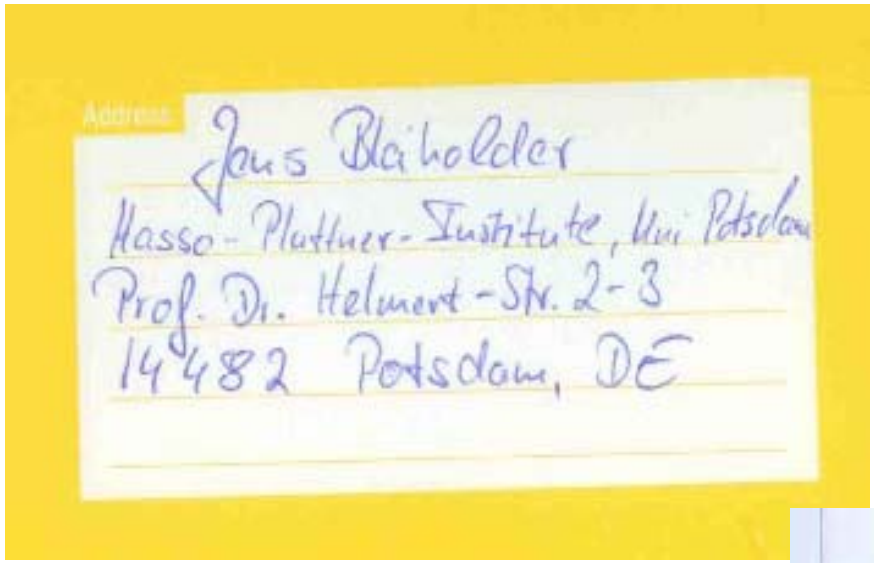
DATA FUSION – RESOLVING DATA CONFLICTS IN INTEGRATION

Tutorial at
VLDB 2009

Xin Luna Dong – AT&T Labs-Research
Felix Naumann – Hasso Plattner Institute (HPI)

Origins of Data Conflicts

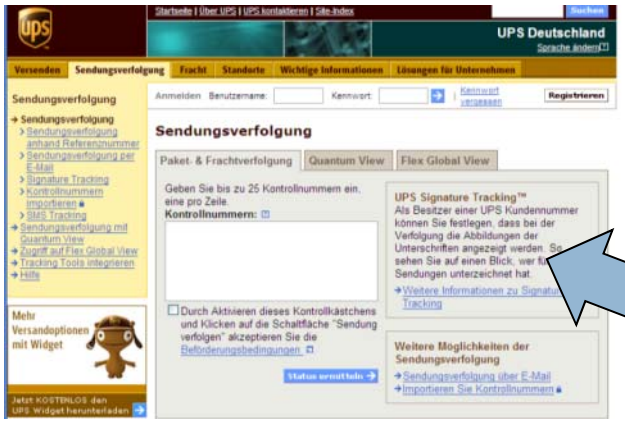


Original

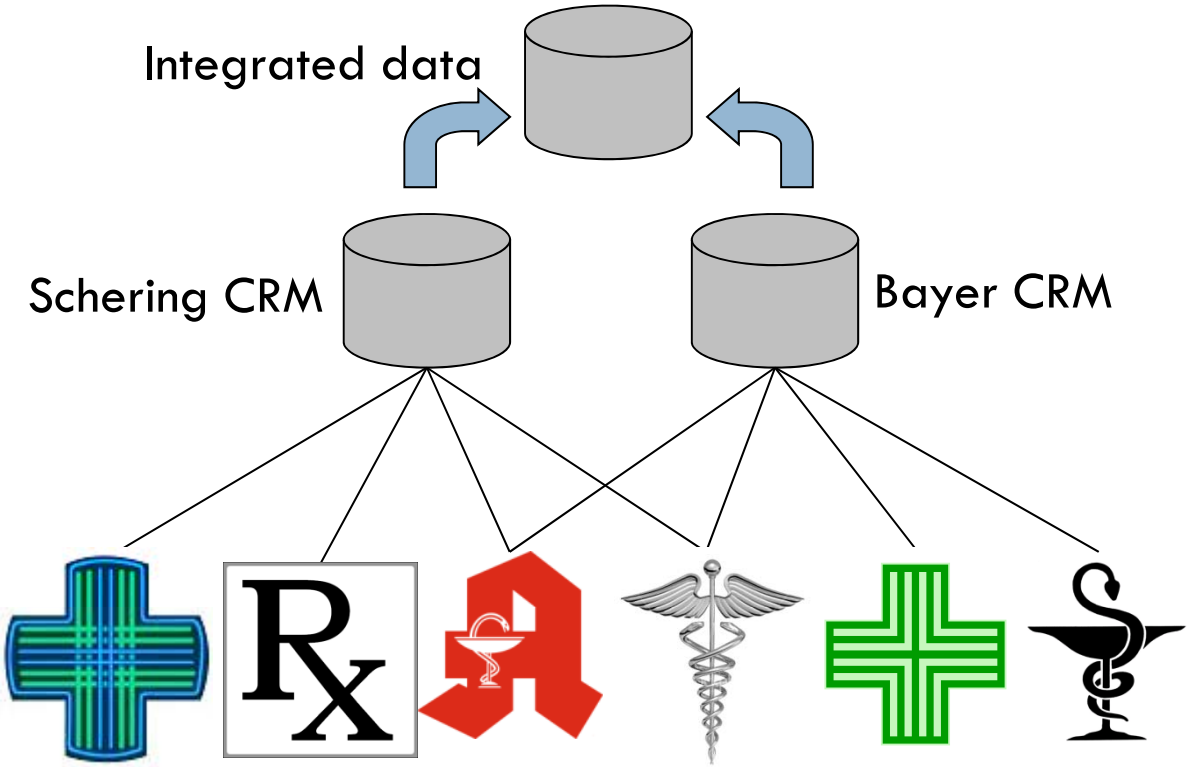
ACM Computing Survey [BN08]

Scanned

Origins of Data Conflicts



Origins of Data Conflicts



Origins of Data Conflicts: German Names

Origins of Data Conflicts: Difficult Names

488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny spear
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spea
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spear
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spear
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears	3 britnesy spear
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears	3 britnetty spea
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxxx spe
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears	3 britnity spear
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey spear
1338 britiny spears	26 btittany spears	9 rbitney spears	4 brbritney spears	3 britnyey spear
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 britterny spea
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey spea
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 brittnney spea
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears	3 brittnyey spea
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityen spears
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney spear
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 brltney spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny spears
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany spear
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay spears
544 brittaney spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney spear
544 brittnay spears	21 biteny spears	8 britnrey spears	4 britaby spears	3 brtitany spear
364 britey spears	21 bratney spears	8 britnty spears	4 britaey spears	3 brtiteny spear
364 brittiny spears	21 britani spears	8 brittner spears	4 britainey spears	3 brtnet spears
329 brtney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny spears
269 bretney spears	21 briteany spears	7 baritney spears	4 britinney spears	3 btney spears
269 britneys spears	21 brittay spears	7 birntey spears	4 britmney spears	3 drittney spear
244 britne spears	21 brittinay spears	7 biteney spears	4 britnear spears	3 pretney spears
244 brytney spears	21 brtany spears	7 bitiny spears	4 britnel spears	3 rbritney spear
220 breatney spears	21 brtiany spears	7 breateny spears	4 britneuy spears	2 barittany spea
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney spea
199 britnney spears	19 brirtney spears	7 brintye spears	4 britnmey spears	2 bbitney spears
163 britnry spears	19 britnaey spears	7 britianny spears	4 brittaby spears	2 bbritny spears

Origins of Intra-Source Conflicts

- No integrity or consistency checks
- Redundant schemata
- Typos, transmission errors, incorrect calculations
- Variants
  - Kantstr. / Kantstrasse / Kant Str. / Kant Strasse
  - Kolmogorov / Kolmogoroff / Kolmogorow
- Typical confusion (OCR)
  - U<->V, 0<->o, 1<->l, etc.
- Obsolete values
  - Different update frequencies, forgotten update

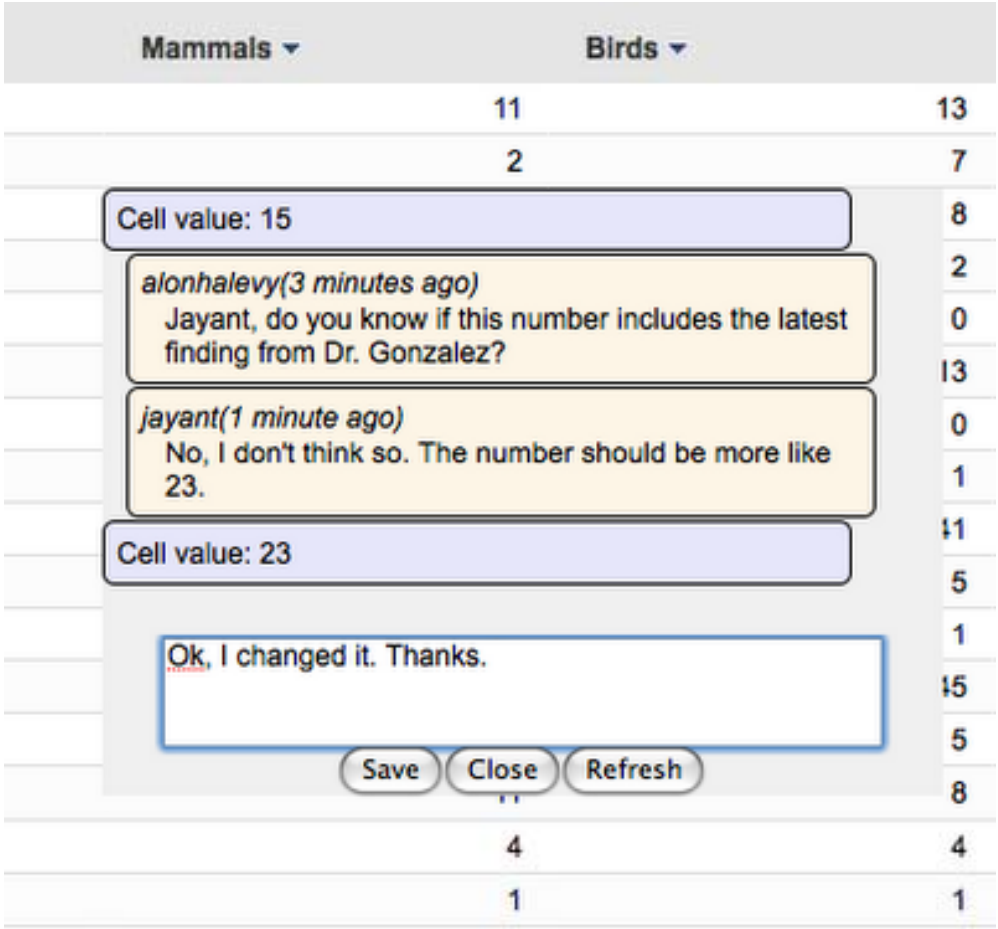
Origins of Inter-Source Conflicts

- Locally consistent but globally inconsistent
- Different data types
- Local spelling variations and conventions
  - Addresses
    - St → Street, Ave → Avenue, etc.
    - R.-Breitscheid-Str. 72 a → Rudolf-Breitscheid.-Str. 72A
    - 128 spellings for Frankfurt am Main
      - Frankfurt a.M., Frankfurt/M, Frankfurt, Frankfurt a. Main, …
  - Names
    - Dr. Ing. h.c. F. Porsche AG
    - Hewlett-Packard Development Company, L.P.
  - Numerical data
    - 10.000 € = 10T EURO = 10k EUR = 10.000,00€ = 10,000.- €
  - Phone numbers, birth dates, etc.

Resolution of Data Conflicts?

- "… focus is on fusing data management and collaboration: **merging** multiple data sources, discussion of the data, querying, visualization, and Web publishing."
- "The power of data is truly harnessed when you combine data from multiple sources. Fusion Tables enables you to fuse multiple sets of data when they are about the **same entities**. In database speak, we call this a **join** on a primary key but the data originates from multiple independent sources."

Web Integration—Google Fusion Tables

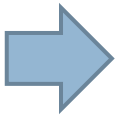


- Allows discussion of values between users

Data Conflict Elimination

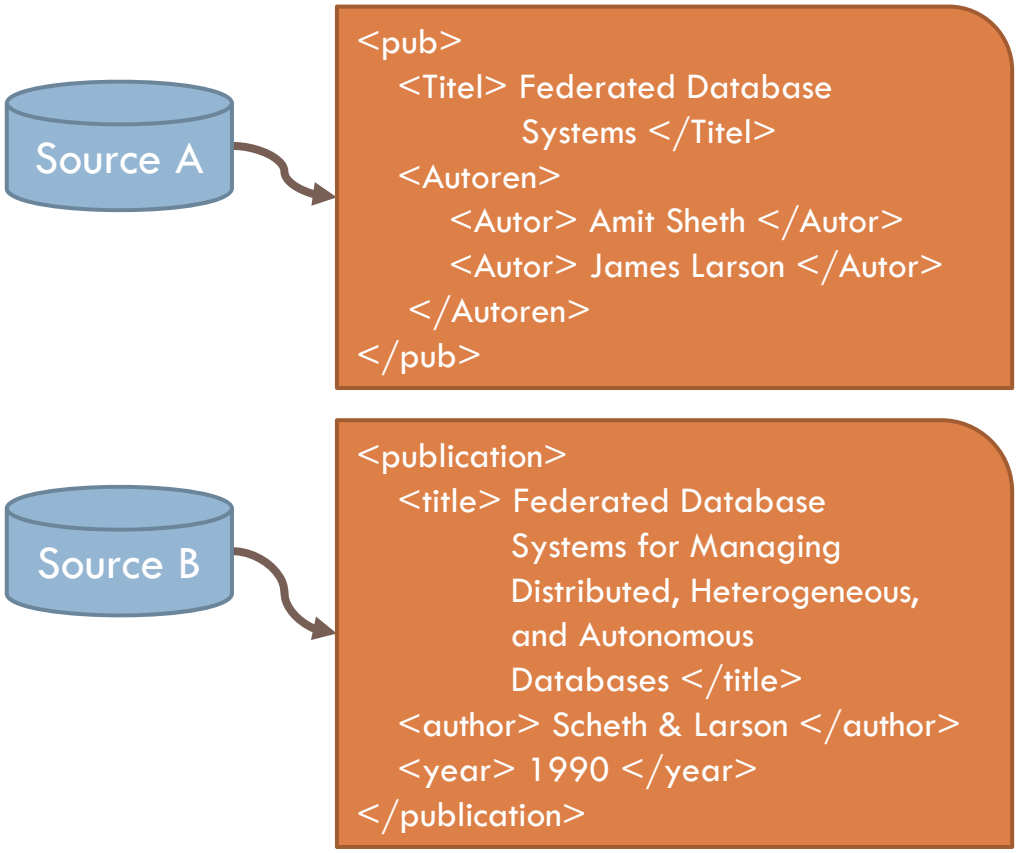
- Error correction
  - Reference tables
    - Cities, countries, products …
    - Similarity measures
  - Standardization and transformation
  - Domain-knowledge (meta data)
    - Conventions (country/region-specific spelling)
    - Ontologies
    - Thesauri, dictionaries for homonyms, synonyms, …
  - Outlier detection and elimination
- And data fusion…

Overview



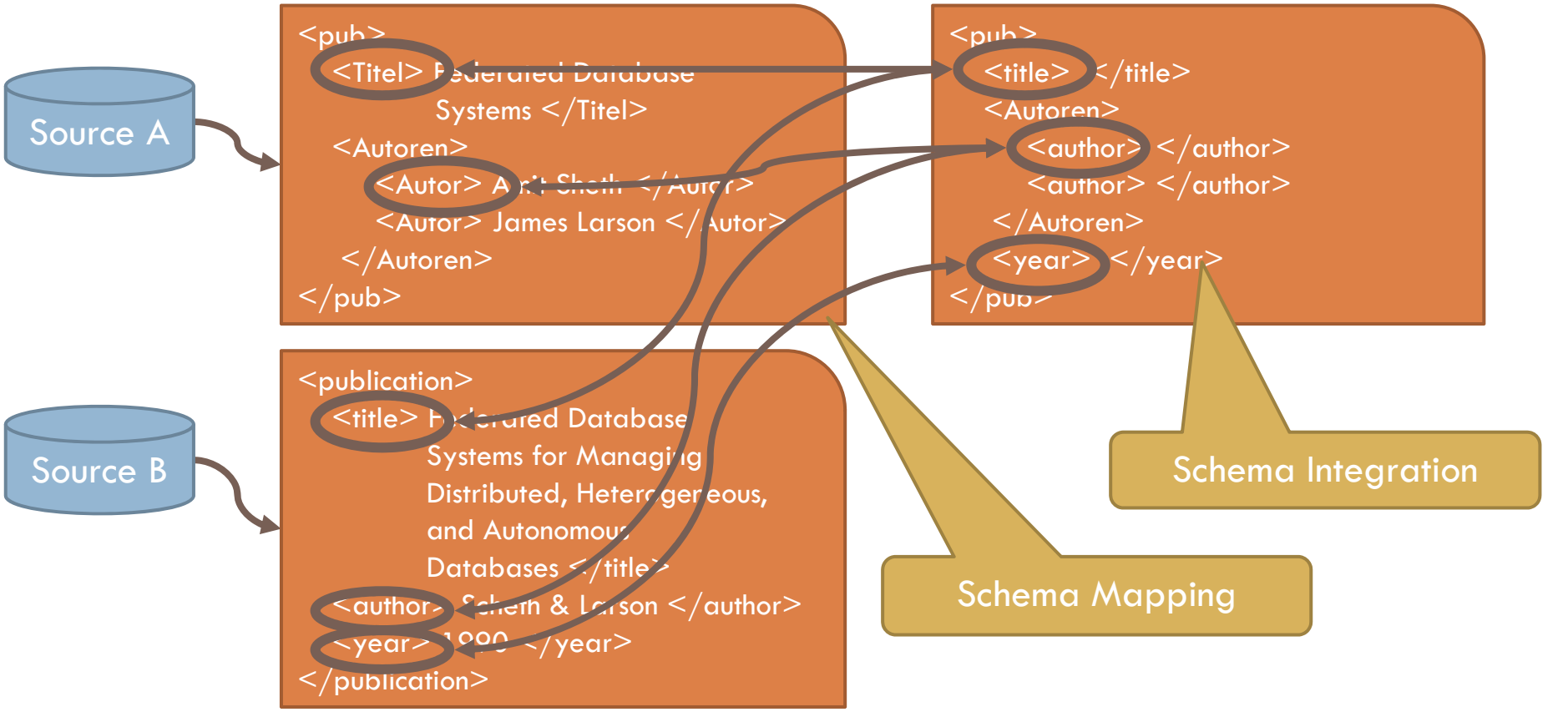
- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems

Information Integration



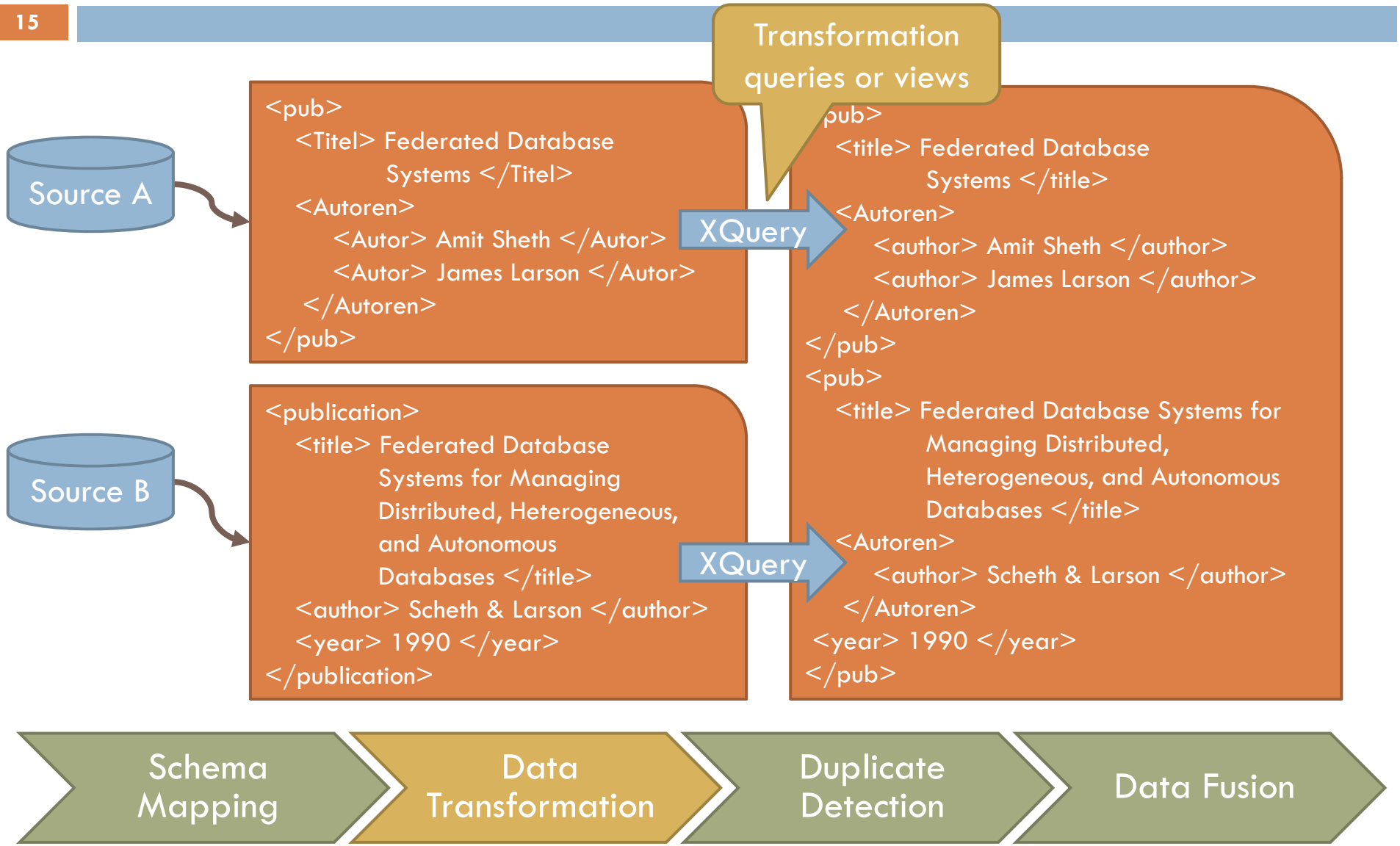
Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

