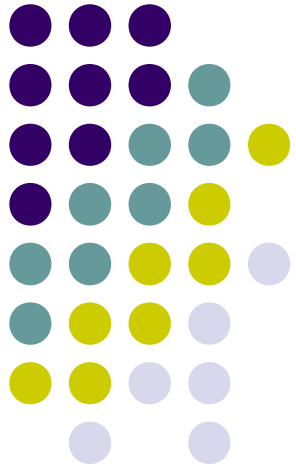
Intranet Mediator
IPRO 356

Information Retrieval Laboratory
Department of Computer Science
http://mediator.iit.edu



ILLINOIS INSTITUTE OF TECHNOLOGY

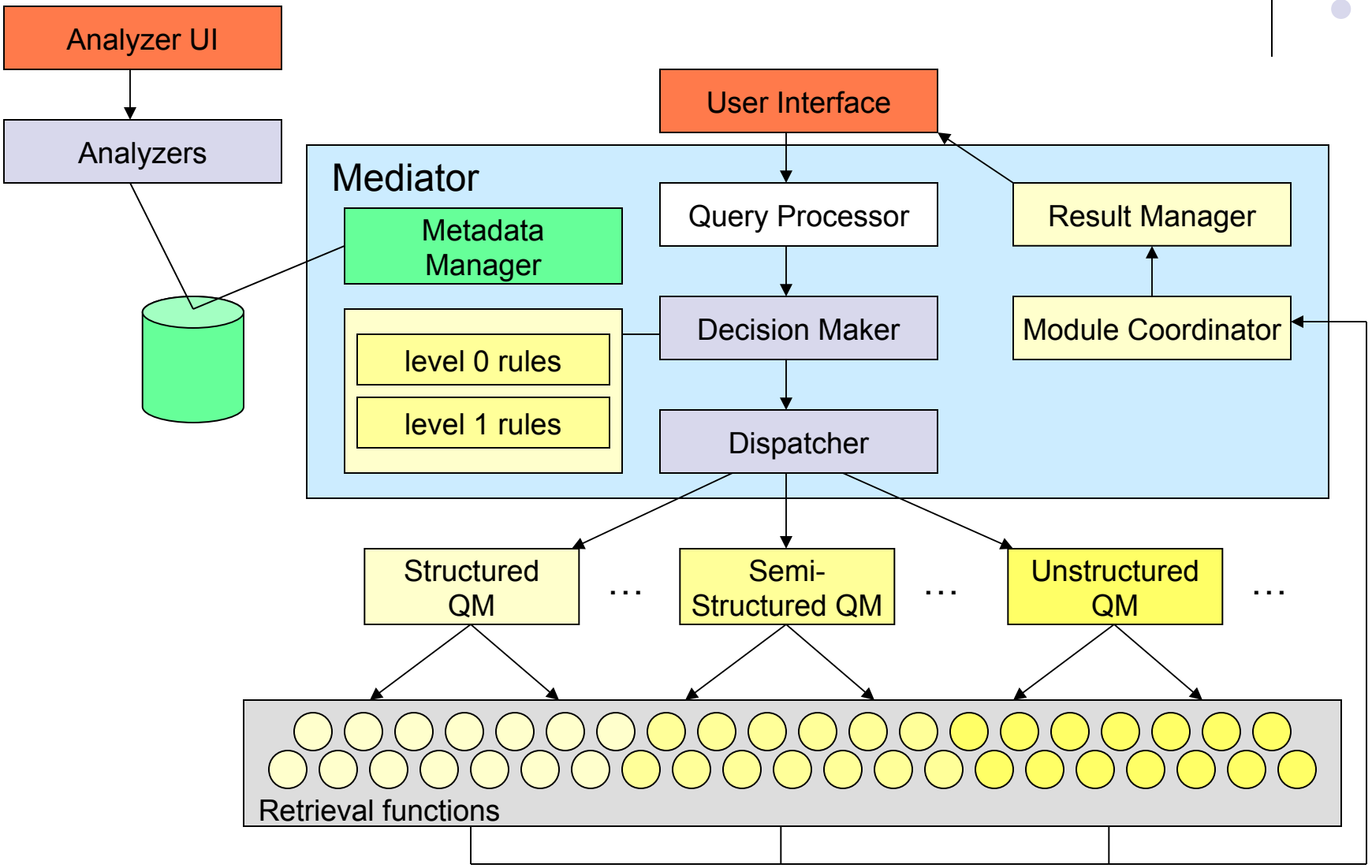
Team

- Faculty Advisor:
  - Dr. David Grossman
- Team Leader:
  - Michael Saelee
- Graduate Advisors:
  - Steve Beitzel
  - Eric Jensen
  - Angelo Pilotto
- Members:
  - Syed Aqeel
  - Robert Guico
  - Joe Prokop
  - Davin Tanabe
  - Dawn Yap
  - Michael Zatopek

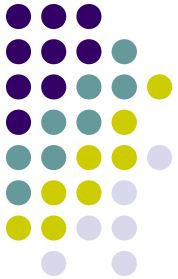
Introduction

- 3 basic high-level data types
  - Structured: name, address, phone number
  - Semi-structured: XML
  - Unstructured: documents, e-mail messages
- Need a “mediator” to accept a natural language query, access all three high-level types of data, and combine results.

