

Toward Reference Reconciliation and Inconsistency Checking in Ontology-based Information Integration

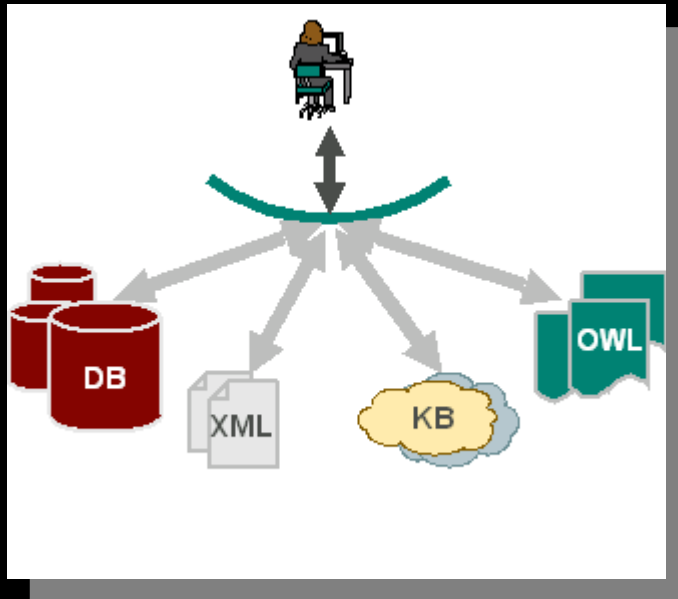
keywords: data, web, ontology, semantic, integration

Presented by:
Paea LePendu
paea@cs.uoregon.edu

Outline

- ⇒ Background
- ⇒ Theoretical Framework
- ⇒ Motivation
- ⇒ Reference Reconciliation & Inconsistency

