Powering the Future: How Multi-Model Databases Empower Intelligent Education & Scientific Discovery?

Sheng Wang

Email: swangcs@whu.edu.cn

http://sheng.whu.edu.cn/

2025-11-8@IEIR-CCNU

Totem Database Lab
http://totemdb.whu.edu.cn/

[1]School of Computer Science
[2]School of Artificial Intelligence
[3]Research Center for Digital and Intelligent Teaching and Education
Wuhan University

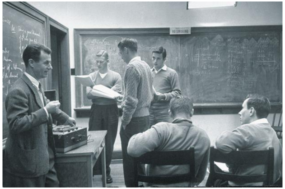
Outline

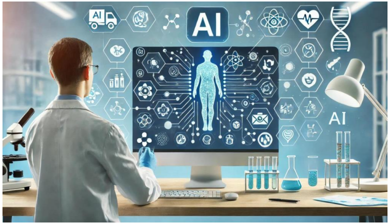
1. Intelligent Edu & SCI: Background and Challenges
2. Multi-model Database System: A Solid Cornerstone for IE&IR
3. Clustering-based Data Summarization
4. Connectivity-aware Spatial Dataset Search
5. Collaboration-driven Scientific Data System
6. What's novel and next in MMDB for IE&IR?

Intelligent Edu & SCI: Background and Challenges

What is going on in universities?

- Education
  - Online Learning
  - Open University
  - Personalization-driven
- Research
  - AI4S
  - Computation-intensive
  - Data-intensive



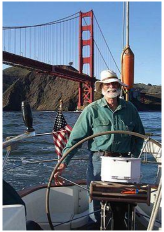


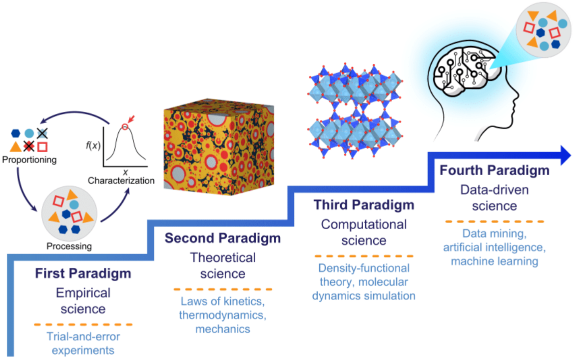
5 paradigms of science:

empirical, theoretical, computational, data-intensive, and AI-driven

- Microsoft researcher Jim Gray described the evolution of scientific paradigms:

James Nicholas Gray (1944 – declared dead in absentia 2012) was an American computer scientist who received the **Turing Award** in 1998 “for seminal contributions to **database** and transaction processing research and technical leadership in system implementation”.

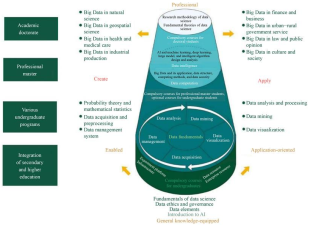




Ioannidis, Yannis. "The 5th paradigm: AI-driven scientific discovery." *Communications of the ACM* 67.12 (2024): 5-5.

Digital education is empowered by AI

- **Education Modernization 2035**
  - This blueprint emphasizes using technology like AI, big data, and VR to create a modern, flexible, and lifelong education system.
  - **National Smart Education Platform:** A free public platform providing curricular resources for students from primary to high school, professional development for teachers, and services for vocational and higher education.
- **Wuhan University's Digital Intelligent Education**
  - "Student-Centered, Data-Driven, AI-Enabled"
  - Seamless integration of technology into teaching, research, and campus life.
  - Aligns with national strategies.



Zhang, Pingwen. "Digital Intelligence Education at Wuhan University: Practice and Innovation." *Frontiers of Digital Education* 2.1 (2025): 1.

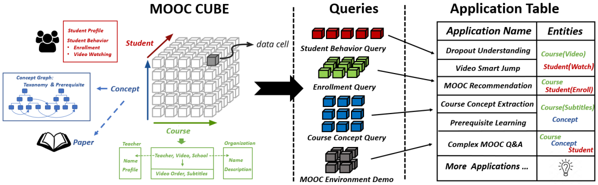
Scientific discovery is empowered by AI

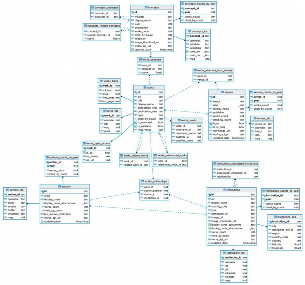
- **Deep Research**
  - Accelerating and Augmenting Human Intelligence
  - Generating and Prioritizing Hypotheses
  - Autonomous Design and Execution of Experiments
  - Modeling and Simulation of Complex Systems

- **Real-World Examples:**
  - **Medicine:** AI is used to discover new antibiotics (like halicin, discovered by an MIT AI model), design personalized cancer treatments by analyzing a patient's tumor, and predict pandemic spread.
  - **Physics:** At the Large Hadron Collider (LHC), AI algorithms sift through petabytes of collision data to find the incredibly rare events that might indicate new physics.
  - **Materials Science:** AI is used to discover new alloys, superconductors, and battery components with specific desired properties.

