Introducing myself

- Heterogeneous data management.
- Imperfect data management.

Dealing with Imperfection in Schema Integration

Some preliminary ideas

Matteo Magnani

Dept. of Computer Science
University of Bologna
Via Mura A.Zamboni, 7
40127 Bologna (Italy)

Nikos Rizopoulos

Dept. of Computing
180 Queen's Gate
South Kensington Campus
Imperial College London
SW7 2AZ London (UK)

Outline

- Imperfect data: an overview.

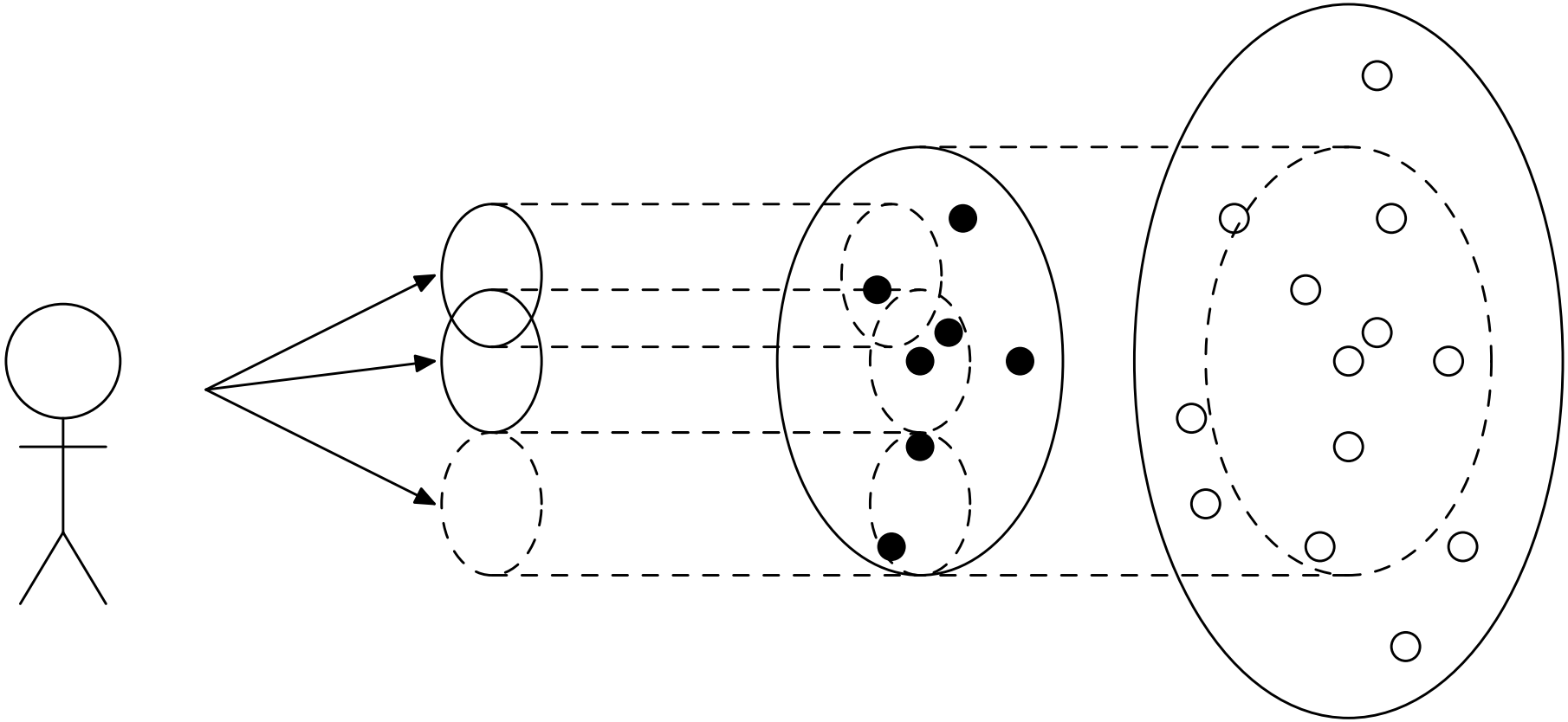
Outline

- Imperfect data: an overview.
- Uncertainty in data integration based on semantic schema matching.
  - Semantic Schema Integration.
  - Adding uncertainty.

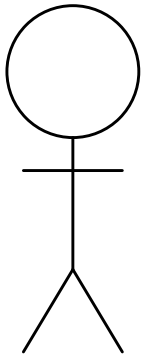
Imperfect Data: An Overview

A taxonomy of imperfection

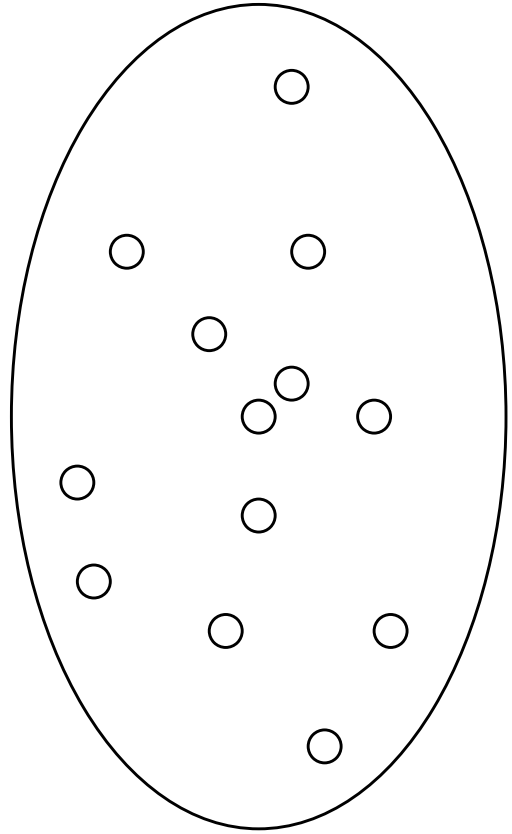
- To identify the main classes of imperfection, we use the following scenario:



(Complete) Absence

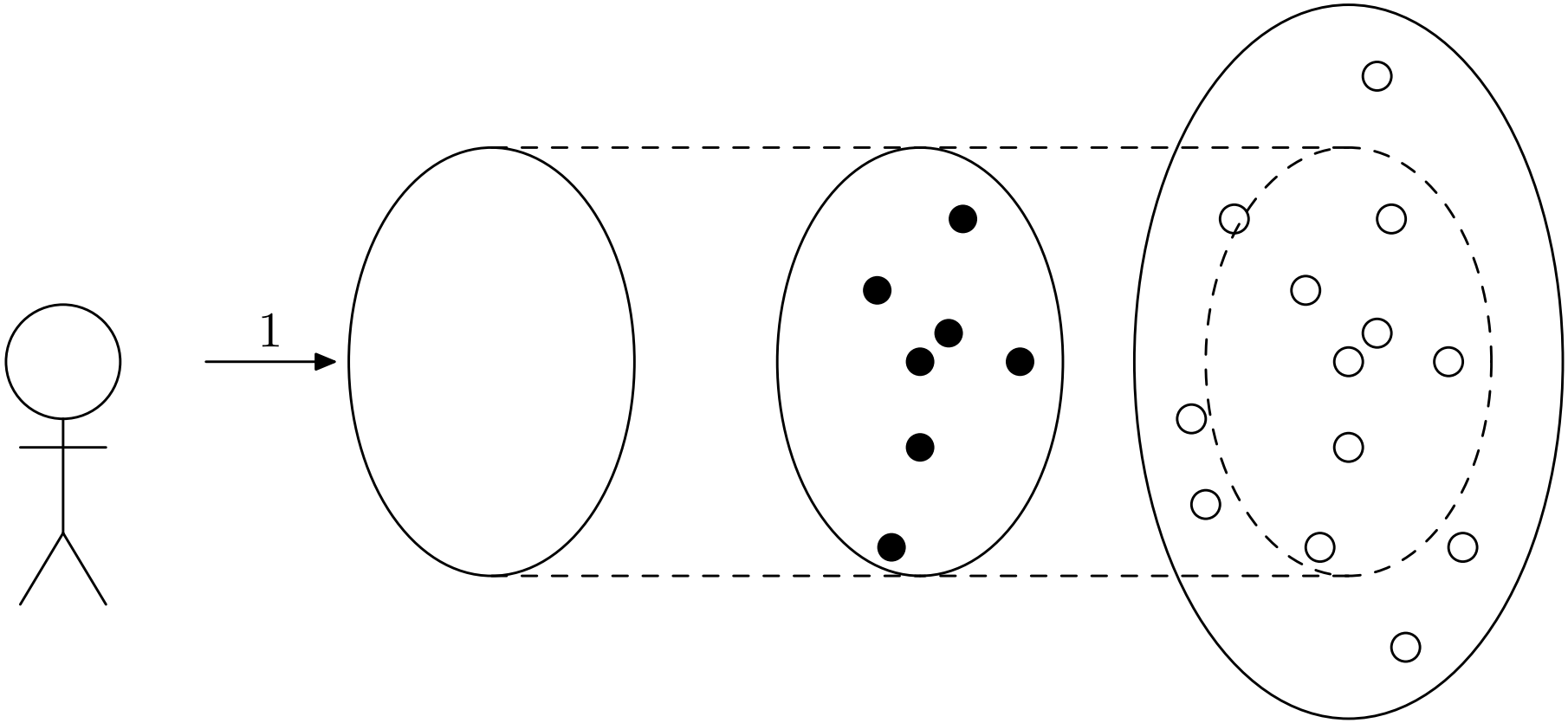


$\emptyset$



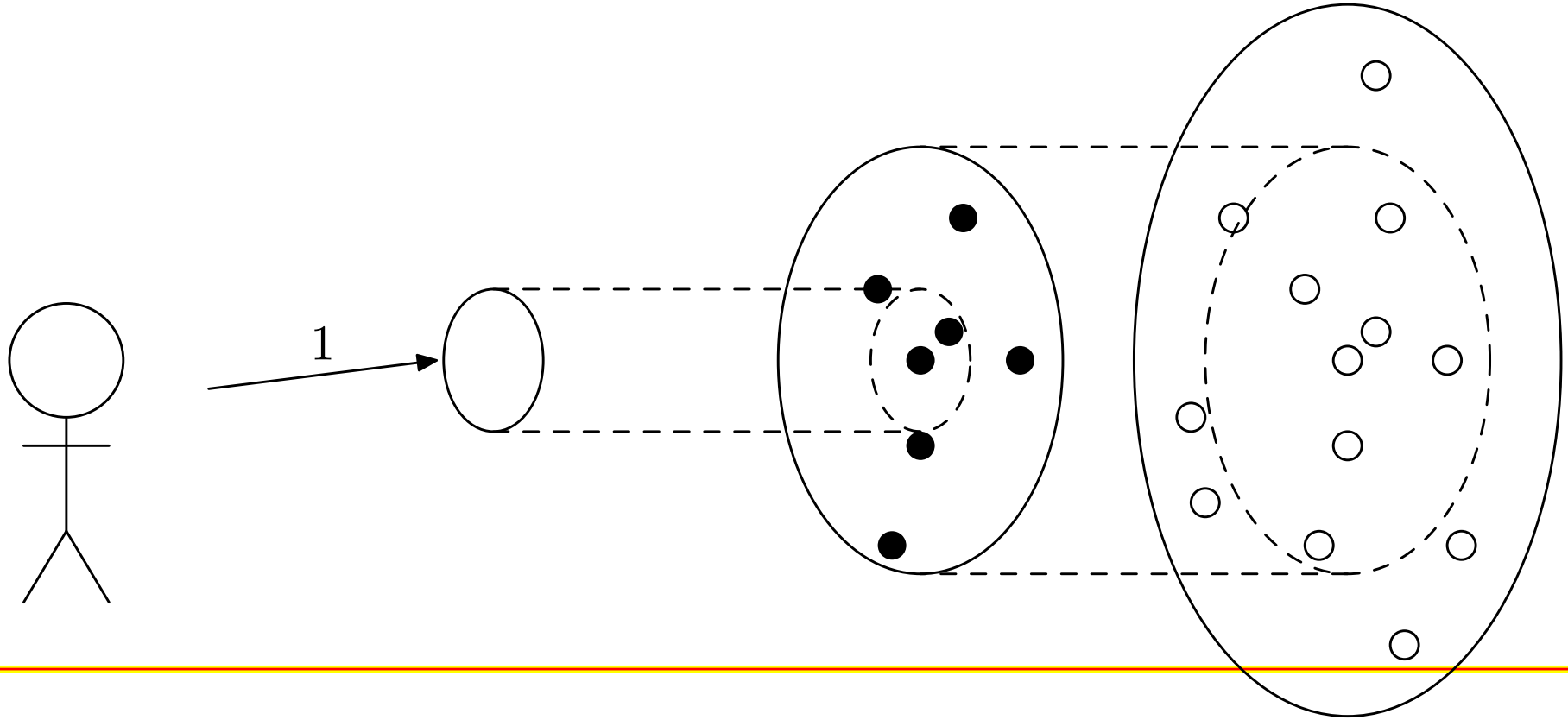
Absence

- All our belief is committed to the known set of alternatives.
- For example: “We do not know the age of John”.



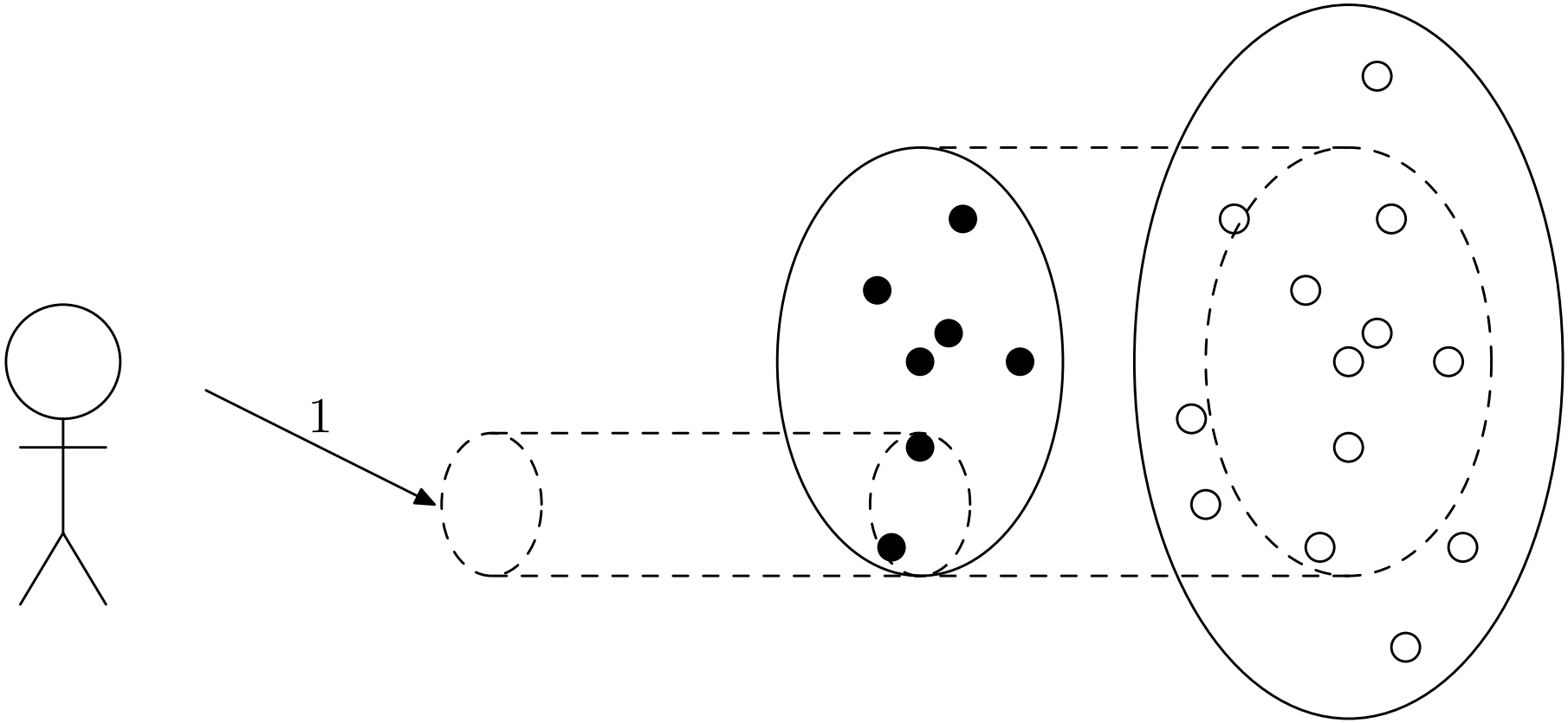
Imprecision: Non-specificity

- Imprecision concerns the cardinality of our believes.
- When we believe in a crisp set, imprecision is called non-specificity.
- For example: “John is between 170 and 180 cm. tall”.



