

Datalog[±]: A Family of Languages for Ontology Querying

Georg Gottlob

Department of Computer Science
University of Oxford

joint work with [Andrea Calì](#), [Thomas Lukasiewicz](#), [Marco Manna](#), [Andreas Pieris](#) et al.

1st Motivation: Ontological Query Answering

- Input: a **knowledge base** $K = \langle D, \Sigma \rangle$
a Boolean **conjunctive query** Q

Question: $K \models Q$ (or, equivalently, $D \cup \Sigma \models Q$)

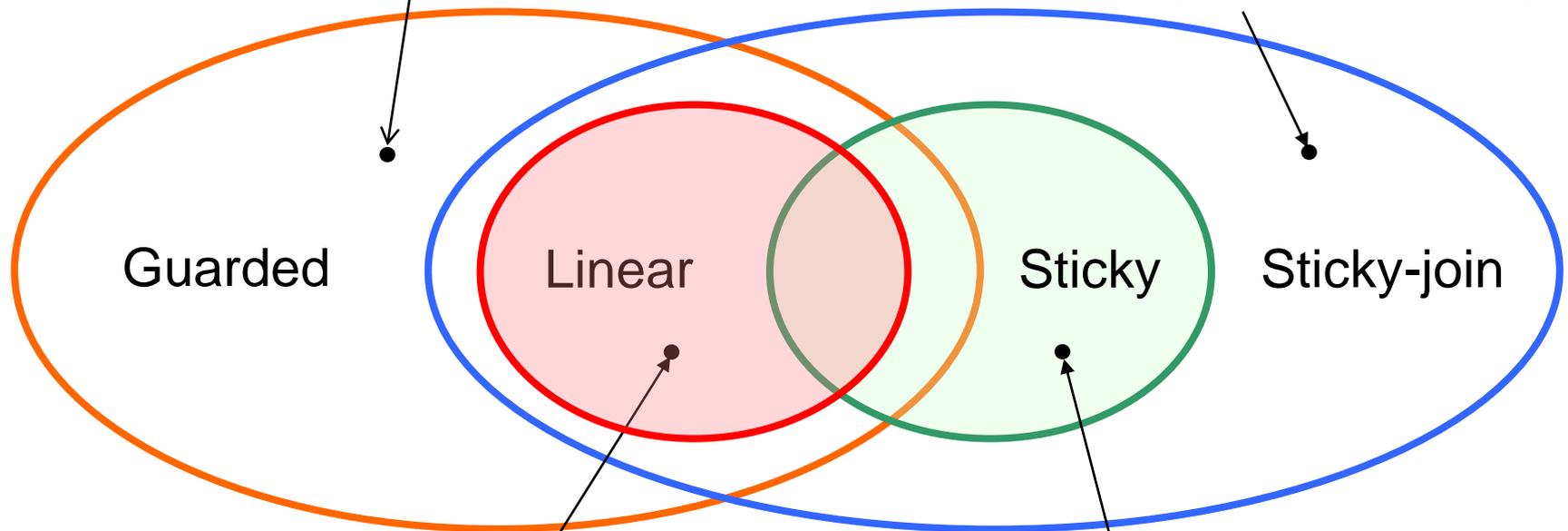
- **Old database problem** – CQ answering over incomplete databases

2nd Motivation: Decidable FO fragments (TGDs)

Good data complexity!

$$\forall X \forall Y \forall Z R(X, Y, Z), S(Y), P(X, Z) \rightarrow \exists W Q(X, W)$$

$$\forall X \forall Y \forall Z R(X, Y), S(Y, Z, Z) \rightarrow \exists W P(Y, W)$$



$$\forall X \forall Y R(X, X, Y) \rightarrow \exists Z R(Y, Y, Z)$$

$$\forall X \forall Y \forall Z S(X, Y), R(Y, Z) \rightarrow \exists W P(Y, W)$$

2nd Motivation: Decidable FO fragments (TGDs)

Good data complexity!

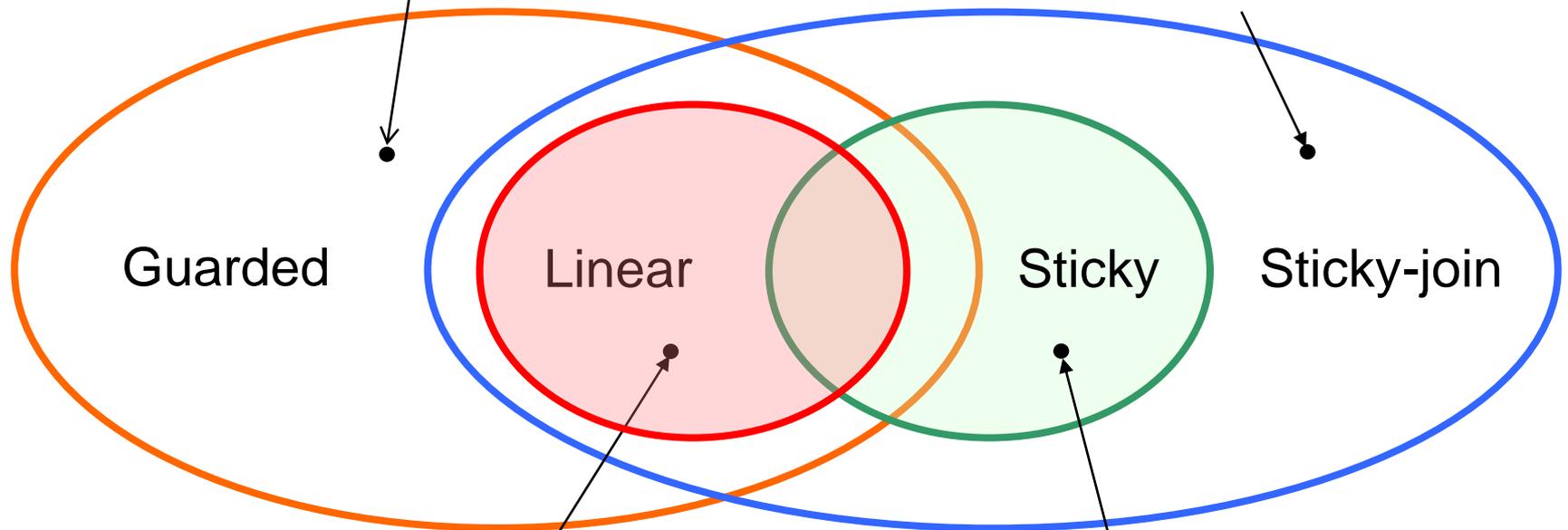
Query: $R(Y, Y, Z), P(Z, W)$

Add rule:

$\forall Y \forall Z \forall W R(Y, Y, Z), P(Z, W) \rightarrow \perp$

$\forall X \forall Y \forall Z R(X, Y, Z), S(Y), P(X, Z) \rightarrow \exists W Q(X, W)$

$\forall X \forall Y \forall Z R(X, Y), S(Y, Z, Z) \rightarrow \exists W P(Y, W)$

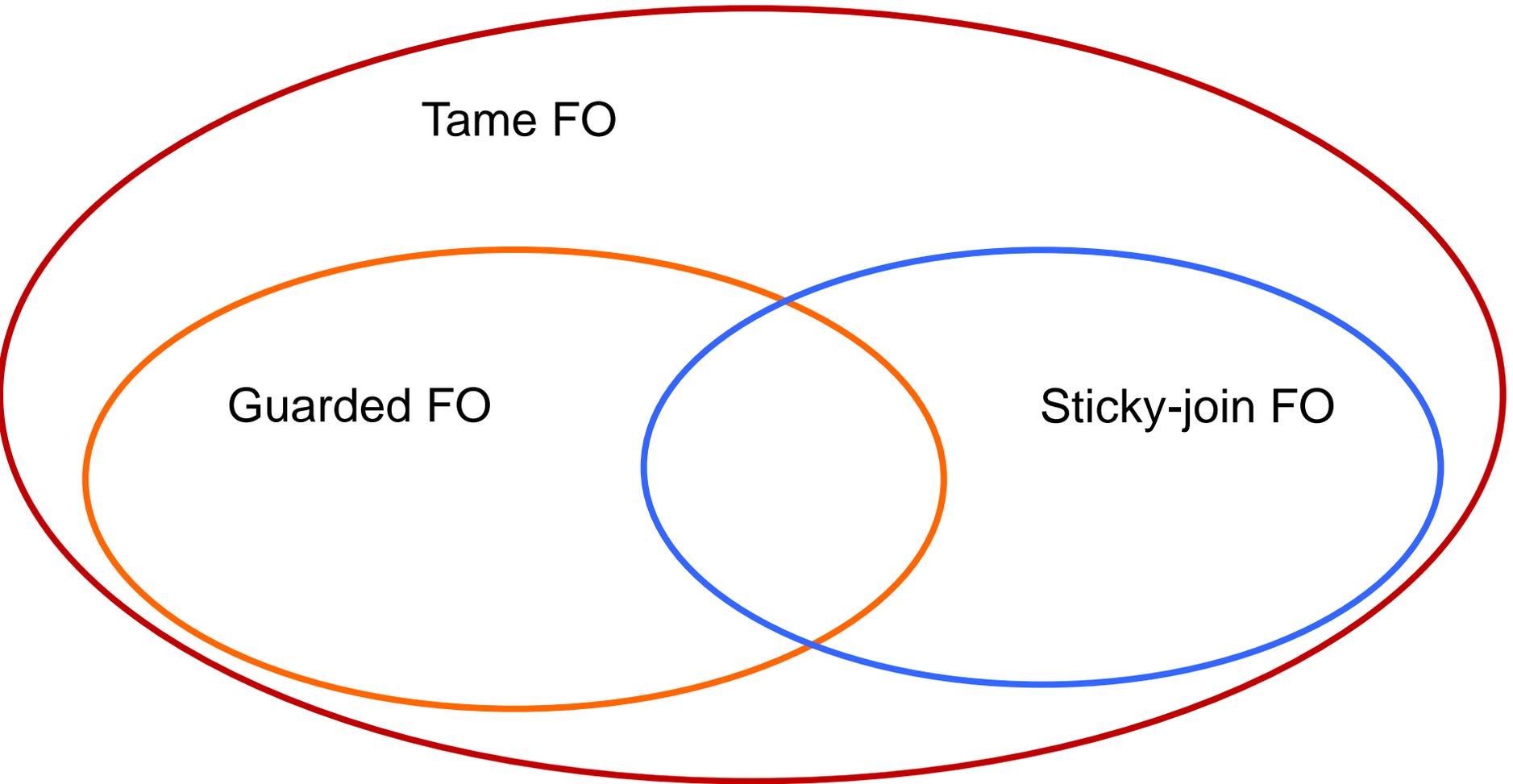


$\forall X \forall Y R(X, X, Y) \rightarrow \exists Z R(Y, Y, Z)$

$\forall X \forall Y \forall Z S(X, Y), R(Y, Z) \rightarrow \exists W P(Y, W)$

3rd Motivation: Decidable FO fragments

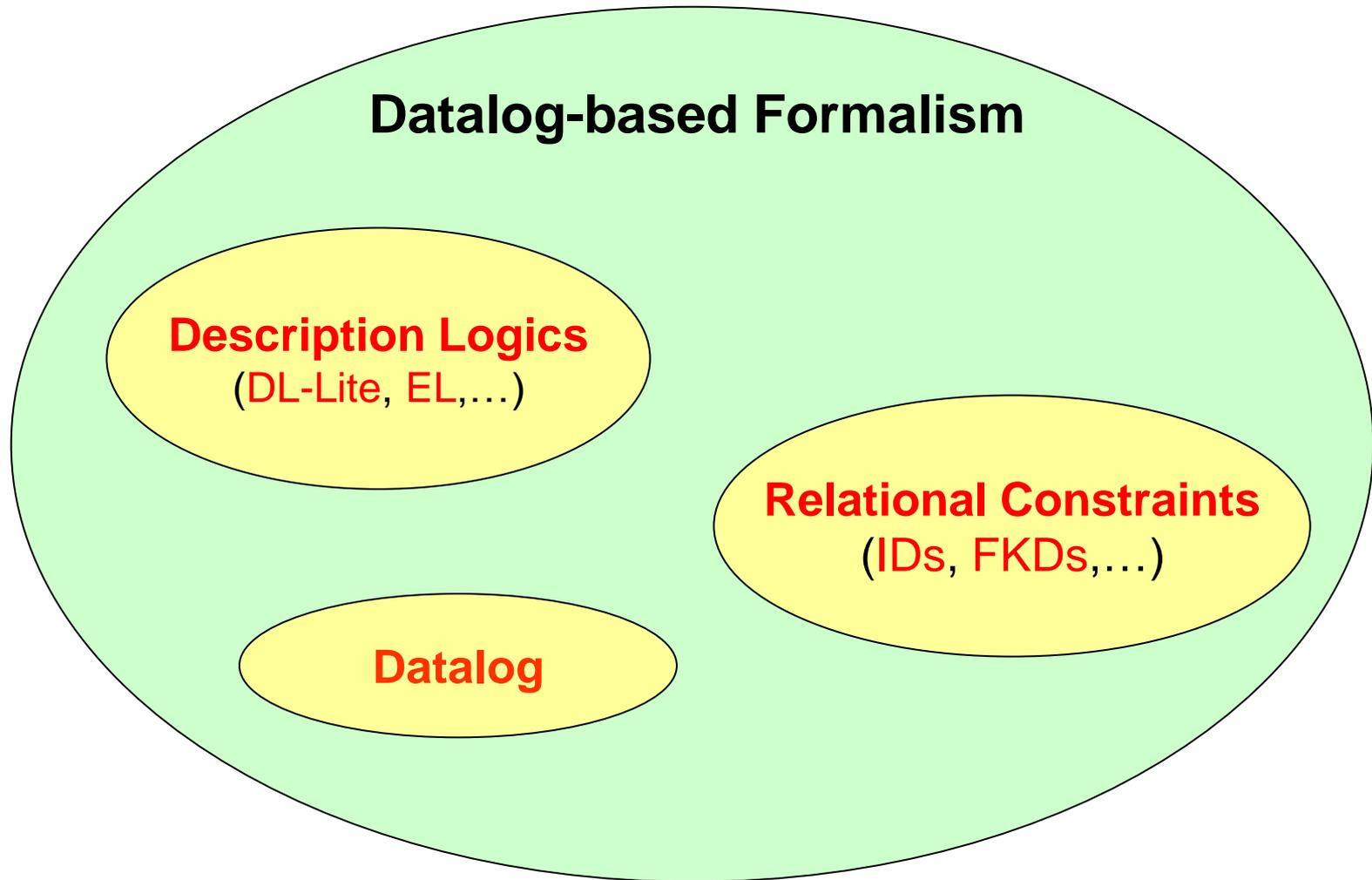
(current work)



