KDD2018

# Crowd-Powered Data Mining

Chengliang Chai
Tsinghua University

Ju Fan
Renmin University

Guoliang Li
Tsinghua University

Jiannan Wang
SFU

Yudian Zheng
Twitter

# Outline

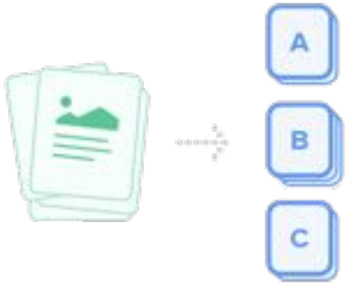
- Crowdsourcing Overview (20min)
- Fundamental Techniques (90min)
  - Quality Control (40min)
  - Cost Control (30min)
  - Latency Control (20min)
- Crowd-powered Data Mining (60min)
  - Crowd-powered Pattern Mining (10min)
  - Crowd-powered Classification (10min)
  - Crowd-powered Clustering (10min)
  - Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - Crowd-powered Knowledge Discovery (20min)
- Challenges (10min)

Part 1

Part 2

# Crowdsourcing：Motivation

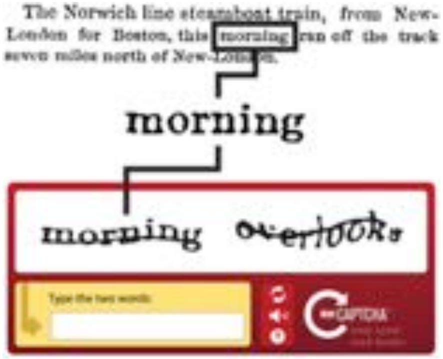
- A new computation model
  - Coordinating the crowd (Internet workers) to do micro-tasks in order to solve computer-hard problems.
- Examples ebay
  - Categorize the products and create product taxonomies from the user’s standpoint.
  - An example question
    - Select the product category of Samsung S7
      - Phone
      - TV
      - Movie



# Crowdsourcing：Applications

- Wikipedia
  - Collaborative knowledge

- reCAPTCHA
  - Digitalizing newspapers


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

- Foldit
  - fold the structures of selected proteins

Click to learn how you contribute to science by playing Foldit.

- App Testing
  - Test apps

# Crowdsourcing: Popular Tasks

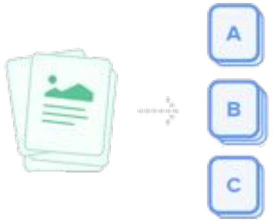
- **Sentiment Analysis**
  - Understand conversation: positive/negative

- **Search Relevance**
  - Return relevant results on the first search

- **Content Moderation**
  - Keep the best, lose the worst

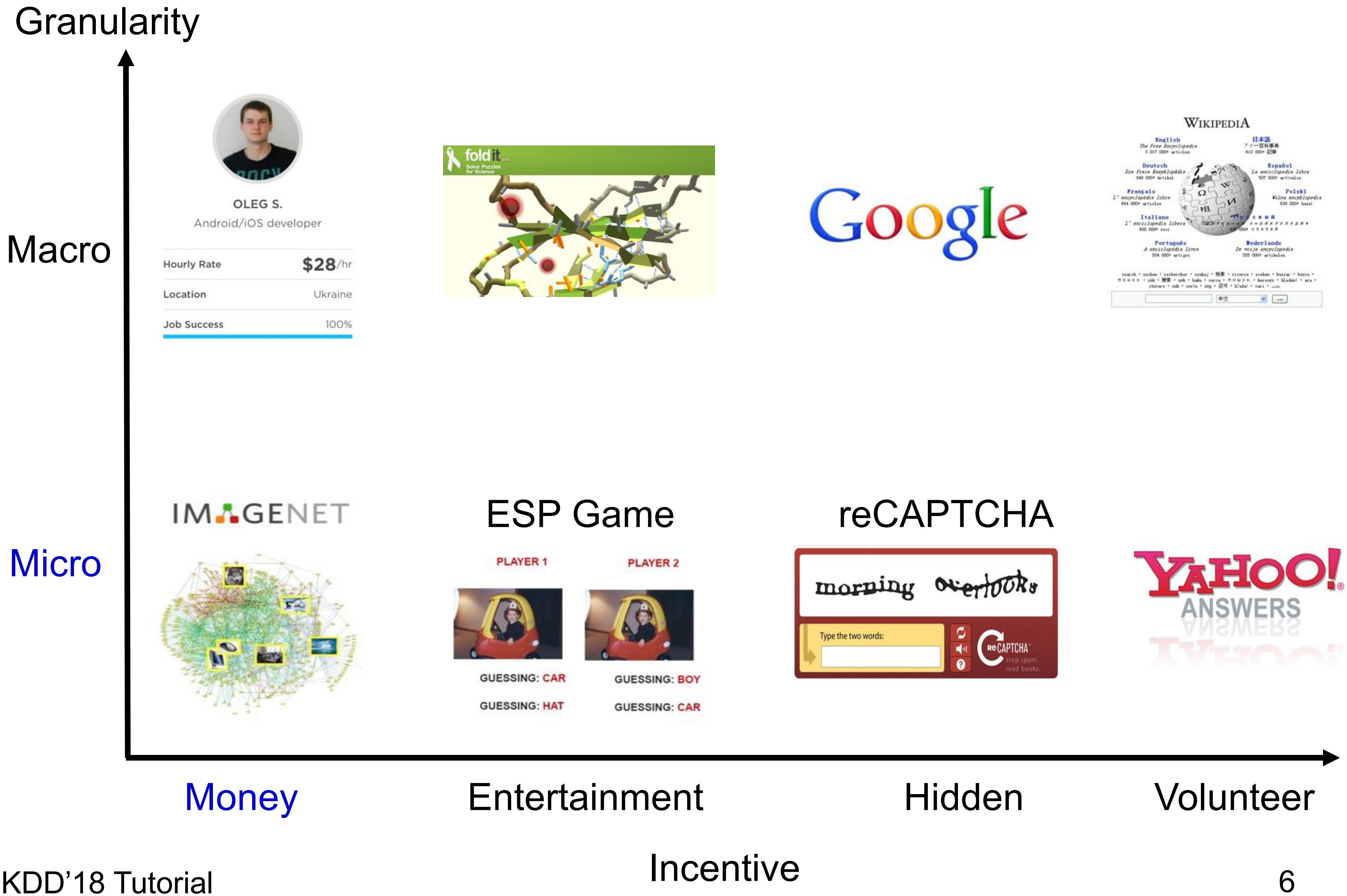
- **Data Collection**
  - Verify and enrich your business data

- **Data Categorization**
  - Organize your data



- **Transcription**
  - Turn images and audio into useful data

# Crowdsourcing Space

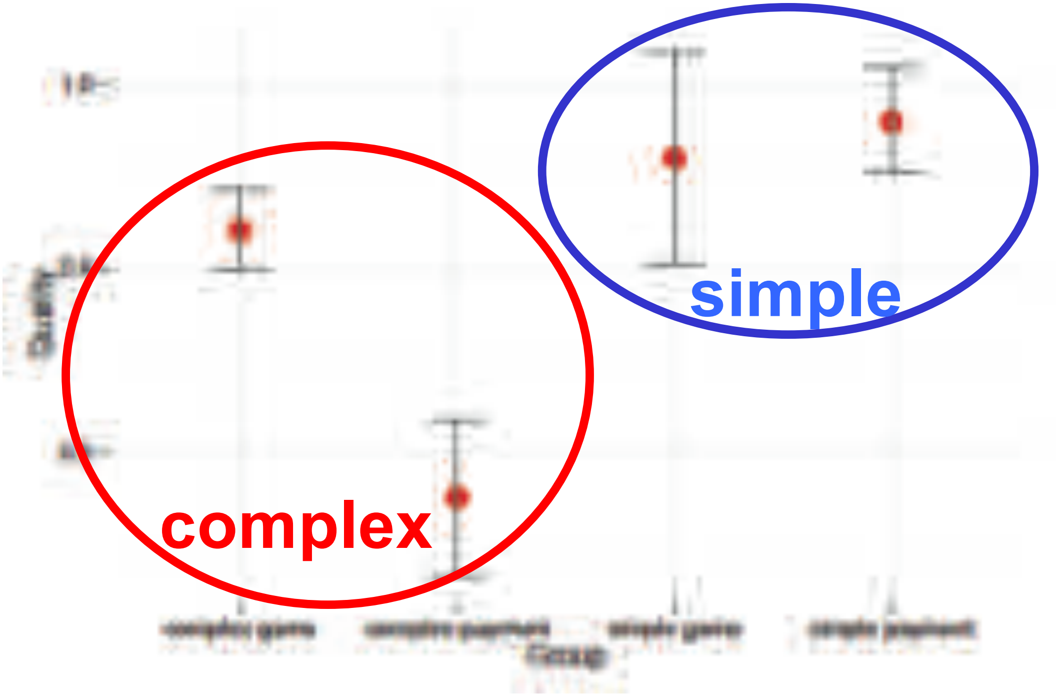


OLEG S.
Android/iOS developer

| | |
|---|---|
| Hourly Rate | $28/hr |
| Location | Ukraine |
| Job Success | 100% |

# Crowdsourcing Category

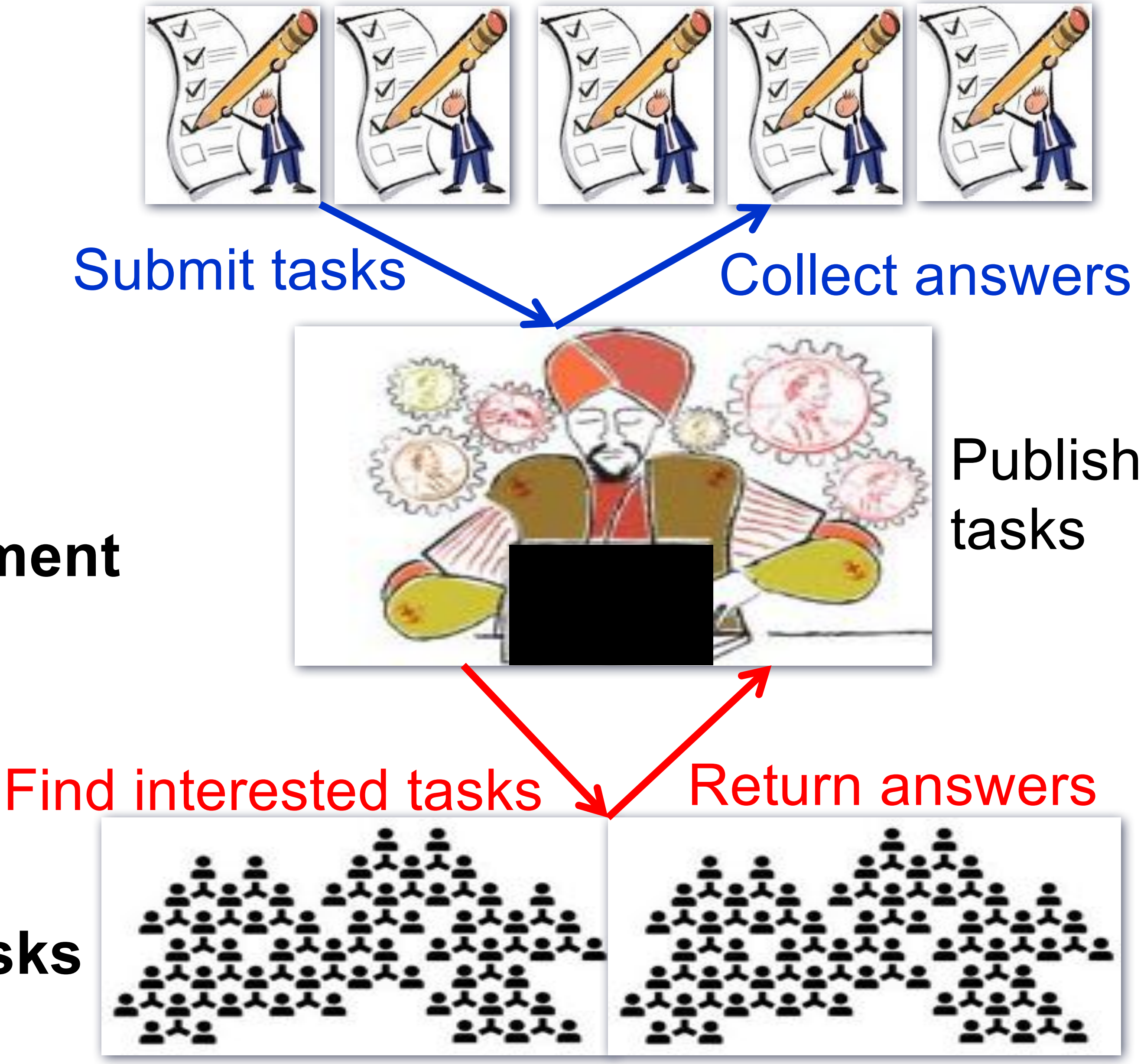
- ○ Game vs Payment
  - – Simple tasks
    - Both payment and game can achieve high quality
  - – Complex tasks
    - Game has better quality



Quality is rather important!

# Crowdsourcing：Workflow

- Requester
  - – Submit Tasks
- Platforms
  - – Task Management
- Workers
  - – Worker on Tasks



[redacted]

Submit tasks

Collect answers

Publish tasks

Find interested tasks

Return answers

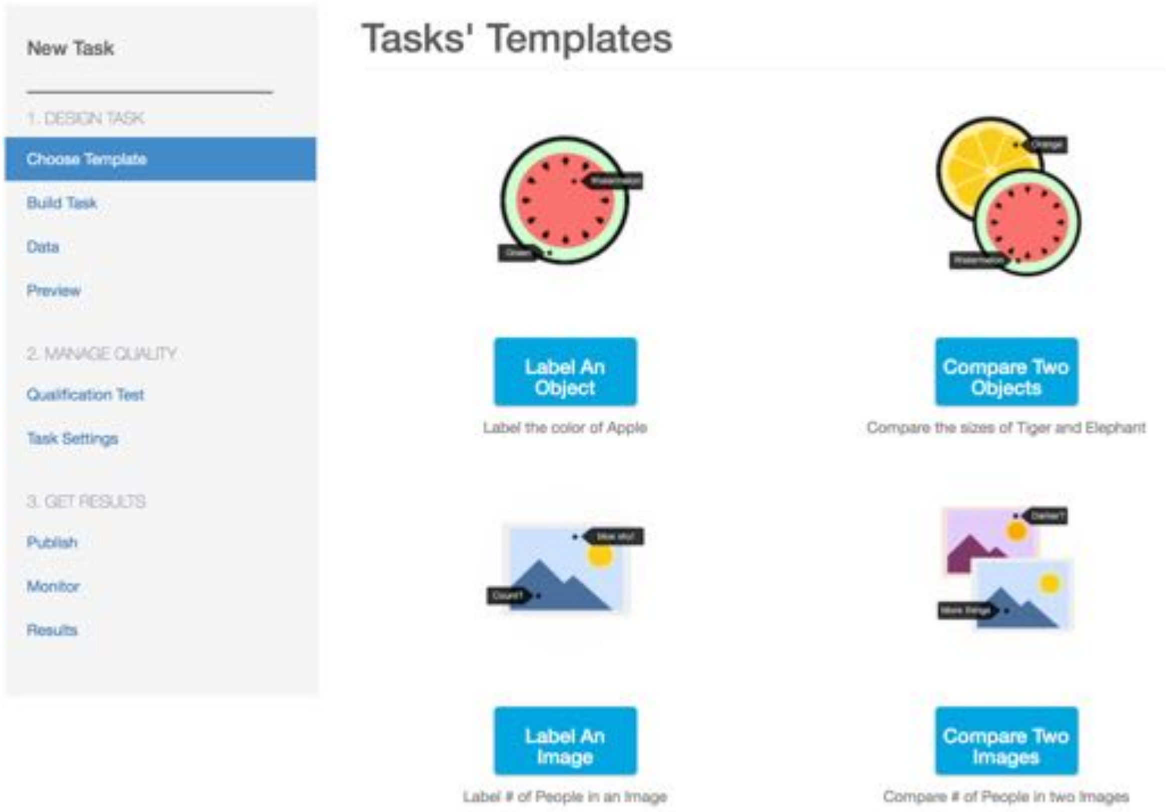
# Crowdsourcing Requester：Workflow

○ **Design Tasks**
  - **Task Type**
  - **Design Strategies**
    - – UI, API, Coding

○ **Upload Data**

○ **Set Tasks**
  - **Price**
  - **Time**
  - **Quality**

○ **Publish Task**
  - **Pay**
  - **Monitor**



Tasks' Templates

Label the color of Apple

Compare the sizes of Tiger and Elephant

Label # of People in an Image

Compare # of People in two Images

# Crowdsourcing Requester：Task Type

## ○ Task Type

Please choose the brand of the phone

- ○ [✓ Apple]
- ○ Samsung
- ○ Blackberry
- ○ Other

What are comment features?

- □ Same band
- □ Same color
- ☑ Similar price
- ☑ Same size

Please fill the attributes of the product

Brand □

Price □

Size □

Camera □

Please submit a picture of a phone with the same size as the left one.

Submit

# Crowdsourcing Requester: Task Design

○ **UI**

Choose the best category for the image

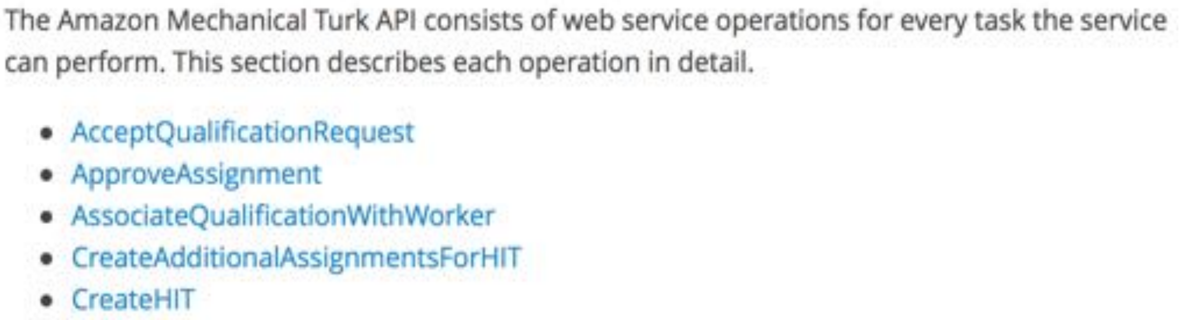
○ Kitchen

○ Bath

○ Living

○ Bed

○ **API**



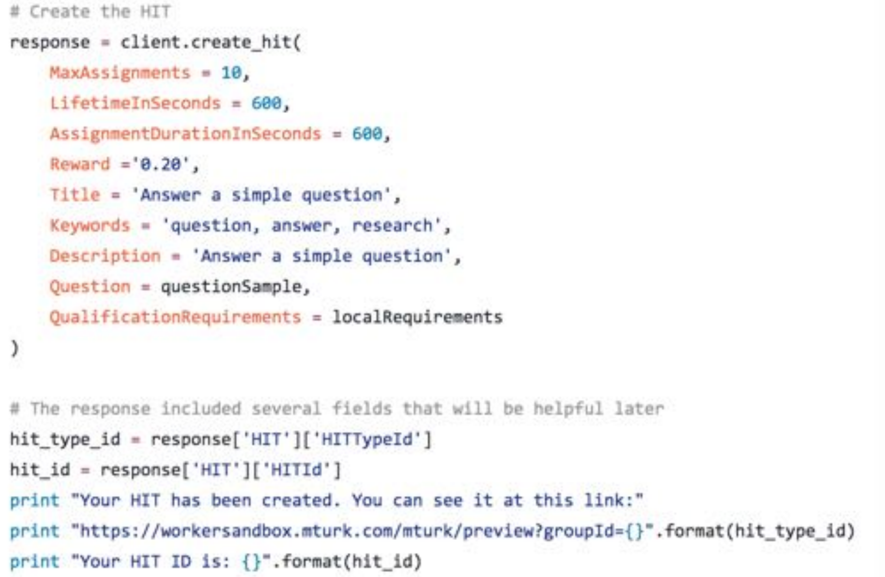
The Amazon Mechanical Turk API consists of web service operations for every task the service can perform. This section describes each operation in detail.

- AcceptQualificationRequest
- ApproveAssignment
- AssociateQualificationWithWorker
- CreateAdditionalAssignmentsForHIT
- CreateHIT

○ **Coding**

**(Your own Server)**

**innerhtml**

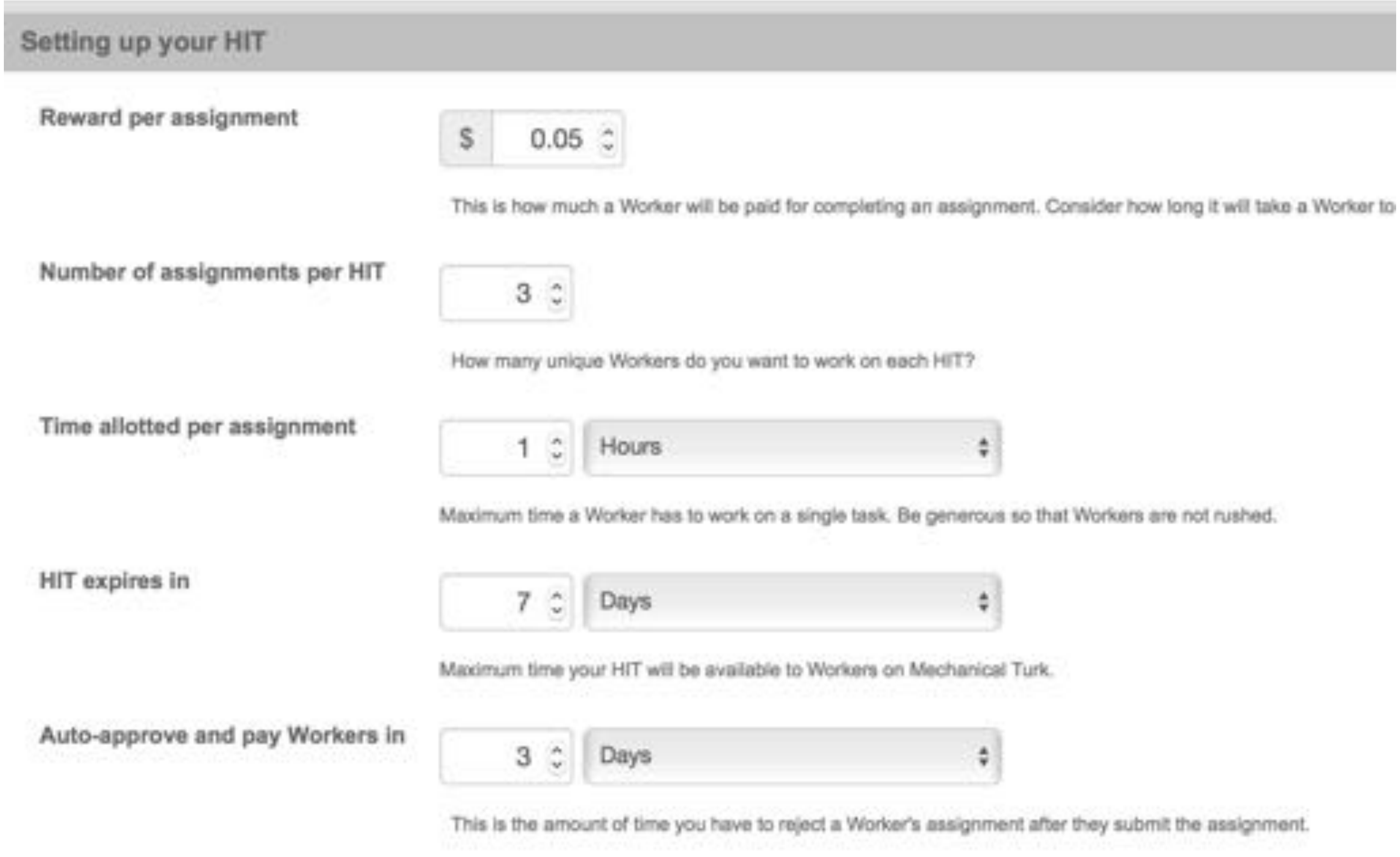


```
# Create the HIT
response = client.create_hit(
    MaxAssignments = 10,
    LifetimeInSeconds = 600,
    AssignmentDurationInSeconds = 600,
    Reward ='0.20',
    Title = 'Answer a simple question',
    Keywords = 'question, answer, research',
    Description = 'Answer a simple question',
    Question = questionSample,
    QualificationRequirements = localRequirements
)

# The response included several fields that will be helpful later
hit_type_id = response['HIT']['HITTypeId']
hit_id = response['HIT']['HITId']
print "Your HIT has been created. You can see it at this link:"
print "https://workersandbox.mturk.com/mturk/preview?groupId={}".format(hit_type_id)
print "Your HIT ID is: {}".format(hit_id)
```

# Crowdsourcing Requester: Task Setting

- ○ **HIT – A group of micro-tasks (e.g., 5)**
- ○ **Price, Assignment, Time**



Setting up your HIT

Reward per assignment: $ 0.05

This is how much a Worker will be paid for completing an assignment. Consider how long it will take a Worker to

Number of assignments per HIT: 3

How many unique Workers do you want to work on each HIT?

Time allotted per assignment: 1 Hours

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

HIT expires in: 7 Days

Maximum time your HIT will be available to Workers on Mechanical Turk.

Auto-approve and pay Workers in: 3 Days

This is the amount of time you have to reject a Worker's assignment after they submit the assignment.

# Crowdsourcing Requester: Task Setting

- ○ **Quality Control**

  – **Qualification test - Quiz**

  Create some test questions to enable a quiz that workers must pass to work on your task.

  – **Hidden test - Training**

  Add some questions with ground truths in your task so workers who get them wrong will be eliminated.

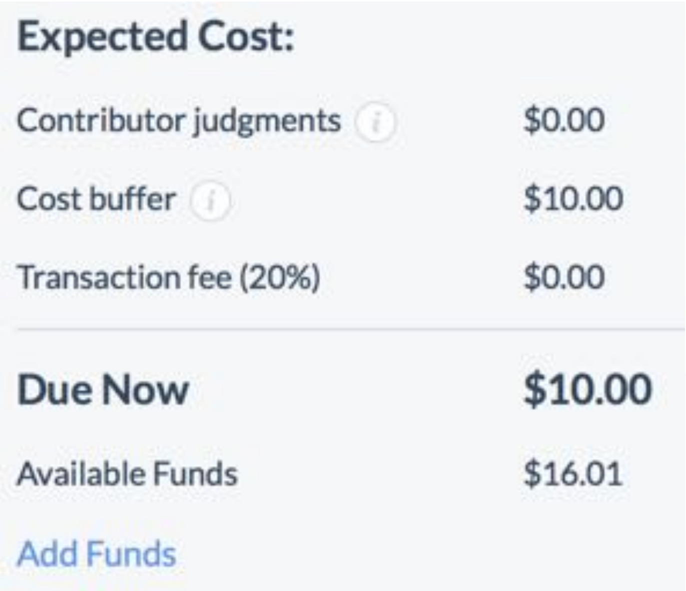
  – **Worker selection**

  Ensure high-quality results by eliminating workers who repeatedly fail test questions in your task

# Crowdsourcing Requester: Publish

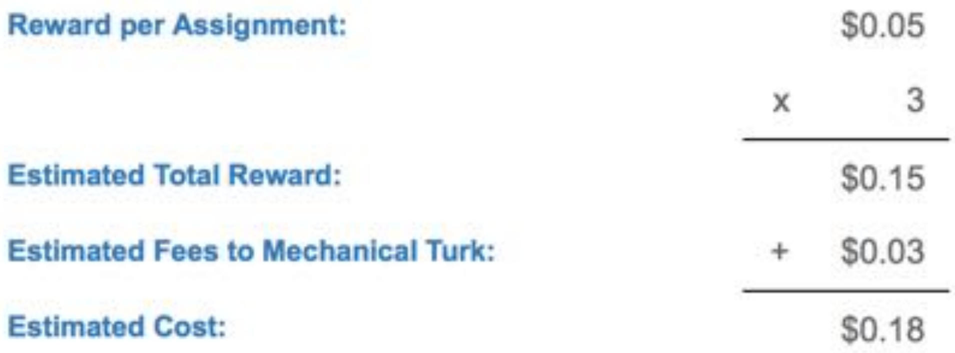
- **Prepay**

cost for workers + cost for platform +cost for test



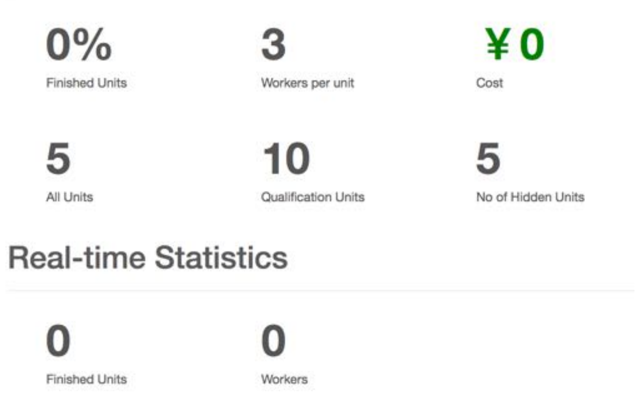
**Expected Cost:**

| | |
|---|---|
| Contributor judgments ⓘ | $0.00 |
| Cost buffer ⓘ | $10.00 |
| Transaction fee (20%) | $0.00 |
| **Due Now** | **$10.00** |
| Available Funds | $16.01 |
| Add Funds | |



| | | |
|---|---|---|
| Reward per Assignment: | | $0.05 |
| | x | 3 |
| Estimated Total Reward: | | $0.15 |
| Estimated Fees to Mechanical Turk: | + | $0.03 |
| Estimated Cost: | | $0.18 |

- **Monitor**



| | | |
|---|---|---|
| 0%<br>Finished Units | 3<br>Workers per unit | ¥0<br>Cost |
| 5<br>All Units | 10<br>Qualification Units | 5<br>No of Hidden Units |

Real-time Statistics

| | |
|---|---|
| 0<br>Finished Units | 0<br>Workers |

# Crowdsourcing: Workers

- Task Selection
- Task Completion
- Workers are not free **Cost**
  - Make Money

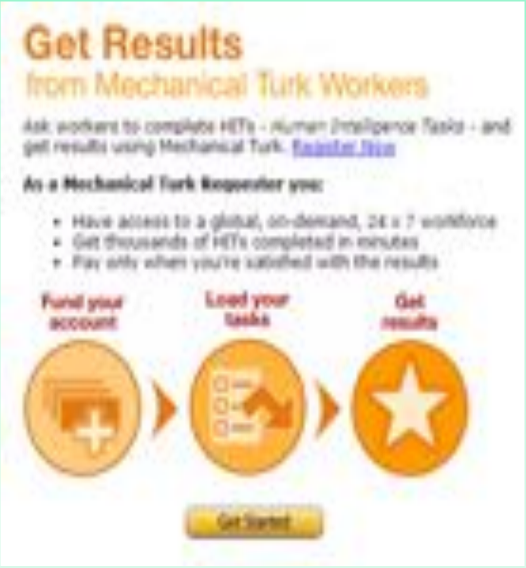
- Workers are not oracle **Quality**
  - Make errors
  - Malicious workers

- Workers are dynamic **Latency**
  - Hard to predict

# Crowdsourcing： Platforms

○ **Amazon Mechanical Turk (AMT)**

□ **Requesters**



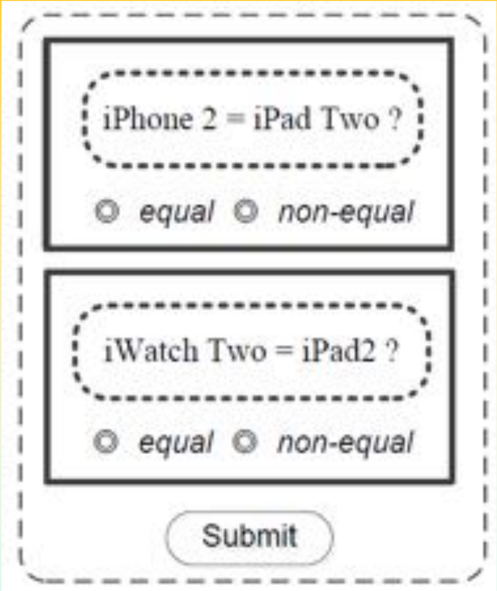
Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

**As a Mechanical Turk Requester you:**

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

□ **HIT** **(k tasks)**



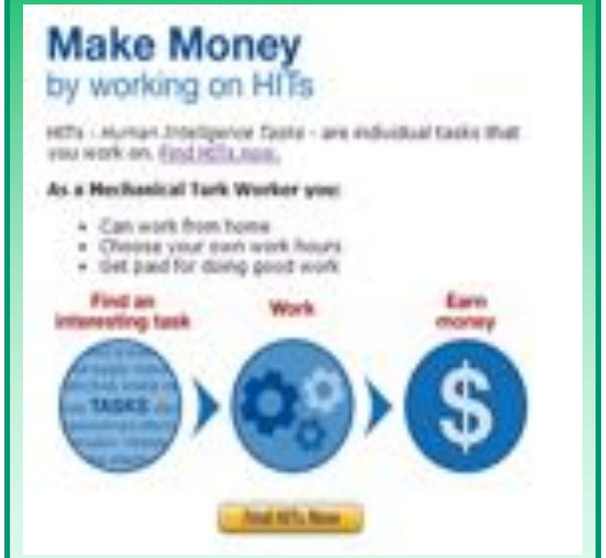
iPhone 2 = iPad Two ?

○ *equal* ○ *non-equal*

iWatch Two = iPad2 ?

○ *equal* ○ *non-equal*

□ **Workers**



Make Money
by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

**As a Mechanical Turk Worker you:**

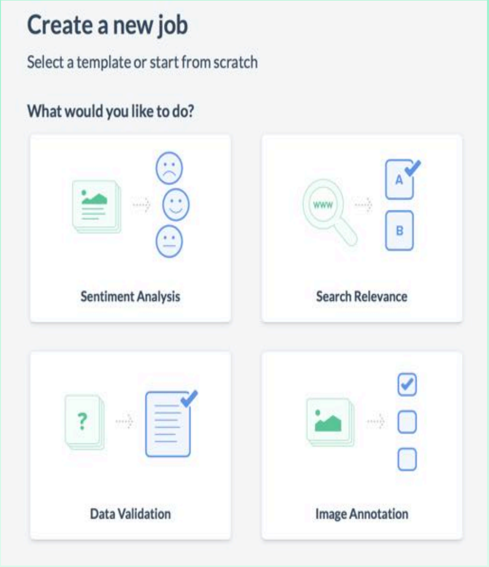
- Can work from home
- Choose your own work hours
- Get paid for doing good work

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

# Crowdsourcing： Platforms

○ CrowdFlower

□ Requesters

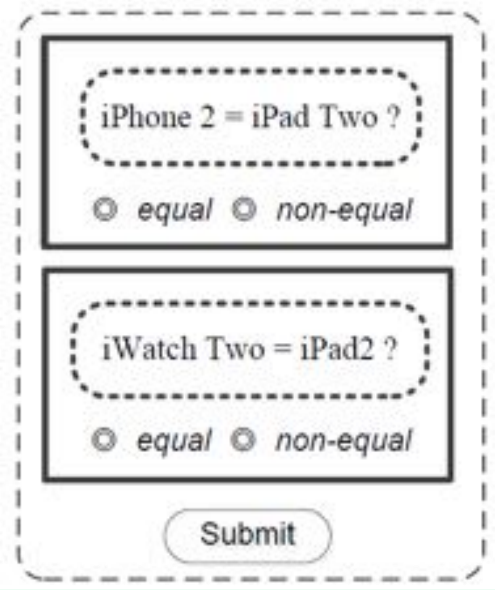


Create a new job

Select a template or start from scratch

What would you like to do?

□ HIT (k tasks)



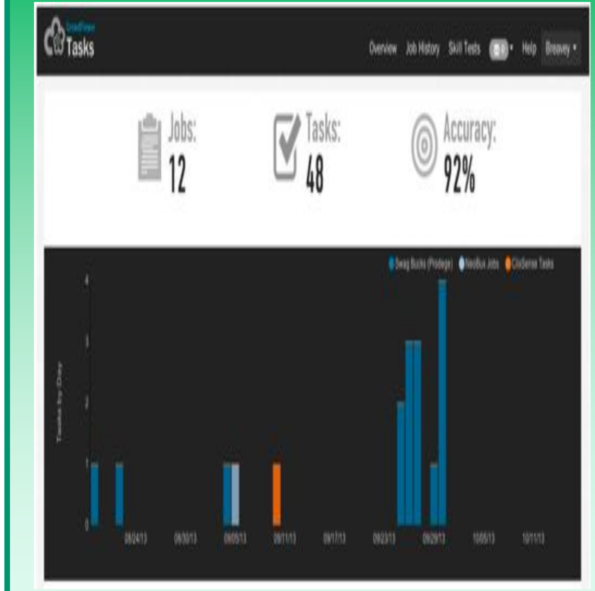
iPhone 2 = iPad Two ?

○ equal ○ non-equal

iWatch Two = iPad2 ?

○ equal ○ non-equal

□ Workers



Jobs: 12

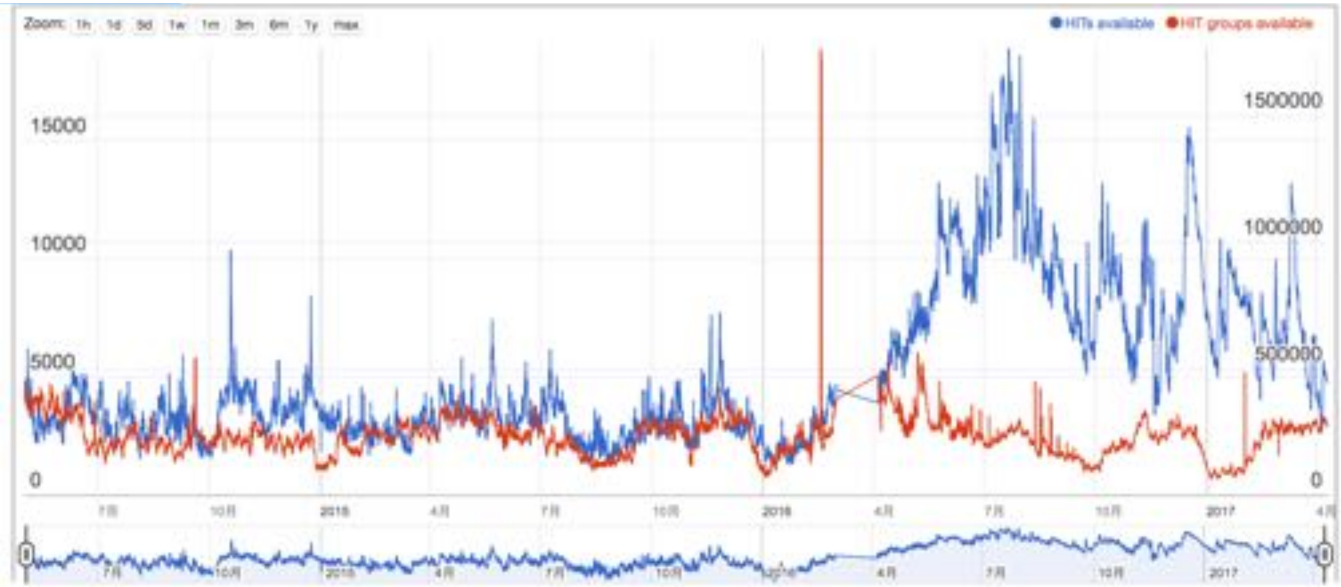
Tasks: 48

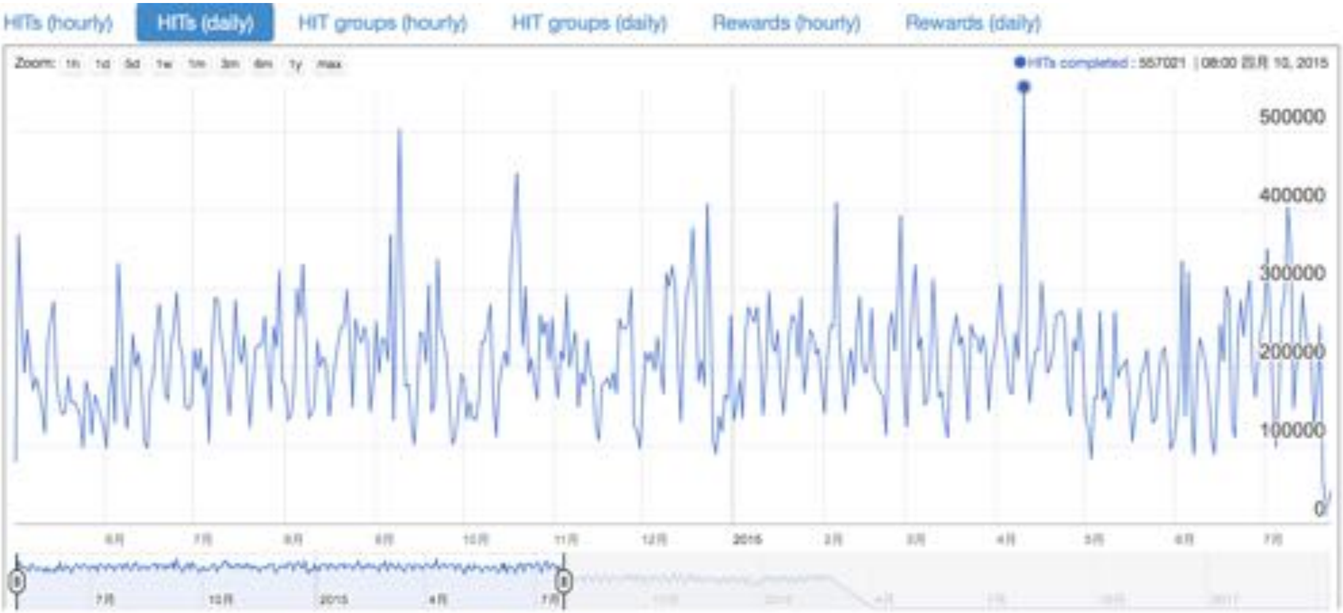
Accuracy: 92%

# AMT vs CrowdFlower

| | AMT | CrowdFlower |
|---|---|---|
| Task Design: UI | √ | √ |
| Task Design: API | √ | √ |
| Task Design: Coding | √ | × |
| Quality: Qualification Test | √ | √ |
| Quality: Hidden Test | × | √ |
| Quality: Worker Selection | √ | √ |
| Task Types | All Types | All Types |

# AMT Task Statistics

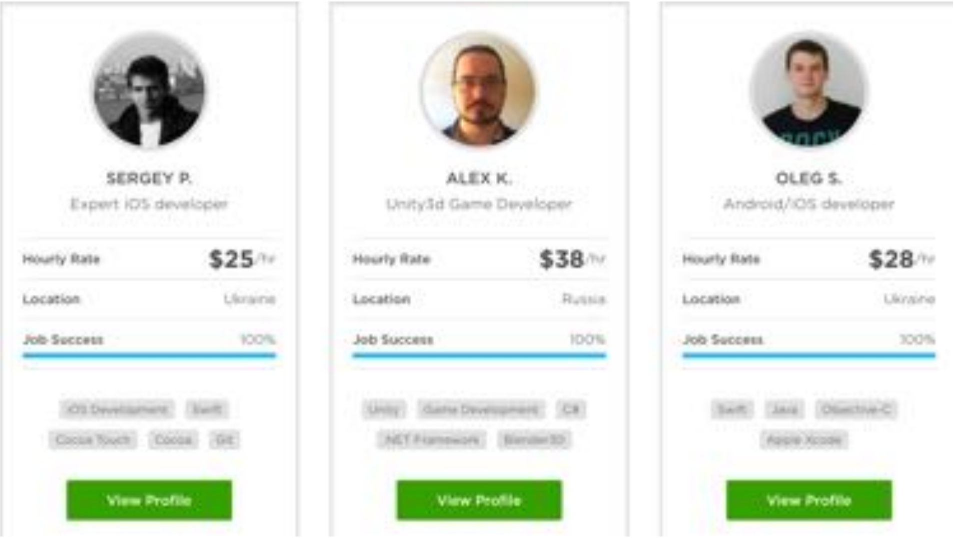




http://www.mturk-tracker.com

# Other Crowdsourcing Platforms

- **Macrotask**
  - **Upwork**
    - https://www.upwork.com
  - **Zhubajie**
    - http://www.zbj.com



SERGEY P.
Expert iOS developer
Hourly Rate $25/hr
Location Ukraine
Job Success 100%

ALEX K.
Unity3d Game Developer
Hourly Rate $38/hr
Location Russia
Job Success 100%

OLEG S.
Android/iOS developer
Hourly Rate $28/hr
Location Ukraine
Job Success 100%

- **Microtask**
  - **ChinaCrowds (cover all features of AMT and CrowdFlower)**
    - http://www.chinacrowds.com

