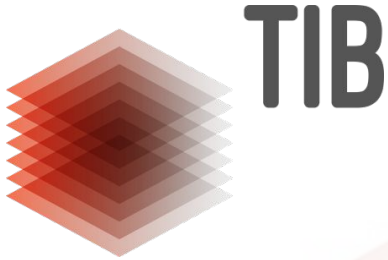
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



TIB

Challenges of Making Data Interoperable during Query Processing

Maria-Esther Vidal
Scientific Data Management Group TIB, Germany
Universidad Simón Bolívar, Venezuela

Leibniz Association

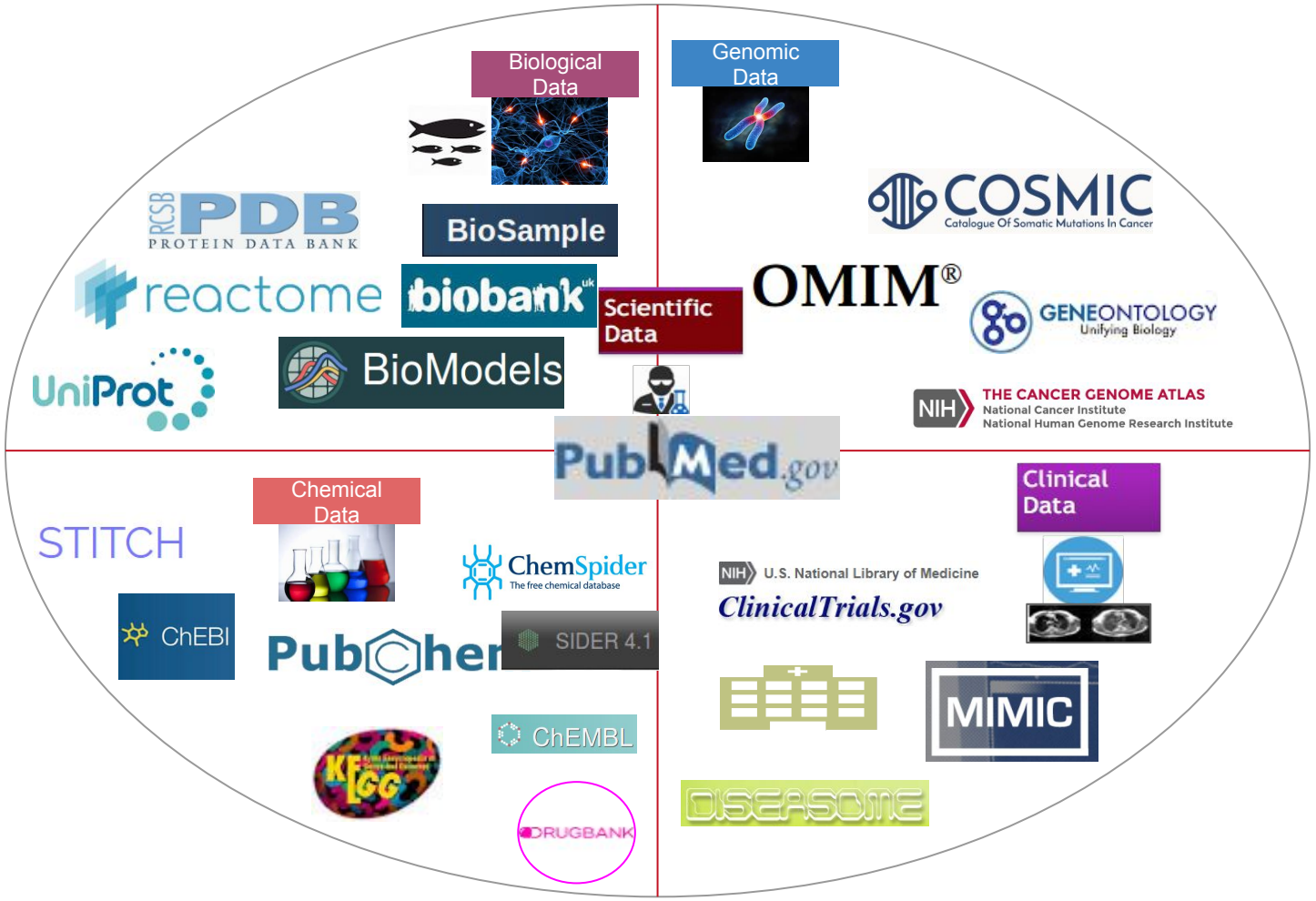
Motivating Example

Query: Drugs with the active substance ***Simvastatin:***

- Name of possible drug targets,
- Chemical formula of a drug,
- Side effects, and
- Disease Name

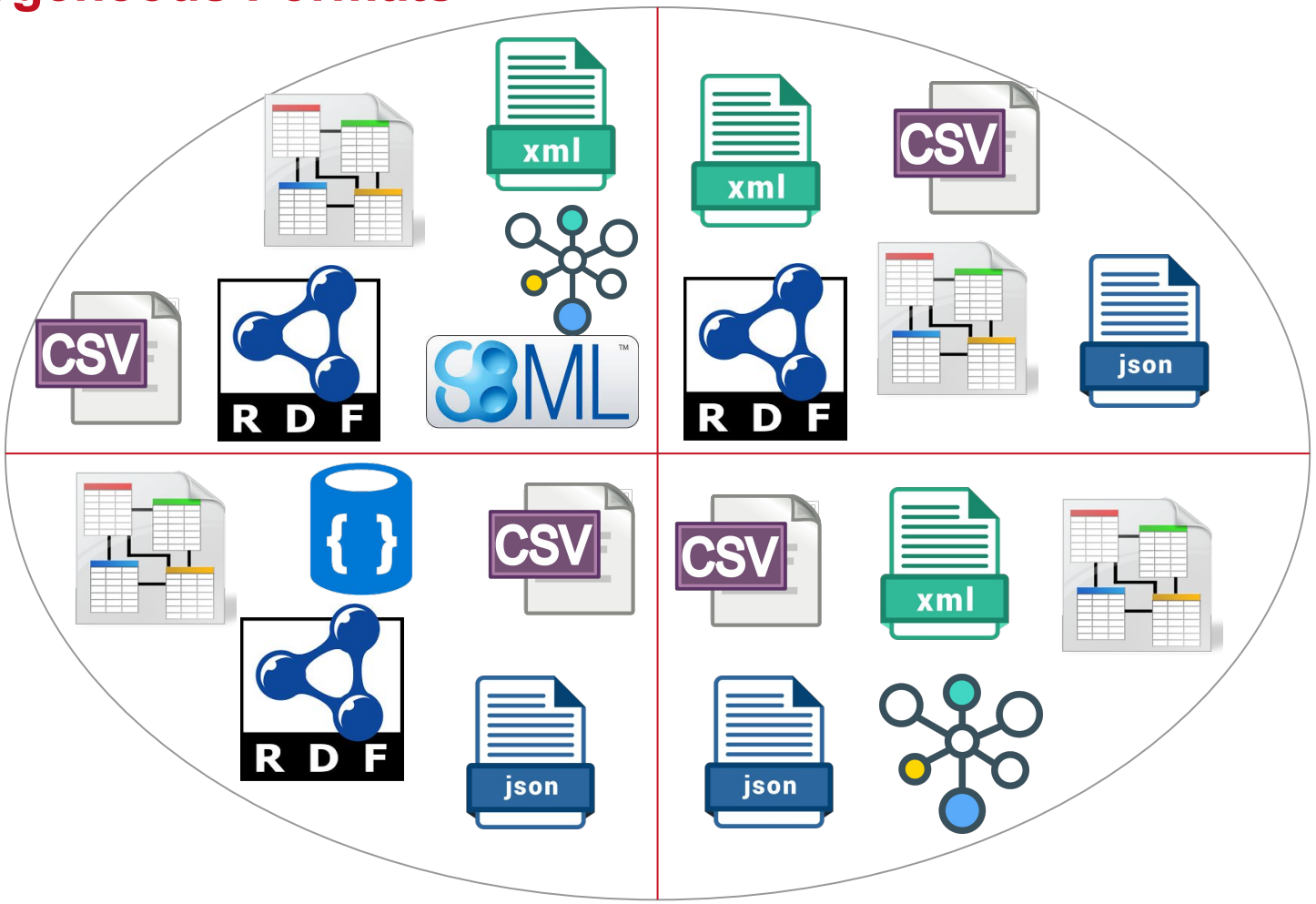
```
SELECT DISTINCT ?drug ?disName ?drugformula ?sename
WHERE {
  t1  ?drug     dailymed:activeIngredient        dailymed:Simvastatin .
  t2  ?drug     dailymed:genericDrug             ?dbdrug .
  t3  ?drug     dailymed:possibleDiseaseTarget   ?disease .
  t4  ?drug     owl:sameAs                       ?sadrug .
  t5  ?disease  rdfs:label                       ?disName
  t6  ?sadrug   sider:sideEffect                 ?seffect .
  t7  ?seffect  sider:sideEffectName             ?sename .
  t8  ?dbdrug   drugbank:chemicalFormula         ?drugformula
}
```

Motivating Example- Available Data Sources



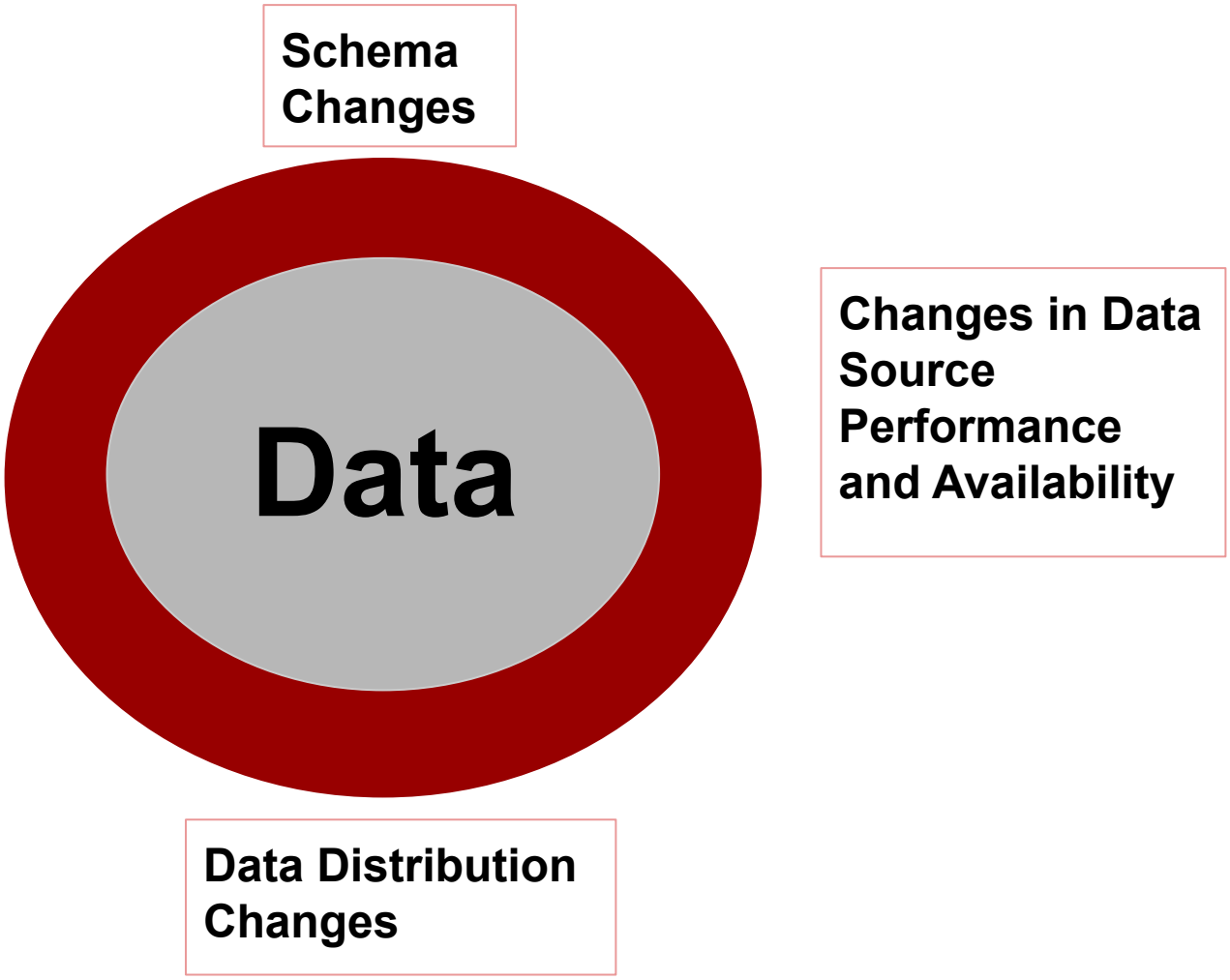
Diverse data sources potentially incomplete and noisy

Motivating Example- Data Sources in Heterogeneous Formats

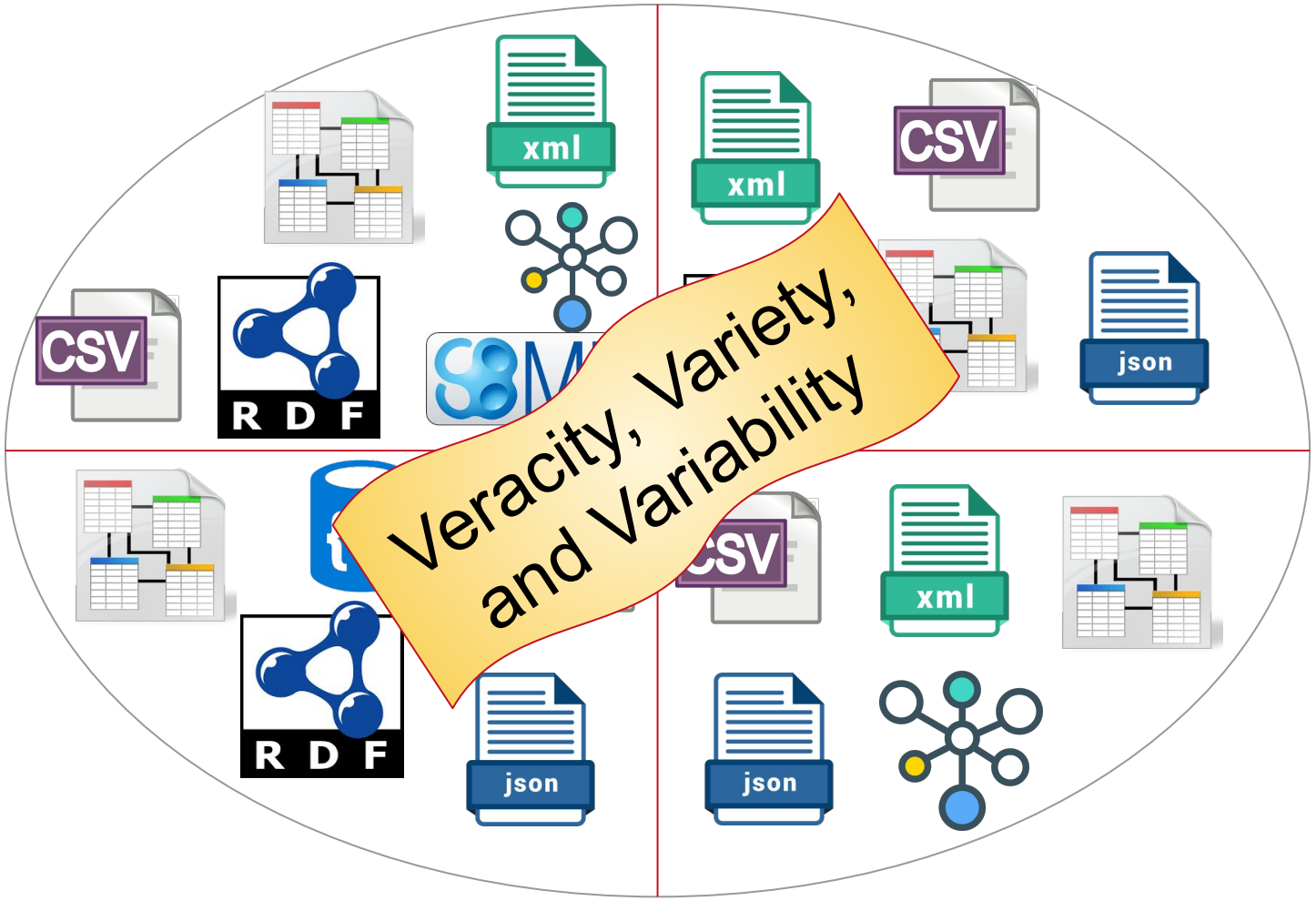


Data sources is diverse formats, e.g., XML, CSV, JSON

Data Evolution….



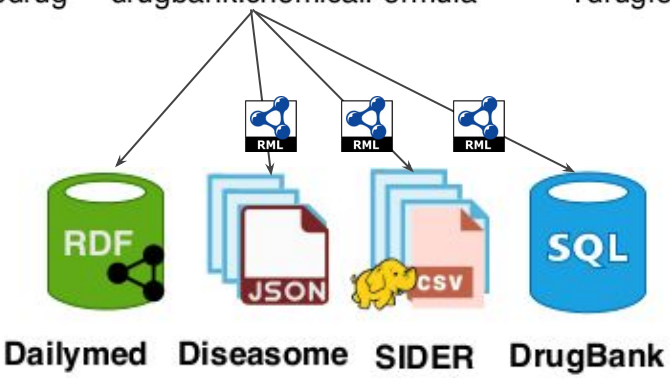
Impacting Data Complexity Dimensions



Query Over Heterogeneous Data Sources

- Query: Drugs with the active substance ***Simvastatin:***
  - Name of possible drug targets,
  - Chemical formula of a drug,
  - Side effects, and
  - Disease Name

```
SELECT DISTINCT ?drug ?disName ?drugformula ?sename
WHERE {
  t1  ?drug     dailymed:activeIngredient          dailymed:Simvastatin .
  t2  ?drug     dailymed:genericDrug               ?dbdrug .
  t3  ?drug     dailymed:possibleDiseaseTarget     ?disease .
  t4  ?drug     owl:sameAs                         ?sadrug .
  t5  ?disease  rdfs:label                         ?disName
  t6  ?sadrug   sider:sideEffect                   ?seffect .
  t7  ?seffect  sider:sideEffectName               ?sename .
  t8  ?dbdrug   drugbank:chemicalFormula           ?drugformula
}
```



Interoperability Issues During Query Processing

RDF DAILYMED

dailymed:798	rdf:type	dailymed:drugs ;
	dailymed:activeIngredient	dming:Simvastatin .
	owl:sameAs	sider:54454 .
	dailymed:genericDrug	drugbank:DB00641 ;
	dailymed:possibleDiseaseTarget	diseasome:319,
		diseasome:2839,
		diseasome:2175 .