(Some) Known Results

Σ	Data Complexity	Combined Complexity
DL-Lite	in AC_0 [Calvanese et al., JAR 07]	NP-complete [Calvanese et al., JAR 07]
EL + ELH	PTIME-complete [Rosati, DL 07]	NP-complete [Rosati, DL 07]
Horn-SHIQ	PTIME-complete [Eiter et al., JELIA 08]	EXPTIME-complete [Eiter et al., JELIA 08]
IDs + FKDs	in AC_0 [Calì et al., IJCAI 03]	PSPACE-complete [Johnson & Klug, JCSS 84]
Datalog	PTIME-complete	EXPTIME-complete

Our Goal



... without losing **tractable data complexity**

Ontological Reasoning and Datalog

DL Assertion	Datalog Rule
<p>Concept Inclusion $emp \sqsubseteq person$</p> <p>Concept Product $sen\text{-}emp \times emp \sqsubseteq moreThan$</p> <p>(Inverse) Role Inclusion $reports^- \sqsubseteq mgr$</p> <p>Role Transitivity $trans(mgr)$</p>	<p>$emp(X) \rightarrow person(X)$</p> <p>$sen\text{-}emp(X), emp(Y) \rightarrow moreThan(X, Y)$</p> <p>$reports(X, Y) \rightarrow mgr(Y, X)$</p> <p>$mgr(X, Y), mgr(Y, Z) \rightarrow mgr(X, Z)$</p>

Ontological Reasoning and Datalog

DL Assertion	Datalog Rule
<p>Concept Inclusion $emp \sqsubseteq person$</p> <p>Concept Product $sen\text{-}emp \times emp \sqsubseteq moreThan$</p> <p>(Inverse) Role Inclusion $reports^- \sqsubseteq mgr$</p> <p>Role Transitivity $trans(mgr)$</p>	<p>$emp(X) \rightarrow person(X)$</p> <p>$sen\text{-}emp(X), emp(Y) \rightarrow moreThan(X, Y)$</p> <p>$reports(X, Y) \rightarrow mgr(Y, X)$</p> <p>$mgr(X, Y), mgr(Y, Z) \rightarrow mgr(X, Z)$</p>
<p>Participation $emp \sqsubseteq \exists report$</p> <p>Disjointness $emp \sqcap customer \sqsubseteq \perp$</p> <p>Functionality $funct(reports)$</p>	<p>$emp(X) \rightarrow \exists Y report(X, Y)$</p> <p>$emp(X), customer(X) \rightarrow \perp$</p> <p>$reports(X, Y), reports(X, Z) \rightarrow Y = Z$</p>

Datalog[±]

- **Datalog**: $\forall \mathbf{X} \forall \mathbf{Y} \Phi(\mathbf{X}, \mathbf{Y}) \rightarrow R(\mathbf{X})$
- **Extend Datalog** by allowing in the head:
 - Existential quantification (\exists) - **TGDs**: $\forall \mathbf{X} \forall \mathbf{Y} \Phi(\mathbf{X}, \mathbf{Y}) \rightarrow \exists \mathbf{Z} \Psi(\mathbf{X}, \mathbf{Z})$
 - Equality atoms (=) - **EGDs**: $\forall \mathbf{X} \Phi(\mathbf{X}) \rightarrow X_i = X_j$
 - Constant false (\perp) - **Negative constraints**: $\forall \mathbf{X} \Phi(\mathbf{X}) \rightarrow \perp$
- But query answering under Datalog[\exists] is **undecidable**
[see, e.g., Beeri & Vardi, **ICALP 81**]
- Datalog[$\exists, =, \perp$] is **syntactically restricted** \rightarrow **Datalog[±]**

The Chase Procedure

Input: Database D , set of TGDs Σ

Output: A model of $D \cup \Sigma$



Σ

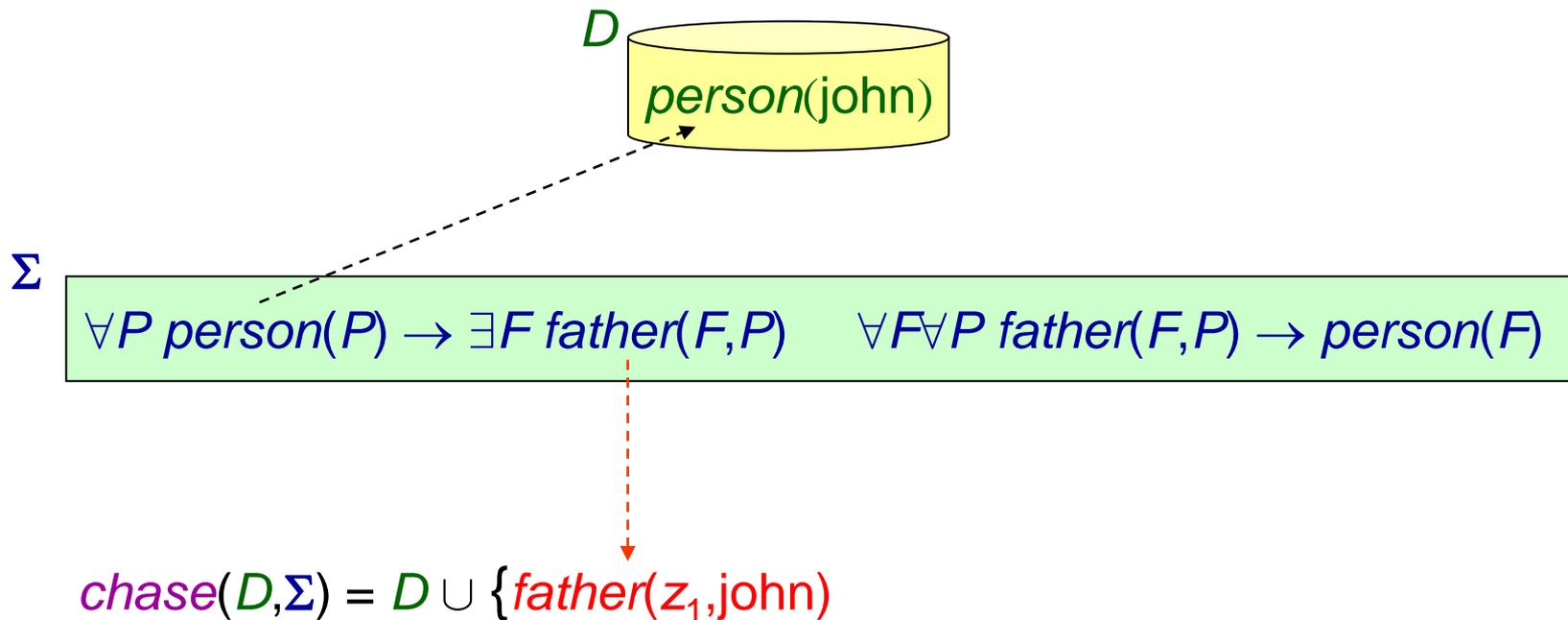
$\forall P \text{ person}(P) \rightarrow \exists F \text{ father}(F,P) \quad \forall F \forall P \text{ father}(F,P) \rightarrow \text{person}(F)$

$\text{chase}(D,\Sigma) = D \cup ?$

The Chase Procedure

Input: Database D , set of TGDs Σ

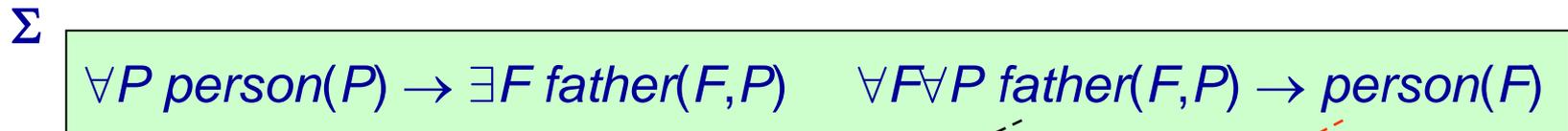
Output: A model of $D \cup \Sigma$



The Chase Procedure

Input: Database D , set of TGDs Σ

Output: A model of $D \cup \Sigma$

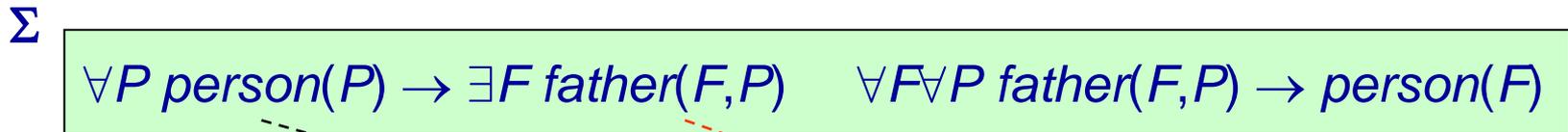


$chase(D, \Sigma) = D \cup \{father(z_1, john), person(z_1)\}$

The Chase Procedure

Input: Database D , set of TGDs Σ

Output: A model of $D \cup \Sigma$



$chase(D, \Sigma) = D \cup \{father(z_1, john), person(z_1), father(z_2, z_1)\}$

The Chase Procedure

Input: Database D , set of TGDs Σ

Output: A model of $D \cup \Sigma$



Σ

$\forall P \text{ person}(P) \rightarrow \exists F \text{ father}(F,P) \quad \forall F \forall P \text{ father}(F,P) \rightarrow \text{person}(F)$

$\text{chase}(D,\Sigma) = D \cup \{\text{father}(z_1,\text{john}), \text{person}(z_1), \text{father}(z_2,z_1), \dots\}$

The Chase Procedure

Input: Database D , set of TGDs Σ

Output: A model of $D \cup \Sigma$



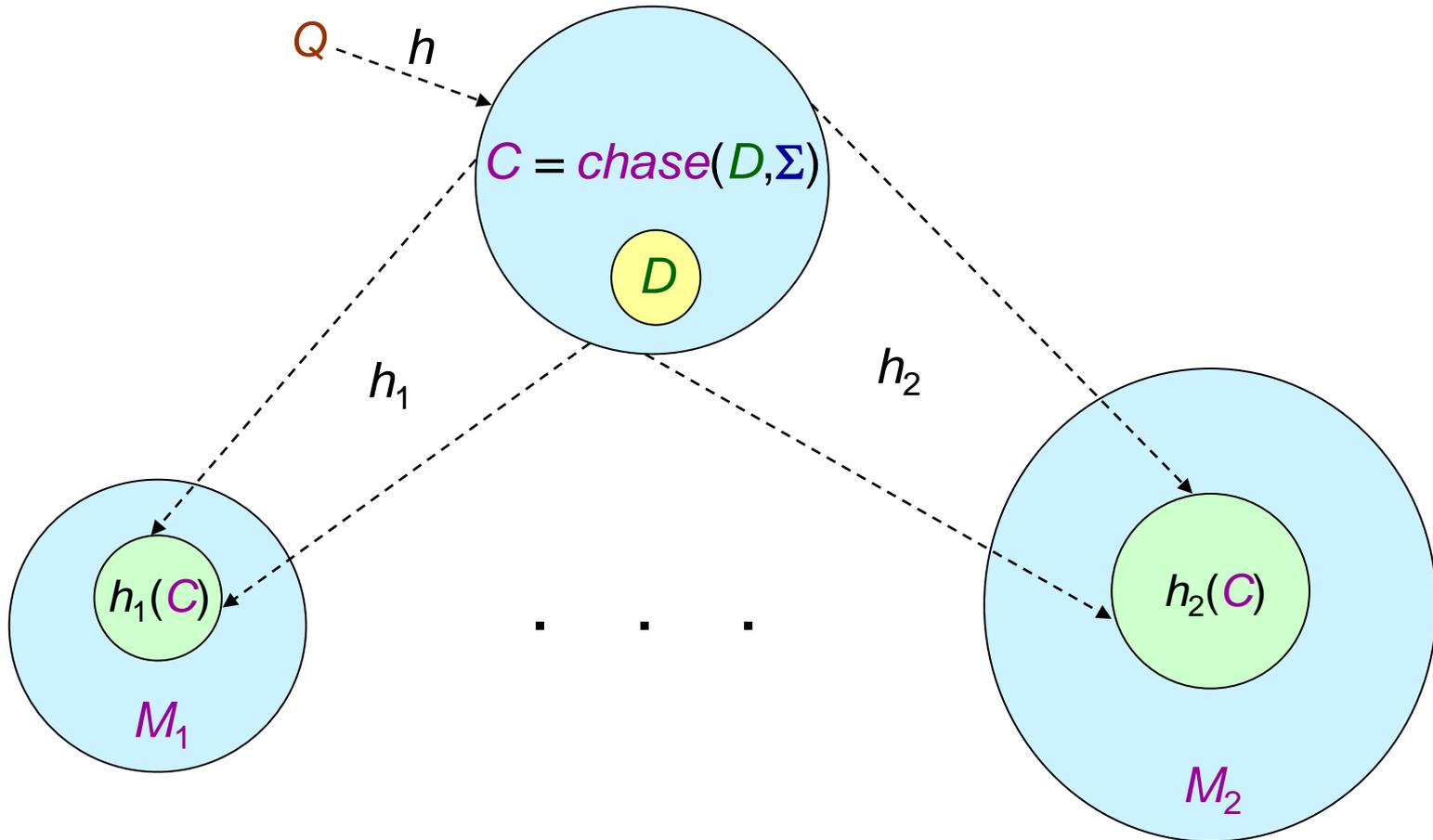
Σ

$\forall P \text{ person}(P) \rightarrow \exists F \text{ father}(F,P) \quad \forall F \forall P \text{ father}(F,P) \rightarrow \text{person}(F)$