Data-centricity represents the fundamental commonality

- 3V of big data
  - **Volume**
  - **Velocity**
  - **Variety**
- MoocCube Dataset
  - With over **1.2 billion** behavioral events, it was one of the largest public MOOC datasets of its time, enabling robust and generalizable machine learning model training.
- OpenAlex Dataset
  - A massive, open-source bibliographic database that aims to catalog the world's scholarly research and the connections between its entities: **works, authors, institutions, concepts, and sources**.
  - **240 million** works, **210 million** authors, and **120,000** concepts, **2-3 million research papers annually**

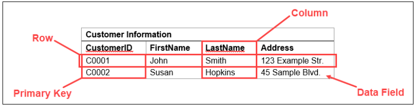




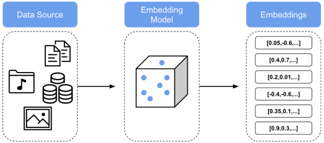
1. Intelligent Edu & SCI: Background and Challenges

What mainly varies are the data models

- **Relational Model** - Traditional table-based structure with rows and columns
- **Document Model** - Stores semi-structured data in JSON, XML, or BSON formats
- **Graph Model** - Represents data as nodes and edges for complex relationships
- **Key-Value Model** - Simple key-value pairs for fast retrieval
- **Vector Model** - **The most popular one in the LLM Era**
  - Captures Meaning and Relationships (Semantics)
  - Enables the Core LLM Architecture (Transformers)
  - Unifies Different Types of Data (Multi-Modality)




```
@article{wang2021survey,
  title={A survey on trajectory
  author={Wang, Sheng and Bao,
  journal={ACM Computing Survey
  volume={54},
  number={2},
  pages={1--36},
  year={2021},
  publisher={ACM New York, NY,
}
```



1. Intelligent Edu & SCI: Background and Challenges

Key tasks in intelligent education & research: 3S

- **Data Summarization**
  - Student **grouping**
  - Literature review
- **Data Search**
  - Scientific Dataset Preparation
  - Scientific Talent Seeking
- **Data System**
  - Google Scholar
  - Google Dataset Search
  - Coursera

ZOOM BREAKOUT ROOMS

Google Dataset Search Beta

Search for Datasets

Google Scholar

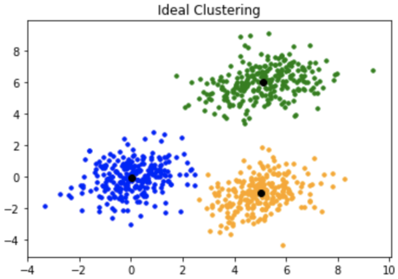
https://www.icourse163.org/

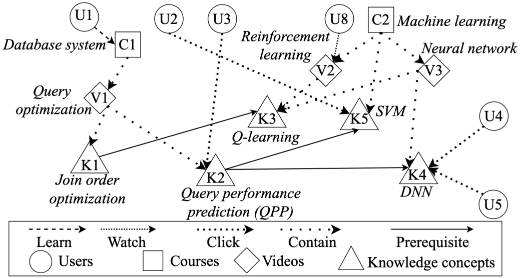
https://www.coursera.org/

Challenge 1: Cross-model data summarization

- Clustering on vector data to divide data points into groups.
  - K-means
  - K-center
  - …

What happens with multi-model data?

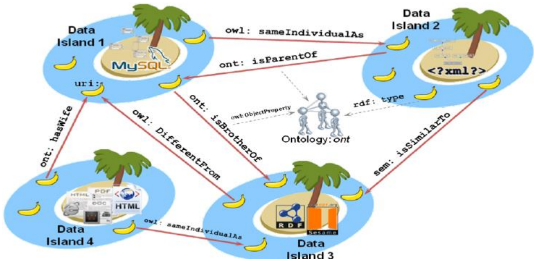


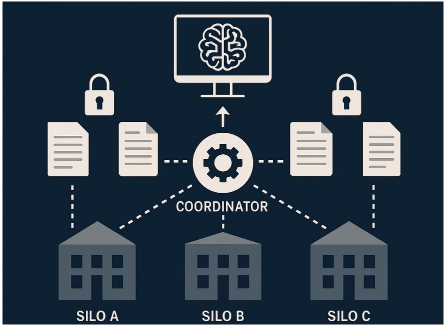


Zhang, Juntao, Sheng Wang, Yuan Sun, and Zhiyong Peng. "Prerequisite-Driven Fair Clustering on Heterogeneous Information Networks." *Proceedings of the ACM on Management of Data* 1, no. 2 (2023): 1–27.

Challenge 2: Cross-silo search over data islands

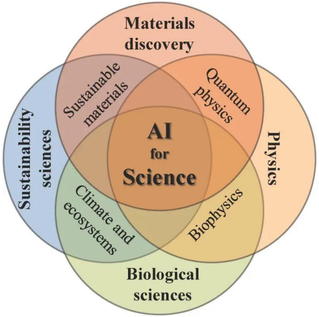
- A data island is an isolated or disconnected store of data within an organization that is not easily accessible to other departments or systems.

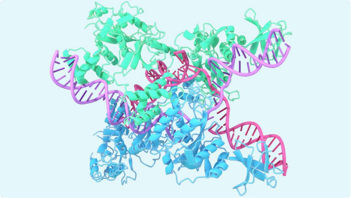


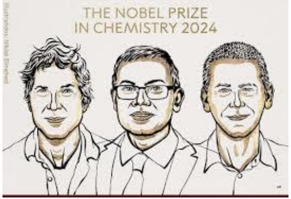


Challenge 3: Cross-disciplinary global research system

- Existing systems exhibit domain specificity owing to inadequate big data processing capabilities.



