Truth Discovery for Passive and Active Crowdsourcing

Jing Gao[1], Qi Li[1], and Wei Fan[2]
[1]SUNY Buffalo; [2]Baidu Research Big Data Lab

Overview

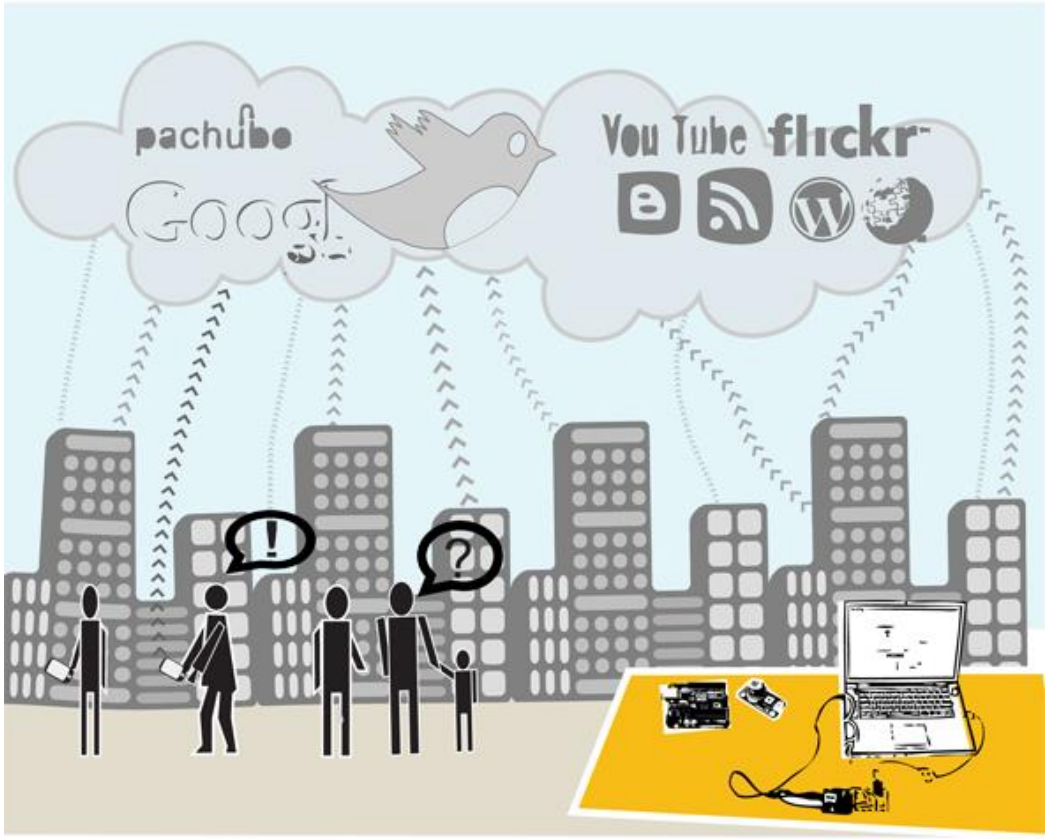
1. Introduction of Crowdsourcing
2. Crowdsourced Data Aggregation for Reliable Information Discovery
3. Passive Crowdsourcing Scenarios
4. Active Crowdsourcing Scenarios
5. Applications and Open Question
6. Resources
7. References

Overview

1. • **Introduction of Crowdsourcing**
2. • Crowdsourced Data Aggregation for Reliable Information Discovery
3. • Passive Crowdsourcing Scenarios
4. • Active Crowdsourcing Scenarios
5. • Applications and Open Question
6. • Resources
7. • References

Motivation

- Huge amounts of data contributed by users (user generated content, user behavioral data, sensory data, ……)
- Crowdsourced data contains valuable information and knowledge
- Inevitable error, noise and conflicts in the data
- Objective: obtain reliable information from crowdsourced data



Passive Crowdsourcing

HealthBoards HEALTH MESSAGE BOARDS MedHelp

"My girlfriend always gets a bad dry skin, rash on her upper arm, cheeks, and shoulders when she is on [**Depo**]. . . . "

"I have had no side effects from [**Depo**] (except ...), but otherwise no rashes..."

DEPO USER1 Bad dry skin
DEPO USER1 Rash
DEPO USER2 No rashes
. . . . . .

RxList MAYO CLINIC

DEPO Rash
. . . . . .

twitter facebook

"Made it through some pretty bad traffic! (John F. Kennedy International Airport (JFK) in New York, NY)"

"Good news....no traffic on George Washington bridge approach from Jersey"

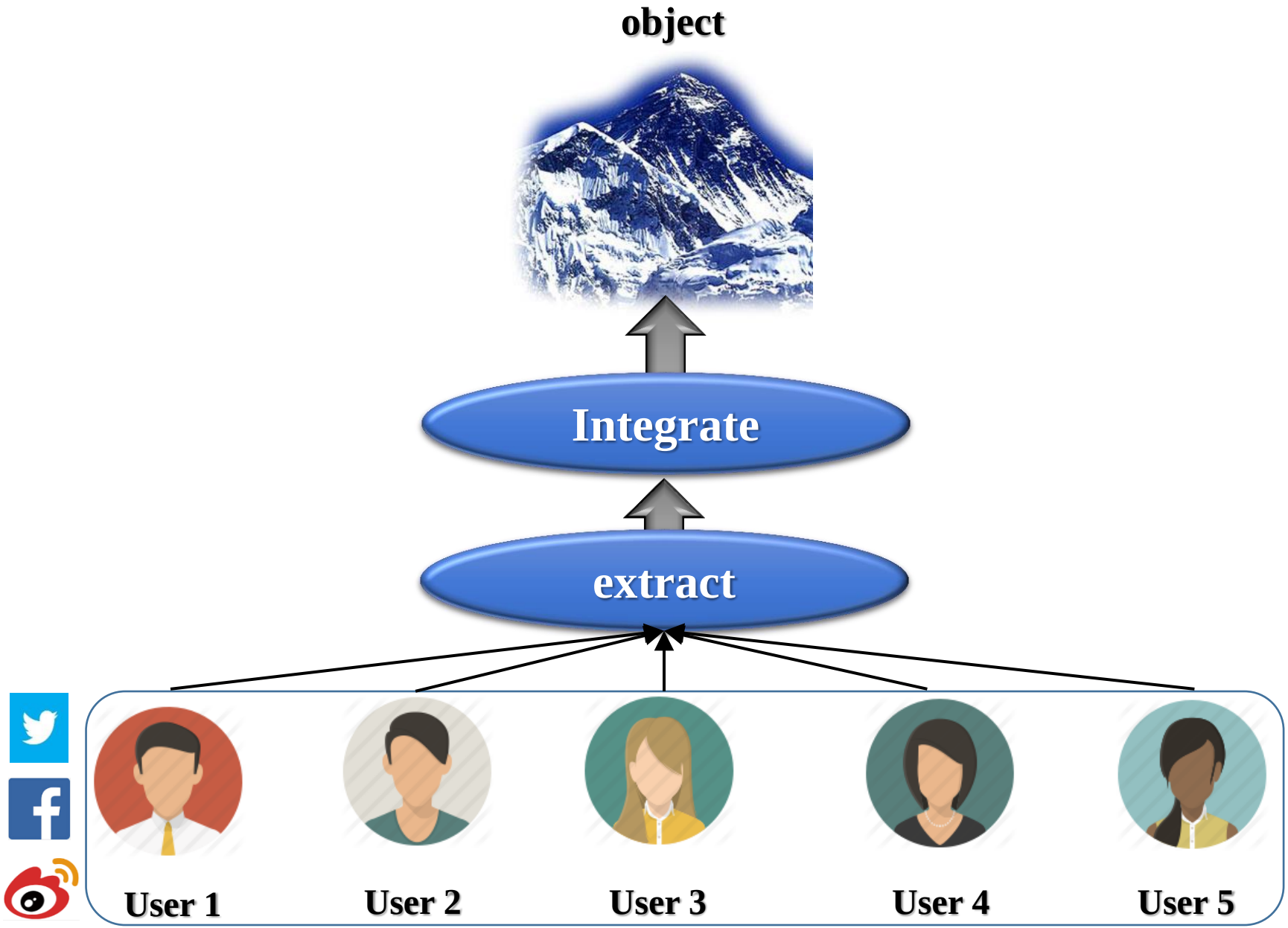
JFK airport Bad Traffic
JFK airport Good Traffic
. . . . . .

THE WHITE HOUSE WASHINGTON

JFK Bad Traffic
. . . . . .

Passive Crowdsourcing

- **Description**
  - Users/Data sources are sharing information on their own.
- **Goal**
  - To extract and integrate relevant information regarding a specific task

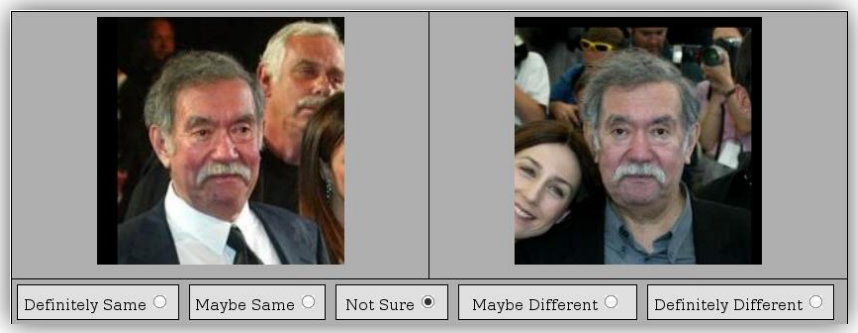

object
Integrate
extract
User 1
User 2
User 3
User 4
User 5

Active Crowdsourcing

amazon mechanical turk™ Artificial Artificial Intelligence

CrowdFlower

Are the two images of the same person?



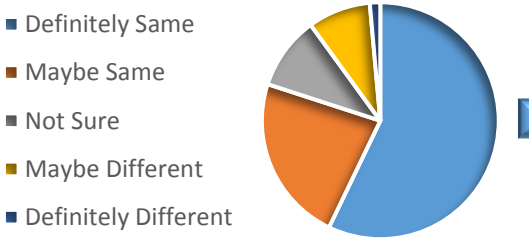

Definitely Same


Maybe Same

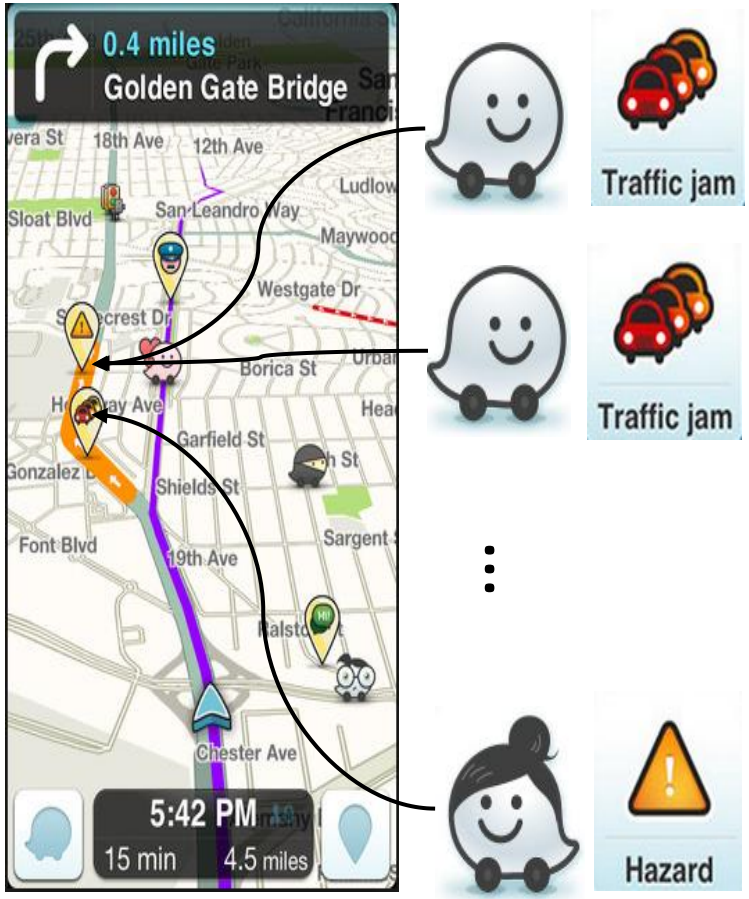
...


Not Sure

Annotation Results



Final Answer:
Same



Active Crowdsourcing

- **Description**
  - Users/Data sources generate information based on requests.
- **Goal**
  - To actively design and collect data for a specific task. And then integrate the information.

Task

query

User 1 User 2 User 3 User 4 User 5

Task

Integration

User 1 User 2 User 3 User 4 User 5

Overview

1. Introduction of Crowdsourcing
2. **Crowdsourced Data Aggregation for Reliable Information Discovery**
3. Passive Crowdsourcing Scenarios
4. Active Crowdsourcing Scenarios
5. Applications and Open Question
6. Resources
7. References

A Straightforward Fusion Solution

- **Voting/Averaging**
  - Take the value that is claimed by majority of the sources
  - Or compute the mean of all the claims
- **Limitation**
  - Ignore source reliability
- **Source reliability**
  - Is crucial for finding the true fact but unknown

Truth Discovery & Crowdsourced Data Aggregation

- **Problem**
  - Input: Multiple conflicting information about the same set of objects provided by various information sources
  - Goal: Discover trustworthy information (i.e., the **truths**) from conflicting data on the same object

Truth Discovery & Crowdsourced Data Aggregation

- **Principle**
  - Infer both truth and source reliability from the data
    - A source is reliable if it provides many pieces of true information
    - A piece of information is likely to be true if it is provided by many reliable sources

Truth Discovery & Crowdsourced Data Aggregation

- A common goal
  - to improve the quality of the aggregation/fusion results
- Via a common method
  - To aggregate by estimating source reliabilities
- Similar principles
  - Data from reliable sources are more likely to be accurate
  - A source is reliable if it provides accurate information
- Mutual challenge
  - Prior knowledge and labels are rarely available

Data Collection and Generation

Truth discovery

- We can't control generation step.
- We only collect.

passive

Crowdsourced data aggregation

- We can control data generation to a certain degree
  - What to ask
  - How to ask
  - How many labels per question

active

Data Format of Claims

Truth discovery

- Data is collected from open domain.
- Can't define data space
  - type of data
  - range of data

Open space

Crowdsourced data aggregation

- Data generation is controlled
- For easier validation of answers, requesters usually choose
  - Multi-choice question
  - Scoring in a range

Closed space

Model Categories

- Statistical model (STA)
  - Generative model (GM)
- Optimization model (OPT)

Statistical Model (STA)

- General goal:
  - **To find the (conditional) probability of a claim being true**
- Source reliability:
  - **Probability(ies) of a source/worker making a true claim**

STA - TruthFinder

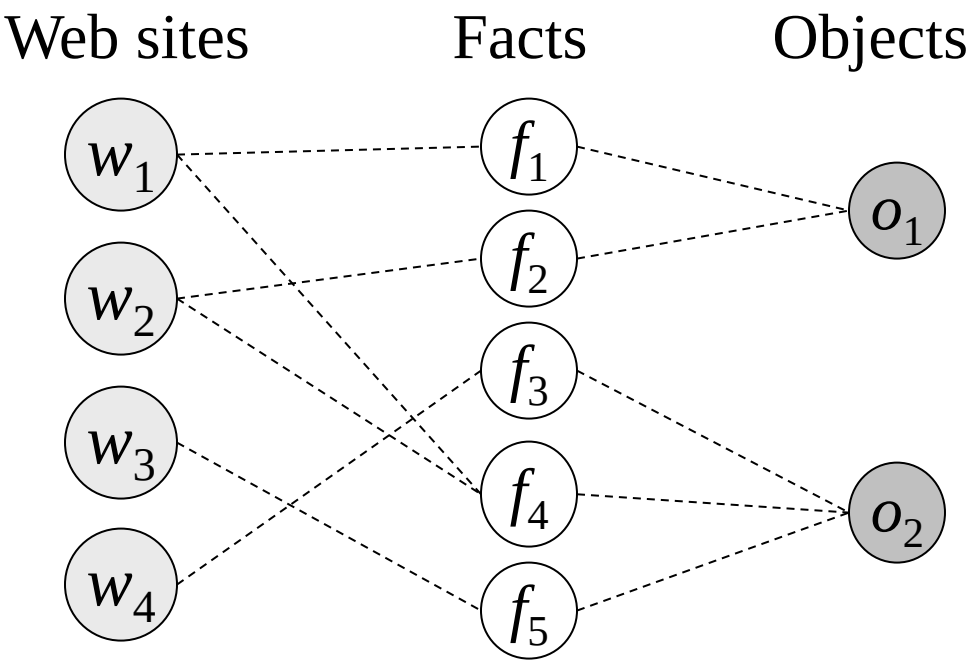
Different websites often provide conflicting information on a subject, e.g., Authors of *"Rapid Contextual Design"*

Online Store	*Authors*
Powell's books	Holtzblatt, Karen
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Cornwall books	Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley

[Yin et al., TKDE'08]

STA - TruthFinder

- Each object has a set of conflictive facts
  - E.g., different author lists for a book
- And each web site provides some facts
- How to find the true fact for each object?



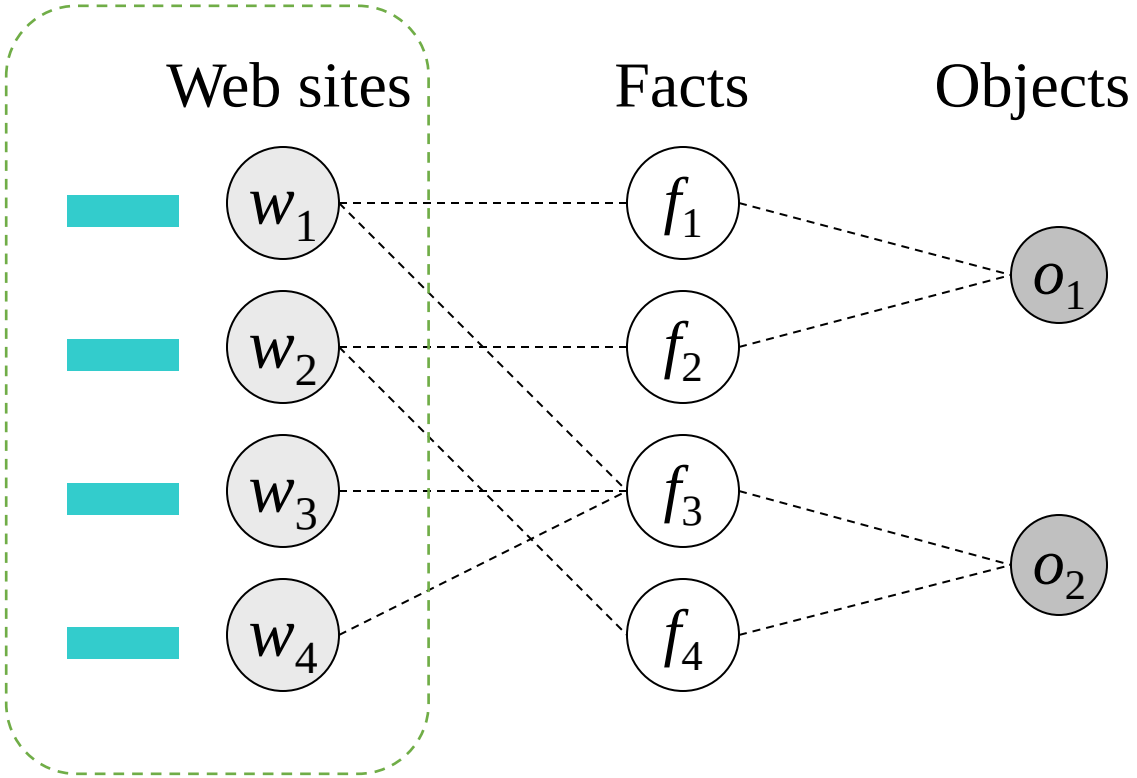
STA - TruthFinder

1. There is usually only one true fact for a property of an object
2. This true fact appears to be the same or similar on different web sites
   - E.g., “Jennifer Widom” vs. “J. Widom”
3. **The false facts on different web sites are less likely to be the same or similar**
   - False facts are often introduced by random factors
4. **A web site that provides mostly true facts for many objects will likely provide true facts for other objects**

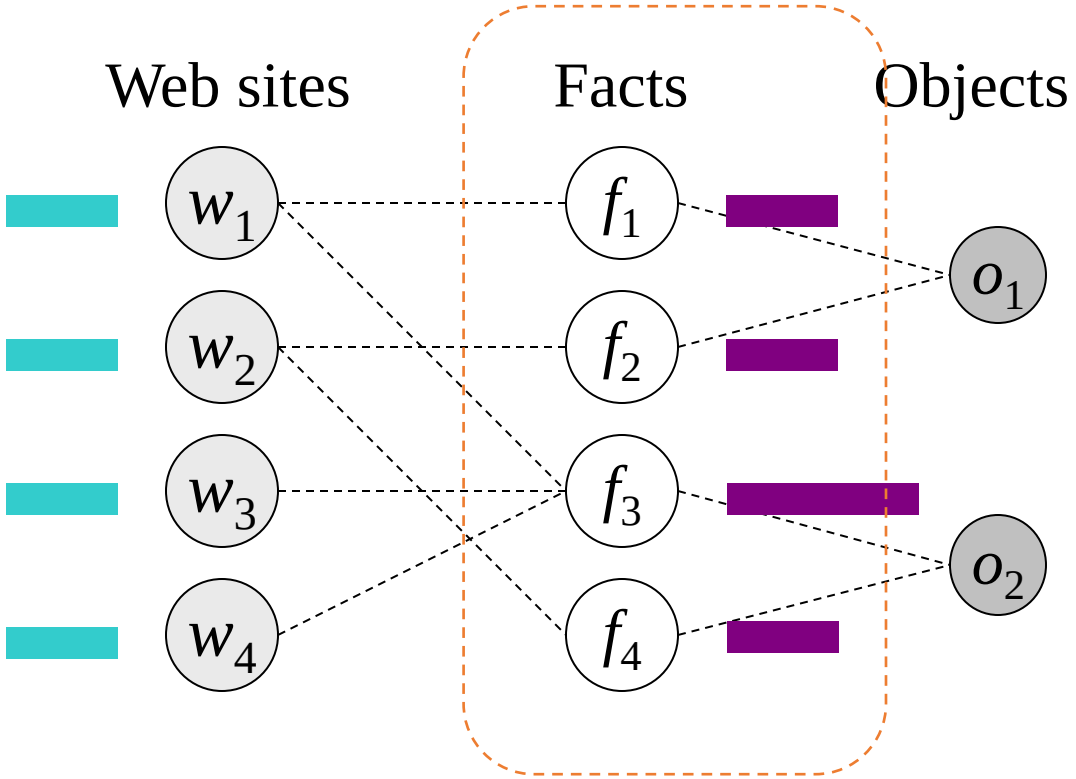
STA - TruthFinder

- *Confidence of facts* ↔ *Trustworthiness of web sites*
  - A fact has *high confidence* if it is provided by (many) trustworthy web sites
  - A web site is *trustworthy* if it provides many facts with high confidence
- Iterative steps
  - Initially, each web site is equally trustworthy
  - Based on the four heuristics, infer fact confidence from web site trustworthiness, and then backwards
  - Repeat until achieving stable state

