# Lab Recommendation Service for Prospective Graduate Students

Team 16
Seung-eon Lee, Je-hyun Park

PROBLEM / WHY IT MATTERS

# Why choosing the right lab matters

## Three key factors in lab selection

**01**

### Research Fit

Does the lab match my interests?
Research topics and publications

**02**

### Research Performance

Does the lab have a strong research record?
Papers and major conferences

**03**

### Career Opportunities

Does it support my career goals?
Alumni destinations and outcomes

**The right choice depends on each student's individual fit.**

PROBLEM / CURRENT APPROACHES

# Limitations of existing approaches

## Personal Referrals

Experience-based advice

Recommendations depend heavily on personal academic networks.

## Search Engine

Finding relevant pages

Hard to compare publications and career outcomes across labs

## Academic Database

Publication metadata

Limited insight into lab size and alumni career paths

## General LLM

Natural-language answers

Evidence, freshness, and completeness are not guaranteed

SOLUTION / SERVICE UI

# Service Process

## 01 USER INPUT

### Research interests and career goals

**Research interests**

**Multimodal AI and LLM-based Recommender Systems**

**Career goal**

Industry R&D · AI Researcher / AI Engineer · Academic Professor

**Additional priorities**

Industry projects · Alumni employment

## 02 RECOMMENDATION

### Top-N Professor and Lab Matches

Research fit · Career · Industry ties

| | | |
|---|---|---|
| 1 | **Prof. Kim (example)**<br>Multimodal Intelligence Lab | 95 /100 |
| 2 | **Prof. Lee (example)**<br>Interactive AI Lab | 89 /100 |
| 3 | **Prof. Park (example)**<br>Machine Learning Theory Lab | 76 /100 |

## 03 DETAILED REPORT

### Lab Intelligence Report

- **PUBLICATION** — Yearly papers and topic trends
- **CITATION** — Citations · h-index · growth
- **PROFESSOR CAREER** — Education and career timeline
- **ACADEMIC INFLUENCE** — Conference roles and coauthors
- **RESEARCH PROJECTS** — Public / industry projects
- **ALUMNI** — Academic and industry careers