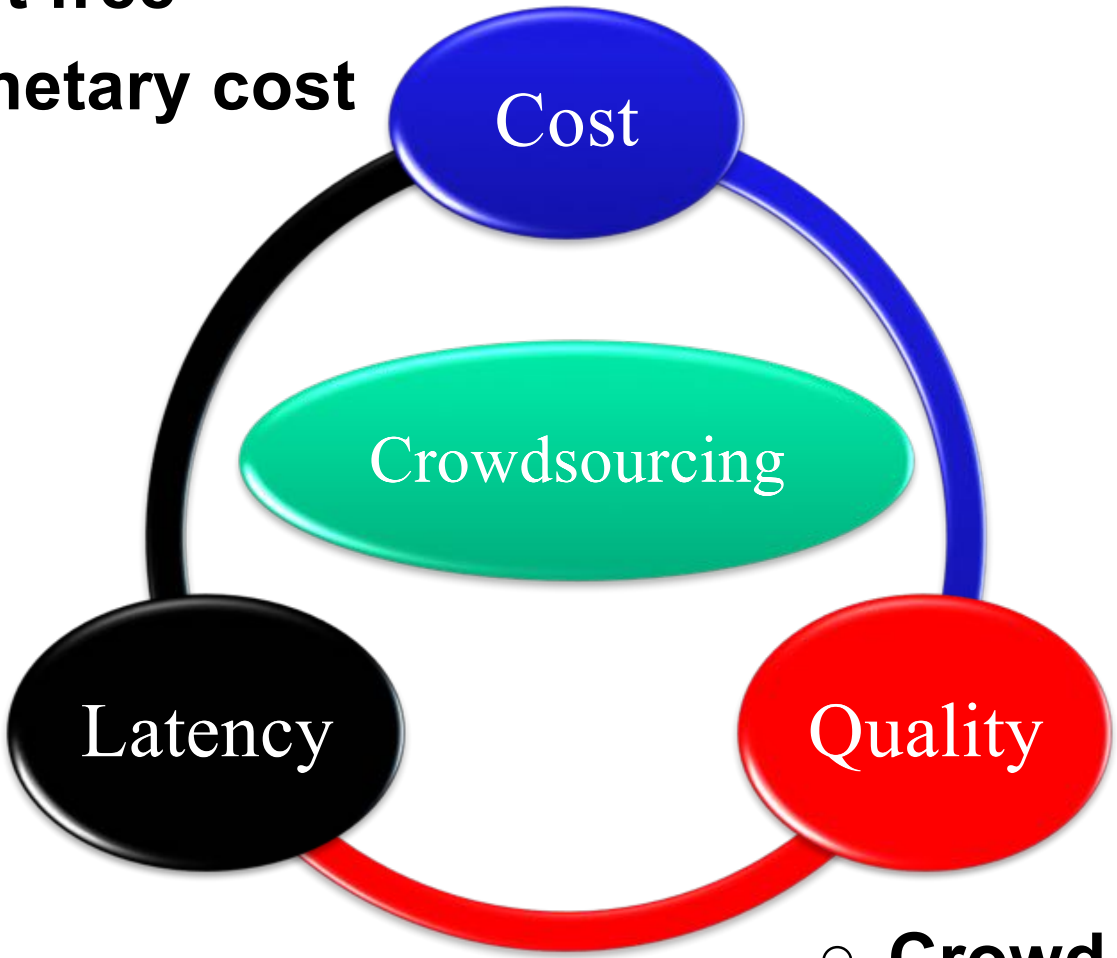
iOS

Android

# Crowdsourcing：Challenges

○ Crowd is not free

○ Reduce monetary cost



○ Crowd is not real-time

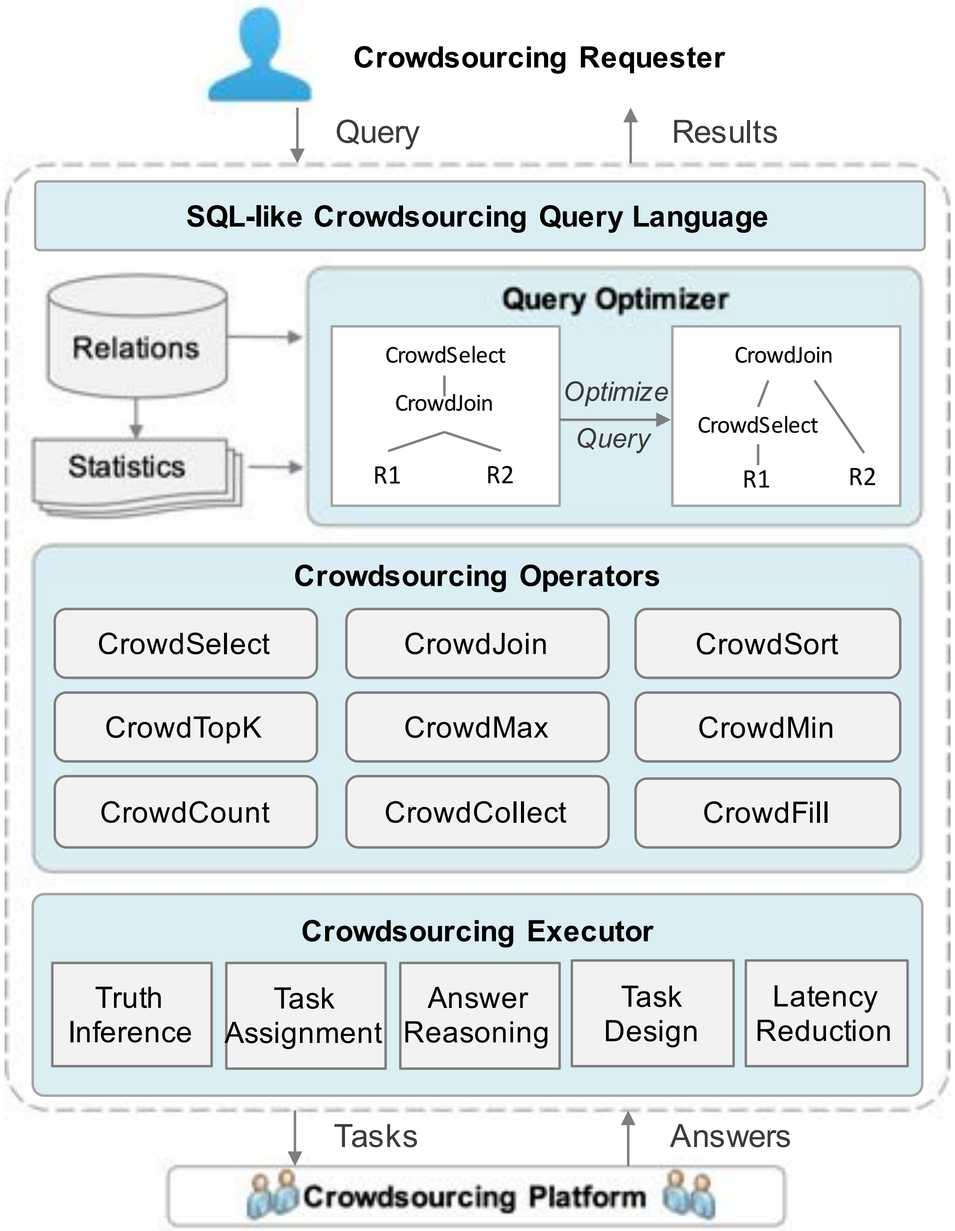
○ Reduce time

○ Crowd may return incorrect answers

○ Improve quality

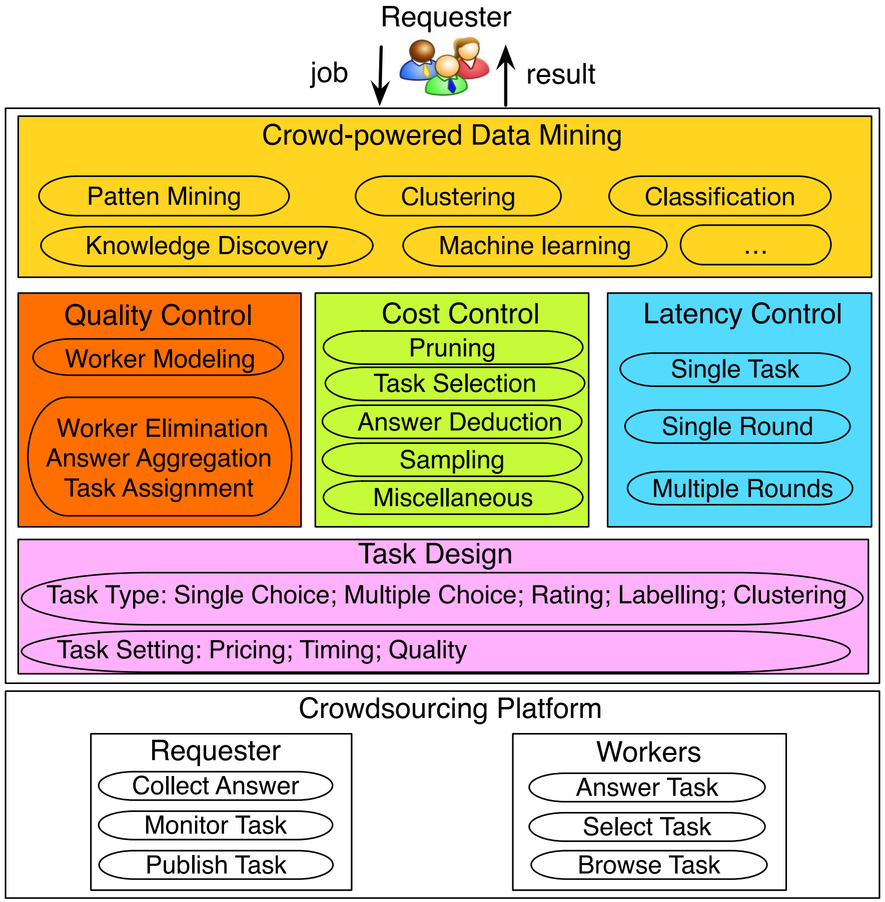
# Crowdsourced Data Management

- **A crowd-powered database system**
  - Users require to write code to utilize crowdsourcing platforms
  - Encapsulates the complexities of interacting with the crowd
  - Make DB more powerful
- **Crowd-powered interface**
- **Crowd-powered Operators**
- **Crowdsourcing Optimization**



# Crowdsourced Data Mining

- Fundamental Optimization
  - Quality Control
  - Cost Control
  - Latency Control
- Crowd-powered Data Mining
  - Classification
  - Cluster
  - Pattern Mining
  - Knowledge Discovery
  - Machine Learning



# Differences with Existing Tutorials

- **SIGMOD' 17**
  - Control quality, cost and latency
  - Design crowdsourced database
- **VLDB'16**
  - Human factors involved in task assignment and completion.
- **VLDB'15**
  - Truth inference in quality control
- **ICDE'15**
  - Individual crowdsourcing operators, crowdsourced data mining and social applications
- **VLDB'12**
  - Crowdsourcing platforms and Design principles
- **Our Tutorial**
  - Crowd-powered data mining

# Outline

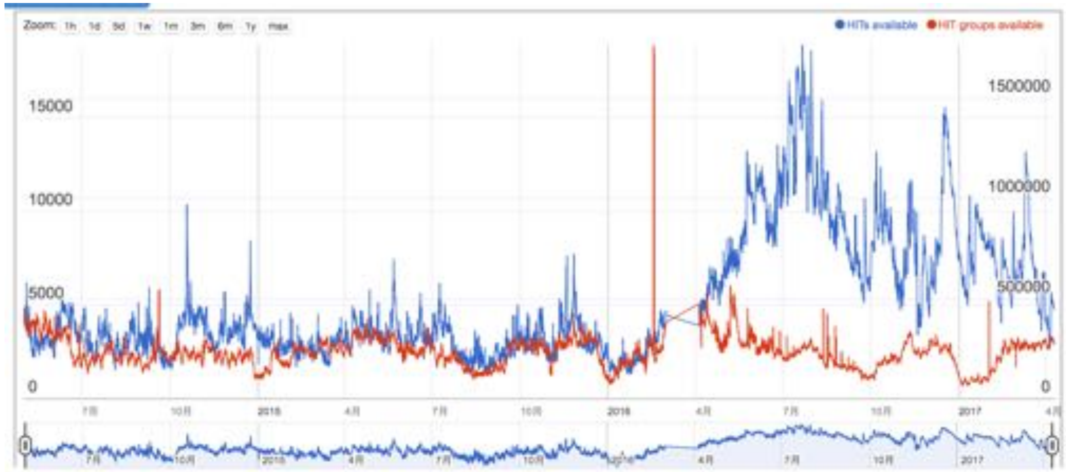
- Crowdsourcing Overview (20min)
- Fundamental Techniques (90min)
  - Quality Control (40min)
  - Cost Control (30min)
  - Latency Control (20min)
- Crowd-powered Data Mining (60min)
  - Crowd-powered Pattern Mining (10min)
  - Crowd-powered Classification (10min)
  - Crowd-powered Clustering (10min)
  - Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - Crowd-powered Knowledge Discovery (20min)
- Challenges (10min)

Part 1

Part 2

# Why Quality Control?

- ○ Huge Amount of Crowdsourced Data



amazonmechanical turk
beta
Artificial Artificial Intelligence

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

- ○ Inevitable noise & error

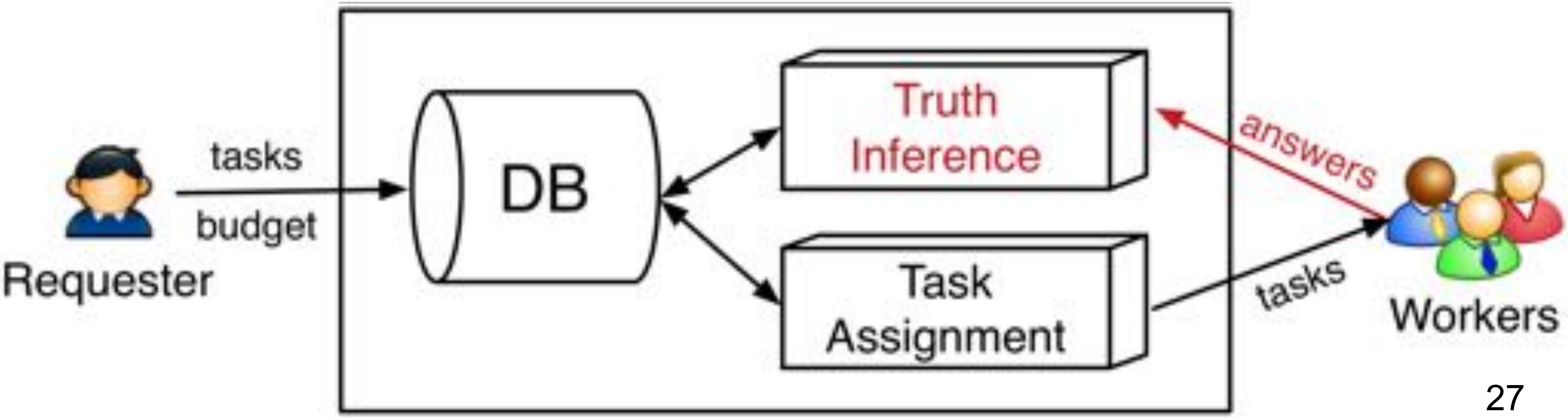
- ○ Goal: Obtain reliable information in Crowdsourced Data

# Crowdsourcing Workflow

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

(1) when a worker comes to the platform, the worker will be assigned to a set of tasks (task assignment);

(2) when a worker accomplishes tasks, the platform will collect answers from the worker (truth inference).



# Outline of Quality Control

☞ ○ **Part I. Truth Inference**

- – Problem Definition
- – Condition 1: with ground truth
  - Qualification Test & Hidden Test
- – Condition 2: without ground truth
  - Unified Framework
  - Differences in Existing Works
  - Experimental Results

○ **Part II. Task Assignment**

- – Problem Definition
- – Differences in Existing Works

# Part I. Truth Inference

- An Example Task

**What is the current affiliation for Michael Franklin ?**

**A. University of California, Berkeley**
**B. University of Chicago**

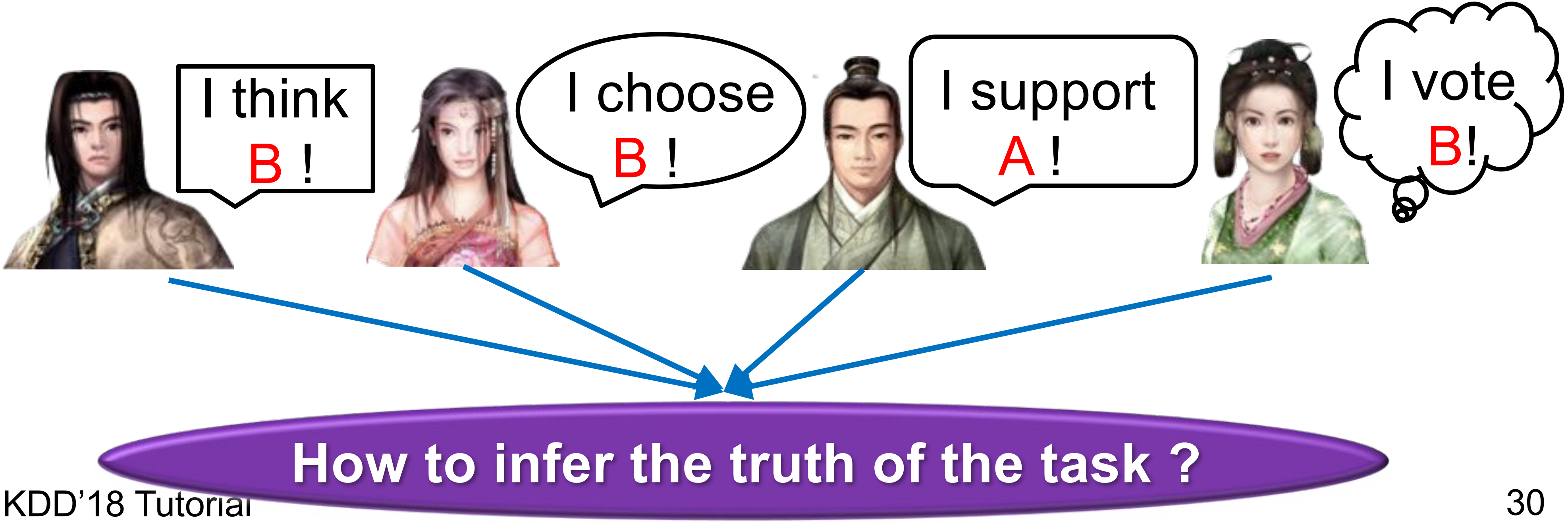
I support
A. UCB !

# Principle: Redundancy

○ **Collect Answers from Multiple Workers**

**What is the current affiliation for Michael Franklin ?**

**A. University of California, Berkeley**
**B. University of Chicago**



**How to infer the truth of the task ?**

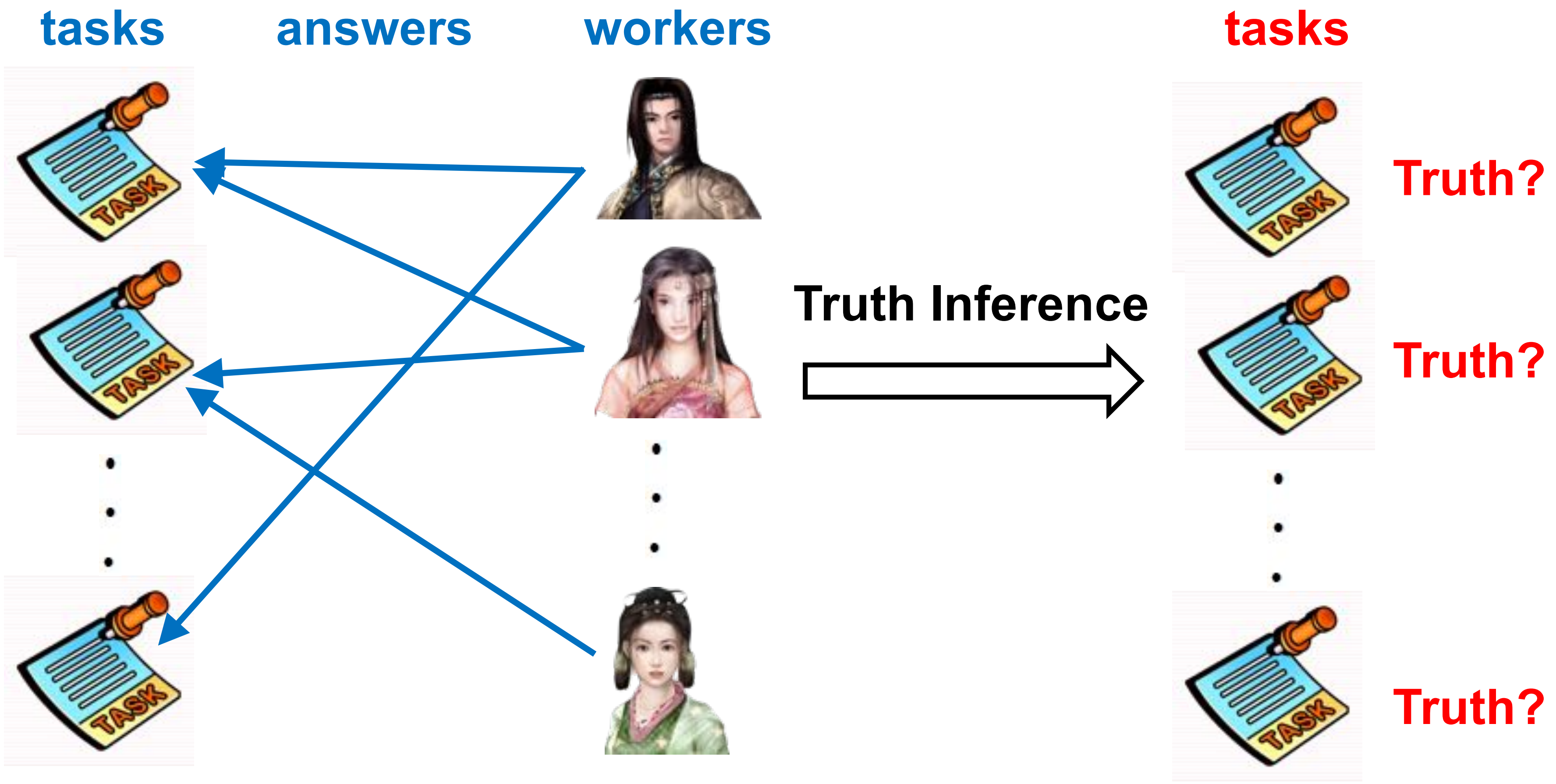
# Outline of Quality Control

- **Part I. Truth Inference**
  - ☞ **Problem Definition**
  - **Condition 1: with ground truth**
    - **Qualification Test & Hidden Test**
  - **Condition 2: without ground truth**
    - **Unified Framework**
    - **Differences in Existing Works**
    - **Experimental Results**

- **Part II. Task Assignment**
  - **Problem Definition**
  - **Differences in Existing Works**

# 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.**

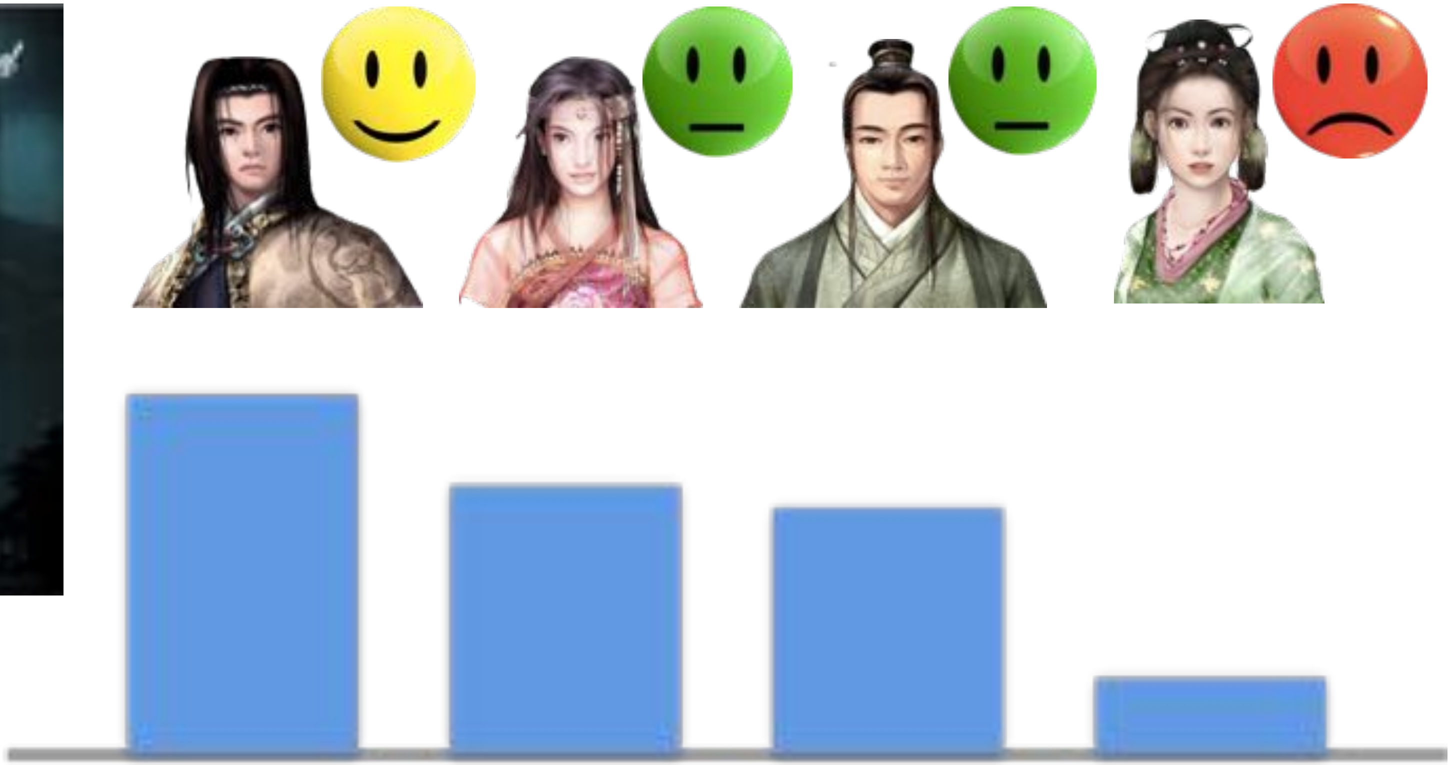
- **Limitation**

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

Expert
Good at Search
Spammer
Random Answer

# 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 ?

- 1. If a small set of tasks with ground truth are known in advance (e.g., refer to experts)

We can estimate each worker’s quality based on the *answering performance for the tasks with known truth*

- 2. If no ground truth is known in advance

The only way is to estimate each worker’s quality based on *the collected answers from all workers for all tasks*

# Outline

- Part I. Truth Inference
  - Problem Definition
  - ☞ Condition 1: with ground truth
    - Qualification Test & Hidden Test
  - Condition 2: without ground truth
    - Unified Framework
    - Existing Works
    - Experimental Results

- Part II. Task Assignment
  - Problem Definition
  - Differences in Existing Works

# 1. A Small Set of Ground Truth is Known

- **Qualification Test (*like an "exam"*)**

amazonmechanical turk
beta Artificial Artificial Intelligence

**Assign the tasks (with known truth) to the worker when the worker comes at first time**
***e.g., if the worker answers 8 over 10 tasks correctly, then the quality is 0.8***

- **Hidden Test (*like a "landmine"*)**

**Embed the tasks (with known truth) in all the tasks assigned to the worker**
***e.g., each time 10 tasks are assigned to a worker, then 10 tasks compose of 9 real tasks (with unknown truth), and 1 task with known truth***

# 1. A Small Set of Ground Truth is Known

○ **Limitations of two approaches** ⚠

**(1) need to know ground truth (may refer to experts);**

**(2) waste of money because workers need to answer these "extra" tasks;**

**(3) as reported (Zheng et al. VLDB'17), these techniques may not improve much quality.**

☞ ***Thus the assumption of "no ground truth is known" is widely adopted by existing works***

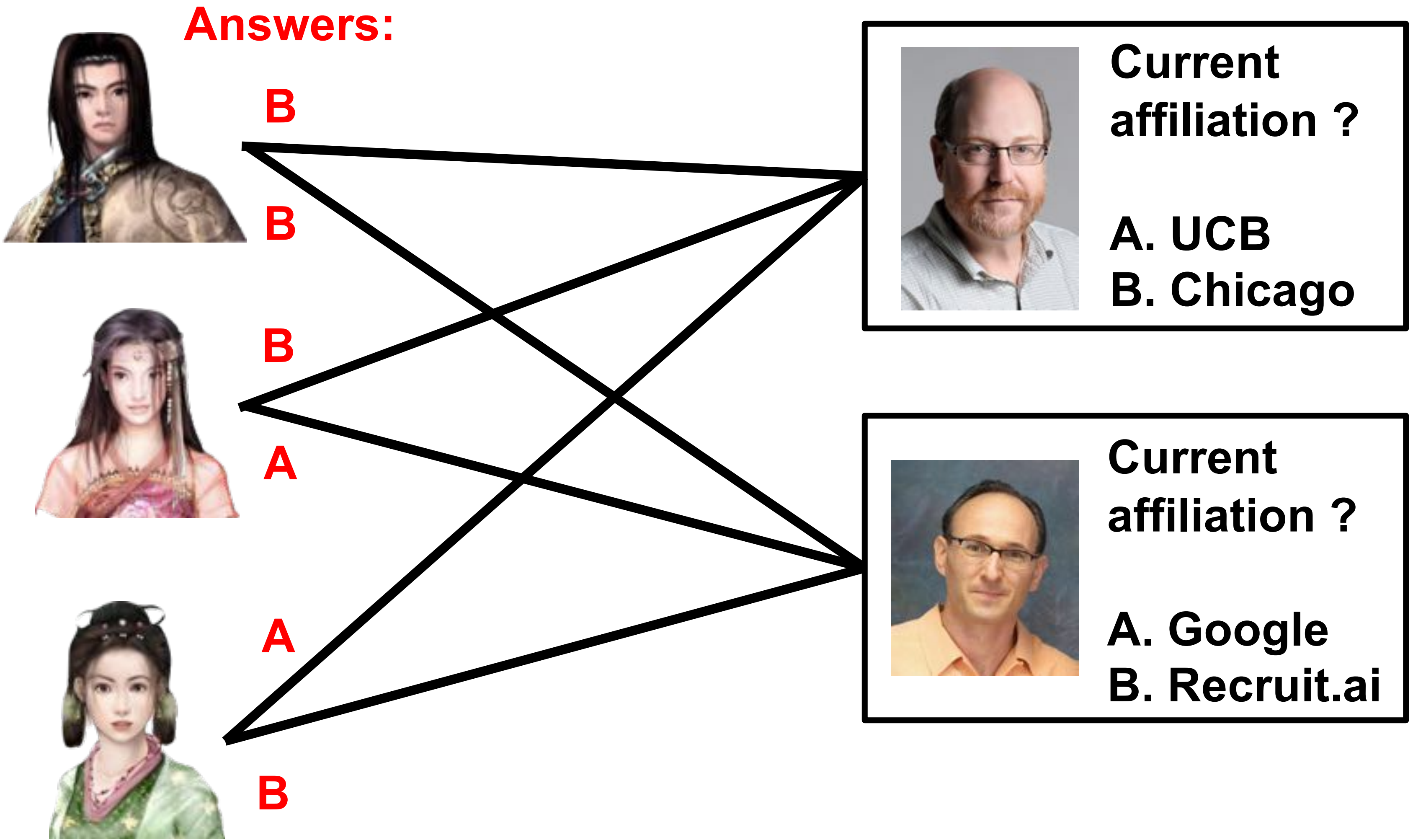
# Outline

- Part I. Truth Inference
  - Problem Definition
  - Condition 1: with ground truth
    - Qualification Test & Hidden Test
  - Condition 2: without ground truth
    - ☞ Unified Framework
    - Existing Works
    - Experimental Results

- Part II. Task Assignment
  - Problem Definition
  - Differences in Existing Works

# 2. If No Ground Truth is Known

- How to know each worker's quality given the collected answers for all tasks ?



Answers:

B

B

B

A

A

B

Current affiliation ?

A. UCB

B. Chicago

Current affiliation ?

