University of Birmingham Research Intelligence Platform Automated Research Discovery & Analysis System

Safi Shamsi

Institute for Data Science and AI

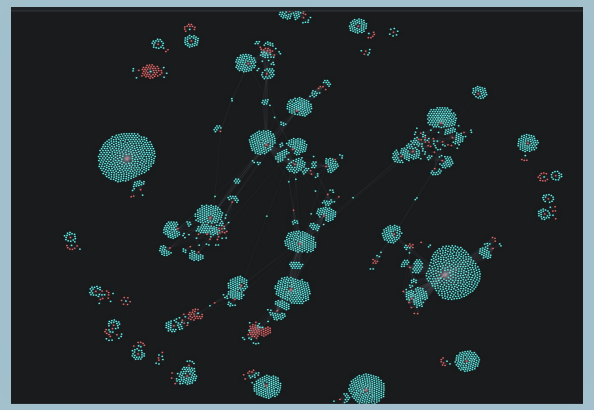
University of Birmingham

Developed in collaboration with the Institute for Data Science and AI

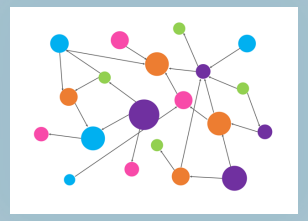
1. Background & Motivation

The Research Collaboration Challenge

- **Problem**: Researchers lack visibility beyond immediate disciplines, limiting cross-disciplinary breakthroughs
- **Scale**: University-wide challenge across all academic departments and career stages
- **Opportunity**: AI + Knowledge Graphs can bridge research recommendations and surface new academic networks.



Project Vision: AI-Powered Research Discovery	**Goal**: Create a knowledge-based system that:	Extracts "who does what" across the entire University	Suggests new academic networks with collaboration potential	Enables GenAI models to provide intelligent research recommendations	Covers complete data-to-knowledge trajectory
Why Knowledge Graphs vs. Traditional Databases?	**Traditional Relational Database Limitations:**	**Rigid Schema**: Fixed table structures can't adapt to evolving research relationships	**Complex Joins**: Multi-hop queries (author → paper → co-author → institution) require expensive joins	**Relationship Focus**: Academic networks are relationship-heavy, not transaction-heavy	**Scalability**: Join operations become prohibitively slow with research network complexity

