Research Portfolio

Minyoung Choe

Ph.D. Candidate in Artificial Intelligence, KAIST

Advisor: Prof. Kijung Shin

Structure-Aware AI Researcher

From complex data to structure-aware models, retrieval systems, and AI agents

- Structured Data Analysis
- Graph & Hypergraph Learning
- Knowledge Graph Construction
- Agentic AutoML
- Retrieval-Augmented Reasoning

minyoung.choe@kaist.ac.kr github.com/young917 https://young917.github.io

KAIST AI
Kim Jaechul Graduate School

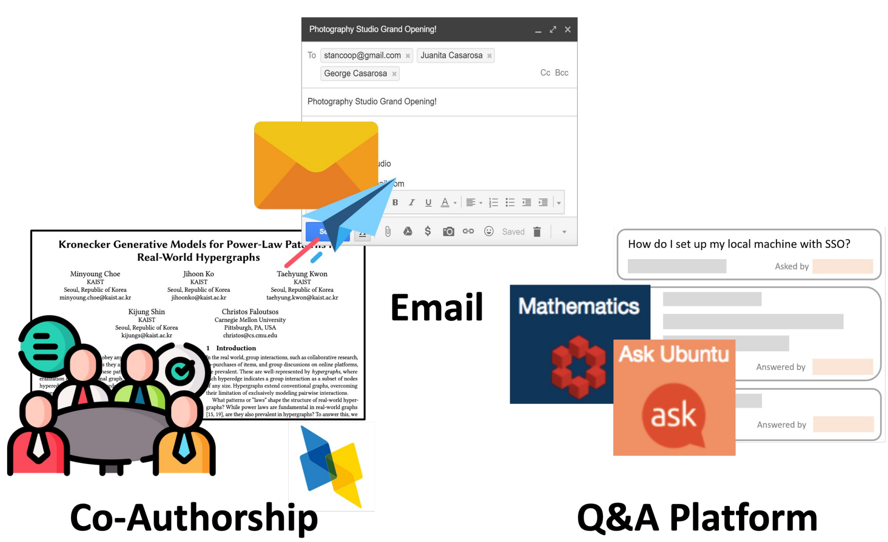
Who I Am: A Structure-Aware AI Researcher

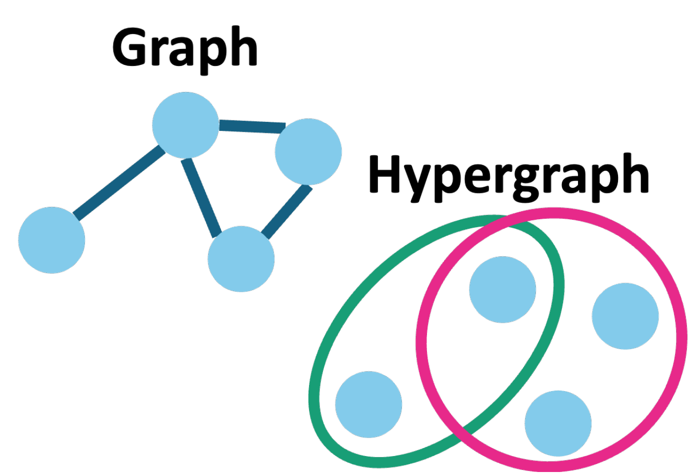
- I study **real-world structured data**, from group interactions to ML pipelines, and scientific knowledge.
- I uncover **structural patterns** in graphs, hypergraphs, and knowledge graphs.
- I design **structure-aware models** and **structured agent memory** for prediction, reasoning, and AutoML.

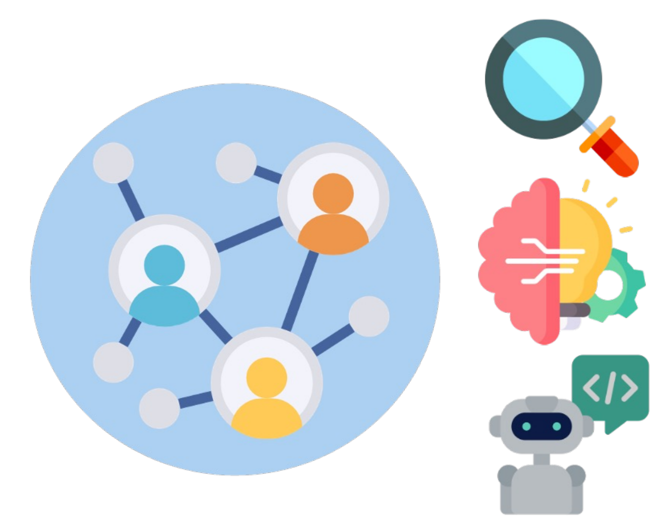
Real-world Data
Co-authorship · Email · Q&A

Structured Representation
Graph · Hypergraph · Knowledge Graph

Structure-Aware AI
Analysis · Modeling · Agent Memory



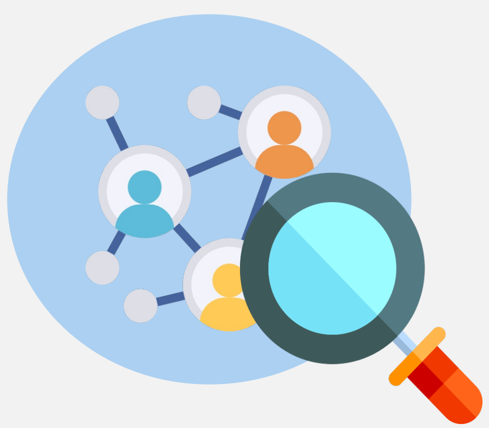




Takeaway: My strength is working with structure end-to-end — from analysis and modeling to AI systems that use it.

Research Trajectory: From Structural Analysis to Agent Memory

- My research has evolved from analyzing real-world structures to designing models and agent memory that use them.



Structural Analysis

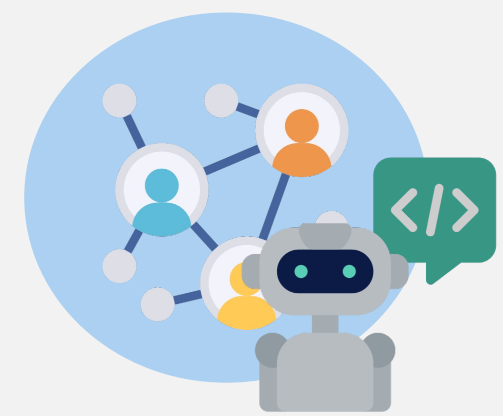
What structural patterns appear in real-world group interactions?

