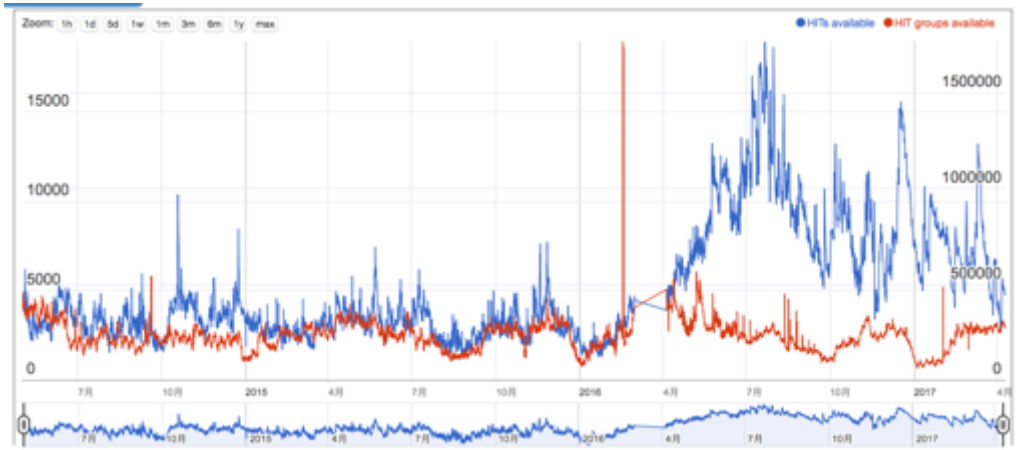
Managing the Quality of Crowdsourced Databases

Yudian Zheng
The University of Hong Kong
zhydhkcws@gmail.com

Why Quality Control?

- **Huge Amount** of Crowdsourced Data



amazon mechanical turk
beta
Artificial Artificial Intelligence

Statistics in AMT:
Over 500K workers
Over 1M tasks

- **Inevitable noise & error**

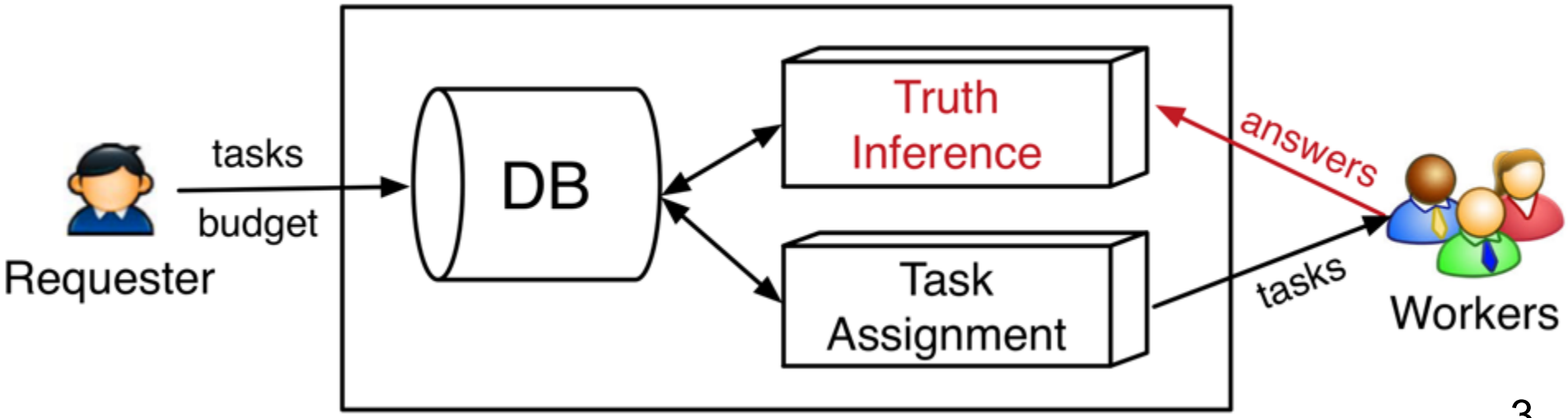
- **Goal: Obtain reliable information in Crowdsourced Data**

Crowdsourcing Workflow

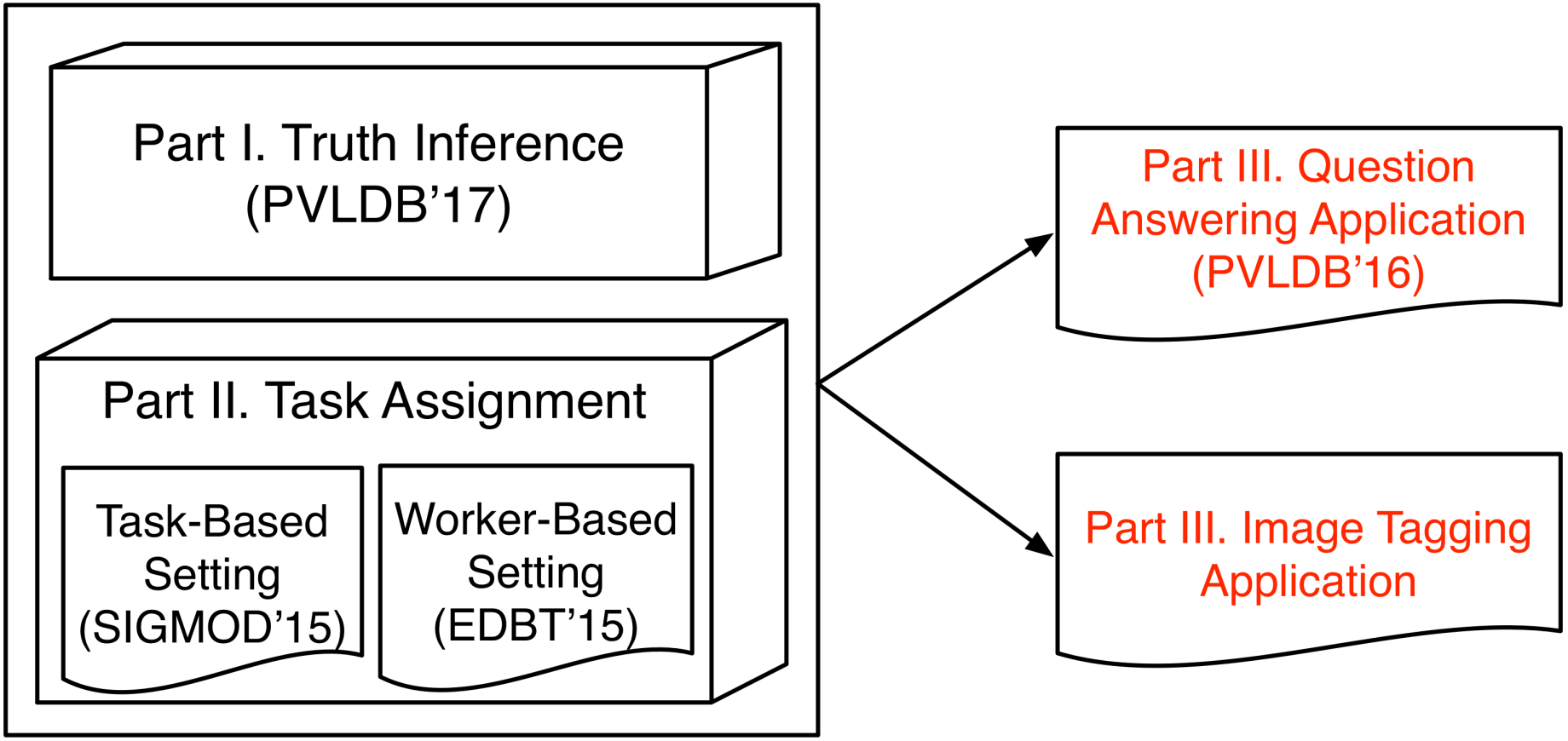
- **Requester** deploys tasks and budget on crowdsourcing platform (e.g., AMT)
- **Workers** interact with platform (**2 phases**)

  Task Assignment: How to assign suitable tasks to appropriate workers?

  Truth Inference: How to aggregate workers' answers and infer the truth of each task?



Overview of the Defense



Outline of the Defense

- **Part I. Truth Inference**
  - **Thorough Analysis and Comparisons (Zheng et al. PVLDB'17)**

- **Part II. Task Assignment**
  - **Task-Based Setting (Zheng et al. SIGMOD'15)**
  - **Worker-Based Setting (Zheng et al. EDBT'15)**

- **Part III. Combinations**
  - **Question-Answering Application (Zheng et al. PVLDB'16)**
  - **Image Tagging Application**

Part I. Truth Inference

- **An Example Task**

Where was ACM SIGMOD 2017 held ?

A. Raleigh 

B. Chicago 

I think
A. Raleigh !

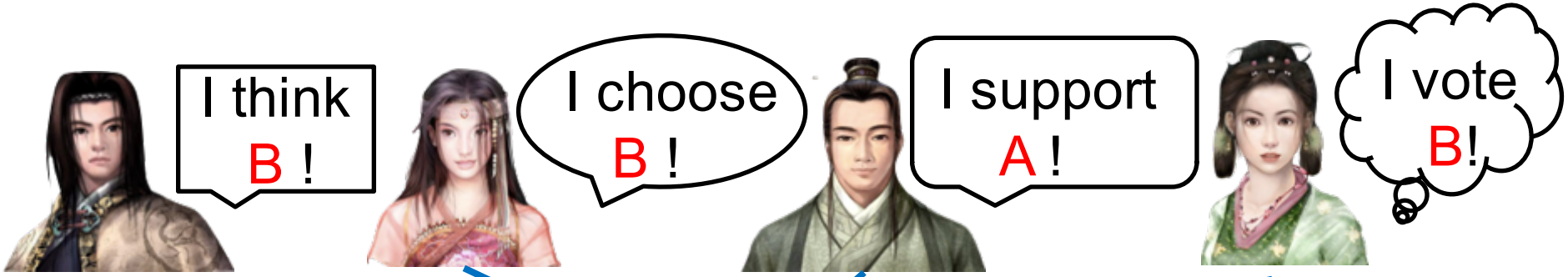
Principle: Redundancy

- **Collect Answers from Multiple Workers**

Where was ACM SIGMOD 2017 held ?

