Advanced Data Management (CSCI 490/680)

Data Fusion

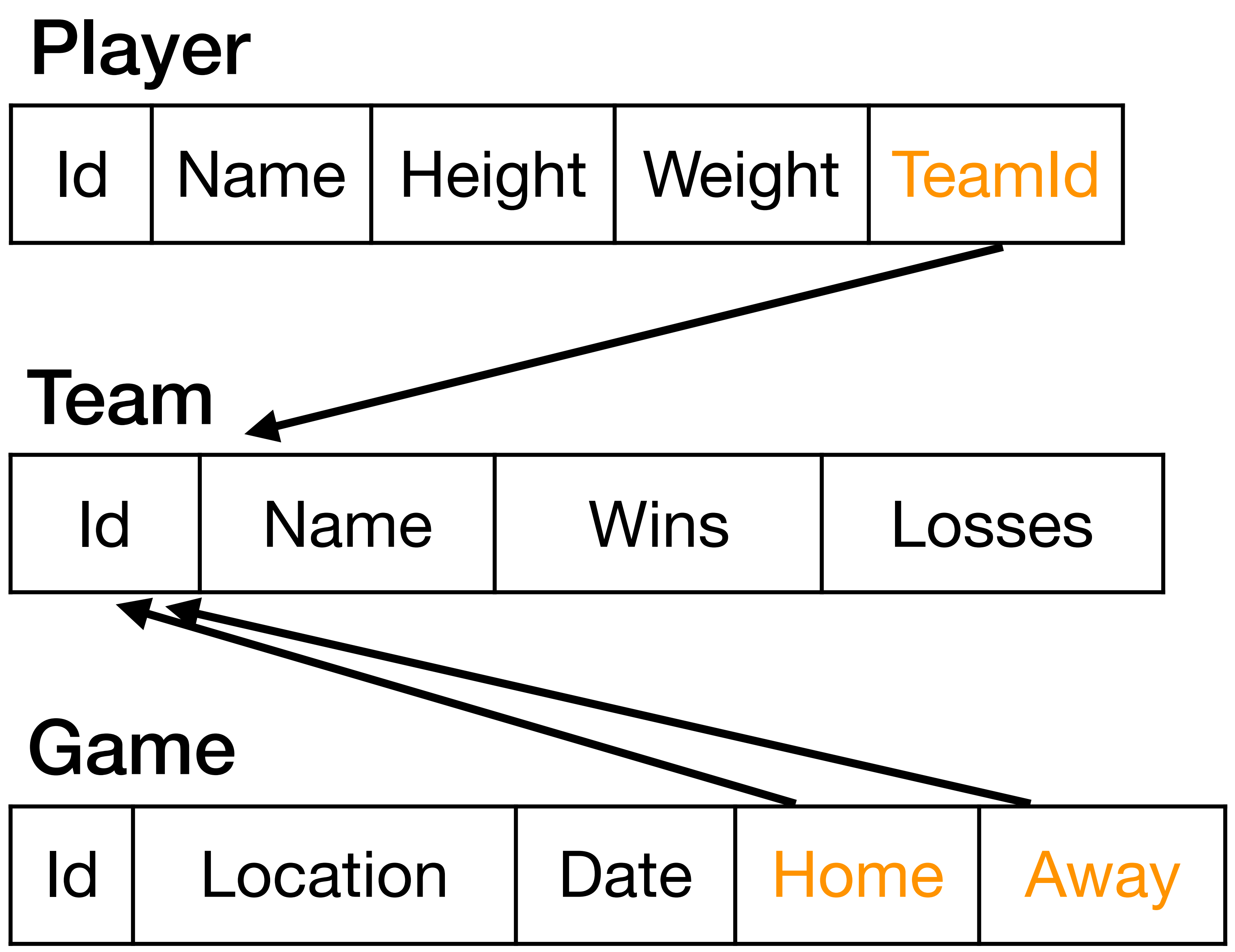
Dr. David Koop

Databases

- Databases:
  - Have been around for years
  - Organize data by tables, allow powerful queries
  - Most support concurrency: allowing multiple users to work with the database at once
  - Provide many features to ensure data integrity, security
- Database Management Systems (DBMS): software that manages databases and facilitates adding, updating, and removing data as well as queries over the data
- Main language used to interact with databases:
  Structured Query Language (SQL)

Football Game Data

- Have each game store the id of the home team and the id of the away team (one-to-one)
- Have each player store the id of the team he plays on (many-to-one)



Player

Id	Name	Height	Weight	TeamId

Team

Id	Name	Wins	Losses

Game

Id	Location	Date	Home	Away

Concatenation

- Take two data frames with the same columns and add more rows
- `pd.concat([data-frame-1, data-frame-2, …])`
- Default is to add rows (`axis=0`), but can also add columns (`axis=1`)
- Can also concatenate Series into a data frame.
- `concat` preserves the index so this can be confusing if you have two default indices (0,1,2,3…)—they will appear twice
    - Use `ignore_index=True` to get a 0,1,2…

Merges (aka Joins)

- Want to join the two tables based on the location and date
- Location and date are the **keys** for the join
- Merges are **ordered**: there is a left and a right side

Game

Id	Location	Date	Home	Away
0	Boston	9/2	1	15
1	Boston	9/9	1	7
2	Cleveland	9/16	12	1
3	San Diego	9/23	21	1

Weather

wId	City	Date	Temp
0	Boston	9/2	72
1	Boston	9/3	68
…	…	…	…
7	Boston	9/9	75
…	…	…	…
21	Boston	9/23	54
…	…	…	…
36	Cleveland	9/16	81

No data for San Diego →

Types of Joins

- Inner: intersection of keys (match on both sides)
- Outer: union of keys (if there is no match on other side, still include with NaN to indicate missing data)
- Left: always have rows from left table (no unmatched right data)
- Right: like left, but with no unmatched left data

Data Merging in Pandas

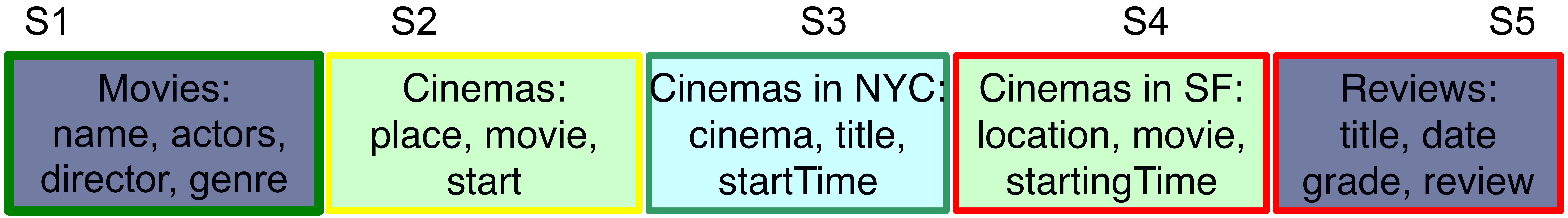
- `pd.merge(left, right, …)`
- Default merge: join on matching column names
- Better: specify the column name(s) to join on via `on` kwarg
  - If column names differ, use `left_on` and `right_on`
  - Multiple keys: use a list
- `how` kwarg specifies type of join `("inner", "outer", "left", "right")`
- Can add suffixes to column names when they appear in both tables, but are not being joined on
- Can also merge using the index by setting `left_index` or `right_index` to `True`

