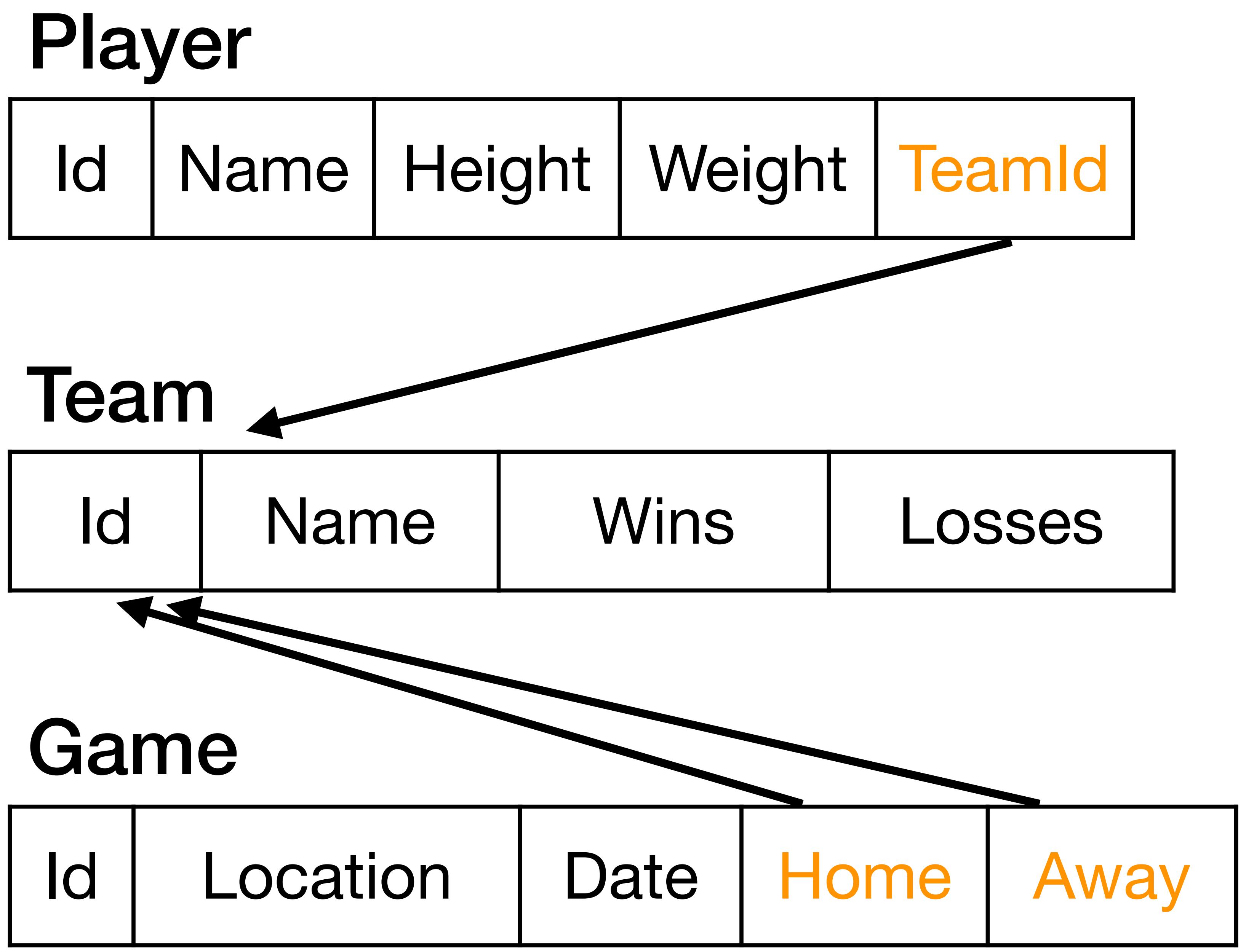
Advanced Data Management (CSCI 640/490)

Data Fusion

Dr. David Koop

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.

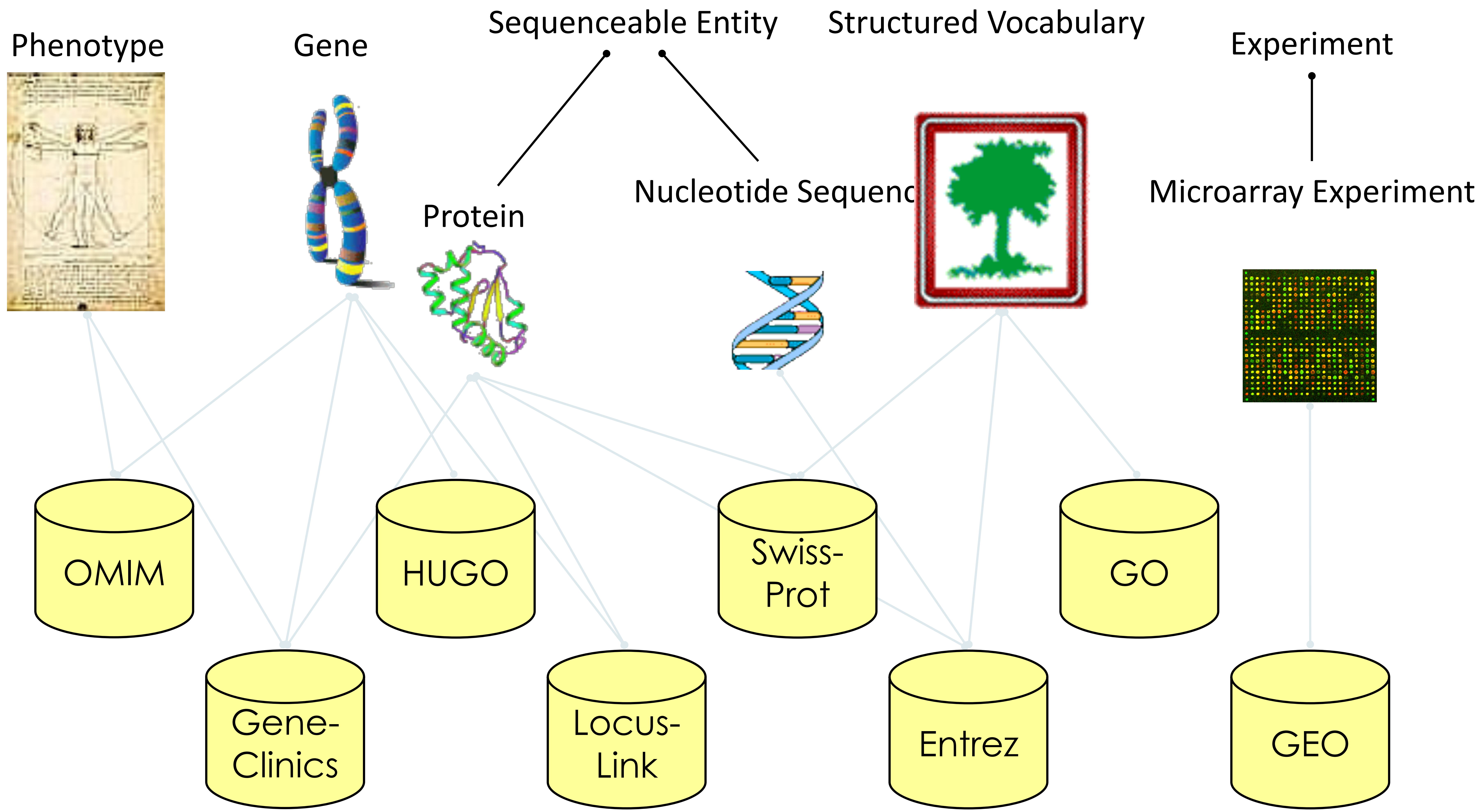
S1	S2	S3	S4	S5
Movies: name, actors, director, genre	Cinemas: place, movie, start	Cinemas in NYC: cinema, title, startTime	Cinemas in SF: location, movie, startingTime	Reviews: title, date grade, review

