華東師範大學
EAST CHINA NORMAL UNIVERSITY

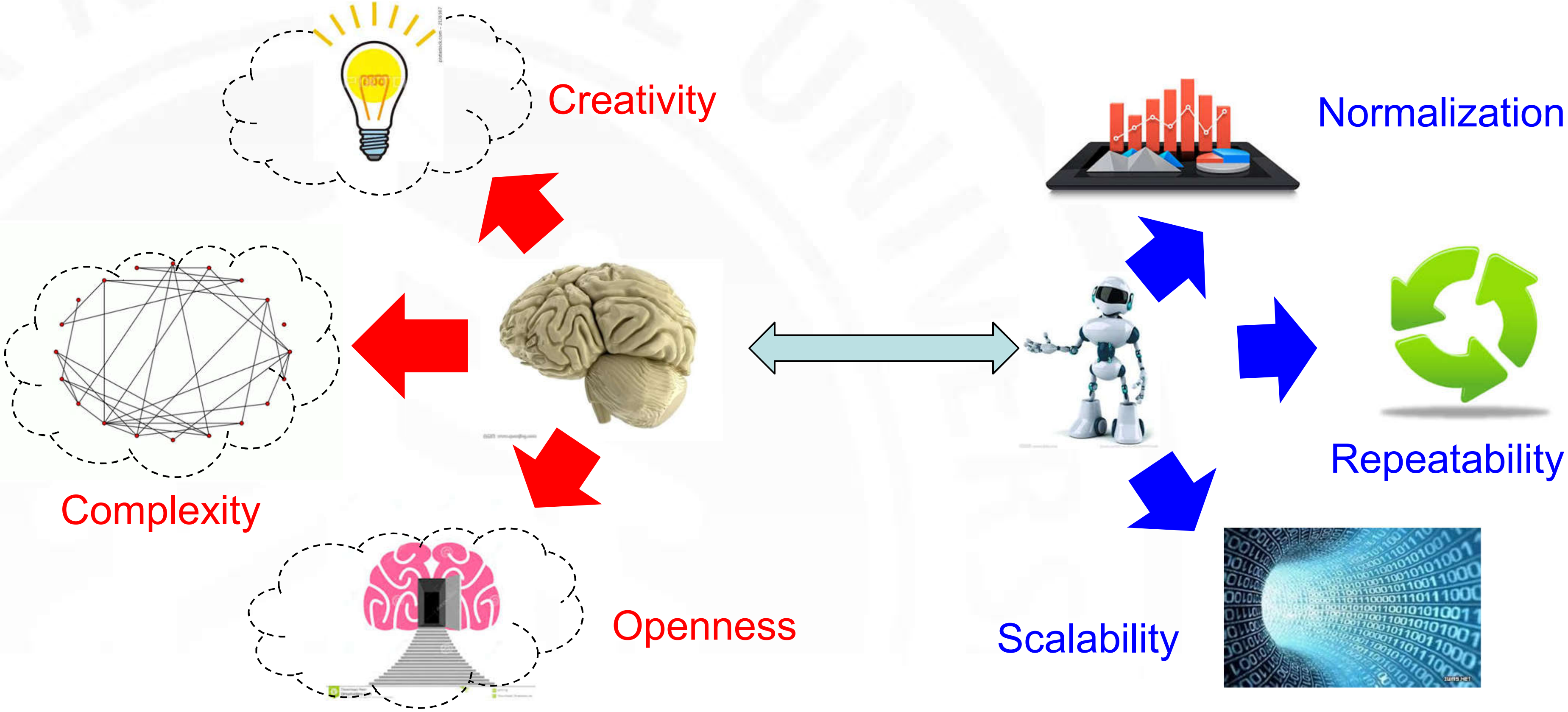
KG Refinement by Knowledge Intensive Crowdsourcing

Xin Lin
CS@Ecnu
xlin@cs.encu.edu.cn

KG Refinement

- Imperfect Data-driven KG Construction
  - Accuracy is not high enough
  - Recall is not high enough
- KG Refinement
  - Auto reasoning
  - Conflict resolution
  - Crowdsourcing

Human Brain and AI



- Human brains may help AI

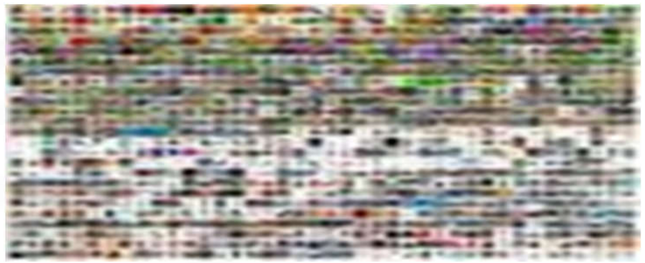
Knowledge-Intensive Crowdsourcing (KIC)

- A branch of crowdsourcing
- To achieve some knowledge-intensive task
- To bridge the gap between AI and human brain

Knowledge-Intensive Crowdsourcing

- Successful applications

CAPTCHAs



ImageNet Labeling

amazon
mechanical turk beta

Amazon MTurk

WIKIPEDIA

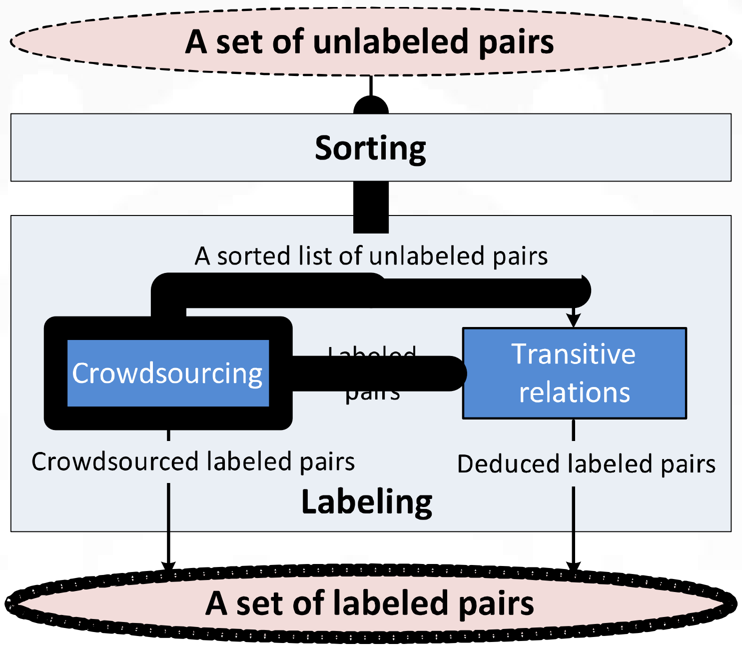
Issues on KIC

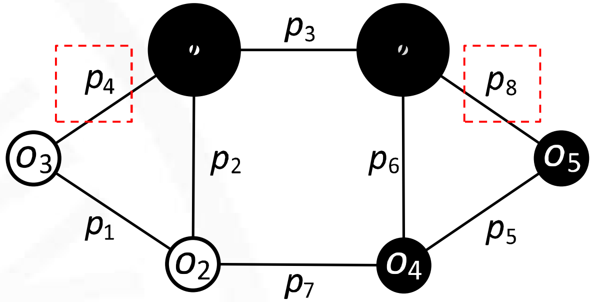
- What
  - to crowdsource?
- Whom
  - to crowdsource?
- How
  - to devise question?
  - to incentivize worker?
  - to control quality?
  - to utilize the crowdsourcing result

What

- Task selection
  - To save monetary and time cost
  - Select the most important task
  - Select the task the human is good at but the computer is not
- Existing work
  - Entity resolution[SIGMOD13] [ICDE15]
  - Schema matching[VLDB13]

Entity Resolution [SIGMOD13]