JSON DISEASOME

```
[{
  "diseaseID": " 319",
  "name": "Diabetes_mellitus",
  "associatedGene": ["ACE", "ABCC8", "TCF1"]
  },{
  "diseaseID": " 2839",
  "name": "Kaposi sarcoma, susceptibility to,
148000",
  "associatedGene": ["IL6", "IFNB2", "BSF2"]
}]
```

SQL DRUGBANK

Drug

accNum	DrugName	formula	pubChemId
DB00641	simvastatin	$C_{25}H_{38}O_5$	54454
DB00295	Morphine	$C_{17}H_{19}NO_3$	5288826

Drug_Target

Drug	Target
DB00641	631
DB00641	1882
DB00295	7683

Target

ID	Name	Gene	UniprotID
631	3-hydroxy-3-methylglutaryl-coenzyme A reductase	HMGCR	P04035
1882	Ras-related C3 botulinum toxin substrate 1	RAC1	P63000
7683	Mu-type opioid receptor	OPRM1	P35372

csv SIDER 4.1

side_effects.csv

DrugID,UMLS_ID,SideEffectName
54454,C0009806,Constipation
54454,C0236071,Throat tightness
54454,C0156404,Menstruation irregular
191,C0012833,Dizziness
191,C0232487, Abdominal discomfort
191,C1956346,Coronary artery disease

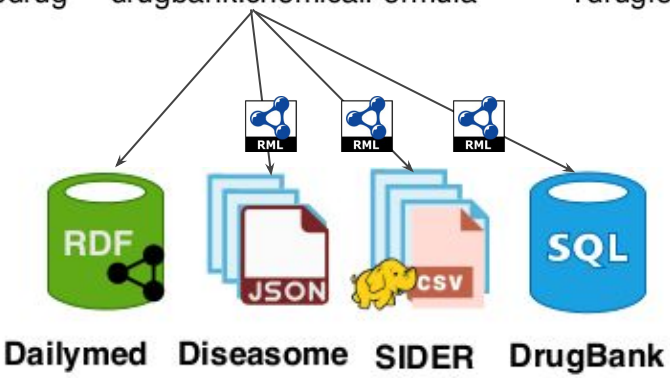
drug_names.csv

ID,DrugName
54454,simvastatin
191,adenosine

Query Over Heterogeneous Data Sources

- Query: Drugs with the active substance ***Simvastatin***:
  - Name of possible drug targets,
  - Chemical formula of a drug,
  - Side effects, and
  - Disease Name

```
SELECT DISTINCT ?drug ?disName ?drugformula ?sename
WHERE {
  t1  ?drug     dailymed:activeIngredient        dailymed:Simvastatin .
  t2  ?drug     dailymed:genericDrug             ?dbdrug .
  t3  ?drug     dailymed:possibleDiseaseTarget   ?disease .
  t4  ?drug     owl:sameAs                       ?sadrug .
  t5  ?disease  rdfs:label                       ?disName
  t6  ?sadrug   sider:sideEffect                 ?seffect .
  t7  ?seffect  sider:sideEffectName             ?sename .
  t8  ?dbdrug   drugbank:chemicalFormula         ?drugformula
}
```



Query must be evaluated against **heterogeneous sources**, that potentially suffer of **quality issues**, and **evolve** over time

Agenda

1. Data Integration Systems
2. Adaptive SPARQL Query Engines
3. Hybrid SPARQL Query Engines

Leibniz Association

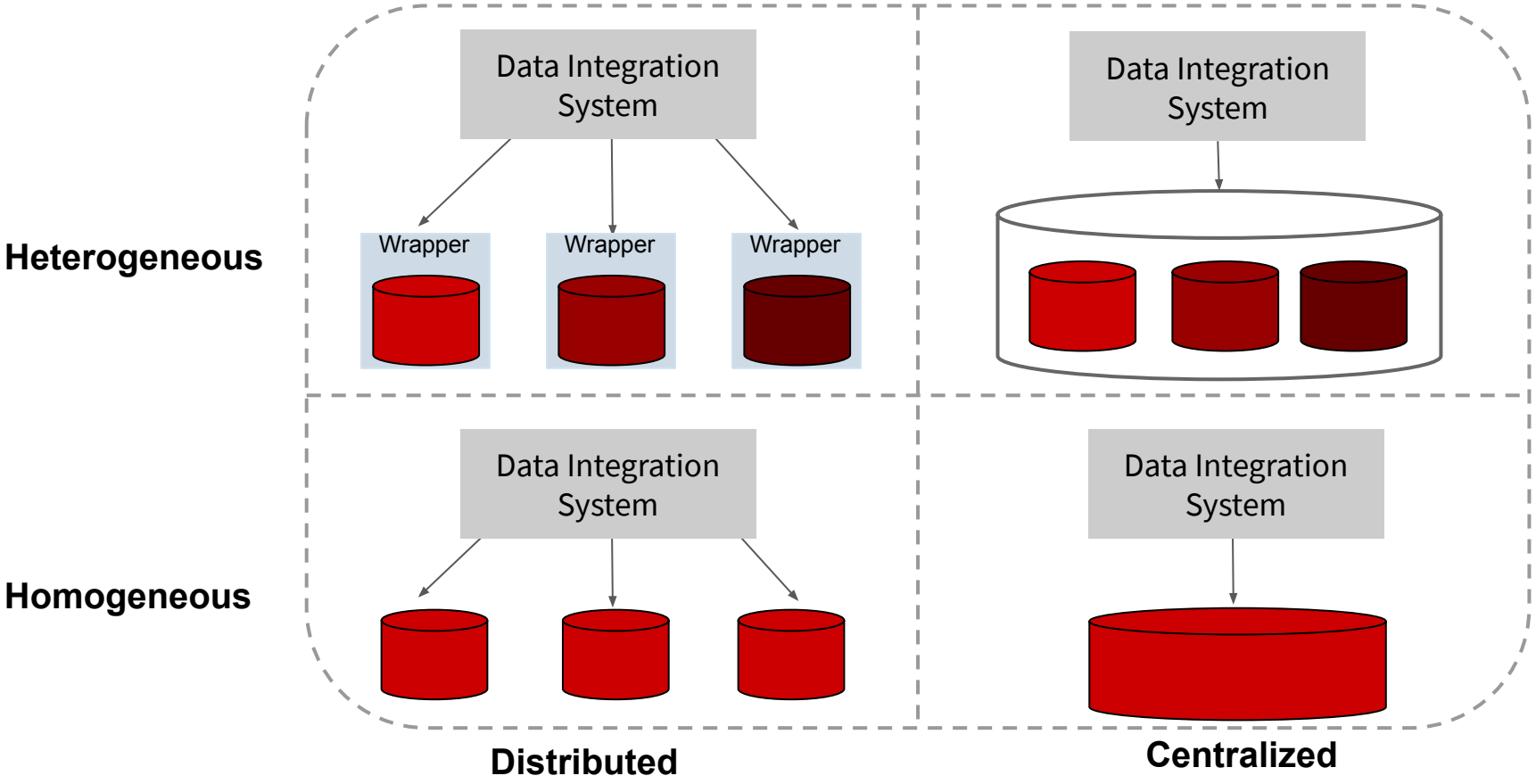
Data Integration Systems

A data integration system **DIS**=<**O**,**S**,**M**>:

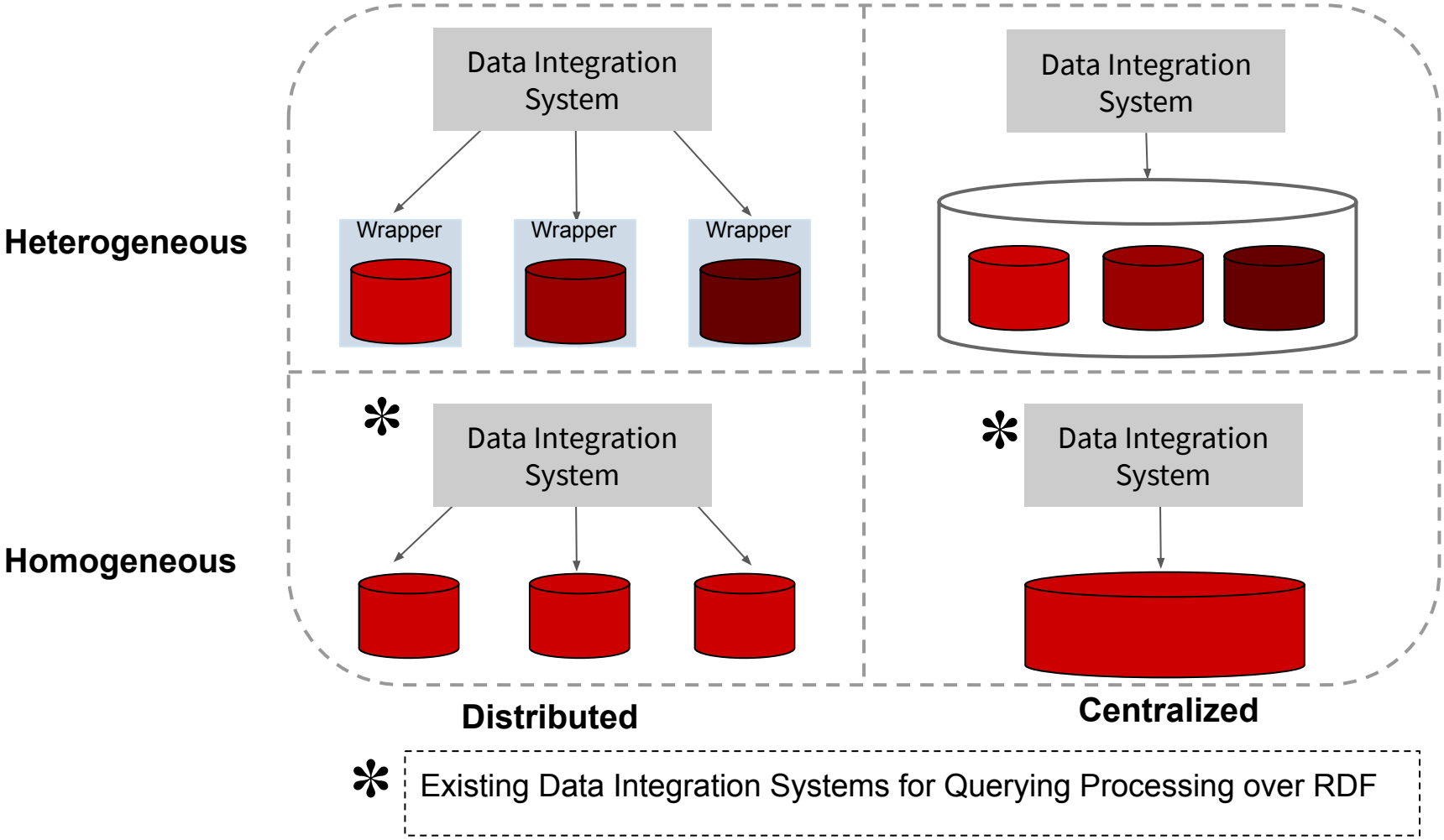
- **O** is a set of general concepts in a general schema (virtual)
- **S** is a set of {**S1**,..,**Sn**} of data sources
- **M** is a set of mappings between sources in **S** and general concepts in **O**

cf. Lenzerini 2002

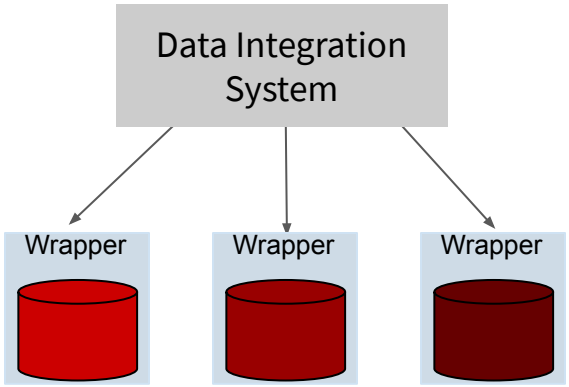
Data Integration Systems



Data Integration Systems



Query Rewriting Problem



Query Rewriting Problem (QRP):

- A query **Q** is a conjunctive query over predicates in **O**
- Find a conjunctive query **Q'** expressed in sources in **S** based on rules in **M**, such that
  - Evaluation of **Q**' produces only answers of **Q**
  - Evaluation of **Q**' produces all the answers of **Q** given the sources in **S**

Theorem [Levy et al. 1995]
To check if there is a valid rewriting **Q'** of **Q** with at most the same number of goals as **Q** is an **NP-complete problem.**

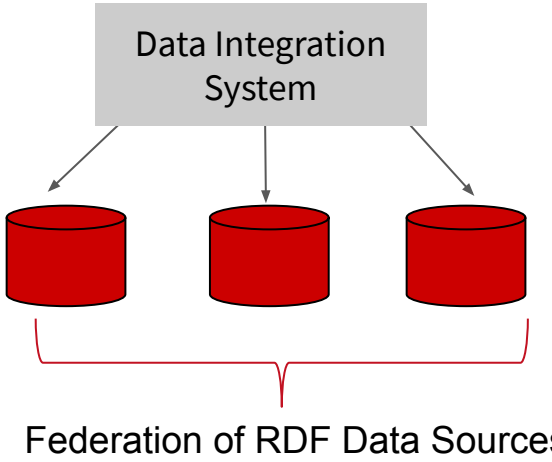
Challenges for Query Processing

Given a query Q in a formal language, i.e., SPARQL

- **Identify** the **relevant** data sources for Q (**Source Selection)**
- **Decompose** Q into subqueries on **relevant** data sources (**Query Decomposition**)
- **Plan** evaluation of **subqueries** against **relevant** data sources (**Query Planning**)
- **Merge** data collected from **relevant** data sources (**Query Execution**)

Relevant data sources for **Q**: **minimal set** of sources **S** from a federation of source **F** such that the answer of evaluating **Q** in **S is the same than** evaluating **Q** in **F**

Federated SPARQL Query Engines



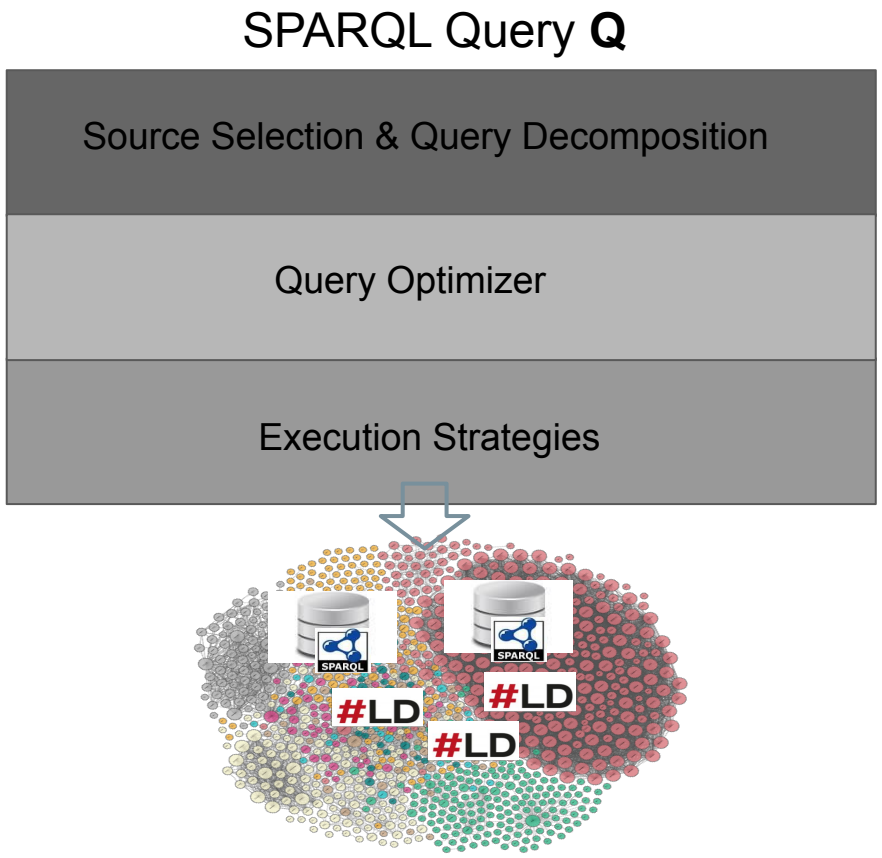
#LD

Web-access interfaces (**unpredictable** behavior) that allow for querying RDF data:

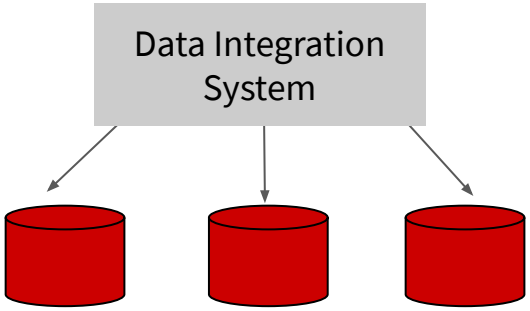
- SPARQL Endpoints: respect **SPARQL** protocol, i.e., **any** SPARQL query
- Linked Data Fragments: limited query capabilities, i.e., **only one** triple pattern

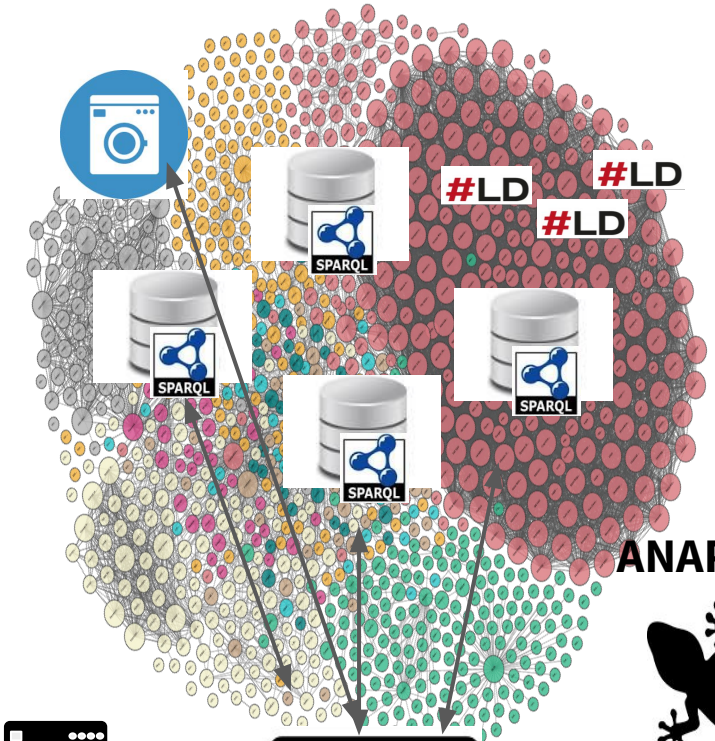
Challenges: Query processing is impacted by different parameters, e.g., **query capabilities, data** fragmentation, dataset **size** and **connectivity**, query **selectivity**, and **current conditions** of the Web-access interfaces

Federated Query Engine



Federated SPARQL Query Engines





Extensions

LILAC[5] FEDRA[6]

Fed-DESATUR[3]

DAW[9] MULDER[10]

HIBISCUS[15]

ANAPSID[1]

FedX[4] Linked Data in a Federation

SPLENDID [3]

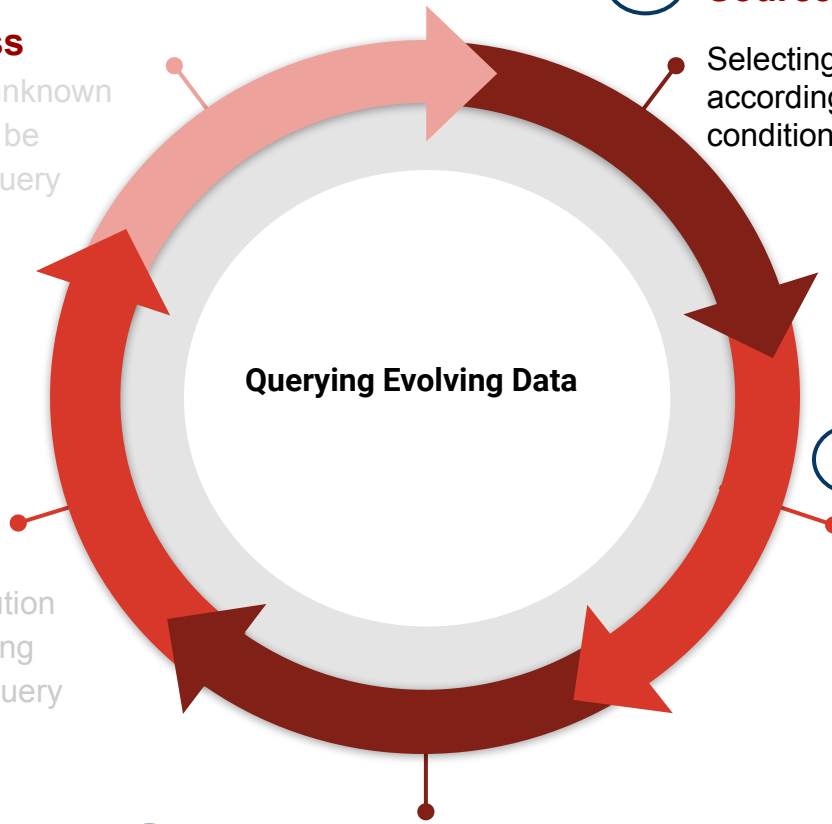
#LD [7]

Network of Linked Data Eddies (nLDE) **[2]**

Ontario [14]

SemaGrow [12]

Required Solutions to Support Evolution



1 **Source Evolution**

Selecting the sources according to their current conditions and availability

2 **Environment Evolution**

Executing queries according to current conditions of the environment

3 **Data Evolution**

Considering the status of the data, e.g., completeness, during the execution of the query

