Datenqualität, -fusion und -herkunft

Komplexe Informationssysteme

Fabian Panse

panse@informatik.uni-hamburg.de

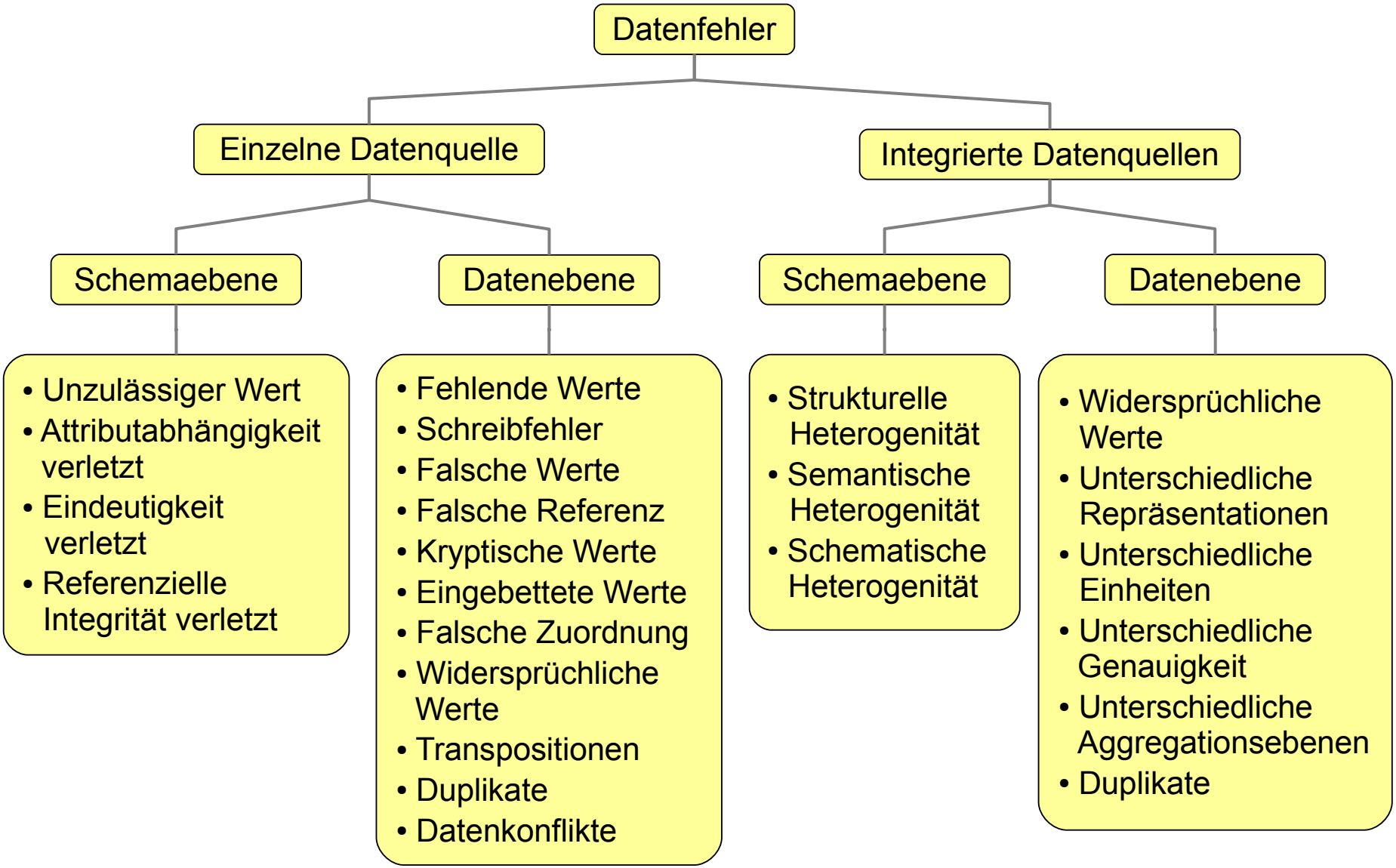
Universität Hamburg

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Datenqualität

Datenfehler [NL06, RD00]



Entstehung von Datenfehlern [NL06]

- Dateneingabe und Erfassung
  (z.B. Dummy-Werte, Falschangaben von Kunden, Tippfehler)
- Alterung
  (z.B. Adressdaten bei Umzug, Name nach Heirat)
- Transformation
  (z.B. falsche oder veraltete Wechselkurse, semantische Heterogenität von korrespondierenden Schemaelementen)
- Integration
  (z.B. Konflikte zwischen Duplikaten)

Auswirkungen von Datenfehlern [NL06]

- Wahrnehmung einer Organisation durch Kunden und Öffentlichkeit:
  - Falsche Preisangaben in Systemen des Einzelhandels kosten Konsumenten in USA jährlich 2,5 Milliarden Dollar.
  - Finanzamt der USA konnte 1992 100.000 Barschecks mit Steuerrückerstattungen aufgrund fehlerhafter Adressangaben nicht zustellen.
  - 2004: in USA von 100.000 Massensendungen durchschnittlich 7.000 aufgrund von Fehlern in Adressen unzustellbar
- Qualität von Unternehmensdaten
  - wichtiger Erfolgsfaktor
  - *garbage-in garbage-out*-Prinzip (z.B. bei Data Warehouse)
  - hohe Kosten für Beseitigung von Datenfehlern
- Datenfehler sind unvermeidlich:
  Fehlerhäufigkeit abhängig vom Vermeidungaufwand

Umgang mit Fehlern [NL06]

- Profiling:
  - Statistiken
  - Musteranalyse (z.B. bei Telefonnummern)
- Assessment
  - Bedingungen, die Daten erfüllen sollen
  - Messung, wie gut diese Bedingungen erfüllt sind
- Maßnahmen zur Fehlerbehebung
- Monitoring (in regelmäßigen Abständen)

Data Scrubbing [NL06]

- einfache Fehler, die nur einzelne Datensätze betreffen
- Normalisierung
  - Schreibweise: Rechtschreibprüfung, *stemming*
  - Adressen
  - Formate, z.B. für Telefonnummern, Datumsangaben, Geldbeträge (Regeln zur automatischen Transformation in Standardformate in kommerziellen Produkten)
- Konvertierung mit Konvertierungsfunktionen
- Fehlende Werte und Ausreißer
  - Profiling-Werkzeug erkennt lückenhafte Datenwertverteilungen (z.B. keine Kunden in Köln)
  - Erfahrungswerte
  - Ausreißer: Erwartungswerte, modellbasierte Ausreißererkennung
- Referenztabellen: einheitliche Schreibweise, Konsistenzprüfung (z.B. Ortsnamen, Bankverbindungen, Handelsregister)

Datenqualität im Integrationssystem

Integrierte Informationssysteme besonders anfällig für Qualitätsprobleme

- Probleme akkumulieren
  - Qualität der Ursprungsdaten (Eingabe, Fremdfirmen, ...)
  - Qualität der Quellsysteme (Konsistenz, ICs, Fehler, ...)
  - Qualität der Integrationsprozesse (Parsen, Transformieren, Mappings)
- Probleme treten (häufig) erst bei integrierter Sicht zu Tage

