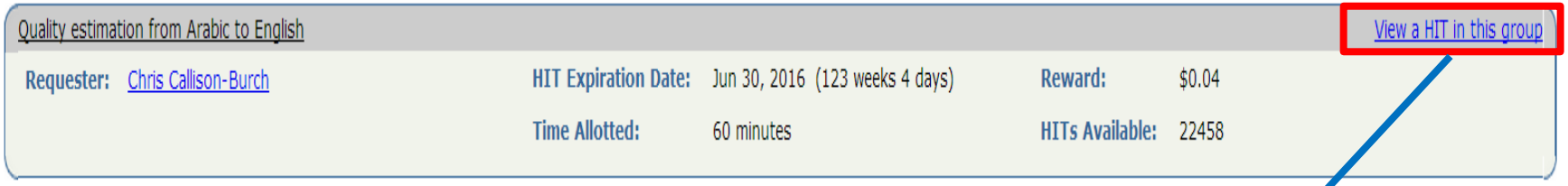




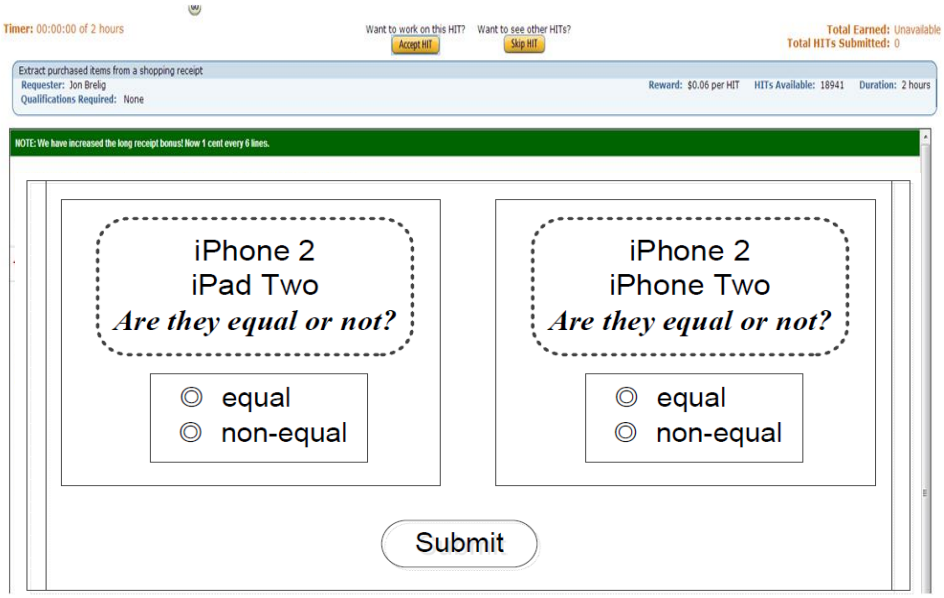
Workers

Human Intelligence Tasks (HITs)

- HIT group on AMT



- A HIT



AMT Statistics

- **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)

- **Types of tasks**

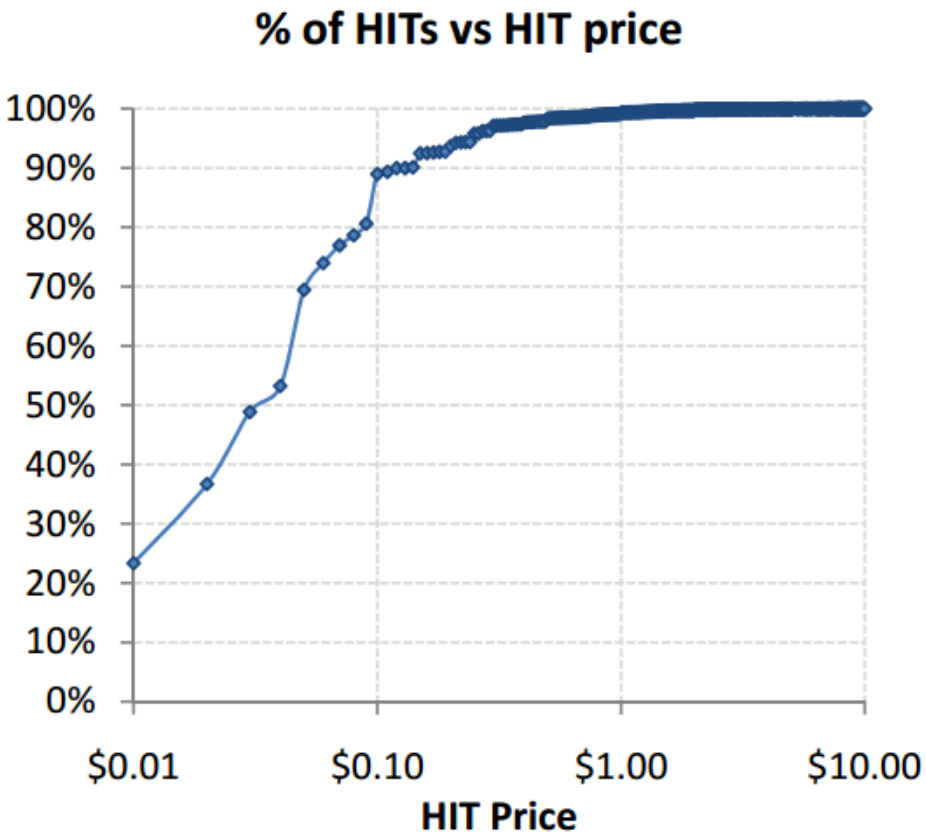
From January 2009 till April 2010, we collected 165,368 HIT groups, with a total of 6,701,406 HITs, from 9,436 requesters. The total value of the posted HITs was $529,259.

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
A1197OGL0WOQ3G	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
AD14NALRDOSN9	retaildata	113	158,206	$2,118	Object classification
A2RFHBFTZHX7UN	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

Ipeirotis, P. G. (2010a). Analyzing the Amazon Mechanical Turk marketplace. ACM XRDS, 17, 16–21.

AMT Statistics (HIT price)

- 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]

Ipeirotis, P. G. (2010a). Analyzing the Amazon Mechanical Turk marketplace.ACM XRDS, 17, 16–21.

AMT APIs

- Application Programming Interfaces

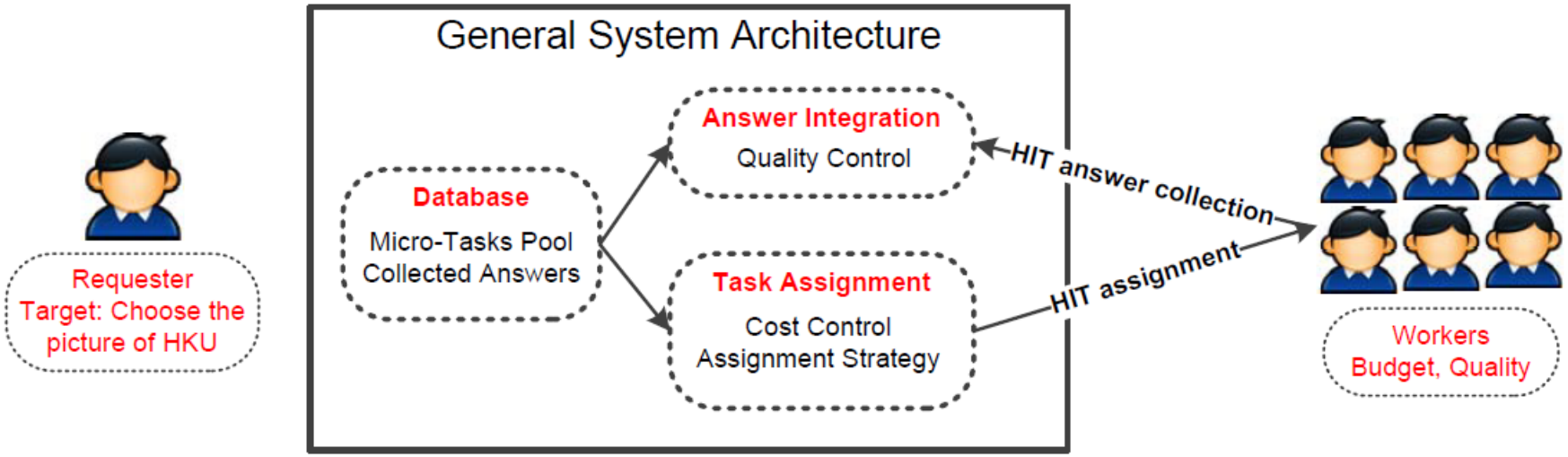
  Official: C# Java Perl Ruby
  Open-source: Python (boto)

- 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



- **1.Answer Integration:**

  How to integrate answers from workers ?

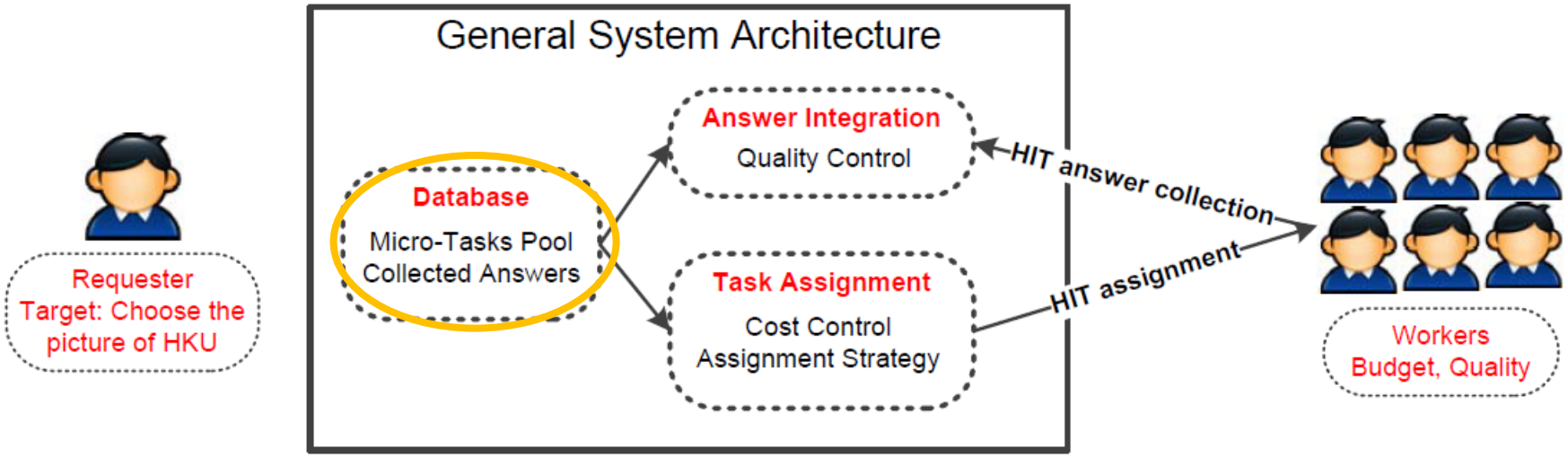
- **2.Task Assignment:**

  Which tasks are chosen to assign to a worker ?

- **3.Database:**

  How to store crowdsourced data?

Crowdsourcing Framework



- **1.Answer Integration:**

  How to integrate answers from workers ?

- **2.Task Assignment:**

  Which tasks are chosen to assign to a worker ?

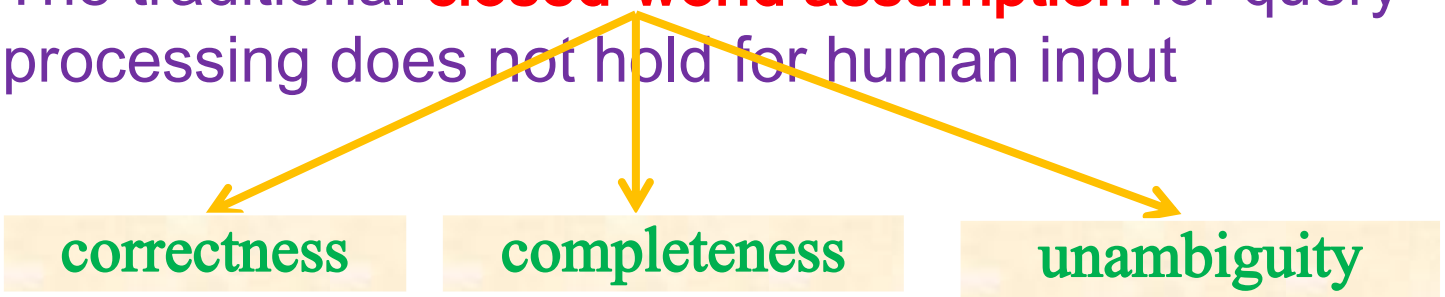
- **3.Database:**

  How to store crowdsourced data?

Crowdsourced Databases

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

- 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

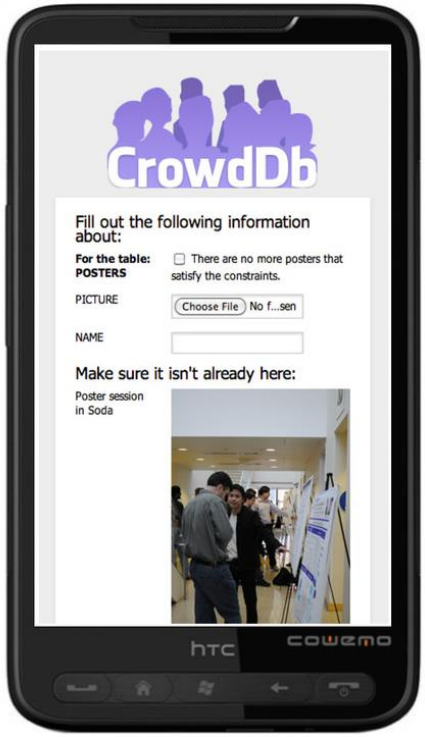
```
SELECT image FROM picture
WHERE topic = "Business Success"
ORDER BY relevance LIMIT 1;
```

cannot decide the relevance

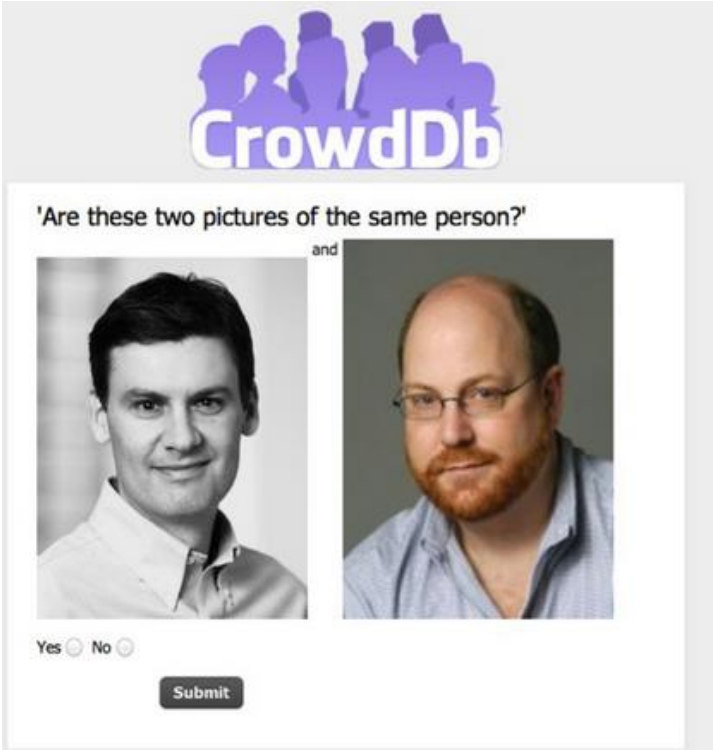
M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

CrowdDB (SIGMOD'11)

- Mobile APP

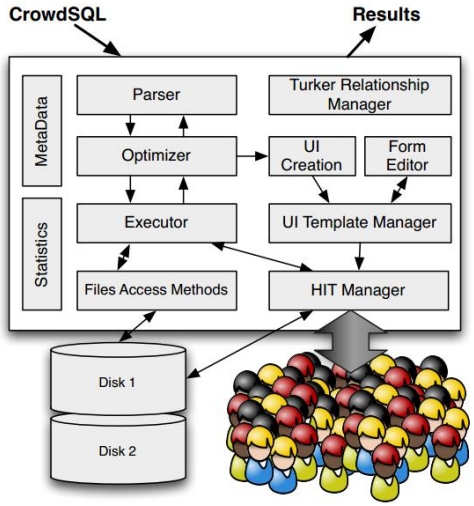


- AMT Platform



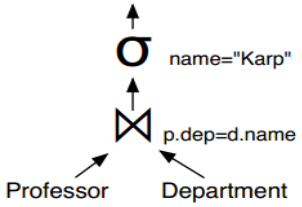
M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.
https://speakerdeck.com/rxin/crowddb-vldb-demo-award-talk

