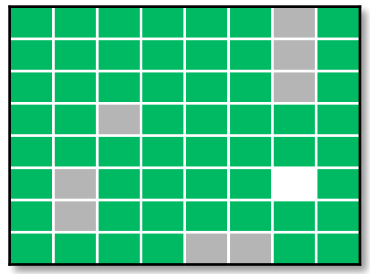
 

Completeness of Queries over Incomplete Databases

Werner Nutt

joint work with Marco Montali, Sergey Paramonov, Simon Razniewski, Ognjen Savkovic, Alex Tomasi, Fariz Darari

(VLDB‘11, CIKM’12, BPM‘13, ISWC‘13)

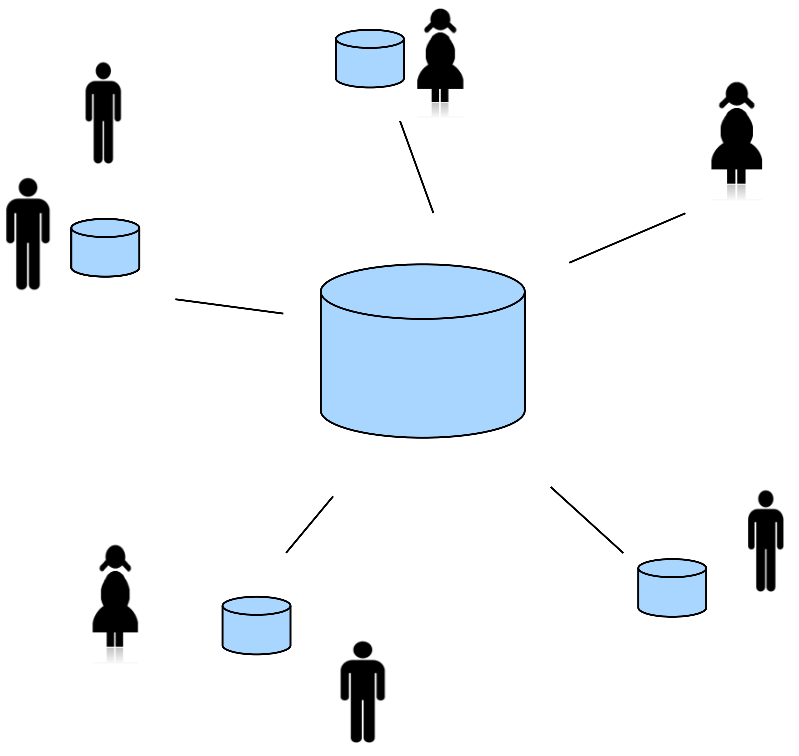
Background

Incompleteness is omnipresent in data management

- Null values in relational databases: Codd 1975
- Representation systems: Imielinski/Lipski
  - 1984 Focus on certain/possible answers
- Query completeness over incomplete databases: little attention

School Data Management in Bolzano

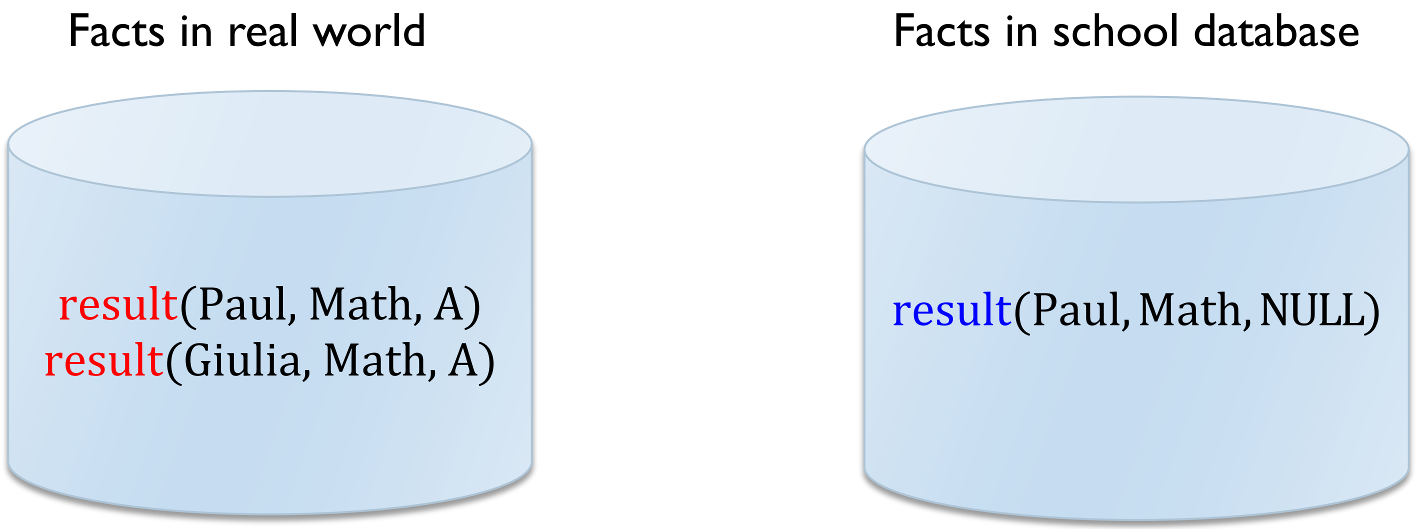
decentrally maintained database („Popcorn“)



generally incomplete

require complete data

Incompleteness in the School Data



Missing information in the school database:

- no entry for Giulia (missing record)
- no grade for Paul (missing value)

Consequence: Query Answers are Incorrect

Query Q: *"How many pupils have grade A in Math?"*

In the real world: Q(result(Paul, Math, A) result(Giulia, Math, A)) = 2

According to available database: Q(result(Paul, Math, NULL)) = 0

→ If data is incomplete, query answers become incorrect.

Why are Data About Pupils Incomplete?

- Data have not yet been copied from the local school database
  to the central database
- The copying procedure has been aborted
- Pupils have been already registered/
  classes have been formed,
  but pupils have not yet been entered into the database
- Some schools (e.g. vocational schools)
  administer student grades with Popcorn,
  others not
- School careers of immigrants are often not captured

But: Data are Partially Complete

- Grades of students at vocational schools are complete …

Business rules

- Grades of students at vocational schools are c
  … after reports have been h

Stadium of a business process

Busin… …esses

- Classes at school X are complete,
  when the classes have been formed
  … and entered into Popcorn

How can we use information about partial completeness? Meta data!

Use Metadata to Guarantee Completeness!

... vocational schools use
the information system of the province
to manage grades

db are complete, e.g.,

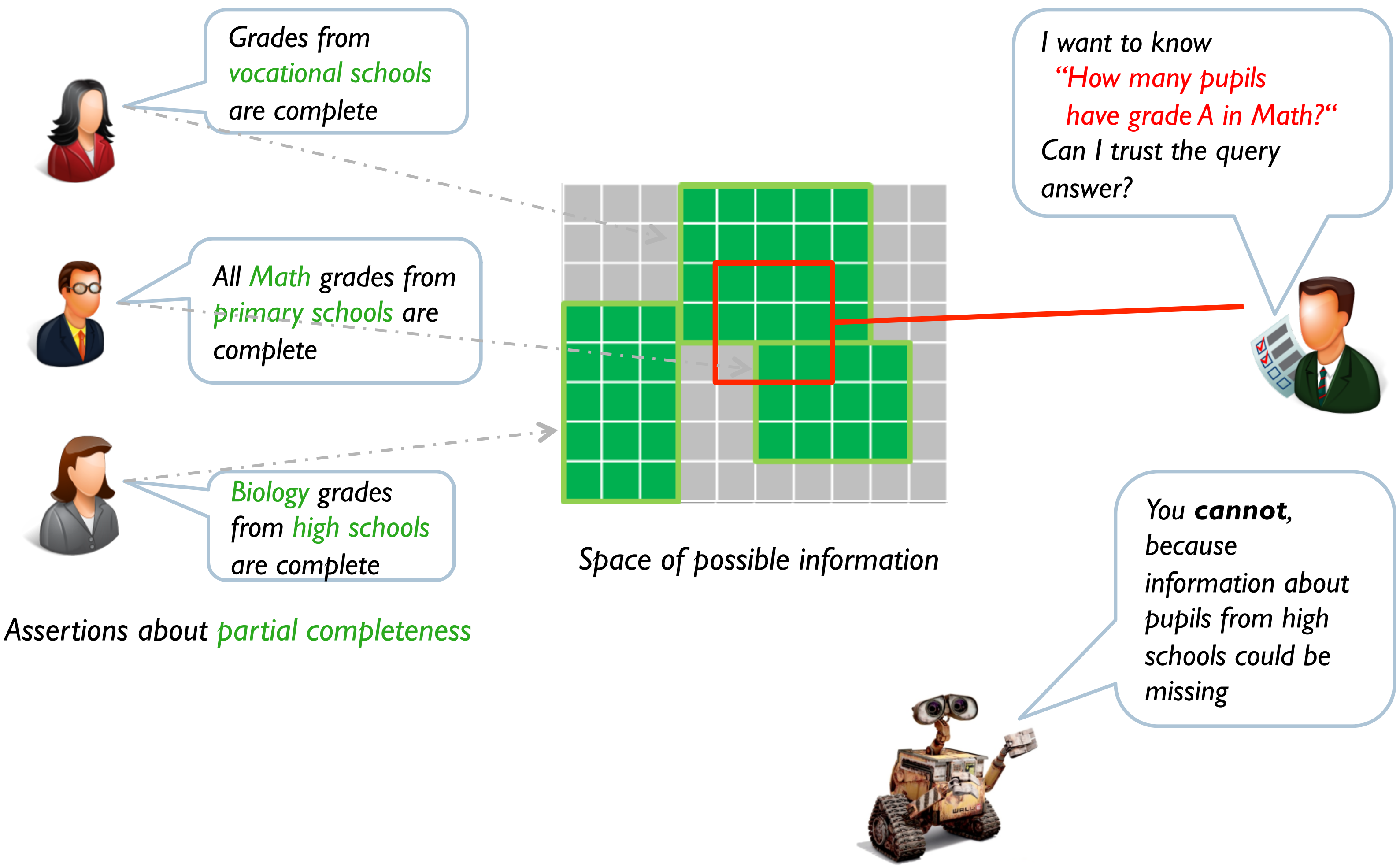
- *"The grades from vocational schools are complete"*

