Innovative Approaches to Data Curation and Retrieval-Augmented Generation: From Annotation and Preparation to Retrieval.

Yushi Sun
Supervised by Prof. Lei Chen

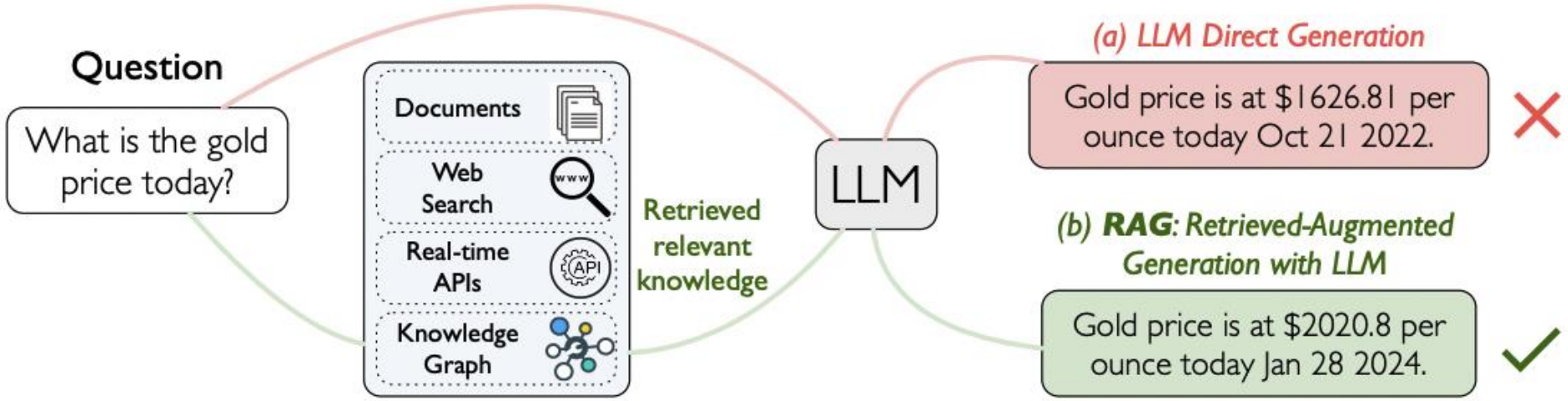
Last updated 2025/6/28

Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Preparation: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- RAG: KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering
- Future Vision and Opportunities

Background: RAG

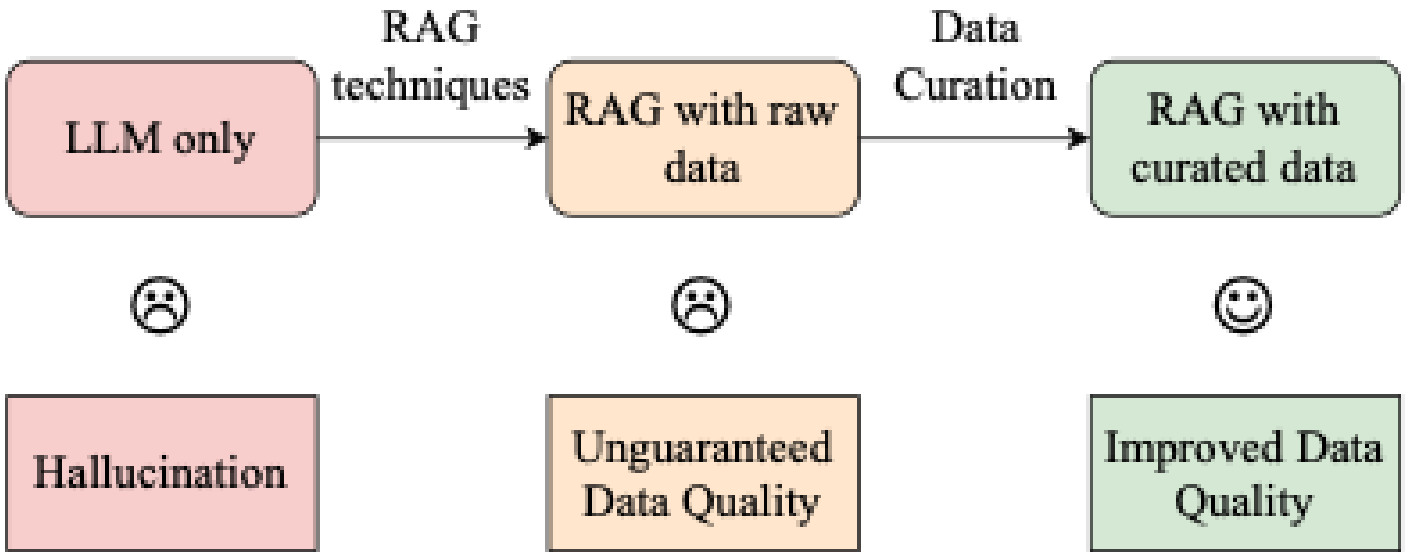
- RAG: an approach that combines retrieval of relevant information with generative capabilities of LLMs to produce more accurate and contextually relevant responses with less hallucination [1, 2].



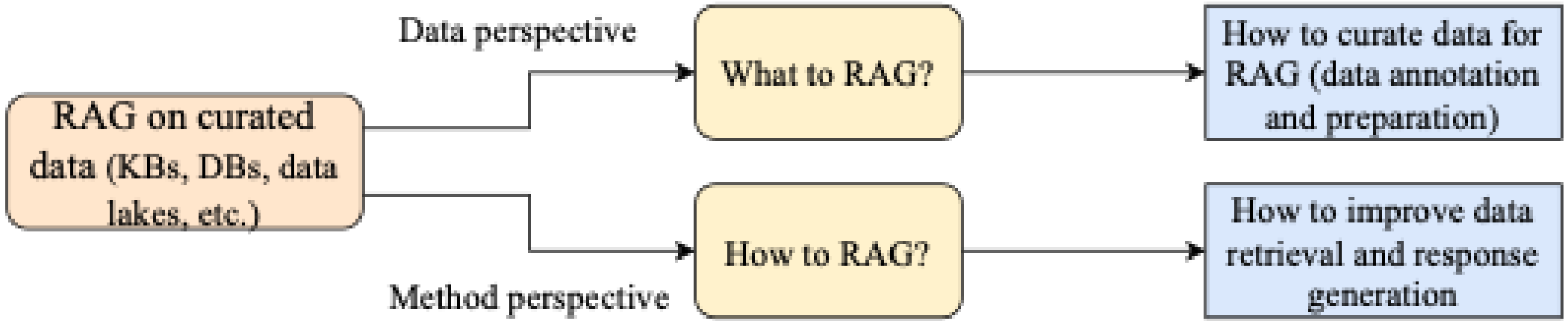
[1] Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R., Jiang, Z., Jiang, Z. and Kong, L., 2025. "CRAG-comprehensive RAG benchmark". Advances in Neural Information Processing Systems, 37, pp.10470-10490.
[2] K. Martineau, "What is retrieval-augmented generation?" Aug. 22, 2023. https://research.ibm.com/blog/retrieval-augmented-generation-RAG (accessed Feb, 24, 2025).

Background: RAG

- Data curation ensures the quality and usability of data, while RAG combines retrieval with generative models for contextually relevant responses. Effective curation improves the data RAG uses, enhancing the accuracy of its outputs.



Overview of my research



My (Co)-first-authored Publications

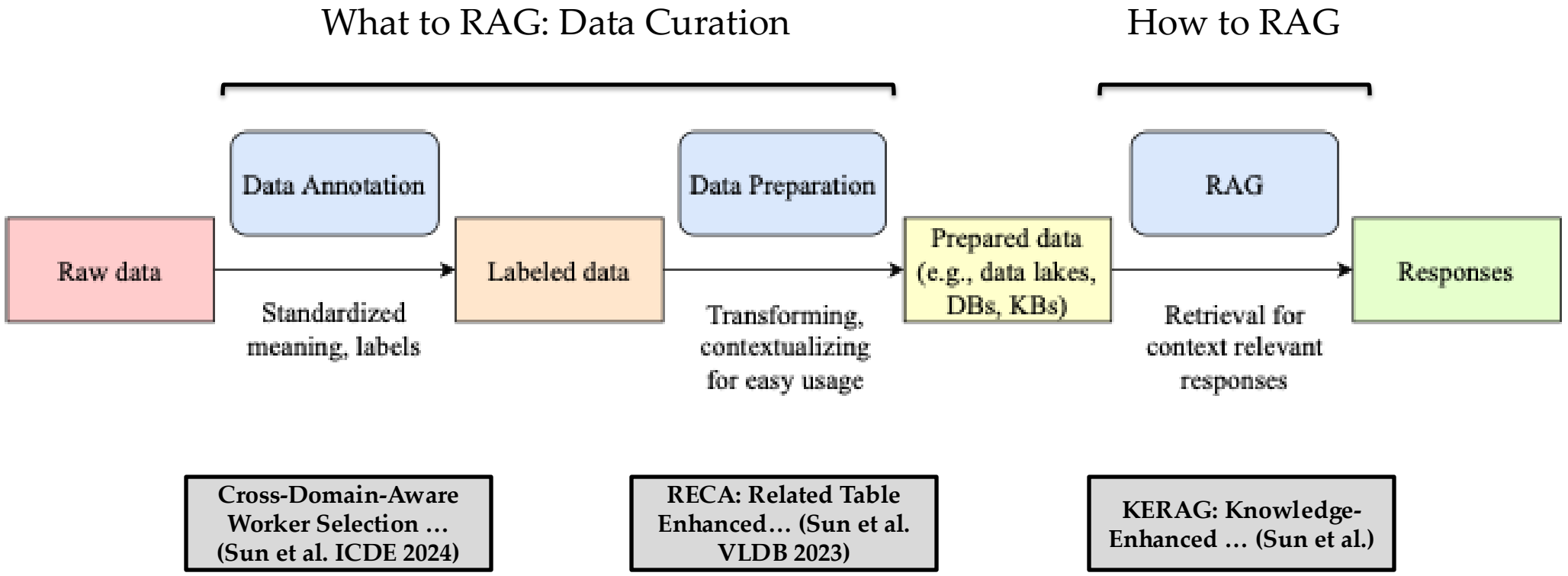
- **Cross-Domain-Aware Worker Selection with Training for Crowdsourced Annotation**, ICDE 2024.
- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework**, VLDB 2023.
- **LakeHopper: Cross Data Lakes Column Type Annotation through Model Adaptation,** Under Submission.
- **Are Large Language Models a Good Replacement of Taxonomies?**, VLDB 2024.
- **CRAG - Comprehensive RAG Benchmark*,** NeurIPS 2024 (used for hosting the KDD Cup 2024).
- **KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering,** Under Submission.

