Seminar Datenqualität
Einführung
27. Oktober 2011

Agenda

Allgemeines

Themen Datenqualität

Seminardurchführung

Themenvergabe

Allgemeines: Seminarablauf

Leistungen:

- 1 Vortrag (vorab in der Konsultation besprechen!)
- Ausarbeitung (ca. 12-15 Seiten pro Teilnehmer
  davon mind. 8 Inhalt)
- Einführungsveranstaltung am 27.10.2011 (heute)
- Präsentationen am 1.12. und 8.12.2011 von 15 - 19 Uhr
- Abgabe der Ausarbeitung 06.02.2012

Bewertung

- Vorträge ~40-50% (20 min Vortrag + 10 min Diskussion)
- Ausarbeitung 50-60%

Allgemeines: Warum dieses Seminar

Wichtige "Soft Skills" oder auch Schlüsselkompetenzen erlernen

- ein wissenschaftliches Papier schreiben
- wissenschaftliche Literatursuche und -analyse
- Vortragsweisen und -stil üben
- "Konferenzflair" erleben
- Arbeit mit entsprechenden Vorlagen (Empfehlung: LaTeX)
- ...

Diese sind genauso wichtig für eine akademische Karriere wie für eine Karriere in der Wirtschaft

Allgemeines: Themengebiete

Wir geben keine speziellen Themen vor, sondern nur größere Themengebiete

Ein Themengebiet wird wie folgt bearbeitet:

- erste Sichtung der aktuellen Forschung (nicht älter als 5 Jahre)
- Auswahl eines spezielleren Themas aus der ersten Sichtung
- die Auswahl nicht zu weit fassen, ein Thema wie "Data Warehousing" kann man nicht erfassen
- das gewählte Thema sollte in den letzten Jahren eine wissenschaftliche Relevanz haben/gehabt haben
- Verständnis und Einordnung des gewählten Themas in die aktuelle Forschung
- kritische Analyse und Abgrenzung des Themas gehört ebenfalls dazu
- ...

Allgemeines: Literatur

- offene Fragen in Bezug auf die Forschung aus bereits besuchten Lehrveranstaltungen
- allgemein große Konferenzen wie VLDB, SIGMOD/PODS, etc.
- Bibliothek Online und gedruckt
- scholar.google.com
- ACM Digital Library
- dblp.uni-trier.de
- IEEE Xplore
- ...
- Related Work in den Papieren und den Portalen ...