Information Integration



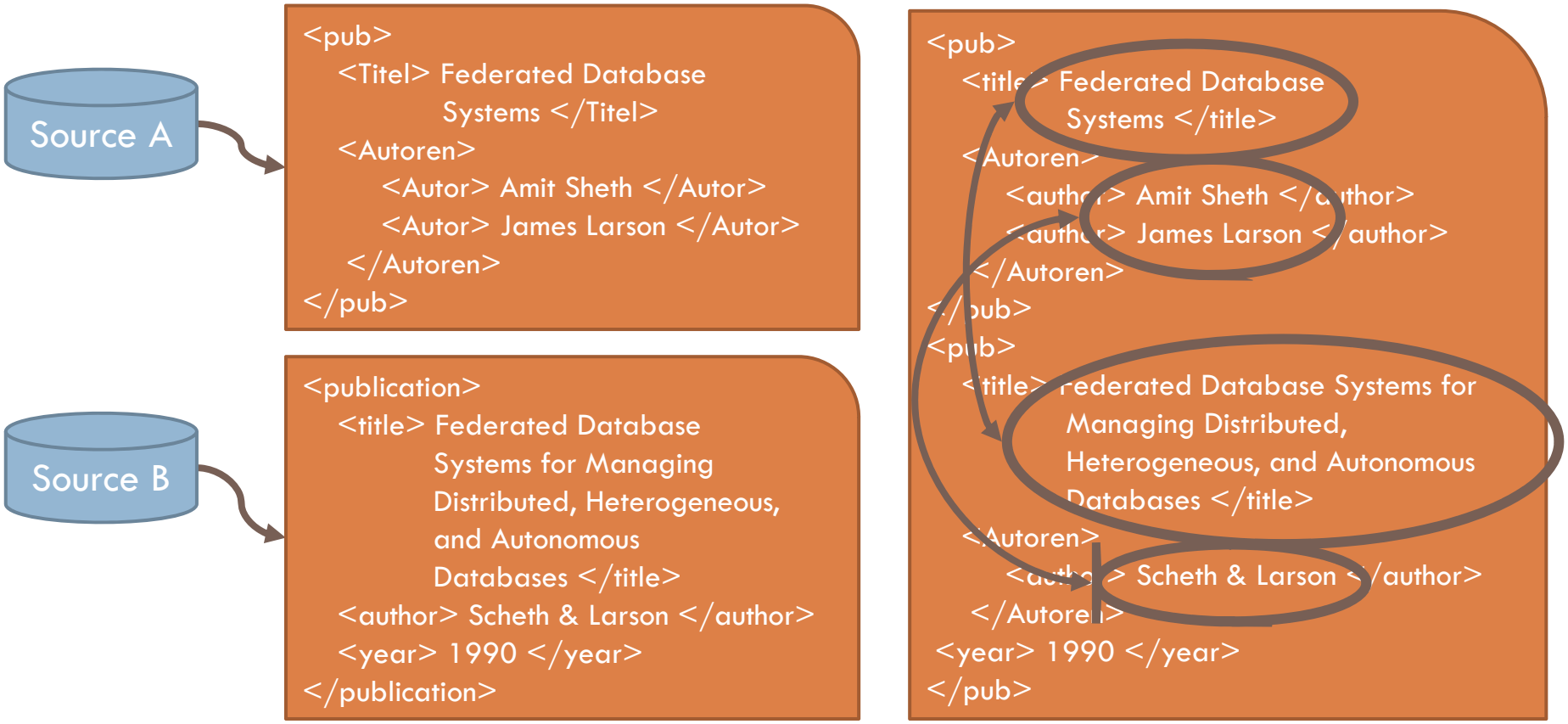
Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

Information Integration

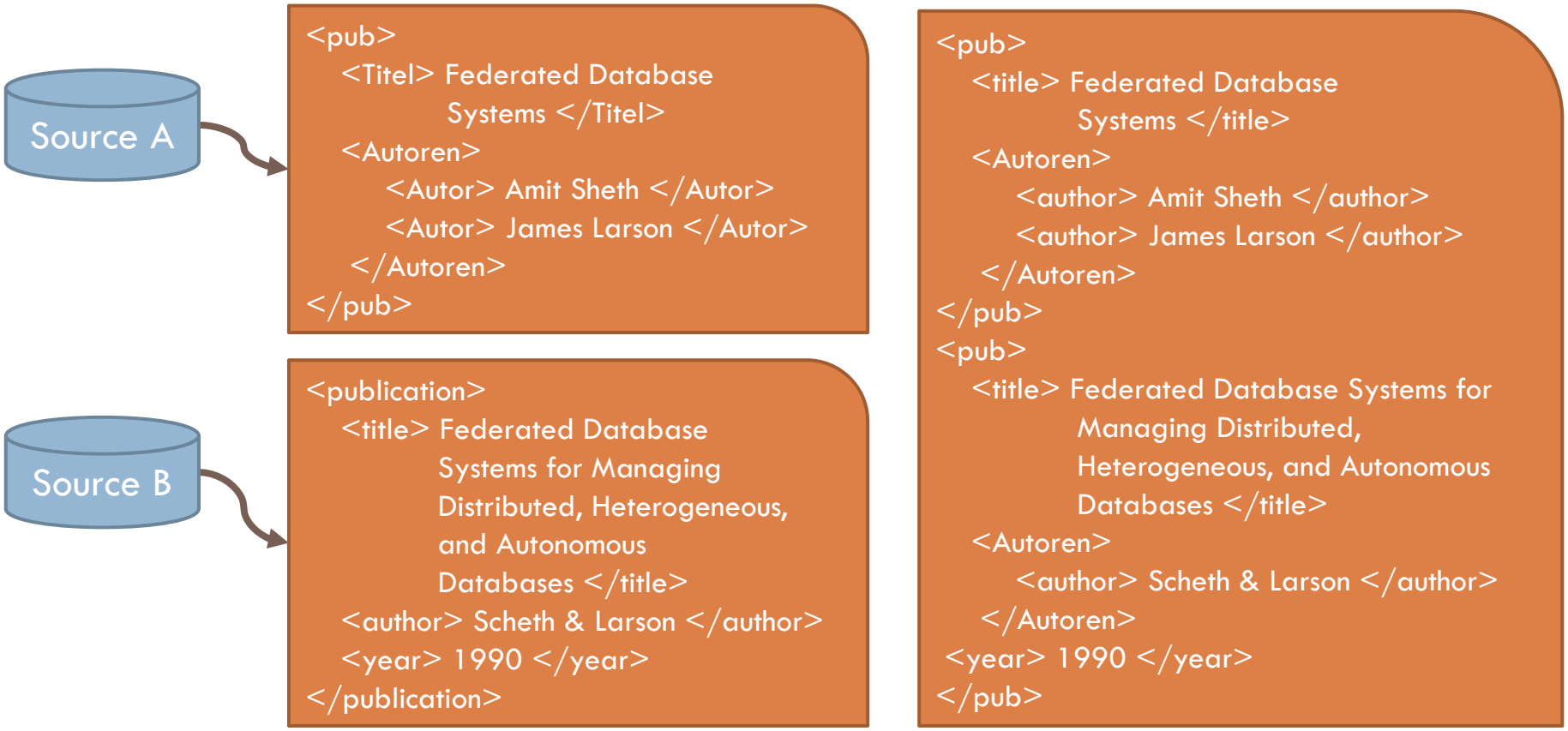


Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

Information Integration

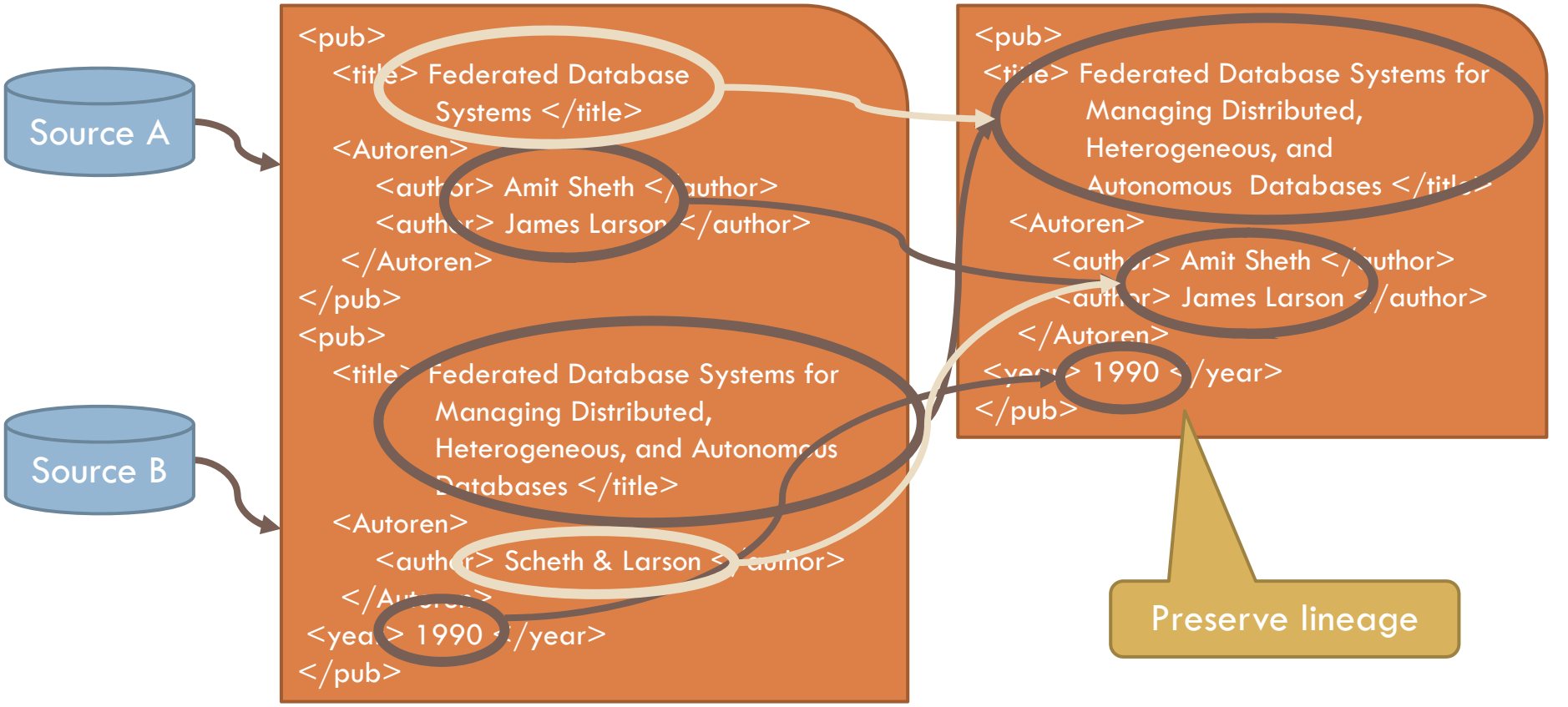


Information Integration



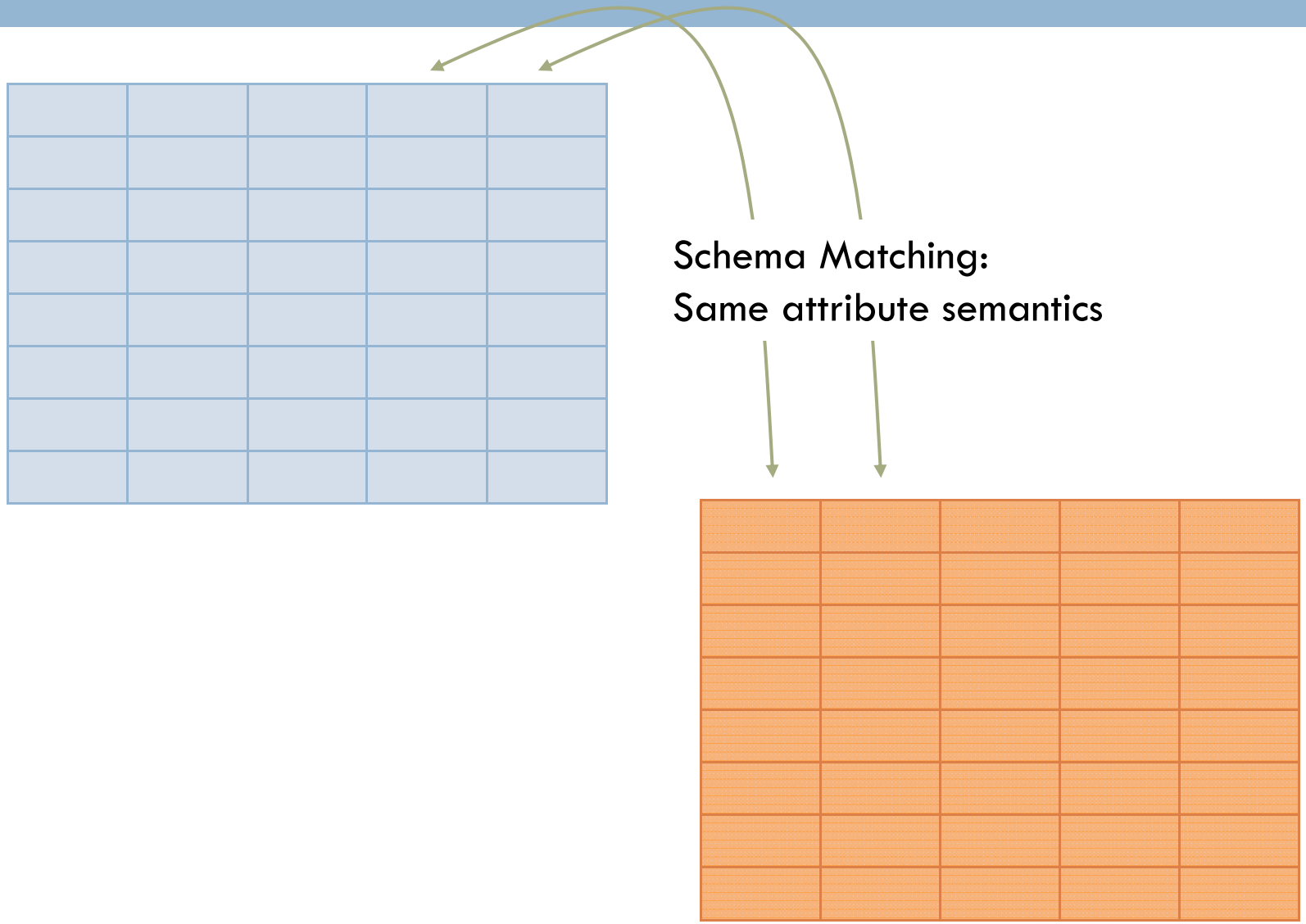
Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

Information Integration

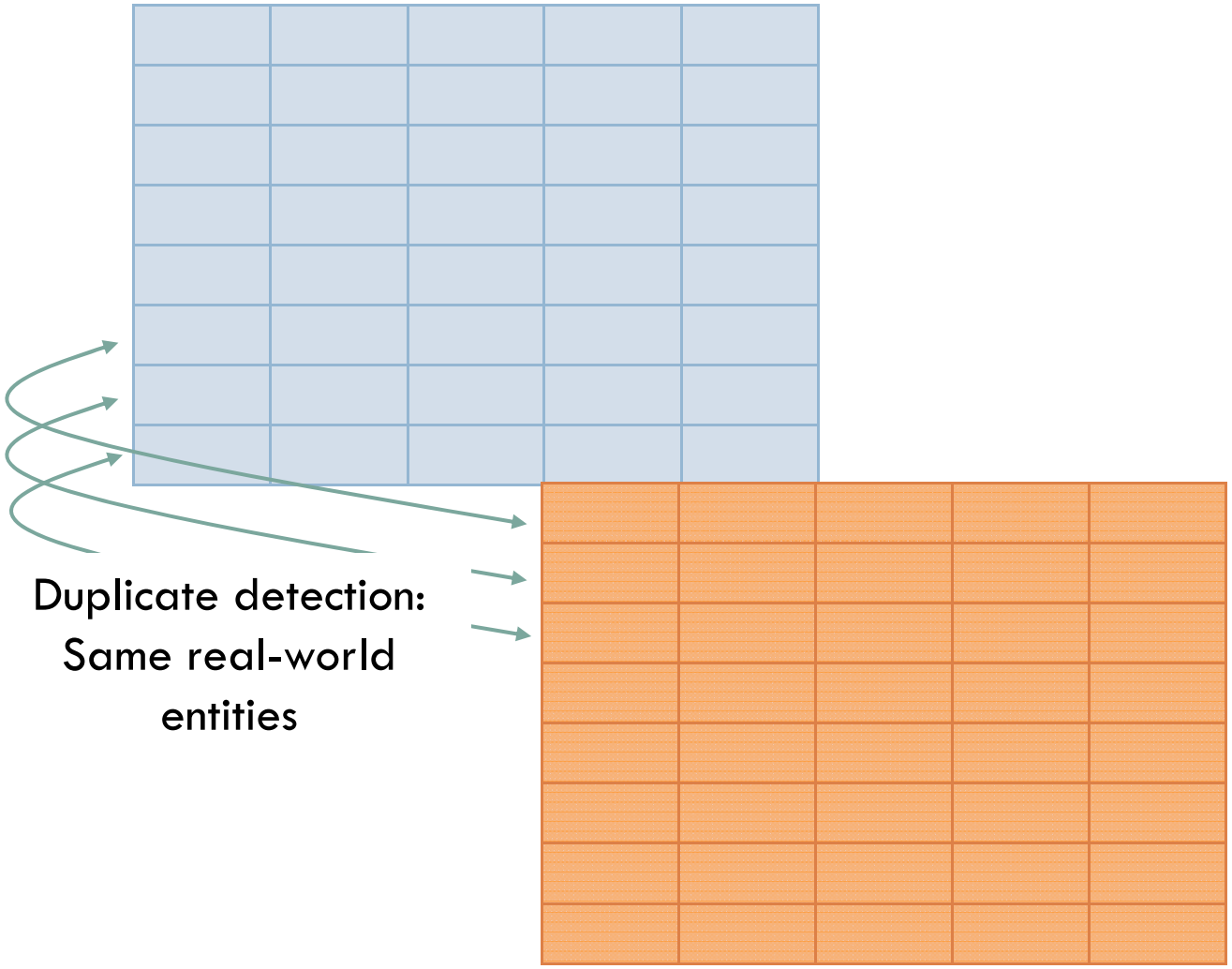


Schema Mapping → Data Transformation → Duplicate Detection → Data Fusion

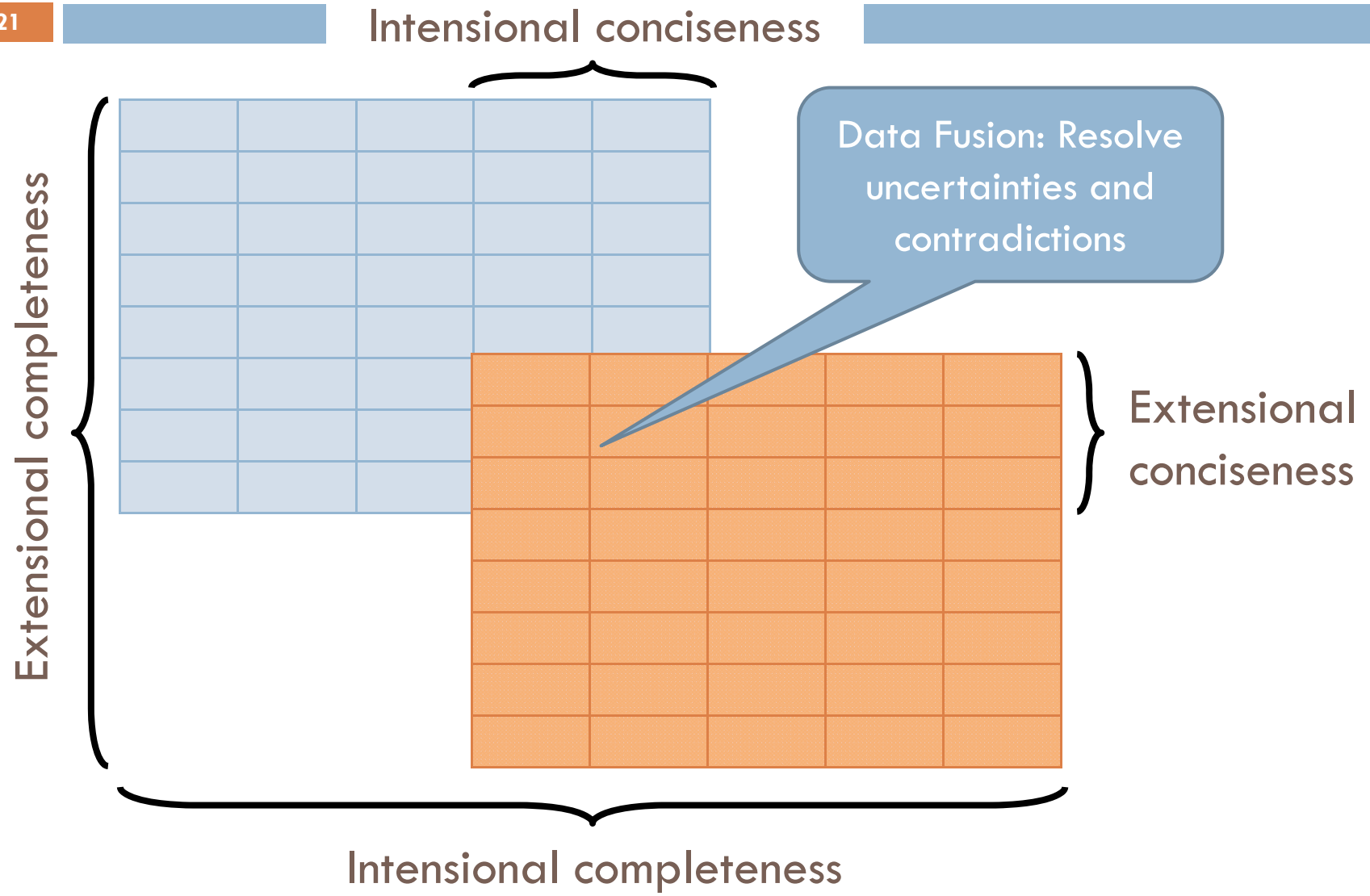
Completeness, Conciseness, and Correctness