STA - TruthFinder

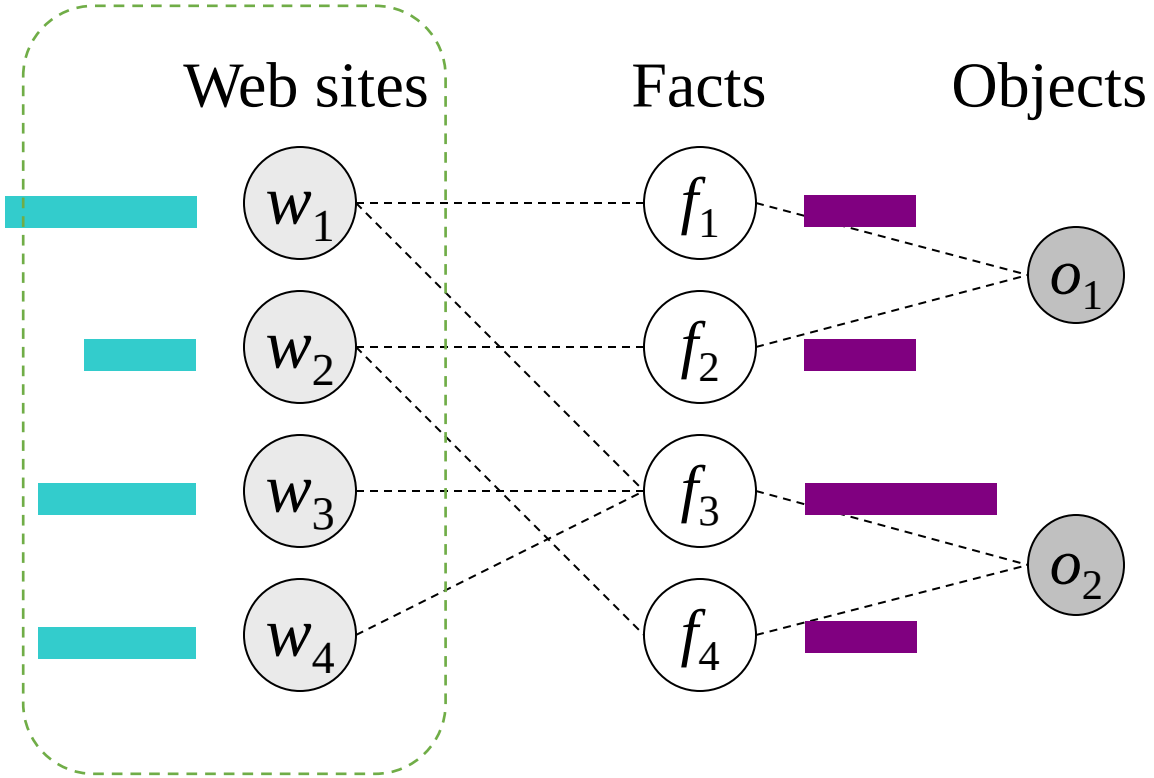


STA - TruthFinder



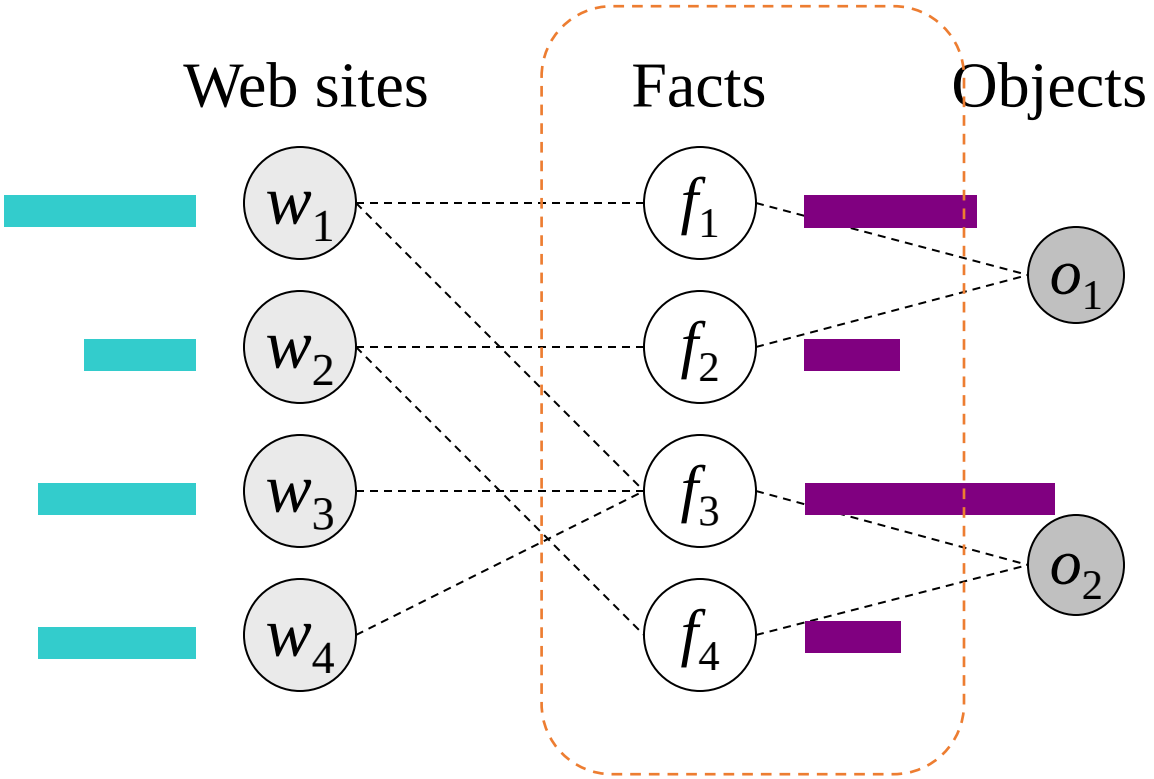
26

STA - TruthFinder



27

STA - TruthFinder



28

STA - TruthFinder

- **The trustworthiness of a web site w: $t(w)$**
  - Average confidence of facts it provides

$$t(w) = \frac{\sum_{f \in F(w)} s(f)}{|F(w)|}$$

Sum of fact confidence

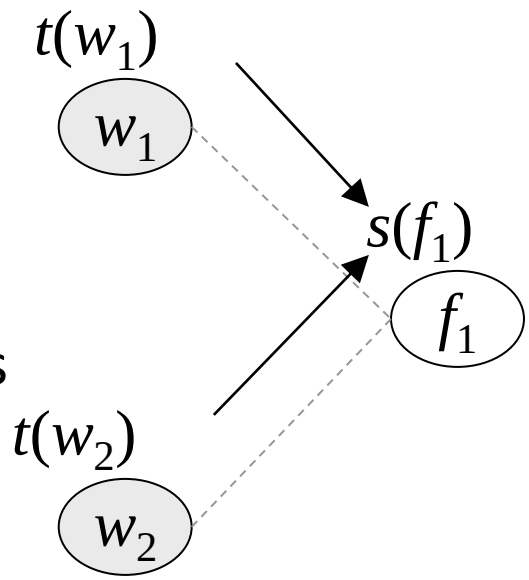
Set of facts provided by w

- **The confidence of a fact f: $s(f)$**
  - One minus the probability that all web sites providing f are wrong

$$s(f) = 1 - \prod_{w \in W(f)} (1 - t(w))$$

Probability that w is wrong

Set of websites providing f

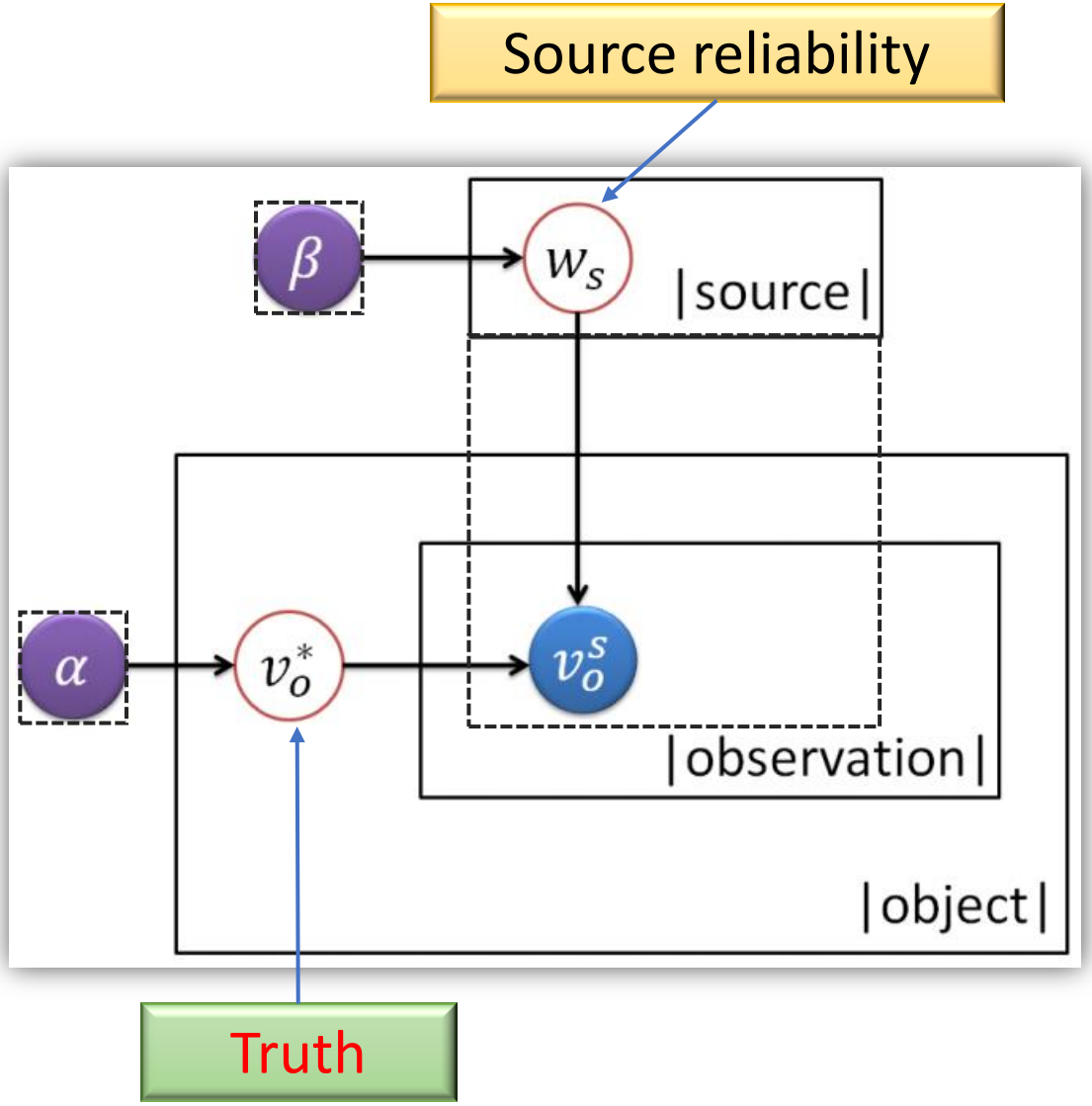


STA - TruthFinder

Type of error	Voting	TruthFinder	Barnes&Noble
Correct	71	85	64
Miss author(s)	12	2	4
Incomplete names	18	5	6
Wrong first/middle names	1	1	3
Has redundant names	0	2	23
Add incorrect names	1	5	5
No information	0	0	2

- Viewing an author list as a fact

Generative Model (GM)



Generative Model (GM)

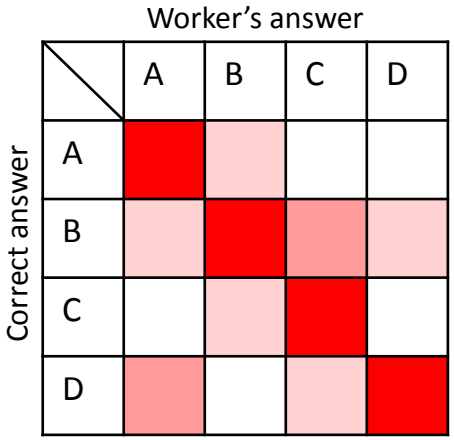
- One of the most popular models
  - GTM [Zhao&Han, QDB'12]
  - LTM [Zhao et al., VLDB'12]
  - MSS [Qi et al., WWW'13]
  - LCA [Pasternack&Roth, WWW'13]
  - TEM [Zhi et al., KDD'15]
  - DS [Dawid&Skene, 1979]
  - GLAD [Whitehill et al., NIPS'09]

  …

GM - Maximum Likelihood Estimation

Multiple choice questions with fixed answer space

For each worker, the reliability is a confusion matrix.



$\pi_{jl}^{(k)}$: the probability that worker k answers l when j is the correct answer.

p_j : the probability that a randomly chosen question has correct answer j.

[Dawid&Skene, 1979]

GM - Maximum Likelihood Estimation

$$likelihood_i^{(k)} | q \text{ is correct} = \prod_{l=1}^{J} \pi_{ql}^{(k)}$$

$$likelihood_i | q \text{ is correct} = \prod_{k}^{K} \prod_{l=1}^{J} \pi_{ql}^{(k)}$$

$$likelihood_i = \prod_{j=1}^{J} \left(p_j \prod_{k}^{K} \prod_{l=1}^{J} \pi_{jl}^{(k)} \right)^{1(j=q)}$$

GM - Maximum Likelihood Estimation

$$likelihood = \prod_{i}^{I} \prod_{j=1}^{J} \left(p_j \prod_{k}^{K} \prod_{l=1}^{J} \pi_{jl}^{(k)} \right)^{1(j_i = q_i)}$$

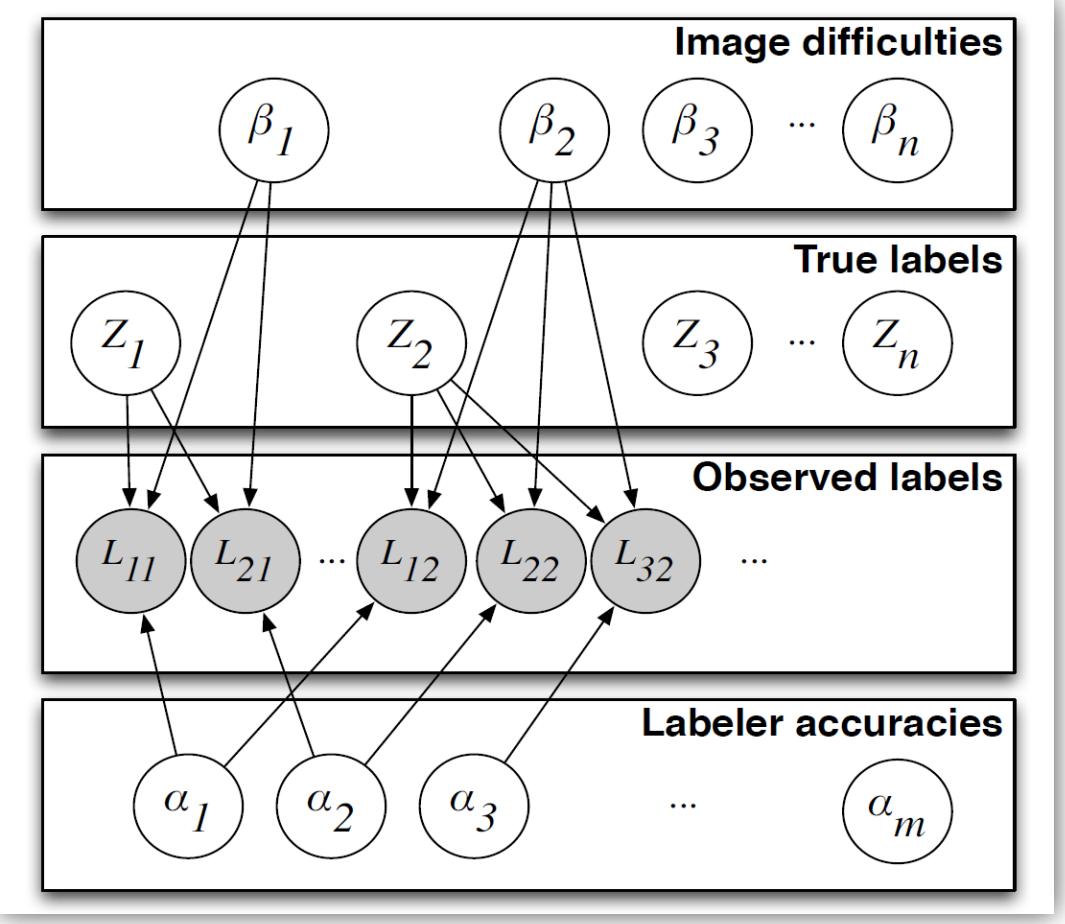
- This is the likelihood if the correct answers (i.e., q_i's) are known.
- What if we don't know the correct answers?
- Unknown parameters are p_j, q, $\pi_{jl}^{(k)}$

EM algorithm

GM - Extension and Theoretical Analysis

- Extensions
  - Naïve Bayesian [Snow et al., EMNLP'08]
  - Finding a good initial point [Zhang et al., NIPS'14]
  - Adding instances' feature vectors [Raykar et al., 2010] [Lakkaraju et al. 2015]
  - Using prior over worker confusion matrices [Raykar et al., 2010][Liu et al., NIPS'12] [Lakkaraju et al. SDM'15]
  - Clustering workers/instances [Lakkaraju et al. SDM'15]
- Theoretical analysis
  - Error bound [Li et al., 2013] [Zhang et al., NIPS'14]

GM - GLAD Model

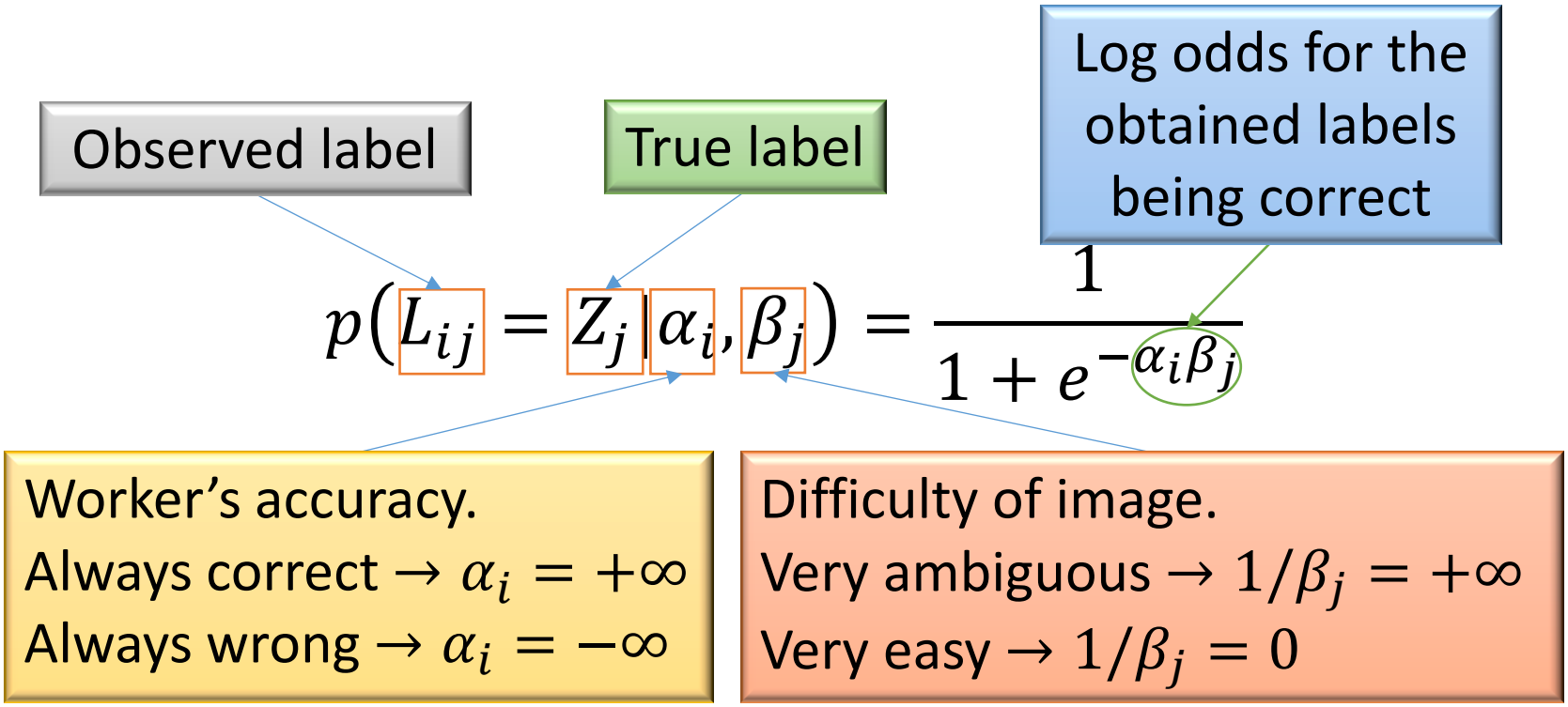


Each image belongs to one of two possible categories of interest, i.e., binary labeling.

Known variables: observed labels.

[Whitehill et al., NIPS'09]

GM - GLAD Model



GM - Latent Truth Model (LTM)

- **Multiple** facts can be **true** for each entity (object)
  - One book may have 2+ authors
- A source can make **multiple claims per entity**, where more than one of them can be true
  - A source may claim a book w. 3 authors
- Sources and objects are **independent** respectively
  - Assume book websites and books are independent
- The majority of data coming from many sources are not erroneous
  - Trust the majority of the claims

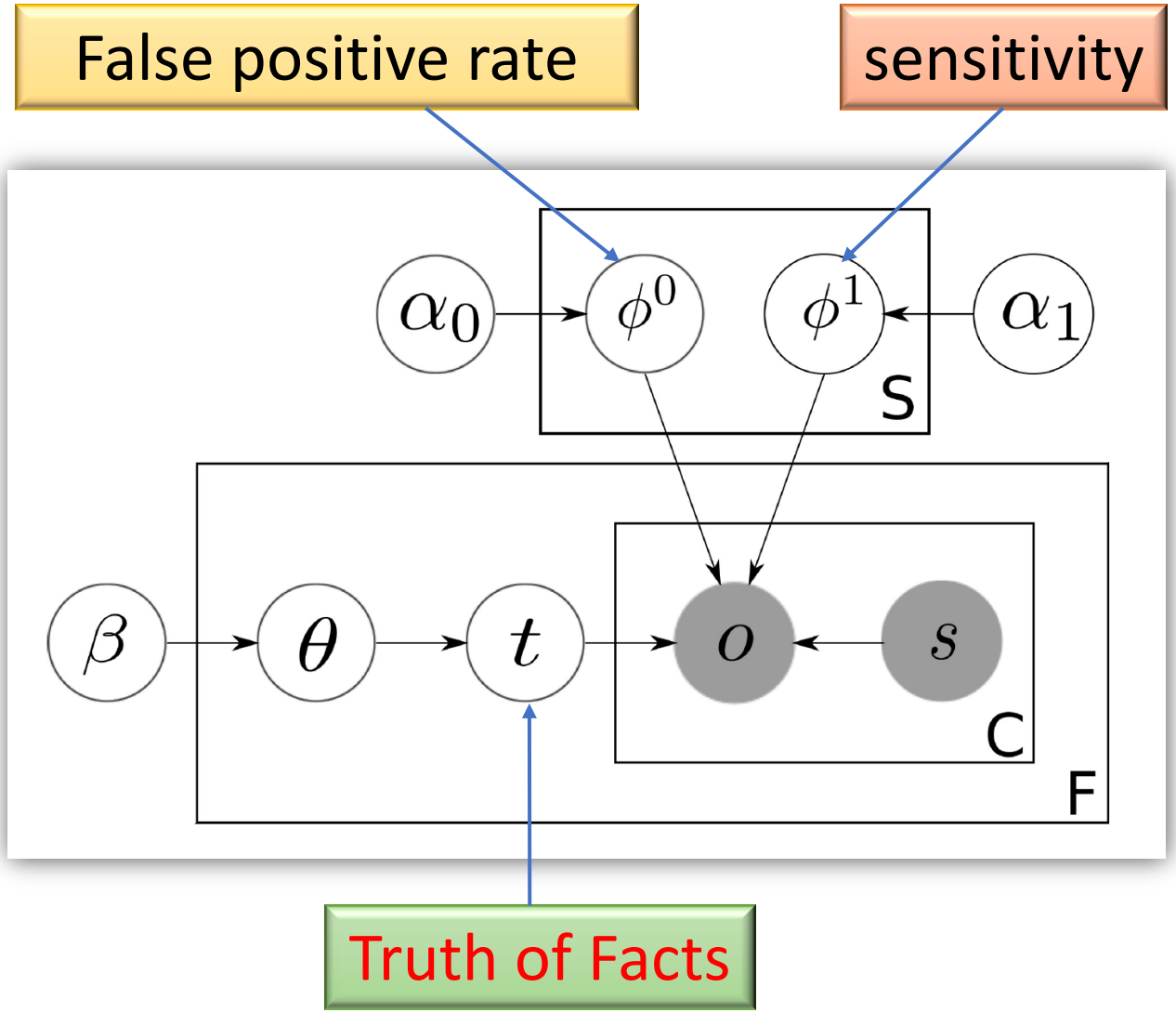
[Zhao et al., VLDB'12]

GM - Latent Truth Model (LTM)

Input			Output
RID	Source	Observation	Truth
1	Barnes&Noble	True	True
1	Brett's Books	True	
1	Ecampus.com	True	
2	Barnes&Noble	True	True
2	Brett's Books	False	
2	Ecampus.com	False	
3	Brett's Books	True	True
…	…	…	

RID	Entity (book)	Attribute (Author)
1	Data Mining: Concepts and Techniques	Jiawei Han
2	Data Mining: Concepts and Techniques	Micheline Kamber
3	Introduction to Algorithms	Thomas H. Cormen

GM - Latent Truth Model (LTM)



GM - Latent Truth Model (LTM)

- For each source k
  - Generate false positive rate (with **strong** regularization, believing most sources have low FPR): $\phi_k^0 \sim Beta(\alpha_{0,1}, \alpha_{0,0})$
  - Generate its sensitivity (1-FNR) with uniform prior, indicating low FNR is more likely: $\phi_k^1 \sim Beta(\alpha_{1,1}, \alpha_{1,0})$
- For each fact f
  - Generate its prior truth prob, uniform prior: $\theta_f \sim Beta(\beta_1, \beta_0)$
  - Generate its truth label: $t_f \sim Bernoulli(\theta_f)$
- For each claim c of fact f, generate observation of c.
  - If f is false, use false positive rate of source:$o_c \sim Bernoulli(\phi_{s_c}^0)$
  - If f is true, use sensitivity of source: $o_c \sim Bernoulli(\phi_{s_c}^1)$

GM - Latent Truth Model (LTM)

Results on book data

	Precision	Recall	FPR	Accuracy	F1
LTM	1.000	0.995	0.000	0.995	0.997
TruthFinder	0.880	1.000	1.000	0.880	0.936
Voting	1.000	0.863	0.000	0.880	0.927

Optimization Model (OPT)

- General model

$$\arg\min_{\{w_s\},\{v_o^*\}} \sum_{o\in O}\sum_{s\in S} g(w_s, v_o^*)$$
$$s.t.\ \delta_1(w_s) = 1, \delta_2(v_o^*) = 1$$

- What does the model mean?
  - The optimal solution can minimize the objective function
  - Joint estimate true claims v_o^* and source reliability w_s under some constraints δ_1, δ_2, … .
  - Objective function $g(\cdot,\cdot)$ can be distance, entropy, etc.