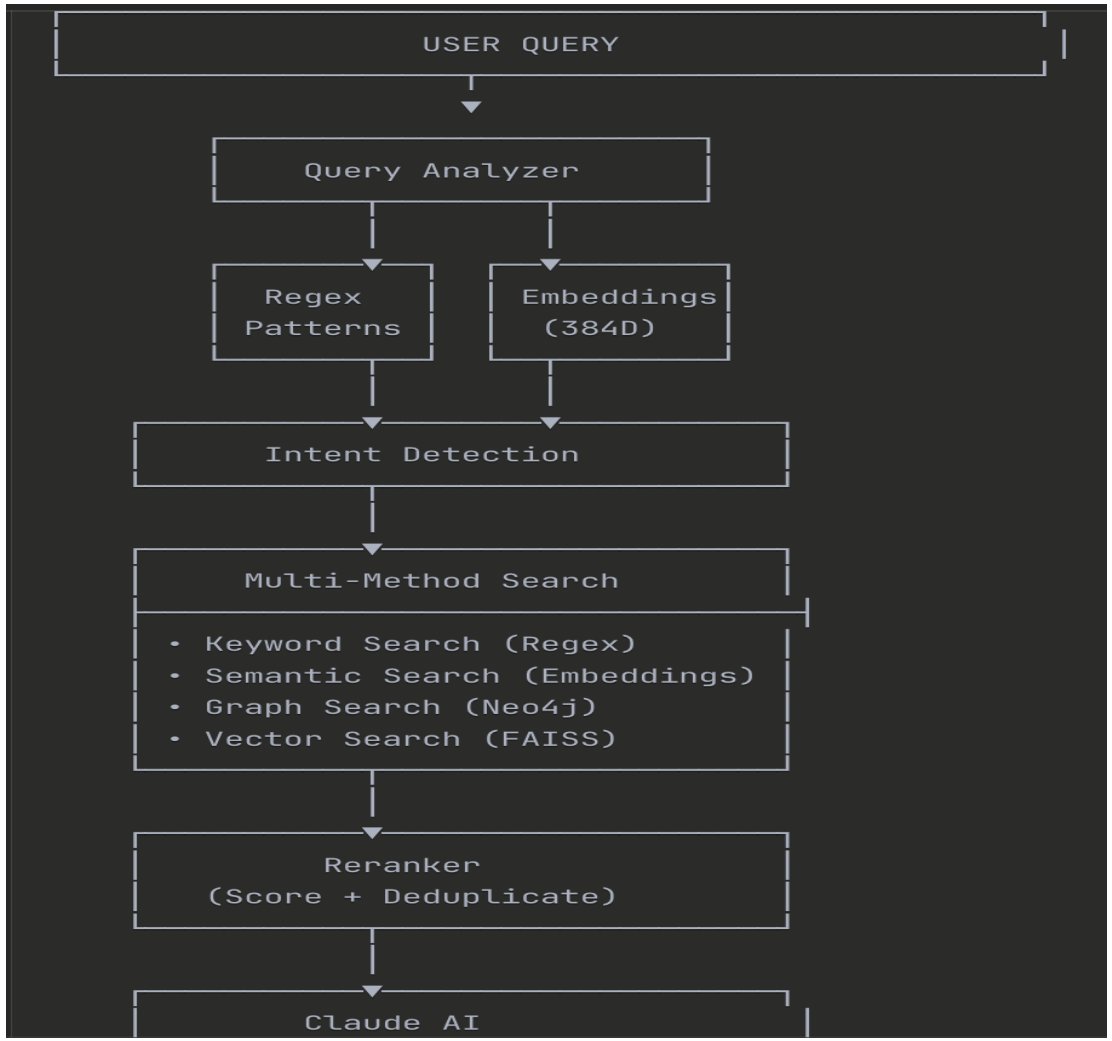


2. Methodology & System Architecture

Complete Data-to-Knowledge Pipeline

Objective:

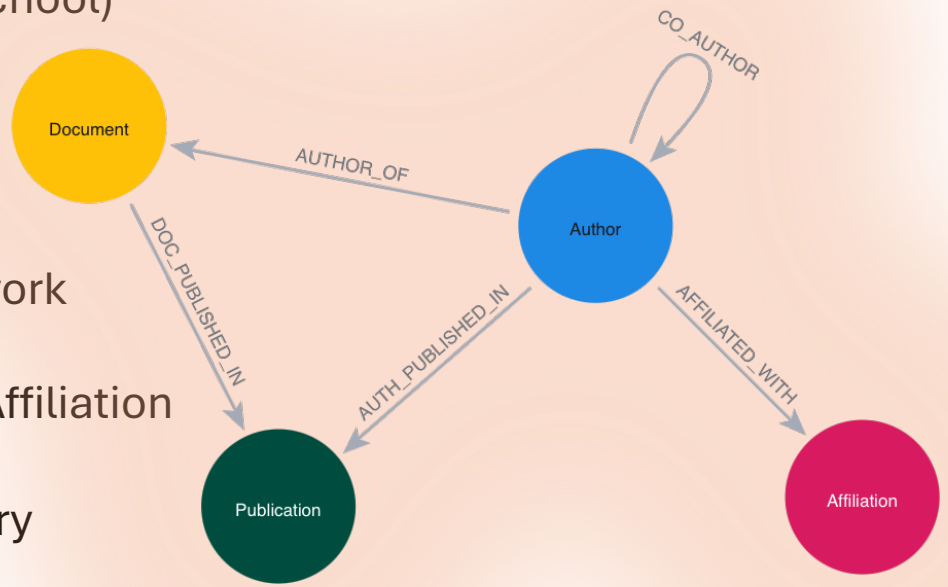
- Build a **Local RAG-based academic recommender system**.
- Enhance it with a **knowledge graph from Scopus-scraped data**.
- Use it to **identify potential research collaborators** at the University of Birmingham.
- Integrate with a **Large Language Model (LLM)** for improved effectiveness and retrieval.