Integration \supseteq Data Exchange \supset Query Answering

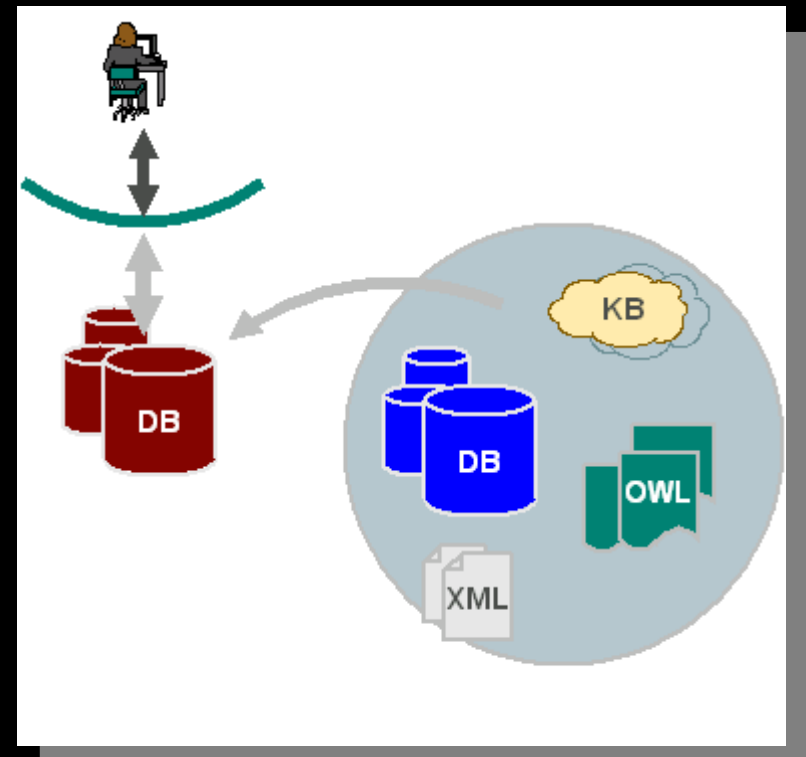


Data Federation

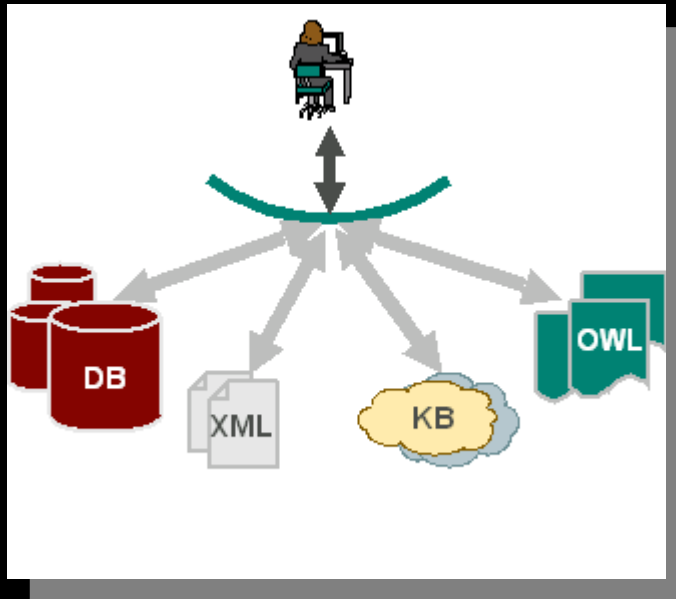
main task: query answering

Data Migration

main task: data translation



The difference is the level of:
(1) Materialization
(2) Constraint Enforcement

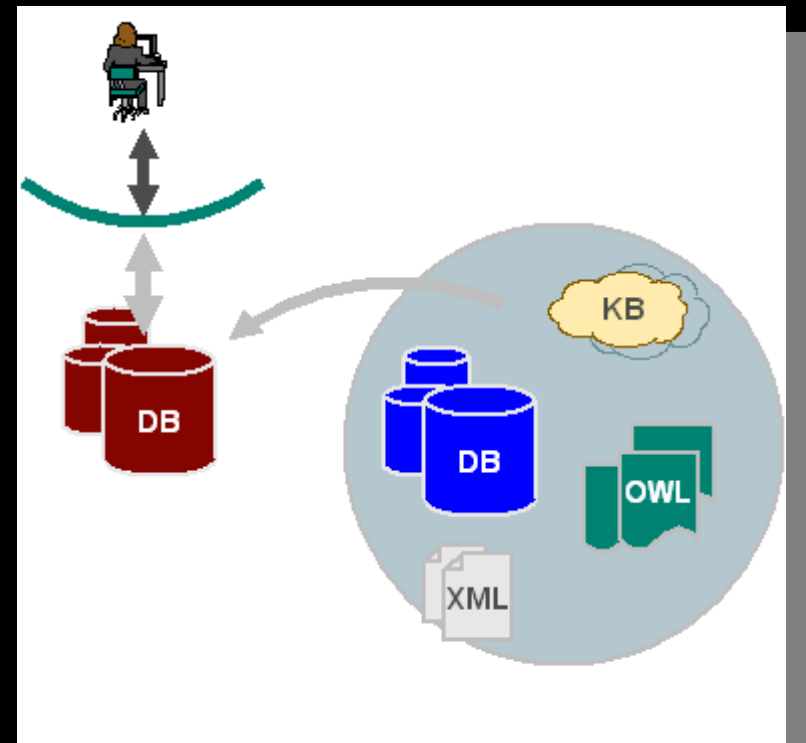


Data Federation

virtual
requirement: on-the-fly, sound

Data Migration

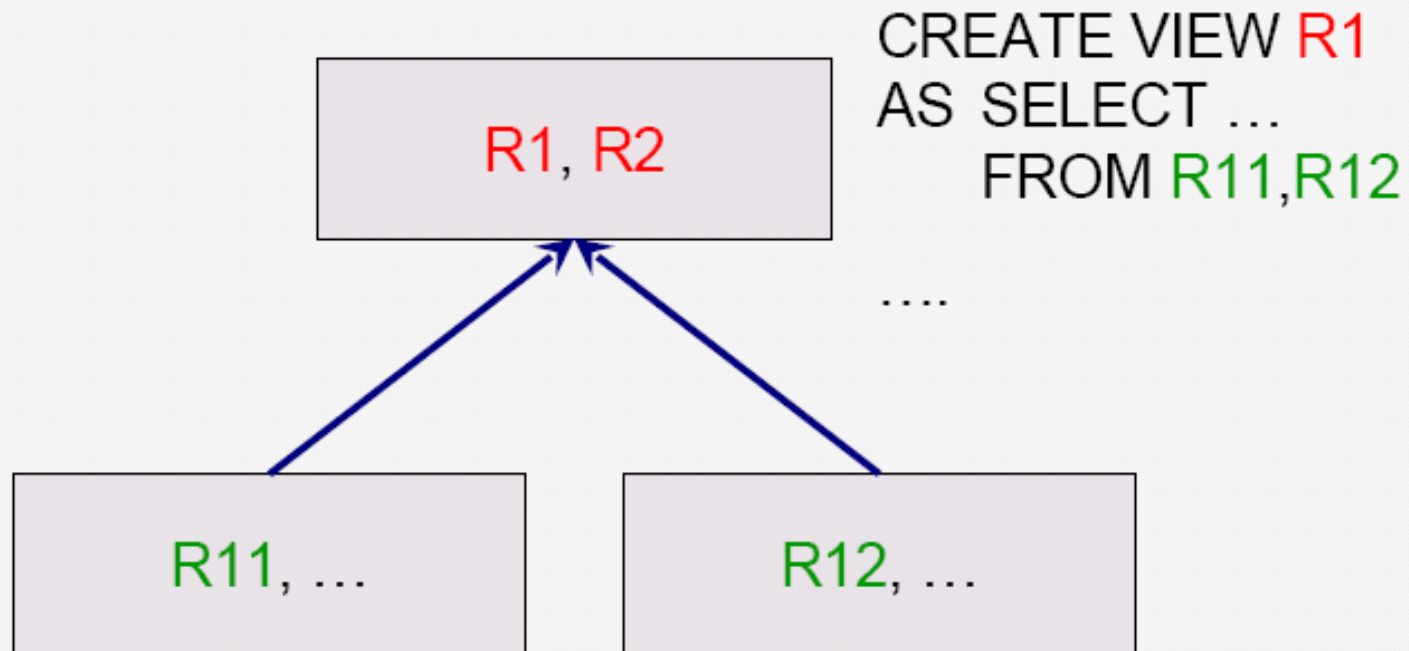
materialized
requirement: consistent at target