Architecture

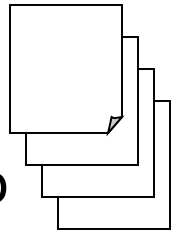


Data Source Acquisition



- Started with 157 questions
- Mapped questions to available sources
- Acquired Sources
  - HUB Events
  - External Sources (weather.com, Mapquest, CTA)
  - Student Directory (McCormick Student Village)
  - Schedule of Courses (Registrar's office)
  - IIT Website, faculty directory (CNS)
- Currently able to answer about 100 questions with about 300 variations.

Source Types



- Unstructured
  - Registration Info
  - TechNews Articles
  - IIT Websites
    - Course Websites
    - Faculty Websites
- Structured
  - Student Directory
  - HUB Events Information
  - Master Calendar
  - Faculty Directory

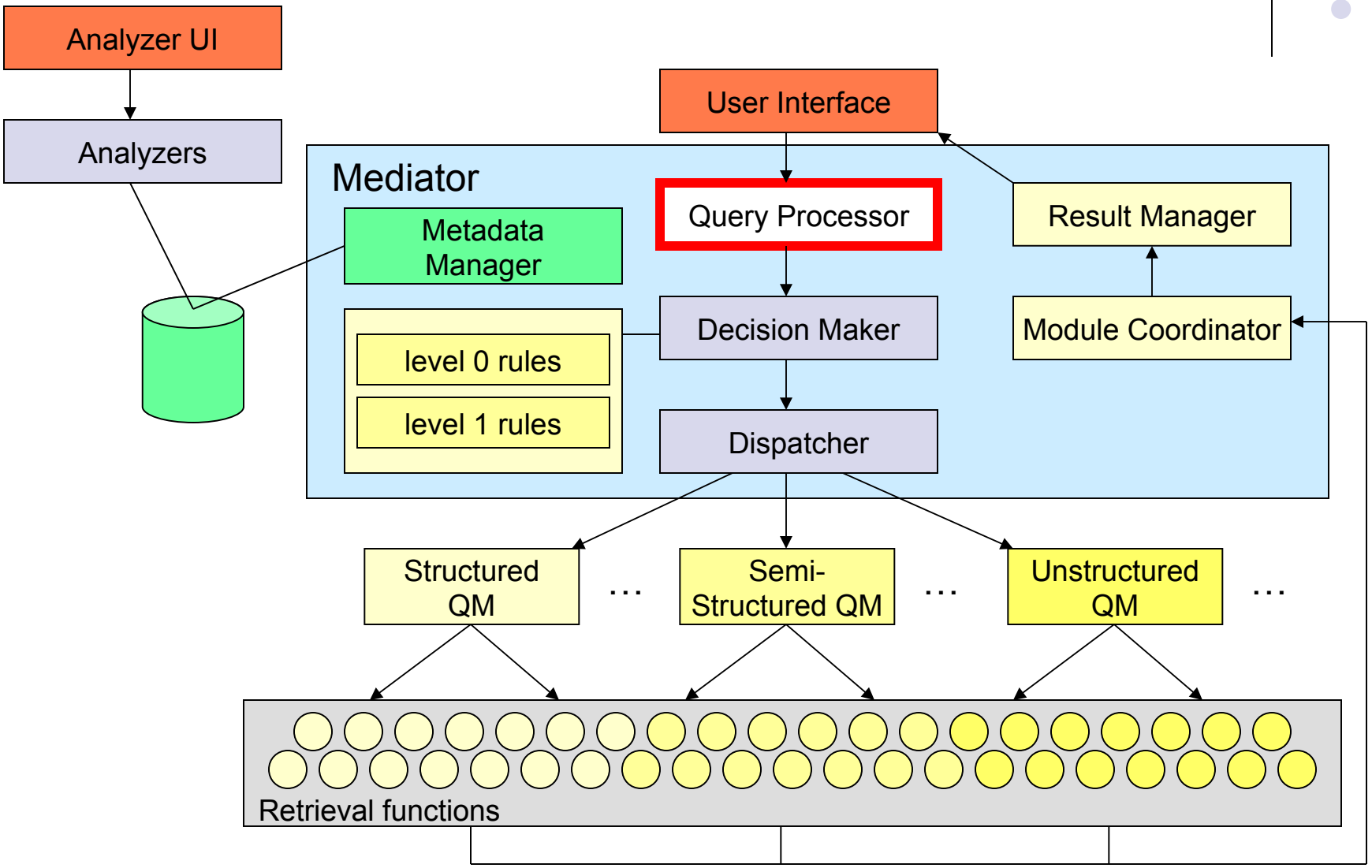
- Semi-Structured
  - IIT Parking Lot Info

```
<?xml version="1.0"
encoding="us-ascii"?>
```