Completeness, Conciseness, and Correctness

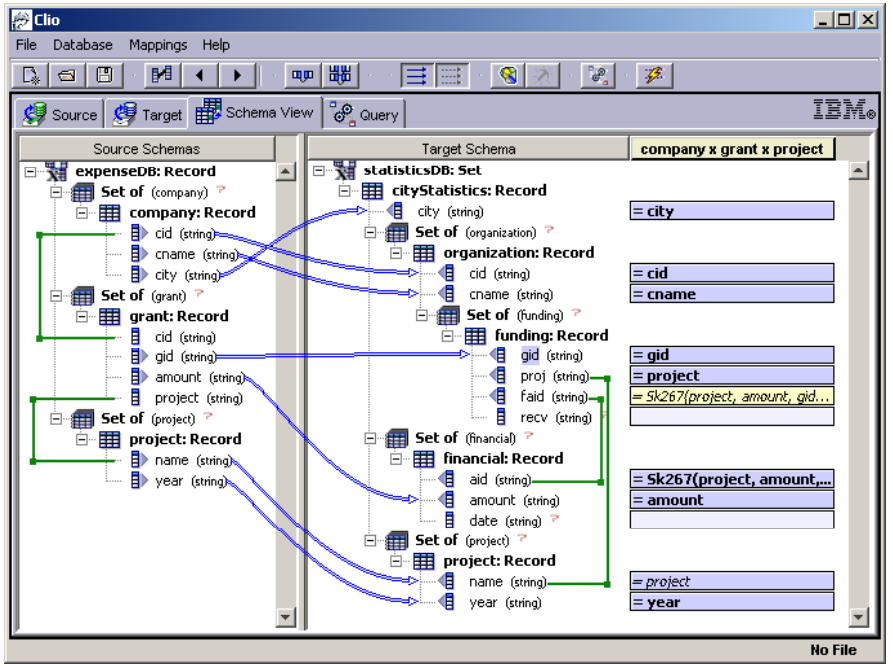


Completeness, Conciseness, and Correctness



Schema Matching

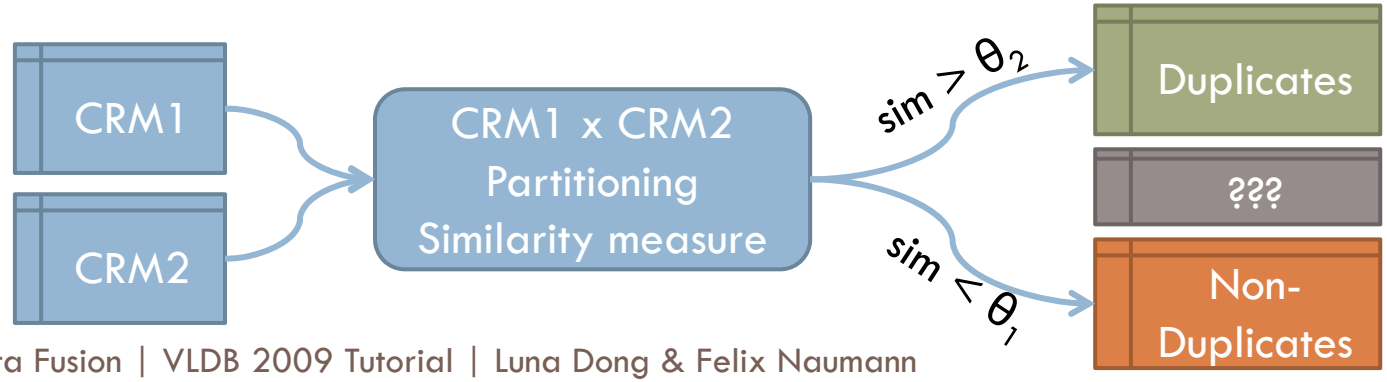
- Problem
  - Given two schemata, find all correspondences between their attributes
- Difficulties
  - Schematic heterogeneity (synonyms & homonyms)
  - Data heterogeneity
  - n:m mappings
  - Transformation functions
  - User interaction



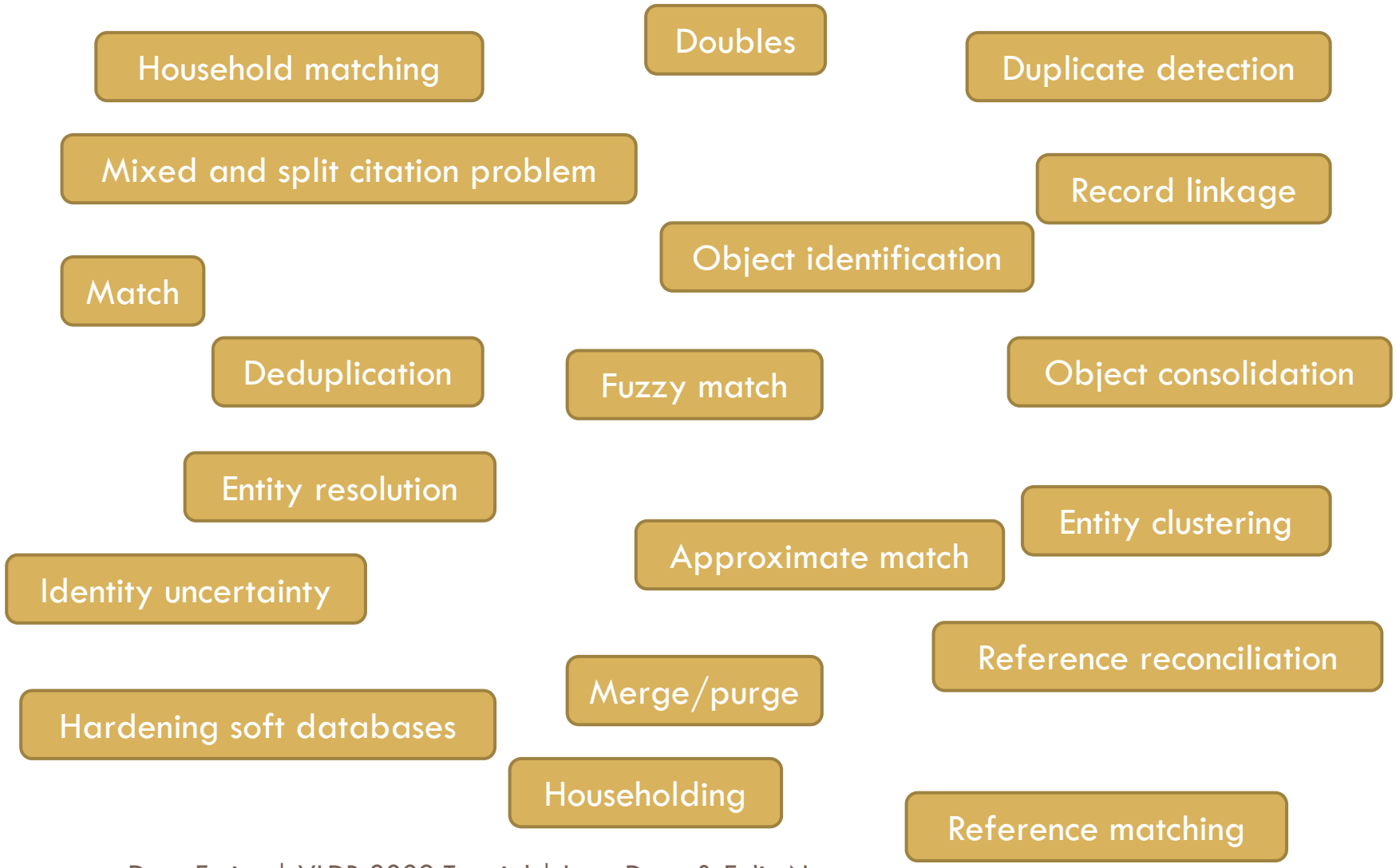
- Then: Derive a schema mapping

Duplicate Detection

- Problem
  - Given one or more data sets, find all sets of objects that represent the same real-world entity.
- Difficulties
  - Duplicates are not identical
    - Similarity measures – Levenshtein, Soundex, Jaccard, etc.
  - Large volume, cannot compare all pairs
    - Partitioning strategies – Sorted neighborhood, Blocking, etc.



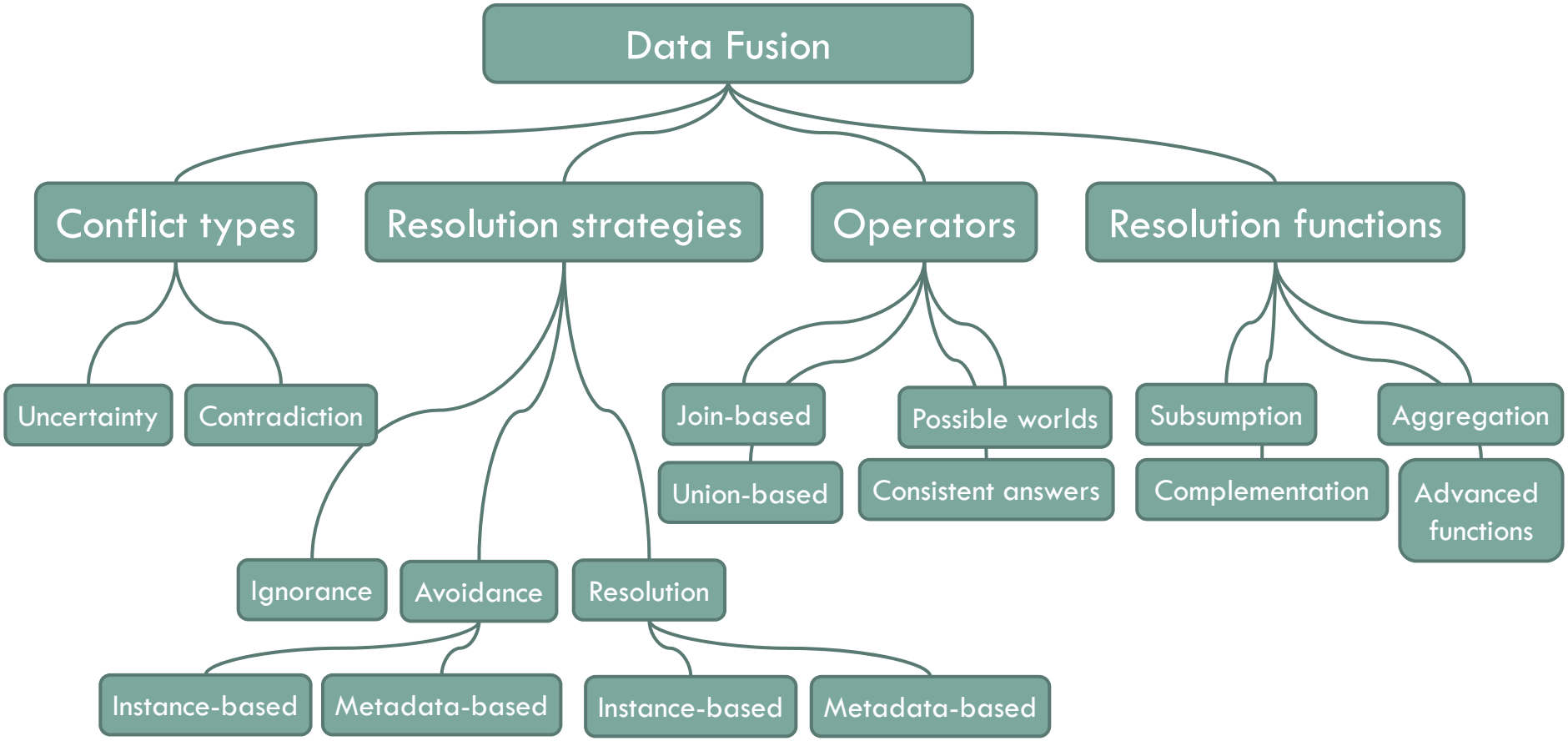
Ironically, "Duplicate Detection" has many Duplicates



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.

The Field of Data Fusion

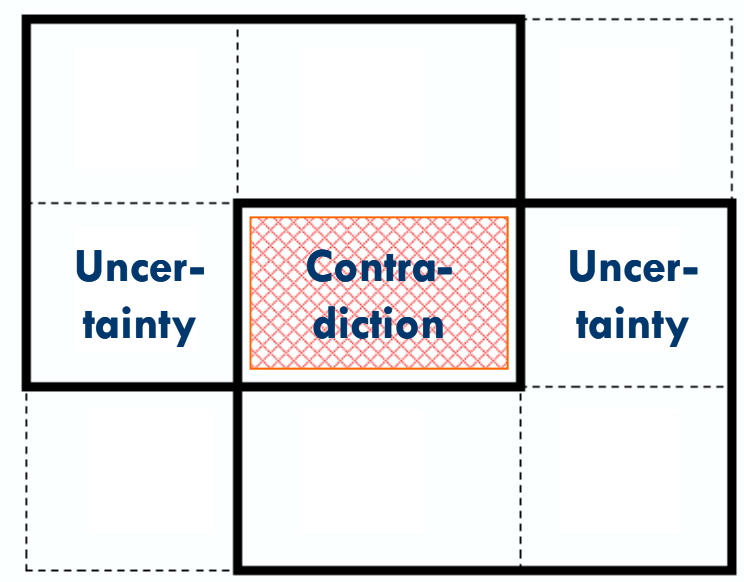


Overview

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems

Uncertainty and Contradiction

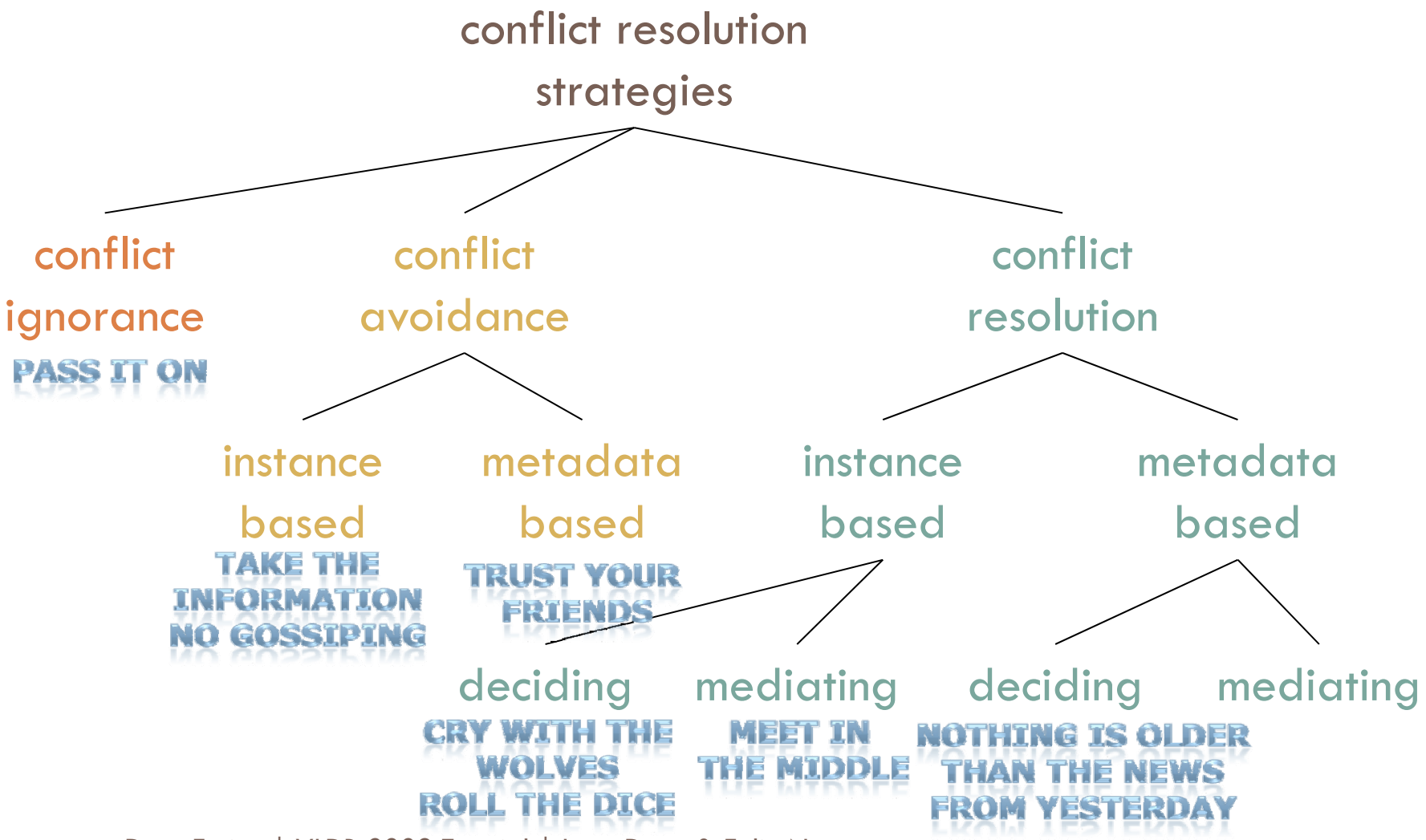
- Uncertainty
  - NULL value vs. non-NULL value
  - "Easy" case
- Contradiction
  - Non-NULL value vs. (different) non-NULL value



Semantics of NULL

- "unknown"
  - There is a value, but I do not know it.
  - E.g.: Unknown date-of-birth
- "not applicable"
  - There is no meaningful value.
  - E.g.: Spouse for singles
- "withheld"
  - There is a value, but we are not authorized to see it.
  - E.g.: Private phone line

