



Datenintegration & Datenherkunft

Varianten der Data Provenance

Wintersemester 2010/11

Melanie Herschel

melanie.herschel@uni-tuebingen.de

Lehrstuhl für Datenbanksysteme, Universität Tübingen

1

Aktueller Stand

Modellierung

- Global-As-View Modellierung
- Local-As-View Modellierung

Anfragen

- Global-As-View Anfragebearbeitung
- Containment und Local-As-View Anfragebearbeitung
- Bucket Algorithmus

Datenintegration

- Duplikaterkennung
- Datenfusion

Datenherkunft (Data Provenance)

- Why-, How-, und Where-Provenance
- Datenherkunft fehlender Daten

2

Provenance – an old problem

W.V. Quine. "Words enough..." *New York Review of Books* XIII(10):3-4, 1969



A well-known encyclopedia, following tradition, incorrectly describes Monaco as having an area of "8 square miles."

A new edition adds "...the length being 2 1/4 miles and the width varying from 165 to 1100 yards."

An editor, spotting the inconsistency, removes the *correct* information from the subsequent edition.

Quelle: Peter Bunemann

3

The area of Monaco today

Most sources	1.95 sq km
www.atlapedia.com	1.94 sq km
military.countrywatch.com	2 sq km

Quelle: Peter Bunemann

4

The population of Monaco today?

• 2004 35,000		http://www.cybevasion.fr/tourisme/monaco.html
• 2004 33,300		http://www.internetworldstats.com/europa2.htm
• 2004 32,270	(July 2004 est.)	http://www.cia.gov/cia/publications/factbook/geos/mn.html
• 2004 32,000		http://www.studentsoftheworld.info/pageinfo_pays.php3?Pays=MCO
• 2004 29,972		http://worldatlas.com/webimage/countrys/europe/mc.htm
• 2003 32,130	(July 2003 est.)	http://www.greenfacts.org/studies/climate_change/index.htm
• 2003 32,130	(mid 2003)	http://www.infoplease.com/ipa/A0004379.html
• 2003 32,000	(July 2003 estimate)	http://www.gesource.ac.uk/worldguide/html/962_people.html
• 2003 30,000		http://www.tifq.ulaval.ca/axl/europe/monaco.htm
• 2002 31,987	(July 2002 est.)	http://www.greekorthodoxchurch.org/wfb2002/monaco/monaco_people.html
• 2001 31,842	(July 2001 est.)	http://wonderclub.com/Atlas/mccia.htm
• 2001 31,842	(July 2001 est.)	http://www.worldfactsandfigures.com/countries/monaco.php
• 2001 31,842	(July 2001 est.)	http://www.workmall.com/wfb2001/monaco/monaco_people.html
• 2000 32,500	(est 2000)	http://www.atlapedia.com/online/countries/monaco.htm
• 2000 32,020	(C 2000-05-03)	http://www.citypopulation.de/Monaco.html
• 2000 32,020	(2000).	http://www.worldtravelguide.net/data/mco/mco.asp
• 2000 32,020	(2000 census[!])	http://www.state.gov/r/pa/ei/bgn/3397.htm
• 2000 31,693	(July 2000 est.)	http://geography.about.com/library/cia/blcmonaco.htm
• 2000 31,693	(July 2000 est.)	http://www.abacci.com/atlas/demography.asp?countryID=269
• 2000 31,693	(July 2000 est.)	http://www.mapquest.com/atlas/?region=monaco
• 2000 31,842		http://www.fact-index.com/m/mo/monaco.html
• 2000 31,842		http://en.wikipedia.org/wiki/Monaco
• 2000 31,700	(e2000m)	http://www.library.uu.nl/wesp/populstat/Europe/monacoc.htm
• 1999 32,149	(July 1999 est.)	http://www.photius.com/wfb1999/monaco/monaco_people.html
• 1999 32,000		http://geography.about.com/library/weekly/aa012599.htm
• 1990 29,972	(1990 census)	http://www.monte-carlo.mc/us/presentation/keyfigur/

Quelle: Peter Bunemann

5

The area and population of Monaco today

- The 2004 figures vary by 17% !!
- Most sites copy from the CIA world fact-book
 - ▶ The US state department does not – and contradicts itself!
- Only two sites give attribution.
- No evidence is given for how the estimates were made.
- The last census appears to have been taken in 1990!