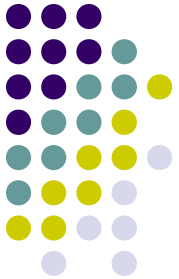
Building	Address	URL
Stuart Building	10 W. 31st St.	Null
Sigma Phi Epsilon	3341 S. Wabash	http://www.sigeps.org

  - IIT Building Information
  - Department Web Site Locations
  - CNS Lab Schedule

Architecture



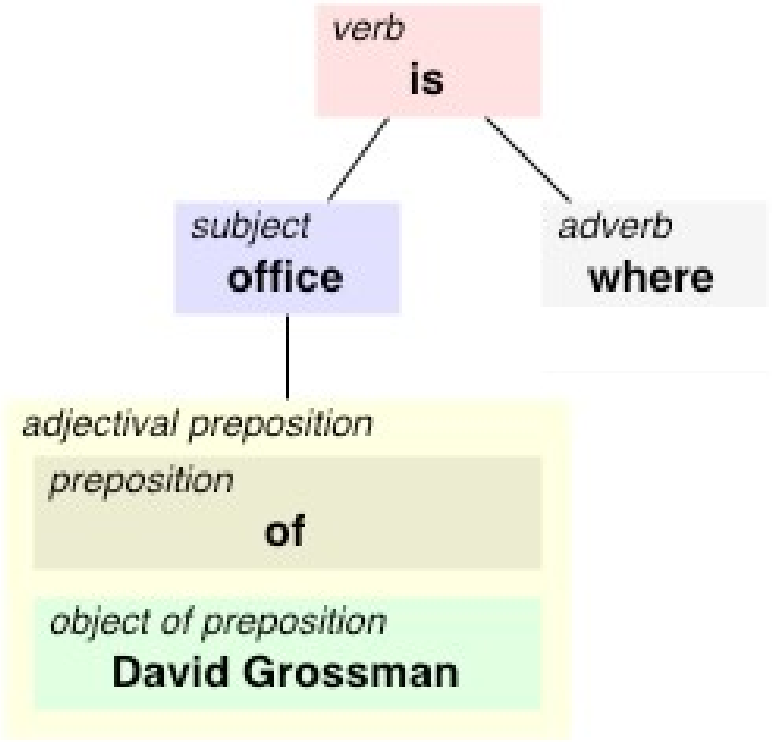
Query Processing: Part-of-speech Tagging



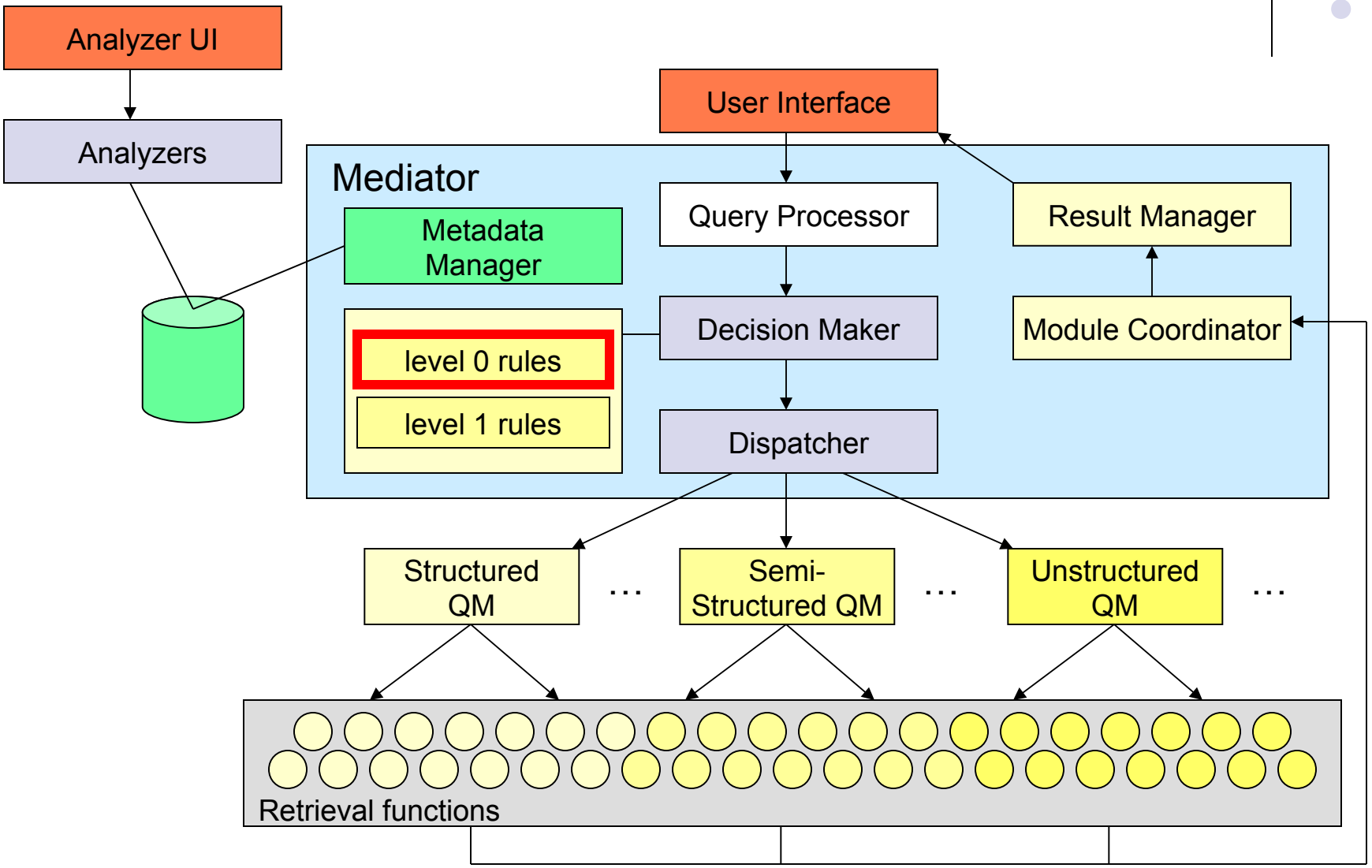
- Brill tagger (University of Pennsylvania)
  - Stochastic tagging
  - Rule-based error correction
- Sample tagged output
  - Where/Adverb is/Verb the/Determiner office/Noun of/Preposition (David Grossman)/Noun Phrase ?/Punctuation