WWW 2021

Structure-Aware Modeling

How can structure be preserved, generated, and used?

WWW 2022 / KDD 2023 / WWW 2025



Structured Agent Memory

How can expert know-how be structured and reused by agents?

KDD Workshop 2026 / Current Research

Research Accomplishment

16
publications in top-tier data mining venues including KDD, WWW, ICDM, and TKDD

6
first-author or co-first-author papers

Takeaway: A coherent research path from structural analysis to models and agent memory.

My Research Style: End-to-End Structure-Aware Research

New Problem Formulation
Define new research problems when existing tasks do not capture the structure of real-world data.

Evidence-Based Hypothesis
Turn observed patterns, metric gaps, and baseline failures into testable hypotheses.

Hypothesis-Driven Design
Design models and algorithms that directly test and use the structural hypothesis.

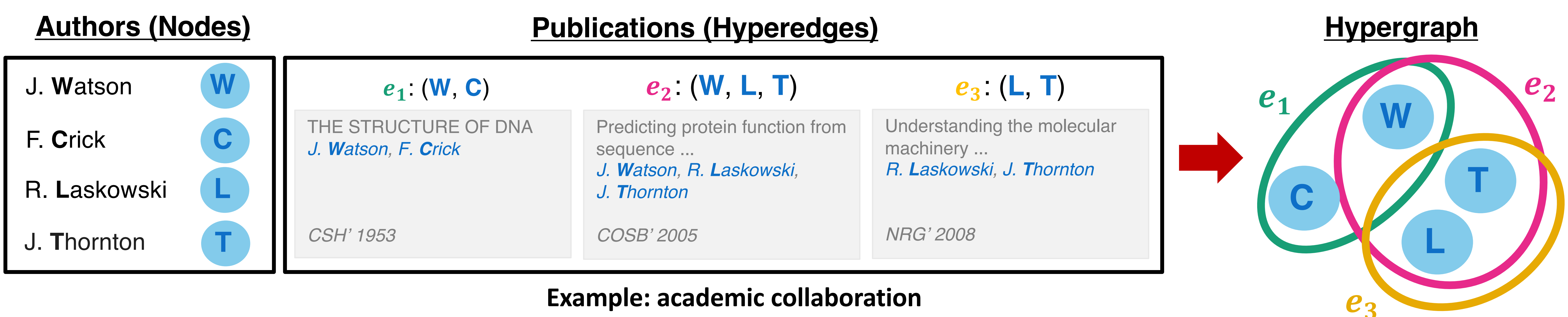
End-to-End Validation
Carry the work from problem definition to implementation, evaluation, iteration, and publication.

Takeaway: I turn structural insight into new research problems and carry them through design, evaluation, and publication.

Research Foundation: Hypergraphs for Modeling Group-Level Interactions

Building research problems from real-world group structure

- **What is a hypergraph?**
  - A **hyperedge** represents **one group interaction** among multiple entities.
- **Why not just graphs?**
  - Graphs split a group into **pairwise edges**, while hypergraphs preserve it as one unit.
- **Why does it matter?**
  - Many **real-world** systems involve groups: co-authorship, email, Q&A, and protein complexes.



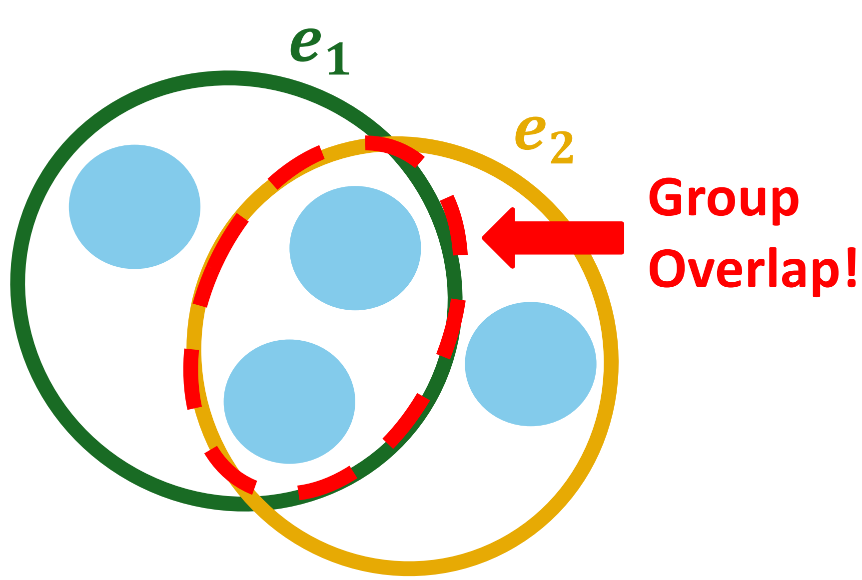
Example: academic collaboration

Each paper forms one hyperedge connecting all co-authors.

Overlap: Understanding How Real-World Groups Overlap

How Do Hyperedges Overlap in Real-World Hypergraphs? — WWW 2021, Co-first Author

01 Research Question



- Real-world groups often **share members**. But graph analysis could not answer:
  How much do groups overlap?

02 Overlap Metric

Metric	Axiom 1	Axiom 2	Axiom 3
Intersection	✗	✗	✗
Union Inverse	✗	✓	✗
Jaccard Index	✗	✗	✗
Overlap Coefficient	✗	✗	✗
Density	✓	✓	✗
Overlapness (Proposed)	✓	✓	✓

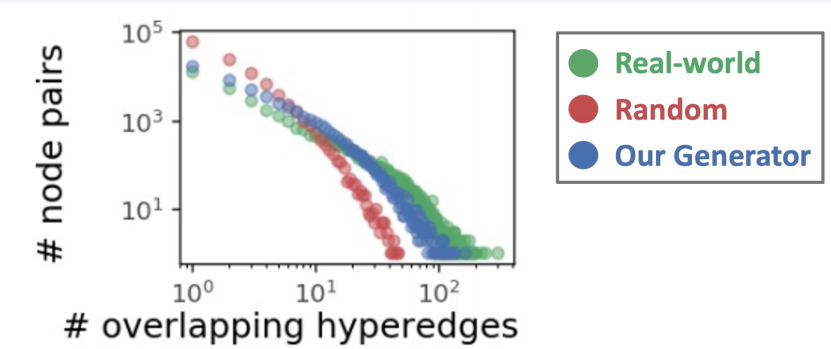
- Defined **overlap measures** to quantify group-level overlap.
- Made overlap patterns **comparable** between real and random data.

