Incomplete information, monotonicity and homomorphism preservation

Amélie Gheerbrant
(Univ. Paris Diderot)

Leonid Libkin
(Univ. of Edinburgh)

Cristina Sirangelo
(LSV, ENS Cachan)

Incomplete information and query answering

- Incomplete information in data: missing / unknown / partially specified data
  - several possible "completions"
- Still one of the most poorly understood aspects of data management
- Query answering
  - over usual databases : model checking $D \vDash Q$
  - over incomplete databases: entailment $R \vDash Q$ for all completions R of D

When can entailment be solved by (straightforward) model checking ?

In a database perspective:

When can we answer queries correctly on incomplete databases by using classical query evaluation engines ?

Model of incompleteness

Employee

Smith
x_1
Brown

Manager

Smith	x_1
x_1	Brown
Brown	x_2

D

- *Const* : a countably infinite set of constants
- *Nulls* : a countably infinite set of variables ranging over *Const* (marked nulls)
- σ : a finite relational signature

Incomplete database over σ (naïve table) [Imielinski, Lipski 84] :

a finite structure of signature σ with domain $\subset$ *Const* $\cup$ *Nulls*

- variables model *unknown* data values

Model of incompleteness

Employee

Smith
x_1
Brown

Manager

Smith	x_1
x_1	Brown
Brown	x_2

D

$\xrightarrow{[\![D]\!]}$

Smith
White
Brown

Smith	White
White	Brown
Brown	Black

Smith
Brown
Black
White

Smith	Brown
Brown	Brown

⋮

Semantics of incompleteness:

$[\![D]\!]$= a set of complete σ-structures

- *Closed world assumption*

  $[\![D]\!]_{CWA}$ = { R over *Const* | R = v(D) for some valuation v: *Nulls* → *Const* }

- *Open world assumption*

  $[\![D]\!]_{OWA}$ = { R over *Const* | R ⊇ v(D) for some valuation v: *Nulls* → *Const* }

- *Weak Closed World assumption* [Reiter 77]

  $[\![D]\!]_{WCWA}$ ={R over *Const* | R ⊇ v(D), dom(R)=dom(v(D)) for v: *Nulls* → *Const*}

Query answering over incomplete databases

For a *Boolean* query Q and an incomplete database D

- Query answering semantics (entailment):

  testing whether $R \models Q$ for all $R \in [\![D]\!]$

  (*certain answers*, in database terminology)

- Usual query answering in db systems (model checking) :

  testing whether $D \models Q$

  (*naïve evaluation*)

- Model-checking solves entailment for Q:

$$\boxed{\text{for all } D \quad D \models Q \quad \text{iff} \quad R \models Q \ \text{ for all } R \in [\![D]\!]}$$

(naïve evaluation works for Q)

  - correct query answering semantics (entailment) , classical query evaluation algorithms (model-checking)
  - clearly not always possible (undecidable vs. PTIME for FO)

A concrete example

"All employees are managers" $Q = \forall x(\ \mathrm{Employee}(x) \rightarrow \exists y\ \mathrm{Manager}(x, y)\ )$

Employee

Smith
x_1
Brown

Manager

Smith	x_1
x_1	Brown
Brown	x_2

$D \models Q$

$\longrightarrow$ $[\![D]\!]_{OWA}$

Smith
White
Brown
Black

Smith	White
White	Brown
Brown	Black

$R \not\models Q$

- Under OWA $\exists\ R \in [\![D]\!]$ s.t $R \not\models Q$: naïve evaluation does not work for Q
- Under CWA $\forall\ R \in [\![D]\!]$ $R \not\models Q$: naïve evaluation works for Q over D

What makes naïve evaluation work?

What makes naïve evaluation work?

What we already know:

Over incomplete relational databases (naïve tables), under the OWA,
if Q is Boolean FO query :

