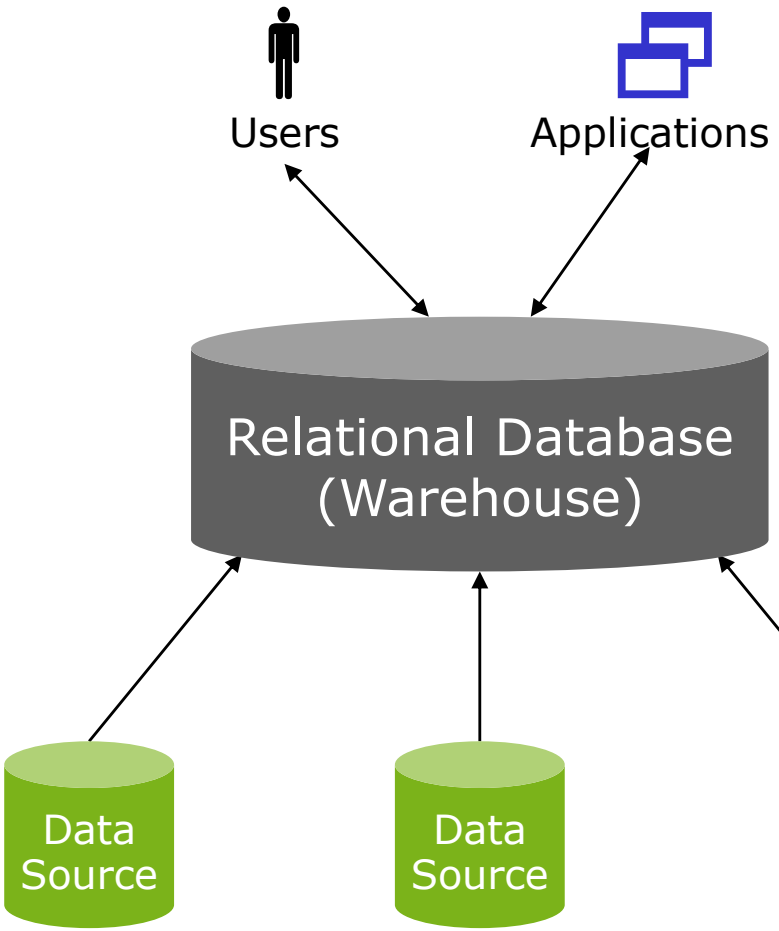
Data Integration Systems: An Introduction

Yannis Velegrakis

Data Warehouse Architecture



OLAP / Decision Support
Data Cubes / Data Mining

ETL Tools
(Extract-Transform-Load)

