LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

Forschungszentrum · Research Center L3S

TIB

PURE: A Privacy Aware Rule-Based Framework over Knowledge Graphs

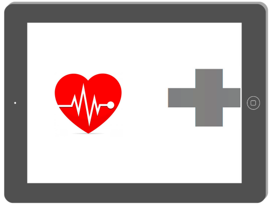
Marlene Goncalves, Maria-Esther Vidal, **Kemele M. Endris**

DEXA 2019 - **August 26 - 29, 2019**
Linz, Austria

Leibniz Universität Hannover

Leibniz Association

Motivation



Electronic
Health Records

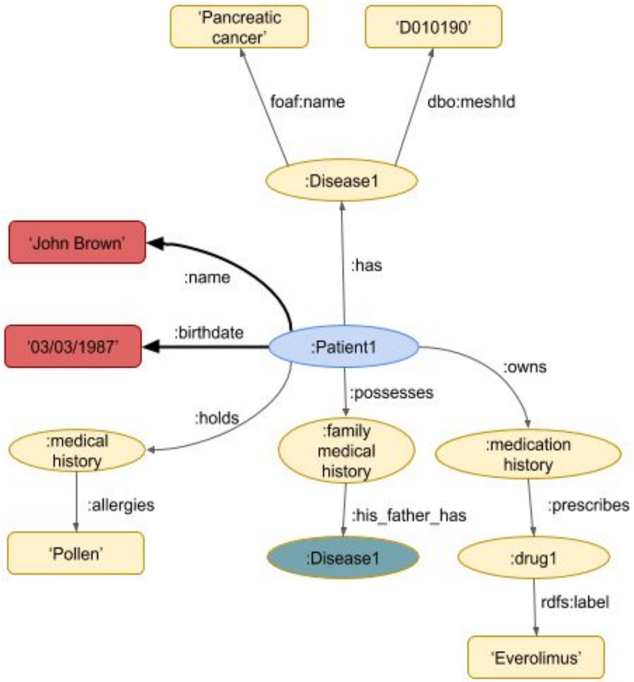
Pancreatic
Cancer

Patient

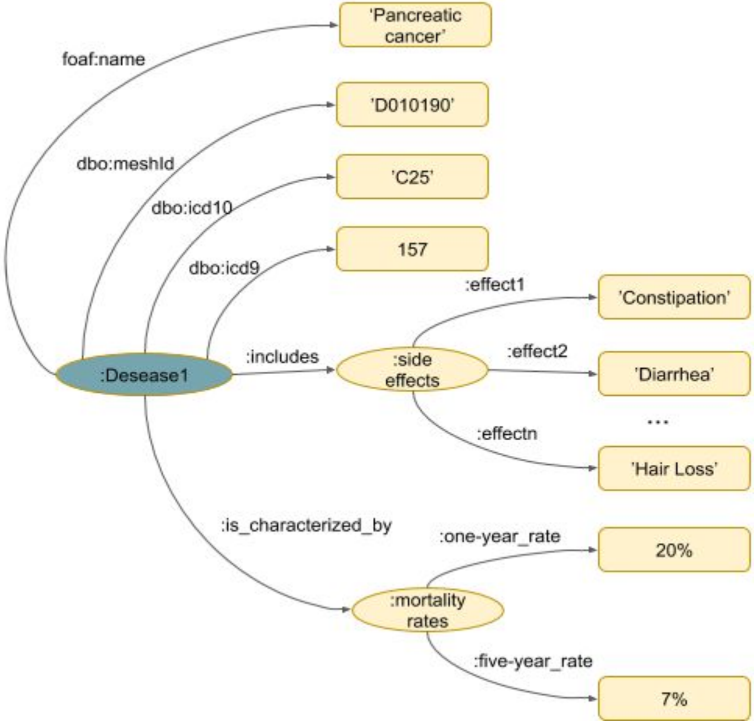
+ DBpedia +

RDF Graphs

Electronic Health Records (EHR)