Naïve evaluation works for Q
$\Updownarrow$
Q is an $\exists$Pos query

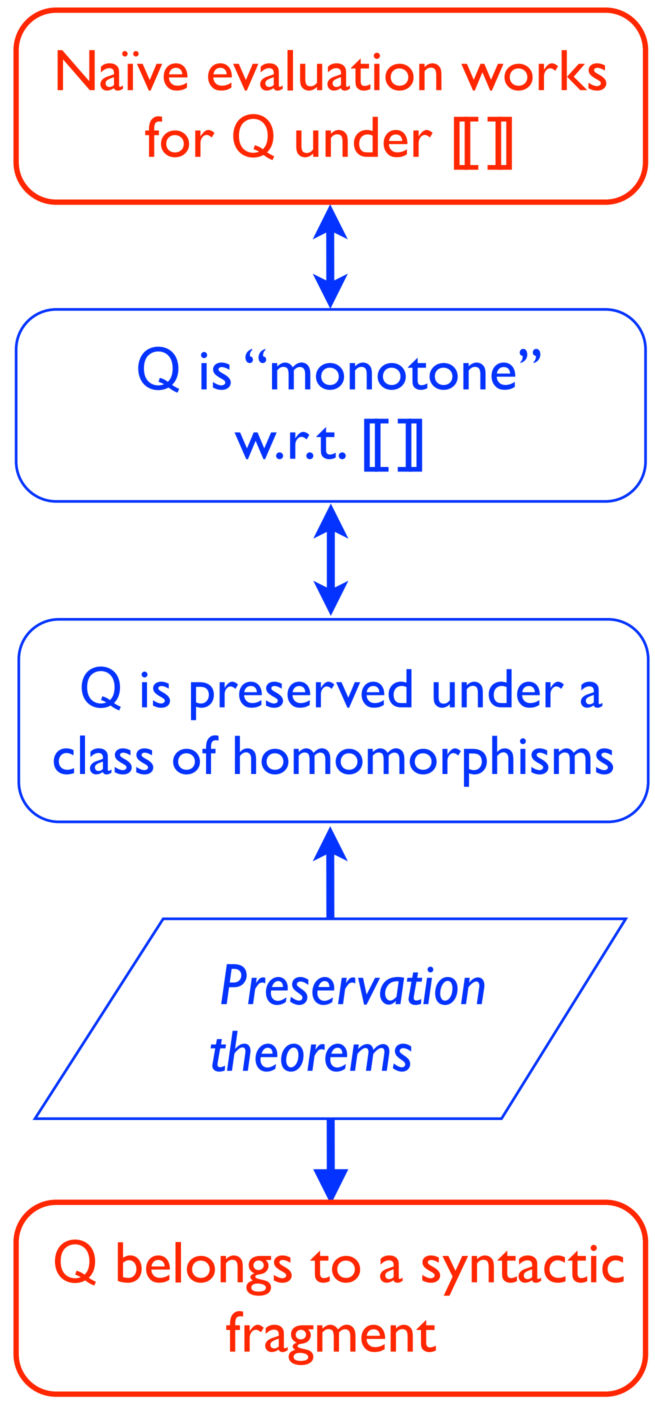
$\exists$Pos : $\exists, \wedge, \vee$ fragment of FO

(Unions of Conjunctive Queries in database terminology)

- The $\Uparrow$ direction [Imielinski, Lipski 84]
- The $\Downarrow$ direction [Libkin 2011] relies on Rossman's homomorphism preservation theorem in the finite

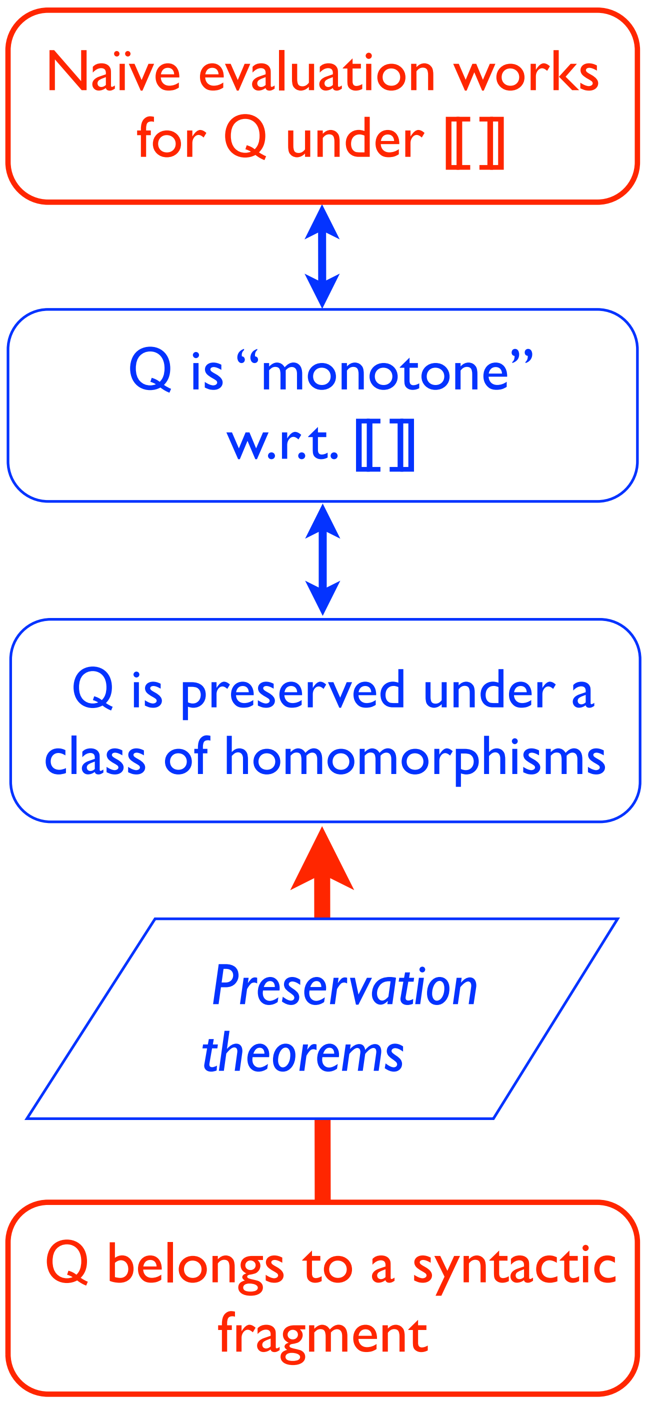
Relating naïve evaluation and syntactic fragments

