Advanced Data Management (CSCI 490/680)

Data Fusion

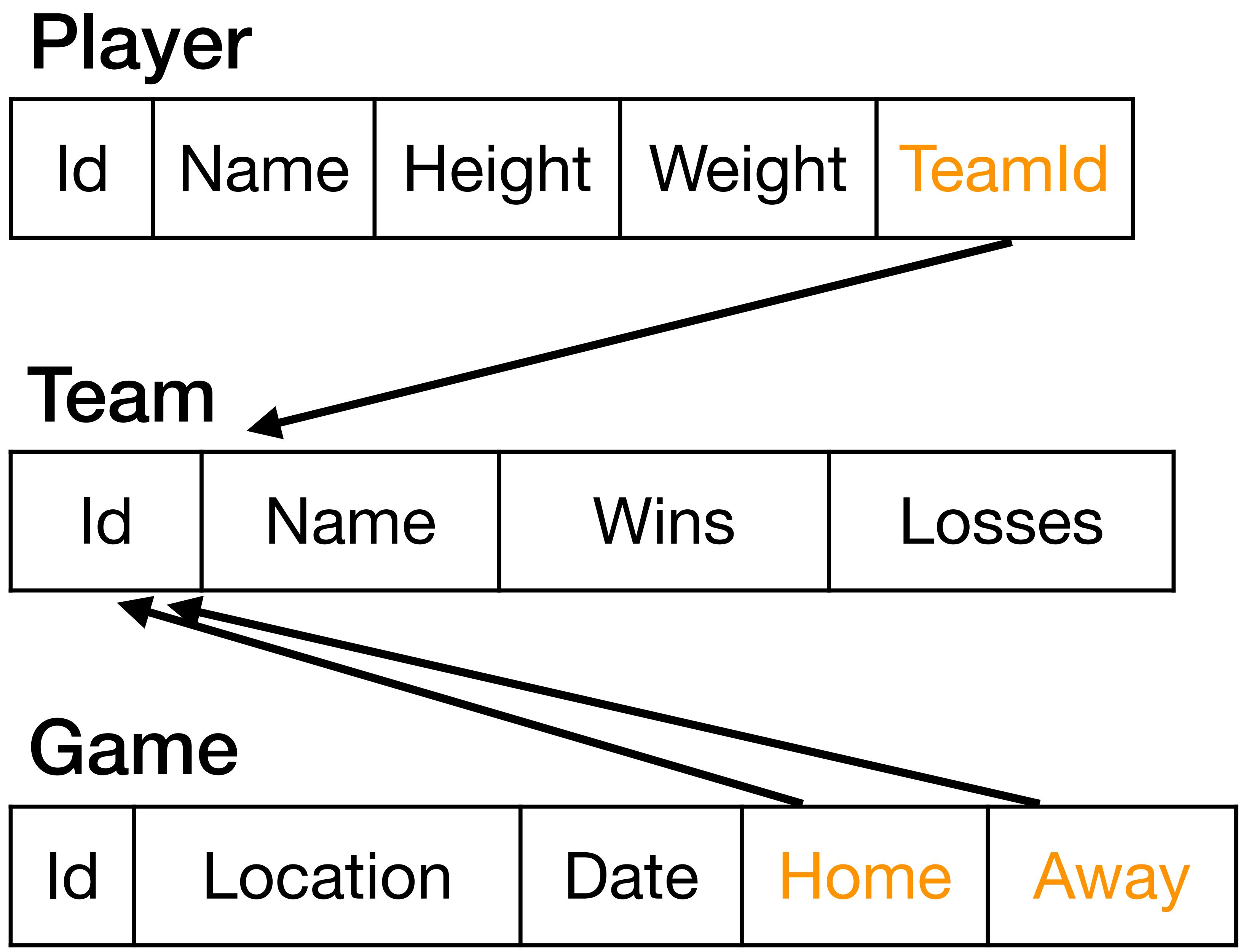
Dr. David Koop

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)



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="Ava DuVernay"
```

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.

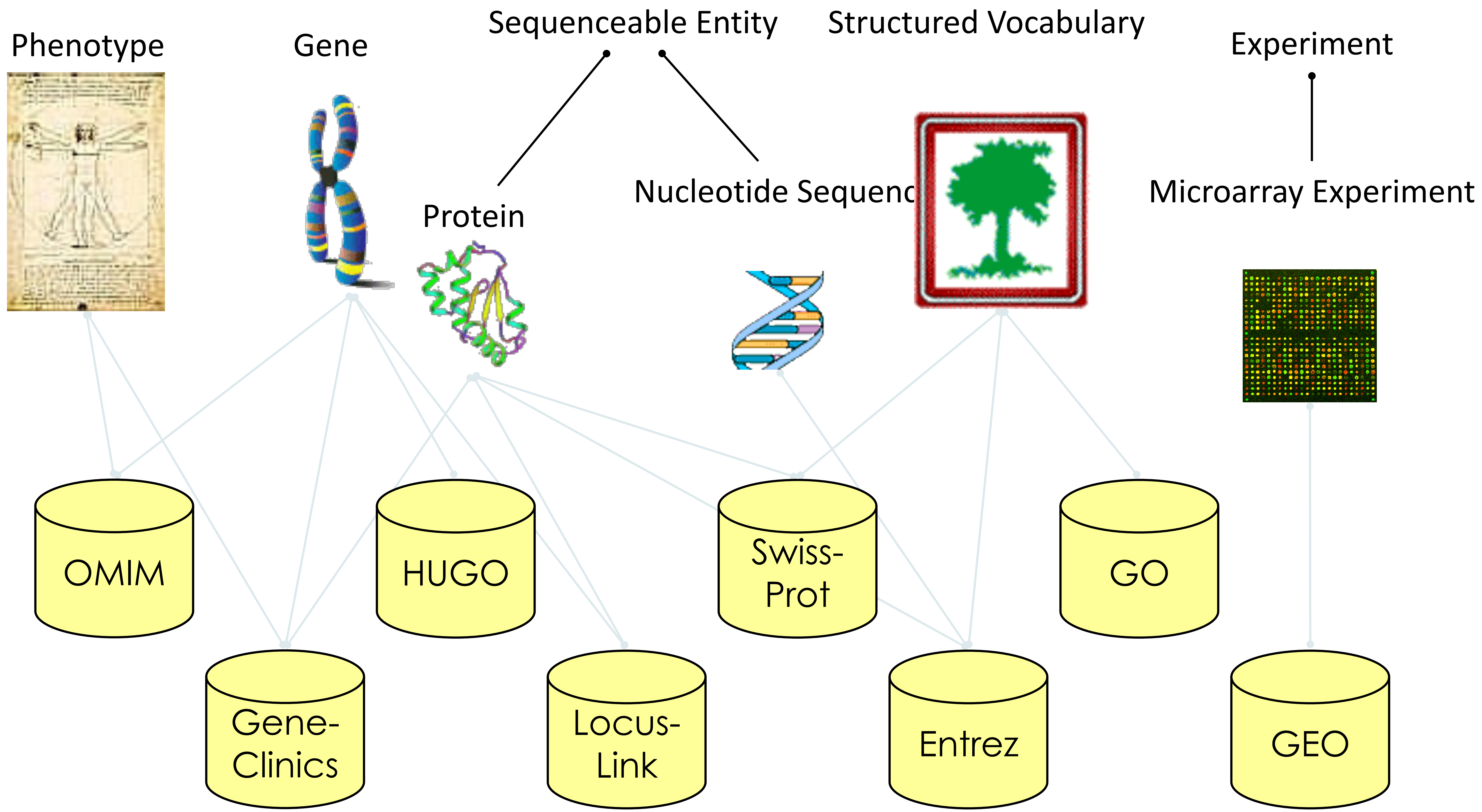
S1	S2	S3	S4	S5
Movies: name, actors, director, genre	Cinemas: place, movie, start	Cinemas in NYC: cinema, title, startTime	Cinemas