A. Raleigh

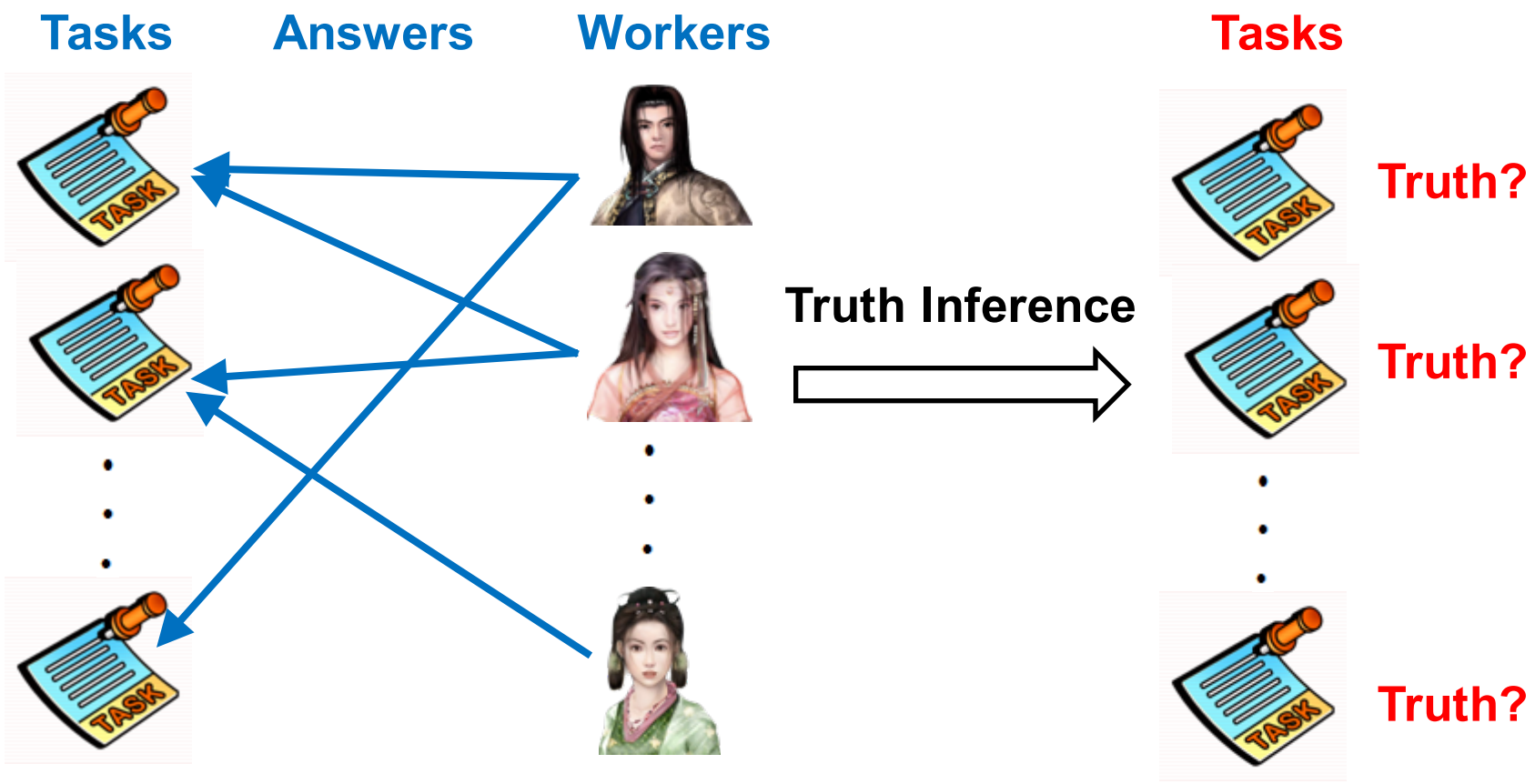
B. Chicago



What is the truth of the task ?

Truth Inference Definition

Given different tasks' answers collected from workers, the target is to infer the truth of each task.



A Simple Solution

- **Majority Voting**

  Take the answer that is voted by the majority (or most) of workers.

- **Limitation**

  Treat each worker equally, neglecting the diverse quality for each worker.

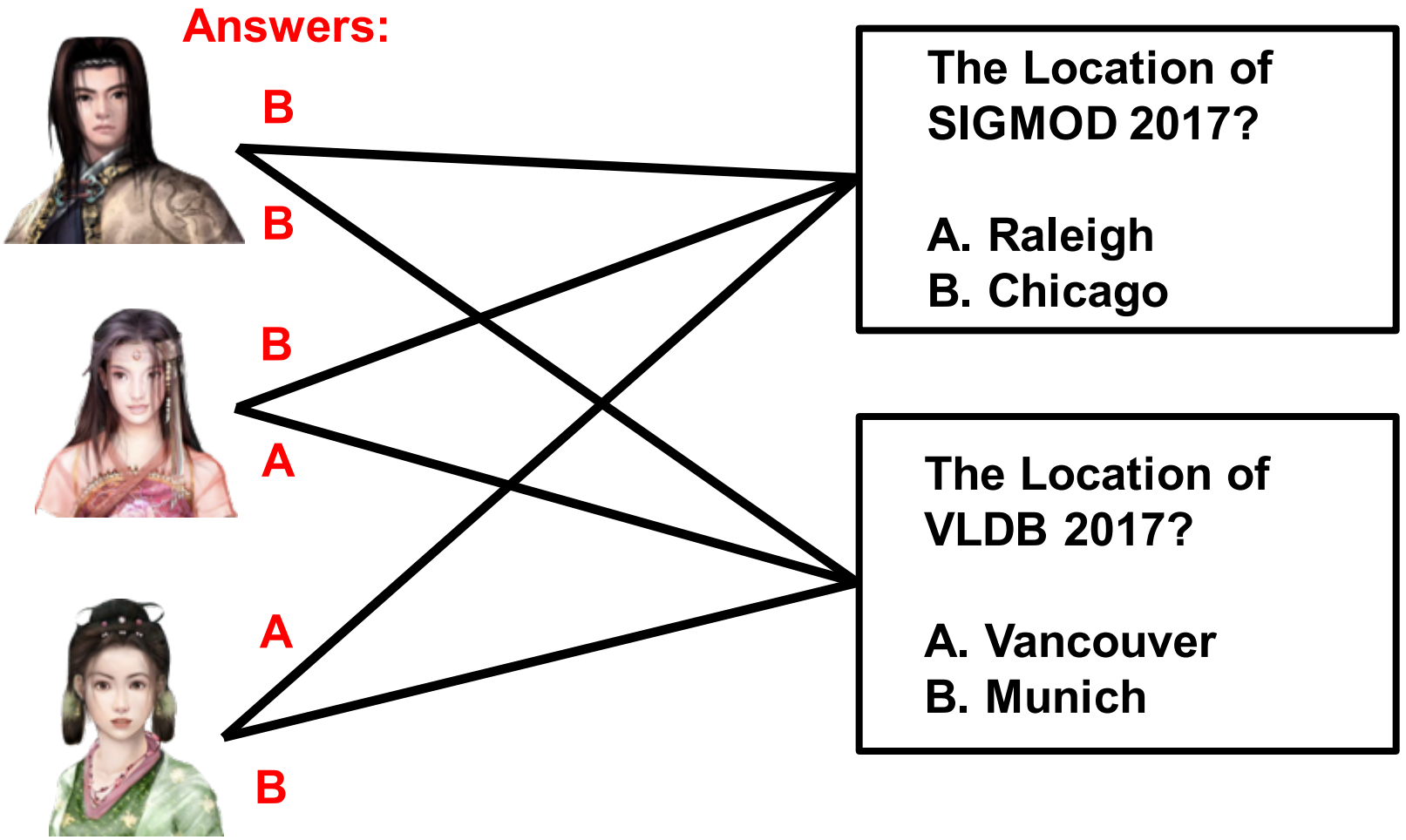
The Key to Truth Inference

- The key is to know **each worker's quality**

Suppose quality of 4 workers are known

How to know worker's quality ?

- Idea: Compute each worker's quality by considering the workers' answers for all tasks



Existing works

- **Classic Method**

D&S [Dawid and Skene. JRSS 1979]

- **Recent Methods**

(1) Machine Learning Community:

GLAD [Whitehill et al. NIPS09], Minimax [Zhou et al. NIPS12], BCC [Kim et al. AISTATS12], LFC [Raykar et al. JLMR10], KOS [Karger et al. NIPS11], VI-BP [Liu et al. NIPS12], VI-MF [Liu et al. NIPS12], LFC_N [Raykar et al. JLMR10]

(2) Database Community:

CATD [Li et al. VLDB14], PM [Li et al. SIGMOD14], iCrowd [Fan et al. SIGMOD15], DOCS [Zheng et al. VLDB17]

(3) Data Mining Community:

ZC [Demartini et al. WWW12], Multi [Welinder et al. NIPS 2010], CBCC [Venanzi et al. WWW14]

Unified Framework in Existing Works (Zheng et al. PVLDB'17)

- **Input:** Workers' answers for all tasks

- **Algorithm Framework:**

Initialize Quality for each worker
While (not converged) {
    Quality for each worker ⇒ Truth for each task ;
    Truth for each task ⇒ Quality for each worker ;
}

