

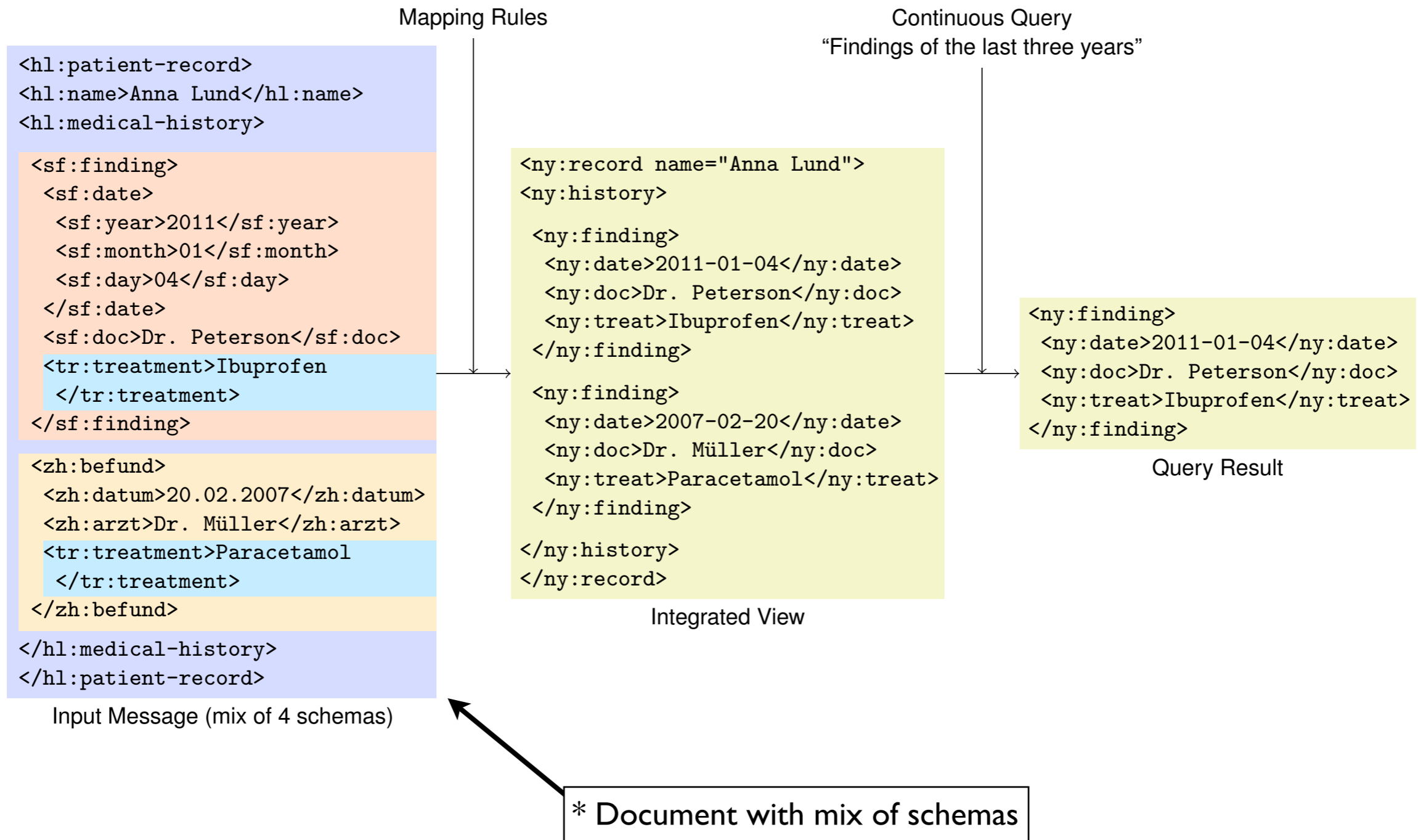
Data Integration Approaches for Highly Heterogenous Data*