- *"The Math grades from primary schools are complete"*

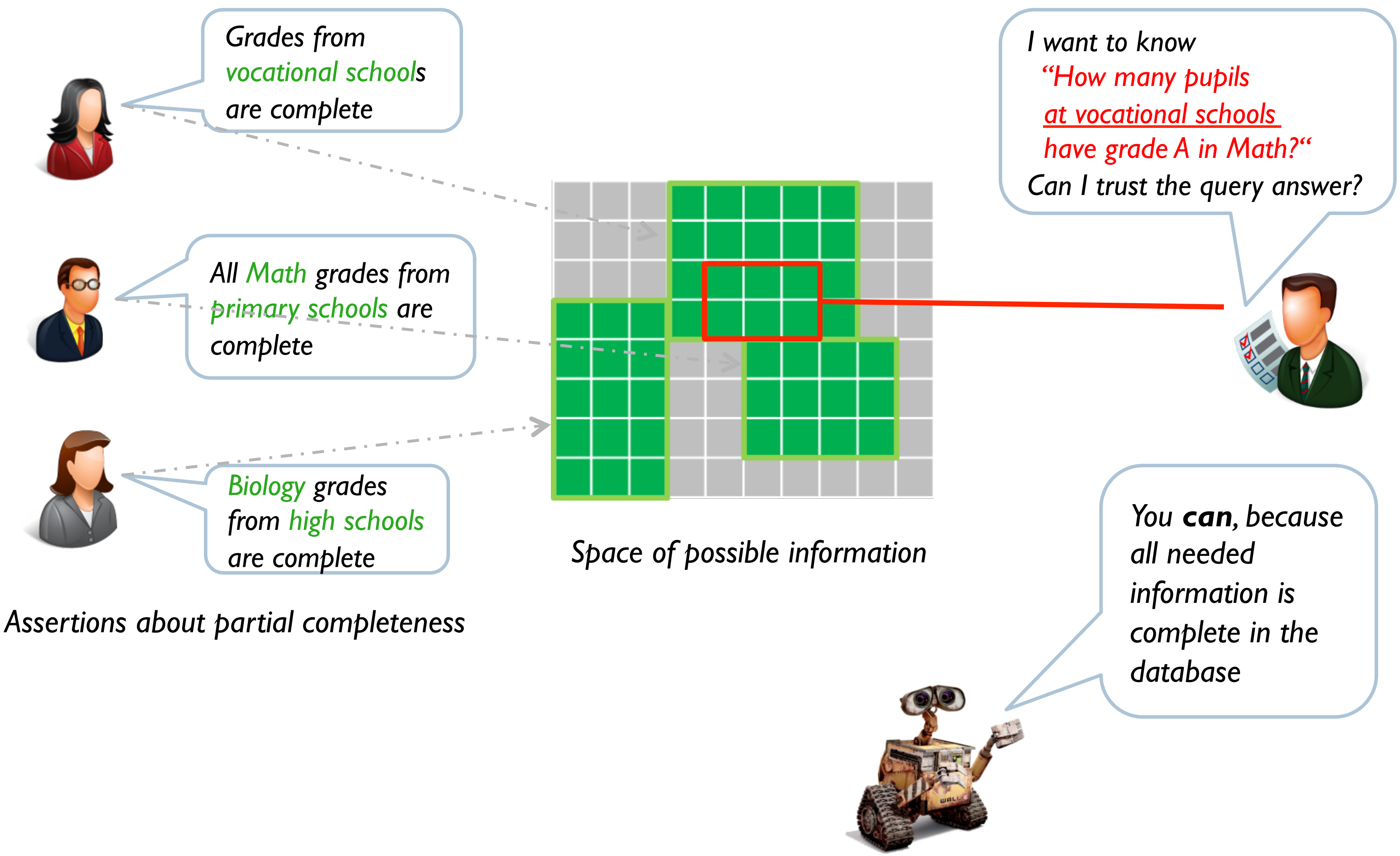
... primary schools
took part in a survey
of Math education

→ Idea: Assess completeness of a query
using completeness assertions for (parts of) tables

Reasoning about Query Completeness

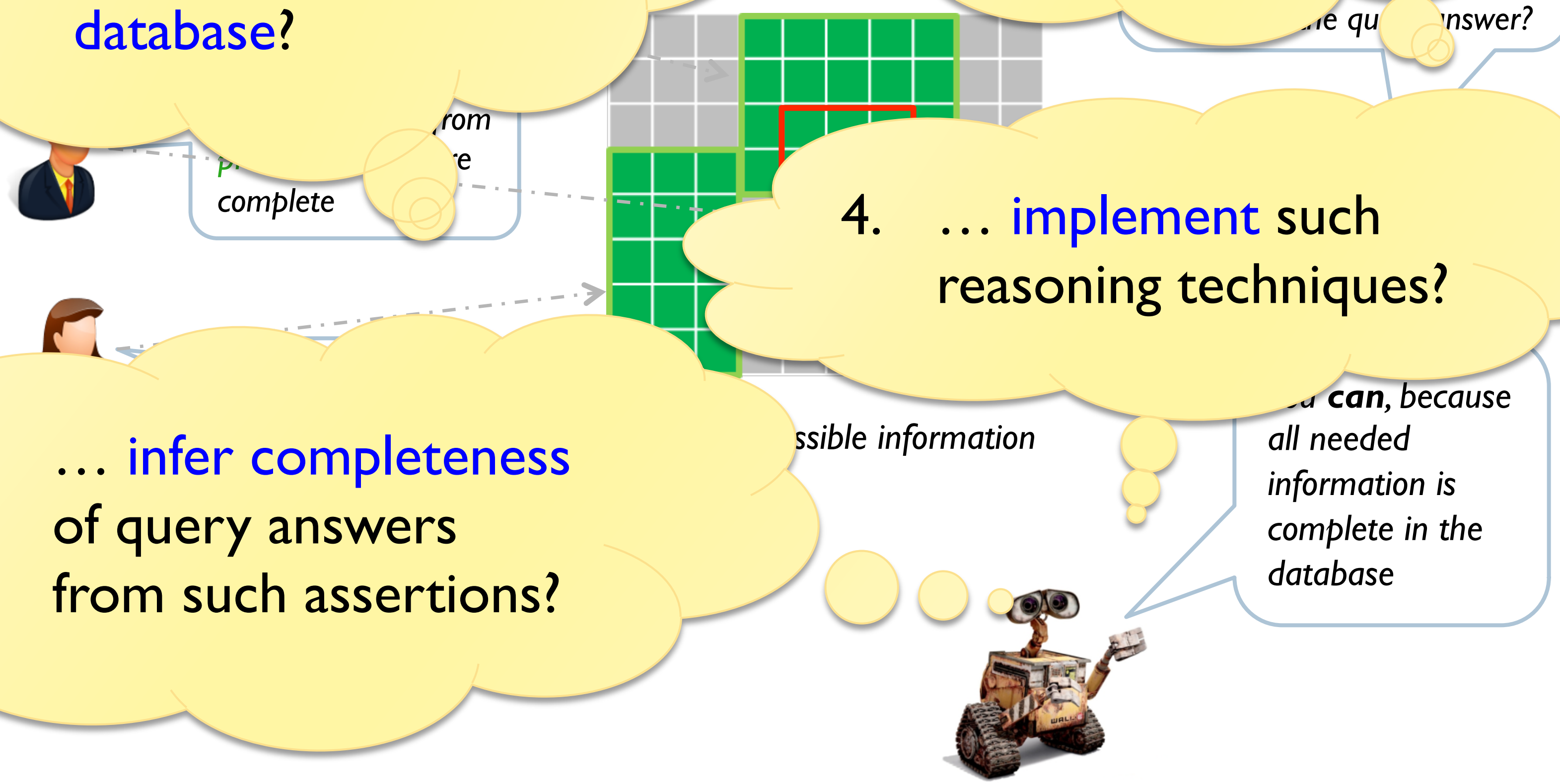


Reasoning about Query Completeness (2)



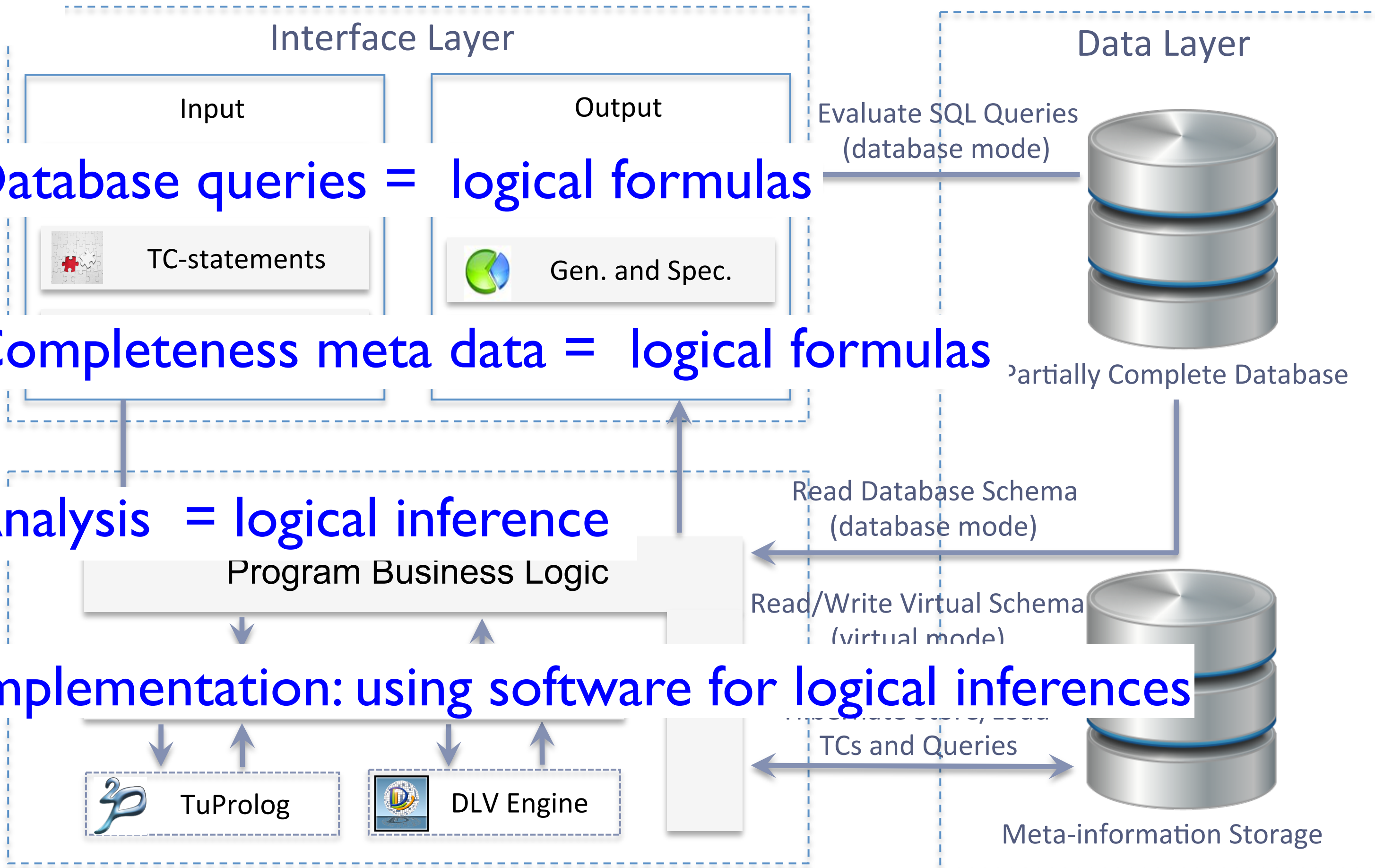
Research Questions: How can one …

1. … formalize completeness of query answers?
2. … assert completeness of parts of a possibly incomplete database?
3. … infer completeness of query answers from such assertions?
4. … implement such reasoning techniques?