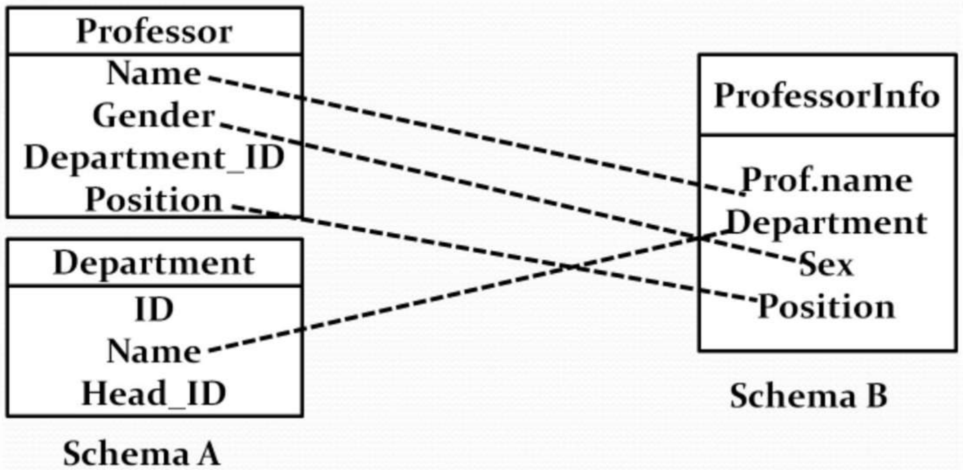




ID	Object
o_1	iPhone 2nd Gen
o_2	iPhone Two
o_3	iPhone 2
o_4	iPad Two
o_5	iPad 2
o_6	iPad 3rd Gen

ID	Object Pairs	Likelihood
p_1	(o_2, o_3)	0.85
p_2	(o_1, o_2)	0.75
p_3	(o_1, o_6)	0.72
p_4	(o_1, o_3)	0.65
p_5	(o_4, o_5)	0.55
p_6	(o_4, o_6)	0.48
p_7	(o_2, o_4)	0.45
p_8	(o_5, o_6)	0.42

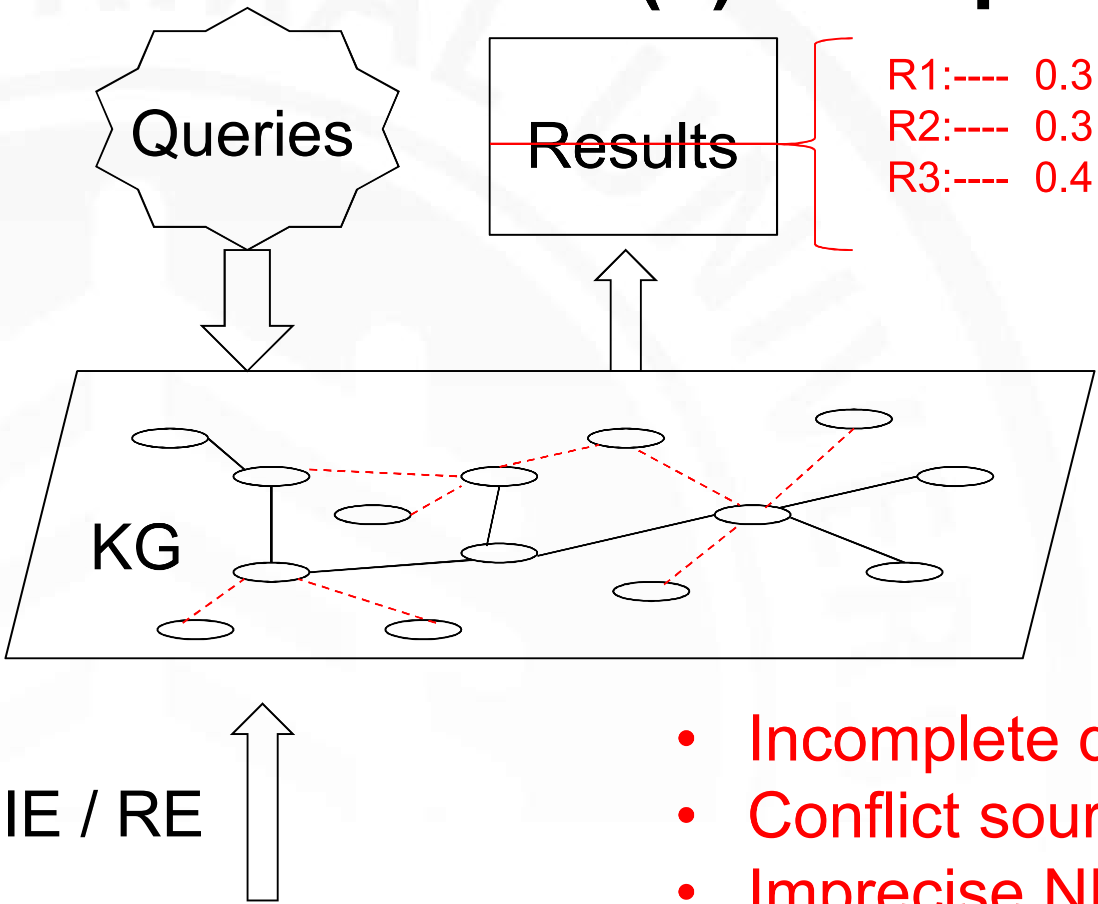
Schema Matching [VLDB13]



Correspondence	probability
c_1=<(Professor)Name,Prof.name>	.75
c_2=<Position, Position>	.7
c_3=<Gender,Sex >	1
c_4=<(Department) Name, Department>	.75
c_5=<(Department)Name,Prof.name>	.25

Possible Matchings	probability
m_1={ <(Professor)Name,Prof.name>, <Position, Position>, <Gender,Sex>, <(Department) Name, Department>}	.45
m_2={ <(Professor)Name,Prof.name>, <Gender, Sex>, <(Department) Name, Department>}	.3
m_3={ ((Department)Name,Prof.name), (Position, Position) (Gender,Sex) }	.25

What——Our work (1)：Graph Cleaning



R1:---- 0.3
R2:---- 0.3
R3:---- 0.4

Ambiguous

Could we make the results less ambiguous?