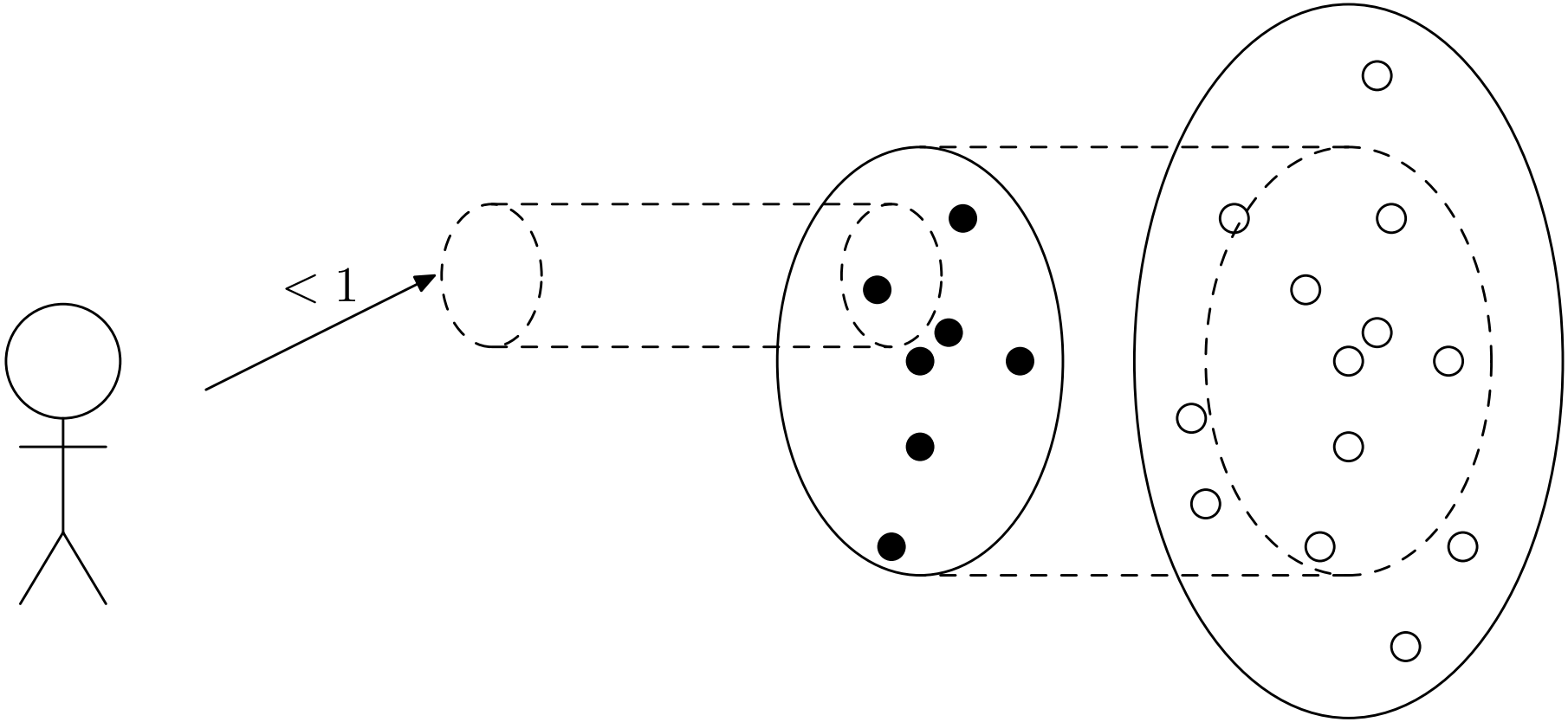
Imprecision: Vagueness

- If a set representing our belief is not crisp, imprecision is called vagueness.
- For example: “John is not very tall”



Uncertainty

- We have uncertainty when we do not commit all our belief.
- For example, “John should be 27 y.o.”.



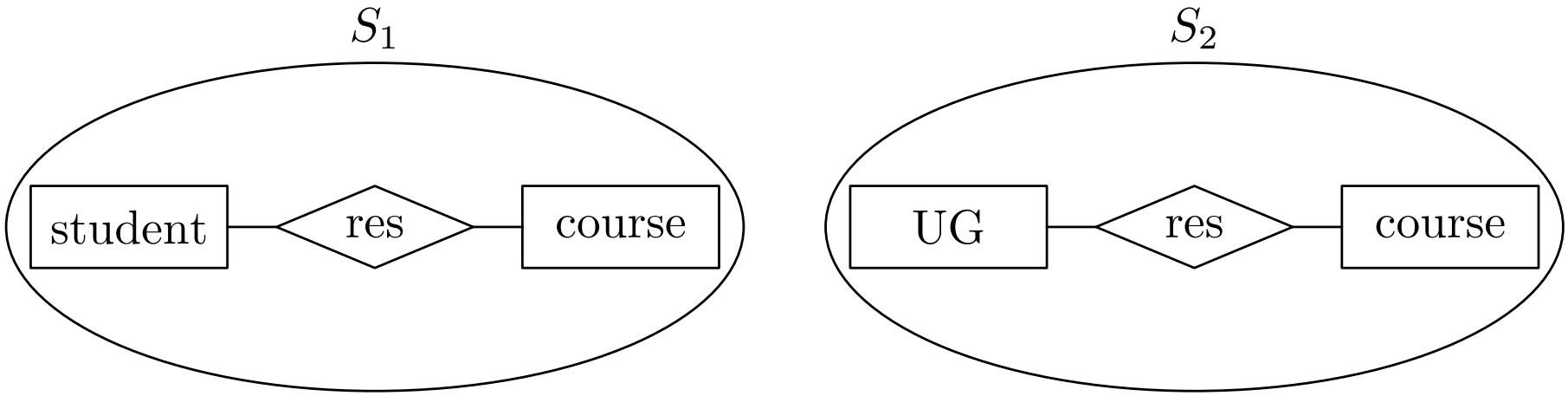
Examples

Type	**Abbr**	**Example** (John's tallness)
Absence	ABS	Not known.
Non-Specificity	NS	Between 180 and 190 cm. 183 or 187 cm.
Vagueness	VAG	Not very tall.
Uncertainty	UN	Perhaps, 183 cm.

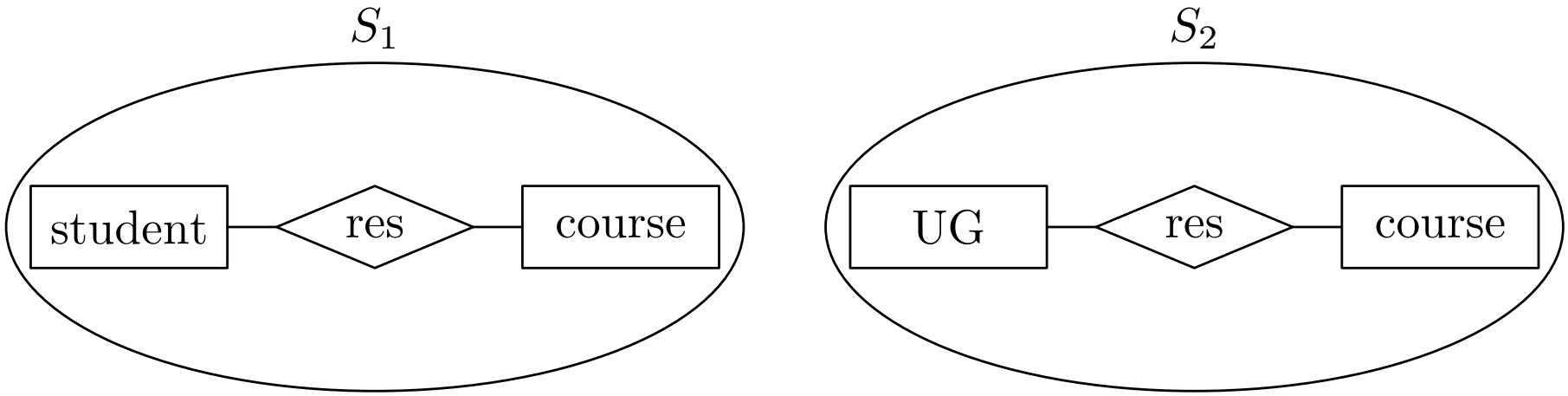
Uncertainty in Schema Integration based on Semantic Schema Matching

