Knowledge-enriched Databases

Querying Relational Databases

List the codes of teaching staff

Lecturer	id	Name
	1	Alice
	2	Bob
	3	Tom
	4	Mary

Course	code	organiser
	CS100	2
	CS200	1
	CS300	4

Q(x) :- TeachingStaff(x,y)

Querying Relational Databases

List the codes of teaching staff

Lecturer	id	Name
	1	Alice
	2	Bob
	3	Tom
	4	Mary

Course	code	organiser
	CS100	2
	CS200	1
	CS300	4

Lecturers are teaching staff

Course organisers are teaching staff

Q(x) :- TeachingStaff(x,y)

Querying Relational Databases

List the codes of teaching staff

Lecturer	id	Name
	1	Alice
	2	Bob
	3	Tom
	4	Mary

Course	code	organiser
	CS100	2
	CS200	1
	CS300	4

$$\forall x \forall y \, (\mathrm{Lecturer}(x,y) \rightarrow \mathrm{TeachingStaff}(x,y))$$

$$\forall x \forall y \, (\mathrm{Course}(x,y) \rightarrow \exists z \, \mathrm{TeachingStaff}(y,z))$$

$Q(x) \ :\!- \ \mathrm{TeachingStaff}(x,y)$

Querying Relational Databases

List the codes of teaching staff

Lecturer	id	Name
	1	Alice
	2	Bob
	3	Tom
	4	Mary

$\Downarrow$

{1, 2, 3, 4}

Course	code	organiser
	CS100	2
	CS200	1
	CS300	4

$\forall x \forall y\ (Lecturer(x,y) \rightarrow TeachingStaff(x,y))$

$\forall x \forall y\ (Course(x,y) \rightarrow \exists z\ TeachingStaff(y,z))$

Q(x) :- TeachingStaff(x,y)

Some Terminology

- Our basic vocabulary:
  - A countable set **Const** of constants - domain of a database
  - A countable set **Nulls** of marked nulls - globally ∃-quantified variables
  - A countable set **Vars** of variables - used in rules and queries
- A term is a constant, marked null, or variable
- An atom has the form $R(t_1,\ldots,t_n)$ - R is an n-ary relation and t_i's are terms
- An instance is a (*possibly infinite*) set of atoms with constants and nulls
- A database is a finite instance with only constants

Syntax of Existential Rules

An existential rule is an expression

$$\forall \mathbf{x} \forall \mathbf{y} \, (\underbrace{\varphi(\mathbf{x},\mathbf{y})}_{\text{body}} \rightarrow \exists \mathbf{z} \, \underbrace{\psi(\mathbf{x},\mathbf{z})}_{\text{head}})$$

- $\mathbf{x}$,$\mathbf{y}$ and $\mathbf{z}$ are tuples of variables of **Vars**
- $\varphi(\mathbf{x},\mathbf{y})$ and $\psi(\mathbf{x},\mathbf{z})$ are (constant-free) conjunctions of atoms

…also known as tuple-generating dependencies and Datalog+/- rules

Semantics of Existential Rules

- An instance J is a model of the rule

$$\sigma \;=\; \forall \mathbf{x} \forall \mathbf{y}\, (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z}\, \psi(\mathbf{x},\mathbf{z}))$$

  written as $J \vDash \sigma$, if the following holds:

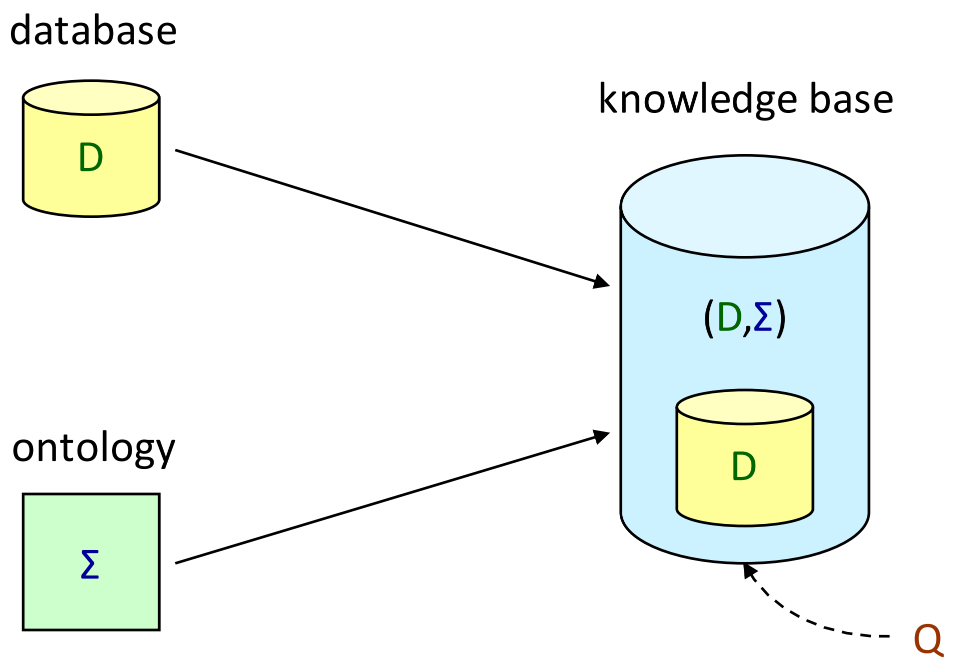
  whenever there exists a homomorphism h such that $h(\varphi(\mathbf{x},\mathbf{y})) \subseteq J$,

  then there exists $g \supseteq h_{|\mathbf{x}}$ such that $g(\psi(\mathbf{x},\mathbf{z})) \subseteq J$

  $\{t \mapsto h(t) \mid t \in \mathbf{x}\}$ - the restriction of h to $\mathbf{x}$

- Given a set Σ of existential rules, J is a model of Σ, written as $J \vDash \Sigma$, if, for each $\sigma \in \Sigma$, $J \vDash \sigma$

Ontology-Based Query Answering (OBQA)



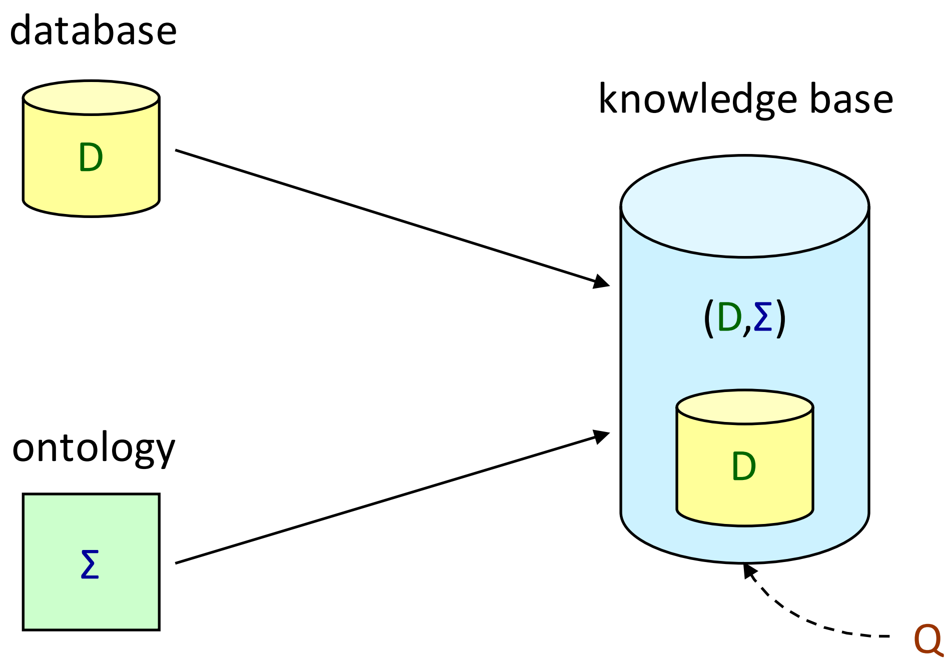
existential rules

$\forall \mathbf{x} \forall \mathbf{y}\ (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z}\ \psi(\mathbf{x},\mathbf{z}))$

conjunctive query

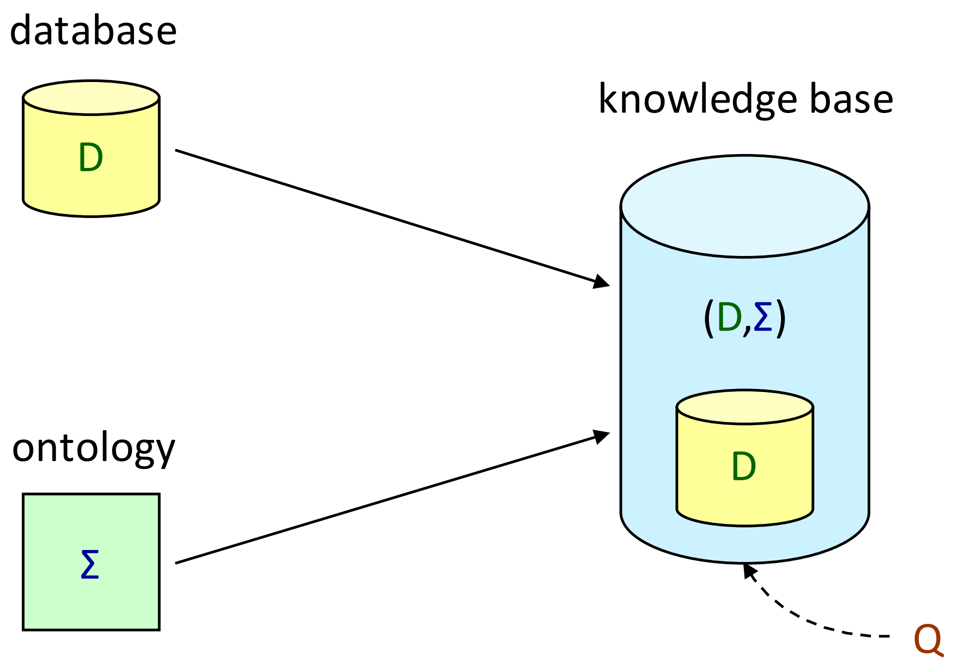
$Q(\mathbf{x})\ \ \text{:-}\ \ R_1(\mathbf{v_1}),\ldots,R_m(\mathbf{v_m})$