[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

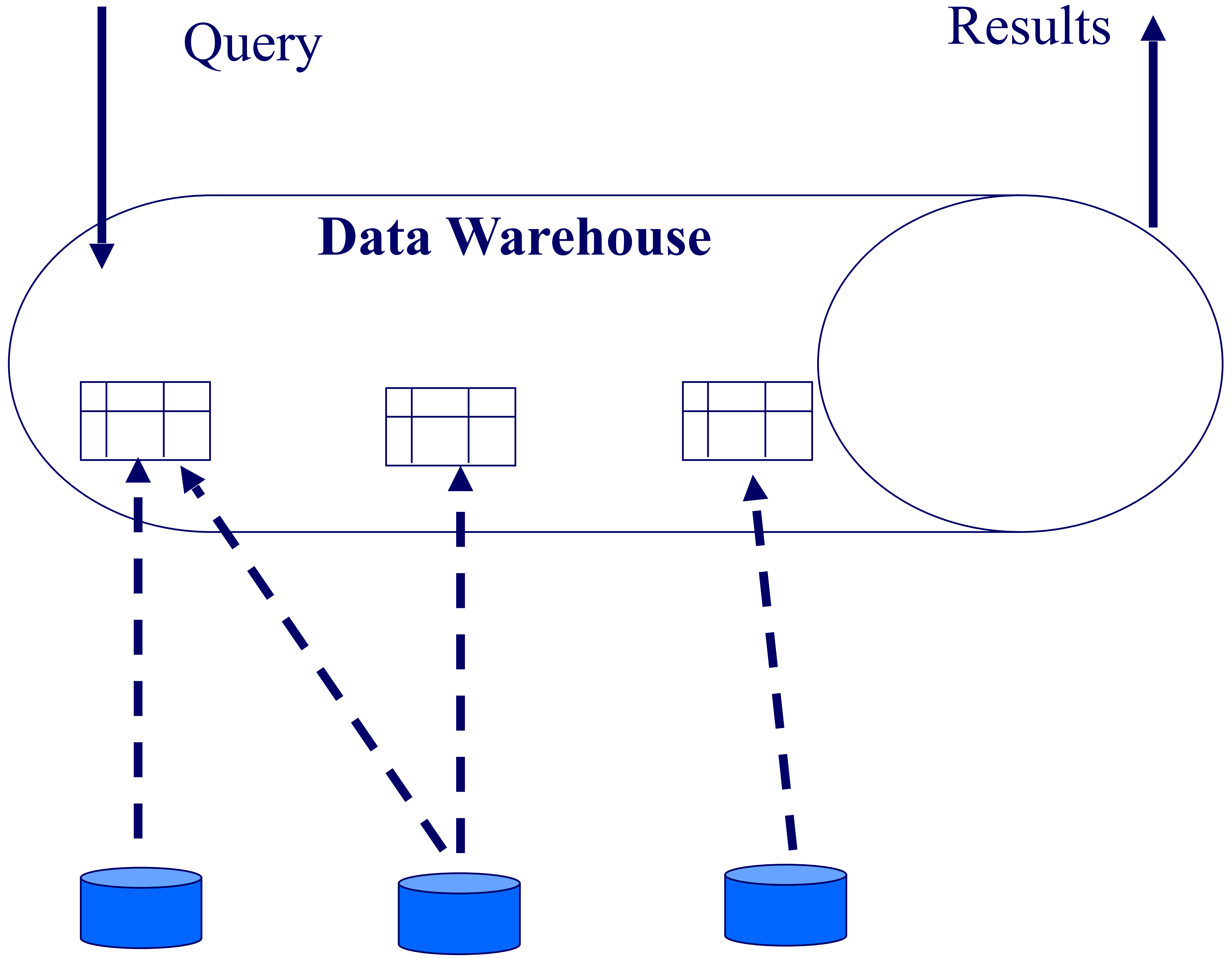
Data Integration Application: Biomedical



[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

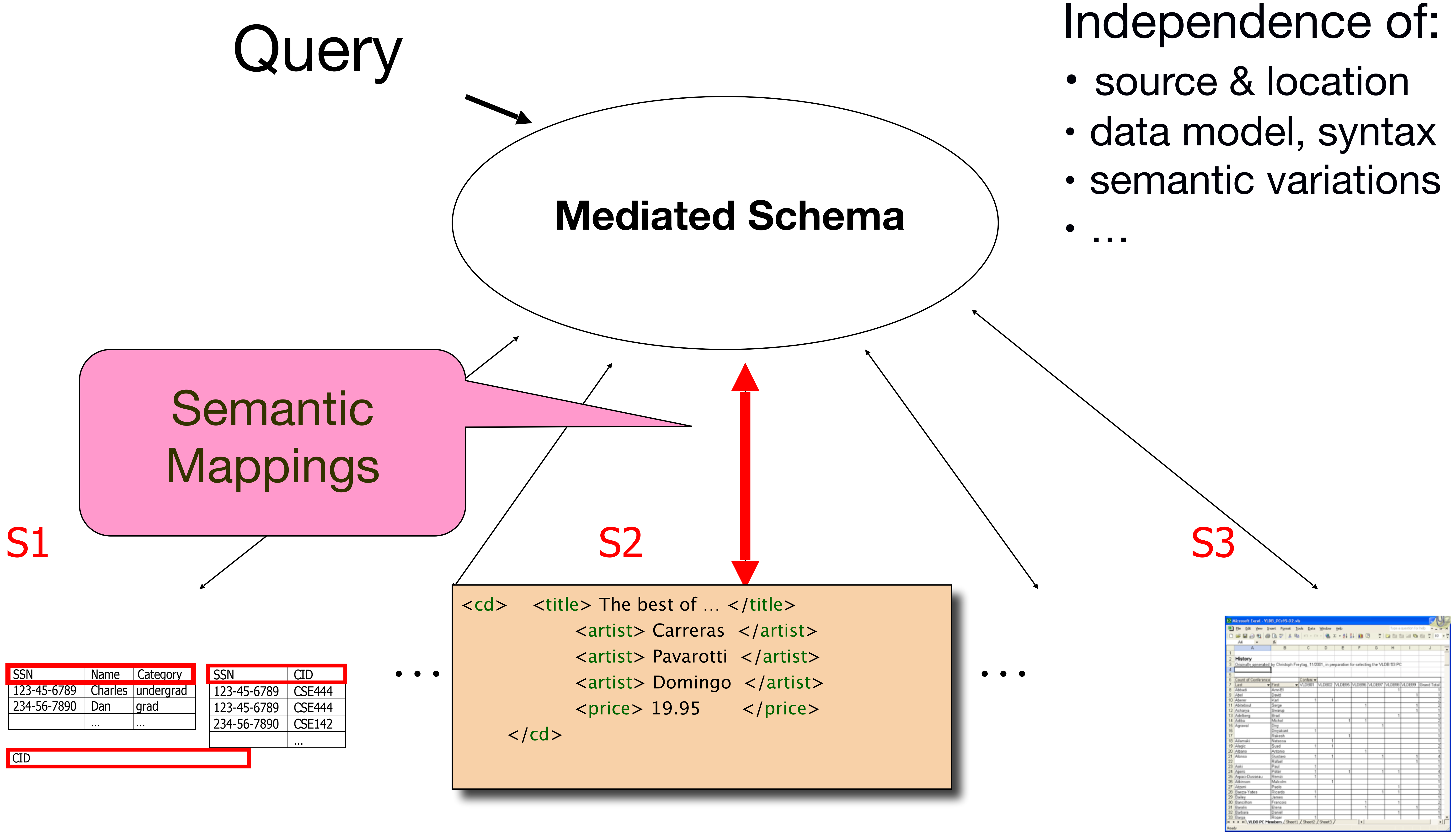


[A. Doan et al., 2012]