with Nikos Rizopoulos

Semantic Schema Integration



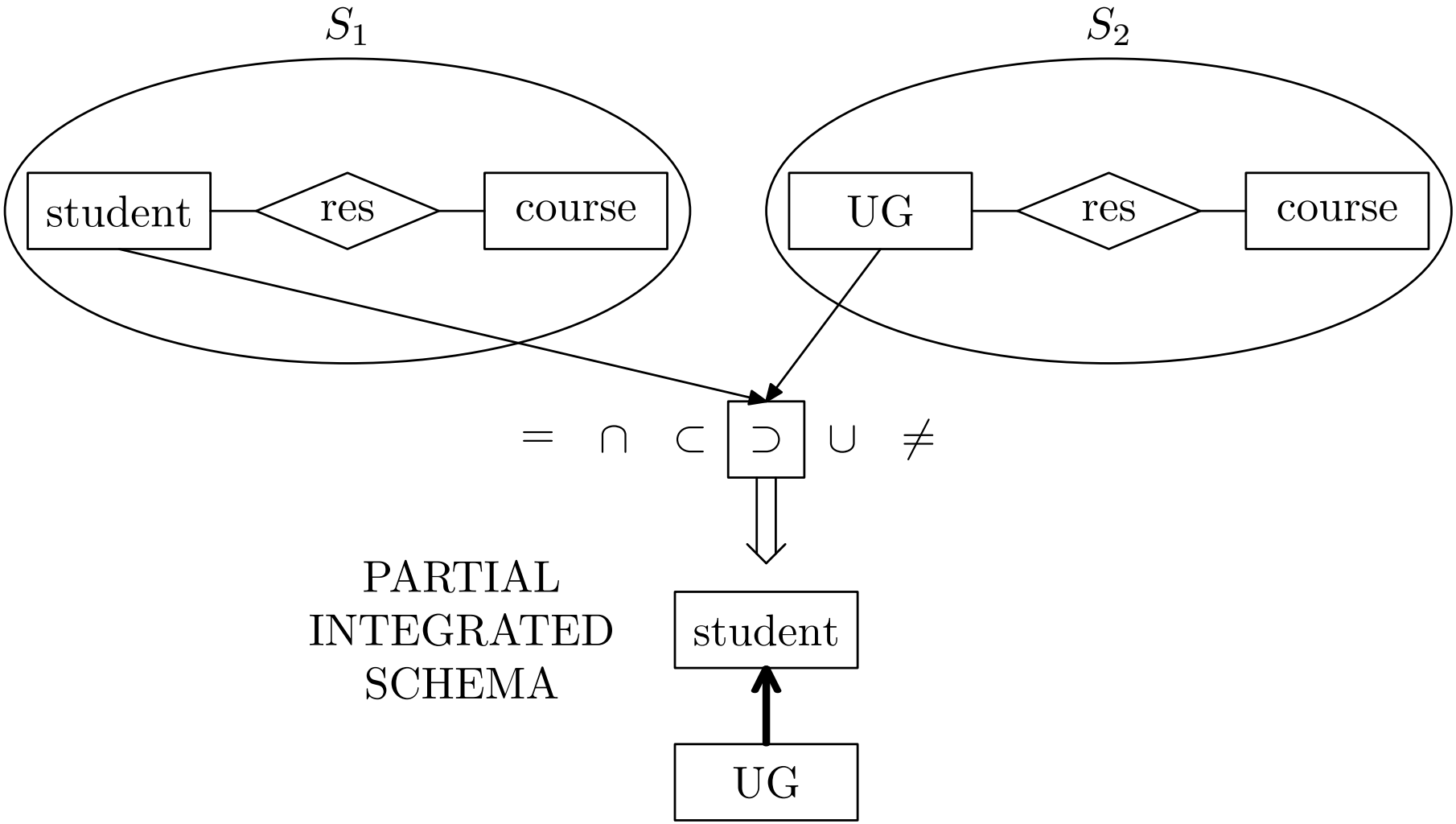
Semantic Schema Integration



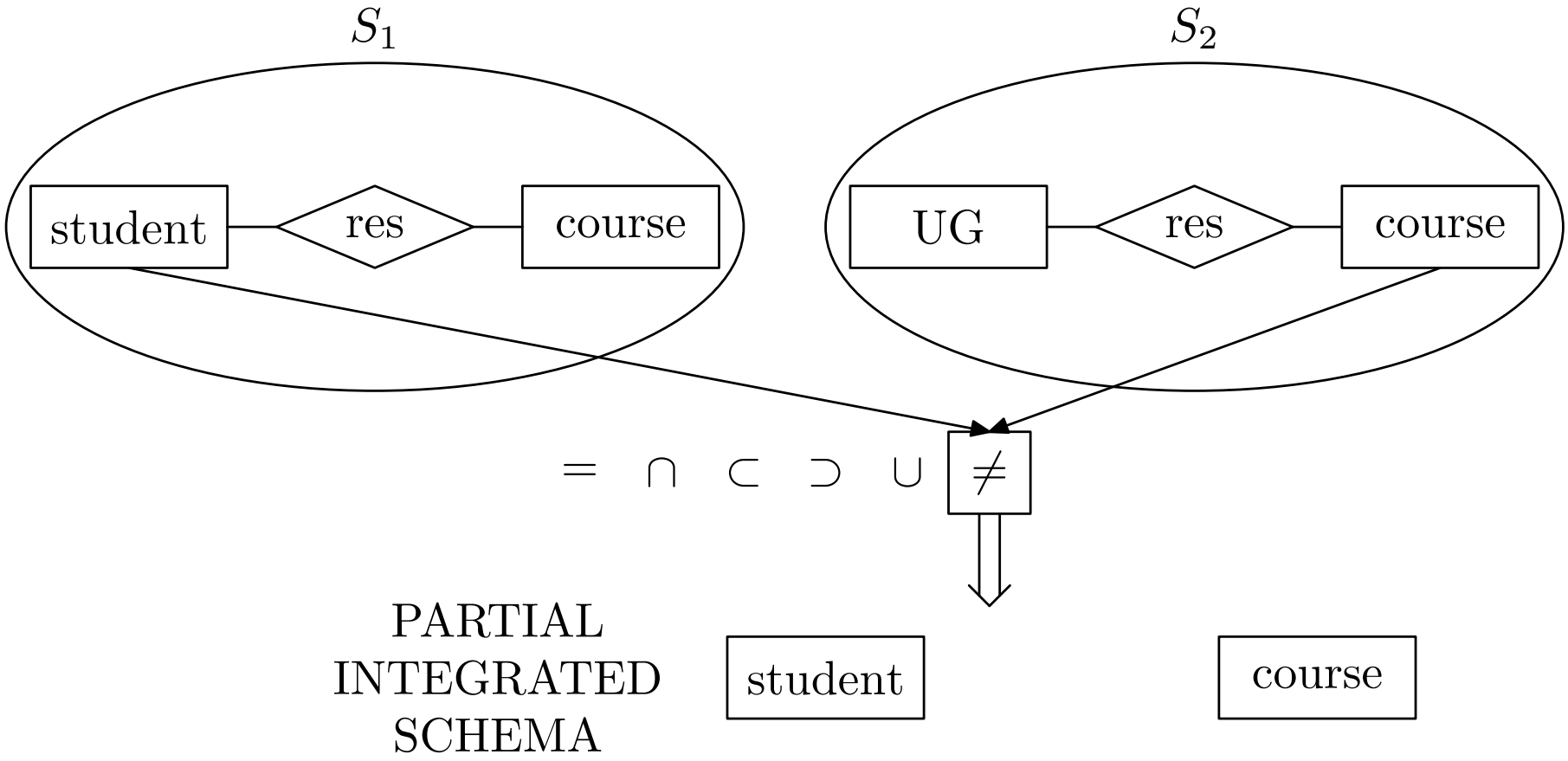
$$= \quad \cap \quad \subset \quad \supset \quad \cup \quad \neq$$