MAGIK (= Managing Incomplete Knowledge)

- Database queries = logical formulas
- Completeness meta data = logical formulas
- Analysis = logical inference
- Implementation: using software for logical inferences



Running Example: Schema

result(name, subject, grade)

pupil(name, age, schoolName, schoolType)

Notation: Databases

Database instances are sets of ground atoms, e.g.,

$$D = \{ \text{result(Paul, Math, NULL)}, \\ \text{result(Giulia, Math, A)}, \\ \text{pupil(Paul, 17, Verdi, Voc)} \},$$

possibly containing NULLs.

Notation: Conjunctive Queries

A single block SQL queries, possibly with DISTINCT,

```
SELECT r.grade
FROM   result r, pupil p
WHERE  r.name = p.name AND
       r.subject = 'Math' AND
       p.age <= 11
```

is expressed as a conjunctive query (CQ), using a Datalog rule:

$$Q(g) \text{ :- } result(n, Math, g), pupil(n, a, sn, st), a \leq 11$$

Notation: Conjunctive Queries (2)

$Q(\underline{x})$:- $L(\underline{x}, \underline{y}), M$

- $L(\underline{x}, \underline{y})$ conjunction of relational atoms
- M conjunction of comparisons
- $\underline{x}$ vector of *distinguished* (= output) variable
- $\underline{y}$ vector of *non-distinguished* (= existen

As a default, we assume set semantics

Query answers (under set semantics):

$$Q(D) = \{\alpha\underline{x} \mid \alpha L \subseteq D,\ \alpha \models M\}$$

Bag semantics: each α contributes a copy of $\alpha\underline{x}$

Possible Completeness Statements

"We get complete answers to the following queries:

- Which pupils have grade A in Math?
- Which pupils from vocational schools have grade A in Math?

Query Completeness Statements

"The database contains

- all subjects and grades of pupils from vocational schools
- all subjects studied by pupils from vocational schools "

Table Completeness Statements

Formalization: Incomplete Database

[Motro 1989]

When talking about incompleteness, we need a complete reference

An *incomplete database* D is a pair of

an ideal database D^i and

an available database D^a

$$D = (D^i, D^a)$$

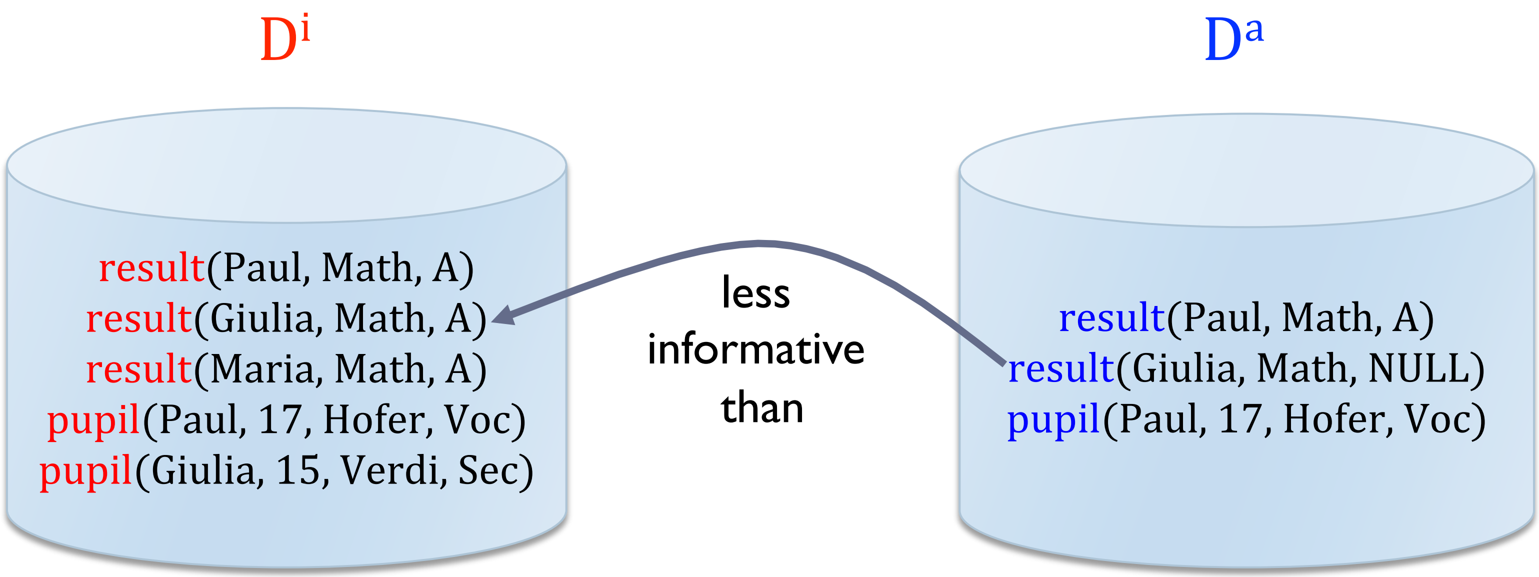
such that

for each record in D^a there is
a "more informative" record in D^i

For databases w/o Nulls, this means

$$D^a \subseteq D^i$$

Example: An Incomplete Database



Formalization: Query Completeness

[Motro 1989]

Query Q

"The answer to Q is complete"

Notation: $\text{Compl}(Q)$

To be precise, we have to distinguish between set and bag semantics

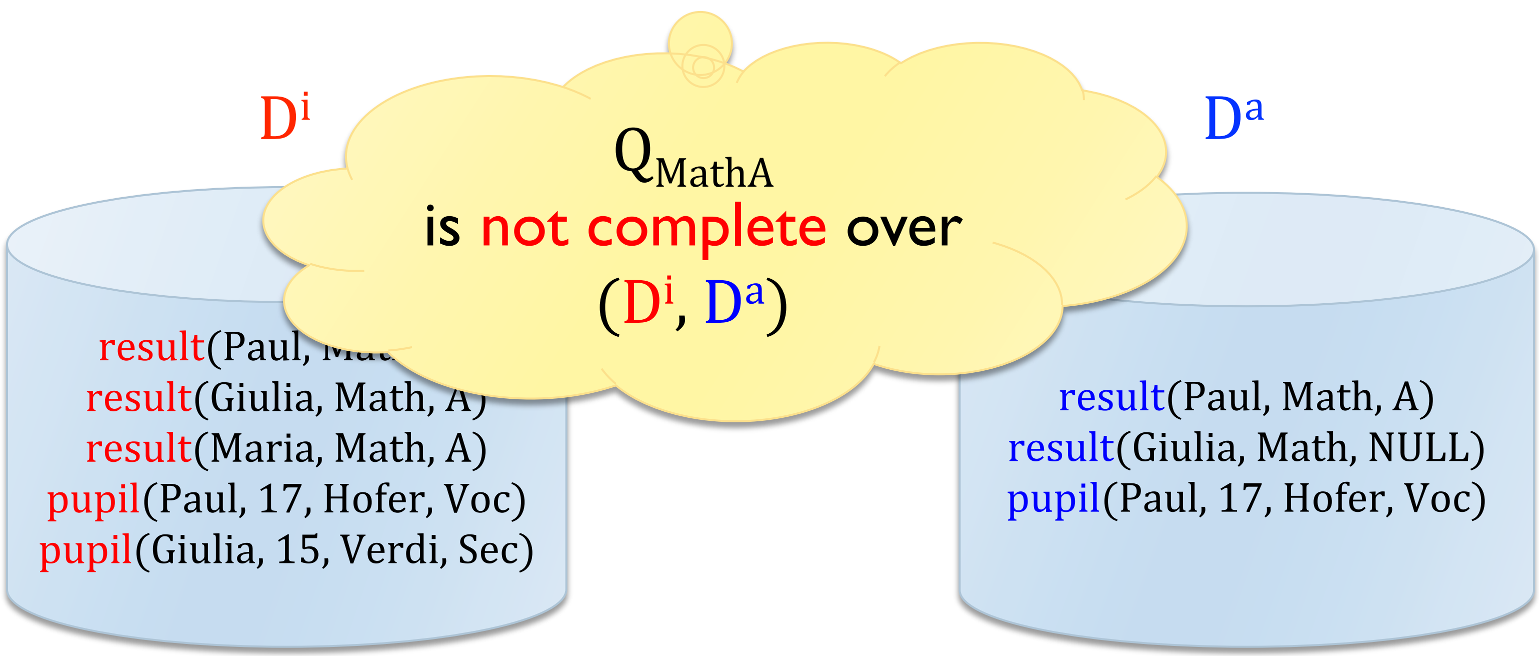
Semantics:

$$(D^i, D^a) \models \text{Compl}(Q) \quad \text{iff} \quad Q(D^i) = Q(D^a)$$

Example: Query Completeness

$$Q_{MathA}(n) \text{ :- } result(n, Math, A)$$

$$Q_{MathA}(D^i) = \{Paul, Giulia, Maria\} \qquad Q_{MathA}(D^a) = \{Paul\}$$



Example: Query Completeness (2)