Open IE / RE

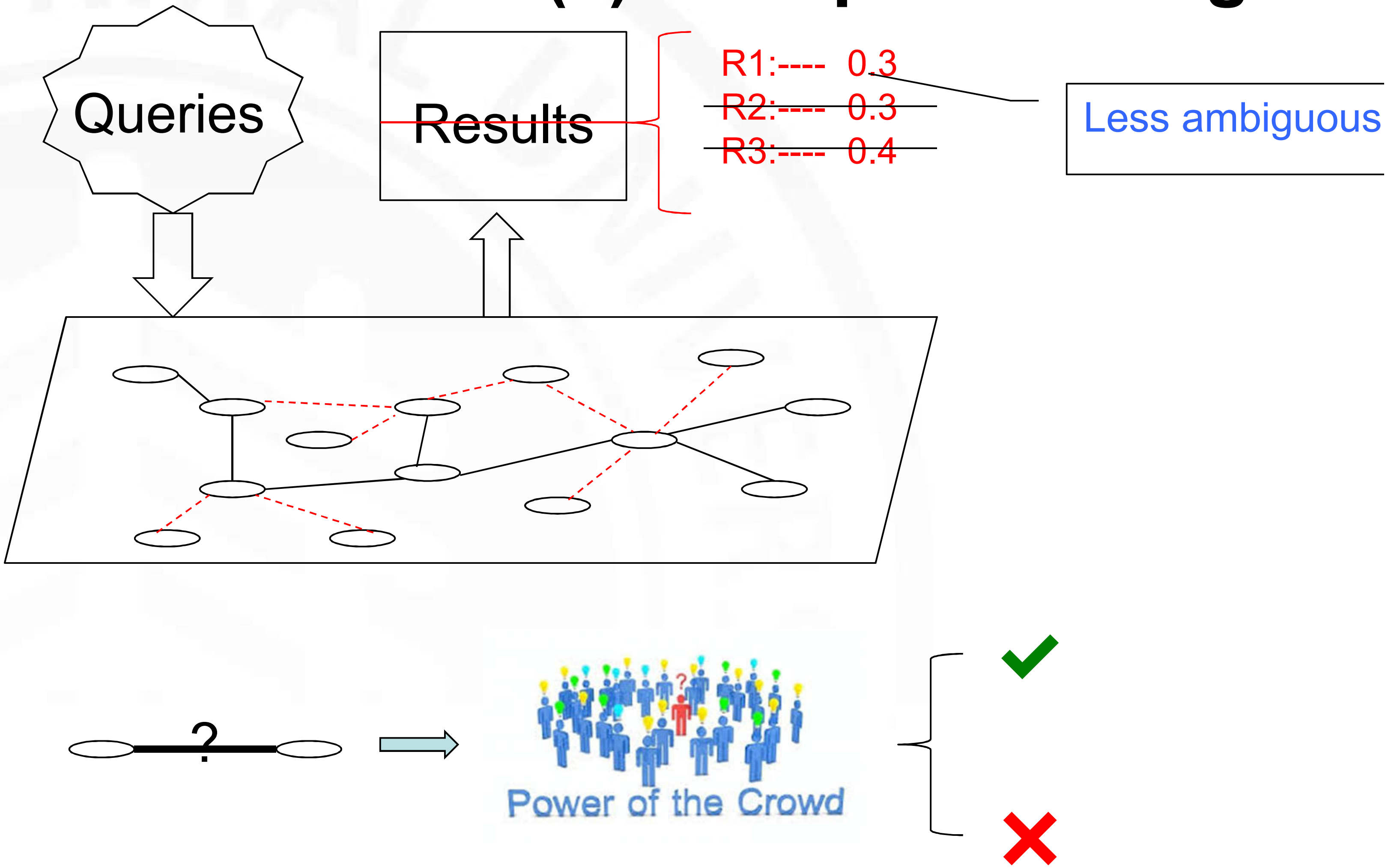
Internet Documents

- Incomplete docs
- Conflict sources
- Imprecise NLP

------------ Uncertain Relationship

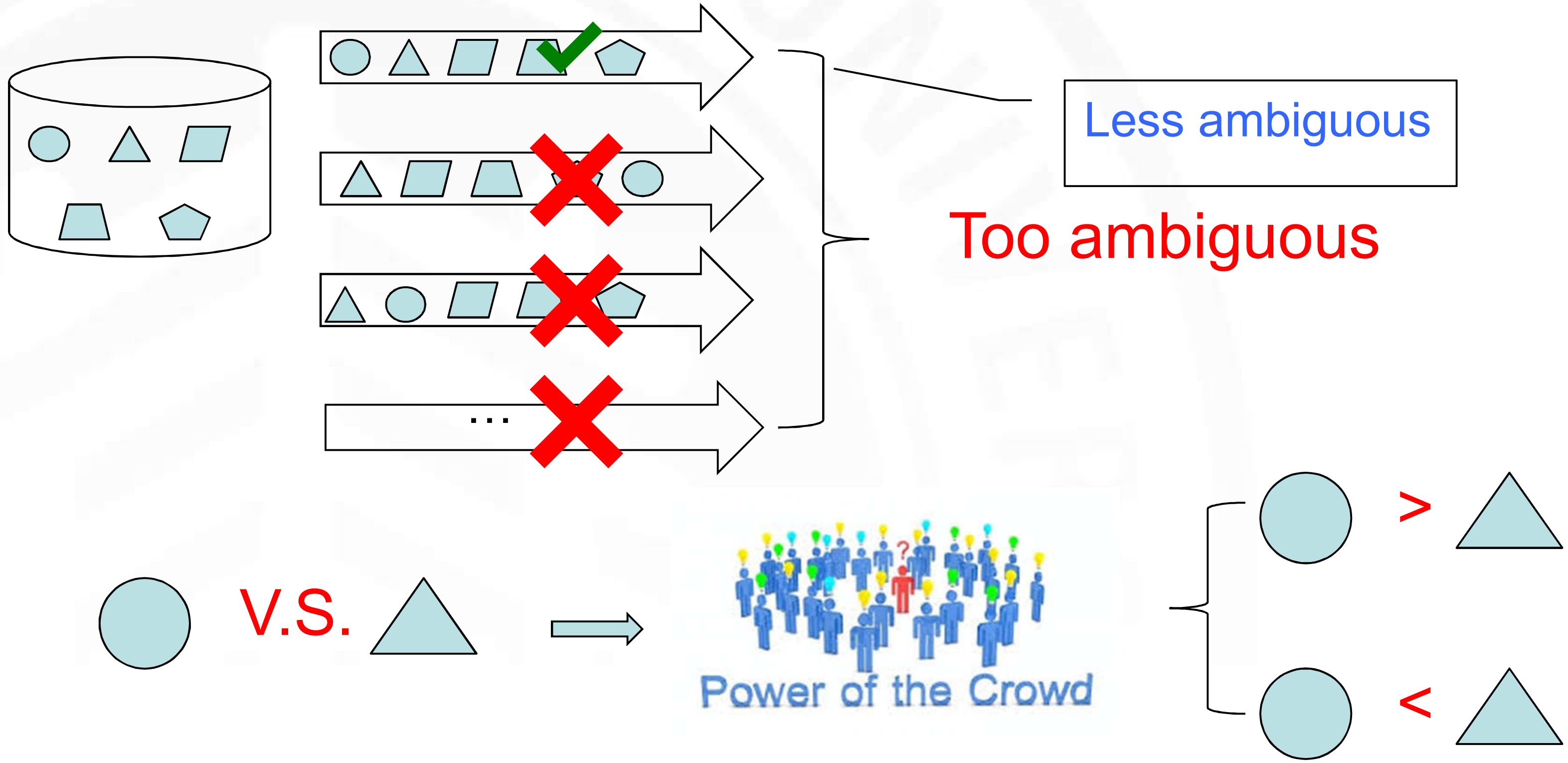
X.Lin, et.al, *Human-Powered Data Cleaning for Probabilistic Reachability Queries on Uncertain Graphs*, TKDE, 2017.

What——Our work (1): Graph Cleaning



X.Lin, et.al, *Human-Powered Data Cleaning for Probabilistic Reachability Queries on Uncertain Graphs*, TKDE, 2017.

What——Our work(2) : Pairwise Top-k cleaning



X.Lin, et.al, *Reducing Uncertainty of Probabilistic Top-k Ranking via Pairwise Crowdsourcing*, TKDE, 2017.