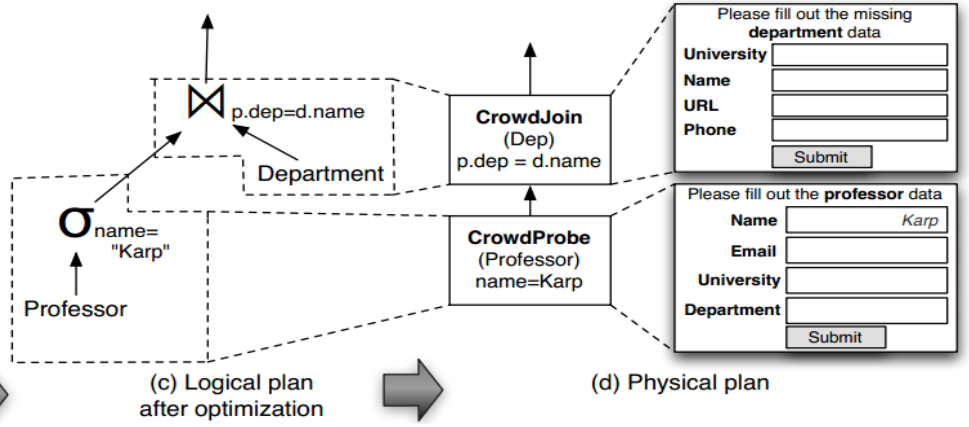
CrowdDB Architecture




```
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"
```

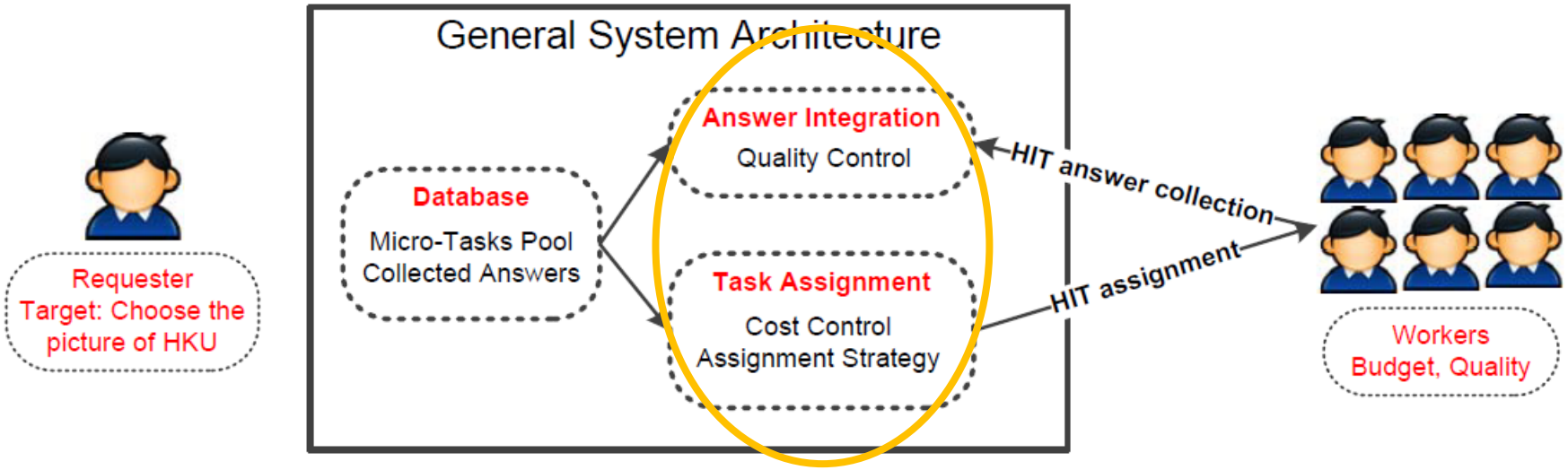




(a) PeopleSQL query → (b) Logical plan before optimization → (c) Logical plan after optimization → (d) Physical plan

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



- **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

- **Format of Questions**

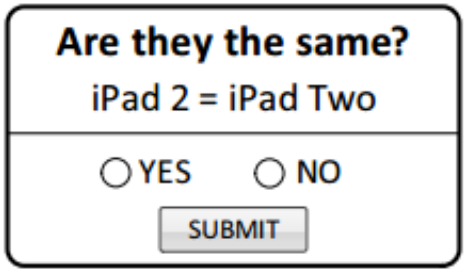
  How are the questions presented to workers?

- **Nature of Answers**

  Does the question have a true answer?

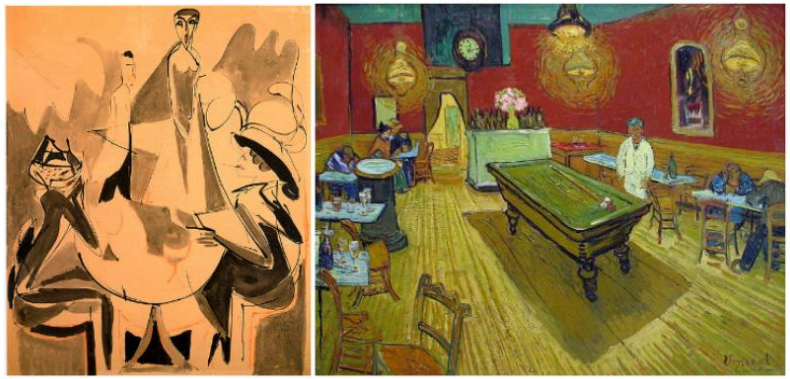
Format of Questions (1)

- Binary Choice Question (BCQ)



ER (Entity Resolution):
Are two entities equal?

- Multiple Choice Question (MCQ)



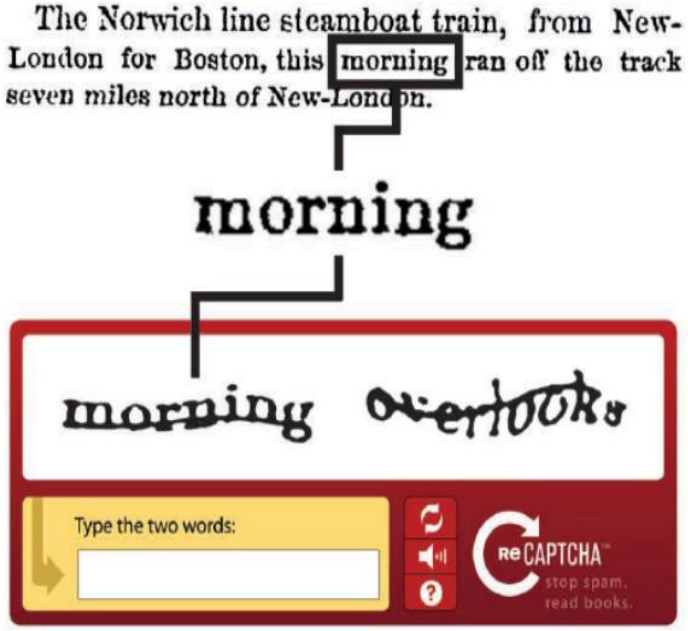
How similar is the artistic style in the paintings above?

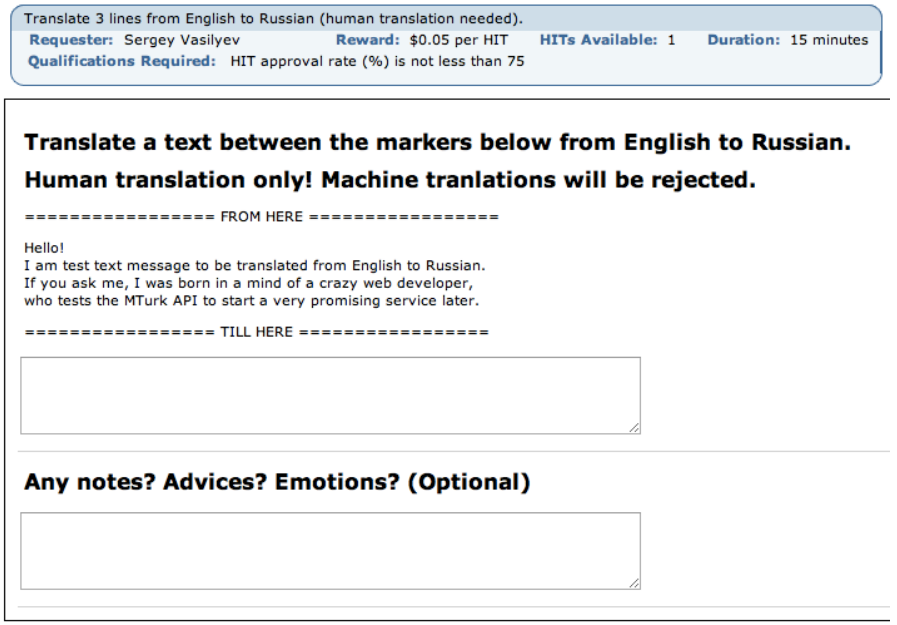
- Very similar
- Somewhat similar
- Neither similar nor dissimilar
- Somewhat dissimilar
- Very dissimilar

CV (Computer Vision):
similarity between two paintings

Format of Questions (2)

- Open questions with text answers





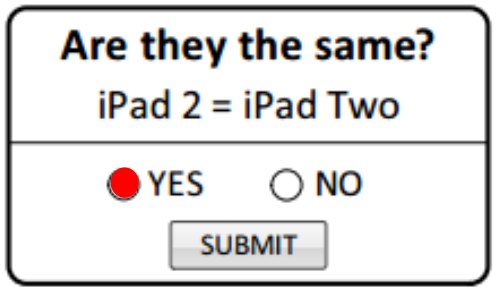
OCR:
reCAPTCHA

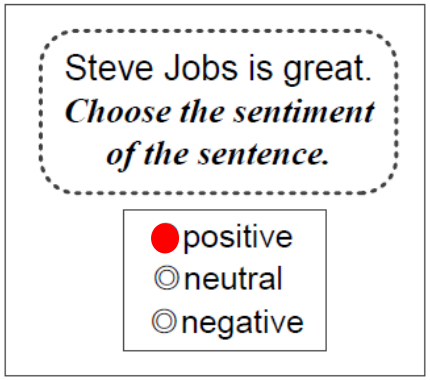
NLP:
language translation

Nature of Answers (1)

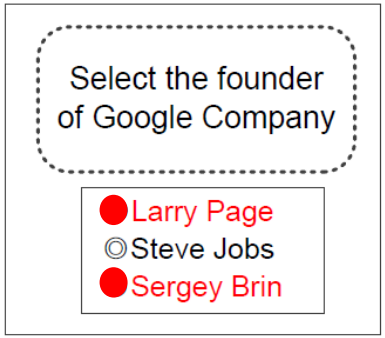
- Ground truth to a question

Single ground truth answer:

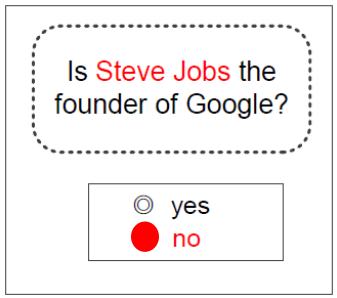


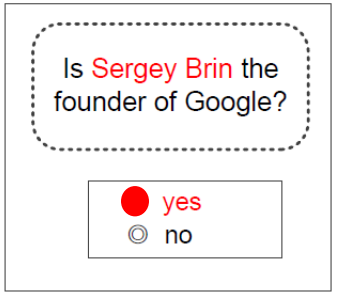


Multiple ground truth answers:



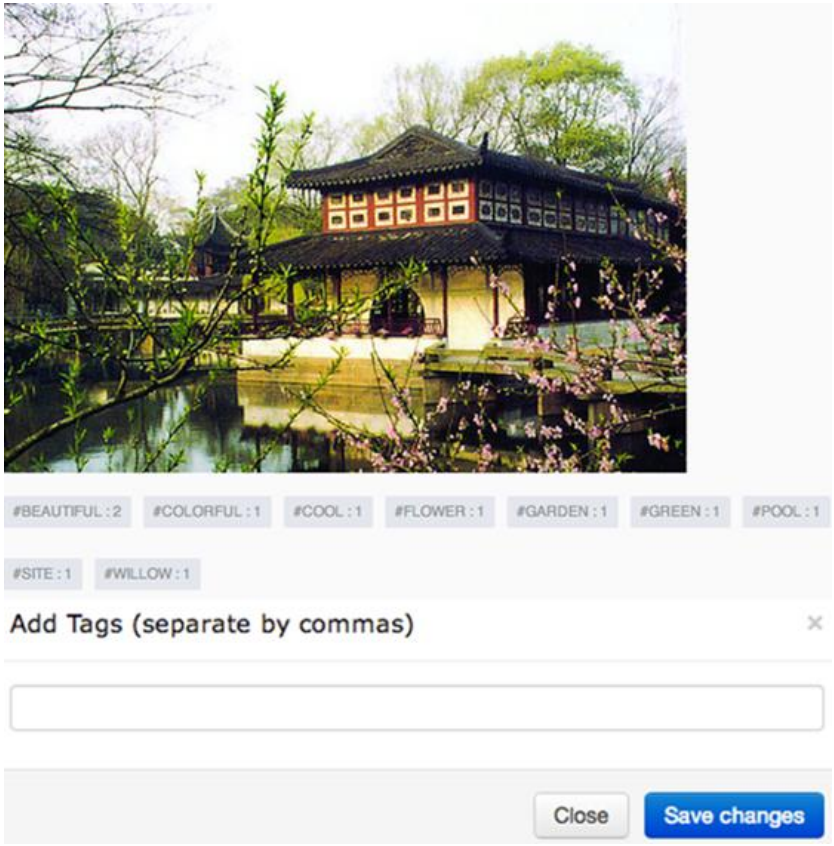
Decomposition →





Nature of Answers (2)

- No ground truth to a question



A Classification of HITs

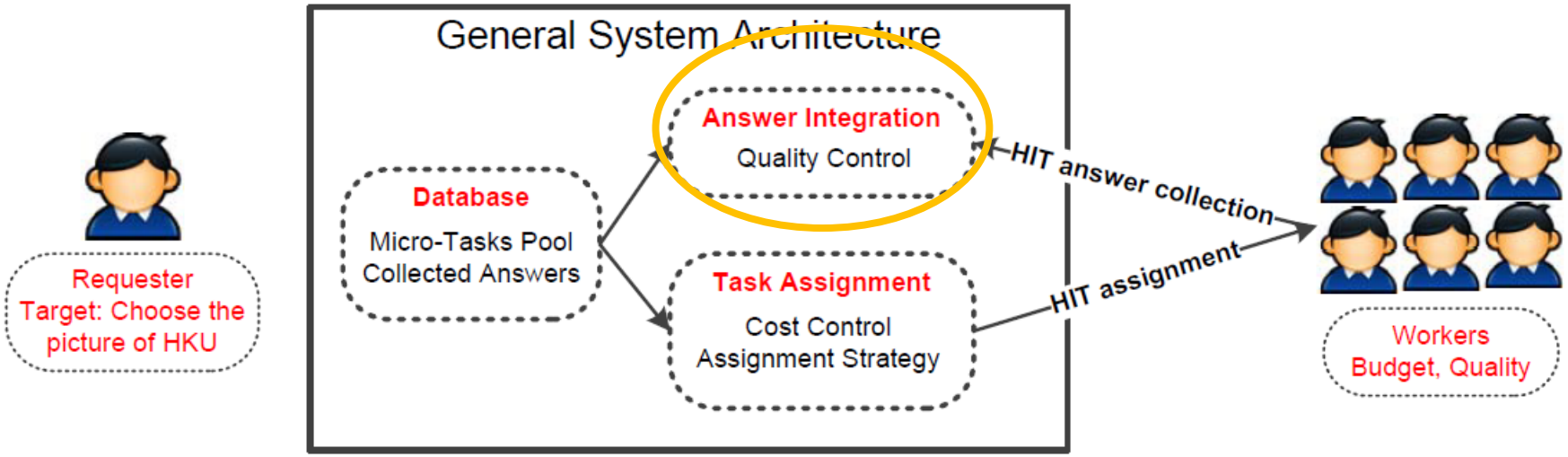
- TWO dimensions

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



- **1.Answer Integration:**

  How to integrate answers from workers ?

- **2.Task Assignment:**

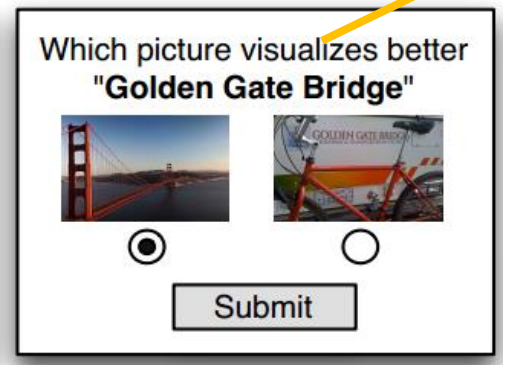
  Which tasks are chosen to assign to a worker ?