$$Q_{MathAVoc}(n) \text{ :- } result(n, Math, A), pupil(n, a, sn, Voc)$$

$Q_{MathAVoc}(D^i) = \{Paul\}$ $\qquad$ $Q_{MathAVoc}(D^a) = \{Paul\}$

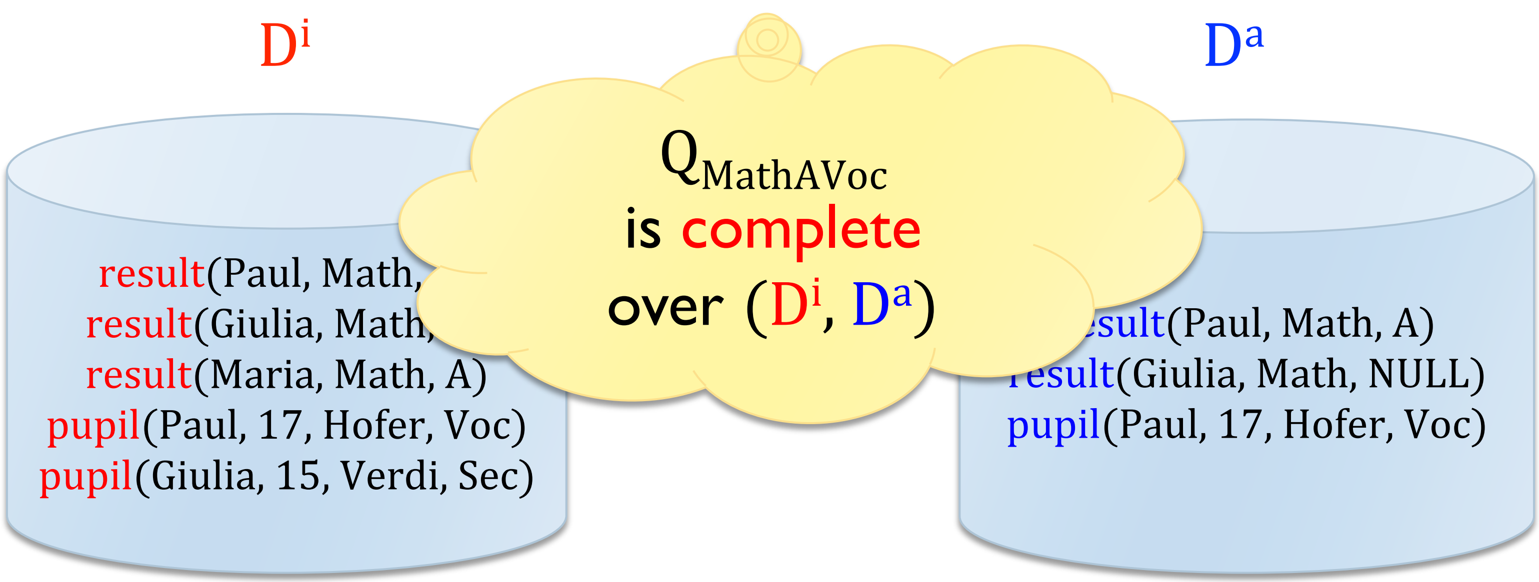


Table Completeness Statements: Idea

"The table result *contains all results of pupils from* ...

means

"If (n,s,g) is a result record according to the ideal db,
and (n, a, sn, Voc) is a pupil record in the ideal db,
then (n,s,g) is in the result table of the available db"

This is a full tuple-generating dependency (TGD)

This can be expressed by the rule

$$result^{i}(n,s,g),\ pupil^{i}(n, a, sn, Voc) \rightarrow result^{a}(n, s, g)$$

We write this table completeness statement as

$$Compl(\ result(n, s, g)\ ;\ pupil(n, a, s, Voc))$$

Idea: an incomplete db satisfies the statement iff it satisfies the rule

Table Completeness Statements [Halevy 96]

A table completeness (TC) statement for a R is an expression

$$\mathrm{Compl}(R(s_1,\ldots, s_n)\ ;\ G)$$

consisting of

- an R-*atom* $R(s_1,\ldots, s_n)$
- a condition G such that $R(s_1,\ldots, s_n), G$ is safe.

G may contain both, relational and built-in atoms

The TC-statement $C = \mathrm{Compl}(R(s_1,\ldots, s_n)\ ;\ G)$ can be seen as a rule

$$r_C \;=\; R^i(s_1,\ldots, s_n), G^i \rightarrow R^a(s_1,\ldots, s_n)$$

Semantics: $(D^i, D^a) \models C$ iff $(D^i, D^a) \models r_C$

Example: TC Statement Satisfaction

$result^i(n, s, g), pupil^i(n, a, sn, Voc) \rightarrow result^a(n, s, g)$

holds over (D^i, D^a)

D^i

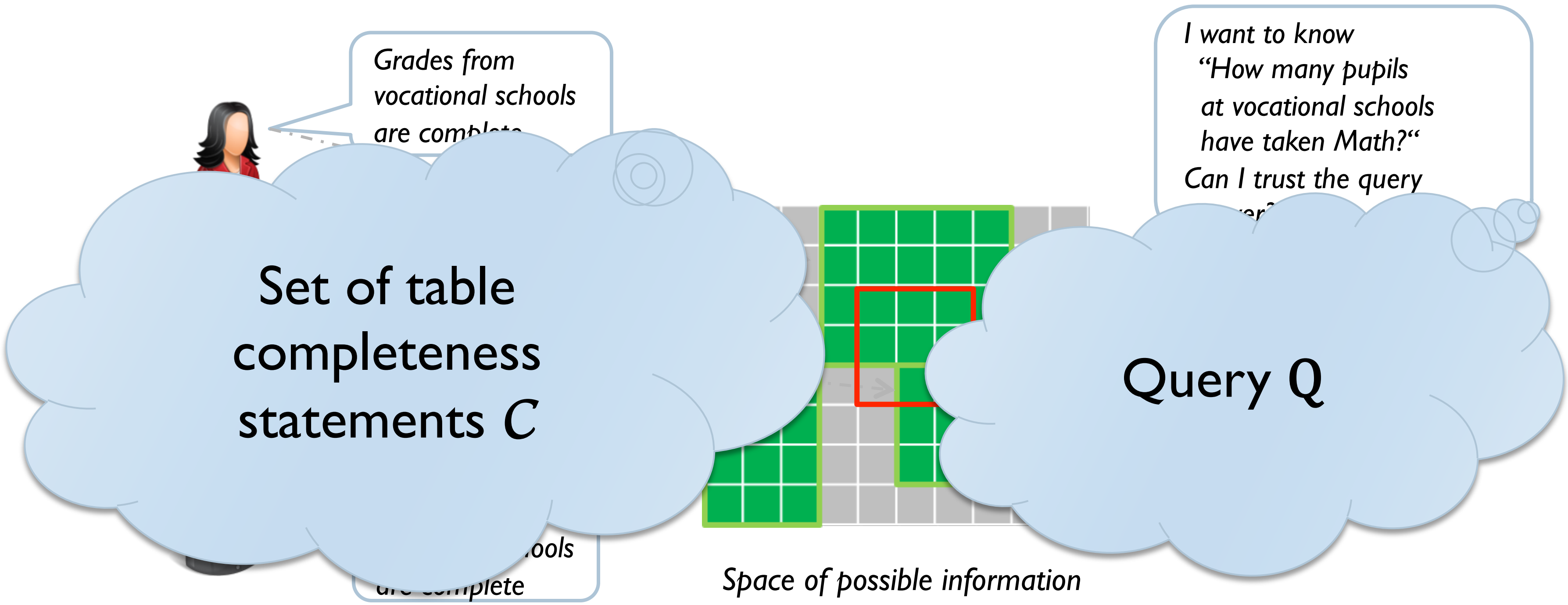
result(Paul, Math, A)
result(Giulia, Math, A)
result(Maria, Math, A)
pupil(Paul, 17, Hofer, Voc)
pupil(Giulia, 15, Verdi, Sec)

D^a

result(Paul, Math, A)
result(Giulia, Math, NULL)
pupil(Paul, 17,Hofer, Voc)

because result(Paul, Math, A) is in D^a

The TC-QC Reasoning Problem



Assertions about partial completeness

$$C \models \text{Compl}(Q) \ ?$$

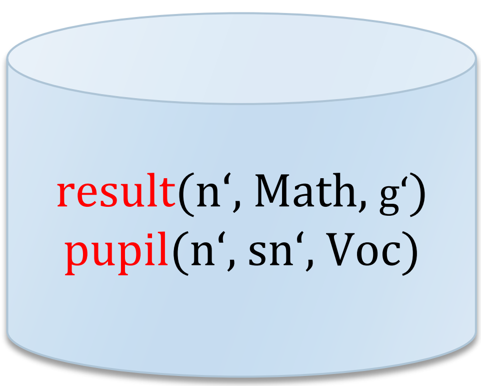
Reasoning: The Principle

“Which pupils at vocational schools had an A in Math?“

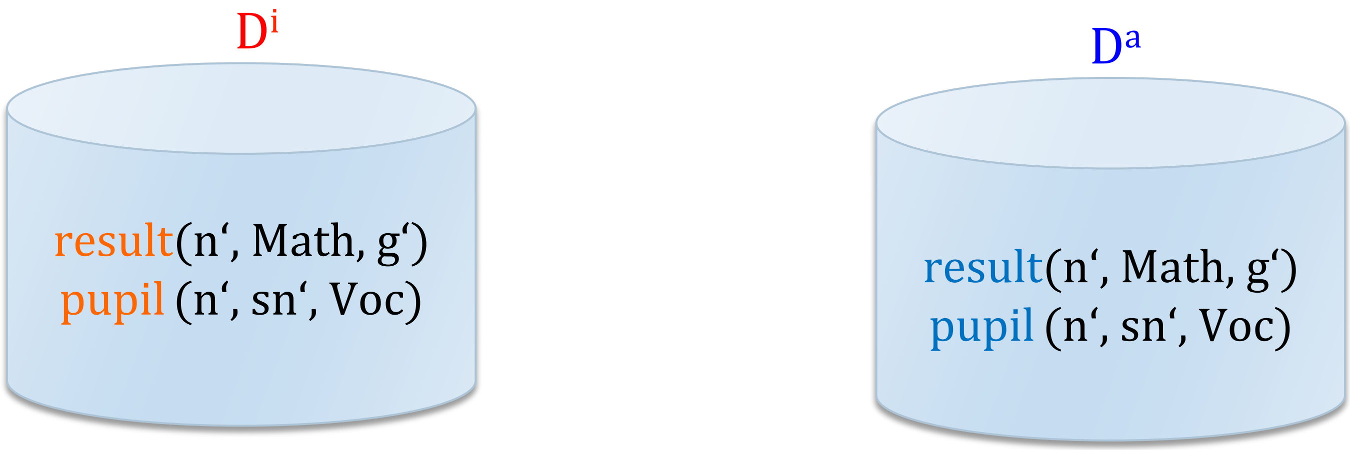
$$Q_{MathAVoc}(n) \text{ :- } result(n, Math, A), pupil(n, sn, Voc)$$