Optimization Model (OPT)

- General model

$$\arg\min_{\{w_s\},\{v_o^*\}} \sum_{o\in O}\sum_{s\in S} g(w_s, v_o^*)$$
$$s.t.\ \delta_1(w_s) = 1, \delta_2(v_o^*) = 1$$

- How to solve the problem?
  - Convert the primal problem to its (Lagrangian) dual form
  - Block coordinate descent to update parameters
  - If each sub-problem is convex and smooth, then convergence is guaranteed

OPT - CRH Framework

$$\min_{\mathcal{X}^{(*)},\, \mathcal{W}} \quad f(\mathcal{X}^{(*)}, \mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{i=1}^{N} \sum_{m=1}^{M} d_m\left(v_{im}^{(*)}, v_{im}^{(k)}\right)$$

$$\text{s.t.} \quad \delta(\mathcal{W}) = 1, \qquad \mathcal{W} \geq 0.$$

Basic idea

- Truths should be close to the observations from reliable sources
- Minimize the overall weighted distance to the truths in which reliable sources have high weights

[Li et al., SIGMOD'14]

OPT - CRH Framework

- **Loss function**
  - d_m: loss on the data type of the *m*-th property
  - Output a high score when the observation deviates from the truth
  - Output a low score when the observation is close to the truth
- **Constraint function**
  - The objective function may go to $-\infty$ without constraints
  - Regularize the weight distribution

OPT - CRH Framework

- **Run the following until convergence**
  - Truth computation
    - Minimize the weighted distance between the truth and the sources' observations

$$v_{im}^{(*)} \leftarrow \arg\min_{v} \sum_{k=1}^{K} w_k \cdot d_m\left(v, v_{im}^{(k)}\right)$$

  - Source reliability estimation
    - Assign a weight to each source based on the difference between the truths and the observations made by the source

$$\mathcal{W} \leftarrow \arg\min_{\mathcal{W}} f(\mathcal{X}^{(*)}, \mathcal{W})$$

OPT - Minimax Entropy

- Workers: $i = 1, 2, \ldots, m$
- Items: $j = 1, 2, \ldots, n$
- Categories: $k = 1, 2, \ldots, c$

Input: response tensor $Z_{m \times n \times c}$

- $z_{ijk} = 1$, if worker i labels item j as category k
- $z_{ijk} = 0$, if worker i labels item j as others (not k)
- $z_{ijk} =$ unknown , if worker i does not label item j

Goal: Estimate the ground truth y_{jl}

[Zhou et al., NIPS'12]

OPT - Minimax Entropy

	item 1	item 2	…	item n
worker 1	z_{11}	z_{12}	…	z_{1n}
worker 2	z_{21}	z_{22}	…	z_{2n}
…	…	…	…	
worker m	z_{m1}	z_{12}	…	z_{mn}

OPT - Minimax Entropy

	item 1	item 2	…	item n
worker 1	π_{11}	π_{12}	…	π_{1n}
worker 2	π_{21}	π_{22}	…	π_{2n}
…	…	…	…	
worker m	π_{m1}	π_{12}	…	π_{mn}

π_{ij} is a vector that presents the underline distribution of the observation.

i.e., z_{ij} is drawn from π_{ij}.

OPT - Minimax Entropy

	item 1	item 2	…	item n
worker 1	π_{11}	π_{12}	…	π_{1n}
worker 2	π_{21}	π_{22}	…	π_{2n}
…	…	…	…	
worker m	π_{m1}	π_{12}	…	π_{mn}

Column constraint: the number of votes per class per item $\sum_i z_{ijk}$ should match $\sum_i \pi_{ijk}$

OPT - Minimax Entropy

	item 1	item 2	…	item n
worker 1	π_{11}	π_{12}	…	π_{1n}
worker 2	π_{21}	π_{22}	…	π_{2n}
…	…	…	…	
worker m	π_{m1}	π_{12}	…	π_{mn}

Row constraint : the empirical confusion matrix per worker $\sum_j y_{jl} z_{ijk}$ should match $\sum_j y_{jl} \pi_{ijk}$

OPT - Minimax Entropy

- If we **know** the true label y_{jl}
- **Maximum** entropy of π_{ijk} under constraints

$$\max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c}\pi_{ijk}\ln\pi_{ijk}$$

$$\text{s.t.} \quad \sum_{i=1}^{m}\pi_{ijk}=\sum_{i=1}^{m}z_{ijk},\ \forall j,k,\ \sum_{j=1}^{n}y_{jl}\pi_{ijk}=\sum_{j=1}^{n}y_{jl}z_{ijk},\ \forall i,k,l,$$

$$\sum_{k=1}^{c}\pi_{ijk}=1,\ \forall i,j,\ \pi_{ijk}\geq 0,\ \forall i,j,k.$$

OPT - Minimax Entropy

- To **estimate** the true label y_{jl}
- **Minimizing** the **maximum** entropy of π_{ijk}

$$\min_{y} \max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk},\ \forall j,k,\ \sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk},\ \forall i,k,l,$$

$$\sum_{k=1}^{c} \pi_{ijk} = 1,\ \forall i,j,\ \pi_{ijk} \geq 0,\ \forall i,j,k,\ \sum_{l=1}^{c} y_{jl} = 1,\ \forall j,\ y_{jl} \geq 0,\ \forall j,l.$$

OPT - Minimax Entropy

- To **estimate** the true label y_{jl}
- **Minimizing** the **maximum** entropy of π_{ijk}

$$\min_{y} \max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \pi_{i} \qquad \sum_{j=1}^{n} \qquad \sum_{l=1}^{n}$$

$$\sum_{k=1}^{c} \pi_{i} \qquad\qquad\qquad\qquad\qquad\qquad 0, \ \forall j, l.$$

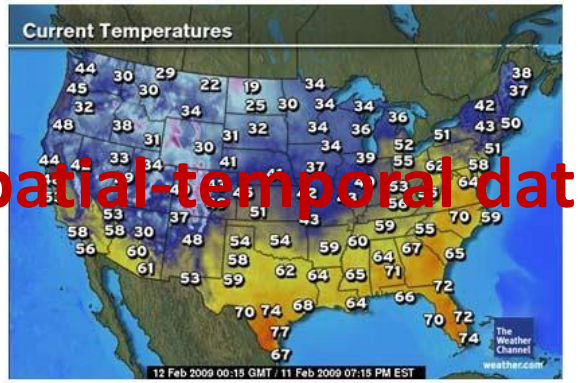
Minimize entropy
is equivalent to
minimizing the KL divergence

Overview

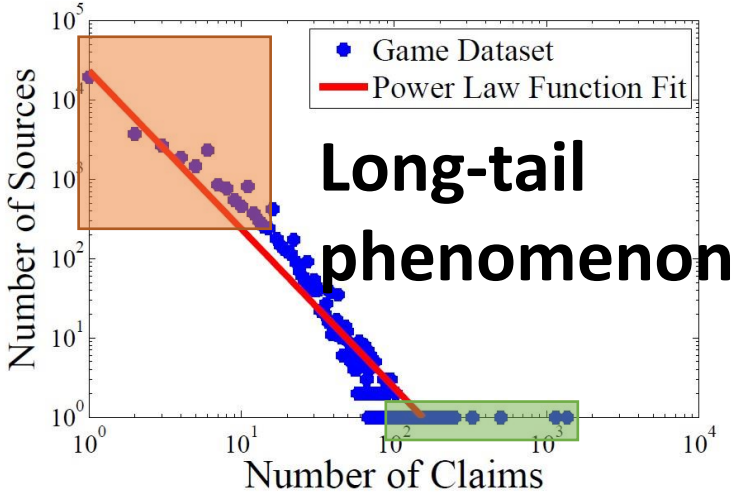
1. Introduction of Crowdsourcing
2. Crowdsourced Data Aggregation for Reliable Information Discovery
3. **Passive Crowdsourcing Scenarios**
4. Active Crowdsourcing Scenarios
5. Applications and Open Questions
6. Resources
7. References

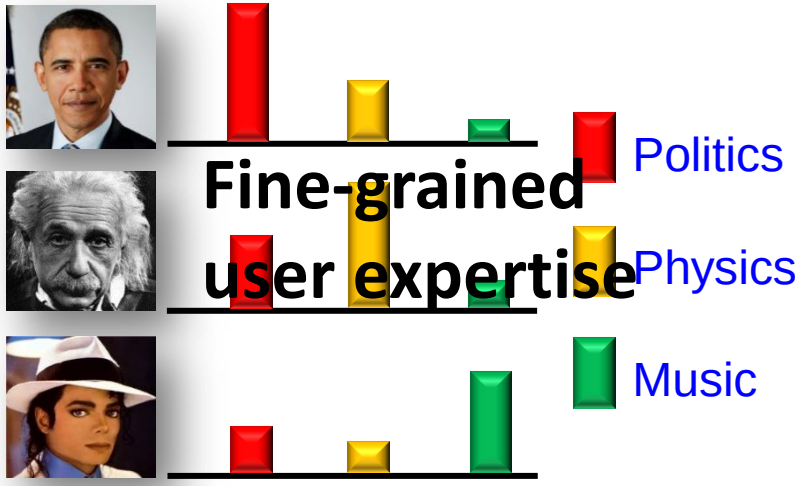
Aggregation of Passively Crowdsourced Data

- More challenges



Weather Condition





Source Correlations

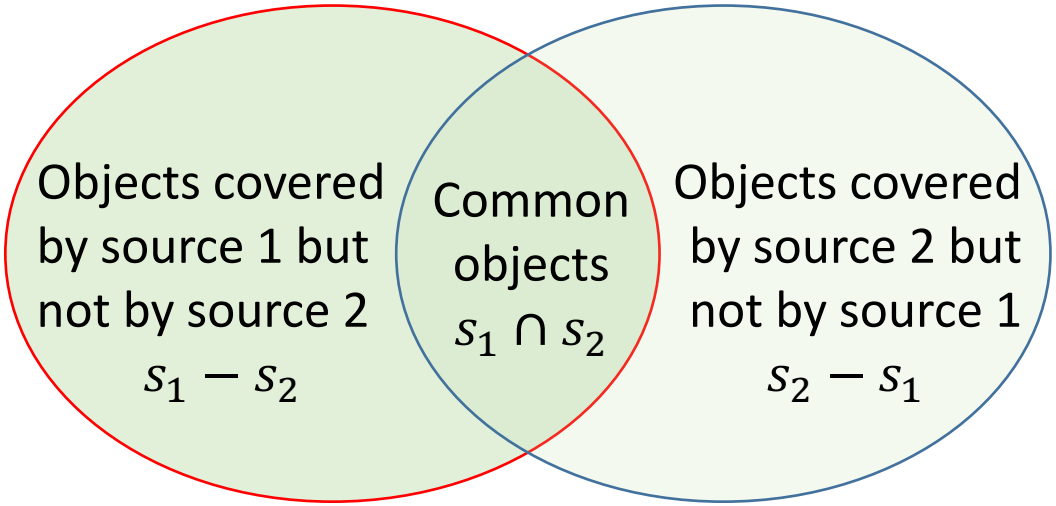
- Many truth discovery methods consider independent sources
  - Sources provide information independently
  - Source correlation can be hard to model
  - However, this assumption may be violated in real life
- Copy relationships between sources
  - Sources can copy information from one or more other sources
- General correlations of sources
  - Sources may provide data from complementary domains (negative correlation)
  - Sources may apply common rules in extraction (positive correlation)

Source Dependency

- Known relationships
  - Apollo-Social [Wang et al., IPSN'14]
    - For a claim, a source may copy from a related source with a certain probability
    - Used MLE to estimate a claim being correct
- Unknown relationships
  - Accu-Copy [Dong et al., VLDB'09a] [Dong et al., VLDB'09b]
  - MSS [Qi et al., WWW'13]
    - Modeled as a PGM
    - Related sources are grouped together and assigned with a group weight

Copy Relationships between Sources

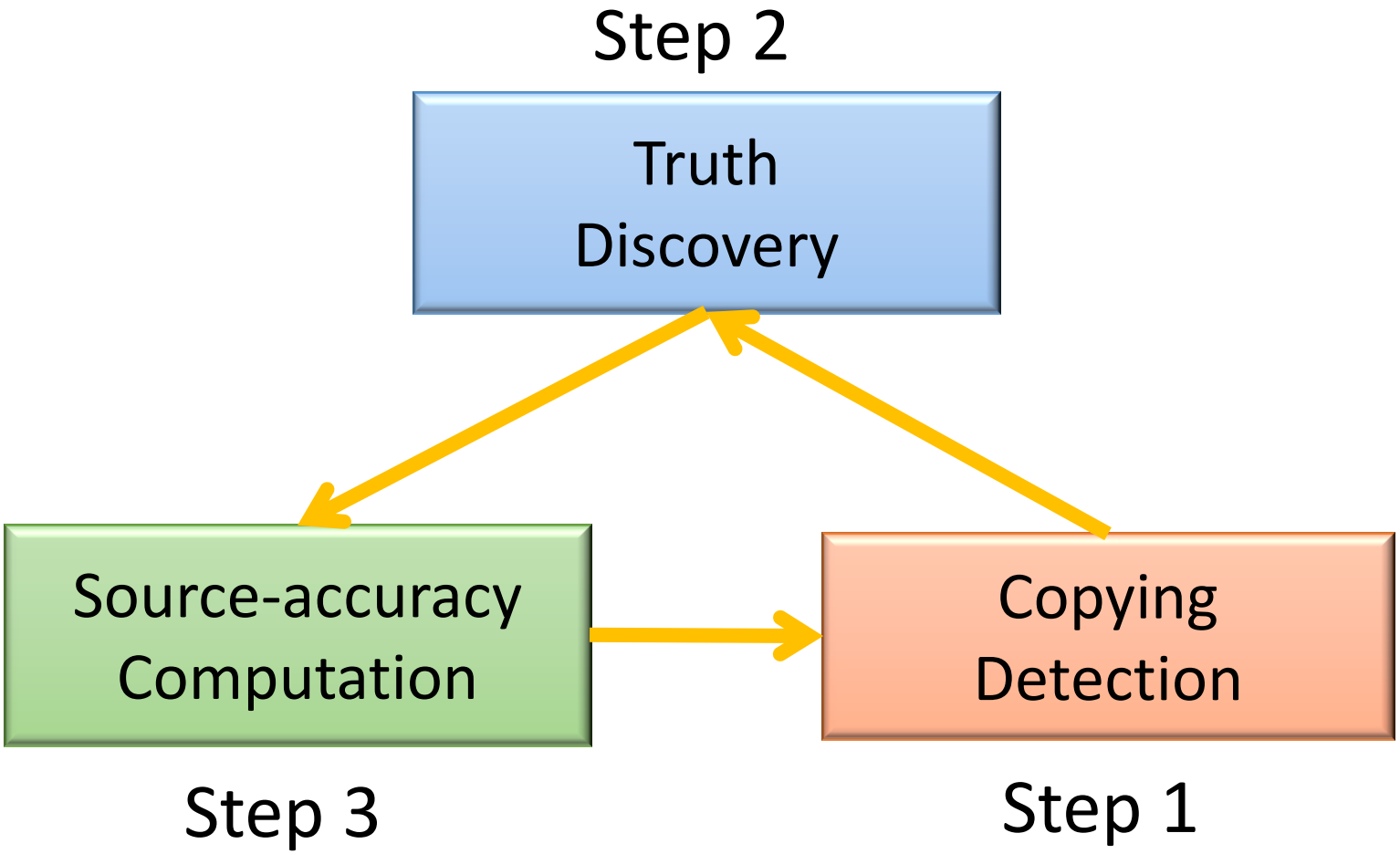
- High-level intuitions for copying detection
  - Common error implies copying relation
    - e.g., many same errors in $s_1 \cap s_2$ imply source 1 and 2 are related
  - Source reliability inconsistency implies copy direction
    - e.g., $s_1 \cap s_2$ and $s_1 - s_2$has similar accuracy, but $s_1 \cap s_2$ and $s_2 - s_1$ has different accuracy, so source 2 may be a copier.



[Dong et al., VLDB'09a] [Dong et al., VLDB'09b] [Pochampally et al., SIGMOD'14]

Copy Relationships between Sources

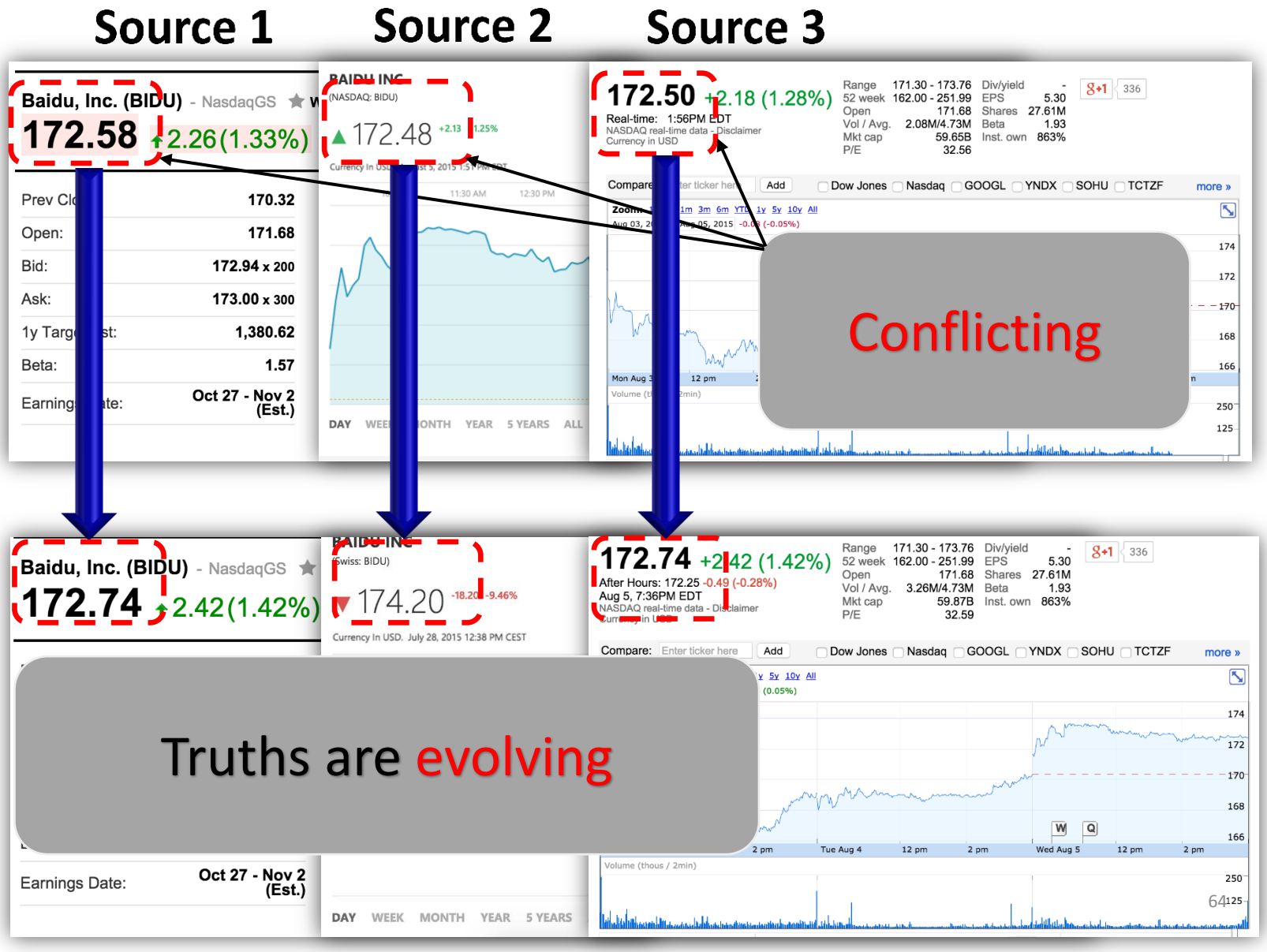
- Incorporate copying detection in truth discovery



Spatial-Temporal Data

- Challenges of dynamic data
  - Efficiency
- Correlation among entities
  - Data smoothness

Real Time Truth Discovery



Real Time Truth Discovery - DynaTD

- Challenges of dynamic data
  - Efficiency: When data comes sequentially, the iterative procedure is time costly
  - Temporal relations exist among entities
  - Source reliability changes: Observed source reliability fluctuates around a certain value.

[Li et al., KDD'15]

Real Time Truth Discovery - DynaTD

- Loss function (similar to [Li et al., SIGMOD'14])

$$L_T = \sum_{t=1}^{T} l_t = \sum_{t=1}^{T} \theta \sum_{s=1}^{S} w_s \sum_{o=1}^{c_t^s} \left(v_{o,t}^s - v_{o,t}^*\right)^2 - \sum_{s=1}^{S} c_t^s \log(w_s)$$

- Solution
  - Equivalence between the optimization problem and the maximization of error likelihood
  - Derive the incremental truth discovery algorithm which can dynamically update source weights and compute truths upon the arrival of new data

Real Time Truth Discovery - DynaTD

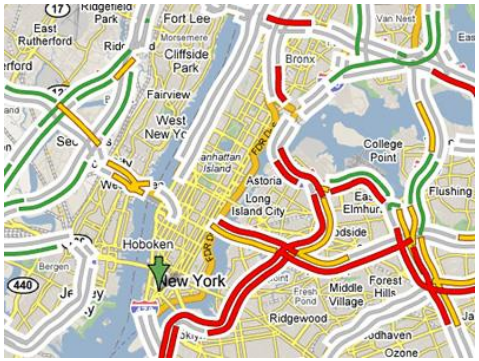
Source reliability evolves over time

Update source reliability based on continuously arriving data:

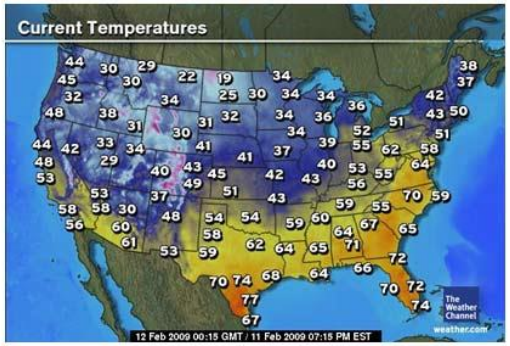
$$p(w_s|e^s_{1:T}) \propto p(e^s_T|w_s)p(w_s|e^s_{1:T-1})$$

Correlation Among Entities

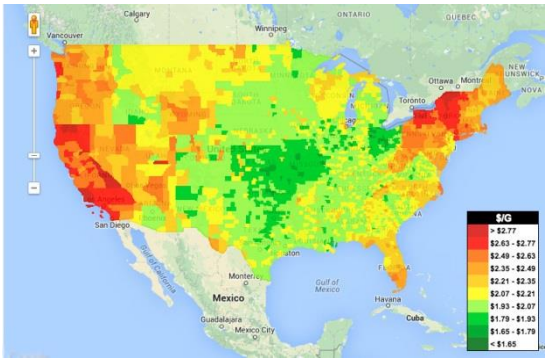
- Example
  - Temporal correlation
  - Spatial correlation
  - Etc.