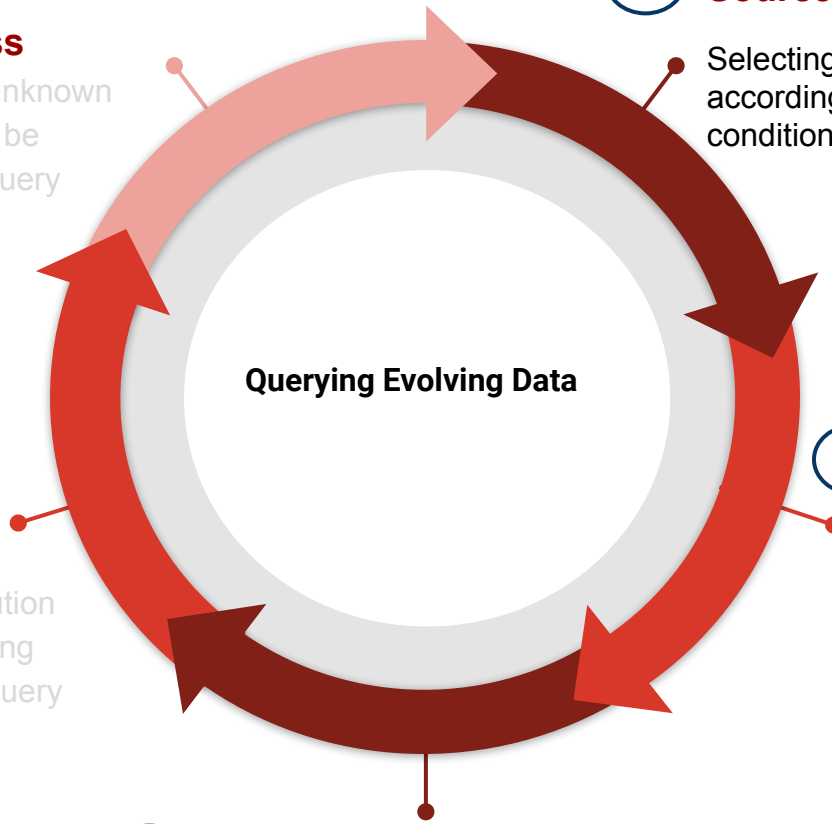
4 **Knowledge Evolution**

Considering the evolution of the knowledge during the execution of the query

5 **Knowledge Incompleteness**

Considering that unknown facts may need to be predicted during query execution

Required Solutions to Support Evolution



1 **Source Evolution**

Selecting the sources according to their current conditions and availability

2 **Environment Evolution**

Executing queries according to current conditions of the environment

3 **Data Evolution**

Considering the status of the data, e.g., completeness, during the execution of the query

4 **Knowledge Evolution**

Considering the evolution of the knowledge during the execution of the query

5 **Knowledge Incompleteness**

Considering that unknown facts may need to be predicted during query execution

Adaptive SPARQL Query Engines

Adapt to Source and Environment Evolution:

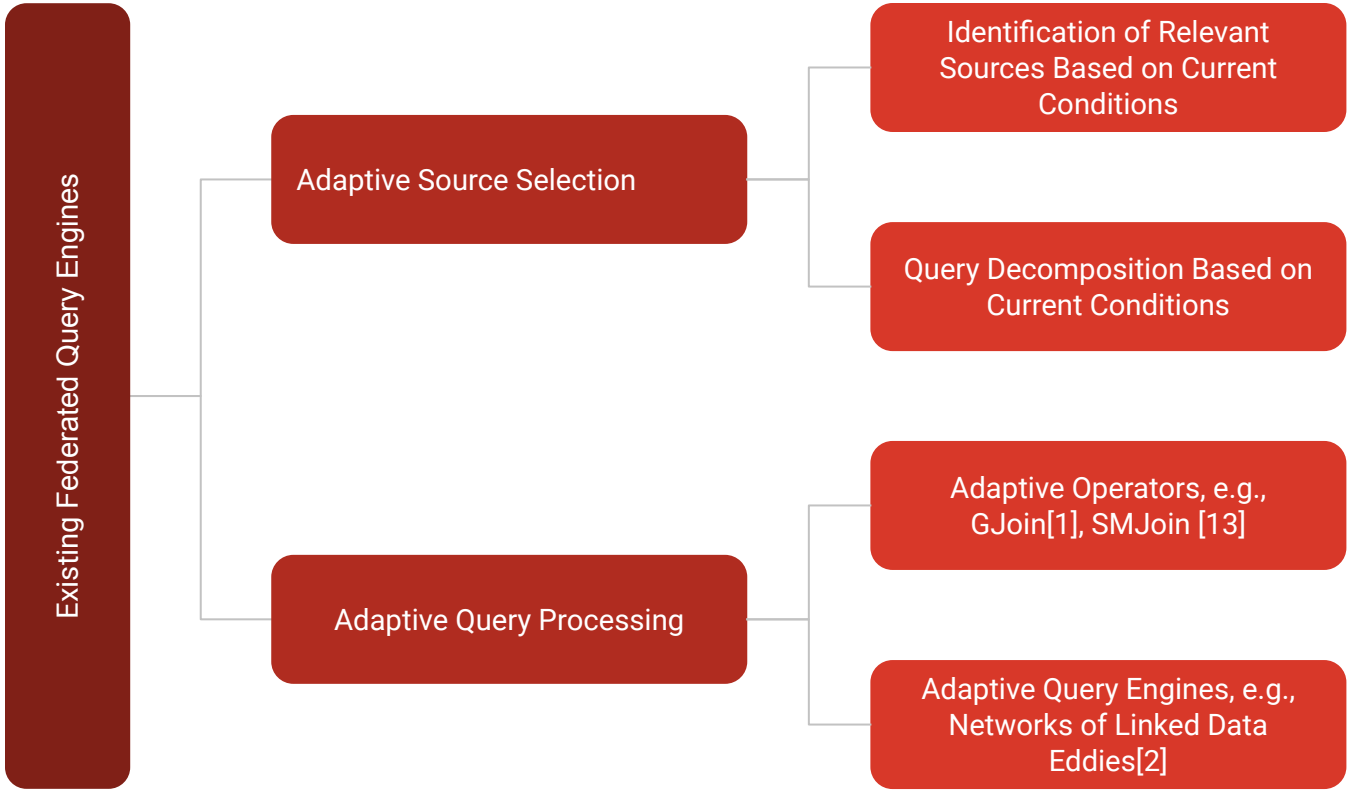
- **Misestimated** or **missing** statistics.
- **Unexpected** correlations.
- **Unpredictable** costs.
- **Dynamically** changing **data**, **workload,** and source **availability**.
- **Changes** at **rates** at which tuples **arrive** from sources
  - **Initial** Delays.
  - **Slow** Delivery.
  - **Bursty** Arrivals.

Adaptivity in Federated Query Processing

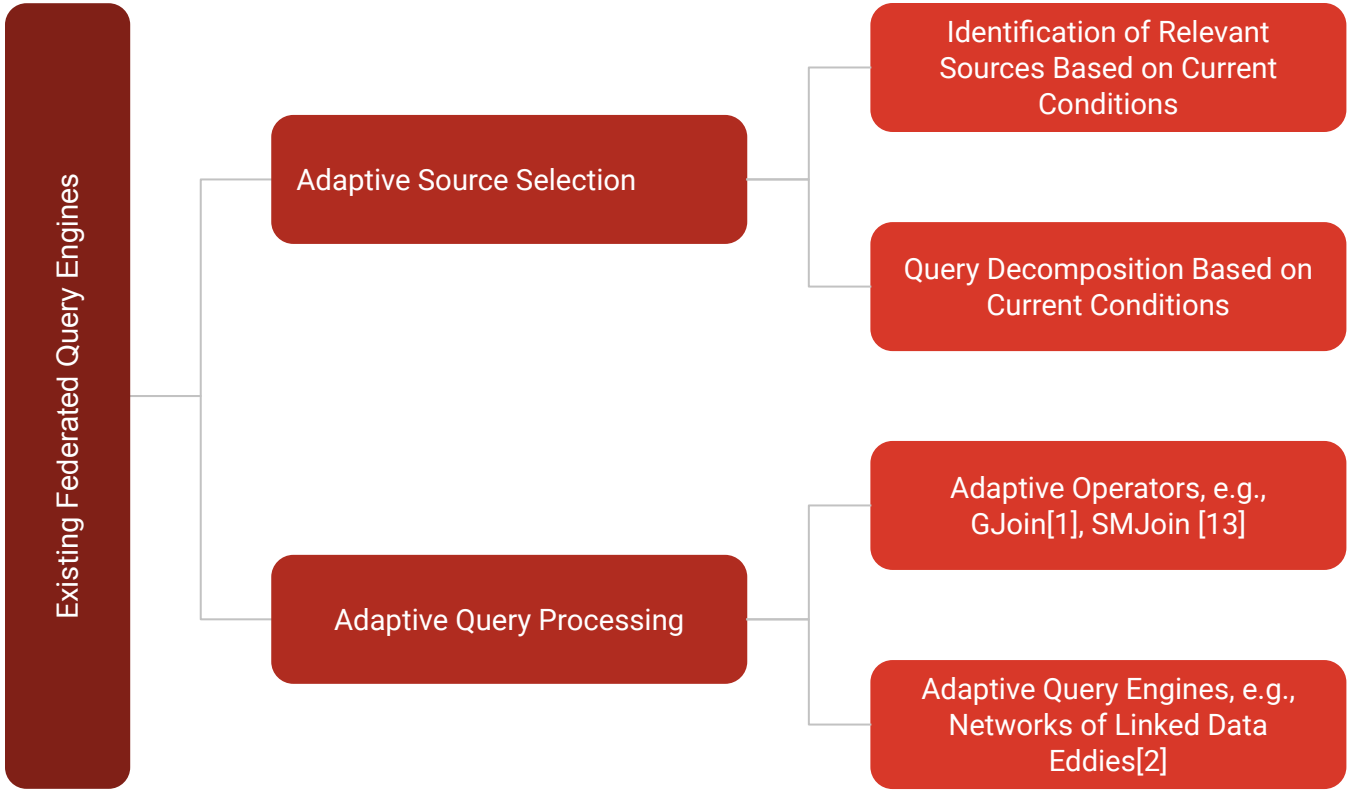
Query Engines able to:

- Change their behavior by learning the behavior of data providers
- Receive and exploit information from the environment
- Use up-to-date information to change their behavior
- Keep iterating over time to adapt their behavior based on the environment conditions

Existing Federated SPARQL Query Engines



Existing Federated SPARQL Query Engines



Only adaptivity to changes in the environment is addressed!!

Adaptivity During Source Selection

No Adaptivity

Coarse-Grained Adaptivity

Fine-Grained Adaptivity

LILAC FEDRA

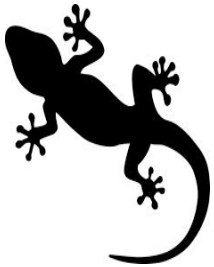
Fed-DESATUR

DAW MULDER

HIBISCUS

SemaGrow

ANAPSID SPLENDID



FedX
Linked Data in a Federation

Source Selection techniques that allow for identifying the sources that can be used to answer a query based on the current conditions of the sources

Adaptivity During Query Execution

No Adaptivity

Fine-Grained Adaptivity

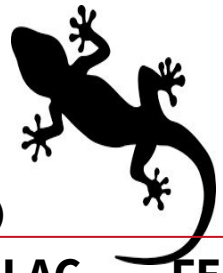
FedX
Linked Data in a Federation

SemaGrow
SPLENDID

ANAPSID

DAW HIBISCUS

Network of Linked Data Eddies (nLDE)



LILAC FEDRA

MULDER

Fed-DESATUR

Implement physical operators and query processing techniques to adjust query schedulers to the conditions of the sources and the network

Evaluation

Dataset: DBpedia 2015 (HDT on top of TPF server), 837M triples

Benchmark 1: **14** high-selective queries (<1000 int. res.)

Benchmark 2: **Four** low-selective queries (>1000 int. res.)

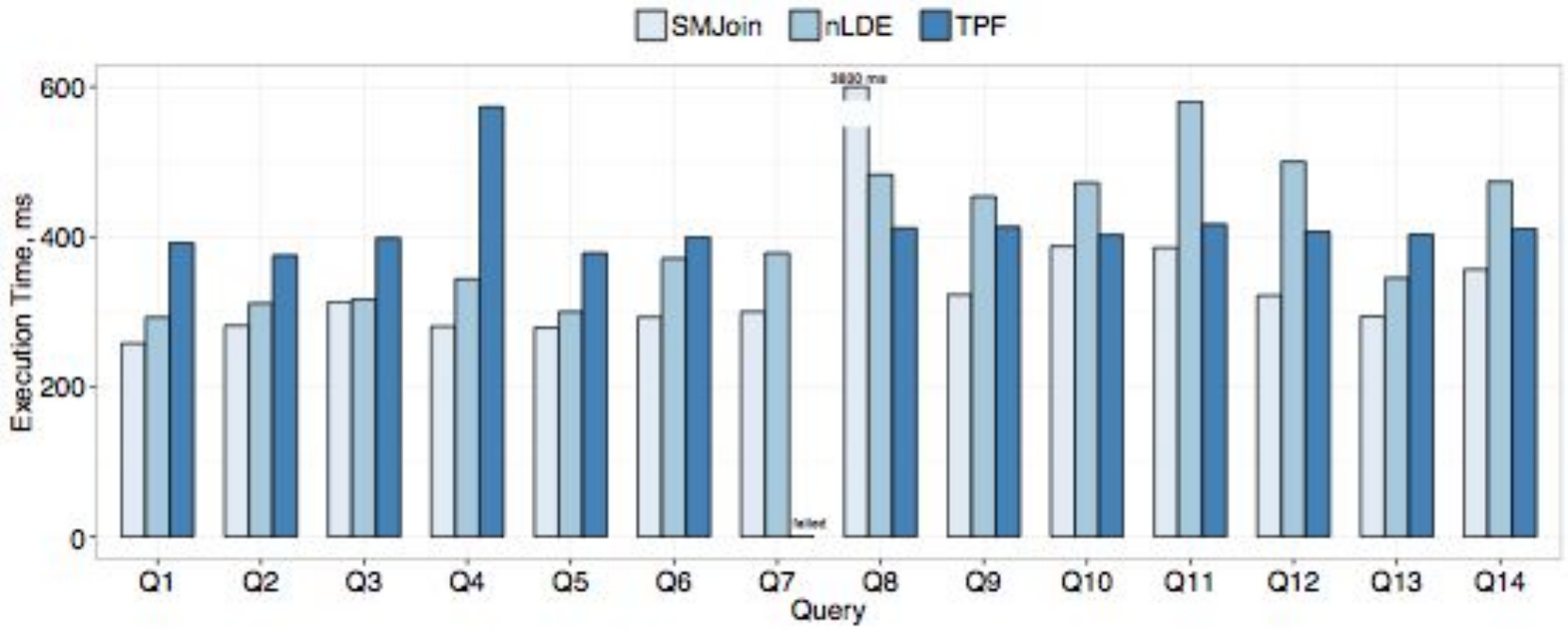
Metrics:

- Execution Time, ms
- Completeness over time, %

Compared tools:

- **TPF**: triple pattern fragment server [7]
- **nLDE**: network of Linked Data Eddies [2]
- **SMJoin**: multi-way join operator for SPARQL [13]

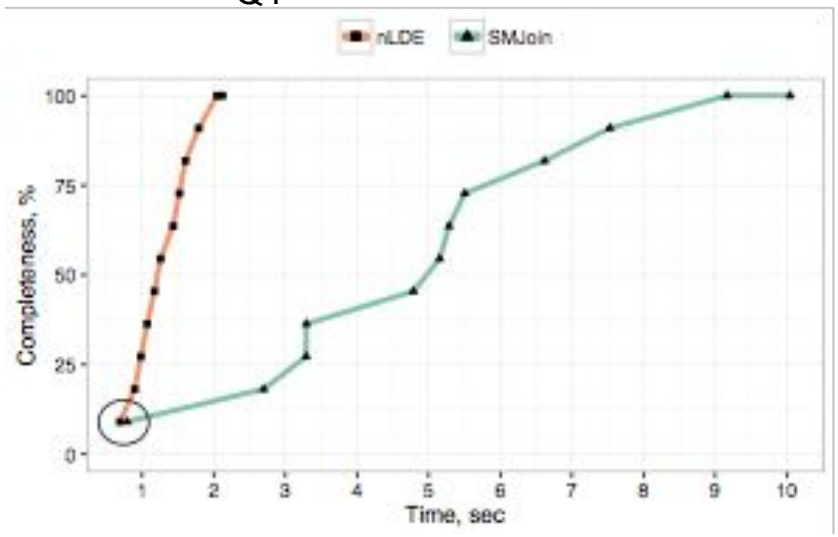
Benchmark 1: High Selective Queries



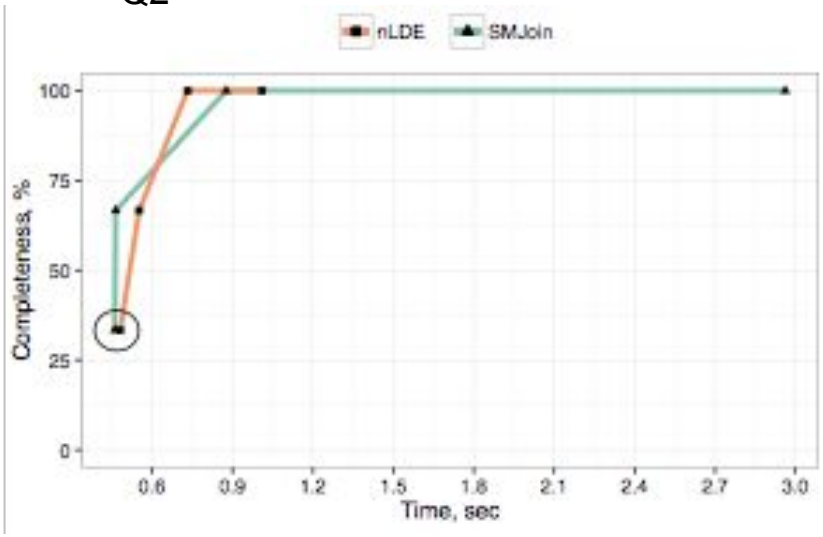
An adaptive approach like SMJoin outperforms other approaches in high-selective queries that produce small number of intermediate results

Benchmark 2: Low Selective Queries

Q1



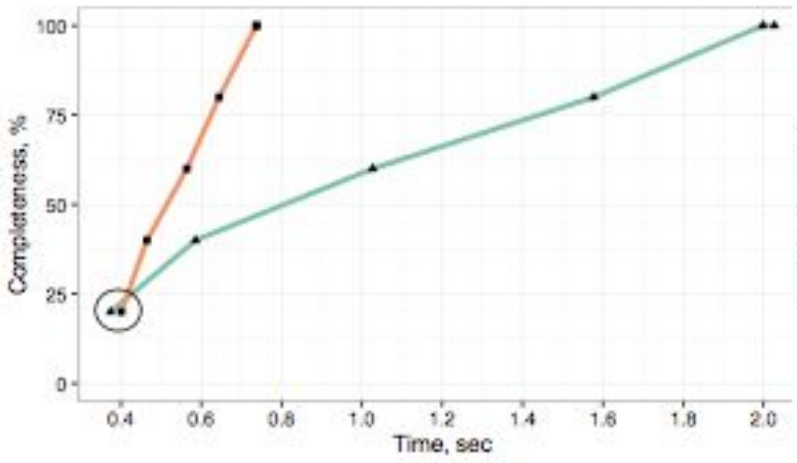
Q2



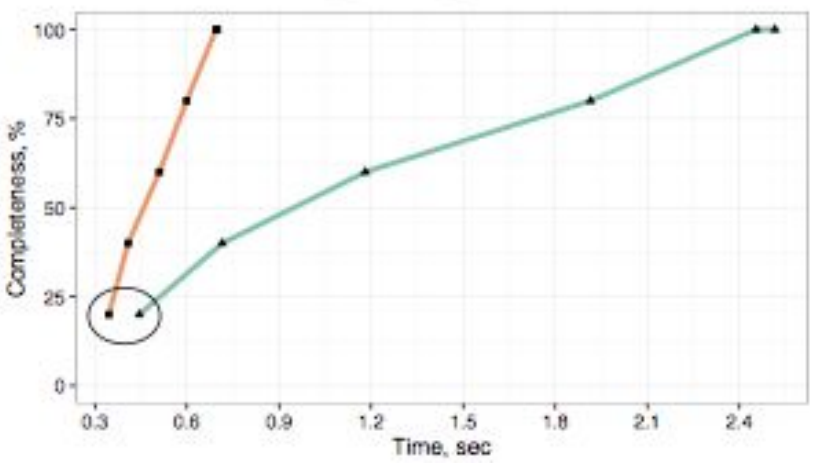
- SMJoin yields the first answer at about the same time as nLDE
- SMJoin has to process more intermediate results
- Q2: results are yielded but all intermediate tuples have to be processed

