Truth Discovery and Crowdsourcing Aggregation: A Unified Perspective

Jing Gao[1], Qi Li[1], Bo Zhao[2], Wei Fan[3], and Jiawei Han[4]
[1]SUNY Buffalo; [2]LinkedIn;
[3]Baidu Research Big Data Lab; [4]University of Illinois

Overview

1. Introduction
2. Comparison of Existing Truth Discovery and Crowdsourced Data Aggregation Setting
3. Models of Truth Discovery and Crowdsourced Data Aggregation
4. Truth Discovery for Crowdsourced Data Aggregation
5. Related Areas
6. Open Questions and Resources
7. References

Overview

1
- Introduction

2
- Comparison of Existing Truth Discovery and Crowdsourced Data Aggregation Setting

3
- Models of Truth Discovery and Crowdsourced Data Aggregation

4
- Truth Discovery for Crowdsourced Data Aggregation

5
- Related Areas

6
- Open Questions and Resources

7
- References

Truth Discovery

- Conflict resolution in data fusion

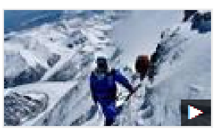
Google what is the height of mount everest

[**Mount Everest** - Wikipedia, the free encyclopedia]
en.wikipedia.org/wiki/**Mount_Everest**
By the same measure of base to summit, **Mount** McKinley, in Alaska, is also taller than **Everest**. Despite its **height** above sea level of only 6,193.6 m (20,320 ft), ...
List of deaths on eight ... - Edmund Hillary - Timeline of climbing Mount - 1996

[**Mt. Everest Height** Mystery May Be Answered : Discovery News]
news.discovery.com/.../**everest**-official-**height**-120301.htm
Mar 1, 2012 – The plunge from 71581 feet was a success. Next up: 120000 feet.

[Facts About **Mt. Everest**]
teacher.scholastic.com/activities/hillary/archive/evefacts.htm
Number of people to successfully climb **Mt. Everest**: 660. Number of people who have died trying to climb **Mt. Everest**: 142. **Height**: 29,028 feet, or 5 and a half ...

[**Mount Everest** by the Numbers: Deaths, Cost to Climb, and More ...]



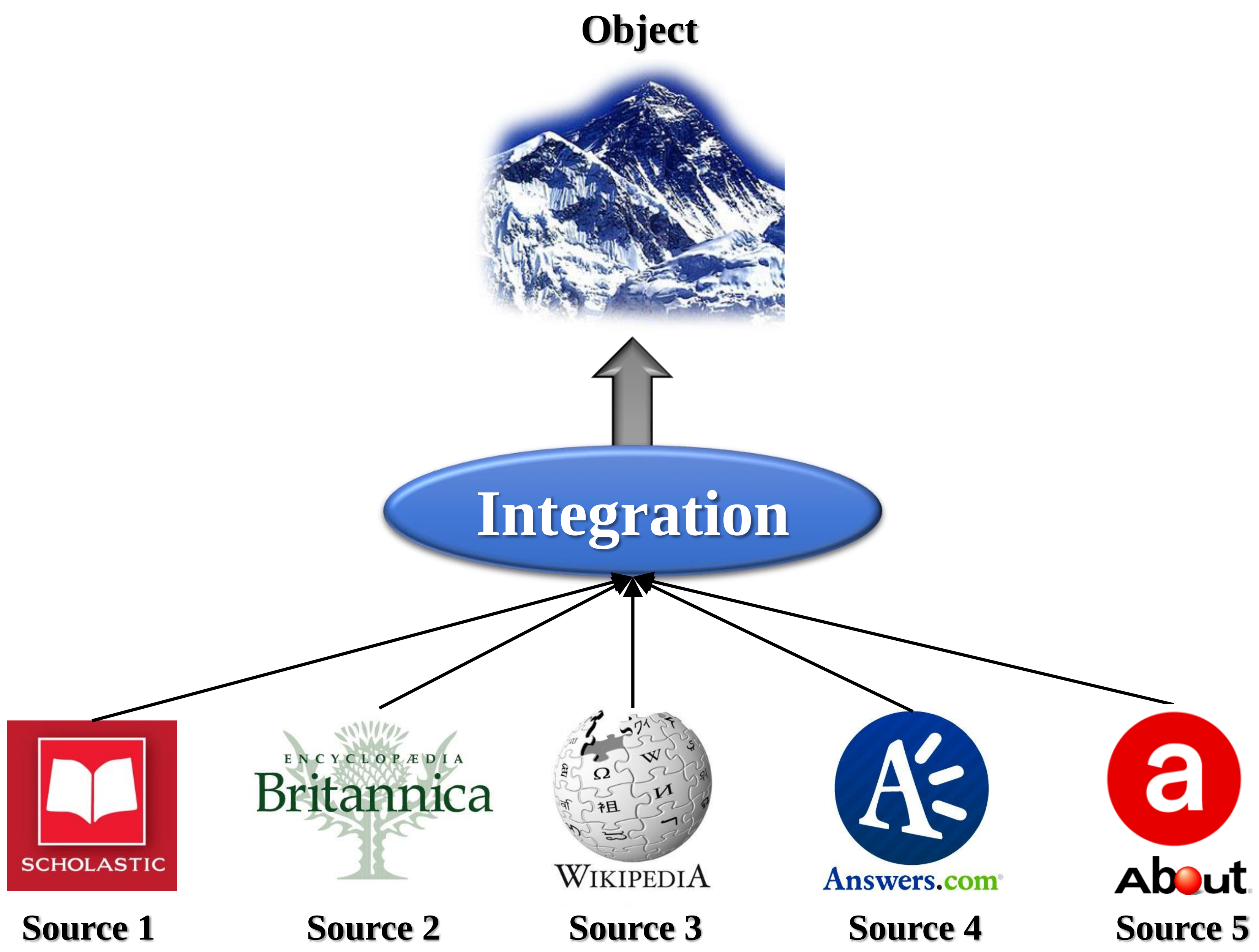
www.thedailybeast.com/.../**mount-everest**-by-th...
May 22, 2012
8,000: **Height** in meters (approximately 26,000 feet) at **Mount Everest's** "death zone," the low-oxygen area above ...

More videos for **what is the height of mount everest** »

[**What is the height of Mount Everest**]
wiki.answers.com › ... › Geography › Landforms › Mountains
Mt. Everest is 29,002 feet hight. And 348,024 inches high. What is the real **height of Mount Everest**? 12,000 ft!!! Everest is, to begin with, 18,000 ft above sea level ...

[**Height of Mount Everest** (Everest, Mount) -- Britannica Online ...]
www.britannica.com/EBchecked/.../**Height-of-Mount-Everest**
The **height of Mount Everest**, according to the most recent and reliable data, is 29035 feet (8850 metres). In 1999 an American survey, sponsored by the (U.S.) ...

[**Mount Everest** - Overview of **Mount Everest**]
geography.about.com › ... › Specific Places of Interest
With a peak **elevation** of 29,035 feet (8850 meters), the top of **Mount Everest** is the world's highest point above sea level. As the world's highest mountain, ...


Object
Integration
SCHOLASTIC
ENCYCLOPÆDIA
Britannica
WIKIPEDIA
Answers.com
About
Source 1
Source 2
Source 3
Source 4
Source 5

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

- What is truth discovery?

Goal:

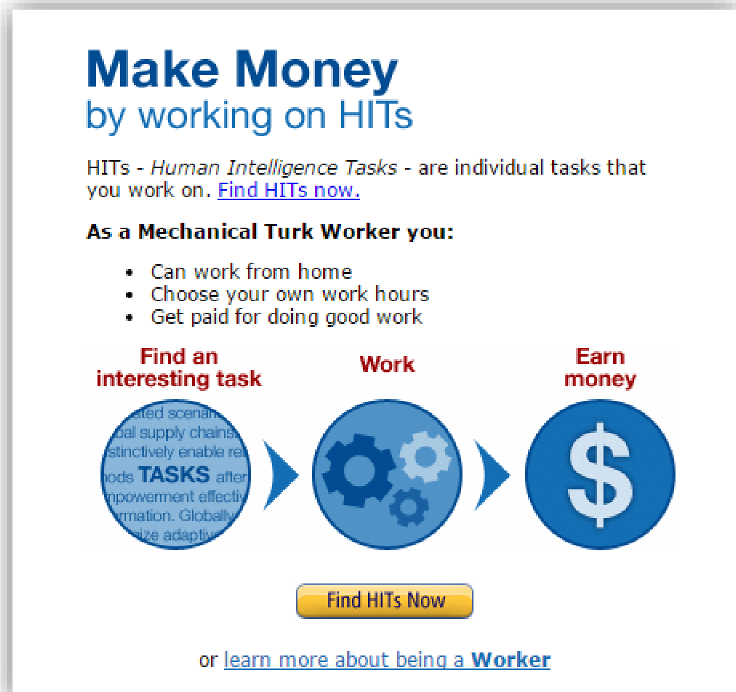
To discover truths by integrating source reliability estimation in the process of data fusion

Crowdsourcing

amazonmechanical turk
Artificial Artificial Intelligence

CrowdFlower •••

requester



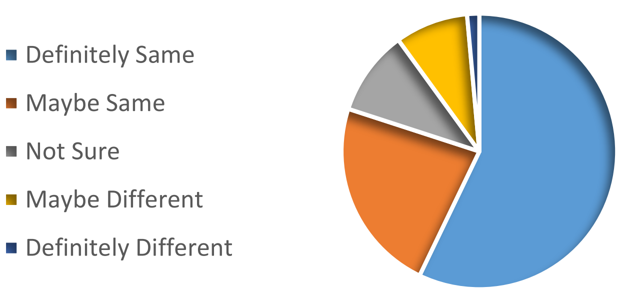
worker

An Example on Mturk

Are the two images of the same person?

Definitely Same Maybe Same … Not Sure

Annotation Results



Final Answer:
Same

Crowdsourced Data Aggregation

- What is crowdsourced data aggregation?
- ***Goal***:

  To resolve disagreement between responses.

