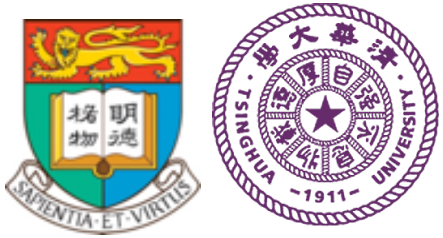


DOCS: A Domain-Aware Crowdsourcing System Using Knowledge Bases

Yudian Zheng, Guoliang Li, Reynold Cheng

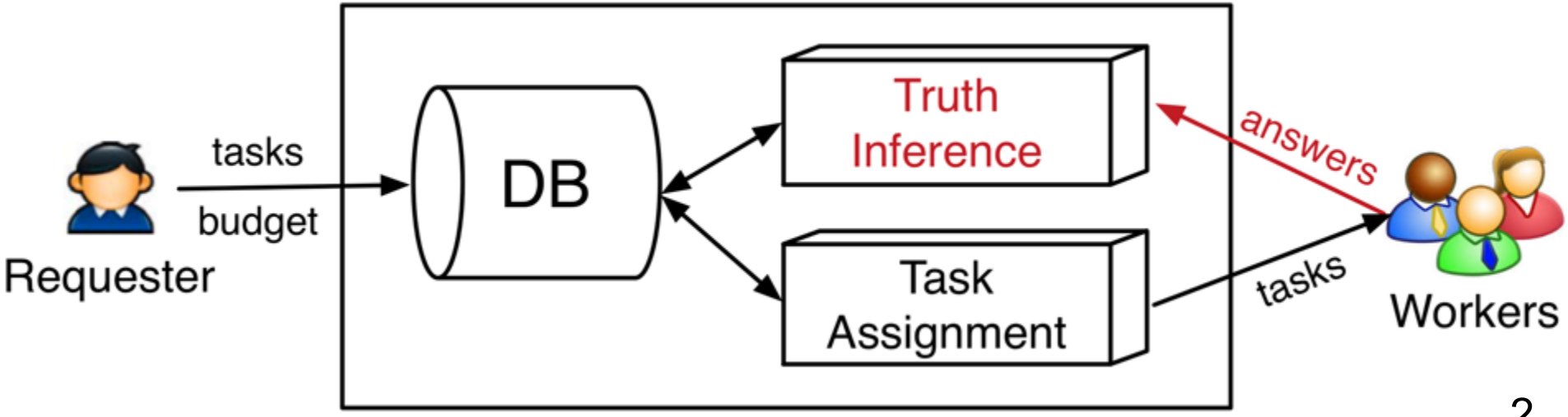
University of Hong Kong, Tsinghua University

Crowdsourcing Workflow

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

  Task Assignment: When a worker comes to the platform, the worker will be assigned to a set of tasks;

  Truth Inference: When a worker accomplishes tasks, the platform will collect answers from the worker.



Question Answering Application

- **Existing Platforms**

Y! ANSWERS

Quora

Baidu 百度

知乎

- **Examples**

Did Michael Jordan win more NBA championships than Kobe Bryant?

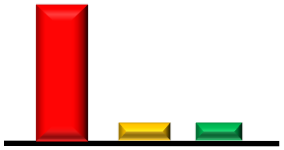
Is there a name for the song that FC Barcelona is known for?

Existing Works Fail in QA tasks

- **Each task is related to different domains**

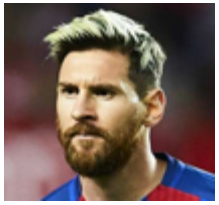
Sports Politics Entertainment

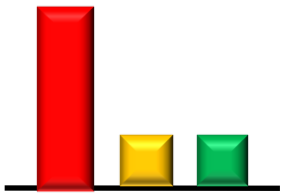
Did Michael Jordan win more NBA championships than Kobe Bryant? → Sports