Ontology-Based Query Answering (OBQA)



$$\text{models}(D,\Sigma) \;=\; \{J \mid J \supseteq D \text{ and } J \models \Sigma\}$$

Ontology-Based Query Answering (OBQA)



$$\text{Answer}(Q,D,\Sigma) \quad = \quad \bigcap_{J \in \text{models}(D,\Sigma)} Q(J)$$

Exercise: Compute the Certain Answers

$D = \{Person(john),\ Person(bob),\ Person(tom),$
$hasFather(john,bob),\ hasFather(bob,tom)\}$

$\Sigma = \{\forall x\, (Person(x) \rightarrow \exists y\, hasFather(x,y)),$
$\forall x \forall y\, (hasFather(x,y) \rightarrow Person(x) \land Person(y))\}$

$Q_1(x,y)$:- $hasFather(x,y)$

$Q_2(x)$:- $hasFather(x,y)$

$Q_3(x)$:- $hasFather(x,y), hasFather(y,z), hasFather(z,w)$

$Q_4(x,w)$:- $hasFather(x,y), hasFather(y,z), hasFather(z,w)$

Exercise: Compute the Certain Answers

$D = \{Person(john),\ Person(bob),\ Person(tom),$
$hasFather(john,bob),\ hasFather(bob,tom)\}$

$\Sigma = \{\forall x\, (Person(x) \rightarrow \exists y\ hasFather(x,y)),$
$\forall x \forall y\, (hasFather(x,y) \rightarrow Person(x) \wedge Person(y))\}$

$Q_1(x,y)$:- $hasFather(x,y)$

$\{(john,bob),\ (bob,tom)\}$

Exercise: Compute the Certain Answers

$D = \{Person(john),\ Person(bob),\ Person(tom),$
$hasFather(john,bob),\ hasFather(bob,tom)\}$

$\Sigma = \{\forall x\,(Person(x) \rightarrow \exists y\ hasFather(x,y)),$
$\forall x \forall y\,(hasFather(x,y) \rightarrow Person(x) \land Person(y))\}$

$Q_2(x)$:- $hasFather(x,y)$

$\{(john), (bob), (tom)\}$

Exercise: Compute the Certain Answers

D = {Person(john), Person(bob), Person(tom),

hasFather(john,bob), hasFather(bob,tom)}

Σ = {$\forall x$ (Person(x) $\rightarrow \exists y$ hasFather(x,y)),

$\forall x \forall y$ (hasFather(x,y) $\rightarrow$ Person(x) $\wedge$ Person(y))}

Q_3(x) :- hasFather(x,y), hasFather(y,z), hasFather(z,w)

{(john), (bob), (tom)}

Exercise: Compute the Certain Answers

$D = \{Person(john),\ Person(bob),\ Person(tom),$

$\quad hasFather(john,bob),\ hasFather(bob,tom)\}$

$\Sigma = \{\forall x\, (Person(x) \rightarrow \exists y\ hasFather(x,y)),$

$\quad \forall x \forall y\, (hasFather(x,y) \rightarrow Person(x) \wedge Person(y))\}$

$Q_4(x,w)$:- $hasFather(x,y), hasFather(y,z), hasFather(z,w)$

{ }

Ontology-Based Query Answering (OBQA)

ontology language based on existential rules

OBQA(**L**)

Input: a database D, a set of existential rules $\Sigma \in$ **L**, a CQ Q/k, a tuple of constants **t** $\in$ **adom**(D)^k

Question: **t** $\in$ Answer(Q,D,Σ)?

BOBQA(**L**)

Input: a database D, a set of existential rules $\Sigma \in$ **L**, a Boolean query Q

Question: is Answer(Q,D,Σ) non-empty?

Theorem: OBQA(**L**) $\equiv_L$ BOBQA(**L**) for every language **L**

($\equiv_L$ means logspace-equivalent)

Data Complexity of BOBQA

input D, fixed Σ and Q

BOBQA[Σ,Q](**L**)

Input: a database D

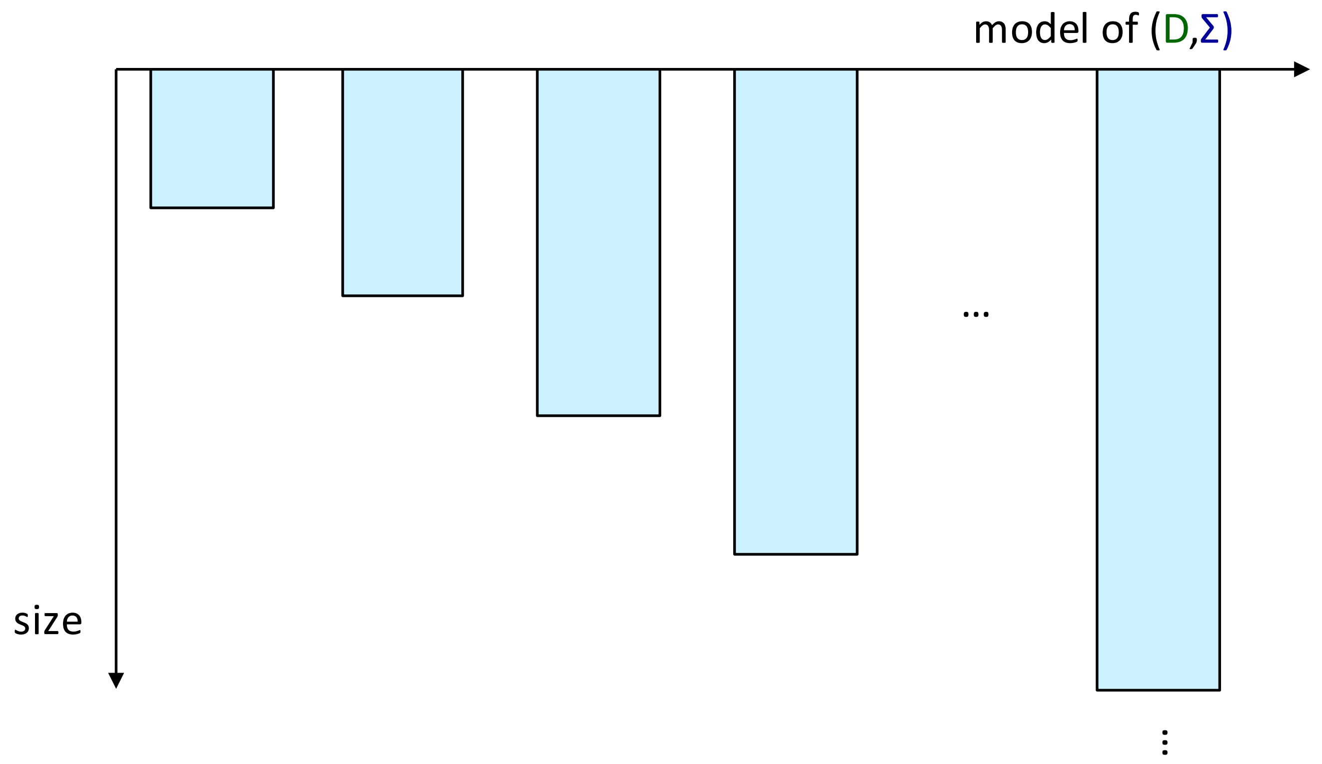
Question: is Answer(Q,D,Σ) non-empty?

Why is OBQA technically challenging?

What is the right tool for tackling this problem?

The Two Dimensions of Infinity

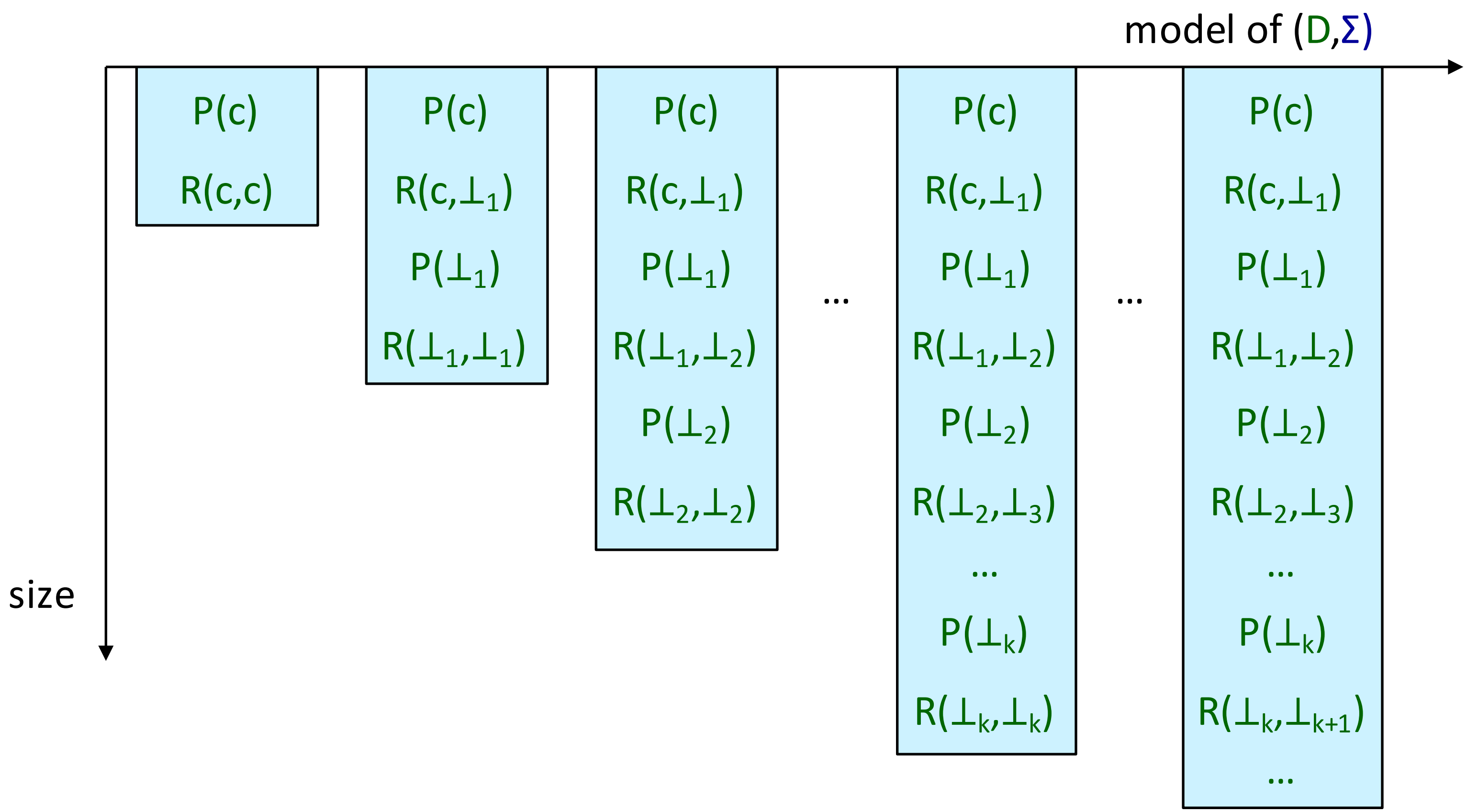
Consider a database D, and a set of existential rules Σ



(D,Σ) admits infinitely many models, of possibly infinite size

The Two Dimensions of Infinity

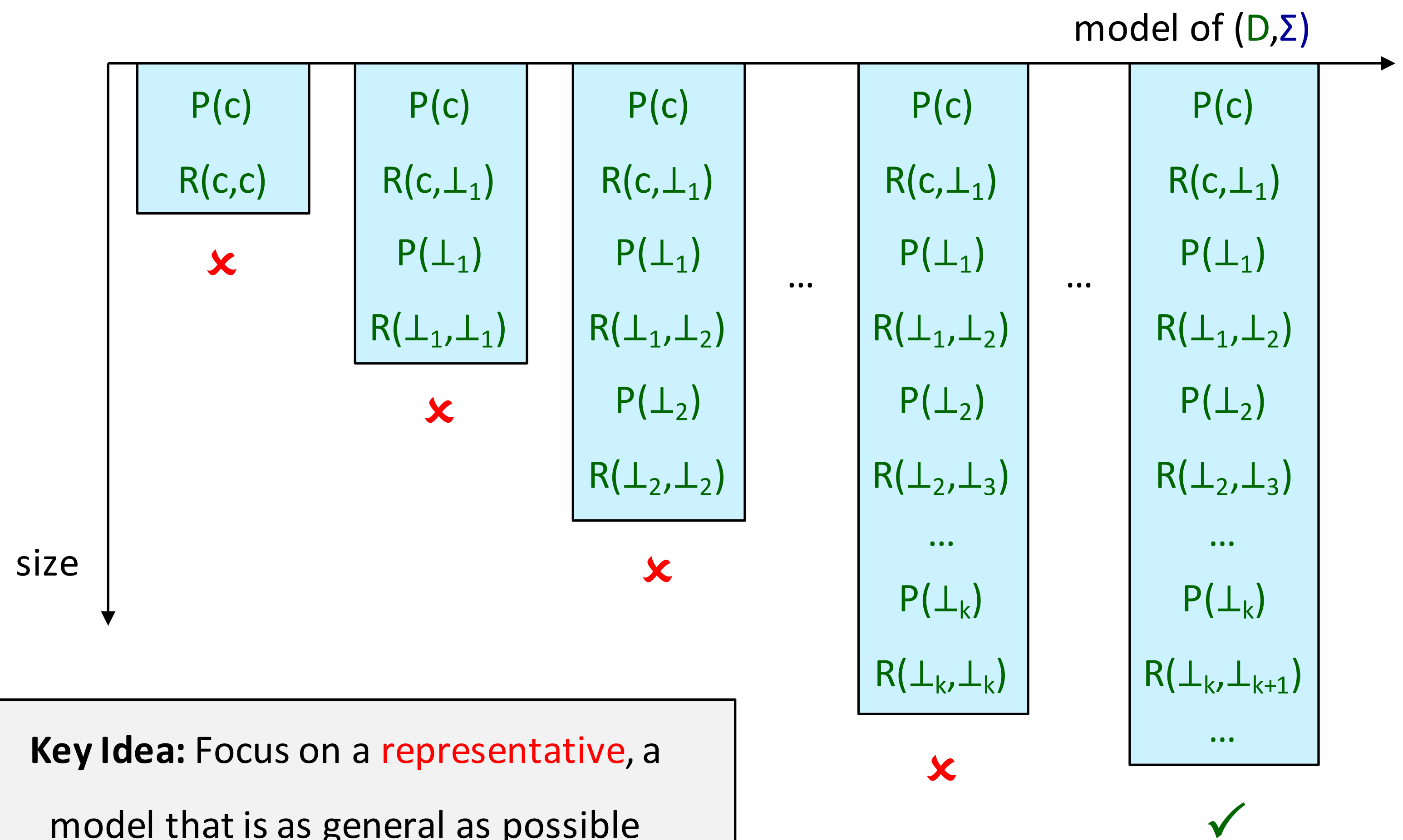
$D = \{P(c)\}$ $\Sigma = \{\forall x\, (P(x) \rightarrow \exists y\, (R(x,y) \wedge P(y)))\}$



$\perp_1, \perp_2, \perp_3, \ldots$ are marked nulls from **Nulls**

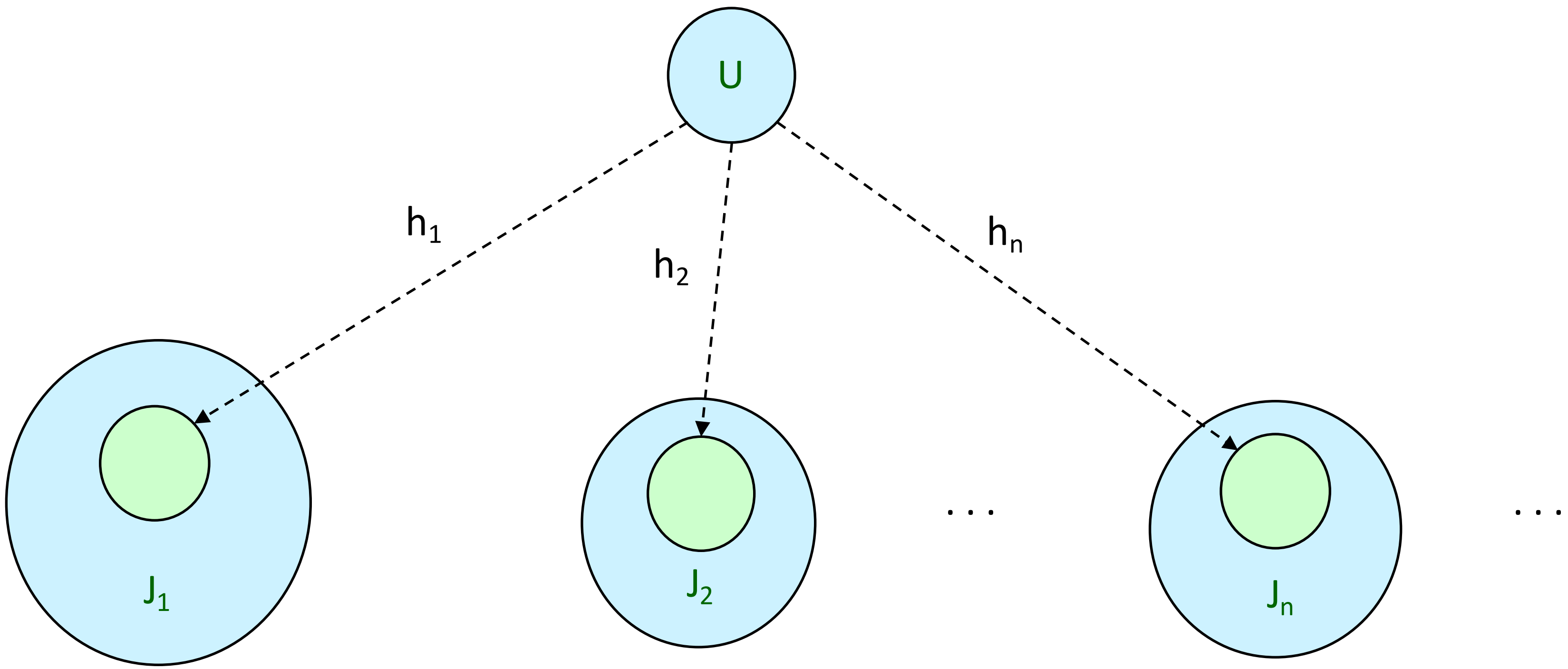
The Two Dimensions of Infinity

$D = \{P(c)\}$ $\Sigma = \{\forall x\, (P(x) \rightarrow \exists y\, (R(x,y) \wedge P(y)))\}$



Key Idea: Focus on a representative, a model that is as general as possible

Universal Models (a.k.a. Canonical Models)



An instance U is a universal model of (D,Σ) if the following holds:

1. U is a model of (D,Σ)

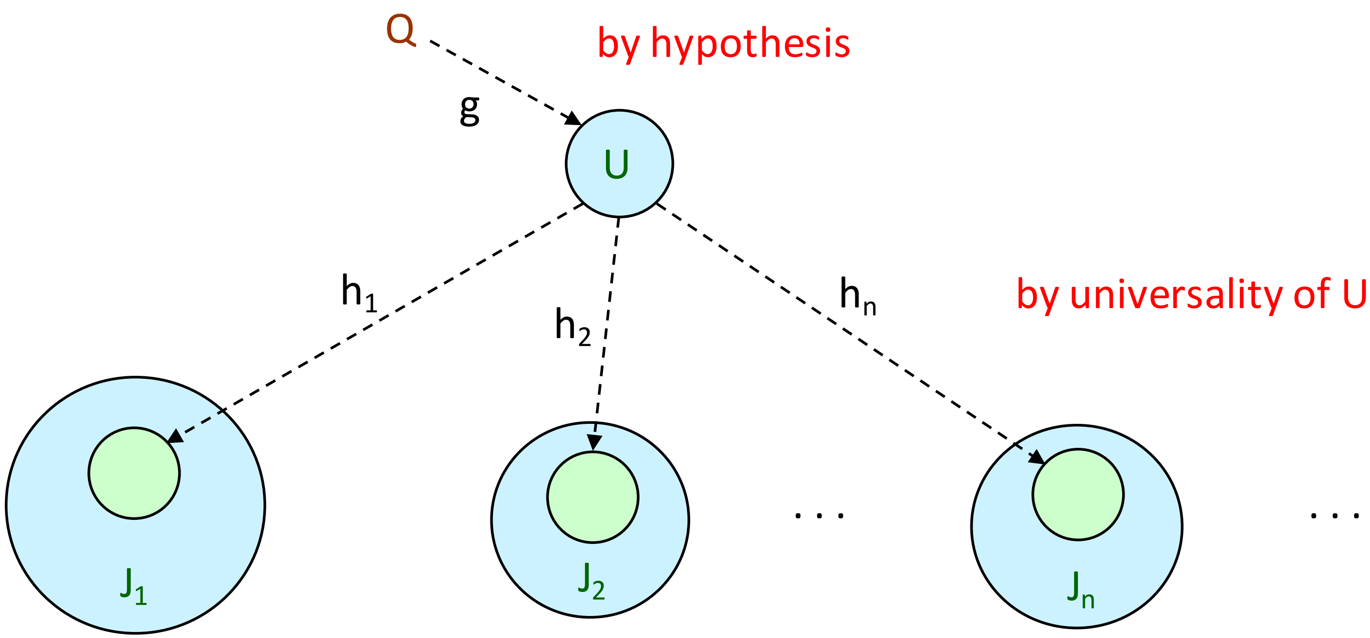
2. for each $J \in$ models(D,Σ), there exists a homomorphism h such that $h(U) \subseteq J$

Query Answering via Universal Models

Theorem: Answer(Q,D,Σ) is non-empty iff Q(U) is non-empty, where U a universal model of (D,Σ)

Proof: (⇒) Trivial since, for every J ∈ models(D,Σ), Q(J) is non-empty

(⇐) By exploiting the universality of U



∀J ∈ models(D,Σ), ∃h such that h(g(Q)) ⊆ J ⇒ ∀J ∈ models(D,Σ), Q(J) is non-empty

⇒ Answer(Q,D,Σ) is non-empty

The Chase Procedure

D

person(john)

Σ

$\forall x\ (\mathrm{Person}(x) \rightarrow \exists y\ (\mathrm{hasParent}(x,y) \wedge \mathrm{Person}(y)))$

chase(D,Σ) = D ∪

The Chase Procedure

D

person(john)

Σ

$\forall x\,(\mathrm{Person}(x) \rightarrow \exists y\,(\mathrm{hasParent}(x,y) \wedge \mathrm{Person}(y)))$

$\mathrm{chase}(D,\Sigma) = D \cup \{\mathrm{hasParent}(\mathrm{john},\perp_1), \mathrm{Person}(\perp_1)$

The Chase Procedure

D

person(john)

Σ

$\forall x\,(\text{Person}(x) \rightarrow \exists y\,(\text{hasParent}(x,y) \wedge \text{Person}(y)))$

$\text{chase}(D,\Sigma) = D \cup \{\text{hasParent}(\text{john},\perp_1), \text{Person}(\perp_1),$

$\text{hasParent}(\perp_1,\perp_2), \text{Person}(\perp_2)$

The Chase Procedure

D

person(john)

Σ

$\forall x\ (Person(x) \rightarrow \exists y\ (hasParent(x,y) \land Person(y)))$

$chase(D,\Sigma) = D \cup \{hasParent(john,\perp_1), Person(\perp_1),$

$hasParent(\perp_1,\perp_2), Person(\perp_2),$

$hasParent(\perp_2,\perp_3), Person(\perp_3)$

The Chase Procedure

D

person(john)

Σ

$$\forall x \, (Person(x) \rightarrow \exists y \, (hasParent(x,y) \wedge Person(y)))$$

$$chase(D,\Sigma) = D \cup \{hasParent(john,\perp_1), Person(\perp_1),$$

$$hasParent(\perp_1,\perp_2), Person(\perp_2),$$

$$hasParent(\perp_2,\perp_3), Person(\perp_3), \ldots$$

infinite instance

The Chase Procedure: Formal Definition

- Chase step - the building block of the chase procedure

- A rule $\sigma = \forall \mathbf{x} \forall \mathbf{y}\, (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z}\, \psi(\mathbf{x},\mathbf{z}))$ is applicable to an instance J if:

    1. There exists a homomorphism h such that $h(\varphi(\mathbf{x},\mathbf{y})) \subseteq J$

    2. There is no $g \supseteq h_{|\mathbf{x}}$ such that $g(\psi(\mathbf{x},\mathbf{z})) \subseteq J$

J = {R(a), P(a,b)}

h = {x ↦ a}

g = {x ↦ a, y ↦ b}

$\forall x\, (R(x) \rightarrow \exists y\, P(x,y))$

✗

J = {R(a), P(b,a)}

h = {x ↦ a}

✗

$\forall x\, (R(x) \rightarrow \exists y\, P(x,y))$

✓

The Chase Procedure: Formal Definition

- Chase step - the building block of the chase procedure

- A rule $\sigma = \forall \mathbf{x} \forall \mathbf{y} \, (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z} \, \psi(\mathbf{x},\mathbf{z}))$ is applicable to an instance J if:

    1. There exists a homomorphism h such that $h(\varphi(\mathbf{x},\mathbf{y})) \subseteq J$

    2. There is no $g \supseteq h_{|\mathbf{x}}$ such that $g(\psi(\mathbf{x},\mathbf{z})) \subseteq J$

- Let $J_+ = J \cup \{g(\psi(\mathbf{x},\mathbf{z}))\}$, where $g \supseteq h_{|\mathbf{z}}$ and $g(\mathbf{z})$ are "fresh" nulls not in J

- The result of applying σ to J is J_+, denoted $J[\sigma,h]J_+$ - single chase step

The Chase Procedure: Formal Definition

- A finite chase of D w.r.t. Σ is a finite sequence

$$D[\sigma_1,h_1]J_1[\sigma_2,h_2]J_2[\sigma_3,h_3]J_3 \cdots J_{n-1}[\sigma_n,h_n]J_n$$

and chase(D,Σ) is defined as the instance J_n

all applicable rules will eventually be applied

- An infinite chase of D w.r.t. Σ is a fair finite sequence

$$D[\sigma_1,h_1]J_1[\sigma_2,h_2]J_2[\sigma_3,h_3]J_3 \ldots J_{n-1}[\sigma_n,h_n]J_n \ldots$$

and chase(D,Σ) is defined as the instance $D \cup J_1 \cup J_2 \cup J_3 \cup \cdots \cup J_n \cup \cdots$

least fixpoint of a monotonic operator - chase step

Chase: A Universal Model

Theorem: chase(D,Σ) is a universal model of (D,Σ)

the result of the chase after $k \geq 0$ applications of the chase step

Proof:

- By construction, chase(D,Σ) ∈ models(D,Σ)
- It remains to show that chase(D,Σ) can be mapped into every other model of (D,Σ)
- Fix an arbitrary instance J ∈ models(D,Σ). We need to show that there exists h such that h(chase(D,Σ)) ⊆ J
- **By induction on the number of applications of the chase step, we show that for every $k \geq 0$, there exists h_k such that $h_k(\mathrm{chase}^{[k]}(D,\Sigma)) \subseteq J$, and h_k is compatible with h_{k-1}**
- Clearly, $h_0 \cup h_1 \cup \cdots \cup h_n \cup \cdots$ is a well-defined homomorphism that maps chase(D,Σ) to J
- The claim follows with $h = h_0 \cup h_1 \cup \cdots \cup h_n \cup \cdots$

Chase: Uniqueness Property

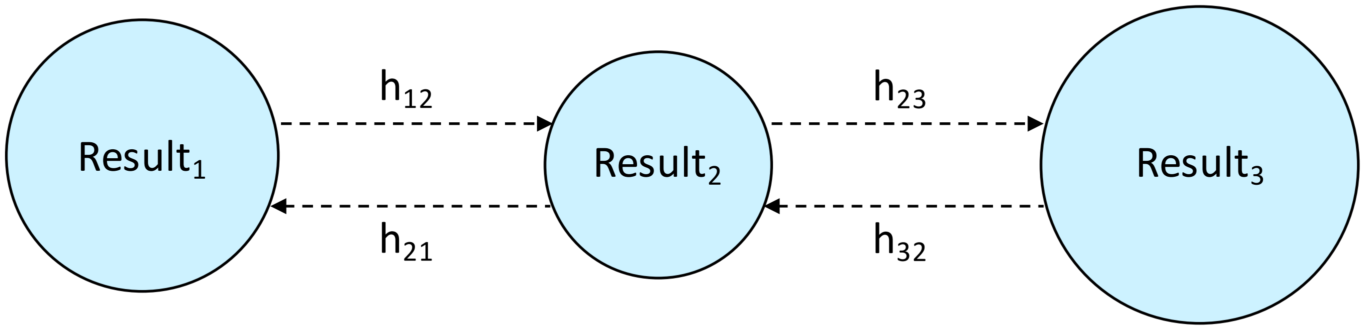
- The result of the chase is not unique - depends on the order of rule application

$D = \{P(a)\}$ $\sigma_1 = \forall x\, (P(x) \rightarrow \exists y\, R(y))$ $\sigma_2 = \forall x\, (P(x) \rightarrow R(x))$

$\text{Result}_1 = \{P(a), R(\perp), R(a)\}$ σ_1 then σ_2

$\text{Result}_2 = \{P(a), R(a)\}$ σ_2 then σ_1

- But, it is unique up to homomorphic equivalence



- Thus, it is unique for query answering purposes

Query Answering via the Chase

Theorem: Answer(Q,D,Σ) is non-empty iff Q(U) is non-empty, where U a universal model of (D,Σ)

&

Theorem: chase(D, Σ) is a universal model of (D,Σ)

$\Downarrow$

Corollary: Answer(Q,D,Σ) is non-empty iff Q(chase(D,Σ)) is non-empty

- We can tame the first dimension of infinity by exploiting the chase procedure
- What about the second dimension of infinity? - the chase may be infinite

Can we tame the second dimension of infinity?

Undecidability of OBQA

arbitrary existential rules

Theorem: OBQA($\exists$**RULES**) is undecidable

Proof Idea : By simulating a deterministic Turing machine with an empty tape.

Encode the computation of a DTM M with an empty tape using a database D, a set Σ of

existential rules, and a Boolean CQ Q such that Answer(Q,D,Σ) is non-empty iff M accepts

Gaining Decidability

By restricting the database

- Answer(Q,{Start(c)},Σ) is non-empty iff the DTM M accepts
- The problem is undecidable even for singleton databases
- No much to do in this direction

By restricting the query language

- Answer(Q :- Accept(x),D,Σ) is non-empty iff the DTM M accepts
- The problem is undecidable already for atomic queries
- No much to do in this direction

By restricting the ontology language

- Achieve a good trade-off between expressive power and complexity
- Field of intense research
- Any ideas?

Source of Non-termination

D

person(john)

Σ

$\forall x\,(Person(x) \rightarrow \exists y\,(hasParent(x,y) \wedge Person(y)))$

$chase(D,\Sigma) = D \cup \{hasParent(john,\bot_1), Person(\bot_1),$

$hasParent(\bot_1,\bot_2), Person(\bot_2),$

$hasParent(\bot_2,\bot_3), Person(\bot_3), \ldots$

1. Existential quantification

2. Recursive definitions

infinite instance

Termination of the Chase

- Drop the existential quantification
  - We obtain the class of **full existential rules**
  - Very close to Datalog

- Drop the recursive definitions
  - We obtain the class of **acyclic existential rules**
  - Also known as non-recursive existential rules

Our Simple Example

D

person(john)

Σ

$$\forall x\,(\mathit{Person}(x) \rightarrow \exists y\,(\mathit{hasParent}(x,y) \wedge \mathit{Person}(y)))$$

$$\begin{aligned}\mathit{chase}(D,\Sigma) = D \cup \{&\mathit{hasParent}(john,\perp_1), \mathit{Person}(\perp_1),\\ &\mathit{hasParent}(\perp_1,\perp_2), \mathit{Person}(\perp_2),\\ &\mathit{hasParent}(\perp_2,\perp_3), \mathit{Person}(\perp_3), \ldots\end{aligned}$$

Existential quantification & recursive definitions

are key features for modelling ontologies

Key Question

We need classes of existential rules such that

- Existential quantification and recursive definition coexist

  $\Rightarrow$ the chase may be infinite

- BOBQA is decidable, and tractable w.r.t. the data complexity

$\Downarrow$

Tame the infinite chase:

Deal with infinite structures without explicitly building them

Linear Existential Rules

- A linear existential rule is an existential rule of the form

$$\forall \mathbf{x} \forall \mathbf{y} \, (\underbrace{P(\mathbf{x},\mathbf{y})}_{\text{single atom}} \rightarrow \exists \mathbf{z} \, \psi(\mathbf{x},\mathbf{z}))$$

- We denote **LINEAR** the class of linear existential rules

- But, is this a reasonable ontology language?

W3C Recommendation

3 OWL 2 QL

The OWL 2 QL profile is designed so that sound and complete query answering is in LOGSPACE (more precisely, in AC^0) with respect to the size of the data (assertions), while providing many of the main features necessary to express conceptual models such as UML class diagrams and ER diagrams. In particular, this profile contains the intersection of RDFS and OWL 2 DL. It is designed so that data (assertions) that is stored in a standard relational database system can be queried through an ontology via a simple rewriting mechanism, i.e., by rewriting the query into an SQL query that is then answered by the RDBMS system, without any changes to the data.

OWL 2 QL is based on the DL-Lite family of description logics [*DL-Lite*]. Several variants of DL-Lite have been described in the literature, and DL-$Lite_R$ provides the logical underpinning for OWL 2 QL. DL-$Lite_R$ does not require the unique name assumption (UNA), since making this assumption would have no impact on the semantic consequences of a DL-$Lite_R$ ontology. More expressive variants of DL-Lite, such as DL-$Lite_A$, extend DL-$Lite_R$ with functional properties, and these can also be extended with keys; however, for query answering to remain in LOGSPACE, these extensions require UNA and need to impose certain global restrictions on the interaction between properties used in different types of axiom. Basing OWL 2 QL on DL-$Lite_R$ avoids practical problems involved in the explicit axiomatization of UNA. Other variants of DL-Lite can also be supported on top of OWL 2 QL, but may require additional restrictions on the structure of ontologies.

3.1 Feature Overview

OWL 2 QL is defined not only in terms of the set of supported constructs, but it also restricts the places in which these constructs are allowed to occur. The allowed usage of constructs in class expressions is summarized in Table 1.

Table 1. Syntactic Restrictions on Class Expressions in OWL 2 QL

Subclass Expressions	Superclass Expressions
a class existential quantification (**ObjectSomeValuesFrom**) where the class is limited to *owl:Thing* existential quantification to a data range (**DataSomeValuesFrom**)	a class intersection (**ObjectIntersectionOf**) negation (**ObjectComplementOf**) existential quantification to a class (**ObjectSomeValuesFrom**) existential quantification to a data range (**DataSomeValuesFrom**)

OWL 2 QL supports the following axioms, constrained so as to be compliant with the mentioned restrictions on class expressions:

- subclass axioms (**SubClassOf**)
- class expression equivalence (**EquivalentClasses**)
- class expression disjointness (**DisjointClasses**)
- inverse object properties (**InverseObjectProperties**)
- property inclusion (**SubObjectPropertyOf** not involving property chains and **SubDataPropertyOf**)
- property equivalence (**EquivalentObjectProperties** and **EquivalentDataProperties**)
- property domain (**ObjectPropertyDomain** and **DataPropertyDomain**)
- property range (**ObjectPropertyRange** and **DataPropertyRange**)
- disjoint properties (**DisjointObjectProperties** and **DisjointDataProperties**)

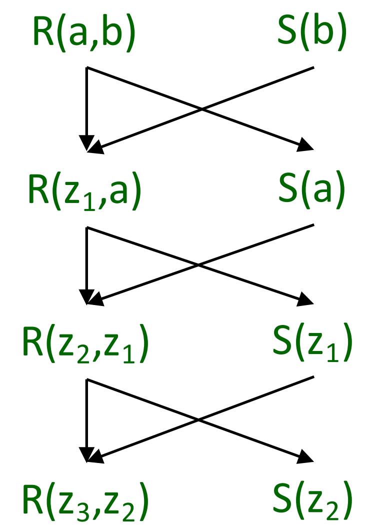
https://www.w3.org/TR/owl2-profiles/#OWL_2_QL

Chase Graph

The chase can be naturally seen as a graph - chase graph

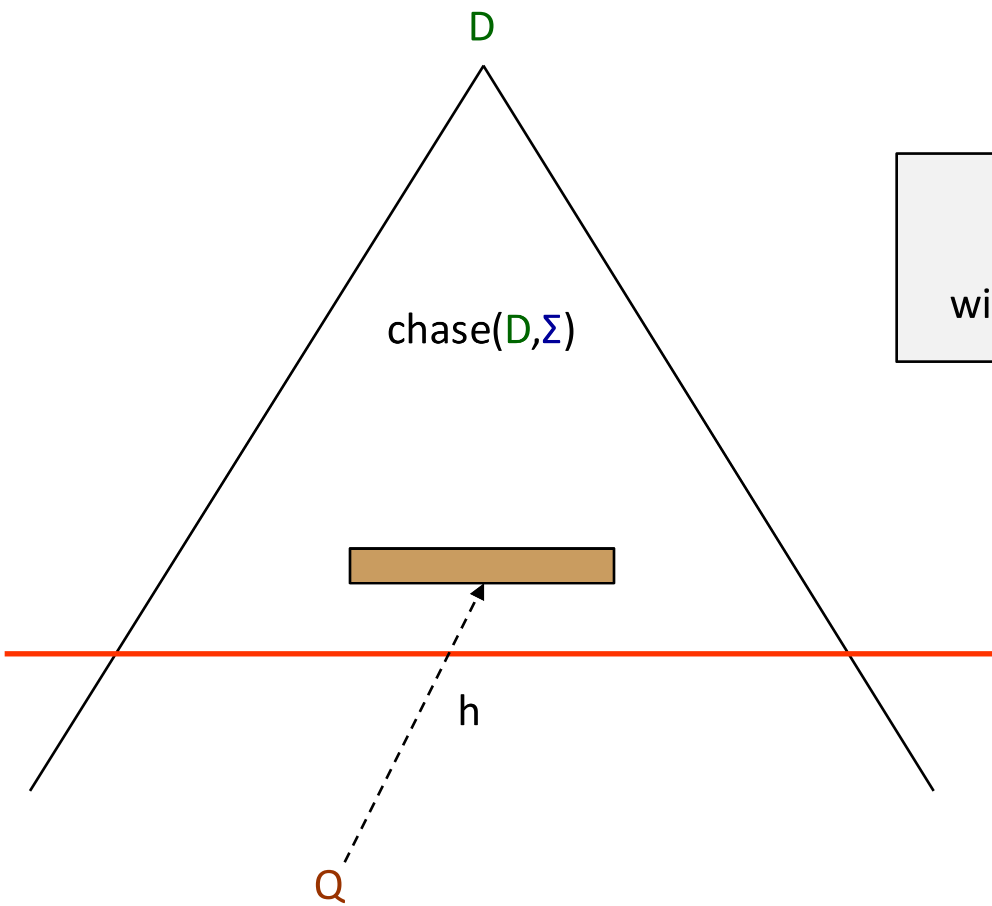
$D = \{R(a,b),\ S(b)\}$

$$\Sigma = \begin{cases} \forall x \forall y\ (R(x,y) \land S(y) \rightarrow \exists z\ R(z,x)) \\ \forall x \forall y\ (R(x,y) \rightarrow S(x)) \end{cases}$$



For **LINEAR** the chase graph is a forest

Bounded Derivation-Depth Property



depth k that does not depend on D

For **LINEAR**, $k = |Q| \cdot m$

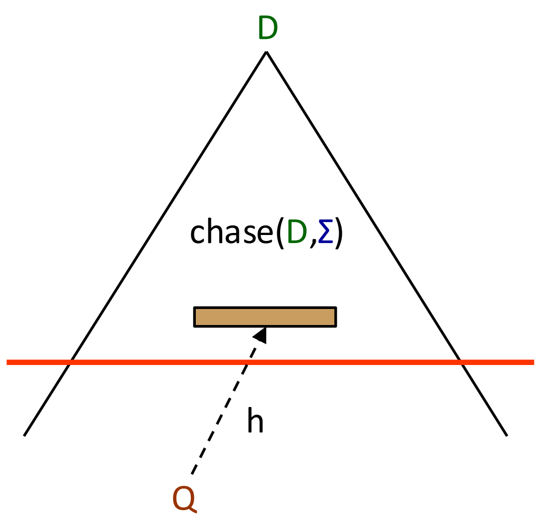
with $m = |sch(\Sigma)| \cdot (2 \cdot \text{maxarity})^{\text{maxarity}}$

chase graph up to depth k

$Q(\text{chase}(D,\Sigma))$ is non-empty $\Rightarrow$ $Q(\text{chase}^k(D,\Sigma))$ is non-empty

The Blocking Algorithm for **LINEAR**

Theorem: BOBQA[Σ,Q](**LINEAR**) is in PTIME for a fixed set Σ, and a Boolean CQ Q



$$k = |Q| \cdot |sch(\Sigma)| \cdot (2 \cdot \text{maxarity})^{\text{maxarity}}$$

The Blocking Algorithm for **LINEAR**

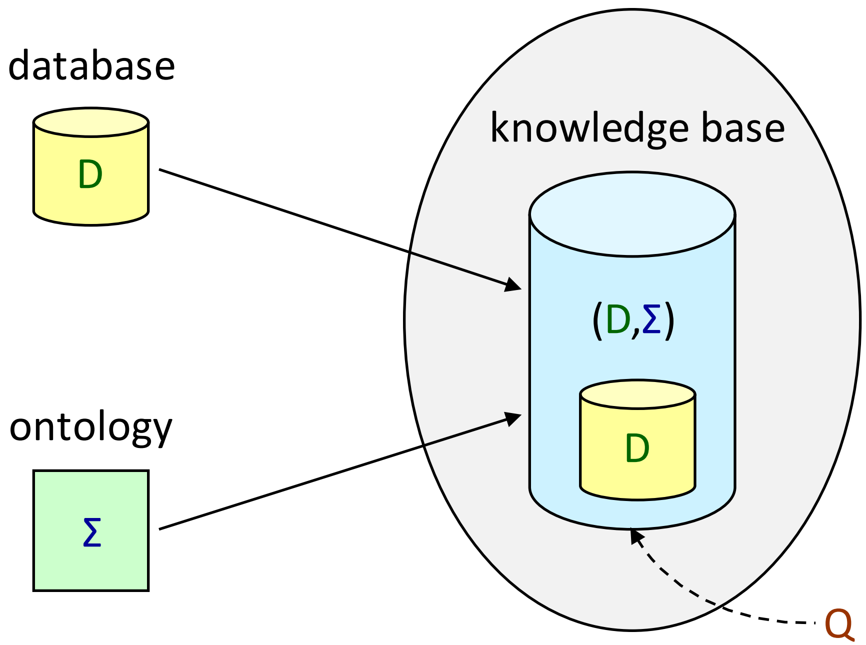
Theorem: BOBQA[Σ,Q](**LINEAR**) is in PTIME for a fixed set Σ, and a Boolean CQ Q

but, we can do better

Theorem: BOBQA[Σ,Q](**LINEAR**) is in LOGSPACE for a fixed set Σ, and a Boolean CQ Q

Scalability in OBQA

Exploit standard RDBMSs - efficient technology for answering CQs



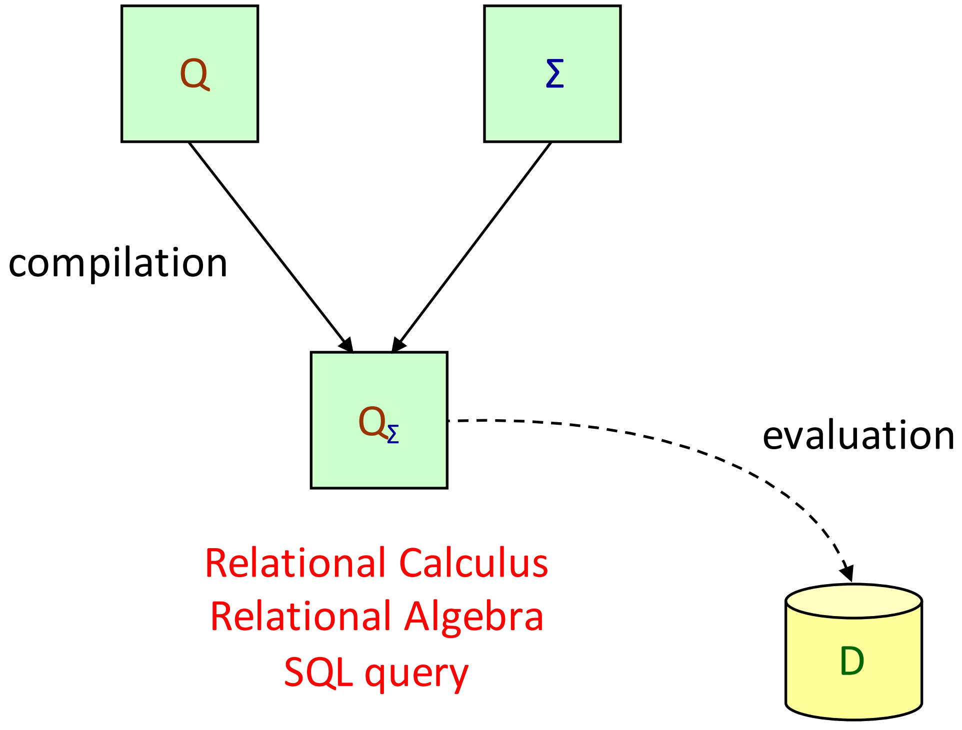
But in the OBQA setting

we have to query a

knowledge base, not just a

relational database

Query Rewriting



Relational Calculus
Relational Algebra
SQL query
...

for every database D, Answer(Q,D,Σ) is non-empty iff Q_Σ(D) is non-empty

Query Rewriting: Formal Definition

Consider a class of existential rules **L**, and a query language **Q**.

BOBQA(**L**) is **Q**-rewritable if, for every $\Sigma \in$ **L** and Boolean CQ Q,

we can construct a query $Q_{\Sigma} \in$ **Q** such that,

for every database D, Answer(Q,D,Σ) is non-empty iff Q_{Σ}(D) is non-empty

NOTE: The construction of Q_{Σ} is database-independent

An Example

$\Sigma = \{\forall x\, (P(x) \rightarrow T(x)),\quad \forall x \forall y\, (R(x,y) \rightarrow S(x))\}$

Q :- S(x), U(x,y), T(y)

Q_Σ = {Q :- S(x), U(x,y), T(y),

Q_1 :- S(x), U(x,y), P(y),

Q_2 :- R(x,z), U(x,y), T(y),

Q_3 :- R(x,z), U(x,y), P(y)}

An Example

$\Sigma = \{\forall x \forall y\, (R(x,y) \land P(y) \rightarrow P(x))\}$

Q :- $P(c)$

$$
\begin{aligned}
Q_\Sigma = \{&Q \text{ :- } P(c),\\
&\quad Q_1 \text{ :- } R(c,y_1),\ P(y_1),\\
&\qquad Q_2 \text{ :- } R(c,y_1),\ R(y_1,y_2),\ P(y_2),\\
&\qquad\quad Q_3 \text{ :- } R(c,y_1),\ R(y_1,y_2),\ R(y_2,y_3),\ P(y_3),\\
&\ldots \}
\end{aligned}
$$

- This cannot be written as a finite first-order query
- It can be written as Q :- $R(c,x),\ R^*(x,y),\ P(y)$, but transitive closure is not FO-expressible

Query Rewriting for **LINEAR**

union of conjunctive queries

Theorem: **LINEAR** is UCQ-rewritable

$\Downarrow$

Theorem: BOBQA[Σ,Q](**LINEAR**) is in LOGSPACE for a fixed set Σ, and a Boolean CQ Q

…it also tells us that for answering CQs in the presence of **LINEAR** ontologies,

we can exploit standard database technology

UCQ-Rewritings

- The standard algorithm for computing UCQ-rewritings performs an exhaustive application of the following two steps:
    1. Rewriting
    2. Minimization

- We are going to see the version of the algorithm that assumes normalized existential rules, where only one atom appears in the head

Normalization Procedure

$$\forall \mathbf{x} \forall \mathbf{y}\, (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z}\, (P_1(\mathbf{x},\mathbf{z}) \wedge \cdots \wedge P_n(\mathbf{x},\mathbf{z})))$$

$$\downarrow$$

$$\forall \mathbf{x} \forall \mathbf{y}\, (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z}\, \mathrm{Auxiliary}(\mathbf{x},\mathbf{z}))$$

$$\forall \mathbf{x} \forall \mathbf{z}\, (\mathrm{Auxiliary}(\mathbf{x},\mathbf{z}) \rightarrow P_1(\mathbf{x},\mathbf{z}))$$

$$\forall \mathbf{x} \forall \mathbf{z}\, (\mathrm{Auxiliary}(\mathbf{x},\mathbf{z}) \rightarrow P_2(\mathbf{x},\mathbf{z}))$$

…

$$\forall \mathbf{x} \forall \mathbf{z}\, (\mathrm{Auxiliary}(\mathbf{x},\mathbf{z}) \rightarrow P_n(\mathbf{x},\mathbf{z}))$$

NOTE : Linearity is preserved, and we obtain an equivalent ontology w.r.t. query answering

UCQ-Rewritings

- The standard algorithm for computing UCQ-rewritings performs an exhaustive application of the following two steps:
    1. Rewriting
    2. Minimization

- We are going to see the version of the algorithm that assumes normalized existential rules, where only one atom appears in the head

Rewriting Step

$\Sigma = \{\forall x \forall y\ (\text{project}(x) \wedge \text{inArea}(x,y) \rightarrow \exists z\ \text{hasCollaborator}(z,y,x))\}$

Q :- hasCollaborator(u,db,v)

$g = \{x \mapsto v, y \mapsto db, z \mapsto u\}$

hasCollaborator(u,db,v)

Thus, we can simulate a chase step by applying a backward resolution step

Q_Σ = {Q :- hasCollaborator(u,db,v),

Q_1 :- project(v), inArea(v,db)}

Unsound Rewritings

$\Sigma = \{\forall x \forall y\, (project(x) \wedge inArea(x,y) \rightarrow \exists z\, hasCollaborator(z,y,x))\}$

Q :- $hasCollaborator(c,db,v)$

(c is a constant)

$g = \{x \mapsto v, y \mapsto db, z \mapsto c\}$

$hasCollaborator(c,db,v)$

After applying the rewriting step we obtain the following UCQ

$Q_\Sigma = \{Q$:- $hasCollaborator(c,db,v),$

Q_1 :- $project(v), inArea(v,db)\}$

Unsound Rewritings

$\Sigma = \{\forall x \forall y\ (\text{project}(x) \wedge \text{inArea}(x,y) \rightarrow \exists z\ \text{hasCollaborator}(z,y,x))\}$

Q :- hasCollaborator(c,db,v)

$$Q_\Sigma = \{Q \text{ :- hasCollaborator(c,db,v)}, \quad Q_1 \text{ :- project(v), inArea(v,db)}\}$$

- Consider the database $D = \{\text{project(a), inArea(a,db)}\}$
- Clearly, $Q_\Sigma(D)$ is non-empty
- However, Answer(Q,D,Σ) is empty since there is no way to obtain an atom of the form hasCollaborator(c,db,_) during the chase

Unsound Rewritings

$\Sigma = \{\forall x \forall y \, (project(x) \land inArea(x,y) \rightarrow \exists z \, hasCollaborator(z,y,x))\}$

Q :- hasCollaborator(c,db,v)

$Q_{\Sigma} = \{$Q :- hasCollaborator(c,db,v),

Q_1 :- project(v), inArea(v,db)$\}$

the information about the constant c in the original query is lost after the application of the rewriting step since c is unified with an ∃-variable

Unsound Rewritings

$\Sigma = \{\forall x \forall y\ (\text{project}(x) \land \text{inArea}(x,y) \rightarrow \exists z\ \text{hasCollaborator}(z,y,x))\}$

Q :- hasCollaborator(v,db,v)

$g = \{x \mapsto v, y \mapsto db, z \mapsto v\}$

hasCollaborator(v,db,v)

After applying the rewriting step we obtain the following UCQ

Q_Σ = {Q :- hasCollaborator(v,db,v),

Q_1 :- project(v), inArea(v,db)}

Unsound Rewritings

$\Sigma = \{\forall x \forall y\ (\text{project}(x) \wedge \text{inArea}(x,y) \rightarrow \exists z\ \text{hasCollaborator}(z,y,x))\}$

Q :- hasCollaborator(c,db,v)

$Q_\Sigma = \{Q$:- hasCollaborator(v,db,v),

Q_1 :- project(v), inArea(v,db)$\}$

- Consider the database $D = \{$project(a), inArea(a,db)$\}$
- Clearly, $Q_\Sigma(D)$ is non-empty
- However, Answer(Q,D,Σ) is empty since there is no way to obtain an atom of the form hasCollaborator(t,db,t) during the chase

Unsound Rewritings

$\Sigma = \{\forall x \forall y\ (project(x) \wedge inArea(x,y) \rightarrow \exists z\ hasCollaborator(z,y,x))\}$

Q :- hasCollaborator(c,db,v)

$Q_\Sigma = \{Q$:- hasCollaborator(c,db,v),

Q_1 :- project(v), inArea(v,db)$\}$

the fact that v in the original query participates in a join is lost after the application of the rewriting step since v is unified with an ∃-variable

Applicability Condition

Consider a Boolean CQ Q, an atom α in Q, and a (normalized) rule σ.

We say that σ is applicable to α if the following conditions hold:

1. head(σ) and α unify via h

2. For every variable x in head(σ):
    1. If h(x) is a constant, then x is a $\forall$-variable
    2. If h(x) = h(y), where y is a shared variable of α, then x is a $\forall$-variable

3. If x is an $\exists$-variable of head(σ), and y is a variable in head(σ) such that x $\neq$ y, then h(x) $\neq$ h(y)

...but, although it is crucial for soundness, may destroy completeness

Incomplete Rewritings

$\Sigma = \{\forall x \forall y\ (\mathrm{project}(x) \wedge \mathrm{inArea}(x,y) \rightarrow \exists z\ \mathrm{hasCollaborator}(z,y,x)),$

$\forall x \forall y \forall z\ (\mathrm{hasCollaborator}(x,y,z) \rightarrow \mathrm{collaborator}(x))\}$

Q :- hasCollaborator(u,v,w), collaborator(u))

Q_Σ = {Q :- hasCollaborator(u,v,w), collaborator(u),

Q_1 :- hasCollaborator(u,v,w), hasCollaborator(u,v',w')

- Consider the database D = {project(a), inArea(a,db)}
- Clearly, Q over chase(D,Σ) = D ∪ {hasCollaborator(z,db,a), collaborator(z)} is non-empty
- However, Q_Σ(D) is empty

Incomplete Rewritings

$\Sigma = \{\forall x \forall y\ (\text{project}(x) \wedge \text{inArea}(x,y) \rightarrow \exists z\ \text{hasCollaborator}(z,y,x)),$

$\forall x \forall y \forall z\ (\text{hasCollaborator}(x,y,z) \rightarrow \text{collaborator}(x))\}$

Q :- hasCollaborator(u,v,w), collaborator(u))

Q_Σ = {Q :- hasCollaborator(u,v,w), collaborator(u),

Q_1 :- hasCollaborator(u,v,w), hasCollaborator(u,v',w')

Q_2 :- project(u), inArea(u,v)

but, we cannot obtain the last query due to the applicablity condition

Incomplete Rewritings

$\Sigma = \{\forall x \forall y\, (\text{project}(x) \wedge \text{inArea}(x,y) \rightarrow \exists z\, \text{hasCollaborator}(z,y,x)),$

$\forall x \forall y \forall z\, (\text{hasCollaborator}(x,y,z) \rightarrow \text{collaborator}(x))\}$

Q :- hasCollaborator(u,v,w), collaborator(u))

Q_Σ = {Q :- hasCollaborator(u,v,w), collaborator(u),

Q_1 :- hasCollaborator(u,v,w), hasCollaborator(u,v',w')

Q_2 :- hasCollaborator(u,v,w) - by minimization

Q_3 :- project(w), inArea(w,v) - by rewriting

$Q_\Sigma(D)$ is non-empty, where D = {project(a), inArea(a,db)}

UCQ-Rewritings

- The standard algorithm for computing UCQ-rewritings performs an exhaustive application of the following two steps:
    1. Rewriting
    2. Minimization

- We are going to see the version of the algorithm that assumes normalized existential rules, where only one atom appears in the head

The Rewriting Algorithm

```
Q_Σ := {Q}
repeat
    Q_aux := Q_Σ
    foreach disjunct q of Q_aux do
    //Rewriting Step
        foreach atom α in q do
            foreach rule σ in Σ do
                if σ is applicable to α then
                    q_rew := rewrite(q,α,σ)   //we resolve α using σ
                    if q_rew does not appear in Q_Σ (modulo variable renaming) then
                        Q_Σ := Q_Σ ∪ {q_rew}
    //Minimization Step
        foreach pair of atoms α,β in q that unify do
            q_min := minimize(q,α,β)   //we apply the MGU of α and β on q
            if q_min does not appear in Q_Σ (modulo variable renaming) then
                        Q_Σ := Q_Σ ∪ {q_min}
until Q_aux = Q_Σ
return Q_Σ
```

Termination

Theorem: The rewriting algorithm terminates under **LINEAR**

Proof Idea:

- **Key observation:** the size of each partial rewriting is at most the size of the given CQ Q
- Thus, each partial rewriting can be transformed into an equivalent query that contains at most ($|Q| \cdot$ maxarity) variables
- The number of queries that can be constructed using a finite number of predicates and a finite number of variables is finite
- Therefore, only finitely many partial rewritings can be constructed - in general, exponentially many

Size of the Rewriting

- Ideally, we would like to construct UCQ-rewritings of polynomial size
- But, the standard rewritng algorithm produces rewritings of exponential size
- Can we do better? **NO!!!**

$\Sigma = \{\forall x\, (R_k(x) \rightarrow P_k(x))\}$ for $k \in \{1,\ldots,n\}$ $\quad Q$:- $P_1(x), \ldots, P_n(x)$

$$Q \text{ :- } P_1(X), \ldots, P_n(X)$$

$$P_1(X) \vee R_1(X) \qquad\qquad P_n(X) \vee R_n(X)$$

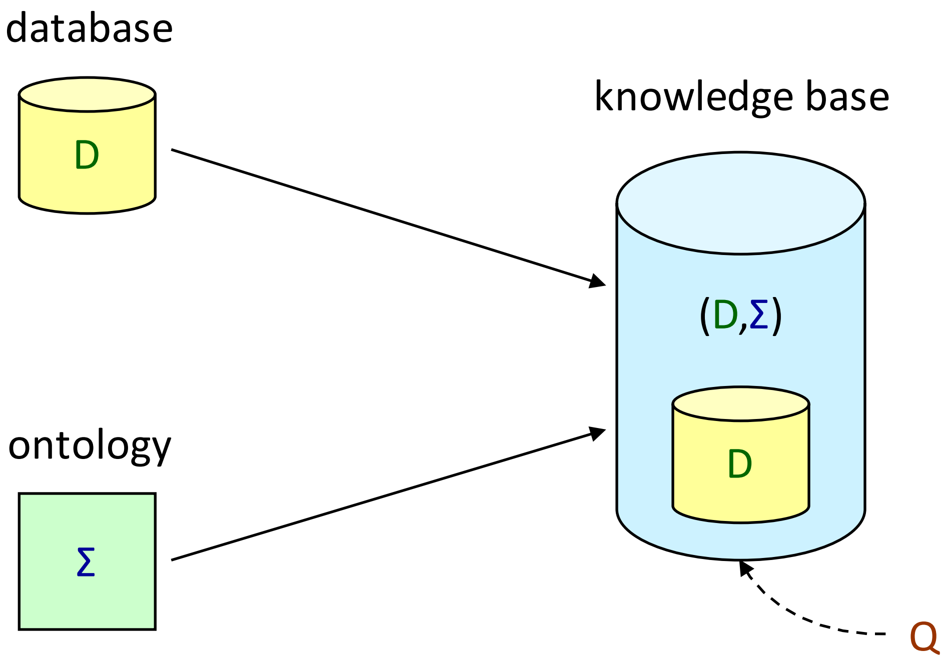
thus, we need to consider 2^n disjuncts

Size of the Rewriting

- Ideally, we would like to construct UCQ-rewritings of polynomial size
- But, the standard rewritng algorithm produces rewritings of exponential size
- Can we do better? **NO!!!**

  - Although the standard rewriting algorithm is worst-case optimal, it can be significantly improved
  - Optimization techniques can be applied in order to compute efficiently small rewritings - field of intense research

Recap



existential rules

$\forall \mathbf{x} \forall \mathbf{y}\, (\varphi(\mathbf{x},\mathbf{y}) \rightarrow \exists \mathbf{z}\, \psi(\mathbf{x},\mathbf{z}))$

conjunctive query

$Q(\mathbf{x})\ \text{:-}\ R_1(\mathbf{v_1}),\ldots,R_m(\mathbf{v_m})$

in general, this is an undecidable problem, but well-behaved ontology languages exists - **LINEAR**