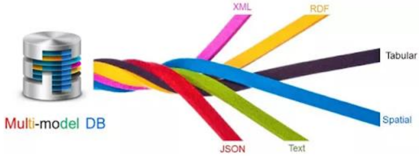


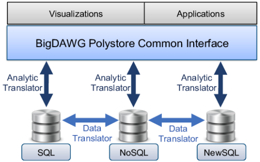


Multi-model Database System:
A Solid Cornerstone for IE&IR

What is the multi-model database system?

- **Definition:**
  - A multi-model database (MMDB) is a database management system designed to support multiple data models against a single, integrated backend.
- **Main feature:**
  - Instead of using separate databases for different types of data, a multi-model database can handle various data formats within one unified system.





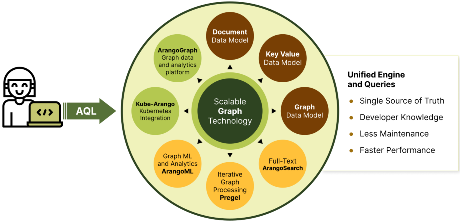
Why is multi-model database system crucial?

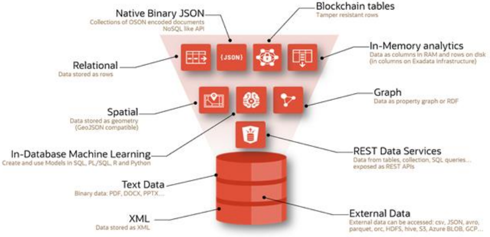
- **Simplified Architecture**
  - No need to manage multiple databases
  - Reduced operational complexity
  - Single system to learn and maintain
- **Improved Performance**
  - Optimized storage and retrieval for each data model
  - Faster queries by leveraging appropriate models
  - Unified query language across models
- **Increased Flexibility**
  - Handles structured, semi-structured, and unstructured data
  - Adapts to changing business requirements
  - Supports diverse data formats without reformatting
- **Cost Efficiency**
  - Reduces infrastructure costs
  - Lower development and maintenance overhead
  - Eliminates data duplication across multiple systems

2014 Turing Award Winner for his contributions to database research

Existing MMDBs: Open-source vs. Oracle

- Popular multi-model databases include:
  - **ArangoDB** - Supports documents, graphs, and key-value pairs
  - **Azure Cosmos DB** - Supports multiple models via different APIs
  - **Couchbase** - Key-value and JSON documents
  - **Redis** - Key-value with extensions for other models
  - **OrientDB** - Graphs, documents, and key-values
- Oracle 26AI
  - AI-native functionality
  - Multi-modal and multi-cloud data processing
  - Performance and security enhancements





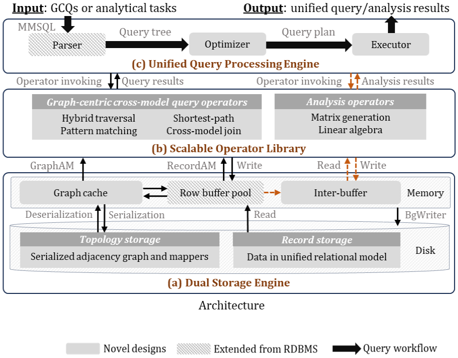
Shortcomings

- Cross-model **analysis operator** is not supported
  - Systems like DuckDB only have plugin-based multi-model support
  - ArangoDB, AgensGraph simply transform multiple models into their first-model
  - None of these systems designed cross-model analysis operators
- Cross-silo distributed data search is not supported
  - Heterogeneous data formats and query languages
  - Absence of federated metadata and global optimization

Framework of our newly developed GredoDB

Our multi-model database works by:

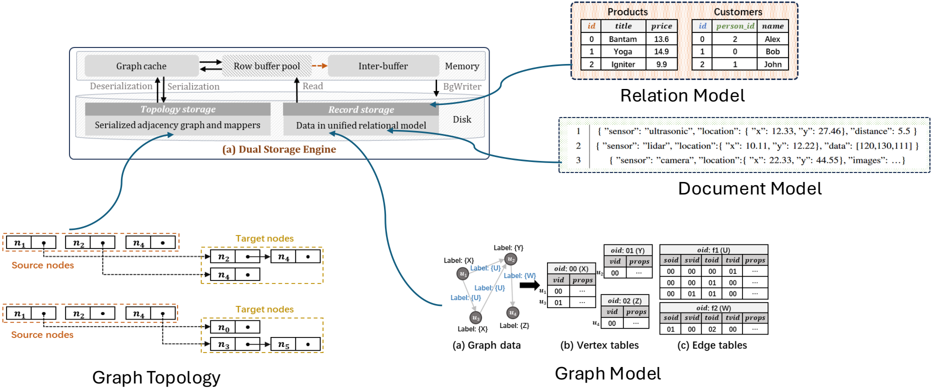
- **Unified Storage**: All data stored in a single representation
- **Easy-to-use Query Language**: One language to access all data models
- **Global Optimization**: Engine optimizes storage and retrieval based on data model



Architecture

Unified Storage Engine for data variety

- Unified Data Modeling in the dual storage engine.



Cross-model Query Optimizations for data volume

- Cross-model Query Optimization refers to the process of optimizing queries that span multiple data models (such as relational, graph, vector, and document data) within a multi-model database system.