A unified framework for relating naïve evaluation and syntactic fragments for several possible semantics:



Relating naïve evaluation and syntactic fragments

A unified framework for relating naïve evaluation and syntactic fragments for several possible semantics:



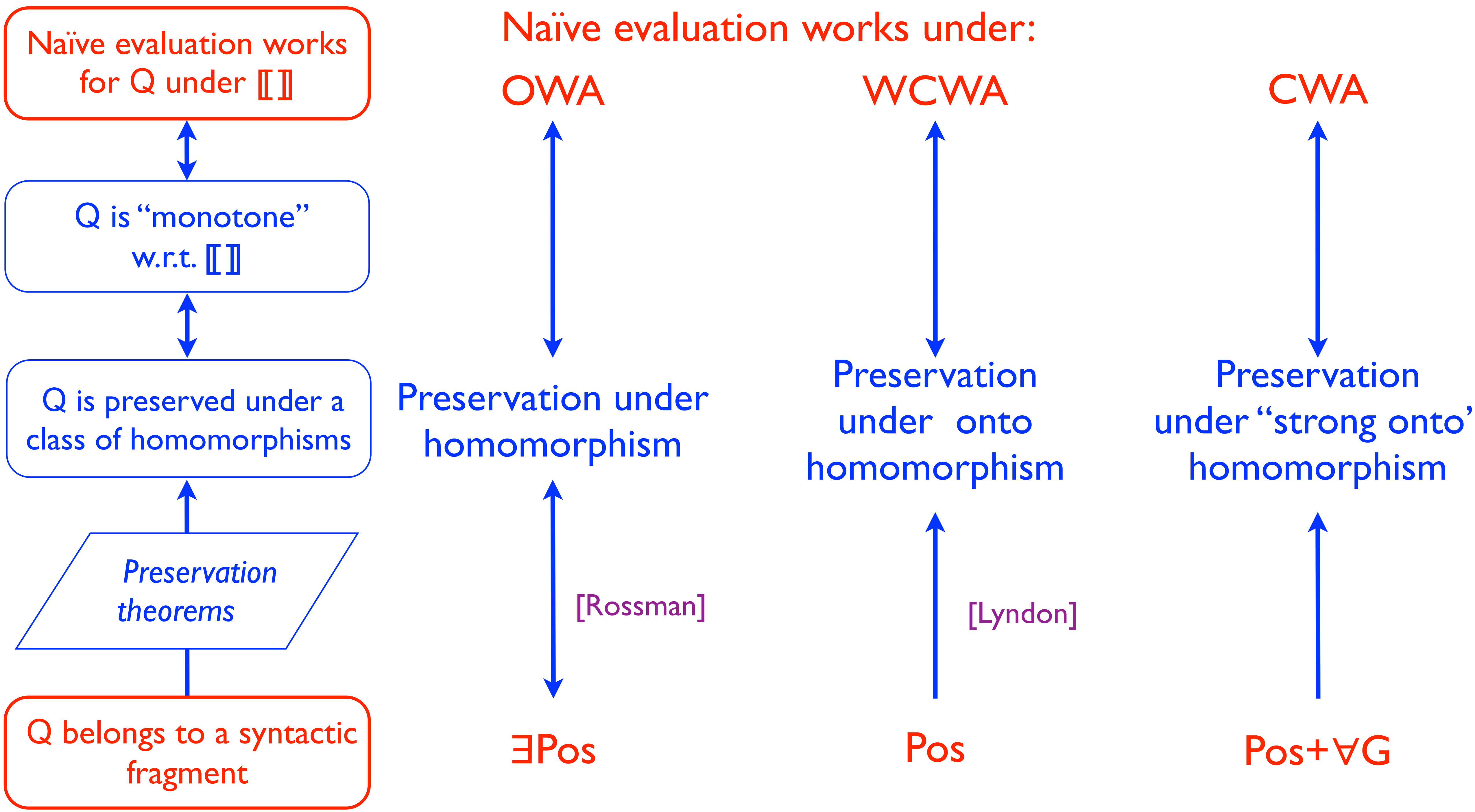
Preservation theorems:

- Usually proved over arbitrary structures (both finite and infinite)
- some fail in the finite
- the direction Syntax ⇒ Preservation always holds in the finite as well

Preservation theorems (even over arbitrary structures) can give us relevant classes of queries where naïve evaluation works

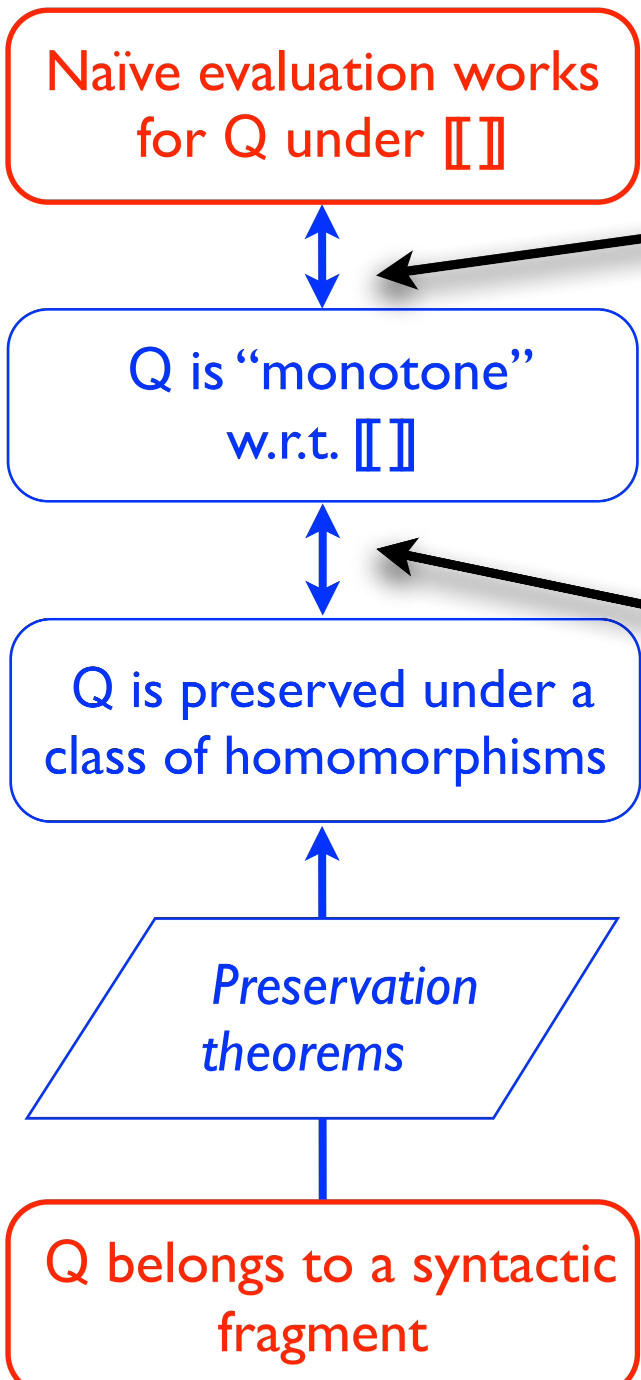
Naïve evaluation and syntactic fragments

