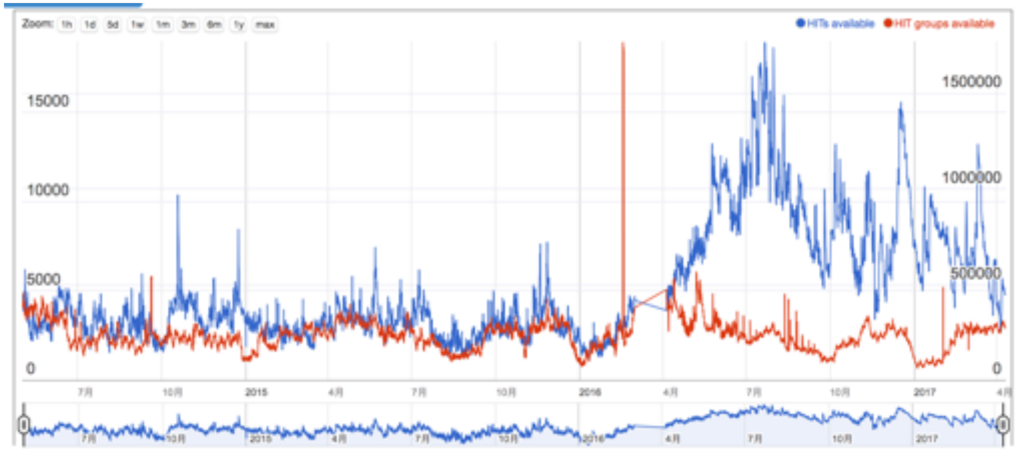
Truth Inference in Crowdsourcing: Is the Problem Solved?

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, Reynold Cheng

University of Hong Kong, Tsinghua University

Why Truth Inference?

- **Huge Amount of Crowdsourced Data**



amazonmechanical turk
Artificial Artificial Intelligence

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

- **Inevitable noise & error**

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

Motivating Example

- An Example Task

Where was ACM SIGMOD 2017 held ?

A. Raleigh

B. Chicago

I think
A. Raleigh !

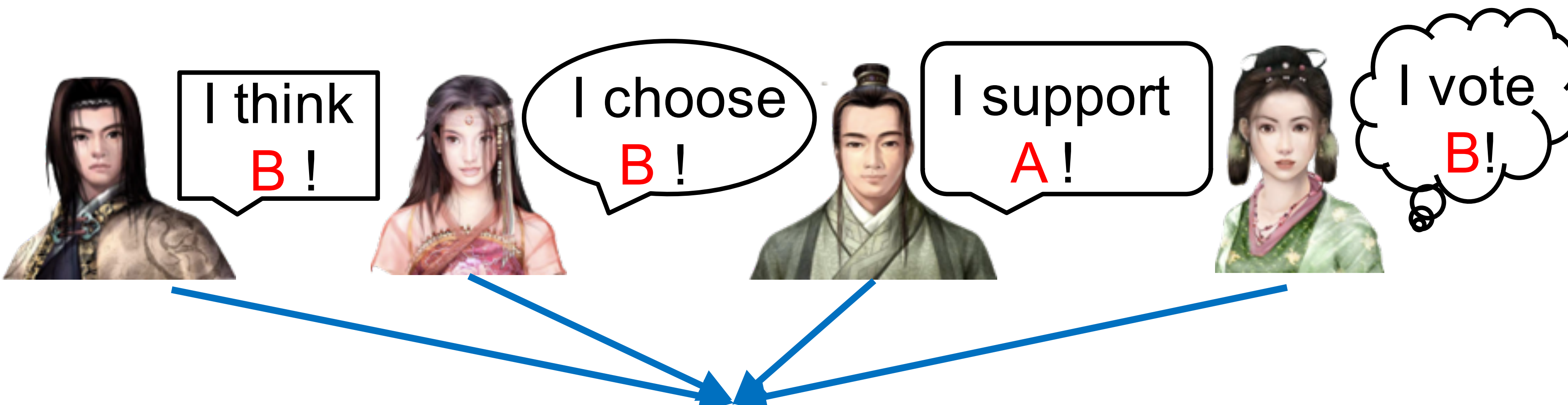
Principle: Redundancy

- Collect Answers from Multiple Workers

Where was ACM SIGMOD 2017 held ?