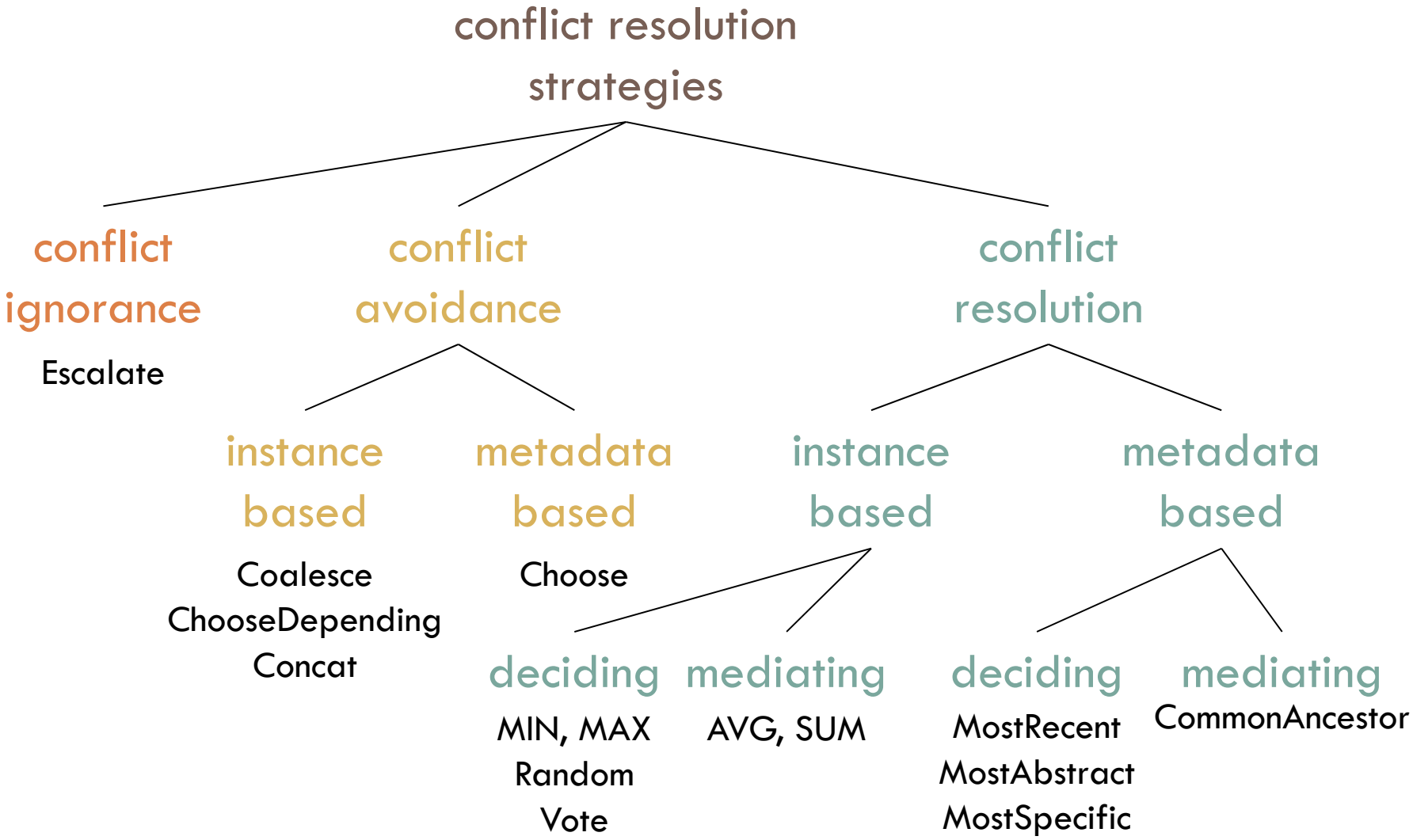
Classification of Strategies



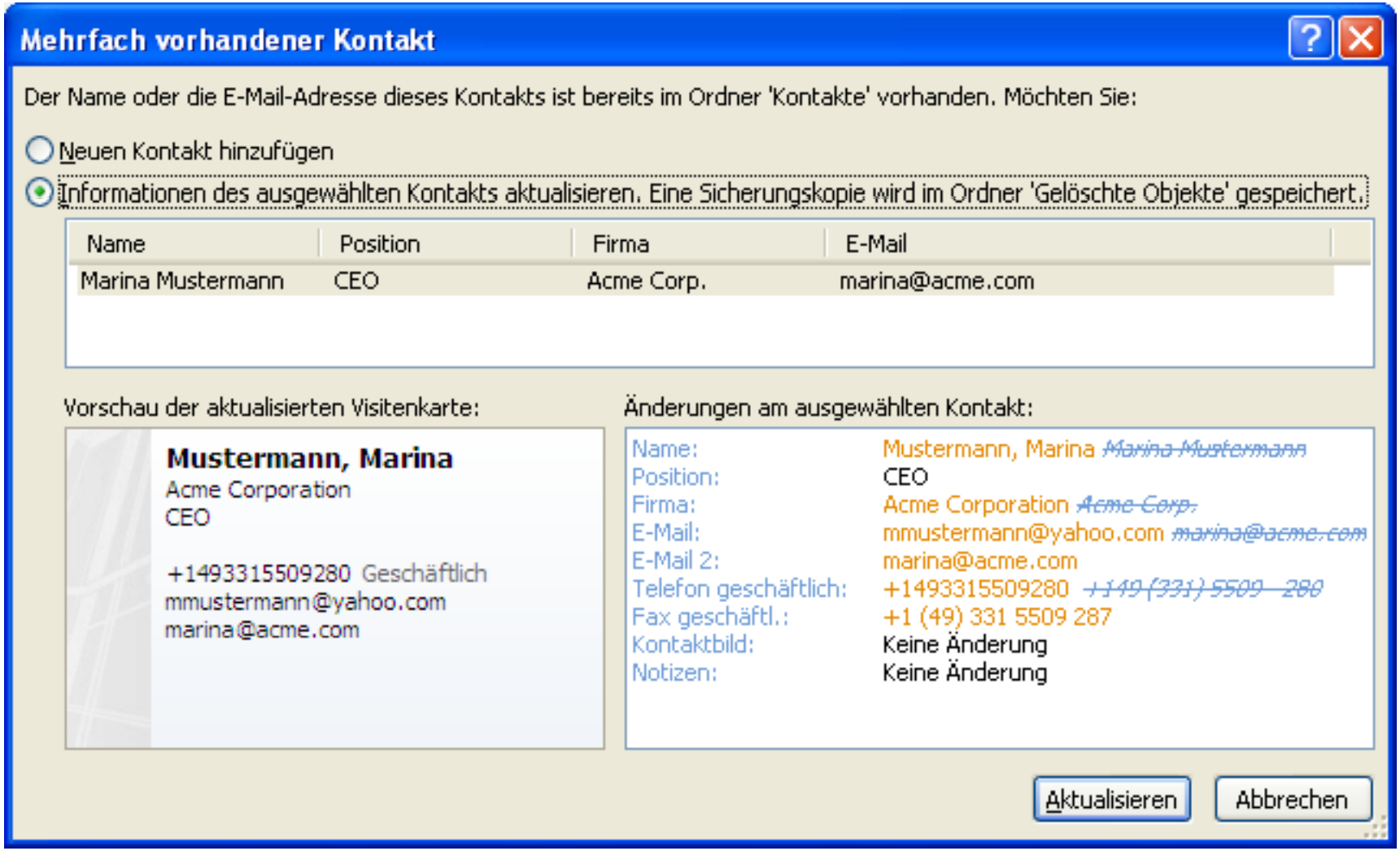
Conflict Resolution Functions

Function	Description	Examples
Min, Max, Sum, Count, Avg	Standard aggregation	NumChildren, Salary, Height
Random	Random choice	Shoe size
Longest, Shortest	Longest/shortest value	First_name
Choose(source)	Value from a particular source	DoB (DMV), CEO (SEC)
ChooseDepending(val, col)	Value depends on value chosen in other column	city & zip, e-mail & employer
Vote	Majority decision	Rating
Coalesce	First non-null value	First_name
Group, Concat	Group or concatenate all values	Book_reviews
MostRecent	Most recent (up-to-date) value	Address
MostAbstract, MostSpecific, CommonAncestor	Use a taxonomy / ontology	Location
Escalate	Export conflicting values	gender
…	…	…

Classification of Functions



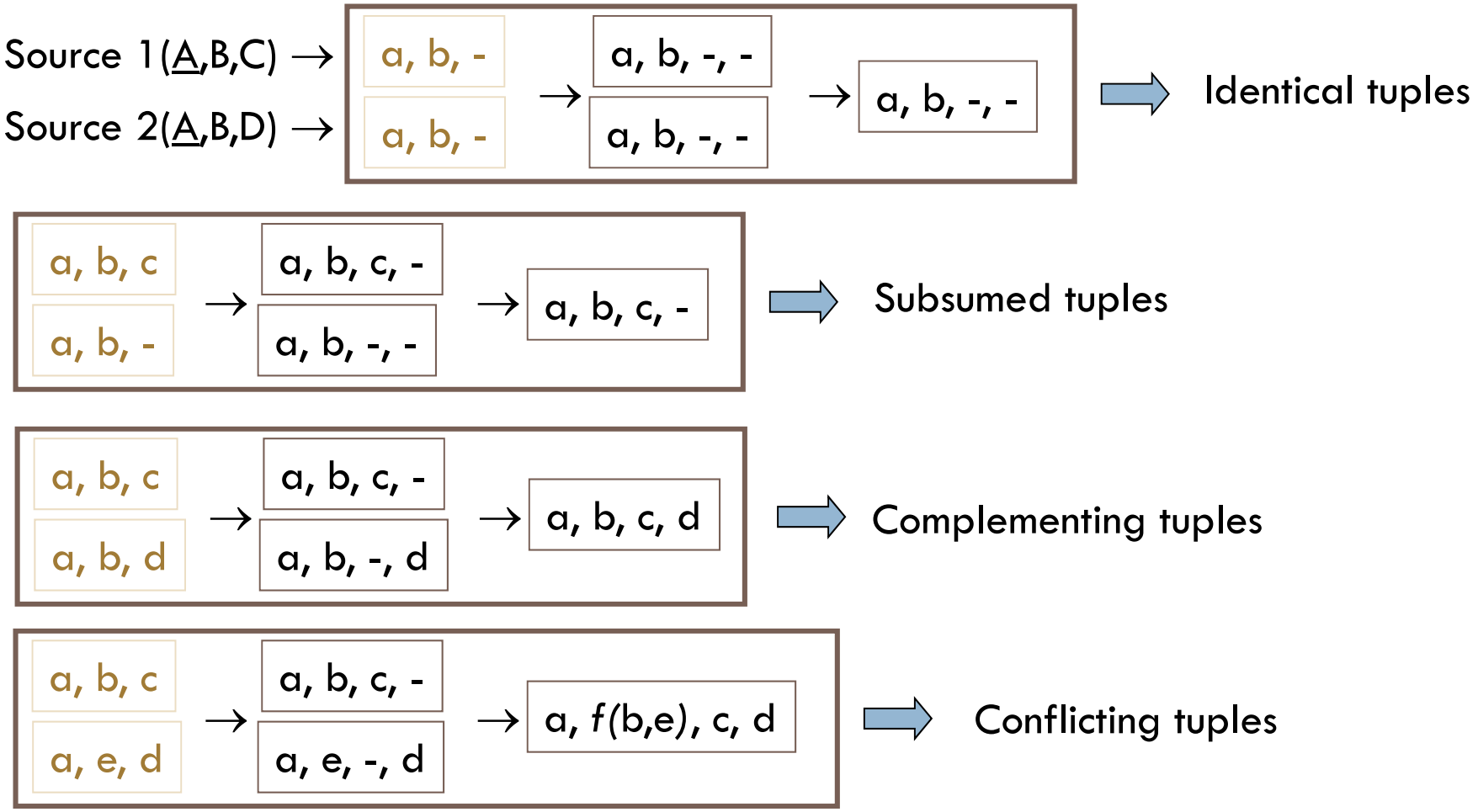
Data Fusion in MS Outlook 2007



Overview

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems

Data Fusion Goals



Relational Operators – Overview

- Identical tuples
  - UNION, OUTER UNION
- Subsumed tuples (uncertainty)
  - MINIMUM UNION
- Complementing tuples (uncertainty)
  - COMPLEMENT UNION, MERGE
- Conflicting tuples (contradiction)
  - Relational approaches: Match, Group, Fuse, …
- Other approaches
  - Possible worlds, probabilistic answers, consistent answers

Minimum Union

- Union: Elimination of exact duplicates
- Minimum Union: Elimination of subsumed tuples
  - Outer union
  - Subsumption
- Rewriting in SQL using DWH extensions (Windows) and assuming existence of favorable ordering [RPZ04]

A	B	C
a	b	c
e	f	g
m	n	o

⊕

A	B	D
a	b	⊥
e	f	h
m	p	⊥

=

A	B	C	D
a	b	c	⊥
a	b	⊥	⊥
e	f	g	⊥
e	f	⊥	h
m	n	o	⊥
m	p	⊥	⊥

A tuple t1 subsumes a tuple t2, if it has same schema, has less NULL-values, and coincides in all non-NULL-values.

R↓

A	B	C	D
a	b	c	⊥
e	f	g	⊥
e	f	⊥	h
m	n	o	⊥
m	p	⊥	⊥

Full Disjunction

- Represents all possible combinations of source tuples
  - Full outer join on all common attributes
    - All combinations for more than two sources
  - Minimum union over results
- Combines complementing tuples (only inter-source)
- Algorithms: [GL94,RU96,CS05]

A	B	C
a	b	c
e	f	g
k	⊥	o
k	m	⊥

$|\bowtie|$

A	B	D
a	b	⊥
e	f	h
m	p	⊥
k	q	r

=

A	B	C	D
a	b	c	⊥
e	f	g	h
m	p	⊥	⊥
k	⊥	o	⊥
k	m	⊥	⊥
k	q	⊥	r

R↓

A	B	C	D
a	b	c	⊥
e	f	g	h
m	p	⊥	⊥
k	⊥	o	⊥
k	m	⊥	⊥
k	q	⊥	r

Complement Union – Proposal

- Elimination of complementing tuples
  - Outer union
  - Complementation
- No known SQL rewriting

A	B	C
a	b	c
e	f	g
m	n	o

⟗

A	B	D
a	b	⊥
e	f	h
m	p	⊥

=

A	B	C	D
a	b	c	⊥
a	b	⊥	⊥
e	f	g	⊥
e	f	⊥	h
m	n	o	⊥
m	p	⊥	⊥

R⇅

A	B	C	D
a	b	c	⊥
e	f	g	h
m	n	o	⊥
m	p	⊥	⊥

A tuple t1 complements a tuple t2, if it has same schema and coincides in all non-NULL-values.

Includes duplicate removal and subsumption

Merge and Prioritized Merge

- Mixes Join and Union to a new operator [GPZ01]
  - Idea: Build two versions for each common attribute, one "favoring" S1, the other "favoring" S2.
  - Nulls in a source are replaced using COALESCE.
  - Fuses complementing tuples, but only for inter-source duplicates
  - Priorization possible: Removes conflicting tuples from right relation.

```
( SELECT R.A, COALESCE(R.B, S.B), R.C, S.D
  FROM R LEFT OUTER JOIN S ON R.A = S.A )
          UNION
( SELECT S.A, COALESCE(S.B, R.B), R.C, S.D
  FROM R RIGHT OUTER JOIN S ON R.A = S.A )
```

Merge and Prioritized Merge

A is real-world ID

A	B	C
a	b	c
e	f	g
m	n	o
m	n	⊥
q	r	s

$|\bowtie$

A	B	D
a	b	⊥
e	f	h
m	p	⊥

$=$

A	B	C	D
a	COAL(b,b)	c	⊥
e	COAL(f,f)	g	h
m	COAL(n,p)	o	⊥
m	COAL(n,p)	⊥	⊥
q	r	s	⊥

$=$

A	B	C	D
a	b	c	⊥
e	f	g	h
m	n	o	⊥
m	n	⊥	⊥
q	r	s	⊥

$\cup$

A	B	C	D
a	b	c	⊥
e	f	g	h
m	p	o	⊥
m	p	⊥	⊥

A	B	C
a	b	c
e	f	g
m	n	o
m	n	⊥
q	r	s

$\bowtie|$

A	B	D
a	b	⊥
e	f	h
m	p	⊥

$=$

A	B	C	D
a	COAL(b,b)	c	⊥
e	COAL(f,f)	g	h
m	COAL(p,n)	o	⊥
m	COAL(p,n)	⊥	⊥

$=$

A	B	C	D
~~a~~	~~b~~	~~c~~	~~⊥~~
~~e~~	~~f~~	~~g~~	~~h~~
m	p	o	⊥
m	p	⊥	⊥

Match Join

- Context: AURORA Project [YÖ99]
- Handles columns individually using projections (with IDs)
- Performs UNION on each column across all sources
- Reassembles using FULL OUTER JOINS
- Uses "conflict-tolerant query model" to query these possible worlds.

```
WITH OU(A,B,C,D) AS (
  ( SELECT A, B, C, NULL AS D FROM U1 )
    UNION
  ( SELECT  A, B, NULL AS C, D FROM U2 ) ),
  B_V (A,B) AS ( SELECT DISTINCT  A, B FROM OU ),
  C_V (A,C) AS ( SELECT DISTINCT A, C FROM OU ),
  D_V (A,D) AS ( SELECT DISTINCT A, D FROM OU ),
SELECT A, B, C, D
FROM B_V FULL OUTER JOIN C_V FULL OUTER JOIN D_V
ON B_V.A=C_V.A AND C_V.A=D_V.A
```

Match Join

- Conflict-tolerant query model
  - Chooses tuples from result of MatchJoin
- Three semantics
  - HighConfidence, RandomEvidence, PossibleAtAll
- Resolution functions
  - SUM, AVG, MAX, MIN, ANY, DISCARD

```
SELECT  ID, Name[ANY], Age[MAX]
FROM    MatchJoin(U1,U2)
WHERE   Age>22
WITH    PossibleAtAll
```

Grouping and Aggregation

- Outer union then group by real-world ID
- Aggregate all other columns using conflict resolving aggregate function
- Efficient implementations
- Catches inter- and intra-source duplicates
- Restricted to built-in aggregate-functions
  - MAX, MIN, AVG, VAR, STDDEV, SUM, COUNT

```
WITH OU AS (
  ( SELECT A, B, C, NULL AS D FROM U1 )
      UNION (ALL)
  ( SELECT A, B, NULL AS C, D FROM U2 ) ),

SELECT   A, MAX(B),  MIN(C), SUM(D)
FROM     OU
GROUP BY A
```

FUSE BY

- SQL extensions to resolve uncertainties and contradictions [BN05,BBB+05]
- FUSE FROM implies OUTER UNION
  - Removes subsumed and duplicate tuples by default
- FUSE BY declares real-world ID
- RESOLVE specifies conflict resolution function from catalog
  - Default: COALESCE
- Implemented on top of relational DBMS "XXL"

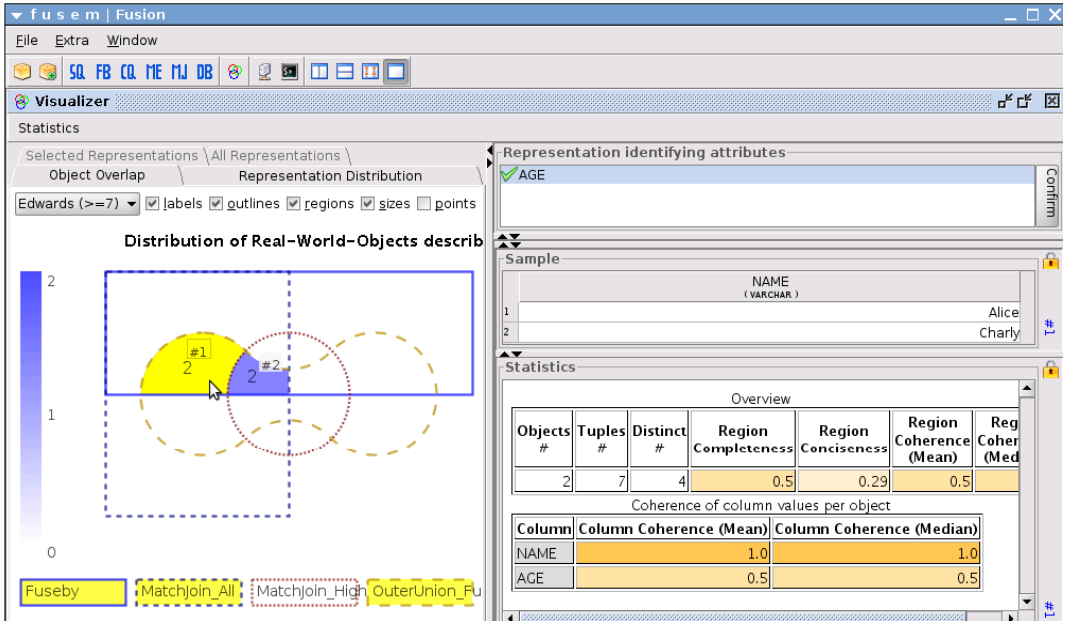
```
SELECT  ID,
        RESOLVE(Title, Choose(IMDB)),
        RESOLVE(Year, Max), RESOLVE(Director),
        RESOLVE(Rating), RESOLVE(Genre, Concat)
FUSE FROM IMDB, Filmdienst
FUSE BY (ID)
ON ORDER Year DESC
```

Summary of Operators

	Duplicates	Subsumed tuples	Complementing tuples	Contradictions
Union, Outer Union	✓	✗	✗	✗
Minimum Union	✓	✓	✗	✗
Full Disjunction	✓	✓	✓ (inter-source)	
Complement Union	✓	✓	✓	✗
Merge	✓	✓ (inter-source)	✓ (inter-source)	✗
MatchJoin + CTQM	✓	✓	✓	✗ ✓
Group By	✓	✓	✓	✓
Fuse By	✓	✓	✓	✓

FuSem

- Tool to query and fuse data from diverse data sources [BDN07]
  - Based on HumMer project [BBB+05].
  - http://www.hpi.uni-potsdam.de/naumann/sites/fusem/
- Explore data and find interesting subsets
- Execute, explore and compare five different data fusion semantics, specified in their respective syntax:
  - SQL (and extensions, such as Subsumption)
  - Merge
  - MatchJoin
  - FuseBy
  - ConQuer



What else is there?