Traffic Condition



Weather Condition



Gas Price

Mobile Sensing





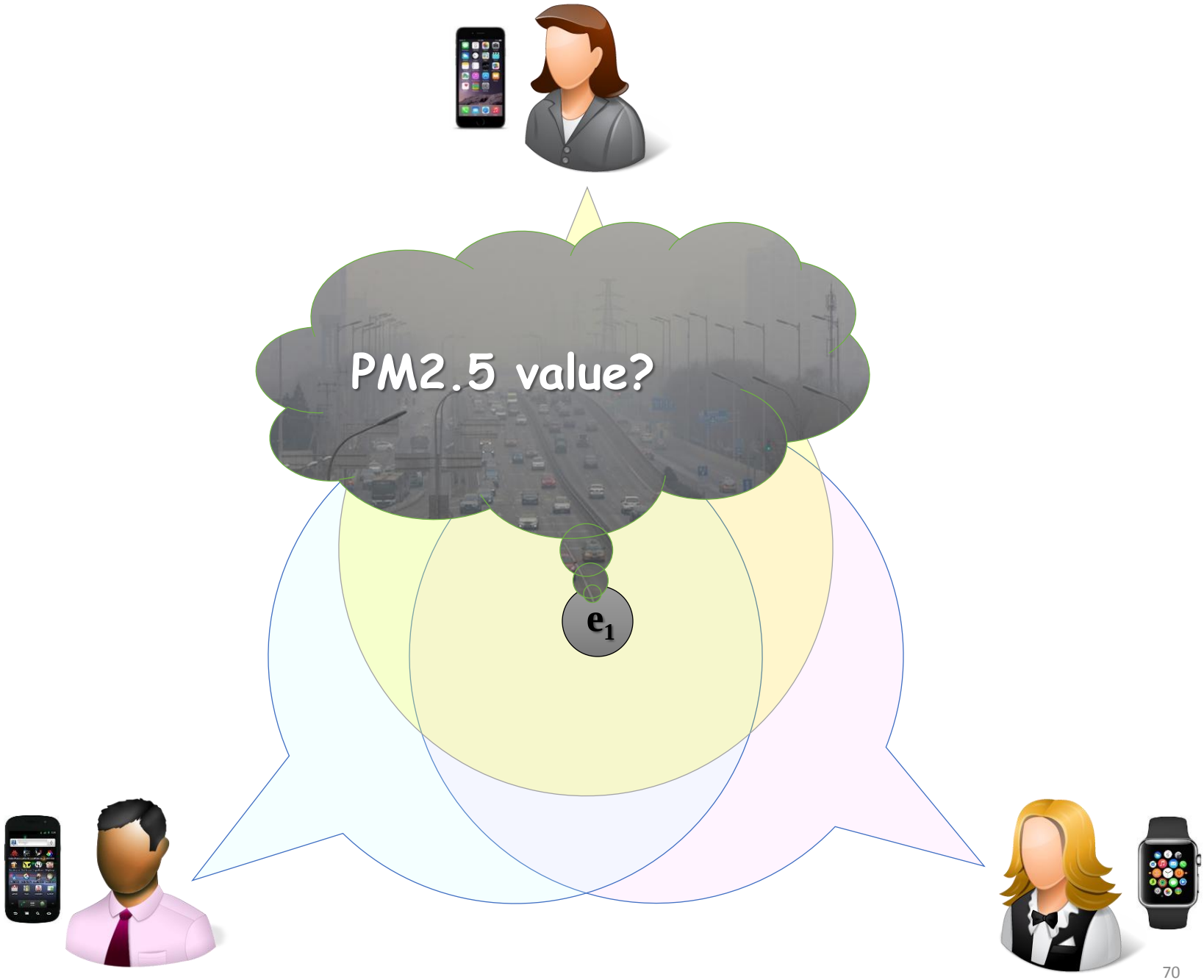


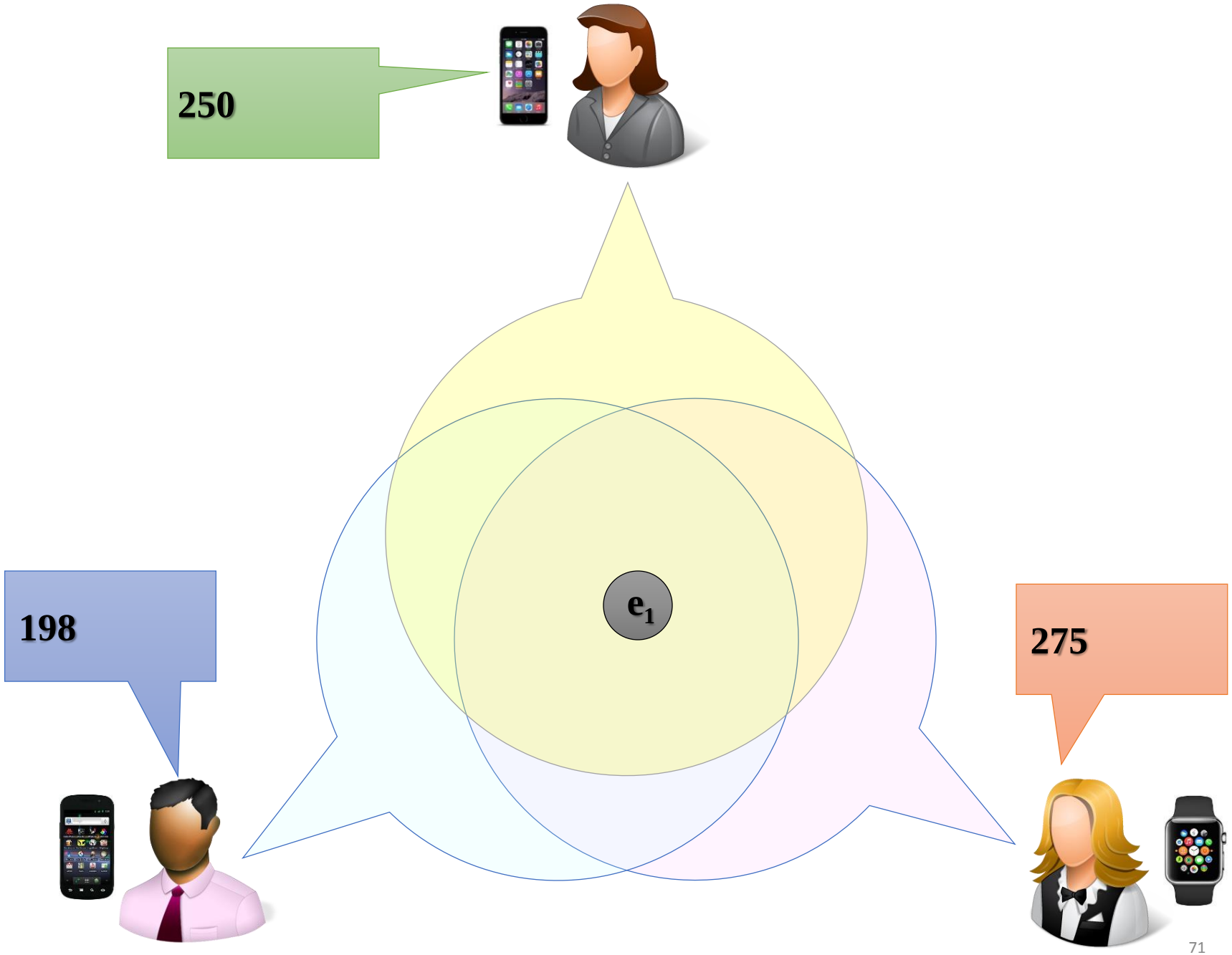




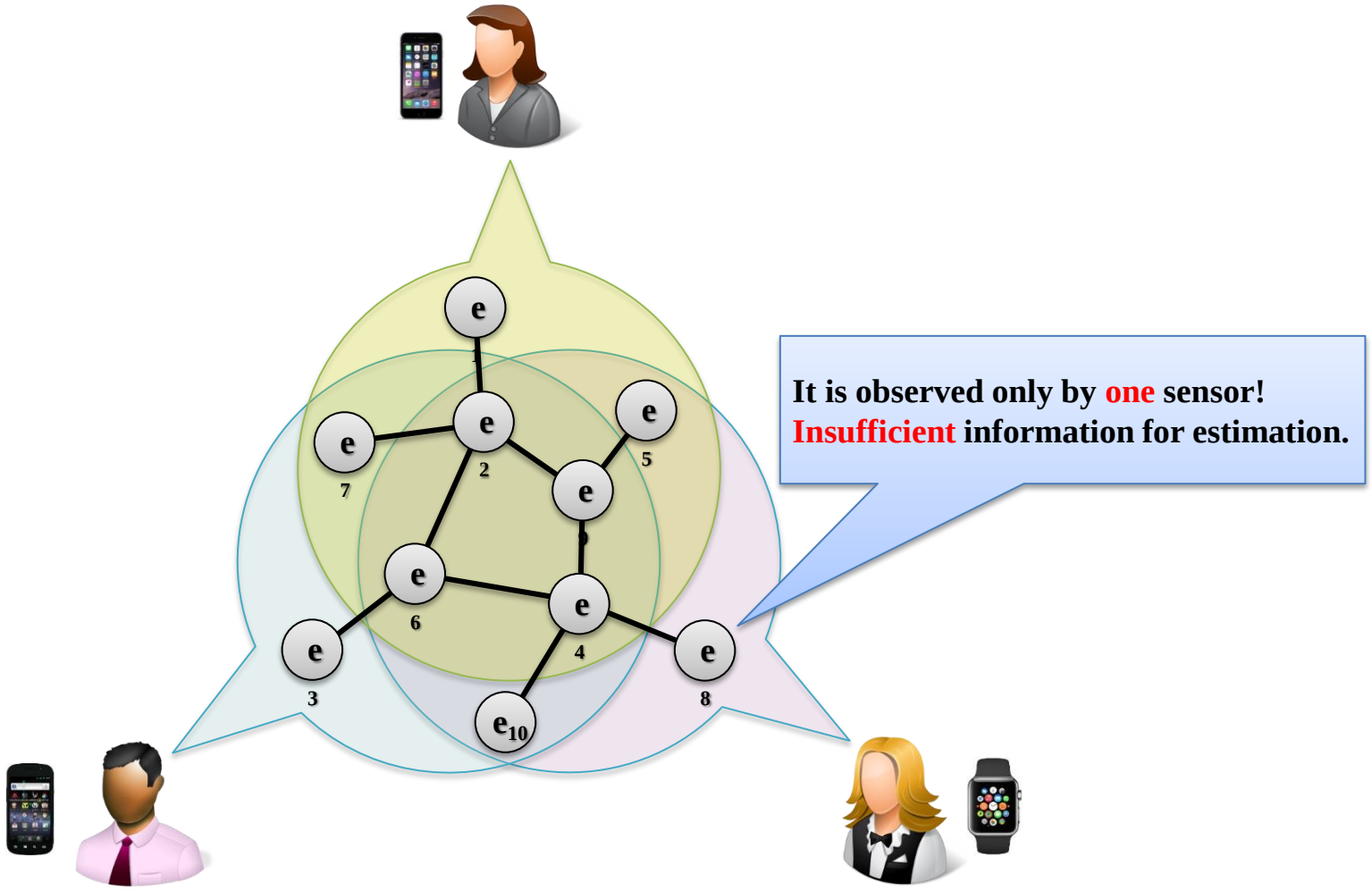


Human Sensor

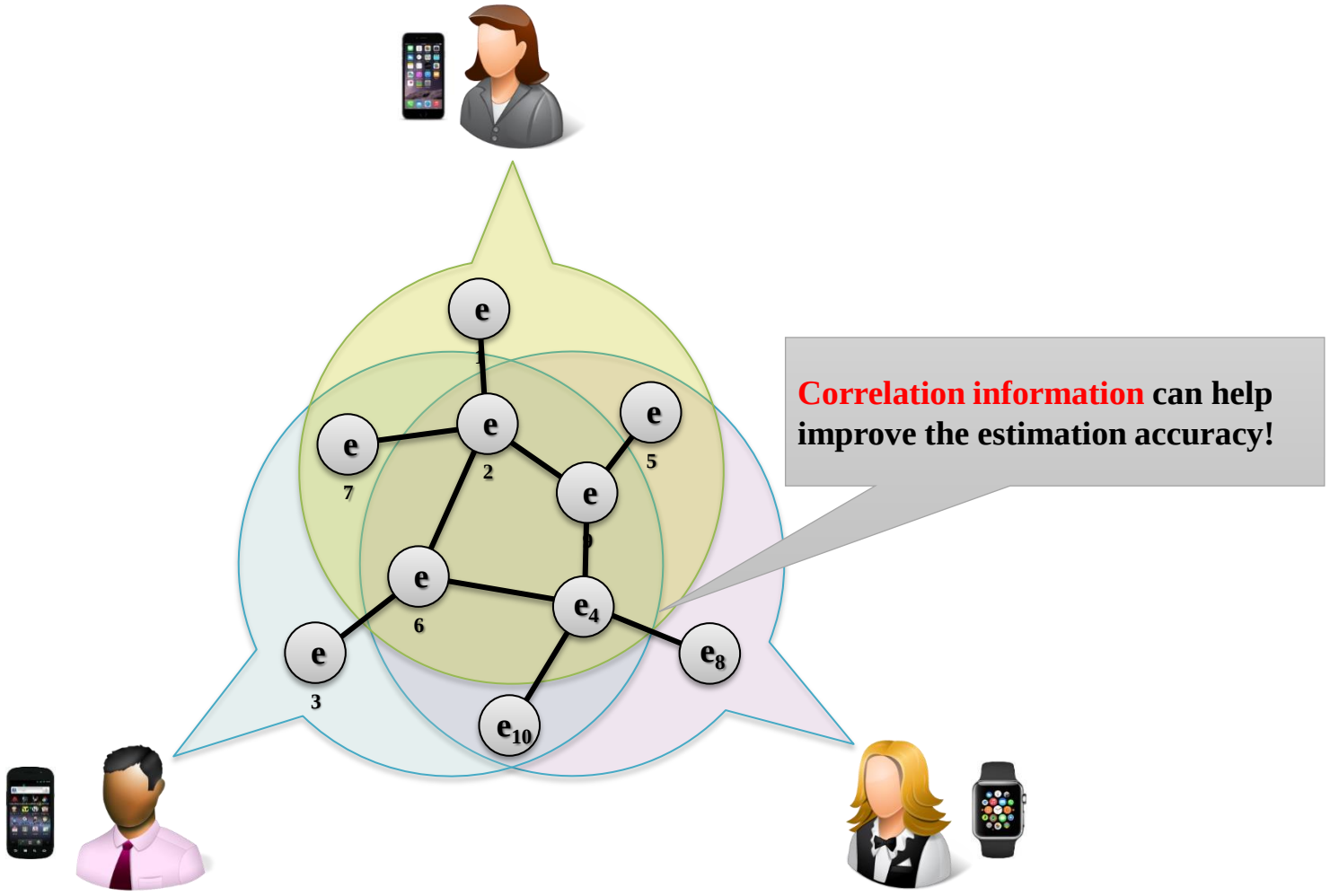

PM2.5 value?
e_1


250
198
e_1
275

Correlation Among Entities

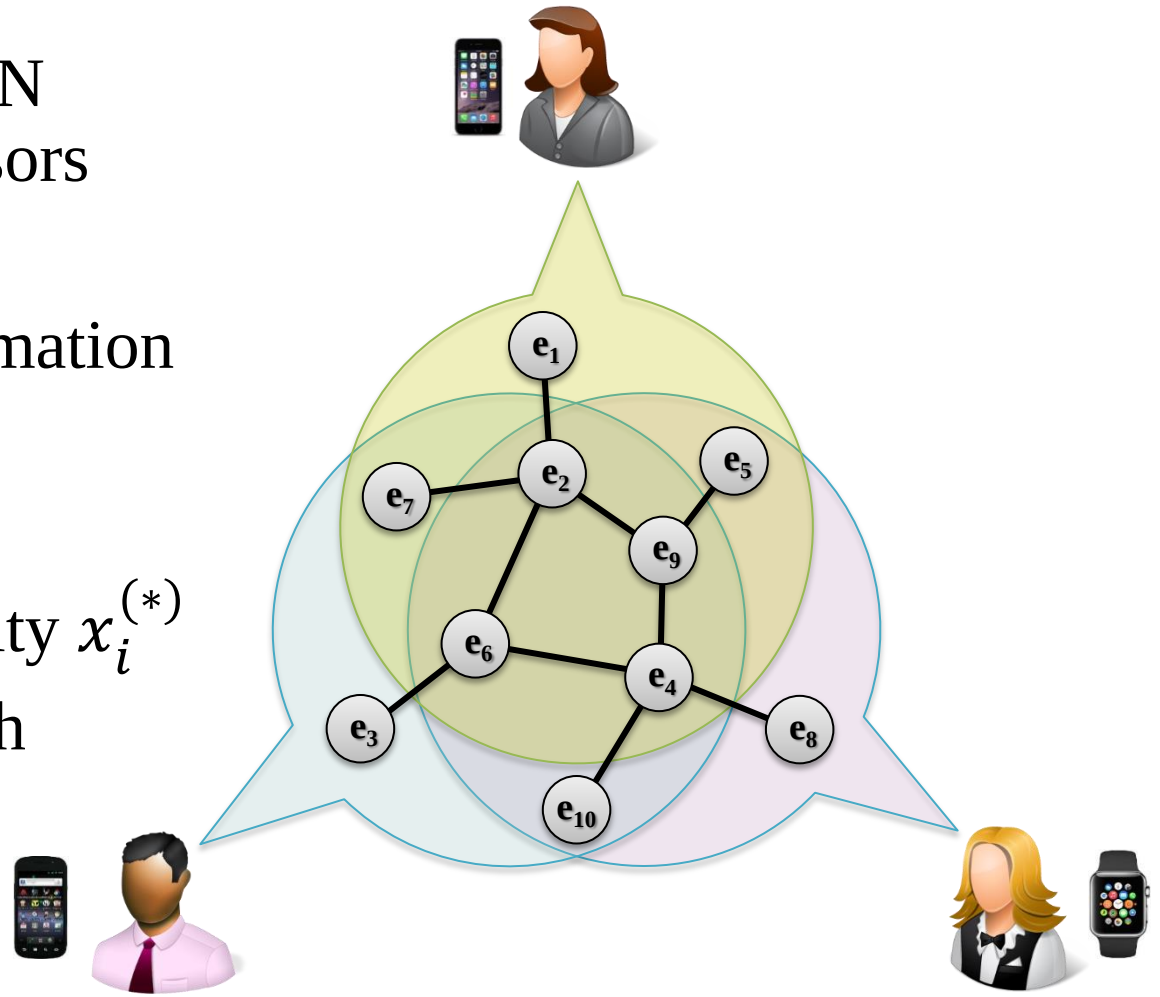


Correlation Among Entities



Correlated Entities – TD-corr

- Input:
  - Observations for N entities by K sensors $x_i^{(k)}$
  - Correlation information among entities
- Output:
  - Truth of each entity $x_i^{(*)}$
  - Reliability of each sensor w_k



[Meng et al., SenSys'15]

Correlated Entities – TD-corr

Variable 1: Sensor Weight
Reliability degree of the information provided by the sensor

Variable 2: Truth
True value of an entity

$$\min_{X^{(*)},W} f(X^{(*)},W) = \sum_{i=1}^{N}\left\{\sum_{k=1}^{K} w_k \left\| x_i^{(*)} - x_i^{(k)} \right\|^2 + \alpha \sum_{i' \in N(i)} S(i,i') \left\| x_{i'}^{(*)} - x_i^{(*)} \right\|^2\right\}$$

$$s.t. \sum_{k=1}^{K} \exp(-w_k) = 1$$

Constraint Function

Similarity Function
Similarity between correlated entities (e.g., Gaussian Kernel)

Regularization with correlation information

Correlated Entities – TD-corr

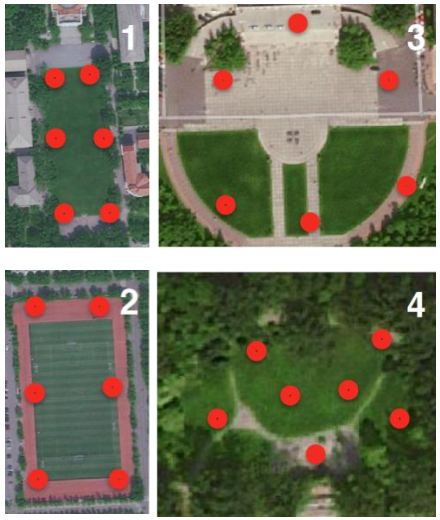
$$\min_{X^{(*)},W} f(X^{(*)},W) = \sum_{i=1}^{N} \left\{ \sum_{k=1}^{K} w_k \left\| x_i^{(*)} - x_i^{(k)} \right\|^2 + \alpha \sum_{i' \in N(i)} S(i,i') \left\| x_{i'}^{(*)} - x_i^{(*)} \right\|^2 \right\}$$

Partition entities into disjoint independent sets $\{I_1, I_2, \ldots, I_J\}$
(there are no correlations within the same set)

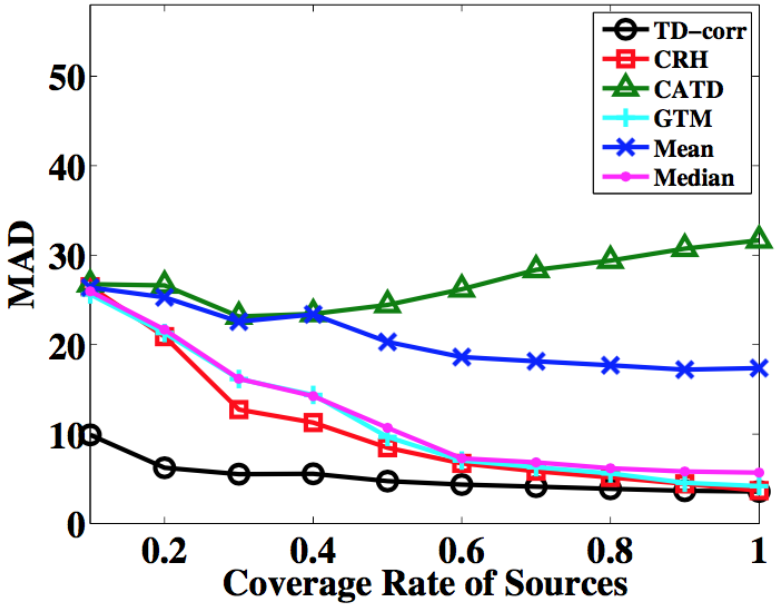
$$\min_{X^{(*)},W} f(X^{(*)},W) = \sum_{I_j \subset I} \sum_{i \in I_j} \left\{ \sum_{k=1}^{K} w_k \left\| x_i^{(*)} - x_i^{(k)} \right\|^2 + \alpha \sum_{i' \in N(i)} S(i,i') \left\| x_{i'}^{(*)} - x_i^{(*)} \right\|^2 \right\}$$

Experiments on Air Quality Sensing System

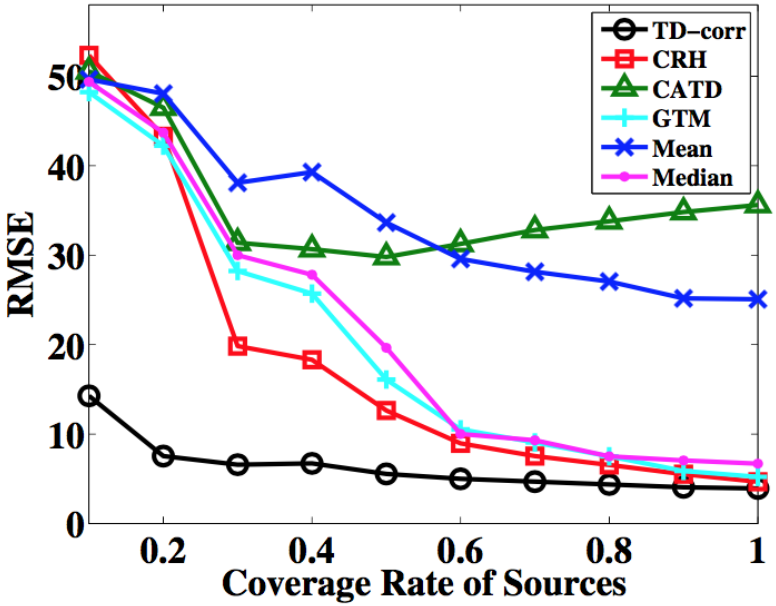
- Air Quality Sensing System
  - Monitor particulate matter with diameter less than 2.5 micron (PM2.5)
  - 14 participants equipped with mini-AQM
  - Ground truth is collected with Thermo
  - Conduct PM2.5 sensing in 4 areas in Tsinghua University



Correlated Entities – TD-corr



(a) Measured by MAD



(b) Measured by RMSE

The proposed method performs better especially when the coverage rates of sensors are low

Long-tail Phenomenon

- Challenge when most sources make a few claims
  - Sources weights are usually estimated as proportional to the accuracy of the sources
  - If long-tail phenomenon occurs, most source weights are not properly estimated.
- Challenge when most entities get a few claims
  - If an entity get very few claims, the estimation of the truth may not be accurate
- Confidence-aware approaches
  - considers the confidence interval of the estimation

Long-tail Phenomenon on Sources Side - CATD

- Assume that sources are independent and error made by source s: $\epsilon_s \sim N(0,\ \sigma_s^2)$
- $\epsilon_{aggregate} = \frac{\sum_{s\in S} w_s \epsilon_s}{\sum_{s\in S} w_s} \sim N\left(0, \frac{\sum_{s\in S} w_s^2 \sigma_s^2}{\left(\sum_{s\in S} w_s\right)^2}\right)$

Without loss of generality, we constrain $\sum_{s\in S} w_s = 1$

- **Optimization**

$$\min_{\{w_s\}} \quad \sum_{s\in\mathcal{S}} w_s^2 \sigma_s^2$$

$$\text{s.t.} \quad \sum_{s\in\mathcal{S}} w_s = 1,$$

$$w_s \geqslant 0, \forall s \in \mathcal{S}.$$

[Li et al., VLDB'15]

Long-tail Phenomenon on Sources Side - CATD

Sample variance is not accurate with small number of samples.