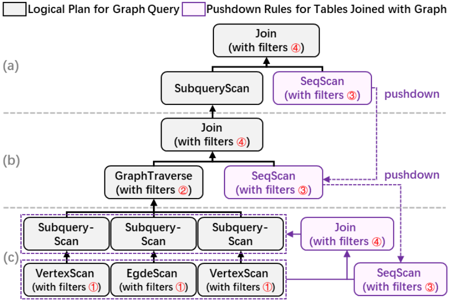
Optimization rules:

① Predicate-level pushdown (on Graph)

② Topology checking with optimized operator

③ Predicate-level pushdown (on Table)

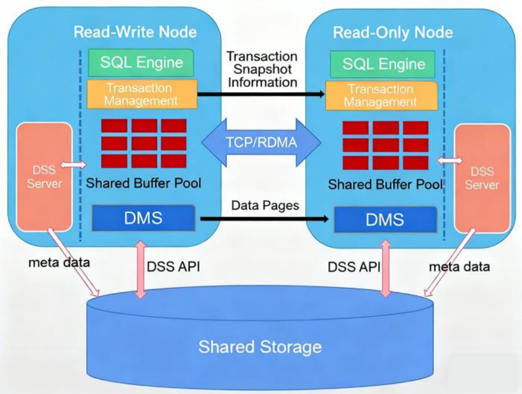
④ Collection-level pushdown



2. Multi-model Database System: A Solid Cornerstone for IE&IR

Distributed multi-model transaction for data velocity

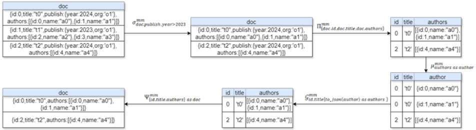
- Distributed architecture for big data scenarios
  - Data is stored in a shared storage
  - Data processing adopts a distributed architecture
- Transaction management based on network protocols
- Coherent cache across nodes, orchestrated by distributed memory service (DMS)



MMSQL: A novel query language for end users

- High usability and user-friendly
  - MMSQL is backward-compatible with SQL
  - Supports parts of the SQL/JSON and SQL/PGQ standards
- Multi-model support
  - Supports multiple models, including relational, document, graph, and vector

```
[ WITH name1 AS subquery1 [, name2 AS subquery2, ... ] ]
SELECT [ DISTINCT ] select_list
[ FROM [ relation [AS alias]
       | document [AS alias] [ UNWIND unnest_fun(col_name) [AS alias] [UNWIND ... ] ]
       | graph MATCH pattern
       | vector [AS alias] ]
  [ JOIN model [ON join_condition [, ... ] ]
              | UNWIND unnest_fun(col_name) [AS alias] [ UNWIND ... ]
  ] [, ... ]
]
[ WHERE query_condition ]
[ GROUP BY group_by_condition ]
[ HAVING group_condition ]
[ ORDER BY [ scalar1 [, ASC | DESC] ] ]
[ LIMIT limit_number ]
```

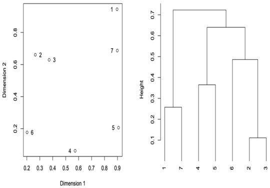



```
select {'from': a.title,'to': b.title , 'similarity': t.vec <-> a_vec.vec} as json_result
    from (select id,vec from Paper_vector p_vec order by p_vec.vec <-> @article_vector limit 10) as t,
        Citations match (a : Paper)-[e: Cite]->(b : Paper),
        Paper_Info p unwind json_array_elements(p.keywords::json) keyword,
        Paper_vector a_vec
    where t.id = b.paper_id and a.paper_id = p.id and p.id = a_vec.id  -- condition of multi-model joining
        and p.fos.name = @fos_name and keyword::varchar = @keyword -- condition of filtering
    order by t.vec <-> a_vec.vec asc
    limit 100
```

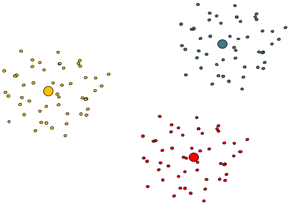
Clustering-based Data Summarization

Why clustering for data summarization?

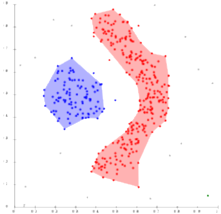
- Clustering is a powerful technique for data summarization because it organizes large volumes of data into meaningful groups, making complex information more manageable and interpretable.



Hierarchical clustering



Partition-based



Density-based

Sheng Wang, Yuan Sun, Zhifeng Bao. On the Efficiency of K-Means Clustering: Evaluation, Optimization, and Algorithm Selection. **PVLDB** 2021 14(2): 163 - 175,

3. Clustering-based Data Summarization

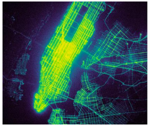
Existing clustering models and broader applications

- The *k*-means algorithm partitions data into clusters with similar characteristics, enabling efficient analysis, compression, and downstream learning.

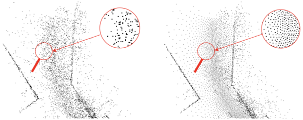
Possible objective:

$$\sum_{C_k}\sum_{x_i \in C_k}(x_i - \mu_k)^2$$

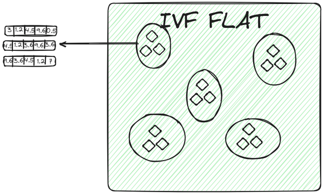
This is the sum of squared errors for each data point x_i, assuming that each x_i is mapped to the closest cluster center μ_k.