A. Raleigh

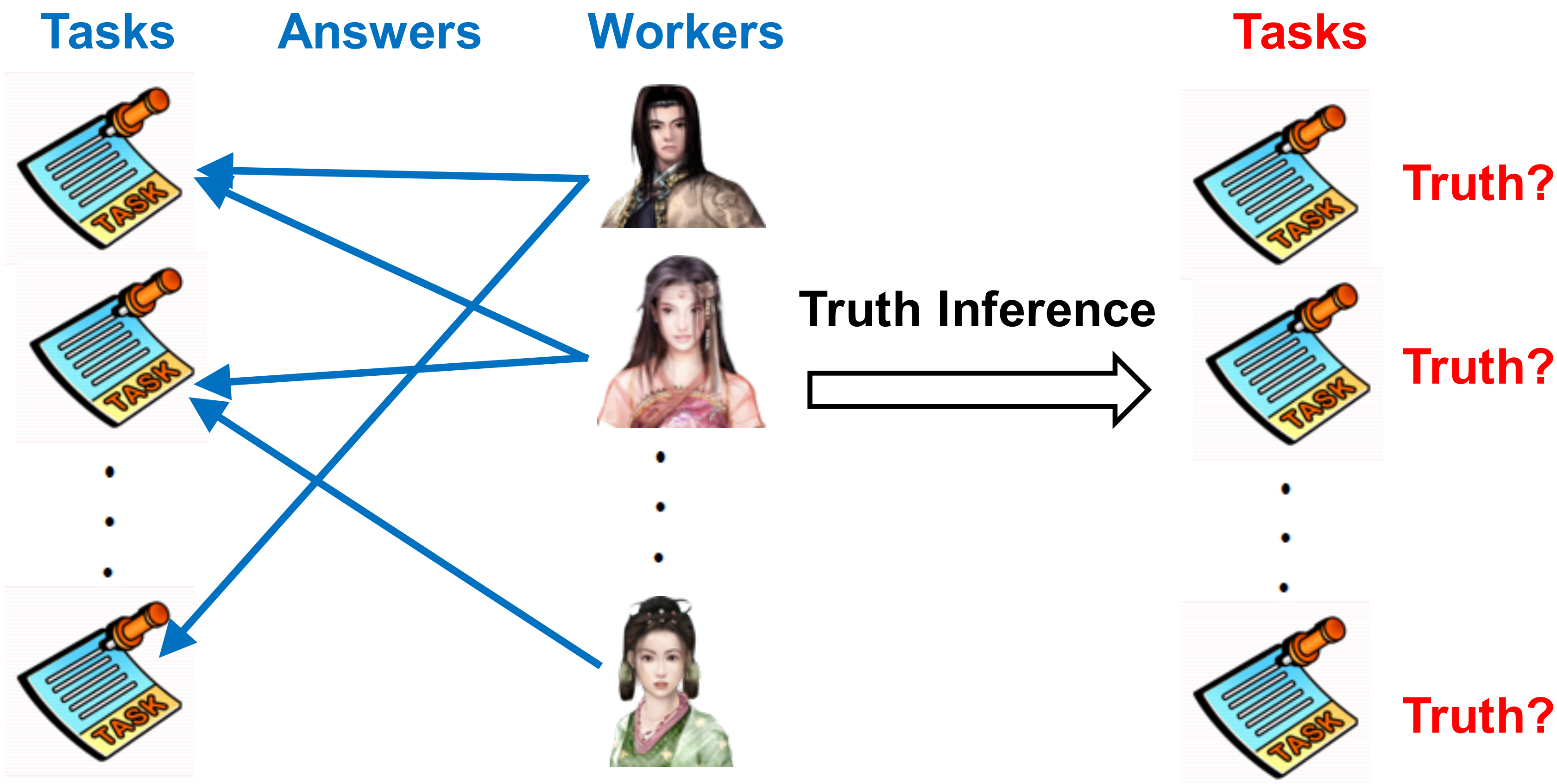
B. Chicago



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**