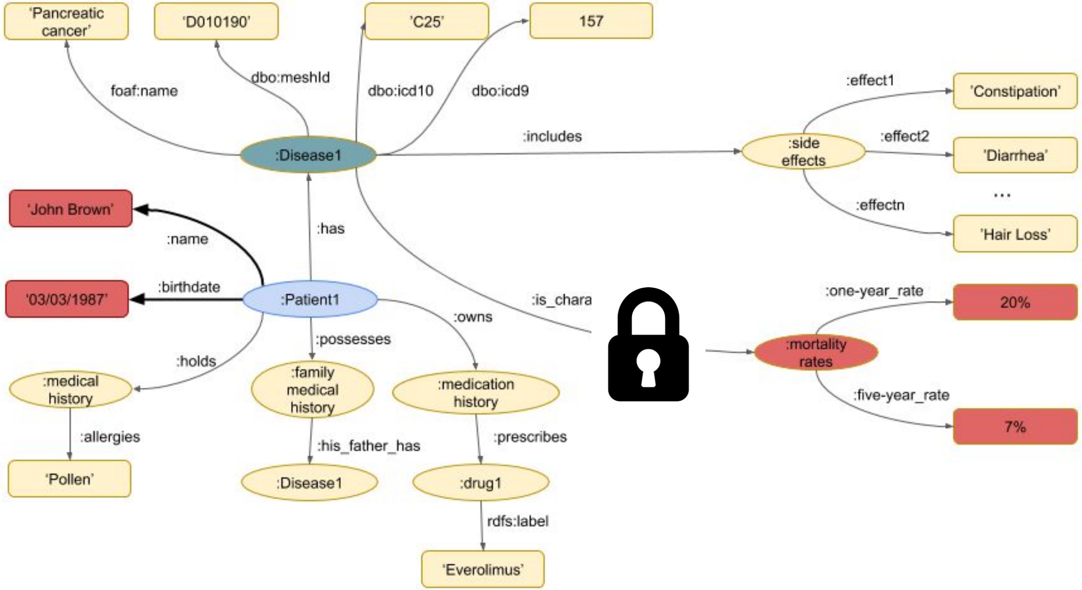


DBpedia



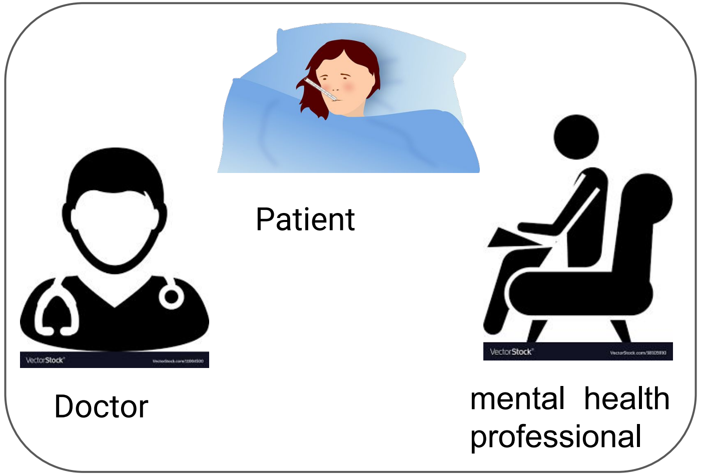
Data Control Access

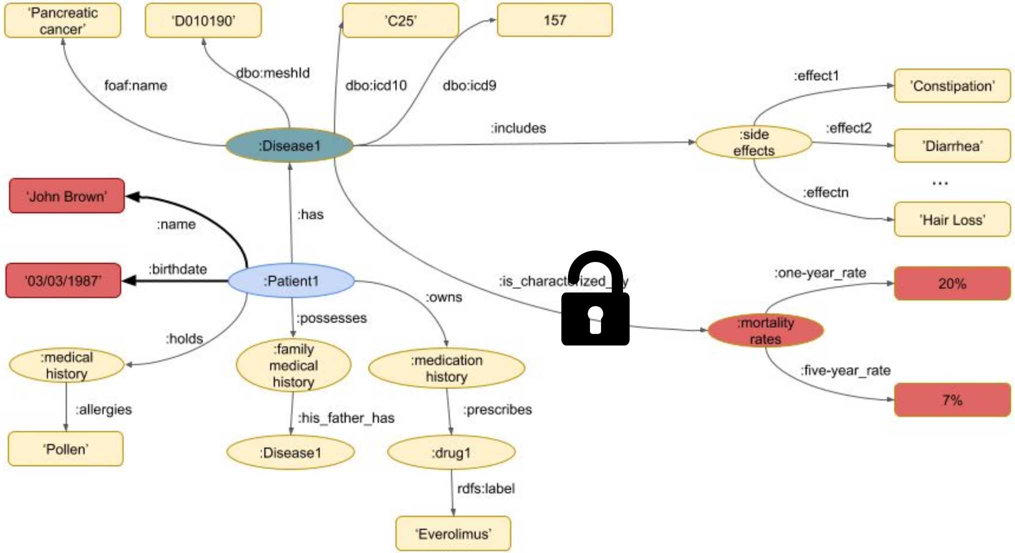
A join between EHR and DBpedia is possible, but mortality rates can not be easy to handle for a patient



Data Control Access

A **complete access** to information is granted when the patient is accompanied by the doctor and mental health professionals!