03 Metric-Driven Observation

- Compared **real-world vs. random**.
- Found **repeated patterns** unique to real-world groups.

04 Pattern-Driven Generation

- Used real-world overlap patterns as a **similarity guide** for generation.
- Built a generator that **reproduces** realistic group overlap.



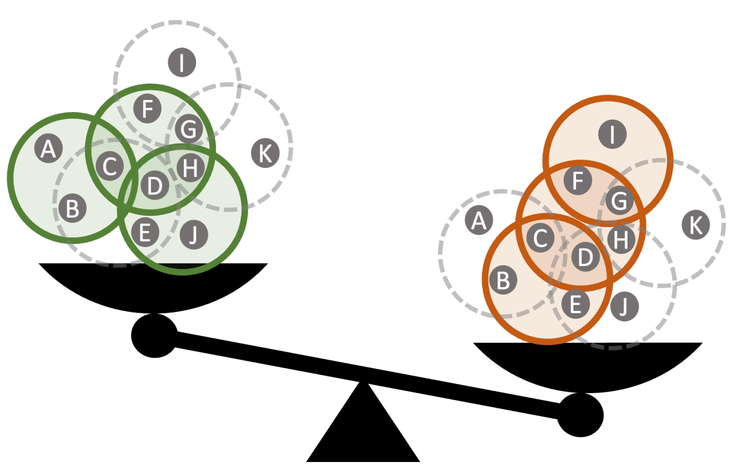
Takeaway: This work turned "groups overlap" into something we can measure, compare, and reproduce.

MiDaS: Sampling Smaller Data Without Losing Structure

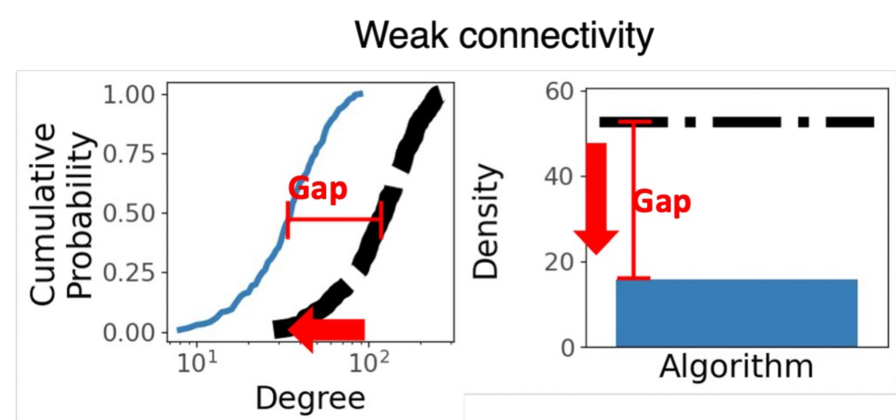
MiDaS: Representative Sampling from Real-world Hypergraphs — WWW 2022, First Author
Representative and Back-In-Time Sampling from Real-world Hypergraphs – TKDD 2024, First Author

01 Research Question

- Large group-structured data is expensive to analyze.
- A smaller sample must preserve the original structure.
- Question: **How can we measure and find the most representative sample?**

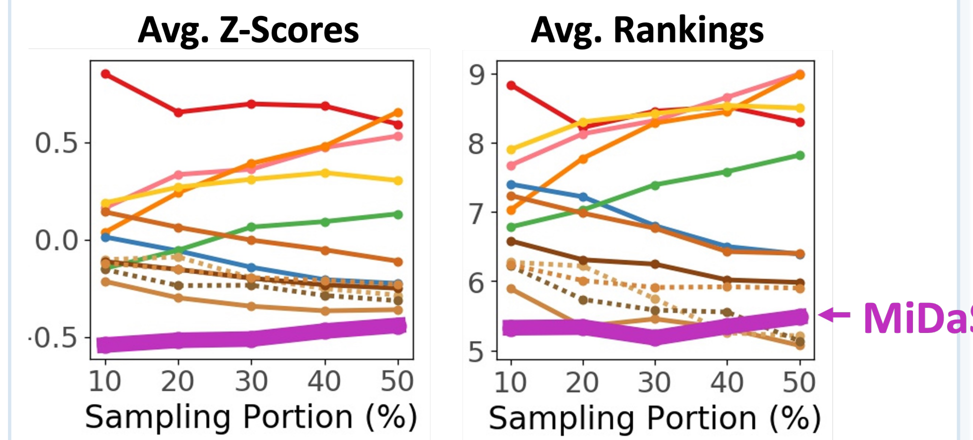


02 Failure → Hypothesis



- **Failure:** Naive sampling missed **important high-degree nodes**. This caused **weak connectivity** and **degree gaps**.
- **Hypothesis**: Preserving important nodes helps preserve structure.

03 Model Design



- **Defined structural criteria** for representative samples.
- Designed a model to sample more hyperedges around **important nodes**.
- **Preserved key structures** effectively across sampling ratios.

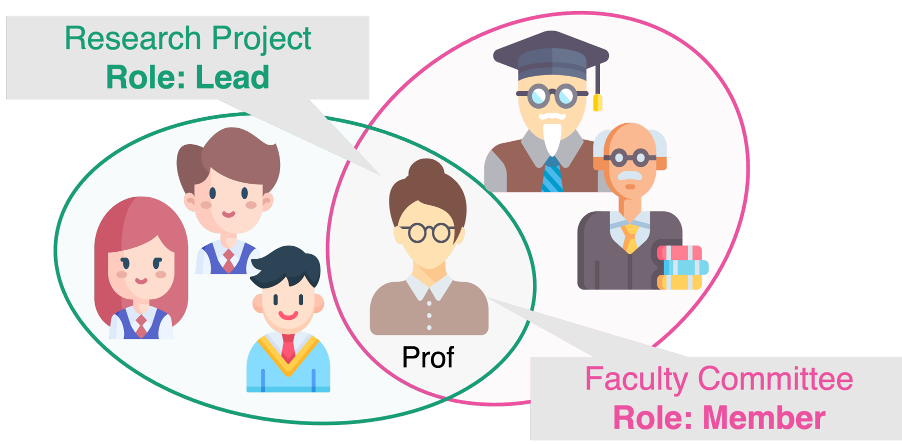
Takeaway: Turned baseline failures into a concrete hypothesis and a structure-preserving sampling method.