A. Google

B. Recruit.ai

# Unified Framework in Existing Works

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

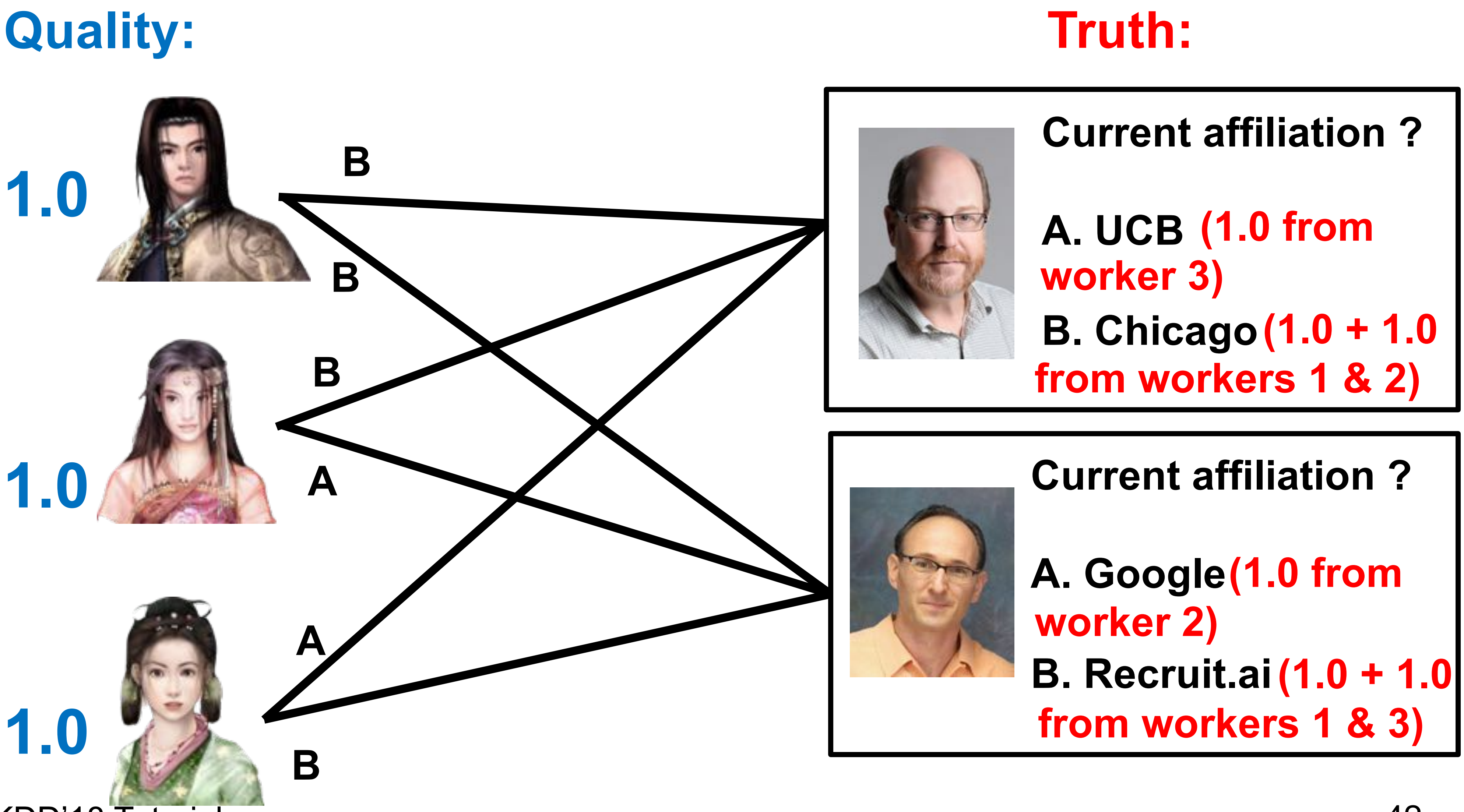
- **Algorithm Framework:**

```
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



Current affiliation ?

A. UCB (1.0 from worker 3)

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

Current affiliation ?

A. Google (1.0 from worker 2)

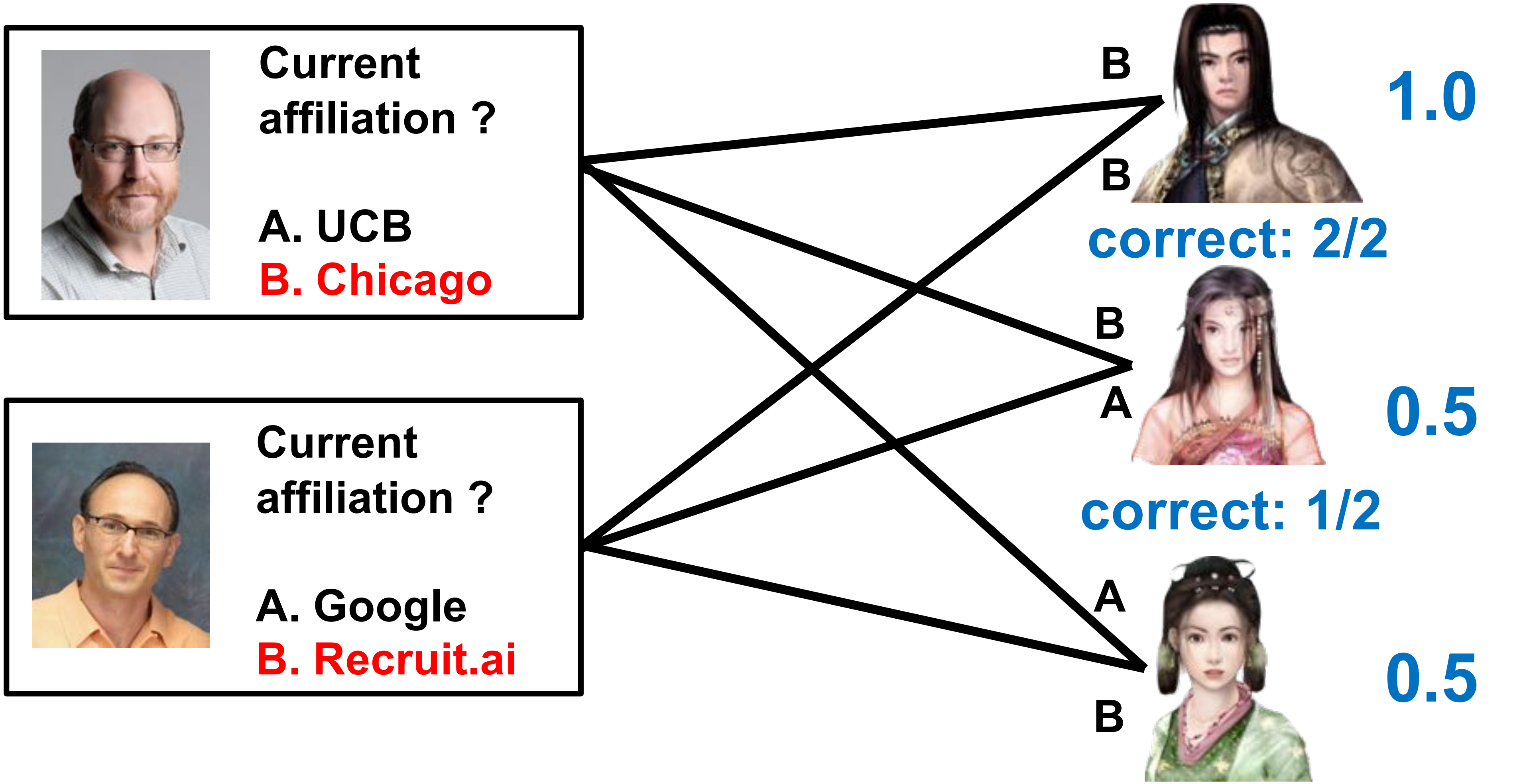
B. Recruit.ai (1.0 + 1.0 from workers 1 & 3)

# Inherent Relationship 2

○ 2. Truth for each task → Quality for each worker

Truth:

Quality:



Current affiliation ?

A. UCB

B. Chicago

Current affiliation ?

A. Google

B. Recruit.ai

1.0

correct: 2/2

0.5

correct: 1/2

0.5

correct: 1/2

# Outline

- ○ **Part I. Truth Inference**
  - – Problem Definition
  - – Condition 1: with ground truth
    - Qualification Test & Hidden Test
  - – Condition 2: without ground truth
    - Unified Framework
    - ☞ Existing Works
    - Experimental Results

- ○ **Part II. Task Assignment**
  - – Problem Definition
  - – Differences in Existing Works

# Existing works

- **Classic Method**

D&S [Dawid and Skene. JRSS 1979]

- **Recent Methods**

**(1) Database Community:**

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

**(2) Data Mining Community:**

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

**(3) 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]

# Differences in Existing works

**Tasks**

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

- ○ **Different Task Models**
  ***How they model each task ?***
  ***E.g., task difficulty …***

**Workers**

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

# Tasks: Different Tasks Types

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

Is Bill Gates currently the CEO of Microsoft ?

○ Yes ○ 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: ……

○ Pos ○ Neu ○ 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

# Tasks: Different Tasks Models

○ **Task Difficulty: a value**

**If a task receives many contradicting (or ambiguous) answers, then it is regarded as a difficult task.**

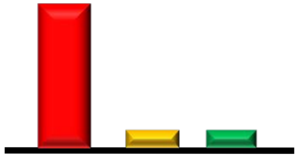
**e.g., Welinder et al. NIPS 2010, Ma et al. KDD16**

○ **Diverse Domains: a vector**

🟥 Sports 🟨 Politics 🟩 Entertainment

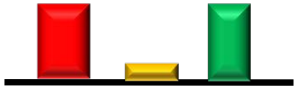
Did Michael Jordan win more NBA championships than Kobe Bryant?

→ Sports



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

→ Sports & Entertainment



# Tasks: Different Task Models (cont’d)

○ **Diverse Domains (cont’d)**

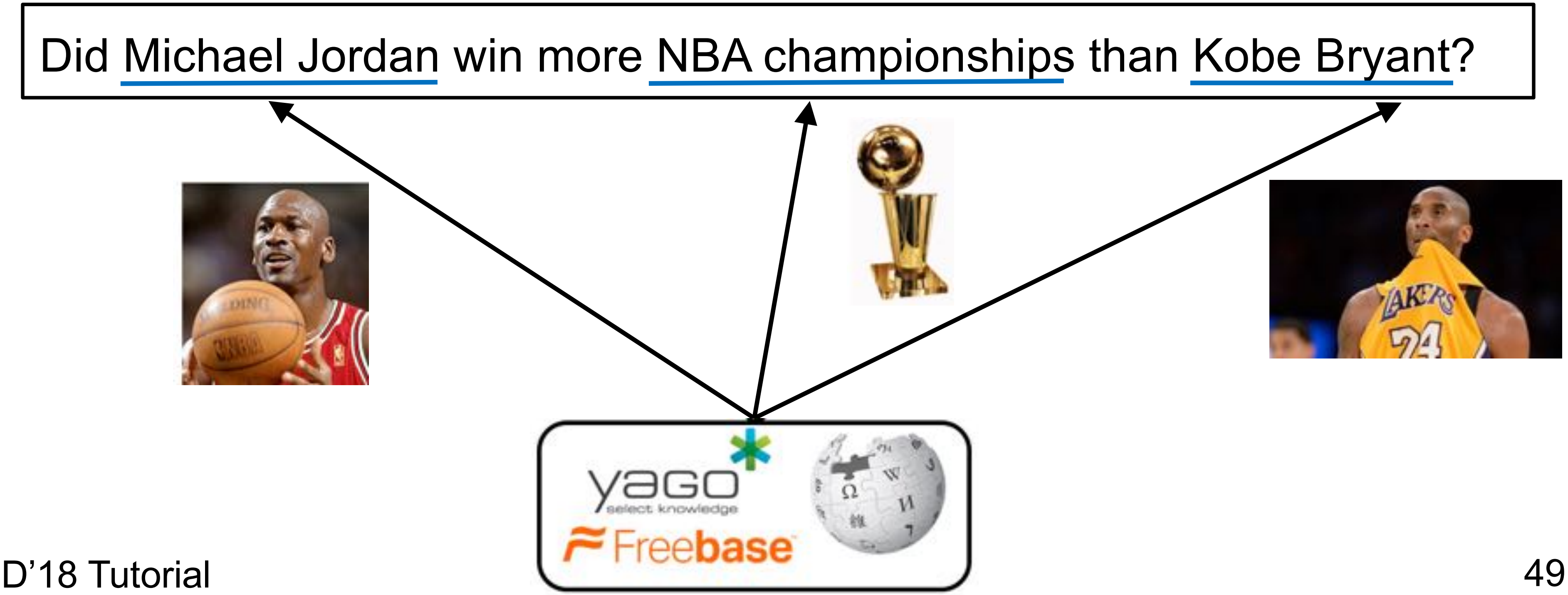
**To obtain the each task’s model:**

**(1) Use machine learning approaches**

**e.g., LDA [Blei e al. JMLR03],**

**TwitterLDA [Zhao et al. ECIR11].**

**(2) Use entity linking (map entity to knowledge bases).**



Did Michael Jordan win more NBA championships than Kobe Bryant?

# Workers: 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 $[p-\varepsilon, p+\varepsilon]$

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

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

# Workers: Different Worker Models (cont’d)

- **Confusion Matrix**: a matrix

  **Capture a worker’s answer for different choices given a specific truth**

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

***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 $\tau$ & Variance $\sigma$ : numerical task**

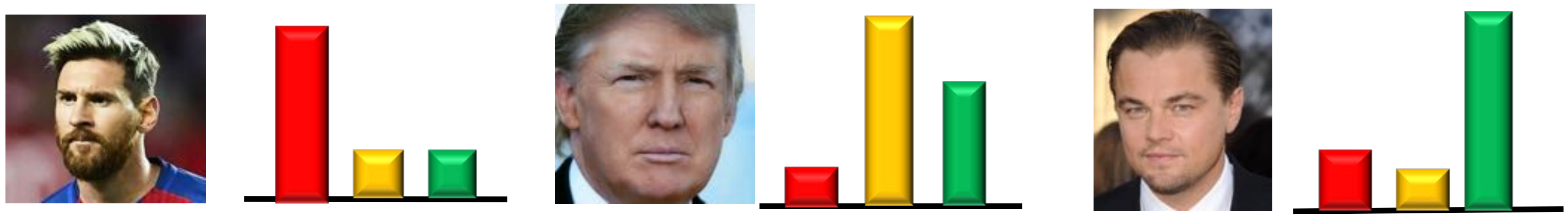
  **Answer follows Gaussian distribution:** $ans \sim N(t+\tau,\sigma)$

  **e.g., Raykar et al. JLMR10**

# Workers: Different Worker Models (cont’d)

- **Quality Across Diverse Domains: a vector**

🟥 Sports 🟨 Politics 🟩 Entertainment



**How to decide the scope of domains ?**

***Idea: Use domains from Knowledge Bases***

**e.g., Ma et al. KDD16, Zheng et al. VLDB17**

# Summary of Truth Inference Methods

| Method | Task Type | Task Model | Worker Model |
|---|---|---|---|
| Majority Voting | Decision-Making Task, Single-Choice Task | No | No |
| Mean / Median | Numeric Task | No | No |
| ZC [Demartini et al. WWW12] | Decision-Making Task, Single-Choice Task | No | Worker Probability |
| GLAD [Whitehill et al. NIPS09] | Decision-Making Task, Single-Choice Task | Task Difficulty | Worker Probability |
| D&S [Dawid and Skene.  JRSS 1979] | Decision-Making Task, Single-Choice Task | No | Confusion Matrix |
| Minimax [Zhou et al. NIPS12] | Decision-Making Task, Single-Choice Task | No | Diverse Domains |
| BCC [Kim et al. AISTATS12] | Decision-Making Task, Single-Choice Task | No | Confusion Matrix |
| CBCC [Venanzi et al. WWW14] | Decision-Making Task, Single-Choice Task | No | Confusion Matrix |
| LFC [Raykar et al. JLMR10] | Decision-Making Task, Single-Choice Task | No | Confusion Matrix |
| CATD [Li et al. VLDB14] | Decision-Making Task, Single-Choice Task, Numeric Task | No | Worker Probability, Confidence |

# Summary of Truth Inference Methods (cont’d)

| Method | Task Type | Task Model | Worker Model |
|---|---|---|---|
| PM [Li et al. SIGMOD14] | Decision-Making Task, Single-Choice Task, Numeric Task | No | Worker Probability |
| Multi [Welinder et al. NIPS 2010] | Decision-Making Task | Diverse Domains | Diverse Domains, Worker Bias, Worker Variance |
| KOS [Karger et al. NIPS11] | Decision-Making Task | No | Worker Probability |
| VI-BP [Liu et al. NIPS12] | Decision-Making Task | No | Confusion Matrix |
| VI-MF [Liu et al. NIPS12] | Decision-Making Task | No | Confusion Matrix |
| LFC_N [Raykar et al. JLMR10] | Numeric Task | No | Worker Variance |
| iCrowd [Fan et al. SIGMOD15] | Decision-Making Task, Single-Choice Task | Diverse Domains | Diverse Domains |
| FaitCrowd [Ma et al. KDD16] | Decision-Making Task, Single-Choice Task | Diverse Domains | Diverse Domains |
| DOCS [Zheng et al. VLDB17] | Decision-Making Task, Single-Choice Task | Diverse Domains | Diverse Domains |

# Outline

- **Part I. Truth Inference**
  - **Problem Definition**
  - **Condition 1: with ground truth**
    - **Qualification Test & Hidden Test**
  - **Condition 2: without ground truth**
    - **Unified Framework**
    - **Existing Works**
    - ☞ **Experimental Results**

- **Part II. Task Assignment**
  - **Problem Definition**
  - **Differences in Existing Works**

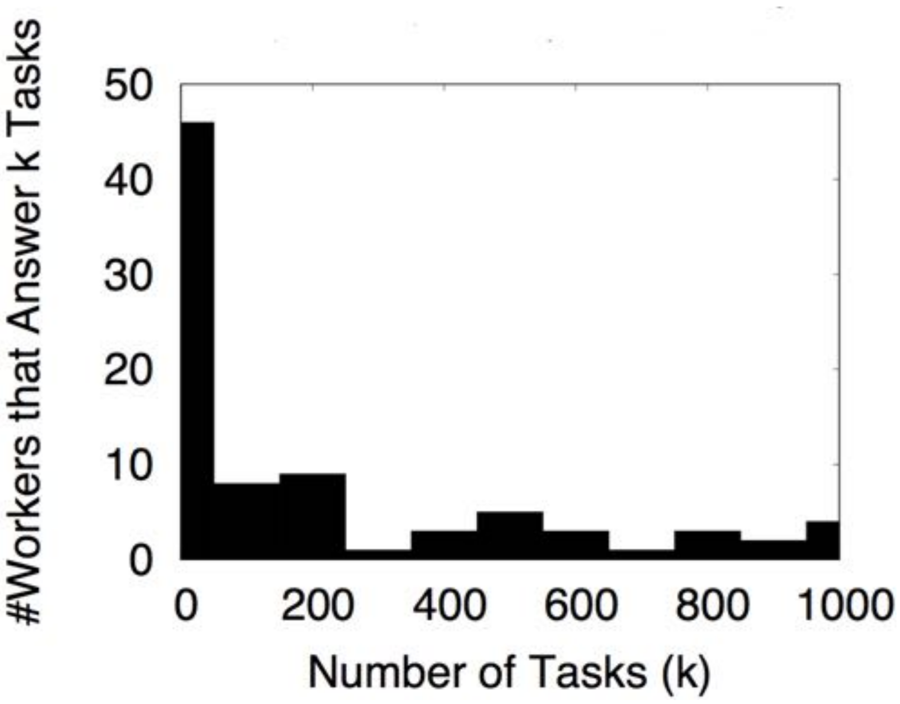
# Experimental Results (Zheng et al. VLDB17)

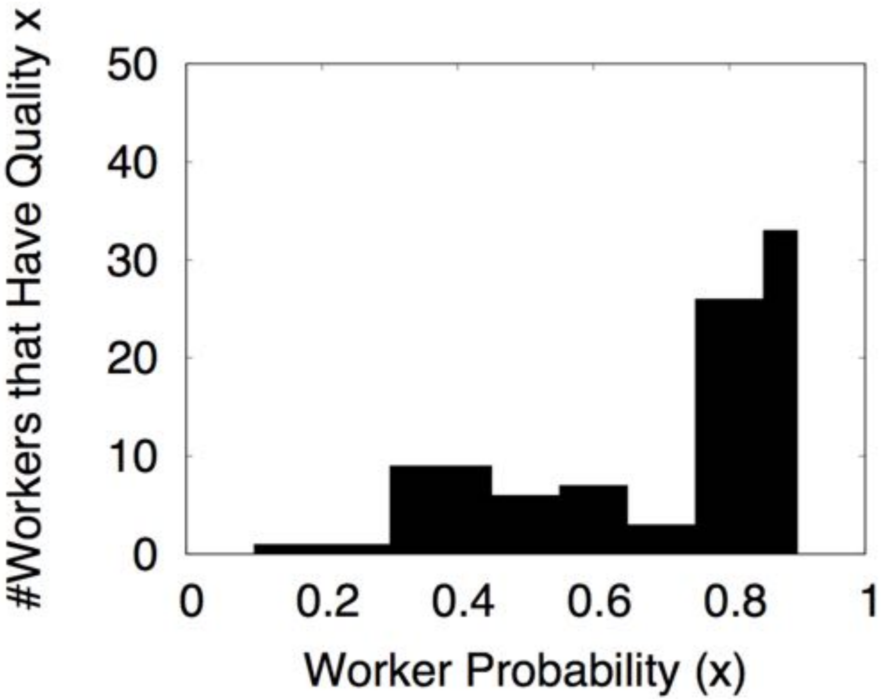
○ **Statistics of Datasets**

| 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 |

# Experimental Results

- Observations (Sentiment Analysis)



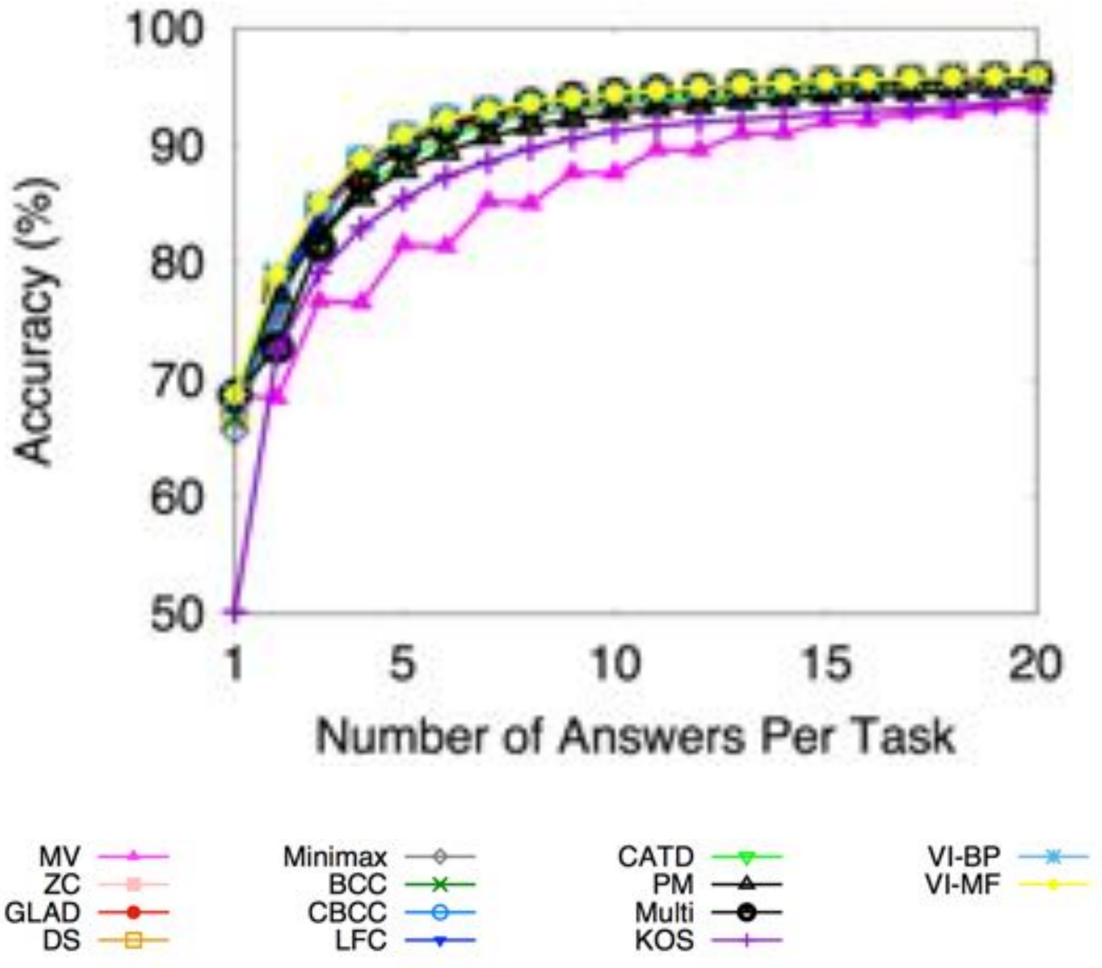


**#workers' answers conform to long-tail phenomenon (Li et al. VLDB14)**

**Not all workers are of very high quality**

# Experimental Results (cont’d)

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



Observations:

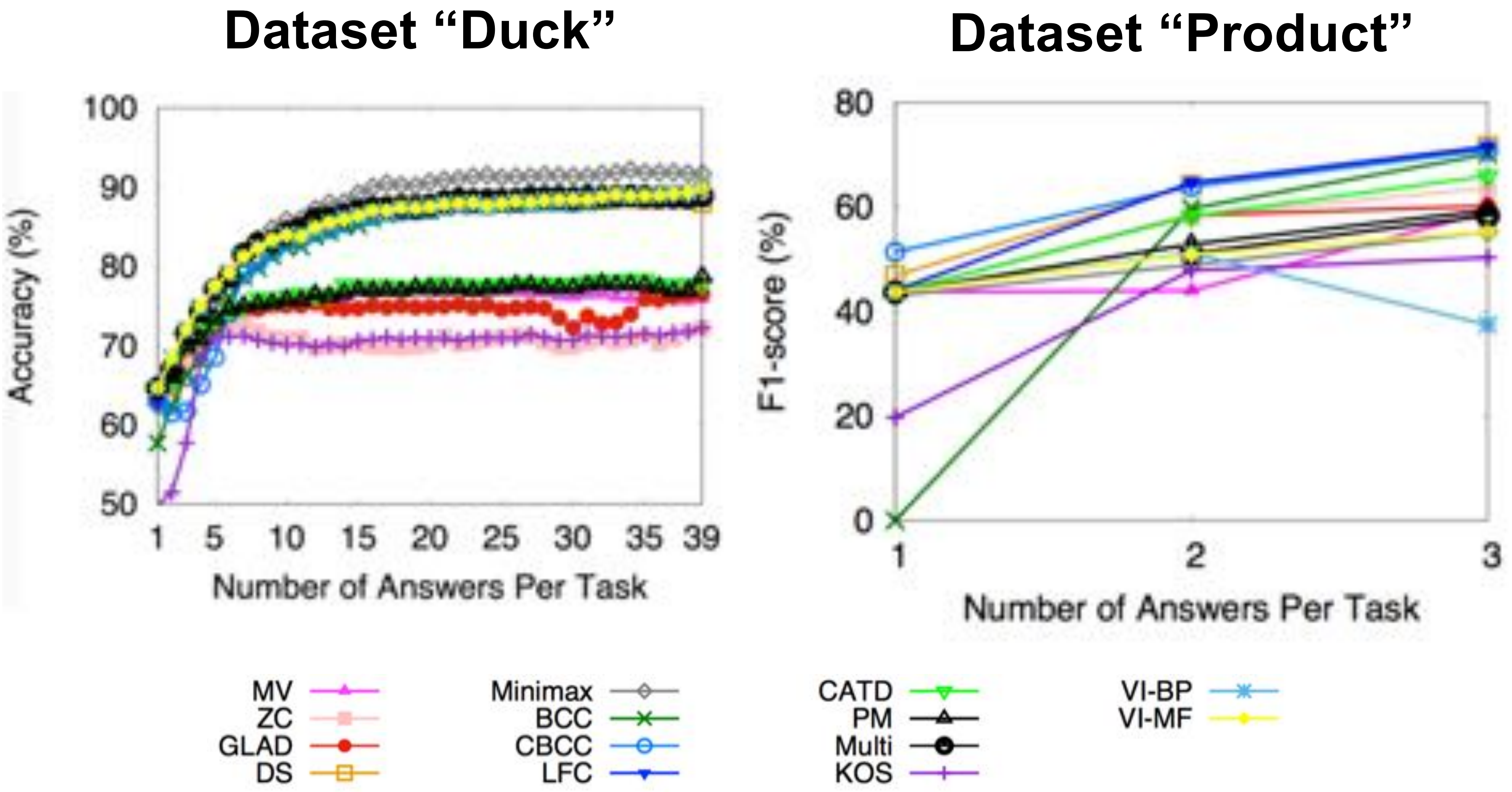
1. The quality increases with #answers;

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).

# Experimental Results (cont’d)

- Performance on more datasets



Dataset “Duck”

Dataset “Product”

# Which method is the best ?

- ○ **Decision-Making & Single-Label Tasks**
  - – "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.

- ○ **Numeric Tasks**
  - – "Mean" since it is robust in practice;
  - – "PM [Li et al. SIGMOD14]" as advanced techniques.

# Take-Away for Truth Inference

- The key to truth is to **compute each worker's quality**

- if some truth is known:

  **qualification test** and **hidden test**;

- if no truth is known:

  (1) relationships between **"quality for each worker"** and **"truth for each task"**

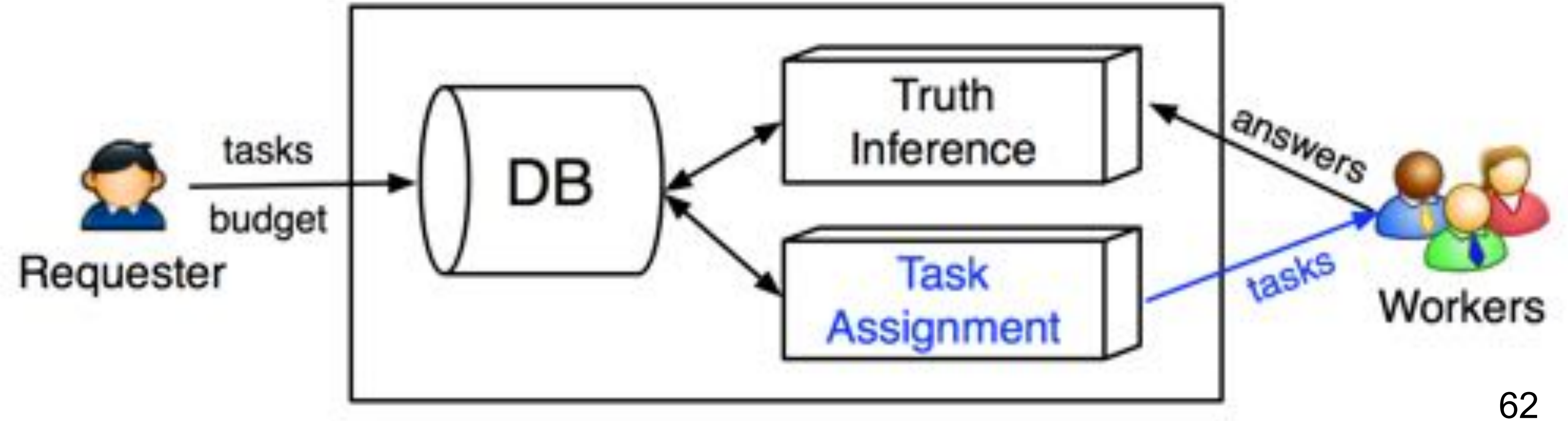
  (2) different **task types & models** and **worker models**

# Crowdsourcing Workflow

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

(1) when a worker comes to the platform, the worker will be assigned to a set of tasks (task assignment);

(2) when a worker accomplishes tasks, the platform will collect answers from the worker (truth inference).



# Part II. Task Assignment

- Existing platforms support online task assignment

amazonmechanical turk beta Artificial Artificial Intelligence ⇒ "External HIT"

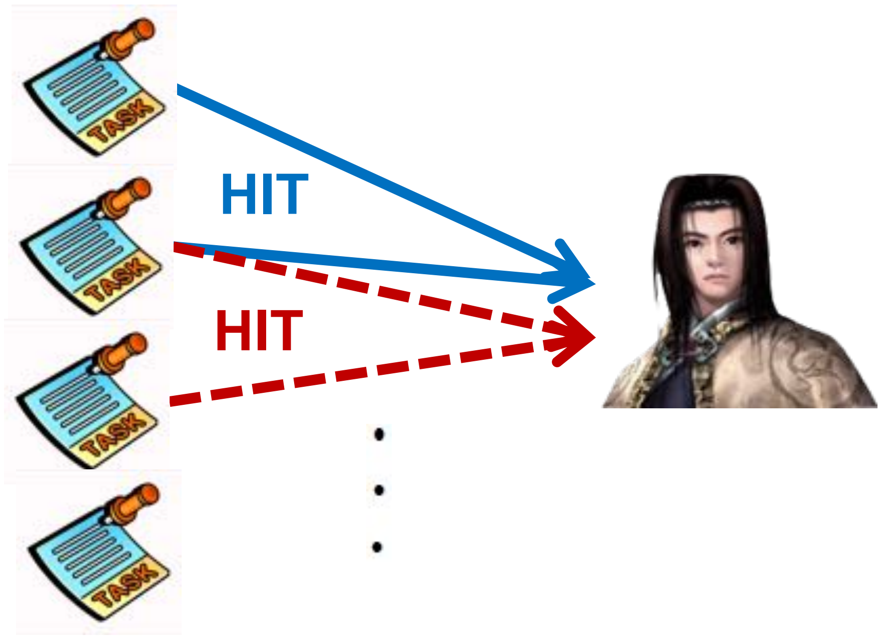
- Intuition: requesters want to wisely use the budgets

**How to allocate suitable tasks to workers?**

# Task Assignment Problem

Given a pool of n tasks, which set of the k tasks should be batched in a HIT and assigned to the worker?

Example:
Suppose we have n=4 tasks, and each time k=2 tasks are assigned as a HIT.



# This problem is complex!

- Simple enumeration:
  “n choose k” combinations

  (n = 100, k = 5) ➔ 100M assignments

- Need efficient (online) assignment

  Fast response to worker’s request

- Develop efficient heuristics

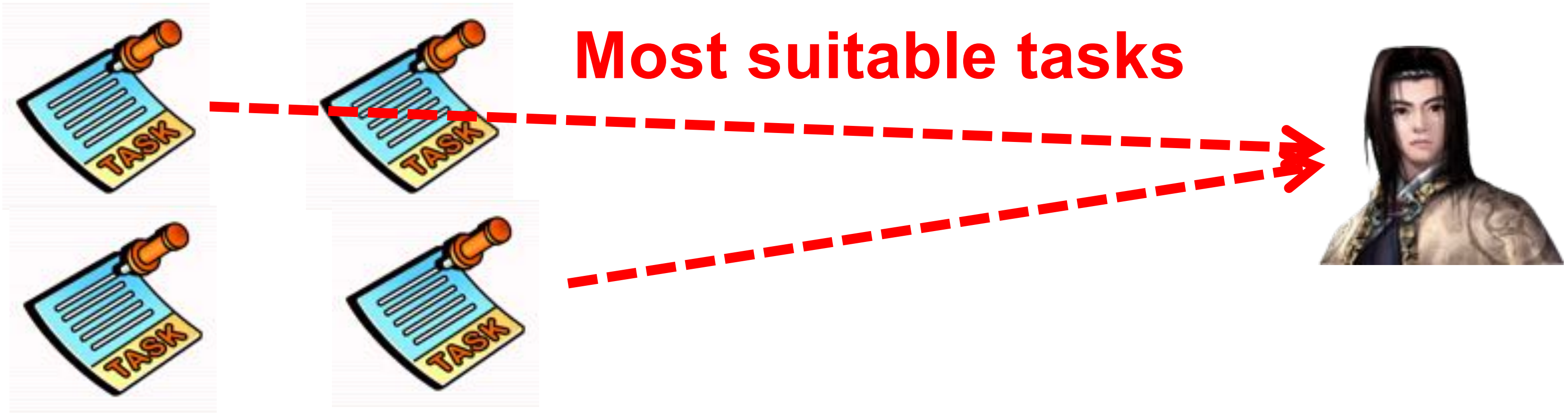
  Assignment time linear in #tasks: O(n)

# Outline

- ○ **Part I. Truth Inference**
  - – **Problem Definition**
  - – **Condition 1: with ground truth**
    - **Qualification Test & Hidden Test**
  - – **Condition 2: without ground truth**
    - **Unified Framework**
    - **Existing Works**
    - **Experimental Results**

- ○ **Part II. Task Assignment**
  - – **Problem Definition**
  - 👉 – **Existing Works**

# Main Idea



Most suitable tasks

**3 factors for characterizing a suitable task:**

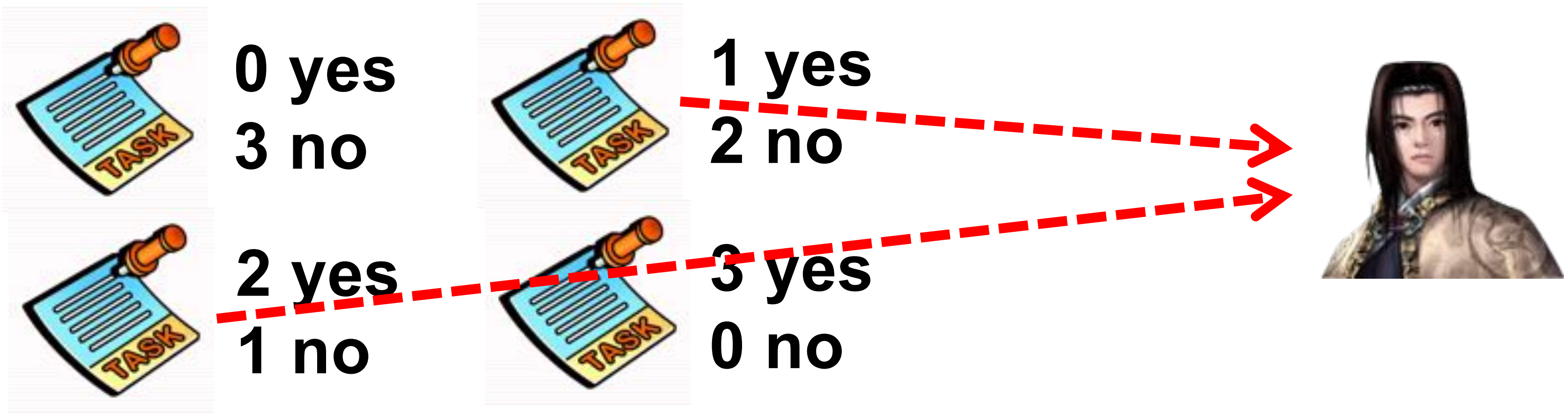
**Answer uncertainty**

**Worker quality**

**Requesters' objectives**

# Factor 1: Answer Uncertainty

○ Consider a decision-making task (yes/no)



0 yes
3 no

1 yes
2 no

2 yes
1 no

3 yes
0 no

○ Select a task whose answers are the most uncertain or inconsistent

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

# Factor 1: Answer Uncertainty

- **Entropy (Zheng et al. SIGMOD15)**
  **Given *c* choices for a task and the distribution of answers for a task** $\vec{p} = (p_1, p_2, \dots, p_c)$
  **The task's entropy is:**

$$H(\vec{p}) = -\sum_{i=1}^{c} p_i \log p_i$$

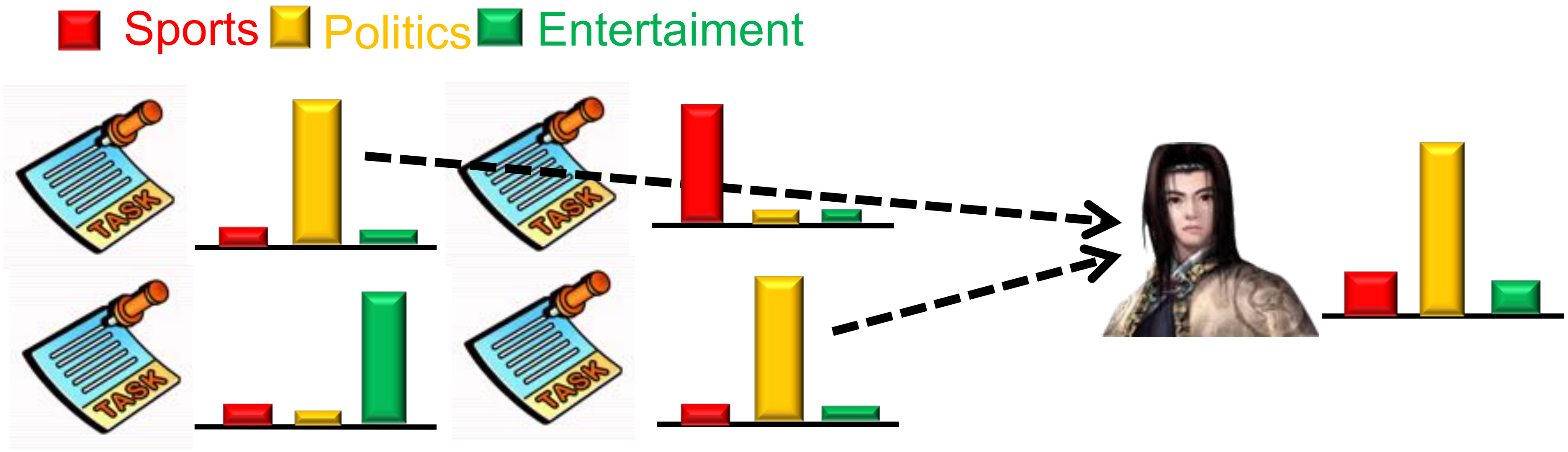
***e.g., a task receives 1 "yes" and 2 "no", then the distribution is (1/3, 2/3), and entropy is 0.637.***

- **Expected change of entropy (Roim et al. ICDE12)**
  **(1/3, 2/3) should be more uncertain than (10/30, 20/30):**

$$E[H(\vec{p}')] - H(\vec{p})$$

# Factor 2: Worker Quality

- Assign tasks to the worker with the suitable expertise



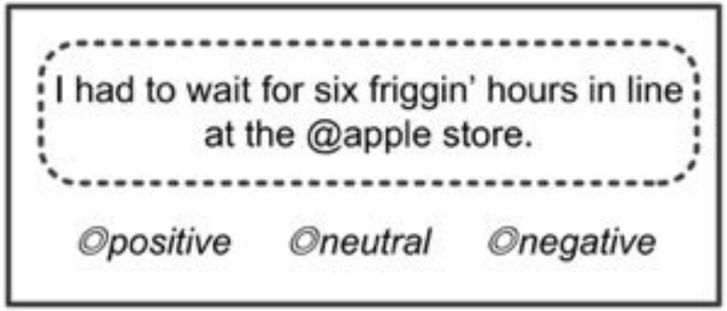
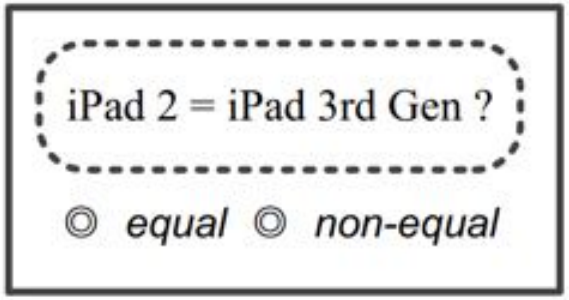
🟥 Sports 🟨 Politics 🟩 Entertaiment

- Uncertainty: consider the matching domains in tasks and the worker

e.g., Ho et al. AAAI12, Zheng et al. VLDB17

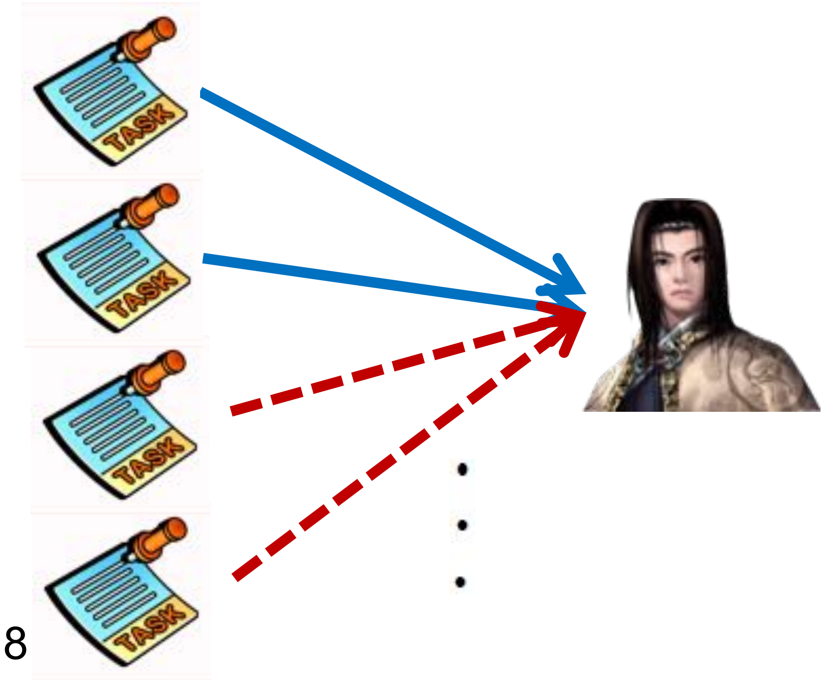
# Factor 3: Objectives of Requesters

- **Requesters may have different objectives (aka "evaluation metric") for different applications**

| **Application** | **Sentiment Analysis** | **Entity Resolution** |
|---|---|---|
| **Task** | <br>I had to wait for six friggin' hours in line at the @apple store.<br>◎positive ◎neutral ◎negative | <br>iPad 2 = iPad 3rd Gen ?<br>◎ *equal* ◎ *non-equal* |
| **Evaluation Metric** | **Accuracy** | **F-score ("equal" label)** |

# Factor 3: Objectives of Requesters

- Solution in QASCA (Zheng et al. SIGMOD15)
  (1) Leverage the answers collected from workers to create a "distribution matrix";
  (2) leverage the "distribution matrix" to estimate the quality improvement for a specific set of selected tasks.
- Idea: Select the best set of tasks with highest quality improvement in the specified evaluation metric.



improvement:

—— : 9%

- - - - : 6%

⋮

# Factor 3: Objectives of Requesters

- **Other Objectives**

**(1) Threshold on entropy (e.g., Li et al. WSDM17)**
**e.g., in the final state, each task should have constraint that its entropy ≥ 0.6.**

**(2) Threshold on worker quality (e.g., Fan et al. SIGMOD15)**
**e.g., in the final state, each task should have overall aggregated worker quality ≥ 2.0.**

**(3) Maximize total utility (e.g., Ho et al. AAAI12)**
**e.g., after the answer is given, the requester receives some utility related to worker quality, and the goal is to assign tasks that maximize the total utility.**

# Task Assignment

| Method | Factor 1: Answer Uncertainty | Factor 2: Worker Quality | Factor 3: Requesters' Objectives |
|---|---|---|---|
| OTA [Ho et al. AAAI12] | Majority | Worker probability | Maximize total utility |
| CDAS [Liu et al. VLDB12] | Majority | Worker probability | A threshold on confidence + early termination of confident tasks |
| iCrowd [Fan et al. SIGMOD15] | Majority | Diverse domains | Maximize overall worker quality |
| AskIt! [Roim et al. ICDE12] | Entropy-based | No | No |
| QASCA [Zheng et al. SIGMOD15] | Maximize specified quality | Confusion matrix | Maximize specified quality |
| DOCS [Zheng et al. VLDB17] | Expected change of entropy | Diverse domains | No |
| CrowdPOI [Hu et al. ICDE16] | Expected change of accuracy | Worker probability | No |
| Opt-KG [Li et al. WSDM17] | Majority | No | ≥ threshold on entropy |

# Take-Away for Task Assignment

- Require online and efficient heuristics

- Key idea: assign the most suitable task to worker, based on:

  (1) uncertainty of collected answers;
  (2) worker quality; and
  (3) requester’ objectives.

# Public Datasets & Codes

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