in SF: location, movie, startingTime	Reviews: title, date grade, review

[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

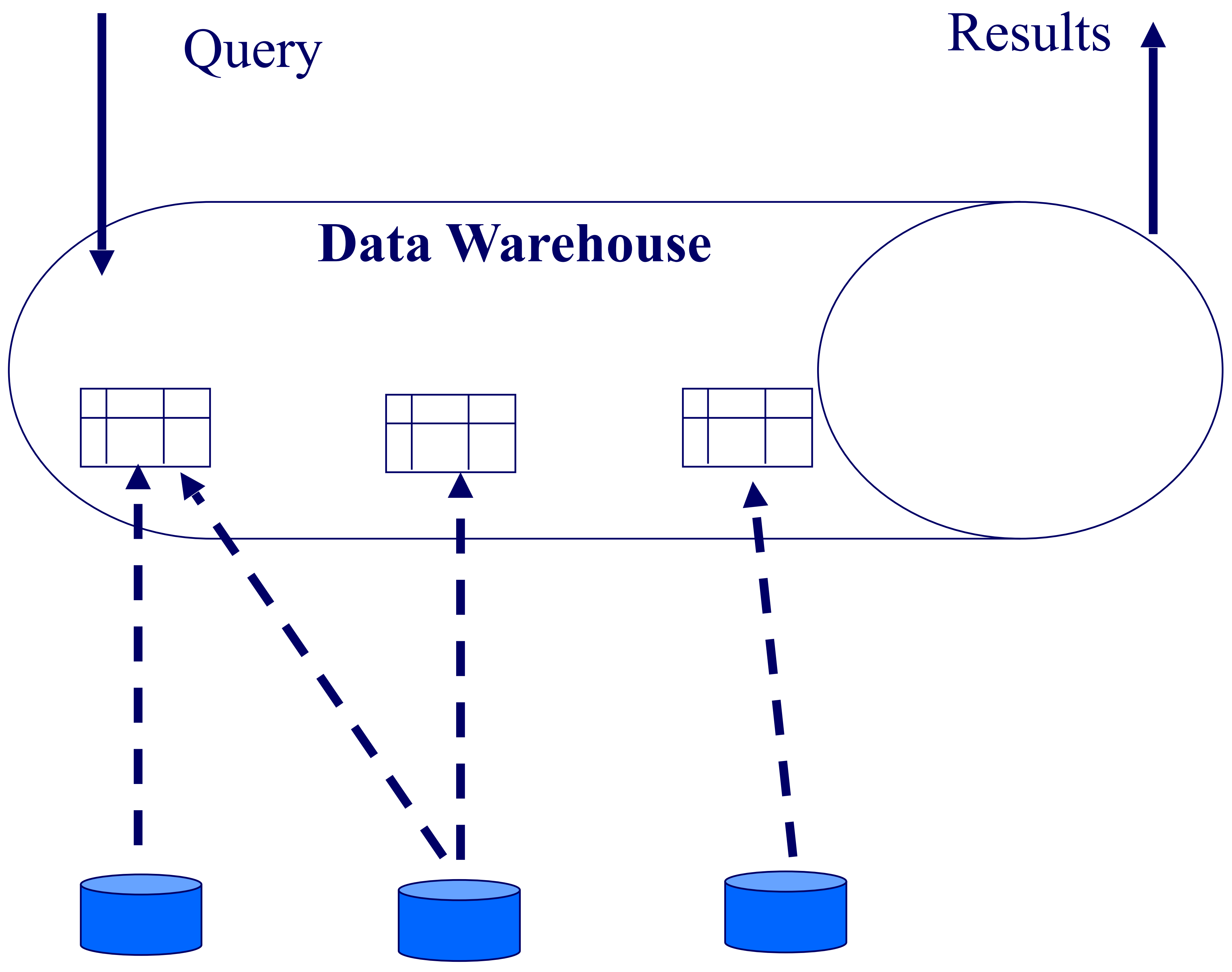
Data Integration Application: Biomedical



[A. Doan et al., 2012]

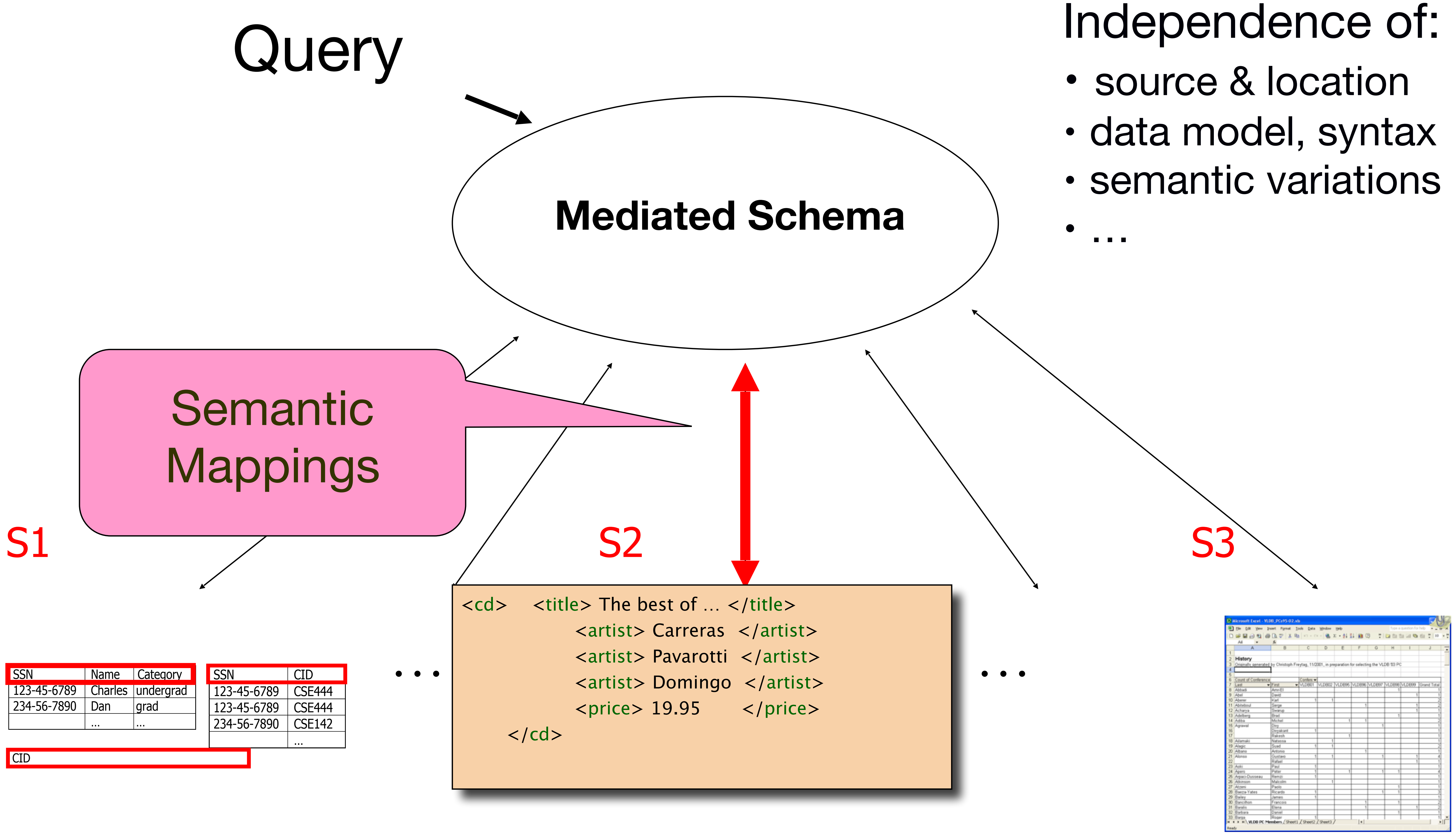
Data Warehouses: Offline Replication

- 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

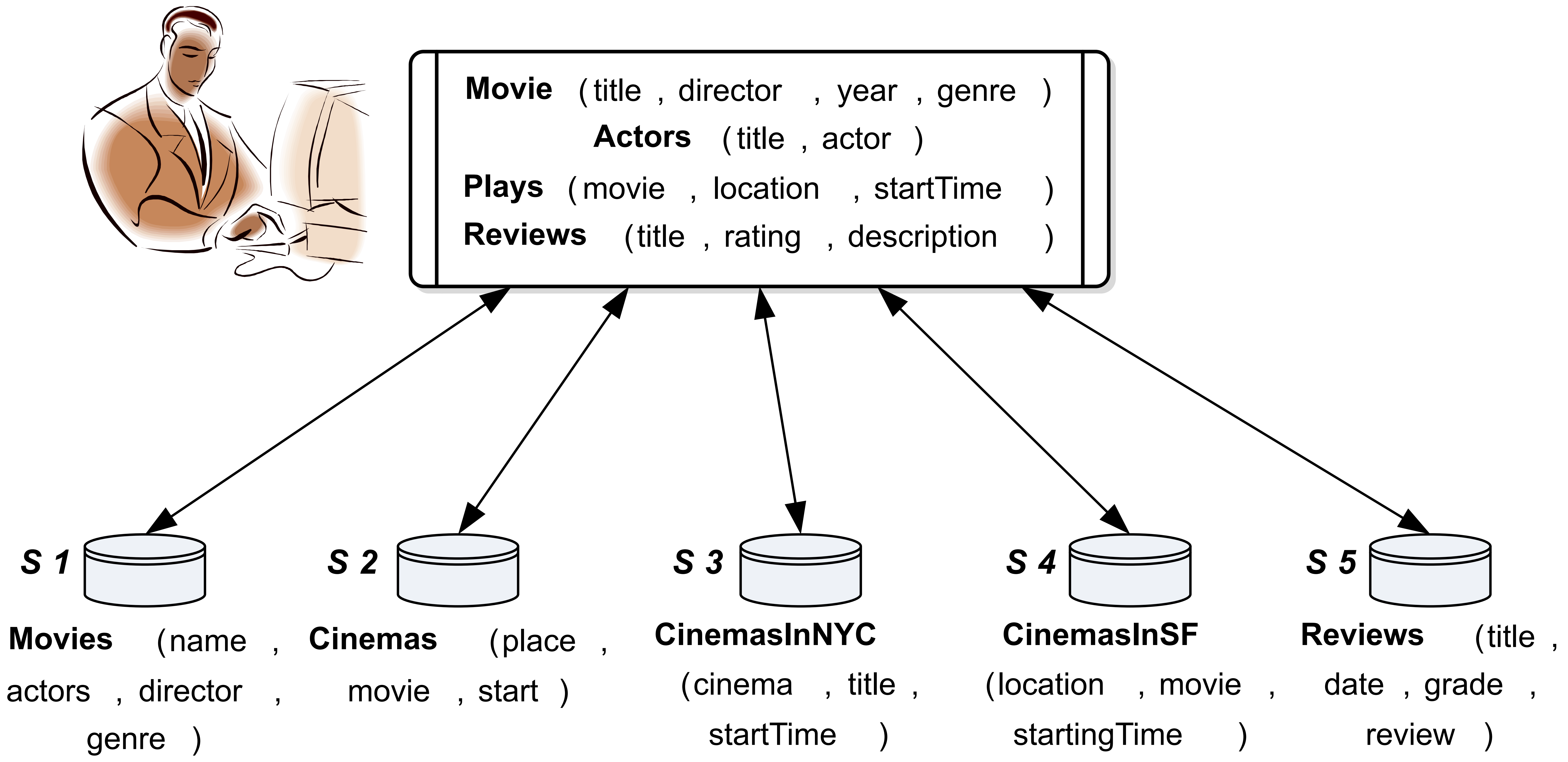


Independence of:

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