WHATsNet: Context-Aware Prediction in Group-Structured Data

Classification of Edge-dependent Labels of Nodes in Hypergraphs — KDD 2023, First Author

01 Research Question

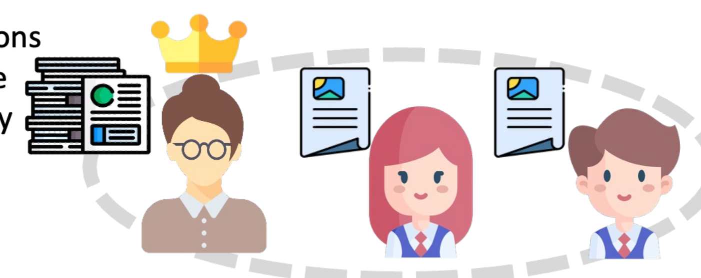
- The same entity can play **different roles** depending on the group.
- Question: **How can we predict the role within each group context?**



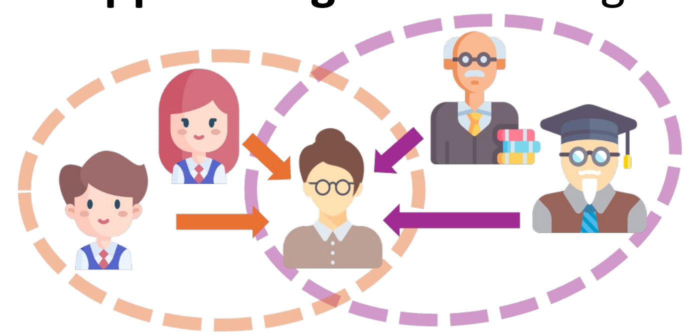
02 Structural Hypothesis

A role depends on both:

- **Structural position** within the group



- **Who appears together** in the group



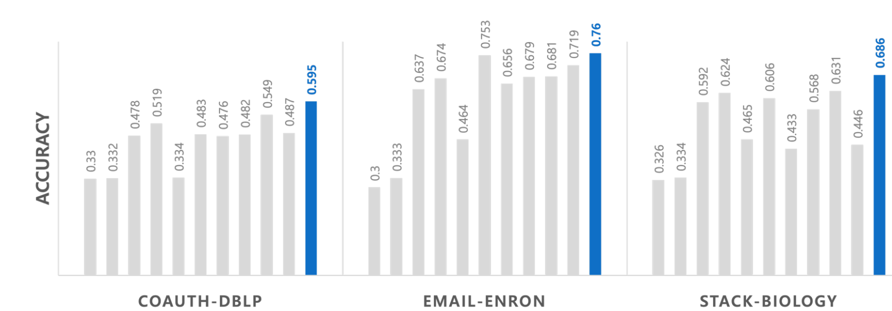
Therefore, each node needs a **group-conditioned representation.**

03 Attention-Based Design

Designed a **hypergraph neural network** with:

- **Positional encoding** from structural rankings
- **Self-attention** within each hyperedge
- **Group summaries** for scalable attention

Result: Learns **group-conditioned node embeddings** for prediction.



Takeaway: This work made group context learnable for predicting context-dependent roles.

HyRec: Explaining Group Growth with a Compact Model

Kronecker Generative Models for Power-Law Patterns in Real-World Hypergraphs — WWW 2025, First Author, Collaboration with Prof. Christos Faloutsos, CMU

01 Research Question

- **Observation:** Real-world group data grows with **repeated structural patterns**.
- **Question:** Can a **compact model** explain and reproduce this growth?

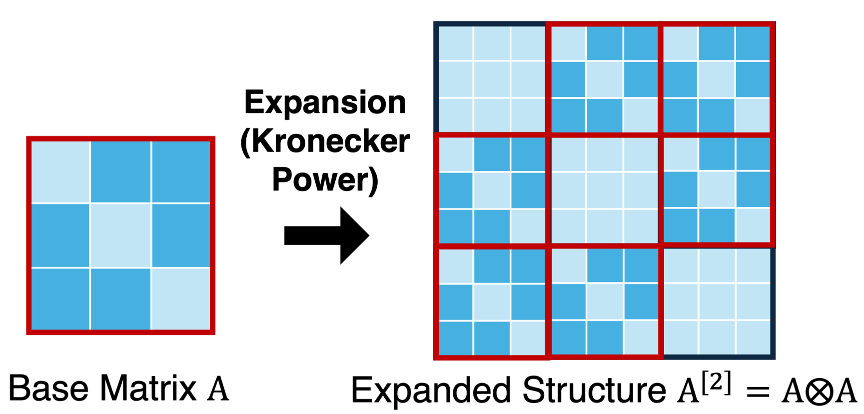
02 Modeling Hypothesis

- **Hypothesis:** A **small base pattern** can be repeatedly expanded to form large group structures.
- **Theoretical result: Kronecker-style expansion** reproduces key structure and growth patterns.

03 Reproduction & Forecasting

- Generated realistic group structures **with few parameters**.
- Used for:
  - **Reproducing** real group data
  - **Forecasting** future growth
- Scales efficiently to large datasets.

Kronecker-style expansion



HyRec matches real data patterns with far fewer parameters

	# Input Parameters		Average Ranking (Across 11 Hypergraph Datasets)									
	Min	Max	Degree	Size	Pair Deg.	Intersect.	Singular Value	Clustering Coef.	Density	Overlapness	Effective Diam.	Average
HyperCL	11,028	2,852,295	**1.455**	**1.000**	3.455	4.909	*2.364*	4.545	3.636	4.364	*3.455*	3.242
HyperFF	**2**	**2**	4.909	4.818	5.000	4.273	5.182	4.182	4.273	3.909	3.818	4.485
HyperPA	11,028	2,852,295	3.000	3.727	*3.000*	4.273	4.545	4.636	4.000	4.273	3.545	3.889
HyperLAP	11,028	2,852,295	*2.636*	**1.000**	**2.636**	**1.636**	3.182	**1.364**	*3.273*	*2.727*	4.727	**2.576**
THera	10,889	1,591,170	4.909	4.091	3.091	*2.364*	4.545	*2.636*	*3.273*	3.182	3.909	3.556
HyRec	*15*	*882*	4.091	4.818	3.818	3.545	**1.182**	3.636	**2.545**	**2.545**	**1.545**	*3.081*

Takeaway: I turned repeated structural patterns into a theory-backed model for realistic generation and forecasting.

Building Structured Memory for Agentic AutoML

- I built an agentic AutoML system and structured scattered ML know-how into reusable pipeline memory.