Data Integration

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
          location="New York"  AND
          director="Woody Allen"
```

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.


S1
Movies:
name, actors,
director, genre

S2
Cinemas:
place, movie,
start

S3
Cinemas in NYC:
cinema, title,
startTime

S4
Cinemas in SF:
location, movie,
startingTime

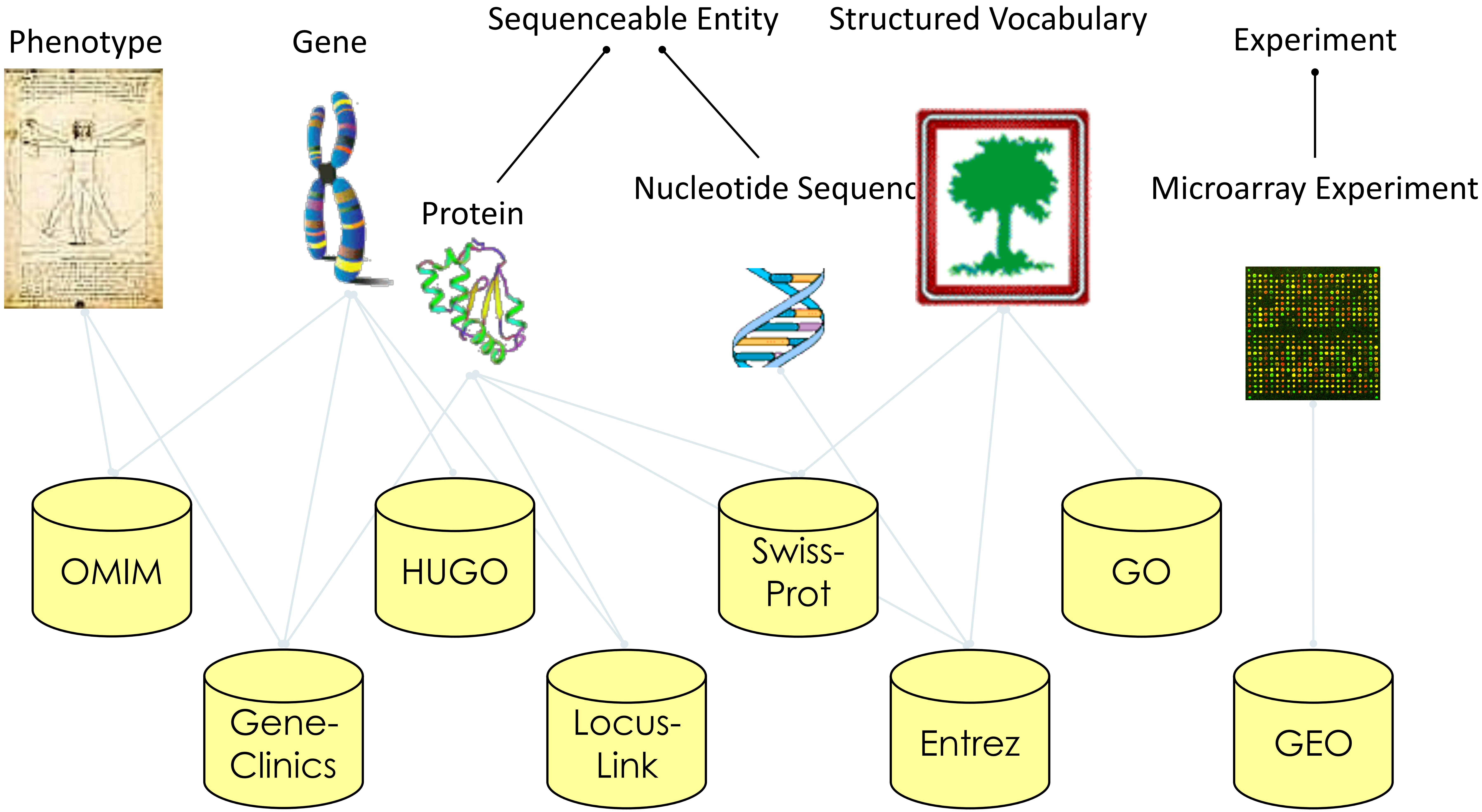
S5
Reviews:
title, date
grade, review

[AH Doan et al., 2012]

Data Integration

- Lots of data sources, how do we answer questions where we need to access data from more than one?
- Schema matching
- Problem of heterogeneity
- AI-Complete problem: difficulty is the same as making computers as intelligent as people
- Two techniques:
  - Mediation
  - Data Warehouses

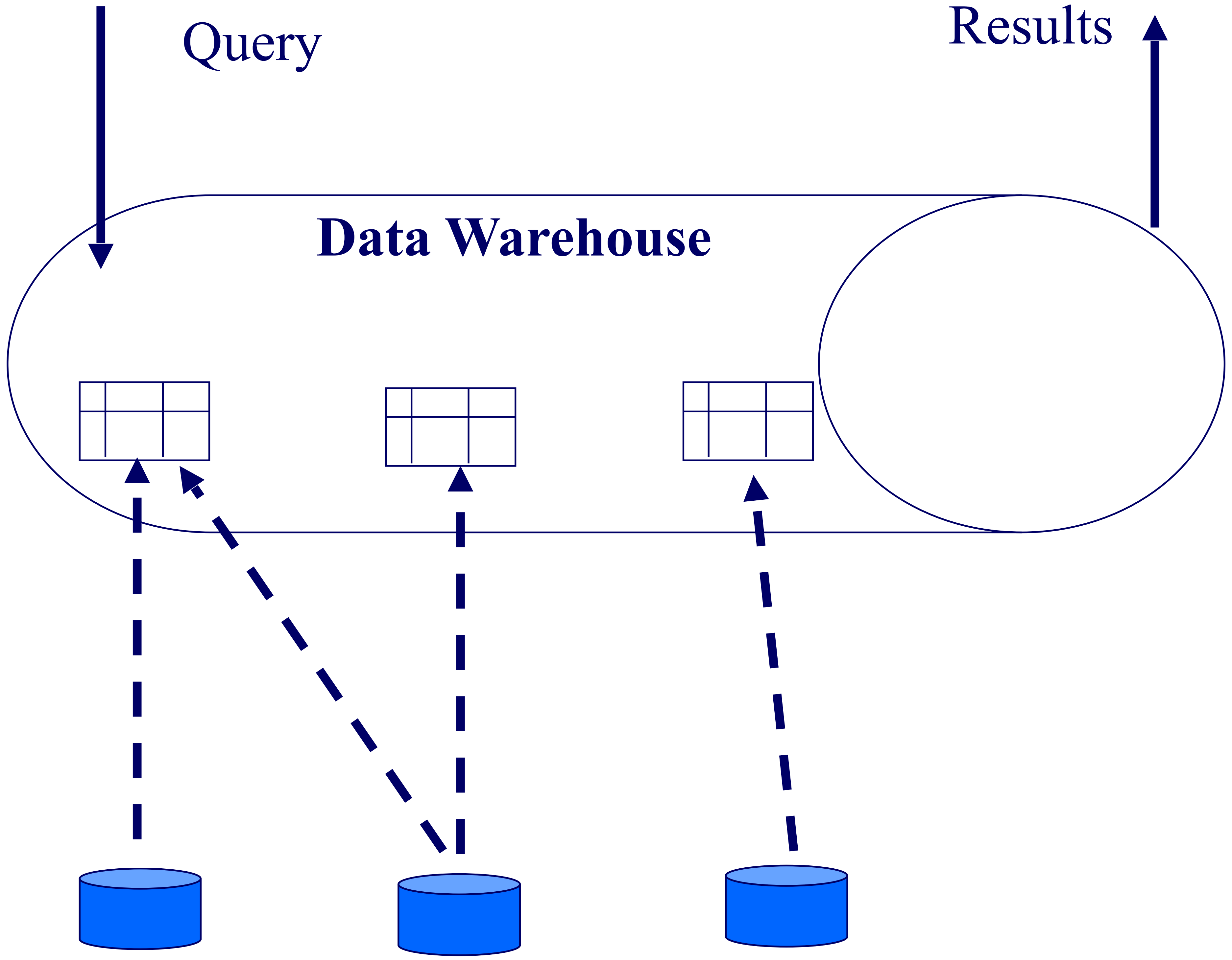
Data Integration Application: Biomedical



[A. Doan et al., 2012]

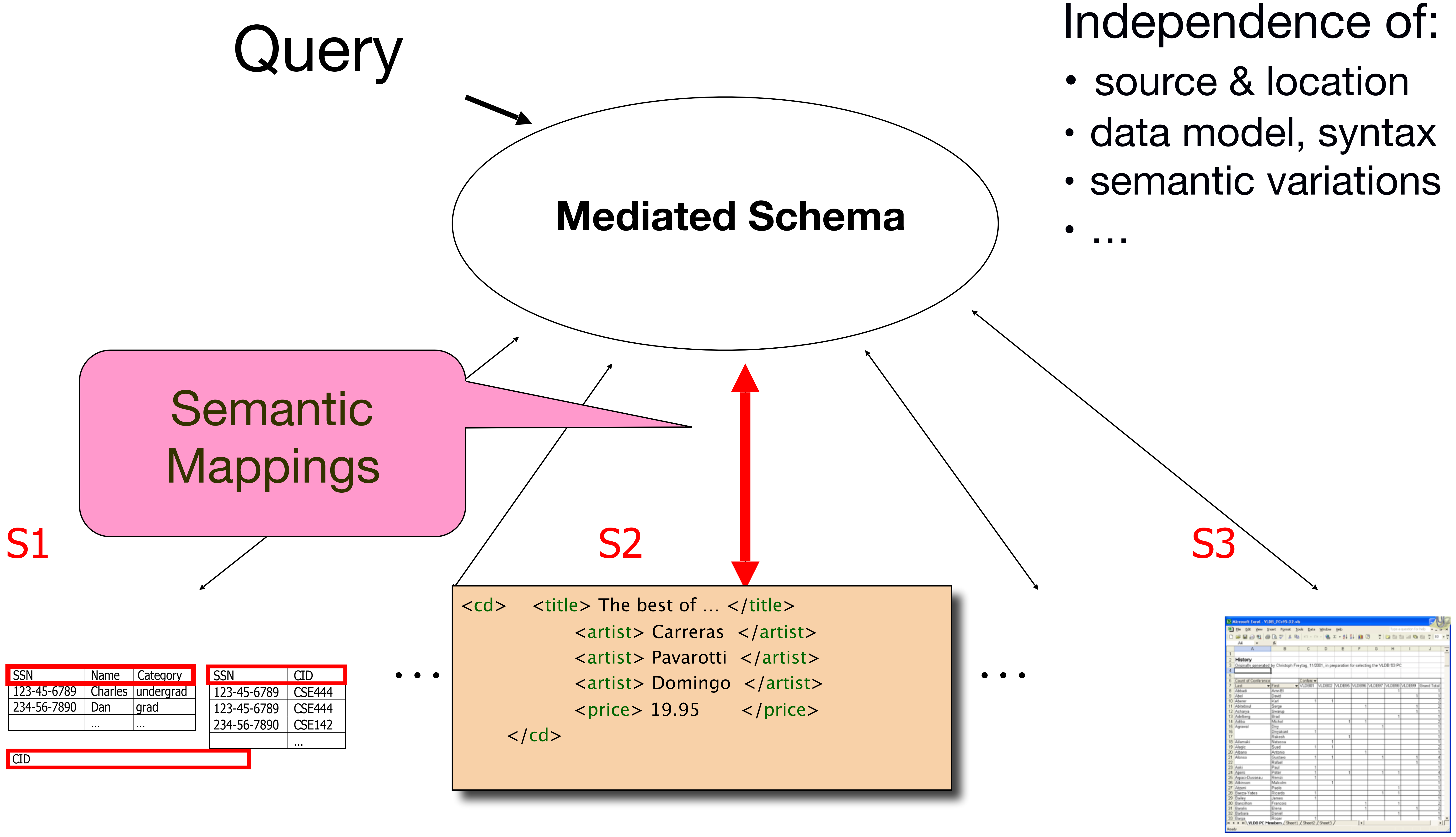
Data Warehouses: Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural mappings in an “ETL tool” to import the data and clean it.
- Periodically copy all of the data from the data sources