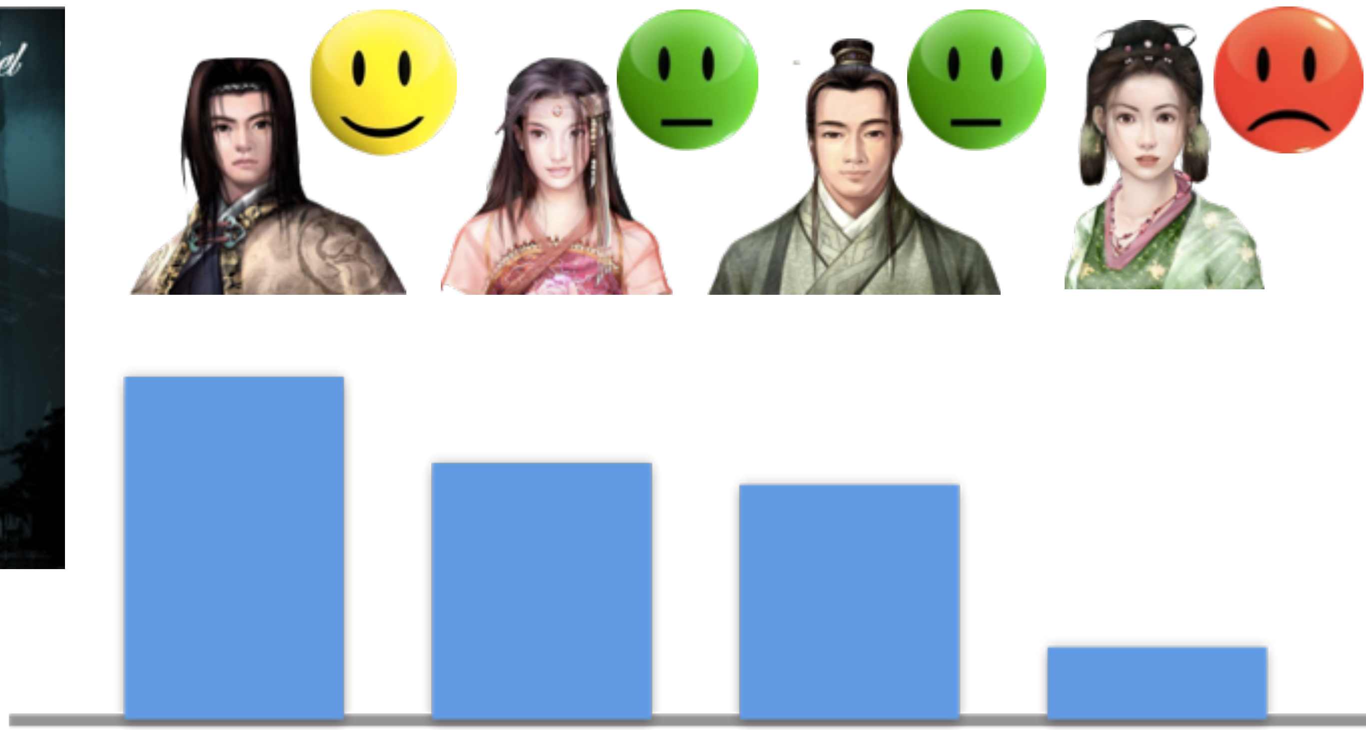
  Take the answer that is voted by the majority (or most) of workers.