Overview

1
- Introduction

2
- **Comparison of Existing Truth Discovery and Crowdsourced Data Aggregation Setting**

3
- Models of Truth Discovery and Crowdsourced Data Aggregation

4
- Truth Discovery for Crowdsourced Data Aggregation

5
- Related Areas

6
- Open Questions and Resources

7
- References

Similarity

- 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

Differences

- Data collection and generation
- Data format of claims

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

Overview

1. Introduction
2. Comparison of Existing Truth Discovery and Crowdsourced Data Aggregation Setting
3. **Models of Truth Discovery and Crowdsourced Data Aggregation**
4. Truth Discovery for Crowdsourced Data Aggregation
5. Related Areas
6. Open Questions and Resources
7. References

Model Categories

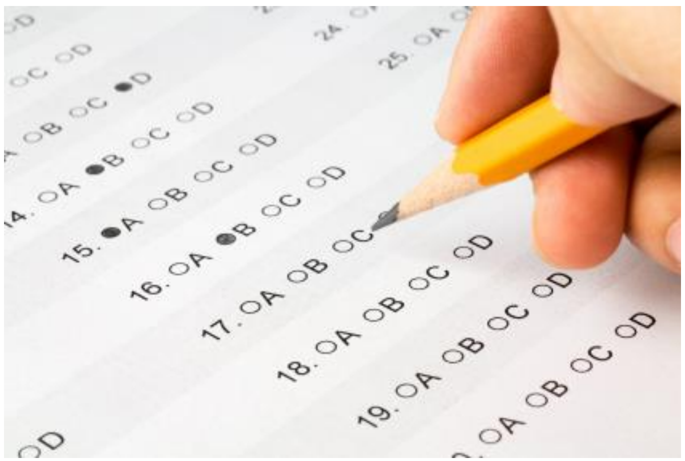
- Statistical model (STA)
- Probabilistic graphical model (PGM)
- Optimization model (OPT)
- Extension (EXT)
    - Source correlation

Statistical Model (STA)

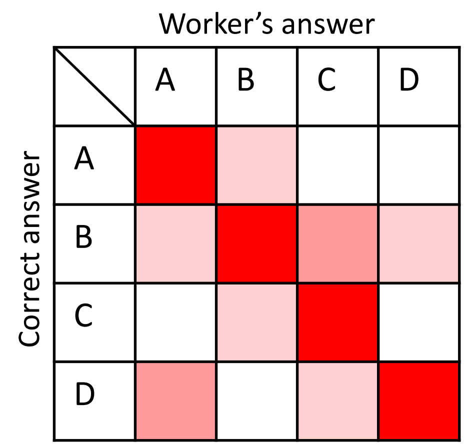
- General goal:
  - **To find the (conditional) probability of a claim being true**
- Source reliability:
  - **Probability(ies) of a source/worker making a true claim**

STA - 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]

STA - Maximum Likelihood Estimation

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

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

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

STA - 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

STA - Extension and Theoretical Analysis

- Extensions
  - Naïve Bayesian [Snow et al., 2008]
  - Finding a good initial point [Zhang et al., 2014]
  - 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., 2012] [Lakkaraju et al. 2015]
  - Clustering workers/instances [Lakkaraju et al. 2015]
- Theoretical analysis
  - Error bound [Li et al., 2013] [Zhang et al., 2014]

STA - TruthFinder

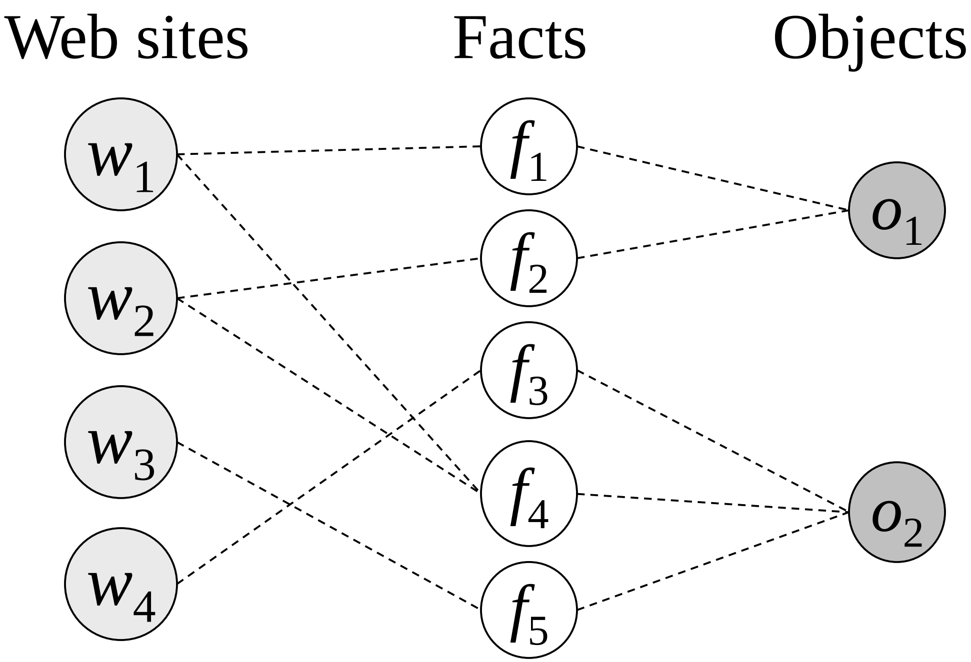
Different websites often provide conflicting information on a subject, e.g., Authors of *"Rapid Contextual Design"*

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., 2008]

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?



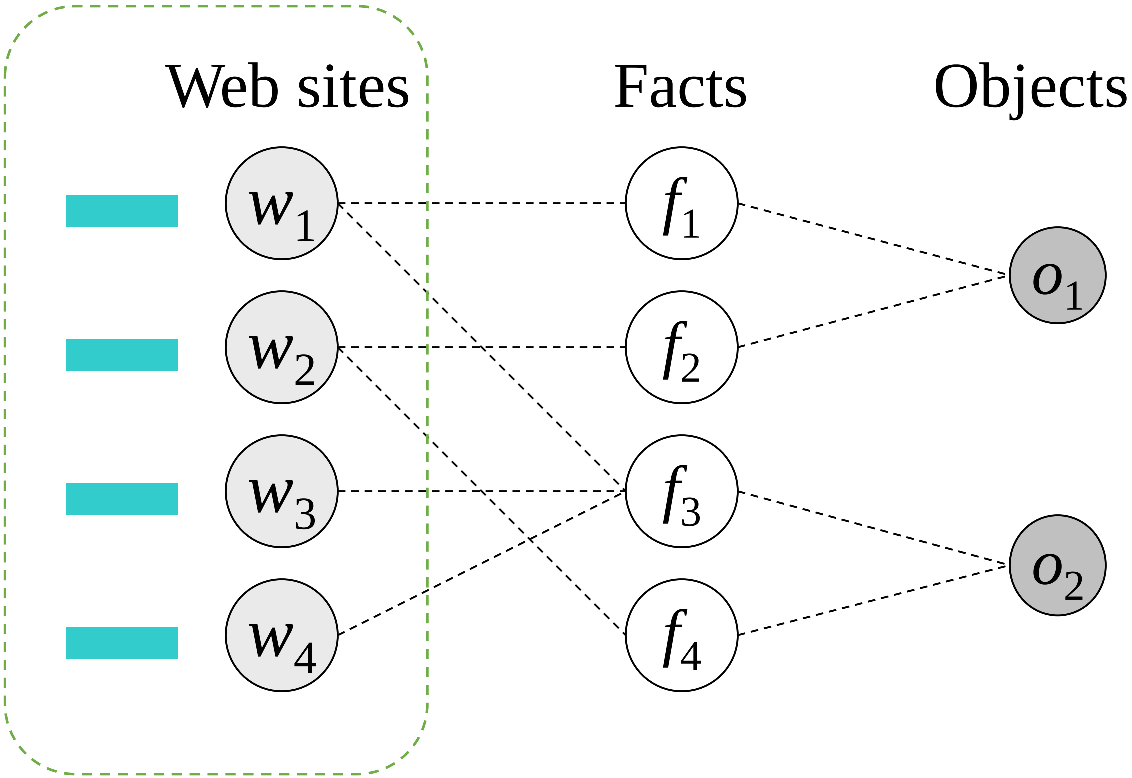
STA - TruthFinder

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**

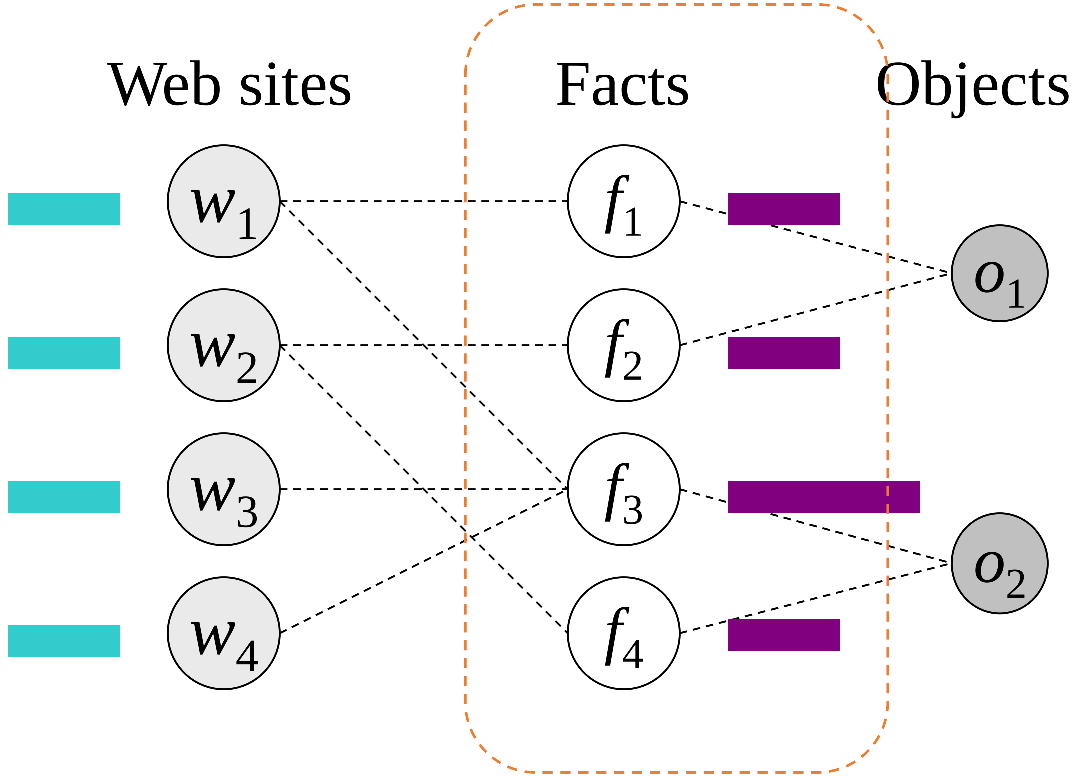
STA - TruthFinder

- <u>*Confidence of facts*</u> ↔ <u>*Trustworthiness of web sites*</u>
  - 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