    - Note that the sources and the warehouse are basically independent at this point



[A. Doan et al., 2012]

Virtual Data Warehouses



SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	...	...

SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	...

CID

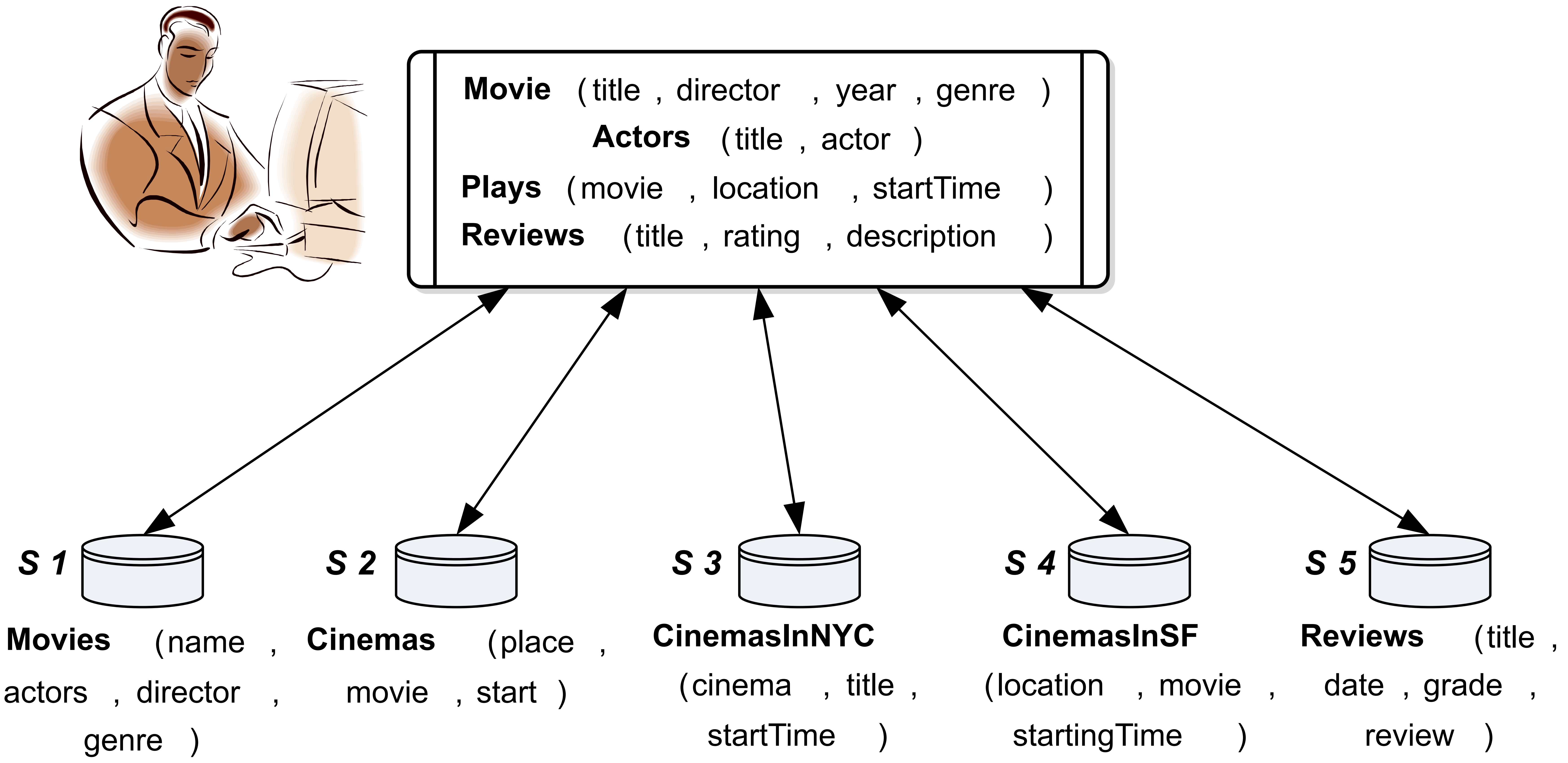
```
<cd>    <title> The best of … </title>
            <artist> Carreras  </artist>
            <artist> Pavarotti  </artist>
            <artist> Domingo  </artist>
            <price> 19.95       </price>
    </cd>
```

Independence of:

- source & location
- data model, syntax
- semantic variations
- …

[A. Doan et al., 2012]

Integrated Schema Example



Movie (title , director , year , genre)
Actors (title , actor)
Plays (movie , location , startTime)
Reviews (title , rating , description)
S 1
S 2
S 3
S 4
S 5
Movies (name , actors , director , genre)
Cinemas (place , movie , start)
CinemasInNYC (cinema , title , startTime)
CinemasInSF (location , movie , startingTime)
Reviews (title , date , grade , review)

[A. Doan et al., 2012]

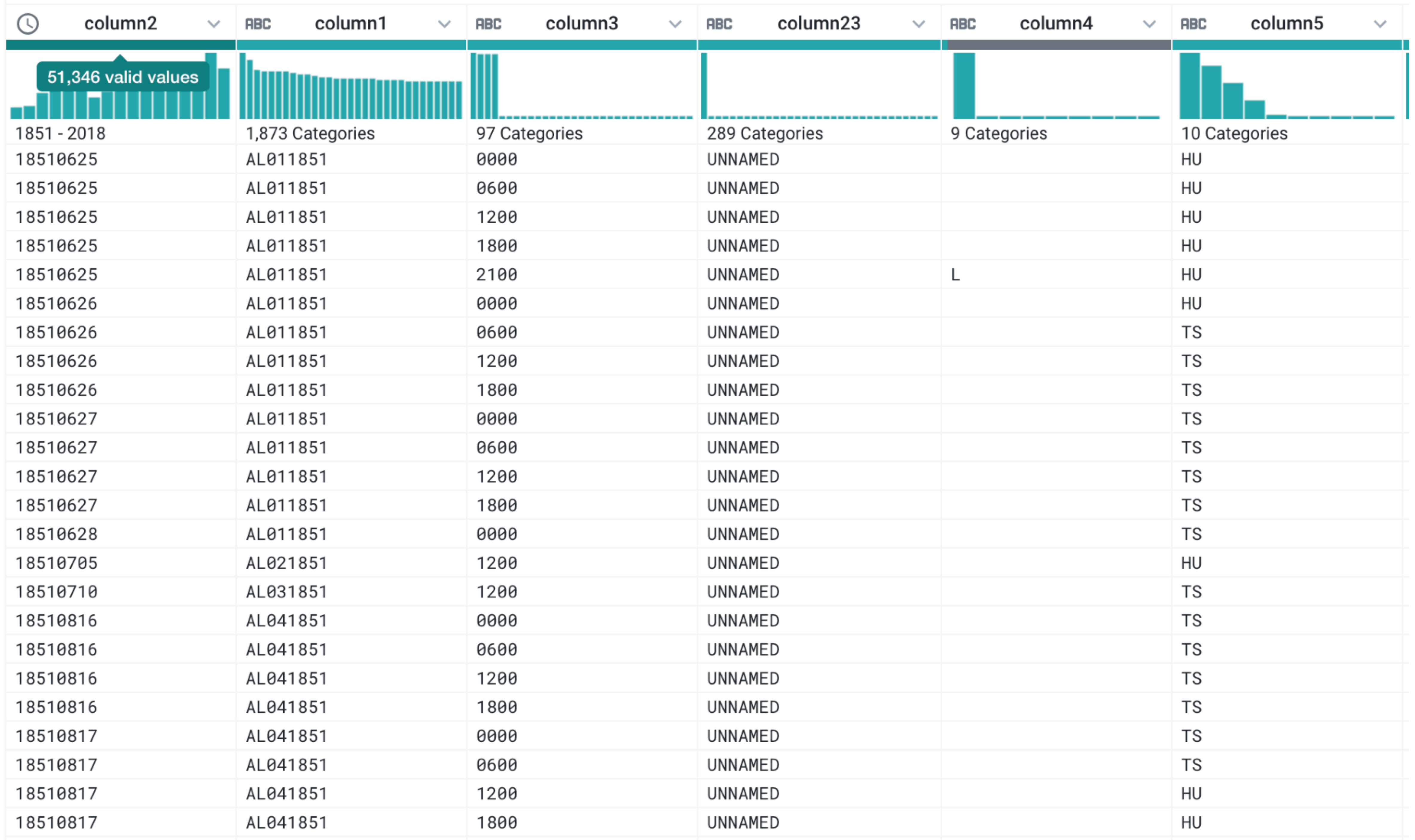