Benchmark 2: Low Selective Queries

Q3

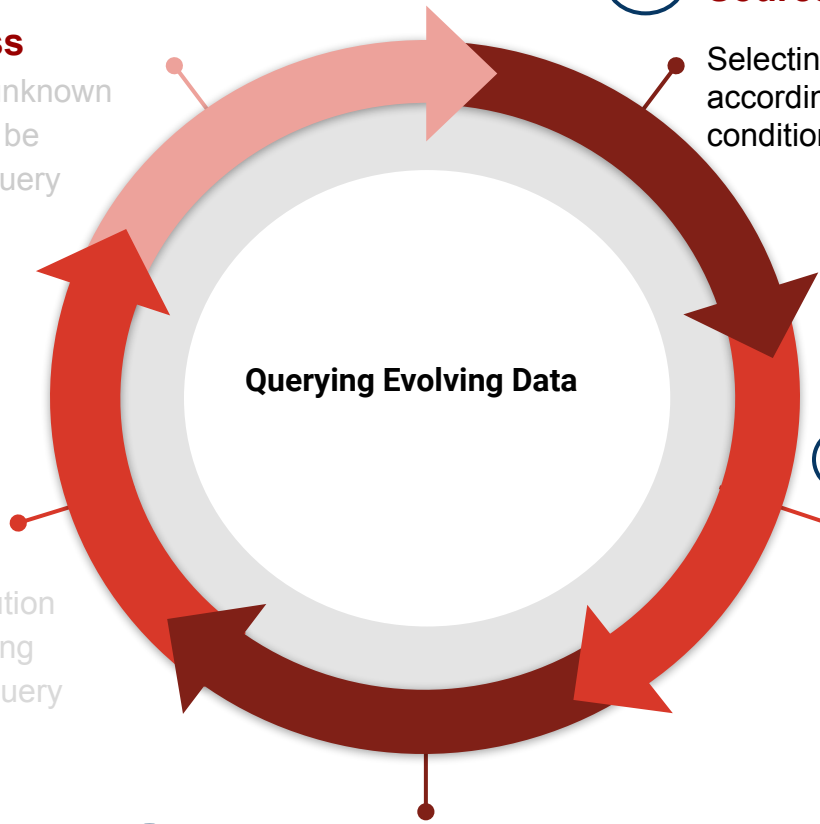


Q4



- SMJoin yields the first answer at about the same time as nLDE
- SMJoin has to process more intermediate results

Required Solutions to Support Evolution



1 **Source Evolution**

Selecting the sources according to their current conditions and availability

2 **Environment Evolution**

Executing queries according to current conditions of the environment

3 **Data Evolution**

Considering the status of the data, e.g., completeness, during the execution of the query

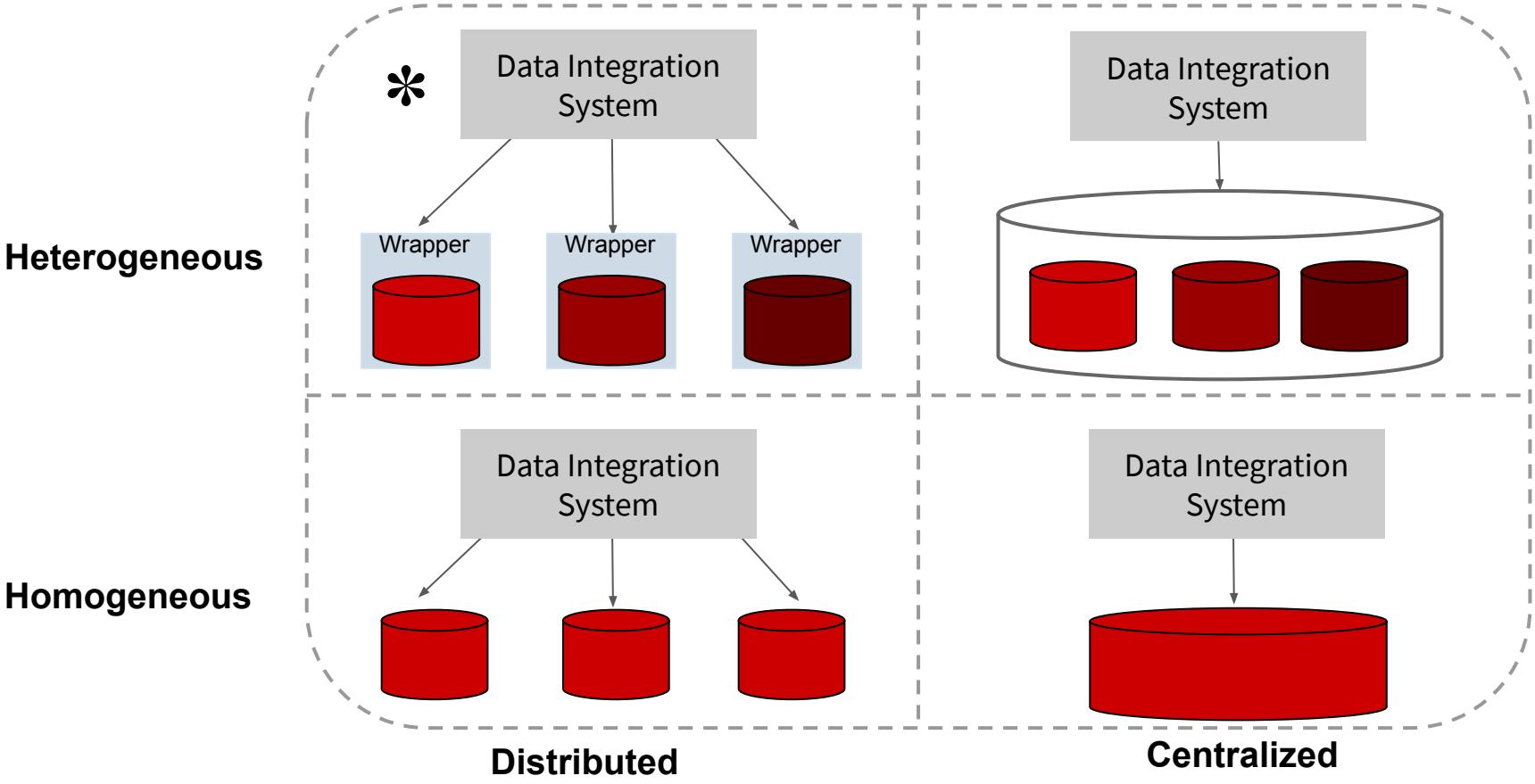
4 **Knowledge Evolution**

Considering the evolution of the knowledge during the execution of the query

5 **Knowledge Incompleteness**

Considering that unknown facts may need to be predicted during query execution

Data Integration Systems



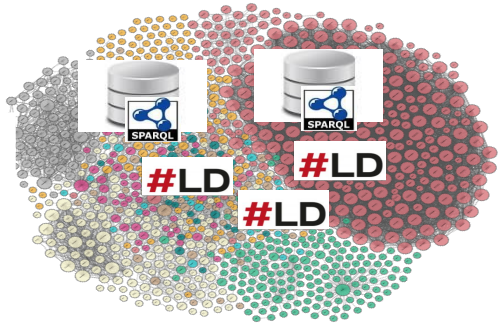
Hybrid Federated Query Engines

SPARQL Query **Q**

Source Selection & Query Decomposition

Query Optimizer

Execution Strategies

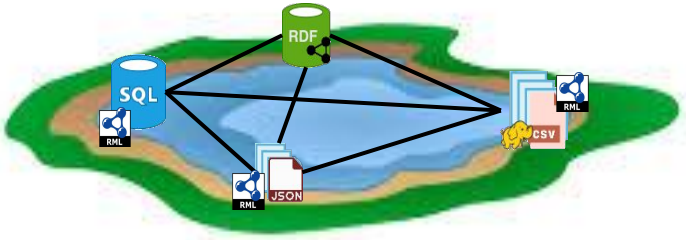


SPARQL Query **Q**

Source Selection & Query Decomposition over Heterogeneous Sources

Query Optimizer

Hybrid Execution Strategies **over Heterogeneous Sources**

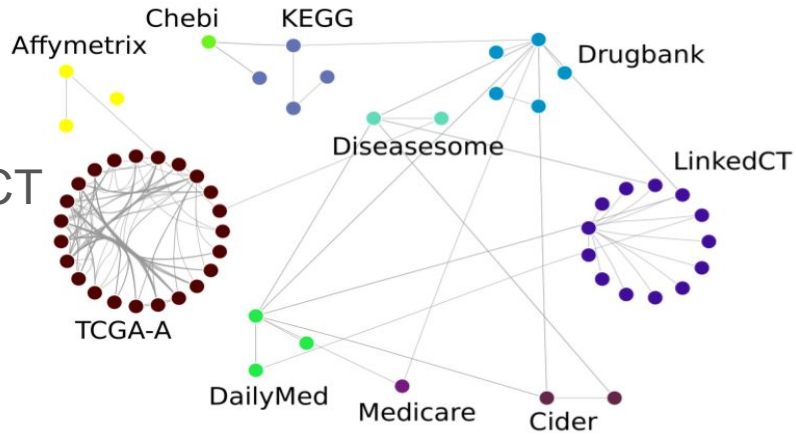


33

Kemele M. Endris, Philipp D. Rohde, Maria-Esther Vidal, Sören Auer: Ontario: Federated Query Processing Against a Semantic Data Lake. DEXA (1) 2019

Experimental Setup

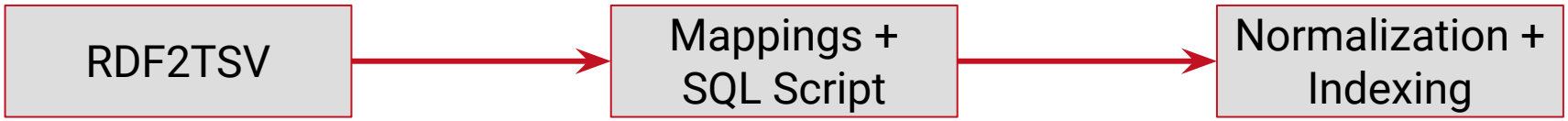
- Benchmark:
  - Life Science Linked Open Data (LSLOD)[15]
  - 10 RDF Data Source
  - 10 Simple Queries
    - UNION, OPTIONAL, DISTINCT
    - 3 - 8 triple patterns
    - 2 - 4 star-shaped sub-queries



#triples	#subjects	#predicates	#objects	RDF file size
96.10 M	8.32 M	742	27.47 M	16.0 GB

15] A. Hasnain, Q. Mehmood, S. Sana e Zainab, M. Saleem, C. Warren, D. Zehra, S. Decker, and D. Rebholz-Schuhmann. Biofed: federated query processing over life sciences linked open data. Journal of Biomedical Semantics, 8(1):13, Mar 2017.

Data Preparation Pipeline



- One NT file per RDF Class
- Transform NT files to TSV files
- Single-value predicates
  - main file of RDF Class
- Multi-value predicates
  - separate file for each multi-value predicate

- Generate RML mappings from the data collected during RDF2TSV
  - one file per RDF Class
- SQL script for creating the relational tables
  - one file per data set
  - data is loaded from TSV with LOAD DATA INFILE command

- Normalization by hand
- 3NF
- Indexes
  - primary keys
  - candidate keys
- Foreign key constraints

Experimental Setup

Experimental Configuration

- 23 Docker containers
  - 10 RDF sources (Virtuoso 6.01.3127)
  - 10 RDB sources (MySQL 5.7)
  - Three engines (FedX, MULDER, Ontario)
- Metrics:
  - **Execution time:** Time elapsed between query submission and retrieval of last answer
  - **Cardinality:** Number of answers produced by the engine
  - **Completeness:** Percentage of answers returned w.r.t the ground truth
  - **Throughput**: number of answers produced per second
  - **dief@t [15]:** Continuous efficiency at time t
    - Area-under-the-curve of the answer traces

[15] Maribel Acosta, Maria-Esther Vidal, York Sure-Vetter: Diefficiency Metrics: Measuring the Continuous Efficiency of Query Processing Approaches. International Semantic Web Conference, 2017

Experimental Setup

Experimental Configuration

- 23 Docker containers
  - 10 RDF sources (Virtuoso 6.01.3127)
  - 10 RDB sources (MySQL 5.7)
  - Three engines (FedX, MULDER, Ontario)
- Metrics:
  - **Execution time:** Time elapsed between query submission and retrieval of last answer
  - **Cardinality:** Number of answers produced by the engine
  - **Completeness:** Percentage of answers returned w.r.t the ground truth
  - **Throughput**: number of answers produced per second
  - **dief@t [15]:** Continuous efficiency at time t
    - Area-under-the-curve of the answer traces

Types of Subqueries

CI: Star-shaped subqueries with no instantiations or filter clauses

CII: Star-shaped subqueries with no instantiations or filter clauses, and defined over an RDF class implemented by joining several relational tables in a data lake

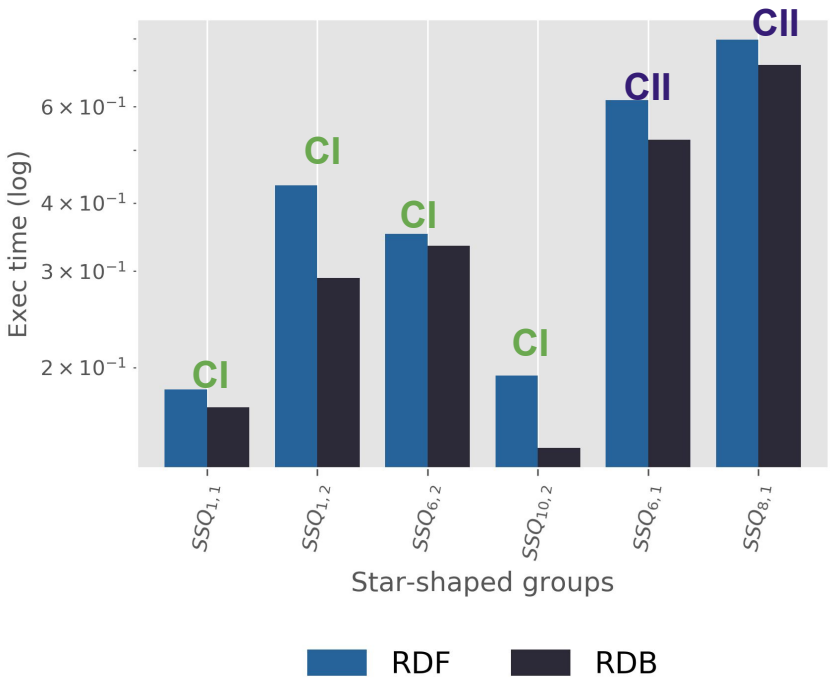
CIII: Star-shaped subqueries with instanstiations in object variables

CIV: Star-shaped subqueries with instantiations or filter clauses, and defined over an RDF class implemented by joining several relational tables in a data lake

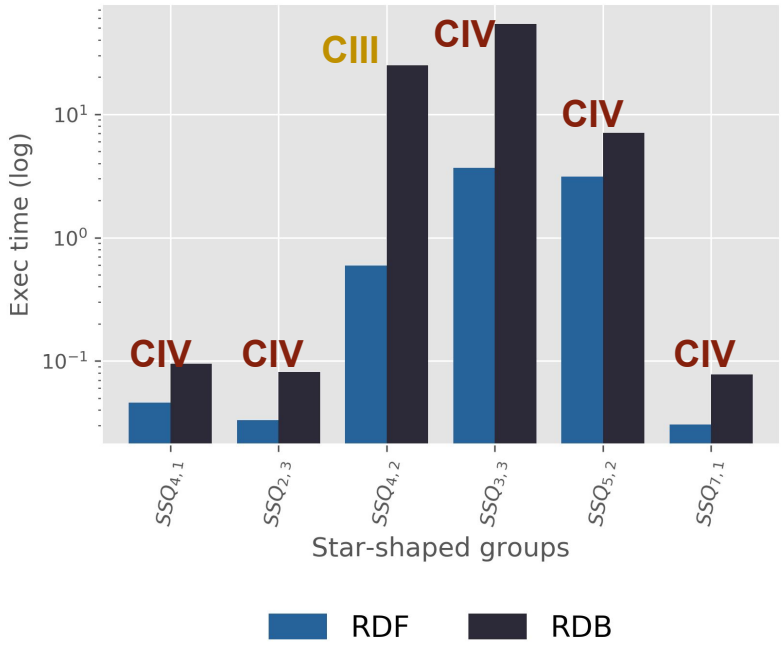
[15] Maribel Acosta, Maria-Esther Vidal, York Sure-Vetter: Diefficiency Metrics: Measuring the Continuous Efficiency of Query Processing Approaches. International Semantic Web Conference, 2017

Exp I: Impact of Star-shaped Groups

Goal: Evaluate the impact of different subqueries--**star-shaped groups (SSQs)**-- on the performance of a query engine.



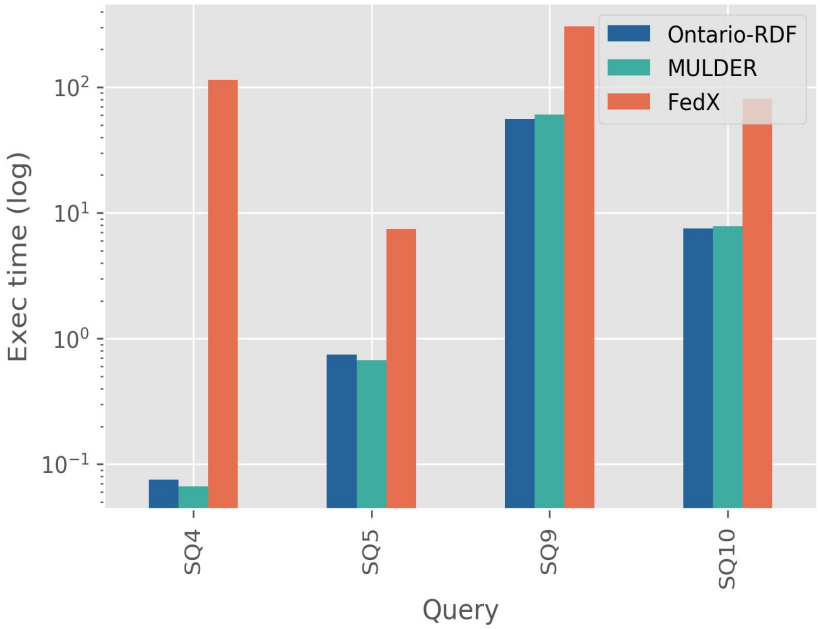
RDB scans a relation or a set of relations, while an **RDF** engine scans over all data. Thus, RDB engines **outperform** RDF engines

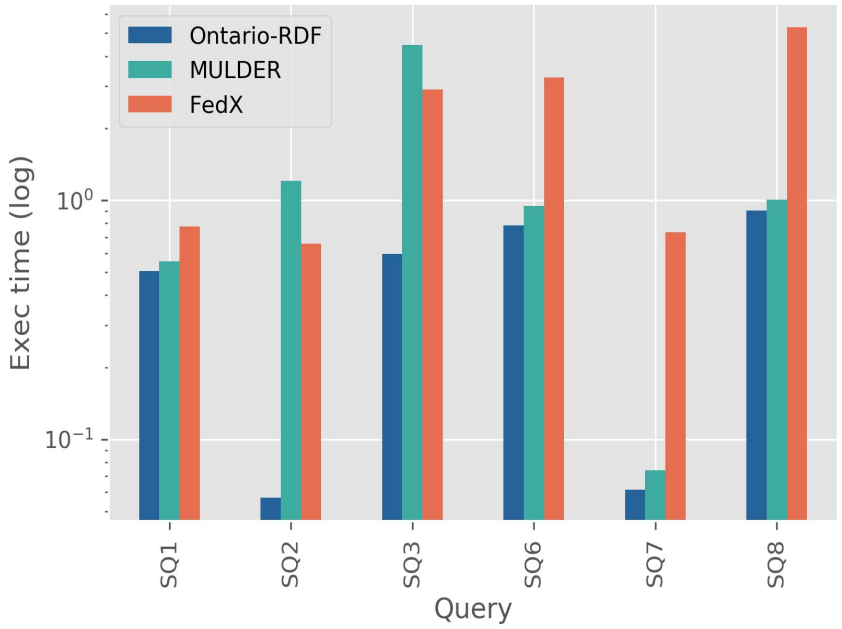


RDB only has indexes on primary keys, while an **RDF** engine has indexes over combinations of subject, predicate, and object. Thus, RDF engines **outperform** RDB engines

Exp II: Impact of Considering Heterogeneity

Goal: Performance of Ontario engine over RDF data sources and the overhead introduced while considering heterogeneity