$\text{chase}(D,\Sigma) = D \cup \{\text{father}(z_1,\text{john}), \text{person}(z_1), \text{father}(z_2,z_1), \dots\}$

infinite instance

Query Answering via Chase

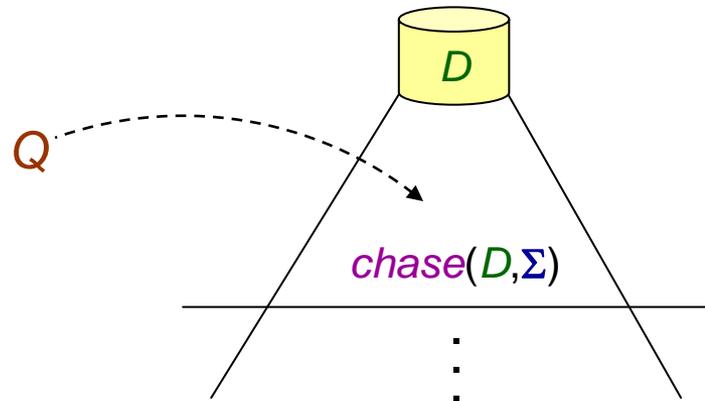


$$D \cup \Sigma \models Q \iff \text{chase}(D, \Sigma) \models Q$$

[see, e.g., Fagin, Kolaitis, Miller & Popa, TCS 05]

Early Positive Result

- Query answering under **IDs** is **decidable** $P(X,Y) \rightarrow \exists Z R(X,Y,Z)$
 - **PSPACE-complete** in combined complexity
 - **NP-complete** for bounded arity



Guardedness

- All \forall -variables occur in one body atom - **guard atom**

$$\forall X \forall Y \forall Z R(X, Y, Z), S(Y), P(X, Z) \rightarrow \exists W Q(X, W)$$

guard



- Chase has **finite treewidth** \Rightarrow decidability of query answering

[Calì, G. & Kifer, **KR 08**]

- Query answering is **P_{TIME}-complete** in data complexity

[Calì, G. & Lukasiewicz, **PODS 09**]

- Properly extends **ELH** (same data complexity)

Ontology Querying

ELH: Popular DL (for biological applications) with **P**TIME data complexity

[Baader, *IJCAI 03* and Rosati, *DL 07*]

ELH TBox	Datalog [±] Representation
$A \sqsubseteq B$	$\forall X A(X) \rightarrow B(X)$
$A \sqcap B \sqsubseteq C$	$\forall X A(X), B(X) \rightarrow C(X)$
$\exists R.A \sqsubseteq B$	$\forall X R(X, Y), A(Y) \rightarrow B(X)$
$A \sqsubseteq \exists R.B$	$\forall X A(X) \rightarrow \exists Y R(X, Y), B(Y)$
$R \sqsubseteq P$	$\forall X \forall Y R(X, Y) \rightarrow P(X, Y)$

Linearity

- Just **one atom** in the body

$$\forall \mathbf{X} \forall \mathbf{Y} R(\mathbf{X}, \mathbf{Y}) \rightarrow \exists \mathbf{Z} \Psi(\mathbf{X}, \mathbf{Z})$$

guard

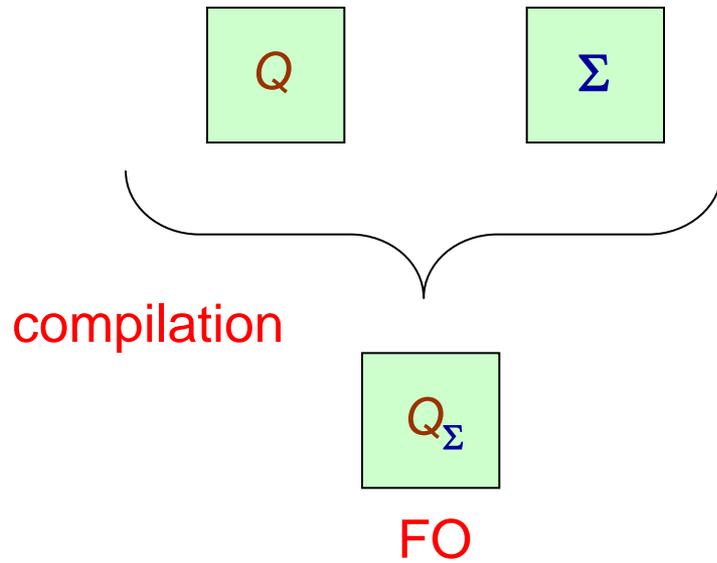


- Linear TGDs are **trivially guarded**

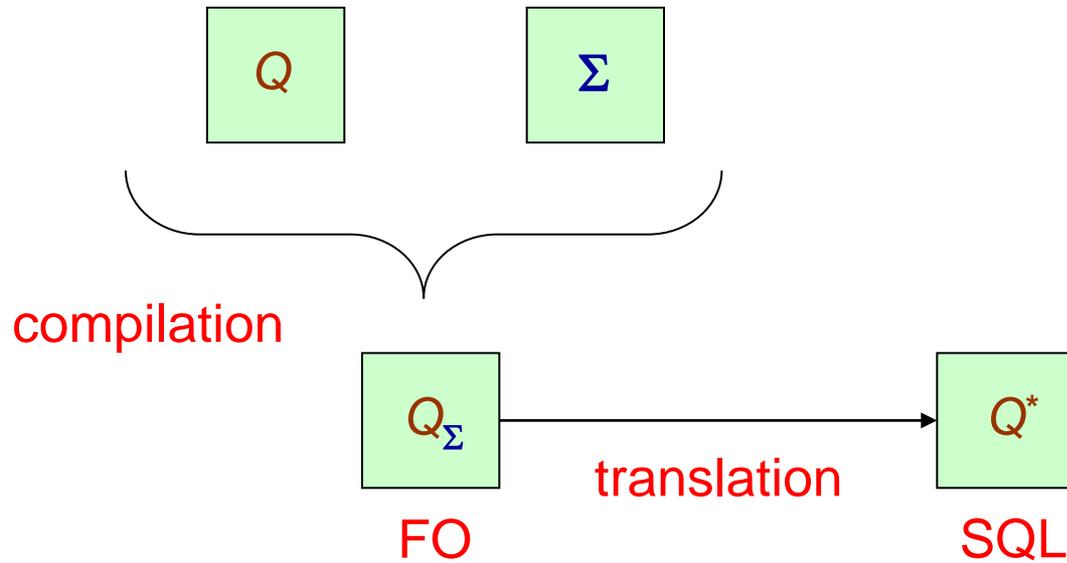
- Query answering is in AC_0 in data complexity (**first-order rewritability**)

[Calì, G. & Lukasiewicz, **PODS 09**]