Why is Data Integration Hard?

- Systems-level reasons:
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing
- Logical reasons:
  - Schema (and data) heterogeneity
- ‘Social’ reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (data fiefdoms)
    - Security, privacy and performance implications

[A. Doan et al., 2012]

Assignment 3

- Data wrangling with
  - Trifacta Wrangler
  - pandas
- Same hurdat2 data
- Start now!
- Due Tuesday, March 3



column2	column1	column3	column23	column4	column5
1851 - 2018	1,873 Categories	97 Categories	289 Categories	9 Categories	10 Categories
18510625	AL011851	0000	UNNAMED		HU
18510625	AL011851	0600	UNNAMED		HU
18510625	AL011851	1200	UNNAMED		HU
18510625	AL011851	1800	UNNAMED		HU
18510625	AL011851	2100	UNNAMED	L	HU
18510626	AL011851	0000	UNNAMED		HU
18510626	AL011851	0600	UNNAMED		TS
18510626	AL011851	1200	UNNAMED		TS
18510626	AL011851	1800	UNNAMED		TS
18510627	AL011851	0000	UNNAMED		TS
18510627	AL011851	0600	UNNAMED		TS
18510627	AL011851	1200	UNNAMED		TS
18510627	AL011851	1800	UNNAMED		TS
18510628	AL011851	0000	UNNAMED		TS
18510705	AL021851	1200	UNNAMED		HU