What to RAG: Data Curation

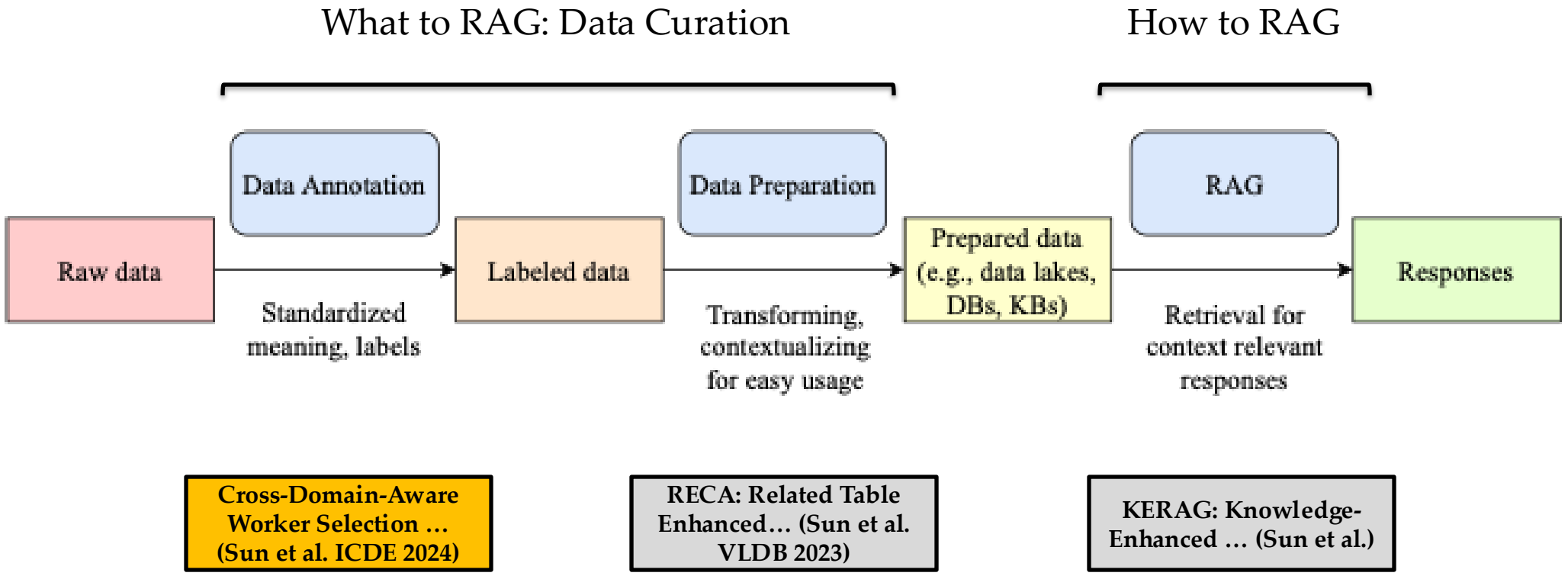
How to RAG

* Indicates co-first-authored work

Workflow – from raw data to responses



Workflow – from raw data to responses



Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Preparation: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- RAG: KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering
- Future Vision and Opportunities

Data Annotation

- Data Annotation: annotating raw data to provide standardized meaning.

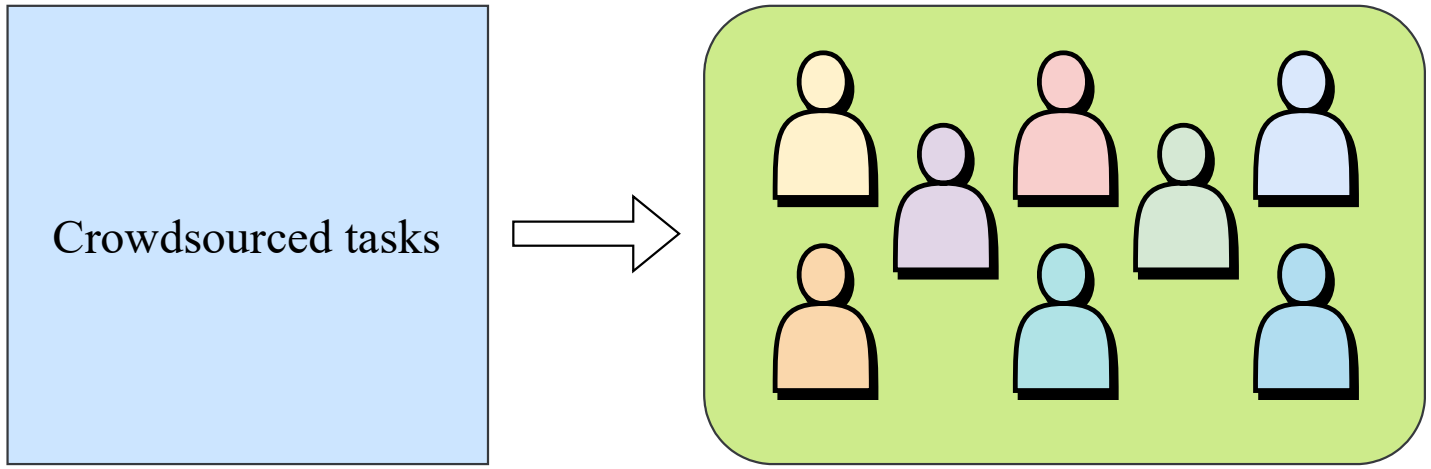
What kind of flower is shown?

petunia	?
morning glory	?
desert rose	?

- The necessity of domain knowledge and the inherent difficulties of the annotation tasks call for a novel cross-domain annotator training and selection scheme.

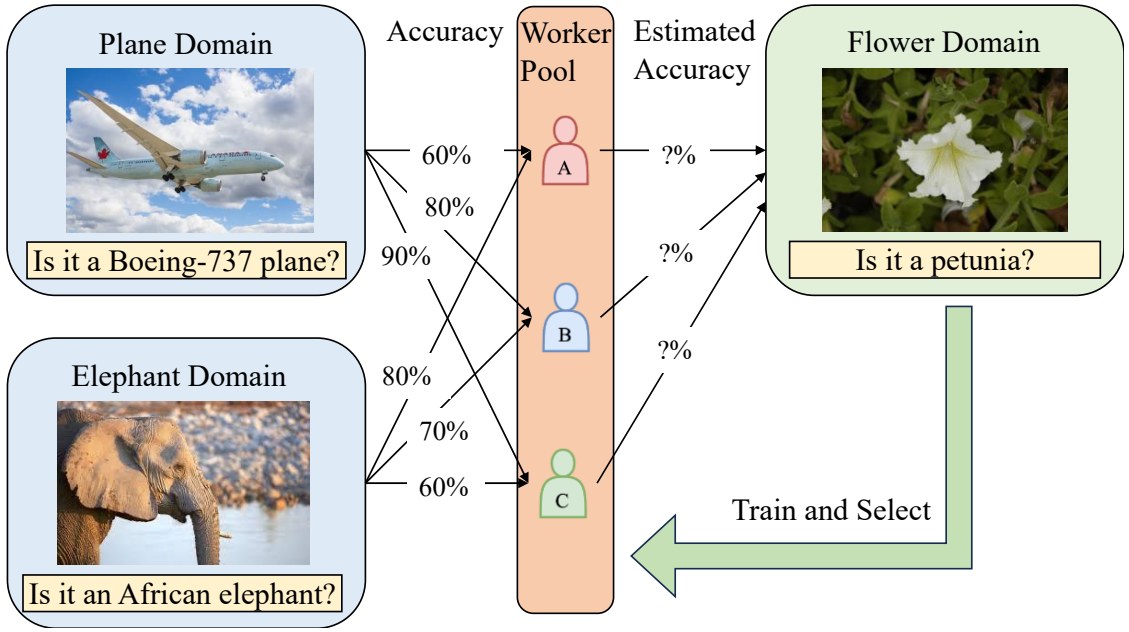
Overview

- **Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation (ICDE 2024)**
  - Crowdsourcing is preferable for obtaining high-quality data labels for large-scale datasets.
  - Worker Selection is important in Crowdsourcing.
  - How to design an allocation scheme for golden questions (questions with ground truth answers that are used for worker training/selection) to **train and select** high-performance crowd workers for the incoming crowdsourced tasks remains a challenge.

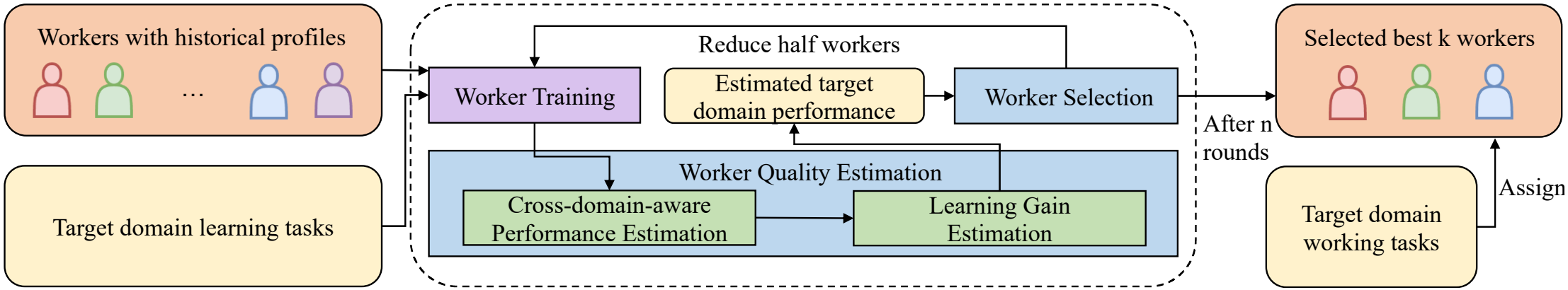


Background

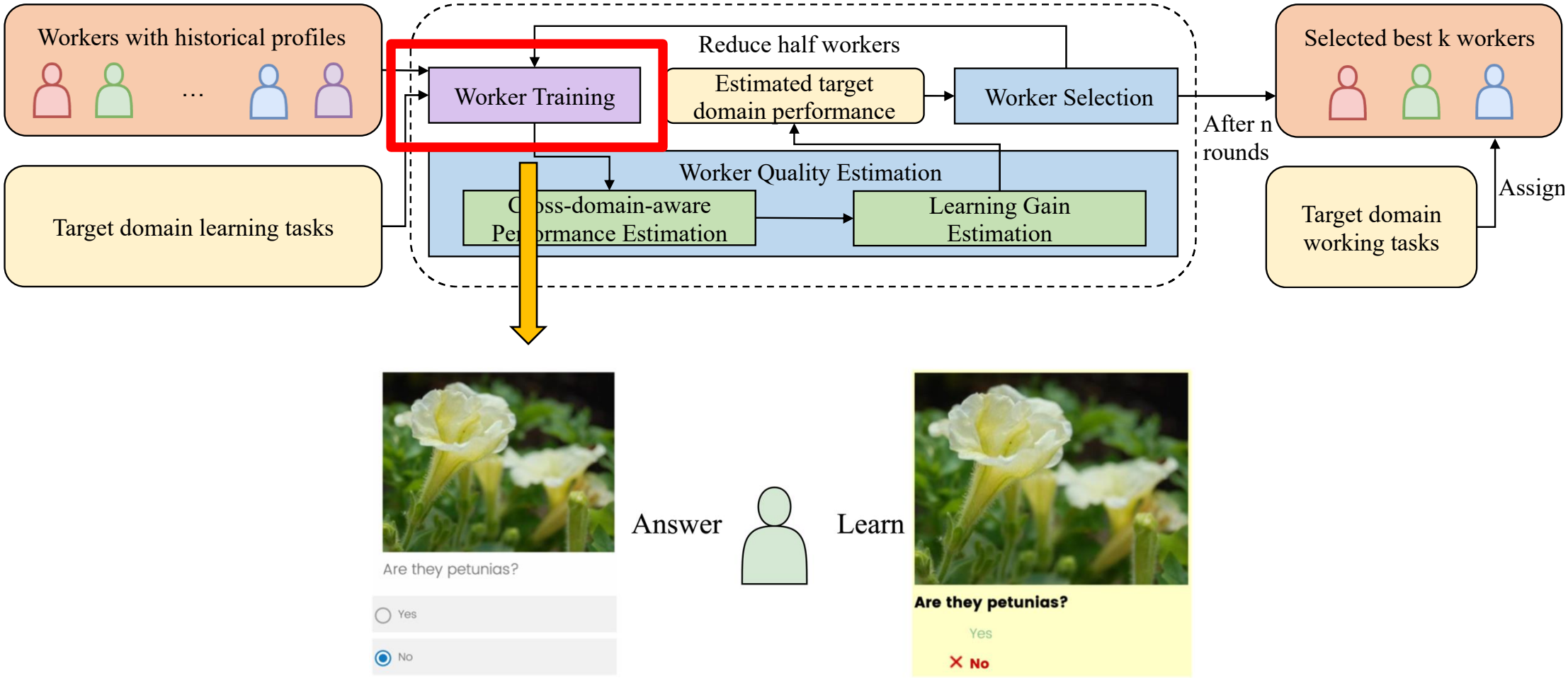
- Many companies such as JD, Alibaba, and Baidu have their commercial crowdsourcing platforms with worker pools, which record the answering history of workers.
- The answering history of workers (prior domain knowledge) can help select high-quality workers when annotating a new domain (target domain task).