Spatial clustering



Point cloud clustering



Yushuai Ji, Zepeng Liu, **Sheng Wang**, Yuan Sun, Zhiyong Peng. On Simplifying Large-Scale Spatial Vectors: Fast, Memory-Efficient, and Cost-Predictable k-means. **ICDE**, pp.863-876, 2025.

Common shortcomings

- Small clusters and large clusters

- Vector-oriented only
  - One typical example is **semantic grouping of students' open-ended answers**.

Student Answers

A: Parabola has highest/lowest point
B: Quadratic has extremum
C: I just follow textbook
D: Symmetry axis -> vertex
E: Curve bends -> turning point

Balanced Clustering

Group 1: A, B (strong)
Group 2: D, E (partial)
Group 3: C (needs support)

Outcome: Equitable group sizes, scalable instruction

Unbalanced Clustering

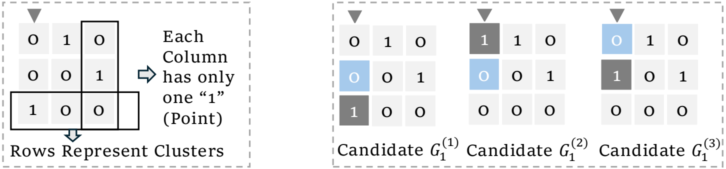
Example:
Group 1: A, B, D, E
Group 2: C alone

Problems:
- One isolated learner
- One giant group
- Hard to differentiate instruction

  - Balanced clustering groups students into similarly sized clusters while preserving semantic similarity, ensuring no student is isolated and teaching remains scalable and fair.

Balanced clustering to generate equal-size clusters

- Our method
  - Loss function: Total loss(G) = Clustering loss(G) + Penalty Term(G)
  - Design a new indicator matrix with rows for classes and columns for points, as shown in Figure (a).
  - Select the optimal indicator matrices $\{G^1, G^2, \ldots, G^k\}$ to minimize the loss, as shown in Figure (b).



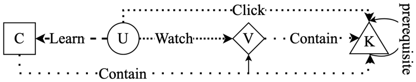
(a) A $\mathbf{3} \times \mathbf{3}$ indicator matrix G

(b) Candidates of G_1

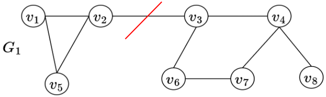
Yushuai Ji, Shengkun Zhu, Shixun Huang, Zepeng Liu, **Sheng Wang**, Zhiyong Peng. Federated and Balanced Clustering for High-dimensional Data. **PVLDB**, 18(11): pp.4032-4044, 2025.

Cross-model clustering in HINs for fair student grouping

- We further achieve balanced clustering in Heterogeneous Information Network (HIN)

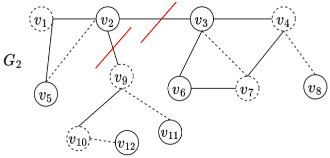


Personalized learning path of user U by watching video V and clicking on knowledge K

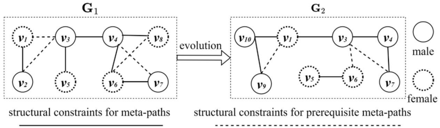


Dividing 8 students into two groups

Balance the ratio of male and female



Support dynamic updates

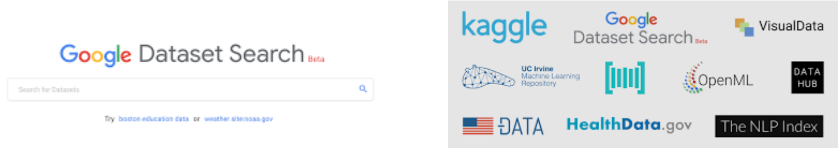


Juntao Zhang, **Sheng Wang**, Yuan Sun, Zhiyong Peng: Prerequisite-driven Fair Clustering on Heterogeneous Information Networks. **SIGMOD** 2023

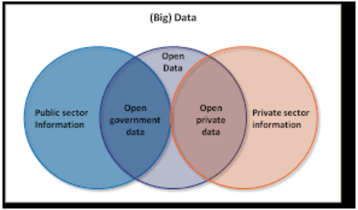
Connectivity-aware Spatial Dataset Search

What is cross-silo dataset search and preparation?

- 90% time is spent on preparing datasets

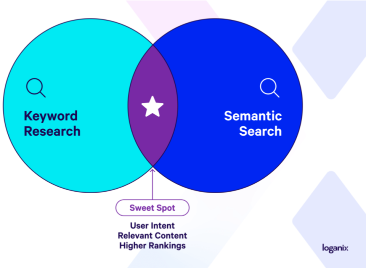


- Data can be stored in distributed systems such as:
  - University library
  - Open government
  - Private owner
  - Data markets