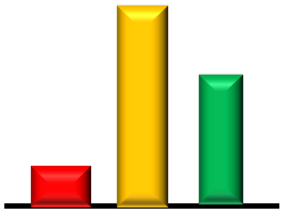


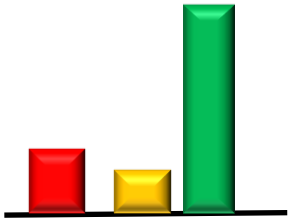
Is there a name for the song that FC Barcelona is known for? → Sports & Entertainment

- **Each worker has diverse qualities over domains**







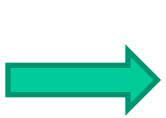


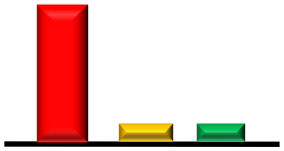
Our Solutions

- **Build domain aware task model**

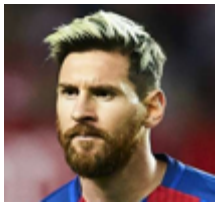
Sports Politics Entertainment

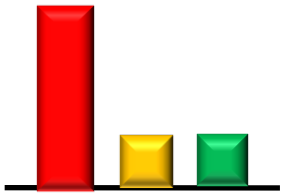
Did Michael Jordan win more NBA championships than Kobe Bryant?