Virtual Data Warehouses

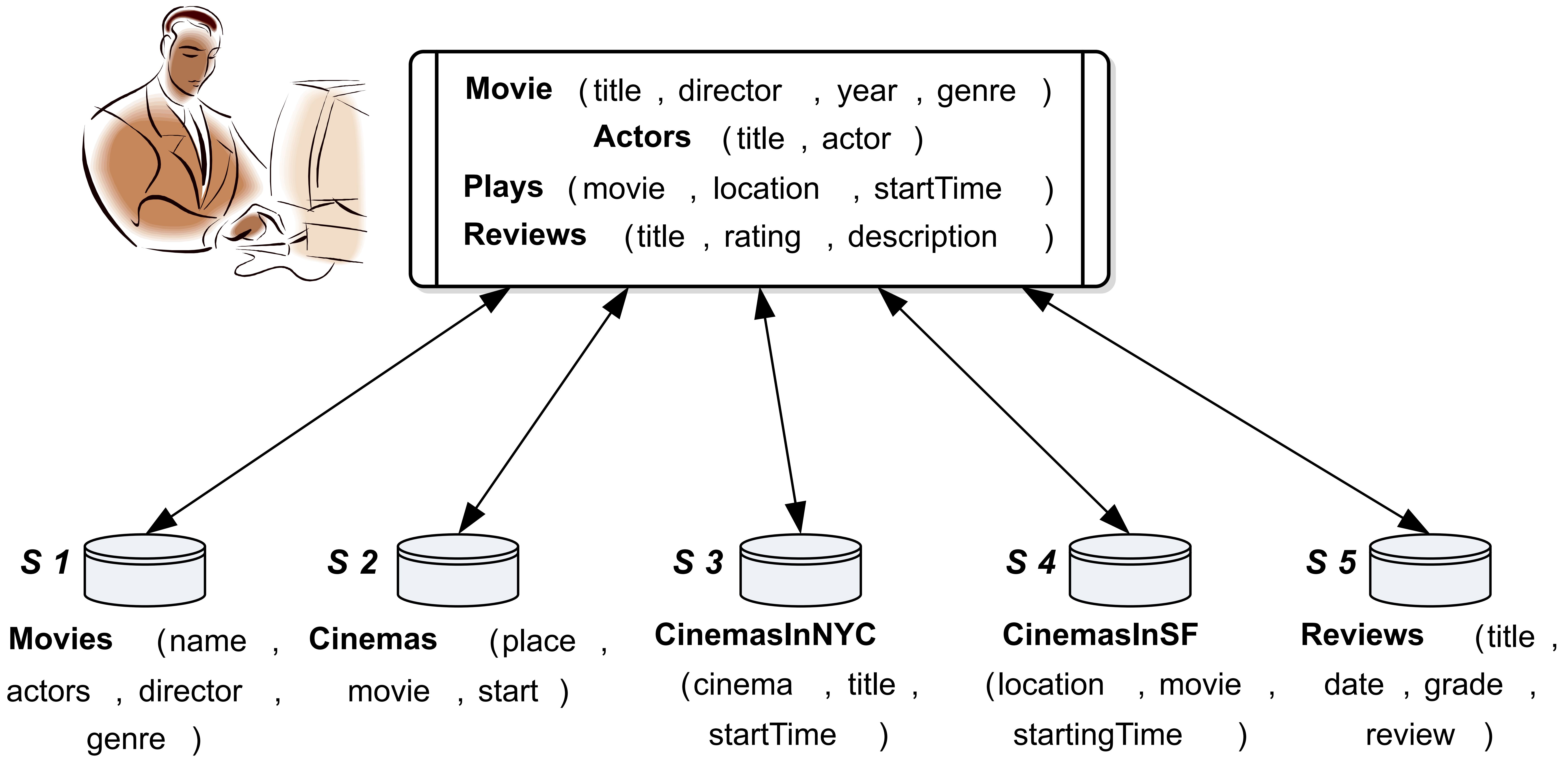
Independence of:

- 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]

Data Fusion

- Definition: Resolving conflicting data and verifying facts.
- Example: "Google, How long is the Mississippi River?"

Mississippi River / Length

2,320 mi

People also search for

Missouri River 2.341K mi

Nile 4.258K mi

Mississippi River

River in the United States of America

4.2 ★★★★ 400 Google reviews

The Mississippi River is the chief river of the second-largest drainage system on the North American continent, second only to the Hudson Bay drainage system. Wikipedia

Discharge: 593,000 cubic feet per second

Basin area: 1.151 million mi²

Source: Lake Itasca

Mouth: Gulf of Mexico

Country: United States of America

Did you know: The Mississippi River is the second-longest river in the US (2,202 mi). wikipedia.org

Mississippi River Facts - Mississippi National River and Recreation ...

https://www.nps.gov/miss/riverfacts.htm

Nov 14, 2017 - The staff of Itasca State Park at the Mississippi's headwaters suggest the main stem of the river is **2,552 miles** long. The US Geologic Survey has published a number of **2,300 miles**, the EPA says it is **2,320 miles** long, and the Mississippi National River and Recreation Area suggests the river's length is **2,350 miles.**

Longest main-stem rivers of the United States

#	Name	Mouth[5]	Length	Source coordinates[11]	Mouth coordinates[11]	Watershed area[12]	Discharge[12]	States, provinces, and image[5][11]
1	Missouri River	Mississippi River	2,341 mi 3,768 km[13]	45°55′39″N 111°30′29″W[14]	38°48′49″N 90°07′11″W	529,353 mi² 1,371,017 km²[15] ‡[n 2]	69,100 ft³/s 1,956 m³/s [n 3]	Montana[s], North Dakota, South Dakota, Nebraska, Iowa, Kansas, Missouri[m]
2	Mississippi River	Gulf of Mexico	2,202 mi 3,544 km[17] [n 4]	47°14′22″N 95°12′29″W[18]	29°09′04″N 89°15′12″W	1,260,000 mi² 3,270,000 km²[19] ‡[n 5]	650,000 ft³/s 18,400 m³/s	Minnesota[s], Wisconsin, Iowa, Illinois, Missouri, Kentucky, Tennessee, Arkansas, Mississippi, Louisiana[m]

[X. L. Dong and T. Rekatsinas, 2018]

Data Fusion

Source Observations

Source	River	Attribute	Value
KG	Mississippi River	Length	2,320 mi
KG	Missouri River	Length	2,341 mi
Wikipedia	Mississippi River	Length	2,202 mi
Wikipedia	Missouri River	Length	2,341 mi
USGS	Mississippi River	Length	2,340 mi
USGS	Missouri River	Length	2,540 mi

Fact

Conflicting value

Source reports
a value for a fact

True Facts

River	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	?

Fact's true value

Goal: Find the **latent** true value of facts.

[X. L. Dong and T. Rekatsinas, 2018]

Data Fusion

Source Observations

Source	River	Attribute	Value
KG	Mississippi River	Length	2,320 mi
KG	Missouri River	Length	2,341 mi
Wikipedia	Mississippi River	Length	2,202 mi
Wikipedia	Missouri River	Length	2,341 mi
USGS	Mississippi River	Length	2,340 mi
USGS	Missouri River	Length	2,540 mi

Fact

Conflicting value

Source reports a value for a fact

True Facts

River	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	?

Fact's true value

Idea: Use *redundancy* to infer the true value of each fact.

[X. L. Dong and T. Rekatsinas, 2018]