Motivation: Datenqualität

Fitness for use

Quelle: Chrisman 1984

Was **verbirgt** sich nun hinter dem Begriff **Datenqualität**

“Fitness for use” = Gebrauchstauglichkeit der Daten,

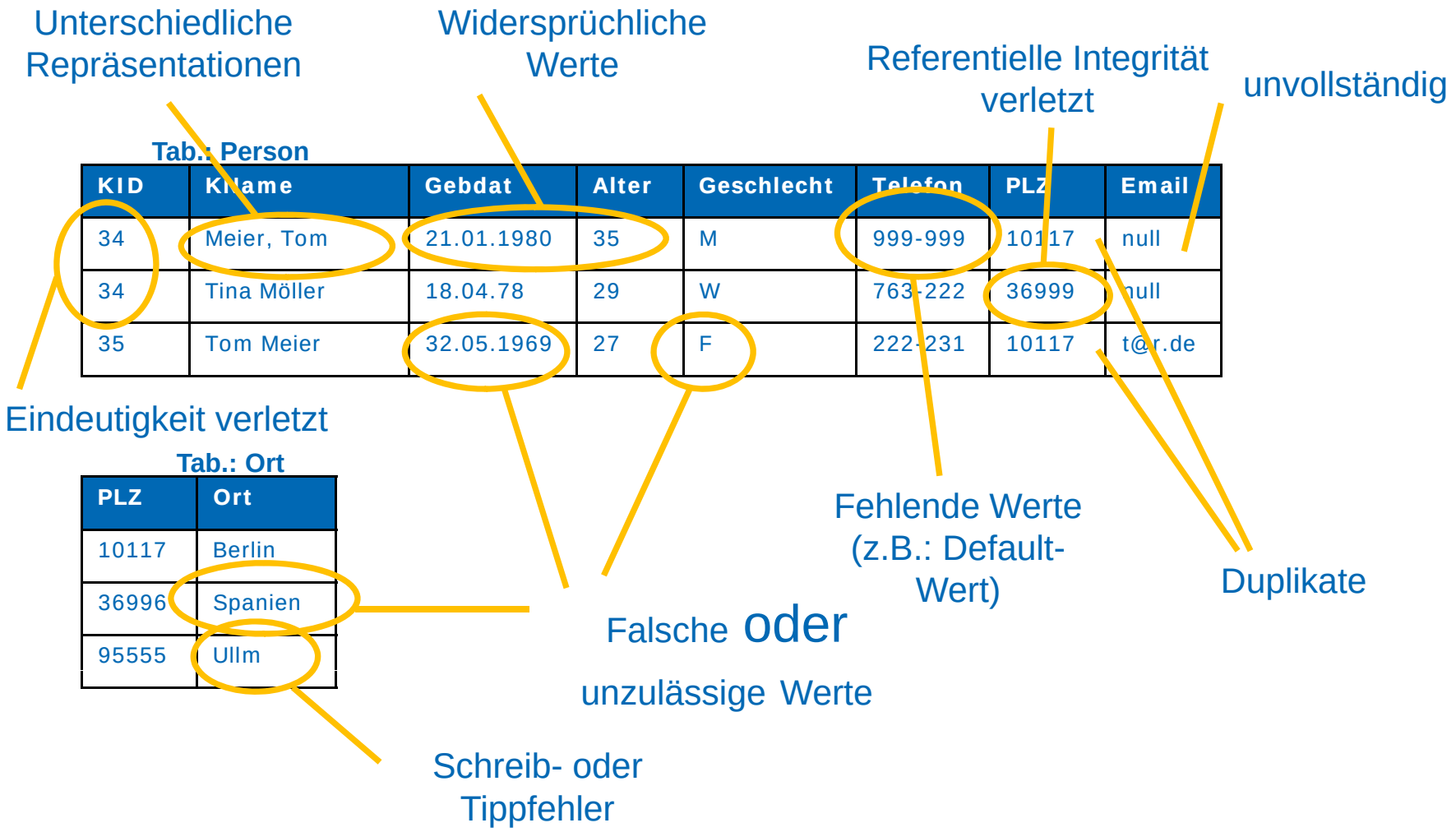
Qualität = **Eignung für den Zweck Datenfitness**

Aktualität von Daten für Bilanzen, Analyse des Kundenverhaltens

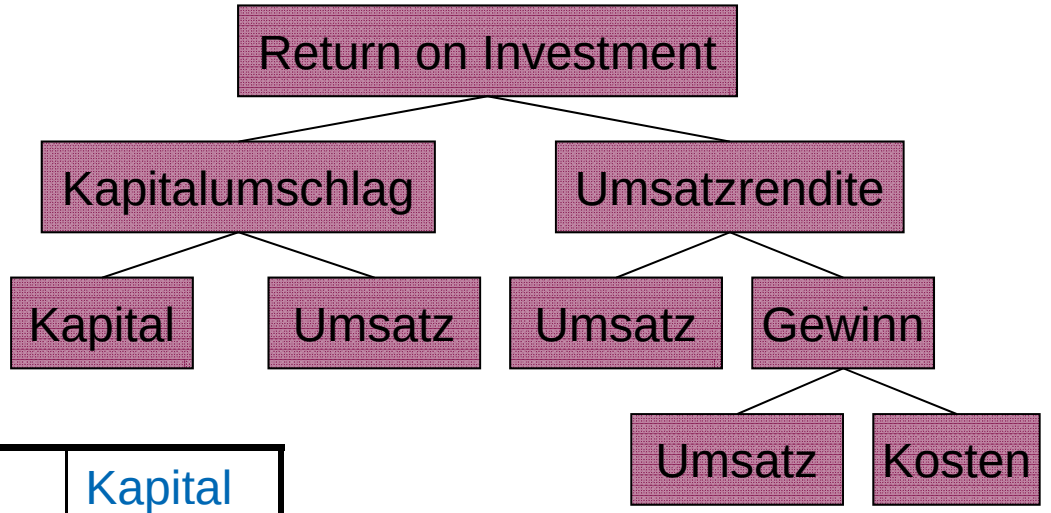
Definition von Eigenschaften von Daten (Qualitätsmerkmale)

