

Entity-Centric Data Management

Philippe Cudré-Mauroux

[eXascale Infolab](#), University of Fribourg
Switzerland

eXascale Infolab

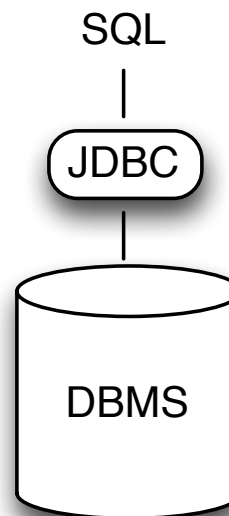
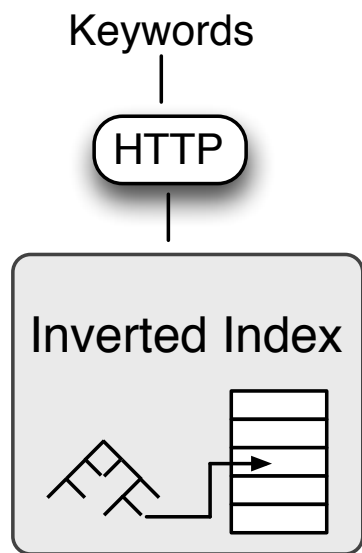
- New lab @ U. of Fribourg, Switzerland
- Financed by Swiss Federal State / private foundations / companies
- Big (non-relational) data management
(Volume, Velocity, **Variety**) (... mostly)



Entities

Information Management

- The story so far:
 - Strict separation between unstructured and structured data management infrastructures



Information Integration

- Information integration is still one of the **biggest** CS problem out there (according to many e.g., Gartner)
 - Information integration typically requires some sort of *mediation*
 1. Unstructured Data: keywords, synsets
 2. Structured Data: global schema, transitive closure of schemas (mostly syntactic)
- ⇒nightmarish if 1 and 2 taken separately, horror marathon if considered together

Entities as Mediation

- Rising paradigm
 - Store information at the entity granularity
 - Integrate information by inter-linking entities
- Advantages?
 - **Coarser** granularity compared to keywords
 - More natural, e.g., brain functions similarly (or is it the other way around?)
 - **Denormalized** information compared to RDBMSs
 - Schema-later, heterogeneity, sparsity
 - Pre-computed joins, “Semantic” linking
- Drawbacks?

Example: Google's Knowledge Graph

Google Inside Search

Home Tips & Tricks Features Search Stories Playground Blog Help

Calendar More

Search

avidbiker80@gmail.com 5 + Share

Search off

Frank Lloyd Wright

Frank Lloyd Wright was an American architect, interior designer, writer and educator, who designed more than 1,000 structures and completed 500 works. [Wikipedia](#)

Born: June 8, 1867, Richland Center

Died: April 9, 1959, Phoenix, AZ

Education: University of Wisconsin-Madison

Spouse: Olgivanna Wright (m. 1928), Maude "Miriam" Wright (m. 1923-1927). [More](#)

Children: John Lloyd Wright, Lloyd Wright

Structures

[Fallingwater](#) [Taliesin](#) [Taliesin West](#) [Robie House](#) [Ennis House](#)

People also search for

[Le Corbusier](#) [Frank Gehry](#) [Louis Sullivan](#) [Ludwig Mies van der Rohe](#) [Walter Gropius](#)

[Report a problem](#)

The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

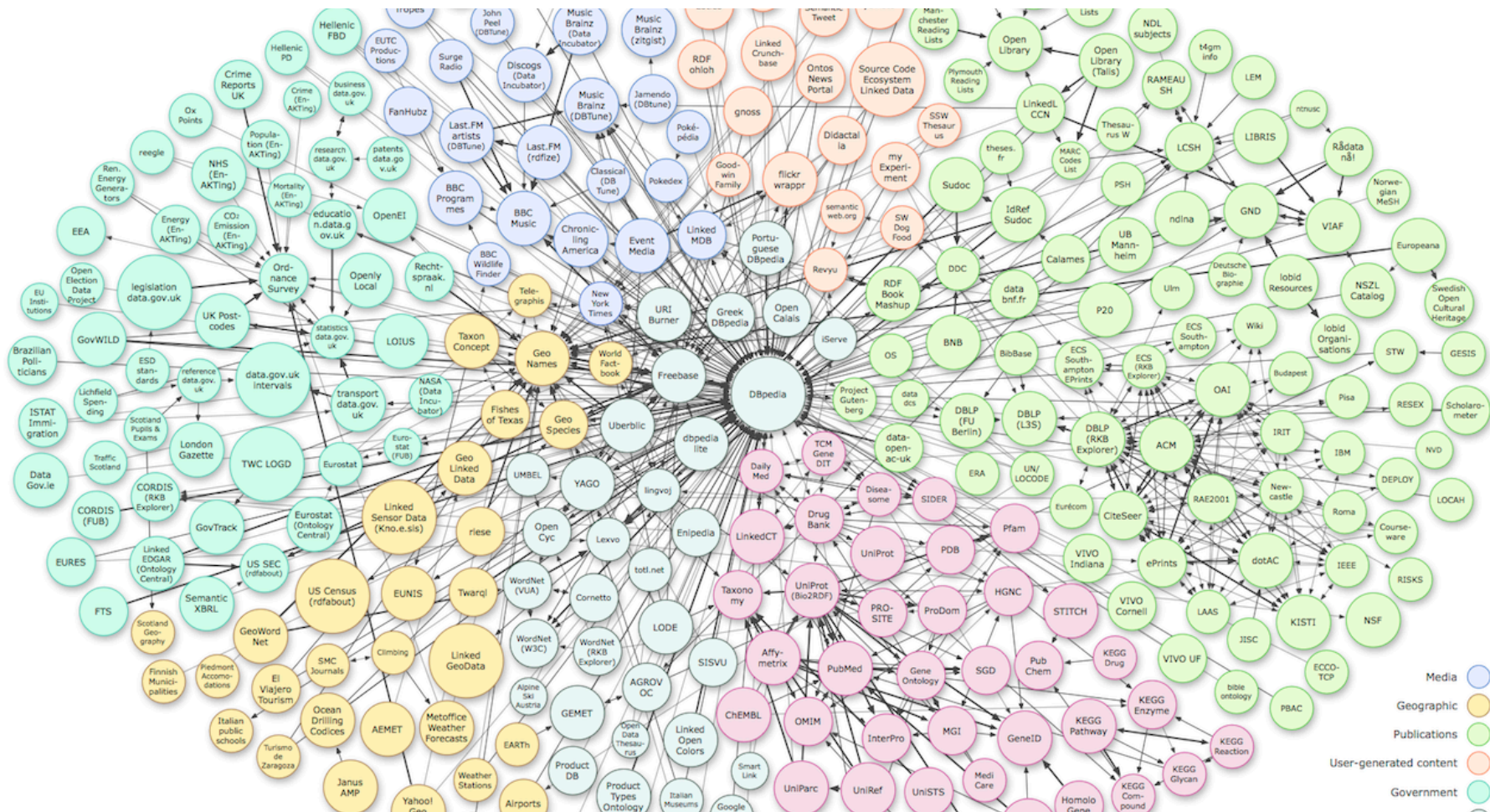
See Discover thought

[www.pbs.org/fllw/](#)
PBS **Frank Lloyd Wright** Web site, a companion to the Ken Burns/Lynn Hirschman film, contains biographical information, drawings and blueprints, analysis of ...

[Images for frank lloyd wright](#) - Report images

7

Example (2): Web Data Integration



<http://data.semanticweb.org/person/philippe-cudre-mauroux/html>

Example (3): BBC Olympics

BBC[News](#)[Sport](#)[Weather](#)[Travel](#)[Future](#)[TV](#)[Radio](#)[More...](#)

SPORT OLYMPICS

[Home](#) | [Football](#) | [Formula 1](#) | [Cricket](#) | [Rugby U](#) | [Rugby L](#) | [Tennis](#) | [Golf](#) [More Sports](#) ▼

[Olympics](#) | [London 2012](#) | [Team GB](#) | [Athletes](#) | [Countries](#) | [Venues](#) | [Guides](#) | [Schedule & Results](#) | [Medals](#) | [Olympic Sports](#) ▼

29 December 2012 Last updated at 00:12 GMT

London 2012



New Year Honours for Games stars

Britain's biggest stars of the London 2012 Olympic and Paralympic Games are recognised in the New Year Honours list.



Spectacular close to London 2012 Games



Best moments of the Olympic Games

BBC Sport's pundits and

Headlines

UK
[How the world saw London 2012](#)

[Lord Coe to stand to be BOA chair](#)

['Greatest ever' GB finish third](#)

[Broadcasting revolution of the digital Olympics](#)

TECHNOLOGY
[Record visits to BBC Sport online](#)

MODERN PENTATHLON
[Murray wins Britain's 65th & final medal](#)

AFRICA
[The effect of Mo Farah's success](#)

[Who will be in Britain's class of 2016?](#)

[How can Rio follow London?](#)

Medal Table

Show me:

| Rank | Country | | | | Total |
|------|----------------------------|----|----|----|-------|
| 1 | United States | 46 | 29 | 29 | 104 |
| 2 | China | 38 | 27 | 23 | 88 |
| 3 | Great Britain & N. Ireland | 29 | 17 | 19 | 65 |
| 33 | Switzerland | 2 | 2 | 0 | 4 |

[View full Switzerland table](#)

Olympics: Key info

[London 2012 Results](#)

[Full medal table](#)

[Team GB medals roll of honour](#)

[In numbers: the 2012 Games](#)

[Live text: London 2012 as it happened](#)

[Get involved in local sport](#)

Focusing on a few Core Problems

1. Extracting entities from text (*ZenCrowd*)
2. Searching for entities
3. Accessing entities (*DNS³*)
4. Storing entities (*Diplodocus[RDF]*)

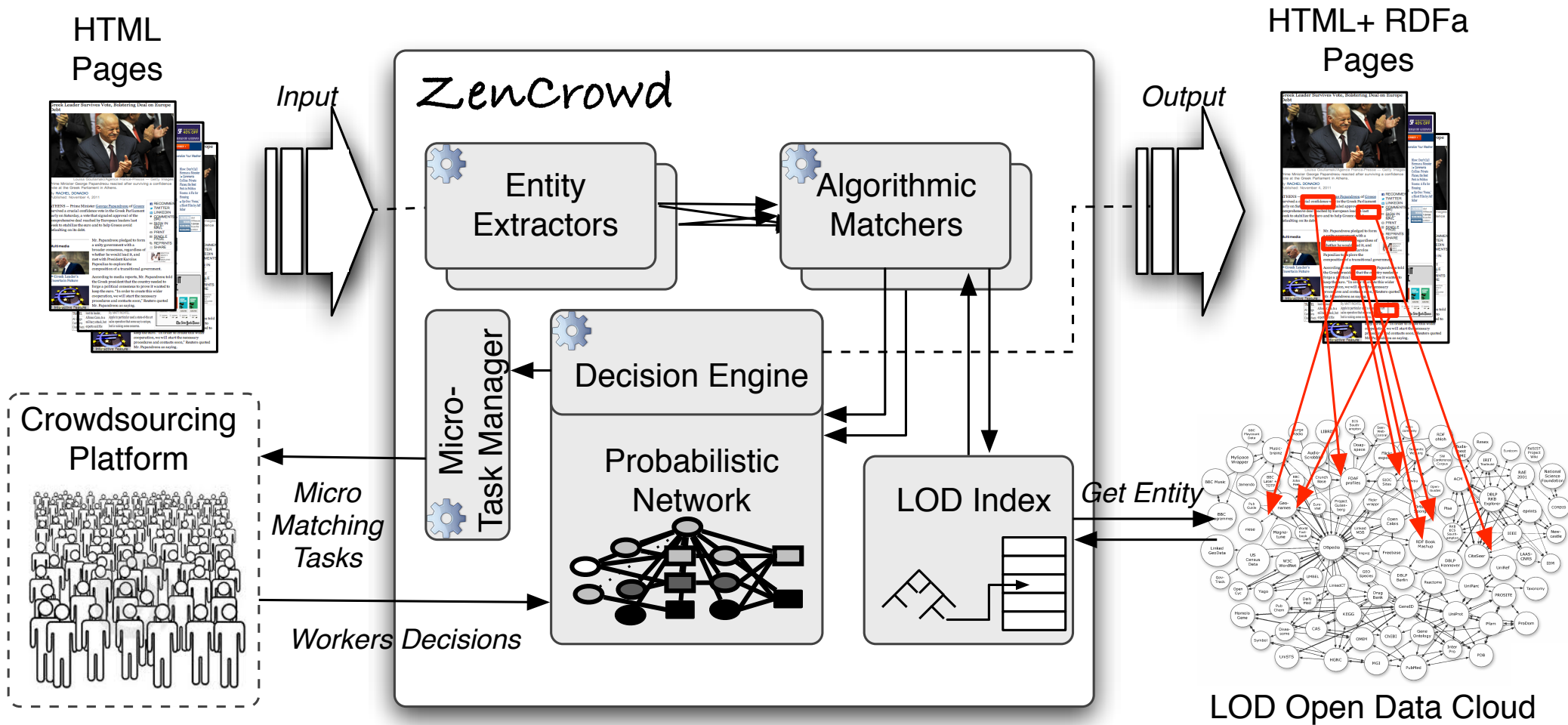
Extracting Entities

- Extracting entities from text is an **old** problem...
 - ... and is extremely **hard**, esp. for machines
- Dozens of approaches have been suggested
- What if
 - We want to **combine** approaches / frameworks?
 - We want to leverage both **human computations** & **algorithms**?

ZenCrowd

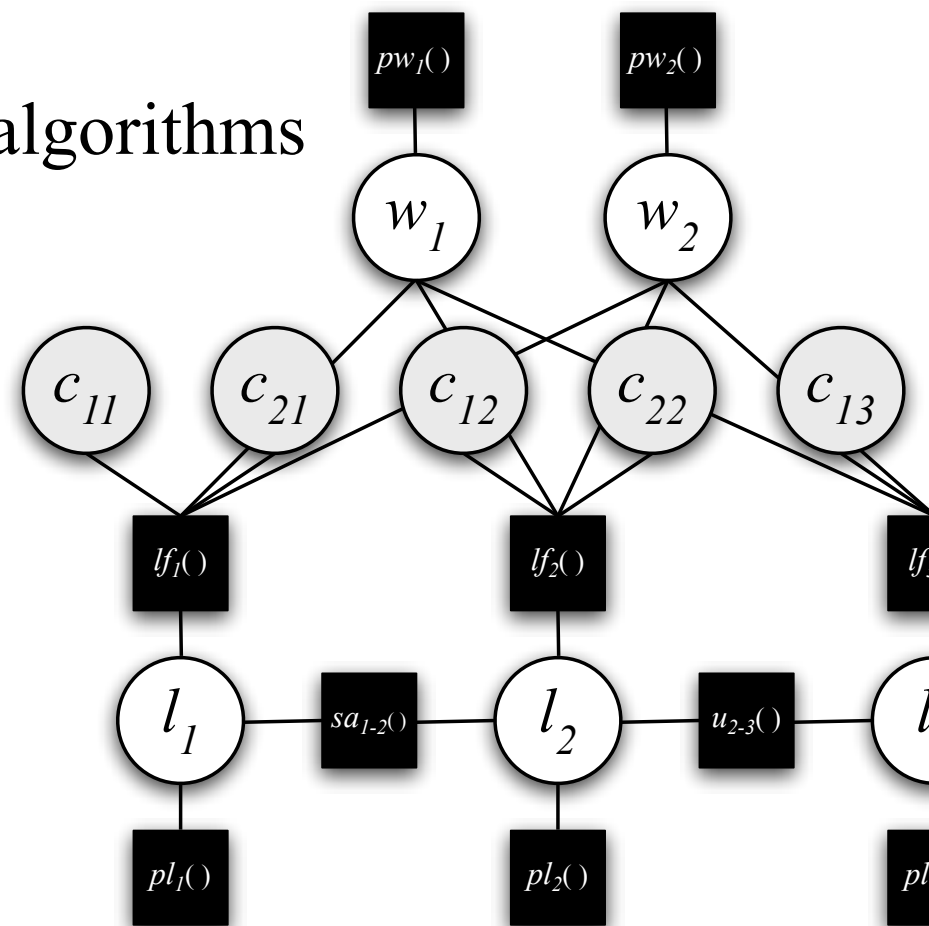
- Extracts entities from text using state-of-the-art techniques
- Uses sets of **algorithmic matchers** to match entities to online concepts
- Uses dynamic templating to create **micro-matching-tasks** and publish them on MTurk
- Combines both algorithmic and human matchers using **probabilistic networks**

ZenCrowd Architecture



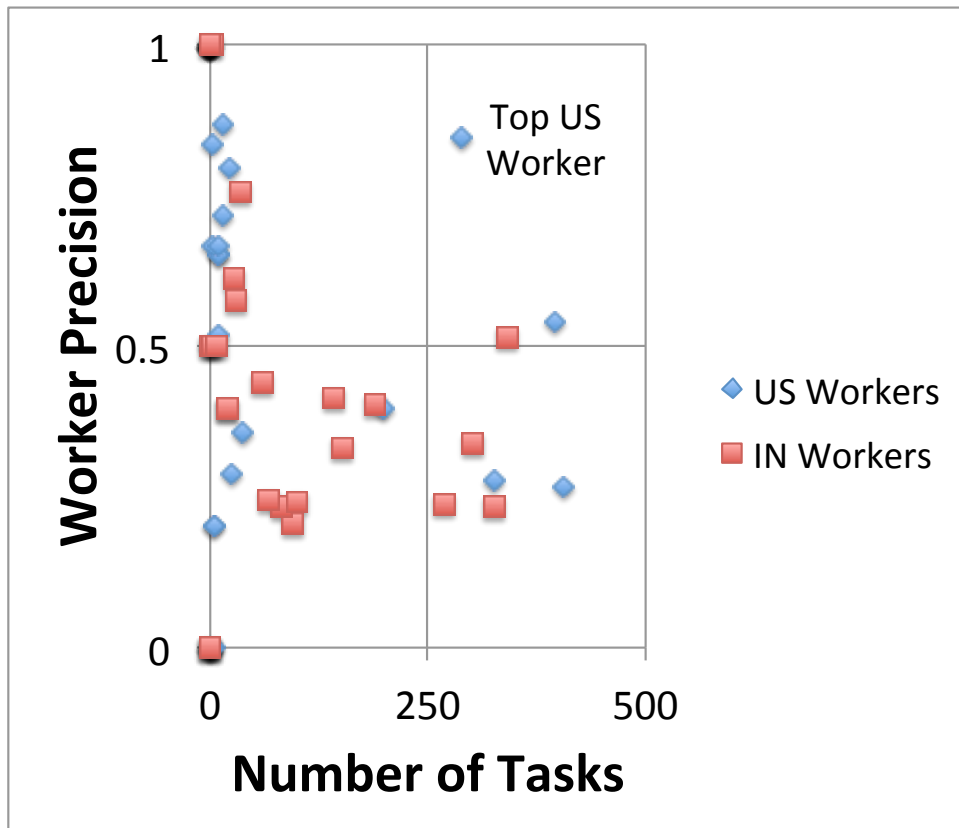
“Black-Magic” Component

- Probabilistic network to integrate a priori & a posteriori information
 - **Agreement** of good turkers & algorithms
 - Learning process
 - Constraints
 - Unicity
 - Equality (SameAs)
 - Giant probabilistic graph
 - Instantiated selectively



Does it Work?

- Improves avg. prec. by 0.14 on average!
 - Minimal crowd involvement
 - Embarrassingly parallel problem

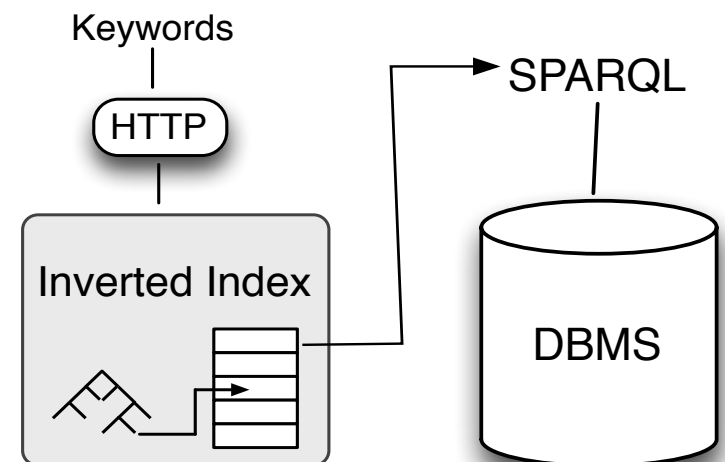
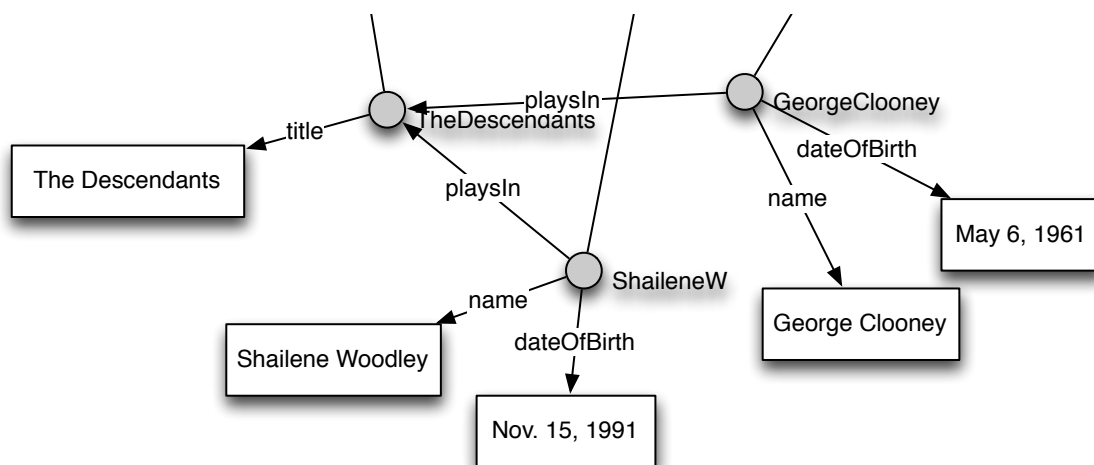


Searching for Entities

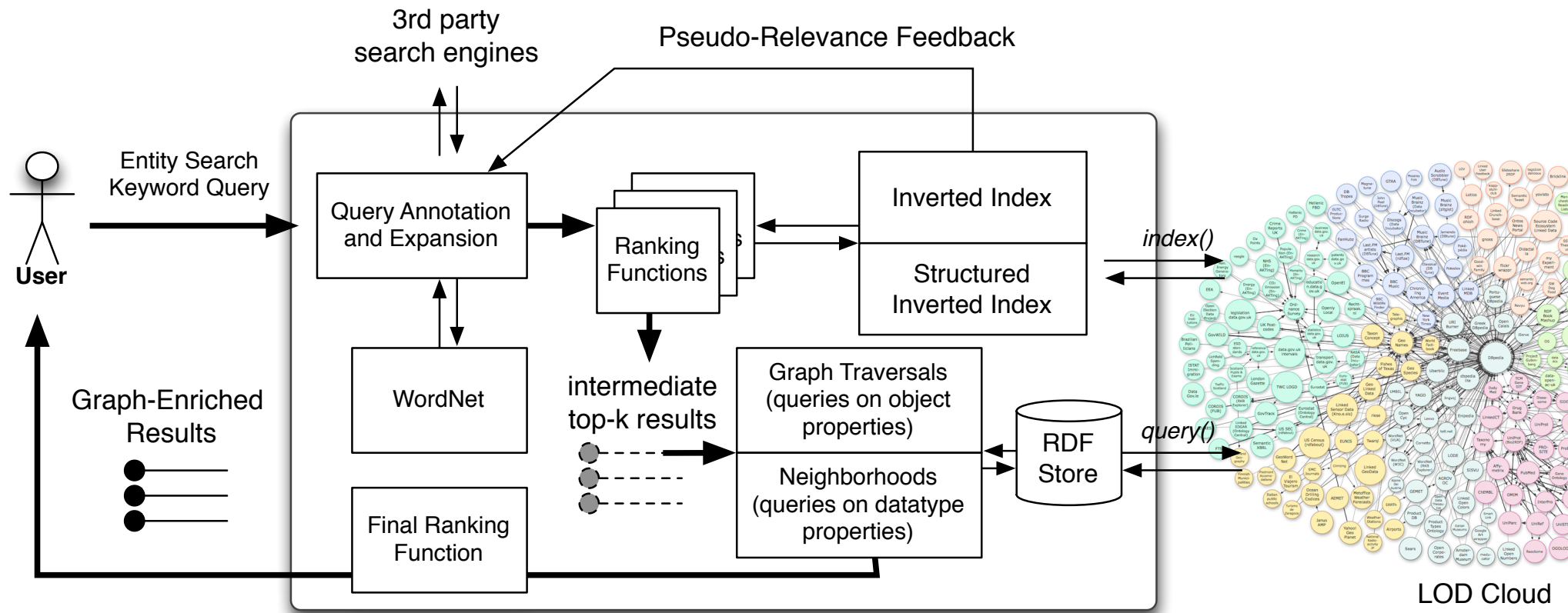
- How can end-users *reach* entities?
 - ⇒ Keyword search
 - On their names or attributes
 - Obviously not ideal
 - BM25 on TREC 2011 AOR: **MAP=0.15, P@10=0.20**
 - Query extension, query completion or pseudo-relevance feedback yield comparable (or worse) results

Hybrid Entity Search

- Main idea: combine unstructured and structured search
 - Inverted index to locate first candidates
 - Graph queries to refine the results
 - **Graph traversals** (queries on object properties)
 - **Graph neighborhoods** (queries on data type properties)



Architecture



Query Refinement

- Finding the **right graph queries** to refine results is a real challenge
 - Which object property to follow? Transitive closures?

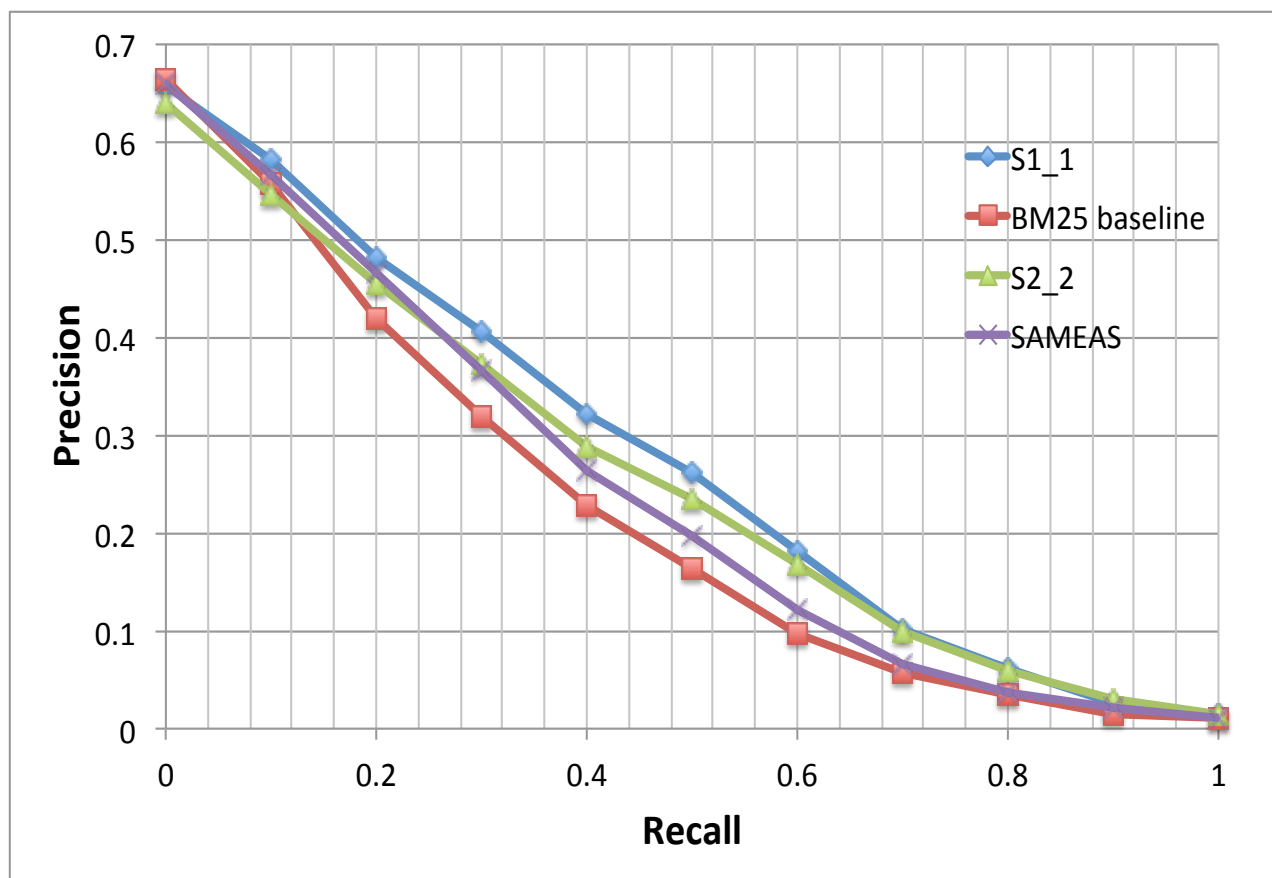
| From retrieved (x) to relevant entities (y) | | | |
|---|---|--------|--------|
| SemSearch 2010 | Object Property | Recall | Prec |
| | <http://dbpedia.org/property/wikilink> | 82.10% | 0.81% |
| | <skos:subject> | 11.75% | 0.7% |
| | <http://www.w3.org/2002/07/owl#sameAs> | 1.60% | 1.54% |
| | <http://dbpedia.org/ontology/artist> | 0.98% | 15.42% |
| | <http://dbpedia.org/property/disambiguates> | 0.68% | 1.98% |
| | <http://dbpedia.org/property/title> | 0.55% | 1.81% |
| | <http://dbpedia.org/ontology/producer> | 0.43% | 2.87% |
| | <http://dbpedia.org/property/region> | 0.43% | 8.37% |
| | <http://dbpedia.org/property/first> | 0.37% | 7.32% |
| | <http://dbpedia.org/property/redirect> | 0.25% | 3.91% |

- Which data type property to take into account? Scope?

| Datatype Property | JW(e',q) | Occurrences |
|--|----------|-------------|
| <http://www.w3.org/2006/03/wn/wn20/schema/lexicalForm> | 0.8449 | 19 |
| <http://dbpedia.org/property/county> | 0.8005 | 17 |
| <http://www.daml.org/2003/02/fips55/fips-55-ont#name> | 0.7674 | 27 |
| <http://www.geonames.org/ontology#name> | 0.7444 | 78 |
| <http://www.actors.org/ontology/portal#full-name> | 0.7360 | 55 |
| <http://dbpedia.org/property/wikiquoteProperty> | 0.7096 | 10 |
| <http://www.w3.org/2004/02/skos/core#prefLabel> | 0.6911 | 158 |
| <http://purl.org/dc/elements/1.1/title> | 0.6711 | 236 |
| <http://sw.opencyc.org/concept/Mx4rwLSVCpwpEbGdrcN5Y29ycA> | 0.6680 | 48 |
| <http://dbpedia.org/property/officialName> | 0.6623 | 54 |

Does it Work?

- Up to 25% improvement over best IR (stat. sign.)
- Very modest impact on latency (17%)



Entity Registries, i.e., Whom to Ask for Entities?

- (Search Engine+) Void + SPARQL end-point
- DOA
- DNS [DNS³]
- Same_as service / Okkam IDs
- P2P Mesh of entities [idMesh]
- (Downscaled) Entity registries?



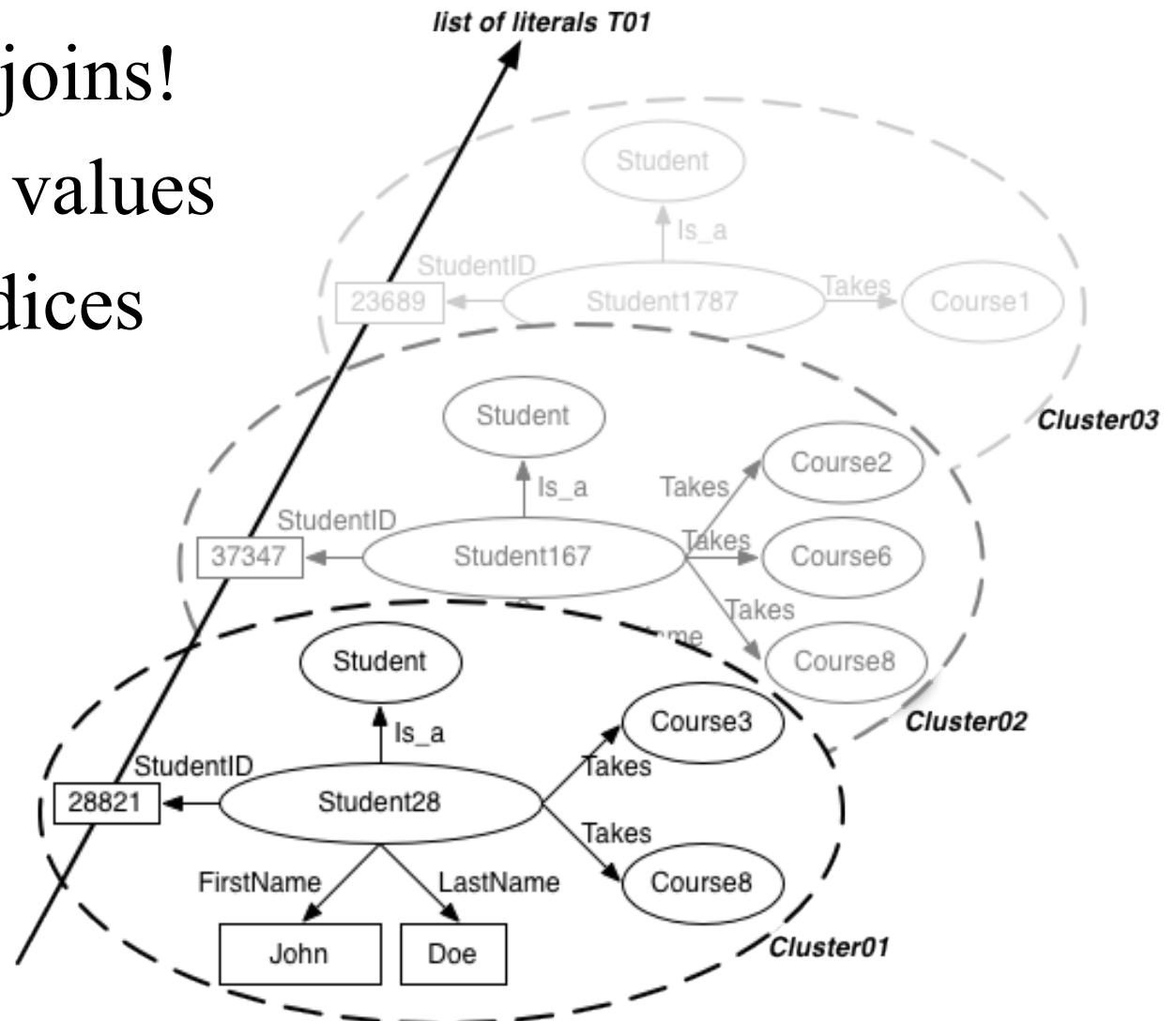
How to Store Entities?

- Fundamental impedance mismatch between graphs of entities and...
 - N-ary / decomposition storage model
 - Inverted Indices
 - Key-value paradigms

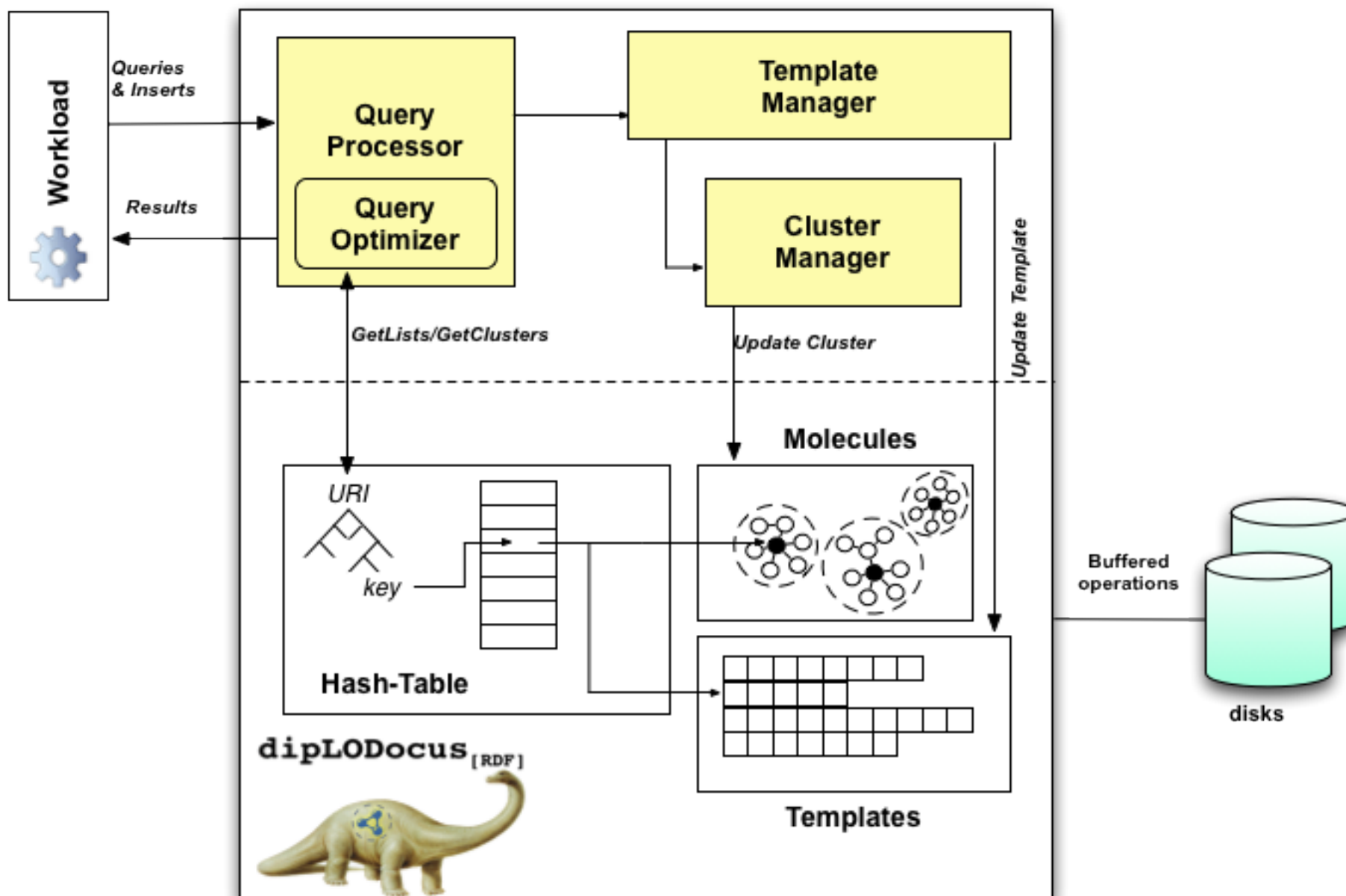
Entity Storage

- Materialize the joins!
- Dense-pack the values
- Provide new indices

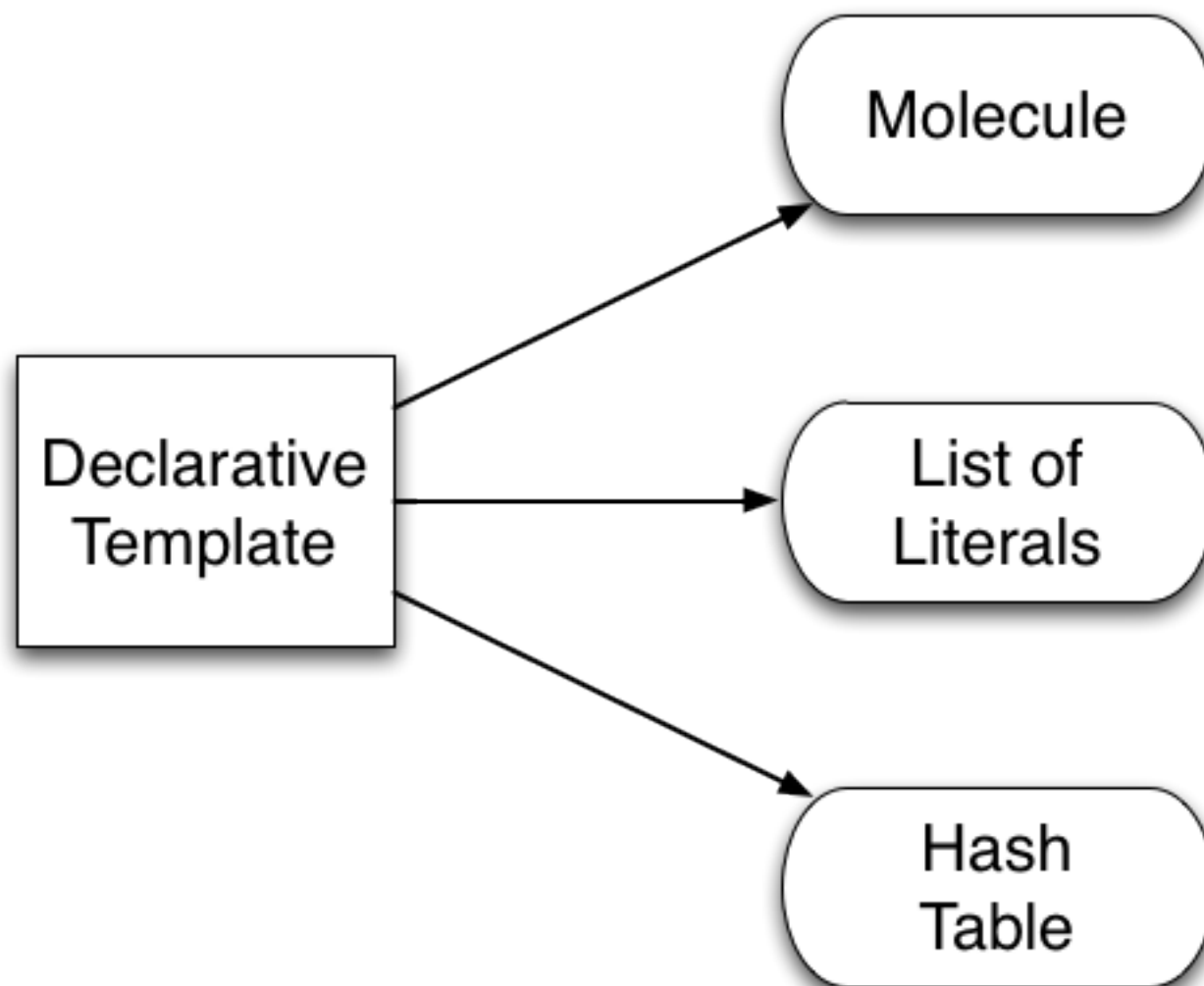
- Co-locate
- Co-locate
- Co-locate



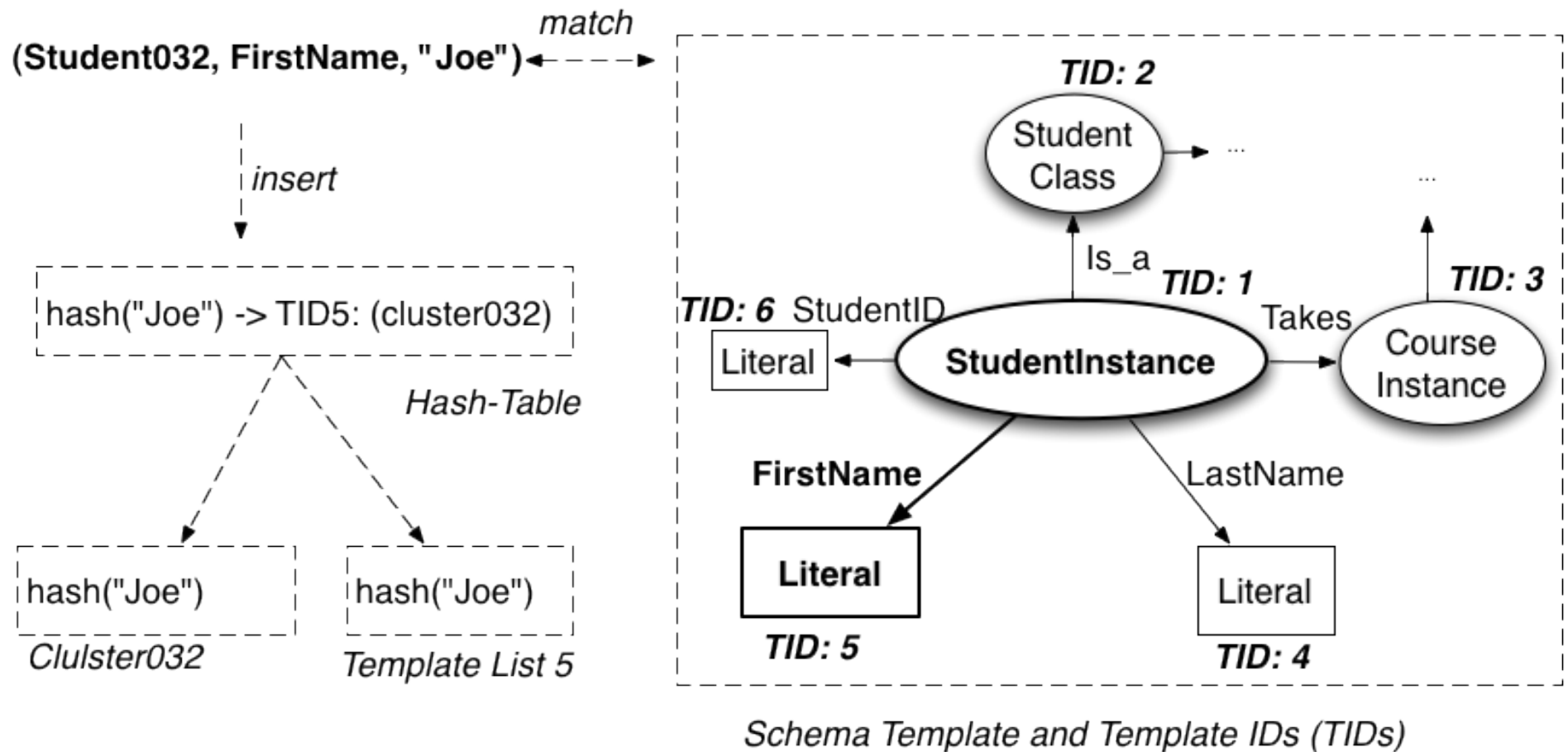
DipLODocus: System Architecture



Main Idea - data structures

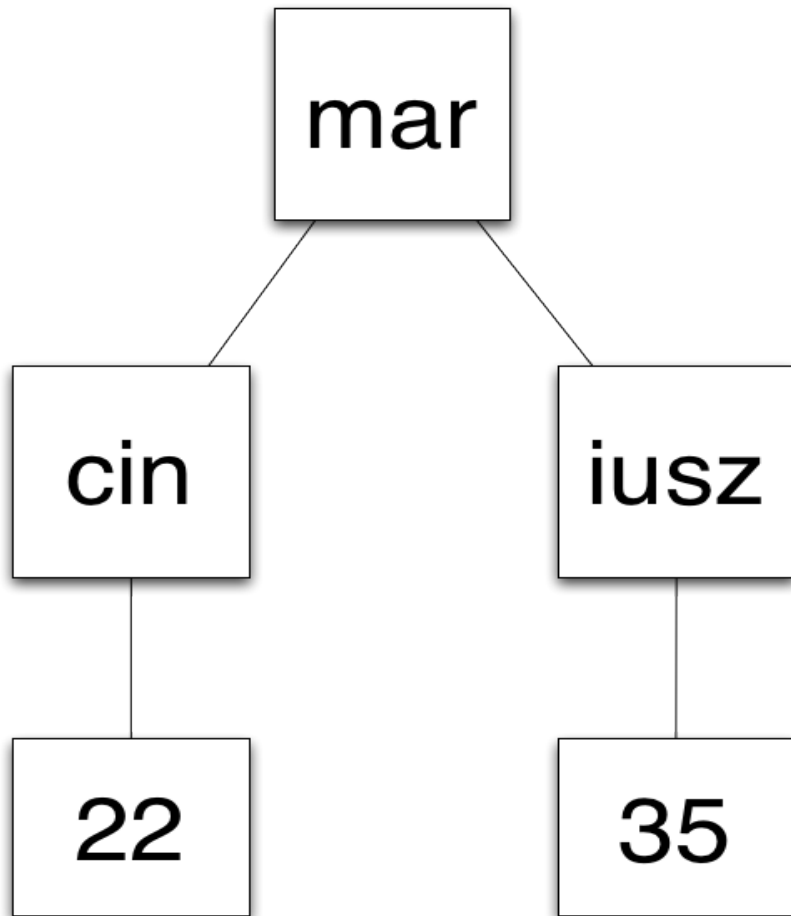


Template Matching (Bottom-up \neq schema)

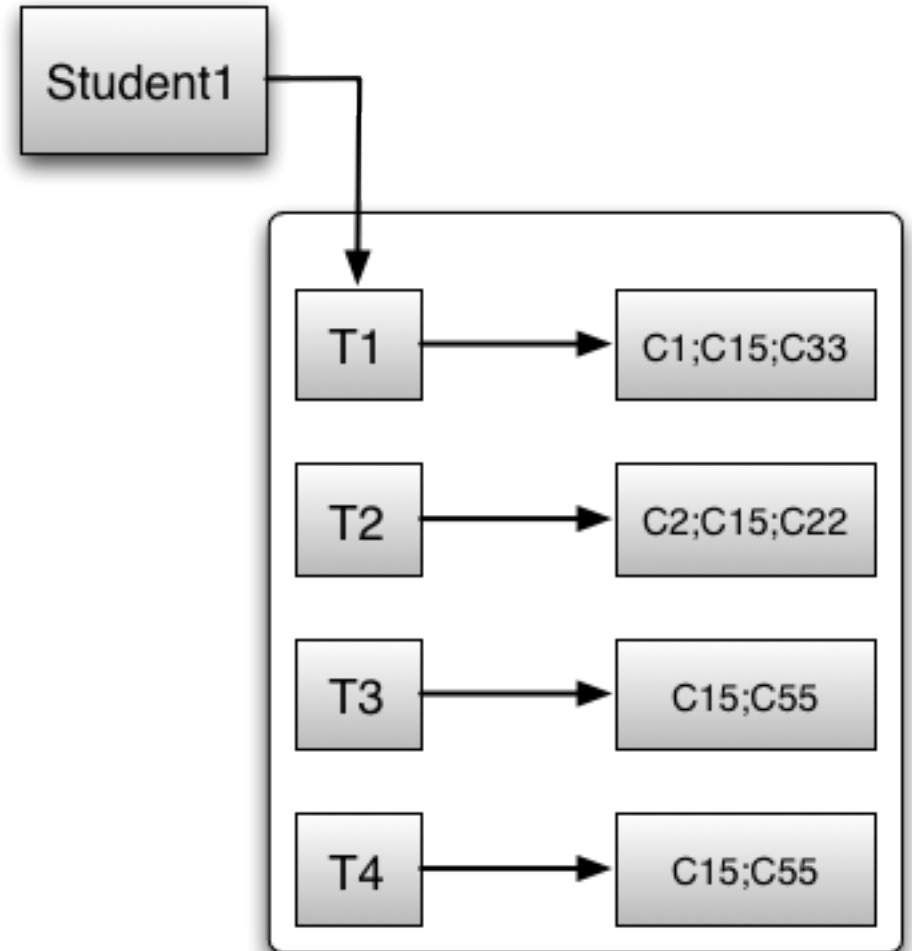


Hash Table

lexicographic tree
to encode URIs



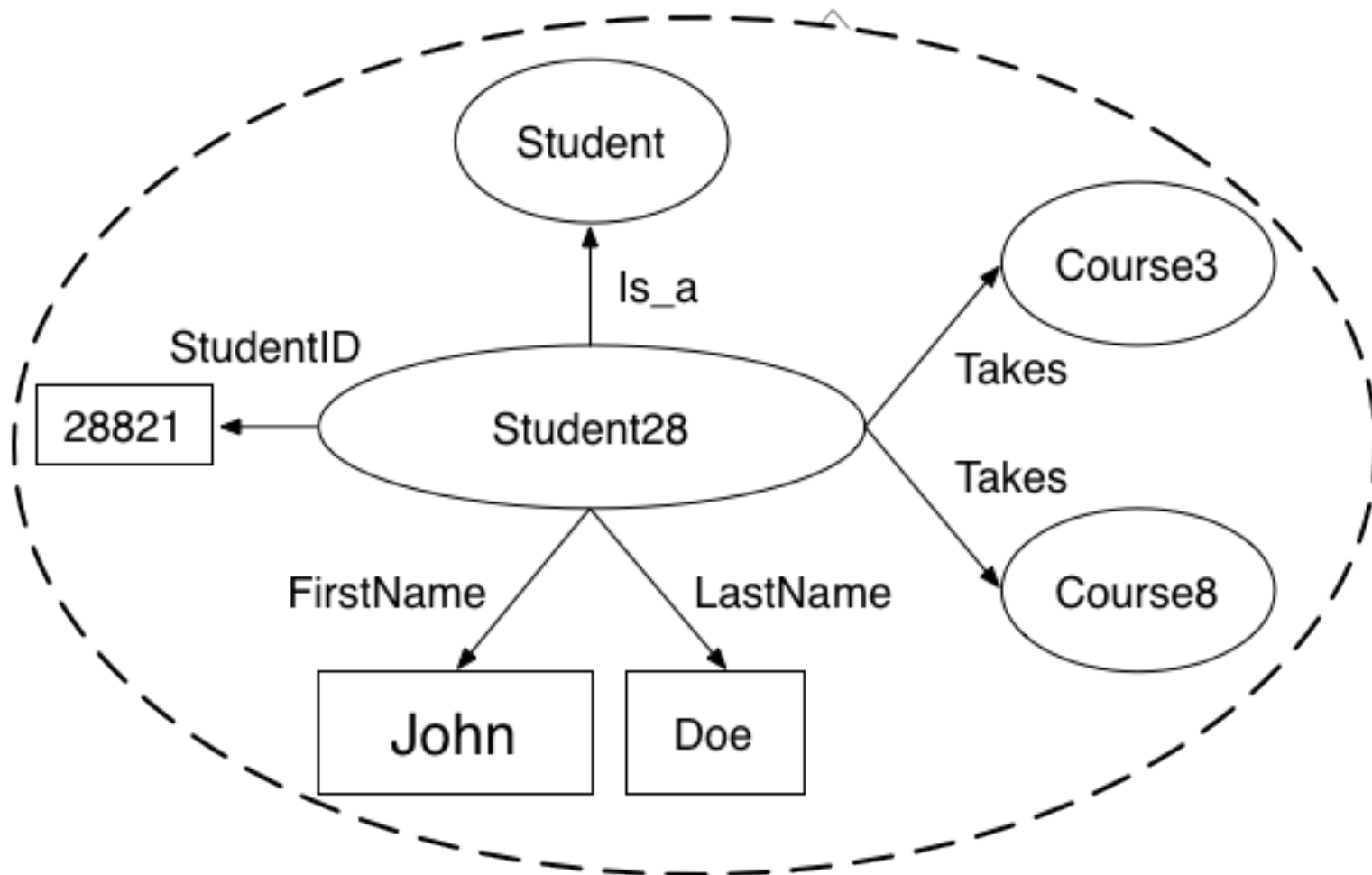
template based
indexing



extremely compact lists of
homologous nodes

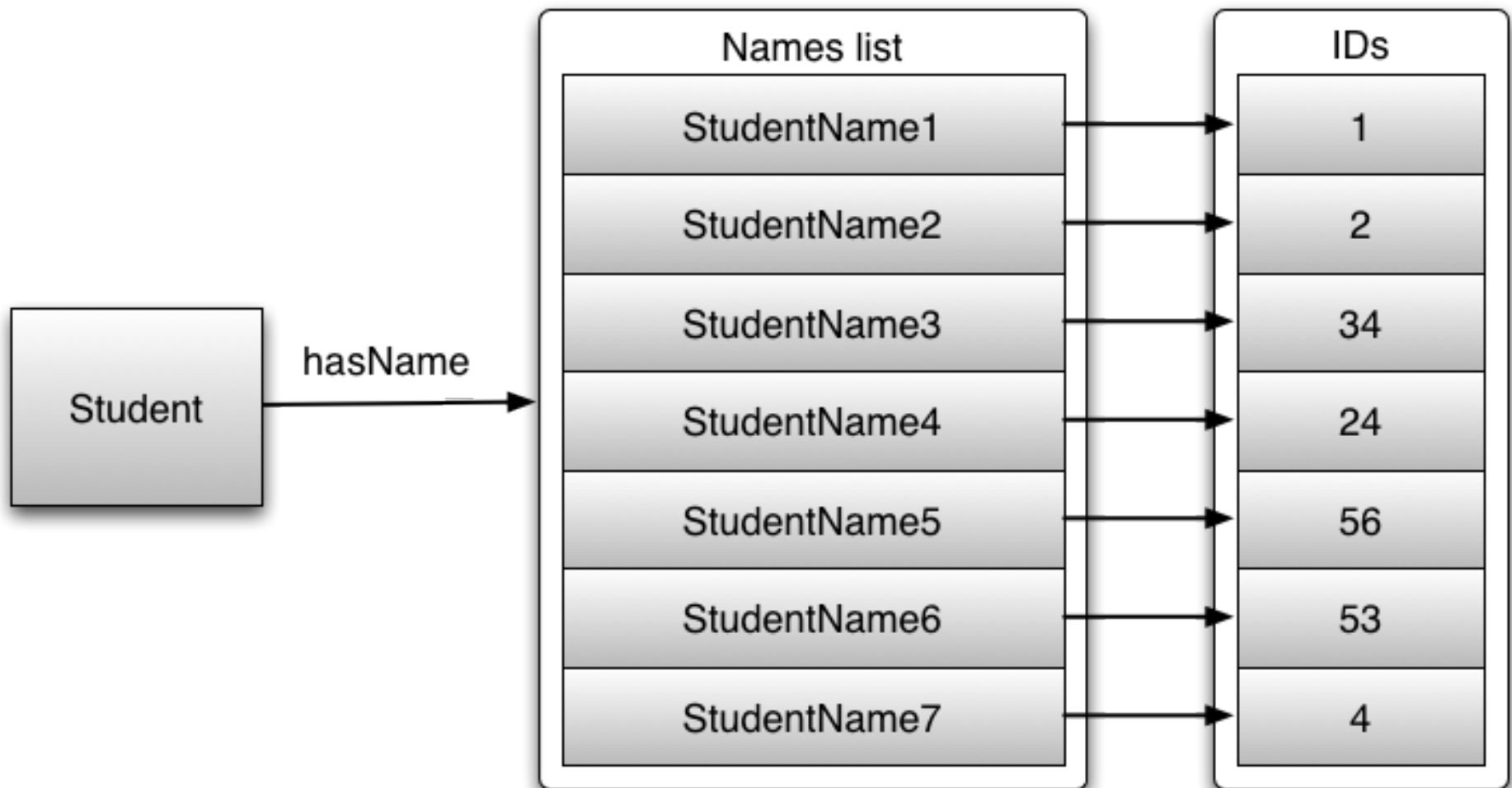
Molecule Clusters

- extremely compact sub-graphs
- precomputed joins



List of Literals

- extremely compact list of sorted values



Basic operations - queries - triple patterns

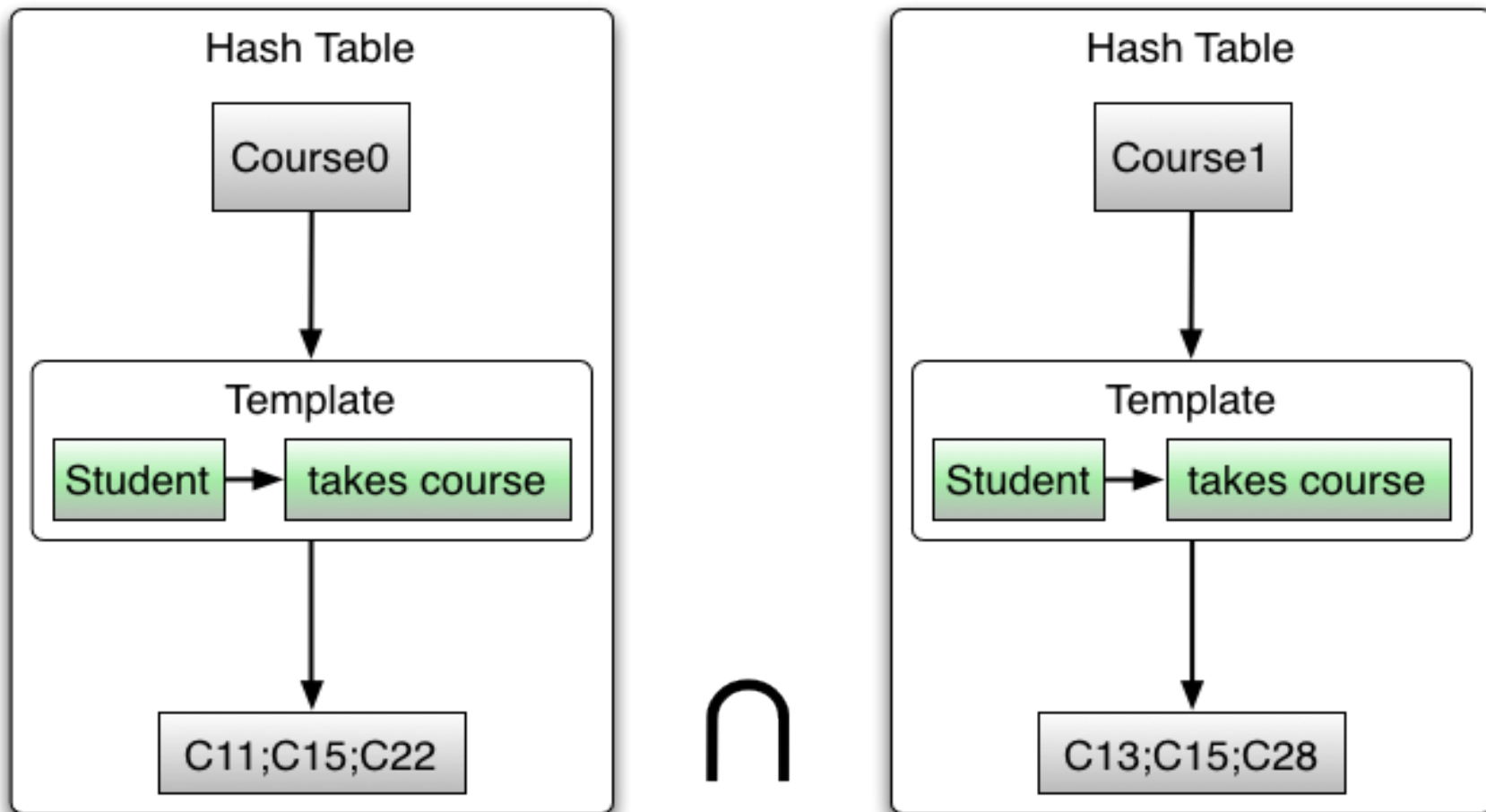
?x type Student.

?x takesCourse Course0.

?x type Student.

?x takesCourse Course0.

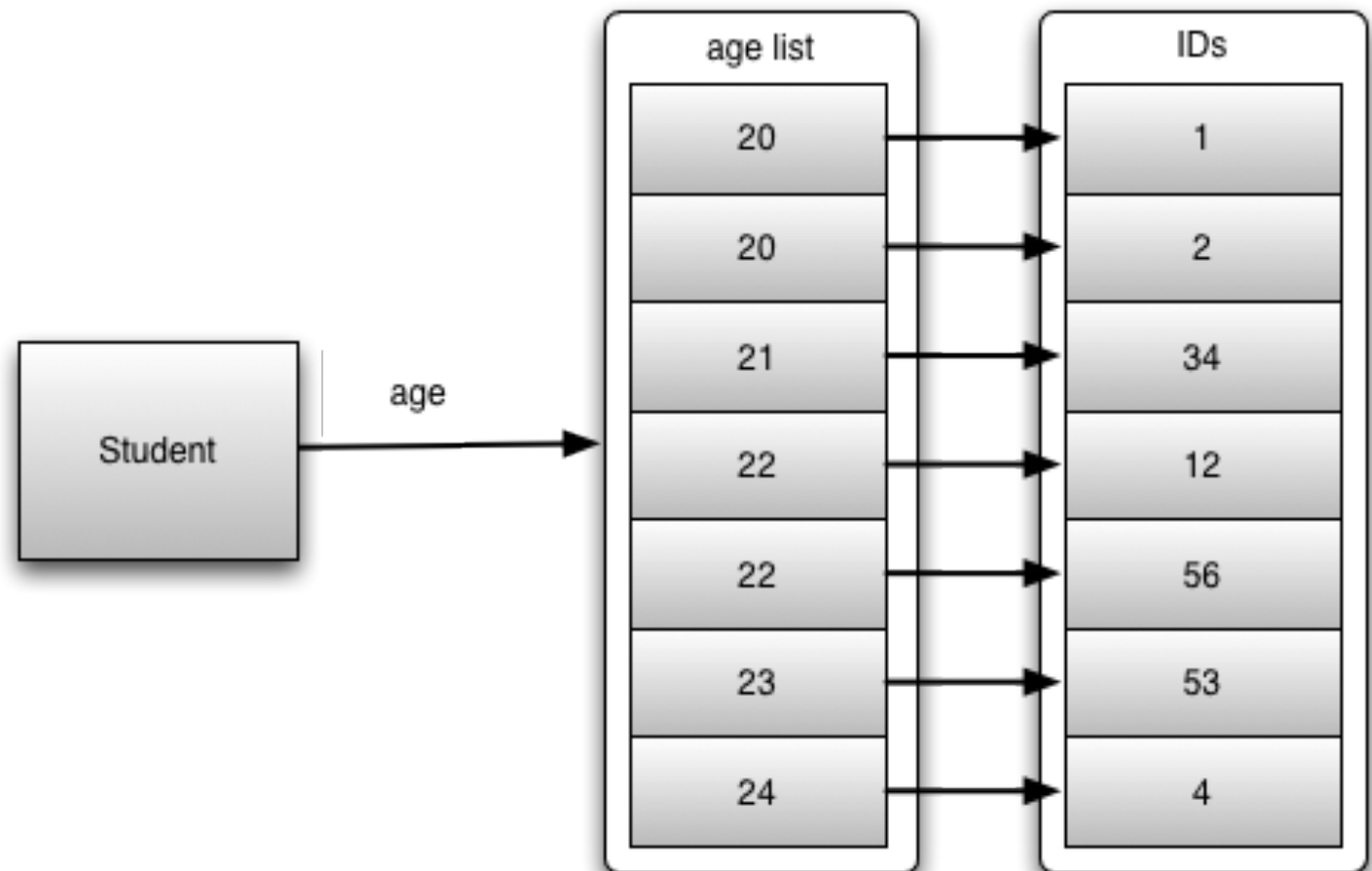
?x takesCourse Course1.



=> intersection of sorted lists

Basic operations - queries aggregates and analytics

?x type Student.
?x age ?y
filter (?y < 21)

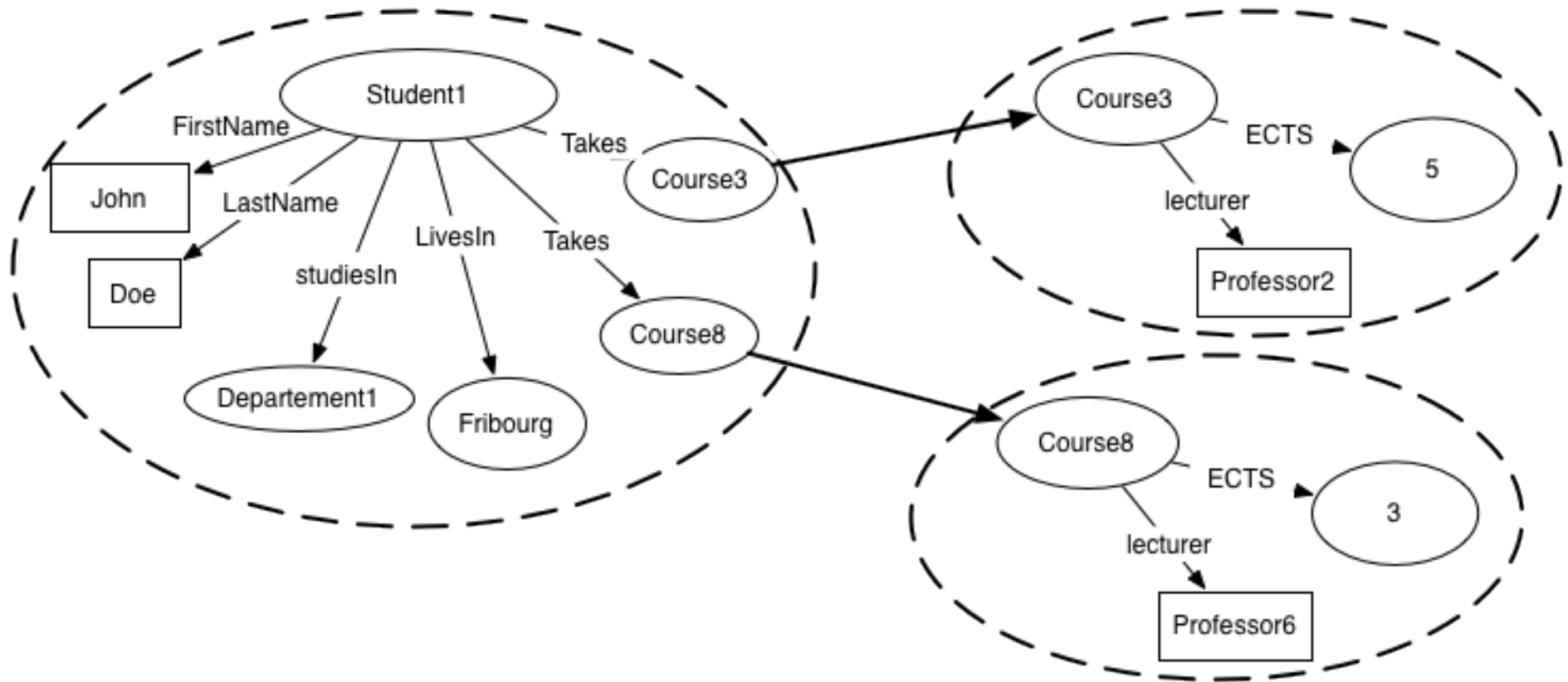


Basic operations - queries - molecule queries

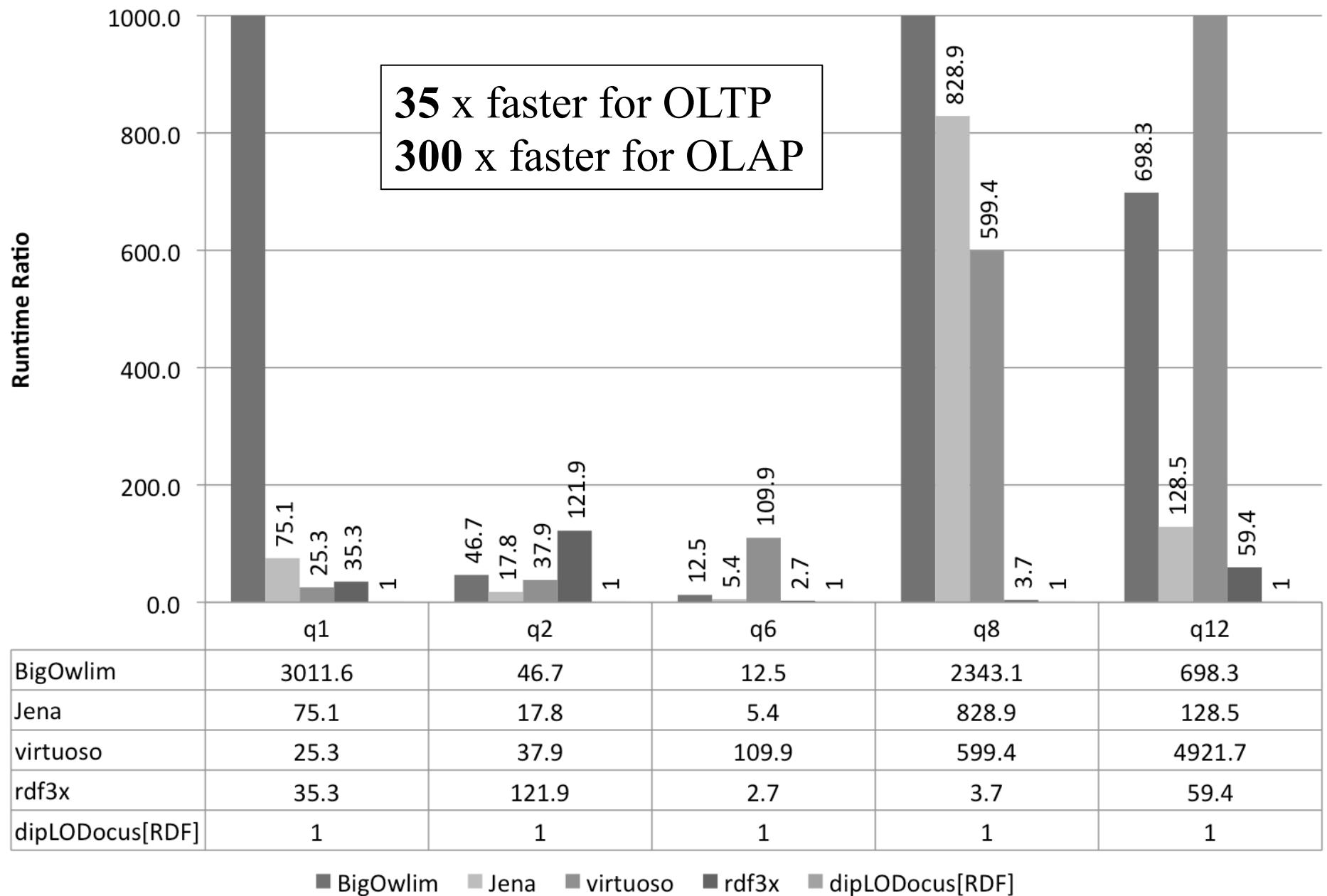
?a name 'Student1'.

?a ?b ?c.

?c ?d ?e.



Results - LUBM - 100 Universities



References

- **ZenCrowd** [WWW 12, WWW 13, VLDB J. 13]
- **idMesh** [WWW 09]
- **Hybrid Entity Search** [SIGIR 12]
- **Downscaling Entity Registries** [Downscale 12]
- **Diplodocus[RDF]** [ISWC 11]
- **Graph Data Management for New Application Domains** [VLDB 11]

<http://exascale.info>