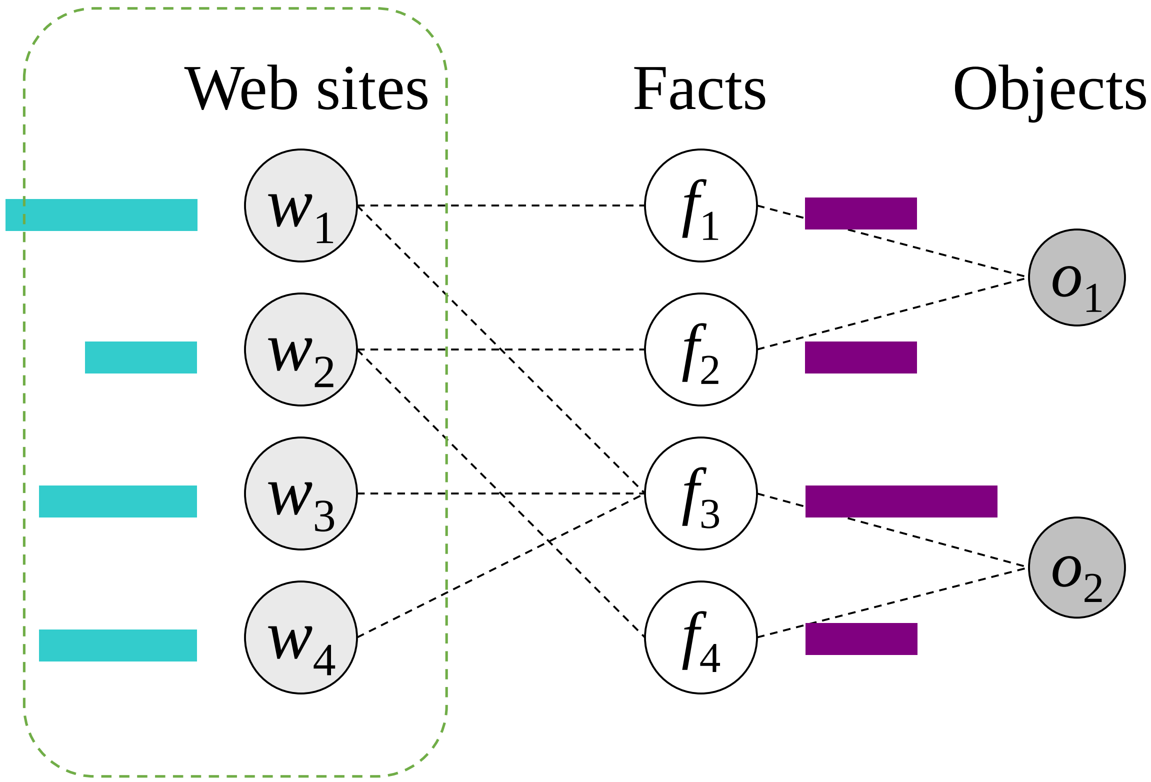
STA - TruthFinder



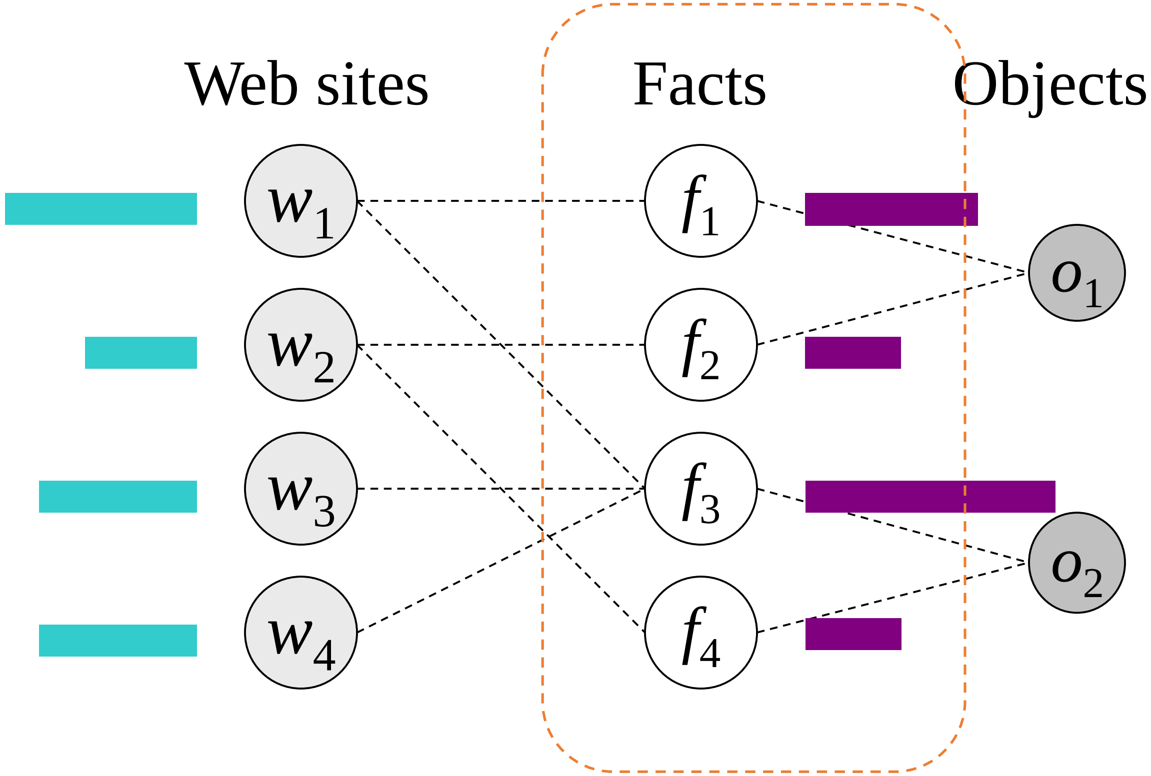
STA - TruthFinder



STA - TruthFinder



STA - TruthFinder



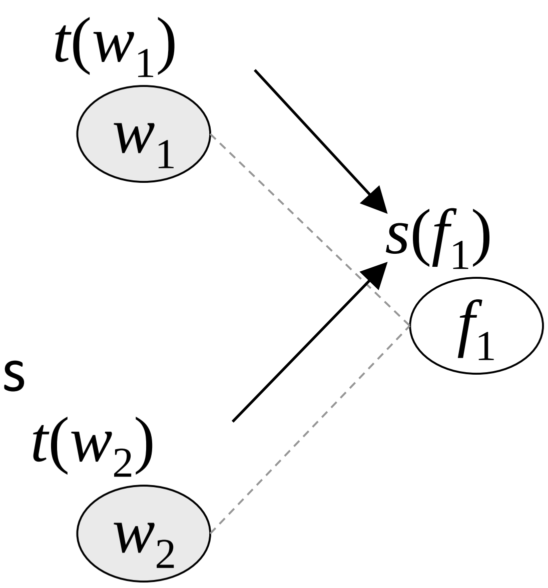
STA - TruthFinder

- **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



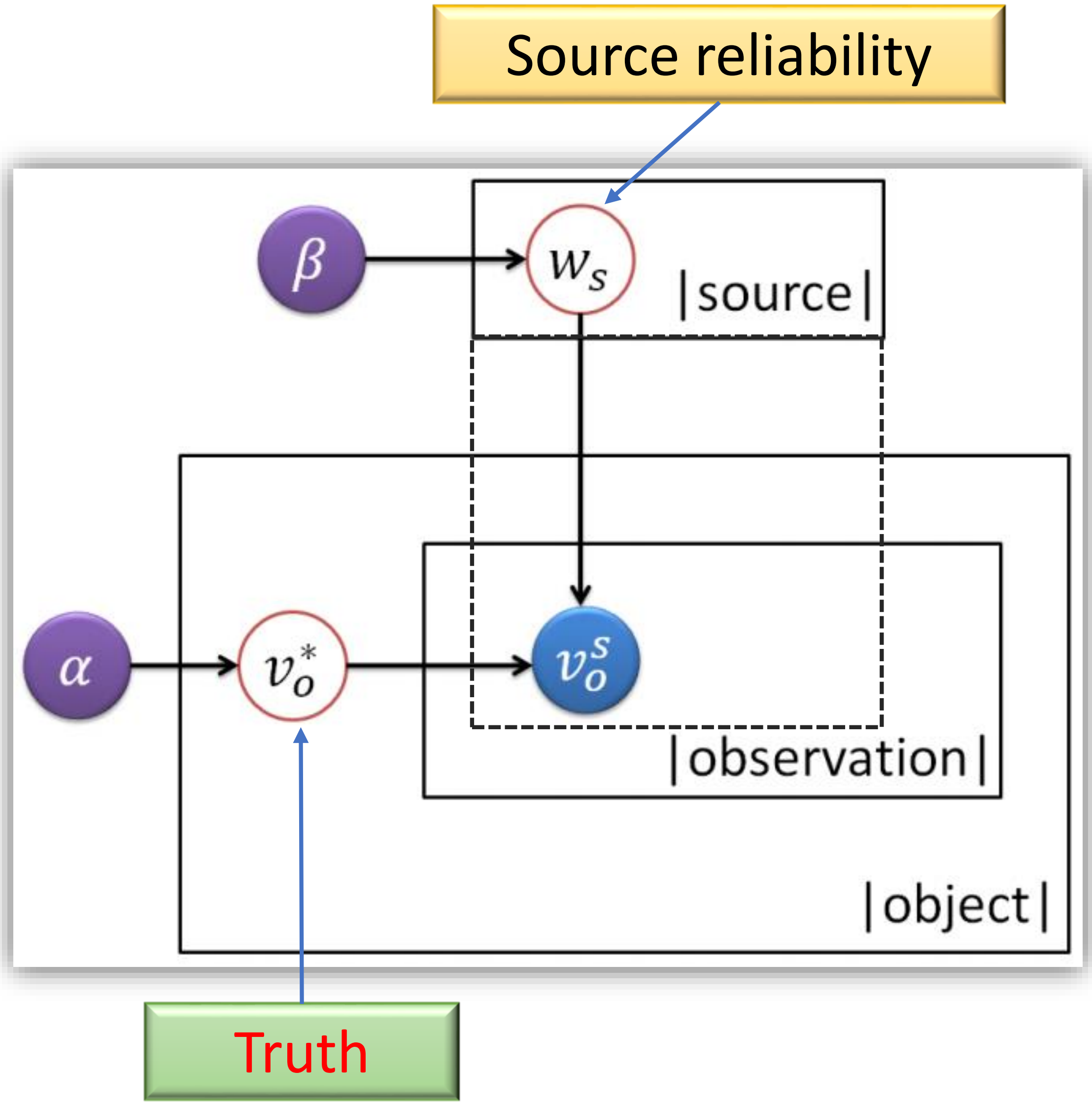
- **The confidence of a fact f: $s(f)$**
  - One minus the probability that all web sites providing *f* are wrong