Data Cleaning

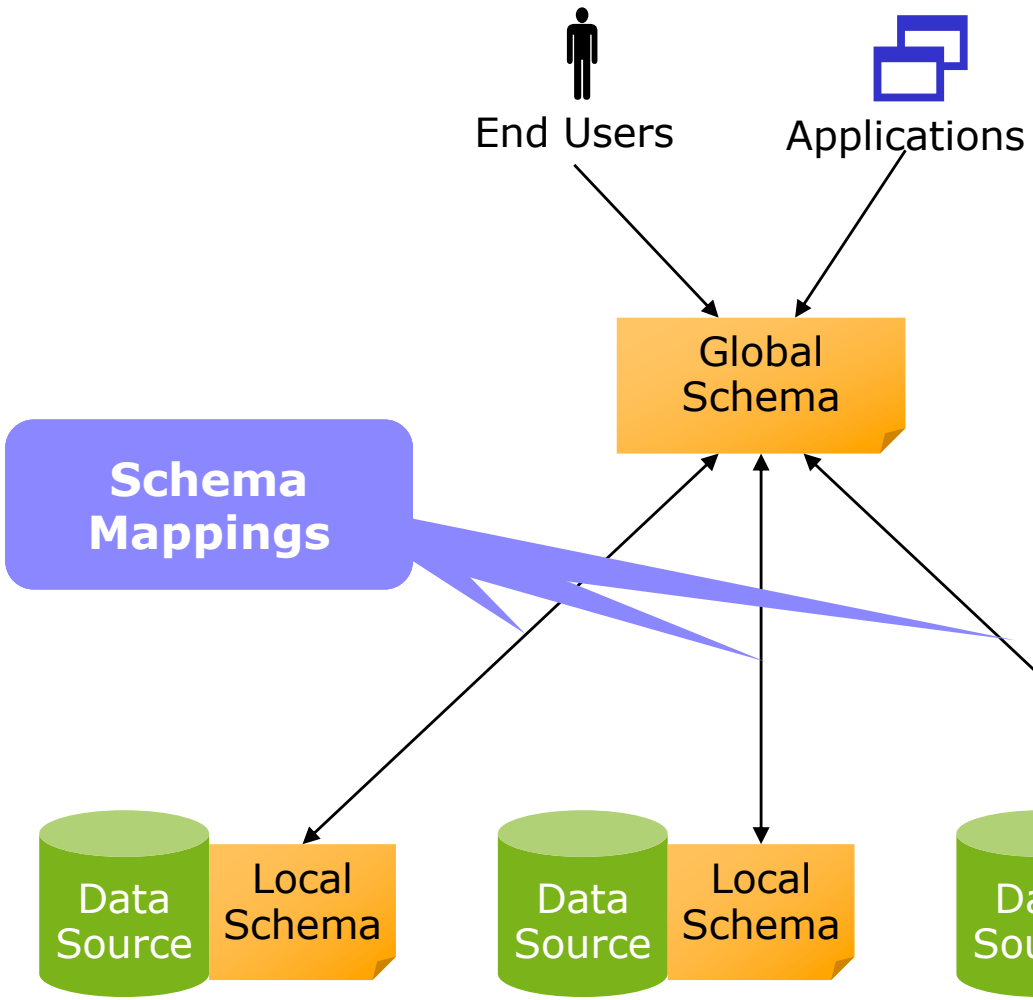
Virtual Integration Architecture

- Leave the data in the sources
- When a query comes in:
  - Determine the relevant sources to the query
  - Break down the query into sub-queries for the sources
  - Get the answers from the sources, filter them if needed and combine them appropriately
- Data is fresh
- Otherwise known as

On Demand Integration

Virtual Integration Architecture

Design-Time

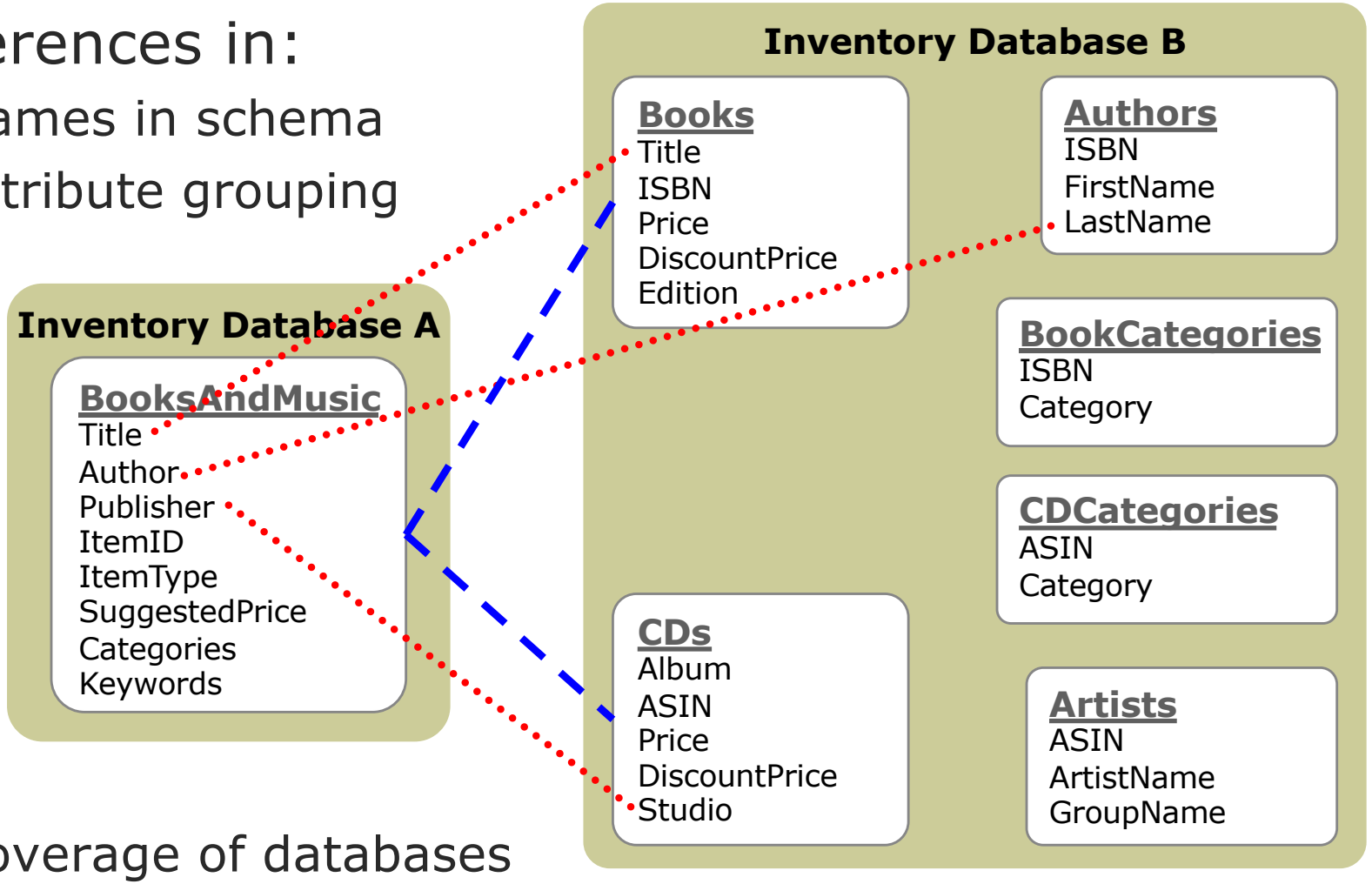


Sources can be:

- Relational DBs
- Excel Files
- Web Sites
- Web Services

Schema Mappings

- Differences in:
  - Names in schema
  - Attribute grouping



  - Coverage of databases
  - Granularity and format of attributes

Issues for Query Processing

Reformulation

Global Schema

Books
Title
ISBN
Price
DiscountPrice
Edition

Local Schema A

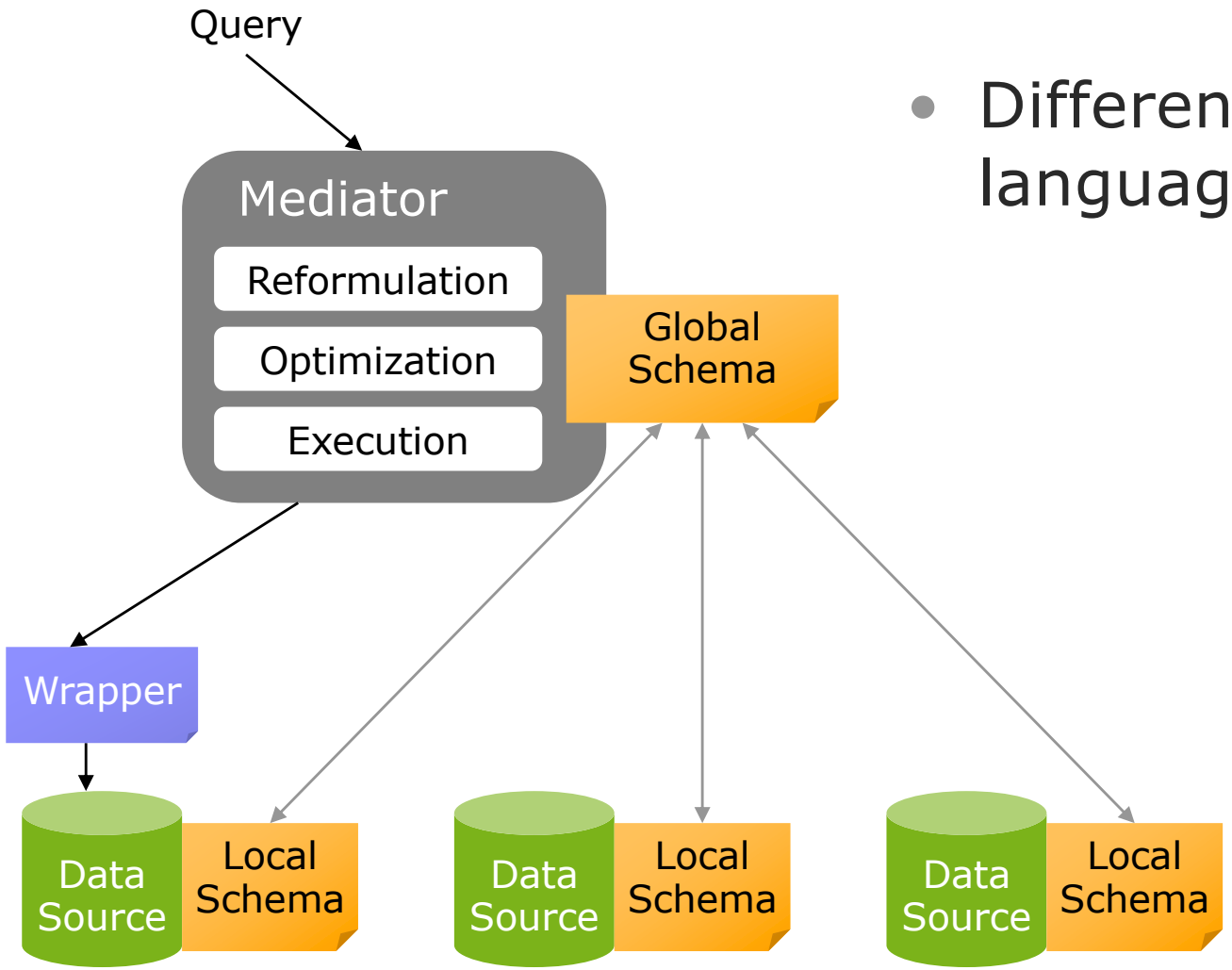
BooksAndMusic
Title
Author
Publisher
ItemID
ItemType
SuggestedPrice
Categories
Keywords

```
SELECT ISBN, Price
FROM Books
WHERE Title = 'on the road'
```

```
SELECT ItemID, SuggestedPrice
FROM BooksAndMusic
WHERE Title = 'on the road'
AND ItemType = 'Books'
```