Quelle: Peter Bunemann

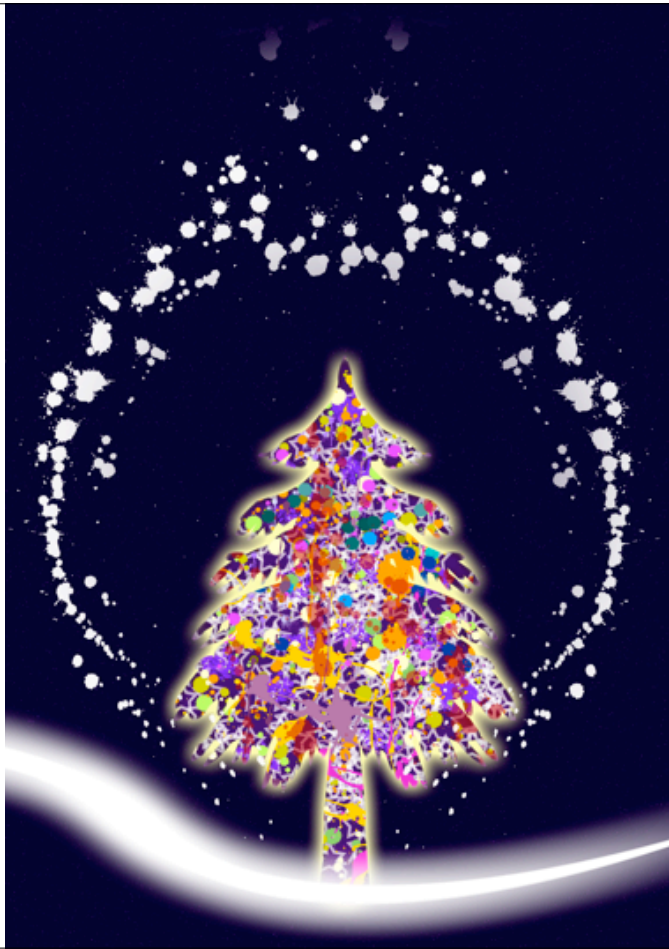
6

Kapitel 10

Datenherkunft

- ➔ Motivation und Klassifikation
- Herkunft existierender Daten
 - Why-Provenance
 - How-Provenance
 - Where-Provenance
- Herkunft fehlender Daten
 - Instanzbasierte Provenance
 - Anfragebasierte Provenance
- Eager vs. Lazy Provenance Berechnung

7



Data Provenance – Motivation

- Data Provenance
 - ▶ Data Provenance ist das Problem, zu Objekten im integrierten System diejenigen Objekte in den Quellen zu bestimmen, aus denen das integrierte Objekt abgeleitet wurde.
 - ▶ Auch: Data Lineage
 - ▶ Auch: Data Pedigree
- Data Warehouses
 - ▶ Datenanalyse
 - ▶ Decision Support
 - ▶ Data Mining
 - ▶ Aggregation

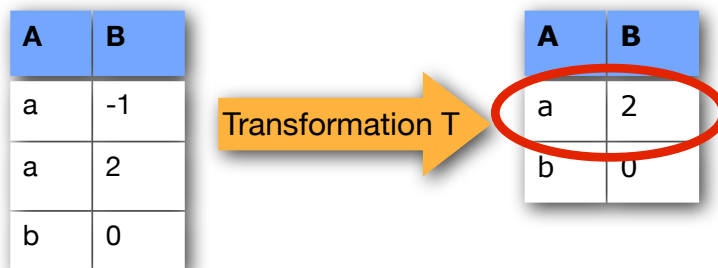


Hilfe durch Data Provenance

Data Provenance – Motivation

- Schwierigkeit des Data Provenance hängt von Transformationen ab
 - ▶ SQL: Leichter aber unrealistisch
 - Data Provenance durch SQL Sichten
 - Data Provenance durch Operatoren der relationalen Algebra
 - ▶ Allgemeine Transformationen: Schwierig aber wichtig
 - Data Provenance durch komplexe, nutzerdefinierte Transformationen
 - Data Provenance durch ETL Prozesse
 - Data Provenance durch Ketten von 60+ Transformationen
- Data Provenance geschieht auf Datenebene.

Data Provenance – Motivation



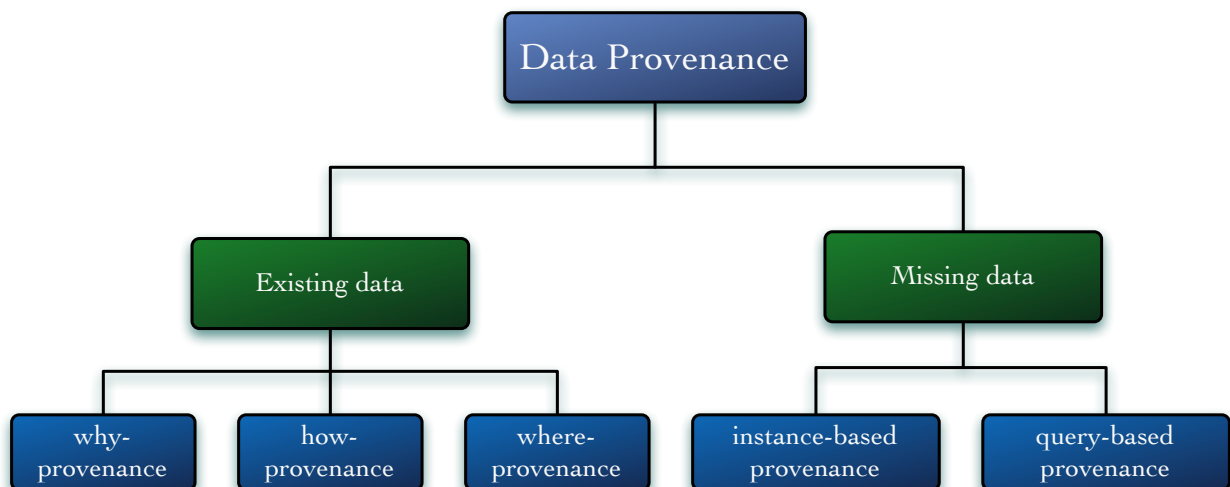
Herkunft (*prov* für *provenance*) des Tupels (a, 2)?

- $T = \sigma_{B \geq 0}$
 - $\Rightarrow \text{prov}(a,2) = \{(a,2)\}$
- $T =$ Gruppierung nach A und Aggregation: 2x SUM(B)
 - $\Rightarrow \text{prov}(a,2) = \{(a,-1); (a,2)\}$
- $T =$ Gruppierung nach A und Aggregation: MAX(B)
 - $\Rightarrow \text{prov}(a,2) = \{(a,2)\}$
- ...