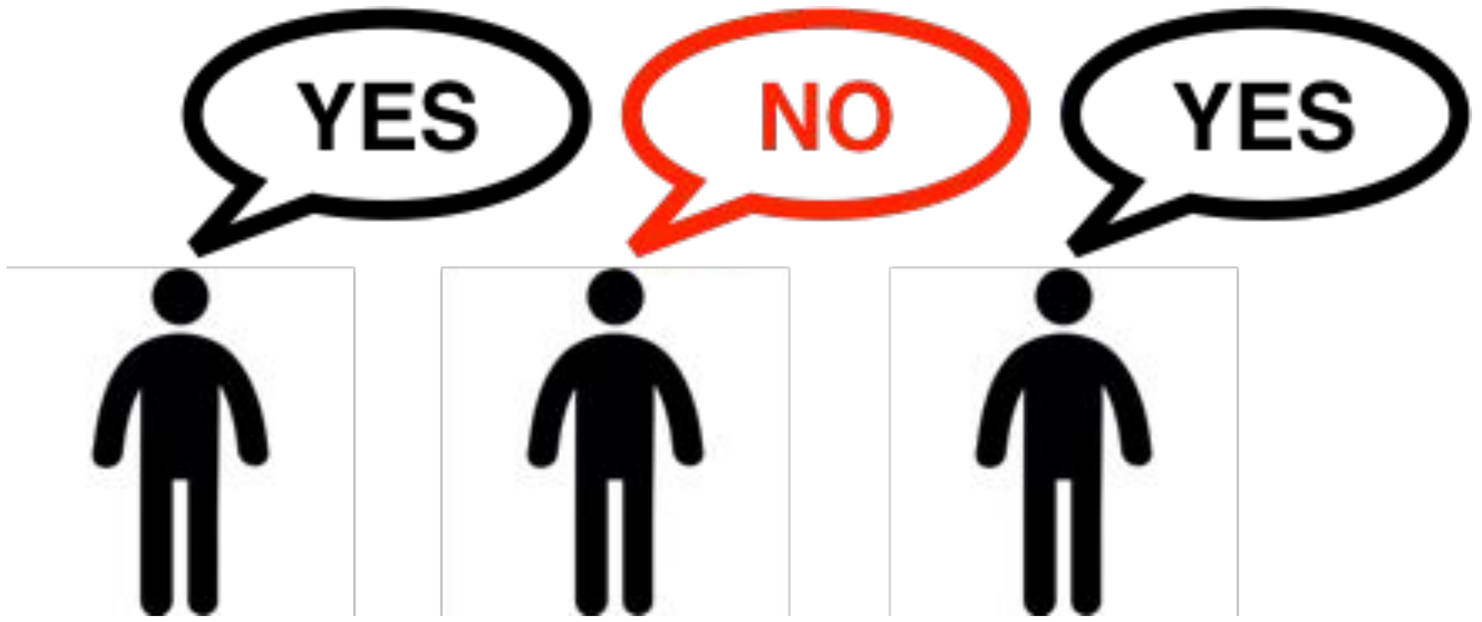
Data Fusion

Source Observations

Source	River	Attribute	Value
KG	Mississippi River	Length	2,320 mi
KG	Missouri River	Length	2,341 mi
Wikipedia	Mississippi River	Length	2,202 mi
Wikipedia	Missouri River	Length	2,341 mi
USGS	Mississippi River	Length	2,340 mi
USGS	Missouri River	Length	2,540 mi

True Facts

River	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	2,341



Majority voting can be limited. What if sources are correlated (e.g., copying)?

Idea: Model source quality for accurate results.

MV's assumptions

1. Sources report values independently
2. Sources are better than chance.

[X. L. Dong and T. Rekatsinas, 2018]

Reading Quiz

Assignment 3

- Salary Data
- Use Pandas (not loops)
- Part 2: CSCI 640 students need to do (b), CSCI 490 students can choose
- Part 5: use melt/pivot or a similar high-level operation

Today

- Due to faculty candidate visit, office hours changed (12:30-1:30pm)
- Faculty candidate talks both today and tomorrow
  - 3:30 to 4:30 FW 201

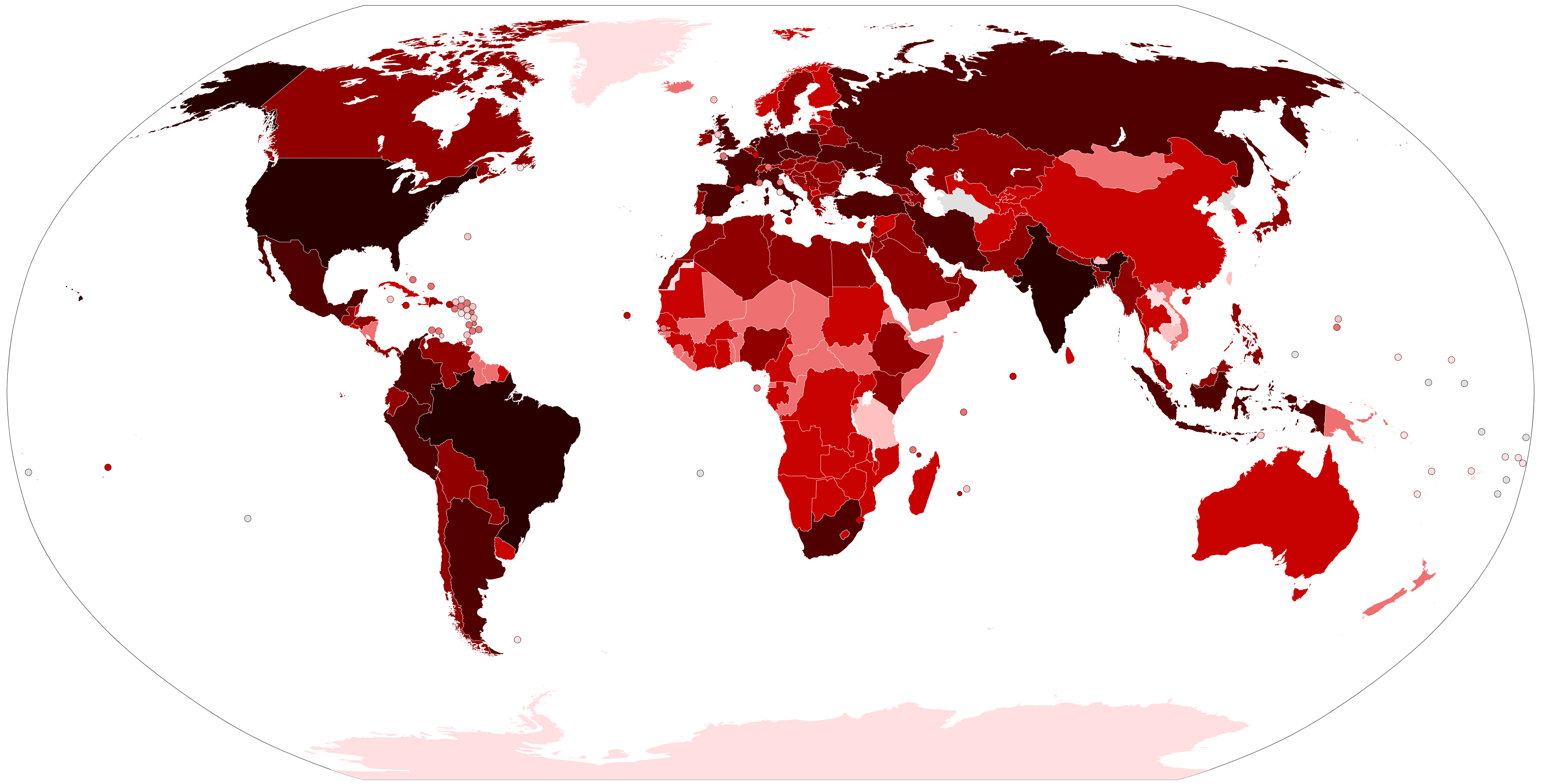
Record Linkage Motivation

- 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)