Data Warehousing 2011
Exercise 6

Martin Hentschel

* Documents with mixes of schemas

Example



Different Approaches

1. Transformation

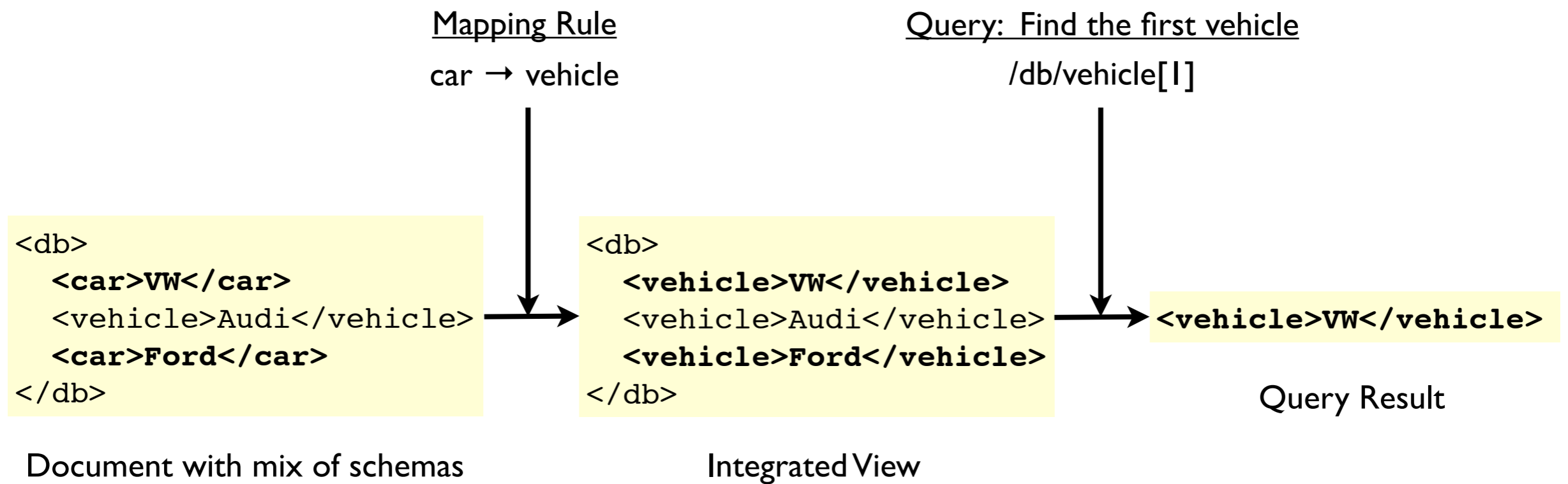
- Transform the document to match your query

2. Query Rewrite

- Transform (rewrite) your query to match the document

3. Any other idea?

Simpler Example



I. Transformation Approach

Transform & materialize

```
<db>  
  <vehicle>VW</vehicle>  
  <vehicle>Audi</vehicle>  
  <vehicle>Ford</vehicle>  
</db>
```

Use original query

```
Query: Find the first vehicle  
/db/vehicle[1]
```

Result

```
<vehicle>VW</vehicle>
```

2. Query Rewrite Approach

Use original data

```
<db>  
  <car>VW</car>  
  <vehicle>Audi</vehicle>  
  <car>Ford</car>  
</db>
```