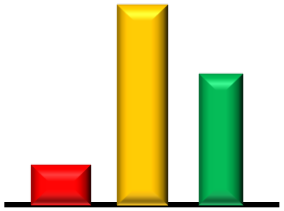


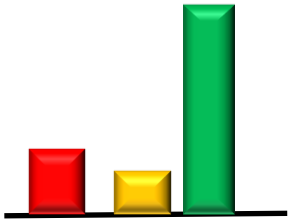


- **Build domain aware worker model**





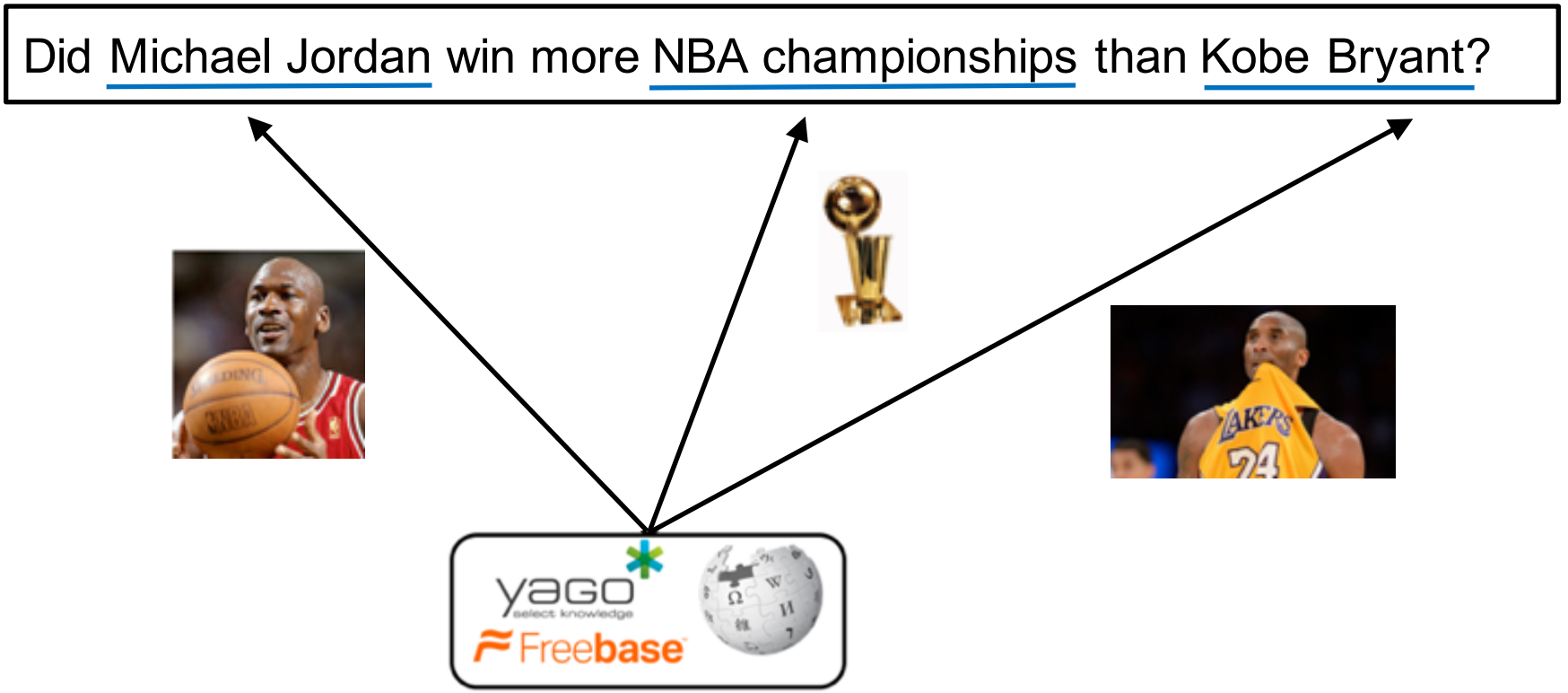




- **Apply them to truth inference and task assignment**

Part I:
Domain Aware Task Model (3 steps)

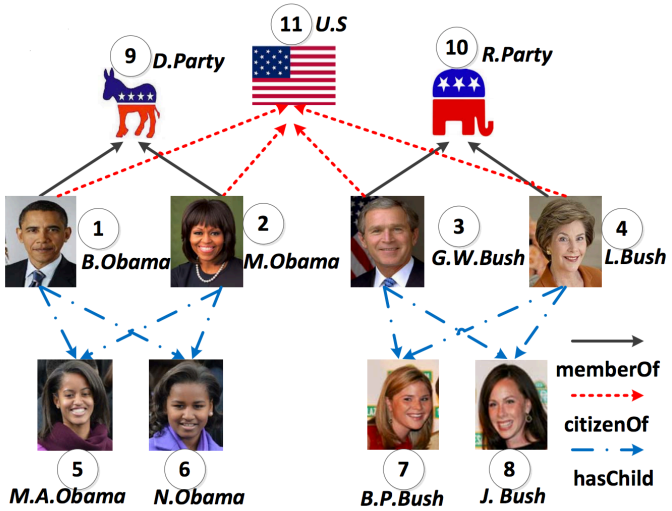
- **Step 1: Entity linking (map entity to knowledge bases)**



W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. TKDE, 27(2):443–460, 2015.

Part I:
Domain Aware Task Model (3 steps)

- **Step 2: Hierarchical domains in knowledge bases**



- **Step 3: For each task, we obtain the task model (a vector of distribution)**

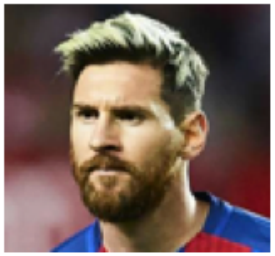
Sports Politics Entertainment

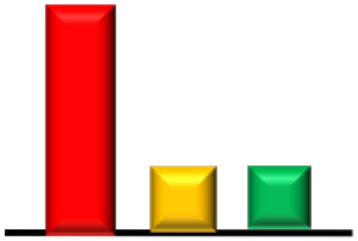
Did Michael Jordan win more NBA championships than Kobe Bryant?

Part II: Domain Aware Worker Model

- **Model each worker using a vector**

Sports Politics Entertainment





Each element in the vector is in the range of (0,1), indicating the expertise of the worker to a specific domain.