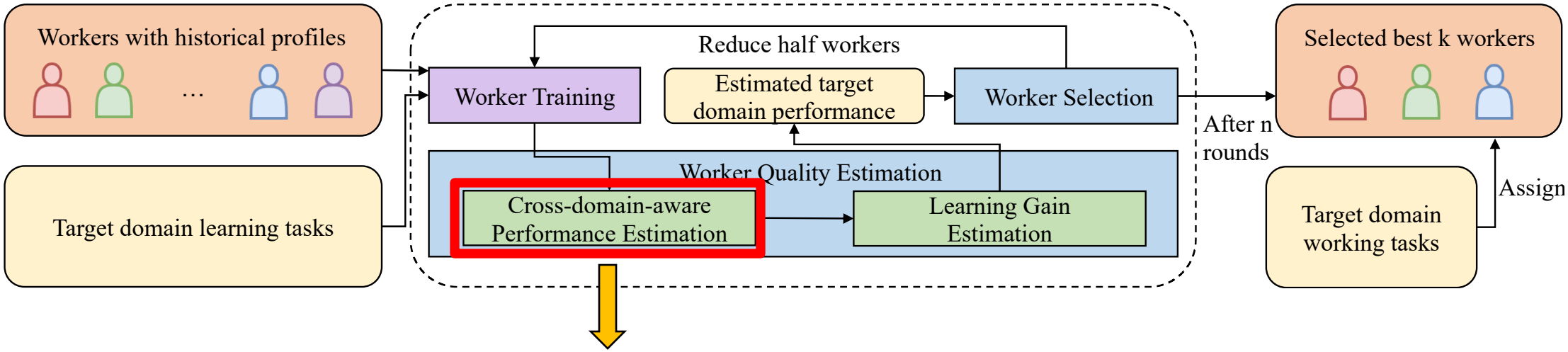
Methodology



Methodology



Methodology



- Multi-variate normal distribution to model the correlation of the crowd-worker as a group over different domains.
- Maximum Likelihood Estimation to estimate the parameters in the distribution based on the worker training results.

Methodology

- Maximum likelihood estimation:

$$\log L = \sum_{i=1}^{|W_c|} \log P(h_{i,T}|h_i)$$
$$= \sum_{i=1}^{|W_c|} \log \int_0^1 h_{i,T}^{C_{i,c}} (1-h_{i,T})^{X_{i,c}} \frac{e^{-\Psi}}{\sqrt{2\pi|\bar{\Sigma}|}} \mathrm{d}h_{i,T}$$
$$= \sum_{i=1}^{|W_c|} \Big[\log \int_0^1 h_{i,T}^{C_{i,c}} (1-h_{i,T})^{X_{i,c}} e^{-\Psi} \mathrm{d}h_{i,T}$$
$$+ \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \log |\bar{\Sigma}| \Big],$$

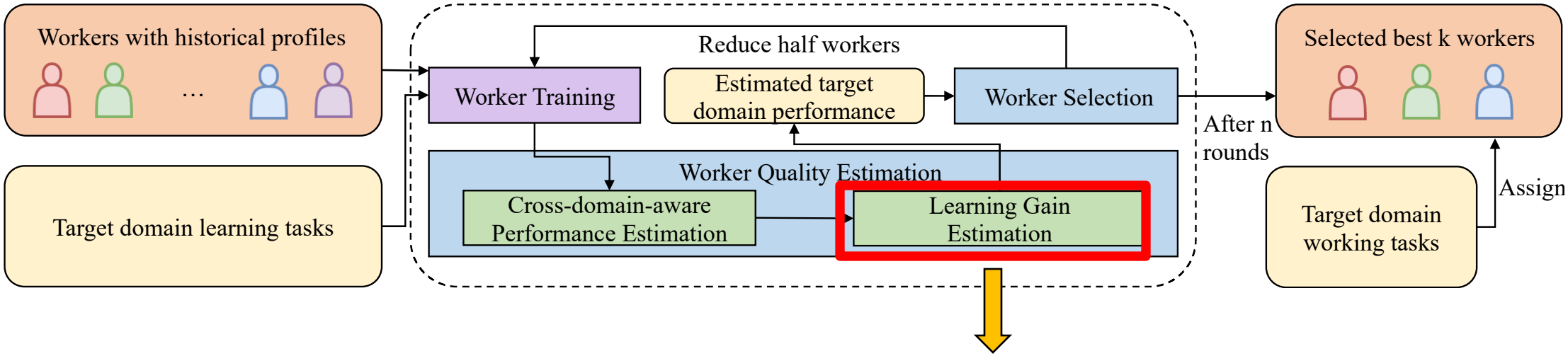
$$\bar{\mu} = \mu_T + \Sigma_{1\times D} \Sigma_{D\times D}^{-1} (h_i - \mu_{1\sim D}),$$
$$\bar{\Sigma} = \Sigma_{1\times 1} - \Sigma_{1\times D} \Sigma_{D\times D}^{-1} \Sigma_{D\times 1},$$

and $\Psi = \frac{(h_{i,T}-\bar{\mu})^{\mathsf{T}}(h_{i,T}-\bar{\mu})}{2\bar{\Sigma}}$.

- Updated annotation accuracy:

$$p_{c,i} = E[h_{i,T}|h_i]$$
$$= \int_0^1 h_{i,T} P(h_{i,T}|h_i) \mathrm{d}h_{i,T}$$
$$= \int_0^1 h_{i,T} \frac{P(h_i, h_{i,T})}{P(h_i)} \mathrm{d}h_{i,T},$$

Methodology



- Item Response Theory (IRT) to model the dynamic worker knowledge change during the training process for each individual worker.

Methodology

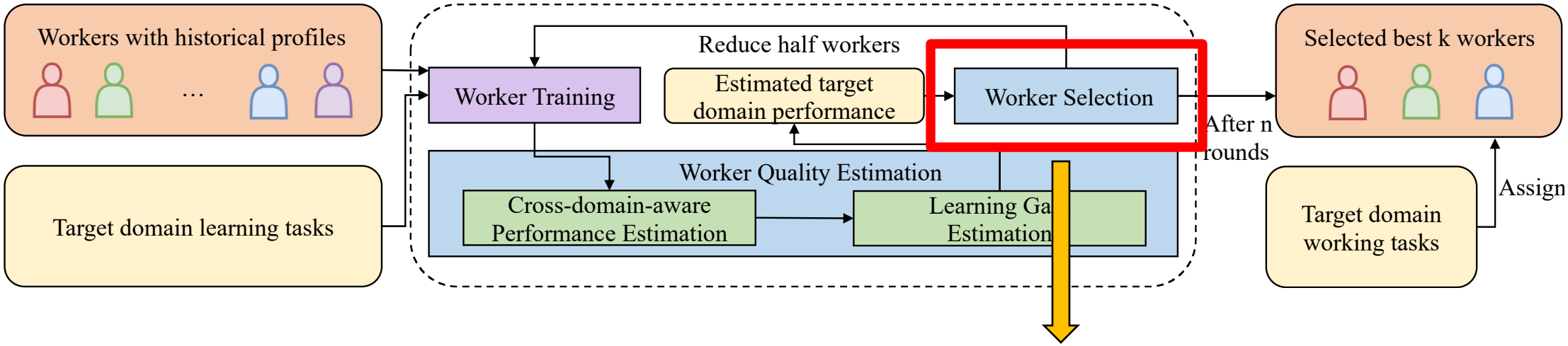
- IRT score:

$$\begin{aligned}\hat{p}_{j,i,d} &= g(\alpha_i, \beta_d, K_j) \\ &= \frac{1}{1 + e^{-(\alpha_i \ln(K_j+1) - \beta_d)}}.\end{aligned}$$

- Update the learning parameter α_i:

$$\alpha_i = \arg\min_{\alpha_i} \left[\sum_{d=1}^{D} (\hat{p}_{1,i,d} - h_{i,d})^2 + \sum_{j=1}^{c} (\hat{p}_{j-1,i,t} - p_{j,i})^2 \right]$$

Methodology



- Medium Elimination, preserve the better half of the workers in the current round and enter the next round.
- Error bound: $O(\sqrt{\frac{nk}{B}\ln\frac{1}{\delta_c}})$.

Datasets

- Datasets:

TABLE II
DATASET STATISTICS

Datasets	|W|	Q	k	total # of batches	B
RW-1	27	10	7	3	540
RW-2	35	10	9	3	700
S-1	40	20	5	7	2400
S-2	50	20	5	7	3000
S-3	80	20	5	15	6400
S-4	160	20	5	31	16000

|W|: number of crowdsourced workers

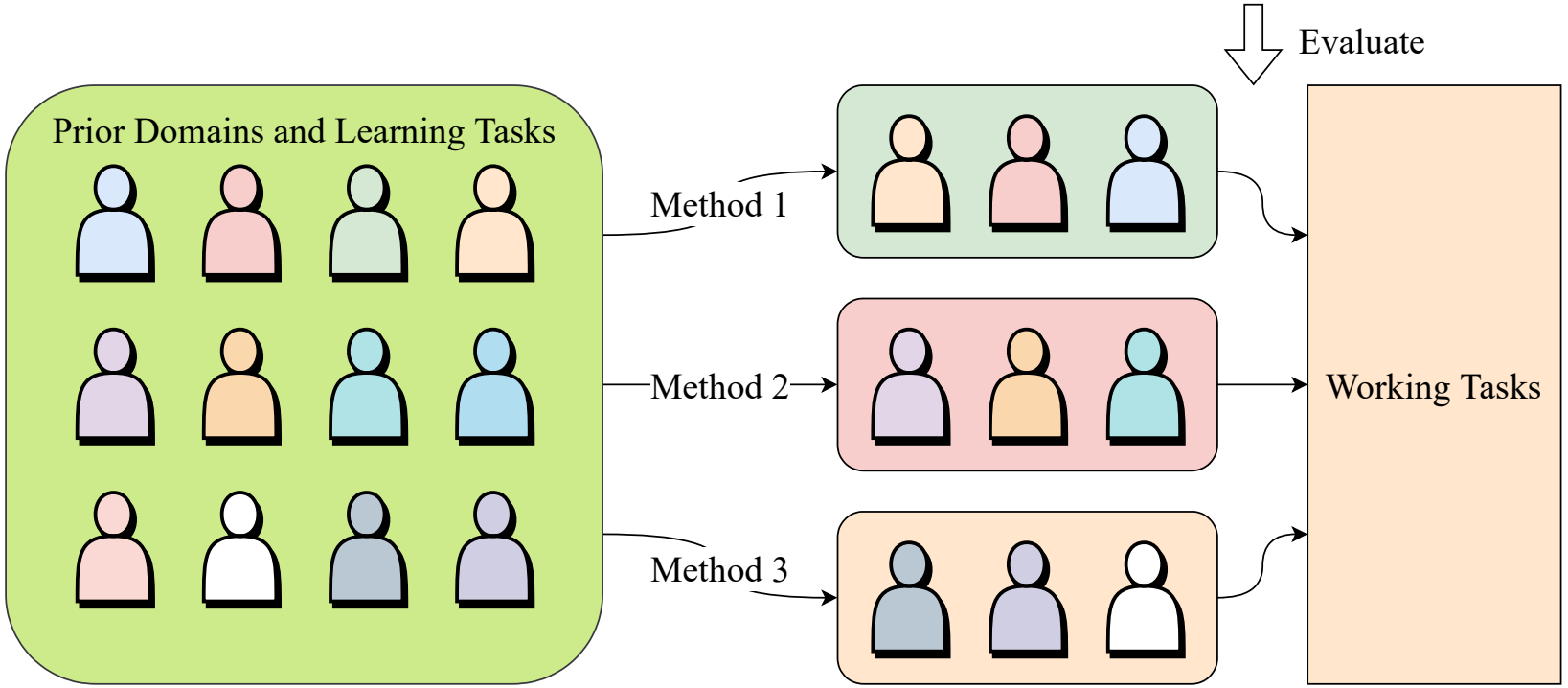
Q: number of learning tasks per batch

k: number of top-k desired workers

B: total worker selection budget

Metrics

- Metric: averaged annotation accuracy of the selected top-k workers on the target domain working task.