Technical Architecture Decisions

Component 1: Intelligent Data Extraction

- Scopus API integration with comprehensive field extraction
- Multi-institutional support (Main, Dubai, Business School, Medical School)
- Robust author metadata capture with fallback methods
- Rate-limited batch processing for production scalability



Component 2: Knowledge Graph Construction

- Neo4j graph database enabling native relationship traversal and network analysis
- Clean schema/ontology consisting: Document, Author, Publication, Affiliation
- Smart deduplication with conflict resolution and alternative names
- Graph-optimized queries supporting multi-hop collaboration discovery
- Birmingham-focused filtering with institutional validation

Component 3: AI-Powered Query System

- LangChain + LangGraph orchestrated RAG system with conversational memory
- Intelligent name conversion with database format auto-detection
- Graph pathfinding using Louvain algorithim for collaboration discovery
- Context-aware follow-up question handling with workflow orchestration

Why Neo4j Graph Database Over Traditional SQL and Vector Databases?

SQL Database Problems:

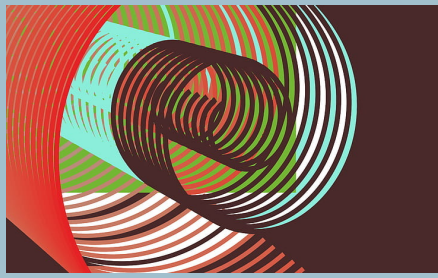
- **Complex Joins**: Finding "researchers 2 degrees from Paolo Missier" requires expensive multi-table joins
- **Rigid Schema**: Can't easily add new relationship types (mentorship, funding collaborations)
- **Poor Performance**: $O(n^3)$ complexity for network queries vs. $O(degree^2)$ in graphs

Vector Database Limitations:

- **Hallucination Risk**: Embeddings can generate plausible but false research connections
- **Black Box**: Can't explain WHY two researchers are connected
- **No Relationship Types**: Treats all connections as similar, missing co-authorship vs. citation distinctions
- **Verification Difficulty**: Hard to validate if suggested collaborations actually exist

Graph Database Advantages:

- **Explainable Results**: Every connection has a clear path (eg: Author → Paper → Co-author or domain)
- **No Hallucination**: Only returns real relationships that exist in the data
- **Relationship Types**: Distinguishes between co-authorship, citations, institutional affiliation



System Performance & Validation

MY RESEARCH ASSISTANT SYSTEM

Data Processing Metrics:

- **Documents Processed**: 421K+ academic papers with full metadata
- **Author Networks**: Complete collaboration graphs with relationship weights
- **Query Response Time**: <8 seconds for complex multi-author searches
- **Memory Efficiency**: Optimized batch processing preventing OOM errors

Based on the intelligent search results, I can identify Paolo Missier's most cited paper and its co-authors:

🎯 Most Cited Paper Details:
Title: "Researching multimorbidity in hospital: can we deliver on the promise of health informatics?"
Citations: 12
Year: 2023