- **Our ideas to initialize each worker's model**

Use qualification test (like an "exam")
i.e., assign the tasks (with known truth) to the worker when the worker comes at first time

amazon mechanical turk
beta
Artificial Artificial Intelligence

Part II: Domain Aware Worker Model (cont'd)

- **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

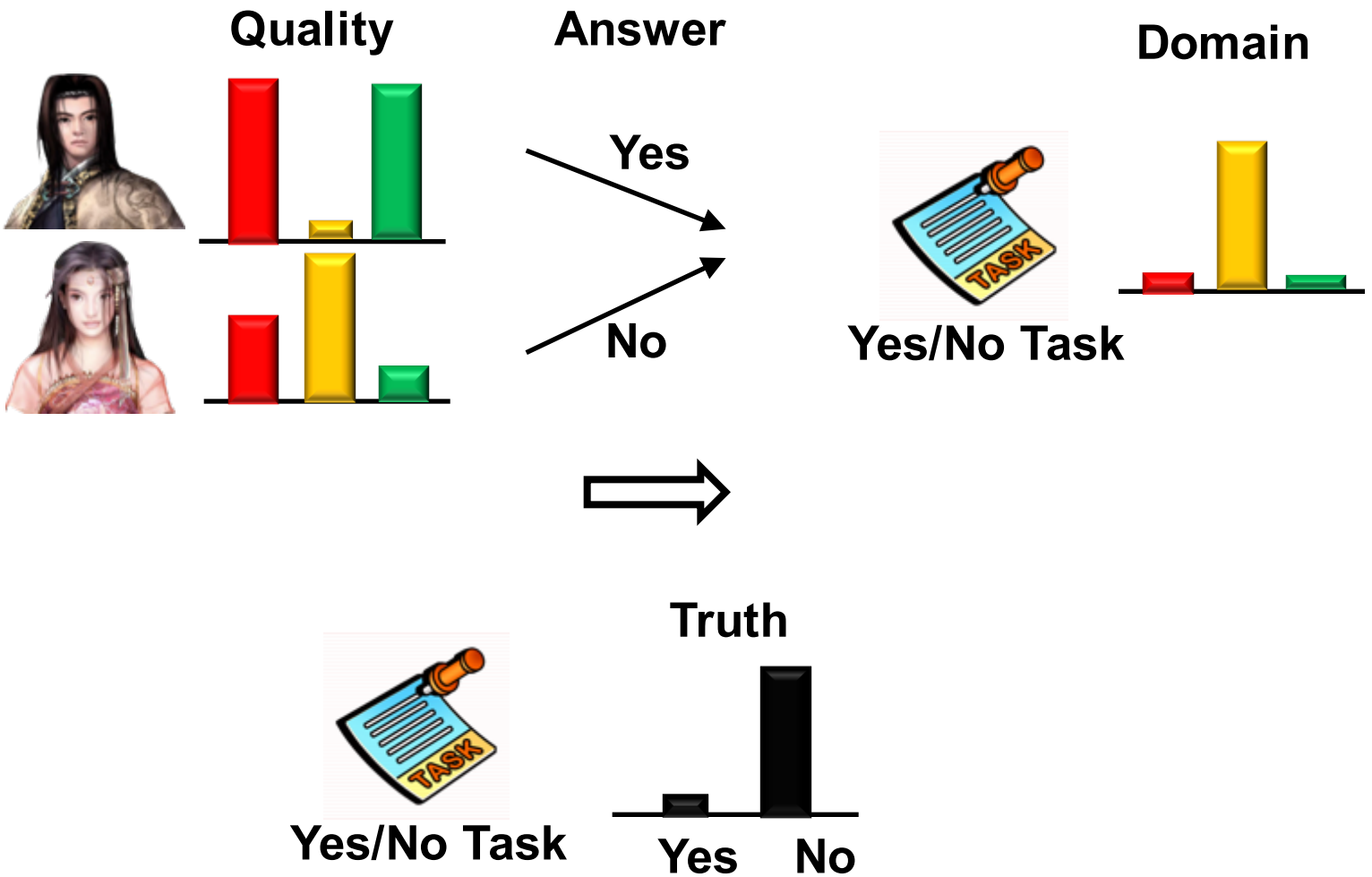
$$\min_{\{n'_k\}} \sum_{k=1}^{m} \frac{n'_k}{n'} \cdot \ln \frac{n'_k \cdot n}{n' \cdot \sum_{i=1}^{n} r_k^{t_i}}$$