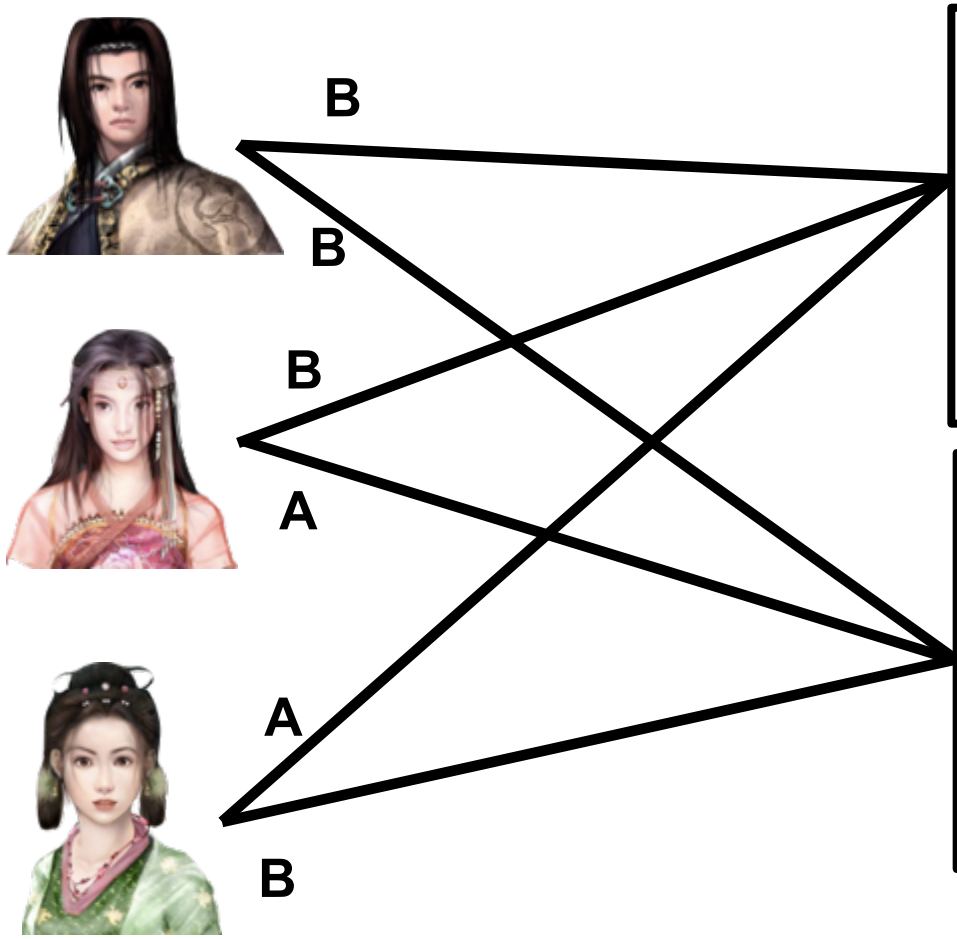
- **Output:** Quality for each worker **and** Truth for each task

Inherent Relationship 1

- 1. Quality for each worker ⟹ Truth for each task

Quality:

(Estimated) Truth:



Location of SIGMOD 2017?

A. Raleigh (1.0 from worker 3)

B. Chicago (1.0 + 1.0 from workers 1 & 2)

Location of VLDB 2017?

A. Vancouver (1.0 from worker 2)

B. Munich (1.0 + 1.0 from workers 1 & 3)

Inherent Relationship 2

- 2. **Truth for each task** ⇒ **Quality for each worker**

(Estimated) Truth:

Location of SIGMOD 2017?

A. Raleigh
B. Chicago

Location of VLDB 2017?

A. Vancouver
B. Munich

Quality:

B
B

1.0

correct: 2/2

B
A

0.5

correct: 1/2

A
B

0.5

correct: 1/2

Differences in Existing works

Tasks

- **Different Task Types**
  - ***What type of tasks they focus on ?***
  - ***E.g., single-label tasks …***

Workers

- **Different Worker Models**
  - ***How they model each worker ?***
  - ***E.g., worker probability (a value) …***

Different Tasks Types

- **Decision-Making Tasks** (yes/no task)

Is Bill Gates currently the CEO of Microsoft ?

O Yes O No

e.g., Demartini et al. WWW12, Whitehill et al. NIPS09, Kim et al. AISTATS12, Venanzi et al. WWW14, Raykar et al. JLMR10

- **Single-Label Tasks** (multiple choices)

Identify the sentiment of the tweet: ……

O Pos O Neu O Neg

e.g., Li et al. VLDB14, Li et al. SIGMOD14, Demartini et al. WWW12, Whitehill et al. NIPS09, Kim et al. AISTATS12

- **Numeric Tasks** (answer with numeric values)

What is the height for Mount Everest ?

[ ] m

e.g., Li et al. VLDB14, Li et al. SIGMOD14

Different Worker Models

- **Worker Probability**: a value $p \in [0,1]$

  The probability that the worker answers tasks correctly
  e.g., a worker answers 8 over 10 tasks correctly, then the worker probability is 0.8.

  e.g., Demartini et al. WWW12, Whitehill et al. NIPS09

- **Confidence Interval**: a range $[p-\varepsilon, p+\varepsilon]$

  ε **is related to the number of tasks answered**
  => the more answers collected, the smaller ε **is.**
  e.g., two workers answer 8 over 10 tasks and 40 over 50 tasks correctly, then the latter worker has a smaller ε.

  e.g., Li et al. VLDB14

Different Worker Models (cont'd)

- **Confusion Matrix**: a matrix

Capture a worker's answer for different choices given a specific truth

	Pos	*Neu*	*Neg*
Pos	0.6	0.2	0.2
Neu	0.3	0.6	0.1
Neg	0.1	0.1	0.8

Given that the truth of a task is "Neu", the probability that the worker answers "Pos" is 0.3.

e.g., Kim et al. AISTATS12, Venanzi et al. WWW14

- **Bias τ & Variance σ : numerical task**

Answer follows Gaussian distribution: $ans \sim N(t+\tau, \sigma)$

e.g., Raykar et al. JLMR10

Summary of Truth Inference Methods

Method	Task Type	Worker Model
Majority Voting	Decision-Making Task, Single-Choice Task	No
Mean / Median	Numeric Task	No
ZC [Demartini et al. WWW12]	Decision-Making Task, Single-Choice Task	Worker Probability
GLAD [Whitehill et al. NIPS09]	Decision-Making Task, Single-Choice Task	Worker Probability
D&S [Dawid and Skene. JRSS 1979]	Decision-Making Task, Single-Choice Task	Confusion Matrix
Minimax [Zhou et al. NIPS12]	Decision-Making Task, Single-Choice Task	Confusion Matrix
BCC [Kim et al. AISTATS12]	Decision-Making Task, Single-Choice Task	Confusion Matrix
CBCC [Venanzi et al. WWW14]	Decision-Making Task, Single-Choice Task	Confusion Matrix
LFC [Raykar et al. JLMR10]	Decision-Making Task, Single-Choice Task	Confusion Matrix

Summary of Truth Inference Methods (cont'd)

Method	Task Type	Worker Model
PM [Li et al. SIGMOD14]	Decision-Making Task, Single-Choice Task, Numeric Task	Worker Probability
Multi [Welinder et al. NIPS 2010]	Decision-Making Task	Worker Bias, Worker Variance
KOS [Karger et al. NIPS11]	Decision-Making Task	Worker Probability
VI-BP [Liu et al. NIPS12]	Decision-Making Task	Confusion Matrix
VI-MF [Liu et al. NIPS12]	Decision-Making Task	Confusion Matrix
LFC_N [Raykar et al. JLMR10]	Numeric Task	Worker Variance
CATD [Li et al. VLDB14]	Decision-Making Task, Single-Choice Task, Numeric Task	Worker Probability, Confidence

Which method is the best ?

- "Majority Voting" if sufficient data is given (each task collects more than 20 answers);
- "D&S [Dawid and Skene JRSS 1979]" if limited data is given (a robust method);
- "Minimax [Zhou et al. NIPS12]" and "Multi [Welinder et al. NIPS 2010]" as advanced techniques.

Summary of Truth Inference (Zheng et al. PVLDB'17)

- **Detailed analysis over 17 existing truth inference methods**
- **Unified truth inference framework**
- **Comparisons over existing works (task model & worker model)**
- **Experimental analysis and open-source codes (http://i.cs.hku.hk/~ydzheng2/crowd_truth_inference/index.html)**

Outline of the Defense

- Part I. Truth Inference
  - Thorough Analysis and Comparisons (Zheng et al. PVLDB'17)

- Part II. Task Assignment
  - Task-Based Setting (Zheng et al. SIGMOD'15)
  - Worker-Based Setting (Zheng et al. EDBT'15)

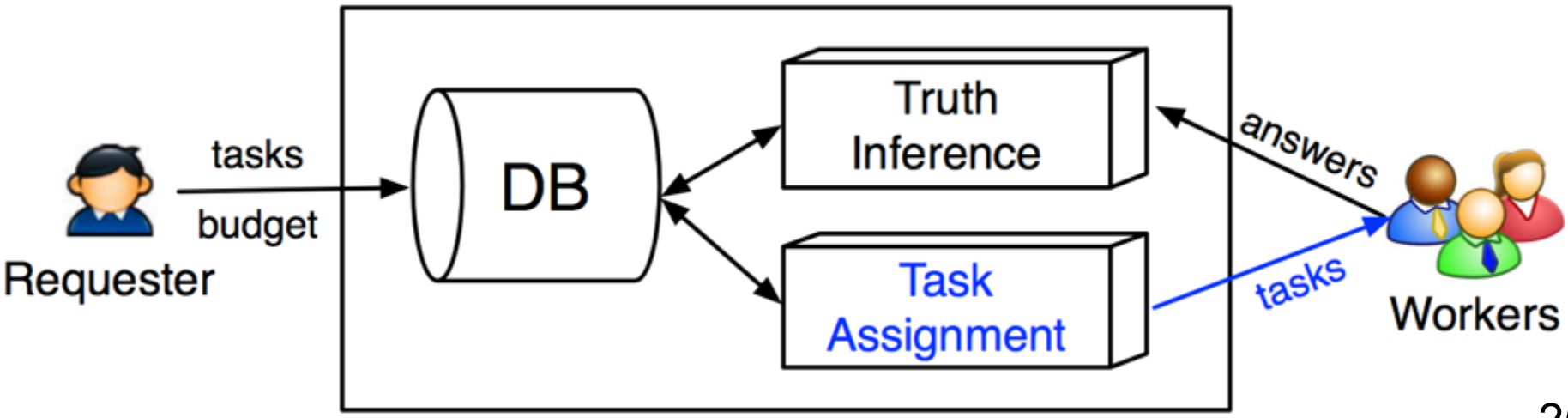
- Part III. Combinations
  - Question-Answering Application (Zheng et al. PVLDB'16)
  - Image Tagging Application