SEMANTIC
RELATIONSHIPS

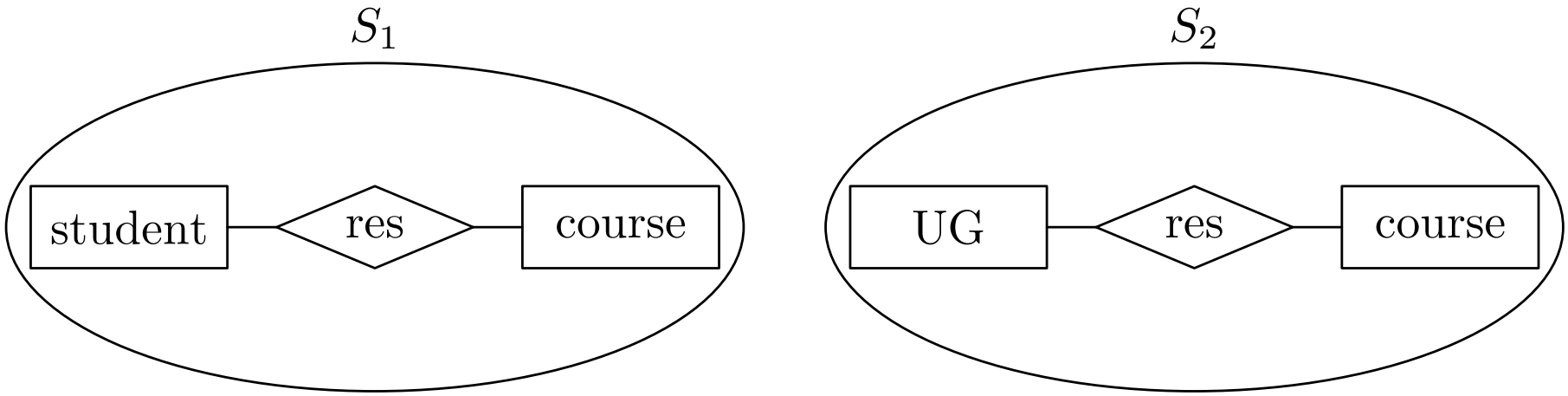
Semantic Schema Integration



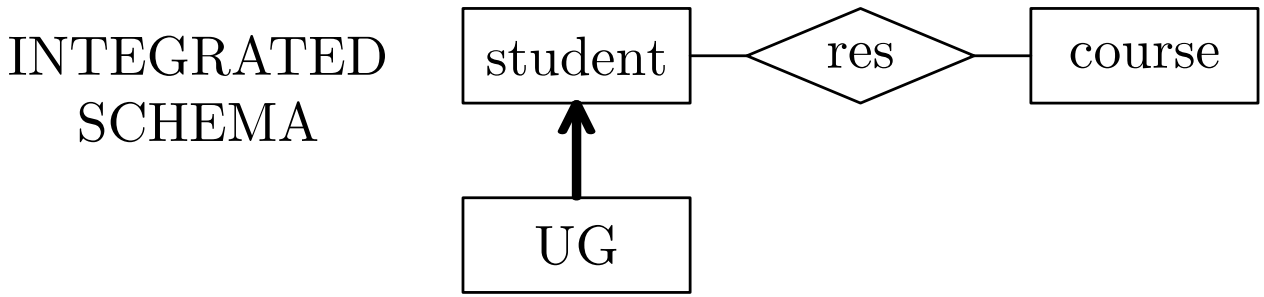
Semantic Schema Integration



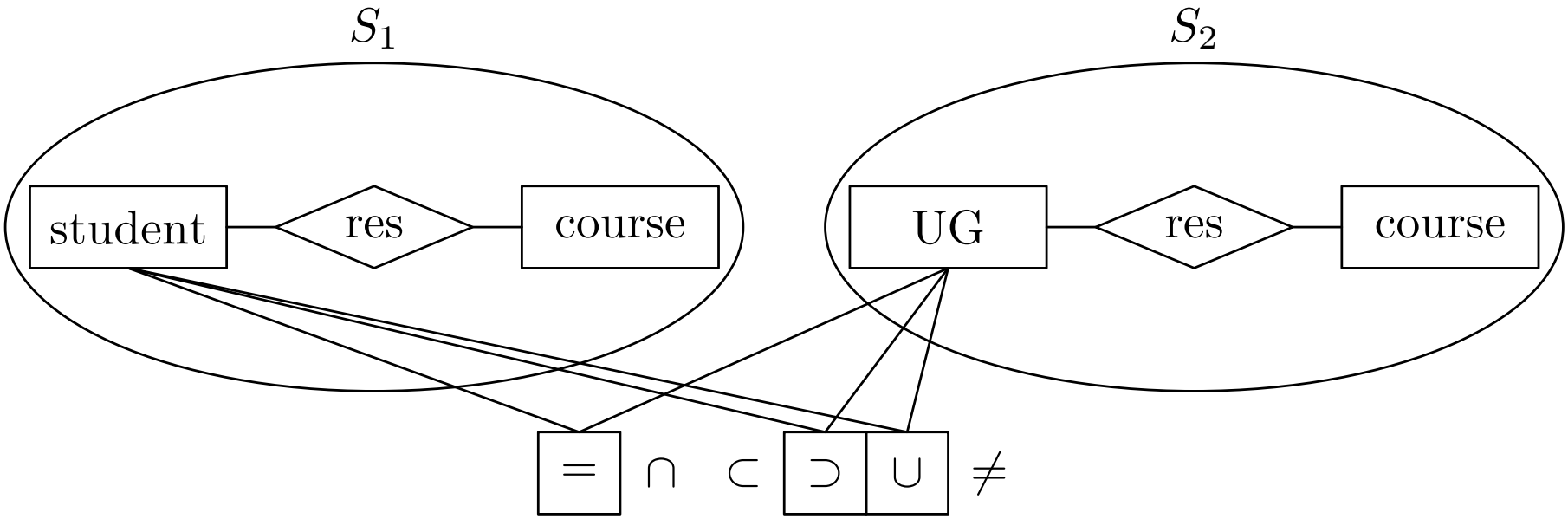
Semantic Schema Integration



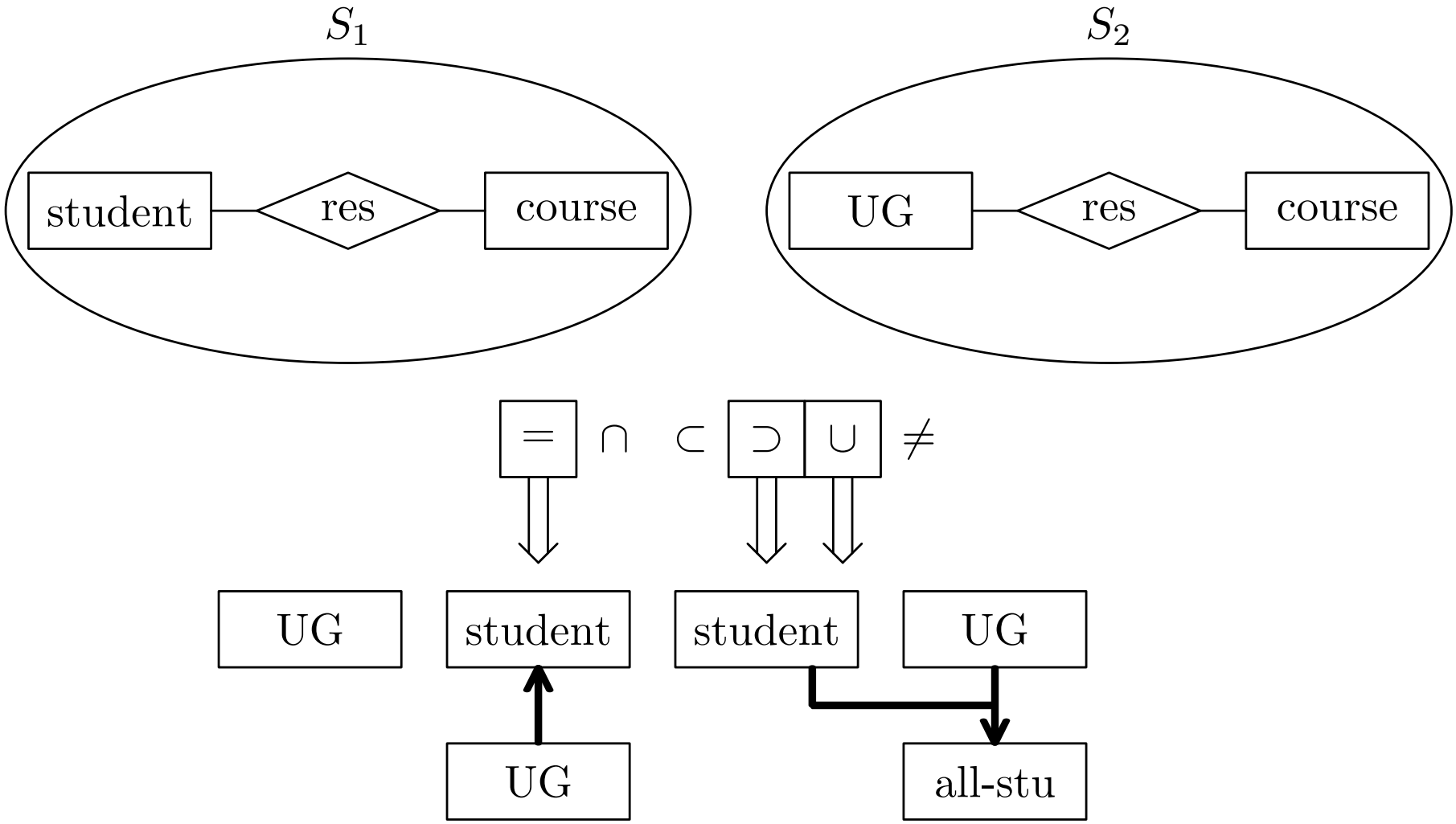
$$= \quad \cap \quad \subset \quad \supset \quad \cup \quad \neq$$