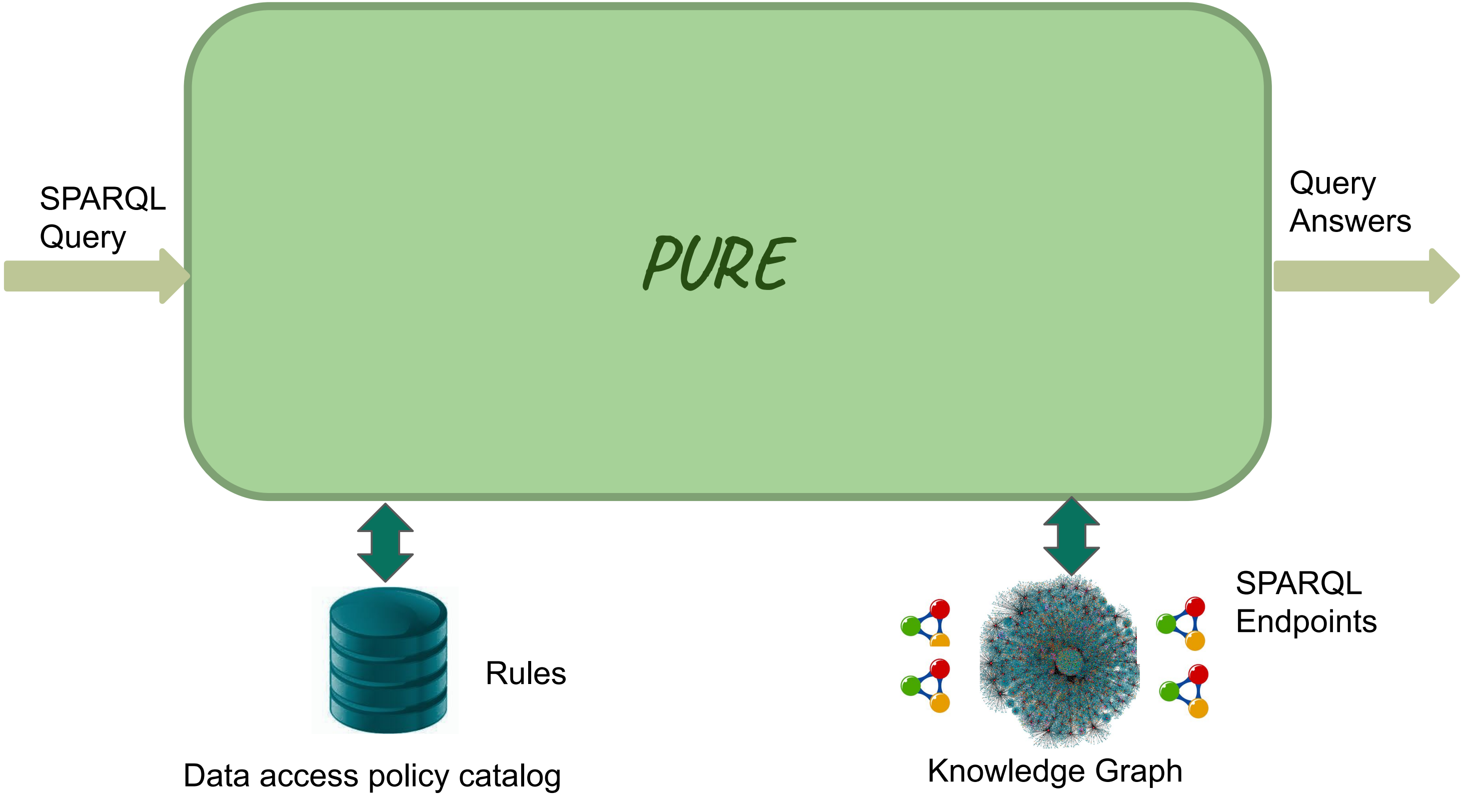
Agenda

1. Related Work
2. PURE: A Privacy Aware Rule-Based Framework
3. Empirical Evaluation
4. Conclusions and Future Work

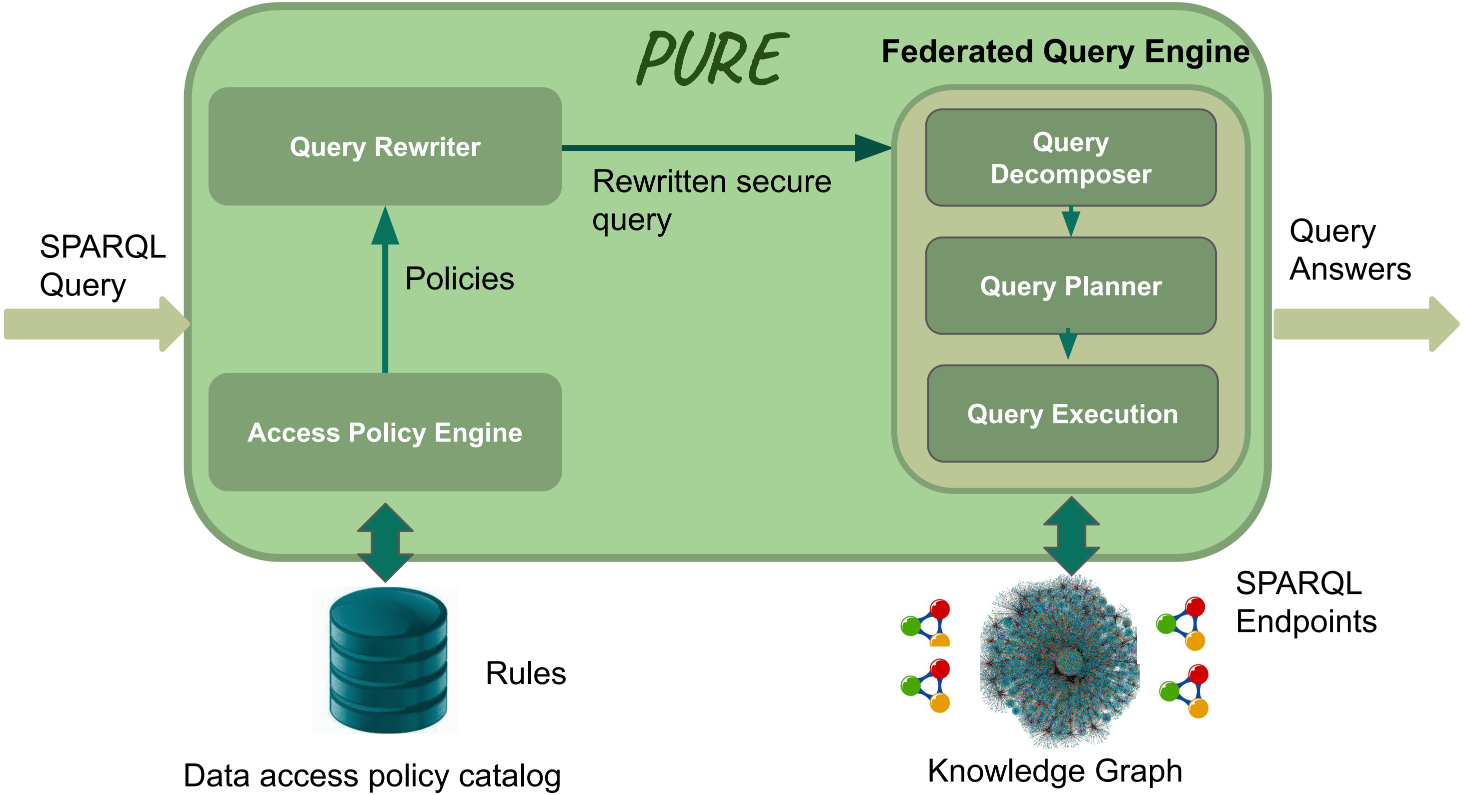
Related Work

- Access control ontologies for RDF data [Costabello & Villata & Gandon 2012; Unbehauen & Frommhold & Martin 2016]
- Access enforcement on centralized or distributed RDF stores [Amini & Jalili 2010] or federated RDF sources [Endris & Almhithawi & Lytra & Vidal & S. Auer 2018; Khan & Saleem & Mehdi & Hogan & Mehmood & Rebholz-Schuhmann & R. Sahay 2017].

PURE: Architecture



PURE: Architecture



PURE: Our Approach

The problem of enforcing data privacy and access regulations (**EDPR**) consists of:

- A **vocabulary** V
- A set of **secrecies** SS
- A user **query** Q over concepts of V

A query Q should be rewritten to secure query Q' if at least one secrecy s in SS is revealed!