Find a range of values that can act as good estimates.

Calculate confidence interval based on

$$\frac{|N_s|\widehat{\sigma_s^2}}{\sigma_s^2} \sim \chi^2(|N_s|)$$

Long-tail Phenomenon on Sources Side - CATD

- Consider the possibly worst scenario of σ_s^2
- Use the upper bound of the 95% confidence interval of σ_s^2

$$u_s^2 = \frac{\sum_{n \in N_s} \left(x_n^s - x_n^{*(0)} \right)^2}{\chi^2_{(0.05,|N_s|)}}$$

Long-tail Phenomenon on Sources Side - CATD

$$\min_{\{w_s\}} \quad \sum_{s\in\mathcal{S}} w_s^2 u_s^2$$

$$\text{s.t.} \quad \sum_{s\in\mathcal{S}} w_s = 1, w_s \geqslant 0, \forall s \in \mathcal{S}.$$

- Closed-form solution:

$$w_s \propto \frac{1}{u_s^2} = \frac{\chi^2_{(0.05,|N_s|)}}{\sum_{n\in N_s}\left(x_n^s - x_n^{*(0)}\right)^2}$$

Long-tail Phenomenon on Sources Side - CATD

Example on calculating confidence interval

Source ID	# Claims	$\hat{\sigma_s^2}$	Confidence Interval (95%)
Source A	200	0.1	(0.0830, 0.1229)
Source B	200	3	(2.4890, 3.6871)
Source C	2	0.1	(0.0271, 3.9498)
Source D	2	3	(0.8133, 118.49)

Long-tail Phenomenon on Sources Side - CATD

Example on calculating source weight

Source ID	$\hat{\sigma}_s^2$	u_s^2	Source Weight (based on $\hat{\sigma}_s^2$)	Source Weight (based on u_s^2)
Source A	0.1	0.1229	0.4839	0.9385
Source B	3	3.6871	0.0161	0.0313
Source C	0.1	3.9498	0.4839	0.0292
Source D	3	118.49	0.0161	0.0010

Long-tail Phenomenon on Sources Side - CATD

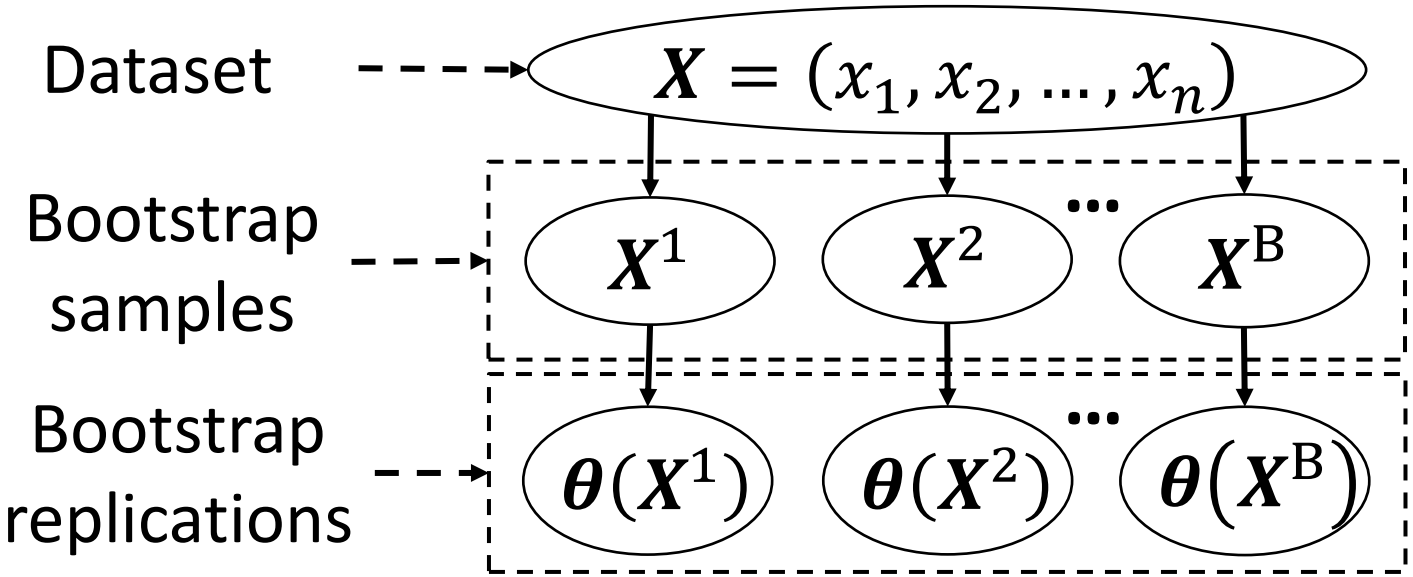
Game dataset

Question level	Error rate of Majority Voting	Error rate of CATD
1	0.0297	**0.0132**
2	0.0305	**0.0271**
3	0.0414	**0.0276**
4	0.0507	**0.0290**
5	0.0672	**0.0435**
6	0.1101	**0.0596**
7	0.1016	**0.0481**
8	0.3043	**0.1304**
9	0.3737	**0.1414**
10	0.5227	**0.2045**

Higher level indicates harder questions

Long-tail Phenomenon on Claim Side - ETCIBoot

- Provide estimation of confidence intervals (i.e., CI) for each entity's truth
- **Bootstrap**

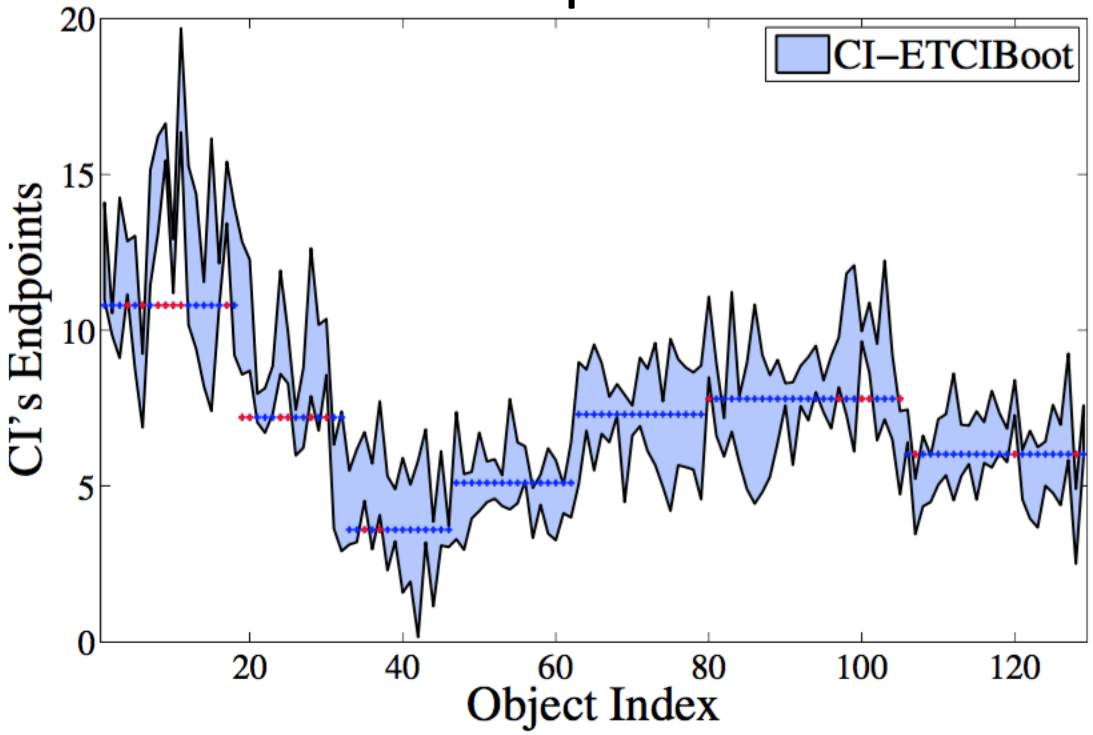


[Xiao et al., KDD'16]

Long-tail Phenomenon on Claim Side - ETCIBoot

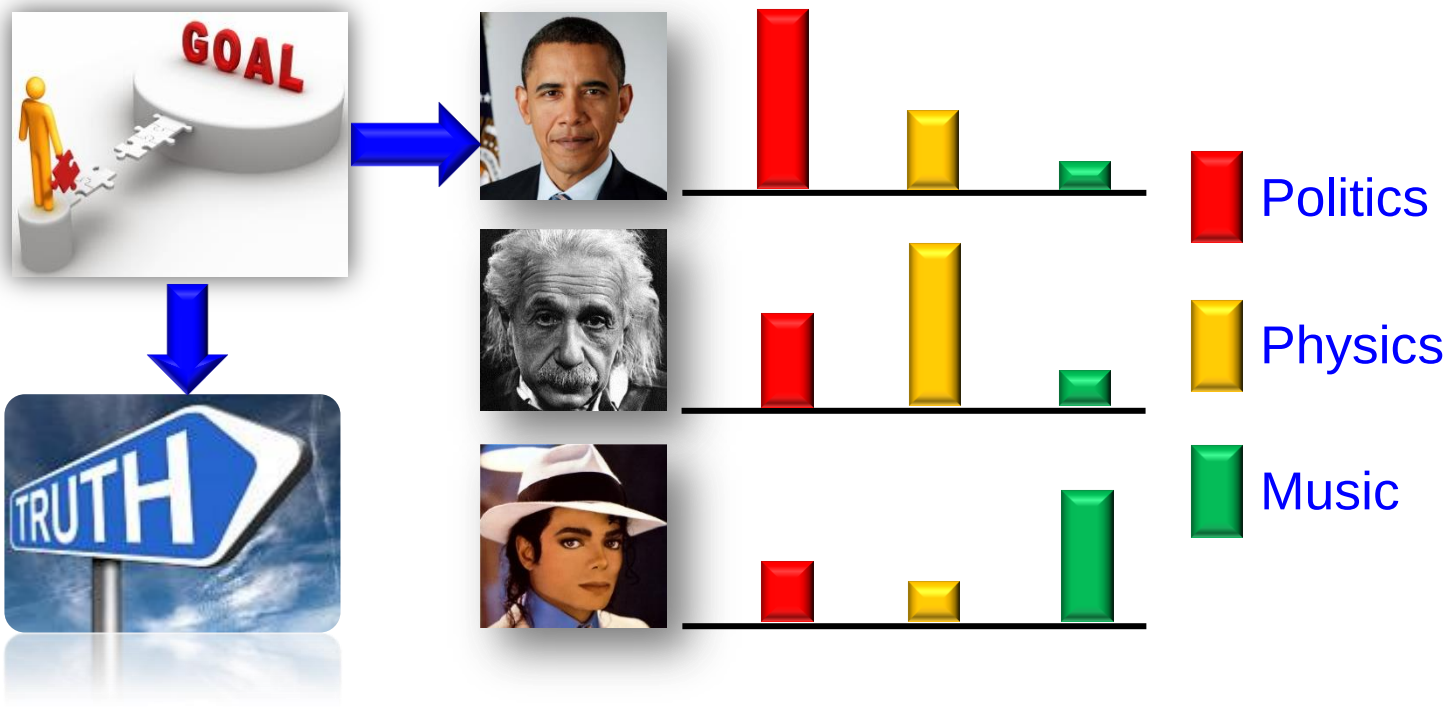
- Derive confidence intervals from bootstrap samples

Confidence Intervals obtained on indoor floorplan dataset



Fine-Grained Truth Discovery - FaitCrowd

- To learn **fine-grained (topical-level) user expertise** and the **truths** from conflicting crowd-contributed answers.
- Topic is learned from question&answer texts



[Ma et al., KDD'15]

Fine-Grained Truth Discovery - FaitCrowd

- **Input**
  - Question Set
  - User Set
  - Answer Set
  - Question Content
- **Output**
  - Questions' Topic
  - Topical-Level Users' Expertise
  - Truths

Question	User			Word	
	u1	u2	u3		
q1	1	2	1	a	b
q2	2	1	2	b	c
q3	1	2	2	a	c
q4	1	2	2	d	e
q5	2		1	e	f
q6	1	2	2	d	f

Topic	Question		
K1	q1	q2	q3
K2	q4	q5	q6

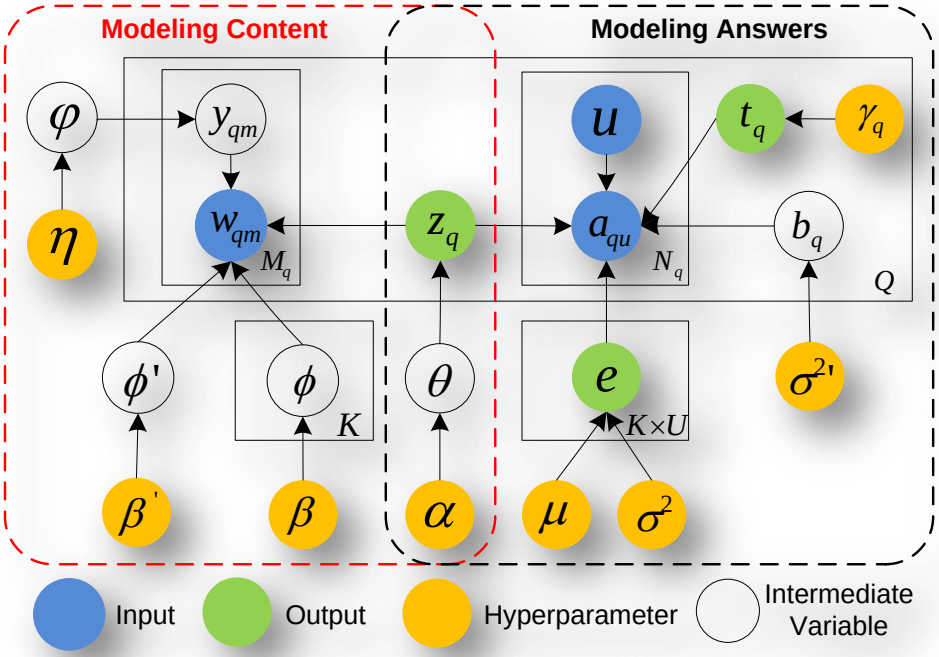
User		u1	u2	u3
Expertise	K1	2.34	2.70E-4	1.00
	K2	1.30E-4	2.34	2.35

Question	q1	q2	q3	q4	q5	q6
Truth	1	2	1	2	1	2

Question	q1	q2	q3	q4	q5	q6
Ground Truth	1	2	1	2	1	2

Fine-Grained Truth Discovery - FaitCrowd

- **Overview**

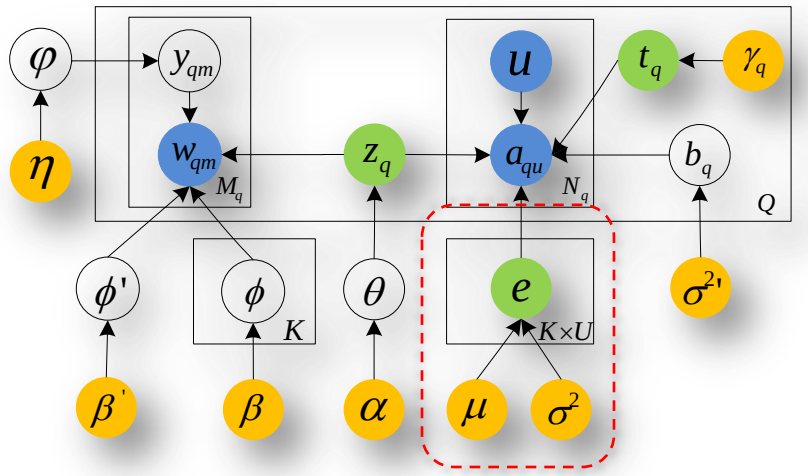


- Jointly modeling question content and users' answers by introducing latent topics.
- Modeling question content can help estimate reasonable user reliability, and in turn, modeling answers leads to the discovery of meaningful topics.
- Learning topics, topic-level user expertise and truths simultaneously.

Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise

$$e_{z_q u} \sim N(\mu, \sigma^2)$$



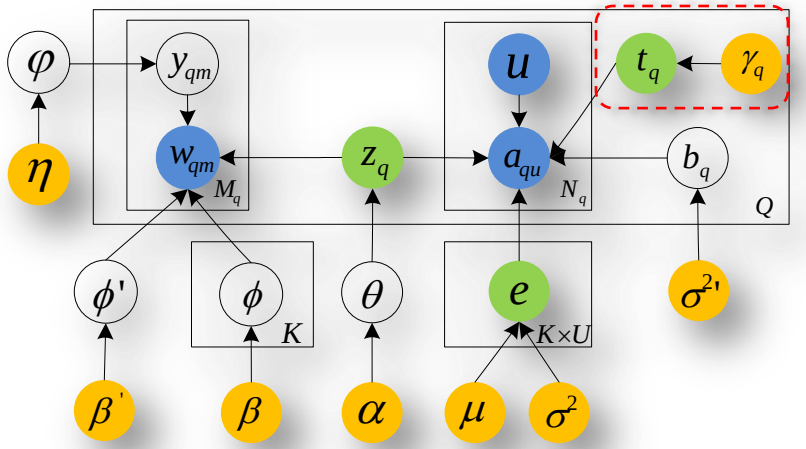
Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise

$$e_{z_q u} \sim N(\mu, \sigma^2)$$

    - Draw the truth

$$t_q \sim U(\gamma_q)$$



Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise

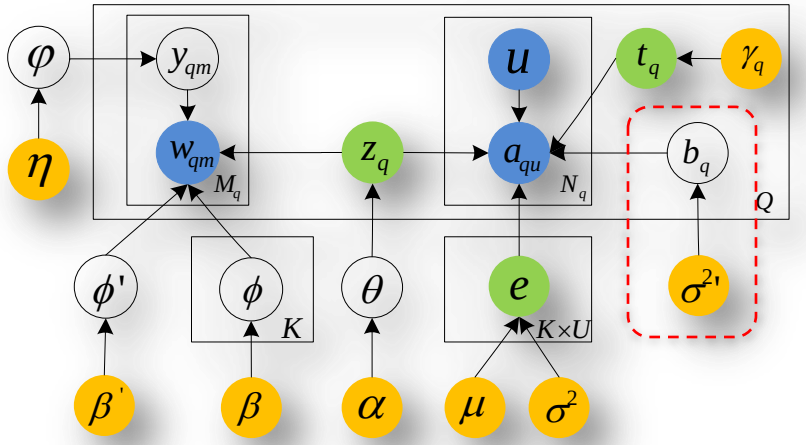
$$e_{z_q u} \sim N(\mu, \sigma^2)$$

    - Draw the truth

$$t_q \sim U(\gamma_q)$$

    - Draw the bias

$$b_q \sim N(0, \sigma^{2'})$$



Fine-Grained Truth Discovery - FaitCrowd

- **Answer Generation**
  - The correctness of a user's answer may be affected by the question's topic, user's expertise on the topic and the question's bias.
    - Draw user's expertise

$$e_{z_q u} \sim N(\mu, \sigma^2)$$

    - Draw the truth

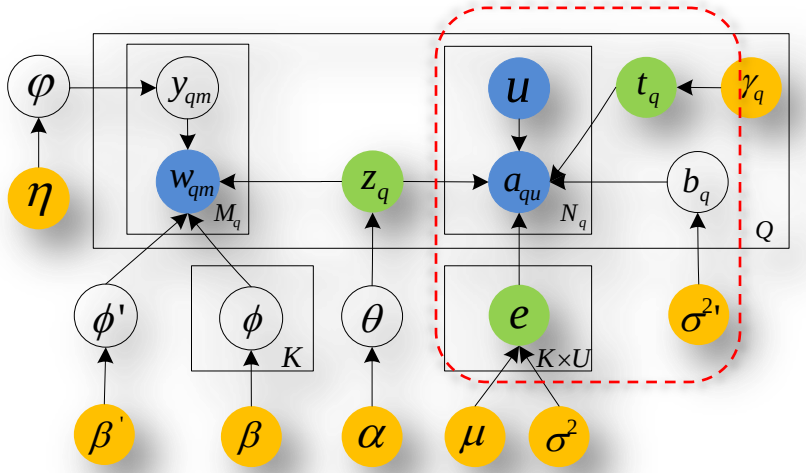
$$t_q \sim U(\gamma_q)$$

    - Draw the bias

$$b_q \sim N(0, \sigma^{2'})$$

    - Draw a user's answer

$$a_{qu}|t_q \sim logistic(e_{z_q u}, b_q)$$



$$e_{z_q u} \uparrow \text{ and } b_q \downarrow \longrightarrow p(a_{qu} = t_q | t_q) \uparrow$$
$$e_{z_q u} \downarrow \text{ and } b_q \uparrow \longrightarrow p(a_{qu} = t_q | t_q) \downarrow$$