Qualität eines Datenproduktes bestimmt durch die Gesamtheit der innewohnenden Merkmale

Motivation: Datenqualität



Motivation: Datenqualität
Ein Kennzahlenbeispiel



Jahresbericht

Umsatz	Gewinn	Kosten	ROI	Kapital
55	10	45	10	60
± 20	± 2	± 20	± 5	± 1

Gewinn = Umsatz - Kosten
Umsatzrendite = Gewinn / Umsatz
Kapitalumschlag = Kapital / Umsatz
ROI = Umsatzrendite / Kapitalumschlag

Fokus Implementierung: H2DB

- 100% JAVA-Datenbank inkl. JDBC API
- Open Source, kleiner Footprint
- In-Memory und Disk-Based DB
- Für Embedded- und Serversysteme
- SQL-Support
- Transaktionsverwaltung und Sicherheitsmechanismen vorhanden
- OBDC über PostreSQL
- Umfangreiche Dokumentation

Fokus Implementierung

Funktionsumfang H2DB

- Grundlegende DB-Funktionalität vorhanden
- Sehr großer Teil des relationalen Teil vom SQL-Standards umgesetzt (DDL, DML, etc.)
- Backup, Locking, Optimierung aus SQL möglich (Grundlagen)
- Wichtigste Datentypen bereits implementiert (int-Varianten, (var-)char, date, CLOB/BLOB,...)
- Aggregate, Numerische-, String-, Zeitfunktionen bereits vorhanden

Themengebiete: Implementierung

Welche Indexstrukturen gibt es in H2DB und wie werden diese genutzt?

- Finden und Abhängigkeiten analysieren
- Herauslösen zu Komponente

Optimierungs- bzw. Kostenfunktion

- DQ-Performance
- Optimierungslösungen

Objektidentifikation

- Unsicherheitsbetrachtung
- Varianten und Möglichkeiten

Themengebiete: Object Identification

Überblick Datentypen und Objektidentifikation

Prozess der Objektidentifikation

- Vorverarbeitung
- Suchraumeingrenzung
- Vergleich

Agency	Identifier	Name	Type of activity	Address	City
Agency 1	CNCBTB765SDV	Meat production of John Ngombo	Retail of bovine and ovine meats	35 Niagara Street	New York
Agency 2	0111232223	John Ngombo canned meat production	Grocer's shop, beverages	9 Rome Street	Albany
Agency 3	CND8TB76SSDV	Meat production in New York state of John Ngombo	Butcher	4, Garibaldi Square	Long Island