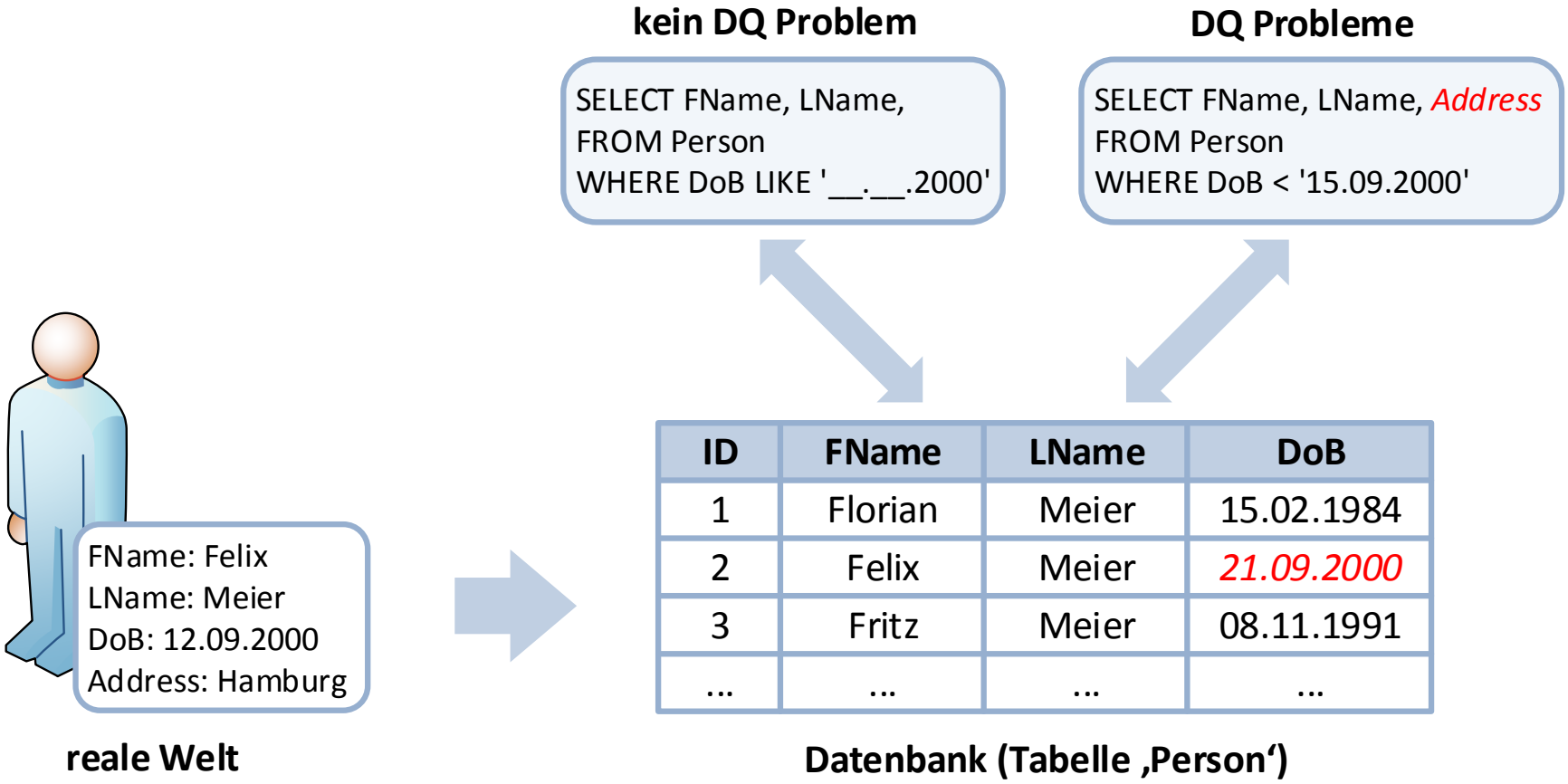
Kosten schlechter Datenqualität

- Unternehmensberatung A.T. Kearny: 25%-40% der operativen Kosten entstehen durch schlechte Datenqualität
- Data Warehouse Institute: Industrie und Verwaltung in den USA verlieren jährlich 600 Milliarden USD
- SAS Studie: Nur 18% der deutschen Betriebe vertrauen ihren Daten
- 80% aller Krankenhaus Datensätze enthalten Fehler

Datenqualität – Fitness for Use

- Gebrauchstauglichkeit der Daten in den Augen der Anwendung

⇒ Beurteilung von Qualität leitet sich aus den individuellen Bedürfnissen der verarbeitenden Anwendungen ab



Inhaltsbasierte DQ Kriterien [WS96]

- Accuracy: ist der Grad, zu dem die Daten korrekt sind
- Completeness: ist der Grad, zu dem Daten nicht fehlen und von ausreichender Breite, Tiefe und Umfang für die jeweilige Aufgabe sind
- Documentation: ist der Umfang und die Nützlichkeit von Metadaten
- Interpretability: ist der Grad, zu dem sich die Daten in geeigneten Sprachen, Symbolen und Einheiten befinden und klaren Definitionen folgen
- Relevancy (or relevance): ist der Grad, zu dem die Daten verwendbar und hilfreich für die jeweilige Aufgabe sind
- Reliability: ist der Grad, zu dem der Benutzer den Informationen vertrauen kann
- Value-Added: ist der Grad, zu dem die Daten dienlich sind und sich Vorteile aus ihrer Nutzung bieten

Technische DQ Kriterien [WS96]

. . . betrifft Software und Hardware

- Accessibility (or availability): ist der Grad, zu dem Daten verfügbar oder leicht und schnell empfangbar sind
- Latency: ist die Zeit in Sekunden, von dem Absenden der Anfrage, bis das erste Datenelement den Benutzer erreicht hat
- Price (cost effectiveness): ist der Geldbetrag, den ein Benutzer für eine Anfrage bezahlen muss
- Response time: misst die Verzögerung in Sekunden zwischen dem Absenden einer Anfrage durch den Benutzer und dem Empfang der kompletten Antwort vom Integrationssystem
- Security: ist der Grad, zu dem der Zugang zu den Daten angemessen beschränkt ist, um deren Sicherheit zu gewährleisten
- Timeliness: ist der Grad, zu dem das Alter der Daten für die jeweilige Aufgabe geeignet ist

Intellektuelle DQ Kriterien [WS96]

. . . betrifft subjektive Aspekte

- Believability: ist der Grad, zu dem die Daten als wahr, real und glaubwürdig angesehen werden
- Objectivity: ist der Grad, zu dem die Daten unvoreingenommen (unbiased) sind
- Reputation: ist der Grad, zu dem der Inhalt bzw. der Ursprung der Daten vertrauenswürdig ist

Instanzbezogene DQ Kriterien [WS96]

. . . betrifft die Darstellung der abgefragten Daten

- Amount of data: ist der Grad, zu dem die Größe oder das Volumen der verfügbaren Daten geeignet ist
- Representational conciseness: ist der Grad, zu dem die Daten kompakt dargestellt werden, ohne zu erdrücken
- Representational consistency: ist der Grad, zu dem die Daten immer im gleichen Format dargestellt werden und mit früheren Daten kompatibel sind
- Understandability (ease of understanding): ist der Grad, zu dem die Daten unzweideutig sind und leicht verstanden werden können
- Verifiability (traceability): ist der Grad, zu dem die Daten gut dokumentiert, belegbar und leicht einer Quelle zugeschrieben werden können

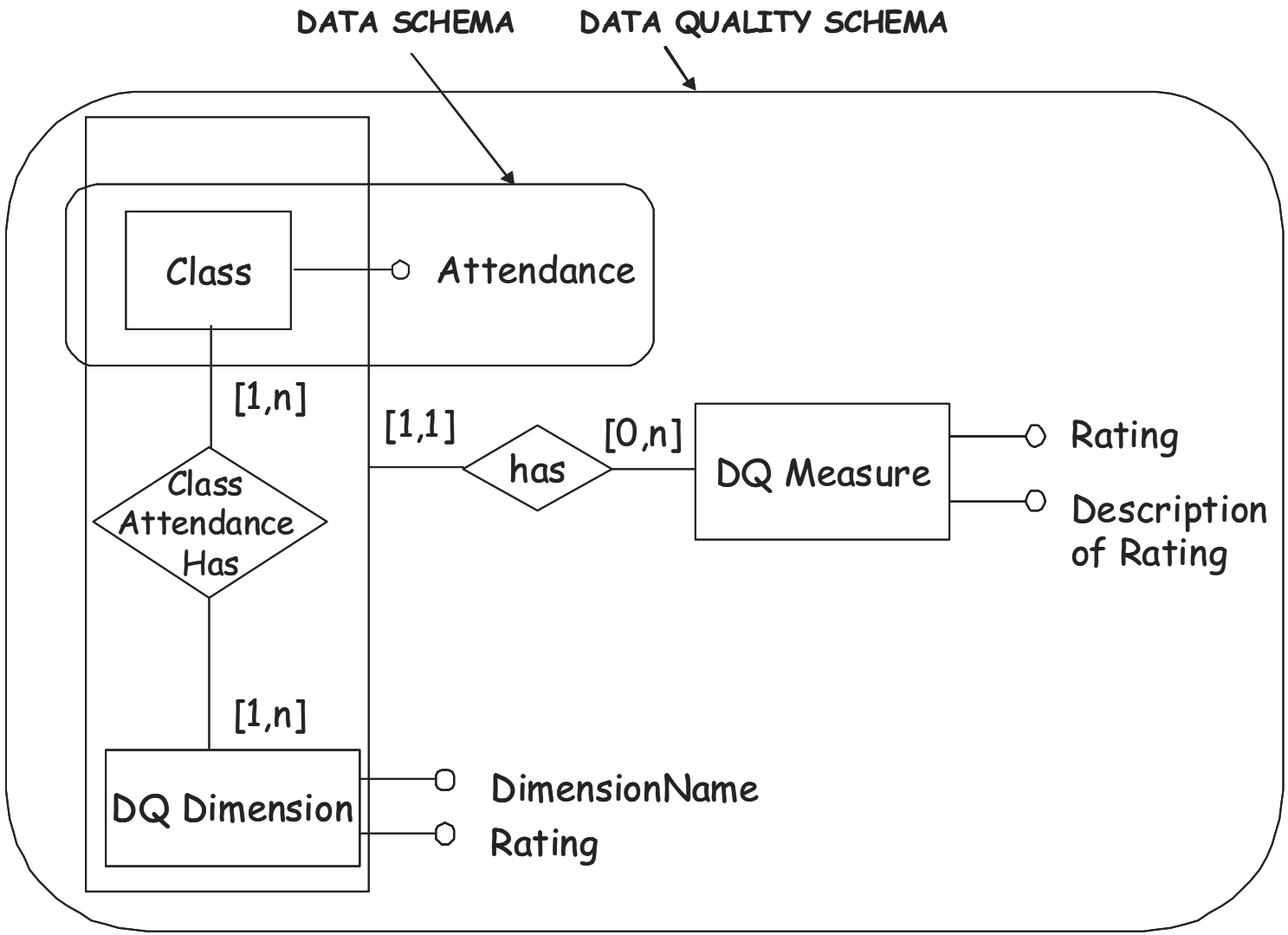
DQ Kriterien – weitere Klassifikationsmodelle [BS06]

