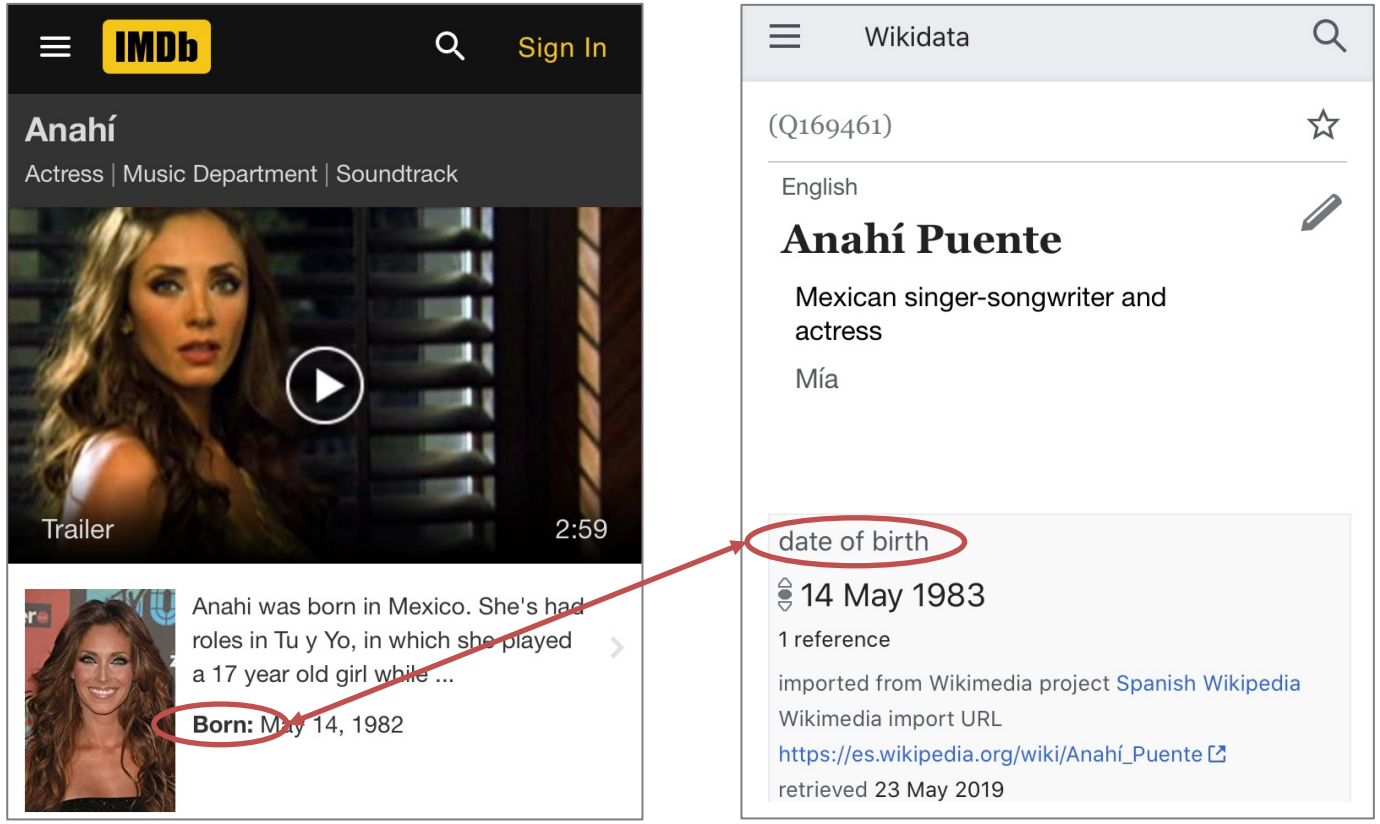
Background – Heterogeneous Data Integration

Schema Alignment

Align schemata of different sources and map to one another the attributes that have the same semantics

