


CAS Information Engineering

Modul: Datenbanken & Data Warehousing

Dozent Manuel Kessler

DWH-4 2024-11-04

Agenda

- 0900 – 0915 Rückblick, Ergänzungen zum Thema data warehousing
- 0915 – 0945 Daten-/Informationsqualität: Einführung
- 0945 – 1015 Grundlagen des «data profiling»

- 1045 – 1200 Aspekte der Duplikaterkennung und -elimination
- 1200 – 1225 Selbständige Übungen
- 1225 – 1230 Ev. letzte Fragen zum Leistungsnachweis, Abschluss



Rückblick – Ergänzungen

Was bisher geschah...

- Multidimensionale Datenmodellierung: Schematypen
 - Sternschema
 - Schneeflockenschema
 - (Fact-Constellation-Schema)
 - Galaxienschema
- Modellierungstypen von Fakten (je nach Auswertungsbedürfnis):
 - Transaction
 - Periodic Snapshot
 - Accumulating Snapshot
- Dimensionen mit Hierarchien

Was bisher geschah...

- Slowly changing dimensions:
 - Typ 1: Überschreiben
 - Typ 2: Historisieren (Gültigkeitsdauern mitführen)
 - Typ 3: Genau einen Vorgängerwert aufbewahren
- Generell gilt: **Strukturelle Änderungen** an Dimensionen (z.B. Zusammenführen unterschiedlicher Produktkataloge, Kontenpläne, ...) stellen grosse Herausforderung an die Analysen und Auswertungen.
- SQL & OLAP:
 - Fensterfunktionen, Partitionierung
 - Erweiterte Möglichkeiten für Gruppierung
 - Statistische Funktionen, Rangfunktionen, ...

Ergänzung: Begriff «data mart»

- Ein Data-Mart ist eine **Kopie** eines Teildatenbestandes eines Data-Warehouse, die für einen bestimmten Organisationsbereich oder eine bestimmte Anwendung oder Analyse erstellt wird. Es kann auch als **Teilansicht** auf das Data-Warehouse oder nicht-persistenter Zwischenspeicher verstanden werden. In der Praxis wird in einigen Fällen der in einem Data-Mart vorhandene Datenbestand auch langfristig vorgehalten (→ «unabhängige Data-Marts»).

[Wikipedia]

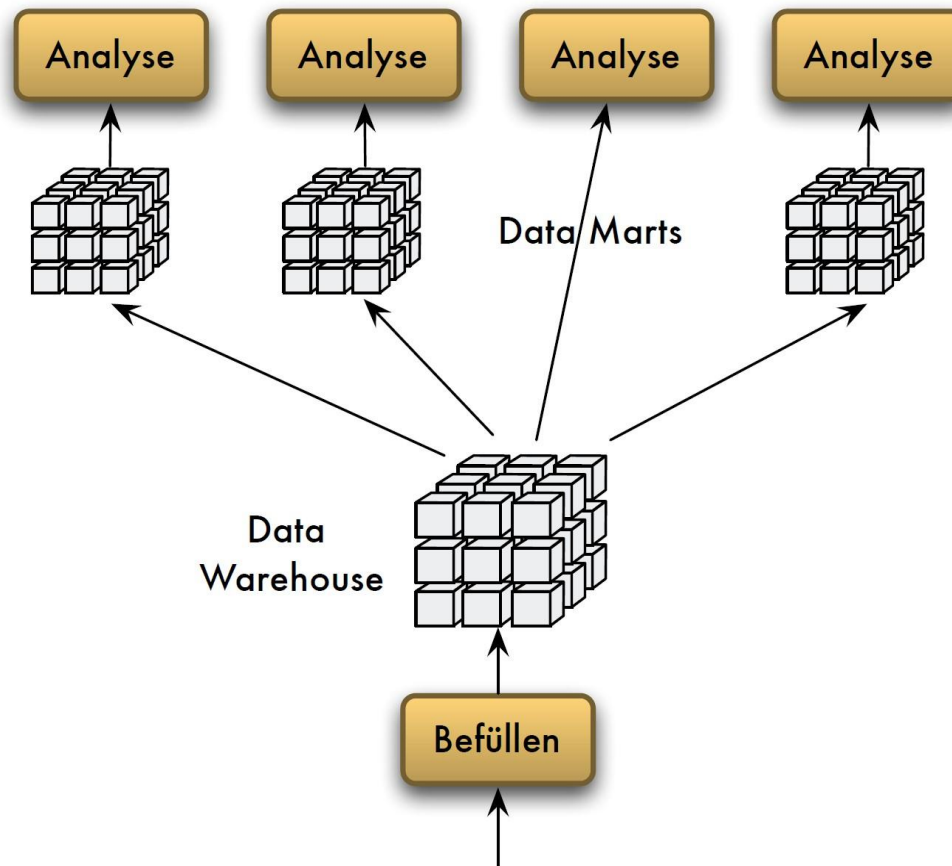
- Aufgabe: Bereitstellung einer inhaltlich **beschränkten** Sicht auf das DWH (z.B. für eine Abteilung).
- Gründe: Eigenständigkeit, Datenschutz, Lastverteilung, Datenvolumen, ...
- Architekturen:
 - Abhängige data marts
 - Unabhängige data marts

Abhängige «data marts»

- Verteilung des Datenbestandes **nach** der Integration und Bereinigung (Basisdatenbank) und Organisation entsprechend den Analysebedürfnissen («Datenwürfel»).
- Data Mart: Nur **Extrakt** (inkl. Aggregation) des Data Warehouse.
- Analysen auf Data Marts **konsistent** zu Analysen auf DWH.
- **Einfache Realisierung**: Replikations- oder Sichtmechanismen des RDBMS.

Abhängige «data marts»

- «Nabe- und Speiche»-Architektur (engl. [hub and spoke](#)):



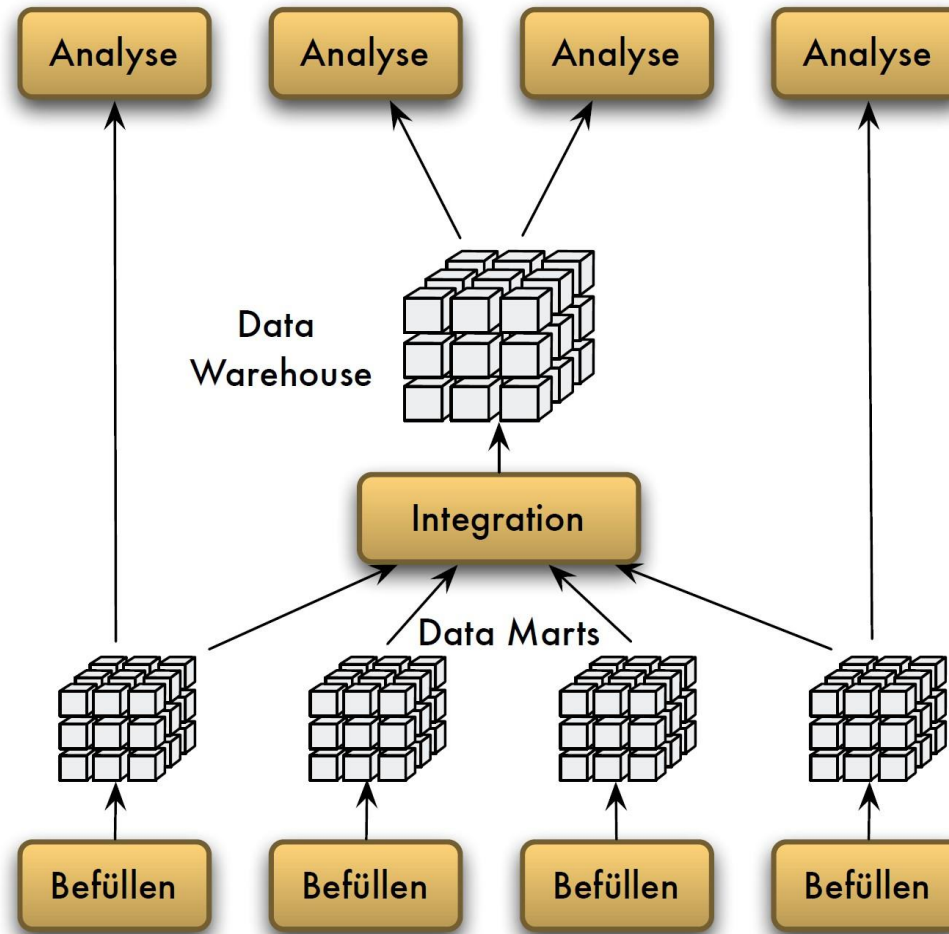
Abhängige «data marts»

- **Strukturelle** Extrakte:
 - Beschränkung auf Teile des Schemas
 - Beispiel: nur bestimmte Kennzahlen oder Dimensionen
- **Inhaltliche** Extrakte:
 - Inhaltliche Beschränkung
 - Beispiel: Nur bestimmte Filialen oder das letzte Jahresergebnis
- **Aggregierte** Extrakte:
 - Verringerung der Granularität
 - Beispiel: Beschränkung auf Monatsergebnisse

Unabhängige «data marts»

- Unabhängig voneinander entstandene «kleine» Data Warehouses (z.B. von einzelnen Organisationseinheiten).
- Nachträgliche Integration und Transformation.
- Probleme:
 - Unterschiedliche Analysesichten (Data Mart, globales Data Warehouse).
 - Konsistenz der Analysen aufgrund zusätzlicher Transformation.

Unabhängige «data marts»



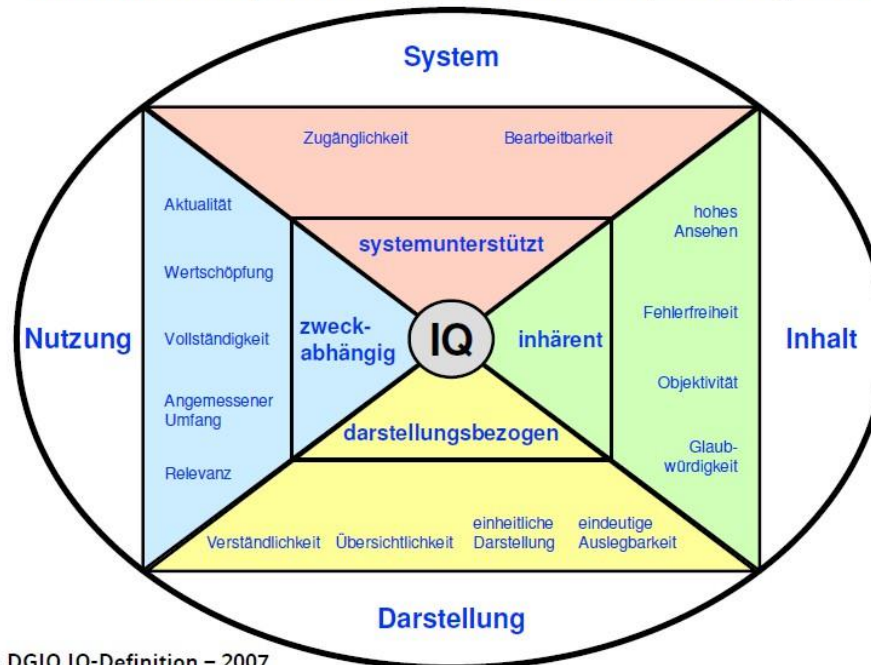


Daten-/Informationsqualität: Einführung & Diskussion

Begriff: Daten-/Informationsqualität

- «**Eignung** von Daten für die **jeweilige** datenverarbeitende **Anwendung**»
- Begriff ist **kontextsensitiv**. Menge von Merkmalen:

Informationsqualität: 15 Dimensionen, 4 Kategorien



Quelle: DGIQ IQ-Definition – 2007

Merkmale

- **Zugänglichkeit (accessibility):**
 - Informationen sind zugänglich, wenn sie anhand einfacher Verfahren und auf direktem Weg für den Anwender abrufbar sind.
 - Beispiel (negativ): Das Kundenstammdaten-System steht dem Back Office aufgrund eines Systemfehlers nicht zur Verfügung. Der Name des Kunden kann nur per telefonischer Rückfrage im Handel erfragt werden.
 - Die angegebenen Beispiele (leicht angepasst) stammen aus folgendem Buch:



ISBN: 978-3-658-09214-6

Merkmale

- Angemessener Umfang (appropriate amount of data):
 - Informationen sind von angemessenem Umfang, wenn die Menge der verfügbaren Information den gestellten Anforderungen genügt.
 - Beispiel (negativ): Für eine Rückfrage bei einem Kunden zu einem Auftrag wird ein Kontaktmanagementsystem aufgerufen. Über die Abfrage mit der Kundennummer oder dem Kundennamen erhält der Mitarbeiter alle bislang erfassten Informationen zum Kunden.