Issues for Query Processing

Query Translation



- Different query languages

Issues for Query Processing

Query Translation

Global Schema

Books
Title
ISBN
Price
DiscountPrice
Edition

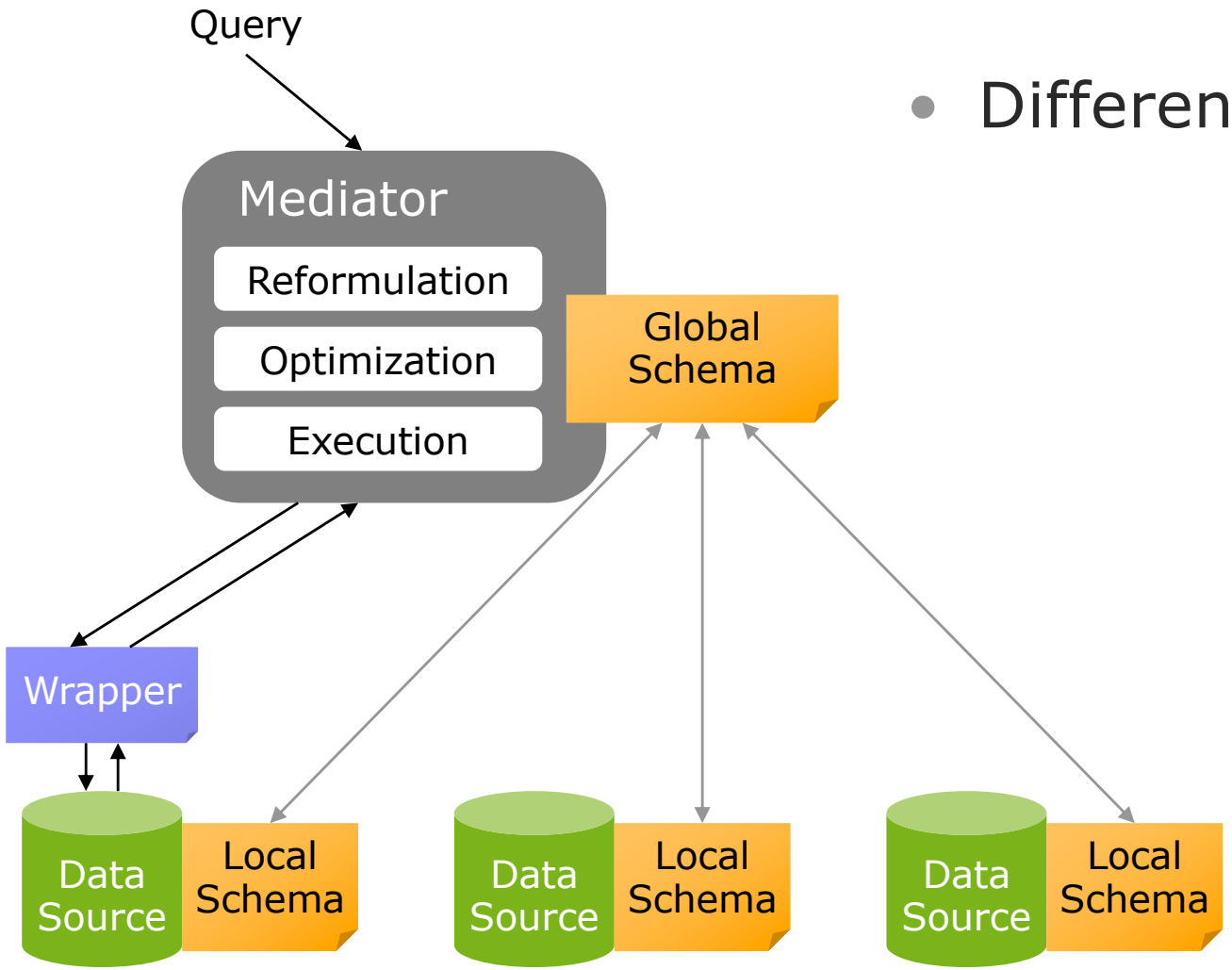
```
SELECT ISBN, Price
FROM Books
WHERE Title = 'on the road'
```