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

Truth Inference

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

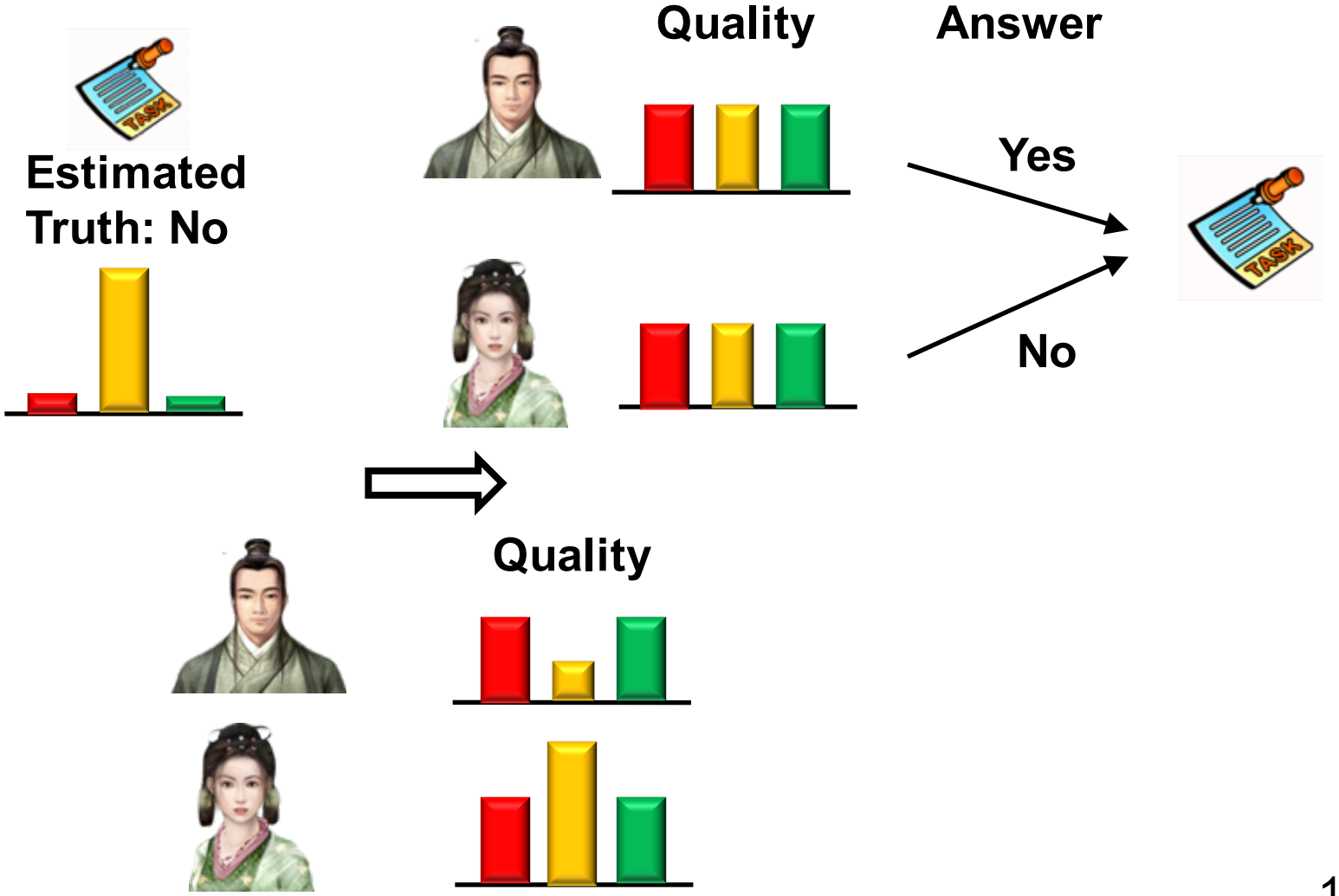
Sports Politics Entertainment



Truth Inference (cont'd)