Probability that w is wrong

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

Set of websites providing f

Probabilistic Graphical Model (PGM)

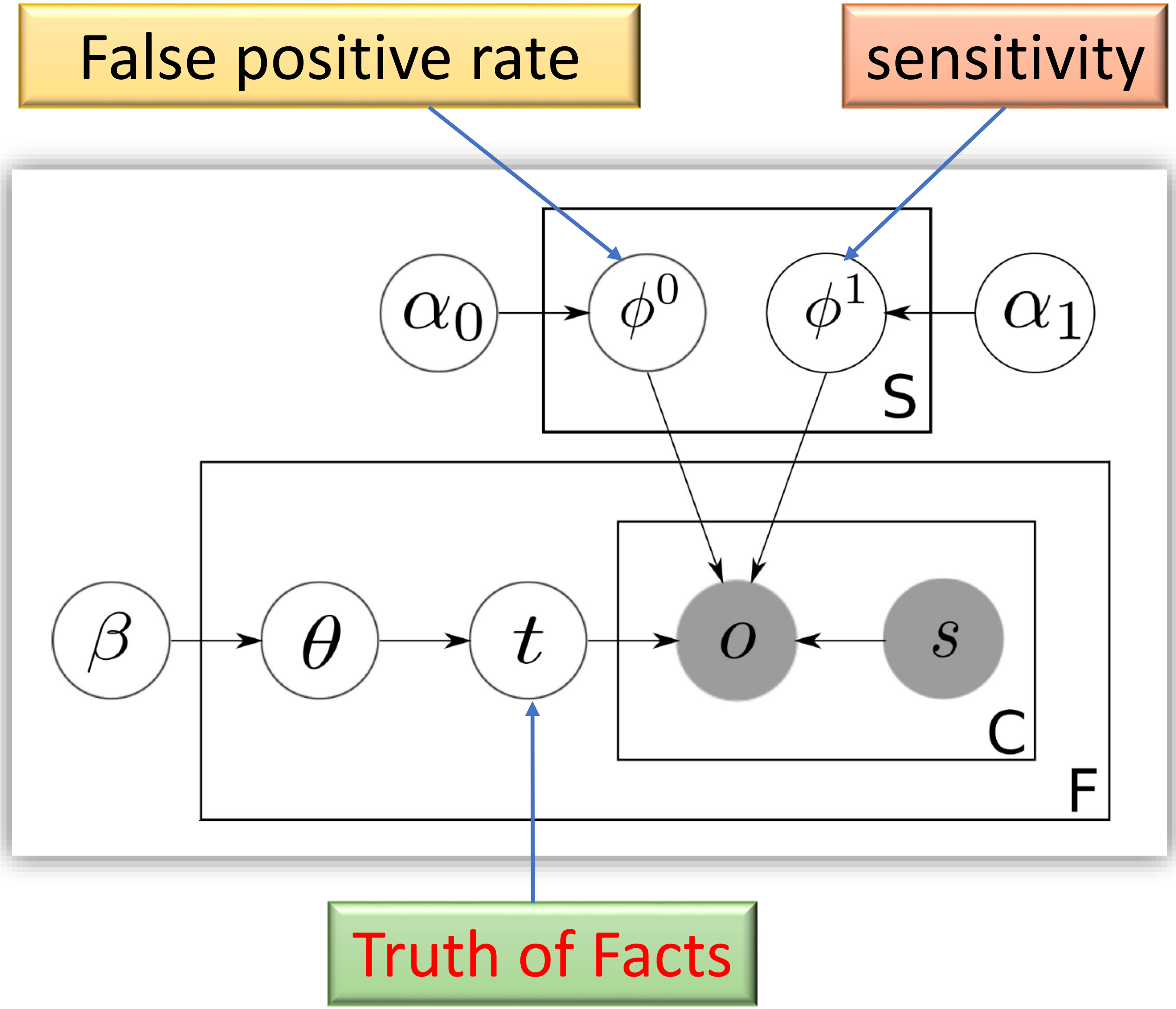


PGM - Latent Truth Model (LTM)

- **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., 2012]

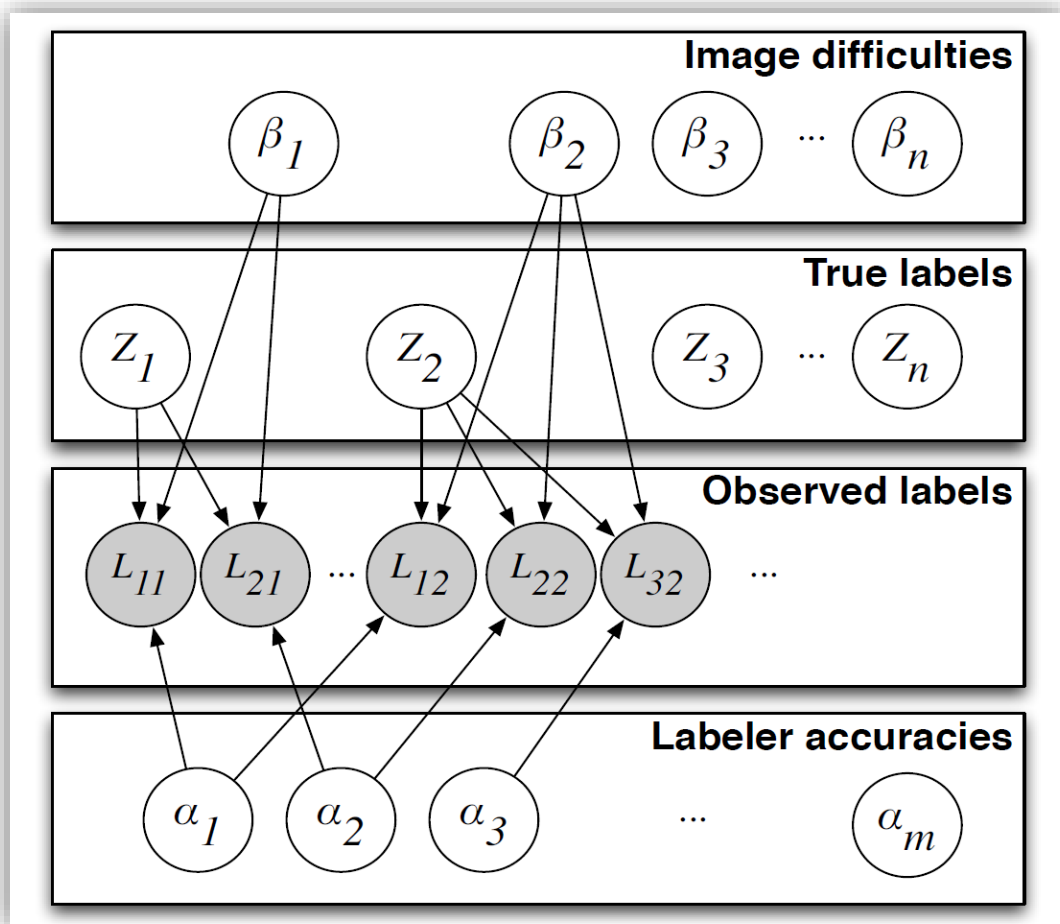
PGM - Latent Truth Model (LTM)



PGM - 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: $o_c \sim Bernoulli(\phi_{s_c}^1)$

PGM - GLAD Model



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

Known variables: observed labels.

