



Datenintegration & Datenherkunft

Datenherkunft fehlender Daten

Wintersemester 2010/11

Melanie Herschel

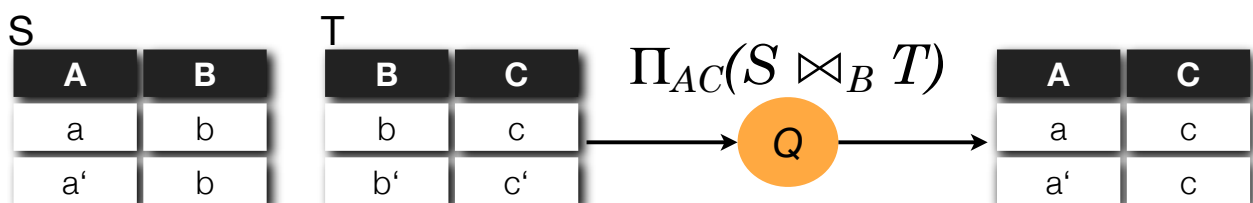
melanie.herschel@uni-tuebingen.de

Lehrstuhl für Datenbanksysteme, Universität Tübingen

1

Erklären fehlender Daten

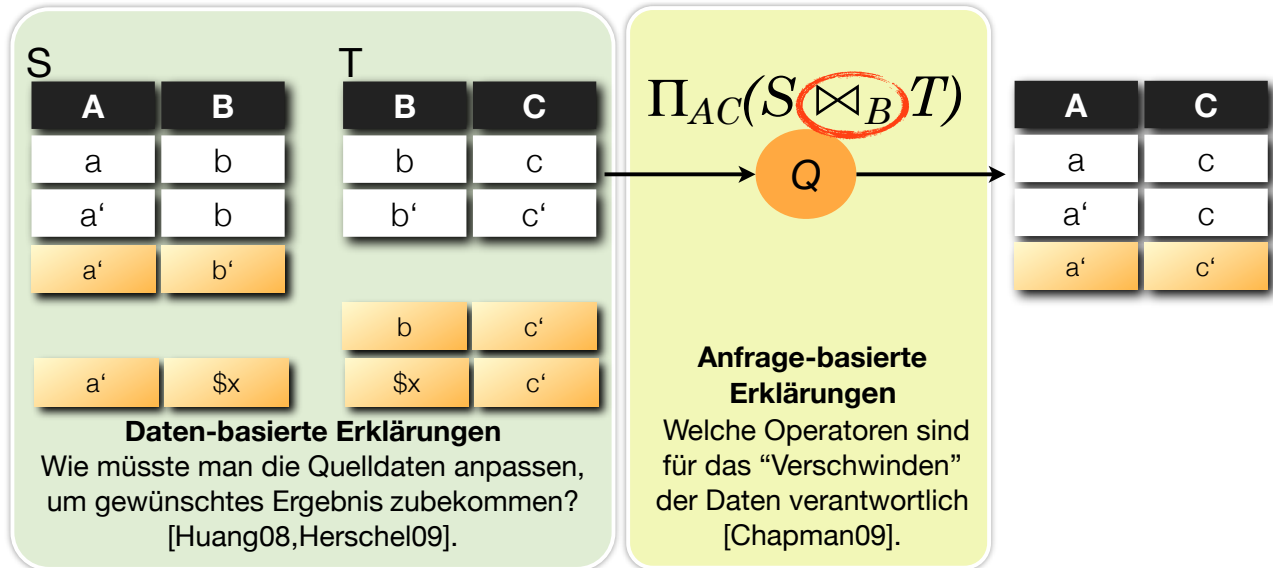
Erkläre, warum bestimmte Daten nicht im Ergebnis einer Anfrage Q sind.



Warum ist (a', c')
nicht im Output?