Part II. Task Assignment

- Requester deploys tasks and budget on crowdsourcing platform (e.g., Amazon Mechanical Turk)
- Workers interact with platform (2 phases)

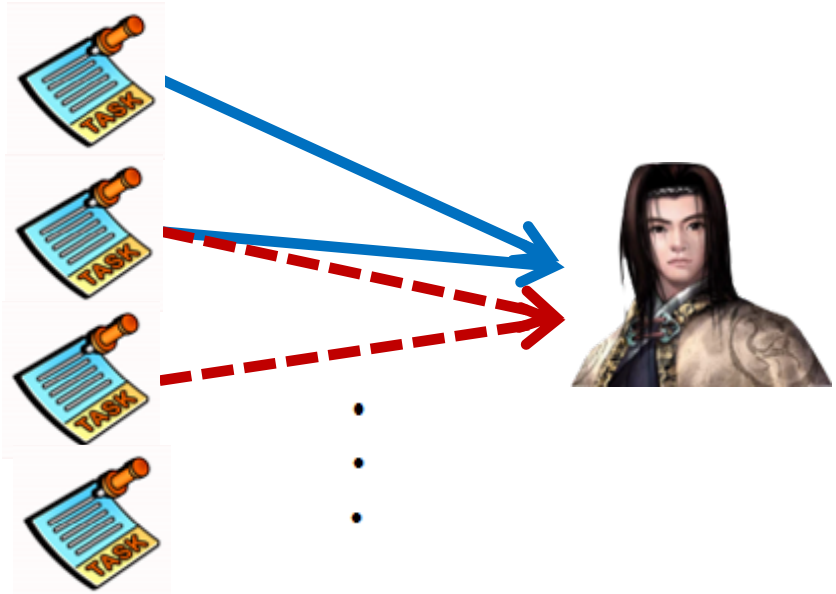
  Task Assignment: How to assign suitable tasks to appropriate workers?

  Truth Inference: How to aggregate workers' answers and infer the truth of each task?

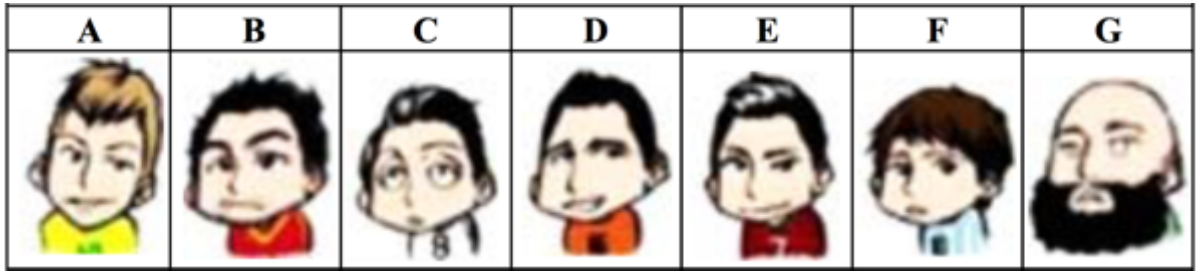


Various Settings in Task Assignment

- (1) **Task-Based Setting**: Given n tasks, select a set of the k tasks and assign to the coming worker.



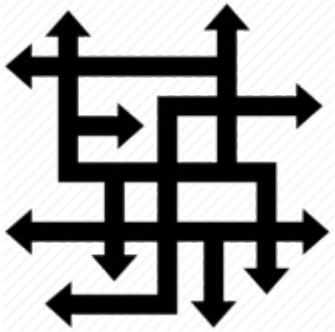
- (2) **Worker-Based Setting**: Given a set of workers, select a subset of workers, such that the task can be completed successfully and economically.



(1) Task-Based Setting

- Simple enumeration considers "n choose k" combinations

  (n = 100, k = 5) ➔ 100M combinations of possible assignments



- Require efficient (online) assignment

  Workers need fast response

- Can we develop efficient heuristics?

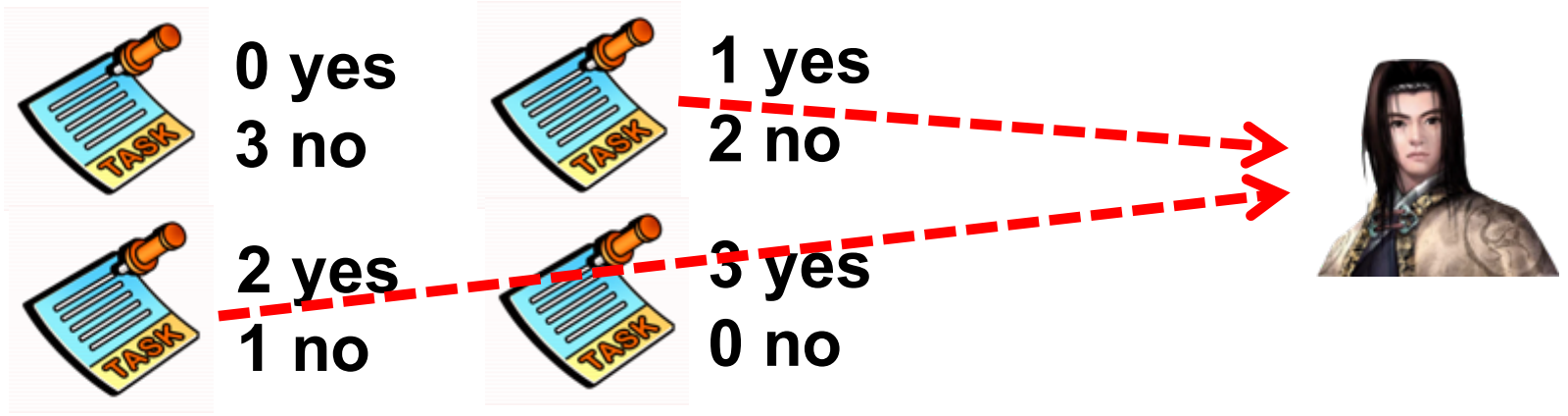
  Assignment time linear in #tasks: O(n)

Existing Works

- Intuition: Consider consistency in the collected answers for each task

  e.g., Liu et al. VLDB12, Roim et al. ICDE12

- Consider Decision-Making Task (yes/no)

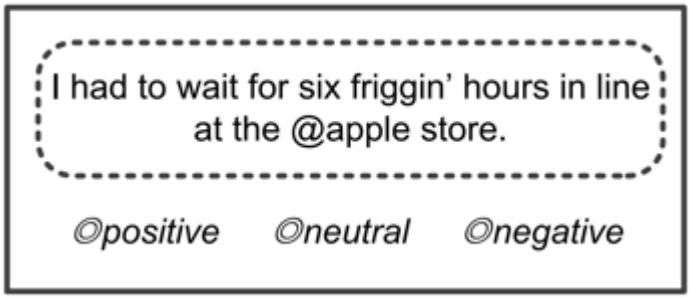
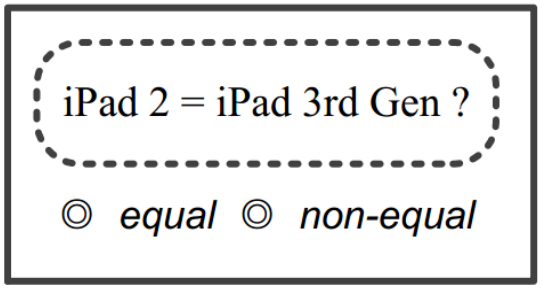


- Limitation: Miss an important factor

  How is the quality defined by an application?

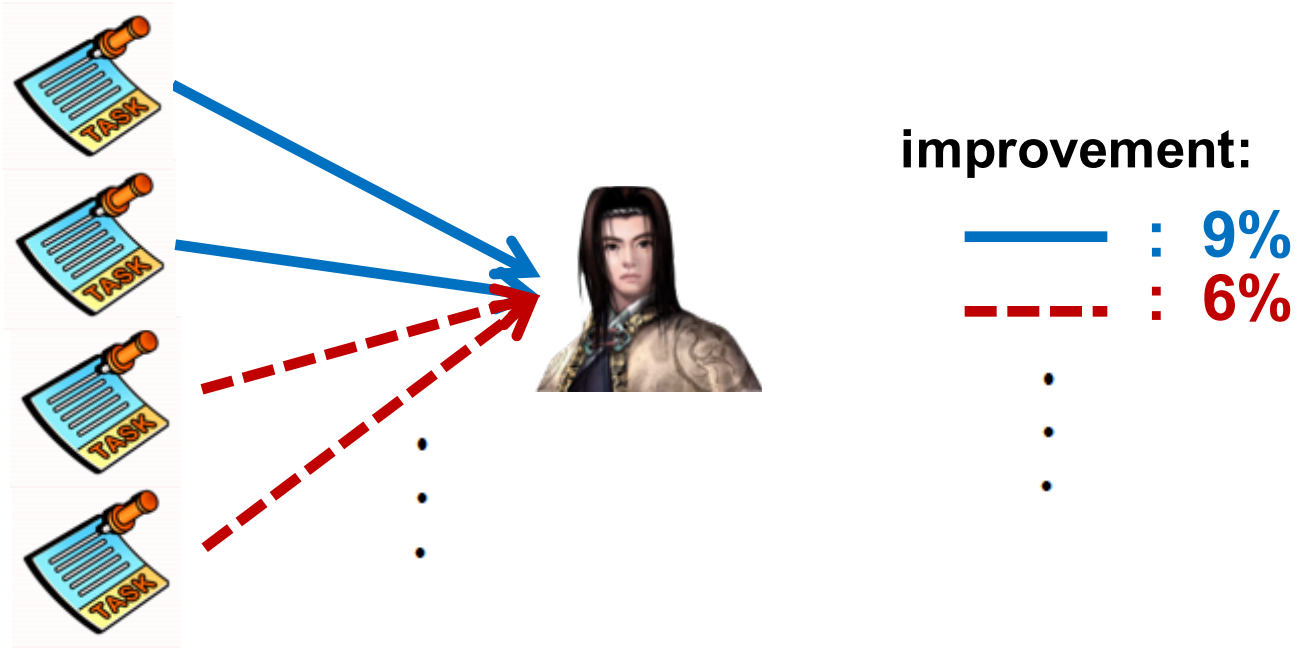
Quality Aware Task Assignment (Zheng et al. SIGMOD'15)

- Different applications may have different qualities (aka "evaluation metric") to optimize

Application	Sentiment Analysis	Entity Resolution
Task		
Evaluation Metric	Accuracy	F-score ("equal" label)

Our Ideas

- Select the best set of tasks with highest quality improvement in the specified evaluation metric.