Baselines

- Baselines: We considered three baselines, Universal Sampling (US), Medium Elimination (ME), and Li et al.
  - US: use the budget for all the workers equally and select the top k workers
  - ME: allocates the budget in rounds and eliminates the workers by half in each round based on the accuracy of the learning tasks
  - Li et al.: compute the correlation between the prior domain historical results with the target domain performance

Experiments

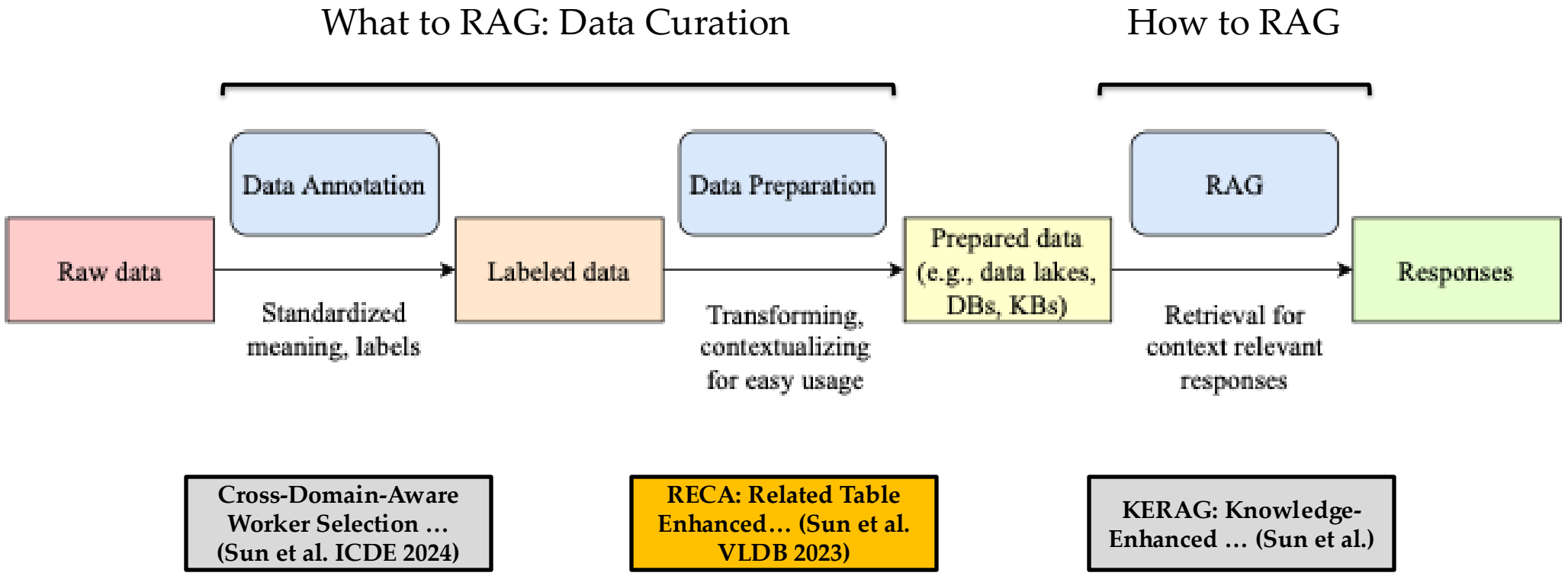
TABLE V
EXPERIMENT RESULTS

	RW-1	RW-2	S-1	S-2	S-3	S-4
US [11], [19]	0.764 (4.5% ↑)	0.956 (0.5% ↑)	0.765 (8.5% ↑)	0.775 (6.8% ↑)	0.815 (4.3% ↑)	0.865 (2.4% ↑)
ME [11], [19]	0.771 (3.5% ↑)	0.944 (1.8% ↑)	0.720 (15.3% ↑)	0.785 (5.5% ↑)	0.795 (6.9% ↑)	0.880 (0.7% ↑)
Li et al. [31]	0.771 (3.5% ↑)	0.936 (2.7% ↑)	0.780 (6.4% ↑)	0.805 (2.9% ↑)	0.845 (0.6% ↑)	0.870 (1.8% ↑)
Ours	**0.798**	**0.961**	**0.830**	**0.828**	**0.850**	**0.886**
Ground Truth	0.914	1.000	0.885	0.875	0.915	0.975

Summary

- We incorporate the **cross-domain knowledge** information and the **dynamic knowledge change** and propose a novel **Median Elimination-based** worker selection with training algorithm to find high-quality workers for data annotation.
- By proposing this method, we improve the effectiveness of data annotation process in the raw data to response workflow.

Workflow – from raw data to responses



Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Preparation: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- RAG: KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering
- Future Vision and Opportunities

Data Preparation

- Data Preparation: cleaning, transforming and organizing data into a format that is suitable for analysis.

Actor Name

Name	Date
Tom Cruise	03/07/1962
Tom Hanks	09/07/1956
Morgan Freeman	01/06/1937

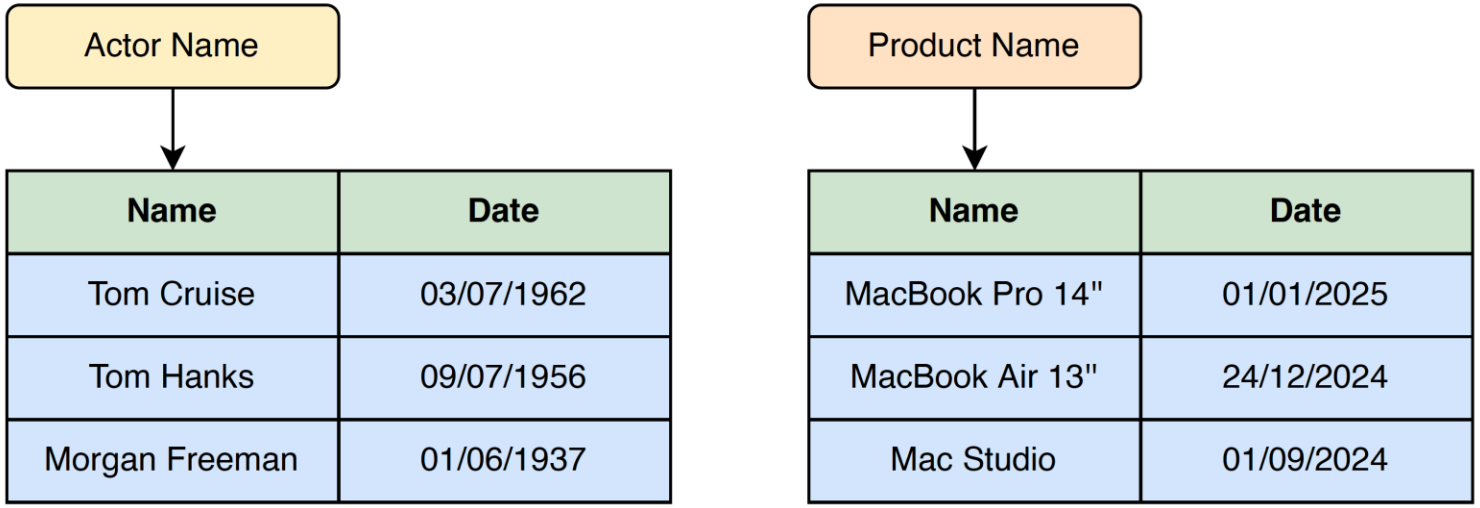
Product Name

Name	Date
MacBook Pro 14"	01/01/2025
MacBook Air 13"	24/12/2024
Mac Studio	01/09/2024

- Column semantic type annotation is one of the core tasks in data preparation.

Overview

- **RECA: Related Tables Enhanced Column Semantic Type Annotation Framework (VLDB 2023)**
- Focus on enhancing **table column semantic type annotation** with **inter-table** context information.



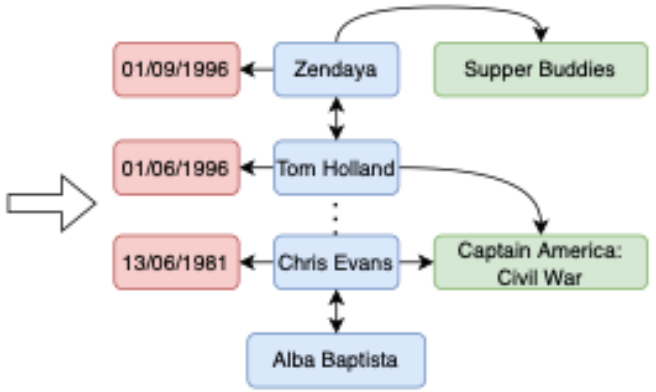
Definition

- (Column semantic type annotation): Given a table T from the data lake D, denote the target column as C_t in T. The column semantic type annotation model W annotates C_t with a semantic type $\bar{y}_t = W\ (C_t, T, D)$, such that $\bar{y}_t$ best fits the semantics of C_t.

Background

- Accurate column semantic type annotation is important for various applications:
  - KG construction, table data RAG, data cleaning, data integration, etc.

Actor	Birthday	Spouse	Works
Zendaya	01/09/1996	Tom Holland	Captain Ameirca: Civil War ...
Tom Holland	01/06/1996	Zendaya	Supper Buddies ...
Chris Evans	13/06/1981	Alba Baptista	Captain Ameirca: Civil War ...
...	...	...	...



KG construction

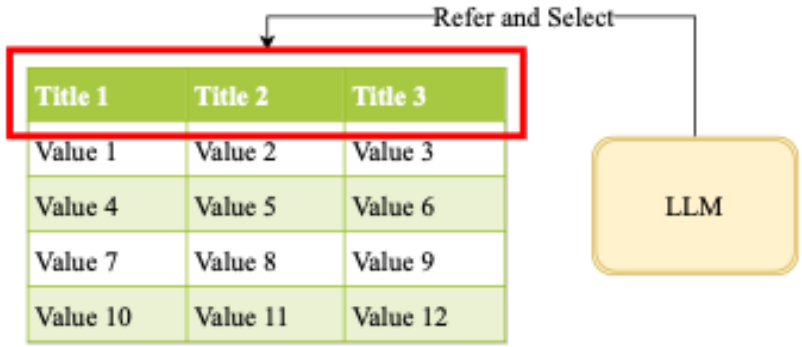


table data RAG

Challenges

- Existing works (Sherlock, Sato, DODUO, TABBIE, etc.) focus on incorporating the inner-table context.
- Our work focuses on the utilization of inter-table context, which is challenging.

?	?	?	?
Amorcito corazón	L. Suárez	D. Olivera	2012-06-10
A Nero Wolfe Mystery	S. M. Kaminsky	M. Chaykin	2002-08-18

WPPD

?	?	?	?
Chōriki Sentai Ohranger	T. Inoue	T. Satō	1996-02-23
Chōjin Sentai Jetman	T. Inoue	T. Wakamatsu	1992-02-14
Brewster Place	M. Angelou	O. Winfrey	1990-05-30
Anne of Green Gables: The Continuing Story	K. Sullivan	J. Crombie	2000-07-30
Angry Boys	C. Lilley	C. Lilley	2011-07-27
Alex Haley's Queen	A. Haley	Ann-Margret	1993-02-18
...	...	...	...