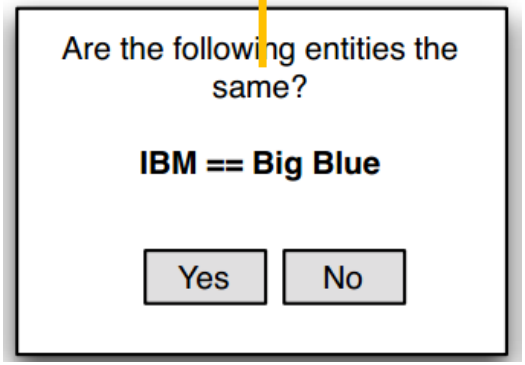
- **3.Database:**

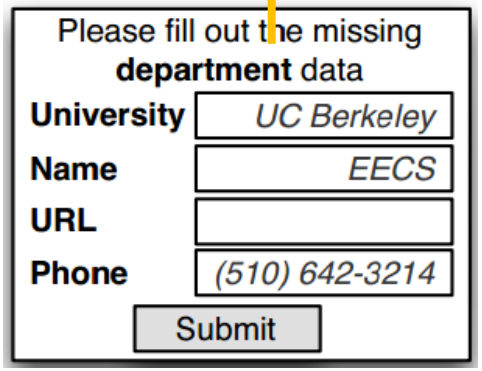
  How to store crowdsourced data?

Answers with Ground Truth

Answer | question	BCQ	MCQ	Open
With Ground Truth			
Without Ground Truth			







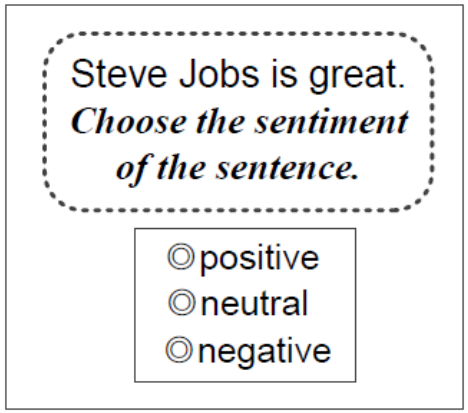
M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Voting Strategies

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

- Half Voting
- Majority Voting
- Bayesian Voting

Example



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

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

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 !

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 ?

- 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. InSIGKDD workshop, pages 64-67,2010.

How to learn worker's quality ?

- **Golden Questions**

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

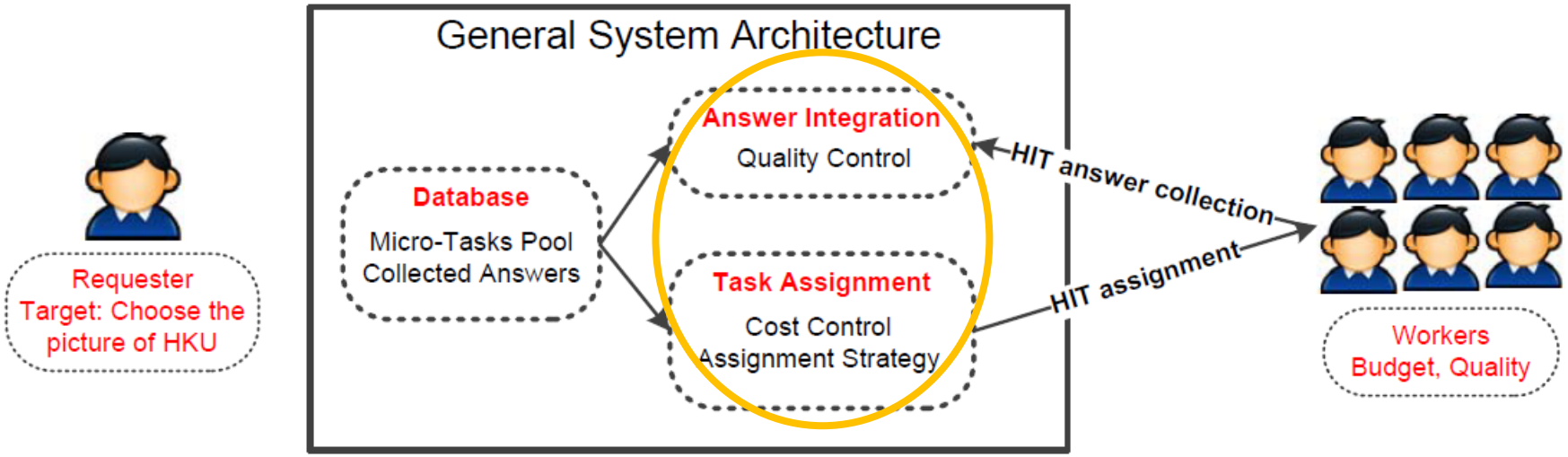
  7 correctly answered questions in 10 golden ones. q= 7/10 =0.7

- **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. InSIGKDD workshop, pages 64-67,2010.

Crowdsourcing Framework



- **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?

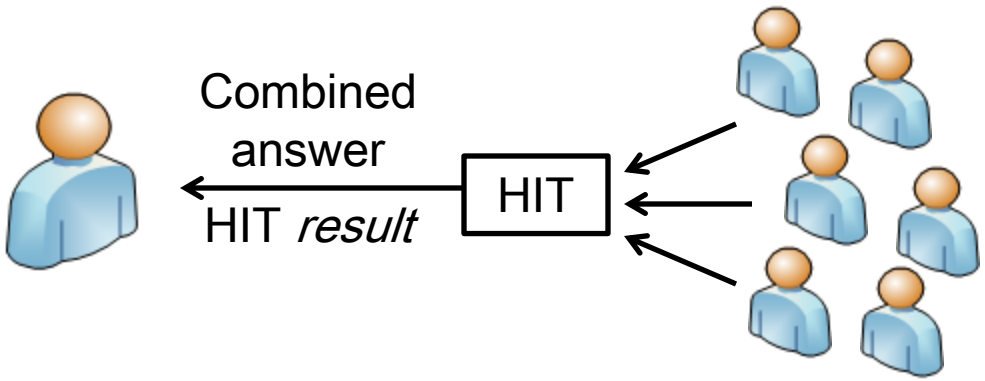
Optimizing Plurality for HITs (CIKM'13)

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

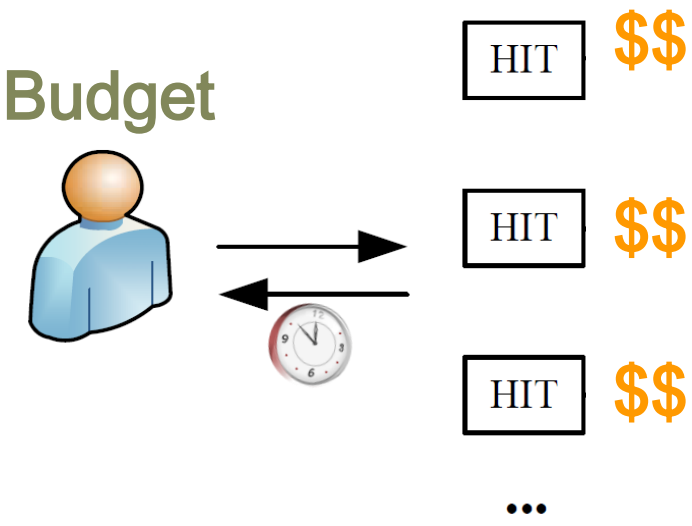
Plurality of HITs

- 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)

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



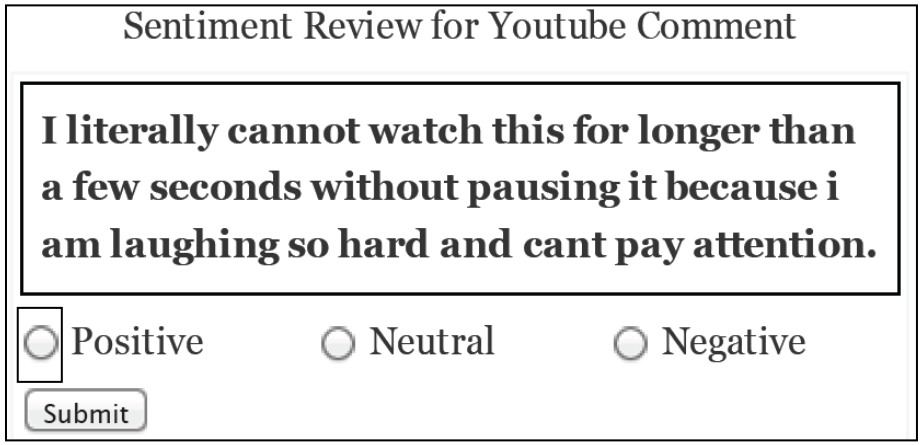
PAP:
wisely assign the right pluralities to various HITs to achieve overall high-quality results

Our Goal

- 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)

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



Data Model

- Set of HITs $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$
- 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

- 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

- 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 $\vec{k} = \{k_1, k_2, \ldots, k_n\}$
- Objective
  - maximize overall average quality

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

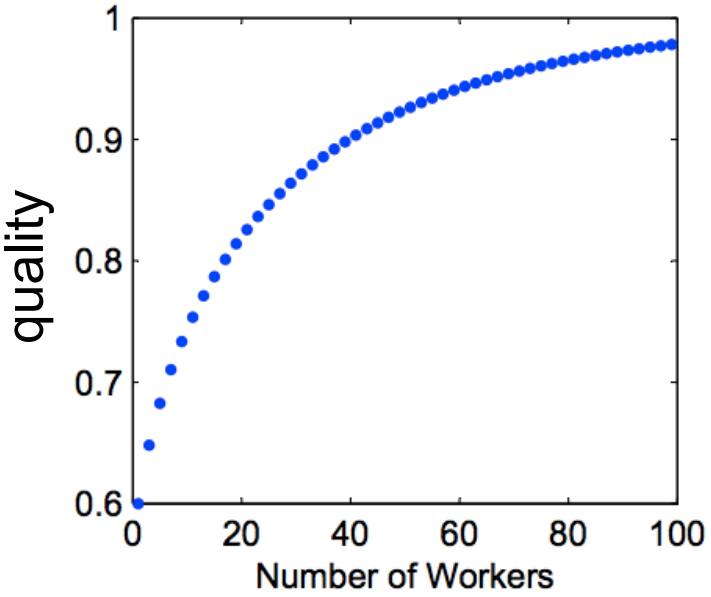
Solutions

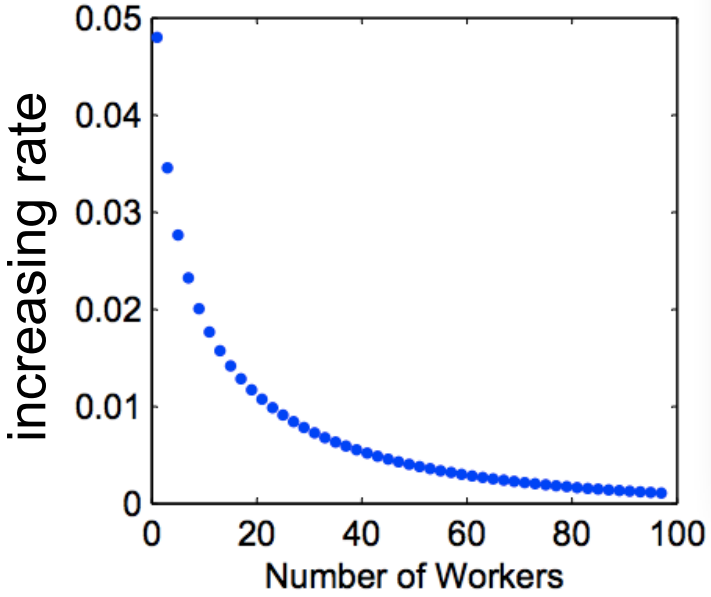
$$\begin{aligned}
&\textbf{given} && B, \mathcal{T} \text{ and } p_i, c_i \text{ for } t_i \in \mathcal{T} \\
&\textbf{maximize} && \mathcal{Q}_t(\mathcal{T}, \vec{k}) \\
&\textbf{subject to} && \sum_{i=1}^{n} k_i c_i \leq B \ \textbf{and} \\
& && k_i = 0 \text{ or a positive odd number}
\end{aligned}$$

- 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

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





Greedy: 2-approximate algorithm

- 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

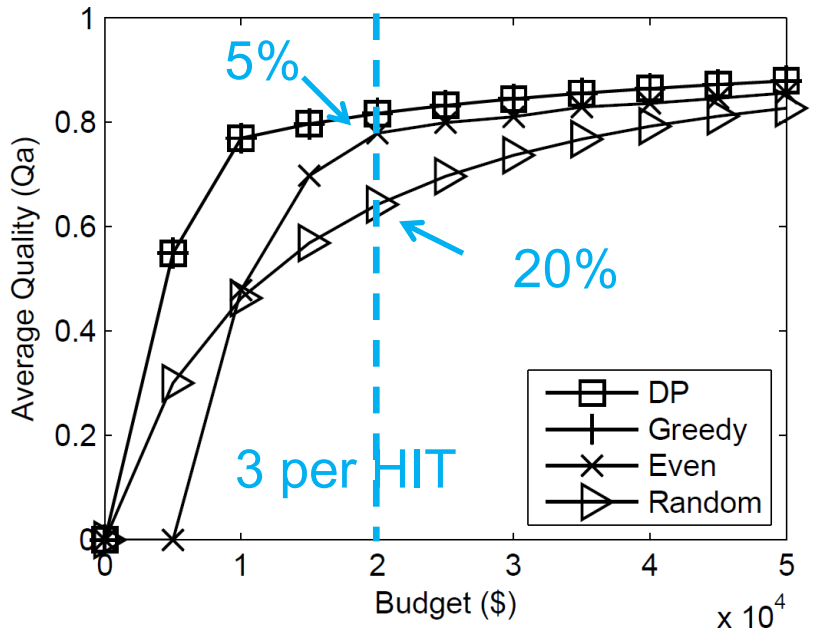
- Observations
  - Many HITs submitted by the same requesters are given the same cost and of very similar nature
- Intuition
  - Group HITs of the same cost and quality function
  - More or less the same plurality for HITs in one group
- Main idea
  - Select a “representative HIT” from each group
  -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

Experiments

- 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

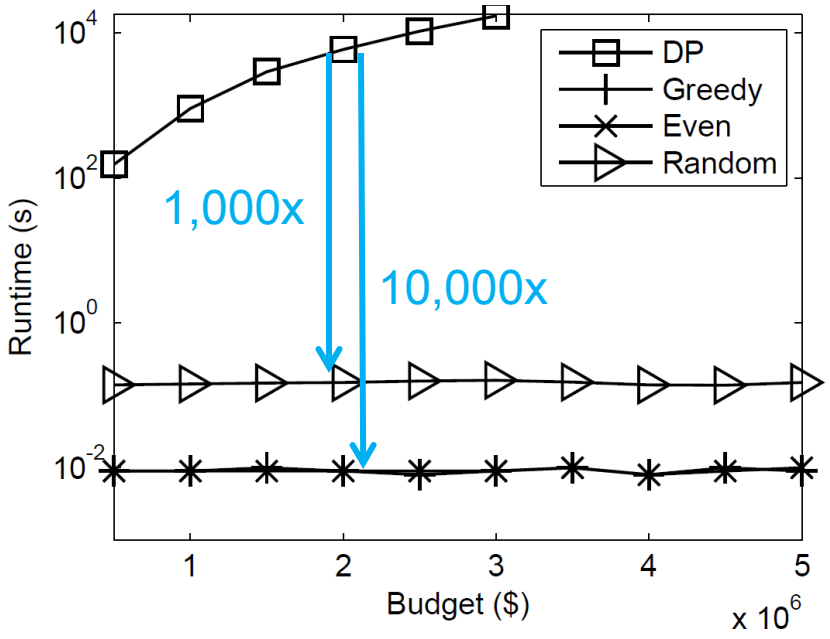
- Competitors
  - *Random*: arbitrarily pick a HIT to increase its plurality until budget is exhausted
  - *Even*: divide the budget evenly across all HITs



Greedy is close-to-optimal in practice

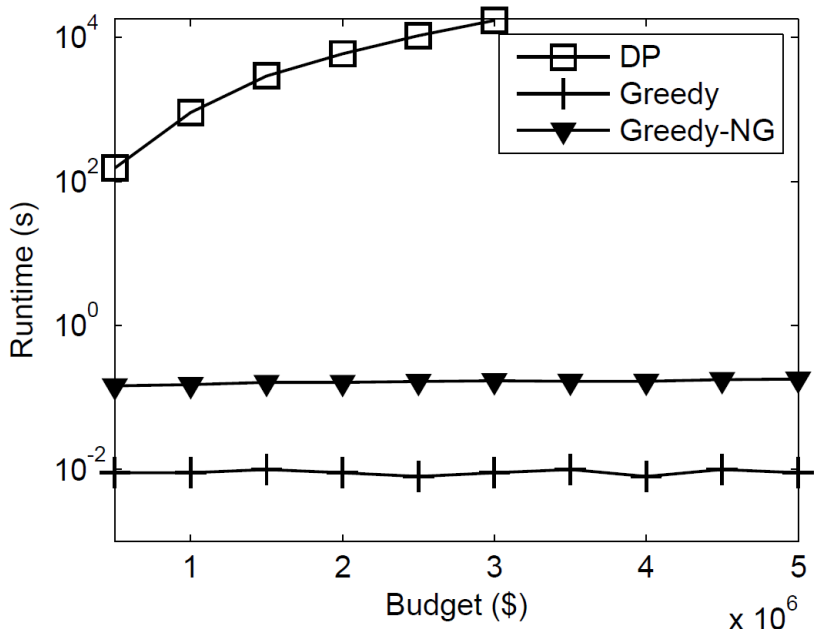
Performance (1)