華東師範大學
EAST CHINA NORMAL UNIVERSITY

Summaries of issue "what"

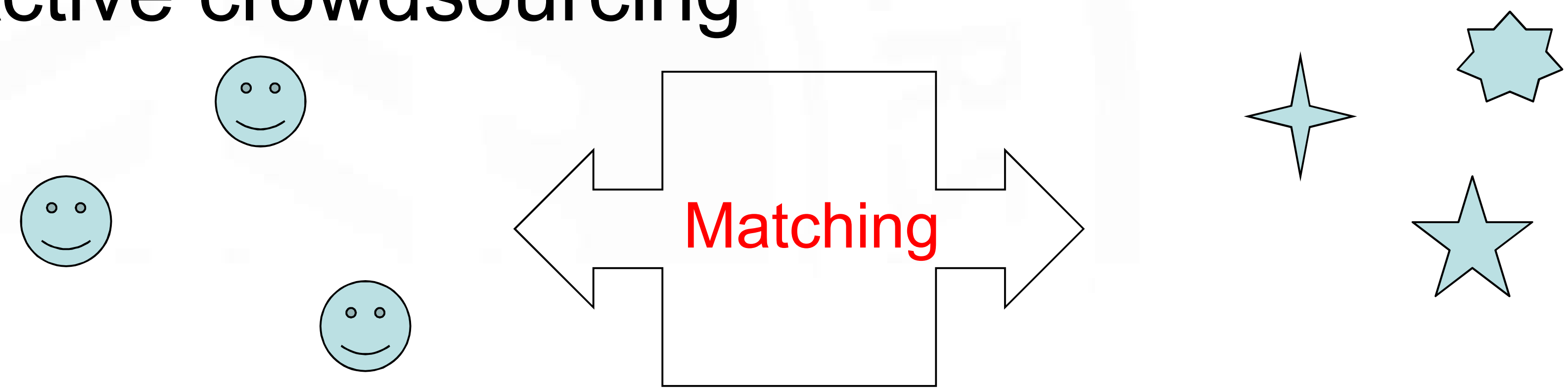
- Local refinement will promote the global quality
- Quantifying the influence is the key issue
- Task independent

Issues on KIC

- What
  - to crowdsource?
- Whom
  - to crowdsource?
- How
  - to devise question?
  - to incentivize worker?
  - to control quality?
  - to utilize the crowdsourcing result

Whom

- Passive crowdsourcing
  - All tasks are *picked up* by the workers
  - Workers are qualified by some golden tasks.
- Active crowdsourcing



User Modeling

Task Modeling

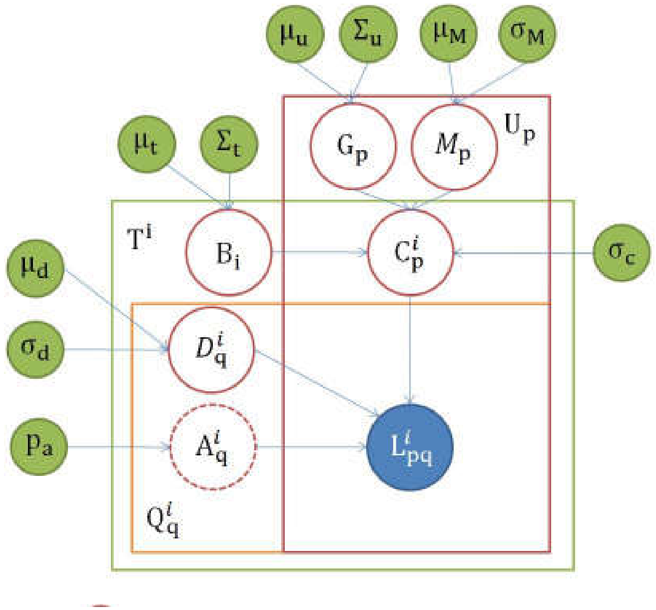
Whom: Active crowdsourcing

- User Modeling
  - Task-history-based modeling
  - Cold start problem
    - Golden task
    - Transfer learning [KDD13b]
- Matching
  - Keyword based
  - Tree based [WWW 16]
  - Vector based [VLDB 16]

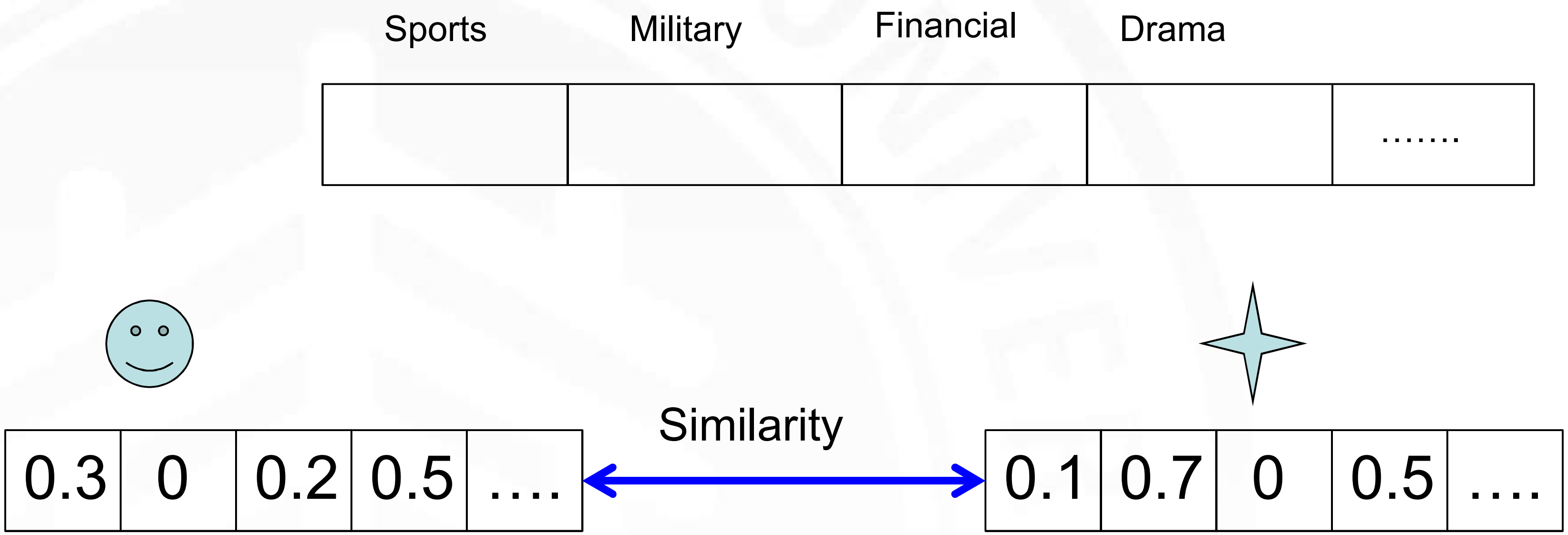
Whom: Active crowdsourcing

- **Task Assignment**
  - Randomly selected
  - Consider other factors (time, worker's quality,etc)
    - Assign the k most uncertain tasks[ICDE 12]
    - Choose the k highest quality workers[SIGMOD 15a]
    - Choose the highest improvement in quality [SIGMOD 15a]
    - …

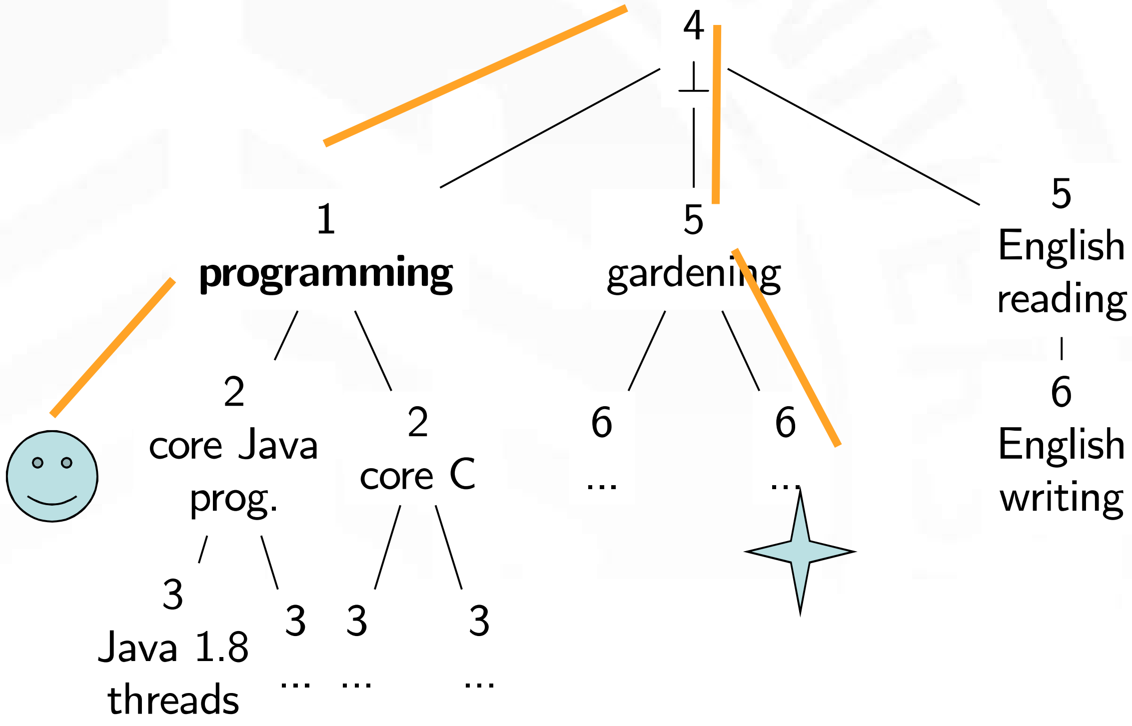
Transfer Learning in Worker Modeling [KDD2013]