Query Processing: Grammar Parsing

- Link parser (Carnegie-Melon University)
- Part-of-grammar identification
  - Subject
  - Verb
  - Objects

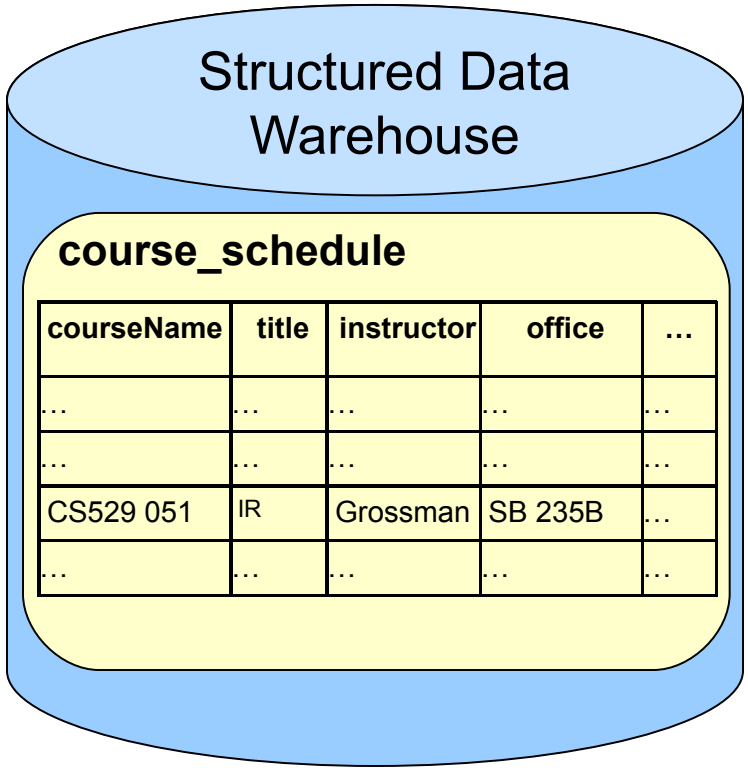


Architecture

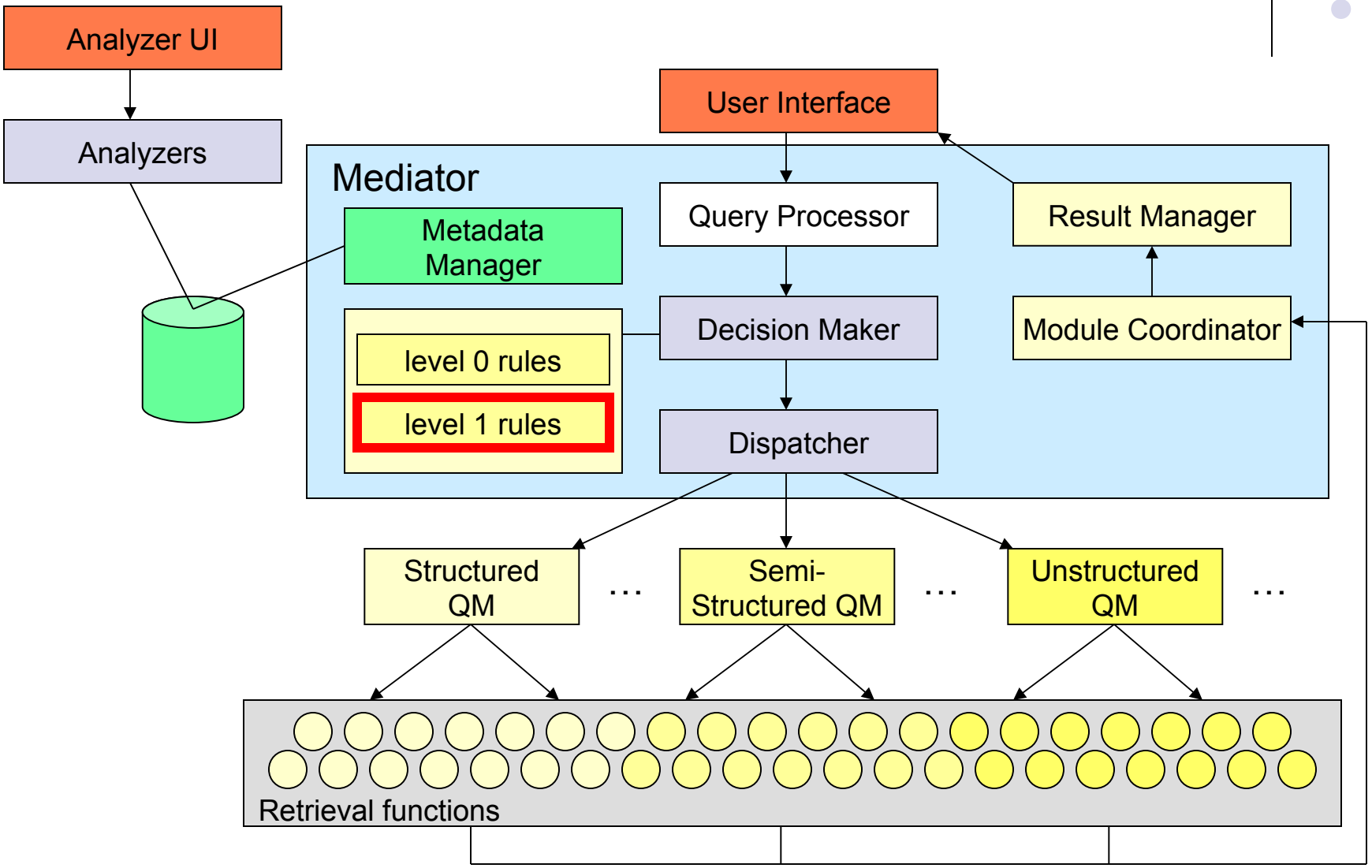


Level 0 Rules: Higher level semantic concepts

- Fueled by Metadata
- Define type of data
  - *Stuart Building* is a **place**
  - *David Grossman* is a **person**
- Allow customization to specific sources
  - *CS529* is a **course**
  - *David Grossman* is a **teacher**



Architecture



Level 1 Rules: Part-of-Speech