- Challenges:
  (1) how to estimate the quality improvement for each set of tasks;
  (2) how to efficiently select the optimal set of tasks.

Solutions to Challenge 1: Quality Estimation

- **Solution Framework**

  (1) Leverage the answers collected from workers to create a "distribution matrix";

	Yes	No
Task 1	0.8	0.2
Task 2	0.6	0.4

In the first task, the probability that the first label to be the truth is 80%.

(2) Compute the quality of "distribution matrix", i.e., the highest expected quality w.r.t. an evaluation metric;

Estimated Truth: Task 1: Yes Task 2: Yes

Ground Truth	Probability	Accuracy
Yes, Yes	0.8 * 0.6 = 0.48	100%
Yes, No	0.8 * 0.4 = 0.32	50%
No, Yes	0.2 * 0.6 = 0.12	50%
No, No	0.2 * 0.4 = 0.08	0%

100% * 0.48 + 50% * 0.32 + 50% * 0.12 + 0% * 0.08 = 70%

(3) Estimate the quality improvement of "distribution matrix" for each set of tasks.

Solutions to Challenge 2: Efficient Selection

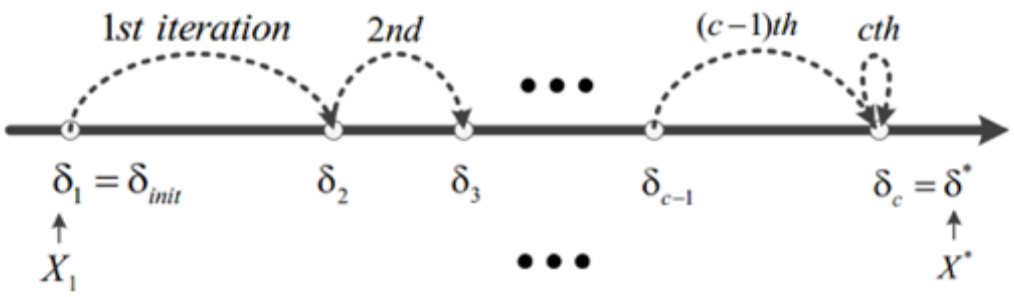
- Focus on 2 evaluation metrics

(1) Accuracy:

Define the benefit of assigning each task;
Choose the task with the highest benefit.

(2) F-score:

Iterative approach (prove that local optimum is global optimum).



The assignment iteratively improves until convergence (global optimal).

Reduce the complexity from $O\left(\binom{n}{k} \cdot n\right)$ to $O(n)$.

Summary of Task-based Assignment (Zheng et al. SIGMOD'15)

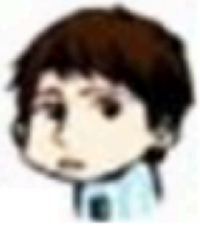
- Consider evaluation metric (Accuracy, F-score) into task assignment
- Use distribution matrix to estimate the quality of improvement
- Devise solutions to select tasks efficiently
- Open-source the task assignment framework (http://i.cs.hku.hk/~ydzheng2/crowd_task_assignment/index.html)

(2) Worker-Based Setting

- **Jury Selection Problem (Cao et al. PVLDB'12)**

Input: (1) A fixed **budget** (e.g., $20);

(2) A set of **workers** (with quality and budget):

A	B	C	D	E	F	G
						
(0.77, $9)	(0.7, $5)	(0.65, $7)	(0.6, $5)	(0.6, $2)	(0.25, $3)	(0.2, $6)

- **Goal: Select a subset of workers (called "Jury") such that the Quality of Jury (called "Jury Quality") is maximized and the overall cost <= the budget.**

B	C	D
		
(0.7, $5)	(0.65, $7)	(0.6, $5)

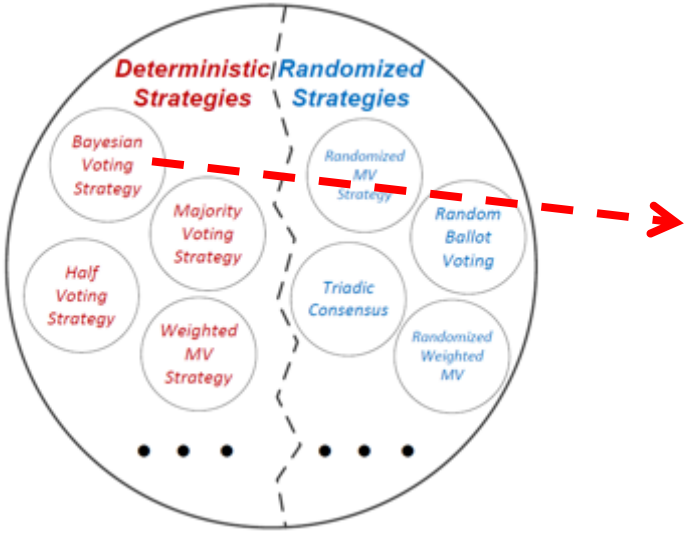
(1) Overall Cost: $5+$7+$5=$18 (<$20);
(2) Jury Quality: JQ({0.7, 0.65, 0.2}).

Jury Quality Computation

- **Jury Quality** in [Cao et al. PVLDB'12]

  JQ({0.7,0.65,0.2}, **Majority Voting**):
  "at least 2 workers answer correctly", i.e.,
  0.7*0.65*0.8 + 0.7*0.35*0.2 + 0.3*0.65*0.2 +
  0.7*0.65*0.2 = 54.3%

- **Optimal Jury Quality** in [Zheng et al. EDBT'15]



"Bayesian Voting"
is Optimal
in Jury Selection Problem!

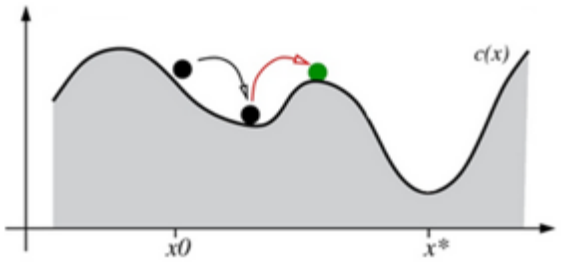
Challenges & Solutions

- **Challenge 1: Computing Optimal Jury Quality (w.r.t. Bayesian Voting) is NP-Hard !**

⇨ **Bucket-Based Approximation Algorithm (Approximation Error within 1% in Polynomial Time)**

- **Challenge 2: Selecting the Best Set of Workers is NP-Hard !**

⇨ **Make use of Simulated Annealing Heuristic which avoids local optimal**



Summary of Worker-based Assignment (Zheng et al. EDBT'15)

- **Define the optimality of "Jury Selection Problem" [Cao et al. PVLDB'12]**
- **Approximate solutions to optimal jury quality estimation (NP-hard)**
- **Heuristic solutions to the optimal jury selection problem (NP-hard)**

Outline of the Defense

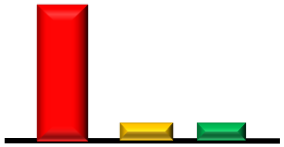
- **Part I. Truth Inference**
  - **Thorough Analysis and Comparisons (Zheng et al. PVLDB'17)**
- **Part II. Task Assignment**
  - **Task-Based Setting (Zheng et al. SIGMOD'15)**
  - **Worker-Based Setting (Zheng et al. EDBT'15)**
- **Part III. Combinations**
  - **Question-Answering Application (Zheng et al. PVLDB'16)**
  - **Image Tagging Application**

Part III. Question Answering Application (Zheng et al. PVLDB'16)

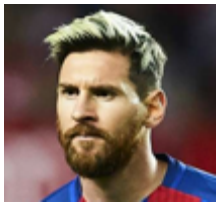
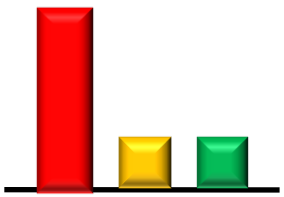
- **Existing methods fail in QA-based tasks**
- **Deeper Analysis in QA-based tasks**

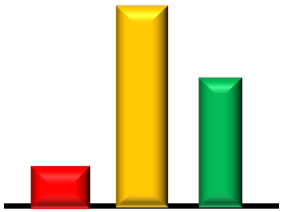
  (1) each task is related to different domains

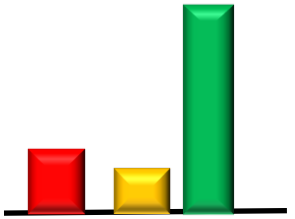
Sports Politics Entertainment

Did Michael Jordan win more NBA championships than Kobe Bryant?	→ Sports	
Is there a name for the song that FC Barcelona is known for?	→ Sports & Entertainment	