[A. Doan et al., 2012]

Integrated Schema Example



[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

- Same Info Wanted data
- Data wrangling with
  - Trifacta Wrangler
  - pandas
- For place, date extraction: 2 regexs, don't try to standardize anything, CS680 need to extract place details, date is EC
- Trifacta # of Rows Issue
- Due Wednesday, March 3

# recid	# order	# date	ABC place	state
1 - 41.23k	1 - 5	1 - 1.87k	5,431 Categories	44 Categories
38575	1	null	MA, BROOKLINE	MA
34452	1	1857	NY, NYC	NY
34453	1	1857	NY, NYC	NY
34454	1	1857	NY, NYC	NY
35259	1	1855	OH, CINCINATTI	OH
37781	1	1864	MA, ABINGTON	MA
37781	2	05/67	MA, BOSTON	MA
37781	3	null	CA	CA
39120	1	null	TX, MILLICAN	TX
34455	1	null	AUSTRALIA	null
34776	1	null	IL, CHICAGO	IL
34881	1	64	NY, BINGHAMPTON, BROOME CO.	NY
35309	1	1860	IL	IL
35537	1	1861	MA, BOSTON	MA
34757	1	null	TN, NASHVILLE	TN
38439	1	null	MA, BOSTON	MA
38439	2	null	CA, SAN FRANCISCO	CA