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

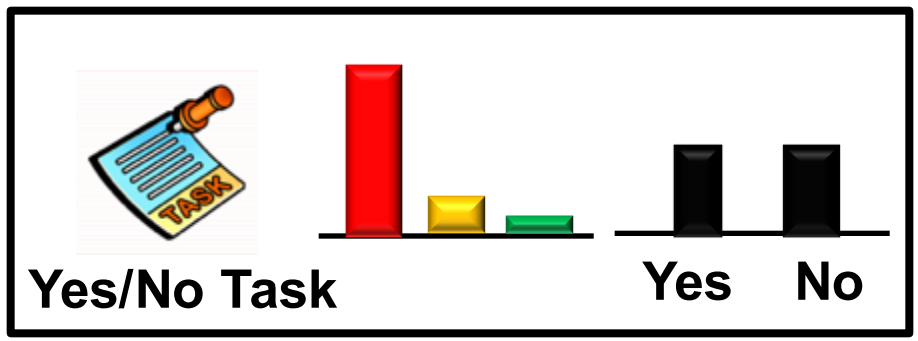
Sports Politics Entertainment

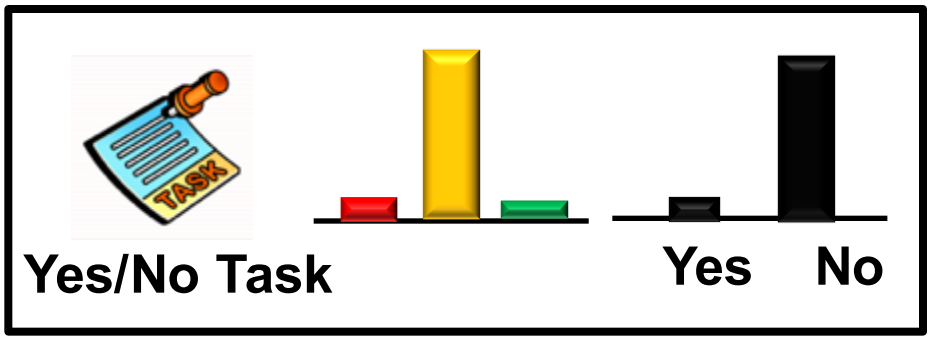


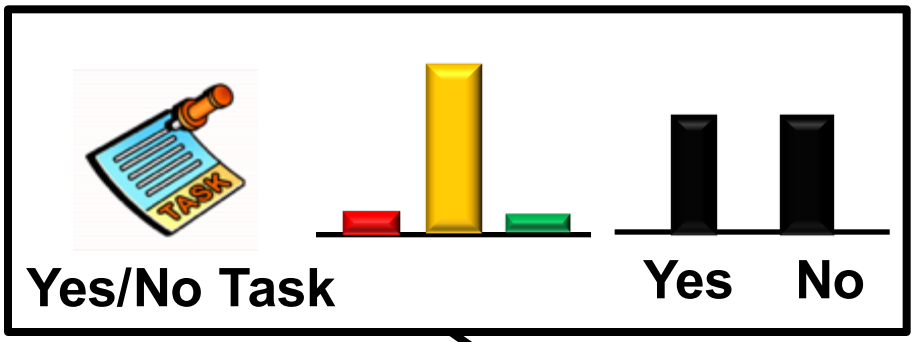
Task Assignment

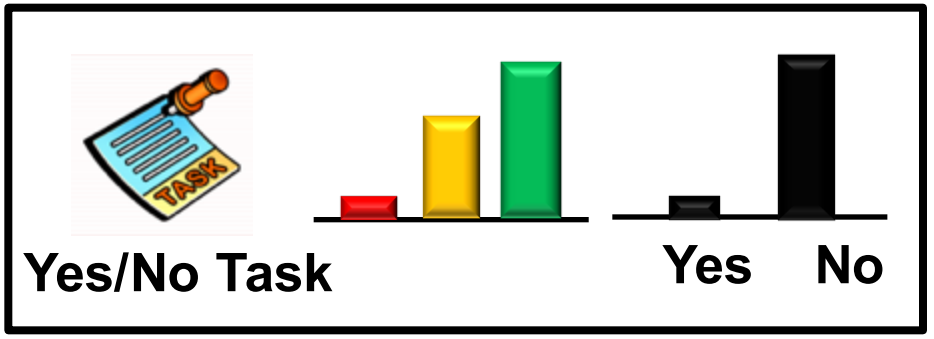
- **Select the most suitable tasks for assignment**

Sports Politics Entertainment









(1) Matching Domains

(2) Answer Uncertainty

Experiments

- **Dataset Setting (1)**

Name	#Tasks	Domains	Description	Example
D_Item: ItemCompare Dataset	360	NBA, Food, Auto, Country	It asks workers to compare between two items	Which food contains more calories, Chocolate or Honey?