Quelle: Hinrichs 2002, S. 97

Empirische Techniken

- Sorted Neighborhood
- Priority Queue

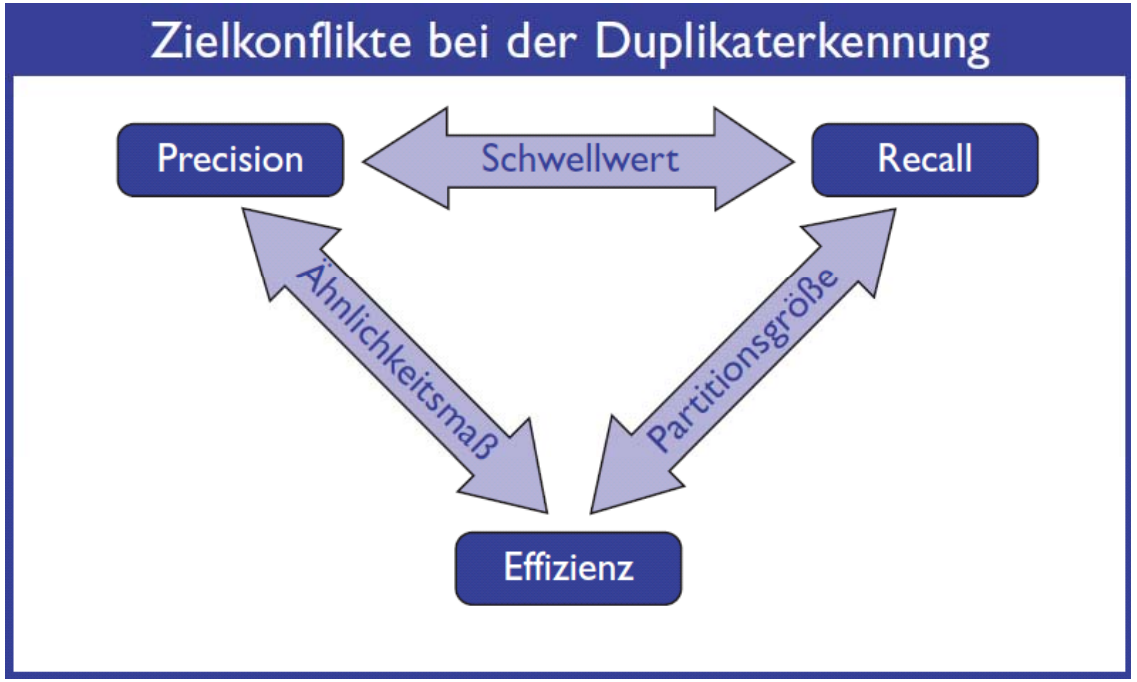
Themengebiete: Edits

- Statistische Data Edits
- Data Cleaning und Data Imputation
- Fellegi Holt (1976) (Statistical Edits), viele Erweiterungen
- Numerische Zusammenhänge

Beispiel-Papiere:

- William E. Winkler (2000) STATE OF STATISTICAL DATA EDITING AND CURRENT RESEARCH PROBLEMS. TR
- Fellegi, I. P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. J. Amer. Statist. Assoc., 71, 17-35.

Themengebiete: Duplikaterkennung



- Ähnlichkeitsmaße
- Suchraumalgorithmen
- Phonetische Algorithmen (Phonetischer Code)

Vortrag

- 20 Minuten Vortrag
- 5-10 Minuten Diskussion/Fragen

- Überziehen: Redner wird abgewürgt
- zu früh: mehr Fragen (ggf. mehr Kritik)

- Rechner wird gestellt, vor Veranstaltung Präsentationen testen und bereitstellen!

Präsentationsrichtlinien

- Kenne Dein(e) Publikum/Zielgruppe

- Rede zum Publikum
- Rede laut und deutlich (langsam)
- Verstecke Dich nicht
- Achte auf Augenkontakt
- Lese nicht vor oder ab
- Übe Dein Timing

- Kenne Dein(e) Publikum/Zielgruppe

Vortrag: Struktur

Stelle Dich selbst und Deinen Background vor (falls nötig)

Erkläre das Ziel des Vortrages frühzeitig

Motiviere Deine Arbeit

Hintergrundwissen soweit wie nötig

Hauptteil: Cohesion! – wichtigsten Ergebnisse – überspringe Details

Zusammenfassung

- Fasse die Hauptpunkte zusammen, Hauptaussage (take-away-message)
- Betone Schlüsse und Konsequenzen

Literatur, wenn in Folien benutzt

Vortrag

20 min, ca. 7 bis 15 Folien

Fontgröße 18, sans-serife Fonts

Name, Titel und Zugehörigkeit auf jeder Folie

Zusammenfassung

- Fasse die Hauptpunkte zusammen, Hauptaussage
- Betone Schlüsse und Konsequenzen auf jeder Folie
- Foliennummern auf jeder Folie
- Nur ein Thema pro Folie
- Farben und Visualisierungen nur wo/wenn nötig
- Vermeide übervolle Folien (> 7 Objekte oder > 36 Wörter)
- Vermeide ganze Sätze, stattdessen fasse Inhalt zusammen (benutze Schlag-/Stichwörter)

Literatur, wenn in Folien benutzt

Warum ein Papier schreiben?