- **Limitation**

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

The Key to Truth Inference

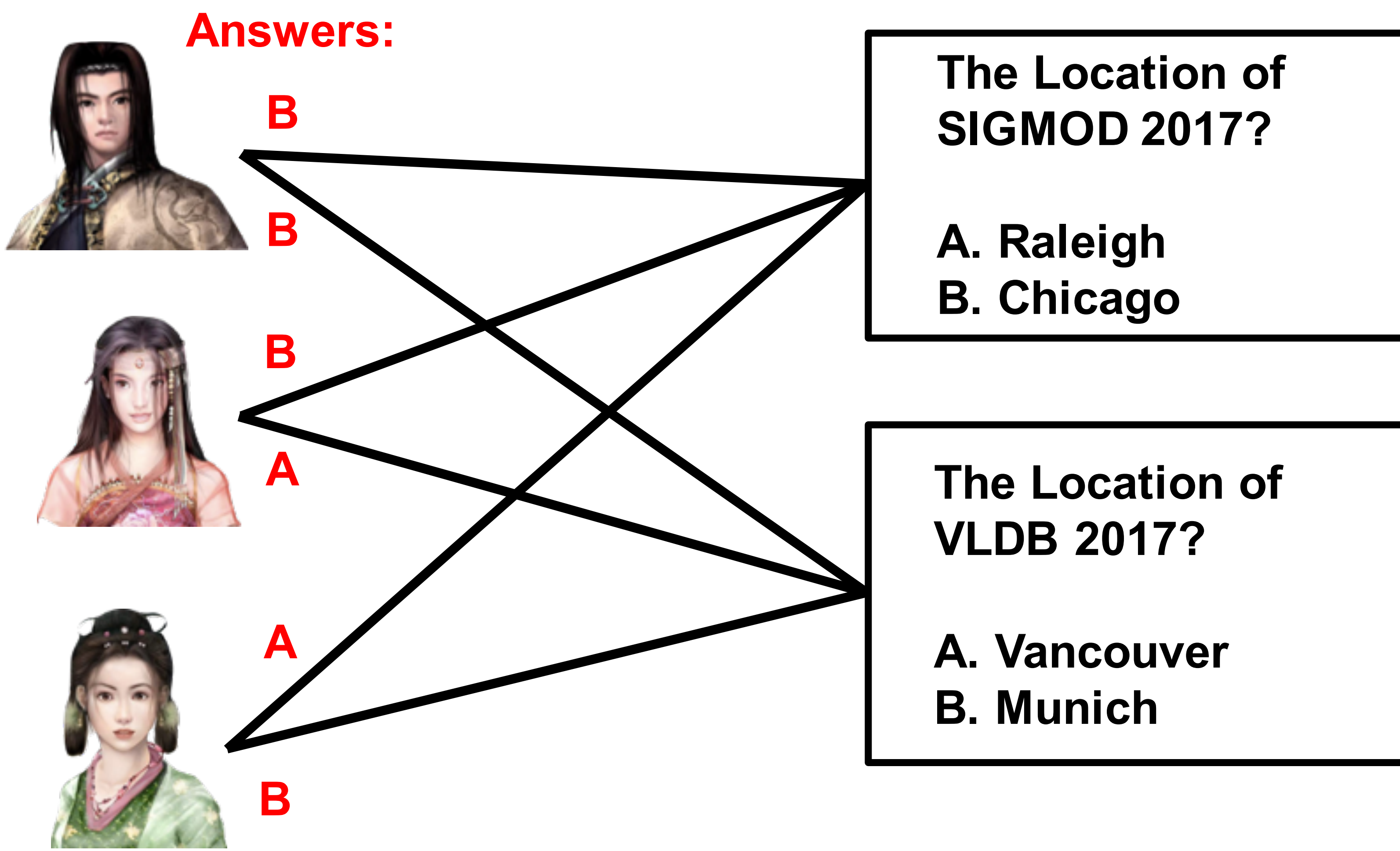
- The key is to know **each worker's quality**



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]

Three Goals in Our Work (Zheng et al. PVLDB'17)

- What are the **similarities** in existing works?

- What are the **differences** in existing works?

- Any **suggestions** to use in practice?