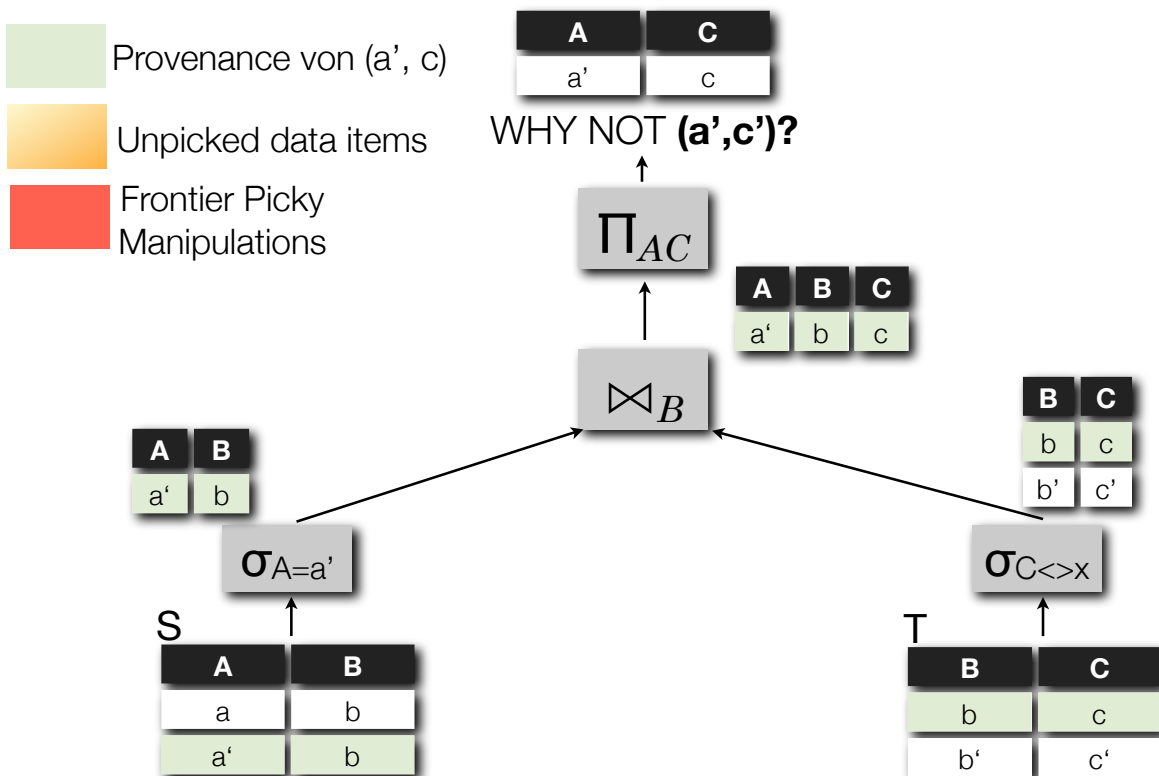
Erklären fehlender Daten

Erkläre, warum bestimmte Daten nicht im Ergebnis einer Anfrage Q sind.