Entity Resolution

Data Fusion



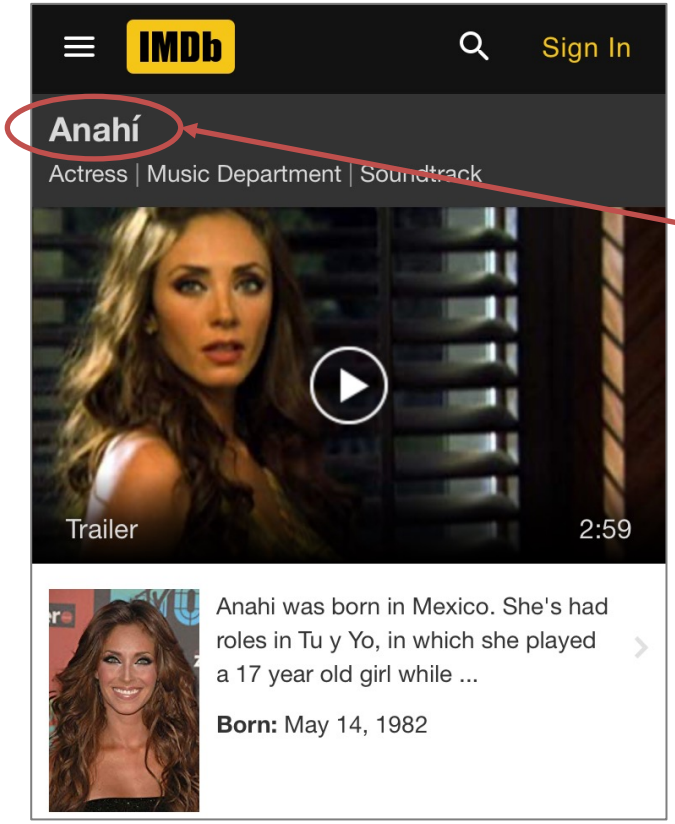
Background – Heterogeneous Data Integration

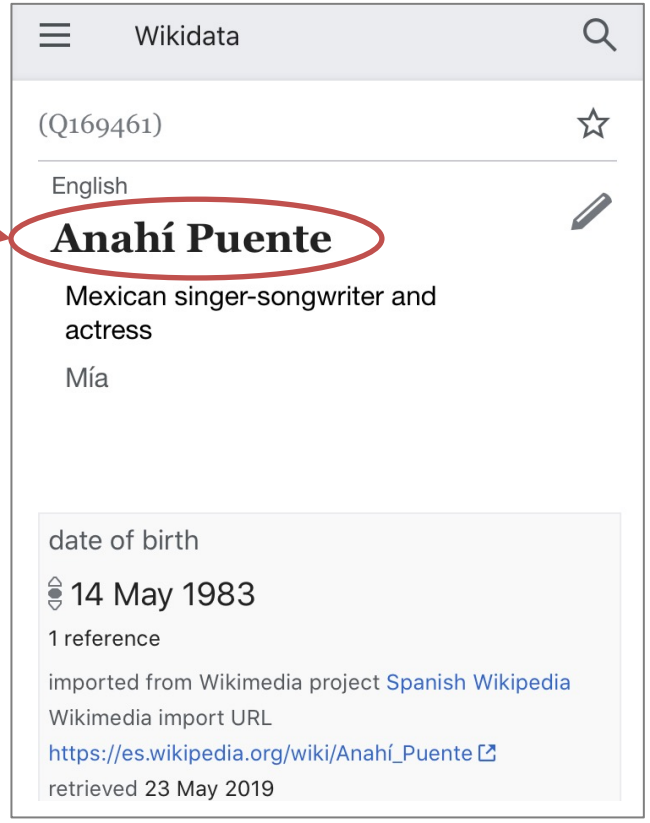
Schema Alignment

Entity Resolution

Find across the data sources the matching records that represent the same entities