18510710	AL031851	1200	UNNAMED		TS
18510816	AL041851	0000	UNNAMED		TS
18510816	AL041851	0600	UNNAMED		TS
18510816	AL041851	1200	UNNAMED		TS
18510816	AL041851	1800	UNNAMED		TS
18510817	AL041851	0000	UNNAMED		TS
18510817	AL041851	0600	UNNAMED		TS
18510817	AL041851	1200	UNNAMED		HU
18510817	AL041851	1800	UNNAMED		HU

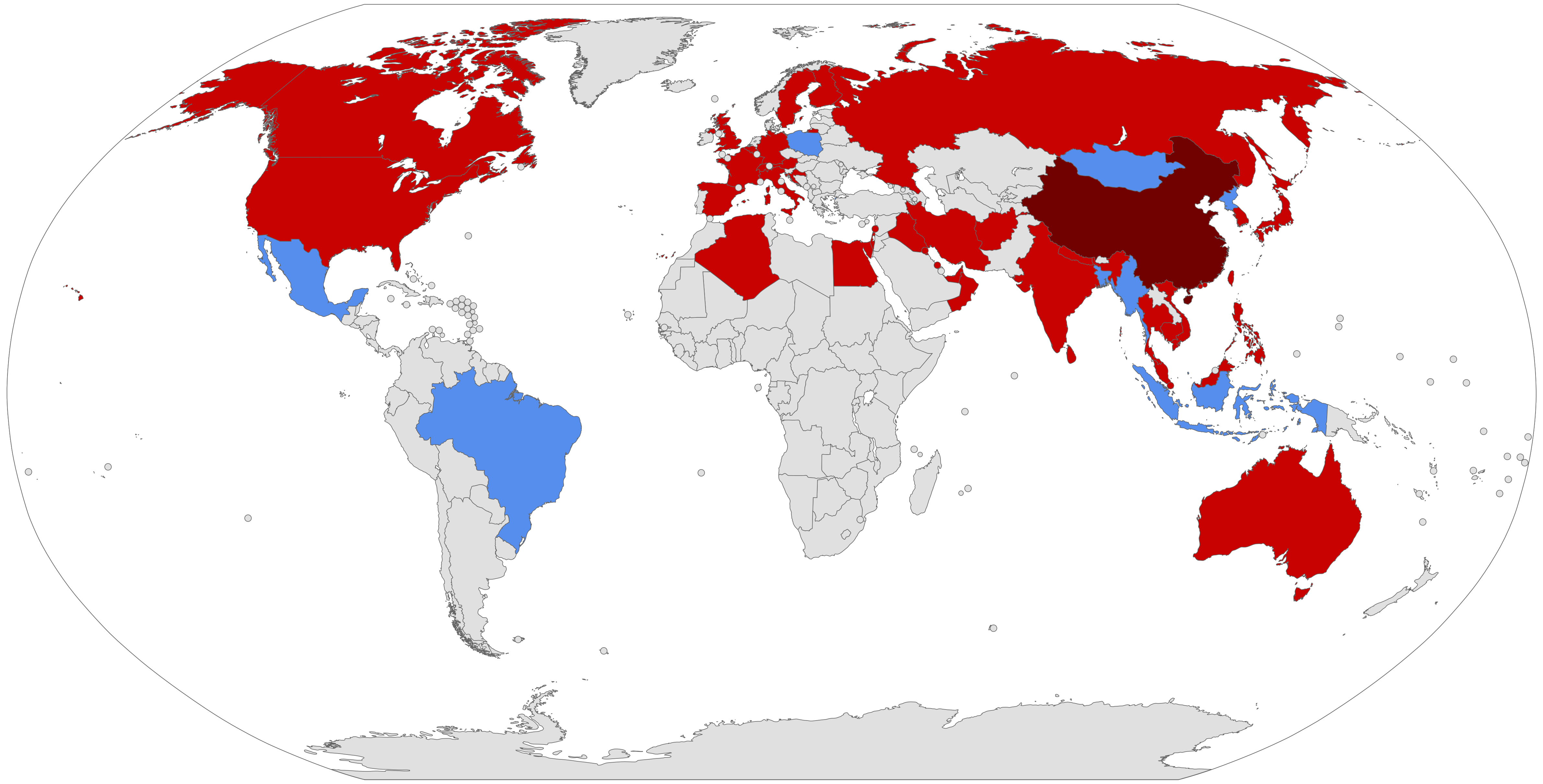
Record Linkage Motivation

- Often data from different sources need to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To improve data quality
  - To enrich data with additional information
- **Lack of unique entity identifiers** means that linking is often based on personal information
- When databases are linked across organisations, maintaining privacy and confidentiality is vital
- The linking of databases is challenged by **data quality**, **database size**, and **privacy concerns**

[P. Christen , 2019]

Motivating Example

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms, ideally in real time
- Data from many different sources will need to be collected (including travel and immigration records; doctors, emergency and hospital admissions; drug purchases; social network and location data; and possibly even animal health data)