Global as View

Source: P. Atzeni [ICDE 2006]

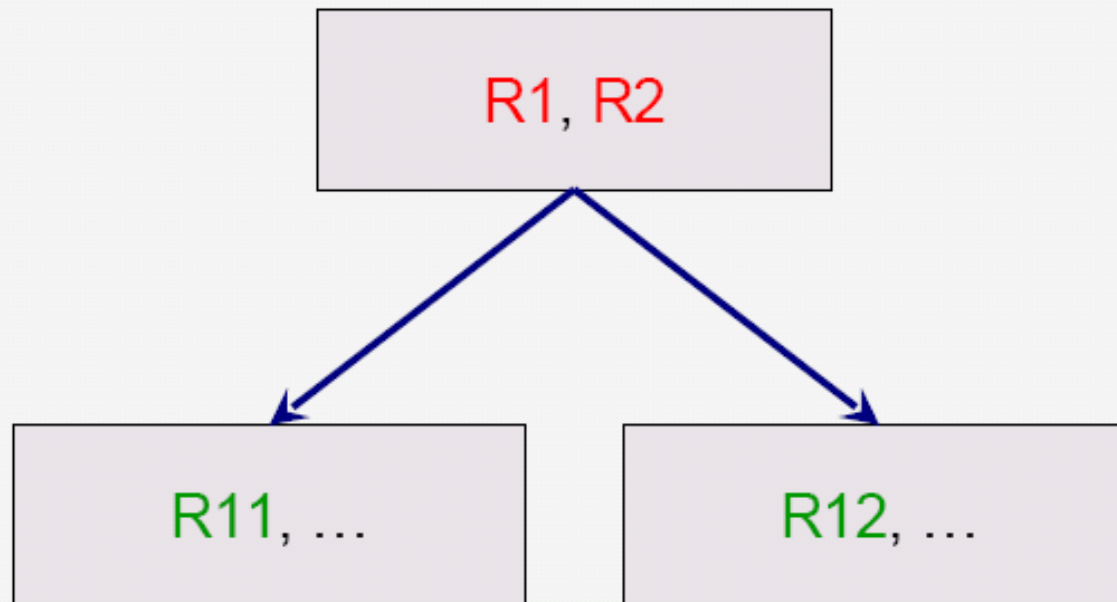
GAV, in integration



Local as View

Source: P. Atzeni [ICDE 2006]

LAV, in integration



```
CREATE VIEW R11  
AS SELECT ...  
FROM R1,R2
```

GAV vs. LAV

Source: Kambhampati & Knoblock [AAAI 2002]

GAV

- Not modular
 - Addition of new sources changes the mediated schema
- Can be awkward to write mediated schema without loss of information
- Query reformulation easy
 - *reduces to view unfolding (polynomial)*
 - Can build hierarchies of mediated schemas
- Best when
 - Few, stable, data sources
 - well-known to the mediator (e.g. corporate integration)
 - **Garlic, TSIMMIS, HERMES**

vs.

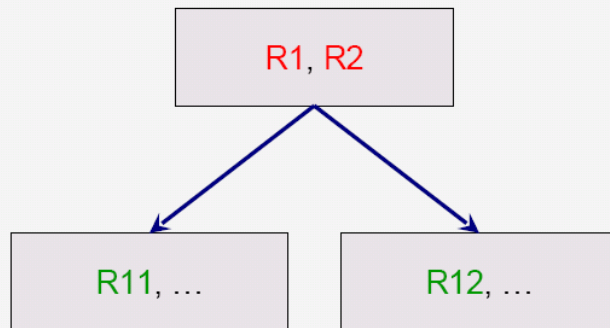
LAV

- Modular--adding new sources is easy
- Very flexible--power of the entire query language available to describe sources
- Reformulation is hard
 - Involves answering queries only using views (can be intractable—see below)
- Best when
 - Many, relatively unknown data sources
 - possibility of addition/deletion of sources
 - **Information Manifold, InfoMaster, Emerac, Havasu**

SQL – GAV & LAV

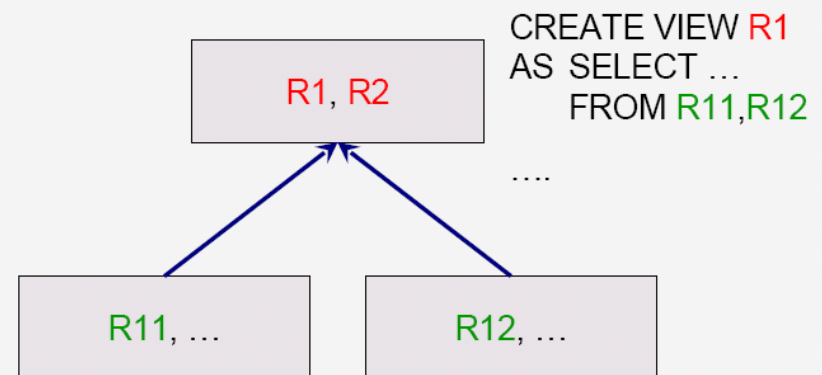
- ➔ GAV:: **R1** AS SELECT x,y,... FROM **R11, R12**
- ➔ LAV:: **R11** AS SELECT x,y,... FROM **R1, R2**