WPPD

Motivation

- Named Entity Schema: table schema generated based on the most frequent named entity type extracted from each column.
- Tables with the same/similar named entity schemata tend to be from the same/similar data source and thus tend to have the same/similar column semantic types.

?	?	?	?
Amorcito corazón	L. Suárez	D. Olivera	2012-06-10
A Nero Wolfe Mystery	S. M. Kaminsky	M. Chaykin	2002-08-18

WPPD

?	?	?	?
Chōriki Sentai Ohranger	T. Inoue	T. Satō	1996-02-23
Chōjin Sentai Jetman	T. Inoue	T. Wakamatsu	1992-02-14
Brewster Place	M. Angelou	O. Winfrey	1990-05-30
Anne of Green Gables: The Continuing Story	K. Sullivan	J. Crombie	2000-07-30
Angry Boys	C. Lilley	C. Lilley	2011-07-27
Alex Haley's Queen	A. Haley	Ann-Margret	1993-02-18
...	...	...	...

WPPD

?	?	?	?
Donkey Kong Country	Nintendo	2006-12-08	2006
F-Zero	Nintendo	2006-12-08	2006
SimCity	Nintendo	2006-12-29	2006
Super Castlevania IV	Konami	2006-12-29	2006
Street Fighter II: The World Warrior	Capcom	2007-01-19	2007
...	...	...	...

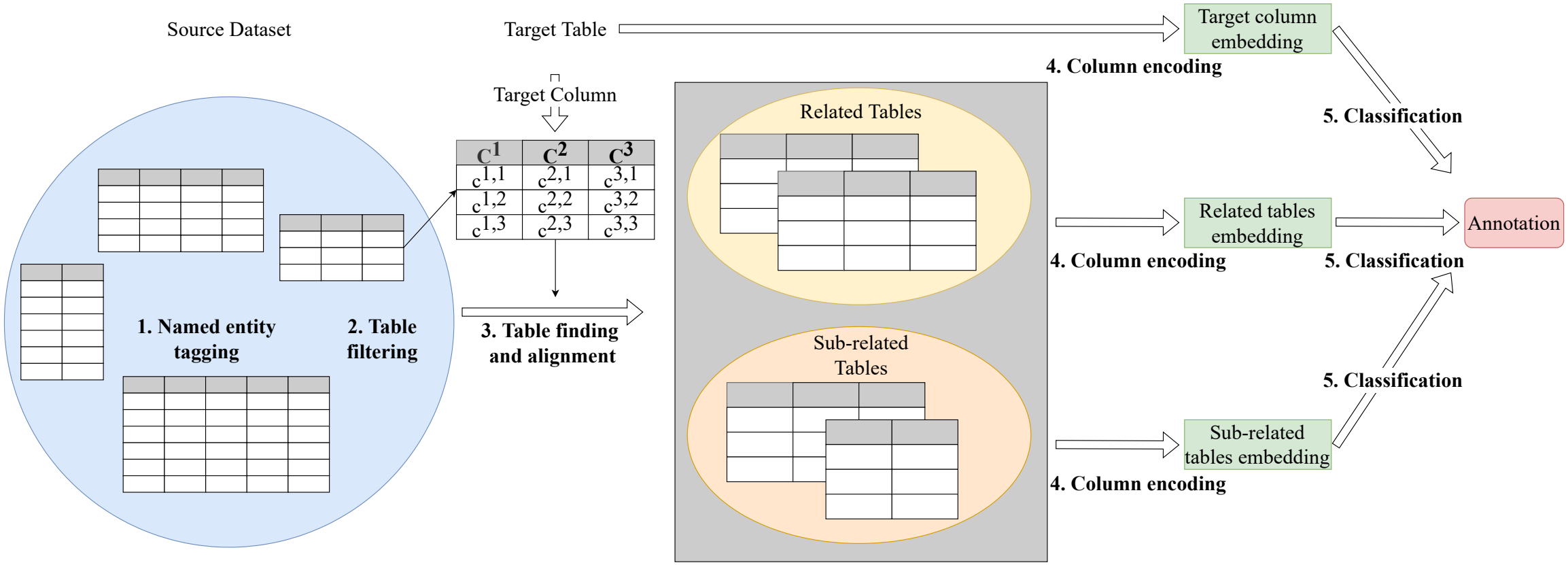
WODD

- W: Work of art; P: Person; D: Date; O: Organization

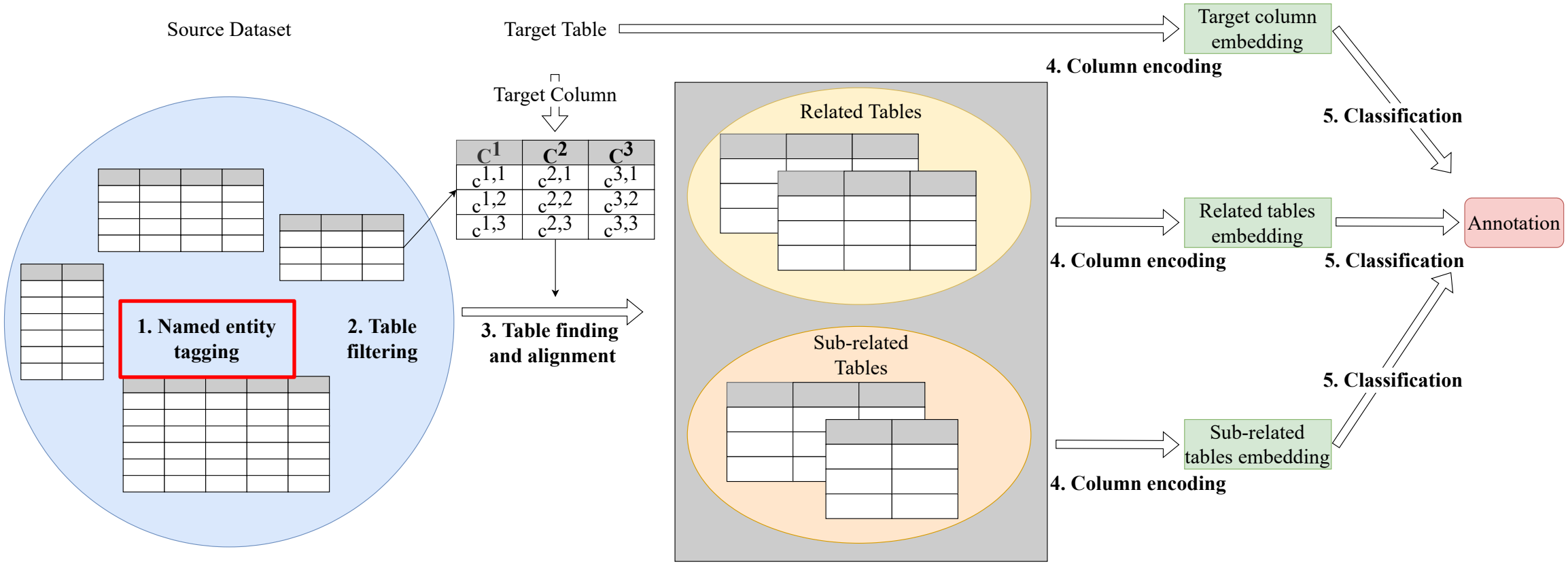
Concepts

- Related Tables: The tables that share the same named entity schema and are similar in content (Jaccard Similarity > δ) with the original table.
- Sub-related Tables: The tables that share a similar named entity schema (the edit distance between their named entity schemata is less than a threshold) and are similar in content (Jaccard Similarity > δ) with the original table.