PURE: Vocabulary

Some examples:

- **patient(Name,Birthdate,Zip,Gender)**
- **disease(Code,Name)**
- **has(X,Y)**
- **etc.**

PURE: Secrecies

- Access policies are expressed using **rules or assertions**.
- For each rule on a **secrecy S_i^{si}**, there is a **mapping** that describes **S_i^{si}** as **a conjunctive query** (**Local-As-View approach**).

e.g. S_1(X,W):-has(X,Y),disease(Y,'PAC'),
is_characterized_by(Y,Z),five_year_rate(Z,W)

where PAC=Pancreatic Cancer

PURE: Queries

- A user query **Q** is a **conjunctive query** in terms of vocabulary concepts.
- Q is **insecure** if there is at least **one insecure rewriting Q’ of Q** with respect to secrecies S.

```
e.g. Q(N,B,Z,G,R):-patient(N,B,Z,G),has(N,Y),disease(Y,'PAC'),
                   is_characterized_by(Y,Z),five_year_rate(Z,R).
```

PURE: Queries

- The query **Q** is **insecure** because there is **one insecure rewriting Q’ of Q** with respect to secrecies S

S_1(X,W):-**has(X,Y),disease(Y,'PAC'),**
is_characterized_by(Y,Z),five_year_rate(Z,W)

Q(N,B,Z,G,R):-patient(N,B,Z,G),**has(N,Y),disease(Y,'PAC'),**
is_characterized_by(Y,Z),five_year_rate(Z,R).

Q can be rewritten using the secrecy S_1:

Q’(N,B,Z,G,R):-patient(N,B,Z,G),**S_1(N,R)**.

PURE: Queries

- The query **Q** is **secure** because there is **one secure rewriting Q’ of Q** with respect to secrecies S

S_1(X,W):-**has(X,Y),disease(Y,'PAC'),**
is_characterized_by(Y,Z),five_year_rate(Z,W)

Q(N,B,Z,G,R):-patient(N,B,Z,G),**has(N,Y),disease(Y,'Cancer'),**
is_characterized_by(Y,Z),five_year_rate(Z,R).

Q can be safely evaluated following a secure rewriting!

Empirical Evaluation

Experimental Setup

The Berlin SPARQL Benchmark (BSBM):

- 200M triples
- 12 queries
- Query Q3 of BSBM was omitted because it is a union query.