41070	2	null	CINCINNATI	null
33438	1	1862	MA, BOSTON	MA
33478	1	10/64	AL, MOBILE	AL
33478	2	null	IL, ST. TRELIA	IL
33940	1	1857	NC	NC
34331	1	02/65	MA, BOSTON	MA
33693	1	null	NY	NY
33693	2	null	CANADAS	null
34306	1	02/65	MA, BOSTON	MA
36900	1	null	PA, PHILADELPHIA	PA
37541	1	null	AUSTRALIA, SIDNEY	null
33485	1	1858	MA, NEW BEDFORD	MA

Quiz

- Login to Blackboard (webcourses.niu.edu)
- Quiz is under Tests & Quizzes
- Reading Quiz - 2021-02-24
- You have **five (5)** minutes to answer the **five (5)** multiple choice questions

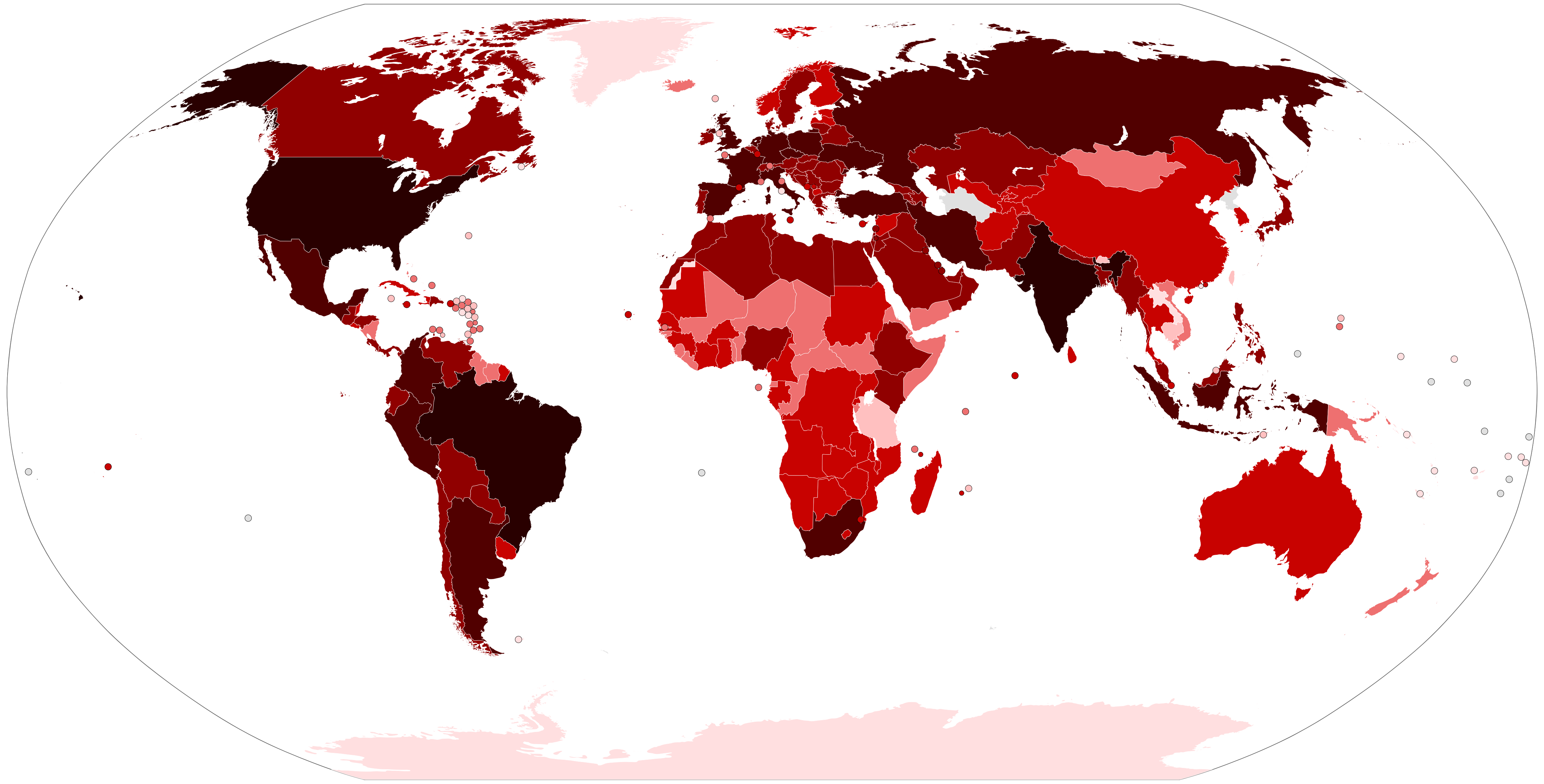
Record Linkage Motivation

- 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)

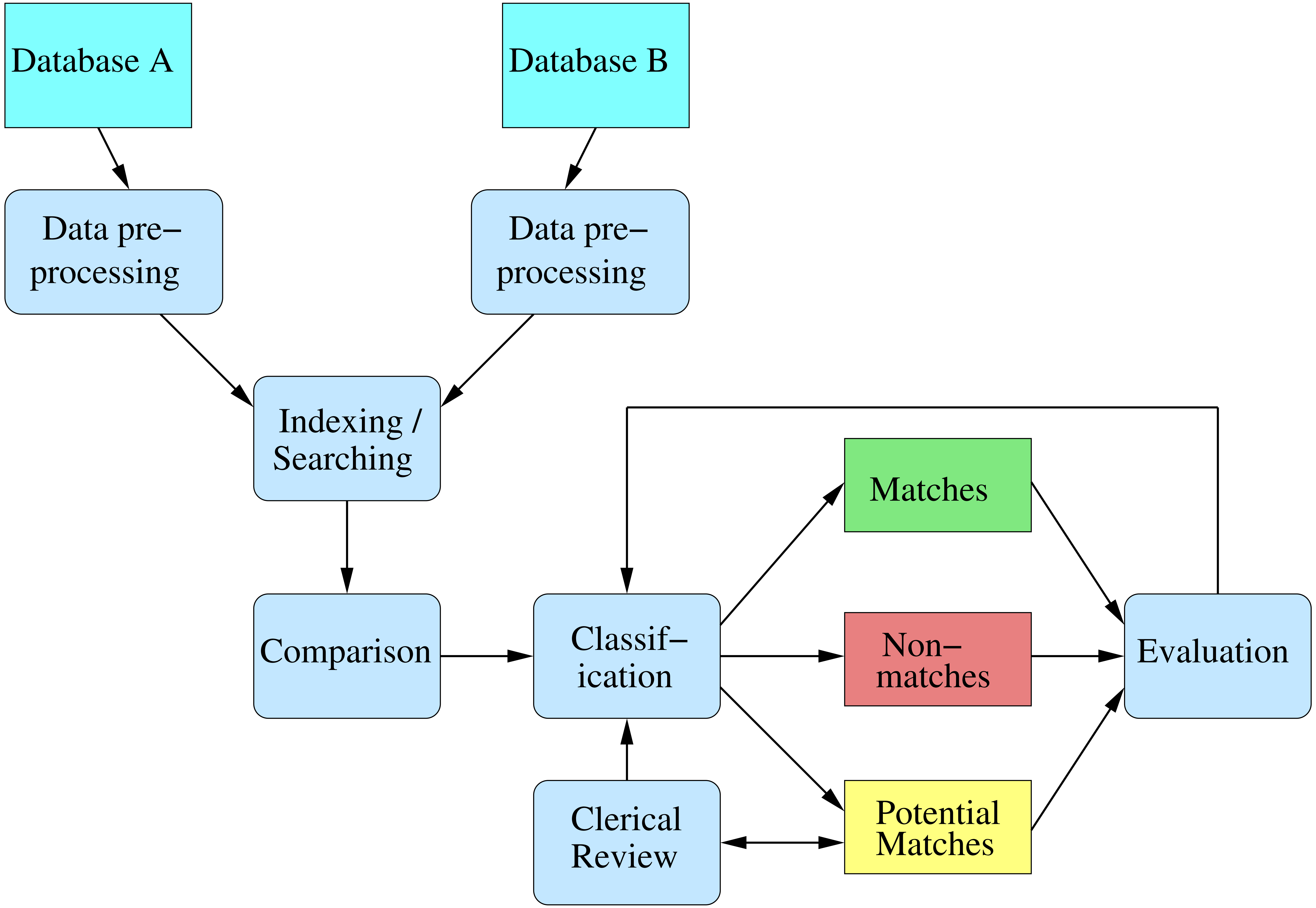


[P. Christen , 2019], image: [Pharexia, Wikipedia]

Record Linkage

P. Christen

Record Linkage Process



[P. Christen , 2019]

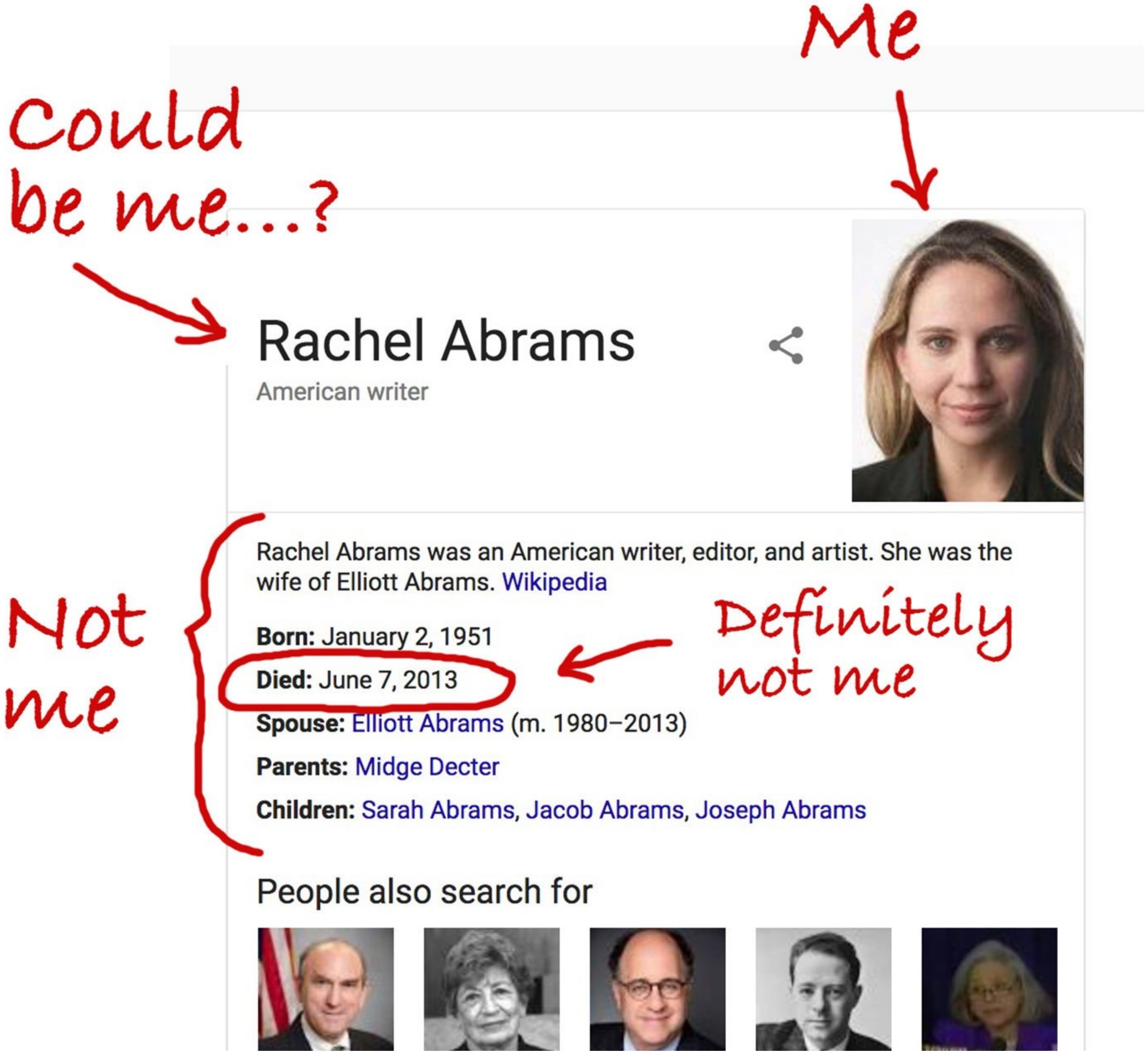
Record Linkage Techniques

- 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."