[Whitehill et al., 2009]

PGM - GLAD Model

Observed label

True label

Log odds for the obtained labels being correct

$$p\left(\boxed{L_{ij}} = \boxed{Z_j} \middle| \boxed{\alpha_i}, \boxed{\beta_j}\right) = \frac{1}{1 + e^{-\alpha_i \beta_j}}$$

Worker's accuracy.
Always correct → $\alpha_i = +\infty$
Always wrong → $\alpha_i = -\infty$

Difficulty of image.
Very ambiguous → $1/\beta_j = +\infty$
Very easy → $1/\beta_j = 0$

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 $\delta_1, \delta_2, \ldots$.
  - 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

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

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

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., 2014]

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., 2012]

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}

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

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

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

OPT - Minimax Entropy

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

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

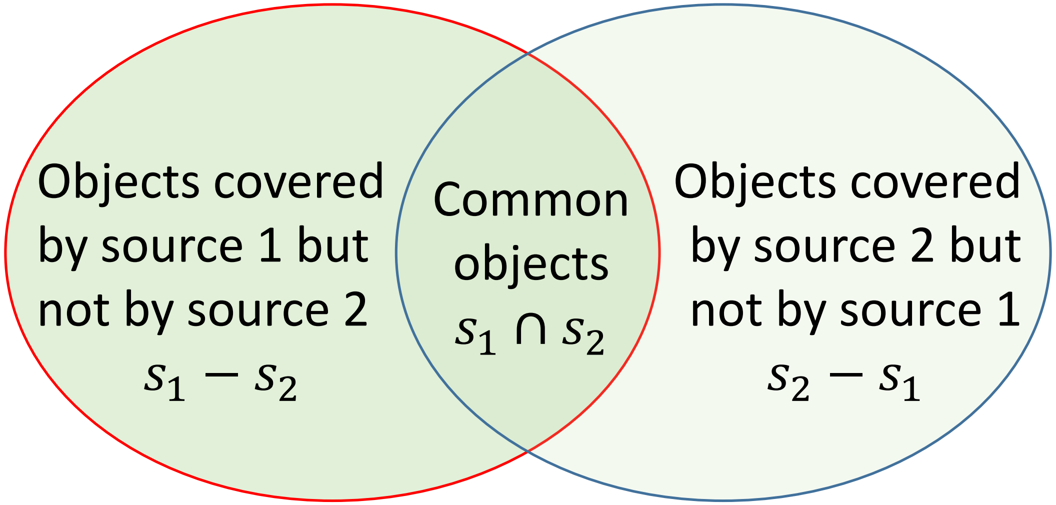
s.t. $\sum_{i=1}^{m} \pi_{i}$ … $0, \forall j, l.$

$\sum_{k=1}^{c} \pi_{i}$ …

Minimize entropy
is equivalent to
minimizing the KL divergence

EXT - Source Correlation

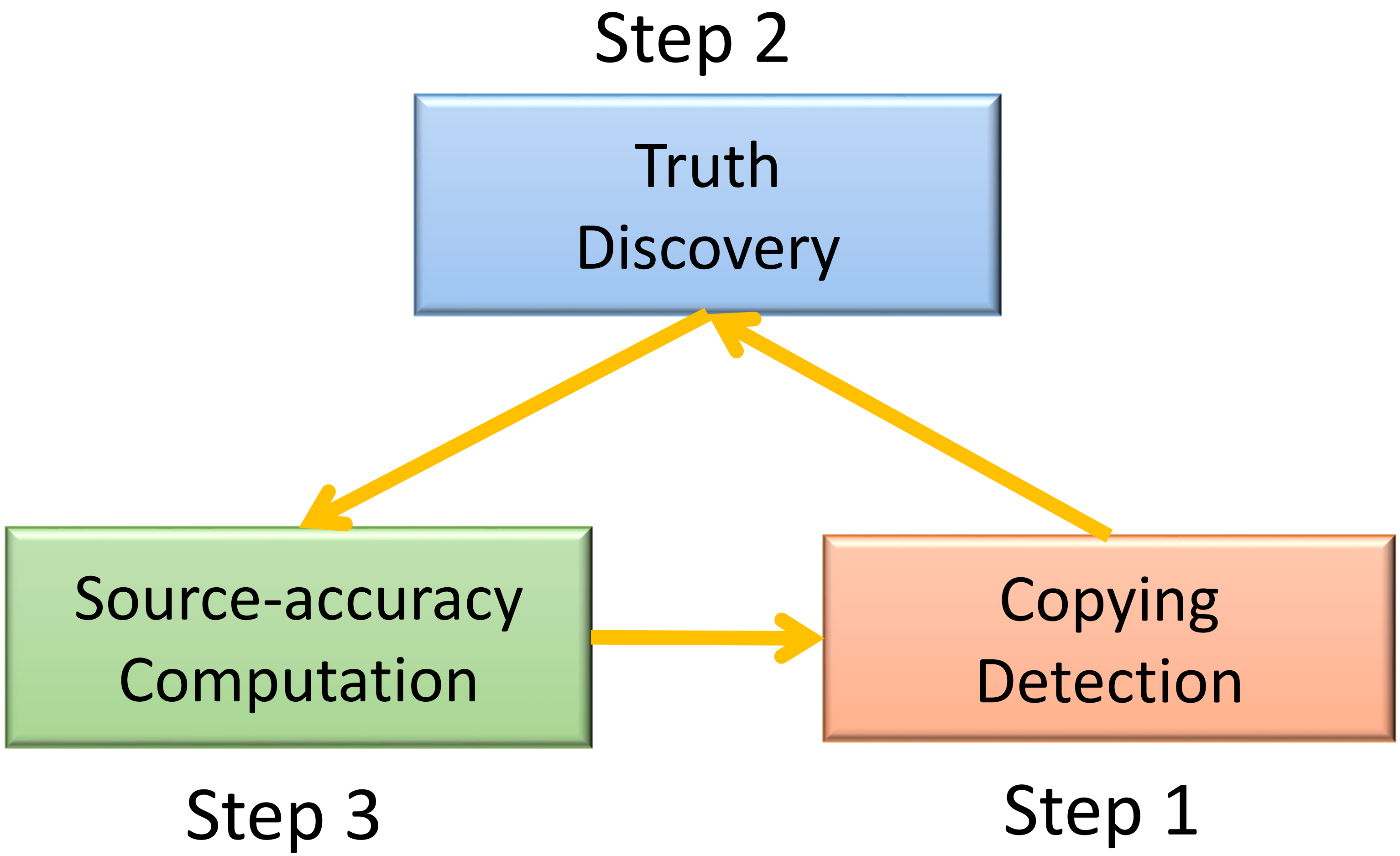
- High-level intuitions for copying detection
  - Common error implies copying relation
    - e.g., many same errors in $s_1 \cap s_2$ imply source 1 and 2 are related
  - Source reliability inconsistency implies copy direction
    - e.g., $s_1 \cap s_2$ and $s_1 - s_2$has similar accuracy, but $s_1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



[Dong et al., 2009a] [Dong et al., 2009b]

EXT - Source Correlation

- Incorporate copying detection in truth discovery



EXT - Source Correlation

- More general source correlations
  - Sources may provide data from complementary domains (negative correlation)
  - Sources may focus on different types of information (negative correlation)
  - Sources may apply common rules in extraction (positive correlation)
- How to detect
  - Hypothesis test of independence using joint precision and joint recall

[Pochampally et al., 2014]

Overview

1. Introduction
2. Comparison of Existing Truth Discovery and Crowdsourced Data Aggregation Setting
3. Models of Truth Discovery and Crowdsourced Data Aggregation
4. **Truth Discovery for Crowdsourced Data Aggregation**
5. Related Areas
6. Open Questions and Resources
7. References

Truth Discovery for Crowdsourced Data Aggregation

- Crowdsourced data
  - Not limited to data collected from Mechanical Turk
  - Can be collected from social media platforms, discussion forums, smartphones, ……
- Truth discovery is useful
  - Open-space passively crowdsourced data
  - Methods based on confusion matrix do not work
- New challenges for truth discovery

Passively Crowdsourced Data

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…"

RxList

MAYO CLINIC

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

DEPO Rash
……

Passively Crowdsourced Data

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
……

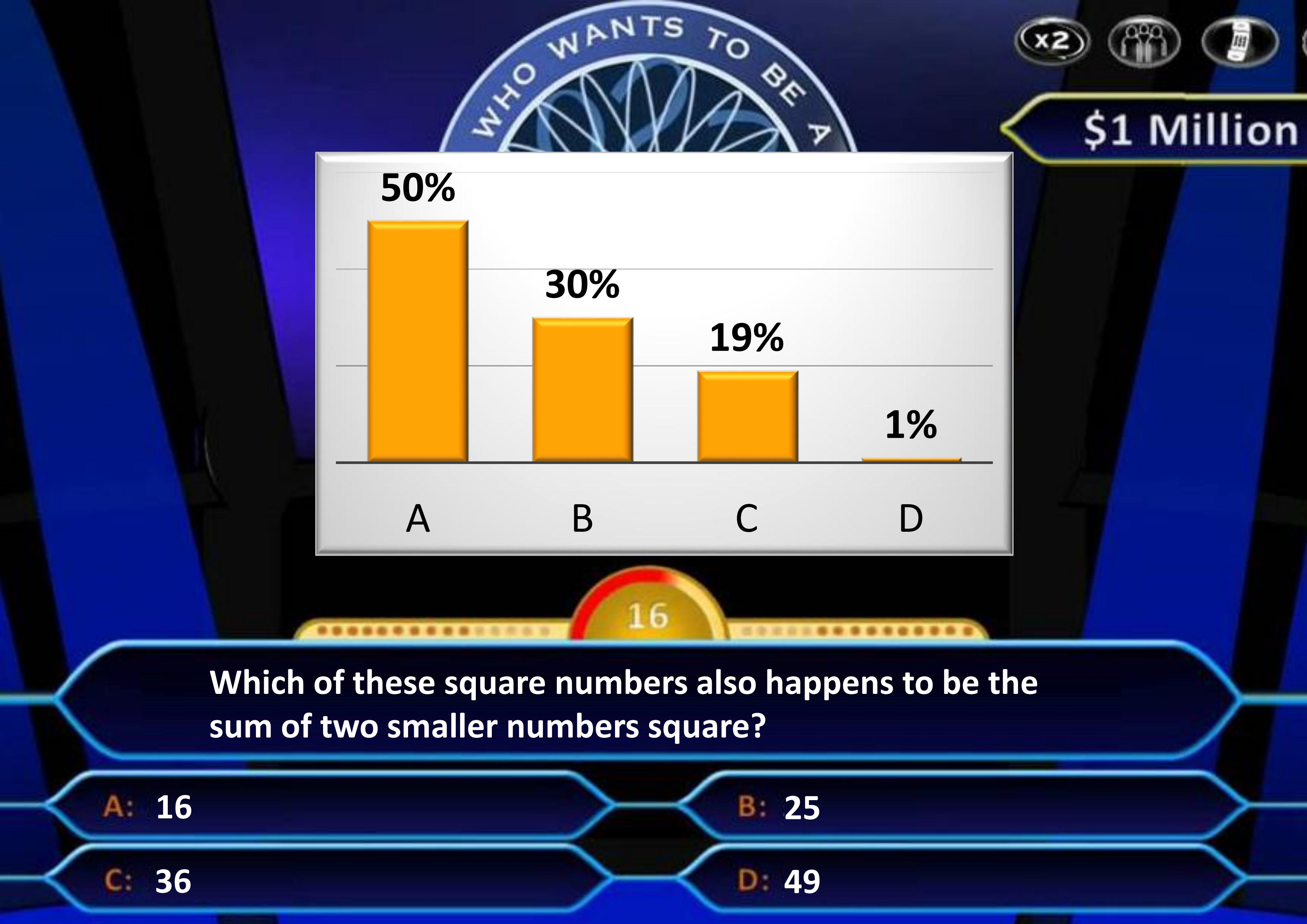
JFK Bad Traffic
……

THE WHITE HOUSE
WASHINGTON

CATD Model

- **Long-tail phenomenon**
  - Most sources only provide very few claims and only a few sources makes plenty of claims.
- **A confidence-aware approach**
  - not only estimates source reliability
  - but also considers the confidence interval of the estimation