The set of rules:

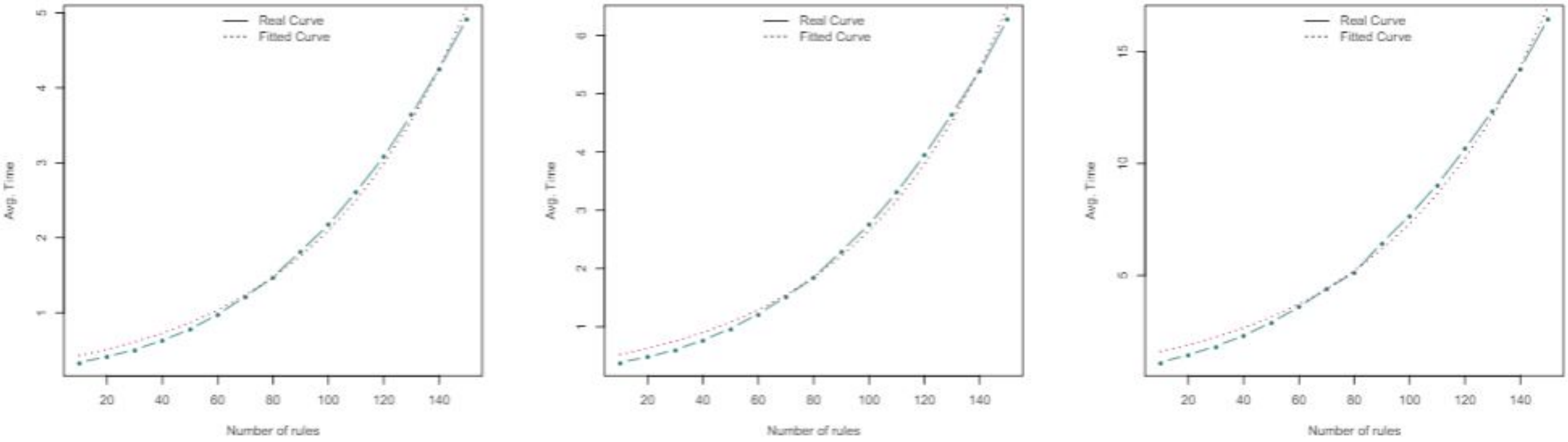
- Randomly generated
- Each one corresponds to a star-join between 1 and 3 predicates.
- The number of rules varies from 10 to 150

MCDSAT: https://github.com/bonetblai/mcdsat

Experiment 1

- Goal: Assess Impact of Number of Rules on Total Execution Time
- Metrics:
    - **Query execution time**: elapsed time in seconds between the submission of a query and the delivery of the answers
    - For each query, total execution time is measured for several configurations of number of rules from 10 to 150