Methodology



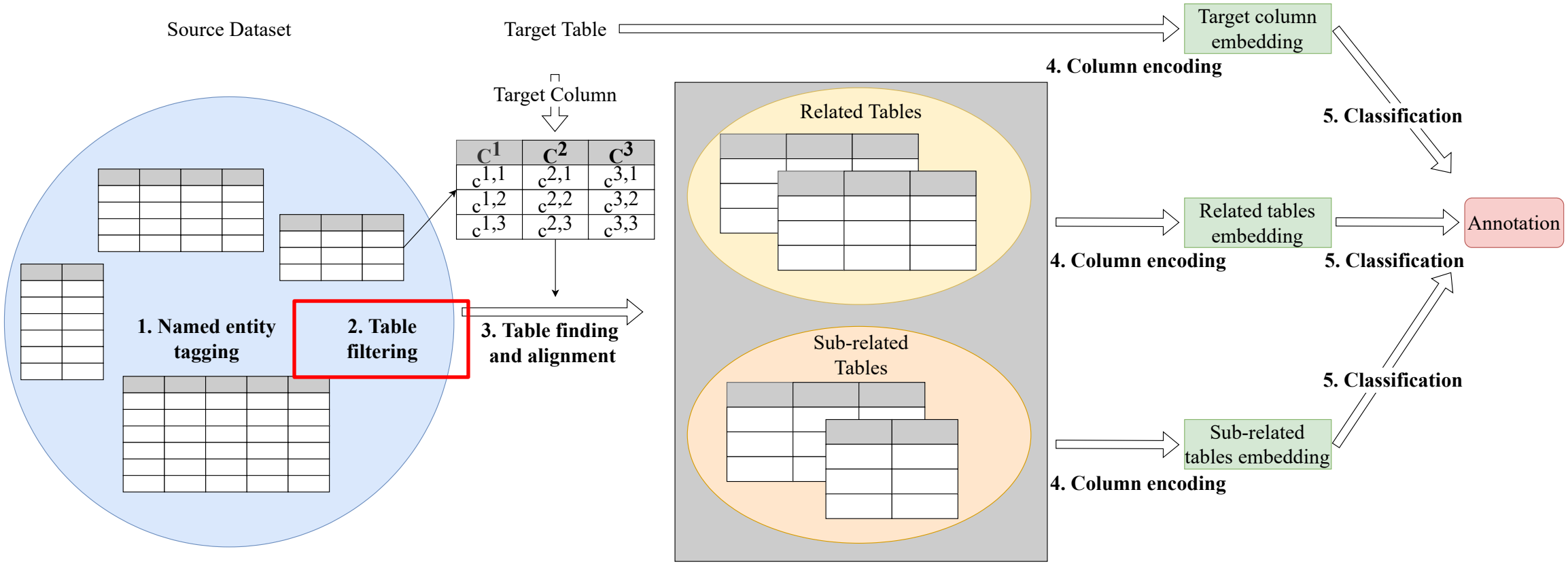
Methodology



Methodology

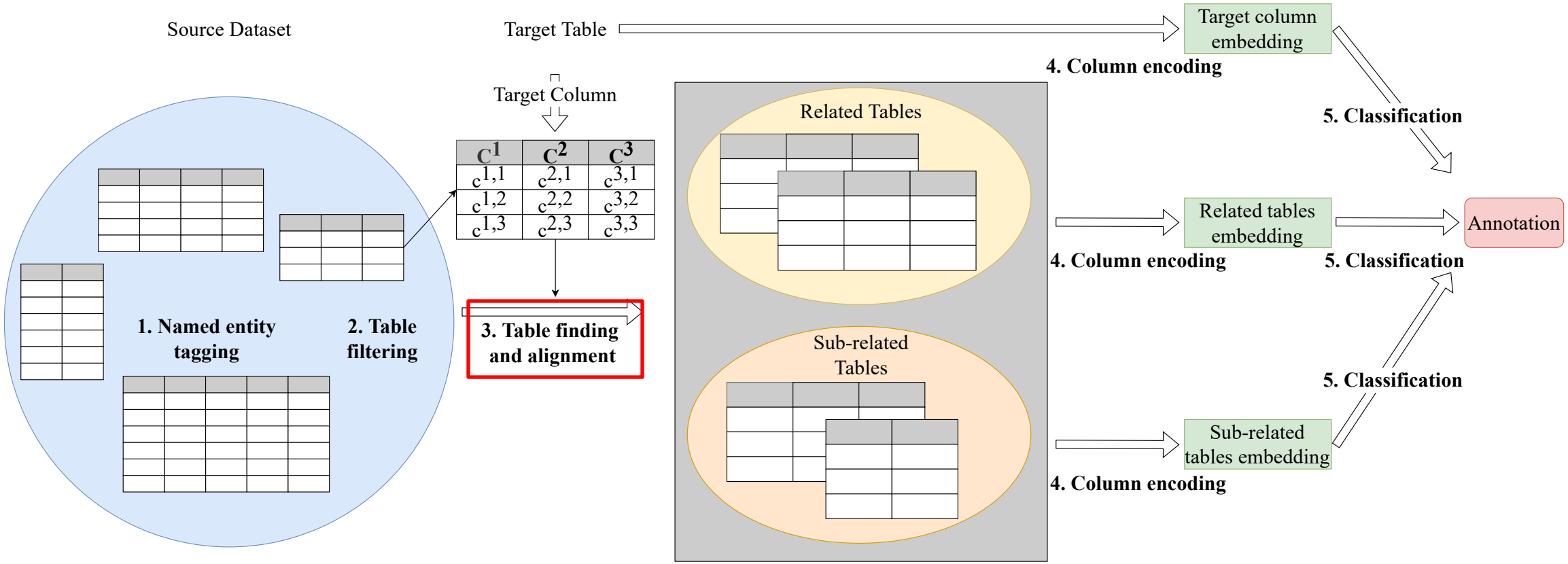
- Given a table *T*, we use the spaCy tagging tool to identify the named entities in each column and tag them.
- We further classify the DATE and PERSON types based on the data format.
  - E.g. DD-MM-YYYY; YYYY; January 16th 2022; 2023
  - E.g. J. K. Rowling; Anna
- We include an additional EMPTY type.
- The most frequent named entity type in each column forms the named entity schema.

Methodology



$$\text{Jaccard}(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

Methodology

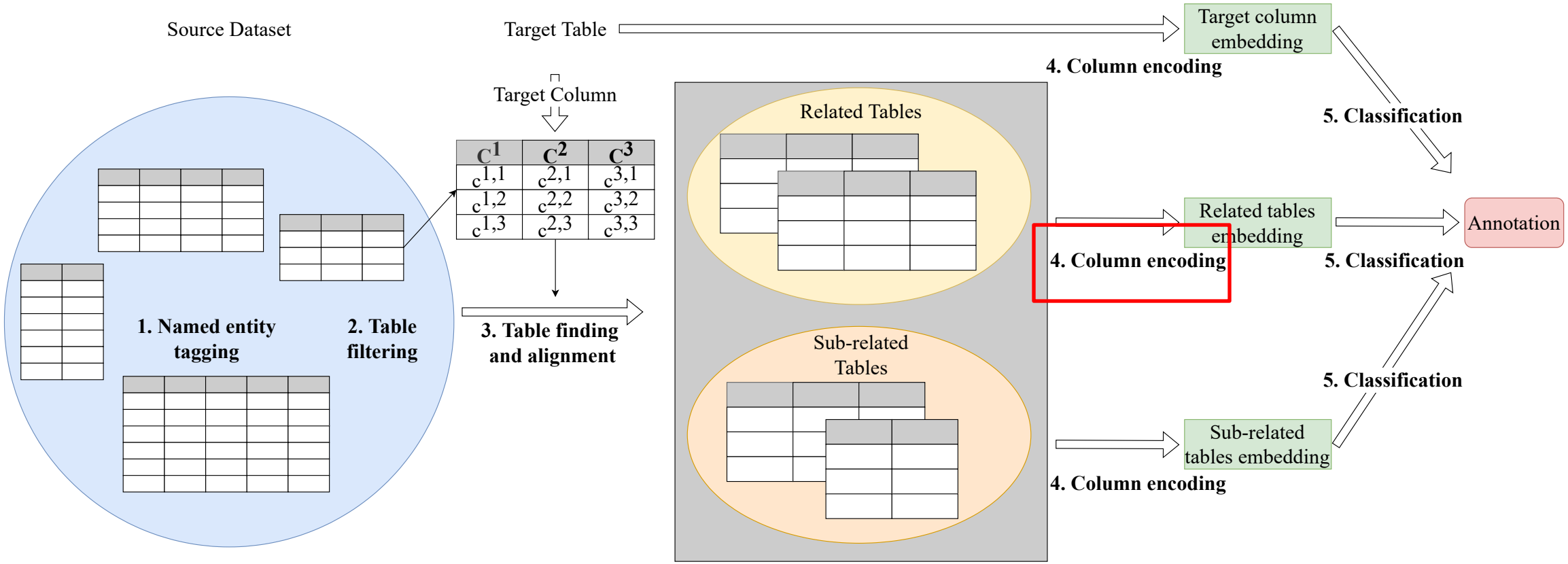


Named Entity Schema & Jaccard Similarity

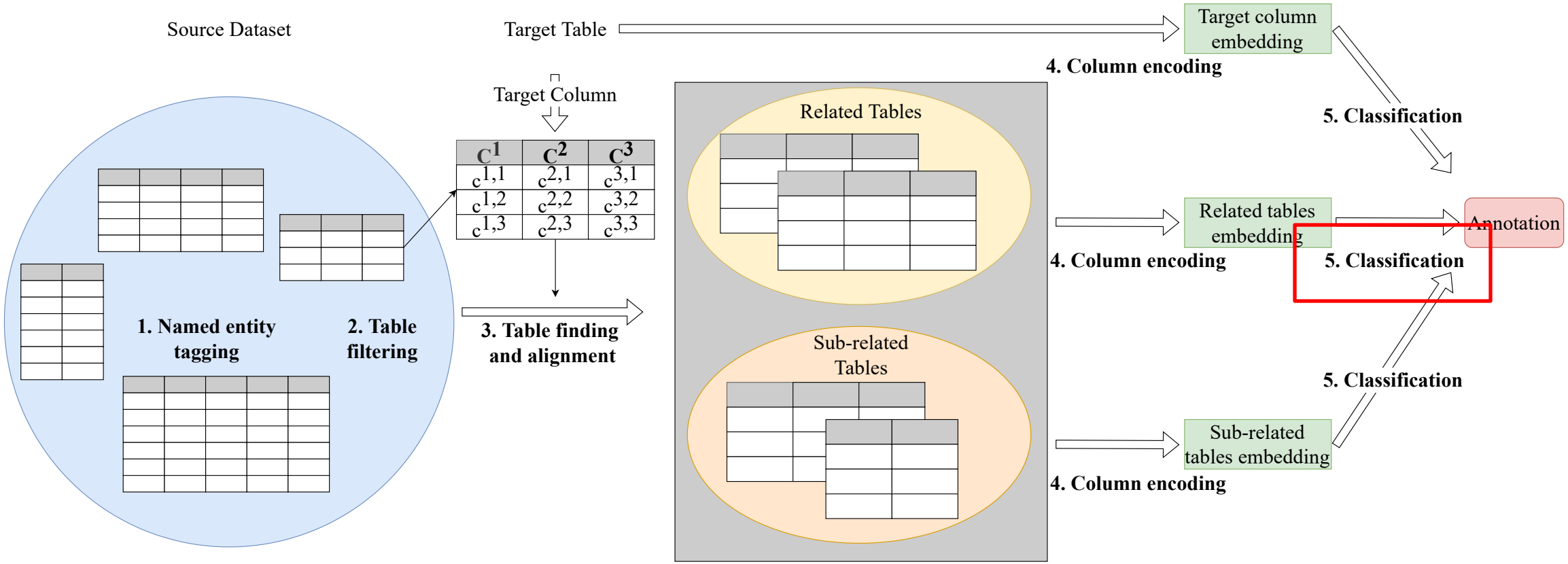
Methodology

- Related tables: candidate tables T_j that share the same named entity schema as T_i.
- Sub-related tables: we consider the following two requirements:
  - Schema similarity: the named entity schemata should not be very different (edit distance less than a threshold).
  - Column location alignment: The named entity type of the target column matches with that of the column at the identical location in the sub-related table.

Methodology



Methodology



$$a_i^t = \alpha * \hat{v}_i^t + \beta * \hat{r}_i^t + \gamma * \hat{x}_i^t$$

Experiments

- Datasets:

	WebTables	Semtab2019
# semantic types	78	275
# tables	32262	3045
# annotated columns	74141	7603
Avg. # rows	20.0	69.0
Avg. # columns	2.3	4.5
Avg. # annotated columns	2.3	2.5

- Metrics:
  - Support-weighted F1: weighted support of per type F1 scores
  - Macro average F1: average of per type F1 scores (emphasize on long-tail types)

Experiments

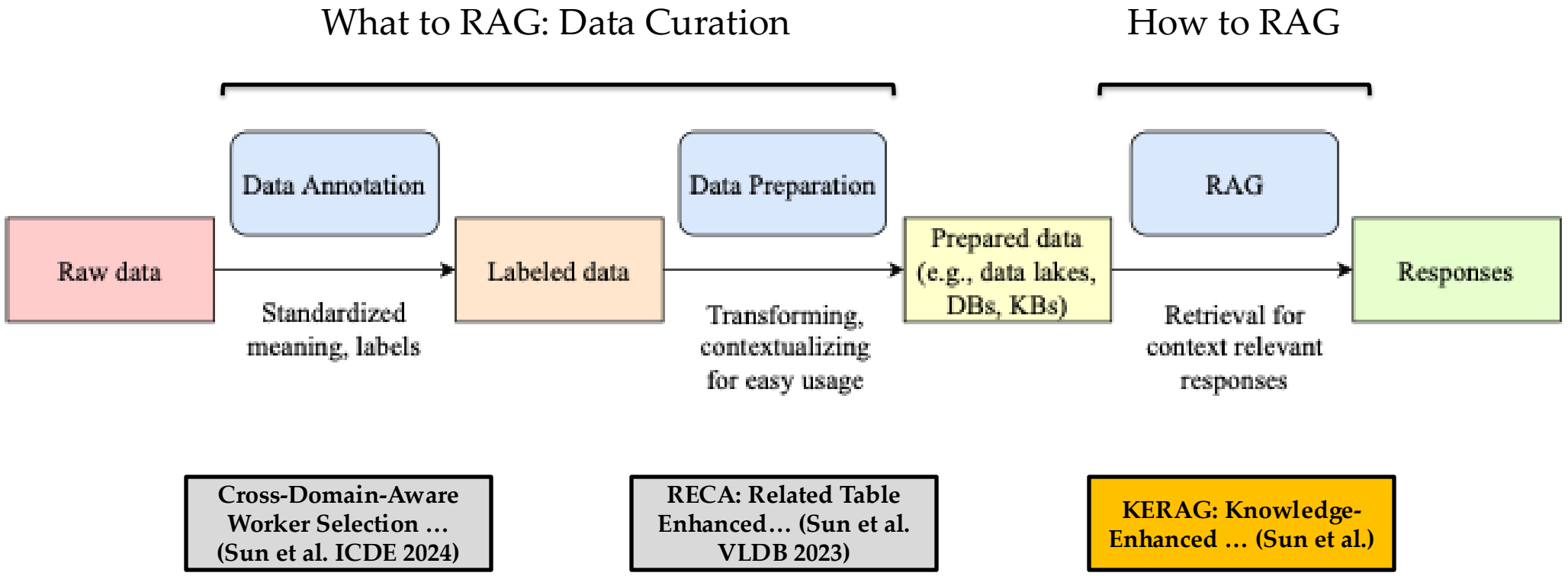
- RECA outperforms all the state-of-the-arts in terms of the F1 scores.