Local Source A

http://www.amazon.com/homepage.html?ItemType=Books&Title=on+the+road

Issues for Query Processing

Data Translation



- Different data models

Issues for Query Processing

Data Translation

Global Schema

Books
Title
ISBN
Price
DiscountPrice
Edition

Title	ISBN	Price	…	…
On the Road	123	10.86	…	…

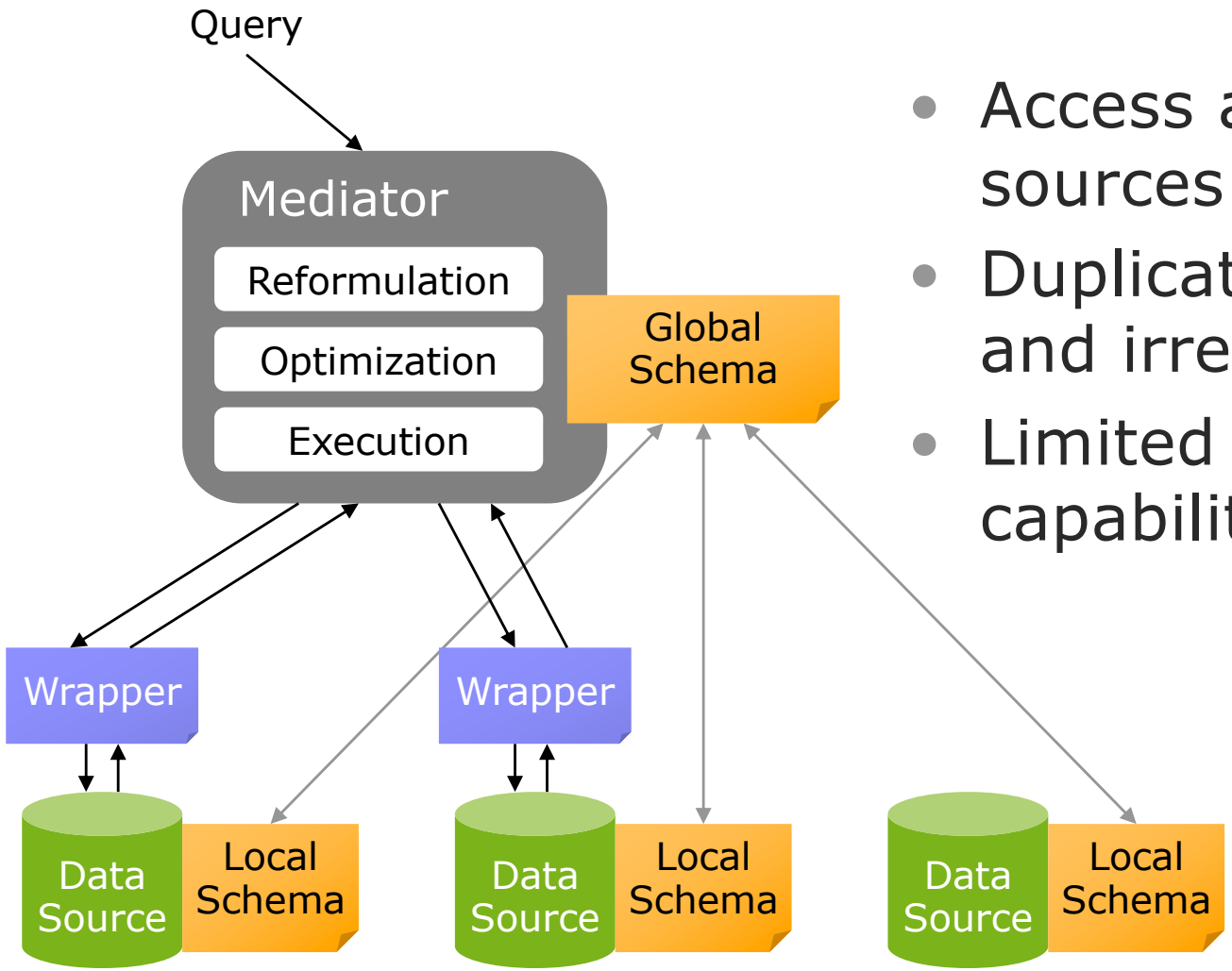
Local Result A

On the Road -- by Jack Kerouac; Paperback (Rate it)
Buy new: $10.86 -- Used & new from: $5.90

```
<table>
  <tr>
    <td>
      <a href=/details?isbn=123>
        <b>On the Road</b>
      </a>
      -- by Jack Kerouac; Paperback
      <br>
      <a href=/details?isbn=123>
        Buy new
      </a>
      :<b class=price>$10.86</b>
    </td>
  </tr>
</table>
```