(2) each worker has diverse qualities over domains

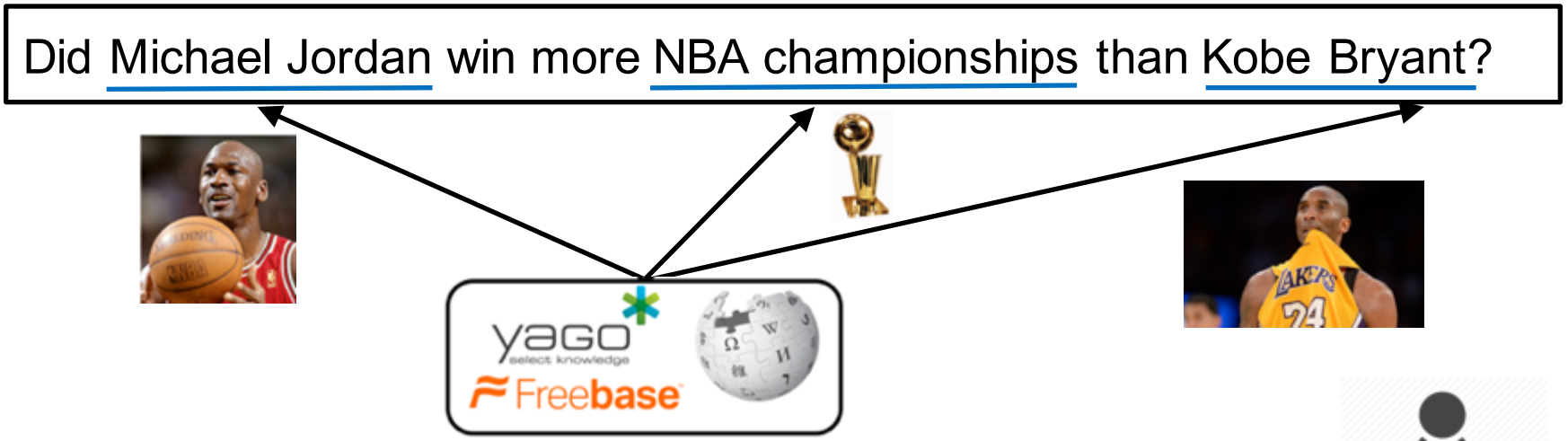







Domain Aware Task Model

- (1) **Entity linking** (map entity to **knowledge bases**)



- (2) **Hierarchical domains in knowledge bases**

- (3) **Obtain the task model (a vector of distribution) for each task**

Sports Politics Entertainment

Did Michael Jordan win more NBA championships than Kobe Bryant?

Domain Aware Worker Model

- **Initialize the model (vector) for a worker**
  Use qualification test (like an "exam") amazonmechanical turk Artificial Artificial Intelligence
  i.e., assign the tasks (with known truth) to the worker when the worker comes at first time

- **Two rules for selecting qualification test**
  (1) each selected task should capture a certain domain

Did Michael Jordan win more NBA championships than Kobe Bryant?	**Good: only related to one domain (sports)**
Is there a name for the song that FC Barcelona is known for?	**Bad: related to multiple domains (both sports & entertainment)**

(2) The domain distribution of selected tasks should approximate the distribution of all tasks

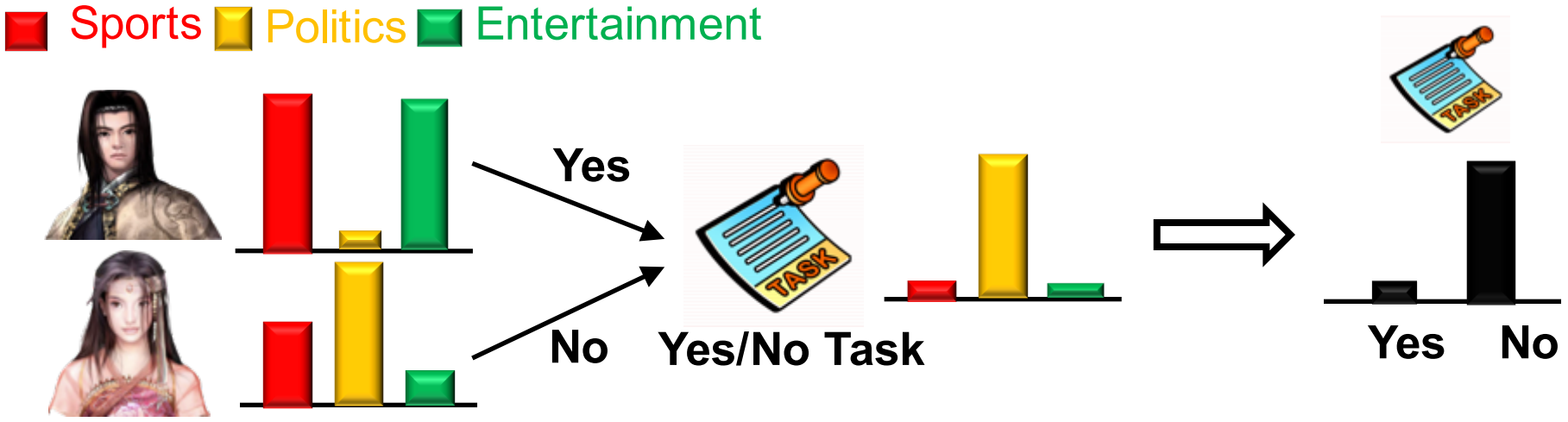
KL-divergence

$$\min_{\{n'_k\}} \sum_{k=1}^{m} \frac{n'_k}{n'} \cdot \ln \frac{n'_k \cdot n}{n' \cdot \sum_{i=1}^{n} r_k^{t_i}}$$

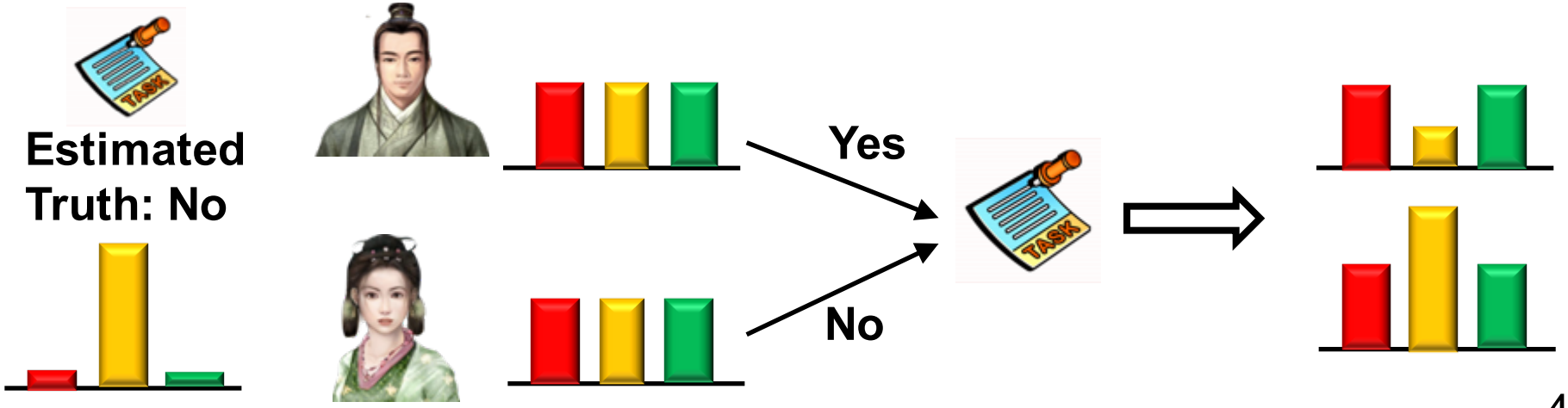
$$\text{s.t.} \sum_{k=1}^{m} n'_k = n' \text{ and } n'_k \in \mathbb{N} \text{ for } 1 \le k \le m.$$

Truth Inference

- **1. Quality for each worker ⇒ Truth for each task**



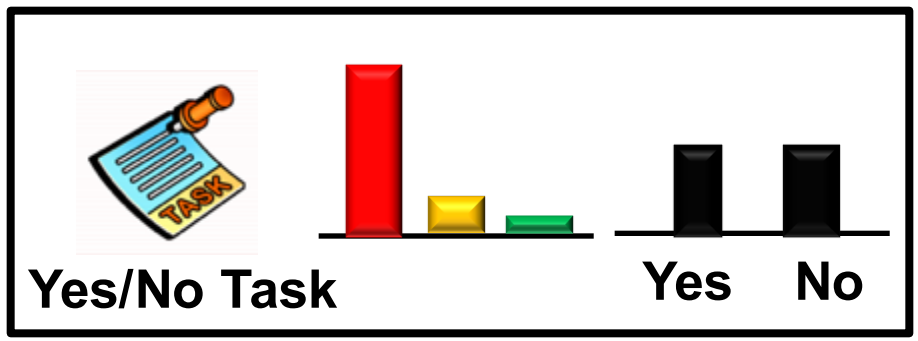
- **2. Truth for each task ⇒ Quality for each worker**