Experiment 1: Impact of Number of Rules

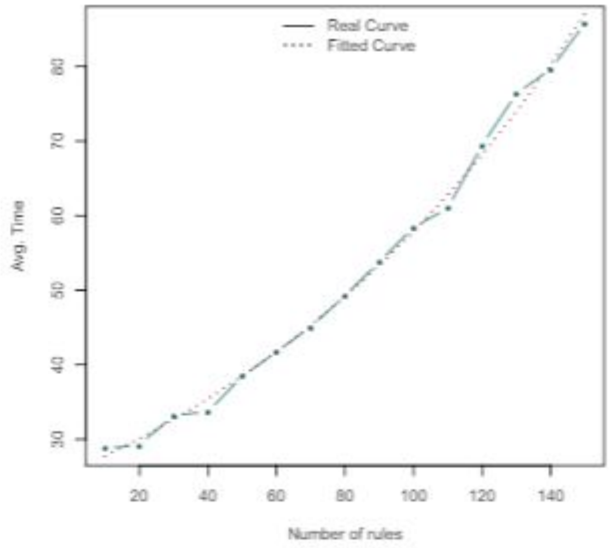


(a) Q8

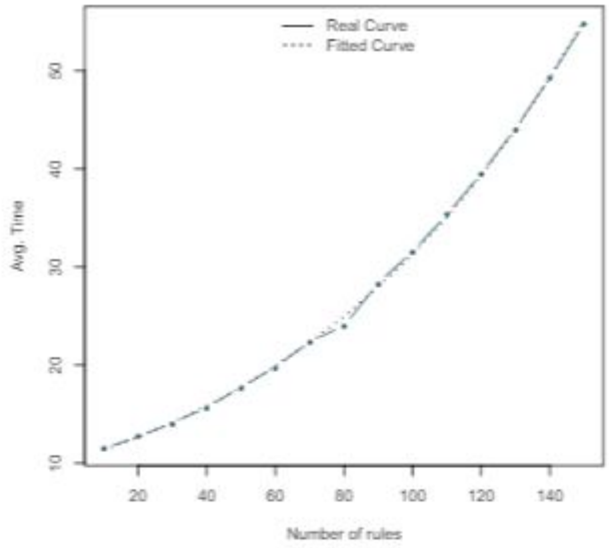
(b) Q9

(c) Q10

Experiment 1: Impact of Number of Rules



(d) Q11



(e) Q12

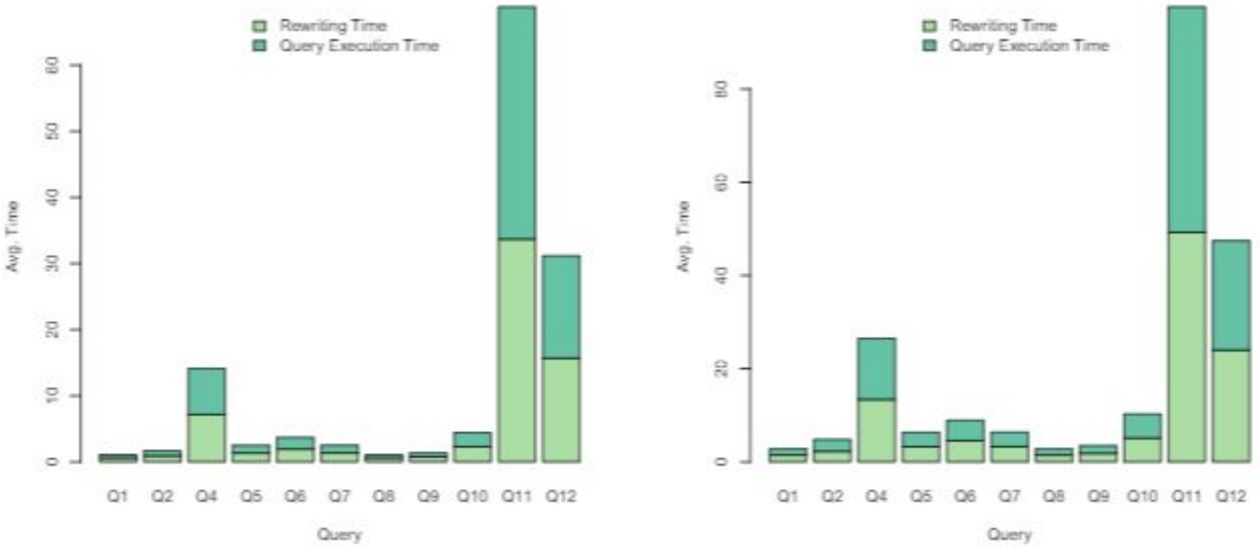
Execution time **grows exponentially** as the number of rules increases

Number of rewritings **blows up exponentially** with the number of views (rules)

Experiment 2

- Goal: Assess Impact of Privacy Validation on Total Execution Time
- Metrics:
  - **Query execution time**: elapsed time in seconds between the submission of a query and the delivery of the answers
  - *Rewriting time* and *Query execution time* are measured for several configurations of number of rules: 40, 80, 120, 150 (time in seconds).