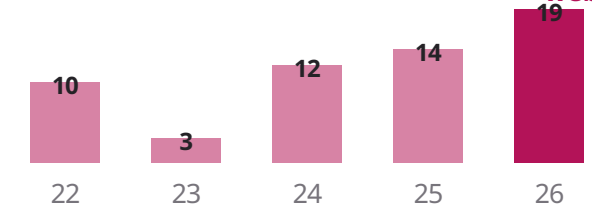
SOLUTION / SERVICE UI

# Lab Intelligence Report

1 **Prof. Kim (example)**
Multimodal Intelligence Lab
95 /100

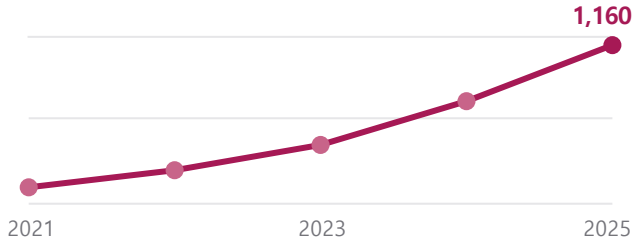
## 01 Publication

Listed on lab website




## 02 Citation

1,160 citations



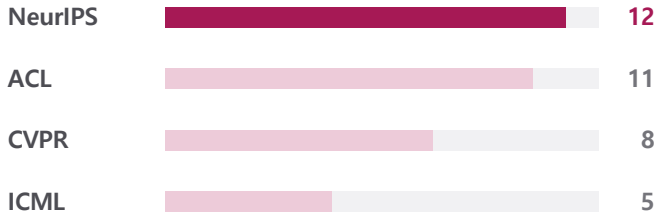

## 03 Professor Career

Career history



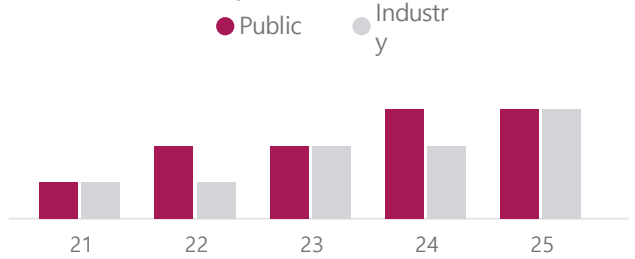

## 04 Academic Influence

Major venues




## 05 Research Projects

2025 · 6 projects




## 06 Alumni Outcomes

Verified 19 / 24

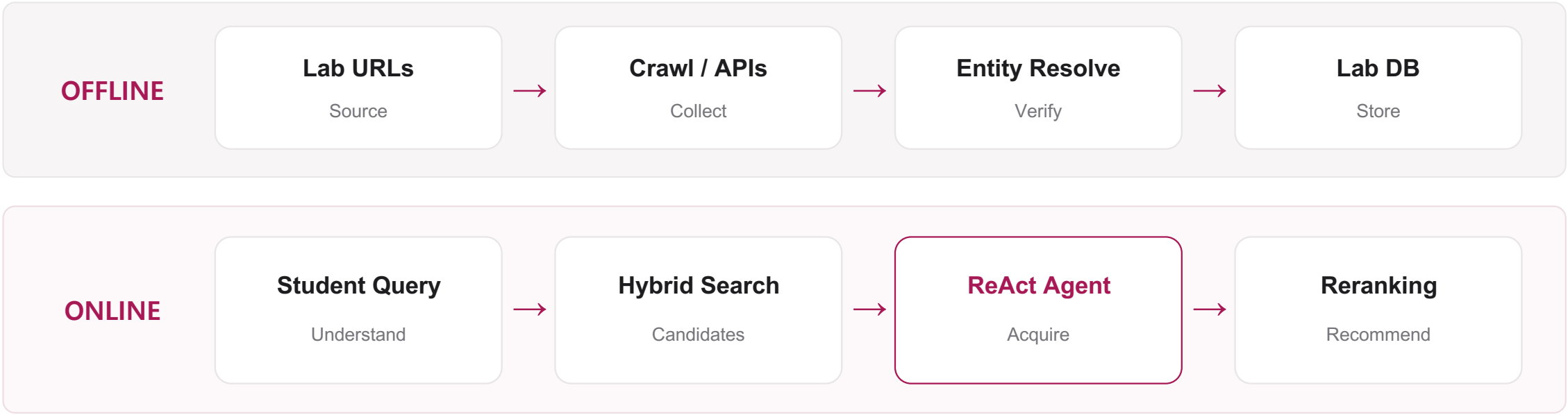



5 unknown; distribution among verified alumni only

SOLUTION / HIGH-LEVEL ARCHITECTURE

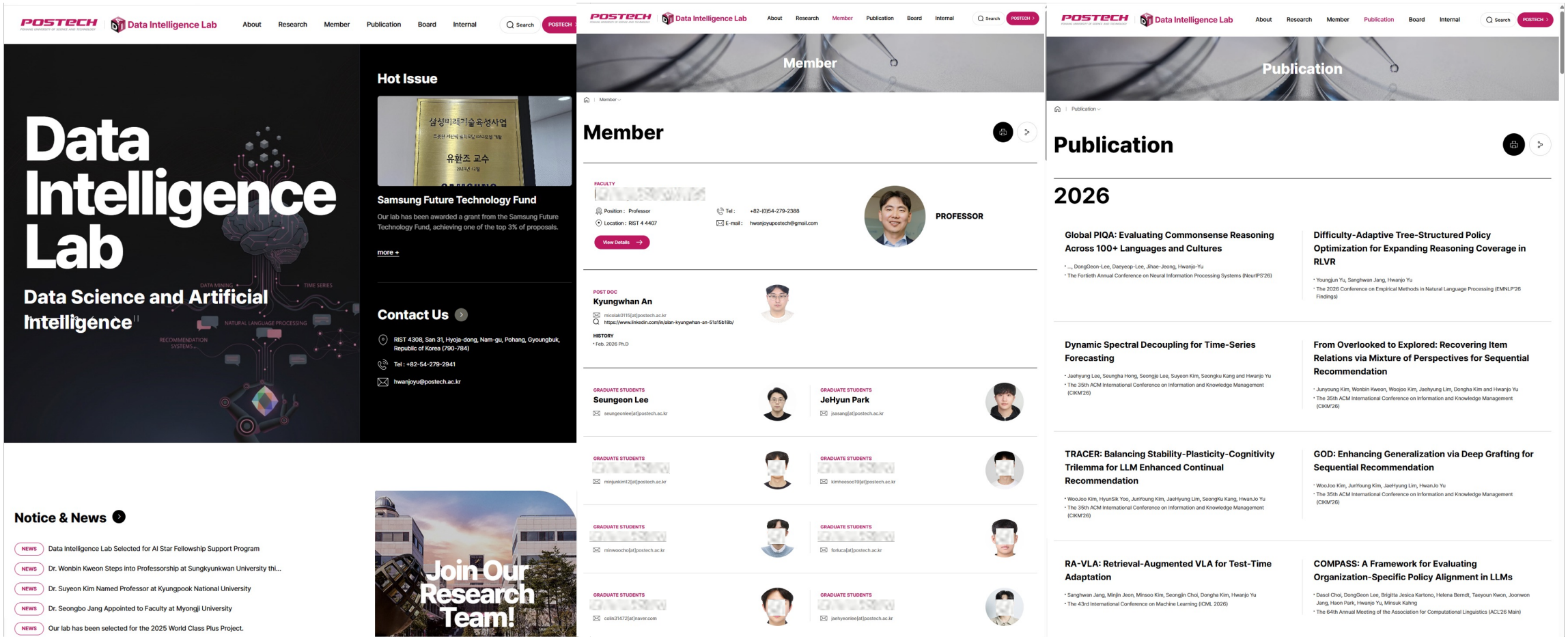
# An integrated database supports personalized recommendations

Build and maintain the core DB offline; retrieve missing evidence on demand.



DATA / SCHEMA & QUALITY

# One lab-centric schema with traceable evidence



DATA / SCHEMA & QUALITY

# One lab-centric schema with traceable evidence



DATA PIPELINE / 01 PAGE DISCOVERY

# Discover pages from one lab URL

Follow in-domain links, identify relevant sections, and build a page inventory before extraction.

**SEED URL** https://di.postech.ac.kr/di/index.do

DISCOVER • CLASSIFY • QUEUE

| MEMBER | PUBLICATION | RESEARCH | PROJECT |
|---|---|---|---|
