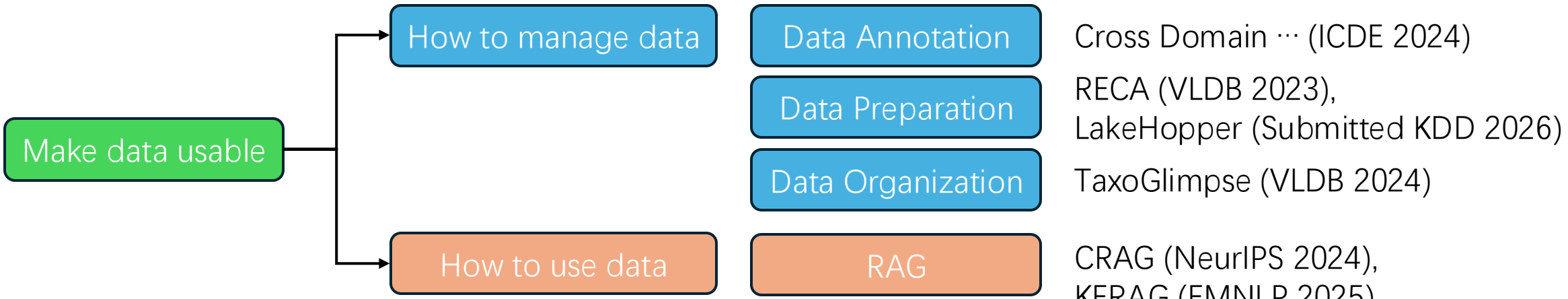
About me

- Yushi Sun (Steve), PhD from HKUST
- Supervised by Prof. Lei Chen
- Currently researcher at Tencent Games
- Interest – make data usable:
  - How to manage data: Data Curation
  - How to use data: Retrieval-Augmented Generation

My Interest

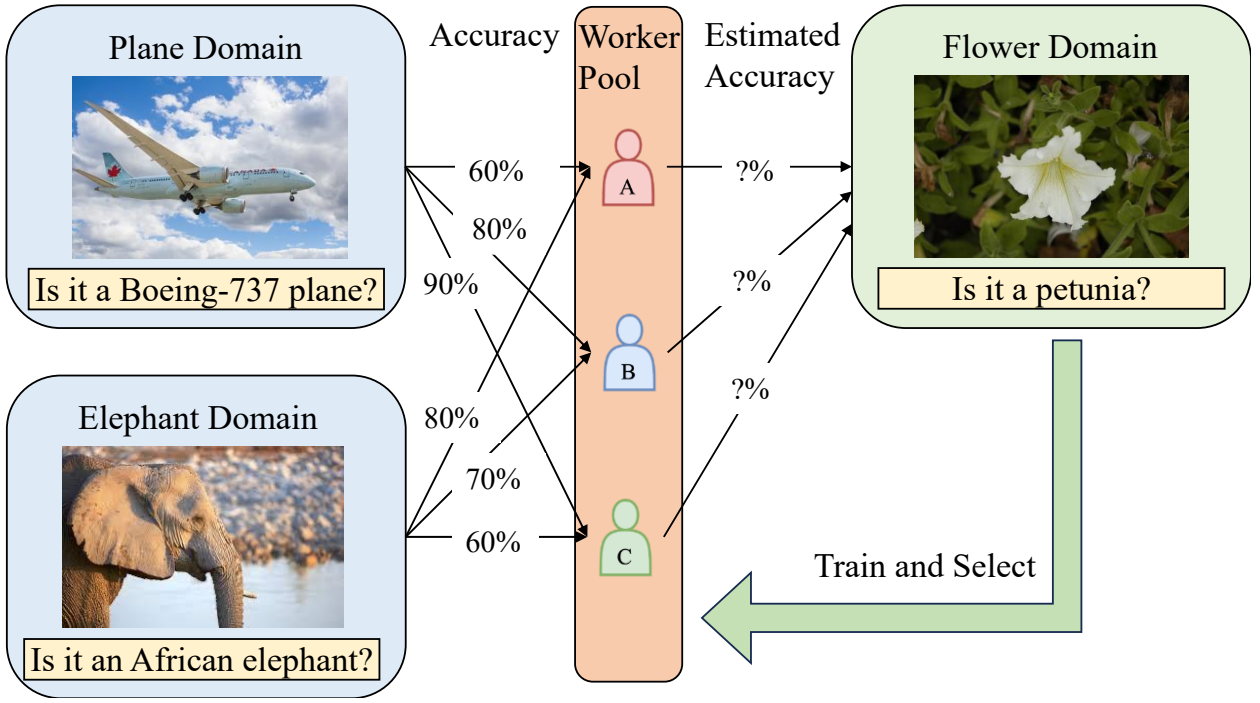

Make data usable
How to manage data
How to use data
Data Annotation
Data Preparation
Data Organization
RAG
Cross Domain ··· (ICDE 2024)
RECA (VLDB 2023),
LakeHopper (Submitted KDD 2026)
TaxoGlimpse (VLDB 2024)
CRAG (NeurIPS 2024),
KERAG (EMNLP 2025),
CacheRAG (Revision VLDB 2026)