- **Implementations of truth inference algorithms** (https://github.com/TsinghuaDatabaseGroup/crowdsourcing/tree/master/truth/src/methods).

- **Implementations of task assignment algorithms** (https://github.com/TsinghuaDatabaseGroup/CrowdOTA).

# 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

[1] 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

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

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

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

[5] 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.

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

[7] 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.

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

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

[10] 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.

# Outline

- ○ Crowdsourcing Overview (20min)
- ○ Fundamental Techniques (90min)
  - – Quality Control (40min)
  - ☞ – Cost Control (30min)
  - – Latency Control (20min)
- ○ Crowd-powered Data Mining (60min)
  - – Crowd-powered Pattern Mining (10min)
  - – Crowd-powered Classification (10min)
  - – Crowd-powered Clustering (10min)
  - – Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - – Crowd-powered Knowledge Discovery (20min)
- ○ Challenges (10min)

Part 1

Part 2

# Cost Control

- **Goal**
  - How to reduce monetary cost?

- **Cost = $\boldsymbol{n}\times\boldsymbol{c}$**
  - $n$: number of tasks
  - $c$: cost of each task

- **Challenges**
  - How to reduce $n$?
  - How to reduce $c$?

# Classification of Existing Techniques

- **How to reduce $n$?**
  - ☞ Task Pruning
  - Answer Deduction
  - Task Selection
  - Sampling

The Database Community

- **How to reduce $c$?**
  - Task Design

The HCI Community

# Task Pruning

- **Key Idea**
  - Prune the tasks that machines can do well

- **Easy Task vs. Hard Task**

**Are they the same?**

IPHONE 6 = iphone 6

**Are they the same?**

IBM = Big Blue

- **How to quantify "difficulty"**
  - Similarity value
  - Match probability

- Jiannan Wang, Tim Kraska, Michael J. Franklin, Jianhua Feng: CrowdER: Crowdsourcing Entity Resolution. VLDB 2012
- Steven Euijong Whang, Peter Lofgren, Hector Garcia-Molina: Question Selection for Crowd Entity Resolution. VLDB 2013

# Task Pruning (cont’d)

- **Workflow (non-iterative)**
  1. Rank tasks based on "difficulty"
  2. Prune the tasks whose difficulty ≤ threshold

- **Pros**
  - Support a **large variety** of applications
- **Cons**
  - Only work for **easy** tasks (i.e., the ones that machines can do well)

# Classification of Existing Techniques

- **How to reduce $n$?**
  - Task Pruning
  - ☞ Answer Deduction
  - Task Selection
  - Sampling

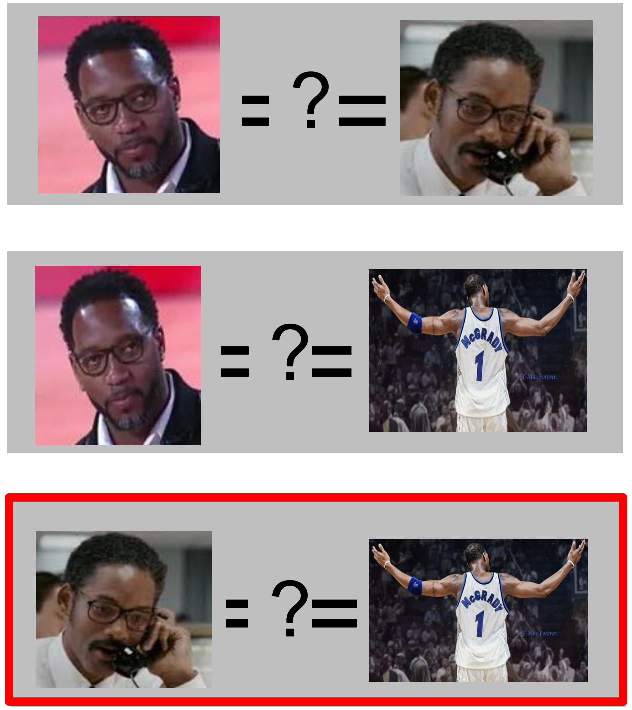
The Database Community

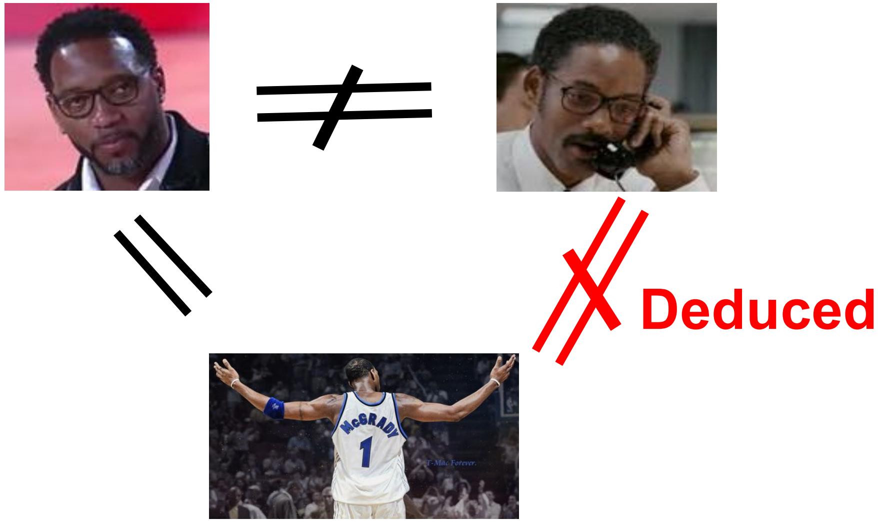
- **How to reduce $c$?**
  - Task Design

The HCI Community

# Answer Deduction

- **Key Idea**
  - Prune the tasks whose answers can be **deduced** from existing crowdsourced tasks
- **Example: Transitivity**



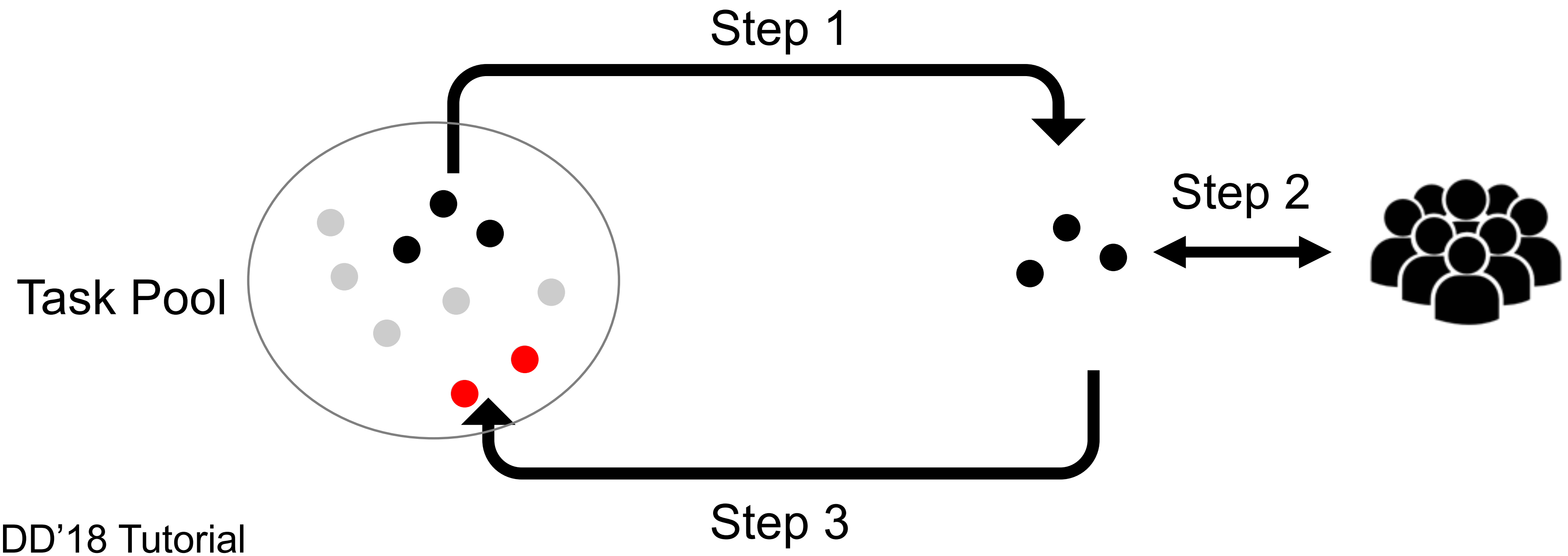


Jiannan Wang, Guoliang Li, Tim Kraska, Michael J. Franklin, Jianhua Feng: Leveraging transitive relations for crowdsourced joins. SIGMOD 2013
Donatella Firmani, Barna Saha, Divesh Srivastava: Online Entity Resolution Using an Oracle. PVLDB 2016

# Answer Deduction (cont'd)

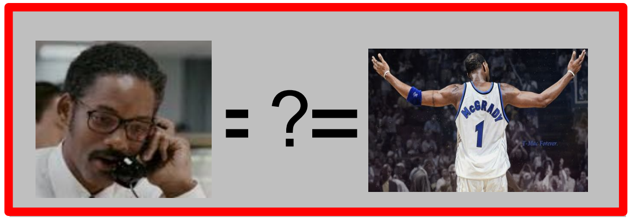
○ **Workflow (iterative)**

1. Pick up some tasks from a task pool
2. Collect answers of the tasks from the Crowd
3. Remove the tasks whose answers can be deduced

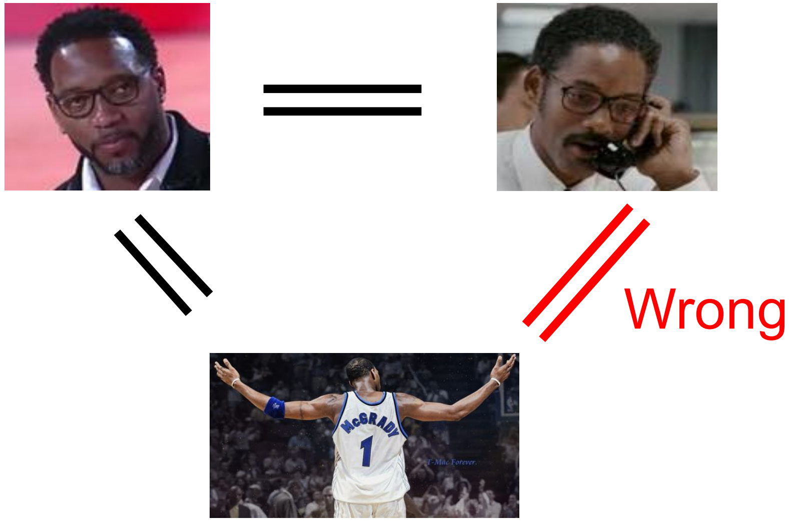


# Answer Deduction (cont'd)

- **Pros**
  - Work for both easy and **hard** tasks



- **Cons**
  - Human errors can be amplified



# Classification of Existing Techniques

- **How to reduce $n$?**
  - Task Pruning
  - Answer Deduction
  - ☞ Task Selection
  - Sampling

The Database Community

- **How to reduce $c$?**
  - Task Design

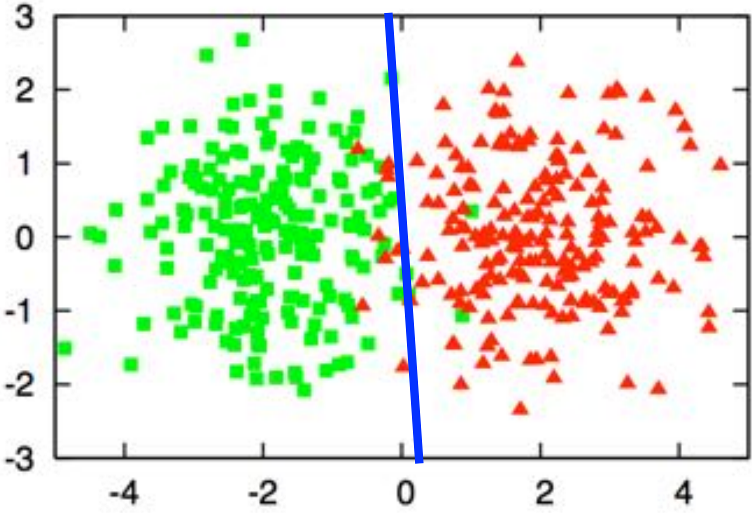
The HCI Community

# Task Selection

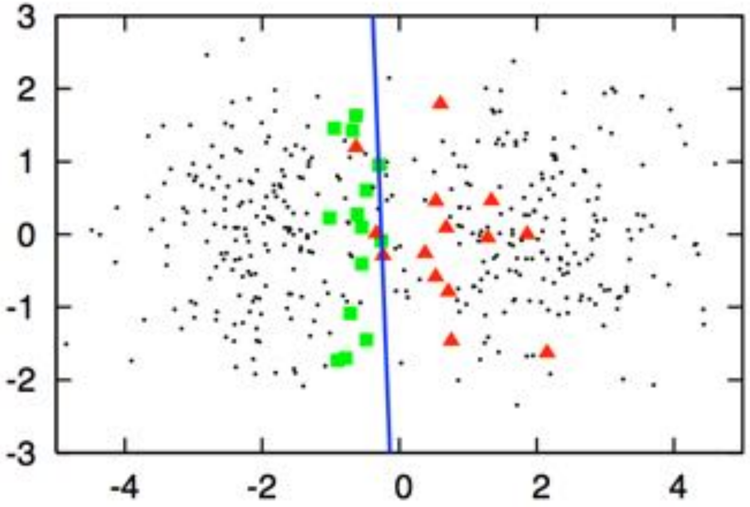
- **Key Idea**
  - Select the most **beneficial** tasks to crowdsource

- **Example 1: Active Learning**
  - Most beneficial for training a model

**Supervised Learning**



**Active Learning**



- Mozafari et al. Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning. PVLDB 2014
- Gokhale et al. Corleone: hands-off crowdsourcing for entity matching. SIGMOD 2014

# Task Selection

- **Key Idea**
  - Select the most **beneficial** tasks to crowdsource

- **Example 2: Top-k**
  - Most beneficial for getting the top-k results

**Which picture visualizes the best SFU Campus?**

Rank by computers

The most beneficial task:

VS.

Xiaohang Zhang, Guoliang Li, Jianhua Feng: Crowdsourced Top-k Algorithms: An Experimental Evaluation. PVLDB 2016

# Task Selection (cont’d)

- **Workflow (iterative)**
  1. Select a set of most beneficial tasks
  2. Collect their answers from the Crowd
  3. Update models and results

- **Pros**
  – Allow for a flexible quality/cost trade-off
- **Cons**
  – Hurt latency (since only a small number of tasks can be crowdsourced at each iteration)

# Classification of Existing Techniques

- **How to reduce $n$?**
  - Task Pruning
  - Answer Deduction
  - Task Selection
  - ☞ Sampling

The Database Community

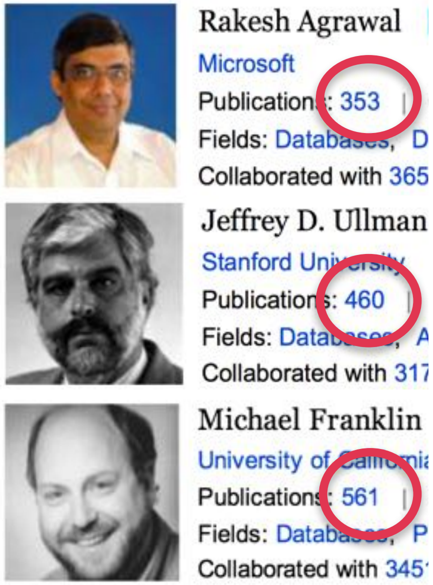
- **How to reduce $c$?**
  - Task Design

The HCI Community

# Sampling

- **Key Idea**
  - Ask the crowd to work on **sample** data
- **Example: SampleClean**

Who published more?



Rakesh Agrawal
Microsoft
Publications: 353 |
Fields: Databases, D
Collaborated with 365

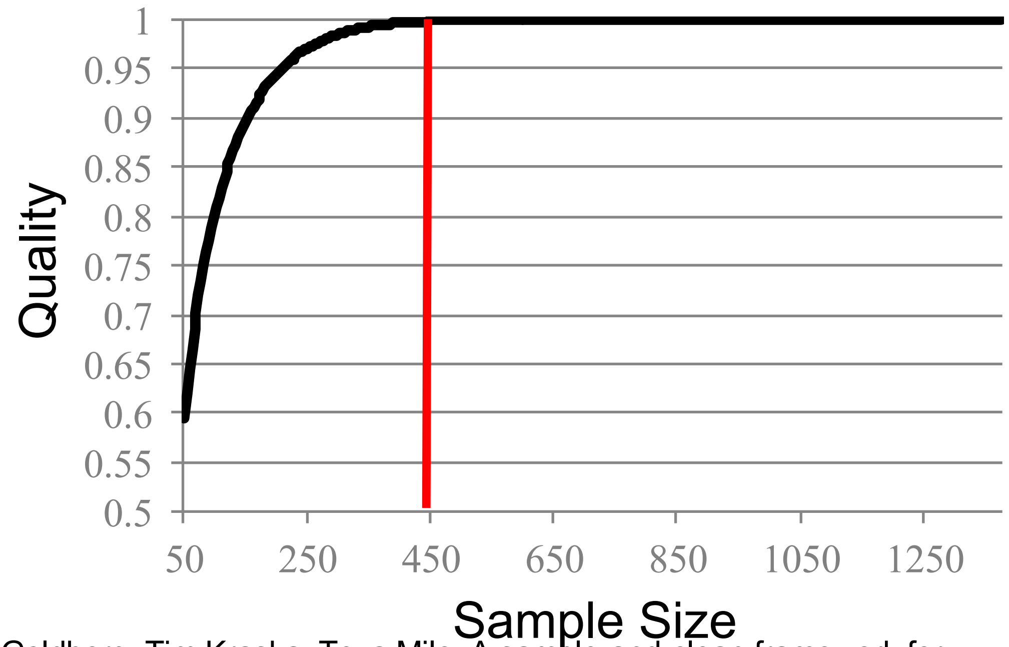
Jeffrey D. Ullman
Stanford University
Publications: 460 |
Fields: Databases, A
Collaborated with 317

Michael Franklin
University of California
Publications: 561 |
Fields: Databases, P
Collaborated with 345

211

255

173



Jiannan Wang, Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Tim Kraska, Tova Milo: A sample-and-clean framework for fast and accurate query processing on dirty data. SIGMOD Conference 2014: 469-480

# Sampling (Cont’d)

- **Workflow (iterative)**
  1. Generate tasks based on a sample
  2. Collect the task answers from the Crowd
  3. Infer the results of the full data

- **Pros**
  - Provable bounds for quality (e.g., the paper count is 211±5 with 95% probability)
- **Cons**
  - Limited to certain applications (e.g., it does not work for max)

# Classification of Existing Techniques

- **How to reduce $n$?**
  - Task Pruning
  - Answer Deduction
  - Task Selection
  - Sampling

**The Database Community**

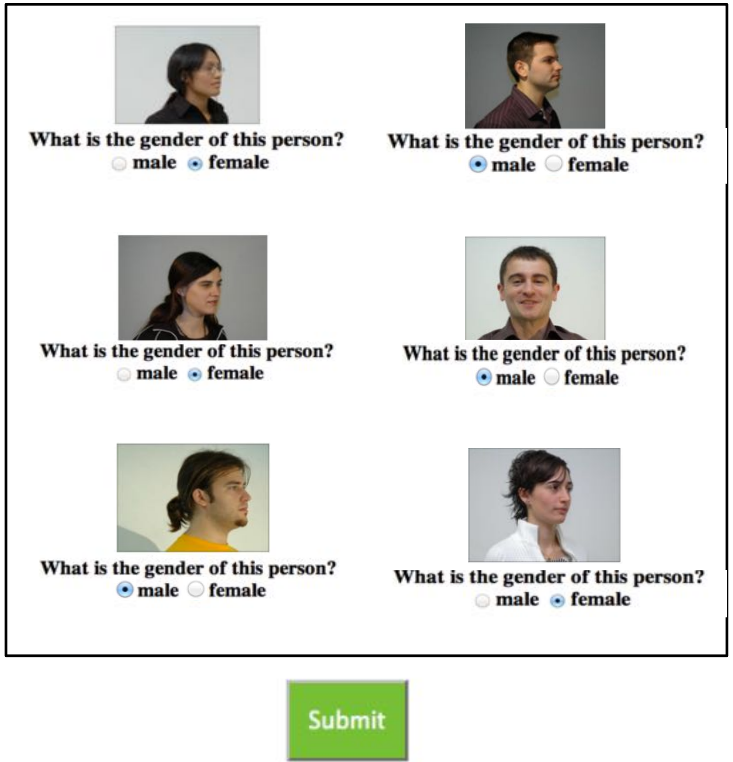
- **How to reduce $c$?**
  - ☞ Task Design

**The HCI Community**

# Task Design (Cont’d)

- **Key Idea**
  - Optimize User Interface

- **Example 1: Count**



What is the gender of this person?
○ male ● female

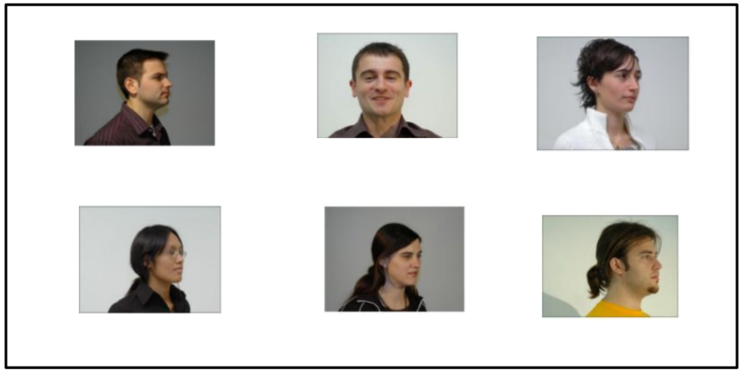
What is the gender of this person?
● male ○ female

What is the gender of this person?
○ male ● female

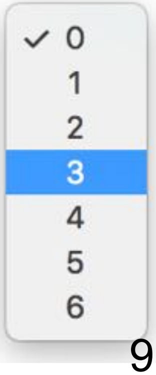
What is the gender of this person?
● male ○ female

What is the gender of this person?
● male ○ female

What is the gender of this person?
○ male ● female



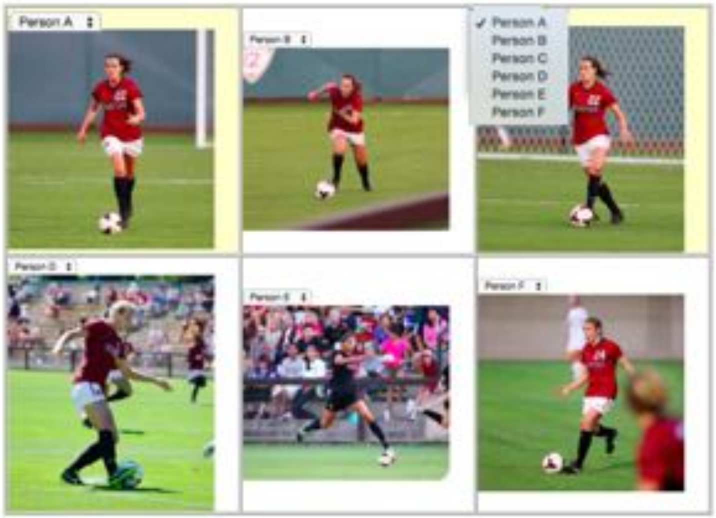
How many are female?



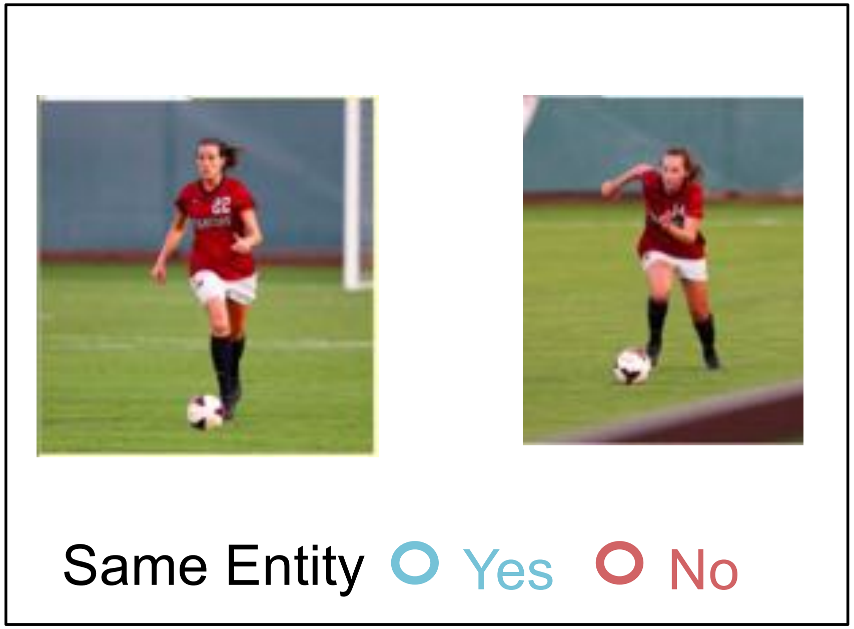
Adam Marcus, David R. Karger, Samuel Madden, Rob Miller, Sewoong Oh: Counting with the Crowd. PVLDB 2012

# Task Design (Cont'd)

- **Key Idea**
  - Optimize User Interface
- **Example 2: Entity Resolution**



Multi-item interface



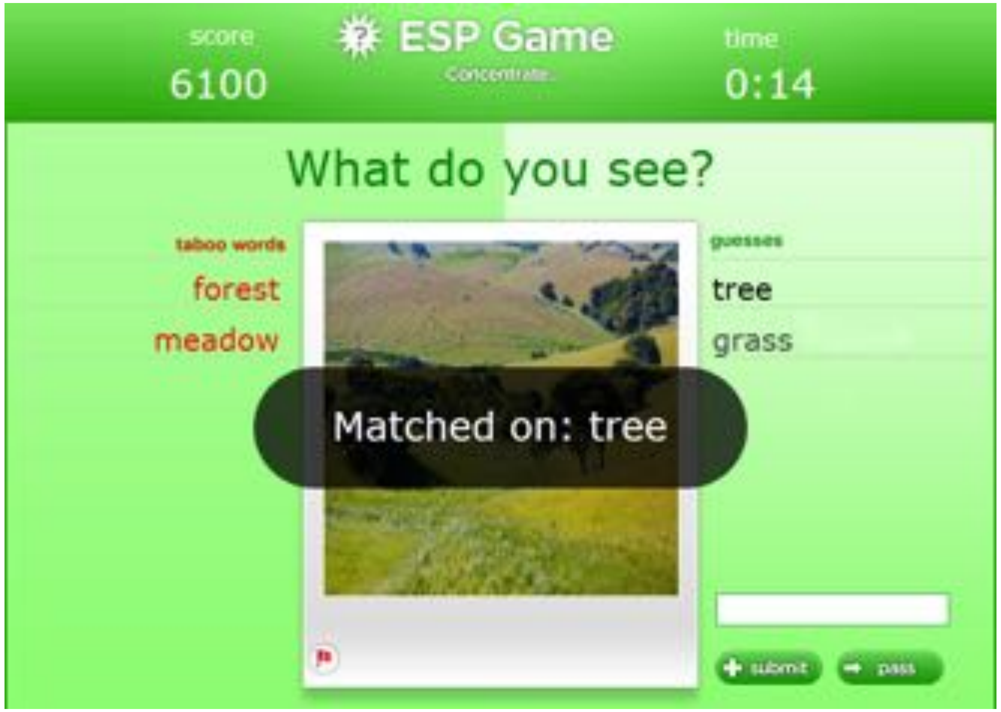
Pairwise interface

Vasilis Verroios, Hector Garcia-Molina, Yannis Papakonstantinou: Waldo: An Adaptive Human Interface for Crowd Entity Resolution. SIGMOD 2017

# Task Design (Cont’d)

- **Key Idea**
  - Optimize User Interface

- **Example 3: Image Labeling**



Luis von Ahn, Laura Dabbish: Labeling images with a computer game. CHI 2004: 319-326

# Summary of Cost Control

- **Two directions**
  - How to reduce n? ← DB
  - How to reduce c? ← HCI

- **DB and HCI should work together**

- **Non-iterative and iterative workflows are both widely used**

# Outline

- ○ Crowdsourcing Overview (20min)
- ○ Fundamental Techniques (90min)
  - – Quality Control (40min)
  - – Cost Control (30min)
  - ☞ – Latency Control (20min)

Part 1

- ○ Crowd-powered Data Mining (60min)
  - – Crowd-powered Pattern Mining (10min)
  - – Crowd-powered Classification (10min)
  - – Crowd-powered Clustering (10min)
  - – Crowd-powered Machine Learning (10min)
    - • Deep learning
    - • Transfer learning
    - • Semi-supervised learning
  - – Crowd-powered Knowledge Discovery (20min)

Part 2

- ○ Challenges (10min)

# Latency Control

- **Goal**
  - How to reduce latency?

- **Latency = $\boldsymbol{n \times t}$**
  - $n$: number ❌ tasks
  - $t$: latency of each task

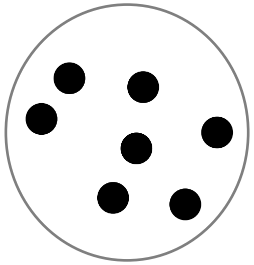
- **Latency =** The completion time of the last task

# Classification of Latency Control

☞ 1. **Single Task**
   - Reduce the latency of a single task

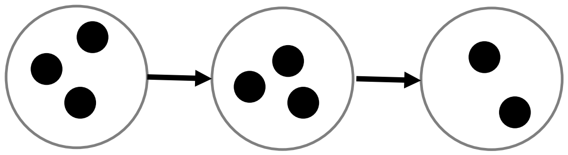
Single task

2. **Single Batch**
   - Reduce the latency of a batch of tasks



Single batch

3. **Multiple Batches**
   - Reduce the latency of multiple batches of tasks



Multiple batches

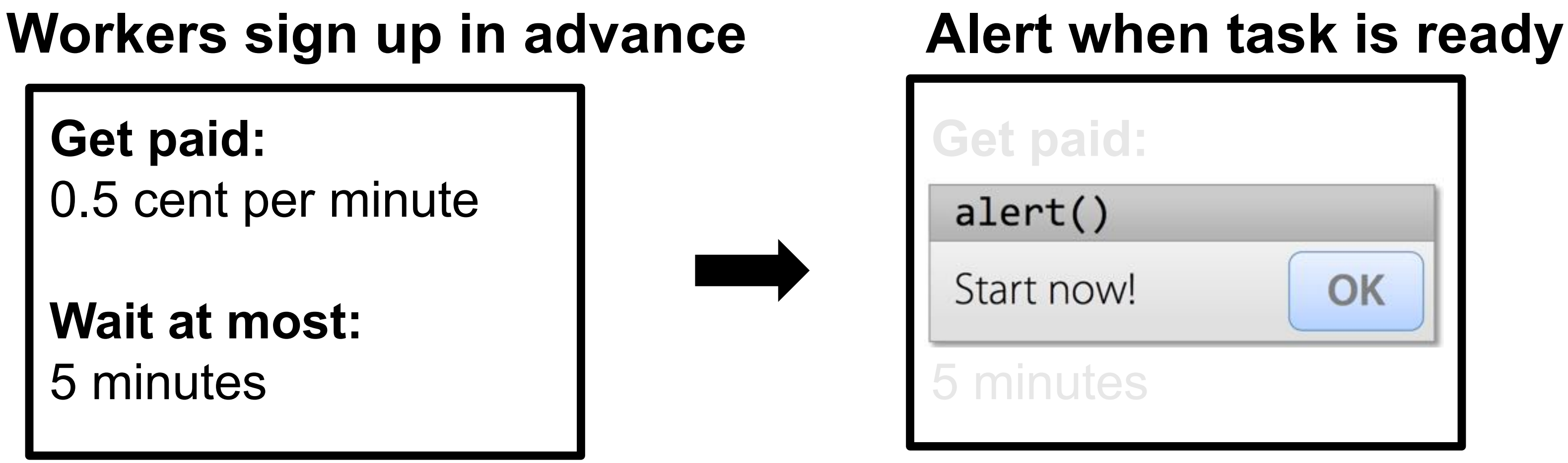
Daniel Haas, Jiannan Wang, Eugene Wu, Michael J. Franklin: CLAMShell: Speeding up Crowds for Low-latency Data Labeling. PVLDB 2015

# Single-Task Latency Control

- **Latency consists of**
  - Phase 1: Recruitment Time
  - Phase 2: Qualification and Training Time
  - Phase 3: Work Time
- **Improve Phase 1**
  - See the next slide
- **Improve Phase 2**
  - Remove this phase by applying other quality control techniques (e.g., worker elimination)
- **Improve Phase 3**
  - Better User Interfaces

# Reduce Recruitment Time

- **Retainer Pool**
  - Pre-recruit a pool of crowd workers



**Workers sign up in advance**

**Get paid:**
0.5 cent per minute

**Wait at most:**
5 minutes

**Alert when task is ready**

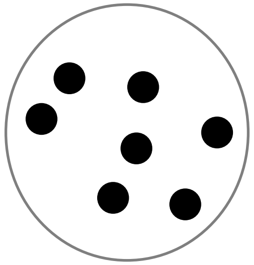
Michael S. Bernstein, Joel Brandt, Robert C. Miller, David R. Karger: Crowds in two seconds: enabling realtime crowd-powered interfaces. UIST 2011

# Classification of Latency Control

1. **Single Task**
   - Reduce the latency of a single task

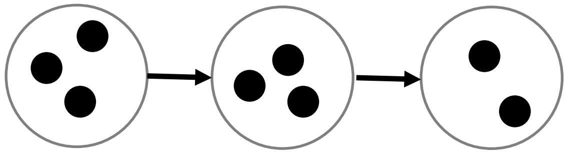
Single task

☞ 2. **Single Batch**
   - Reduce the latency of a batch of tasks



Single batch

3. **Multiple Batches**
   - Reduce the latency of multiple batches of tasks



Multiple batches

Daniel Haas, Jiannan Wang, Eugene Wu, Michael J. Franklin: CLAMShell: Speeding up Crowds for Low-latency Data Labeling. PVLDB 2015

# Single-Batch Latency Control

- **Idea 1: Pricing Model**
  - Model the relationship between task price and completion time

- **Predict worker behaviors** [1,2]
  - Recruitment Time
  - Work Time
- **Set task price**
  - Fixed Pricing [2]
  - Dynamic Pricing [3]

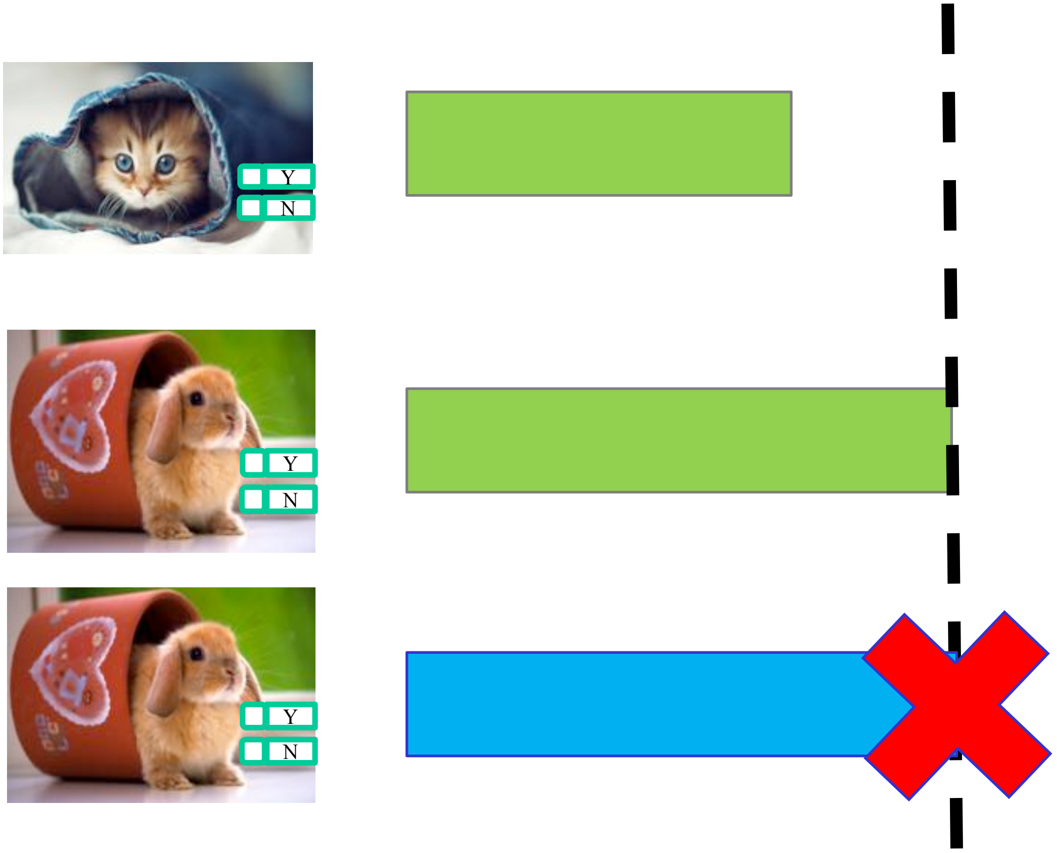
[1]. Wang et al. Estimating the completion time of crowdsourced tasks using survival analysis models. CSDM 2011

[2]. S. Faradani, B. Hartmann, and P. G. Ipeirotis. What's the right price? pricing tasks for finishing on time. In AAAI Workshop, 2011

[3]. Y. Gao and A. G. Parameswaran. Finish them!: Pricing algorithms for human computation. PVLDB 2014.

# Single-Batch Latency Control

- **Idea 2: Straggler Mitigation**
  - Replicate a task to multiple workers and return the result of the fastest worker



Straggler mitigation (e.g., MapReduce, Spark)

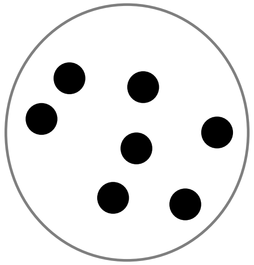
Daniel Haas, Jiannan Wang, Eugene Wu, Michael J. Franklin: CLAMShell: Speeding up Crowds for Low-latency Data Labeling. PVLDB 2015

# Classification of Latency Control

1. **Single Task**
   - Reduce the latency of a single task

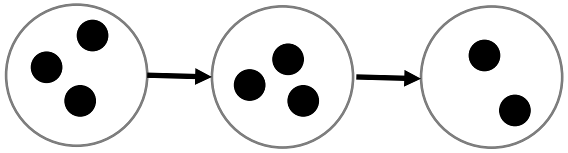
Single task

2. **Single Batch**
   - Reduce the latency of a batch of tasks



Single batch

☞ 3. **Multiple Batches**
   - Reduce the latency of multiple batches of tasks

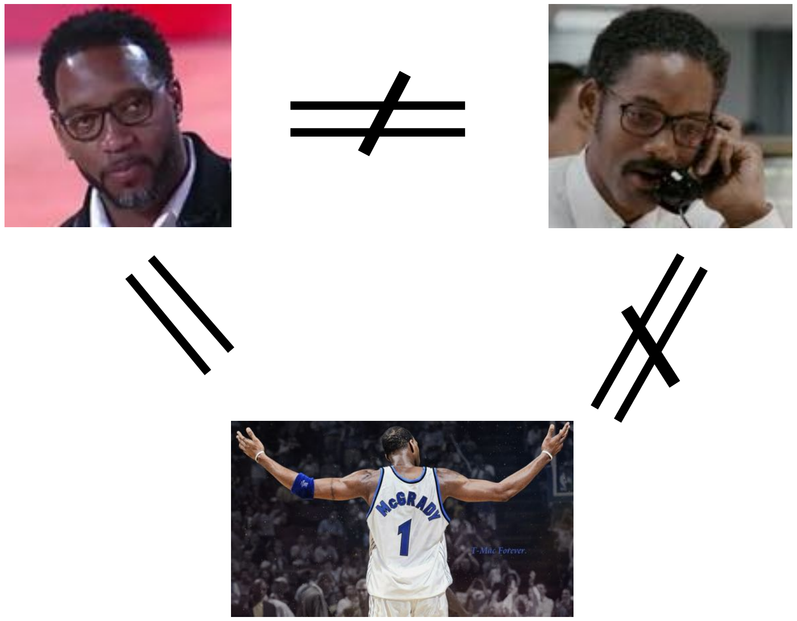


Multiple batches

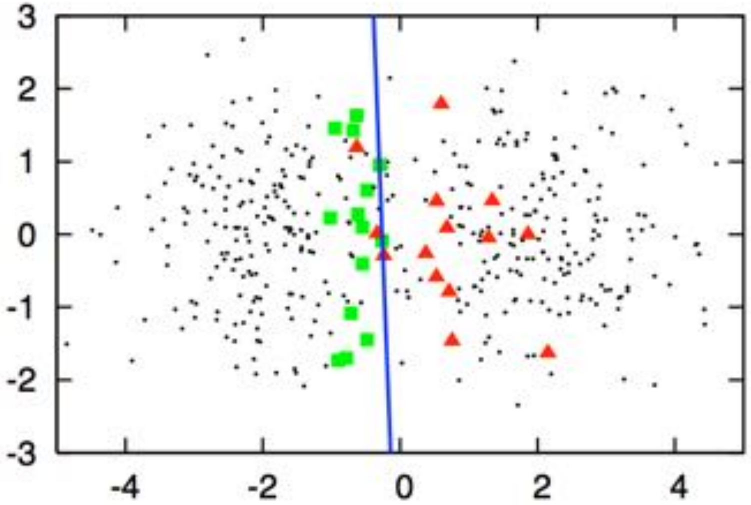
Daniel Haas, Jiannan Wang, Eugene Wu, Michael J. Franklin: CLAMShell: Speeding up Crowds for Low-latency Data Labeling. PVLDB 2015

# Multiple-Batches Latency Control

- **Why multiple batches?**
  - To save cost
    - Answer Deduction (e.g., leverage transitivity)
    - Task Selection (e.g., active learning)



**Active Learning**



# Multiple-Batches Latency Control

- **Two extreme cases**
  - Single task per batch: high latency
  - All tasks in one batch: high cost
- **Idea 1**
  - Choose the maximum batch size that does not hurt cost [1,2]
- **Idea 2**
  - Model as a latency budget allocation problem [3]

1. Jiannan Wang, Guoliang Li, Tim Kraska, Michael J. Franklin, Jianhua Feng: Leveraging transitive relations for crowdsourced joins. SIGMOD 2013
2. D. Sarma, A. G. Parameswaran, H. Garcia-Molina, and A. Y. Halevy. Crowd-powered find algorithms. ICDE 2014.
3. Verroios et al.. tdp: An optimal latency budget allocation strategy for crowdsourced MAXIMUM operations. SIGMOD 2015

# Summary of Latency Control

- **Latency**
  - The completion time of the last task
- **Classification of Latency Control**