01 Problem

AutoML is an ML pipeline composition problem

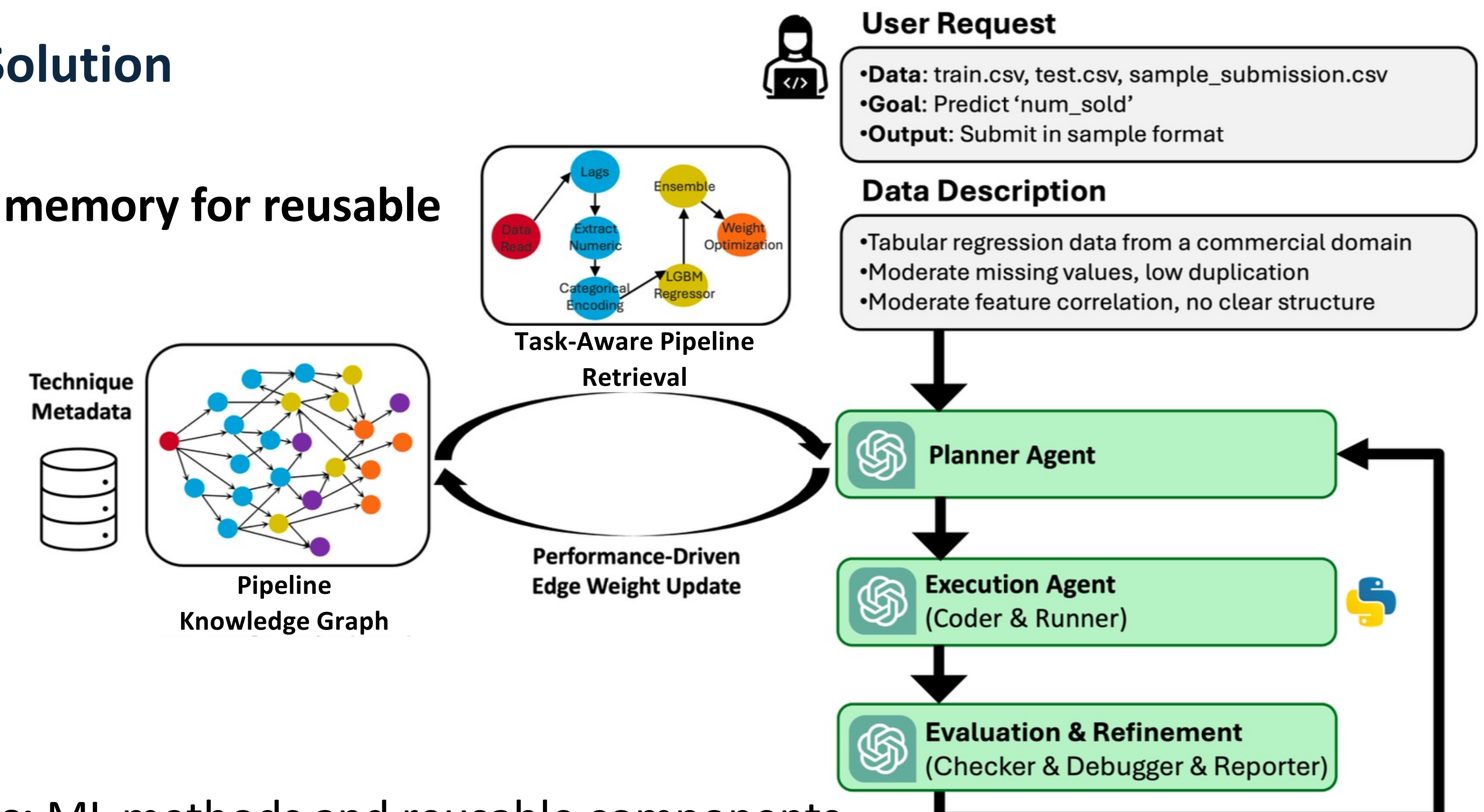
ML solutions require linked decisions: preprocess → feature → model → train → evaluate

Text-only RAG loses pipeline structure

It retrieves techniques as text, but often misses: step order, dependencies, technique combinations, performance feedback

02 Goal & Solution

Build structured memory for reusable ML pipelines



- **Technique nodes:** ML methods and reusable components
- **Dependency edges:** dataflow, step order, and method relations
- **Task-aware weights:** performance signals for reusable ML pipelines
- **Agent usage:** retrieve, plan, execute, debug, and update ML pipelines

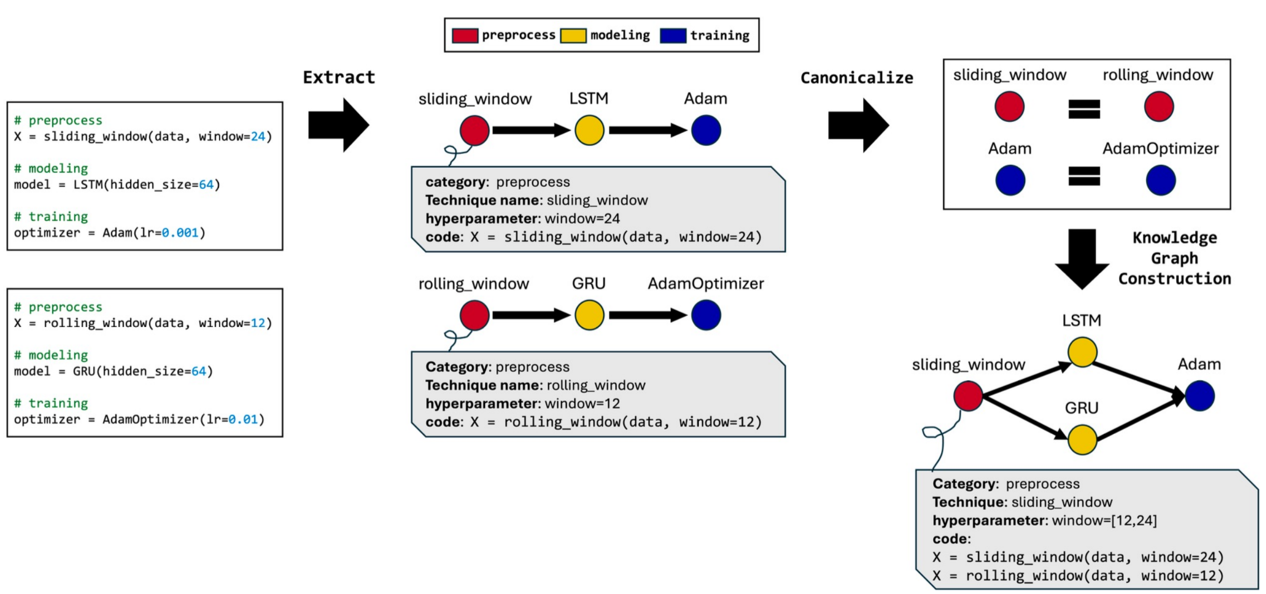
Takeaway: I built a multi-agent AutoML system with structured, reusable memory for ML pipelines.