How to manage data

Data Annotation

Cross-domain-aware Worker Selection with Training for Crowdsourced Annotation (ICDE 2024)

Task: Cross domain worker selection for data annotation.
Challenges / Limitations:
1) cross domain knowledge estimation.
2) dynamic worker knowledge change.
Solution:
A medium elimination-based worker selection methods to select crowd workers on the target domain based on the history and the golden target questions.
1) multi-variate normal distribution for modeling.
2) Item response theory to model worker learning progress.

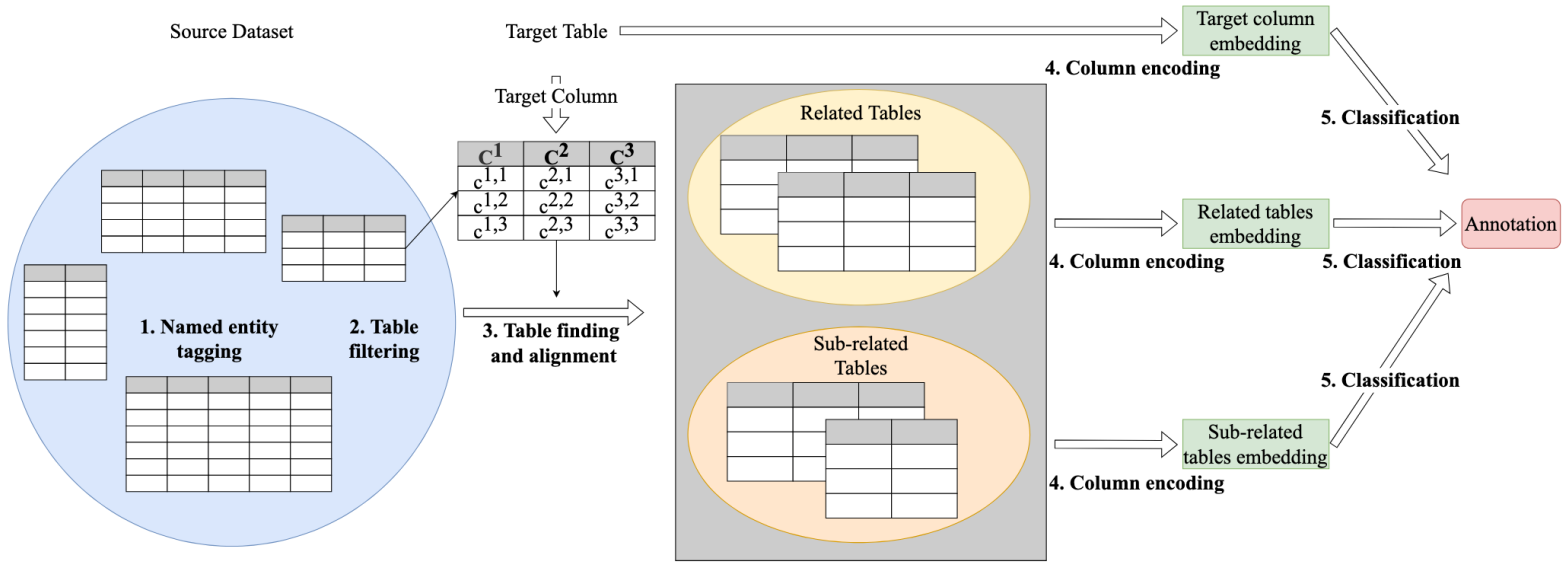


How to manage data

Data Preparation

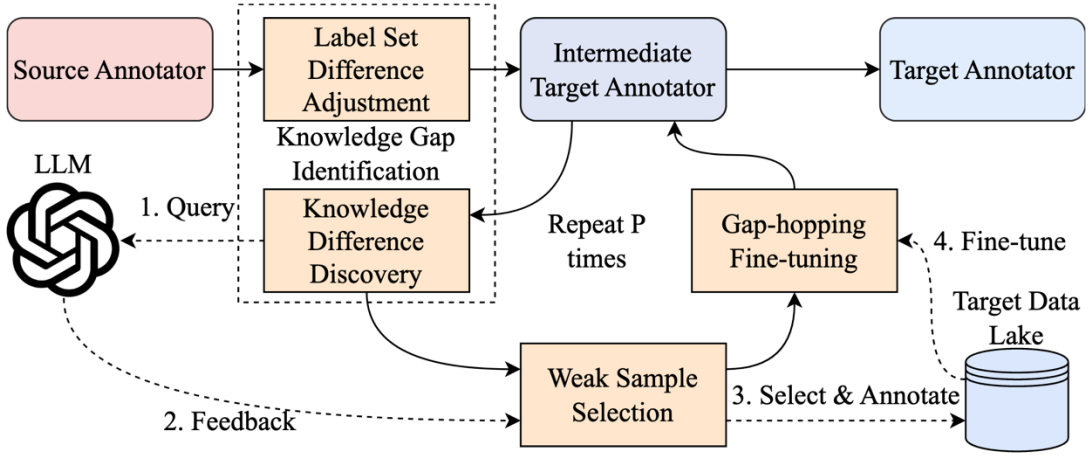
RECA (VLDB 2023)
Task: Column Semantic Type Annotation
Challenges / Limitations:
1) Inter-table context information is neglected.

Solution:
A BERT-based inter-table-context-aware solution for table annotation.
1) A named entity table schema for inter-table context alignment.



LakeHopper (Submitted KDD 2026)
Task: Cross Data Lake Column Semantic Type Annotation
Challenges / Limitations:
1) Cross data lake table annotation performance is poor.
2) Retrain annotator is too expensive.
3) Reuse of existing trained table annotation models is neglected.

Solution:
A PLM-LLM collaborated framework that identifies the difficult samples from the target data lake.
1) The world knowledge of LLM helps the PLM-based table annotators identify the useful weak samples from the target data lake.
2) An incremental gap-hopping finetuning mechanism that helps the transfer.



How to manage data

Data Organization

TaxoGlimpse (VLDB 2024)
Task: Benchmarking the performance of LLM on ontology learning.
Challenges \ Limitations:
New paradigm of data organization in the era of LLMs?
The performance of LLMs in internalizing the ontology information of different data entities is unknown.
Solution and Insights:
We create a benchmark to systematically evaluate the performance of LLMs on multiple domains and multiple levels in the taxonomies.

1) LLMs are good at common domain taxonomy knowledge, weak on specialized taxonomies.