Bekanntgeben von neuen Errungenschaften/Erfahrungen

- Publizieren = Ultimatives Ergebnis wissenschaftlicher Arbeit
- Forschung ist nie beendet, solange sie nicht publiziert wurde

Andere (z.B. Community) über die eigene Arbeit informieren

- Anerkennung/Beachtung
- Kontakte, wertvolle Zusammen-/Mitarbeit

Bekomme Feedback

- extern, unabhängig, anonym

Was gehört in ein Papier?

Man schreibt für den Leser (insb. Gutachter)!

Kommunikation zwischen dem Leser und einem selbst

Bedenke den Hintergrund/das Wissen der Leser

Habe immer die Evaluierungskriterien in Gedanken:

- Orginaler Beitrag
- Signifikantes Problem
- Signifikante Lösung
- Aussagekräftige (robuste) Ergebnisse
- Hochqualitative Präsentation

Aufbau eines Papiers

- Titel
- Zusammenfassung
- Einleitung/ -führung
- Verwandte Arbeiten (auch nach Diskussion üblich/möglich)
- Eigene Arbeit (evtl. mehrere (Unter-)Kapitel)
- Evaluierung
- Diskussion
- Schlussfolgerung und Ausblick
- Referenzen/Literatur

Stil

- Cohesion (Zusammenhang)
- Roter Faden/Gedankenfluß
- In sich geschlossen
- Say what you’re saying before saying it
- Vermeidung bloßer Beschreibungen

Don‘t

- Fehlende Motivation
- Unklare Ziele, unklarer Beitrag
- Fehlende Begründung
- Endlose Diskussionen, ungenutzter Background
- Fehlende Cohesion
- Das große Bild fehlt (einfach nur Details)
- Fehlende Schlussfolgerung oder Ergebnisse
- Umgangssprache, fehlendes (Hintergrund-)Wissen
- Fehlende verwandte Arbeiten
- Max. Seiten (hier 15) überschreiten, aber auch nicht nur Hälfte nutzen
- Nicht an Vorlage halten

Themenvergabe

Thema	Bearbeiter	Thema	Bearbeiter

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

Danke für die Aufmerksamkeit

wwwiti.cs.uni-magdeburg.de/iti_db
Dr. V. Köppen: vkoeppen@ovgu.de
A. Lübcke: andreas.luebcke@ovgu.de

www.ovgu.de

Literatur

Hinrichs, H.: *Datenqualitätsmanagement in Data-Warehouse-Systemen,* Universität Oldenburg, Diss., 2002.

BACKUP

Themengebiete: Data Quality Dimensions

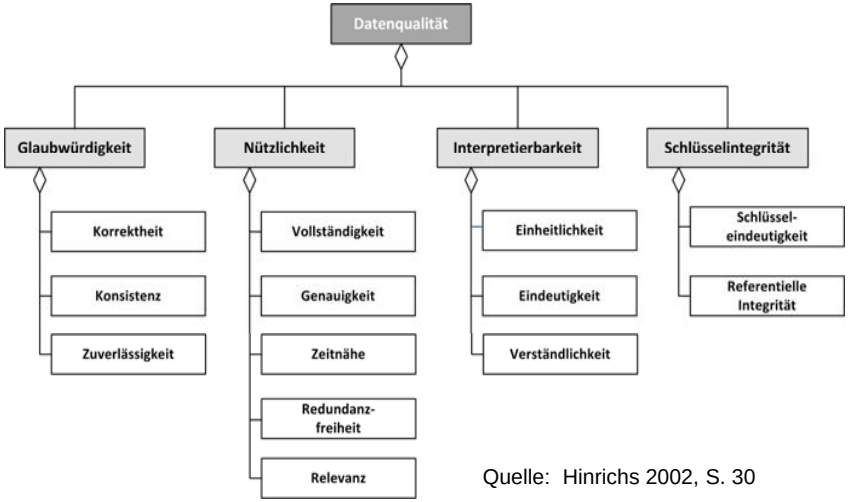
Überblick Datenqualitätsdimensionen

- Klassifikation

Dimensionen von Datenqualität

- Glaubwürdigkeit,
- Nützlichkeit,
- Interpretierbarkeit,
- Integrität
- Zeitbasiert
- Konsistenz
- Schema bedingte Dimensionen