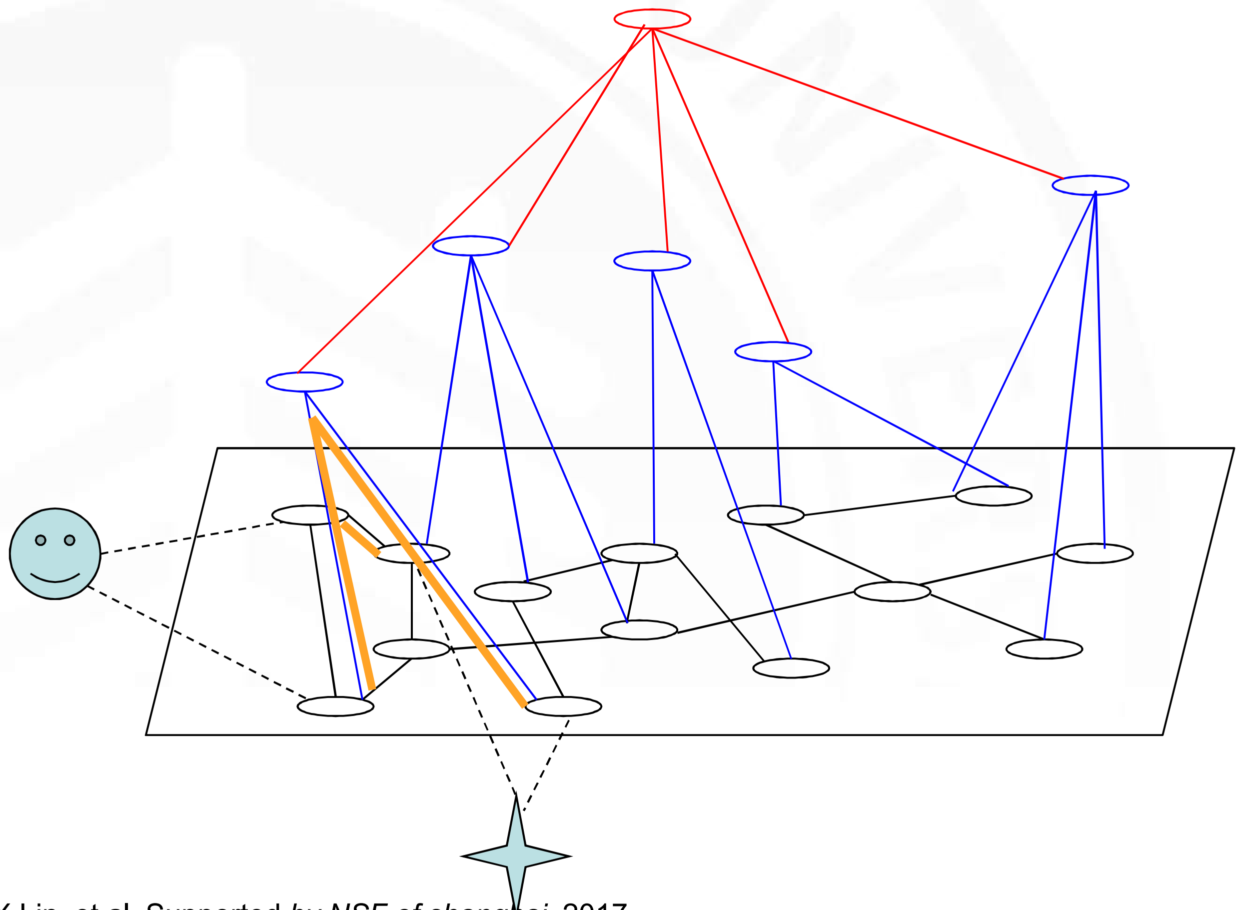
Domain-based matching [VLDB2016]



Tree-based matching [WWW16]



Whom—Our work: Graph+Tree-based



X.Lin, et.al, Supported *by NSF of shanghai*, 2017.

Issues on KIC

- What
  - to crowdsource?
- Whom
  - to crowdsource?
- How
  - to devise question?
  - to incentivize worker?
  - to control quality?
  - to utilize the crowdsourcing result

How to devise question?

- Explicit crowdsourcing
- Implicit crowdsourcing

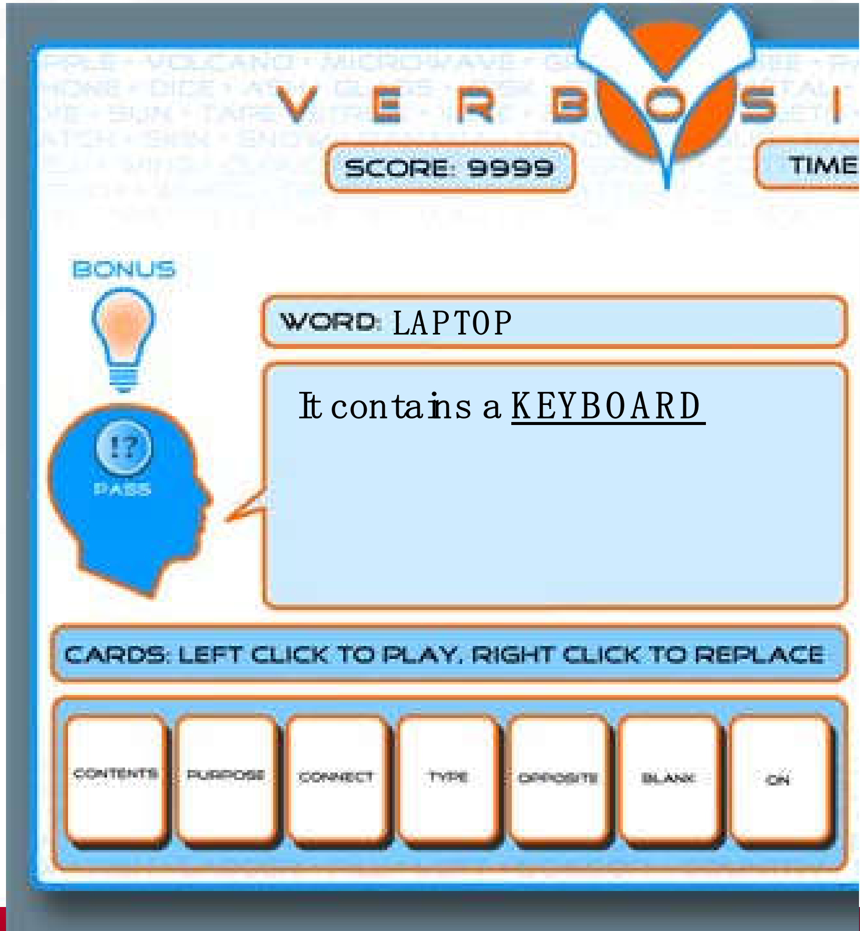
Devise questions

- Explicit crowdsourcing
  - Traditional guidelines:
    - 1. Small piece of task is preferred
    - 2. Yes-or-No > Choice >Blank filling
    - 3. Less cooperation is preferred
    - 4. Good UI is preferred
  - New research points:
    - Should tradeoff the cost and accuracy
      - Mix multi-choice and Yes-or-no [SIGMOD 17]
    - Should devise the workflow of Crowdsourcing