Fine-Grained Truth Discovery - FaitCrowd

Game dataset

Question level	Majority Voting	CATD	FaitCrowd
1	0.0297	0.0132	**0.0132**
2	0.0305	0.0271	**0.0271**
3	0.0414	0.0276	**0.0241**
4	0.0507	0.0290	**0.0254**
5	0.0672	0.0435	**0.0395**
6	0.1101	0.0596	**0.0550**
7	0.1016	0.0481	**0.0481**
8	0.3043	0.1304	**0.0870**
9	0.3737	0.1414	**0.1010**
10	0.5227	0.2045	**0.1136**

Overview

1. Introduction of Crowdsourcing
2. Crowdsourced Data Aggregation for Reliable Information Discovery
3. Passive Crowdsourcing Scenarios
4. **Active Crowdsourcing Scenarios**
5. Applications and Open Questions
6. Resources
7. References

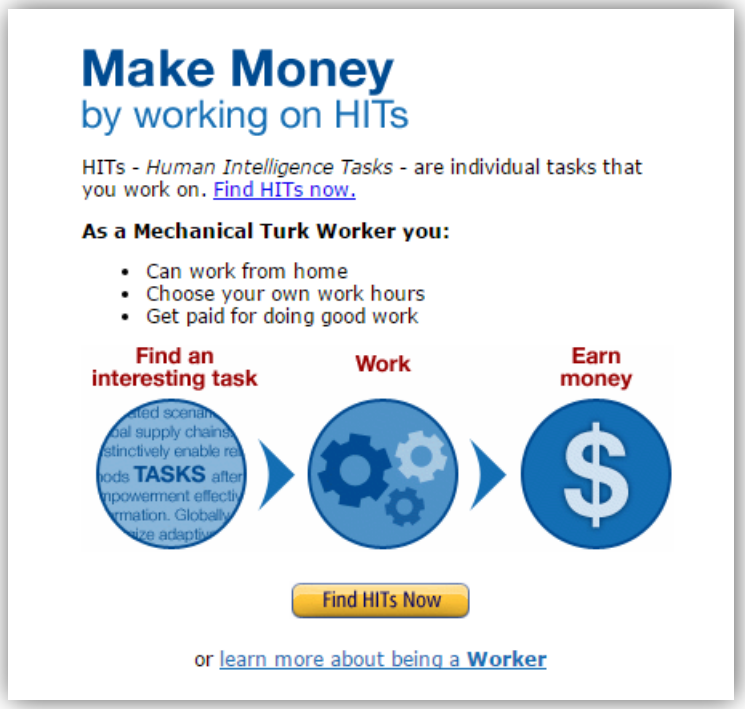
Active Crowdsourcing

amazon mechanical turk beta — Artificial Artificial Intelligence

CrowdFlower

• • •

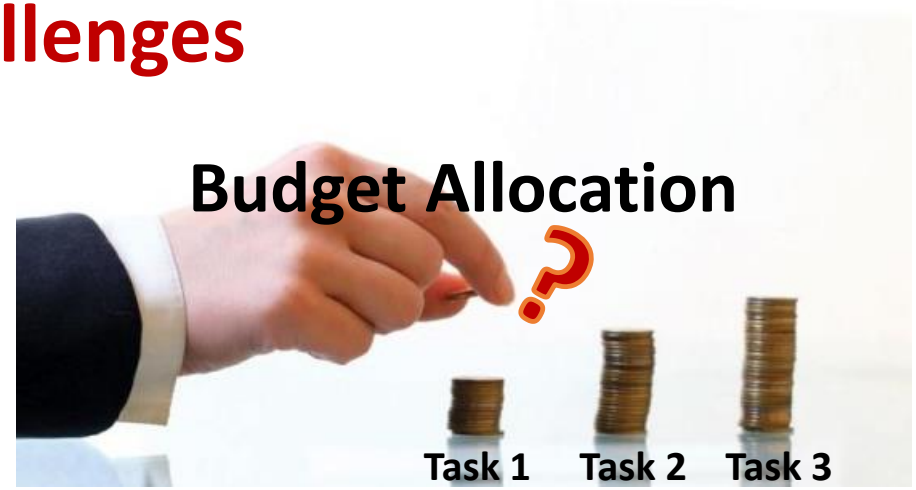
requester



worker

Active Crowdsourcing Scenarios

- **Challenges**



Budget Allocation

- Since active crowdsourcing costs money, we need to use the budget wisely.
- Budget allocation
  - Which instance should we query for labels?
  - Which worker should we choose for a certain task?
- Goal
  - To maximize utility (eg. overall accuracy)

Maximize Accuracy – Opt-KG

- Need to estimate the labeling ambiguity for each instance on the fly
- Intuition:
  - avoid spending much budget on fairly easy instances
  - avoid spending much budget on few highly ambiguous instances
- Ideally
  - put those few highly ambiguous instances aside to save budget
  - estimate the reliability of each worker on the fly
  - allocate as many labeling tasks to reliable workers as possible

[Chen et al., ICML'13]

Problem Settings

- N independent binary instances
- True label $Z_i \in \{+1, -1\}$
- Instance difficulty: $\theta_i = P(Z_i = +1)$
  - relative frequency of +1 appears when the number of workers approaches infinity
  - $P(Z_i = +1) \approx 0.5$ means the instance is hard
- Workers are noiseless (for basic model)
  - $P(y_{ij} = +1) = \theta_i$, where y_{ij} is worker j's label for instance i
  - Labels for instance i are i.i.d. from Bernoulli(θ_i)

Bayesian setting

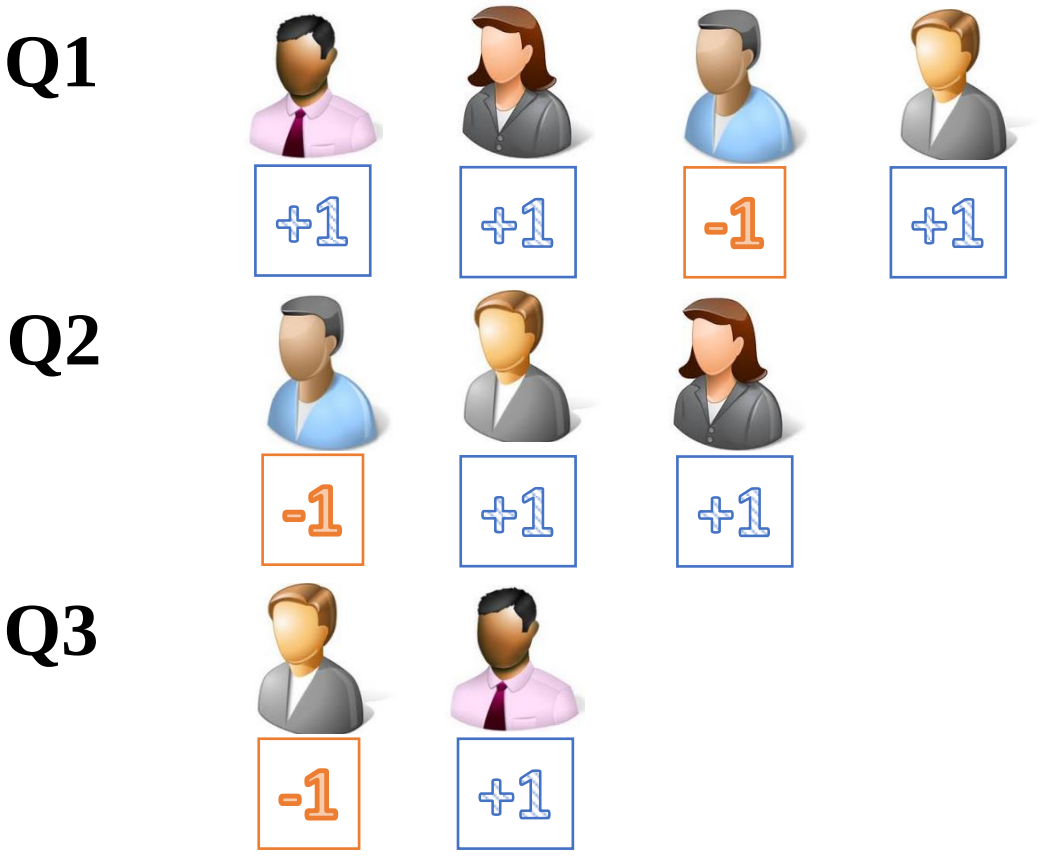
- θ_i is drawn from a known Beta prior distribution Beta(a_i^0, b_i^0)
- It means we have a_i^0 positive and b_i^0 negative pseudo-labels for the i-th instance at the initial stage
- Posterior:
  - $\text{Beta}\left(a_{i_t}^{t+1}, b_{i_t}^{t+1}\right) = \begin{cases} \text{Beta}(a_{i_t}^t + 1, b_{i_t}^t), \text{ if } y_{i_t} = 1 \\ \text{Beta}(a_{i_t}^t, b_{i_t}^t + 1), \text{ if } y_{i_t} = -1 \end{cases}$

Maximize Accuracy – Opt-KG

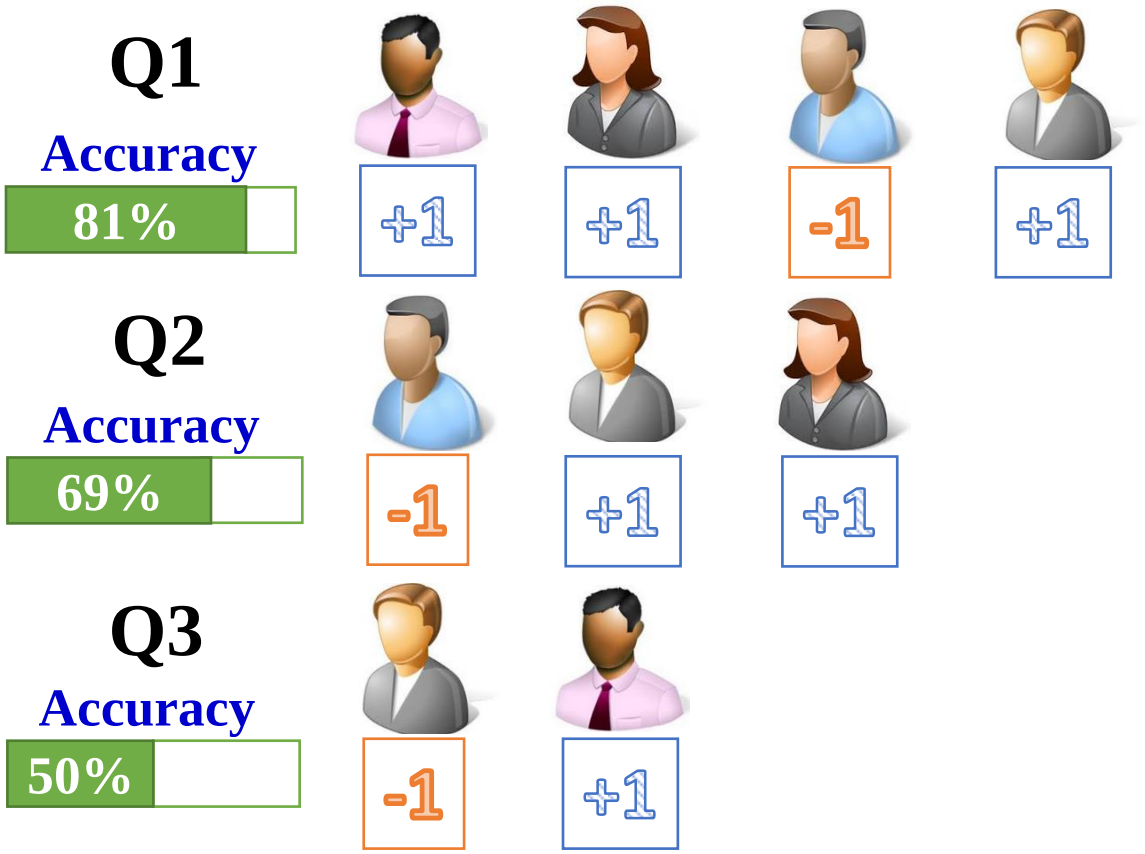
- Formally, maximizes the expected accuracy taken over the sample paths $\left(i_0, y_{i_0}, \ldots, i_{T-1}, y_{i_{T-1}}\right)$ generated by a policy π
- Stage-wise Rewards:
  - Get label $+1$: $R_{it}^{+1}(a,b) = h\big(I(a+1,b)\big) - h(I(a,b))$
  - Get label -1: $R_{it}^{-1}(a,b) = h\big(I(a,b+1)\big) - h(I(a,b))$
  - Where $h(x) = \max(x, 1-x)$,
    $I(a,b)$ is the cdf of Beta(a,b) at $x = 0.5$
- Greedy strategy

$$R(S^t, i^t) = \max(R_{it}^{+1}, R_{it}^{-1})$$

Maximize Accuracy – Opt-KG

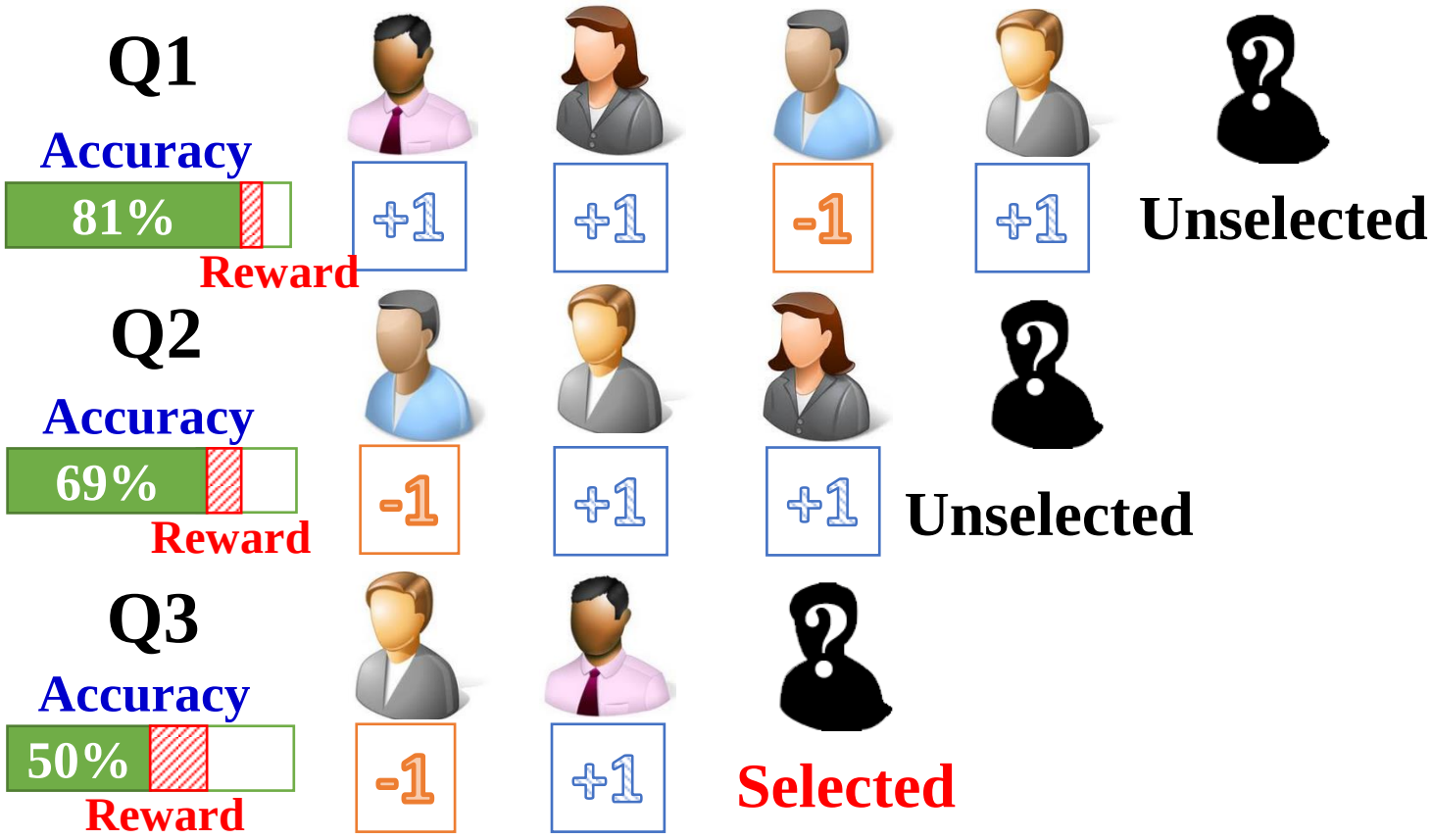


Maximize Accuracy – Opt-KG



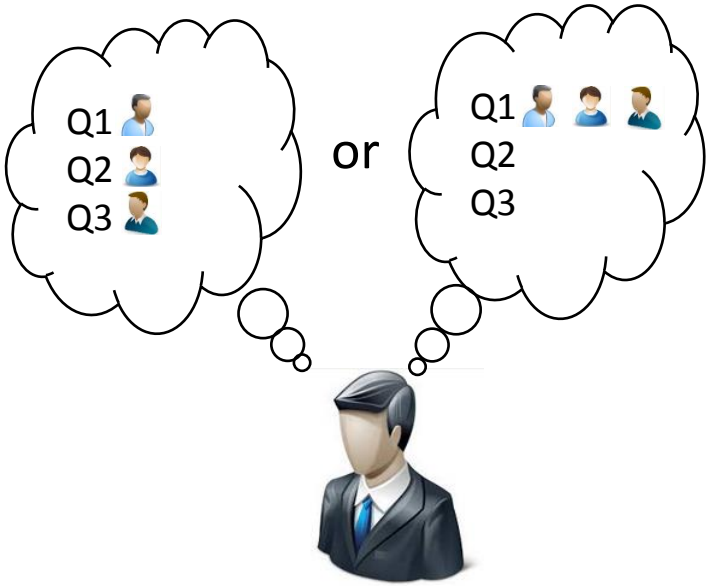
Maximize Accuracy – Opt-KG

Maximize Accuracy – Opt-KG

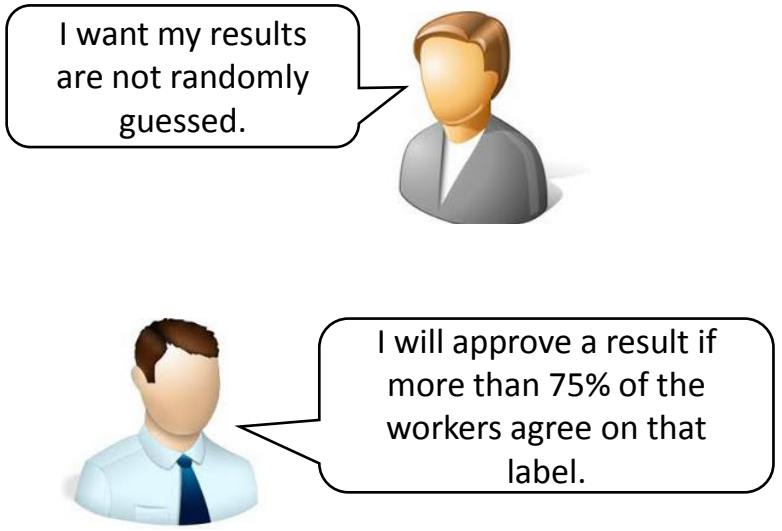


Challenges Under a Tight Budget

Quantity and Quality Trade-off



Different Requirements of Quality



[Li et al., WSDM'16]

Maximize Quantity – Requallo

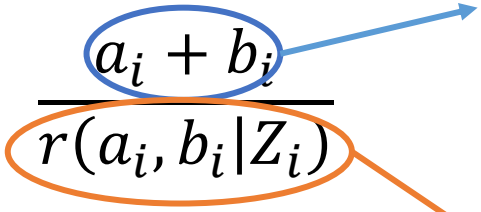
- Inputs
  - Requester's requirement
  - The budget
    - T: the maximum amount of labels can be afforded
- Goal
  - Label as many instances as possible which achieve the requirement under the budget

Examples of Requirement

- Minimum ratio
  - Approve the result on an instance if $a_i : b_i \geq c$ or $b_i : a_i \geq c$
  - Equivalent to set a threshold on entropy
- Hypothesis test
  - Fisher exact test to test if the labels are randomly guessed
  - Calculate the p-value, and approve the result if $p\text{−value} < \alpha$

Completeness

- Ratio between the observed total vote counts and the minimum count of labels it needs to achieve the requirement.
- Denoted as:



Observed total vote counts

Minimum count to achieve the requirement

Completeness

- Ratio between the observed total vote counts and the minimum count of labels it needs to achieve the requirement.
- Example:
  - $a_i = 3, b_i = 1$, requirement is the minimum ratio of 4
  - If $Z_i = +1$, completeness=$\frac{3+1}{4+1} = \frac{4}{5}$
  - If $Z_i = -1$, completeness=$\frac{3+1}{3+12} = \frac{4}{15}$

Expected Completeness

$$V_i(a_i, b_i)$$
$$= P(Z_i = +1 | a_i, b_i) \frac{a_i + b_i}{r(b_i)}$$
$$+ P(Z_i = -1 | a_i, b_i) \frac{a_i + b_i}{r(a_i)}$$

Completeness given that the true label is $+1$

Completeness given that the true label is -1

where

$$r(b_i) = r(a_i, b_i | Z_i = +1),$$
$$r(a_i) = r(a_i, b_i | Z_i = -1)$$

Maximize Quantity – Requallo

- The goal is to label instances as many as possible that achieve the requirement of quality.
- Stage-wise reward

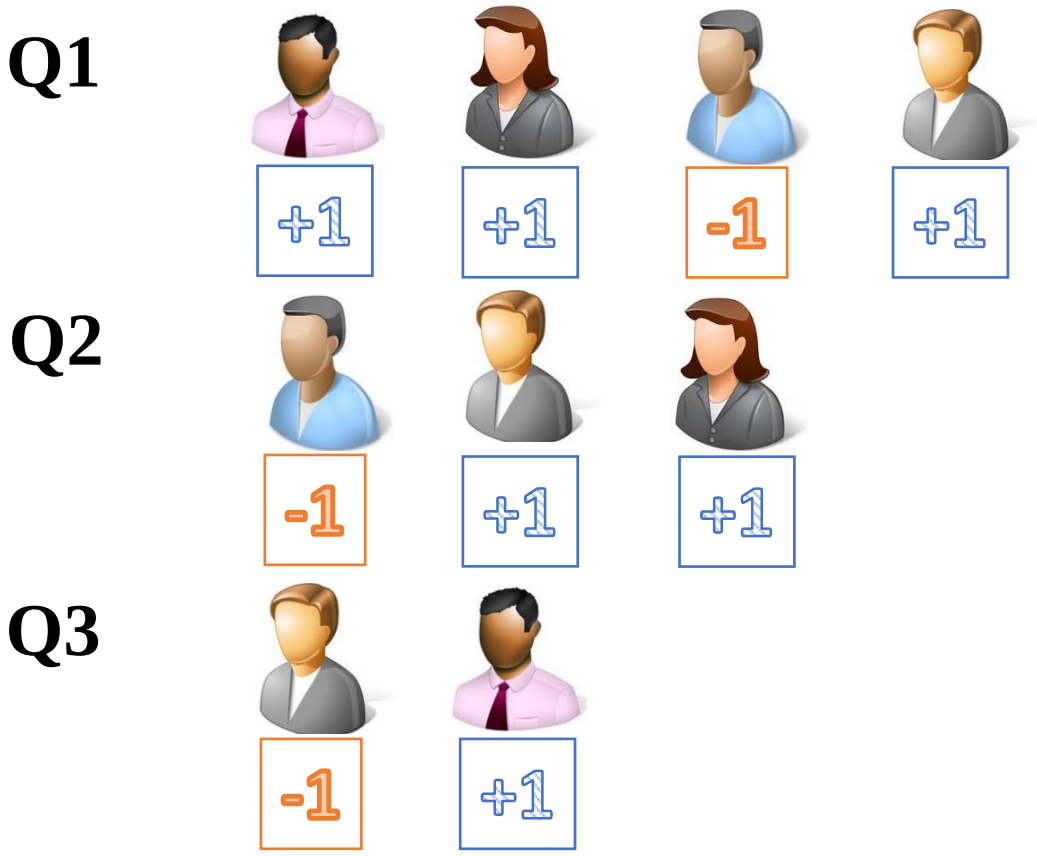
$$R_{i^t}^{+1} = V_{i^t}\left(a_{i^t}^t + 1, b_{i^t}^t\right) - V_{i^t}\left(a_{i^t}^t, b_{i^t}^t\right)$$
$$R_{i^t}^{-1} = V_{i^t}\left(a_{i^t}^t, b_{i^t}^t + 1\right) - V_{i^t}\left(a_{i^t}^t, b_{i^t}^t\right)$$