Part I:
Unified Framework in Existing Works

- 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

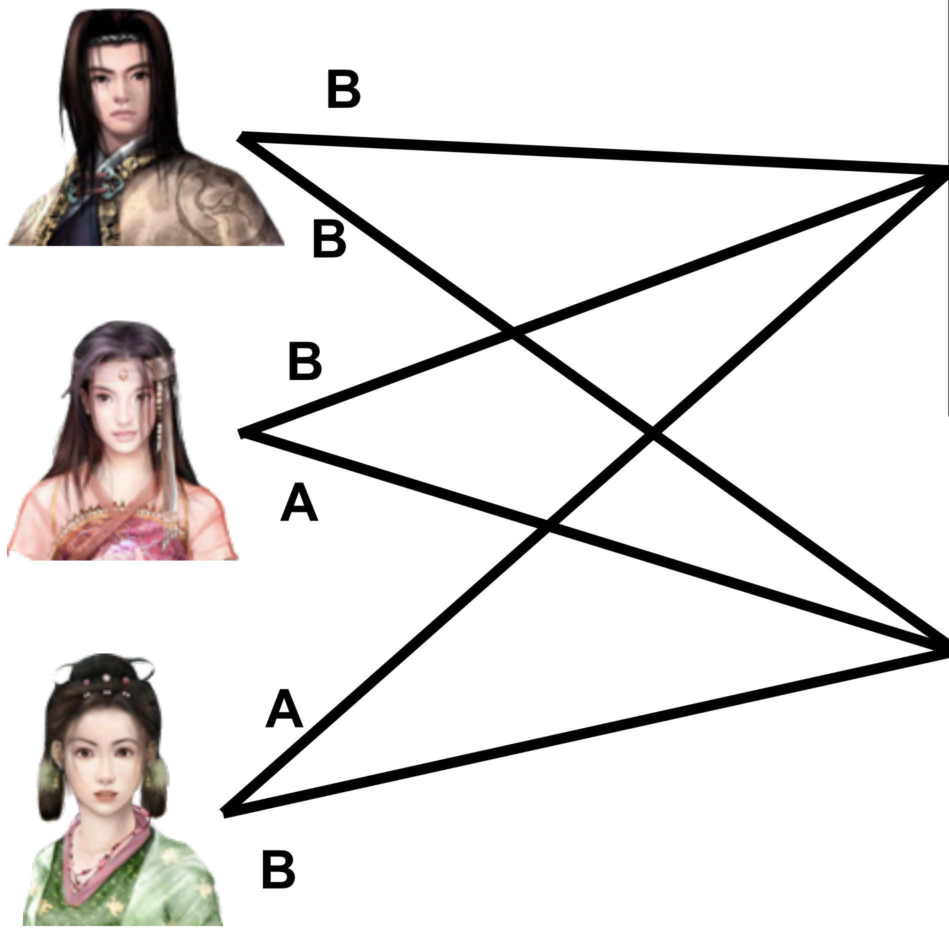
Quality:

(Estimated) Truth:

1.0

1.0

1.0



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

Part II: 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) …

Objectives

- **Different Objective Functions**
  What type of objectives they use?
  E.g., Graphical Model…

(1) 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

(2) 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

(2) Different Worker Models (cont'd)

- **Confusion Matrix**: a matrix

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

$$\begin{array}{c} \\ Pos \\ Neu \\ Neg \end{array} \begin{array}{c} \begin{array}{ccc} Pos & Neu & Neg \end{array} \\ \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \end{array}$$

*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

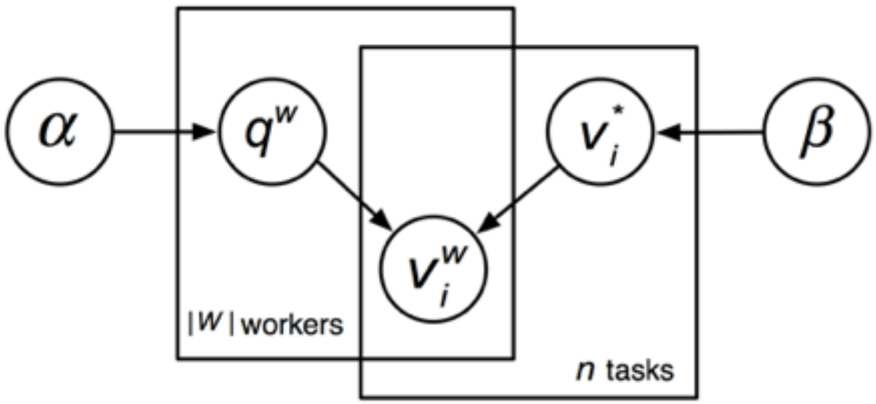
- **Bias τ & Variance σ** : numerical task

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

e.g., Raykar et al. JLMR10

(3) Different Objective Functions

- **PGM**, or Probabilistic Graphical Model (e.g., D&S [David & Skene JRSS 1979])



=> Likelihood: $$\prod_{i=1}^{n} \sum_{z \in \{\mathrm{T}, \mathrm{F}\}} \Pr(v_i^* = z) \cdot \prod_{w \in \mathcal{W}^i} \Pr(v_i^w \,|\, q^w, v_i^* = z)$$

- **Optimization** (self-defined objective function, e.g., PM [Li et al. SIGMOD14])

$$\min_{\{q^w\}, \{v_i^*\}} f(\{q^w\}, \{v_i^*\}) = \sum_{w \in \mathcal{W}} q^w \cdot \sum_{t_i \in \mathcal{T}^w} d(v_i^w, v_i^*)$$