Existing systems for dataset search and preparation

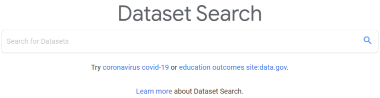
- Keyword-based search still dominates



The Dataverse® Project

Open source research data repository software

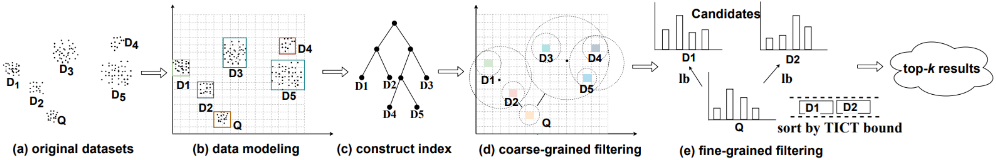
北京大学 开放研究数据平台
Peking University Open Research Data



[1] https://dataverse.org/
[2] https://opendata.pku.edu.cn/
[3] https://datasetsearch.research.google.com/
[4] https://auctus.vida-nyu.org/

Spatial dataset search by a given exemplar dataset

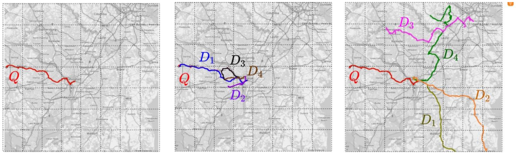
- A user has a small dataset on hand and expects to find relevant dataset to augment ：
  - Q: the input query dataset
  - E.g., which is the most similar dataset to Q among D_1, D_2, …, D_5?



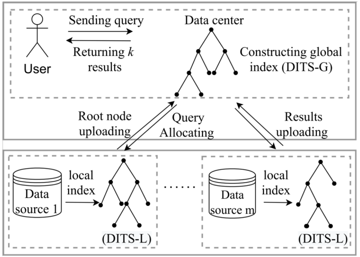
Wenzhe Yang, **Sheng Wang**, Yuan Sun, Zhiyong Peng: Fast Dataset Search with Earth Mover's Distance. Proc. VLDB Endow. 15(11): 2517-2529 (2022)

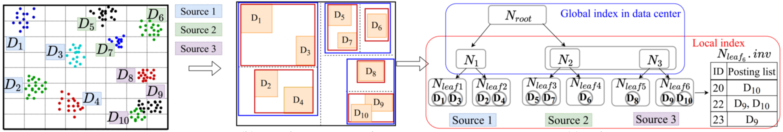
Cross-silo dataset search without leaking original datasets

- When the datasets are from multi sources



(a) Query dataset (b) Overlap joinable search (c) Coverage joinable search



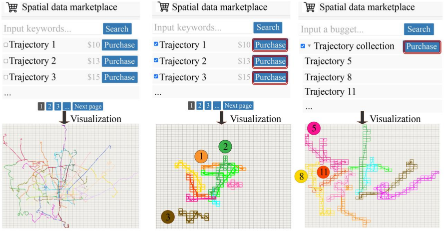


(a) Data source (b) Top-down construction (c) Index structure

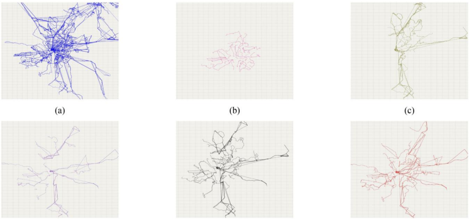
Wenzhe Yang, **Sheng Wang**, Yuan Sun, Zhiyu Chen, Zhiyong Peng: Joinable Search over Multi-source Spatial Datasets: Overlap, Coverage, and Efficiency. ICDE 2025. **(CCF A)**

Connectivity-aware for dataset preparation

- Not just similar, but should be also connected and cover more space
  - Coverage: maximize the area that the newly formed datasets cover
  - Connectivity: the searched datasets should be linked to each other



(a) Individual recommendation (b) Random acquisition (c) Collection recommendation

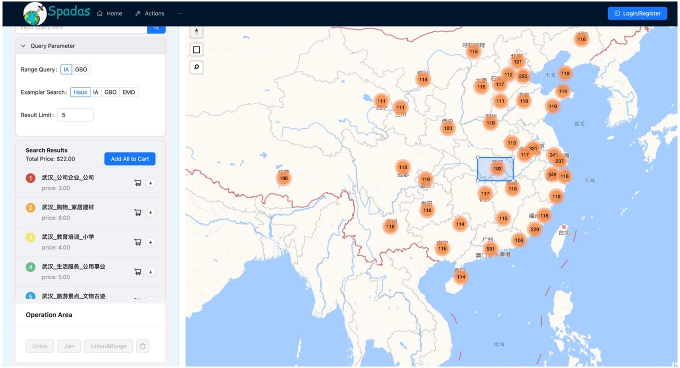

(a) (b) (c) (d) (e) (f)

Wenzhe YANG, Shixun HUANG, Sheng WANG, Zhiyong PENG. Budgeted spatial data acquisition: when coverage and connectivity matter. *Front. Comput. Sci.*, 2026, 20(3):

4. Connectivity-aware Spatial Dataset Search

Spadas: a demo of spatial dataset search and preparation

- 80% datasets have geographic information
- Key functions
  - Range query
  - Multiple similarity measures
  - Open dataset search
  - Upload private datasets for sale



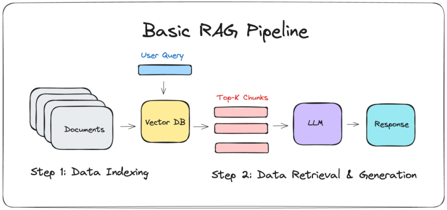
http://sheng.whu.edu.cn/spadas/

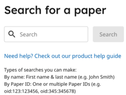
Collaboration-driven Scientific Data System

5. Collaboration-driven Scientific Data System

Cross-disciplinary scientific discovery

- Existing systems
  - DBLP, Google scholar
  - Keywords-based
- Why vector similarity search?
  - Semantic-aware
  - Retrieval Augmented Generation (RAG)
- Billion-scale search
  - Paper search
  - Talent search





Vector Similarity Search

Why does Vector Similarity Search Work?