Merkmale

- **Glaubwürdigkeit (believability):**
 - Informationen sind glaubwürdig, wenn Zertifikate einen hohen Qualitätsstandard ausweisen oder die Informationsgewinnung und -verbreitung mit hohem Aufwand betrieben werden.
 - Beispiel (positiv): Eine vom Bundesamt für Statistik herausgegebene Informationsbroschüre zur Bevölkerungsentwicklung besitzt eine hohe Glaubwürdigkeit, und zwar unabhängig davon, inwieweit die Daten vollständig, fehlerfrei, eindeutig auslegbar, objektiv richtig, aktuell und verständlich sind.

Merkmale

- **Vollständigkeit (completeness):**
 - Informationen sind vollständig, wenn sie nicht fehlen und zu den festgelegten Zeitpunkten in den jeweiligen Prozess-Schritten zur Verfügung stehen.
 - Beispiel (negativ): Obwohl der Vertrieb des Unternehmens nach Bundesländern strukturiert ist, lässt sich für ein Bundesland kein Vertriebsbeauftragter ermitteln.

Merkmale

- Übersichtlichkeit (concise representation):
 - Informationen sind übersichtlich, wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.
 - Beispiel (positiv): Währungs-Wechselkursentwicklung der letzten drei Monate. Übersichtliche Darstellung als Candle-Stick-Chart (dadurch auf einen Blick: Eröffnungskurs, Schlusskurs, Höchst- und Niedrigstkurs sowie Richtung zwischen Eröffnungs- und Schlusskurs je Handelstag).

Merkmale

- Einheitliche Darstellung (consistent representation):
 - Informationen sind einheitlich dargestellt, wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden.
 - Beispiel (negativ): Darstellung des Geschlechts einer Person.
Wertemenge: w, f, m; wobei w = weiblich und f = female

Merkmale

- **Bearbeitbarkeit (ease of manipulation):**
 - Informationen sind leicht bearbeitbar, wenn sie leicht zu ändern und für unterschiedliche Zwecke zu verwenden sind.
 - Beispiel (positiv): Die E-Mail-Adresse bei einem online-Bestellvorgang ist als mailto:-Link angegeben. Es kann wahlweise der Link angeklickt und der E-Mail-Client gestartet oder die E-Mail-Adresse kann kopiert und an anderer Stelle eingefügt werden.

Merkmale

- Fehlerfreiheit (free of error):
 - Informationen sind fehlerfrei, wenn sie mit der Realität übereinstimmen.
 - Beispiel (negativ): Eingabefehler in einem Warenwirtschaftssystem führen zu Abweichungen vom tatsächlichen Warenbestand, was zu Lieferengpässen wegen fehlender Nachbestellung führt.

Merkmale

- **Eindeutige Auslegbarkeit (interpretability):**
 - Informationen sind eindeutig auslegbar, wenn sie in gleicher, fachlich korrekter Art und Weise begriffen werden.
 - Beispiel (negativ): Die Erfassung der in einem Aufgabenbereich eingesetzten Arbeitszeit erfolgt in Tagen, wobei nicht definiert ist, ob es sich um die kalendarische Dauer von Beginn bis Ende der Arbeit oder die netto eingesetzten Arbeitstage handelt.

Merkmale

- **Objektivität (objectivity):**
 - Informationen sind objektiv, wenn sie streng sachlich und wertfrei sind.
 - Beispiel (positiv): Für eine Wetterprognose werden die Wetterdaten der letzten Jahre an einem bestimmten Standpunkt benötigt. Eine Übersicht enthält nur die Temperaturangaben (d.h. ohne Zusätze wie „gutes Wetter“ oder „schlechtes Wetter“).

Merkmale

- **Relevanz (relevancy):**
 - Informationen sind relevant, wenn sie für den Anwender notwendige Informationen liefern.
 - Beispiel (negativ): Personendaten, die in CRM-Systemen als Kontakt zu Firmenkunden dienen. Sehr geringe bzw. keine Relevanz hat die Augenfarbe „graugrün“.

Merkmale

- Hohes Ansehen (reputation):
 - Informationen sind hoch angesehen, wenn die Informationsquelle, das Transportmedium und das verarbeitenden System im Ruf einer hohen Vertrauenswürdigkeit und Kompetenz stehen.
 - Beispiel (positiv): Die aus einer Call-Center-Applikation übernommenen Telefonnummern genießen ein hohes Ansehen, wenn die Erfahrung gemacht wurde, dass in dem System nur Telefonnummern gespeichert werden, unter denen die entsprechende Person erreicht wurde.

Merkmale

- **Aktualität (timeliness):**
 - Informationen sind aktuell, wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden.
 - Beispiel (positiv): Währungswechselkurse werde in einem Händlerinformationssystem alle 1–3 Sekunden an die geänderten Marktdaten angepasst. Dies erlaubt die Nutzung der Wechselkursinformation für kurzfristige Kauf- oder Angebotsentscheidungen.

Merkmale

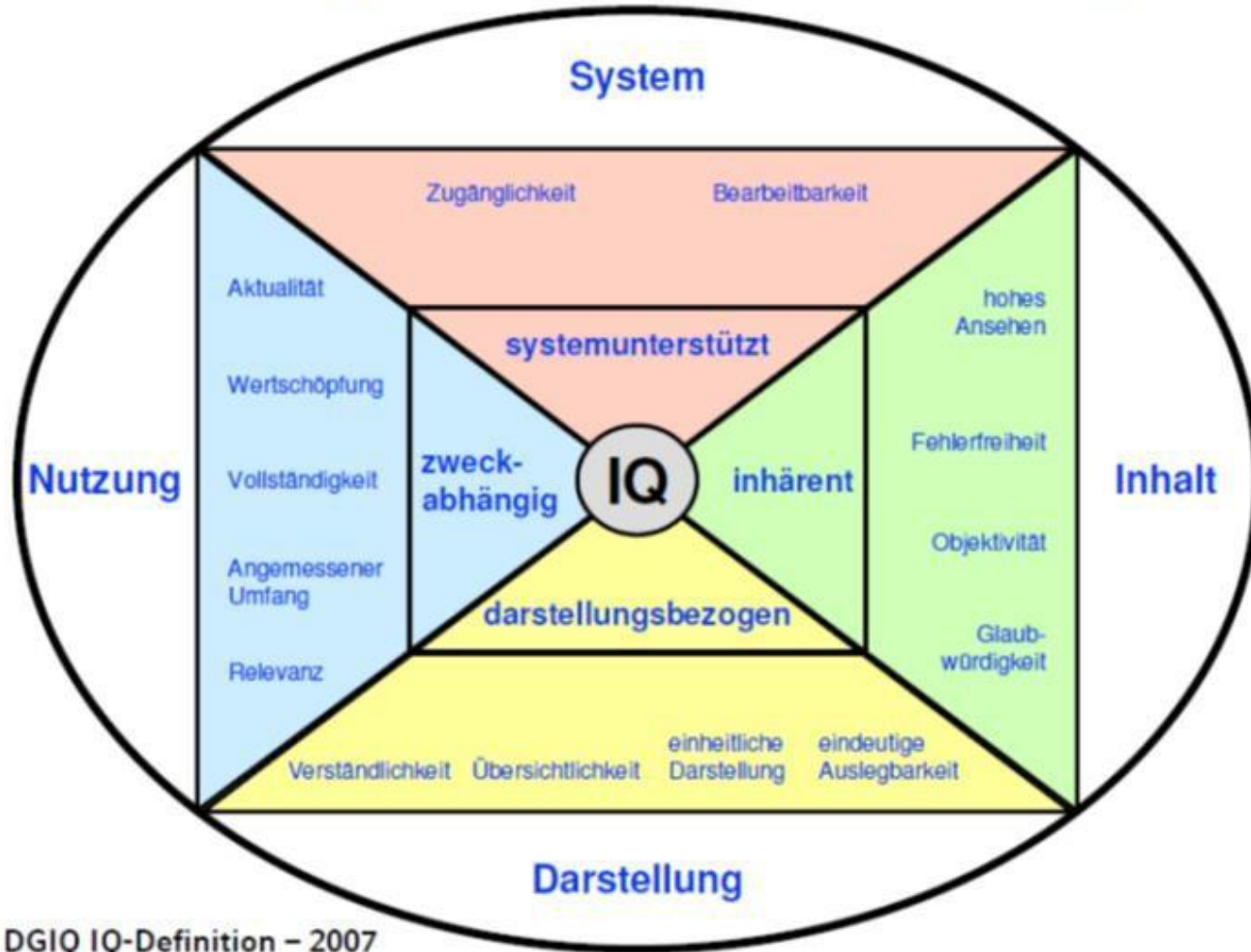
- Verständlichkeit (understandability):
 - Informationen sind verständlich, wenn sie unmittelbar von den Anwendern verstanden und für deren Zwecke eingesetzt werden können.
 - Beispiel (negativ): Der Wohnort eines Kunden, an den Ware gesendet werden soll, ist als GPS-Koordinate „642.85/156.50“ erfasst.

Merkmale

- Wertschöpfung (value-added):
 - Informationen sind wertschöpfend, wenn ihre Nutzung zu einer quantifizierbaren Steigerung einer monetären Zielfunktion führen kann.
 - Beispiel (positiv): Angaben zu Personen, die potenzielle Kunden sind, in Bezug auf die Zielfunktion Umsatz: Der Nachname hat eine hohe Wertschöpfung, da durch die personalisierte Ansprache bei Direktmarketing der Erfolg (z. B. Bestellwahrscheinlichkeit, Umsatz) deutlich gesteigert werden kann.

