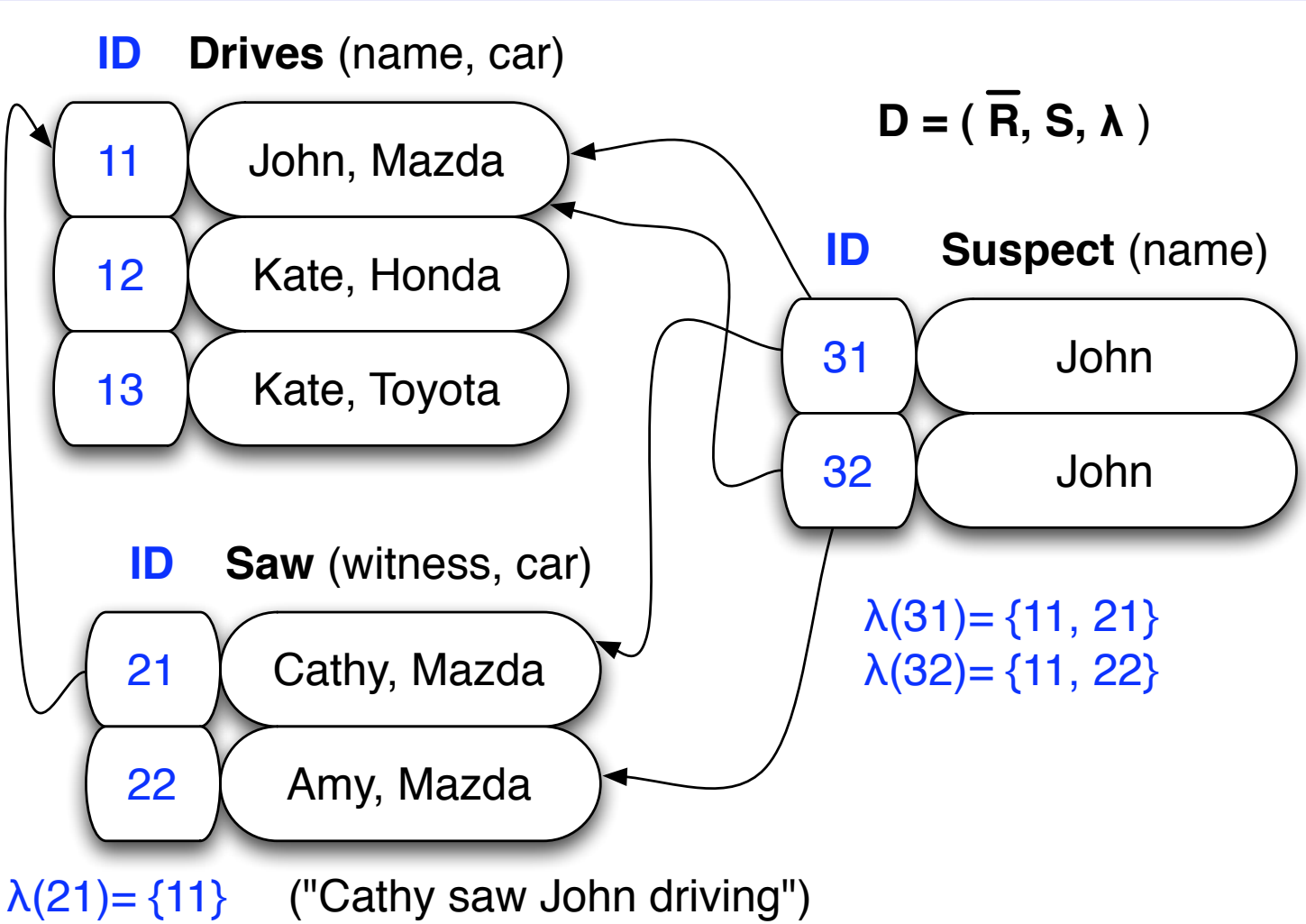
Query Containment for Databases with Uncertainty and Lineage

Foto N. Afrati Angelos Vasilakopoulos

National Technical University of Athens

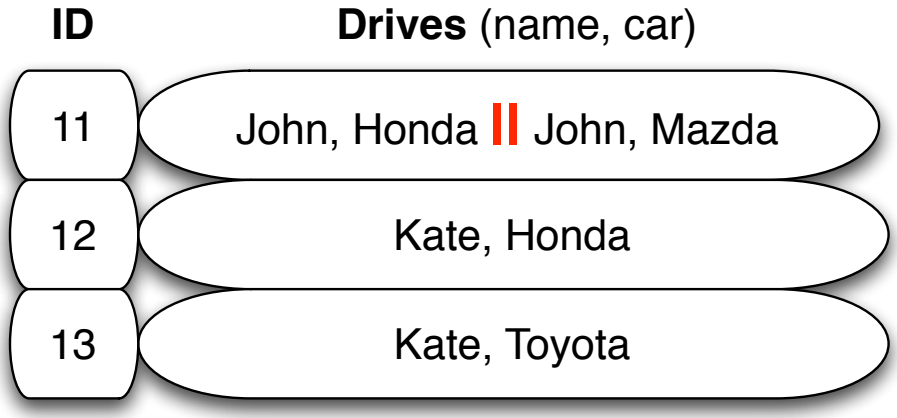
July 15, 2013

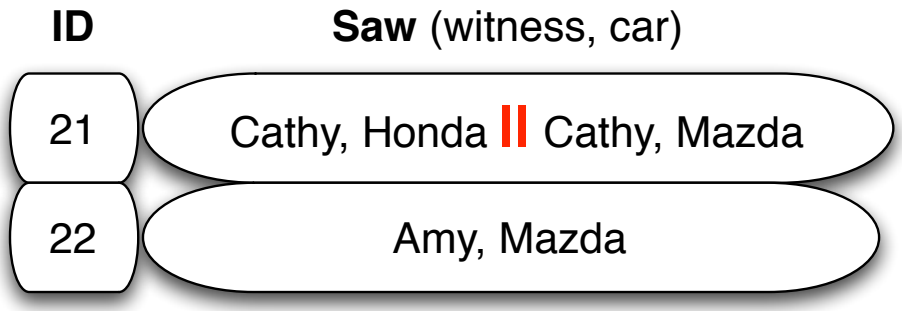
Databases with Lineage - LDBs



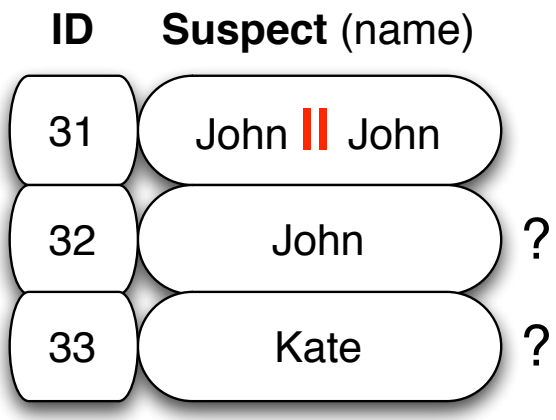
$$Suspect(name) : -Drives(name, car), Saw(witness, car)$$

Databases with Uncertainty and Lineage - ULDBs





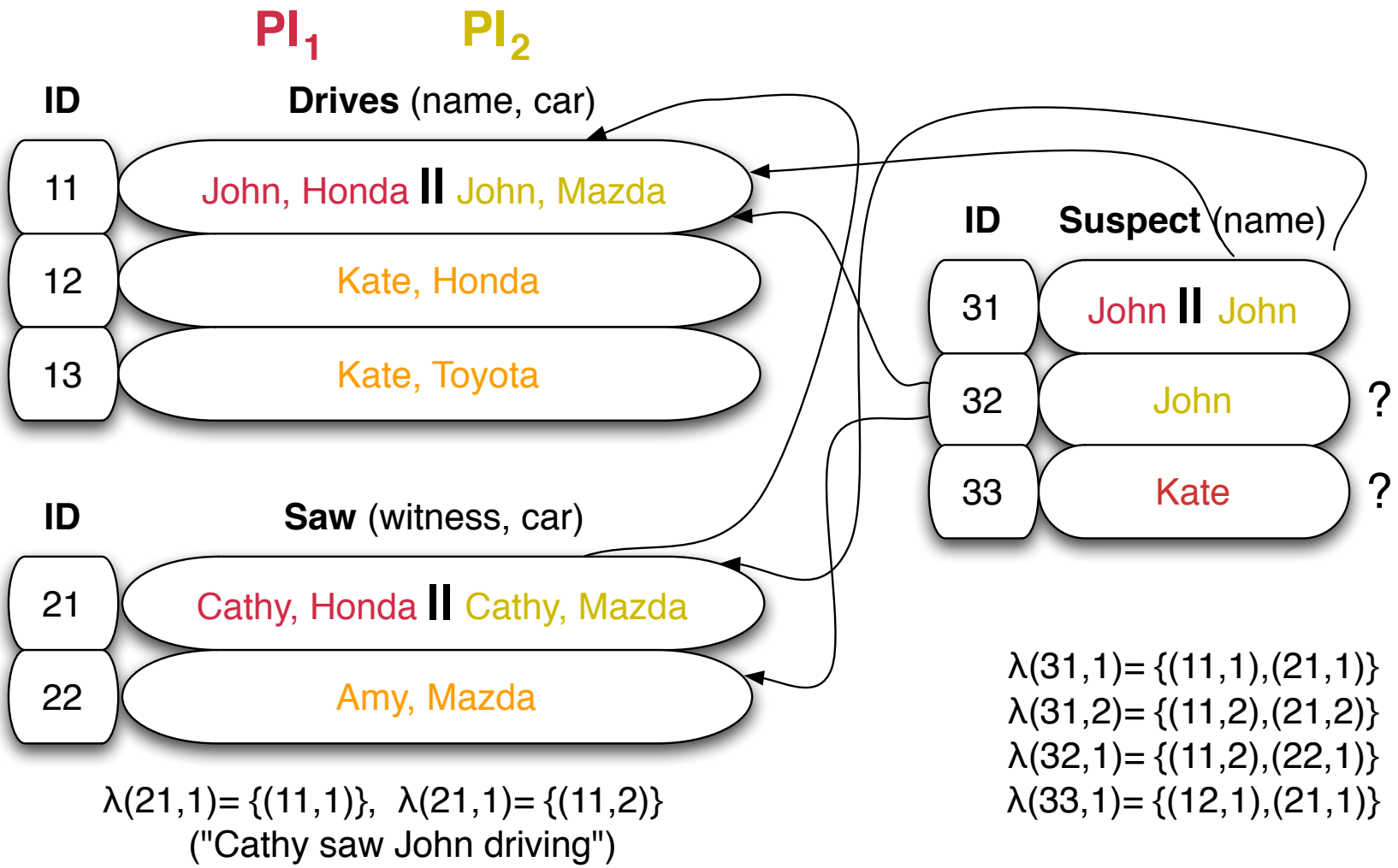
$\lambda(21,1) = \{(11,1)\}$, $\lambda(21,1) = \{(11,2)\}$
("Cathy saw John driving")



$\lambda(31,1) = \{(11,1),(21,1)\}$
$\lambda(31,2) = \{(11,2),(21,2)\}$
$\lambda(32,1) = \{(11,2),(22,1)\}$
$\lambda(33,1) = \{(12,1),(21,1)\}$

$$Suspect(name) :- Drives(name, car), Saw(witness, car)$$

Possible Instances - PIs



$$Suspect(name) :- Drives(name, car), Saw(witness, car)$$

Query Containment

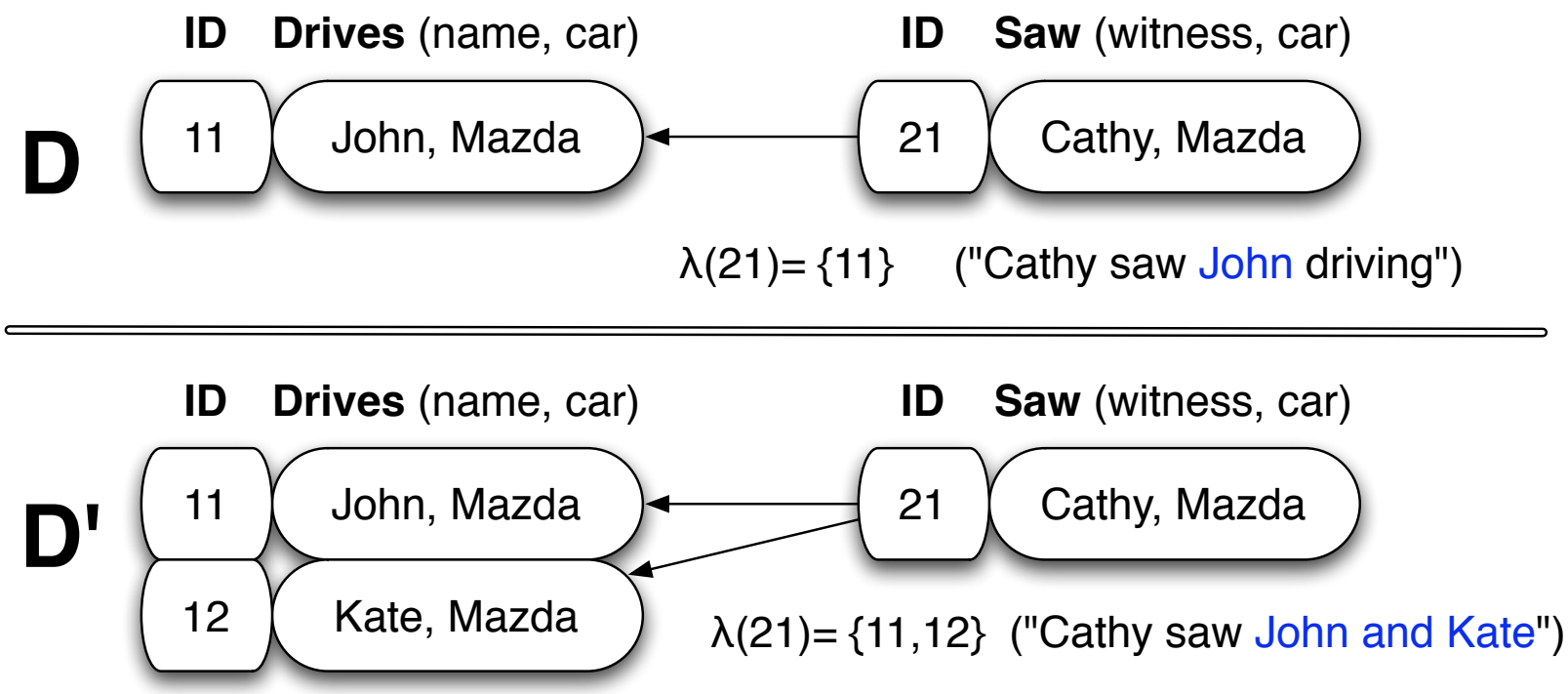
A query Q_1 is said to be contained in a query Q_2 if for every database D, database $Q_1(D)$ is contained in database $Q_2(D)$.

For ordinary databases, a database D_1 is contained in a database D_2 if the tuples of every relation in D_1 are contained in the corresponding relation of D_2 as a set.

A relation of ULDB however semantically is not a set; it represents a set of possible LDB instances - PIs.

A possible instance PI, is an LDB and does not only contain a set of tuples, but a bag of tuples with different lineage information.

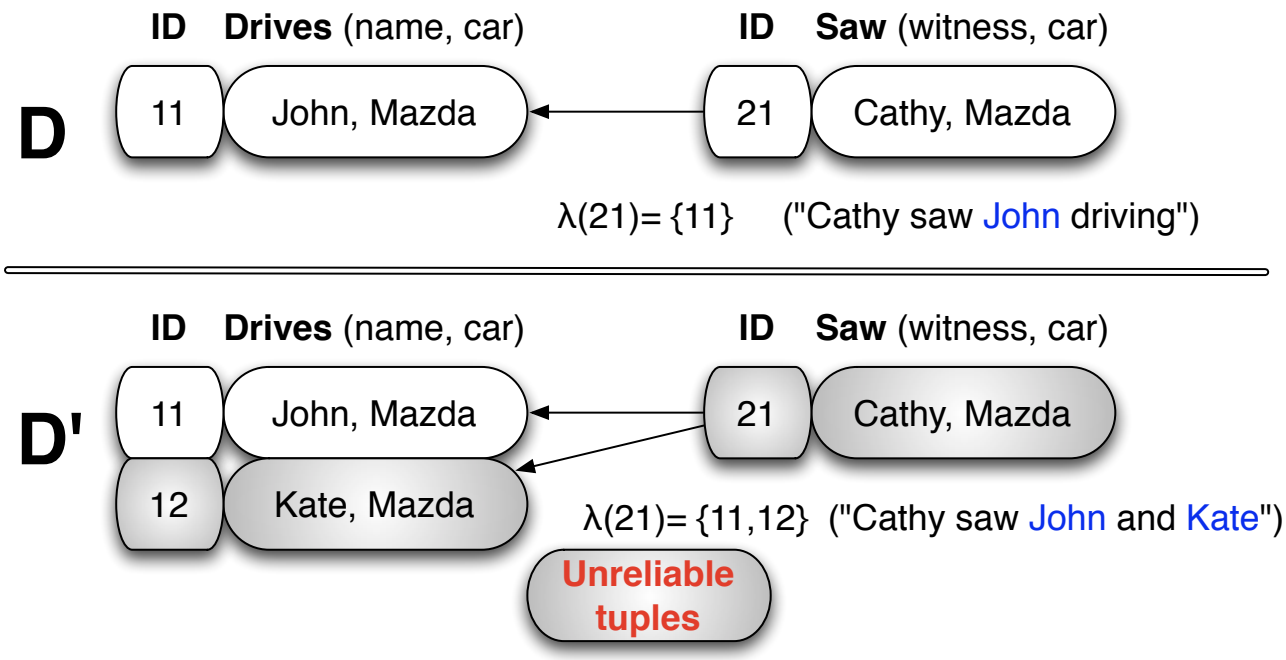
LDB Database containment



- $D \subseteq D'$ under data set containment.
- $D \not\subseteq D'$ if we take lineage into account
  ($12 \in \lambda'(21)$, but $12 \notin \lambda(21)$, so $\lambda'_B(21) \not\subseteq \lambda_B(21)$) .

When does linage matter?

If tuple 12 (*Kate*, *Mazda*) was considered unreliable then tuple 21 would be unreliable in D', while still reliable in D.



A system might want to delete unreliable data. Then even data set containment $D \subseteq D'$ will no longer hold.

Goals

LDB and ULDB Database containment

- Introduce several variants of LDB database containment.
- Define corresponding different kinds of ULDB Database containment.
- Discuss some cases where each semantics may be suitable.
- Study exact interrelationship among them as concerns implication of database containment.

Query Containment

- Study ULDB Query containment for Conjunctive Queries (CQs) under each of the semantics.
- Computational complexity.

Semantics #1: Data containment - $\subseteq_{Data}$

Data LDB Containment $\subseteq_{Data}$

Let $D = (\bar{R}, S, \lambda)$ and $D' = (\bar{R}', S', \lambda')$ be two LDBs, where $\bar{R}$ and $\bar{R}'$ have the same schemas. We say that D is *Data* LDB-contained in D' (denoted as $D \subseteq_{Data} D'$), if:

1. $S_- \subseteq S'_-$.
2. For every relation $R_i \in D$ and its corresponding $R'_i \in D'$ the following holds:
if $t \in R_i$ then there exists a tuple with data t in R'_i.

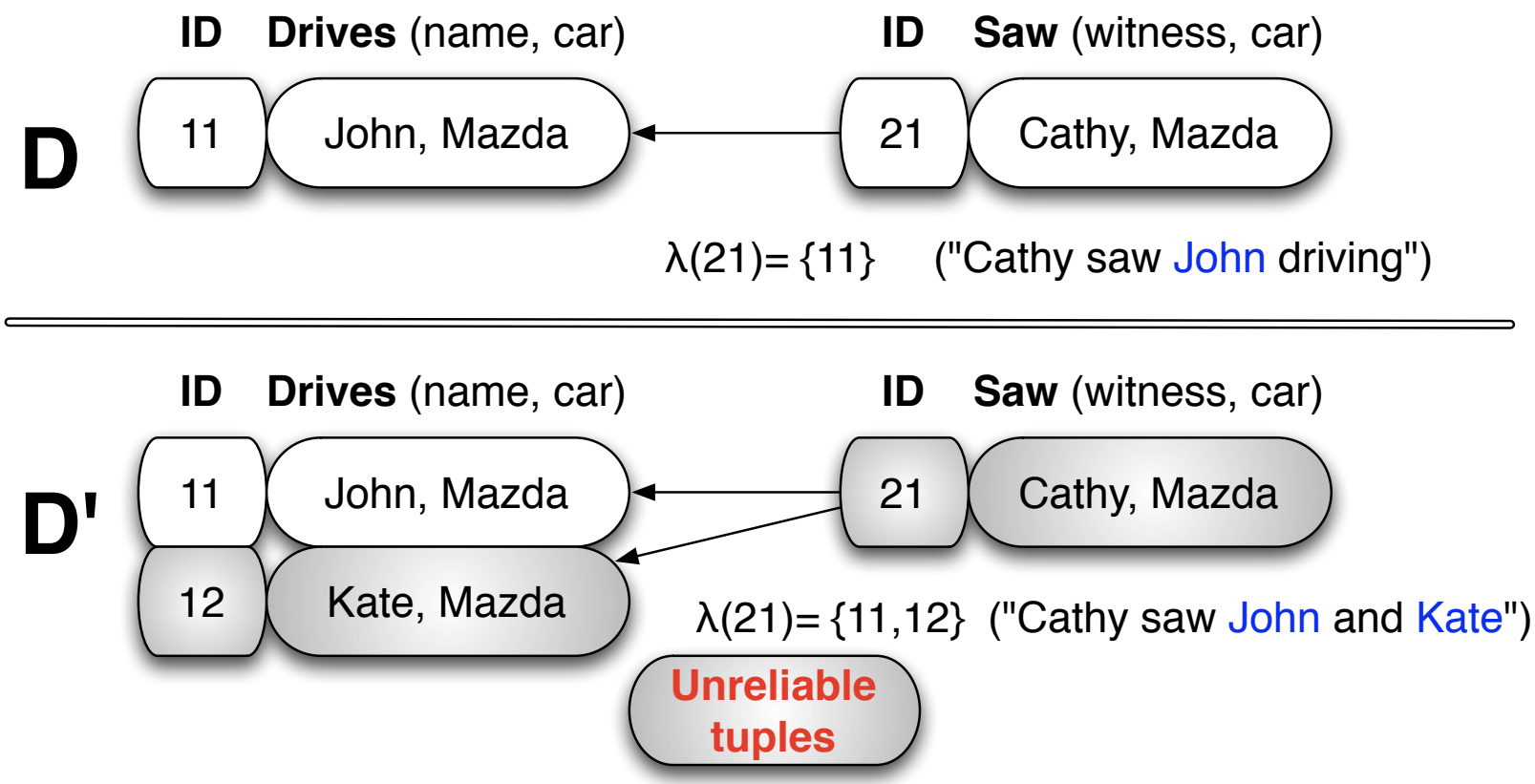
Semantics #2: Contained Base Lineage - $\subseteq_{CBase}$

Contained Base Lineage (CBase-lineage)
LDB Containment $\subseteq_{CBase}$

Let $D = (\bar{R}, S, \lambda)$ and $D' = (\bar{R}', S', \lambda')$ be two LDBs, where $\bar{R}$ and $\bar{R}'$ have the same schemas. We say that D is *CBase-Lineage* LDB-contained in D' (denoted as $D \subseteq_{CBase} D'$), if:

1. $S_{-} \subseteq S'_{-}$.
2. For every relation $R_i \in D$ and its corresponding $R'_i \in D'$ the following holds:
if $t \in R_i$ then there exists a tuple with data t in R'_i and $COND_2$: $\lambda'_B(t) \subseteq \lambda_B(t)$.

LDB Database containment



Under Semantics #1: $D \subseteq_{Data} D'$
Under Semantics #2: $D \not\subseteq_{CBase} D'$

$12 \in \lambda'(21)$, but $12 \notin \lambda(21)$, so $\lambda'_B(21) \not\subseteq \lambda_B(21)$.

Semantics #3: Trio/Transitive Closure - $\subseteq_{TR}$

Semantics #3: Trio/Transitive Closure of Lineage Containment (TR-lineage - $\subseteq_{TR}$).

Let $D = (\bar{R}, S, \lambda)$ and $D' = (\bar{R}', S', \lambda')$ be two LDBs, where $\bar{R}$ and $\bar{R}'$ have the same schemas. We say that D is *TR-lineage* LDB-contained in D' (denoted as $D \subseteq_{TR} D'$), if:

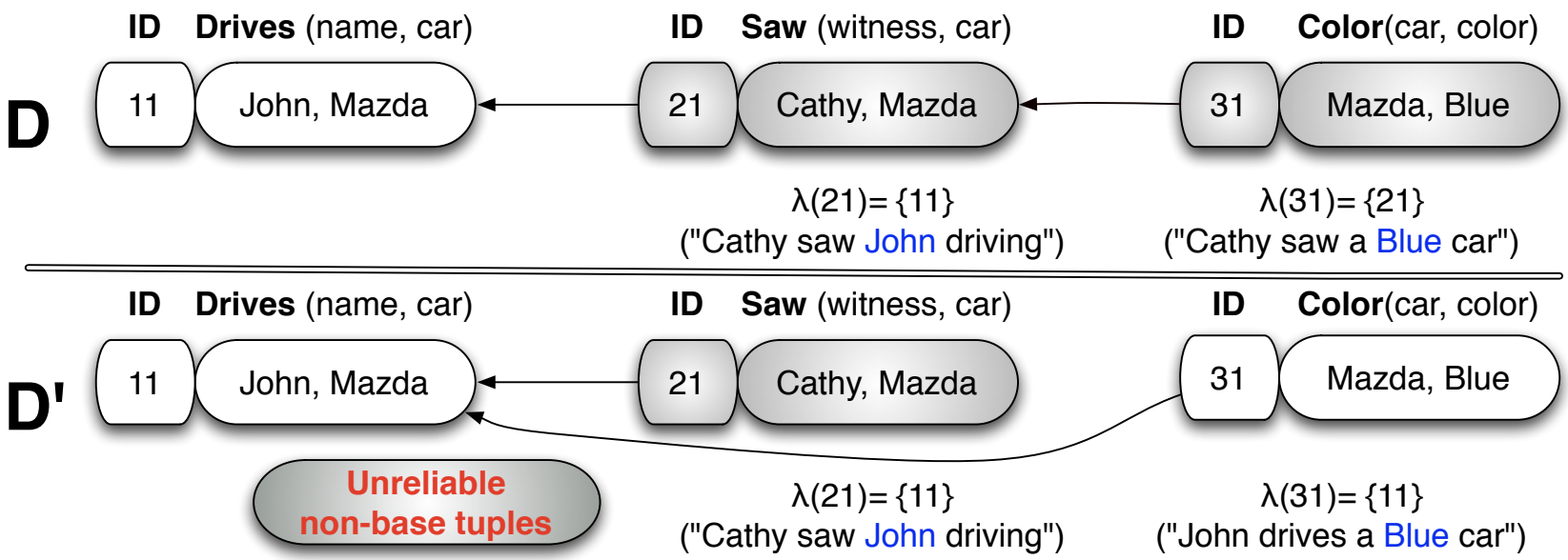
1. $S_- \subseteq S'_-$.
2. For every relation $R_i \in D$ and its corresponding $R'_i \in D'$ the following holds:

if $t \in R_i$ then there exists a tuple with data t in R'_i and $COND_3$: $\lambda(t) \subseteq \lambda'^*(t)$.

Semantics #3: Trio/Transitive Closure - $\subseteq_{TR}$

Semantics #3: Trio/Transitive Closure of Lineage Containment (TR-lineage - $\subseteq_{TR}$).

The additional condition is: $COND_3$: $\lambda(t) \subseteq \lambda'^*(t)$.



$D \subseteq_{Data} D'$, $D \subseteq_{CBase} D'$, $D \not\subseteq_{TR} D'$

Semantics #4: Same Base-lineage - $\subseteq_{SBase}$

Semantics #4: Same Base-Lineage (SBase-lineage) LDB Containment $\subseteq_{SBase}$

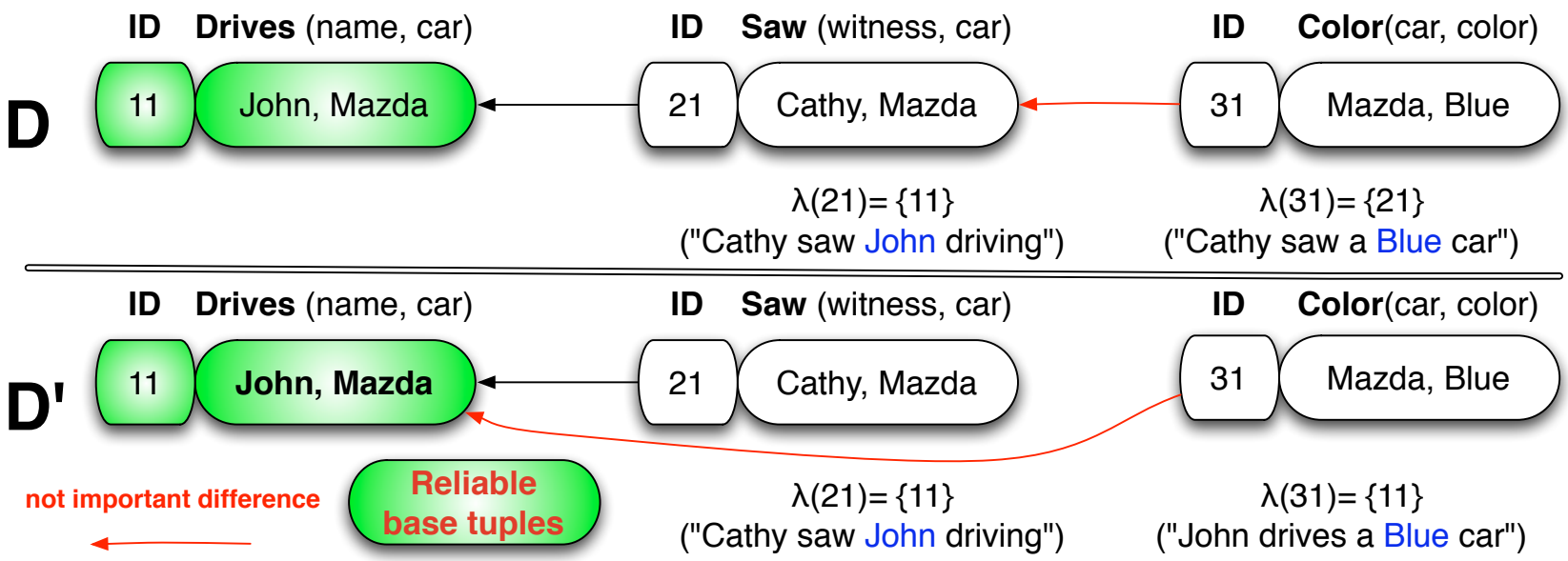
Let $D = (\bar{R}, S, \lambda)$ and $D' = (\bar{R}', S', \lambda')$ be two LDBs, where $\bar{R}$ and $\bar{R}'$ have the same schemas. We say that D is *Same Base-Lineage* LDB-contained in D' (denoted as $D \subseteq_{SBase} D'$), if:

1. $S_- \subseteq S'_-$.
2. For every relation $R_i \in D$ and its corresponding $R'_i \in D'$ the following holds:

if $t \in R_i$ then there exists a tuple with data t in R'_i and $COND_4$: $\lambda'_B(t) = \lambda_B(t)$.

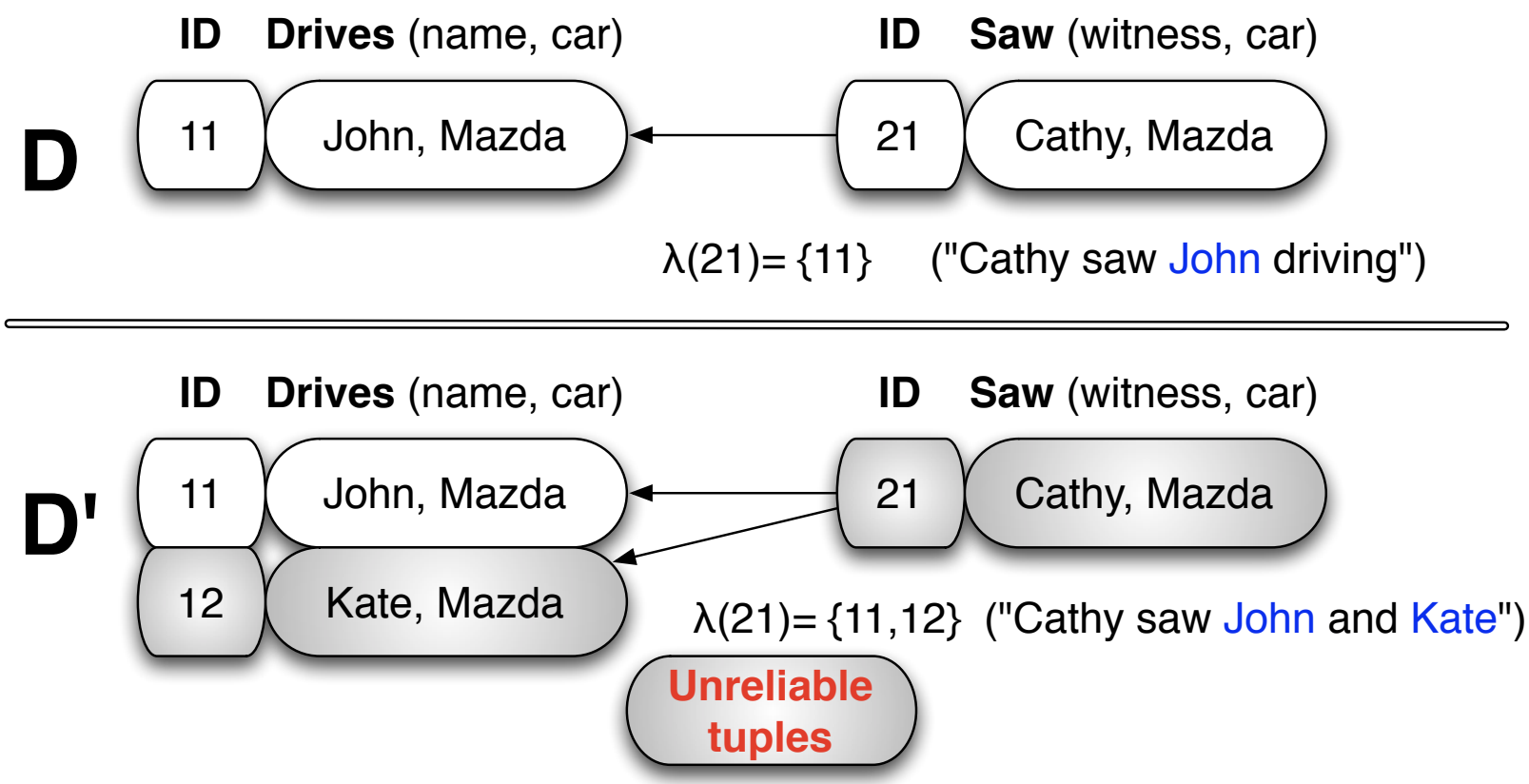
Semantics #4: Same Base-Lineage (SBase-lineage) LDB Containment $\subseteq_{SBase}$

The additional condition is: $COND_4$: $\lambda'_B(t) = \lambda_B(t)$.



$$D \subseteq_{Data} D', \quad D \subseteq_{CBase} D',$$
$$D \not\subseteq_{TR} D', \quad D \subseteq_{SBase} D', \quad D \not\subseteq_{Same} D'$$

SBase Lineage is also important in ULDB data exchange.



$$D \subseteq_{Data} D', \quad D \not\subseteq_{CBase} D',$$
$$D \subseteq_{TR} D', \quad D \not\subseteq_{SBase} D', \quad D \not\subseteq_{Same} D'$$

Semantics #5: Same-Lineage LDB Containment $\subseteq_{Same}$

Semantics #5: Same-Lineage LDB Containment $\subseteq_{Same}$

Let $D = (\bar{R}, S, \lambda)$ and $D' = (\bar{R}', S', \lambda')$ be two LDBs, where $\bar{R}$ and $\bar{R}'$ have the same schemas. We say that D is *Same Lineage* LDB-contained in D' (denoted as $D \subseteq_{Same} D'$), if:

1. $S_- \subseteq S'_-$.
2. For every relation $R_i \in D$ and its corresponding $R'_i \in D'$ the following holds:

if $t \in R_i$ then there exists a tuple with data t in R'_i and $COND_5$: $\lambda'(t) = \lambda(t)$.

Adding Uncertainty

ULDB Database Containment

Let $\subseteq_L$ denote a variant of LDB containment.
Let U and U' be two ULDB's. We say that U is L-contained in U' (denoted with $\subseteq_L$) if:
i) for every possible instance D_i of U there exists a possible instance D'_j of U' such that: $D_i \subseteq_L D'_j$, and
ii) for every possible instance D'_j of U' there exists a possible instance D_i of U such that: $D_i \subseteq_L D'_j$.

ULDB Query Containment

Let $\subseteq_L$ denote a variant of LDB containment.
A query Q_1 is ULDB contained in a query Q_2 if for every ULDB U we have that: $Q_1(U) \subseteq_L Q_2(U)$.

Our Results 1:

Comparison of Different Semantics

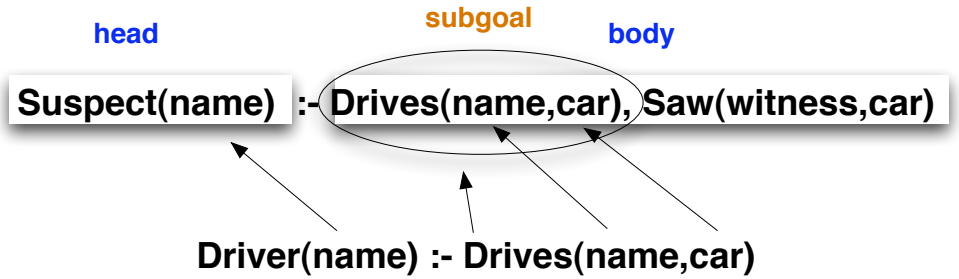
#	Semantics	Implies DB cont.
1	Data	-
2	CBase- Lineage	1
3	TR- Lineage	1
4	SBase- Lineage	1, 2
5	Same Lineage	1, 2, 3, 4

Our Results 2:

Complexity of ULDB Conjunctive Query Containment

#	Semantics	CQ Containment Test	Complexity
1	Data	Containment Mapping	NP-Complete
2	CBase- Lineage	Containment Mapping	NP-Complete
3	TR- Lineage	Onto Containment Mapping	NP-Complete
4	SBase- Lineage	Onto Containment Mapping	NP-Complete
5	Same Lineage	Onto Containment Mapping	NP-Complete

Containment mapping and subgoal-onto containment mapping



Containment Mapping $h: Q' \to Q$

$h : values(Q') \to values(Q)$:

- $\forall$ constants c: $h(c) = c$, $h(head(Q')) = head(Q)$
- every atom in the body of Q' is mapped to an atom of the body of Q with the same predicate.

Subgoal-onto Containment Mapping

A containment mapping from Q' to Q is subgoal-onto if we additionally have that the set of images of all the subgoals of Q' contains every subgoal of the body of Q.

Semantics #6: Uncertain Equality containment - $\subseteq_E$

A new kind of containment for uncertain databases with no lineage was defined in:
"Foundations of uncertain-data integration.
P. Agrawal, A. D. Sarma, J. Ullman, and J. Widom. VLDB 2010."

Informally equality containment $U_1 \subseteq_E U_2$ means that if we throw away from the possible worlds of U_2 all tuples that do not appear in any possible world of U_1, then the resulting possible worlds are the worlds of U_1.

Example (subgoal-onto containment mapping is not good)

$U = \{\{(a,a)\},\{(b,b)\},\{(a,b),(b,a)\}\},\ a \neq b$
$Q_1(x) :- R(x,x)$
$Q_2(x) :- R(x,y).$
$\exists h : Q_2 \to Q_1$: subgoal-onto

$Q_1(U) = \{\{a\},\{b\},\emptyset\}$
$Q_2(U) = \{\{a\},\{b\},\{a,b\}\}$

CQ Containment test and complexity:

Given two conjunctive queries Q_1 and Q_2 we have that $Q_1 \subseteq_E Q_2$ iff there exists a containment mapping h: $Q_2 \to Q_1$ and a containment mapping h': $Q_1 \to Q_2$.
In addition checking whether $Q_1 \subseteq_E Q_2$ is NP-complete.

Thank you