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



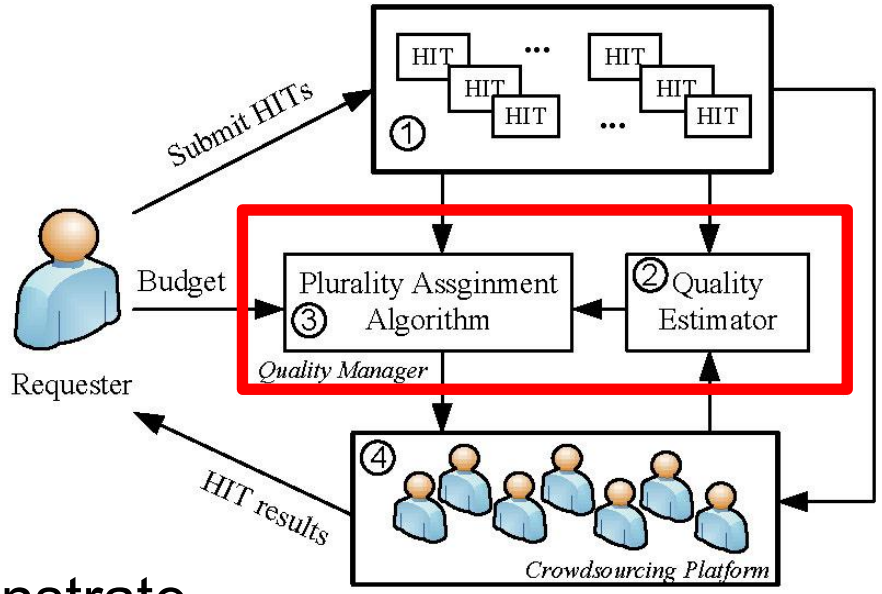
Performance (2)

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



Other HIT Types

- 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

- Objective
  - obtain a complete set of distinct elements for a set query
- Quality function

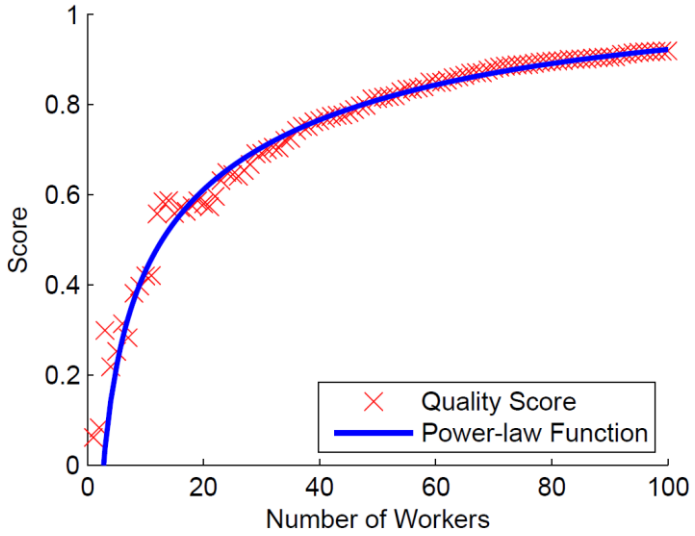
$$\zeta_i(k) = \hat{f}_0 \left[1 - \left(1 - \frac{1 - \hat{C}}{\hat{f}_0} \right)^{k - k_0} \right]$$

- Satisfy monotonicity and diminishing return
  - Greedy can be applied

Crowdsourced Enumeration Queries. MJ Franklin et al. ICDE'13

Tagging Query

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



On Incentive-based Tagging. X. Yang et al. ICDE'13

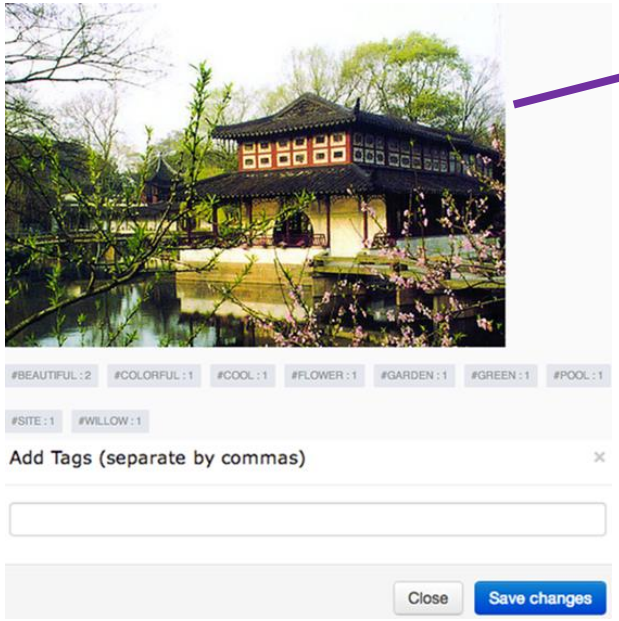
Summary

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

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.

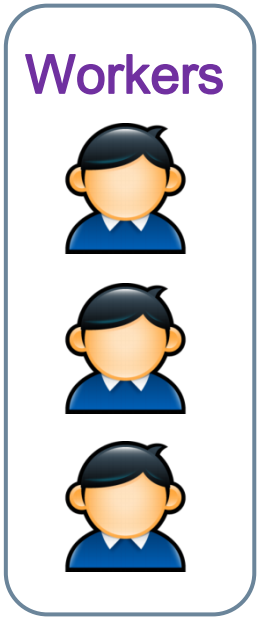
On incentive-based tagging (ICDE'13)

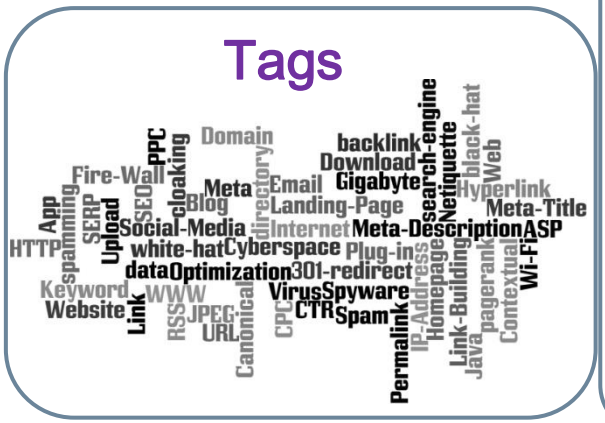
2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

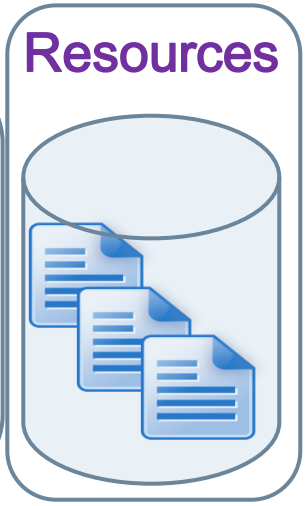
- How to define Tag Data Quality
- Incentive-Based Tagging
- How to select pictures for workers to tag

Collaborative Tagging Systems

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







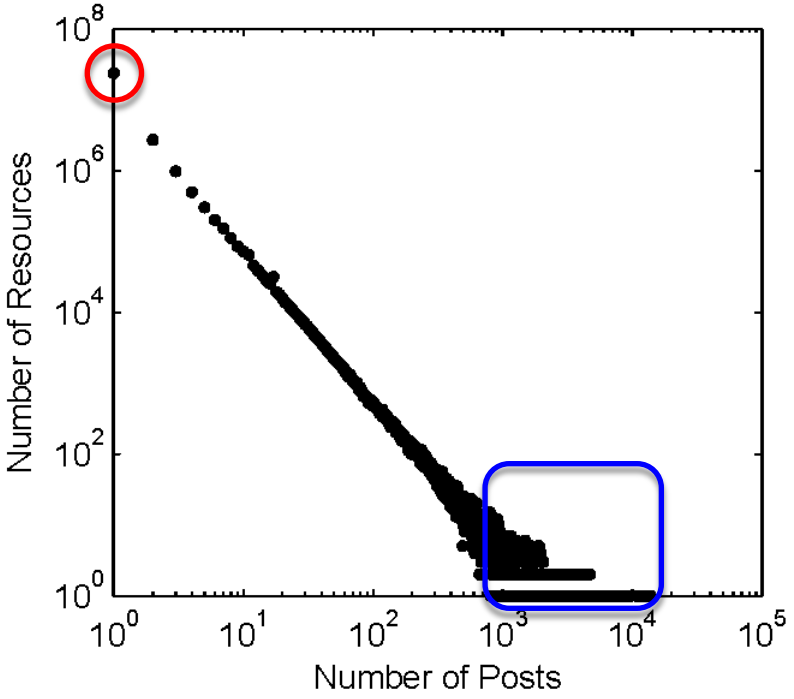
Applications with Tag Data

- 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

- **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



Analyzing Social Bookmarking Systems: A del.icio.us Cookbook. ECAI Mining Social Data Workshop. 2008

Under-Tagging

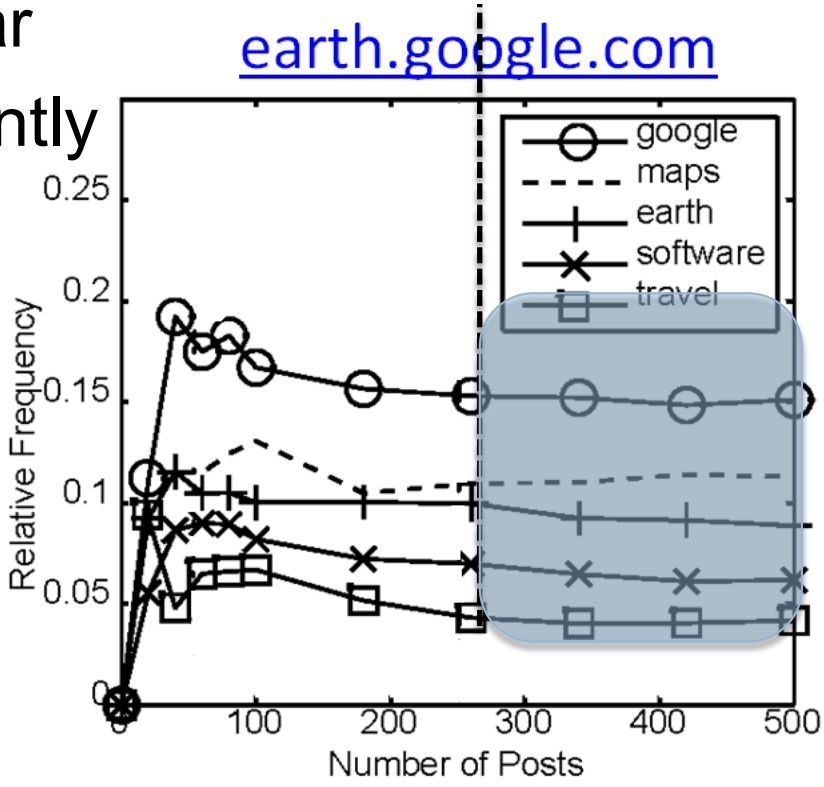
- Resources with **very few** posts have **low quality** tag data
- Low quality of one single post
  - Irrelevant to the resource
    - {3dmax}
  - Not cover all the aspects
    - {geography, education}
  - Don't know which tag is more important
    - {maps, education}

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

Relative Frequency

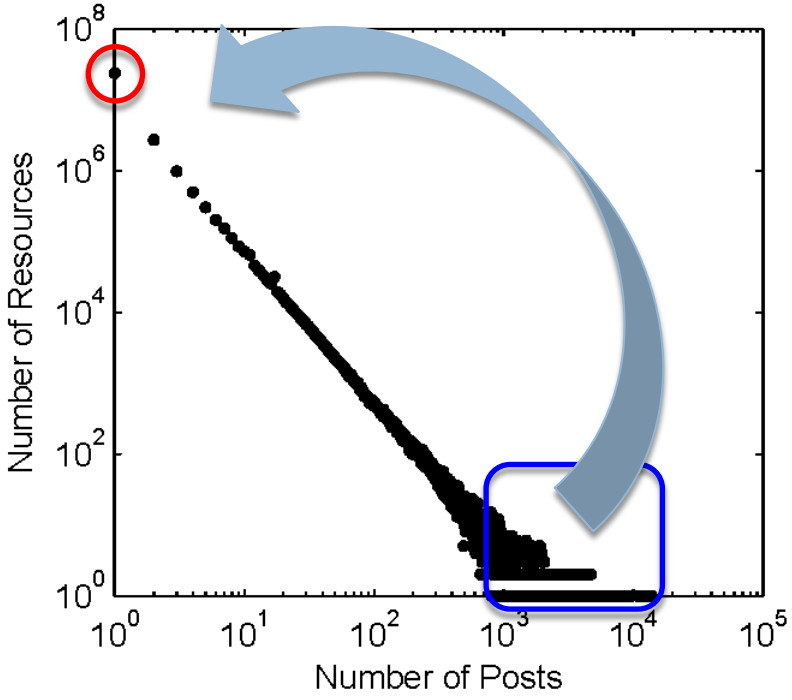
- Relative occurrence frequency of tag t can reflect its importance
  - Irrelevant Tags rarely appear
  - Important tags occur frequently
- Relative Frequency vs. no. of posts
  - >=250, stable

Tagging Efforts are **Wasted**!



Incentive-Based Tagging

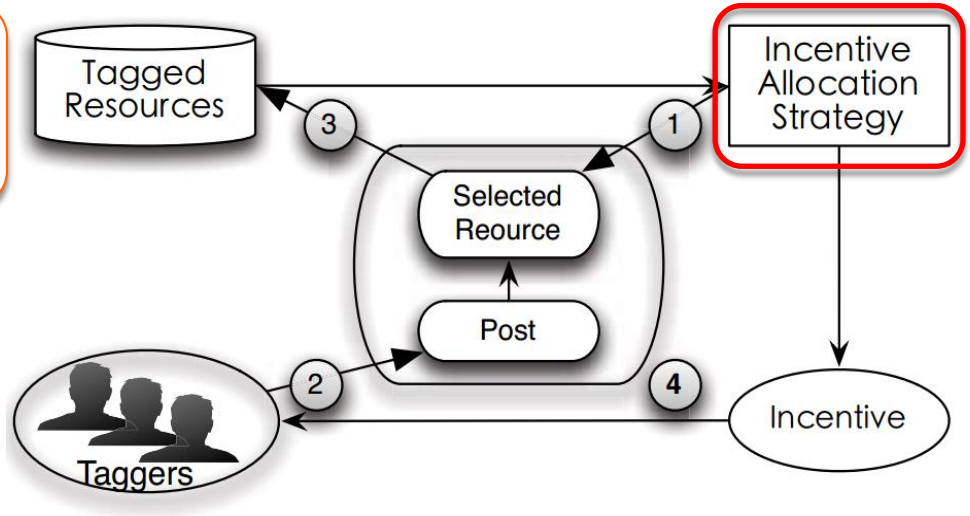
- Guide users' tagging effort
  - **Reward** users for annotating under-tagged resources
- Reduce the number of under-tagged resources
- Save the tagging efforts wasted in over-tagged resources



Incentive-Based Tagging (cont'd)