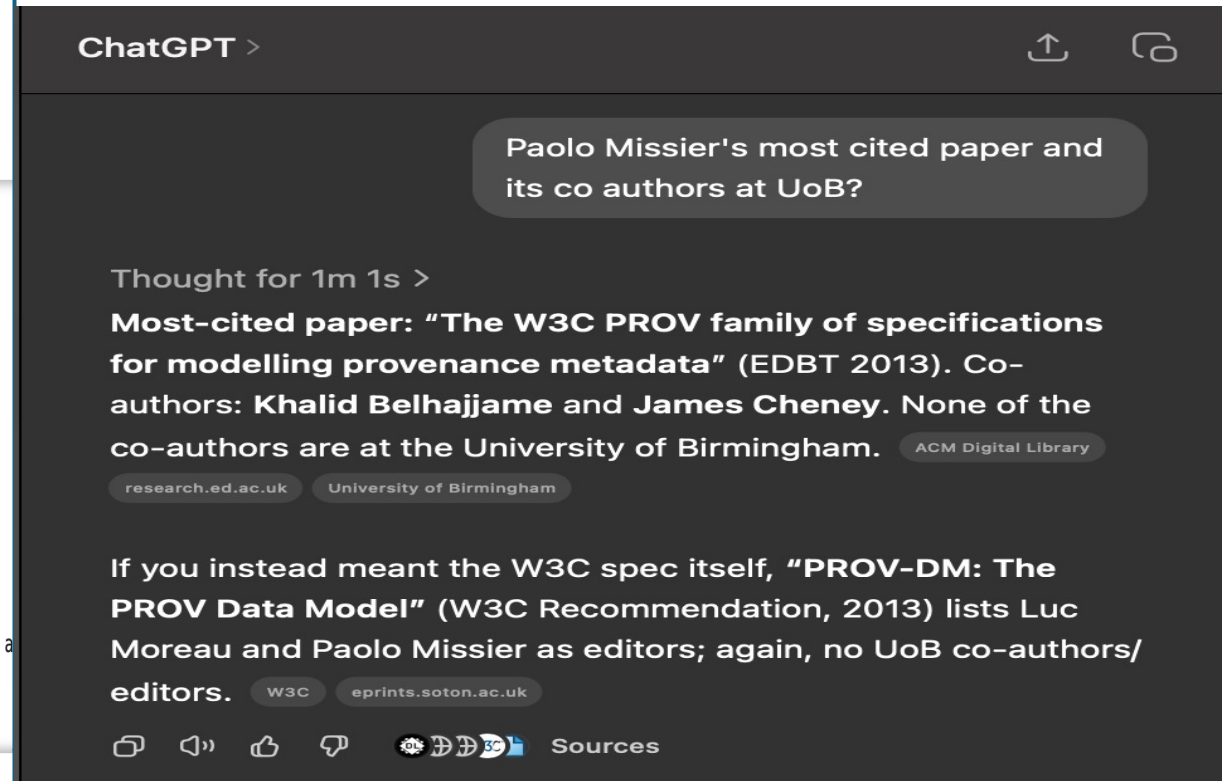
👥 The 5 Co-authors are:
1. Sapey E.
2. Witham M.
3. Cooper R.
4. Robinson S.M.
5. Sayer A.A.

This paper has the highest citation count (12) among the papers found in the database for Paolo Missier. The collaboration is classified as a "Medium Team" with a total of 6 authors including Missier himself.

Would you like to know more about any of these co-authors or their other collaborations with Paolo Missier?

A generic response by a LLM (OpenAI)

- **Query Response Time**: >60 seconds for complex multi-author searches
- **Efficiency:** The response is totally not what we needed.



System Performance & Validation

REGEX SYSTEM

1.Regex only matches exact patterns, hence can miss on certain data during retrieval from the database.
2.Regex fails with synonyms or paraphrases, but hybrid search understands variations in human language while still allowing rule-based filtering.