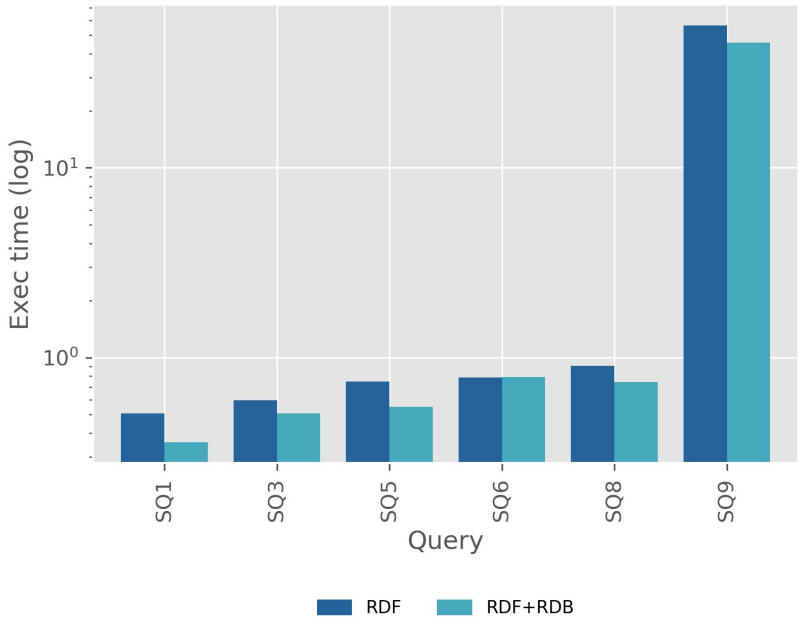


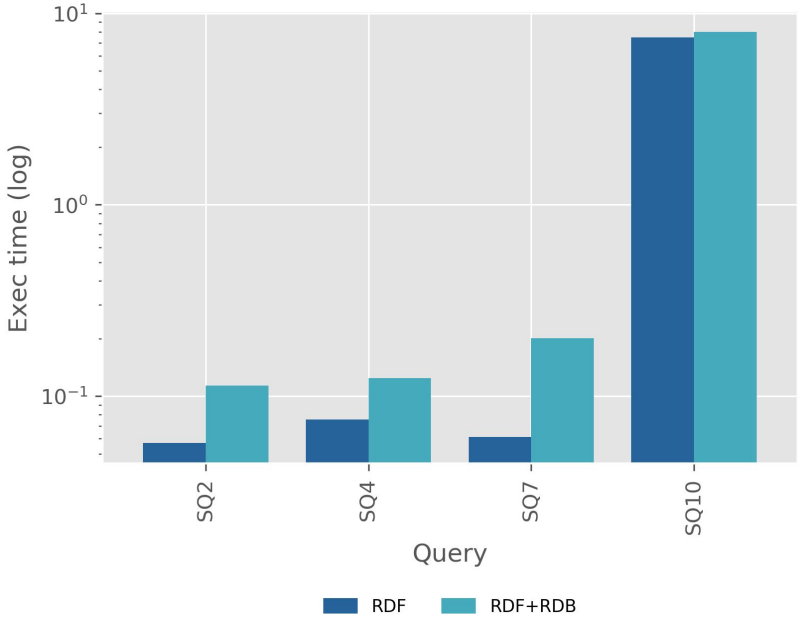


Ontario pays the price of **considering heterogeneous data sources**. Ontario outperforms both **FedX and MULDER** by generating efficient plans and using optimization rules tailored for RDF sources on the rest of the queries

Exp III: Impact of Heterogeneity

Goal: Performance of Ontario over heterogeneous sources, i.e., RDF and RDB



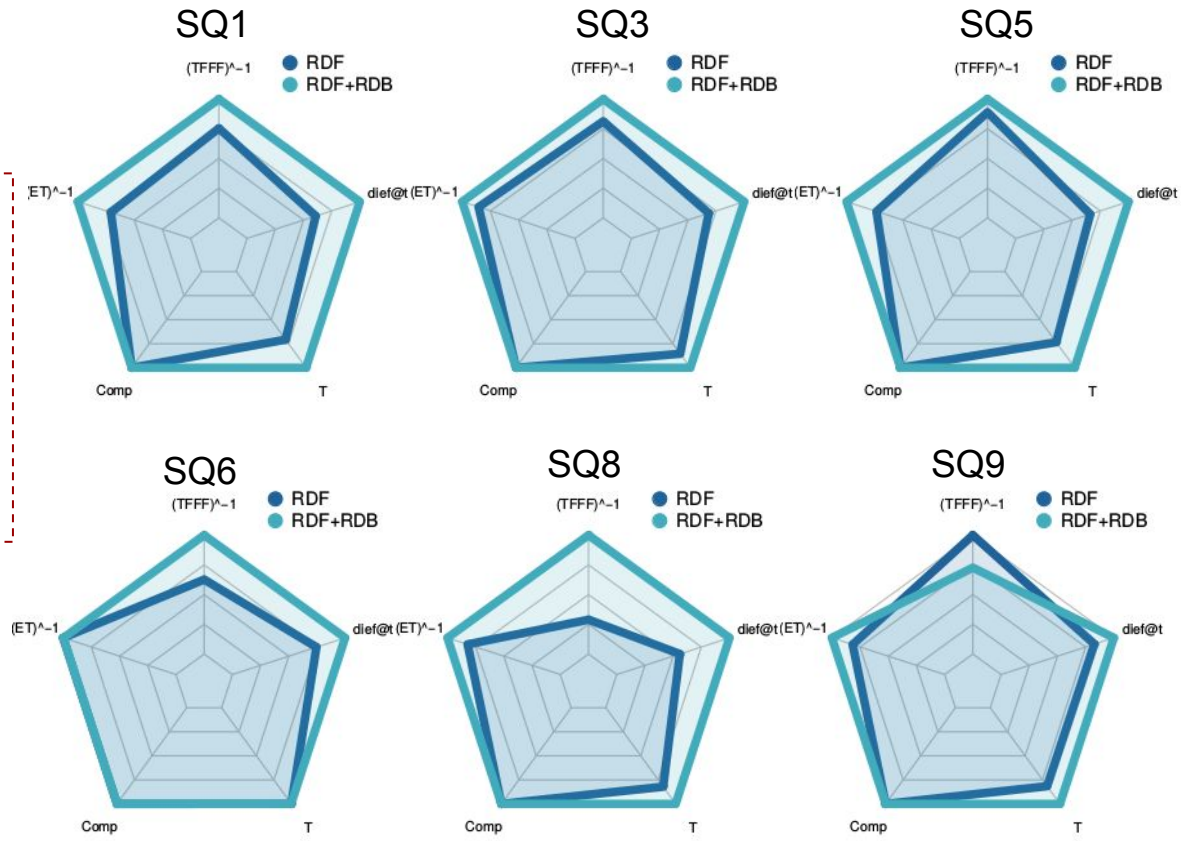


Characteristics of the queries impact on the performance of the federated query engine. **Ontario** is able to identify according to the data source implementations which is the most effective plan.

Exp IV: Measuring Continuous Efficiency

Goal: Performance of Ontario in producing continuous answers.

Characteristics of the queries impact on the performance of the federated query engine. **Ontario** is able to identify according to the data source implementations which is the most effective plan.



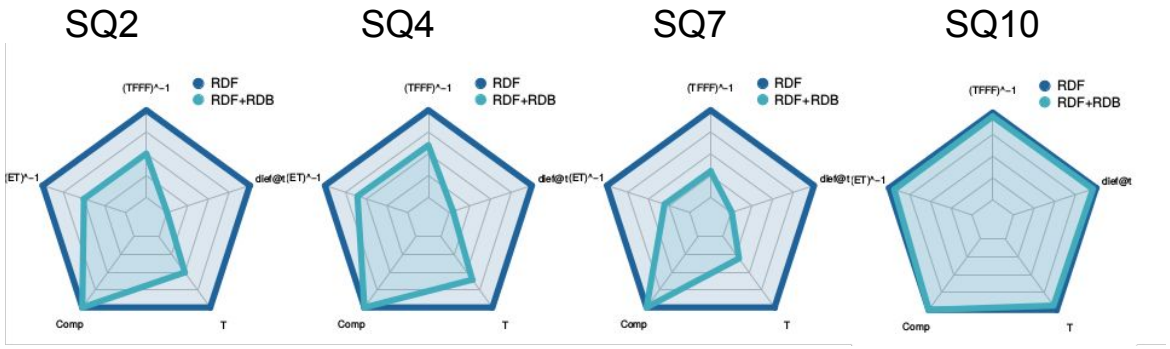
Queries composed of SSQs in CI or CII

Higher is Better!

Exp IV: Measuring Continuous Efficiency

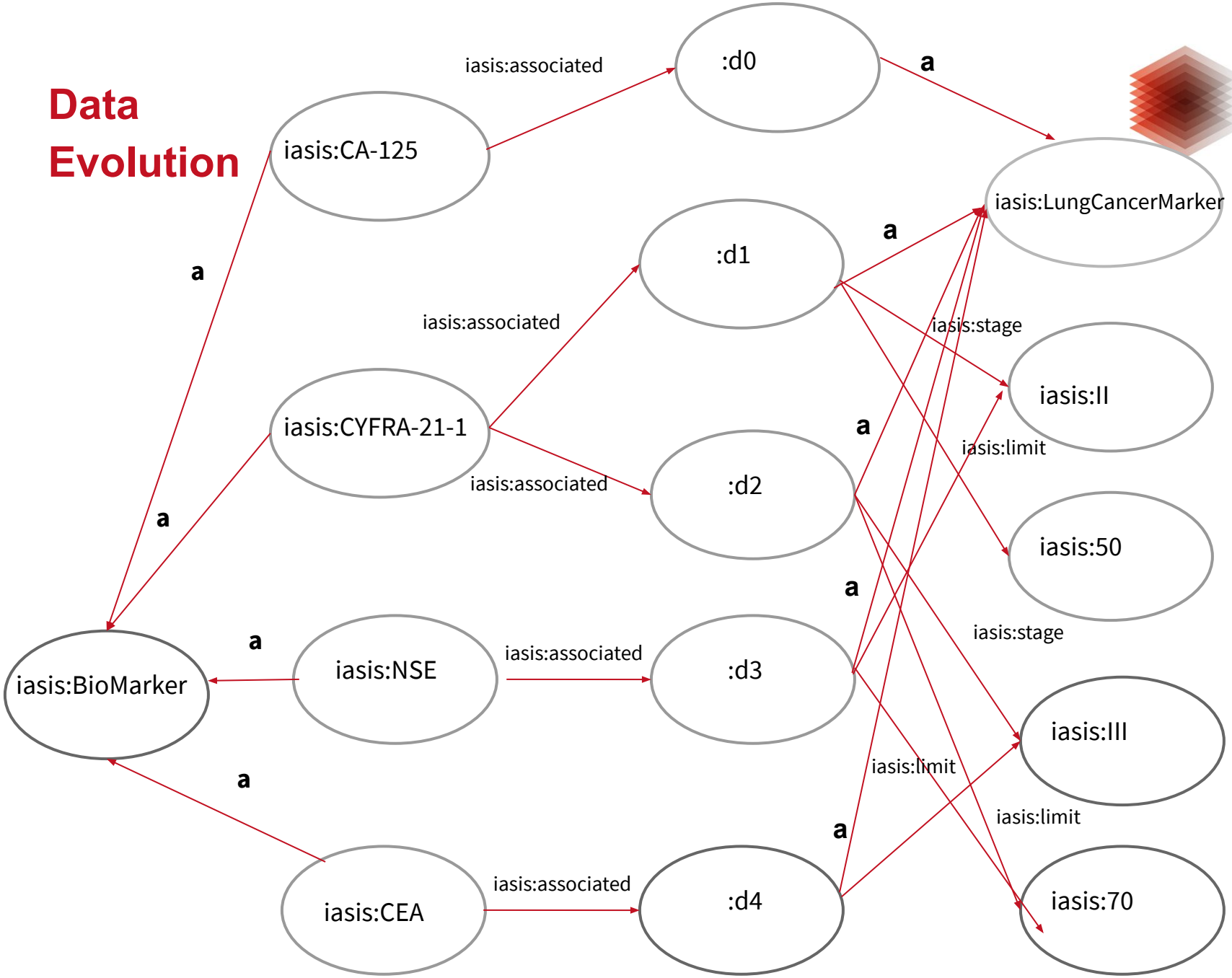
Goal: Performance of Ontario in producing continuous answers.

Characteristics of the queries impact on the performance of the federated query engine. **Ontario** is able to identify according to the data source implementations which is the most effective plan.



Queries composed of SSQs in CIII or CIV **Higher is Better!**

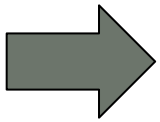
Data Evolution



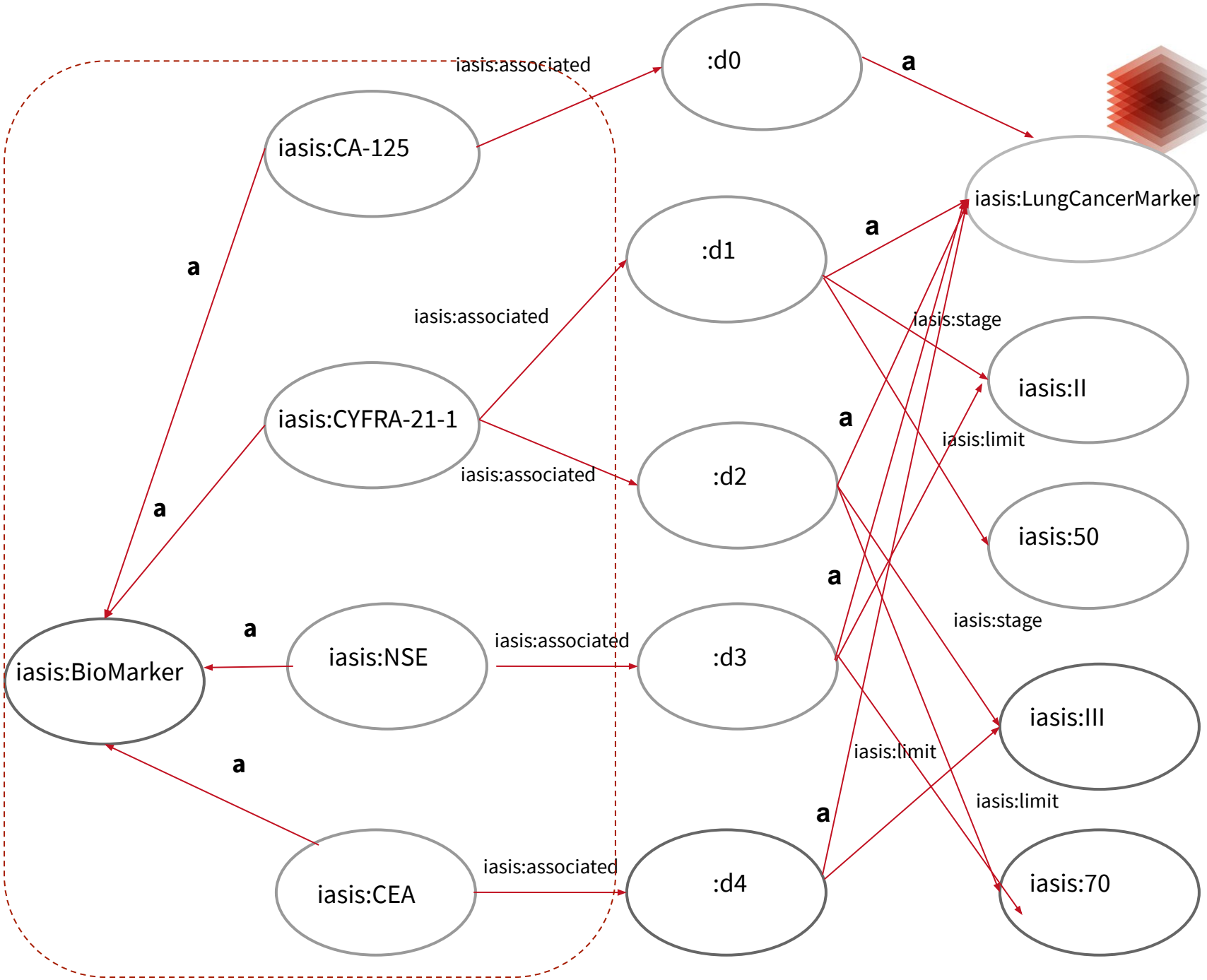
Data Changes….

Lung Cancer Biomarkers?

```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?id  ?stage ?limit
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
  ?bm iasis:associated ?obs .
  ?bm iasis:limit ?limit .
  ?bm iasis:stage ?stage
  ?id iasis:associated ?bm .
  }
```



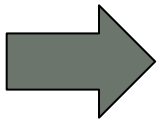
iasis:CYFRA-21-1	iasis:II	iasis:50
iasis:CYFRA-21-1	iasis:III	iasis:70
iasis:NSE	iasis:III	iasis:70


iasis:associated
:d0
a
iasis:CA-125
iasis:LungCancerMarker
a
:d1
a
iasis:associated
iasis:stage
iasis:II
iasis:CYFRA-21-1
a
iasis:limit
iasis:associated
:d2
iasis:50
a
a
iasis:stage
iasis:BioMarker
a
iasis:NSE
iasis:associated
:d3
iasis:III
iasis:limit
a
iasis:limit
a
iasis:CEA
iasis:associated
:d4
iasis:70

Data Changes….

Lung Cancer Biomarkers?

```
PREFIX iasis:<http://iasis/vocab/>
SELECT distinct ?id
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
  ?id iasis:associated ?bm .
  }
```