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

Quality Metric for Tag Data



Data Model

- 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

{maps, education}

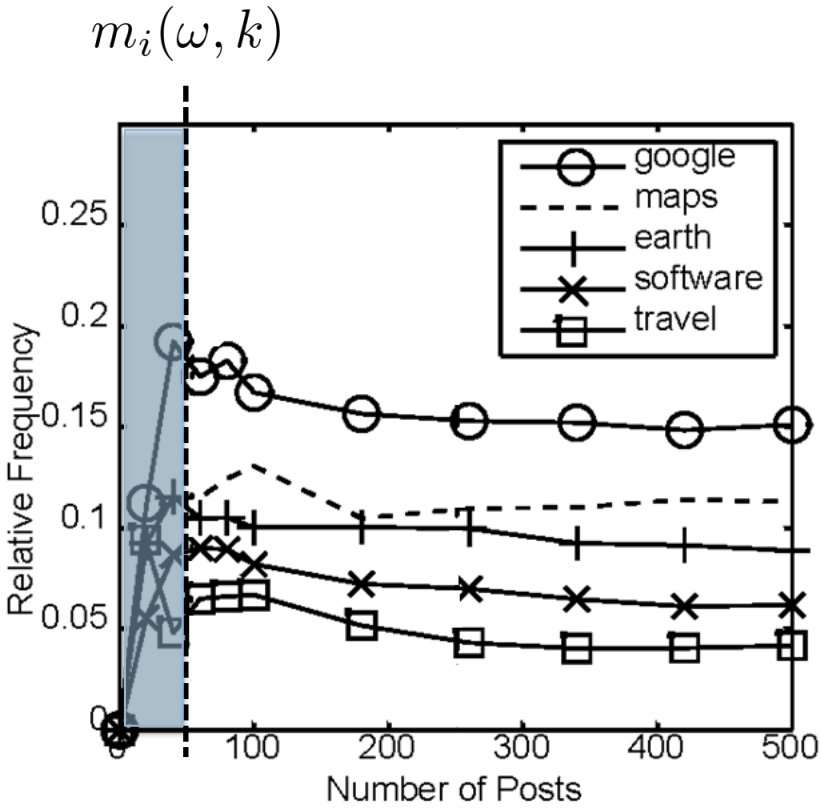
{geography, education}

{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

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



Quality

- For one resource r_i with k posts
  - Similarity between its current *rfd* and its stable *rfd*

$$q_i(k) = s(\vec{F}_i(k), \hat{\varphi}_i)$$

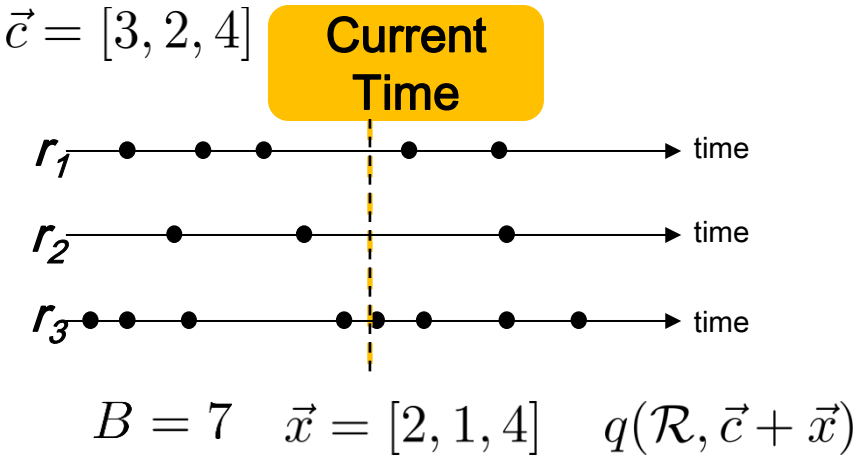
- For a set of resources R
  - Average quality of all the resources

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

Incentive-Based Tagging

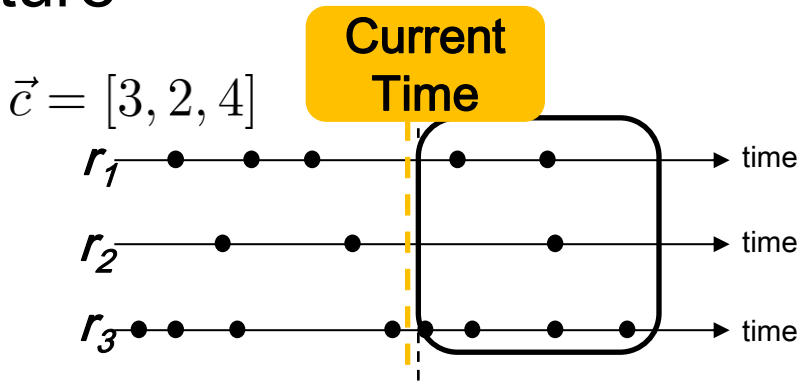
- Input
  - A set of resources $\mathcal{R}$
  - Initial posts $\vec{c}$
  - Budget B
- Output
  - Incentive assignment $\vec{x}$
  - how many new posts should r_i get
- Objective
  - Maximize quality

$$q(\mathcal{R}, \vec{c} + \vec{x})$$



Incentive-Based Tagging (cont'd)

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



Strategy Framework

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

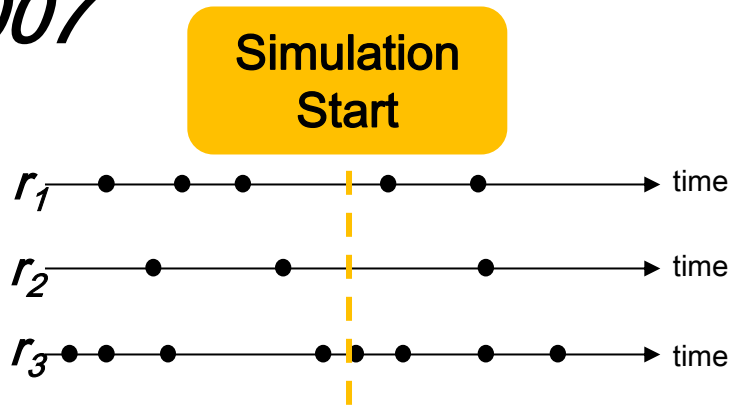
1: **for** $i \leftarrow 1$ to n **do** $x[i] \leftarrow 0$
2: INIT()
3: **while** $B > 0$ **do**
4: $i_0 \leftarrow$ CHOOSE()
5: Present r_{i_0} for a tagger to tag
6: The tagger completes a post task on r_{i_0}
7: UPDATE()
8: $x[i_0] \leftarrow x[i_0] + 1$, $B \leftarrow B - 1$
return $\vec{x}$

Implementing *CHOOSE()*

- **F**ree **C**hoice (FC)
  - Users freely decide which resource they want to tag.
- **R**ound **R**obin (RR)
  - The resources have even chance to get posts.
- **F**ewest **P**ost First (FP)
  - Prioritize Under-Tagged Resources
- **M**ost **U**nstable First (MU)
  - Resources with unstable rfds' need more posts
  - Window size
- Hybrid (FP-MU)

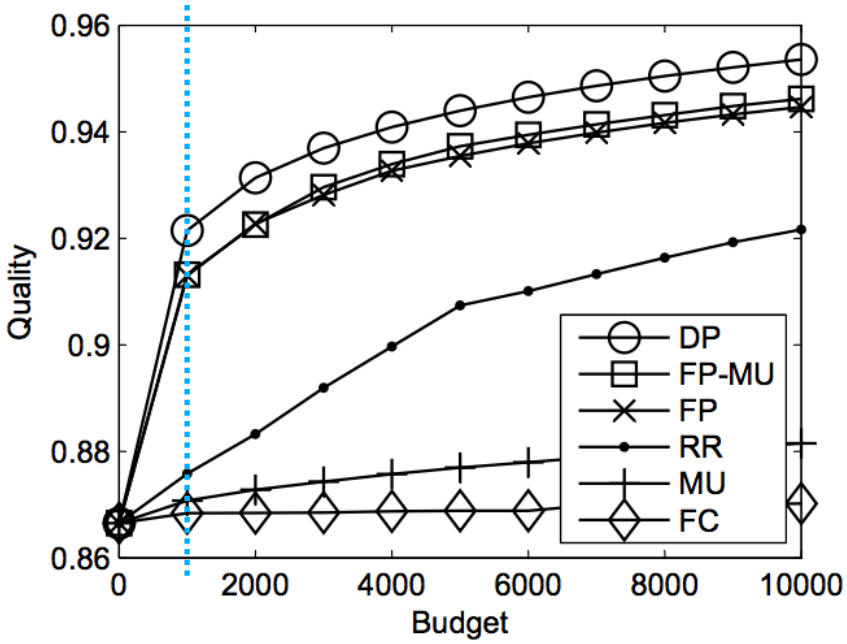
Setup

- ***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***)



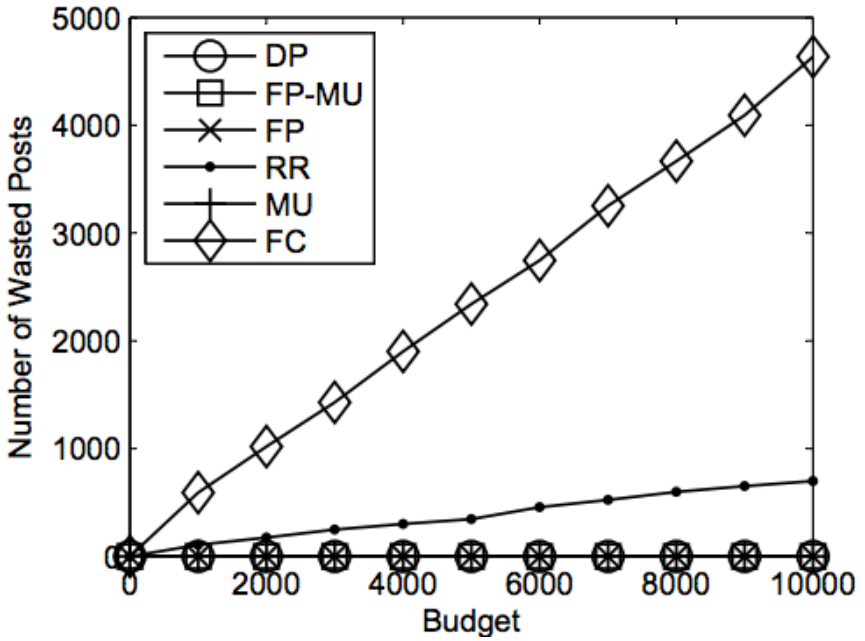
Quality vs. Budget

- 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

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



Top-10 Similar Sites (Cont'd)

- 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

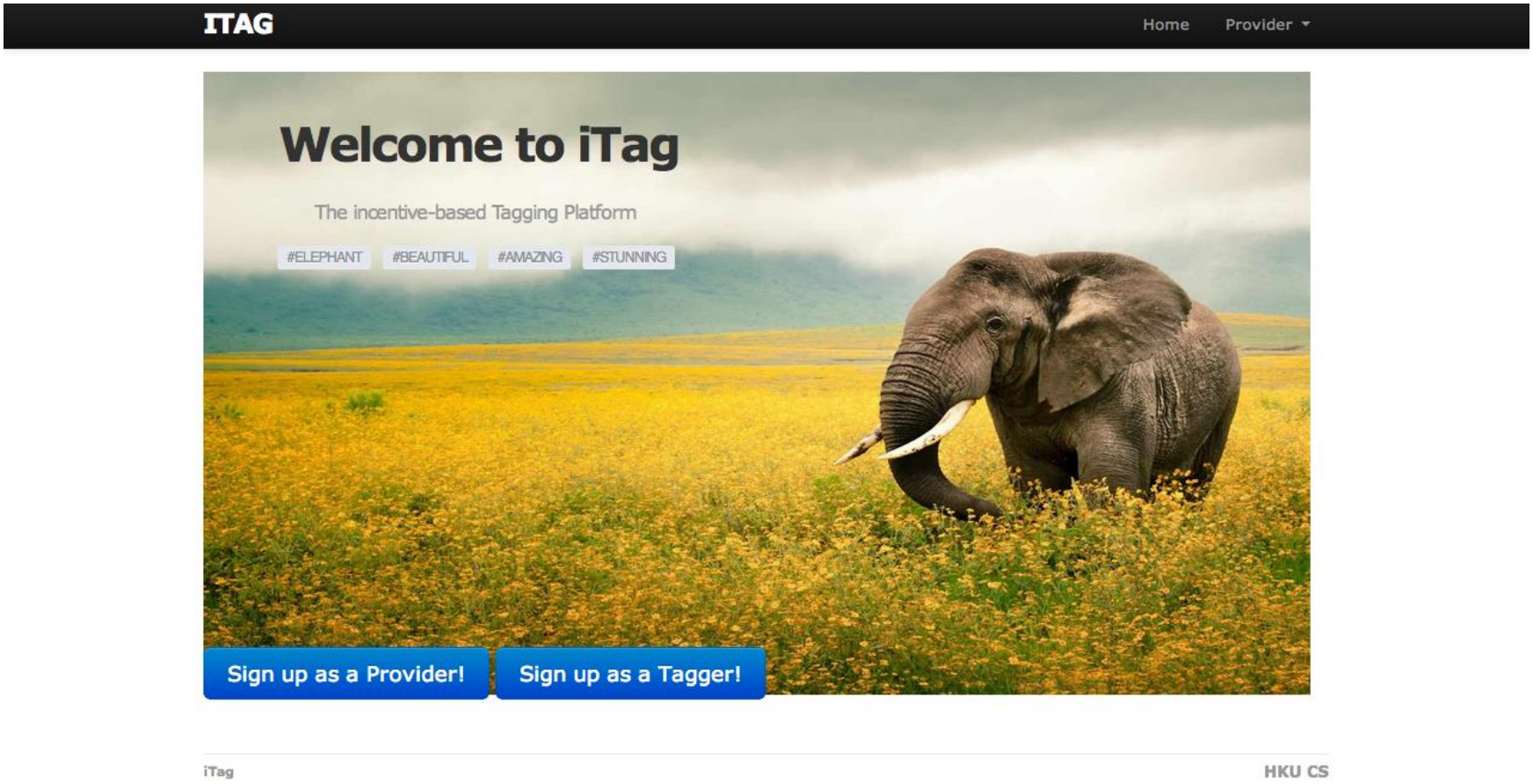
Rank	$3 + 4(\text{by } \text{FC}_{10000}) = 7$ posts
1	physicsclassroom.com ✔
2	hyperphysics.phy-astr.gsu.edu/hbas/hframe.html ✔
3	practicalphysics.org ✔
4	hyperphysics.phy-astr.gsu.edu/hbas/hframe.html ✔
5	javaranch.com
6	robocode.sourceforge.net
7	www.onjava.com
8	java.sun.com
9	jguru.com
10	kickjava.com

Top-10 Similar Sites (Cont'd)

- 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

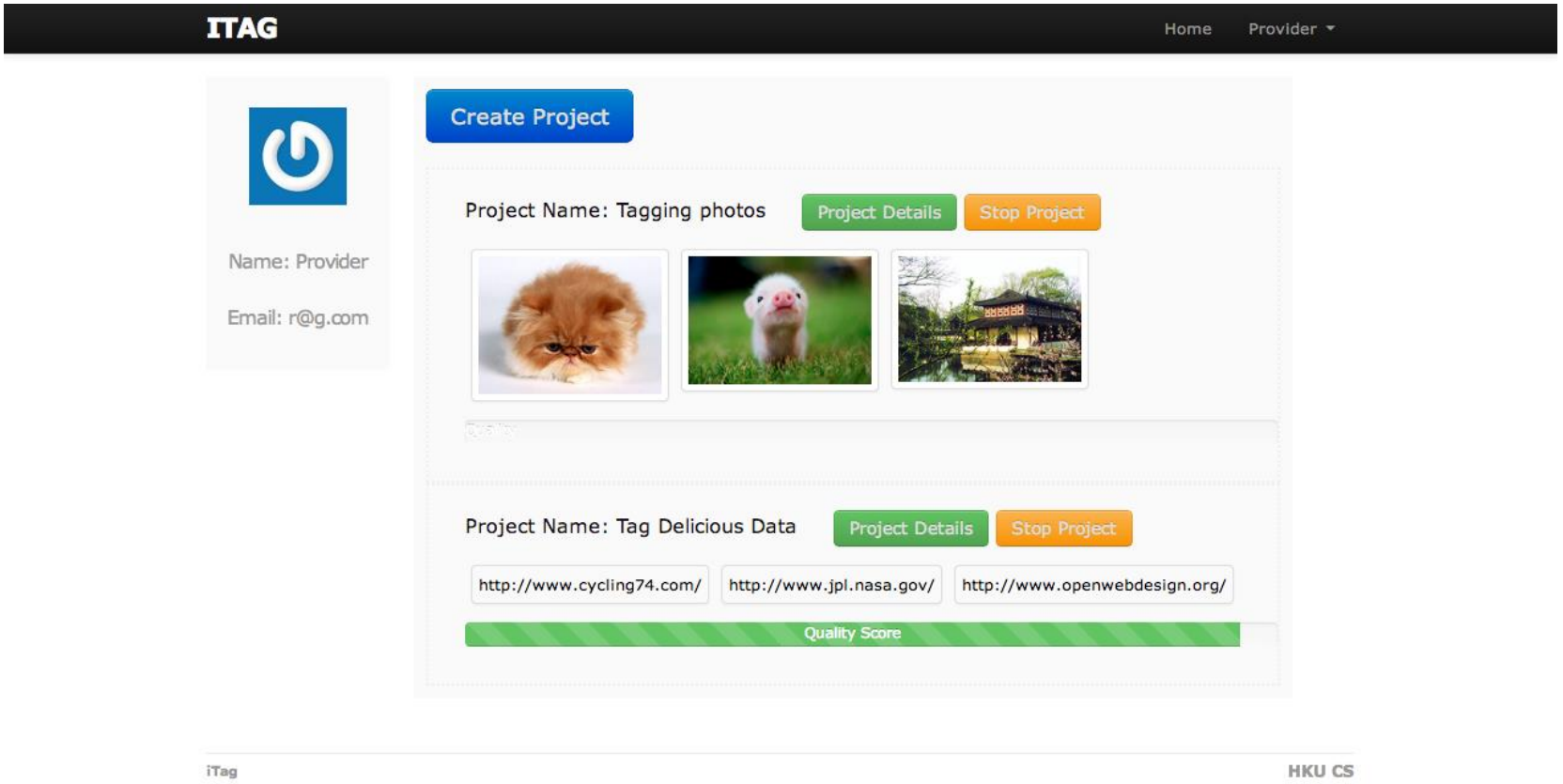
Rank	$3 + 11$ (by FP_{10000}) $= 14$ posts
1	physicsclassroom.com
2	practicalphysics.org
3	hyperphysics.phy-astr. gsu.edu/hbas/hframe.html
4	hyperphysics.phy-astr. gsu.edu/hbas/hph.html
5	physicsforums.com
6	upscale.utoronto.ca/ GeneralInterest/Harrison/Flash
7	bethe.cornell.edu
8	grc.nasa.gov/WWW/K-12/Numbers/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

iTag (1) [ICDE'14 demo]