Data Fusion





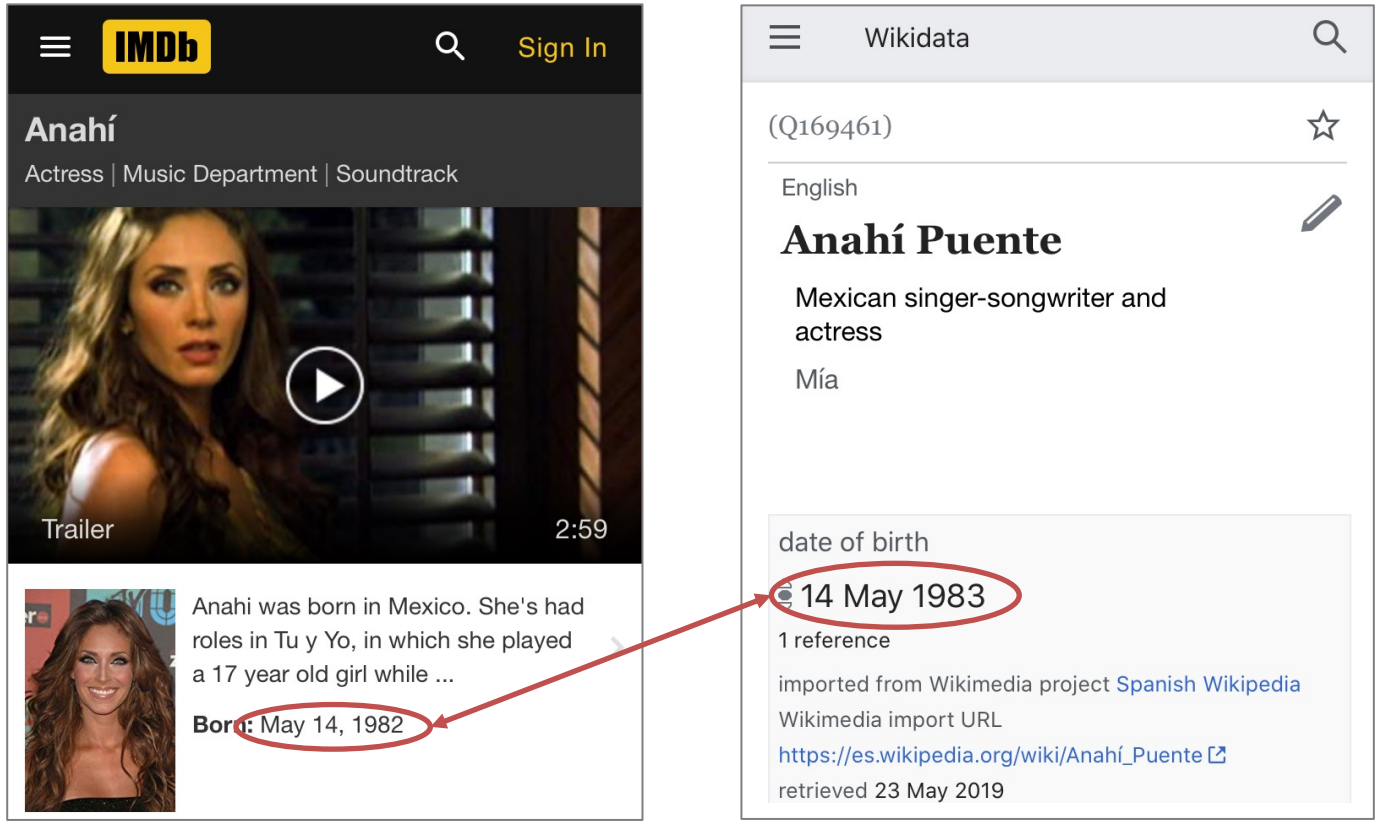
Background – Heterogeneous Data Integration