  - Single-Task
    - Retainer Pool
    - Better UIs
  - Single-Batch
    - Pricing Model
    - Straggler Mitigation
  - Multiple-Batches
    - Batch size

# Two Take-Away Messages

- **There is no free lunch**
  - Cost control
    - Trades off quality (or/and latency) for cost
  - Latency control
    - Trades off quality (or/and cost) for latency

- **Learn from other communities**
  - Task Design (from HCI)
  - Straggler Mitigation (from Distributed System)

# Reference – Cost Control

1. Y. Amsterdamer, S. B. Davidson, T. Milo, S. Novgorodov, and A. Somech. Oassis: query driven crowd mining. In SIGMOD, pages 589–600. ACM, 2014
2. X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In WSDM, pages 193–202, 2013
3. G. Demartini, D. E. Difallah, and P. Cudre-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In WWW, pages 469–478, 2012.
4. B. Eriksson. Learning to top-k search using pairwise comparisons. In AISTATS, pages 265–273, 2013.
5. C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. W. Shavlik, and X. Zhu. Corleone: hands-off crowdsourcing for entity matching. In SIGMOD, pages 601–612, 2014.
6. A. Gruenheid, D. Kossmann, S. Ramesh, and F. Widmer. Crowdsourcing entity resolution: When is A=B? Technical report, ETH Zurich.
7. S. Guo, A. G. Parameswaran, and H. Garcia-Molina. So who won?: dynamic max discovery with the crowd. In SIGMOD, pages 385–396, 2012.
8. H. Heikinheimo and A. Ukkonen. The crowd-median algorithm. In HCOMP, 2013.
9. S. R. Jeffery, M. J. Franklin, and A. Y. Halevy. Pay-as-you-go user feedback for dataspace systems. In SIGMOD, pages 847–860, 2008.
10. H. Kaplan, I. Lotosh, T. Milo, and S. Novgorodov. Answering planning queries with the crowd. PVLDB, 6(9):697–708, 2013.
11. A. R. Khan and H. Garcia-Molina. Hybrid strategies for finding the max with the crowd. Technical report, 2014.
12. A. Marcus, D. R. Karger, S. Madden, R. Miller, and S. Oh. Counting with the crowd. PVLDB, 6(2):109–120, 2012.
13. B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden. Scaling up crowd-sourcing to very large datasets: a case for active learning. PVLDB, 8(2):125–136, 2014.
14. A. G. Parameswaran, A. D. Sarma, H. Garcia-Molina, N. Polyzotis, and J. Widom. Human-assisted graph search: it’s okay to ask questions. PVLDB, 4(5):267–278, 2011.

# Reference – Cost Control

15. T. Pfeiffer, X. A. Gao, Y. Chen, A. Mao, and D. G. Rand. Adaptive polling for information aggregation. In AAAI, 2012.
16. B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar. Crowdsourced enumeration queries. In ICDE, pages 673–684, 2013.
17. V. Verroios and H. Garcia-Molina. Entity resolution with crowd errors. In ICDE, pages 219–230, 2015.
18. N. Vesdapunt, K. Bellare, and N. N. Dalvi. Crowdsourcing algorithms for entity resolution. PVLDB, 7(12):1071–1082, 2014.
19. J. Wang, T. Kraska, M. J. Franklin, and J. Feng. CrowdER: crowdsourcing entity resolution. PVLDB, 5(11):1483–1494, 2012.
20. J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraska, and T. Milo. A sample-and-clean framework for fast and accurate query processing on dirty data. In SIGMOD, pages 469–480, 2014.
21. J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng. Leveraging transitive relations for crowdsourced joins. In SIGMOD, 2013.
22. S. Wang, X. Xiao, and C. Lee. Crowd-based deduplication: An adaptive approach. In SIGMOD, pages 1263–1277, 2015.
23. S. E. Whang, P. Lofgren, and H. Garcia-Molina. Question selection for crowd entity resolution. PVLDB, 6(6):349–360, 2013.
24. T. Yan, V. Kumar, and D. Ganesan. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In MobiSys, pages 77–90, 2010.
25. P. Ye, U. EDU, and D. Doermann. Combining preference and absolute judgements in a crowd-sourced setting. In ICML Workshop, 2013.
26. C. J. Zhang, Y. Tong, and L. Chen. Where to: Crowd-aided path selection. PVLDB, 7(14):2005–2016, 2014.

# Reference – Latency Control

1. J. P. Bigham et al. VizWiz: nearly real-time answers to visual questions. UIST, 2010.
2. M. S. Bernstein, J. Brandt, R. C. Miller, and D. R. Karger. Crowds in two seconds: enabling realtime crowd-powered interfaces. UIST, 2011.
3. M. S. Bernstein, D. R. Karger, R. C. Miller, and J. Brandt. Analytic Methods for Optimizing Realtime Crowdsourcing. Collective Intelligence, 2012.
4. Y. Gao and A. G. Parameswaran. Finish them!: Pricing algorithms for human computation. PVLDB, 7(14):1965–1976, 2014
5. S. Faradani, B. Hartmann, and P. G. Ipeirotis. What's the right price? pricing tasks for finishing on time. In AAAI Workshop, 2011.
6. D. Haas, J. Wang, E. Wu, and M. J. Franklin. Clamshell: Speeding up crowds for low-latency data labeling. PVLDB, 9(4):372–383, 2015
7. A. D. Sarma, A. G. Parameswaran, H. Garcia-Molina, and A. Y. Halevy. Crowd-powered find algorithms. In ICDE, pages 964–975, 2014
8. V. Verroios, P. Lofgren, and H. Garcia-Molina. tdp: An optimal-latency budget allocation strategy for crowdsourced MAXIMUM operations. In SIGMOD, pages 1047–1062, 2015.
9. T. Yan, V. Kumar, and D. Ganesan. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In MobiSys, pages 77–90, 2010.

# Outline

- Crowdsourcing Overview (20min)
- Fundamental Techniques (90min)
  - Quality Control (40min)
  - Cost Control (30min)
  - Latency Control (20min)

Part 1

- Crowd-powered Data Mining (60min)
  - ☞ Crowd-powered Pattern Mining (10min)
  - Crowd-powered Classification (10min)
  - Crowd-powered Clustering (10min)
  - Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - Crowd-powered Knowledge Discovery (20min)

Part 2

- Challenges (10min)

# Crowd-Powered Pattern Mining

○ **Typical Crowdsourcing Tasks (fixed choices)**

**What is the current affiliation for Michael Franklin ?**

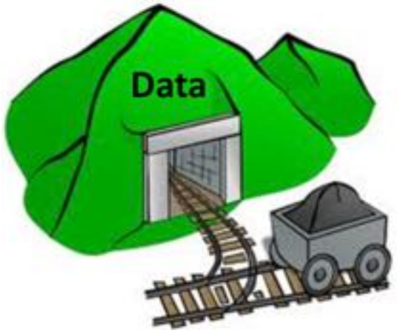
A. **University of California, Berkeley**
B. **University of Chicago**

○ **Crowd Pattern Mining**

**Find out what is *interesting* and *important* in some specific domains (e.g., medicines, habits)**

# Classic Pattern Mining

- Significant data pattern are identified using data mining techniques



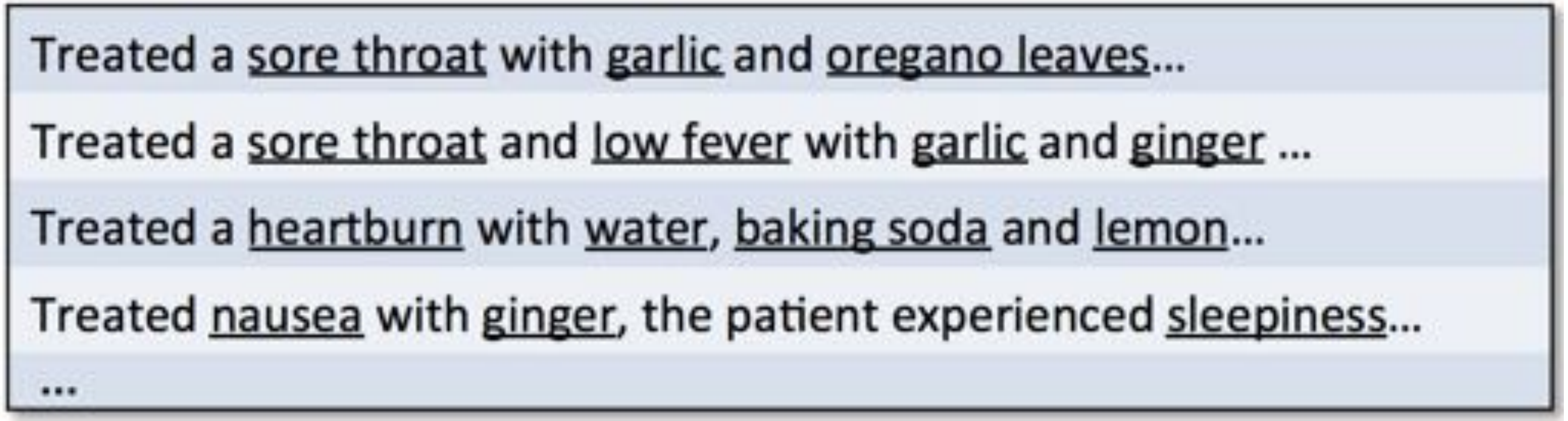
- A useful type of data pattern: association rules
  e.g.,
  *catch cold*
  to
  *sleep more,*
  *drink hot water,*
  *eat pills*

- Is it possible to mine from the crowd?

# User Modeling

- **A set of Users** $U$

- **Each User** $u \in U$ **has a (hidden) database**



Treated a sore throat with garlic and oregano leaves...

Treated a sore throat and low fever with garlic and ginger ...

Treated a heartburn with water, baking soda and lemon...

Treated nausea with ginger, the patient experienced sleepiness...

...

# User Modeling (cont'd)

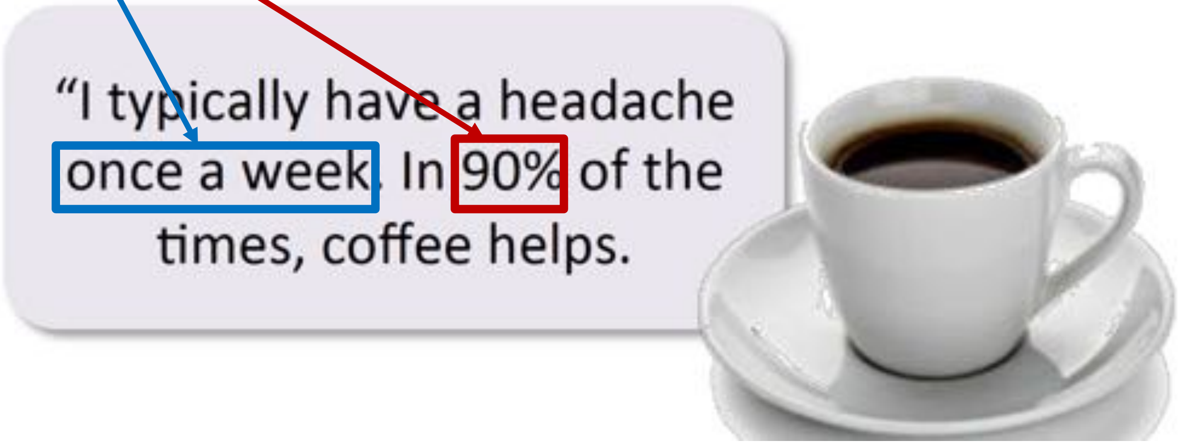
○ Each Rule $X \to Y$ in database is associated with

User Support

$$\mathrm{supp}_u(X \to Y) := \frac{|\{t \in D_u | X \cup Y \subseteq t\}|}{|D_u|}$$

User Confidence

$$\mathrm{conf}_u(X \to Y) := \frac{|\{t \in D_u | X \cup Y \subseteq t\}|}{|\{t \in D_u | X \subseteq t\}|}$$

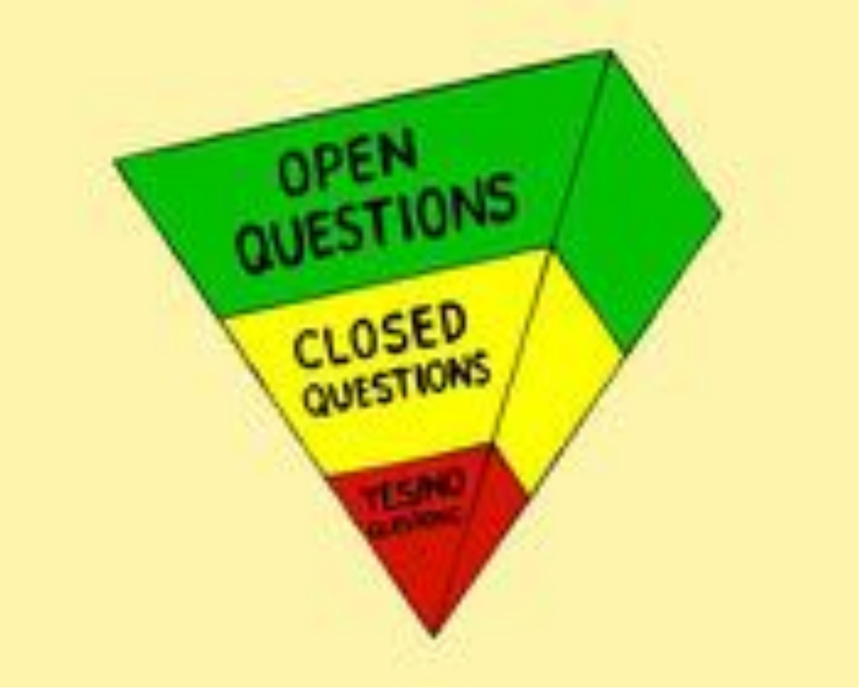


"I typically have a headache once a week. In 90% of the times, coffee helps.

# Question Modeling

- For each user's (hidden) database
  - *It's hard for the user to recall every detail*
  - *But the user can often provide useful summaries e.g., "When I catch cold, I often sleep more, drink hot water and eat pills"*

- Question Types
  - *Open Questions, e.g., "tell me about an illness and how you will treat it"*
  - *Closed Questions, e.g., "when you catch a cold, how often do you drink hot water?"*



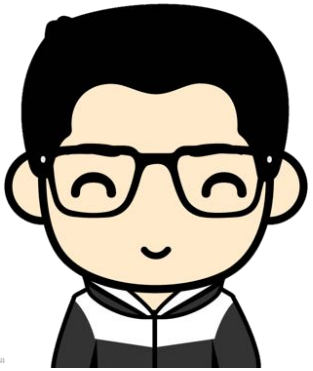
# Question Modeling (cont’d)

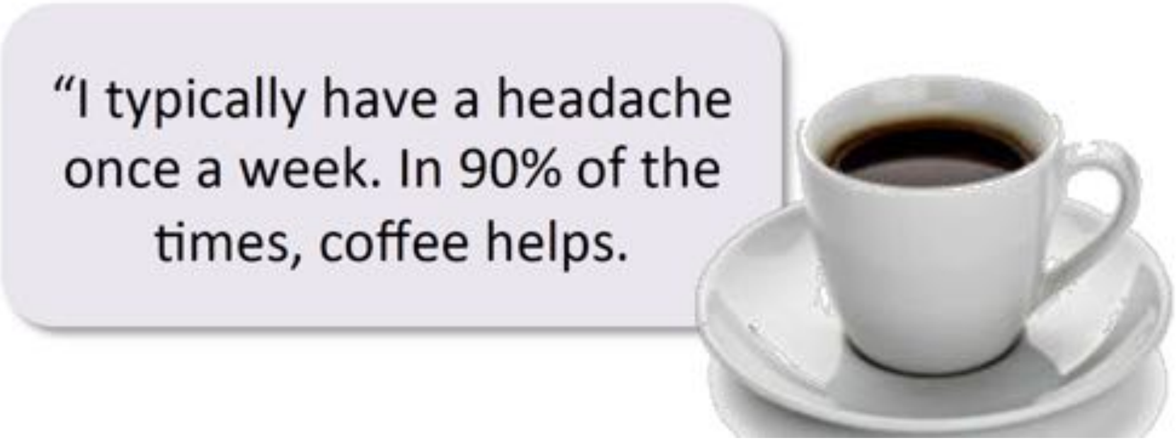
○ **Open Questions:** $? \xrightarrow{?} ?$

**Answer:** ***an arbitrary rule with its (approximate) user support and confidence***

○ **Closed Questions:**

**Answer:** ***(approximate) user support and confidence***





“I typically have a headache once a week. In 90% of the times, coffee helps.

# Goals of Crowd Mining

- **Overall Goals**

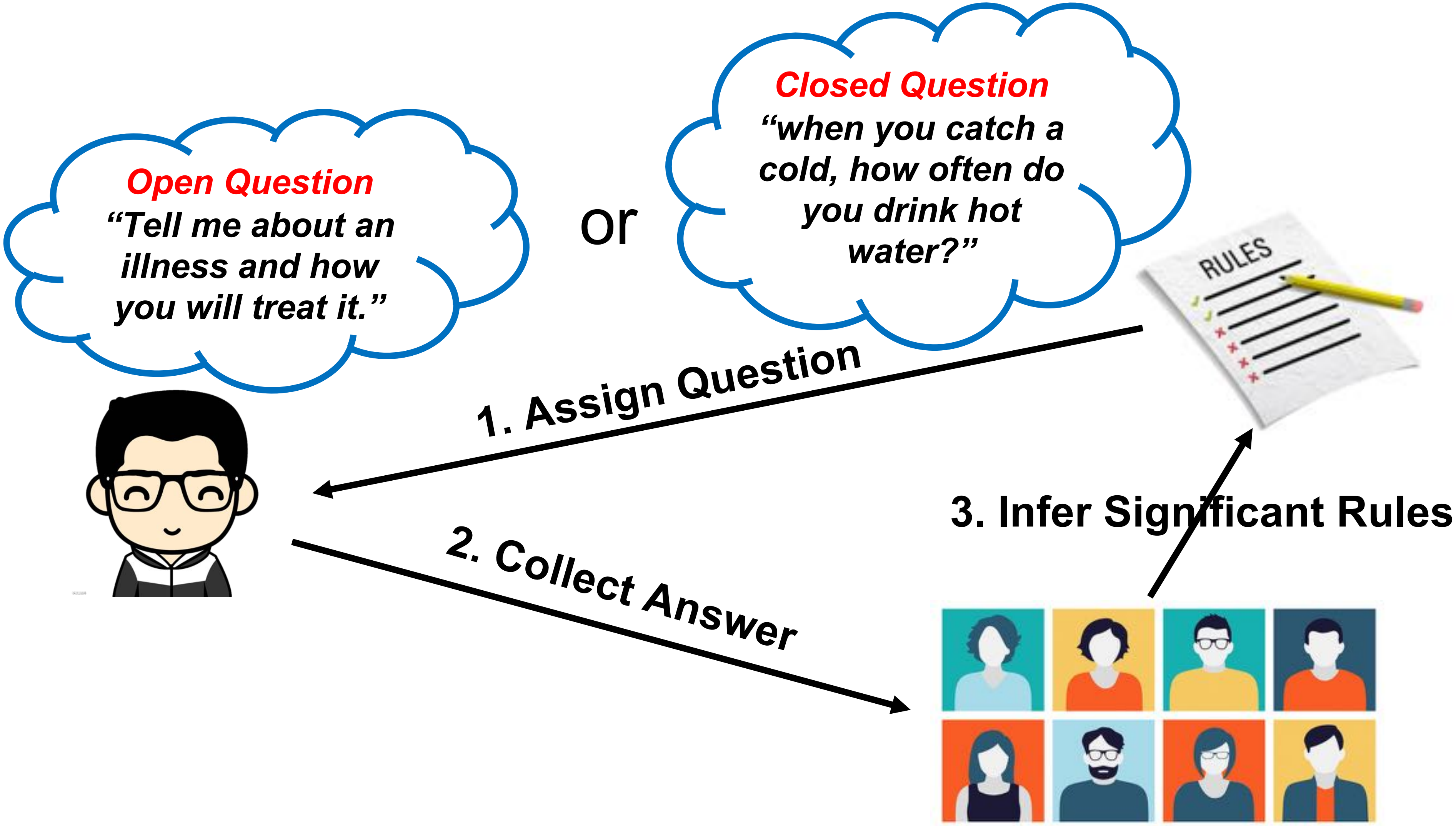
***Ask the smallest number of questions to find the significant rules***

***Rules where the user support and user confidence are above some pre-defined thresholds***

***e.g., user support > 0.4, user confidence > 0.7***

# Overall Framework

○ **Finding significant rules in illness**



***Open Question***

***“Tell me about an illness and how you will treat it.”***

or

***Closed Question***

***“when you catch a cold, how often do you drink hot water?”***

**1. Assign Question**

**2. Collect Answer**

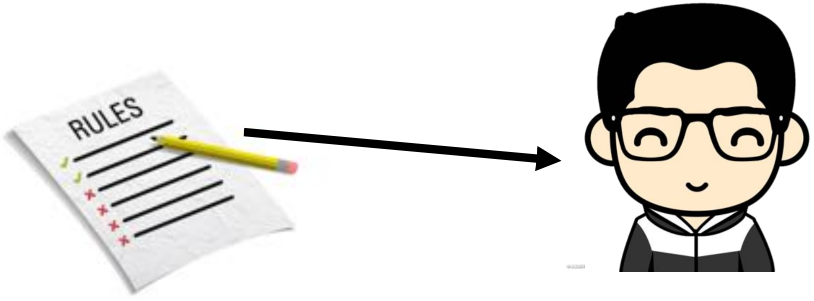
**3. Infer Significant Rules**

# Two Important Problems

- **Aggregation Problem**

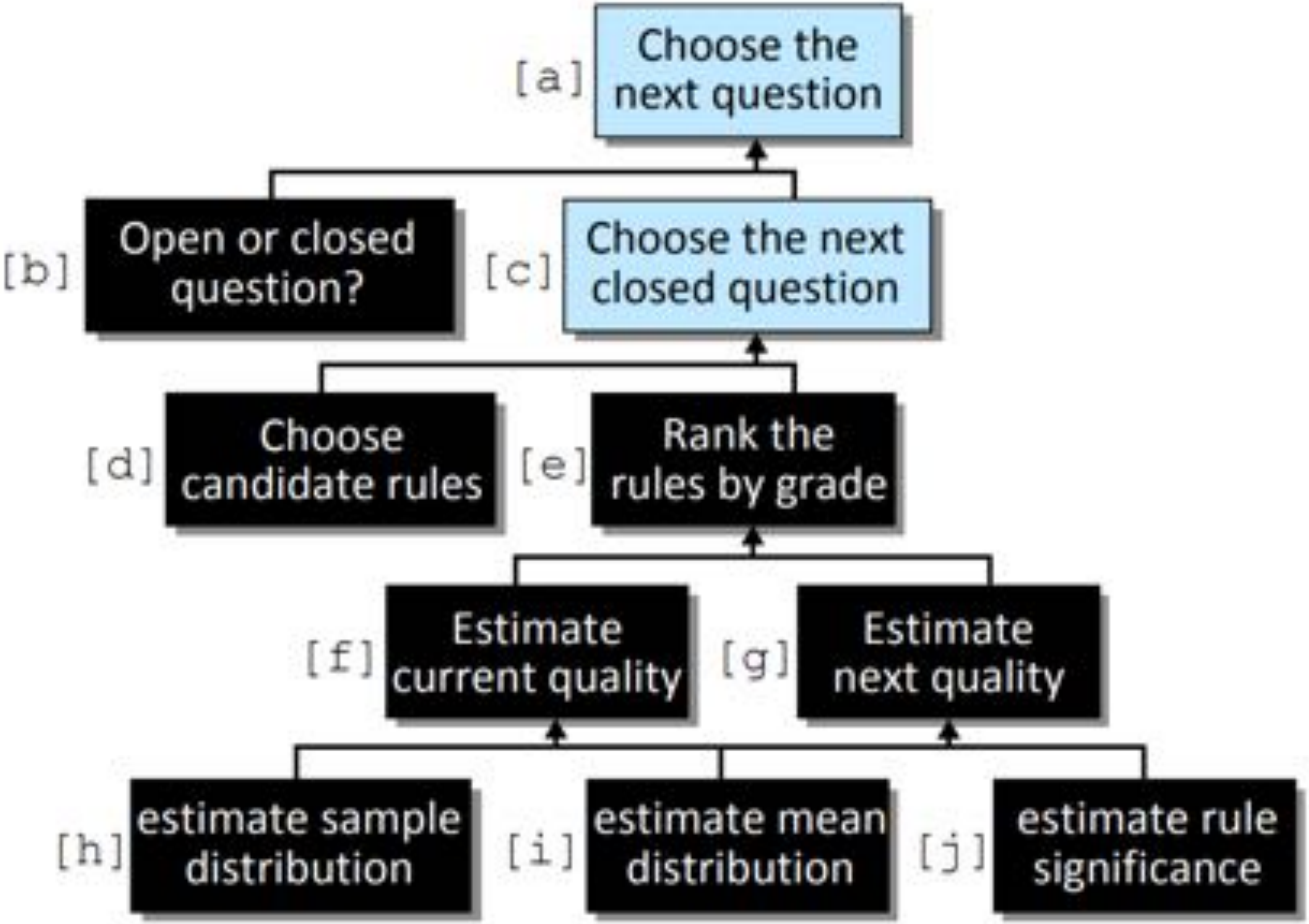
*How to compute the significant rules based on workers’ answers?*

- **Assignment Problem**



*Which rule should be chosen to assign when a worker comes?*

# Solution Framework


[a] Choose the next question
[b] Open or closed question?
[c] Choose the next closed question
[d] Choose candidate rules
[e] Rank the rules by grade
[f] Estimate current quality
[g] Estimate next quality
[h] estimate sample distribution
[i] estimate mean distribution
[j] estimate rule significance

# Aggregation Problem

- Estimating **Sample Mean**

**Define a rule $r : A \rightarrow B$ , its support $s$ , confidence $c$**

**The sample mean $f_r(s,c)$ follows the distribution**

$$f_r \sim \mathcal{N}(\mu, \frac{1}{N}\Sigma)$$

**where $N$ is #answers, $\mu$ is the mean, $\Sigma$ is the covariance**

- Estimating **Rule Significance**

**Define $\theta_s$ and $\theta_c$ as the thresholds for support and confidence, then the significance is represented as**

$$sig(r) = \int_{\theta_s}^{\infty} \int_{\theta_c}^{\infty} f_r(s,c) \ \mathrm{dc} \ \mathrm{ds}$$

# Assignment Problem

- Estimate Current Quality for Each Rule r

  e.g., $Q = \text{sig}(r)$, defined above

- Estimate Next Quality for Each Rule r

  Generate a new sample based on the current distribution, and estimate ***expected next quality*** based on the sample: $Q' = E[\ \text{sig}(r) \mid \text{sample}\ ]$

- Final Ranking of Rules

  Rank the rules based on the values of $Q' - Q$

# Outline

- ○ Crowdsourcing Overview (20min)
- ○ Fundamental Techniques (90min)
  - – Quality Control (40min)
  - – Cost Control (30min)
  - – Latency Control (20min)

Part 1

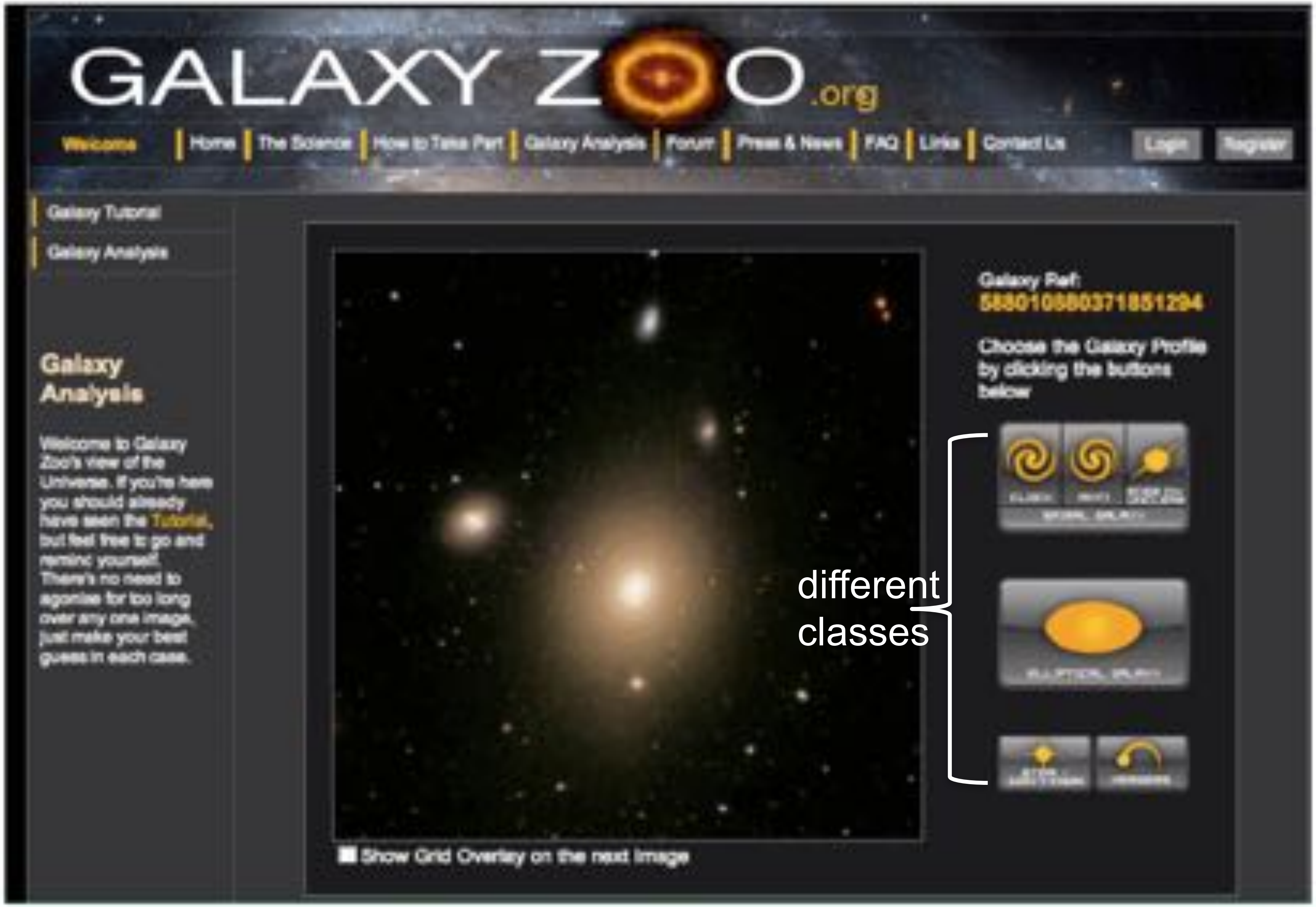
- ○ Crowd-powered Data Mining (60min)
  - – Crowd-powered Pattern Mining (10min)
  - ☞ – Crowd-powered Classification (10min)
  - – Crowd-powered Clustering (10min)
  - – Crowd-powered Machine Learning (10min)
    - • Deep learning
    - • Transfer learning
    - • Semi-supervised learning
  - – Crowd-powered Knowledge Discovery (20min)
- ○ Challenges (10min)

Part 2

# Crowd-powered Classification

Galaxy Zoo

# Crowd-powered Classification



**Galaxy Analysis**

Welcome to Galaxy Zoo's view of the Universe. If you're here you should already have seen the Tutorial, but feel free to go and remind yourself. There's no need to agonise for too long over any one image, just make your best guess in each case.

Galaxy Ref:
588010880371851294

Choose the Galaxy Profile by clicking the buttons below

□ Show Grid Overlay on the next image

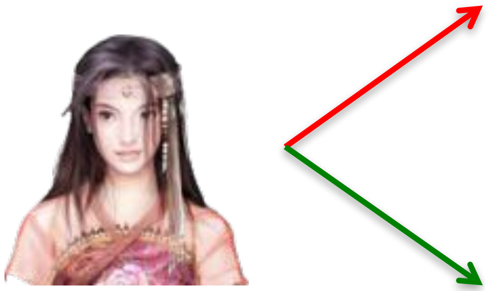
different classes

# Crowd-powered Classification

- Overview
  - Machine Learning-based Model
    - Model workers' quality, answers and features
  - Hierarchical Taxonomy
    - Classification based on taxonomy
  - Scale up to large dataset
    - Use active learning approach

# Truth Inference Model

A Two-coin Model:



False positive rate:

$$\beta^j := \Pr[y^j = 0 | y = 0].$$

j-th worker's answer

True positive rate:

True label

$$\alpha^j := \Pr[y^j = 1 | y = 1].$$

Limitation of existing truth inference models:

- Only consider the answers.
- Neglect the features on tasks.

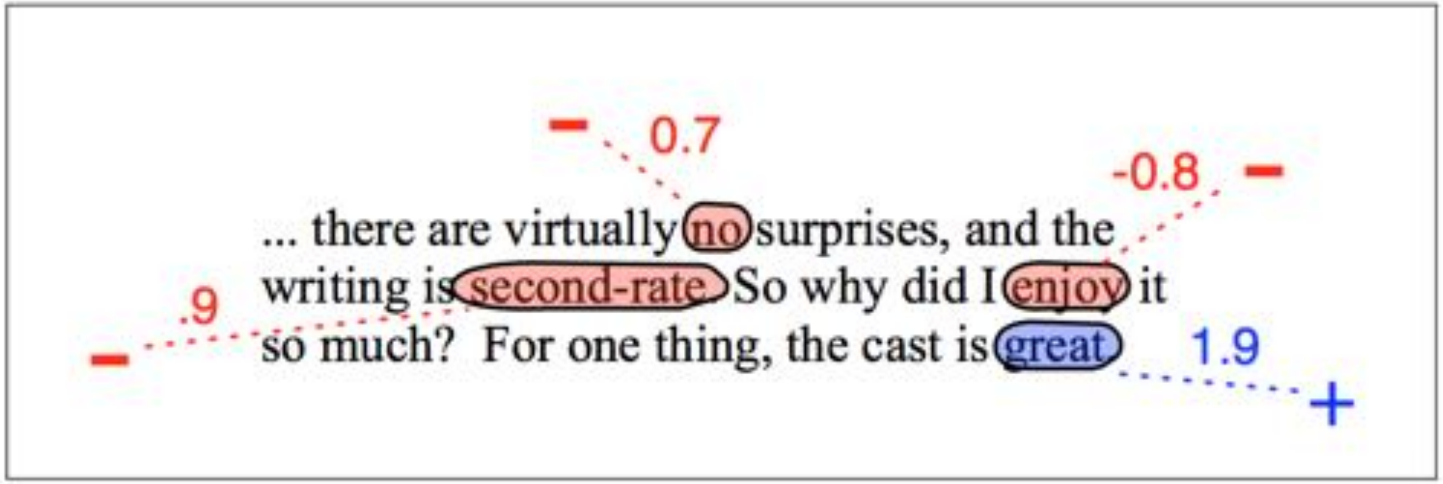
# Classification based on features

Logistic regression model: consider features of data itself

$$\Pr[y=1|\boldsymbol{x},\boldsymbol{w}] = \sigma(\boldsymbol{w}^{\top}\boxed{\boldsymbol{x}}) \qquad \sigma(z) = 1/(1+e^{-z})$$

features of the instance

Sentiment classification example：



... there are virtually no surprises, and the writing is second-rate. So why did I enjoy it so much? For one thing, the cast is great.

# Maximum Likelihood Estimator

Learning problem:

Given observed training data *D* with *N* instances from *R* workers, the task is to

- Estimate the weight vector *w.*
- Estimate the true/false positive rate of each worker.
- Infer the true classification of each instance.

Solved by EM

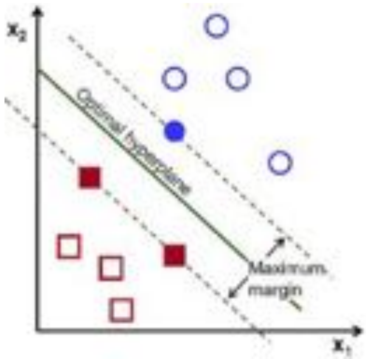
$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \Pr[y_i^1, \ldots, y_i^R | \boldsymbol{x}_i, \theta]. \quad \theta = \{\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$$

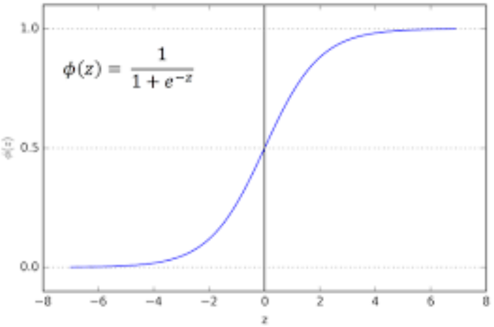
$$\begin{aligned}\Pr[\mathcal{D}|\theta] &= \prod_{i=1}^{N} \{\Pr[y_i^1, \ldots, y_i^R | y_i = 1, \boldsymbol{\alpha}]\Pr[y_i = 1 | \boldsymbol{x_i}, \boldsymbol{w}] \\ &+ \Pr[y_i^1, \ldots, y_i^R | y_i = 0, \boldsymbol{\beta}]\Pr[y_i = 0 | \boldsymbol{x_i}, \boldsymbol{w}]\}.\end{aligned}$$

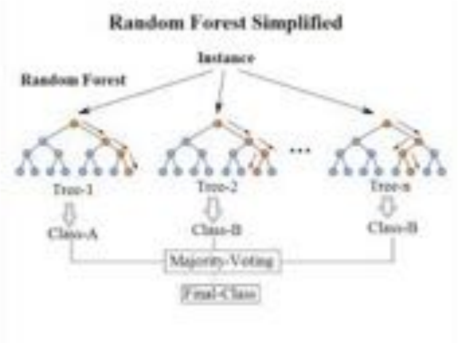
Vikas C. Raykar et.al. Learning from the Crowd. JMLR 2010