Summary of Truth Inference Methods

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

Summary of Truth Inference Methods (cont'd)

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

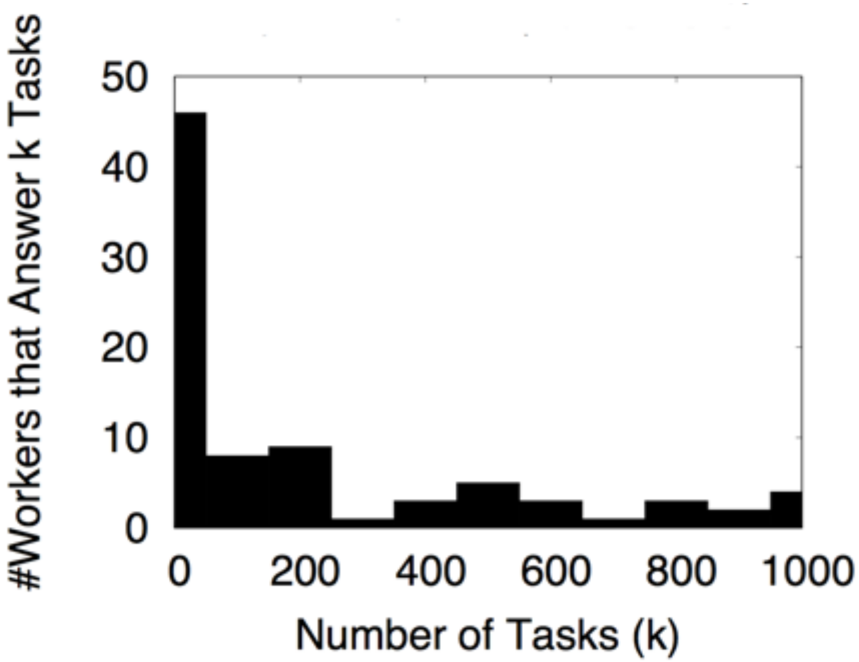
Part III:
Experiments and Analysis

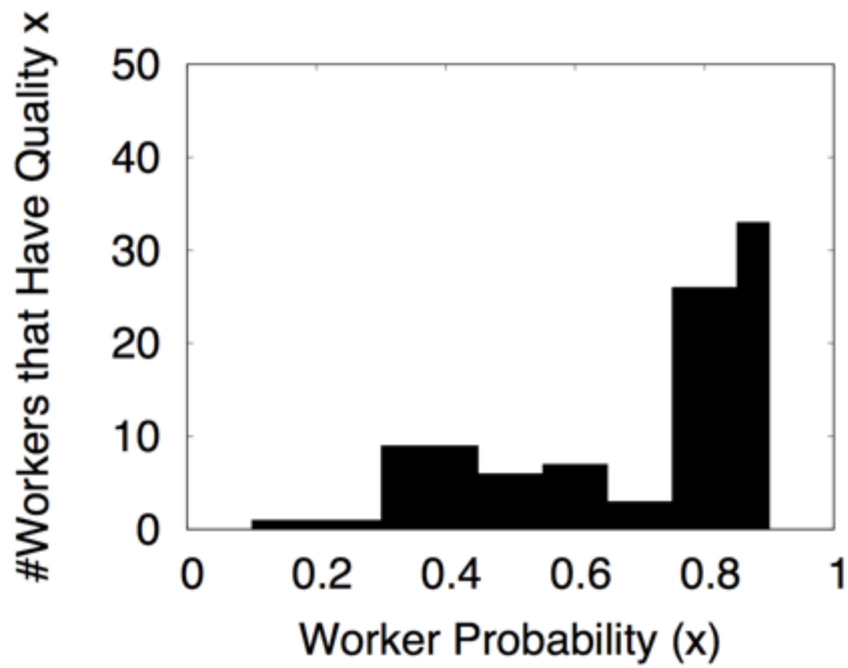
- **Statistics of Datasets**

Dataset	# Tasks	# Answers Per Task	# Workers	Description
Sentiment Analysis [Zheng et al. VLDB17]	1000	20	185	Given a tweet, the worker will identify the sentiment of the tweet
Duck [Welinder et al. NIPS10]	108	39	39	Given an image, the worker will identify whether the image contains a duck or not
Product [Wang et al. VLDB12]	8315	3	85	Given a pair of products, the worker will identify whether or not they refer to the same product