Category	Dimension	Definition: the extent to which ...
Intrinsic	Beleivability	data are accepted or regardedas true, real and credible
	Accuracy	data are correct, reliable and certified free of error
	Objectivity	data are unbiased and impartial
	Reputation	data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task at hand
	Completeness	data are of sufficient depth, breadth, and scope for the task at hand
	Appropriate amount of data	the quantity or volume of available data is appropriate
Representational	Intepretability	data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
	Representational consistency	data are always presented in the same format and are compatible with the previous data
	Concise representation	data are compactly represented without behing overwhelmed
Accessibility	Accessibility	data are available or easily and quickly retrieved
	Access security	access to data can be restricted and hence kept secure

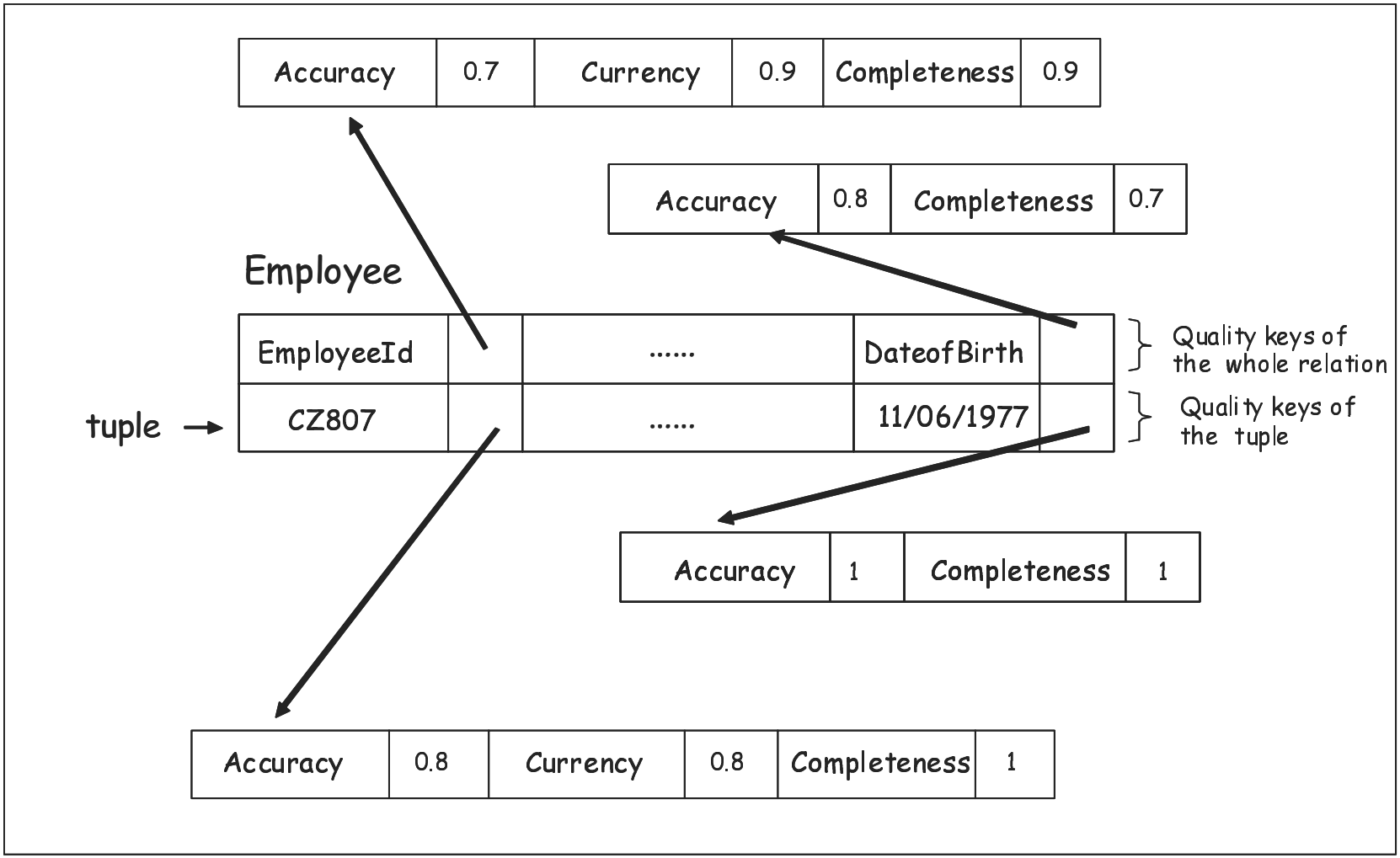
DQ Kriterien – weitere Klassifikationsmodelle [BS06]

Dimension Name	Type of dimension	Definition
Accuracy	data value	Distance between v and v', considered as correct
Completeness	data value	Degree to which values are present in a data collection
Currency	data value	Degree to which a datum is up-to-date
Consistency	data value	Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules
Appropriateness	data format	One format is more appropriate than another if it is more suited to user needs
Interpretability	data format	Ability of the user to interpret correctly values from their format
Portability	data format	The format can be applied to as a wide set of situations as possible
Format precision	data format	Ability to distinguish between elements in the domain that must be distinguished by users
Format flexibility	data format	Changes in user needs and recording medium can be easily accommodated
Ability to represent null values	data format	Ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain
Efficient use of memory	data format	Efficiency in the physical representation. An icon is less efficient than a code
Representation consistency	data format	Coherence of physical instances of data with their formats

Modellierung der Datenqualität (Metamodell) [BS06]

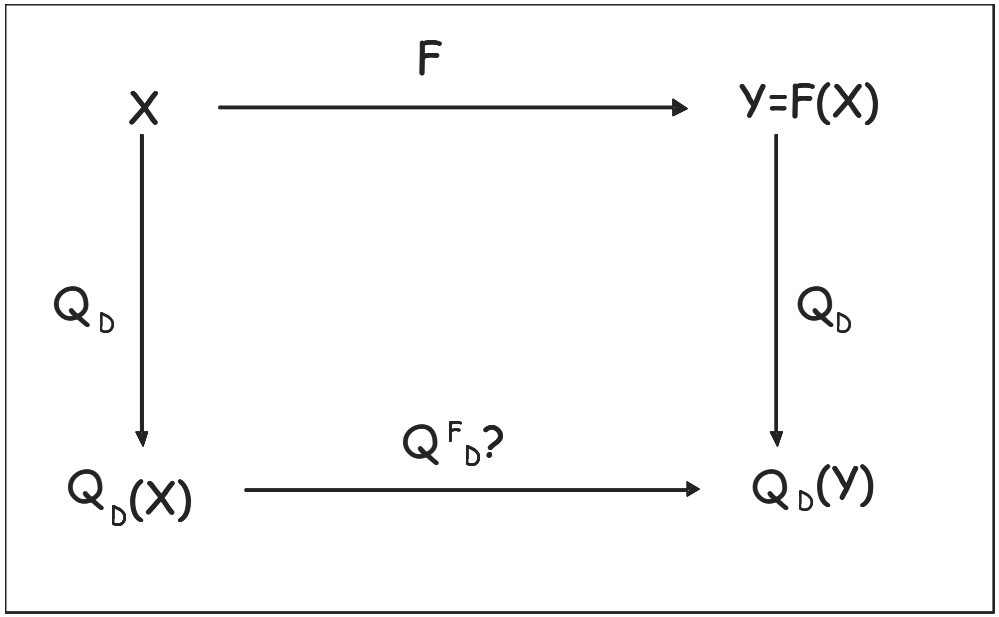


Modellierung der Datenqualität (Datenmodell) [BS06]



Komposition von Qualitätswerten [BS06]

- Ziel: Abschätzung der Qualtität eines Transformations-ergebnisses (Wert $Q_D(Y)$) basierend auf den Qualitätswerten der Eingabedaten (Wert $Q_D(X)$)
- Transformation (Mapping F) kann u.a. sein:
  - Maßnahme zur Qualitätssteigerung
  - Integration mehrerer Quellen



Accuracy (Korrektheit bzw. Genauigkeit)