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

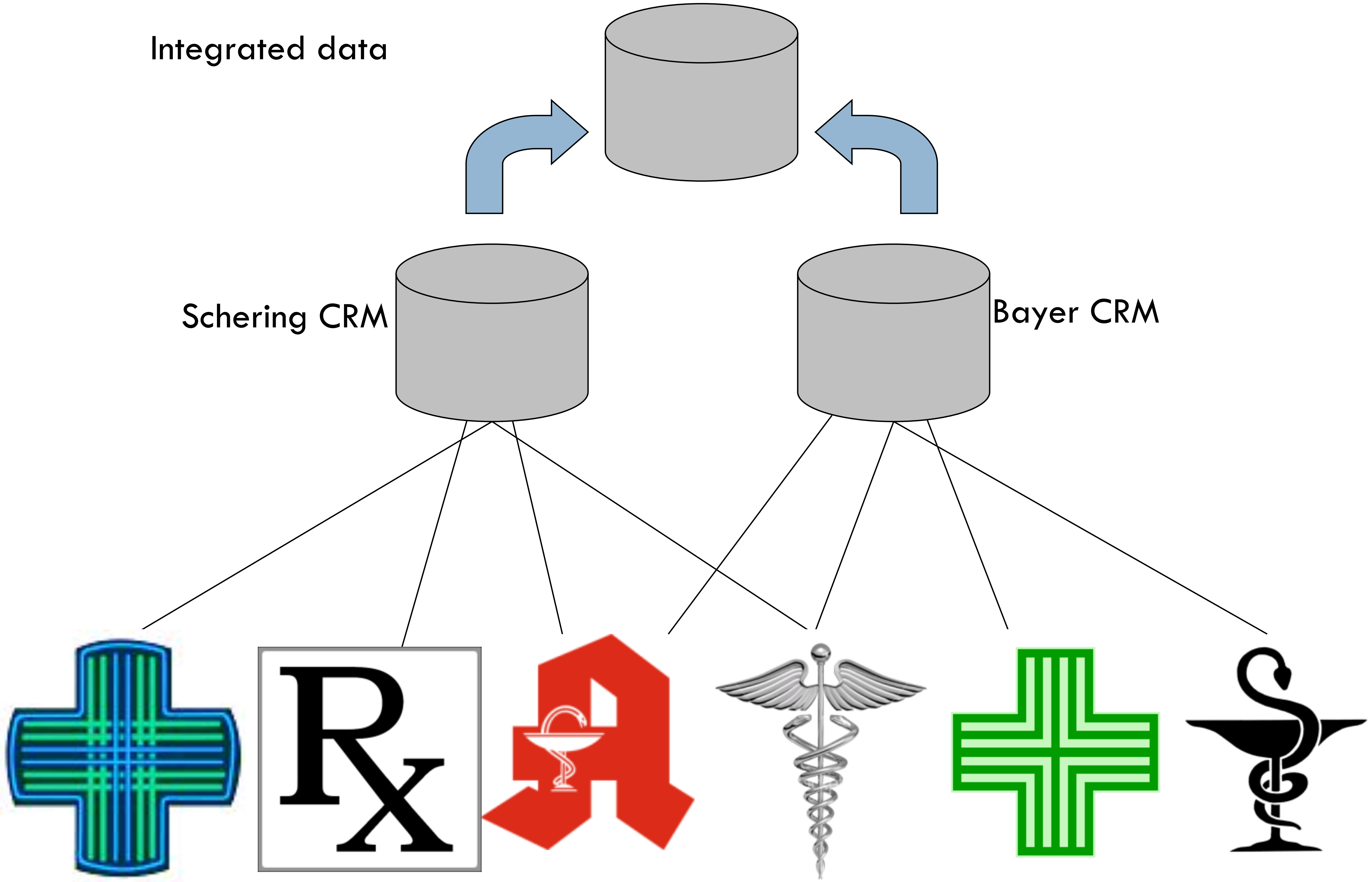
Data Fusion— Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann

Data Fusion Summary

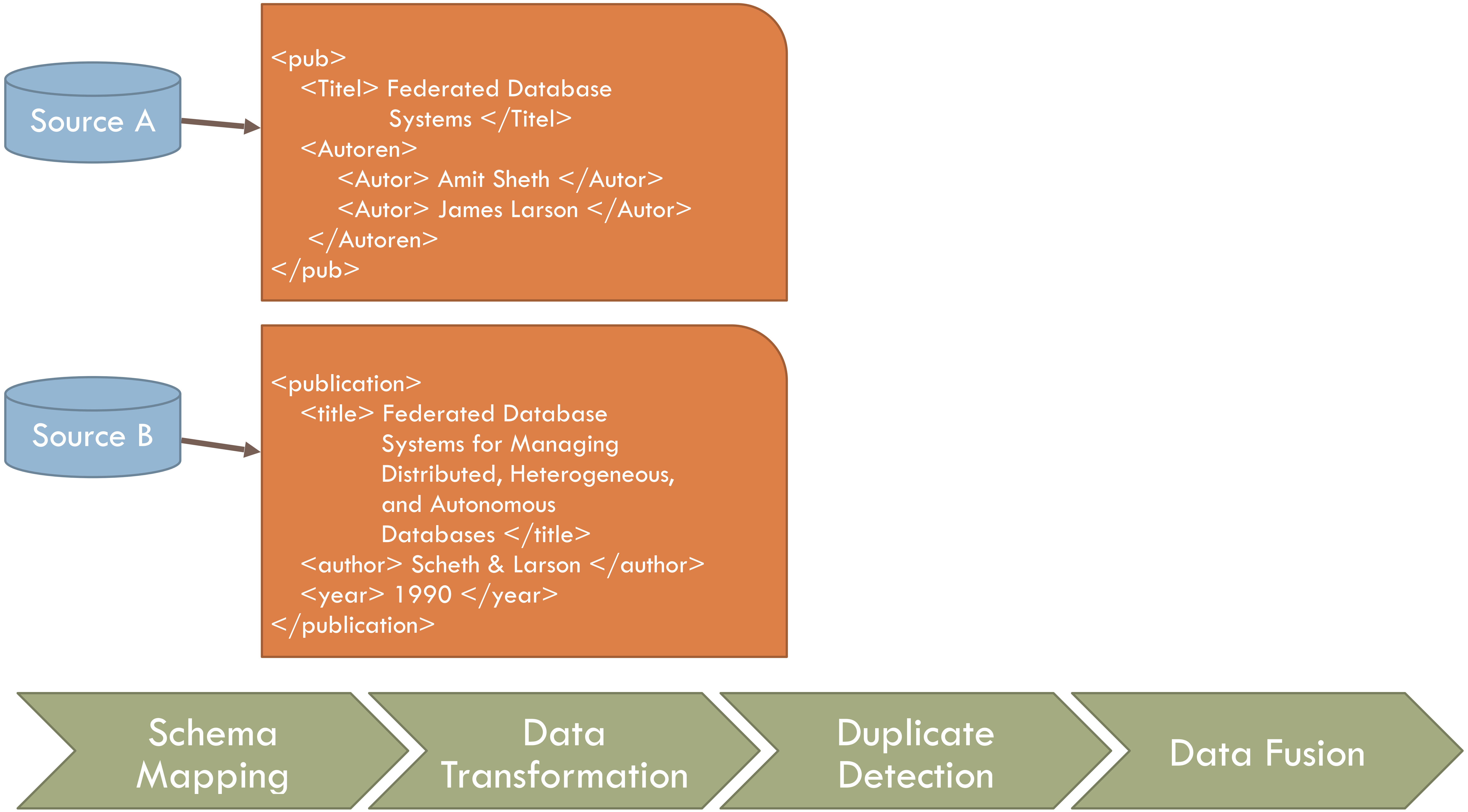
- Conflict resolution strategies
- "Truth-discovery" techniques
  - Accuracy
  - Freshness
  - Dependence
- Fusion Issues
  - Accuracy
  - Efficiency
  - Usability
  - How fusion fits with the rest of data integration?