Schema Alignment

Entity Resolution

Data Fusion

Merge the records that have been regarded as matching in the previous phase



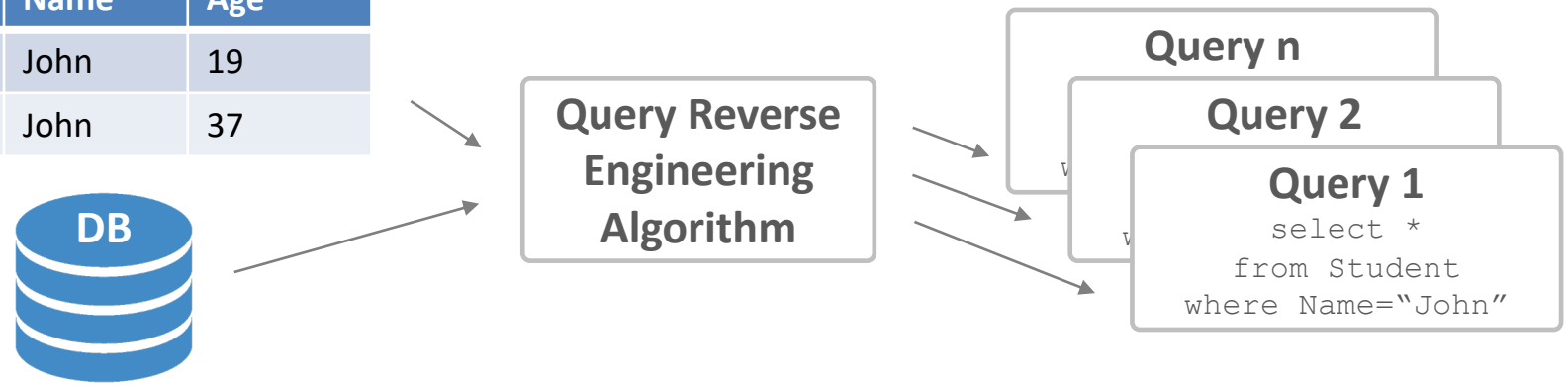
Background – Query Reverse Engineering

Expert user of DB

ID	Name	Age
1	John	19
2	John	37

Non-expert user of DB

ID	Name	Age
1	John	19
2	John	37



Schema Mapping – State of the Art

Source 1

Departure

Client	Town
John Doe	Milan

Source 2

Customer

Name	City of residence
John Doe	Milan

Ticket

Customer	Departure City
John Doe	Milan

Flight

Pilot	Arrival City
John Doe	Milan

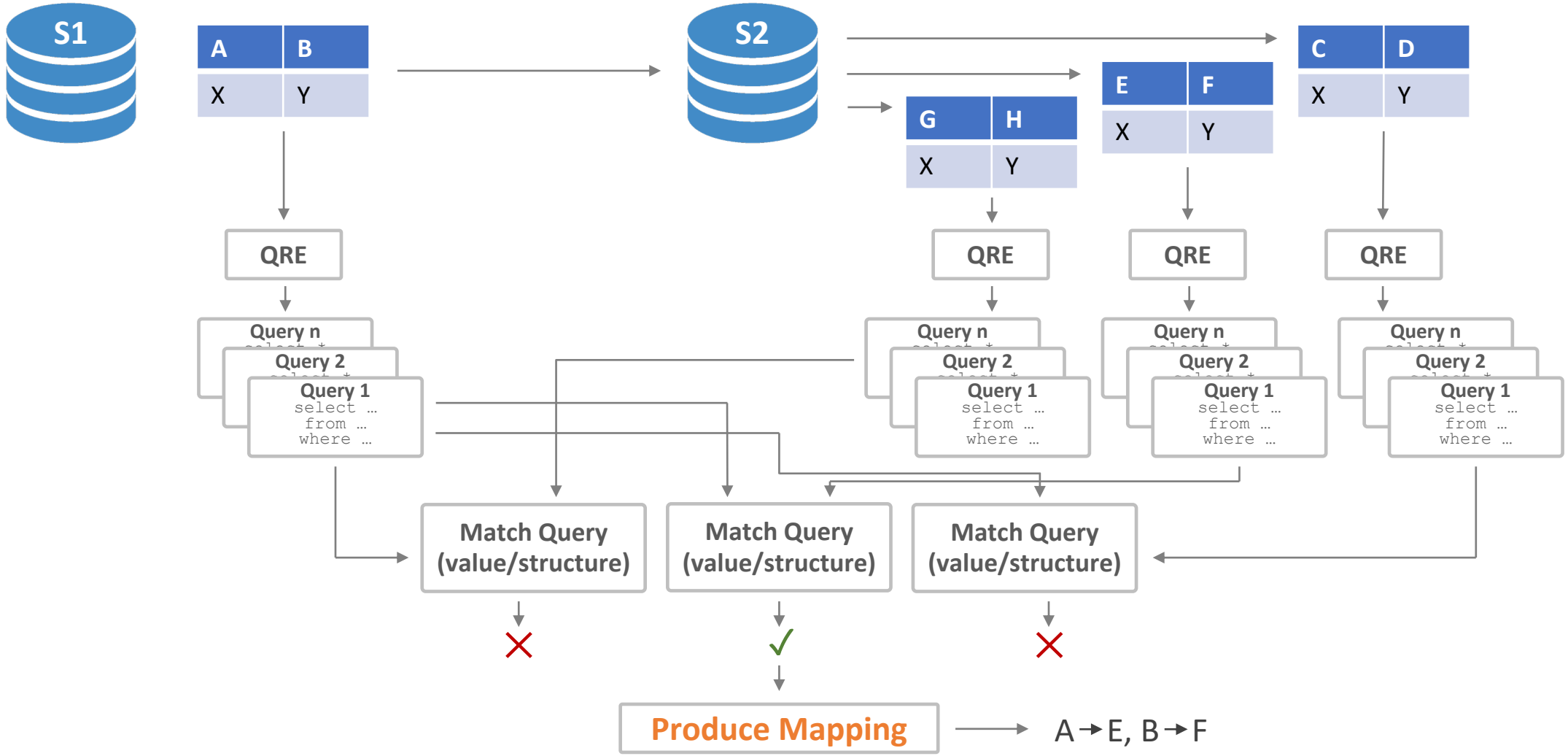
Schema Mapping tools for DB Administrator:

- Require in-dept understanding of the semantics of multiple schemas
- Not suitable when a large number of sources has to be integrated

Interactive Schema Mapping tools for not-technical users:

- Require user interaction
- Prohibitive when the number of sources, or the number of attributes to be integrated is high
- Not suitable for discovering complex mappings

Schema Mapping with QRE



Schema Mapping with QRE

Film

Id	Title	Director	Studios
100	Titanic	James Cameron	London
200	Schindler's List	Steven Spielberg	London

Finance

Film	Revenue
100	2 mld
200	1.5 mld

```
select Director, Studios
from Film join Finance on Id=Film
where Revenue>=1.5mld
```

Crew

Name	Role	Birthplace	Birth_date
James Cameron	Director	London	1954
Steven Spielberg	Director	London	1946

```
select Name, Birthplace
from Crew
where Role='Director' & Birth_date<1955
```

Movie

Title	Movie_Director	Year	Location	Box_Office
Titanic	James Cameron	1997	London	2 mld
Schindler's List	Steven Spielberg	1993	London	1.5 mld

```
select Movie_Director, Location
from Movie
where Box_office>=1.5mld and Year<1998
```

Next steps: test of the methodology on two datasets: movie dataset, museum dataset