[Li et al., 2015a]



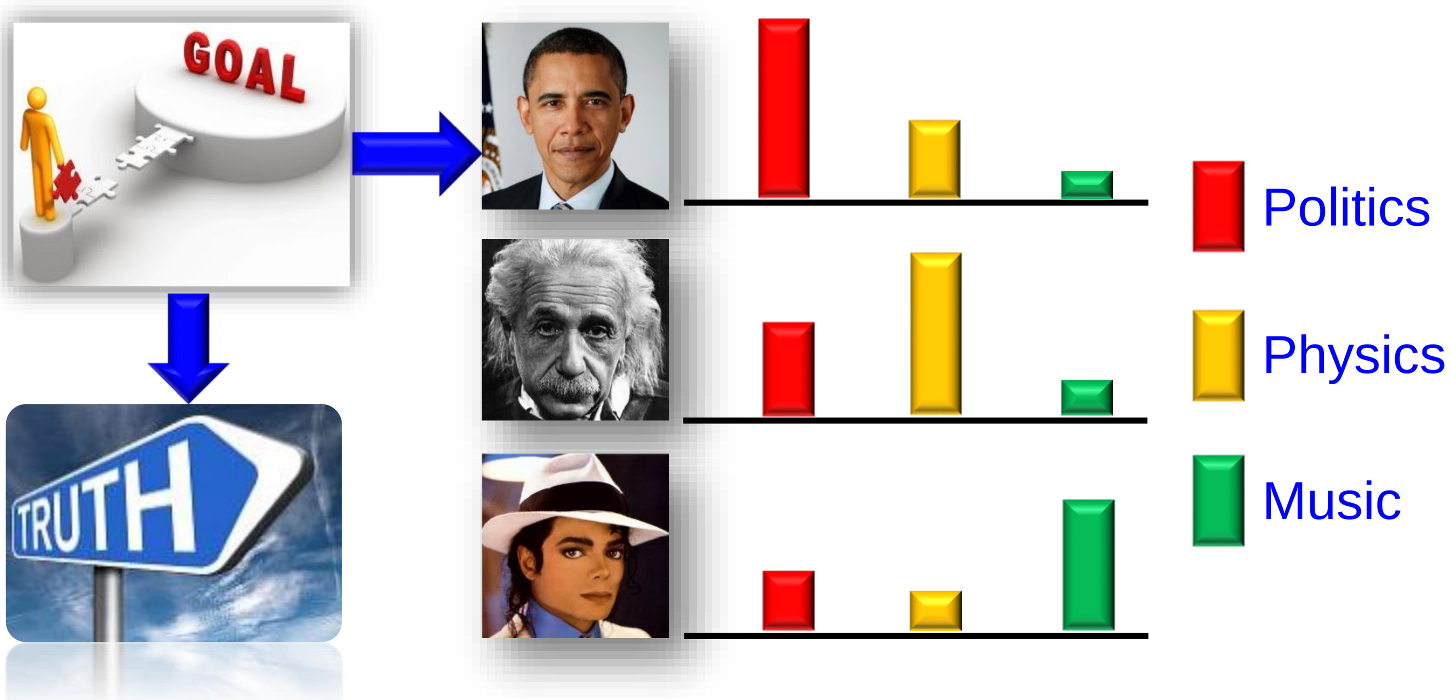
https://www.youtube.com/watch?v=BbX44YSsQ2I

Error Rate Comparison on Game Data

Question level	Majority Voting	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

FaitCrowd

- **Goal**
  - To learn **fine-grained (topical-level) user expertise** and the **truths** from conflicting crowd-contributed answers.



[Ma et al., 2015]

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

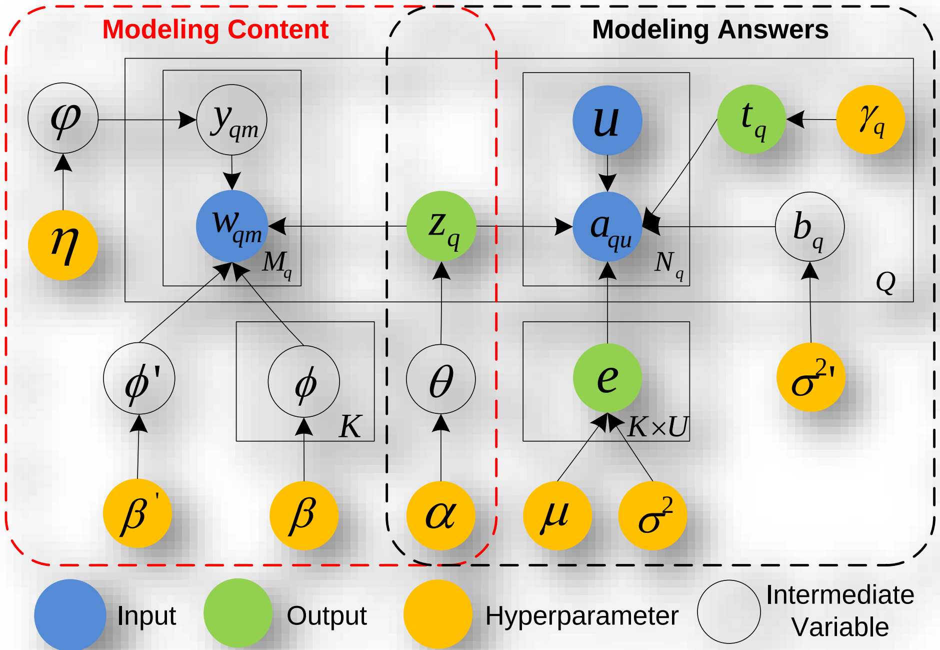
User		u1	u2	u3
Expertise	K1	2.34	2.70E-4	1.00
	K2	1.30E-4	2.34	2.35

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

FaitCrowd

- **Overview**

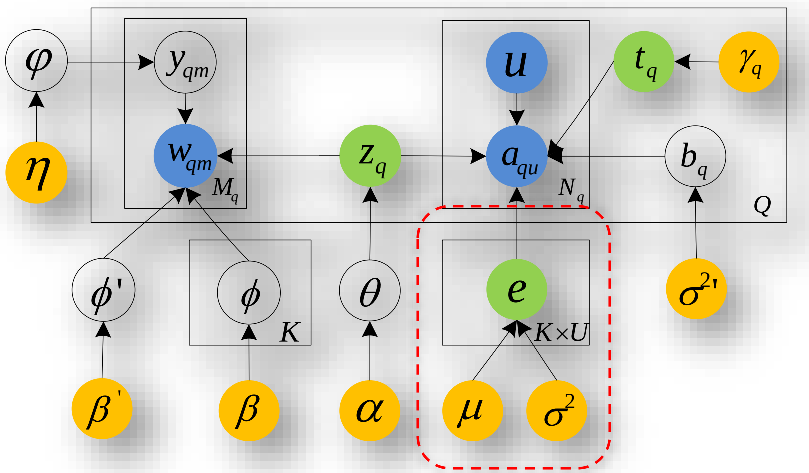


- Jointly modeling question content and users' answers by introducing latent topics.
- Modeling question content can help estimate reasonable user reliability, and in turn, modeling answers leads to the discovery of meaningful topics.
- Learning topics, topic-level user expertise and truths simultaneously.