	Semtab2019 dataset		WebTables dataset	
Model names	Support-weighted F1	Macro average F1	Support-weighted F1	Macro average F1
Sherlock [15]	0.646 ± 0.006	0.440 ± 0.009	0.844 ± 0.001	0.670 ± 0.010
TaBERT [35]	0.768 ± 0.011	0.413 ± 0.019	0.896 ± 0.005	0.650 ± 0.011
TABBIE [16]	0.799 ± 0.013	0.607 ± 0.011	0.929 ± 0.003	0.734 ± 0.019
DODUO [30]	0.820 ± 0.009	0.630 ± 0.015	0.928 ± 0.001	0.742 ± 0.012
RECA	**0.853** ± 0.005	**0.674** ± 0.007	**0.937** ± 0.002	**0.783** ± 0.014

Summary

- We propose RECA for column semantic type annotation. RECA extracts and leverages inter-table context to enhance the annotation quality of the target column. A novel named entity schema was designed to efficiently align related and sub-related tables, which resolves the difficulty of incorporating inter-table context.
- RECA enhances the performance of one of the core tasks in data preparation, which provides contextualized labels for tabular data, improving the usability of tabular data in the downstream applications such as RAG.

Workflow – from raw data to responses

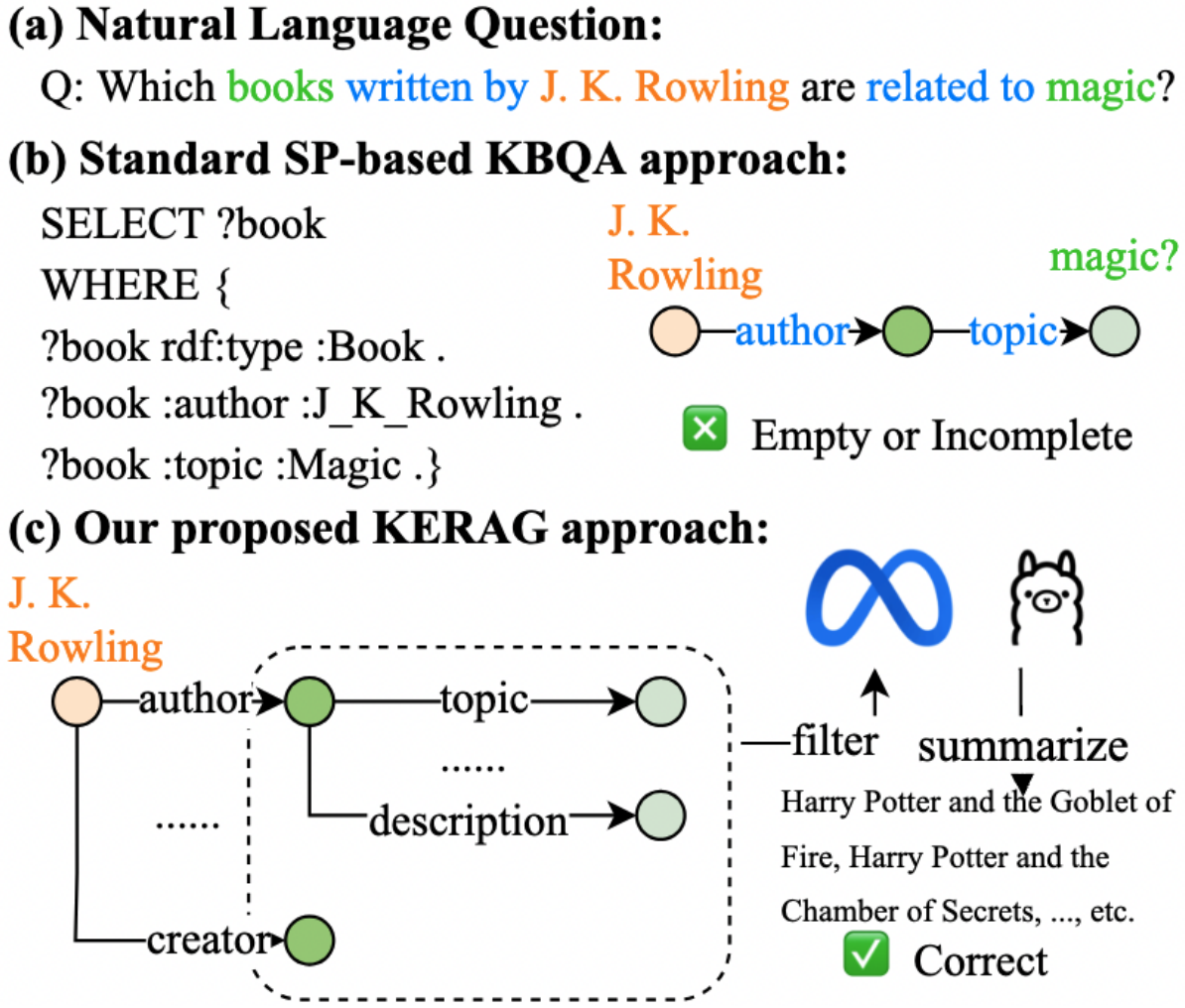


Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Preparation: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- RAG: KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering
- Future Vision and Opportunities

Overview

- **KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering (Under Submission)**
  - KB is considered as an important knowledge source for RAG due to its conciseness in knowledge, clarity in semantics, and efficiency in querying.
  - Existing KBQA approaches mainly rely on semantic parsing, which requires rigorous schemas and has low tolerance to parsing errors.
  - The powerful summarization ability of LLMs opens up opportunity for us to relax the strict retrieval process of the semantic parsing approaches. → high coverage, reduced complexity in query parsing.



Challenges

- Knowledge overloading: some head entities may contain too much relevant information (over 2M attributes [3]).
- Multi-hop retrieval boundary: how to determine the boundary of multi-hop retrieval is challenging.
- Complex query handling: how to handle the complex queries that require aggregation / reasoning is challenging.

[3] X. L. Dong, et al., "From Data Fusion to Knowledge Fusion," *Proceedings of the VLDB Endowment*, vol. 7, no. 10, pp. 881-892, Jun. 2014, doi:https://doi.org/10.14778/2732951.2732962.

KB-based RAG

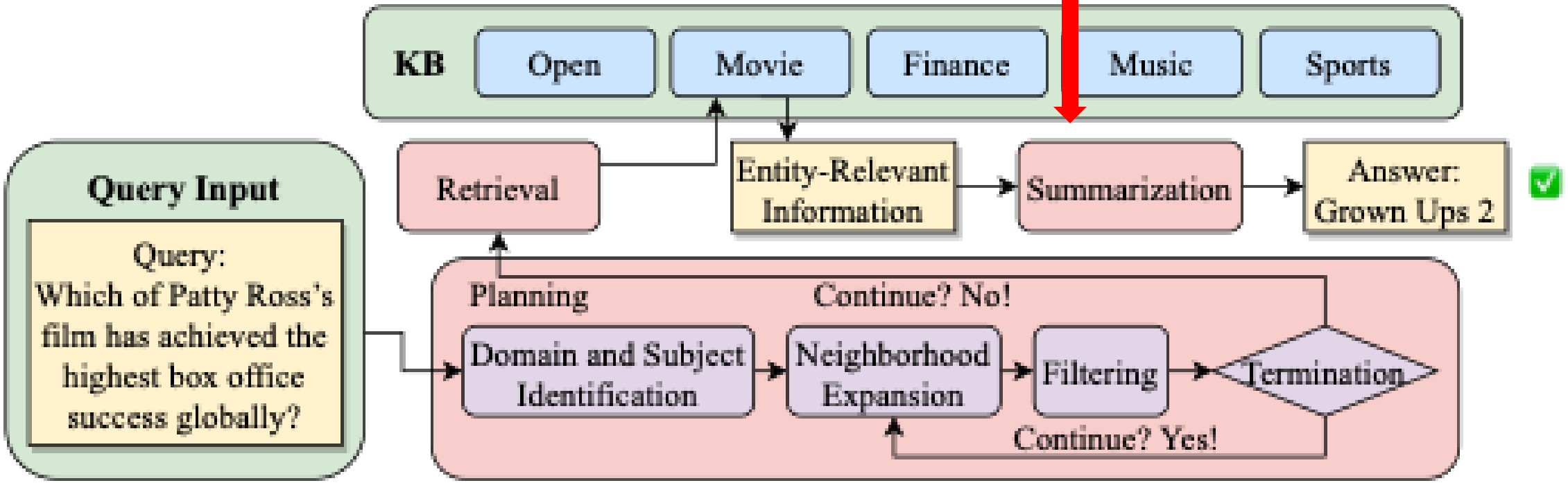
- A **Knowledge Base (KB)** K can be defined as $K = (E, R, D)$, where E is a set of entities, R is a set of relations and optionally D is a set of domains covered by K. Each entity $e_i \in E$ can have properties. Each relation $r_j \in R$ can be represented as a binary relation $R: E \times E \longrightarrow \mathbb{B}$. A KB is normally accompanied with an ontology, describing the entity types and relationships between different types of entities.
- **KB-based RAG**: Given a natural language question Q with a KB K, KB-based RAG aims to answer the question Q with knowledge in K.

Methodology

Entity level parsing + predicate level planning → flexibility across different KBs

Complex query handling

CoT-based summarizer finetuning



LLM as agent, operate on KB ontology instead of the actual KB content

Knowledge overloading and multi-hop retrieval boundary

Methodology – Planning (Domain and Subject Identification)

- LLM is leveraged to detect the question domain (if multiple KBs with different domains are involved) and the subject entity.

Planning Prompts Skeleton

You are given a Query [Q], please extract the main entity [E] from the Query.
Determine the domain the query is about. The domain should be one of the following: [Set of KG domains].
[Examples] (Optional)
[Further Instructions] (Optional)

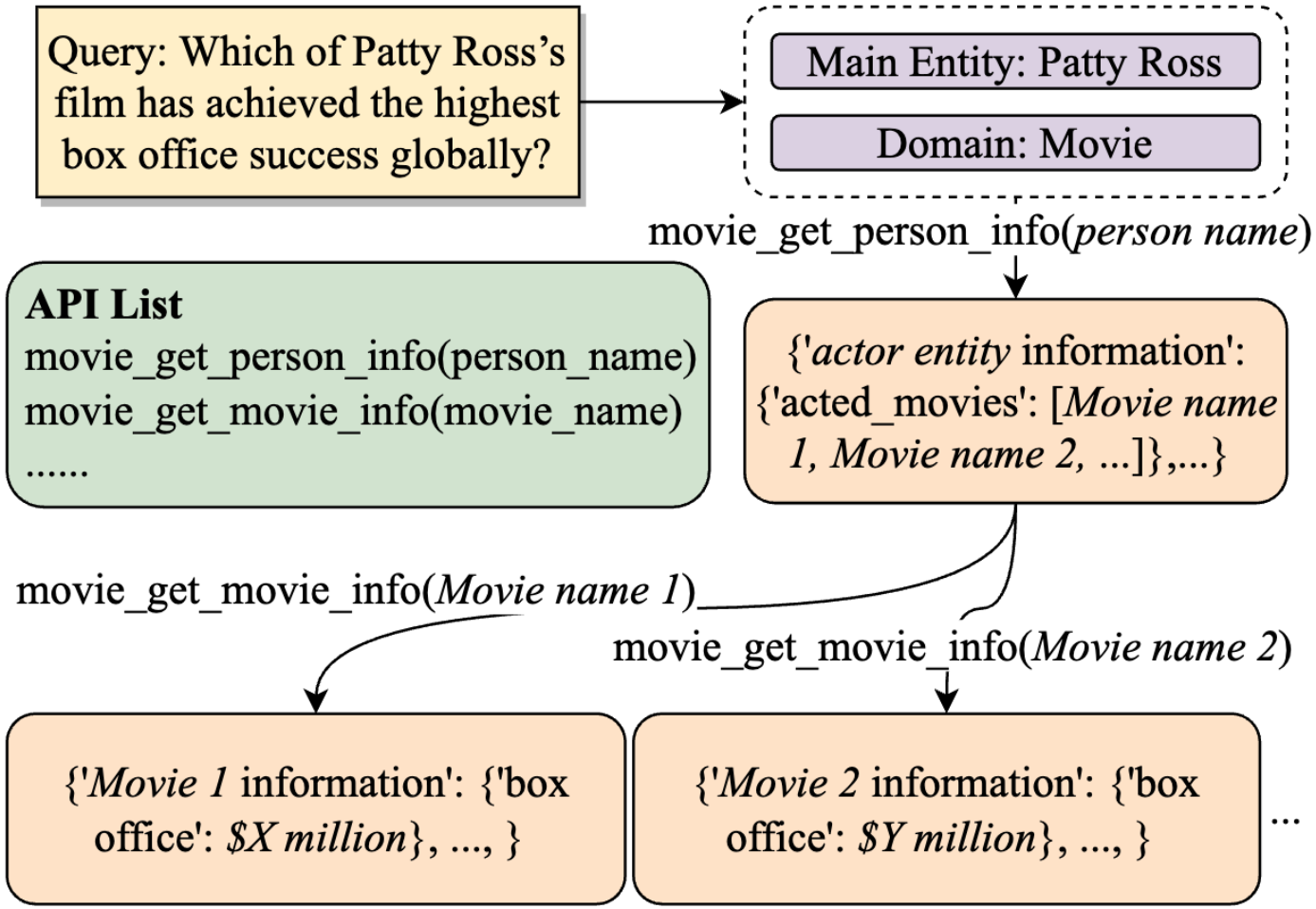
Extract subject entity

Detect domain for KG alignment

Extract relevant information based on application scenarios (e.g., time, period, etc.)