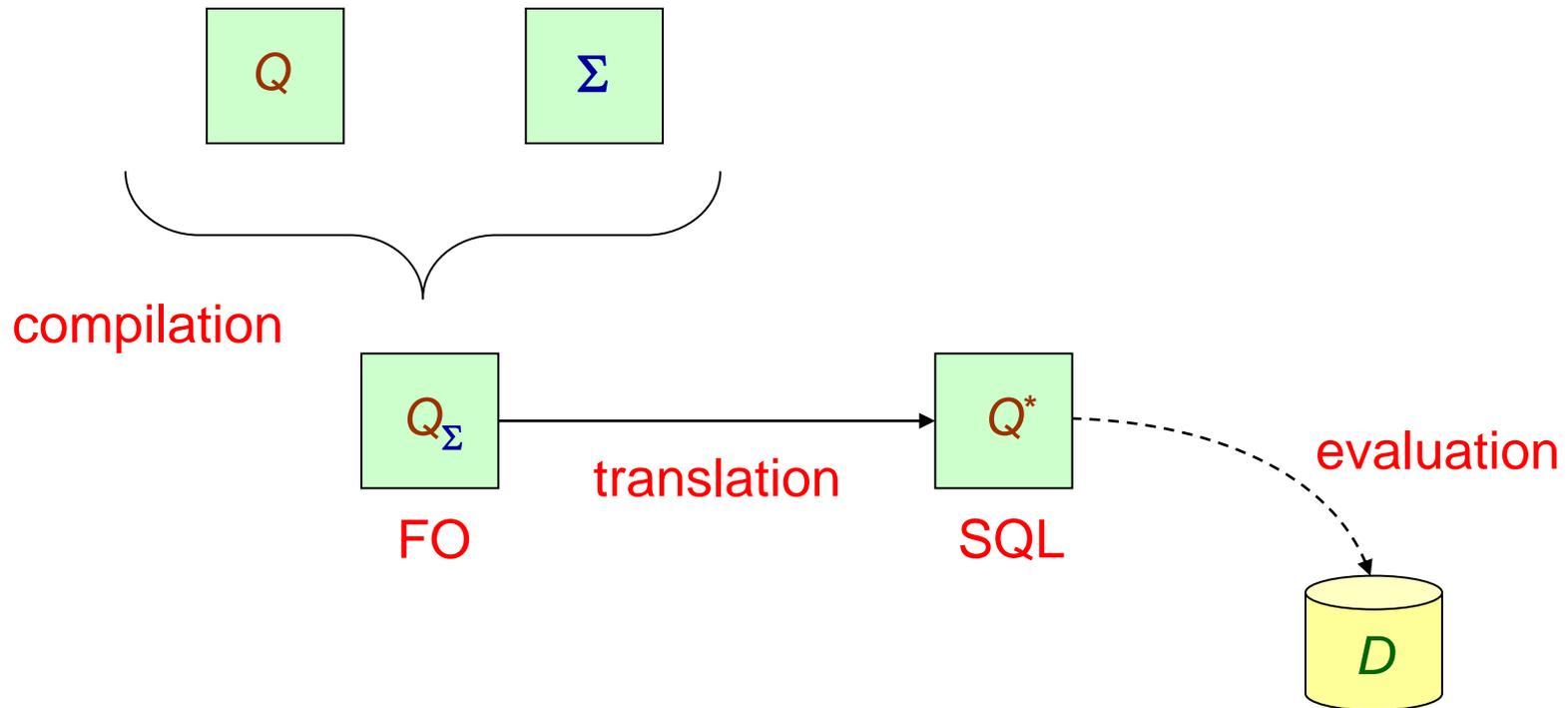
First-Order Rewritable TGDs



First-Order Rewritable TGDs



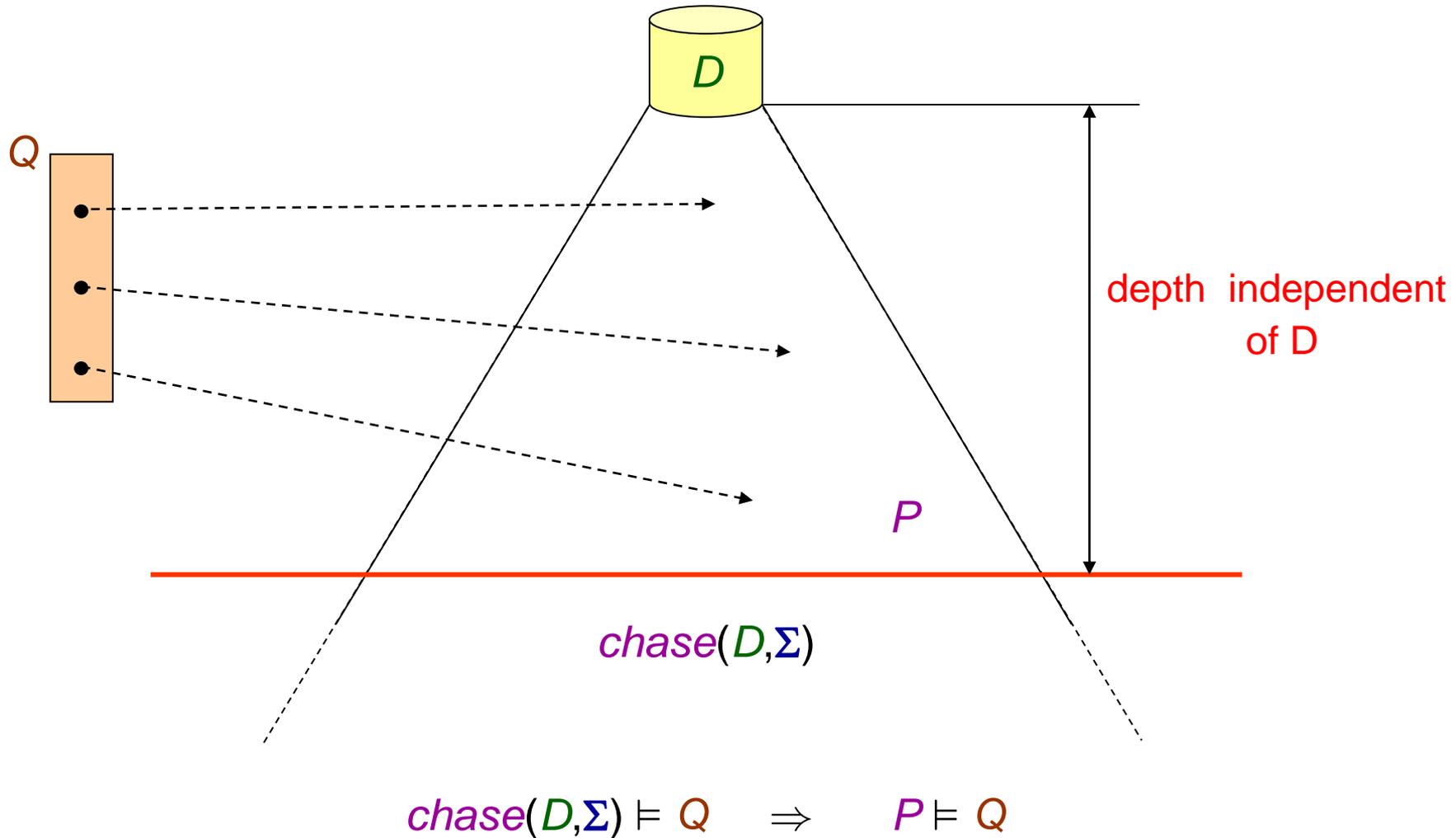
First-Order Rewritable TGDs



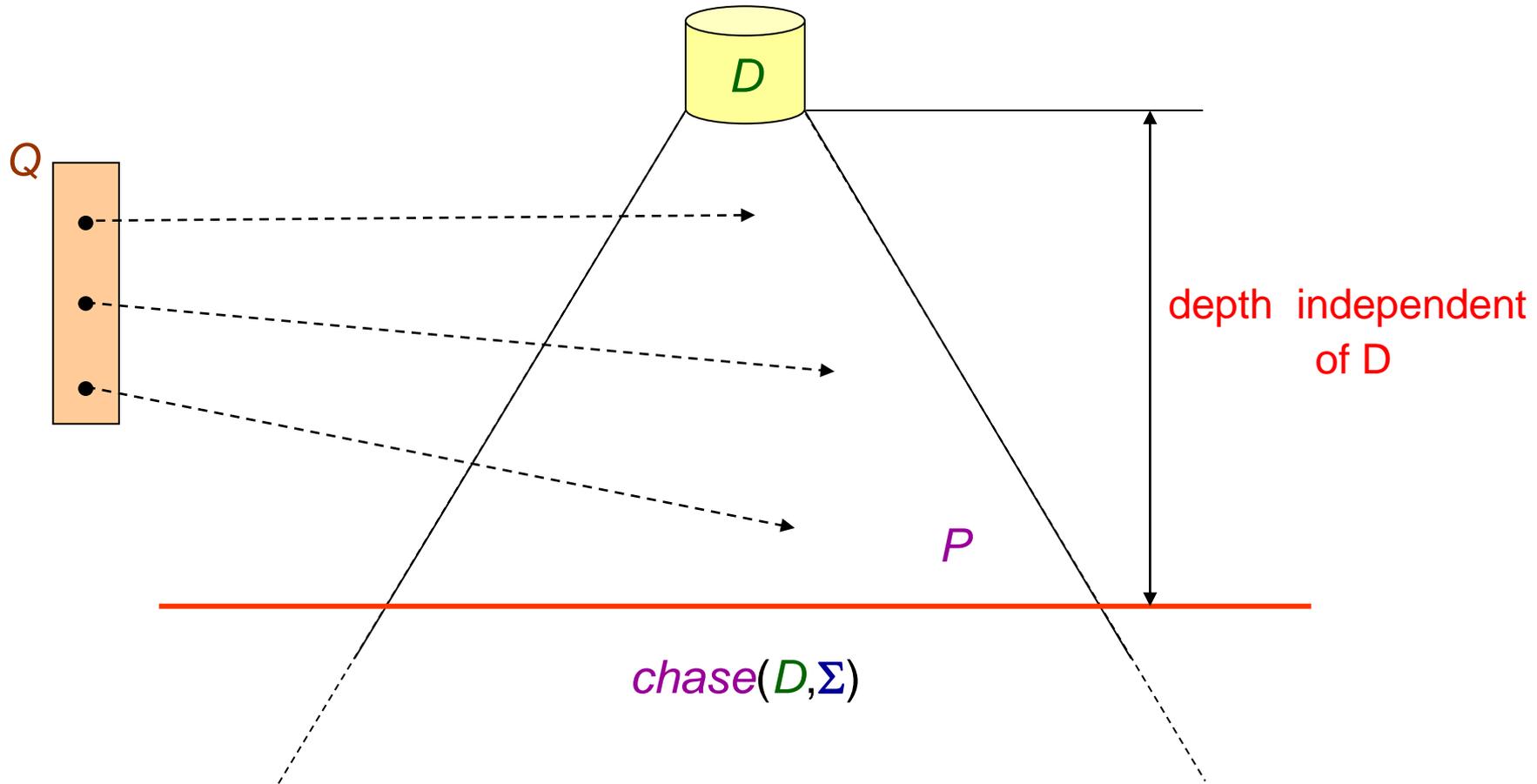
$$\forall D (D \cup \Sigma \models Q \Leftrightarrow D \models Q^*)$$

Query answering is in AC_0 in data complexity [Vardi, PODS 95]

Bounded Derivation-Depth Property (BDDP)



Bounded Derivation-Depth Property (BDDDP)



BDDDP \Rightarrow First-Order Rewritability

Linearity

- Just **one atom** in the body

$$\forall \mathbf{X} \forall \mathbf{Y} R(\mathbf{X}, \mathbf{Y}) \rightarrow \exists \mathbf{Z} \Psi(\mathbf{X}, \mathbf{Z})$$

trivially guard



- Linear TGDs **are** trivially guarded
- Query answering is in AC_0 in data complexity (**first-order rewritability**)
[Calì, Gottlob & Lukasiewicz, **PODS 09**]
- Properly extends **DL-Lite** (same data complexity)

Ontology Querying

DL-Lite: Popular family of DLs with AC_0 data complexity (OWL 2 QL)

[Calvanese, De Giacomo, Lembo, Lenzerini & Rosati, *JAR 07*]

DL-Lite TBox	Datalog [±] Representation
$A \sqsubseteq B$	$\forall X A(X) \rightarrow B(X)$
$A \sqsubseteq \exists R$	$\forall X A(X) \rightarrow \exists Y R(X, Y)$
$\exists R \sqsubseteq A$	$\forall X \forall Y R(X, Y) \rightarrow A(X)$
$R \sqsubseteq P$	$\forall X \forall Y R(X, Y) \rightarrow P(X, Y)$

But...

- What about **joins** in rule bodies?

$\forall A \forall D \forall P \text{ runs}(D,P), \text{area}(P,A) \rightarrow \exists E \text{ employee}(E,D,P,A)$



- What about the DL assertion **concept product**?

$\forall E \forall M \text{ elephant}(E), \text{mouse}(M) \rightarrow \text{biggerThan}(E,M)$

But...

- What about **joins** in rule bodies?

$$\forall A \forall D \forall P \text{ runs}(D, P), \text{ area}(P, A) \rightarrow \exists E \text{ employee}(E, D, P, A)$$



- What about the DL assertion **concept product**?

$$\forall E \forall M \text{ elephant}(E), \text{ mouse}(M) \rightarrow \text{ biggerThan}(E, M)$$

No tree-like models guaranteed

$$\forall X \forall Y R(X, Y) \rightarrow \exists Z R(Y, Z)$$

$$\forall X \forall Y R(X, Y) \rightarrow S(X)$$

$$\forall X \forall Y S(X), S(Y) \rightarrow P(X, Y)$$

Infinitely many symbols in S

P forms an **infinite clique**

Stickiness

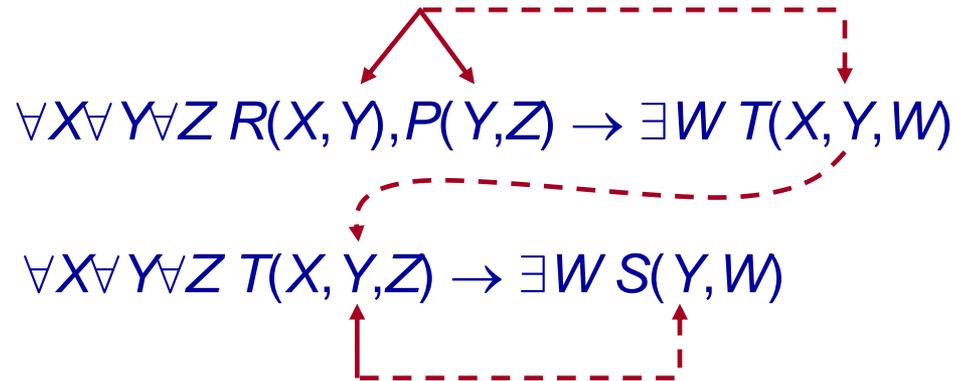
$\forall A \forall D \forall P \text{ runs}(D, P), \text{area}(P, A) \rightarrow \exists E \text{ employee}(E, D, P, A)$



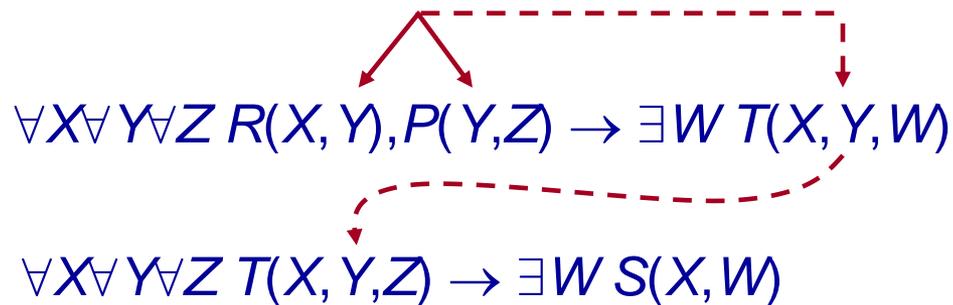
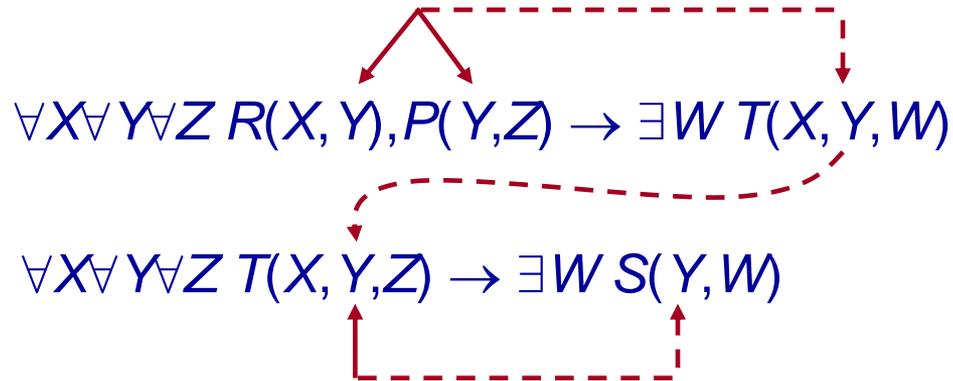
$\forall E \forall M \text{ elephant}(E), \text{mouse}(M) \rightarrow \text{biggerThan}(E, M)$



Stickiness



Stickiness



Stickiness: Marking Procedure

Initial Marking: mark all occurrences of body-variables that do not appear in **all** head-atoms

$$\forall V \forall W R_1(V, W) \rightarrow \exists X \exists Y \exists Z R_2(W, X, Y, Z)$$

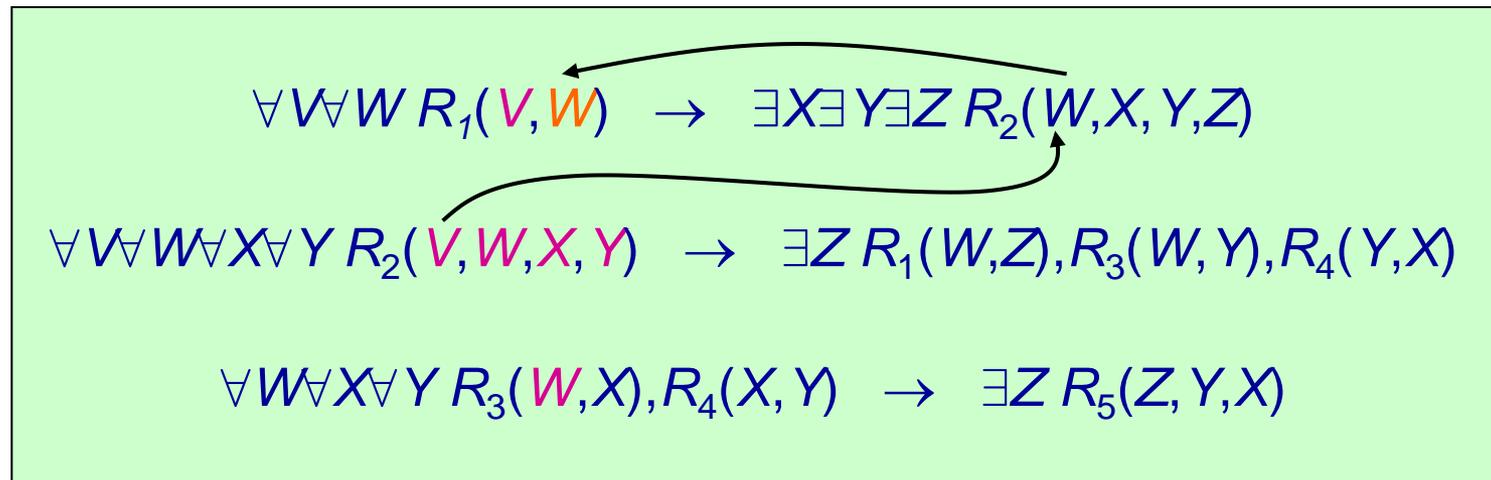
$$\forall V \forall W \forall X \forall Y R_2(V, W, X, Y) \rightarrow \exists Z R_1(W, Z), R_3(W, Y), R_4(Y, X)$$

$$\forall W \forall X \forall Y R_3(W, X), R_4(X, Y) \rightarrow \exists Z R_5(Z, Y, X)$$

Stickiness: Marking Procedure

Initial Marking: mark all occurrences of body-variables that do not appear in **all** head-atoms