Issues for Query Processing

Query Execution

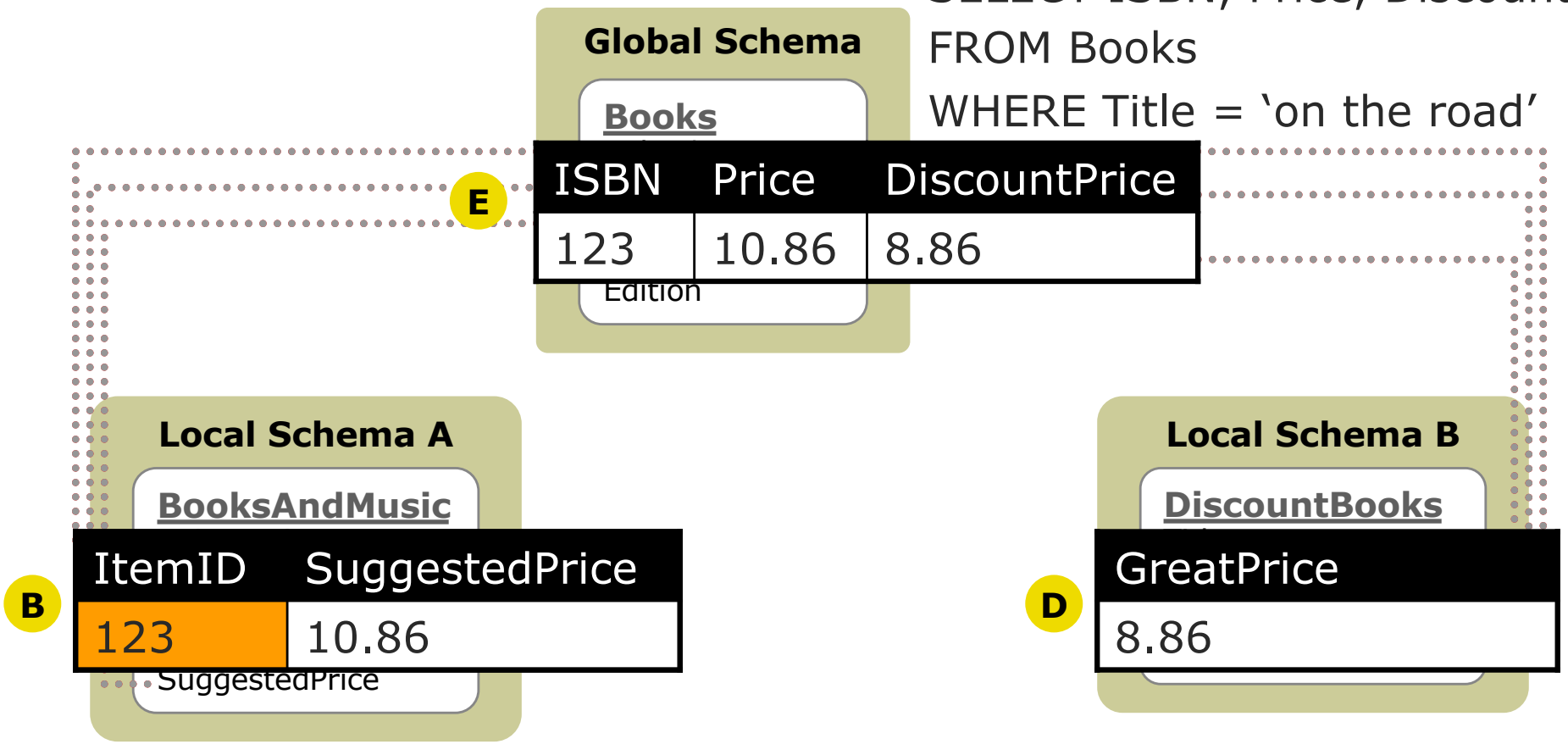


- Access as many data sources as needed
- Duplicate/redundant and irrelevant data
- Limited query capabilities

Issues for Query Processing

Limited Query Capabilities

SELECT ISBN, Price, DiscountPrice
FROM Books
WHERE Title = 'on the road'