Methodology – Planning (Neighborhood Expansion)

- Explore the schema of the KB.
- Understand the relations and entity types in the next step.
- Prepare for the next hop.



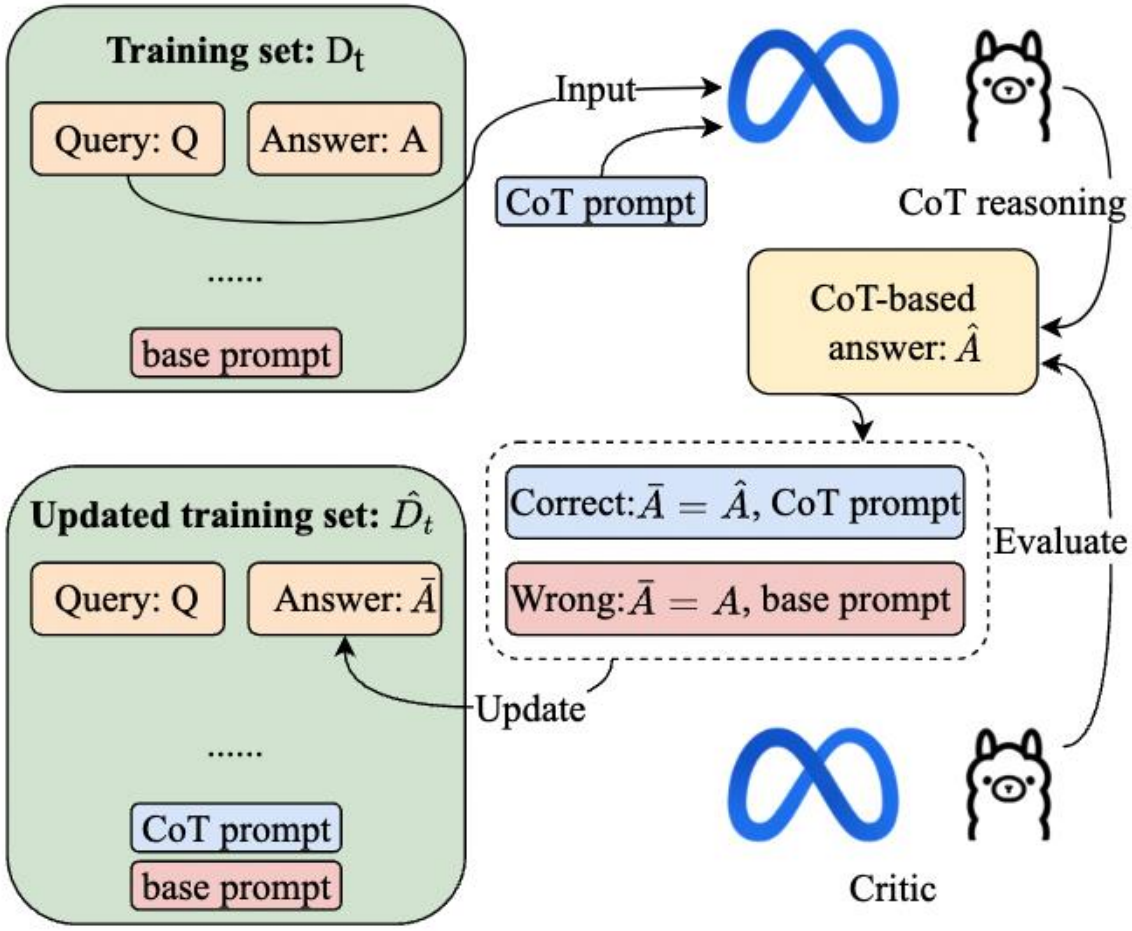
Methodology – Planning (Filtering and Termination)

- This step decides:
  - Which neighborhood content is absolutely irrelevant for answering the question
  - Whether we have sufficient information for answering the question
- Avoid knowledge overloading and decide the multi-hop boundary

Filtering Prompts Skeleton

Given the user query [Q] with detected main entity [E] and the current [retrieval plan], you will also have access to the following functions [a set of functions/predicates with ontology] to perform additional hops based on the current plan.
Decide which functions/predicates are relevant to answer the query based on the ontology and if the composed retrieval plan sufficient for the question answering.
[Examples] (Optional)
[Further Instructions] (Optional)

Methodology - Summarization



Summarization Prompts Skeleton

Please provide a brief answer to the question [Q] based on your knowledge and the following content. Answer "I don't know" if you are not confident of your answer. Please think step by step.
[Retrieved Content (Triple form)]

Experiments on CRAG, Head2Tail, and QALD-10-en

CRAG (API-based KB)

Model	Accu.	Hall.	Miss.	Truth.
GPT-4o	0.341	0.090	0.569	0.251
Llama	0.306	0.080	0.614	0.227
GPT-4o (tool)	0.362	**0.047**	0.592	0.315
Llama (tool)	0.220	0.057	0.723	0.163
apex (Ouyang et al., 2024)	0.652	0.194	0.154	0.458
db3 (Xia et al., 2024)	0.510	0.173	0.317	0.337
KERAG (8B)	0.713	0.208	0.080	0.505
KERAG	**0.732**	0.202	**0.066**	**0.529**

Head2Tail (SPARQL-based KB)

Model	Accu.	Hall.	Miss.	Truth.
GPT-4o	0.502	0.160	0.338	0.342
Llama	0.452	0.163	0.385	0.290
GPT-4o (tool)	0.799	0.035	0.166	0.764
Llama (tool)	0.752	**0.031**	0.217	0.721
WikiSP (Xu et al., 2023)	0.858	0.066	0.076	0.782
StructGPT (Jiang et al., 2023)	0.895	0.105	**0.000**	0.790
KERAG	**0.908**	0.049	0.043	**0.860**

We adopted an auto-evaluation scheme based on Llama-3.1-70B-Instruct for CRAG and Head2Tail datasets. For the QALD-10-en dataset, we align the evaluation with the exact match scheme used by ToG for fair comparison.

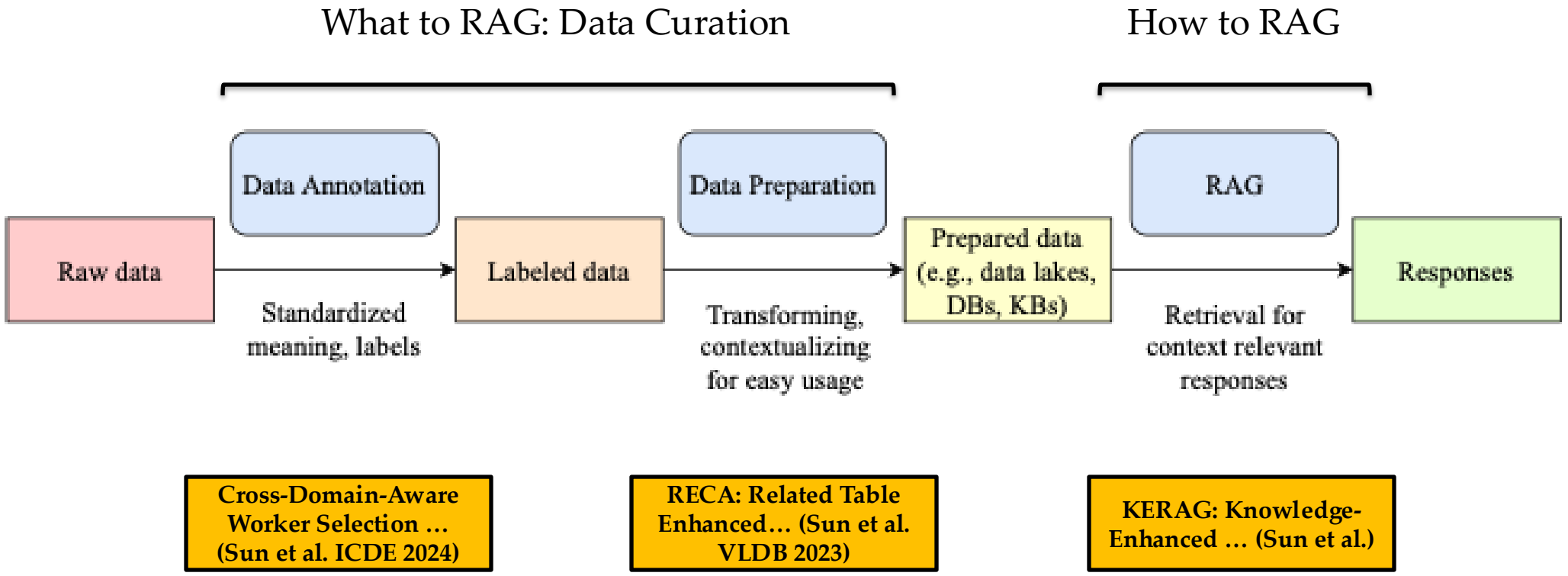
QALD-10-en (SPARQL-based KB)

Model	Accuracy
ToG (Sun et al., 2024a)	0.502
ToG-2 (Ma et al., 2025)	0.541
KERAG	**0.558**

Summary

- In this work, we proposed KERAG, an LLM-based KB RAG solution, which resolves the limitations of existing semantic parsing KBQA approaches.
- Our KERAG approach deploys LLMs as agent to understand and explore the KB; introduces a CoT-based summarizer finetuning scheme, and generalizes well across different types of KBs.
- This work represents our exploration on how to transform curated data (KBs) into valuable responses for QA.

Workflow – from raw data to responses



Outline

- Background
- Data Annotation: Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation
- Data Preparation: RECA: Related Tables Enhanced Column Semantic Type Annotation Framework
- RAG: KERAG: Knowledge-Enhanced Retrieval-Augmented Generation for Advanced Question Answering
- Future Vision and Opportunities

Research Opportunities: Multi-source RAG

- Existing KB-based RAG solutions mainly focus on a single KB as the knowledge source.
- In reality, answering a single question may require the RAG system to jointly consider multiple knowledge sources to generate correct answers (less touched by the community).
- These sources can have different modalities and different trustworthiness.
- Consistency and/or conflicts could exist in different knowledge sources.
- Currently, I am working on an advanced LLM-based router for multi-source RAG.