Three well known semantics as instances of our framework



Naïve evaluation and syntactic fragments

The framework is much more general



Holds in a very general setting:

- arbitrary domain of database objets equipped with a semantic function
- do not even need to be relational databases
- very mild assumptions on the domain ("*saturated subdomain*")

Holds for a whole class of relational semantics:

- semantics described by relations on structures
- general notion of homomorphism based on these relations

Naïve evaluation and syntactic fragments

Beyond OWA, CWA and WCWA:

Naïve evaluation works for Q under $[\![\ ]\!]$

Q is "monotone" w.r.t. $[\![\ ]\!]$

Q is preserved under a class of homomorphisms

Preservation theorems

Q belongs to a syntactic fragment

Naïve evaluation works under:

Powerset semantics	Minimal semantics
Preservation under unions of strong onto homomorphisms	over cores Preservation under unions of minimal homomorphisms
$\exists\mathrm{Pos}{+}\forall\mathrm{G}^{\mathrm{bool}}$	$\exists\mathrm{Pos}{+}\forall\mathrm{G}^{\mathrm{bool}}$

Reference

Details in our paper:

"When is Naive Evaluation Possible?" PODS 2013

by Amélie Gheerbrant, Leonid Libkin and Cristina Sirangelo

Conclusions and future work

- A general framework for relating naïve evaluation and syntactic fragments
  - applied to (generalizations of) existing relational semantics
- All results extend to non-boolean relational queries
- Extend to other data models
  - more complex form of relational incompleteness (e.g. conditional tables), incomplete trees, incomplete graphs
- Preservation theorems
  - new notions of preservation, candidate fragments, preservation theorems in the infinite?
  - do they hold in the finite?
- Extend to other languages: fixed-point, fragments of SO, etc.
- Naïve evaluation over restricted instances/ in the presence of constraints

Syntactic fragments

- **Pos :** FO without negation (but with $\forall$)
  - Pos = FO queries preserved under onto homomorphisms over arbitrary structures (Lyndon positivity theorem)
- **Pos+$\forall$G :** Positive fragment with Universal Guards

$$\varphi := \top \mid \bot \mid R(\bar{x}) \mid x = y \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \exists x \varphi \mid \forall x \varphi \mid$$

$$\left(\forall \bar{x} \; (\; G(\bar{x}) \to \varphi \;) \quad \begin{array}{l} \text{with} \\ G : \text{a relation or equality symbol} \\ \bar{x} : \text{a tuple of distinct variables} \end{array} \right)$$

  - preserved under strong onto homomorphisms, a good syntax
  - extends [Keisler '65] (complex syntactic restrictions, one binary relation only)

The most general setting: database domains

Database domain: a quadruple $\langle \mathcal{D}, C, [\![\ ]\!], \approx \rangle$

	description	example
$\mathcal{D}$: a set	database objects (complete and incomplete)	all naïve relational instances over a fixed schema σ
C : a subset of $\mathcal{D}$	complete database objects	all complete relational instances over σ
$[\![\ ]\!] : \mathcal{D} \to 2^C$	semantics of incompleteness	$[\![\ ]\!]_{\mathrm{OWA}}$, $[\![\ ]\!]_{\mathrm{CWA}}$, etc.
$\approx$: an equivalence relation on $\mathcal{D}$	equivalence of objects (w.r.t. queries)	isomorphism of relational instances