- Consistent Query Answering
  - Avoid conflicts and report only certain tuples
    - Those that appear in every repair [FFM05]
- "Possible worlds" models
  - Build all possible solutions, annotated with likelihood
    - Yes/No/Maybe [DeM89]
    - Probability value [LSS94]
  - Probabilistic databases [SD05]
    - Extend algebra to produce probabilities
    - Extend query language to query and export probabilities

Overview

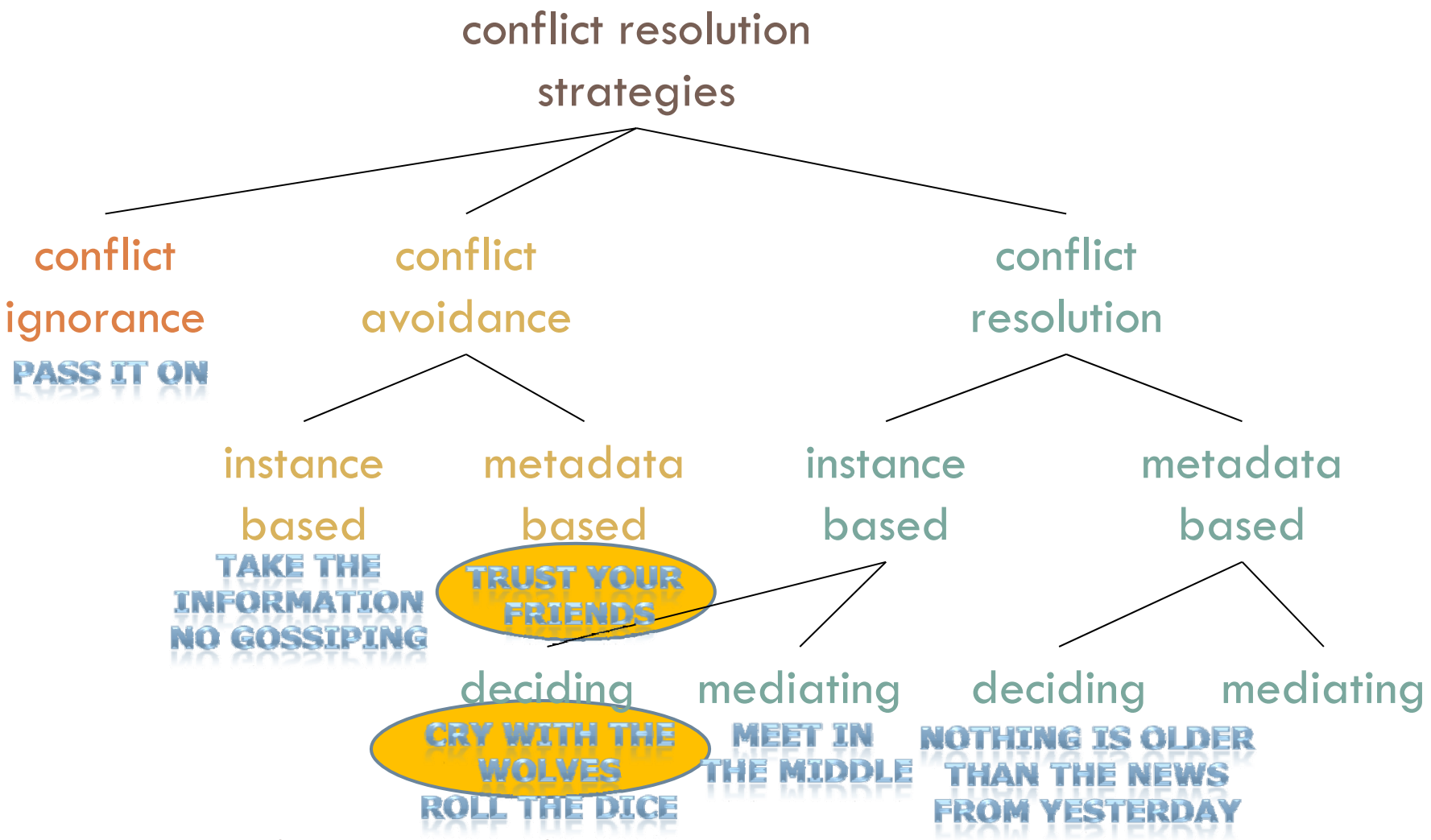
- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators

- Advanced truth-discovery techniques
- Existing data fusion systems
- Open problems

Outline

- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Data fusion in existing integration systems
- Open problems

Basic Strategies



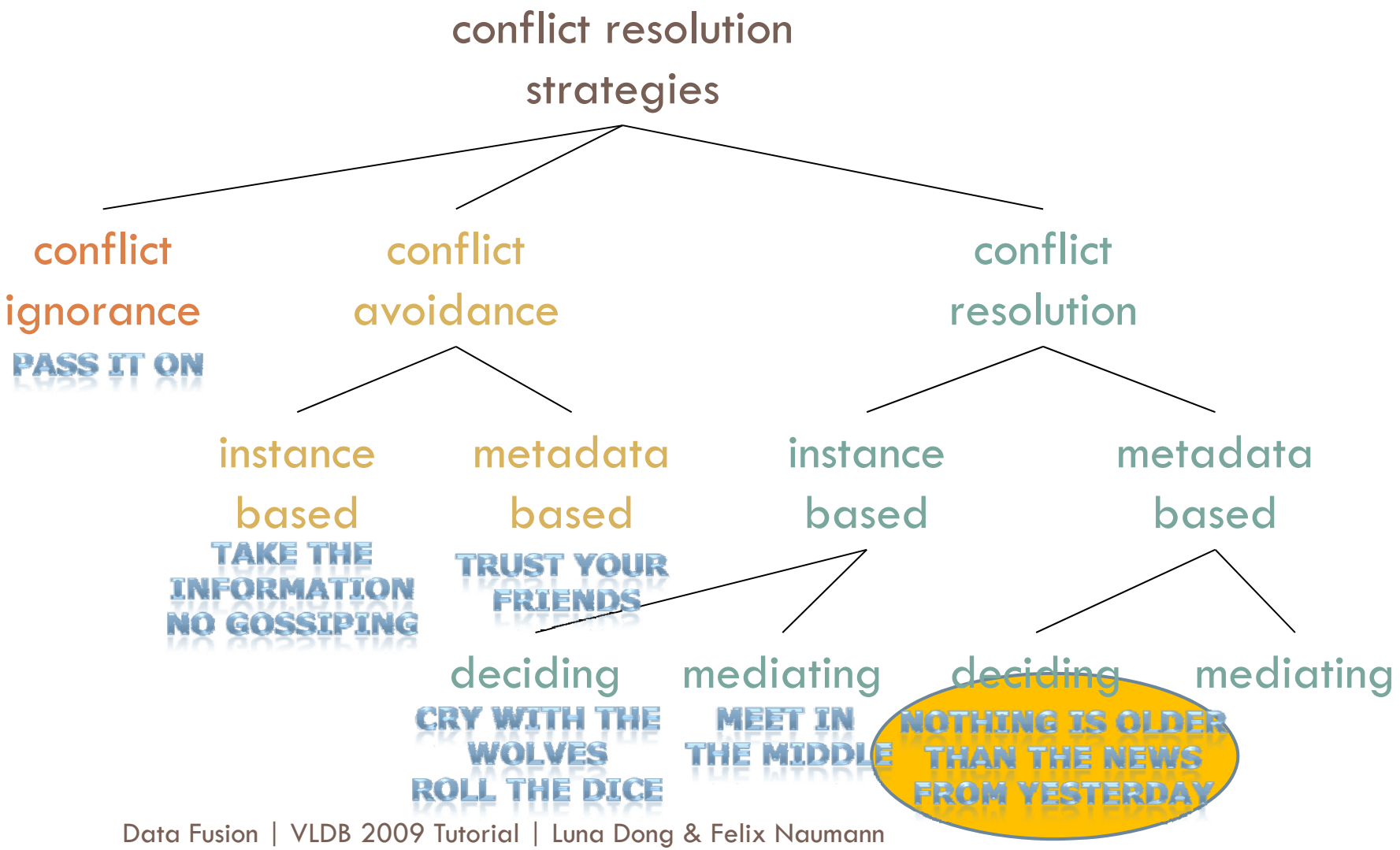
Intuitions

- Data sources are of different quality and we trust data from accurate sources more

Intuitions

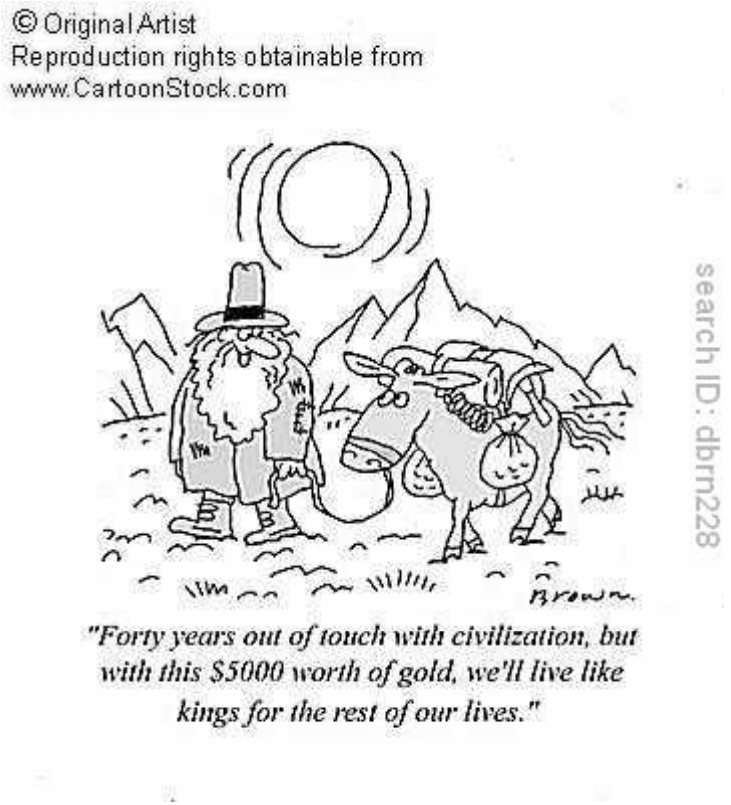
- Data sources are of different quality and we trust data from accurate sources more

Basic Strategies



Intuitions

- Data sources are of different quality and we trust data from accurate sources more
- The real world is dynamic and the true value often evolves over time
  - E.g., person affiliation, business contact phone

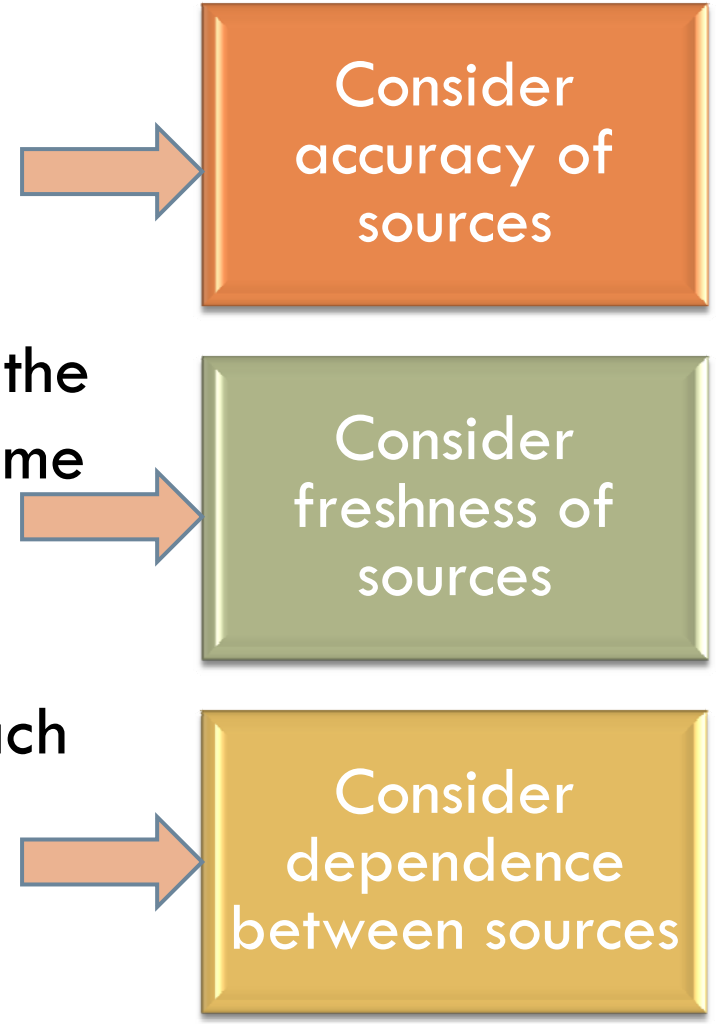


Intuitions

- Data sources are of different quality and we trust data from accurate sources more
- The real world is dynamic and the true value often evolves over time
  - E.g., person affiliation, business contact phone
- Data sources can copy from each other and errors can be propagated quickly

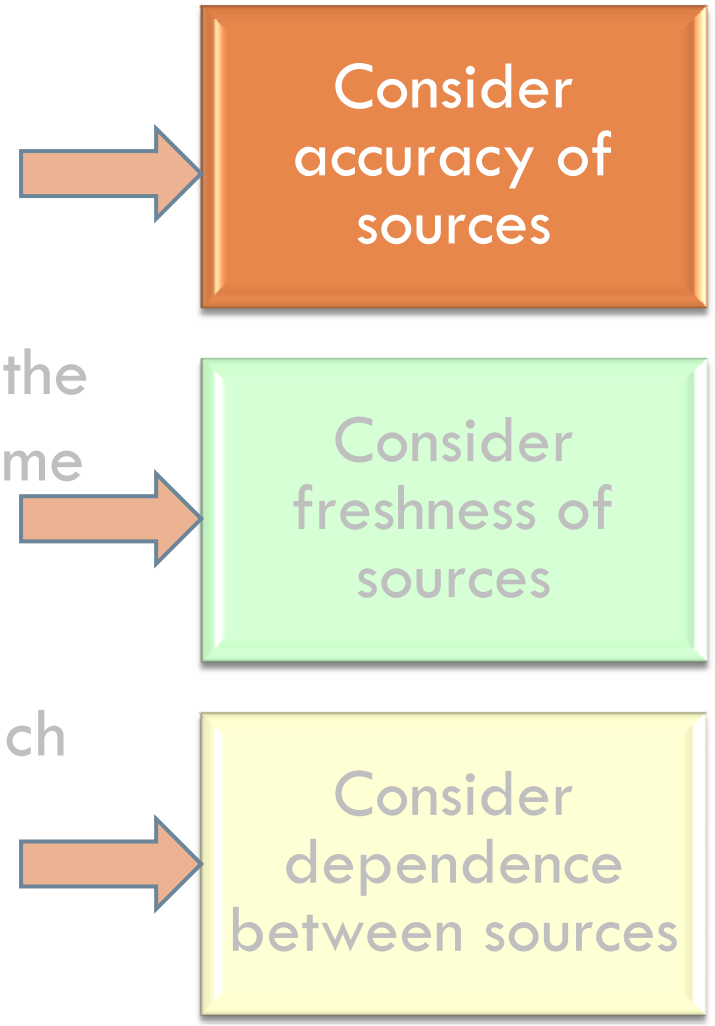
Advanced Truth-Discovery Techniques

- Data sources are of different quality and we trust data from accurate sources more
- The real world is dynamic and the true value often evolves over time
  - E.g., person affiliation, business contact phone
- Data sources can copy from each other and errors can be propagated quickly



Advanced Truth-Discovery Techniques

- Data sources are of different quality and we trust data from accurate sources more
- The real world is dynamic and the true value often evolves over time
  - E.g., person affiliation, business contact phone
- Data sources can copy from each other and errors can be propagated quickly



Trust Accurate Sources

- Considering accuracy can often improve truth discovery

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

Trust Accurate Sources

- Considering accuracy can often improve truth discovery

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

S1 is more accurate; trusting it more can help find the correct affiliation for Carey

Trust Accurate Sources

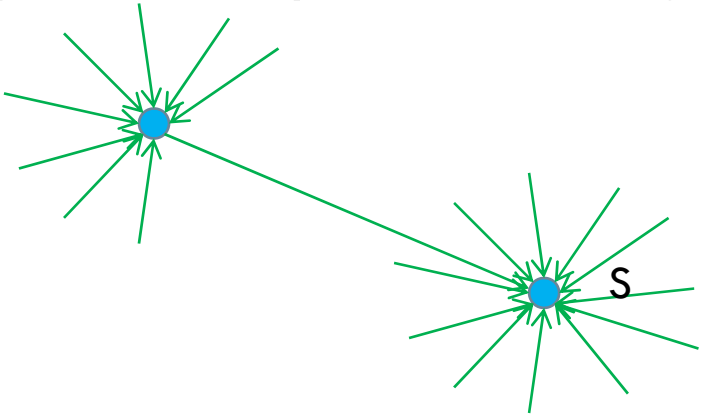
- Considering accuracy can often improve truth discovery

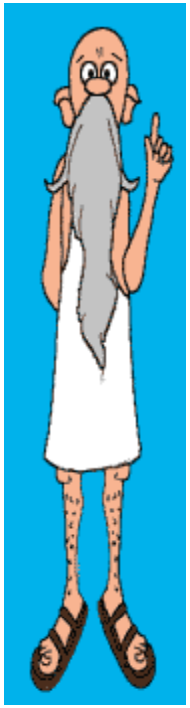
	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

S1 is more accurate; trusting it more can help find the correct affiliation for Carey

Find Trustable Sources (I)

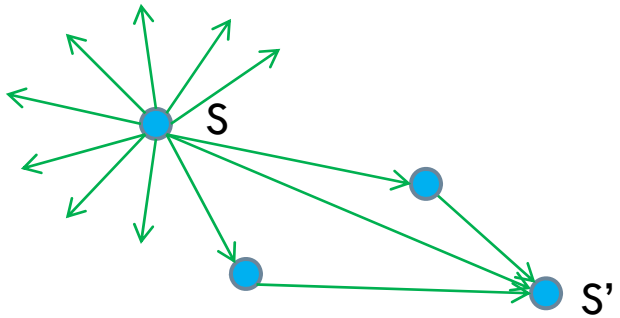
- Deciding authority based on link analysis and source popularity
  - Survey: "Link analysis ranking: algorithms, theory, and experiments" [Borodin et al., 05]
  - PageRank [Brin and Page, 98]
  - Authority-hub analysis [Kleinberg, 98]





Find Trustable Sources (II)

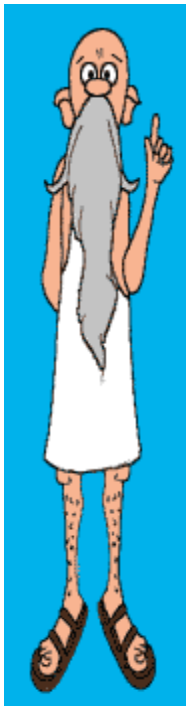
- Assign a global trust rating to each data source based on its behavior in a P2P network
  - TrustMe [Singh and Liu, 03]
  - EigenTrust [Kamvar et al., 03]
    - Peer i&j:



$$s_{ij} = sat(i, j) - unsat(i, j)$$

$$c_{ij} = \frac{\max(s_{ij}, 0)}{\sum_j \max(S_{ij}, 0)}$$

$$t_{ij} = \sum_k c_{ik} c_{kj}$$



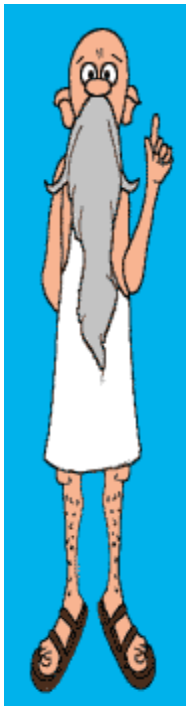
Find Trustable Sources (III)

- Compute accuracy of sources
  - Corroborating answers from web sources [Wu and Marian, 07]
  - TruthFinder [Yin et al., 07]
  - Solomon [Dong et al., 09a]

$$A(S) = \underset{v \in \overline{V}(S)}{Avg}\ P(v)$$

$\overline{V}(S)$ -values provided by S; P(v)-pr of value v being true

How to compute P(v)?



Apply Source Accuracy in Truth Discovery

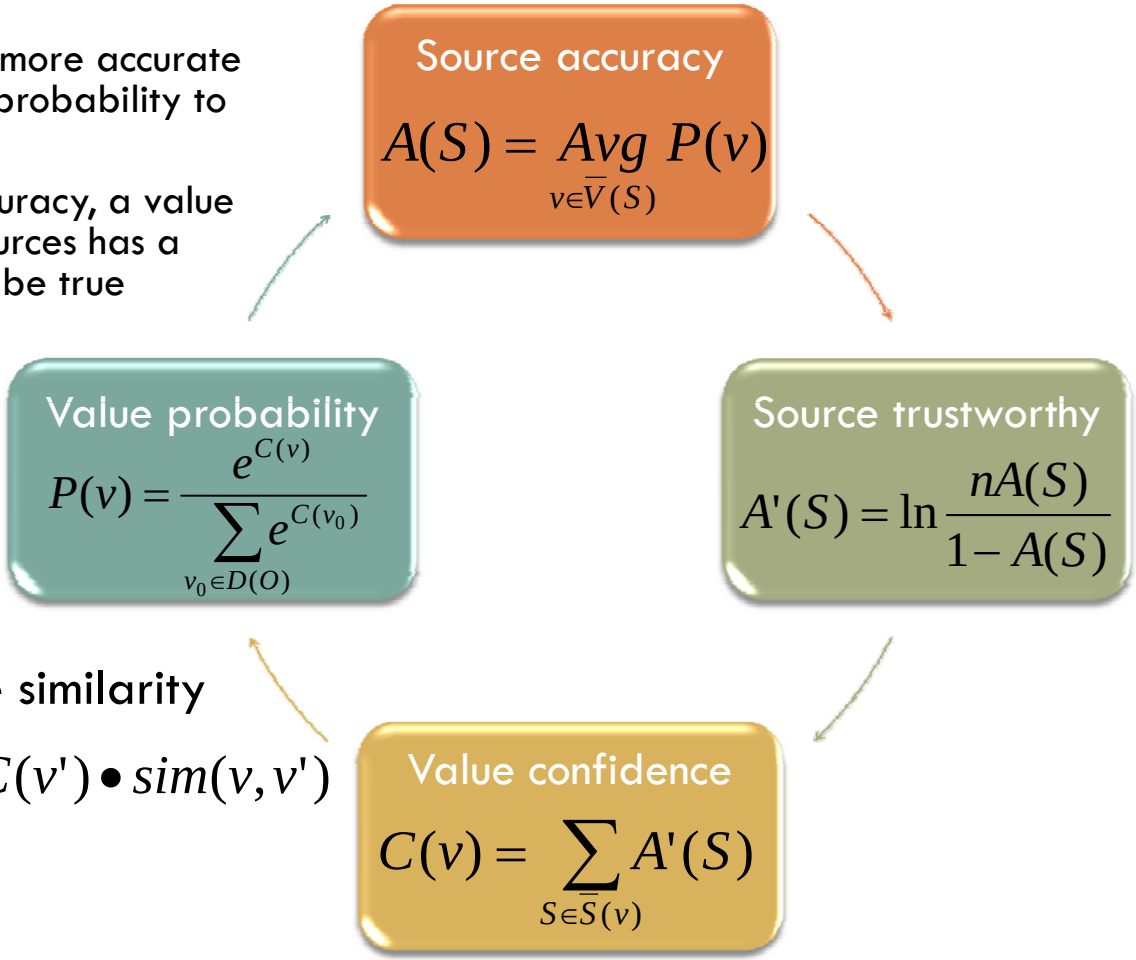
[Yin et al., 07] [Dong et al., 09a]

- Input:
  - Object O
  - Dom(O)={v_0,v_1,…,v_n}
  - Observation Φ on O
- Output: Pr(v_i true | Φ) for each i=0,…, n (sum up to 1)
- According to the Bayes Rule, we need to know Pr(Φ | v_i true)
  - Assuming independence of sources, we need to know Pr(Φ(S) | v_i true)
  - If S provides v_i : Pr(Φ(S) | v_i true) =A(S)
  - If S does not provide v_i : Pr(Φ(S) | v_i true) =(1-A(S))/n

Model and Algorithm [Dong et al., 09a]