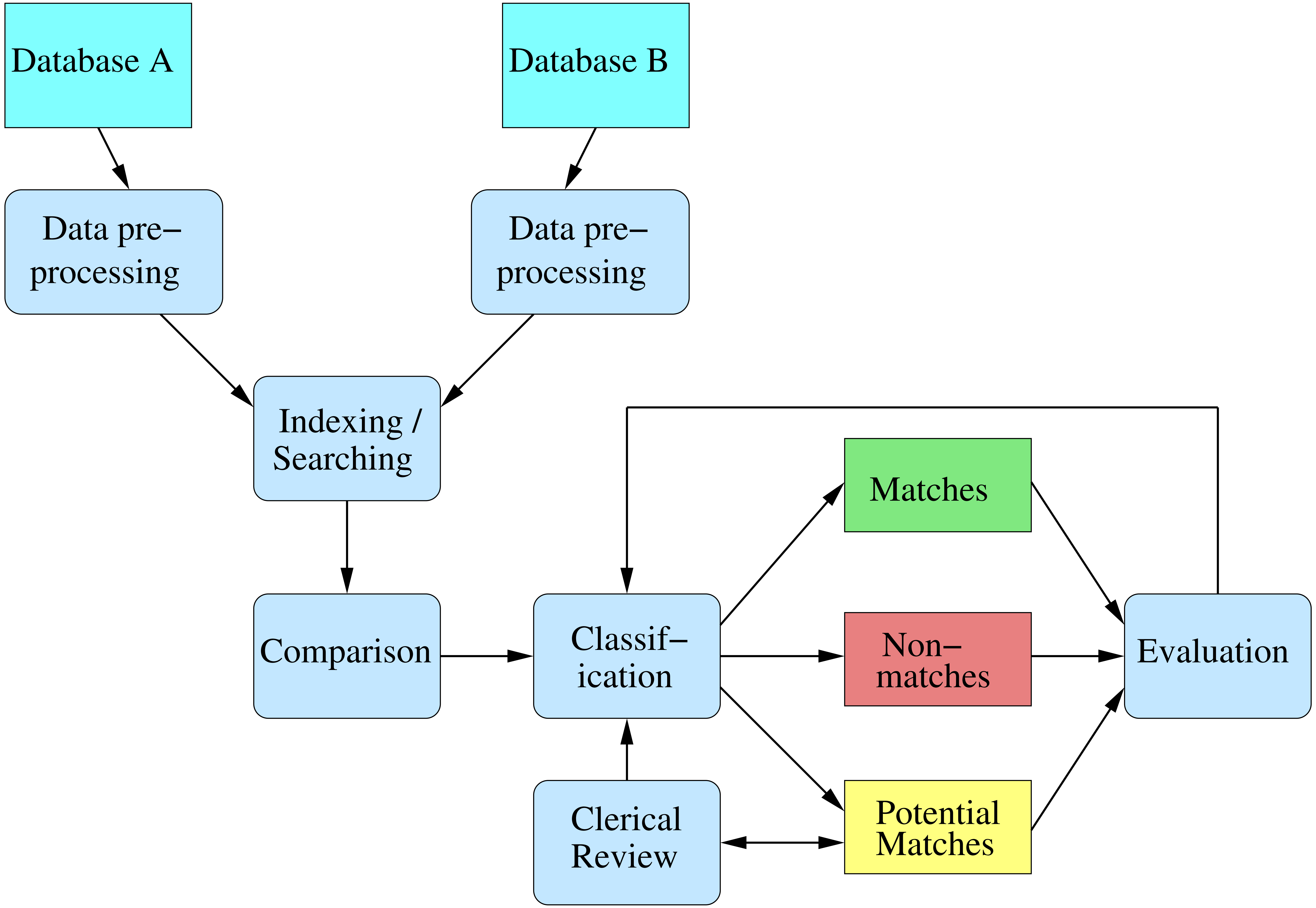


[P. Christen , 2019], image: [Pharexia, Wikipedia]

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]

Record Linkage/Entity Resolution Recipe

- Problem: Link references to the same entity
- Short Answers:
  - Random Forest with attribute similarity features
  - Deep Learning to handle text and noise
  - End-to-end solutions still being worked on

[X. L. Dong and T. Rekatsinas, 2018]

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

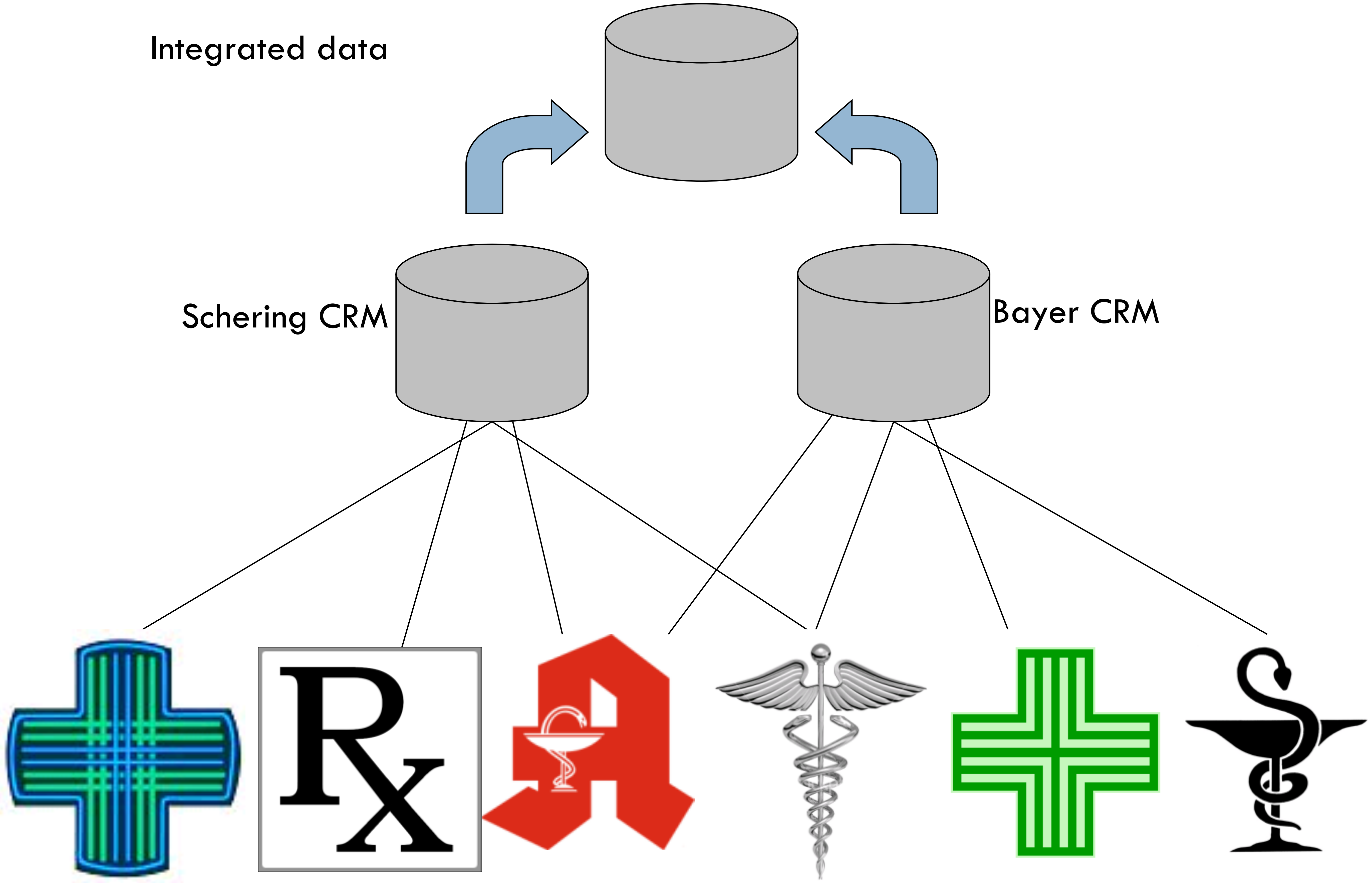
Data Fusion— Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann

Data Fusion Summary

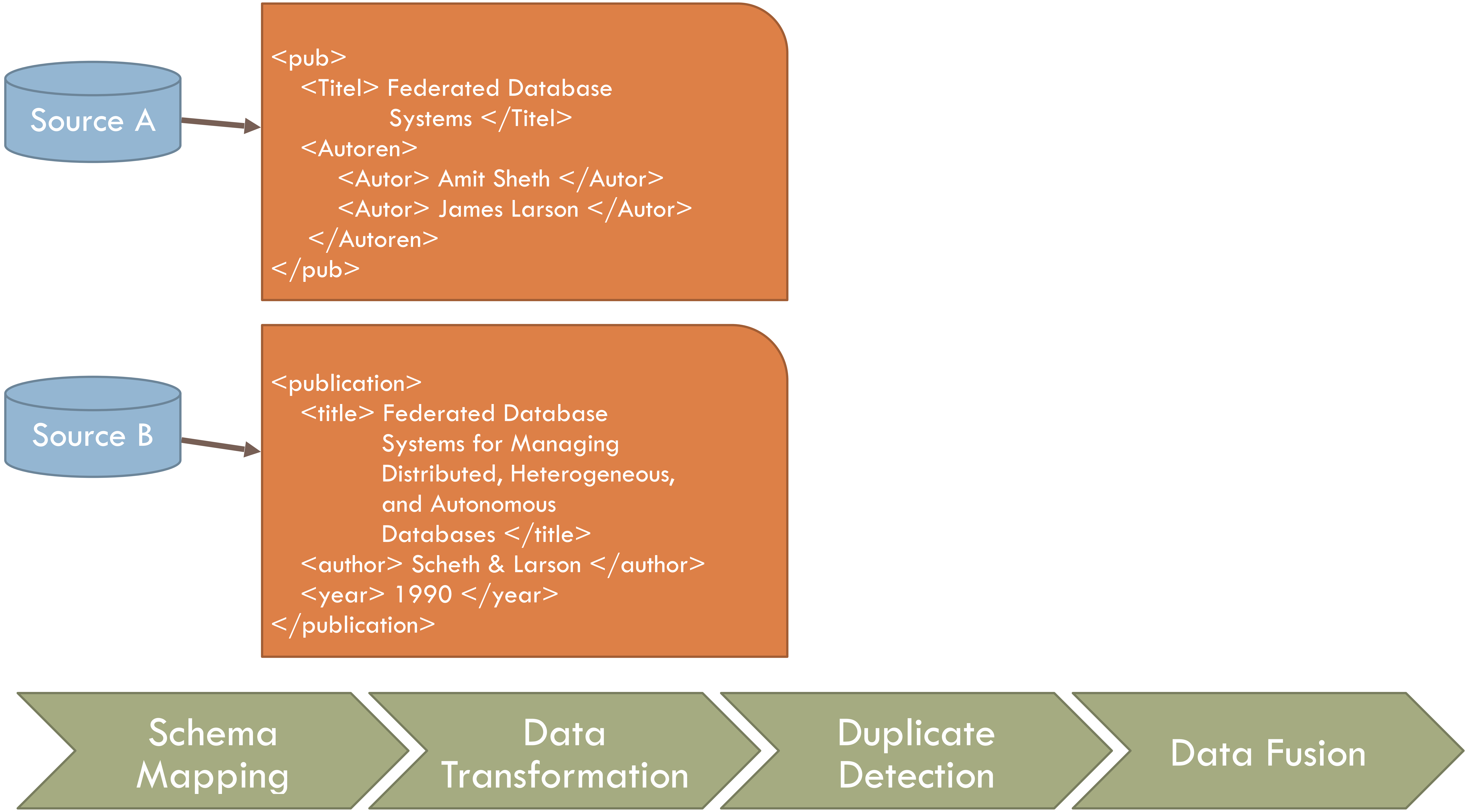
- Conflict resolution strategies
- "Truth-discovery" techniques
  - Accuracy
  - Freshness
  - Dependence
- Fusion Issues
  - Accuracy
  - Efficiency
  - Usability
  - How fusion fits with the rest of data integration?

Data Conflicts



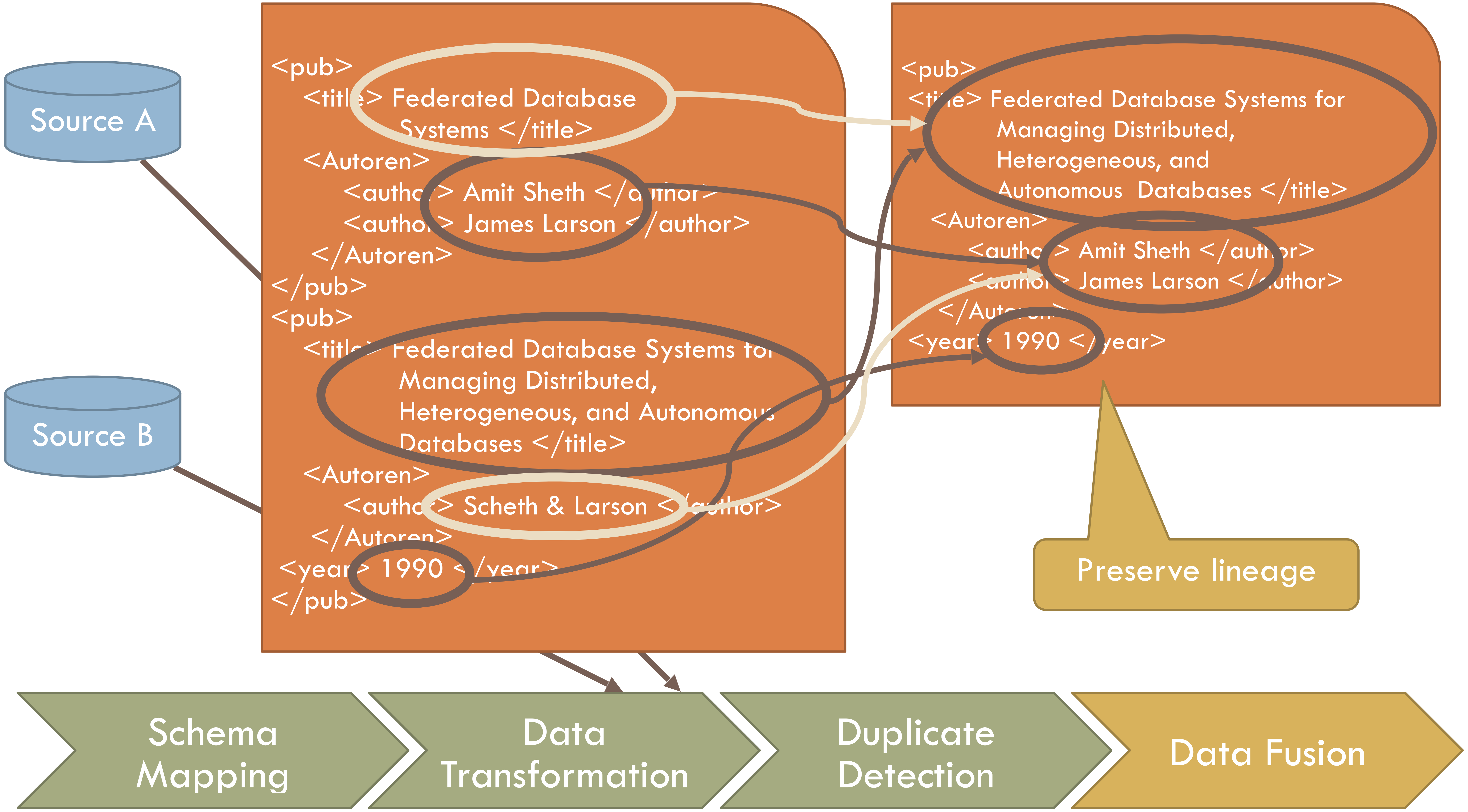
[L. Dong and F. Naumann, 2009]

Information Integration



[L. Dong and F. Naumann, 2009]