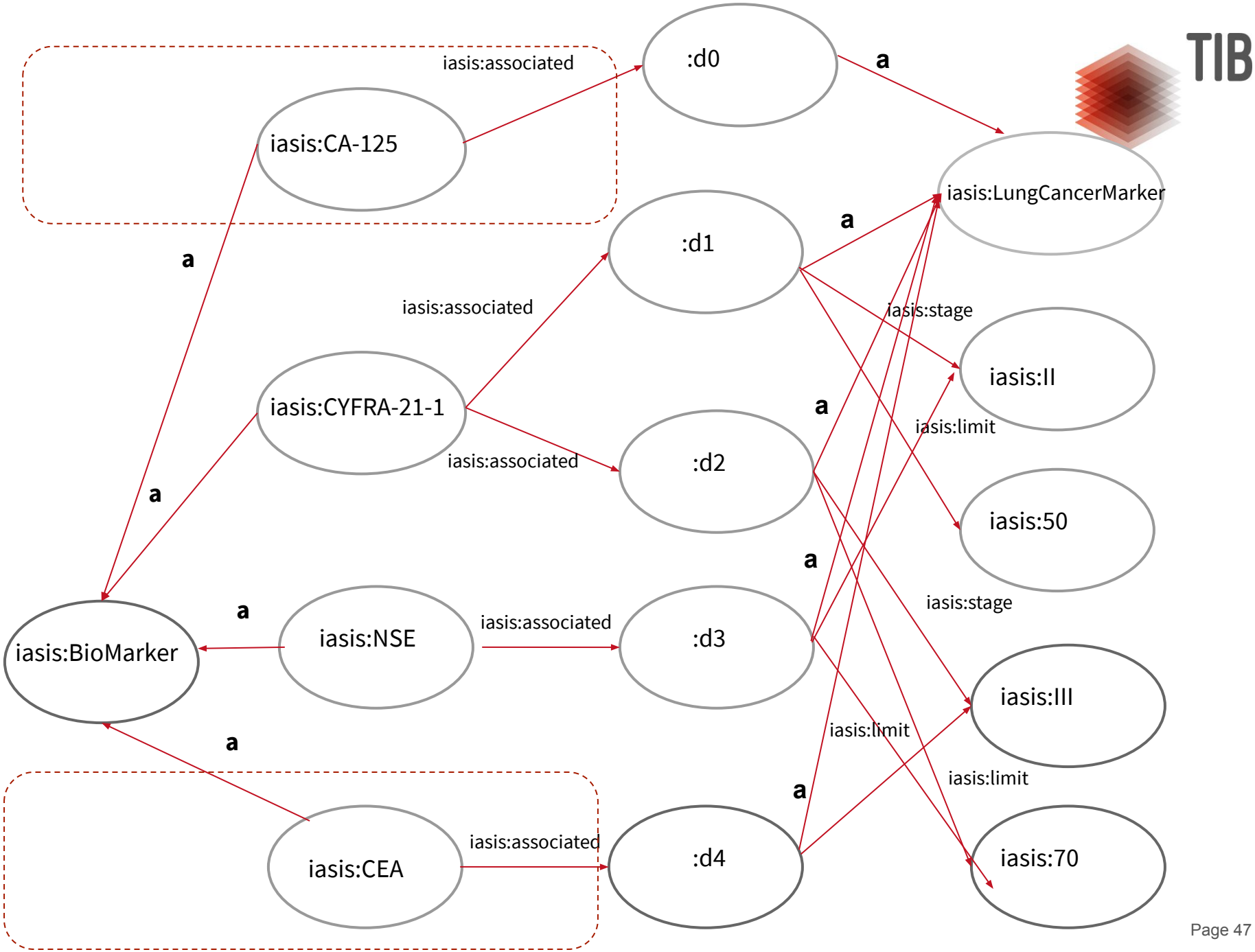


iasis:CYFRA-21-1

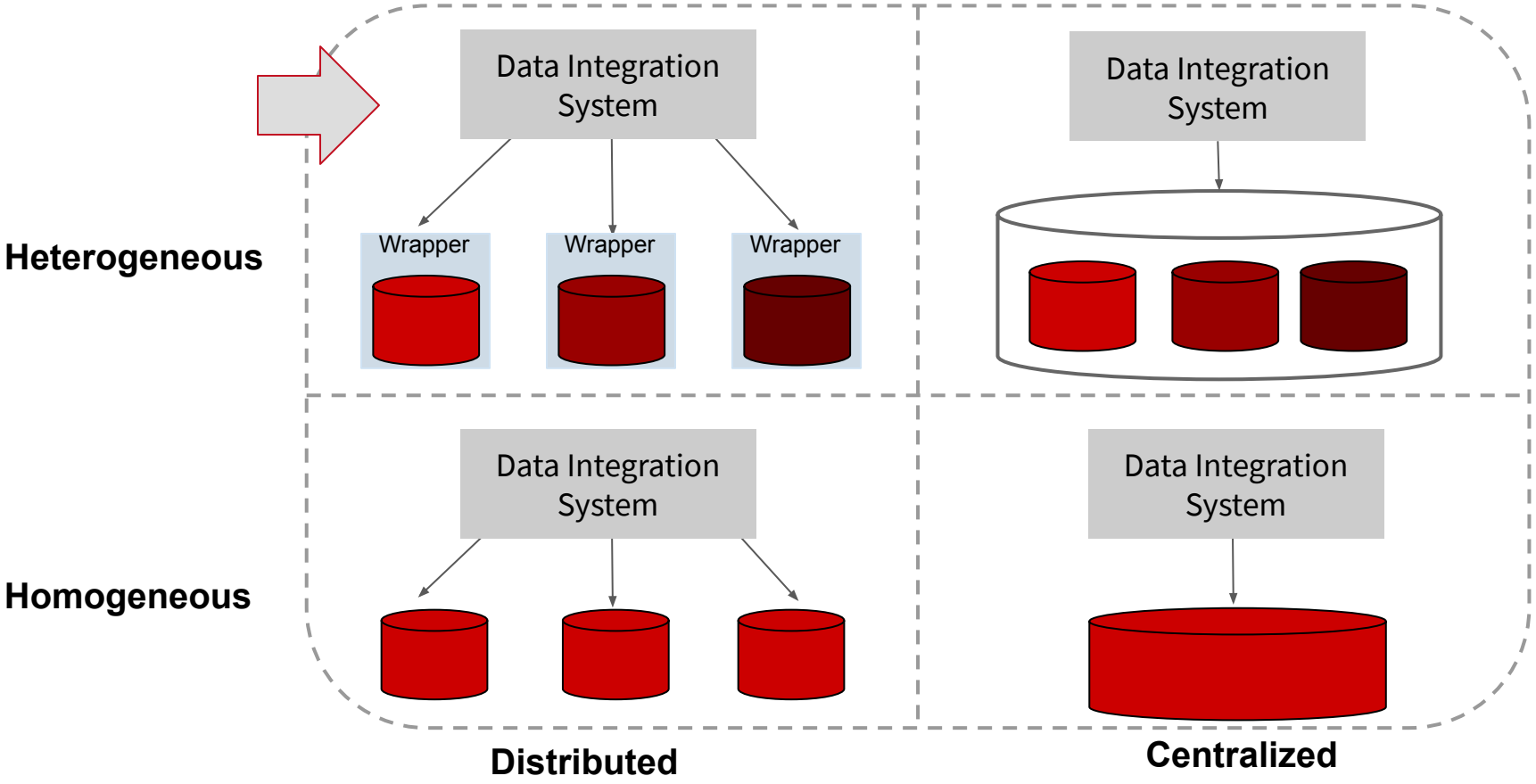
iasis:CEA

iasis:NSE

iasis:CA-125


iasis:associated
:d0
a
iasis:CA-125
iasis:LungCancerMarker
a
:d1
a
iasis:associated
iasis:stage
iasis:II
iasis:CYFRA-21-1
a
iasis:limit
iasis:associated
:d2
iasis:50
a
a
iasis:stage
a
iasis:BioMarker
iasis:NSE
iasis:associated
:d3
iasis:III
iasis:limit
a
iasis:limit
a
iasis:CEA
iasis:associated
:d4
iasis:70

Data and Knowledge Evolution



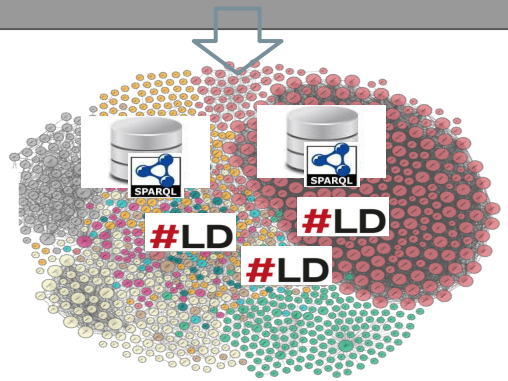
Hybrid Federated Query Engines

SPARQL Query **Q**

Source Selection & Query Decomposition

Query Optimizer

Execution Strategies

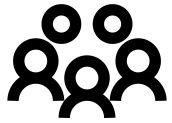


SPARQL Query **Q**

Source Selection & Query Decomposition

Query Optimizer

Hybrid Execution Strategies
Crowd Microtask Manager



Crowd

M. Acosta, E. Simperl, F. Flöck, M.-E. Vidal: HARE: A Hybrid SPARQL Enhancing answer completeness of SPARQL queries via crowdsourcing. J. Web Sem. 45: 41-62 (2017)

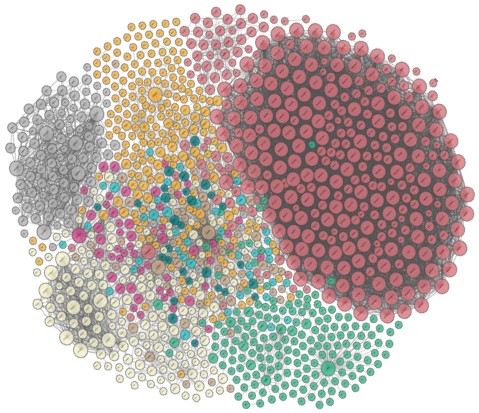
Hybrid Query Processing

Lung Cancer Biomarkers?

```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?id  ?stage ?limit
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
  ?bm iasis:associated ?obs .
  ?bm iasis:limit ?limit .
  ?bm iasis:stage ?stage
  ?id iasis:associated ?bm .
   }
```

```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?id
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
   ?bm iasis:associated ?obs .
   ?id iasis:associated ?bm .
     ?bm iasis:stage ?stage
}
```

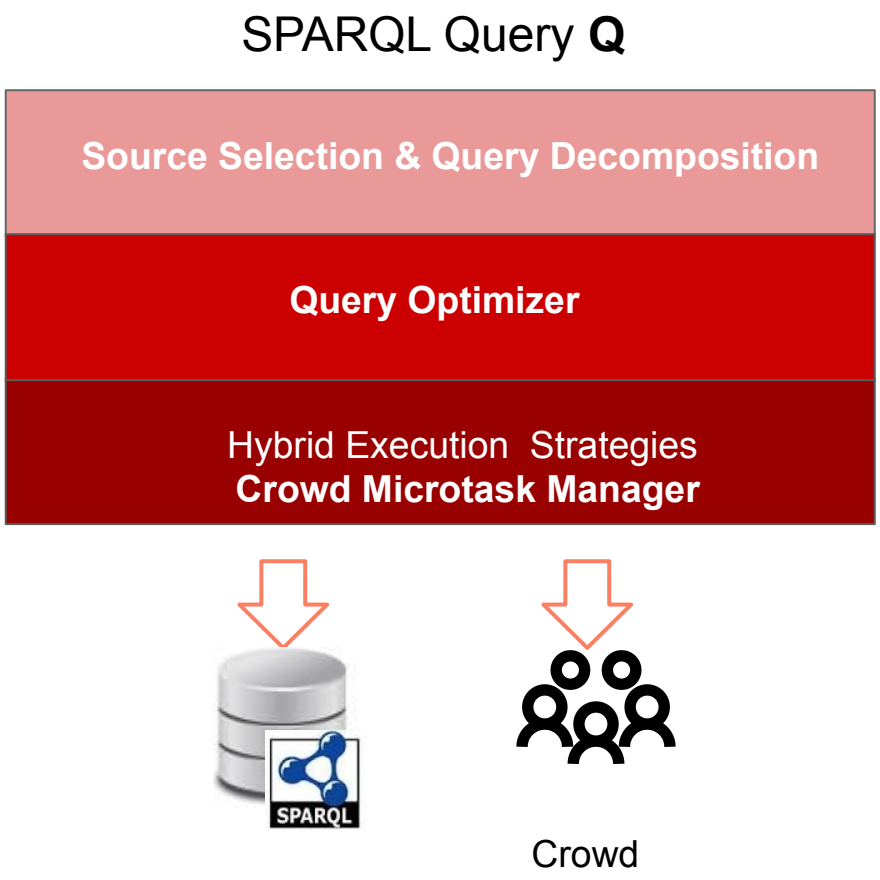
```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?limit
 WHERE {
?bm iasis:limit ?limit .
  ?bm iasis:stage ?stage
  ?id iasis:associated ?bm .
```



Crowd

HARE: A Hybrid Query Engine

- **Completeness model** to estimate dataset completeness
- **Crowd knowledge bases** to capture crowd answers about missing data
- **Query engine** that combines **knowledge** in knowledge bases and **estimates** from the **completeness** model to **decompose** and plan sub-query execution
- **Microtask manager** that exploits **metadata** to **crowdsource** subqueries as **microtasks** and update the **knowledge bases** according to the **crowd answers**



M. Acosta, E. Simperl, F. Flöck, M.-E. Vidal: HARE: A Hybrid SPARQL Enhancing answer completeness of SPARQL queries via crowdsourcing. J. Web Sem. 45: 41-62 (2017)

HARE Microtasks

Metadata is utilized by the microtask manager to **automatically** generate **well-described** crowd tasks
Microtasks are submitted to **crowdsourcing platforms**, e.g., CrowdFlower or Mechanical Turk
Answers collected from the **crowd** are represented as **structured data**

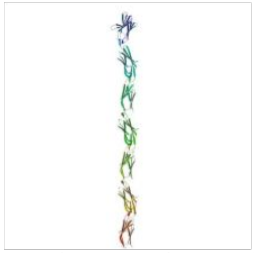
What is the **value** of the **Marker CEA** for **Lung Cancer Stage III**?

Search in Google: Carcinoembryonic antigen

Short Description: Carcinoembryonic antigen (CEA) describes a set of highly related glycoproteins involved in cell adhesion. CEA is normally produced in gastrointestinal tissue during fetal development, but the production stops before birth. Therefore, CEA is usually present only at very low levels in the blood of healthy adults. However, the serum levels are raised in some types of cancer, which means that it can be used as a tumor marker in clinical tests. Serum levels can also be elevated in heavy smokers.

Wikipedia Page: https://en.wikipedia.org/wiki/Carcinoembryonic_antigen

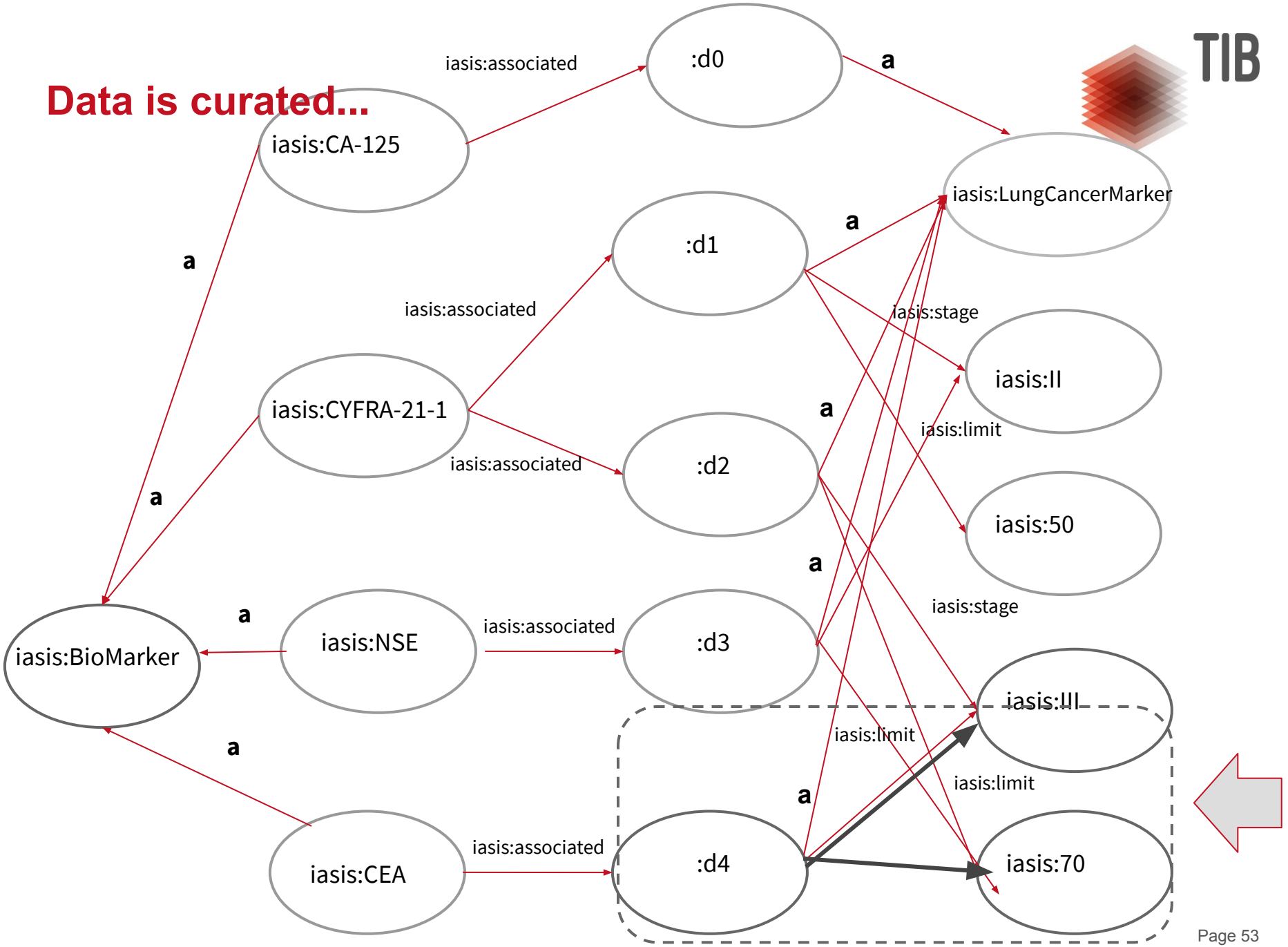
Picture:



Does the Marker CEA have a value for Lung Cancer Stage III?

- Yes
- No
- I do not know

Data is curated...

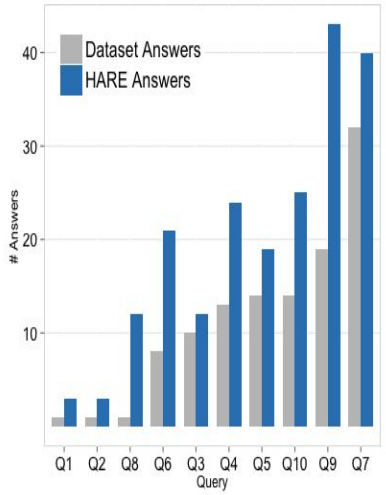


Experimental Study - Set Up

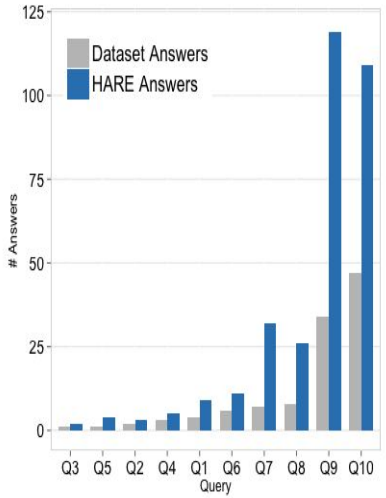
- **Benchmark:** 50 queries against DBpedia (v. 2014).
  - Ten queries in five different knowledge domains:
    History, Life Sciences, Movies, Music, and Sports.
- **Implementation details:**
  - HARE is implemented in Python 2.7.6,
  - The crowd is reached via CrowdFlower.
- **Crowdsourcing configuration:**
  - Four different RDF triples per task, 0.07 US$ per task.
  - At least three judgments were collected per task.
- Total RDF triple patterns crowdsourced: 502
- Total answers collected from the crowd: 1,609

Experimental Evaluation

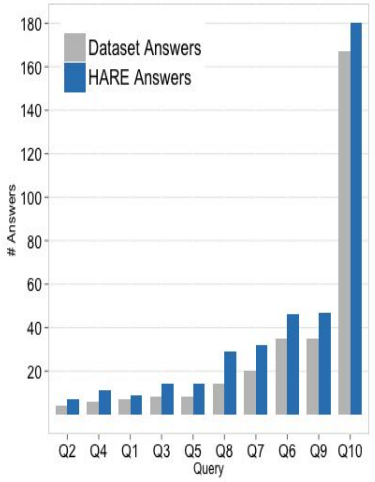
Sports



Music



Life Sciences

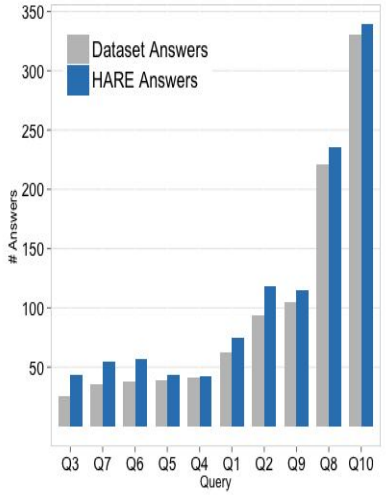


Crowdsourced answers and answers **collected** from DBpedia

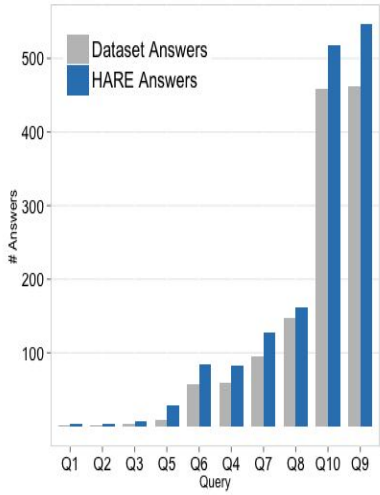
HARE **identifies** subqueries with **incomplete answers**