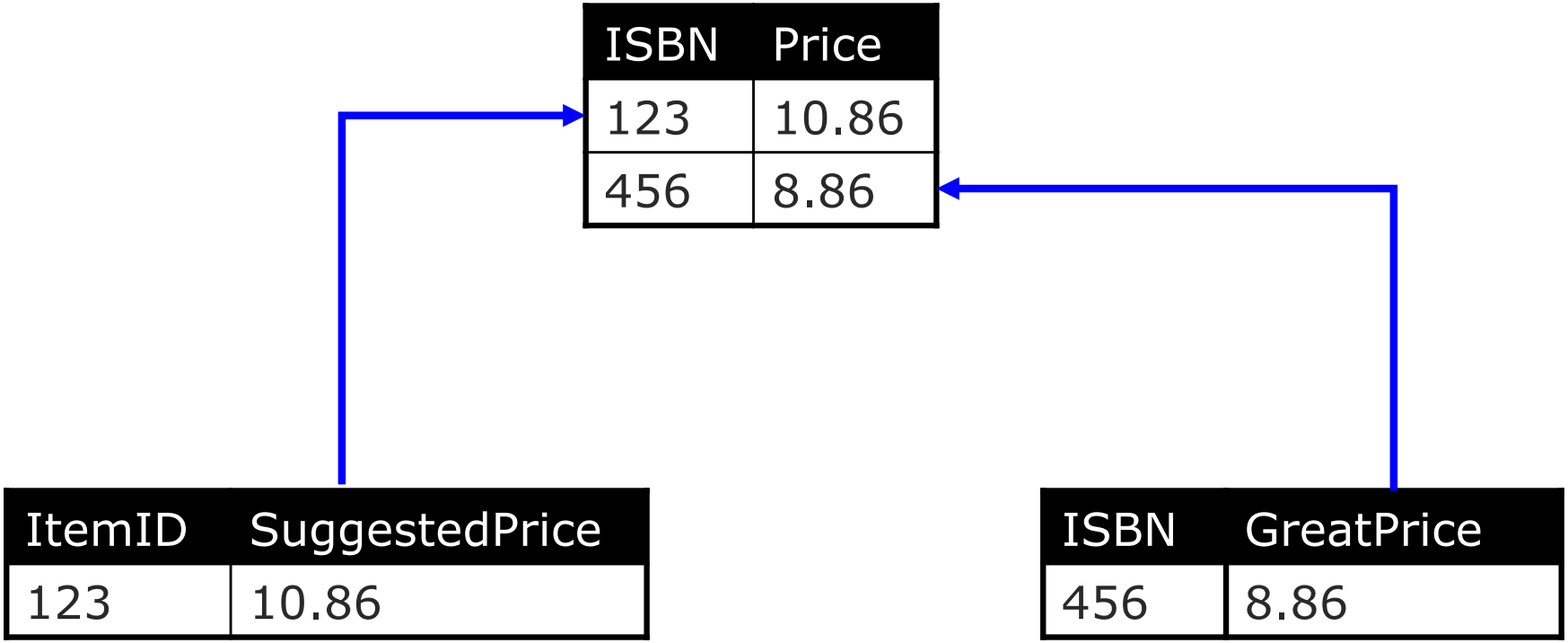


A SELECT ItemID, SuggestedPrice
FROM BooksAndMusic
WHERE Title = 'on the road'

C SELECT GreatPrice
FROM DiscountBooks
WHERE ISBN = 123

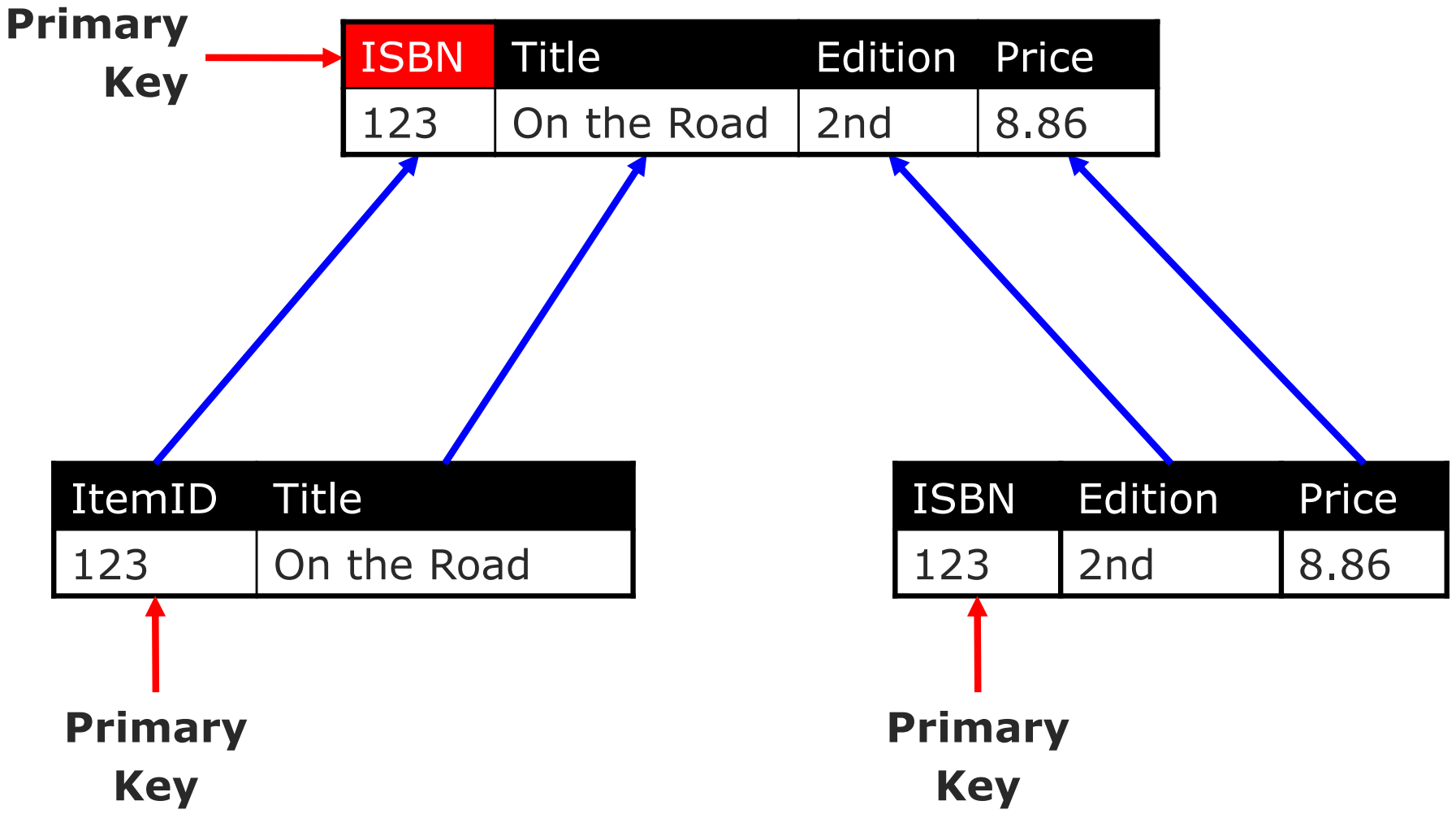
Issues for Query Processing

Query Answering (Union)