LAV, in integration



```
CREATE VIEW R11  
AS SELECT ...  
FROM R1,R2
```

GAV, in integration

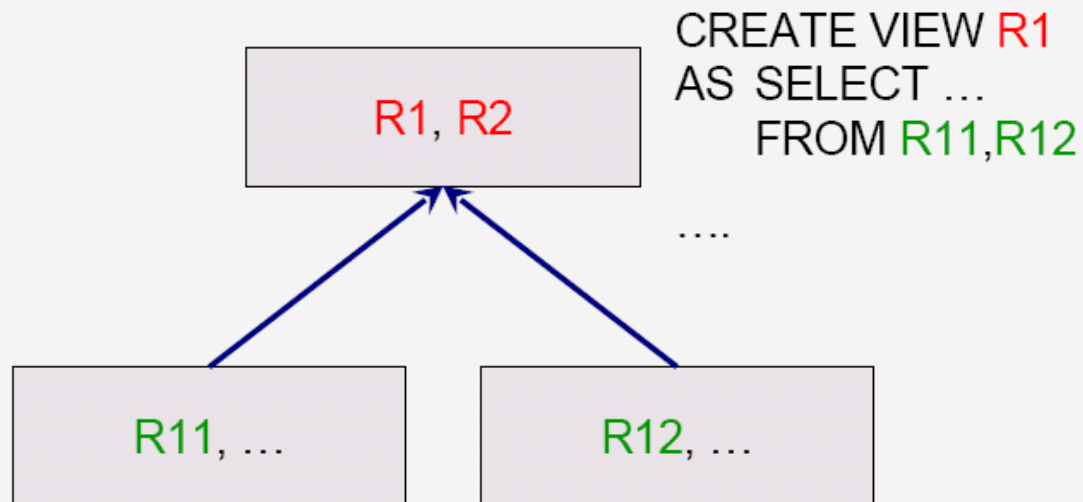


```
CREATE VIEW R1  
AS SELECT ...  
FROM R11,R12  
....
```