From ML Code to Graph and Hypergraph Agent Memory

- I structure ML code into two levels of memory: step-level graphs and block-level hypergraphs.

Step-Level Graph Construction

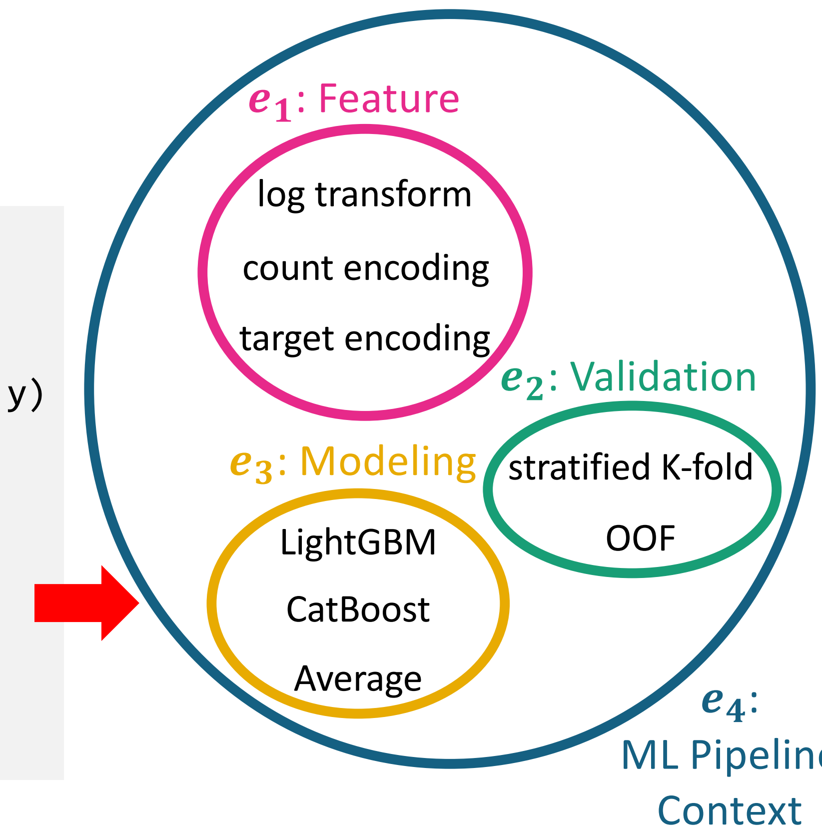
- **Extract ML pipelines** from kaggle/github code.
- **Link steps** by dataflow, order, and dependencies.
- **Normalize** method aliases into reusable technique nodes.
- **Store implementation context:** code and parameters.



Block-Level Knowledge Hypergraph

Hyperedges capture techniques that work together as one reusable block.

```
# Feature block
X["log_price"] = log1p(X["price"])
X["cat_count"] = count_encode(X["category"])
X["cat_target"] = target_encode(X["category"], y)
# Validation block
folds = StratifiedKFold(n_splits=5)
oof_pred = np.zeros(len(X))
# Modeling block
lgb_pred = train_lightgbm(X, y, folds)
cat_pred = train_catboost(X, y, folds)
final_pred = average([lgb_pred, cat_pred])
```



- **Hyperedge memory:** preserves useful technique combinations.
- **Neighbor augmentation:** retrieves connected blocks as planning context.
- **Feedback propagation:** updates related blocks after execution results.

Takeaway: I structure ML code into graph and hypergraph memory that agents can retrieve, reuse, and improve.

Structured Memory Guides Better AutoML Search

- I ran a controlled comparison: the same AutoML agent with vs. without structured memory.
- Goal: test whether structured memory changes the agent's modeling direction.

	Description	Baseline (ML-Master)	Our (ML-Master + Structured Memory)
champs-scalar-coupling ↓	molecule regression	-0.38	**-0.78**
yelp-dataset ↓	link prediction	1.24	**1.15**
goodbooks-10k ↓	rating prediction	0.90	**0.89**

↓ *Lower is better*

Case Study: champs-scalar-coupling

- **Task challenge:** scalar coupling depends on atomic interactions and 3D molecular geometry.

Baseline Agent

ML-Master without structured memory

- **Feature-based molecular modeling:**
- Pair-distance features + LightGBM

Retriever-Augmented Agent (Our)

ML-Master with structured memory

- **Geometry-aware molecular modeling:**
- DimeNet-style GNN with pair-index readout
- Attentive kNN GNN

Takeaway: Structured graph/hypergraph memory helped the agent explore more domain-aware modeling directions.

Finding Similar Past Satellite Videos for Forecasting Support

- I built a retrieval system that helps forecasters find similar past weather cases from large satellite video archives.

01 Problem

No Visual Similarity Labels

Historical satellite videos have no ground-truth pairs for “similar weather cases.”

Massive Historical Satellite Video Archive

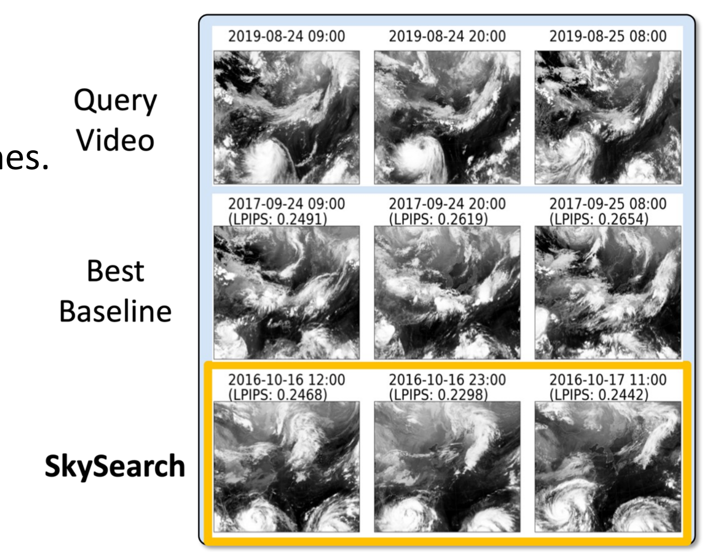
Decades of high-resolution videos make past-case search impractical.