- Greedy strategy

$$R(S^t, i^t) = \max(R_{i^t}^{+1}, R_{i^t}^{-1})$$

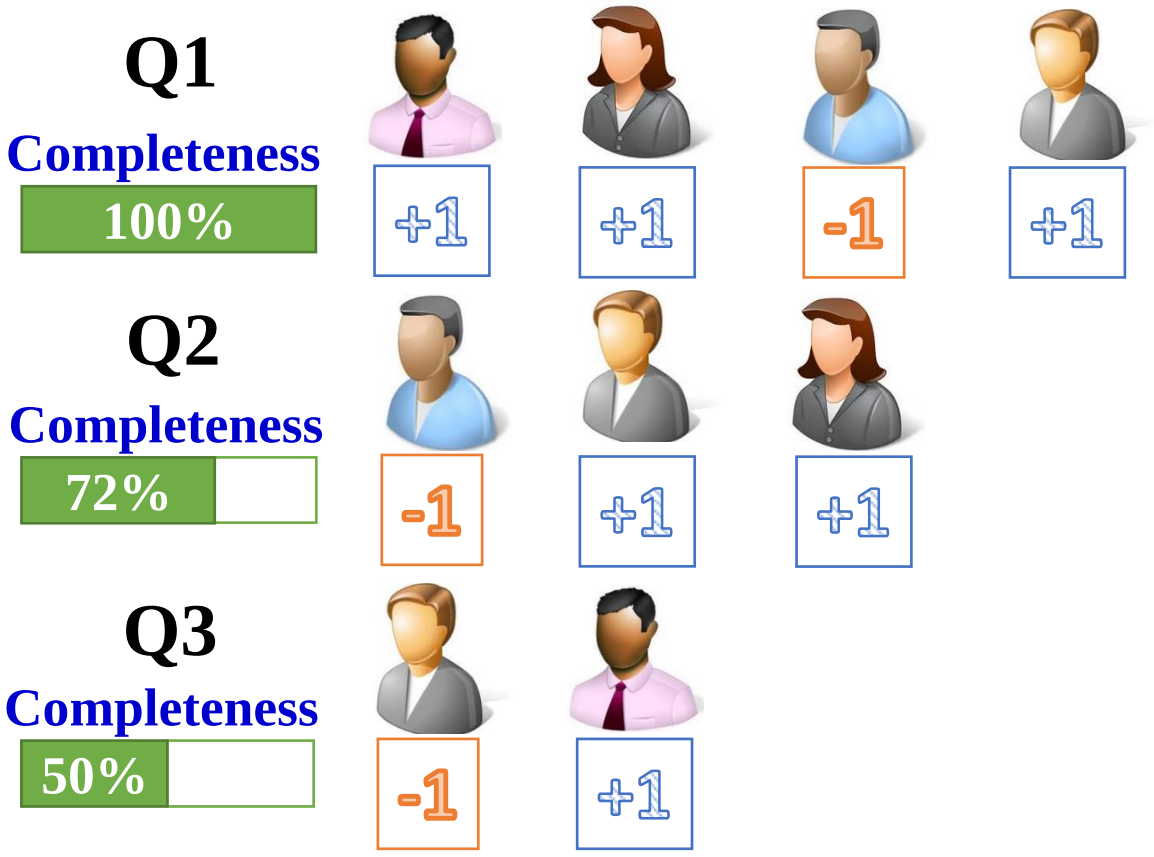
Maximize Quantity – Requallo

Requirement: Minimum Ratio of 3



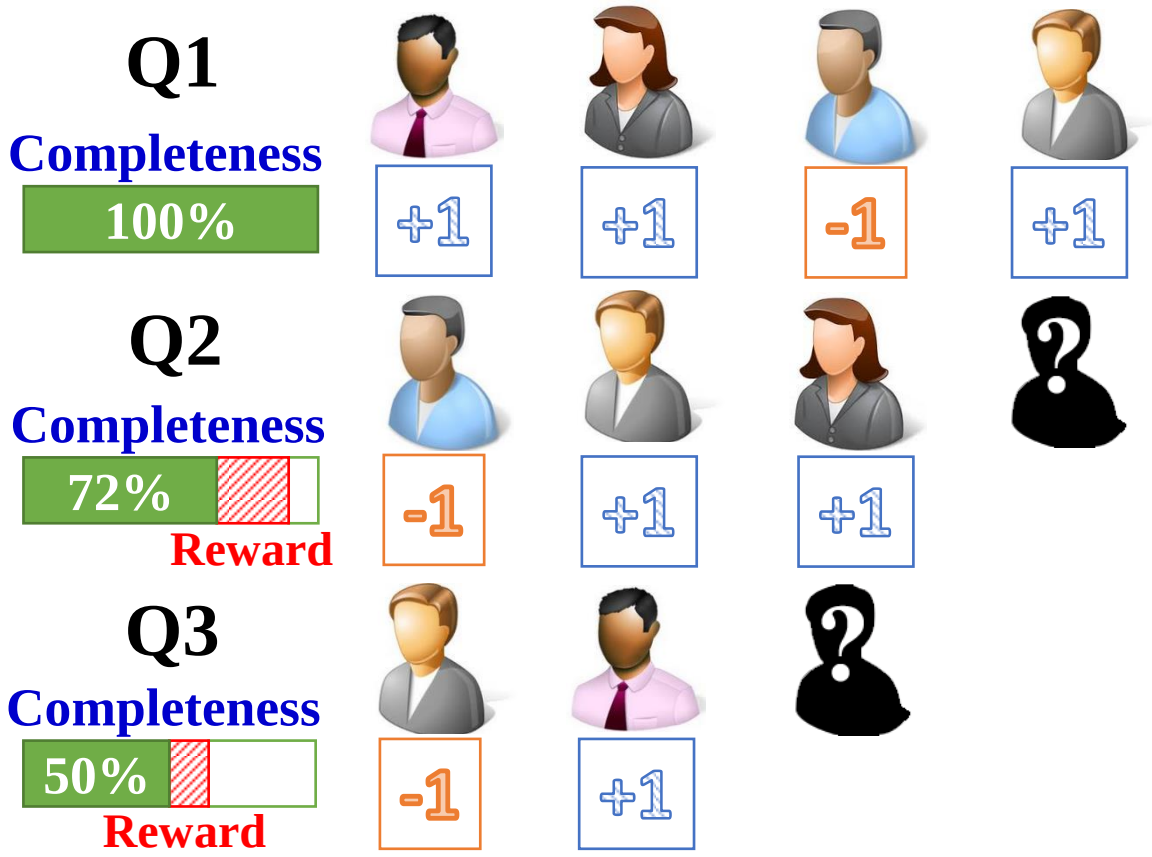
Maximize Quantity – Requallo

Requirement: Minimum Ratio of 3

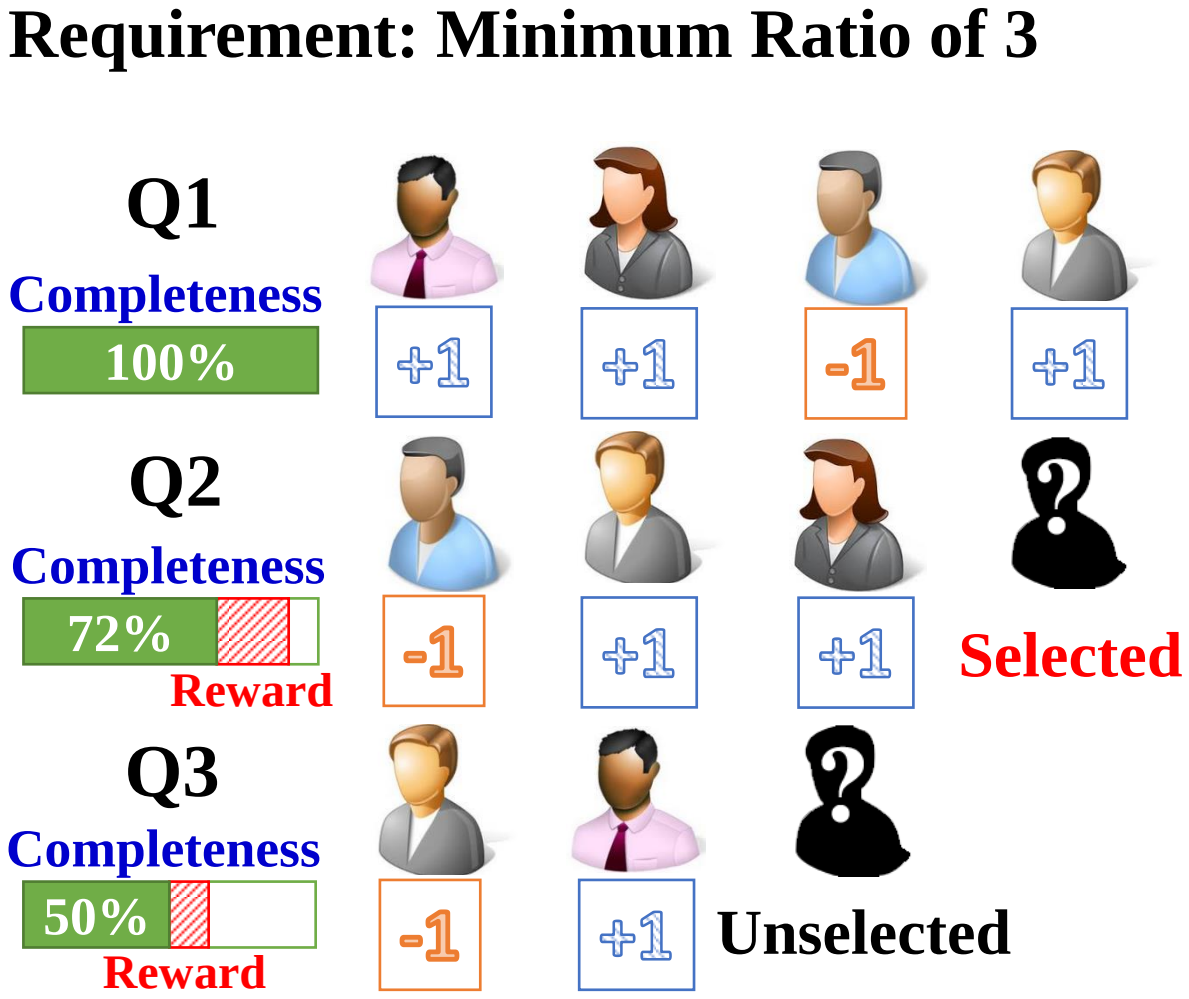


Maximize Quantity – Requallo

Requirement: Minimum Ratio of 3



Maximize Quantity – Requallo



Crowdsourcing for Machine Learning

- Crowdsourced labels for machine learning
  - Labeling by machine can save more money
  - Pros: labeling is cheap
  - Cons: workers are noisy
  - Solution: reduce noise in annotations

Crowdsourcing for Active Learning

- Active learning
  - Motivation: budget
  - Goal: query as few instances as possible to train a good classifier
  - Which to query? The most "informative" instances
    - Uncertainty, density, influence,…
- Active learning with crowdsourced labels
  - Workers are weak oracles
  - Instances can be queried multiple times
  - Which to query? How to query?

Active Learning with Crowdsourced Labels

- Strategy 1: Query the "best" worker [Yan et al., ICML'11]
- Strategy 2: Repeat labeling [Mozafari et al., VLDB'14]
  - Once an instance is queried, query multiple workers
- Strategy 3: Joint design
  - Jointly consider model uncertainty and label uncertainty
  - Model uncertainty × label uncertainty
    [Sheng et al., KDD'08]
  - Model uncertainty + label uncertainty
    [Zhao et al., PASSAT'11]

Incentive Mechanism

- Goal:
  - Design payment mechanisms to incentivize workers
  - A win-win strategy
- For requesters
  - Get the optimal utility (eg. quality, profit, etc) for their expense
- For workers
  - Get maximal payment if they follow the rules

Incentive Mechanism – Double or Nothing

- Incentive compatibility
  - To encourage the worker to skip the questions about which she is unsure
  - Reason: for the questions that a worker is not sure of, her answers could be very unreliable
- No-free-lunch
  - If all the questions attempted by the worker are answered incorrectly, then the payment must be zero

[Shah&Zhou, NIPS'15]

Incentive Mechanism – Double or Nothing

- Input
  - Confidence thresho
  - Budget μ
  - Evaluations $(x_1, \ldots, x_G) \in \{-1, 0, +1\}^G$ of the worker's answers to the G gold standard questions

Wrong Skipped Correct

- Let $C = \sum_{i=1}^{G} \mathbf{1}\{x_i = +1\}$ and $W = \sum_{i=1}^{G} \mathbf{1}\{x_i = -1\}$

Number of correct answers

Number of wrong answers

- Payment:

$$\mu T^{G-C} \mathbf{1}\{W = 0\}$$

Example

- Input
  - Confidence threshold $T = \frac{1}{2}$
  - Budget $\mu = 80$ cents
  - Number of gold standard questions $G = 3$
- Incentive mechanism description

  The reward starts at 10 cents. For every correct answer in the 3 gold standard questions, the reward will double. However, if any of these questions are answered incorrectly, then the reward will become zero. So please use the "I'm not sure" option wisely.

Incentive Mechanism – Double or Nothing

- Analysis
  - This payment mechanism is incentive-compatible and satisfies the no-free-lunch condition
  - This payment mechanism is the only incentive-compatible mechanism that satisfies the no-free-lunch condition
- Optimality against spamming behavior
  - This payment mechanism minimizes the expected payment to a worker who answers all questions uniformly at random

Incentive Mechanism in Crowd Sensing

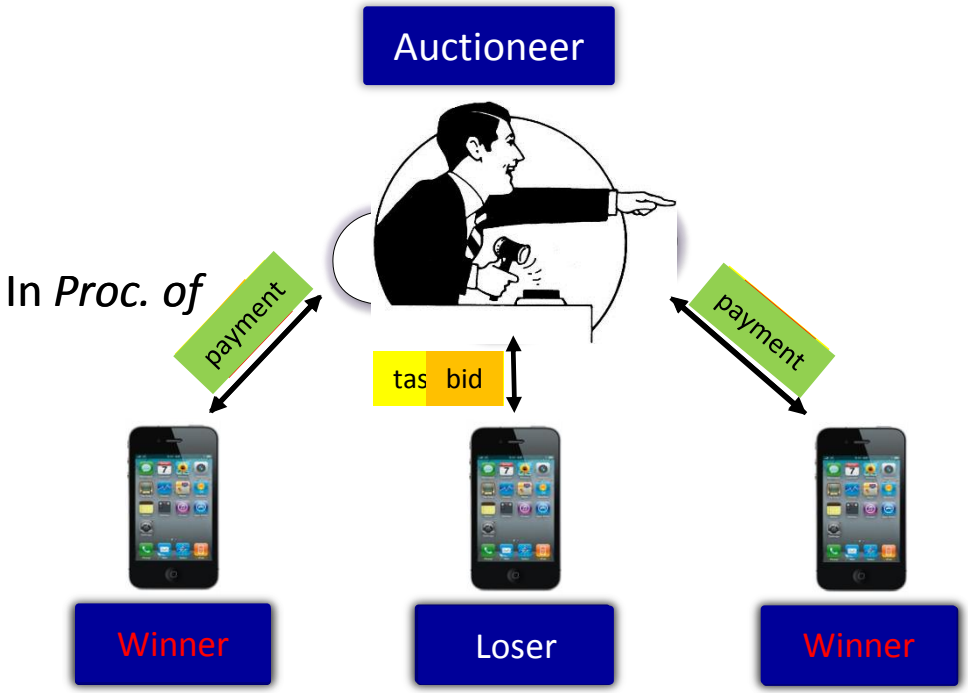
The platform acts as the auctioneer and workers act as bidders.

The platform announces the set of sensing tasks, $\mathcal{T} = \{\tau_1, \cdots, \tau_M\}$.

Every worker i submits a bid tuple (Γ_i, b_i) to the platform. (Γ_i: bidding bundle[1], b_i: bidding price).

The platform determines the set of winners $\mathcal{S}$ and the payment profile $\mathbf{p}$ to all winners.

Winners send sensory data to the platform and the platform sends payments to winners.



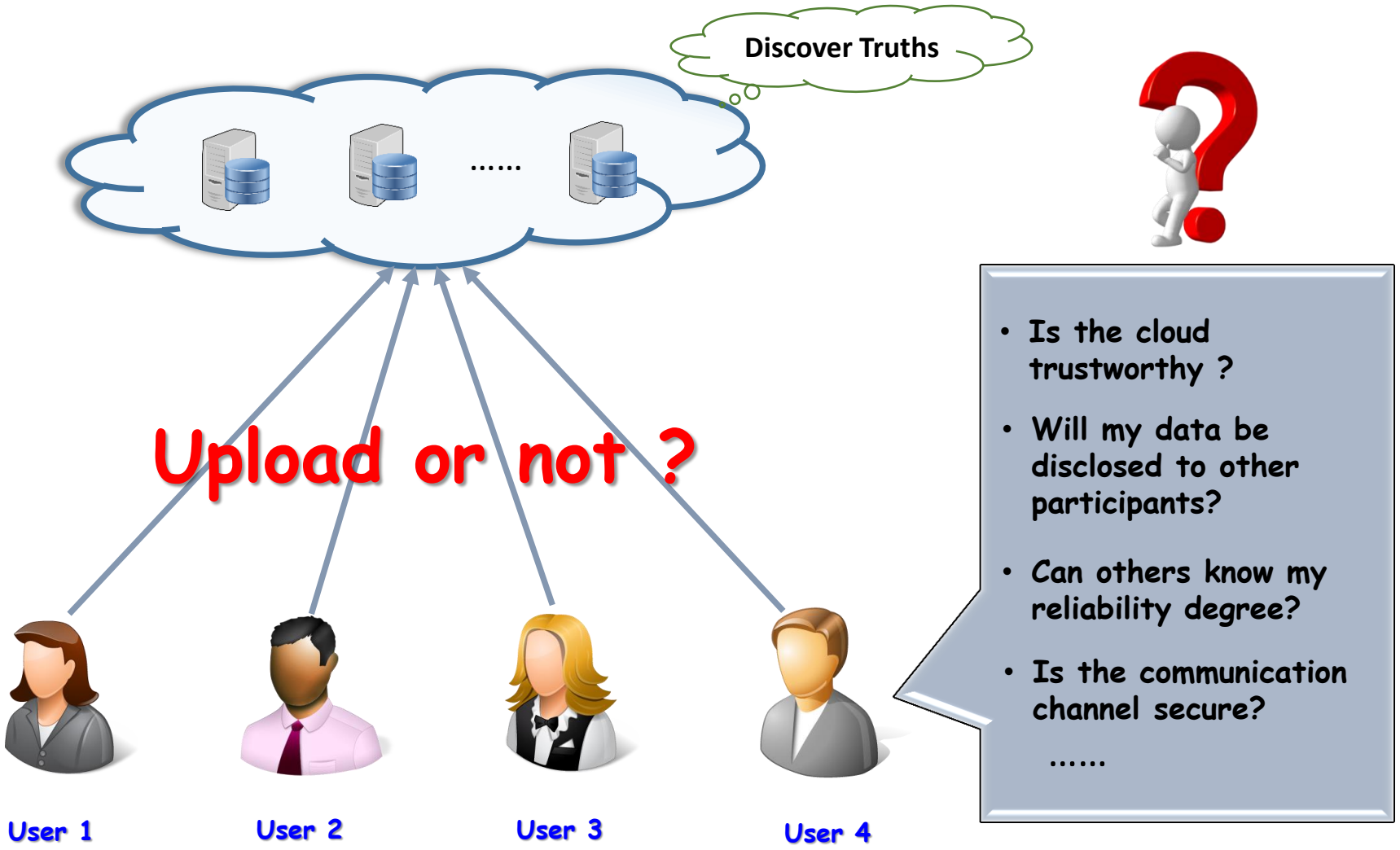
[Jin et al., MobiHoc'15]

1. Bundle refers to a set of tasks.

Incentive Mechanism in Crowd Sensing

- Game theory based design
- Analysis
  - With proper functions
  - This auction is individual rational
    - A mechanism is individual rational if and only if the user's utility (payment – cost) is non-negative is satisfied for every user
  - This auction is truthful
    - Truthfulness means that each worker submits to the platform his truly interested tasks, and a bidding price equal to his true cost for executing these tasks

Privacy Concerns



Privacy Concerns

- Sensitive personal information
  - Health data of patients
  - Locations of participants
  - Answers for special questions
  - ……
- User's reliability degree is also sensitive
  - Inferring personal information
  - Maliciously manipulating data price

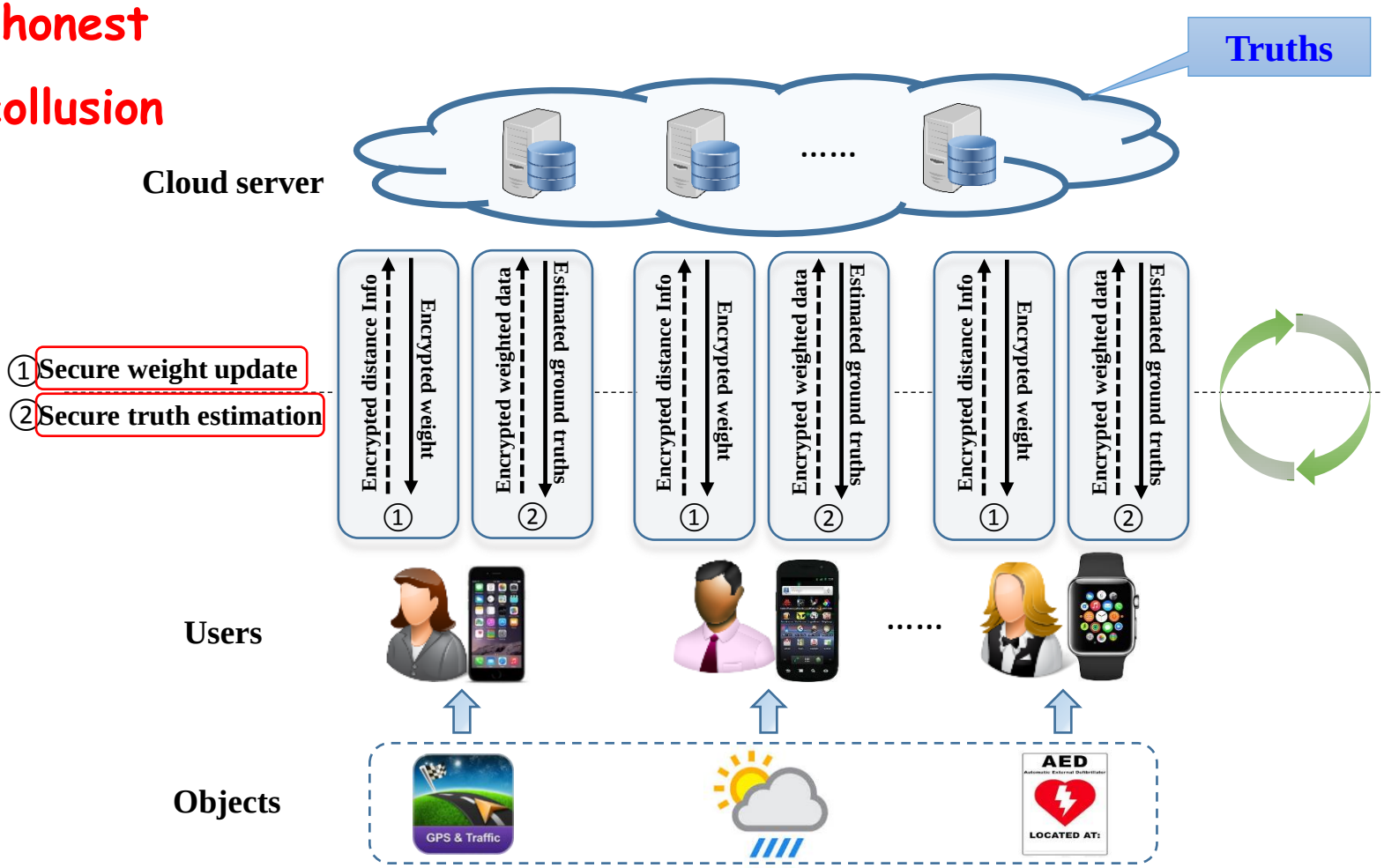
Problem Setting

- Worker-private label aggregation problem
  - Input: Crowd labels $\{y_{ij}\}_{i=1,j=1}^{N,M}$
  - Output: Estimated true labels $\{y_i\}_{i=1}^{N}$
  - Subject to: labels and reliabilities are kept worker-private
- Worker-private
  - Worker j's w_j is worker-private if others cannot determine w_j uniquely