Properties

- A value provided by more accurate sources has a higher probability to be true
- Assuming uniform accuracy, a value provided by more sources has a higher probability to be true



- Consider value similarity

$$C^*(v) = C(v) + \rho \sum_{v' \neq v} C(v') \bullet sim(v, v')$$

- Continue until source accuracy converges

An Example

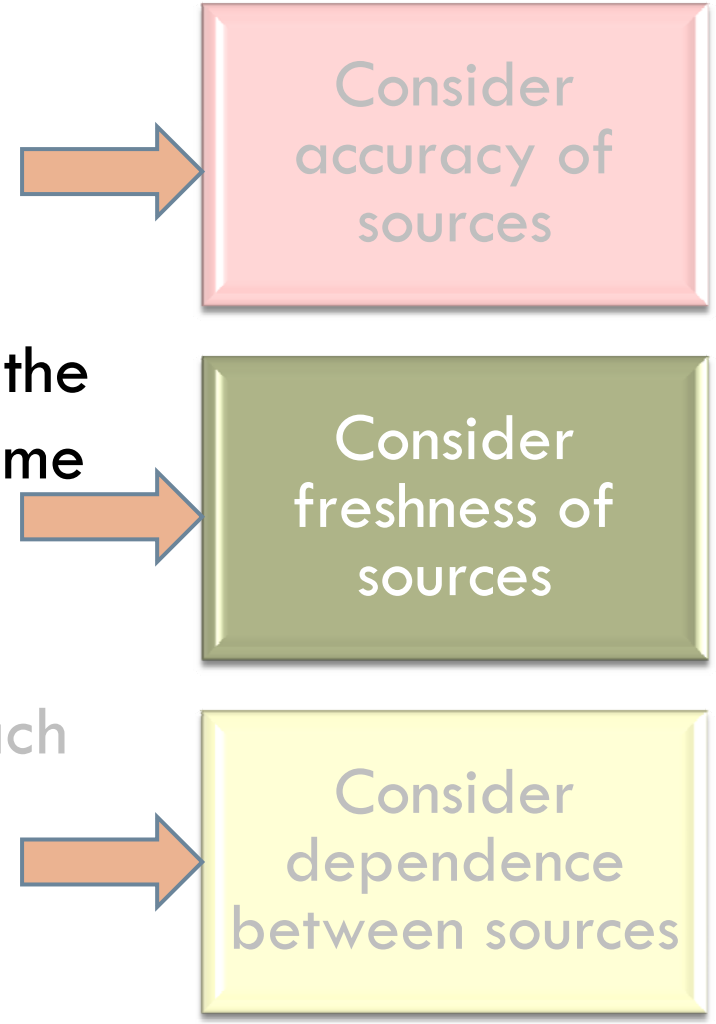
	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

Accuracy	S1	S2	S3
Round 1	.69	.57	.45
Round 2	.81	.63	.41
Round 3	.87	.65	.40
Round 4	.90	.64	.39
Round 5	.93	.63	.40
Round 6	.95	.62	.40
Round 7	.96	.62	.40
Round 8	.97	.61	.40

Value Confidence	Carey		
	UCI	AT&T	BEA
Round 1	1.61	1.61	1.61
Round 2	2.40	1.89	1.42
Round 3	3.05	2.16	1.26
Round 4	3.51	2.23	1.19
Round 5	3.86	2.20	1.18
Round 6	4.17	2.15	1.19
Round 7	4.47	2.11	1.20
Round 8	4.76	2.09	1.20

Advanced Truth-Discovery Techniques

- ☑ Data sources are of different quality and we trust data from accurate sources more
- ☐ The real world is dynamic and the true value often evolves over time
  - E.g., person affiliation, business contact phone
- ☐ Data sources can copy from each other and errors can be propagated quickly



A Dynamic World

- True values can evolve over time
  - A subtle third case: *out-of-date*

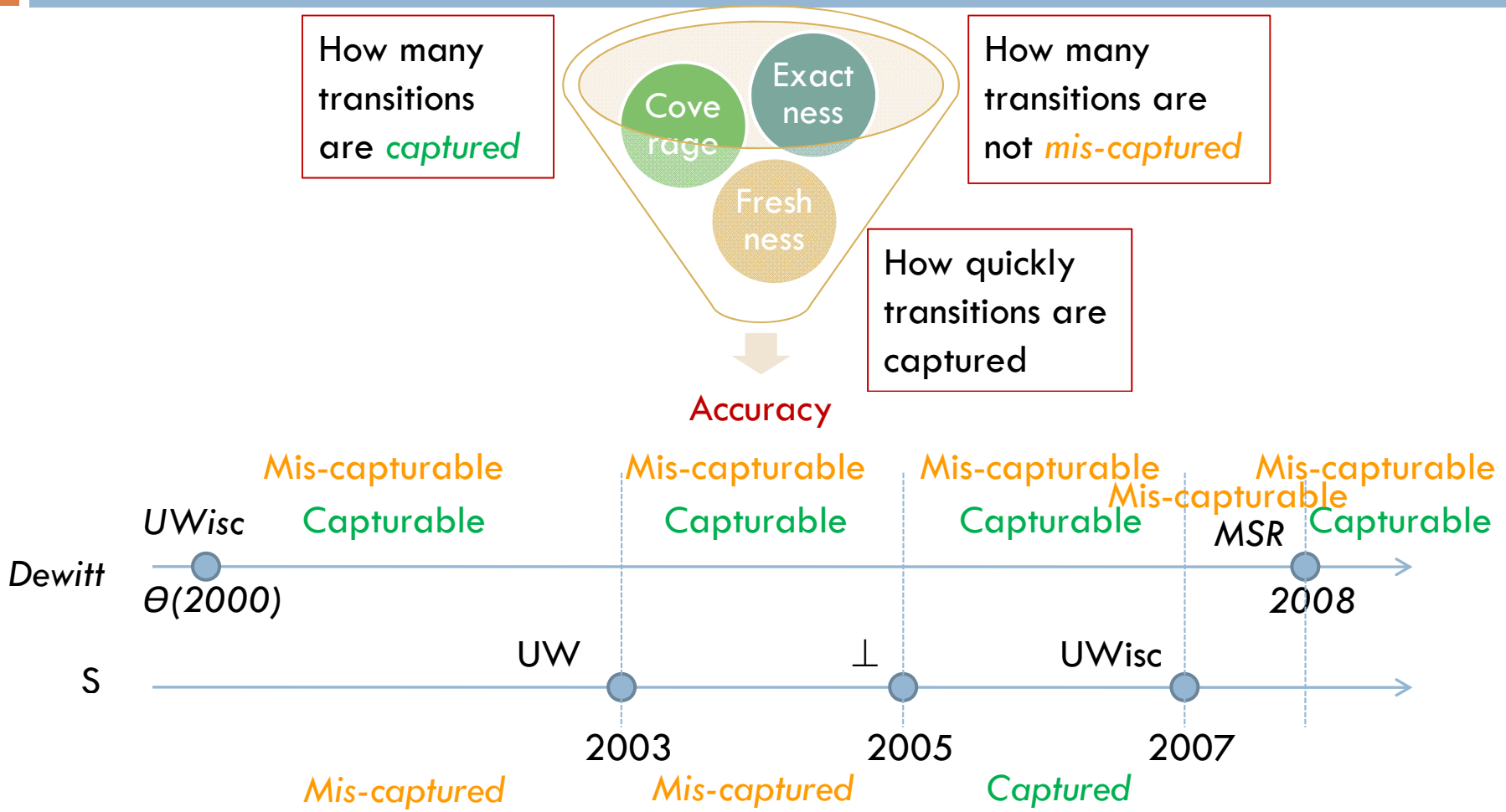
	S1	S2	S3
Stonebraker	(03, MIT)	(00, Berkeley)	(06, MIT)
Dewitt	(09, MSR)	(08, MSR)	(01, UWisc)
Bernstein	(00, MSR)	(00, MSR)	(01, MSR)
Carey	(09, UCI)	(05, AT&T)	(06, BEA)
Halevy	(07, Google)	(05, Google)	(06, UW)

A Dynamic World

- True values can evolve over time
  - A subtle third case: *out-of-date*
- Low-quality data can be caused by different reasons

	S1	S2	S3
Stonebraker (Θ, Berkeley), (02, ***MIT***)	**(03, MIT)**	**(00, Berkeley)** OUT-OF-DATE!	(01, Berkeley) **(06, MIT)**
Dewitt (Θ, UWisc), (08, ***MSR***)	(00, UWisc) **(09, MSR)**	(00, UW) (01, UWisc) **(08, MSR)**	**(01, UWisc)** OUT-OF-DATE!
Bernstein (Θ, ***MSR***)	**(00, MSR)**	**(00, MSR)**	**(01, MSR)**
Carey (Θ, Propell), (02, BEA), (08, ***UCI***)	(04, BEA) **(09, UCI)**	**(05, AT&T)** ERR!	**(06, BEA)** OUT-OF-DATE!
Halevy (Θ, UW), (05, ***Google***)	(00, UW) **(07, Google)**	(00, UWisc) (02, UW) **(05, Google)**	(01, UWisc) **(06, UW)** SLOW!

Refine Accuracy of Sources [Dong et al., 09b]



Coverage = #Captured/#Capturable (e.g., ¼=.25)

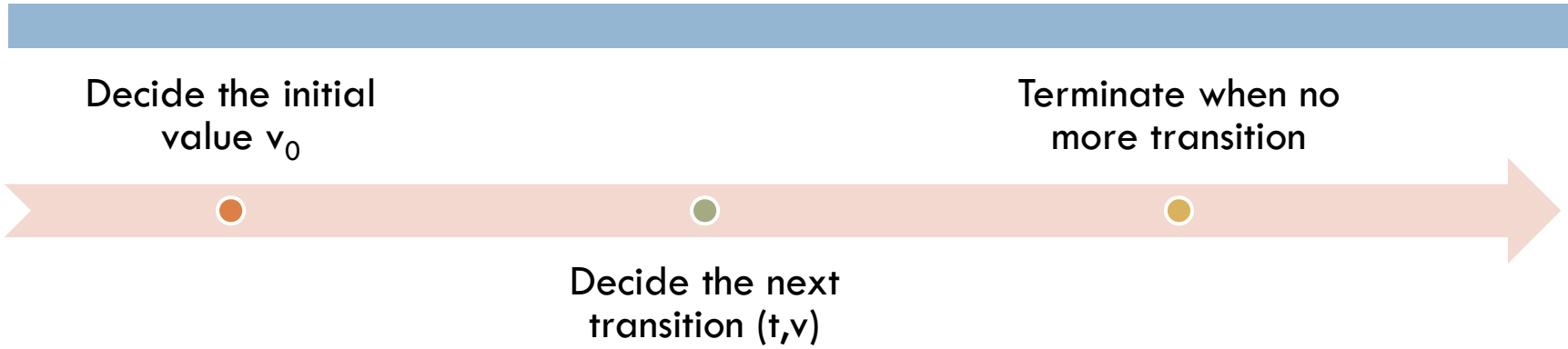
Exactness = 1-#Mis-Captured/#Mis-Capturable (e.g., 1-2/5=.6)

Freshness(Δ) = #(Captured w. length<=Δ)/#Captured (e.g., F(0)=0, F(1)=0, F(2)=1/1=1…)

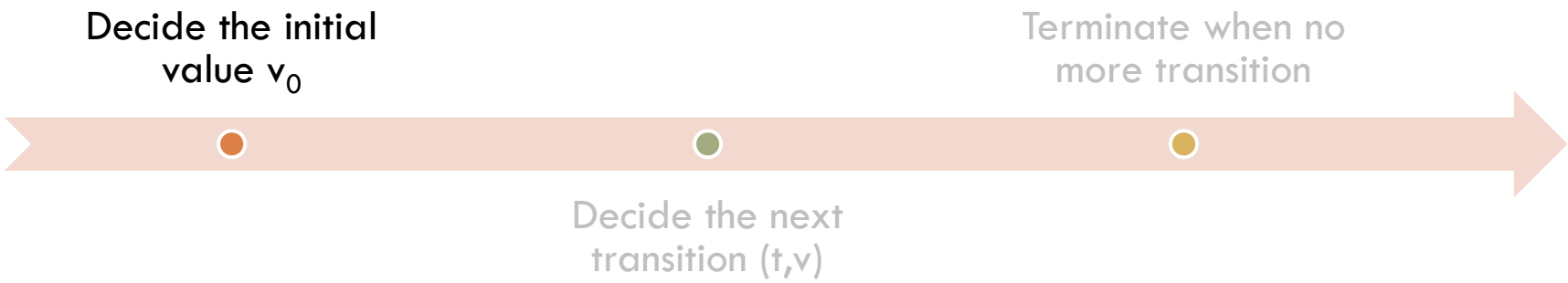
Freshness Measures in Other Work

- Other work on data freshness: Compare a materialized view with the original source
  - [Peralta, Ph.D. Thesis'06]: timeliness, currency
  - [Guo et al., 05]: completeness, consistency, currency
  - [Olston and Widom, 05]: divergence
  - [Labrinidis and Roussopoulos, 04]: QoD(freshness)
  - [Theodoratos and Bouzeghoub, 01]: consistency
  - [Cho and Garcia-Molina, 00]: freshness, age

Discover Evolving True Values



Discover Evolving True Values

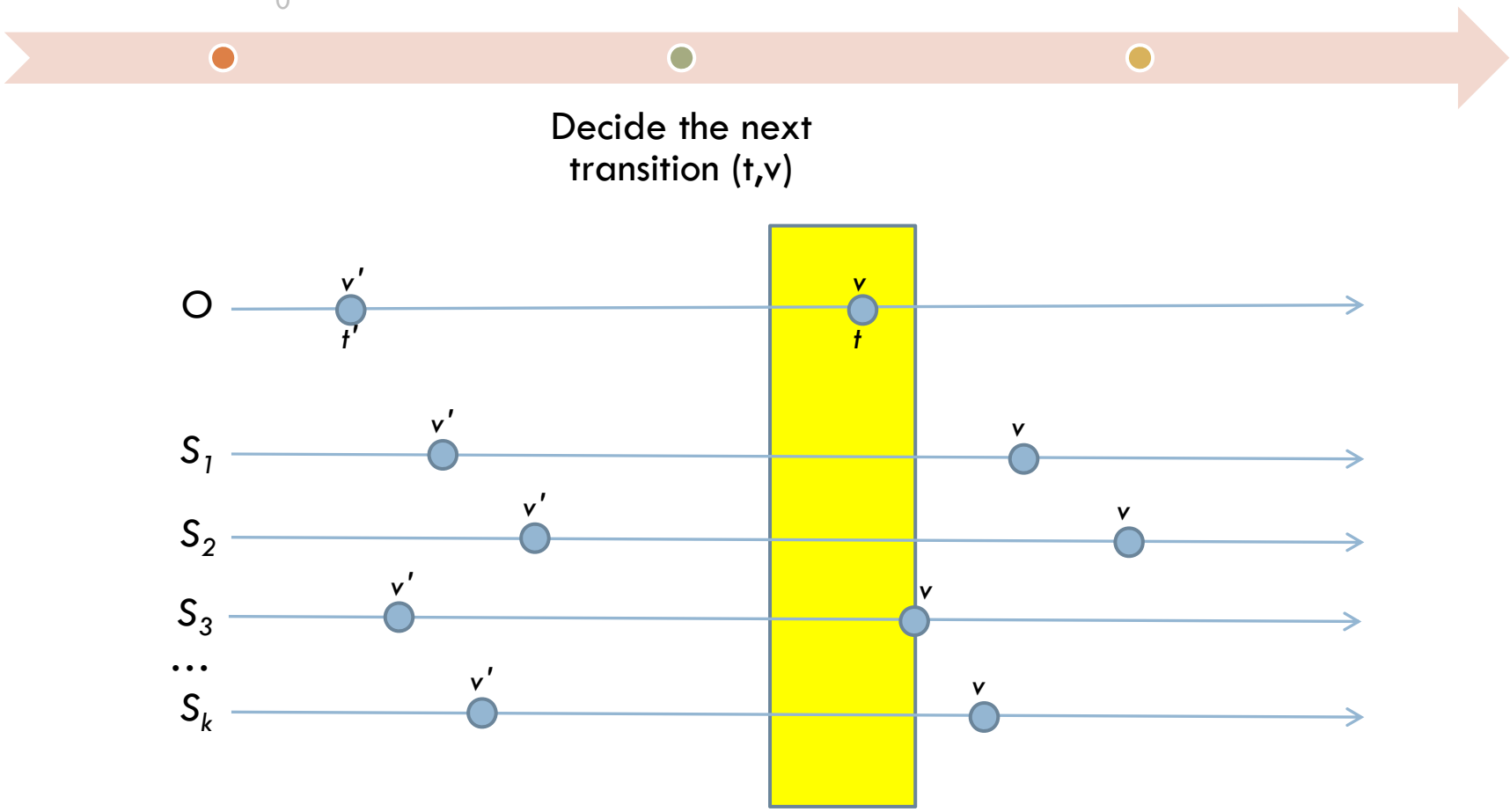


- Decide the initial value: according to the Bayes Rule, we need to know
  - $Pr(\Phi(S)|v_i)$ for each value v_i
    - If S provides v_i : E(S)C(S)
    - If S does not provide any value: E(S)(1-C(S))
    - If S provides another value: (1-E(S))/n
  - $Pr(\Phi(S)|\perp)$—the object does not exist initially
    - If S does not provide any value: E(S)
    - If S provides a value : (1-E(S))/(n+1)

Discover Evolving True Values

Decide the initial value v_0

Terminate when no more transition



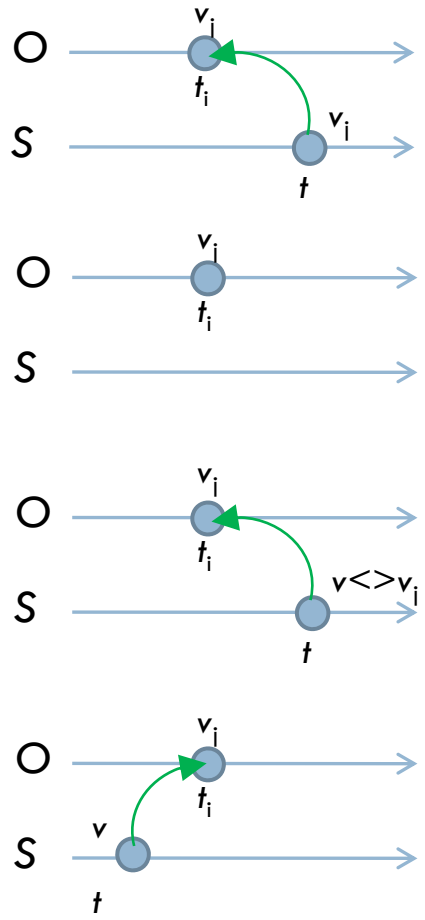
Discover Evolving True Values

Decide the initial value v_0

Terminate when no more transition

Decide the next transition (t,v)

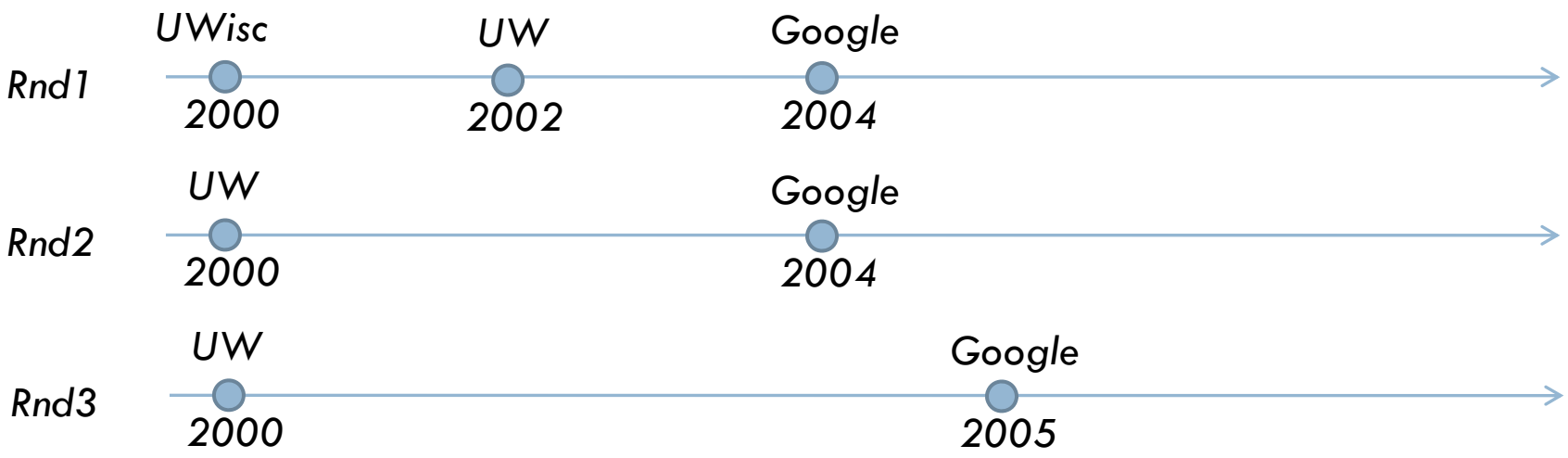
- Decide the next transition (t,v): according to the Bayes Rule, we need to know
  - $Pr(\Phi(S)|(t_i,v_i))$ for each time t_i and value v_i
    - If S provides v_i at time t : $E(S)C(S)F(S, t-t_i)$
    - If S does not update any more: $E(S)(1-C(S)F(S, t_n-t_i))$
    - If S makes a wrong update: $(1-E(S))/n(t_n-t')$ (t_n—the last obs point, t'—time of the prev update)
  - $Pr(\Phi(S)|$no more transition$)$: similarly computed
    - If S does not update any more: E(S)
    - If S makes an update: $(1-E(S))/(n+1)(t_n-t')$