- Sentence matching
  - office of (Noun Phrase) :: staff_office($1)
  - Matches the partial sentence “office/Noun of/Preposition (David Grossman)/Noun Phrase”

Weaknesses of Part-of-Speech

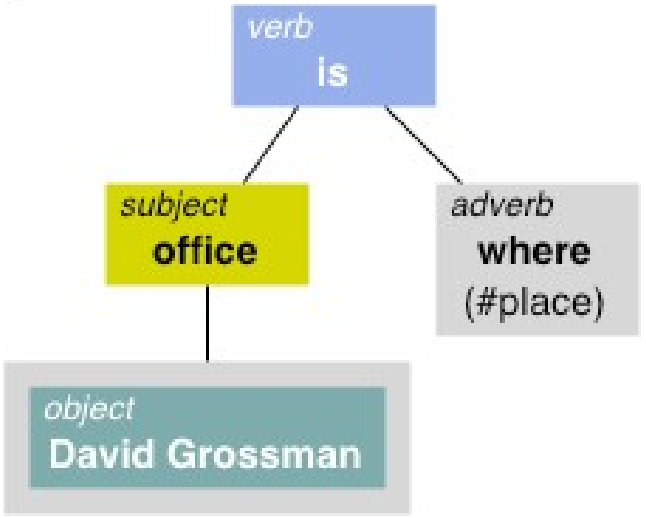
- Sentences with same or similar meanings but different word order
  - Where is the office of David Grossman?
  - Where is David Grossman's office?
  - David Grossman's office is where?
- Different rules for different structures
  - office of (Noun Phrase) :: staff_office($1)
  - (Noun Phrase)'s office :: staff_office($1)