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

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



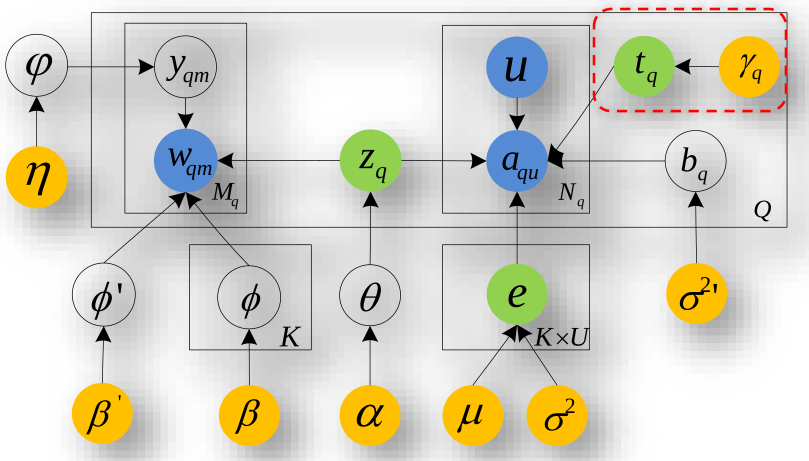
FaitCrowd

- **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

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

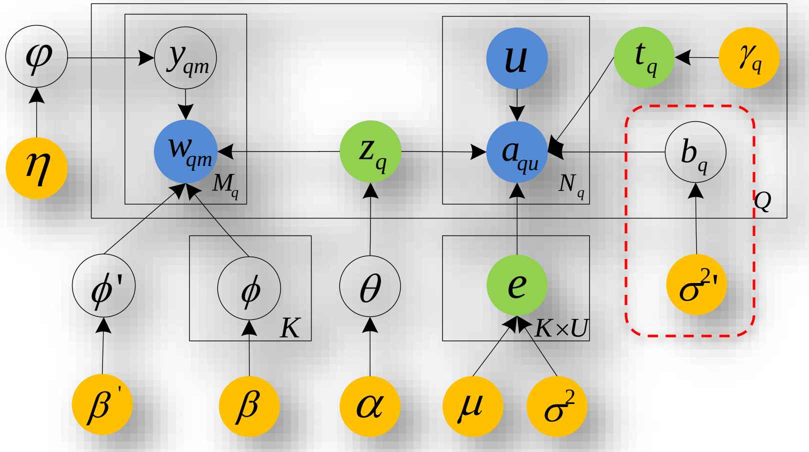
    - Draw the truth

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



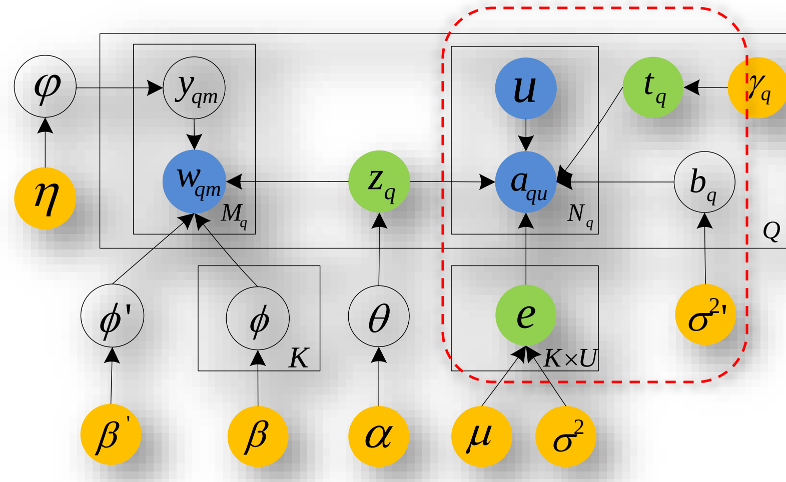
FaitCrowd

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



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
      $$b_q \sim N(0, \sigma^{2\prime})$$
    - Draw a user's answer
      $$a_{qu} | t_q \sim logistic(e_{z_q u}, b_q)$$



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

Overview

1. Introduction
2. Comparison of Existing Truth Discovery and Crowdsourced Data Aggregation Setting
3. Models of Truth Discovery and Crowdsourced Data Aggregation
4. Truth Discovery for Crowdsourced Data Aggregation
5. **Related Areas**
6. Open Questions and Resources
7. References

Related Areas

- Information integration and data cleaning
  - Data fusion and data integration
    - schema mapping
    - entity resolution
    - ➢ They can be deemed as the pre-step of Truth Discovery
  - Sensor data fusion
    - ➢ Difference: the sources are treated indistinguishably
  - Data cleaning
    - ➢ Difference: single source VS multi-source

Related Areas

- Active Crowdsourcing
  - Designing of crowdsourcing applications
  - Designing of platforms
  - Budget allocation
  - Pricing mechanisms

Related Areas

- Ensemble learning
  - Integrate different machine learning models
  - ➢ Difference: supervised VS unsupervised
- Meta analysis
  - Integrate different lab studies
  - ➢ Difference: weights are calculated based on sample size
- Information trustworthiness analysis
  - Rumor detection
  - Trust propagation
  - ➢ Difference: input may contain link information or features extracted from data

Overview

Open Questions

- Data with complex relations
  - Spatial and temporal
- Evaluation and theoretical analysis
- Information propagation
- Privacy preserving truth discovery
- Applications
  - Health-oriented community question answering

Health-Oriented Community Question Answering Systems

Quality of Question-Answer Thread

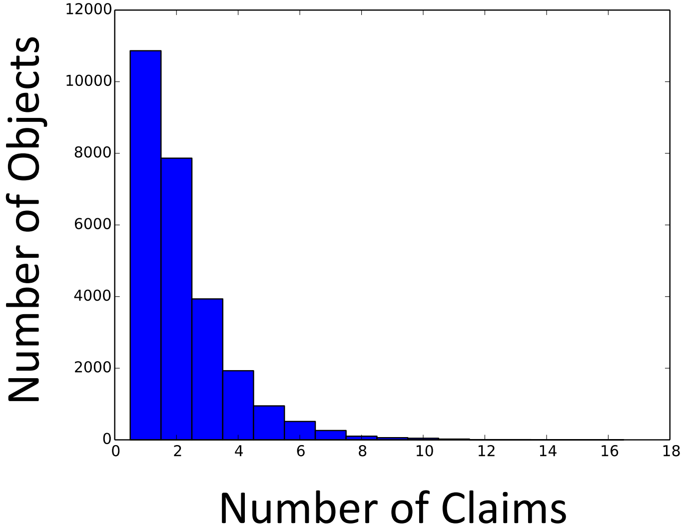
Impact of Medical Truth Discovery

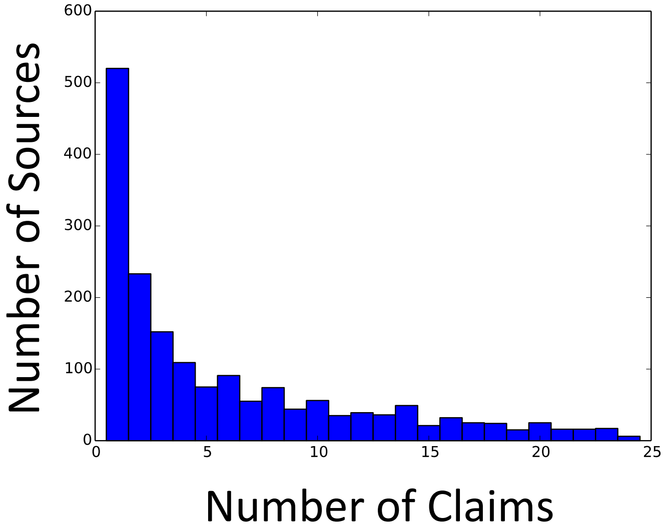
Challenge (1): Noisy Input

- Raw textual data, unstructured
- Error introduced by extractor
- New data type: textual data

Challenge (2): Long-tail Phenomenon

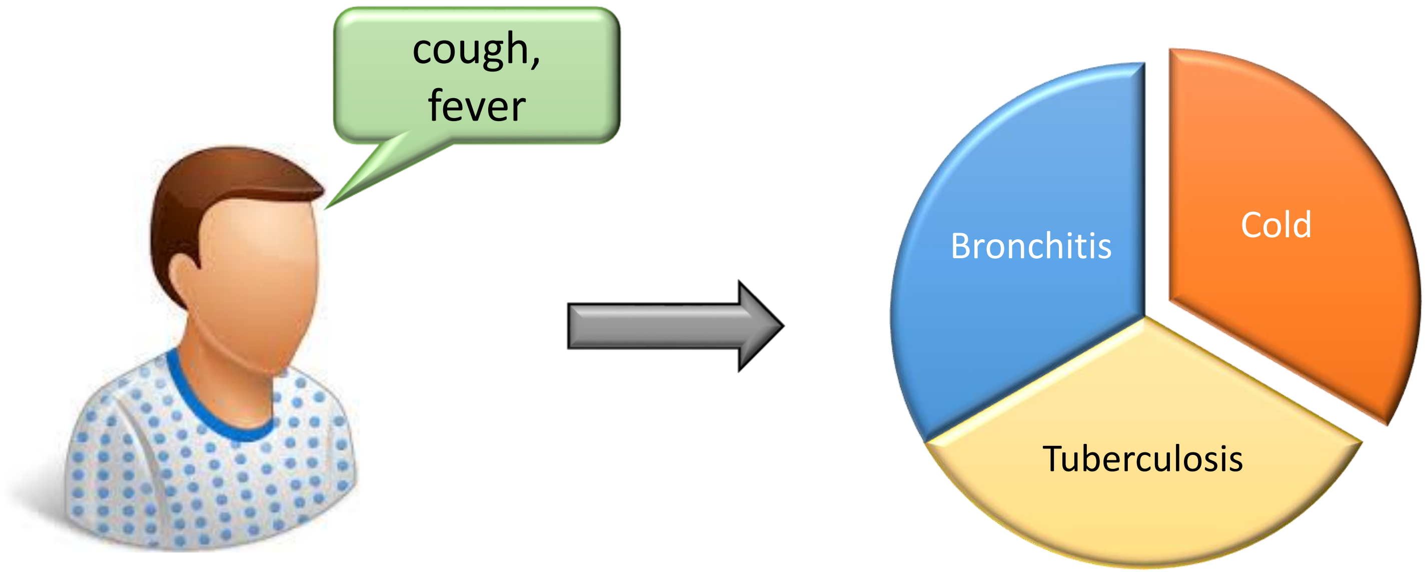
- Long-tail on source side
  - Each object still receives enough information.
- Long-tail on both object and source sides
  - Most of objects receive few information.