Daten-/Informationsqualität

Informationsqualität: 15 Dimensionen, 4 Kategorien



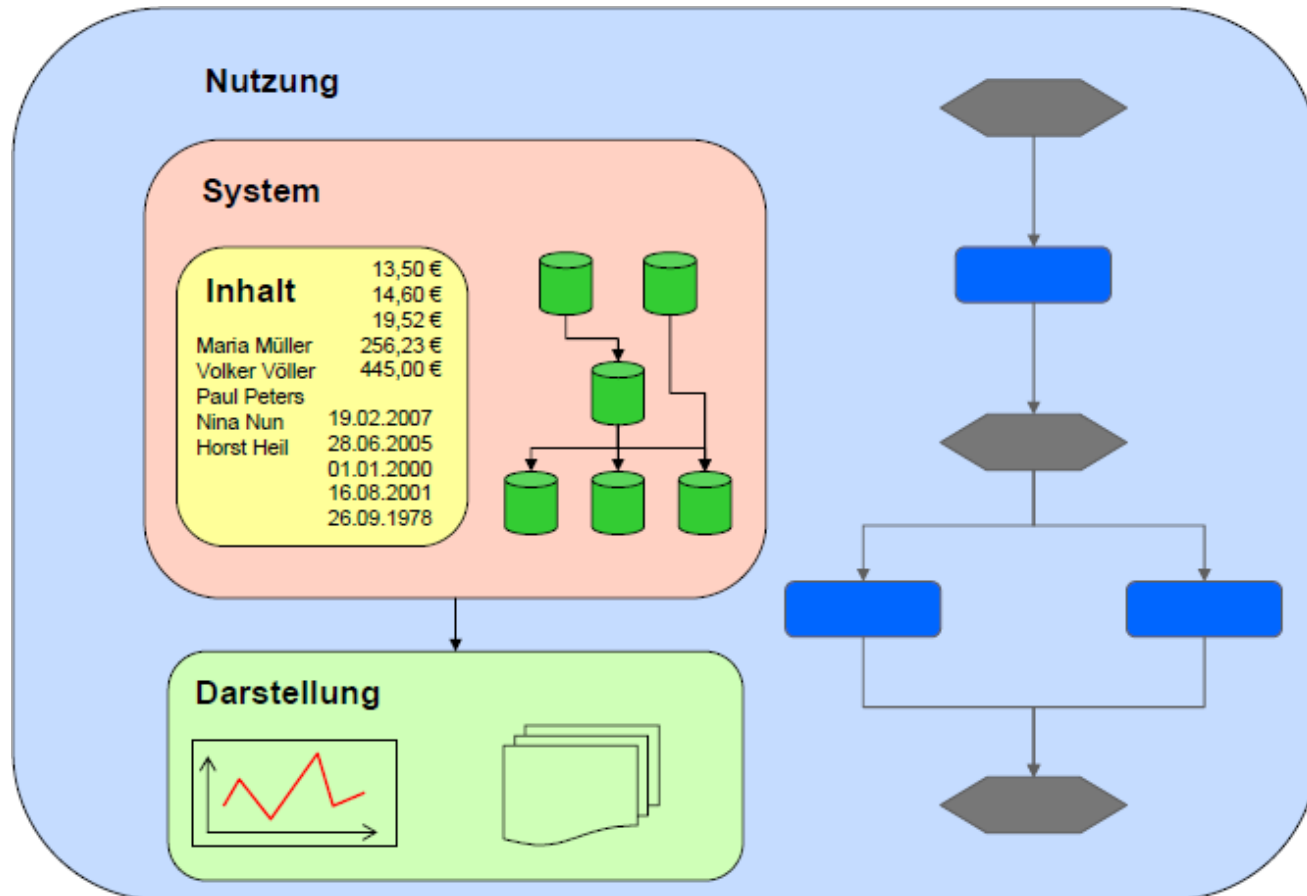
Quelle: DGIQ IQ-Definition – 2007

Daten-/Informationsqualität

- Allen Ordnungsbegriffen ist gemeinsam: Sie führen zu einem bestimmten **Untersuchungsgegenstand**, der der Beurteilung der Qualität dient.
- Betrachtet werden:
 - Inhalt = Datenfeldinhalt
 - System = Datenverarbeitendes System, Oberfläche
 - Darstellung = Ergebnisse (Reports, Statistiken etc.)
 - Nutzung = Kontext der Daten / Prozesse
- Darstellung der gegenseitigen Abhängigkeit und Beeinflussung.
- Alle Untersuchungsgegenstände sind a priori **gleich wichtig**. Eine Gewichtung kann nur durch den **Anwender** vorgenommen werden.

Daten-/Informationsqualität

- IQ-Definitionen: Untersuchungsgegenstände



Daten-/Informationsqualität

- In der Regel geht es um ein **Optimieren** von mehreren Merkmalen.
- Man kann nur verbessern was man **messen** kann, d.h. wichtige Merkmale müssen **quantitativ** erfasst werden.
- Es bestehen dann grundsätzlich zwei Varianten:
 1. **Akzeptieren** und bewusster Umgang mit der vorhandenen Qualität (z.B. bei extern bezogenen Daten).
 2. Erkennen von Problembereichen und gezieltes **Verbessern**:
 - Erfassungsfehler korrigieren.
 - Dubletten erkennen und fusionieren.
 - Abgleichen mit Referenzdaten.
 - ...



Grundlagen des «data profiling»

Data Profiling – Begriff

- Data Profiling bezeichnet den ~~weitgehend automatisierten~~ Prozess zur **Analyse** vorhandener Datenbestände durch unterschiedliche Analysetechniken. Durch das Data Profiling werden die **existierenden Metadaten** zu den Echtdaten **validiert** und **neue Metadaten** identifiziert.
- Data profiling is the process of examining the data available in an existing data source (e.g. a database or a file) and **collecting statistics** and information about that data.

[Wikipedia]

- Data profiling refers to the activity of creating small but informative **summaries of a database**.

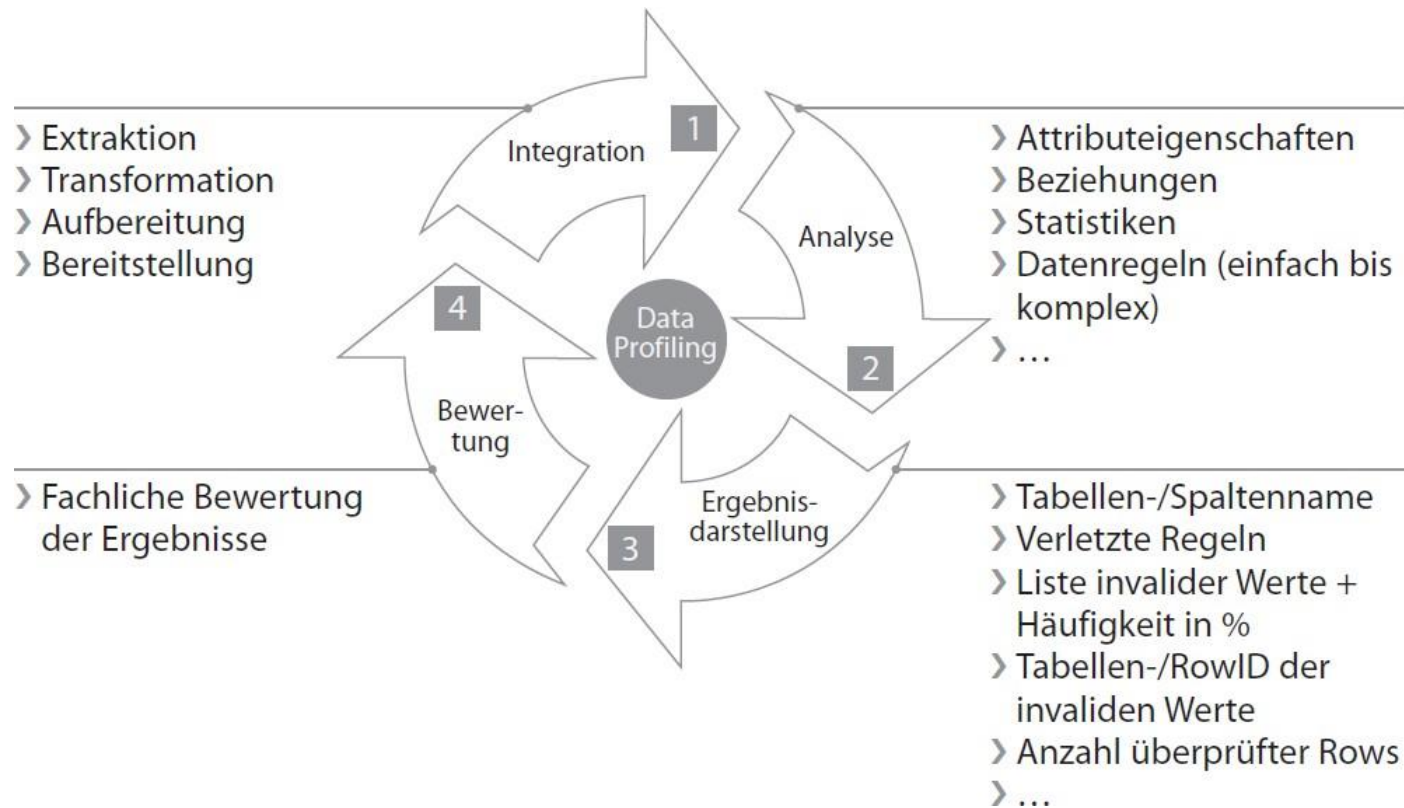
[Ted Johnson, Encyclopedia of Database Systems]

Data Profiling – Anwendungsfälle

- Data cleansing (data scrubbing, data cleaning, ...):
 - Erkennen von Mustern, Duplikaten, ...
- Query-Optimierung:
 - Statistiken, Mengengerüste, Histogramme, ...
- Datenintegrationen (ETL, ...), Datenübernahmen:
 - Verständnis über die Ausgangsdaten gewinnen, («DB-reverse engineering»)
 - Erkennen von Abhängigkeiten, Überlappungen, Duplikaten, ...
- Data analytics:
 - **Verständnis** für die zu analysierenden Daten **gewinnen**.
- ...

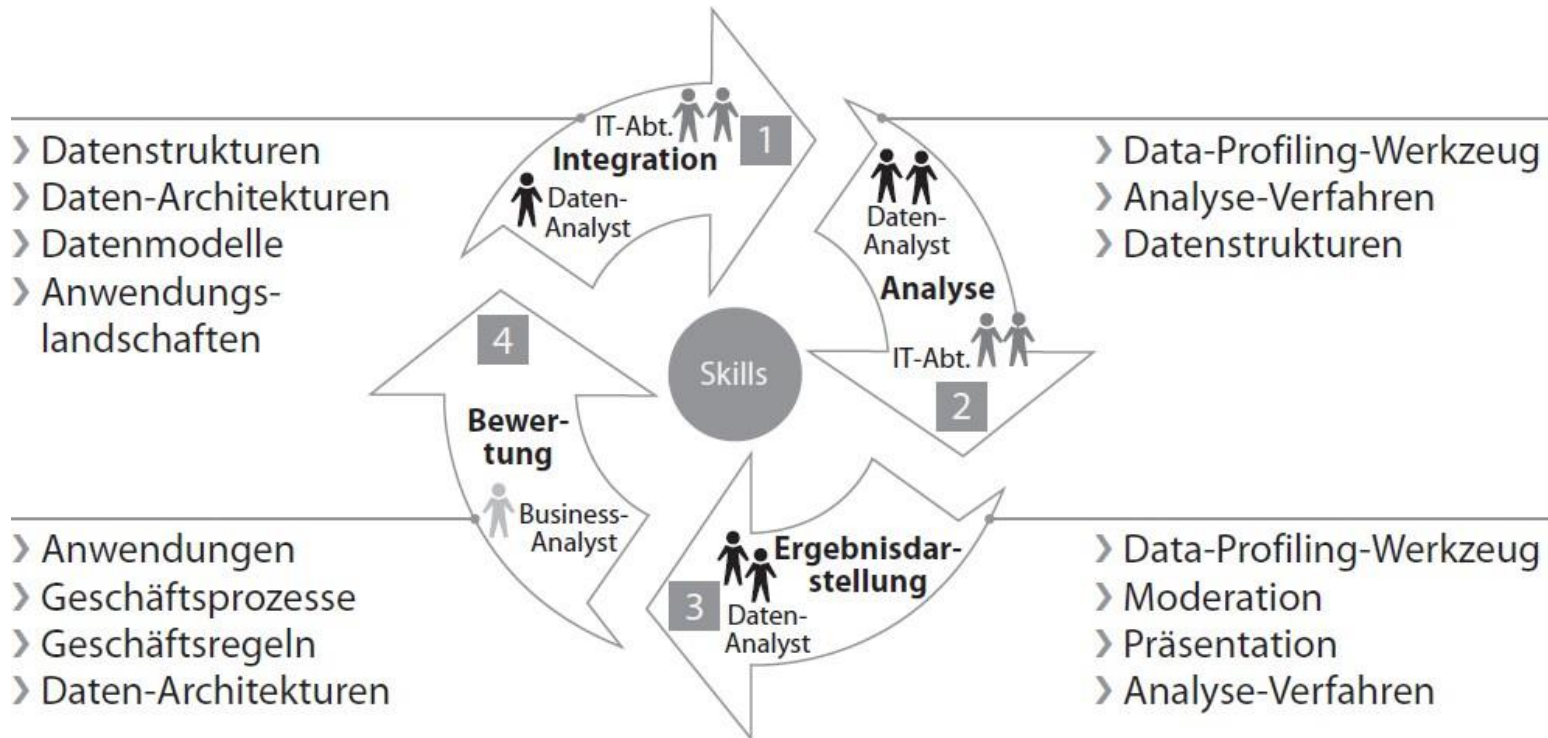
Data Profiling – Prozess

- Iterativer Prozess:



Data Profiling – Prozess

- Zusammensetzung und Skills eines Data-Profiling-Teams:



Data Profiling – Erkenntnisse

- **Werkzeugunterstützte** Analyse der bereitgestellten Daten unter Anwendung verschiedener Data-Profiling-Verfahren.
- Hochgradig **iteratives** und **interaktives** Vorgehen:
 - Wahl der zu analysierenden Daten und von geeigneten Verfahren
 - Analyse der Daten
 - Darstellung, Dokumentation, Interpretation der Ergebnisse
 - Identifikation von Problembereichen
 - Gewinnen von **handlungsrelevanten Erkenntnissen**, Korrekturmassnahmen umsetzen (je nach Art der Fehlerart)

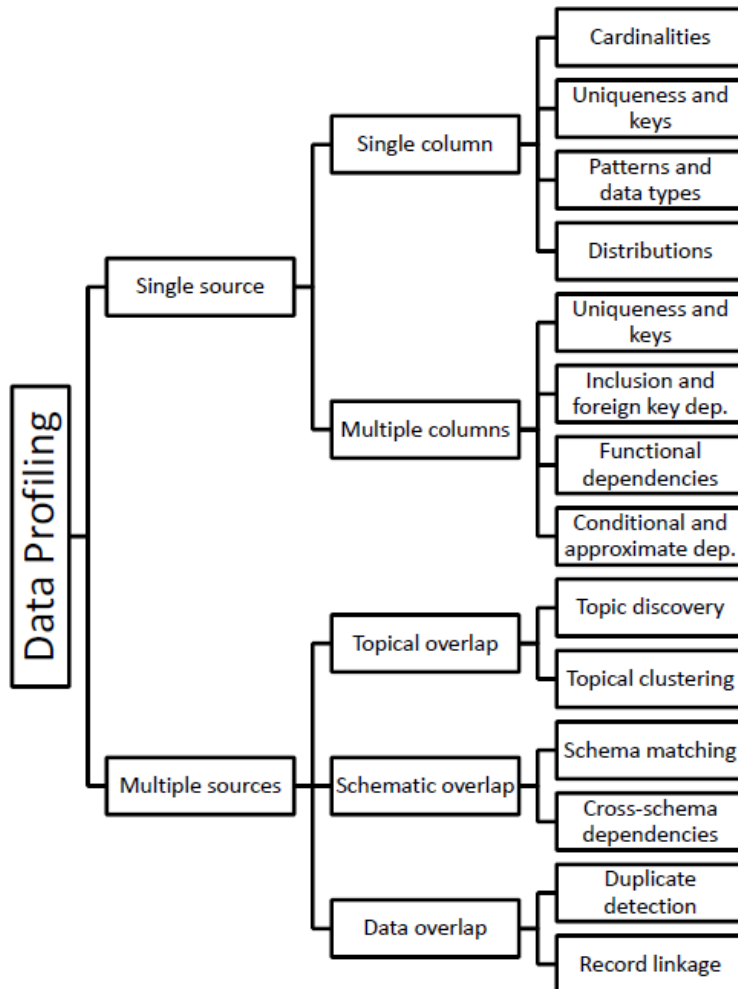
Data Profiling – Fehlerarten

- Fehler 1. Art:
 - Maschinell erkennbar
 - Maschinell behebbar
 - Beispiel: falsch formatierter Datumswert
- Fehler 2. Art:
 - Maschinell erkennbar
 - Nur manuell behebbar
 - Beispiel: fehlendes Geburtsdatum

Data Profiling – Resultate vermitteln

- Die Resultate müssen mit den Fachpersonen der entsprechenden Geschäftsbereiche diskutiert werden. Dabei ist es wichtig, nicht mit IT-Begriffen zu arbeiten, sondern die Probleme in die **Sprache der Anwender** zu übersetzen:
- Falsch:
 - «In der Tabelle B19 fehlen bei 4.3% der Tupel die Fremdschlüsselwerte zur Tabelle K_7»
- Besser:
 - «Es hat sich gezeigt, dass 4.3% aller Bestellpositionen keinem Kunden zugeordnet werden können»
- Die betroffenen Datenwerte sind dabei (zumindest konkrete **Beispiele**) den Anwendern vorzulegen.

Data Profiling – «Messobjekte»



- Achtung:
 - Die meisten Publikationen, Verfahren, Tools etc. basieren auf **strukturierten** Daten.
 - Semi-strukturierte Daten (XML, JSON, ...) müssen ggf. entsprechend umformatiert werden oder die genannten Verfahren sind anzupassen bzw. es sind u.U. ganz andere Verfahren anzuwenden!

Data Profiling – mühsam!

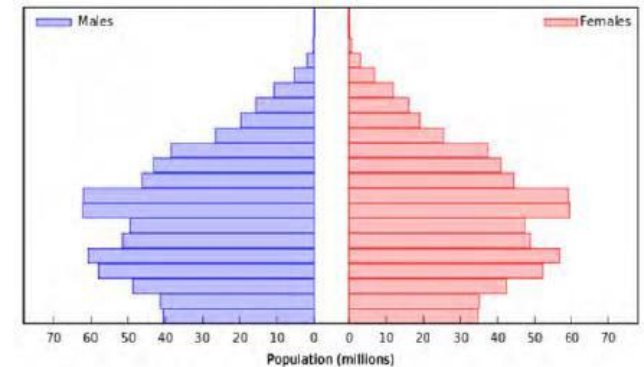
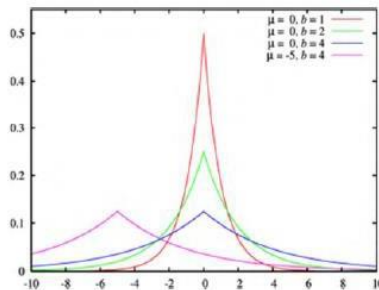
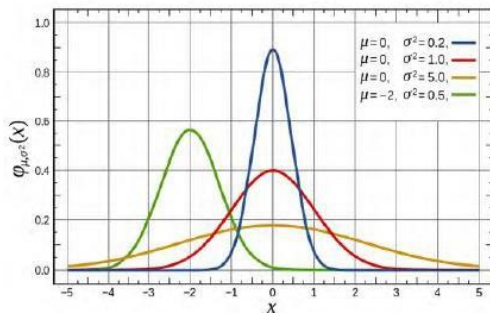
- In aller Regel liefern Einzelanalysen nur ein unvollständiges Bild von der Datenqualität. Sie können aber bereits wichtige **Problembereiche** aufdecken und sind oft notwendig, um beispielsweise notwendige Korrekturen zu erkennen um die Daten überhaupt in ein Data Warehouse laden zu können.
- Komplexere Analysen erfordern ein **vertieftes Wissen** der zu Grunde liegenden **Geschäftsregeln**.
- Achtung: Data Profiling ist ein «mühsames» Geschäft. Sehr oft liegen den Datenproblemen **menschliche Ursachen** zu Grunde. Das Aufzeigen und Diskutieren von Mängeln und deren Behebung benötigt oft viel «Fingerspitzengefühl».

Data Profiling – Verfahren

- Beispiele von Standard-Analysen auf **Attributebene**:
 - Attributnamen → Semantik erkennen (Beizug allfälliger Dokumentationen, Entwicklungsrichtlinien u.a.)
 - Oft lässt sich aus dem Attributnamen nicht erkennen, was der Inhalt ist («Status», «CUSTOM01», «PID», ...)
 - Datentypen (physisch ↔ logisch), Abweichungen erkennen (z.B. falsch benutzte Typen, alphanumerische Zeichen, dort wo nur Zahlen vorkommen sollten, ungültige Kalenderdaten, ...)
 - Muster (z.B. bei Telefonnummern, E-Mail-Adressen, AHV-Nr., IBAN, ...)
 - Zulässige Wertebereiche («Anzahl Kinder = -1», ...)

Data Profiling – Verfahren

- Beispiele von Standard-Analysen auf **Attributebene**:
 - Statistische Angaben:
 - Max-Länge, Min-Länge, Ø-Länge (bei alphanumerischen Daten)
 - NULL-Werte, Spezialwerte, «defaults» («99.99.9999», ...), führende Blanks, ...
 - Max-Wert, Min-Wert, Ø-Wert (bei numerischen Daten)
 - Wahrscheinlichkeitsverteilungen für numerische Werte, Histogramme
 - ...



Data Profiling – Verfahren

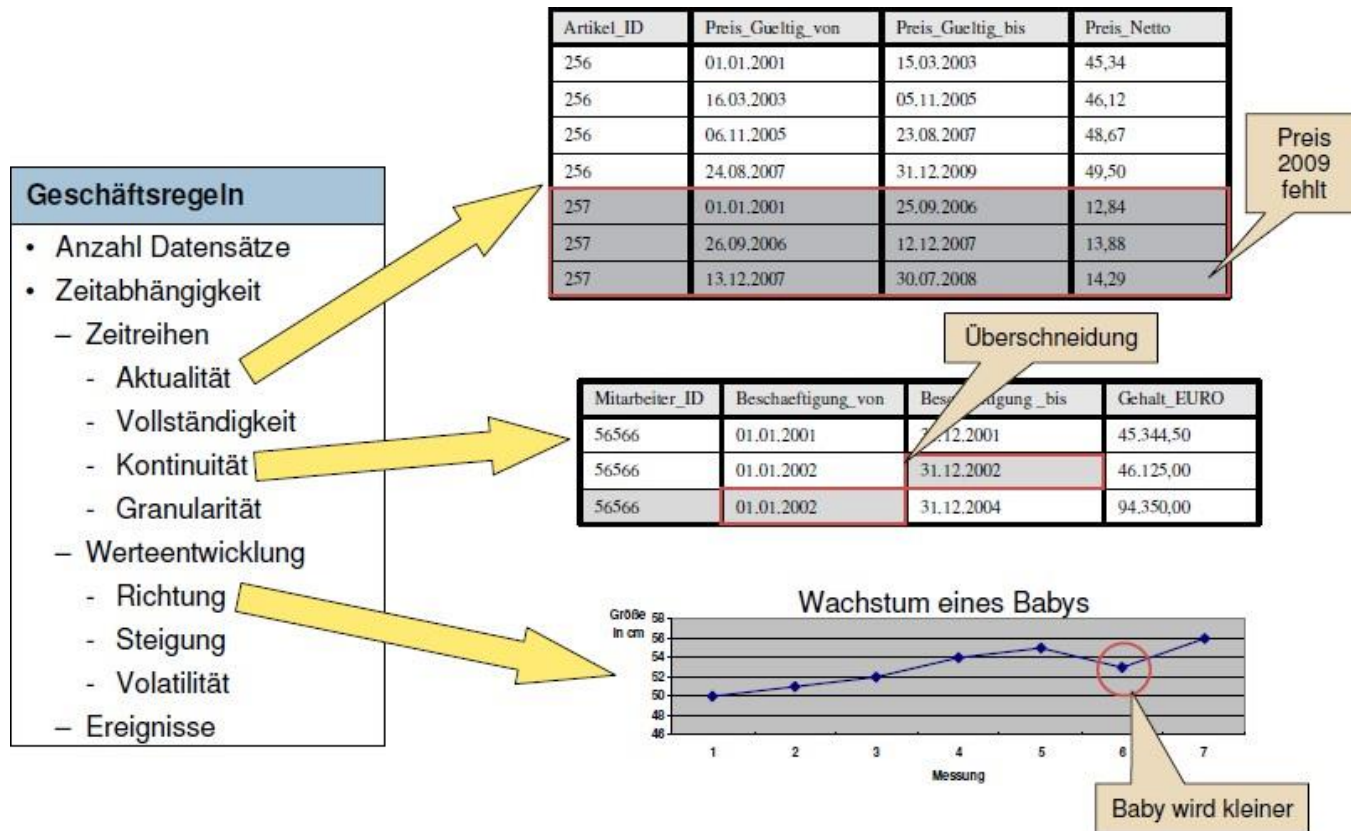
- Beispiele von Analysen auf **Tupelebene**:
 - Eindeutigkeit, Schlüsseleigenschaften:
 - Einzelattribut
 - Attributkombinationen
 - Funktionale Abhängigkeiten:
 - $A \rightarrow B$: Wenn zwei Tupel denselben Wert haben in Attribut A haben sie dann auch denselben Wert in Attribut B?
 - Berechnete Werte (Anzahl * Preis = Umsatz, ...)
 - Übereinstimmungen (PLZ \leftrightarrow Ort)
 - Duplikate

Data Profiling – Verfahren

- Beispiele von Analysen auf **Tabellenebene**:
 - Referentielle Integrität, Schlüsselwerte (Primär- / Fremdschlüsselbeziehungen)
 - Abgleich von Werten mit Referenztabelle
(Währungen, Länderkennzeichen, Artikelcodes, Warengruppen, ...)
 - Kardinalitäten («eine Bestellung darf max. 20 Positionen umfassen», ...)
 - Datenvolumen (ungefähre Anzahl Tupel, erwartetes Wachstum, ...)
 - ...