Datalog - GAV

→ GAV:: $R1(x,y,...) :- R11(x,y,...) \wedge R12(x,y,...)$

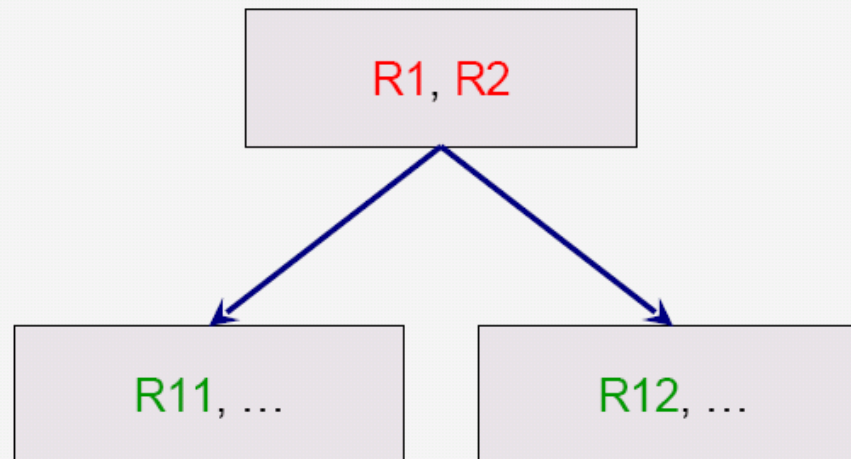
GAV, in integration



Datalog - LAV

⇒ LAV:: $R11(x,y,\dots) :- R1(x,y,\dots) \wedge R2(x,y,\dots)$

LAV, in integration



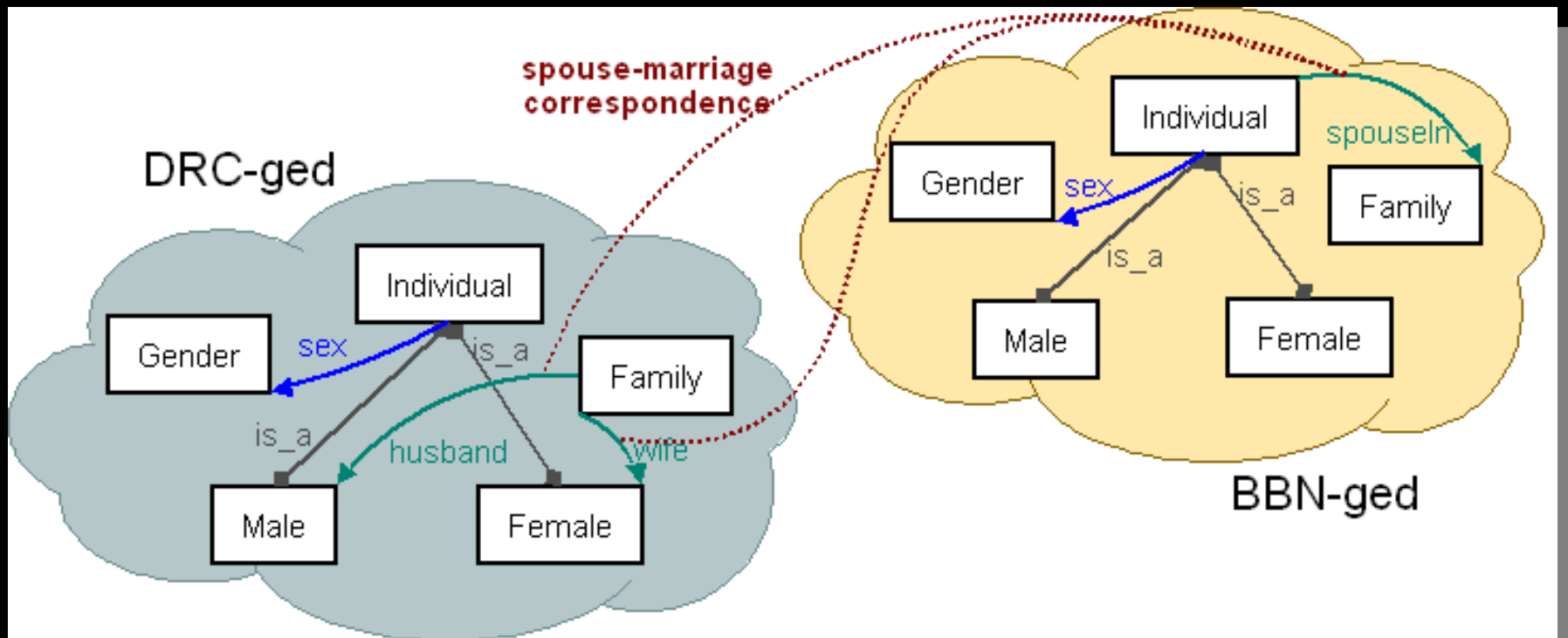
```
CREATE VIEW R11  
AS SELECT ...  
FROM R1,R2
```

Outline

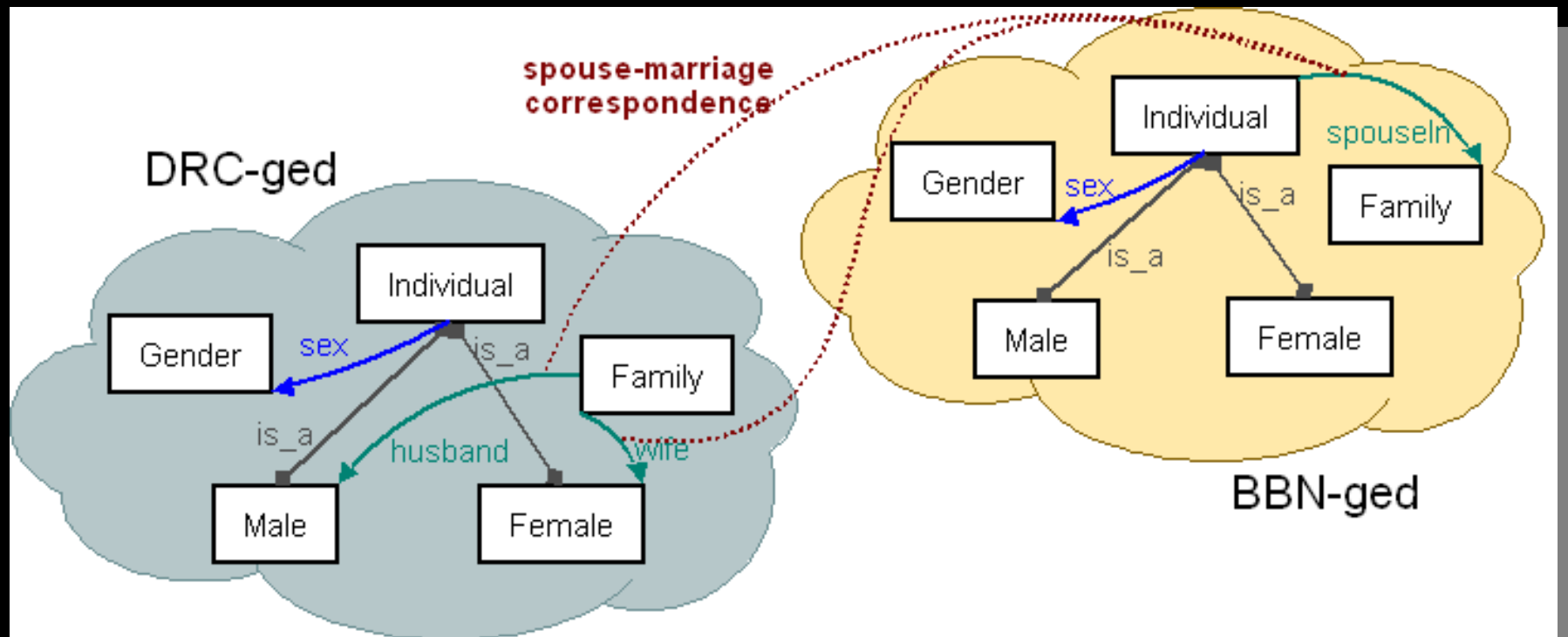
- ⇒ Background
- ⇒ Theoretical Framework
- ⇒ Motivation
- ⇒ Reference Reconciliation & Inconsistency