1. Assume $Q_{MathAVoc}$ returns n' over D^i

2. See which facts must be in D^i



Reasoning: The Principle (2)



3. Use table completeness to derive facts in D^a

“All results of pupils at vocational schools are available“

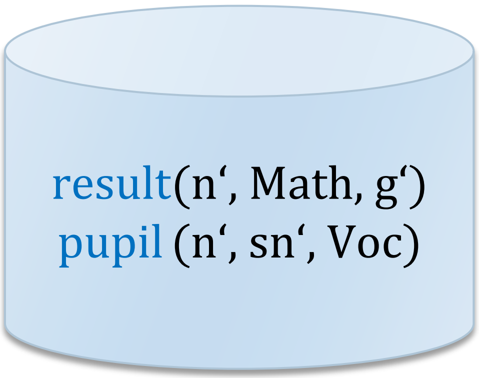
$\text{result}^i(n, s, g), \text{pupil}^i(n, sn, Voc) \rightarrow \text{result}^a(n, s, g)$

“All pupils are available“

$\text{pupil}^i(n, sn, st) \rightarrow \text{pupil}^a(n, sn, st)$

Reasoning: The Principle (3)

D^a



4. Query the available database *“Pupils at vocational schools with an A in Math“*

$Q_{MathAVoc}(D^a) = \{n'\}$ → n' is also in $Q(D^a)$

Conclusion: $Q_{MathAVoc}$ is complete given the table completeness statements

TC-Transformation

To $C = \mathrm{Compl}(R(\underline{s})\,;\,G)$ we associate the query

$$Q_C(\underline{s}) \text{ :- } R(\underline{s})\,,\,G$$

and the transformation on db instances

$$T_C(D) := \{\, R(\underline{t}) \mid \underline{t} \in Q_C(D) \,\}$$

For a set $\mathcal{C}$ of TC statements we define the transformation

$$T_{\mathcal{C}}(D) := \bigcup_{C \in \mathcal{C}} T_C(D)$$

TC-Transformations: Properties

- $(D, T_C(D))$ is an incomplete database
- $(D, T_C(D)) \models C$
- $(D^i, D^a) \models C$ iff $T_C(D^i) \subseteq D^a$

In other words:

- $(D, T_C(D))$ is the least incomplete database
  - with ideal db D
  - that satisfies C

TC-QC Reasoning: Relational Case

Let

- C set of relational TC statements
- $Q(\underline{x})$:- L relational query
- L′ := *frozen version* of L

variables $\underline{x}$, $\underline{y}$ considered as constants $\underline{x}'$, $\underline{y}'$

Theorem:

$$C \models \text{Compl}(Q) \quad \text{iff} \quad \underline{x}' \in Q(T_C(L'))$$

What if C or Q contain comparisons?

Example: TC-QC with Comparisons

Query: $Q_{pupil}(n)$:- pupil(n, a, sn, st)

C = { $C_{\leq 10}$: pupil^{i}(n, a, sn, s

$C_{>10}$: pupil^{i}(n, a, sn, s

- We retrieve n′ in all 3 cases
- The cases cover all possibilities

→ Q is complete wrt *C*

How can we chase L′ = { pupil(n′, a′, sn′, st′) } with C ?

Idea: Case analysis!

Substitute "*representative values*" for $a' < 10$, $a' = 10$, > 10

Substitution yields: [a′/9]L′ = { pupil(n′, 9, sn′, st′) }

to which we can apply $C_{\leq 10}$ …

TC-QC Reasoning with Comparisons

Let

- $\mathcal{C}$ set of TC statements with comparisons
- $Q(\underline{\mathbf{x}})$:- L, M
- Γ set of representative value substitutions for $\mathcal{C}$, Q

Theorem: The following are equivalent

- □ $\mathcal{C} \models \text{Compl}(Q)$
- □ $\gamma\underline{\mathbf{x}}' \in Q(T_{\mathcal{C}}(\gamma L'))$ for all $\gamma \in \Gamma$

Set Semantics vs. Bag Semantics

$Q(\underline{x})$:- L query

$(D^i, D^a) \models \mathrm{Compl}_{set}(Q)$

iff every answer of Q over D^i is returned over D^a, too

iff $\alpha L \subseteq D^i \Rightarrow$ ex. β s.th. $\beta L \subseteq D^a$ and $\beta\underline{x} = \alpha\underline{x}$

$(D^i, D^a) \models \mathrm{Compl}_{bag}(Q)$

iff every answer of Q over D^i is returned over D^a *the same number of times*

iff $\alpha L \subseteq D^i \Rightarrow \alpha L \subseteq D^a$

"no assignments get lost"

TC-QC Reasoning for Bag Semantics

Let

- $\mathcal{C}$ set of TC statements with comparisons
- $Q(\underline{x})$:- L, M
- Γ set of representative value substitutions for $\mathcal{C}$, Q

Theorem:

$$\mathcal{C} \models \mathrm{Compl}_{bag}(Q) \quad \text{iff} \quad \gamma L' \subseteq T_{\mathcal{C}}(\gamma L') \quad \text{for all} \quad \gamma \in \Gamma$$

Corollary: If $\mathcal{C}$ has no comparisons, then:

$$\mathcal{C} \models \mathrm{Compl}_{bag}(Q) \quad \text{iff} \quad L' \subseteq T_{\mathcal{C}}(L')$$

Complexity

Classes of conjunctive queries:

- CQ: Conjunctive queries with comparisons over dense orders
- RQ: Relational conjunctive queries (i.e., without comparisons)
- LCQ: Linear conjunctive queries (i.e., without self-joins)
- LRQ: Linear relational conjunctive queries

TC-QC$_{bag}$ - Complexity

		Query Language			
		LRQ	LCQ	RQ	CQ
TC Statement Language	LRQ	in PTIME	in PTIME	NP	NP
	RQ	in PTIME	in PTIME	NP	NP
	LCQ	coNP	coNP	Π^P_2	Π^P_2
	CQ	coNP	coNP	Π^P_2	Π^P_2

Note, the axes are asymmetric:

- NP appears with repeated relation symbols in the query
- coNP appears with comparisons in the TC statements

TC-QC$_{set}$ - Complexity

		Query Language			
		LRQ	LCQ	RQ	CQ
TC Statement Language	LRQ	in PTIME	in PTIME	NP	Π^P_2
	RQ	in PTIME	in PTIME	NP	Π^P_2
	LCQ	coNP	coNP	Π^P_2	Π^P_2
	CQ	coNP	coNP	Π^P_2	Π^P_2

Intuition: the query has to be contained in the TC-statements …

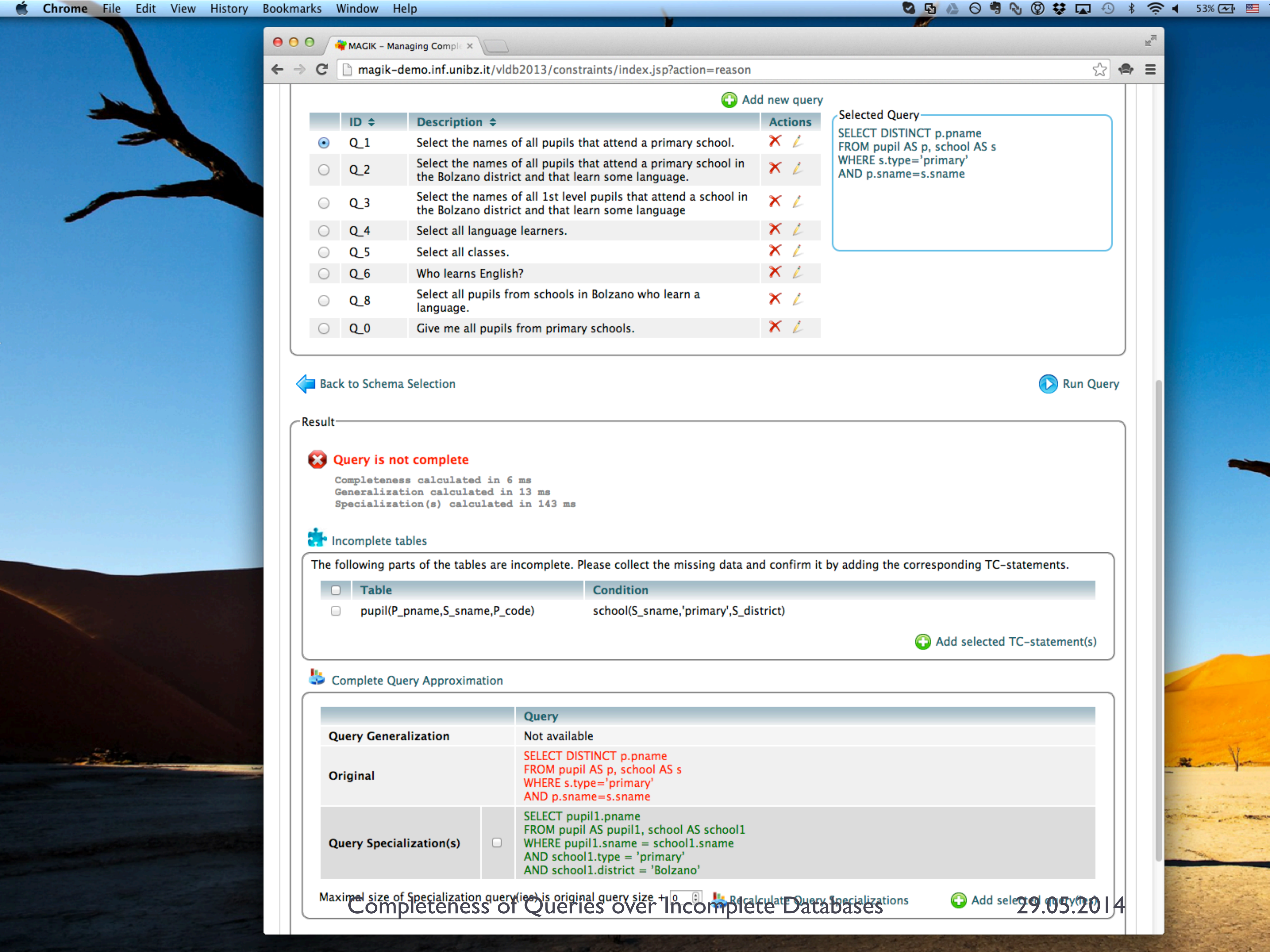
… but that does not explain it all

How Can One Implement Completeness Reasoning?