Rewrite query

```
Query: Find the first vehicle  
/db/(vehicle | car)[1]
```

Result

```
<car>VW</car>
```

Union expression



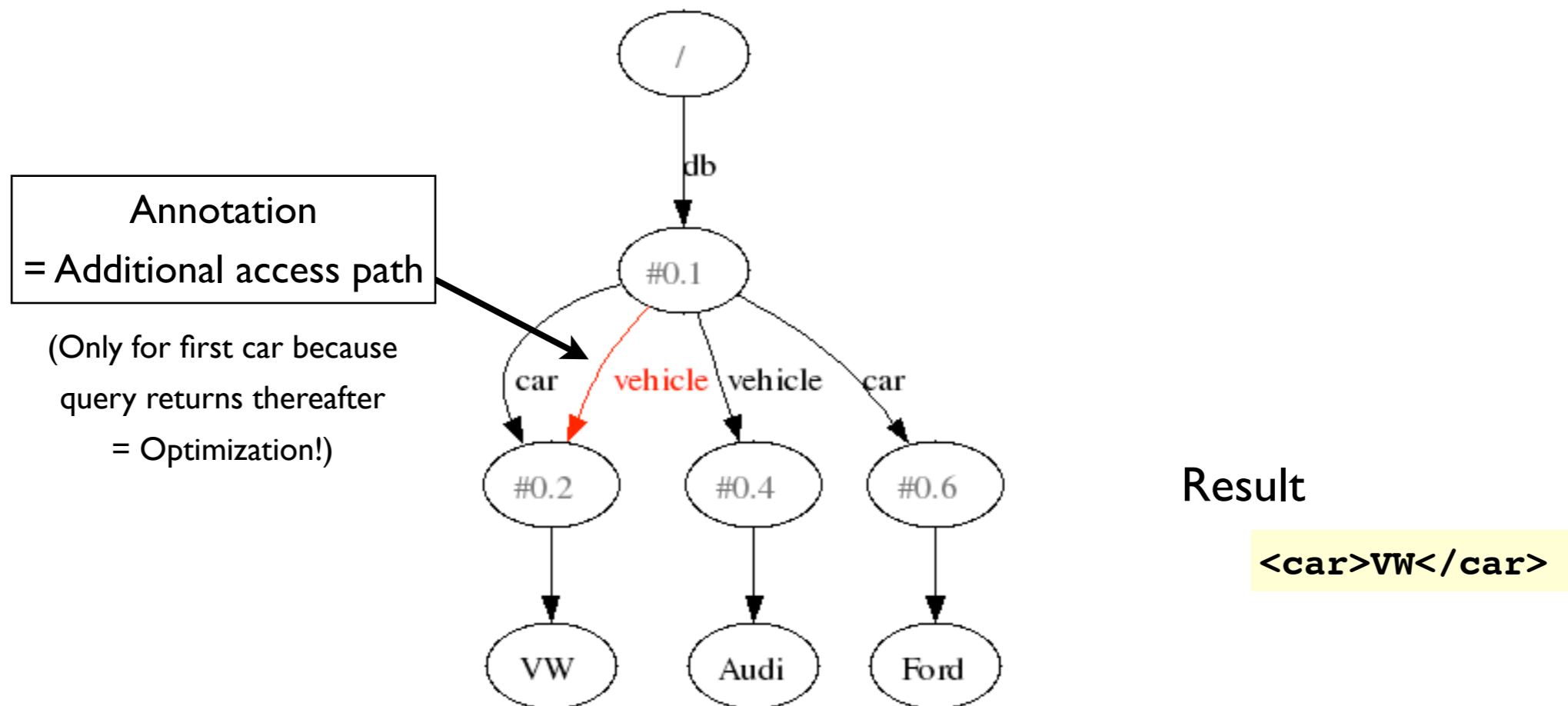
3. New Approach: Mapping Data to Queries

Use original data & original query

```
<db>  
  <car>VW</car>  
  <vehicle>Audi</vehicle>  
  <car>Ford</car>  
</db>
```