Level 1 Rules: Subject-Verb-Object

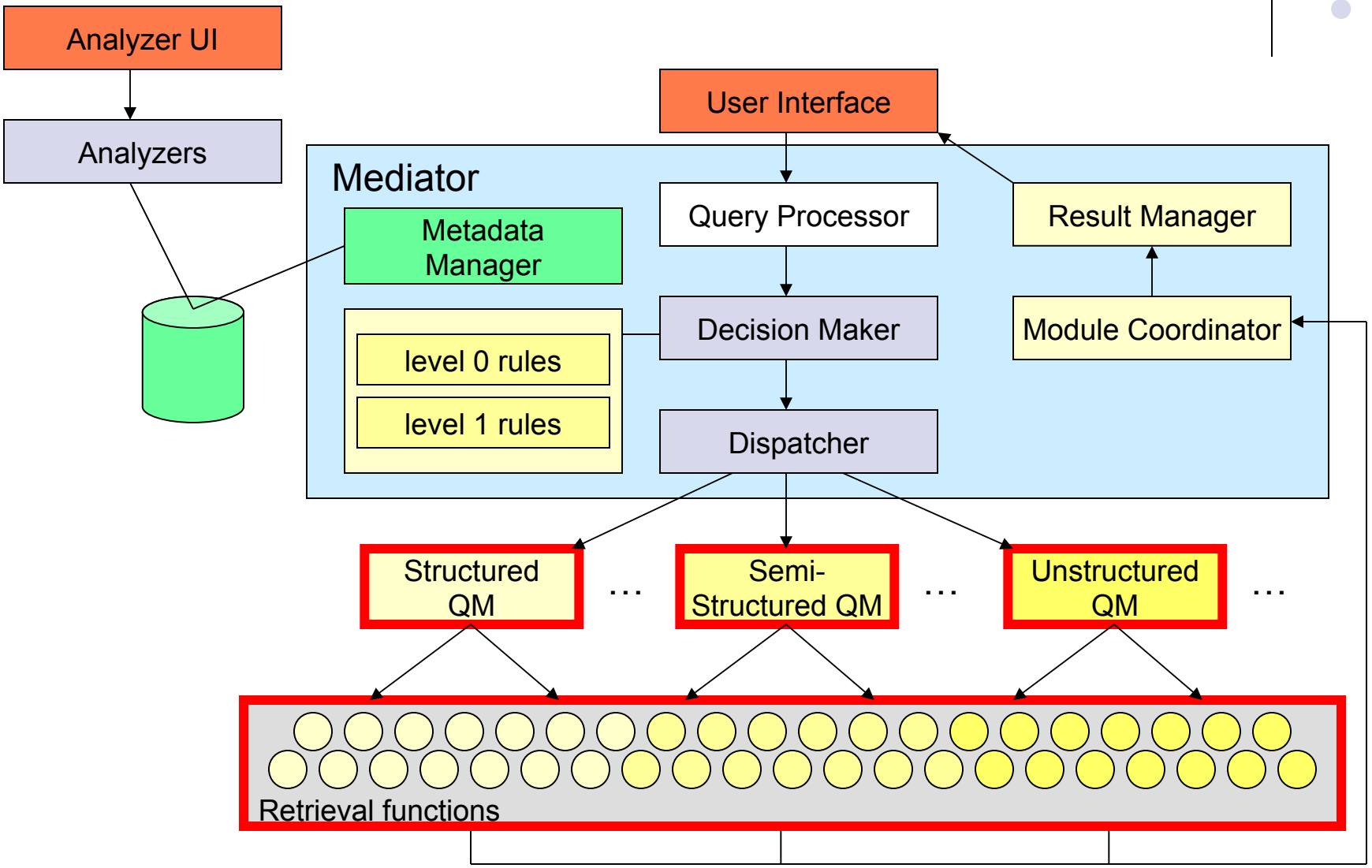
- **S**ubject-**V**erb-**O**bject identification
  - Where is the office of David Grossman?
  - Where is David Grossman's office?
  - David Grossman's office is where?

Notice how all three sentences can be parsed with one rule.

```
staff_office ($person) {
    S = "office";
    V = "is" | "are";
    O = $person AS
            %teacher;
    QT = #place;
}
```



Architecture



Retrieval Function Mapping

Where is the office of David Grossman?

SVO parse of the query:

S: office

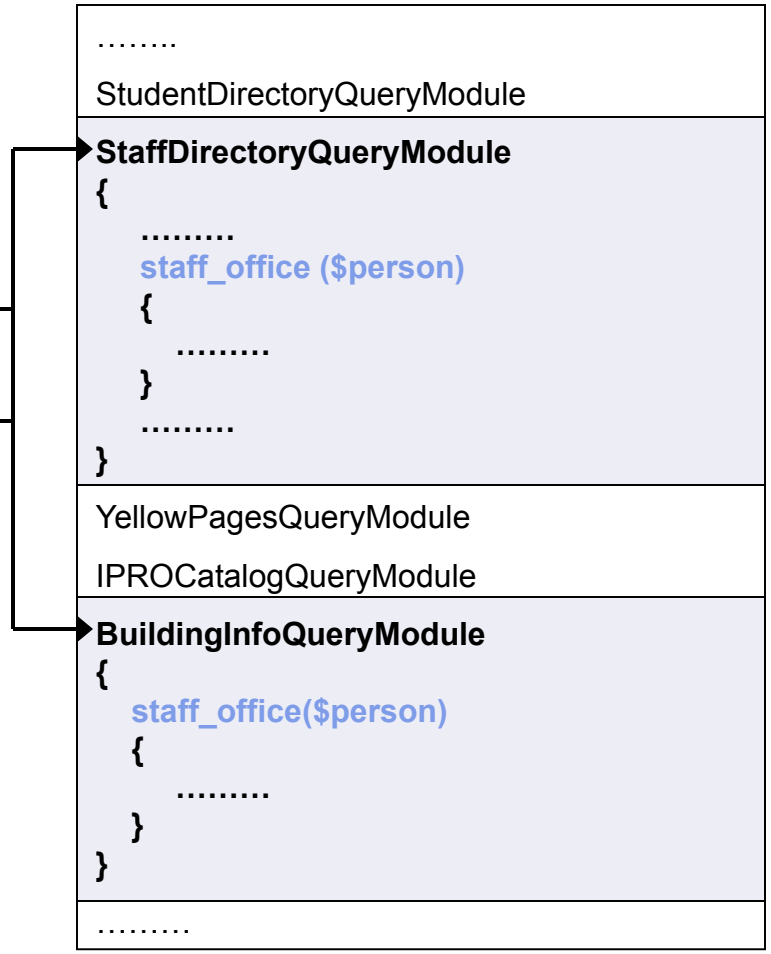
V: is

O: [David Grossman]

office of (Noun Phrase) :: staff_office($1)

staff_office ($person)

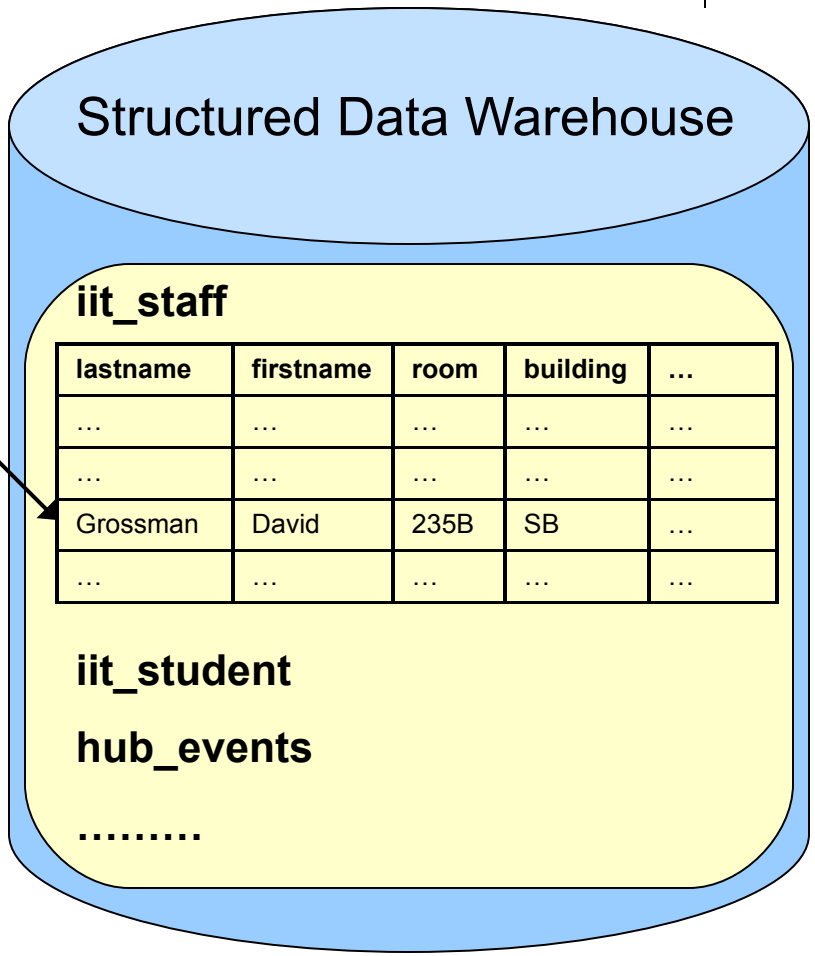
```
{
    S = “office”;
    V = “is” | “are”;
    O = $person AS %teacher;
    QT = #place;
}
```