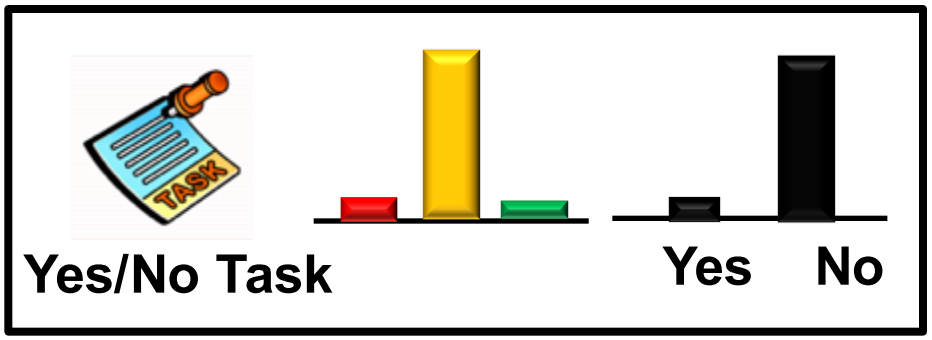


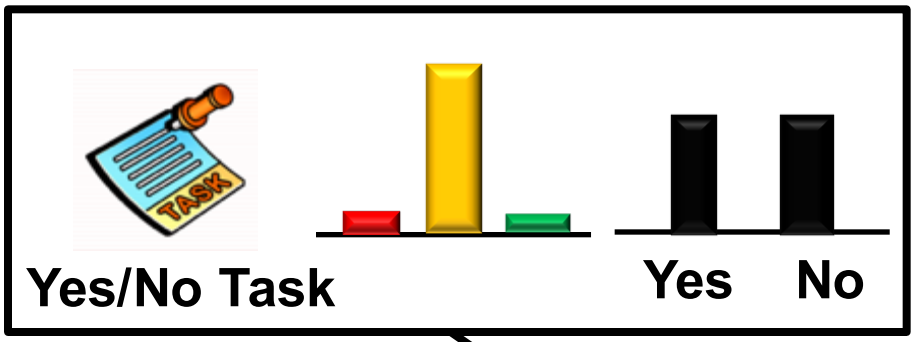
Task Assignment

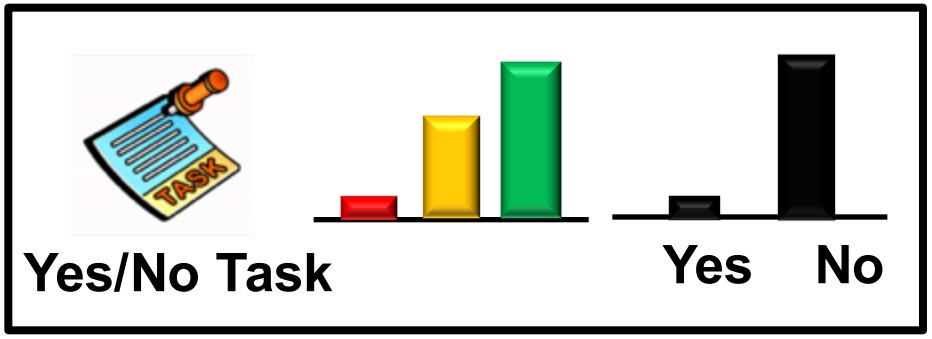
- **Select the most suitable tasks for assignment**

Sports Politics Entertainment









(1) Matching Domains

(2) Answer Uncertainty

Summary of Domain Aware Crowdsourcing (Zheng et al. PVLDB'16)

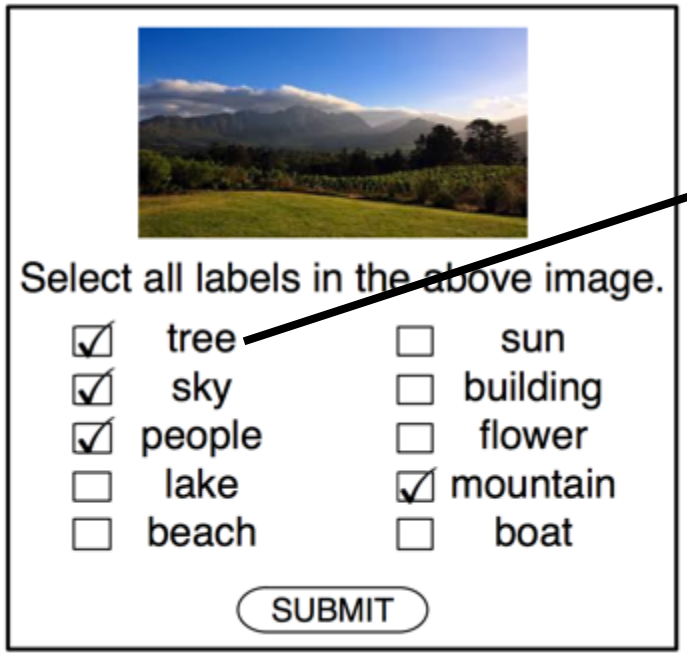
- Consider the domain aware task model and worker model
- Design solutions to accurately estimate the task model and worker model
- Incorporate task model and worker model in truth inference and task assignment

Outline of the Defense

- **Part I. Truth Inference**
  - **Thorough Analysis and Comparisons (Zheng et al. PVLDB'17)**
- **Part II. Task Assignment**
  - **Task-Based Setting (Zheng et al. SIGMOD'15)**
  - **Worker-Based Setting (Zheng et al. EDBT'15)**
- **Part III. Combinations**
  - **Question-Answering Application (Zheng et al. PVLDB'16)**
  - **Image Tagging Application**

Part III. Image Tagging Application

- **Multi-Label Tasks**



Decomposition:

Is "tree" a label in the above image ?
YES **NO**

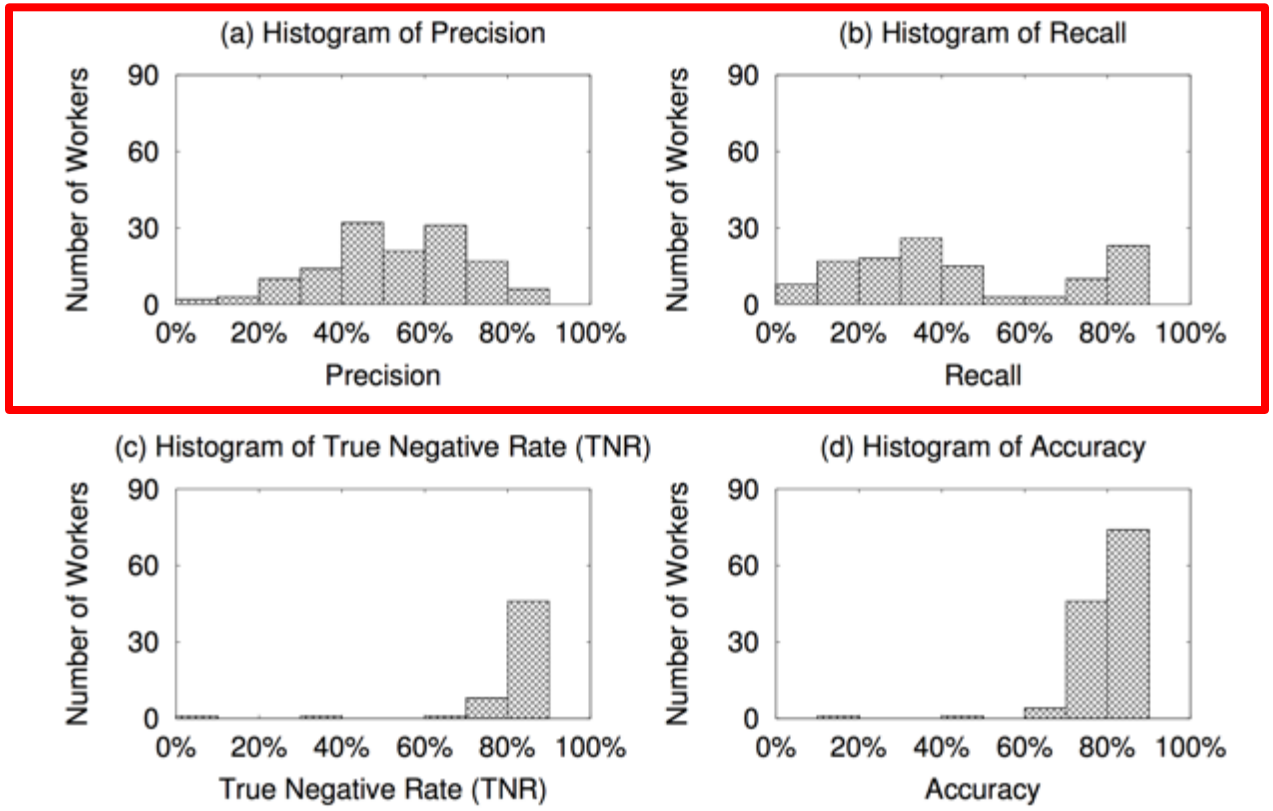
•
•
•

10 tasks

- **Naïve Solution: Multi-Label Tasks will enable six times of improvement in terms of human computation time, without sacrificing much quality. (J. Deng et al. SIGCHI'14)**

Ideas in Our Solutions

- **Worker Quality (Precision, Recall)**



- **Correlation in Labels**

e.g., If "sun" is a correct label, then the probability of "sky" is a correct label is very high.

Summary of Multi-Label Crowdsourcing

- **Design an effective worker model in answering multi-label tasks**
- **Consider label correlations in multi-label tasks**
- **Incorporate worker model and label correlations in truth inference and task assignment**

Reference – Truth Inference

[1] ZenCrowd: G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In WWW, pages 469–478, 2012.

[2] EM: A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. J.R.Statist.Soc.B, 30(1):1–38, 1977.

[3] Most Traditional Work (D&S): A.P.Dawid and A.M.Skene. Maximum likelihood estimation of observererror-rates using em algorithm. Appl.Statist., 28(1):20–28, 1979.

[4] iCrowd: J. Fan, G. Li, B. C. Ooi, K. Tan, and J. Feng. icrowd: An adaptivecrowdsourcing framework. In SIGMOD, pages 1015–1030, 2015.

[5] J. Gao, Q. Li, B. Zhao, W. Fan, and J. Han. Truth discovery andcrowdsourcing aggregation: A unified perspective. VLDB, 8(12):2048–2049, 2015

[6] CrowdPOI: H. Hu, Y. Zheng, Z. Bao, G. Li, and J. Feng. Crowdsourced poi labelling:Location-aware result inference and task assignment. In ICDE, 2016.

[7] P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazonmechanical turk. In SIGKDD Workshop, pages 64–67, 2010.

[8] M. Joglekar, H. Garcia-Molina, and A. G. Parameswaran. Evaluating thecrowd with confidence. In SIGKDD, pages 686–694, 2013.