Propagation Step: propagate the marking to body-variables via the head-atoms



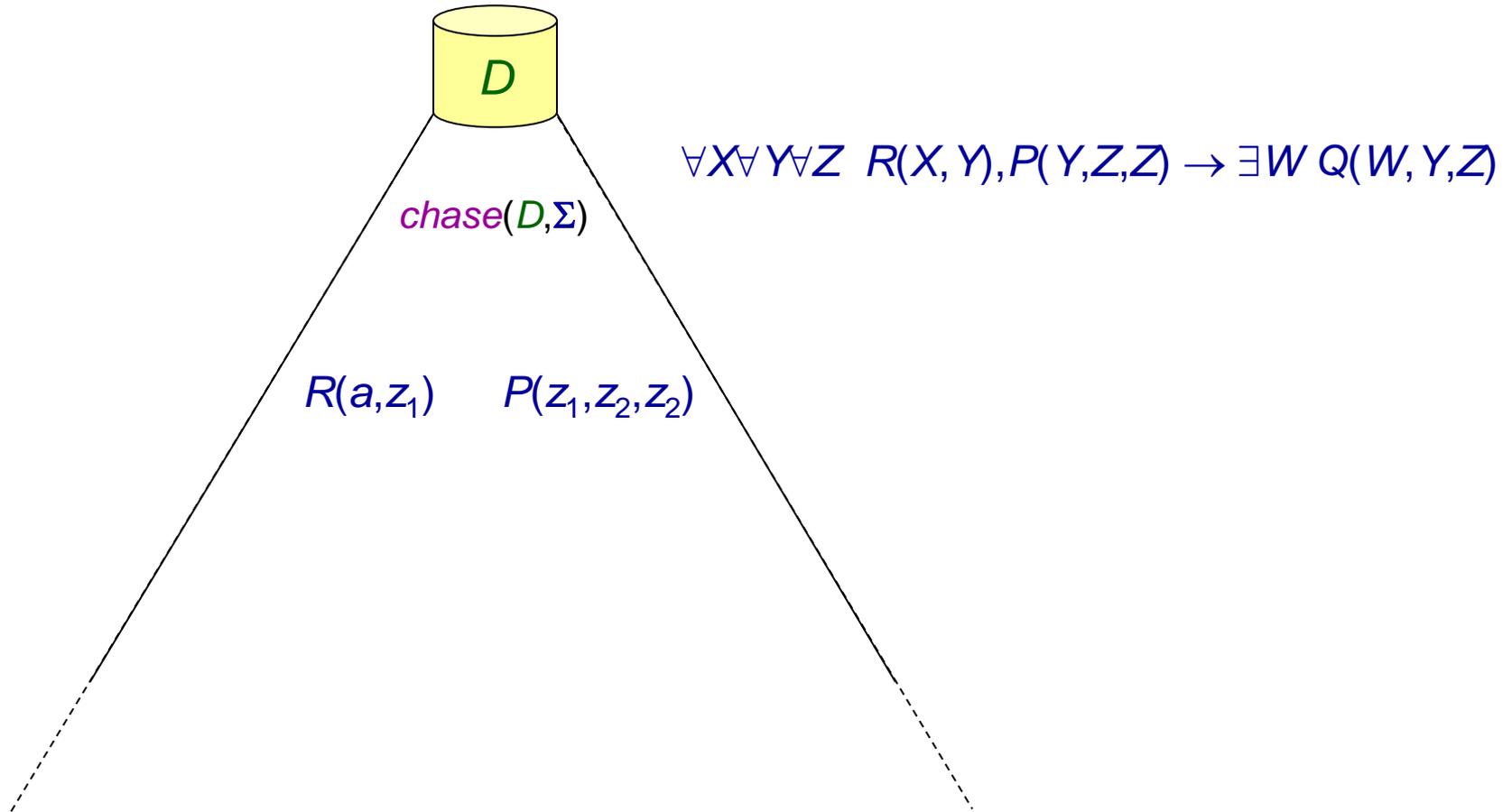
Stickiness

- Marked variables occur **just once** in the body of each TGD

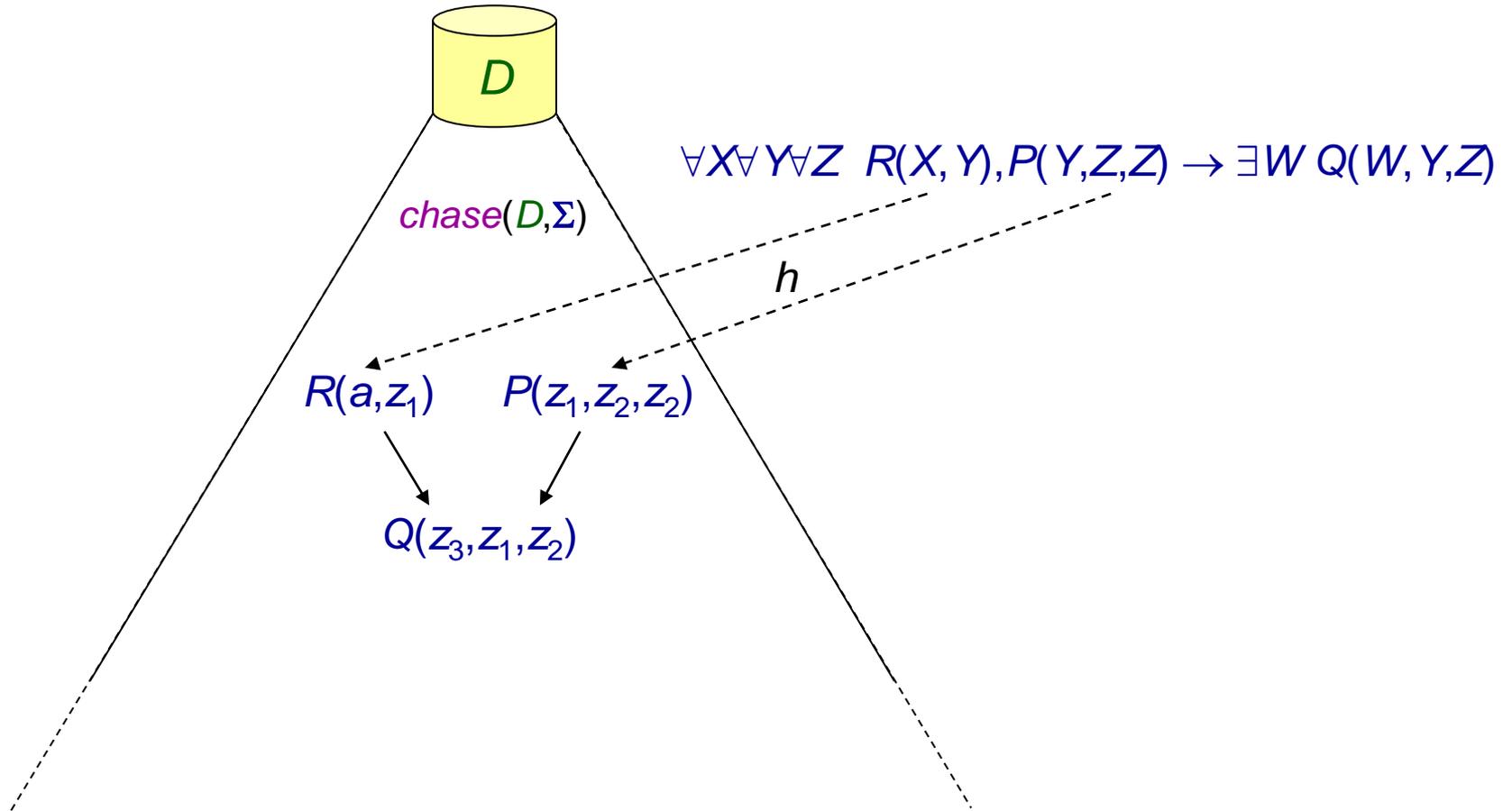
$$\begin{aligned} \forall V \forall W R_1(V, W) &\rightarrow \exists X \exists Y \exists Z R_2(W, X, Y, Z) \\ \forall V \forall W \forall X \forall Y R_2(V, W, X, Y) &\rightarrow \exists Z R_1(W, Z), R_3(W, Y), R_4(Y, X) \\ \forall W \forall X \forall Y R_3(W, X), R_4(X, Y) &\rightarrow \exists Z R_5(Z, Y, X) \end{aligned}$$

- The chase has the **sticky property** (backward-resolution **terminates**)