Data Conflicts



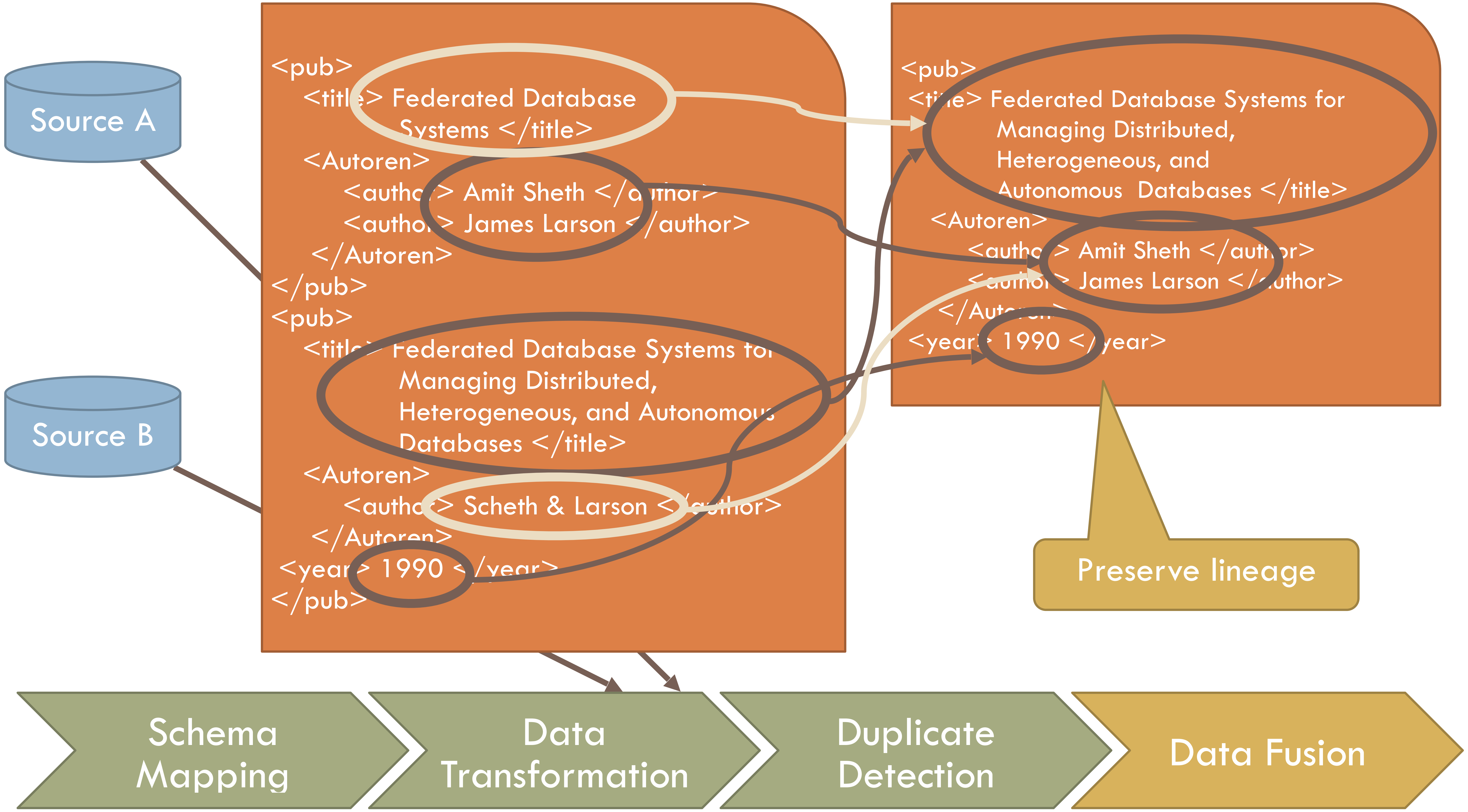
[L. Dong and F. Naumann, 2009]

Information Integration



[L. Dong and F. Naumann, 2009]

Information Integration

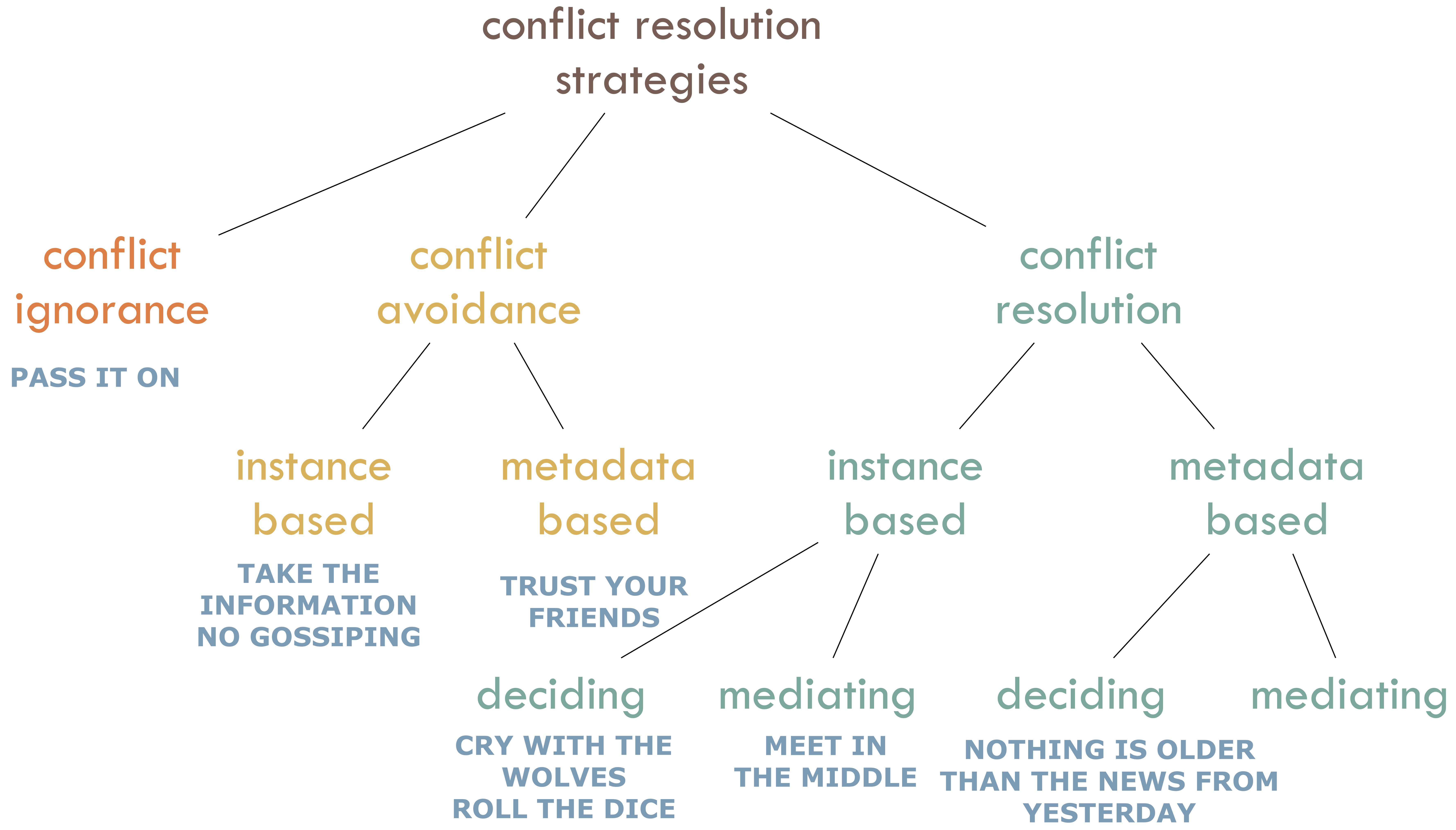


[L. Dong and F. Naumann, 2009]

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