# Extensions

- Easy to use any classifier and handle missing labels.





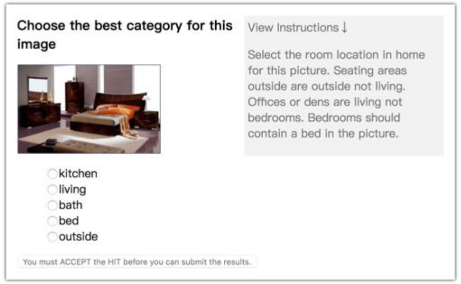


- A beta prior for workers

I trust her more

$$
\begin{aligned}
\Pr[\alpha_j | a_1^j, a_2^j] &= \mathrm{Beta}(\alpha_j | a_1^j, a_2^j). \\
\Pr[\beta_j | b_1^j, b_2^j] &= \mathrm{Beta}(\beta_j | b_1^j, b_2^j).
\end{aligned}
$$

- Easy to extend to multi-class classification



Choose the best category for this image

View Instructions ↓

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

○ kitchen
○ living
○ bath
○ bed
○ outside

You must ACCEPT the HIT before you can submit the results.

$$\alpha_{ck}^j := \Pr[y^j = k | y = c]$$

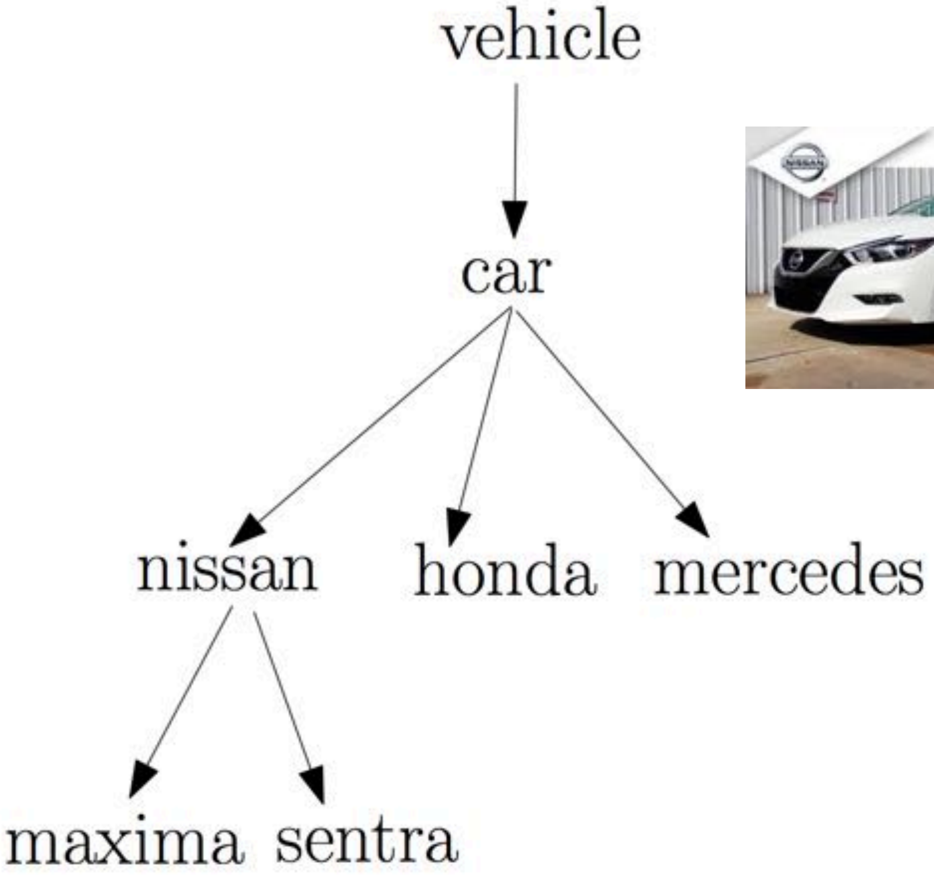
Given the true class *c*, worker *j* assigns class *k* to an instance

# Crowd-powered Classification

- Overview
  - Machine Learning-based Model
    - Model workers’ quality, answers and features
  - Hierarchical Taxonomy
    - Classification based on taxonomy
  - Scale up to large dataset
    - Use active learning approach

# Classification on Hierarchical Taxonomy

Categorize an image into one of the classes of the hierarchical taxonomy



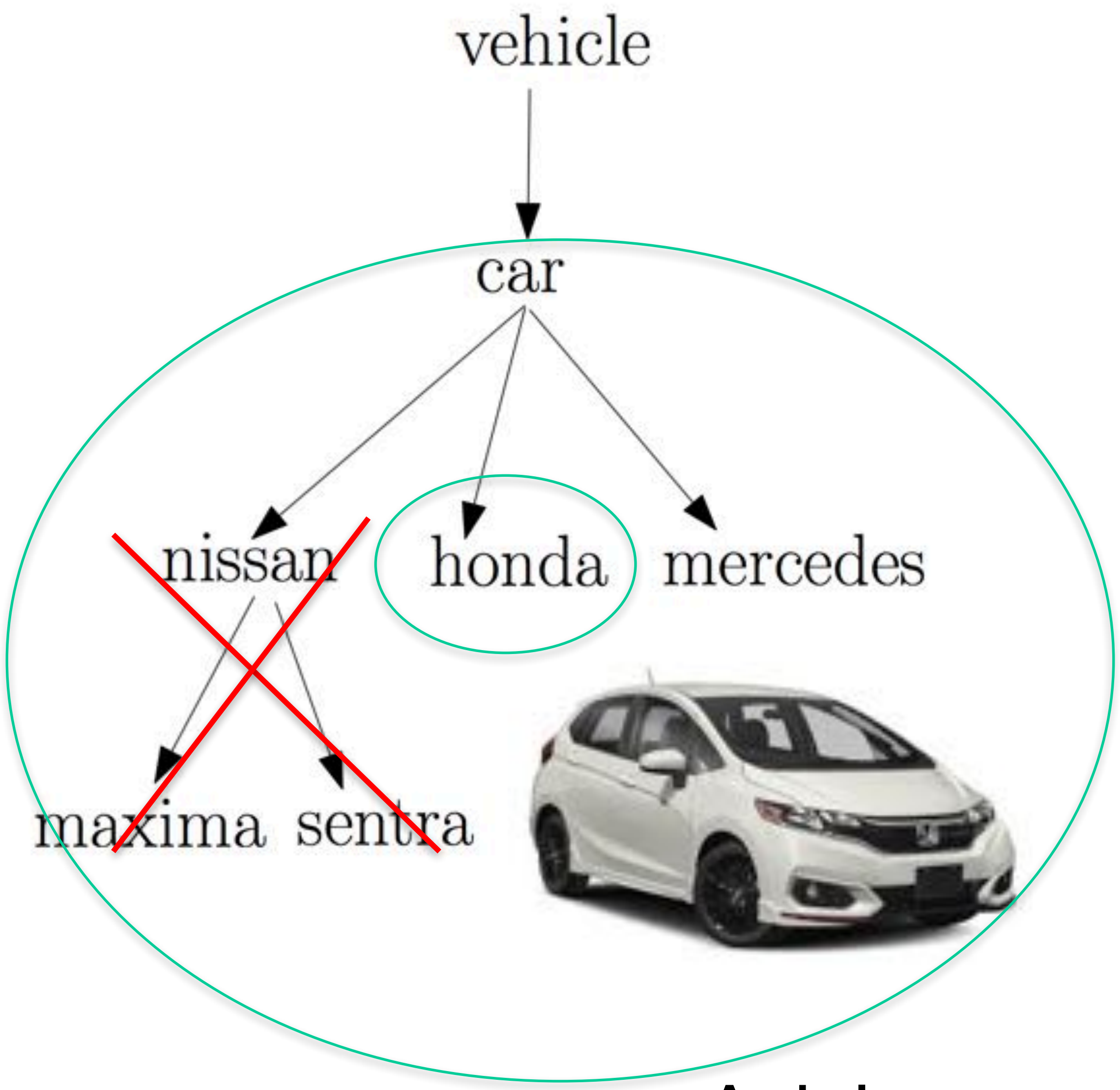
Is it a car?

Is it a Nissan car？

Is it Maxima？

## Application

- Image Categorization
- Manual Curation
- Debugging of Workflows



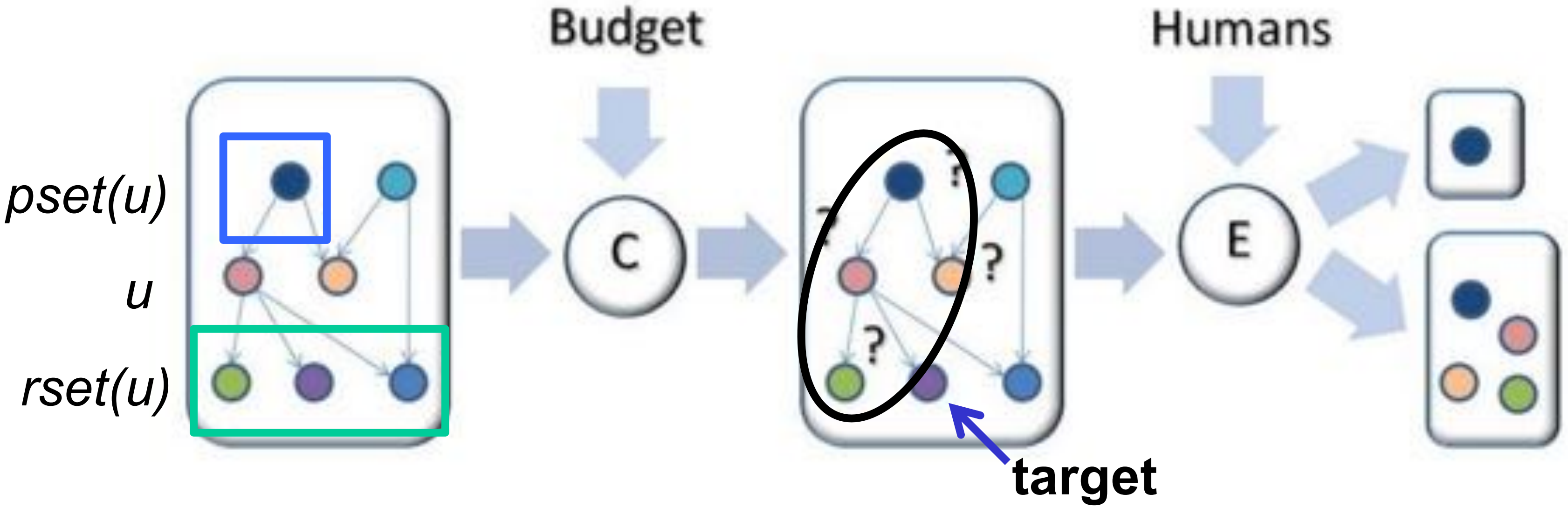
Is it a car?

Is it a Nissan car？

Is it a Honda car？

Ask leaves: negative answers
Ask root: positive answers
Ask middle nodes:more information

# Solution Overview



Candidate set:

$$\textit{cand}(\{u\}, U^*) = \begin{cases} V - \textit{rset}(u) & \mathsf{q}(\{u\}, U^*) = \mathrm{NO} \\ V - \textit{pset}(u) & \mathsf{q}(\{u\}, U^*) = \mathrm{YES} \wedge \textit{Multi} \\ \textit{rset}(u) & \mathsf{q}(\{u\}, U^*) = \mathrm{YES} \wedge \textit{Single} \end{cases}$$

Size of the largest candidate set when the target node could be any node in V:

$$\mathsf{wcase}(N) = \max_{u_i \in V} |\mathsf{cand}(N, u_i)|$$

Find a set of *N* to minimize *wcase(N)*

# Crowd-powered Classification

- Overview
  - Machine Learning-based Model
    - Model workers’ quality, answers and features
  - Hierarchical Taxonomy
    - Classification based on taxonomy
  - Scale up to large dataset
    - Use active learning approach

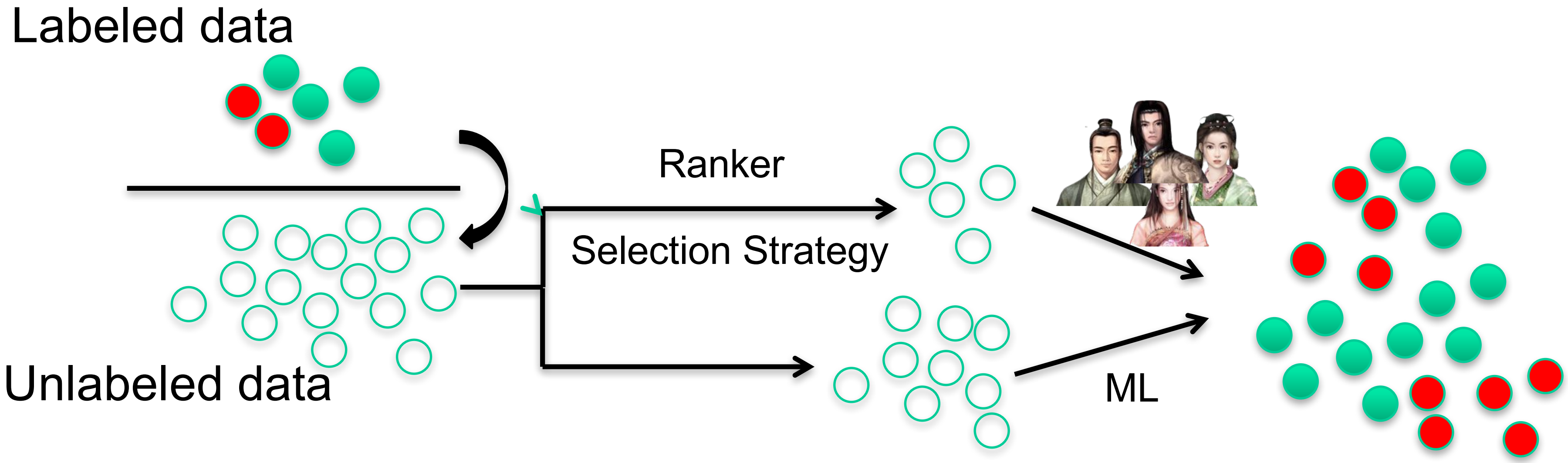
# Scaling up to large dataset

Solutions that solely rely on crowdsourcing are always limited to small datasets.

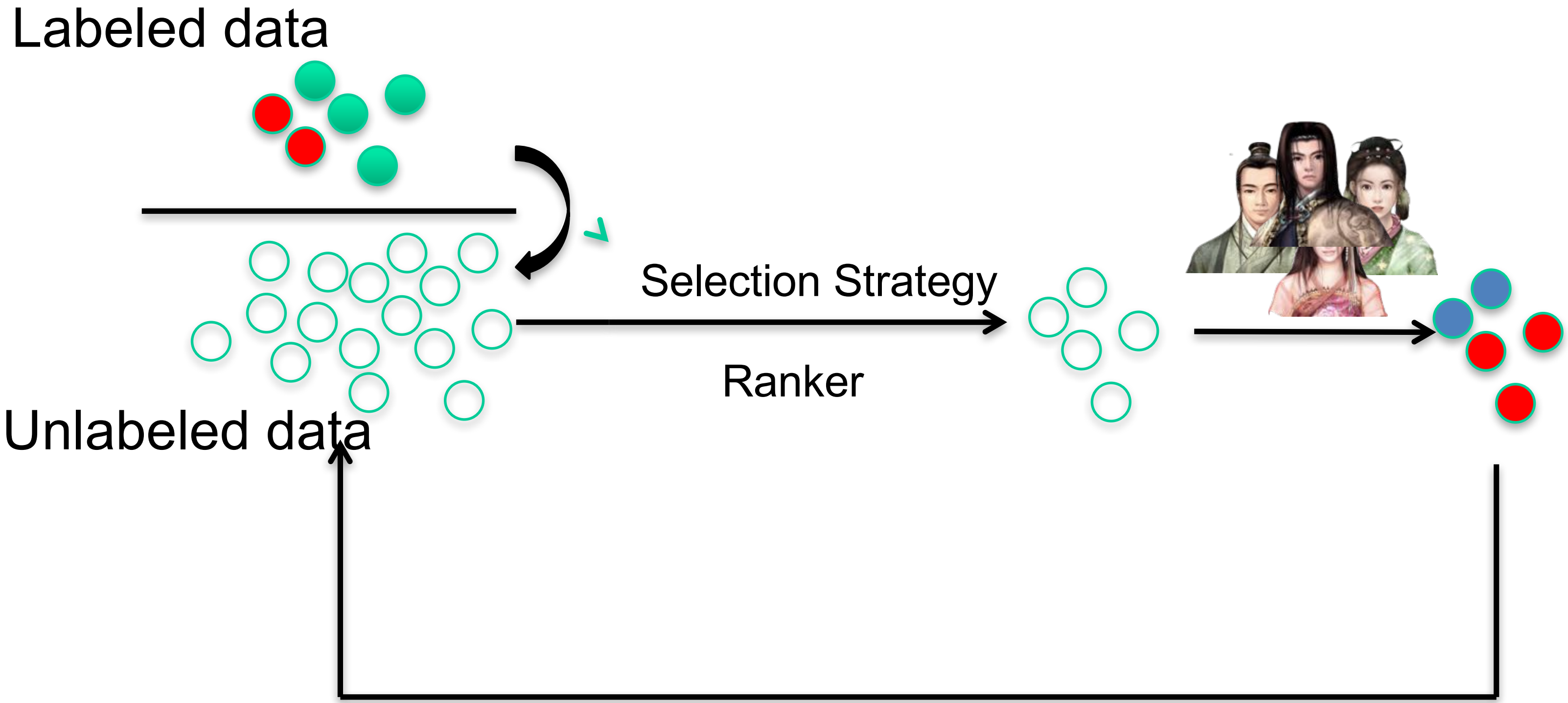
Active Learning

- Generality: can use any classifier
- Black-box treatment of classifier
- Batching: request multiple labels at a time.
- Noise management: Handling human errors.

# Upfront Scenario in Active Learning



# Iterative Scenario in Active Learning

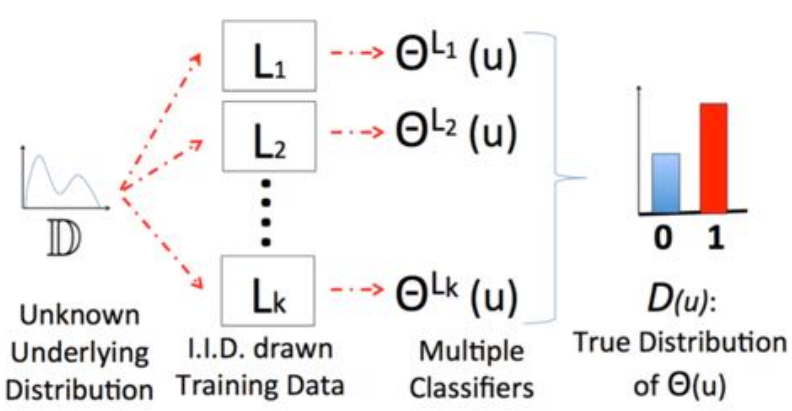


Barzan Mozafari et.al Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning VLDB 2014

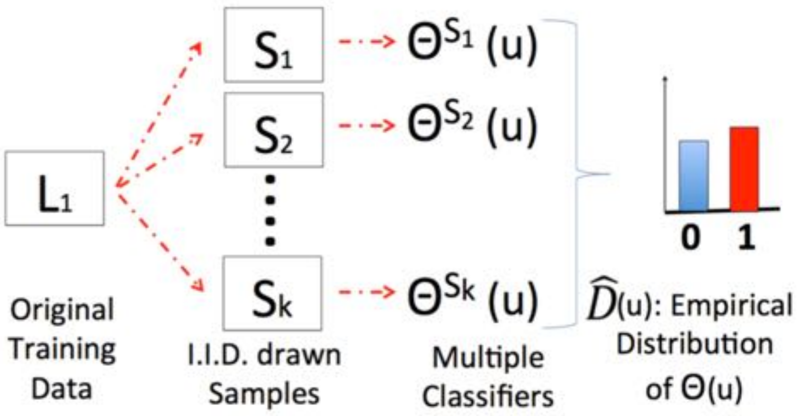
# Ranker

Uncertainty Algorithm: use bootstrap to verify errors of classifiers

$\theta^L(u)$: $\theta$: the classifier $L$: Training data $u$: data point to be predicted



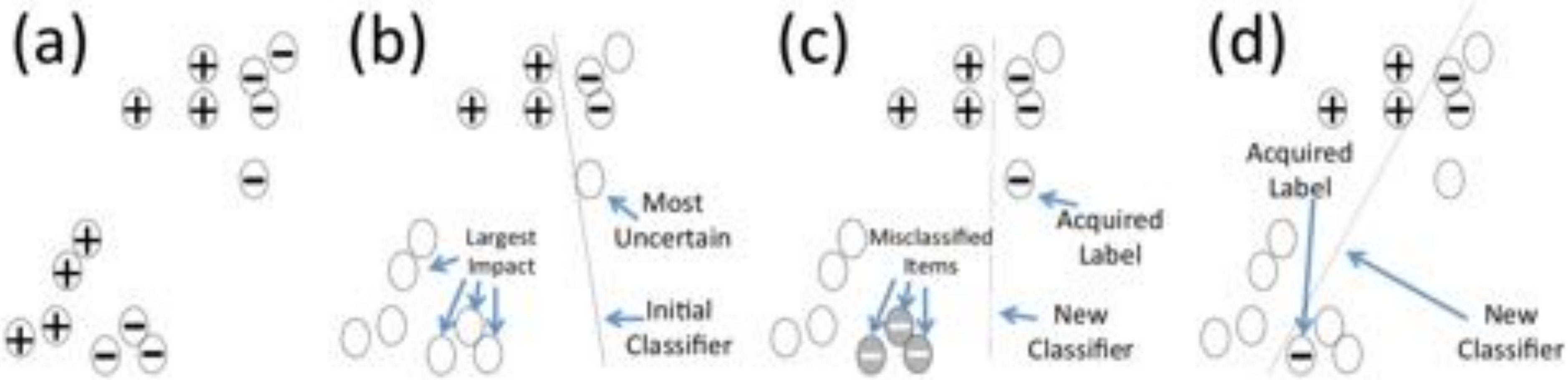
(a) Ideal computation of $D(u)$



(b) Bootstrap computation

# Ranker

MinExpError Algorithm: consider both uncertain and large impact data points



$$
\begin{aligned}
MinExpError(u) &= \hat{p}(u)\hat{e}_{\text{right}} + (1 - \hat{p}(u))\hat{e}_{\text{wrong}} \\
&= \hat{e}_{\text{wrong}} - \hat{p}(u)(\hat{e}_{\text{wrong}} - \hat{e}_{\text{right}})
\end{aligned}
$$

# Take-Away for Crowd Classification

- Different datasets need different classification approaches
  - Simple truth inference approach
  - Feature-based classification using the crowd
  - Hierarchical Taxonomy
  - Large datasets
- Handling human errors

# Outline

- Crowdsourcing Overview (20min)
- Fundamental Techniques (90min)
  - Quality Control (40min)
  - Cost Control (30min)
  - Latency Control (20min)
- Crowd-powered Data Mining (60min)
  - Crowd-powered Pattern Mining (10min)
  - Crowd-powered Classification (10min)
  - ☞ Crowd-powered Clustering (10min)
  - Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - Crowd-powered Knowledge Discovery (20min)
- Challenges (10min)

Part 1

Part 2

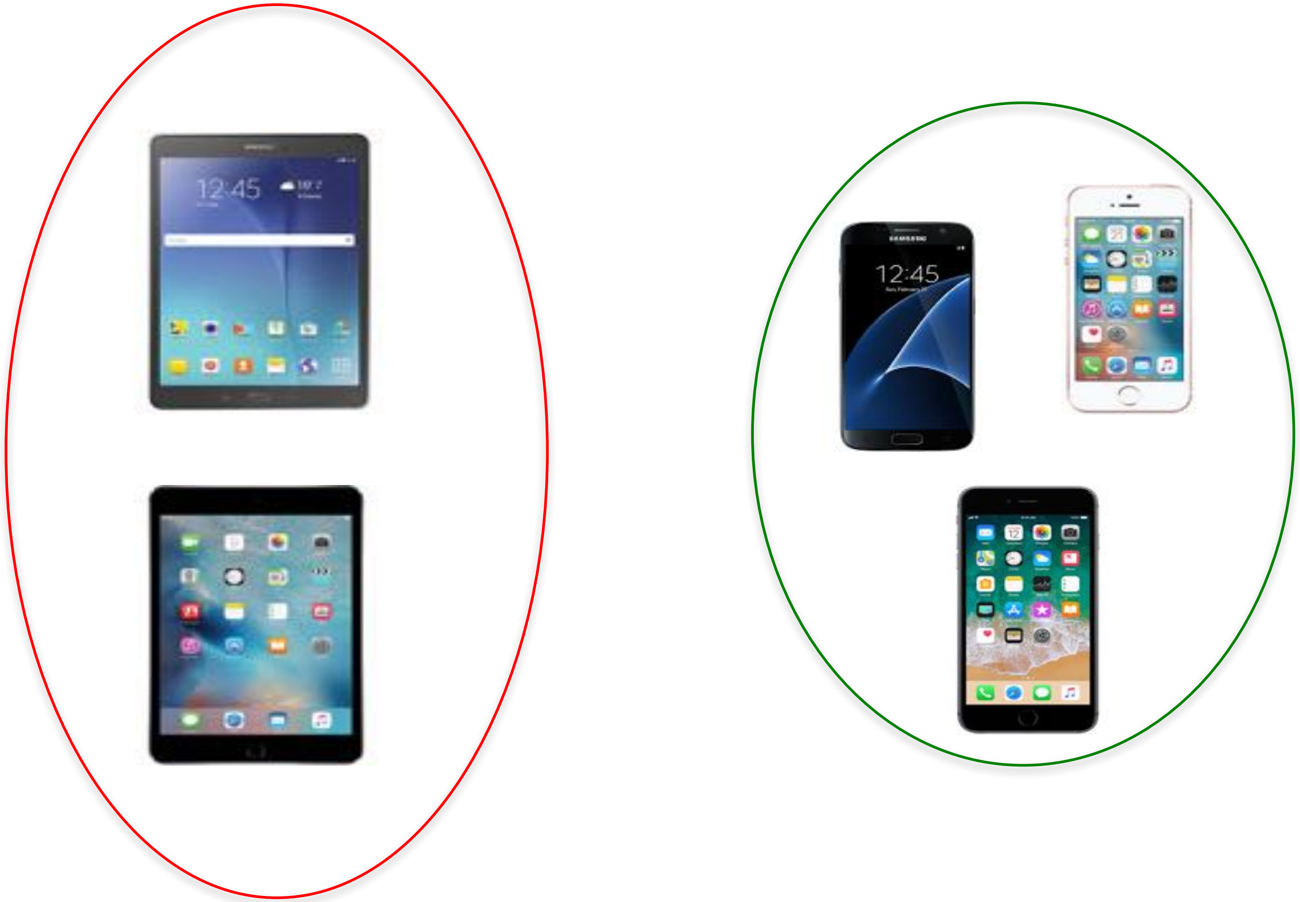
# Crowd-powered Clustering

Easy to cluster by machine

Hard to cluster by machine

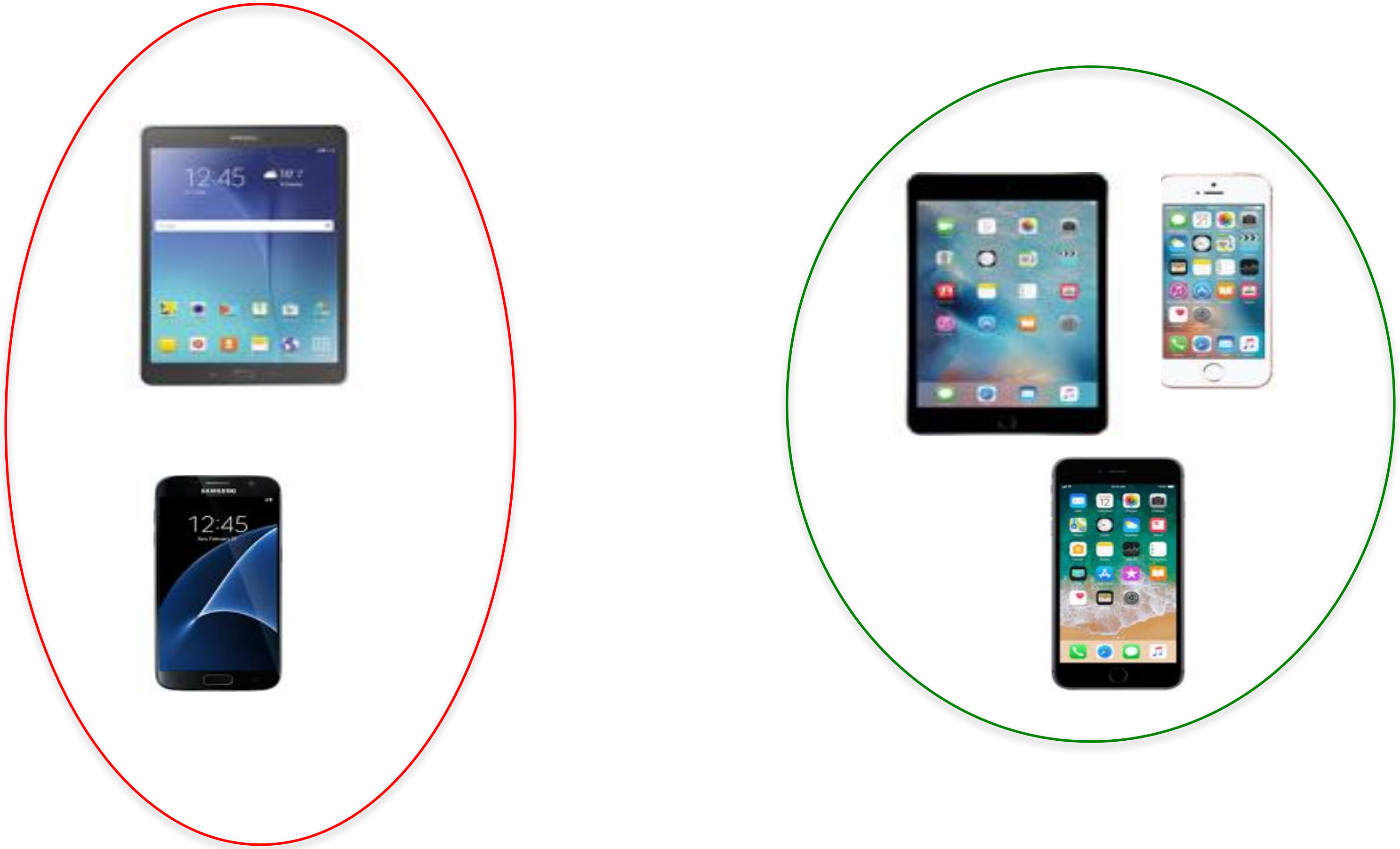
# Clustering based on different human insights

## Crowd may cluster by types of products



# Clustering based on different human insights

## Crowd may cluster by brands of products



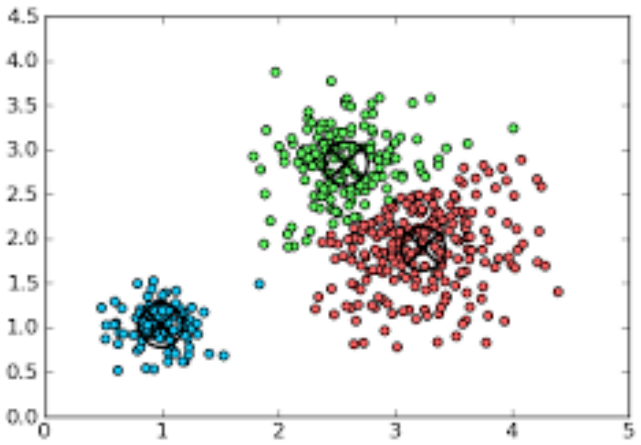
# Crowd-powered Clustering

- Overview
  - Kmeans-based Model
  - Generative Model based on different human insights

# A K-means Based Approach

Standard K-means Algorithm：

**Assign：** Given a set of items $C \subset D$ and an item $x \in D$， find the item $c \in C$ that is the closest to $x$ according to the distance function $d$



**Update：** Given a set of items $C \subseteq D$ , find the “center” of C,that is,the item x $\in$ C that minimizes $\sum_{c \in C} d(x,c)$

Hannes Heikinheimo et.al The Crowd-Median Algorithm HCOMP 2013

# Crowd-based Solution

**Assign:** Show the worker all items in C,*as well as the item x ∈ D, and ask her to pick one in C that resembles x the most.*

Which one resembles the left pad most ?

**Update:**

Which one differs the other two most ?

- Pick about 20% of triplets from D
- Out of three shown items pick one that appears to be different from the two others.
- Compute a penalty score defined as the number of times the item was chosen to be "different".
- Return the item having the lowest penalty score

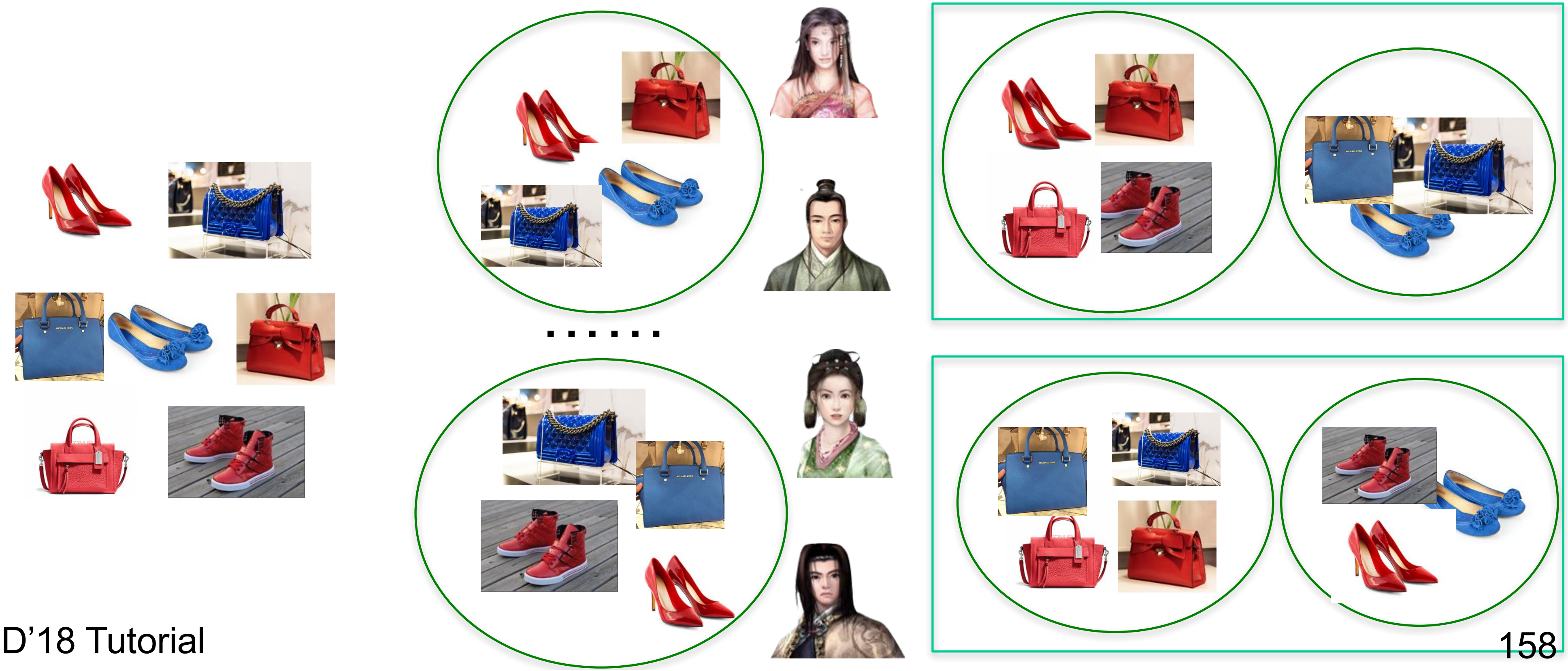
# Crowd-powered Clustering

- Overview
  - Kmeans-based Model
  - Generative Model based on different human insights

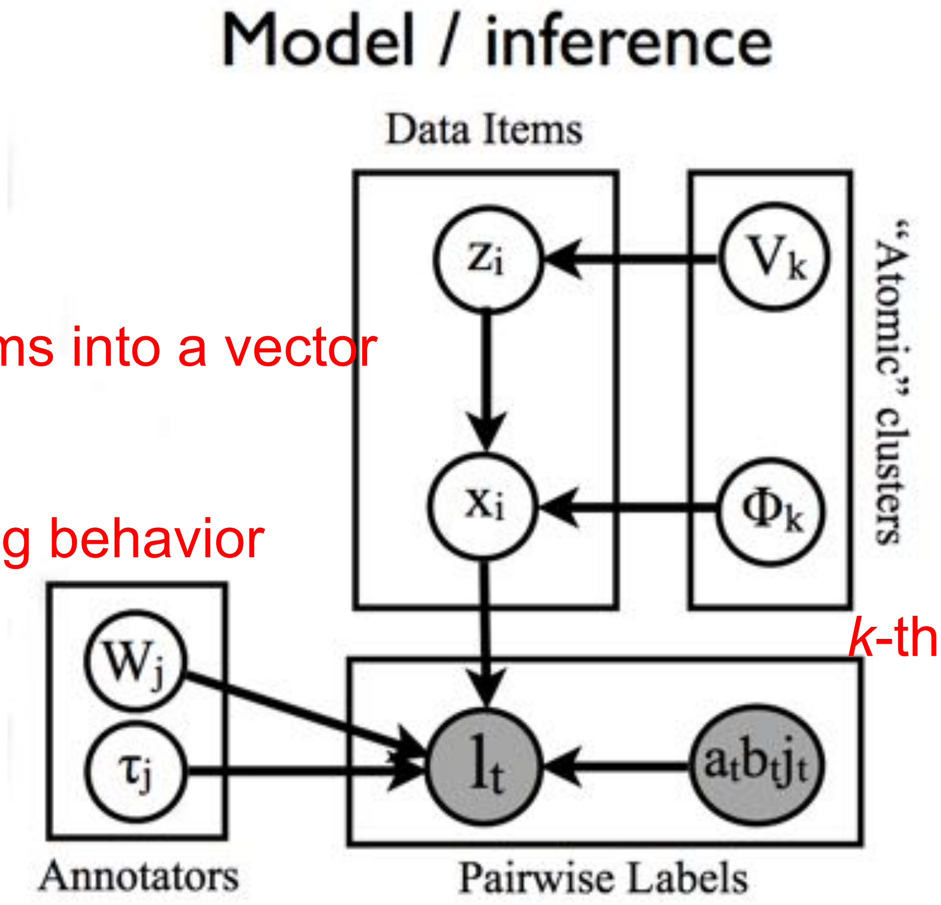
# Generative Model based on different human insights

## Workflow

- Sample  a number of small groups of items
- Leverage the crowd to cluster these small groups
- Aggregate the crowd answers and infer the true clusters of the dataset



# Aggregation: Generative Model



Embedding items into a vector

Workers' labeling behavior

*k*-th Gaussian atomic cluster

Crowd annotators

Ryan Gomes et.al Crowdclustering NIPS 2011

# Take-Away for Crowd Clustering

- Challenges
  - We can’t let users to see all items in the datasets !
- Key ideas:
  - Sample small groups and show them to the crowd
  - Infer the truth based on different clusters

# Outline

- ○ Crowdsourcing Overview (20min)
- ○ Fundamental Techniques (90min)
  - – Quality Control (40min)
  - – Cost Control (30min)
  - – Latency Control (20min)
- ○ Crowd-powered Data Mining (60min)
  - – Crowd-powered Pattern Mining (10min)
  - – Crowd-powered Classification (10min)
  - – Crowd-powered Clustering (10min)
  - – Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - – Crowd-powered Knowledge Discovery (20min)
- ○ Challenges (10min)

Part 1

Part 2

# Machine Learning with Crowd

- Overview
  - Deep learning from the crowd
    - A crowd layer
  - Transfer Learning using the Crowd
    - Crowd selection on Twitter
  - Semi-supervised Learning using the Crowd
    - Training using crowds and unlabeled data
  - HMM-based Crowd Model
    - Model workers' behaviors with different rewards

# Deep Learning from the Crowd

- Classification or regression for items with high dimension features

  deep learning

- Large training data

  Crowdsourcing

- Need to consider workers' reliability

  EM algorithm

Rodrigues et.al. Deep Learning from Crowds. AAAI 2018

# Deep Learning from the Crowd

$$p(\mathcal{D}, \mathbf{z}|\mathbf{\Theta}, \{\mathbf{\Pi}^r\}_{r=1}^R) = \prod_{n=1}^N p(z_n|\mathbf{x}_n, \mathbf{\Theta}) \prod_{r=1}^R p(y_n^r|z_n, \mathbf{\Pi}^r).$$

## EM for deep learning

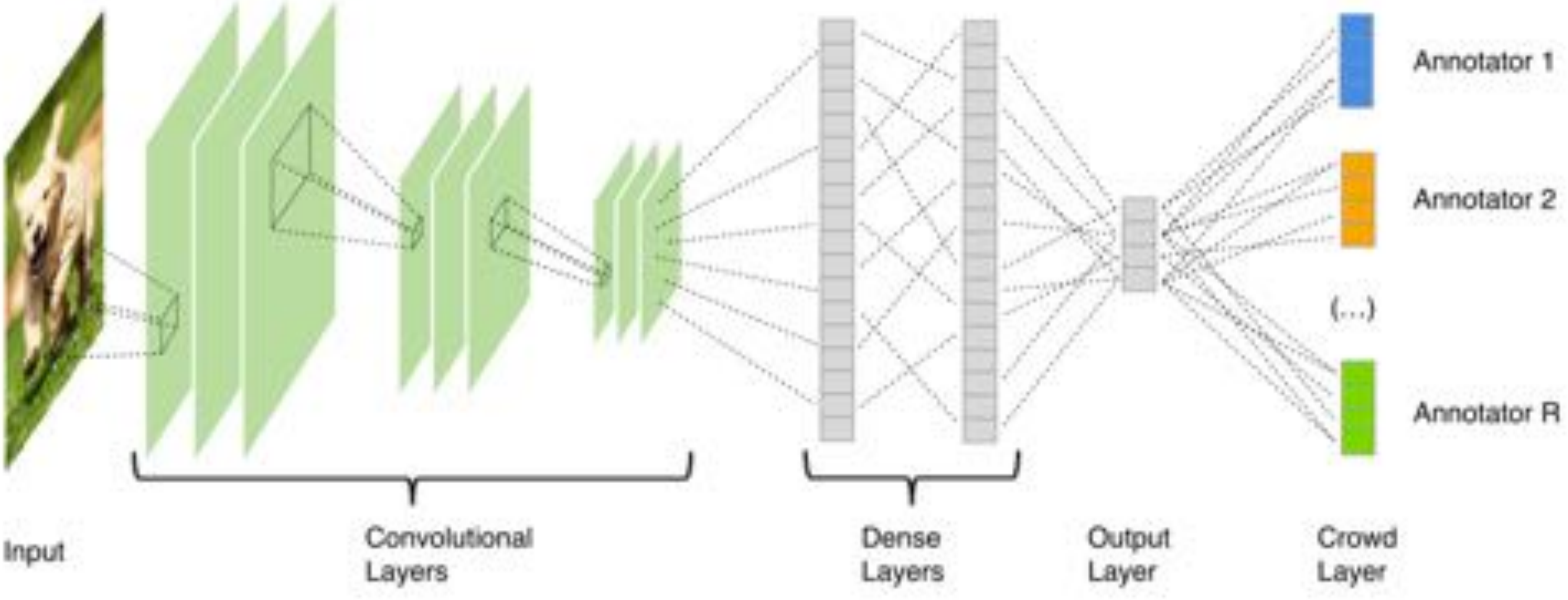
Estimate the parameters using Deep Neural Network in M step

- One EM iteration per mini-batch——No enough evidence for annotators’ reliabilities.

- Many EM iterations until converge——Large computational overhead

# Deep Neural Network

Provide noisy training data



- Account for unreliable annotators
- Correct systematic biases

# Machine Learning with Crowd

- Overview
  - Deep learning from the crowd
    - A crowd layer
  - Transfer Learning using the Crowd
    - Crowd selection on Twitter
  - Semi-supervised Learning using the Crowd
    - Training using crowds and unlabeled data
  - HMM-based Crowd Model
    - Model workers' behaviors with different rewards

# Crowd-Selection using Transfer Learning

Given a question, how to select workers to answer ?

Early Approaches: select randomly on well-defined crowd platform.

New trend: utilize social network as crowd platform, eg: ask your followings or followers on Twitter.

# Challenges

- Limited Expertise Information

  Infer the user expertise based on tweets.

- Large Volume of Tweets

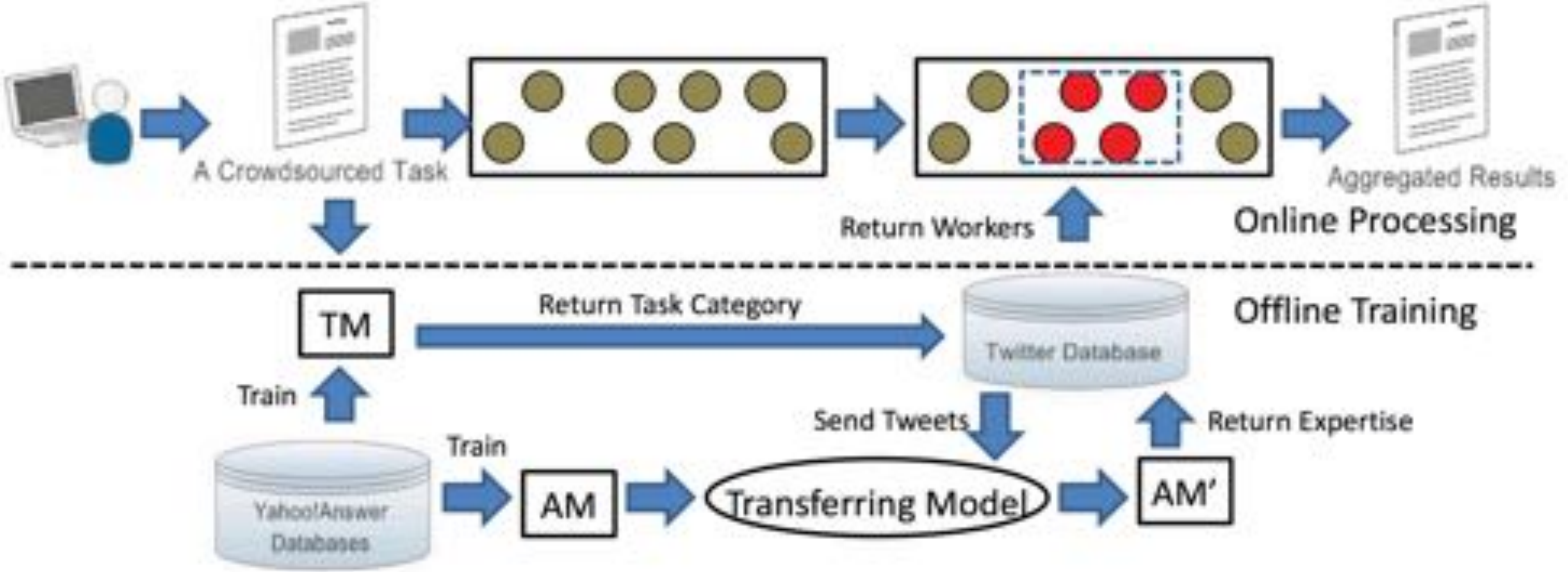
  Transfer learning from other sources.

- Requiring Online Crowd Selection

  Training offline and processing online.

Zhao et.al. A Transfer Learning based Framework of Crowd-Selection on Twitter. KDD’13

# System Overview



TM: A naïve Bayes' model based on categorized tasks from Yahoo! Answer.

AM: A naïve Bayes' model based on categorized answers from Yahoo! Answer

# Transfer Learning

Some notations

*$D_c$*: categorized answers from Yahoo!;
*$D_u$*: uncategorized ones on Twitter.
*$a \in D_c$*: an answer, can be represented as a bag of words.
*c*: a category, each answer *a* corresponds to a category *c.*
*w*: a word come from a corpus.

Basic Model: Naïve Bayes

$$\begin{aligned} p_{D_c}(c|a) &\propto p_{D_c}(c) \cdot p_{D_c}(a|c) \\ &= p_{D_c}(c) \prod_{w \in a} p_{D_c}(w|c). \end{aligned}$$

Transfer Learning Model: EM Algorithm

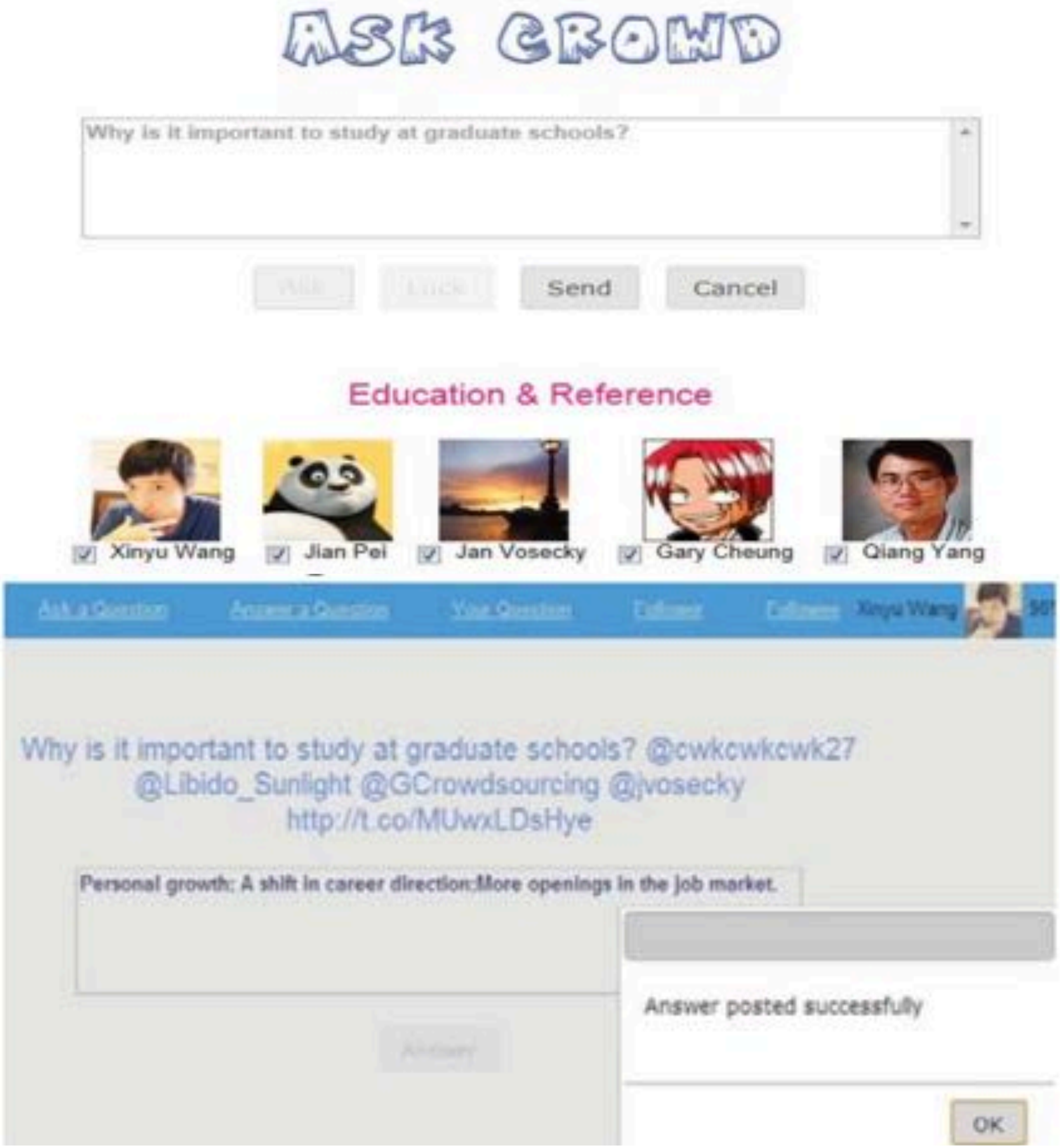
**E-step**: estimate the posterior probability of the category of tweets in *$D_u$*

$$p_{D_u}(c|d) \propto p_{D_u}(c) \prod_{w \in d} p_{D_u}(w|c).$$

**M-step**: estimate the parameter of the model AM'

$$p_{D_u}(c) \quad p_{D_u}(w|c)$$

# Selection Process



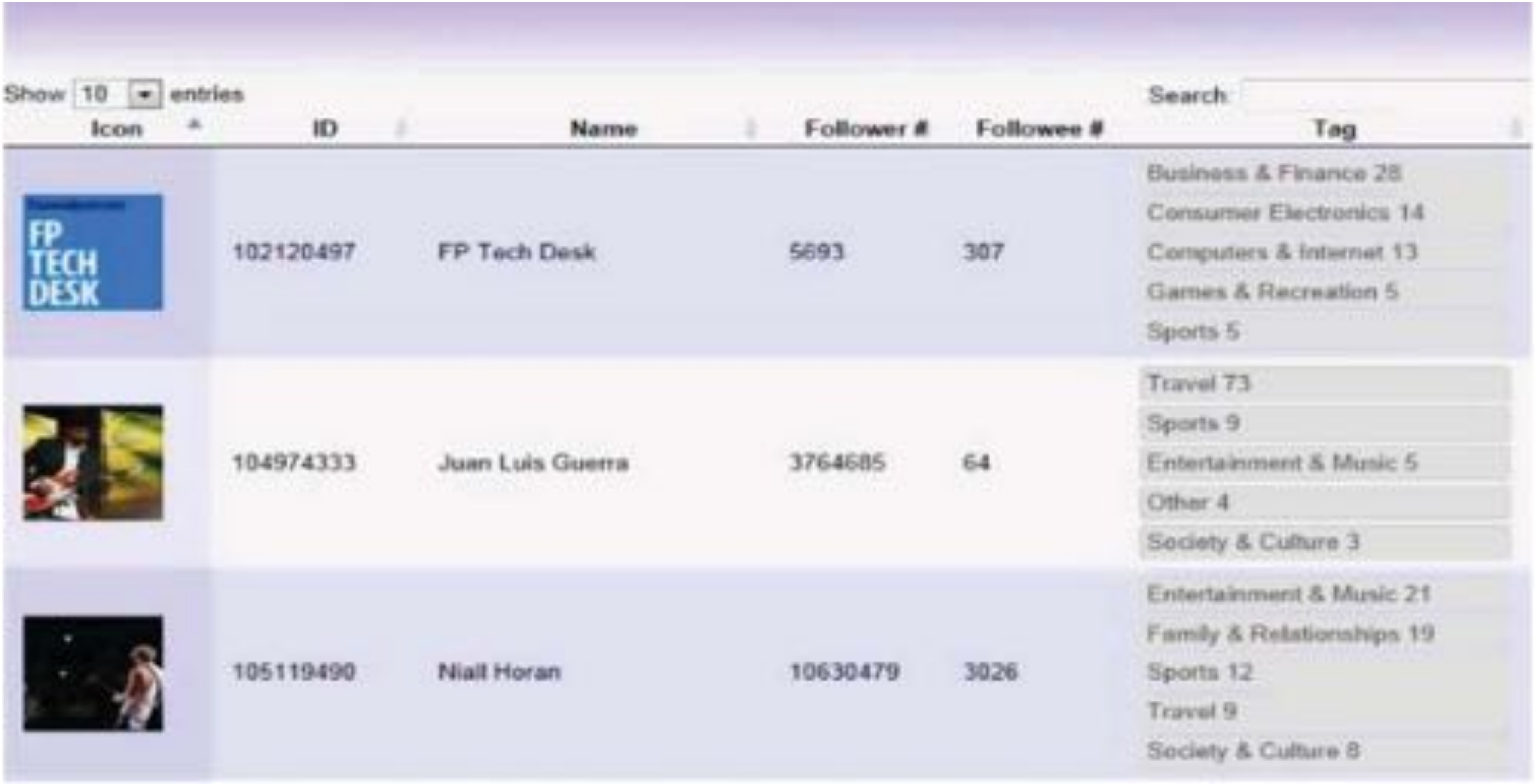
Why is it important to study at graduate schools?

Why is it important to study at graduate schools? @cwkcwkcwk27 @Libido_Sunlight @GCrowdsourcing @jvosecky http://t.co/MUwxLDsHye

Personal growth: A shift in career direction:More openings in the job market.

Answer posted successfully

# Selection Process



Show 10 entries

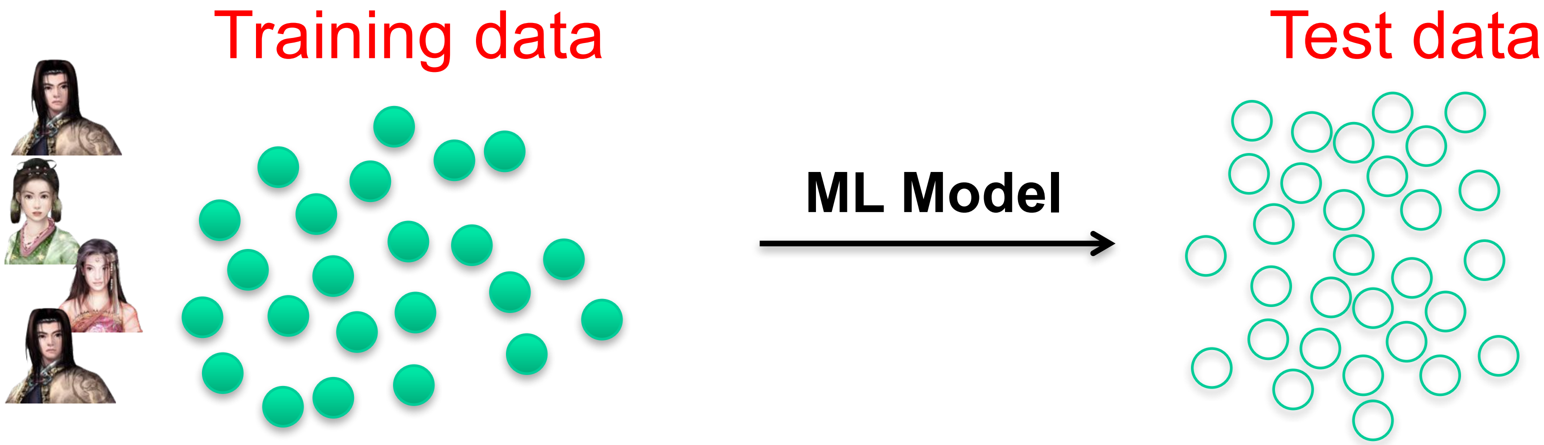
Search:

| Icon | ID | Name | Follower # | Followee # | Tag |
|---|---|---|---|---|---|
| | 102120497 | FP Tech Desk | 5693 | 307 | Business & Finance 28<br>Consumer Electronics 14<br>Computers & Internet 13<br>Games & Recreation 5<br>Sports 5 |
| | 104974333 | Juan Luis Guerra | 3764685 | 64 | Travel 73<br>Sports 9<br>Entertainment & Music 5<br>Other 4<br>Society & Culture 3 |
| | 105119490 | Niall Horan | 10630479 | 3026 | Entertainment & Music 21<br>Family & Relationships 19<br>Sports 12<br>Travel 9<br>Society & Culture 8 |

# Machine Learning with Crowd

- Overview
  - Deep learning from the crowd
    - A crowd layer
  - Transfer Learning using the Crowd
    - Crowd selection on Twitter
  - Semi-supervised Learning using the Crowd
    - Training using crowds and unlabeled data
  - HMM-based Crowd Model
    - Model workers’ behaviors with different rewards

# Semi-supervised Learning from Crowds

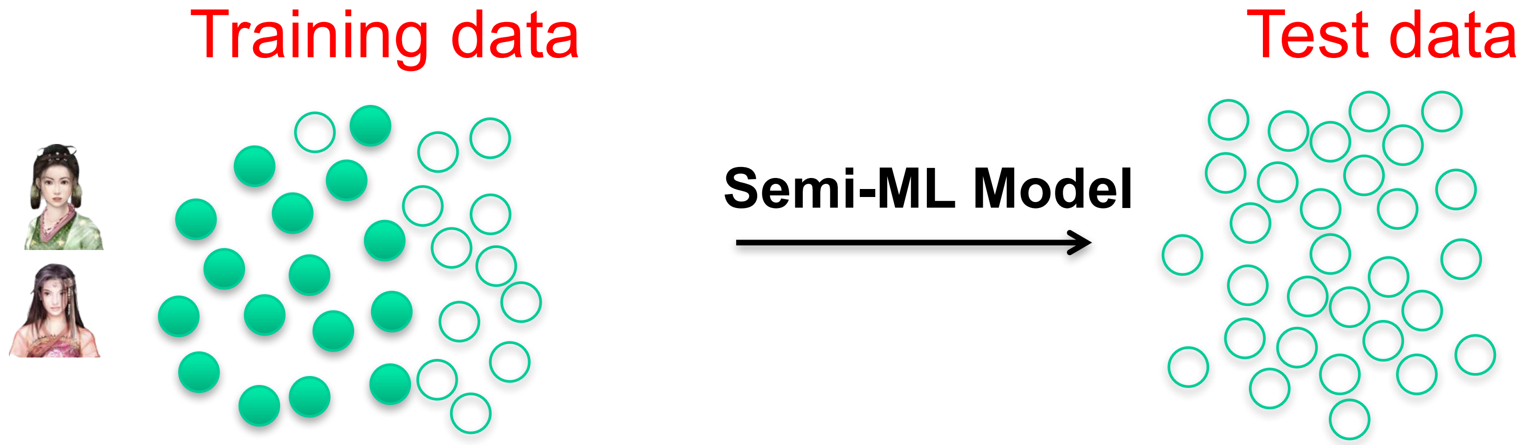


Training data

ML Model

Test data

Huge amount of data labeled by crowd workers.



Training data

Semi-ML Model

Test data

Use labeled and unlabeled data to train

# Semi-supervised Learning from Crowds

How can we utilize unlabeled data?

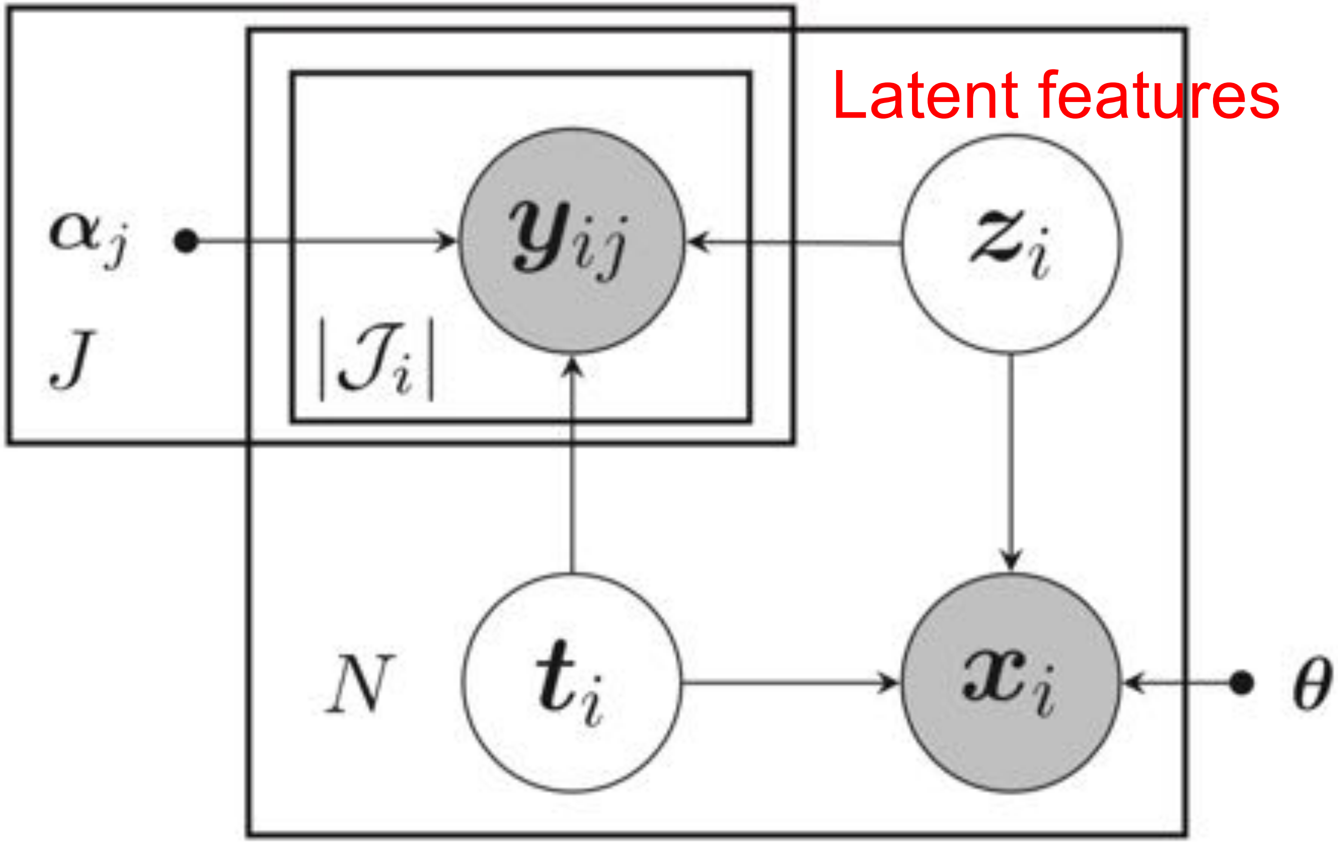
Unlabeled data $\xrightarrow{\textbf{Modeled by}}$ Latent features / Distribution of data

Worker label $\xrightarrow{\textbf{Based on}}$ Latent features / True labels

Latent variables

# Graphical Model

Worker's answer



Latent features

True label

Data point

Atarashi et.al. Semi-supervised Learning from Crowds Using Deep Generative Models AAAI'18

# Machine Learning with Crowd

- Overview
  - Deep learning from the crowd
    - A crowd layer
  - Transfer Learning using the Crowd
    - Crowd selection on Twitter
  - Semi-supervised Learning using the Crowd
    - Training using crowds and unlabeled data
  - HMM-based Crowd Model
    - Model workers’ behaviors with different rewards

# HMM-based Crowd Model

Q1 Q2 Q3 Q4 Q5

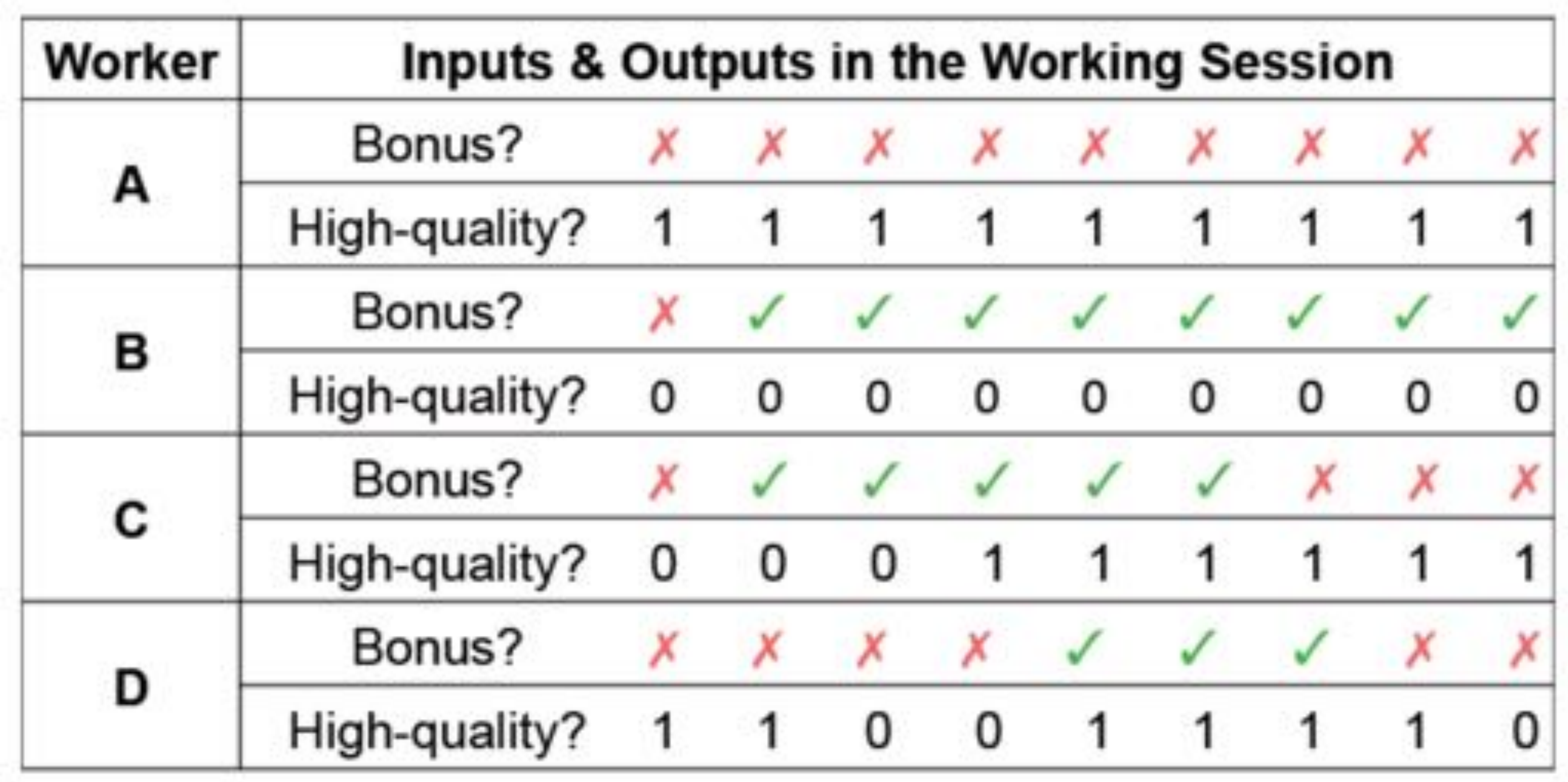
Incent or not

Low quality

High quality

.......

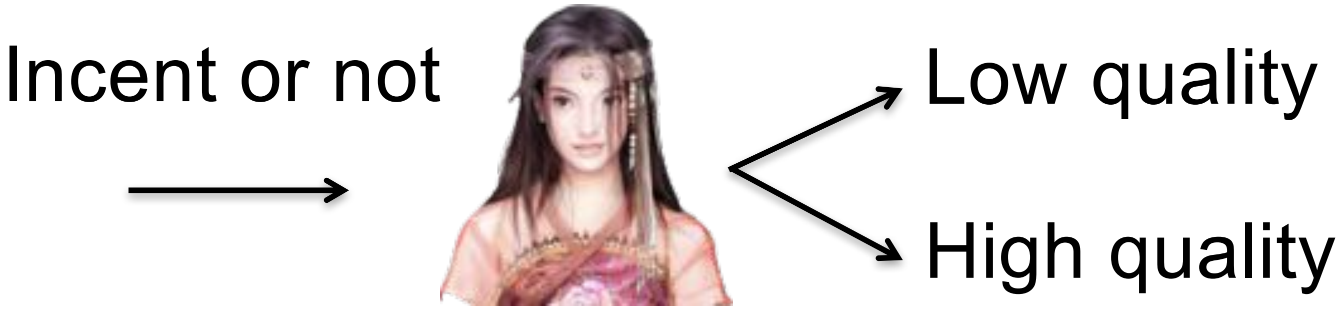
# An Incentive-based Model



| Worker | Inputs & Outputs in the Working Session | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | Bonus? | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | High-quality? | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| B | Bonus? | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | High-quality? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | Bonus? | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| | High-quality? | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| D | Bonus? | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ |
| | High-quality? | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |

Model with a Input-output Hidden Markov Model

- Inputs: $a_t \in \{0,1\}$, $t = 1, 2, \cdots, T$, with 0 representing bonus is not placed on the task.
- Outputs: $x_t \in \{0,1\}$, t = 1, 2, $\cdots$ ,$T$, with 0 representing an incorrect (or low-quality) answer for the task.
- Hidden States: $z_t \in \{1,2,\cdots,K\}$
- Transition probability: $P(z_t|z_{t-1}, a_t)$
- Emission probability: $P_e(x_t|z_t, a_t)$



Incent or not

Low quality

High quality

# Take-Away Messages

- **Crowdsourcing can be utilized well on machine learning tasks**
  - E.g., Provide labeled data in  deep learning, semi-supervised learning and transfer learning.
- **Key challenges in crowd-powered machine learning tasks**
  - Human may make mistakes
  - We need huge amount of labeled data, which is costly.
- **Solutions**
  - Quality control methods.
  - Utilize unlabeled data and other data sources.

# Outline

- Crowdsourcing Overview (20min)
- Fundamental Techniques (90min)
  - Quality Control (40min)
  - Cost Control (30min)
  - Latency Control (20min)

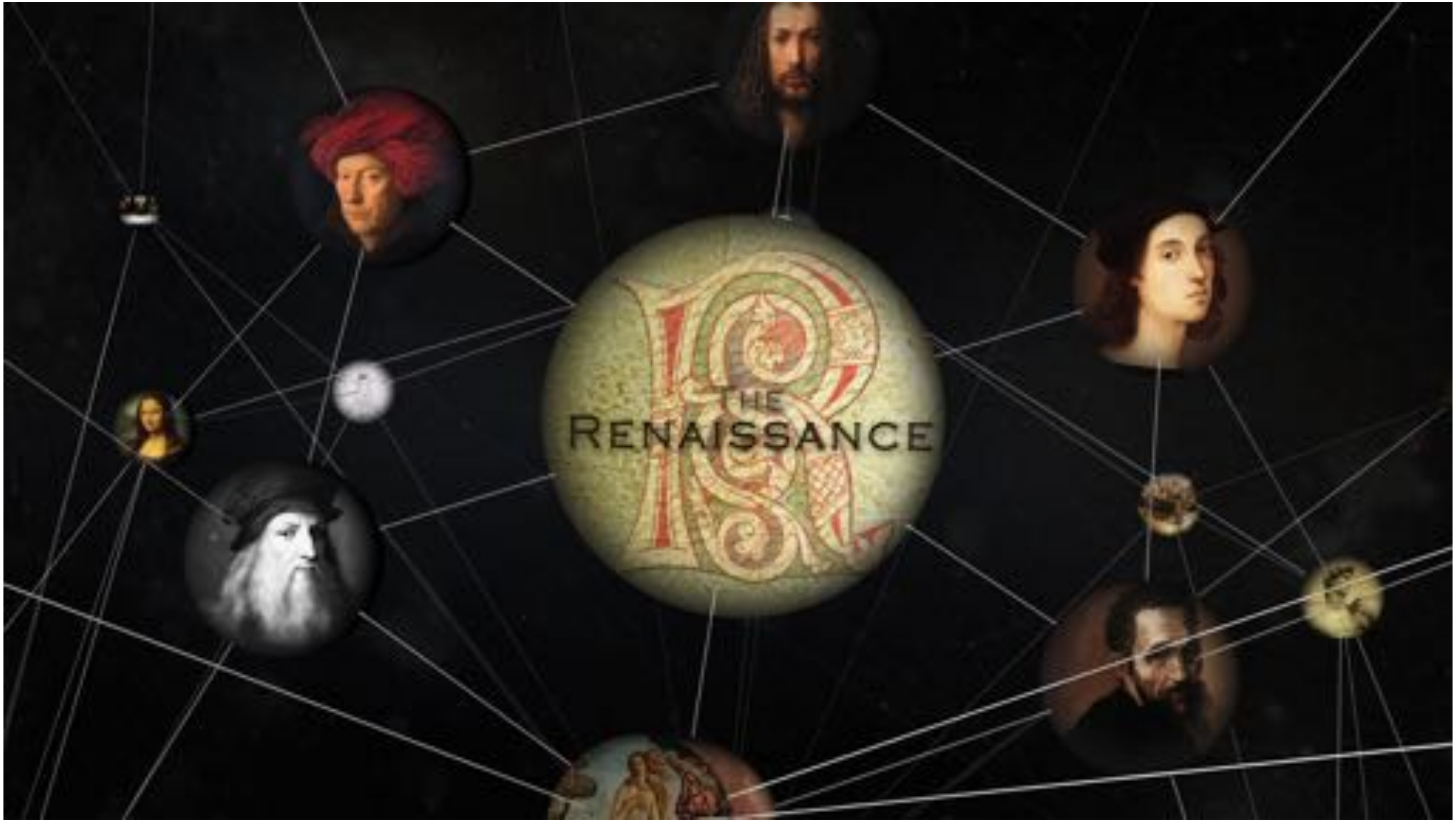
Part 1

- Crowd-powered Data Mining (60min)
  - Crowd-powered Pattern Mining (10min)
  - Crowd-powered Classification (10min)
  - Crowd-powered Clustering (10min)
  - Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - ☞ Crowd-powered Knowledge Discovery (20min)

Part 2

- Challenges (10min)

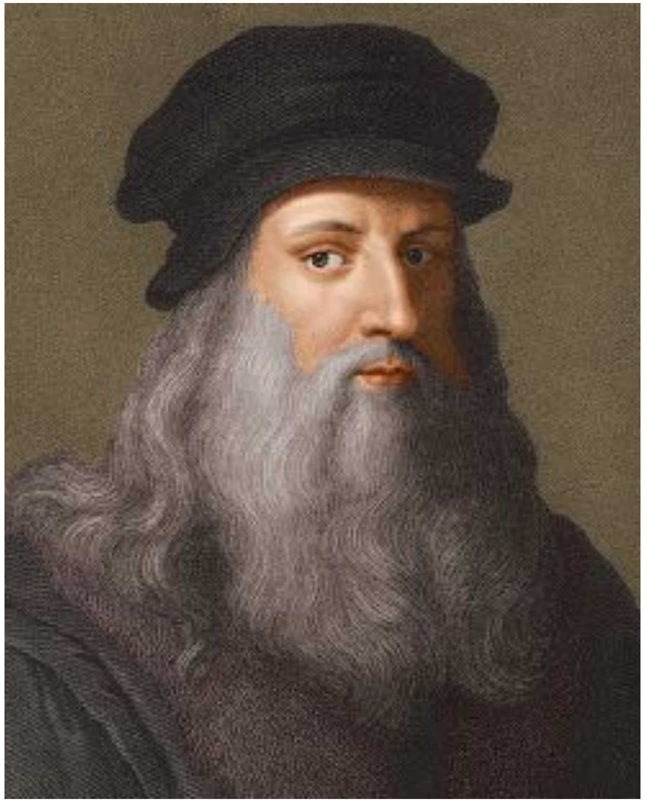
# Knowledge Base (KB)

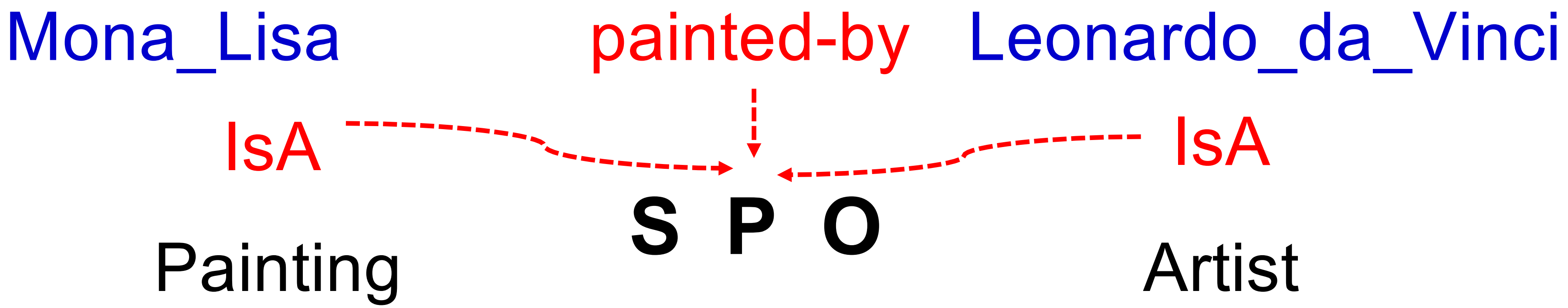


A semantically-organized and machine-readable collection of entities, classes, and SPO facts (attributes, relations)

# Subject-Predicate-Object Facts

painted-by





Mona_Lisa painted-by Leonardo_da_Vinci

IsA

IsA

S P O

Painting

Artist

# Opportunity and Challenge

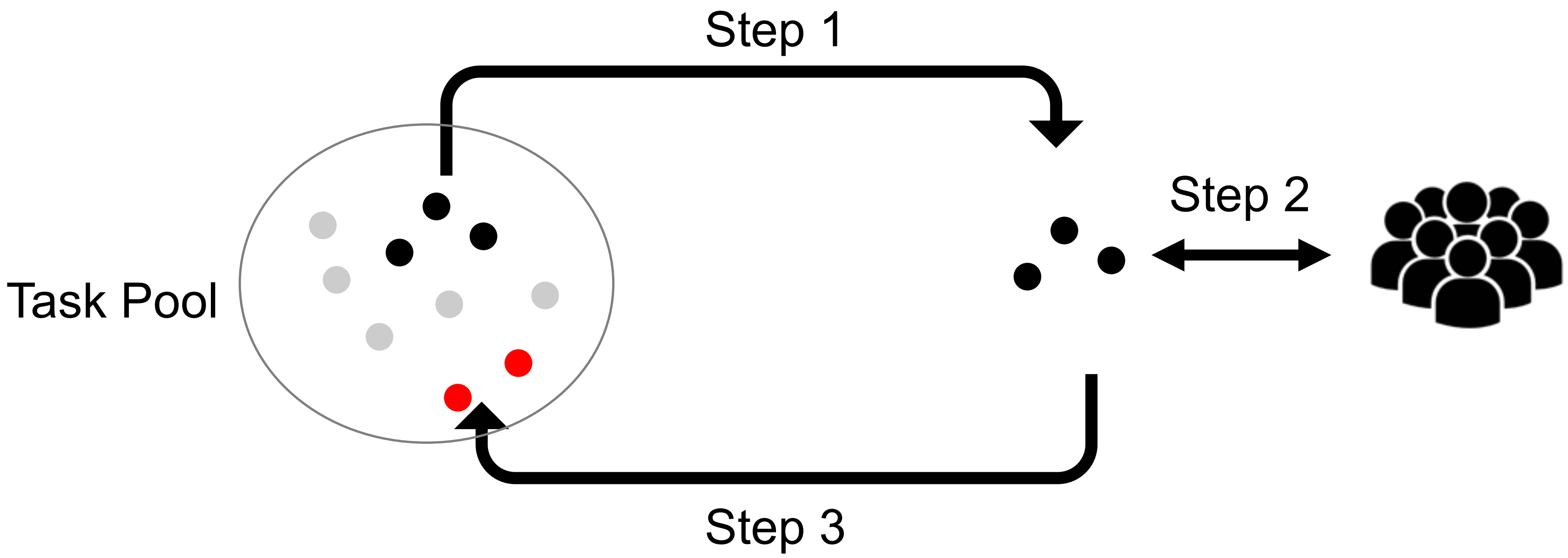
- **Humans are much better than machine on many KB-related tasks**
  - **Extracting SPO facts from a sentence**
  - **Aligning entities across two different KBs**
  - **Enriching KB by matching external sources**

- **However, It is not affordable to do exhaustive crowdsourcing for large-scale KBs**

**$$$ !!!**

# General Idea

- **Machine-Crowd Hybrid Approach**
  - **Before Crowdsourcing:** assigning the most "beneficial" tasks to the crowd
  - **After Crowdsourcing:** utilizing the crowdsourcing result to help infer the rest of tasks



# Crowd-Powered Knowledge Discovery

- Overview
  - Crowd-Powered Knowledge Acquisition
    - Extracting missing attributes of entities or relations among entities using crowd
  - Crowd-Powered Entity Alignment
    - Aligning entities across KBs using crowd
  - Crowd-Powered KB Enrichment
    - Matching web tables to KB using crowd
  - Crowd-Powered Entity Collection
    - Collecting missing entities in KB using crowd

# Knowledge Acquisition (KA)

- **Extracting SPO Facts from raw text**

The Mona Lisa is a half-length portrait painting by the Italian Renaissance artist Leonardo da Vinci…

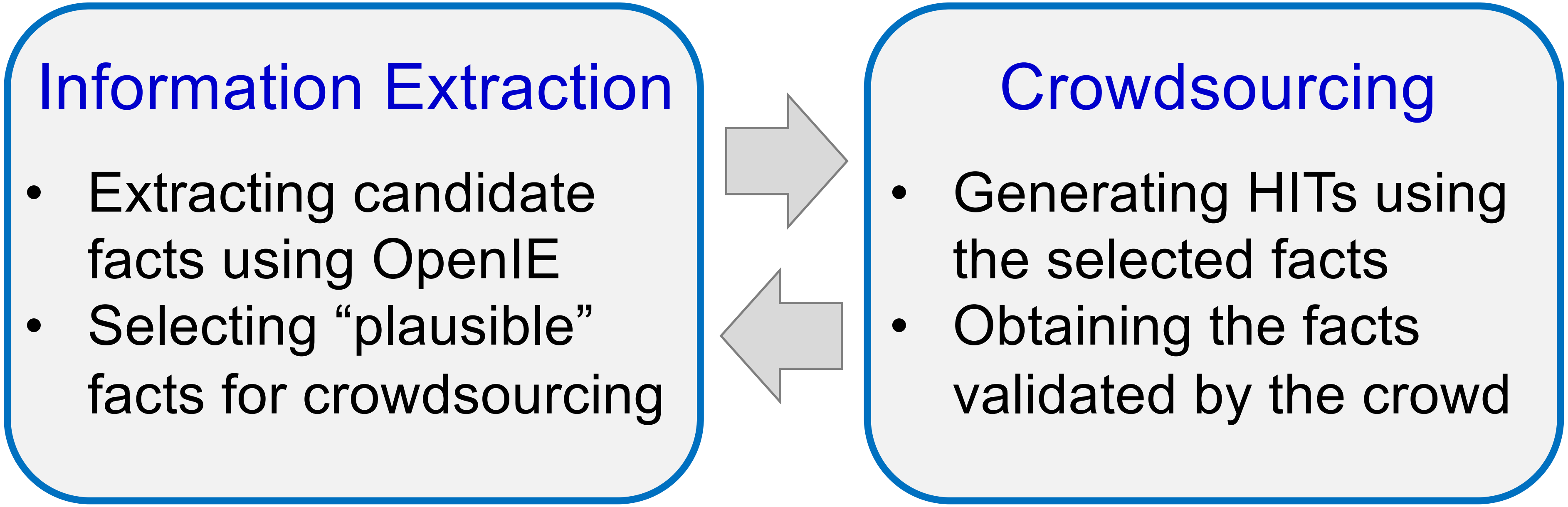
↓

Mona_Lisa Author Leonardo_da_Vinci

- **Existing approach: Information Extraction**
  - E.g., OpenIE using NLP techniques
  - Limitations: noisy or duplicated SPO facts, such as "(Mona Lisa, by, Leonardo da Vinci)", "(Mona Lisa, drew-by, Leonardo da Vinci)", etc.

# The HIGGINS Approach

- **Employing Crowdsourcing for KA comes with opportunities**
  - Human is good at identifying SPO facts
- **However, crowdsourcing alone cannot carry the burden of large-scale KA**



Information Extraction

- Extracting candidate facts using OpenIE
- Selecting “plausible” facts for crowdsourcing

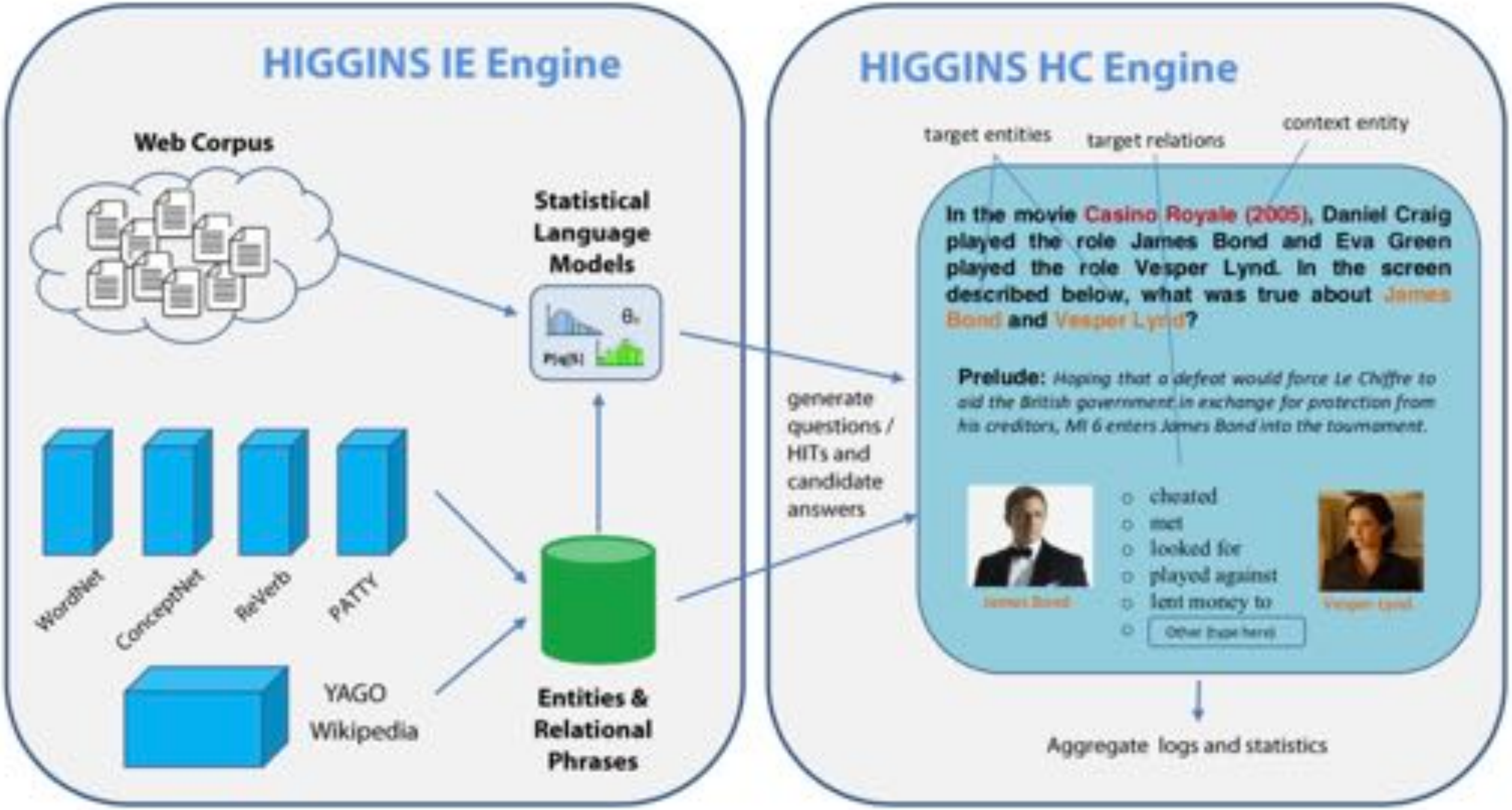
Crowdsourcing

- Generating HITs using the selected facts
- Obtaining the facts validated by the crowd

S. K. Kondreddi, P. Triantafillou, G. Weikum: Combining information extraction and human computing for crowdsourced knowledge acquisition. ICDE 2014

# The HIGGINS Approach

- **Architecture**
  - IE Engine + HC (Crowdsourcing) Engine



In the movie Casino Royale (2005), Daniel Craig played the role James Bond and Eva Green played the role Vesper Lynd. In the screen described below, what was true about James Bond and Vesper Lynd?

**Prelude:** *Hoping that a defeat would force Le Chiffre to aid the British government in exchange for protection from his creditors, MI 6 enters James Bond into the tournament.*

- ○ cheated
- ○ met
- ○ looked for
- ○ played against
- ○ lent money to
- ○ Other (type here)

# The HIGGINS Approach

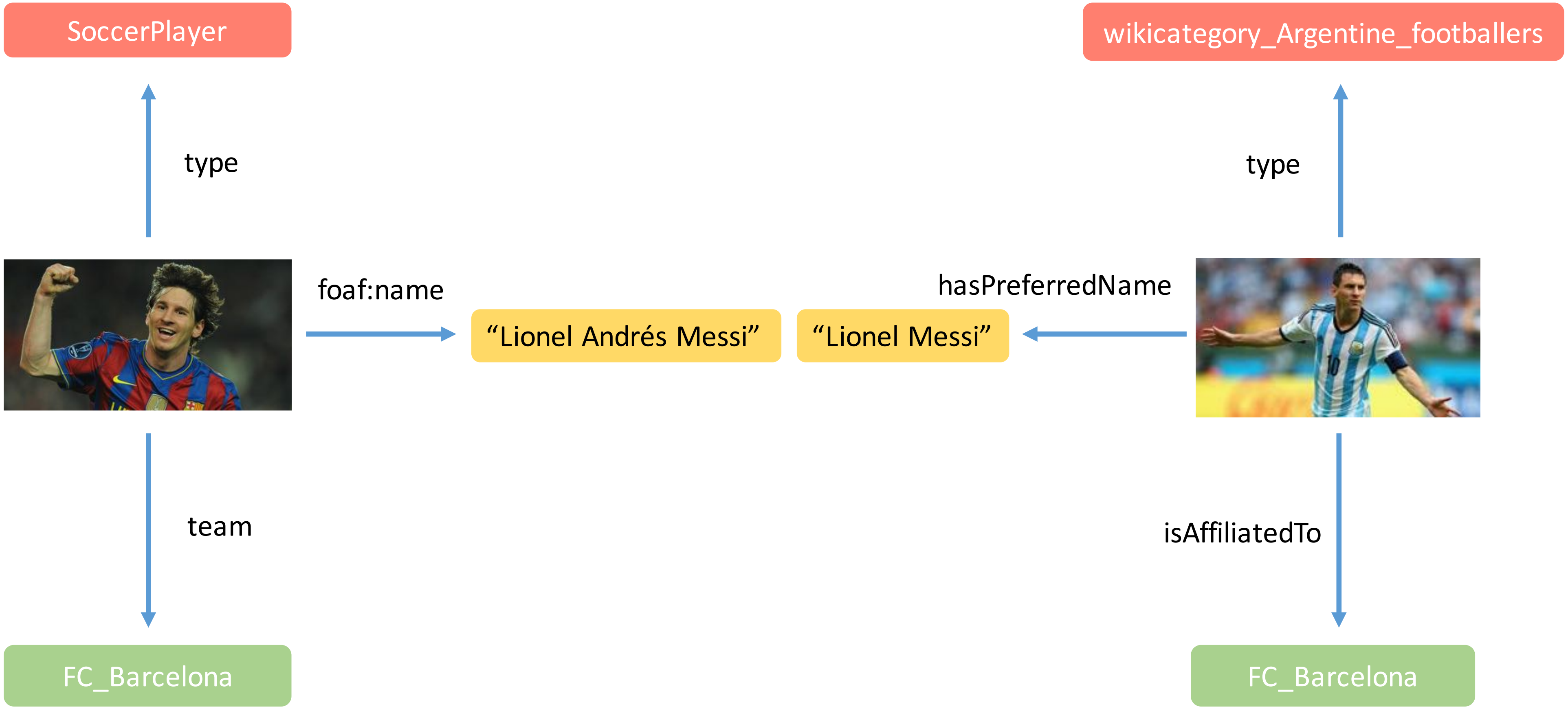
- **HIGGINS IE Engine**
  - **Identifying** entity occurrence, e.g., noun phrases
  - **Detecting** relational phrases that contains two entities using lexicon-syntactic patterns like verbal phrases
  - **Pruning** unpromising candidates using dependency
- **HIGGINS Crowdsourcing Engine**
  - **Question Generation**: providing context information to the crowd, e.g., popular movies/books she knows
  - **Candidate Answer Generation:** suggesting a small number (e.g., 5) of candidate answers by considering criteria like phrase relatedness & diversification
  - **HIT Design:** pre-defined question templates plugged with judiciously selected context cues

# Crowd-Powered Knowledge Discovery

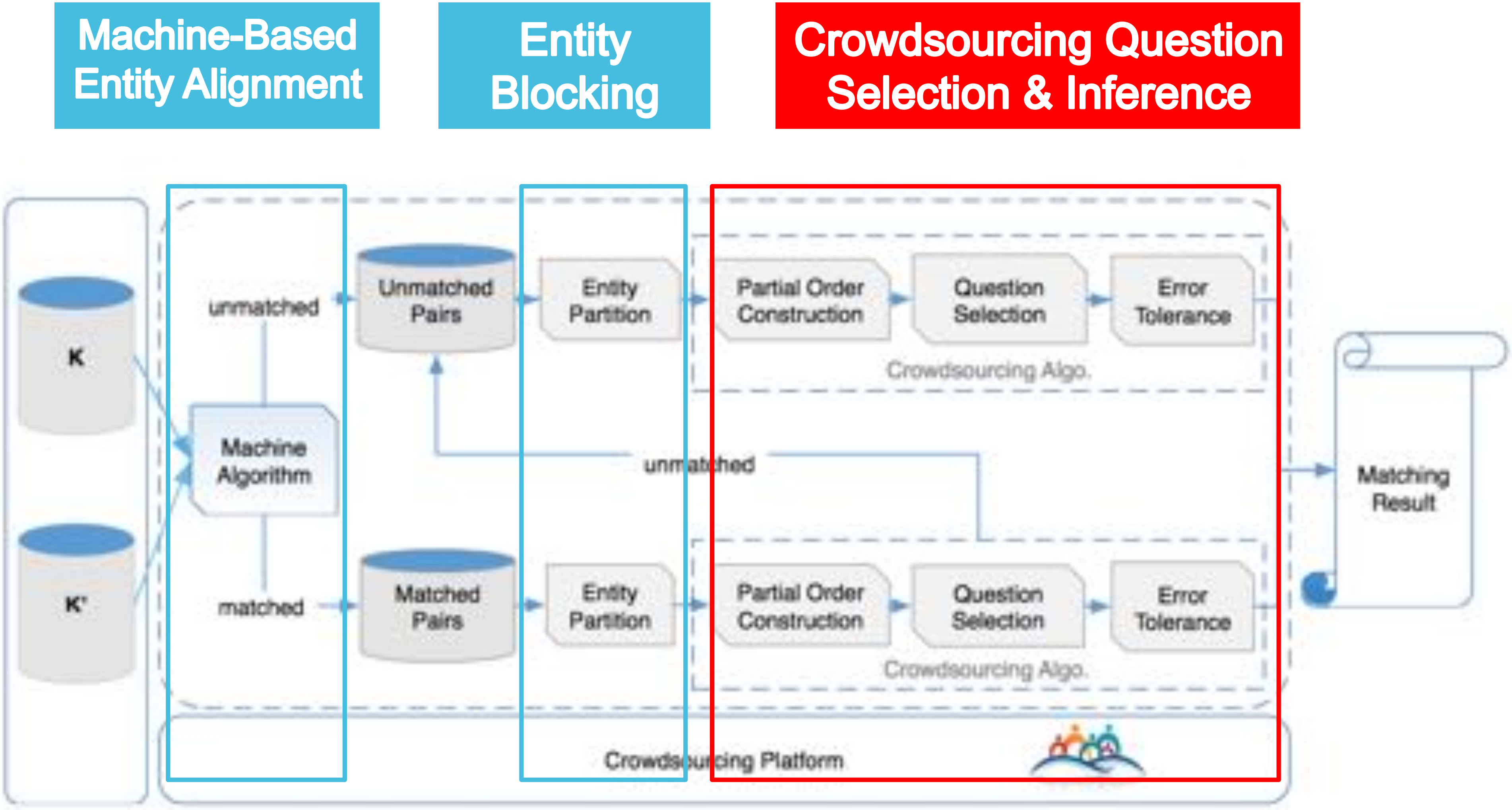
- Overview
  - Crowd-Powered Knowledge Acquisition
    - Extracting missing attributes of entities or relations among entities using crowd
  - Crowd-Powered Entity Alignment
    - Aligning entities across KBs using crowd
  - Crowd-Powered KB Enrichment
    - Matching web tables to KB using crowd
  - Crowd-Powered Entity Collection
    - Collecting missing entities in KB using crowd

# Entity Alignment

- Given two KBs, the entity alignment problem is to find the pairs of entities across the KBs that refer to the same real-world entity.

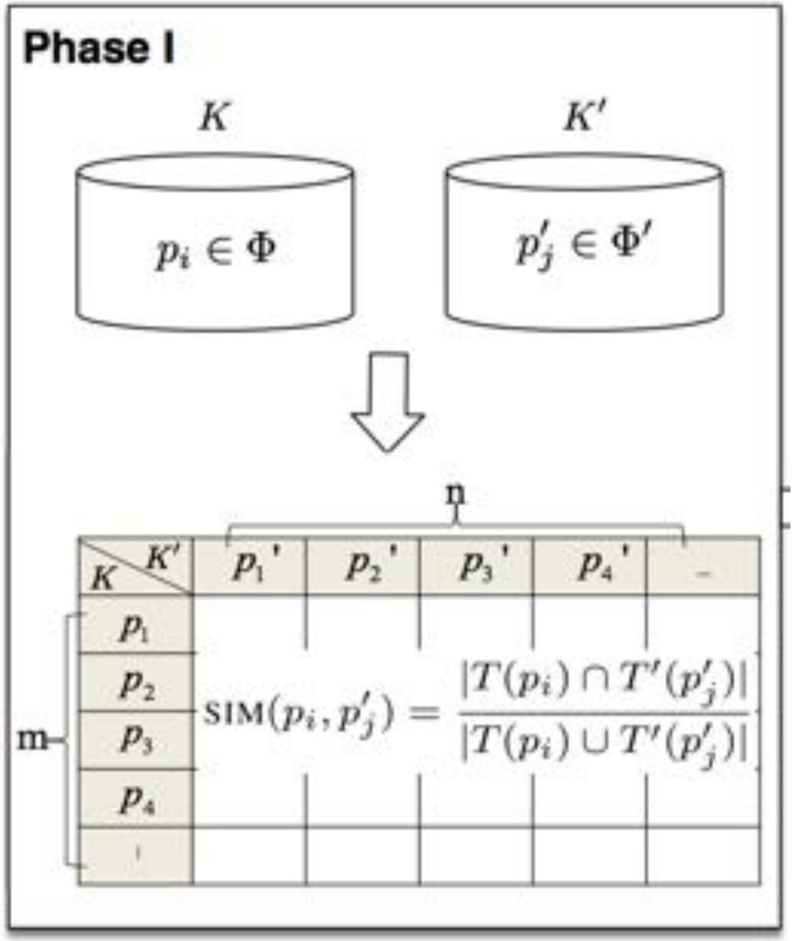


# The HIKE Approach



Y. Zhuang, G. Li, Z. Zhong, J. Feng: Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases. CIKM 2017.

# Predicate-Based Blocking



Considering two KBs K and K′, Hike computes the similarity $SIM(p_i, p'_j)$ between any predicate $p_i$ from K and any $p'_j$ from K′

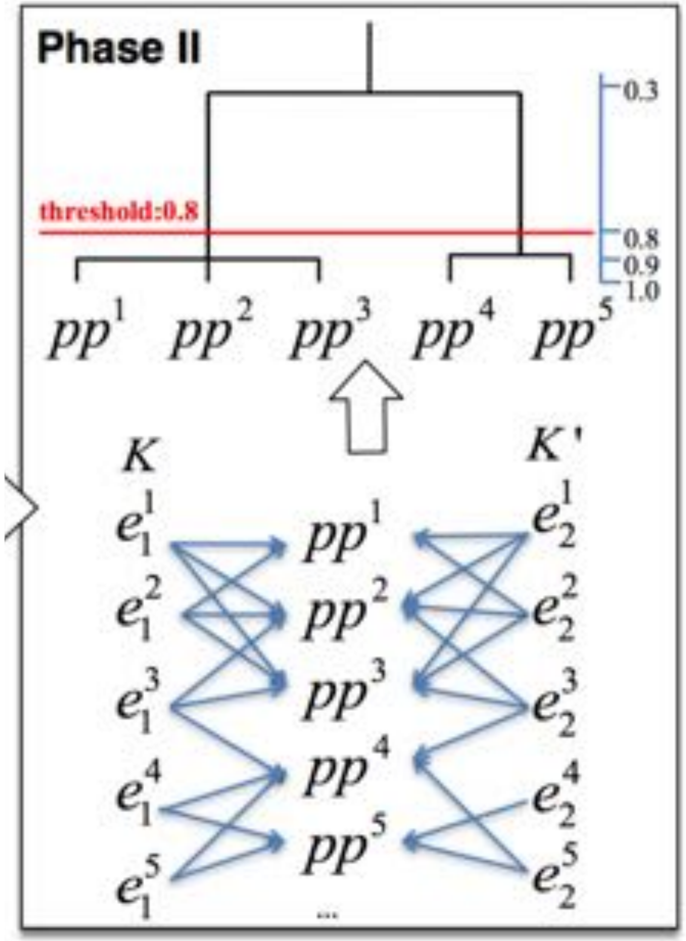
The similarity is based on the overlap between the triple sets corresponding to the predicates

$$\mathrm{SIM}(p_i, p'_j) = \frac{|T(p_i) \cap T'(p'_j)|}{|T(p_i) \cup T'(p'_j)|}$$

Then, how to partition predicates based on the pairwise similarities?

**Phase I:** producing predicate pairs using similarity

# Predicate-Based Blocking


Phase II
threshold:0.8
0.3
0.8
0.9
1.0
$pp^1$ $pp^2$ $pp^3$ $pp^4$ $pp^5$
$K$
$K'$
$e_1^1$ $e_1^2$ $e_1^3$ $e_1^4$ $e_1^5$
$pp^1$ $pp^2$ $pp^3$ $pp^4$ $pp^5$ ...
$e_2^1$ $e_2^2$ $e_2^3$ $e_2^4$ $e_2^5$

Step 1 – Find matching predicates

- For each $p_i \in K$ ($p_i' \in K'$), find its most similar predicate $p_j' \in K'$ ($p_j \in K$).
- Each of such predicate pair is called a matching predicate pair

Step 2 – Compute similarity between matching predicate pairs

$$\rho(pp^i, pp^j) = \frac{\cos(S(pp^i), S(pp^j)) + \cos(S'(pp^i), S'(pp^j))}{2}$$

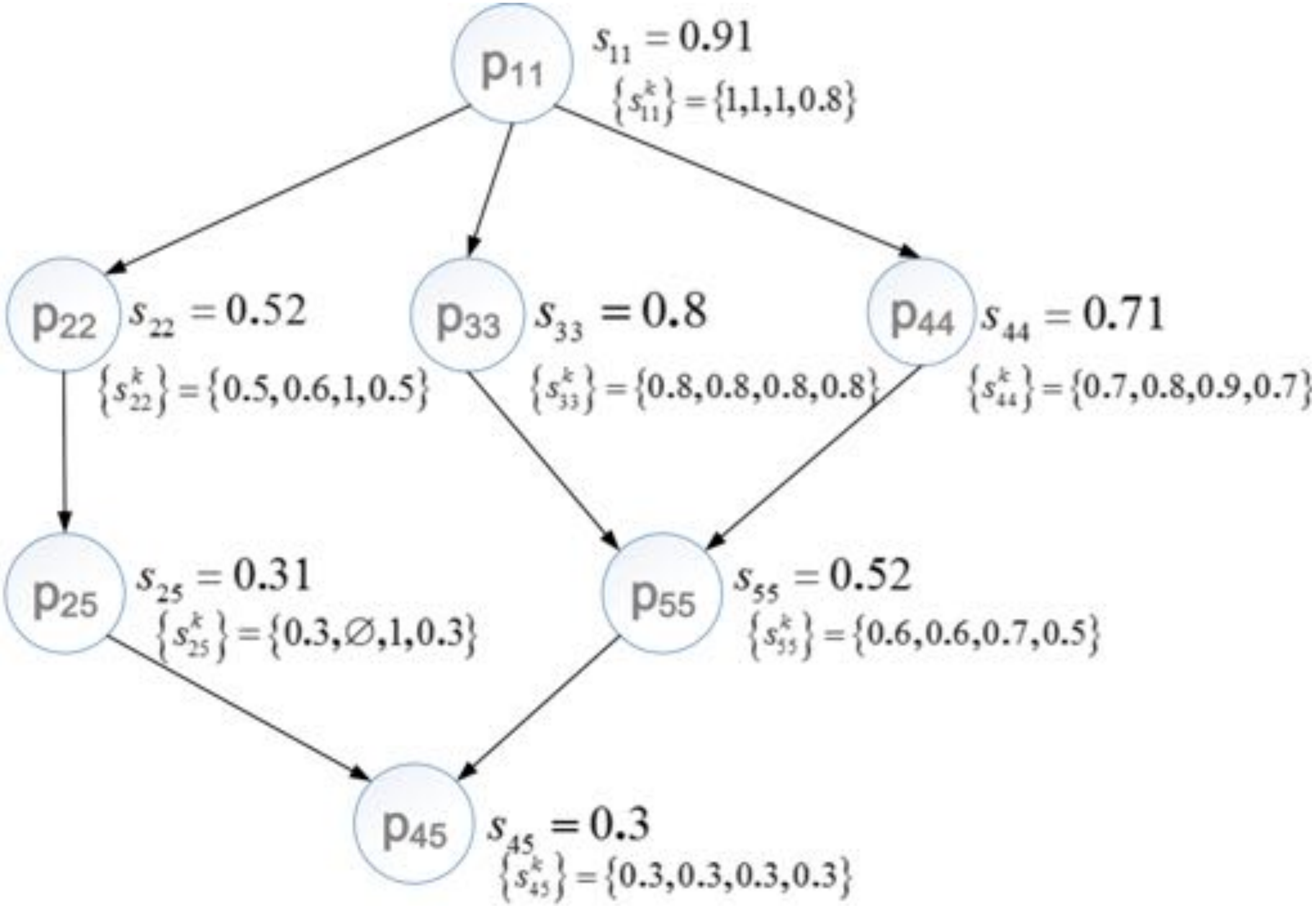
Step 3 – Apply hierarchical agglomerative clustering (HAC) algorithm

**Phase II:** partition KBs by clustering predicate pairs

# Crowd Question Selection

- **Question selection based “partial orders”**

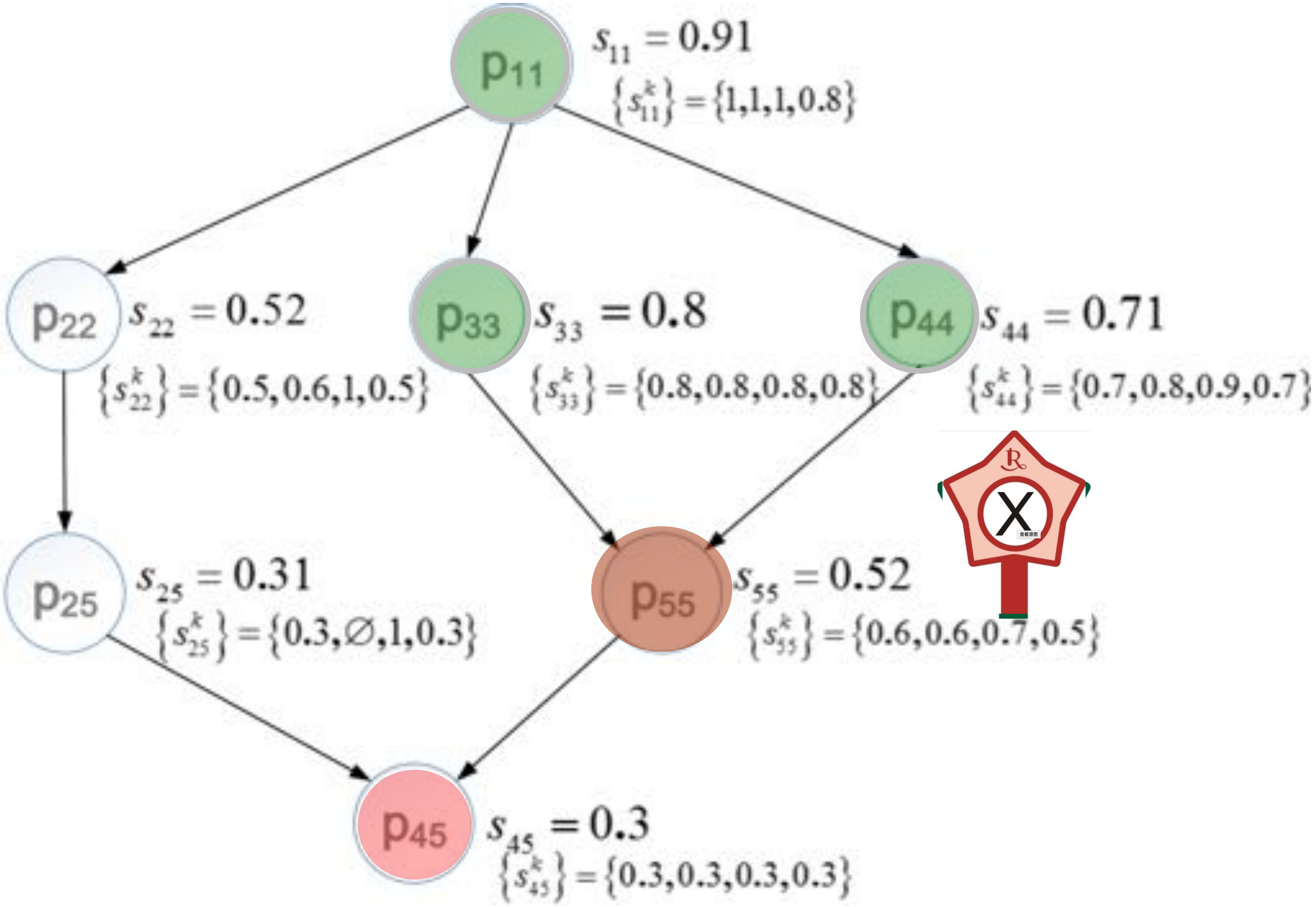
  Suppose we have 5 entities in each KB whose predicate pairs are {⟨name,name⟩,⟨birth_place,born_in⟩,⟨birth_date,dob⟩, ⟨article, article⟩}


p11
$s_{11} = 0.91$
$\{s_{11}^k\} = \{1,1,1,0.8\}$
p22
$s_{22} = 0.52$
$\{s_{22}^k\} = \{0.5,0.6,1,0.5\}$
p33
$s_{33} = 0.8$
$\{s_{33}^k\} = \{0.8,0.8,0.8,0.8\}$
p44
$s_{44} = 0.71$
$\{s_{44}^k\} = \{0.7,0.8,0.9,0.7\}$
p25
$s_{25} = 0.31$
$\{s_{25}^k\} = \{0.3,\varnothing,1,0.3\}$
p55
$s_{55} = 0.52$
$\{s_{55}^k\} = \{0.6,0.6,0.7,0.5\}$
p45
$s_{45} = 0.3$
$\{s_{45}^k\} = \{0.3,0.3,0.3,0.3\}$

# Crowd Question Selection

- **Question selection based “partial orders”**

  Suppose we have 5 entities in each KB whose predicate pairs are {⟨name,name⟩,⟨birth_place,born_in⟩,⟨birth_date,dob⟩, ⟨article, article⟩}

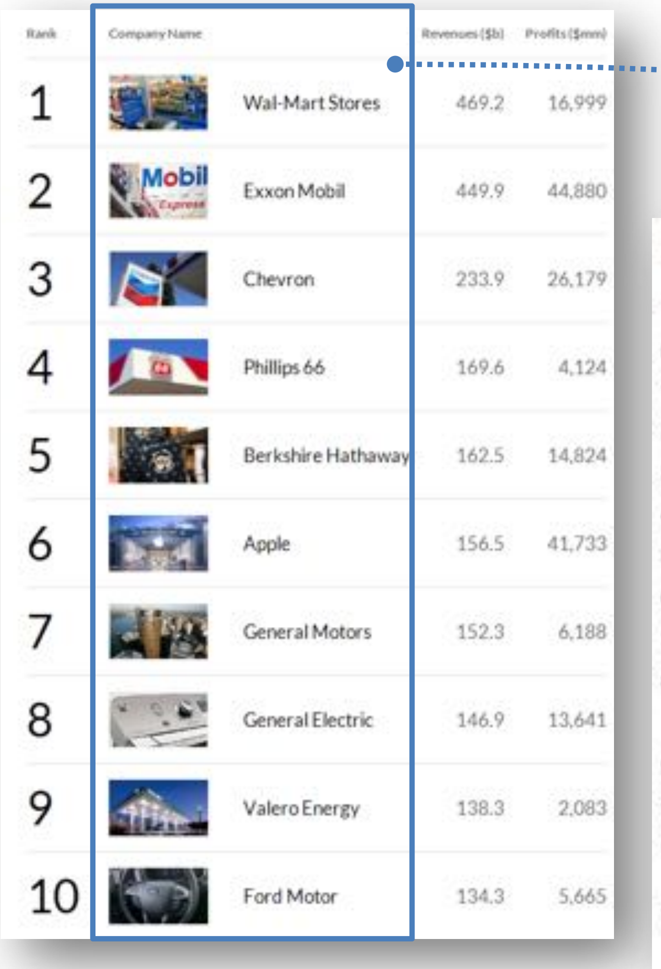

p11
$s_{11} = 0.91$
$\{s_{11}^k\} = \{1,1,1,0.8\}$
p22
$s_{22} = 0.52$
$\{s_{22}^k\} = \{0.5,0.6,1,0.5\}$
p33
$s_{33} = 0.8$
$\{s_{33}^k\} = \{0.8,0.8,0.8,0.8\}$
p44
$s_{44} = 0.71$
$\{s_{44}^k\} = \{0.7,0.8,0.9,0.7\}$
p25
$s_{25} = 0.31$
$\{s_{25}^k\} = \{0.3,\varnothing,1,0.3\}$
p55
$s_{55} = 0.52$
$\{s_{55}^k\} = \{0.6,0.6,0.7,0.5\}$
X
p45
$s_{45} = 0.3$
$\{s_{45}^k\} = \{0.3,0.3,0.3,0.3\}$

A sketch map of partial order set

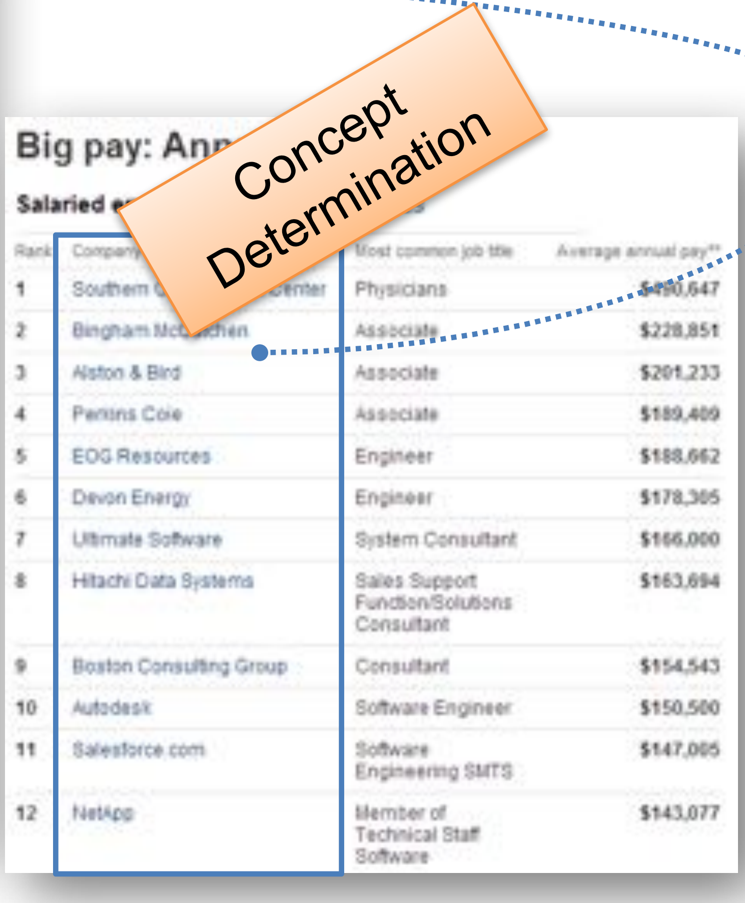
# Crowd-Powered Knowledge Discovery

- Overview
  - Crowd-Powered Knowledge Acquisition
    - Extracting missing attributes of entities or relations among entities using crowd
  - Crowd-Powered Entity Alignment
    - Aligning entities across KBs using crowd
  - Crowd-Powered KB Enrichment
    - Matching web tables to KB using crowd
  - Crowd-Powered Entity Collection
    - Collecting missing entities in KB using crowd

# Enriching KB using Web Tables



| Rank | Company Name | | Revenues ($b) | Profits ($mm) |
|---|---|---|---|---|
| 1 | | Wal-Mart Stores | 469.2 | 16,999 |
| 2 | | Exxon Mobil | 449.9 | 44,880 |
| 3 | | Chevron | 233.9 | 26,179 |
| 4 | | Phillips 66 | 169.6 | 4,124 |
| 5 | | Berkshire Hathaway | 162.5 | 14,824 |
| 6 | | Apple | 156.5 | 41,733 |
| 7 | | General Motors | 152.3 | 6,188 |
| 8 | | General Electric | 146.9 | 13,641 |
| 9 | | Valero Energy | 138.3 | 2,083 |
| 10 | | Ford Motor | 134.3 | 5,665 |

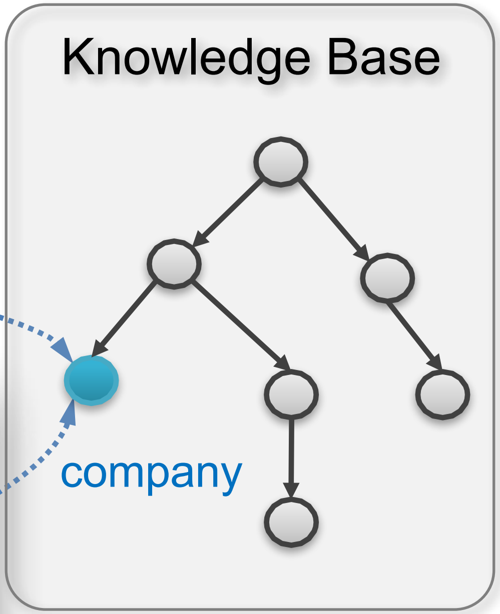


Big pay: Ann

Salaried e

| Rank | Company | Most common job title | Average annual pay** |
|---|---|---|---|
| 1 | Southern C Center | Physicians | $490,647 |
| 2 | Bingham Mc chen | Associate | $228,851 |
| 3 | Alston & Bird | Associate | $201,233 |
| 4 | Perkins Coie | Associate | $189,409 |
| 5 | EOG Resources | Engineer | $188,662 |
| 6 | Devon Energy | Engineer | $178,305 |
| 7 | Ultimate Software | System Consultant | $166,000 |
| 8 | Hitachi Data Systems | Sales Support Function/Solutions Consultant | $163,694 |
| 9 | Boston Consulting Group | Consultant | $154,543 |
| 10 | Autodesk | Software Engineer | $150,500 |
| 11 | Salesforce.com | Software Engineering SMTS | $147,005 |
| 12 | NetApp | Member of Technical Staff Software | $143,077 |

Concept Determination

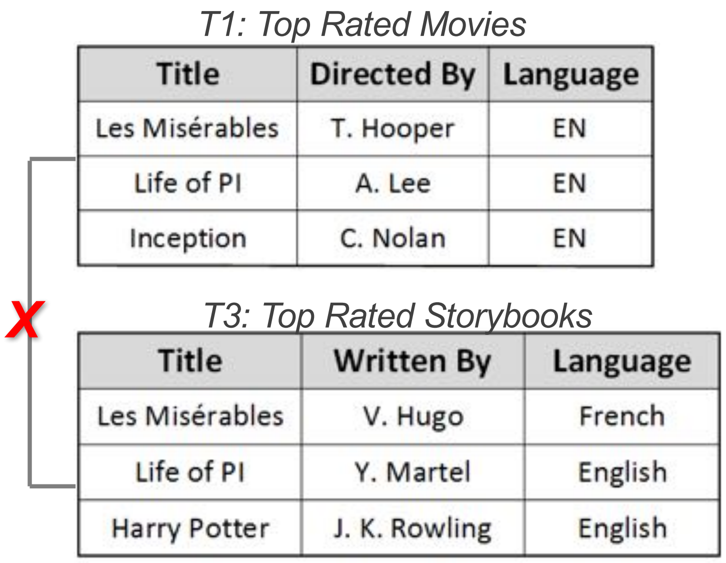


# Prior Work on Concept Determination

- **Table annotation techniques**
  - **Annotate web table columns with concepts in KB**
  - **Pure machine-based algorithm**
  - **Limitation:**
    - **Not suitable for some inherently difficult columns**

Accuracy on 1,166 randomly selected columns

| Approach | Accuracy |
|---|---|
| G.Limaye et al. VLDB'10 | 58.7% |
| P. Venetis et al. VLDB'11 | 52.1% |



*T1: Top Rated Movies*

| Title | Directed By | Language |
|---|---|---|
| Les Misérables | T. Hooper | EN |
| Life of PI | A. Lee | EN |
| Inception | C. Nolan | EN |

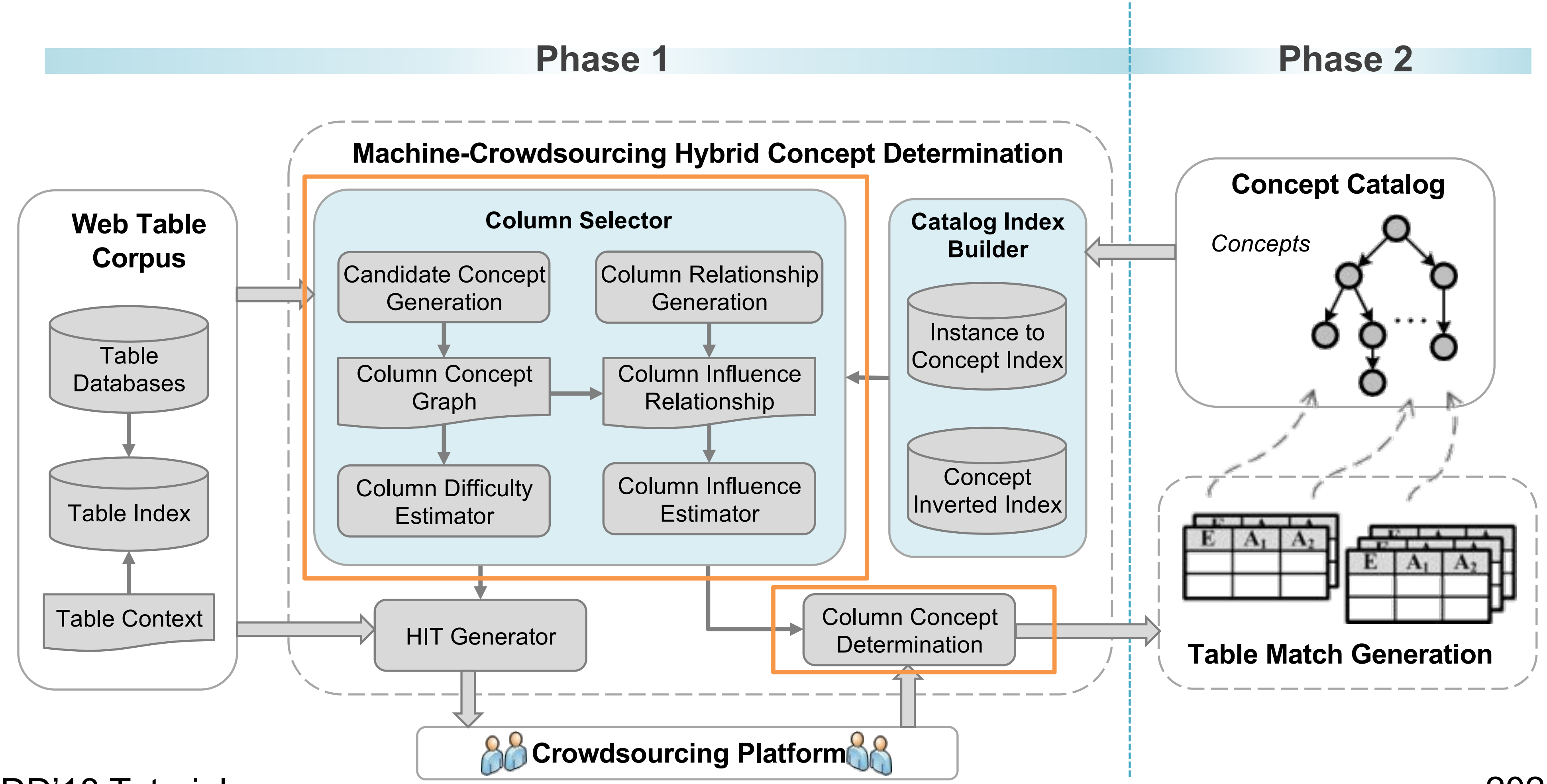
*T3: Top Rated Storybooks*

| Title | Written By | Language |
|---|---|---|
| Les Misérables | V. Hugo | French |
| Life of PI | Y. Martel | English |
| Harry Potter | J. K. Rowling | English |

G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. PVLDB, 2010.
P. Venetis, A. Y. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. PVLDB, 2011.

# The CROWDWT Approach

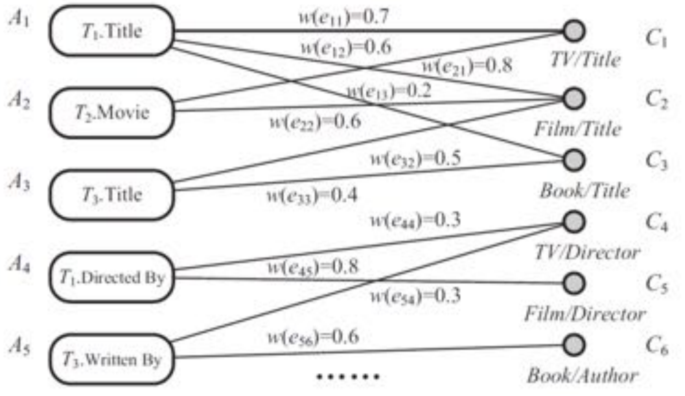
- Machine: Generate candidate matched concepts for each column
- Crowd: Verify the candidate matches



Phase 1
Phase 2
Machine-Crowdsourcing Hybrid Concept Determination
Web Table Corpus
Table Databases
Table Index
Table Context
Column Selector
Candidate Concept Generation
Column Relationship Generation
Column Concept Graph
Column Influence Relationship
Column Difficulty Estimator
Column Influence Estimator
Catalog Index Builder
Instance to Concept Index
Concept Inverted Index
Concept Catalog
Concepts
HIT Generator
Column Concept Determination
Table Match Generation
E
A1
A2
Crowdsourcing Platform

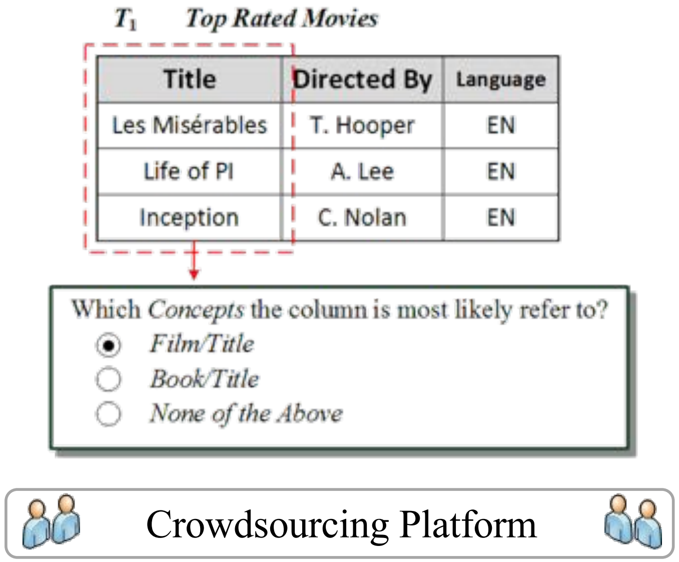
# Machine-Crowdsourcing Hybrid Framework

- ○ **Machine:**
  - – **Generate candidate matched concepts for each column**
- ○ **Crowd:**
  - – **Verify the candidate matches**



Machine

**Candidates**



$T_1$ *Top Rated Movies*

| Title | Directed By | Language |
|---|---|---|
| Les Misérables | T. Hooper | EN |
| Life of PI | A. Lee | EN |
| Inception | C. Nolan | EN |

Which *Concepts* the column is most likely refer to?

● *Film/Title*
○ *Book/Title*
○ *None of the Above*

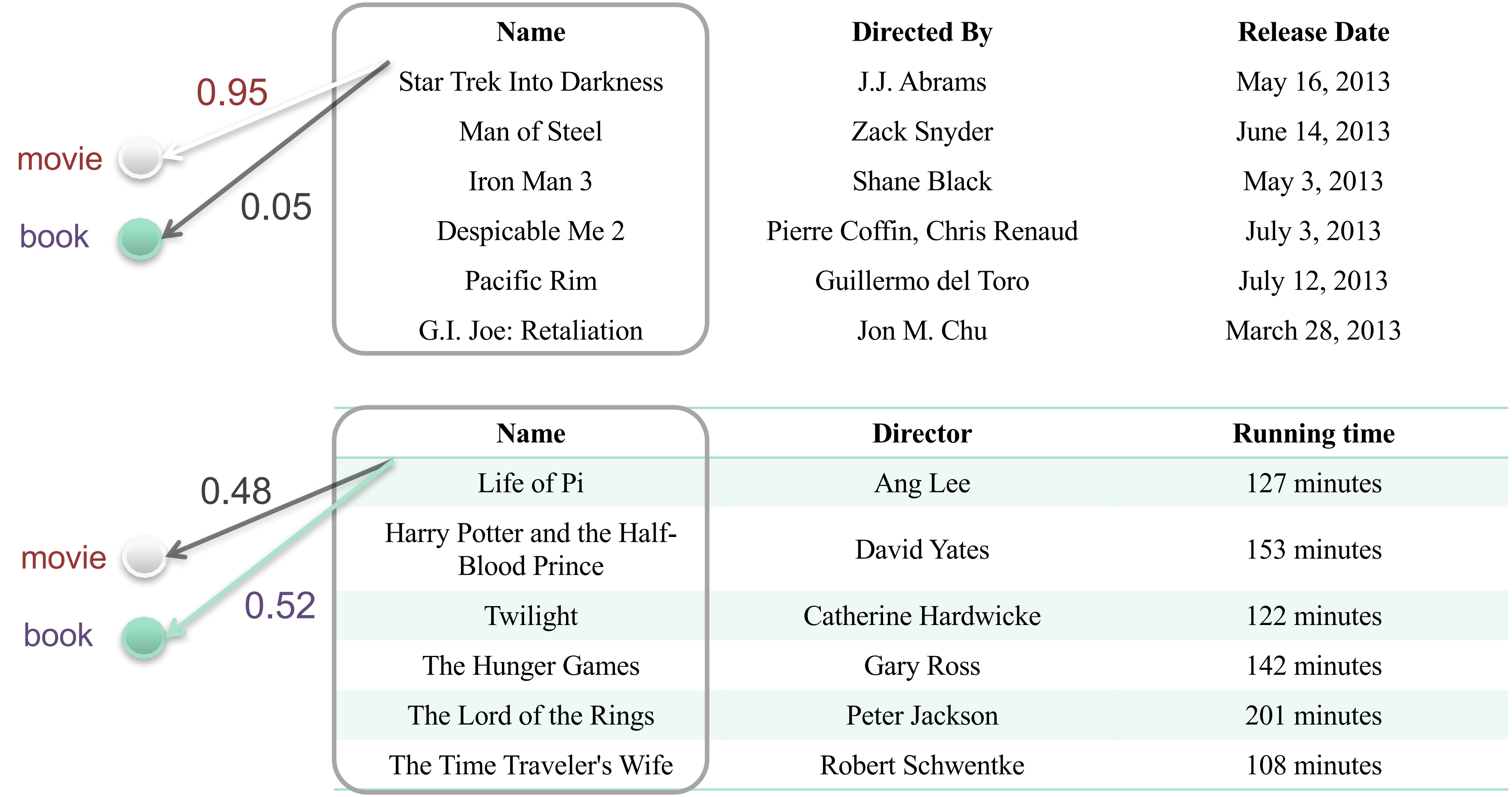
Crowdsourcing Platform

# Crowdsourcing Column Selection

- Selecting the most “beneficial” columns
  - Factor 1: Column difficulty
    - Columns that are difficult for machines
  - Factor 2: Column influence
    - Columns, if verified, would have greater influence on inferring the concepts of other columns

# Crowdsourcing Column Selection

## Column Difficulty



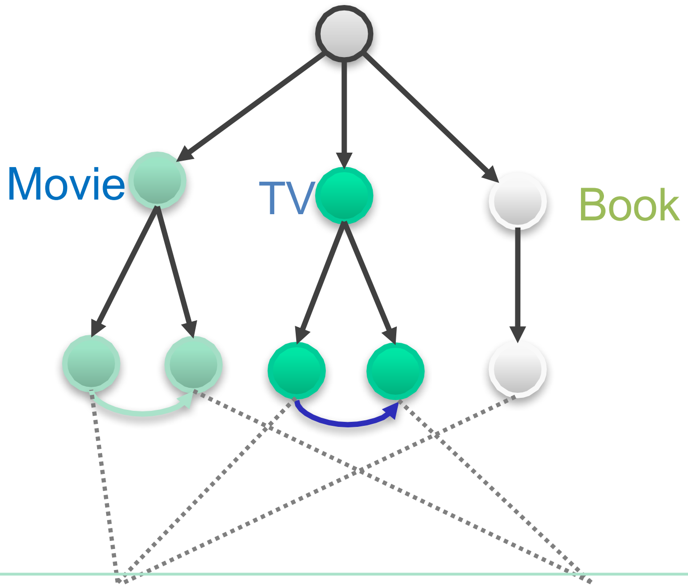
| Name | Directed By | Release Date |
|---|---|---|
| Star Trek Into Darkness | J.J. Abrams | May 16, 2013 |
| Man of Steel | Zack Snyder | June 14, 2013 |
| Iron Man 3 | Shane Black | May 3, 2013 |