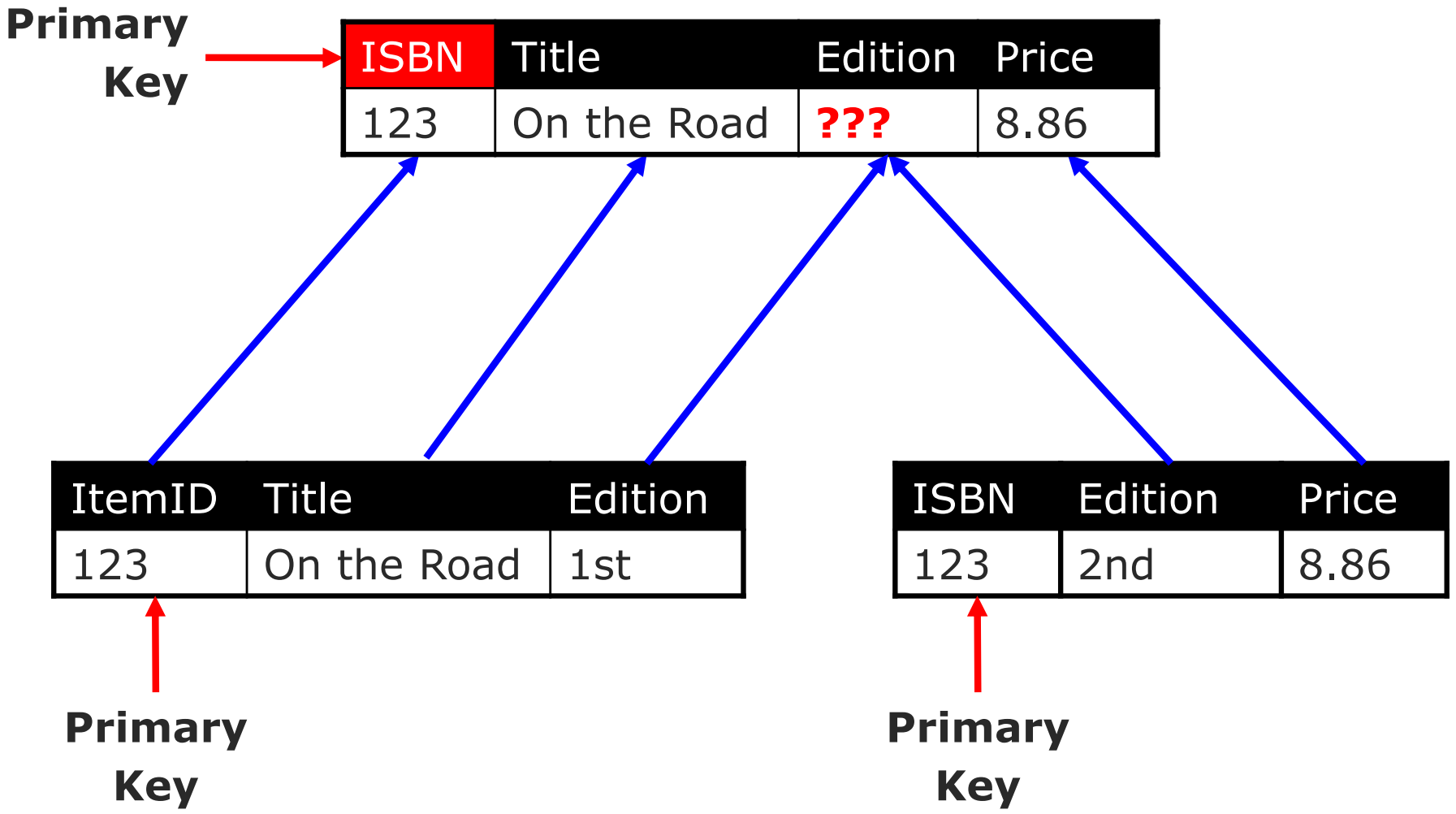
Issues for Query Processing

Query Answering (Merge)



Issues for Query Processing

Query Answering (Inconsistencies)



References for more info

- **Information integration**
  - Maurizio Lenzerini
    - *Eighteenth International Joint Conference on Artificial Intelligence, IJCAI 2003*
    - *Invited Tutorial*
- **Data Integration: a Status Report**
  - Alon Halevy
    - *German Database Conference (BTW), 2003*
    - *Invited Talk*

Federated Learning