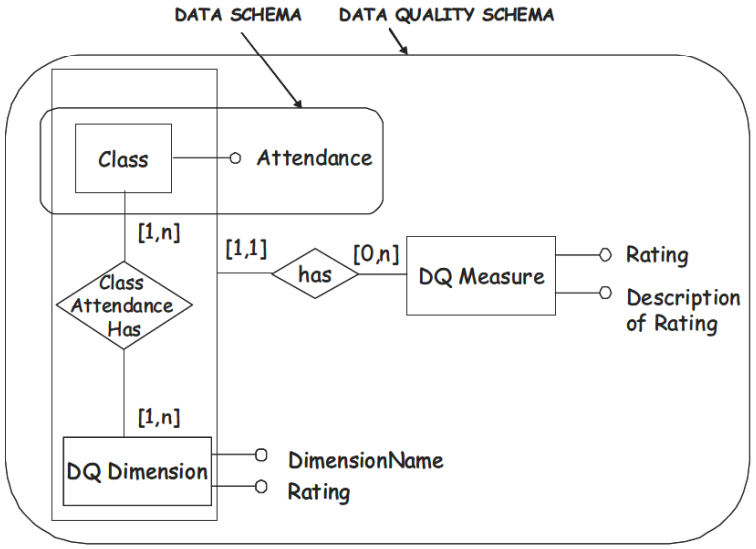
Ansätze zur Definition der Dimensionen



Quelle: Hinrichs 2002, S. 30

Themengebiete: Data Quality Models

Überblick Modelle der Datenqualität



Quelle: Hinrichs 2002, S. 30

Strukturierte Datenmodelle

- Konzeptuelle Modelle
- Logische Beschreibungen
- Data Provenance

Semistrukturierte Datenmodelle

Management IS Modelle

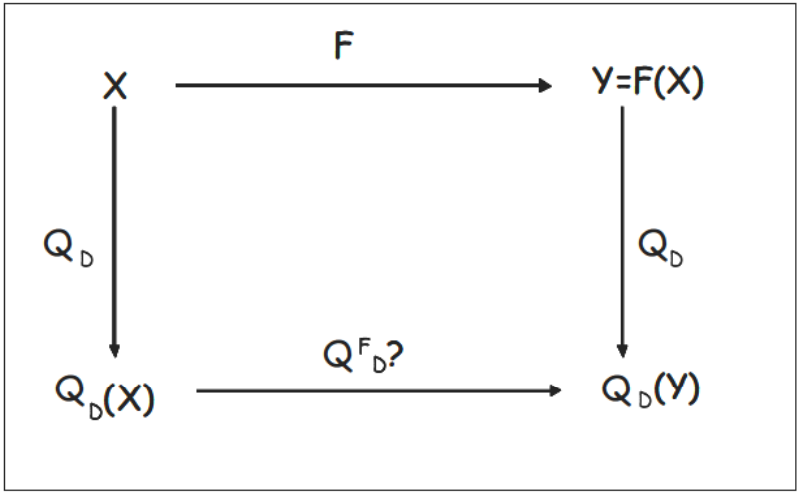
- IP-MAP

Themengebiete: Data Quality Steps

Überblick Aktivitäten und Techniken

Komposition

- Modelle
- Dimensionen
- Genauigkeit und Vollständigkeit



Quelle: Hinrichs 2002, S. 30

Lokalisation und Korrektur

- Inkonsistenzen
- Unvollständige Daten
- Ausreißer