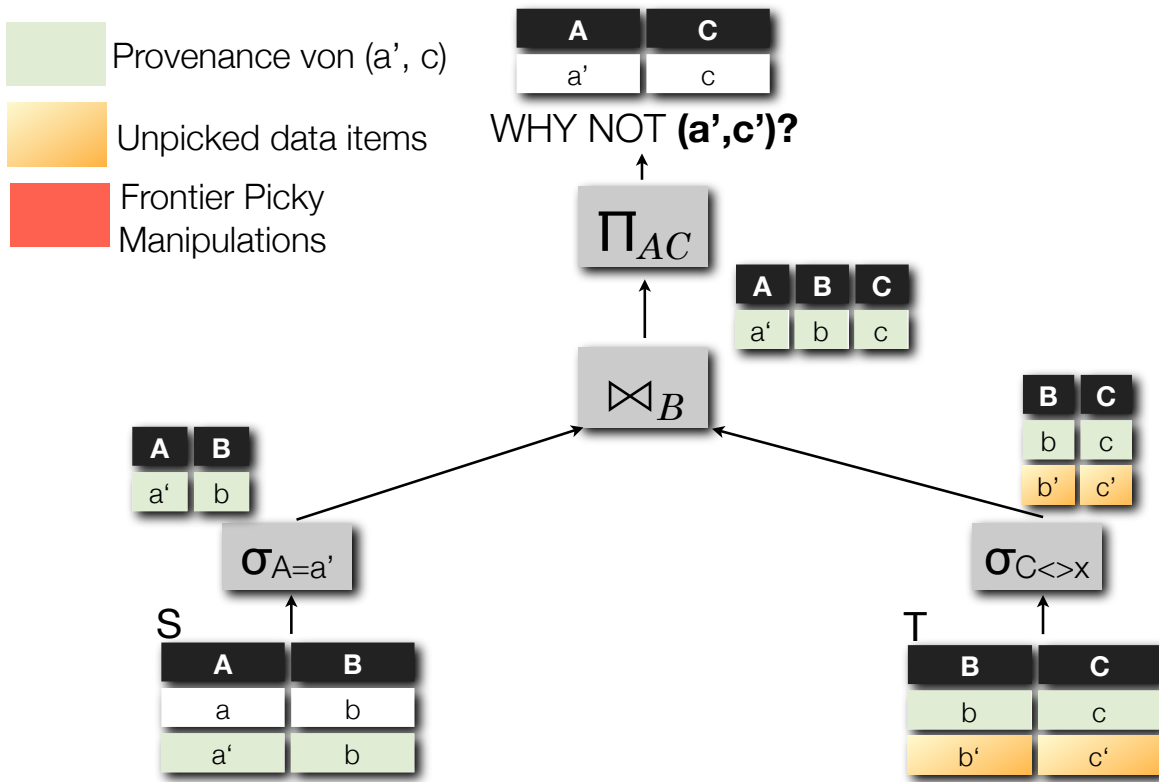
Erklären fehlender Daten

Anfrage-basierte Erklärungen [Chapman09]



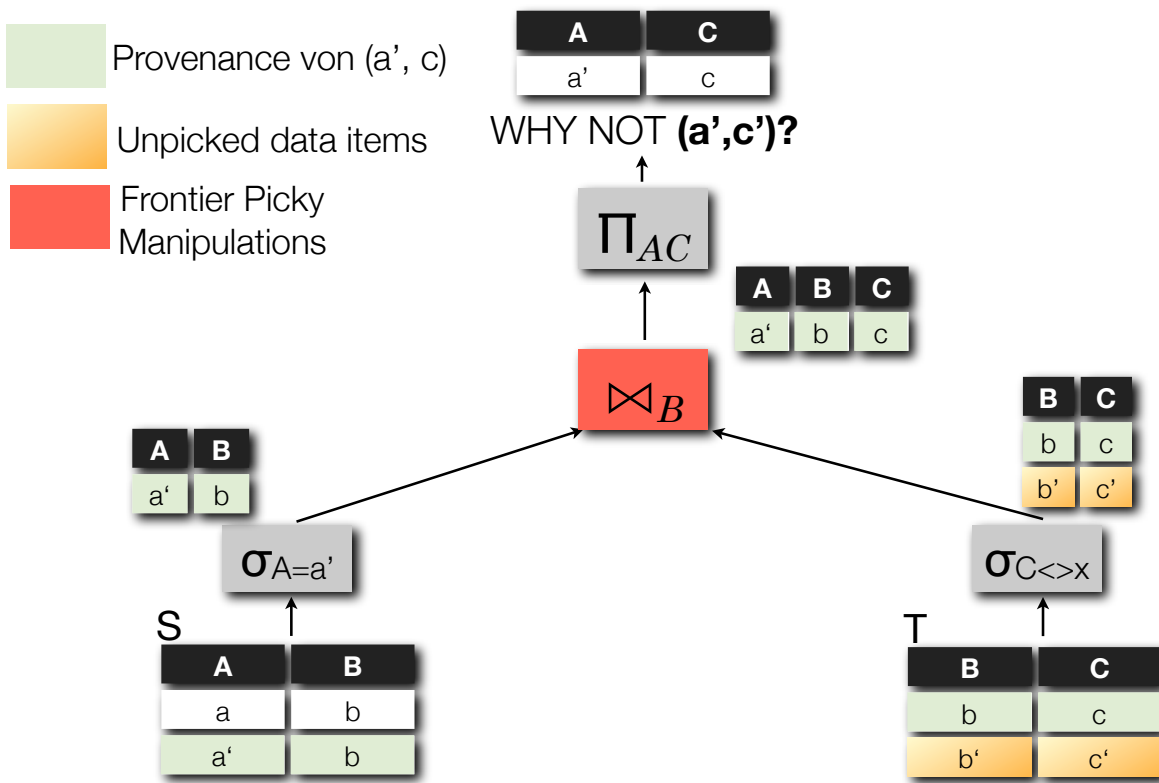
Erklären fehlender Daten

Anfrage-basierte Erklärungen [Chapman09]