Ontologies

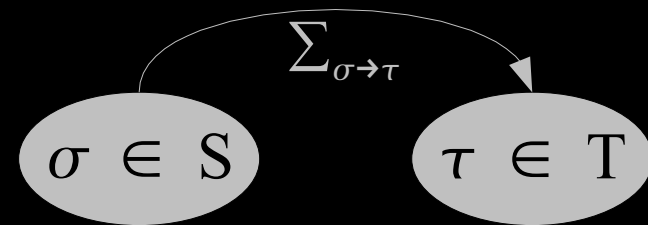
- ➔ Types (square)
- ➔ Properties (colored arrow)
- ➔ Axioms (dotted line)



Bridging Axioms

$$\forall x.y.z, \text{ DRC-ged: } Family(z) \wedge Male(x) \wedge Female(y) \wedge husband(z, x) \wedge wife(z, y) \\ \Rightarrow \text{ BBN-ged: } Family(z) \wedge spouseIn(x, z) \wedge spouseIn(y, z)$$


Logical Framework



- ⇒ Main idea:
 - Translation as **entailment**
 - **Inference** as the implementation

- ⇒ Data Translation (\Rightarrow_D):

$(\Sigma_{\sigma \rightarrow \tau}, \sigma) \Rightarrow_D \tau$ *only if* $(\Sigma_{\sigma \rightarrow \tau}, \sigma) \not\models \tau$

$(\Sigma_{\sigma \rightarrow \tau}, \sigma) \Rightarrow_D \tau \iff (\Sigma_{\sigma \rightarrow \tau}, \sigma) \vdash \tau \rightarrow (\Sigma_{\sigma \rightarrow \tau}, \sigma) \not\models \tau$

- ⇒ Query Answering and Translation (\Rightarrow_Q):

$(\Sigma_{\sigma \rightarrow \tau}, \sigma?) \Rightarrow_Q \tau?$ *only if* $(\Sigma_{\sigma \rightarrow \tau}, \theta(\tau?)) \not\models \theta(\sigma?)$

$(\Sigma_{\sigma \rightarrow \tau}, \sigma?) \Rightarrow_Q \tau? \iff (\Sigma_{\sigma \rightarrow \tau}, \theta(\tau?)) \vdash \theta(\sigma?) \rightarrow (\Sigma_{\sigma \rightarrow \tau}, \theta(\tau?)) \not\models \theta(\sigma?)$

Outline

- ⇒ Background
- ⇒ Theoretical Framework
- ⇒ **Motivation**
- ⇒ Reference Reconciliation & Inconsistency

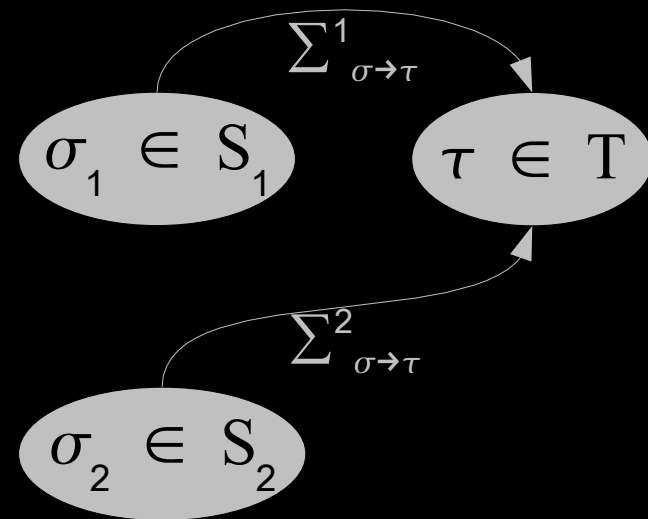
Multi-Source Query Answering

⇒ Target query:
(fullname x ?y) ^ (phone x ?z)

⇒ Answers from Source S_1 :
{y/John Doe, z/509-3333}
{y/James Doe, z/541-9999}
{y/Kathy U, z/206-2345}

⇒ Answers from Source S_2 :
{y/J. Doe, z/509-3334}
{y/J. Doe, z/541-4444}
{y/J. Doe, z/541-9999}
{y/K. Unbehaum, z/206-2345}

⇒ *So now what?*



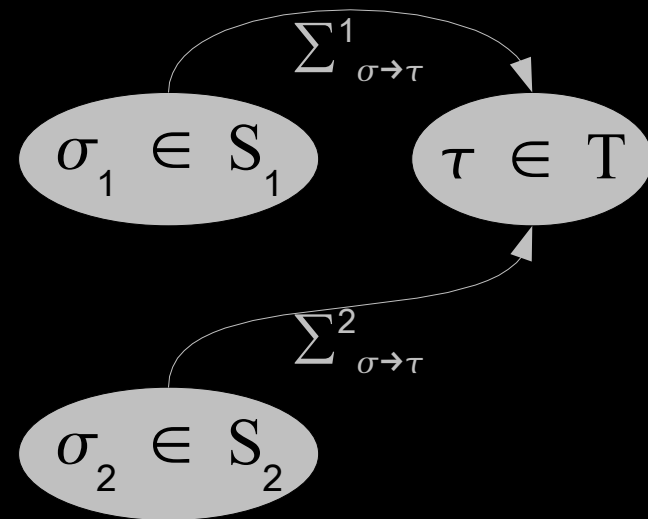
So now what?

- ➔ Characterizing inconsistency:
 - Eliminating redundancy
 - Augmenting missing data
 - Correcting wrong data

Outline

- ⇒ Background
- ⇒ Theoretical Framework
- ⇒ Motivation
- ⇒ Reference Reconciliation & Inconsistency

Step 1: Translation and Answering



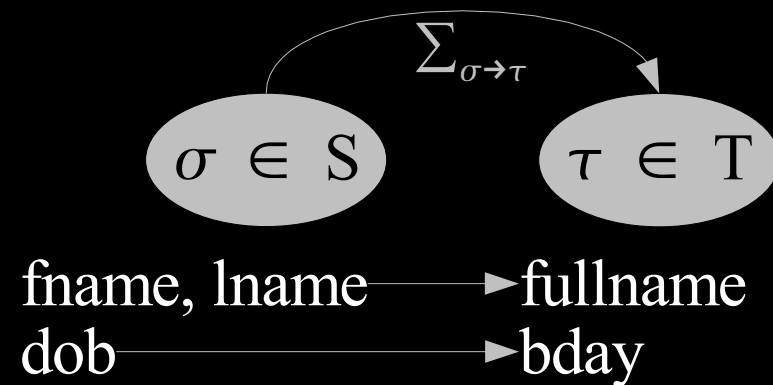
⇒ Σ^1 :

- $(\sigma:\text{fname } x \text{ fn}) \wedge (\sigma:\text{lname } x \text{ ln}) \rightarrow (\tau:\text{fullname } x \text{ concat}(\text{fn}, \text{ln}))$
- $(\sigma:\text{telephone } x \text{ t}) \rightarrow (\tau:\text{phone } x \text{ t})$

⇒ Σ^2 :

- $(\sigma:\text{fullname } x \text{ fn}) \rightarrow (\tau:\text{fullname } x \text{ fn})$
- $(\sigma:\text{phone } x \text{ t}) \rightarrow (\tau:\text{phone } x \text{ t})$

Step 2: Reference Reconciliation ①



- ⇒ S^1 : (candidate key)
 - $(\text{fname } x \text{ fn}) \wedge (\text{lname } x \text{ ln}) \wedge (\text{fname } y \text{ fn}) \wedge (\text{lname } y \text{ ln}) \rightarrow (= x y)$

- ⇒ T: (candidate key)
 - $(\text{fullname } x \text{ fn}) \wedge (\text{bday } x \text{ b}) \wedge (\text{fullname } y \text{ ln}) \wedge (\text{bday } y \text{ b}) \rightarrow (= x y)$

- ⇒ Reconciliation Rule as a form of *logical weakening*:
 - Σ^1 :
 - $(\text{fname } x \text{ fn}) \wedge (\text{lname } x \text{ ln}) \wedge (\text{fname } y \text{ fn}) \wedge (\text{lname } y \text{ ln}) \wedge (\text{dob } x \text{ b}) \wedge (\text{dob } y \text{ b}) \rightarrow (= x y)$

Step 1 + 2: Intertwining translation and reconciliation

⇒ Target query:

$(\tau:\text{fullname } x \text{ ?}y) \wedge (\tau:\text{phone } x \text{ ?}z)$

⇒ Translation:

RULE: $(\sigma:\text{fname } x \text{ fn}) \wedge (\sigma:\text{lname } x \text{ ln}) \rightarrow (\tau:\text{fullname } x \text{ concat(fn,ln))$

RULE: $(\sigma:\text{telephone } x \text{ t}) \rightarrow (\tau:\text{phone } x \text{ t})$

=====

Source query: $(\sigma:\text{fname } x \text{ ?fn}) \wedge (\sigma:\text{lname } x \text{ ?ln}) \wedge (\sigma:\text{telephone } x \text{ ?t})$

⇒ Reconciliation:

RULE: $(\text{fullname } x \text{ fn}) \wedge (\text{bday } x \text{ b}) \wedge (\text{fullname } y \text{ ln}) \wedge (\text{bday } y \text{ b}) \rightarrow (= x y)$

⇒ Query Weakening:

RULE: $(\sigma:\text{dob } x \text{ b}) \rightarrow (\tau:\text{bday } x \text{ b})$

=====

$(\sigma:\text{fname } x \text{ ?fn}) \wedge (\sigma:\text{lname } x \text{ ?ln}) \wedge (\sigma:\text{telephone } x \text{ ?t}) \wedge (\sigma:\text{dob } x \text{ ?b})$

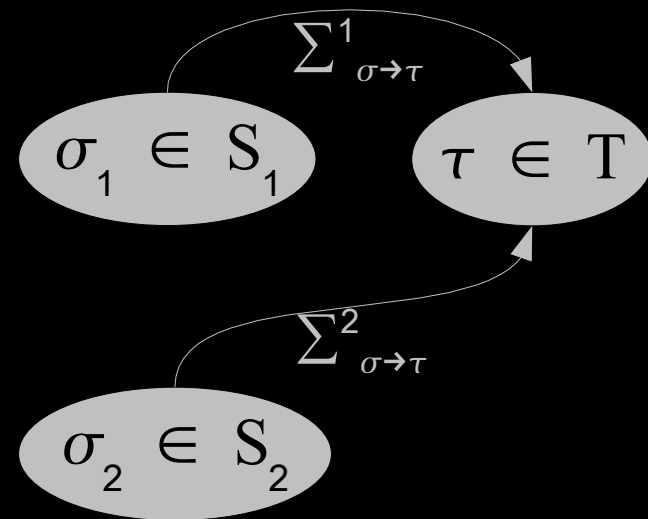
Step 2: Reference Reconciliation ①

S_1

{y/John Doe, b/8-29-1974, z/509-3333}
{y/James Doe, b/4-13-1970, z/541-9999}
{y/Kathy U, b/2-28-1977, z/206-2345}

S_2

{y/J. Doe, b/8-29-1974, z/509-3334}
{y/J. Doe, b/4-13-1970, z/541-4444}
{y/J. Doe, b/4-13-1970, z/541-9999}
{y/K. Unbehaum, b/2-28-1977, z/206-2345}



Step 3: Inconsistency ?

→ Redundancy

$\{y/\text{James Doe, b}/4\text{-}13\text{-}1970, z/541\text{-}9999\} +$
 ~~$\{y/\text{J. Doe, b}/4\text{-}13\text{-}1970, z/541\text{-}9999\}$~~

→ Augmentation

$\{y/\text{Kathy U, b}/2\text{-}28\text{-}1977, z/206\text{-}2345\} +$
 $\{y/\text{K. Unbehau, b}/2\text{-}28\text{-}1977, z/206\text{-}2345\}$

$\{y/\text{Kathy Unbehau, b}/2\text{-}28\text{-}1977, z/206\text{-}2345\}$

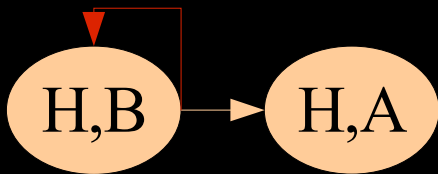
→ Incorrectness

$\{y/\text{John Doe, b}/8\text{-}29\text{-}1974, z/509\text{-}3333\} \neq$
 $\{y/\text{J. Doe, b}/8\text{-}29\text{-}1974, z/509\text{-}3334\}$

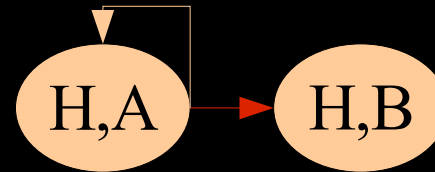
Finally... future work

- ➔ Complexity: expressiveness vs. computability
- ➔ Complexity: open vs. closed worlds

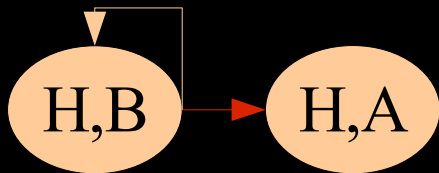
$$H(x_1, x_2) \rightarrow \exists y H(x_2, y)$$



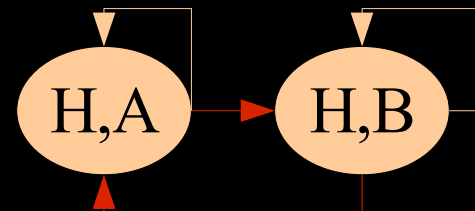
$$H(x_1, x_2) \rightarrow \exists y H(x_1, y)$$



$$H(x_1, x_2) \rightarrow \exists y H(y, x_2)$$



$$\begin{aligned} H(x_1, x_2) &\rightarrow \exists y H(x_1, y) \\ H(x_1, x_2) &\rightarrow \exists y H(y, x_2) \end{aligned}$$



References

- ➔ M. Lenzerini. Data integration: A theoretical perspective. [PODS 2002]
- ➔ P. Kolaitis. Schema Mappings, Data Exchange, and Metadata Management. [PODS 2005]
- ➔ X. Dong et al. Reference Reconciliation in Complex Information Spaces. [SIGMOD 2005]
- ➔ D. Dou & P. LePendu. Ontology-based Integration for Relational Databases. [ACMSAC 2006]

Thank you.