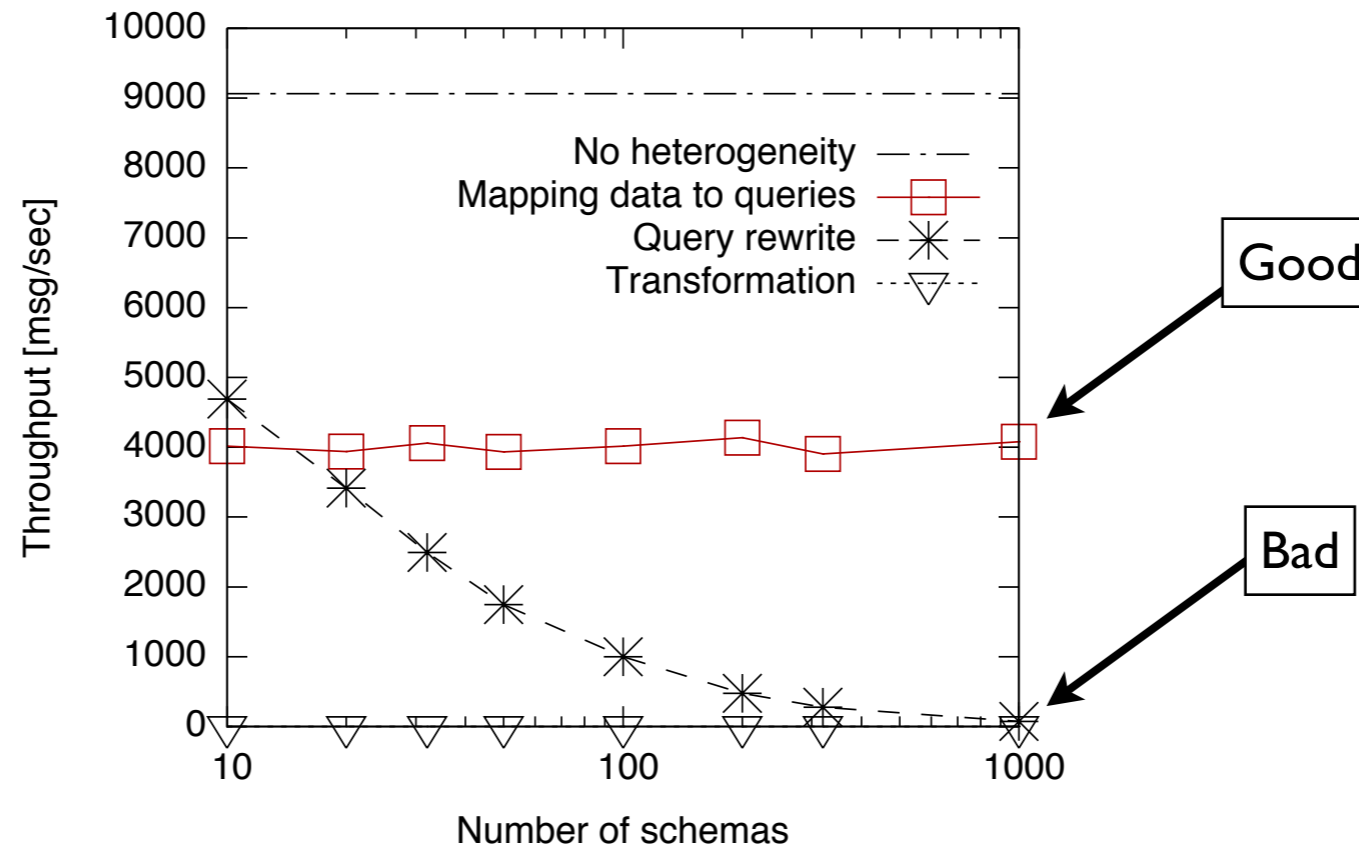
Query: Find the first vehicle
/db/vehicle[1]

Integrate data at runtime using annotations



Experimental Comparison

100x performance improvements when scaling schemas



Throughput, XMark Benchmark Query I, Mixed Schemas

Mapping Data to Queries

Demo

Exercise

Use example health record

- XML with mixed schemas

Implement queries using integrated view

- Return the findings of the last three years
- Return the doctor of the oldest finding
- List all treatments the patient has ever received

Execute queries on health record

- Using transformation approach
- Using query rewrite approach

Lesson to learn: Good & bad features of these approaches