D_4D: 4 Domain Dataset	400	NBA, Car, Film, Mountain	It asks workers about tasks on a certain domain	Did Michael Jordan win more NBA championships than Kobe Bryant?

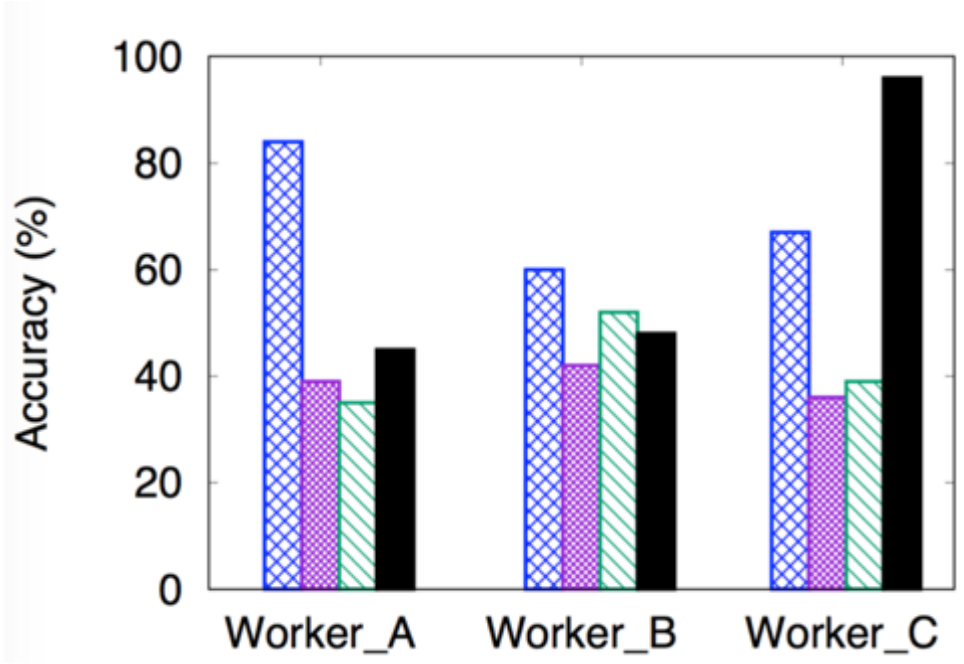
Experiments

- **Dataset Setting (2)**

Name	#Tasks	Domains	Description	Example
D_QA: Yahoo QA Dataset	1000	Domains in Yahoo Answers	It asks workers tasks on Yahoo Answers	Where does chili originate from, Texas or Turkey?
D_SFV: SFV Dataset (a NLP dataset)	328	Domains in Yahoo Answers	It asks workers the attribute of a person, where the answers are collected from different QA systems	What is the age of Bill Gates?