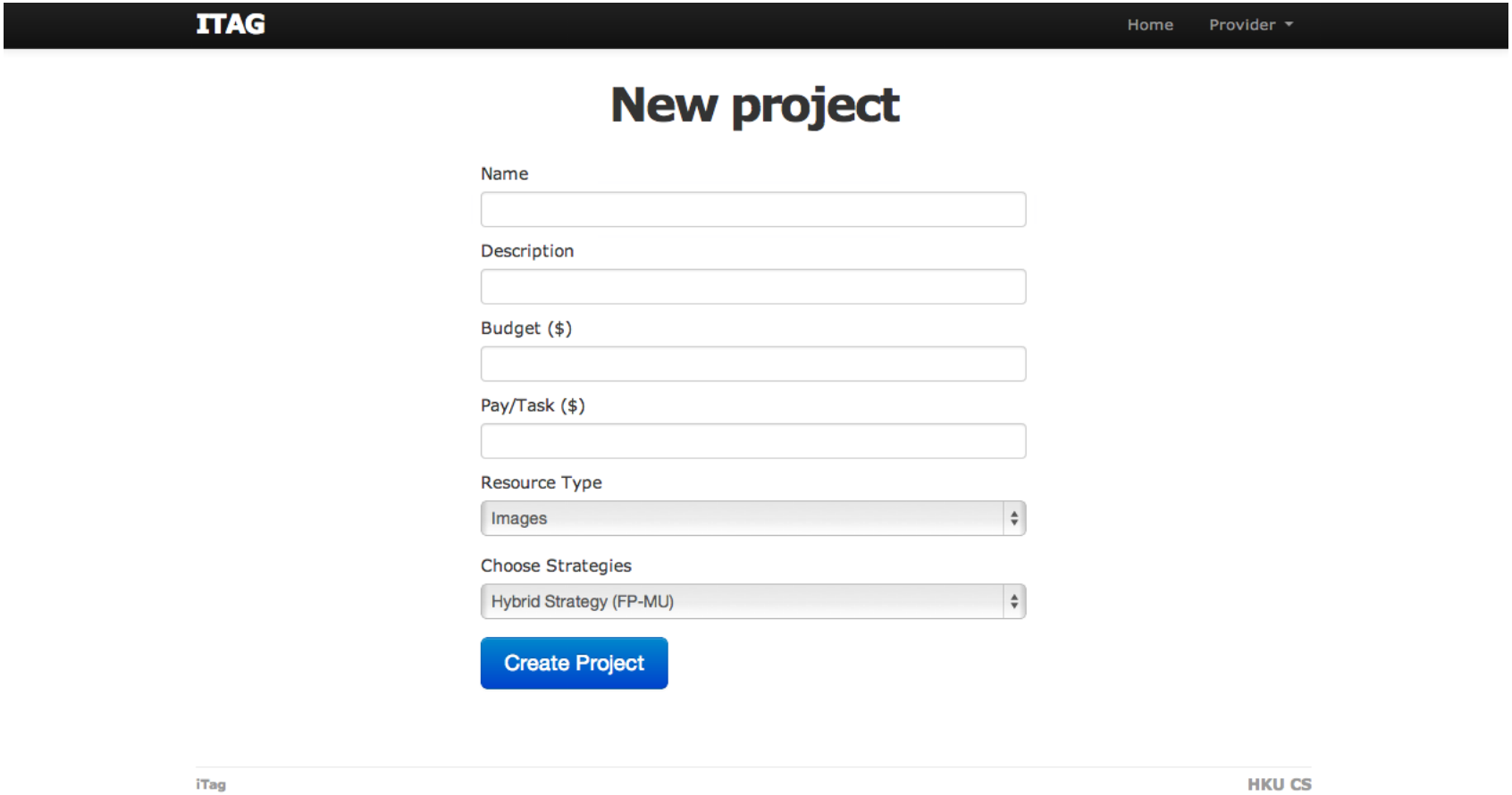
- Welcome page for provider

iTag (2)



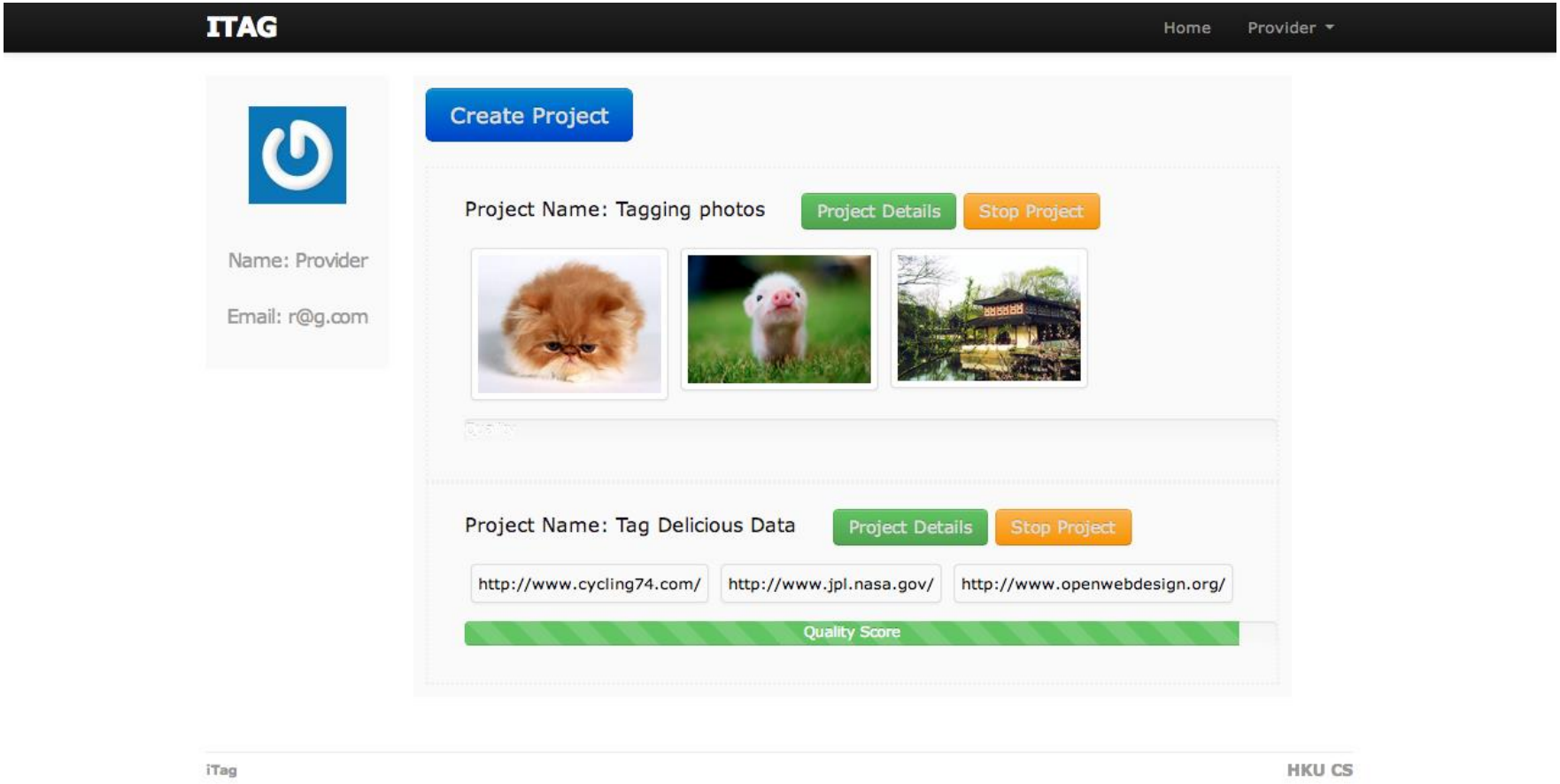
- Existing projects created by the Provider

iTag (3)



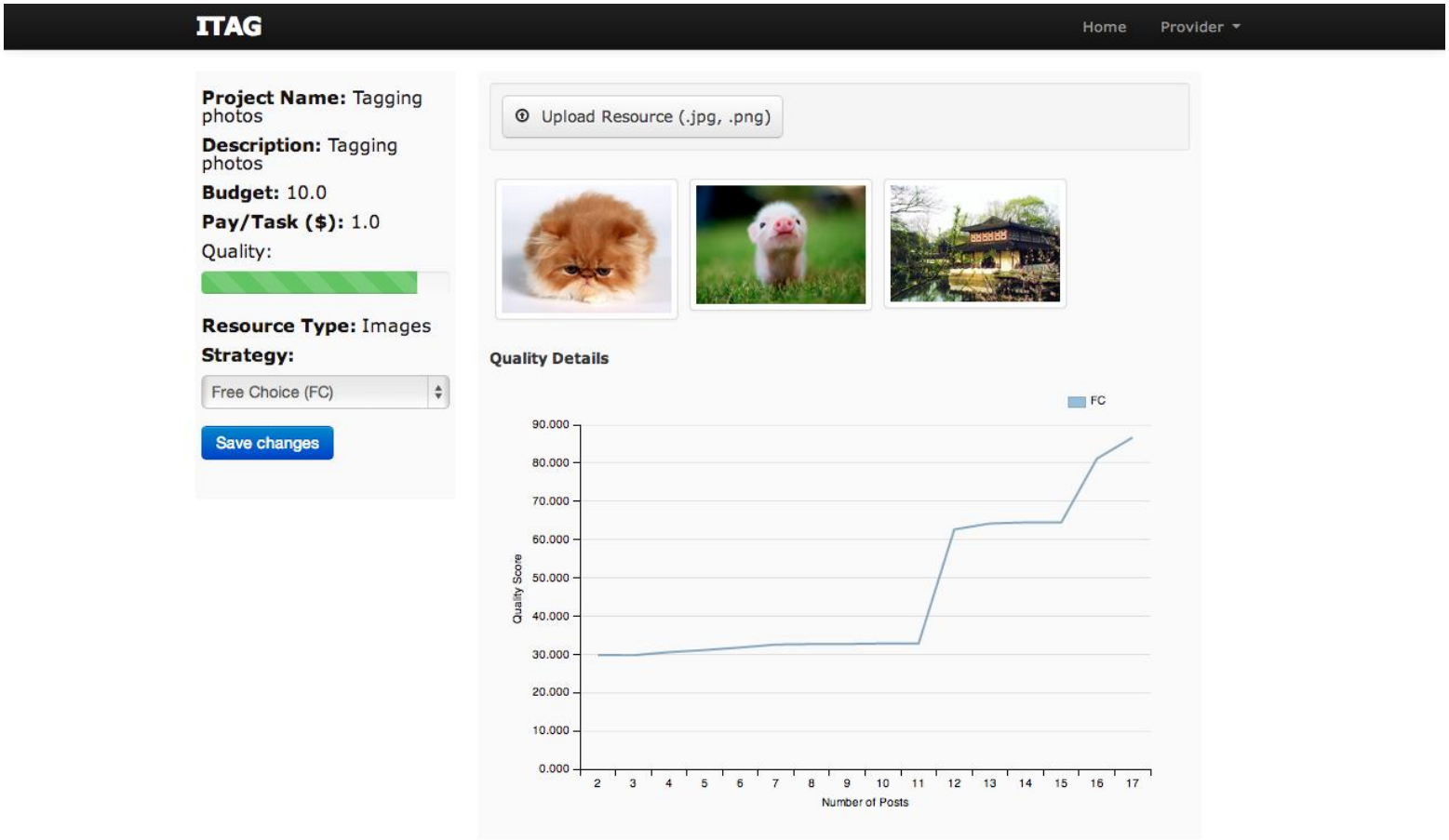
- Create a new project

iTag (4)



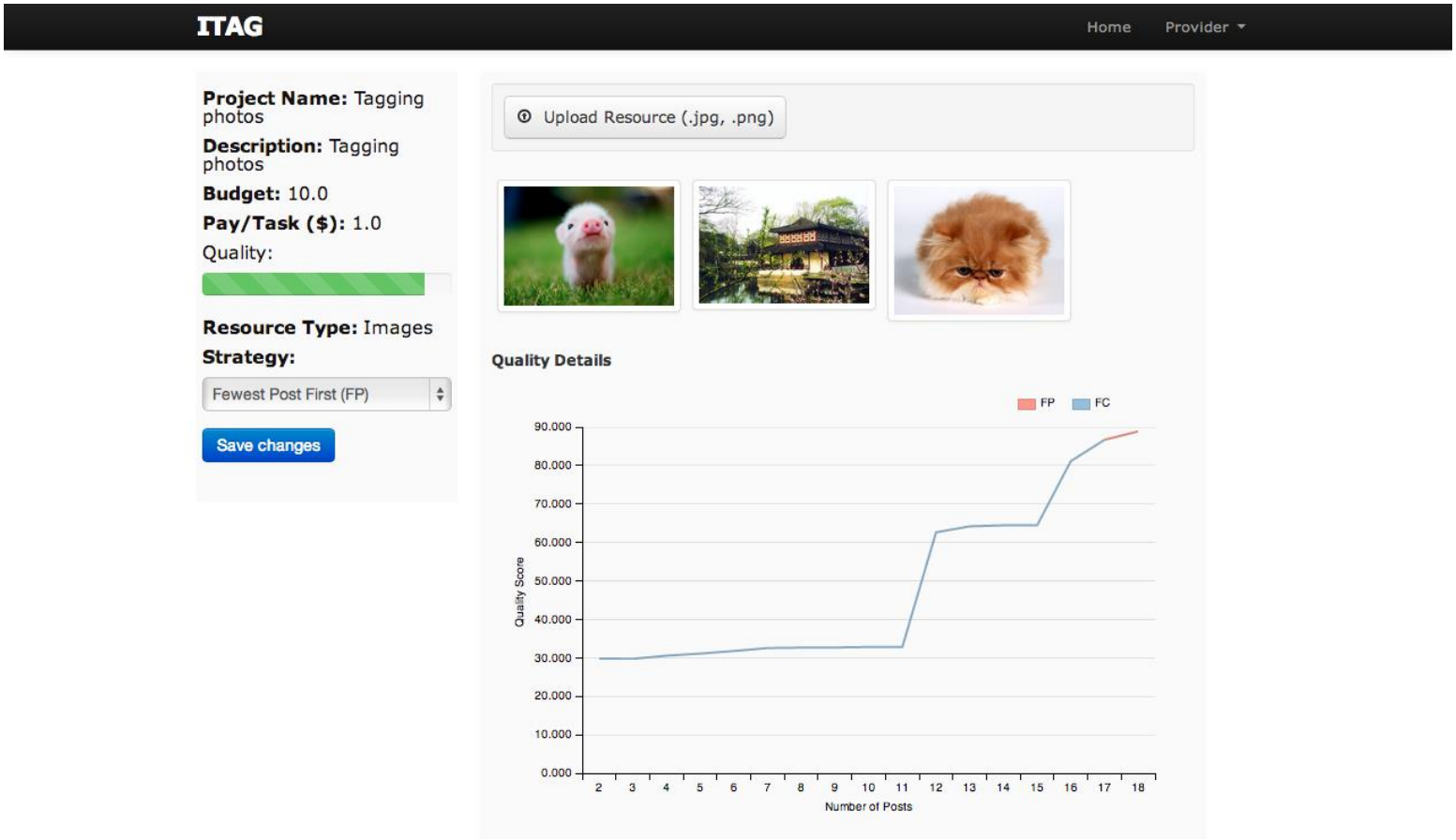
- A previously created project

iTag (5)