Hybrid query processing **enhances** query answer **completeness**

Movies

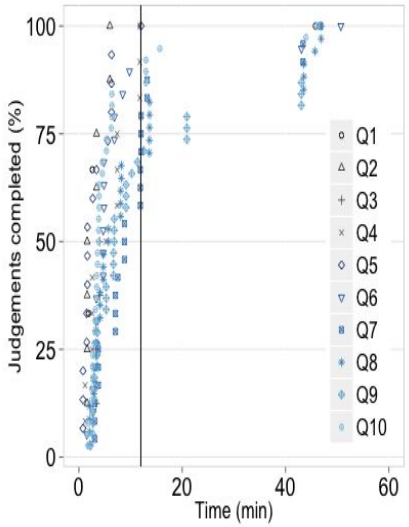


History

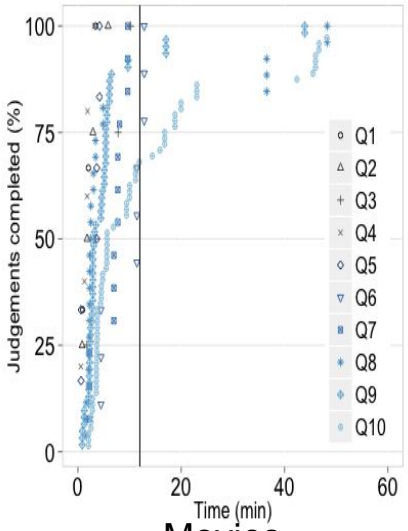


Experimental Evaluation

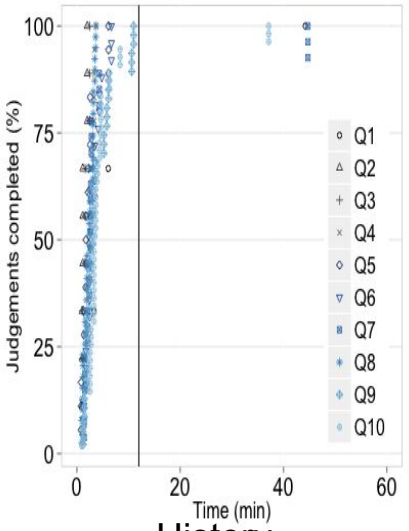
Sports



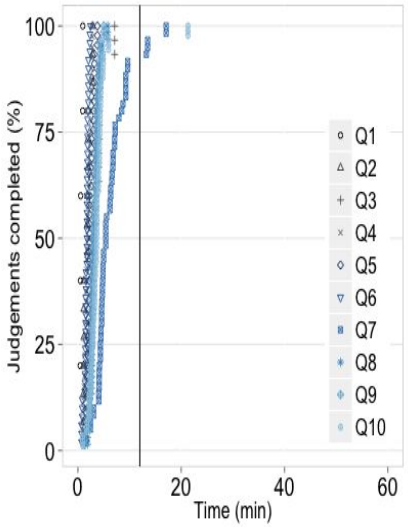
Music



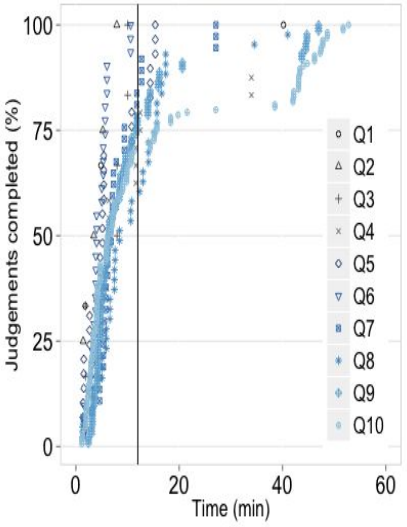
Life Sciences



Movies



History



HARE **is able** to produce more than 75% of the answers at the 12th minute

Experimental Evaluation

Precision

	Sports	Music	Life Sciences	Movies	History
Q1	1.00	1.00	0.67	0.88	1.00
Q2	1.00	1.00	1.00	0.96	1.00
Q3	1.00	1.00	0.89	0.79	0.67
Q4	0.55	0.67	1.00	1.00	0.96
Q5	0.86	0.67	1.00	1.00	0.95
Q6	0.69	0.83	1.00	1.00	0.96
Q7	1.00	0.63	0.71	1.00	0.57
Q8	1.00	0.67	0.88	0.94	0.72
Q9	0.46	0.73	1.00	1.00	0.64
Q10	0.92	0.49	1.00	1.00	0.95
Avg	**0.85**	**0.77**	**0.91**	**0.96**	**0.84**

Recall

	Sports	Music	Life Sciences	Movies	History
Q1	1.00	1.00	1.00	0.47	1.00
Q2	1.00	0.29	1.00	1.00	1.00
Q3	1.00	1.00	1.00	1.00	1.00
Q4	0.83	1.00	1.00	1.00	1.00
Q5	1.00	0.86	1.00	1.00	1.00
Q6	1.00	1.00	1.00	1.00	0.96
Q7	1.00	1.00	1.00	1.00	0.84
Q8	1.00	1.00	1.00	1.00	0.78
Q9	1.00	1.00	1.00	1.00	0.92
Q10	1.00	1.00	1.00	1.00	0.98
Avg	**0.98**	**0.91**	**1.00**	**0.95**	**0.95**

The crowd exhibits heterogeneous performance within domains.
This supports the importance of HARE triple-based approach.

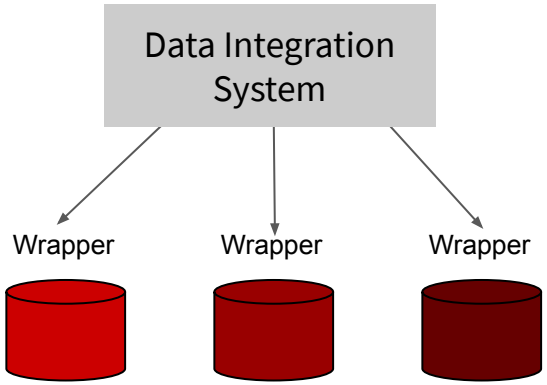
Applications

http://project-iasis.eu/

https://www.bigmedilytics.eu/

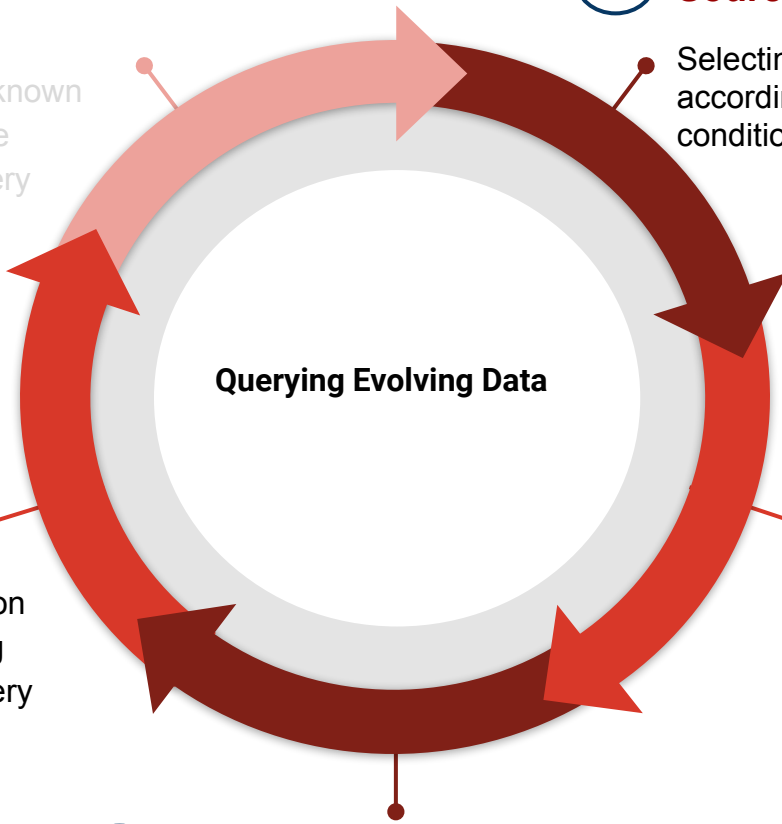
https://qualichain-project.eu/

Lessons Learned



- **Hybrid data integration** systems allow for the adaptation of the system to the conditions of the data sources
- **Hybrid data integration** systems enable the integration of **heterogeneous** data sources
- **Wisdom of the crowd** can contribute the evolution of the knowledge

Required Solutions to Support Evolution



5

Knowledge Incompleteness

Considering that unknown facts may need to be predicted during query execution

1

Source Evolution

Selecting the sources according to their current conditions and availability

4

Knowledge Evolution

Considering the evolution of the knowledge during the execution of the query

2

Environment Evolution

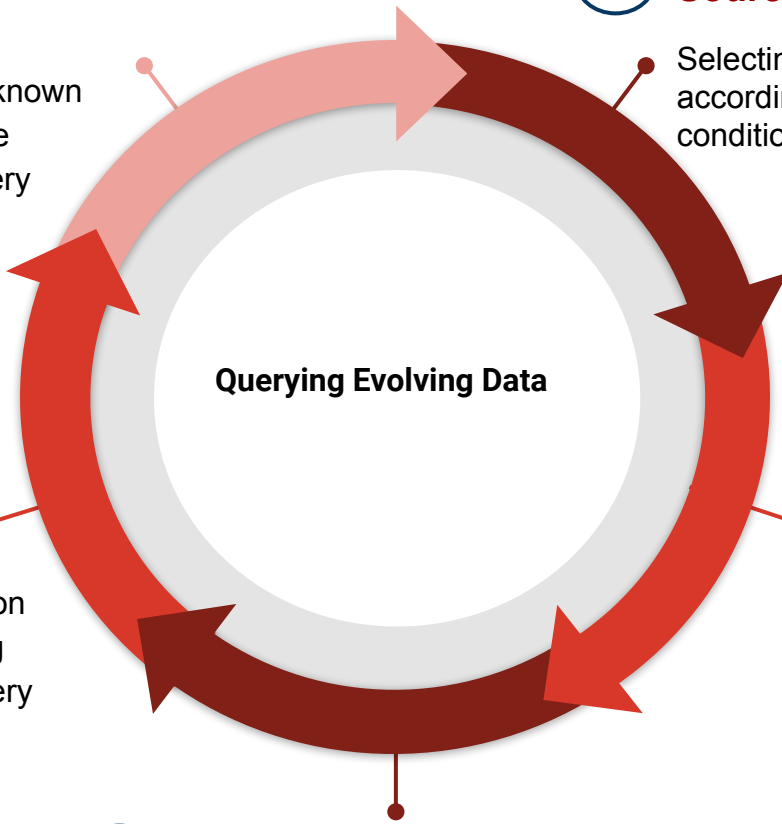
Executing queries according to current conditions of the environment

3

Data Evolution

Considering the status of the data, e.g., completeness, during the execution of the query

Required Solutions to Support Evolution



5

Knowledge Incompleteness

Considering that unknown facts may need to be predicted during query execution

1

Source Evolution

Selecting the sources according to their current conditions and availability

2

Environment Evolution

Executing queries according to current conditions of the environment

3

Data Evolution

Considering the status of the data, e.g., completeness, during the execution of the query

4

Knowledge Evolution

Considering the evolution of the knowledge during the execution of the query

Knowledge Evolution

Aroney RS, Dermody WC, Aldenderfer P, Parsons P, McNitt K, Marangos PJ, Whitacre MY, Ruddon RW, Wiernik PH, Aisner J

Cancer Treatment Reports [01 Jun 1984, 68(6):859-866]

Type: Research Support, U.S. Gov't, P.H.S., Research Support, Non-U.S. Gov't, Journal Article

Abstract

To correlate serial biomarkers and disease activity in carcinoma of the lung, carcinoembryonic antigen (CEA), neuron-specific enolase (NSE), adrenocorticotropic hormone (ACTH), C3-derived protein (C3DP-C), and LDH were assayed in 43 patients with small cell lung carcinoma (SCLC) and in 20 patients with non-small cell lung cancer (NSCLC) (15 with adenocarcinoma, three with squamous cell carcinoma, and two with mixed histology). Disease status after treatment was rated as one of the following: complete response, partial response, minor regression, stable disease, and progressive disease. Significant correlations between disease status and markers in SCLC were found for CEA, NSE, LDH, and ACTH. In NSCLC, only CEA and LDH showed significant correlation. Marker-marker correlations were significant in SCLC for CEA and NSE (P less than 0.05), CEA and LDH (P = 0.01), and NSE and LDH (P less than 0.01); in NSCLC none were significant. None of the markers exhibited significant correlations with specific metastatic sites. Certain biomarkers (CEA, NSE, and LDH in SCLC; CEA and LDH in NSCLC) can be used alone or in combination to monitor disease activity but appear to be no more sensitive than standard clinical investigational methods.

Funding

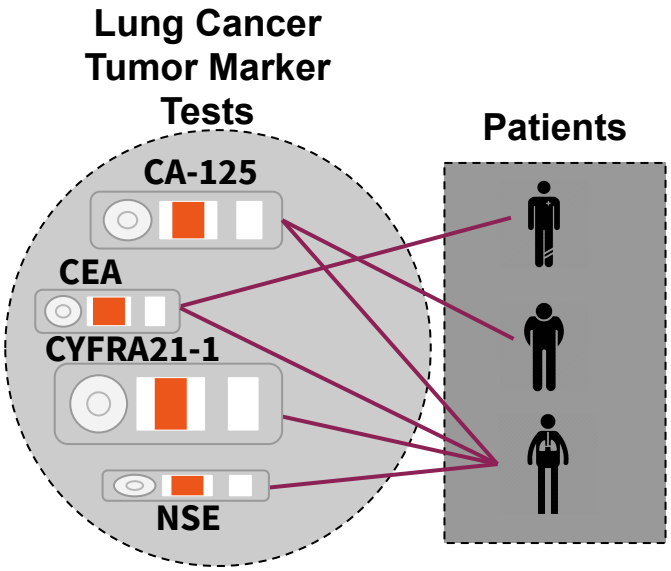
Knowledge Evolution

Table 4. Comparative levels of lung cancer biomarkers in blood plasma of patients with non-small-cell lung carcinoma (NSCLC) and small-cell lung carcinoma (SCLC) and healthy people.

Tumor-Associated Protein	NSCLC	SCLC	Normal
LDH	525.079 ± 24.817 ng mL^{-1} [134]	209.880 ± 161.322 ng mL^{-1} [134]	<245 ng mL^{-1} [134]
CRP	25.079 ± 24.817 ng mL^{-1} [134]	14.935 ± 21.078 ng mL^{-1} [134]	<8 ng mL^{-1} [134]
CEA	51.493 ± 77.529 ng mL^{-1} [134] 78.5 ng mL^{-1} [23] ≥ 100 ng mL^{-1} [65]	25.074 ± 40.957 [134]	<5.0 ng mL^{-1} 5.0 ng mL^{-1} [23,61] <20.9 ng mL^{-1} 6.5 ng mL^{-1} [66]
NSE	13.638 ± 5.571 ng mL^{-1} [134] >6.4 ng mL^{-1} [19] 5–35 ng mL^{-1} 17.95 ng mL^{-1} [61] 0–170 ng mL^{-1} [23]	62.972 ± 63.012 [134] 50.8 ng mL^{-1} [61] 15–173 ng mL^{-1} [23]	15.7–17.1 ng mL^{-1} 15.2 ng mL^{-1} 13 ng mL^{-1} [65]
CYFRA21-1	12.447 ± 15.814 ng mL^{-1} [134] 81.7 ng mL^{-1} [23]	6.418 ± 9.567 ng mL^{-1} [134]	<3.3 ng mL^{-1} [134] 3.3 ng mL^{-1} [35] 3.3 ng mL^{-1} [61,65] 0.5 ng mL^{-1} [65] 2.0 ng mL^{-1} [23]
SCCA	0.22–3.79 ng mL^{-1} [61] 0.5–1.7 >2 ng mL^{-1} [135]	0.15 ng mL^{-1} [61]	1.5 ng mL^{-1} [23]
TPS	0–3842 ng mL^{-1} [136]	12.5–773 ng mL^{-1} [23]	34.9 ng mL^{-1} UL^{-1} [23]
ProGRP	<35 pg mL^{-1} [22]	>200 pg mL^{-1} [22]	<35 pg mL^{-1} [22]

Zamay TN, Zamay GS, Kolovskaya OS, et al. Current and Prospective Protein Biomarkers of Lung Cancer. Cancers. 2017;9(11):155. doi:10.3390/cancers9110155.

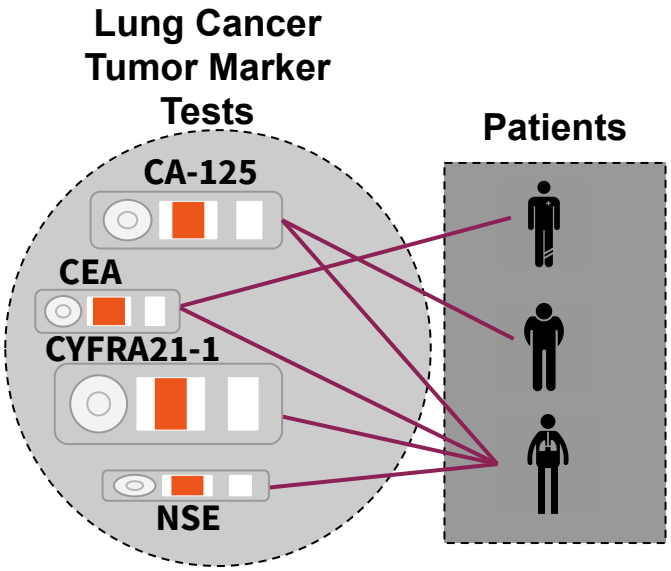
How can Knowledge Evolution help?



Level of the **Lung Cancer Biomarkers** in the patients with Lung Cancer?

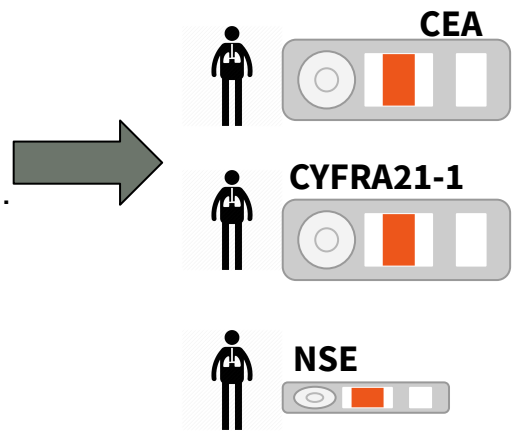
```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?id ?date ?level
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
  ?bm iasis:associated ?obs .
  ?bm  iasis:limit ?limit .
  ?obs iasis:level ?level .
  ?obs iasis:date ?date .
  ?obs iasis:patient ?id .
  ?id iasis:diagnostic iasis:LungCancer .
        FILTER (?level > ?limit)
  }
```

How can Knowledge Evolution help?

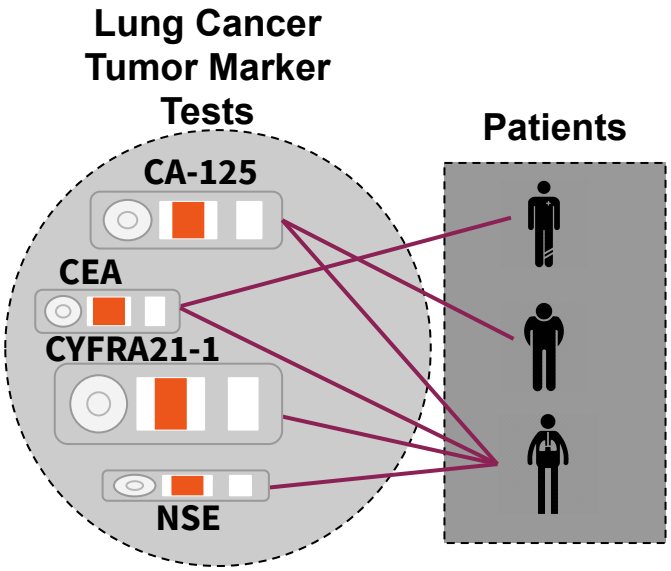


Level of the **Lung Cancer Biomarkers** in the patients with Lung Cancer?

```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?id ?date ?level
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
  ?bm iasis:associated ?obs .
  ?bm  iasis:limit ?limit .
  ?obs iasis:level ?level .
  ?obs iasis:date ?date .
  ?obs iasis:patient ?id .
  ?id iasis:diagnostic iasis:LungCancer .
        FILTER (?level > ?limit)
 }
```



How can Knowledge Evolution help?