Data Provenance – Motivation

- Zusätzliche Schwierigkeiten
 - ▶ Runtime overhead
 - ETL
 - Bei virtueller Integration
 - ▶ Speicherbedarf
 - Metadaten
 - ▶ Transformationen
 - Einzel
 - In Ketten
 - In (azyklischen) Graphen
- Trade-off zwischen Nutzen und Kosten

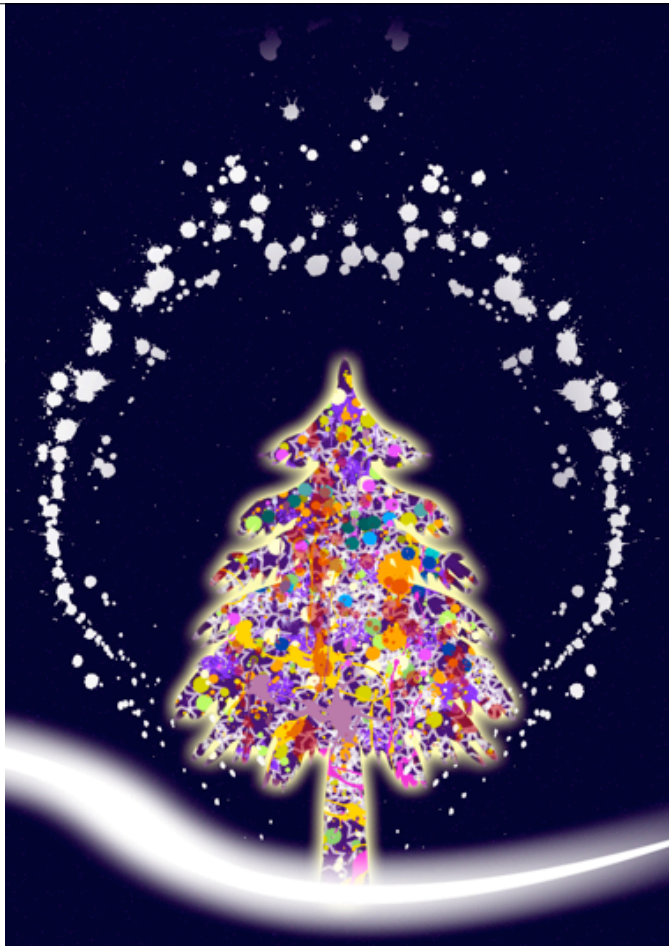
Arten des Data Provenance



Kapitel 10

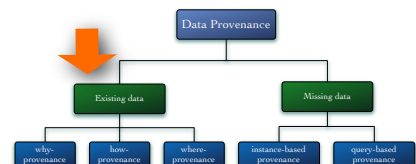
Datenherkunft

- Motivation und Klassifikation
- Herkunft existierender Daten
 - Why-Provenance
 - How-Provenance
 - Where-Provenance
- Herkunft fehlender Daten
 - Instanzbasierte Provenance
 - Anfragebasierte Provenance
- Eager vs. Lazy Provenance Berechnung



13

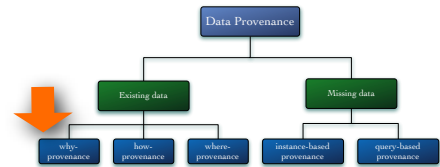
Datenherkunft existierender Daten



- Gegeben:
 - ▶ Eine Menge von Datenquellen D
 - ▶ Eine Datentransformation T
 - ▶ Das Ergebnis $T(D)$ der Ausführung von T über die Instanz von D .
- Datenherkunft existierender Daten (nach [CCT09])
 - ▶ Entspricht der Herkunft eines Tupels $t \in T(D)$
 - ▶ Why-provenance: Aus welchen Quelldaten in D kombiniert sich t ?
 - ▶ How-provenance: Wie werden die Tupel aus D kombiniert, um t zu produzieren?
 - ▶ Where-provenance: Aus welchen Quellen in D wurden Werte in t kopiert.

Why-Provenance

Beispiel



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Transformation T

```
SELECT DISTINCT
    a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

Touren

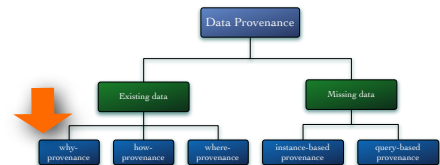
	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

Ergebnis von T:

Name	Telefon
BayTours	415-1200
HarborCruz	831-3000

Why-Provenance

Beispiel



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Transformation T

```
SELECT DISTINCT
    a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

Touren

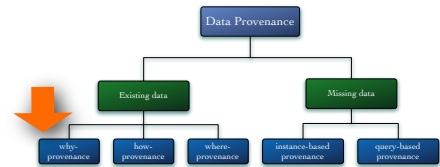
	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

Ergebnis von T:

Name	Telefon
BayTours	415-1200
HarborCruz	831-3000

Herkunft?

Why-Provenance nach Cui & Widom [CW03]

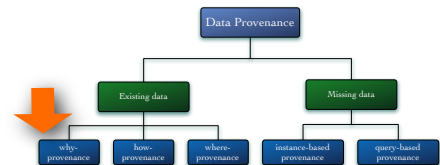


Witness von t

Ein **witness** für ein Tupel $t \in T(D)$ ist eine Teilmenge der Tupel in D , die ausreicht, um t zu produzieren.

- Im Beispiel ist die Why-Provenance des Tupels $t = (\text{HarborCruz}, 831-3000)$ die Tupelmengemenge $\{t_2, t_7\}$.
- Bedeutung
 - Die Tupel in dieser Menge tragen zur Bildung von t bei.
 - t_2 und t_7 sind sogenannte Zeugen (*witnesses*) für t .
 - Kein weiteres Tupel in D trägt zu t bei.
- Problem: in einigen Fällen ist diese Definition zu ungenau, denn die Why-Provenance beschreibt Tupel, die zum Ergebnis beitragen können, aber nicht zwangsläufig müssen.

Why-Provenance nach Cui & Widom [CW03]



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Transformation T

```
SELECT DISTINCT
    a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

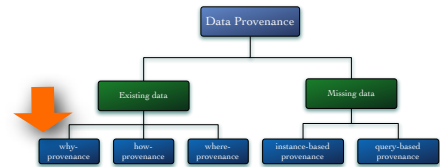
Touren

	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

Ergebnis von T:

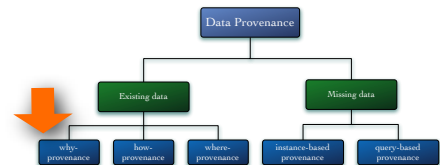
	Name	Telefon
	BayTours	415-1200
Herkunft?	HarborCruz	831-3000

Why-Provenance nach Cui & Widom [CW03]



- Im Beispiel ist die Why-Provenance des Tupels $t =$ (BayTours, 415-1200) die Tupelmengemenge $\{t_1, t_5, t_6\}$.
- Dabei sind entweder $\{t_1, t_5\}$ oder $\{t_1, t_6\}$ ausreichend, um t zu erzeugen.
- Diese Unterscheidung ist nicht Teil der Why-Provenance von Cui & Widom, jedoch von Buneman et. al.

Why-Provenance nach Buneman et. al. [BKT01]



Witness basis von $t =$ why-provenance

Eine **witness basis** für ein Tupel $t \in T(D)$ ist eine Menge von witnesses von t , die der Anfragesemantik entsprechen, indem sie jeweils eine Tupelkombination beschreiben.

T in Datalog:

$$Q(n, t) :- A(n, o, t), T(n, z, 'boot', p)$$


Allgemeine Transformation, die zu $t =$ (BayTours, 415-1200) führt:

$$Q('BayTours', '415-1200') :- \\ A('BayTours', o, '415-1200'), T('BayTours', z, 'boot', p)$$

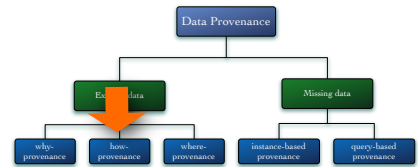

Zwei Übereinstimmungen
zu diesem Muster:

- 1) $\{t_1, t_5\}$
- 2) $\{t_1, t_6\}$



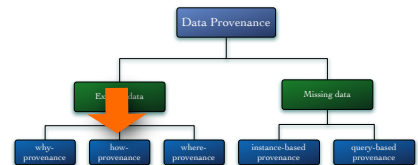
Witness basis von $t =$
 $\{ \{t_1, t_5\}, \{t_1, t_6\} \}$

How-Provenance



- Why-Provenance = Welche Tupel tragen zur Bildung eines Ergebnistupels bei.
- Unklar ist, wie das Ergebnistupel aus den Eingabedaten in der Why-Provenance kombiniert wird.
- How-Provenance liefert eine Beschreibungsart (*provenance semi-rings* [GKT07]), die zeigt, wie Tupel in der Why-Provenance von einer Transformation kombiniert werden.

How-Provenance



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Why- vs. How-Provenance

why-provenance: {{t1}, {t1, t3}}
 ABER: in erstem witness ist unklar, dass t1 zweimal durch die Transformation verwendet wird.

Transformation T

```
SELECT e.ziel, a.Telefon
FROM   Agenturen a,
       (SELECT name, ort AS ziel
        FROM   Agenturen a
        UNION
        SELECT name, zielort AS ziel
        ) e
WHERE  a.name = e.name
```

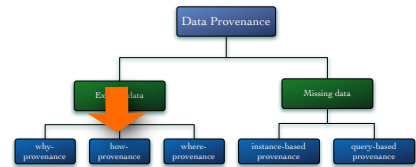
Touren

	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

Ergebnis von T

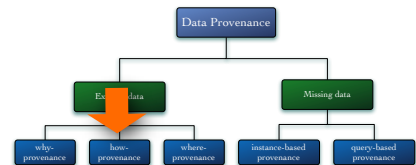
Ziel	Telefon
San Francisco	415-1200
Santa Cruz	831-3000
Santa Cruz	415-1200
Monterey	415-1200
Monterey	831-3000
Carmel	831-3000

How-Provenance



- Polynom beschreibt How-Provenance
- Multiplikation (\times) = Kombination durch Join
- Addition ($+$) = Beschreibung einer alternativen *witness basis*.
- Z.B. How-Provenance von (San Francisco, 415-1200) ist $t1 \times (t1 + t3) = t1 \times t1 + t1 \times t3$, was bedeutet, dass $t1$ entweder zweimal verwendet wird (witness 1) oder $t1$ mit $t3$ kombiniert wird (witness 2).
- Aus der How-Provenance lässt sich die Why-Provenance stets ableiten, die Gegenrichtung gilt jedoch nicht.

How-Provenance



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Transformation T

```

SELECT e.Ziel, a.Telefon
FROM   Agenturen a,
       (SELECT name, ort AS ziel
        FROM   Agenturen a
        UNION
        SELECT name, zielort AS ziel
        ) e
WHERE  a.name = e.name
    
```

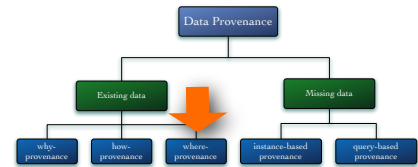
Touren

	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

Ergebnis von T

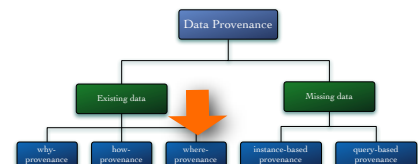
Ziel	Telefon	how-provenance
San Francisco	415-1200	$t1 \times (t1 + t3)$
Santa Cruz	831-3000	$t2 \times t2$
Santa Cruz	415-1200	$t1 \times (t4 + t5)$
Monterey	415-1200	$t1 \times t6$
Monterey	831-3000	$t1 \times t7$
Carmel	831-3000	$t1 \times t8$

Where-Provenance



- Beschreibt, aus welcher Quelle Daten kopiert wurden.
- Im Gegensatz zu Why-Provenance, die den Zusammenhang zwischen Quell- und Zieltupeln beschreibt, beschreibt Where-Provenance die Beziehung zwischen Quell- und Zielorten.
- Im relationalen Modell ist der Ort z.B. die Zelle einer Tabelle.
- Die Where-Provenance von Daten am Ort O in $T(D)$ besteht aus Orten in D .
- Bezieht man sich auf ein Attribut eines bestimmten Tupels in $T(D)$, so befindet sich die Where-Provenance innerhalb der *witness basis* dieses Tupels, wenn der Wert nicht von T selbst berechnet wird.

Where-Provenance



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Warum nicht t7.name?

Touren

	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

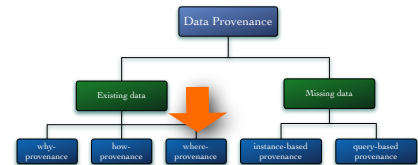
Transformation T

```
SELECT DISTINCT
  a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

Ergebnis von T:

Name	Telefon
BayTours	415-1200
HarborCruz	831-3000

Where-Provenance



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000

Why- vs. Where-Provenance

Why-Provenance: { {t1,t5}, {t1, t6} }
 Where-Provenance: {t5.Name, t6.Name}

Transformation T

```
SELECT DISTINCT
  t.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Type = 'boot'
```

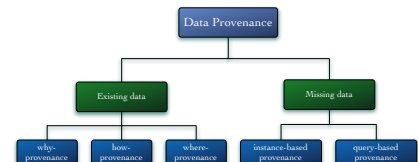
Touren

	Name	Zielort	Typ	Preis
t3	BayTours	San Francisco	cable car	\$50
t4	BayTours	Santa Cruz	bus	\$100
t5	BayTours	Santa Cruz	boot	\$250
t6	BayTours	Monterey	boot	\$400
t7	HarborCruz	Monterey	boot	\$200
t8	HarborCruz	Carmel	zug	\$90

Ergebnis von T:

Name	Telefon
BayTours	415-1200
HarborCruz	831-3000

Zusammenfassung Herkunft Existierender Daten

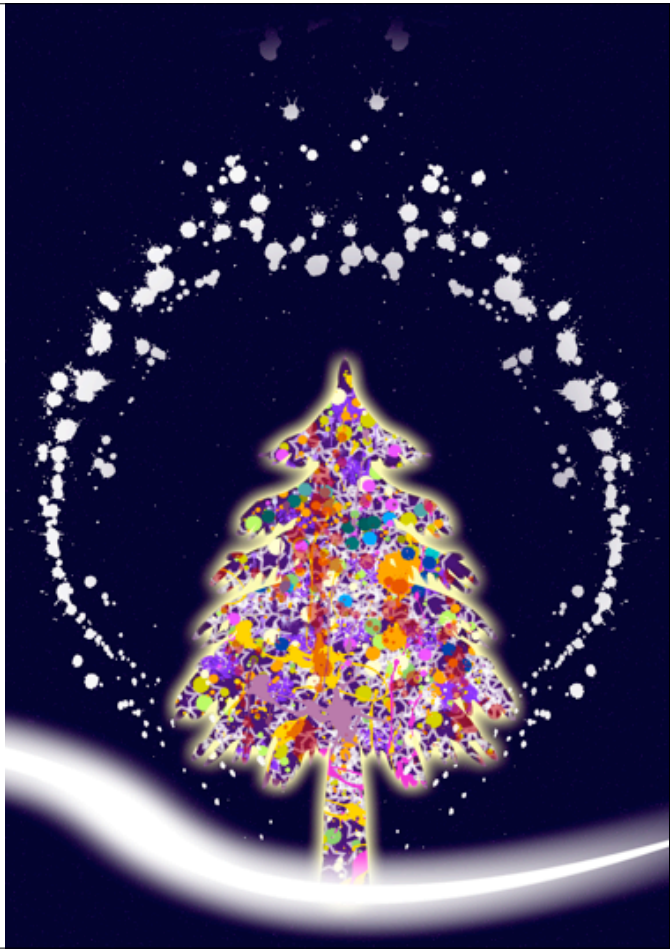


- Gegeben:
 - ▶ Eine Menge von Datenquellen D
 - ▶ Eine Datentransformation T
 - ▶ Das Ergebnis T(D) der Ausführung von T über die Instanz von D.
- Datenherkunft existierender Daten
 - ▶ Entspricht der Herkunft eines Tupels $t \in T(D)$
 - ▶ Why-provenance identifiziert welche Quelltuplel in D zur Bildung von t verwendet werden (*witness basis*).
 - ▶ How-provenance beschreibt in Form eines Polynoms, wie welche Tuplel kombiniert werden, um t zu produzieren?
 - ▶ Where-provenance beschreibt Beziehungen zwischen Orten, an denen Daten residieren. Bei relationalen Daten wird z.B. die Herkunftszelle eines bestimmten Attributwerts von t bestimmt.

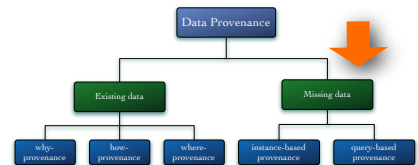
Kapitel 10

Datenherkunft

- Motivation und Klassifikation
- Herkunft existierender Daten
 - Why-Provenance
 - How-Provenance
 - Where-Provenance
- Herkunft fehlender Daten
 - Instanzbasierte Provenance
 - Anfragebasierte Provenance
- Eager vs. Lazy Provenance Berechnung



Erklären fehlender Daten



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

S

A	B
a	b
a'	b
a'	b'
a'	\$x

T

B	C
b	c
b'	c'
b	c'
\$x	c'

Instanzbasierte Erklärungen
Wie müsste man die Quelldaten anpassen, um gewünschtes Ergebnis zu bekommen?
[Huang08, Herschel09].

$\Pi_{AC}(S \bowtie B T)$

Q

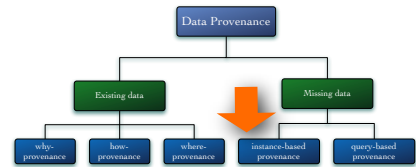
A	C
a	c
a'	c
a'	c'

Anfragebasierte Erklärungen
Welche Operationen sind für das "Verschwinden" von Daten verantwortlich?
[Chapman09].

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

Instanzbasierte Erklärungen

Beispiel



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000
t3	CoastTours	Monterey	123-4567

Transformation T

```
SELECT DISTINCT
  a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

Touren

	Name	Zielort	Typ	Preis
t4	BayTours	San Francisco	cable car	\$50
t5	BayTours	Santa Cruz	bus	\$100
t6	BayTours	Santa Cruz	boot	\$250
t7	BayTours	Monterey	boot	\$400
t8	HarborCruz	Monterey	boot	\$200
t9	HarborCruz	Carmel	zug	\$90
t10	CoastTours	Monterey	bus	\$80

Ergebnis von T:

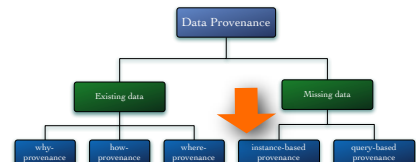
Name	Telefon
BayTours	415-1200
HarborCruz	831-3000

Warum fehlt folgendes Tupel
Gründe in den Quelldaten .suchen

CoastTours	123-4567
------------	----------

Instanzbasierte Erklärungen

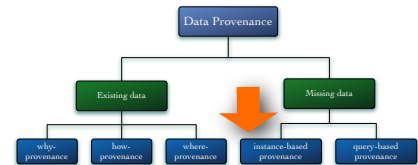
Beispiel



Notizen

Instanzbasierte Erklärungen

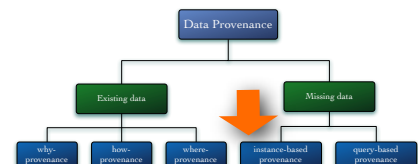
Kommentare



- Welche Änderungen der Datenquellen D sind erlaubt?
 - ▶ Einfügen von Tupeln?
 - ▶ Ändern von Attributwerten?
 - ▶ Entfernen von Tupeln?
 - ▶ Oben genannte Operationen können auch eingeschränkt angewendet werden.
 - Z.B. Updates sind nicht auf Agenturen.Name und Touren.Name zulässig.
- Welche Änderungen der Datenquellen sind sinnvoll?
 - ▶ Im Allgemeinen sollten nur die Änderungen in einer Erklärung auftauchen, die wirklich nötig sind.
 - ▶ Also z.B. kein Einfügen eines existierenden Tupels oder keine Änderung von Attributen, die nicht zur Bildung des gewünschten Tupels beitragen.

Instanzbasierte Erklärungen

Beispiel



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000
t3	CoastTours	Monterey	123-4567

Transformation T

```
SELECT DISTINCT
  a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

Touren

	Name	Zielort	Typ	Preis
t4	BayTours	SF	cable car	\$50
t5	BayTours	Santa Cruz	bus	\$100
t6	BayTours	Santa Cruz	boot	\$250
t7	BayTours	Monterey	boot	\$400
t8	HarborCruz	Monterey	boot	\$200
t9	HarborCruz	Carmel	zug	\$90
t10'	CoastTours	Monterey	bus --> boot	\$80
	CoastTours	?	boot	?

Update

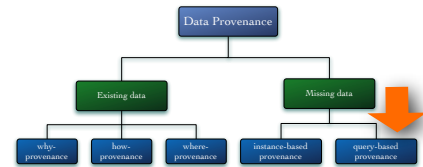
Insert

Ergebnis von T:

Name	Telefon
BayTours	415-1200
HarborCruz	831-3000
CoastTours	123-4567

Anfragebasierte Erklärungen

Beispiel



Agenturen

	Name	Ort	Telefon
t1	BayTours	San Francisco	415-1200
t2	HarborCruz	Santa Cruz	831-3000
t3	CoastTours	Monterey	123-4567

Transformation T

```
SELECT DISTINCT
  a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Typ = 'boot'
```

Touren

	Name	Zielort	Typ	Preis
t4	BayTours	San Francisco	cable car	\$50
t5	BayTours	Santa Cruz	bus	\$100
t6	BayTours	Santa Cruz	boot	\$250
t7	BayTours	Monterey	boot	\$400
t8	HarborCruz	Monterey	boot	\$200
t9	HarborCruz	Carmel	zug	\$90
t10	CoastTours	Monterey	bus	\$80

Ergebnis von T:

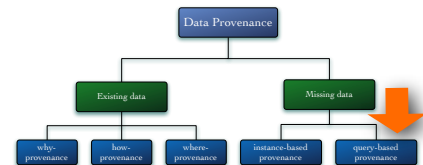
Name	Telefon
BayTours	415-1200
HarborCruz	831-3000

Warum fehlt folgendes Tupel
Gründe in der Anfrage suchen.

CoastTours	123-4567
------------	----------

Anfragebasierte Erklärungen

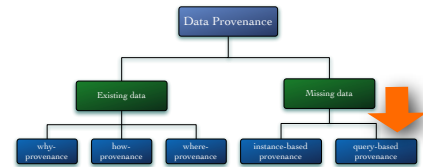
Beispiel



Notizen

Anfragebasierte Erklärungen

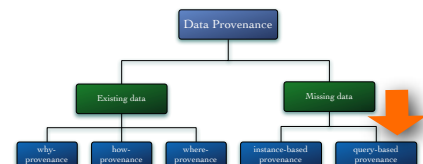
Kommentare



- Betrachtung der Anfrage als Operatorbaum.
- Sind überhaupt Quelldaten vorhanden, deren Kombination zum fehlenden Tupel führen könnte?
- An welchen Operatoren gehen diese Tupel verloren? D.h., die Tupel
 - tauchen in der Eingabe auf
 - aber kein Tupel der Ausgabe hat eine Datenherkunft, die diesen Tupeln entspricht.
- Es sind i.A. mehrere anfragebasierte Erklärungen möglich.

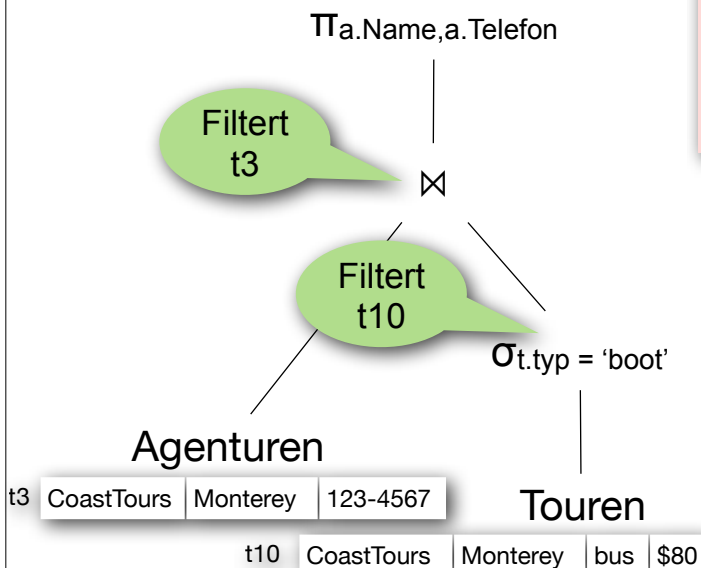
Anfragebasierte Erklärungen

Beispiel



Transformation T

```
SELECT DISTINCT
    a.Name, a.Telefon
FROM Agenturen a, Touren t
WHERE a.name = t.name
AND t.Type = 'boot'
```



Ergebnis von T:

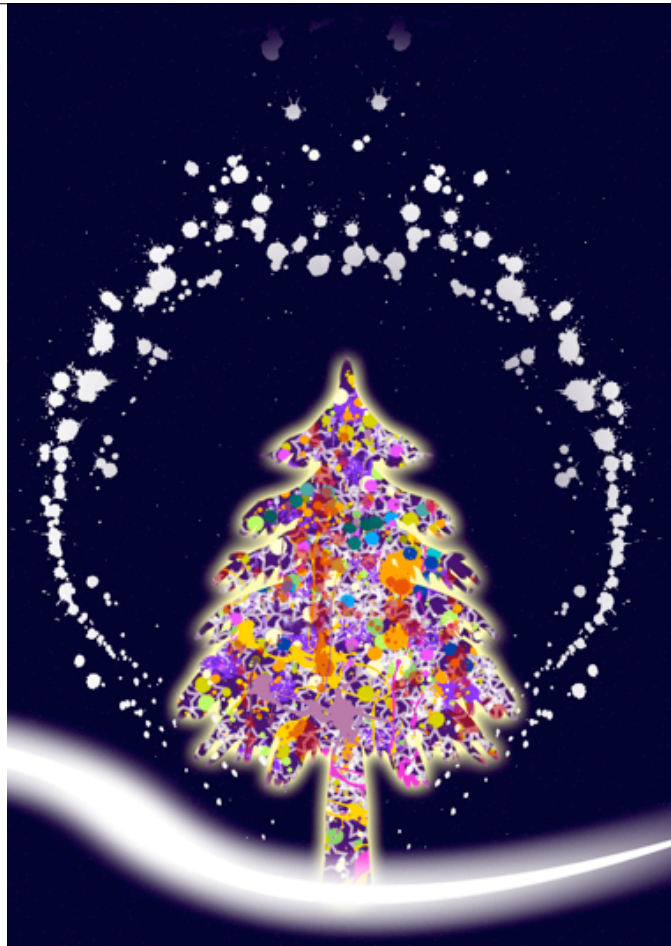
Name	Telefon
BayTours	415-1200
HarborCruz	831-3000
CoastTours	123-4567

Kapitel 10

Datenherkunft

- Motivation und Klassifikation
- Herkunft existierender Daten
 - Why-Provenance
 - How-Provenance
 - Where-Provenance
- Herkunft fehlender Daten
 - Instanzbasierte Provenance
 - Anfragebasierte Provenance

➔ Eager vs. Lazy Provenance Berechnung

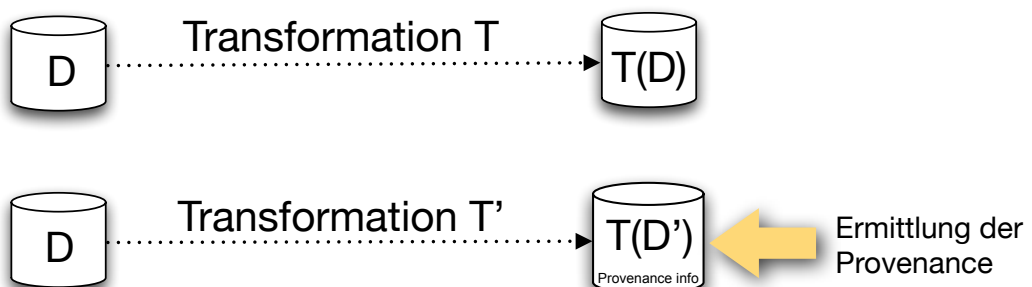


39

Eager Provenance Berechnung

Eager Provenance Berechnung

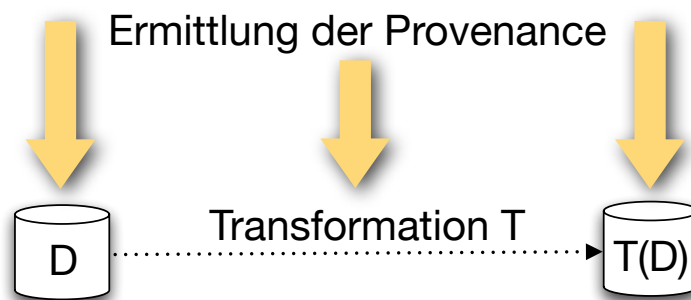
Transformation T wird in eine Transformation T' umgeschrieben, so dass $T'(D)$ sowohl $T(D)$ als auch Provenance-Information enthält.



Eager Provenance Berechnung

Lazy Provenance Berechnung

Provenance-Information wird bei Bedarf aus T, D und T(D) berechnet.



41

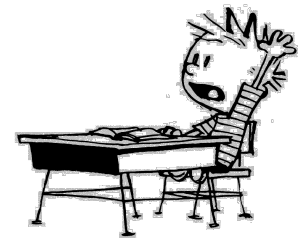
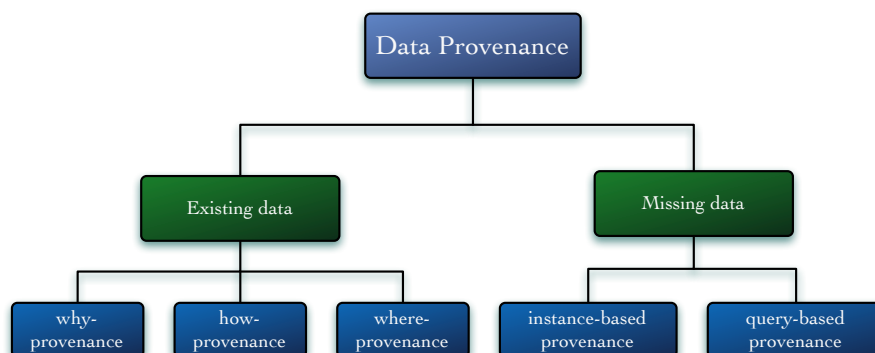
Vor- und Nachteile

	Eager	Lazy
Vorteile	<ul style="list-style-type: none"> • Provenance direkt aus Transformationsergebnis berechenbar → Schnellerer Zugriff auf Provenance-Information 	<ul style="list-style-type: none"> • Kann auf existierende Systeme ohne teures re-engineering angewendet werden. • Keine zusätzlichen Speicherkosten. • Keine längere Anfragebearbeitung.
Nachteile	<ul style="list-style-type: none"> • Komplexere Anfrage → Längere Anfragebearbeitung • Größerer Speicherbedarf bei Materialisierung des Transformationsergebnisses. 	<ul style="list-style-type: none"> • Komplexe Berechnung der Provenance

42

Zusammenfassung

- Data Provenance ist das Problem, zu Objekten im integrierten System diejenigen Objekte in den Quellen zu bestimmen, aus denen das integrierte Objekt abgeleitet wurde.
- Verschiedene Arten der Data Provenance
- Verschiedene Berechnungsarten (Eager vs. Lazy)



43

Literatur

- Survey
 - ▶ [CCT09] J. Cheney, L. Chiticariu, W.C. Tan. *Provenance in Databases: Why, How, and Where*. Foundations and Trends in Databases, vol. 1, number 4. 2009
- Datenherkunft existierender Daten
 - ▶ [CW03] Y. Cui, J. Widom: Lineage tracing for general data warehouse transformations. VLDB Journal 12(1). 2003
 - ▶ [BKT01] P. Buneman, S. Khanna, W.C. Tan. Why and where: A characterization of data provenance. International Conference on Database Theory (ICDT), 2001 .
 - ▶ [GKT07] T.J. Green, G. Karvounarakis, V. Tannen. Provenance Semirings. ACM Symposium on Principles of Database Systems (PODS). 2007.
- Datenherkunft fehlender Daten
 - ▶ [CJ09] A. Chapman, H.V. Jagadish. *Why Not?* International Conference on the Management of Data (SIGMOD). 2009.
 - ▶ [HH10] M. Herschel, M.A. Hernandez. *Explaining Missing Answers to SPJUA Queries*. Proceedings of the VLDB Endowment (PVLDB), Volume 3. 2010.

44