| Despicable Me 2 | Pierre Coffin, Chris Renaud | July 3, 2013 |
| Pacific Rim | Guillermo del Toro | July 12, 2013 |
| G.I. Joe: Retaliation | Jon M. Chu | March 28, 2013 |

| Name | Director | Running time |
|---|---|---|
| Life of Pi | Ang Lee | 127 minutes |
| Harry Potter and the Half-Blood Prince | David Yates | 153 minutes |
| Twilight | Catherine Hardwicke | 122 minutes |
| The Hunger Games | Gary Ross | 142 minutes |
| The Lord of the Rings | Peter Jackson | 201 minutes |
| The Time Traveler's Wife | Robert Schwentke | 108 minutes |

# Crowdsourcing Column Selection

## Column Influence

*Intra-table influence*

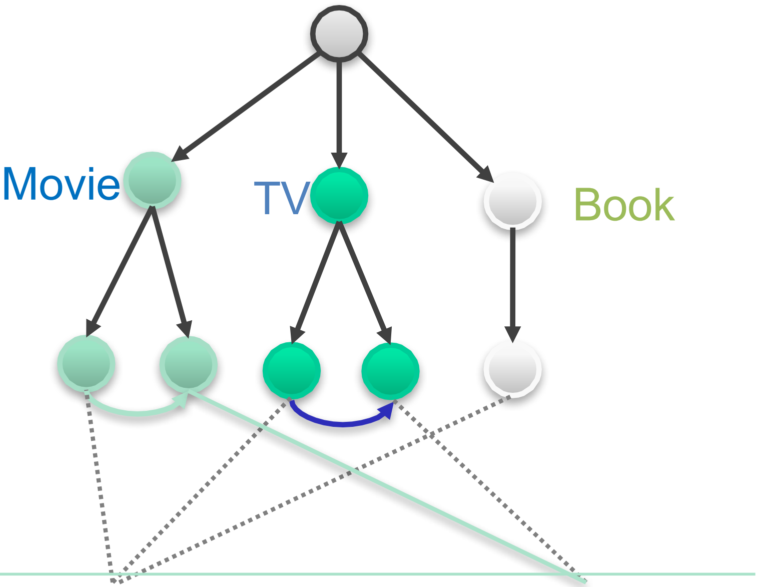


| Name | Director |
|---|---|
| Life of Pi | Ang Lee |
| Harry Potter and the Half-Blood Prince | David Yates |
| Twilight | Catherine Hardwicke |
| The Hunger Games | Gary Ross |
| The Lord of the Rings | Peter Jackson |
| The Time Traveler's Wife | Robert Schwentke |

# Crowdsourcing Column Selection

## Column Influence

*Intra-table influence*

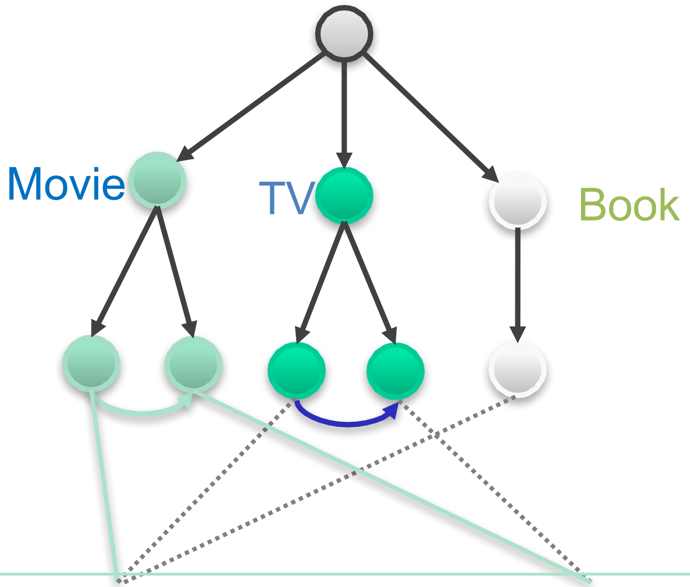


| Name | Director |
|---|---|
| Life of Pi | Ang Lee |
| Harry Potter and the Half-Blood Prince | David Yates |
| Twilight | Catherine Hardwicke |
| The Hunger Games | Gary Ross |
| The Lord of the Rings | Peter Jackson |
| The Time Traveler's Wife | Robert Schwentke |

# Crowdsourcing Column Selection

## Column Influence

*Intra-table influence*

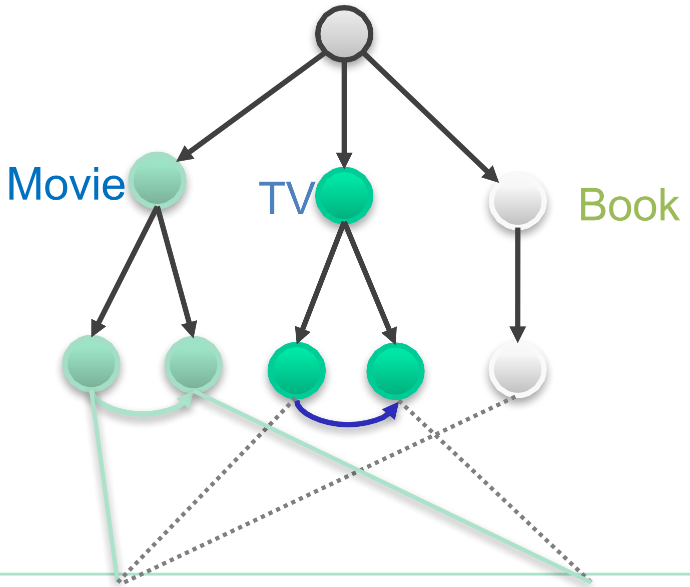


| Name | Director |
|---|---|
| Life of Pi | Ang Lee |
| Harry Potter and the Half-Blood Prince | David Yates |
| Twilight | Catherine Hardwicke |
| The Hunger Games | Gary Ross |
| The Lord of the Rings | Peter Jackson |
| The Time Traveler's Wife | Robert Schwentke |

# Crowdsourcing Column Selection

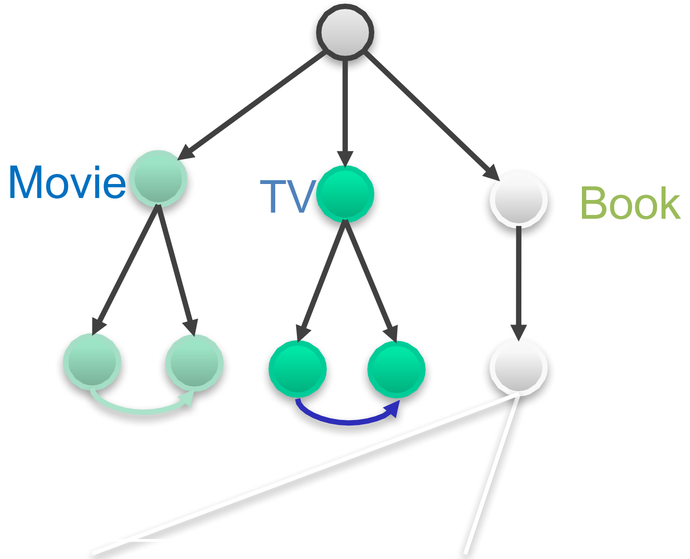
## Column Influence

*Intra-table influence*



| Name | Director |
|---|---|
| Life of Pi | Ang Lee |
| Harry Potter and the Half-Blood Prince | David Yates |
| Twilight | Catherine Hardwicke |
| The Hunger Games | Gary Ross |
| The Lord of the Rings | Peter Jackson |
| The Time Traveler's Wife | Robert Schwentke |

*Inter-table influence*



| Name | Title |
|---|---|
| Life of Pi | Clockwork Princess |
| Harry Potter and the Half-Blood Prince | Time Traveler's Wife |
| Boneshaker | Harry Potter |
| The Hunger Games | Boneshaker |
| Clockwork Princess | The Hunger Games |
| The Time Traveler's Wife | Life of Pi |

# Crowd-Powered Knowledge Discovery

- **Overview**
  - **Crowd-Powered Knowledge Acquisition**
    - **Extracting missing attributes of entities or relations among entities using crowd**
  - **Crowd-Powered Entity Alignment**
    - **Aligning entities across KBs using crowd**
  - **Crowd-Powered KB Enrichment**
    - **Matching web tables to KB using crowd**
  - **Crowd-Powered Entity Collection**
    - **Collecting missing entities in KB using crowd**

# Crowdsourced Entity Collection

We want to get all names of ACTIVE NBA players. You will be requested to give us the DIFFERENT names.

**NO.1 Name**

**NO.2 Name**

**NO.3 Name**

- Applications
  - Knowledge Base Construction
  - Enterprise Data Collection
  - Cardinality Estimation

# Challenges

We want to get all names of ACTIVE NBA players. You will be requested to give us the DIFFERENT names.

*R*={Steven Curry, Kevin Durant, Michael Jordan, Russell Westbrook, Steven Curry}

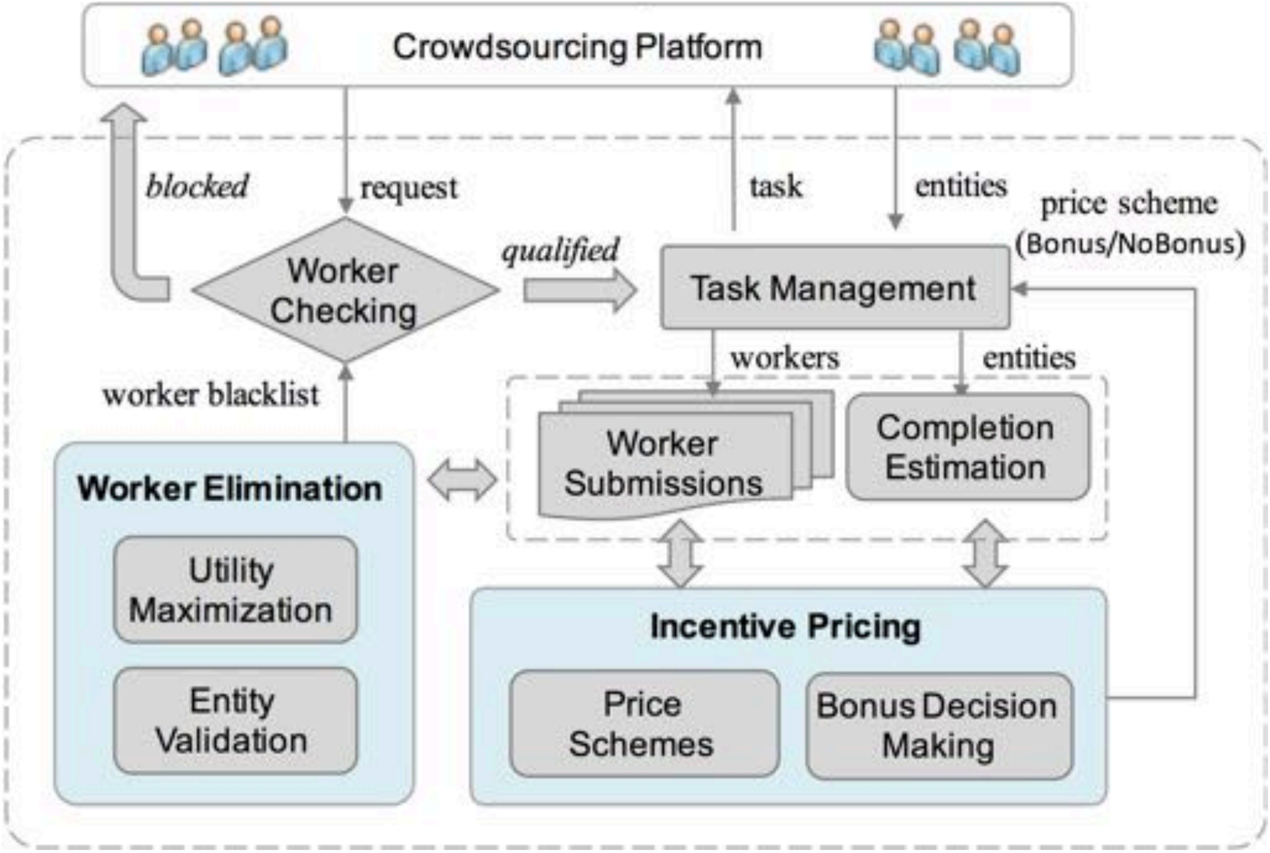
*O*={Steven Curry, Kevin Durant, Michael Jordan, Russell Westbrook, …}

Precision=3/4

Recall=3/450 Unknown !!!

- **Objectives**
  - **Correct**
  - **Complete**
  - **Less-Duplicate**

# The CrowdEC Approach



- Worker Elimination

  Eliminate low quality workers.

  Avoid many duplicated answers.

- Incentive Pricing

  Encourage workers to provide distinct answers

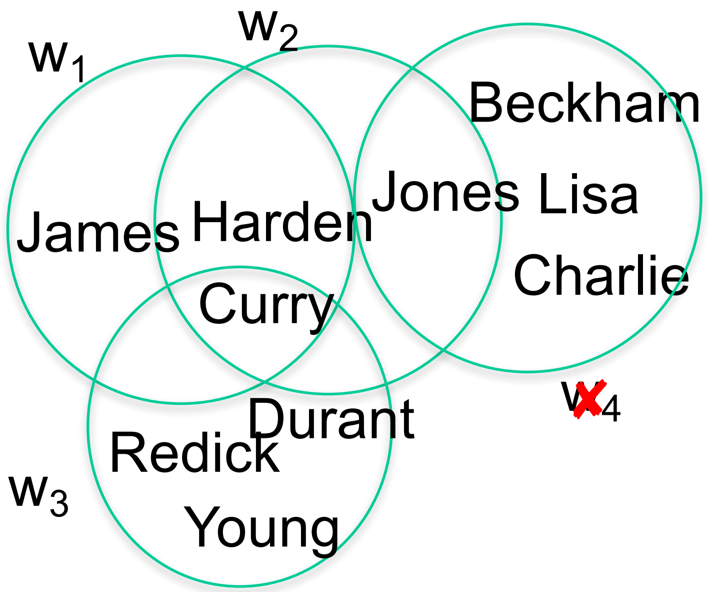
Chengliang Chai, Ju Fan, Guoliang Li: Incentive-based Entity Collection using Crowdsourcing. ICDE 2018
Ju Fan, Zhewei Wei, Dongxiang Zhang, Jingru Yang, and Xiaoyong Du: Distribution-Aware Crowdsourced Entity Collection. TKDE 2017

# Worker Elimination

- Worker Quality
- Worker Distinctness

*Answers set by worker j*

$$(|\cup \mathcal{R}_j|)/(\sum |\mathcal{R}_j|)$$



Given v1=3, v2=1 and v3=6,

$D_{\{w1,w2,w3\}}(7x(3+1+6))/(3+3+4)=7$

$D_{\{w1,w3\}}=(6x(3+6))/(3+4)=7.7$

$$D_{\mathcal{W}} = \frac{\left|\bigcup_{w_j \in \mathcal{W}} \mathcal{R}_j\right| \boxed{\sum_{w_j \in \mathcal{W}} v_j}}{\sum_{w_j \in \mathcal{W}} |\mathcal{R}_j|}$$

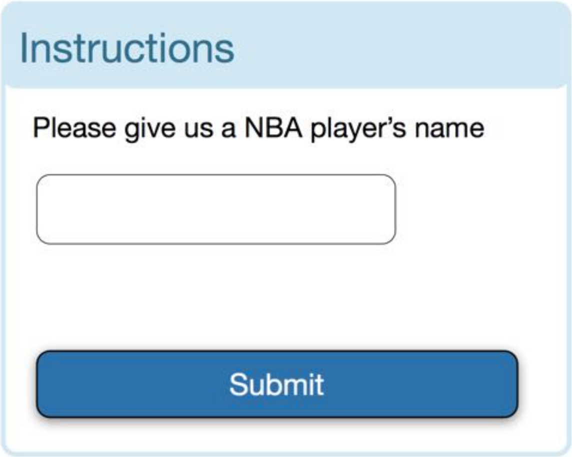
*throughput by worker j*

# Incentive Pricing

- Pricing Schema
- Optimization

NoBonus Schema:

Collect one entity at a time, with a basic reward

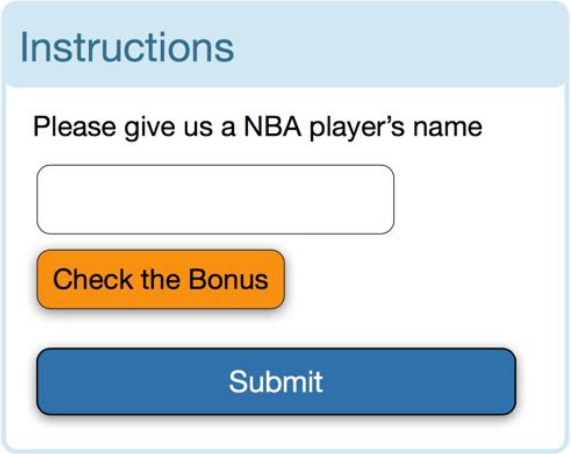


Instructions

Please give us a NBA player's name

Bonus Schema:

Collect multiples entities at a time.  We reward *the bonus.* if there is  a distinct answer, otherwise we reward the same as NoBonus Schema .



Instructions

Please give us a NBA player's name

# Incentive Pricing

## Pricing Schema ( Example)

Given a task with a bonus schema, a worker gives answer {James, Curry, Durrant}.

Given $C_r$=\$1 and $C_b$=\$0.5, Bonus Schema costs: \$1.5 ; NoBonus Schema costs:\$3

## How to choose between them ? (Intuitive ideas)

- At the beginning, Nobonus schema is better.
- With the #entities accumulating, encouragement should begin.
- When it almost completes, encouragement seems useless
- For workers who are positive to Bonus schema, we can give more incentive tasks

# Take-Away Messages

- **Crowdsourcing can perform well on many knowledge discovery tasks**
  - E.g., knowledge extraction, alignment, enrichment and entity collection
- **Key challenge of crowdsourced knowledge discovery is crowd cost control.**
  - Not affordable to do exhaustive crowdsourcing for large-scale KBs
- **Solutions**
  - Task selection & Answer reduction
  - Incentive mechanism for pricing

# Reference – Crowd-powered Data Mining

[1] Yael Amsterdamer, Susan B. Davidson, Anna Kukliansky, Tova Milo, Slava Novgorodov, Amit Somech: Managing General and Individual Knowledge in Crowd Mining Applications. CIDR 2015

[2] Yael Amsterdamer, Anna Kukliansky, Tova Milo: NL2CM: A Natural Language Interface to Crowd Mining. SIGMOD Conference 2015: 1433-1438

[3] Yael Amsterdamer, Susan B. Davidson, Tova Milo, Slava Novgorodov, Amit Somech: Ontology Assisted Crowd Mining. PVLDB 7(13): 1597-1600 (2014)

[4] Yael Amsterdamer, Susan B. Davidson, Tova Milo, Slava Novgorodov, Amit Somech: OASSIS: query driven crowd mining. SIGMOD Conference 2014: 589-600

[5] Yael Amsterdamer, Yael Grossman, Tova Milo, Pierre Senellart: Crowd mining. SIGMOD Conference 2013: 241-252

[6] Lei Chen, Dongwon Lee, Tova Milo: Data-driven crowdsourcing: Management, mining, and applications. ICDE 2015: 1527-1529

[7] Vikas C. Raykar, Jeremy Magruder . Learning from the Crowd. JMLR 2010 Volume 122, Issue 563, Pages 957-989

[8] Aditya Parameswaran et. al Human-Assisted Graph Search: It’s Okay to Ask Questions VLDBJ 2011, Volume 4 Issue 5, Pages 267-278

[9] Barzan Mozafari , Purna Sarker, Michael Franklin, Michael Jordan, Samuel Madden Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning VLDB 2014. Volume 8 Issue 2.

# Reference – Crowd-powered Data Mining

[10] Hannes Heikinheimo Antti Ukkonen The Crowd-Median Algorithm HCOMP 2013

[11] Hannes Ryan Gomes , Peter Welinder, Andreas Krause, Pietro Perona Crowdclustering NIPS 2011 Pages 558-566

[12] S. K. Kondreddi, P. Triantafillou, G. Weikum: Combining information extraction and human computing for crowdsourced knowledge acquisition. ICDE 2014

[13] Y. Zhuang, G. Li, Z. Zhong, J. Feng: Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases. CIKM 2017.

[14] G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and searching web tables using entities, types and relationships. PVLDB, 2010.

[15] P. Venetis, A. Y. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. PVLDB, 2011.

[16] Chengliang Chai, Ju Fan, Guoliang Li: Incentive-based Entity Collection using Crowdsourcing. ICDE 2018

[17] Ju Fan, Zhewei Wei, Dongxiang Zhang, Jingru Yang, and Xiaoyong Du: Distribution-Aware Crowdsourced Entity Collection. TKDE 2017

[18] Filipe Rodriguesl, Francisco Pereira Deep Learning from Crowds. AAAI 2018

[19]Yaosheng Yang, Meishan Zhang, Wenliang Chen, Wei Zhang Haofen Wang, Min Zhang. Adversarial Learning for Chinese NER from Crowd Annotations AAAI 2018

[20]Zhou Zhao, Da Yan, Wilfred Ng, Shi Gao. A Transfer Learning based Framework of Crowd-Selection on Twitter. KDD'13, Pages 1514-1517

[21] Kyohei Atarashi, Satoshi Oyama, Masahito Kurihara Semi-supervised Learning from Crowds Using Deep Generative Models AAAI'18

# Outline

- Crowdsourcing Overview (20min)
- Fundamental Techniques (90min)
  - Quality Control (40min)
  - Cost Control (30min)
  - Latency Control (20min)

Part 1

- Crowd-powered Data Mining (60min)
  - Crowd-powered Pattern Mining (10min)
  - Crowd-powered Classification (10min)
  - Crowd-powered Clustering (10min)
  - Crowd-powered Machine Learning (10min)
    - Deep learning
    - Transfer learning
    - Semi-supervised learning
  - Crowd-powered Knowledge Discovery (20min)

Part 2

☞ Challenges (10min)

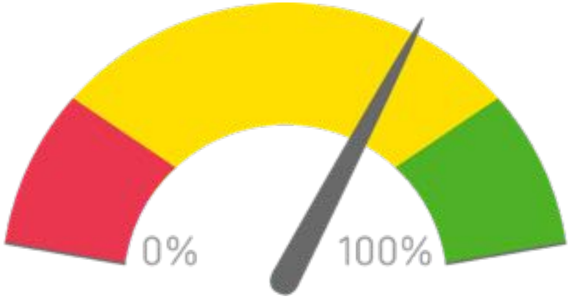
# The Crowdsourcing Challenges

- Benchmarking
- Large-Scale Data Annotation
- Outlier Detection
- Truth Inference
- Incentive Mechanism
- Scalability
- Privacy
- Macro-Tasks

# 1. Benchmarking

- **Database Benchmarks**

  **TPC-C, TPC-H, TPC-DI,…**



- **Crowdsourcing**
  **No standard benchmarks**

- **Existing public datasets (link) are inadequate**

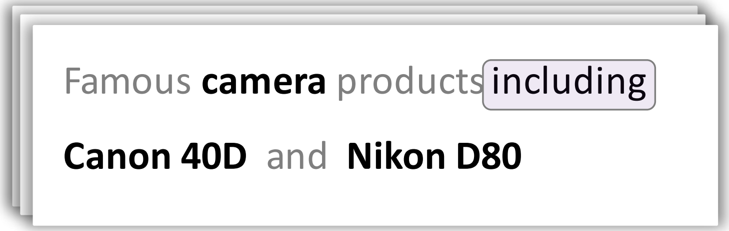
# 1. Benchmarking

- Existing public datasets are inadequate, because:

- Each task often receives 5 or less answers
- Most tasks are single-label tasks
- Very few numeric tasks
- Lack ground truth
  - Expensive to get ground truth for 10K tasks

# 2. Large-Scale Data Annotation

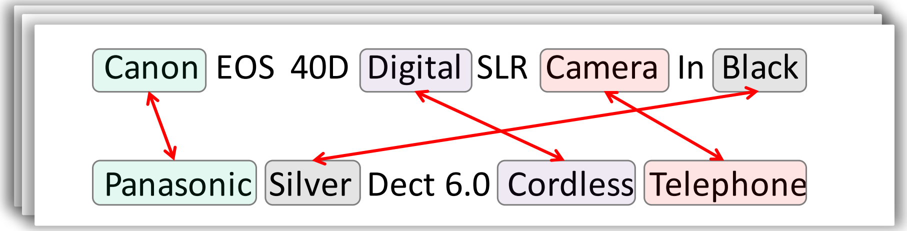
- **It is indispensable to obtain large-scale annotated datasets with high quality for many applications**
  - Creating large training sets for many DM tasks



Famous **camera** products including

**Canon 40D** and **Nikon D80**

**Entity Extraction**



Canon EOS 40D Digital SLR Camera In Black

Panasonic Silver Dect 6.0 Cordless Telephone

**Entity Matching**

- **Utilizing crowdsourcing to annotate tuple-by-tuple**
  - Hard to scale to datasets with tens of thousands to millions of tuples

- **Leverage labeling rules automatically generated**
  - Some rules may be noisy and it is hard to consolidate rules with diverse quality

**Utilizing crowdsourcing for rule generation?**

# 3. Outlier Detection

○ **Machine only outlier detection methods may not work well on many datasets.**

○ **It is hard to select appropriate similarity metrics, features and algorithms.**

○ **Human can help, but it is challenging (1) to design tasks to ask, (2) to guide human to infer the similarity metrics, and (3) combine the results of different approaches.**

# 4. Truth Inference

○ **Not fully solved**
**(Zheng et al. VLDB17)**

○ **We have surveyed 20+ methods:**

**(1) No best method;**

**(2) The oldest method (David & Skene JRSS 1979) is the most robust;**

**(3) No robust method for numeric tasks (the baseline “Mean” performs the best !)**

# 5. Incentive Mechanism

- **Existing crowdsourcing quality control is based on fixed payment**
- **Can we design payment mechanisms to incentivize workers to work better?**

- **Challenging Questions**
  - How to make the smallest possible payment to spammers
  - How to design incentive-compatible mechanism
  - How to support self-correction mechanisms
  - …

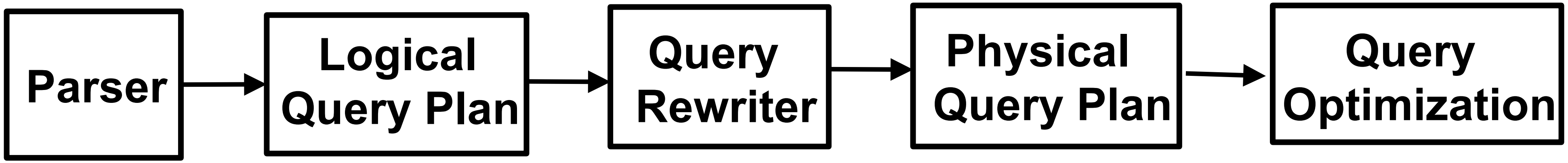
# 6. Scalability

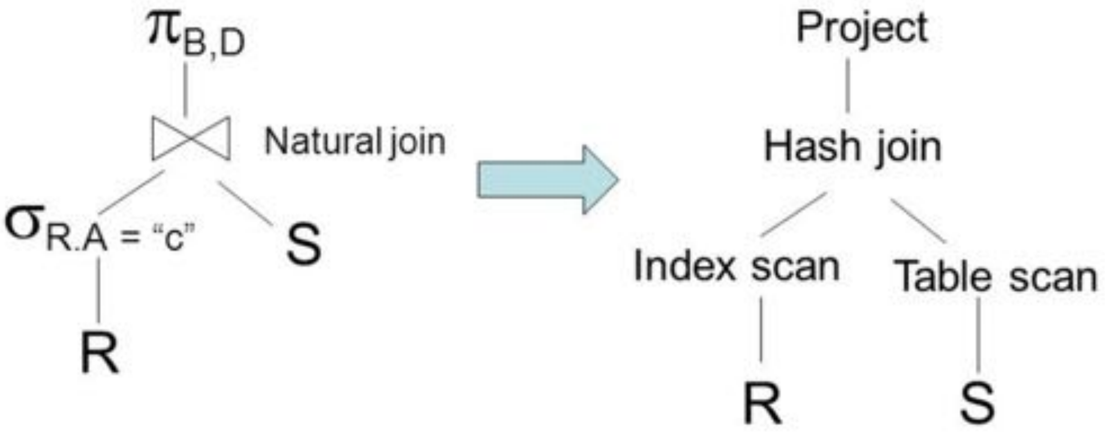
- ○ **Hard to Scale in Crowdsourcing to tackle the 3Vs of Big Data?**

- ○ **(1) workers are expensive;**
  **(2) answers can be erroneous;**
  **(3) existing works focus on specific problems, e.g., active learning (Mozafari et al. VLDB14), entity matching (Gokhale et al. SIGMOD14).**

# 6. Scalability: Query Optimization

○ **Query Processing in Traditional RDBMS**





IBM DB2

PostgreSQL

MySQL

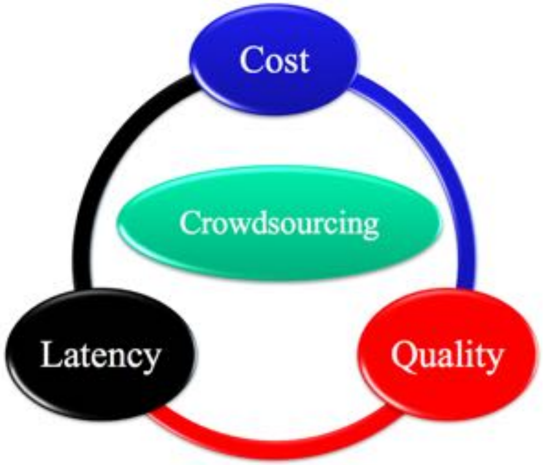
SYBASE

Microsoft SQL Server

# 6. Scalability: Query Optimization

- Query optimization in crowdsourcing is challenging:

(1) handle 3 optimization objectives



(2) humans are more unpredictable than machines

# 7. Privacy

- ○ (1) **Requester**

  **Wants to protect the privacy of their tasks from workers**

  ***e.g., tasks may contain sensitive attributes, e.g., medical data.***

# 7. Privacy

- (2) **Workers**

  **Want 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.***

# 8. Macro-Tasks

- Existing works focus on simple micro-tasks

Is Bill Gates currently the CEO of Microsoft ?
○ Yes ○ No

Identify the sentiment of the tweet: ……
○ Pos ○ Neu ○ Neg

- Hard to perform big and complex tasks, e.g., writing an essay

  (1) macro-tasks are hard to be split and accomplished by multiple workers;
  (2) workers may not be interested to perform a time-consuming macro-task.

# Thanks !
# Q & A

**Chengliang Chai**
Tsinghua University

**Ju Fan**
Renmin University

**Guoliang Li**
Tsinghua University

**Jiannan Wang**
SFU

**Yudian Zheng**
Twitter