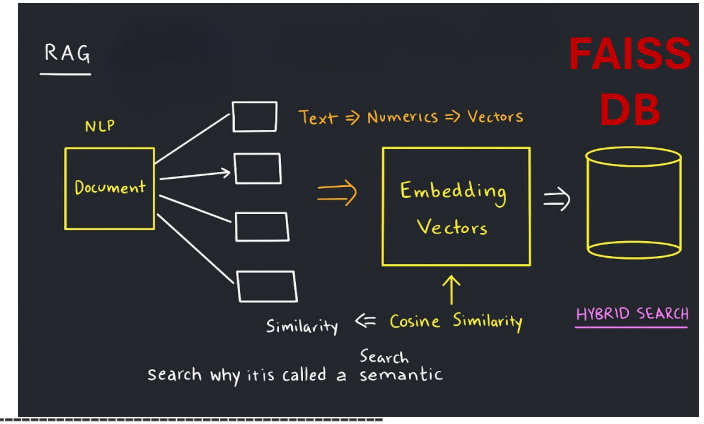
Database Search Results
Based on the database search, 15 prolific ML researchers at the University of Birmingham were identified with a total of 31 papers and 501 citations. Here are the key researchers and their work:

1. **COVID-19 ML Research Group** (Yang J., Soltan M., Zhu T., Clifton D.A., Thickett D.R., Eyre D.W., Soltan A.A.S.)
- Each has 2 papers with 54 total citations
- Focus on federated learning and AI-driven COVID-19 screening
- Key papers:
  - "A scalable federated learning solution for secondary care..."
  - "Real-world evaluation of rapid and laboratory-free COVID-19 triage..."

2. **Railway ML Applications** (Kaewunruen S., Huang J.)
- Kaewunruen: 3 papers, 29 citations
- Research on automated diagnostics and passenger comfort
- Key papers:
  - "Automated Prognostics and Diagnostics of Railway Tram Noises"

HYBRID SYSTEM REGEX+SEMANTIC

1.hybrid search (vector + keyword) captures meaning and context, retrieving results even if exact words differ.
2.Higher accuracy in retrieval



Based on the provided evidence, there are several prominent ML researchers affiliated with institutions in Birmingham:

Key researchers:
1. Yao X. - Affiliated with both Birmingham City University and University of Birmingham, with 63 published papers

2. Tino P. - Has multiple affiliations across Birmingham institutions including:
- University of Birmingham
- Birmingham Children's Hospital
- Queen Elizabeth Hospital Birmingham
- Birmingham Health Partners
With 49 published papers

3. Kaewunruen S. - Affiliated with University of Birmingham, with 38 published papers

4. Gkoutos G.V. - Has extensive affiliations across Birmingham institutions including:
- University of Birmingham (including College of Medical and Dental Sciences)

4. Conclusion & Future Work

Current Technical Achievements

Complete Data-to-Knowledge Pipeline:

•Robust multi-source data extraction with intelligent field mapping
•Production-ready knowledge graph with refined ontology and embeddings and community clustering.
•LangChain/LangGraph orchestrated AI system with conversational capabilities
•Graph-based collaboration discovery using network algorithms like Louvain with workflow management

Birmingham-Specific Optimization:

•Multi-institutional data integration (Main, Dubai, Business School, Medical)
•Institution-focused filtering with precision validation
•Local research network analysis with collaboration strength measurement
•Scalable architecture supporting university-wide deployment

Limitations:

- **Embedding Model Dependency**: Fixed to MiniLM-L6-v2 (384D) model; changing models requires complete re-indexing of all vectors.
- **Query Processing Time**: 2-3 second latency may be insufficient for real-time applications requiring sub-second responses
- **Language Restriction**: System only processes English-language papers; multilingual research is excluded
- **Citation Bias**: Ranking algorithm favors highly-cited older papers over potentially innovative recent research
- **Memory Requirements**: Loading 248,078 author embeddings requires significant RAM; may not run on resource-constrained systems

Future Work:

- **Cross-lingual Search**: No support for cross-language information retrieval.
- **Dynamic Ontology**: Knowledge graph schema is fixed; cannot adapt to new relationship types
- **Incremental Indexing**: System cannot add new papers without complete re-indexing for real world deployment
- **Personalization**: No user preference learning or search history utilization