2) LLMs are good at root level taxonomy structure, weak on leaf levels.
3) A novel data organization structure - neural-symbolic structure: internalize the root level and common domain data in LLMs, keep the leaf level and specialized domain data in explicit triple forms.

How to use data

RAG

CRAG (NeurIPS 2024)

Task: Benchmarking the performance of LLMs and RAG systems in answering questions with different timeliness, domains, and popularities

Challenges / Limitations:

Lack of systematic evaluation of LLMs and RAG systems in terms of questions with: 1) different timeliness; 2) different domains; 3) different question types; 4) different popularities.

Solution and Insights:

We create a benchmark to systematically evaluate the performance of LLMs and RAG solutions:

1) Existing solutions are far from perfect in terms of real-time and fast-changing questions.
2) Complex questions such as aggregation questions are difficult for existing methods.
3) Hallucination issue is severe and greatly influence the trustworthiness of the answers.

KERAG (EMNLP 2025)

Task: KG-based RAG

Challenges / Limitations:

KG-based QA offers high precision, but often suffers from low recall. Can we boost recall without sacrificing QA accuracy?

Solution:

We proposed KERAG, a novel KG-based RAG pipeline, which retrieves information at the entity level, rather than the triple level done by traditional methods.

KERAG shows how KG subgraph search + LLM reasoning can yield more complete and accurate answers (+7% accuracy over SOTA, +10-21% over GPT-4o tool model on CRAG).

CacheRAG (Revision VLDB 2026)

Task: KG-based RAG

Challenges / Limitations:

Existing RAG solution often ignores the value of experience replay and continual learning. Can we further boost the performance of KG-based RAG solutions by introducing experience history cache and continual learning?

Solution:

We proposed CacheRAG, a novel cache-based KG RAG solution that utilizes a caching structure to provide useful experience for the LLM-based KG retrieval planner to continuously learn from the QA process.

Our method boosts the SOTA performance by 13-18% on CRAG.