Data Profiling – Verfahren

- Probleme, die aufgrund von Geschäftsregeln gefunden werden können:



Data Profiling – Kosten / Nutzen

- **Nutzen** solcher Analysen:
 - «Gefühl» entwickeln für die Daten
 - Kategorisierung («was kommt wie oft vor»)
 - Erkennen von Ausreissern
 - Verletzungen von Geschäftsregeln erkennen
 - ...

⇒ Basis für nachfolgende Datenbereinigung («Verbesserung der Datenqualität»)
- **Kosten** der Analysen nicht vernachlässigen!
- Achtung: Manche Verfahren können aus Komplexitätsgründen oft nur Näherungswerte liefern (z.B. funktionale Abhängigkeiten, Inklusionsabhängigkeiten, Schlüssel u.a.)!

Data Profiling – Tools

- Stark wachsender Markt. Oft kombinierte Werkzeuge für data profiling / data cleansing / data quality / ...
- Beispiele:
 - IBM InfoSphere Information Analyzer:
 - <http://www.ibm.com/software/data/infosphere/information-analyzer/>
 - Talend Open Studio for Data Quality:
 - <http://www.talend.com/products/data-quality>
 - Microsoft SQL Server Integration Services Data Profiling Task and Viewer:
 - <http://msdn.microsoft.com/en-us/library/bb895310.aspx>
 -

Data Profiling – Tools

- In der Regel sehr viele features:
 - Num rows
 - Min value length
 - Median value length
 - Max value length
 - Avg value length
 - Precision of numeric values
 - Scale of numeric values
 - Quartiles
 - Basic data types
 - Num distinct values ("cardinality")
 - Percentage null values
 - Data class and data type
 - Uniqueness and constancy
 - Single-column frequency histogram
 - Multi-column frequency histogram
 - Pattern discovery
 - Soundex frequencies
 - Benford Law Frequency
 - Single column primary key discovery
 - Multi-column primary key discovery
 - Single-column FK discovery
 - Multi-column FK discovery
 - ...

Data Profiling – Tools

- Mängel solcher Werkzeuge:
 - Benutzung:
 - Komplex zu konfigurieren
 - Resultate schwierig zu verwalten, zu analysieren und zu interpretieren
 - Skalierung:
 - Hauptspeicherbasiert
 - SQL basiert
 - Effizienz:
 - Kaffeepause, Mittagessen, Nachtlauf, ...?
 - Funktionalität:
 - Oft auf einfache Untersuchungen beschränkt
 - Überprüfen \neq Erkennen

Data Profiling – Bemerkungen

- Die Ausgangslage ist meistens «diffus», man weiss am Anfang nicht genau wonach man sucht.
- Es empfiehlt sich daher so früh wie möglich mit realen Daten erste Untersuchungen anzustellen (d.h. VOR dem Entwickeln einer ETL-Komponente, VOR dem Durchführen einer Migration, ...).
- Solche Untersuchungen generieren in der Regel rasch sehr viele **Einzelergebnisse**, die analysiert, verdichtet und präsentiert bzw. interpretiert werden müssen.
- Man sollte sich auf das Notwendige konzentrieren, z.B. auf Probleme, die ein Laden der Daten in ein Data Warehouse verunmöglichen können.

Data Profiling – Bemerkungen

- Die Analysen müssen **handlungsrelevante** Ergebnisse liefern:
 - **Korrektur** von falschen Daten.
 - Wahrscheinlichkeit **reduzieren**, dass dieselben Probleme erneut auftauchen:
 - Qualitätsbewusstsein schaffen
 - Schulung
 - Qualitätsprozesse etablieren (regelmässiges Fehlerfeedback)
 - ...
 - Wiederauftreten **verhindern**:
 - Anpassen von OLTP-Anwendungen, z.B. durch Implementation notwendiger Integritätsbedingungen oder Eingabeprüfungen
 - Problembehebung im Rahmen eines ETL-Prozesses
 - ...

⇒ Data profiling ist in der Regel nur ein (erster) Schritt!

Data Profiling – Bemerkungen

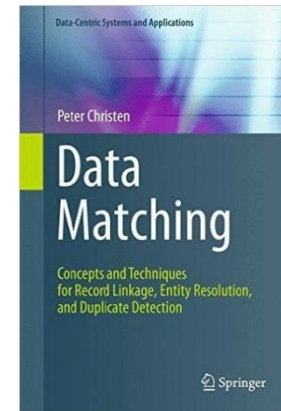
- Werkzeuge sind nützlich, aber nur ein Hilfsmittel.
- Das Verbessern der Datenqualität ist eine **zeitintensive** Daueraufgabe und muss in entsprechende **Prozesse** eingebunden werden!
- Literaturhinweise:



ISBN: 978-3-86490-042-6



ISBN: 978-3-8348-1453-1



ISBN: 978-3-6423-1163-5

F. Naumann: Data Profiling Revisited (zu finden auf Moodle).



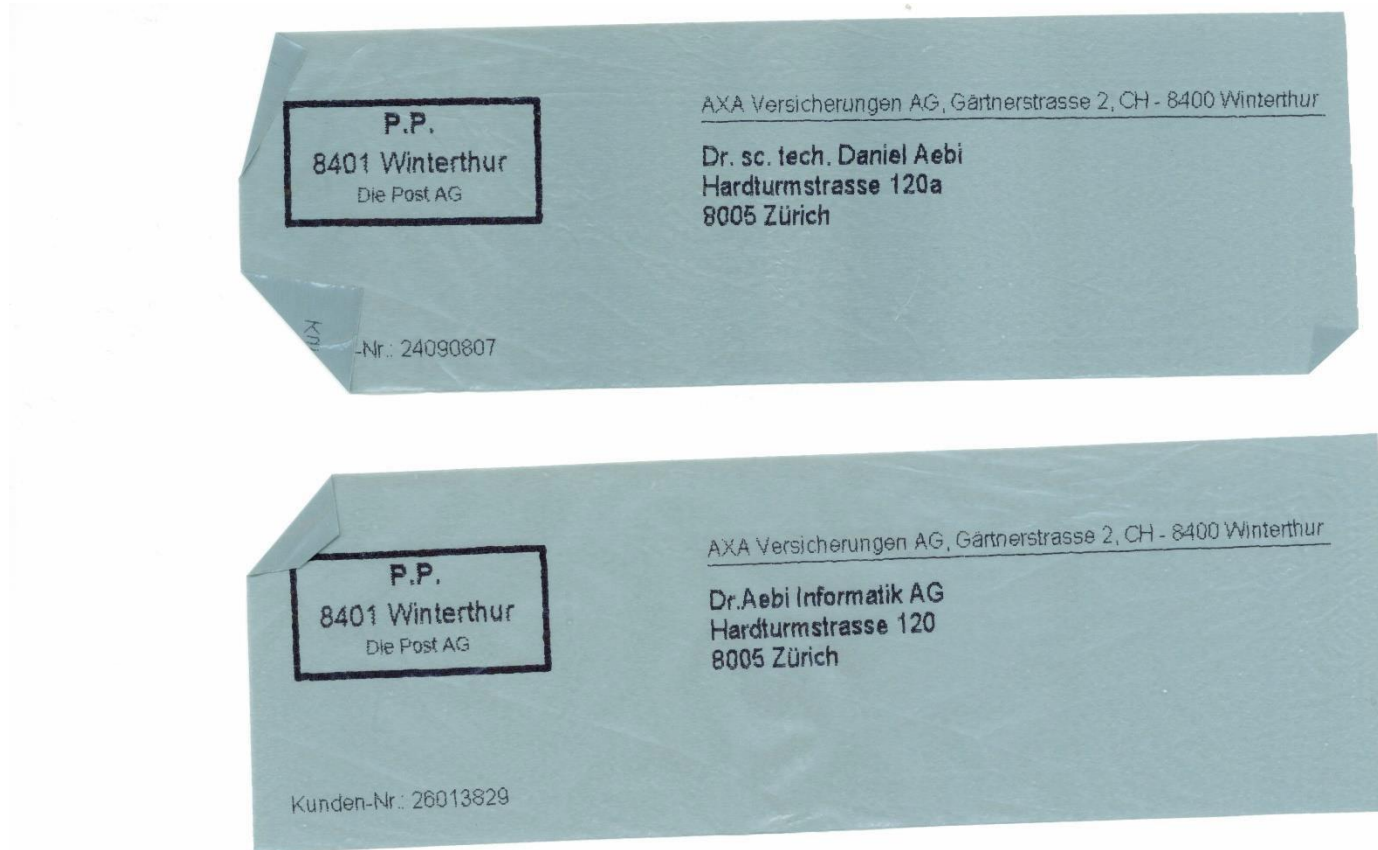
Aspekte der Duplikaterkennung und -elimination

Duplikaterkennung

- Duplikate:
 - Allgemein: Kopien oder Replikate vom Original
 - Bezogen auf Datenbanken: **Mehrere Datensätze** die **dasselbe Realweltobjekt** darstellen.
- Probleme:
 - Die Repräsentationen sind **nicht identisch** (Aufgabe sonst trivial)
 - ⇒ Ähnlichkeitsmasse nötig, Vergleiche auf Attribut und Tupelebene
 - Die **Datenmengen** sind **gross**
 - ⇒ Hoher Aufwand, um Duplikate zu identifizieren

Duplikaterkennung

- Erkennung anspruchsvoll:



Entstehung von Duplikaten

- Bei einer Datenquelle:
 - Mehrfach geführte Kunden im System
 - Verschiedene Repräsentationen desselben Produktes
 - Doppelt gebuchte Bestellungen
 - ...
- Bei mehreren Quellen:
 - Zusammenführung von Daten aus unterschiedlichen Systemen
 - ...
- Häufige Gründe:
 - Tipp-/Erfassungs- und Übertragungsfehler, unterschiedliche Schreibweisen und Aktualitäten «Fuzzy Duplicates»
 - Informationen über Realweltobjekte (Kunden, Mitarbeiter, ...) in verschiedenen Systemen benötigt

Auswirkungen von Duplikaten

- Erhöhte Kosten / Risiken
 - Herstellungs- und Distributionsaufwand für Mailings
 - Kein Mengenrabatt bei Bestellungen gleicher Artikel
 - Kreditlimiten nicht erkennbar
 - ...
- Imageverlust
 - Warum kriege ich den Katalog der Firma XY immer doppelt?
 - ...