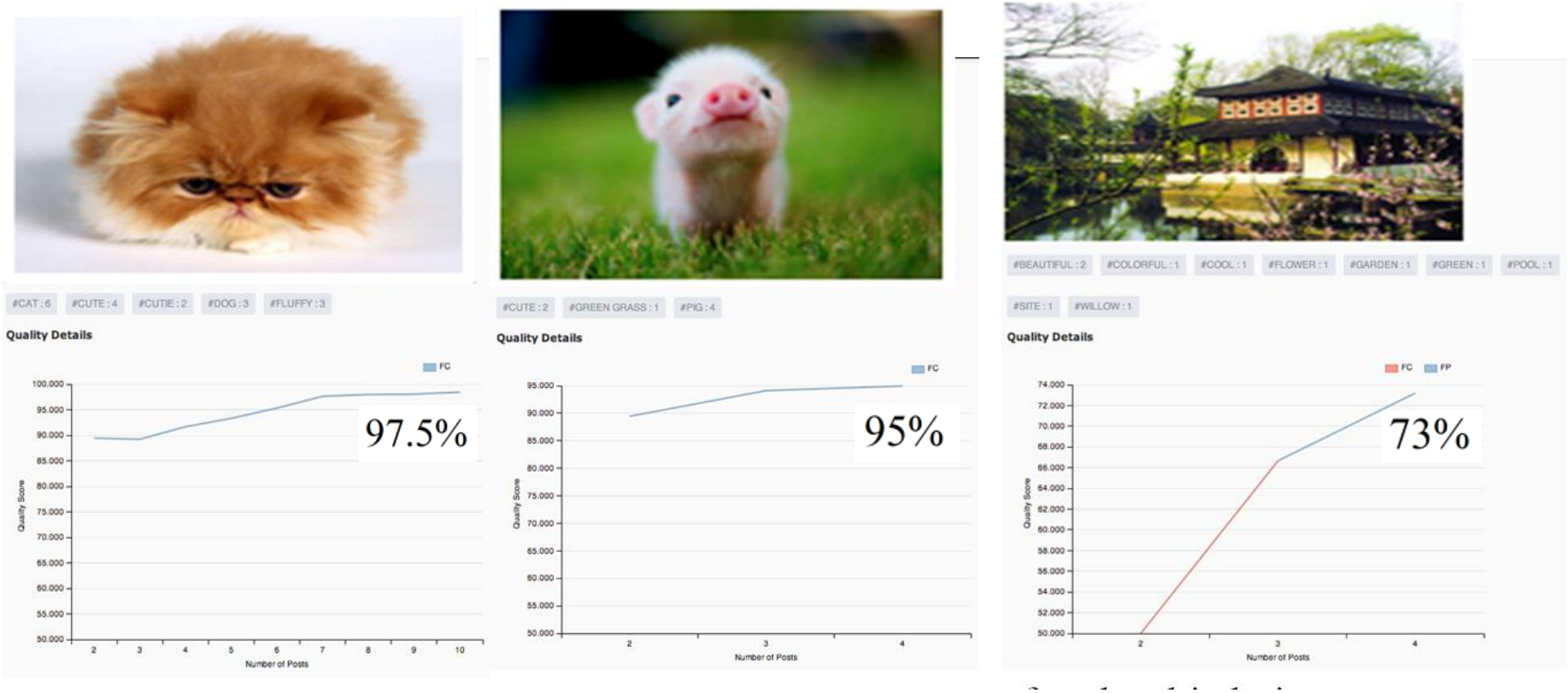
- The quality score by FC

iTag (6)



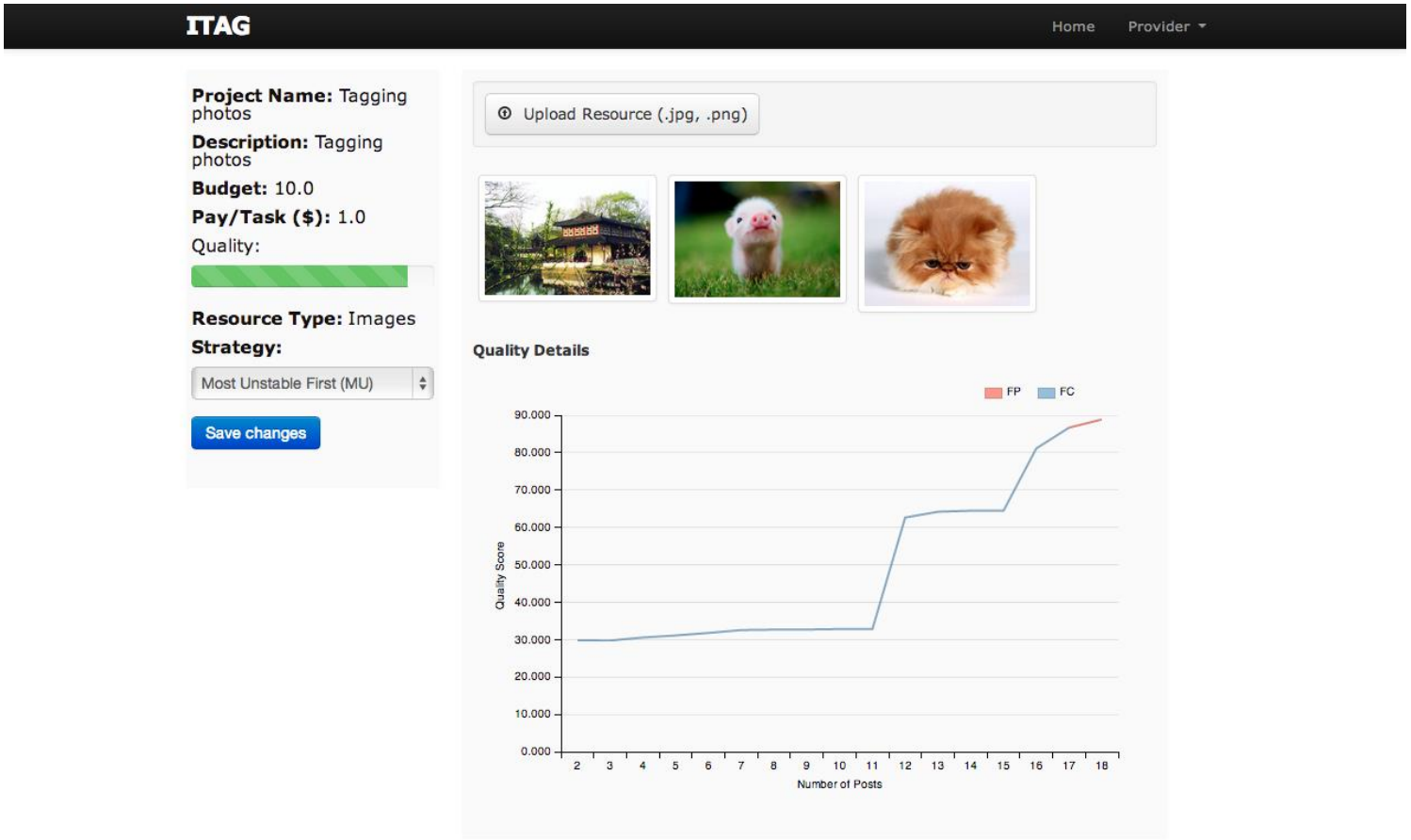
- The quality score with different strategies

iTag (7)



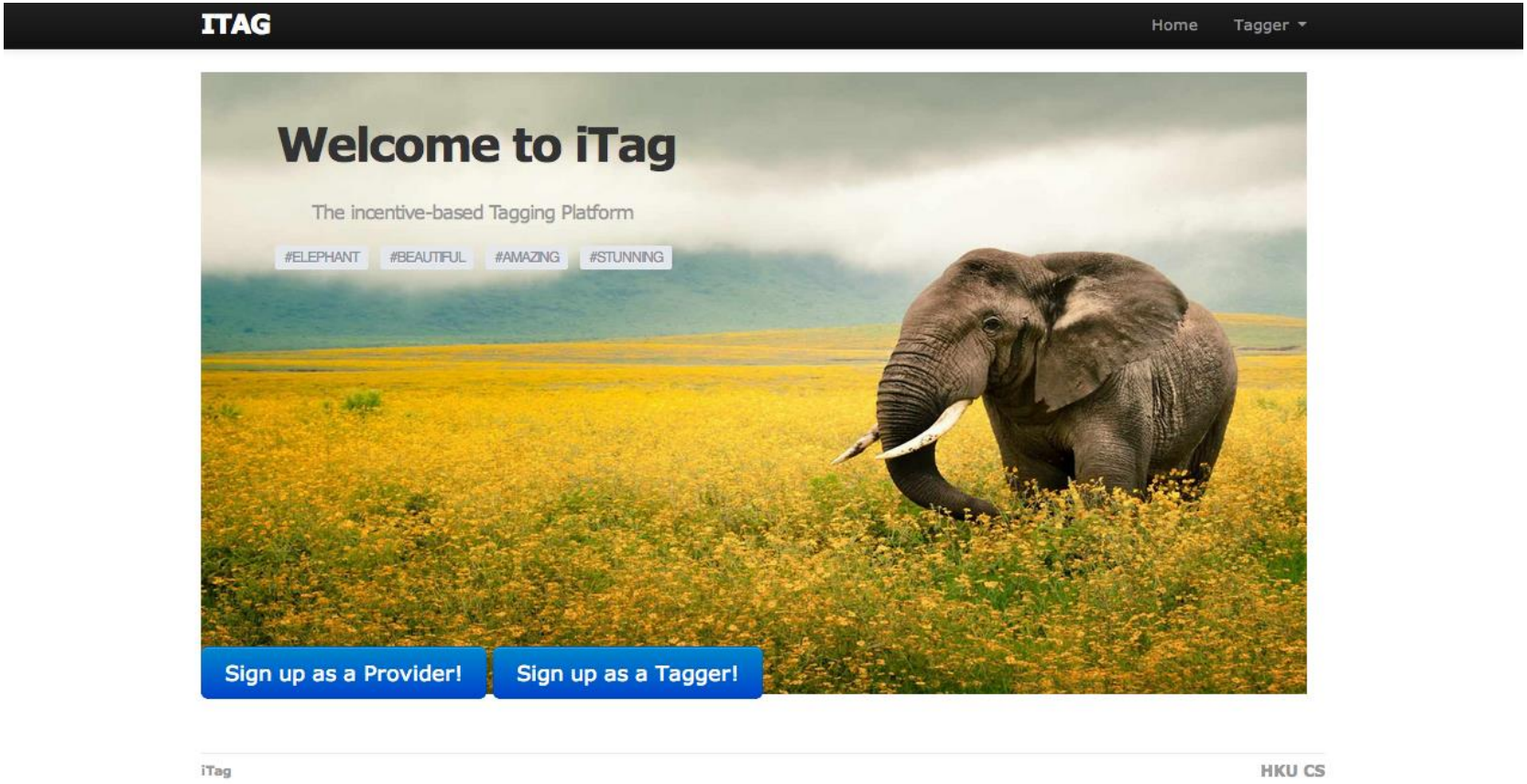
- The quality scores for three pictures

iTag (8)



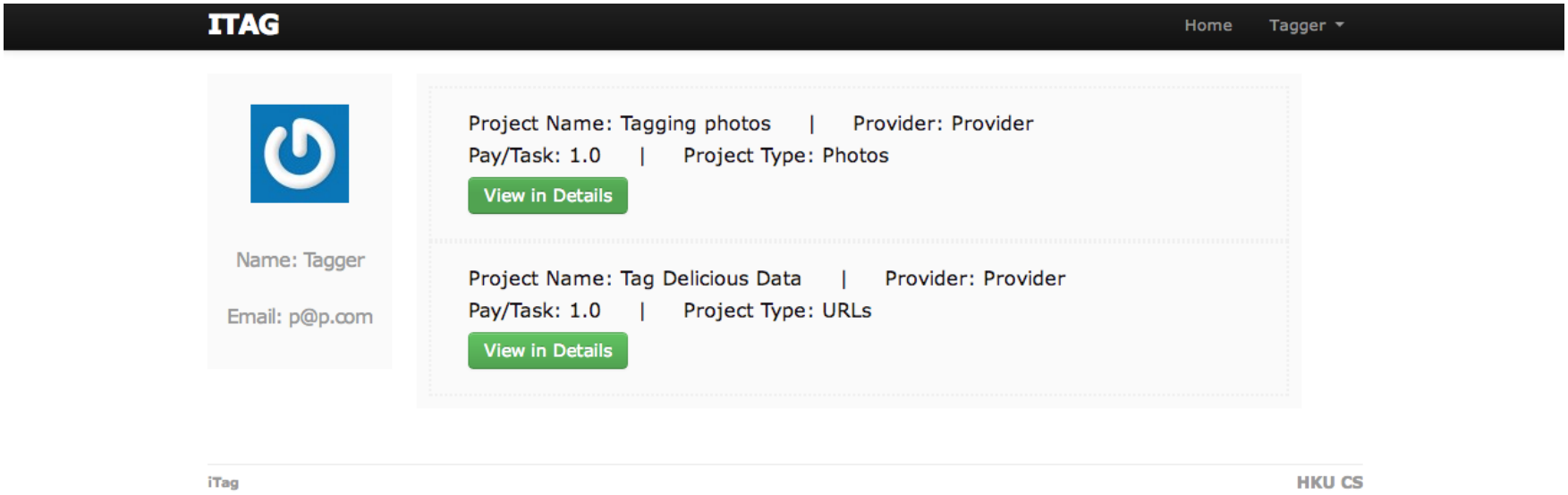
- Tagging sequence for MU

iTag (9)



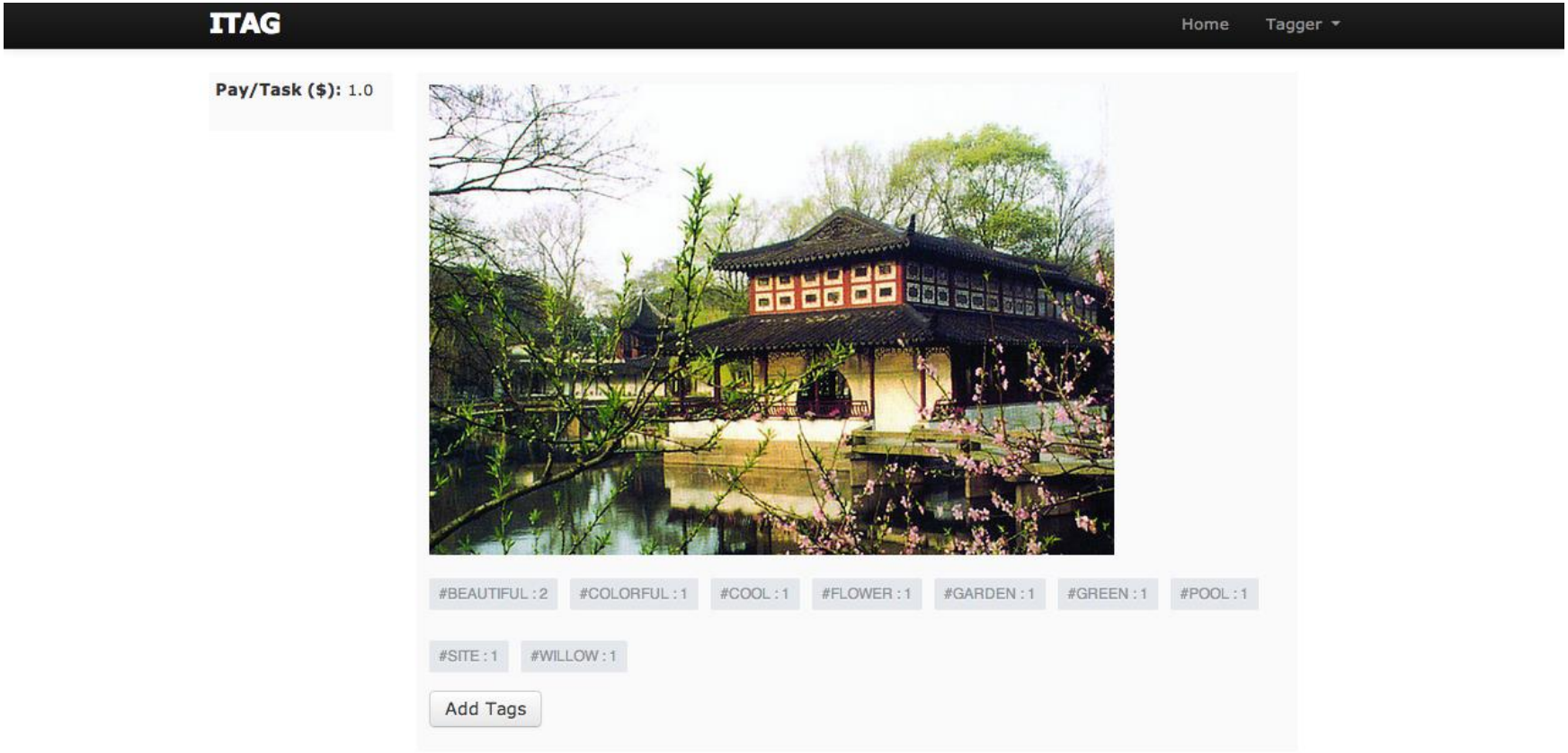
- Welcome page for Tagger

iTag (10)