Idea: Map reasoning tasks to a generic reasoner

Candidate reasoners:

- SMT (SAT modulo theories) solvers ?
  - encoding may be of exp. size for Π^P_2 problems
- Disjunctive Logic Programming with Answer Set Semantics ?
  - can express all Π^P_2 problems
  - demo implementation for
    - conjunctive queries
    - finite domain constraints
    - keys and (acyclic) foreign keys


MAGIK – Managing Comple
magik-demo.inf.unibz.it/vldb2013/constraints/index.jsp?action=reason
Add new query
ID
Description
Actions
Q_1 Select the names of all pupils that attend a primary school.
Q_2 Select the names of all pupils that attend a primary school in the Bolzano district and that learn some language.
Q_3 Select the names of all 1st level pupils that attend a school in the Bolzano district and that learn some language
Q_4 Select all language learners.
Q_5 Select all classes.
Q_6 Who learns English?
Q_8 Select all pupils from schools in Bolzano who learn a language.
Q_0 Give me all pupils from primary schools.
Selected Query
SELECT DISTINCT p.pname
FROM pupil AS p, school AS s
WHERE s.type='primary'
AND p.sname=s.sname
Back to Schema Selection
Run Query
Result
Query is not complete
Completeness calculated in 6 ms
Generalization calculated in 13 ms
Specialization(s) calculated in 143 ms
Incomplete tables
The following parts of the tables are incomplete. Please collect the missing data and confirm it by adding the corresponding TC-statements.
Table
Condition
pupil(P_pname,S_sname,P_code)
school(S_sname,'primary',S_district)
Add selected TC-statement(s)
Complete Query Approximation
Query
Query Generalization
Not available
Original
SELECT DISTINCT p.pname
FROM pupil AS p, school AS s
WHERE s.type='primary'
AND p.sname=s.sname
Query Specialization(s)
SELECT pupil1.pname
FROM pupil AS pupil1, school AS school1
WHERE pupil1.sname = school1.sname
AND school1.type = 'primary'
AND school1.district = 'Bolzano'
Maximal size of Specialization query(ies) is original query size +
Recalculate Query Specializations
Add selected query(ies)

Completeness on the Semantic Web

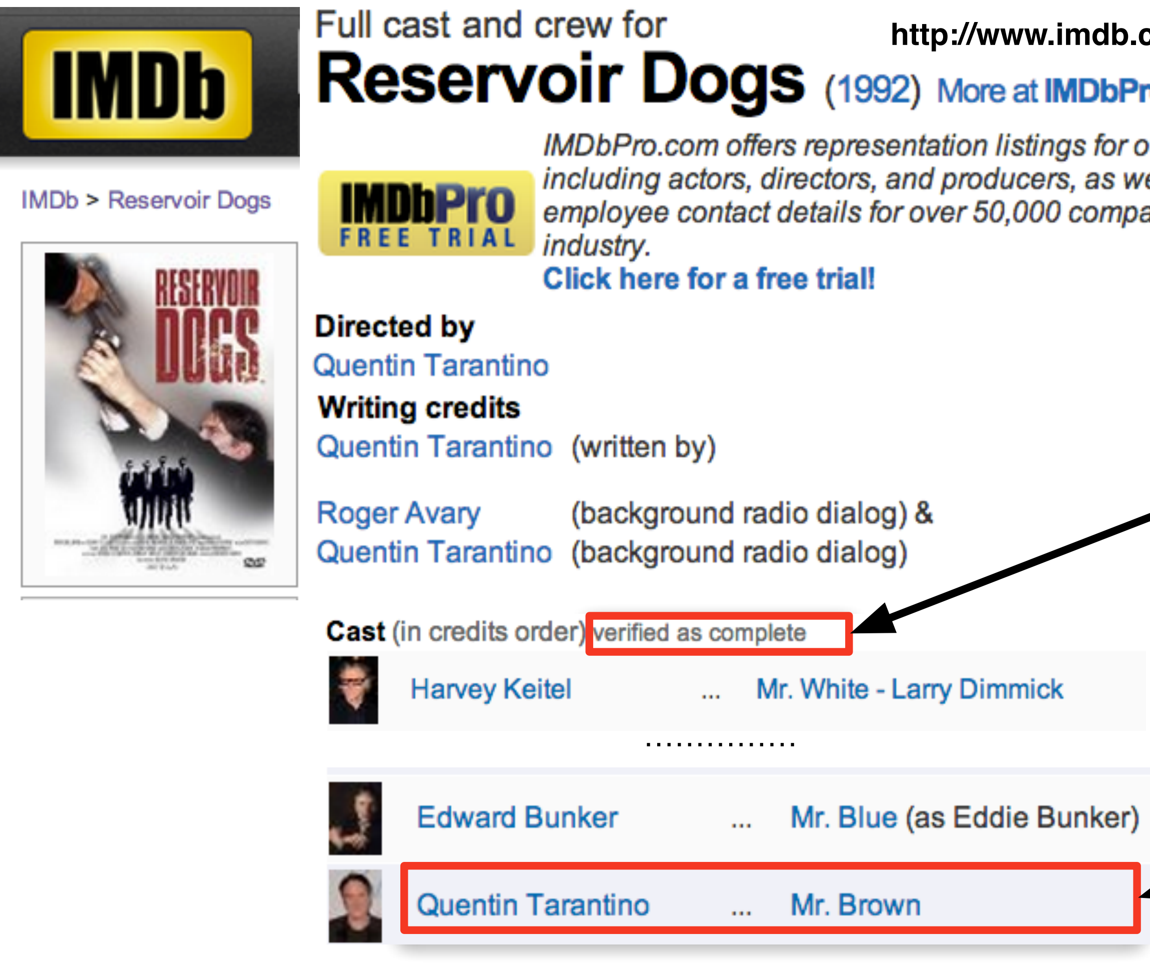
DBPedia Misses Some Facts …

dbpedia.org/page/Reservoir_Dogs

dbpedia-owl:runtime	▪ 5940 (xsd:double)
dbpedia-owl:starring	▪ dbpedia:Chris_Penn ▪ dbpedia:Tim_Roth ▪ dbpedia:Lawrence_Tierney ▪ dbpedia:Steve_Buscemi ▪ dbpedia:Harvey_Keitel ▪ dbpedia:Michael_Madsen

IMDB Has Completeness Guarantees

http://www.imdb.com/title/tt0105236/fullcredits?ref_=tt_ov_st_sm#cast

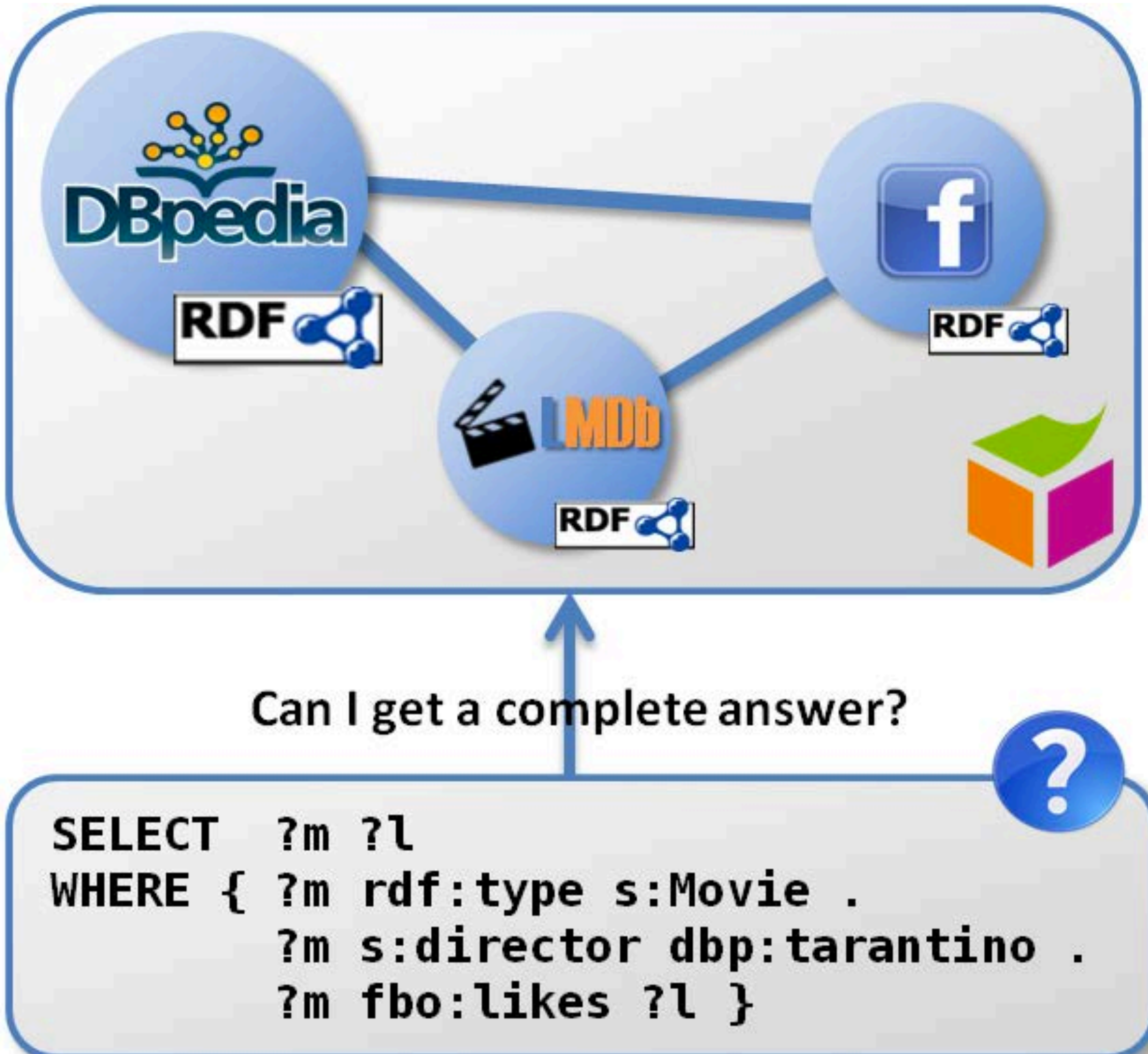


Completeness statement about the IMDB data source

Quentin Tarantino was the character Mr. Brown

If Completeness Info Were Available in RDF …

Federated Framework



Completeness of SPARQL Queries over RDF Sources

- Completeness statements in RDF
- Reasoning algorithms for queries with
  - **`DISTINCT`**
  - **`OPT`**
  - over RDFS sources
- Generation of queries with **`SERVICE`** calls
  over federated sources
- Prototypical implementation using Apache Jena
  `http://rdfcorner.wordpress.com`

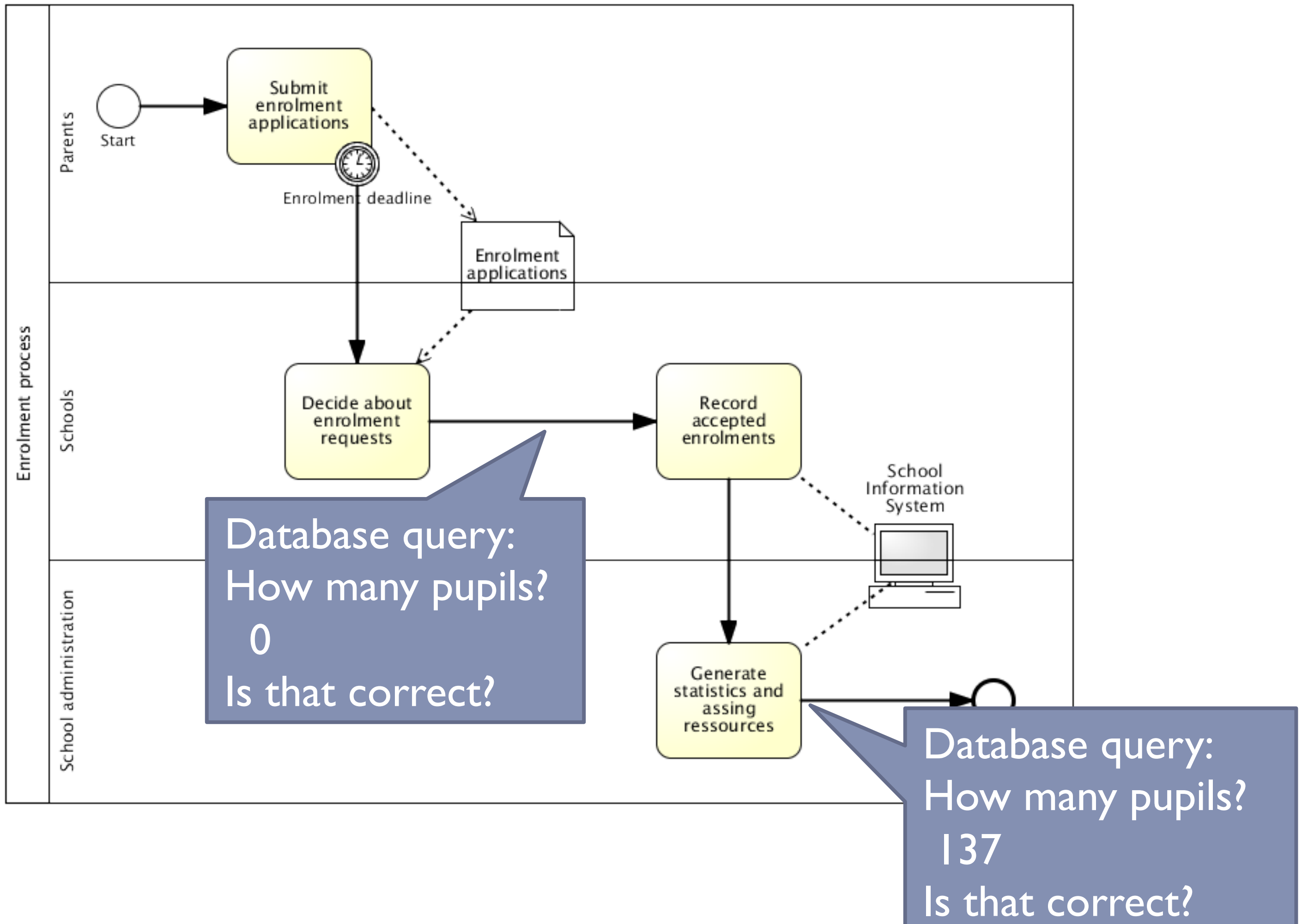
Verifying Query Completenessover Processes

- Data often created following processes
- Many processes are executed only partially formal
  (pen & paper, email, phone, …)

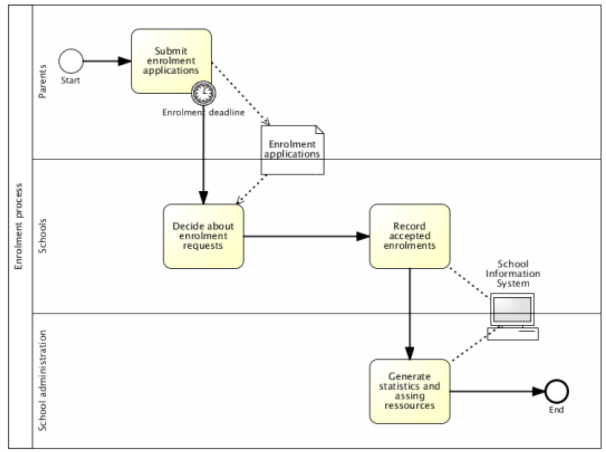
➔ Valid information may be stored in databases
  with delays

➔ Database content is of questionable completeness

Enrolment Process in a School



Observation



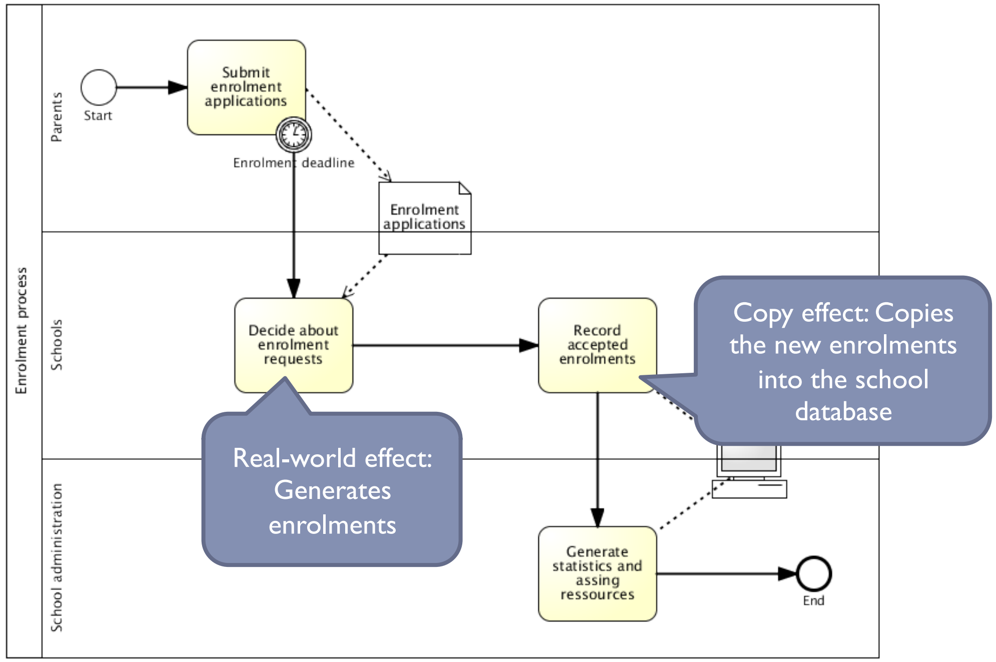
- At some points, new facts in the real world have not yet been stored

  ➔ queries may give wrong answers

- At other points, all facts that hold in the real world have been stored

  ➔ queries give correct answers

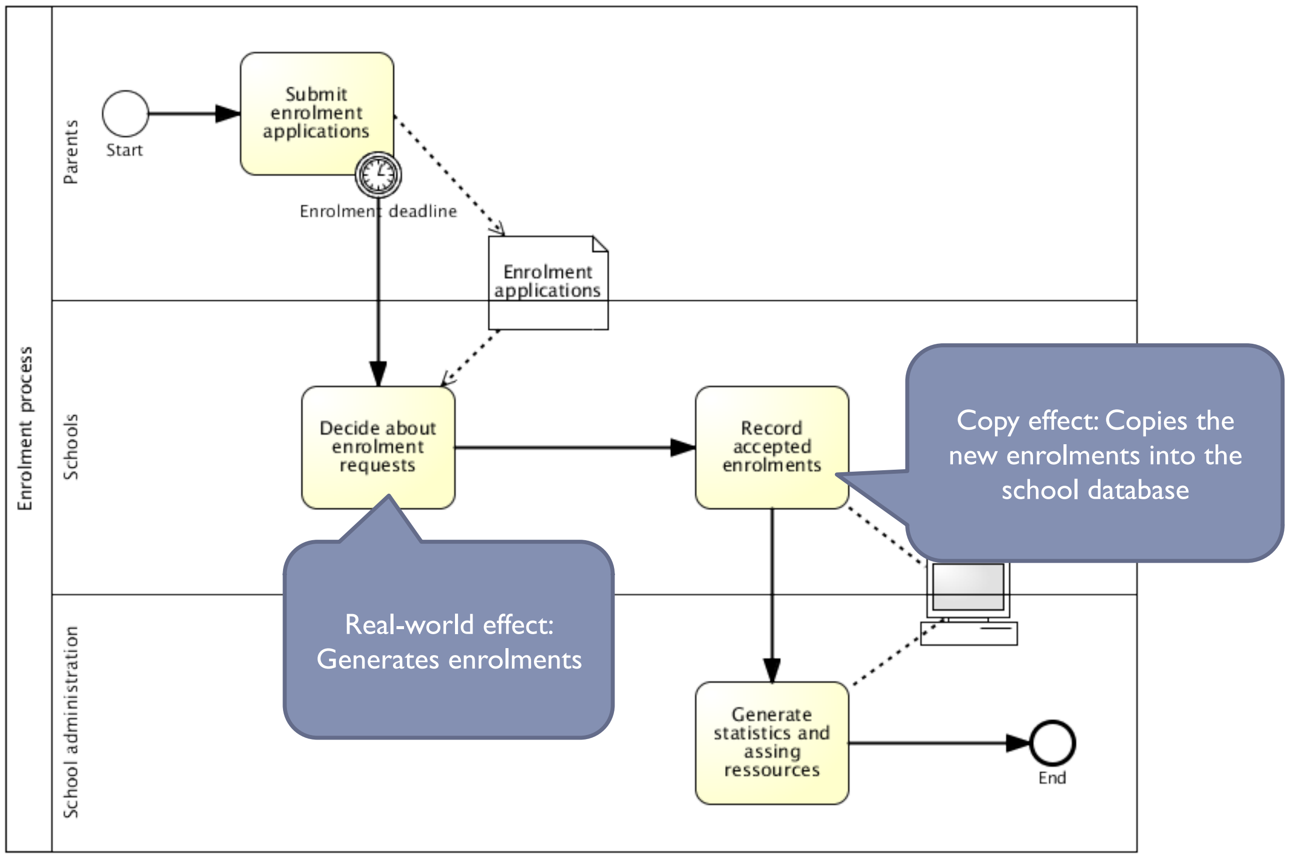
Real-world and Copy Effects



Real-world effect: $\text{pupil}^{rw}(n, s) \leftsquigarrow \text{request}^{rw}(n, s)$

Copy effect: $\text{pupil}^{rw}(n, s) \rightarrow \text{pupil}^{is}(n, s)$

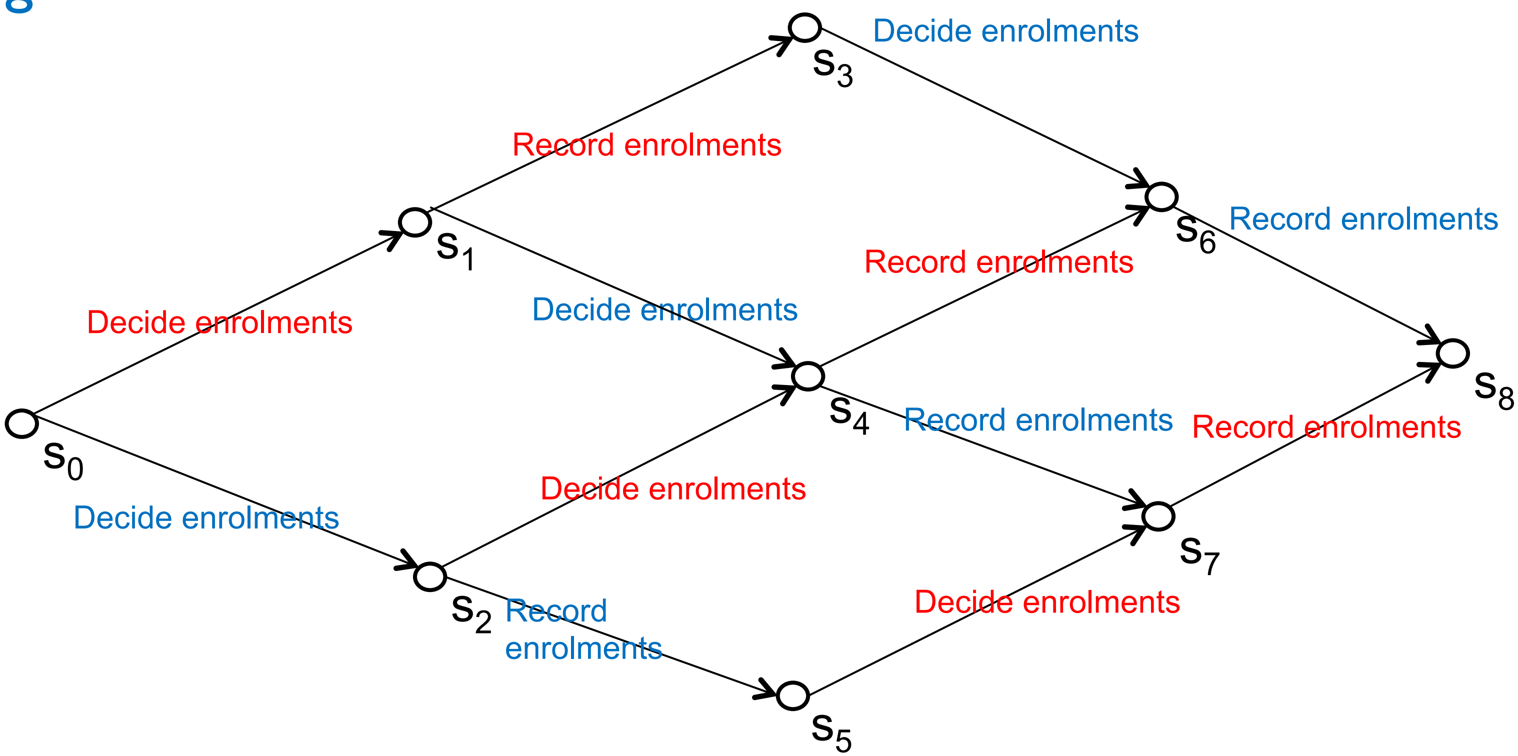
Transition Systems for Process Instances



Transition Systems for Process Instances

Two concurrent process instances:

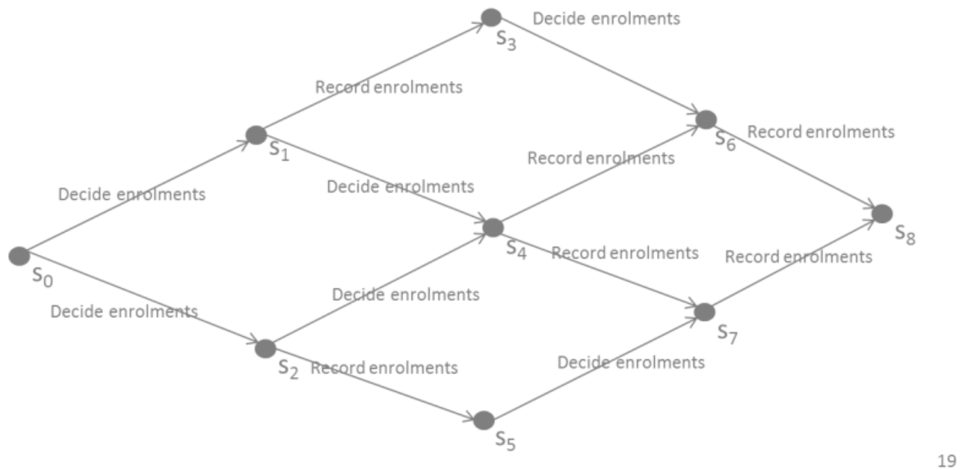
- Middle School A
- High School B



Completeness Verification

Given

- Process description
- State S
- Query Q



Question

Is it safe to pose the query Q in state S against the information system database?

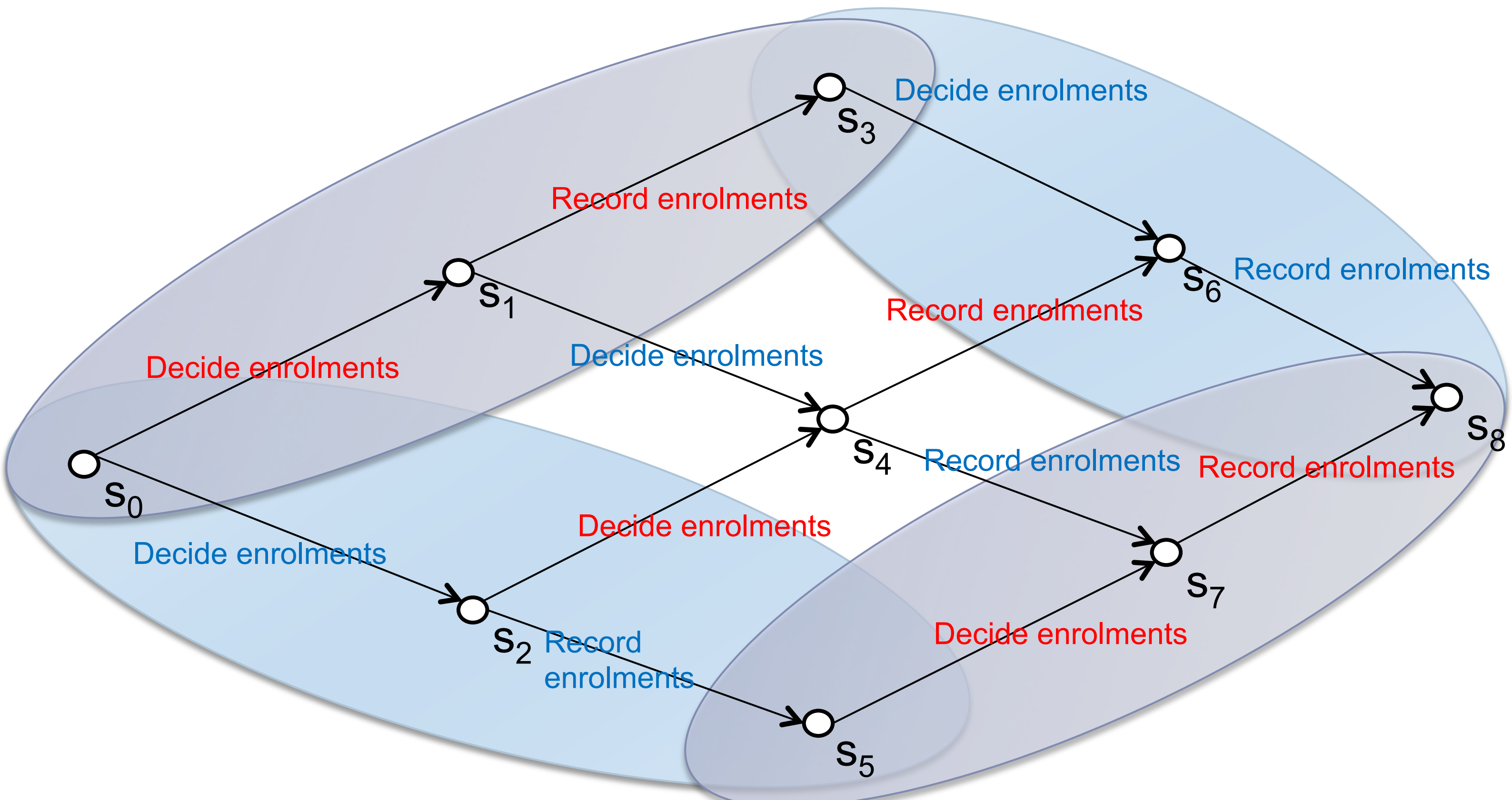
Verification: Example Revisited

Middle School A

High School B

How many middle school pupils?

How many high school pupils?



Possible Applications

- Annotation of statistics and KPI with completeness information
- Process mining (trace analysis) - to validate whether queries over traces return the real state of the process
- Auditing – to verify whether the information about the real-world is properly stored

Conclusion

- Framework for statements about completeness of
  - query answers
  - (projections of) parts of db tables
- Complexity of TC-QC Reasoning
- Implementation based on DLV answer set programming engine
- Application to
  - Semantic Web
  - Business Processes

Questions?