Devise questions

- Implicit crowdsourcing
  - **Gamification**
    - Common sense knowledge acquisition[CHI06]
    - Spatial Positions[AIIDE 14]
  - **Collecting Secretly**
    - CAPTCHAS
    - Auto Image Annotation [MTA 14]
    - Visual Focus [TMM14]
  - **Make Use of Psychological Characteristic**
    - Curiosity[CHI16]
    - Micro-diversions[CSCW 15]

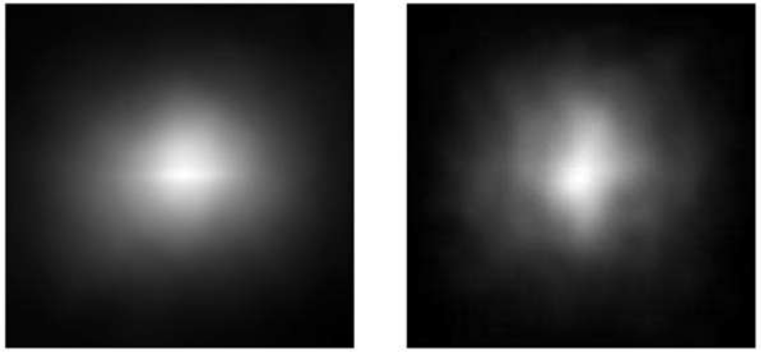
Common knowledge acquisition



Templates:

- ___ is a kind of ___.
- ___ is used for ___.
- ___ is typically near/in/on ___ .
- ___ is the opposite of ___ / ___ is related to ___ .

Touch Saliency & Visual Focus



Touch Visual

Heat Map Touch

Heat Map Visual

Implicit crowdsourcing

- Guidance of implicit crowdsourcing
  - Provide the task unconsciously
  - Workers are Users
  - First purpose should match user's demands, while second purpose should match the crowdsourced task.
  - First purpose is always the most important.
  - Motivate the crowds with Curiosity

Issues on KIC

- What
  - to crowdsource?
- Whom
  - to crowdsource?
- How
  - to devise question?
  - to incentivize workers?
  - to control quality?
  - to utilize the crowdsourcing result

Taxonomy of incentives

- Incentives of crowdsourcing
  - Reputation
    - Rewards
    - Penalty
  - Joy
    - Game
  - Financial Rewards
  - Social Influence

Taxonomy of incentives

- Financial rewards
  - Static rewards
  - Dynamic rewards
  - Conditional rewards
    - Competition
    - By accuracy
    - By agreement

Taxonomy of incentives

Social Influence

Strong connection

Weak connection

華東師範大學
EAST CHINA NORMAL UNIVERSITY

Our works

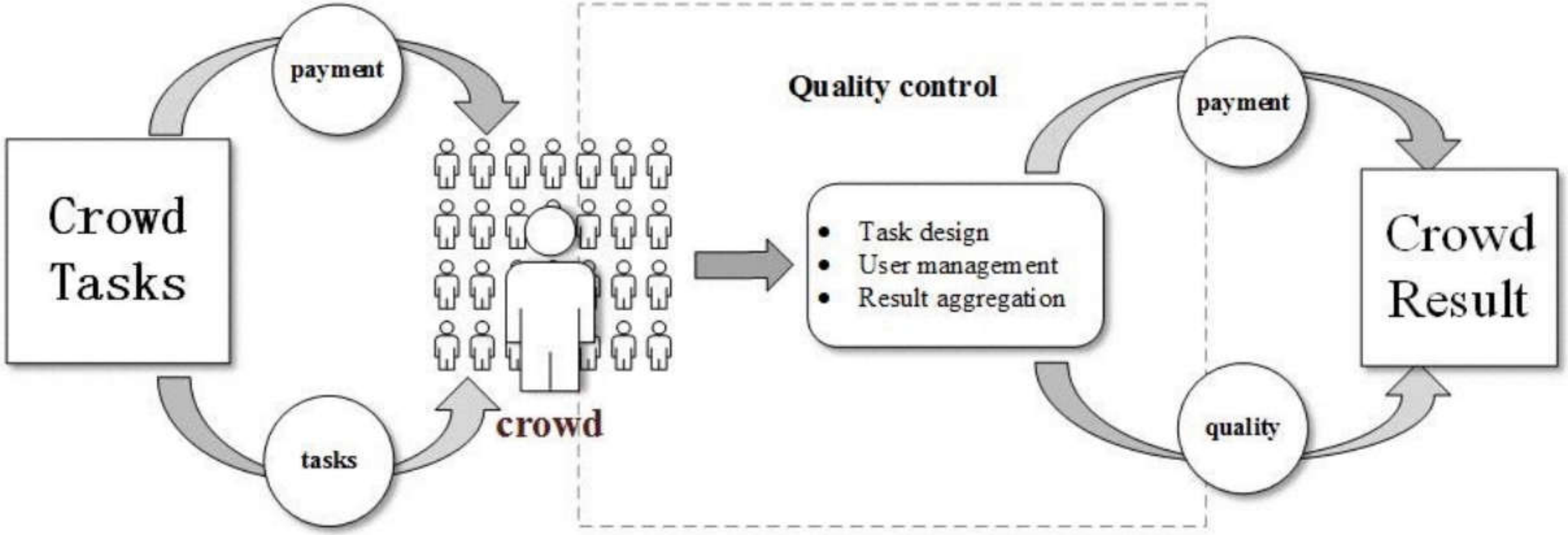
- 1. Weak connection performance better than strong connection for short-term tasks
- 2. Hybrid incentive in different phrases

Issues on KIC

- What
  - to crowdsource?
- Whom
  - to crowdsource?
- How
  - to devise question?
  - to incentivize workers?
  - to control quality?
  - to utilize the crowdsourcing result