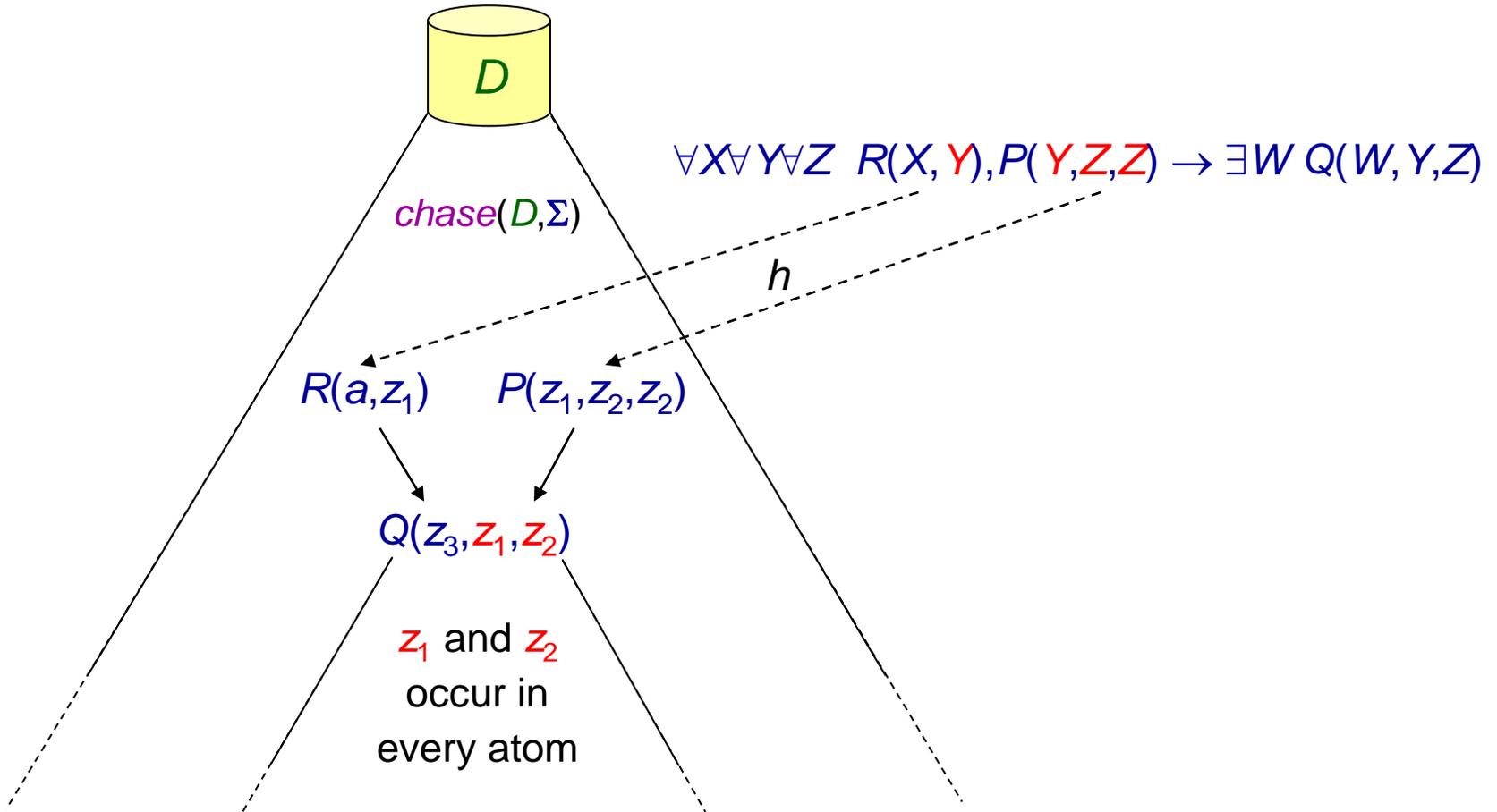
Sticky Property



Sticky Property



Sticky Property



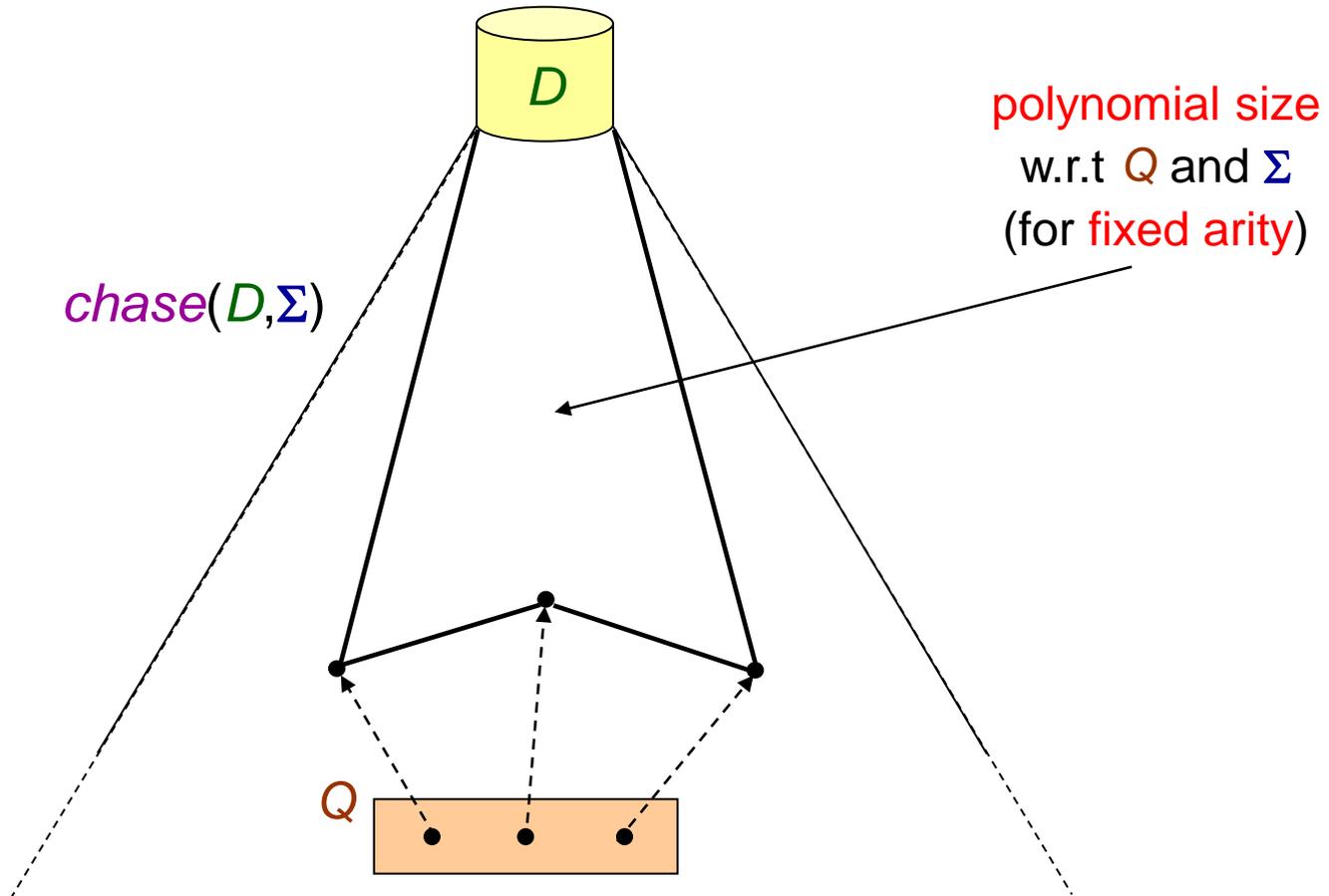
Stickiness

- Marked variables occur **just once** in the body of each TGD

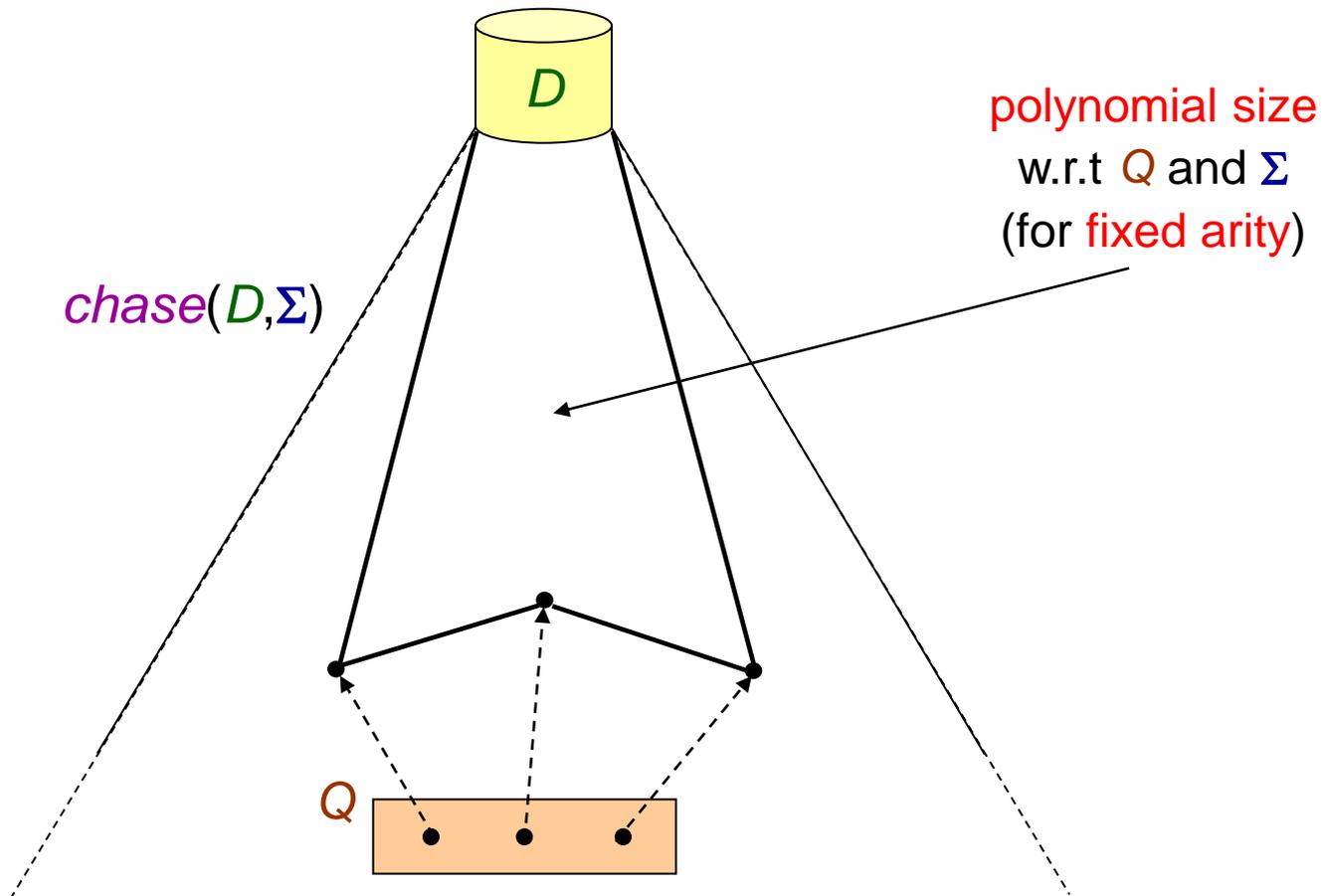
$$\begin{aligned} \forall V \forall W R_1(V, W) &\rightarrow \exists X \exists Y \exists Z R_2(W, X, Y, Z) \\ \forall V \forall W \forall X \forall Y R_2(V, W, X, Y) &\rightarrow \exists Z R_1(W, Z), R_3(W, Y), R_4(Y, X) \\ \forall W \forall X \forall Y R_3(W, X), R_4(X, Y) &\rightarrow \exists Z R_5(Z, Y, X) \end{aligned}$$

- The chase has the **sticky property** (backward-resolution **terminates**)
- Query answering is in AC_0 in data complexity (**first-order rewritability**)
[Cali, G. & Pieris, VLDB 10]
- Properly extends **DL-Lite** (same data complexity)

Polynomial Witness Property (PWP)

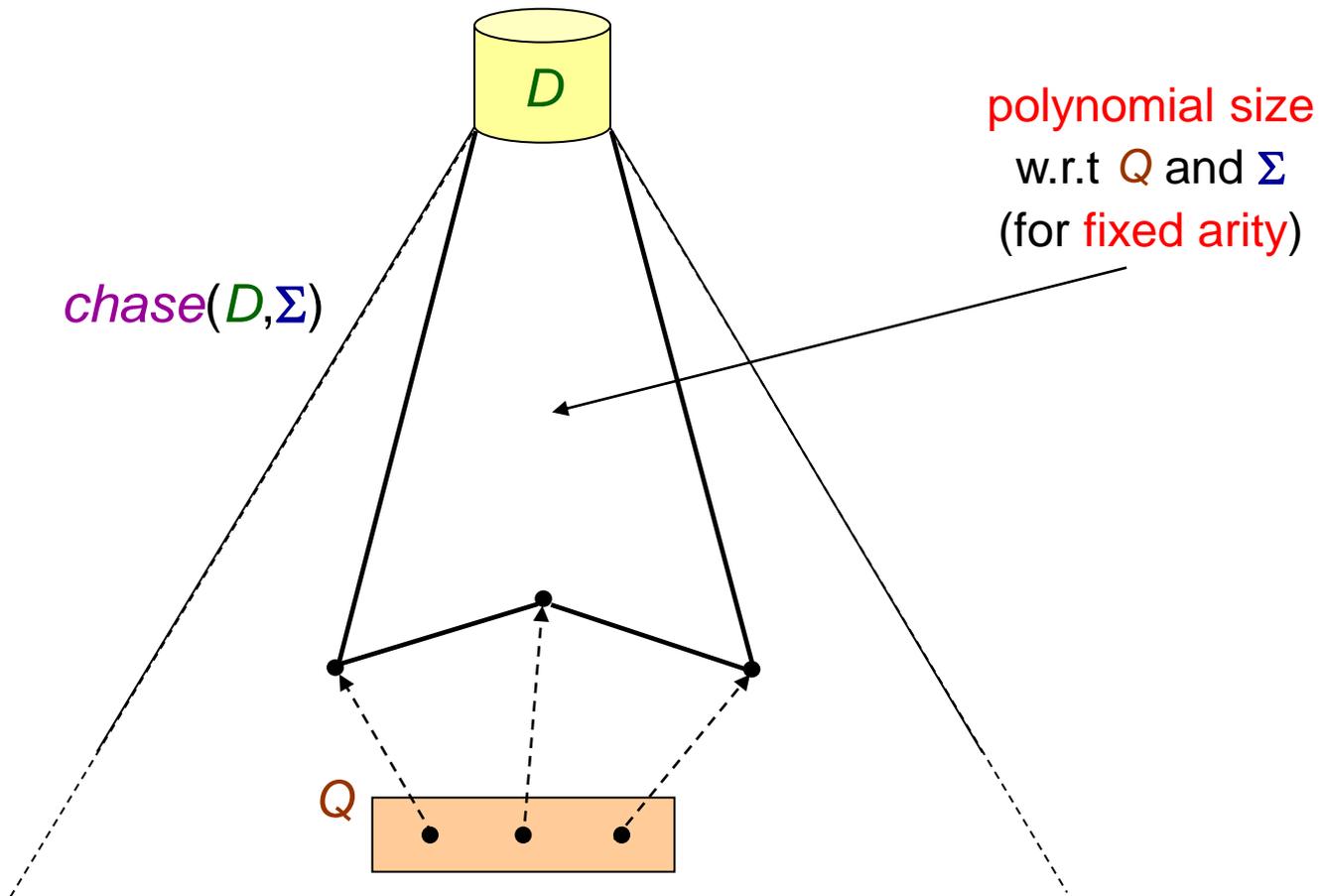


Polynomial Witness Property (PWP)



PWP \Rightarrow non-recursive Datalog rewriting of polynomial size

Polynomial Witness Property (PWP)



Holds for the first-order rewritable Datalog[±] fragments (**linear** and **sticky**)

Finite Controllability

$$D \cup \Sigma \models Q \stackrel{?}{\Leftrightarrow} D \cup \Sigma \models_{\text{fin}} Q$$

- Holds for **inclusion dependencies**

[Rosati, **PODS 06**]

- Holds for **guarded TGDs** (in fact, for the **guarded fragment**)

[Bárány, Gottlob & Otto, **LICS 10**]

- Holds for **sticky TGDs**

[Gogacz & Marcinkowski, \rightarrow **next talk**]

Additional Features

- **EGDs**, e.g., $\forall X \forall Y \forall Z \text{ reports}(X, Y), \text{ reports}(X, Z) \rightarrow Y = Z$
Non-Conflicting EGDs: do not interact with TGDs
Preliminary check **without** adding complexity
- **Negative constraints**, e.g., $\forall X \text{ emp}(X), \text{ customer}(X) \rightarrow \perp$
Check **without** adding complexity

Additional Features

- **EGDs**, e.g., $\forall X \forall Y \forall Z \text{ reports}(X, Y), \text{ reports}(X, Z) \rightarrow Y = Z$
Non-Conflicting EGDs: do not interact with TGDs
Preliminary check **without** adding complexity
- **Negative constraints**, e.g., $\forall X \text{ emp}(X), \text{ customer}(X) \rightarrow \perp$
Check **without** adding complexity

Finite controllability **does not hold**

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

$$\Sigma = \left\{ \begin{array}{l} \forall X \forall Y R(X, Y) \rightarrow \exists Z R(Y, Z) \\ \forall X \forall Y \forall Z R(Y, X), R(Z, X) \rightarrow Y = Z \end{array} \right\}$$

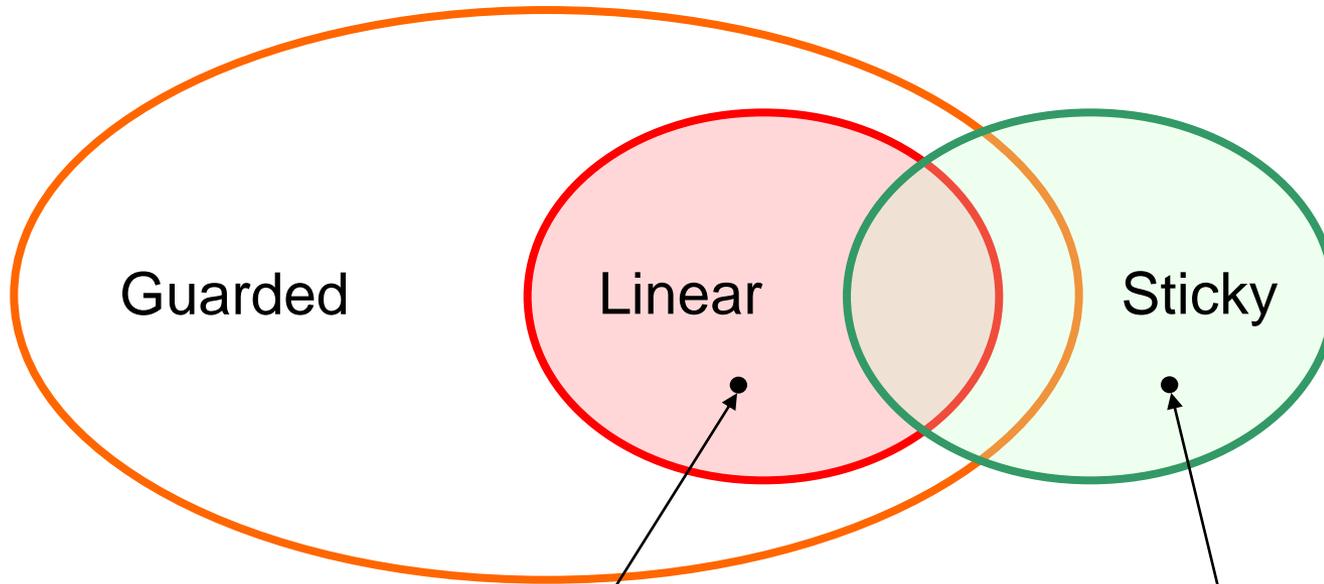
$$Q \leftarrow R(A, a)$$

$$D \cup \Sigma \not\models Q$$

but

$$D \cup \Sigma \models_{\text{fin}} Q$$

Datalog[±]: Overview



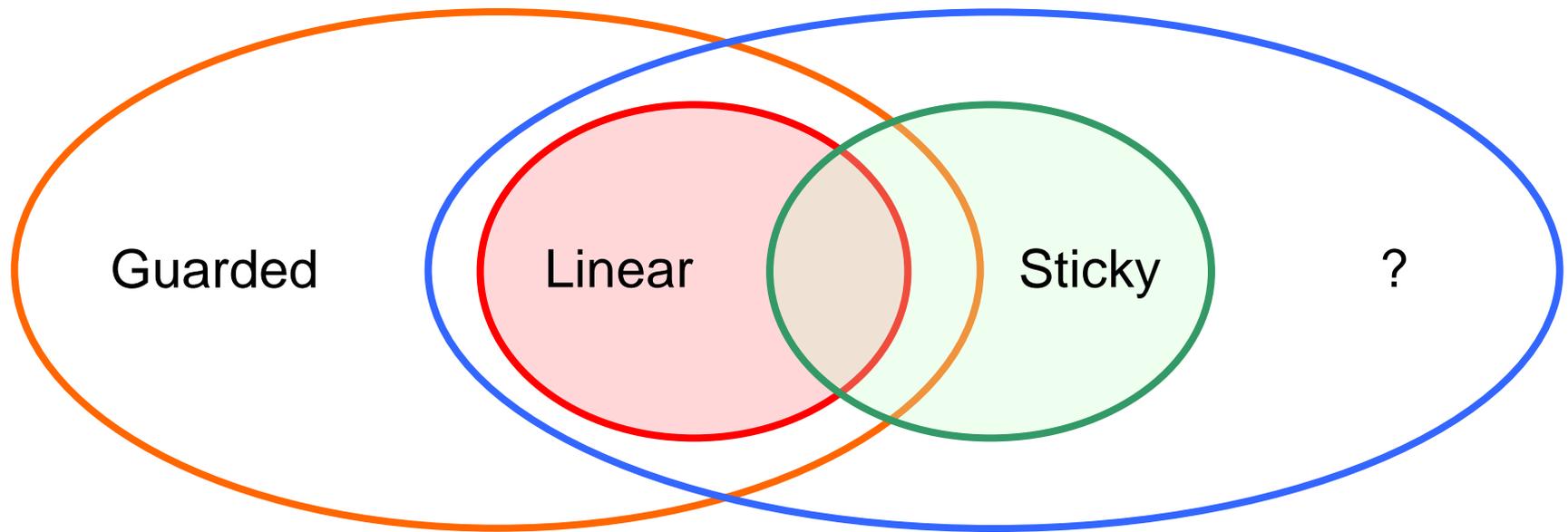
$$\forall X \forall Y R(X, X, Y) \rightarrow \exists Z R(Y, Y, Z)$$

An arrow points from this formula to the "Linear" region of the Venn diagram.

$$\forall X \forall Y \forall Z \forall W S(X, Y), R(Z, W) \rightarrow \exists V S(Y, V), R(W, V)$$

An arrow points from this formula to the "Sticky" region of the Venn diagram.