Privacy-Preserving on Crowdsourced Data

Semi-honest

Non-collusion



Privacy Concerns – WPLC protocol

- Worker-private latent class protocol (WPLC protocol)
  - Model: [Dawid&Skene, 1979]
  - Secure inference:
    - E-step: Requester & workers estimate $\{y_i\}$ by secure computation
    - M-step: Each worker updates the confusion matrix secretly
- Privacy-Preserving Truth Discovery Protocol (PPTD Protocol)
  - Model: CRH [Li et al., SIGMOD'14]
  - Secure inference:
    - Secure Weight Update
    - Secure Truth Estimation

[Kajino et al., 2014] [Miao et al., 2015]

Overview

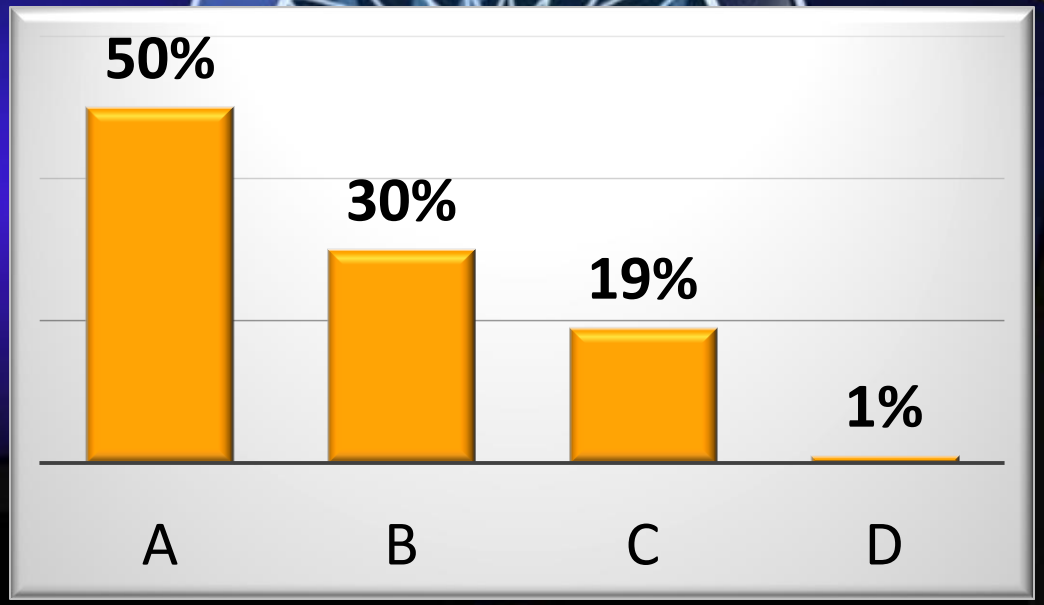
1. Introduction of Crowdsourcing
2. Crowdsourced Data Aggregation for Reliable Information Discovery
3. Passive Crowdsourcing Scenarios
4. Active Crowdsourcing Scenarios
5. **Applications and Open Question**
6. Resources
7. References

Applications

- Wisdom of the crowd
- Slot filling
- Social Sensing
  - Indoor floorplan reconstruction
- Mobile sensing
  - Environmental monitoring
- Community Question Answering
  - Healthcare

Wisdom of the Crowd

- Who want to be a millionaire?
- Smart phone app to collect the players' answers in real time
- We have 2,103 questions, 37,029 users, and 214,849 answers
- The error rate of the truth discovery method is reduced by more than half of voting



16

Which of these square numbers also happens to be the sum of two smaller numbers?

A: 16

B: 25

C: 36

D: 49

https://www.youtube.com/watch?v=BbX44YSsQ2I

Slot Filling

- Extracted from Slot Filling Validation (SFV) task of the NITS Text Analysis Conference Knowledge Base Population (TAC-KBP) track
- Each system in the competition is a data source
- Each slot filling query is an object
- Goal: to find the best slot filling results

[Yu et al., COLING'14]

Slot Filling

Table 1: Example Questions of Slot Filling Task

	Question
q_1	What's the age of Ramazan Bashardost?
q_2	What's the country of birth of Ramazan Bashardost?
q_3	What's the province of birth of Ramazan Bashardost?
q_4	What's the age of Marc Bolland?
q_5	What's the country of birth of Marc Bolland?
q_6	What's the age of Stuart Rose?
q_7	What's the country of birth of Stuart Rose?
q_8	What's the province of death of Stuart Rose?

Method	SF2013			SF2014		
	PREC	REC	F1	PREC	REC	F1
TEM	0.78	0.82	**0.80**	0.65	0.69	**0.67**
Vot	0.38	0.90	0.54	0.35	0.89	0.51
Vot*E*	0.85	0.54	0.66	0.62	0.54	0.58
Find	0.39	0.92	0.55	0.37	0.93	0.53
Find*E*	0.88	0.54	0.67	**0.67**	0.51	0.58
Ave	0.40	**0.93**	0.56	0.37	0.93	0.53
Ave*E*	**0.90**	0.53	0.66	0.66	0.54	0.60
Inv	0.33	0.77	0.46	0.31	0.78	0.44
Inv*E*	0.82	0.49	0.62	0.50	0.44	0.47
PInv	0.12	0.29	0.17	0.25	0.63	0.35
PInv*E*	0.75	0.52	0.62	0.61	0.74	**0.67**
3Est	0.38	0.90	0.54	0.37	**0.94**	0.53
3Est*E*	0.74	0.57	0.65	0.62	0.58	0.60
LCA	0.37	0.87	0.52	0.35	0.90	0.51
LCA*E*	0.85	0.51	0.63	0.63	0.54	0.58
LTM	0.41	0.87	0.56	0.37	0.80	0.51
EM	0.35	0.72	0.47	0.43	0.88	0.57

[Zhi et al., KDD'15]

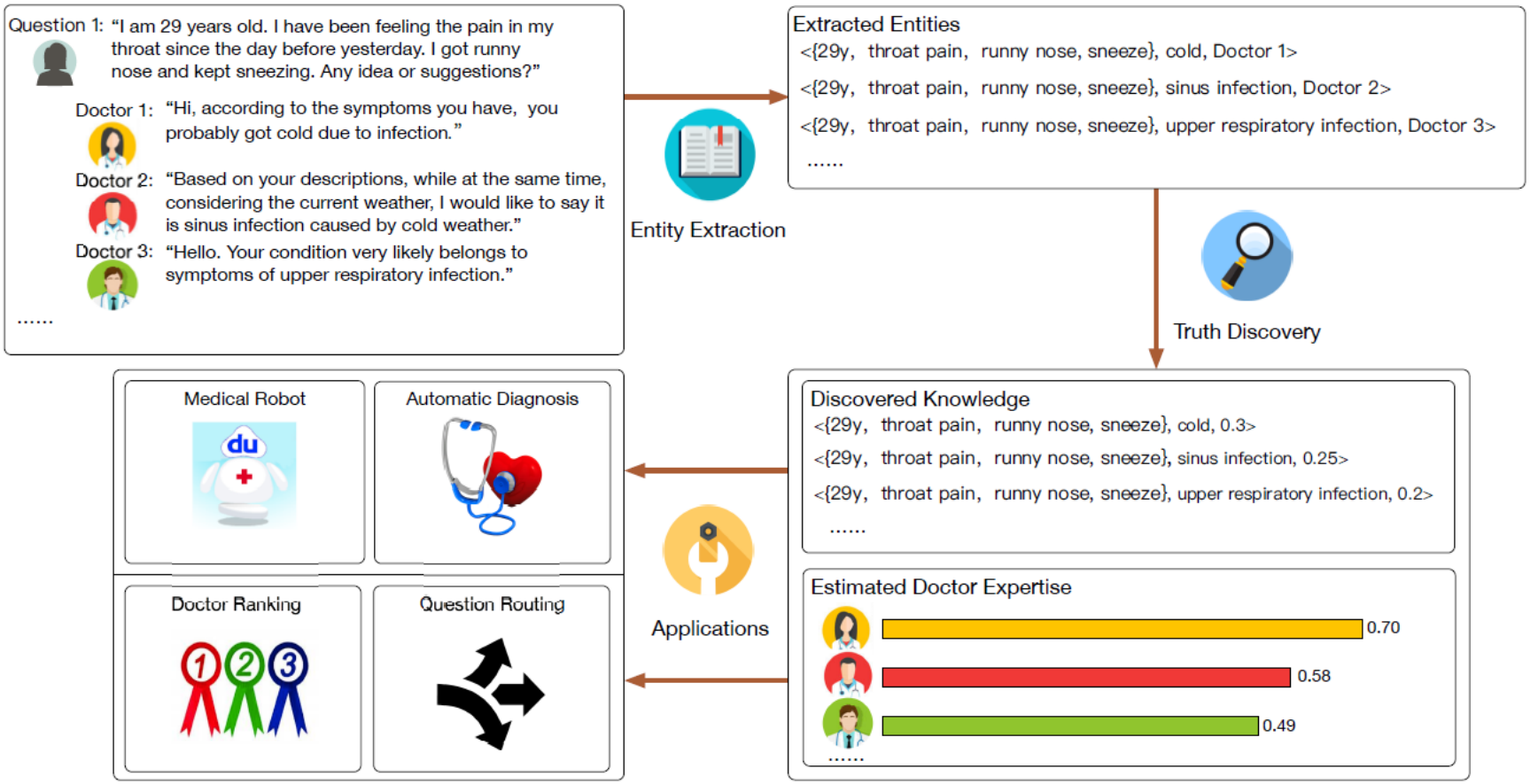
Indoor Floorplan Reconstruction

- Automatic floorplan construction system:
  infer the information about the building floorplan from the movement traces of a group of smartphone users
- One specific task:
  to estimate the distance between two indoor points (e.g., a hallway segment)
- We develop an Android App that can estimate the walking distances of a smartphone user
  - We have 247 users walking on 129 segments

Health-Oriented Community Question Answering Systems

Quality of Question-Answer Thread

Medical Knowledge Extraction System



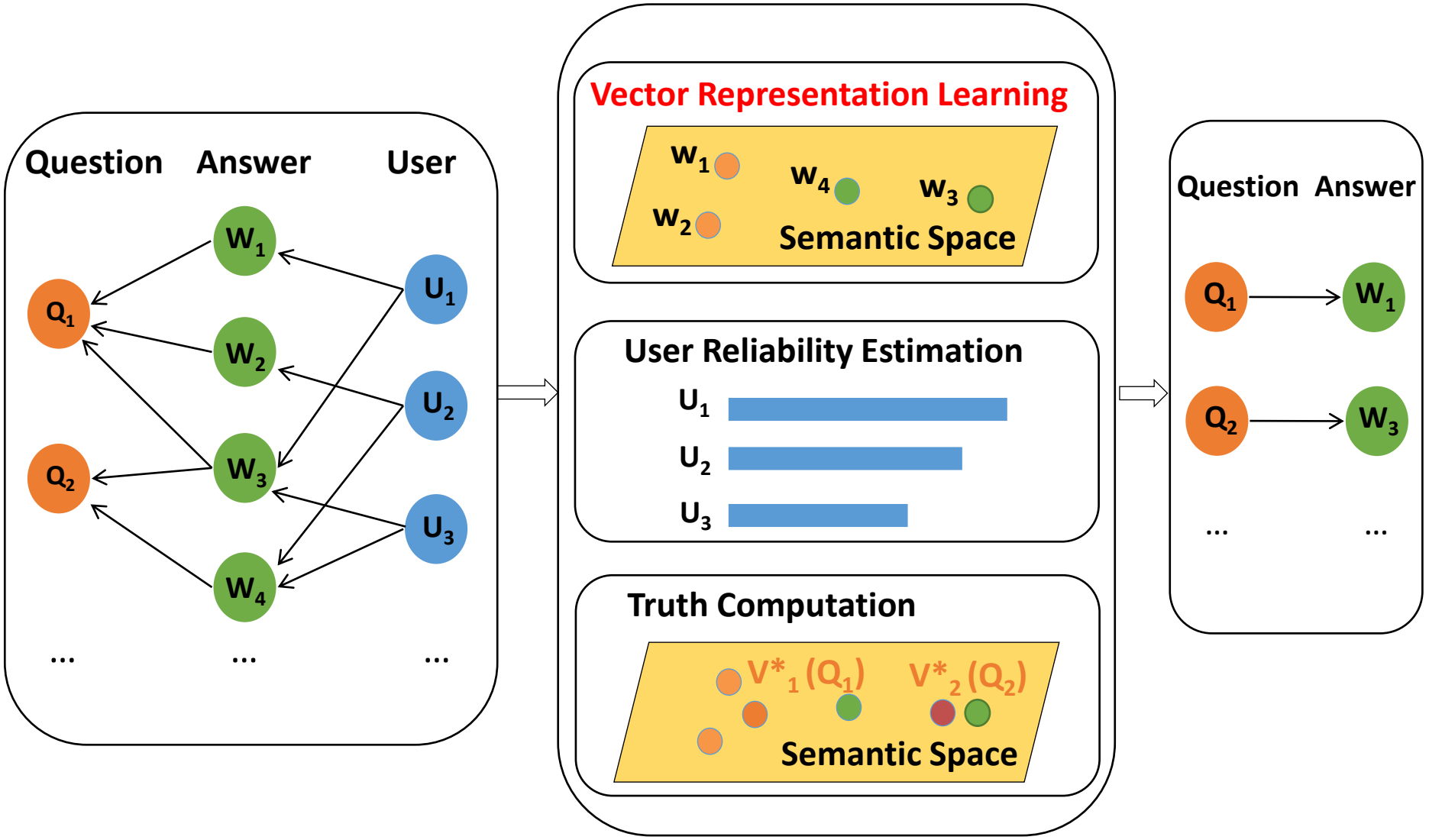
Overview of the Medical Knowledge Extraction (MKE) System. The illustrative example is translated from *xywy.com*, a Chinese medical crowdsourced question answering website.

[Li et al., TBD'16][Li et al., WSDM'17]

Challenges in Knowledge Extraction Systems

- Raw textual data, unstructured
  - Semantic meanings of texts
  - Solution: vector representation
- Long-tail phenomenon
  - Solution: merge similar questions
- Truths can be multiple, and they are correlated with each other
  - Solution: using the similarities between the vector representations of texts

Semantic Truth Discovery Method



Case Study

Symptom	Diagnosis	Possibilities without considering semantic correlations	Possibilities with considering semantic correlations
40 years old, sneezing, running noise	Common cold	0.3253	0.3022
	Allergic rhinitis	0.5556	0.3565
	Rhinitis	0.1190	0.3412
10 years old, chest pain, short of breath, limply	Anemia	0.4946	0.3271
	Enteritis	0.4946	0.3630
	Diarrhea	0.0071	0.3097

Case Study

Symptom	Diagnosis	Possibilities without considering semantic correlations	Possibilities with considering semantic correlations
40 years old, sneezing, running noise	Common cold	0.3253	0.3022
	Allergic rhinitis	0.5556	0.3565
	Rhinitis	0.1190	0.3412
10 years old, chest pain, short of breath, limply	Anemia	0.4946	0.3271
	Enteritis	0.4946	0.3630
	Diarrhea	0.0071	0.3097

Overview

1. Introduction of Crowdsourcing
2. Crowdsourced Data Aggregation for Reliable Information Discovery
3. Passive Crowdsourcing Scenarios
4. Active Crowdsourcing Scenarios
5. Applications and Open Questions
6. **Resources**
7. References

Available Resources

- Survey for truth discovery
  - [Li et al., 2015b]
  - [Waguih et al., 2015]
  - [Waguih et al., 2014]
  - [Li et al., 2012]
  - [Gupta&Han, 2011]
- Survey for crowdsourcing
  - [Zhang et al., 2016]
  - [Hung et al., 2013]
  - [Sheshadri&Lease, 2013]

Available Resources

- Truth discovery data and software
  - http://lunadong.com/fusionDataSets.htm
  - http://cogcomp.cs.illinois.edu/page/resource_view/16
  - http://www.cse.buffalo.edu/~jing/software.htm
- Crowdsourced data aggregation data and software
  - https://sites.google.com/site/amtworkshop2010/data-1
  - http://ir.ischool.utexas.edu/square/index.html
  - https://sites.google.com/site/nlpannotations/
  - http://research.microsoft.com/en-us/projects/crowd
  - http://ceka.sourceforge.net/

- These slides are available at

http://www.cse.buffalo.edu/~jing/talks.htm

References

[Chen et al., ICML'13] X. Chen, Q. Lin, and D. Zhou. Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourcing. In *Proc. of International Conference on Machine Learning*, pages 64-72, 2013.

[Dawid&Skene, 1979] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society, Series C, pages 20–28, 1979.

[Dong et al., VLDB'09a] X. L. Dong, L. Berti-Equille,and D. Srivastava. Integrating conflicting data: The role of source dependence. In *Proc. of VLDB Endow*., pages 550–561, 2009.

[Dong et al., VLDB'09b] X. L. Dong, L. Berti-Equille,and D. Srivastava. Truth discovery and copying detection in a dynamic world. In *Proc. of VLDB Endow*., pages 550–561, 2009.

[Gupta&Han, 2011] M. Gupta and J. Han. Heterogeneous network-based trust analysis: A survey. ACM SIGKDD Explorations Newsletter, 13(1):54–71, 2011.

[Hung et al., 2013] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In Web Information Systems Engineering, pages 1–15, 2013.

[Jin et al., MobiHoc'15] H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu. Quality of Information Aware Incentive Mechanisms for Mobile Crowd Sensing Systems. In *Proc. of ACM Symposium on Mobile Ad Hoc Networking and Computing*, 2015.

[Kajino et al., 2014] H.Kajino, H. Arai, and H. Kashima. Preserving worker privacy in crowdsourcing. Data Mining and Knowledge Discovery, 28(5-6), 1314-1335, 2014.

[Lakkaraju et al., SDM'15] H. Lakkaraju, J. Leskovec, J. Kleinberg, and S. Mullainathan. A Bayesian framework for modeling human evaluations. In *Proc. of the SIAM International Conference on Data Mining*, 2015.

[Li et al., 2013] H. Li, B. Yu, and D. Zhou. Error rate analysis of labeling by crowdsourcing. In ICML Workshop: Machine Learning Meets Crowdsourcing, 2013.

[Li et al., VLDB'15] 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. In *Proc. of VLDB Endow*., 8(4), 2015.

[Li et al., SIGMOD'14] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving Conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proc. of the ACM SIGMOD International Conference on Management of Data*, pages 1187–1198, 2014.

[Li et al., WSDM'16] Q. Li, F. Ma, J. Gao, L. Su, and C. J. Quinn. Crowdsourcing High Quality Labels with a Tight Budget. In *Proc. of ACM International Conference on Web Search and Data Mining*, 2016.

[Li et al., VLDB'12] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? In *Proc. of VLDB Endow.*, 6(2):97–108, 2012.

[Li et al., 2015] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 2015.

[Li et al. TBD'16] Y. Li, C. Liu, N. Du, W. Fan, Q. Li, J. Gao, C. Zhang, and H. Wu. Extracting Medical Knowledge from Crowdsourced Question Answering Website. IEEE Transactions on Big Data, 2016.

[Li et al. WSDM'17] Y. Li, N. Du, C. Liu, Y. Xie, W. Fan, Q. Li, J. Gao, and H. Sun. Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts. In *Proc. of ACM International Conference on Web Search and Data Mining*, 2017.

[Liu et al., NIPS'12] Q. Liu, J. Peng, and A. Ihler. Variational Inference for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 692–700, 2012.

[Ma et al., KDD'15] F. Ma, Y. Li, Q. Li, M. Qui, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[Meng et al., SenSys'15] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng. Truth Discovery on Crowd Sensing of Correlated Entities. In *Proc. of the ACM Conference on Embedded Networked Sensor Systems*, 2015.

[Miao et al., SenSys'15] C. Miao, W. Jiang, L. Su, Y. Li, S. Guo, Z. Qin, H. Xiao, J. Gao, and K. Ren. Cloud-Enabled Privacy-Preserving Truth Discovery in Crowd Sensing Systems. In *Prof of the ACM Conference on Embedded Networked Sensor Systems*, 2015.

[Mozafari et al., VLDB'14] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden. Scaling up crowd-sourcing to very large datasets: a case for active learning. In *Proc. of VLDB Endow*., 8(2):125–136, 2014.

[Mukherjee et al., KDD'14] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: credibility of user statements in health communities. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, 2014.

[Pasternack&Roth, WWW'13] J. Pasternack and D. Roth. Latent credibility analysis. In *Proc. of the International Conference on World Wide Web*, pages 1009–1020, 2013.

[Pochampally et al., SIGMOD'14] R. Pochampally, A. D. Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In *Proc. of the ACM SIGMOD International Conference on Management of Data*, pages 433–444, 2014.

[Qi et al., WWW'13] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proc. of the International Conference on World Wide Web*, pages 1041–1052, 2013.

[Raykar et al., 2010] V. C. Raykar, S. Yu, L. H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. The Journal of Machine Learning Research, 11: 1297–1322, 2010.

[Shah&Zhou, NIPS'15] N. Shah and D. Zhou. Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing. In *Advances in Neural Information Processing Systems*, 2015.

[Sheng et al., KDD'08] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

[Sheshadri&Lease, HCOMP'13] A. Sheshadri and M. Lease. SQUARE: A benchmark for research on computing crowd consensus. In *Proc. of the AAAI Conference on Human Computation*, pages 156–164, 2013.

[Snow et al., EMNLP'08] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast—but is it good? Evaluating non-expert Annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263, 2008.

[Waguih et al., 2014] D. A. Waguih and L. Berti-Equille. Truth discovery algorithms: An experimental evaluation. arXiv preprint arXiv:1409.6428, 2014.

[Waguih et al., ICDE'15] D. A. Waguih, N. Goel, H. M. Hammady, and L. Berti-Equille. Allegatortrack: Combining and reporting results of truth discovery from multi-source data. In *Proc. of the IEEE International Conference on Data Engineering*, 2015.

[Whitehill et al., NIPS'09] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, pages 2035–2043, 2009.

[Xiao et al., KDD'16] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang. Towards Confidence in the Truth: A Bootstrapping based Truth Discovery Approach. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016.

[Yan et al., ICML'11] Y. Yan, G. M. Fung, R. Rosales, and J. G. Dy. Active learning from crowds. In *Proc. of the International Conference on Machine Learning*, pages 1161–1168, 2011.

[Yin et al., TKDE'08] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering, 20(6): 796–808, 2008.

[Yu et al., COLING'14] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proc. of the International Conference on Computational Linguistics*, 2014.

[Zhang et al., NIPS'14] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet EM: A provably Optimal Algorithm for Crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 1260–1268, 2014.

[Zhang et al., 2015] J. Zhang, V. S. Sheng, B. A. Nicholson, and X. Wu. CEKA: A Tool for Mining the Wisdom of Crowds. Journal of Machine Learning Research, vol. 16, page 2853–2858.

[Zhang et al., 2016] J. Zhang, X. Wu, and V.S. Sheng. Learning from crowdsourced labeled data: a survey. Artificial Intelligence Review. 2016:1-34.

[Zhao&Han, QDB'12] B. Zhao, and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. of the VLDB workshop on Quality in Databases*, 2012.

[Zhao et al., VLDB'12] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. In *Proc. of VLDB Endow*., 5(6):550–561, 2012.

[Zhao et al., PASSAT'11] L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In PASSAT and SocialCom, pages 728–733, 2011.

[Zhi et al., KDD'15] S. Zhi, B. Zhao, W. Tong, J. Gao, D. Yu, H. Ji, and J. Han. Modeling truth existence in truth discovery. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[Zhou et al., NIPS'12] D. Zhou, J. Platt, S. Basu, and Y. Mao. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2012.