Auswirkungen von Duplikaten

- Verfälschte Analysen und Statistiken (Verkauf, Lager, ...)
 - Umsatz für Kunde/Produkt X
 - ...
- Generell
 - Erhöhte (IT-)Kosten
 - Reduzierte Kundenzufriedenheit
 - Erhöhte Risiken
 - ...
- Zwei zu lösende Probleme:
 1. Erkennen von Duplikaten
 2. Eliminieren von Duplikaten (Fusion)

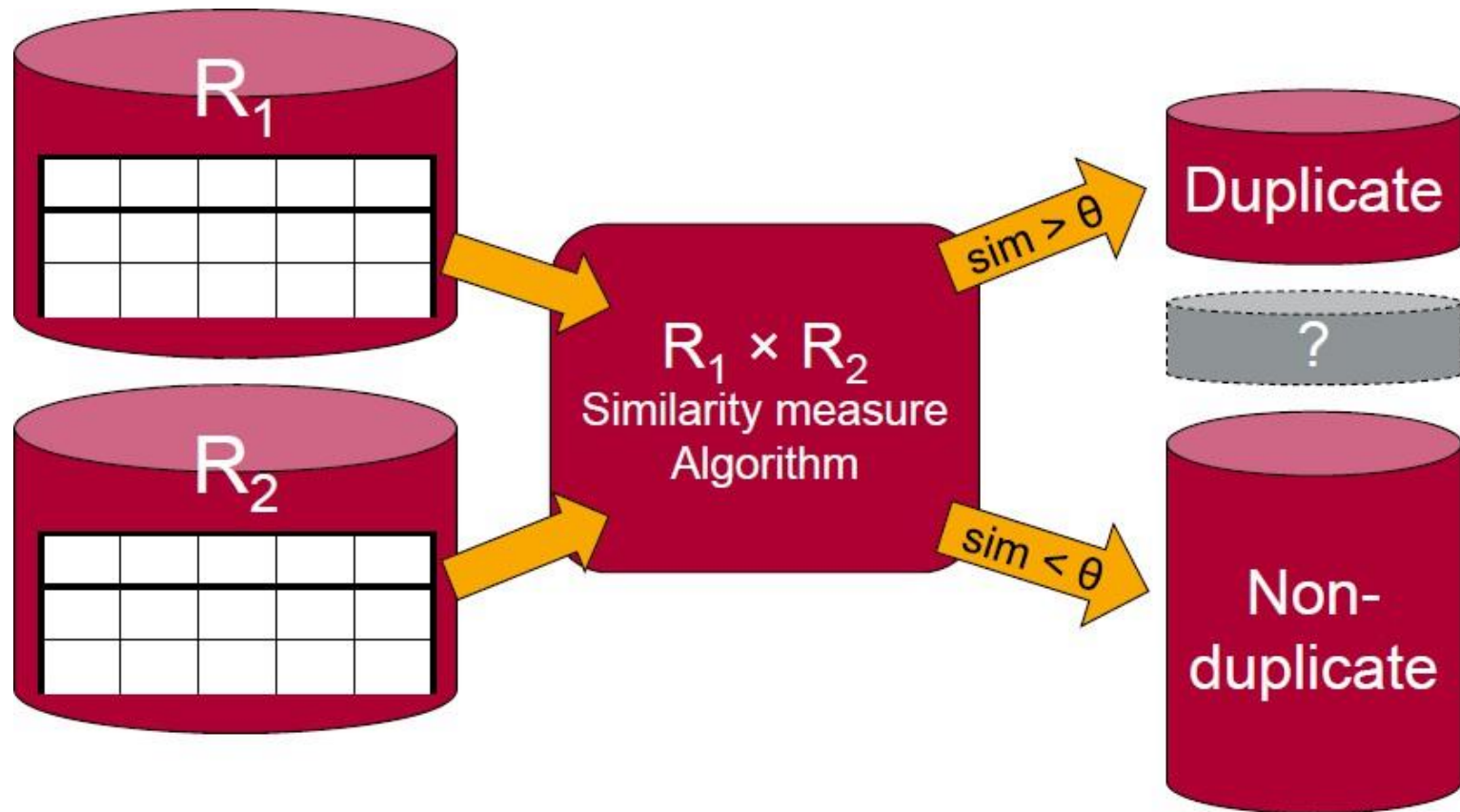
Duplikaterkennung

- Beispiele:

Tippfehler	Chinois on Main	2709 Main Street	Santa Monica
	Chniois on Main	2909 Main Street	Santa Monca
Abkürzungen	Foobar Holding	Flurstrasse 10	Oetwil an der Limmat
	Foobar Hldg	Flurstr. 10	Oetwil a. d. Limmat
Abweichende Bezeichnungen	Four Seasons	854 Seventh Avenue	New York City
	4 Seasons Grill Room	854 7th Ave. between 54th and 55th Sts.	New York
Untergeordnete Ortsbezeichnungen	Grill on the Alley	9560 Dayton Way	Los Angeles
	Grill on the Alley	9560 Dayton Way	Beverly Hills
Zweisprachigkeit	Déborah François	Quellgasse 4	Biel
	Déborah François	Rue de la Source 4	Bienne
Unvollständige Angaben	Grill on the Alley	9560 Dayton Way	Los Angeles
	Grill on the Alley	(null)	Beverly Hills
Vertauschte Werte	Thomas	Blake	Santa Monica
	Blake	Thomas	Santa Monica

Erkennung von Duplikaten

- Grundprinzip: Paarweiser Vergleich aller zu untersuchender Datensätze.



Erkennung von Duplikaten

- Grundprinzip: Paarweiser Vergleich aller zu untersuchender Datensätze.
- Problem: Performanz, Aufwand wächst **quadratisch** (d.h. mit $O(n^2)$).
- In der Regel geht es nicht um das Finden von exakt gleichen Datensätzen (das wäre trivial) sondern um das Erkennen von **Ähnlichkeiten**.
- Es werden also Verfahren benötigt, um die Ähnlichkeit von Datensätzen zu bestimmen, sogenannte **Ähnlichkeitsmasse**. Für diese sind geeignete Schwellwerte festzulegen («ab welcher Übereinstimmung ist es ein Duplikat?»)

Anforderungen

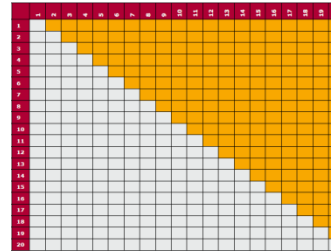
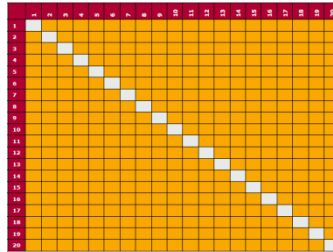
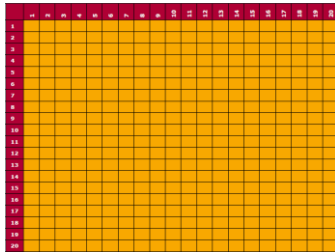
- Effektivität:
 - Effektivität (von lateinisch effectivus „bewirkend“) bezeichnet das Verhältnis von erreichtem Ziel zu definiertem Ziel (**Zielerreichungsgrad**).
 - Dies ist unabhängig vom zur Zielerreichung nötigen Aufwand. **Effektiv arbeiten bedeutet, so zu arbeiten, dass ein Ergebnis erreicht wird**, dass das gesteckte Ziel mindestens erreicht, möglichst sogar noch weit übererfüllt.
 - **Die Menge der erkannten Duplikate sollte möglichst der wirklichen Menge an Duplikaten entsprechen («alle erkennen»).**
 - Die Effektivität hängt von dem gewählten Ähnlichkeitsmass und dem gewählten Schwellwert ab.

Anforderungen

- Effizienz:
 - Verhältnis zwischen erreichtem Erfolg und dafür benötigten Mitteleinsatz. Das Ziel ist, **mit einem möglichst geringen Aufwand einen gegebenen Ertrag zu erreichen** oder mit einem gegebenen Aufwand einen möglichst grossen Ertrag zu erreichen.
 - Die Laufzeit des Verfahrens zur Erkennung von Duplikaten sollte mit der **Anzahl untersuchter Datensätze «vernünftig» skalieren**. Die Zeit für das Erhalten von Resultaten – auch bei der Untersuchung von Hunderttausenden von Datensätzen – sollte im Stunden-, nicht im Tagebereich liegen («schnell erkennen»).
 - Die Effizienz kann durch geeignete Auswahl der zu vergleichenden Datensätze stark beeinflusst werden.

Beurteilung von Verfahren

- Bei der Erkennung von Duplikaten muss jeder Datensatz mit jedem anderen «verglichen» werden. Dies bedeutet einen sehr grossen Aufwand ($\frac{n^2 - n}{2}$ Vergleiche). Bei einer Tabelle mit 10'000 Datensätzen entspricht das etwa 50 Mio. Vergleichen.



- Mögliche Resultate:
 - Korrekt erkannte Paare von Duplikaten
 - Aber auch:
 - Vorhandene Duplikate werden nicht als solche erkannt
 - Paare von Datensätzen werden fälschlicherweise als Duplikate erkannt

Beurteilung von Verfahren

- Falsch-positiv, falsch-negativ:
 - Bei einer Klassifizierung werden Objekte anhand von bestimmten Merkmalen durch einen Klassifikator in verschiedene Klassen eingeordnet. Der Klassifikator macht dabei im allgemeinen Fehler, ordnet also in manchen Fällen ein Objekt einer falschen Klasse zu.

Mögliche Ergebnisse der Duplikaterkennung			
		Realität	
		Duplikat	Nicht Duplikat
Methode	Duplikat	true-positive	false-positive
	Nicht Duplikat	false-negative	true-negative

Beurteilung von Verfahren

- Zur Messung der Güte eines Verfahrens werden zwei Masse (die aus dem Bereich data retrieval stammen) verwendet:

- **Precision**: Anteil an «true-positives» an allen von der Methode erkannten Duplikaten:

$$\text{precision} = \frac{|\text{true-positives}|}{|\text{true-positives}| + |\text{false-positives}|}$$

- Ein hoher Wert für «precision» bedeutet also, dass die gefundenen Duplikate mit hoher Wahrscheinlichkeit auch tatsächlich Duplikate sind. Hohe «precision» wird erreicht durch «strenge» Ähnlichkeitsmasse und hohe Schwellwerte.
- In der Regel kann den maschinell erhaltenen Ergebnissen bei hoher precision vertraut werden.

Beurteilung von Verfahren

- Zur Messung der Güte eines Verfahrens werden zwei Masse (die aus dem Bereich data retrieval stammen) verwendet:

- **Recall**: Anteil an «true-positives» an allen Duplikaten:

$$\text{recall} = \frac{|\text{true-positives}|}{|\text{true-positives}| + |\text{false-negatives}|}$$

Ein hoher Recall heisst, dass viele der tatsächlichen Duplikate gefunden wurden, allerdings möglicherweise auch viele Falsche. Ein hoher «recall» wird durch tolerante Ähnlichkeitsmasse und tiefe Schwellwerte erreicht.

- In der Regel müssen die Ergebnisse bei hohem recall manuell überprüft werden.

Beurteilung von Verfahren

- Um ein Verfahren zu beurteilen braucht man einen **Referenzdatenbestand** («Goldstandard») von dem genau bekannt ist, was Duplikate sind. Ein solcher Vergleichsdatenbestand ist aber fast nie verfügbar!
- Um den (quadratisch wachsenden) Aufwand zu reduzieren vergleicht man in der Praxis in der Regel nicht «alles mit allem».
- Man unterscheidet zwei Verfahren (die auch kombiniert zur Anwendung gelangen können):
 - Partitionierung
 - Sorted neighbourhood, windowing

Beurteilung von Verfahren

- Partitionierung: Aufteilen der zu untersuchende Menge an Datensätzen in **Partitionen** und nur Paare **innerhalb einer Partition** vergleichen.
- Diese auch als «blocking» bekannte Strategie verschlechtert tendenziell den Recall, da Duplikate möglicherweise über mehrere Partitionen verteilt liegen und damit nicht mehr gefunden werden. Die Wahl der richtigen Partitionierung ist deshalb von entscheidender Bedeutung.
- Beispiel: Vergleich von Adressen nur innerhalb von PLZ-Bereichen.