- Nähe des gegebenen Datenwertes v zu dem korrekten Wert v' der modellierten Wirklichkeit
  - Boolesche Modellierung: 1 wenn korrekt ($v \equiv v'$), 0 sonst
  - Differenzierte Modellierung: Verwendung eines Ähnlichkeitsmaßes, d.h. $sim(v, v')$
- Als modellierte Wirklichkeit wird eine Referenztabelle verwendet (z.B. Adressangaben der Post)
- *Fitness for Use* modelliert in $\equiv$ oder *sim* (z.B. zwei Daten sind äquivalent wenn ihre Jahreszahlen identisch sind)
- Korrektheit der Datenbank berechnet sich aus den Korrektheitswerten der einzelnen Datenwerte (z.B. avg)
- Alternativ: Korrektheit der Datenbank als durchschnittliche Korrektheit der Anfrageergebnisse

Accuracy (Korrektheit bzw. Genauigkeit) [BS06]

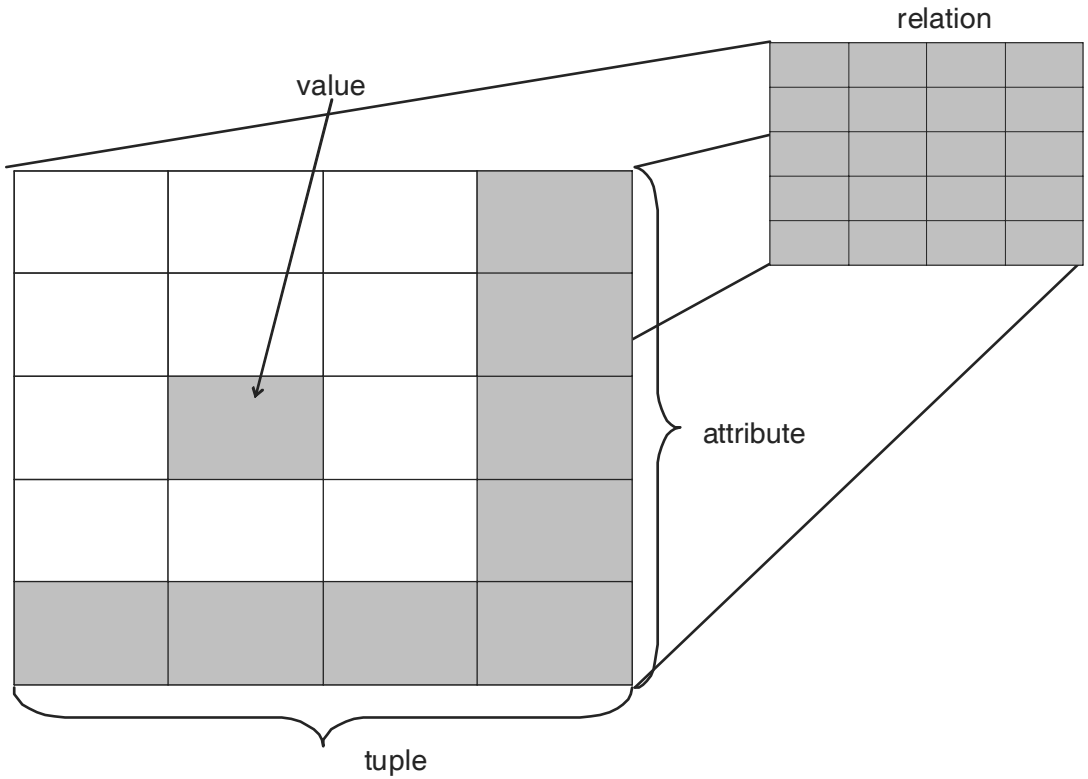
- Inkorrekte Werte können verschieden schwere Auswirkungen haben
- *Strong accuracy error:* Anteil an Tupeln bei denen Datenfehler eine korrekte Identifikation verhindern

$$\sum_{i=1}^{N} \frac{\beta((q_i > 0) \wedge (s_i = 1))}{N}$$

  - $\beta(X) = 1$ wenn X wahr ist und 0 sonst
  - $s_i = 0$ wenn Tupel t_i einem Tupel der Referenztabelle zugeordnet werden kann und 1 sonst
  - $q_i = 1$ wenn Tupel t_i einen inkorrekten Wert beinhaltet und 0 sonst
- Da q_i nicht auf Basis der Referenztabelle berechnet werden kann, sind andere Lösung notwendig (z.B. Vergleich mit fixem Wertebereich)

Completeness (Vollständigkeit) [NL06, BS06]

- Deckung: extensionale Vollständigkeit (d.h. Anteil an intendierten Tupeln die vorhanden sind)
- Dichte: intensionale Vollständigkeit (d.h. Anteil an Nicht-Nullwerten)
- Vollständigkeit: Dichte $\times$ Deckung



Datenqualitätsmaße – Anforderungen [HKK07]

- Normierung:
  - erhöht Interpretierbarkeit und Vergleichbarkeit
  - meist auf Intervall $[0, 1]$
- Kardinale Skalierung:
  - absolute bzw. relative Veränderungen sind interpretierbar (z.B. doppelt so viele Fehler $\Rightarrow$ Korrektheitswert halbiert sich)
  - Interpretation eines Qualitätsunterschieds hängt nur vom Wert und nicht vom Bereich ab (d.h. Differenz von 0.1 ist immer gleich bedeutend, egal ob zwischen 0.1 und 0.2 oder 0.9 und 1.0)
  - unterstützt eine Betrachtung der zeitlichen Entwicklung
- Sensibilisierbarkeit:
  - ermöglicht zielgerichte Konfigurierung auf Anwendungsfall (z.B. Gewichtung von Attributen)

Datenqualitätsmaße – Anforderungen [HKK07]

- Aggregierbarkeit:
  - Messung auf verschiedenen Granularitätsebenen (z.B. Attributwert-, Tupel-, Relationen-, Datenbankebene)
  - Wert einer Ebene ergibt sich aus Werten der darunterliegenden Ebene (z.B. Korrektheit einer Relation basierend auf Korrektheit der Tupel)
- Operationalisierbarkeit mittels Messverfahren:
  - Messverfahren welche die Metriken operationalisieren
  - ermöglicht praktische Anwendung
- Fachliche Interpretierbarkeit:
  - Metrikergebnisse sollen fachlich interpretierbar sein (bspw. als Anteil der korrekt erfassten Attributwerte in einer Datenbank)
  - ermöglicht Dritten die Ergebnisse nachzuvollziehen

Kardinale Skalierung – Beispiel [HKK07]

Verbesserung der Korrektheit $Q_{Korr.}(w_i, w_j)$	Notwendige Veränderung von $d(w_i, w_j)$
$0.0 \rightarrow 0.5$	$\infty \rightarrow 1.0$
$0.5 \rightarrow 1.0$	$1.0 \rightarrow 0.0$

- Distanzmaß d: Hamming, Levenshtein
- Korrektheitsmaß:

$$Q_{Korr.}(w_i, w_j) = \frac{1}{d(w_i, w_j) + 1}$$

- Skaliert nicht kardinal
  ($1.0 \rightarrow 0.0$ entspricht einem unterschiedlichen Zeichen)

Kardinale Skalierung – Beispiel [HKK07]

Verbesserung der Korrektheit $Q_{Korr.}(w_i, w_j)$	Notwendige Veränderung von $d(w_i, w_j)$
$0.0 \rightarrow 0.5$	$1.0 \rightarrow 0.5$
$0.5 \rightarrow 1.0$	$0.5 \rightarrow 0.0$

- Distanzmaß d_1: Hamming, Levenshtein
- Korrektheitsmaß:

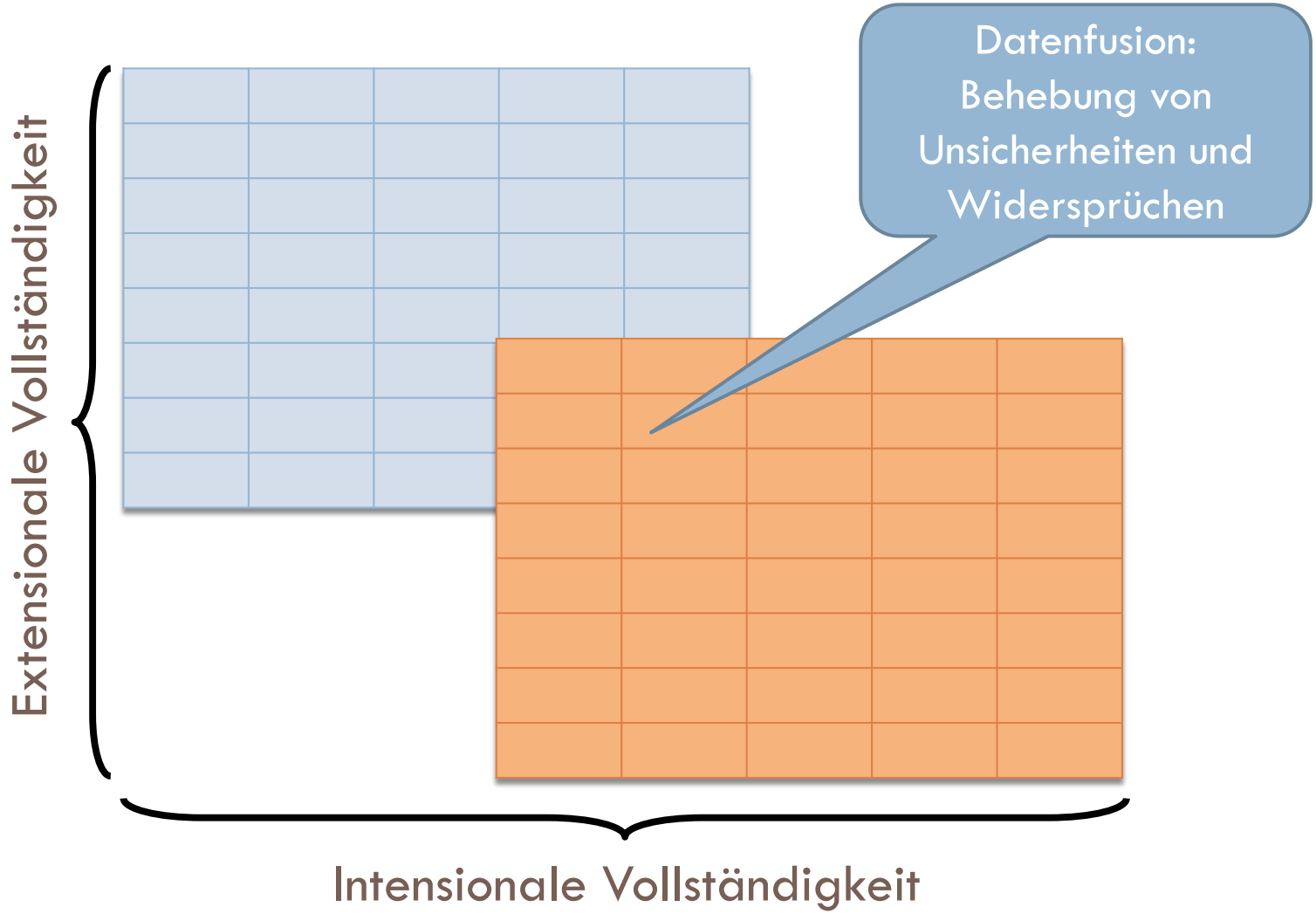
$$Q_{Korr.}(w_i, w_j) = 1 - d(w_i, w_j)$$

$$\text{mit} \quad d(w_i, w_j) = \frac{d_1(w_i, w_j)}{max(|w_i|, |w_j|)}$$

- Skaliert kardinal

Datenfusion

Vollständigkeit, Korrektheit von Integrationsergebnissen

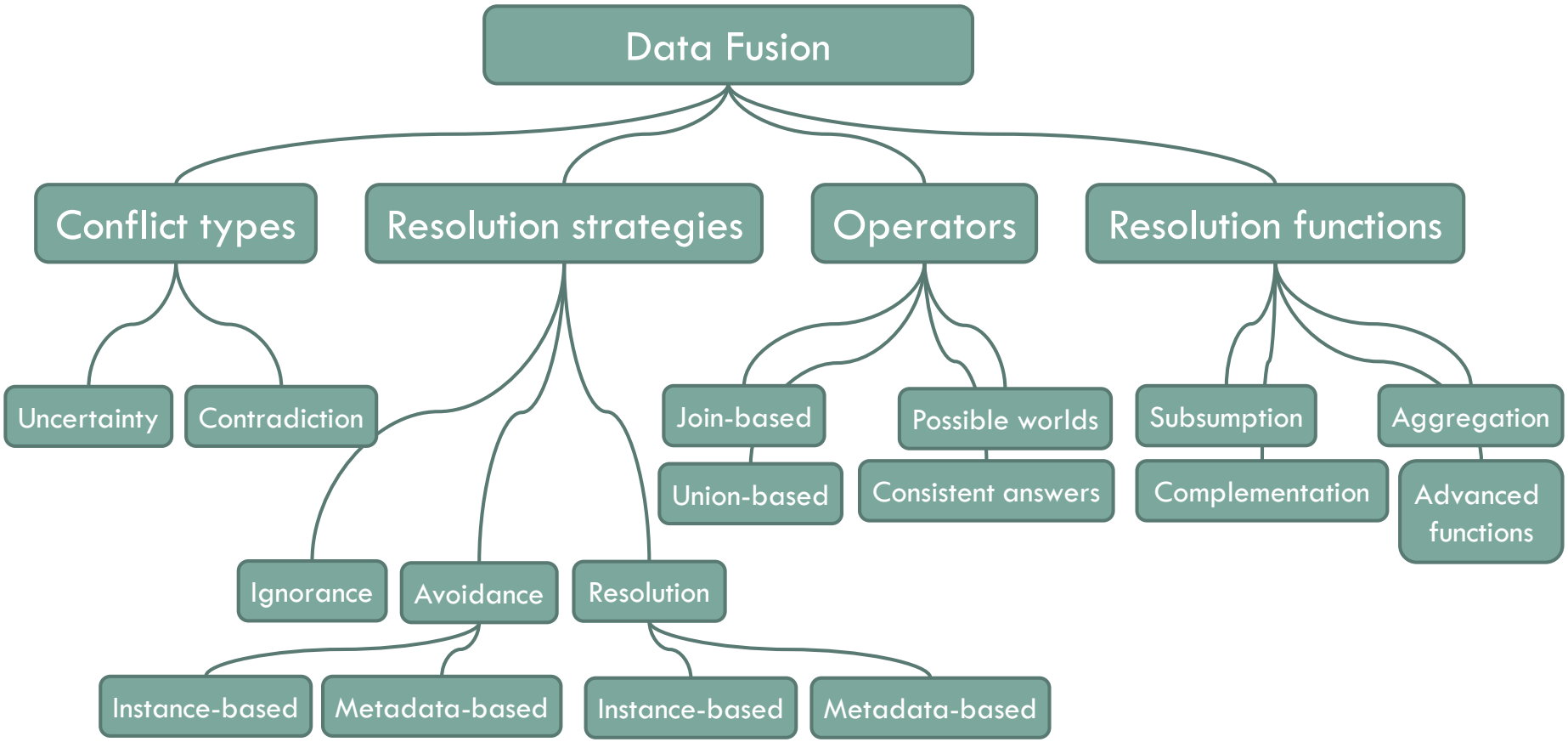


Quelle: Luna Dong, Felix Naumann [DN09]

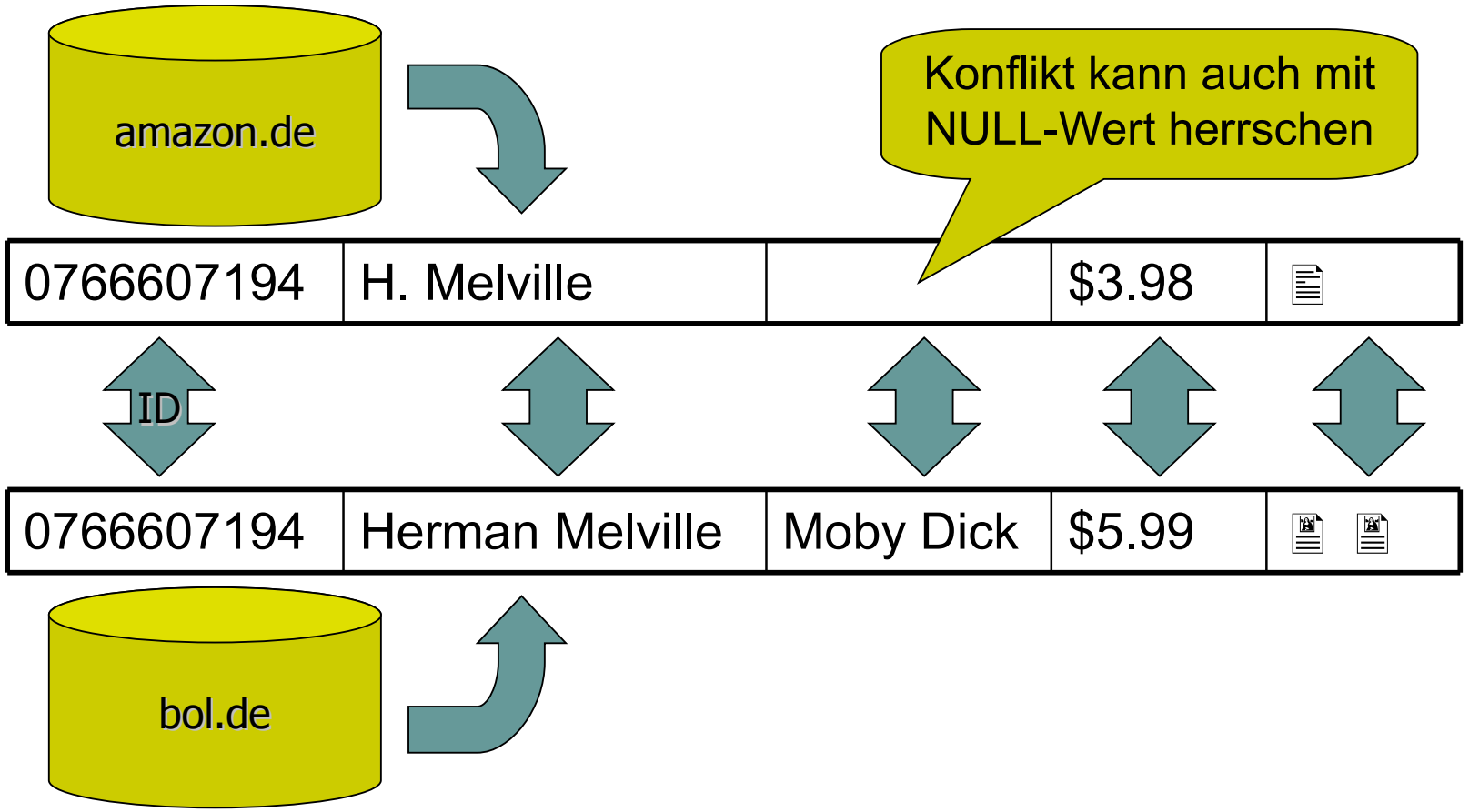
Überblick Datenfusion [NL06]