Experiments and Analysis (cont'd)

- **Observations (Sentiment Analysis)**



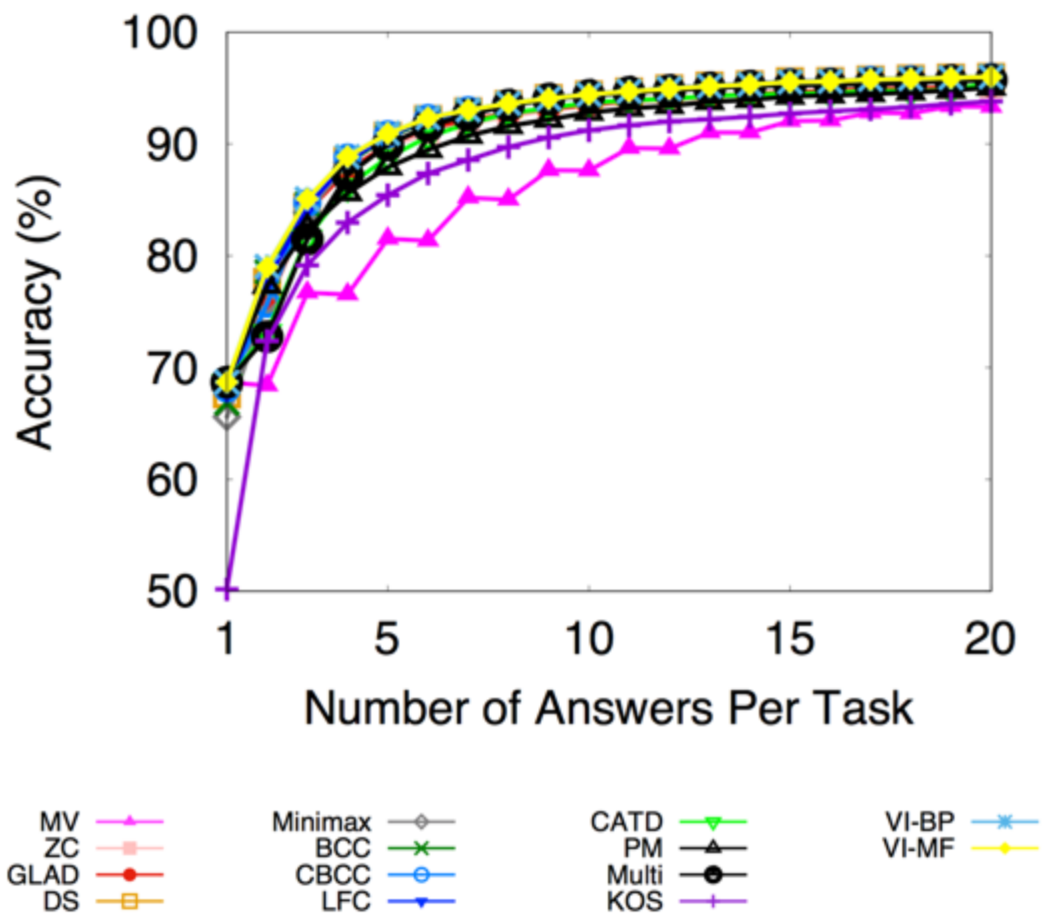


#workers' answers conform to long-tail phenomenon

Not all workers are of very high quality

Experiments and Analysis (cont'd)

- **Change of Quality vs. #Answers (Sentiment Analysis)**



Observations:

1. The quality increases with #answers;

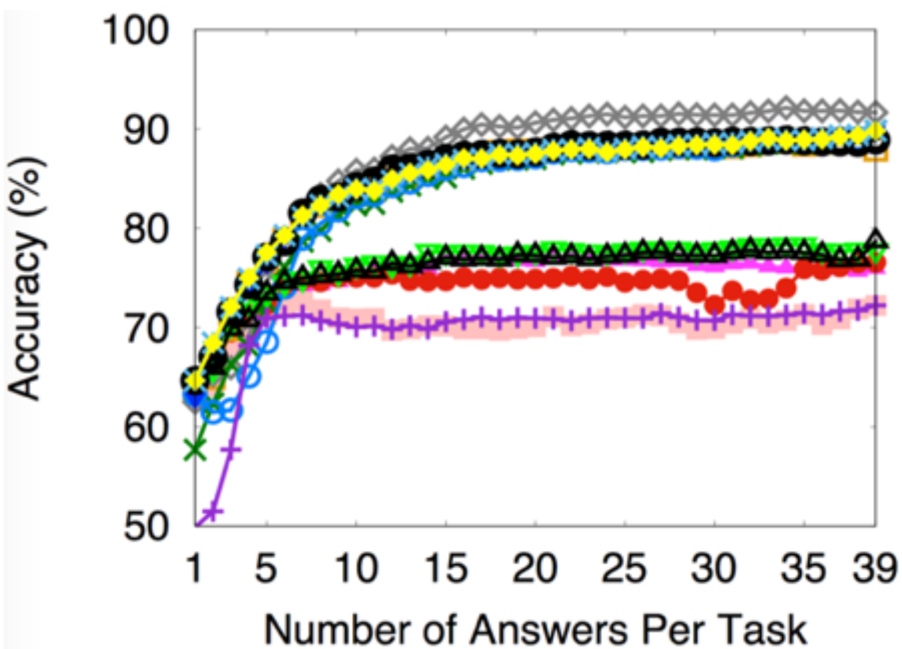
2. The quality improvement is significant with few answers, and is marginal with more answers;

3. Most methods are similar, except for Majority Voting (in pink color).

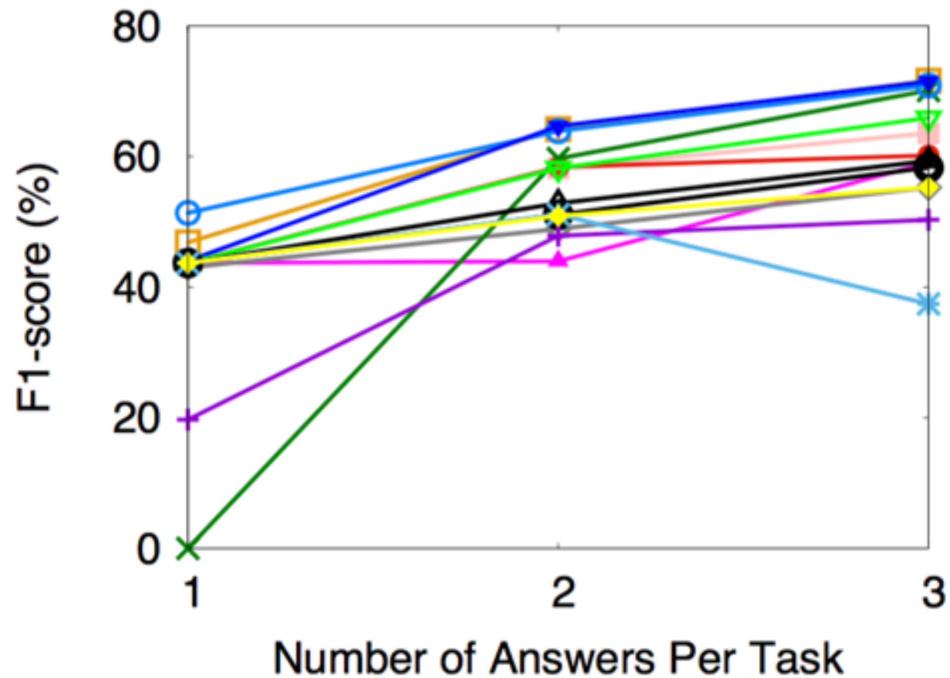
Experiments and Analysis (cont'd)

- Performance on more datasets

Dataset "Duck"



Dataset "Product"



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

- **The key to truth is to know each worker's quality;**
- **Unified Framework: Relationships between "quality for each worker" and "truth for each task";**
- **Different task types, worker models and objectives**

Open-Source Datasets & Codes

- **Public crowdsourcing datasets:** http://i.cs.hku.hk/~ydzheng2/crowd_survey/datasets.html

- **Implementations of truth inference algorithms:** http://i.cs.hku.hk/~ydzheng2/crowd_truth_inference/index.html

Reference

[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 (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 (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.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, Reynold Cheng

University of Hong Kong, Tsinghua University