| /di/mem/member.do | /di/pub/publication.do | /di/res/research_area.do | /di/res/project.do |
| Faculty · Students · Alumni | Title · Authors · Venue | Research topics | Projects · Dates · Funding |

**Next: visit each page → find repeated content blocks → emit one JSON record per item**

Illustrative discovery workflow based on the official DI navigation · Preserve original URL and crawl timestamp.

DATA PIPELINE / 02 STRUCTURED EXTRACTION

# Extract typed JSON from each page

Parse repeated member and publication entries into JSONL, preserving their original wording and source.

**MEMBER PAGE**

/di/mem/member.do


```
{
  "type": "member",
  "name": "Seungeon Lee",
  "role_raw": "GRADUATE STUDENTS",
  "email_raw": "seungeonlee[at]postech.ac.kr",
  "source": "/di/mem/member.do"
}
```


**PUBLICATION PAGE**

/di/pub/publication.do


```
{
  "type": "paper",
  "title": "RA-VLA: Retrieval-Augmented VLA for Test-Time
Adaptation",
  "year_raw": "2026",
  "venue_raw": "The 43rd International Conference on Machine
Learning (ICML 2026)",
  "source": "/di/pub/publication.do"
}
```


**Page HTML → repeated content blocks → JSON Schema validation → JSONL records (one row per entity)**

Actual DI page entries; JSON structure is an extraction example. Keep the original text, full source URL, and fetched_at.