Datalog[±]: Overview



without losing **first-order rewritability** and **PWP**

Sticky-Join TGDs

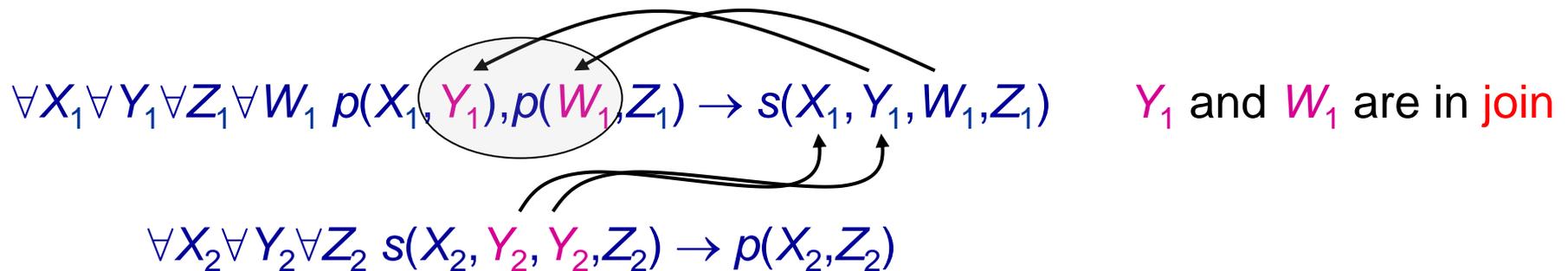
- Marked variables can appear more than once in a **single atom**
- Providing that the following **does not happen**:

$$\forall X_1 \forall Y_1 \forall Z_1 \forall W_1 p(X_1, Y_1), p(W_1, Z_1) \rightarrow s(X_1, Y_1, W_1, Z_1)$$
$$\forall X_2 \forall Y_2 \forall Z_2 s(X_2, Y_2, Y_2, Z_2) \rightarrow p(X_2, Z_2)$$

Y_1 and W_1 are in **join**

Sticky-Join TGDs

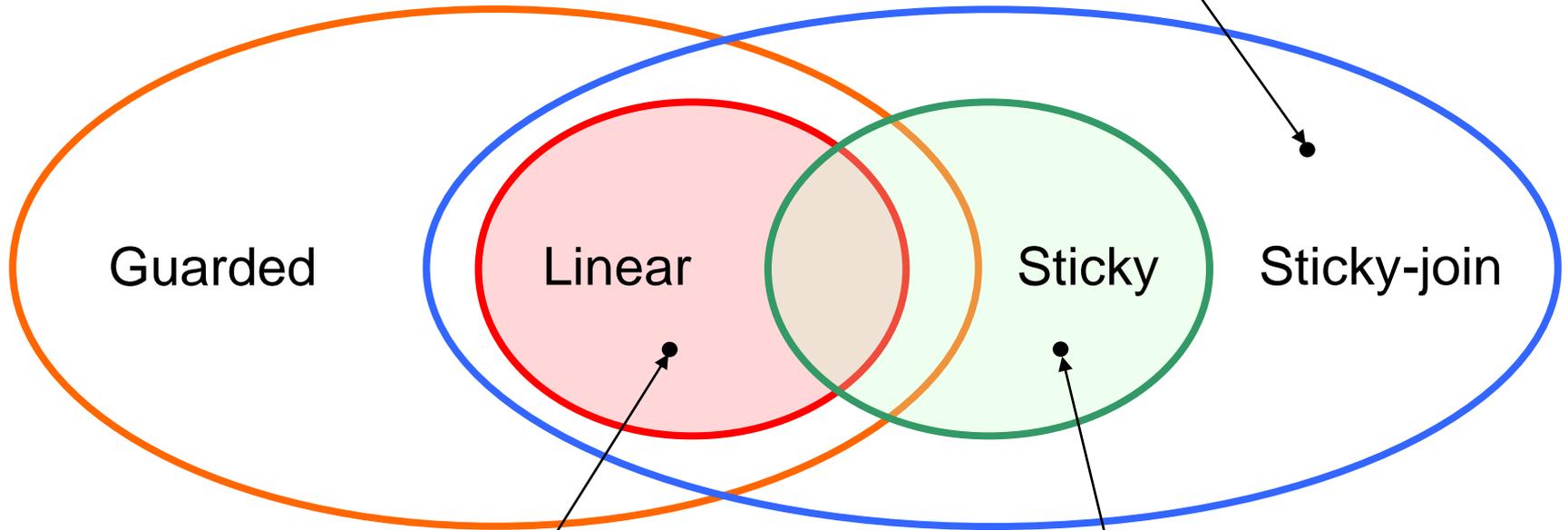
- Marked variables can appear more than once in a **single atom**
- Providing that the following **does not happen**:



- **Same complexity** as sticky TGDs and **enjoy PWP**
- But harder to identify them - **PSPACE-complete**

Datalog[±]: Overview

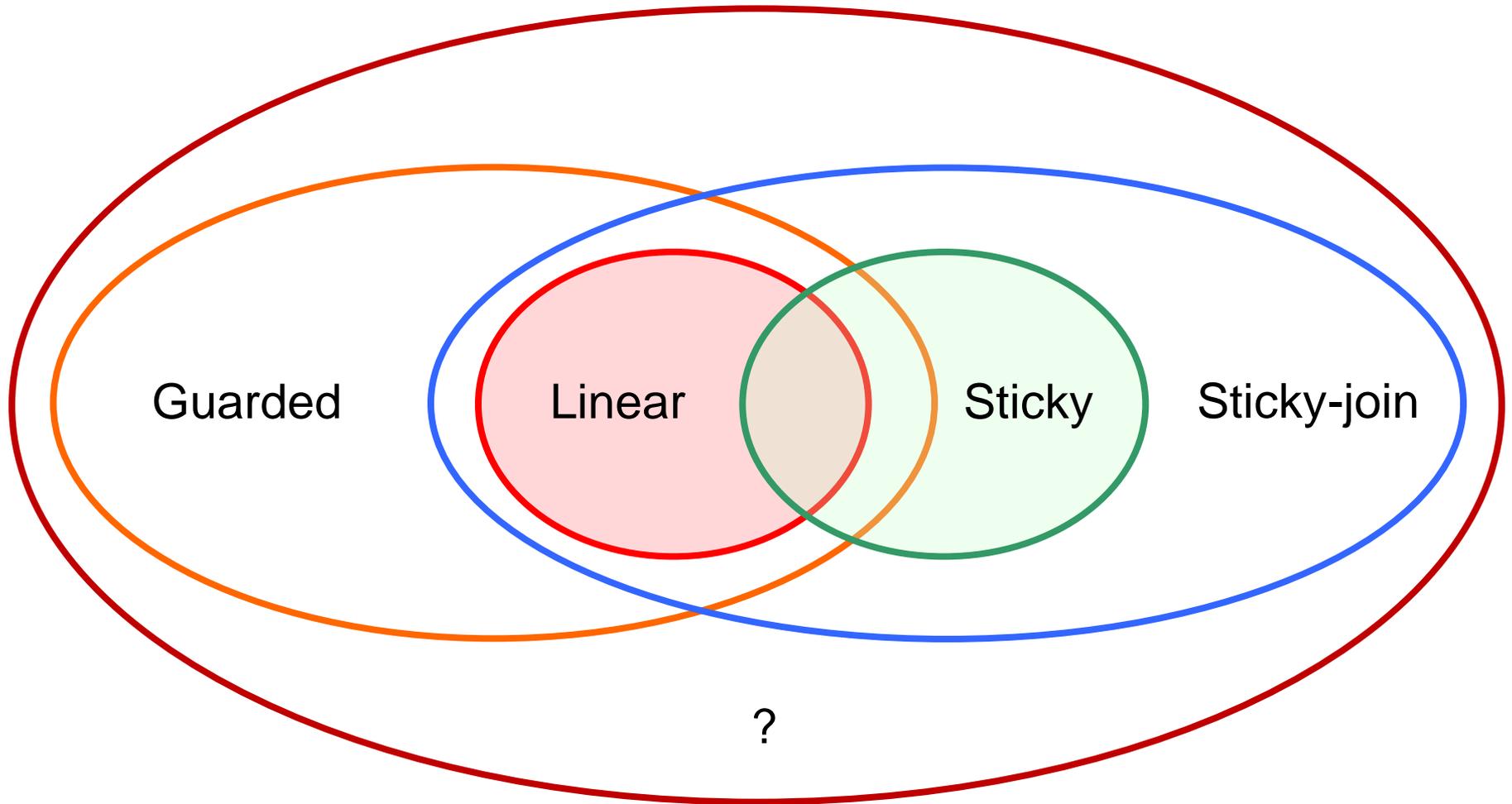
$$\forall X \forall Y \forall Z R(X, Y), S(Y, Z, Z) \rightarrow \exists W P(Y, W)$$



$$\forall X \forall Y R(X, X, Y) \rightarrow \exists Z R(Y, Y, Z)$$

$$\forall X \forall Y \forall Z S(X, Y), R(Y, Z) \rightarrow \exists W P(Y, W)$$

Datalog[±]: Overview



without losing **PTIME** data complexity

Tame TGDs (TTGDs)

- Check that each rule is **guarded*** or **sticky-join** (and thus sticky)

$$\forall X \forall Y \forall Z R(X, Y) \rightarrow \exists Z R(Y, Z)$$

sticky-join and guarded*

$$\forall X \forall Y T(X), R(X, Y) \rightarrow T(Y)$$

Guarded* not sticky-join

$$\forall X \forall Y T(X), T(Y) \rightarrow \exists Z P(X, Y, Z)$$

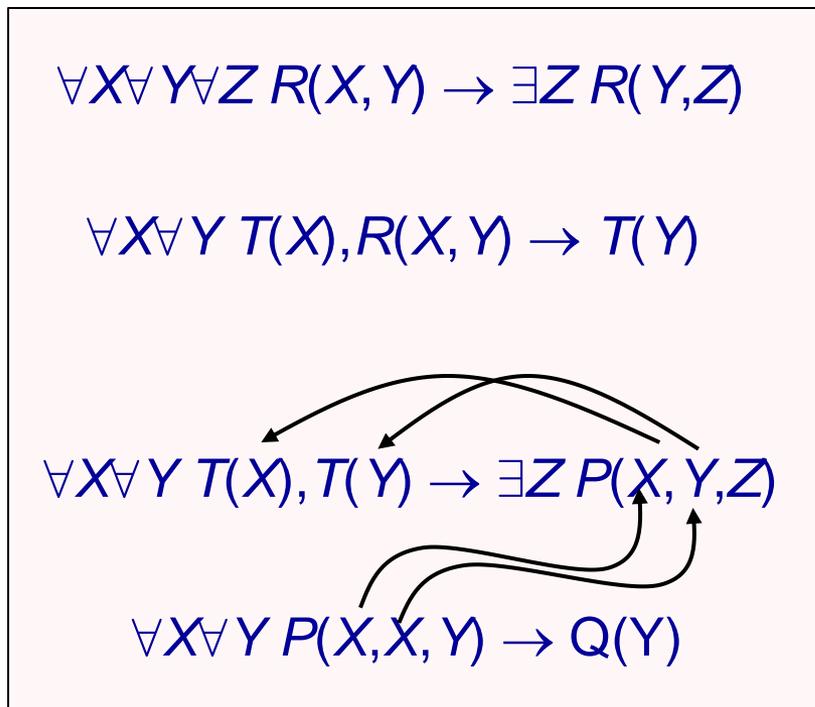
sticky-join not guarded

- **Not first-order rewritable** due to the second rule
- Chase has **infinite treewidth** since P forms an infinite clique

**) Predicates in heads of non-guarded rules are unguarded. Atoms whose predicate is unguarded are not allowed to serve as guards!*