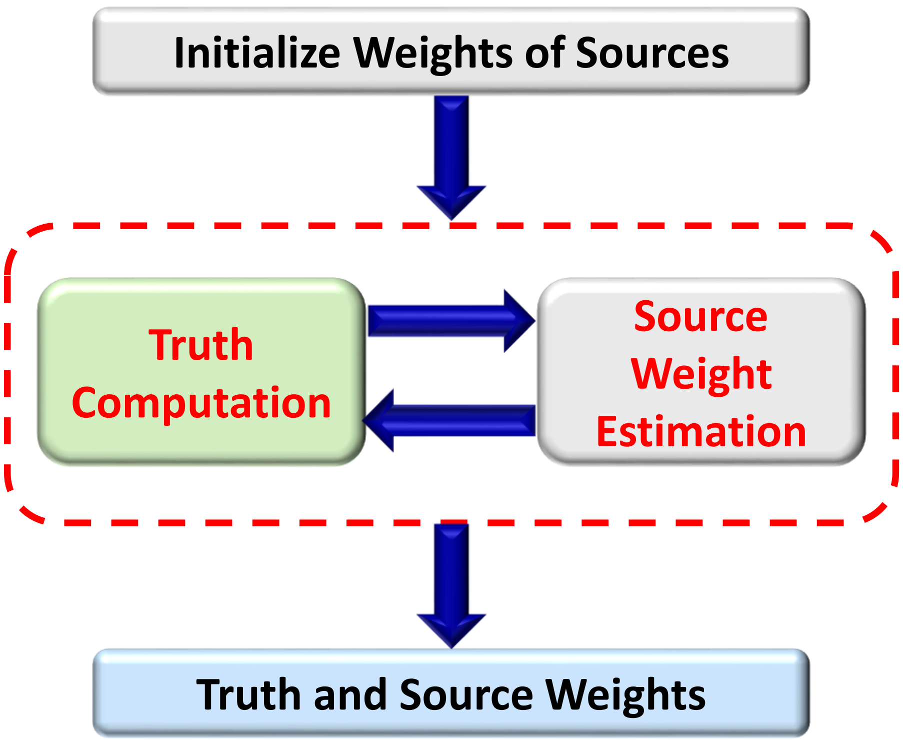
Challenge (3): Multiple Linked Truths

- Truths can be multiple, and they are linked with each other.



Challenge (4): Efficiency Issue

- Truth Discovery
  - iterative procedure

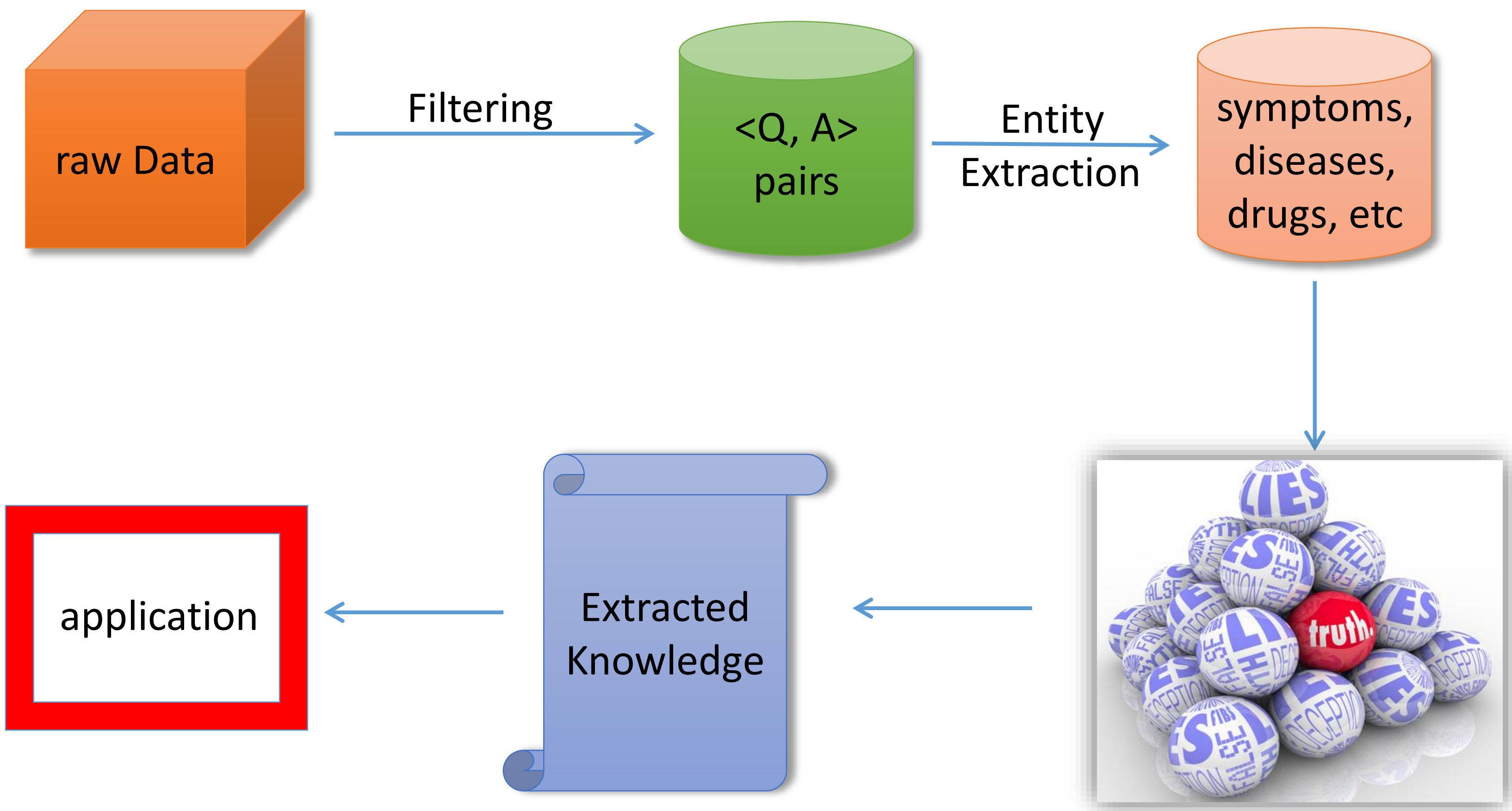


- Medical QA
  - large-scale data

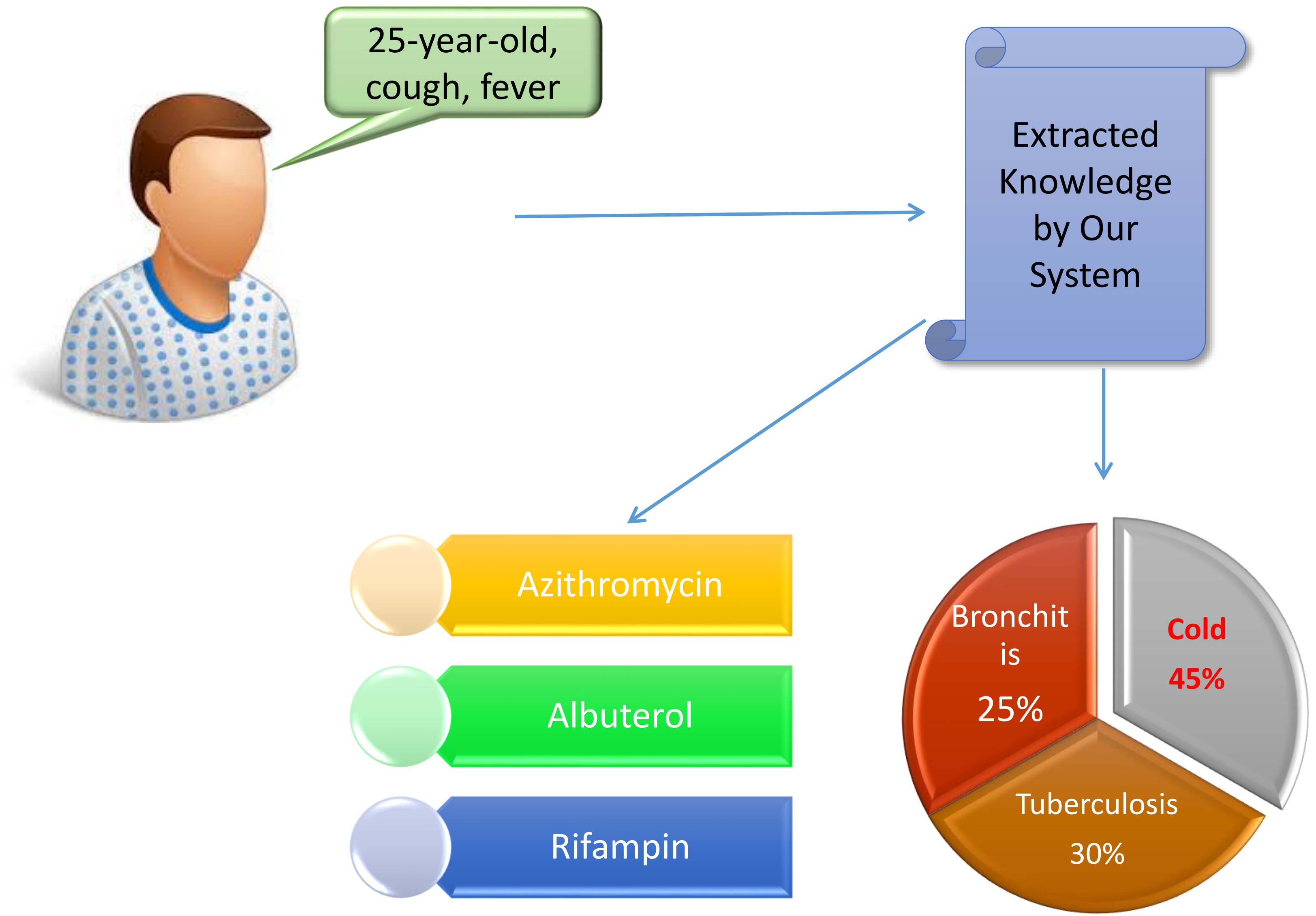
One Chinese Medical Q&A forum:

- millions of registered patients
- hundreds of thousands of doctors
- thousands of new questions per day

Overview of Our System



Preliminary Result: Example



Available Resources

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

Available Resources

- Truth discovery data and code
  - 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 code
  - 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

- These slides are available at

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

References

[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.

[Snow et al., 2008] 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.

[Zhang et al., 2014] 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.

[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.

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

[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.

[Lakkaraju et al., 2015] 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.

[Yin et al., 2008] 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.

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

[Whitehill et al., 2009] 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.

[Zhou et al., 2012] 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.

[Li et al., 2014] 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.

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

[Dong et al., 2009b] X. L. Dong, L. Berti-Equille,and D. Srivastava. Truth discovery and copying detection in a dynamic world. *PVLDB,* pages 550–561, 2009.

[Pochampally et al., 2014] 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.

[Li et al., 2015a] 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), 2015.

[Mukherjee et al., 2014] 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.

[Ma et al., 2015] 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.

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

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

[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., 2015] 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.

[Li et al., 2015b] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *arXiv preprint arXiv:1505.02463*, 2015.

[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.

[Sheshadri&Lease, 2013] 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.