DATA PIPELINE / 03 PREPROCESS & LOAD

# Normalize raw data before database storage

Standardize fields and URLs while preserving original source values for traceability.

## 01 EMAIL NORMALIZATION

| RAW | | CLEAN |
|---|---|---|
| seungeonlee[at]postech.ac.kr | → | **seungeonlee@postech.ac.kr** |

Convert [at] and [dot] only in email fields.

## 02 CONFERENCE NORMALIZATION

| RAW | | CLEAN |
|---|---|---|
| The 43rd International Conference on Machine Learning (ICML 2026) | → | **ICML · 2026** |

Keep raw venue and conference edition (43rd).

## 03 MEMBER ROLE NORMALIZATION

| RAW | | CLEAN |
|---|---|---|
| GRADUATE STUDENTS | → | **graduate_student** |

Map known role labels; flag unknown values.

## 04 URL NORMALIZATION

| RAW | | CLEAN |
|---|---|---|
| /di/mem/member.do | → | **https://di.postech.ac.kr/di/mem/member.do** |

Resolve relative URLs; strip fragments; deduplicate.

**CLEAN DB** **MEMBER(name, email, role, source_url) · PAPER(title, year, venue_code, source_url)**

Checks: DOI/title deduplication · year/venue consistency · missing → NULL · original value + source provenance

DATA PIPELINE / 04 CLEAN DB RECORDS

# Store multiple normalized records in relational tables

After preprocessing, each extracted item becomes one row linked to the DI lab and its source page.

## MEMBER · 4 SAMPLE ROWS

| ID | name | email | role |
|---|---|---|---|
| M001 | Seungeon Lee | seungeonlee@postech.ac.kr | graduate_student |
| M002 | JeHyun Park | jsasang@postech.ac.kr | graduate_student |
| M003 | Minjun Kim | minjunkim12@postech.ac.kr | graduate_student |
| M004 | Heesoo Kim | kimheesoo19@postech.ac.kr | graduate_student |

source: /di/mem/member.do

## PAPER · 3 SAMPLE ROWS

| ID | title | venue | year |
|---|---|---|---|
| P001 | RA-VLA: Retrieval-Augmented VLA for Test-Time Adaptation | ICML | 2026 |
| P002 | Global PIQA: Evaluating Commonsense Reasoning Across 100+ Languages and Cultures | NeurIPS | 2026 |
| P003 | Dynamic Spectral Decoupling for Time-Series Forecasting | CIKM | 2026 |

source: /di/pub/publication.do

Shared lab_id = DI · Each row retains source_url & fetched_at · IDs are illustrative; entries are not exhaustive

ROADMAP / MILESTONES

# Service development roadmap

Seven-week plan | From lab web data to SQL ETL, LLM analysis, survey UI, and integration tests

| Period | Phase | Main activities |
|---|---|---|
| **Week 1** | Scope and requirements | Confirm lab URLs; define target fields, crawl boundaries, and collection policy |
| **Weeks 2–3** | Raw data collection | Discover subpages; extract members, publications, and research; write JSONL with source metadata |
| **Week 4** | SQL schema and ETL | Design LAB/PAPER/ALUMNI/TOPIC tables; normalize IDs, handle missing values, and deduplicate |
| **Week 5** | LLM paper analysis | Extract and review topics/keywords; generate yearly publication and topic summaries |
| **Week 6** | Service integration | Connect LAB and user inputs; implement Hybrid Search and a prototype Evidence Gate |
| **Week 7** | Quality review and demo | Audit data coverage and provenance; test the end-to-end pipeline; prepare demo |

Timeline may be adjusted based on page coverage, extraction quality, and integration results.

# Thank You for Listening