Experiments

- **Worker Characteristics on Dataset D_Item**

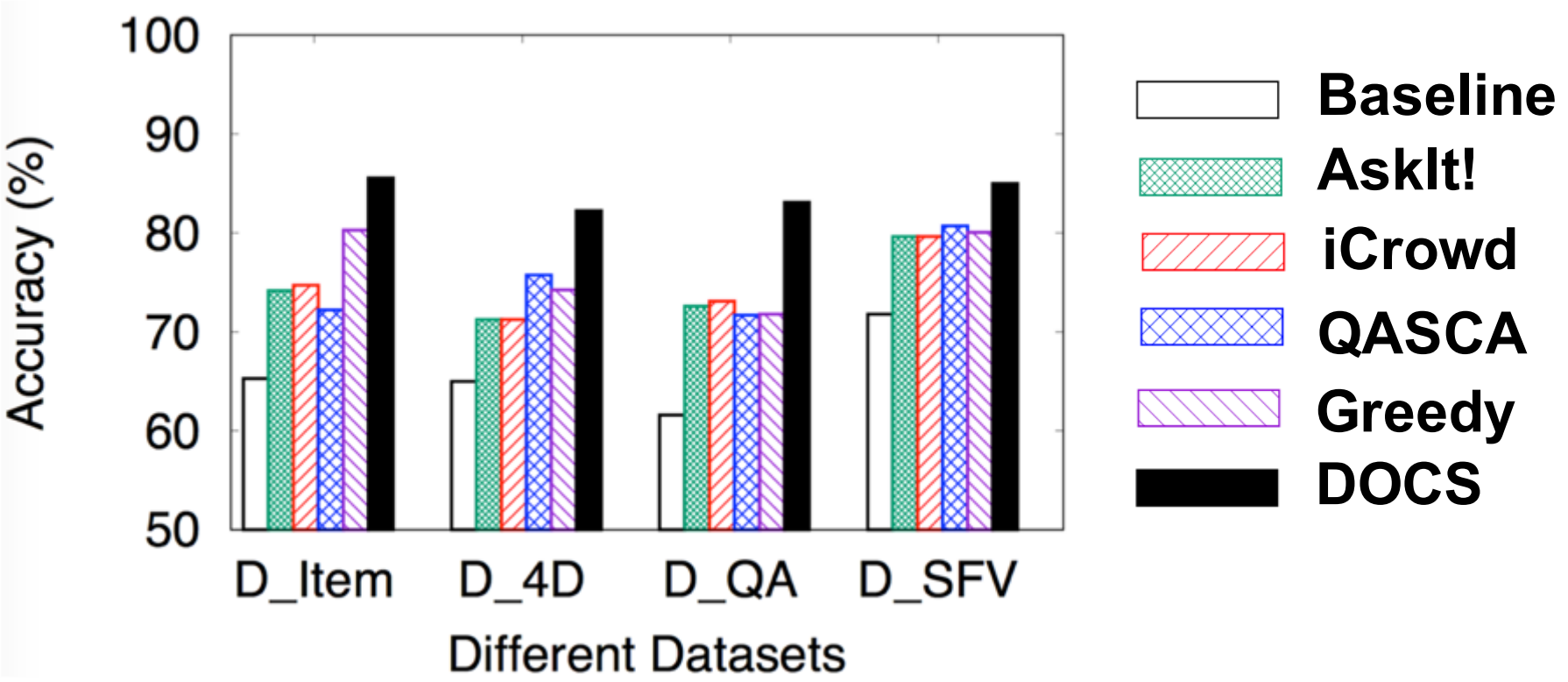


Each worker has diverse quality
Over different domains

Estimated worker quality is close
to real worker quality

Experiments

- **System Comparisons**



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.
iCrowd: J. Fan, G. Li, B. C. Ooi, K. Tan, and J. Feng. icrowd: An adaptivecrowdsourcing framework. In SIGMOD, pages 1015–1030, 2015.
QASCA: Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, Jianhua Feng. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications. SIGMOD 2015.

Summary

- **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**

Yudian Zheng, Guoliang Li, Reynold Cheng

University of Hong Kong, Tsinghua University