Result Retrieval

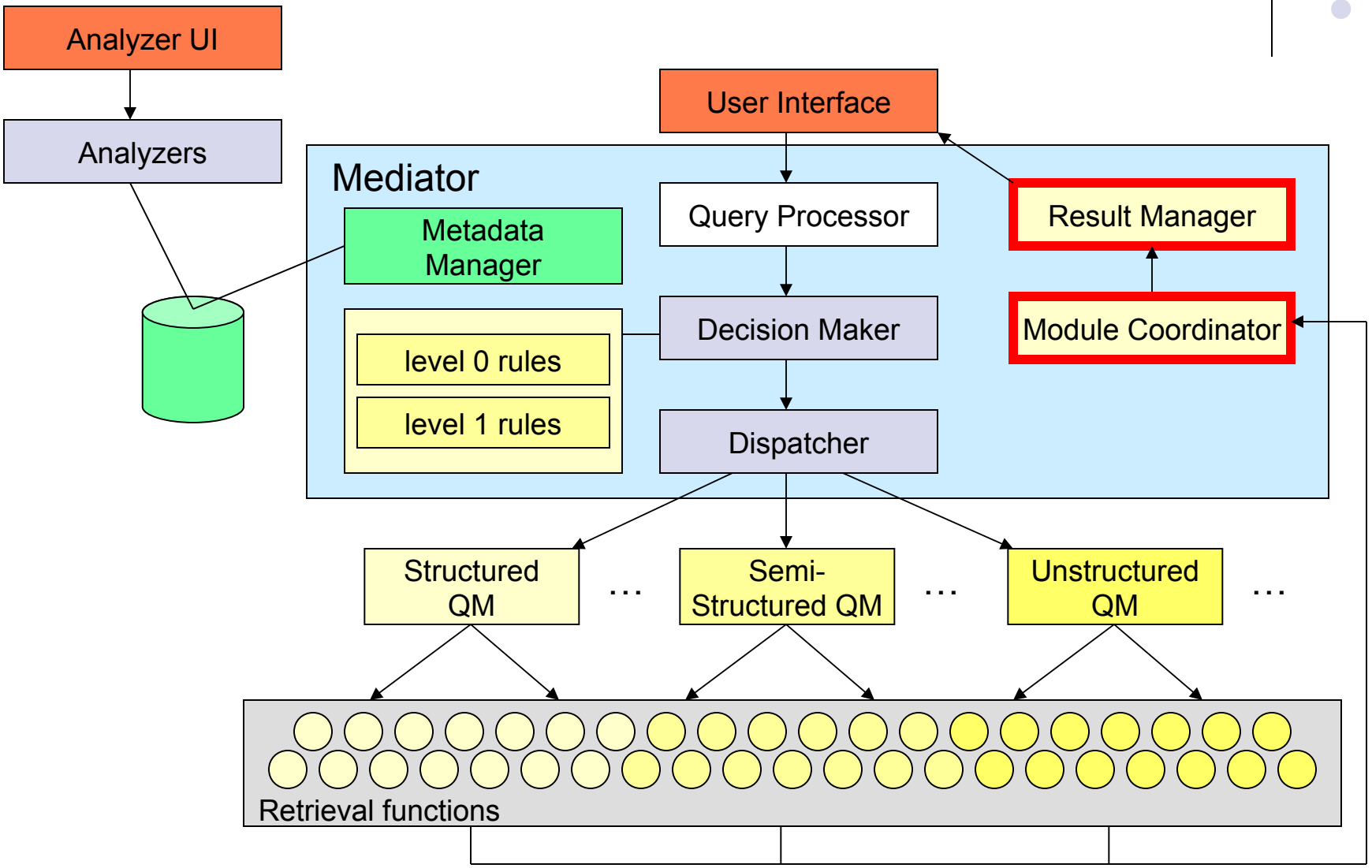
```
StaffDirectoryQueryModule
{
    staff_office ($person)
    {
        ……
    }
}
```

SQL

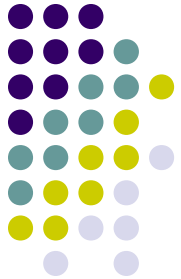


- Connect to the database
- Specify the table name for this source
- Query the table and try to get an **answer**
- Return the results

Architecture

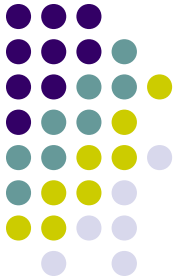


Result Unification



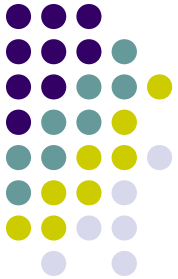
- Sources provide limited ranking and weighting of results
- Weights from disparate retrieval functions must be unified by weighting via:
  - Rule precision
  - Retrieval function precision
  - Source type

Experimental Evaluation



- 350 test queries about IIT
  - Built regression tester to verify them
- 22 Data Sources
  - 14 structured
    - Faculty and Student phone directories
  - 2 semi-structured
    - Course registration information
  - 6 unstructured
    - IIT website

Summary



- Mediator now runs numerous queries on numerous data sources.
- Source acquisition is *much* easier due to new architecture rule language and retrieval functions.
- Ready for commercial quality prototypes.
- Demo can now be turned over to CNS for production use on the IIT website.