Quality Control

Overview



- Task Design
- Worker Organization Model
- Result aggregation

Quality Control

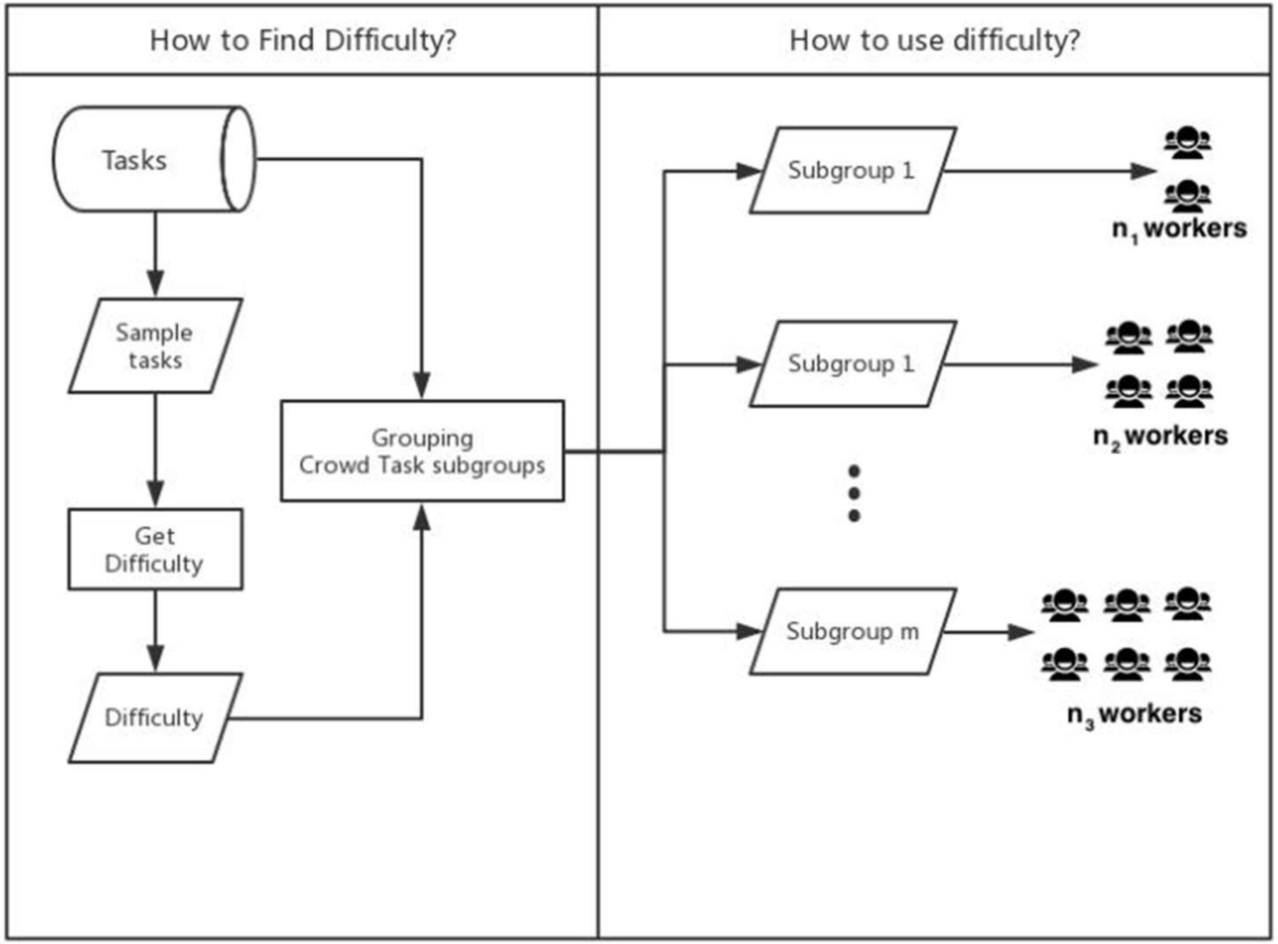
- Task design
  - Anti-malicious strategy [CHI15]
  - Add feedback mechanism[CSCW14]
- User management
  - Similar to the company management model

Quality Control

- Result Aggregation
  - Golden standard datasets
    - Dynamically insert golden tasks
    - Using golden tasks to test users
  - Redundancy-based strategy
    - Basic Majority Voting
    - Weighted Voting
  - Two-Stage strategy [KDD13a]

Our Work

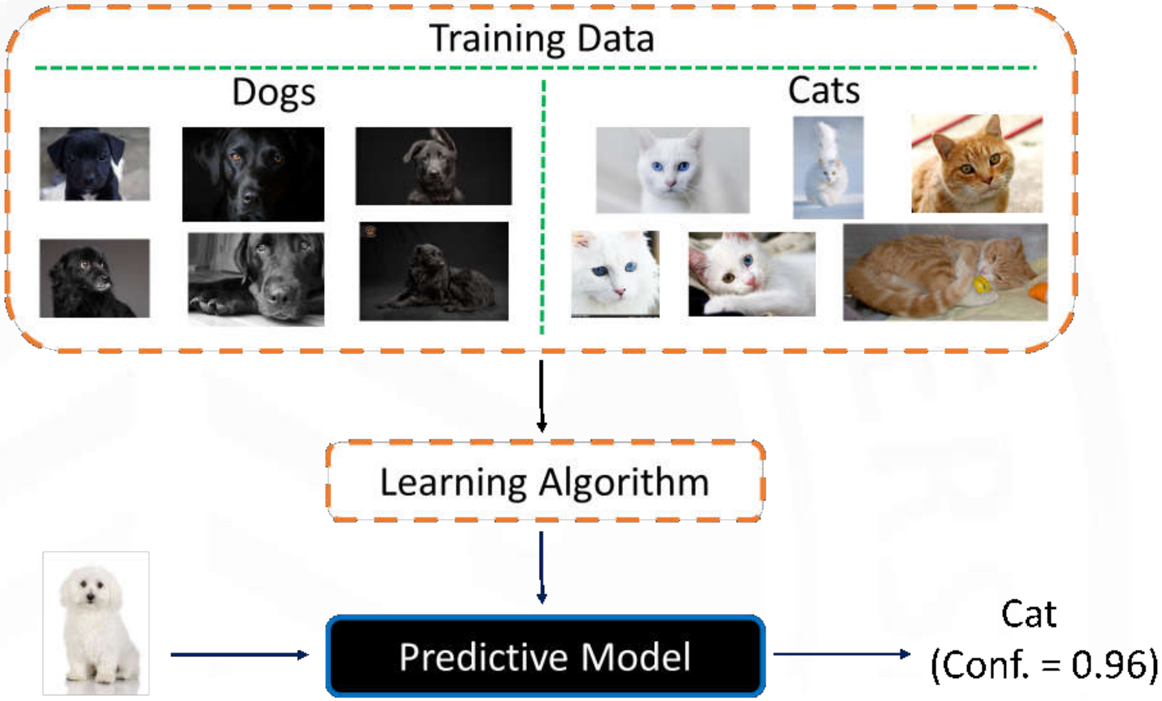
- Difficulty-based task assignment [Group 2018]