Semantic similarity ≈ geometric closeness in the vector space.

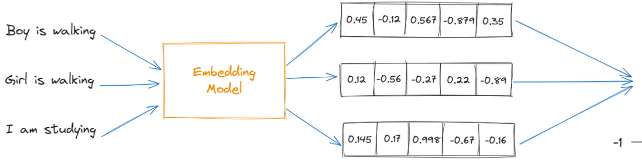
Represent texts as dense vectors.

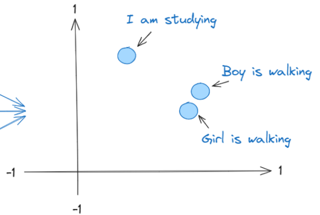
Retrieve by encoding a query, then finding nearest neighbors in a vector index.

Toy example (semantic proximity):

- distance(Boy is walking | Girl is walking) ≈ 0.02
- distance(Boy is walking | I am studying) ≈ 0.34
- distance(Girl is walking | I am studying) ≈ 0.40

The embeddings of "Boy is walking" and "Girl is walking" demonstrate higher cosine similarity due to their semantic correspondence.





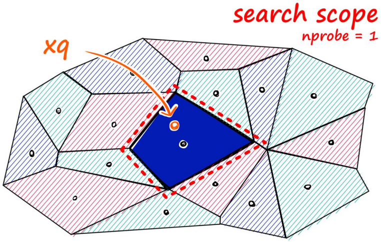
Yushuai Ji, Sheng Wang, Zhiyu Chen, Yuan Sun, Zhiyong Peng. Updatable Balanced Index for Fast On-device Search with Auto-selection Model. ICDE 2026 (CCF A)

LorIndex for efficient index routing

Why is indexing necessary for vector search?

Brute-force vector search requires computing similarity with every vector in the database, resulting in high complexity that becomes prohibitively expensive for large-scale datasets. (Especially literature data)

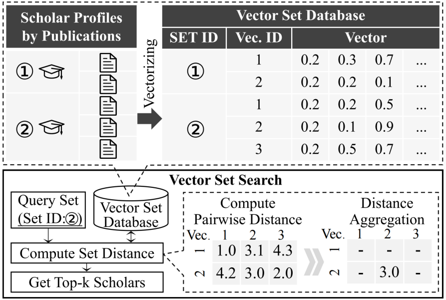
- Partition N vectors into K cells via clustering.
- At query time: compute distances to K centroids; scan only nprobe lists.
- Consequently, **only about *nprobe/K* of the corpus** must be scanned, yielding significant acceleration.



Yiqi Li, Sheng Wang, Zhiyu Chen, Zhiyong Peng. Efficient Low-Rank Index Routing for High-Dimensional Approximate Nearest Neighbor Search . Information Processing and Management. (CCF B)

Vector Set Search

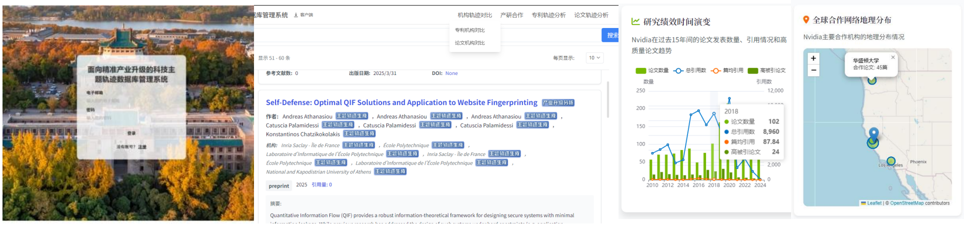
- **Problem:** Match researchers by their publication portfolios
  - Each scholar publishes multiple papers across different topics
  - Query: Find experts whose research profile matches project needs
  - Challenge: Scholars cannot be represented by a single vector
- **Why set-level matching?**
  - Researchers have **multi-faceted expertise**
  - Research **evolution over time** reflected in paper collections
  - Single-vector aggregation loses **topic diversity** and **depth signals**



Yiqi Li, **Sheng Wang**, Zhiyu Chen, Shangfeng Chen, Zhiyong Peng: Approximate Vector Set Search: A Bio-Inspired Approach for High-Dimensional Spaces. ICDE 2025. **(CCF A)**

Stotra: a demo for global scientific discovery

End to end technology-theme trajectory capabilities for intelligence centers, S&T platforms, industry institutes, and university research offices: algorithms, data governance, sharing, intelligent retrieval and recommendation, analytics, and reporting. Vector search plays an important role.



http://db4s.swangdbs.com/

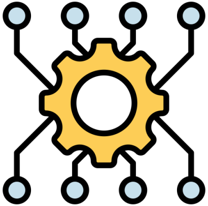
What's novel and next in MMDB for IE&IR?

Multi-model Database for IE&IR: 3Vs, 3S and 3C

- **3Vs**
  - Variety
  - Volume
  - Velocity
- **3Ss**
  - Summarization
  - Search
  - System
- **3Cs**
  - Clustering
  - Connectivity
  - Collaboration

Data

User



Algorithm

What is next for MMDB to be more applicable to IE&IR?

- **Future**
  - **The Convergence of Education and Research:** Research-Informed Teaching, Teaching-Informed Research
  - **"Integration of science & technology, education, and talent"**
- **Fairness**
  - **Algorithmic Bias:** AI systems can perpetuate and amplify existing biases in curricula and assessment if not carefully designed and audited.
  - **Digital Divide:** Equitable access to technology and fair algorithms are crucial to prevent a new form of inequality.
- **Federated**
  - **Data Privacy & Ethics:** Using student data for analytics (Learning Analytics) must be balanced with strong privacy protections and ethical guidelines.

Acknowledgement

- Grants

国家自然科学基金委员会 N S F C

中华人民共和国科学技术部 Ministry of Science and Technology of the People's Republic of China

"十四五"国家重点研发计划

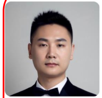
湖北省科学技术厅 Department of Science and Technology of Hubei Province

蚂蚁集团 ANT GROUP

oppo

- My Students

We have three Ph.D. candidates entering the job market next year. Inquiries are welcome!


Yiqi Li*
PhD Student


Yushuai Ji
PhD Student


ShengKun Zhu
PhD Student


Feiteng Nie
PhD Student


Pengyue Li
PhD Student


Zepeng Liu
MS Student


Zekun Tang*
MS Student


Yiming Wang
MS Student


Yue Wang
MS Student


Xinxin Huang
MS Student


Tianhao Lin
PhD Student


Hang Xue
PhD Student


Shangfeng Chen
MS Student


Yuhui Lai
MS Student


Hao Liu
MS Student


Zhongming Liao
MS Student


Xuetao Liu
MS Student


Yiting Tang
MS Student


Ziyi Wang
MS Student


Xingyu Qu
BS Student (Hongyi Class)

References

- Ji, Yushuai, Zepeng Liu, Sheng Wang, Yuan Sun, and Zhiyong Peng (2025). "On Simplifying Large-Scale Spatial Vectors: Fast, Memory-Efficient, and Cost-Predictable k-means". In: *ICDE*.
- Ji, Yushuai, Sheng Wang, Zhiyu Chen, Yuan Sun, and Zhiyong Peng (2026). "Updatable Balanced Index for Fast On-device Search with Auto-selection Model". In: *ICDE*.
- Ji, Yushuai, Shengkun Zhu, Shixun Huang, Zepeng Liu, Sheng Wang, and Zhiyong Peng (2025). "Federated and Balanced Clustering for High-dimensional Data". In: *Proceedings of the VLDB Endowment* 18.11, pp. 4032–4044.
- Li, Yiqi, Sheng Wang, Zhiyu Chen, Shangfeng Chen, and Zhiyong Peng (2025). "Approximate Vector Set Search: A Bio-Inspired Approach for High-Dimensional Spaces". In: *ICDE*.
- Li, Yiqi, Sheng Wang, Zhiyu Chen, and Zhiyong Peng (2026). "Efficient low-rank index routing for high-dimensional approximate nearest neighbor search". In: *Information Processing & Management* 63.2, p. 104459.

- Yang, Wenzhe, Sheng Wang, Zhiyu Chen, Yuan Sun, and Zhiyong Peng (2025). “Joinable search over multi-source spatial datasets: overlap, coverage, and efficiency”. In: *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE, pp. 585–598.
- Yang, Wenzhe, Sheng Wang, Yuan Sun, and Zhiyong Peng (2022). “Fast Dataset Search with Earth Mover’s Distance”. In: *PVLDB* 15.11.
- Zhang, Juntao, Sheng Wang, Yuan Sun, and Zhiyong Peng (2023). “Prerequisite-driven Fair Clustering on Heterogeneous Information Networks”. In: *Proc. ACM Manag. Data* 1.1, pp. 1–27.
- Zhu, Shengkun, Feiteng Nie, Jinshan Zeng, Sheng Wang, Yuan Sun, Yuan Yao, Shangfeng Chen, Quanqing Xu, and Chuanhui Yang (2025). “FedAPM: Federated Learning via ADMM with Partial Model Personalization”. In: *KDD*.
- Zhu, Shengkun, Jinshan Zeng, Yuan Sun, Sheng Wang, Xiaodong Li, and Zhiyong Peng (2026). “Efficient k-means with Individual Fairness via Exponential Tilting”. In: *PVLDB*.
- Zhu, Shengkun, Jinshan Zeng, Sheng Wang, Quanqing Xu, Yuan Sun, Zhifeng Yang, Chuanhui Yang, and Zhiyong Peng (2023). “F3KM: Federated, Fair, and Fast k-means”. In: *Proc. ACM Manag. Data* 1.4, pp. 1–25.

Q & A

swangcs@whu.edu.cn

Recent Selected Papers

(Ji, Liu, et al., 2025; Ji, Wang, et al., 2026; Ji, Zhu, et al., 2025; Li, Wang, Z. Chen, S. Chen, et al., 2025; Li, Wang, Z. Chen, and Peng, 2026; Yang, Wang, Z. Chen, et al., 2025; Yang, Wang, Sun, et al., 2022; Zhang et al., 2023; Zhu, Nie, et al., 2025; Zhu, Zeng, Sun, et al., 2026; Zhu, Zeng, Wang, et al., 2023)