02 Goal & Solution

Build SkySearch to retrieve similar past cases without similarity labels

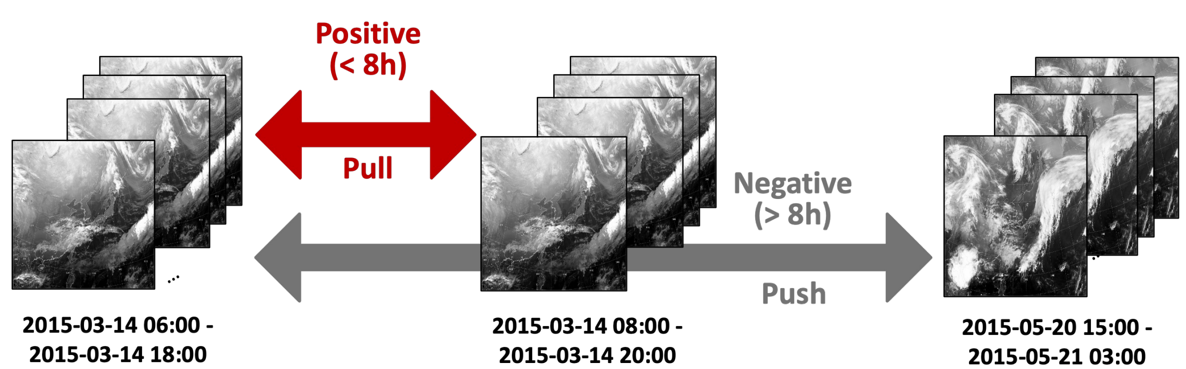
- **Temporal self-supervision:** nearby timestamps become positives because weather changes gradually.
- **Fast graph search:** search a latent KNN graph instead of brute-forcing the full archive.
- **Future-aware query:** use current + predicted frames for longer-range matching.

Retrieved similar weather cases



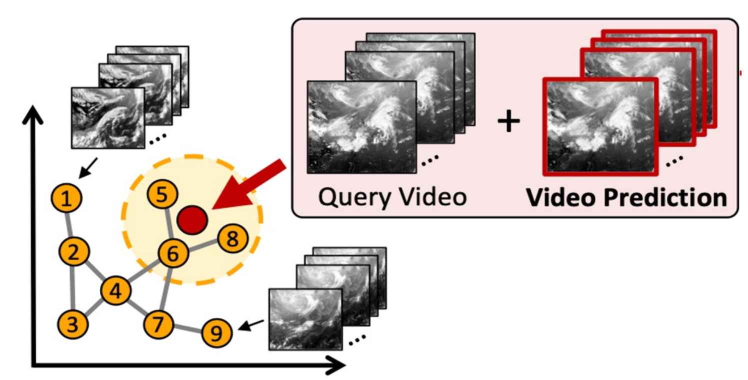
Self-supervised training

nearby videos are positives; distant videos are negatives.



Graph Search with Future-Aware Query

Search the latent KNN graph using current + predicted frames.



Takeaway: I built self-supervised satellite video retrieval for a deployed forecasting support system.

Multi-View Retrieval for Time-Series LLM Prediction

- I used time-series foundation models as retrievers and analyzed which evidence formats best support LLM prediction.

01 Multi-View Retrieval

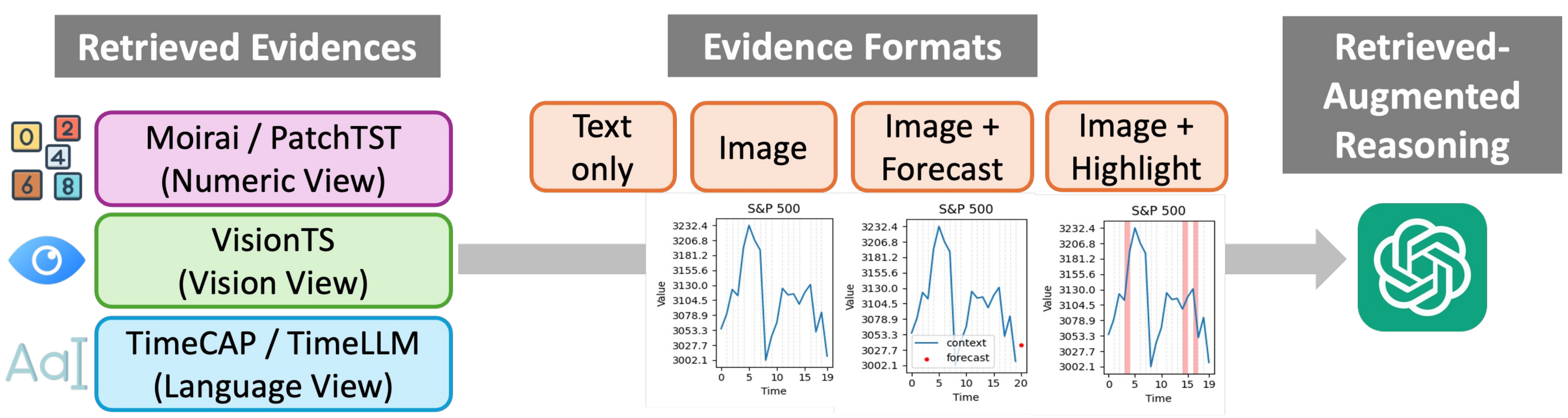
- Retrieve similar cases using foundation models:
  - **Numeric:** value-level similarity
  - **Visual:** shape and trend cues
  - **Language:** semantic/domain context