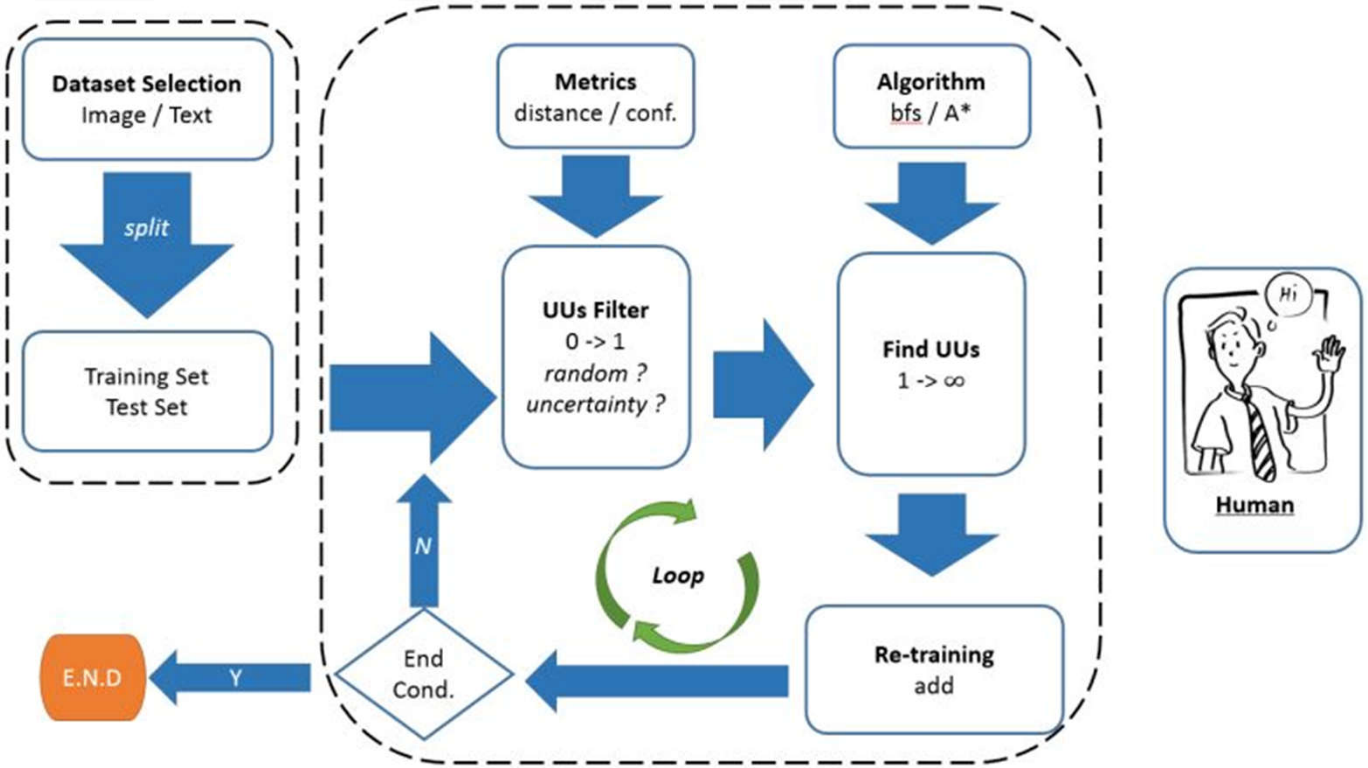
Issues on KIC

- What
  - to crowdsource?
- Whom
  - to crowdsource?
- How
  - to devise question?
  - to incentivize workers?
  - to control quality?
  - to utilize the crowdsourcing result

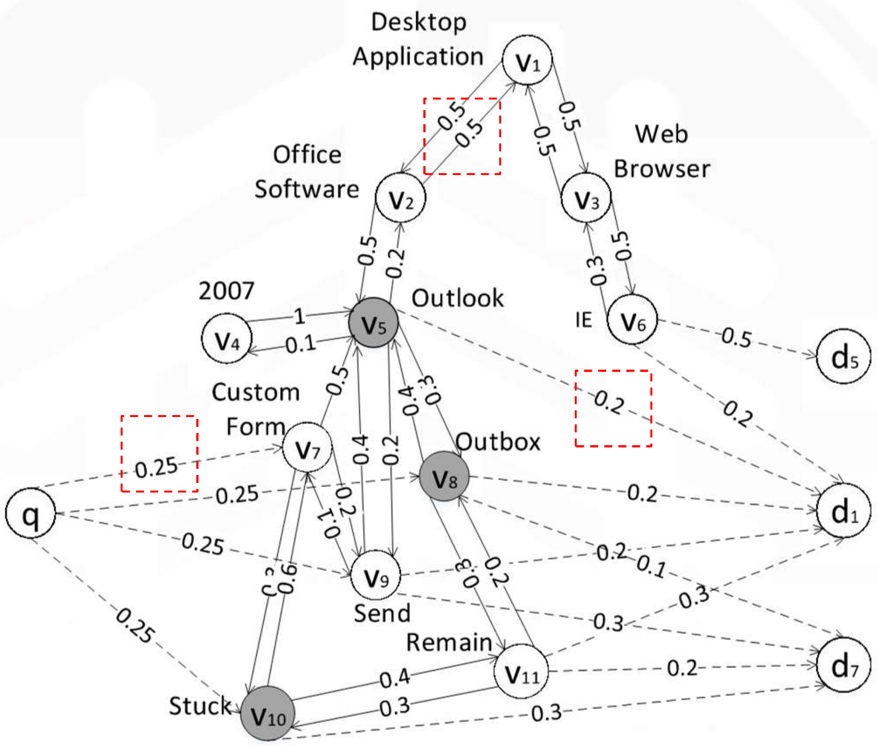
Our work (1): Finding unknown unknowns

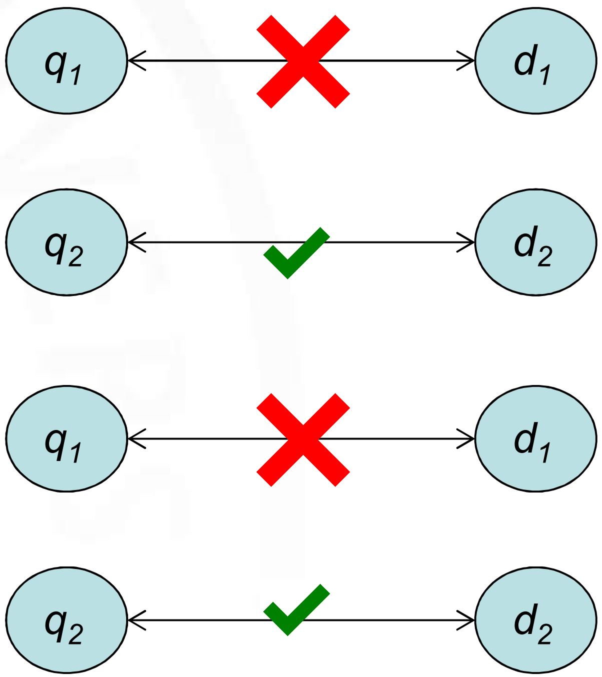


Our work (1): Finding unknown unknowns



Our work : Crafting KG via QA FeedBacks





華東師範大學
EAST CHINA NORMAL UNIVERSITY

Thank you!

References

- [SIGMOD13] J. Wang, et.al. Leveraging Transitive Relations for Crowdsourced Joins.
- [ICDE15] V. Verroios, et.al. Entity resolution with crowd errors.
- [VLDB15] C. Zhang, et.al. Reducing uncertainty of schema matching via crowdsourcing.
- [VLDB11] A. Marcus, et.al. Human-powered sorts and joins.
- [WWW16] P. Mavridis, et.al. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing.
- [VLDB16] Y. Zheng, et.al. DOCS: Domain-Aware Crowdsourcing System.
- [KDD13a] Y. Baba. Statistical Quality Estimation for General Crowdsourcing Tasks.
- [KDD13b] K.Mo. Cross-task Crowdsroucing.

References

- [ICDE12] R. Boim, et.al. Asking the right questions in crowd data sourcing.
- [SIGMOD15a] J. Fan. ICrowd: An adaptive crowdsourcing framework.
- [SIGMOD15b] Y. Zheng, et.al. Qasca: A quality-aware task assignment system for crowdsourcing applications.
- [WWW14] G. Goel, et.al. Allocating tasks to workers with matching constraints: truthful mechanisms for crowdsourcing markets.
- [SIGMOD17] V. Verroios, et.al. Waldo: An Adaptive Human Interface for Crowd Entity Resolution.
- [TMM14] B. Ni,et al. Touch Saliency: Characteristics and Prediction[J]. IEEE Transactions on Multimedia, 2014, 16(6):1779-1791.
- [AIIDE 14] R. Hodhod, et.al. Toward Generating 3D Games with the Help of Commonsense Knowledge and the Crowd.
- [MTA 14] K. Ntalianis, et al. Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution[J]. Multimedia Tools and Applications, 2014, 69(2):397-421.

References

- [CHI06] L. Ahn, et.al. Verbosity: A Game for Collecting Common-Sense Facts.
- [CHI16] E. Law, et al. Curiosity Killed the Cat, but Makes Crowdwork Better.
- [CSCW 15] P. Dai, et al. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions.
- [CSCW 14]L. Yu, et.al. A Comparison of Social, Learning, and Financial Strategies on Crowd Engagement and Output Quality.
- [CHI 15] U.Gadiraju, et.al, Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys.