D2R2: Disk-oriented Deductive Reasoning in a RISC-style RDF Engine

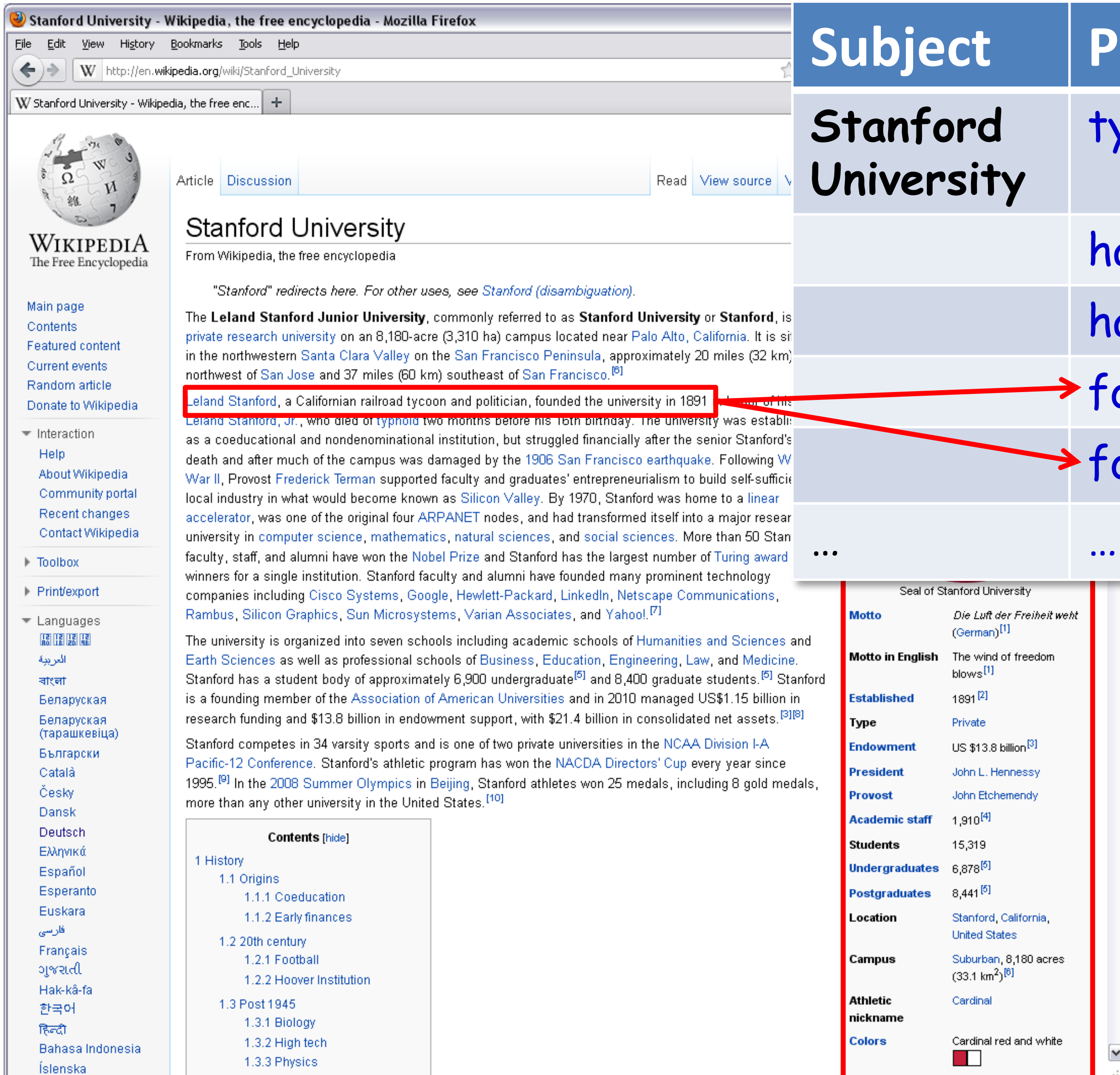
RuleML

Nov. 03, 2011

Mohamed Yahya
Martin Theobald
{myahya,mtb}@mpi-inf.mpg.de

Max-Planck Institute for Informatics,
Saarbrücken, Germany

Background: Information Extraction


Stanford University - Wikipedia, the free encyclopedia - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://en.wikipedia.org/wiki/Stanford_University

Stanford University - Wikipedia, the free enc...

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages

العربية
বাংলা
Беларуская
Беларуская (тарашкевіца)
Български
Català
Česky
Dansk
Deutsch
Ελληνικά
Español
Esperanto
Euskara
فارسی
Français
ગુજરાતી
Hak-kâ-fa
한국어
हिन्दी
Bahasa Indonesia
Íslenska

Article Discussion Read View source

Stanford University

From Wikipedia, the free encyclopedia

"Stanford" redirects here. For other uses, see Stanford (disambiguation).

The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is private research university on an 8,180-acre (3,310 ha) campus located near Palo Alto, California. It is si in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of San Jose and 37 miles (60 km) southeast of San Francisco.[6]

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1891 Leland Stanford, Jr., who died of typhoid two months before his 16th birthday. The university was establi as a coeducational and nondenominational institution, but struggled financially after the senior Stanford's death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following W War II, Provost Frederick Terman supported faculty and graduates' entrepreneurialism to build self-sufficie local industry in what would become known as Silicon Valley. By 1970, Stanford was home to a linear accelerator, was one of the original four ARPANET nodes, and had transformed itself into a major resear university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stan faculty, staff, and alumni have won the Nobel Prize and Stanford has the largest number of Turing award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including Cisco Systems, Google, Hewlett-Packard, LinkedIn, Netscape Communications, Rambus, Silicon Graphics, Sun Microsystems, Varian Associates, and Yahoo!.[7]

The university is organized into seven schools including academic schools of Humanities and Sciences and Earth Sciences as well as professional schools of Business, Education, Engineering, Law, and Medicine. Stanford has a student body of approximately 6,900 undergraduate[5] and 8,400 graduate students.[5] Stanford is a founding member of the Association of American Universities and in 2010 managed US$1.15 billion in research funding and $13.8 billion in endowment support, with $21.4 billion in consolidated net assets.[3][8]

Stanford competes in 34 varsity sports and is one of two private universities in the NCAA Division I-A Pacific-12 Conference. Stanford's athletic program has won the NACDA Directors' Cup every year since 1995.[9] In the 2008 Summer Olympics in Beijing, Stanford athletes won 25 medals, including 8 gold medals, more than any other university in the United States.[10]

Contents [hide]

1 History
1.1 Origins
1.1.1 Coeducation
1.1.2 Early finances
1.2 20th century
1.2.1 Football
1.2.2 Hoover Institution
1.3 Post 1945
1.3.1 Biology
1.3.2 High tech
1.3.3 Physics

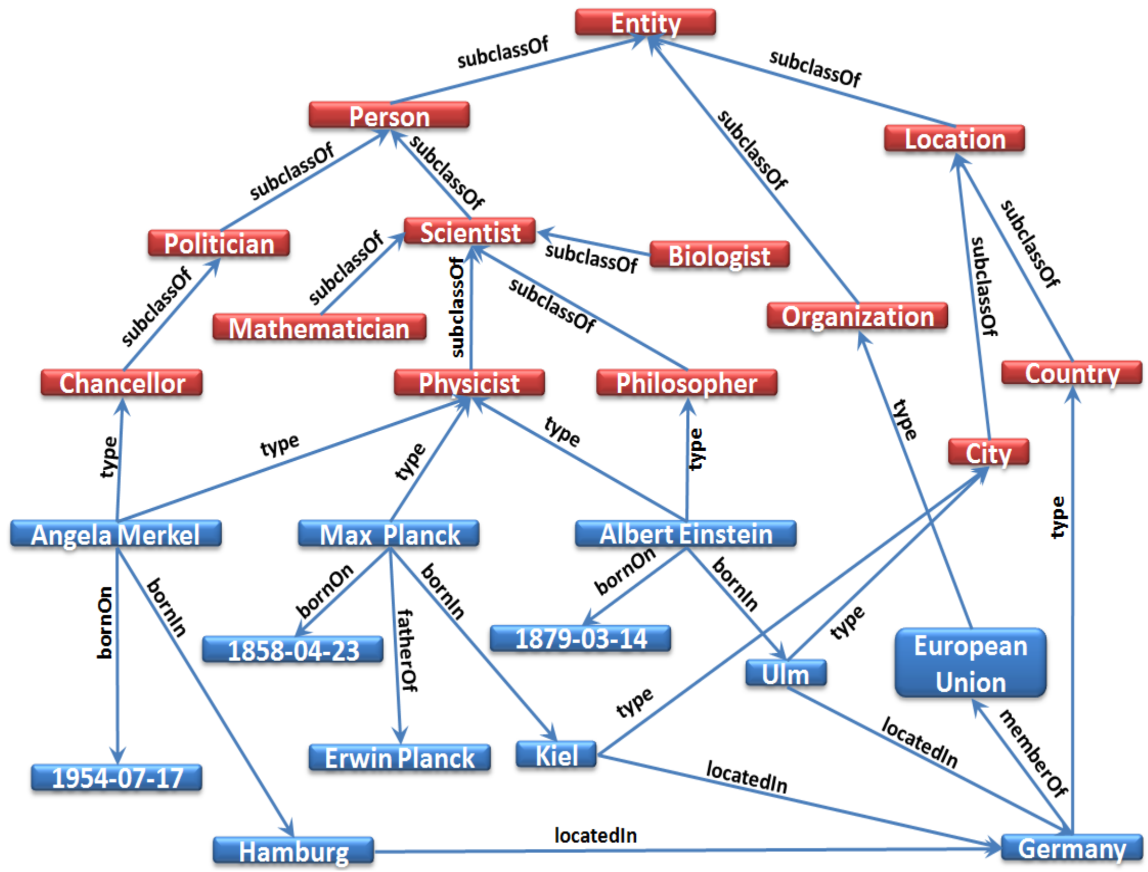
Seal of Stanford University

Motto	Die Luft der Freiheit weht (German)[1]
Motto in English	The wind of freedom blows[1]
Established	1891[2]
Type	Private
Endowment	US $13.8 billion[3]
President	John L. Hennessy
Provost	John Etchemendy
Academic staff	1,910[4]
Students	15,319
Undergraduates	6,878[5]
Postgraduates	8,441[5]
Location	Stanford, California, United States
Campus	Suburban, 8,180 acres (33.1 km²)[6]
Athletic nickname	Cardinal
Colors	Cardinal red and white

Subject	Predicate	Object
Stanford University	type	Private University
	hasPresident	J.L.Hennessy
	hasStudents	15,319
	foundedBy	L.Stanford
	foundedIn	1891
…	…	…

YAGO Knowledge Base

- Combine knowledge from **WordNet** & **Wikipedia**
- Additional **Gazetteers** (geonames.org)
- Part of the **Linked-Data** cloud



- **YAGO2**: >30 Mio RDF facts in base relations, >95% precision

Motivation (1/2)

Answers:
- bornIn(Bill,NY)

Incompleteness

Query:
q = ?← bornIn(Bill,?y)

Deduction Rules:
- r1: bornIn(?x,?y) ← marriedTo(?x,?z) ∧ bornIn(?z,?y)
- r2: bornIn(?x,?y) ← represents(?x,?y)

handcrafted or inductively learned rules

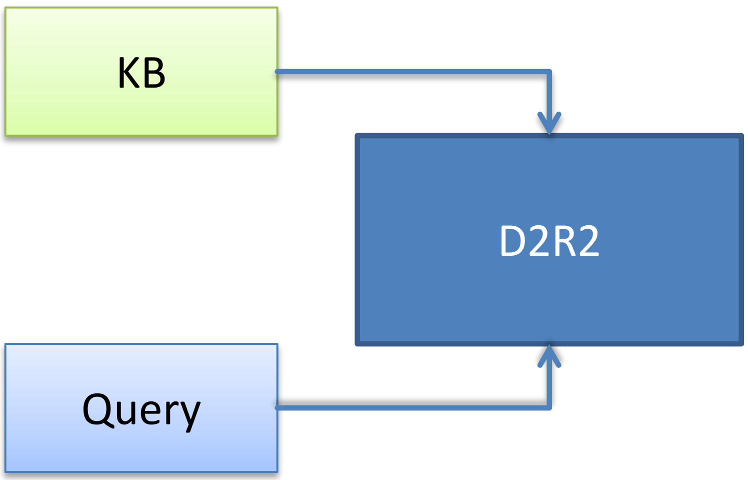
RDF Facts:
- f1: marriedTo(Bill,Hillary)
- f2: represents(Hillary,NY)

...

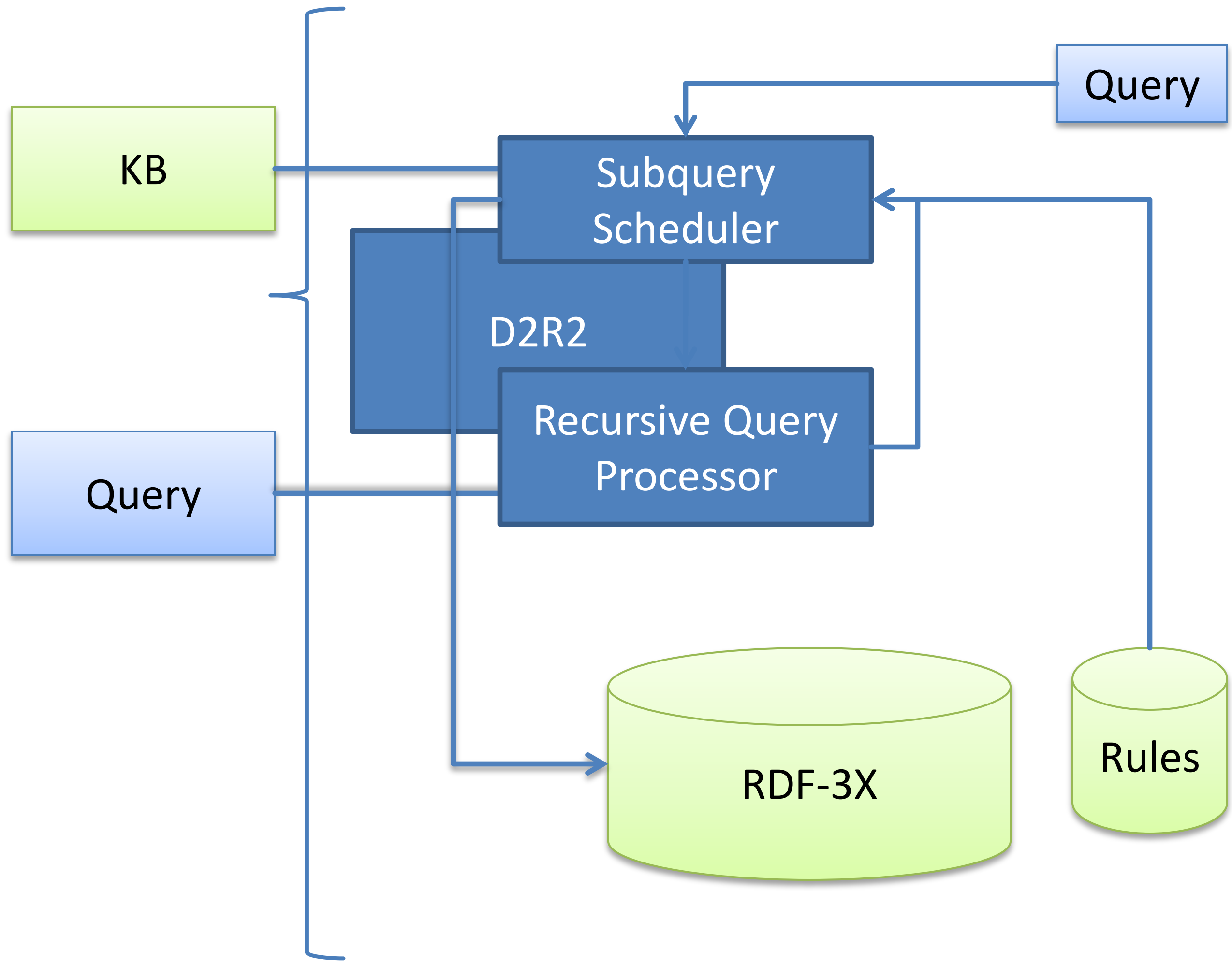
$>>10^7$ disk-resident facts

Knowledge Base

Motivation (2/2)

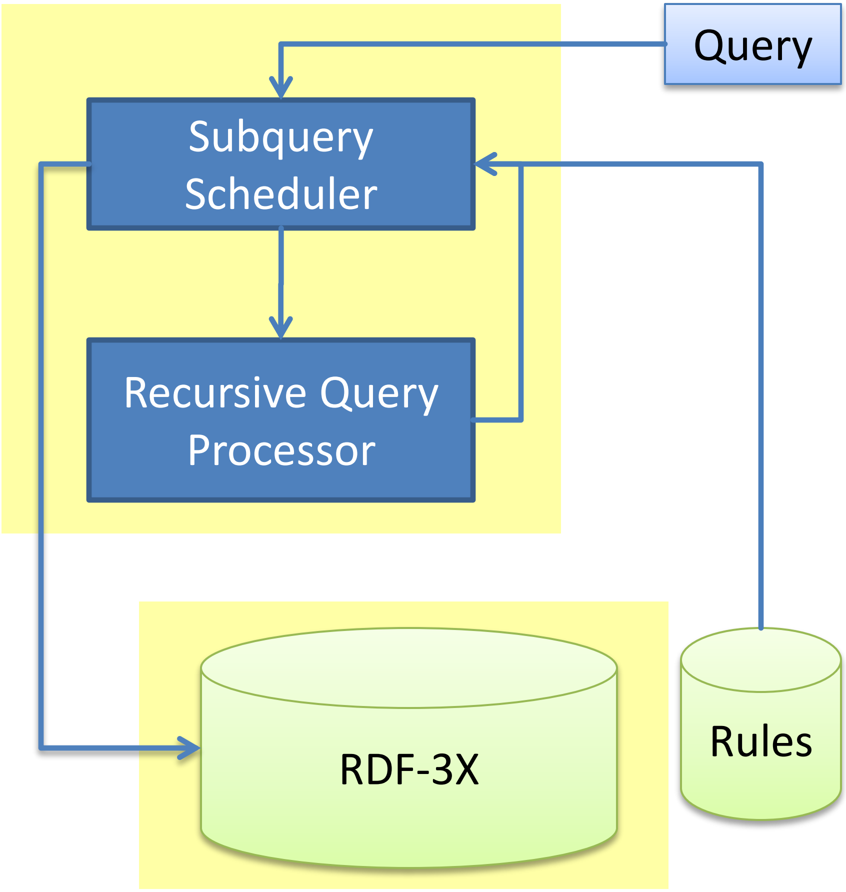


Motivation (2/2)



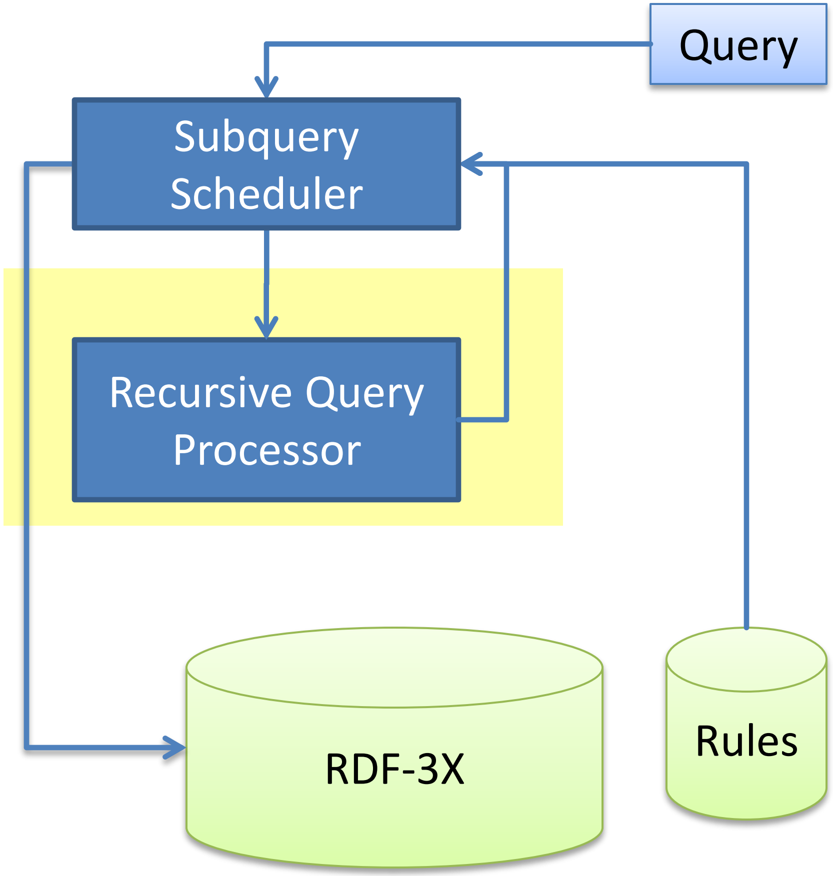
Outline

- Recursive query processing:
  - QSQR
  - Chaining & dynamic subquery scheduling
- Adding deductive reasoning to an RDF-engine
- Experiments



Outline

- Recursive query processing:
  - QSQR



QSQR

Algorithm 1: QSQR(D, q)

Input: A Datalog program D and an intensional query q
Input: The global *ans_* and *input_* relations
Output: All answers for q

```
begin
    Set all ans_ relations to be empty
    Set (R^γ, J) to be the generalized query corresponding to q
    repeat
        Set all input_ relations to be empty
        Call QSQR_EVAL_GENERALIZED(D, (R^γ, J))
    until Until no ans_ relation has changed in the last iteration;
    return All answers for q by performing a selection on ans_R^γ using J
```

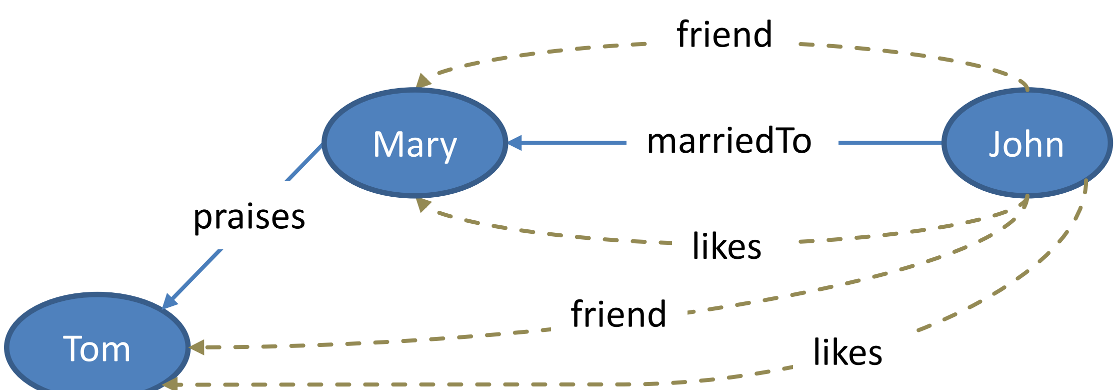
- Query(q):
  - Q = ? ← likes(John,?y)

[Abiteboul, Hull, Vianu: *Foundations of Databases*, 1995]

- Rules(D):
  - r_1: likes(?x,?y) ← friend(?x,?y)
  - r_2: friend(?x,?y) ← marriedTo(?x,?y)
  - r_3: friend(?x,?y) ← likes(?x,?z) ∧ praises(?z,?y)
- Facts(D):
  - f_1: marriedTo(John,Mary)
  - f_2: praises(Mary,Tom)

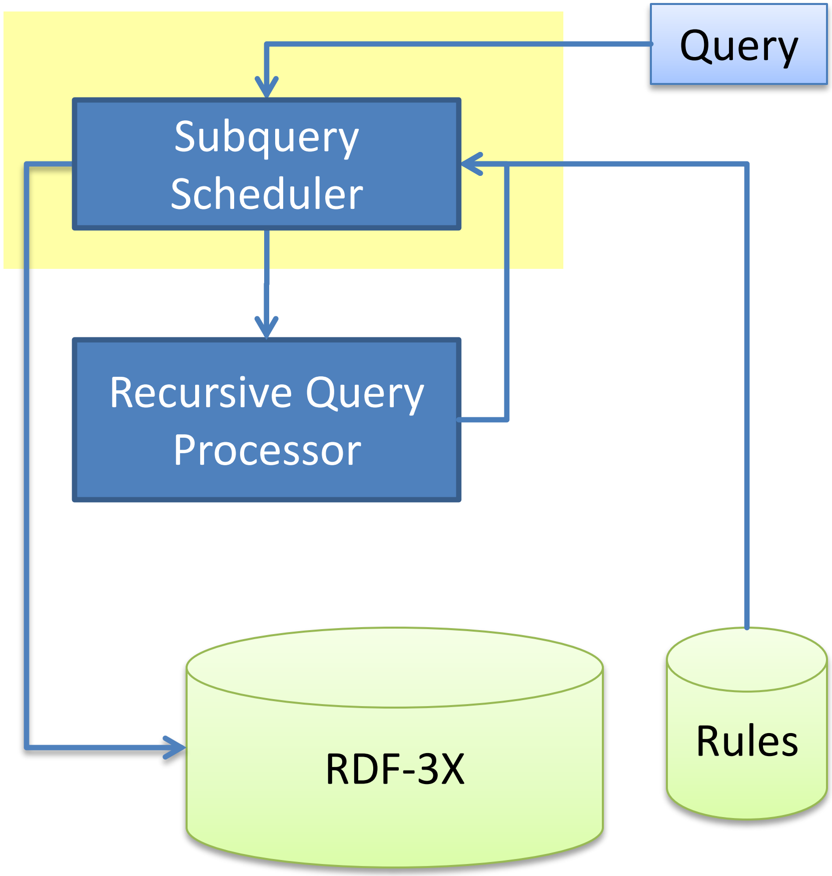
input_likes	input_friend
<John,?>	<John,?>

ans_friend	ans_likes
<John, Mary>	<John, Mary>
<John, Tom>	<John, Tom>



Outline

- Recursive query processing:
  - QSQR
  - Chaining & dynamic subquery scheduling



QSQR + Extensions

1. **Chaining**: make use of the underlying engine's index structures & ability to perform joins

2. **Dynamic scheduling**: next set of atoms to be evaluated in a subquery is determined after current set is evaluated

QSQR + Extensions – Chaining

$I_1(?x,?y) \leftarrow I_2(?x,?y) \wedge E_1(?x,?z) \wedge E_2(?y,?z)$

Two ways to evaluate: $? \leftarrow I_1(a,?y)$

1. Naïve
   a. $E_1(a,?z)$ get bindings for ?z
   b. $E_2(?y,?z)$ for every ?z
   c. $I_2(a,?y)$ for every ?y

Nested Loop Joins

1. Grouping: better utilization of query optimizer
   a. $E_1(a,?z) \wedge E_2(?y,?z)$
   b. $I_2(a,?y)$ for every ?y

More choices:
(SMJ , HJ, NLJ)
Merging, Hashing

QSQR + Extensions – Dynamic Scheduling

$I_1(?x,?y) \leftarrow I_2(?x,?y) \wedge E_1(?x,?z) \wedge E_2(?y,?z)$

Just because the rule places I_2 first, doesn't mean it has to be evaluated first:

Declarative paradigm: <u>What</u> vs. <u>How</u>

Extend QSQR to support *dynamically* deciding which (set of) literals to evaluate next.

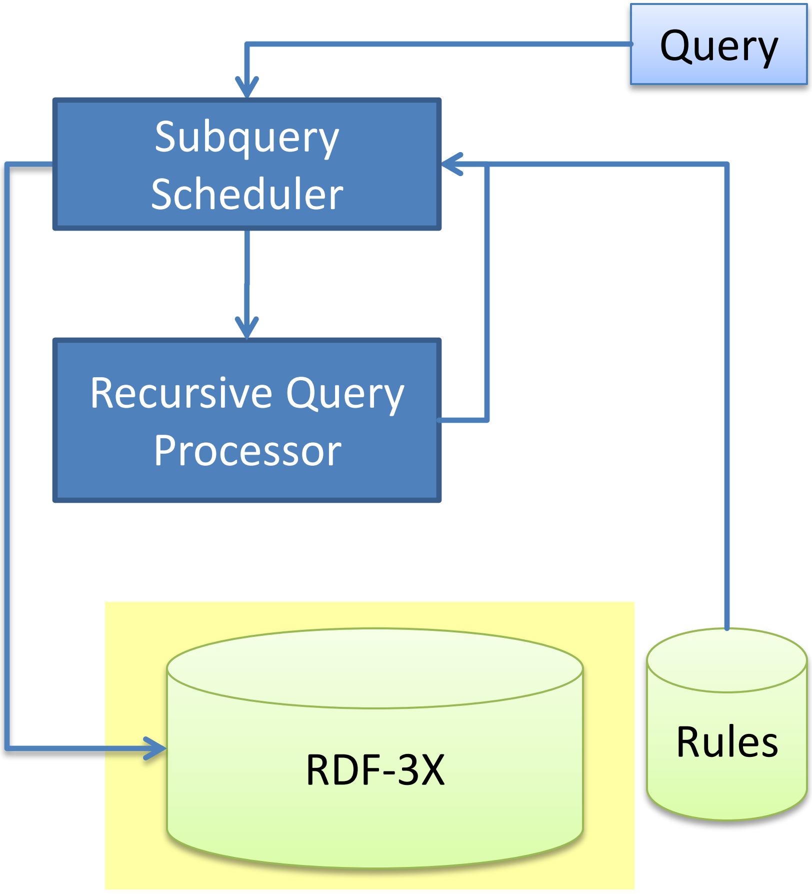
Algorithm 3: QSQR_EVAL_RULE(D, φ, gq)

Input: A Datalog program D, a rule φ and a generalized query $gq = (R^\gamma, T)$
Input: The global $ans\_$ and $input\_$ relations

begin
- Remove from T all tuples that do not unify with the head of the rule
- Initialize sup from T
- $SQ_{remaining} = \varphi.body$
- $SQ_{current} = \phi$
- **while** $SQ_{remaining} \neq \phi$ **do**
  - $(SQ_{current}, SQ_{remaining})$ = SELECTION_FUNCTION($SQ_{remaining}, \ldots$)
  - $sup = sup \bowtie$ EVALUATE($SQ_{current}$, sup)
  - PROJECT(sup)
- Add tuples produced for sup into the global variable ans_R^γ

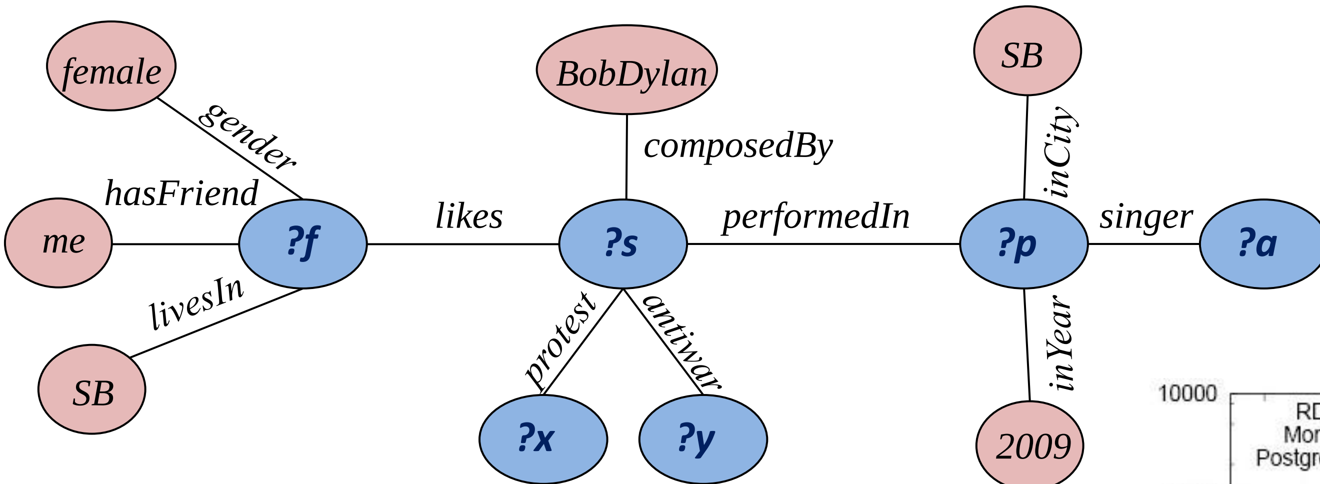
Outline

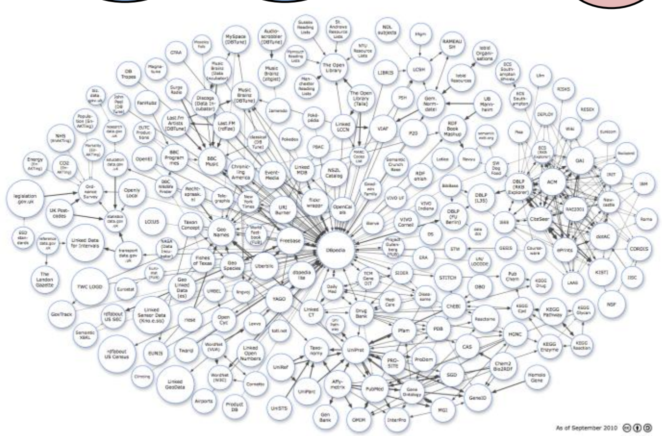
- Recursive query processing:
  - QSQR
  - Chaining & dynamic subquery scheduling
- Adding deductive reasoning to an RDF-engine

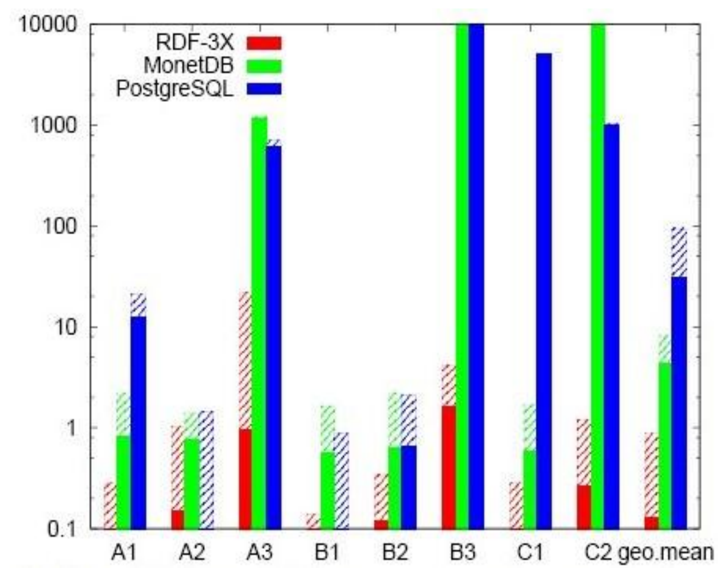


RDF-3X Engine [T. Neumann et al.: VLDB'08, SIGMOD'09]

- No-tuning RISC, versioning, online updates, transactions
- Aggressive indexing (15 SPO permutations & projections)
- Fast DP-based join-order optimization for up to 20-30 joins!
- Aggressive sideways information passing

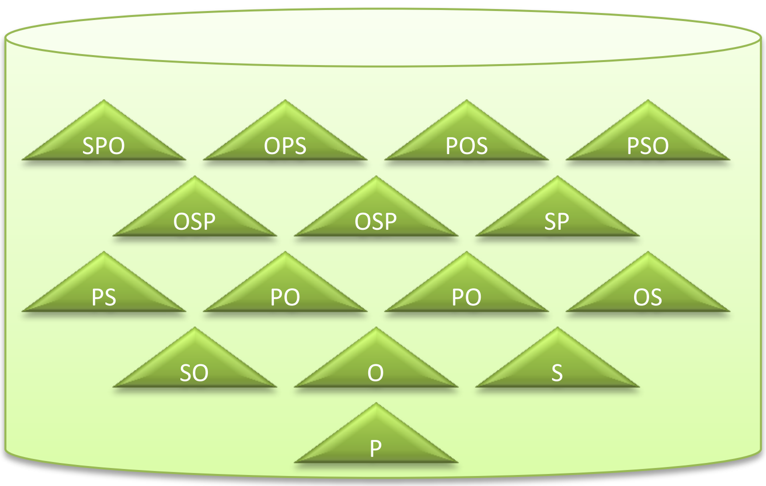






RDF-3X (1/2)

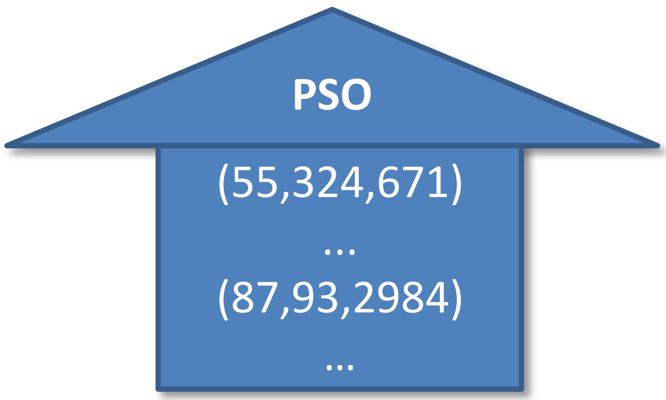
- 15 subsets & permutations of SPO attributes stored in exhaustive B^+-tree indexes:



- Disk-oriented statistics, relying on the indexes above.
- Index-only tables: B^+-tree leafs contain data & statistics.

RDF-3X (2/2)

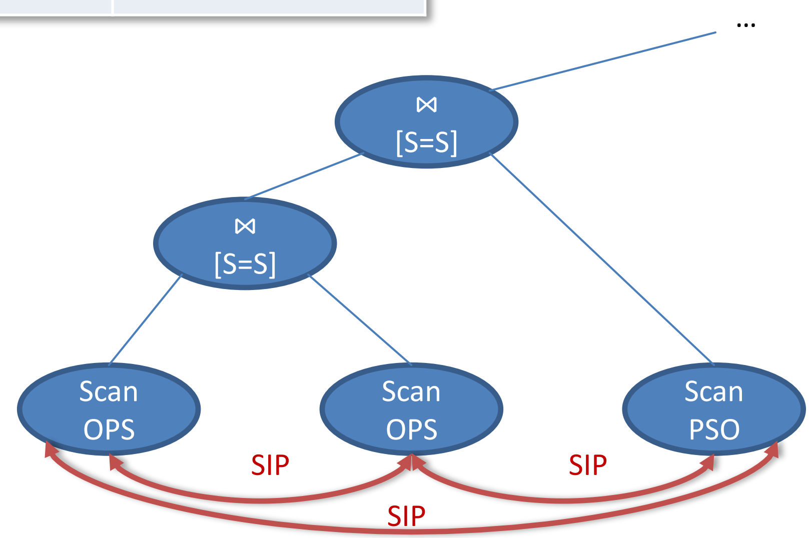
- Operators see integers:



Dictionary

String	ID
324	Hillary_Clinton
55	livesIn
671	New_York
93	Bill_Clinton
87	presidentOf
2984	USA

- Sideways information passing (SIP):
  - Operators communicate across query plan to skip pages.



RDF-3X Integration (1/2)

Our query patterns do not always match what RDF-3X expects.

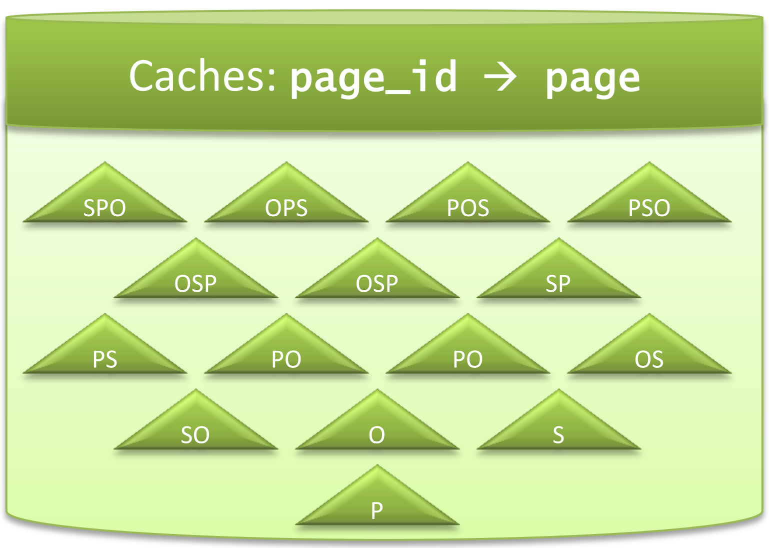
- **Small subqueries** (single literal/atom) are very common:
  - » Bypass RDF-3X query optimizer & employ own precompiled plans.

    $I_1(?x,?y) \leftarrow I_2(?x,?y) \wedge E_2(?y,\mathbf{?z}) \wedge I_3(?z,?y)$

- **Predicates** are always given in our context:
  - » Reduce the number of plans RDF-3X considers: only POS, PSO, PO and PS needed (others still used for statistics).

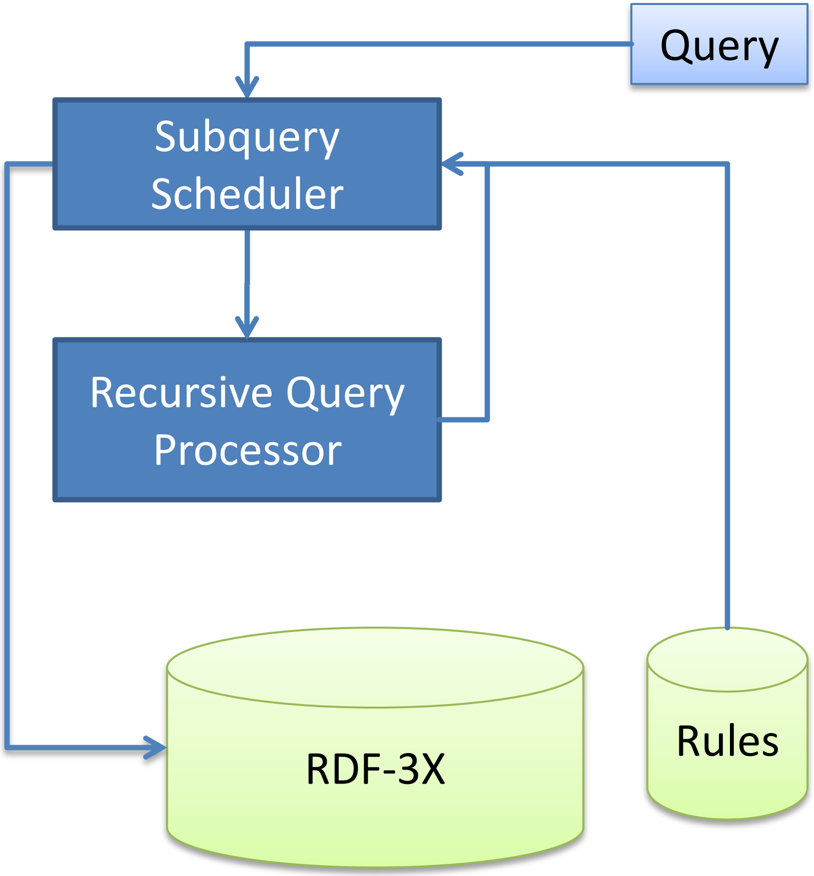
RDF-3X Integration (2/2)

- During **recursion**, same pages can be requested multiple times:
  - » Add caching on top of compressed indexes.



Outline

- Recursive query processing:
  - QSQR
  - Chaining & dynamic subquery scheduling
- Adding deductive reasoning to an RDF-engine
- Experiments



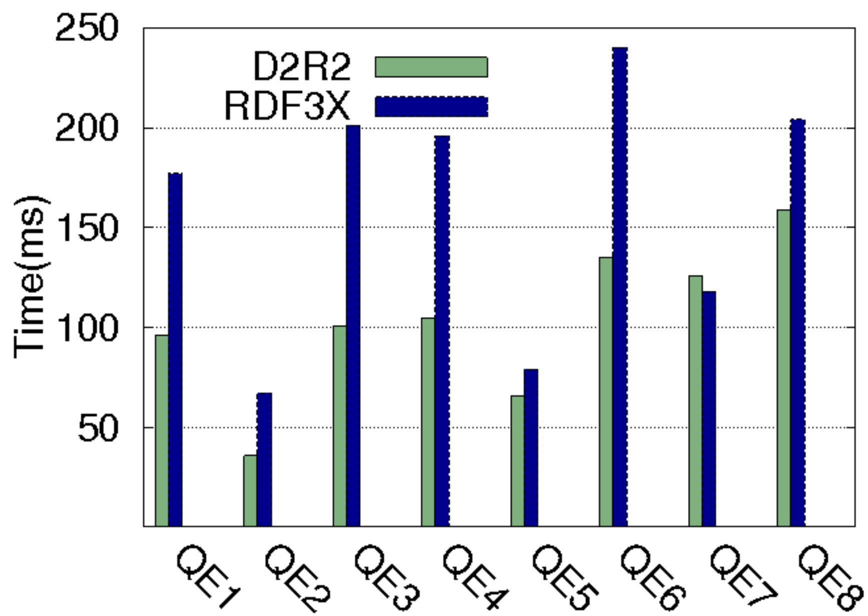
Experiments (1/5)

- Datasets:
  - **YAGO1** – 22 Mio RDF facts (3.8 GB)
    → 16 partly recursive rules
  - **LUBM** – Scale factor 1: 100k RDF facts (5.7 MB)
    → single recursive rule (*subOrganizationOf*)

- D2R2 implemented in C++ on top of RDF-3X

- Other systems:
  - **IRIS** Reasoner (incl. PostgreSQL 8.4 backend)
  - **Jena** Semantic Web Framework (incl. TDB 0.8.7 backend)

Experiments (2/5)

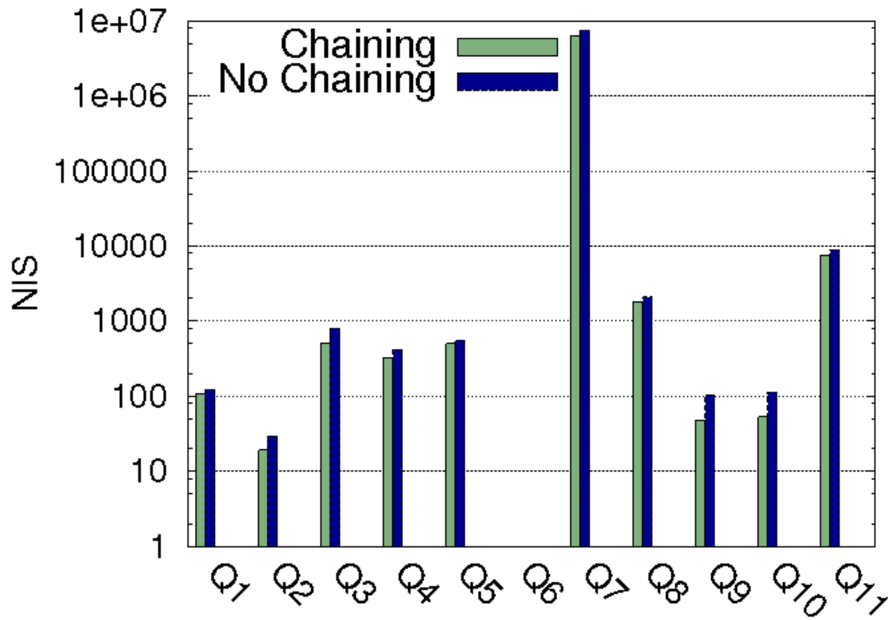


YAGO setting, no rules

QE1 = ? ← bornIn(Tipper Gore; ?y);
QE3 = ? ← isMarriedTo(Al Gore; ?y); bornIn(?y; ?z);
QE6 = ? ← actedIn(Schwarzenegger; ?y); actedIn(?x; ?y); bornIn(?x; ?z);

Experiments (3/5)

NIS: Number of Intermediate Subqueries



YAGO setting, recursive rules

R1: bornIn(?x; ?z) ← isCitizenOf(?x; ?y); locatedIn(?z; ?y); livesIn(?x; ?z);

R2: bornIn(?c; ?y) ← livesIn(?x; ?y); livesIn(?z; ?y); isMarriedTo(?x; ?z); hasChild(?x; ?c); hasChild(?z; ?c);

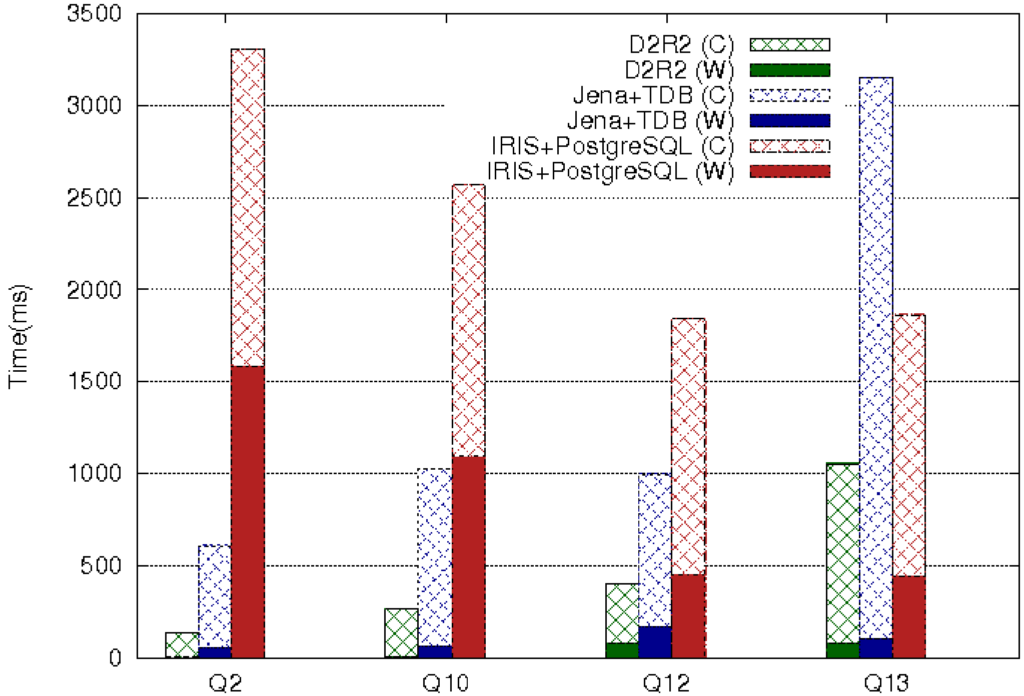
R3: livesIn(?y; ?z) ← isMarriedT o(?x; ?y); livesIn(?x; ?z);