The most general setting: database domains

Over $\langle \mathcal{D}, C, [\![\ ]\!], \approx \rangle$

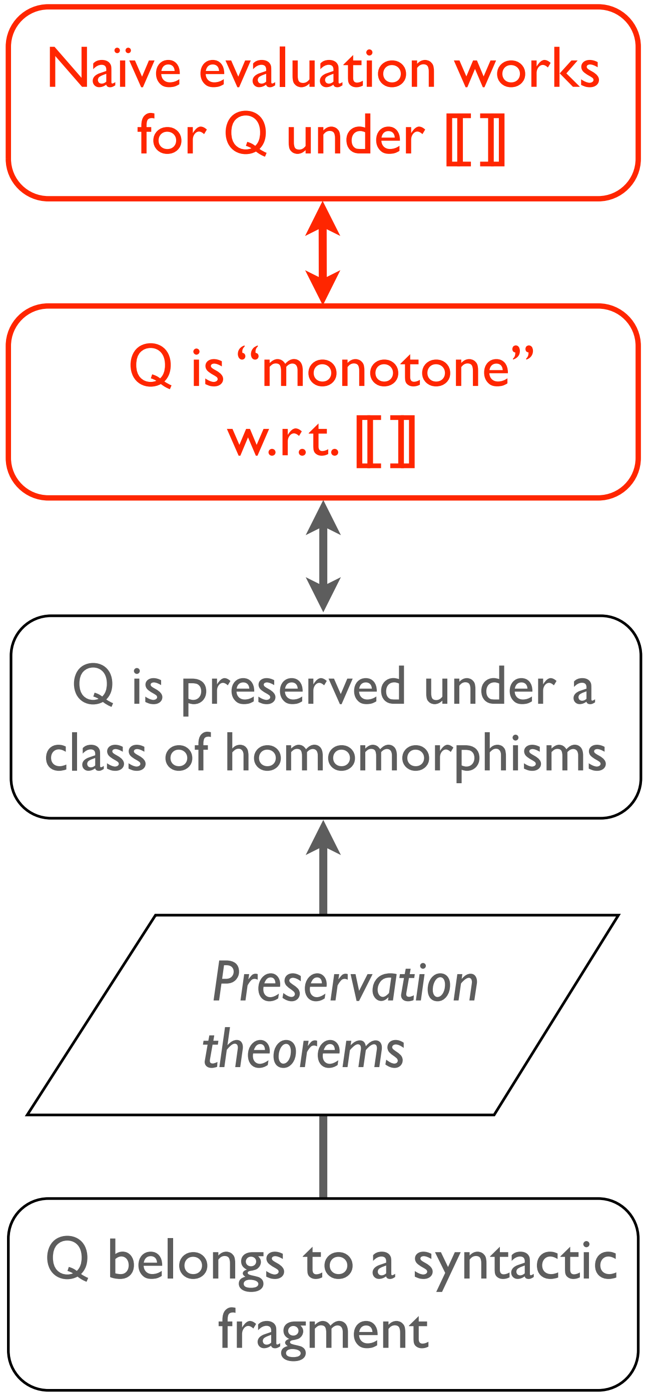
- Boolean query: $Q : \mathcal{D} \to \{\mathit{true}, \mathit{false}\}$
  - Q generic : $x \approx y$ implies $Q(x)=Q(y)$
  - Q monotone w.r.t. $[\![\ ]\!]$: $y \in [\![x]\!]$ implies $Q(x) \Rightarrow Q(y)$
- Certain answers for $x \in \mathcal{D}$:

$$\mathrm{cert}(Q, x) = \bigwedge_{c \in [\![x]\!]} Q(c)$$

- Naïve evaluation works for Q :

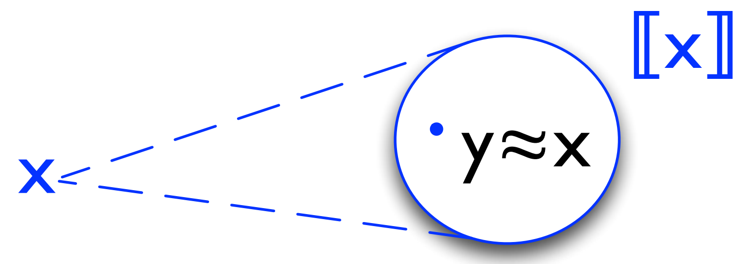
$$\text{For all } x \in \mathcal{D} \quad Q(x) = \mathrm{cert}(Q, x)$$

Naïve evaluation and monotonicity



Saturation property for $\langle\, \mathcal{D}, C\,, [\![\ ]\!]\,, \approx \,\rangle$:

For all $x \in \mathcal{D}$ there exists $y \in [\![x]\!]$ $y \approx x$



holds for most common semantics

Proposition

Over a saturated database domain,
if Q is a generic Boolean query:

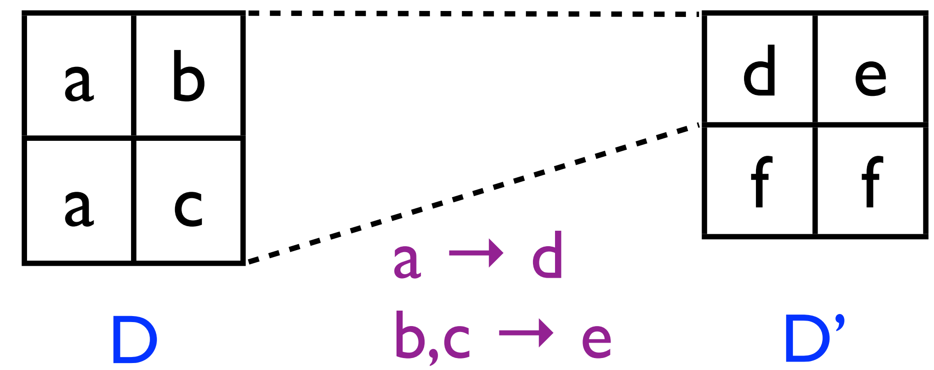
Naïve evaluation works for Q iff
Q is monotone w.r.t. $[\![\ ]\!]$

Homomorphisms

Homomorphism $D \rightarrow D'$:

a mapping $h: dom(D) \rightarrow dom(D')$ s.t.

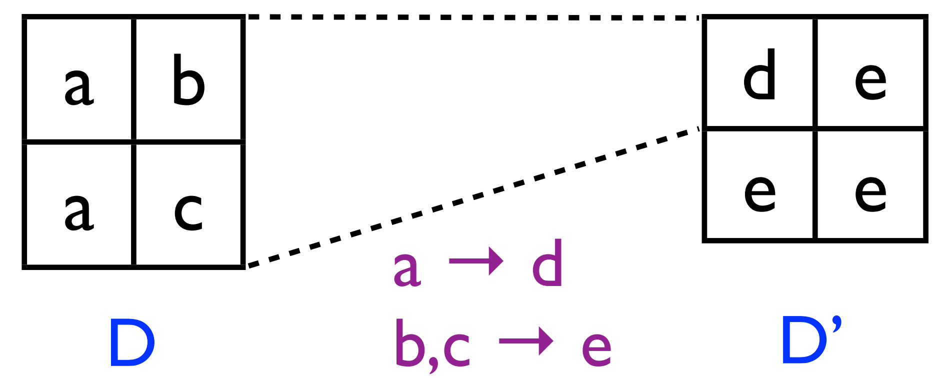
$h(D) \subseteq D'$



Onto homomorphism $D \rightarrow D'$:

a homomorphism $h: D \rightarrow D'$ s.t.

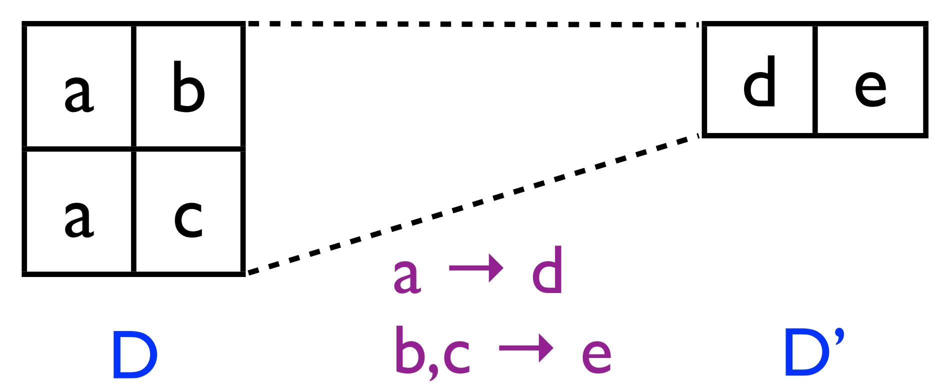
$h(dom(D)) = dom(D')$



Strong onto homomorphism $D \rightarrow D'$:

a homomorphism $h: D \rightarrow D'$ s.t.

$h(D) = D'$



Homomorphisms

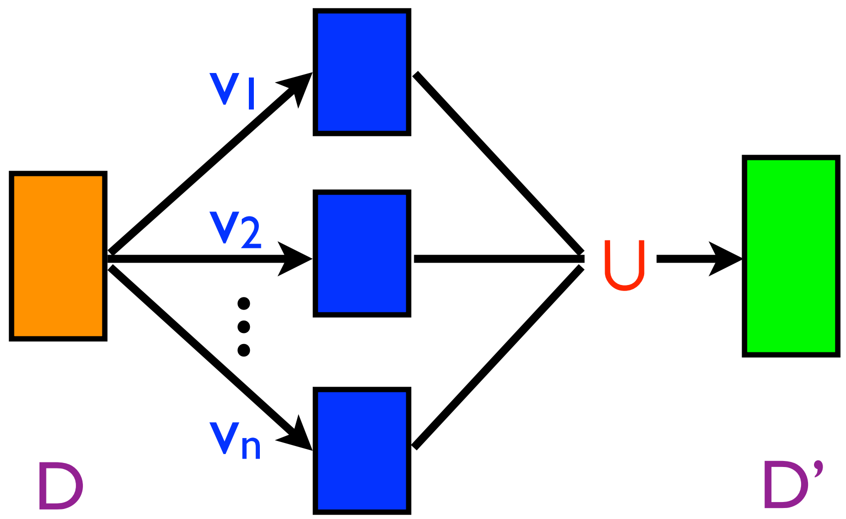
- Union of strong onto homomorphisms $D \rightarrow D'$: $\bigcup_i h_i(D) = D'$