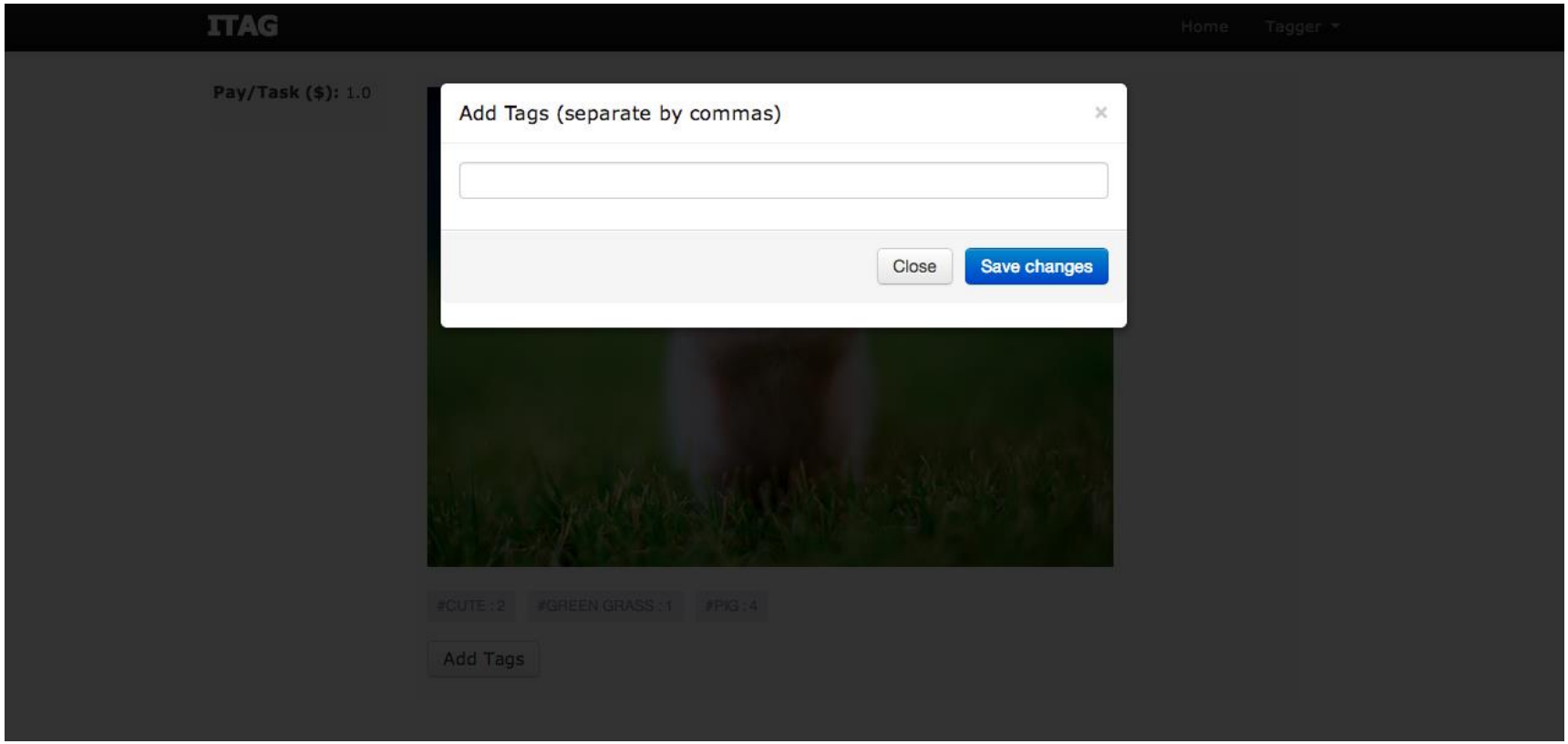
- Select a project to tag

iTag (11)



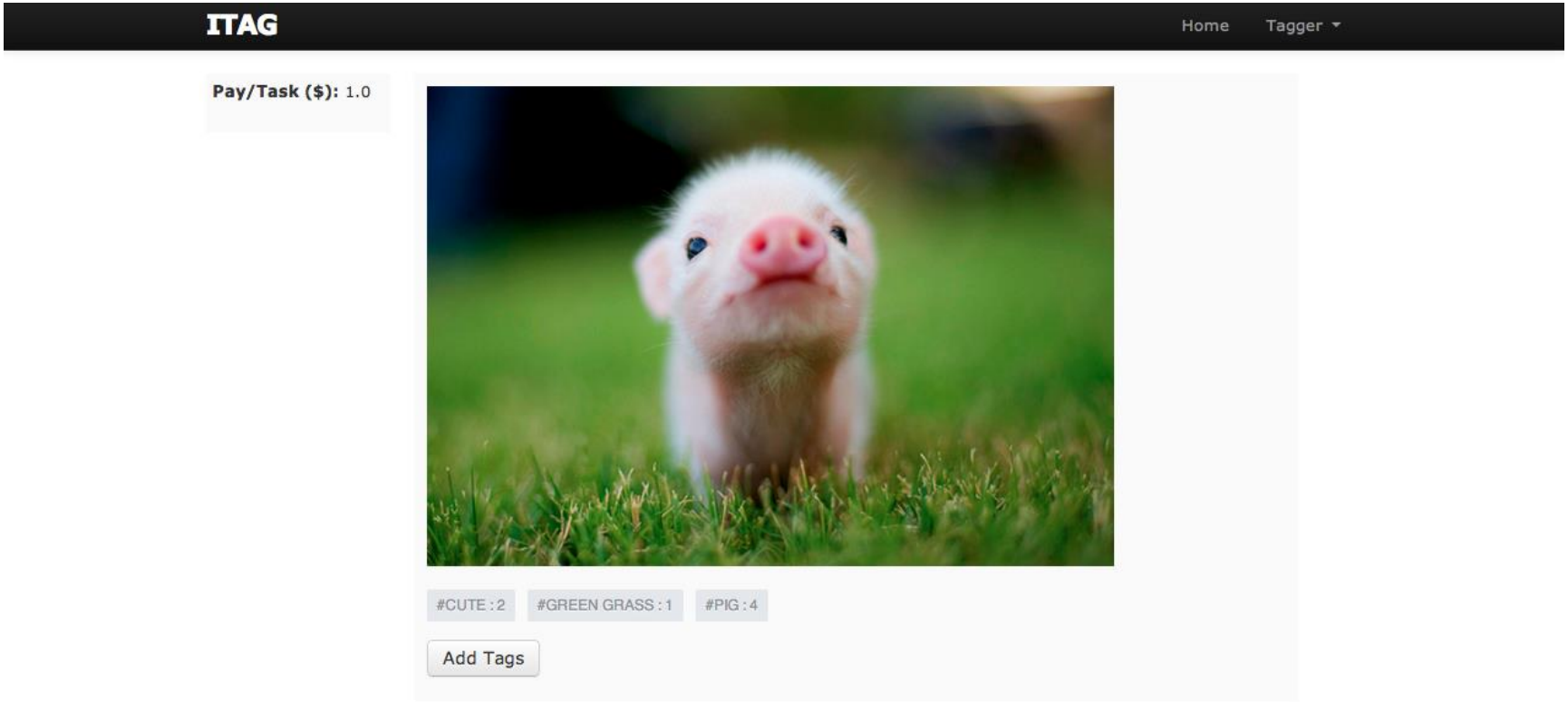
- A picture with existing tags

iTag (12)



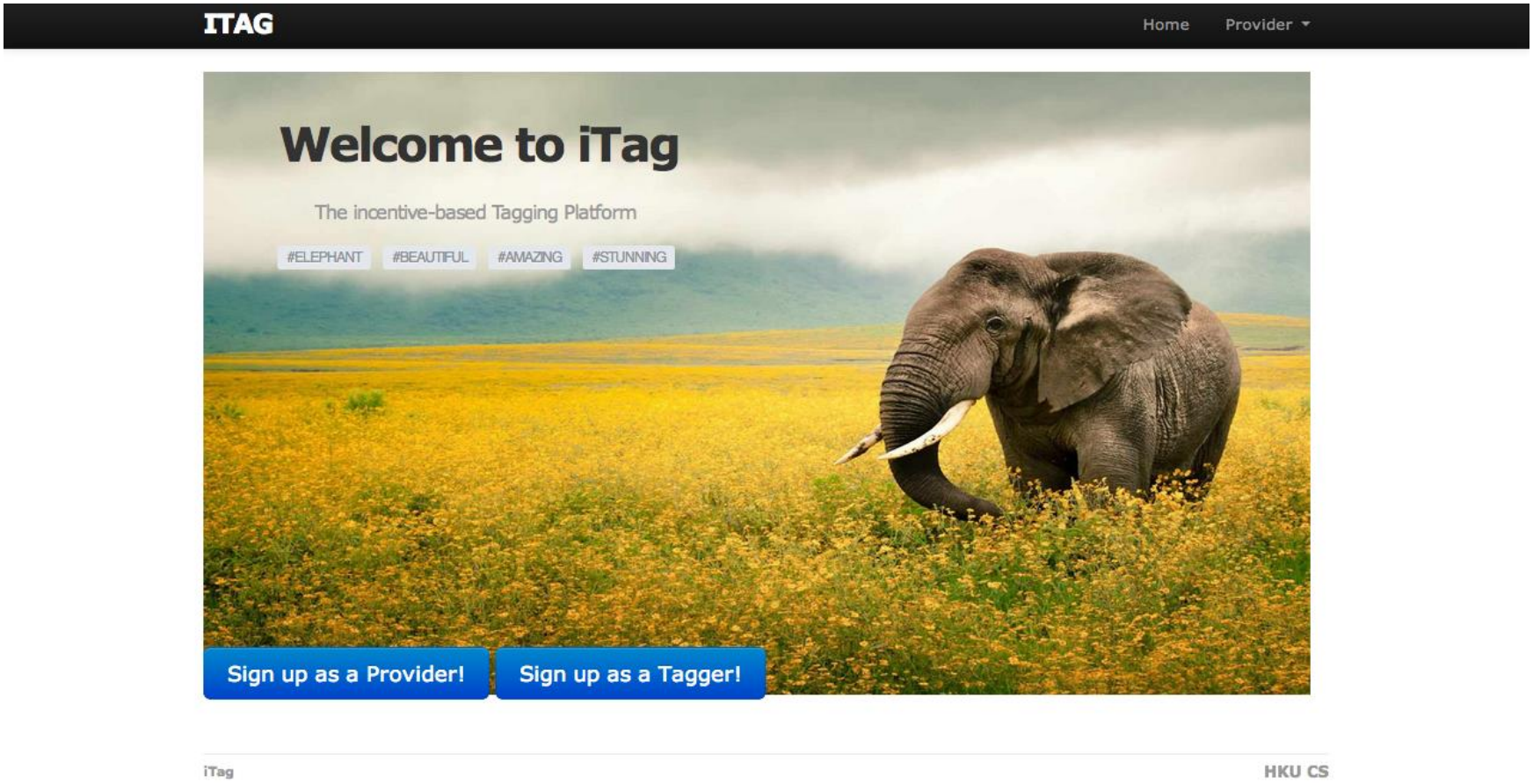
- Add your tags for the picture

iTag (13)



- Another picture for you to tag

iTag (14)

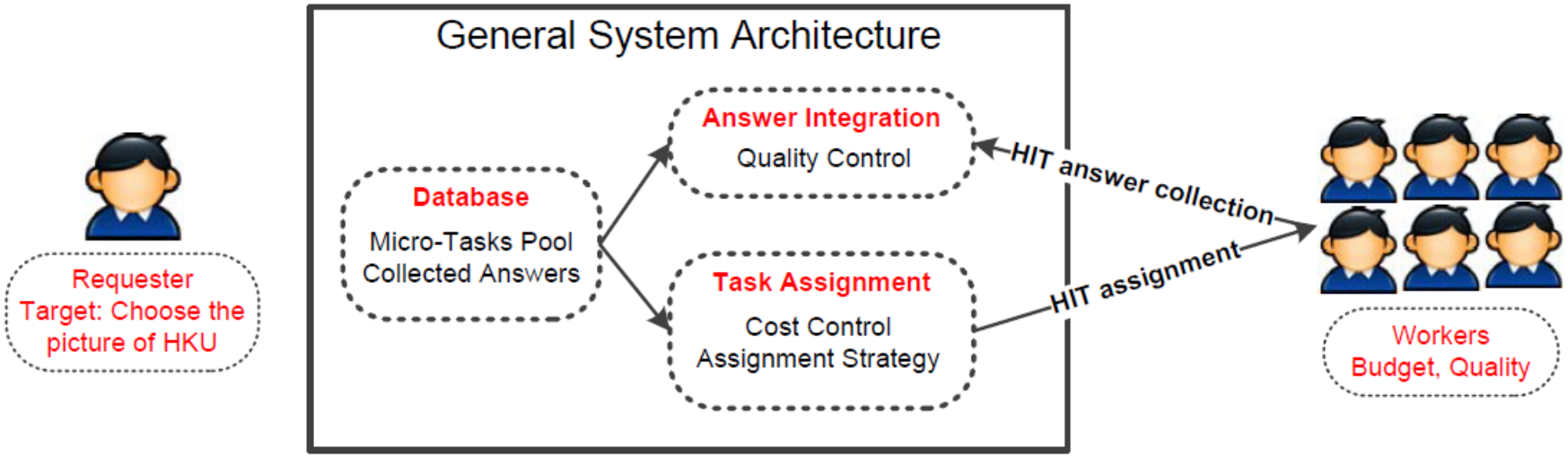

ITAG
Home
Provider
Welcome to iTag
The incentive-based Tagging Platform
#ELEPHANT
#BEAUTIFUL
#AMAZING
#STUNNING
Sign up as a Provider!
Sign up as a Tagger!
iTag
HKU CS

Summary

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



- **1.Answer Integration:**

  How to integrate answers from workers ?

- **2.Task Assignment:**

  Which tasks are chosen to assign to a worker ?

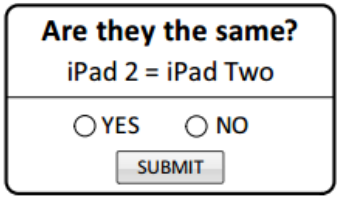
- **3.Database:**

  How to store crowdsourced data?

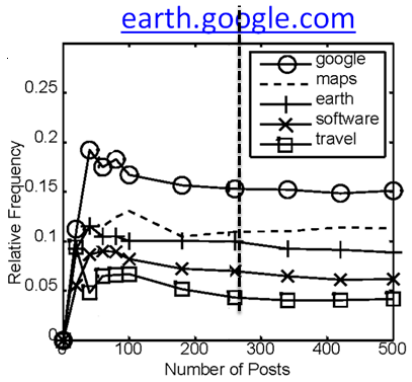
Review:Answer Integration

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

- Half Voting [CIKM'13]



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



Review:Task Assignment

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 **C**hoice, Round Robin, FPF,…

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

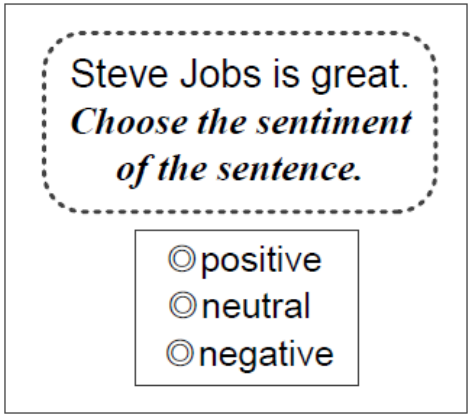
Challenge 1: Manage answer integration & task assignment

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



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

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

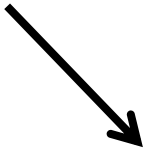
Traditional and Crowdsourced database does not provide support for handling uncertain data

Open questions

- 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

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

D. Suciu, D. Olteanu, C. R´e, and C. Koch. Probabilistic Databases . Morgan-Claypool, 2011.

Types of Uncertainty

- 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

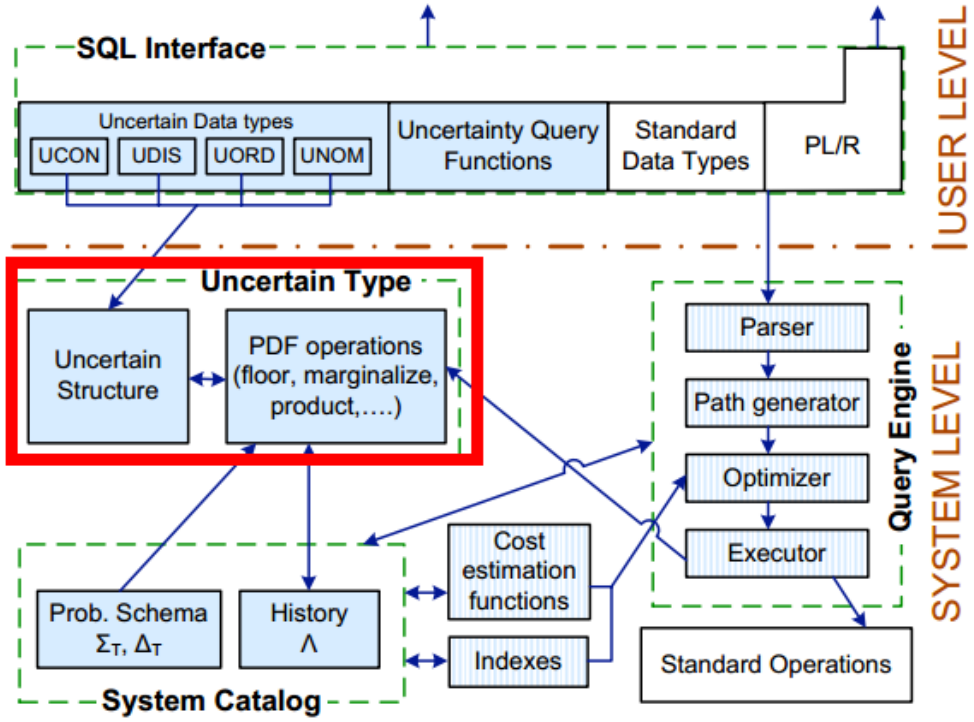
No.	sentiment		
1	positive 0.66	Neutral 0.32	Negative 0.02

D. Suciu, D. Olteanu, C. R´e, and C. Koch. Probabilistic Databases . Morgan-Claypool, 2011.

Orion

- 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

- **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