Tame TGDs

- The following set of TGDs is **not tame**



sticky-join and guarded*

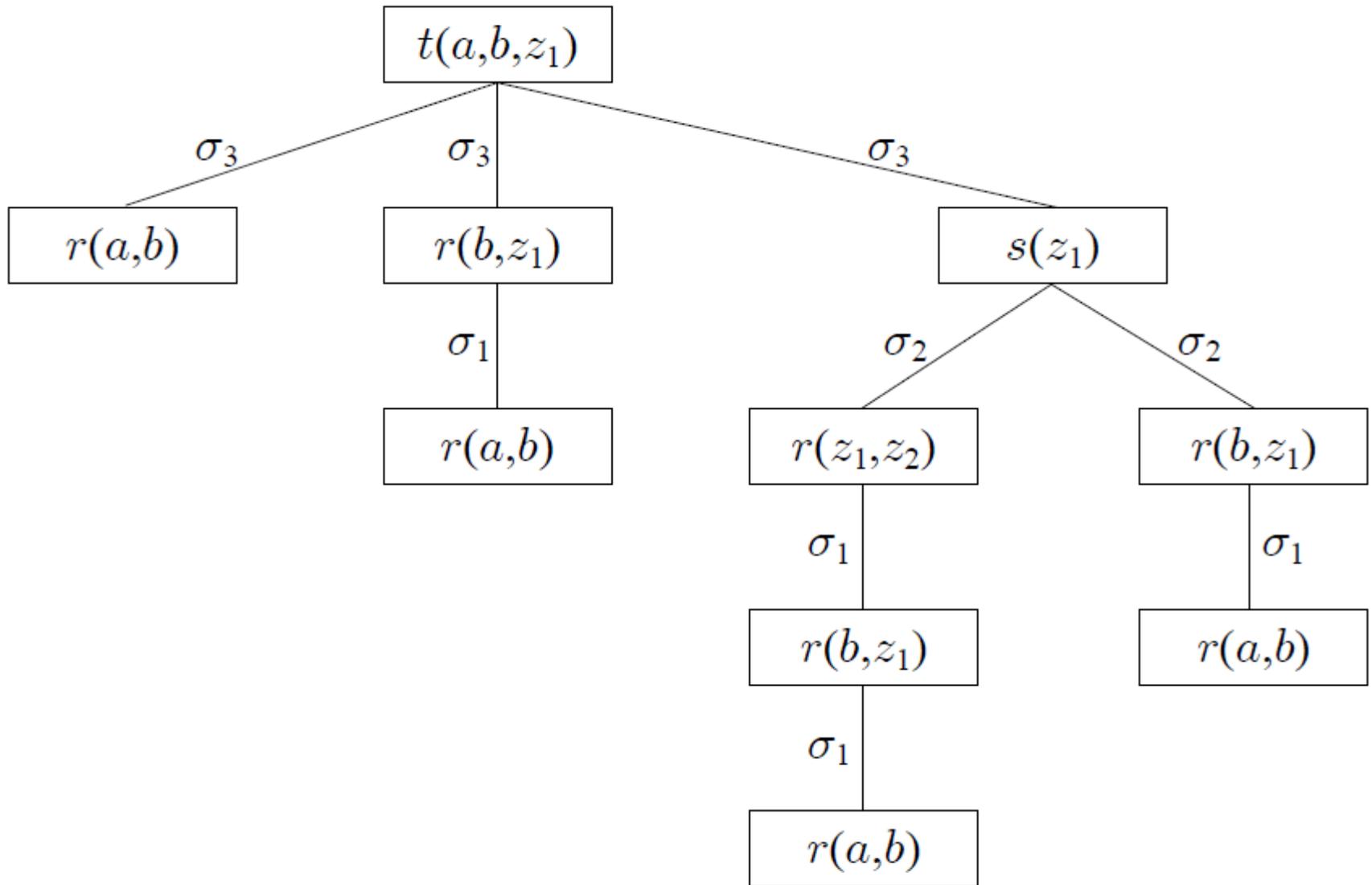
Guarded* not sticky-join

neither sticky-join nor guarded

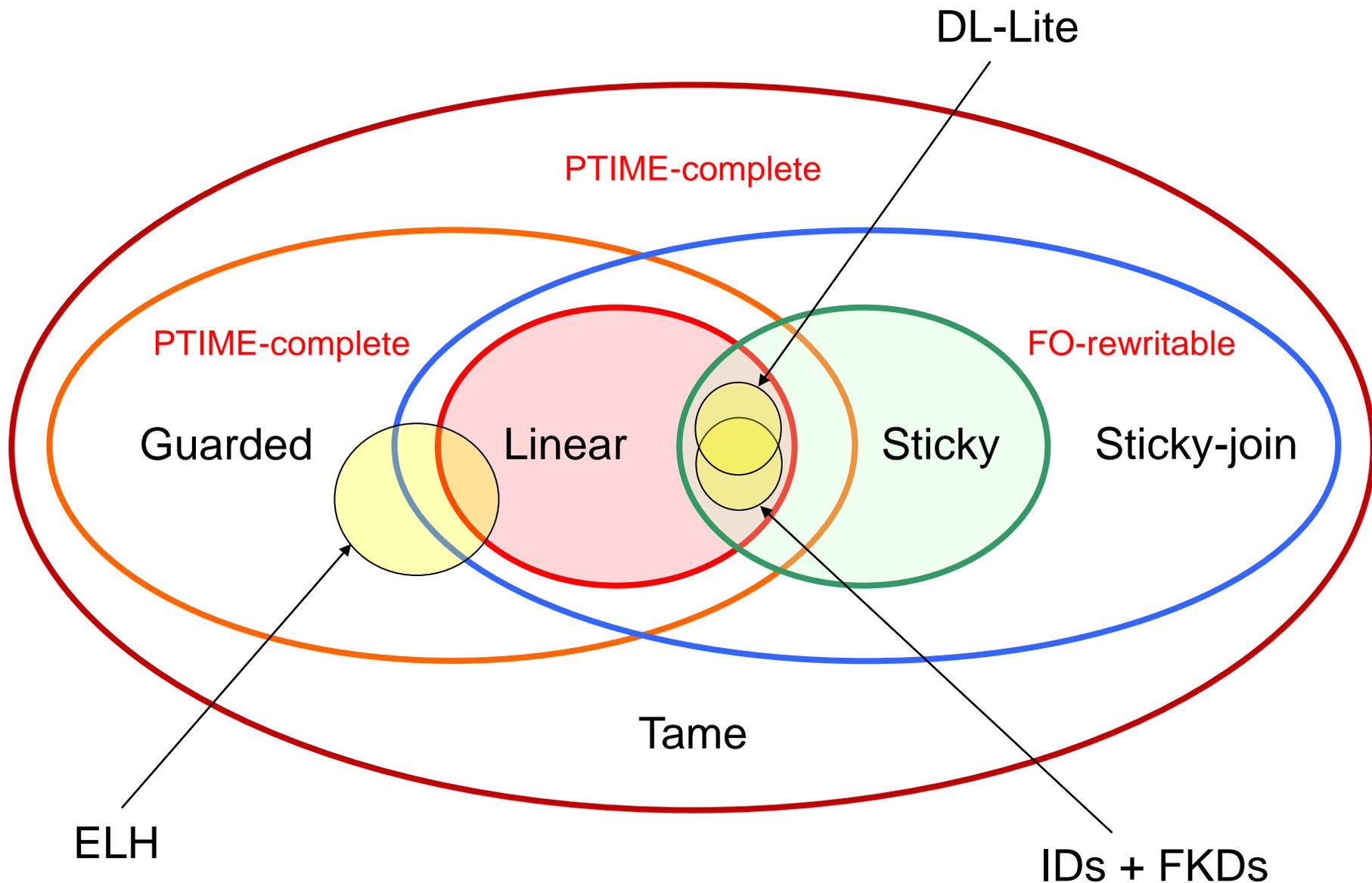
sticky-join and guarded

*) *Predicates in heads of non-guarded rules are unguarded. Atoms whose predicate is unguarded are not allowed to serve as guards!*

No bounded tw model, but bounded tw resolution proof scheme!



Datalog[±]: Overview



Datalog[±]: Summary of Complexity Results

	Data	Fixed Σ	Combined
Guarded	PTIME	NP	2EXPTIME
Linear	in AC_0	NP	PSPACE
Sticky	in AC_0	NP	EXPTIME
Sticky-Join	in AC_0	NP	EXPTIME
Tame	PTIME	NP	2EXPTIME

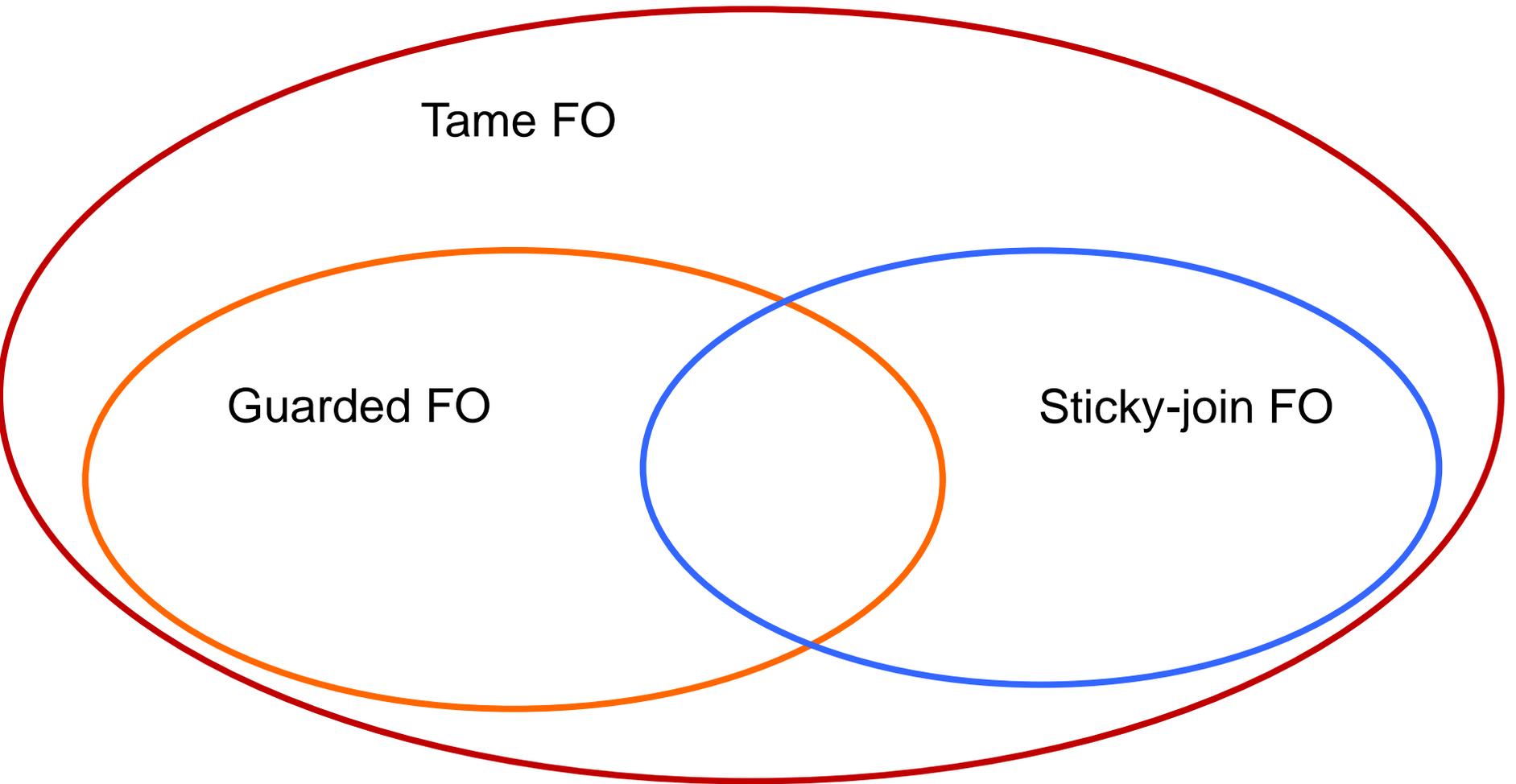
Same complexity with negative constraints and non-conflicting EGDs

current work

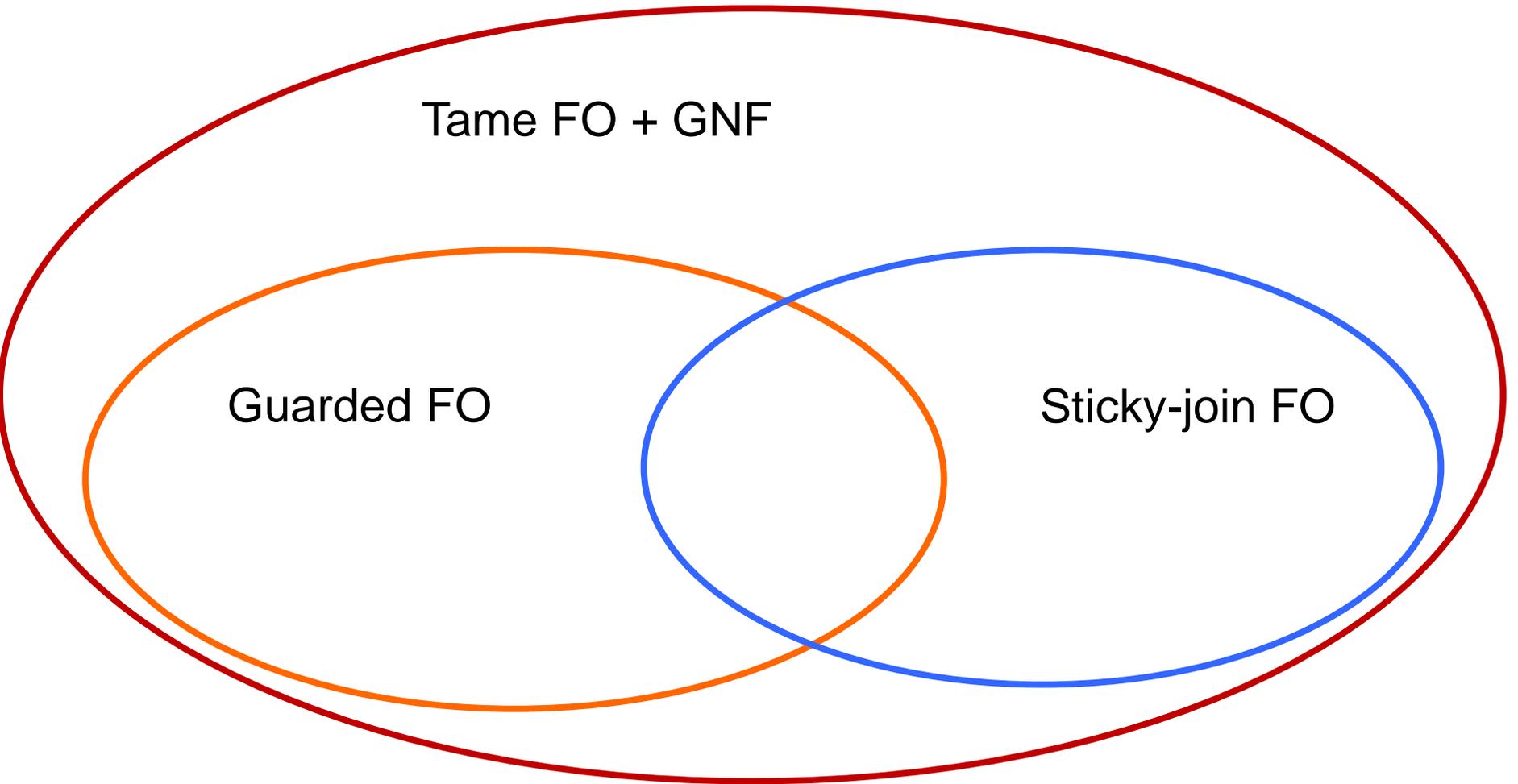
Finite controllability of Tamed TGDs:

Reduction to Sticky TGDs

current work



current work



Tame FO + GNF

Guarded FO

Sticky-join FO

Thank you!