Information Integration

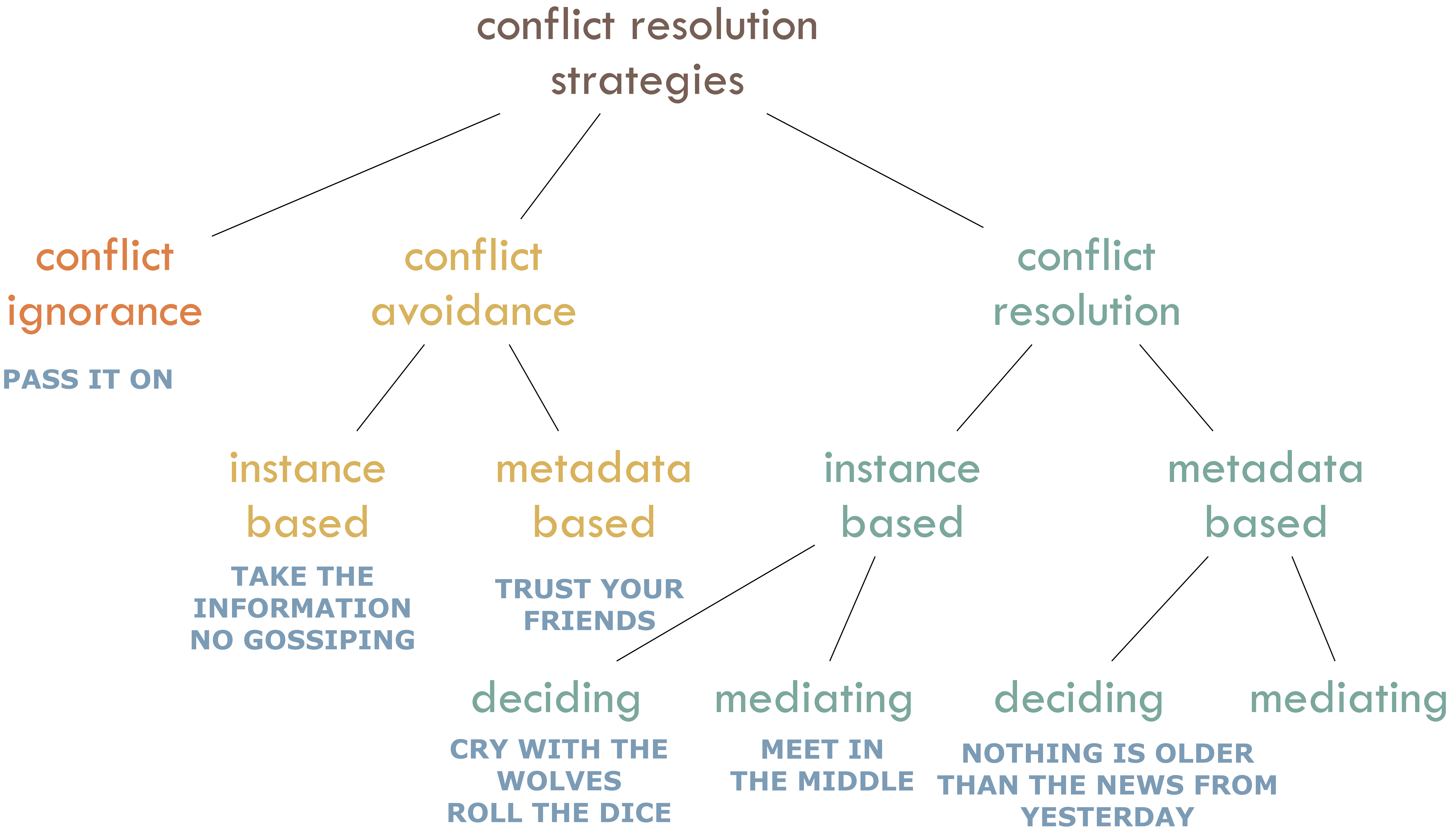


[L. Dong and F. Naumann, 2009]

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

Example Problem

[X L Dong et al., 2009]

Example Problem

	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

[X L Dong et al., 2009]

Naive Voting Works

	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

[X L Dong et al., 2009]

Naive Voting Only Works if Data Sources are Independent

[X L Dong et al., 2009]

Naive Voting Only Works if Data Sources are Independent

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

S4 and S5 copy from S3

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

S4 and S5 copy from S3

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

2. With only a snapshot it is hard to decide which source is a copier.

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

1. Sharing common data does not in itself imply copying.

2. With only a snapshot it is hard to decide which source is a copier.

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

3. A copier can also provide or verify some data by itself, so it is inappropriate to ignore all of its data.

[X L Dong et al., 2009]

Source Dependence

- Source dependence: two sources S and T deriving the same part of data directly or transitively from a common source (can be one of S or T).
  - Independent source
  - Copier
    - copying part (or all) of data from other sources
    - may verify or revise some of the copied values
    - may add additional values
- Assumptions
  - Independent values
  - Independent copying
  - No loop copying

[X L Dong et al., 2009]

Core Case

- Conditions
  - Same source accuracy
  - Uniform false-value distribution
  - Categorical value
- Proposition: W. independent "good" sources, Naïve voting selects values with highest probability to be true.

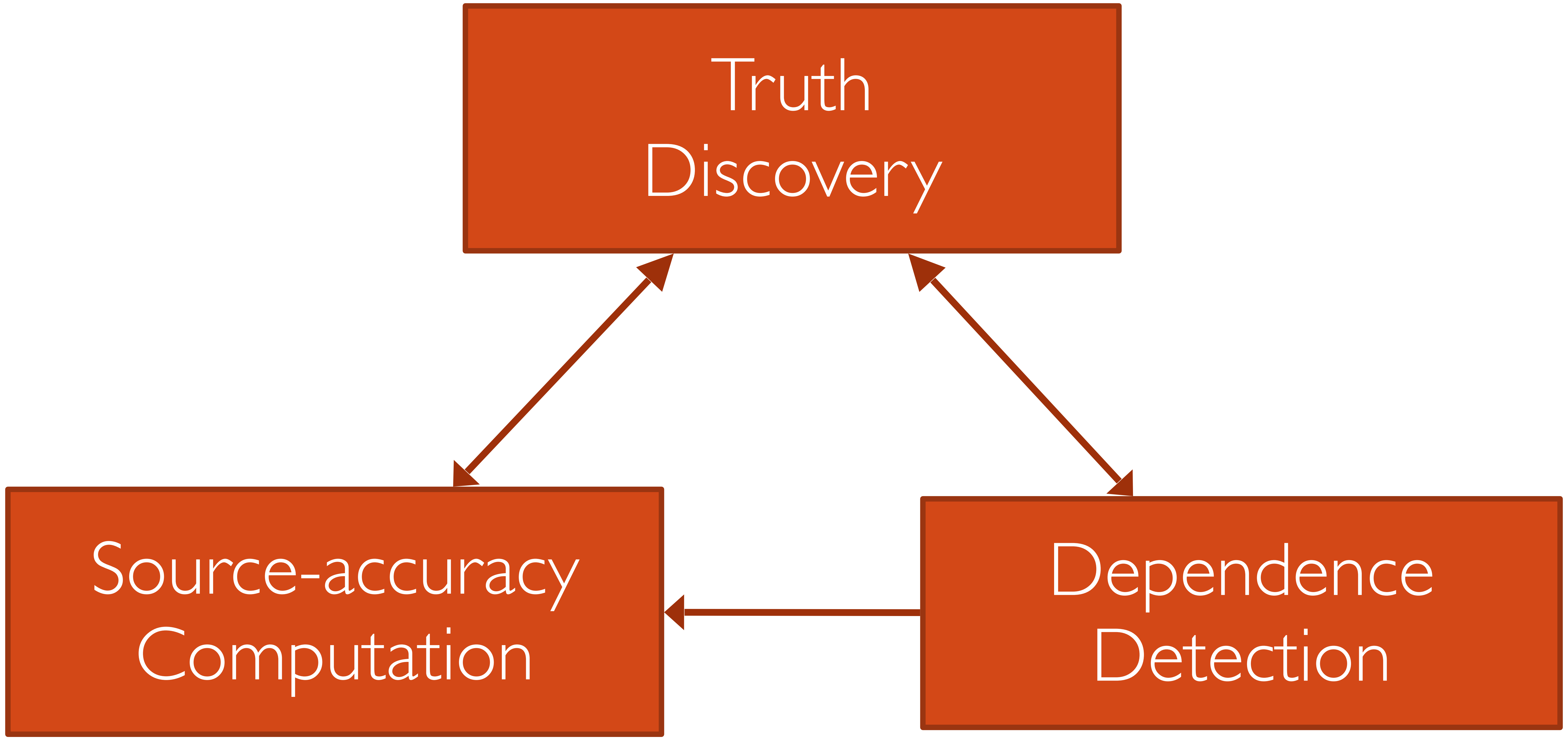
[X L Dong et al., 2009]

Ideas

- If two sources share a lot of false values, they are more likely to be dependent.
- S1 is more likely to copy from S2, if the accuracy of the common data is highly different from the accuracy of S1.

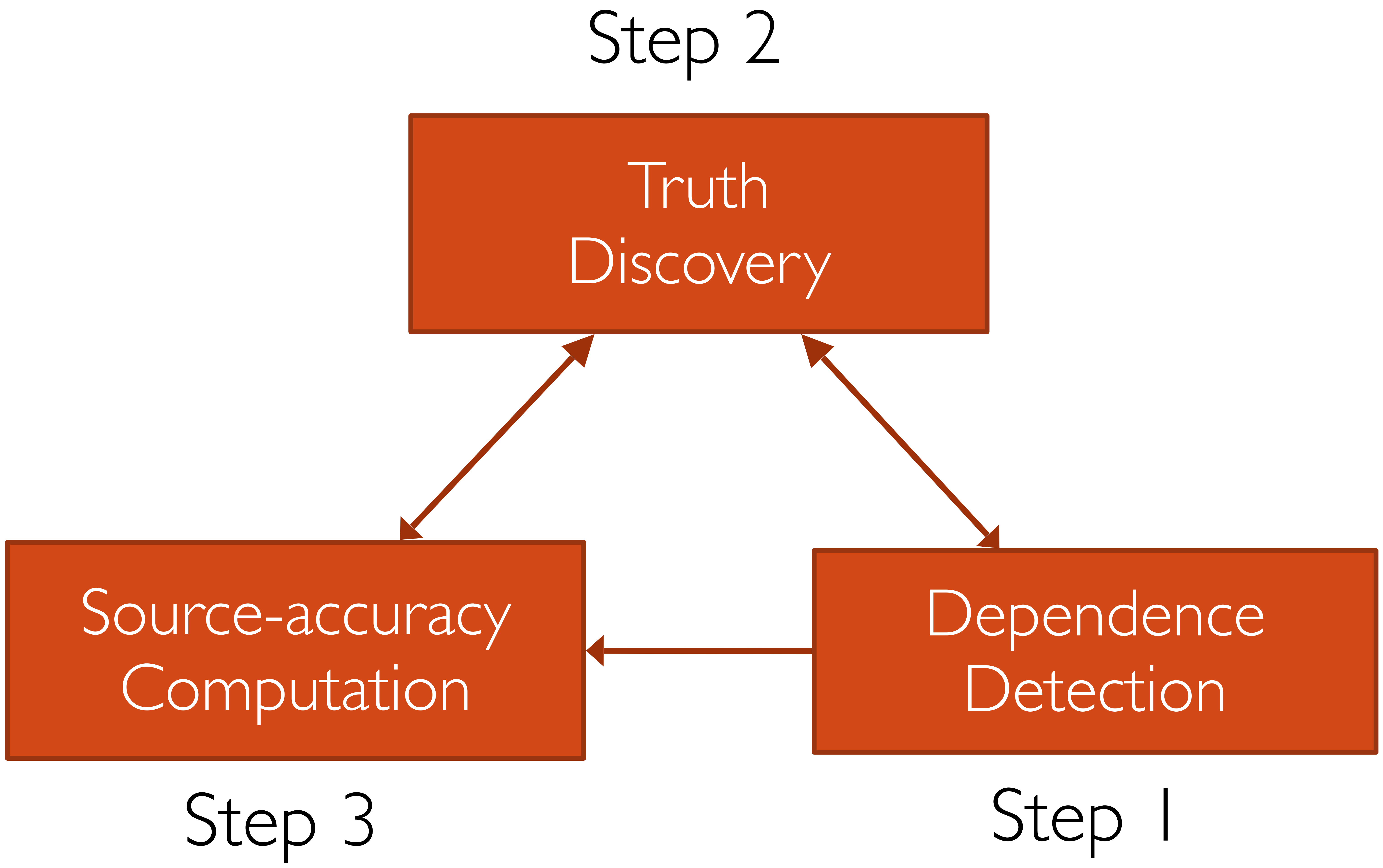
[X L Dong et al., 2009]

Combining Accuracy and Dependence



[X L Dong et al., 2009]

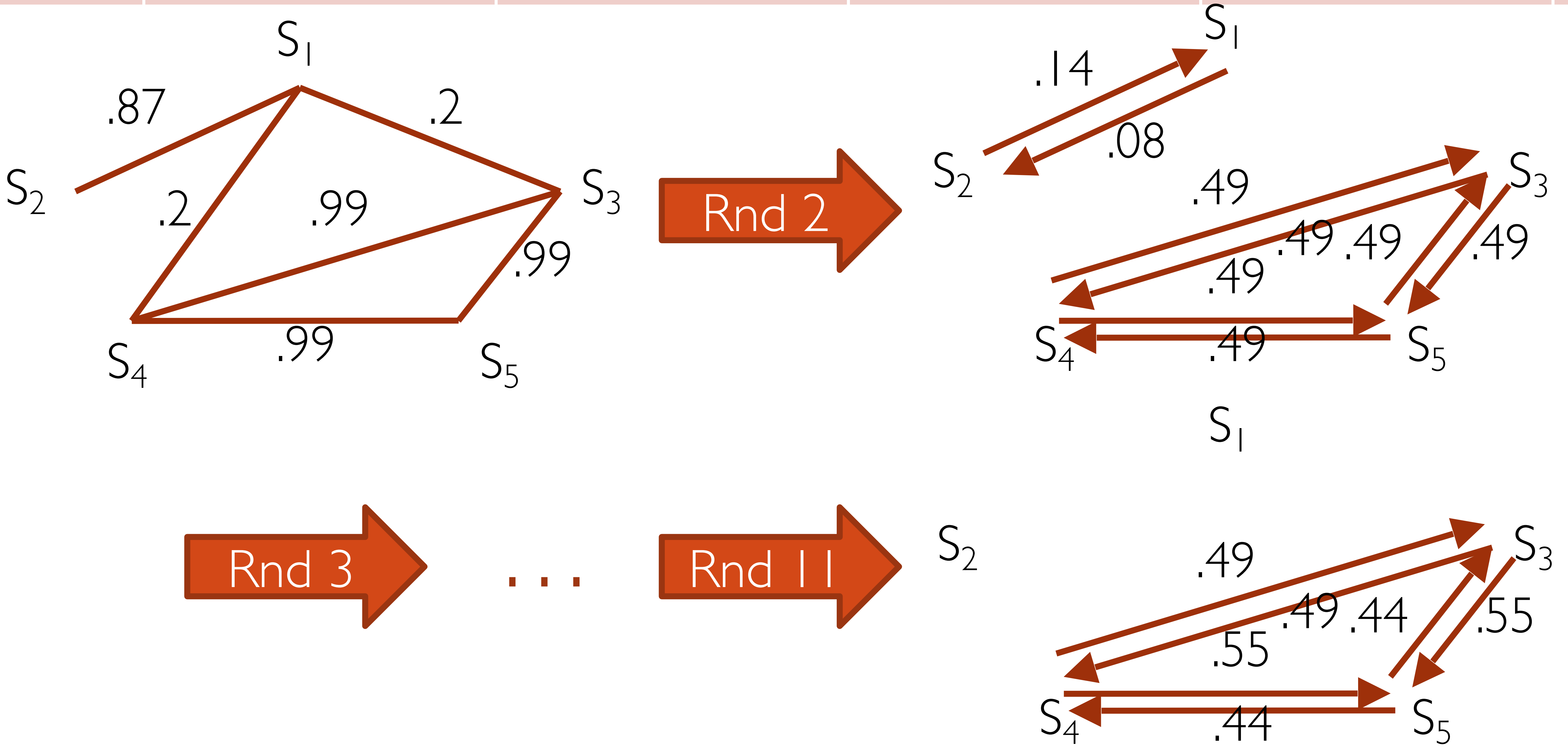
Combining Accuracy and Dependence



[X L Dong et al., 2009]

The Motivating Example

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW



[X L Dong et al., 2009]

The Motivating Example

Accuracy	S1	S2	S3	S4	S5
Round 1	.52	.42	.53	.53	.53
Round 2	.63	.46	.55	.55	.55
Round 3	.71	.52	.53	.53	.37
Round 4	.79	.57	.48	.48	.31
…	…	…	…	…	…
Round 11	.97	.61	.40	.40	.21

Value Confidence	Carey			Halevy	
	UCI	AT&T	BEA	**Google**	UW
Round 1	1.61	1.61	2.0	2.1	2.0
Round 2	1.68	1.3	2.12	2.74	2.12
Round 3	2.12	1.47	2.24	3.59	2.24
Round 4	2.51	1.68	2.14	4.01	2.14
…	…	…	…	…	…
Round 11	4.73	2.08	1.47	6.67	1.47

[X L Dong et al., 2009]