R4: isMarriedTo(?x; ?y) ← hasChild(?x; ?z); hasChild(?y; ?z); notEquals(?x; ?y);

…

Q1 = ? ← bornIn(Al Gore; ?x);

Experiments (4/5)



YAGO setting, single recursive rule (*ancestor*)

Q2: ? ← isMarriedTo(Woody Allen; ?x);

Q10: ? ← directed(Martin Scorsese; ?x); actedIn(?y; ?x); actedIn(?z; ?x);notEquals(?y; ?z);

Q12: ? ← ancestor(?x; Charles; Prince of Wales);

Q13: ? ← ancestor(Edward IV of England; ?x);

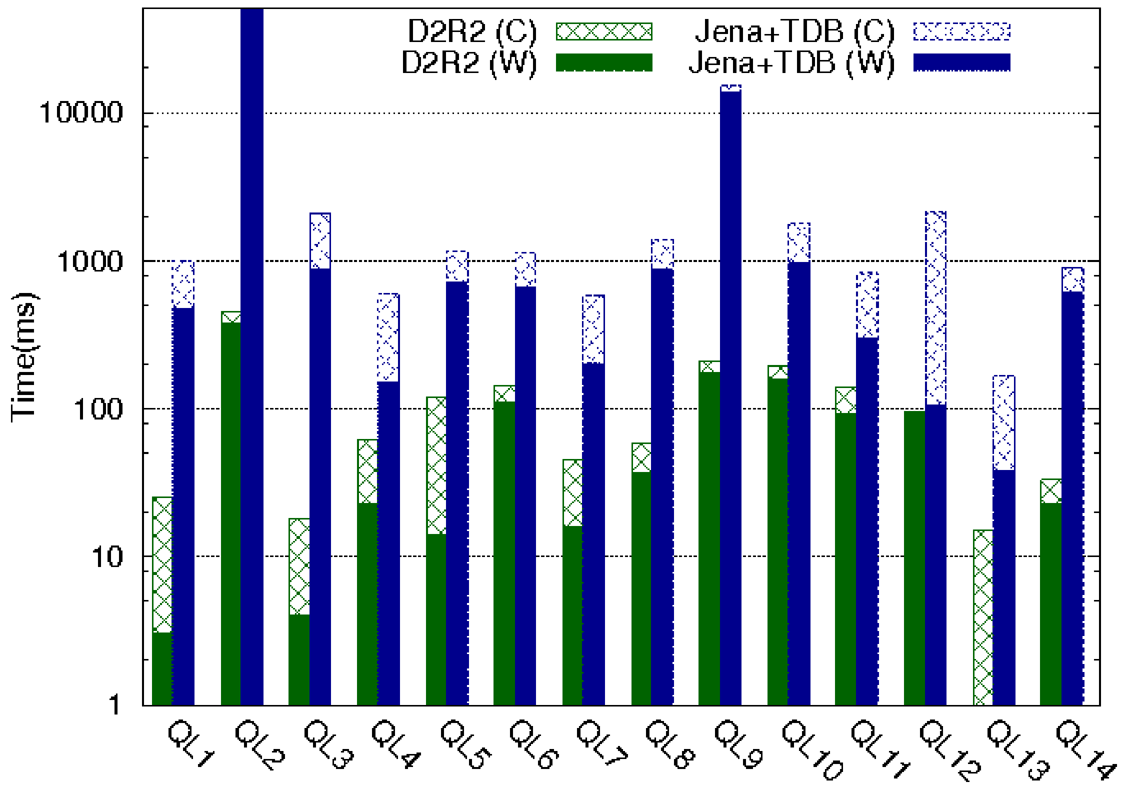
…

R1: <u>ancestor</u>(?x,?y) ← parent(?x,?y);

R2: <u>ancestor</u>(?x,?y) ← <u>ancestor</u>(?x,?z); parent(?z,?y);

…

Experiments (5/5)

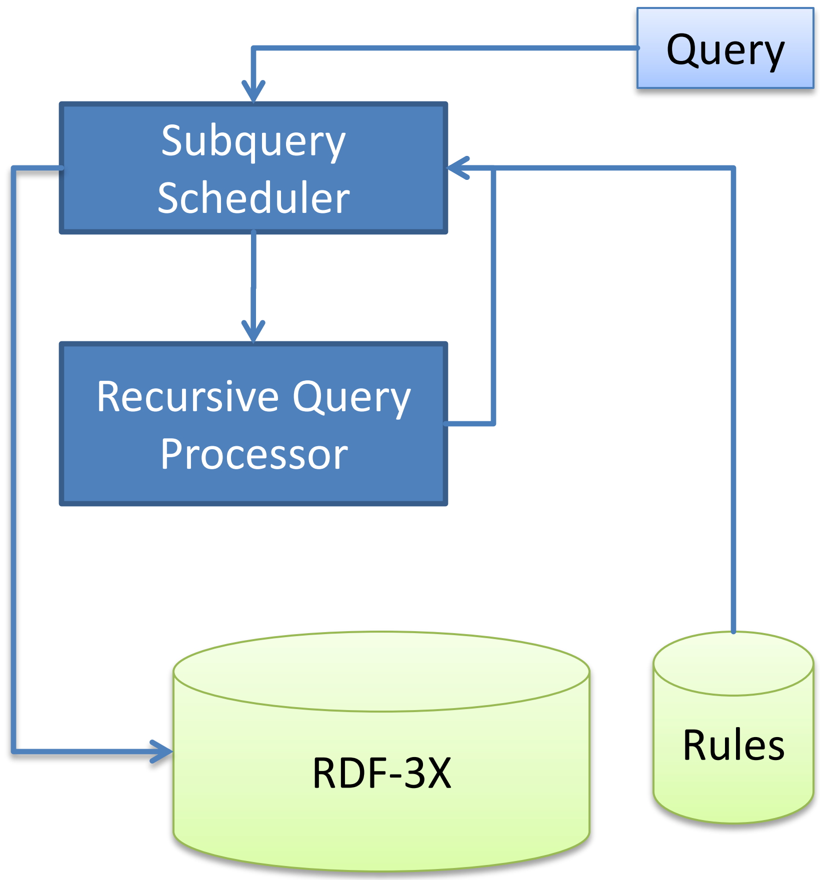


LUBM setting

- R1: type(?x; Professor) ← type(?x; Full Professor);
- R2: subOrganizationOf(?x; ?z) ← subOrganizationOf(?x; ?y); subOrganizationOf(?y; ?z);
- R3: hasAlumnus(?x; ?y) ← degreeF rom(?y; ?x);
- …
- QL1: ? ← type(?x; GraduateStudent); takesCourse(?x; http://www:Department0:University0:edu=GraduateCourse0);

Conclusions

- Setting
  - Large disk-resident RDF collections & deductive reasoning
- Recursive queries
  - QSQR with extensions: *chaining* & *dynamic scheduling*
- Facts store
  - RDF-3X with modifications for our setting
- <u>Current work</u>
  - Distributed reasoning
  - Lineage tracing & probabilistic inference ("soft rules")



Thank You!

Questions?