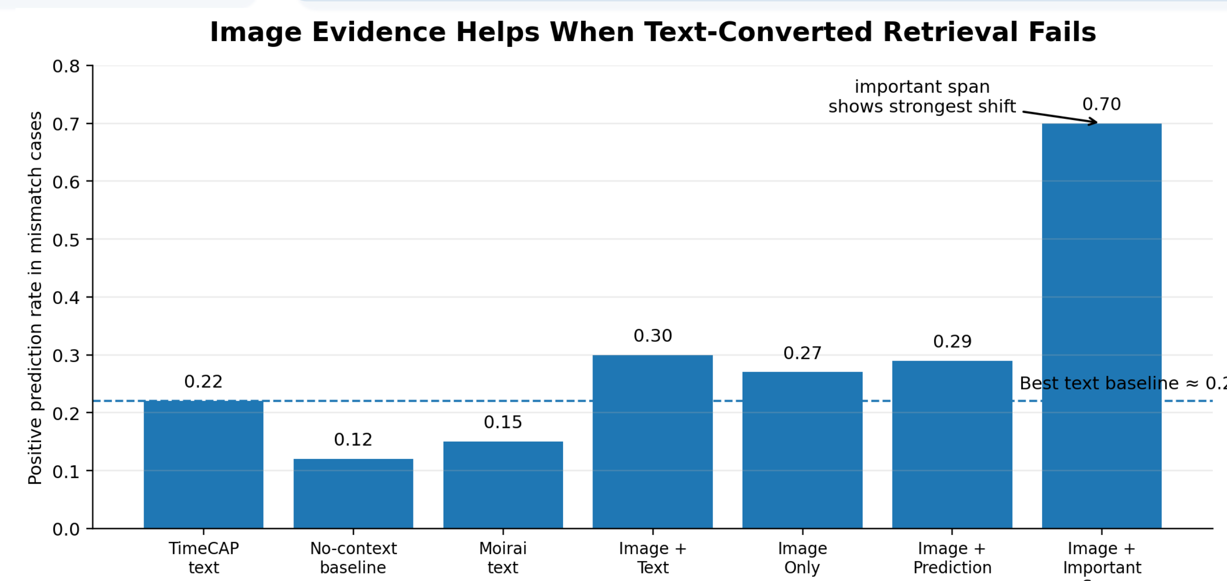
02 Evidence Formats

- Convert retrieved cases into different LLM inputs:
  - **Text summary, Image, Image + forecast, Image + highlight**

03 Key Finding

- No single view or format worked best across all datasets.
- **Image inputs helped** when text summaries failed.
- **Highlighted spans** produced the **largest** prediction change.





Takeaway: I showed how retrieval views and evidence formats shape LLM-based prediction.

Scientific Evidence Reasoning for R&D Support

- I identified key **scientific QA requirements** and designed an **evaluation set** for R&D evidence reasoning.

01 Problem

Scientific QA often misses realistic evidence reasoning

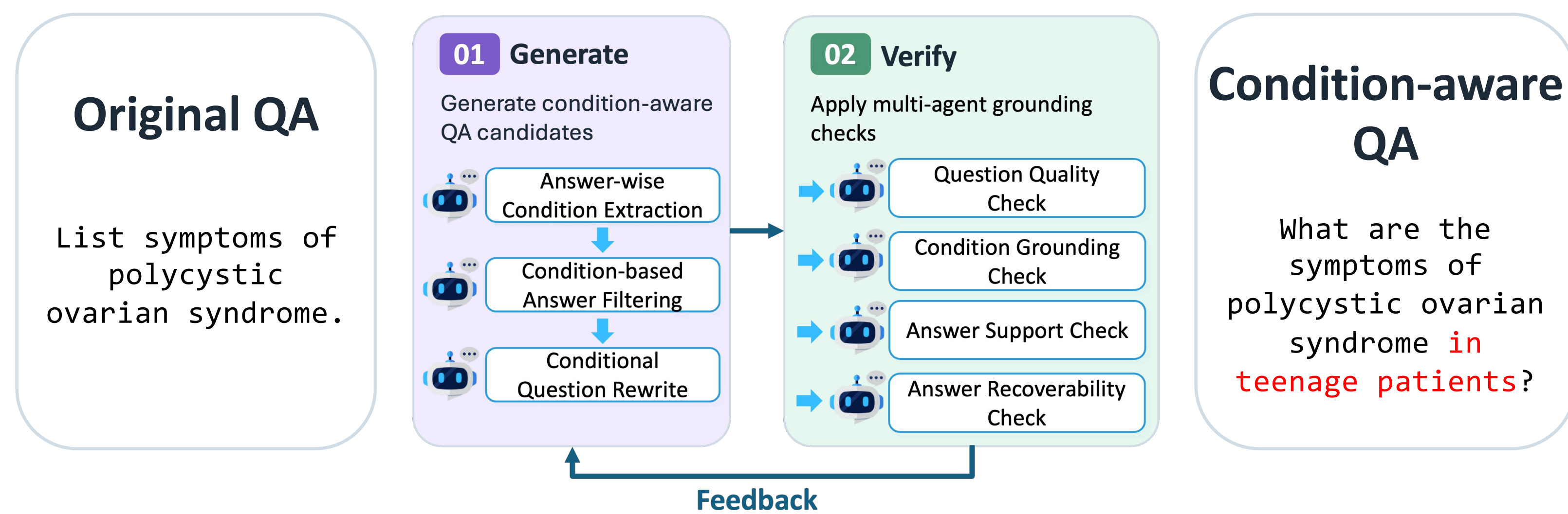
- Many QA tasks focus on **direct fact recall**
- Real findings depend on **conditions and contexts**
- Evidence is often **scattered or inconsistent across papers**

Condition-dependent	"Drug A reduces tumor growth in mice, but not in humans"
Multi-document	Paper 1: Drug A → lowers Inflammation Paper 2: Lower inflammation → improves Disease B Q: Relationship between Drug A and Disease B?

02 Goal & Solution

Build **paper-grounded QA data** with **multi-agent verification**

- Ground each QA pair in **explicit paper evidence**
- Transform questions into **condition-aware reasoning** tasks
- Use **multi-agent checks** for quality, grounding, and answer support



Takeaway: I turned scientific QA challenges into paper-grounded evaluation data with multi-agent verification.

Selected Publications

Hypergraph Structure Analysis

Group-interaction Analysis

How Do Hyperedges Overlap in Real-World Hypergraphs? — Patterns, Measures, and Generators
WWW 2021 · Co-first Author

Hypergraph Structure Modeling

Sampling · Generation · Context-aware prediction

Kronecker Generative Models for Power-Law Patterns in Real-World Hypergraphs
WWW 2025 · First Author

Representative and Back-In-Time Sampling from Real-world Hypergraphs
TKDD 2024 · First Author

Classification of Edge-dependent Labels of Nodes in Hypergraphs
KDD 2023 · First Author

MiDaS: Representative Sampling from Real-world Hypergraphs
WWW 2022 · First Author

Structured Agent Memory

Structured workflow memory for AutoML agents

TacitFlow: Learning Workflow Representations for Tacit-Knowledge-Grounded Machine Learning Engineering Agents
KDD 2026 Workshop AIDataScientist Co-author

Deployed AI System

Self-supervised satellite video retrieval for operational forecasting

SkySearch: Satellite Video Search at Scale
KDD 2025 Industry Track · Co-first Author

Structured Data Applications

Beyond hypergraphs: EHR and financial networks

Temporal Graph Networks for Graph Anomaly Detection in Financial Networks
AAAI 2024 Workshop · Co-author

GraphEHR: Heterogeneous Graph Neural Network for Electronic Health Records
IJCAI 2024 Workshop · Co-first Author