Beurteilung von Verfahren

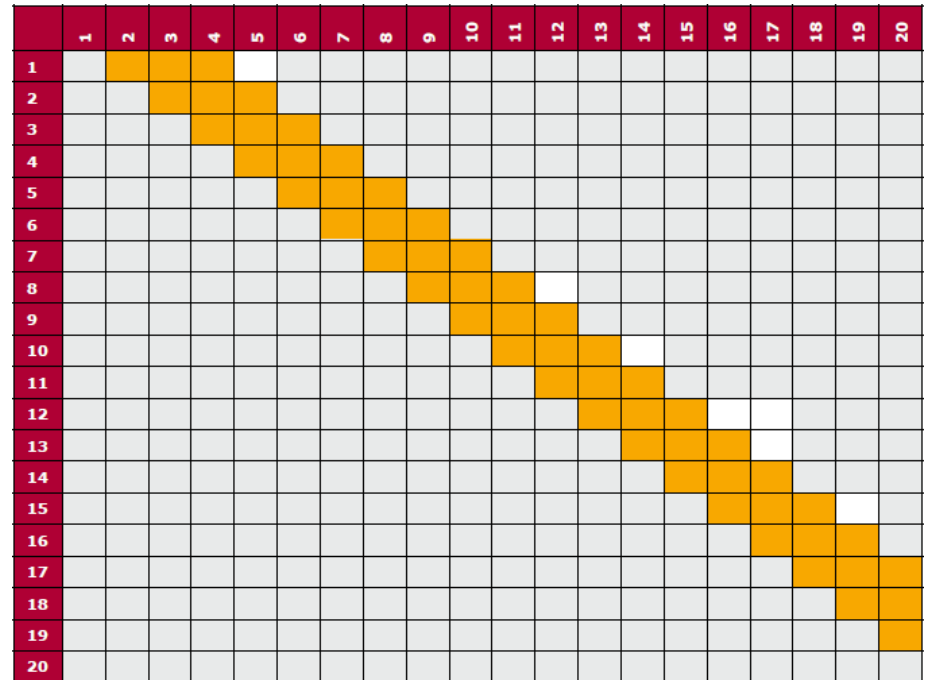
- Sorted neighbourhood, windowing: Man sortiert den Datenbestand zuerst so, dass ähnliche Datensätze näher beieinanderliegen. Dann verschiebt man ein 'Fenster' über den sortierten Datenbestand. Es sind zwei Probleme zu lösen:

- Sortierkriterium
- Fenstergrösse

- Literaturempfehlung:

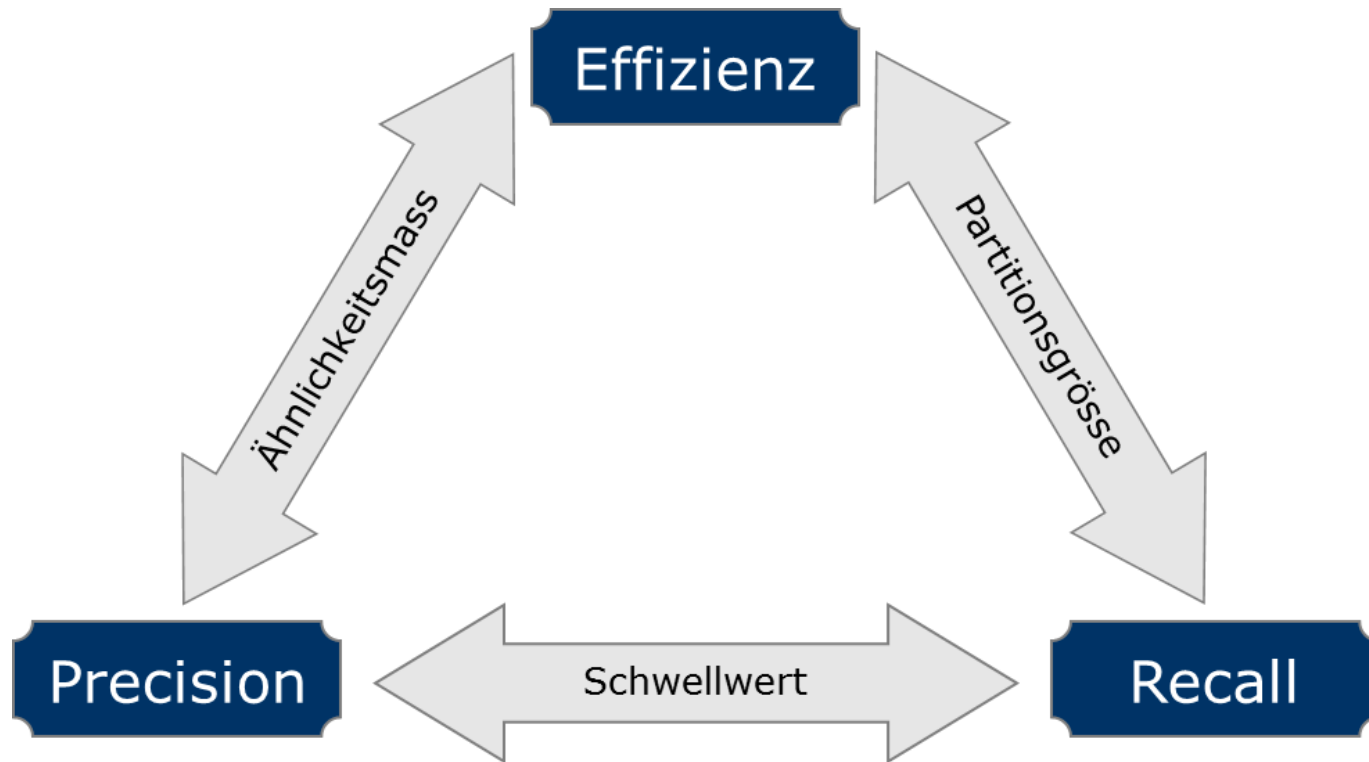


ISBN 978-3-8348-1772-3



Beurteilung von Verfahren

- Es liegen folgende Zielkonflikte vor:



Beurteilung von Verfahren

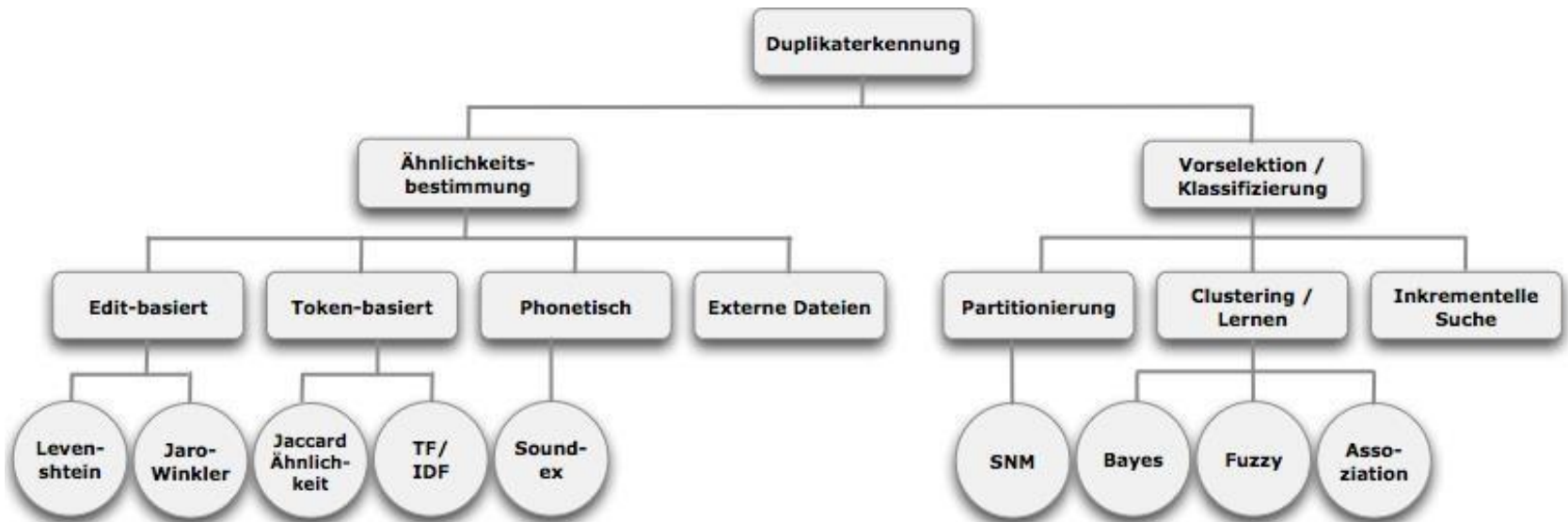
- **Ähnlichkeitsmass** und **Schwellwert** sind Parameter die sich direkt auf die Güte der Duplikaterkennung auswirken. «Tolerante» Methoden führen zu vielen «Treffern» aber auch vielen «Fehlalarmen» (und umgekehrt).
- Die zu wählenden Parameter (Ähnlichkeitsmass und Schwellwert) hängen somit stark von der Anwendung ab.

Ähnlichkeitsmasse

- Ein **Ähnlichkeitsmass** vergleicht **zwei Datensätze**. In der Regel wird ein Gesamtmasse für Datensätze aus Massen für die einzelnen Attributwerte zusammengesetzt.
- **Ähnlichkeitsmasse** sind in höchstem Masse **anwendungsabhängig**.
- Beispiel: Tipp-/Erfassungsfehler
 - Typischer menschlicher Fehler: **69** ↔ **96** ("Buchstabenverdrehen")
 - Typischer maschineller Fehler (OCR): **B**aumann ↔ **8**aumann
- Für viele Datentypen lassen sich in konkreten Anwendungen spezialisierte Ähnlichkeitsmasse definieren.

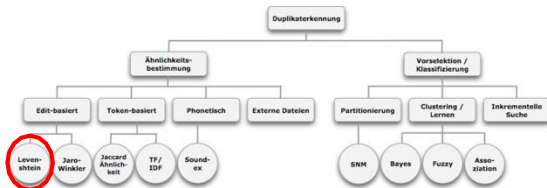
Ähnlichkeitsmasse

- Es sind viele verschiedene Masse bekannt:



Ähnlichkeitsmasse

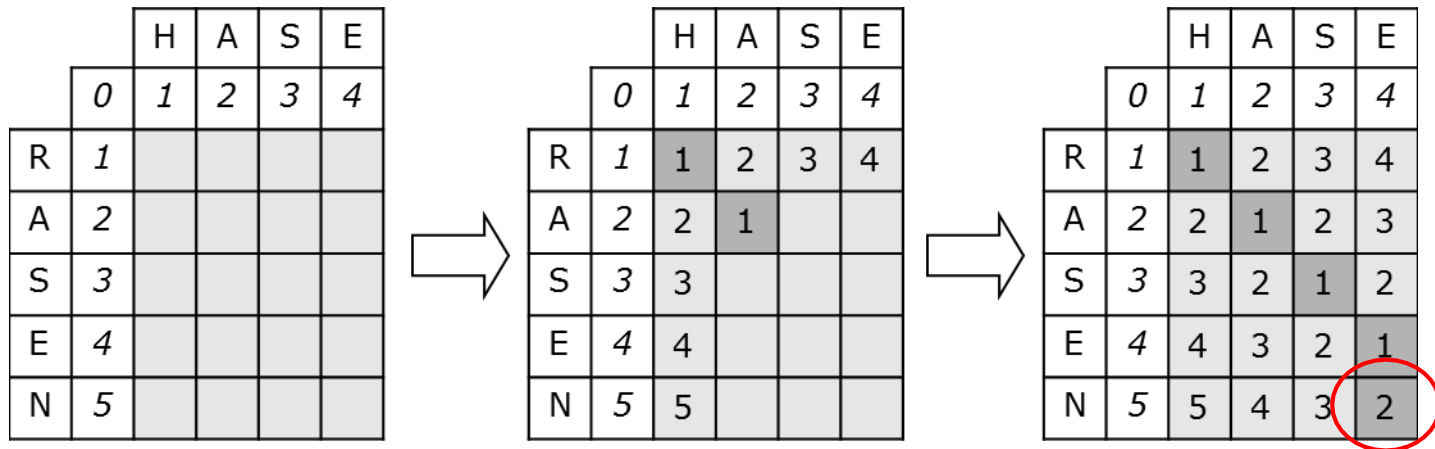
- Edit-basierte Masse:



- Buchstabenweiser Vergleich** zweier Zeichenketten. Je ähnlicher die Buchstabenmenge und ihre Anordnung innerhalb der Zeichenketten sind, desto höher ist die Gesamtähnlichkeit.