- Nach Duplikaterkennung: Kombination von Duplikaten, so dass im Ergebnis keine Entität der Realwelt mehr als einmal repräsentiert ist
- Ergebnisdatensätze angereichert mit Datenwerten aus unterschiedlichen Quellen (intensionale Komplementierung)
- Datenwerte aus unterschiedlichen Quellen können sich widersprechen (Datenkonflikte)

Überblick Datenfusion [DN09]



Datenkonflikte



Ursachen für Konflikte innerhalb einer Datenquelle [DN09]

- Kein Integritäts- oder Konsistenzcheck
- Redundanzen im Schema (z.B. Geburtsdatum und Alter)
- Tippfehler, Übertragungsfehler, falsche Berechnungen
- Verschiedene Varianten
  - Kantstr. / Kantstrasse / Kant Str. / Kant Strasse
  - Kolmogorov / Kolmogoroff / Kolmogorow
- Typische Vertauschungen (OCR)
  - U $\leftrightarrow$ V, 0 $\leftrightarrow$ o, 1 $\leftrightarrow$ I, etc.
- Obsolete Werte
  - Verschiedene Aktualisierungshäufigkeit, vergessene Aktualisierungen

Ursachen für Konflikte zwischen Datenquellen [DN09]

- Lokal konsistent, aber global inkonsistent (z.B. Primärschlüssel)
- Unterschiedliche Datentypen
- Lokale Schreibvarianten und Konventionen
  - Adressen
    - St → Street, Ave → Avenue, etc.
    - R.-Breitscheid-Str. 72 a → Rudolf-Breitscheid.-Str. 72A
    - 128 Schreibweisen für Frankfurt am Main (z.B. Frankfurt a.M., Frankfurt/M, Frankfurt, Frankfurt a. Main)
  - Namen
    - Dr. Ing. h.c. F. Porsche AG
    - Hewlett-Packard Development Company, L.P.
  - Numerische Daten
    - 10.000 €= 10T EURO = 10k EUR = 10.000,00 €= 10,000.-€
  - Telefonnummern, Geburtsdaten, etc.
- Lokal gebräuchliche Wörter (z.B. verkuzelt, rummähren)

Nullwert Semantiken

- unbekannt
  - Es gibt einen Wert, den kennen wir aber nicht
  - Zum Beispiel: Unbekannte Adresse
- nicht zutreffend (*not applicable*)
  - Es gibt keinen sinnhaften Wert
  - Zum Beispiel: Ehepartner bei Singles/Unverheirateten
- vorenthalten (*withheld*)
  - Es gibt einen Wert, wir sind aber nicht authorisiert ihn zu sehen
  - Zum Beispiel: Private Telefonnummer
- nicht bekannt ob fehlend, nicht zutreffend oder vorenthalten
  - *No information* Semantik
- ANSI/X3/SPARC interim report von 1975: 14 verschiedene Nullwert Semantiken

Konfliktbeziehungen [NL06]

1. Gleichheit: Die Tupel sind in allen Attributwerten identisch.
2. Subsumption: Ein Tupel subsumiert ein anderes, wenn es weniger Nullwerte hat und in jedem Attribut mit einem Nicht-Nullwert den gleichen Wert wie das andere Tupel besitzt.
3. Komplementierung: ein Tupel komplementiert ein anderes, wenn keines der beiden das andere subsumiert und wenn es für jedes Attribut mit einem Nicht-Nullwert entweder den gleichen Wert wie das andere Tupel hat oder das andere Tupel an dieser Stelle einen Nullwert besitzt.
4. Widerspruch (*contradiction*): Bei Widersprüchen gibt es mindestens ein Attribut, in dem beide Tupel unterschiedliche Werte haben, die nicht `NULL` sind.

Konfliktbeziehungen – Beispiele [NL06]

Film					
	ID	**titel**	**regisseur**	**jahr**	**studio**
t_1	1	Alien	Scott	1980	Fox
t_2	1	Alien	Scott	1980	Fox
t_3	1	Alien	Scott	1980	`null`
t_4	1	Alien	`null`	1980	`null`
t_5	1	Alien	Scott	1982	MGM
t_6	1	Alien	`null`	1980	MGM

- Welche Tupel sind gleich? (t_1 und t_2)
- Welches Tupel subsumiert welches andere Tupel?
  (t_3 und t_6 subsumieren beide t_4)
- Welche Tupel komplementieren einander? (t_3 und t_6)
- Welche Tupel widersprechen sich? (t_3 undt_5, t_1 und t_6, etc.)