Map of the 2019–20 coronavirus outbreak (as of 25 February 2020):

🟫 Region of origin

🟥 Confirmed cases reported

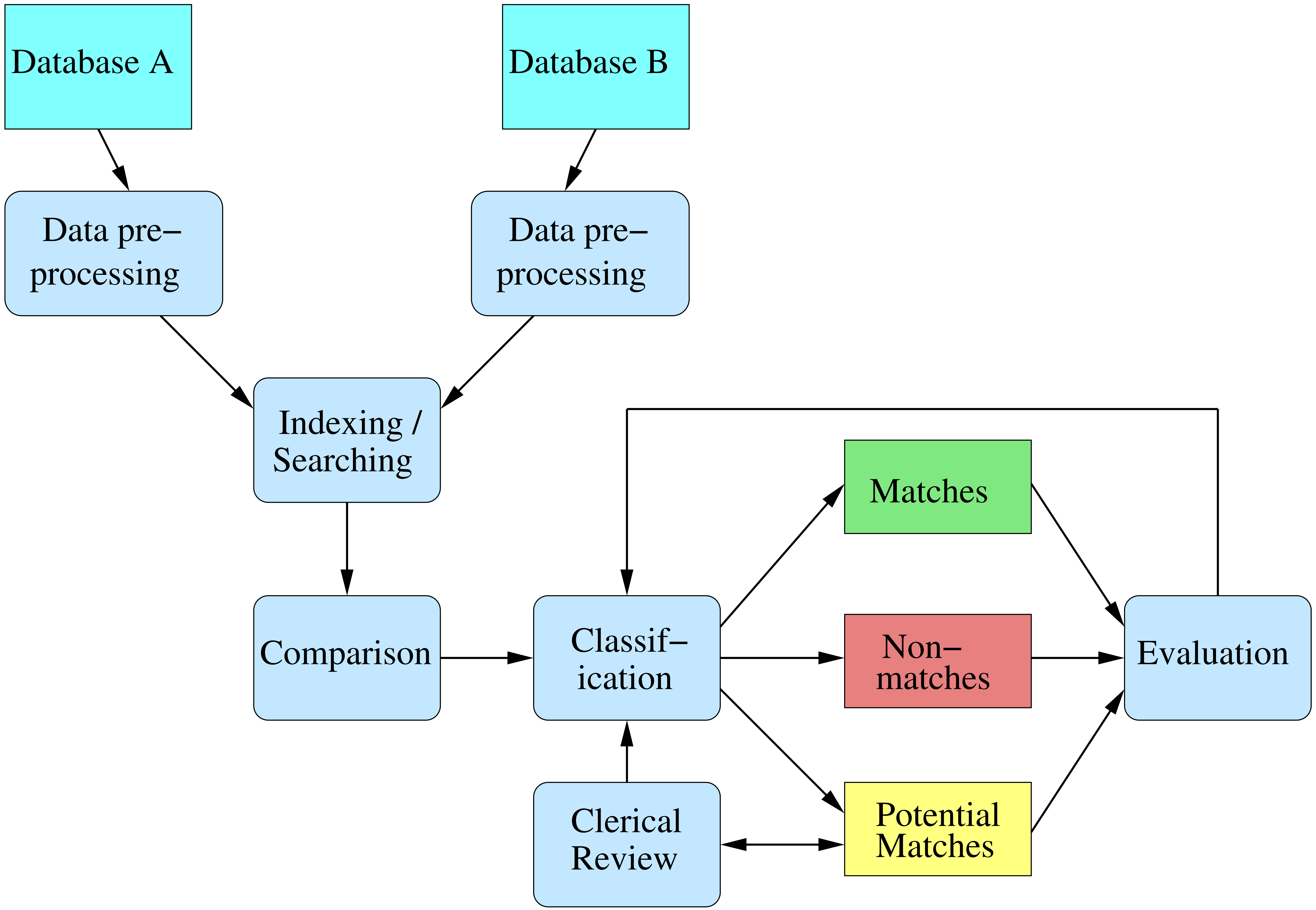
🟦 Suspected cases reported

[P. Christen , 2019], image: [Pharexia, Wikipedia]

Record Linkage

P. Christen

Record Linkage Process



[P. Christen , 2019]

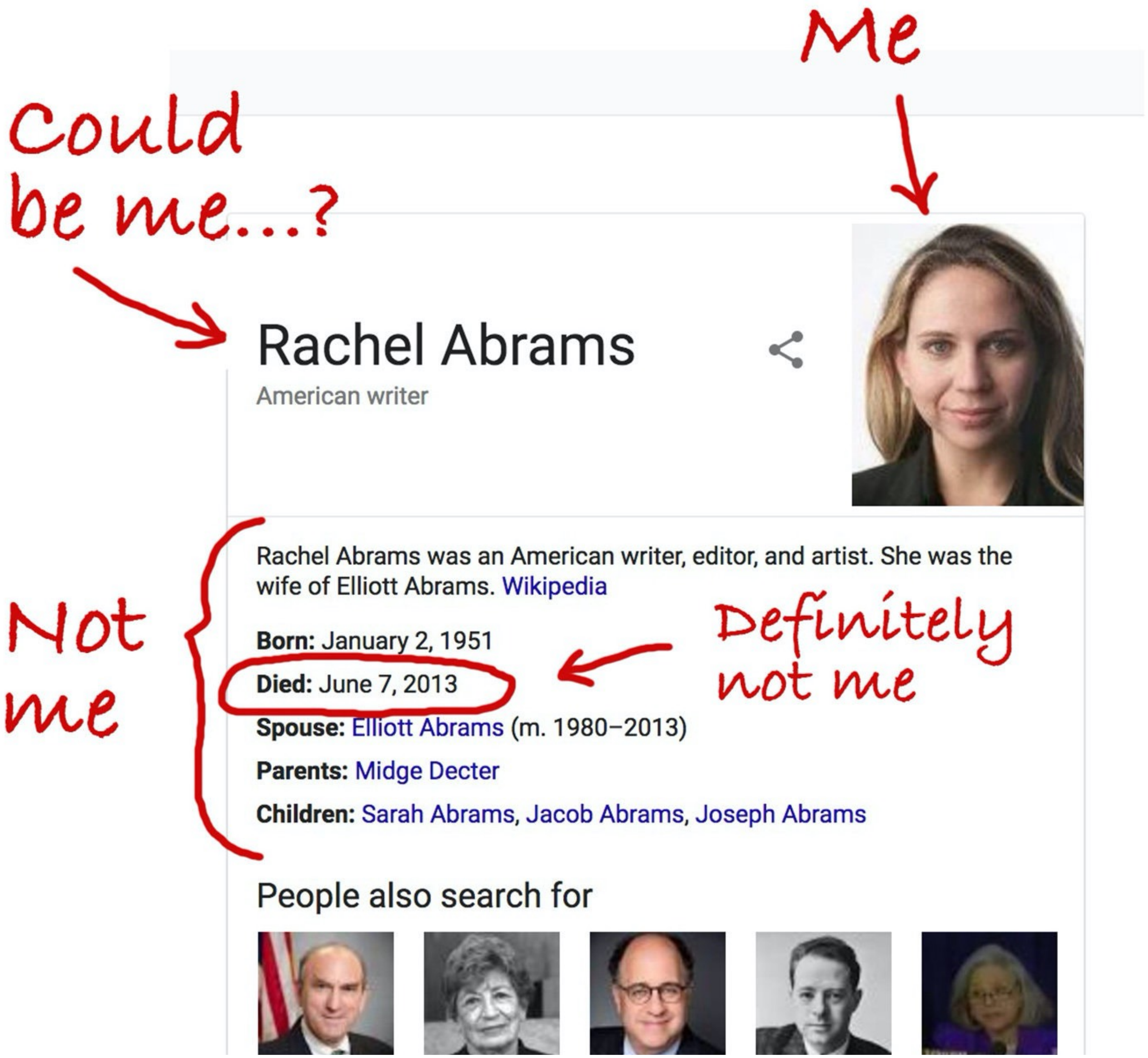
Record Linkage Techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage [Fellegi and Sunter, 1969]
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes
- “Computer science” approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

[P. Christen , 2019]

Data Matching & Data Fusion

- Google Thinks I'm Dead (I know otherwise.) [R. Abrams, NYTimes, 2017]
- Not only Google, but also Alexa:
  - "Alexa replies that Rachel Abrams is a sprinter from the Northern Mariana Islands (which is true of someone else)."
  - "He asks if Rachel Abrams is deceased, and Alexa responds yes, citing information in the Knowledge Graph panel."



Rachel Abrams

American writer