Experiment 2: Privacy Validation on Total Execution Time

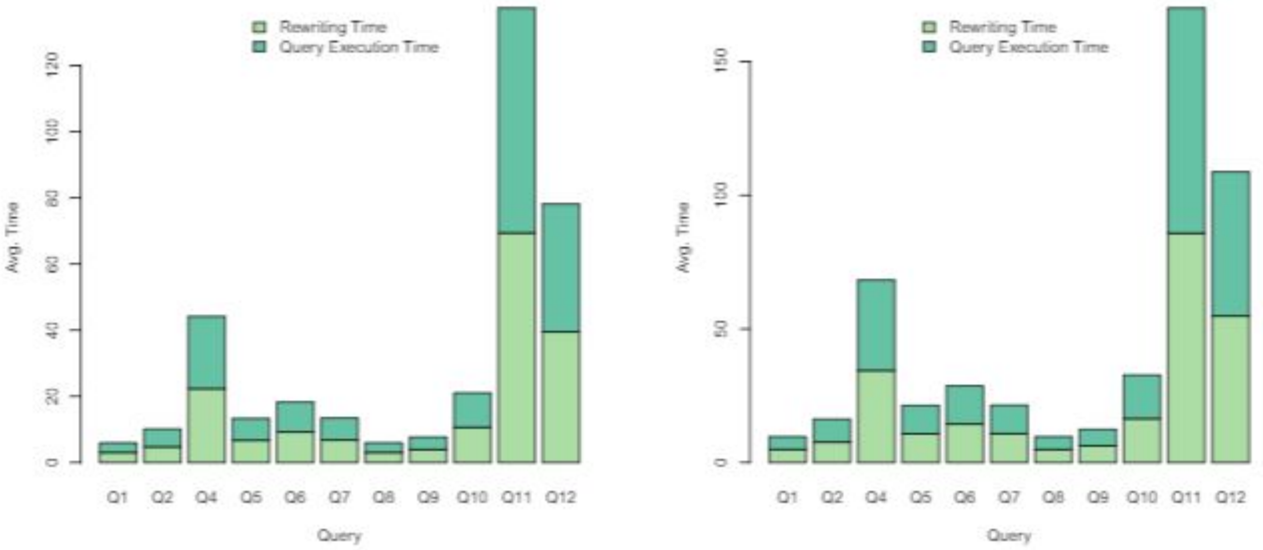


(a) 40 rules

(b) 80 rules

Average query rewriting time is **approximately 50%** of average total execution time

Experiment 2: Privacy Validation on Total Execution Time



(c) 120 rules

(d) 150 rules

Conclusions and Lessons Learned

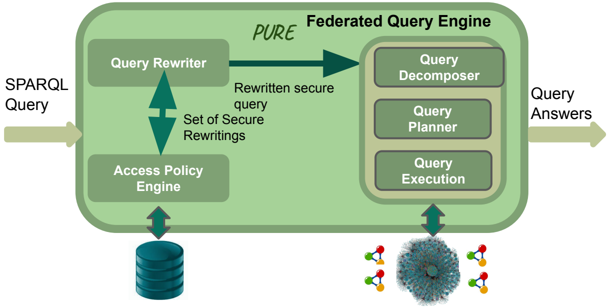
Data privacy policies can be described in terms of **LAV rules**

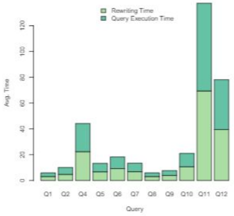
PURE is a privacy-aware rule-based federated query engine

Enforcing data privacy and access control is **costly**

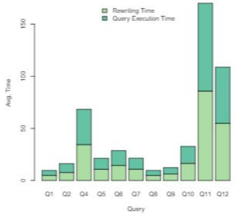
Execution time **grows exponentially** as the number of rules increases

Rewriting time is **approximately 50%** of average total execution time

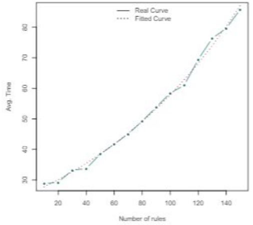




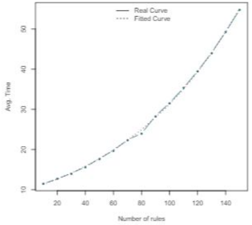
(c) 120 rules



(d) 150 rules



(d) Q11



(e) Q12

Thank You!