- D-minimal homomorphism h on D :
  there exists no h', preserving all constants preserved by h, s.t. $h'(D) \subsetneq h(D)$

- Union of minimal homomorphisms $D \rightarrow D'$: $\bigcup_i h_i(D) = D'$

  with $h_1 \ldots h_n$ D-minimal and preserving the same constants

Powerset semantics



Powerset CWA

$D' \in (\!|D|\!)_{CWA}$ iff

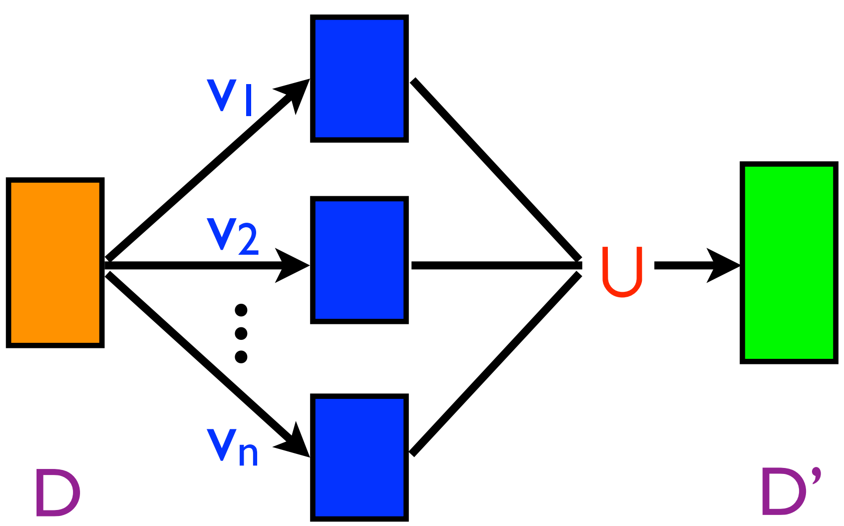
$\exists$ valuations $v_1, \ldots v_n$ $D' = \cup_i v_i(D)$

- Extend (and generalize) an ordering-based semantics of Codd databases
- Naïve evaluation $\leftrightarrow$ Monotonicity $\leftrightarrow$ Preservation continues to hold
- Under the powerset CWA the needed notion is preservation under *unions of strong onto homomorphisms* (i.e. homomorphisms $D \rightarrow \bigcup_{i=1}^{n} h_i(D)$)
- An FO fragment preserved under this relationship: $\exists Pos+\forall G^{bool}$

> Corollary
>
> Naïve evaluation works for $\exists Pos+\forall G^{bool}$ Boolean queries under $(\!| \cdot |\!)_{CWA}$

Minimal semantics

- A special form of powerset semantics, finds its roots in [Minker '82]
- Later modified and adopted as *data exchange* semantics (GCWA* [Hernich'11])
- We define it here for arbitrary incomplete instances:



Minimal Powerset CWA

$D' \in (\!|D|\!)^{min}_{CWA}$ iff

$\exists$ *D-minimal* valuations $v_1, \ldots v_n$

$D' = \cup_i \, v_i(D)$

A valuation v on D is D-minimal if there is no valuation v' s.t. $v'(D) \subsetneq v(D)$

- Under the minimal powerset CWA the saturation property does not hold
- Cores come to the rescue: naive evaluation recovered over cores

Non-Boolean queries

All results can be lifted to non-boolean relational queries.

- unified technique: reduction to the boolean case

For k-ary FO queries , $k \geq 0$

Semantics	Naïve evaluation works for
OWA	$\exists$Pos
WCWA	Pos
CWA	Pos+$\forall$G
Powerset CWA	$\exists$Pos+$\forall$G^{bool}
Min Powerset CWA	$\exists$Pos+$\forall$G^{bool} iff Q(D)=Q(core(D))