Erklären fehlender Daten

Anfrage-basierte Erklärungen [Chapman09]



Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Eingabe

- Eine Select-Project-Join (SPJ) Anfrage Q
- Eine Quelldatenbank D
- Ein Tuple t , $t \notin Q(D)$
- Eine Menge von Tabellen bzw. Attributen, die unveränderbar sind (Änderungen dieser Tabellen können nicht in Erklärungen vorkommen).

Ausgabe

- Eine Menge von Erklärungen, wobei eine Erklärung eine Kombination existierender, neuer, und geänderter Tupel ist.
- ➔ Nötige insert und update Operationen, damit t in $Q(D)$ erscheint.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

$Q =$

SELECT	A, C
FROM	S, T
WHERE	S.B = T.B

SELECT A, C
FROM S, T
WHERE S.B = T.B

$D =$

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

$t =$

a'	c'
----	----

unveränderbare Tabellen = $\{S\}$

unveränderbare Attribute = $\{C\}$

Algorithmus: Formulierung der Erklärungsanfrage

1. Erweiterung von Q mit Prädikaten, die t beschreiben.
2. Erweiterung der **SELECT** Klausel um Attribute aller in der **FROM** Klausel referenzierter Tabellen.
3. Passe **WHERE** Klausel den veränderbaren Tabellen an.
4. Erweitere veränderbare Tabellen mit einem *proxy tuple*.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Q = **SELECT** A, C
FROM S, T
WHERE S.B = T.B

SELECT A, C
FROM S, T
WHERE S.B = T.B
AND S.A = a'
AND T.C = c'

D =

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

t =

a'	c'
----	----

unveränderbare Tabellen = {S}

unveränderbare Attribute = {C}

Algorithmus: Formulierung der Erklärungsanfrage

1. Erweiterung von Q mit Prädikaten, die t beschreiben.
2. Erweiterung der SELECT Klausel um Attribute aller in der FROM Klausel referenzierter Tabellen.
3. Passe WHERE Klausel den veränderbaren Tabellen an.
4. Erweitere veränderbare Tabellen mit einem proxy tuple.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Q = **SELECT** A, C
FROM S, T
WHERE S.B = T.B

SELECT ... AS S_A, ... AS S_B,
... AS T_B, ... AS T_C
FROM S, T
WHERE S.B = T.B
AND S.A = a'
AND T.C = c'

D =

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

t =

a'	c'
----	----

unveränderbare Tabellen = {S}

unveränderbare Attribute = {C}

Algorithmus: Formulierung der Erklärungsanfrage

1. Erweiterung von Q mit Prädikaten, die t beschreiben.
2. Erweiterung der SELECT Klausel um Attribute aller in der FROM Klausel referenzierter Tabellen.
3. Passe WHERE Klausel den veränderbaren Tabellen an.
4. Erweitere veränderbare Tabellen mit einem proxy tuple.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Q =

```
SELECT A, C
FROM S, T
WHERE S.B = T.B
```

D =

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

t =

a'	c'
----	----

unveränderbare Tabellen = {S}

unveränderbare Attribute = {C}

```
SELECT ... AS S_A, ... AS S_B,
... AS T_B, ... AS T_C
FROM S, T
WHERE S.A = a'
AND (T.C = c' OR T.C IS NULL)
```

Algorithmus: Formulierung der Erklärungsanfrage

1. Erweiterung von Q mit Prädikaten, die t beschreiben.
2. Erweiterung der SELECT Klausel um Attribute aller in der FROM Klausel referenzierter Tabellen.
3. Passe WHERE Klausel den veränderbaren Tabellen an.
4. Erweitere veränderbare Tabellen mit einem proxy tuple.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Q =

```
SELECT A, C
FROM S, T
WHERE S.B = T.B
```

D =

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

t =

a'	c'
----	----

unveränderbare Tabellen = {S}

unveränderbare Attribute = {C}

```
SELECT ... AS S_A, ... AS S_B,
... AS T_B, ... AS T_C
FROM S,
(SELECT * FROM T UNION
SELECT NULL, NULL FROM dual) T
WHERE S.A = a'
AND (T.C = c' OR T.C IS NULL)
```

Algorithmus: Formulierung der Erklärungsanfrage

1. Erweiterung von Q mit Prädikaten, die t beschreiben.
2. Erweiterung der SELECT Klausel um Attribute aller in der FROM Klausel referenzierter Tabellen.
3. Passe WHERE Klausel den veränderbaren Tabellen an.
4. Erweitere veränderbare Tabellen mit einem proxy tuple.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Q =

```
SELECT A, C
FROM S, T
WHERE S.B = T.B
```

D =

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

t =

a'	c'
----	----

unveränderbare Tabellen = {S}

unveränderbare Attribute = {C}

```
SELECT S.A AS S_A, S.B AS S_B,
T.B ||->|| S.B AS T_B,
T.C ||->|| c' AS T_C
FROM S,
(SELECT * FROM T UNION
SELECT NULL, NULL FROM dual) T
WHERE S.A = a'
AND (T.C = c' OR T.C IS NULL)
```

Algorithmus: Formulierung der Erklärungsanfrage

1. Erweiterung von Q mit Prädikaten, die t beschreiben.
2. Erweiterung der SELECT Klausel um Attribute aller in der FROM Klausel referenzierter Tabellen.
3. Passe WHERE Klausel den veränderbaren Tabellen an.
4. Erweitere veränderbare Tabellen mit einem proxy tuple.

Erklären fehlender Daten

Daten-basierte Erklärung [Huang08]

Q =

```
SELECT A, C
FROM S, T
WHERE S.B = T.B
```

D =

S		T	
A	B	B	C
a	b	b	c
a'	b	b'	c'

t =

a'	c'
----	----

unveränderbare Tabellen = {S}

unveränderbare Attribute = {C}

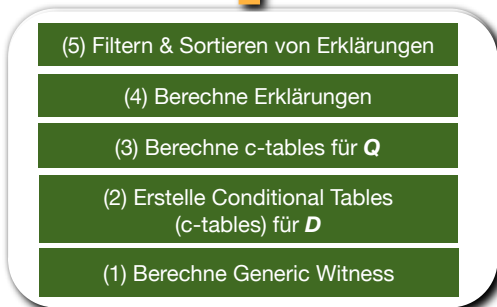
```
SELECT S.A AS S_A, S.B AS S_B,
T.B ||->|| S.B AS T_B,
T.C ||->|| c' AS T_C
FROM S,
(SELECT * FROM T UNION
SELECT NULL, NULL FROM dual) T
WHERE S.A = a'
AND (T.C = c' OR T.C IS NULL)
```

S_A	S_B	T_B	T_C
a'	b	b'->b	c'->c'
a'	b	null->b	null->c'

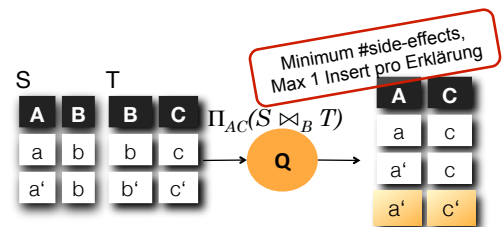
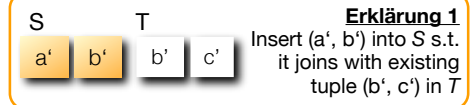
Der Artemis Algorithmus

Überblick

Menge Erklärungen X



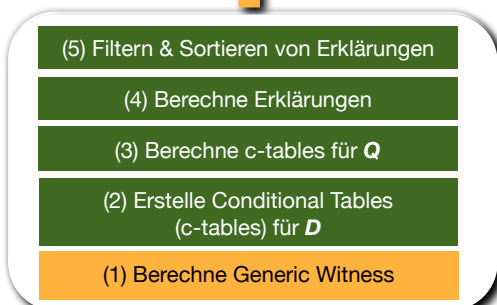
- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E



Der Artemis Algorithmus

Überblick

Menge Erklärungen X



- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

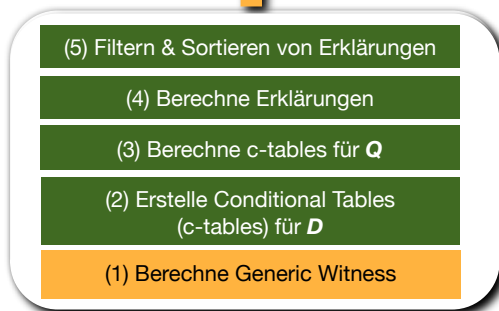
$Q: V(v_a, v_c) :- R(v_a, v_b), S(v_b, v_c)$

$E = \{(a', c')\}$

Der Artemis Algorithmus

Überblick

Menge Erklärungen X



- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

$$Q: V(a', c') :- R(a', v_b), S(v_b, c')$$

$$E = \{(a', c')\}$$

Generic Witness:
 $R(a', \$x), S(\$x, c')$

Der Artemis Algorithmus

Überblick

Menge Erklärungen X



- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

S^C

A	B	con
a	b	TRUE
a'	b	TRUE
a'	$\$x1$	$\$x1 \neq b$

T^C

B	C	con
b	c	TRUE
b'	c'	TRUE
$\$x2$	c'	$\$x2 \neq b'$

$$Q: V(a', c') :- R(a', v_b), S(v_b, c')$$

$$E = \{(a', c')\}$$

Generic Witness:
 $R(a', \$x), S(\$x, c')$

Der Artemis Algorithmus

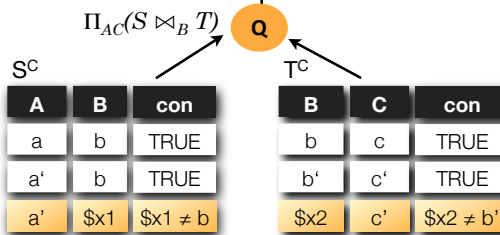
Überblick

Menge Erklärungen X

- (5) Filtern & Sortieren von Erklärungen
- (4) Berechne Erklärungen
- (3) Berechne c-tables für Q
- (2) Erstelle Conditional Tables (c-tables) für D
- (1) Berechne Generic Witness

- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

A	C	con
a	c	TRUE
a'	c	TRUE
a	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x1 = b' \wedge \$x1 \neq b$
a'	c'	$\$x1 = \$x2 \wedge \$x2 \neq b' \wedge \$x1 \neq b$



$$Q: V(a', c') :- R(a', v_b), S(v_b, c')$$

$$E = \{(a', c')\}$$

Generic Witness:
 $R(a', \$x), S(\$x, c')$

Der Artemis Algorithmus

Überblick

Menge Erklärungen X

- (5) Filtern & Sortieren von Erklärungen
- (4) Berechne Erklärungen
- (3) Berechne c-tables für Q
- (2) Erstelle Conditional Tables (c-tables) für D
- (1) Berechne Generic Witness

- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

A	C	con
a	c	TRUE
a'	c	TRUE
a	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x1 = b' \wedge \$x1 \neq b$
a'	c'	$\$x1 = \$x2 \wedge \$x2 \neq b' \wedge \$x1 \neq b$

Seiteneffekt →

Matches zu (a', c') →

Der Artemis Algorithmus

Überblick

Menge Erklärungen X

(5) Filtern & Sortieren von Erklärungen

(4) Berechne Erklärungen

(3) Berechne c-tables für Q

(2) Erstelle Conditional Tables (c-tables) für D

(1) Berechne Generic Witness

- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

Constraint Satisfaction Problem CSP1

tuple (a', c') existiert

AND

minimale Anzahl Seiteneffekte

Seiteneffekt

Matches zu

(a', c')

A	C	con
a	c	TRUE
a'	c	TRUE
a	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x1 = b' \wedge \$x1 \neq b$
a'	c'	$\$x1 = \$x2 \wedge \$x2 \neq b' \wedge \$x1 \neq b$

Der Artemis Algorithmus

Überblick

Menge Erklärungen X

(5) Filtern & Sortieren von Erklärungen

(4) Berechne Erklärungen

(3) Berechne c-tables für Q

(2) Erstelle Conditional Tables (c-tables) für D

(1) Berechne Generic Witness

- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

Constraint Satisfaction Problem CSP1

$\$x2 = b \wedge \$x2 \neq b'$

AND

minimale Anzahl Seiteneffekte

$\$x2 = b \wedge \$x2 \neq b'$

$\$x1 = \$x2 \wedge \$x2 \neq b' \wedge \$x1 \neq b$

$\$x1 = b' \wedge \$x1 \neq b$

Seiteneffekt

Matches zu

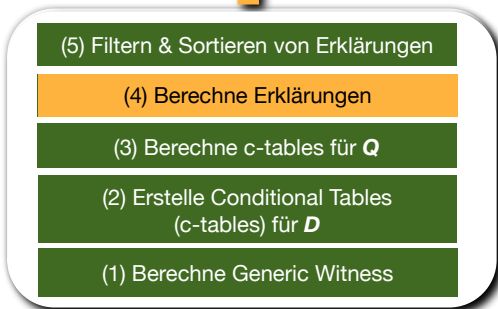
(a', c')

A	C	con
a	c	TRUE
a'	c	TRUE
a	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x1 = b' \wedge \$x1 \neq b$
a'	c'	$\$x1 = \$x2 \wedge \$x2 \neq b' \wedge \$x1 \neq b$

Der Artemis Algorithmus

Überblick

Menge Erklärungen X



- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E

Output constraint solver für CSP1
 $\$x2 = b$, 1 side-effect

Weitere Ergebnisse
 CSP2: $\$x1 = b'$, 0 side-effects
 CSP3: $\$x1 = \$x2$, $\$x1 \neq b$, $\$x2 \neq b'$,
 0 side-effects

A	C	con
a	c	TRUE
a'	c	TRUE
a	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x2 = b \wedge \$x2 \neq b'$
a'	c'	$\$x1 = b' \wedge \$x1 \neq b$
a'	c'	$\$x1 = \$x2 \wedge \$x2 \neq b' \wedge \$x1 \neq b$

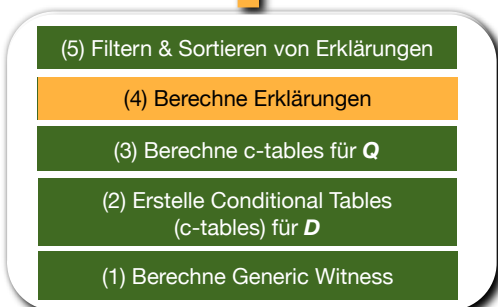
Seiteneffekt → (points to the row with a, c')

Matches zu (a', c') → (points to the rows with a', c')

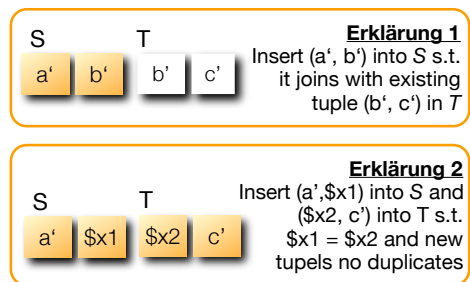
Der Artemis Algorithmus

Überblick

Menge Erklärungen X



- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E



Output constraint solver für CSP1
 $\$x2 = b$, 1 side-effect

Weitere Ergebnisse
 CSP2: $\$x1 = b'$, 0 side-effects
 CSP3: $\$x1 = \$x2$, $\$x1 \neq b$, $\$x2 \neq b'$,
 0 side-effects

Der Artemis Algorithmus

Überblick

Menge Erklärungen X

(5) Filtern & Sortieren von Erklärungen

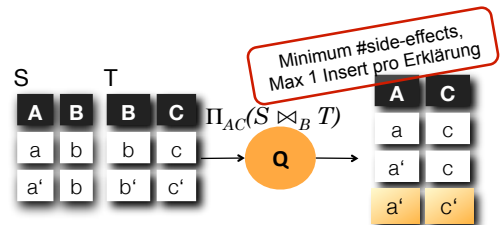
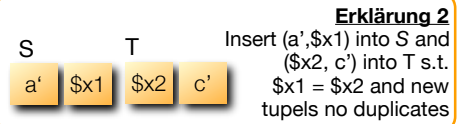
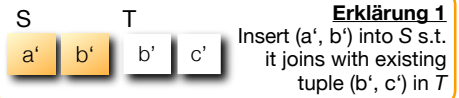
(4) Berechne Erklärungen

(3) Berechne c-tables für Q

(2) Erstelle Conditional Tables (c-tables) für D

(1) Berechne Generic Witness

- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E



Der Artemis Algorithmus

Überblick

Menge Erklärungen X

(5) Filtern & Sortieren von Erklärungen

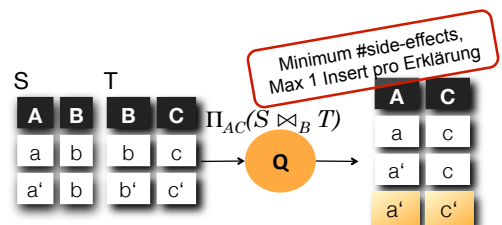
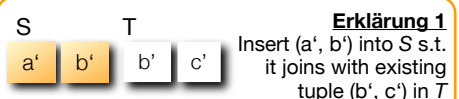
(4) Berechne Erklärungen

(3) Berechne c-tables für Q

(2) Erstelle Conditional Tables (c-tables) für D

(1) Berechne Generic Witness

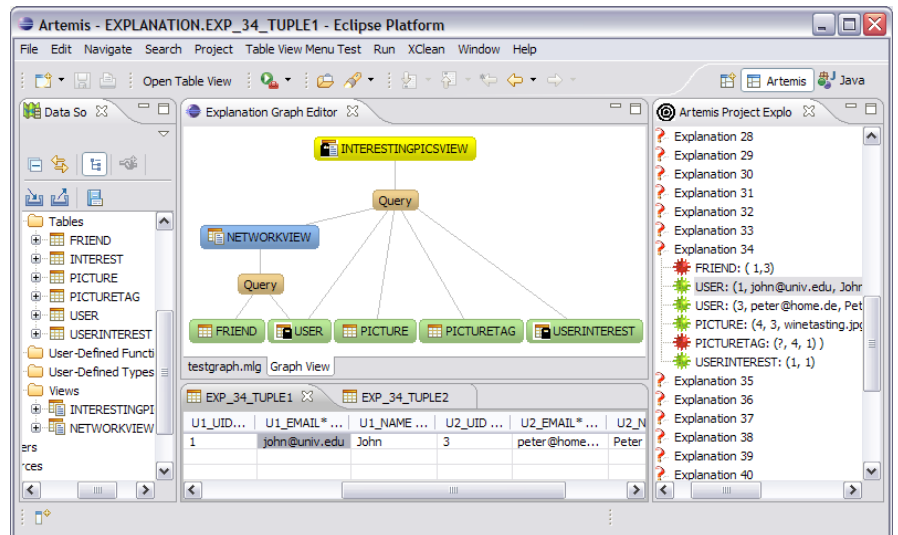
- 1) Quelldatenbank D
- 2) Menge an SPJU Anfragen Q
- 3) Menge der fehlenden Tupel E



Der Artemis Algorithmus

Implementierung [Herschel09]

- Implementierung als Eclipse Plugin.
- Data Tools Platform (DTP) Erweiterung für DB Anbindung.
- Eigene Perspective, Views, Launcher, ...
- Minion als Constraint Solver.



Zusammenfassung

- Why-Not Algorithmus (anfragebasierte Erklärungen)
- Missing-Answers Algorithmus (instanzbasierte Erklärungen durch Anfrageumschreibung)
- Artemis-Algorithmus (instanzbasierte Erklärungen durch *conditional tables* und *constraint solver*)