Rachel Abrams was an American writer, editor, and artist. She was the wife of Elliott Abrams. Wikipedia

Born: January 2, 1951

Died: June 7, 2013

Spouse: Elliott Abrams (m. 1980–2013)

Parents: Midge Decter

Children: Sarah Abrams, Jacob Abrams, Joseph Abrams

People also search for

Data Integration and Data Fusion

- Data Integration: focus on integrating data from different sources
- When sources are orthogonal, no problems
- What happens when two sources provide the same type of information and they **conflict**?
- Data Fusion: create a single object while resolving conflicting values

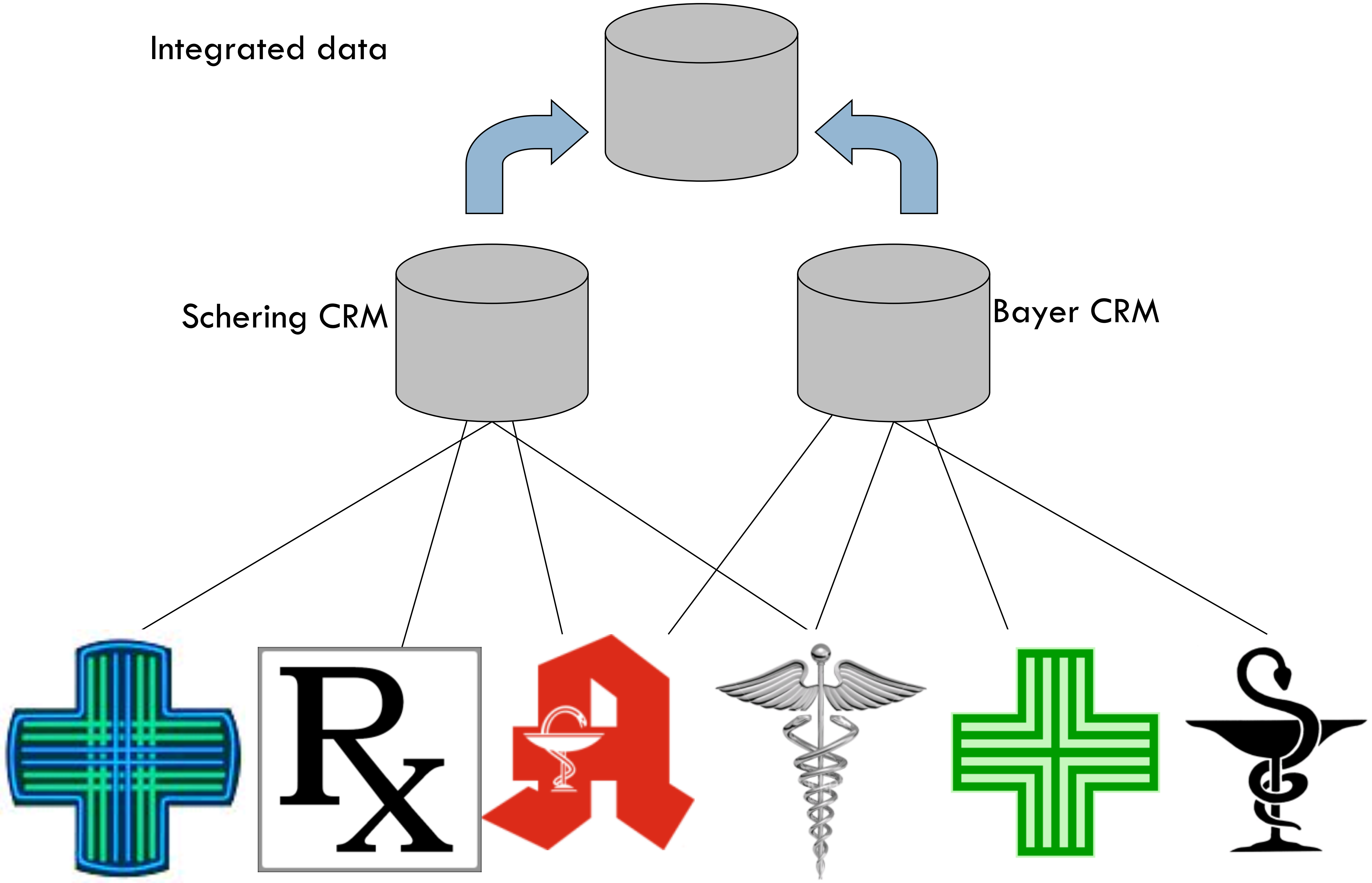
Data Fusion— Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann

Data Fusion Summary

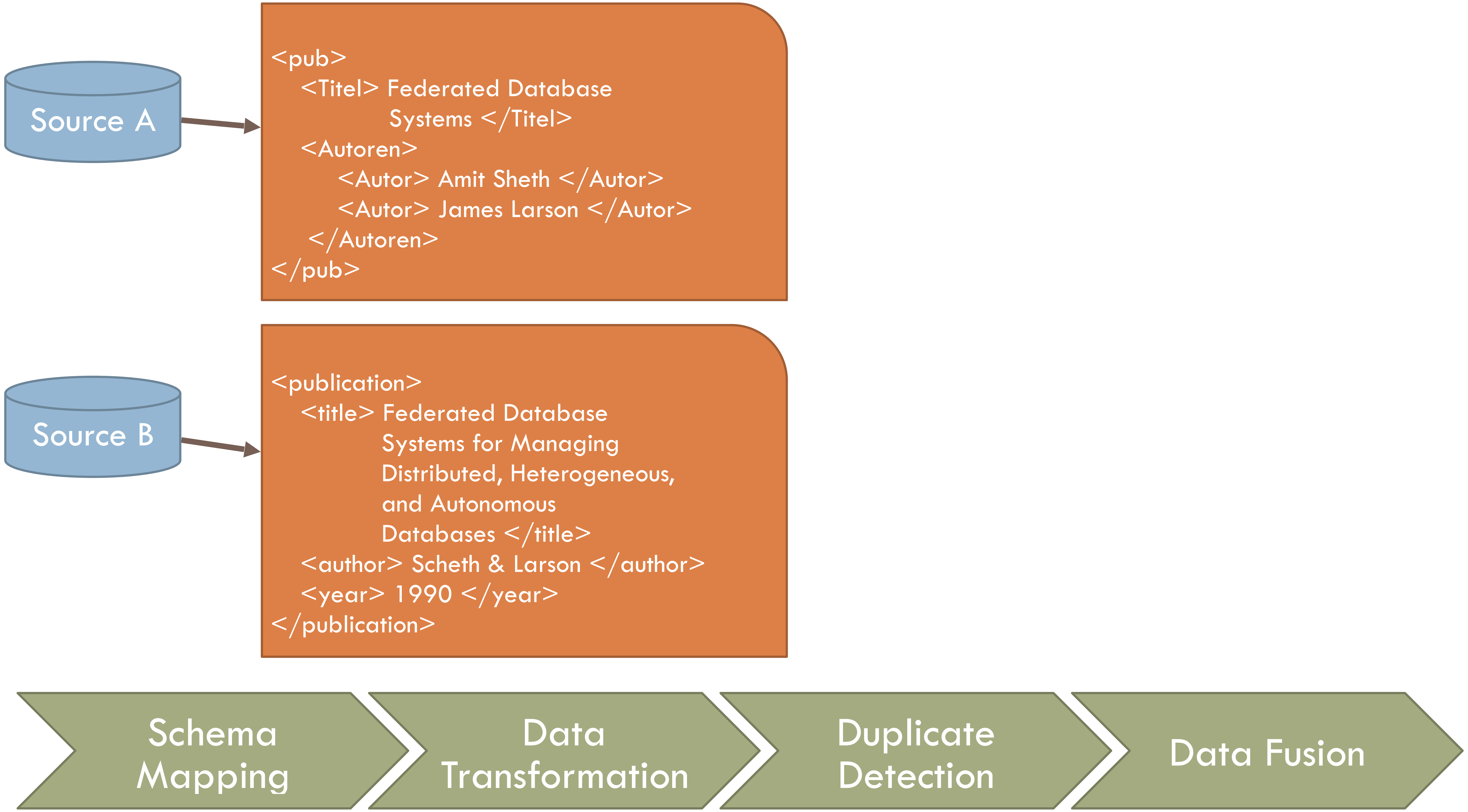
- Conflict resolution strategies
- "Truth-discovery" techniques
  - Accuracy
  - Freshness
  - Dependence
- Fusion Issues
  - Accuracy
  - Efficiency
  - Usability
  - How fusion fits with the rest of data integration?

Data Conflicts



[L. Dong and F. Naumann, 2009]

Information Integration

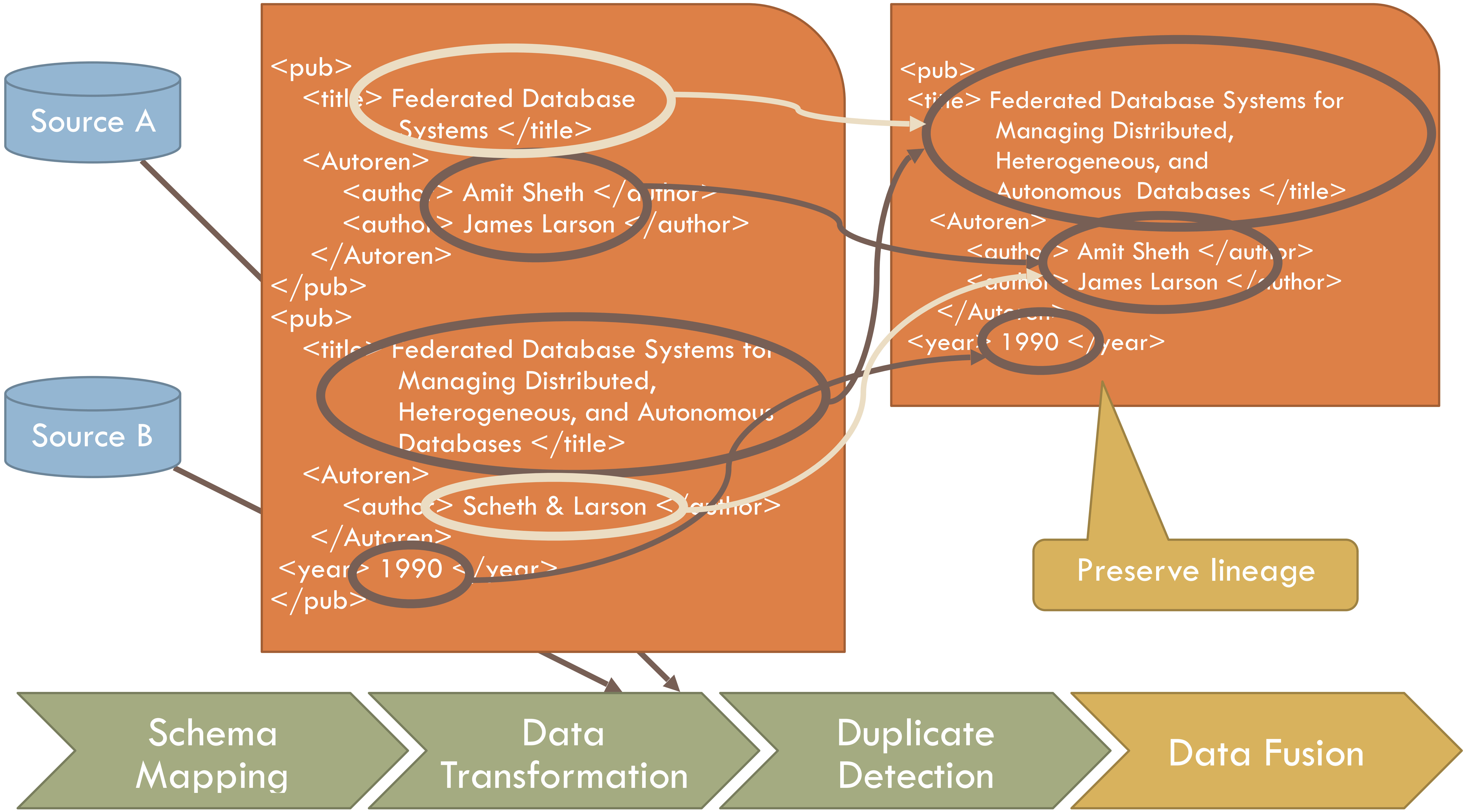