[9] G. Li, J. Wang, Y. Zheng, and M. J. Franklin. Crowdsourced datamanagement: A survey. TKDE, 28(9):2296–2319, 2016.

[10] CATD: Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. PVLDB,8(4):425–436, 2014.

[11] PM: Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts inheterogeneous data by truth discovery and source reliability estimation. InSIGMOD, pages 1187–1198, 2014.

[12] KOS / VI-BP / VI-MF: Q. Liu, J. Peng, and A. T. Ihler. Variational inference for crowdsourcing. In NIPS, pages 701–709, 2012.

[13] CDAS: X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. CDAS: Acrowdsourcing data analytics system. PVLDB, 5(10):1040–1051, 2012

Reference – Truth Inference (cont'd)

[14] FaitCrowd: F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han.Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In KDD, pages 745–754. ACM, 2015.

[15] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. Journal of Machine Learning Research,13:491–518, 2012.

[16] V. C. Raykar, S. Yu, L. H. Zhao, A. K. Jerebko, C. Florin, G. H. Valadez,L. Bogoni, and L. Moy. Supervised learning from multiple experts: whom totrust when everyone lies a bit. In ICML, pages 889–896, 2009.

[17] LFC: V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, andL. Moy. Learning from crowds. JMLR, 11(Apr):1297–1322, 2010.

[18] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, Reynold Cheng. Truth Inference in Crowdsourcing: Is the Problem Solved? VLDB 2017.

[19] DOCS: Yudian Zheng, Guoliang Li, Reynold Cheng. DOCS: A Domain-Aware Crowdsourcing System Using Knowledge Bases. VLDB 2017.

[20] CBCC: M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi.Community-based bayesian aggregation models for crowdsourcing. In WWW,pages 155–164, 2014.

[21] Minimax: D. Zhou, S. Basu, Y. Mao, and J. C. Platt. Learning from the wisdom ofcrowds by minimax entropy. In NIPS, pages 2195–2203, 2012.

[22] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi. Inferring groundtruth from subjective labelling of venus images. In NIPS, pages 1085–1092,1994.

[23] Multi: P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In NIPS, pages 2424–2432, 2010.

[24] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In NIPS, pages 2035–2043, 2009.

[25] BCC: H.-C. Kim and Z. Ghahramani. Bayesian classifier combination. In AISTATS, pages 619–627, 2012.

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

Reference – Truth Inference (cont'd)

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan):993–1022, 2003.
[28] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In ECIR, pages 338–349, 2011.
[29] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. PVLDB, 6(2):37–48, 2012.
[30] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava. Online data fusion. PVLDB, 4(11):932–943, 2011.
[31] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan):993–1022, 2003.
[32] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In ECIR, pages 338–349, 2011.

Reference – Task Assignment

[33] CDAS: X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. CDAS: Acrowdsourcing data analytics system. PVLDB, 5(10):1040–1051, 2012

[34] OTA: C.-J. Ho and J. W. Vaughan. Online task assignment in crowdsourcingmarkets. In AAAI, 2012.

[35] QASCA: Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, Jianhua Feng. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications. SIGMOD 2015.

[36] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment forcrowdsourced classification. In ICML, pages 534–542, 2013.

[37] CrowdPOI: H. Hu, Y. Zheng, Z. Bao, G. Li, and J. Feng. Crowdsourced poi labelling:Location-aware result inference and task assignment. In ICDE, 2016.

[38] DOCS: Yudian Zheng, Guoliang Li, Reynold Cheng. DOCS: A Domain-Aware Crowdsourcing System Using Knowledge Bases. VLDB 2017.

[39] AskIt: R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W. C. Tan. Asking the right questions in crowd data sourcing. In ICDE, 2012.

[40] iCrowd: J. Fan, G. Li, B. C. Ooi, K. Tan, and J. Feng. icrowd: An adaptivecrowdsourcing framework. In SIGMOD, pages 1015–1030, 2015.

[41] Opt-KG: Qi Li, Fenglong Ma, Jing Gao, Lu Su, and Christopher J Quinn, Crowdsourcing High Quality Labels with a Tight Budget, WSDM 2016.

[42] Jing Gao, Qi Li, Bo Zhao, Wei Fan, and Jiawei Han, Enabling the Discovery of Reliable Information from Passively and Actively Crowdsourced Data, KDD'16 tutorial.

[43] J. Deng, O. Russakovsky, J. Krause, M. S. Bernstein, A. Berg, and L. Fei-Fei. Scalable multi-label annotation. In SIGCHI, 2014.

[44] Mozafari, B., Sarkar, P., Franklin, M., Jordan, M., & Madden, S. (2014). Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, *8*(2), 125-136.

Reference – Task Assignment (cont'd)

[45] Gokhale, C., Das, S., Doan, A., Naughton, J. F., Rampalli, N., Shavlik, J., & Zhu, X. (2014, June). Corleone: Hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 601-612). ACM.

[46] Zhao Chen, Rui Fu, Ziyuan Zhao, Zheng Liu, Leihao Xia, Lei Chen, Peng Cheng, Caleb Chen Cao, Yongxin Tong, and Chen Jason Zhang. "gMission: A General Spatial Crowdsourcing Platform." In Proceedings of the VLDB 2014 Endowment 7, no. 13.

Future Work: Scalability

- **Hard to Scale in Crowdsourcing: Tackle the 3Vs of Big Data?**

- **Existing works focus on specific problems, e.g., active learning [Mozafari et al. PVLDB'14], entity matching [Gokhale et al. SIGMOD'14].**

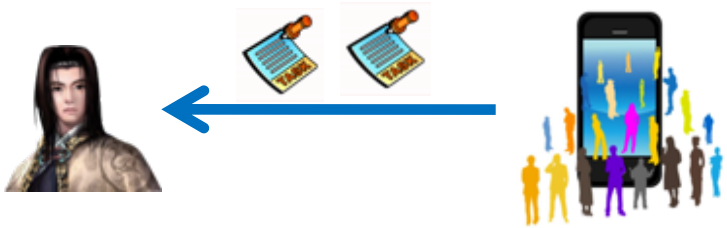
Future Work: Privacy

- (1) Requester wants to protect the privacy of their tasks from workers
  e.g., tasks may contain sensitive attributes, e.g., medical data.

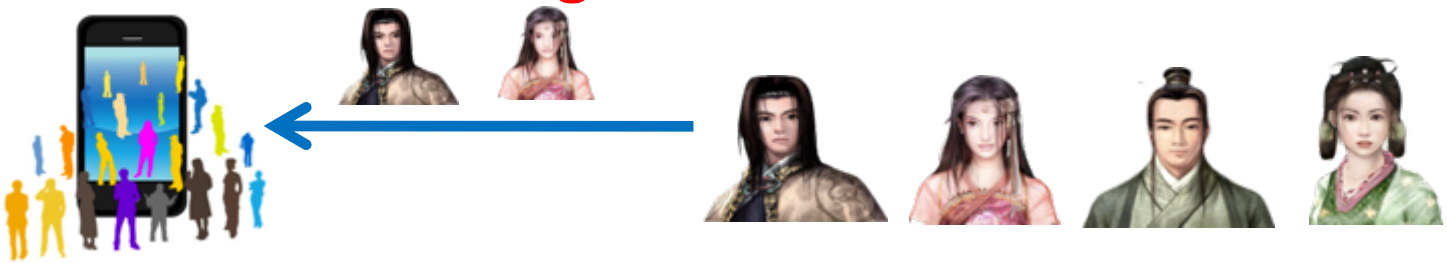
- (2) Workers wants to have privacy-preserving requirement & worker profile
  e.g., personal info of workers can be inferred from the worker's answers, e.g., location, gender, etc.

Future Work: Mobile Crowdsourcing

- Emerging mobile crowdsourcing platforms
  e.g., gMission (HKUST), ChinaCrowd (Tsinghua)
- Challenges
  (1) Other factors (e.g., spatial distance, mobile user interface) affect workers' latency and quality;
- (2) Different mechanisms
  traditional crowdsourcing platforms: workers request tasks from the platform;



for mobile crowdsourcing platform: only workers close to the crowdsourcing task can be selected.

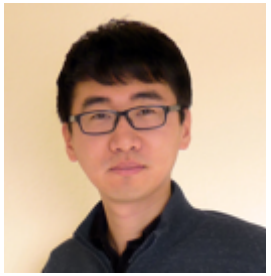


Brief Summary for Me

- Crowdsourcing and Other Projects
- Publish Over 10 Papers in Distinguished Venues

- Tutorial on Crowdsourcing (SIGMOD 2017)
- Research Internships

Thanks for all the collaborators !











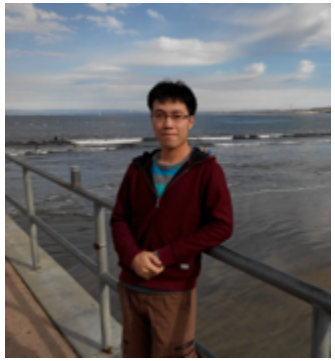


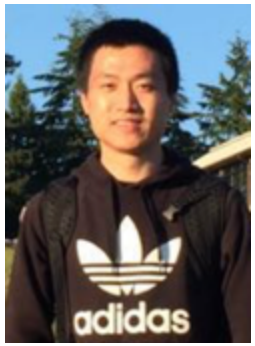


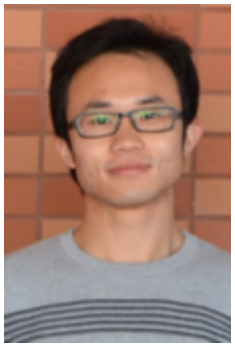
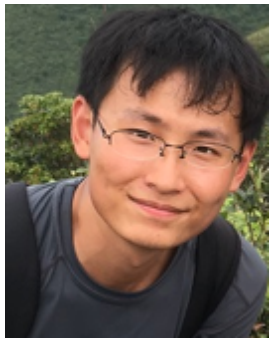

et al.

Thanks for the lab mates !




et al.

Yudian Zheng
The University of Hong Kong
zhydhkcws@gmail.com