An Example

	S1	S2	S3
Halevy (Θ, UW), (05, ***Google***)	(00, UW) **(07, Google)**	(00, UWisc) (02, UW) **(05, Google)**	(01, UWisc) **(06, UW)**

Affiliation for Halevy:



Advanced Truth-Discovery Techniques

- ☑ Data sources are of different quality and we trust data from accurate sources more → Consider accuracy of sources
- ☑ The real world is dynamic and the true value often evolves over time → Consider freshness of sources
  - E.g., person affiliation, business contact phone
- ☐ Data sources can copy from each other and errors can be propagated quickly → Consider dependence between sources

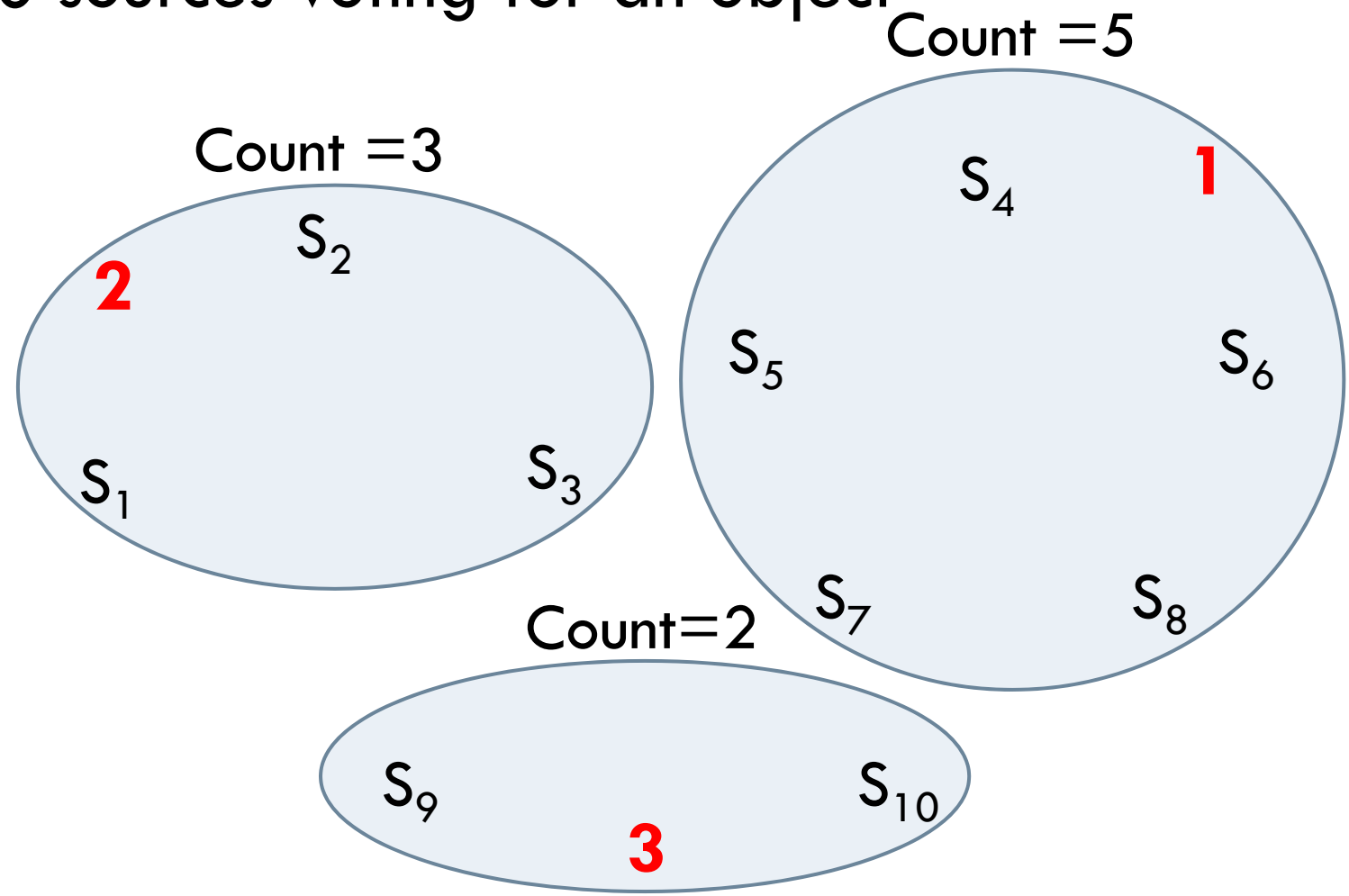
Copied Data Can Change Truth Discovery Results

- Previous methods assume source independence

	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

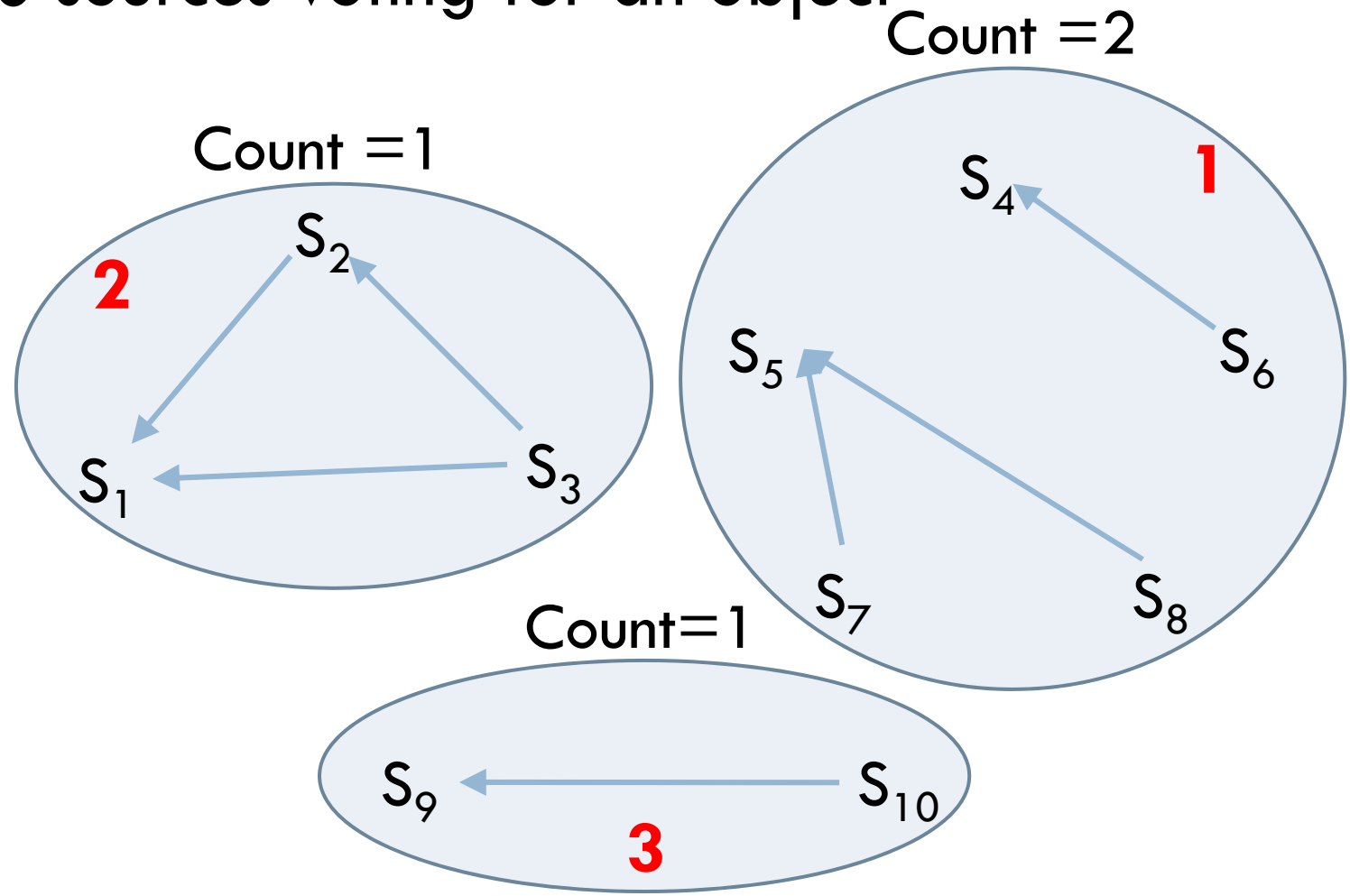
Voting for Independence Sources

- 10 sources voting for an object



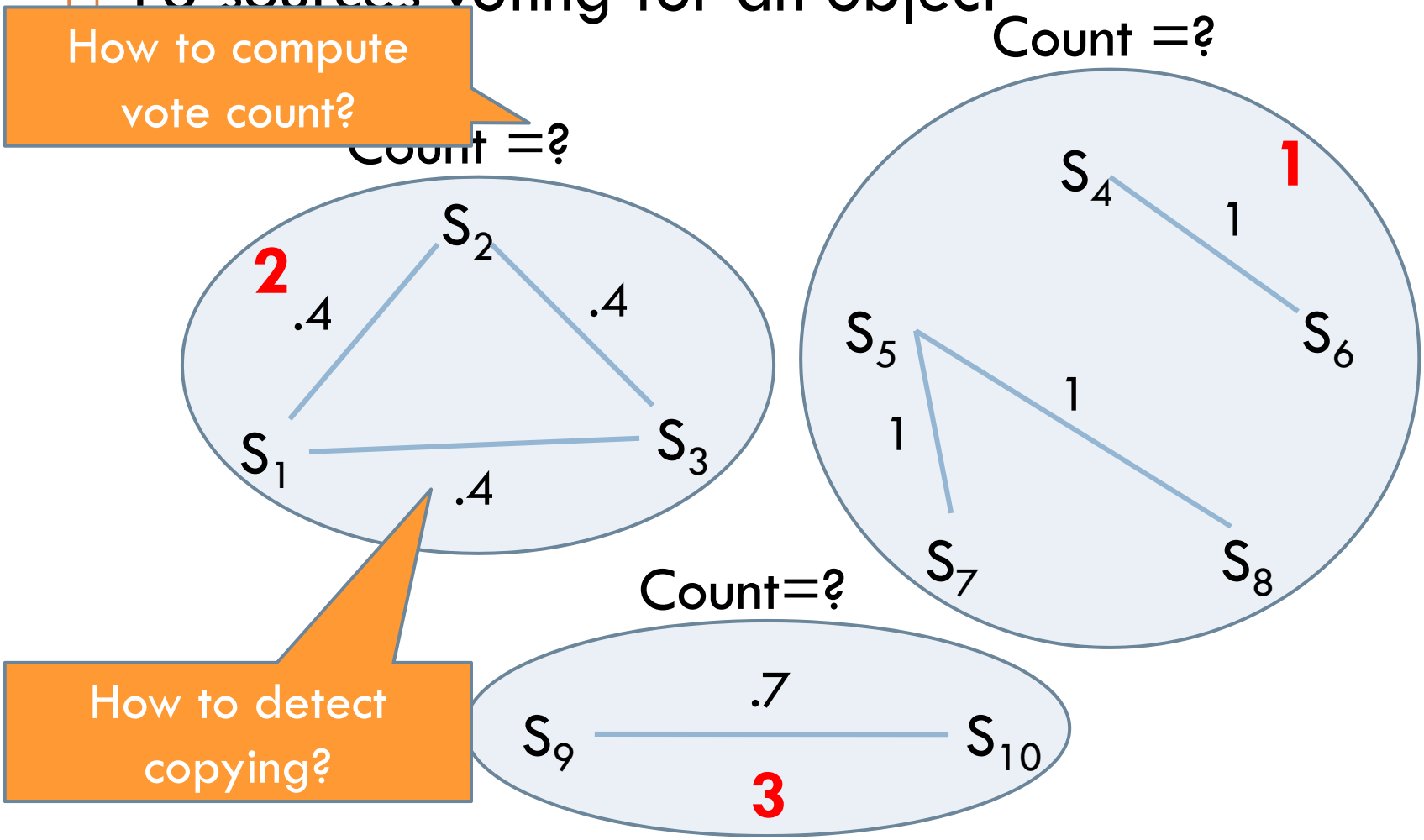
Voting w. Knowledge of Copying

- 10 sources voting for an object



Voting w. Probabilistic Copying

- 10 sources voting for an object



Considering Dependence

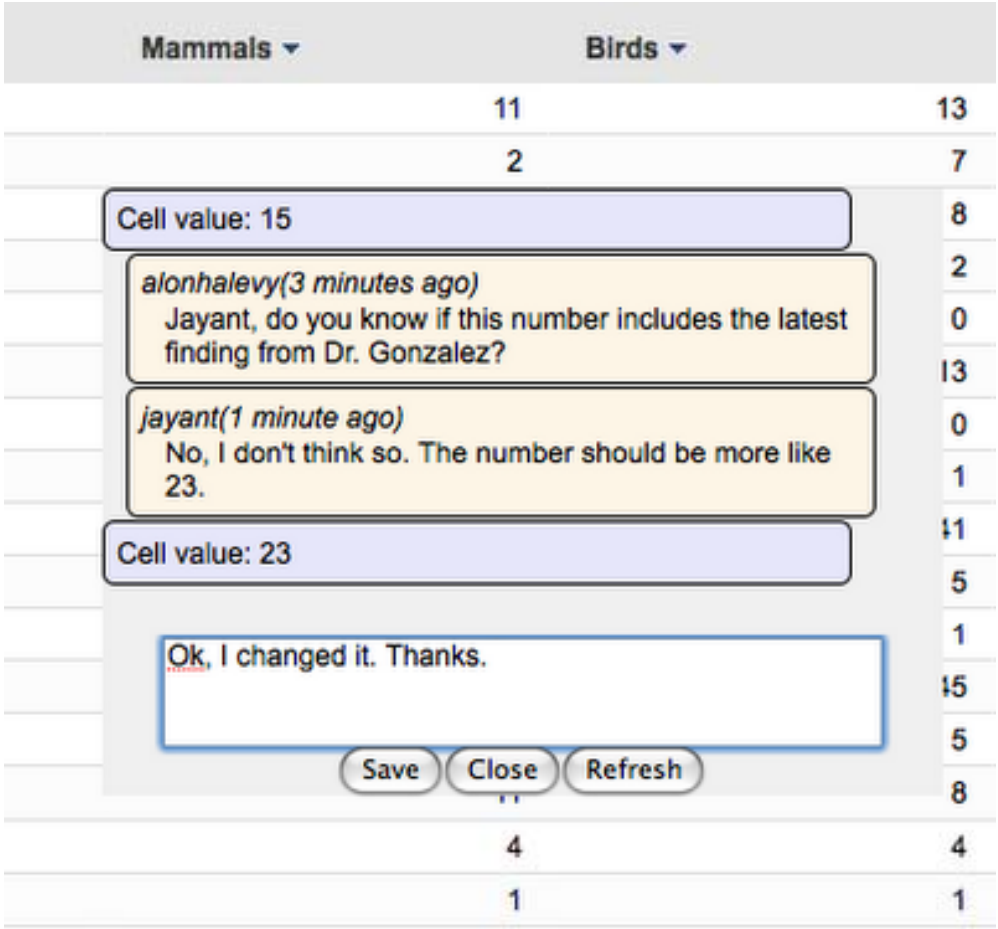
- Opinion pooling: combine probability distribution from multiple experts
  - Combination of opinions [Chang, Ph.D. thesis'85]
  - Reconciliation of probability distributions [Lindley, 83]
  - Updating of belief in the light of someone else's opinion [French, 80]
- Data fusion w. source dependence
  - [Dong et al., 09a][Dong et al., 09b]

See Tomorrow's talks in "Data Integration I"

Outline

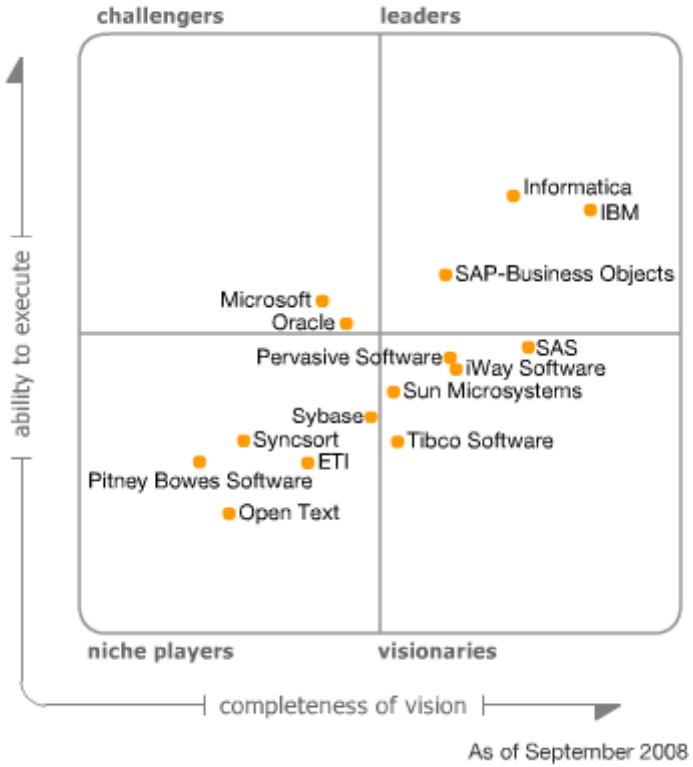
- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Data fusion in existing integration systems
- Open problems

Web Integration—Google Fusion Tables



- Allows discussion of values between users

Commercial DI Tools



Source: Gartner

- Typical ETL tools support rule-based fusion
  - IIS (IBM Information Server)
  - SSIS (Microsoft's SQL Server Integration Services)
  - Etc.

See details in survey [Bleiholder and Naumann, 08]

Research DI Systems w. Awareness of Data Conflicts

System	Conflict types	Methodology	Strategy	Specification
Multibase	Schematic, data	Resolution	Choose, Avg, Min, Max, Sum, …	Manually, in query
Hermes	Schematic, data	Resolution	MostRecent, Choose	Manually, in mediator
Fusionplex	Schematic, object, data	Resolution	MostRecent, Min, Max, Avg, …	Manually, in query
HumMer	Schematic, object, data	Resolution	MostAbstract, Vote, Min, ChooseDepen…	Manually, in query
Ajax	Schematic, object, data	Resolution	Various	Manually, in workflow definition
TSIMMIS	Schematic, data	Avoidance	Choose	Manually, rules in mediator
SIMS/Ariadne	Schematic, data	Avoidance	Choose	Automatically
Infomix	Schematic, data	Avoidance	onlyConsistentValue	Automatically
Hippo	Schematic, object, data	Avoidance	onlyConsistentValue	Automatically
ConQuer	Schematic, object, data	Avoidance	onlyConsistentValue	Automatically
Rainbow	Schematic, object, data	Avoidance	onlyConsistentValue	Automatically
Pegasus	Schematic, data	Ignorance	Escalate	Manually
Nimble	Unknown	Ignorance	Escalate	Manually
Carnot	Schematic	Ignorance	Escalate	Automatically
InfoSleuth	Schematic	Ignorance	Escalate	Unknown
Potter's Wheel	Schematic	Ignorance	Escalate	Manually, transformation

 See details in survey [Bleiholder and Naumann, 08]

Other DI Systems

- Research DI systems
  - Trio: including accuracy and lineage into data model
  - Information Manifold
  - Garlic
  - Disco (Distributed Information Search Component)
  - etc.
- Peer data management systems
  - Orchestra: allowing multiple viewpoints
  - Hyper: isolating the minimum amount of data to reach consistency

See details in survey [Bleiholder and Naumann, 08]

Outline

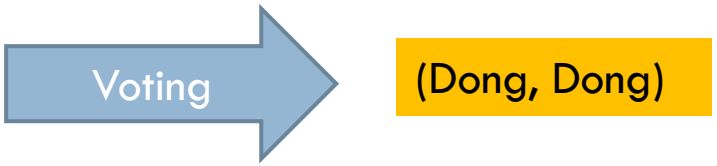
- Data fusion in the integration process
- Foundations of data fusion
  - Conflict resolution strategies and functions
  - Conflict resolution operators
- Advanced truth-discovery techniques
- Data fusion in existing integration systems
- Open problems

Open Problems

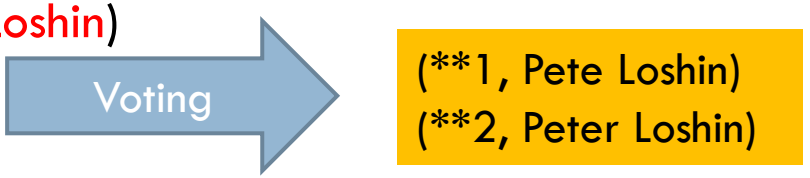
- Accuracy of fusion
- Efficiency of fusion
- Usability of fusion
- Interaction with other components of data integration

Accuracy of Fusion (I)

- **Challenge 1**: Correlated values
  - E.g.1, (firstName, lastName) from 4 sources
    - S1: (Xin, Dong)
    - S2: (Xin Luna, Dong)
    - S3: (Dong, Xin)
    - S4: (Dong, Xin Luna)



  - E.g.2, (ISBN, authors) from 3 sources
    - S1: (**1, Peter Loshin) (**2, Peter Loshin)
    - S2: (**1, Pete Loshin)
    - S3: (**1, Pete Loshin)



- **Current effort:** ChooseDepending(val, col)
- **Directions:** consider correlation at the attribute level and at the instance level.