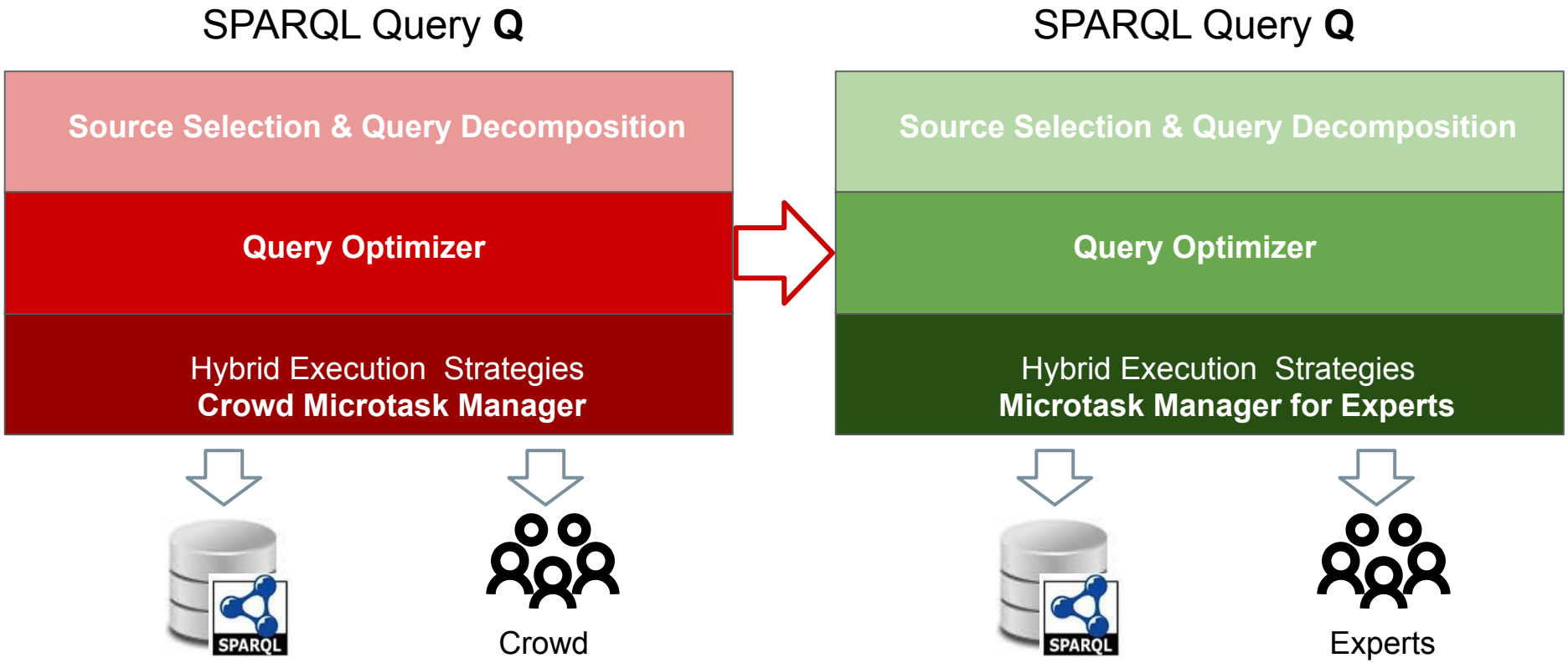


Level of the **Lung Cancer Biomarkers** in the patients with Lung Cancer?

```
PREFIX iasis:<http://iasis/vocab/>
SELECT ?id ?date ?level
 WHERE {
  ?bm  a  iasis:LungCancerBiomarker .
  ?bm iasis:associated ?obs .
  ?bm  iasis:limit ?limit .
  ?obs iasis:level ?level .
  ?obs iasis:date ?date .
  ?obs iasis:patient ?id .
  ?id iasis:diagnostic iasis:LungCancer .
        FILTER (?level > ?limit)
  }
```

EMPTY

Future Hybrid Federated Query Engines



Knowledge Completeness Evolution

Biomarkers associated with Brain Metastasis

- Ki-67 expression
- low caspase-3 expression
- high vascular endothelial growth factor C expression, and low E-cadherin expression

Knowledge Completeness Evolution

Biomarkers associated with Brain Metastasis

- Ki-67 expression
- low caspase-3 expression
- high vascular endothelial growth factor C expression, and low E-cadherin expression

Prediction Process

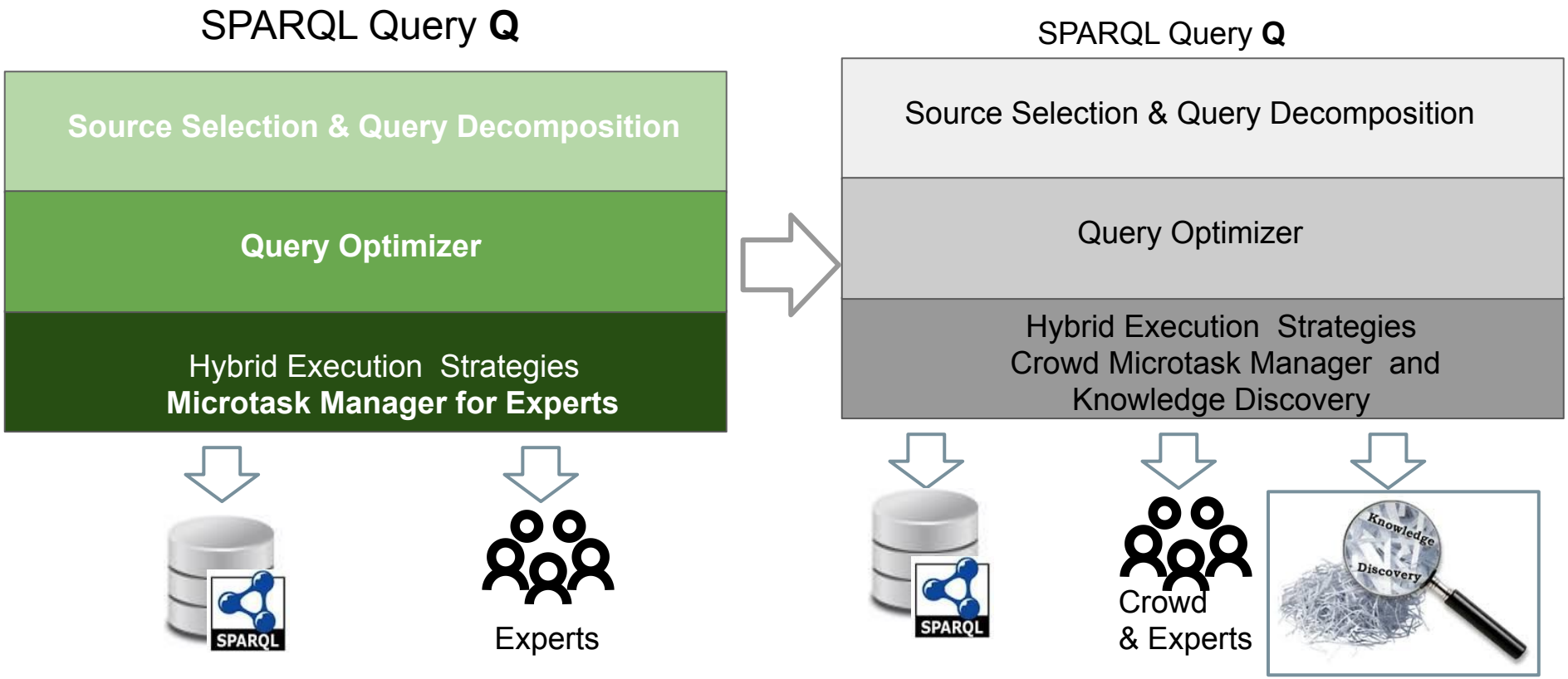
Prediction methods to determine “similar cancers” associated with the same biomarkers

- Non-small cell lung cancer (NSCLC)
- Breast cancer

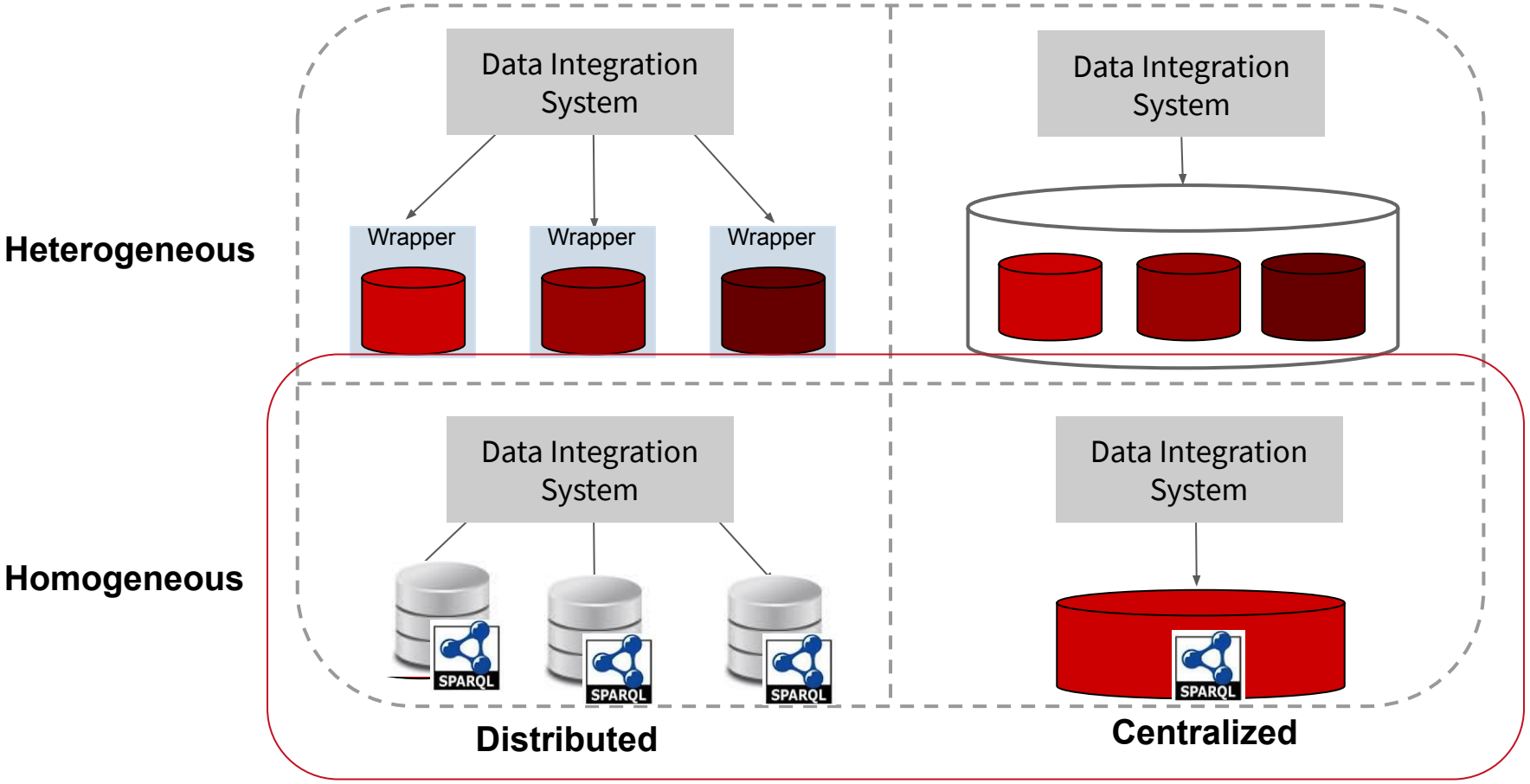
Examples of Predictions….

Prediction Task	Goal
Drug-Drug Interactions	Adverse Drug Events
Drug Side-Effect Interactions	Adverse Drug Reactions
Drug-Target Interactions	Drug Effectiveness
Disease Biomarkers	Disease Early Detection
Disease Mutations	Disease Early Detection and Drug Effectiveness

Future Hybrid Federated Engines

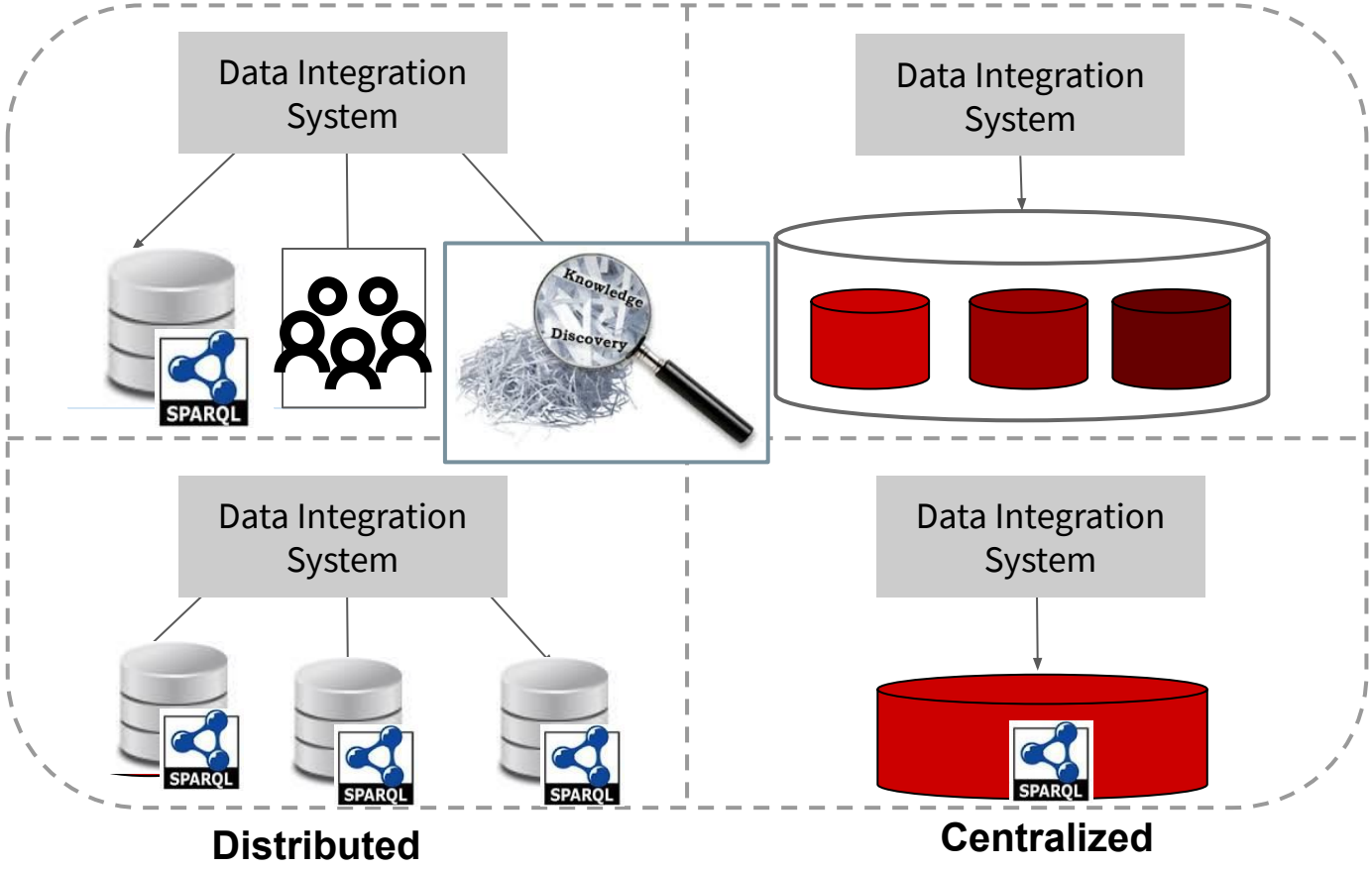


Data Integration Systems



Existing Approaches have focused on adaptive techniques to support SPARQL Query Processing over RDF Data Sources

Data Integration Systems



Future Approaches require to be focused on techniques to support data and knowledge evolution of RDF Data Sources

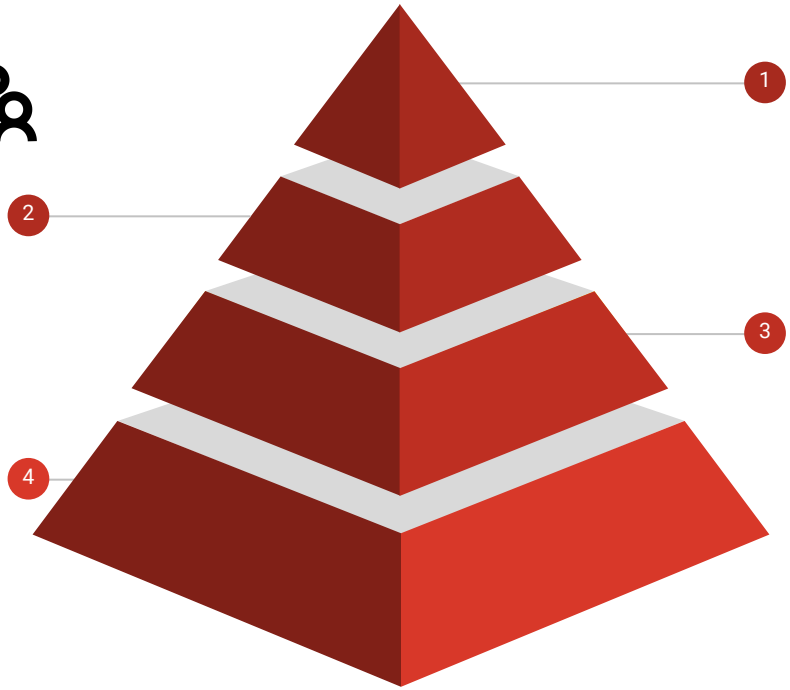
Future Hybrid Query Engines

Knowledge Prediction

Knowledge discovery techniques able to "***predict unknown facts"*** to complete RDF data sources..

Data Curation

Crowd based techniques able to exploit "***public domain"*** knowledge to complete RDF data sources.



Knowledge Curation

Crowd based techniques able to exploit "***specialized knowledge"*** to complete RDF data sources.

RDF Data Sources

Adaptive query processing techniques able to *adjust query execution* schedulers to current conditions of the data sources.


Leibniz
Universität
Hannover
Forschungszentrum · Research Center
L3S
LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY
TIB

Our Team at the Scientific Data Management Group

Prof, Dr.
Maria-Esther Vidal

PostDoc

Dr. Ingo Keck

Senior Researcher

Akhilesh Vyas

Research Assistants

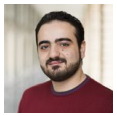
Samaneh Jozashoor

Ariam Rivas

Maria Isabel Castellanos

Kemele Endris

Farah Karim



Ahmad Sakor

Philipp Rohde

Visiting Researchers

Lucie-Aimée Kaffee

David Chaves

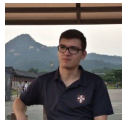
Master Research Assistants

Enrique Iglesias

Monica Figuera

Collaborators

Dr. Maribel Acosta



Dr. Michael Galkin

Dr.Diego Collarana

Dr. Irlan Grangle

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

TIB

Thank you!
Questions

Contact
Maria-Esther Vidal
Maria.Vidal@tib.eu

Leibniz Association

References

[1] Maribel Acosta, Maria-Esther Vidal, Tomas Lampo, Julio Castillo, Edna Ruckhaus: ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints. International Semantic Web Conference (2011)
[2] Maribel Acosta, Maria-Esther Vidal: Networks of Linked Data Eddies: An Adaptive Web Query Processing Engine for RDF Data. International Semantic Web Conference (2015)
[3] Olaf Görlitz, Steffen Staab: SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. COLD (2011)
[4] Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, Michael Schmidt: FedX: Optimization Techniques for Federated Query Processing on Linked Data. International Semantic Web Conference (2011)
[5] Gabriela Montoya, Hala Skaf-Molli, Pascal Molli, Maria-Esther Vidal: Decomposing federated queries in presence of replicated fragments. J. Web Sem. (2017)
[6] Gabriela Montoya, Hala Skaf-Molli, Pascal Molli, Maria-Esther Vidal: Federated SPARQL Queries Processing with Replicated Fragments. International Semantic Web Conference (2015)
[7] Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, Pieter Colpaert: Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. J. Web Sem.(2016)
[8] Maria-Esther Vidal, Simón Castillo, Maribel Acosta, Gabriela Montoya, Guillermo Palma: On the Selection of SPARQL Endpoints to Efficiently Execute Federated SPARQL Queries. Trans. Large-Scale Data- and Knowledge-Centered Systems 25: 109-149 (2016)

References

[9] Muhammad Saleem, Axel-Cyrille Ngonga Ngomo, Josiane Xavier Parreira, Helena F. Deus, Manfred Hauswirth: DAW: Duplicate-AWare Federated Query Processing over the Web of Data. International Semantic Web Conference (2013)
[10] Kemele M. Endris, Mikhail Galkin, Ioanna Lytra, Mohamed Nadjib Mami, Maria-Esther Vidal, Sören Auer: MULDER: Querying the Linked Data Web by Bridging RDF Molecule Templates. International Conference on Database and Expert Systems Applications (2017)
[11] Muhammad Saleem, Axel-Cyrille Ngonga Ngomo: HiBISCuS: Hypergraph-Based Source Selection for SPARQL Endpoint Federation. Extended Semantic Web Conference (2014)
[12] SemaGrow: Optimizing federated SPARQL queries Angelos Charalambidis, Antonis Troumpoukis and Stasinos Konstantopoulos In Proceedings of the 11th International Conference on Semantic Systems (SEMANTiCS 2015)
[13] Mikhail Galkin, Kemele M. Endris, Maribel Acosta, Diego Collarana, Maria-Esther Vidal, Sören Auer: SMJoin: A Multi-way Join Operator for SPARQL Queries. SEMANTICS 2017: 104-111
[14] Kemele M. Endris, Philipp D. Rohde, Maria-Esther Vidal, Sören Auer: Ontario: Federated Query Processing Against a Semantic Data Lake. DEXA 2019: 379-395
[15] Maribel Acosta, Maria-Esther Vidal, York Sure-Vetter: Diefficiency Metrics: Measuring the Continuous Efficiency of Query Processing Approaches. International Semantic Web Conference, 2017