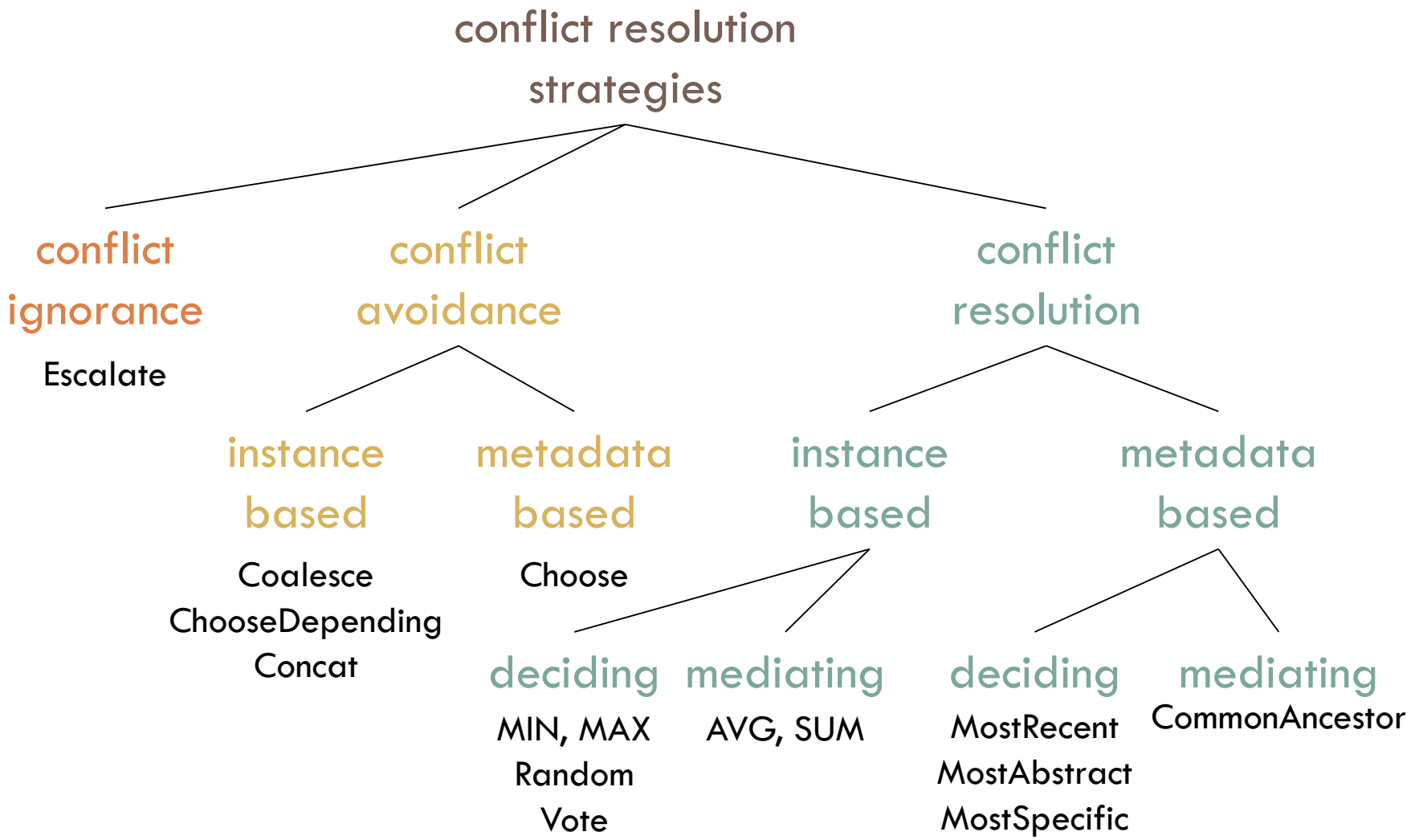
Konfliktlösungsfunktion

$$f(x,y) := \begin{cases} \texttt{null} & \text{wenn} \quad x = \texttt{null} \wedge y = \texttt{null} \\ x & \text{wenn} \quad y = \texttt{null} \wedge x \neq \texttt{null} \\ y & \text{wenn} \quad x = \texttt{null} \wedge y \neq \texttt{null} \\ \underbrace{g(x,y)}_{\text{interne Konfliktlösungsfkt.}} & \text{sonst} \end{cases}$$

Interne Konfliktlösungsfunktionen [DN09]

Function	Description	Examples
Min, Max, Sum, Count, Avg	Standard aggregation	NumChildren, Salary, Height
Random	Random choice	Shoe size
Longest, Shortest	Longest/shortest value	First_name
Choose(source)	Value from a particular source	DoB (DMV), CEO (SEC)
ChooseDepending(val, col)	Value depends on value chosen in other column	city & zip, e-mail & employer
Vote	Majority decision	Rating
Coalesce	First non-null value	First_name
Group, Concat	Group or concatenate all values	Book_reviews
MostRecent	Most recent (up-to-date) value	Address
MostAbstract, MostSpecific, CommonAncestor	Use a taxonomy / ontology	Location
Escalate	Export conflicting values	gender
…	…	…

Interne Konfliktlösungsfunktionen [DN09]



Minimum Union [DN09]

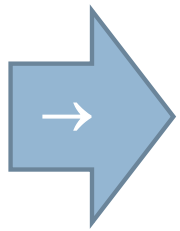
A	B	C
a	b	c
e	f	g
m	n	o

$\dot{\cup}$

A	B	D
a	b	⊥
e	f	h
m	p	⊥

=

A	B	C	D
a	b	c	⊥
a	b	⊥	⊥
e	f	g	⊥
e	f	⊥	h
m	n	o	⊥
m	p	⊥	⊥



A	B	C	D
a	b	c	⊥
e	f	g	⊥
e	f	⊥	h
m	n	o	⊥
m	p	⊥	⊥

- Union: Eliminierung exakter Duplikate
- Minimum Union: Eliminierung subsumierter Tupel

Weitere Relationale Operatoren:

- Full Disjunction: Minimum Union über Full-Outer Join mehrerer Quellpaare
- Complement Union: Outer-Union mit Komplementierung

Weitere Aspekte

- ICAR Eigenschaften binärer Konfliktlösungsfunktionen:
  - Idempotenz (d.h. $g(x, x) = x$)
  - Kommutativität (d.h. $g(x, y) = g(y, x)$)
  - Assoziativität (d.h. $g(z, g(x, y)) = g(x, g(y, z))$)
  - Repräsentativität (d.h. $x \approx z \Rightarrow g(x, y) \approx z$)
- Lernende Verfahren zur Konfliktbehebung
- Erkennen von Kopiemustern zwischen Datenquellen
- Aufbau und Nutzen von Trust Mappings

Datenherkunft

Data Lineage – Definition

Umgangssprachlich:

Definition (Data Lineage) Data Lineage (auch Data Provenance) ist das Problem, zu Objekten im integrierten Informationsystem diejenigen Objekte in den Quellen zu bestimmen, aus denen das integrierte Objekt abgeleitet wurde.

Formal:

Definition (Data Lineage) Sei D eine Menge von Datenquellen. Sei T eine Transformation von Daten aus D. Sei weiter $O = T(D)$ und $o \in O$. Dann ist die *Lineage* $lin(o) \subseteq D$ die Menge der Eingabewerte, die zur Ausgabe o beiträgt. Wenn $O^* \subseteq O$, dann $lin(O^*, D) = \cup_{o \in O^*} lin(o, D)$

Data Lineage – Motivation und Probleme

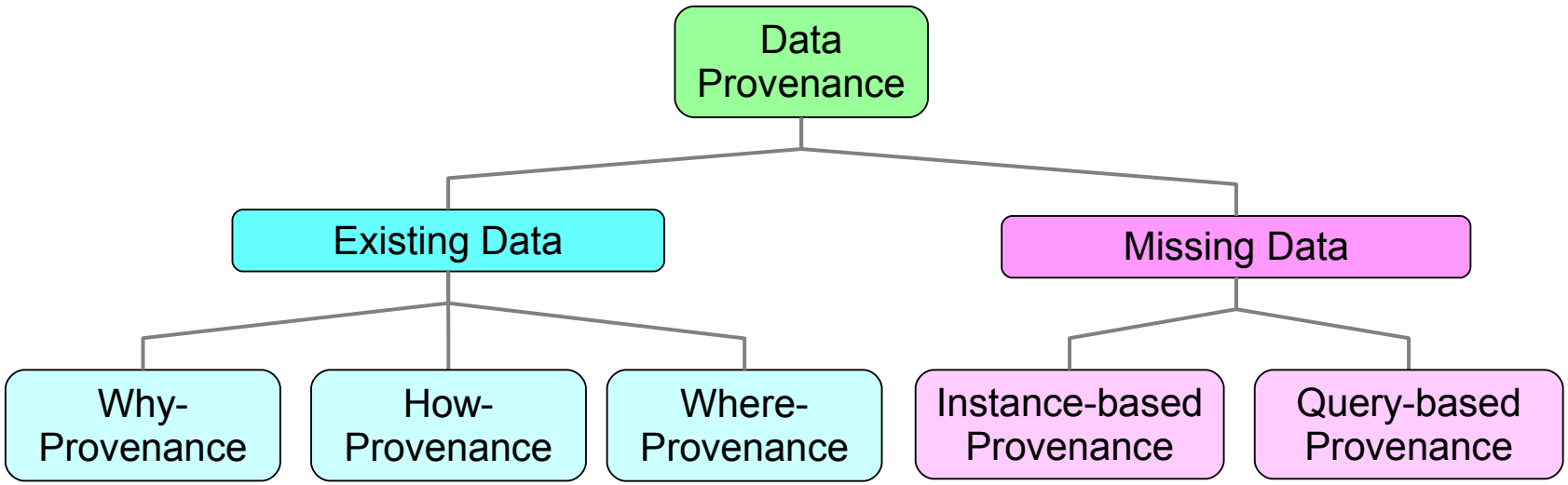
Motivation:

- Rückwärtsgerichtete Analyse von Datenabhängigkeiten
  - Warum gehört/fehlt ein Tupel zu/in meinem Anfrageergbnis?
  - Warum haben meine Ergebnistupel die Werte die sie besitzen?
- Vorwärtsgerichtete Analyse von Datenabhängigkeiten
  - Wie wird mein Anfrageergebnis beeinflusst wenn ich ein Tupel/Attributwert ändere/einfüge/lösche? (besonders interessant für materialisierte Sichten)

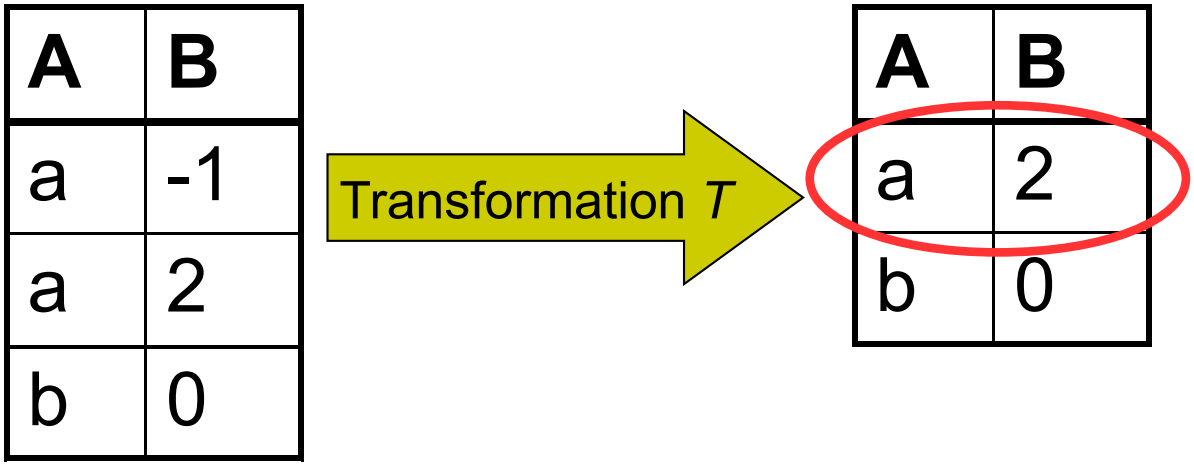
Probleme:

- Runtime Overhead
- Speicherbedarf
- Trade-off zwischen Nutzen und Kosten

Data Lineage – Taxonomie



Data Lineage – Beispiel

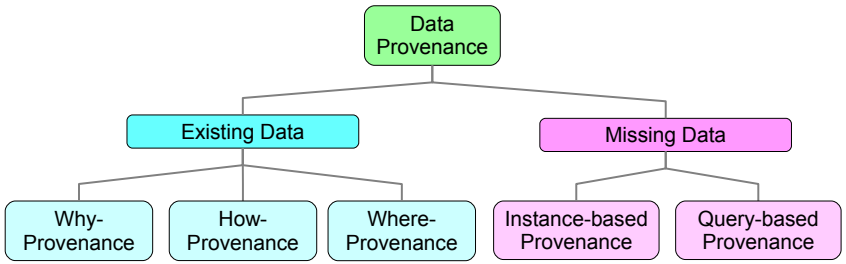


Herkunft des Tupels $(a, 2)$

- $T = \sigma_{B \geq 0}$
  $\Rightarrow lin((a, 2)) = \{(a, 2)\}$
- $T =$ Gruppierung nach A und Aggregation: $2 \times sum(B)$
  $\Rightarrow lin((a, 2)) = \{(a, -1), (a, 2)\}$
- $T =$ Gruppierung nach A und Aggregation: $max(B)$
- ...

Data Lineage – Übersicht [CCT09]

- Why-provenance: Aus welchen Quelldaten kombiniert sich ein Ergebnistupel?
- How-provenance: Wie werden die Tupel aus den Quellen kombiniert, um ein Ergebnistupel zu produzieren?
- Where-provenance: Aus welchen Quellen wurden Werte in ein Ergebnistupel kopiert?
- Instanzbasierte Erklärungen: Wie müsste man die Quelldaten anpassen, um gewünschtes Ergebnis zubekommen?
- Anfragebasierte Erklärungen: Welche Operatoren sind für das Verschwinden der Daten verantwortlich [CJ09]?



Why- vs. 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.

Why- vs. How-Provenance – Beispiel

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 *tuple set* ist unklar, dass t1 zweimal 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
        FROM Touren
       ) 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 × (t1 + t3)
Santa Cruz	831-3000	t2 × t2
Santa Cruz	415-1200	t1 × (t4 + t5)
Monterey	415-1200	t1 × t6
Monterey	831-3000	t1 × t7
Carmel	831-3000	t1 × t8

Where-Provenance

- Beschreibt, aus welchem Ort innerhalb einer 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.

Why- vs. Where-Provenance – Beispiel

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}

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
       t.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

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 Quelltupel in D zur Bildung von t verwendet werden (*witness basis*).
  - How-provenance beschreibt in Form eines Polynoms, wie welche Tupel 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.

Erklären fehlender Daten

- Erkläre, warum bestimmte Daten nicht im Ergenbis einer Anfrage auftauchen.
- Unterscheidung
  - Instanzbasierte Erklärungen
  - Anfragebasierte Erklärungen

Instanzbasierte Erklärungen – Beispiel

Agenturen

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

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

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

Warum fehlt folgendes Tupel
Gründe in den Quelldaten .suchen

CoastTours	123-4567

Quelle: Melanie Herschel, Universität Stuttgart

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 – Ergänzungen

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	Update
	CoastTours	?	boot	?	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

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

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

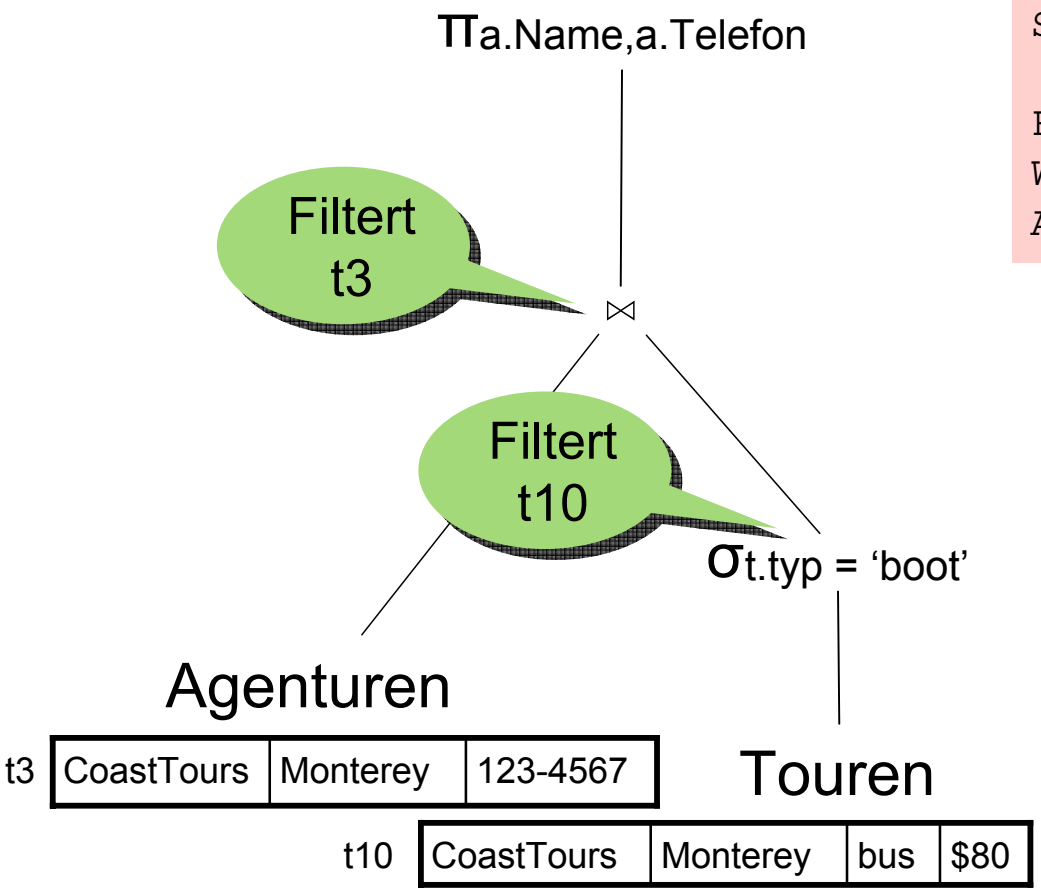
Warum fehlt folgendes Tupel Gründe in der Anfrage suchen.

CoastTours	123-4567

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 – Anfragebaum



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
CoastTours	123-4567

Literatur I

[BS06] Carlo Batini and Monica Scannapieco.
Data Quality: Concepts, Methodologies and Techniques.
Data-Centric Systems and Applications. Springer, 2006.

[CCT09] Cheney, Chiticariu, and Tan.
Provenance in databases: Why, how, and where.
In *Foundations and Trends in Databases*,1(4), 2009.

[CJ09] A. Chapman and H.V. Jagadish.
Why not?
In *Proc. of the ACM Int. Conf. on Management of Data* (SIGMOD), 2009.

[DN09] Luna Dong and Felix Naumann.
Data Fusion.
VLDB Tutorial, 2009.

[GKT07] T.J. Green, G. Karvounarakis, and V. Tannen.
Provenance semirings.
In *Proc. of the Symposium on Principles of Database Systems* (PODS), 2007.

Literatur II

[HKK07] Bernd Heinrich, Marcus Kaiser, and Mathias Klier.
Metrics for Measuring Data Quality - Foundations for an Economic Data Quality Management.
In *ICSOFT* (ISDM/EHST/DC), 2007, pages 87-94.

[LN06] Ulf Leser and Felix Naumann.
Informationsintegration.
dpunkt.verlag, 2006.
In German.

[RD00] Erhard Rahm and Hong-Hai Do.
Data cleaning: Problems and current approaches.
In *IEEE Data Engineering Bulletin* 23(4), 2000.

[WS96] Richard Y. Wang and Diane M. Strong.
Beyond accuracy: What data quality means to data consumers.
In *Journal on Management of Information Systems* 12(4), 1996.