```
<pub>
   <Titel> Federated Database
            Systems </Titel>
   <Autoren>
       <Autor> Amit Sheth </Autor>
       <Autor> James Larson </Autor>
    </Autoren>
</pub>
```

```
<publication>
   <title> Federated Database
            Systems for Managing
            Distributed, Heterogeneous,
            and Autonomous
            Databases </title>
   <author> Scheth & Larson </author>
   <year> 1990 </year>
</publication>
```

[L. Dong and F. Naumann, 2009]

Information Integration



Source A

Source B

```
<pub>
  <title> Federated Database
     Systems </title>
  <Autoren>
     <author> Amit Sheth </author>
     <author> James Larson </author>
  </Autoren>
</pub>
<pub>
  <title> Federated Database Systems for
     Managing Distributed,
     Heterogeneous, and Autonomous
     Databases </title>
  <Autoren>
     <author> Scheth & Larson </author>
  </Autoren>
 <year> 1990 </year>
</pub>
```

```
<pub>
 <title> Federated Database Systems for
        Managing Distributed,
        Heterogeneous, and
        Autonomous  Databases </title>
  <Autoren>
     <author> Amit Sheth </author>
     <author> James Larson </author>
   </Autoren>
 <year> 1990 </year>
```

Preserve lineage

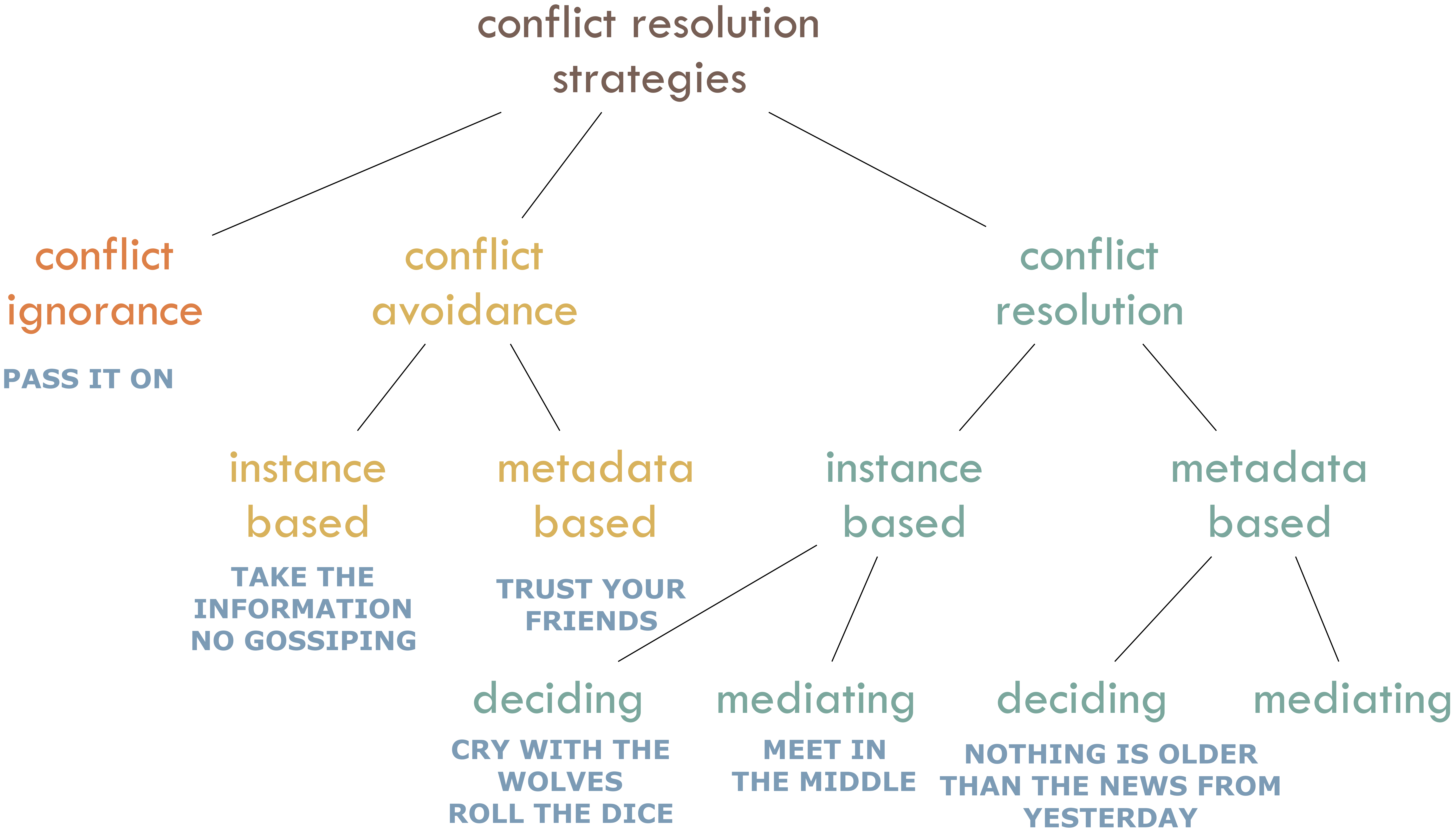
Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

[L. Dong and F. Naumann, 2009]

Data Fusion

- Problem: Given a duplicate, create a single object representation while resolving conflicting data values.
- Difficulties:
  - Null values: Subsumption and complementation
  - Contradictions in data values
  - Uncertainty & truth: Discover the true value and model uncertainty in this process
  - Metadata: Preferences, recency, correctness
  - Lineage: Keep original values and their origin
  - Implementation in DBMS: SQL, extended SQL, UDFs, etc.

Conflict Resolution Strategies



[L. Dong and F. Naumann, 2009]

Integrating Conflicting Data: The Role of Source Dependence

X. L. Dong, L. Berti-Equille, and D. Srivastava

Discussion

- What is the paper's main contribution?
- Do you buy the argument? Any issues with the experiments?
- Can you think of any scenarios where the proposed technique will fail?
- Questions?