Accuracy of Fusion (II)

- **Challenge 2:** Different formating styles
  - E.g., (ISBN, authors) from 4 sources

Src1
(**1, Pete Loshin)
(**2, Dennis Suhanovs)
(**3, Zhigang Xiang, Roy A Plastock)
(**4, Peter Aiken, David M Allen)…

Src2
(**1, Pete Loshin)
(**2, Dennis Suhanovs)
(**3, Zhigang Xiang, Roy Plastock)
(**4, Peter Aiken, David Allen)…

Skip middle-names

Src3
(**1, Pete Loshin)
(**2, Dennis Suhanovs)
(**3, Zhigang Xiang)
(**4, David Allen, Peter Aiken)…

Accuracy of Fusion (II)

- **Challenge 2:** Different formating styles
  - E.g., (ISBN, authors) from 4 sources

Src1
(**1, Pete Loshin)
(**2, Dennis Suhanovs)
(**3, Zhigang Xiang, Roy A Plastock)
(**4, Peter Aiken, David M Allen)…

Src2
(**1, Pete Loshin)
(**2, Dennis Suhanovs)
(**3, Zhigang Xiang, Roy Plastock)
(**4, Peter Aiken, David Allen)…

Skip middle-names

Src3
(**1, Pete Loshin)
(**2, Dennis Suhanovs)
(**3, Zhigang Xiang)
(**4, David Allen, Peter Aiken)…

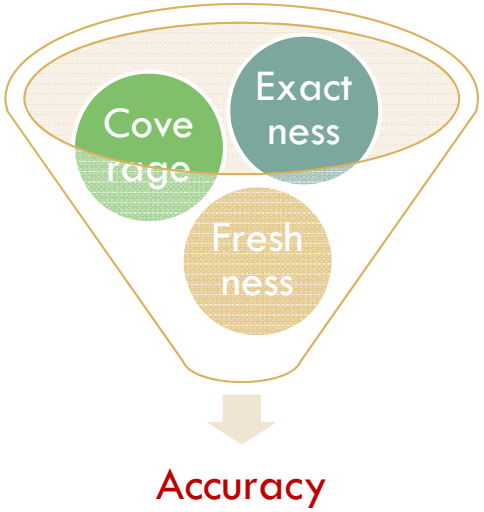
Src4
(**1, Pete Loshin)
(**2, Dennis uhanovs)
(**3, Zhigang Xiang)
(**4, Peter Aiken)…

Only first-authors

- **Current effort:** consider value similarity
- **Directions:** consider formatting styles used by each source.

Accuracy of Fusion (III)

- **Challenge 3:** Source profiling
  - **Current effort:** accuracy (coverage, exactness, freshness)
  - Data properties can be different for different categories of data
    - Source A is a vertical source on restaurants
    - Source B knows very well about NYC
  - Data properties can evolve over time
    - Source C improves its data over time
- **Directions:** partition data into different portions and profile on each portion



Efficiency of Fusion (I)

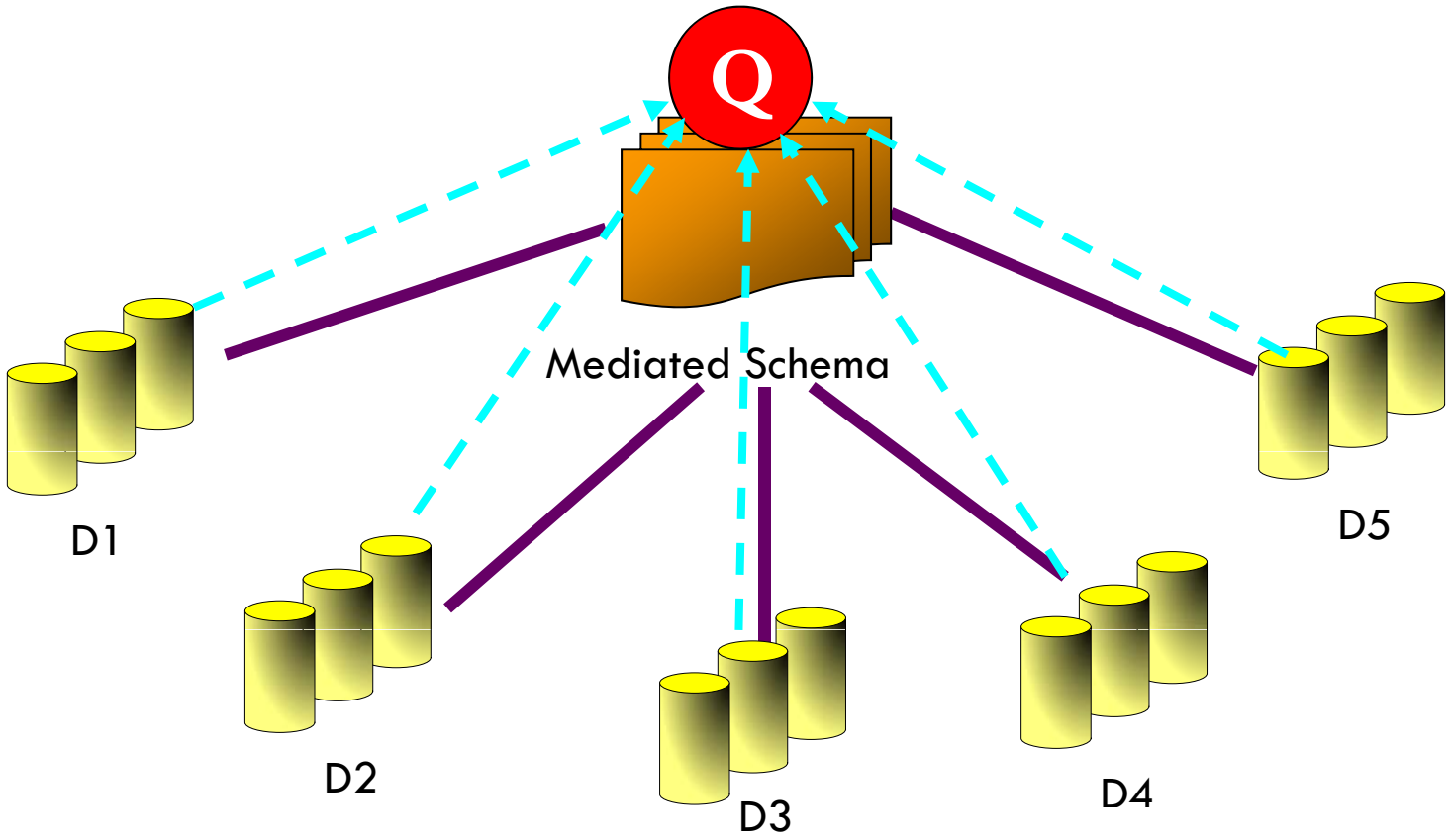
- **Challenge 4:** Incremental fusion
  - When we have more data sources (e.g., Src4) or lose some data sources, shall we do data fusion from scratch?

Src1	Src2
(**1, Pete Loshin) (**2, Dennis Suhanovs) (**3, Zhigang Xiang, Roy A Plastock) (**4, Peter Aiken, David M Allen)…	(**1, Pete Loshin) (**2, Dennis Suhanovs) (**3, Zhigang Xiang, Roy Plastock) (**4, Peter Aiken, David Allen)…
Src3	**Src4**
(**1, Pete Loshin) (**2, Dennis Suhanovs) (**3, Zhigang Xiang) (**4, David Allen, Peter Aiken)…	(**1, Pete Loshin) (**2, Dennis uhanovs) (**3, Zhigang Xiang) (**4, Peter Aiken)…

  - When more data come, shall we start from scratch?
- **Directions:** maintain metadata or statistics, retain data lineage.

Efficiency of Fusion (II)

- **Challenge 5**: Runtime fusion
  - In some applications fusing data upfront is infeasible



- **Directions**: maintain source profiles by sampling; emphasize efficiency.

Usability of Fusion (I)

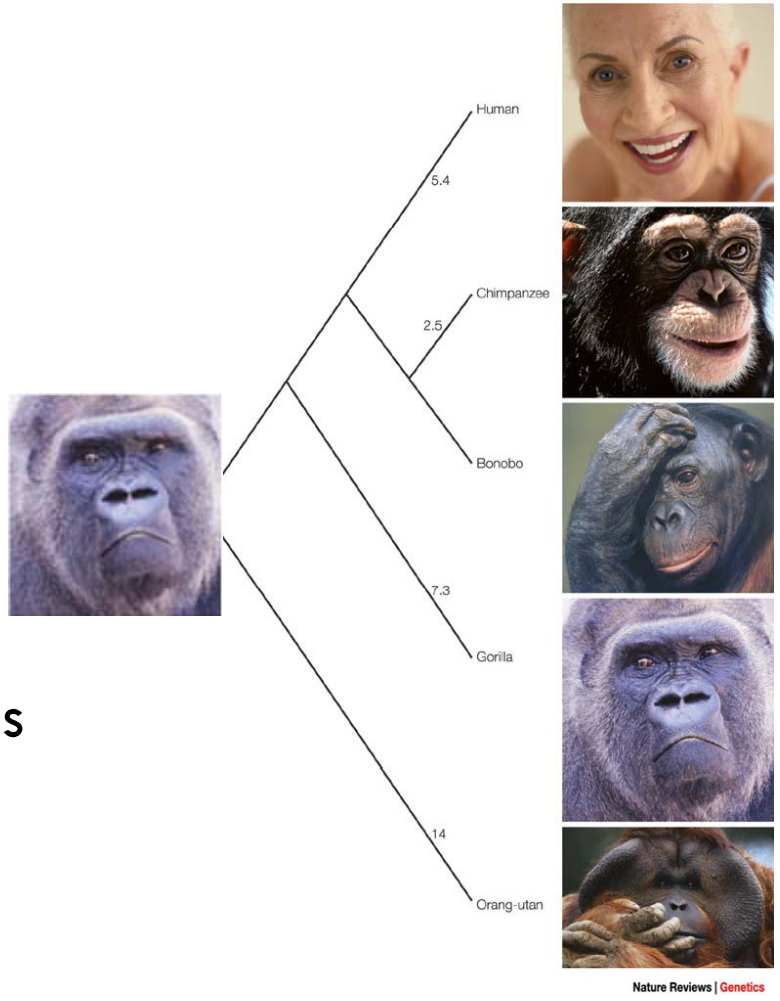
- **Challenge 6:** Personalized fusion
  - Express preference on certain sources
  - Emphasize certain property; e.g., up-to-date vs. high coverage
  - Use certain formats; e.g., full author list vs. only first author
- **Current effort:**
  - Function choose(src)
  - Operator Prioritized-Merge
- **Directions:**
  - A language to express such user preferences
  - Algorithms for efficient execution.

Usability of Fusion (II)

- **Challenge 7**: User feedback
  - Correct certain errors
- **Directions:**
  - Critical questions that can best improve the fusion results
  - A way for users to browse source data and fusion results, and correct mistakes
  - Quickly fixing errors and propagation to related items

Usability of Fusion (III)

- **Challenge 8**: Data Lineage
  - Legal requirement
  - Application requirement: e.g., fusing two customers
  - HCI requirement: HOW did you merge the data? And WHY?
- **Directions:**
  - Effective representation of lineage information
  - Explanation of merging decisions
  - Effective way to find disappeared data items
  - Reversibility and repeatability



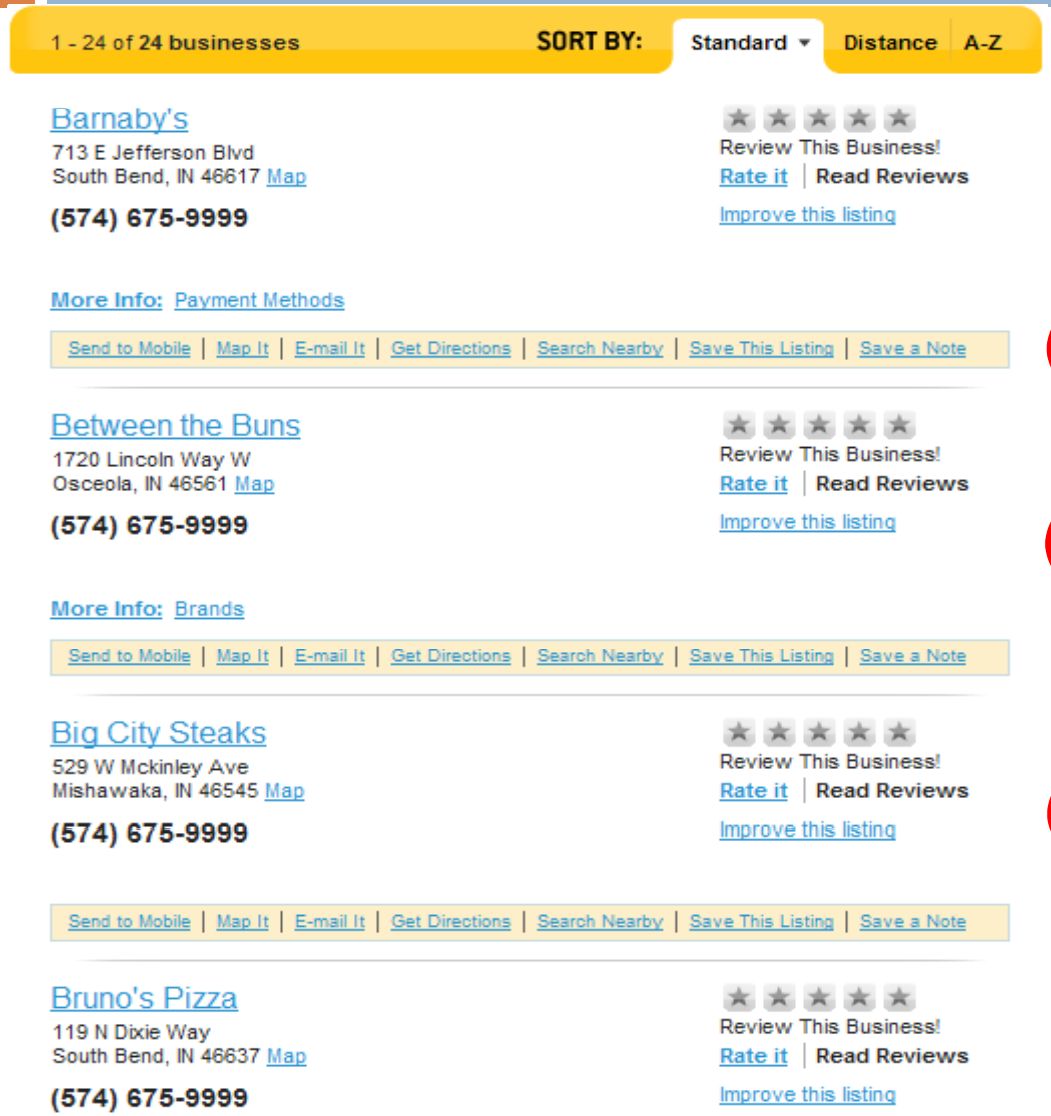
Interaction with Other Components of DI (I)

- **Challenge 9**: Fuse data w. different schemas
  - E.g., Contact information from three sources
    - S1: (pid = "1", work phone = "1234", home phone = "8765", mobile phone = "4321")
    - S2: (pid = "1", daytime phone = "1234", evening phone = "4321")
    - S3: (pid = "1", phone = "4321")
- **Directions:** Combine data fusion w. schema matching

Interaction with Other Components of DI (II)

- **Challenge 10:** Distinguish wrong values from alternative representations of correct values
  - E.g., A quiz

A Quiz



Which type of listing are they?

A: are the same business

B: are different businesses sharing the same phone#

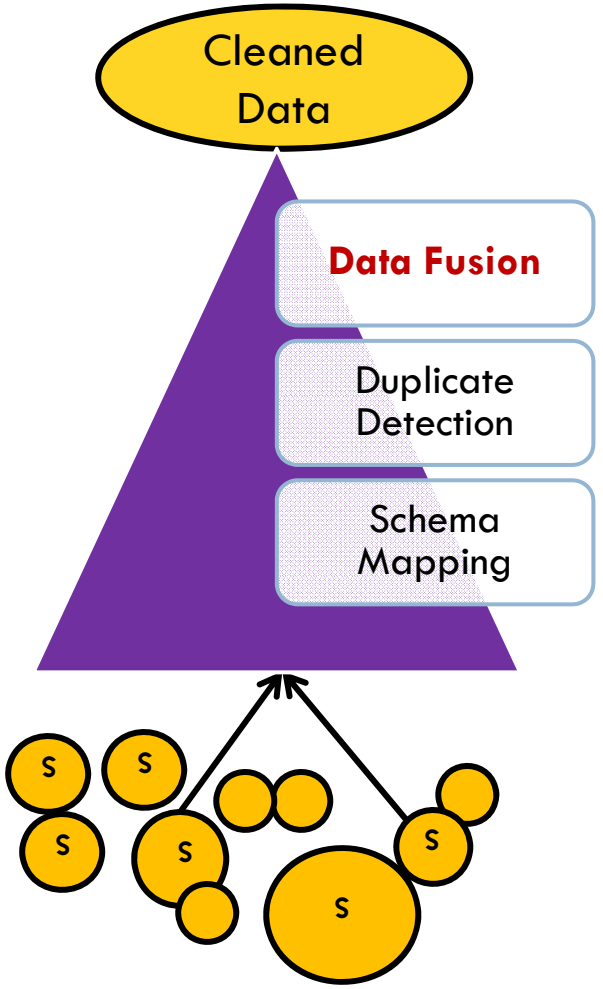
C: are different businesses, only one with correct phone#

Interaction with Other Aspects of DI (II)

- **Challenge 10:** Distinguish wrong values from alternative representations of correct values
  - E.g., A quiz
- **Directions:** Combine data fusion w. record linkage

Conclusions

- Foundations
  - Strategies and functions
  - Operators
- Advanced techniques
  - Consider accuracy
  - Consider freshness
  - Consider dependence
- Open problems
  - Accuracy
  - Efficiency
  - Usability
  - Interaction with other components of DI



References

- **Survey**
  - [BN08] J. Bleiholder, F. Naumann. Data Fusion. ACM Computing Survey 2009.
- **Foundations of Fusion**
  - [BBB+05] A. Bilke, J. Bleiholder, C. Böhm, K. Draba, F. Naumann and M. Weis. Automatic Data Fusion with HumMer. VLDB demo 2005.
  - [BDN07] Jens Bleiholder, Karsten Draba, and Felix Naumann. FuSem - Exploring Different Semantics of Data Fusion (demo) VLDB demo 2007.
  - [BN05] J. Bleiholder, F. Naumann. Declarative Data Fusion - Syntax, Semantics, and Implementation. ADBIS 2005.
  - [CS05] S. Cohen and Y. Sagiv. An incremental algorithm for computing ranked full disjunctions. PODS 2005.
  - [DeM89] DeMichiel, L. G. Resolving database incompatibility: An approach to performing relational operations over mismatched domains. TKDE 1989.
  - [FFM05] Ariel Fuxman, Elham Fazli, Renee J. Miller. ConQuer: Efficient Management of Inconsistent Databases. SIGMOD 2005.
  - [GL94] C. A. Galindo-Legaria. Outerjoins as disjunctions. SIGMOD 1994.
  - [GPZ01] S. Greco, L. Pontieri, and E. Zumpano. Integrating and managing conflicting data. International Andrei Ershov Memorial Conference on Perspectives of System Informatics, 2001.
  - [LSS94] Lim, E.-P., Srivastava, J., and Shekhar, S. Resolving attribute incompatibility in database integration: An evidential reasoning approach. ICDE 1994.
  - [RPZ04] J. Rao, H. Pirahesh, and C. Zuzarte. Canonical abstraction for outerjoin optimization. SIGMOD 2004.
  - [RU96] A. Rajaraman and J. D. Ullman. Integrating information by outerjoins and full disjunctions. PODS1996.
  - [SD05] Dan Suciu and Nilesh Dalvi. Probabilistic Databases. Tutorial at SIGMOD 2005.
  - [YÖ99] L. L. Yan and M. T. Özsu. Conflict tolerant queries in AURORA. CoopIS 1999.

References (con't)

- **Advanced truth-discovery techniques**

- [ESD+09] L. Berti-Equille, A. D. Sarma, X. L. Dong, A. Marian, and D. Srivastava. Sailing the information ocean with awareness of currents: Discovery and application of source dependence. In *CIDR, 2009*.
- [BRR+05] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM TOIT, 5:231–297, 2005*.
- [BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30(1–7):107–117, 1998*.
- [CG00] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. In *SIGMOD, 2000*.
- [Cha85] K. Chang. *Combination of opinions: the expert problem and the group consensus problem. PhD thesis, University of* California, Berkeley, 1985.
- [DBS09a] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. VLDB, 2009.
- [DBS09b] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. In VLDB, 2009
- [French80] S. French. Updating of belief in the light of someone else's opinion. *Jour. of Roy. Statist. Soc. Ser. A, 143:43–48, 1980*.
- [GLR05] H. Guo, P.- A. Larson, and R. Ramakrishnan. Caching with 'good enough' currency, consistency, and completeness. In *VLDB, 2005*.
- [KSG03] S. Kamvar, M. Schlosser, and H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. In *Proc. of WWW, 2003*.
- [Kle98] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA, 1998*.
- [Lin83] D. Lindley. Reconciliation of probability distributions. *Oper. Res., 31:866–880, 1983*.

References (con't)

- [LR04] A. Labrinidis and N. Roussopoulos. Exploring the tradeoff between performance and data freshness in database-driven web servers. *VLDB J., 13(3):240–255, 2004.*
- [OW05] C. Olston and J. Widom. Efficient monitoring and querying of distributed, dynamic data via approximate replication. *IEEE Data Eng. Bull., 28(1):11–18, 2005.*
- [SL03] A. Singh and L. Liu. TrustMe: anonymous management of trust relationships in decentralized P2P systems. In *IEEE Intl. Conf. on Peer-to-Peer Computing, 2003.*
- [TB01] D. Theodoratos and M. Bouzeghoub. Data currency quality satisfaction in the design of a data warehouse. *Int. J. ooperative Inf. Syst., 10(3):299–326, 2001.*
- [WM07] M. Wu and A. Marian. Corroborating answers from multiple web sources. In *Proc. of WebDB, 2007.*
- [YHY07] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the Web. In *Proc. Of SIGKDD, 2007.*