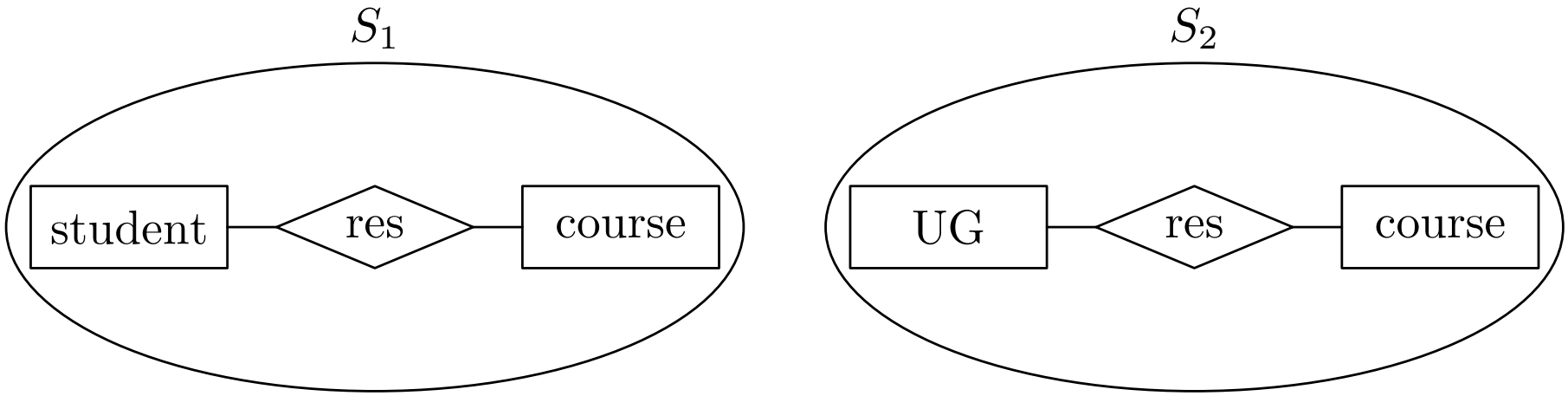
Dealing with uncertainty



Dealing with uncertainty



Dealing with uncertainty



?

Four main problems

- Identify **uncertain relationships**.

Four main problems

- Identify **uncertain relationships**.
- Produce **uncertain partial integrated schemas**.

Four main problems

- Identify **uncertain relationships**.
- Produce **uncertain partial integrated schemas**.
- Put together the uncertain partial integrated schemas, to obtain an **uncertain integrated schema**.

Four main problems

- Identify **uncertain relationships**.
- Produce **uncertain partial integrated schemas**.
- Put together the uncertain partial integrated schemas, to obtain an **uncertain integrated schema**.
- **Query** the database.

Uncertain relationships

- Dempster–Shafer's theory to represent believes in relationships.
- $\Theta = \{=, \cap, \subset, \supset, \cup, \neq\}$

Uncertain relationships

- Dempster–Shafer's theory to represent believes in relationships.
- $\Theta = \{=, \cap, \subset, \supset, \cup, \neq\}$
  - $m(\{=\}) = .4$
  - $m(\{\supset\}) = .3$
  - $m(\{\subset\}) = .3$

Uncertain relationships

- Dempster–Shafer's theory to represent believes in relationships.
- $\Theta = \{=, \cap, \subset, \supset, \cup, \neq\}$
  - $m(\{\cup, \neq\}) = 1$

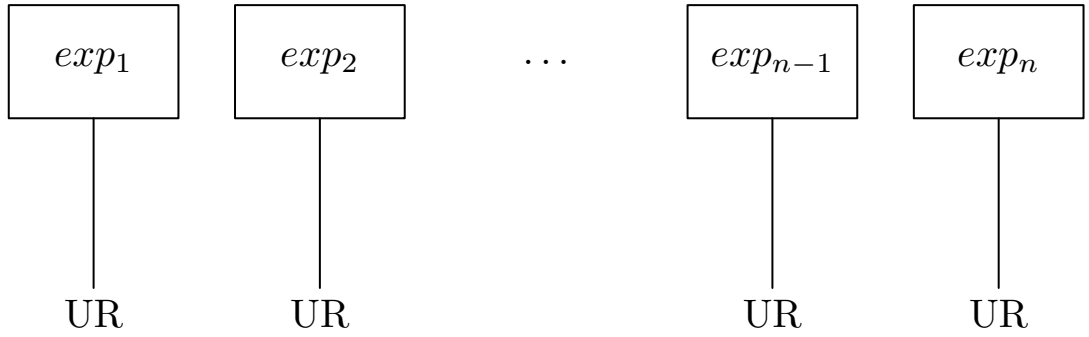
Uncertain relationships

- Dempster–Shafer's theory to represent believes in relationships.
- $\Theta = \{=, \cap, \subset, \supset, \cup, \neq\}$
  - $m(\{=\}) = .2$
  - $m(\Theta) = .8$

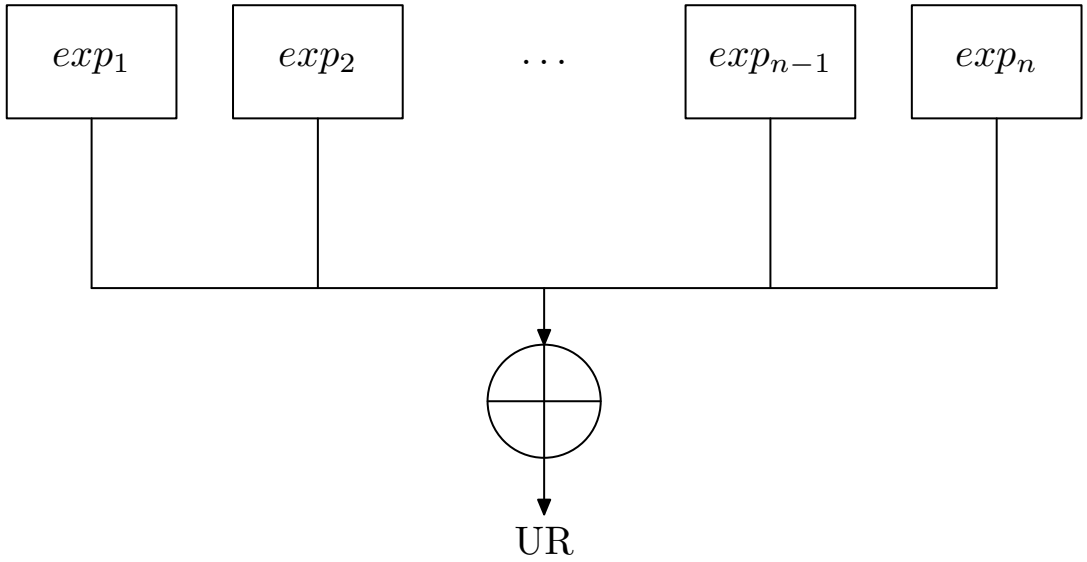
Uncertain relationships

- Dempster–Shafer's theory to represent believes in relationships.
- $\Theta = \{=, \cap, \subset, \supset, \cup, \neq\}$
  - $m(\Theta) = 1$

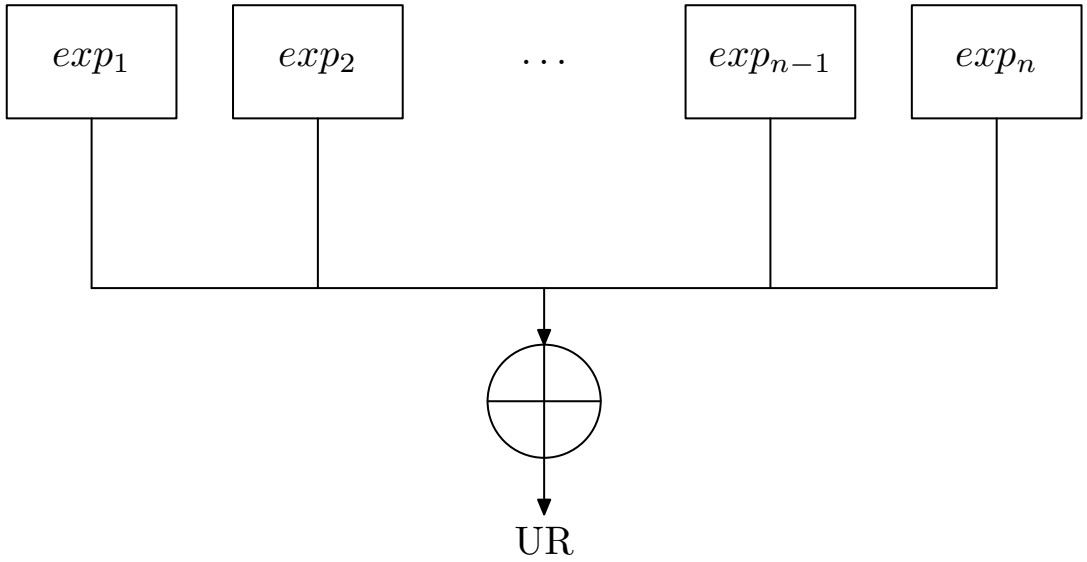
Architecture



Architecture

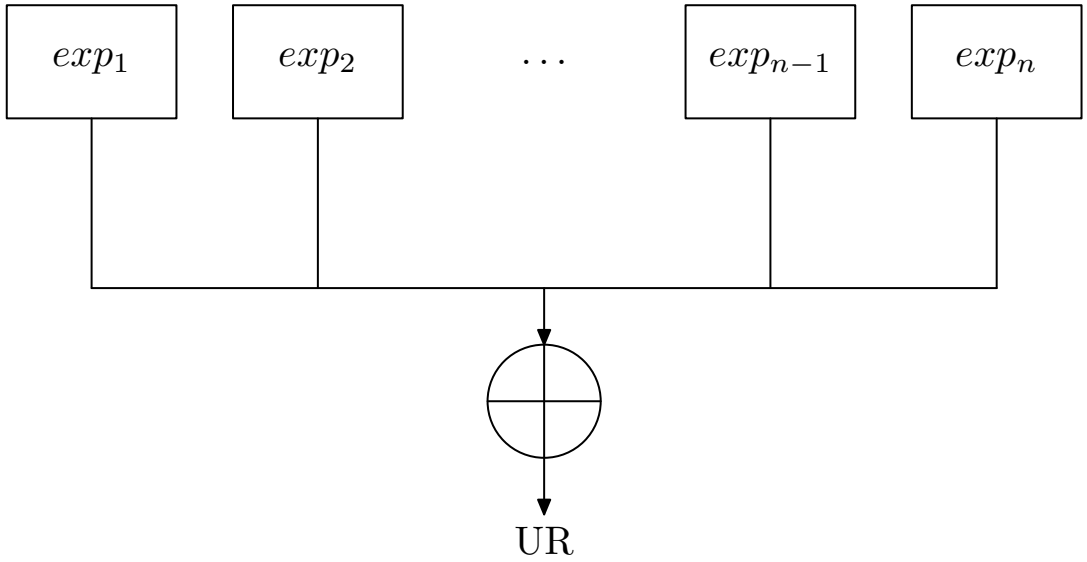


Architecture



- Scalable.

Architecture



- Scalable.
- Experts can be software agents or humans.

Example

Exp$_1$ (Cardinality):

- $m(\{\cap, \supset, \cup, \neq\}) = 1$

Example

Exp$_1$ (Cardinality):

- $m(\{\cap, \supset, \cup, \neq\}) = 1$

Exp$_2$ (Thesaurus):

- $m(\{\supset\}) = .5$
- $m(\{=\}) = .2$
- $m(\Theta) = .3$

Example

Exp$_1$ (Cardinality):

- $m(\{\cap, \supset, \cup, \neq\}) = 1$

Exp$_2$ (Thesaurus):

- $m(\{\supset\}) = .5$
- $m(\{=\}) = .2$
- $m(\Theta) = .3$

Exp$_3$ (Human):

- $m(\{=, \supset, \cup\}) = 1$

Example

Exp$_1$ (Cardinality):

- $m(\{\cap, \supset, \cup, \neq\}) = 1$

Exp$_2$ (Thesaurus):

- $m(\{\supset\}) = .5$
- $m(\{=\}) = .2$
- $m(\Theta) = .3$

Exp$_3$ (Human):

- $m(\{=, \supset, \cup\}) = 1$

$\oplus$ (Combination):

- $m(\{\supset\}) = \frac{5}{8}$
- $m(\{\supset, \cup\}) = \frac{3}{8}$

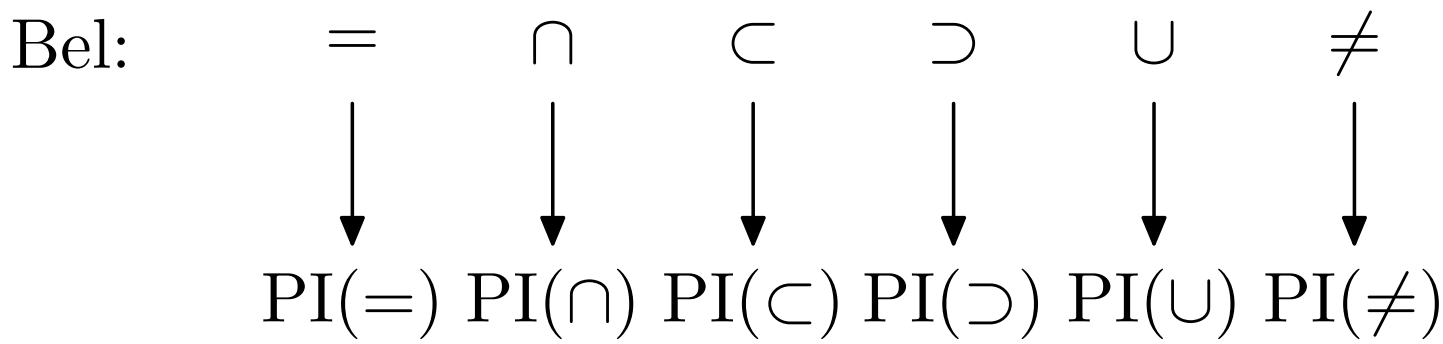
Unc. partial integrated schemas

- Given two objects, some belief is committed to each possible relationship between them.

$$\text{Bel:} \qquad = \qquad \cap \qquad \subset \qquad \supset \qquad \cup \qquad \neq$$

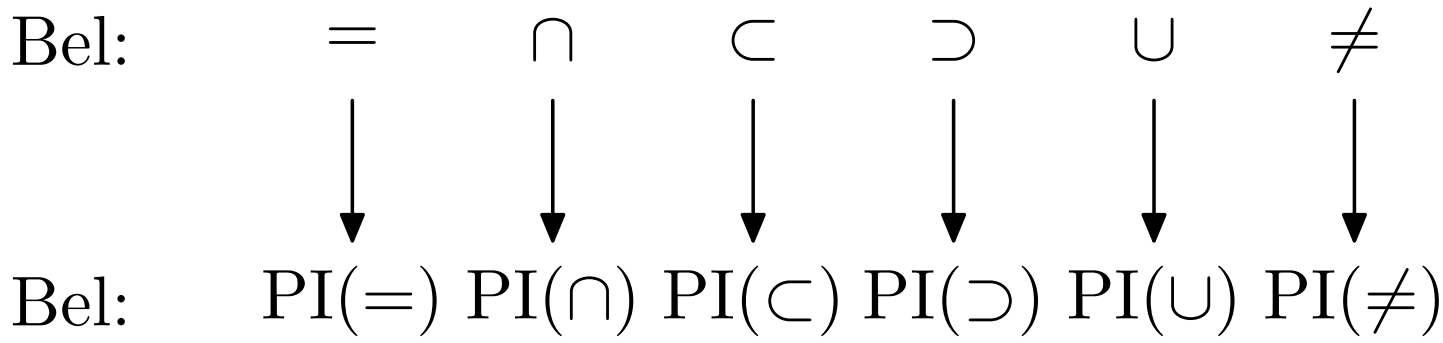
Unc. partial integrated schemas

- Given two objects, some belief is committed to each possible relationship between them.
- From each possible relationship we can obtain a partial integrated schema.



Unc. partial integrated schemas

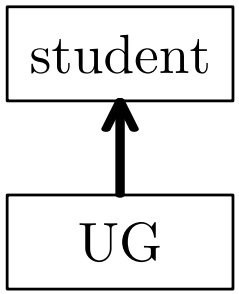
- Given two objects, some belief is committed to each possible relationship between them.
- From each possible relationship we can obtain a partial integrated schema.
- The belief committed to each partial integrated schema is the same previously committed to the corresponding possible relationship.



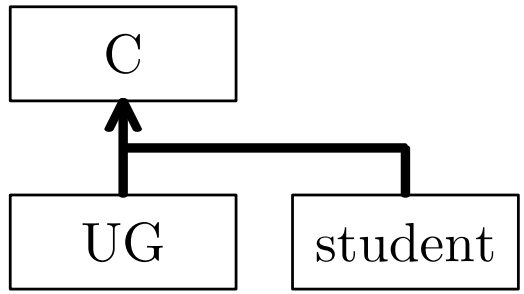
Example

- Uncertain relationship (Student-UG):
  - $m(\{\supset\}) = \frac{5}{8}$
  - $m(\{\supset, \cup\}) = \frac{3}{8}$
- Uncertain partial integrated schema:
  - $m(\{\mathrm{PI}(\supset)\}) = \frac{5}{8}$
  - $m(\{\mathrm{PI}(\supset), \mathrm{PI}(\cup)\}) = \frac{3}{8}$

$\mathrm{PI}(\supset)$



$\mathrm{PI}(\cup)$



Uncertain integrated schemas

Main idea: take all possible combinations of uncertain partial integrated schemas.

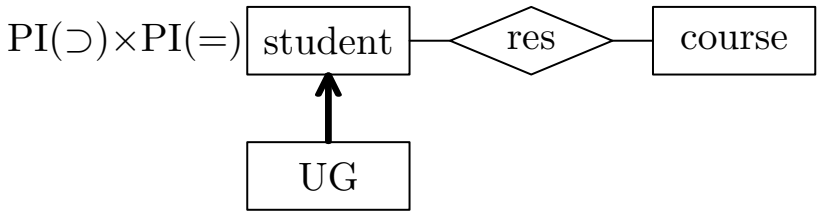
- $= (A, B), \neq (A, B)$
- $= (B, C), \neq (B, C)$

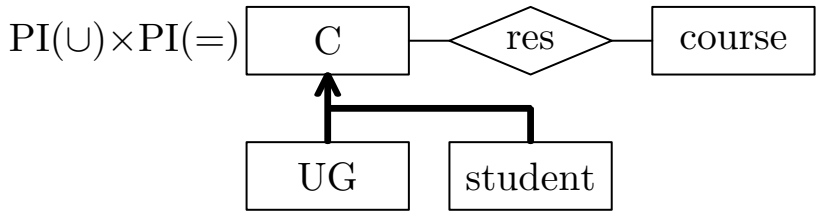
A, B	B, C
$=$	$=$
$=$	$\neq$
$\neq$	$=$
$\neq$	$\neq$

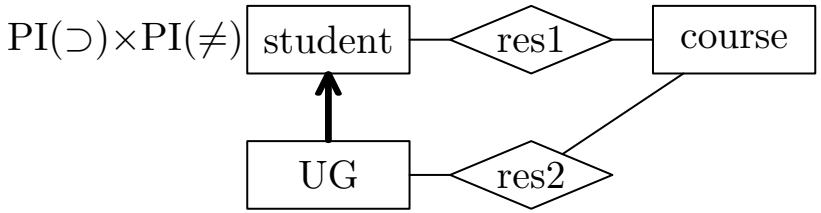
Some issues:

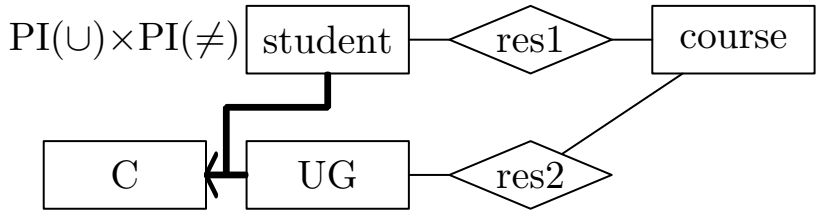
- Compact representation.
- Dependencies.

Uncertain integrated schemas



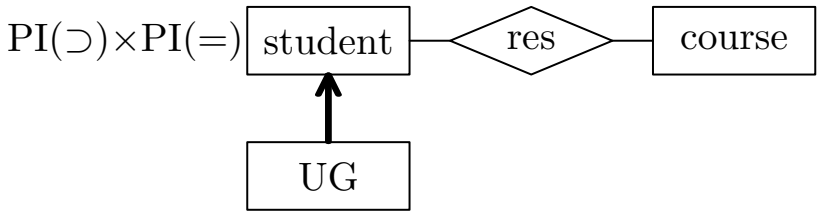


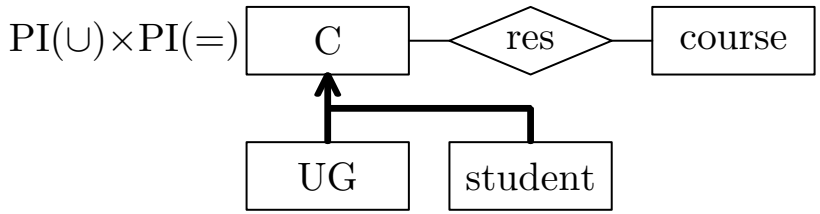


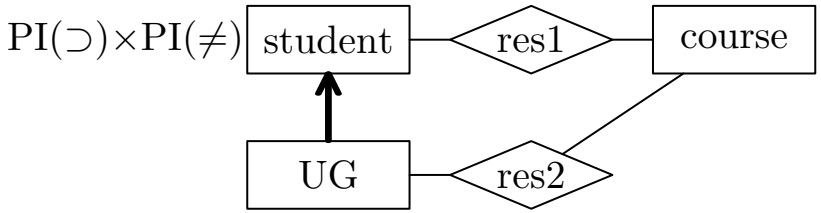


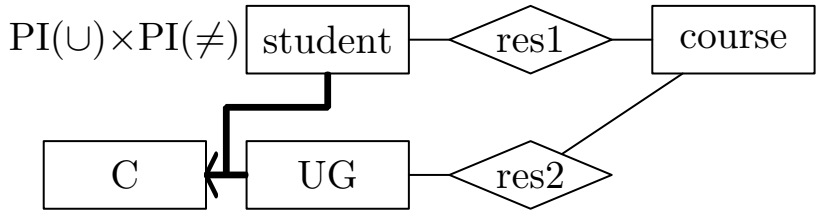
- Uncertain relationship (Student–UG):
  - $m(\{\supset\}) = \frac{5}{8}$, $m(\{\supset, \cup\}) = \frac{3}{8}$
- Uncertain relationship (res–res):
  - $m(\{=\}) = \frac{1}{3}$, $m(\{\neq\}) = \frac{2}{3}$

Uncertain integrated schemas









- Uncertain partial integrated schema:
  - $m(\{\mathsf{PI}(\supset)\} \times \{\mathsf{PI}(=)\}) = \frac{5}{24}$
  - $m(\{\mathsf{PI}(\supset),\mathsf{PI}(\cup)\} \times \{\mathsf{PI}(=)\}) = \frac{3}{24}$
  - $m(\{\mathsf{PI}(\supset)\} \times \{\mathsf{PI}(\neq)\}) = \frac{5}{12}$
  - $m(\{\mathsf{PI}(\supset),\mathsf{PI}(\cup)\} \times \{\mathsf{PI}(\neq)\}) = \frac{3}{12}$

Uncertain integrated schemas

- $= (A, B), \neq (A, B)$
- $= (B, C), \neq (B, C)$
- $= (A, C), \neq (A, C)$

Uncertain integrated schemas

- $= (A, B), \neq (A, B)$
- $= (B, C), \neq (B, C)$
- $= (A, C), \neq (A, C)$

A, B	B, C	A, C	Allowed
$=$	$=$	$=$	
$=$	$=$	$\neq$	
$=$	$\neq$	$=$	
$\neq$	$=$	$=$	
$\neq$	$=$	$\neq$	
$=$	$\neq$	$\neq$	
$\neq$	$\neq$	$=$	
$\neq$	$\neq$	$\neq$	

Uncertain integrated schemas

- $= (A, B), \neq (A, B)$
- $= (B, C), \neq (B, C)$
- $= (A, C), \neq (A, C)$

A, B	B, C	A, C	Allowed
$=$	$=$	$=$	Y
$=$	$=$	$\neq$	N
$=$	$\neq$	$=$	N
$\neq$	$=$	$=$	N
$\neq$	$=$	$\neq$	Y
$=$	$\neq$	$\neq$	Y
$\neq$	$\neq$	$=$	Y
$\neq$	$\neq$	$\neq$	Y

Querying the database

- No idea. . .
- It should not be very difficult to define the semantics of a query.
- Efficiency problems.
  - Cardinality reduction.
  - Compact query plans.

Concluding remarks

- Uncertainty is one of many possible types of imperfection/ignorance.
- We start our investigation from a method of schema integration based on semantic schema matching.
- In real cases of data integration, it can be difficult to identify certain semantic relationships.
- We have presented some preliminary ideas on how to extend this method to deal with uncertainty.

Discussion

Dealing with Imperfection in Schema Integration

Some preliminary ideas

Matteo Magnani

Dept. of Computer Science
University of Bologna
Via Mura A.Zamboni, 7
40127 Bologna (Italy)

Nikos Rizopoulos

Dept. of Computing
180 Queen's Gate
South Kensington Campus
Imperial College London
SW7 2AZ London (UK)