Edit-basiertes Mass: Levenshteindistanz

- Definition: Summe der minimalen Kosten aller benötigten Editier-Operationen (Einfügen, Löschen, Ersetzen), um eine Zeichenkette in eine andere zu überführen.
- Beispiel: Überführung von RASEN in HASE (R→H, N löschen)
(Kosten für Einfügen, Löschen und Ersetzen je 1):



Edit-basiertes Mass: Levenshteindistanz

- Levenshtein-Distanz-Berechnung zwischen s_1 und s_2 :
 1. Initialisiere Matrix M der Grösse $(|s_1| + 1) \times (|s_2| + 1)$
 2. Fülle Matrix: $M_{i,0} = i$ und $M_{0,j} = j$
 3. Rekursion:
$$M_{i,j} = \begin{cases} M_{i-1,j-1}, & \text{falls } s_1[i] = s_2[j] \\ 1 + \min(M_{i-1,j}, M_{i,j-1}, M_{i-1,j-1}), & \text{sonst} \end{cases}$$
 4. Levenshtein-Distanz: $M_{|x|,|y|}$

Edit-basiertes Mass: Levenshteindistanz

- Um die Edit-Distanz in eine Ähnlichkeitsmass umzuwandeln wird sie zunächst auf die Länge des längeren der beiden Strings normalisiert und anschliessend von 1 abgezogen:

$$\text{sim}_{\text{Levenshtein}}(S_1, S_2) := 1 - \frac{\text{ed}_{\text{Levenshtein}}(S_1, S_2)}{\max(|S_1|, |S_2|)}$$

- Damit entspricht ein Wert von 1 zwei identischen Strings und ein Wert von 0 bezeichnet zwei komplett unterschiedliche Strings.
- Es gibt zahlreiche Implementationen (auch in SQL & Python):

https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance

Edit-basiertes Mass: Levenshteindistanz

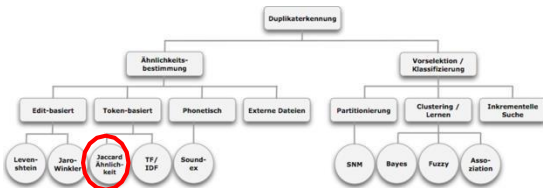
- Die Edit-Distanz **allein** ist aber als Ähnlichkeitsmass für zwei Tupel **ungeeignet**:

Hans	Muster	Bergstr. 17	<i>NULL</i>	29
Hans	Muster	Bergstr. 17	Angestellter	30

- Die Edit-Distanz der konkatenierten Werte ist 14
- Die Ähnlichkeit demnach $1 - 14/35 \sim 0.4$
- Edit-Distanzen werden deshalb üblicherweise in eine **Regelmenge** eingebettet, z.B.:
IF t1.nachname = t2.nachname
AND sim(t1.vorname, t2.vorname) > 0
AND t1.adresse = t2.adresse
THEN t1 ist Duplikat von t2

Ähnlichkeitsmasse

- Token-basierte Masse:



- «Wortweiser» Vergleich zweier Zeichenketten.
- Prüfen der Übereinstimmung ganzer Wörter.

Token-basiertes Mass: Jaccard-Distanz

- «Peter Müller» und «Müller, Peter» haben eine Edit-Distanz von 13
→ ungeeignet.
- Für zusammengesetzte Zeichenketten sind **token-basierte Masse** besser geeignet.
- Einfachste Möglichkeit zur Aufteilung in Token: durch Sonderzeichen (Leer- oder Satzzeichen, Bindestriche, etc.). Aber Achtung: Solche kommen allerdings auch in «Yahoo!», «.NET», «C#», «C++», ... vor!
- Andere Möglichkeit: Bildung von sogenannten **n-Grammen**, das sind Teilstrings der Länge n.

Beispiel: 3-Gramme des Wortes Rasen: {Ras, ase, sen}

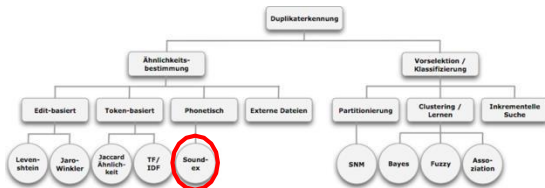
Token-basiertes Mass: Jaccard-Distanz

- Ein häufig verwendetes tokenbasiertes Ähnlichkeitsmass ist die sogenannte **Jaccard-Ähnlichkeit**. Sie vergleicht die Anzahl der gemeinsamen Token beider Strings mit der Anzahl aller Token der beiden Strings.
- Seien T_1 und T_2 die Tokenmengen der beiden Strings S_1 und S_2 :

$$sim_{Jaccard}(S_1, S_2) := \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

Ähnlichkeitsmasse

- Phonetische Masse:



- Indexieren von Worten basierend auf Aussprache.
- Phonetische Verfahren stellen keine Ähnlichkeits- oder Distanzfunktion dar, sondern bestimmen Mengen **ähnlich klingender Wörter**.

Phonetisches Mass: Soundex

- Soundex ist ein **phonetischer Algorithmus** zur **Indizierung** von Wörtern und Phrasen nach ihrem Klang in der englischen Sprache. Gleich klingende Wörter sollen dabei zu einer identischen Zeichenfolge kodiert werden. Entwickelt ~1918 (Patent) in den USA, zur Indizierung von Familiennamen bei der Volkszählung.
- Der **Soundex-Code** für ein Wort besteht aus seinem ersten Buchstaben, gefolgt von drei Ziffern, die die nach dem Anfangsbuchstaben folgenden Konsonanten des Wortes repräsentieren. Ähnliche Laute besitzen den gleichen Code (B, F, P und V werden z. B. alle mit der Ziffer „1“ codiert).
- Es gibt auch für die deutsche Sprache angepasste Versionen.

Phonetisches Mass: Soundex

- Jeder **Soundex-Code** besteht aus einem **Buchstaben** gefolgt von **drei Ziffern**, z. B. W213 für Wikipedia. Hat das zu codierende Wort so viele Buchstaben, dass man mehr Ziffern erzeugen könnte, bricht man nach der dritten Ziffer ab. Hat das Wort zu wenige Buchstaben, füllt man die letzten Ziffern mit Nullen auf.
- Es gilt:
 - Vokale sowie die Buchstaben «H», «W» und «Y» werden ignoriert.
 - Direkt aufeinander folgende Buchstaben mit gleicher Kodierung werden nur einmal verwendet.
 - Ist der Code kürzer als vier Zeichen, wird er mit Nullen aufgefüllt.
- Bsp: Meyer, Maier und Meier → «M600»

Ziffer	Repräsentierte Buchstaben
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Datenfusion

- **Datenfusion**, auch Ergebnis- oder Datenintegration, Deduplication, Duplicate Elimination und Record Merging, etc... ist der **zweite Schritt** nach der Duplikaterkennung.
- Ziel: **Kombination** von Duplikaten, so dass im Ergebnis kein Objekt der Realwelt mehr als einmal repräsentiert wird.
- Selektion von «Survivor» genügt nicht – es sollen ja nicht nur Duplikate entfernt, sondern möglichst Datensätze **angereichert** werden.
- Datensätze (innerhalb einer oder mehreren Quellen) können sich ergänzen oder widersprechen!

Datenfusion

- Beim paarweisen Vergleich können insgesamt folgende **vier Konfliktsituationen** auftreten:
 - **Gleichheit**: Die Tupel sind in allen Attributwerten gleich. T1 und T2 in der Abbildung sind gleich.
 - **Subsumption**: Ein Tupel subsummiert ein anderes, wenn es weniger NULLs hat und in jedem Attribut mit einem Nicht-NULL den gleichen Wert wie das andere Tupel besitzt. Es enthält also «mehr» Information. T1 und T2 subsummieren jeweils sowohl T3 als auch T4.

Film	ID	Titel	Regis.	Jahr	Studio
T1	1	Alien	Scott	1980	Fox
T2	1	Alien	Scott	1980	Fox
T3	1	Alien	Scott	1980	NULL
T4	1	Alien	NULL	1980	Fox
T5	1	Alien	Scott	1982	MGM

Datenfusion

- Beim paarweisen Vergleich können insgesamt folgende vier Konfliktsituationen auftreten:
 - **Komplementierung:** Ein Tupel komplementiert ein anderes, wenn keines der beiden das andere subsumiert und wenn es für jedes Attribut mit einem Nicht-NULL entweder den gleichen Wert wie das andere Tupel hat oder das andere Tupel an dieser Stelle ein NULL besitzt. Die beiden Tupel ergänzen sich. In der Abbildung komplementieren sich T3 und T4.

Film	ID	Titel	Regis.	Jahr	Studio
T1	1	Alien	Scott	1980	Fox
T2	1	Alien	Scott	1980	Fox
T3	1	Alien	Scott	1980	NULL
T4	1	Alien	NULL	1980	Fox
T5	1	Alien	Scott	1982	MGM

Datenfusion

- Beim paarweisen Vergleich können insgesamt folgende vier Konfliktsituationen auftreten:
 - **Konflikt:** In allen anderen Situationen stehen zwei Tupel in Konflikt. Bei Konflikten gibt es mindestens ein Attribut, in dem beide Tupel unterschiedliche Werte haben, die nicht NULL sind. T5 ist in Konflikt mit jedem anderen Tupel der Tabelle.

Film	ID	Titel	Regis.	Jahr	Studio
T1	1	Alien	Scott	1980	Fox
T2	1	Alien	Scott	1980	Fox
T3	1	Alien	Scott	1980	NULL
T4	1	Alien	NULL	1980	Fox
T5	1	Alien	Scott	1982	MGM

Datenkonflikte

- Verschiedene Strategien sind möglich:

Lösungsstrategie	Beschreibung
<i>Pass It On</i>	Konflikt weiterreichen (i)
<i>Consider All Possibilities</i>	Kombination aller möglichen Attributwerte (i)
<i>Take The Information</i>	Bevorzugung von nicht-NULL-Werten (v)
<i>No Gossiping</i>	Nur konsistente Resultate anzeigen (v)
<i>Trust Your Friends</i>	Priorisierung bestimmter Quellen (v)
<i>Cry With The Wolves</i>	Nach Erscheinungshäufigkeit (Wert) auswählen (a)
<i>Roll The Dice</i>	Zufällige Wahl (Wert) treffen (a)
<i>Meet In The Middle</i>	Durchschnittswert bilden (a)
<i>Nothing is older than yesterdays news</i>	Den jüngsten Wert wählen (a)

- (i) ignorieren; (v) vermeiden; (a) auflösen

Duplikate – Bemerkungen

- Duplikaterkennung und Fusion sind anspruchsvolle Tätigkeiten.
- Man muss sich in der Regel **iterativ** an eine genügend gute Lösung durch Experimente **herantasten**:
 - Wahl des Ähnlichkeitsmasses (ev. Kombination mehrerer)
 - Partitionierung («was mit was vergleichen»)
 - Umgang mit den Ergebnissen ist oft schwierig zu entscheiden
 - Es ist mit viel **Handarbeit** zu rechnen, insbesondere bei der Fusion!



Geleitete Übungen



Leistungsnachweis DB + DWH

- Wissen alle was Sie zu tun haben?
- Wissen alle mit wem Sie ggf. zusammenarbeiten?
- Termin: Abgabe bis spätestens am 25.11.
- Achtung: **Ergebnisse NICHT als Entwurf einreichen.**